DATA INFERENCE AND APPLIED MACHINE LEARNING

OCTOBER 20, 2021

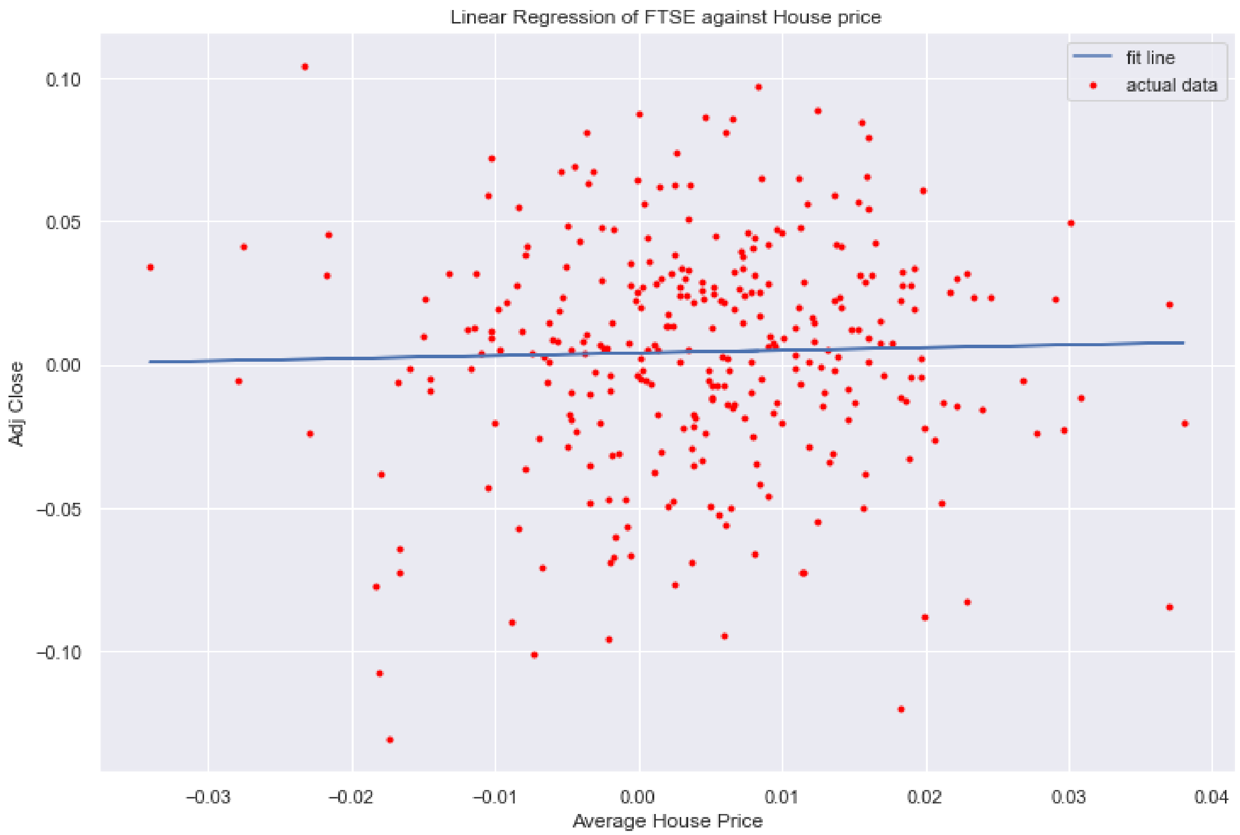
DEJI ADEBAYO

Python Libraries used:

- **Numpy**
- **Pandas**
- **matplotlib**
- **Math**
- **Scipy**
- **Statistics**
- **Stastsmodels**
- **Seaborn**
- **itertools**

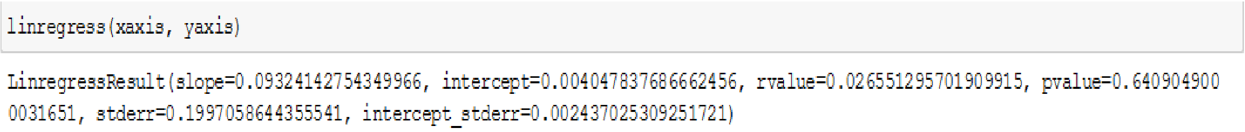
Question 1

I started by loading the data sets **house price.xls** (as independent variable) and FTSE100.csv (as dependent variable) into a data frame. For the house price dataframe I renamed the column with no header or title as Date and I selected the needed columns (Date, and Average house price) from the dataset loaded based on the needed date range, while for the FTSE100 dataframe, I converted the date column from date time to timestamp and selected the needed columns (Date and Adj Close) and sorted the data by date in the data frame. Having done the mentioned cleaning/manipulation on the two datasets (House Price and FTSE), I merged the two datasets on date column, and I calculated the percentage change on Average house price and Adj close for the merged data. I computed the linear regression and fitting by importing the sklearn library. The graph below was plotted using the dependent and independent variables (xaxis and yaxis) in the program.



After plotting the graph above I used the Lin regression imported from SciPy.Stats library to calculate values for the model; The image below shows the values obtained after computing lin regression. From the values obtained we have the model below:

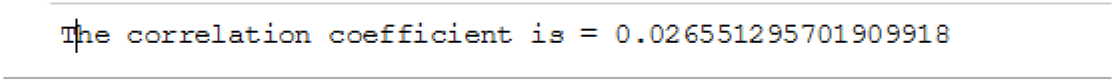
Y = 0.00404 + 0.0932x1



```
linregress(xaxis, yaxis)
```

```
LinregressResult(slope=0.09324142754349966, intercept=0.004047837686662456, rvalue=0.026551295701909915, pvalue=0.640904900
0031651, stderr=0.1997058644355541, intercept_stderr=0.002437025309251721)
```

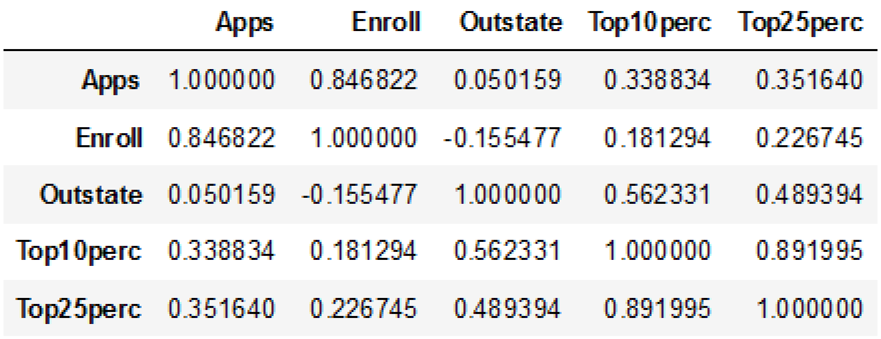
For this question, the null hypothesis will be: “**There is no significant relationship between the two variables.”** And alternative hypothesis is: “There is a significant relationship between the two variables.” From the obtained results (correlation coefficient= 0.0266 and P-value= 0.6409) I can say that there is no significant relationship between these 2 variables, hence accepting the null hypothesis. The image below shows the correlation coefficient value obtain.



```
The correlation coefficient is = 0.026551295701909918
```

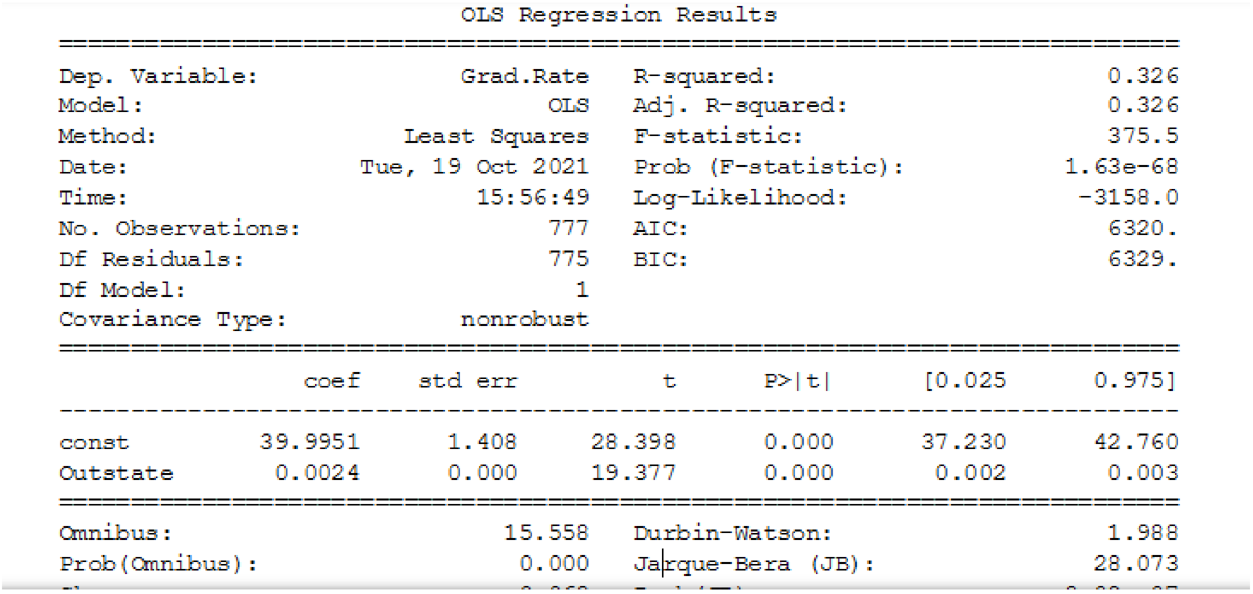
Question 2 Solution:

a) I started by reading the dataset college.csv and extracting all the necessary columns for this question into a data frame. Next, I calculated the correlation coefficient on this data and obtained the table below:



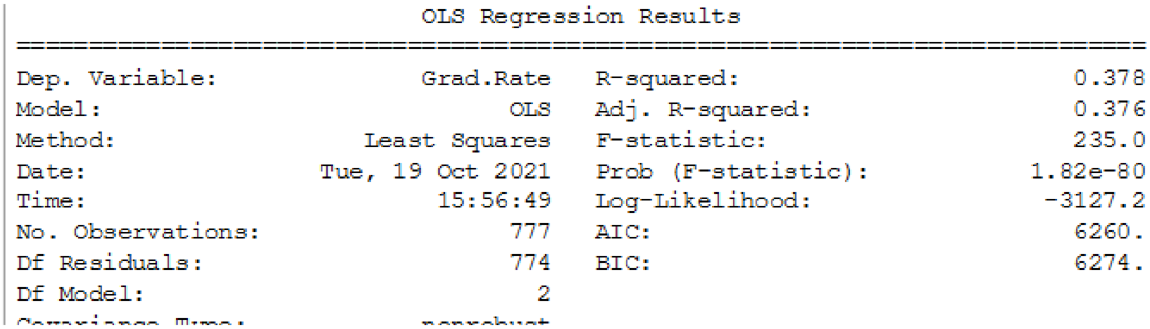
	Apps	Enroll	Outstate	Top10perc	Top25perc
Apps	1.000000	0.846822	0.050159	0.338834	0.351640
Enroll	0.846822	1.000000	-0.155477	0.181294	0.226745
Outstate	0.050159	-0.155477	1.000000	0.562331	0.489394
Top10perc	0.338834	0.181294	0.562331	1.000000	0.891995
Top25perc	0.351640	0.226745	0.489394	0.891995	1.000000

b) Next, I calculated the linear regression model using the stepwise function and below are the sample values obtained from the stepwise function summary:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Grad.Rate   R-squared:                       0.326
Model:                            OLS   Adj. R-squared:                  0.326
Method:                 Least Squares   F-statistic:                     375.5
Date:                Tue, 19 Oct 2021   Prob (F-statistic):           1.63e-68
Time:                        15:56:49   Log-Likelihood:                -3158.0
No. Observations:                 777   AIC:                             6320.
Df Residuals:                     775   BIC:                             6329.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.9951      1.408     28.398      0.000      37.230      42.760
Outstate       0.0024      0.000     19.377      0.000       0.002       0.003
==============================================================================
Omnibus:                       15.558   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.073
```

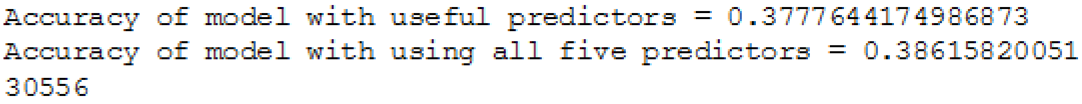
c) To know the predictor variable useful for predicting the graduation rate, I used the Bic function, and I found out that the important variables useful for predicting the graduation rate are percentageTop25 and outstate and I obtained them after using the model and it selected those with high correlation coefficient that are useful to what it wants to predict. And in addition, using stepwise it gave the same variables.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Grad.Rate   R-squared:                       0.378
Model:                            OLS   Adj. R-squared:                  0.376
Method:                 Least Squares   F-statistic:                     235.0
Date:                Tue, 19 Oct 2021   Prob (F-statistic):           1.82e-80
Time:                        15:56:49   Log-Likelihood:                -3127.2
No. Observations:                 777   AIC:                             6260.
Df Residuals:                     774   BIC:                             6274.
Df Model:                           2
Covariance Type:            nonrobust
```

d) Yes, the set of predictor variables would be useful in predicting the graduation rate, because BIC (Bayesian information criteria) select the variables that are useful and necessary for the model while predicting.

e) Comparing the accuracy of the model using useful predictors (Outstate and Top25perc) against the accuracy of the model using all five predictors, it appears that the model with smallest values is the most accurate. The image below shows the values obtained when the models are compared:



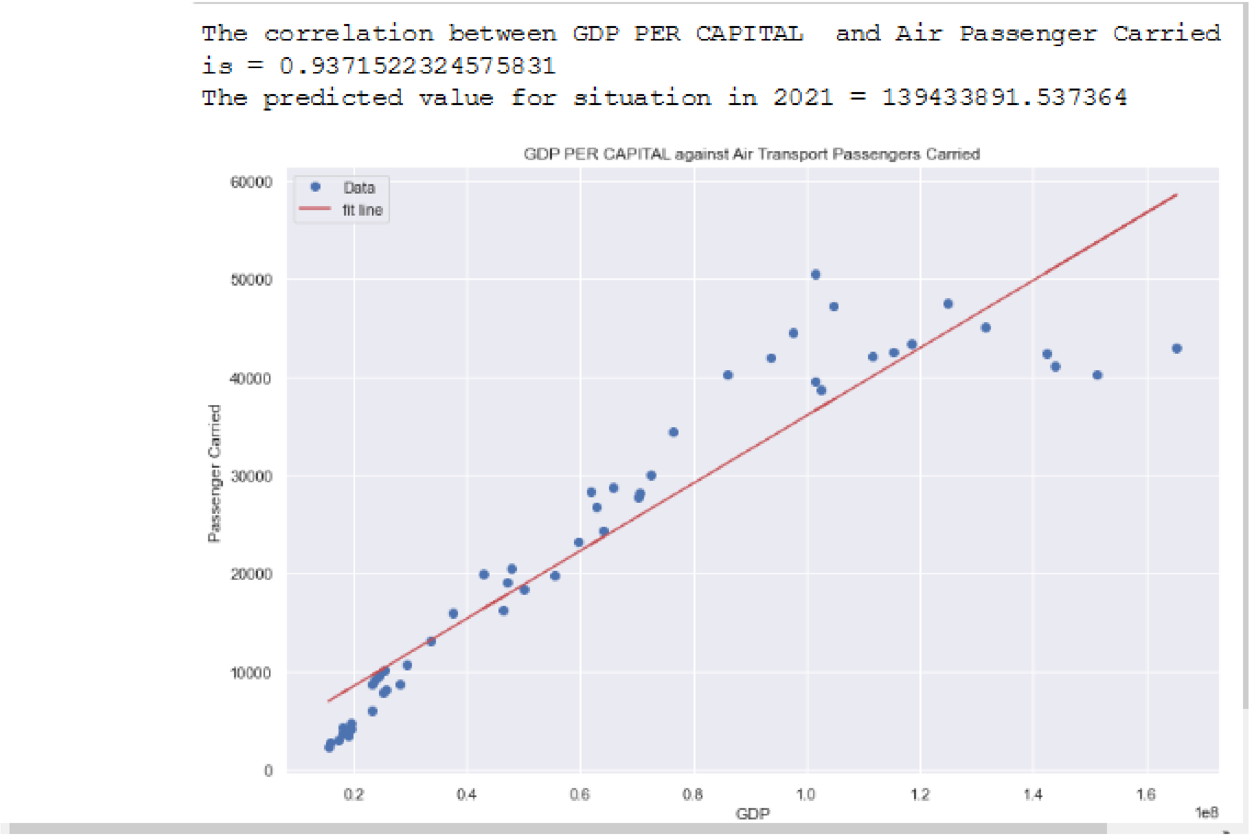
```
Accuracy of model with useful predictors = 0.3777644174986873
Accuracy of model with using all five predictors = 0.38615820051
30556
```

f) Given a set of predictors corresponding to Carnegie Mellon University, the graduation rate value that should predict the accurate model is: 89.20112305346854

Question 3 Solution:

Using world bank indicator data, I downloaded two datasets Air Transport, passengers carried and GDP per capita respectively, and I selected United Kingdom as my case study. This study aim to Check if the increase in GDP affected the increase in passengers in air transport. Next, I scattered GDP vs air transport, passengers carried then calculated the correlation coefficient to see the relationship between these variables. I continued by predicting the situation in 2021, the image below shows the predicted value obtained for situation in 2021

Using the correlation coefficient **(0.9371)** obtained, I assume that there exists a high dependence relationship between increase in GDP and the increase of passengers carried in air transport in United Kingdom over the years.



```
The correlation between GDP PER CAPITAL  and Air Passenger Carried
is = 0.9371522324575831
The predicted value for situation in 2021 = 139433891.537364
```

Source Datasets:

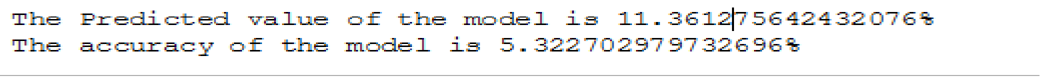
https://data.worldbank.org/indicator/IS.AIR.PSGR

https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Question 4 Solution:

I connected to Quandl using the api key. This was done to get the unemployment data with the code (ODA/ISR_LUR) provided. After connecting to the dataset, Next, I reset the index and selected the needed data into a data frame using the location or position of the dataset. I realized the date provided in the dataset is in a date time format which will not be good for the prediction required, hence, I changed the date format from data time to ordinal format.

Next, I used NumPy library to reshape or transform the output and input variable into a two-dimensional array and I stored each data into a variable (xaxis and yaxis). This variable was used to calculate the linear regression of the employment data. In other to estimate the likely rate of unemployment in the year 2020, I extracted the needed data for the year 2020, where I also perform the following operations as stated earlier: reset index, change date from data time to ordinal format and transform the x variable to two dimensions. Next, I, predict the likely rate of unemployment in the year 2020 and I calculated MAPE (Mean Absolute Percentage Error) which was what I used to obtain the accuracy of the estimate. The image below shows the values obtained.



```
The Predicted value of the model is 11.361275642432076%
The accuracy of the model is 5.322702979732696%
```