UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Vizualizace mnohorozměrných dat s R

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Karel Hron Ph.D.**
Vypracoval(a): **Tadeáš Václavek**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Matematika–ekonomie se zaměřením na bankovnictví/pojišťovnictví
Forma studia: prezenční
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Tadeáš Václavek

Název práce: Vizualizace mnohorozměrných dat s R

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron Ph.D.

Rok obhajoby práce: 2017

Abstrakt: Vizualizace mnohorozměrného datového souboru je základním nástrojem průzkumové statistické analýzy, kdy si chceme udělat představu o struktuře dat, resp. o vztazích mezi proměnnými. Cílem bakalářské práce bude na názorných příkladech seznámit s užívanými grafickými prostředky pro vizualizaci mnohorozměrných dat s využitím statistického softwaru R a provést jejich případné srovnání. Za tímto účelem jsou využita data z Major League Baseball, nejslavnější baseballové ligy na světě.

Klíčová slova: data, vizualizace, software R, baseball, jádrové odhady hustoty, bodový graf, chí-graf, biplot, trellis grafika

Počet stran: 54

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Tadeáš Václavek

Title: Visualization of multivariate data with R

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron Ph.D.

The year of presentation: 2017

Abstract: Visualization of multidimensional data set is a basic tool for exploratory statistical analysis if we want to get an idea about the structure of data or about relationships between variables. The aim of this thesis is to present used graphical tools for visualization of multidimensional data in illustrative examples using the statistical software R and to compare these tools if needed. For this purpose real data from Major League Baseball, the most famous baseball league, are utilized.

Key words: data, visualization, software R, baseball, kernel density estimation, scatterplot, chiplot, biplot, trellis graphics

Number of pages: 54

Number of appendices: 2

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne 24. dubna 2017 .....

podpis

Poděkování

Rád bych poděkoval zejména vedoucímu bakalářské práce doc. RNDr. Karlu Hronovi, Ph.D. za obětavost a čas, který mi věnoval při konzultacích. Dále bych rád poděkoval své rodině a blízkým za trpělivost, kterou se mnou měli při tvorbě této práce.

Obsah

Úvod

Obsahem této práce je popsat metody vizualizace statistických dat, neboť pouze na základě získaných dat je často velmi obtížné vyvodit konkrétní závěry. Jako data jsou v této práci použity základní baseballové statistiky. Již na úvod je třeba zmínit, že ke každé použité metodě je přiložen i příslušný kód ve statistickém softwaru R.

V první kapitole se seznámíme s pravidly baseballu, charakteristikou pozic a hlavními baseballovými statistikami. O tyto znalosti se budeme opírat i v dalších kapitolách. Dále se budeme zabývat jednorozměrnými daty. Tato data si porovnáme pomocí základních číselných charakteristik a vykreslíme si jádrové odhady hustoty. Kapitola třetí pojednává o datech dvourozměrných. Postupně bude představen bodový graf a chí-graf. V závěru kapitoly si opět ukážeme příslušné jádrové odhady hustoty. V poslední kapitole jsou představeny grafické nástroje pro vizualizaci vícerozměrných dat.

Toto téma jsem si vybral ze dvou důvodů. Zaprvé mě zejména zaujalo to, že je práce zaměřená spíše na praxi než na čistou teorii. Druhým důvodem je možnost aplikovat výše zmíněné metody na baseballová data a dozvědět se o nich něco více. Jako cíl jsem si dal zjistit a dokázat, zda se různé skupiny hráčů liší v útočné činnosti.

Bakalářská práce byla vysázena v systému LaTeX 2_ε.

Kapitola 1

Základy baseballu

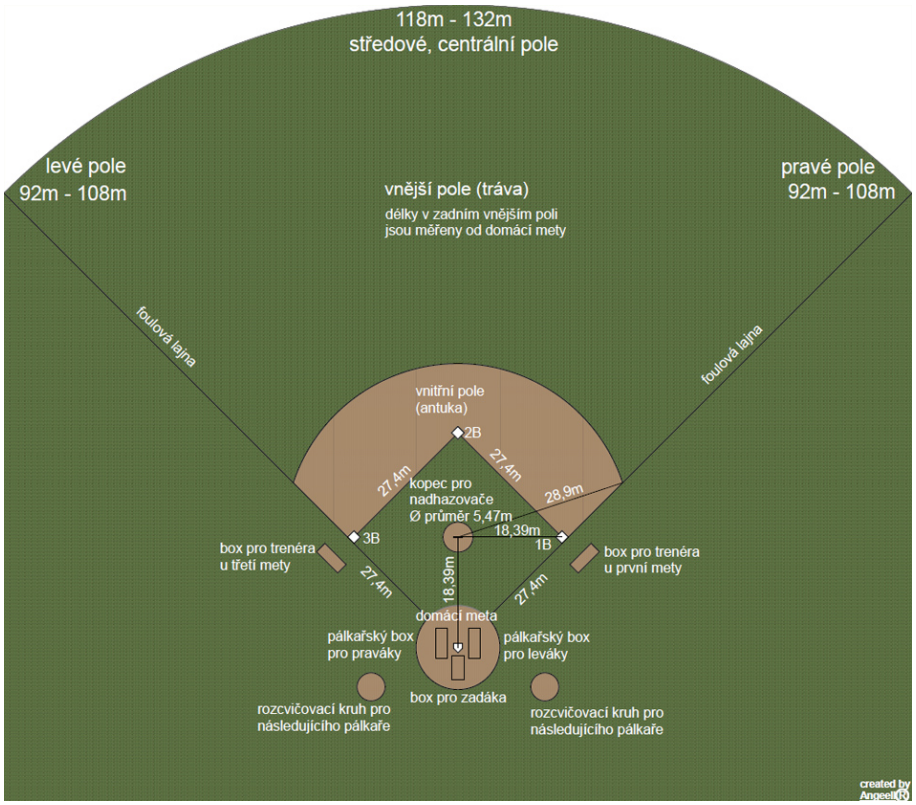
V této kapitole se postupně seznámíme se vším podstatným pro pochopení dat užitých v této bakalářské práci. Začneme se základními pravidly baseballu, užívanou terminologií, charakteristikou jednotlivých pozic, na kterých hráči nastupují, vlastnostmi jednotlivých hráčů a v neposlední řadě s hlavními statistikami v baseballu. Představíme si také nejslavnější baseballovou ligu na světě, ze které pochází veškerá data. Při tvorbě této kapitoly bylo čerpáno zejména z [1], [17] a [6].

1.1. Pravidla baseballu

Baseball je pálkovací hra, oblíbená zejména v Severní a Jižní Americe, Asii a momentálně zažívá svůj rozmach i v evropských zemích. Nastupují proti sobě dva týmy skládající se z devíti hráčů[1], kteří se snaží pro svůj tým získat co nejvíce bodů a zároveň zamezit soupeři ve skórování. Bod může získat pouze tým, který je právě v útoku. Nadhazovač hodí míč směrem na pálkaře, který se jej snaží zasáhnout a odpálit do vymezené výseče tak, aby jej druhý tým nemohl vyautovat. To je možné udělat několika způsoby. První možností je, že bránící hráč zachytí míč přímo ze vzduchu, než se dotkne země. Dále pak příhozem na první metu dříve, než se na ni dostane hráč, který odpálil, nebo tečováním hráče mezi metami. Další možností je tzv. strikeout, kdy pálící hráč třikrát „prokouká“

[1] Ve většině soutěží má každý tým právo využít tzv. DH rule, kdy nadhazovač nechodí odpalovat, ale zastupuje ho suplující pálkař. Tento tým má tedy hráčů deset.

nebo mine dobře nadhozený míč. Útočící hráč se snaží dostat postupně přes první, druhou a třetí metu až na metu domácí, za což si jeho tým připíše bod. Jestliže se hráč dostane na některou z prvních třech met a nemůže pokračovat dále, čeká na odpal svého spoluhráče, který jde pálit po něm. Útok končí, jakmile bránící tým zahraje tři auty. Potom se týmy vymění a tým v obraně jde útočit neboli pálit a naopak útočící jde do obrany. Hra je rozdělena na devět směn, přičemž jedna směna znamená vystřídání týmu jak v útoku, tak i v obraně. Baseball se hraje na speciálním hřišti tvaru čtvrtkruhu, které je rozděleno na vnitřní pole, tzv. infield, a vnější pole, tzv. outfield.



Obrázek 1.1: Baseballové hřiště s popisky [6]

1.2. Charakteristika pozic

V této kapitole se pokusíme poodkrýt vlastnosti jednotlivých hráčů a popsat pozice, na kterých hrají. Jak již bylo řečeno, baseball hraje devět hráčů, přičemž

každý má svoji vlastní roli v týmu. Některé se od sebe výrazně odlišují zejména s ohledem na obrannou činnost.

- **Catcher** neboli zadák je hráč, který chytá míč od nadhazovače a řídí celou obrannou hru. Catchera hrají především hráči vybaveni vůdcovskými schopnostmi a je požadováno, aby vynikali zejména v obranné hře. V útoku se od nich čeká průměrný či podprůměrný výkon, jelikož mnoho času na tréninku stráví komunikací s vnitřním polem a nadhazovačem.

- **First basemen** se často překládá jako první meta nebo prvometař. Tuto pozici zastávají zejména urostlejší hráči, jejichž cílem je chytat příhozy od ostatních vnitřních polařů. Hráče hrající na této pozici bychom mohli zařadit do skupiny tzv. corners. V útoku jsou to lídři týmu, jejichž úkolem je hlavně pálka. Obvykle vynikají zejména v homerunech a RBI, stejně jako třetí meta.

- **Third baseman** - tuto pozici zastávají stejně jako první metu hlavně vynikající pálkaři. Jejich pálkařské statistiky jsou velice podobné těm prvometařským, s tím rozdílem, že je kladen důraz i na jejich obrannou hru. Vyznačují se silnou rukou, jelikož přihazují stejně jako shortstop na nejdelší vzdálenost. Stejně tak jako prvometaři jsou zařazeni do skupiny corners.

- **Shortstop a second baseman** jsou jedni z nejvíce atletických hráčů na hřišti, rychlí a dynamičtí, jejichž úkolem je zastávat obrannou hru. V průběhu zápasu na ně jde nejvíce odpalů a tomu je podřízen i jejich trénink. Chytají mnoho míčů po zemi a rovněž trénují i souhru mezi sebou. Obě dvě pozice se označují jako tzv. middle infield.

- **Center fielder**, někdy označován jako středopolař, je hráč nastupující ve středu zadního pole. Vyznačuje se hlavně rychlostí a lehce nadprůměrnou útočnou hrou. V obraně je jeho nejdůležitějším úkolem chytat tzv. flyballs, tedy míče letící vzduchem.

- **Left filder a right fielder** neboli levý a pravý zadopolař. Pozice podobné střednímu poli.

- **Pitcher**, česky nadhazovač, je nejdůležitějším hráčem na baseballovém hřišti. On rozehrává hru a jeho výkon často rozhoduje o výsledku zápasu. Útočné hře se nevěnuje téměř vůbec. Jak jsme zmínili dříve, za nadhazovače chodí pálit suplující pálkař.

- **DH** bývá česky překládán jako suplující pálkař. Celá jeho práce spočívá v útoku. Do obrany se nedostane a tak může celý svůj trénink věnovat nácviku pálky.

Pro účely této bakalářské práce jsme jednotlivé posty rozdělili do tří skupin. Tou první je vnitřní pole, kam bychom mohli zařadit catchery, shortstopy a hráče hrající na druhé metě. Od této skupiny se budou očekávat slabší pálkařské výsledky. Druhou skupinou budou tzv. corners, tedy první a třetí mety. Pálkařské statistiky by u těchto hráčů měly výrazně převýšit ostatní pozice. Poslední skupinou je vnější pole, kde jsou zahrnuty všechny tři pozice v zadním poli. Jejich pálkařské statistiky jsou největší neznámou, ale lze předpokládat, že by měly být obecně lepší než ty u infieldu, ale horší než u již zmiňovaných corners. Nadhazovače a suplující pálkaře v této práci nebudeme brát v potaz, jelikož nadhazovači pálit nechodí a suplujícího pálkaře nelze zařadit podle obranné činnosti.

1.3. Hlavní baseballové statistiky

Dále se seznámíme s těmi nejdůležitějšími útočnými statistikami. Konkrétní data jsou vybrána z nejznámější baseballové ligy na světě zvané MLB tedy *Major League Baseball*, se kterou se seznámíme v další podkapitole. Budeme se zabývat pouze pálkařskými statistikami, tedy těmi útočnými. Pro potřeby této bakalářské práce jsme vybrali prvních 240 hráčů, seřazených podle počtu odehraných zápasů. Jedna sezóna MLB má 162 zápasů, není však obvyklé, aby jeden hráč odehrál za sezónu všechny.

- **At bats (AB)** - start na pálce. Za jeden zápas by měl hráč v průměru nastoupit k pěti startům na pálce a za sezónu je to průměrně 470.

- **Hits (H)** - odpal. Jako dobrý odpal se počítá tzv. single, double, triple a homerun. Velice důležitá statistika zejména pro výpočet AVG.

- **Doubles (2B)** - dvoumetový odpal. Jak již název napovídá, jedná se o odpal, kdy hráč doběhne přímo na druhou metu. 2B je velice důležitá statistika, podle které se pozná silový pálkař, který kromě dobrých pálkařských schopností musí být i rychlý, aby na druhou metu stihl doběhnout.

- **Triples (3B)** - třímetový odpal. Je podobnou statistikou jako dvoumetový odpal s tím rozdílem, že hráč musí doběhnout až na třetí metu. Třímetový odpal se ovšem vidí jen velmi ojediněle, jelikož ti nejlepší hráči v této statistice dosahují maximálně deseti odpalů za sezónu.

- **Homeruns (HR)** - nepřekládá se. Je jednou z nejvíce ceněných statistik v baseballu, kdy hráč musí odpálit míč až za homerunový plot, čímž může automaticky oběhnout všechny mety a získává tak pro svůj tým bod. Druhá možnost je oběhnout na jeden odpal do hřiště všechny mety a pak se jedná o tzv. inside the park homerun. To se ovšem může povést jen po chybě některého ze zadních polařů. Za vynikající, se považuje, když hráč odpálí přes třicet homerunů za sezónu. Není však výjimkou, že nejlepší hráči odpálí i přes čtyřicet pět homerunů.

- **Base on balls (BB)** - meta zdarma. Jestliže nadhazovač hodí pálkaři čtyři špatné nadhozy, hráč postupuje bez odpalu na první metu. Této statistice vládnou především hráči menšího vzrůstu, jelikož je pro nadhazovače těžší trefit se do jejich strike zóny.[2]

- **Hit by pitch (HBP)** - trefení. Pálkař trefený nadhozem se automaticky posouvá na první metu.

[2]Strike zóna je pomyslný čtverec vymezený nad domácí metou a mezi koleny a páskem pálkaře.

- **Batting average (AVG)** - pálkařský průměr. Vypočítá se jako podíl počtu hitů ku počtu at bats. Za velmi kvalitní se považuje průměr kolem 0,280. Tato statistika by se dala považovat za tu nejdůležitější, jelikož jsou v ní započteny všechny výše zmíněné statistiky a udává nám celkový pohled na útočnou sílu pálkaře.

- **Runs bated in (RBI)** - počet stažených bodů. Této statistice většinou vévodí hlavně siloví pálkaři. RBI se připíše pálkaři za každého hráče, který doběhne pro bod po jeho úspěšném odpalu. Pálkař může dosáhnout až čtyř RBI na jeden odpal. Takovýto odpal se nazývá grand slam a vidí se velmi ojediněle. Aby hráč dosáhl grand slamu, musí odpálit homerun a zároveň musí být obsazená první, druhá a třetí meta. Všichni tito hráči včetně pálkaře tedy po homerunu dobíhají pro bod.

- **Sacrifice fly (SF)** - sebeobětovací odpal. Po dlouhém odpalu, který je chycen zadopolařem ze vzduchu, je pálkař samozřejmě aut, ale jestliže je obsazena třetí meta, tak po chycení míče může doběhnout pro bod. Pálkaři se tak připisuje RBI a SF.

- **On-base percentage (OBP)** - úspěšnost postupu na metu. Statistika velmi podobná výše zmíněnému pálkařskému průměru. Vypočítá se jako podíl hitů, met zdarma a trefení ku počtu startů na pálce, met zdarma, trefení a sebeobětovacích odpalů, tedy (H+BB+HBP)/(AB+BB+HBP+SF).

1.4. Major League Baseball

Nejslavnější baseballová liga na světě, která se hraje od roku 1869 v severní Americe. Momentálně je rozdělená na dvě části, a to na americkou a národní ligu. Týmy v každé lize jsou pak ještě rozděleny do tří divizí. MLB, jak se této lize zkráceně říká, hraje 29 týmů z USA doplněných jedním týmem z Kanady. Každá sezóna je rozdělena na dvě části. Tou první je regular season, ve které každý tým odehraje 162 zápasů. Osm nejlepších týmů postoupí do postseason, kde se nejlepší

tým z americké a národní ligy utká ve finále o titul. Mezi nejslavnější týmy patří New York Yankees, kteří vyhráli titul 27 krát, St. Luis Cardinals s jedenácti tituly, Boston Red Sox nebo San Francisco Giants s osmi tituly. Veškeré statistické údaje použité v této práci pochází právě z této ligy. Čerpány byly z internetové stránky mlb.com, kde se dají najít veškeré informace o hráčích a týmech již od sezóny 1876.

Kapitola 2

Vizualizace jednorozměrných dat

V této kapitole se budeme zabývat jednorozměrnými daty a pomocí základních číselných charakteristik se je budeme snažit porovnat. Pokusíme se také hodnoty některého statistického znaku vykreslit pomocí jádrových odhadů hustoty, které si podrobněji vysvětlíme.

2.1. Základní číselné charakteristiky

Jednorozměrná data, jak již název napovídá, nám vznikají z naměřených hodnot jedné proměnné statistického znaku. Naše data budou obsahovat důležité pálkařské statistiky, které již byly představeny dříve. Ve třech tabulkách se nejdříve pokusíme poodkrýt základní rozdíly mezi třemi uvedenými skupinami pálkařů, tedy corners, vnitřními polaři a vnějšími polaři. K tomu nám poslouží číselné charakteristiky. Postupně budeme porovnávat aritmetický průměr, směrodatnou odchylku (označenou v tabulkách jako sd z anglického standard deviation), medián, modus, šikmost a špičatost. Jednotlivé charakteristiky, které si v průběhu podkapitoly vysvětlíme a ukážeme jejich výpočet, jsme spočítali pomocí softwaru R. V této podkapitole bylo čerpáno zejména z knihy [13] a knihoven [14] , [10].

	průměr	sd	medián	modus	šikmost	špičatost
AVG	0,2647	0,0262	0,2615	0,259	-0,01837	2,8617
HR	14,9778	10,2749	13	10	0,9782	3,6047
2B	23,8778	8,5497	23	25	0,1884	2,5567
RBI	55,9111	23,1855	51,5	41	0,4478	2,3826
H	125,0222	36,2934	128	172	-0,2129	2,1745
OBP	0,3271	0,0337	0,3205	0,314	0,8227	4,4153

Tabulka 2.1: Charakteristiky - vnější pole

	průměr	sd	medián	modus	šikmost	špičatost
AVG	0,2589	0,0309	0,2615	0,265	-0,2156	3,1501
HR	10,3889	6,2002	9,5	7	0,5383	2,6922
2B	22,4	7,7920	22	23	0,3130	2,9912
RBI	50,7333	17,3496	49	44	0,1716	2,7261
H	118,9667	34,7606	116,5	132	0,1767	2,8521
OBP	0,3132	0,0329	0,3145	0,334	-0,4177	3,3276

Tabulka 2.2: Charakteristiky - vnitřní pole

	průměr	sd	medián	modus	šikmost	špičatost
AVG	0,2660	0,0277	0,268	0,287	0,0800	2,8419
HR	20,8167	10,0347	18,5	18	0,4852	2,9075
2B	27,7	7,8076	29	33	-0,0322	2,3209
RBI	72,2833	23,6852	73	73	0,0619	3,0324
H	131,3667	34,4969	139,5	163	-0,2925	2,1009
OBP	0,3372	0,0380	0,329	0,307	0,8664	4,4096

Tabulka 2.3: Charakteristiky - corners

Podívejme se nejdříve na nejzákladnější, ale také nejdůležitější pálkařskou statistiku, kterou je AVG (pálkařský průměr). Jak vidíme, tak největší aritmetický průměr AVG mají dle předpokladu corners, ale zajímavý je jen malý odstup vnějších polařů. Dále si ukážeme, jak ve statistickém softwaru R spočítáme jednotlivé číselné charakteristiky. Nejprve je třeba si do R načíst příslušný datový soubor. To provedeme příkazem:

`> baseball <- read.csv("Baseball.csv", sep=";", header=TRUE, as.is=1)` nebo nám pomůže například program RStudio, který má speciální

funkci pro načtení dat. Dále bude v našem případě potřeba vybrat jen určitou část souboru. Například vybereme ze všech hráčů pouze vnitřní polaře:

```
> y=Baseball[Baseball$pos=="INF",]
```

Teď už je pouze třeba si ze skupiny vnitřních polařů vybrat hodnoty, pro které budeme chtít počítat příslušnou charakteristiku, řekněme aritmetický průměr. Zvolíme například pálkařský průměr, tedy AVG, ze kterého si vytvoříme vektor následujícím způsobem:

```
> z=y[,"avg"]
```

Nyní už se dostáváme k samotnému příkazu pro výpočet aritmetického průměru z našeho vytvořeného vektoru z. Ten vypadá následovně:

```
> mean(z)
```

Při ukázkách výpočtů dalších charakteristik budeme předpokládat, že soubor je již načten, skupina vybrána a také máme vytvořen vektor z.

Dalšími daty, kterými se budeme v této části zabývat, jsou H neboli hity. I zde opět vládnou corners, i když rozdíl není tak markantní jako například u HR. Zajímavá se zde zdá směrodatná odchylka (sd), která nám udává, jak moc jsou dané hodnoty koncentrovány kolem aritmetického průměru. Vidíme, že u všech tří skupin je tato hodnota poměrně vysoká, což znamená, že budou hodnoty této proměnné poměrně variabilní. Výpočet směrodatné odchylky v R je opět velmi jednoduchý a postačí nám za tímto účelem následující příkaz:

```
R> sd(z)
```

Dále můžeme porovnat HR (homeruny) a 2B (dvoumetové odpaly). Z tabulek 2.1 - 2.3 lze vyčíst výraznou převahu corners nad vnějšími polaři a zejména pak nad vnitřními polaři. Zajímavá je také výrazně vyšší hodnota modusu a mediánu u corners oproti ostatním dvěma skupinám. Modus je hodnota, která se v daném statistickém souboru vyskytuje nejčastěji. V našem případě je modus HR u corners více než dvojnásobný než u vnitřních polařů a výrazně vyšší je i modus 2B. Medián je hodnota, která dělí vzestupně seřazené výsledky na dvě stejně velké poloviny. Výpočet hodnot těchto dvou číselných charakteristik za pomocí softwaru R je velmi jednoduchý. Pro medián máme přímo vloženou funkci median,

kterou do R zadáme následně:

```
> median(z)
```

K výpočtu modu budeme muset do R nainstalovat knihovnu `modeest`. Dále zadáme funkci mfv (most frequent value), která je součástí této knihovny. Výsledný kód v softwaru bude vypadat následovně:

```
R> mfv(z)
```

Předposledním statistickým znakem, kterým se budeme zabývat, je RBI. Opět dle předpokladu vládnou corners, ale zdá se, že zde není tak velký rozdíl mezi vnitřními a vnějšími polaři. Při bližším prozkoumání tabulek najdeme také koeficient šikmosti, který nám udává, zda jsou hodnoty kolem středu, reprezentovaného aritmetickým průměrem, rozděleny souměrně nebo jsou nějak asymetrické. Zde vidíme, že zmíněný koeficient šikmosti je u vnějších polařů kladný a téměř trojnásobný oproti koeficientu u vnitřních polařů. Dá se tedy předpokládat, že mezi vnějšími polaři bude více nadprůměrných hráčů ve statistice RBI. I zde se při výpočtu v R neobejdeme bez některé z volitelných knihoven. V tomto případě si nainstalujeme knihovnu `moments` a zadáme následující příkaz:

```
R> skewness(z)
```

Jako poslední si rozebereme OBP, tedy procentuální úspěšnost dosažení alespoň první mety. Tato statistika je velmi úzce propojená s AVG a dává nám podobné výsledky. Můžeme tedy říci, že corners jsou opět nejlepší. Zde si porovnáme koeficient špičatosti. Tento koeficient nám udává koncentraci naměřených hodnot kolem středu ve srovnání s normálním rozdělením pravděpodobnosti. Z tabulek můžeme vyčíst, že rozdělení OBP u corners a vnějších polařů je špičatější než u vnitřních polařů, tedy tyto hodnoty budou více koncentrovány kolem jejich středu. Pro výpočet je stejně jako u koeficientu šikmosti potřebná knihovna `moments`. Kód pro výpočet bude vypadat následovně:

```
R> kurtosis(z)
```

2.2. Jádrové odhady hustoty

Ve druhé části vizualizace jednorozměrných dat si představíme jádrové odhady hustoty. Nejprve se seznámíme s potřebnou teorií a zavedeme si několik důležitých pojmů. Potom některé hustoty pomocí softwaru R vykreslíme a uvedeme i příslušné kódy. Na konci kapitoly zkusíme vykreslené odhady porovnat. Hlavním zdrojem v teoretické části byla kniha [2] a práce [12], v praktické pak knihovna [4].

Jak již název napovídá, jádrové odhady hustoty jsou pojmenovány podle funkcí, které nazýváme jádra. Nejprve si proto zadefinujeme pojem jádro.

Definice 2.2.1 *Jádrem nazveme libovolnou funkci*

$$\mathbf{K} : (\mathbb{R}, \mathcal{B}) \to [0, +\infty),$$

která je symetrická, ohraničená a pro niž platí

$$\int_{-\infty}^{+\infty} \mathbf{K}(x)dx = 1$$

a

$$\lim_{x \to \pm\infty} |x| \cdot \mathbf{K}(x) = 0,$$

kde $\mathcal{B}$ je σ-algebra borelovských množin na přímce.

Definice 2.2.2 *Nechť $\{h_n\}_{n=1}^{\infty}$ je posloupnost kladných čísel taková, že*

$$\lim_{n\to\infty} h_n = 0 \quad a \quad \lim_{n\to\infty} nh_n = \infty,$$

$\mathbf{K}(x)$ je jádro, $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s hustotou f(x). Jádrový odhad hustoty f(x) je definován vztahem:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbf{K}\left(\frac{x - X_i}{h_n}\right), \quad x \in \mathbb{R}.$$

Parametr h_n zde hraje roli měřítka. Nazývá se šířka okna nebo také vyhlazovací parametr. Třemi nejčastěji používanými jádry jsou obdélníkové,

$$\mathbf{K}(x) = \begin{cases} \frac{1}{2}, \, |x| < 1 \\ \\ 0, \, \text{jinak}, \end{cases}$$

trojúhelníkové,

$$\mathbf{K}(x) = \begin{cases} 1 - |x|, \, |x| < 1 \\ \\ 0, \qquad \text{jinak}, \end{cases}$$

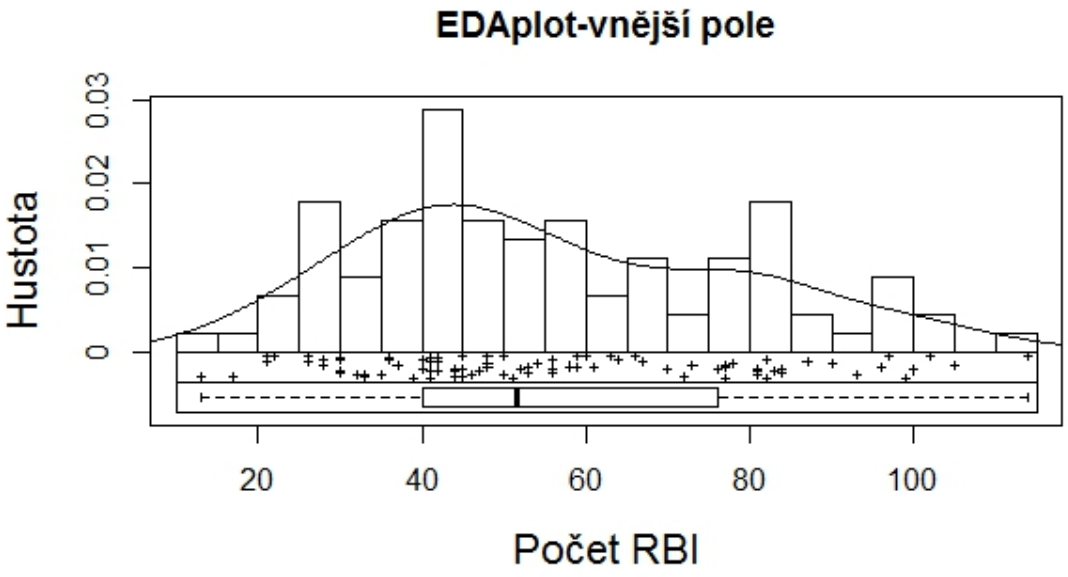
nebo Gaussovo

$$\mathbf{K}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R}.$$

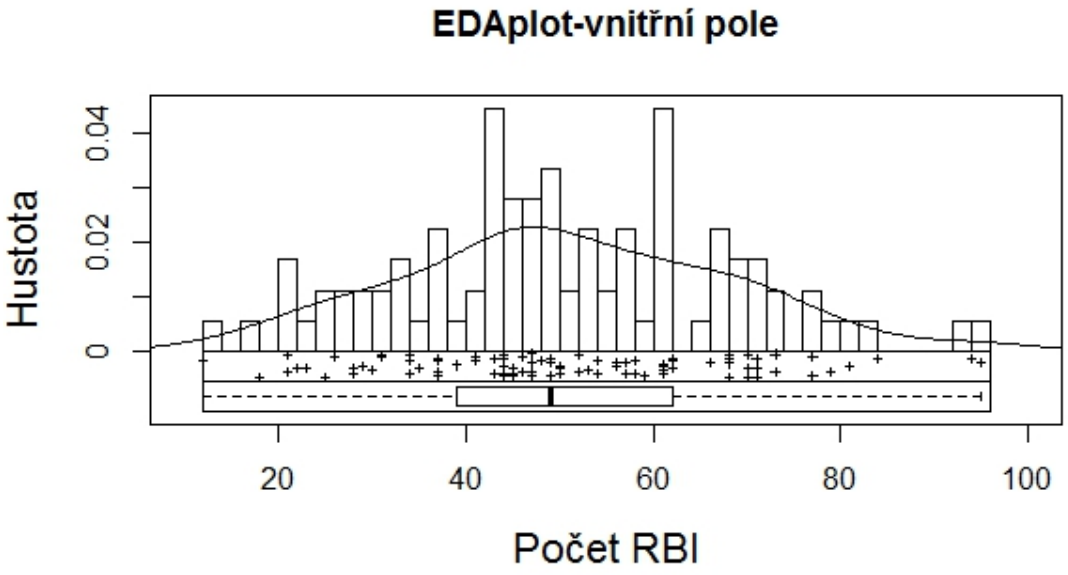
Nyní si pro každou skupinu hráčů vykreslíme jeden graf, tzv. EDAplot (z angl. exploratory data analysis plot), který v sobě obsahuje základní grafické nástroje pro vizualizaci jednorozměrných dat. EDAplot je tedy velice zajímavý grafický nástroj, jelikož v sobě zahrnuje hned čtyři typy grafů. Kromě našich jádrových odhadů hustoty dokáže EDAplot znázornit také boxplot neboli krabičkový graf (někdy označován též jako box and whisker plot), histogram a bodový graf. Podrobný popis těchto tří grafů můžeme najít například v [2] a [7]. Jako výchozí jádro je v EDAplotu nastaveno Gaussovo, se kterým budeme pracovat i my. Nyní si ukážeme, jak takový EDAplot sestrojit v R. Předpokládejme, že máme načtený soubor s daty a vytvořen známý vektor z, který jsme zkonstruovali již v první kapitole. K vytvoření tohoto grafu budeme potřebovat stáhnout knihovnu `StatDA`. Výsledný kód, který byl použit,

```
R> edaplot(z,H.freq=FALSE,box=TRUE,H.breaks=30,S.pch=3,S.cex=0.4,
+    P.xlim=c(11,114),D.lwd=1.5,P.log=FALSE,
+    P.main="EDAplot - vnější pole",P.xlab="Počet RBI",
+    P.ylab="Hustota",B.pch=3,B.cex=0.5),
```

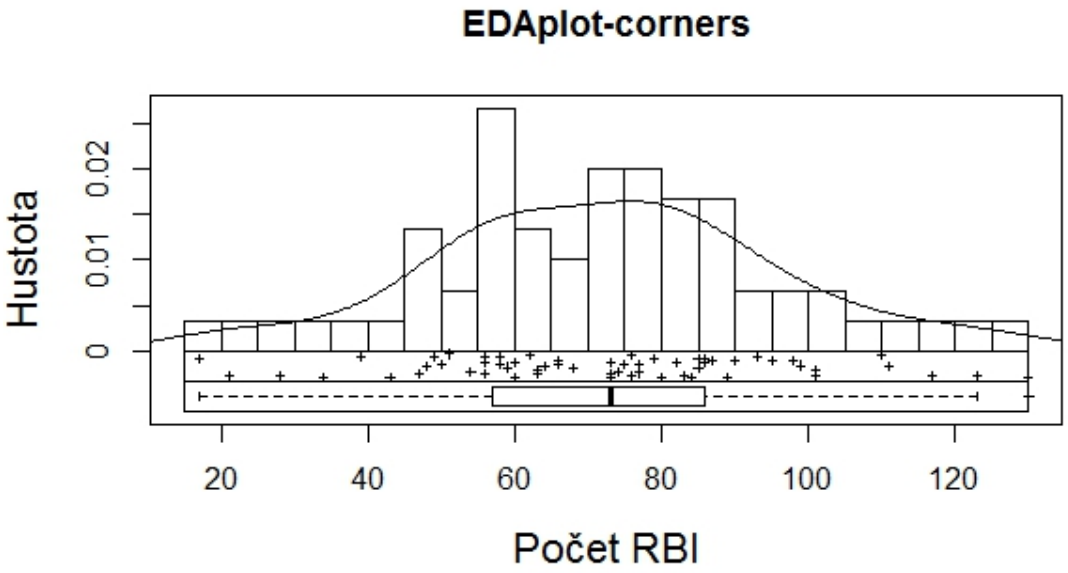
vykreslí EDAplot pro naše data:



Obrázek 2.1: EDAplot - vnější pole



Obrázek 2.2: EDAplot - vnitřní pole



Obrázek 2.3: EDAplot - corners

Zkusme provést jednoduché srovnání těchto tří grafů. Začneme s funkcemi hustot. Na grafu, který znázorňuje corners (obrázek 2.3), je zcela patrné, že křivka nabývá vrcholu zhruba v bodě 75. Zde lze vidět jasnou dominanci této skupiny hráčů nad ostatními, jelikož vrcholy dalších dvou skupin se nachází okolo bodů 40 až 45. Dále můžeme z grafů vyčíst, že všechny tři křivky začínají totožně cca v bodě 15, ale každá končí jinde. Konkrétně u vnějších polařů v daleko vzdálenějším bodě než u vnitřních polařů, což svědčí o jejich vyšších maximálních hodnotách RBI. Když se nyní podíváme na boxplot, který se nachází ve spodní části grafů, najdeme v obdélníku vyznačen medián příslušného statistického souboru. Snadno se lze přesvědčit, že hodnoty, které jsme vypočítali v tabulkách dříve, naprosto přesně odpovídají těm, které můžeme z našich boxplotů vyčíst. Můžeme tedy celkově říci, že EDAplot podpořil náš předpoklad, který vzešel z číselných charakteristik a ukázal, že corners jsou ve statistice RBI opravdu nejlepší.

Kapitola 3

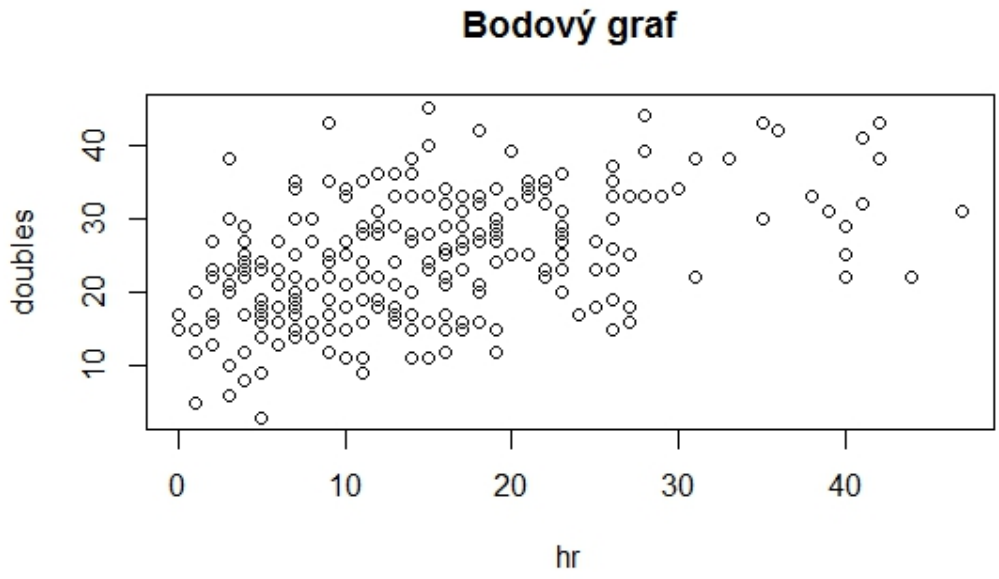
Vizualizace dvourozměrných dat

V této kapitole se budeme zabývat grafickým znázorněním dvourozměrných dat. Začneme se základním nástrojem pro vizualizaci, kterým je bodový graf. Dále si ukážeme jak zjistit zda na sobě dvě proměnné závisí či nikoliv. S tím nám pomůže tzv. chí-graf. V závěru si opět vykreslíme jádrový odhad hustoty, tentokrát však pro data mající dvě proměnné. Jako literatura zde byla použita kniha [2], článek [16] a knihovny [3], [18].

3.1. Bodový graf

Jednoduchý bodový graf (anglicky *scatter plot*), označován někdy též jako korelační diagram, byl používán už v 18. století a má mnoho předností. Nejčastěji se používá pro dvourozměrná data, ale jak si ukážeme v následující kapitole o mnohorozměrných datech, lze vytvořit i tzv. 3D bodový graf. Nyní se však budeme zabývat pouze standardní verzí. Bodový graf, jak již název napovídá, je tvořen množinou bodů. Na osu x se nanášejí hodnoty první proměnné a na osu y hodnoty té druhé. Z bodového grafu lze i odhadnout, zda na sobě jsou dvě proměnné závislé či nikoliv. My ho budeme používat pro porovnání našich tří skupin, tedy vnitřních polařů, vnějších polařů a corners. Nejprve si, ale jeden jednoduchý bodový graf vykreslíme. Jelikož jde opravdu o základní grafický nástroj, tak není potřeba stahovat žádnou novou knihovnu. Bodový graf vykreslíme následujícím příkazem:

```
R> plot(doubles ~ hr,data=Baseball,main="Bodový graf")
```



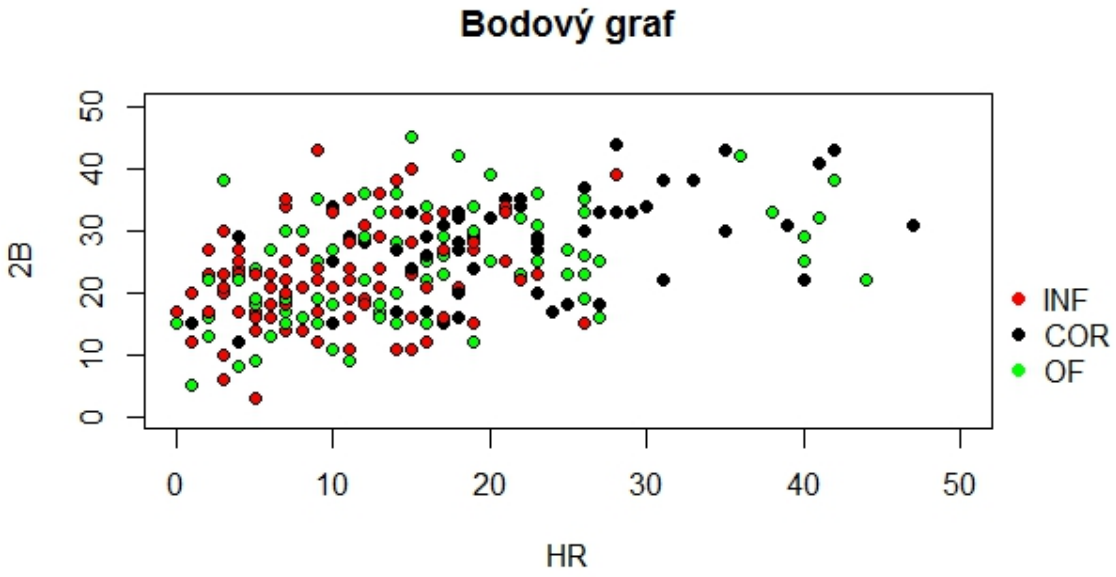
Obrázek 3.1: Bodový graf

Tento bodový graf znázorňuje vztah mezi proměnnými 2B a HR. V grafu jsou zahrnuty všechny tři skupiny, ale pro naši potřebu by bylo velmi vhodné tyto skupiny nějak odlišit. Můžeme například použít pro každou skupinu jiný typ symbolu. Tato varianta by však byla v našem případě velmi nevhodná, neboť výsledný graf by byl nepřehledný, a proto použijeme barevné odlišení. Nyní si ukážeme příslušný kód v R,

```
R> plot(Baseball$hr, Baseball$doubles, pch=21,
+    bg=c("black","red","green")[unclass(Baseball$pos)],
+    main="Bodový graf",xlab="HR",ylab="2B",
+    xlim=c(0,50),ylim=c(0,50))
```

a přidáme si legendu,

```
R> legend(par()$usr[2], mean(par()$usr[3:4]),
+    c("INF","COR","OF"), xpd=T, bty="n", pch=19,
+    col=c("Red","Black","Green"))
```



Obrázek 3.2: Bodový graf

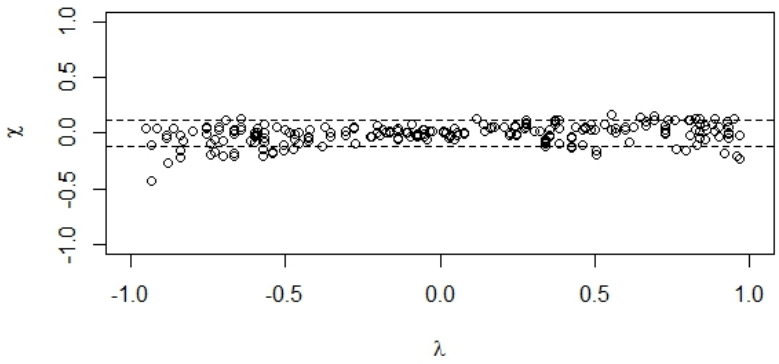
S tímto grafem se nám již bude pracovat mnohem lépe. Jak můžeme vidět, tak svislá osa nám udává počet 2B a vodorovná počet HR. Dle legendy můžeme zjistit, že červené body značí hráče, jenž reprezentují skupinu vnitřních polařů, černé corners a zelené vnější polaře (v legendě byly z úsporných důvodů použity anglické zkratky). Z grafu můžeme například vyčíst, že červené body jsou nejvíce koncentrovány v levé polovině grafu, což dokazuje, že vnitřní polaři odpalují nejméně homerunů. Spíše v levé polovině se také nacházejí body zelené barvy, které značí vnější polaře. Nejvíce bodů na pravé straně a zároveň nejméně na levé je černých, což dokazuje, že corners jsou ve statistice homerunů nejlepší. Podívejme se nyní na statistiku 2B. Zde nelze jednoznačně určit, jestli se výše nachází spíše body zelené nebo černé. Zdá se ale, že body znázorňující vnitřní polaře se nacházejí nejníže. Jak již bylo na začátku této podkapitoly řečeno, tak z bodového grafu můžeme vyčíst, zda jsou na sobě veličiny závislé či nikoliv. Můžeme říci, že v našem grafu lze vidět jakousi lehkou závislost, ale není natolik zřejmá, aby měla dostatečnou vypovídající hodnotu. V další podkapitole si proto ukážeme způsob, jakým lze závislost zkoumat daleko přesněji.

3.2. Chí-graf

I když je bodový graf hlavním nástrojem pro pozorování závislosti mezi dvojicí proměnných, tak je někdy velmi obtížné tuto závislost z grafu vyčíst. O tom jsme se ostatně mohli přesvědčit v předchozím grafu. Tento problém lze velmi pěkně vyřešit pomocí tzv. chí-grafu navrženého v [5]. Chí-graf je vlastně náš známý bodový graf, který tvoří dvojice hodnot (λ_i, χ_i). Hodnota $\lambda_i \in (-1, 1)$ je normovanou vzdáleností bodu (x_i, y_i) od dvourozměrného mediánu datového souboru. Kladné hodnoty λ_i znamenají, že hodnoty x_i a y_i jsou obě větší v porovnání s jejich příslušnými mediány, nebo obě menší. Naopak záporné hodnoty λ_i odpovídají tomu, že jsou hodnoty proměnných na opačných stranách jejich příslušných mediánů. Hodnoty $\chi_i \in (-1, 1)$ jsou odmocninou statistiky χ^2, získané z tabulky velikosti 2×2, kterou získáme rozdělením dat pomocí vztahů $x_{\cdot 1} \leq x_{i1}$ a $x_{\cdot 2} \leq x_{i2}$. Za předpokladu nezávislosti jsou tyto hodnoty asymptoticky normální s nulovou střední hodnotou.

Nyní si ukážeme jak chí-graf sestrojit v softwaru R. Nejprve si vykreslíme graf, který nám ukáže nezávislost, a poté graf dvou závislých proměnných. K tomu je potřeba nainstalovat si knihovnu `MVA`. Teď už můžeme psát velmi jednoduchý kód pro sestrojení chí-grafu:

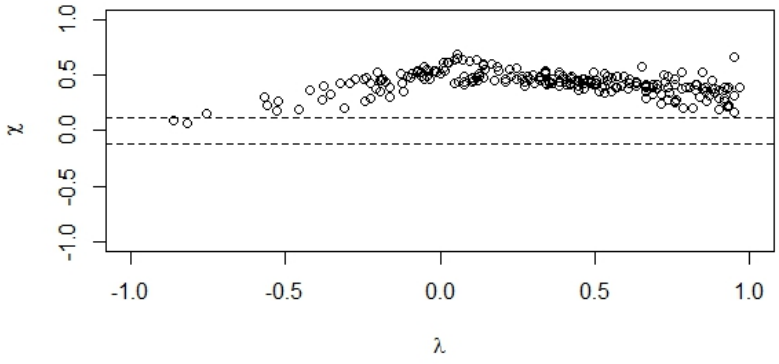
```
R> with(Baseball, chiplot(hr, avg))
```



Obrázek 3.3: Chí-graf pro AVG a HR

Když se podíváme na tento graf, můžeme zde vidět dvě horizontální rovnoběžky. V našem případě téměř všechny body leží mezi nimi a hodnota χ se pohybuje kolem 0. Jelikož hodnota χ nám udává závislost, tak můžeme říci, že proměnné HR a AVG jsou na sobě zcela nezávislé.

Nyní si vykreslíme druhý chí-graf. Tentokrát pro proměnné AVG a OBP. Tyto dvě statistiky jsou si velmi podobné, neboť OBP v sobě kromě met zdarma (BB) a trefení (HBP) zahrnuje již zmíněné AVG. Měli bychom tedy vidět graf, který se od toho předchozího bude lišit.



Obrázek 3.4: Chí-graf pro AVG a OBP

Jak vidíme, tak tento graf naplňuje náš předpoklad. Hodnoty proměnné χ se pohybují někde kolem hodnoty 0,5 a to znamená, že proměnné AVG a OBP jsou na sobě závislé.

3.3. Jádrové odhady hustoty pro dvourozměrná data

Jádrové odhady hustoty jsme si představili již v kapitole 2.2 a nyní zde ukážeme, jak vykreslit graf pro dvourozměrná data.

Definice 3.3.1 *Dvourozměrný jádrový odhad hustoty pro data* $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ *je definován jako*

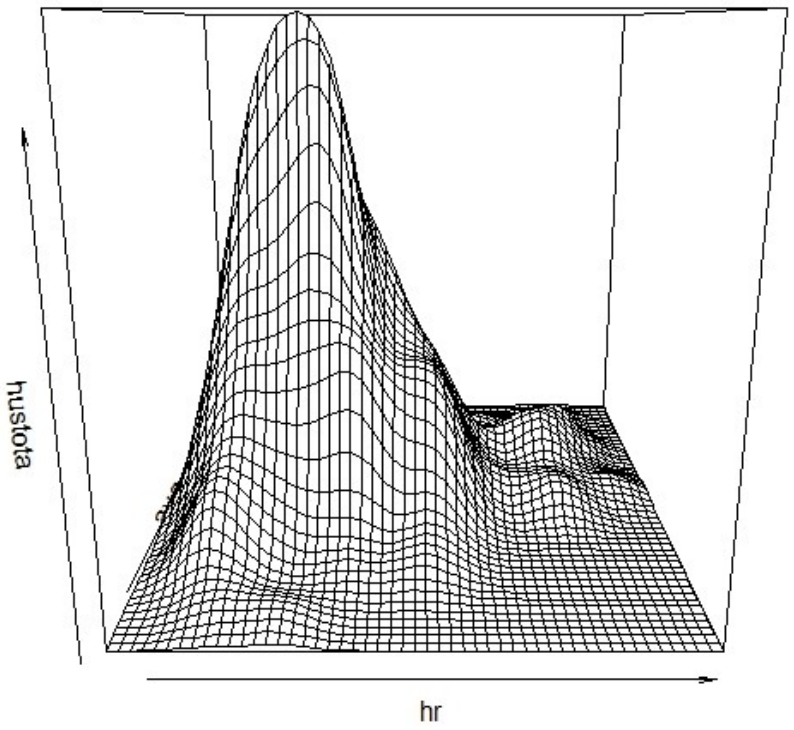
$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} \mathbf{K}\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right), \quad x, y \in \mathbb{R}.$$

Jako jádro opět použijeme Gaussovo, které bude nyní vypadat následovně:

$$\mathbf{K}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}, \quad x, y \in \mathbb{R}.$$

Tento graf bude trojrozměrný a použijeme proměnné HR a AVG. K vytvoření tohoto grafu budeme potřebovat nainstalovat knihovnu `KernSmooth`. Nyní už můžeme graf sestrojit následujícím kódem:

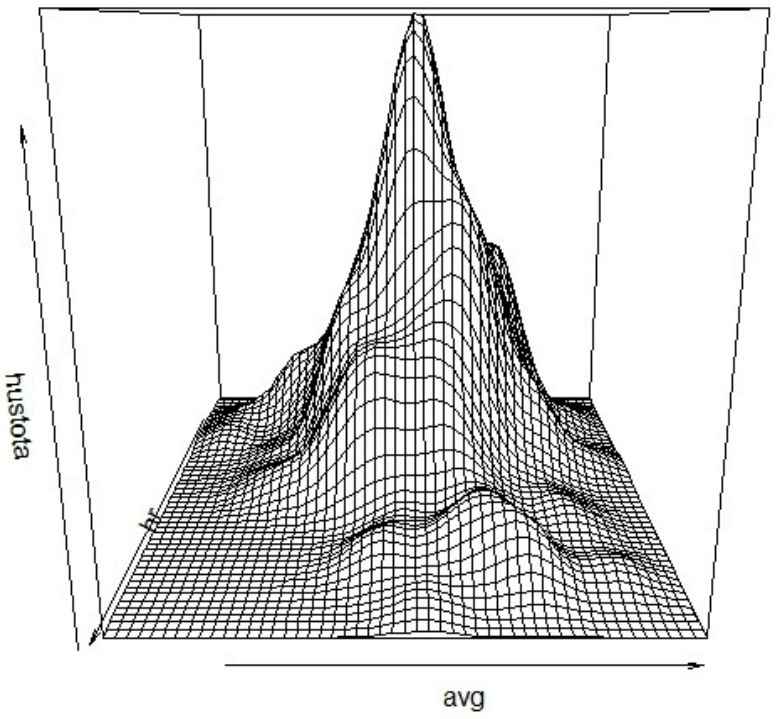
```
R> Baseball=Baseball[,c("avg","hr")]
R> a = bkde2D(Baseball, bandwidth = sapply(Baseball, dpik))
R> plot(Baseball, xlab = "avg",ylab = "hr")
R> contour(x = a$x1, y = a$x2,z = a$fhat, add = TRUE)
R> persp(x = a$x1, y = a$x2, z = a$fhat,
+    xlab = "avg",
+    ylab = "hr",
+    zlab = "hustota")
```



Obrázek 3.5: Trojrozměrný jádrový odhad hustoty pro AVG a HR

Úplně v levém dolním rohu můžeme vidět lehký náznak jakéhosi hrbolu. Ten reprezentuje hráče s opravdu nízkým AVG, kteří se však téměř přiblížili k průměrnému počtu homerunů. Největší hrbol, který se nachází na levé straně z pohledu osy HR a zhruba ve středu osy AVG, reprezentuje skoro všechny hráče. Velmi zajímavě vypadá skupina tří malých hrbolů v pravé horní části, které znázorňují ty nejlepší hráče ligy, jenž vynikají jak v HR, tak v AVG.

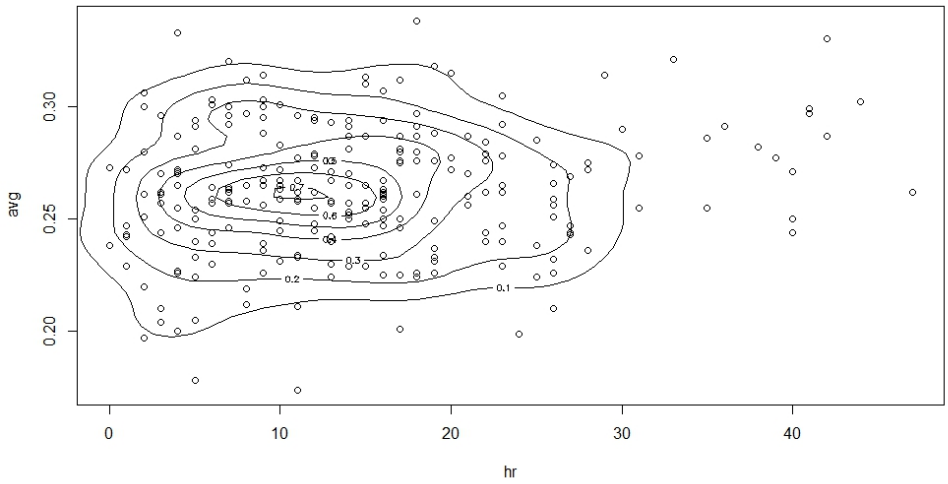
Pojďme se nyní podívat na tento graf z jiné perspektivy. Za tímto účelem pouze přidáme do původního kódu nový parametr `theta`, který otočí graf podle svislé osy. Zvolme například `theta=90`.



Obrázek 3.6: Trojrozměrný jádrový odhad hustoty pro AVG a HR otočený o 90°

Jelikož jsme zvolili `theta=90`, tak se původní graf otočil o 90° doleva. V novém grafu si lze ukázat několik nových skutečností. Například, že v předchozím grafu jsme skupinu tří hrbolků v pravé horní části označili za skupinu nejlepších hráčů ligy. Nyní si ale může všimnout, že v pravém dolním rohu se objevil nový hrbolek, který byl dříve zastíněn. Ten reprezentuje ty opravdu nejlepší hráče celé soutěže. Dále se zdá, že vpravo za hlavním a největším hrbolem, je schován jeden menší. Ten zobrazuje hráče s nadprůměrným AVG, kteří jsou však velmi podprůměrní v HR.

Je zřejmé že, díky nové perspektivě se můžeme o grafu dozvědět další užitečné informace. Proto se na tento graf podíváme ještě ze třetí perspektivy. S tím nám pomůže tzv. vrstevnicový graf, který ukazuje data z pohledu shora. Pro vykreslení tohoto grafu stačí použít první čtyři řádky z předchozího kódu.



Obrázek 3.7: Vrstevnicový graf pro AVG a HR otočený o 90°

Z obrázku vidíme, že vrstevnicový graf se skládá z obyčejného bodového grafu, do kterého jsou zasazeny vrstevnice, jež znázorňují odhadnuté hustoty pro dvourozměrná data. Jednotlivá čísla pak označují funkční hodnoty hustot. Vidíme, že většina dat se nachází na levé straně grafu, proto se zde objevily zmiňované vrstevnice. Naopak díky malému počtu hráčů s vysokým počtem HR, nevidíme na pravé straně žádné vrstevnice. Díky vrstevnicovému grafu je možné si všimnout, že nejvyšší funkční hodnota hustoty je 0,7. Tedy nejčastěji mají hráči okolo 11 HR a AVG kolem 0,26.

Kapitola 4

Vizualizace mnohorozměrných dat

V této poslední kapitole si ukážeme, jak zkonstruovat grafy pro mnohorozměrná data. Postupně se seznámíme se čtyřmi grafickými nástroji, kterými je možné tyto data vykreslit. Nejprve se budeme opět zabývat již dříve zmíněným bodovým grafem, tentokrát však pro tři proměnné. Následovat bude biplot a dále si ukážeme, jak znázornit více bodových grafů pomocí tzv. matice bodových grafů. Nakonec si představíme tzv. mřížový (trellis) graf. V této kapitole byla jako hlavní literatura opět použita kniha [2], bakalářská práce [9] a knihovny [11] a [15].

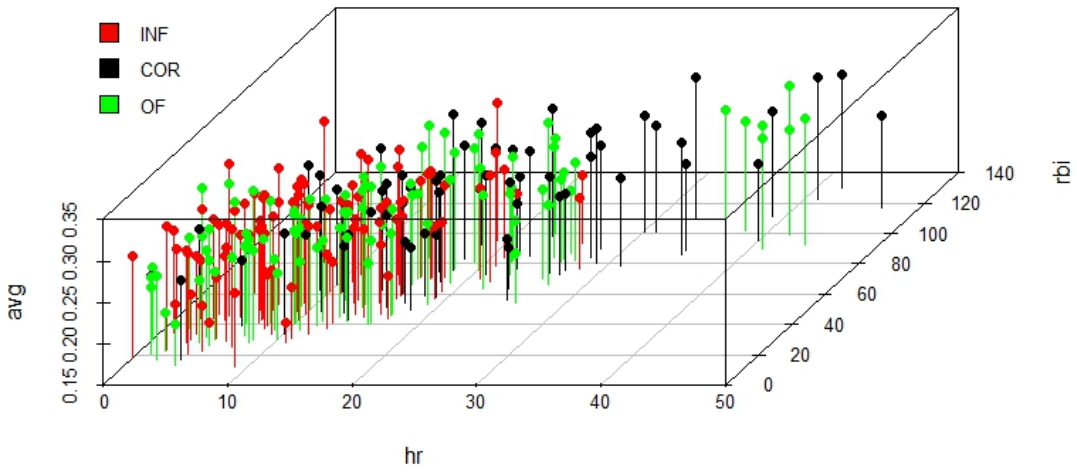
4.1. Trojrozměrný bodový graf

Tento velmi pěkný grafický nástroj je jakýmsi vylepšením obyčejného bodového grafu. Vylepšení spočívá v možnosti přidat ještě jednu proměnnou, což vede k dalšímu zkvalitnění vizualizace a interpretace datového souboru. Graf je opět tvořen množinou bodů, jejichž umístění navíc závisí na již zmíněné třetí proměnné. Nyní si sestrojíme trojrozměrný bodový graf pro proměnné HR, RBI a AVG. Hráče i zde rozdělíme do našich tří skupin, podobně jako tomu bylo v grafu 3.2. K vykreslení budeme potřebovat knihovnu `scatterplot3d`. Výsledný kód potom vypadá následovně:

```
R> with(Baseball,scatterplot3d(hr, rbi, avg,
+    color = c("Black","Red","Green")[pos],
+    type = "h",pch=19,angle = 66))
```

Přidáme ještě legendu:

```
R> legend("topleft", inset=.01, bty="n", cex=0.8,
+    c("INF", "COR", "OF"), fill=c("red", "black", "green"))
```



Obrázek 4.1: 3D bodový graf pro data HR, RBI a AVG

A takto vypadá výsledný graf s legendou (v legendě opět anglické zkratky), který si nyní popíšeme. Můžeme si všimnout, že velké množství červených bodů se nachází v levé spodní části, což vypovídá o tom, že vnitřní polaři jsou v útočné činnosti opravdu nejslabší, jak jsme již několikrát ukázali. Když se nyní podíváme na zelené body značící vnější polaře a body barvy černé, které reprezentují corners, tak se může zdát, že zde velký rozdíl není. Při bližším pohledu ale nacházíme jisté rozdíly. Zejména v pravé horní části je černých bodů téměř dvojnásobek. Naopak v levé dolní částí je znatelná převaha bodů zelených. To jen potvrzuje naše výsledky z předešlých kapitol. Jelikož rozdíly v AVG mezi skupinami nejsou nijak markantní, nelze z toho grafu o AVG moc vyčíst.

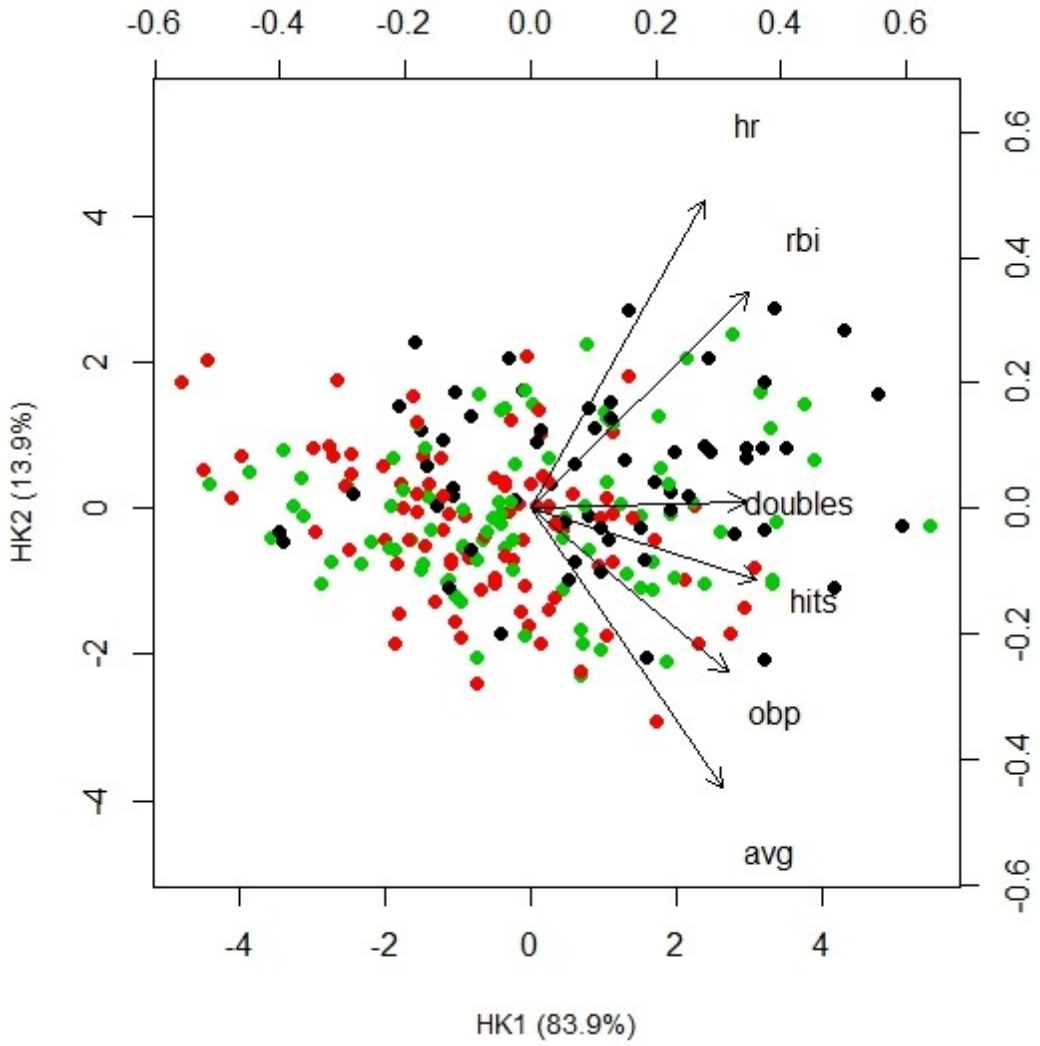
4.2. Biplot

Nyní si představíme velmi zajímavý a v současnosti hojně používaný grafický nástroj, který se nazývá biplot. Autorem této metody z roku 1971 je K. R. Gabriel. Biplot zobrazuje formou bodů a šipek skóry a zátěže prvních dvou hlavních komponent (HK), tedy slouží k redukci dimenze mnohorozměrných dat. Jelikož se hlavní komponenty snaží o vysvětlení co největší části celkové variability datového souboru (u druhé komponenty s podmínkou kolmosti na komponentu první), je součástí popisu os biplotu vždy též příslušný procentuální údaj. Biplot je tak v podstatě bodový graf, ve kterém jsou jednotliví hráči zobrazeni jako body. Proměnné jsou v biplotu zobrazeny ve formě šipek. Z důvodu rozsahu této bakalářské práce se dále teorií, která je popsána v řadě publikací, zabývat nebudeme, s jeho konstrukcí se seznámíme v kontextu zvoleného konkrétního datového souboru. Podrobnosti o biplotu lze najít například v [8] nebo [9]. Pojďme se podívat, jak biplot sestrojit v softwaru R. Jelikož náš statistický soubor obsahuje i sloupce jako například jméno hráče, pozici nebo počet zápasů, které v biplotu vykreslit samozřejmě nechceme, musíme nejdříve vybrat pouze proměnné, které je možné zobrazit. Aby měl biplot nějakou vypovídající hodnotu, bude také potřeba provést normování původních proměnných. Vytvoříme si tedy nový statistický soubor, který bude obsahovat pouze námi zvolená normovaná data. Soubor vytvoříme následovně:

```
R> Baseball3=scale(Baseball[,5:10])
```

Teď nám již nic nebrání vytvořit biplot. Použijeme k tomu vytvořenou funkci `biplot.color`, kterou lze nalézt v příloze B.

```
R> biplot.color(princomp(Baseball3)$scores[,1:2],
+    princomp(Baseball3)$loadings[,1:2],
+    col=as.numeric(Baseball[,2]),
+    pch=16,xlabMY="a",ylabMY="b")
```



Obrázek 4.2: Biplot

Takto vypadá výsledný biplot. Jak již bylo zmíněno, jednotlivé proměnné jsou vyznačeny šipkami, přičemž kosinus úhlu mezi dvěma šipkami aproximuje hodnotu příslušného korelačního koeficientu. Délky šipek jsou pak přibližně rovny rozptylům jednotlivých proměnných. Když se tedy podíváme například na dvojici proměnných HR a RBI, můžeme si všimnout, že úhel mezi příslušnými šipkami je celkem malý, tedy že hodnota jejich korelačního koeficientu se bude blížit k jedné. To znamená, že na sobě budou spíše lineárně závislé. Stejně tak na tom je i dvojice AVG a OBP, což jsme si dokázali už v kapitole 3.2 o chí-grafu. Podívejme se nyní na jednotlivé body. Černé znázorňují corners, červené vnitřní polaře a zelené

vnější polaře. Osu x, představující první hlavní komponentu, lze interpretovat jako úspěšnost hráčů v jednotlivých statistikách. Nejlepší hráči se nacházejí na pravé straně. Z grafu vidíme, že nejvíce bodů na pravé straně má barvu zelenou a černou. Tedy se nám opět potvrzuje, že corners a vnější polaři jsou pálkařsky lepší než vnitřní polaři, jejichž červené body se nachází více vlevo. S interpretací druhé hlavní komponenty nám výrazně pomůže směr šipek. Můžeme totiž říci, že hráči v horní polovině biplotu budou lepší v HR a RBI, zatímco ti ve spodní polovině v H, AVG a OBP.

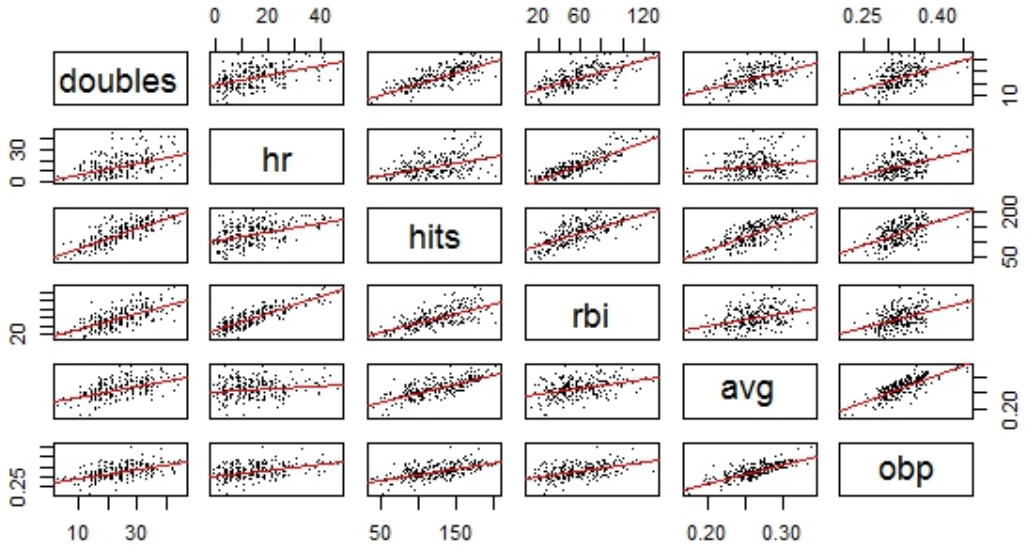
4.3. Matice bodových grafů

V této předposlední části si představíme nástroj, který umožňuje zobrazit bodové grafy pro více proměnných současně. Řekněme, že budeme chtít vykreslit bodové grafy pro kombinace proměnných 2B, HR, H, RBI, AVG a OBP. To umožňuje matice bodových grafů (anglicky *scatterplot matrix*). Můžeme si navíc každý graf proložit přímkou, která může leccos napovědět o vztahu mezi proměnnými. Jelikož je tento graf jeden ze základních, nebude potřeba stahovat žádnou další knihovnu. Nás samozřejmě nezajímá bodový graf z počtu odehraných zápasů a startů na pálce, vybereme proto pouze některá data. To provedeme vytvořením nového souboru, který pojmenujeme například Base.

```
R> Base=Baseball[,c(5:10)]
```

Teď už můžeme vykreslit matici grafů.

```
R> pairs(Base, panel = function (x, y, ...)
+    {points(x, y, ...) abline(lm(y ~ x), col = "red")},
+    pch = ".", cex = 1.5)
```



Obrázek 4.3: Matice bodových grafů

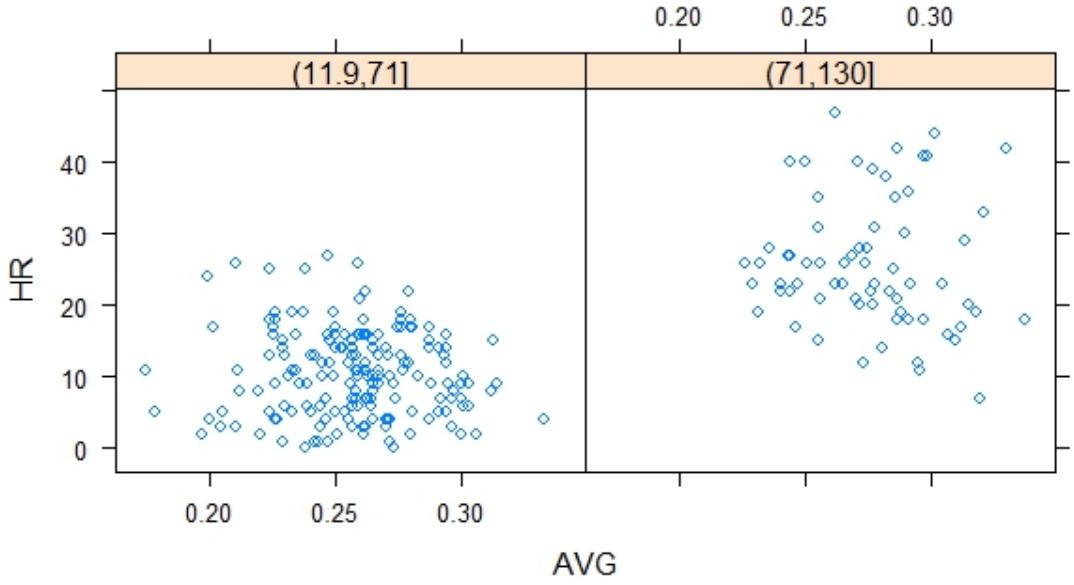
A tady máme výslednou matici. Jak vidíme, všechny grafy jsou přehledně uspořádány vedle sebe a lze je vzájemně porovnávat. Můžeme si třeba všimnout, které dvojice proměnných jsou na sobě závislé. Určitě to jsou proměnné AVG a OBP, což jsme si ukázali už v kapitole o chí-grafu. Dále se silně závislá zdá být dvojice HR a RBI, což dává smysl, jelikož nejvíce RBI dosáhnou hráči s největším počtem HR. Jako závislé se jeví také dvojice AVG a H nebo třeba 2B a H. Jako méně závislé můžeme jmenovat například HR a AVG, jelikož hráči s nejvyšším AVG jsou většinou kontaktní pálkaři, zatím co nejvíce HR dosahují pálkaři siloví.

4.4. Trellis grafika

Mřížový neboli trellis graf se opět podobá již několikrát zmiňovanému bodovému grafu. Trellis graf nám vznikne tak, že obyčejný bodový graf dvou proměnných rozdělíme na dva nebo více grafů podle třetí proměnné. V této práci se budeme zabývat variantou, kdy náš bodový graf proměnných HR a AVG rozdělíme na dvě části pomocí proměnné RBI. Pro vytvoření takovéhoto grafu

v programu R bude zapotřebí nainstalovat knihovnu `lattice`. Trellis graf potom vytvoříme následovně:

```
R> plot(xyplot(hr ~ avg| cut(rbi, 2), data = Baseball,
+    xlab="AVG",ylab="HR"))
```



Obrázek 4.4: Trellis graf

Takto vypadá výsledný graf, ve kterém nyní budeme pozorovat hráče jako celek. Můžeme si všimnout, že nám opravdu vyšly dva bodové grafy, které jsou rovnoměrně rozdělené pomocí hodnot proměnné RBI. V levé části máme bodový graf hráčů, kteří odpálili 12 až 71 RBI. Naopak na pravé straně můžeme vidět hráče, kteří dosáhli více než 71 RBI. Na první pohled je markantní, že největší počet hráčů se nachází na levé straně. Zdá se, že tito hráči dosahují menšího počtu HR než skupinka v pravé části grafu. To si můžeme vysvětlit velkou závislostí mezi veličinami HR a RBI. Skupinu, která se nachází v pravé horní části grafu, můžeme označit za nejlepší pálkaře, neboť dominují ve všech třech sledovaných statistikách.

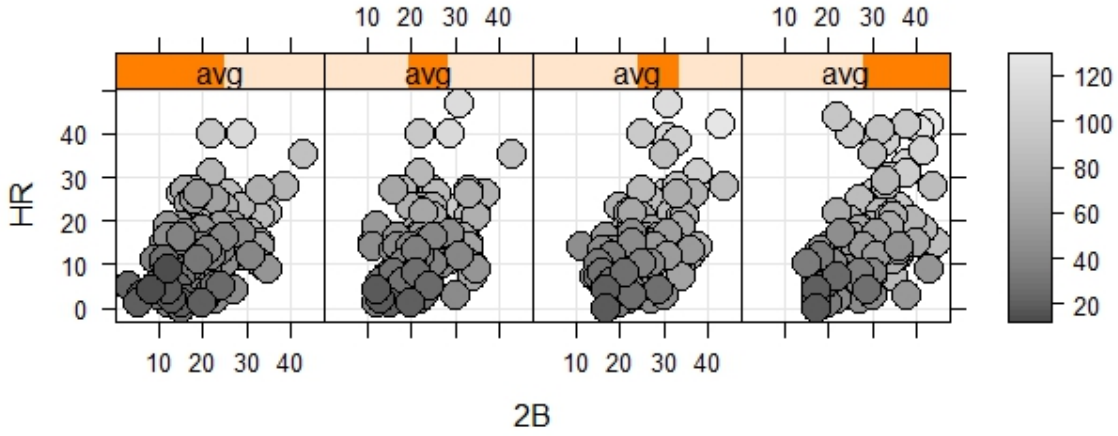
Na konci této práce si ještě představíme poslední graf. Bude jím opět trellis graf, ale nyní si ukážeme, co vše do něj lze přidat. Použijeme čtyři proměnné

a to AVG, HR, 2B a RBI. Jelikož jde pouze o ukázku, co vše se dá s trellis grafikou vytvořit, nebudeme se zde zabývat interpretací a jeho funkce si popíšeme na vykresleném grafu. Jako knihovnu opět použijeme `lattice`. Nejdříve bude potřeba v softwaru R nadefinovat několik nezbytných objektů.

```
R> rbi.col = gray.colors(100)[cut(Baseball$rbi,
+    100, label = FALSE)]
R> rbi.ord = rev(order(Baseball$rbi))
R> Baseball$color = rbi.col
R> Baseball.ordered = Baseball[rbi.ord, ]
```

Nyní již můžeme sestrojit výsledný kód:

```
xyplot(hr ~ doubles | avg, data = Baseball.ordered,
+    aspect = "iso", groups = color, cex = 2, col = "black",
+    panel = function(x, y, groups, ..., subscripts) {
+    fill = groups[subscripts]
+    panel.grid(h = -1, v = -1)
+    panel.xyplot(x, y, pch = 21, fill = fill, ...)},
+    legend = list(right = list(fun = draw.colorkey,
+    args = list(key = list(col = gray.colors,
+    at = rbi.breaks), draw = FALSE))),
+    xlab = "2B", ylab = "HR")
```



Obrázek 4.5: Trellis graf

Zde vidíme výsledný graf, který se tentokrát skládá ze čtyř oken. Vodorovná osa udává počet 2B a svislá počet HR. Navíc v horní části každého okna vidíme, jak jsou hráči rozdělení podle AVG. Dále si lze všimnout, že každý bod má jiný odstín šedé barvy. To značí počet RBI, kterého každý hráč dosáhl. S orientací v těchto odstínech nám pomáhá přiložená legenda. Na závěr bych rád zmínil, že ne vždy je vhodné použití takto složitých grafických nástrojů. Ve většině případů se nám naopak budou jednodušší grafy interpretovat lépe.

Závěr

V úvodu jsem uvedl, že za cíl této bakalářské práce si kladu zjistit, která ze tří skupin baseballových hráčů je nejlepší v útočné činnosti. Tento úkol nebyl vůbec jednoduchý a nejvíce času na této práci jsem strávil sběrem a analýzou dat, ale myslím si, že tohoto cíle se podařilo dosáhnout. V průběhu práce jsem několikrát dokázal, že nejlepší skupinou jsou po řadě corners, vnější polaři a vnitřní polaři. Musím říci, že ze zkušeností, které jsem za 15 let věnování se baseballu získal, jsem tento výsledek očekával. Někdy mě však překvapily některé malé nebo naopak velké rozdíly mezi zmíněnými skupinami.

Celou tvorbou této práce mě provázel statistický software R a musím přiznat, že začátky pro mě byly velmi obtížné. Nakonec jsem ale rád za získané zkušenosti a nabyté znalosti, které jsem u práce získal. Všem, kteří plánují s tímto softwarem pracovat, jej mohu jen doporučit.

Příloha A

Statistiky hráčů nastupujících v základní části MLB sezóny 2015.

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Machado, M	COR	162	633	30	35	181	86	0,286	0,359
Seager, K	COR	161	623	37	26	166	74	0,266	0,328
Andrus, E	INF	160	596	34	7	154	62	0,258	0,309
Davis, C	COR	160	573	31	47	150	117	0,262	0,361
Longoria, E	COR	160	604	35	21	163	73	0,27	0,328
Rizzo, A	COR	160	586	38	31	163	101	0,278	0,387
Calhoun, K	OF	159	630	23	26	161	83	0,256	0,308
Cespedes, Y	OF	159	633	42	36	184	105	0,291	0,328
Goldschmidt, P	COR	159	567	38	33	182	110	0,321	0,435
Pillar, K	OF	159	586	31	12	163	56	0,278	0,314
Trout, M	OF	159	575	32	41	172	90	0,299	0,402
Cabrera, M	OF	158	629	36	12	172	77	0,273	0,314
Donaldson, J	COR	158	620	41	41	184	123	0,297	0,371
Fielder, P	COR	158	613	28	23	187	98	0,305	0,378
Hosmer, E	COR	158	599	33	18	178	93	0,297	0,363
Martinez, J	OF	158	596	33	38	168	102	0,282	0,344
Mauer, J	COR	158	592	34	10	157	66	0,265	0,338
Votto, J	COR	158	545	33	29	171	80	0,314	0,459
Arenado, N	COR	157	616	43	42	177	130	0,287	0,323
Blackmon, C	OF	157	614	31	17	176	58	0,287	0,347
Bruce, J	OF	157	580	35	26	131	87	0,226	0,294
Dozier, B	INF	157	628	39	28	148	77	0,236	0,307
Frazier, T	COR	157	619	43	35	158	89	0,255	0,309
Granderson, C	OF	157	580	33	26	150	70	0,259	0,364
McCutchen, A	OF	157	566	36	23	165	96	0,292	0,401
Pollock, A	OF	157	609	39	20	192	76	0,315	0,367
Pujols, A	COR	157	602	22	40	147	95	0,244	0,307
Aybar, E	INF	156	597	30	3	161	44	0,27	0,301
Bogaerts, X	INF	156	613	35	7	196	81	0,32	0,355
Cano, R	INF	156	624	34	21	179	79	0,287	0,334
Desmond, I	INF	156	583	27	19	136	62	0,233	0,29
Fowler, D	OF	156	596	29	17	149	46	0,25	0,346
Gonzales, A	COR	156	571	33	28	157	90	0,275	0,35
Headley, C	COR	156	580	29	11	150	62	0,259	0,324
Markakis, N	OF	156	612	38	3	181	53	0,296	0,37

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Gregorius, D	INF	155	525	24	9	139	56	0,265	0,318
Parra, G	OF	155	547	36	14	159	51	0,291	0,328
Peralta, J	INF	155	579	26	17	159	71	0,275	0,334
Semien, M	INF	155	556	23	15	143	45	0,257	0,31
Abreu, J	COR	154	613	34	30	178	101	0,29	0,347
Altuve, J	INF	154	638	40	15	200	66	0,313	0,353
Carpenter, M	COR	154	574	44	28	156	84	0,272	0,365
Castellanos, N	COR	154	549	33	15	140	73	0,255	0,303
Heyward, J	OF	154	547	33	13	160	60	0,293	0,359
Kemp, M	OF	154	596	31	23	158	100	0,265	0,312
Kinsler, I	INF	154	624	35	11	185	73	0,296	0,342
Ramirez, A	INF	154	583	33	10	145	62	0,249	0,285
Santana, C	COR	154	550	29	19	127	85	0,231	0,357
Bautista, J	OF	153	543	29	40	136	114	0,25	0,377
Eaton, A	OF	153	610	28	14	175	56	0,287	0,361
Forsythe, L	INF	153	540	33	17	152	68	0,281	0,359
Gonzales, C	OF	153	554	25	40	150	97	0,271	0,325
Harper, B	OF	153	521	38	42	172	99	0,33	0,46
Marte, S	OF	153	580	30	19	167	81	0,288	0,337
Polanco, G	OF	153	593	35	9	152	52	0,256	0,32
Suzuki, I	OF	153	398	5	1	91	21	0,229	0,282
Cruz, N	OF	152	590	22	44	178	93	0,302	0,369
Peterson, J	INF	152	528	23	6	126	52	0,239	0,314
Plouffe, T	COR	152	573	35	22	140	86	0,244	0,307
Revere, B	OF	152	592	22	2	181	45	0,306	0,342
Solarte, Y	COR	152	526	33	14	142	63	0,27	0,32
Bryant, K	COR	151	559	30	26	153	99	0,274	0,368
Castro, S	INF	151	547	24	11	146	68	0,267	0,298
Galvis, F	INF	151	559	14	7	147	50	0,263	0,302
Gardner, B	OF	151	571	26	16	148	66	0,259	0,343
Kiermaier, K	OF	151	505	25	10	133	40	0,263	0,298
Pederson, J	OF	151	480	19	26	101	54	0,21	0,346
Walker, N	INF	151	543	32	16	145	71	0,267	0,327
Alvarez, P	COR	150	437	18	27	106	76	0,243	0,32
LeMahieu, D	INF	150	564	21	6	170	61	0,301	0,358
Posey, B	INF	150	557	28	19	177	95	0,318	0,379
Upton, J	OF	150	542	26	26	136	81	0,251	0,336
Wong, K	INF	150	557	28	11	146	61	0,262	0,321
Choo, S	OF	149	555	32	22	153	82	0,276	0,375

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Duffy, M	COR	149	573	28	12	169	77	0,295	0,334
Lawrie, B	COR	149	562	29	16	146	60	0,26	0,299
Lind, A	COR	149	502	32	20	139	87	0,277	0,36
Peralta, D	OF	149	462	26	17	144	78	0,312	0,371
Reddick, J	OF	149	526	25	20	143	77	0,272	0,333
Coghlan, C	OF	148	440	25	16	110	41	0,25	0,341
Escobar, A	INF	148	612	20	3	157	47	0,257	0,293
Garcia, A	OF	148	553	17	13	142	59	0,257	0,309
Phillips, B	INF	148	588	19	12	173	70	0,294	0,328
Herrera, O	OF	147	495	30	8	147	41	0,297	0,344
Moustakes, M	COR	147	549	34	22	156	82	0,284	0,348
Norris, D	INF	147	515	33	14	129	62	0,25	0,305
Owings, C	INF	147	515	27	4	117	43	0,227	0,264
Simmons, A	INF	147	535	23	4	142	44	0,265	0,321
Encarnacio, E	COR	146	528	31	39	146	111	0,277	0,372
Morrison, L	COR	146	457	15	17	103	54	0,225	0,302
Betts, M	OF	145	597	42	18	174	77	0,291	0,341
Gordon, D	INF	145	615	24	4	205	46	0,333	0,359
Moss, B	COR	145	469	24	19	106	58	0,226	0,304
Miller, B	INF	144	438	22	11	113	46	0,258	0,329
Rollins, J	INF	144	517	24	13	116	41	0,224	0,285
Beltre, A	COR	143	567	32	18	163	83	0,287	0,334
Cabrera, A	INF	143	505	28	15	134	58	0,265	0,315
Crawford, B	INF	143	507	33	21	130	84	0,256	0,321
Lagares, J	OF	143	441	16	6	114	41	0,259	0,289
Ethier, A	OF	142	395	20	14	116	53	0,294	0,366
Perez, S	INF	142	531	25	21	138	70	0,26	0,28
Russel, A	INF	142	475	29	13	115	54	0,242	0,307
Segura, J	INF	142	560	16	6	144	50	0,257	0,281
Trumbo, M	OF	142	508	23	22	133	64	0,262	0,31
Bourn, M	OF	141	425	15	0	101	30	0,238	0,31
Kipnis, J	INF	141	565	43	9	171	52	0,303	0,372
Maybin, C	OF	141	505	18	10	135	59	0,267	0,327
Braun, R	OF	140	506	27	25	144	84	0,285	0,356
Cain, L	OF	140	551	34	16	169	72	0,307	0,361
Gose, A	OF	140	485	24	5	123	26	0,254	0,321
Reynolds, M	COR	140	382	21	13	88	48	0,23	0,315
Young, C	OF	140	318	20	14	80	42	0,252	0,32
Escobar, Y	COR	139	535	25	9	168	56	0,314	0,375

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Hunter, T	OF	139	521	22	22	125	81	0,24	0,293
Rodriguez, S	COR	139	224	12	4	55	17	0,246	0,281
Taylor, M	OF	138	472	15	14	108	63	0,229	0,282
Belt, B	COR	137	492	33	18	138	68	0,28	0,356
Brantley, M	OF	137	529	45	15	164	84	0,31	0,379
Flores, W	INF	137	483	22	16	127	59	0,263	0,295
Jones, A	OF	137	546	25	27	147	82	0,269	0,308
Ramirez, Ar	COR	137	475	31	17	117	75	0,246	0,297
Rasmus, C	OF	137	432	23	25	103	61	0,238	0,314
Jackson, A	OF	136	491	25	9	131	48	0,267	0,311
Molina, Y	INF	136	488	23	4	132	61	0,27	0,31
Smith, S	OF	136	395	31	12	98	42	0,248	0,33
Vogt, S	INF	136	445	21	18	116	71	0,261	0,341
Byrd, M	OF	135	506	25	23	125	73	0,247	0,29
Duda, L	COR	135	471	33	27	115	73	0,244	0,352
McCann, B	INF	135	465	15	26	108	94	0,232	0,32
Venable, W	OF	135	349	13	6	85	33	0,244	0,32
Ahmed, N	INF	134	421	17	9	95	34	0,226	0,275
Beltran, C	OF	133	478	34	19	132	67	0,276	0,337
Marisnick, J	OF	133	339	15	9	80	36	0,236	0,281
Napoli, M	COR	133	407	20	18	91	50	0,224	0,324
Pagan, A	OF	133	512	21	3	134	37	0,262	0,303
Inciarte, E	OF	132	524	27	6	159	45	0,303	0,338
Moreland, M	COR	132	471	27	23	131	85	0,278	0,33
Murphy, D	OF	132	361	18	10	102	50	0,283	0,318
Smoak, J	COR	132	296	16	18	67	59	0,226	0,299
Valbuena, L	COR	132	434	18	25	97	56	0,224	0,31
Drew, S	INF	131	383	16	17	77	44	0,201	0,271
Schumaker, S	OF	131	244	20	1	59	21	0,242	0,306
Suzuki, K	INF	131	433	17	5	104	50	0,24	0,296
Cervelli, F	INF	130	450	17	7	133	44	0,296	0,37
Hechavarria, A	INF	130	470	17	5	132	48	0,281	0,315
Murphy, Dan	INF	130	499	38	14	140	73	0,281	0,322
Asche, C	OF	129	425	22	12	104	39	0,245	0,294
Bour, J	COR	129	409	20	23	107	73	0,262	0,321
Carter, C	COR	129	391	17	24	78	64	0,199	0,307
Giavotella, J	INF	129	453	25	4	123	49	0,272	0,318
Holt, B	INF	129	454	27	2	127	45	0,28	0,349
Howard, R	COR	129	467	29	23	107	77	0,229	0,277

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Martin, R	INF	129	441	23	23	106	77	0,24	0,329
Prado, M	COR	129	500	22	9	144	63	0,288	0,338
Goins, R	INF	128	376	16	5	94	45	0,25	0,318
Guyer, B	OF	128	332	21	8	88	28	0,265	0,359
Gyorko, J	INF	128	421	15	16	104	57	0,247	0,297
Ramos, W	INF	128	475	16	15	109	68	0,229	0,258
Tulowitzki, T	INF	128	486	27	17	136	70	0,28	0,337
Escobar, E	INF	127	409	31	12	107	58	0,262	0,309
Hernandez, C	INF	127	405	20	1	110	35	0,272	0,339
Kang, J	COR	126	421	24	15	121	58	0,287	0,355
Realmuto, J	INF	126	441	21	10	114	47	0,259	0,29
Robinson, C	COR	126	309	15	10	84	34	0,272	0,358
Sandoval, P	COR	126	470	25	10	115	47	0,245	0,292
Turner, J	COR	126	385	26	16	113	60	0,294	0,37
Yelich, C	OF	126	476	30	7	143	44	0,3	0,366
Zobrist, C	INF	126	467	36	13	129	56	0,276	0,359
Burns, B	OF	125	520	18	5	153	42	0,294	0,334
Canha, M	OF	124	441	22	16	112	70	0,254	0,315
Infante, O	INF	124	440	23	2	97	44	0,22	0,234
Ozuna, M	OF	123	459	27	10	119	44	0,259	0,308
Rosario, E	OF	122	453	18	13	121	50	0,267	0,289
Davis, K	OF	121	392	16	27	97	66	0,247	0,323
DeSchields, D	OF	121	425	22	2	111	37	0,261	0,344
Freese, D	COR	121	424	27	14	109	56	0,257	0,323
Fuld, S	OF	120	290	16	2	57	22	0,197	0,276
Gonzales, M	INF	120	344	18	12	96	34	0,279	0,317
Iglesias, J	INF	120	416	17	2	125	23	0,3	0,347
Ordor, R	INF	120	426	21	16	111	61	0,261	0,316
Sanchez, C	INF	120	389	23	5	87	31	0,224	0,268
Sogard, E	INF	120	372	12	1	92	37	0,247	0,294
Cabrera, M	COR	119	429	28	18	145	76	0,338	0,44
Uribe, J	COR	119	360	17	14	91	43	0,253	0,32
Amarista, A	INF	118	324	10	3	66	30	0,204	0,257
Espinosa, D	INF	118	367	21	13	88	37	0,24	0,311
Francoeur, J	OF	118	326	16	13	84	45	0,258	0,286
Freeman, F	COR	118	416	27	18	115	66	0,276	0,37
Tomas, Y	OF	118	406	19	9	111	48	0,273	0,305
Bourjos, P	OF	117	195	8	4	39	13	0,2	0,29
Cuddyer, M	OF	117	379	18	10	98	41	0,259	0,309

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Kendrick, H	INF	117	464	22	9	137	54	0,295	0,336
Hill, A	COR	116	313	18	6	72	39	0,23	0,295
Mercer, J	INF	116	394	21	3	96	34	0,244	0,293
Paulsen, B	COR	116	325	19	11	90	49	0,277	0,326
Reyes, J	INF	116	481	25	7	132	53	0,274	0,31
Tejada, R	INF	116	360	23	3	94	28	0,261	0,338
Blanco, G	OF	115	327	19	5	95	26	0,291	0,368
Gomez, C	OF	115	435	29	12	111	56	0,255	0,314
Grandal, Y	INF	115	355	12	16	83	47	0,234	0,353
De Aza, A	OF	114	325	17	7	85	35	0,262	0,333
Gennett, S	INF	114	375	18	6	99	29	0,264	0,294
Hamilton, B	OF	114	412	8	4	93	28	0,226	0,274
Hardy, J	INF	114	411	14	8	90	37	0,219	0,253
Harrison, J	COR	114	418	29	4	120	28	0,287	0,327
McCann, J	INF	114	401	18	7	106	41	0,264	0,297
Cron, C	COR	113	378	17	16	99	51	0,262	0,3
Montero, M	INF	113	347	11	15	86	54	0,248	0,345
Pierzinsky, A	INF	113	407	24	9	122	49	0,3	0,339
Davis, R	OF	112	341	16	8	88	30	0,258	0,306
DeJesus, D	OF	112	288	9	5	67	30	0,233	0,297
Flowers, T	INF	112	331	12	9	79	12	0,239	0,295
Perez, H	COR	112	263	15	1	64	21	0,243	0,257
Zunino, M	INF	112	350	11	11	61	28	0,174	0,23
Ellsbury, J	OF	111	452	15	7	116	33	0,257	0,318
Johnson, K	INF	111	310	11	14	82	47	0,265	0,314
Teixeira, M	COR	111	392	22	31	100	79	0,255	0,357
Castillo, W	INF	110	342	15	19	81	57	0,237	0,296
Rivera, R	INF	110	298	14	5	53	26	0,178	0,213
Souza Jr., S	OF	110	373	15	16	84	40	0,225	0,318
Ackley, D	OF	108	238	11	10	55	30	0,231	0,284
Pena, B	INF	108	333	17	0	91	18	0,273	0,334
Spangenberg, C	INF	108	303	17	4	82	21	0,271	0,333
Utley, C	INF	107	373	21	8	79	39	0,212	0,286
Barnes, B	OF	106	255	13	2	64	17	0,251	0,314
Blanco, A	INF	106	233	22	7	68	25	0,292	0,36
Chisenhall, L	OF	106	333	19	7	82	44	0,246	0,294
Guerrero, A	OF	106	219	9	11	51	36	0,233	0,261
Pennington, C	INF	105	210	6	3	44	21	0,21	0,298

Player	Pos	G	AB	2B	HR	H	RBI	AVG	OBP
Ramirez, H	OF	105	401	12	19	100	53	0,249	0,291
Rios, A	OF	105	385	22	4	98	32	0,255	0,287
Castro, J	INF	104	337	19	11	71	31	0,211	0,283
Gordon, A	OF	104	354	18	13	96	48	0,271	0,377
Grichuk, R	OF	103	323	23	17	89	47	0,276	0,329
Hundley, N	INF	103	366	21	10	110	43	0,301	0,339
Lucroy, J	INF	103	371	20	7	98	43	0,264	0,326
Descalso, D	INF	101	185	3	5	38	22	0,205	0,283
Joseph, C	INF	100	320	16	11	75	49	0,234	0,299
Panik, J	INF	100	382	27	8	119	37	0,312	0,378
Correa, C	INF	99	387	22	22	108	68	0,279	0,345

Příloha B

Funkce pro vytvoření biplotu v softwaru R.

```
"biplot.color" = function (x, y, var.axes = TRUE, col,
cex = rep(par("cex"), 2), xlabs = NULL, ylabs = NULL,
expand = 1, xlim = NULL, ylim = NULL, arrow.len = 0.1,
pch=NULL, pch.cex=1, xlabMY, ylabMY, ...)
{
n <- nrow(x)
p <- nrow(y)
if (is.null(xlabs)) {
xlabs <- dimnames(x)[[1]]
if (is.null(xlabs))
xlabs <- 1:n
}
xlabs <- as.character(xlabs)
dimnames(x) <- list(xlabs, dimnames(x)[[2]])
if (missing(ylabs)) {
ylabs <- dimnames(y)[[1]]
if (is.null(ylabs))
ylabs <- paste("Var", 1:p)
}
ylabs <- as.character(ylabs)
dimnames(y) <- list(ylabs, dimnames(y)[[2]])
if (length(cex) == 1)
cex <- c(cex, cex)
if (missing(col)) {
col <- par("col")
if (!is.numeric(col))
col <- match(col, palette())
col <- c(col, col + 1)
```

```
}
else if (length(col) == 1)
col <- c(col, col)
unsigned.range <- function(x) c(-abs(min(x)), abs(max(x)))
rangx1 <- unsigned.range(x[, 1])
rangx2 <- unsigned.range(x[, 2])
rangy1 <- unsigned.range(y[, 1])
rangy2 <- unsigned.range(y[, 2])
if (missing(xlim) && missing(ylim))
xlim <- ylim <- rangx1 <- rangx2 <- range(rangx1, rangx2)
else if (missing(xlim))
xlim <- rangx1
else ylim <- rangx2
ratio <- max(rangy1/rangx1, rangy2/rangx2)/expand
on.exit(par(oldpar))
oldpar <- par(pty = "s")
#####################
#nadefinovaný graf
plot(x, type = "p", xlim = xlim, ylim = ylim, col = col,
pch=pch, cex=pch.cex, xlab = xlabMY, ylab = ylabMY, ...)
#####################
#žádný text u objektů
par(new = TRUE)
plot(y, axes = FALSE, type = "n", xlim = xlim * ratio,
ylim = ylim * ratio, xlab = "", ylab = "", col = col, ...)
axis(3, col = 1)
axis(4, col = 1)
box(col = 1)
#####################
#text proměnných
```

```
text(y, labels = ylabs, cex = cex[2], col = 1, ...)
if (var.axes)
####################
#barva šipek znázorňujících proměnné
arrows(0, 0, y[, 1] * 0.8, y[, 2] * 0.8, col = 1,
length = arrow.len) invisible()
}
```

Literatura

[1] *Baseball stats* [online]. [cit. 2016-04-01]. Dostupné z: mlb.mlb.com/stats/sortable.jsp.

[2] Everitt, B., Hothorn, T.: *An Introduction to Applied Multivariate Analysis with R,* Springer, New York, 2011.

[3] Everitt, B., Hothorn, T.: *Package 'MVA'* [online]. 2015, [cit. 2016-10-26]. dostupné z: https://cran.r-project.org/web/packages/MVA/MVA.pdf.

[4] Filzmoser, P.: *Package 'StatDA'* [online]. 2015, [cit. 2016-10-22]. dostupné z: https://cran.r-project.org/web/packages/StatDA/StatDA.pdf.

[5] Fisher, N.I., Switzer, P.: *Chi-plots for assessing dependence*. Biometrika 72 (1985), s. 253-265.

[6] Háněl, T.: *Rozměry baseballového hřiště* [online]. [cit. 2016-10-16]. Obrázek ve formátu JPG. Dostupné z: http://www.mlbcz.com/hriste.php.

[7] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistiky,* Univerzita Palackého v Olomouci, Olomouc, 2013.

[8] Johnson, R., Wichern, D.: *Applied Multivariate Statistical Analysis* Pearson, 2008.

[9] Kalivodová, A.: *Biplot a jeho aplikace,* Bakalářská práce, přírodovědecká fakulta UP, Olomouc, 2010.

[10] Komsta, L.: *Package 'moments'* [online]. 2015, [cit. 2016-10-16]. dostupné z: https://cran.r-project.org/web/packages/moments/moments.pdf.

[11] Ligges, U.: *Package 'scatterplot3d"* [online]. 2016, [cit. 2016-10-28]. dostupné z: https://cran.r-project.org/web/packages/scatterplot3d/scatterplot3d.pdf.

[12] Orava, J.: *Jádrové odhady a binární data,* Bakalářská práce, Přírodovědecká fakulta MU, Brno, 2006.

[13] Owen, V.J.: *The R guide,* Department of Mathematics and Computer Science University of Richmond, Richmond, 2010.

[14] Poncet, P.: *Package 'modeest'* [online]. 2012, [cit. 2016-10-16]. dostupné z: https://cran.r-project.org/web/packages/modeest/modeest.pdf.

[15] Sarkar, D.: *Package 'lattice'* [online]. 2016, [cit. 2016-11-14]. dostupné z: https://cran.r-project.org/web/packages/lattice/lattice.pdf.

[16] Spanhel, F.: *Chi Plots* [online]. 2012, [cit. 2016-10-20]. dostupné z: http://www.statistik.lmu.de/∼robinzoni/posters/poster−chiplots.pdf.

[17] Süss, V.: *Softball a baseball,* Grada, Praha, 2003.

[18] Wand, M., Ripley, B.: *Package 'KernSmooth'* [online]. 2015, [cit. 2016-10-26]. dostupné z: https://cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf.