

# Predicting House Prices

Oliver Tomondy, Friedrich Winkelbauer

28/12/2021

## Ziele

- Das Ziel dieser Arbeit ist es, einen Datensatz mit Seattle Häuser zu analysieren und Preis der Häuser in der Stadt Seattle vorherzusagen
- ...
- 

## Libraries

```
#install.packages("corrplot")
#install.packages("plumber", "rjson")
##install.packages(c("cowplot", "googleway", "ggplot2", "ggrepel",
##"ggspatial", "libwgeom", "sf", "rnaturalearth", "rnaturalearthdata"))
library(zoo)
library(corrplot)
library(tidyverse)
library(tidygraph)
library(igraph)
library(ggplot2)
library(ggraph)
library(rnaturalearth)
library(rnaturalearthdata)
```

## Datenaufbereitung

Zuerst lesen wir die Daten ein. Wir verwenden dafür `read_delim` anstatt `read_csv` um den Spaltentyp zu schätzen.

```
data = read_delim("data/house_sales.csv", delim=",")
data = data %>% as_tibble()
```

Wir entfernen einen Ausreißer, der wahrscheinlich nur eine Fehleingabe war.

```
data = data %>% subset(bedrooms != 33)
```

# Explorative Datenanalyse

Unsere Datensatz enthält Informationen über 21.613 Häuser in der US-amerikanischen Stadt Seattle. Jedes Haus ist durch eine ID gekennzeichnet und ist durch 19 Merkmale beschrieben. Unten findet man einen Überblick dieser Merkmale.

```
summary(data)
```

```
##          id          date          price
## Length:21612    Min.   :2014-05-02 00:00:00    Min.   : 75000
## Class :character 1st Qu.:2014-07-22 00:00:00    1st Qu.: 321838
## Mode  :character Median :2014-10-16 00:00:00    Median : 450000
##          Mean   :2014-10-29 04:46:26    Mean   : 540084
##          3rd Qu.:2015-02-17 00:00:00    3rd Qu.: 645000
##          Max.   :2015-05-27 00:00:00    Max.   :7700000
## bedrooms      bathrooms      sqft_living      sqft_lot
## Min.   : 0.000    Min.   :0.000    Min.   : 290    Min.   : 520
## 1st Qu.: 3.000    1st Qu.:1.750    1st Qu.: 1426    1st Qu.: 5040
## Median : 3.000    Median :2.250    Median : 1910    Median : 7619
## Mean   : 3.369    Mean   :2.115    Mean   : 2080    Mean   : 15107
## 3rd Qu.: 4.000    3rd Qu.:2.500    3rd Qu.: 2550    3rd Qu.: 10688
## Max.   :11.000    Max.   :8.000    Max.   :13540    Max.   :1651359
## floors      waterfront      view      condition
## Min.   :1.000    Min.   :0.000000    Min.   :0.0000    Min.   :1.000
## 1st Qu.:1.000    1st Qu.:0.000000    1st Qu.:0.0000    1st Qu.:3.000
## Median :1.500    Median :0.000000    Median :0.0000    Median :3.000
## Mean   :1.494    Mean   :0.007542    Mean   :0.2343    Mean   :3.409
## 3rd Qu.:2.000    3rd Qu.:0.000000    3rd Qu.:0.0000    3rd Qu.:4.000
## Max.   :3.500    Max.   :1.000000    Max.   :4.0000    Max.   :5.000
## grade      sqft_above      sqft_basement      yr_built
## Min.   : 1.000    Min.   : 290    Min.   : 0.0    Min.   :1900
## 1st Qu.: 7.000    1st Qu.:1190    1st Qu.: 0.0    1st Qu.:1951
## Median : 7.000    Median :1560    Median : 0.0    Median :1975
## Mean   : 7.657    Mean   :1788    Mean   : 291.5    Mean   :1971
## 3rd Qu.: 8.000    3rd Qu.:2210    3rd Qu.: 560.0    3rd Qu.:1997
## Max.   :13.000    Max.   :9410    Max.   :4820.0    Max.   :2015
## yr_renovated      zipcode      lat      long
## Min.   : 0.00    Min.   :98001    Min.   :47.16    Min.   : -122.5
## 1st Qu.: 0.00    1st Qu.:98033    1st Qu.:47.47    1st Qu.: -122.3
## Median : 0.00    Median :98065    Median :47.57    Median : -122.2
## Mean   : 84.41    Mean   :98078    Mean   :47.56    Mean   : -122.2
## 3rd Qu.: 0.00    3rd Qu.:98118    3rd Qu.:47.68    3rd Qu.: -122.1
## Max.   :2015.00    Max.   :98199    Max.   :47.78    Max.   : -121.3
## sqft_living15      sqft_lot15
## Min.   : 399    Min.   : 651
## 1st Qu.:1490    1st Qu.: 5100
## Median :1840    Median : 7620
## Mean   :1987    Mean   : 12769
## 3rd Qu.:2360    3rd Qu.: 10083
## Max.   :6210    Max.   :871200
```

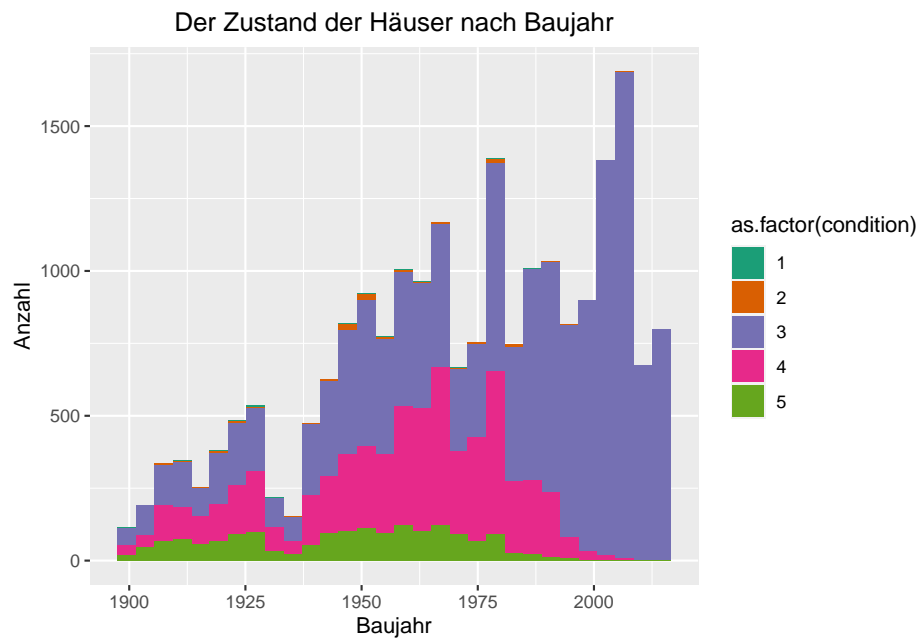
- Der durchschnittliche Preis eines Hauses im Datensatz beträgt 540.084 US-Dollar. Das teuerste Haus kostet 7.7 Millionen US-Dollar.
- Die Wohnfläche beträgt durchschnittlich 2.080 Quadraftfuß, was in ca. 193 Quadratmeter ist.
- Die Median Größe eines Grundstücks beträgt 7.618 Quadratfuß, wobei das größte Grundstück 1.651.359 Quadratfuß hat.
- Die Häuser in unserem Datensatz haben außerdem durchschnittlich 3.4 Zimmer und 2.25 Badezimmer.
- Von Mehr als 20 Tausend Häuser liegen nur 163 am Wasser.

Schauen wir uns nun weitere Statistiken graphisch an:

## 1. Zustand der Häuser nach Baujahr

```
data %>%  
  ggplot(aes(x=yr_built, fill=as.factor(condition))) +  
  geom_histogram() +  
  ggtitle("Der Zustand der Häuser nach Baujahr") +  
  xlab("Baujahr") + ylab("Anzahl") +  
  scale_fill_brewer(palette = "Dark2") +  
  theme(plot.title = element_text(hjust = 0.5))
```

## 'stat\_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

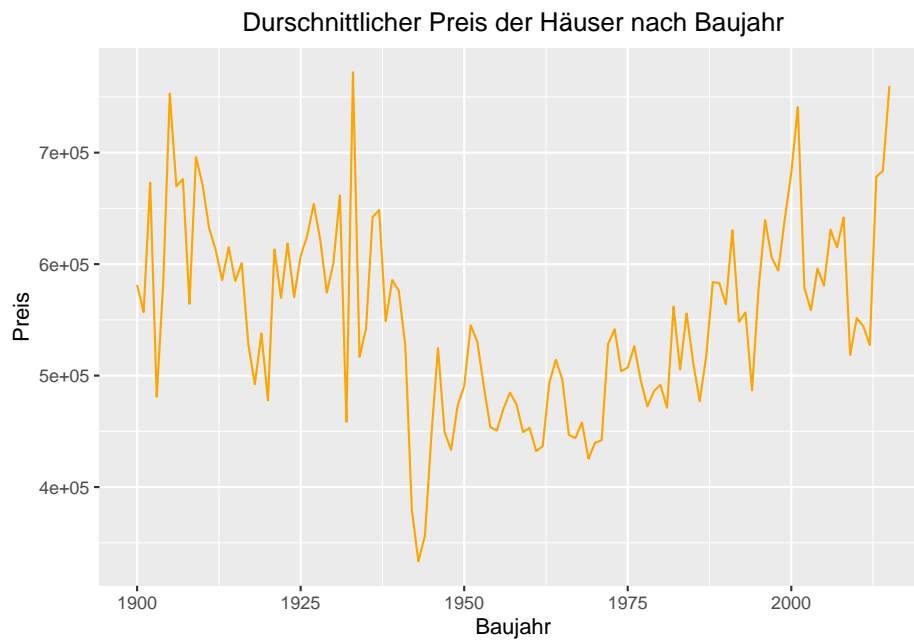


- Auf dem Diagramm ist zu sehen, dass

---

## 2. Durchschnittlicher Preis der Häuser nach Baujahr

```
data %>%
  ggplot(aes(x=yr_built,y=price)) +
  geom_line(stat = "summary", fun = "mean", color="orange") +
  ggtitle("Durschnittlicher Preis der Häuser nach Baujahr") +
  xlab("Baujahr") + ylab("Preis") +
  scale_fill_brewer(palette = "Dark2") +
  theme(plot.title = element_text(hjust = 0.5))
```



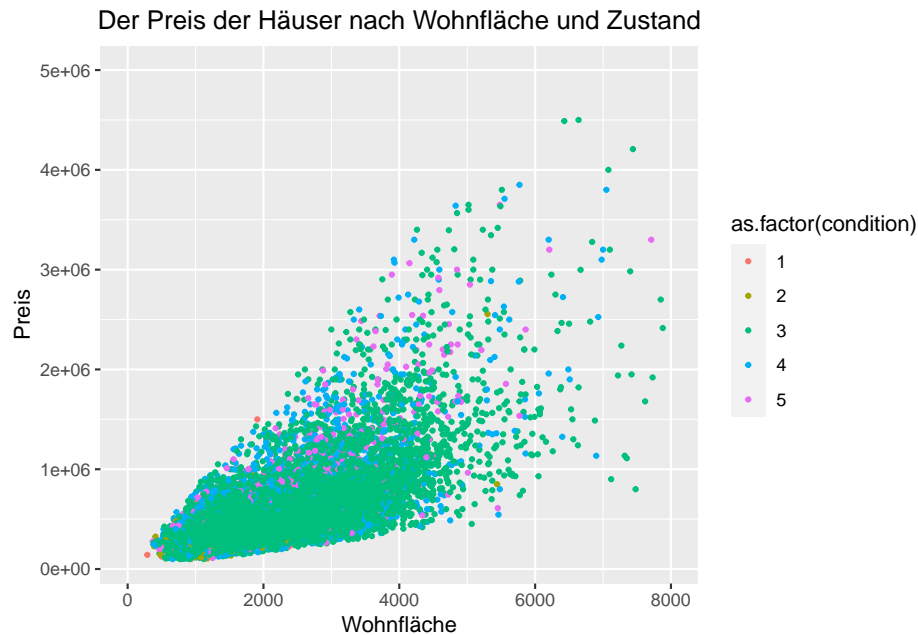
- Auf dem Diagramm sehen wir den durchschnittlichen Verkaufspreis der Häuser nach dem Baujahr.
- Auffällig ist, dass Häuser, die zwischen 1900 und 1930 gebaut wurden, durchschnittlich einen höheren Preis haben als Häuser, die zwischen den Jahren 1945 und 1980 gebaut wurden. Erst ganz junge Häuser, die am Ende des 20. Jahrhunderts und am Anfang des 21. Jahrhunderts gebaut wurden, sind wieder teurer.

---

### 3. Preis der Häuser nach Wohnfläche und Zustand

```
data %>%
  ggplot(aes(x=sqft_living,y=price, colour =as.factor(condition))) +
  geom_point(size=0.8) +
  ggtitle("Der Preis der Häuser nach Wohnfläche und Zustand") +
  xlab("Wohnfläche") +
  ylab("Preis") +
  scale_fill_brewer(palette = "Dark2") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylim(100000, 5000000) +
  xlim(0, 8000)
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```

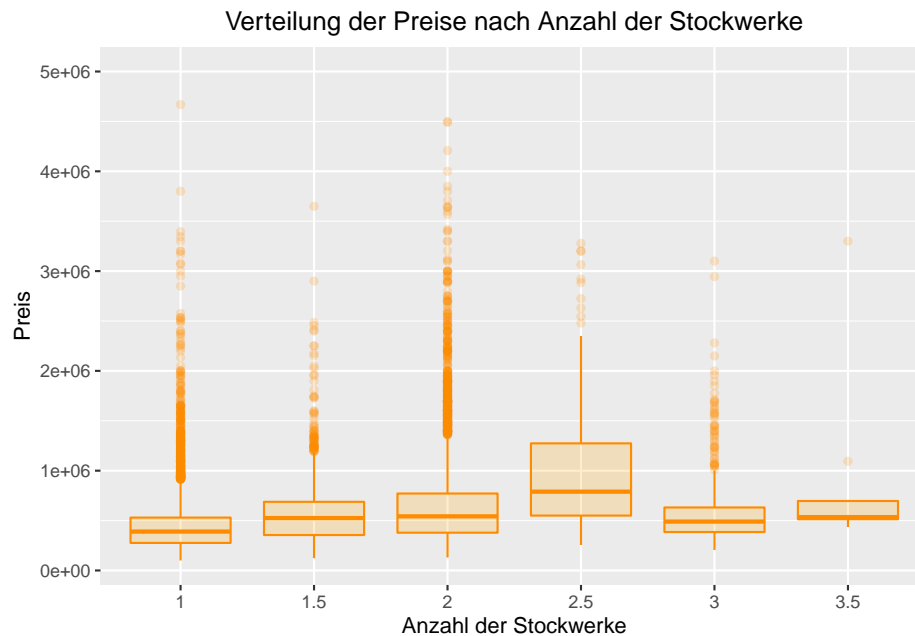


- Das Diagramm zeigt den Preis der Häuser nach Wohnfläche und Zustand.
- Es lässt sich deutlich erkennen, dass mit steigender Wohnfläche auch der Preis für ein Haus steigt.
- Leider kann man nicht deutlich sehen, ob der Zustand auch eine Rolle beim Preis des Hauses spielt. Es ist lediglich zu beobachten, dass die Mehrheit der Häuser in einem mittleren Zustand sind.

#### 4. Verteilung der Preise nach Anzahl der Stockwerke

```
data %>%
  ggplot(aes(x=as.factor(floors), y=price)) +
  geom_boxplot(color="darkorange", fill="orange", alpha=0.2) +
  ylim(100000, 5000000) +
  ggtitle("Verteilung der Preise nach Anzahl der Stockwerke") +
  xlab("Anzahl der Stockwerke") + ylab("Preis") +
  scale_colour_brewer(palette = "Dark2") +
  theme(plot.title = element_text(hjust = 0.5))
```

## Warning: Removed 32 rows containing non-finite values (stat\_boxplot).



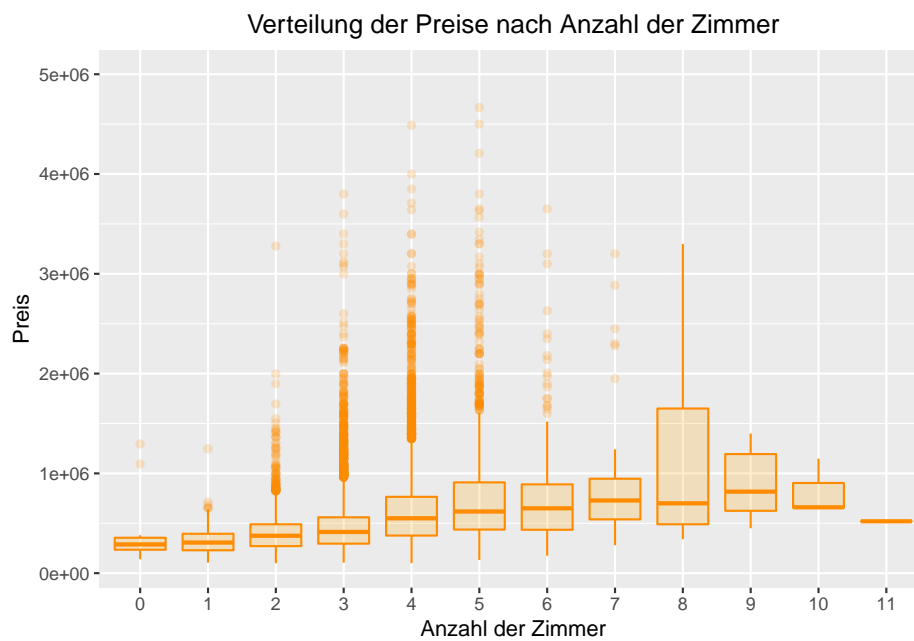
- Das Diagramm zeigt die Verteilung der Hauspreise nach der Anzahl der Stockwerke.
- Es lässt sich erkennen, dass je mehr Stockwerke das Haus hat, desto höher der Preis ist. Interessenterweise gilt dieser Trend nur bis zu 2.5 Stockwerken. Wenn ein Haus 3 oder 3.5 Stockwerke hat, ist der Preis durchschnittlich niedriger als bei Häusern mit nur 2.5 Stockwerken.
- Vielleicht lässt sich eine deutlichere Tendenz bei der Anzahl der Zimmer feststellen.



---

## 5. Verteilung der Preise nach Anzahl der Zimmer

```
data %>%  
  ggplot(aes(x=as.factor.bedrooms), y=price)) +  
  geom_boxplot(color="darkorange", fill="orange", alpha=0.2) +  
  ylim(100000, 5000000) +  
  ggtitle("Verteilung der Preise nach Anzahl der Zimmer") +  
  xlab("Anzahl der Zimmer") + ylab("Preis") +  
  scale_colour_brewer(palette = "Dark2") +  
  theme(plot.title = element_text(hjust = 0.5))
```

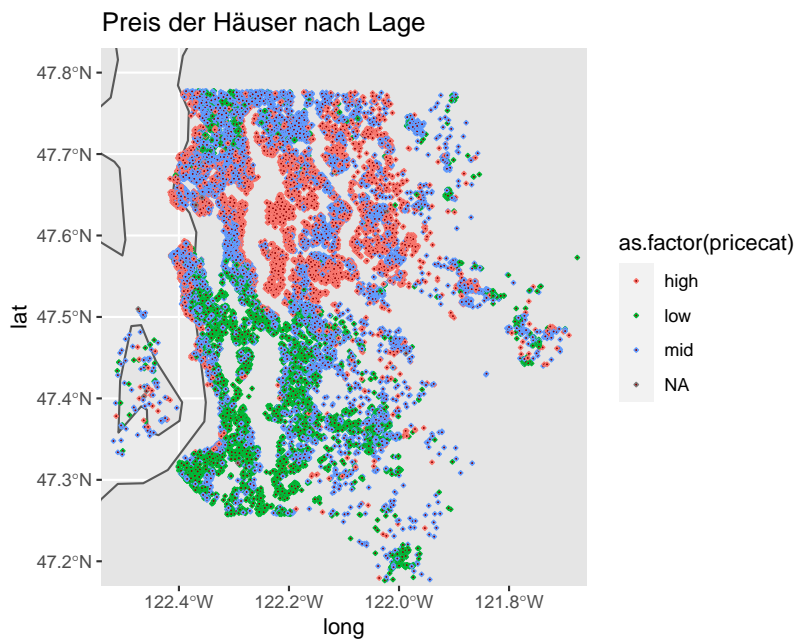


- Auf der Grafik können wir die Verteilung der Hauspreise nach der Anzahl der Zimmer beobachten.
- Es gibt einen klaren aufsteigenden Trend: Also je mehr Zimmer ein Haus hat, desto mehr wird er wahrscheinlich kosten.

## 6. Preis der Häuser nach Lage

```
data = data %>% mutate(pricecat = case_when(
  price < 321950 ~ 'low',
  price < 645000 ~ 'mid',
  price > 645000 ~ 'high'
))

world <- ne_countries(scale = "medium", returnclass = "sf")
ggplot(data = world) +
  geom_sf() +
  geom_point(data = data, aes(x = long, y = lat, color=as.factor(pricecat)), size = 0.5,
    shape = 23, fill = "darkred") +
  ggtitle("Preis der Häuser nach Lage")+
  coord_sf(xlim = c(-122.5, -121.7), ylim = c(47.20, 47.8))
```



•

## 7. Korrelation der einzelnen Merkmale mit Preis

```
datacor = data %>% select(-c("id", "date", "pricecat"))
corrplot(cor(datacor$price, datacor), method="number", diag = FALSE, tl.cex = 0.8,
  number.cex=0.75, cl.pos = "n", title="Korrelation der Merkmale mit dem Preis")
```

## Korrelation der Merkmale mit dem Preis

1	bedrooms	0.32
	bathrooms	0.53
	sqft_living	0.70
	sqft_lot	0.09
	floors	0.26
	waterfront	0.27
	view	0.40
	condition	0.04
	grade	0.67
	sqft_above	0.61
	sqft_basement	0.32
	yr_built	0.05
	yr_renovated	0.13
	zipcode	-0.05
	lat	0.31
	long	0.82
	sqft_living15	0.59
	sqft_lot15	0.08

# Modelierung

...