Homework 2: Exploratory Data Analysis in IPython

Submitted By: Santwana

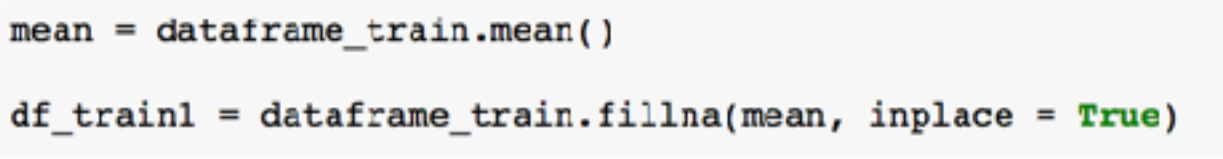
Student ID: 111498518

This Homework revolves around the prediction of the price that a real estate property would sell for. I joined the Kaggle challenge and submitted my model. Below is a summary on what I did to achieve it.

Step 1. The first step is to import all the relevant libraries for the implementation.

Step 2. Next step is to obtain the “Properties” dataset into a DataFrame.

Step 3. Further, I did Data Cleaning. The data set should have no inconsistencies. For example, it should have all numerical values and there should be no unknown/ NaN values. The code below solves the purpose. I changed the datatype from object type to Float64 for all numerical variables. Then I replaced the NaN values with the mean value for that column. All the inconsistencies found should be removed.



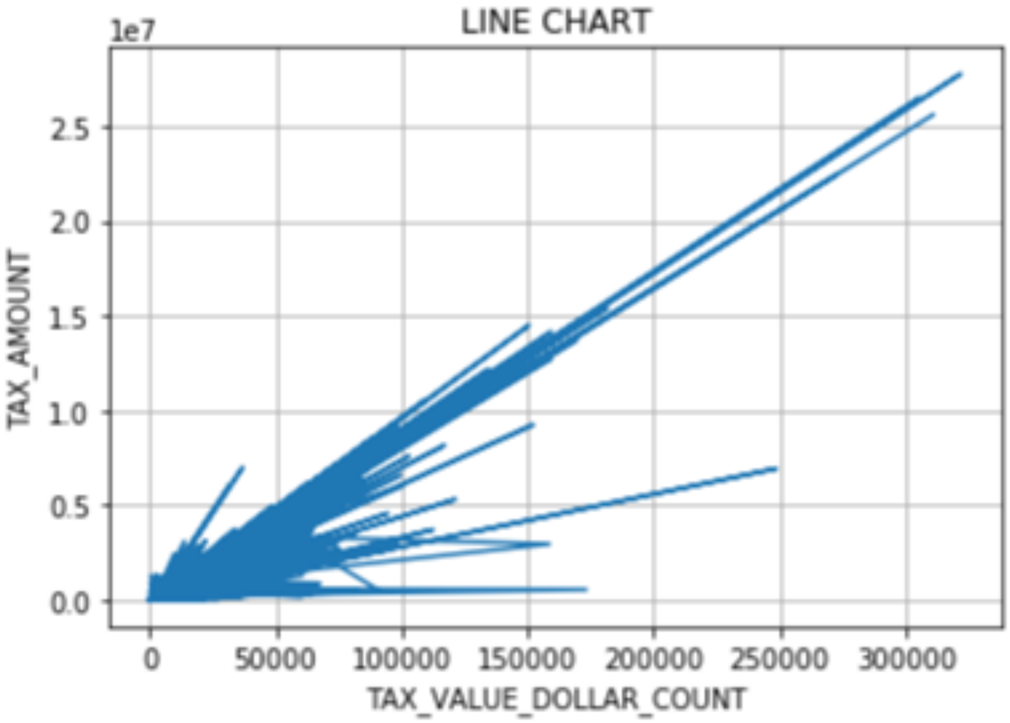
```
mean = dataframe_train.mean()

df_train1 = dataframe_train.fillna(mean, inplace = True)
```

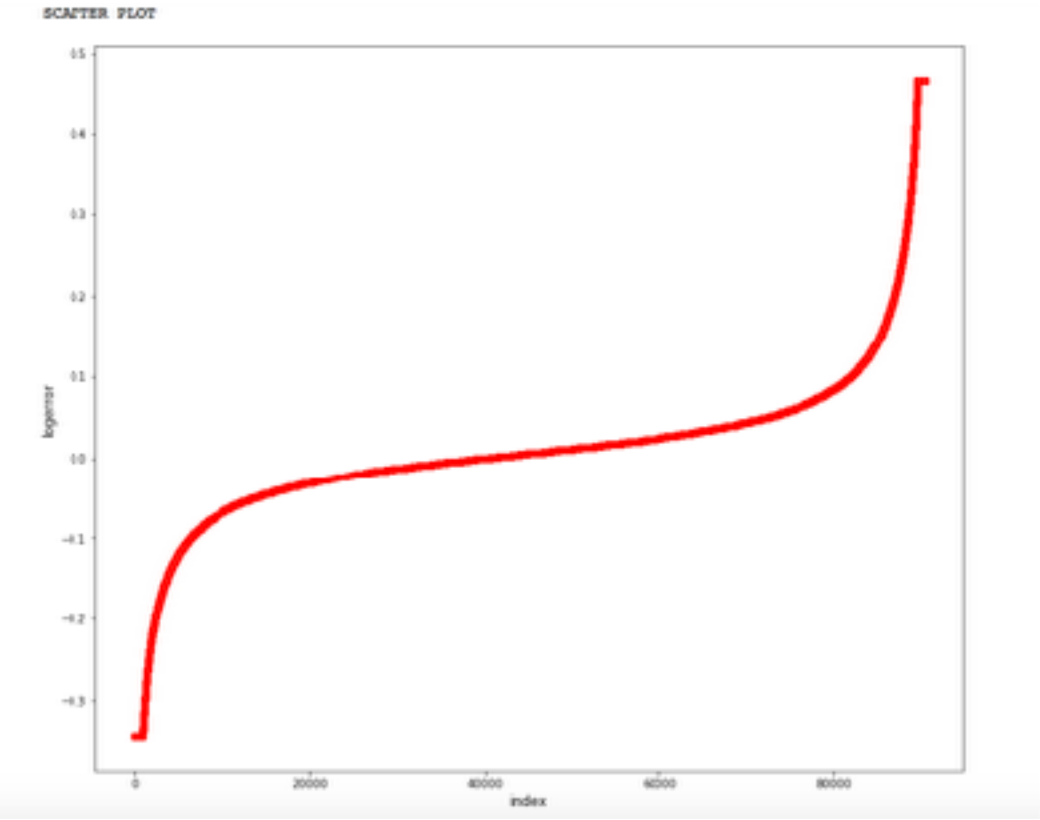
Question 1: Pairwise Pearson correlation analysis
The correlation analysis can be done on numerical values. The idea is to first separate the numerical variables for property so that we can do the analysis on pairs of valid attributes.
Here, I have done a correlation analysis on variables which I found interesting by Pearson method. The correlation between (x, y) is the same as of (y, x), thus, I calculated the correlation coefficients for such ordered pair just once. The strongest positive correlation came out to be 1 and -0.637.

Question 2: Data Visualisation

(i) Line Chart: I plotted a line chart between taxvaluedollarcnt and taxamnt variables. Below is the plot I got.

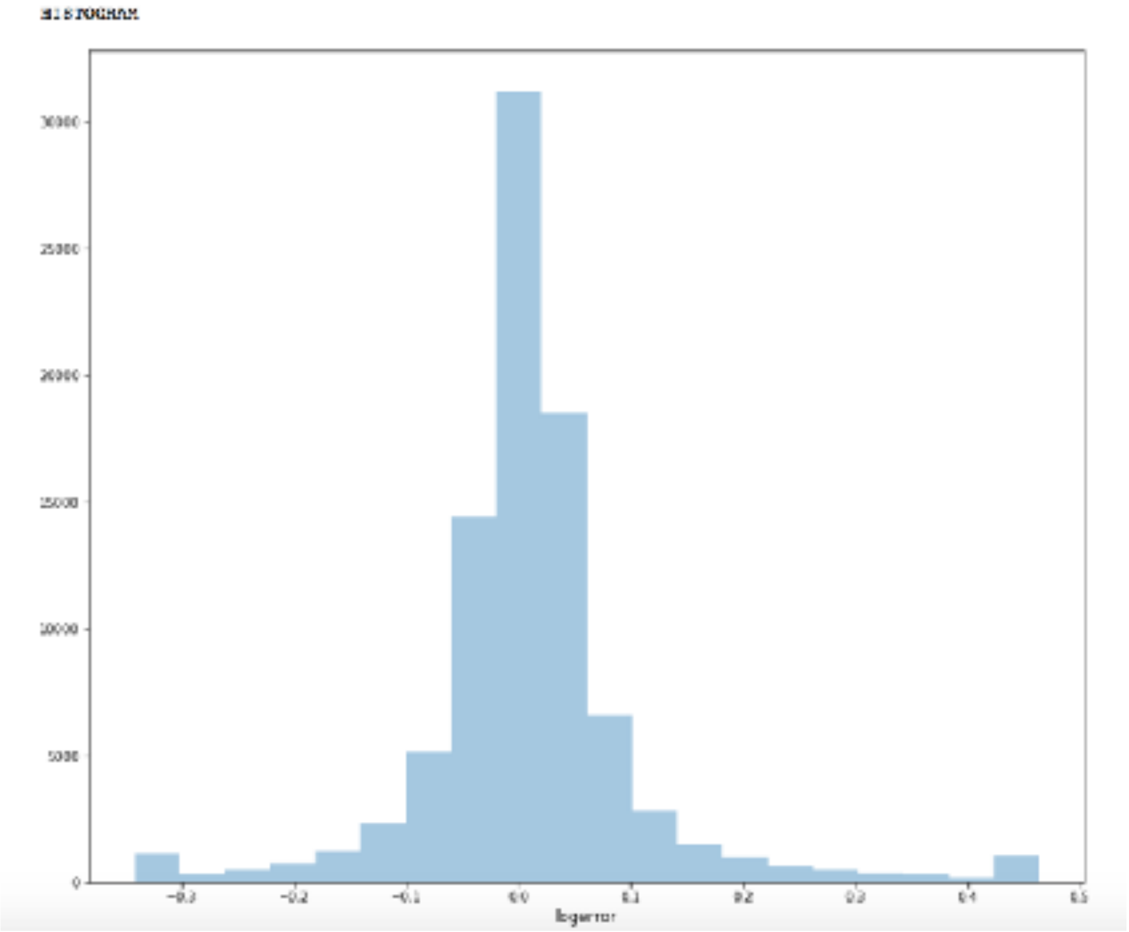


(ii) Scatter Plot
The Scatter plot is between index and logerror. From the Scatter plot, I observed that there is a strong positive correlation between the aforesaid variables, which



is of incrementing nature. The plot is non linear. Also, log error tends to have almost same value with increasing index.

(iii) Histogram
Below is a snapshot of the histogram I plotted.



Question 3: Setting up a Linear Regression Model:

Question 5: Predicting log error for instances at file sample_submission.csv & submitting on Kaggle:

Question 6: My favourite model:
As part of question 3, I have setup a simple linear regression model of the variables.The correlation function is used to measure the strength of the linear relationship between two variables.In the cases when such a linear relationship which exists is not enough, Linear Regression comes into the picture.
There are two types of data in statistical domain:

- Independent variables- data that can be controlled directly.
- Dependent variables - data that cannot be controlled directly.

Parameters are added to the model for estimating the output. When the relationship between dependent and independent variable can be expressed in a straight line, it is said to be Linear. Many parameters which we strongly believe to be linearly associated with the price are area and taxes.

The linear regression models are not accurate instead, they approximate the relationship between dependent and independent variables in a straight line. However, this has the scope of errors since its an approximation.

In my case for the current implementation, the mean squared error came out to be 0.02, which is a very decent value.

I encountered few surprising outcomes:
1. Variance score for the model calculated was almost zero, which I doubt and consider to be a case of over fitting. Thus the prediction seems to be very accurate. I found Linear Regression model to be quite efficient in prediction of the property value.
2. The scatter plot showed constant behaviour of log error wrt a range of values of index.
3. From the histogram I noticed, that almost 70% of out data shows quite a small log error which ranges from -0.1 to + 0.1