

REPORT OF ASSIGNMENT 4

**Carnegie
Mellon
University
Africa**

Submitted By: Peace Ekundayo Bakare

Course: Data, Inference, and Applied Machine Learning

Libraries Used:

Pip Installed quandl, pandas, matplotlib, numpy

1. *Numpy as peacenp for calculations.*
2. *Import stats from scipy.*
3. *Matplotlib.pyplot as peaceplt for scatter plot diagram.*
4. *Pandas as peacepd for dataframes and operations*
5. *LinearRegression from sklearn.linear_model*
6. *Train_test_split from sklearn.model_selection*
7. *Mean_squared_error, mean_absolute_percentage_error from sklearn.metrics*
8. *Datetime*
9. *Scipy.stats*
10. *Seaborn*
11. *SARIMAX from statsmodels.tsa.statespace.sarimax*

QUESTION 1 – Linear Regression with one explanatory variable

Download, Read and Cleaning of the Dataset

After downloading the two datasets, *Monthly.xls* and *FTSE100.csv*, I read it into Google Collaboratory (Colab) tool using the appropriate pandas functions *read_excel* and *read_csv* respectively.

At first, I observed that the UK House pricing data does not have the date column named, so I renamed the column. In addition, I filtered the data using *start_date* set to 1991-01-01, and *end_date* set to 2016-12-31, to select the data within the needed period.

For the FTSE100 data, I observed that the date column data type is of the object type, I had to convert it to datetime format, and sort the date using the Date column in ascending order. The data was in descending order i.e. 2016 to 1991.

(a) Identification of Dependent and Independent variables

Dependent Variable – FTSE100 Index data

Independent Variable – Monthly House prices data in the UK

Calculate the monthly returns for each variable

To calculate the monthly returns for each variable, *df_UK_house_prices_within_2016* and *df_FTSE100_index_sorted*, I used the formula $r(t) = \frac{p(t)}{p(t-1)-1}$, creating a new column 'Monthly Returns' in the dataframe, using the 'Average House Price' column in the, *df_UK_house_prices_within_2016* dataframe and the 'Adj Close' column in the *df_FTSE100_index_sorted* data frame.

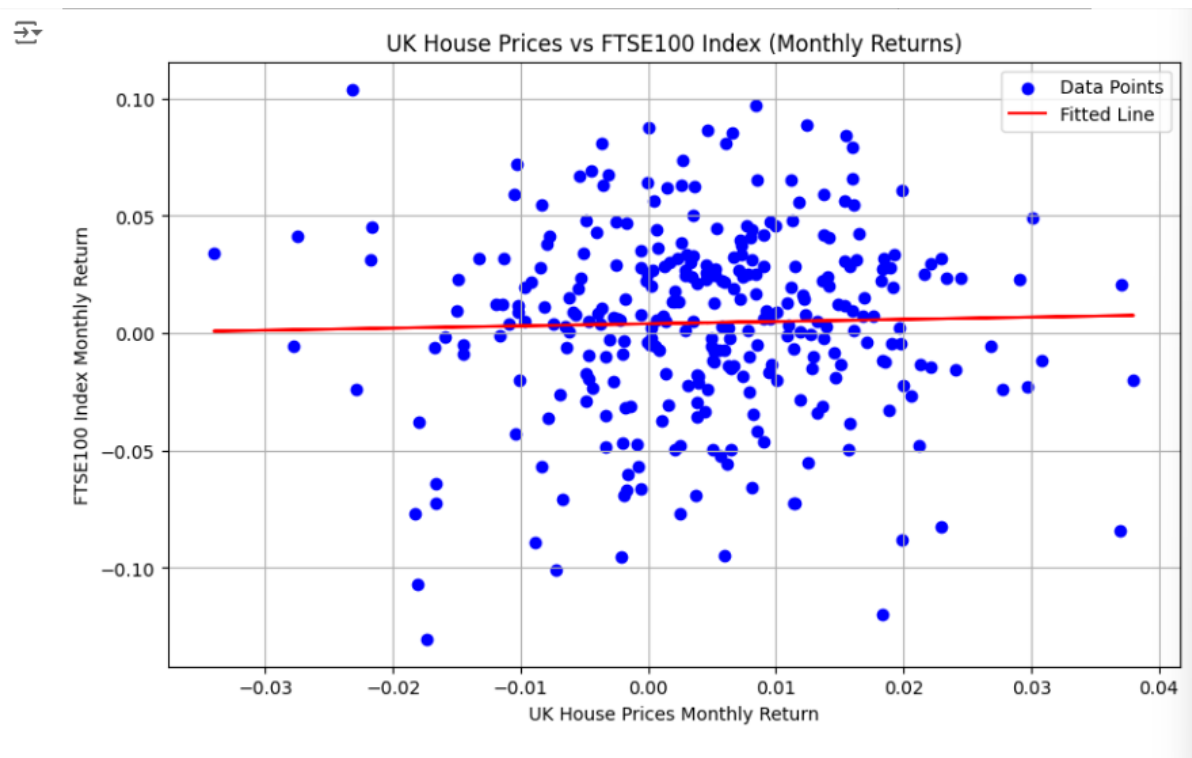
Create the regression model

I used the *linregress* function in the *scipy.stats* library to find the slope, intercept, *r_value*, *p_value* and standard error. The function takes in the dependent and independent variable data.

Plot the actual and predictions on a scatter plot

I plotted the graph of UK House Prices vs. FTSE100 Index (Monthly Returns), having the FTSE100 Index Monthly Return on the y-axis and UK House Prices Monthly Return on the

x-axis. For proper visualization, the scatter plot sits on a grid, with a legend showing the data points representation and the fitted line.



Calculate the correlation coefficients

I used the *corrcoef* function in the numpy library to calculate the correlation coefficient. The value gotten is the same as the *r_value* gotten when I used *linregress* function in the scipy.stats library. The value of the correlation coefficient gotten is **0.026551295701909918**.

(b) Interpret your Results

The correlation coefficient gotten, $r = 0.02655$ is close to 0 than to -1 or +1. Hence, it indicates that the linear relationship between the dependent variable, Monthly House price data in the UK, and the independent variable, FTSE100 index data is **WEAK**. Correlation coefficient suggests the strength and direction of the linear relationship between two variables.

The wide scattering of the data points also suggests the weakness between the two variables. In addition, there is a fair clustering around (0, 0) data points, which suggests that there is low variability in the monthly returns for the two variables.

From the regression line, it can be seen that it is almost flat, which suggests that the slope is very small i.e. an insignificant change in the monthly returns of the House prices data and FTSE100 index.

Conduct a hypothesis test between the Dependent and Independent variables

Null hypothesis $H_0 : slope = 0$

Alternative hypothesis $H_1 : slope \neq 0$

The Null Hypothesis states that there is no significant relationship between the dependent and independent variables, hence slope = 0

The Alternative Hypothesis states that there is a significant relationship between the two variables.

Since the $p_value < 0.05$, the null hypothesis should be rejected indicating that there is a significant relationship between the monthly house prices data in the UK and the FTSE100 index data.

QUESTION 2 – Linear Regression with multiple explanatory variables

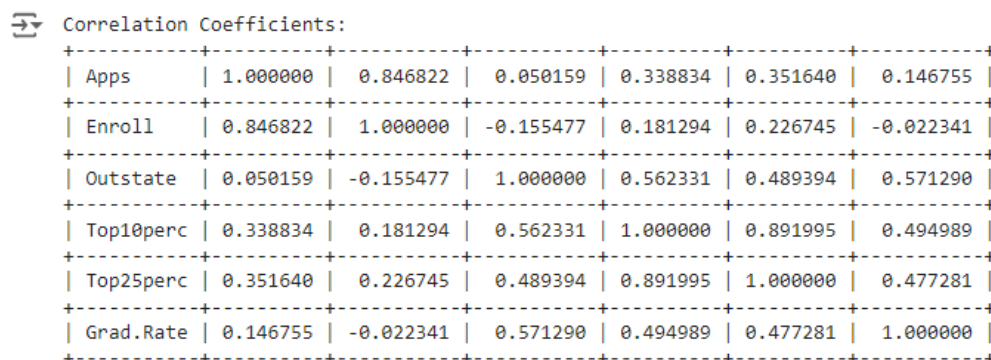
Download, Read and Cleaning of the Dataset

After reading the appropriate dataset *College.csv* using the *read_csv* function of the pandas library, and understanding what the dataset looks like, I renamed the column that contains the Universities as “Universities”.

I selected the columns needed as mentioned in the question namely, Apps, Enroll, Outstate, Top10perc, Top25perc, and Grad.Rate, and formed a dataframe for these columns, *df_college_needed*.

(a) Calculate the Correlation Coefficients of the Variables

I calculated the correlation coefficients between the variables, the number of Applications *Apps*, the number of students that enrolled *Enroll*, the number of out of the state students *Outstate*, the percentage of students that were admitted who were in the top 10% *Top10perc*, the percentage of students that were admitted who were in the top 25% *Top25perc* and the graduation rate *Grad.Rate*. I have presented the values in a tabular format.



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

(b) Use Stepwise to build the Linear Regression model

Using Stepwise to build the linear regression model involves an iterative process of adding or removing the predictor variables **on the basis of their statistical significance**. I set significant thresholds, *in_threshold* as 0.05 and *out_threshold* as 0.10. The **p-value** is the statistical significance here. A predictor is significant enough to be added if its p-value is less than the *in_threshold*. Also, a predictor is insignificant to be kept in the list if its p-value is higher than the *out_threshold*. Predictors are kept in a list if their significance remains until no more predictors can be added or removed from the list. I adapted the logic from a resource found on a GitHub repository [1]. The entire process can be summarized into 4 procedures:

- (i) Initialization – starting out with an empty model i.e. no predictors

- (ii) Forward Selection – evaluation of the p-value of each predictor that is currently not in the model and addition of the predictor with the lowest p-value
- (iii) Backward Elimination – evaluation of the p-value of each predictor that is currently in the model and removal of the predictor with the highest p-value.
- (iv) Iteration – repetition of the forward selection and backward elimination process until there is no more predictors that can be added or removed based on the thresholds earlier set.

(c) Useful Predictor Variables

The Useful Predictor variables were just two (2). They are:

- (i) **Outstate** – the number of out of the state students
- (ii) **Top25perc** – the percentage of students that were admitted who were in the top 25% in their class.

These are the variables that remained in the list i.e. final model after the stepwise selection process highlighted above. Their p-values are statistically significant in relation to the graduation rate, in that they were lower than the *out_threshold* set.

(d) Model Selection using Bayesian Information Criterion (BIC)

I calculated the Bayesian Information Criterion using the Ordinary Least Squares (OLS) function from the *statsmodels* library. I printed the summary of the fitted model using the *summary()* function, which includes the coefficients, p-values, R-squared, bic and other statistics. The relevant part of the statistics is the bic attribute which gives the BIC value for the model. The value returned is **6,274.3329824422635**

```

OLS Regression Results
=====
Dep. Variable:          Grad.Rate      R-squared:                0.378
Model:                  OLS            Adj. R-squared:           0.376
Method:                 Least Squares   F-statistic:             235.
Date:                   Tue, 15 Oct 2024 Prob (F-statistic):       1.82e-80
Time:                   17:30:24        Log-Likelihood:          -3127.2
No. Observations:      777             AIC:                    6260.
Df Residuals:          774             BIC:                    6274.
Df Model:               2
Covariance Type:       nonrobust
=====
                    coef    std err          t      P>|t|      [0.025    0.975]
-----
const          33.0860      1.607      20.593   0.000      29.932     36.240
Outstate       0.0019      0.000     13.658   0.000      0.002     0.002
Top25perc     0.2255      0.028      7.995   0.000      0.170     0.282
=====
Omnibus:                25.071      Durbin-Watson:           1.945
Prob(Omnibus):          0.000      Jarque-Bera (JB):        48.404
Skew:                   0.189      Prob(JB):                3.08e-11
Kurtosis:               4.163      Cond. No.                3.69e+04
=====

```

The entire process for using BIC to select the best model to predict the graduation rate involves:

- (i) Fitting the model using the selected predictor variables.
- (ii) Display of the model summary to have a better understanding of its performance
- (iii) Calculation of the BIC value to determine the balance between complexity and fitness.

The Lower the BIC value the better the model.

Answer: YES! The set of predictor variables would be useful in predicting the graduation rate if I were to use BIC to select the model because:

- (i) BIC suggests a good fit with fewer parameters.
- (ii) BIC balances model fitness and complexity, penalizing models with more parameters to avoid overfitting.
- (iii) When we compare the BIC value of the model using the selected predictor variables, Outstate and Top25perc with the BIC value of the model that used all the predictor variables, the BIC value of the former would be lower, and the lower the BIC value the better the model.

(e) Comparing Model Accuracy

To compare the accuracy of the model, I calculated the Mean Squared Error (MSE) using all the predictor variables and the Mean Squared Error using the selected predictor variables. The lower MSE indicates a more accurate model. So, the more accurate model is the model with the selected predictor variables – Outstate and Top25perc.

The MSE using all the predictor variables is **128.68965237614051**

The MSE using the selected predictor variables is **121.50146022950815**

To determine the MSE, I first split the data into test data and train data.

(f) Prediction of Graduation Rate for Carnegie Mellon University

I predicted the graduation rate for CMU using the most accurate model that I have identified, by extracting the data for Carnegie Mellon University and making the prediction.

The Predicted graduation rate for Carnegie Mellon University is **86.83968987**.

QUESTION 3

The aim of the study is to examine the impact of travel services (percentage of commercial service exports) on the GDP per capita of a country and predict the travel services in 2021.

Problem Statement:

The aim of this study is to examine how travel services (% of commercial service exports) affect a country's GDP per capita. Nigeria was used as a case study. In addition, the study forecast the travel services for the year 2021 to provide insights into trends that might affect economic performance.

Data Sources:

Both Dataset was downloaded from **World Bank Indicators**.

- Travel Services (% of commercial service exports) – ([Travel Services](#))
- GDP per capita (current US\$) – ([GDP per capita](#))

Assumptions:

1. The data from the World Bank Indicators is accurate and up-to-date.
2. GDP per capita is used as a determinant for national economic prosperity.
3. Travel services percentage reflects the strength of the tourism and related service sectors.
4. Linear relationships exist between travel services and GDP per capita for the purpose of the regression modeling.
5. Forecasting for 2021 is performed using the available data until 2020 and assumes that there is no sudden shocks in the economy.

Methodology, Statistical Analysis and Results

1. Data Preprocessing

I removed the unnecessary columns such as “Indicator Name”, “Indicator Code”. Also, I melted each of the datasets *df_gdp_per_capita* and *df_transport_data*, to a long format with the “Year” column as a common key. Then I merged the two datasets on “Country Name” and “Country Code” using inner join. Finally, I extracted Nigeria's data from the list of countries, for a focused time-series analysis.

2. Time-Series Analysis

I visualized trends for travel services and GDP per capita over the years using a line graph.

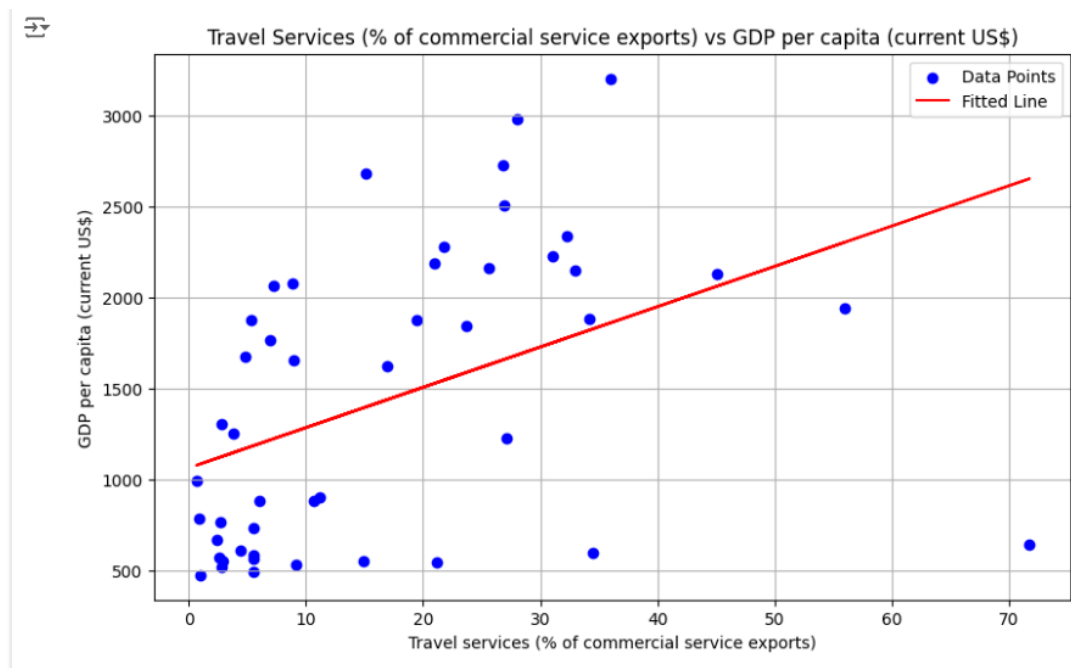
↓



From the time series above, it shows that travel services remains relatively flat and low throughout the period from 1978 to 2020, suggesting that travel services contribute only minimally to overall export activities. The divergence in trends suggests that while GDP per capita experienced periods of growth, the travel services sector has not been a major driver of economic performance during the observed years.

In addition, I conducted statistical modeling through linear regression to measure the correlation between travel services and GDP per capita.

↓



The scatter plot shows the relationship between travel services and GDP per capita. The red fitted line indicates a positive linear trend, suggesting a **weak to moderate positive correlation**. As the percentage of travel services increases, GDP per capita generally tends to increase as well, although the data points are widely scattered around the trend line. This dispersion suggests that other factors beyond travel services may significantly influence GDP per capita, and the correlation is not very strong.

3. Modeling and Forecasting

I built a linear regression model to evaluate the relationship between GDP per capita and travel services. Also, I used the ARIMAX () Model to forecast travel services for 2021 with GDP per capita as an exogenous variable. I used 80% of the dataset as training data.

Here is the summary of the ARIMAX model

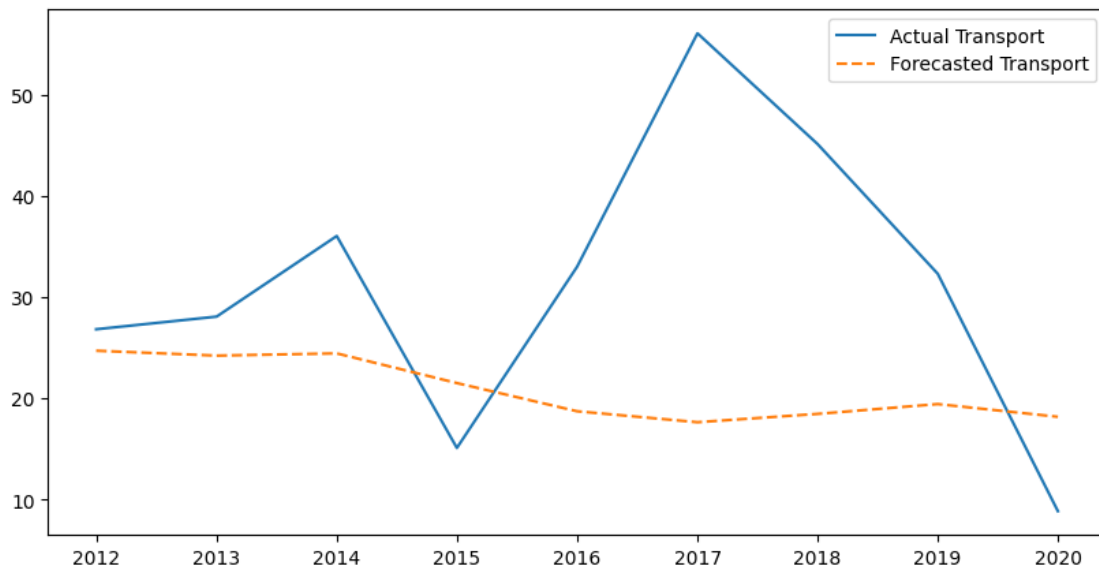
→

SARIMAX Results						
=====						
Dep. Variable:	Transport Data	No. Observations:	35			
Model:	SARIMAX(1, 1, 1)	Log Likelihood	-134.853			
Date:	Wed, 16 Oct 2024	AIC	277.706			
Time:	18:16:01	BIC	283.811			
Sample:	01-01-1977	HQIC	279.788			
	- 01-01-2011					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

GDP per Capita	0.0048	0.013	0.373	0.709	-0.020	0.030
ar.L1	0.5089	0.203	2.504	0.012	0.110	0.907
ma.L1	-0.9986	10.150	-0.098	0.922	-20.893	18.896
sigma2	151.7798	1541.059	0.098	0.922	-2868.640	3172.200
=====						
Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	326.58			
Prob(Q):	0.90	Prob(JB):	0.00			
Heteroskedasticity (H):	0.12	Skew:	3.35			
Prob(H) (two-sided):	0.00	Kurtosis:	16.63			

From the SARIMAX Model used (1, 1, 1) terms for both Autoregressive (AR), Moving Average (MA), and differencing components. The GDP per capita has a coefficient of 0.0048, but a high p-value of 0.709, suggesting that it is not statistically significant in explaining the dependent variable, Travel services (Transport Data).

ARIMAX RMSE: 17.741676950523487



The RMSE value of 17.741676950523487 suggests the magnitude of the prediction error. It is not acceptable because there is significant deviation from the actual value of the Travel services in 2021.

Prediction of the situation in 2021

The actual value of the transport data (travel services) in 2021 is **7.250054**. The predicted value is **21.537**, which shows that the model overestimates the actual value by a significant margin. I'll call the margin, Error (Absolute) $|21.537 - 7.250| = 14.287$. The large deviation shows that the model is not accurately capturing the underlying patterns or the influence of the external regressor (GDP per capita) for the year 2021.

There are possible causes of the overestimation, such as Overfitting, incomplete training, ARIMAX model limitations, insufficiency of GDP alone as a strong predictor for transport data etc.

QUESTION 4

Download, Read and Cleaning of the Dataset

I downloaded the data from Canvas named *Israeli_Unemployment_Rate.csv* and read it using *read_csv* function of the pandas library as *df_Israel_Unemployment*.

Processing the data

I converted the date column to datetime format, filtered the dataframe with *start_date* and *end_date* set at 1980-12-31 and 2013-09-02 respectively. I used the function *toordinal()* function in python to create a new column off the date column, called “Date Ordinal” and converted it to ordinal.

Fit Linear Regression Model

I identified the “Date Ordinal” as the independent variable X, and the “Value” column as the dependent variable y. I used the *LinearRegression()* function and fit the model.

To Predict the rate of Unemployment by 2020

I ensured that the year 2020 is in the same format as X by using the *toordinal()* function to do the conversion. I used the model to predict the unemployment rate in 2020. I got an approximated value of 11.94.

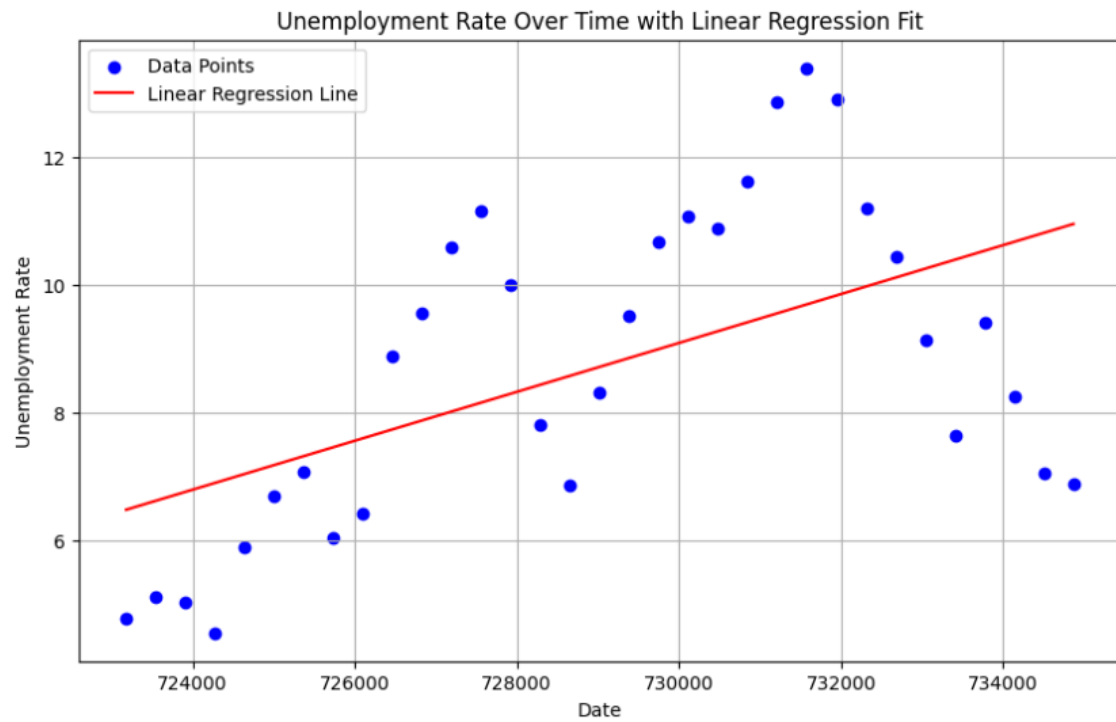
Evaluate Performance of the model

To evaluate the performance of the model, I calculated the Mean Absolute Percentage Error (MAPE) first by obtaining the predicted of y using the X value, and then calling the function *mean_absolute_percentage_error()* from the *sklearn.metrics* library.

I have also plotted a graph of Unemployment rate over time using data points from 1980-12-31 to 2013-09-02, also indicating a linear regression line fitted to the data.

The blue dots represent unemployment rate for specific dates in the dataset. The spread shows the variation in unemployment rates over the years. The line slopes upward indicating an increase in unemployment rate over the years. The implication is that there was increasing challenges faced by the labor market over the period identified.

11



References

- [1] A. R. V, *AakkashVijayakumar/stepwise-regression*. (Mar. 27, 2024). Python. Accessed: Oct. 15, 2024. [Online]. Available: <https://github.com/AakkashVijayakumar/stepwise-regression>