Analyse der erfolgreichsten Beachvolleyballteams der Welt mit Hilfe von R

Sonstige Beteiligung Angewandte Programmierung

Mark Ebinger

FOM – Master of Science – Big Data & Business Analytics

Stuttgart, 30.07.2022

Agenda

1. Warum Beachvolleyballdaten?
2. Wie bin ich vorgegangen?
   a. Data Understanding
   b. Data Preparation
   c. Data Modelling & Evaluation
3. Fazit

Agenda

1. **Warum Beachvolleyballdaten?**
2. Wie bin ich vorgegangen?
   a. Data Understanding
   b. Data Preparation
   c. Data Modelling & Evaluation
3. Fazit

Warum Beachvolleyballdaten?

- Große Leidenschaft für (Beach-)Volleyball

- Wer sind die besten Beachvolleyballteams der Welt?
- Bin ich mit 178 cm zu klein?
- Bin ich mit 28 Jahren schon zu alt?

Agenda

1. Warum Beachvolleyballdaten?
2. **Wie bin ich vorgegangen?**
   a. **Data Understanding**
   b. Data Preparation
   c. Data Modelling & Evaluation
3. Fazit

Data Understanding

- Datensatz aus Kaggle mit 78.756 Datensätze und 65 Spalten
- Beachvolleyball-Profispiele 2000 - 2019

```
# CRISP-DM - Schritt 2: Data Understanding

# Schritt 2.1: Daten laden

beach1 <- read.csv("C:/Users/Mark/Karriere/2021-2024_Master Big Data/2_Semester/

# Schritt 2.2.: Erster Blick auf die Daten

head(beach1)
tail(beach1)
summary(beach1)
str(beach1)
```

Filter | Cols: « < 1 - 50 > »

	circuit	tournament	country	year	date	gender	match_num	w_player1	w_p1_birthdate	w_p1_age	w_p
1	AVP	Huntington Beach	United States	2002	2002-05-24	M	1	Kevin Wong	1972-09-12	29.69473	
2	AVP	Huntington Beach	United States	2002	2002-05-24	M	2	Brad Torsone	1975-01-14	27.35661	
3	AVP	Huntington Beach	United States	2002	2002-05-24	M	3	Eduardo Bacil	1971-03-11	31.20329	
4	AVP	Huntington Beach	United States	2002	2002-05-24	M	4	Brent Doble	1970-01-03	32.38604	
5	AVP	Huntington Beach	United States	2002	2002-05-24	M	5	Albert Hannemann	1970-05-04	32.05476	

Agenda

1. Warum Beachvolleyballdaten?
2. **Wie bin ich vorgegangen?**
   a. Data Understanding
   b. **Data Preparation**
   c. Data Modelling & Evaluation
3. Fazit

Data Preparation

```
# Schritt 3.1: Nur relevante Turnierdaten herausfiltern

beach2 <- subset(beach1, beach1$circuit=="FIVB"
                 & beach1$gender=="M"
                 & beach1$bracket!="Bronze Medal"
                 & beach1$bracket!="Qualifier Bracket"
                 & beach1$bracket!="Qualifier Playoff"
                 & beach1$bracket!= "Country Quota Matches"
                 & beach1$bracket!= "Lucky Losers")
```

Herausforderungen in Daten:

- Pro Spieler eine Spalte → 2 Spieler pro Team = 2 Spalten
- Pro Match eine Zeile:
  - Spalten pro Gewinnerteam
  - Spalten pro Verliererteam

	w_player1	w_p1_birthdate	w_p1_age	w_p1_hgt	w_p1_country	w_player2	w_p2_birthdate	w_p2_age	w_p2_hgt
1	Kevin Wong	1972-09-12	29.69473	79	United States	Stein Metzger	1972-11-17	29.51403	75
2	Brad Torsone	1975-01-14	27.35661	78	United States	Casey Jennings	1975-07-10	26.87201	75
3	Eduardo Bacil	1971-03-11	31.20329	74	Brazil	Fred Souza	1972-05-13	30.02875	79
4	Brent Doble	1970-01-03	32.38604	78	United States	Karch Kiraly	2060-11-03	41.55236	74
5	Albert Hannemann	1970-05-04	32.05476	75	United States	Jeff Nygaard	1972-08-03	29.80424	80

	l_player1	l_p1_birthdate	l_p1_age	l_p1_hgt	l_p1_country	l_player2	l_p2_birthdate	l_p2_age	l_p2_hgt	l_p2_count
	Chuck Moore	1973-08-18	28.76386	76	United States	Ed Ratledge	1976-12-16	25.43463	80	United Stat
	Mark Paaluhi	1971-03-08	31.21150	75	United States	Nick Hannemann	1972-01-12	30.36277	78	United Stat
	Adam Jewell	1975-06-24	26.91581	77	United States	Collin Smith	1975-05-26	26.99521	76	United Stat
	David Swatik	1973-02-14	29.27036	76	United States	Mike Mattarocci	1969-10-05	32.63244	80	United Stat
	Adam Roberts	1976-01-25	26.32717	73	United States	Jim Walls	1978-03-26	24.16153	75	United Stat

Data Preparation

```
# Schritt 3.2: Spalte mit Sieg = 1 & Niederlage = 0
beach2$win <- 1
beach2$lost <- 0

# Schritt 3.3: Teams mit Teamdaten erzeugen
beach2$wteam <- paste(beach2$w_player1, beach2$w_player2, sep="/")
beach2$avg_w_age <- (beach2$w_p1_age + beach2$w_p2_age)/2
beach2$avg_w_heightcm <- ((beach2$w_p1_hgt + beach2$w_p2_hgt)/2)*2.54

beach2$lteam <- paste(beach2$l_player1, beach2$l_player2, sep="/")
beach2$avg_l_age <- (beach2$l_p1_age + beach2$l_p2_age)/2
beach2$avg_l_heightcm <- ((beach2$l_p1_hgt + beach2$l_p2_hgt)/2)*2.54

# Schritt 3.4: Eine Datentabelle nur mit Siegerdaten

beach2winner <- subset(beach2, select=c(year, wteam, w_p1_country,avg_w_age, avg_w_heightcm, win))

# Schritt 3.5: Eine Datentabelle nur mit Verliererdaten

beach2loser <- subset(beach2, select=c(year, lteam, l_p1_country,avg_l_age, avg_l_heightcm, lost))

# Schritt 3.6: Gleiche Bezeichnung der Überschriften
names(beach2winner) <- c("year", "team", "country", "age", "height", "win_lost")
names(beach2loser) <- c("year", "team", "country", "age", "height", "win_lost")

# Schritt 3.7: Daten zusammenfassen in einem Datensatz untereinander
beach3 <- rbind(beach2winner, beach2loser)
```

Agenda

1. Warum Beachvolleyballdaten?
2. **Wie bin ich vorgegangen?**
   a. Data Understanding
   b. Data Preparation
   c. **Data Modelling & Evaluation**
3. Fazit

Data Modelling & Evaluation

Wer sind die besten Beachvolleyballteams der Welt?

```
# CRISP-DM - Schritt 4: Modelling

# Frage 4.1: Was waren die erfolgreichsten Teams?

topteams <- beach3 %>%
  group_by (team, country) %>%
  summarize(height = mean(height) ,wins = sum(win_lost), games = n())

topteams$winning_rate <- toptea

topteams50 <- topteams %>% filt
```

	team	country	height	wins	games	winning_rate
1	Oleg Stoyanovskiy/Viacheslav Krasilnikov	Russia	200.66	48	59	0.8135593
2	Phil Dalhausser/Todd Rogers	United States	196.85	255	314	0.8121019
3	Anders Mol/Christian Sorum	Norway	189.23	108	133	0.8120301
4	Emanuel Rego/Tande Ramos	Brazil	194.31	76	94	0.8085106
5	Emanuel Rego/Ricardo Santos	Brazil	195.58	435	544	0.7996324
6	Jonas Reckermann/Julius Brink	Germany	193.04	139	174	0.7988506
7	Alison Cerutti/Emanuel Rego	Brazil	196.85	173	218	0.7935780
8	Alison Cerutti/Harley Marques	Brazil	198.12	53	67	0.7910448
9	Phil Dalhausser/Sean Rosenthal	United States	199.39	86	109	0.7889908
10	Alison Cerutti/Bruno Oscar Schmidt	Brazil	194.31	175	222	0.7882883

Data Modelling & Evaluation
Bin ich zu klein?

```
# Frage 4.3.2: Sind größere Teams erfolgreicher als kleinere Teams?
summary(topteams100)

summary(lm(topteams100$winning_rate~topteams100$height, data=topteams100))
```

```
     team              country             height          wins            games          winning_rate
Length:185         Length:185         Min.   :185.4   Min.   : 13.00   Min.   : 51.0   Min.   :0.2388
Class :character   Class :character   1st Qu.:191.8   1st Qu.: 33.00   1st Qu.: 72.0   1st Qu.:0.4259
Mode  :character   Mode  :character   Median :194.3   Median : 53.00   Median :103.0   Median :0.5192
                                      Mean   :194.4   Mean   : 74.56   Mean   :132.6   Mean   :0.5283
                                      3rd Qu.:196.8   3rd Qu.: 83.00   3rd Qu.:166.0   3rd Qu.:0.6211
                                      Max.   :205.7   Max.   :435.00   Max.   :544.0   Max.   :0.8136
                                      NA's   :3
>
```

Data Modelling & Evaluation
Bin ich zu klein?

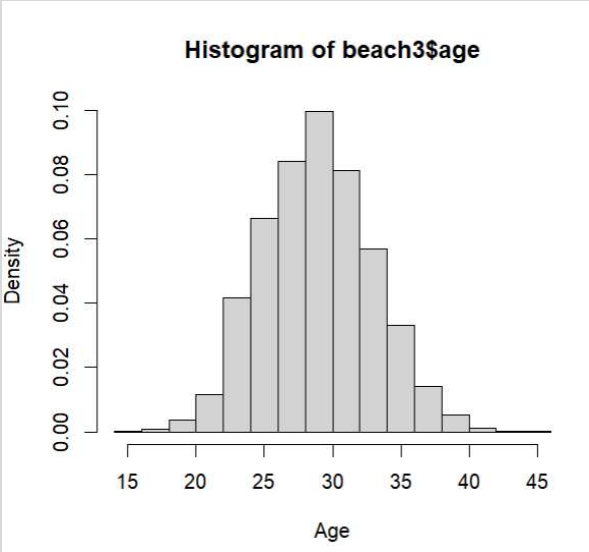
```
# Frage 4.3.2: Sind größere Teams erfolgreicher als kleinere Teams?
summary(topteams100)

summary(lm(topteams100$winning_rate~topteams100$height, data=topteams100))
#P-Wert: Hohe Signifikanz 0.00161
#Regressionskoeffizenz: 0.008314 --> 1 cm größer erhöht die Siegquote um 0,8 %
#Bestimmtheitsmaß:0,05393
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.34237 -0.09177 -0.01274  0.08235  0.32588

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.087020   0.504714  -2.154  0.03259 *
topteams100$height  0.008314   0.002595   3.203  0.00161 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Data Modelling & Evaluation
Bin ich schon zu alt?

```r
# Frage 4.3.1: Exkurs Histogram Altersverteilung

hist(beach3$age, freq = FALSE, xlab="Age")
mean(beach3$age)

# Frage 4.3.1: Sind erfahrenere Teams (über 30 Jahre) erfolgreicher als junge Teams?
beach3$age_group <- ifelse(beach3$age>30, "experienced", "young & wild")

t.test(win_lost~age_group, data = beach3)

#Ergebnis: Ja, höchste Signifikanz
# Erfahren: 0.53 vs. Young & Wild: 0.48
```

```
        Welch Two Sample t-test

data:  win_lost by age_group
t = 8.3492, df = 30157, p-value < 2.2e-16
alternative hypothesis: true difference in means between group experienced and group young & wild is not equal to 0
95 percent confidence interval:
 0.03409653 0.05501641
sample estimates:
 mean in group experienced mean in group young & wild
                 0.5283271                  0.4837706
```



Agenda

1. Warum Beachvolleyballdaten?
2. Wie bin ich vorgegangen?
   a. Data Understanding
   b. Data Preparation
   c. Data Modelling & Evaluation
3. **Fazit**

Fazit

- Zu Beginn Datensatz prüfen, ob Problemstellung wirklich beantwortet werden kann → Zahlreiche Möglichkeiten in R, Daten zu bearbeiten
- Die erfolgreichsten Teams kommen aus Russland, USA, Norwegen, Brasilien und Deutschland
- Ja, ich bin zu klein → Es gibt aber auch zahlreiche andere Einflüsse
- Nein, ich bin im besten Beachvolleyballalter → Erfahrung kann ein Sieggarant sein


London 2012
2
GER
1
GER