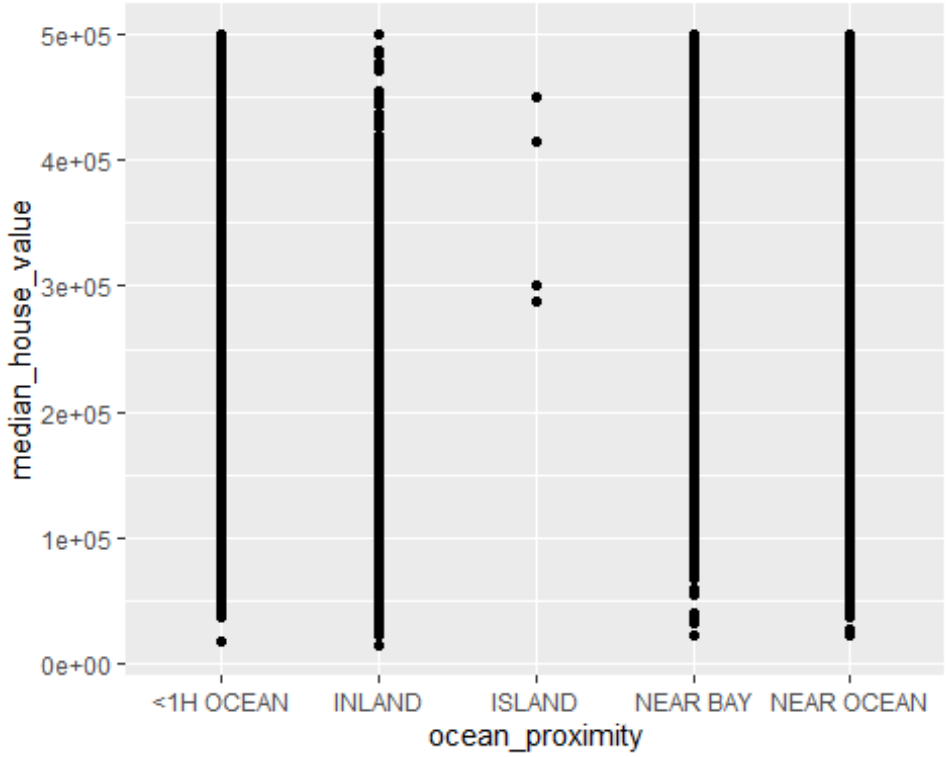
I have got three datasets, that contain data points in different formats.

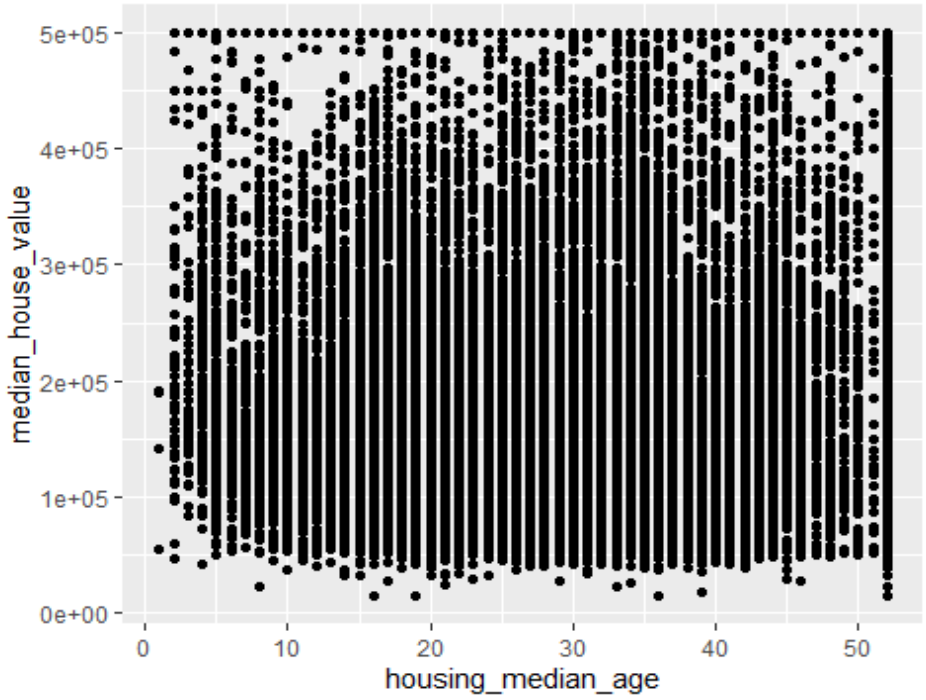
- ⇨ London Housing prices dataset has just about 11 column variables and some of those variables do not seem to have relation for being picked as predictor (Sequence ID, Property name, Blank / Invalid Location values and in some cases consist of partial street address). Also, the sale prices are in Pounds, which may not be relevant. So decided not to use the dataset.

- ⇨ California Housing prices dataset has 10 columns including median house value and population, ocean proximity etc. there was not much significance and data was almost equally distributed. Thus not enough variables available for analysis. Also, when looked at the plots between. The total number of rooms, bedrooms data is not in standard format and would skew other dataset info, if merged with it. Hence decided not to use this dataset.

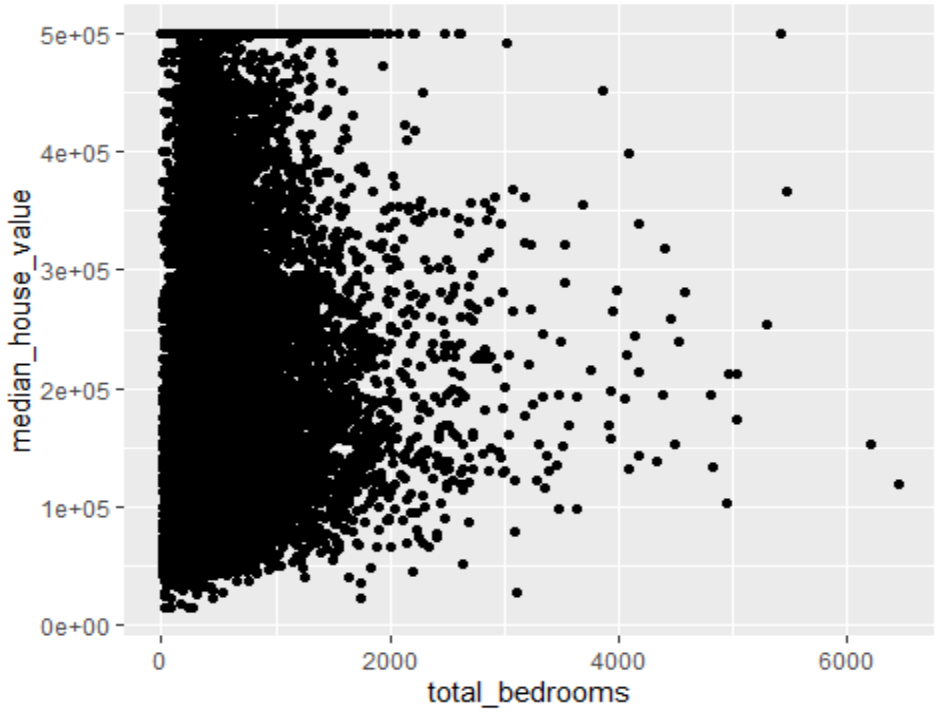
  Median House value price ranges are similar in each of the ocean proximity category.



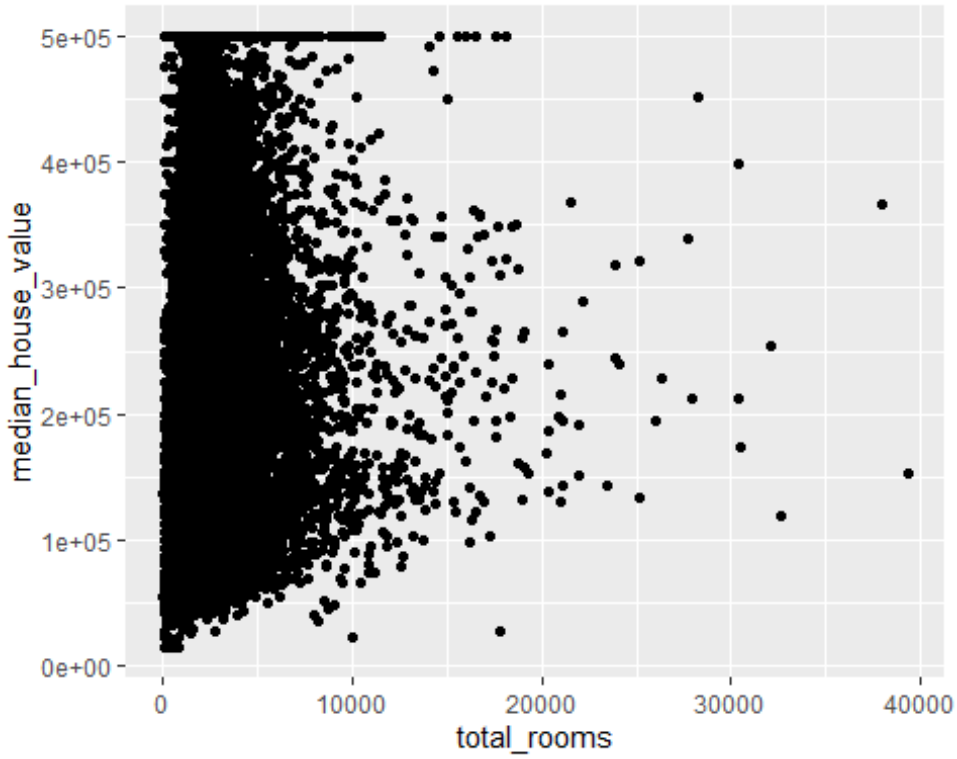
All ranges of Housing median age as well seem to have median house values price ranges.



Total bedrooms variable does not have standard set of values and many data points are clustered between 0 thru 1800 and fall into all the price range categories.



Similar case with total rooms category where, Total rooms variable does not have standard set of values and many data points are clustered between 0 thru 10000 and fall into all the price range categories.

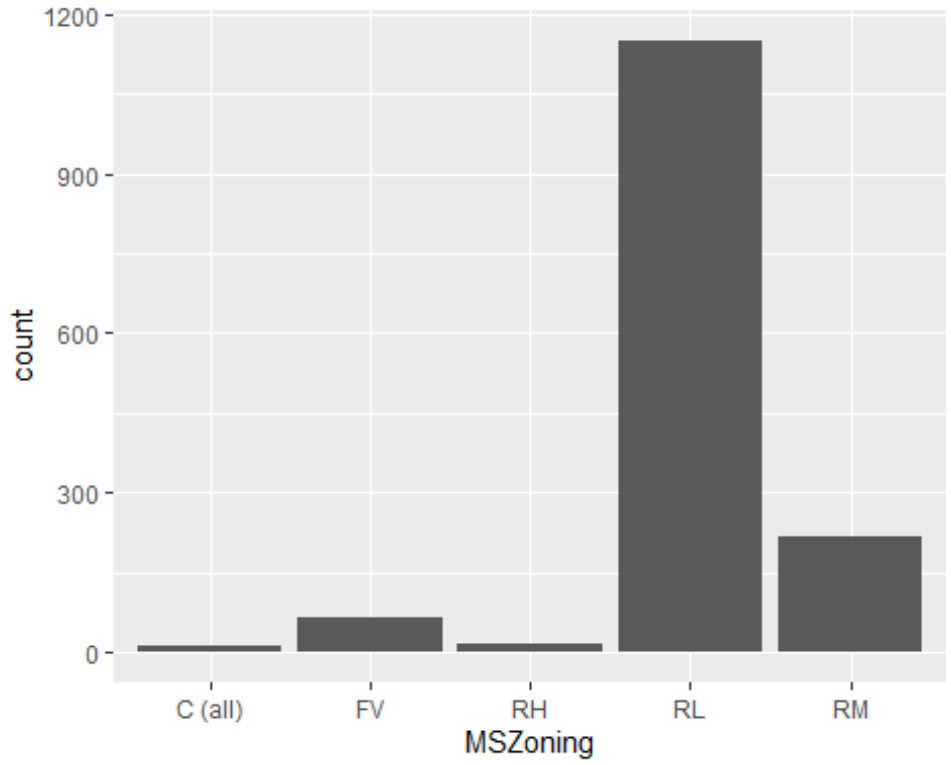


- US Housing prices dataset has plenty of variables (81 in all) and hence I will be using this dataset for my analysis purpose. I have manually looked at the data variables and also used some of the plots to understand the data points / variables. I have tried to capture this information below. Noticed that TotalBsmtSF is the addition of BsmtFinSF1, BsmtFinSF2 and BsmtUnfSF. Similarly 1stFlrSF and 2ndFlrSF columns values are combined into and it is present in variable GrLivArea.

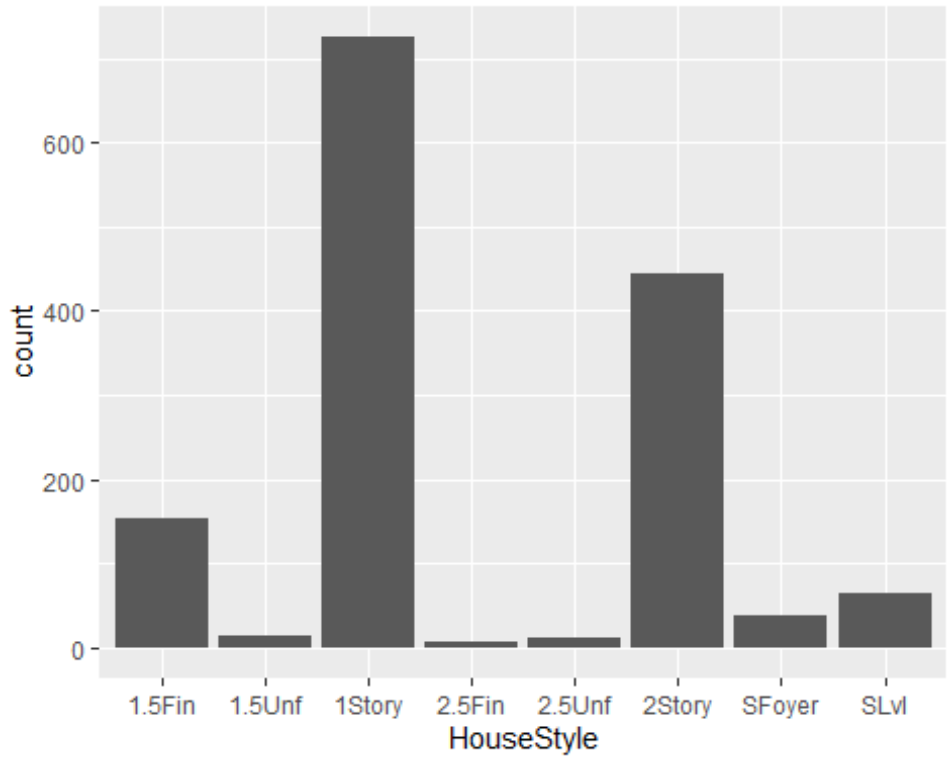
Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

- First I imported the entire US Housing prices dataset as is and used summary() on the data frame. Looking at the data manually and some of the plots for variables considered, below plots were observed.

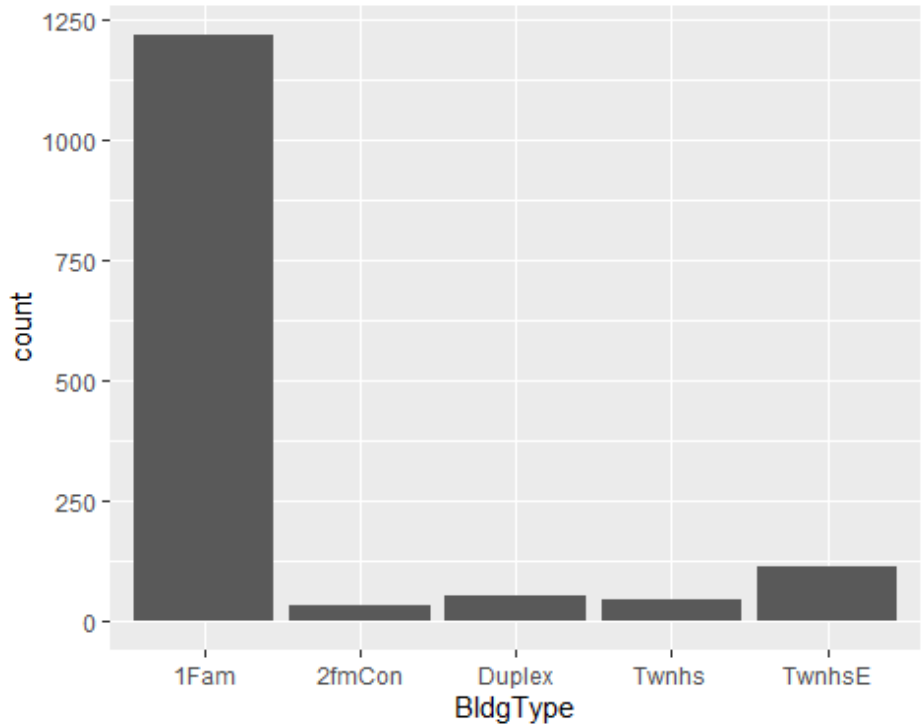
MS Zoning variable Histogram. Most of the values are in RL category and some RM values.



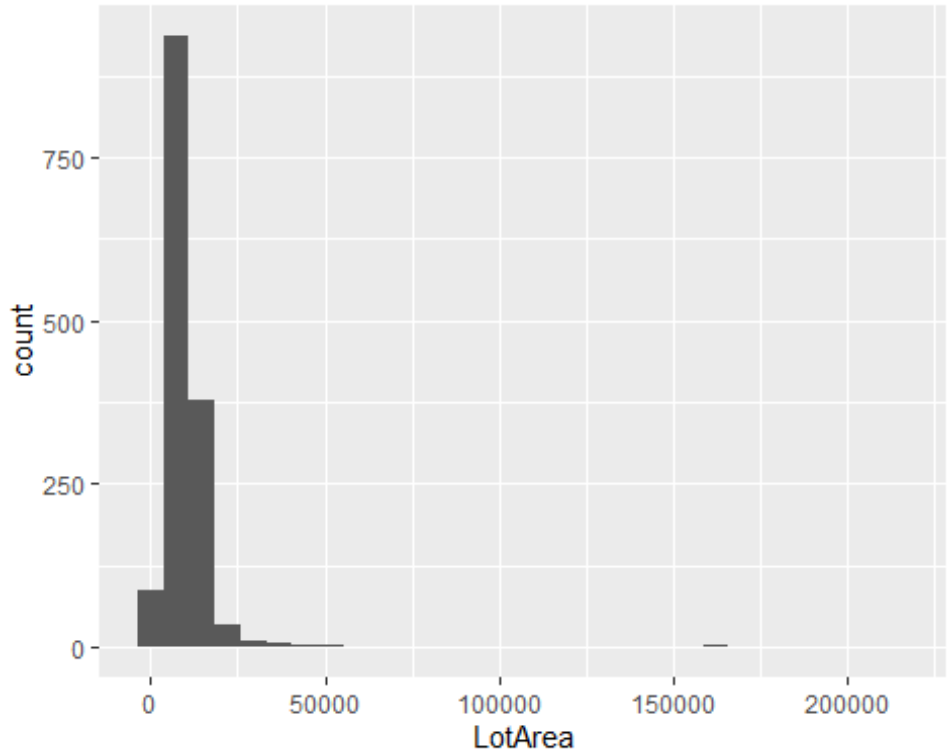
Majority of the houses are 1Story and 2Story houses, while next is 1.5Fin category is major.



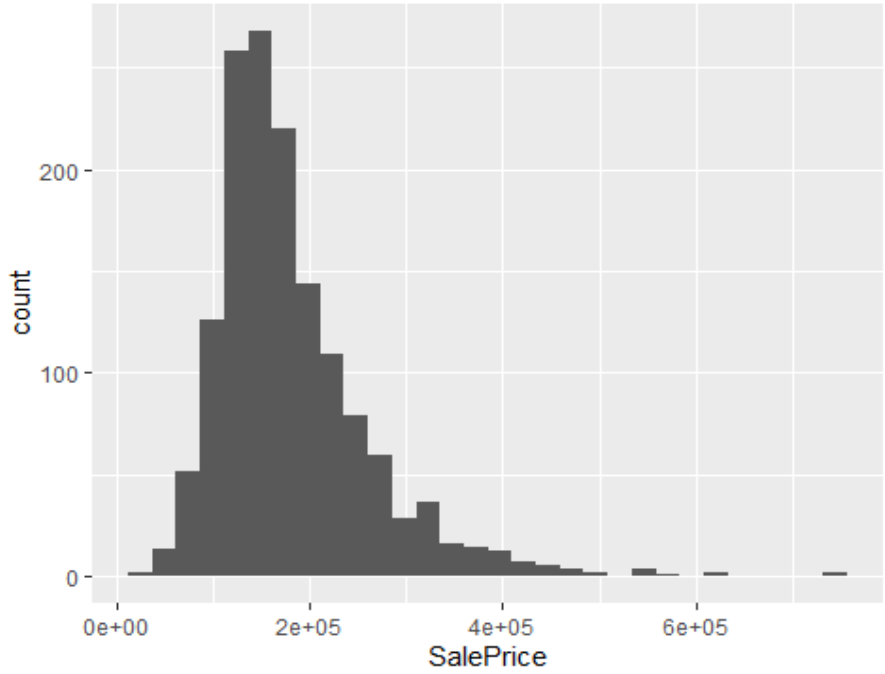
BldgType variable shows most of the houses being 1Family.



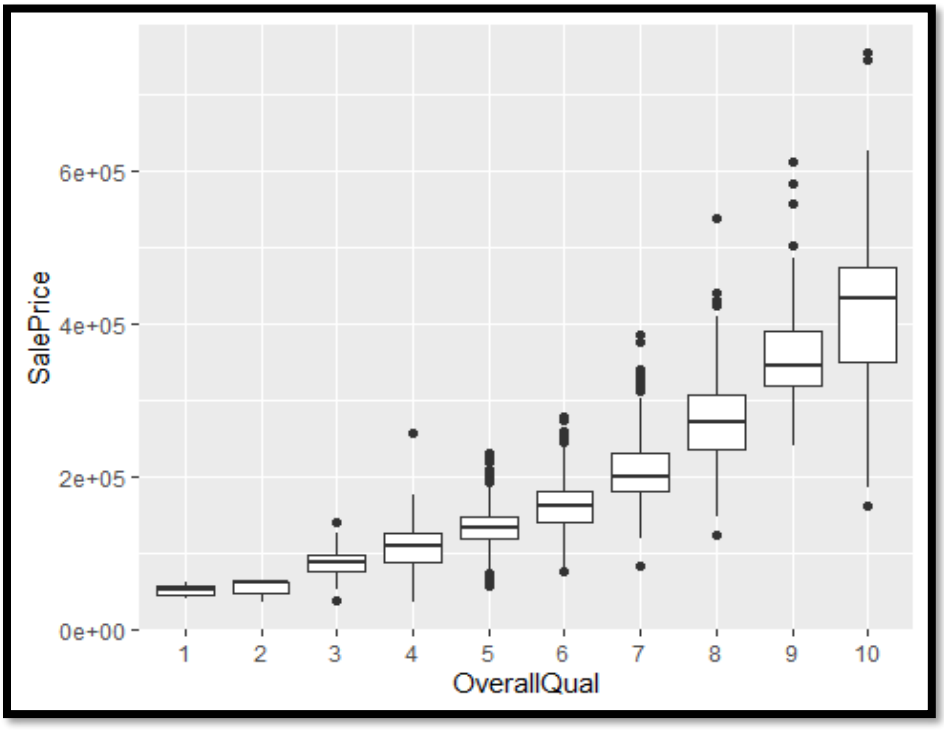
Looking at LotArea Histogram, there are outliers and range beyond 25000 sqft – 26000 sqft could be omitted.



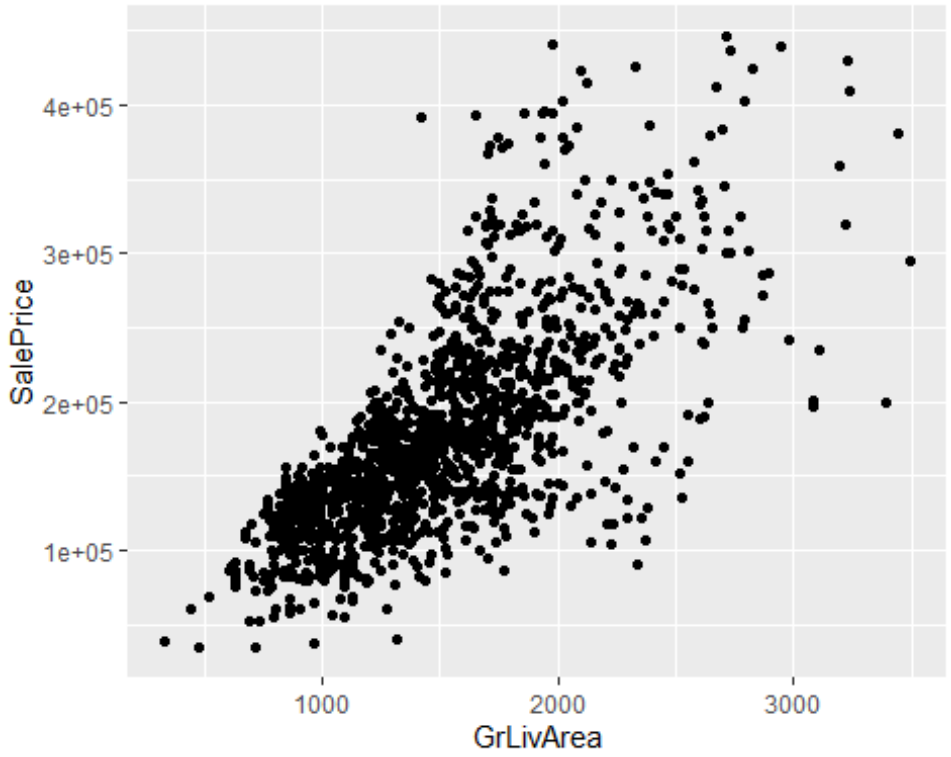
Looking at Histogram of Sale Prices, we see there are some outliers in the Sale prices, So I am planning to drop values above $450,000

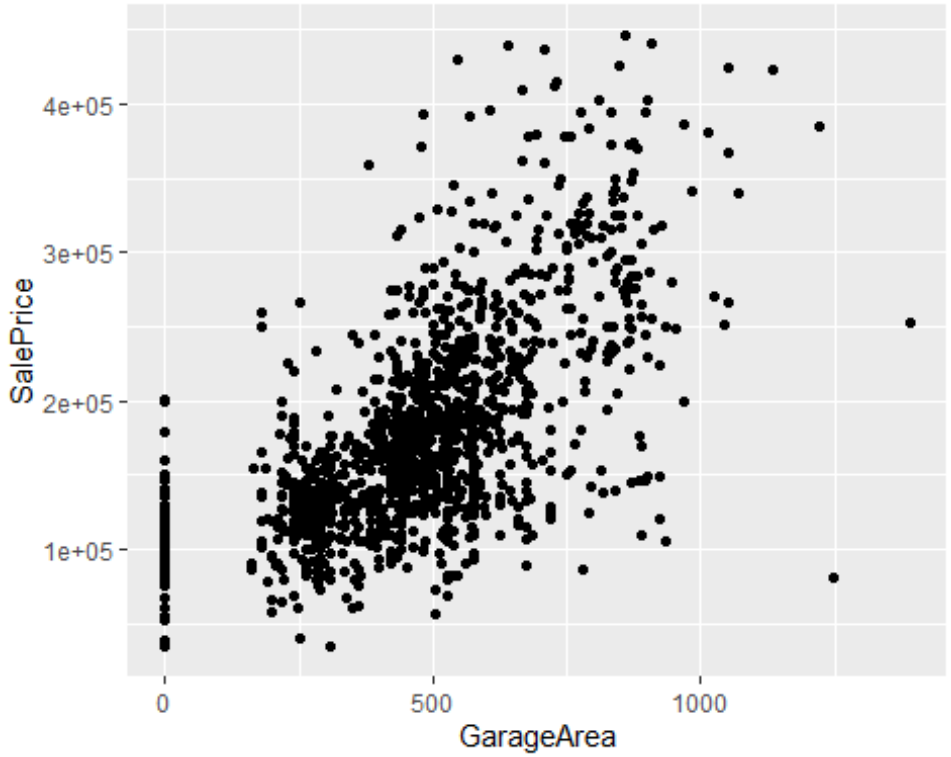


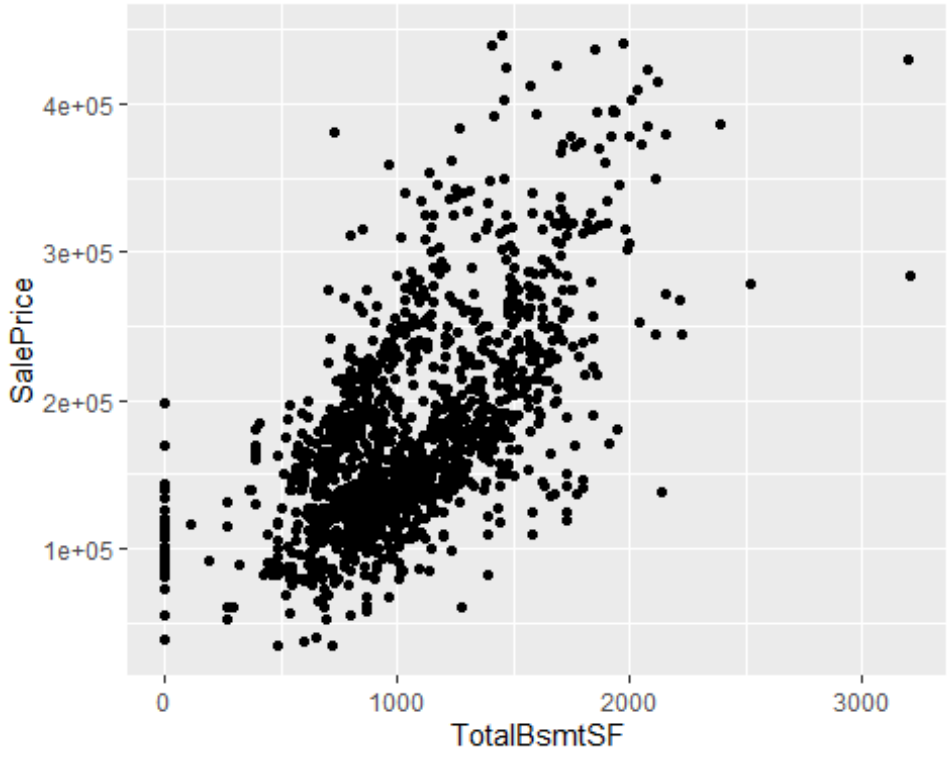
Overall Quality variable, shows very close correlation with Sale Prices.

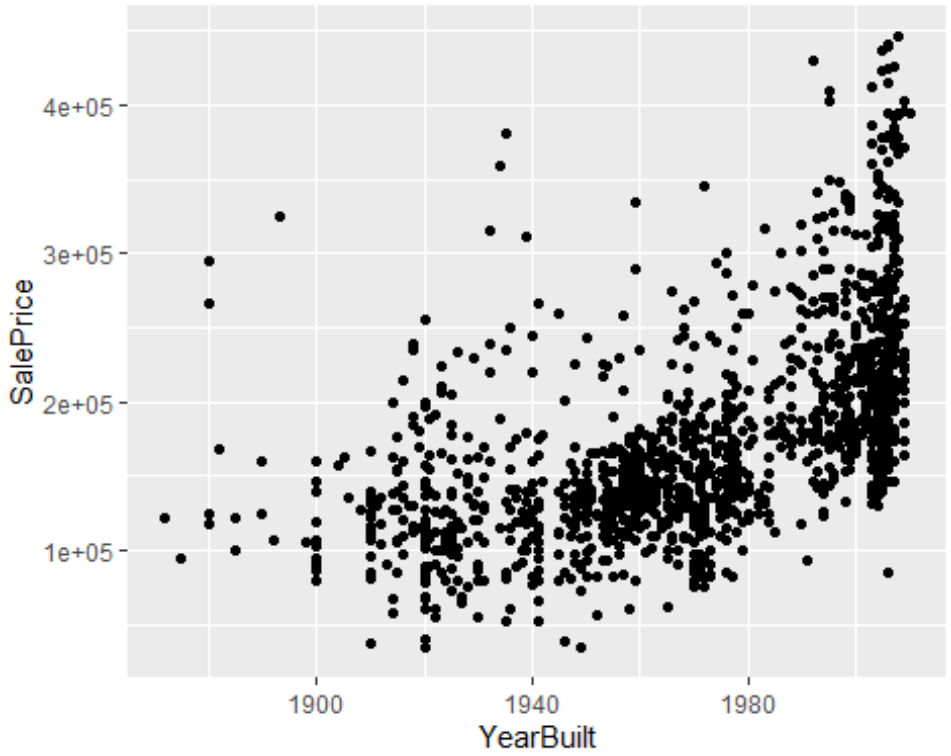


Other variables such as GrLivingArea, GarageArea, TotalBsmtSF and YearBuilt have shown good amount of correlation with Sale Price, looking below scatter plots. So, I will be using them.






SalePrice
4e+05
3e+05
2e+05
1e+05
0
1000
2000
3000
TotalBsmtSF

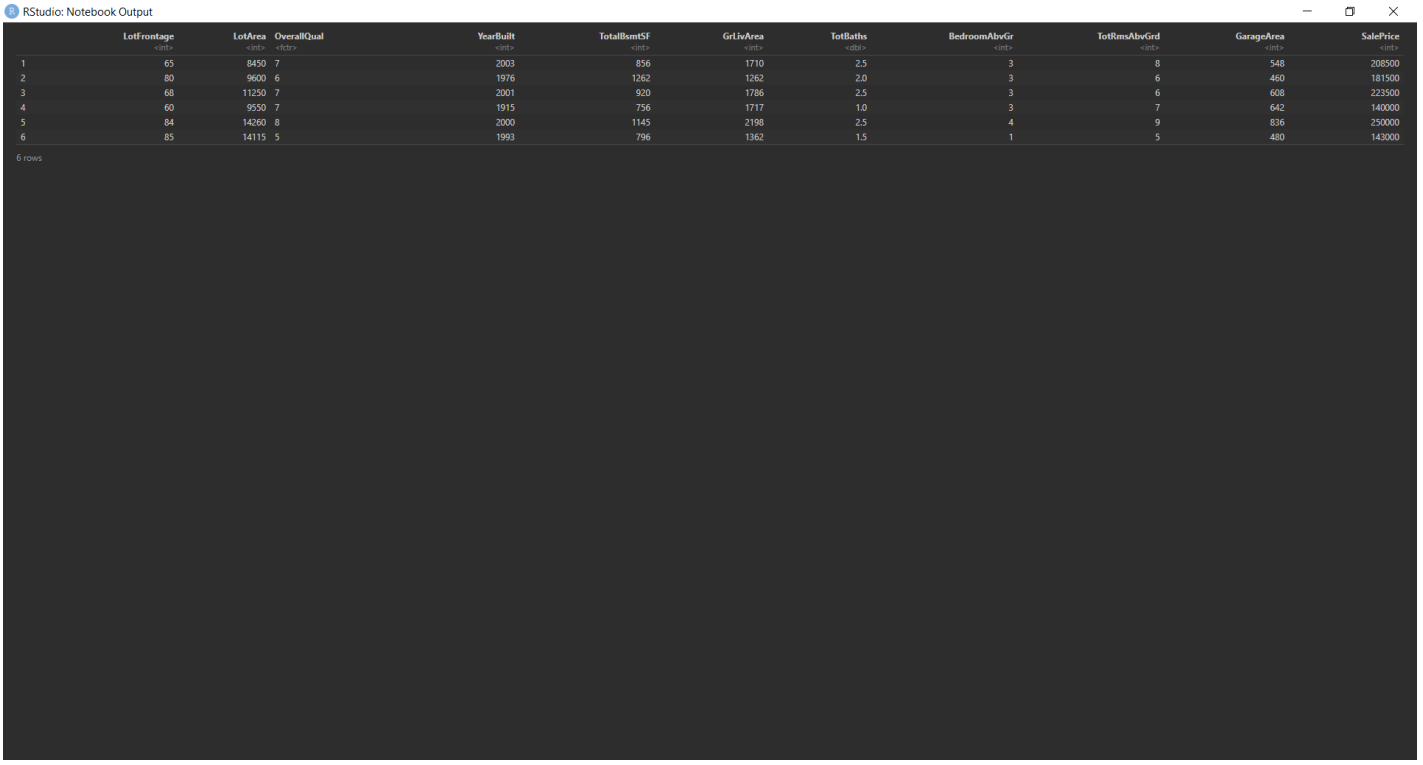

SalePrice
4e+05
3e+05
2e+05
1e+05
1900
1940
1980
YearBuilt

Also, majority of the houses have 1 Kitchen (more than 95%). Hence did not consider Kitchen Above Ground (KitchenAbvGr) variable.

From general experience, total number of bedrooms, bathrooms (full and half) and overall total number of rooms tend to affect the housing prices as well. So, I have kept them in new subset data frame to be continued for analysis.

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

- ⇨ I have used head() command on the new subset data frame for the condensed data.

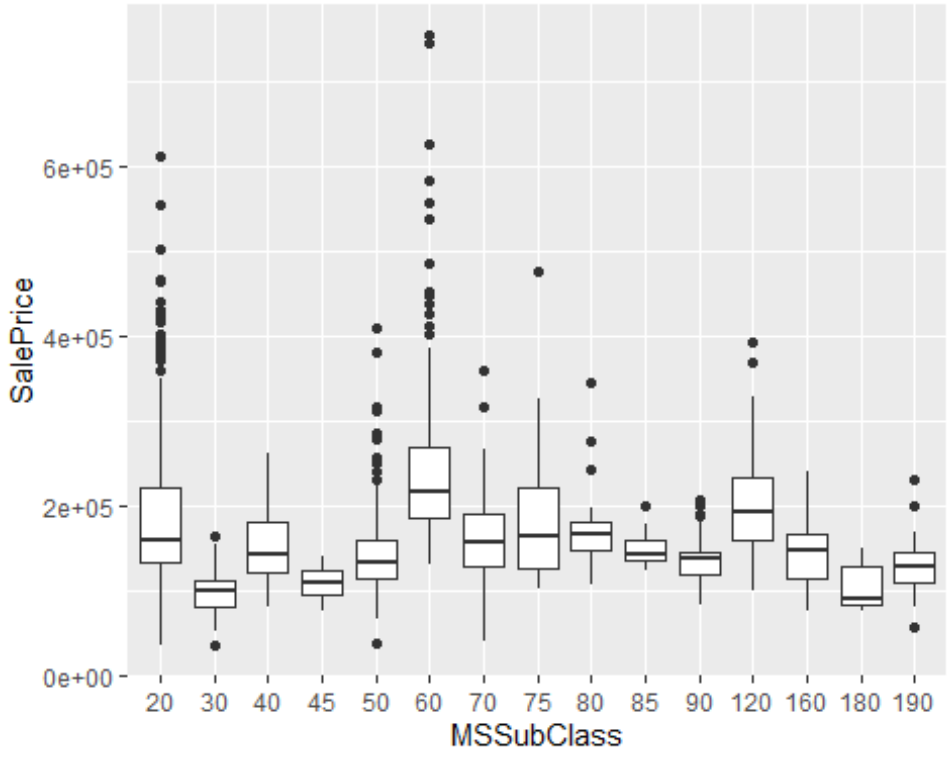

RStudio: Notebook Output

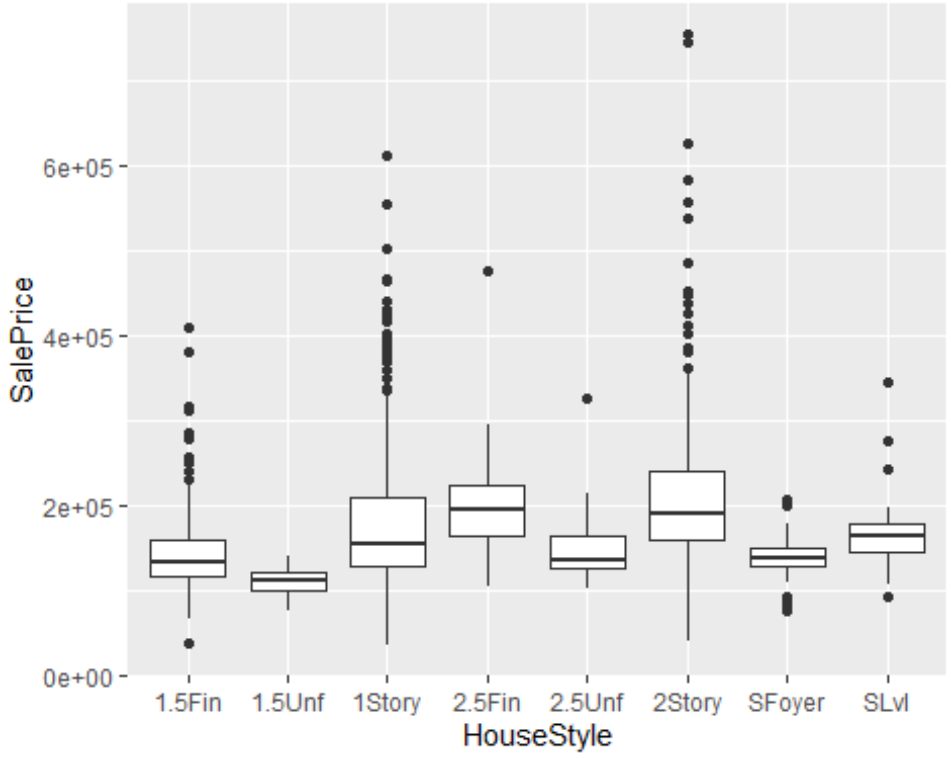
	LotFrontage <int>	LotArea <int>	OverallQual <fctr>	YearBuilt <int>	TotalBsmtSF <int>	GrLivArea <int>	TotBaths <dbl>	BedroomAbvGr <int>	TotRmsAbvGrd <int>	GarageArea <int>	SalePrice <int>
1	65	8450	7	2003	856	1710	2.5	3	8	548	208500
2	80	9600	6	1976	1262	1262	2.0	3	6	460	181500
3	68	11250	7	2001	920	1786	2.5	3	6	608	223500
4	60	9550	7	1915	756	1717	1.0	3	7	642	140000
5	84	14260	8	2000	1145	2198	2.5	4	9	836	250000
6	85	14115	5	1993	796	1362	1.5	1	5	480	143000

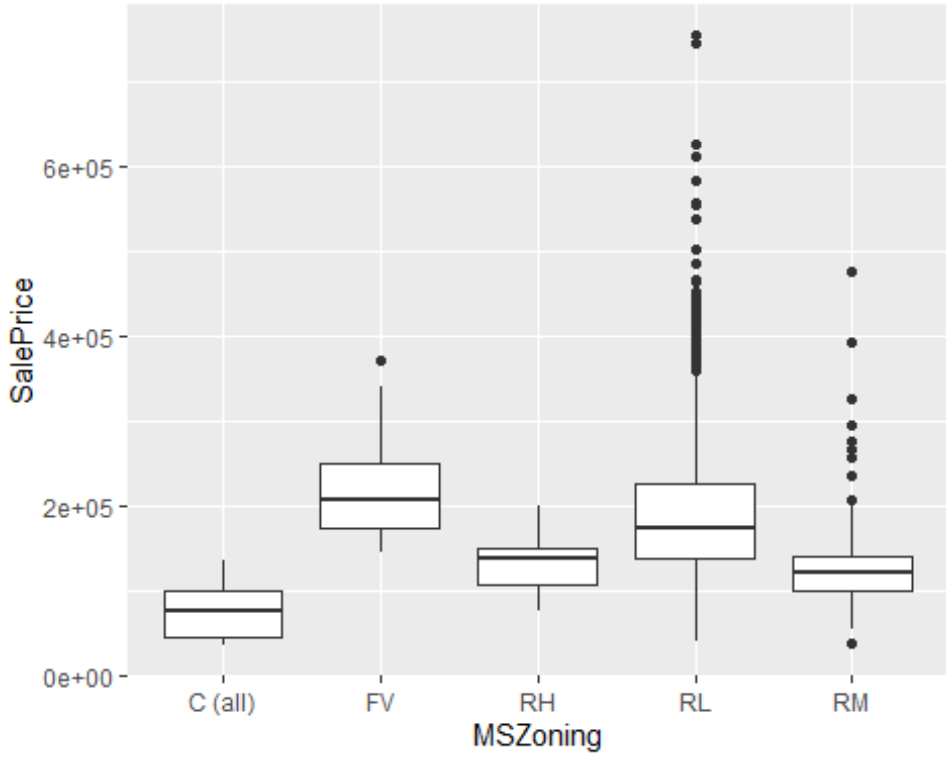
6 rows

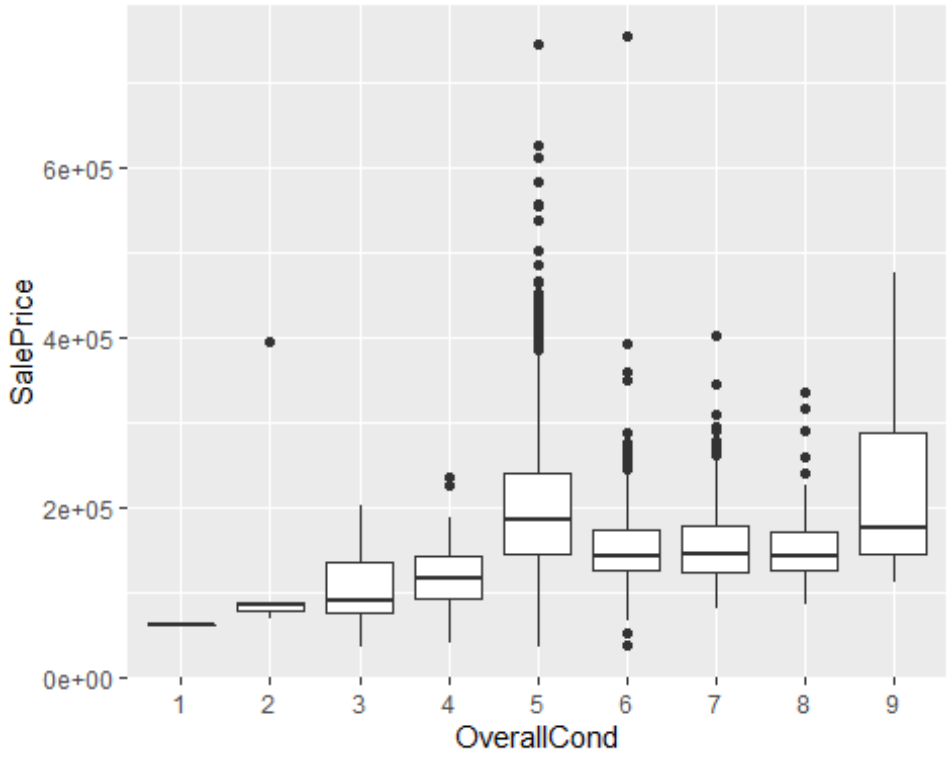
What do you not know how to do right now that you need to learn to import and cleanup your dataset?

- ⇨ Currently, I have not included some of the variables like MSSubClass, HouseStyle, MSZoning and OverallCond amongst others in my dataset as I did not see much significance – the Sale Price values were spread out amongst all corresponding sub-categories within these column / variables. Below are sample Box Plots for each category.

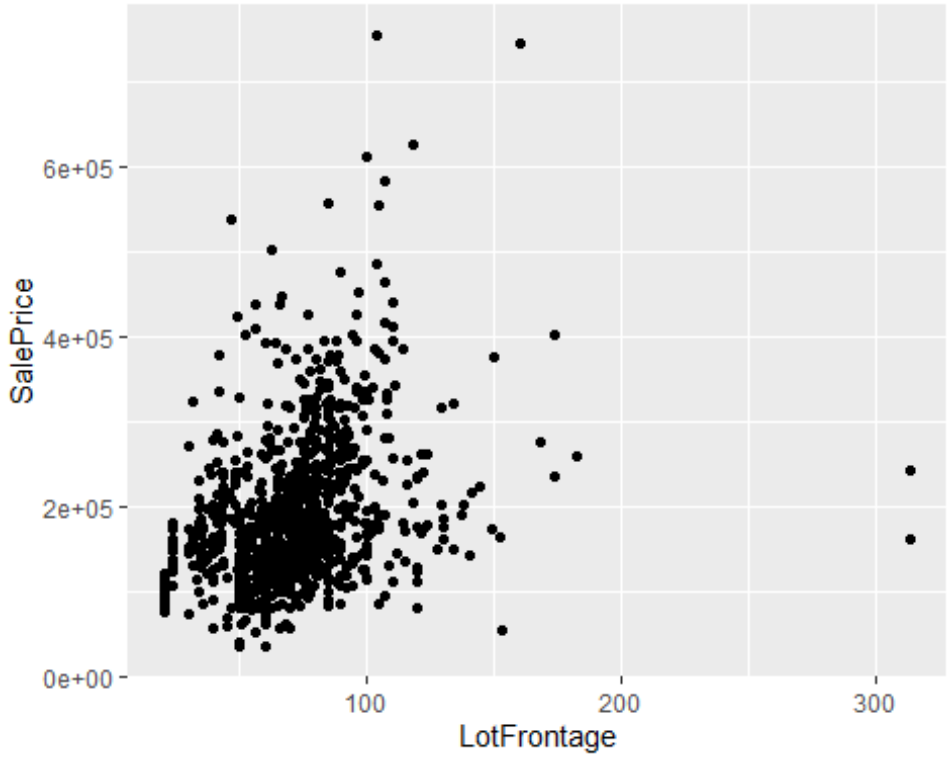

SalePrice
6e+05
4e+05
2e+05
0e+00
20
30
40
45
50
60
70
75
80
85
90
120
160
180
190
MSSubClass


SalePrice
6e+05
4e+05
2e+05
0e+00
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle


SalePrice
6e+05
4e+05
2e+05
0e+00
C (all)
FV
RH
RL
RM
MSZoning


SalePrice
6e+05
4e+05
2e+05
0e+00
1
2
3
4
5
6
7
8
9
OverallCond

I have kept Lot Frontage area in the subset variable for now. However I need to check further whether to continue using it or we can probably delete it for final analysis – it seems to show some level of correlation with house pricing, but may not be significant.



Discuss how you plan to uncover new information in the data that is not self-evident.

- ⇨ While Continuing to work through the analysis and identifying the relations between Sale Price and various factors, I will probably need to perform couple of iterations to see if any additional variables need to be considered. At the moment, I the variables which have not been considered currently, may need to be revisited later, if the need be. E.g. Masonry Veneer area, Exterior, RoofStyle, Quality, Foundation etc.. I am sure some of these will play role in real life. But I did not want to complicate the model to begin with.

What are different ways you could look at this data to answer the questions you want to answer?

- ⇨ As Discussed to some extent earlier with help of various plots and answers to previous questions, I would look at the different plots to confirm the appropriate variables to be used for housing price analysis. In addition to the plots, I would be using some of correlation / covariance techniques and the regression models to ensure the relevant factors are used for arriving at conclusion for predicting the housing prices.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

- ⇨ I planned on adding Full Bath and Half Bath data columns to create new variable as TotalBaths, based on colleague student's feedback in Week 10. I feel this will be helpful. Since I am using US Housing prices dataset – single file, I do not plan on joining different data frame, for the reasons explained at the very beginning. So, I will be using the newly created subset data frame itself.

How could you summarize your data to answer key questions?

- ⇨ I am using summary() function to summarize the data. Along with this, I will be using Correlation methods (one of the Pearson, Spearman and Kendall) deriving corresponding test statistics and p-values. After running the regression models, I will be running some tests on these models, checking standardized residuals, perform analysis of variance, and with help of plots(), hist() functions evaluating the model.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

- ⇨ I am using following types of plots as shown earlier for my analysis. I considered using GGally pair wise plots. However, it may not be needed at the moment since Sale Price is the only outcome variable.

    - ➢ Histograms for variable evaluation
    - ➢ Box Plots for patterns evaluation
    - ➢ Scatter Plots for checking relation
    - ➢ Q-Q plots for model evaluation

What do you not know how to do right now that you need to learn to answer your questions?

- ⇨ I would like to know whether current word document format is sufficient for final submission. I noticed RMarkdown is needed as well. So, will need to work towards that. I probably may not be aware of all the efficient techniques in R or other packages to make life easier for professional life work. I will need to explore more to learn on those.

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

- ⇨ I plan on using Logistic Regression / Multiple Linear regression techniques to answer the questions about housing price prediction, as these will be more relevant and helpful

using various data variables as factors in my current dataset. I am not planning on using K-means or K-Nearest-Neighbor machine learning algorithms as these may not align with my current methods for the Housing prices predictions. However, I would think on this further to see if any other needs arise. If someone would like to provide any suggestions, I would be happy to work in that direction and explore.