Week 10 - Final Project Step 2

Ganesh Kale

May 23, 2021

Plan to complete final project:

1. Load the required packages
2. Load three data sets into three different data frames
3. Join all three data sets into one data set
4. Perform Exploratory Data Analysis -
    - 1. Check for missing data or na data in data frame and correct them.
    - 2. View the data summary and data structure - data types
    - 3. View descriptive statistics - min, max, std dev, median, count etc
    - 4. Plot the scatter charts to check the relationship among the features and possible outliers
    - 5. Plot the distribution charts for features to check data normality
    - 6. Plot Box chart to see outliers
    - 7. Fix the outliers from the data set based on the feature and data type
    - 8. Test for multicollinearity based on scatter plots
    - 9. Remove the highly correlated features and unwanted features from data set.
    - 10. Transform the data and or club nultiple columns to create one which makes more impact.
    - 11. Test for the correlation between features and target variable.
    - 12. Finalize the features for regression
    - 13. Transform the data if required for feature variables
    - 14. Create training, validation and testing data sets by using random or straitified sampling menthods
    - 15. Select the appropriate method of regression
    - 16. Run the regression model on training data set
    - 17. Evaluate the regression result summary
    - 18. Change or update features or transform data if required based on regression result and re-train the model
    - 19. Tune the model on validation data set and evaluate result summary
    - 20. Test the model on test data set and validate the result summary and accuracy of it
    - 21. Validate the assumption of regression based on result summary
    - 22. Plot the residual charts and fitted regression data set (prediction vs actual)
    - 23. Conclude the prediction based on result

Load the required packages

```
library(dplyr)
library(ggplot2)
```

Load three data sets in 3 different data frames

Austin TX housing main data set and first 5 records

```
house_main <- read.csv("project data/austinHousingData.csv")
head(house_main, 5)
```

```
##         zpid         city         streetAddress zipcode
## 1  111373431 pflugerville  14424 Lake Victor Dr   78660
## 2  120900430 pflugerville    1104 Strickling Dr   78660
## 3 2084491383 pflugerville   1408 Fort Dessau Rd   78660
## 4  120901374 pflugerville    1025 Strickling Dr   78660
## 5   60134862 pflugerville 15005 Donna Jane Loop   78660
##
## 1
## 2
## 3 Under construction - estimated completion in August 2019.  The Pioneer features an expansive open
## 4
## 5
##   latitude longitude propertyTaxRate garageSpaces hasAssociation hasCooling
## 1 30.43063 -97.66308            1.98            2           TRUE       TRUE
## 2 30.43267 -97.66170            1.98            2           TRUE       TRUE
## 3 30.40975 -97.63977            1.98            0           TRUE       TRUE
## 4 30.43211 -97.66166            1.98            2           TRUE       TRUE
## 5 30.43737 -97.65686            1.98            0           TRUE       TRUE
##   hasGarage hasHeating hasSpa hasView      homeType parkingSpaces yearBuilt
## 1      TRUE       TRUE  FALSE   FALSE Single Family             2      2012
## 2      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 3     FALSE       TRUE  FALSE   FALSE Single Family             0      2018
## 4      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 5     FALSE       TRUE  FALSE   FALSE Single Family             0      2002
##   latestPrice numPriceChanges latest_saledate latest_salemonth latest_saleyear
## 1      305000               5          9/2/19                9            2019
## 2      295000               1        10/13/20               10            2020
## 3      256125               1         7/31/19                7            2019
## 4      240000               4          8/8/18                8            2018
## 5      239900               3        10/31/18               10            2018
##                                 latestPriceSource lotSizeSqFt livingAreaSqFt
## 1 Coldwell Banker United, Realtors - South Austin        6011          2601
## 2                                  Agent Provided        6185          1768
## 3                                  Agent Provided        7840          1478
## 4                                  Agent Provided        6098          1678
## 5                                  Agent Provided        6708          2132
##   numOfBathrooms numOfBedrooms numOfStories
## 1              3             4            2
```

```
## 2             2             4           1
## 3             2             3           1
## 4             2             3           1
## 5             3             3           2
```

Austin TX housing area school infromation data set and first 5 records

```
house_school <-  read.csv("project data/austinHousingData_school_Info.csv")
head(house_school,5)
```

```
##         zpid numOfPrimarySchools numOfElementarySchools numOfMiddleSchools
## 1  111373431                   1                      0                  1
## 2  120900430                   1                      0                  1
## 3 2084491383                   0                      2                  1
## 4  120901374                   1                      0                  1
## 5   60134862                   1                      0                  1
##   numOfHighSchools avgSchoolDistance avgSchoolRating avgSchoolSize
## 1                1          1.266667        2.666667          1063
## 2                1          1.400000        2.666667          1063
## 3                1          1.200000        3.000000          1108
## 4                1          1.400000        2.666667          1063
## 5                1          1.133333        4.000000          1223
##   MedianStudentsPerTeacher
## 1                       14
## 2                       14
## 3                       14
## 4                       14
## 5                       14
```

Austin TX house features data set and first 5 records

```
house_features <- read.csv("project data/austinHousingData_features.csv")
head(house_features,5)
```

```
##         zpid numOfPhotos numOfAccessibilityFeatures numOfAppliances
## 1  111373431          39                          0               5
## 2  120900430          29                          0               1
## 3 2084491383           2                          0               4
## 4  120901374           9                          0               0
## 5   60134862          27                          0               0
##   numOfParkingFeatures numOfPatioAndPorchFeatures numOfSecurityFeatures
## 1                    2                          1                     3
## 2                    2                          0                     0
## 3                    1                          0                     1
## 4                    2                          0                     0
## 5                    1                          0                     0
##   numOfWaterfrontFeatures numOfWindowFeatures numOfCommunityFeatures
## 1                       0                   1                      0
## 2                       0                   0                      0
## 3                       0                   0                      0
```

```
## 4                       0                   0                       0
## 5                       0                   0                       0
##                                                homeImage
## 1  111373431_ffce26843283d3365c11d81b8e6bdc6f-p_f.jpg
## 2  120900430_8255c127be8dcf0a1a18b7563d987088-p_f.jpg
## 3 2084491383_a2ad649e1a7a098111dcea084a11c855-p_f.jpg
## 4  120901374_b469367a619da85b1f5ceb69b675d88e-p_f.jpg
## 5   60134862_b1a48a3df3f111e005bb913873e98ce2-p_f.jpg
```

Join all three datasets into one data set and display first 5 rows of final data set

```
temp <- house_main %>% inner_join(house_features, by = 'zpid')
house <- temp %>% inner_join(house_school, by = "zpid")
head(house, 5)
```

```
##         zpid         city         streetAddress zipcode
## 1  111373431 pflugerville  14424 Lake Victor Dr   78660
## 2  120900430 pflugerville    1104 Strickling Dr   78660
## 3 2084491383 pflugerville   1408 Fort Dessau Rd   78660
## 4  120901374 pflugerville    1025 Strickling Dr   78660
## 5   60134862 pflugerville 15005 Donna Jane Loop   78660
##
## 1
## 2
## 3 Under construction - estimated completion in August 2019.  The Pioneer features an expansive open
## 4
## 5
##   latitude longitude propertyTaxRate garageSpaces hasAssociation hasCooling
## 1 30.43063 -97.66308            1.98            2           TRUE       TRUE
## 2 30.43267 -97.66170            1.98            2           TRUE       TRUE
## 3 30.40975 -97.63977            1.98            0           TRUE       TRUE
## 4 30.43211 -97.66166            1.98            2           TRUE       TRUE
## 5 30.43737 -97.65686            1.98            0           TRUE       TRUE
##   hasGarage hasHeating hasSpa hasView      homeType parkingSpaces yearBuilt
## 1      TRUE       TRUE  FALSE   FALSE Single Family             2      2012
## 2      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 3     FALSE       TRUE  FALSE   FALSE Single Family             0      2018
## 4      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 5     FALSE       TRUE  FALSE   FALSE Single Family             0      2002
##   latestPrice numPriceChanges latest_saledate latest_salemonth latest_saleyear
## 1      305000               5          9/2/19                9            2019
## 2      295000               1        10/13/20               10            2020
## 3      256125               1         7/31/19                7            2019
## 4      240000               4          8/8/18                8            2018
## 5      239900               3        10/31/18               10            2018
##                                 latestPriceSource lotSizeSqFt livingAreaSqFt
## 1 Coldwell Banker United, Realtors - South Austin        6011           2601
## 2                                  Agent Provided        6185           1768
## 3                                  Agent Provided        7840           1478
## 4                                  Agent Provided        6098           1678
## 5                                  Agent Provided        6708           2132
##   numOfBathrooms numOfBedrooms numOfStories numOfPhotos
```

```
## 1              3             4            2          39
## 2              2             4            1          29
## 3              2             3            1           2
## 4              2             3            1           9
## 5              3             3            2          27
##   numOfAccessibilityFeatures numOfAppliances numOfParkingFeatures
## 1                          0               5                    2
## 2                          0               1                    2
## 3                          0               4                    1
## 4                          0               0                    2
## 5                          0               0                    1
##   numOfPatioAndPorchFeatures numOfSecurityFeatures numOfWaterfrontFeatures
## 1                          1                     3                       0
## 2                          0                     0                       0
## 3                          0                     1                       0
## 4                          0                     0                       0
## 5                          0                     0                       0
##   numOfWindowFeatures numOfCommunityFeatures
## 1                   1                     0
## 2                   0                     0
## 3                   0                     0
## 4                   0                     0
## 5                   0                     0
##                                          homeImage numOfPrimarySchools
## 1  111373431_ffce26843283d3365c11d81b8e6bdc6f-p_f.jpg                   1
## 2  120900430_8255c127be8dcf0a1a18b7563d987088-p_f.jpg                   1
## 3 2084491383_a2ad649e1a7a098111dcea084a11c855-p_f.jpg                   0
## 4  120901374_b469367a619da85b1f5ceb69b675d88e-p_f.jpg                   1
## 5   60134862_b1a48a3df3f111e005bb913873e98ce2-p_f.jpg                   1
##   numOfElementarySchools numOfMiddleSchools numOfHighSchools avgSchoolDistance
## 1                      0                  1                 1          1.266667
## 2                      0                  1                 1          1.400000
## 3                      2                  1                 1          1.200000
## 4                      0                  1                 1          1.400000
## 5                      0                  1                 1          1.133333
##   avgSchoolRating avgSchoolSize MedianStudentsPerTeacher
## 1        2.666667          1063                       14
## 2        2.666667          1063                       14
## 3        3.000000          1108                       14
## 4        2.666667          1063                       14
## 5        4.000000          1223                       14
```

Cleaning data by removing unwanted columns such as property dscription, city name, address, home images, number of photos etc.

```
house <- subset(house, select = -c(zpid,city,streetAddress,description,numOfPhotos,homeImage,latitude,l
```

View na or null values in data set

Based on the below, we can see there are no na or null values in the data set

```
sum(is.na(house))
```

```
## [1] 0
```

View the data summary and structure

Data Summary:

```
summary(house)
```

```
##     zipcode      propertyTaxRate  garageSpaces    hasAssociation
##  Min.   :78617   Min.   :1.980   Min.   : 0.000   Mode :logical
##  1st Qu.:78727   1st Qu.:1.980   1st Qu.: 0.000   FALSE:7164
##  Median :78739   Median :1.980   Median : 1.000   TRUE :8007
##  Mean   :78736   Mean   :1.994   Mean   : 1.229
##  3rd Qu.:78749   3rd Qu.:1.980   3rd Qu.: 2.000
##  Max.   :78759   Max.   :2.210   Max.   :22.000
##  hasCooling      hasGarage       hasHeating        hasSpa
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:274       FALSE:6825      FALSE:149       FALSE:13972
##  TRUE :14897     TRUE :8346      TRUE :15022     TRUE :1199
##
##
##
##   hasView          homeType         parkingSpaces      yearBuilt
##  Mode :logical   Length:15171       Min.   : 0.000   Min.   :1905
##  FALSE:11716     Class :character   1st Qu.: 0.000   1st Qu.:1974
##  TRUE :3455      Mode  :character   Median : 1.000   Median :1993
##                                     Mean   : 1.225   Mean   :1989
##                                     3rd Qu.: 2.000   3rd Qu.:2006
##                                     Max.   :22.000   Max.   :2020
##   latestPrice       numPriceChanges   lotSizeSqFt        livingAreaSqFt
##  Min.   :    5500   Min.   : 1.000   Min.   :1.000e+02   Min.   :   300
##  1st Qu.:  309000   1st Qu.: 1.000   1st Qu.:6.534e+03   1st Qu.:  1483
##  Median :  405000   Median : 2.000   Median :8.276e+03   Median :  1975
##  Mean   :  512768   Mean   : 3.033   Mean   :1.191e+05   Mean   :  2208
##  3rd Qu.:  575000   3rd Qu.: 4.000   3rd Qu.:1.089e+04   3rd Qu.:  2687
##  Max.   :13500000   Max.   :23.000   Max.   :1.508e+09   Max.   :109292
##  numOfBathrooms   numOfBedrooms    numOfStories   numOfAccessibilityFeatures
##  Min.   : 0.000   Min.   : 0.00   Min.   :1.000   Min.   :0.00000
##  1st Qu.: 2.000   1st Qu.: 3.00   1st Qu.:1.000   1st Qu.:0.00000
##  Median : 3.000   Median : 3.00   Median :1.000   Median :0.00000
##  Mean   : 2.683   Mean   : 3.44   Mean   :1.467   Mean   :0.01299
##  3rd Qu.: 3.000   3rd Qu.: 4.00   3rd Qu.:2.000   3rd Qu.:0.00000
##  Max.   :27.000   Max.   :20.00   Max.   :4.000   Max.   :8.00000
##  numOfAppliances  numOfParkingFeatures numOfPatioAndPorchFeatures
##  Min.   : 0.000   Min.   :0.00         Min.   :0.0000
##  1st Qu.: 2.000   1st Qu.:1.00         1st Qu.:0.0000
##  Median : 3.000   Median :2.00         Median :0.0000
##  Mean   : 3.475   Mean   :1.71         Mean   :0.6634
##  3rd Qu.: 4.000   3rd Qu.:2.00         3rd Qu.:1.0000
##  Max.   :12.000   Max.   :6.00         Max.   :8.0000
```

```
##  numOfSecurityFeatures numOfWaterfrontFeatures numOfWindowFeatures
##  Min.   :0.0000        Min.   :0.000000        Min.   :0.0000
##  1st Qu.:0.0000        1st Qu.:0.000000        1st Qu.:0.0000
##  Median :0.0000        Median :0.000000        Median :0.0000
##  Mean   :0.4669        Mean   :0.002768        Mean   :0.2085
##  3rd Qu.:1.0000        3rd Qu.:0.000000        3rd Qu.:0.0000
##  Max.   :6.0000        Max.   :2.000000        Max.   :4.0000
##  numOfCommunityFeatures numOfPrimarySchools numOfElementarySchools
##  Min.   :0.00000        Min.   :0.0000      Min.   :0.00000
##  1st Qu.:0.00000        1st Qu.:1.0000      1st Qu.:0.00000
##  Median :0.00000        Median :1.0000      Median :0.00000
##  Mean   :0.01885        Mean   :0.9407      Mean   :0.04917
##  3rd Qu.:0.00000        3rd Qu.:1.0000      3rd Qu.:0.00000
##  Max.   :8.00000        Max.   :2.0000      Max.   :2.00000
##  numOfMiddleSchools numOfHighSchools avgSchoolDistance avgSchoolRating
##  Min.   :0.000      Min.   :0.0000   Min.   :0.200     Min.   :2.333
##  1st Qu.:1.000      1st Qu.:1.0000   1st Qu.:1.100     1st Qu.:4.000
##  Median :1.000      Median :1.0000   Median :1.567     Median :5.779
##  Mean   :1.036      Mean   :0.9768   Mean   :1.838     Mean   :5.780
##  3rd Qu.:1.000      3rd Qu.:1.0000   3rd Qu.:2.267     3rd Qu.:7.000
##  Max.   :3.000      Max.   :2.0000   Max.   :9.000     Max.   :9.500
##  avgSchoolSize  MedianStudentsPerTeacher
##  Min.   : 396   Min.   :10.00
##  1st Qu.: 966   1st Qu.:14.00
##  Median :1287   Median :15.00
##  Mean   :1237   Mean   :14.86
##  3rd Qu.:1496   3rd Qu.:16.00
##  Max.   :1913   Max.   :19.00
```

Data Structure:

```
str(house)
```

```
## 'data.frame':    15171 obs. of  35 variables:
##  $ zipcode                   : int  78660 78660 78660 78660 78660 78660 78660 78660 78660 78617 ...
##  $ propertyTaxRate           : num  1.98 1.98 1.98 1.98 1.98 1.98 1.98 1.98 1.98 1.98 ...
##  $ garageSpaces              : int  2 2 0 2 0 2 0 0 0 2 ...
##  $ hasAssociation            : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ hasCooling                : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ hasGarage                 : logi  TRUE TRUE FALSE TRUE FALSE TRUE ...
##  $ hasHeating                : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ hasSpa                    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hasView                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ homeType                  : chr  "Single Family" "Single Family" "Single Family" "Single Family"
##  $ parkingSpaces             : int  2 2 0 2 0 2 0 0 0 2 ...
##  $ yearBuilt                 : int  2012 2013 2018 2013 2002 2020 2016 2002 2002 2013 ...
##  $ latestPrice               : int  305000 295000 256125 240000 239900 309045 315000 219900 225000 19
##  $ numPriceChanges           : int  5 1 1 4 3 2 2 2 1 1 ...
##  $ lotSizeSqFt               : num  6011 6185 7840 6098 6708 ...
##  $ livingAreaSqFt            : int  2601 1768 1478 1678 2132 1446 2432 1422 1870 1422 ...
##  $ numOfBathrooms            : num  3 2 2 2 3 2 3 3 2 3 ...
##  $ numOfBedrooms             : int  4 4 3 3 3 3 4 3 3 3 ...
##  $ numOfStories              : int  2 1 1 1 2 1 2 2 2 2 ...
```

```
##  $ numOfAccessibilityFeatures: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ numOfAppliances           : int  5 1 4 0 0 3 3 3 2 3 ...
##  $ numOfParkingFeatures      : int  2 2 1 2 1 1 1 1 1 2 ...
##  $ numOfPatioAndPorchFeatures: int  1 0 0 0 0 2 0 0 1 0 ...
##  $ numOfSecurityFeatures     : int  3 0 1 0 0 2 0 0 1 0 ...
##  $ numOfWaterfrontFeatures   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ numOfWindowFeatures       : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ numOfCommunityFeatures    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ numOfPrimarySchools       : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ numOfElementarySchools    : int  0 0 2 0 0 0 0 0 0 0 ...
##  $ numOfMiddleSchools        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ numOfHighSchools          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ avgSchoolDistance         : num  1.27 1.4 1.2 1.4 1.13 ...
##  $ avgSchoolRating           : num  2.67 2.67 3 2.67 4 ...
##  $ avgSchoolSize             : int  1063 1063 1108 1063 1223 1223 1051 1223 1223 1615 ...
##  $ MedianStudentsPerTeacher  : int  14 14 14 14 14 14 12 14 14 14 ...
```

View of the data set - sample 10 records

Sample 10 records from the house data frame:

```
house %>% sample_n(10)
```

```
##    zipcode propertyTaxRate garageSpaces hasAssociation hasCooling hasGarage
## 1    78754            1.98            0           TRUE       TRUE     FALSE
## 2    78749            1.98            0          FALSE       TRUE     FALSE
## 3    78729            2.21            2           TRUE       TRUE      TRUE
## 4    78741            1.98            0          FALSE       TRUE     FALSE
## 5    78746            1.98            2          FALSE       TRUE      TRUE
## 6    78732            1.98            3           TRUE       TRUE      TRUE
## 7    78744            1.98            0          FALSE       TRUE     FALSE
## 8    78745            1.98            0          FALSE       TRUE     FALSE
## 9    78747            1.98            2           TRUE       TRUE      TRUE
## 10   78745            1.98            0          FALSE      FALSE     FALSE
##    hasHeating hasSpa hasView      homeType parkingSpaces yearBuilt latestPrice
## 1        TRUE  FALSE   FALSE Single Family             0      2012      229995
## 2        TRUE  FALSE   FALSE Single Family             0      1986      330000
## 3        TRUE  FALSE   FALSE Single Family             2      1998      454000
## 4        TRUE  FALSE   FALSE Single Family             0      1958      225000
## 5        TRUE  FALSE    TRUE Single Family             2      1980      455000
## 6        TRUE   TRUE   FALSE Single Family             3      2005      849900
## 7        TRUE  FALSE    TRUE Single Family             0      1974      165000
## 8        TRUE  FALSE   FALSE Single Family             0      1979      279900
## 9        TRUE  FALSE   FALSE Single Family             2      2005      269000
## 10      FALSE  FALSE   FALSE         Condo             0      2008      309000
##    numPriceChanges lotSizeSqFt livingAreaSqFt numOfBathrooms numOfBedrooms
## 1                1      4356.0           1346            2.0             3
## 2                2      6926.0           1576            3.0             3
## 3                3      7318.0           2776            3.0             4
## 4                1      8712.0           1560            2.0             4
## 5                6     12196.8           2877            2.0             4
## 6                2     18295.2           4135            4.0             4
## 7                1      6446.0           1321            2.0             3
```

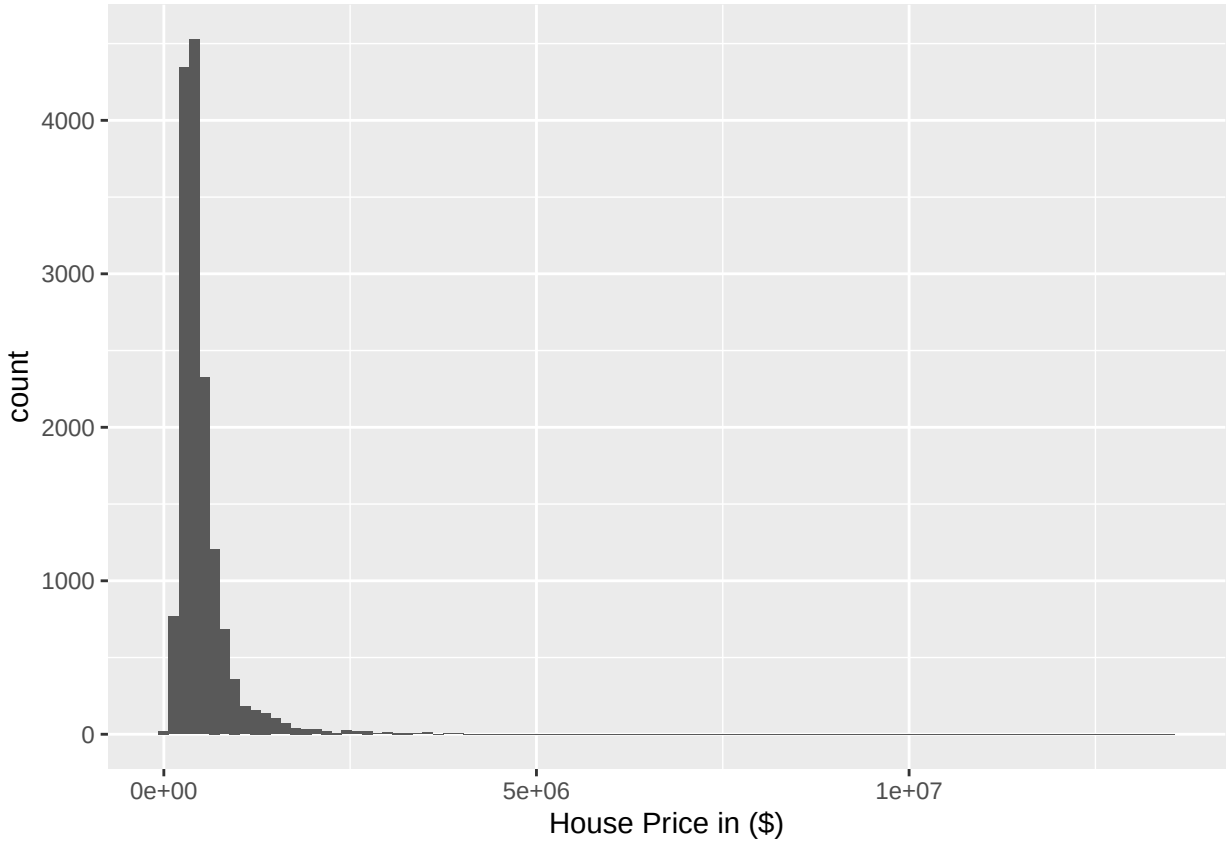
```
## 8               2      7840.0            999           2.0             3
## 9               2      3615.0           2068           3.0             3
## 10              2      8407.0           1650           2.5             2
##    numOfStories numOfAccessibilityFeatures numOfAppliances numOfParkingFeatures
## 1             1                         0               4                    1
## 2             2                         0               3                    1
## 3             2                         0               4                    3
## 4             1                         0               0                    0
## 5             2                         0               2                    3
## 6             2                         0               8                    2
## 7             1                         0               7                    1
## 8             1                         0               5                    1
## 9             2                         0               1                    2
## 10            1                         0               0                    0
##    numOfPatioAndPorchFeatures numOfSecurityFeatures numOfWaterfrontFeatures
## 1                           0                     0                      0
## 2                           2                     1                      0
## 3                           0                     1                      0
## 4                           0                     0                      0
## 5                           2                     0                      0
## 6                           0                     0                      0
## 7                           0                     0                      0
## 8                           0                     0                      0
## 9                           1                     1                      0
## 10                          0                     0                      0
##    numOfWindowFeatures numOfCommunityFeatures numOfPrimarySchools
## 1                    0                     0                   1
## 2                    0                     0                   1
## 3                    0                     0                   1
## 4                    0                     0                   1
## 5                    1                     0                   0
## 6                    0                     0                   1
## 7                    0                     0                   1
## 8                    0                     0                   1
## 9                    0                     0                   1
## 10                   0                     0                   1
##    numOfElementarySchools numOfMiddleSchools numOfHighSchools avgSchoolDistance
## 1                       0                  1                2          3.600000
## 2                       0                  1                1          1.066667
## 3                       0                  1                1          1.266667
## 4                       0                  1                1          3.433333
## 5                       1                  1                0          0.400000
## 6                       0                  1                1          2.633333
## 7                       0                  1                1          1.566667
## 8                       0                  1                1          0.800000
## 9                       0                  1                1          2.966667
## 10                      0                  1                1          1.600000
##    avgSchoolRating avgSchoolSize MedianStudentsPerTeacher
## 1         4.333333          1066                      12
## 2         6.666667          1460                      16
## 3         5.666667          1402                      12
## 4         3.333333          1561                      13
## 5         9.000000          1600                      14
## 6         8.333333          1533                      17
```

```
## 7         3.333333          1317                      14
## 8         3.333333           926                      13
## 9         5.333333          1506                      15
## 10        3.333333           792                      13
```

Visualization:

'latestPrice' is the target variable, we are going to predict based on the predictors. View the distribution of target variable. Based on histogram it seems the target variable is not normally distributed and there are outliers that is higher house prices but house counts are less.

```
ggplot(data = house, aes(latestPrice)) + geom_histogram(bins = 100) + xlab("House Price in ($)")
```



Questions for future steps.

Based on the data set, there are several features which may or may not have direct impact on sale price. In order to choose the jet predictors, need to find the suitable regression method to onley keep relevant predictors.

1. What regression method is suitable to choose the predictors?
2. Which variables need data transformation?
3. How to handle biased data ?

4. Which features will give more sense after clubbing together?
5. What is school size and how to interpret those numbers?

What information is not self-evident?

The school size feature does not give much information about impact on house price or area. There house features, which explains the how many appliances or features avaible in house but number does not give exact picture on what are those appliances. Also property tax rate does not give information about those are current rates or the time of the data recorded or when last tax was paid by house owner.

What are different ways you could look at this data?

The data is explored using statistical methods and using visualizations. To see the basic stats, data summary will be used that provides information about abasic stats. If the linear regression is not the right choice then how can non-linear regression or any other regression used to solve the business problem. Different models will be used to check which one is producing the better result or compared the results of the models.

How do you plan to slice and dice the data?

The detailed plan is mentioned in the starting of this document.

How could you summarize your data to answer key questions?

The main problem is to predict the house price, in order to tackle this problem I am going to find answers to below questions - 1. Identify the variables from the data sets affecting the house prices, e,g, bedrooms, lotsize, yearbuilt, school zone etc. 2. Train the multiple regression model that quantitatively relates house prices with predictors. 3. To calculate the accuracy of the model to see how goo fit it is?

What types of plots and tables will help you to illustrate the findings to your questions?

For checking the distribution of the data , I am going to use histograms, histograms are also show the outliers in the distribution.Scatter plot to check multicollinearity among the predictors and correlation between target and predictor variables, this plot is also helpful to see how regression line fitted to actual data versus predicted data. Box plot will help identify outliers easily from the features.

1. Scatter plot - To see the correlation between two variables and possible outliers
2. Histogram - To check data distribution for the different features
3. Q-Q plot - To check standard normal distribution
4. correlation matrix - To check the correlation of all the features from the data set
5. Boxplot - To check outliers and quantiles
6. Density plot - To check the distribution
7. Bar Chart - Data distribution of categorical variables.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

To solve the housing price prediction problem, I am going to use machine learning techniques. Basically there are three machine learning techniques, out of three the first which is Supervised Machine Learning, I am going to incorporate to solve this problem. The reason I am going to use Supervised machine Learning because this technique is used to predict the outcomes accurately and in data set the outcome variable already exist that means we have labels ready and supervised machine learning is basically used on labelled datasets.