661991197

Shiwei Huang

Data Analytics

Assignment 4

1. For any one of the Brooklyn, Manhattan, Queens sales datasets, perform the following:

a). Describe the type of patterns or trends you might look for and how you plan to model them. Describe any exploratory data analysis you performed. Include plots and other descriptions. Min. 2-3 sentences (2%)

I am picking Brooklyn data for this data analysis project. I identified the column that can be acted as an independent variable and the column that can be used as a dependent variable.

A screenshot of a cell phone

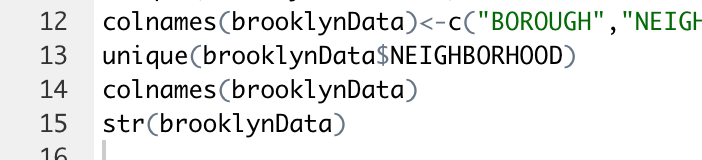
Description automatically generated

I want to get a unique neighborhood from the Brooklyn data:

And here is what I got:

A screenshot of text

Description automatically generated



A close up of a newspaper

Description automatically generated

The rest of the code snippets:

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

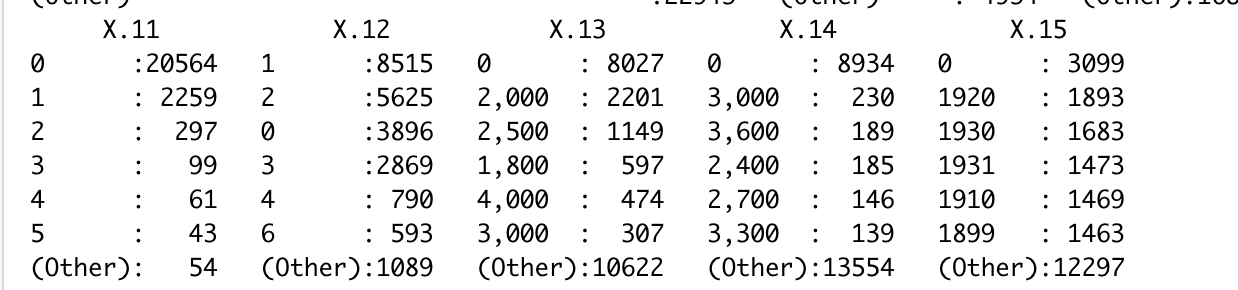
Description automatically generated

A screenshot of a cell phone

Description automatically generated

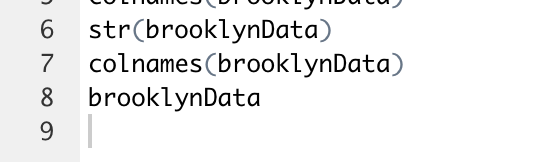
A screenshot of a cell phone

Description automatically generated



A screenshot of a cell phone

Description automatically generated



A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

b). Pick one or more models (these need not be restricted to the models you’ve learned so far [multivariate regression, KNN, K-Means]) to explore the chosen data. Interpret the model fits and indicate significance. Describe any cleaning you had to do and why. Min. 2-3 sentences (3%)

I want to use the multivariable regression model. I want to predict sales pricing and all the other variables.

For the data cleaning, I removed extra rows.

Then I removed extra columns.

I analyzed sale price column and removed all the rows with 0 as a value for sale price column.

Results:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A picture containing animal

Description automatically generated

2. For your chosen dataset:

a). Apply the model(s) to predict quantities of interest (that you choose). Describe (contingency table) or plot the predictions. Min. 2-3 sentences (4000-level 5%, 6000-level 3%)

I predicted sale pricing with all the variables.

Then I selected only the variables with the appropriate p-value or high significance.

A screenshot of a cell phone

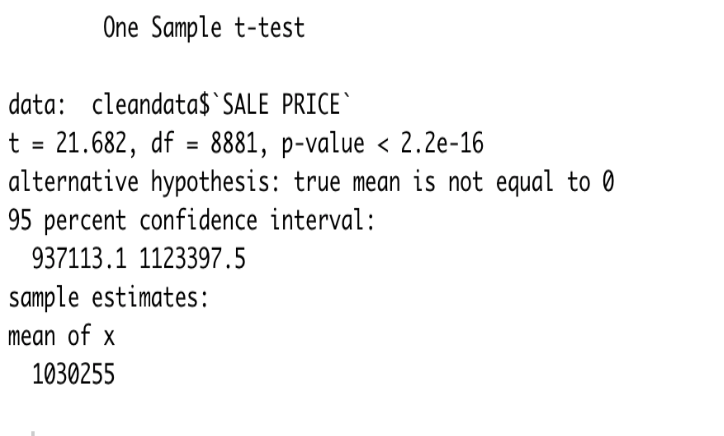
Description automatically generated

b). Examine the fit(s). Perform a significance test that is suitable for the variables you are investigating and describe the results. Min. 2-3 sentences (4000-level 4%, 6000-level 3%)

The results of my prediction is:

For the significance test, I performed T-test on sales price:

By Calling: t.test(cleandata$’SALE PRICE’)



A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated

c). Discuss any observations you had about the datasets/ variables, other data in the dataset and/or your confidence in the result. Min 1-2 sentences (1%)

The result is not as accurate as I thought it would be. Other models may give more accurate results.

3. 6000-level question (3%). Draw conclusions from this study – about the model type and suitability/ deficiencies. Describe what worked and why/ why not. Min. 4-5 sentences

My conclusion: the sale price might be dependent on the square feet, blocks, tax classes.

The multivariate regression model helped in finding the column on which the sale price might be the most dependent on.

Deficiencies: the model I predicted is not as accurate as I thought it would be because it showed a high significance with some columns and some rows. The prediction might not be accurate due to my data cleaning method.