Projekt: Datenanalyse mit R

Jonas Haberstroh

2022-11-14

Dataunderstanding und Zielsetzung

In dem folgenden Report wird ein Datensatz analysiert welcher Immobiliendaten beinhaltet. Dieser wird zuerst erläutert, anschließend wird das Ziel des Reports definiert und zuletzt folgt die Analyse und Überprüfung der Hypothese.

Beschreibung des Datensatzes

Der Datensatz besteht aus 2342 Zeilen und 21 Spalten. Die Beschreibung des Datesatzes folgt unter dem folgenden Code Block.

```
# Import der Daten
df_immobilien <- read.csv(file="Dataset/dataset_immobilien.csv", sep=";")
print(paste("Shape: (", nrow(df_immobilien), ",", ncol(df_immobilien), ")"))
```

```
## [1] "Shape: ( 2342 , 21 )"
```

```
# Beschreibung des Datensatzes
head(df_immobilien)
```

```
##   A_Index AnzahlZimmer Ausbaustufe Baeder BaederKG Baujahr EG_qm Garage_qm
## 1    2358            3     1 Ebene      2        1    1992   125        49
## 2     266            2     1 Ebene      2        1    2010   170        79
## 3    1169            2     1 Ebene      2        0    2015   119        40
## 4    2564            2    2 Ebenen      3        1    2015    64        40
## 5     169            3     1 Ebene      2        0    2021   103        39
## 6     435            2     1 Ebene      2        1    1988    89        27
##   Garagen Gesamteindruck Keller_Typ_qm Keller_qm  Kellerhoehe
## 1       2              3            88       116          Gut
## 2       3              3           141       168          Gut
## 3       2              3             0       119          Gut
## 4       2              3            48        64          Gut
## 5       2              3             3       103          Gut
## 6       1              4            10        89 Durchschnitt
##            Kellertyp      Lage OG_qm Umgebaut Verkaufsjahr Verkaufsmonat
## 1     Guter Wohnraum Bezirk 19     0     1992         2021             6
## 2     Guter Wohnraum Bezirk 16     0     2010         2020             7
## 3             Rohbau Bezirk 18     0     2015         2018             3
## 4     Guter Wohnraum Bezirk 18    73     2016         2020            10
## 5     Guter Wohnraum  Bezirk 8     0     2021         2022             3
## 6 Mittlerer Wohnraum Bezirk 17     0     1988         2018             4
##   Wohnflaeche_qm Z_Verkaufspreis
## 1            125          187500
## 2            170          350000
## 3            119          171750
## 4            138          154000
## 5            103          213899
## 6             89          137500
```

1. **A_Index:** Eindeutige Identifikationsnummer, nicht fortlaufend (durch Sampling in die ausgegebenen und zurückgehaltenen Daten)
2. **Anzahl Zimmer:** Gesamtanzahl der Zimmer (keine Küchen und Bäder eingerechnet)
3. **Ausbaustufe:** Anzahl der Ebenen oberhalb des Kellers
    - 1 Ebene
    - 2 Ebenen
4. **Baeder:** Anzahl der Badezimmer die nicht im Kellergeschoss (KG) liegen, Toiletten eingerechnet

5. **BaederKG:** Analog Baeder, aber im KG
6. **Baujahr:** Jahr in dem das Gebäude gebaut wurde
7. **EG_qm:** Größe der Wohnfläche in qm im Erdgeschoss
8. **Garage_qm:** Größe der Garage in qm
9. **Garagen:** Anzahl der Fahrzeuge, die in der Garage abgestellt werden können
10. **Gesamteindruck:** Eindruck des Gesamtzustandes des Gebäudes insgesamt
    - 5 Sehr gut
    - 4 Gut
    - 3 Durchschnitt
    - 2 Schlecht
    - 1 Sehr schlecht
11. **Keller_Typ_qm:** Anzahl der qm im Typ des Kellers (siehe „Kellertyp" unten)
12. **Keller_qm:** Anzahl der qm des gesamten Kellers
13. **Kellerhoehe:** Höhe des Kellers
    - Sehr gut: ca. 250 cm
    - Gut: ca. 225 cm
    - Durchschnitt: ca. 200 cm
    - Schlecht: ca. 175 cm
    - Sehr schlecht: niedriger als 175 cm
    - Keine Angabe: kein Keller
14. **Kellertyp:** Typ des Kellers
    - Guter Wohnraum
    - Mittlerer Wohnraum
    - Kein Wohnraum
    - Freizeitraum
    - Niedrige Qualität
    - Rohbau
15. **Lage:** Bezirk, in dem die Immobilie steht
16. **OG_qm:** Quadratmeter des Geschosses oberhalb des OG
17. **Umgebaut:** Jahr, in dem größere Umbauten / Anbauten / Renovierungen stattfanden, wenn keine durchgeführt wurden entspricht dies dem Baujahr Verkaufsjahr: Jahr des Verkaufs
18. **Verkaufsmonat:** Monat des Verkaufs
19. **Wohlflaeche_qm:** Wohnfläche in qm
20. **Z_Verkaufspreis:** Verkaufspreis in Euro

Zielsetzung und Hypothesen

Der Datensatz soll dazu verwendet werden Aufschluss über den Zusammenhang bestimmter Attribute mit dem Kaufpreis zu bekommen. Basierend auf den Erkenntnissen des Reports sollen grundlegende Regeln abgeleitet werden, welche bei der Investition in eine Immobilie beachtet werden sollten.

Die folgenden Hypothesen sollen dabei genauer betrachtet werden:

- **In welchen Bezirken erlebt der Kaufpreis der Immobilien das größte Wachstum?**

  In dieser Frage soll Aufschluss über den Zusammenhang zwischen dem Gebiet der Immobilie und dem Kaufpreis gewonnen werden. Welche Gebiete haben besonders teure Immobilien? Zusätzlich ist für eine Investition wichtig, in Gebieten zu kaufen in welchen sich der Kaufpreis in Zukunft erhöhen wird. Somit stellt sich die Frage welche Gebiete in der Vergangenheit eine

Preiserhöhung erfahren haben.

- **Aus welchem Baujahr sind Häuser besonders begehrt?**

  Mit dieser Frage sollen Zeitperioden herausgefunden werden in welchen Häuser aufgrund ihres Baujahrs einen hohen Preis erzielen.

  Die Erwartung ist, das neue Häuser generell einen höheren Verkaufspreis erzielen. Jedoch können alte Häuser, welche Renoviert sind für liebhaber Attraktive Immobilien sein, welche demnach einen hohen Verkaufspreis erzielen.

Laden der Daten

In diesem Block werden die Daten geladen und Fehler in diesen behoben.

```
str(df_immobilien)
```

```
## 'data.frame':    2342 obs. of  21 variables:
##  $ A_Index        : int  2358 266 1169 2564 169 435 961 1540 892 5 ...
##  $ AnzahlZimmer   : int  3 2 2 2 3 2 3 3 3 3 ...
##  $ Ausbaustufe    : chr  "1 Ebene" "1 Ebene" "1 Ebene" "2 Ebenen" ...
##  $ Baeder         : int  2 2 2 3 2 2 1 3 3 1 ...
##  $ BaederKG       : int  1 1 0 1 0 1 1 1 1 0 ...
##  $ Baujahr        : int  1992 2010 2015 2015 2021 1988 1988 2013 2011 1973 ...
##  $ EG_qm          : int  125 170 119 64 103 89 75 70 80 97 ...
##  $ Garage_qm      : int  49 79 40 40 39 27 26 53 38 24 ...
##  $ Garagen        : int  2 3 2 2 2 1 1 2 2 1 ...
##  $ Gesamteindruck : int  3 3 3 3 3 4 4 3 3 3 ...
##  $ Keller_Typ_qm  : int  88 141 0 48 3 10 15 40 74 66 ...
##  $ Keller_qm      : int  116 168 119 64 103 89 75 70 78 97 ...
##  $ Kellerhoehe    : chr  "Gut" "Gut" "Gut" "Gut" ...
##  $ Kellertyp      : chr  "Guter Wohnraum" "Guter Wohnraum" "Rohbau" "Guter Wohnraum" ...
##  $ Lage           : chr  "Bezirk 19" "Bezirk 16" "Bezirk 18" "Bezirk 18" ...
##  $ OG_qm          : int  0 0 0 73 0 0 0 72 65 0 ...
##  $ Umgebaut       : int  1992 2010 2015 2016 2021 1988 1988 2013 2012 1973 ...
##  $ Verkaufsjahr   : int  2021 2020 2018 2020 2022 2018 2022 2020 2019 2018 ...
##  $ Verkaufsmonat  : int  6 7 3 10 3 4 2 9 4 7 ...
##  $ Wohnflaeche_qm : int  125 170 119 138 103 89 75 142 145 97 ...
##  $ Z_Verkaufspreis: int  187500 350000 171750 154000 213899 137500 215852 190500 180000 146000 ...
```

Die Datentypen der in dem Dataframe enthaltenen Features entsprechen den in der Beschreibung des Datensatzes angebenen Werten. Somit muss hier keine Anpassung vorgenommen werden. Jedoch fällt hier auf dass der Verkaufspreis nur Integer Werte beinhaltet. Dies ist interessant, da Preise generell stetig sind, jedoch bei hohen Beträgen nicht weiter verwunderlich.

```
library(skimr)
skim(df_immobilien)
```

Data summary

Name	df_immobilien
Number of rows	2342
Number of columns	21
_____	
Column type frequency:	
character	4
numeric	17
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Ausbaustufe	0	1	7	8	0	3	0
Kellerhoehe	0	1	1	13	0	6	0
Kellertyp	0	1	1	18	0	7	0
Lage	0	1	1	9	0	28	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
A_Index	0	1	1302.70	752.04	0	649.25	1310.5	1951.75	2602	▇▇▇▇▇
AnzahlZimmer	0	1	2.86	0.82	0	2.00	3.0	3.00	8	▁▇▂▁▁
Baeder	0	1	1.91	0.81	0	1.00	2.0	2.00	6	▇▇▅▁▁
BaederKG	0	1	0.49	0.53	0	0.00	0.0	1.00	3	▇▇▁▁▁
Baujahr	0	1	1980.76	29.63	1884	1964.00	1982.0	2009.00	2022	▁▁▂▅▇
EG_qm	0	1	95.86	31.63	28	73.00	89.0	114.00	324	▇▇▂▁▁
Garage_qm	0	1	38.29	17.39	0	26.00	39.0	48.00	126	▂▇▅▂▁
Garagen	0	1	1.71	0.74	0	1.00	2.0	2.00	5	▂▇▂▁▁
Gesamteindruck	0	1	3.45	0.66	1	3.00	3.0	4.00	5	▁▃▇▃▁
Keller_Typ_qm	0	1	36.76	35.74	0	0.00	32.0	61.00	194	▇▃▂▁▁
Keller_qm	0	1	86.24	34.91	0	66.00	81.5	105.00	272	▂▇▃▁▁
OG_qm	0	1	28.54	35.73	0	0.00	0.0	59.00	175	▇▃▂▁▁
Umgebaut	0	1	1994.70	20.43	1962	1976.00	2002.0	2014.00	2022	▆▂▁▃▇
Verkaufsjahr	0	1	2019.87	1.32	2018	2019.00	2020.0	2021.00	2022	▇▁▁▇▇
Verkaufsmonat	0	1	6.13	2.67	1	4.00	6.0	8.00	12	▃▅▇▅▅
Wohnflaeche_qm	0	1	125.01	41.59	28	93.00	121.0	146.00	380	▃▇▅▁▁
Z_Verkaufspreis	0	1	180443.51	77722.39	13100	130000.00	161950.0	213882.00	755000	▇▇▂▁▁

Alle Features besitzen kein Nullwerte und sind vollständig. Nun müssen die Features im einzelnen betrachtet werden um potentielle Null Werte zu entfernen. wDer Datensatz besitzt zusätzlich keine Ordnung der Daten.

```
cat_cols <- c("Ausbaustufe", "Gesamteindruck", "Kellerhoehe", "Kellertyp", "Lage")
for(cat_col in cat_cols) {
  print(paste("Unique values of ", cat_col, ":"))
  print(unique(df_immobilien[cat_col])[, cat_col])
}
```

```
## [1] "Unique values of  Ausbaustufe :"
## [1] "1 Ebene"  "2 Ebenen" "3 Ebenen"
## [1] "Unique values of  Gesamteindruck :"
## [1] 3 4 2 5 1
## [1] "Unique values of  Kellerhoehe :"
## [1] "Gut"           "Durchschnitt"  "0"             "Sehr gut"
## [5] "Schlecht"      "Sehr Schlecht"
## [1] "Unique values of  Kellertyp :"
## [1] "Guter Wohnraum"     "Rohbau"             "Mittlerer Wohnraum"
## [4] "Niedrige Qualität"  "0"                  "Freizeitraum"
## [7] "Kein Wohnraum"
## [1] "Unique values of  Lage :"
##  [1] "Bezirk 19" "Bezirk 16" "Bezirk 18" "Bezirk 8"  "Bezirk 17" "Bezirk 6"
##  [7] "Bezirk 23" "Bezirk 9"  "Bezirk 15" "Bezirk 20" "Bezirk 14" "Bezirk 1"
## [13] "Bezirk 24" "Bezirk 21" "Bezirk 22" "Bezirk 7"  "Bezirk 4"  "Bezirk 25"
## [19] "Bezirk 5"  "Bezirk 26" "Bezirk 27" "Bezirk 12" "Bezirk 2"  "Bezirk 13"
## [25] "Bezirk 10" "Bezirk 3"  "Bezirk 11" "0"
```

Die beiden kategorischen Features Kellerhoehe und Kellertyp beinhalten bei fehlendem Keller den Wert “0” dieser wird umbenannt. Das Feature Lage beinhaltet auch einen “0” Wert, da dieser sich jedoch nicht aus der Beschreibung des Datensatzes erschließt, handelt es sich dabei um Outlier, die entfernt werden müssen.

```
# Umbennenung kategorische Werte
df_immobilien["Kellerhoehe"][df_immobilien["Kellerhoehe"] == "0"] <- "Kein Keller"
df_immobilien["Kellertyp"][df_immobilien["Kellertyp"] == "0"] <- "Kein Keller"
# Entfernen outlier Lage
df_immobilien <- df_immobilien[df_immobilien["Lage"] != "0", ]
```

Transformation der Daten

In dem folgenden Block werden die Daten transformiert/bearbeitet um diese auf die Analyse vorzubereiten. Der Datensatz soll bereinigt und neue Features erstellt werden, falls dies der Analyse dienlich ist.

```
df_immobilien$Ausbaustufe <- factor(df_immobilien[, "Ausbaustufe"])
df_immobilien$Gesamteindruck <- factor(df_immobilien[, "Gesamteindruck"])
df_immobilien$Kellerhoehe <- factor(df_immobilien[, "Kellerhoehe"])
df_immobilien$Kellertyp <- factor(df_immobilien[, "Kellertyp"])
df_immobilien$Lage <- factor(df_immobilien[, "Lage"])
```

Die kategorischen Features (Ausbaustufe, Gesamteindruck, Kellerhoehe, Kellertyp, Lage) werden im Anschluss in Faktoren konvertiert.

```
data.frame(unclass(summary(df_immobilien)), check.names = FALSE, stringsAsFactors = FALSE)
```

```
##            A_Index    AnzahlZimmer     Ausbaustufe          Baeder
## X   Min.   :   0   Min.   :0.000   1 Ebene :1335   Min.   :0.000
## X.1 1st Qu.: 649   1st Qu.:2.000   2 Ebenen: 978   1st Qu.:1.000
## X.2 Median :1310   Median :3.000   3 Ebenen:  28   Median :2.000
## X.3 Mean   :1303   Mean   :2.855              <NA> Mean   :1.906
## X.4 3rd Qu.:1952   3rd Qu.:3.000              <NA> 3rd Qu.:2.000
## X.5 Max.   :2602   Max.   :8.000              <NA> Max.   :6.000
## X.6           <NA>            <NA>            <NA>            <NA>
##             BaederKG        Baujahr           EG_qm        Garage_qm
## X   Min.   :0.0000   Min.   :1884   Min.   : 28.00   Min.   :  0.00
## X.1 1st Qu.:0.0000   1st Qu.:1964   1st Qu.: 73.00   1st Qu.: 26.00
## X.2 Median :0.0000   Median :1982   Median : 89.00   Median : 39.00
## X.3 Mean   :0.4917   Mean   :1981   Mean   : 95.87   Mean   : 38.29
## X.4 3rd Qu.:1.0000   3rd Qu.:2009   3rd Qu.:114.00   3rd Qu.: 48.00
## X.5 Max.   :3.0000   Max.   :2022   Max.   :324.00   Max.   :126.00
## X.6             <NA>            <NA>             <NA>             <NA>
##             Garagen Gesamteindruck    Keller_Typ_qm        Keller_qm
## X   Min.   :0.000         1:   4   Min.   :  0.00   Min.   :  0.00
## X.1 1st Qu.:1.000         2:  50   1st Qu.:  0.00   1st Qu.: 66.00
## X.2 Median :2.000         3:1328   Median : 32.00   Median : 82.00
## X.3 Mean   :1.707         4: 796   Mean   : 36.77   Mean   : 86.25
## X.4 3rd Qu.:2.000         5: 163   3rd Qu.: 61.00   3rd Qu.:105.00
## X.5 Max.   :5.000             <NA> Max.   :194.00   Max.   :272.00
## X.6            <NA>            <NA>             <NA>             <NA>
##              Kellerhoehe                Kellertyp             Lage
## X   Durchschnitt :1101   Freizeitraum      :242   Bezirk 15: 385
## X.1 Gut          : 954   Guter Wohnraum    :658   Bezirk 6 : 208
## X.2 Kein Keller  :  63   Kein Keller       : 63   Bezirk 20: 199
## X.3 Schlecht     :  74   Kein Wohnraum     :234   Bezirk 8 : 155
## X.4 Sehr gut     : 147   Mittlerer Wohnraum:372   Bezirk 22: 134
## X.5 Sehr Schlecht:   2   Niedrige Qualität :132   Bezirk 9 : 120
## X.6                 <NA> Rohbau            :640   (Other)  :1140
##                OG_qm       Umgebaut   Verkaufsjahr    Verkaufsmonat
## X   Min.   :  0.00   Min.   :1962   Min.   :2018   Min.   : 1.000
## X.1 1st Qu.:  0.00   1st Qu.:1976   1st Qu.:2019   1st Qu.: 4.000
## X.2 Median :  0.00   Median :2002   Median :2020   Median : 6.000
## X.3 Mean   : 28.52   Mean   :1995   Mean   :2020   Mean   : 6.128
## X.4 3rd Qu.: 59.00   3rd Qu.:2014   3rd Qu.:2021   3rd Qu.: 8.000
## X.5 Max.   :175.00   Max.   :2022   Max.   :2022   Max.   :12.000
## X.6             <NA>            <NA>            <NA>             <NA>
##     Wohnflaeche_qm  Z_Verkaufspreis
## X    Min.   : 28   Min.   : 13100
## X.1  1st Qu.: 93   1st Qu.:130000
## X.2  Median :121   Median :162000
## X.3  Mean   :125   Mean   :180462
## X.4  3rd Qu.:146   3rd Qu.:213899
## X.5  Max.   :380   Max.   :755000
## X.6           <NA>             <NA>
```

Die Wertebereiche der numerishen Features weisen keine Auffälligkeiten auf. Aufgrund dieser Auswertung können noch keine Fehler in den Daten erkannt werden. Wenn man die Features betrachtet , die Daten über die beschaffenheit der Zimmer beschreibt (AnzahlZimmer, Garagen, Baeder, BaederKG) starten diese bei 0 und beinhalten keine nicht Integer Werte. Betrachten wir die Zeitangaben (Baujahr, Verkaufsjahr, Verkaufsmonat) sind diese Integer Werte. Aus dem Wertebereich der Jahrszahlen können auch keine unplausiblen Outlier erkannt werden. Die Angabe des Monats beinhaltet nur die zugelassenen 0-12 Integerwerte.

```
# Entfernen von Immobilien mit ungültigem Baujahr
df_immobilien <- df_immobilien[df_immobilien["Baujahr"] <= df_immobilien["Umgebaut"], ]
# QM bei nicht bestehendem OG
df_immobilien <- df_immobilien[(df_immobilien["OG_qm"] > 0) & ((df_immobilien["Ausbaustufe"] == "2 Eben
en") | (df_immobilien["Ausbaustufe"] == "3 Ebenen")) | (df_immobilien["OG_qm"] == 0) & (df_immobilien["
Ausbaustufe"] == "1 Ebene"), ]
# Bestehende Garage mit keinen QM
df_immobilien <- df_immobilien[(df_immobilien["Garage_qm"] > 0) & (df_immobilien["Garagen"] > 0), ]
# kein bestehender Keller trotz befüllter Keller Features
df_immobilien <- df_immobilien[((df_immobilien["Kellerhoehe"] == "Kein  Keller") & ((df_immobilien["Kel
ler_Typ_qm"] == 0) | (df_immobilien["Keller_qm"] == 0) |(df_immobilien["Kellertyp"] == "Kein Keller") |
(df_immobilien["BaederKG"] == 0)) | (df_immobilien["Kellerhoehe"] != "Kein Keller")), ]
```

Wir entfernen logisch nicht plausible Werte aus dem Datensatz mithilfe von subsetting. Die logischen Schlüsse werden im Anschluss erklärt:

1. Die Immobilie wurde renoviert bevor diese erbaut wurden
2. Die Immobilie besitzt kein Obergeschoss, jedoch wird eine qm Anzahl angegeben oder die Immobilie besitzt ein Obergeschoss und es ist keine qm Anzahl angegeben.
3. Die Immobilie besitzt eine Garage, diese besitzt jedoch keine qm Anzahl.
4. Es ist kein Keller vorhanden, jedoch sind die Features des Kellers befüllt.

Analyse der Daten

Im folgenden Block werden die Daten visualisiert und analysiert. Im Anschluss wird ein Fazit gezogen.

In welchen Bezirken erlebt der Kaufpreis der Immobilien das größte Wachstum?

Zunächst wird die Compound Annual Growth Rate für die Bezirke berechnet. Dies soll Aufschluss über die Preisentwicklung der Bezirke über die vier abgebildeten Verkaufsjahre geben.

```
# Berechnen der Compound annual growth rate
max_year <- max(df_immobilien$Verkaufsjahr)
min_year <- min(df_immobilien$Verkaufsjahr)
n = max_year - min_year
# group mean per year
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

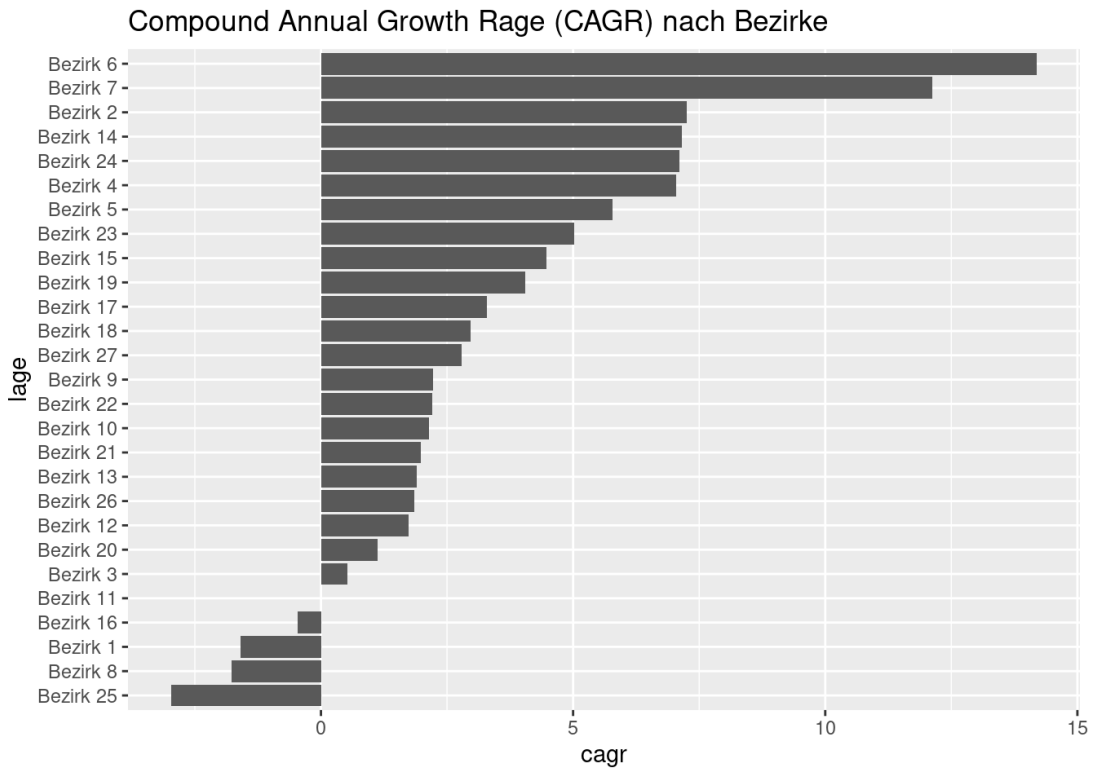
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_immobilien_grouped_lage <- df_immobilien %>% group_by(Lage, Verkaufsjahr) %>% summarise(mean_year=me
an(Z_Verkaufspreis))
```

```
## `summarise()` has grouped output by 'Lage'. You can override using the
## `.groups` argument.
```

```
cagr <- c()
for (bezirk in levels(df_immobilien_grouped_lage$Lage)) {
  min_year = min(df_immobilien_grouped_lage[df_immobilien_grouped_lage["Lage"] == bezirk, ]$Verkaufsjah
r)
  max_year = max(df_immobilien_grouped_lage[df_immobilien_grouped_lage["Lage"] == bezirk, ]$Verkaufsjah
r)
  growth <- ((df_immobilien_grouped_lage[(df_immobilien_grouped_lage["Lage"] == bezirk & df_immobilien_
grouped_lage["Verkaufsjahr"] == max_year), ]$mean_year / df_immobilien_grouped_lage[(df_immobilien_grou
ped_lage["Lage"] == bezirk & df_immobilien_grouped_lage["Verkaufsjahr"] == min_year), ]$mean_year)**(1
/(max_year-min_year))-1)*100
  cagr <- append(cagr, growth)
}
lage <- levels(df_immobilien_grouped_lage$Lage)
df_cagr_lage <- tibble(lage, cagr)
```

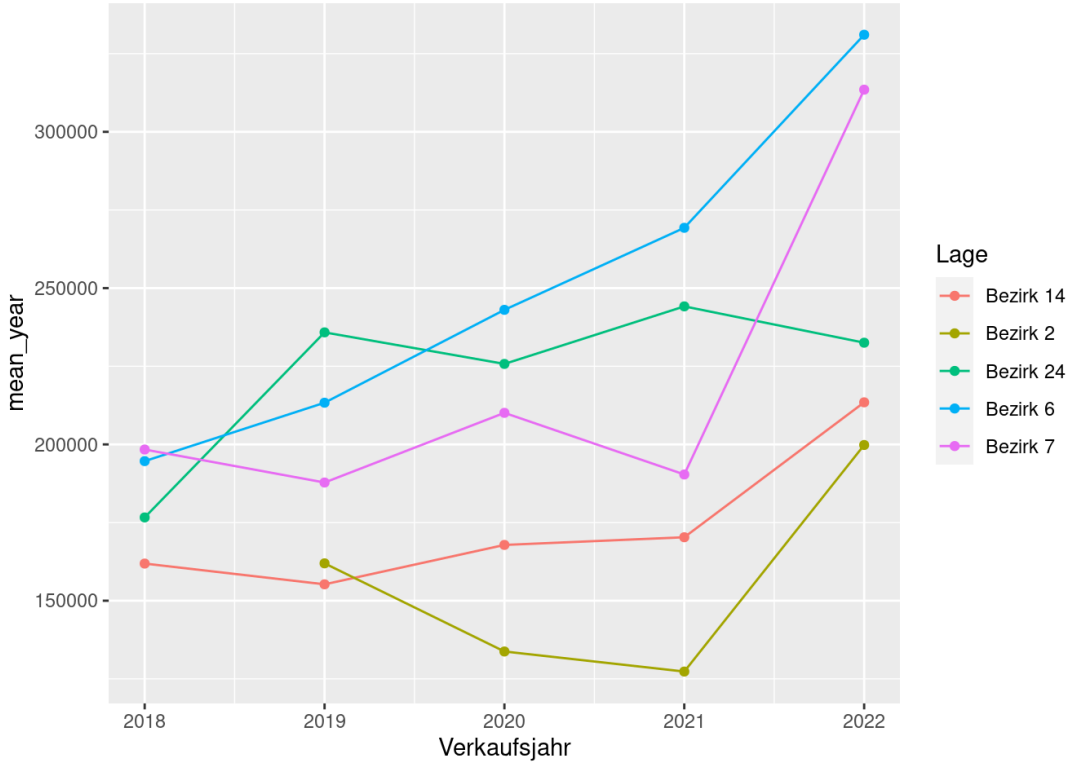
```
library(ggplot2)
p <- ggplot(df_cagr_lage, aes(reorder(lage, cagr), cagr))+geom_col( width = 0.9)+labs(x="lage",title="C
ompound Annual Growth Rage (CAGR) nach Bezirke")
p <- p + coord_flip()
p
```



Der vorangehende Barplot zeigt die CAGR nach Bezirken pro Jahr. Hier kann man sehen das die Bezirke sechs und sieben überdurchschnittliches Wachstum erfahren haben. Die Bezirke 1, 8 und 25 dagegen ngatives Wachstum erfahren haben. Dem Wachstum entsprechend empfielt es sich Immobilien aus den Top 5 aus dieser Grafik zu kaufen. Um weitere Erkenntnis über die Preisentwicklung zu bekommen wird in der nachfolgenden die Entwicklung der Top 5 Wachstumsstärksten Bezirke über die vier Verkaufsjahre betrachtet.

```
df_cagr_lage <- arrange(df_cagr_lage, cagr)

p <- ggplot(data=filter(df_immobilien_grouped_lage, Lage %in% tail(df_cagr_lage, 5)$lage), aes(x=Verkau
fsjahr, y=mean_year, group=Lage, color=Lage))+geom_point()+geom_line()
p
```

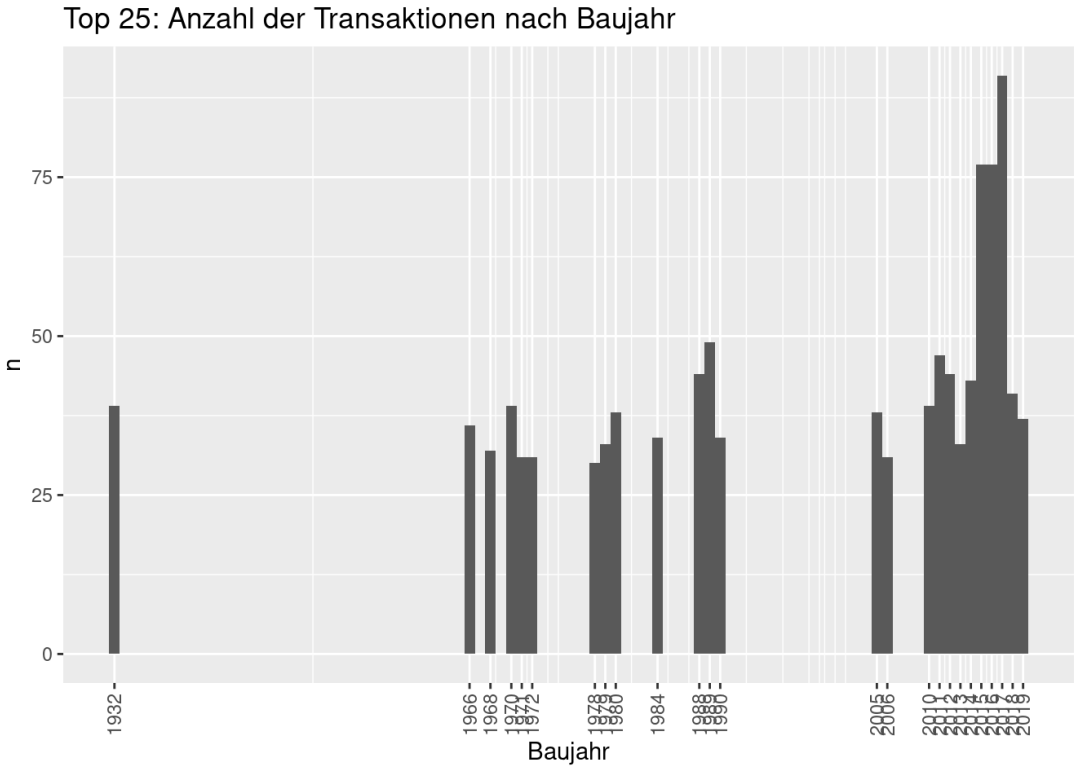


Betrachtet man den durchschnittlichen Verkaufspreis der Top 5 wachstumsstärksten Bezirke über die vier Verkaufsjahre, erkennt man das Immobilien aus den Bezirken sechs und vierundzwanzig im Durchschnitt die höchsten Verkaufspreise erzielen. Dabei hat der durchschnittliche Verkaufspreis in Bezirk sieben ein enormes Wachstum im Jahr 2022 erfahren. Die Bezirke sechs und vierzehn erfuhren stetiges Wachstum pber die abgebildeten Jahre und sind somit eine sicherere Investition als beispielsweise die Bezirke zwei und sieben, bei welchen der Verkaufspreis stark schwankt.

Somit kann man je nach Investorprofil entweder in die stabilen Bezirke sechs, vierzehn und vierundzwanzig investieren oder bei Risikofreudigeren Investoren un die Bezirke zwei und sieben.

Aus welchem Baujahr sind Häuser besonders begehrt?

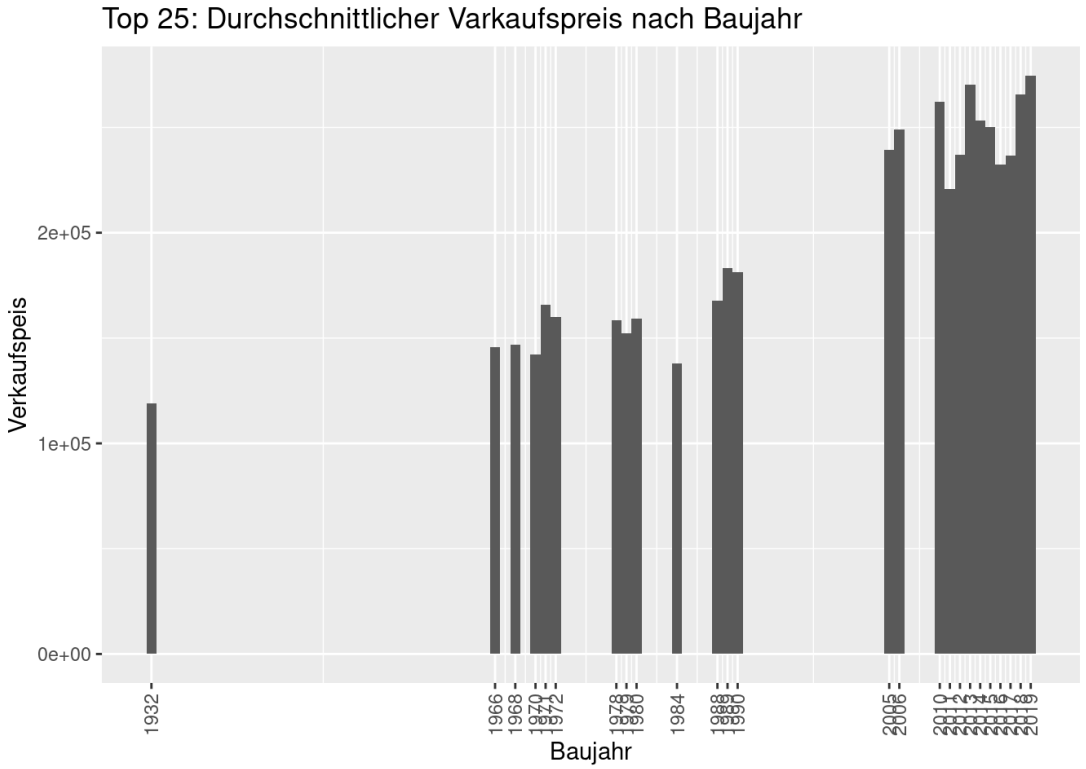
```
df_immobilien_grouped_Baujahr <- df_immobilien %>% group_by(Baujahr) %>% tally()
df_immobilien_grouped_Baujahr <- arrange(df_immobilien_grouped_Baujahr, n)
p <- ggplot(tail(df_immobilien_grouped_Baujahr, 25), aes(Baujahr, n))+geom_col( width = 1)+labs(x="Bauj
ahr",title="Top 25: Anzahl der Transaktionen nach Baujahr")+theme(axis.text.x = element_text(angle = 9
0, vjust = 0.5, hjust=1))+scale_x_continuous("Baujahr", labels = as.character(tail(df_immobilien_groupe
d_Baujahr, 25)$Baujahr), breaks = tail(df_immobilien_grouped_Baujahr, 25)$Baujahr)
p
```



Im vorangehenden Balkendiagramm werden die Top 25 Baujahre mit den meisten Transaktionen dargestellt. Hier kann man erkennen dass vor allem Häuser aus dem 21 Jahrhundert viele Transaktionen erfahren. Auch die Zeitperiode zwischen 1966 und 1990 erfahren viele Transaktionen. Daraus kann man auf die Beliebtheit der Gebäude schließen. Im folgenden Plot bertachten wir den durchschnittlichen verkaufspreis der Jahre.

```
df_immobilien_grouped_Baujahr_preis <- filter(df_immobilien, Baujahr %in% tail(df_immobilien_grouped_Ba
ujahr, 25)$Baujahr) %>% group_by(Baujahr) %>% summarise(mean_Baujahr=mean(Z_Verkaufspreis))
```

```
p <- ggplot(df_immobilien_grouped_Baujahr_preis, aes(Baujahr, mean_Baujahr))+geom_col( width = 1)+labs
(x="Baujahr", y="Verkaufspeis",title="Top 25: Durchschnittlicher Varkaufspreis nach Baujahr")+theme(axi
s.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+scale_x_continuous("Baujahr", labels = as.ch
aracter(tail(df_immobilien_grouped_Baujahr_preis, 25)$Baujahr), breaks = tail(df_immobilien_grouped_Bau
jahr_preis, 25)$Baujahr)
p
```



Betrachten wir die im Vorigen erläuterten Top 25 Baujahre erkennen wir das neuere Immobilien generell einen höheren Verkaufspreis erzielen als alte Immobilien.

Somit sollte in Immobilien aus dem 21 Jahrhundert investiert werden, da diese die höchsten Verkaufspreise und die meisten Transaktionen aufweisen. Von dieser Empfehlung sind Häuser, die interessant für Liebhaber sind. Im diesem Falle ist diese empfelung nichtmehr anwendbar.

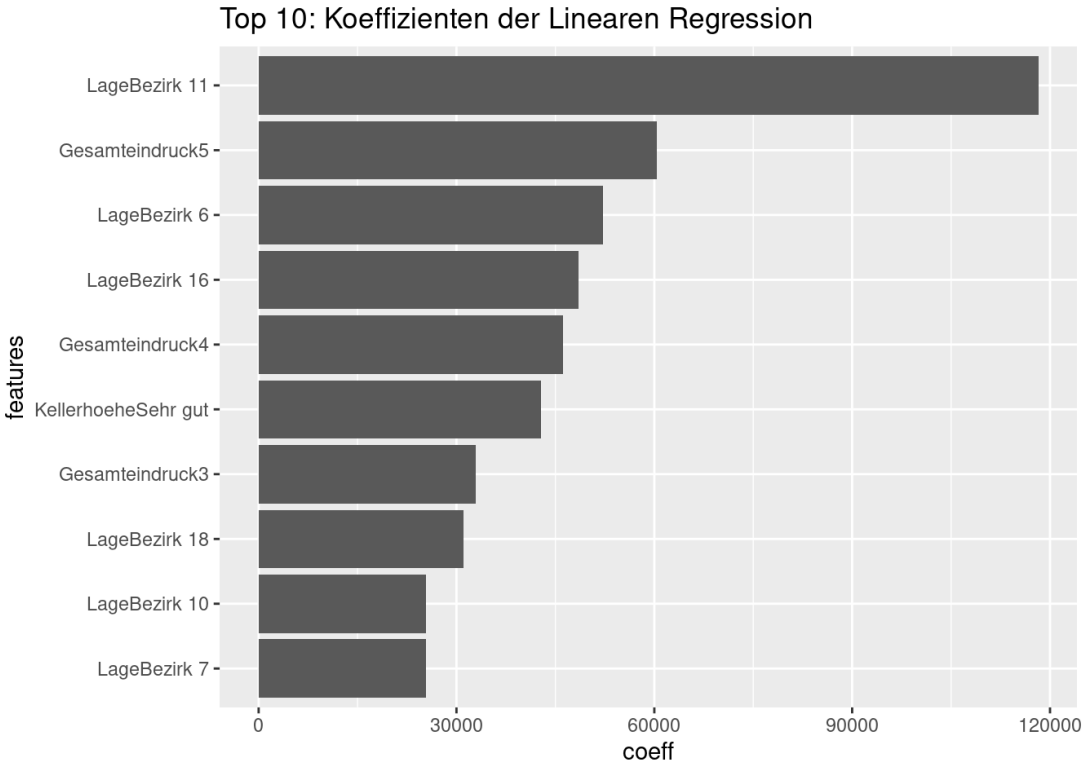
Lineare Regression des Datensatzes

Im folgenden wird eine Lineare Regression durchgeführt um Features festzustellen, welche den größten Einfluss auf den Kaufpreis des Hauses haben.

```
library(readxl)
num_cols <- unlist(lapply(df_immobilien, is.numeric))
df_immobilien_numeric <- df_immobilien[ , num_cols]
features <- colnames(df_immobilien_numeric)
target <- c("Z_Verkaufspreis")
features <- features[! features %in% target]

linearmodel <- lm(Z_Verkaufspreis~AnzahlZimmer+Ausbaustufe+Baeder+BaederKG+Baujahr+EG_qm+Garage_qm+Gara
gen+Gesamteindruck+Keller_Typ_qm+Keller_qm+Kellerhoehe+Kellertyp+Lage+OG_qm+Umgebaut+Verkaufsjahr+Verka
ufsmonat+Wohnflaeche_qm, data = df_immobilien)
```

```
features <- c("Intercept", "AnzahlZimmer", "Ausbaustufe2 Ebenen", "Ausbaustufe3 Ebenen","Baeder","Baede
rKG","Baujahr","EG_qm","Garage_qm","Garagen","Gesamteindruck2","Gesamteindruck3",
"Gesamteindruck4","Gesamteindruck5","Keller_Typ_qm","Keller_qm","KellerhoeheGut","KellerhoeheSchlech
t","KellerhoeheSehr gut","KellerhoeheSehr Schlecht","KellertypGuter Wohnraum","KellertypKein Wohnrau
m","KellertypMittlerer Wohnraum","KellertypNiedrige Qualität","KellertypRohbau","LageBezirk 10","LageBe
zirk 11","LageBezirk 12","LageBezirk 13","LageBezirk 14","LageBezirk 15","LageBezirk 16","LageBezirk 1
7","LageBezirk 18","LageBezirk 19","LageBezirk 2","LageBezirk 20","LageBezirk 21","LageBezirk 22","Lage
Bezirk 23","LageBezirk 24","LageBezirk 25","LageBezirk 26","LageBezirk 27","LageBezirk 3","LageBezirk
4","LageBezirk 5","LageBezirk 6","LageBezirk 7","LageBezirk 8","LageBezirk 9 ","OG_qm","Umgebaut","Verk
aufsjahr","Verkaufsmonat","Wohnflaeche_qm")
coeff <- summary(linearmodel)$coefficients[, "Estimate"]
df_coeff <- tibble(features, coeff)
df_coeff <- arrange(df_coeff, coeff)
p <- ggplot(tail(df_coeff, 10), aes(reorder(features, coeff), coeff))+geom_col( width = 0.9)+labs(x="fe
atures",title="Top 10: Koeffizienten der Linearen Regression")
p <- p + coord_flip()
p
```



Wir betrachten die Top 10 Features, welche den Größten Einfluss auf den Kaufpreis haben. Dabei stellt sich heraus, dass die Lage der Immobilie den größten Einfluss auf den Wert dieser hat. Von den Top 10 Features sind sechs Standorte (One Hot Encoded). Wenn eine Immobile im Bezirk 11 ist erhöht dies den Kaufpreis um etwa 118 Tausend Euro. Dies ist in sofern überraschend, da dieser Bezirk nicht unter den Top 5 Bezirke mit größtem Wachstum befinet. Zusätzlich erfuhr der Bezirk 11 überhaupt kein Wachstum innerhalb der vier Jahre.

Der nächstwichtigste Faktor ist der Gesamteindruck der Immobilie. Drei der zehn Features sind dabei One Hot encodete Gesamteindruck Features. Wenn die Immobilie den Gesamteindruck 5 (sehr gut) besitzt erhöht dies den Preis um etwa 60 Tausend Euro.

Das drittwichtigste Feature ist die Kellerhöhe. Eine Kellerhöhe von Sehr gut (250 cm) entspricht hierbei eine Preiserhöhung von etwa 42700 Euro.

Somit sollten beim Kauf einer Immobilie vor allem auf die Features der Lage, des Gesamteindrucks und der Kellerhöhe geachtet werden.

Zusammenfassung

Es hat sich ergeben das die Bezirke 2, 6, 7, 14 und 24 das größte Wachstum erfahren haben. Und wenn man den Trend dieser betrachtet, entwickelt sich der durchschnittliche Kaufpreis weiterhin nach oben. Somit sin Investitionen in diese Bezirke empfehlenswert.

Die Untersuchung der Baujahre hat ergeben das sich die Hypothese bestätigt hat und neue Häuser einen höheren Kaufpreis erzielen. Das Liebhaber-Phänomen ist aus den Daten nicht erkennbar und somit für die Kaufentscheidung irrelevant. Dies ist überraschend da die Vermutung aufgestellt wurde, dass alte Häuse im Wert steigen, da diese Attraktiv für Liebhaber sind.

Die Lineare Regression hat ergeben, dass im wesentliche die drei Features Lage, Gesamteindruck und Kellerhöhe den Größten Einfluss auf den Verkaufspreis haben. Demnach ist auf diese Features besonders Wert zu legen