Housing

Include libraries:

```
library(tidyverse)
library(ggplot2)
library(readr)
```

Introduction

The goal of this thesis is to study and predict housing prices (response variable) based on predictor variables that provide data on the physical characteristics of properties, such as their size, age, location, etc.

Exploratory data analysis:

Lets have a short look at our data set:

```
#read data in and have a look at it:
house <- read_csv("data/realestate.csv")
head(house)
```

```
## # A tibble: 6 x 12
##      ID  Price  Sqft Bedroom Bathroom Airconditioning Garage  Pool
##   <int>  <int> <int>   <int>    <int>           <int>  <int> <int>
## 1     1 360000  3032       4        4               1      2     0
## 2     2 340000  2058       4        2               1      2     0
## 3     3 250000  1780       4        3               1      2     0
## 4     4 205500  1638       4        2               1      2     0
## 5     5 275500  2196       4        3               1      2     0
## 6     6 248000  1966       4        3               1      5     1
## # ... with 4 more variables: YearBuild <int>, Quality <int>, Lot <int>,
## #   AdjHighway <int>
```

```
#we only work with "house":
attach(house)

#delete column "ID":
house <- house[,-1]

#dim of house:
dim(house)
```

```
## [1] 522  11
```

```
names(house)
```

```
##  [1] "Price"           "Sqft"            "Bedroom"
##  [4] "Bathroom"        "Airconditioning" "Garage"
##  [7] "Pool"            "YearBuild"       "Quality"
## [10] "Lot"             "AdjHighway"
```

```r
#is "Pool" binary or does t count the number of pools?
house %>% select(Pool) %>% distinct()
```

```
## # A tibble: 2 x 1
##    Pool
##   <int>
## 1     0
## 2     1
```

```r
#which values does "Quality" take?:
house %>% select(Quality) %>% distinct()
```

```
## # A tibble: 3 x 1
##   Quality
##     <int>
## 1       2
## 2       3
## 3       1
```

```r
#which values does "AdjHighway" take?:
house %>% select(AdjHighway) %>% distinct()
```

```
## # A tibble: 2 x 1
##   AdjHighway
##        <int>
## 1          0
## 2          1
```

```r
View(house)
```

We store our data set "realestat.csv" into a variable called house. The first insight is that the column "ID" is not needed, as it only stores the index of the houses. As a result, we can delete this column. The data set now contains 11 columns and 522 rows. The variables are as follows: "Price" (quantitative; the price of the property in question), "Sqft" (quantitative; size), "Bedroom" (quantitative; number of bedrooms), "Bathroom" (quantitative; number of bathrooms), "Airconditioning" (qualitative; yes: 1, no: 0), "Garage" (quantitative; number of garages (or parking spots)), "Pool" (qualitative; yes: 1, no: 0), "YearBuild" (qualitative; year in which the house was built), "Quality" (qualitative; the build-quality of the house on a scale of 1 to 3), "Lot" (quantitative, ??????????????????????????), and finally "AdjHighway"(qualitative; 1: next to a highway, 0: not close or next to a highway).

Resonse variable: