home_data_for_ml_course

Presley Hunter

Introduction

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

#Loading Packaging

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(infer)
library(GGally)
```

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

Loading Dataset

```
HousePrice = read_csv("home1460.csv")
```

```
## Rows: 1452 Columns: 54
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (25): MSZoning, Street, LotShape, Utilities, LotConfig, LandSlope, Neigh...
## dbl (29): Id, MSSubClass, OverallQual, OverallCond, YearBuilt, YearRemodAdd,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#Exploring the Dataset

```r
glimpse(HousePrice)
```

```
## Rows: 1,452
## Columns: 54
## $ Id             <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ MSSubClass     <dbl> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, 20~
## $ MSZoning       <chr> "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RM", "~
## $ Street         <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pave",~
## $ LotShape       <chr> "Reg", "Reg", "IR1", "IR1", "IR1", "IR1", "Reg", "IR1",~
## $ Utilities      <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "AllP~
## $ LotConfig      <chr> "Inside", "FR2", "Inside", "Corner", "FR2", "Inside", "~
## $ LandSlope      <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl",~
## $ Neighborhood   <chr> "CollgCr", "Veenker", "CollgCr", "Crawfor", "NoRidge", ~
## $ Condition1     <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "Norm"~
## $ Condition2     <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "Norm",~
## $ BldgType       <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam",~
## $ HouseStyle     <chr> "2Story", "1Story", "2Story", "2Story", "2Story", "1.5F~
## $ OverallQual    <dbl> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4, 5~
## $ OverallCond    <dbl> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 5~
## $ YearBuilt      <dbl> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931, 1~
## $ YearRemodAdd   <dbl> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950, 1~
## $ RoofStyle      <chr> "Gable", "Gable", "Gable", "Gable", "Gable", "Gable", "~
## $ RoofMatl       <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg", ~
## $ Exterior1st    <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "VinylSd", ~
## $ Exterior2nd    <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Shng", "VinylSd", ~
## $ MasVnrType     <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace", "None"~
## $ MasVnrArea     <dbl> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 306,~
## $ ExterQual      <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA", "TA", "~
## $ ExterCond      <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "~
## $ Foundation     <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", "Wood", ~
## $ BsmtFinSF1     <dbl> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 9~
## $ BsmtFinSF2     <dbl> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CentralAir     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", ~
## $ `1stFlrSF`     <dbl> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077,~
## $ `2ndFlrSF`     <dbl> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142, 0~
## $ LowQualFinSF   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ GrLivArea      <dbl> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774, 1~
## $ BsmtFullBath   <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1~
## $ BsmtHalfBath   <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ FullBath       <dbl> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1~
## $ HalfBath       <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1~
## $ BedroomAbvGr   <dbl> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3~
## $ KitchenAbvGr   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1~
## $ KitchenQual    <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA", "TA", "~
## $ TotRmsAbvGrd   <dbl> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6, ~
## $ Functional     <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ",~
## $ Fireplaces     <dbl> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0~
## $ GarageCars     <dbl> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2~
## $ GarageArea     <dbl> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, ~
## $ PavedDrive     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", ~
## $ WoodDeckSF     <dbl> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 16~
## $ PoolArea       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ MiscVal        <dbl> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 700~
## $ MoSold         <dbl> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3, 1~
```

```
## $ YrSold         <dbl> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 2~
## $ SaleType       <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "~
## $ SaleCondition  <chr> "Normal", "Normal", "Normal", "Abnorml", "Normal", "Nor~
## $ SalePrice_1000 <dbl> 208.5, 181.5, 223.5, 140.0, 250.0, 143.0, 307.0, 200.0,~
```

```
anyNA(HousePrice)
```

```
## [1] FALSE
```

In my dataset there are 1452 rows and there are 54 variables. My data set does not contain missing values.

Research Question and Hypotheses

My first research question is: Does neighborhood, the year the house was built, and HouseStyle predict the SalePrice of the house?, and the second is: Does the 1st and 2nd floor square footage affect the price of the house? The variables the first question will involve will be: Neighborhood, YearBuilt, HouseStyle; with the variables of the second research question being: 1stFlrSF, 2ndFlrSF. The main response/target variable is SalePrice.

Exploratory Data Analysis

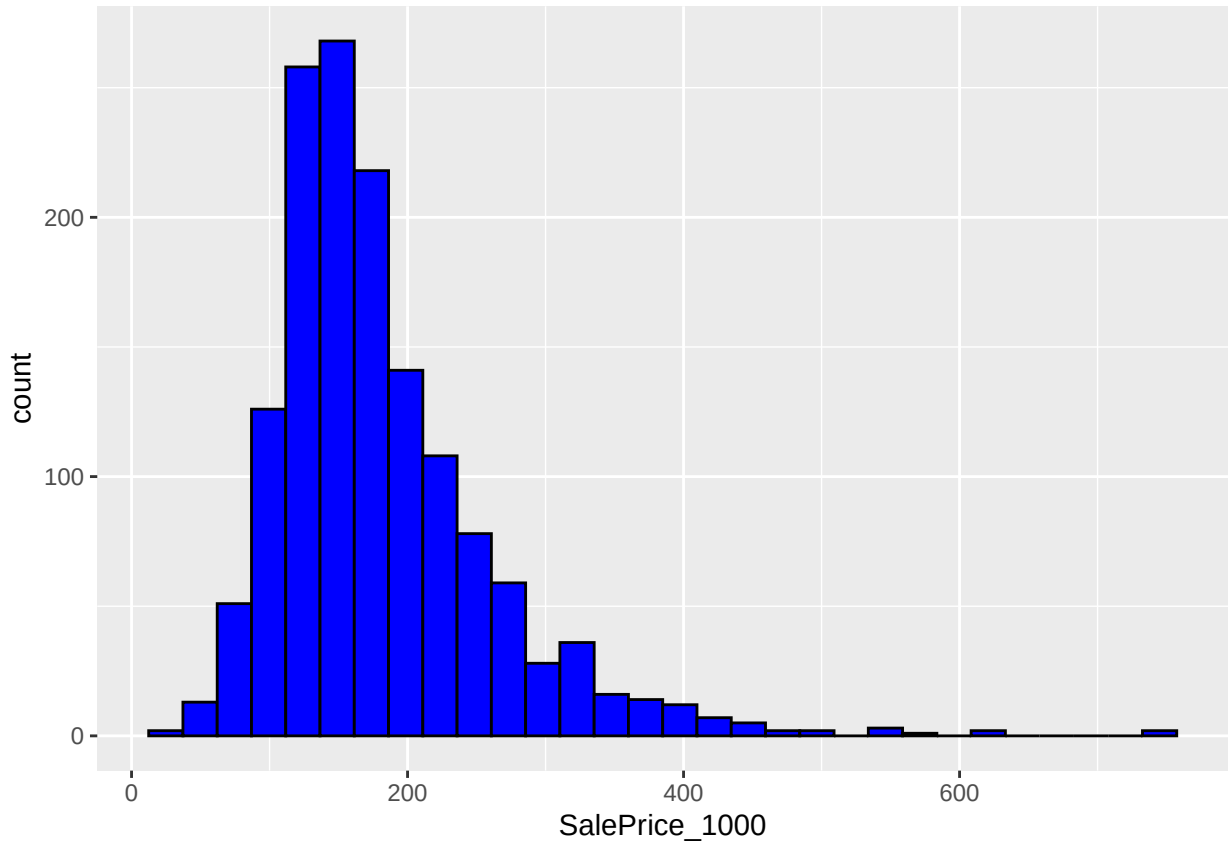
Here we compute and report summary statistics (e.g., mean, sd, and five number summary) for summarizing the distribution of the response variable:

```
HousePrice %>%
  summarise(Mean_Price = mean(SalePrice_1000),
            SD_Price = sd(SalePrice_1000),
            Min_Price = min(SalePrice_1000),
            Q1_Price = quantile(SalePrice_1000, .25),
            Median_Price = median(SalePrice_1000),
            Q3_Price = quantile(SalePrice_1000, .75),
            Max_Price = max(SalePrice_1000)
  )
```

```
## # A tibble: 1 x 7
##   Mean_Price SD_Price Min_Price Q1_Price Median_Price Q3_Price Max_Price
##        <dbl>    <dbl>     <dbl>    <dbl>        <dbl>    <dbl>     <dbl>
## 1       181.     79.3      34.9     130.         163.      214       755
```

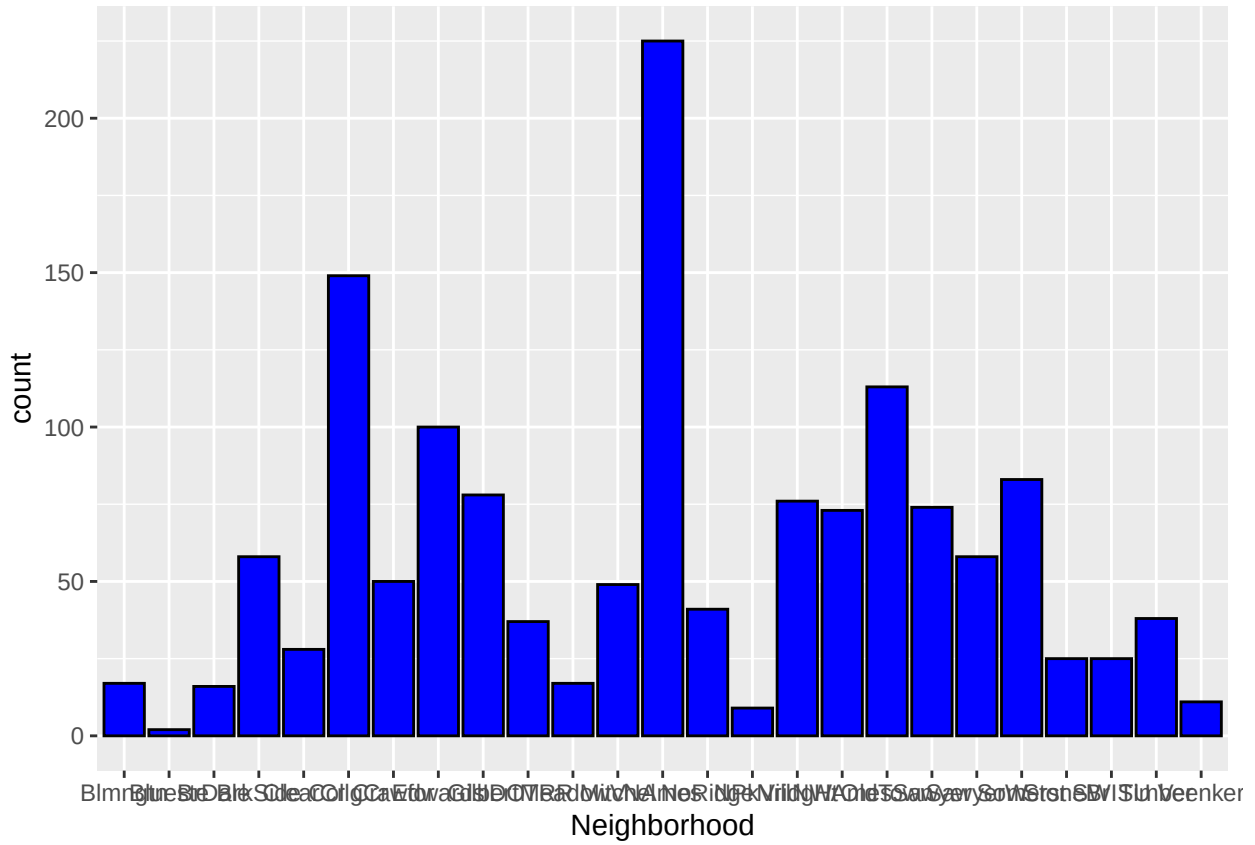
```
ggplot(data = HousePrice, aes(x =SalePrice_1000)) + geom_histogram(color ="black", fill="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

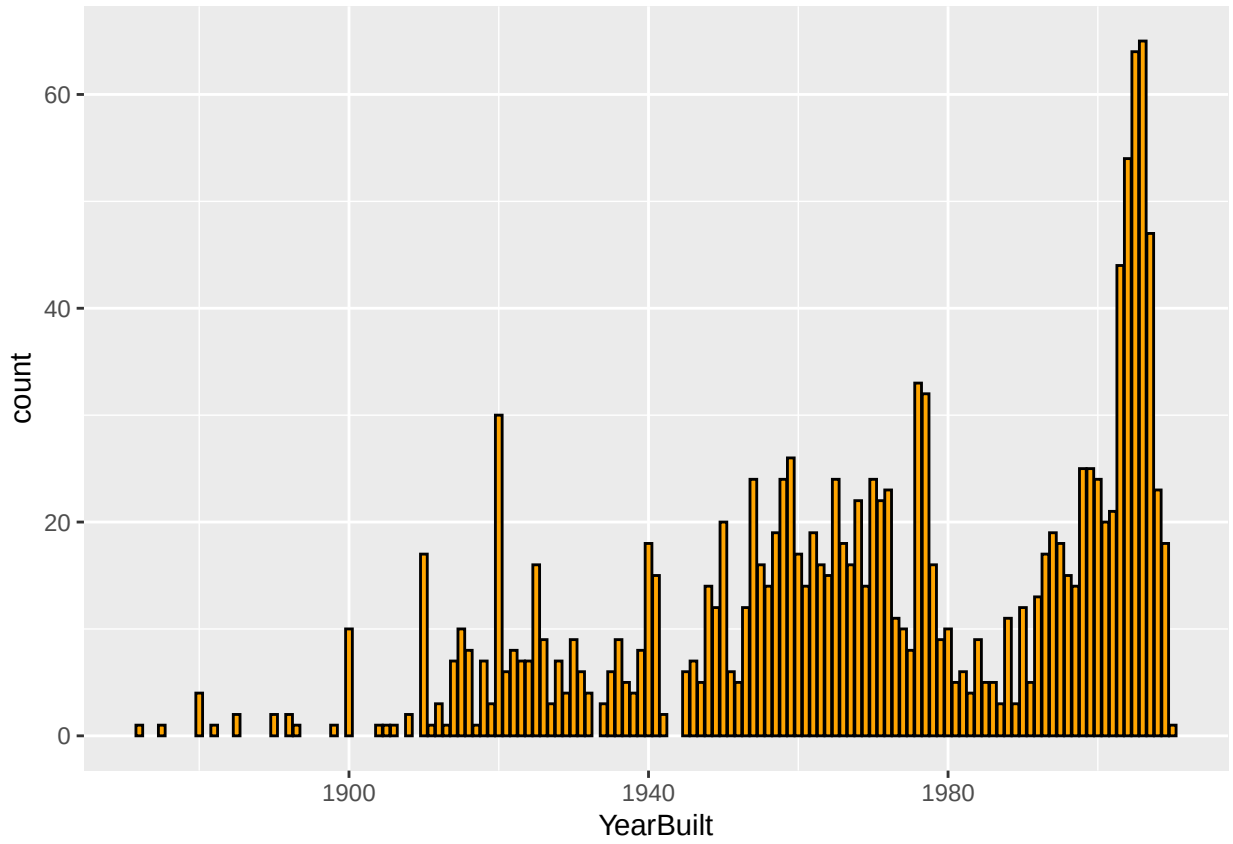


In this section I create a few graphs to display distributions and relationships between variables.

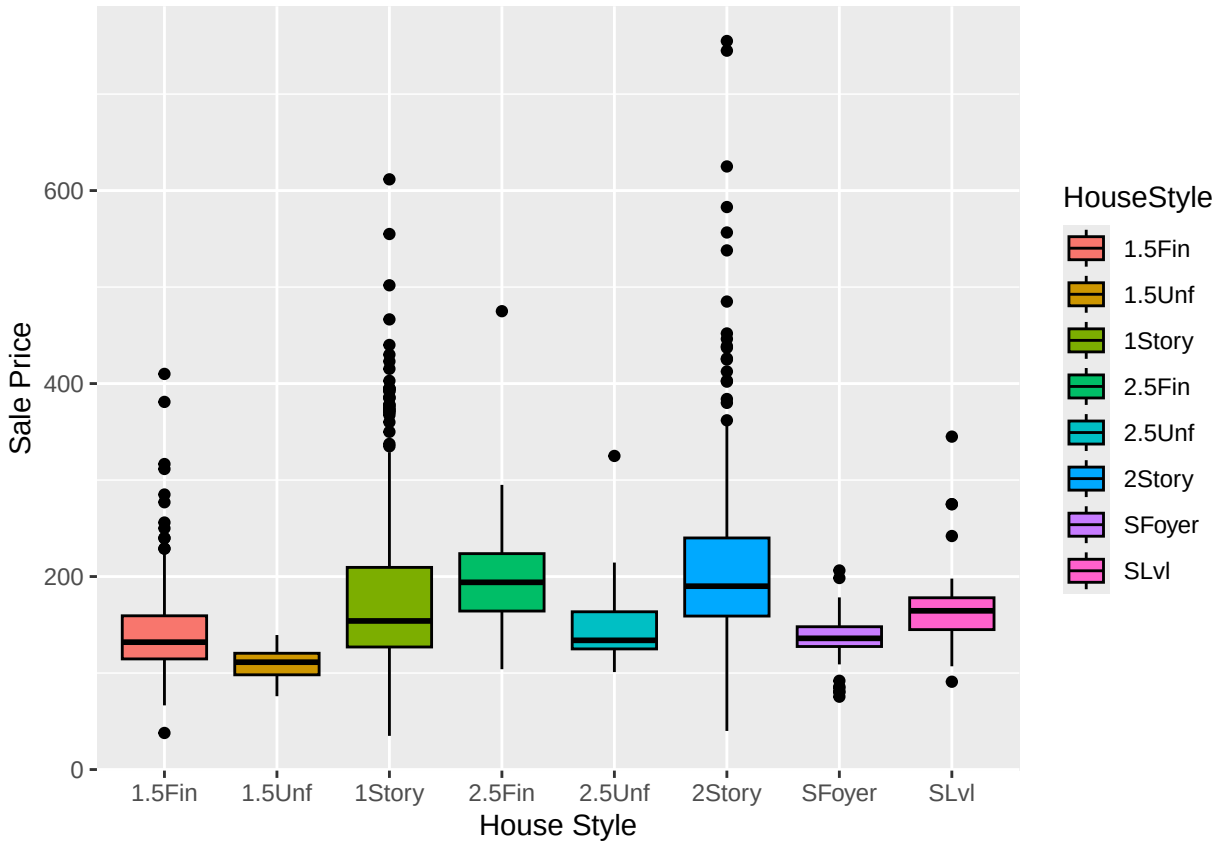
```
ggplot(data = HousePrice, aes(x =Neighborhood)) + geom_bar(color ="black", fill="blue")
```




```
ggplot(data = HousePrice, aes(x = YearBuilt)) + geom_bar(color ="black", fill="orange")
```

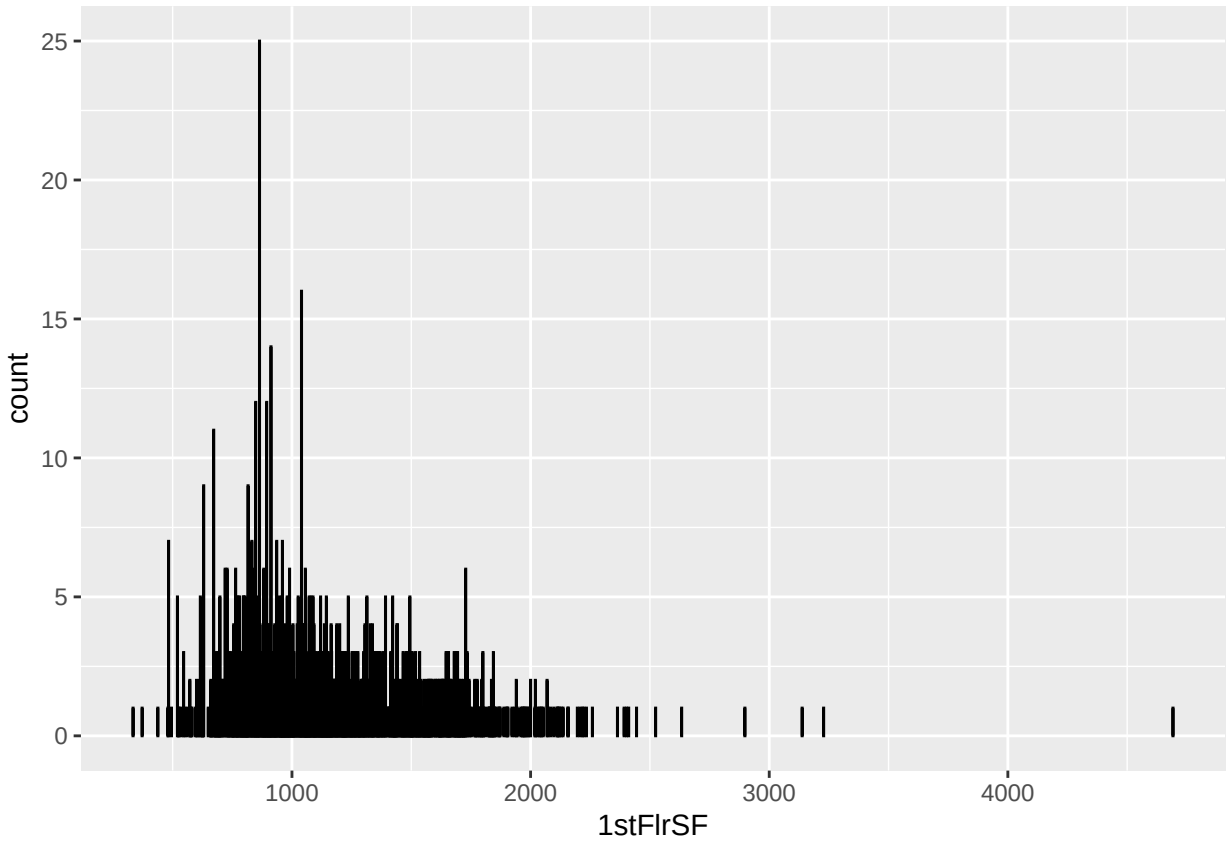



```
ggplot(data = HousePrice, aes(x = HouseStyle, y = SalePrice_1000, fill=HouseStyle)) + geom_boxplot(colo
```



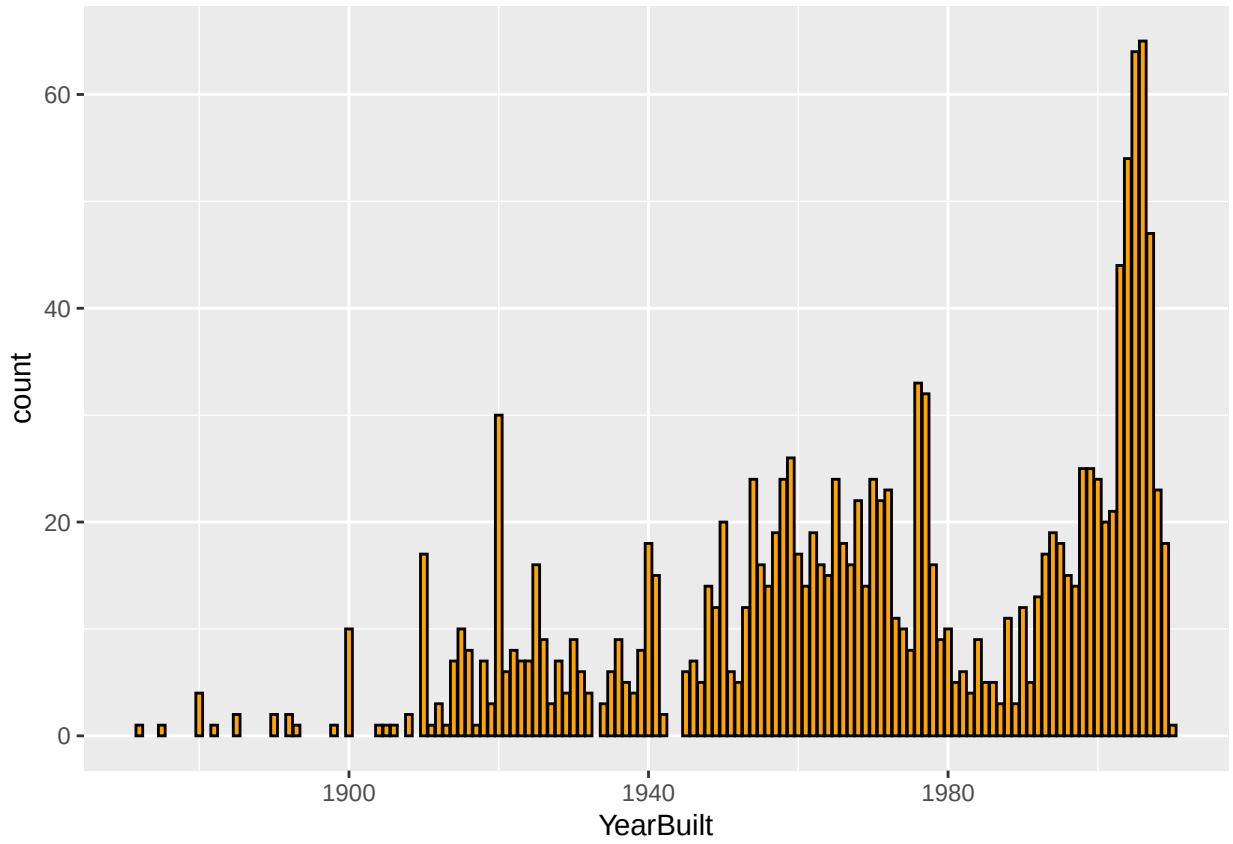

```
ggplot(data = HousePrice, aes(x =`1stFlrSF`, fill = `1stFlrSF`)) + geom_bar(color ="black")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

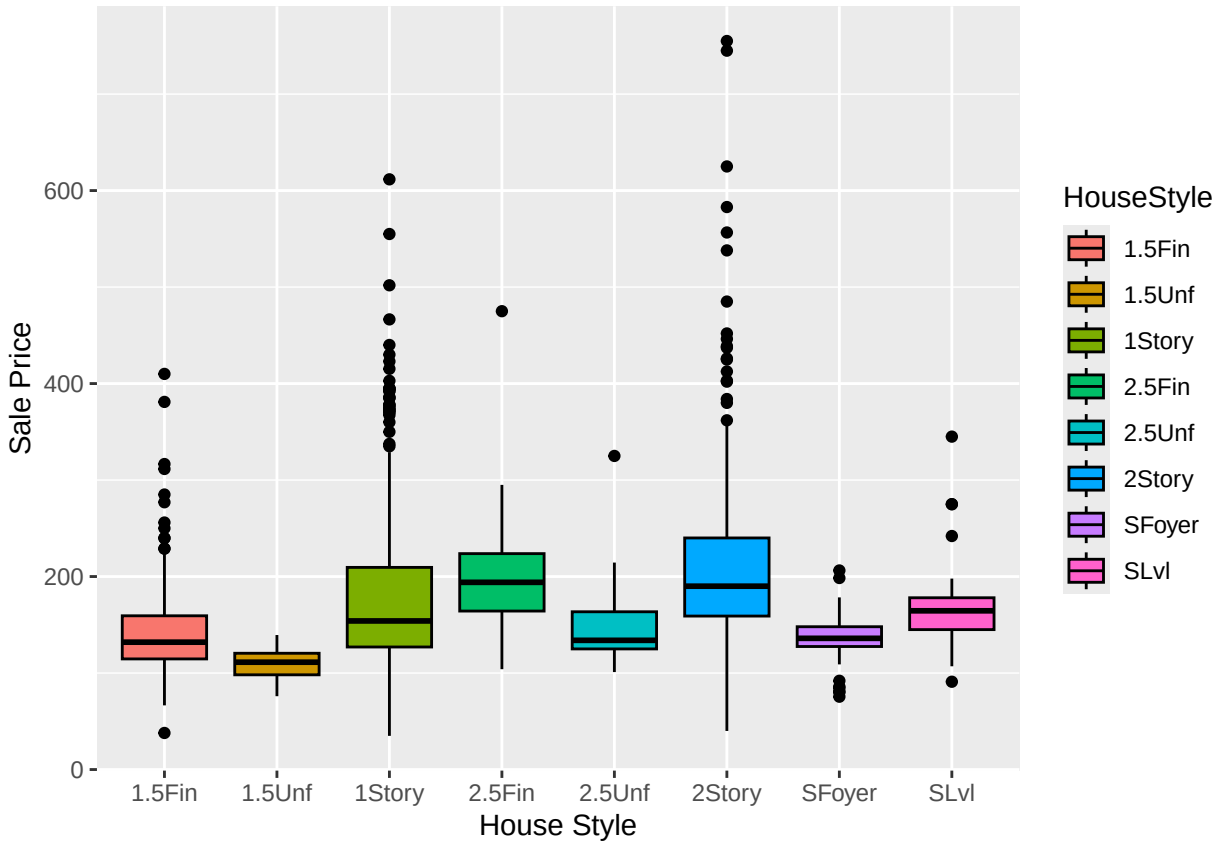


And then I added a few more additiona graphs to display the association between the response variable and two explanatory variables:

```
ggplot(data = HousePrice, aes(x = YearBuilt)) + geom_bar(color ="black", fill="orange")
```

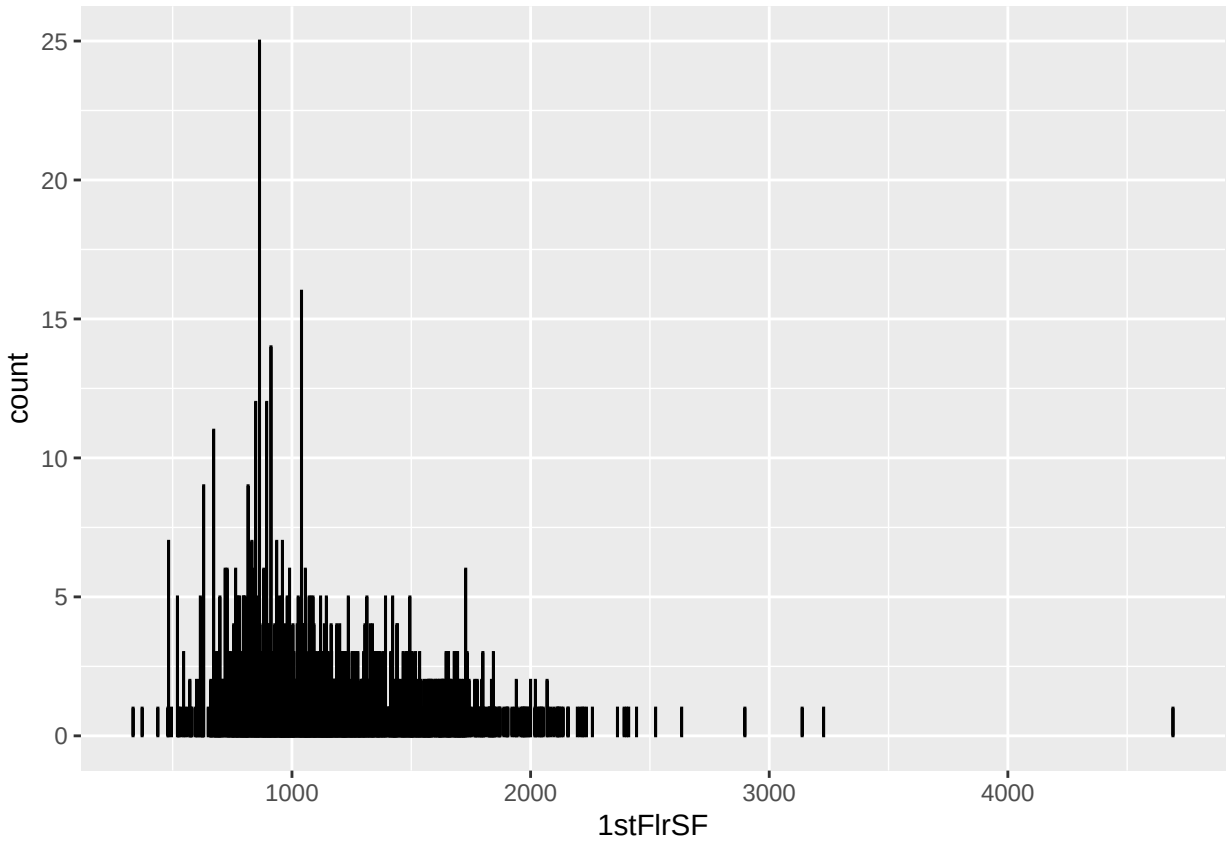



```
ggplot(data = HousePrice, aes(x = HouseStyle, y = SalePrice_1000, fill=HouseStyle)) + geom_boxplot(colo
```




```
ggplot(data = HousePrice, aes(x =`1stFlrSF`, fill = `1stFlrSF`)) + geom_bar(color ="black")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



DAP Part 2

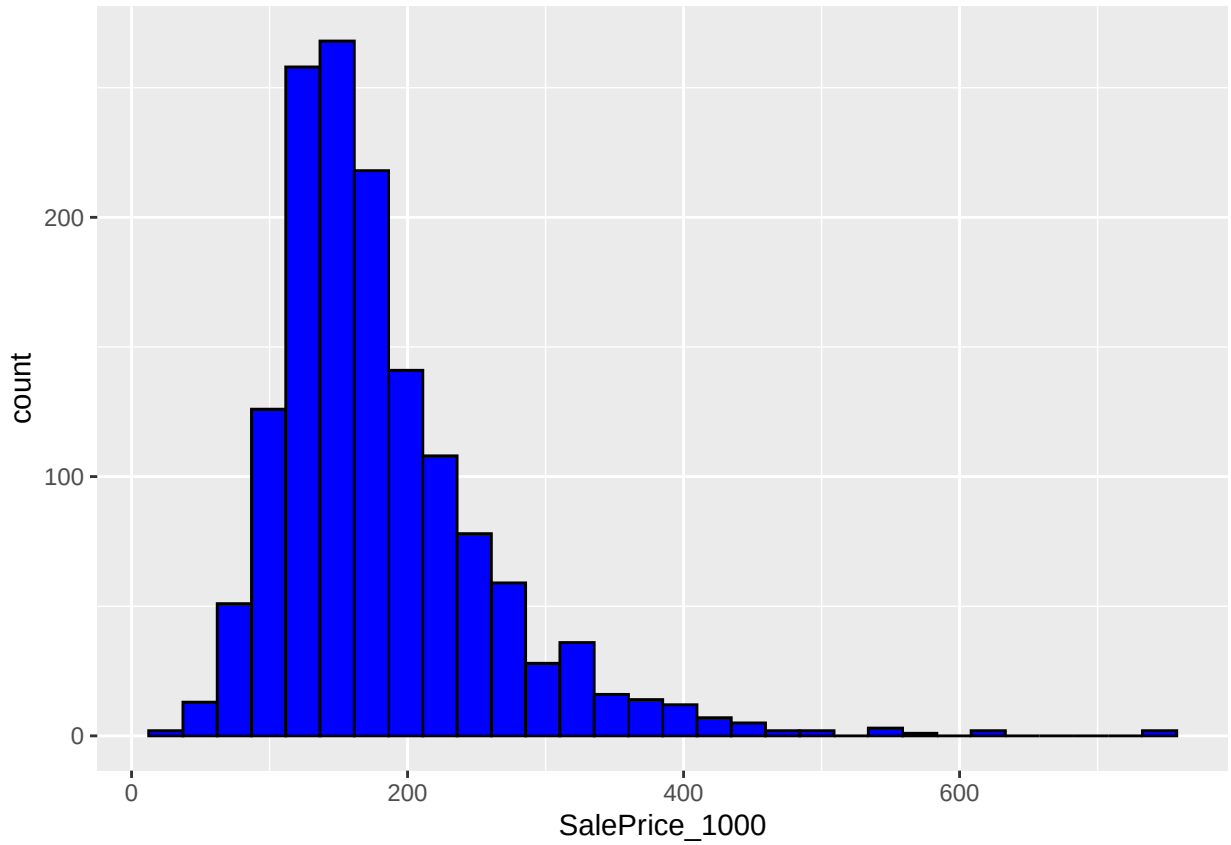
Here once again, I report the summary statistics:

```
HousePrice %>%
  summarise(Mean_Price = mean(SalePrice_1000),
            SD_Price = sd(SalePrice_1000),
            Min_Price = min(SalePrice_1000),
            Q1_Price = quantile(SalePrice_1000, .25),
            Median_Price = median(SalePrice_1000),
            Q3_Price = quantile(SalePrice_1000, .75),
            Max_Price = max(SalePrice_1000)
  )
```

```
## # A tibble: 1 x 7
##   Mean_Price SD_Price Min_Price Q1_Price Median_Price Q3_Price Max_Price
##        <dbl>    <dbl>     <dbl>    <dbl>        <dbl>    <dbl>     <dbl>
## 1       181.     79.3      34.9     130.         163.      214       755
```

```
ggplot(data = HousePrice, aes(x =SalePrice_1000)) + geom_histogram(color ="black", fill="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



When checking the assumptions, the distribution of my response variable is right-skewed, but the ,sample size is sufficiently large as it is it 1452, so the Central Limit Theorem applies.

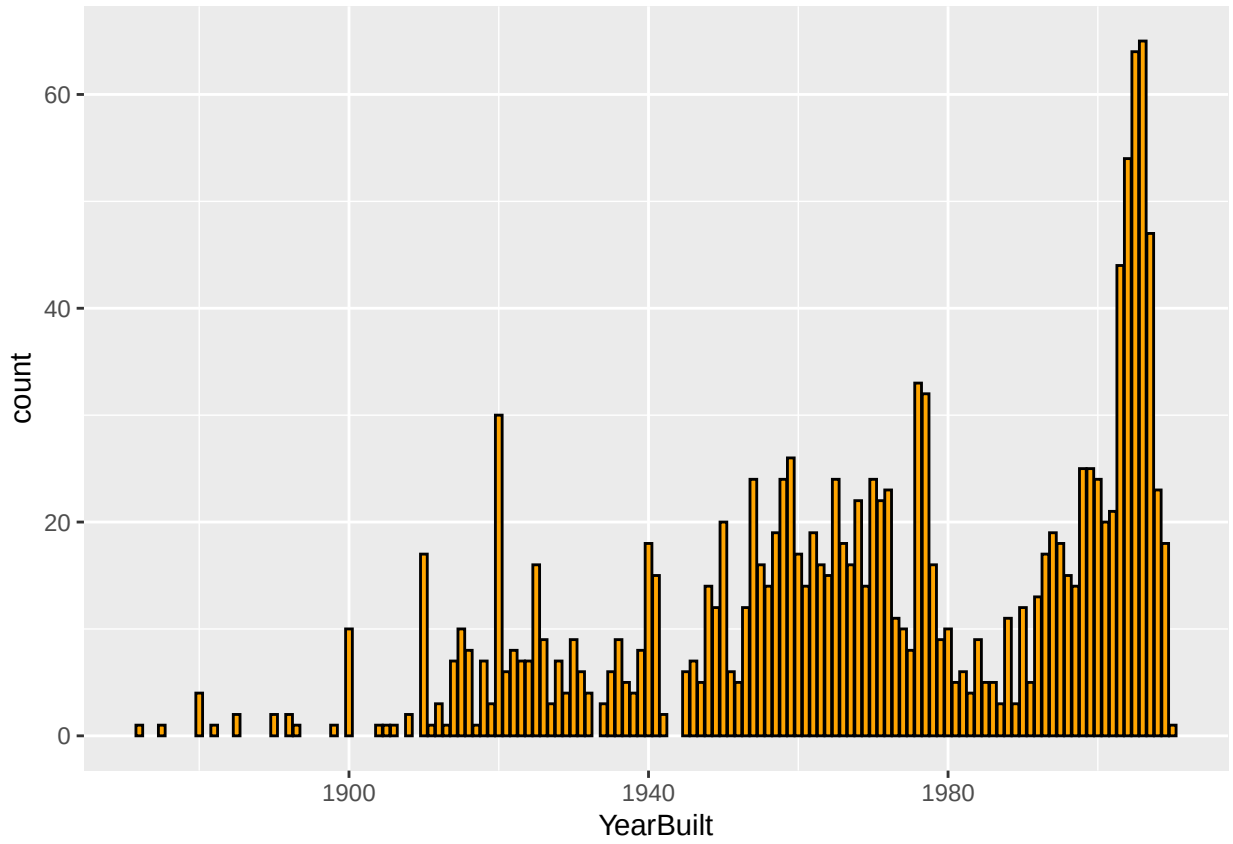
```
HousePrice %>%
  filter(!is.na(SalePrice_1000)) %>%
  t_test(response = SalePrice_1000,
         conf_int = TRUE,
         conf_level = .95,
         mu = 180.6151,
         alternative = "two-sided")
```

```
## # A tibble: 1 x 7
##    statistic  t_df p_value alternative estimate lower_ci upper_ci
##        <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1 -0.0000176  1451    1.00 two.sided       181.     177.     185.
```

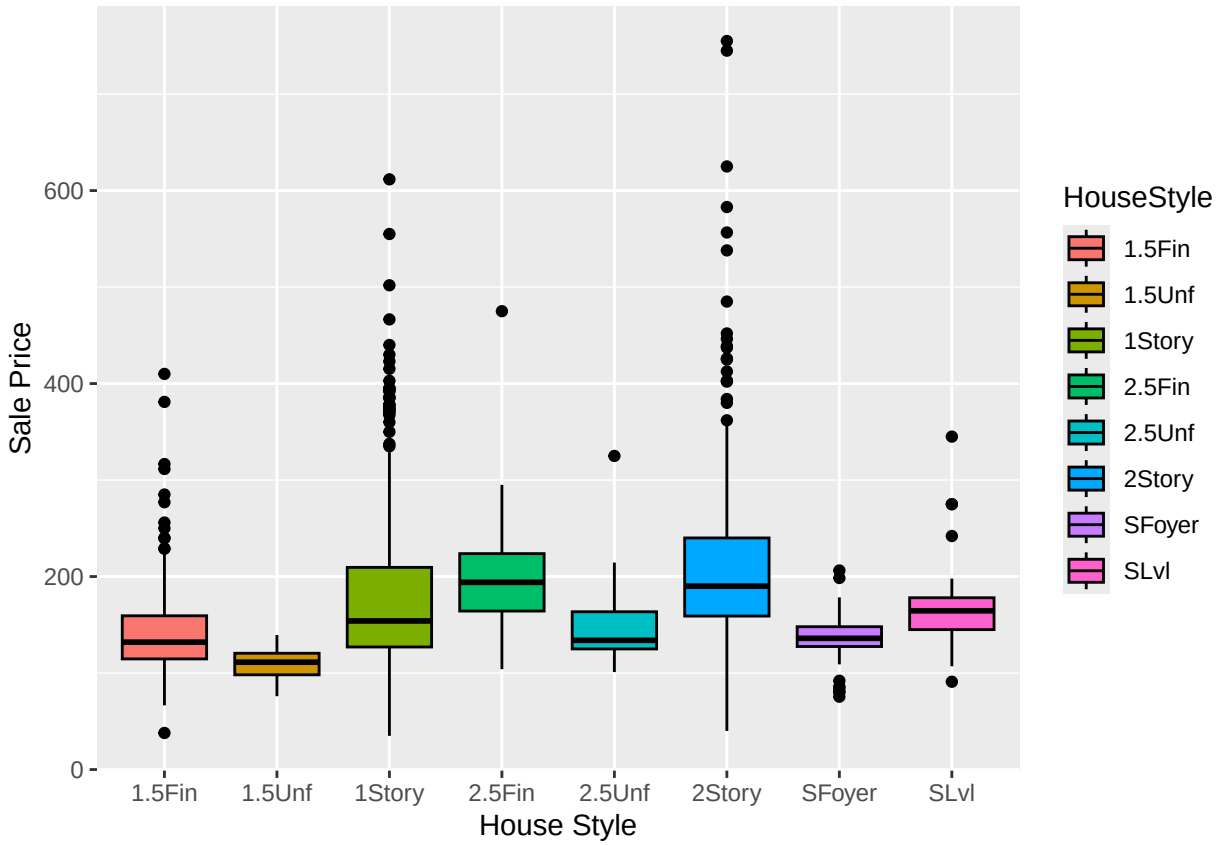
I am 95% confident that the true population mean price of a house is between 176.5336 and 184.6966 thousands of dollars.

The following graphs are a few from the earlier part of the report with a few being a new additions for a better display of categorical variables.

```
ggplot(data = HousePrice, aes(x = YearBuilt)) + geom_bar(color ="black", fill="orange")
```

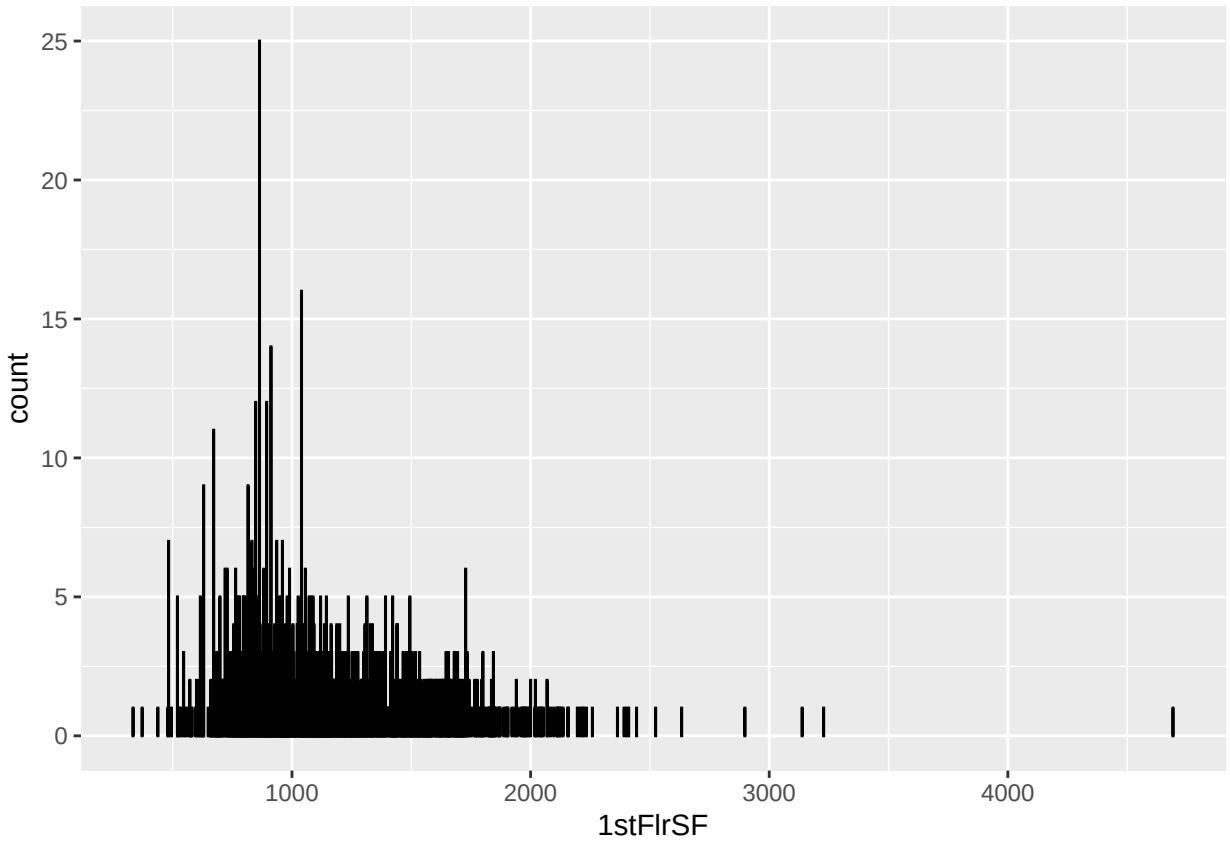



```
ggplot(data = HousePrice, aes(x = HouseStyle, y = SalePrice_1000, fill=HouseStyle)) + geom_boxplot(colo
```

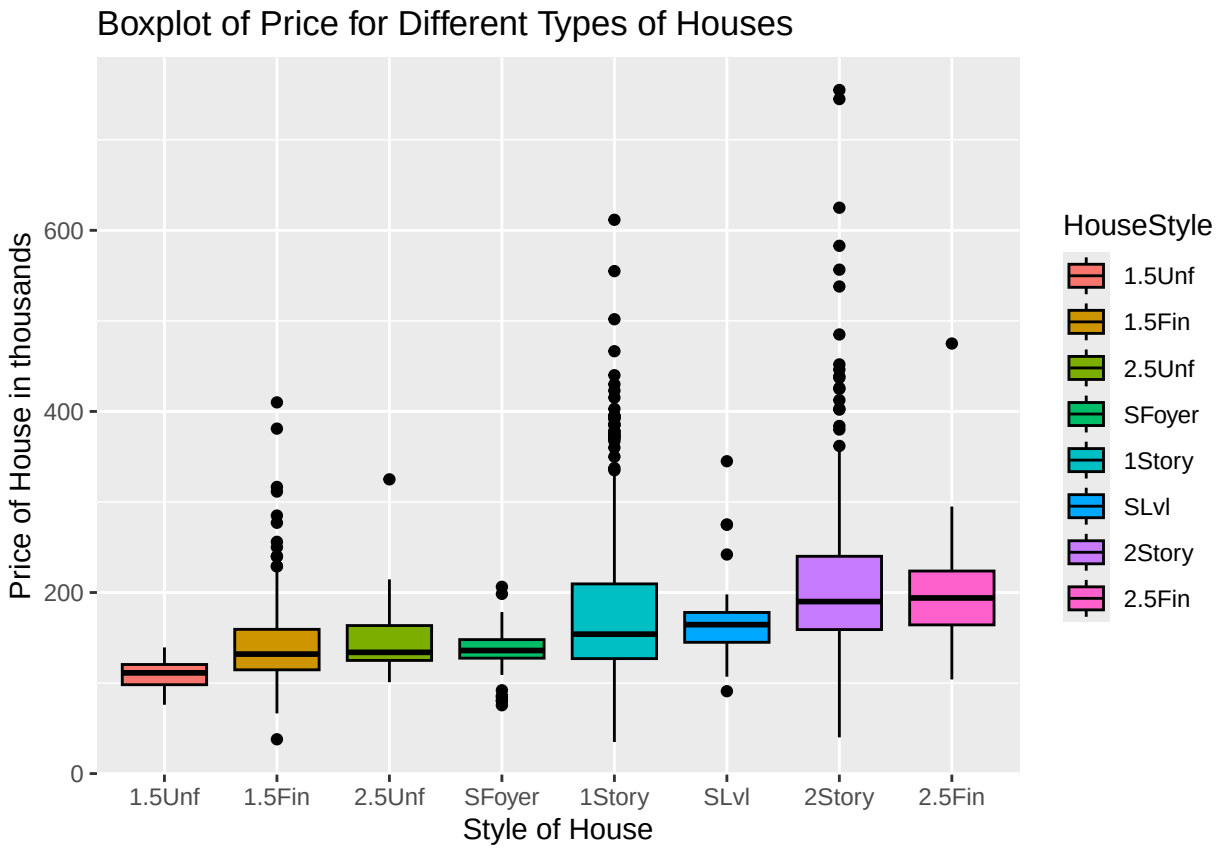



```
ggplot(data = HousePrice, aes(x =`1stFlrSF`, fill = `1stFlrSF`)) + geom_bar(color ="black")
```

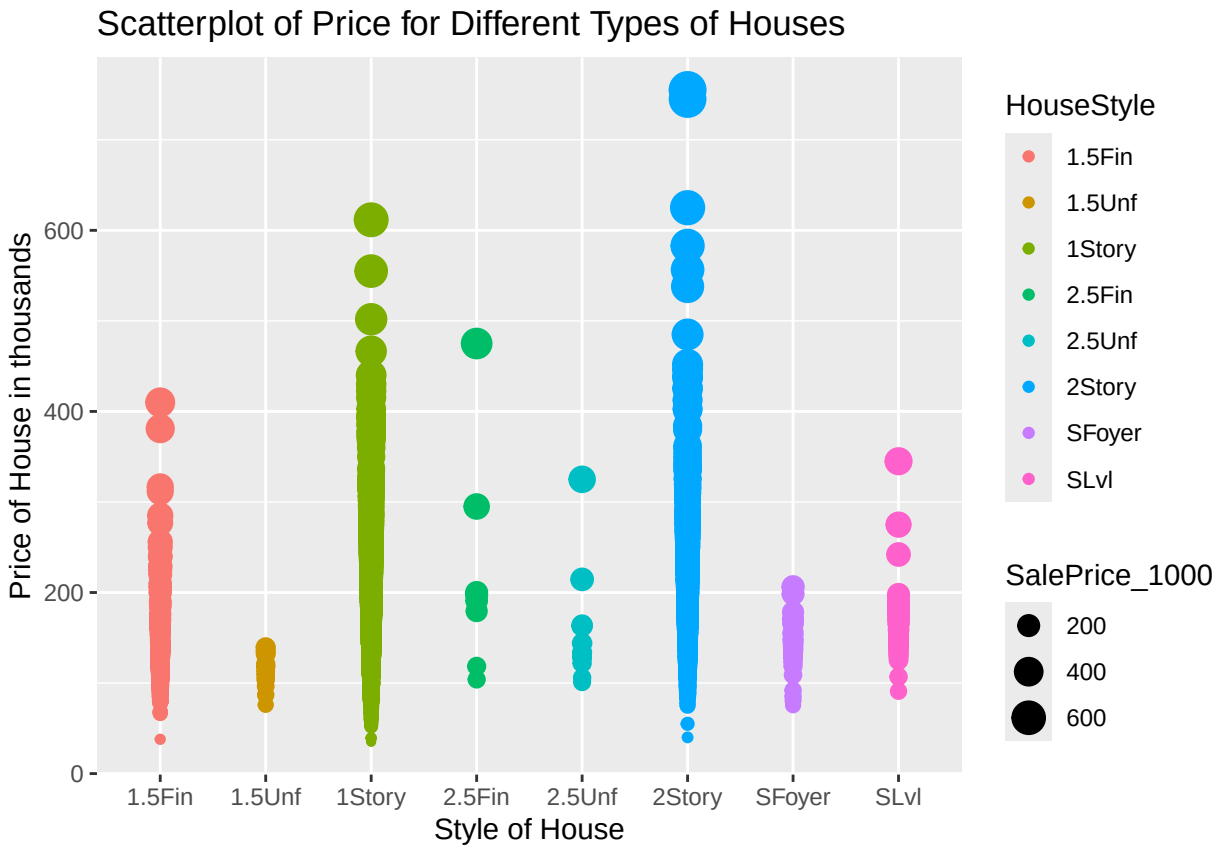
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
##DAP Part 2 Plots
HousePrice %>%
  mutate(HouseStyle = reorder(HouseStyle,SalePrice_1000,median)) %>%
ggplot(aes(x = HouseStyle, y=SalePrice_1000, fill = HouseStyle)) + geom_boxplot(color ="black") +labs(x=
```




```
HousePrice %>%
  ggplot(aes(x=HouseStyle, y=SalePrice_1000, col=HouseStyle)) + geom_point(aes(size=SalePrice_1000)) + 
```




```
HousePrice %>%
  group_by(HouseStyle) %>%
  summarise(Mean_Price = mean(SalePrice_1000),
            SD_Price = sd(SalePrice_1000),
            Median_Price = median(SalePrice_1000),
  )
```

```
## # A tibble: 8 x 4
##   HouseStyle Mean_Price SD_Price Median_Price
##   <chr>           <dbl>    <dbl>        <dbl>
## 1 1.5Fin           143.     54.3         132
## 2 1.5Unf           110.     19.0         111.
## 3 1Story           175.     76.6         154
## 4 2.5Fin           220     118.          194
## 5 2.5Unf           157.     63.9         134.
## 6 2Story           210.     87.6         190
## 7 SFoyer           135.     30.5         136.
## 8 SLvl             167.     38.3         164.
```

```
HousePrice %>%
  group_by(Neighborhood) %>%
  summarise(Mean_Price = mean(SalePrice_1000),
            SD_Price = sd(SalePrice_1000),
            Median_Price = median(SalePrice_1000),
  )
```

```
## # A tibble: 25 x 4
##    Neighborhood Mean_Price SD_Price Median_Price
```

```
##    <chr>             <dbl>    <dbl>        <dbl>
##  1 Blmngtn            195.     30.4         191
##  2 Blueste            138.     19.1         138.
##  3 BrDale             104.     14.3         106
##  4 BrkSide            125.     40.3         124.
##  5 ClearCr            213.     50.2         200.
##  6 CollgCr            198.     51.5         196.
##  7 Crawfor            211.     69.6         209.
##  8 Edwards            128.     43.2         122.
##  9 Gilbert            193.     36.1         181
## 10 IDOTRR             100.     33.4         103
## # i 15 more rows
```

```
HousePrice %>%
  filter(!is.na(SalePrice_1000)) %>%
  t_test(response = SalePrice_1000,
         explanatory = Neighborhood,
         order = c("ClearCr","Edwards"),
         conf_int = TRUE,
         conf_level = .90,
         alternative = "two-sided")
```

```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      8.09  38.9 7.40e-10 two.sided       84.3     66.8     102.
```

I am 90% confident that the true population difference in price between the neigborhoods, ClearCr and Edwards is between 66.77133(in thousands) dollars and 101.9201(in thousands) dollars.

Part 3

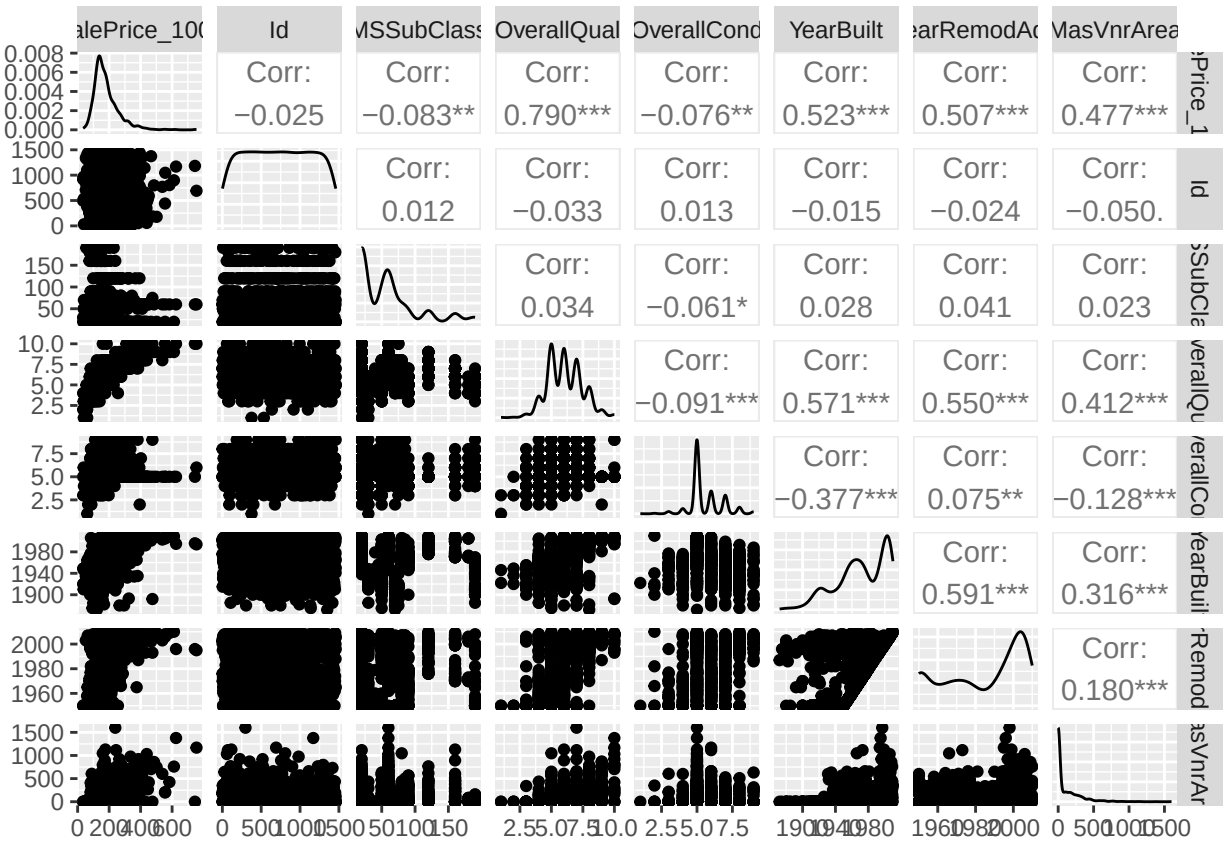
Correlation

```
glimpse(HousePrice)
```

```
## Rows: 1,452
## Columns: 54
## $ Id            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ MSSubClass    <dbl> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, 20~
## $ MSZoning      <chr> "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RM", "~
## $ Street        <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pave",~
## $ LotShape      <chr> "Reg", "Reg", "IR1", "IR1", "IR1", "IR1", "Reg", "IR1",~
## $ Utilities     <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "AllP~
## $ LotConfig     <chr> "Inside", "FR2", "Inside", "Corner", "FR2", "Inside", "~
## $ LandSlope     <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl",~
## $ Neighborhood  <chr> "CollgCr", "Veenker", "CollgCr", "Crawfor", "NoRidge", ~
## $ Condition1    <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "Norm"~
## $ Condition2    <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "Norm",~
## $ BldgType      <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam",~
## $ HouseStyle    <chr> "2Story", "1Story", "2Story", "2Story", "2Story", "1.5F~
## $ OverallQual   <dbl> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4, 5~
## $ OverallCond   <dbl> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 5~
## $ YearBuilt     <dbl> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931, 1~
## $ YearRemodAdd  <dbl> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950, 1~
```

```
## $ RoofStyle      <chr> "Gable", "Gable", "Gable", "Gable", "Gable", "Gable", "~
## $ RoofMatl       <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg", ~
## $ Exterior1st    <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "VinylSd", ~
## $ Exterior2nd    <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Shng", "VinylSd", ~
## $ MasVnrType     <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace", "None"~
## $ MasVnrArea     <dbl> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 306,~
## $ ExterQual      <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA", "TA", "~
## $ ExterCond      <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "~
## $ Foundation     <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", "Wood", ~
## $ BsmtFinSF1     <dbl> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 9~
## $ BsmtFinSF2     <dbl> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CentralAir     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", ~
## $ `1stFlrSF`     <dbl> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077,~
## $ `2ndFlrSF`     <dbl> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142, 0~
## $ LowQualFinSF   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ GrLivArea      <dbl> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774, 1~
## $ BsmtFullBath   <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1~
## $ BsmtHalfBath   <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ FullBath       <dbl> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1~
## $ HalfBath       <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1~
## $ BedroomAbvGr   <dbl> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3~
## $ KitchenAbvGr   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1~
## $ KitchenQual    <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA", "TA", "~
## $ TotRmsAbvGrd   <dbl> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6, ~
## $ Functional     <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ",~
## $ Fireplaces     <dbl> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0~
## $ GarageCars     <dbl> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2~
## $ GarageArea     <dbl> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, ~
## $ PavedDrive     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", ~
## $ WoodDeckSF     <dbl> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 16~
## $ PoolArea       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ MiscVal        <dbl> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 700~
## $ MoSold         <dbl> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3, 1~
## $ YrSold         <dbl> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 2~
## $ SaleType       <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "~
## $ SaleCondition  <chr> "Normal", "Normal", "Normal", "Abnorml", "Normal", "Nor~
## $ SalePrice_1000 <dbl> 208.5, 181.5, 223.5, 140.0, 250.0, 143.0, 307.0, 200.0,~
```

```
HousePrice %>%
  select(SalePrice_1000, Id, MSSubClass, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrArea)
  ggpairs()
```



In my dataset, OverallQuality has the strongest correlation with the response variable SalesPrice_1000 with a correlation value of 0.790. The next strongest correlations with the SalePrice are YearBuilt and YearRemodAdd.

Predictive Modeling

Develop a multiple linear regression model to predict the outcome (response) variable using all the relevant explanatory variables.

```
Model <- lm(SalePrice_1000 ~ OverallQual, data=HousePrice)
summary(Model)
```

```
##
## Call:
## lm(formula = SalePrice_1000 ~ OverallQual, data = HousePrice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.78  -29.40   -1.73   21.15  397.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.6745     5.7739  -16.57   <2e-16 ***
## OverallQual  45.3456     0.9242   49.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 48.63 on 1450 degrees of freedom
## Multiple R-squared:  0.6241, Adjusted R-squared:  0.6238
## F-statistic:  2407 on 1 and 1450 DF,  p-value: < 2.2e-16
```

The linear regression model is $SalePrice_1000 = -95.6745 + 45.3456 * OverallQual$. To interpret the y-intercept when the OverallQual is 0, the model predicts the price of the car to be negative 95.6745 thousand dollars. For the slope, the model predicts that the price will increase by 45.3456 thousand dollars for every 1 unit increase

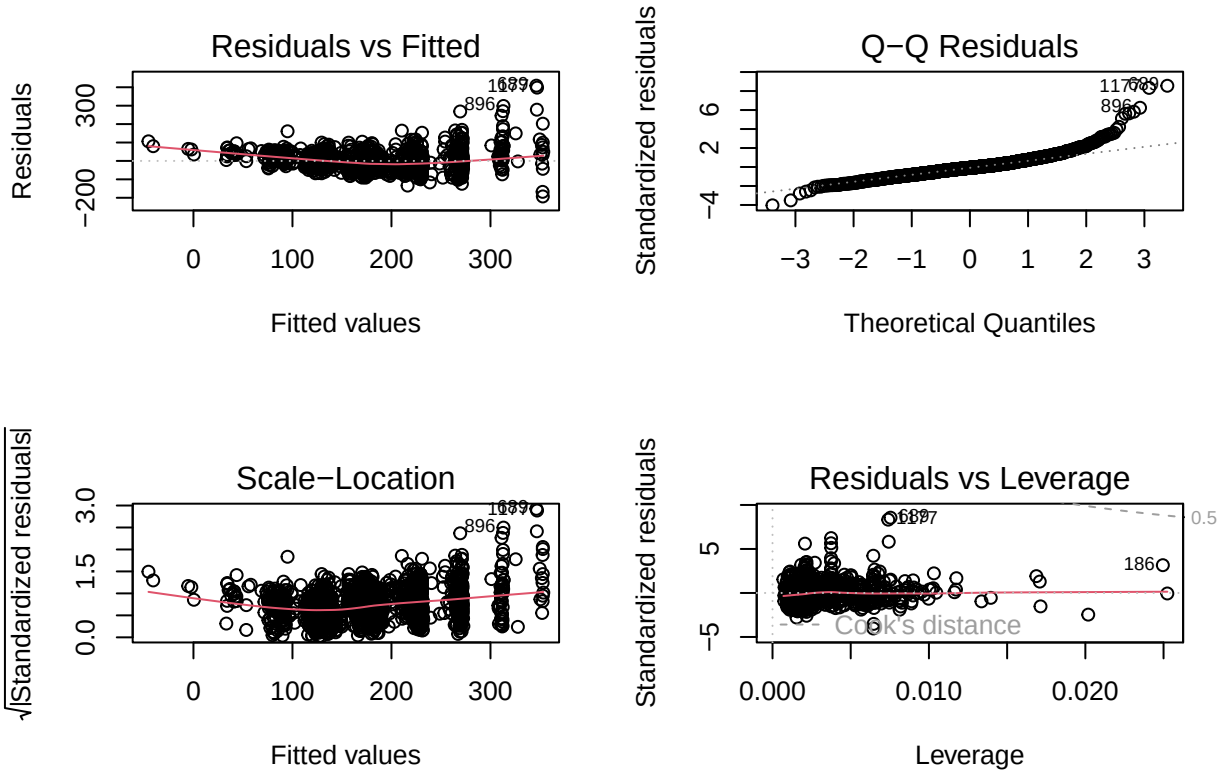
Multiple Linear Regression Model

```
Model2 <- lm(SalePrice_1000 ~ OverallQual + YearBuilt + YearRemodAdd,  data=HousePrice)
summary(Model2)
```

```
##
## Call:
## lm(formula = SalePrice_1000 ~ OverallQual + YearBuilt + YearRemodAdd,
##     data = HousePrice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -192.87  -27.74   -4.01   19.73  408.63
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.026e+03  1.476e+02  -6.952 5.43e-12 ***
## OverallQual   4.051e+01  1.172e+00  34.564  < 2e-16 ***
## YearBuilt     1.991e-01  5.551e-02   3.587 0.000346 ***
## YearRemodAdd  2.861e-01  7.974e-02   3.587 0.000345 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.95 on 1448 degrees of freedom
## Multiple R-squared:  0.635,  Adjusted R-squared:  0.6342
## F-statistic: 839.6 on 3 and 1448 DF,  p-value: < 2.2e-16
```

The multiple linear regression model is $SalePrice_1000 = -1.026e+03 + 4.051e+01 * OverallQual + 1.991e-01 * YearBuilt + 2.861e-01 * YearRemodAdd$. The response variable is affected in a positive way by the variables with a positive slope, that is, OverallQual, YearBuilt, YearRemodAdd. The adjusted R-squared value tells us that 63.42% of the total variation in the response variable (SalePrice_1000) is explained by the explanatory variables.

```
par(mfrow=c(2,2))
plot(Model2)
```



Methodology

Type of model

Pros and Cons of the model

Results

Output from R

Interpretation

Explain the results

Conclusion

Inference of the results

Discussion

Possible improvement to the project

Reference