Project Part 1

Reference 1 for source of the data [1]

Data Explanation

The data used in this project were collcted and were used to examine the relationship between the sale price of a real estate and three independent variables: appraised land value of the real estate, appraised value of improvements on the real estate, and the neighborhood in which the real estate is listed. The data consist of the appraised land values and improvement values and sale prices for real estates sold in city of Tampa, Florida, during the period between May 2008 and June 2009. The data pertains to eight neighborhoods (Hyde Park, Cheval, Hunter's Green, Davis Isles, Avila, Carrollwood, Tampa Palms, and Town & Country), which are relatively homogeneous but differ sociologically and in real estate types and values. The sales and appraisal data were recorded along with the neighborhood that the real estate was sold at. In total, the data has 350 observations. In the data, the numbers represent thousands of dollars. For the project, the relevant variables are pertained to 5 variables: SALES, LAND, IMP, TOTVAL, and NBHD. SALES variable is the sales price of the property in thousands of dollars. LAND variabe is the value of land in thousands of dollars. IMP variable is the value of improvements in thousands of dollars. TOTVAL variable is the sum of LAND and IMP variables, representing the total value of the real estate in terms of appraised land and improvement values. NBHD variable is the neighborhood for which the property was sold at.

Data Characteristics

The data is a sample data as the population of interest is the real estate market. The data for the project was collected by the property appraiser's office of Hillsborough County, Florida. As stated in Data Explanation section, the data collected sales prices and appraised land and improvement values of real estates sold in 8 neighborhoods from May 2008 and June 2009. The sample selection method was non-probability sampling. The samples chosen were not chosen at random. To be specific, this is a purposive sampling as the neighborhoods chosen were all relatively homogeneous.

[1] The data set that I will be using will be a private data set that I have permission to use. If needed, please contact me.

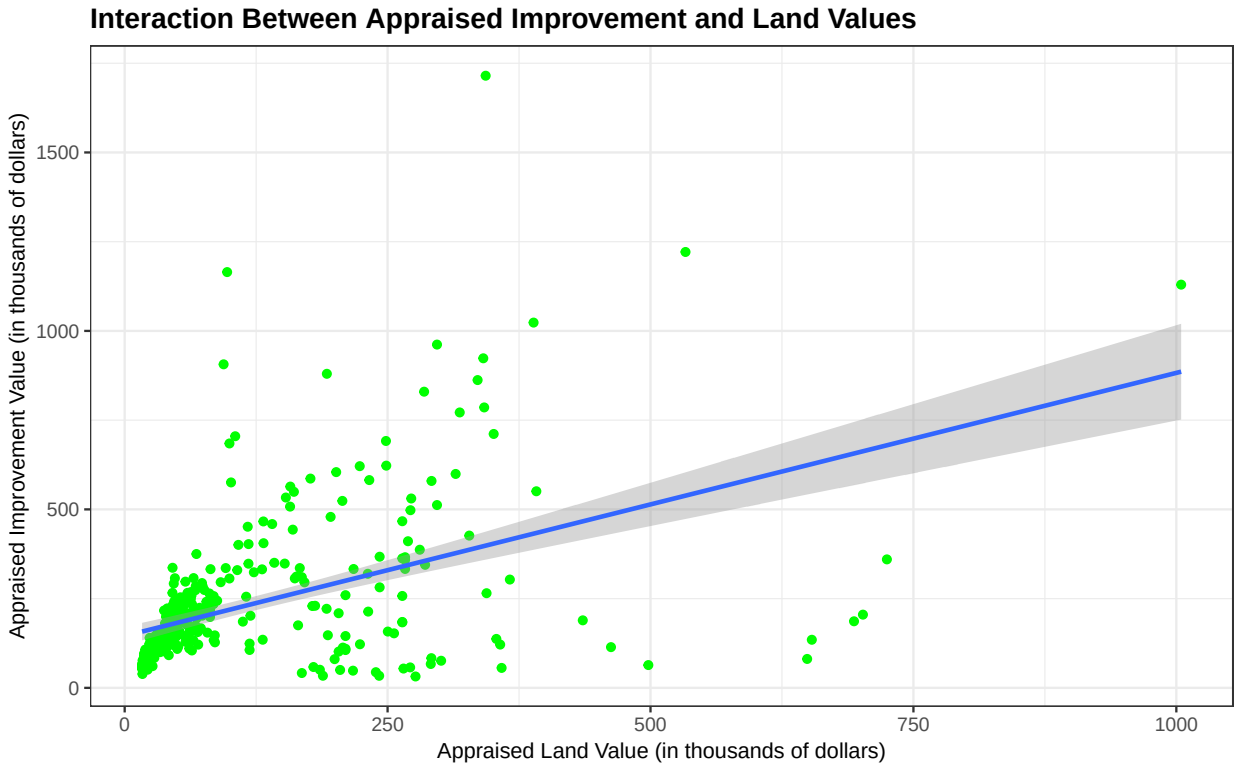
Data Issues

The most important issue is that the samples were not chosen at random. Although non-probability sampling does provide advantages of convenience and cost, it does not allow to estimate how different the sample statics will be from the population parameters. Another issue is that the sample size for one of the neighborhoods is small. For the Avila neighborhood, the sample size is only 12, which would definitely cause issues with measurements of the center. For that neighborhood, the data will most-likely to be skewed and thus the mean will be heavily affected by the outlier along with the variance.

Data Summary

The first set of variables that will be analyzed is LAND and IMP variables.

```
ggplot(sale2,aes(x=LAND,y=IMP)) +geom_point(colour="Green") +geom_smooth(method=lm)+
  ggtitle("Interaction Between Appraised Improvement and Land Values")+
  labs(x="Appraised Land Value (in thousands of dollars)",
       y="Appraised Improvement Value (in thousands of dollars)")+
  theme_bw()+theme(plot.title=element_text(face="bold"),axis.title.y
                   =element_text(size=10),axis.title.x=element_text(size=10))
```



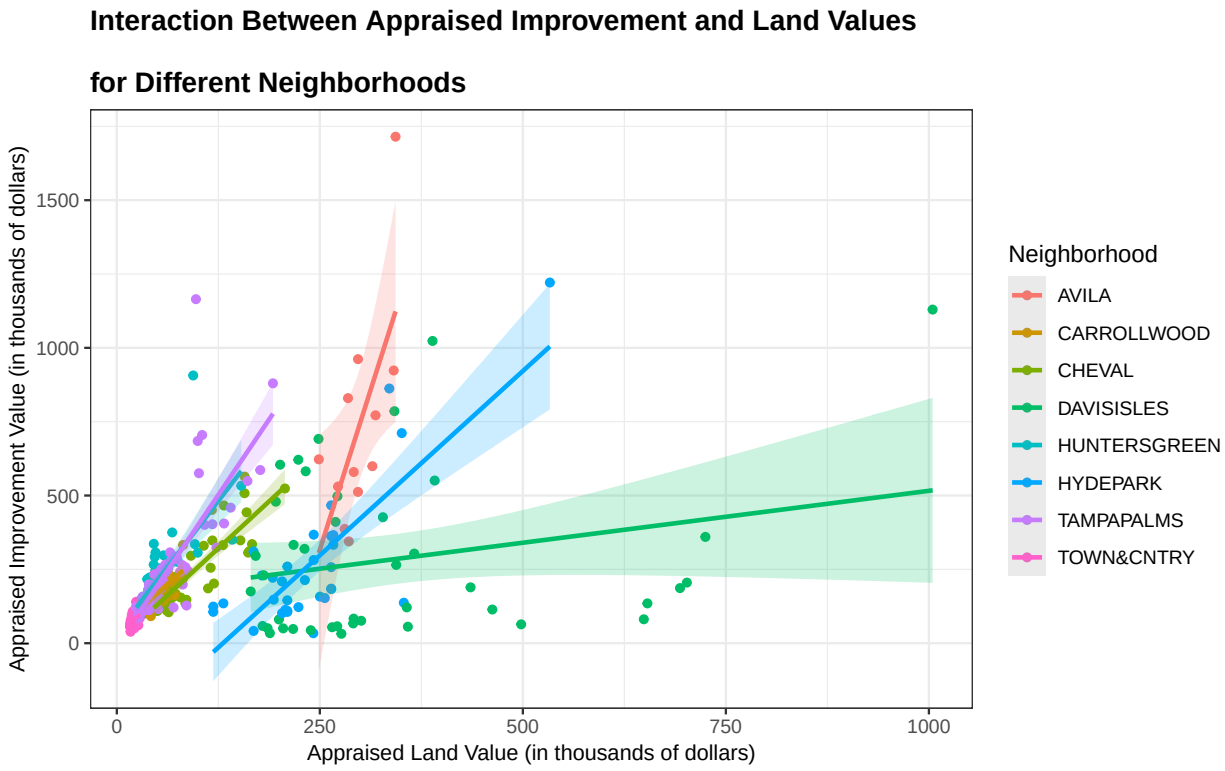
Based on the graph, the general trend is that the appraised improvement value of a real estate increases as the appraised land value increases. However, based on the correlation test, these two variables do not seem to have strong correlation with each other and have rather weak positive correlation.

```
cor(sale2$LAND,sale2$IMP)
```

```
## [1] 0.4615883
```

For different neighborhoods, the regression lines look quite different as seen below.

```
ggplot(sale2,aes(x=LAND,y=IMP,colour = NBHD)) +geom_point() +
  stat_smooth(method="lm", aes(fill=NBHD),alpha=0.2)+
  ggtitle("Interaction Between Appraised Improvement and Land Values
          \nfor Different Neighborhoods")+
  labs(x="Appraised Land Value (in thousands of dollars)",
       y="Appraised Improvement Value (in thousands of dollars)",color="Neighborhood")+
  theme_bw() +theme(plot.title=element_text(face="bold"),axis.title.y=element_text
                    (size=10),axis.title.x=element_text(size=10))+ guides(fill=FALSE)
```



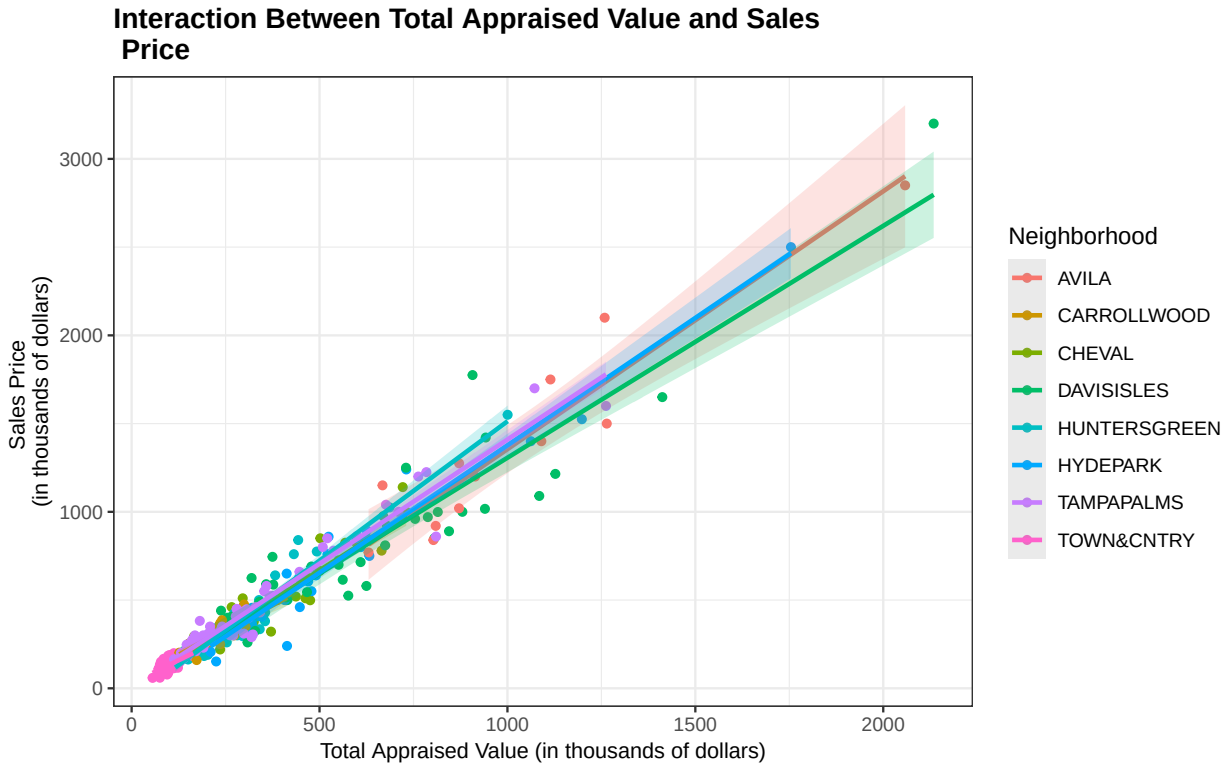
Reference 2 for stat_smooth() function #### Reference 3 for guides(fill=F) function It can be seen that the linear regression lines for each neighborhood differ quite significantly from one another. Especially for Davis Isles neighborhood, the regression line has a very wide confidence band, so perhaps the one green data sample in the far right is an outlier. Based on the data analysis done above, it can be concluded that appraised imrpovement and land values interact with each other. The next two variables that will be tested are TOTVAL and SALES.

```
cor(sale2$TOTVAL,sale2$SALES)
```

```
## [1] 0.9726988
```

Based on the correlation test, it appears that they have strong positive correlation.

```
ggplot(sale2,aes(x=TOTVAL,y=SALES,colour=NBHD))+geom_point()+stat_smooth(method="lm",
 aes(fill=NBHD),alpha=0.2)+ggtitle("Interaction Between Total Appraised Value and Sales
 Price")+labs(x="Total Appraised Value (in thousands of dollars)",y="Sales Price
(in thousands of dollars)",color="Neighborhood")+theme_bw()+
theme(plot.title=element_text(face="bold"),
axis.title.y=element_text(size=10),axis.title.x=element_text(size=10))+guides(fill=FALSE
```



Based on the graph above, it can be seen that total appraised value does affect the sales price of property and does have a very strong positive correlation. Thus, it can be concluded that total appraised value affects the sales price. The next variable that will be analyzed is NBHD. The analysis below is a statistical summary of sales price of real estates for different neighborhoods. The interaction between total appraised value and neighborhood will not be analyzed since affect of different neighborhoods on sales price is much more important to analyze and analysis of neighborhood and total appraised value will yield similar results as the sales price analysis.

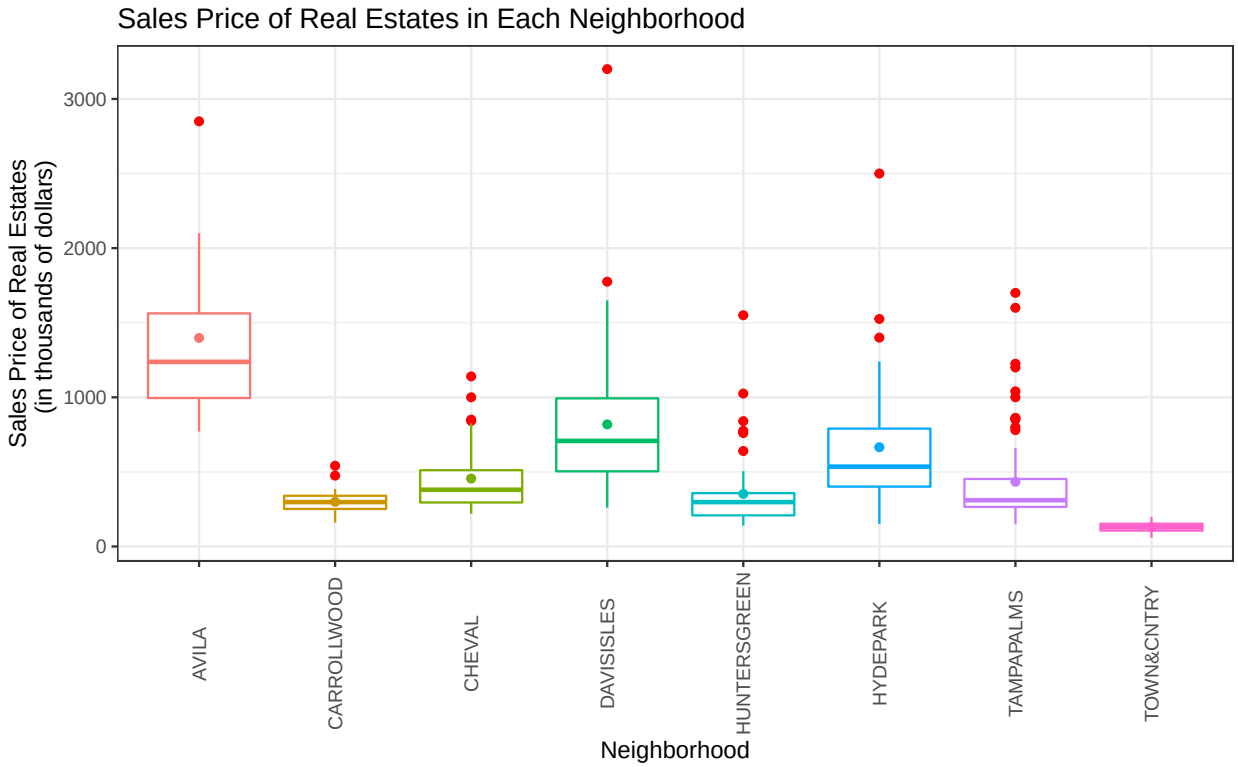
```
tapply(sale2$SALES,sale2$NBHD,summary)
```

```
## $AVILA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   770.0   995.1  1237.5  1398.0  1562.5  2850.0
##
## $CARROLLWOOD
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   160.0   251.5   297.5   299.4   340.0   541.0
##
## $CHEVAL
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   219.9   295.0   380.2   455.4   511.2  1140.0
##
## $DAVISISLES
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   259.9   503.8   707.5   818.2   993.0  3200.0
##
## $HUNTERSGREEN
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   140.0   208.5   297.5   352.8   357.5  1550.0
##
## $HYDEPARK
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   152.0   401.2   535.0   665.3   790.0  2500.0
##
## $TAMPAPALMS
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   149.0   265.2   310.0   433.3   452.5  1700.0
##
## $`TOWN&CNTRY`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    59.1   106.7   130.0   130.3   152.5   199.0
```

Based on the statistical summary, it appears that for different neighborhoods, the median sale prices differ as well. Perhaps the sales price of real estate is indeed affected by the neighborhood that the real estate is sold at. Below is a boxplot of the statistical summary for different neighborhoods.

```
ggplot(sale2,aes(x=NBHD,y=SALES,colour=NBHD))+geom_boxplot(outlier.color="red")+
  stat_summary(fun.y="mean",geom="point")+
ggtitle("Sales Price of Real Estates in Each Neighborhood")+
labs(x="Neighborhood",y="Sales Price of Real Estates
     (in thousands of dollars)", color="Neighborhood")+theme_bw()+
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5))+guides(color=F)
```



The boxplot above shows that for different neighborhoods have different measures of center. The red dots indicate outliers and the dot in the box represents the mean. Based on the statistical summary, it can be concluded that neighborhoods do affect the sales price of real estates.

Conclusions

Based on the analysis, it can be concluded that the appraised values of land and improvements are positively correlated to the sale price and can be used to predict the sales price of real estates. The neighborhood that the property was sold at also seems to have an affect on sales price as the median for sales price for each neighborhood differ significantly from each other; thus, it will be necessary to incorporate neighborhood as a factor when predicting sales price of a real estate.

References

1. https://www.hcpafl.org/
2. https://stackoverflow.com/questions/32800716/ggplot-stat-smooth-change-look-of-multiple-bands
3. https://stackoverflow.com/questions/35618260/remove-legend-ggplot-2-2