**DSC 520 Final Project**

Name: Conrad Ibanez

Week 12 Updates: 3/2/2019

**Section 1**

* Explain what your interests are in the data sets identified.
* Identify the Packages that are needed for your project.
* Original source where the data was obtained is cited and, if possible, hyperlinked.
* Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

For my Final project, I am considering the following data sets.

1) https://www.kaggle.com/zillow/zecon#State\_time\_series.csv

I am interested in Zillow data as I may be selling a home in the near future. I have heard that the best time to sell is in the Spring/Summer time before schools starts in the Fall, so I want to see if the data upholds that and can indicate which period is the best time to sell. I will likely narrow the data by state or if necessary by county/zip.

2)https://s3.amazonaws.com/amazon-reviews-pds/readme.html - https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon\_reviews\_us\_Baby\_v1\_00.tsv.gz

I am interested in many things related to baby products as I have one child and another on the way. I would like to analyze baby product reviews in more detail from Amazon since I have purchased many products on the site and have relied on many customer reviews.

3) https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories

I am interested in investing in various stocks. I recently opened a Robinhood account and started investing in stocks of my own choosing. Since this data set contains historical data for stock prices for many companies, I would like to focus on companies within the same industry, perhaps automobile or telecommunications, where competition is fierce and there could be a relationship between stock prices among those companies.

Since the data in these sets are tab delimited or in csv format, I will use the readr package to import the data once I have downloaded the files. I will also use the ggplot2 package to create various graphs of the data sets.

With regard to the following:

a. Importing the data- There should be no issues as the data files are tab delimited or csv. I would need to explore the possibility of importing the data directly from a url source file instead of having to download locally.

b. Identifying and reviewing the schema (codebook)- I do not think any of the data sets have a formal codebook. The Zillow data set has a data dictionary and the Amazon data set has a readme webpage with a lot of information. The data set for stocks has some notes.

c. Identifying missing data- While briefly scanning the data sets, there does not seem to be issues with missing data. Some of the data sets are quite large so further analysis must be done to properly handle missing data if required.

d. Learn about the data visually (plotting) and numerically (descriptive stats)- I will need to do further analysis of the data sets to determine the best way to visualize the data through various graphs and plotting. Additionally, I will need to determine which descriptive stats will be useful for the analysis.

**Section 2**

* Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
* Provide a concise explanation of how you plan to address this problem statement.
* Discuss how your proposed approach will address (fully or partially) this problem.
* List at least 6 research questions you aim to answer.
* Explain how your analysis may help the consumer of your research findings.
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?
* What do you not know how to do right now that you need to learn to answer your research questions?

Problem Statement: Selling a house is not likely a familiar process or a common occurrence for many people. However when it does happen, when is the best time to sell? What factors may have an influence for sellers to get top dollar for the homes they are selling? Home ownership is probably one of the biggest investments for many people, so it is important to maximize the price if an owner does decide to sell.

Data sets from Redfin will be analyzed to determine relationships with regard to the price of homes. Variables such as median sale price, number of homes sold, inventory, number of new listings and others will be investigated. Calculations, plots and correlation data for these variables will be completed.

Approach: The following research questions will be examined:

1. What factors influence the best time to sell?
2. Which time of the year yields the highest prices for homes?
3. Is there a relationship between inventory level and the prices of homes?
4. Is there a specific time such as a particular month where inventory levels are the highest?
5. To get top-dollar, does that mean a homeowner may also need to wait longer to sell?

Consumers of this research will get a better understanding of the variables that can affect home prices. They can use the findings to consider when would be the best time to sell their homes to get the best prices in the market.

Scatter plots will be useful to help discover relationships among variables. Averages, covariance, correlation coefficients and other metrics will be calculated to help illustrate the findings of the research questions.

For this project, Redfin site (<https://www.redfin.com/blog/data-center>) will be the source of the data sets as there seems to be more workable data and lacks missing data that the Zillow data set had. Additionally, the site allows you to do initial filtering of the data set that you want to use which brings up some questions regarding the scope of the data and how much data should be considered in the analysis. Should data sets in the same city/multiple cities, same state/multiple states, and/or the overall national numbers, etc. be considered?

Also, there are likely other factors influencing pricing of homes that are beyond what the current Redfin data set holds. I am curious if maybe those should also be investigated and included as part of the project. For example, the historical data for interest rates or the employment numbers for the areas in consideration could have an impact.

Regarding the question- ‘What do you not know how to do right now that you need to learn to answer your research questions?’ As I mentioned previously, I would like to get a better understanding on how to determine whether a data set is good. In addition to the question I had regarding which locations to consider, it appears Redfin data only goes back to 2012. Will that be good enough for analysis? When data sets get bigger, would it be better to set up a local database and use R to query it? If not, is there a way to persist data in R after it has already been read in. When I tried to analyze the data set from Zillow that had very large files and seemed to have millions of rows, R Studio appeared to be hanging. I am curious if someone else is facing issues working with very large data sets.

A peer in class also brought up potentially using maps layout to present data. I am not sure if that is more of an advanced topic than what this class should be, but considering this project is about real estate and location is an important factor in home buying and selling, it may be a good idea to explore the possibility of using maps for the data. If time permits, I will look into incorporating ggmap library for displaying data against the physical locations.

Other topics that I may need to look into more are advanced data manipulations and calculations. After completing some of the assignments, it seemed R had a robust set of library and functions that are equivalent to what can be accomplished using database queries, stored procedures, and temp tables. Those abilities may be useful for this project.

**Section 3**

* Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.
* 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.
* What do you not know how to do right now that you need to learn to import and cleanup your dataset?

I did more exploration on the home sales data from Redfin after receiving feedback that I may need to run processing of the data set on a server since it contained millions of rows. Fortunately, the Redfin data set has features and capabilities using Tableau Public. I downloaded the Tableau Public client app and began browsing the data. For now, I have decided to limit the data set to the state of Georgia where I live and expand the data set as needed for the analysis. Tableau allowed me to filter and export the data for Georgia to a CSV file. Instead of having millions of rows which R Studio could not process on my local machine, I now have about 50K records for analysis which is more manageable.

After reading the CSV file, the str() function gave me the following metrics on the initial data set: 48960 obs. of 55 variables. I ran mean() on the attributes that would help me answer some of my research questions:

> mean(redfin$Homes.Sold) [1] 112.9739

> mean(redfin$Inventory) [1] NA

> mean(redfin$New.Listings) [1] NA

The ‘NA’ indicated that there are missing values for inventory and new listings. Cleaning data is necessary to avoid these types of errors in calculations. I created a new data set omitting the records with missing data for those attributes using the following code:

redfinClean <- na.omit(redfin[, c("Homes.Sold", "Inventory", "New.Listings")])

My clean data set has the following metrics: str(redfinClean) 'data.frame': 48900 obs. of 55 variables. This means 60 rows were removed. This should be fine as it accounts for only a small percentage of the overall data set and without the data for these variables, the record would not be useful in this project’s analysis. I ran the mean for the new data set and got values for all:

> mean(redfinClean$Homes.Sold) [1] 113.1049

> mean(redfinClean$Inventory) [1] 329.1

> mean(redfinClean$New.Listings) [1] 123.8579

I will follow this same logical process for cleaning other potential records if more attributes are required for my analysis. There may be instances where keeping records with null values is better for analysis. In those cases, setting attributes to functions to handle missing data will be required.

As far as things I would like to know more, I will continue to use Tableau to browse the overall data set from Redfin to see if there are additional data for analysis in this project and possibility of expanding to other states/locations. There have also been discussions in the class thread as far as omitting data and whether there is some standard process for making decisions. My peers have mentioned a Data Camp course for cleaning up data which I plan to complete that may provide some additional direction.

**Section 4**

* Discuss how you plan to uncover new information in the data that is not self-evident.
* What are different ways you could look at this data to answer the questions you want to answer?
* 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, learn and use an R package not covered so far in the course?
* How could you summarize your data to answer key questions?
* What types of plots and tables will help you to illustrate the findings to your questions?
* What do you not know how to do right now that you need to learn to answer your questions?
* Do you plan on incorporating any machine learning techniques?

I plan to uncover new information in the data by generating various plots and analyzing the graphs. Please see the accompanying R Markdown generated Word document for additional details. By generating different scatterplots, we can determine relationships between variables. We see from the scatterplot of Median Sale Price vs Median List Price that there is a positive relationship between the two variables, and a linear regression model and analysis can be done to further analyze the relationship.

I also generated time series charts since most of my project questions involved questions on when would be the best time to sell a house. I had to convert the text format of the date to a date format by using dplyr and lubridate libraries. The time series charts are based on the 30097 zip code, which was data obtained by using filter() function on the overall data set. I wanted to simplify things by using one particular zip code before doing further analysis, but the data set may be too small. I may need to expand to other zip codes near and around 30097. I will look into using ggmap library to display information within a context of a map.

In order to answer key questions, I will need to break down the dates into more detail including months. Therefore, I will look into categorizing the data into periods of months. By doing so, I will be able to calculate the differences in prices and various inventory levels per month and analyze the impacts of the time period to the variables.

Again, plots such as scatterplots, time series, and various regression models should help to illustrate the findings to the questions of this project. Pricing tables with respect to the different months should also be useful. I will need to understand more on how to transform some of the data and manipulate text values into formats that can be used for processing. There may be other libraries such as the one I used to convert the date text that may be helpful for this project.

As for machine learning, I am looking forward to learning the techniques. I am not familiar with machine learning at this point, so I will have to wait on determining if they can be applied to this project. Perhaps based on machine learning, the best time to sell a home can be determined or pricing as it relates to the list versus what is the actual sale price received can be projected better.

**Section 5 Summary**

Problem Statement: Selling a house is not likely a familiar process or a common occurrence for many people. However when it does happen, when is the best time to sell? What factors may have an influence for sellers to get top dollar for the homes they are selling? Home ownership is probably one of the biggest investments for many people, so it is important to maximize the price if an owner does decide to sell.

We attempt to answer the following questions:

1. What factors influence the best time to sell?

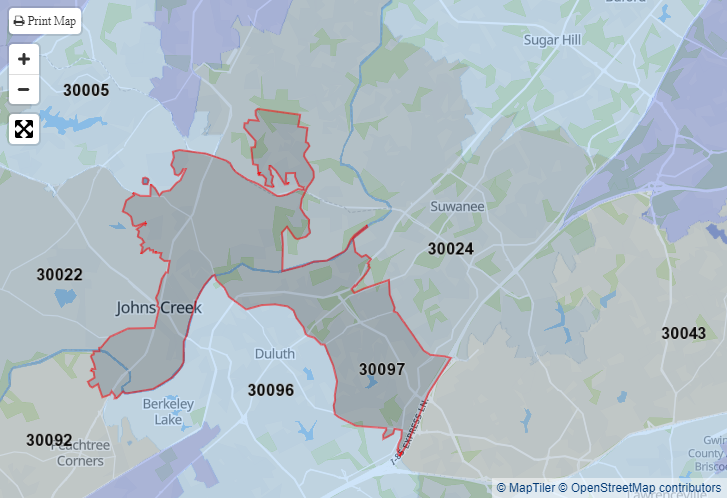
2. Which time of the year yields the highest prices for homes?

3. Is there a relationship between inventory level and the prices of homes?

4. Is there a specific time such as a particular month where inventory levels are the highest?

5. To get top-dollar, does that mean a homeowner may also need to wait longer to sell?

In this project, housing sales data from Redfin from 2011 to 2018 is analyzed. A subset of the data is selected with a focus on the city of Duluth, GA, with a zip code of 30097. We define the Area 30097 to also include the bordering zip codes of 30096, 30092, 30024, 30022, 30005, and 30043.



Based on linear regression line on scatterplots of Median Sale Price and other variables, there is a negative relationship between Median Sale Price with most other variables such as Inventory, Homes Sold, New Listings, and Pending Sales. That is when these variables increase, Median Sale Price decreases. However, there is a strong positive relationship between Median Sale Price and Median List Price. This makes sense in that as Median List Price increases, then Median Sale Price also increases.

We observe the following from Time Series Graphs regarding the different variables (Please see R Markdown report for graphs and more details):

* The Median Sale Price is highest in the month of June, followed by May, July and then April. The Median Sale Price is lowest in the month of November, followed by December and January.
* Median List Price is highest in March, followed by April and February. It is lowest in November, followed by October and December.
* Inventory is at the highest in April, followed by May and March. Inventory is at the lowest in October.
* Homes Sold is highest in the month of June, followed by May, July and then April. Homes sold is lowest in the month of January, followed by December and November.
* Pending Sales is at the highest in April, followed by March and May. Pending Sales is at the lowest in November.
* New Listings is at the highest in April, followed by March and May. New Listings is at the lowest in November.

We gain the following interesting insights and implications to the consumer (target audience) of your analysis.

Based on a Redfin article, it usually takes 50-60 days for a mortgage loan to close. This could explain why certain variables such as Median List Price, Inventory, Pending Sales, and New Listings seem to be generally at the highest in the months of March, April, and May, while Median Sale Price and Homes Sold are the highest in June followed by May and July.

The Redfin data supports the initial assumption that the best time to sell a home is in the Spring or Summer (March - August) in order to maximize the Sale Price. It would be best to list a home in the Spring or even early Spring (February – April) and hope to complete the sale by Summer (May – July).

Inventory is highest in April which could indicate home sellers taking advantage of potentially more buyers and possibly buyers willing to pay a premium for homes. There seems to be a strong relationship between Inventory levels and the Sale Price with respect to the time of year.

Unfortunately, the Redfin data does not include metrics on the amount of time closing took or how many offers were made and fell through to answer the question of whether the homeowner may also need to wait longer to sell to get top dollar. Observations seem to indicate that if a seller waits too long and enters the months where activity is slow and variables are lowest, that the seller may need to wait until the more promising months cycle again.

In comparing the months of different years, we see that the housing market has recovered very well since the housing crisis. It is great to sell in 2018 compared to 2011 as Median Sale Price and Median List Price have increased significantly over the years since then. There were high inventory levels in 2011 and 2012 possibly due to the lower demand for homes or higher foreclosures.

This analysis focused on the 30097-zip code for Duluth, GA, and neighboring zip codes. It would probably be good to compare each zip code for better insights or compare the area against the metrics for the entire state or maybe comparing different states. I had planned to use a mapping library to associate the metrics with location but did not have the time, so that could be an improvement. Additionally, other variables such as zip code median income, employment/unemployment, and other data could be other factors influencing home prices that could further be investigated.