# HARVARD EXTENSION SCHOOL

EXT CSCI E-106 Model Data Class Group Project Template

Author Kaleo Pudim Author Luciano Carvalho Author Ibrahim Hashim
Author Bethun Bhowmik Author Mohanish Kashiwar Author Seymur Hasanov

23 November 2023

**Abstract**

This is the location for your abstract. It must consist of two paragraphs.

## Contents

## House Sales in King County, USA data to be used in the Final Project

| Variable | Description |
|---|---|
| id | **Unique ID for each home sold (it is not a predictor)** |
| date | *Date of the home sale* |
| price | *Price of each home sold* |
| bedrooms | *Number of bedrooms* |
| bathrooms | *Number of bathrooms, where ".5" accounts for a bathroom with a toilet but no shower* |
| sqft_living | *Square footage of the apartment interior living space* |
| sqft_lot | *Square footage of the land space* |
| floors | *Number of floors* |
| waterfront | *A dummy variable for whether the apartment was overlooking the waterfront or not* |
| view | *An index from 0 to 4 of how good the view of the property was* |
| condition | *An index from 1 to 5 on the condition of the apartment,* |
| grade | *An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 has a high-quality level of construction and design.* |
| sqft_above | *The square footage of the interior housing space that is above ground level* |
| sqft_basement | *The square footage of the interior housing space that is below ground level* |
| yr_built | *The year the house was initially built* |
| yr_renovated | *The year of the house's last renovation* |
| zipcode | *What zipcode area the house is in* |
| lat | *Latitude* |
| long | *Longitude* |
| sqft_living15 | *The square footage of interior housing living space for the nearest 15 neighbors* |
| sqft_lot15 | *The square footage of the land lots of the nearest 15 neighbors* |

## Instructions:

0. Join a team with your fellow students with appropriate size (Four Students total)
1. Load and Review the dataset named "KC_House_Sales'csv
2. Create the train data set which contains 70% of the data and use set.seed (1023). The remaining 30% will be your test data set.
3. Investigate the data and combine the level of categorical variables if needed and drop variables as needed. For example, you can drop id, Latitude, Longitude, etc.
4. Build a regression model to predict price.
5. Create scatter plots and a correlation matrix for the train data set. Interpret the possible relationship between the response.
6. Build the best multiple linear models by using the stepwise selection method. Compare the performance of the best two linear models.
7. Make sure that model assumption(s) are checked for the final model. Apply remedy measures (transformation, etc.) that helps satisfy the assumptions.
8. Investigate unequal variances and multicollinearity. If necessary, apply remedial methods (WLS, Ridge, Elastic Net, Lasso, etc.).
9. Build an alternative model based on one of the following approaches to predict price: regression tree, NN, or SVM. Check the applicable model assumptions. Explore using a logistic regression.
10. Use the test data set to assess the model performances from above.
11. Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model.
12. Create a model development document that describes the model following this template, input the name of the authors, Harvard IDs, the name of the Group, all of your code and calculations, etc..:

## Due Date: December 18th, 2023 at 11:59 pm EST

**Notes No typographical errors, grammar mistakes, or misspelled words, use English language All tables need to be numbered and describe their content in the body of the document All figures/graphs need to be numbered and describe their content All results must be accurate and clearly explained for a casual reviewer to fully understand their purpose and impact Submit both the RMD markdown file and PDF with the sections with appropriate explanations. A more formal document in Word can be used in place of the pdf file but must include all appropriate explanations.**

Executive Summary

This section will describe the model usage, your conclusions and any regulatory and internal requirements. In a real world scneario, this section is for senior management who do not need to know the details. They need to know high level (the purpose of the model, limitations of the model and any issues).

## I. Introduction (5 points)

*This section needs to introduce the reader to the problem to be resolved, the purpose, and the scope of the statistical testing applied. What you are doing with your prediction? What is the purpose of the model? What methods were trained on the data, how large is the test sample, and how did you build the model?*

— in-progress —

This project aims to develop a predictive model for house prices in King County, USA. The model's purpose is to provide an analytical tool for understanding the key factors influencing property values in this area, which is significant for various stakeholders in the real estate sector.

Our primary data source is the 'KC_House_Sales' dataset, which includes detailed information on various house attributes. **(Add here the hypothesize of initial exploration on house prices)**.

The methodology adopted for this project combines conventional statistical techniques with modern data analytics methods. Initial models will be based on linear regression, with further exploration into more complex approaches like neural networks and decision trees, depending on their performance in preliminary analyses.

To ensure robust model training and validation, the dataset has been divided into a training set (70%) and a test set (30%). This division is crucial for evaluating the model's effectiveness in making predictions on new, unseen data.

## I. Description of the data and quality (15 points)

*Here you need to review your data, the statistical test applied to understand the predictors and the response and how are they correlated. Extensive graph analysis is recommended. Is the data continuous, or categorical, do any transformation needed? Do you need dummies?*

### Data Overview

```
df_house = read.csv("KC_House_Sales.csv")

head(df_house)
```

```
##           id            date           price bedrooms bathrooms sqft_living
## 1 7129300520 20141013T000000    $221,900.00         3      1.00        1180
## 2 6414100192 20141209T000000    $538,000.00         3      2.25        2570
## 3 5631500400 20150225T000000    $180,000.00         2      1.00         770
## 4 2487200875 20141209T000000    $604,000.00         4      3.00        1960
## 5 1954400510 20150218T000000    $510,000.00         3      2.00        1680
## 6 7237550310 20140512T000000  $1,225,000.00         4      4.50        5420
##   sqft_lot floors waterfront view condition grade sqft_above sqft_basement
## 1     5650      1          0    0         3     7       1180             0
## 2     7242      2          0    0         3     7       2170           400
## 3    10000      1          0    0         3     6        770             0
## 4     5000      1          0    0         5     7       1050           910
## 5     8080      1          0    0         3     8       1680             0
## 6   101930      1          0    0         3    11       3890          1530
##   yr_built yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1     1955            0   98178 47.5112 -122.257          1340       5650
## 2     1951         1991   98125 47.7210 -122.319          1690       7639
## 3     1933            0   98028 47.7379 -122.233          2720       8062
## 4     1965            0   98136 47.5208 -122.393          1360       5000
## 5     1987            0   98074 47.6168 -122.045          1800       7503
## 6     2001            0   98053 47.6561 -122.005          4760     101930
```

### Data Types, Categories and Cleaning

```
#removing `id` column
df_house = subset(df_house, select = -id)

#converting `price` to numeric
df_house$price <- as.numeric(gsub("[\\$,]", "", df_house$price))

##cleaning `date` and adding `year`,`month`,`day` columns
df_house$date <- as.POSIXct(df_house$date, format = "%Y%m%d")
df_house$year <- as.numeric(format(df_house$date, "%Y"))
df_house$month <- as.numeric(format(df_house$date, "%m"))
df_house$day <- as.numeric(format(df_house$date, "%d"))

head(df_house)
```

```
##         date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront
## 1 2014-10-13  221900        3      1.00        1180     5650      1          0
## 2 2014-12-09  538000        3      2.25        2570     7242      2          0
## 3 2015-02-25  180000        2      1.00         770    10000      1          0
## 4 2014-12-09  604000        4      3.00        1960     5000      1          0
```

```
## 5 2015-02-18  510000        3      2.00        1680     8080      1          0
## 6 2014-05-12 1225000        4      4.50        5420   101930      1          0
##   view condition grade sqft_above sqft_basement yr_built yr_renovated zipcode
## 1    0         3     7       1180             0     1955            0   98178
## 2    0         3     7       2170           400     1951         1991   98125
## 3    0         3     6        770             0     1933            0   98028
## 4    0         5     7       1050           910     1965            0   98136
## 5    0         3     8       1680             0     1987            0   98074
## 6    0         3    11       3890          1530     2001            0   98053
##       lat     long sqft_living15 sqft_lot15 year month day
## 1 47.5112 -122.257          1340       5650 2014    10  13
## 2 47.7210 -122.319          1690       7639 2014    12   9
## 3 47.7379 -122.233          2720       8062 2015     2  25
## 4 47.5208 -122.393          1360       5000 2014    12   9
## 5 47.6168 -122.045          1800       7503 2015     2  18
## 6 47.6561 -122.005          4760     101930 2014     5  12
```

```
#converting all to numeric...ignores date column
ndf_house = df_house[sapply(df_house, is.numeric)]

head(ndf_house)
```

```
##     price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
## 1  221900        3      1.00        1180     5650      1          0    0
## 2  538000        3      2.25        2570     7242      2          0    0
## 3  180000        2      1.00         770    10000      1          0    0
## 4  604000        4      3.00        1960     5000      1          0    0
## 5  510000        3      2.00        1680     8080      1          0    0
## 6 1225000        4      4.50        5420   101930      1          0    0
##   condition grade sqft_above sqft_basement yr_built yr_renovated zipcode
## 1         3     7       1180             0     1955            0   98178
## 2         3     7       2170           400     1951         1991   98125
## 3         3     6        770             0     1933            0   98028
## 4         5     7       1050           910     1965            0   98136
## 5         3     8       1680             0     1987            0   98074
## 6         3    11       3890          1530     2001            0   98053
##       lat     long sqft_living15 sqft_lot15 year month day
## 1 47.5112 -122.257          1340       5650 2014    10  13
## 2 47.7210 -122.319          1690       7639 2014    12   9
## 3 47.7379 -122.233          2720       8062 2015     2  25
## 4 47.5208 -122.393          1360       5000 2014    12   9
## 5 47.6168 -122.045          1800       7503 2015     2  18
## 6 47.6561 -122.005          4760     101930 2014     5  12
```

**Stat Summary - in progress**

[Add here the initial analysis of the summary, if applicable]

```
str(ndf_house)
```

```
## 'data.frame':    21613 obs. of  22 variables:
##  $ price       : num  221900 538000 180000 604000 510000 ...
##  $ bedrooms    : int  3 3 2 4 3 4 3 3 3 3 ...
##  $ bathrooms   : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot    : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
##  $ floors      : num  1 2 1 1 1 1 2 1 1 2 ...
##  $ waterfront  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
##  $ year         : num  2014 2014 2015 2014 2015 ...
##  $ month        : num  10 12 2 12 2 5 6 1 4 3 ...
##  $ day          : num  13 9 25 9 18 12 27 15 15 12 ...
```

```
summary(ndf_house)
```

```
##      price            bedrooms        bathrooms      sqft_living
##  Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
##  1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
##  Median : 450000   Median : 3.000   Median :2.250   Median : 1910
##  Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080
##  3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
##  Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540
##     sqft_lot           floors        waterfront            view
##  Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :   7618   Median :1.500   Median :0.000000   Median :0.0000
##  Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343
##  3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
##    condition         grade          sqft_above   sqft_basement
##  Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0
##  1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0
##  Median :3.000   Median : 7.000   Median :1560   Median :   0.0
##  Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5
##  3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0
##  Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0
##     yr_built     yr_renovated       zipcode           lat
##  Min.   :1900   Min.   :   0.0   Min.   :98001   Min.   :47.16
##  1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47
##  Median :1975   Median :   0.0   Median :98065   Median :47.57
##  Mean   :1971   Mean   :  84.4   Mean   :98078   Mean   :47.56
##  3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68
##  Max.   :2015   Max.   :2015.0   Max.   :98199   Max.   :47.78
##       long        sqft_living15    sqft_lot15          year
##  Min.   :-122.5   Min.   : 399   Min.   :   651   Min.   :2014
##  1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100   1st Qu.:2014
##  Median :-122.2   Median :1840   Median :  7620   Median :2014
##  Mean   :-122.2   Mean   :1987   Mean   : 12768   Mean   :2014
##  3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083   3rd Qu.:2015
##  Max.   :-121.3   Max.   :6210   Max.   :871200   Max.   :2015
##      month             day
##  Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 4.000   1st Qu.: 8.00
##  Median : 6.000   Median :16.00
##  Mean   : 6.574   Mean   :15.69
##  3rd Qu.: 9.000   3rd Qu.:23.00
##  Max.   :12.000   Max.   :31.00
```

```r
summary(ndf_house$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75000  321950  450000  540088  645000 7700000
```

### Graphical Analysis - in progress - Seymur

*[Add here the initial analysis]*

```r
library(GGally)
```

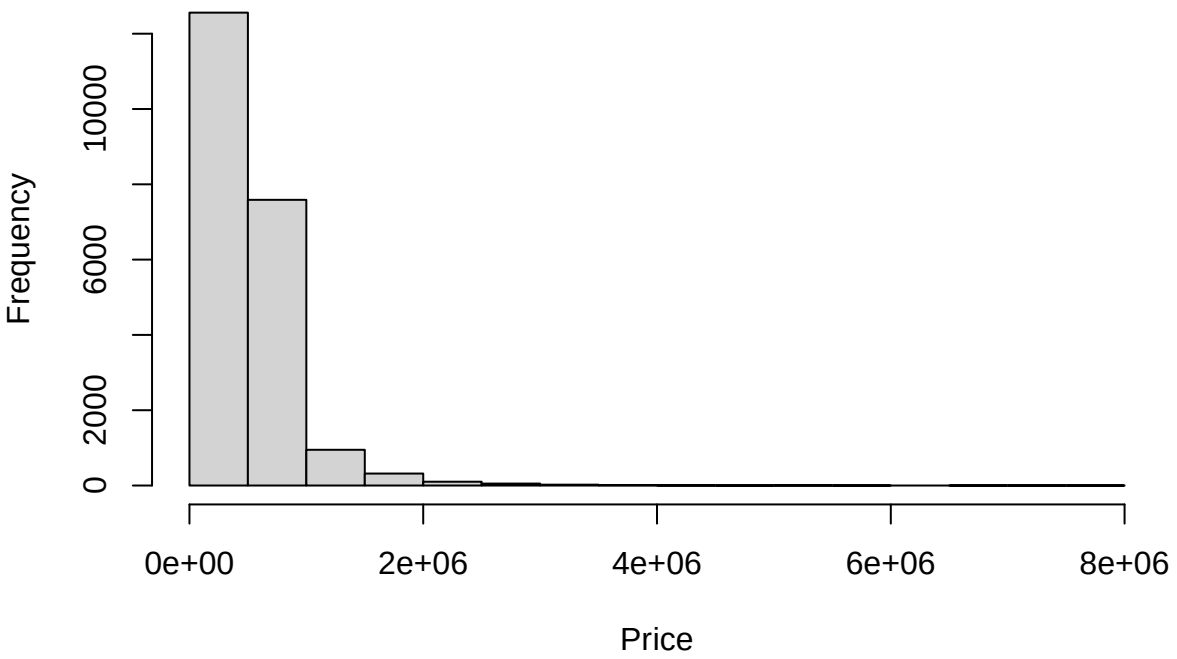
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
# Pairwise scatter plot with correlation coefficients
ggpairs(ndf_house, columns = c("price", "sqft_living", "bedrooms", "bathrooms", "grade"),
        title = "Pairwise Scatter Plots with House Price")
```


*[Add here the initial analysis]*

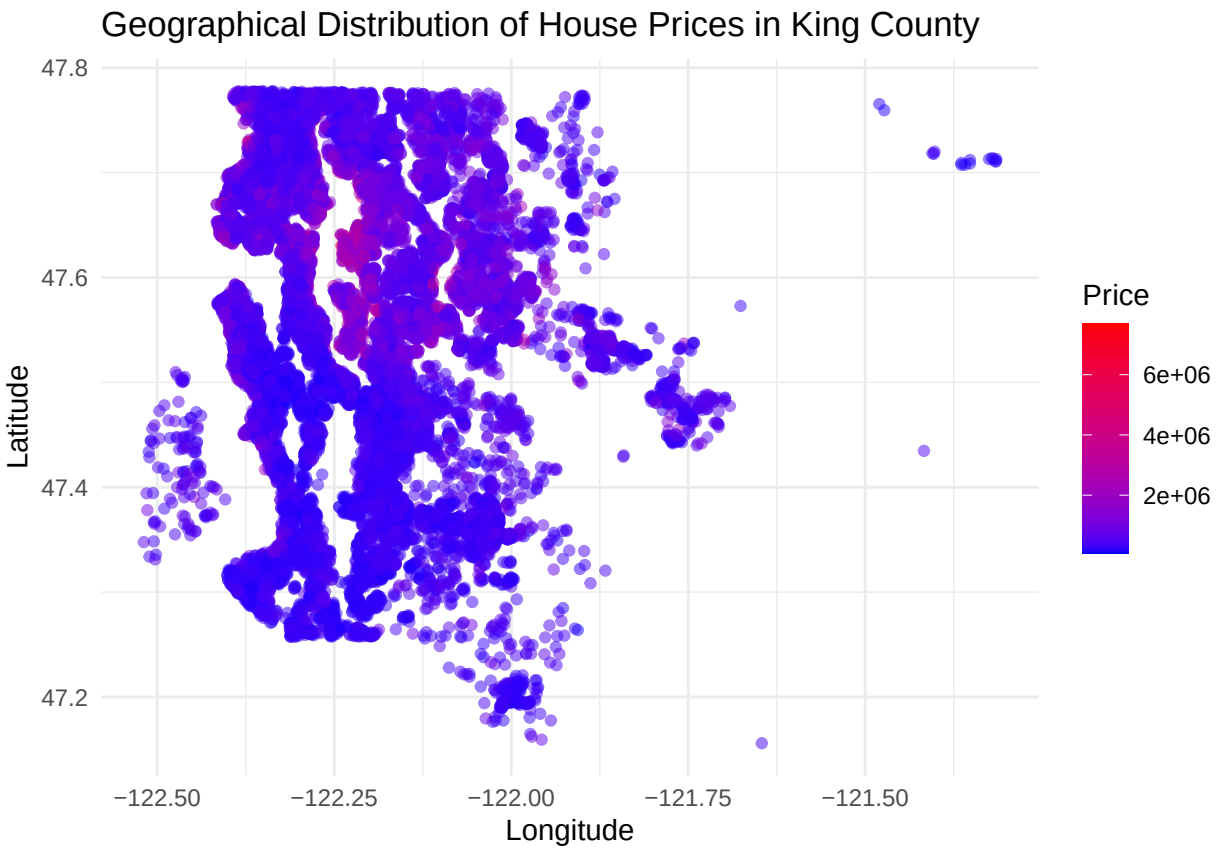
```r
hist(ndf_house$price, main = "Histogram of House Prices", xlab = "Price")
```




*[Add here the initial analysis]*

```
library(ggplot2)

# Scatter plot of properties
ggplot(data = ndf_house, aes(x = long, y = lat, color = price)) +
  geom_point(alpha = 0.5) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Geographical Distribution of House Prices in King County",
       x = "Longitude", y = "Latitude", color = "Price") +
  theme_minimal()
```




**Correlation Analysis - in progress**

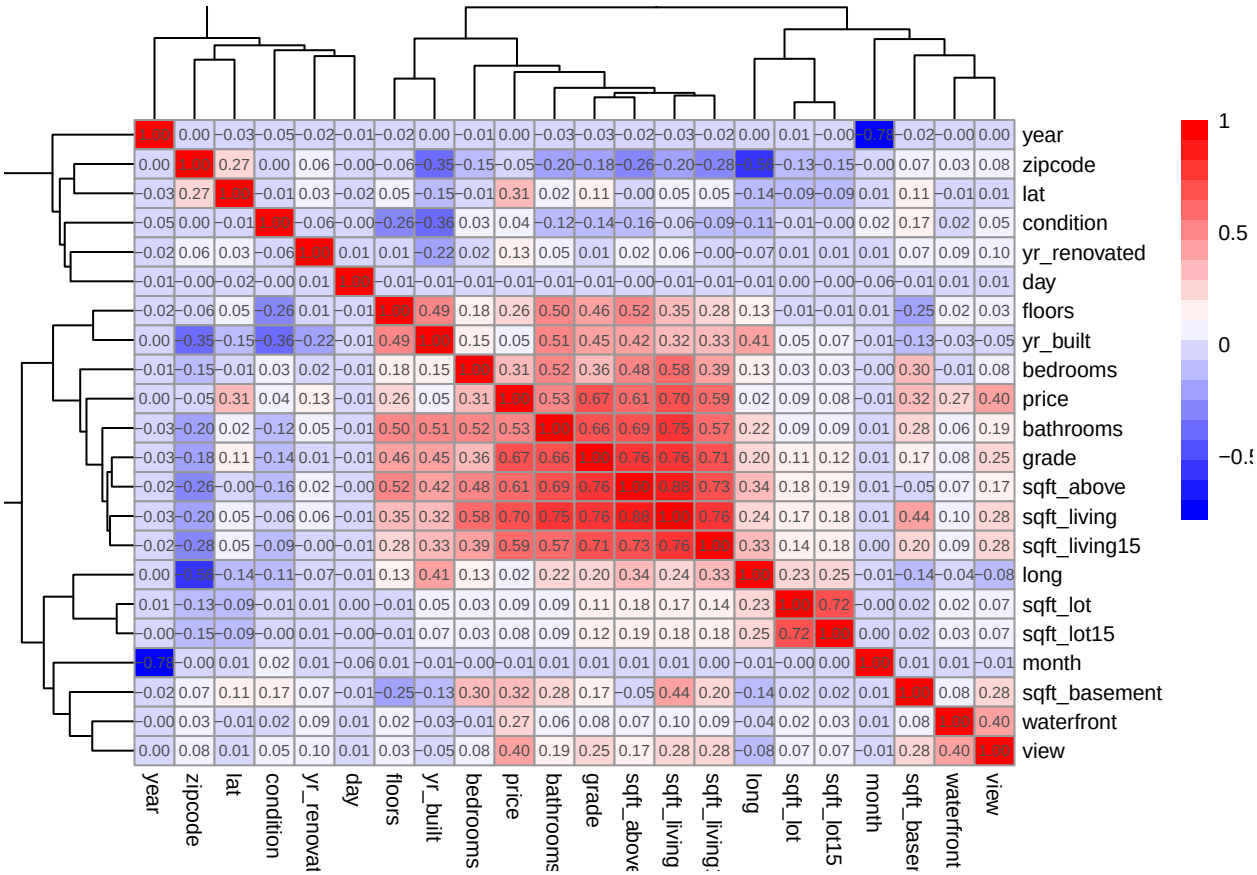
***[Add here the initial analysis of the correlation matrix]***

```
library(pheatmap)

#correlation matrix
cor_matrix = cor(ndf_house)
correlation_df = as.data.frame(cor_matrix)

#new dataframe with variables 'highly' (>0.2) correlated with price
x_high = subset(ndf_house, select = c(price,view,waterfront,sqft_basement,sqft_living,
                                      sqft_living15,sqft_above,grade,bathrooms,bedrooms,floors,lat))

pheatmap(cor_matrix,
         color = colorRampPalette(c("blue", "white", "red"))(20),
         main = "correlation matrix heatmap",
         fontsize = 8,
         cellwidth = 15,
         cellheight = 11,
         display_numbers = TRUE
)
```




**Creating a new dataframe with variables 'highly' (>0.2) correlated with price**

```
head(x_high)
```

```
##     price view waterfront sqft_basement sqft_living sqft_living15 sqft_above
## 1  221900    0          0             0        1180          1340        1180
## 2  538000    0          0           400        2570          1690        2170
## 3  180000    0          0             0         770          2720         770
## 4  604000    0          0           910        1960          1360        1050
## 5  510000    0          0             0        1680          1800        1680
## 6 1225000    0          0          1530        5420          4760        3890
##   grade bathrooms bedrooms floors     lat
## 1     7      1.00        3      1 47.5112
## 2     7      2.25        3      2 47.7210
## 3     6      1.00        2      1 47.7379
## 4     7      3.00        4      1 47.5208
## 5     8      2.00        3      1 47.6168
## 6    11      4.50        4      1 47.6561
```

**Data Transformation**

**Dummy Variables**

**Summary of the intro section**

## III. Model Development Process (15 points)

*Build a regression model to predict price. And of course, create the train data set which contains 70% of the data and use set.seed (1023). The remaining 30% will be your test data set. Investigate the data and combine the level of categorical variables if needed and drop variables. For example, you can drop id, Latitude, Longitude, etc.*

## IV. Model Performance Testing (15 points)

*Use the test data set to assess the model performances. Here, build the best multiple linear models by using the stepwise both ways selection method. Compare the performance of the best two linear models. Make sure that model assumption(s) are checked for the final linear model. Apply remedy measures (transformation, etc.) that helps satisfy the assumptions. In particular you must deeply investigate unequal variances and multicollinearity. If necessary, apply remedial methods (WLS, Ridge, Elastic Net, Lasso, etc.).*

## V. Challenger Models (15 points)

*Build an alternative model based on one of the following approaches to predict price: regression tree, NN, or SVM. Explore using a logistic regression. Check the applicable model assumptions. Apply in-sample and out-of-sample testing, backtesting and review the comparative goodness of fit of the candidate models. Describe step by step your procedure to get to the best model and why you believe it is fit for purpose.*

## VI. Model Limitation and Assumptions (15 points)

*Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model. Validate your models using the test sample. Do the residuals look normal? Does it matter given your technique? How is the prediction performance using Pseudo R^2, SSE, RMSE? Benchmark the model against alternatives. How good is the relative fit? Are there any serious violations of the model assumptions? Has the model had issues or limitations that the user must know? (Which assumptions are needed to support the Champion model?)*

## VII. Ongoing Model Monitoring Plan (5 points)

*How would you picture the model needing to be monitored, which quantitative thresholds and triggers would you set to decide when the model needs to be replaced? What are the assumptions that the model must comply with for its continuous use?*

## VIII. Conclusion (5 points)

*Summarize your results here. What is the best model for the data and why?*

## Bibliography (7 points)

*Please include all references, articles and papers in this section.*

## Appendix (3 points)

*Please add any additional supporting graphs, plots and data analysis.*