EVALYNE LWOBA

DIAML

ASSIGNMENT 4

ANDREW ID: elwoba

Libraries used

```
import numpy as np
import pandas as pd
import math
from scipy import stats
import matplotlib.pyplot as plt
from scipy.stats import linregress
from scipy.optimize import curve_fit
from scipy.stats import pearsonr
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
import time
from sklearn.metrics import mean_squared_error
import datetime
plt.rcParams["figure.figsize"]=(10,8)
```

Question 1

a) I used the data from the last assignment (Homework 3) in this homework. I created a sub data frame using only the required columns. Set index and extracted data for the required period. I calculated monthly returns on FTSE100 and Housing data frames. I created a scatter plot of the monthly returns of both data sets.
Using linregress function from scipy I created a linear regression model using FTSE100 and UK monthly house prices as the independent variable. The following statistics were generated from the fitted linear regression model:

```
slope: 0.09324142754349966 p-value 0.6409049000031651
std_err 0.1997058644355541 intercept: 0.004047837686662456 r_value
0.026551295701909915
```

I called the .fit() method on the lin regress model on x variables and generated corresponding y variables by running a prediction on the fit x variables. This was stored on a varible which was used to plot the estimated regression line.

b) I used Pearsons correlation function to calculate the correlation coefficient of the two variables . The result is:

```
Pearsons correlation: 0.0266
```

This result from person correlation infers that the FTSE100 monthly returns has a very weak correlation with the UK monthly house prices. There is possibly no relation between the two variables.

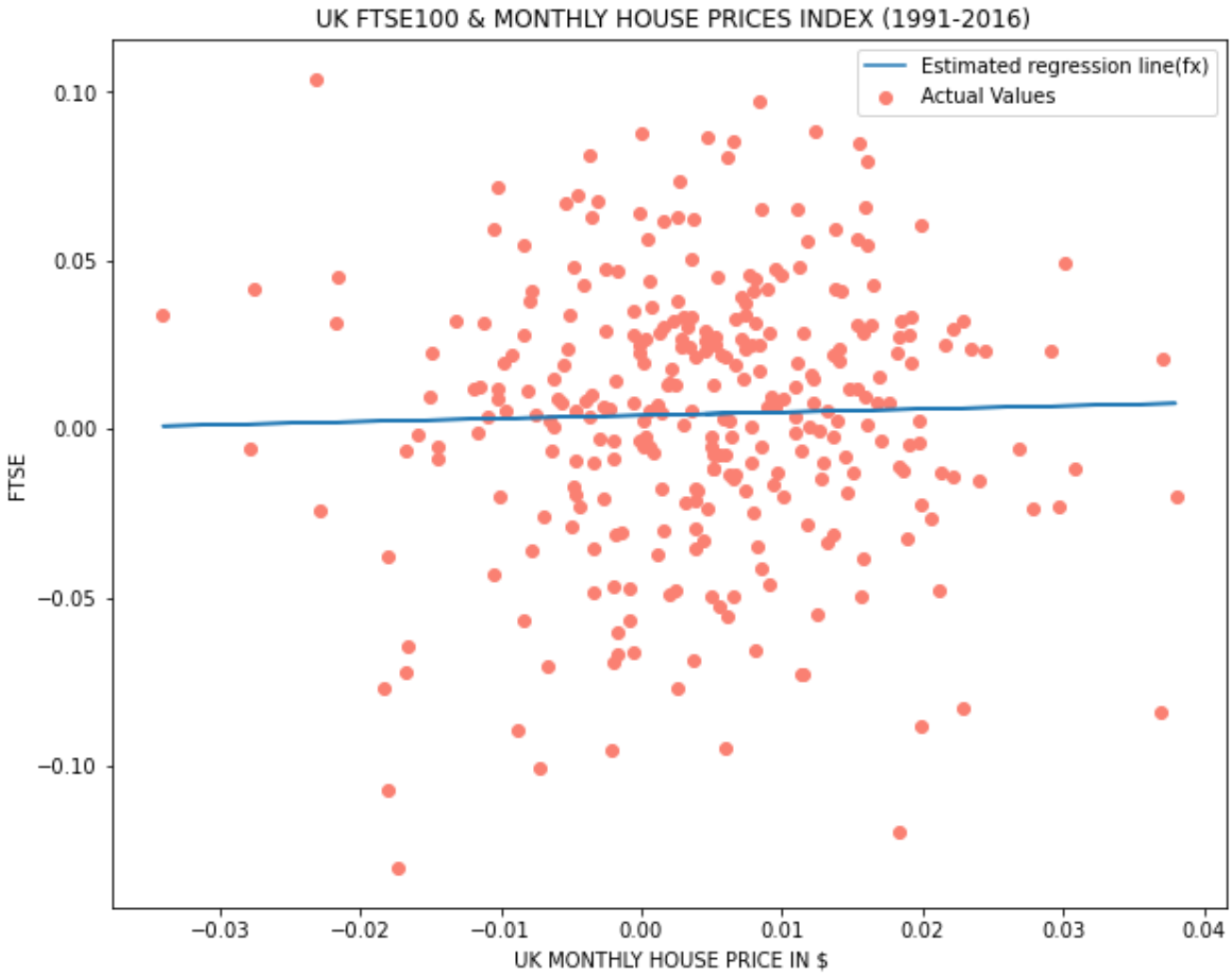
c) In order to do Hypothesis testing I have 2 hypothesis statements:

H0: There IS NOT a significant linear relationship(correlation) between x and y in the population

H1:There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population

From the linregress calculations, we have a **p-value of 0.6409** which is greater that our alpha of alpha of 0.05. this clearly means we accept the null hypothesis infer that there **IS NOT a significant linear** relationship(correlation) between FTSE100 and monthly housing prices in the population.

Result Graph Q1.



Question 1

Graph 1 Findings

The graph above shows no relationship between the UK monthly house prices and FTSE.

Question 2

Methodology and results

I loaded the required data set and put it in my working environment. I created a sub data frame using the required columns. The general statistics of the data are as below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Grad.Rate   R-squared (uncentered):                   0.946
Model:                            OLS   Adj. R-squared (uncentered):              0.946
Method:                 Least Squares   F-statistic:                              2707.
Date:                Tue, 19 Oct 2021   Prob (F-statistic):                        0.00
Time:                        15:22:26   Log-Likelihood:                         -3243.1
No. Observations:                 777   AIC:                                      6496.
Df Residuals:                     772   BIC:                                      6519.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Apps       -5.856e-07      0.000     -0.002      0.998      -0.001       0.001
Enroll         0.0015      0.001      1.251      0.211      -0.001       0.004
Outstate       0.0032      0.000     20.955      0.000       0.003       0.003
Top10perc     -0.5961      0.061     -9.795      0.000      -0.716      -0.477
Top25perc      0.8178      0.046     17.721      0.000       0.727       0.908
==============================================================================
Omnibus:                       26.233   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               55.792
Skew:                           0.157   Prob(JB):                     7.67e-13
Kurtosis:                       4.275   Cond. No.                     1.47e+03
==============================================================================
```

Using corr() and generated a matrix of correlation of all the required variables.

Correlation Result

	Apps	Enroll	Outstate	Top10perc	Top25perc	Grad.Rate
Apps	1.000000	0.846822	0.050159	0.338834	0.351640	0.146755
Enroll	0.846822	1.000000	-0.155477	0.181294	0.226745	-0.022341
Outstate	0.050159	-0.155477	1.000000	0.562331	0.489394	0.571290
Top10perc	0.338834	0.181294	0.562331	1.000000	0.891995	0.494989
Top25perc	0.351640	0.226745	0.489394	0.891995	1.000000	0.477281
Grad.Rate	0.146755	-0.022341	0.571290	0.494989	0.477281	1.000000

I used backward stepwise regression to choose the best features. This model keeps features that does reduce the model error significantly. Once a feature is dropped, it cannot be added back into the model. The model dropped 1 feature (Top10prec) and retained the other 4 which we used to test model accuracy.

Results

```
Drop Top10perc                      with p-value 0.446344
Backward stepwise generated useful variables for prediction are: ['Apps', 'Enroll', 'Outstate', 'Top25perc']
```

I also used LassoLars model and BIC criterion to choose the best features. This model dropped the fourth feature (Top25Perc) which had a p value of 0 wchich meants it didn't have any relationship with the graduation rate. The following were the p value results:

```
['Apps', 'Enroll', 'Outstate', 'Top10perc'] were retained


[ 0.00082806 -0.00293584  0.00186158  0.          0.17922103]
```

To test the accuracy I used 3 different models. The first model using MAPE score on all the 5 features generated an accuracy result as below:

MAPE score using all 5 features : 29.31197418109906

The second model on best features of backward stepwise generated the following results.

MAPE score using backward stepwise useful features: 29.4263004020622

The third model on the best BIC features generated the following results:

MAPE score using BIC best features : 29.274957446567957

From the above the best model is backward stepwise which has the lowest value of the 29.2750

To predict the CMU graduation rate, I used backward stepwise regression model to calculate the predicted graduation rate. This model predicted a value which is greater than the actual value of 74.

```
CMU graduation rate is: [89.082926]
```

Question 3

From the world bank indicators data bank, I downloaded the following data sets and imported into my working environment.

a. Air transport Freight(Million tons per km)
b. Trade (% of GDP)

I extracted the 2018 column for both data sets and merged to form a sub dataset. I transposed my data frame to make it easier to manipulate and also have dates as a column. I dropped some rows and renamed the required columns. I then converted my data frames values from an object to numeric values for ability to perform statistical analysis. I plotted a scatter to visualize the pattern.

In order to assess any trends in the data sets, I formed the following hypothesis statements:

H0 : There is no significant linear relationship or correlation between air transport freight and the Trade(% of GDP) of a country

H1: There is a significant linear relationship or correlation between air transport freight and the Trade(% of GDP) of a country.

For analysis I performed the following calculation:

a. I used persons correlation from scipy.stats library to calculate the correlation between the two variables. The result is:

```
Pearsons correlation: -0.051
```

From this results we can infer that there is almost no linear correlation between the two variables.

a. I also calculate statistics by using linregress() method from scipy library by passing the x and y variables in the model Fitting on the x variable gave the below output:

```
slope: -0.01986101852061476 p-value 0.8298008337934937
std_err 0.09106172318718633 intercept: 64.38862475299034
r_value: -0.05134004996723577
```

To plot a line of regression, I used the following formula:

R = c +(x*p) where:

R is the line of regression

C is the constant (Intercept)

P is the slope

X is the values from variable x in our dataset

I used the results to plot the regression line with R as the x_variable and X as my y_variable.

b. To do a prediction for 2021, I created 2 models using Lineregressor method from sklearn.Liner_model library:

I reshaped my two varibles (x = trage gdp data while y=air transport) to form a 2D array which I fit in the Linear Regressor model. I created another variable to hold the date values. First model was used to fit date and gdp. I then used the model to do a prediction on 2021 which gave me the predicted gdp value for 2021. Secondly, I fitted the second model with gdp and freight data as my xand y. I used the initial x predicted value(gdp) to predict the Freight value (y predicted).The resulting values are:

```
2021 predicted values are GDP: [[49.21822527]]
                        Air Freight:  [[63.41110067]]
```

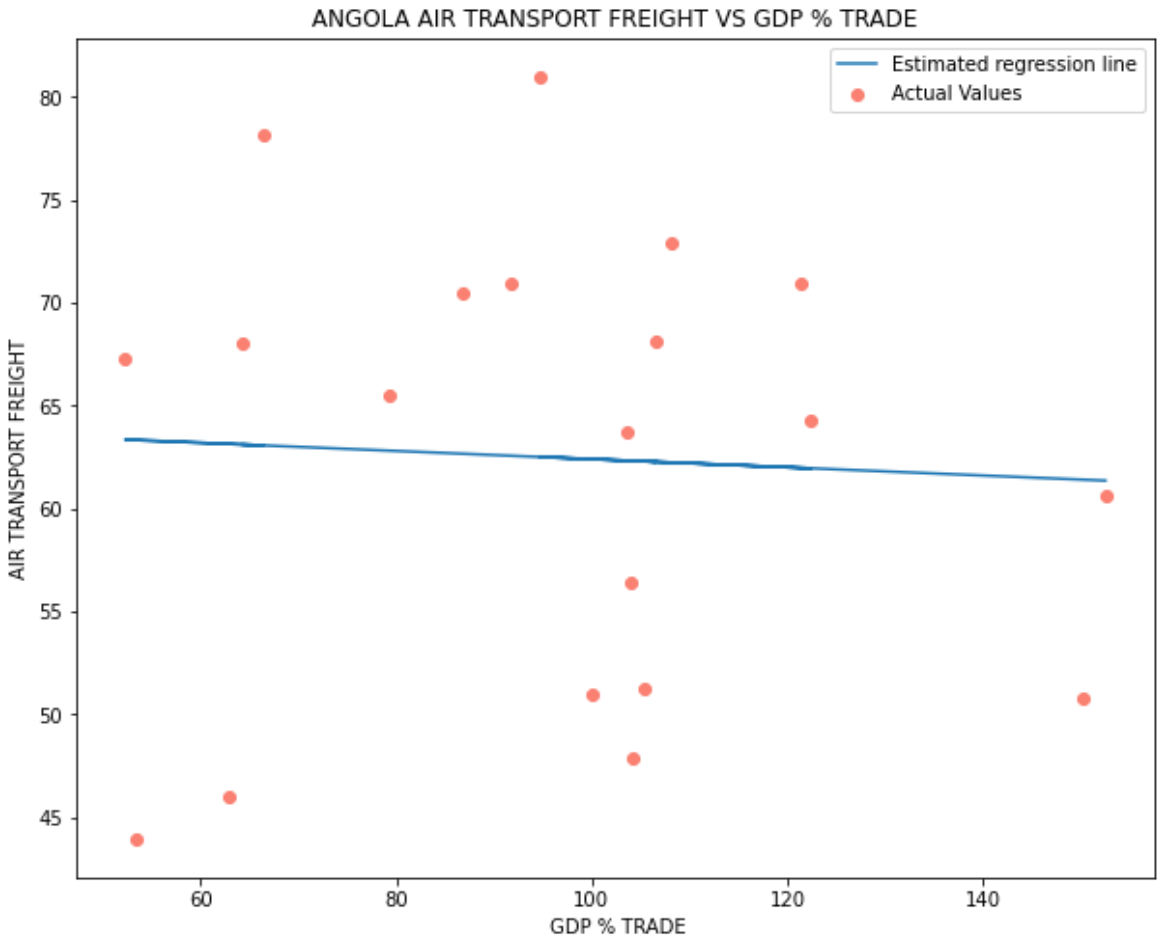
To test my hypothesis statements; I used the p value generated from the model above and the given alpha of 0.05. If :

p_value > alpha then we accept the null hypothesis but if

P_value is < alpha then we reject the null hypothesis

In this case p_ value is **p-value 0.8298008337934937** which is greater than **alpha of 0.05**. we therefore accept the null hypothesis and agree that there is not a significant linear relationship between air freight transport and the trade (% of GDP) of a country.

Graph Q3



Question 3

.

Finding.

The graph shows that there is a very weak negative nonlinear relationship between GDP % Trade and Air freight transportation in 2018.

Question 4.

I downloaded the data from quandl and set the start and end date in order to have a data set with the required date range. I created 2 arrays from the data set and created my X and Y variables. To be able to do some data manipulations, I changed the date column from a string format toa number using the toordinal function from datetime library. The .reshape() from numpy was also essential in creating an array for both variables.

I created a variable to hold the 2020 prediction date and converted it to a number using the toordinal function.

Data fitting is used in python to estimate the best representation of the actual data points through new predicted points. In our case I used a lineregression function from sklearn library to create a model to fit my data. I passed my X and Y variables into the model. To calculate the 2020 unemployement prediction, I passed the prediction date as a variable in the model and stored it in a variable as a 2D array.

The MAPE (mean absolute percentage error) is used to test the accuracy of a model as a percentage of the error. This is calculated by:

((y_actual – y_predicted) / y_actual |) * 100

A mape of 23% means that there is a 23% variance between the actual and predicted value.A

Results:

```
Intercept: [-221.49399182] slope : [[0.00031561]]
R squared: 0.2113665859769598

Unemployment rate prediction for 2020 is [[11.36127564]]
The MAPE score of the model is: 23.71 %
```

Findings

MAPE shows the accuracy as a percentage of the model error. This makes it easier to understand the statistics of a model.in our case our MAPE result is For example, if the MAPE is 23.71%, on average, which means the future forecast is off by 23.71%.

Reference

1. GDP data: https://data.worldbank.org/indicator/IS.AIR.GOOD.MT.K1
2. Air Freight Transport: https://data.worldbank.org/indicator/NE.EXP.GNFS.ZS?view=chart
3. Regression: https://machinelearningmastery.com/probabilistic-model-selection-measures/
4. Backward stepwise : https://github.com/AakkashVijayakumar/stepwise-regression/blob/master/stepwise_regression/step_reg.py
5. MAPE: https://www.statology.org/mape-python/