EDA

Tuan Pham

2023-04-27

1. Load the Dataset

```
# Load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library("scales")
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# Load a dataset
data <- read_csv('C:/Users/arsen/OneDrive/Desktop/ames.csv')
```

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## 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.
```

```
head(data)
```

```
## # A tibble: 6 x 81
##      Id MSSubClass MSZoning LotFr~1 LotArea Street Alley LotSh~2 LandC~3 Utili~4
##   <dbl>      <dbl> <chr>      <dbl>   <dbl> <chr>  <chr> <chr>   <chr>   <chr>
## 1     1         60 RL            65    8450 Pave   <NA>  Reg     Lvl     AllPub
## 2     2         20 RL            80    9600 Pave   <NA>  Reg     Lvl     AllPub
## 3     3         60 RL            68   11250 Pave   <NA>  IR1     Lvl     AllPub
## 4     4         70 RL            60    9550 Pave   <NA>  IR1     Lvl     AllPub
## 5     5         60 RL            84   14260 Pave   <NA>  IR1     Lvl     AllPub
## 6     6         50 RL            85   14115 Pave   <NA>  IR1     Lvl     AllPub
## # ... with 71 more variables: LotConfig <chr>, LandSlope <chr>,
## #   Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>, BldgType <chr>,
## #   HouseStyle <chr>, OverallQual <dbl>, OverallCond <dbl>, YearBuilt <dbl>,
## #   YearRemodAdd <dbl>, RoofStyle <chr>, RoofMatl <chr>, Exterior1st <chr>,
## #   Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <dbl>, ExterQual <chr>,
## #   ExterCond <chr>, Foundation <chr>, BsmtQual <chr>, BsmtCond <chr>,
## #   BsmtExposure <chr>, BsmtFinType1 <chr>, BsmtFinSF1 <dbl>, ...
```

```
#View a structure of the data
str(data)
```

```
## spc_tbl_ [1,460 x 81] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id           : num [1:1460] 1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : num [1:1460] 60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr [1:1460] "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : num [1:1460] 65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : num [1:1460] 8450 9600 11250 9550 14260 ...
##  $ Street       : chr [1:1460] "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr [1:1460] NA NA NA NA ...
##  $ LotShape     : chr [1:1460] "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr [1:1460] "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr [1:1460] "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr [1:1460] "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr [1:1460] "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr [1:1460] "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr [1:1460] "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr [1:1460] "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr [1:1460] "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr [1:1460] "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : num [1:1460] 7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : num [1:1460] 5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : num [1:1460] 2003 1976 2001 1915 2000 ...
##  $ YearRemodAdd : num [1:1460] 2003 1976 2002 1970 2000 ...
##  $ RoofStyle    : chr [1:1460] "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr [1:1460] "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr [1:1460] "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr [1:1460] "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr [1:1460] "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : num [1:1460] 196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr [1:1460] "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr [1:1460] "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr [1:1460] "PConc" "CBlock" "PConc" "BrkTil" ...
```

```
##  $ BsmtQual     : chr [1:1460] "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr [1:1460] "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr [1:1460] "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr [1:1460] "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : num [1:1460] 706 978 486 216 655 ...
##  $ BsmtFinType2 : chr [1:1460] "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2   : num [1:1460] 0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : num [1:1460] 150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : num [1:1460] 856 1262 920 756 1145 ...
##  $ Heating      : chr [1:1460] "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr [1:1460] "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr [1:1460] "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr [1:1460] "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 1stFlrSF     : num [1:1460] 856 1262 920 961 1145 ...
##  $ 2ndFlrSF     : num [1:1460] 854 0 866 756 1053 ...
##  $ LowQualFinSF : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : num [1:1460] 1710 1262 1786 1717 2198 ...
##  $ BsmtFullBath : num [1:1460] 1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : num [1:1460] 0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : num [1:1460] 2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : num [1:1460] 1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : num [1:1460] 3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : num [1:1460] 1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : chr [1:1460] "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : num [1:1460] 8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr [1:1460] "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : num [1:1460] 0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr [1:1460] NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr [1:1460] "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : num [1:1460] 2003 1976 2001 1998 2000 ...
##  $ GarageFinish : chr [1:1460] "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : num [1:1460] 2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : num [1:1460] 548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr [1:1460] "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr [1:1460] "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr [1:1460] "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : num [1:1460] 0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : num [1:1460] 61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: num [1:1460] 0 0 0 272 0 0 0 228 205 0 ...
##  $ 3SsnPorch    : num [1:1460] 0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : num [1:1460] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr [1:1460] NA NA NA NA ...
##  $ Fence        : chr [1:1460] NA NA NA NA ...
##  $ MiscFeature  : chr [1:1460] NA NA NA NA ...
##  $ MiscVal      : num [1:1460] 0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : num [1:1460] 2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : num [1:1460] 2008 2007 2008 2006 2008 ...
##  $ SaleType     : chr [1:1460] "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr [1:1460] "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : num [1:1460] 208500 181500 223500 140000 250000 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
```

```
##   ..   MSSubClass = col_double(),
##   ..   MSZoning = col_character(),
##   ..   LotFrontage = col_double(),
##   ..   LotArea = col_double(),
##   ..   Street = col_character(),
##   ..   Alley = col_character(),
##   ..   LotShape = col_character(),
##   ..   LandContour = col_character(),
##   ..   Utilities = col_character(),
##   ..   LotConfig = col_character(),
##   ..   LandSlope = col_character(),
##   ..   Neighborhood = col_character(),
##   ..   Condition1 = col_character(),
##   ..   Condition2 = col_character(),
##   ..   BldgType = col_character(),
##   ..   HouseStyle = col_character(),
##   ..   OverallQual = col_double(),
##   ..   OverallCond = col_double(),
##   ..   YearBuilt = col_double(),
##   ..   YearRemodAdd = col_double(),
##   ..   RoofStyle = col_character(),
##   ..   RoofMatl = col_character(),
##   ..   Exterior1st = col_character(),
##   ..   Exterior2nd = col_character(),
##   ..   MasVnrType = col_character(),
##   ..   MasVnrArea = col_double(),
##   ..   ExterQual = col_character(),
##   ..   ExterCond = col_character(),
##   ..   Foundation = col_character(),
##   ..   BsmtQual = col_character(),
##   ..   BsmtCond = col_character(),
##   ..   BsmtExposure = col_character(),
##   ..   BsmtFinType1 = col_character(),
##   ..   BsmtFinSF1 = col_double(),
##   ..   BsmtFinType2 = col_character(),
##   ..   BsmtFinSF2 = col_double(),
##   ..   BsmtUnfSF = col_double(),
##   ..   TotalBsmtSF = col_double(),
##   ..   Heating = col_character(),
##   ..   HeatingQC = col_character(),
##   ..   CentralAir = col_character(),
##   ..   Electrical = col_character(),
##   ..   `1stFlrSF` = col_double(),
##   ..   `2ndFlrSF` = col_double(),
##   ..   LowQualFinSF = col_double(),
##   ..   GrLivArea = col_double(),
##   ..   BsmtFullBath = col_double(),
##   ..   BsmtHalfBath = col_double(),
##   ..   FullBath = col_double(),
##   ..   HalfBath = col_double(),
##   ..   BedroomAbvGr = col_double(),
##   ..   KitchenAbvGr = col_double(),
##   ..   KitchenQual = col_character(),
##   ..   TotRmsAbvGrd = col_double(),
```

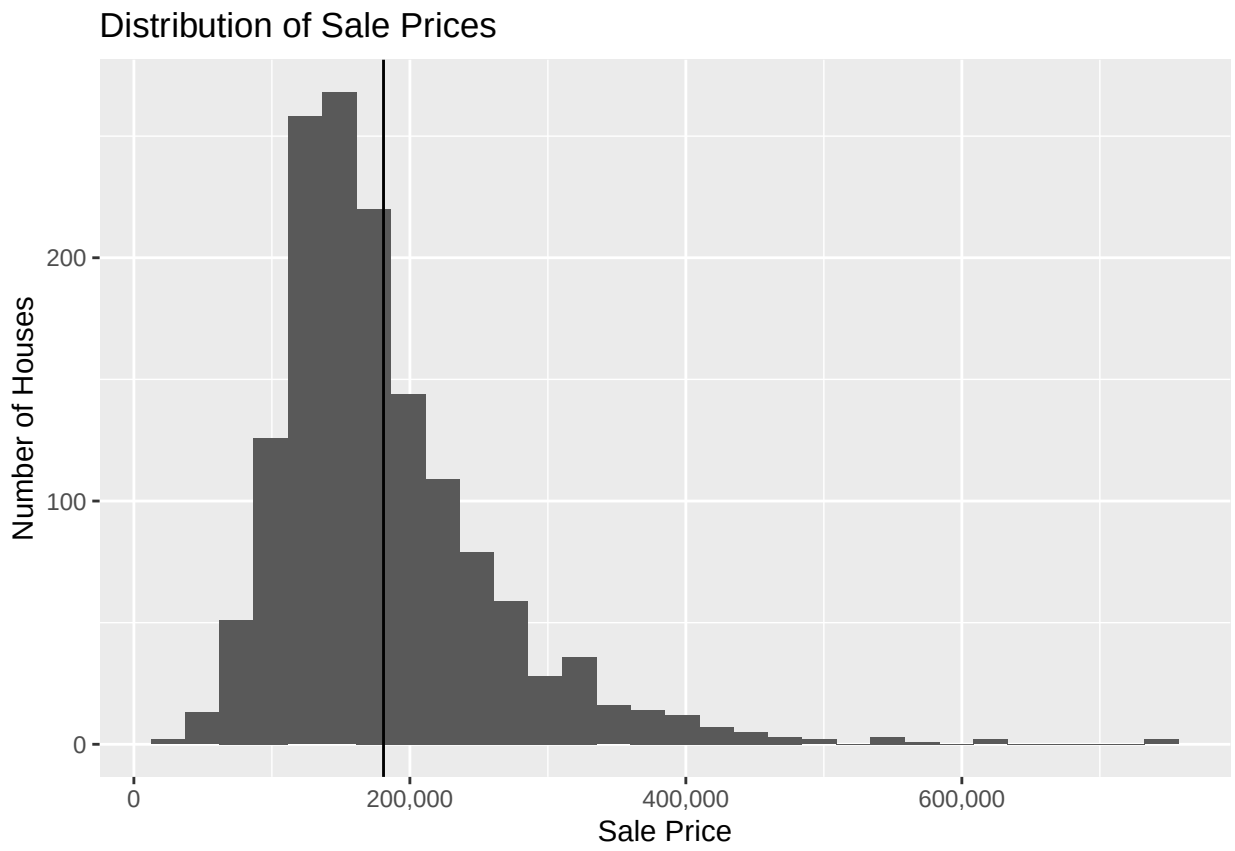
```
##   ..   Functional = col_character(),
##   ..   Fireplaces = col_double(),
##   ..   FireplaceQu = col_character(),
##   ..   GarageType = col_character(),
##   ..   GarageYrBlt = col_double(),
##   ..   GarageFinish = col_character(),
##   ..   GarageCars = col_double(),
##   ..   GarageArea = col_double(),
##   ..   GarageQual = col_character(),
##   ..   GarageCond = col_character(),
##   ..   PavedDrive = col_character(),
##   ..   WoodDeckSF = col_double(),
##   ..   OpenPorchSF = col_double(),
##   ..   EnclosedPorch = col_double(),
##   ..   `3SsnPorch` = col_double(),
##   ..   ScreenPorch = col_double(),
##   ..   PoolArea = col_double(),
##   ..   PoolQC = col_character(),
##   ..   Fence = col_character(),
##   ..   MiscFeature = col_character(),
##   ..   MiscVal = col_double(),
##   ..   MoSold = col_double(),
##   ..   YrSold = col_double(),
##   ..   SaleType = col_character(),
##   ..   SaleCondition = col_character(),
##   ..   SalePrice = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

2. Explore Data Distributions

Sale Price

```
#Set up a plot
ggplot(data, aes(x=SalePrice)) +
  geom_histogram()+
  scale_x_continuous(labels = comma)+
  geom_vline(xintercept=mean(data$SalePrice), color="black") +
  labs(title="Distribution of Sale Prices", x="Sale Price", y="Number of Houses")
```

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




```
#Descriptive statistics of SalePrice
summary(data$SalePrice)
```

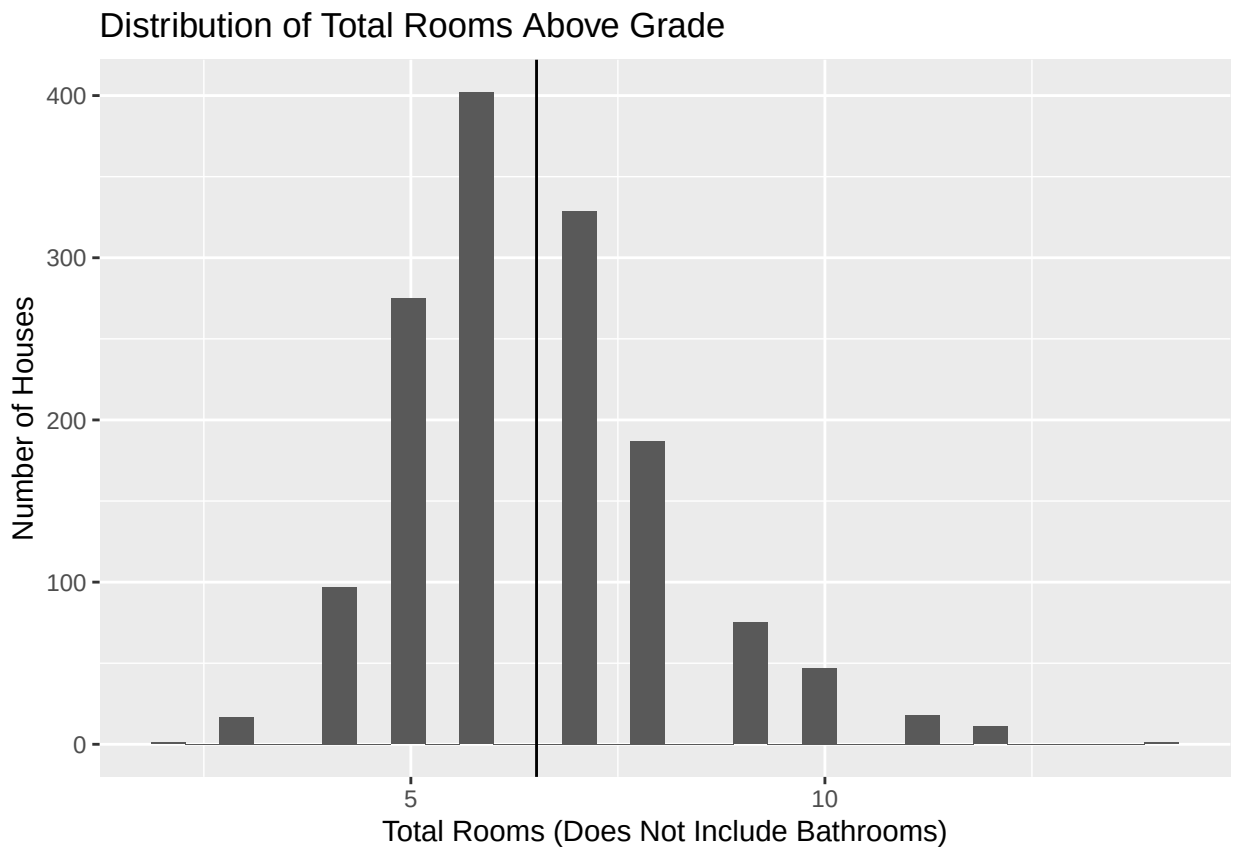
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

Looks like a log normal distribution. Most houses in this sample are clustered around the median value of $163,000, but the higher-end homes are pulling the mean up to over $180,000

Total Rooms Above Grade

```
#Set up a plot
ggplot(data,aes(x=TotRmsAbvGrd)) +
  geom_histogram()+
   geom_vline(xintercept=mean(data$TotRmsAbvGrd), color="black") +
  labs(title="Distribution of Total Rooms Above Grade", x="Total Rooms (Does Not Include Bathrooms)", y=
```

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




```
#Descriptive Statistics of TotRmsAbvGrd
summary(data$TotRmsAbvGrd)
```

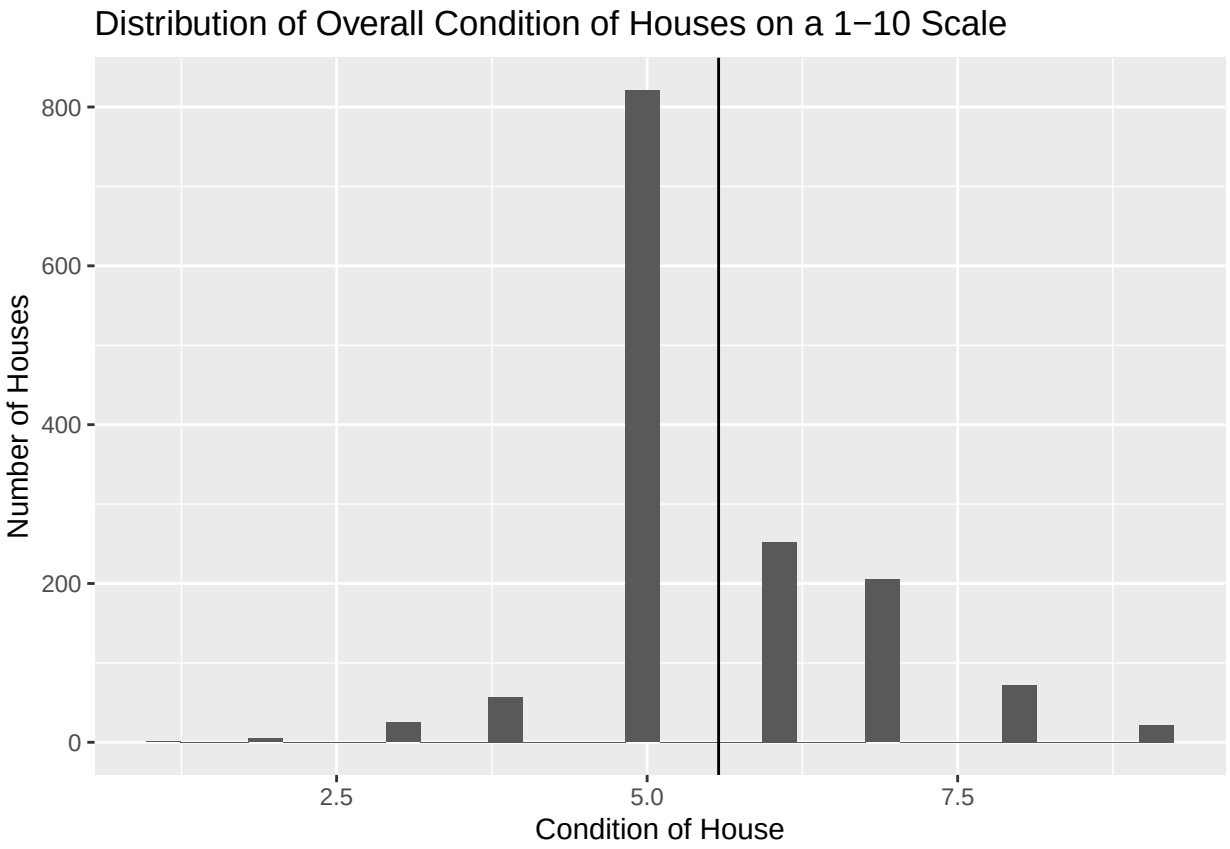
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.000   6.000   6.518   7.000  14.000
```

The number of rooms in houses is approximately normally distributed, with a mean and median around 6 rooms. There are some houses with twice as many rooms as the average, but overall the distribution is less skewed than the sale price distribution

Overall Condition

```
#Set up a plot
ggplot(data,aes(x=OverallCond)) +
  geom_histogram()+
   geom_vline(xintercept=mean(data$OverallCond), color="black") +
  labs(title="Distribution of Overall Condition of Houses on a 1-10 Scale", x="Condition of House", y="
```

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




```
#Descriptive Statistics of OverallCond
summary(data$OverallCond)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   5.000   5.575   6.000   9.000
```

Most homes have a condition of 5. It seems like we should treat this as a categorical rather than numeric variable, since the difference between conditions is so abrupt

3. Explore Differences between Subsets

```
#Create subsets based on specific conditions
below_average_condition <- data %>% filter(OverallCond < 5)
average_condition <- data %>% filter(OverallCond == 5)
above_average_condition <- data %>% filter(OverallCond > 5)
```

```
#Set up a plot
ggplot() +
  geom_histogram(data=above_average_condition, aes(x=SalePrice, fill="above average condition"), alpha=
  geom_histogram(data=average_condition, aes(x=SalePrice, fill="average condition"), alpha=0.3) +
  geom_histogram(data=below_average_condition, aes(x=SalePrice, fill="below average condition"), alpha=
  scale_x_continuous(labels = comma) +
  labs(title="Distributions of Sale Price Grouped by Condition", x="Sale Price", y="Number of Houses",
  guides(fill = guide_legend(reverse=TRUE)) +
  scale_fill_manual(values=c("cyan", "gray", "yellow"), labels=c("above average condition", "average co
```

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

First, we note again that the majority of the houses have average condition, then about 1/3 have above average condition, then less than 10% have below average condition.

As we might expect, the average condition therefore contains houses across a broader spectrum of the sale price range than either the below-average or above-average houses.

Another unsurprising finding is that below-average condition houses have a price distribution that is much lower than average or above-average condition houses.

But what might be surprising is that above-average condition houses do not seem to have higher average sale prices than average condition houses. In fact, above-average condition houses seem more clustered around a particular price range, especially the $100,000 to $200,000 range, whereas average condition houses are more frequent above $200,000. We might want to investigate further to understand what kinds of houses are rated as above-average condition, since this goes against a standard assumption that better condition would mean higher cost.

4. Explore Correlations

The most positively correlated with SalePrice

```
#Extract numeric columns from the dataset
numeric_cols <- sapply(data, is.numeric)
numeric_cols
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##          TRUE          TRUE         FALSE          TRUE          TRUE
##        Street         Alley      LotShape   LandContour     Utilities
##         FALSE         FALSE         FALSE         FALSE         FALSE
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##         FALSE         FALSE         FALSE         FALSE         FALSE
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##         FALSE         FALSE          TRUE          TRUE          TRUE
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##          TRUE         FALSE         FALSE         FALSE         FALSE
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##         FALSE          TRUE         FALSE         FALSE         FALSE
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##         FALSE         FALSE         FALSE         FALSE          TRUE
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##         FALSE          TRUE          TRUE          TRUE         FALSE
##     HeatingQC    CentralAir    Electrical      1stFlrSF      2ndFlrSF
##         FALSE         FALSE         FALSE          TRUE          TRUE
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##          TRUE          TRUE          TRUE          TRUE          TRUE
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##          TRUE          TRUE          TRUE         FALSE          TRUE
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##         FALSE          TRUE         FALSE         FALSE          TRUE
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##         FALSE          TRUE          TRUE         FALSE         FALSE
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch     3SsnPorch
##         FALSE          TRUE          TRUE          TRUE          TRUE
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##          TRUE          TRUE         FALSE         FALSE         FALSE
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##          TRUE          TRUE          TRUE         FALSE         FALSE
##     SalePrice
##          TRUE
```

```
data_numeric <- data[, numeric_cols]
data_numeric
```

```
## # A tibble: 1,460 x 38
##       Id MSSub~1 LotFr~2 LotArea Overa~3 Overa~4 YearB~5 YearR~6 MasVn~7 BsmtF~8
##    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1     1      60      65    8450       7       5    2003    2003     196     706
##  2     2      20      80    9600       6       8    1976    1976       0     978
##  3     3      60      68   11250       7       5    2001    2002     162     486
##  4     4      70      60    9550       7       5    1915    1970       0     216
##  5     5      60      84   14260       8       5    2000    2000     350     655
##  6     6      50      85   14115       5       5    1993    1995       0     732
##  7     7      20      75   10084       8       5    2004    2005     186    1369
##  8     8      60      NA   10382       7       6    1973    1973     240     859
##  9     9      50      51    6120       7       5    1931    1950       0       0
## 10    10     190      50    7420       5       6    1939    1950       0     851
## # ... with 1,450 more rows, 28 more variables: BsmtFinSF2 <dbl>,
## #   BsmtUnfSF <dbl>, TotalBsmtSF <dbl>, `1stFlrSF` <dbl>, `2ndFlrSF` <dbl>,
## #   LowQualFinSF <dbl>, GrLivArea <dbl>, BsmtFullBath <dbl>,
```

```
## #   BsmtHalfBath <dbl>, FullBath <dbl>, HalfBath <dbl>, BedroomAbvGr <dbl>,
## #   KitchenAbvGr <dbl>, TotRmsAbvGrd <dbl>, Fireplaces <dbl>,
## #   GarageYrBlt <dbl>, GarageCars <dbl>, GarageArea <dbl>, WoodDeckSF <dbl>,
## #   OpenPorchSF <dbl>, EnclosedPorch <dbl>, `3SsnPorch` <dbl>, ...
```

```
# Get a list of correlations with SalePrice, sorted from smallest to largest
correlation_series <- sort(cor(data_numeric)[,'SalePrice'])

# Select second to last correlation, since the highest (last)
# correlation will be SalePrice correlating 100% with itself
max_corr_value <- correlation_series[length(correlation_series) - 1]
max_corr_column <- names(correlation_series)[length(correlation_series) - 1]

# Print the most positively correlated column and its maximum correlation value
cat("Most Positively Correlated Column: ", max_corr_column, "\n")
```

```
## Most Positively Correlated Column:  OverallQual
```

```
cat("Maximum Correlation Value: ", max_corr_value, "\n")
```

```
## Maximum Correlation Value:  0.7909816
```

The most negatively correlated with SalePrice

```
# Print the most negatively correlated column and its minimum correlation value
min_corr_value <- correlation_series[1]
min_corr_column <- names(correlation_series)[1]

cat("Most Negatively Correlated Column: ", min_corr_column, "\n")
```

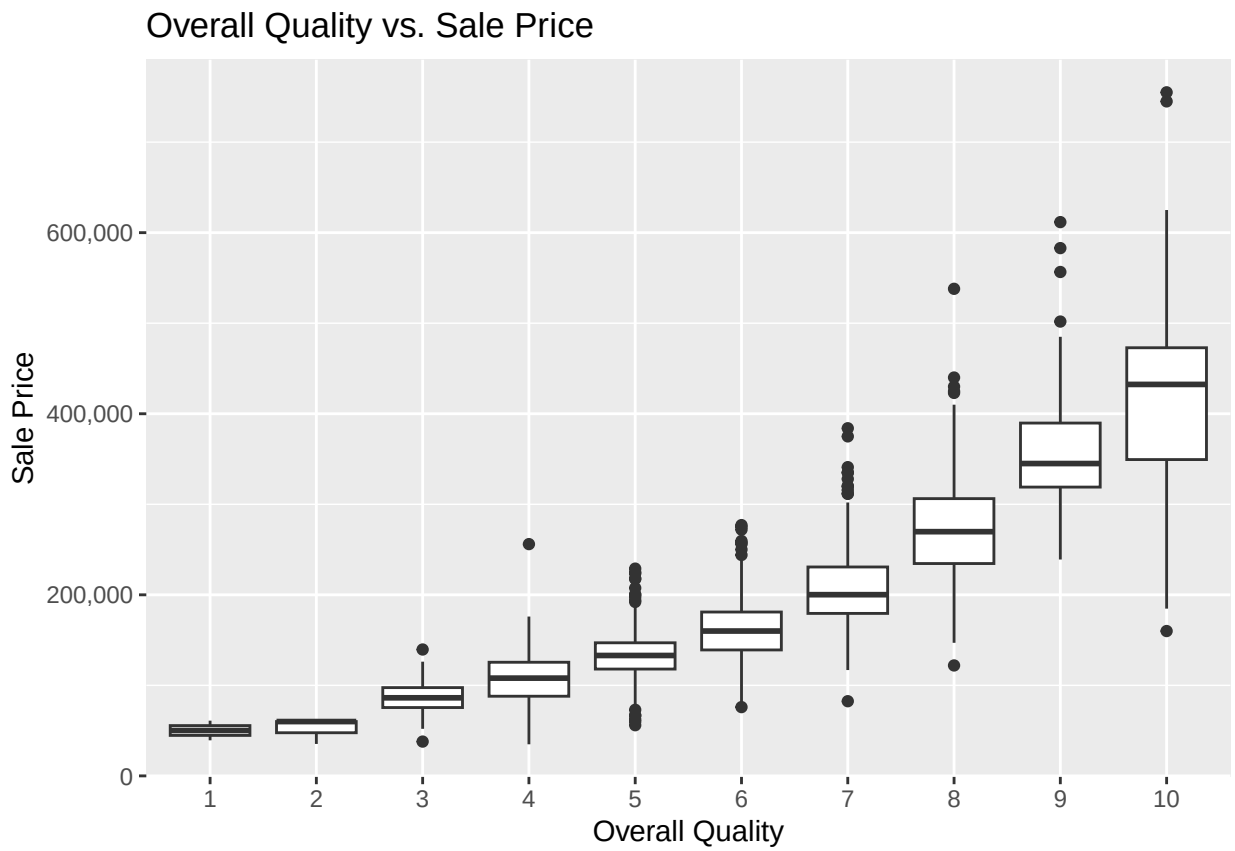
```
## Most Negatively Correlated Column:  KitchenAbvGr
```

```
cat("Minimum Correlation Value: ", min_corr_value, "\n")
```

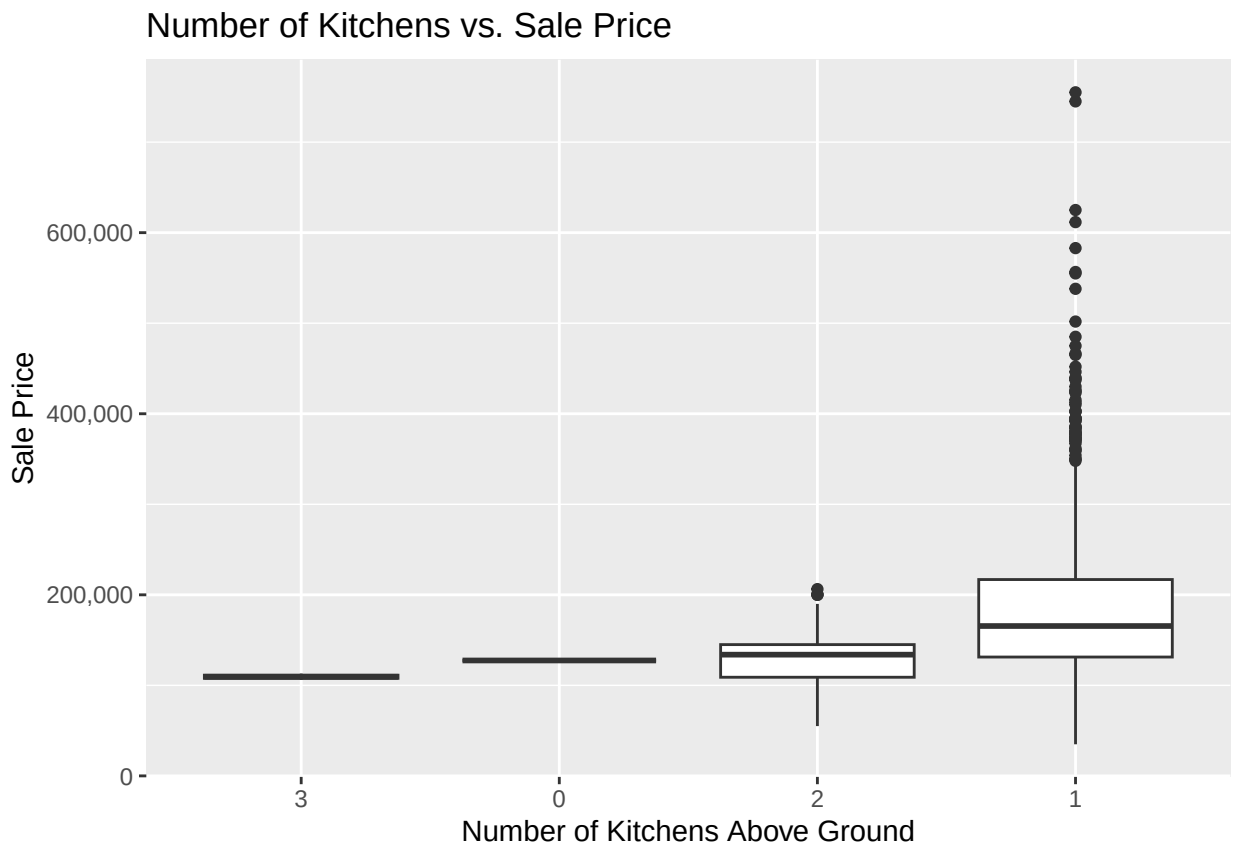
```
## Minimum Correlation Value:  -0.1359074
```

Once we have enough information, we creates box-plots of the relevant columns

```
# Plot distribution of column with highest correlation
ggplot(data, aes(x = reorder(as.character(OverallQual),SalePrice), y = SalePrice)) +
  geom_boxplot() +
  scale_y_continuous(labels = comma) +
  labs(title = "Overall Quality vs. Sale Price",
       x = "Overall Quality",
       y = "Sale Price")
```




```
# Plot distribution of column with most negative correlation
ggplot(data, aes(x = reorder(as.character(KitchenAbvGr),SalePrice), y = SalePrice)) +
  geom_boxplot() +
  scale_y_continuous(labels = comma) +
  labs(title = "Number of Kitchens vs. Sale Price",
       x = "Number of Kitchens Above Ground",
       y = "Sale Price")
```



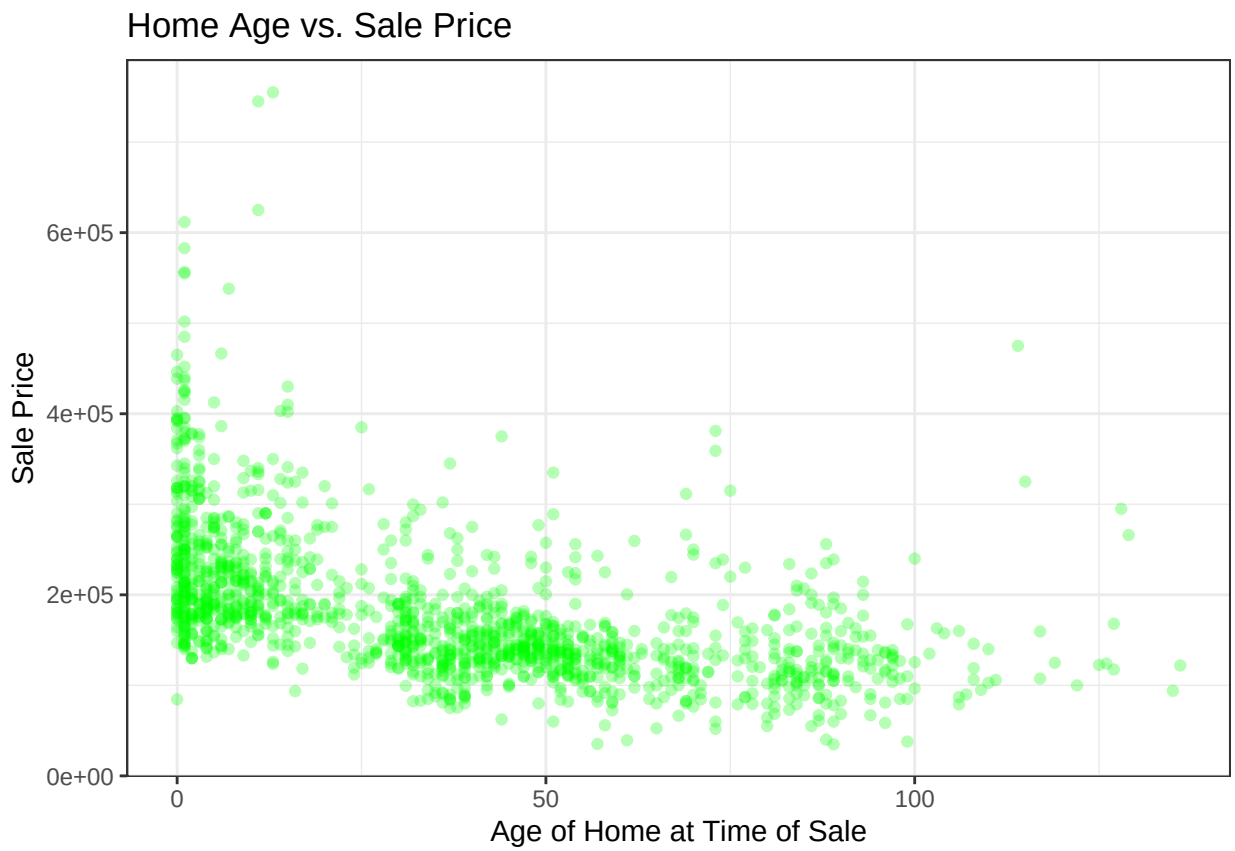
5. Engineer and Explore a New Feature

```
table(data$YrSold)
```

```
##
## 2006 2007 2008 2009 2010
##  314  329  304  338  175
```

```
#Create a new column
df <- data %>%
  mutate(Age = YrSold - YearBuilt)

#Set up a plot
ggplot(df, aes(x = Age, y = SalePrice)) +
  geom_point(alpha = 0.3, color = "green") +
  labs(title = "Home Age vs. Sale Price",
       x = "Age of Home at Time of Sale",
       y = "Sale Price") +
  theme_bw()
```



In general, newer houses appear to be more valuable, with value increasing as homes age. Interestingly the variance seems to increase once the home age goes over 100 years, with several above-average sale prices and fewer home sales in general.

We are also seeing potential housing booms and busts over the past decades, indicated by e.g. relatively few 20-year-old houses compared to 25-year-old houses being sold. We might find something interesting if we investigae this further.