VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

ÚSTAV SOUDNÍHO INŽENÝRSTVÍ

INSTITUTE OF FORENSIC ENGINEERING

SOFTWAROVÉ ŘEŠENÍ PRO TRŽNÍ KOMPARATIVNÍ OCEŇOVÁNÍ V REALITNÍ PRAXI

SOFTWARE SOLUTIONS FOR COMPARATIVE MARKET VALUATION IN THE REAL ESTATE PRACTICE

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Štěpán Skovajsa**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Vítězslava Hlavinková, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
ÚSTAV SOUDNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Student:	**Bc. Štěpán Skovajsa**
Studijní program:	Soudní inženýrství
Studijní obor:	Realitní inženýrství
Vedoucí práce:	**Ing. Vítězslava Hlavinková, Ph.D.**
Akademický rok:	2017/18
Ústav/odbor:	Ústav soudního inženýrství

Ředitel ústavu Vám v souladu se zákonem č. 111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Softwarové řešení pro tržní komparativní oceňování v realitní praxi

Stručná charakteristika problematiky úkolu:

Návrh vhodného grafického rozhraní potenciální aplikace (front-end). Stanovení kritérií selekce vhodných porovnávacích objektů. Vytvoření databázového modelu a serverové části (back-end). Naplnění databáze vzorovými objekty a ověření funkčnosti a otestování věrohodnost výsledku.

Cíle diplomové práce:

Komplexní návrh a realizace webové aplikace pro porovnávání bytů v realitní praxi a také návrh centrálního uložiště pro srovnávací objekty.

Seznam literatury:

BRADÁČ, A.; a kol. Teorie a praxe oceňování nemovitých věcí, první vydání. Brno: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o., 2016, 790 p. ISBN 978-80-7204-930- 1.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2017/18.

V Brně, dne 20. 10. 2017

L. S.

doc. Ing. Aleš Vémola, Ph.D.
ředitel

Abstrakt

Tato diplomová práce pojednává o tržní komparativní metodě, statistickém zpracování a ve své praktické části se snaží nastínit, jak lze využít výpočetní techniku pro automatizaci získávání srovnávacích nemovitostí a navrhuje algoritmus pro rychlé ocenění. V závěru praktické části porovnává výstupy z programu nejen s inzeráty, ale i se skutečně zobchodovanými nemovitostmi.

Abstract

This diploma thesis is about real estate market comparative appraisement, statistical approach and how to use software for data gathering and processing in its practical part. Practical part also includes draft of appraising algorithm. In the end of practical part, there are check of application's output and comparison of computed prices to both real traded prices and advertisement prices.

Klíčová slova

realitní trh, reality, nemovitosti, byty, software, aplikace, oceňování, cena, tržní cena, porovnání, databáze, klient, server, geocoding, google maps, javascript, node, react, graphql

Keywords

real estate market, real estates, real properties, flats, software, application, appraisal, price, market price, comparation, database, client, server, geocoding, google maps, javascript, node, react, graphql

Bibliografická citace

SKOVAJSA, Š. Softwarové řešení pro tržní komparativní oceňování v realitní praxi. Brno: Vysoké učení technické v Brně, Ústav soudního inženýrství, 2018. 71 s. Vedoucí diplomové práce Ing. Vítězslava Hlavinková, Ph.D.

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje.

V Brně dne 24. 5. 2018

……

Podpis diplomanta

Poděkování

Touto cestou bych velmi rád poděkoval Ing. Vítězslavě Hlavinkové, Ph.D. za její čas, trpělivost, ochotu a praktické poznámky ke zpracování této práce.

OBSAH

1 ÚVOD

Oceňování je činnost stará asi jako samy peníze. Lidé se od pradávna snažili ocenit různé druhy majetku (dobytek, práci, jídlo, materiál) kvůli různým účelům (obchodování, později odvod daně, vyměřování sankcí). Zatím však stále neexistuje a zřejmě nikdy existovat ani nebude absolutně objektivní metodika, která by dokázala reflektovat všechny aspekty potřebné pro ocenění včetně všech zájmů zúčastněných, protože to ani není reálné. Zpravidla chce kupující co nejvíce ušetřit a prodávající co nejvíce profitovat. Je třeba stanovit „spravedlivou" cenu a za tu se pokládá ta tržní. Na to se v praxi používají empirické nebo statisticky aproximované postupy, které nám alespoň co nejvíce možně promítnout „spravedlivou" cenu, a přitom nejsou příliš složité nebo zdlouhavé. A právě složitost a zdlouhavost je problém, který se v poslední době, s nástupem moderní informační infrastruktury, začal odbourávat. V dnešní době, kdy je hardware i software dostupný běžnému smrtelníkovi, lze i banální problémy řešit rychleji, přesněji a jednodušeji.

Jelikož mě zajímá chování trhu s nemovitostmi v širší geografické rovině a mám pracovní zkušenosti s vyvíjením aplikací a softwaru celkově, rozhodl jsem se tyto oblasti propojit a vytvořit software, který dokáže oceňovat nemovitosti. Cílem však není vytvořit nejpřesnější algoritmus oceňování nemovitostí na světě, ale pochopit a navrhnout strojové zpracování této úlohy a prezentovat zde cestu k tomuto poznání.

1.1 CÍLE PRÁCE

Cílem této práce je vytvořit algoritmus pro oceňování bytů a následně tento algoritmus implementovat a vytvořit tak software, který dokáže, dle zadaných parametrů, ocenit nemovitost. Je tedy zapotřebí uvést, kromě teorie realitního inženýrství, i trochu teorie z informačních technologií, která dopomůže k lepšímu pochopení textu. Nebudu zde však uvádět žádné konkrétní implementace, ani ukázky kódu, nýbrž diagramy popisující jednotlivé algoritmy.

Budu zde vycházet spíše z obecných předpokladů, statistiky a vlastního úsudku než z legislativy, která sice nabízí cenové předpisy, nicméně nemusí dostatečně reflektovat tržní

situaci, protože to ani není jejím účelem. Vše se však budu snažit v co největší míře logicky zdůvodnit.

Pro jednoduchost jsem se omezil pouze na srovnávaní bytů v rámci Jihomoravského kraje, nicméně nebyl by problém algoritmus upravit na zpracování nemovitostí v obecnější rovině a v širší geografické působnosti.

2 TRH A JEHO VLIV NA VNÍMÁNÍ CENY A HODNOTY

2.1 TRH

Trh je obecně prostředí, kde dochází ke směně zboží a služeb mezi nakupujícím a prodávajícím. Je samozřejmé, že do tržního prostředí vstupuje i stát, a to především tak, aby chránil zájmy jak zúčastněných, tak i zájmy své (např. výběr daně).

Dříve se ve velké míře praktikoval barter, tj. směnný obchod, kdy se zboží či služba vyměnila za jiné zboží či službu. Problém byl však v tom, že ne vždy si obě strany vyhověly svou nabídkou, resp. ne vždy disponovala jedna ze stran tím, o co měla zájem druhá strana. Proto se později začaly používat peníze a bylo tedy potřeba kvantifikovat finanční sumu, která reflektuje hodnotu daného zboží nebo služby. Vznikl tak obor oceňování majetku.

2.2 CENA VERSUS HODNOTA

V tržním ocenění je mezi cenou a hodnotou velmi tenká bariéra. Pod cenou si většinou představujeme peněžitou částku, která je určena pro prodej či nákup dané věci. Naproti tomu hodnota je vnímána více subjektivně, protože stejné věci mohou rozdílné osoby přiřazovat odlišnou, mnohdy markantně odlišnou, hodnotu. Ve spoustě případů tedy hodnota reflektuje užitek vlastníka, resp. uživatele, dané věci. Například můj pes, kterého mám velmi rád, má pro mě nevyčíslitelnou hodnotu, kdežto pro souseda, který je rušen jeho štěkáním, bude mít takřka nulovou hodnotu. Jeho cena, pokud si prohlédneme inzerci prodeje štěňat, však bude víceméně stálá, protože prodejci se zpravidla snaží získat co největší marži, aby maximalizovali svůj užitek, ale zároveň musí reflektovat poptávku trhu. Musí tedy cenu nastavit tak, aby jim pokryla náklady, marži a zároveň bylo zboží cenově dostupné pro potenciální kupce.

2.3 DRUHY CEN

2.3.1 Cena zjištěná (administrativní, úřední)

Cena stanovená dle cenového předpisu za účely stanovených zákonem, což je například výběr daně, dědické řízení, vyvlastnění, konkurz firmy apod.

2.3.2 Cena pořizovací (historická)

Cena, za kterou byla daná věc pořízena v době jejího pořízení. U nemovitostí to může být i náklad, za který byla nemovitost postavena. Zpravidla tato cena zahrnuje i náklady na pořízení. U koupené nemovitostí by to mohl být například poplatek realitní kanceláři za zprostředkování obchodu a právních služeb.

2.3.3 Cena reprodukční (reprodukční pořizovací cena)

Částka, za kterou se daná nebo podobná věc dala pořídit v době ocenění. Nepočítá se zde s opotřebením. U nemovitostí se zpravidla určuje reprodukcí vynaložených nákladů nebo pomocí technickohospodářských ukazatelů (THU).

2.3.4 Cena obvyklá (obecná, tržní)

Je zpravidla cena, která nás zajímá nejčastěji. Jedná se totiž o cenu, za kterou běžně nakupujeme věci v tržním prostředí. Reflektuje nabídku a poptávku, tedy reálné chování trhu. Tato cena je tedy jakýmsi průměrem, za který by byl schopný průměrný prodávající prodat danou nemovitost průměrnému kupujícímu, což znamená, že se do této ceny nepromítají mimořádné vlivy a okolnosti (např.: prodávající a kupující jsou rodina, prodávající je v tísni apod.). (1)

Nabídkové tržní ceny

Jsou ceny udávané nabídkovou stranou trhu. V případě nemovitostí se jedná o ceny, za které nám realitní kanceláře a jiní prodejci nabízejí nemovitosti. Pokud je trh přehlcen, mezní užitek klesá a tím i ochota potenciálních kupců nakupovat, což v konečném důsledku nutí

prodávající (resp. stranu nabídky) snižovat cenu. Pokud je naopak na trhu daná komodita exkluzivní, má zpravidla vyšší hodnotu a pokud je dostatečně poptávaná, tak i cenu.

Nevýhodou tedy je, že nemusí adekvátně reflektovat poptávku trhu a často pak nabídková cena nemovitosti s postupem času klesá, dokud se neobjeví zájem ze strany poptávky. Naopak výhodou nabídkových cen je snadná dostupnost.

Realizované tržní ceny

Realizované ceny jsou ceny, za které skutečně proběhl prodej nebo pronájem nemovitostí. Oproti nabídkovým cenám však nejsou dost dobře dostupné, protože nepatří zpravidla mezi veřejně dostupné informace.

2.4 DRUHY HODNOT

2.4.1 Věcná hodnota (časová cena)

Je v podstatě reprodukční cena snížená o adekvátní opotřebení, které odpovídá průměrnému opotřebení podobné věci, eventuálně ještě o náklady na opravy, které věc uvedou do znovupoužitelného stavu.

2.4.2 Výnosová hodnota

Je produktem výnosové metody ocenění, což znamená, že vychází z potenciálních budoucích čistých výnosů, které jsou diskontovány. (1)

3 OCEŇOVÁNÍ POROVNÁVACÍM (KOMPARATIVNÍM) ZPŮSOBEM

Oceňování chápeme jako přiřazování určitému předmětu ocenění peněžitou částku, za kterou prodejce nabízí, kupec nakupuje nebo slouží k takovým účelům, jako například odvod daně. Pro oceňování nemovitostí se používají v zásadě 3 metody, případně jejich kombinace:

- Nákladové (výdajové) metody – předmět ocenění je oceněn dle nákladu, které byly do něj vloženy

- Výnosové (příjmové) metody – předmět ocenění je oceněn na základě budoucích benefitů, které z něj plynou
- Porovnávací (komparační) metody – předmět ocenění je oceněn na základě porovnání se stejnými nebo podobnými předměty (1)

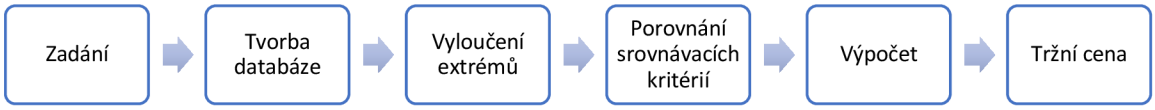
3.1 SPECIFIKA OCEŇOVÁNÍ NEMOVITOSTÍ

Nemovitosti, na rozdíl od sériových výrobků, jsou specifické svou unikátností a navzájem velmi rozdílnými parametry, což jejich ocenění komplikuje. Nejzásadnější rozdíl mezi nemovitostmi a ostatními komoditami je, že nemovitosti jsou nepřemístitelné. Z toho plyne, že zeměpisná poloha má nejzásadnější vliv. Má totiž vliv i na další parametry jako například atmosférické podmínky (vítr, déšť), bezpečnostní podmínky (kriminalita), kulturní podmínky, sociální podmínky, dopravní dostupnost, goodwill dané lokality. Z toho vyplývá, že v milionářské čtvrti New Yorku bude bez pochyby cena za m^2 z prodeje bytu větší než na okraji Bangladéše. Při oceňování nemovitostí porovnávacím způsobem je tedy nutné pamatovat na to, aby srovnávací nemovitosti pocházeli ze stejné, nebo poměrově podobné lokality. (1)

3.2 POROVNÁVACÍ ZPŮSOB A NEMOVITOSTI

Nemovitost, u niž známe parametry, ale neznáme cenu, proto ji zjišťujeme se nazývá oceňovaná nemovitost. Nemovitost, u niž známe cenu i ostatní její parametry a ze které vycházíme při porovnání vlastností a cen, nazýváme srovnávací nemovitost. Obecně by mělo platit, že čím více informací o nemovitostech máme, tím jsme lépe schopni dosáhnout přesnějšího výsledku.

Výsledná tržní cena oceňované nemovitosti TCO tedy vychází z tržních cen srovnávacích nemovitostí TCS_i. Postup ocenění můžeme znázornit tímto diagramem:



Obrázek 1 - postup ocenění porovnávacím způsobem (vlastní)

Na začátku tedy, jakmile víme, co oceňujeme, vytvoříme databázi podobných nemovitostí v okolí. Může se však stát, že si do databáze zaneseme srovnávací nemovitost, do

jejíž ceny se promítá nějaká mimořádná okolnost, která cenu zkresluje a nemovitost je pak buď podceněna nebo „nadceněna“. Takovéto extrémy je potřeba vyloučit, jinak bychom nedosáhli tržní ceny, do které by se neměly promítat *zvláštní vlivy*. Když už máme finální podobu databáze, můžeme přistoupit k samotnému porovnání. Porovnáváme nemovitosti na základě srovnávacích kritérií, které vyjadřují, jak moc je daná vlastnost oproti oceňované nemovitosti lepší nebo horší.

3.3 DATABÁZE NEMOVITOSTÍ

Databáze slouží jako podklad pro ocenění, kdy výslednou cenu stanovíme na základě cen srovnávacích nemovitosti. Srovnávací nemovitosti v databázi by měly splňovat následující podmínky:

- měly by se nacházet v okolí oceňované nemovitosti
- měly by být podobného typu a vlastností
- známé tržní ceny TCS_i by měly časově spadat k datu ocenění a měly by být věrohodné (nejideálněji ceny ze skutečně realizovaných obchodů)
- výměra srovnávací nemovitosti VS_i musí být ve stejných jednotkách, aby seděla co do rozměru i jednotková tržní cena $JTCS_i$ vypočtena ze vztahu (1):

$$\boldsymbol{JTCS_i = \frac{TCS_i}{VS_i}} \tag{1}$$

3.4 POROVNÁNÍ VLASTNOSTÍ A VÝPOČET TRŽNÍ CENY

Nemovitosti nejsou stejné, proto musíme jejich odlišnosti zohlednit. To provádíme tak, že si zvolíme několik srovnávacích kritérií a ty pak mezi sebou porovnáme. Jako kritéria u nemovitostí volíme například:

- druh a účel objektu
- poloha objektu
- vnitřní dispozice
- materiály (zpravidla u prvků dlouhodobé životnosti)
- stáří (resp. doba od poslední rekonstrukce)
- dopravní dostupnost

3.4.1 Koeficient odlišnosti

Koeficient odlišnosti vyjadřuje míru odlišnosti jedné konkrétní vlastnosti mezi oceňovanou a srovnávací nemovitostí. Označujeme jej K_i pro kritérium i. Pokud je roven 1, je daná vlastnost obou nemovitostí na stejné úrovni. Je nutné na začátku stanovit konvenci, která nám řekne, zda nabývá hodnot větších než 1, pokud je vlastnost srovnávací nemovitosti lepší než oceňované nebo tomu bude naopak, a této konvence se držet v průběhu celého výpočtu. (1)

3.4.2 Index odlišnosti

Vyjadřuje podíl více vlastností, resp. zahrnuje vliv několika koeficientů odlišnosti na výsledný rozdíl v ceně. Označujeme jej I_j pro nemovitost j. Můžeme jej tedy spočítat jako prostý součin, kde n je počet srovnávacích kritérií:

$$I_j = \prod_{i=1}^{n} K_i \tag{2}$$

U takto vypočítaného indexu odlišnosti předpokládáme, že má každé kritérium stejnou váhu. To samé by platilo i pro vztah daný aritmetickým průměrem:

$$I_j = \frac{\sum_{i=1}^{n} K_i}{n} \tag{3}$$

Pokud bychom chtěli zohlednit váhu jednotlivých kritérií, mohli bychom použít vztah pro vážený průměr, kde w_i je váha kritéria:

$$I_j = \frac{\sum_{i=1}^{n} K_i w_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

3.4.3 Rozdělení metod dle počtu kritérií

Monokriteriální

K porovnání se používá pouze jedno kritérium. U nemovitostí by však takovéto ocenění mohlo být velmi nepřesné. Proto se spíše uchylujeme k využití multikriteriálních metod.

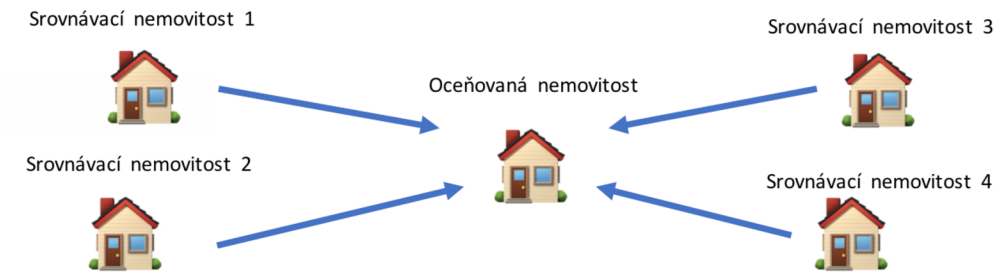
Multikriteriální

K porovnání se využije vícero kritérií. Zde je důležité, abychom znali hodnotu kritéria pro oceňovanou i srovnávací nemovitosti, jinak takové kritérium nemůžeme pro ocenění využít. (1)

3.4.4 Rozdělení metod dle postupu

Metoda přímého porovnání

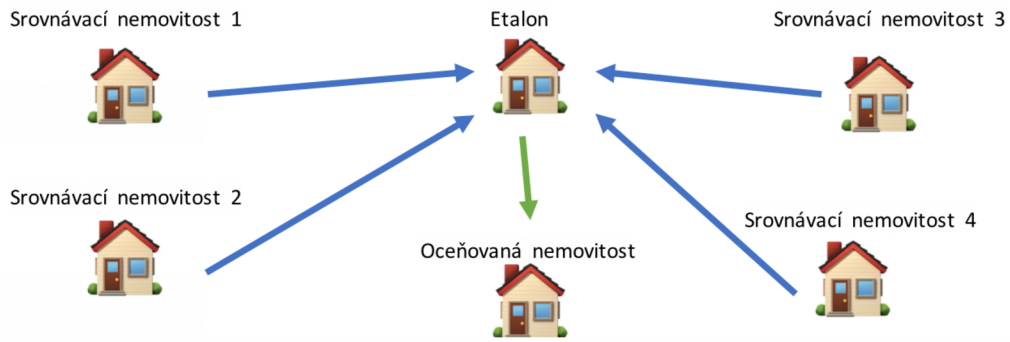
Spočívá v přímém porovnání vlastností srovnávacích nemovitostí s vlastnostmi oceňované nemovitosti.



Obrázek 2 - metoda přímého porovnání (vlastní)

Metoda nepřímého porovnání

Oproti přímé metodě obsahuje mezikrok, kdy se ze srovnávacích nemovitostí vytvoří jakýsi standard (etalon) a teprve s ním je oceňovaná nemovitost porovnána. Je vhodná zejména pro oceňovaní více typově podobných oceňovaných nemovitostí. (1)



Obrázek 3 - metoda nepřímého porovnání (vlastní)

4 VYUŽITÍ STATISTICKÉHO PŘÍSTUPU

4.1 MOTIVACE

Vím, že existuje mnoho přístupů pro oceňování nemovitostí, některé vycházejí z empirie, některé z bližšího zkoumání konkrétního trhu. Každý odhadce může volit odlišné metody a přijít k podobnému závěru nebo taky nemusí. Vše závisí na náhodě, kdy, kdo a kde provede dané ocenění a jakých metod využije. 10 odhadců může mít 10 názorů, což sice není nic podivuhodného, protože zde hraje roli i určitá míra subjektivity, ale může to být v některých situacích matoucí. Zvláště v dnešní době, kdy jsou události kolem nás čím dál rychlejší, je těžší se v čím dál větším proudu dat orientovat. Ceny mohou dosahovat extrémnějších výkyvů vlivem nestálosti trhu, krizemi, změnami úrokových sazeb, a proto je i aktualizace oceňovacích přístupů nesnadným úkolem.

Statistika je vědou o sběru, organizování a interpretací údajů, které nazýváme data. Metody statistiky umožňují rozlišovat pravdivé výroky od nepravdivých na základě pravděpodobnosti, vychází z matematické statistiky, která je větví aplikované matematiky. Jedná se tedy o zobecněné modely chování různých náhodných jevů. (2) Díky metodám statistiky (a následně informačních systému) budeme schopni:

- Vybrat data (populaci) pro porovnání
- Popsat data, vytvořit grafy závislostí veličin
- Vyloučit nevalidní data (extrémy)
- Vytvořit a popsat matematický model ocenění – aby byl strojově automatizovatelný
- Posoudit přesnost matematického modelu ocenění

Zároveň statistickým přístupem docílíme deterministočnosti, tzn. pro stejný vstup, při stejných podmínkách, dostaneme pokaždé stejný výstup, takže z modelu se vytratí prvek náhody.

4.2 ZÁKLADNÍ POJMY

Populace neboli *základní soubor* reprezentuje množinu všech dat, která odpovídají nám zadaným kritériím. Pokud například řekneme, že zkoumáme tržní cenu garáží ve Zlíně, měli

bychom brát v úvahu jako populaci všechny nedávno proběhlé transakce s těmito typy nemovitostí ve Zlíně.

Výběrový soubor neboli *výběr* naproti tomu reprezentuje množinu jedinců, kteří odpovídají specifičtějšímu kritériu. Pokud nás bude zajímat cena konkrétní garáže na Lesní čtvrti ve Zlíně, budou pro nás výběrovým souborem okolní garáže v této oblasti, které mají podobné vlastnosti jako garáž, kterou chceme ocenit, resp. jejich obchodované ceny.

Statistická jednotka je jedinec, dále nedělitelná entita ve výběrovém souboru, resp. v populaci. V našem případě by to byla naposledy zobchodovaná cena konkrétní garáže na Lesní čtvrti ve Zlíně.

Můžeme vyjádřit vztah mezi *populací*, *výběrem* a *jednotkou* matematicky. Výběr je podmnožinou populace. *Statistická jednotka* je podmnožinou *výběru*, a protože je tato relace tranzitivní, je *statistická jednotka* zároveň podmnožinou *populace*.

Aby toho však nebylo málo, máme tu ještě *statistický znak* neboli *náhodnou veličinu*, která vyjadřuje sledovanou míru, resp. hodnoty, kterých nabývá. Právě podle hodnot, kterých nabývá ji můžeme rozdělit na:

- Kvantitativní – vyjadřují počet, nabývá spojitých nebo diskrétních hodnot (např. velikost garáže v m^2, vzdálenost do centra města v km), jedná se tedy o číselné údaje
- Kvalitativní – nabývají pouze diskrétních hodnot velmi specifických intervalů (např. barva fasády; materiál – dřevo, cihla, beton; logické závěry ano/ne - např.: zda je v garáži přípojka i na 320 V), jedná se tedy o výčty (enumerátory – viz dále Datové typy) (2)

4.3 CHARAKTERISTIKY NÁHODNÉ VELIČINY

Náhodné veličiny ve výběru a populaci lze popisovat hromadně mnoha způsoby, nicméně všechny veličiny se dají rozdělit v podstatě do 2 kategorií, a to *míry polohy* a *míry variability* (*rozptýlenosti*).

4.3.1 Míry polohy

Výběrový průměr

Nazývaný taktéž jako aritmetický průměr, označovaný většinou $\bar{x}$, je součet všech hodnot vydělený počtem hodnot n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

reprezentuje tedy takovou hodnotu, která by byla společná zachování stejného počtu jedinců a při stejné sumě:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{6}$$

Nedostatek průměru je, že špatně reprezentuje nehomogenní skupiny, zvlášť u malého výběru srovnávacích nemovitostí s velkou odchylkou od velikosti (např. předražené malé byty budou průměrnou jednotkovou cenu značně zvedat). (2)

Střední hodnota

Nazývá se taktéž populační průměr a je váženým aritmetickým průměrem daného výběru s předpokladem, že obdobně se chová i populace. Proto se tedy používá tam, kde nás zajímá průměr z celé populace, avšak nemáme dostatek dat, abychom prostý aritmetický průměr z populace determinovat. Lze jej určit ze vztahu (3):

$$\bar{x} = \frac{\sum_{j=1}^{k} n_j * \bar{x}_j}{n} \tag{7}$$

kde n je celkový počet prvků, k je počet tříd, n_j je počet prvků v j-té třídě a $\bar{x}_j$ je charakteristická hodnota pro danou třídu. Z toho můžeme vyvodit relativní četnost třídy j jako n_j/n.

4.3.2 Míry variability (rozptýlenosti)

Variační rozpětí

Nejjednodušší míra rozptýlenosti, reflektuje pouze rozdíl maxima a minima. Je tedy nejhrubší mírou variability (3):

$$\boldsymbol{R = x_{max} - x_{min}} \tag{8}$$

Rozptyl

Rozptyl je důležitou charakteristikou variability a doplňuje průměr. Jak jsme si již řekli, průměr špatně reflektuje nehomogenní skupiny, díky rozptylu tedy jsme schopni určit, jak moc nesourodá náhodná veličina je (2). Rozptyl populace s^2 určíme ze vztahu:

$$\boldsymbol{s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2} \tag{9}$$

Ze vzorce je tedy patrné, že se jedná o průměr všech kvadrátů odchylek od aritmetického průměru. Pro výběr používáme upravený vztah (3):

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2} \tag{10}$$

Směrodatná odchylka

Vzhledem k tomu, že rozptyl je průměr kvadrátů, vyvstala potřeba mít veličinu, která bude ve stejných jednotkách, jako průměr a jiné veličiny. Směrodatná odchylka je tedy odmocněný rozptyl. Díky tomu získáme veličinu, která je ve stejném měřítku jako průměr. Označujeme ji prostě s a spočítáme ji (2):

$$\boldsymbol{s = \sqrt{s^2}} \tag{11}$$

Za s^2 dosazujeme rozptyl buď výběrový nebo populační v závislosti na tom, kterou směrodatnou odchylku potřebujeme.

4.4 VYLOUČENÍ EXTRÉMŮ (OUTLIERS EXCLUSION)

4.4.1 Motivace

Extrémní hodnoty jsou hodnoty, které se velmi liší od průměrných hodnot a mohou nám zkreslovat měření. Měli by se proto vyskytovat vzácně. V populaci, pokud je dostatečně velká a homogenní, nemají téměř žádný vliv, nicméně v malých výběrech, se kterými zpravidla při oceňování pracujeme, mohou například zvětšovat rozptyl a tím pádem i způsobovat nepřesnost regresní analýzy.

Příliš drahé (luxusní) nebo naopak příliš levné (rozpadlé nebo jinak neobyvatelné) nemovitosti nám mohou značně zkreslovat výsledek a tím pádem tyto musíme z výběru vyloučit. Při oceňování tímto softwarem tedy předpokládáme, že oceňujeme běžný uživatelný byt v přiměřené hodnotě.

4.4.2 Grubbsův test

Předpokladem Grubbsova testu je fakt, že testovaný soubor odpovídá Gaussovu (normálnímu) rozdělení pravděpodobnosti. Grubbsův test se vyznačuje iterativním postupem, tzn. je třeba jej opakovat, dokud není splněna podmínka, že nevylučujeme žádnou z hodnot (viz dále).

Postup jedné iterace je následovný:

1) Vypočítá se aritmetický průměr $\overline{x}$ a střední hodnota s ze všech testovaných hodnot souboru
2) Vypočítáme testovací kritéria pro maximální a minimální testovanou hodnotu. Výpočet testovacího kritéria T_i pro i-tou hodnotu:

$$\boldsymbol{T_i = \frac{|x_i - \overline{x}|}{s}} \tag{12}$$

3) Vypočteme (nebo dohledáme v tabulce) kritickou hodnotu T_{crit}:

$$\boldsymbol{T_{crit} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{(\frac{\alpha}{n}, n-2)}}{n-2+t^2_{(\frac{\alpha}{n}, n-2)}}}} \tag{13}$$

kde:

n … počet hodnot v souboru

$t^2_{(\frac{\alpha}{n},n-2)}$… kritická hodnota pro Studentovo rozdělení umocněná 2

Ze vzorce také vyplývá, že $n \geq 3$, protože pro Studentovo rozdělení není nižší stupeň volnosti než 1.

4) Porovnáme testovací kritéria T_i s kritickou hodnotou T_{crit} a platí:
   - $T_i > T_{crit} \Longrightarrow$ hodnotu ze souboru vyloučíme a provedeme test ještě jednou od druhého kroku. Musíme mít při tom na paměti, že je třeba znovu vypočítat průměr a směrodatnou odchylku již bez vyloučené hodnoty.
   - $T_i \leq T_{crit} \Longrightarrow$ hodnotu nevylučujeme. Test je u konce, pokud podmínka projde pro minimální a zároveň i pro maximální hodnotu.

4.5 REGRESNÍ ANALÝZA

Regresní analýza patří k nejpoužívanějším metodám statistické analýzy vícerozměrných dat. Slouží v podstatě k nalezení obecného výpočetního modelu, kdy na základě známých veličin, které nazýváme *vysvětlující proměnné* (*regresory*) se snažíme predikovat neznámou hodnotu (*vysvětlovaná proměnná* neboli *odezva*). V případě multikriteriální lineární regresní analýzy budeme vycházet ze základního vztahu:

$$\boldsymbol{Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e} \tag{14}$$

Kde Y je vysvětlovaná proměnná, β_1, β_1, … β_k jsou neznámé parametry, které se snažíme s pomocí regresní analýzy zjistit, X_1, X_1, … X_k jsou regresory a e je náhodná chybová složka. V tomto případě se dá tvrdit, že β_j je váha, která určuje, jaký má náhodná veličina X_j na vysvětlovanou proměnnou vliv. Vzhledem k tomu, že náhodnou chybovou složku e dopředu neznáme, budeme ji u predikované hodnoty $\hat{y}$ zanedbávat:

$$\boldsymbol{\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} \tag{15}$$

resp. zjednodušeněji pomocí funkce:

$$\boldsymbol{\hat{y} = f(x_1, x_2, \ldots\ x_k) = f(x)} \tag{16}$$

Později však náhodnou chybu můžeme dopočítat po odečtení od reálné hodnoty y z tohoto vztahu:

$$\boldsymbol{y = \hat{y} + e = f(x_1, x_2, \ldots x_k) + e} \tag{17}$$

Cílem je tedy najít parametry β_1, β_1, … β_k, tak aby náhodná chybová složka e byla minimální. (3)

4.5.1 Residuální součet čtverců (residual sum of squares)

Využijeme právě při testování parametrů β_1, β_1, … β_k na n vzorcích. Vychází z metody nejmenších čtverců a snaží kvantifikovat celkovou chybu pro konkrétní vektor parametrů β:

$$\boldsymbol{RSS = E(\beta) = e^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2} \tag{18}$$

po rozepsání vektorů β, a x dostaneme následující tvary:

$$\begin{aligned} \boldsymbol{RSS = E(\beta_1, \beta_2, \ldots, \beta_k)} &= \boldsymbol{\sum_{i=1}^{n} \left[y_i - f(x_{i,1}, x_{i,2}, \ldots, x_{i,k})\right]^2} \\ &= \boldsymbol{\sum_{i=1}^{n} \left[y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \cdots - \beta_k x_{i,k}\right]^2} \end{aligned} \tag{19}$$

$E(\beta)$ je tedy funkcí, které se snažíme najít vektor parametrů β, tak aby její výstup byl limitní k 0, což znamená, že hledáme první derivaci této funkce podle β rovnu 0:

$$\boldsymbol{\frac{dE(\beta)}{d\beta} = 0} \tag{20}$$

Respektive nulovost prvních parciálních derivací:

$$\boldsymbol{\frac{dE(\beta_1, \beta_2, \ldots, \beta_k)}{d\beta_1} = \frac{dE(\beta_1, \beta_2, \ldots, \beta_k)}{d\beta_2} = \cdots = \frac{dE(\beta_1, \beta_2, \ldots, \beta_k)}{d\beta_k} = 0} \tag{21}$$

5 INFORMAČNĚ-TECHNOLOGICKÁ VÝCHODISKA PRÁCE

5.1 DATOVÉ TYPY

Pro snadnou zpracovatelnost dat stroje rozlišují datové typy. Každý údaj, který existuje, má nějaký datový typ. Datový typ závisí na druhu hodnot, kterých může daný údaj nabývat. V našem případě se jedná o tyto datové typy:

- číslo (*number*) jak celé, tak i desetinné, ze statistického hlediska se jedná o kvantitativní proměnnou
- text, resp. *řetězec znaků* (*string*) – libovolně dlouhá sekvence znaků
- logická hodnota (*boolean*) – hodnota *1* nebo *0*, resp. *pravda* nebo *lež*, resp. *ano* nebo *ne*. Jedná se o kvalitativní proměnné.
- pole (*array*) - znamená, že údaj může nabývat více hodnot nějakého datového typu; například záliby jedince mohou být reprezentovány polem *stringů* následujících řetězců „turistika“, „pletení“, „křížovky“, „rychlá auta“.
- výčet, enumerátor *(enum)* – výběr z předem stanovených hodnot; v našem případě to může být typ materiálu nosné konstrukce, kde předem řekneme, že může nabývat pouze takových hodnot: panelová, cihlová, skelet, dřevo, smíšená; a žádnou další hodnotu nepřipouštíme.

  Taktéž z hlediska zpracování se jedná o kvalitativní proměnné.

5.2 DATA MINING NEBOLI DOLOVÁNÍ DAT

Je poměrně mladé odvětví, někdy také nazývané dobývání znalostí, které se snaží ve velké „změti“ dat nacházet určité vzorce nebo zajímavé modely chování. Dnešní doba je často charakterizována jako „informační“, neboť informace hrají důležitější roli než kdy dřív. Zatímco statistika nám pomáhá v informacích hledat „pravdu“ a určuje, jak data popisovat, abychom z nich získali potřebné informace, data mining popisuje metody, jak vůbec data získat, zpracovat, vyčistit, roztřídit a uchovat. Samozřejmě pro spoustu úkonů a algoritmů může využívat právě statistických metod. Životní cyklus problému data miningové úlohy je následující:

- Obchodní/praktický krok – formulování zadaní, tj. proč to vůbec potřebujeme
- Datový krok – zkoumání možných zdrojů získání dat, příprava na analýzu
- Analytický krok – vytvoření statistických modelů a vyhodnocování dat
- Aplikační krok – zjištěné poznatky se aplikují v praxi
- Kontrolní – je třeba zjistit, zda jsou naše poznatky správné a jsou tedy přínosem

Celý postup se dá znázornit cyklicky, protože pokud při kontrole zjistíme, že je něco špatně, opakujeme celý postup, popř. jej opakujeme od určitého kroku.

Samotný proces data miningu by se dal rozepsat do následujících kroků:

- Definici problému
- Určení zdroje dat
- Příprava dat
- Volba a tvorba modelu (4)

5.3 DATOVÝ SKLAD (DATA WAREHOUSE)

Cílem datové skladu je poskytovat souhrnnější informace jako podklady pro analýzy a rozhodování v managementu a jiných institucích. Datový sklad je tedy proces, který z mnohdy nesourodých dat z nesourodých zdrojů sestaví zřejmé a čitelné přehledy. Jako vstup do tohoto procesu můžou být všelijaké tabulky, relační databáze, ale např. i papírové podklady. Výstupem je pak zpravidla nějaký report, ať už ve formě týdenního emailu nebo jiného upozornění.

Základem datových skladů je ETL systém (Extract-Transform-Load). Jeho úkolem je data extrahovat z konkrétních zdrojů, zkontrolovat kvalitu a konzistenci, přizpůsobit tak, aby mohlo být využito více nehomogenních zdrojů, a nakonec data nabídnout koncovému uživateli.

ETL systém, ačkoliv je ve většině případů před koncovým uživatelem ukryt, může spotřebovat až 70 % zdrojů potřebných pro implementaci a udržování datového skladu. ETL je pro nás tedy nezanedbatelnou součástí, která má na starosti především: (5)

- Odstranit nedostatky v datech a napravit chybějící data
- Poskytovat zdokumentovatelné metriko pro měření důvěryhodnosti dat
- Umět uschovat zpracovaná data

- Umět data zpracovat z více nesourodých zdrojů
- Přizpůsobit data, aby byla čitelná pro koncového uživatele

V této práci je prezentován software, který by se dal nazývat datovým skladem, protože splňuje tyto podmínky, využívá data mining pro zpracování dat a v konečném důsledku nabízí data uživateli přehledně prezentovány.

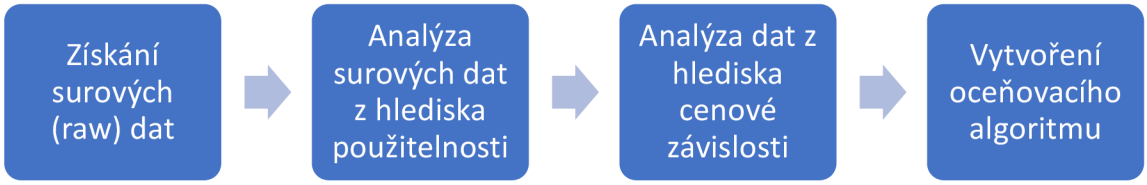
6 PRAKTICKÁ ČÁST

6.1 ÚVOD K PRAKTICKÉ ČÁSTI

Jak již bylo zmíněno, pokud chceme oceňovat tržní cenu komparativním způsobem, potřebujeme nejdříve databázi nemovitostí, které jsou podobné, ideálně ze stejné lokality a jejich prodejní tržní cenu, která byla zobchodována nejlépe v co nejbližší současnosti. Bohužel však jsou tato data těžko dostupná, resp. těžko dostupná ve velkém množství pro nějakou analýzu a zjištění, od čeho se vlastně tržní cena jednotlivých nemovitostí vlastně odvíjí. Pokud však budeme makroekonomicky předpokládat, že cena se odvíjí od poptávky a nabídky, resp. nachází se v rovnovážném bodě, tedy tam, kde nabídka je stejná jako poptávka, a inzerce reflektuje současnou nabídku, tak můžeme i předpokládat, že inzerované ceny jsou podobné jako tržní ceny.

Nejsnadněji získatelná inzertní data jsou realitní portály na internetu. Jejich velká výhoda je, že obsahují velké množství inzerátů, které jsou snadno dostupné. Hlavní výhodou pro nás navíc je, že se dají dost dobře strojově zpracovat.

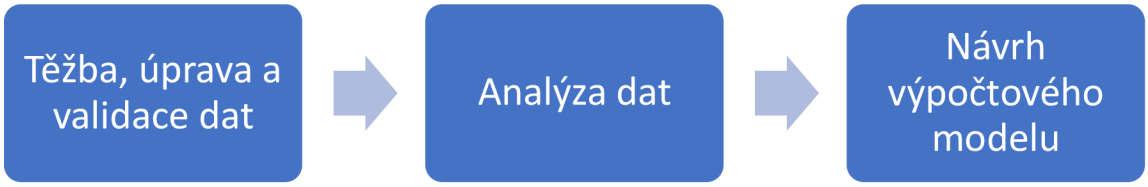
Dopředu bohužel nevíme, na kterých parametrech je cena bytu závislá a co hůř, nevíme ani hromadně jaké všechny parametry a v jakém zastoupení v jednotlivých inzerátech nalezneme. Proto jsem se držel následujícího postupu:



Obrázek 4 - flow analyzování dat (vlastní)

Základem všeho je získání dat, na kterých můžeme dělat rozbor toho, který parametr je, v jakém zastoupení, a jakých hodnot nabývá. Jakmile máme srovnaná a použitelná data, můžeme zkoumat závislosti ceny nemovitosti na jednotlivých parametrech nemovitosti. Na základě těchto poznatků lze pak sestrojit oceňovací algoritmus, který bude tyto parametry reflektovat.

Problém tedy můžeme rozdělit na tři části, které jsou reflektovány v následujících 3 kapitolách 2. úrovně.



Obrázek 5 - přístup ke stanovení výpočtového modelu (vlastní)

6.2 TĚŽBA, ÚPRAVA A VALIDACE DAT

6.2.1 Definice problému

Proto, abychom dokázali ocenit nějakou nemovitost porovnávacím způsobem, potřebujeme údaje o proběhlých transakcích s tímto druhem nemovitostí. V tomto případě zejména cenu, zda nebylo s touto nemovitostí obchodováno v mimořádné okolnosti, tedy jestli například na kupujícího nebyl vyvíjen nátlak nebo se do ceny nepromítají rodinné poměry mezi prodávajícím a kupujícím, zároveň by však tato informace o ceně neměla být příliš zastaralá. Dále potřebujeme znát i nějaké bližší vlastnosti, které danou nemovitost charakterizují. Problémem však bývá to, že tyto informace nebývají dost dobře dostupné. Případně nebývají vhodné (například údaje o nemovitostech z jiných lokalit s jinými okolními poměry).

6.2.2 Určení zdroje dat

Vzhledem k tomu, že potřebujeme pro minimalizaci chyby, co největší množství informací, je nasnadě vyzkoušet internet a realitní weby, které by měly reflektovat tržní ceny, jako zdroj dat.

Pro naše účely je potřeba, aby aplikace dokázala přečíst údaje z jednotlivých inzerátů a brala je v potaz při oceňování. Je zde však několik problémů:

1. každý realitní web napříč internetem vypadá jinak a má jiný vzhled a jinou strukturu informací
2. dva inzeráty na stejném webu mohou mít odlišnout strukturu a nabízet poněkud rozdílné spektrum informací (např. u jednoho inzerátu se nalézá údaj o parkování; u druhého nikoliv, z důvodu, že jej inzerent nezadal)
3. pokud bychom museli pro každé ocenění projít každý inzerát znovu a znovu, trvalo by takové oceňování nesmírně dlouho
4. některý z realitních serveru by mohl být zrovna offline (například při údržbě či výpadku) a my bychom měli o zdroj méně, pokud by ocenění proběhlo v reálném čase
5. některé inzeráty se mohou vyskytovat duplicitně, museli bychom zpětně porovnávat téměř každý inzerát s každým pro danou lokalitu napříč několika realitními weby, což by bylo výpočetně značně náročné, nehledě na počet požadavků na zdrojové servery

Problém č. 1 a 2 se dá výřešit tak, že inzeráty transformujeme na jednotný formát. K tomu abychom toto provedli však potřebujeme nashromáždit velké množství dat a analyzovat, jaké možné parametry nemovitostí se v inzerátech vyskytují a jakých mohou nabývat hodnot. Pak je třeba porovnat jednotlivé, vyskytující se, parametry, zda se vyskytují na všech webech. Pokud se například informace o vytápění vyskytuje jen na jednom webu, není pro nás relevantní, protože tento parametr nemůžeme porovnávat s inzerátem z jiného webu, který tento parametr neobsahuje.

Probém č. 3 a 4 vyřešíme tím, že zřídíme lokální uložiště inzerátů, které bude reprezentovat stav nabídky prodeje bytů v Jihomoravském kraji v čase t. V tomto uložišti budou

uložena data v unifikovaném formátu, abychom mohli snadno data analyzovat a využívat pro ocenění.

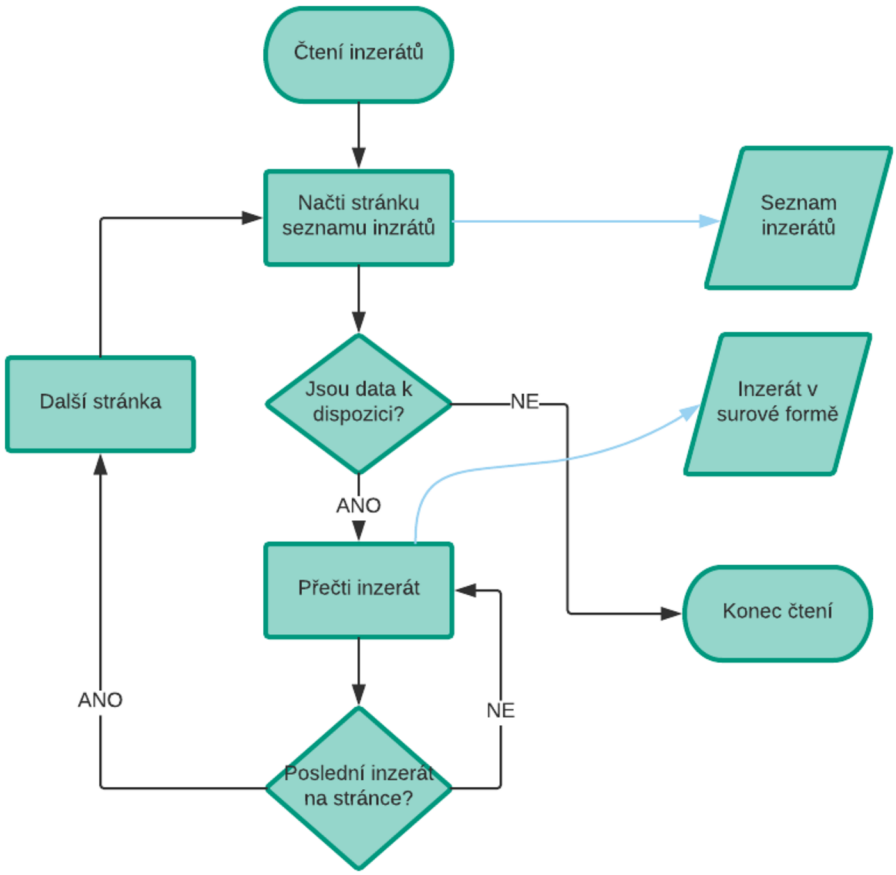
Poslední 5. problém částečně řeší řešení problému č. 3 a 4, avšak je nutné stanovit kritéria pro vyloučení duplicit (viz dále).

Pro tvorbu přechodné databáze bytů byly vybrány následující realitní portály (zdroje dat):

- Reality.idnes.cz
- Realitymix.centrum.cz
- Sreality.cz

Vybrány byly především pro jejich snadnou dostupnost, dostatek bytů a informací o nich a taky pro jejich častou aktualizovanost.

Čtení dat probíhá analogicky jako u člověka, s tím, že algoritmus mezi inzeráty nerozlišuje a dívá se postupně na každý inzerát, než narazí na konec stránky. Poté se snaží přepnout další stránku, pokud je to možné. V opačném případě je čtení považováno za ukončené. Na následujícím obrázku je celý proces znázorněn.



Obrázek 6 - proces získávání dat (vlastní)

6.2.3 Analýza surových dat

Poté, co jsou data přečtena v surovém formátu z jednotlivých zdrojových webů, je třeba projít jednotlivé parametry a zjistit, zdali má vůbec cenu se daným parametrem zabývat. Můžeme rovnou vyloučit parametry, které:

- jsou pro nás obtížně strojově zpracovatelné
- nabývají příliš heterogenních hodnot
- mají hodně malý výskyt
- některý ze zdrojových webů nedisponuje tímto nebo podobným parametrem

Většinou se jedná o parametry, kde má inzerent, při vkládání inzerátu, svobodu napsat cokoliv, takže hodnoty tohoto parametru jsou značně nedeterministické. Toto velmi záleží na konkrétním inzertním webu a jeho grafickém rozhraní, resp. způsobu vkládání inzerátů.

Někdo by mohl namítnout, že při současném pokroku ve strojovém učení a umělé inteligenci existují metody, jak si poradit i s tímto problémem, nicméně využití těchto technologií by značně překročilo rámec této práce, neřkuli by bylo na samostatnou práci.

Přítomnost jednotlivých parametrů v inzerátech

Abychom věděli, zda se nějakým parametrem zabývat a zkoumat jej, potřebujeme vědět, jak moc se v jednotlivých inzerátech tento parametr objevuje. V následující tabulce nalezneme přehled výskytu podle zdrojového portálu. Hodnota 400 je pro nás nejpříznivější, protože ze 400 přečtených inzerátů, je daný parametr obsažen v každém z nich a můžeme s přítomností tohoto parametru na 100 % počítat.

Název parametru	Identifikátor	reality.idnes.cz	realitymix.centrum.cz	sreality.cz
Vybavení	equipment	270	127	195
Vytápění	heatingType	202	140	224
Vlastnictví	ownershipType	399	336	400
Stav bytu	flatCondition	380	400	400
Číslo podlaží	floorNum	381	400	400
Balkón/lodžie/terasa	balcony	247	165	63
Adresa	address	400	400	400
Počet pokojů	numRooms	400	400	400
ID inzerátu	theirId	398	400	337
Typ konstrukce	constructionType	397	400	400
Plocha	size	396	400	400
Poznámka k ceně	priceNote	400	400	260
Popis	description	400	400	400
Cena	price	400	400	381
PENB	energyEfficiencyRating	374	285	372

Sklep	basement	240	189	290
Výtah	elevator	123	0	275
Parkování	parking	136	0	110
Stav budovy	buildingCondition	133	0	0
Počet podlaží budovy	floorsTotal	0	371	0
Doprava	traffic	0	174	224
Bezbariérový přístup	barrierFreeAccess	0	17	0
Poslední změna	lastUpdate	0	0	400

Tabulka 1 - četnost přítomnosti jednotlivých atributů v inzerátech (vlastní)

Jak vidíme, tak počet podlaží budovy se vyskytuje pouze *realitymix.centrum.cz*, proto s ním nebudeme počítat. Totéž platí pro bezbariérový přístup a poslední změnu, které se vyskytují pouze na jednom z portálů. Výtah a parkování nejsou zase přítomné na jednom z webů. Totéž i doprava. Z toho plyne, že dále musíme prozkoumat všechny přečtené hodnoty z inzerátu a analyzovat jejich možnosti využití, resp. jakých hodnot nabývají, zda jsou pro nás užitečné a zda jsme schopni je zpracovat. Hodnoty označené ❌ nesplňují tyto podmínky a nezahrnujeme je dále k parametrům srovnávacích nemovitostí. V opačném případě je parametr označen ✅ a můžeme jej dále statisticky analyzovat nebo využít pro kontrolu duplicit.

Zkoumání hodnot jednotlivých parametrů

- ❌ Vybavení (*equipment*) – hodnoty, kterých nabývá nejsou sice tak různorodé, nicméně u *realitymix.centrum.cz* se tento parametr nachází pouze 127x, takže s ním nebudeme počítat.
- ❌ Vytápění (*heatingType*) – opět u *realitymix.centrum.cz* se nachází pouze 140x, takže tento budeme taktéž ignorovat.
- ✅ Vlastnictví (*ownershipType*) – má poměrně velkou četnost u všech řešených realitních webů a nabývá vesměs pouze hodnot: osobní, družstevní a jiné; můžeme tedy tento parametr brát v potaz.

- ✅ Stav bytu (*flatCondition*) – má téměř 100 % četnost a nabývá pouze hodnot: dobrý stav, novostavba, po rekonstrukci, před rekonstrukcí, ve výstavbě, projekt. Rozhodně bereme tento parametr v úvahu.
- ✅ Číslo podlaží (*floorNum*) – taktéž se vyskytuje téměř u každého inzerátu. Jedná se navíc o číselný údaj, takže s ním můžeme počítat.
- ❌ Balkón/lodžie/terasa (*balcony*) – vzhledem k různorodosti nabývajících hodnot (některé weby uvádějí plochu, některé jen ano/ne) a nízké četnosti (nejvíce 247, nejméně 63), tento parametr nebereme v úvahu.
- ✅ Adresa (*address*) – pro nás důležitý parametr, které se vyskytuje v každém inzerátu. Z hlediska ocenění patří mezi ty nejdůležitější, rozhodně bereme v potaz.
- ✅ Počet pokojů (*numRooms*) – opět parametr, který se vyskytuje u každého inzerátu a může korelovat s jinými parametry, má smysl se jím tedy dále zabývat.
- ✅ ID inzerátu (*theirId*) – jedno z kritérií pro kontrolu duplicit. Pokud se snažíme uložit inzerát a máme již inzerát ze stejného zdroje a se stejným id, nemusí být sporu o tom, že se jedná o duplicitní záznam.
- ✅ Typ konstrukce (*constructionType*) – vyskytuje se v podstatě u všech inzerátů a může hrát nezanedbatelnou roli na ceně nemovitosti a také může být zajímavé srovnávat ceny mezi panelovými a zděnými byty.
- ✅ Plocha (*size*) – Jeden z rozhodujících parametrů pro ocenění. Inzeráty bez ceny budou ignorovány.
- ❌ Poznámka k ceně (*priceNote*) – ačkoliv obsahuje důležité informace o ceně, obsahuje velmi heterogenní informace, zřejmě se taktéž jedná o pole, do kterého může uživatel napsat cokoliv. Vzhledem k tomuto, bylo by třeba využít složitějšího algoritmu na dešifrování poznámky. Proto se zde tímto parametrem nebudeme zabývat. Co ale však říci můžeme je to, že ceny jsou ve většině případů bez daně a provizí, což je informace, kterou využijeme později.
- ❌ Popis (*description*) – bohužel do popisku může inzerent napsat cokoliv a jak již bylo zmíněno, tato práce se nezaměřuje na využití umělé inteligence, která by dokázala danému

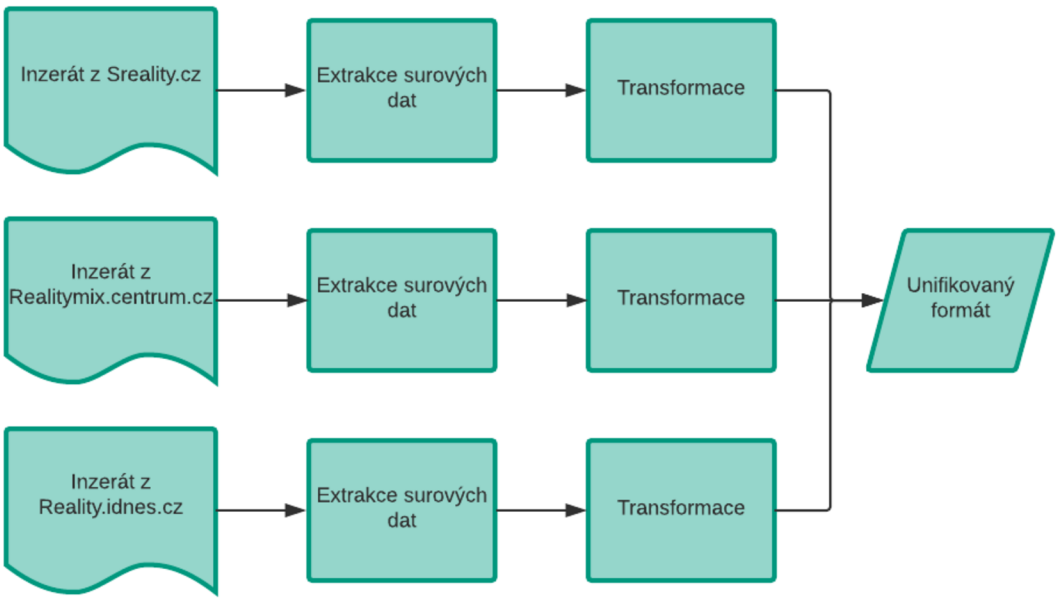
textu porozumět a vyvodit z něj další parametry nemovitosti. Proto tento parametr nebudeme brát v úvahu.

- ✅ Cena (*price*) – rozhodující údaj, stejně jako plocha.
- ❌ PENB (*energyEfficiencyRating*) – má sice častý výskyt, ale naneštěstí se v drtivé většině inzerátů vyskytuje PENB třída G. Důvodem je legislativní rámec: „Nepředá-li majitel bytu zprostředkovateli energetický štítek, musí ho zprostředkovatel inzerovat s energetickou třídou G – mimořádně nehospodárná budova.“ (6) Navíc u sreality.cz tento parametr nabývá více různorodých hodnot.
- ✅ Sklep (*basement*) – ačkoliv se vyskytuje u mnoha inzerátů v mnoha různých podobách, dá se předpokládat, že inzeráty, kde chybí, sklep nemají. Můžeme jej tedy brát jako logickou hodnotu ano/ne.
- ❌ Výtah (*elevator*) – u *realitymix.centrum.cz* se tento parametr nevyskytuje vůbec, proto jej nebereme v potaz.
- ❌ Parkování (*parking*) - u *realitymix.centrum.cz* se tento parametr opět nevyskytuje vůbec, proto jej nebereme v potaz.
- ❌ Stav budovy (*buildingCondition*) – parametr vyskytující se jen u *sreality.cz*, vyloučíme jej.
- ❌ Počet podlaží budovy (*floorsTotal*) – parametr, který se vyskytuje paradoxně pouze u *realitymix.centrum.cz*. Nemá smysl se jím zabývat.
- ❌ Doprava (*traffic*) – nedostupné u sreality.cz. Nebude dále řešeno.
- ❌ Bezbariérový přístup (*barrierFreeAccess*) – vyskytuje se pouze u *realitymix.centrum.cz* a ve velmi malém výskytu. Z toho lze usuzovat, že bezbariérové byty ještě nejsou úplně trendem. Proto tento parametr také vynecháme, nicméně do budoucna bude mít bez pochyby nemalý vliv na cenu, kvůli potřebným stavebním opatřením.
- ❌ Poslední změna (*lastUpdate*) – Mohlo by mít vliv na oceňovací algoritmus, který by zahrnul stáří inzerátu a jeho vliv na cenu, nicméně se bohužel tento údaj vyskytuje pouze u *sreality.cz*.

6.2.4 Transformace na unifikovaný formát

Účelem transformace dat je získat je ve formátu, který nám vyhovuje a je pro nás dobře zpracovatelný. V našem případě je třeba adekvátně interpretovat co nejvíce parametrů z inzerátů tak, aby si napříč jednotlivými weby jejich hodnoty byly konkurenceschopné. To znamená, že pokud na webu *A* i *B* bude parametr „druh vlastnictví“ nabývat hodnot „osobní“, „družstevní“, měl by tento parametr stejných hodnot nabývat i na webu *C*, byť třeba v jiné formulaci. Transformací se pak data dostanou do deterministické podoby (*unifikovaný formát*), kde formulace nabývají předem definovaných a navzájem si odpovídajících hodnot, výměry a ceny jsou ve stejných jednotkách a celkově jsou data takto snadno zpracovatelná.

Jedná se tedy o proces, který by se dal vyjádřit jednoduchou funkcí $y = f(x)$, kde x jsou vstupní data, f je transformační funkce a y jsou data vystupující. Jelikož je každý ze zdrojových webů jiný, má každý zdroj svou transformační funkci.



Obrázek 7 - flow extrakce a transformace (vlastní)

6.2.5 Kontrola duplicit

Kontrola duplicit je nelehký úkol, neboť je komplikován nejednoznačností parametrů. Můžeme mít například shodné nemovitosti s rozdílnou adresou, protože jedna bude mít uvedeno

„Brno, Jaselská 5“, nemovitost *B* pak třeba jen „Brno, Jaselská“. Můžeme tedy zvolit cenu a plochu jako srovnávací parametr duplicit, nicméně není vyloučeno, že někde jinde nebude inzerována rozdílná nemovitost se stejnou cenou i plochou. Může nastat i situace, kdy bude jedna a tatáž nemovitost inzerována s rozdílnou cenou u konkurenčních realitních webů. Vzhledem k rozsahu této práce budeme předpokládat, že tento problém zde nenastává nebo nastává pouze minimálně. Jako duplicitní zde tedy považujeme inzeráty, které pro jednoduchost splňují všechny následující podmínky:

- Mají stejnou adresu
- Mají stejnou cenu
- Mají stejnou plochu
- Mají stejný počet místností

6.2.6 Přiřazení lokality

Pro nás lidi je adresa jako posloupnost znaků srozumitelná a zřejmě bychom se dokázali pomocí ní dopravit na konkrétní místo. Bohužel však pro stroj je to nesrozumitelný řetězec znaků, pod kterým si nedokáže nic představit. Pro to, abychom dokázali s adresou strojově pracovat, musíme ji dostat do strojově čitelného formátu. Ideálně do číselné podoby. Tímto se nám nabízí GPS souřadnice, nad kterými již stroj dokáže provádět matematické operace jako je třeba počítání vzdáleností.

GPS souřadnice

„Globální polohový systém (Global Positioning System) je vojenský družicový polohový systém provozovaný Ministerstvem obrany Spojených států amerických, s jehož pomocí je možno určit geografickou polohu přijímače nacházejícího se kdekoliv na Zemi nebo nad Zemí s přesností jednotek metrů a také čas s přesností na jednotky nanosekund.“ (7)

GPS souřadnice přijímače signálu jsou reprezentovány zeměpisnou šířkou (latitude), zeměpisnou délkou (longitude) a výškou (altitude), která nás však v této práci zajímat nebude.

Geocoding

Je to proces, jehož smyslem je převést adresu (jako například „Purkyňova 464/118, 612 00 Brno“) na zeměpisné souřadnice (resp. GPS souřadnice). V této práci využijeme tohoto instrumentu pro převod inzerované adresy z textového formátu na GPS souřadnice, které následně budeme strojově zpracovávat. (8)

Google Maps API

Asi není velice třeba představovat službu Google Maps od společnosti Google. Jako největší poskytovatel internetových map nabízí i aplikační rozhraní, se kterým je možno komunikovat se servery Google Maps a využívat tak těchto služeb bez přístupu přes grafické rozhraní.

V našem případě tedy využijeme Google Maps API pro geocoding. Google ve své dokumentaci uvádí, že umožňuje zdarma zpracovat pouze 2 500 požadavků za den, což však pro naše účely dostačuje. Do požadavku tedy vložíme adresu, odešleme požadavek serveru a server by nám měl odpovědět v následujícím JSON formátu (8):

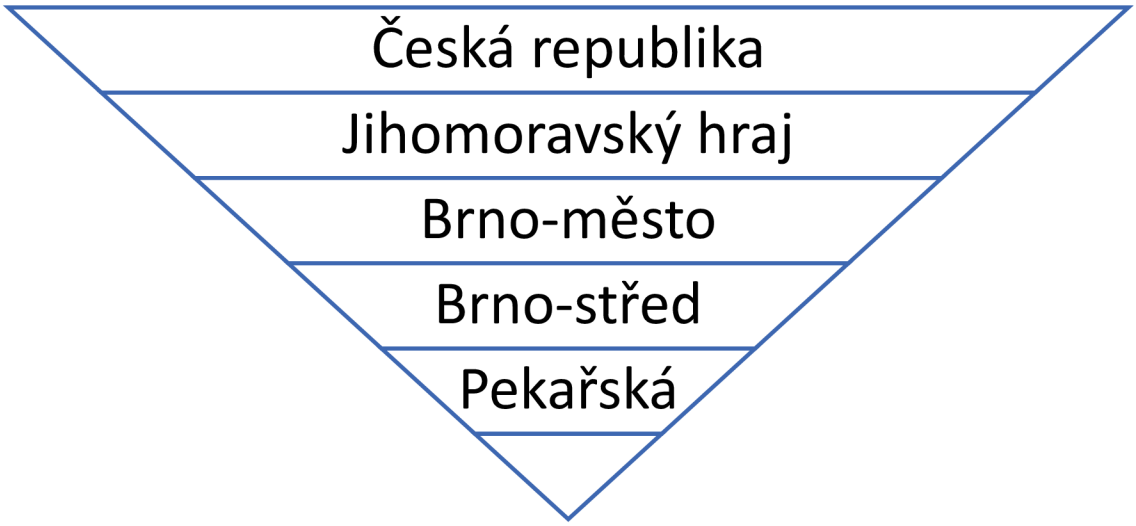
```
results[]: {
 types[]: string,
 formatted_address: string,
 address_components[]: {
   short_name: string,
   long_name: string,
   postcode_localities[]: string,
   types[]: string
 },
 partial_match: boolean,
 place_id: string,
 postcode_localities[]: string,
 geometry: {
   location: LatLng,
   location_type: GeocoderLocationType
   viewport: LatLngBounds,
   bounds: LatLngBounds
 }
}
```

Obrázek 8 - šablona odpovědi Google Maps geocodingu (8)

Z toho nás bude zajímat *geometry*, které uchovává samotné pozice a *address_components*, které nám poskytuje informace o adrese, tedy jednotlivé složky adresy

ve stylu {$stát, kraj, obec,$ část $obce, ulice$}, což se nám vyplatí, pokud budeme chtít do budoucna jednotlivé inzeráty filtrovat například podle obce.

Abychom dokázali adresy snadno strojově zpracovávat, budeme si je ukládat jako pole názvů jednotlivých geografických celků od největšího po nejmenší. Například tedy:



Obrázek 9 - strojově zpracovatelná forma rozložené adresy (vlastní)

Pak pro nás nebude problém mezi sebou tento parametr porovnávat. Pokud budeme chtít například nemovitosti z celého Brna-střed, tak porovnáme první 4 položky tohoto pole, resp. 4 položku, zda se rovná řetězci „Brno-střed".

Vzdálenost od středu regionu

Čistě ze zvědavosti mě zajímala korelace mezi jednotkovou cenou a vzdáleností do středu většího územního celku, než je samotná obec, ale zároveň do menšího, než je krajská úroveň. Nabízí se tedy vzdálenost do středu regionu. Jelikož jsem si vybral Jihomoravský kraj pro analýzu, jedná se konkrétně o tyto regiony: *Brno-město, Hodonín, Blansko, Vyškov, Brno-venkov, Znojmo, Břeclav.*

Jak ale zjistit pozici centra daného regionu? Jak již bylo řečeno, z *address_components* jsme schopni zjistit obec, kde se nemovitost nachází, pak již stačí zavolat geocoding metodu Google Maps API a zjistit pozici obce, kdy Google Maps zpravidla vrací střed dané obce, pokud hledáte pouze danou obec a žádnou konkrétní ulici, městskou část, či číslo domu.

Výpočet ortodromy

Z analytické geometrie známe vztah pro výpočet vzdálenosti dvou bodů. Bohužel však máme k dispozici pouze zeměpisné šířky a délky těchto bodů, které jsou udávány v úhlech, takže bychom dostali pouze úhlovou vzdálenost, která nám nic moc neřekne o tom, jak jsou reálně od sebe místa vzdálena. Potřebujeme tedy z těchto úhlů vypočítat vzdálenost na zemském povrchu. V geografické kartografii zcela postačuje výpočet na kulové ploše. Nejkratší vzdálenost dvou bodů na ploše koule se tedy nazývá ortodroma. Pro její výpočet se využívá kosinová věta sférické trigonometrie:

$$\cos \boldsymbol{D} = (\sin \boldsymbol{\varphi_A} \sin \boldsymbol{\varphi_B}) + (\cos \boldsymbol{\varphi_A} \cos \boldsymbol{\varphi_B} \cos|\boldsymbol{\lambda_A} - \boldsymbol{\lambda_B}|)$$
$$\boldsymbol{d} = \frac{\boldsymbol{2\pi r}}{\boldsymbol{360}} \boldsymbol{D} \qquad (22)$$

kde φ_A a φ_B jsou zeměpisné šířky bodů, λ_A a λ_B jejich délky, r je poloměr koule (v našem případě poloměr Země, tedy 6378 km). d je výsledná vzdálenost. (9)

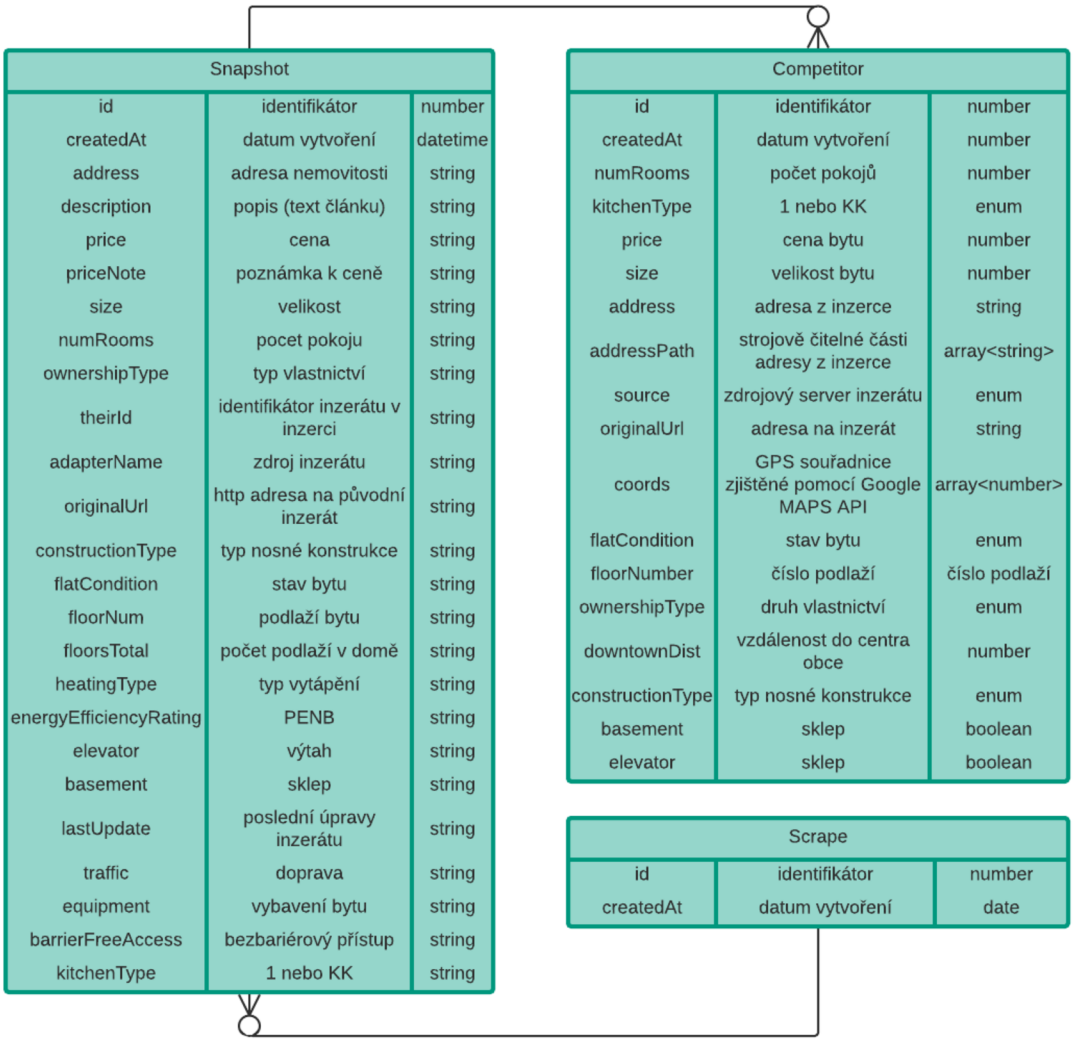
6.2.7 Databázové schéma

Na základě potřeb vzniklo následující schéma. Je reprezentováno 3 druhy entit:

Scrape – označuje jednu dávku čtení dat z inzerátů. Je reprezentována pouze datem vytvoření a identifikátorem a slouží ke shlukování jednotlivých *snapshotů*.

Snapshot – reprezentuje surová data přečtená z jednoho inzerátu. Sloužilo pouze v analytické části, kdy nebylo dopředu jasné, který parametr nabývá, jakých hodnot a bylo je tedy potřeba před transformací uložit a analyzovat. Na jejich základě pak byla vytvořena transformační funkce.

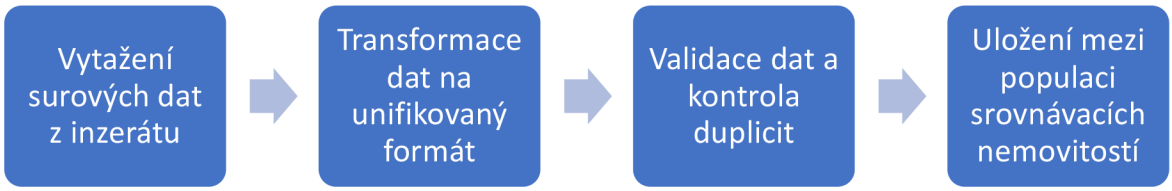
Competitor – představuje finální podobu inzerátu v naši databázi. Inzerát je již uložen v kompaktním homogenním stavu, kdy je porovnatelný s inzeráty z jiných zdrojů.



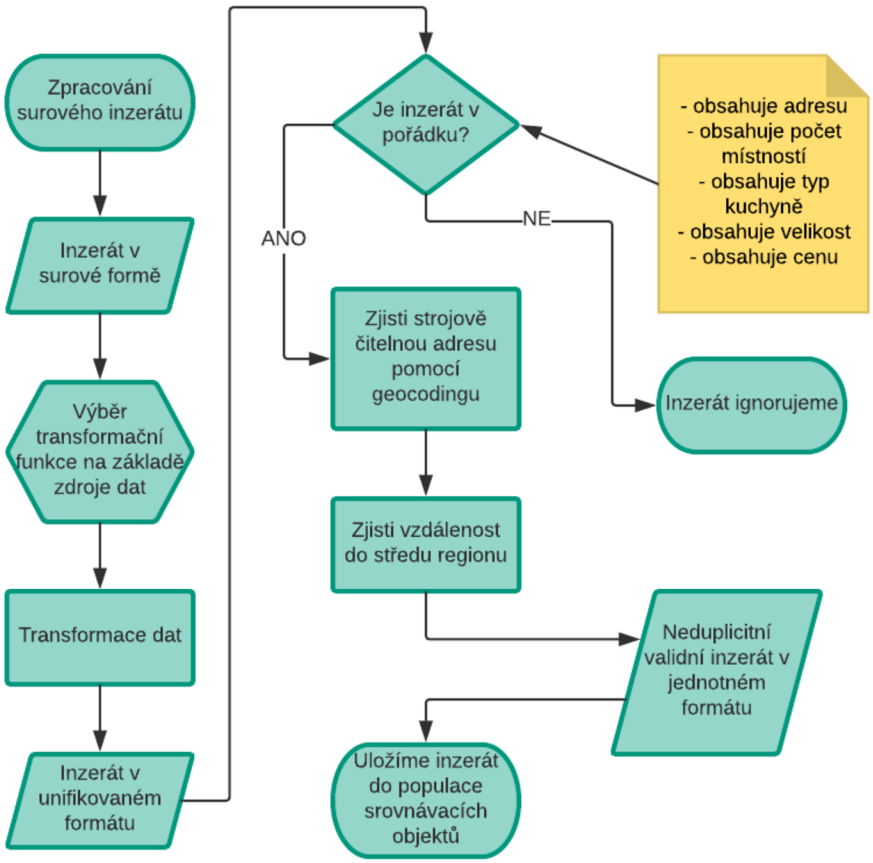
Obrázek 10 – ER diagram (vlastní)

6.2.8 Shrnutí

Aplikace tedy po přečtení dat z inzerátu (vytvoření *snapshotu*) provede transformaci, abychom měli data v homogenním formátu a mohli s daty zacházet stejně (jedna z vlastností datových skladů). Po transformaci se musí data zkontrolovat, zda obsahují validní údaje, následně se provede kontrola duplicit a pokud je inzerát v pořádku, je mu přiřazena pomocí geocodingu strojově čitelná lokalita. Nakonec se vše zapíše do databáze (vytvoření *competitora*) pro pozdější analytické potřeby a samozřejmě pro potřeby ocenění.



Obrázek 11 – flow pro těžbu, úpravu a uchování dat (vlastní)



Obrázek 12 - proces zpracování surových dat (vlastní)

6.2.9 Zdrojová data

V následující tabulce vidíme přehled o použitých datech. Bylo by zde vhodné využít analýzu rozptylů (ANOVA), ale pro jednoduchost budeme předpokládat, že populace je homogenní. Odpovídaly by tomu i směrodatné odchylky jednotkových cen, které se příliš neliší. Průměry mohou být vychýleny některými extrémy (viz kapitola 4.3.1 – výběrový průměr).

Source	Num	Average price	Average unit price	Minimal unit price	Maximal unit price	Standard deviation
sreality.cz	338	3 519 311	55 084	13 978	121 223	16 848
realitymix.centrum.cz	334	3 599 430	51 078	10 938	109 489	16 941
reality.idnes.cz	291	3 396 148	49 181	11 803	95 600	16 681

Obrázek 13 - statistika zdrojových dat (vlastní)

Srovnávací objekty byly pořízeny k datu: úterý 8. května 2018 v 17:05, a veškeré přehledy, tabulky, grafy a obrázky v této práci z nich vycházejí. Z celkových 1200 přečtených inzerátu algoritmus vyhodnotil 963 jako použitelných, což odpovídá ~ 80 %.

6.3 ANALÝZA POPULACE SROVNÁVACÍCH NEMOVITOSTÍ

V této části jsou řešeny závislosti jednotlivých parametrů a jejich vliv na cenu. Jako populaci zde chápeme prodej z inzerce pro celý Jihomoravský kraj, která bude srovnána s výběrovou množinou bytů Brno-město.

6.3.1 Analýza výčtů (enumerátorů)

Jedná se o analýzu průměrných jednotkových cen u parametrů, které nabývají předem daných hodnot. Pokud zde budeme mluvit o jednotkové ceně, myslíme cenu za m^2. Číslo v závorce reprezentuje počet jedinců, kteří odpovídají dané hodnotě konkrétního parametru.

Průměrné ceny dle typu vlastnictví

Z následujícího obrázku je patrné, že je inzerenty požadována v průměru vyšší cena, pokud prodávají byt v osobním (*personal*) vlastnictví. To může být dáno tím, že pří koupi družstevního (*society*) bytu nejste vedeni jako vlastníci bytu v katastru, nemůžete s ním volně nakládat, máte pouze nárok na smlouvu na dobu neurčitou. (10) Toto může být problémem i při žádosti o úvěr.

Obrázek 14 - průměrné inzertní ceny dle typu vlastnictví (vlastní)

Průměrné ceny dle stavu bytu

Stav nemovitosti hraje nepochybně roli na její ceně. Totéž platí i pro byty. Je otázka, zda inzeráty na realitních serverech dostatečně dobře reflektují reálný stav. Nicméně z následujícího obrázku lze vyvodit pár závěru, které lze i logicky odůvodnit.

Obrázek 15 - průměrné inzertní ceny dle stavu bytu (vlastní)

Nejhůře jsou na tom pochopitelně budovy před rekonstrukcí (*before reconstruction*), neboť jejich koupě bude zahrnovat dodatečné náklady na rekonstrukci. Naopak nejlépe jsou na tom v Brně městě byty, které jsou teprve ve fázi projektu nebo již ve výstavbě (*building*), které dle dostupné inzerce, je nadstandardně vybaveno a v mnoha případech se v inzerci vyskytují vizualizace. Mimo Brno město jsou to zase na druhou stranu nové byty (novostavby), kde zřejmě komfort nebude hrát asi tak velkou roli jako možnost se dříve nastěhovat.

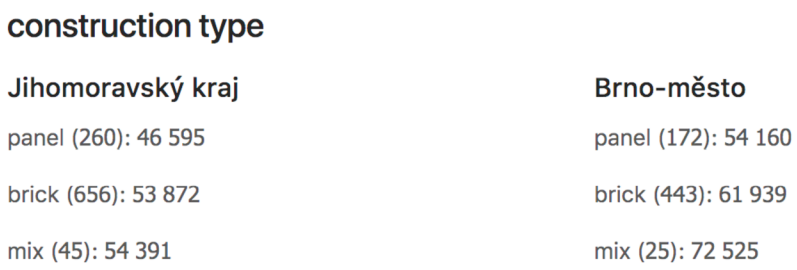
Dále na tom nejsou špatně byty po rekonstrukci, kde se dá očekávat podobný komfort jako u novostaveb, nicméně dle rčení „i když nasadíte kozlovi smoking, pořád to bude kozel“, můžete mít stále na paměti, že to může být starší stavba, která po čase vykáže majoritní poruchy. V Brně-město můžeme vidět, že tento rozdíl není tak markantní, jako v celém Jihomoravském

kraji. Důvodem může být to, že centrum města je již z velké části zastavěno a jsou možné ve větší míře pouze rekonstrukce.

Mezi stavy dobrý (*good*) a velmi dobrý (*very good*) asi mnoho inzerentů nerozlišuje. Může to být dáno subjektivním dojmem ze stavu bytu. Zatímco stavy před rekonstrukcí, ve výstavbě apod. jsou naprosto zřejmé a snadno určitelné, a proto je zřejmě mezi nimi větší cenový rozdíl.

Průměrné ceny dle typu konstrukce

Z inzerce je patrné (viz následující obrázek), že cena je závislá na druhu použitého materiálu, neboť panelové stavby jsou levnější než zděné. Důvodem může být fakt, že o panelové stavby není taková poptávka, neboť konstrukční vlastnosti panelových staveb (převážně tepelné a zvukové parametry) nejsou zpravidla tak dobré jako u novějších zděných staveb. Jako další nevýhoda může být vnímáno umístění panelových staveb, které se v drtivé většině případů stavěli na sídlištích s větší hustotou obyvatelstva.



Obrázek 16 - průměrné inzertní ceny dle typu konstrukce (vlastní)

Objevuje se zde i kategorie „mix“. Po menším průzkumu jsem došel k závěru, že někteří inzerenti využívají této kategorie pro byty v řadové městské zástavbě, což by vysvětlovalo i nejvyšší průměrnou cenu. V jiných případech to může být lenost inzerenta daná neznalostí nebo nezájmem o jaký typ konstrukce se vlastně jedná.

Průměrné ceny dle přítomnosti sklepu

Co můžeme s určitostí říct, tak že v inzerci bytů v Jihomoravském kraji převažují byty se sklepem, které jsou však průměrnými cenami níže než byty bez sklepu. Tento jev bych si vysvětloval tím, že více dnešních novostaveb je bez sklepu, tudíž je většina nepodsklepených domů novějších a tím i dražších.

Obrázek 17 - průměrné inzertní ceny dle přítomnosti sklepu

Průměrné ceny dle přítomnosti oddělené kuchyně

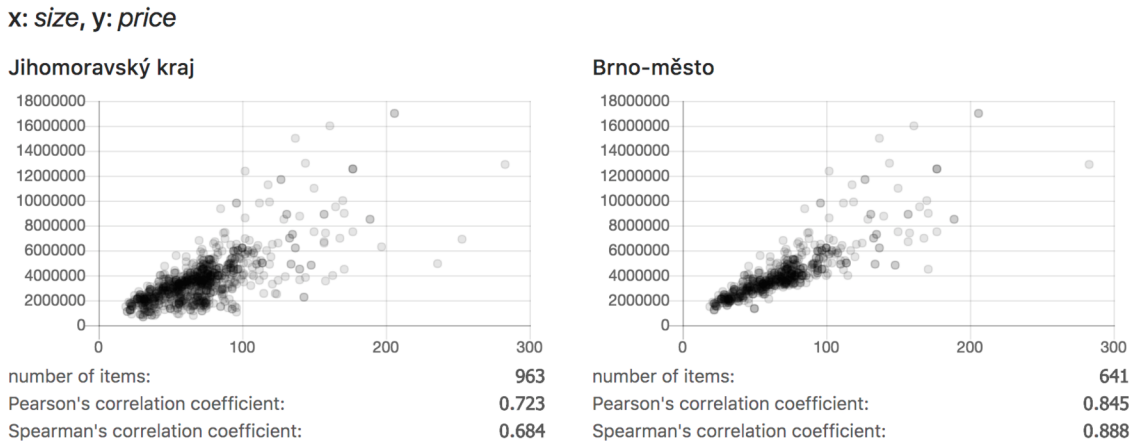
Zde je vidět diametrální rozdíl v cenách bytů se samostatnou kuchyní a KK. Mým názorem je, že je podobný trend jako u sklepů, kdy spousta dnešních novostaveb má obývací nebo jiný pokoj spojen s kuchyní.

Obrázek 18 - průměrné inzertní ceny dle samostatnosti kuchyně (vlastní)

6.3.2 Analýza závislostí

Závislost cen bytů na jejich ploše

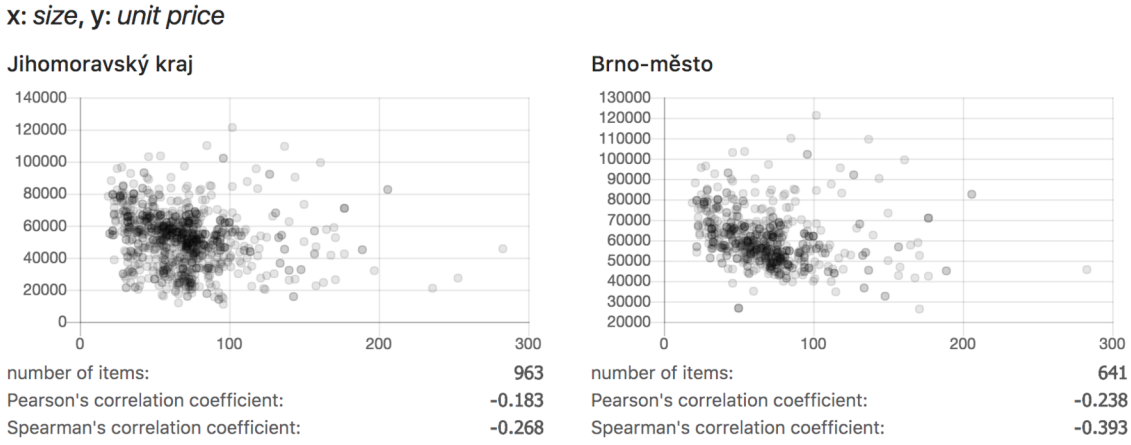
Na následujícím obrázku je patrná poměrně jasná závislost ceny na ploše nemovitosti. V Brně-městě je pak tato korelace ještě více patrnější.



Obrázek 19 - závislost ceny na ploše dle inzerce (vlastní)

Závislost jednotkových cen bytů na jejich ploše

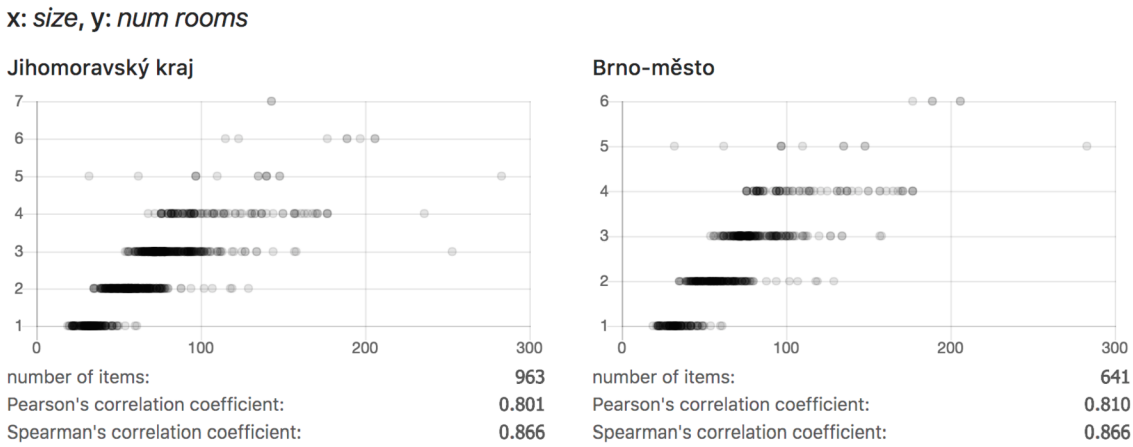
A z dalšího obrázku vyplývá, že existuje vztah i mezi plochou a jednotkovou cenou. Není úplně překvapující, že menší byty jsou co do jednotkové ceny dražší. Je o ně větší poptávka a jsou i dobrým investičním artiklem pro pronájem. (11)



Obrázek 20 - závislost jednotkové ceny na ploše bytu dle inzerce (vlastní)

Závislost počtu místností na velikosti bytu

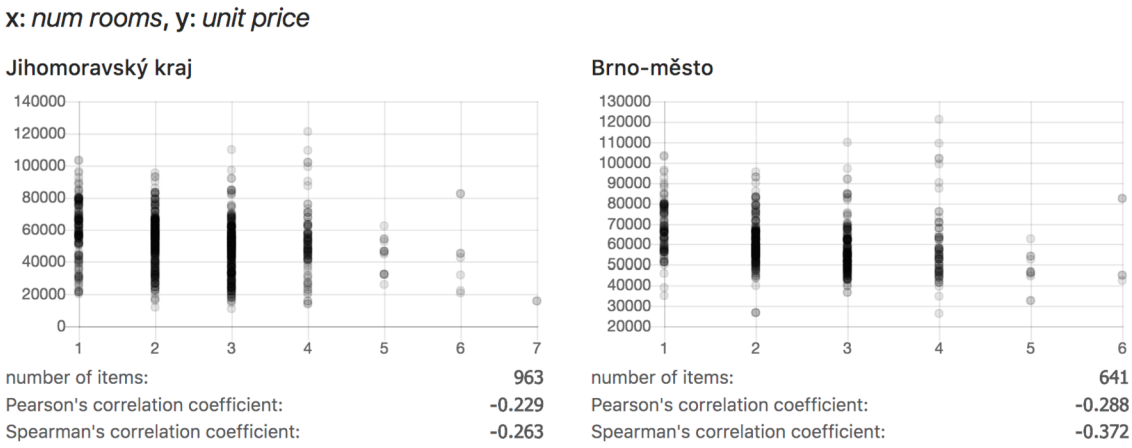
Tento obrázek asi nebude pro nikoho překvapující, ukazuje zcela jasnou závislost počtu místností na velikosti bytu. Lze však vidět i některé extrémy (velký být, málo místností; malý byt, hodně místností).



Obrázek 21 - závislost počtu místností na velikosti bytu dle inzerce (vlastní)

Závislost jednotkové ceny na počtu místností

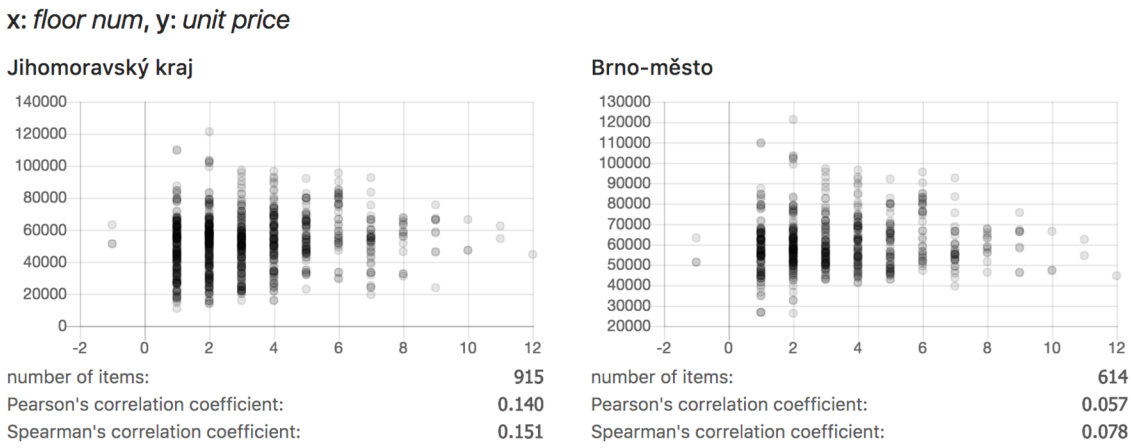
Vzhledem k tomu, že existuje vztah mezi velikostí a počtem místností, velikostí a jednotkovou cenou, jak jsme si před tím ukázali, je pravděpodobné, že bude existovat i určitý vztah mezi počtem místností a jednotkovou cenou. Následující obrázek ukazuje, že tomu tak částečně je. Jedná se tedy o multikolinearitu.



Obrázek 22 - závislost jednotkové ceny na počtu místností dle inzerce (vlastní)

Závislost jednotkové ceny na čísle podlaží

Zde bylo nejprve nutné odfiltrovat jednotky, které nemají číslo podlaží uvedeno, proto jsou zde menší počty položek. Vidíme, že korelace není nijak významná v celém Jihomoravském kraji, natož přímo v Brně, kde je ještě bezvýznamnější.

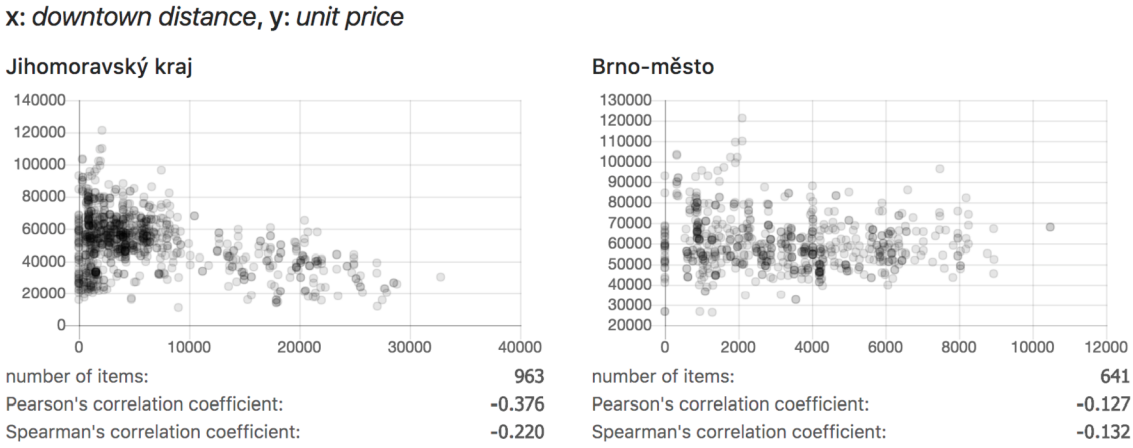


Obrázek 23 - závislost jednotkové ceny na čísle podlaží dle inzerce (vlastní)

Zajímavé by však mohlo být porovnat s korelací jednotkové ceny a číslem podlaží v domech bez výtahu. Bohužel však, jak jsem před tím uvedl, přítomnost výtahu jakožto parametr u *realitymix.centrum.cz* neexistuje, proto jsem parametr zamítnul.

Závislost jednotkové ceny na vzdálenosti do centra regionu

Jedná se čistě o experimentální údaj, získaný strojovým výpočtem, kdy je centrum regionu stanoveno pomocí služby Google Maps, poté je spočítána ortodroma k místu, kde se nachází konkrétní nemovitost.



Obrázek 24 - závislost jednotkové ceny na vzdálenosti do centra regionu (vlastní)

I skrze tato fakta, lze z uvedeného grafu pozorovat mírnou korelaci, zvláště pak u Jihomoravského kraje jako celku. Vzhledem k vysoké neurčitosti se však v práci tímto dále zabývat nebudu, avšak pokud by se přehodnotil model stanovení centra regionu, popř. jiného

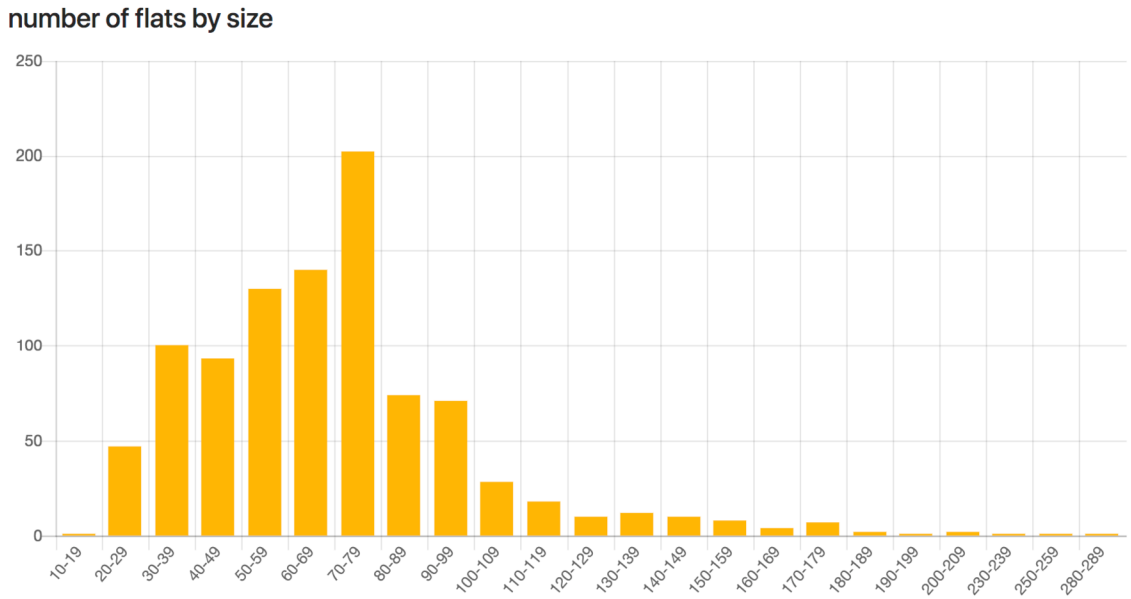
centra, stálo by za další prozkoumání, případně zvážení, jakým způsobem toto kritérium začlenit do výpočtového modelu tržní ceny.

6.3.3 Analýza četností

Analýza četností nám může prozradit informace o tom, jaká je nabídka, resp. jaké typy bytů se v nabídce vyskytují nejčastěji. Například podle (11) je z hlediska investic do nemovitostí nevýhodnější poptávat byty menších rozměrů, protože jsou snadněji pronajmutelné. Navíc v poměru nájemného na m^2 bývá menší byt lukrativnější, protože obecná představa je platit za bydlení co nejméně. Proto tak i potenciální % výnos může být větší u menšího bytu.

Četnost bytů dle plochy

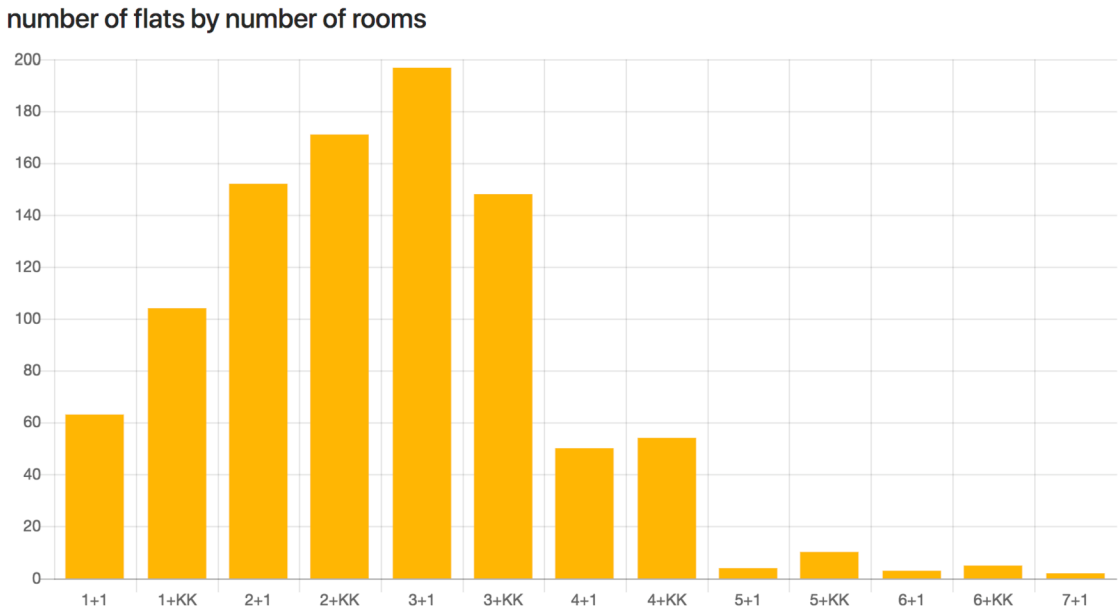
Rozdělení má asymetrický charakter, nedá se tedy zobecnit Gaussovou křivkou. Vidíme, že nejčastěji vyskytující byty (modus) se pohybují mezi 70-79 m^2. Dá se říct, že je to odpovídající byt pro průměrnou rodinu se 2 dětmi, partu studentů nebo i mladý pár s dítětem, popř. dožívající seniory. Menší byty jsou pak levnější varianty modusu, eventuálně byty pro jednotlivce. Větší byty lze považovat za vyšší standard, protože je zde menší nabídka, která zřejmě plyne z menší poptávky.



Obrázek 25 - počty bytů podle kategorií ploch v m^2 (vlastní)

Četnost bytů podle počtu místností

Další obrázek doplňuje předešlý. Jak jsme si ukázali před tím, na obrázku se vztahem mezi počtem místností a velikostí, můžeme očekávat, že histogram bude vypadat obdobně jako předešlý.



Obrázek 26 - počty bytů podle počtu místností a typu kuchyně (vlastní)

6.4 VÝPOČTOVÝ MODEL PRO TRŽNÍ OCENĚNÍ

Zde je popsán základní myšlenkový pochod při sestavování algoritmu pro oceňování a jak tento algoritmus funguje. Samotný základ procesu tržního ocenění je prostý a spočívá v tom, že obdržíme nějaké vstupní parametry, proběhne ocenění a získáme výstup, jak je znázorněno na následujícím obrázku:

Obrázek 27 - základní oceňovací workflow (vlastní)

6.4.1 Stanovení výběrového souboru

Předním vstupním parametrem pro ocenění je lokalita, neboť podle (1) je to nejzákladnější srovnávací kritérium. V této fázi se nabízí dva přístupy, jak vybrat vhodné okolní nemovitosti:

1. Na základě geografické vzdálenosti – prosté určení vzdálenosti mezi GPS souřadnicemi a výběru těch nejbližších.
2. Výběr na základě strojově čitelné adresy – preferovat nemovitosti ve stejné ulici, části obce, obce.

Vybral jsem si variantu č. 2, neboť v případě první varianty může dojít k tomu, že bude do výběru zahrnuta i nemovitost, která ačkoliv s oceňovanou nemovitostí sousedí, nachází se v ulici, která cenově neodpovídá. Např. Brno-Cejl vs. Brno-střed.

Proto tedy potřebujeme nejprve vstupní parametr adresy upravit na strojově čitelnou adresu. Využijeme zde opět geocoding, stejně jako při transformaci populace srovnávacích objektů. Nyní můžeme již strojově čitelné adresy srovnávacích nemovitostí porovnávat se strojově čitelnou adresou oceňovaného objektu.

Procházíme tedy postupně zprava položky adresy oceňované nemovitosti a porovnáváme je se srovnávacími nemovitostmi. Pokud dojde ke shodě, nemovitost je přidána do výběru. Pokud již máme dostatečný počet srovnávacích nemovitostí ve výběru, algoritmus končí. Pokud nemovitostí máme více, než potřebujeme, seřadíme je dle vzdálenosti a přebytečné vyloučíme na základě nejvzdálenějších (viz 1. varianta).

6.4.2 Vyřazení extrémů

Pro vyřazení extrémů využívám Grubbsův test, který již byl popsán v teoretické části. Vyřazení probíhá na základě jednotkové ceny, protože, jako poměr ceny a plochy, lépe reflektuje extrémy. Extrémy jsou vyřazovány dříve, než jsou vypočteny jednotlivá kritéria, protože při testování mi vycházeli lehce příznivější výsledky.

6.4.3 Navržený výpočtový model

Hned z kraje se nabízí několik přístupů a to:

1. Využít klasické ocenění přímým porovnáním, kdy si stanovíme kritéria, na základě nich nemovitosti srovnáme a poté upravíme cenu dle indexu odlišnosti.
2. Využijeme regresi pro stanovení koeficientů polynomu, který poté využijeme pro ocenění.
3. Využijeme metod strojového učení pro hledání řešení (neuronové sítě, genetické algoritmy, …)

Určitě by se daly vymyslet i další výpočetní modely, jak nalézt tržní cenu, pro naše účely však postačí tyto. 3. přístup je celý na samostatnou práci v této oblasti, takže jsem jej vyloučil okamžitě. 2. přístup vyžaduje pokročilejší analytické a statistické metody jako je analýza rozptylů atd., takže jsem jej také vyloučil. Zbyl nám tedy 1. „znalecký" přístup.

Pokud znalec hodnotí kritérium mezi oceňovanou a srovnávací nemovitostí, dokáže se zamyslet, která nemovitost je na tom lépe a stanovit míru tohoto rozdílu, která je reflektována v koeficientu odlišnosti. Pro stroj je tento úkol o dost obtížnější, protože umí pracovat jen s čísly a pojmy jako „nemovitosti", „kritérium" nebo „lokalita" jsou pro něj cizí. Musíme tedy vybrat srovnávací kritéria a vymyslet způsob, jakým vyjádřit tato jednotlivá kritéria pomocí čísel. Na základě předchozí kapitoly, kde byly uvedeny některé statistické údaje, jsem se rozhodl využít tato kritéria:

- Plocha
- Počet místností
- Typ kuchyně (1 nebo KK)
- Typ konstrukce
- Stav bytu

Pokud si vzpomeneme na kapitolu o datových typech, tak víme, že plocha a počet místností jsou číselné hodnoty, zbytek jsou výčtové typy, které nabývají předem daných hodnot. U číselných hodnot tedy stanovení koeficientu odlišnosti probíhá tímto jednoduchým způsobem:

$$\boldsymbol{K_i = \frac{S_i}{C_i}} \tag{23}$$

Kde S_i je hodnota kritéria oceňované nemovitosti (*subject*) a C_i je hodnota číselného kritéria srovnávací nemovitosti (*comparable*). Pokud budou hodnoty stejné (např. stejná plocha), bude koeficient K_i roven 1. Pokud naše oceňovaná nemovitosti bude lepší, co do konkrétního kritéria, bude koeficient > 1.

Problém nastává u kritérií, které nemají číselný charakter. V našem případě se jedná o skupinu výčtových typů. Zde jsem se rozhodl využít statistiky z předešlé kapitoly a věřit, že uvedený průměr u každého prvku každého výčtu, při daném množství vzorků, dostatečně reflektuje cenový rozdíl. Může to znít trochu těžkopádně, proto je nasnadě příklad:

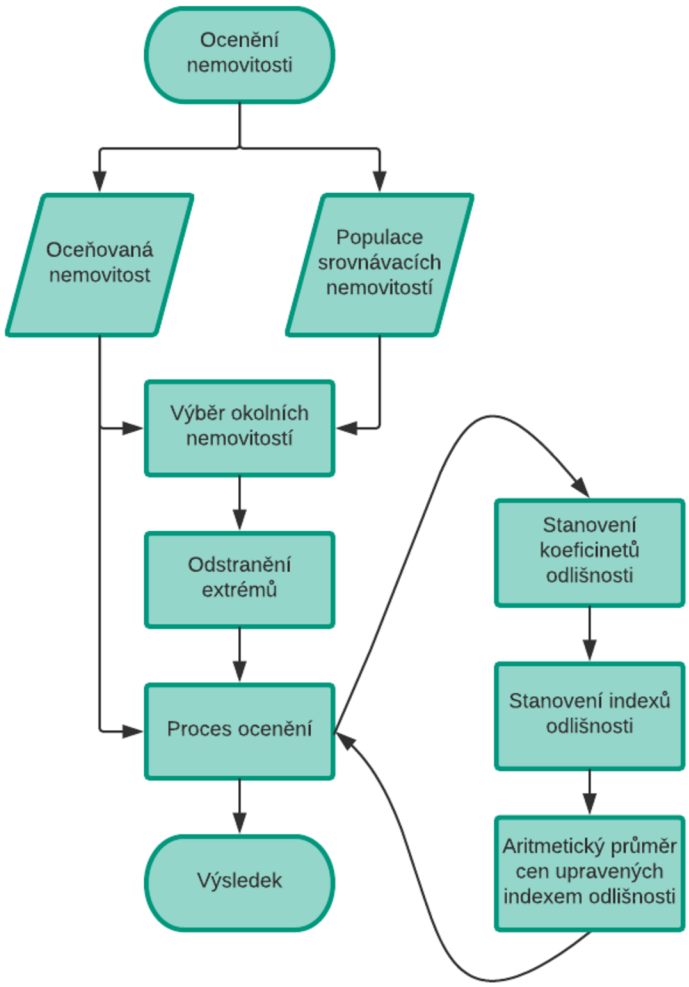
Řekněme, že kritériem je typ konstrukce, že naše oceňovaná nemovitost je zděna, ale srovnávací je panelová. Najdeme si tedy, že průměr pro zděné byty v Jihomoravském kraji je 53 872 Kč/m^2, průměr pro panelové stavby v Jihomoravském kraji je 46 595 Kč/m^2. Nyní již máme číselné hodnoty, které můžeme pohodlně dosadit do předešlého vzorce a získat tak koeficient odlišnosti i pro výčty.

Takto stanovené koeficienty odlišnosti bohužel nereflektují váhu daného kritéria. Pokud totiž koeficient určuje znalec, může jeho volbou udávat nejen odlišnost, ale i váhu s ohledem na ostatní kritéria. Stroj jako takový kritéria nerozeznává a vyhodnocuje je nezávisle na sobě. Proto jsem se rozhodl počítat index odlišnosti I_j váženým průměrem, kde váhy jsou předem stanoveny. Hodnoty vah by se měly určit nějakým aproximačním či regresním způsobem, kdy budeme zkoušet váhy různě přibližovat a hledat minimální chybu. Tento způsob však není jednoduchý a překročil by rámec této práce, proto jsem se rozhodl váhy nastavit ručně s ohledem na minimalizaci chyby.

Závislost ceny na ploše bytu je sice signifikantní, nemyslím však, že by byla směrodatná vzhledem k testování, rozhodl jsem se pracovat s celkovými cenami, namísto jednotkových. Plocha je pouze srovnávací kritérium s velkou vahou. Konečná cena P je vypočtena ze vztahu:

$$\boldsymbol{P = \sum_{j=1}^{m} P_{c,j} I_j} \qquad (24)$$

kde $P_{c,j}$ je inzertní cena j-té srovnávací nemovitosti a I_j je její index odlišnosti.



Obrázek 28 - vývojový diagram oceňovacího procesu (vlastní)

6.5 TESTOVÁNÍ PŘESNOSTI OCEŇOVACÍHO MODELU

Testování není jednoduchá záležitost ve světě samotného software a představuje poměrně rozsáhlý obor. Základem pro testování je vědět, jak se má testovaný subjekt chovat a jaké má vykazovat vlastnosti. V našem oboru to tedy představuje ověřit funkcionalitu na nějaké nemovitosti, u niž známe cenu a výsledek s touto cenou porovnat. Problémem je však takové nemovitosti sehnat. U jednotlivých inzerátů totiž může docházet k výskytu nadproporcionálních cen, které je nutno redukovat koeficientem redukce na pramen ceny. Na druhou stranu, vzhledem k dostatku dat, je větší šance eliminovat extrémy. V případě proběhlých transakci, mnoho snadno dostupných dat není.

Někdo by mohl namítat, proč nebyla využita nějaká konvenční tržní oceňovací metoda, tržní cena nebyla stanovena pomocí ní a toto nebylo porovnáno s výstupem z aplikace. Je to z toho důvodu, že zde může panovat zaujatost vůči vlastnímu výtvoru.

Rozhodl jsem se tedy pro hybridní řešení, kdy se navzájem mezi sebou porovnají všechny inzeráty a zároveň zde bude nastíněno ocenění na několika reálně proběhlých transakcích.

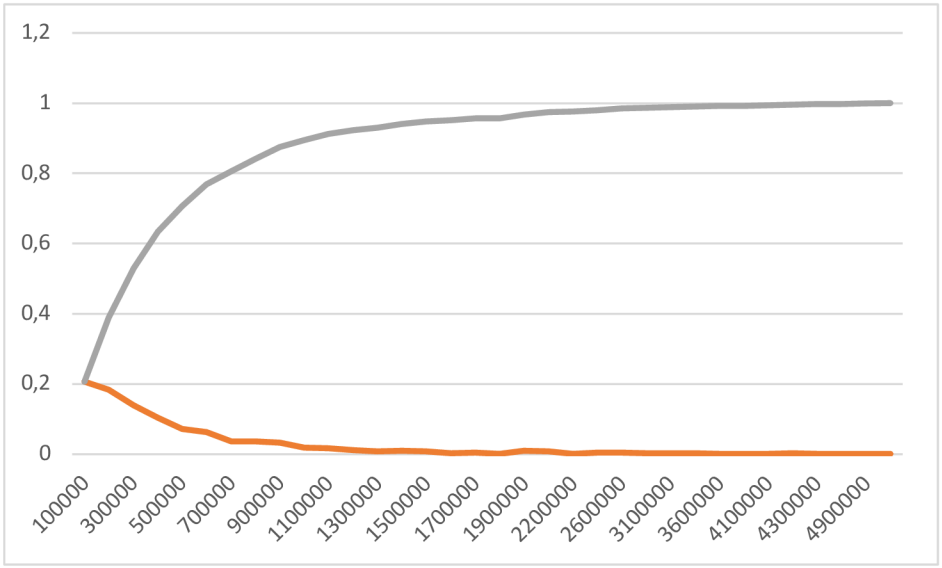
6.5.1 Testování na inzerátech

Vzhledem k tomu, že zobchodované nemovitosti, jejich ceny a další parametry se obtížně shánějí, rozhodl jsem se k testování přesnosti využít populaci všech srovnávacích nemovitostí, takže se nemovitosti porovnají mezi sebou. Výhodu tohoto přístupu vidím v tom, že mám k dispozici velké množství srovnávacích nemovitostí, takže lze provést tolik testů, kolik máme nemovitostí v databázi. Další výhodou je, že se pohybujeme v inzertních cenách, takže nepotřebujeme zde některou z cen přepočítávat koeficientem redukce na pramen ceny, čímž se vyhneme možnému zkreslení výsledné chyby.

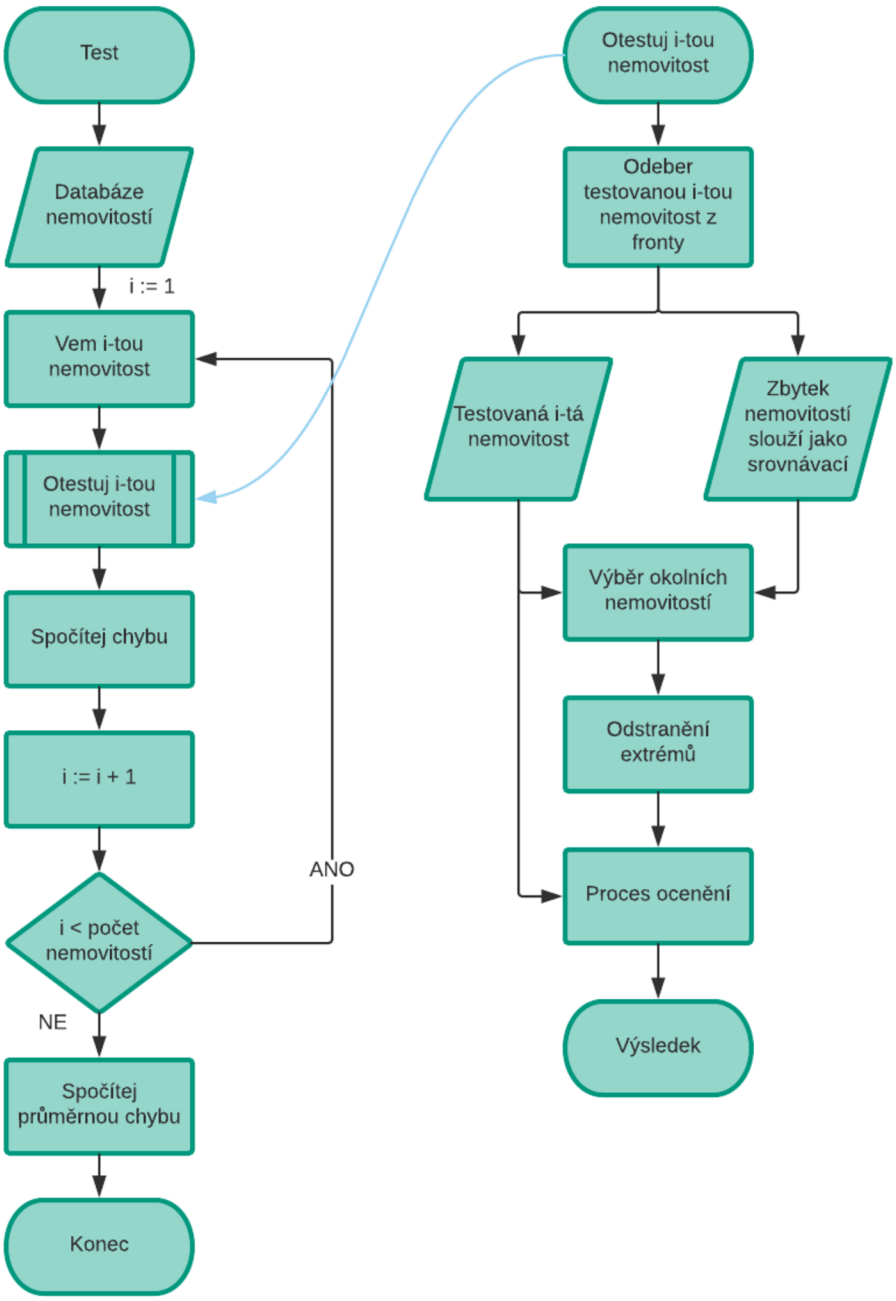
Po každém porovnání určíme chybu jako absolutní rozdíl odhadnuté ceny a inzerované ceny:

$$\boldsymbol{chyba} = |\boldsymbol{odhadnutá\ cena} - \boldsymbol{inzerovaná\ cena}| \tag{25}$$

Když tedy test spustíme, proběhne tolik ocenění, kolik máme srovnávacích nemovitostí v databázi (v našem případě 963). Na následujícím obrázku na spodní křivce vidíme v relativních četnostech, jak nejmenší chyba se vyskytuje nejvíce a postupně, se zvyšujícím se rozdílem částek, relativní četnost klesá. Horní křivka představuje kumulativní relativní četnost. Dá se tedy podle obrázku tvrdit, že s pravděpodobností 80 % bude chyba v odhadu tržní ceny nemovitosti cca 700 000 Kč.



Obrázek 29 - graf relativní četnosti a kumulativní relativní četnosti chyb (vlastní)



Obrázek 30 - Vývojový diagram testování přesnosti na vlastní DB (vlastní)

6.5.2 Testování na skutečně proběhlých transakcích

Pro názornost je zde i test na skutečně zobchodovaných nemovitostech. Jedna z nevýhod je tedy nízký počet dostupných údajů, další nevýhodou je nutnost přepočtu koeficientem

redukce na pramen ceny, abychom mohli výsledné ceny porovnat a určit chybu v odhadu. Koeficient redukce na pramen ceny jsem se rozhodl určit dle (12), z tab. 70, hodnota pro byty v Brně 0,91 a byt v Šatově 0,82.

Adresa	Počet pokojů	Druh vlastnictví	Materiál	Velikost	Stav bytu	Zobchodovaná cena bez daní a provizí	Vypočtená cena	Vypočtená cena po redukci	Rozdíl
Brno, Rybkova	4+1	Družstevní	Zděný	102	Po rekonstrukci	5 100 000	5 181 883	4 715 513	384 487
Brno, Cejl	2+kk	Osobní	Zděný	48	Po rekonstrukci	2 400 000	3 074 318	2 797 629	397 629
Šatov, Znojmo	3+1	Osobní	Zděný	100	Velmi dobrý	2 500 000	2 791 066	2 288 674	211 326

Tabulka 2 - porovnání skutečných cen s vypočtenými (vlastní)

Z použitých nemovitostí lze tedy konstatovat, že maximální chyba je cca 400 000 Kč, což by odpovídalo 63 % na našem předešlém obrázku, které v tomto případě představují 100 % odhad ceny pro tuto chybu. Z toho lze usuzovat, že tento test dopadl příznivěji než testování na svých datech. Vzhledem ale k počtu použitých nemovitosti z reálných obchodů, se na tuto informaci nedá spolehnout. Pro ověření tohoto tvrzení by bylo tedy třeba nasbírat další informace o reálných obchodech a udělat podrobnější analýzu.

7 ZÁVĚR

Výhody aplikace, jež byla zpracována v rámci této diplomové práce:

- Rychlé a snadné ocenění bytu v Jihomoravském kraji
- Agregovaná data na jednom místě
- Determinismus – při zachování stejné populace srovnávacích nemovitostí (databáze), stejného zdrojového kódu a stejných vstupních parametrech, je cena při ocenění vždy stejná
- Přehledná statistika cen v jednotlivých lokalitách, analýza jednotlivých parametrů
- Možnosti dalšího rozšíření funkcionality

Známé slabé stránky a nedostatky:

- Výsledná cena odpovídá inzertním, je tedy nutné ji redukovat koeficientem redukce na pramen ceny, abychom získali výslednou tržní cenu
- Databázi je nutné obnovit vždy, pokud chceme aktuální údaje
- Pokud některý z inzertních webů upraví grafické rozhraní, je nutné upravit kód aplikace
- Vlivem malé angažovanosti strojového učení a nedostatečným uvědoměním inzerentů může docházet k chybné interpretaci dat. Například existují inzeráty, kde je do adresy inzerentem zadaná zajímavější lokalita (např. centrum města), kdežto v popisku je skutečná adresa, která už v tak dobré lokalitě není (např. Brno-Jundrov).

Tím, že se jedná o inzerované ceny, které bývají mnohdy vyšší vlivem marží nebo touze po zisku inzerenta, tak i výsledná tržní cena bude tímto faktorem zkreslená. Podle (11) se navýšení ceny pohybuje mezi 5 a 25 %. Pro „narovnání“ tedy používáme koeficient redukce na pramen ceny, abychom získali reálnější představu o tržní ceně, která může být takto skutečně zobchodována.

Samozřejmě jako každý rozsáhlejší program, tak i tento může obsahovat chyby, které mohou způsobovat mírné zkresleni statistik a tím pádem i samotnou tržní cenu. Jako kritická místa bych považoval:

- Kontrola duplicit
- Výběr srovnávacích objektů
- Na pevno nastavené váhy jednotlivých srovnávacích kritérií

Teprve až rozsáhlejší testování v delším časovém horizontu odhalí ty nejvíce majoritní chyby. Bohužel však v rámci diplomové práce není dostatek času na vývoj a důkladné testování.

7.1 MOŽNOSTI DALŠÍHO ROZVOJE

Trh jako takový je nesmírně zajímavý. Ten realitní o to víc, protože nemovitosti patří k nejdražším komoditám, zároveň nemovitosti v dlouhodobém horizontu rostou na ceně, což je jedna z věcí proč se vyplatí nemovitosti vlastnit. Ačkoliv jsou nemovitosti poměrně málo likvidní, dají se během doby vlastnění pronajímat a tím mohou generovat další profit. Jaká je tedy vztah mezi tržní cenou nemovitosti a cenou nájemného? Má vliv počet obyvatel nebo hustota obyvatelstva vliv na cenu nemovitosti?

S potenciálním nasazením umělé inteligence (AI) a sofistikovanějších algoritmů by se potenciálně dalo docílit přesnějších statistik a relevantnějším údajům, které se mnohdy skrývají například v popisku. Dalším využitím AI by mohlo být rozpoznávání fotek, které by se dalo využít na testování duplicit a hodnocení kvality inzerátu. Podle kvality inzerátu by se pak mohla nastavovat váha na výslednou cenu při ocenění. V neposlední řadě bude taky třeba rozklíčovat poznámku k ceně, která obsahuje poměrně cenné informace (zda je cena s/bez DPH, zda je v ní zahrnuta provize RK apod.)

7.2 VYUŽITÍ

Prakticky tento software může sloužit k rychlému, orientačnímu odhadu ceny. Pokud by se z něj udělala mobilní aplikace, mohl by pomáhat nakupujícím při prohlídkách bytu, kdy mohou zjistit, zda se je realitní kancelář nesnaží „napálit"; realitním makléřům pro stanovení optimální tržní ceny a investorům hledat vhodné nemovitosti k investicím, u kterých dosáhnou co nejvyšších výnosů. Pokud totiž stanovíme výnos pro jednoduchost jako:

$$\boldsymbol{výnos} = \frac{\boldsymbol{roční\ zisk\ z\ nájemného}}{\boldsymbol{pořizovací\ cena}} \tag{26}$$

Pak nejvyššího výnosu získáme buď vysokým výnosem z nájemného nebo nižší pořizovací cenou. Nikdo však nebude radostí skákat do stropu, když mu dáte vysoký nájem, proto je nižší pořizovací cena nasnadě. V této situaci může tento nástroj tedy velmi pomoct.

7.3 POUŽITÉ NÁSTROJE

Pro vývoj této aplikace jsem využil jazyk *JavaScript*, jak pro klientskou část, tak pro serverovou část. Aplikace byla vyvinuta na operačním systému Mac OS, a ačkoliv by měla být multiplatformní, nebyla na jiném operačním systému testována. Aplikace je nazvána *Anubis*, což v podstatě vystihuje, co aplikace dělá:

- *Aggregation Network Utility* – agregace dat v rámci internetu
- *Broker Inquire System* – zpracování dat pro zainteresovaného uživatele

7.3.1 Serverová část

Běží v prostředí *Node.js (v8.9.4)*, využívá grafové databáze *Neo4j*. Serverová část je v podstatě zákulisí celé aplikace, protože má na starosti vše, co je zde zmíněno, tedy:

- Těžbu dat
- Veškeré zpracování dat
- Oceňování

Server pro komunikaci nabízí *GraphQL API*, které je taktéž navrženo společností Facebook, Inc.

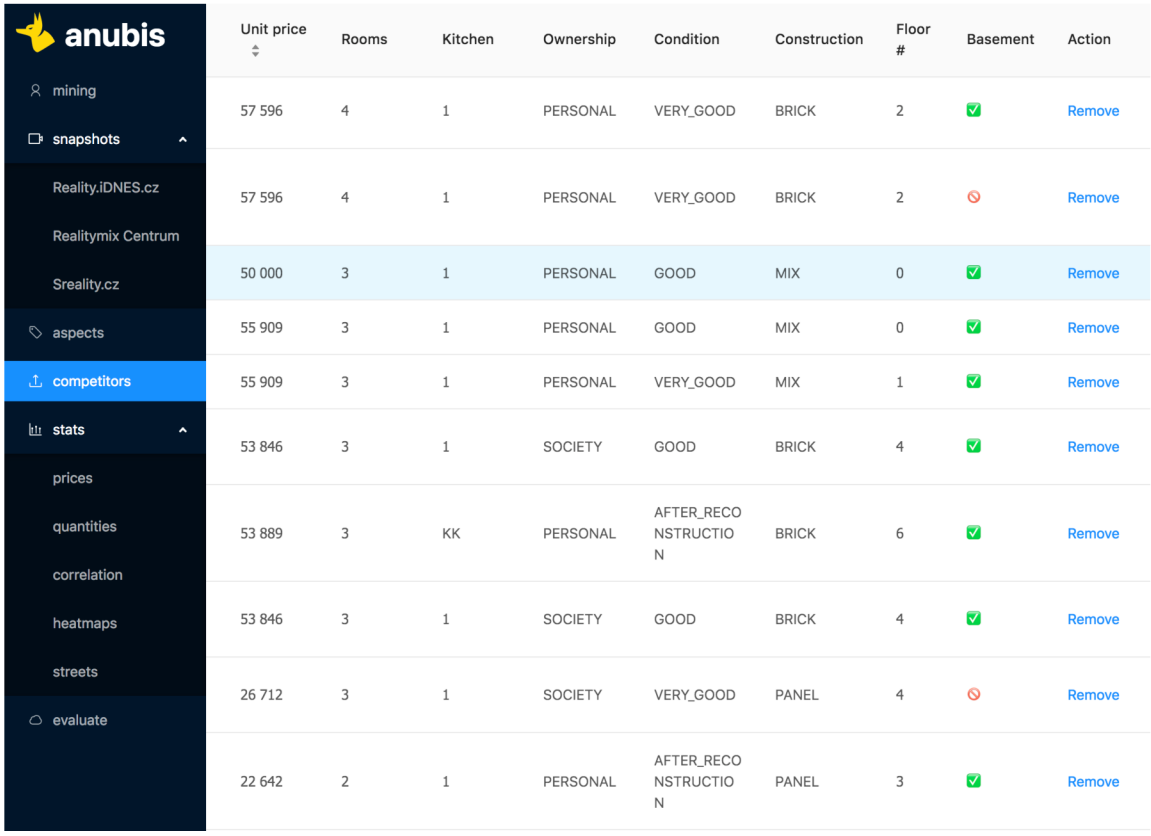
7.3.2 Klientská část

Tvoří grafické rozhraní (*user interface – UI*) pro komunikaci uživatele se serverem. Pro tvorbu jsem využil knihovny *React* od společnosti Facebook, Inc. Pro samotné ovládací prvky byla použita knihovna grafických komponent Ant design a jako jazyk UI jsem zvolil angličtinu pro možnost budoucího rozšiřování. V případě klientské části je jazyk JavaScript interpretován přímo konkrétním webovým prohlížečem. Já jsem pro testování používal Google Chrome 66.0.3359.181 (Official Build) (64-bit).

Klientská část je špičkou ledovce, její hlavní smysl je prezentovat data uživateli, má tedy na starosti:

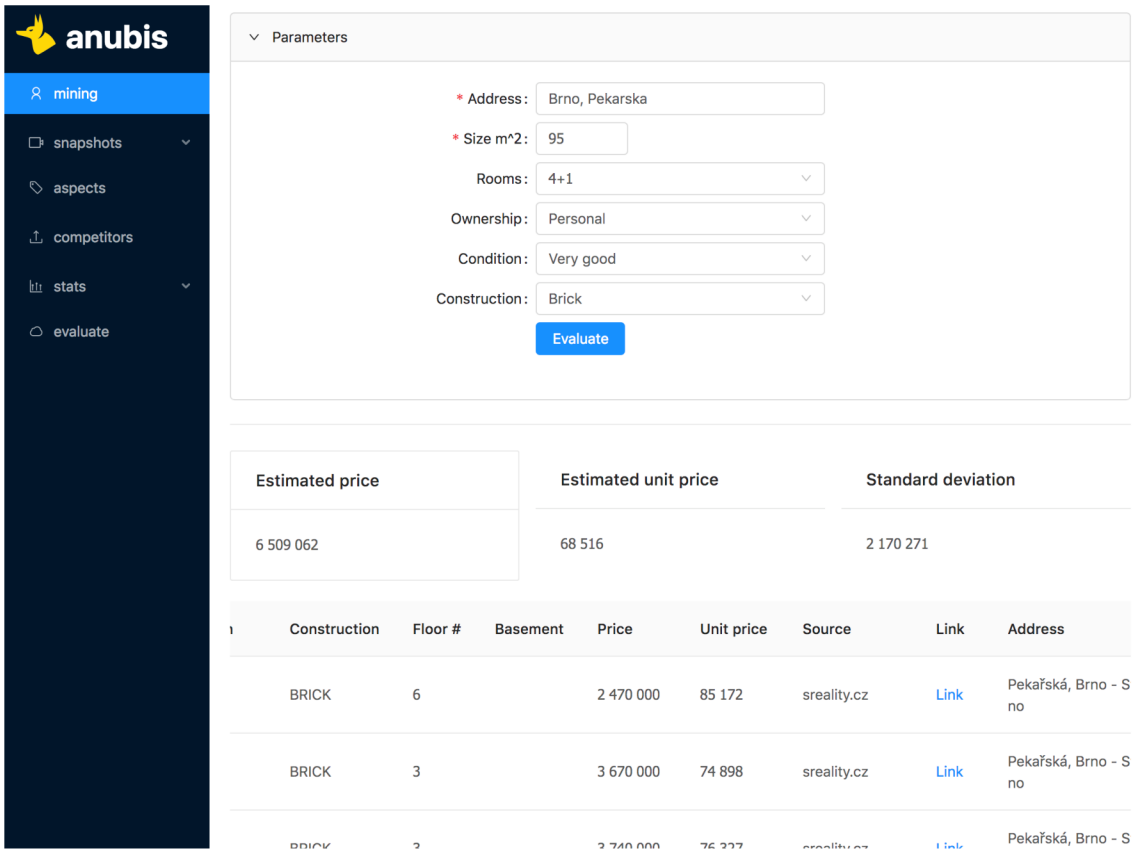
- Prezentace statistik v přehledné formě
- Zobrazovat formulář pro ocenění a umožňovat jeho komunikaci se serverem

Tato část aplikace je tedy závislá na serveru a tvoří jen jakousi vrstvu mezi ním a uživatelem. Na následujícím obrázku lze vidět ukázku grafického rozhraní, zde konkrétně tabulku populace srovnávacích nemovitostí.



Obrázek 31 - ukázka klientské části – tabulka srovnávacích nemovitostí (vlastní)

K samotnému ocenění pak slouží následující formulář, který po odeslání zobrazí přibližnou cenu, přibližnou jednotkovou cenu a směrodatnou odchylku cen srovnávacích nemovitostí. Taktéž nechybí ani tabulka s výpisem srovnávacích nemovitostí, které byly využity pro konkrétní ocenění.



Obrázek 32 - ukázka klientské části – proces ocenění (vlastní)

7.3.3 Shrnutí

Aplikace, ač se úkol na začátku nezdál být příliš složitý, nakonec nabyla poměrně sofistikované podoby. Její rozsah je ~5 500 řádků zdrojového *JavaScript* kódu i přes to, že na spoustu dílčích částí bylo využito různých knihoven. Díky nutnosti hostovat serverovou část, což vyžaduje netriviální znalosti ze základů Unixových systémů a počítačových sítí, není tato aplikace určena pro koncové klienty, resp. běžné uživatele. Z tohoto důvodu ani neexistuje snadno distribuovatelná a snadno instalovatelná verze, což ani nebylo předmětem práce.

Existuje však možnost aplikaci provozovat na nějakém vzdáleném serveru a přistupovat přes webový prohlížeč jen ke klientské části. Bohužel toto taktéž není jednoduché řešení z hlediska realizace a finanční náročnosti. Z tohoto důvodu jsou v této práci prezentovány, jak tento software funguje a výstupy z něj, a ne software samotný.

8 SEZNAM OBRÁZKŮ

9 SEZNAM TABULEK

10 SEZNAM ZKRATEK A POJMŮ

Zkratky

AI – artificial intelligence

DPH – daň z přidané hodnoty

RK – realitní kancelář

ER – entity relationship

THU – technickohospodářský ukazatel

RSS – residual sum of squares

ETL – extract, transform, load

ID – identifikátor

GPS – global positioning systém

API – application program interface

JSON – javascript object notation

PENB – průkaz energetické náročnosti budovy

UI – user interface (grafické rozhraní)

KK – kuchyň, která součástí obytné místnosti (není samostatně)

Pojmy

populace (srovnávacích nemovitostí) – veškeré nemovitosti zjištění z inzerátu, které má aplikace k dispozici

výběr (srovnávacích nemovitostí) – nemovitosti, které byli algoritmem vybrány pro konkrétní případ ocenění

flow (procesu) – označuje posloupnost kroků činnosti

11 SEZNAM POUŽITÝCH ZDROJŮ

1. BRADÁČ, Albert a kol. *Teorie a praxe oceňování nemovitých věcí.* Brno : Akademické nakladatelství CERM, s.r.o. Brno, 2016. ISBN 978-80-7204-930-1.

2. HENDL, Jan a kol. *Statistika v aplikacích.* Praha : Portál, s.r.o., 2014. ISBN 978-80-262-0700-9.

3. NEUBAUER, Jiří, SEDLAČÍK, Marek a KŘÍŽ, Oldřich. *Základy statistiky.* Praha : Grada Publishing, a.s., 2016. ISBN 978-80-247-5786-5.

4. HAN, Jiawei, KAMBER, Micheline a PEI, Jian. *Data Mining.* Waltham, MA : Morgan Kaufmann Publishers, 2011. ISBN 978-0-12-381479-1.

5. KIMBALL, Ralph a CASERTA, Joe. *The Data Warehouse ETL Toolkit.* Indianapolis, IN : Wiley Publishing, Inc., 2004. eISBN 0-764-57923-1.

6. Energetický štítek při prodeji bytu v osobním vlastnictví. *Energitický štítek domu.* [Online] https://www.energeticky-stitek-domu.cz/kdo-musi-mit-enereticky-stitek/pri-prodeji-bytu.

7. Global Positioning System. *Wikipedia.* [Online] https://cs.wikipedia.org/wiki/Global_Positioning_System.

8. Google, Inc. Geocoding Service. *Google Maps - API docs.* [Online] https://developers.google.com/maps/documentation/javascript/geocoding.

9. KRTIČKA, Luděk. *Úvod do kartografie.* Ostrava : Ostravská univerzita v Ostravě, 2007. ISBN 978-80-7368-344-3.

10. ZUZÁK, Vladimír. Družstevní byt: Je váš, nebo není? *Peníze.cz.* [Online] https://www.penize.cz/nemovitosti/299983-druzstevni-byt-je-vas-nebo-neni.

11. TEMROVÁ, Pavla. *Realitní kuchařka.* Ostrava : AMOS repro, spol. s r. o., 2017. ISBN 978-80-260-5163-3.

12. CUPAL, Martin. *Vliv koeficientu redukce na zdroj ceny na výsledný index odlišnosti při komparativní metodě oceňování nemovitostí.* Brno : Ústav soudního inženýrství, 2010.

12 SEZNAM PŘÍLOH

Příloha č. 1 – Export populace použitých srovnávacích nemovitostí