# Problem 3 Part A

*Problem Statement:* Performing EDA on a real world data set from RealDirect.com for Brooklyn city.

**Data:**

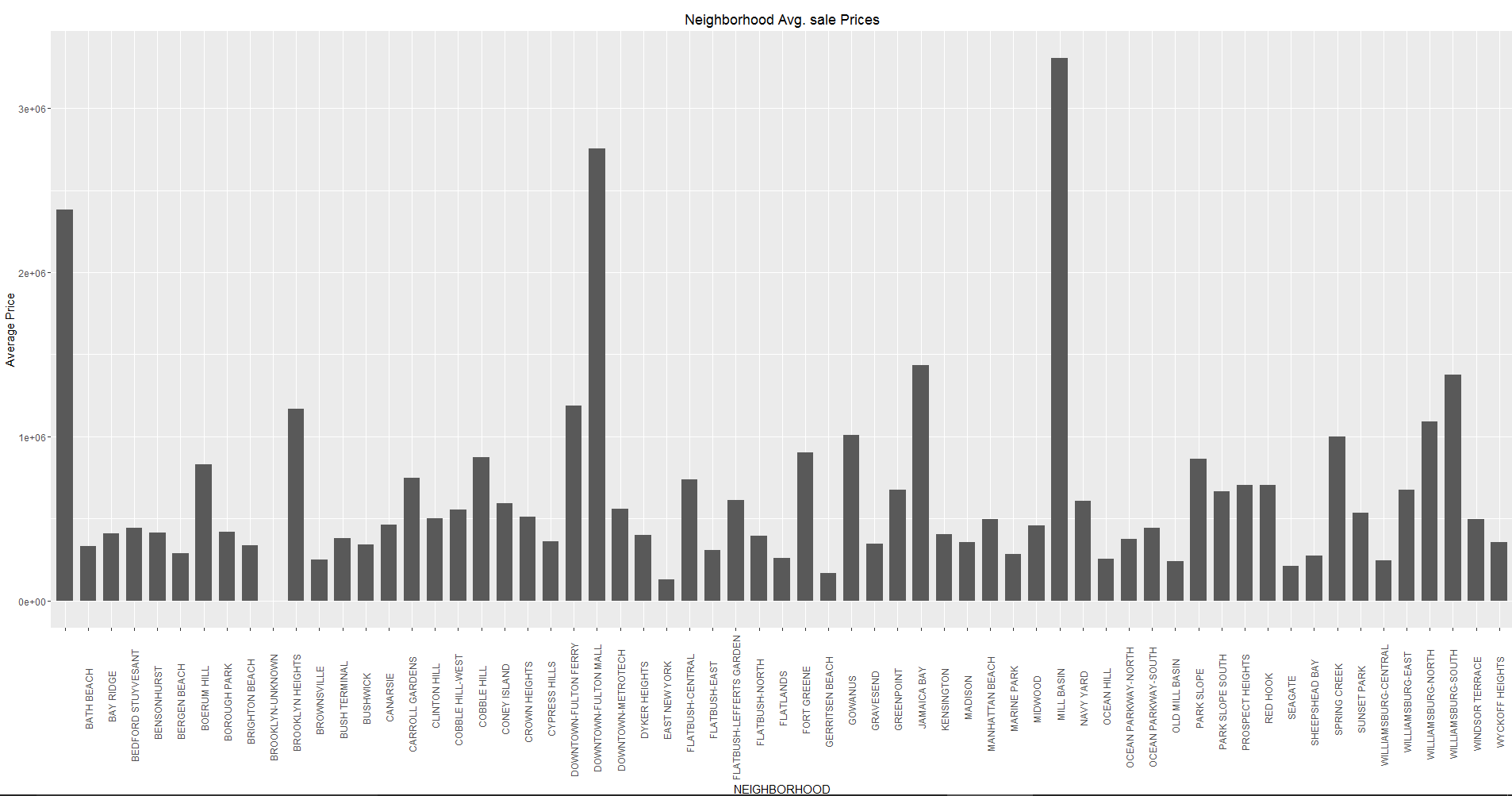
The Given dataset represent the housing sales from August 2012 to August 2013. Each row represents a single house entry/listing on realdirect.com and its features represented by values in respective columns. In our analysis we are considering the columns which help us perform some EDA on the data.

**Simple Data analysis on Brooklyn Dataset:**

Initially we format the data, by saying format I mean, changing the format of columns like Price, Date and represent numerically so that we can process it. Now we want to view the sales rate per square feet (gross) for Brooklyn City. This can be achieved by plotting the square feet in X axis and respective prices for that house in Y axis.

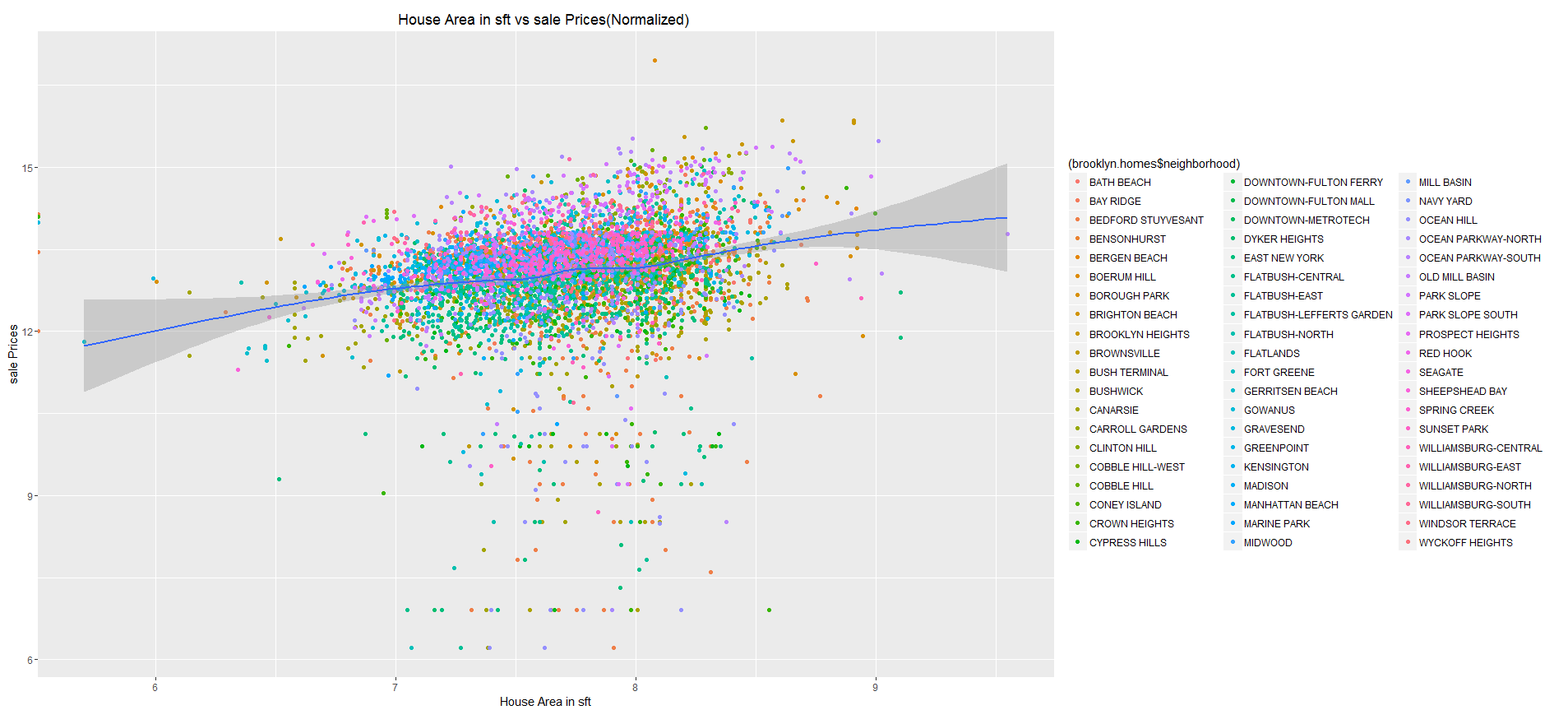
Analysis graph for Brooklyn Dataset:

**1. Average sales prices in Neighborhoods (for housing data only):**



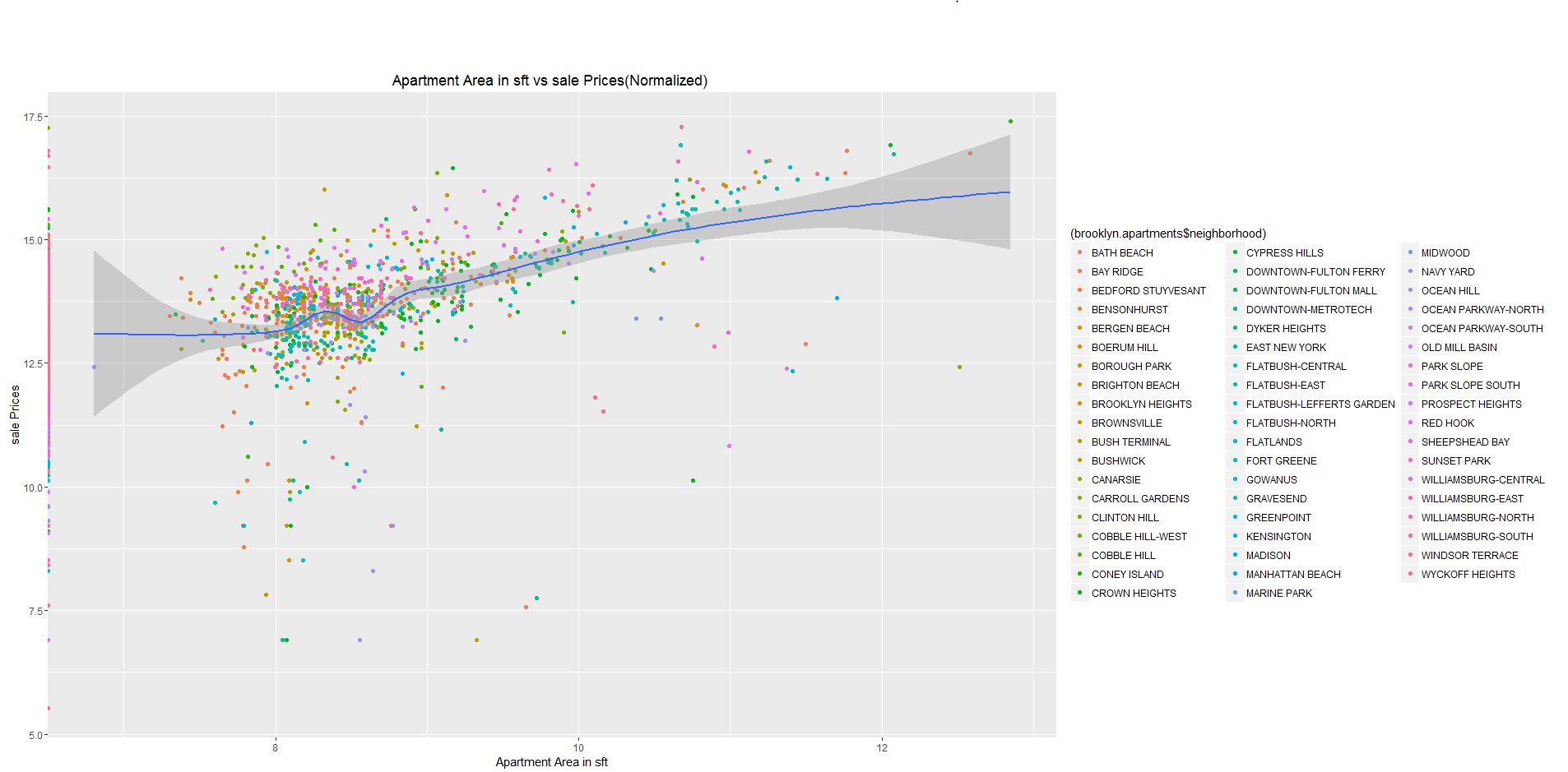
The above plot shows us what are the average prices of Houses (1BR, 2BR, 3BR only) in all the neighborhood of Brooklyn city.

2. **Predict Sale Price with House area using Existing data for all Neighborhoods (Housing Data):**



Using the Existing data we can predict the price for a new listing based on the features and past data. Now we can also predict for different types of property more accurately by considering the data relevant to that type only. Here we are trying to estimate the House prices variation with Area Using the Existing subset of Family (1BR, 2BR, 3BR) data only for all Neighborhoods. We can also summarize our analysis to particular Neighborhood alone as well

**3. Predict Sale Price with Apartment area using Existing data for all Neighborhoods (Apartment Data)**

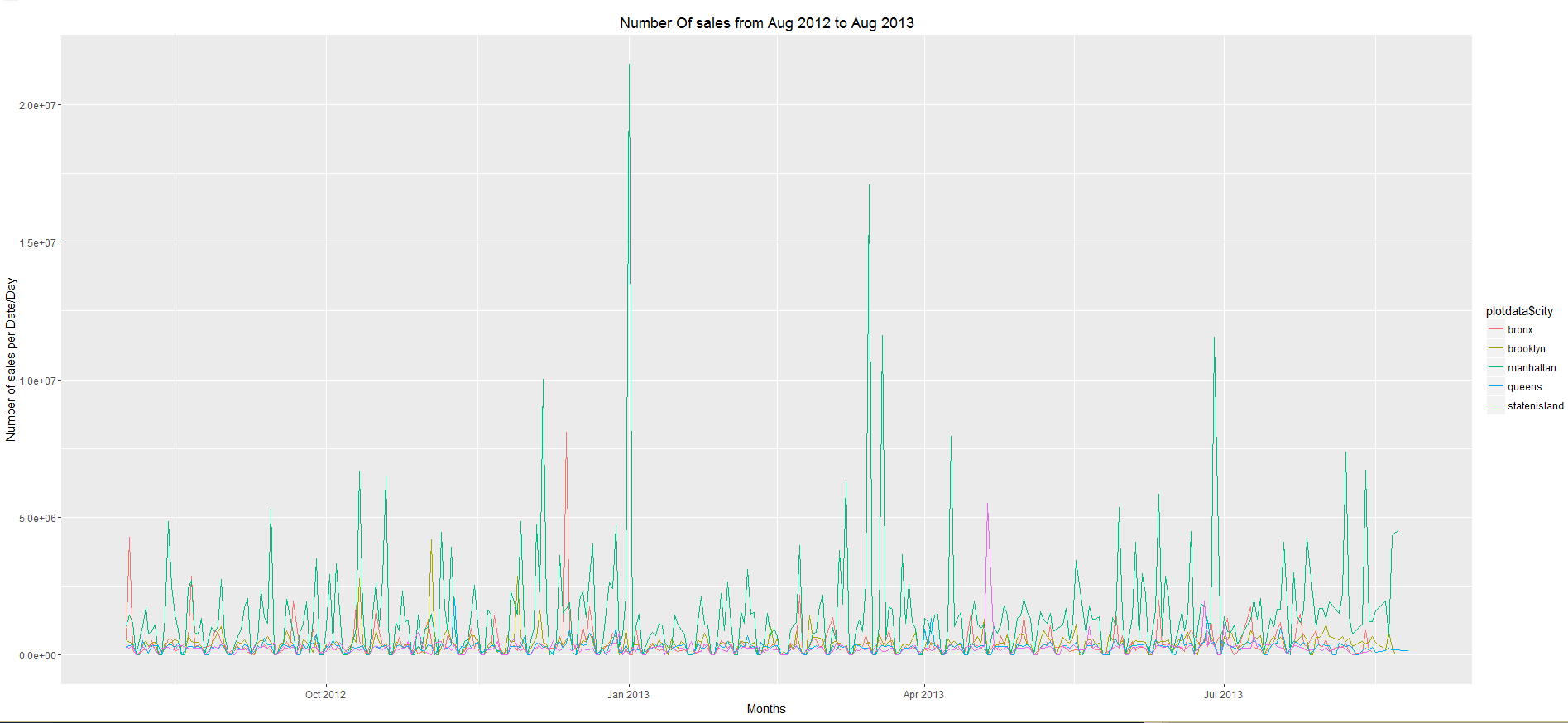


Now we can also predict for different types of property more accurately by considering the data relevant to that type only. Here we are trying to estimate the Apartment prices variation with Area Using the Existing subset of Apartment data only for all Neighborhoods. We can also summarize our analysis to particular Neighborhood alone as well

# Problem 3 Part B

For all the 5 Datasets, we can do the sales comparison. The below plot does that, by plotting number of sales per city on Y axis over the year 2012-2013. We can observe that Manhattan has highest sales overall followed by Brooklyn.

**Sales per Day compared for each city from August 2012 to August 2013**



# Conclusion

*Q1. Suggested Data Plan/Model:*

In an ideal scenario we would like to store every transaction that happens across. Transaction means a sale, irrespective of how it happens (with realdirect or without realdirect listing)

We can use this data to predict the expected selling prices for the new listing based on features and implementing some ranking algorithm.

Also we can separate the data on daily basis and categorize them based on type of property and features of property.

*Q2. Uses of Data collected:*

Collecting housing data for various cities would enable us to estimate the housing prices for a given house based on locality (neighborhood) and also compute price per square feet for each locality based on Condition of the house.

*Q3. Summarizing Data:*

I summarized data by Sales per Day compared for each city from August 2012 to August 2013

This can also be referred to as yearly sales report for all cities compared.

*Q4 People to speak to:*

Taking to a real-estate agent would help us understand how real-estate works

*Q5: Domain Knowledge:*

Domain knowledge will be helpful in accurately analyzing the data and also gathering relevant data. Currently we did not deal with the TAX fields in the datasets because we I do not have any domain experience about how to interpret that data. A domain expert would have utilized those fields to make some interesting observations. Doing some research about how real-estate works, gave me a better understanding and helped me to accurately summarize data.

Reference: http://home.howstuffworks.com/real-estate/buying-home/realtor4.htm

*Q6: Recommended Best practices:*

From an online website view, we need to update sales prices based on inflation difference over years for accurate listing prediction/suggestion to customers and dealers.

References:

My submission with for this project includes references from below links:

1. http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

2. http://stackoverflow.com/questions/11370323/learning-to-understand-plyr-ddply

3. http://ww2.coastal.edu/kingw/statistics/R-tutorials/dataframes.html

4. R-Cookbook

5. http://stackoverflow.com/questions/19655431/reading-multiple-csv-files-in-r