Running head: ASSIGNMENT 3

Assignment 3

Yuqi Zheng

AD 571 Business Analytics Foundations

Boston University

2019 Fall

10/11/2019

- Task 1: Provide several descriptive statistics for real estate sales in SUNNYSIDE.
- Task 2: Perform k-means clustering, comparing SUNNYSIDE to other neighborhoods. Choose at least three KPI.
- Task 3: Using t-test to test the hypothesis that, starting in 2009, the average residential property of SUNNYSIDE costs more, less, or a different amount in WHITESTONE.

Table of Contents

Executive Summary

This report uses R to provide several descriptive statistics for real estate sales in SUNNYSIDE. Besides, uses k-means clustering and t-test to compare SUNNYSIDE to other neighborhoods.

After analyzing, it indicates that SUNNYSIDE has a very large sale price, which is $1983213120, and it also sales 1604 units since 2009. To analyze a neighborhood is stable or not, we need to see SUNNYSIDE's mean sale price and standard deviation. These two data are $1129941 and 2760438 respectively.

Not only by finding the standard deviation of sale price, but also using k-means clustering. They are all indicates SUNNYSIDE is a quite stable neighborhood comparing to other neighborhoods since it has a low standard deviation and median sale price. The sales situation, unit price and standard deviation of SUNNYSIDE could be clearly seen in the Appendix 4.

In addition, comparing to WHITESTONE, SUNNYSIDE is more profitable since 2009 because its mean sale price is 1129940.8, which is larger than WHITESTONE's 732785.9. Thus, the recommendation would be SUNNYSIDE based on t-test analysis.

Descriptive statistics for real estate in SUNNYSIDE

The total number of sales means the total unit sales in SUNNYSIDE since 2009. It should use the filter function to filter neighborhood name to SUNNYSIDE; year should be large and equal to 2009 and filter out data with 0 for sale price and square feet. Using the summarise function to show the sum of the sale price is $1983213120, and the unit of sales is 1604 by using n = n( ).

Excepting filter data as the same steps in the above, it should add one more step to filter the residential Status. Then it is easily getting the mean sale price and gross square footage for residential properties in SUNNYSIDE since 2009 by using functions mean (GrossSqFt) and mean (SalePrice). Thus, the mean sale price is $1129941, and the mean gross square feet is $4124.841.

The five-number summary for both sale price and gross square feet for residential properties in SUNNYSIDE since 2009 could be found by using the fivenum function. Besides using the fivenum function, quantile function also could present the five-number summary for both sale price and gross square feet. The difference between the fivenum function and quantile function is quantile function provides the percentages, but the fivnum function does not. Thus, the gross square feet' sample minimum is 484; the lower quartile is 1390, the median is 1864, the upper quartile is 2617, and the sample maximum is 93000. The sale price's sample minimum is $1, the lower quartile is $425000, the median is $590000, the upper quartile is $850000, and the sample maximum is $28796875.

There are four types of Status in SUNNYSIDE; they are commercial, mixed, residential, and other. The proportion for each Status is 0.05882353, 0.04380476, and 0.89737171. By using the unique function, unique(df.Historical$ Status), to see what Status SUNNYSIDE has, there are residential, mixed. Commercial and NA. But R could not display the proportion number of other Status, as Appendix 1 shows.

It is easy to use sd function, sd(SalePrice), to calculate the standard deviation of the sale price. By using this function, the standard deviation of the sale price for residential in SUNNYSIDE since 2009 is 2760438.

The result of the correlation between the sale price and gross square feet for residential properties in SUNNYSIDE since 2009, could be found by using cor(df.Historical2[c(-1)]). The result should be found in Appendix 2.

K-means clustering, comparing SUNNYSIDE and other neighborhoods

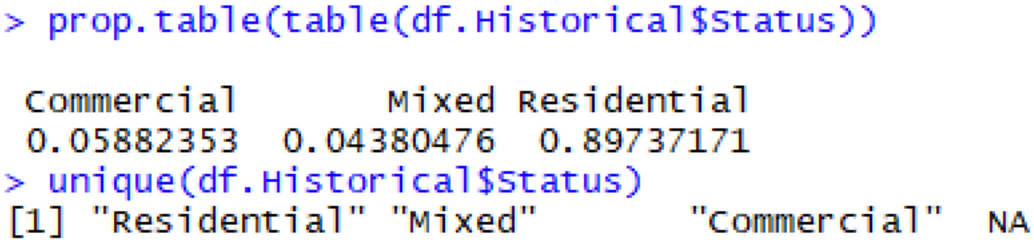
Using k-means clustering to group together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters could let the data be more visualized. Here, k represents the number of clusters. First, it plots total sale price, entire land, and unit price of all residential since 2009 by k-means clustering as Appendix 3. After filter the outlier, cluster 4 gets more land with a lower unit price than cluster 1 around the same total sale price. Cluster 2 has the lowest unit price than other clusters. Comparing cluster 2, cluster 3 has the lowest whole land at a higher price. Thus, the total sale price of cluster 3 is almost the weakest among those clusters. Cluster 5 has the most elevated entire land and total sale price, but the unit price of cluster 5 is not the most expensive. Therefore, cluster 5 using the low cost to sell more lands and got the highest total sale price.

Median sale price, the standard deviation of sales, and the amount of one gross square foot of residential could be the three KPI to compare SUNNYSIDE to other neighborhoods. As Appendix 4 shows, the x=Median sale price, y=Standard deviation, and the unit=One gross square foot of residential. The red dot represents SUNNYSIDE. Cluster 3 has the highest unit price comparing to other clusters. Under the same range of median sale price, cluster 4 has a more significant standard deviation than other clusters, which means the correlation of cluster 4 is smaller too. Cluster 2's standard deviation is more minor, with a lower unit price than Cluster 1, which implies cluster 2 is more stable than other clusters. Cluster 1 is also durable due to the lower standard deviation, but its price is in the middle place. SUNNYSIDE is similar to cluster 2. Because SUNNYSIDE has a small standard deviation and median sale price, but it's the unit price. It is also shallow. Overall, cluster 5 changes intensely, and clusters 1 and 2 are more stable. Cluster 3 has the highest unit price and median sale price. Cluster 4 is kind of stable with a central unit price and median sale price.

T-test for the average residential property costs in SUNNYSIDE and WHITESTONE

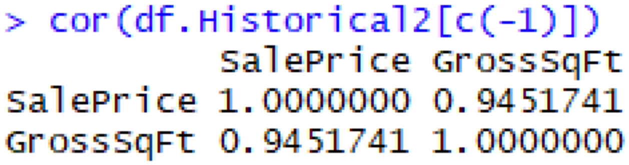
Comparing SUNNYSIDE and WHITESTONE by t-test for the average residential property costs by conf.level=0.95, SUNNYSIDE costs more since its mean sale price is $1129940.8, but WHITESTONE is $732785.9. Under conf.level=0.95, 95% confidence interval between SUNNYSIDE and WHITESTONE 252108.9 to 542200.9, which means 95% SUNNYSIDE will cost more than WHITESTONE between range 252108.9 to 542200.9. As Appendix 5 shows, the p-value of conf.level=0.95 is 9.054e-08. Comparing with conf.level=0.95, conf.level=0.99 has a lower p-value, 4.527e-08. But the confidence interval is changed to 224950.2 to infinity, which means 99% SUNNYSIDE will cost more than WHITESTONE between range 224950.2 to infinity, as Appendix 6 shows.

Appendix



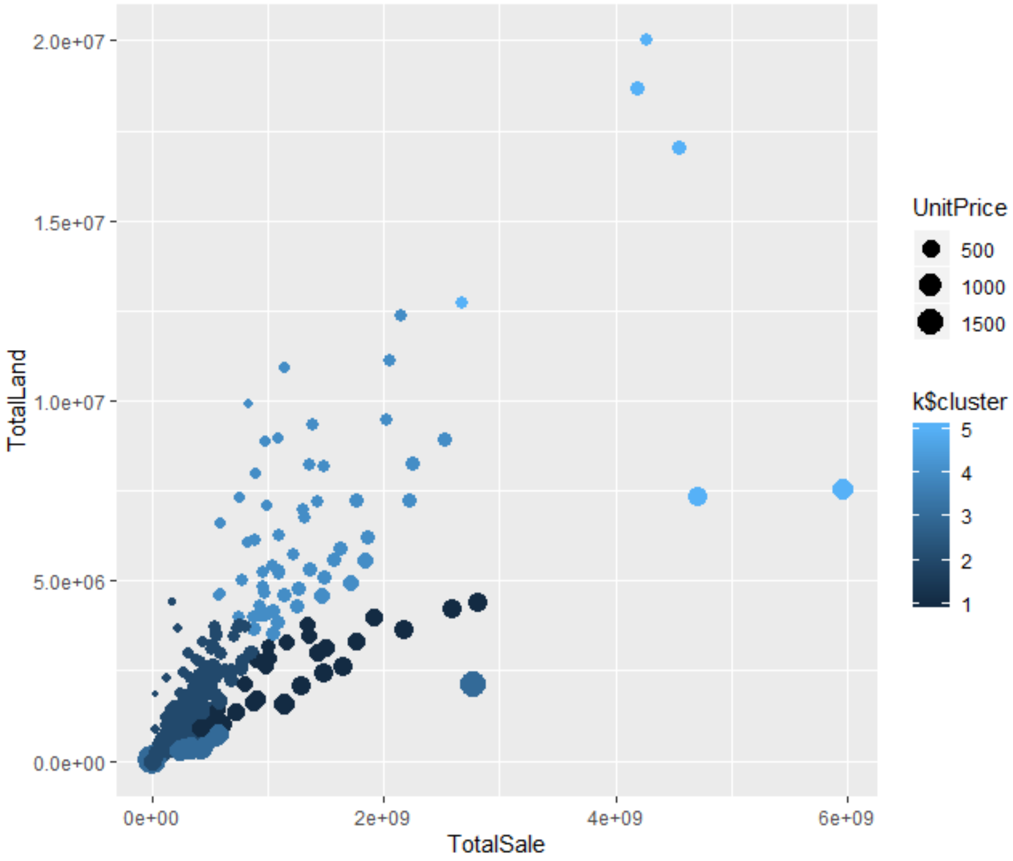
```
> prop.table(table(df.Historical$Status))

 Commercial       Mixed Residential 
 0.05882353  0.04380476  0.89737171 
> unique(df.Historical$Status)
[1] "Residential" "Mixed"       "Commercial"   NA
```

Appendix 1

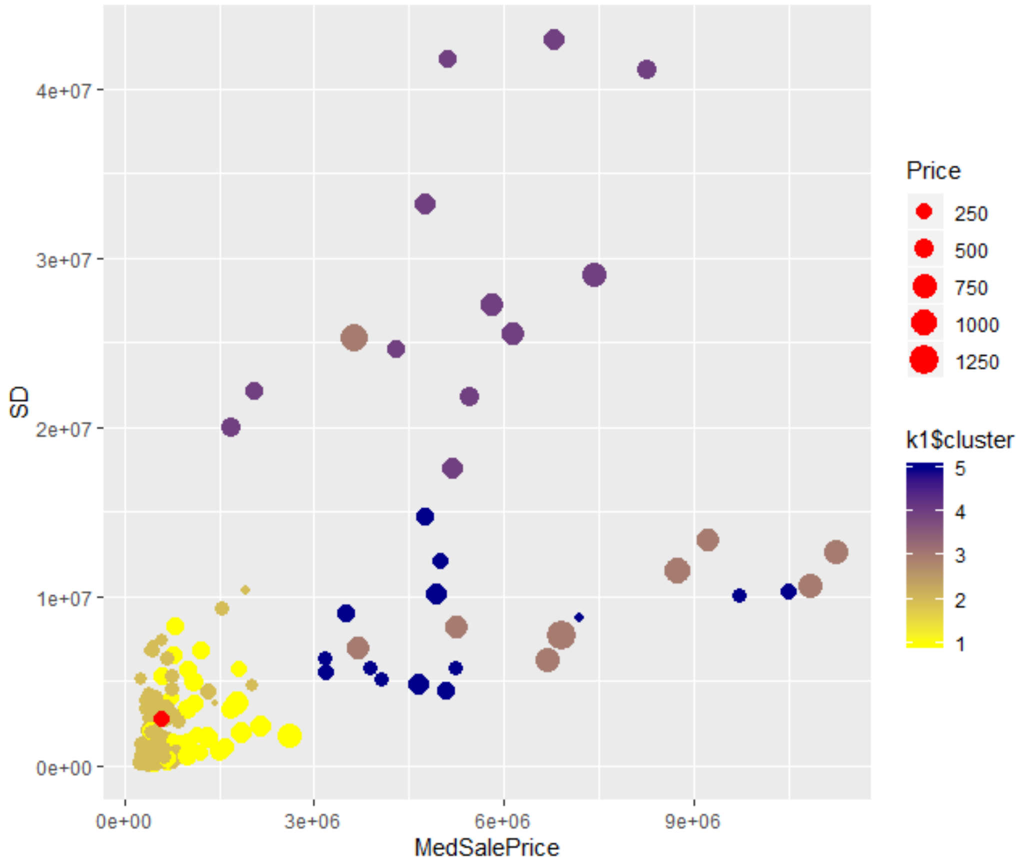


```
> cor(df.Historical2[c(-1)])
          SalePrice GrossSqFt
SalePrice 1.0000000 0.9451741
GrossSqFt 0.9451741 1.0000000
```

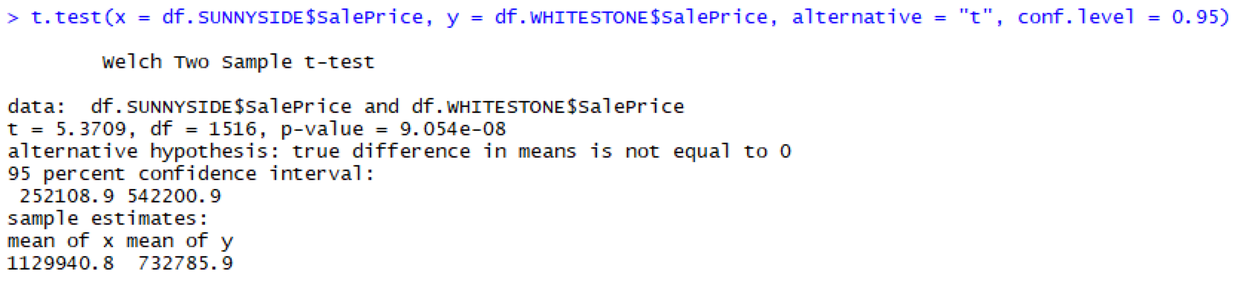
Appendix 2



Appendix 3

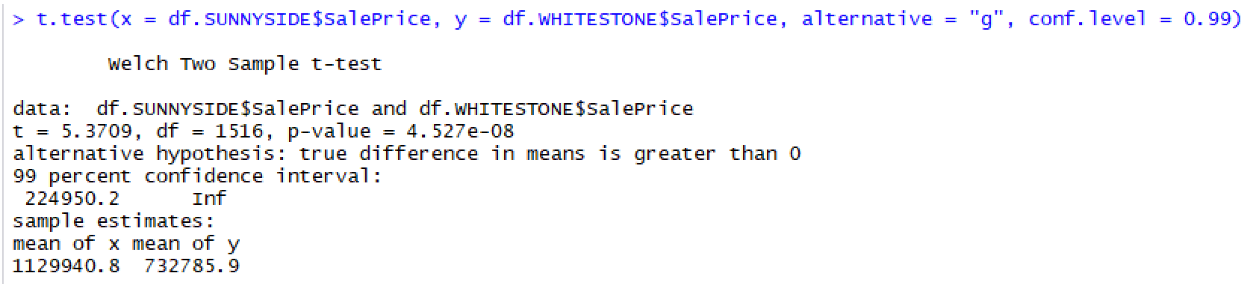


Appendix 4



```
> t.test(x = df.SUNNYSIDE$SalePrice, y = df.WHITESTONE$SalePrice, alternative = "t", conf.level = 0.95)

        Welch Two Sample t-test

data:  df.SUNNYSIDE$SalePrice and df.WHITESTONE$SalePrice
t = 5.3709, df = 1516, p-value = 9.054e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 252108.9 542200.9
sample estimates:
mean of x mean of y 
1129940.8  732785.9 
```

Appendix 5



```
> t.test(x = df.SUNNYSIDE$SalePrice, y = df.WHITESTONE$SalePrice, alternative = "g", conf.level = 0.99)

        Welch Two Sample t-test

data:  df.SUNNYSIDE$SalePrice and df.WHITESTONE$SalePrice
t = 5.3709, df = 1516, p-value = 4.527e-08
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 224950.2      Inf
sample estimates:
mean of x mean of y 
1129940.8  732785.9 
```

Appendix 6