DIAML: Assignment 4

by Anthony Rizkallah

Table of contents

1 Python Packages

2 Linear Regression with on explanatory variable

2.1 Steps

To identify whether the performance of FTSE100, the index tracking the top 100 stocks in the British market, has any correlation to house prices, there is certain logic that has to be used. Finding correlation in the short term or long term requires the usage of returns or actual values, respectively. FTSE100 and Housing Prices are both supposed to increase with time due to the nature of the economy– inflation will increase hence purchasing power will decrease (or prices will always go up). So, using the returns of each will describe their inherent correlation, or its lack of, to verify whether the relationship is strong, moderate, or weak.

The equation that describes the rate of return is

$$r(t) = \frac{p(t)}{p(t-1)} - 1$$

where $r(t)$ is the rate of return at time t, $p(t)$ is the actual value at time t, and $p(t-1)$ is the actual value at time $t-1$.

Next, identifying the dependent and independent variables relies on making an assumption for the study. For this specific case, an assumption could be made that housing prices represent the general trend of investment capital flowing in and out of it, depending on whether the returns in the stock market are high or low. Another one is higher house appreciation enables investors to lend more money to invest in the market. Hence, FTSE100 is the dependent variable that is affected by the general state of the economy (Housing Prices).

Calculating the correlation coefficient between the two variables, FTSE100 and Average House Price is

$$\rho_{\text{FTSE,AHP}} = \frac{\sum (\text{FTSE100}_i - \overline{\text{FTSE100}})(\text{AHP}_i - \overline{\text{AHP}})}{\sqrt{\sum (\text{FTSE100}_i - \overline{\text{FTSE100}})^2 \sum (\text{AHP}_i - \overline{\text{AHP}})^2}}$$

where AHP is Average House Price Return, $\overline{\text{FTSE100}}$ and $\overline{\text{AHP}}$ is the mean value for each return, and FTSE_i and AHP_i are the individual values of returns.

Based on the previous analysis:

H_0 : There is no relationship between FTSE100 returns and Average House Prices.

H_1 : There is a relationship between FTSE100 returns and Average House Prices.

To verify the alternative hypothesis, then the p-value must be less than the tolerance threshold of $\alpha = 0.05$ to make sure that the relationship does not exist by chance, but rather an actual relationship.

Then, if the Null Hypothesis H_0 is rejected, the visualization of the linear regression graph would be relevant in predicting the returns of FTSE100 based on Housing Price Returns. The linear regression equation is

$$y_n = a + b x_n + \epsilon_n$$

where a is the intercept, b is the slope, and ϵ_n are the model errors or residuals.

2.2 Results

First, the correlation factor and p-value between Average House Returns and FTSE100 Returns is

Correlation Coefficient	P-Value
0.035	0.605

Table 1. Correlation Coefficient and P-Value of House Returns and FTSE100 Returns

The p-value was calculated based on a two-tailed test presuming that there was either a correlation (positive or negative) or no correlation. The correlation coefficient is very near to 0, meaning that there is no identifiable correlation among the two datasets.

Next, Table 2 contains the values after linearly regressing FTSE100 onto Average House Price, and Figure 1 shows the regression line on the scatter plot.

Slope	Intercept	Coefficient of Correlation	P-value	Standard Error
0.111	0.006	0.035	0.605	0.214

Table 2. Statistical values from regressing FTSE100 onto Average House Price

Figure 1. FTSE100 returns vs Housing returns

By observing the values in Table 2: First, the correlation coefficient is 0.035, indiciating a very weak positive relationship between Average House Price returns and FTSE100 returns. Second, the p-value associated with this relationship is 0.605, which is greater than the tolerance threshold of 0.05, indicating that the Null Hypothesis H_0 can not be rejected.

When observing the graph in Figure 1: First, the scatter plot shows that the FTSE100 returns with respect to Housing Returns don't have a specific trend, or one that is worthy of noting, as the dots don't follow a specific flow, but rather look random (with some exceptions). Second, the linear regression line is close to zero, confirming that the correlation between the two variables is week. A linear regression along $y = 0$ represents no statistically significant relationship.

Finally, H_0 holds– there is no relationship between FTSE100 returns and Average House Price returns, first because the p-value is higher than the tolerance $\alpha = 0.05$, and second, the regression line is hovering just above 0, also verifiying that there is no relationship between the two.

3 Linear regression with multiple explanatory variables

3.1 Steps

To build a model that can predict the graduation rate of universities, choosing certain variables to build the model is required. Based on the ones mentioned in the question (Applications Received, Enrolled Students, Number of Out of State Students, Admitted Top 10% in their classes, Admitted Top 25% in their classes), it is worth exploring their relevance as to why those might have been chosen.

Applications Received: Its relevance is that a high application volume could indicate high interest in the institution.

Enrolled Students: This metric would have been better off if it were a percentage of total applications received rather than just enrolled students.

Number of Out of State Students: This one is particularly interesting because students who come from out of state might not want to go back to their homes, so they might decide to proceed to graduation.

Admitted Top 10%/25% in their classes: This could be relevant because it reflects the academic competency of the students being admitted, and being in the top 10%/25% could mean that they won't have an issue graduating.

To find out whether those variables are significant, their correlation coefficients with the graduation rates must be calculated. Hence,

$$\rho_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where x_i is the value of each of the independent variables, $\bar{x}$ is the mean value of each independent variable, and y_i and $\bar{y}$ is for the dependent variable (Graduation Rate), respective to the former. At face value, if the coefficient of correlation is moderate to high, then there exists a relationship between it and the independent variable. Moreover, if the p-value of the variable is less than $\alpha = 0.05$, then this variable is accepted because it has statistical significance rejecting the Null Hypothesis H_0 (Variable x_1 has no statistical significance), and if it is higher than $\alpha = 0.15$, then the Alternative Hypothesis H_1 (Variable x_1 has statistical significance) is rejected.

However, to build a mutli-variable model that contains multiple explanatory variables, there are multiple methods to follow. For this case, stepwise regression and BIC were chosen as two methods to decide between different models.

Stepwise Regression Methodology:

1. Choose the variable (x_1) with the lowest p-value after calculating the correlation coefficient of all variables as a preliminary step (see previous section)

2. Regress the dependent variable on x_1 and x_2, x_1 and x_3 … x_1 and x_p. The variables in the model that have a p-value higher than the exit alpha $\alpha_{\text{exit}} = 0.15$ (chosen for this problem) are removed (Backward step).

3. Assuming that x_1 and x_2 remained, regress the independent variable onto the model containing both variables and repeat step 2 to evaluate the p-values.

4. Re-iterate the process until no addition nor removal changes the p-values.

Stepwise BIC Methodology:

1. Similar to stepwise regression, but rather choose the variable that has the lowest BIC of the model (x_1)

2. Regress the dependent variable on x_1 and x_2, x_1 and x_3 ... x_1 and x_p. The model that has the lowest BIC among the other models is chosen.

3. Assuming that x_1 and x_2 remained, regress the independent variable onto the model containing both variables and repeat step 2 to evaluate the BIC.

4. Re-iterate the process until no addition nor removal changes the BIC values.

The formula for calculating BIC is

$$BIC = N \ln\left(\frac{\text{RSS}}{N}\right) + k \ln N$$

where BIC is Bayesian Information Criteria, N is the number of observations (data points), k is the number of parameters, and RSS is the residual sum of squares.

Both processes at the end will keep the significant variables with respect to each model.

3.2 Results

3.2.1 Part a

Variables	**Correlation Coefficient**	**P-value**
Applications Received	0.146755	4.018556×10^{-5}
Enrolled Students	-0.022341	0.5340568
Admitted Top 10% of their classes	0.494989	2.897974×10^{-49}
Admitted Top 25% of their classes	0.477281	1.872333×10^{-45}
Number of Out of State Students	0.571290	1.628927×10^{-68}

Table 3. Initial Correlation Coefficients and P-values of Applications Received, Enrolled Students, Number of Out of State Students, Admitted Top 10% in their classes, Admitted Top 25% in their classes

3.2.2 Part b and c

Variables	**P-value**
Applications Received	0.0018626667929462599
Enrolled Students	0.003935533913809995
Admitted Top 25% of their classes	$2.3872299542646965\times10^{-12}$
Number of Out of State Students	$5.978384201445189\times10^{-30}$

Table 4. Stepwise Model Variable Summary

In the stepwise regression process, the method of adding and removing variables (Section 3.1 - Stepwise Regression Methodology) yielded those variables. Important to note is the variable of "Enrolled students" which initially had a p-value higher than $\alpha_{\text{entry}} = 0.05$ and $\alpha_{\text{exit}} = 0.15$, was later introduced to the model with a p-value lower than 0.05. This is what the stepwise regression model does– introducing variables at a later time with other variables might turn out to be overall significant.

3.2.3 Part d

Variables	**BIC**
Admitted Top 25%, Number of Out of State Students	6,274.3330

Table 5. BIC Model Variable and BIC Summary

In the stepwise BIC process, the method of adding and removing variables (Section 3.1 - Stepwise BIC Methodology) yielded those variables. Important to note that the variables that survived are the ones with the lowest p-values in the stepwise regression model.

3.2.4 Part e

Model	R^2	BIC	RMSE
Stepwise	0.385696	6277.675824	13.454801
BIC	0.377764	6274.332982	13.541383
All 5 Predictors	0.946047	6519.454820	15.719798

Table 6. Model Comparison

The three statistical values R^2, BIC, and $RMSE$ can be used to evaluate the most accurate model. First, higher R^2 values signify that higher variance in the data is explained, and in this aspect, the "All 5 Predictors" model is best, although it might be too high which means that it might be overfitted. On the BIC metric, the "BIC" model trivially is the best as it was chosen using the "BIC" model. On "RMSE", the Stepwise model is slightly better than that of "BIC".

Assuming that the "All 5 Predictors" model having a very high R^2 is not chosen due to overfitting, the second one in line is "Stepwise" in both R^2 and RMSE.

3.2.5 Part f

	Stepwise	BIC	All 5 Predictors
Actual: Carnegie Mellon University Graduation Rate	74%	74%	74%
Prediction: Carnegie Mellon University Graduation Rate	89.13%	87.09%	95.80%
Error	20.45%	17.60%	29.60%

Table 7. Carnegie Mellon Graduation Rate Predictions

It turns out that the most accurate model is the BIC and not the Stepwise. This could mean one of two things: Either CMU has certain variables that is specific to it that Stepwise could not have captured, slightly overfitting certain variables. On the other hand, the BIC model favors simplicity and generalization, which was more accurate in predicting the graduation rate at CMU. This does not mean that BIC will succeed at predicting other schools, but it rather suggests that the trade-off of simplicity works for CMU.

4 Open Study

4.1 Study Setup

The trend in the domain of transport to be studied is whether the increase in miles driven by vehicles, increases the sales of auto parts.

The datasets are provided on the St. Louis Federal Reserve with indicators:

- **Vehicle Miles Traveled**[1]**:** TRFVOLUSM227NFWA
- **Retail Sales: Motor Vehicle and Parts Dealers**[2]**:** MRTSSM441USN

H_0 : Increase in vehicle miles traveled does not increase the Motor Vehicle and Parts sales.

1. **https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA**
2. **https://fred.stlouisfed.org/series/MRTSSM441USN**

H_1 : Increase in vehicle miles traveled increases the the Motor Vehicle and Parts sales.

Moreover, I not only wanted to study the long-term trend, but also the short-term. To do so, I find the rate of return (or the percent increase/decrease) in each data point using

$$r(t) = \frac{p(t)}{p(t-1)} - 1$$

Next, I studied the correlation between both datasets actual values and the rate of returns of each value in the datasets. This gives insight into the general and the short-term trends. The correlation coefficient is:

$$\rho_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Then, to predict the auto-part sales in 2021, the regression model is trained on a dataset from 2014 to 2019, and then tested on the remaining dataset (2020 onwards). The reason I chose to train the model between 2014 and 2019 and not go back all the way to 1990 is because there were recessions in 1990, 2001, and 2008, and this does not represent the real trend up to 2019. Worthy of noting is the recession in 2020, and with the severity of this recession and the quick bounce-back, the model might fail to accurately predict the numbers, but it should predict the trend.

Using the linear regression formula of

$$y_n = a + bx_n + \epsilon_n$$

where a is the intercept, b is the slope, and ϵ_n are the model errors or residuals.

4.2 Study

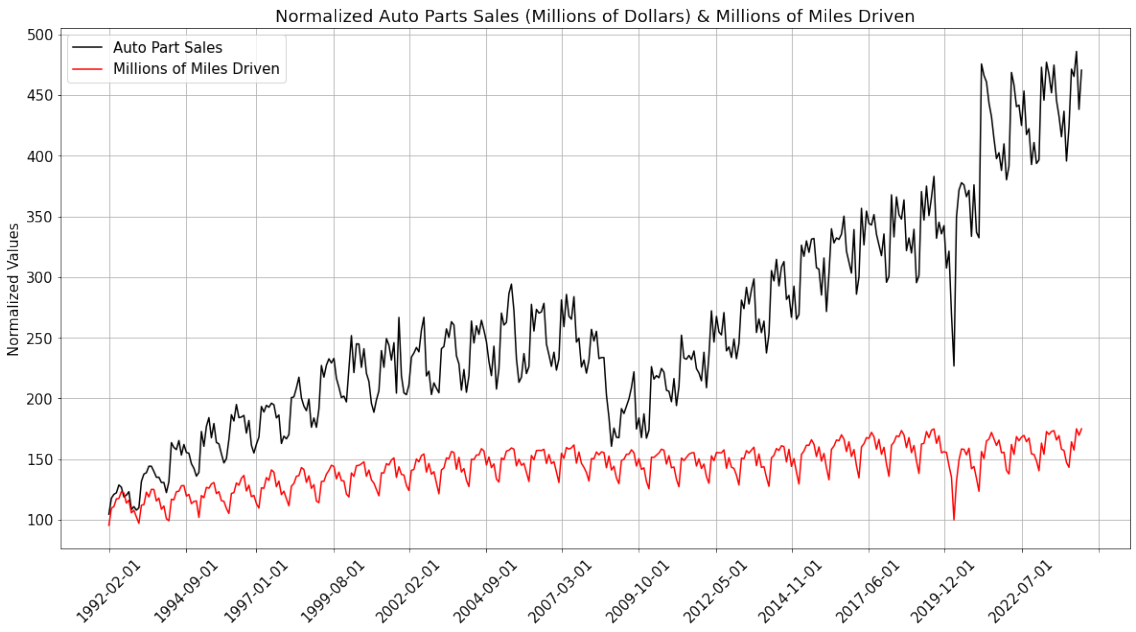

Figure 2. Normalized Auto Parts Sales (Millions of Dollars) & Millions of Miles Driven

By observation, the trend in Figure 2 suggests that the there is a trend that exists both seasonally (short-term) and generally. Both trends have increased, but the Auto Parts and Vehicles have increased much more than the Vehicle Miles Driven.

	Actual Values	% Returns
Correlation Coefficient	0.785	0.761
P-value	8.354×10^{-83}	6.341×10^{-75}

Table 8. Correlation Coefficient and P-value of Actual Values and % Returns of Auto Parts Sales and Vehicle Miles Driven

The following figures (3 and 4) show the scatter plots and the regression lines for actual values and % returns. The reason both trends have to be studied is because the nature of the economy is such that as it grows, more autoparts and vehicles are sold, and as the population grows, the miles driven increase. By ensuring that both the % change and the general trend are statistically significant, then the argument against choosing miles driven to describe sales in autoparts can be counter-argued by verifying that the sales follow the trend of vehicle miles driven.

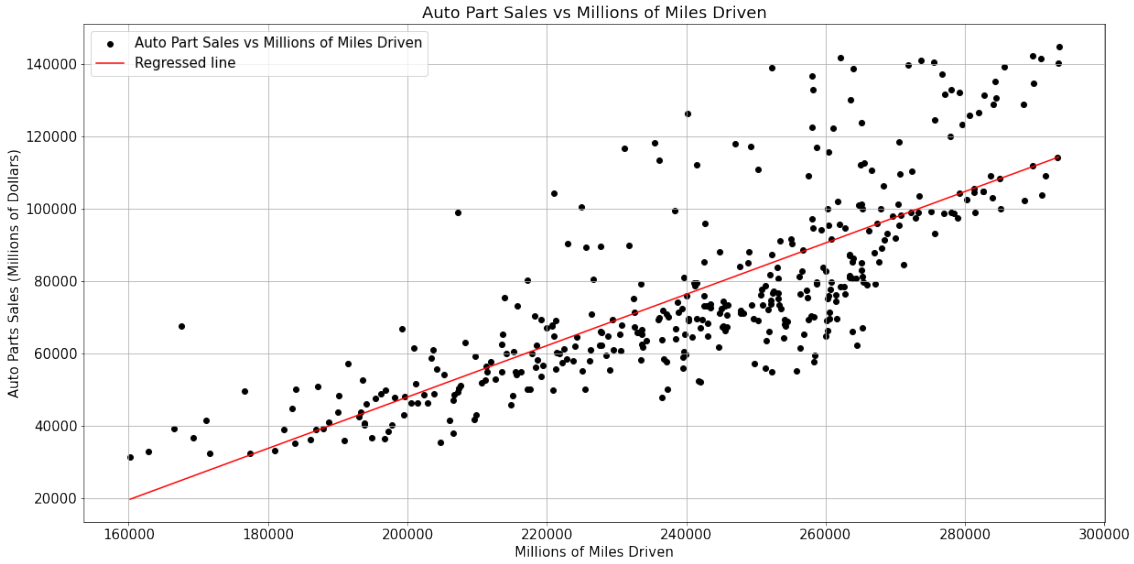

Figure 3. Auto Parts Sales vs Millions of Miles Driven

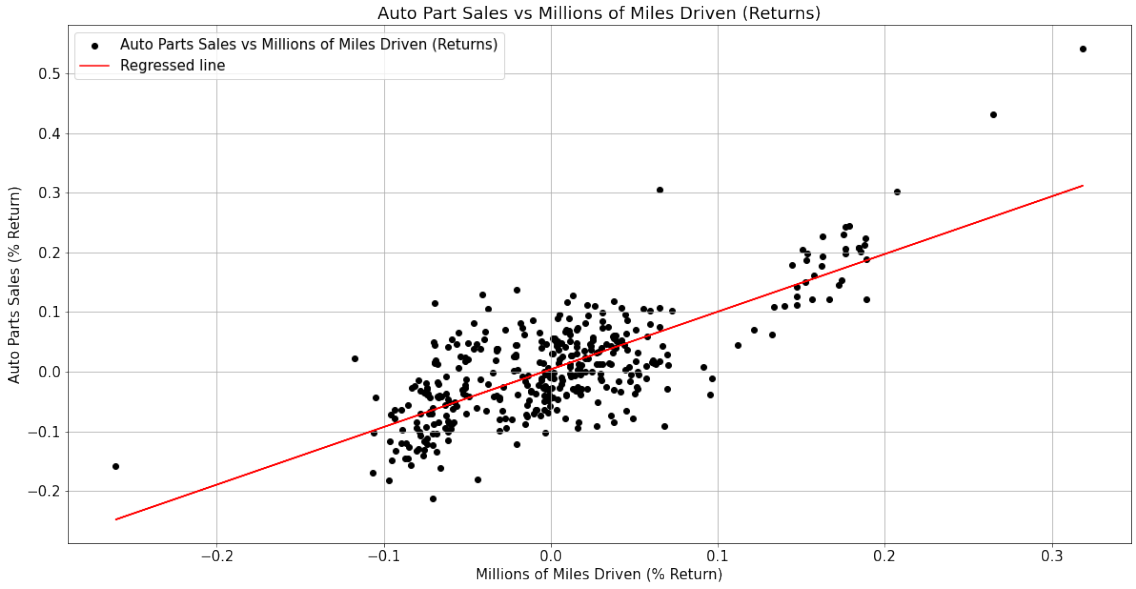

Figure 4. Auto Parts Sales vs Millions of Miles Driven (Returns)

Figure 3 shows that as the miles driven by vehicles increases, the sales of auto parts increases. Figure 4 shows that as the miles driven by vehicles fluctuates, the auto parts sales fluctuate. Both graphs show a strong correlation between the two variables and the p-values for both (Table 8) shows that there is a statistical significance for both, not just chance.

4.3 Prediction

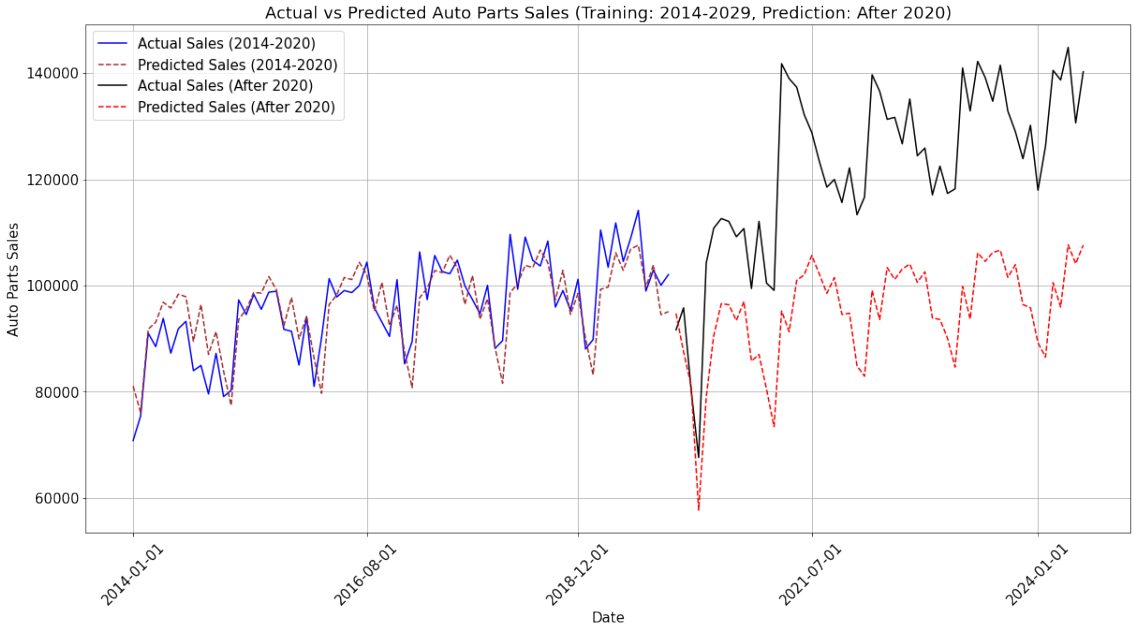

Figure 5. Actual vs Predicted Auto Part and Vehicles Sales with prediction after 2020

Date	**Actual Sales** (MM of Dollars)	**Predicted Sales** (MM of Dollars)	**% Error**
2021-01-01	100,460.0	80,472.744301	-19.895735
2021-02-01	99,091.0	73,406.912759	-25.919697
2021-03-01	141,782.0	95,203.051913	-32.852512
2021-04-01	138,960.0	91,293.998416	-34.301959
2021-05-01	137,361.0	100,988.625759	-26.479404
2021-06-01	132,167.0	102,008.862124	-22.818206
2021-07-01	128,836.0	105,674.567414	-17.977454
2021-08-01	123,405.0	102,162.493048	-17.213652
2021-09-01	118,527.0	98,536.088692	-16.866124
2021-10-01	119,978.0	101,480.085923	-15.417755
2021-11-01	115,627.0	94,555.181972	-18.223960
2021-12-01	122,160.0	94,791.384555	-22.403909

Table 9. Actual vs Predicted Auto Parts and Vehicles Sales

The model was capable of predicting the trend, but not the intensity of the sales. After the 2020 recession, it was unable to predict the bounce-back value. This means that the intensity of the recession and how long it lasts might be another predictor to add to the model to enhance the intensity of the values as the economy comes back from a recession.

5 Model Fitting and Prediction

5.1 Steps

To predict the unemployment rate in Israel in 2020, it follows the general methodology of first collecting the data, analyzing it, linearly regressing the dependent variable (Unemployment Rate)

onto Time (Ordinal) (Similar to previous sections).

Using the linear regression formula of

$$y_n = a + bx_n + \epsilon_n$$

where a is the intercept, b is the slope, and ϵ_n are the model errors or residuals.

Then using data from 1980 to 2013, calculating the MAPE would be the accuracy of the model. One could use RMSE or MSE, the only difference is the math. MAPE formula is

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where n is the number of data points, A_t is the actual value, and F_t is the forecast value.

5.2 Results

The unemployment rate from 1980 to 2024 is represented in Figure 6.

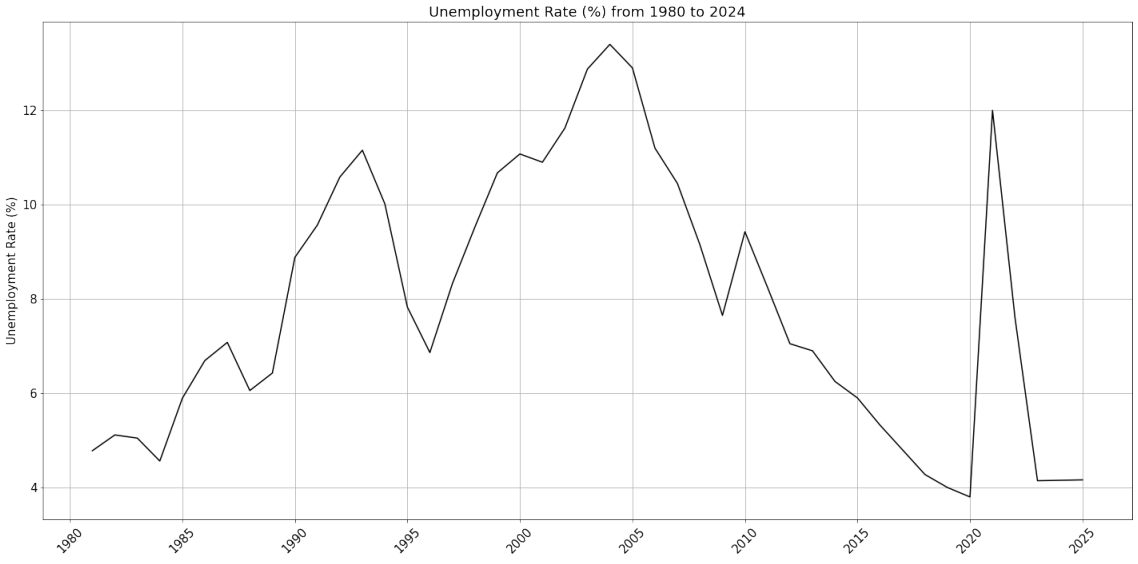

Figure 6. Unemployment Rate (%) from 1980 to 2024

Correlation Coefficient	-0.13429694800639072
Pvalue	0.3791118657294632

Table 10. Correlation Coefficient and P-value of Unemployment Rate over time

The values in Table 10 show that there is a weak correlation between unemployment rate and time (trivially if not studying values that are affected by time e.g. the moon's position in solar system). The pvalue is also greater than 0.05. For the premise of this question, there is nothing that can be done, but if it is being studied openly, there are other predictors that should be used. Moreover, since cyclicality was brought up, it is worth studying the autocorrelation of the unemployment rate to check for seasonality (Figure 7).

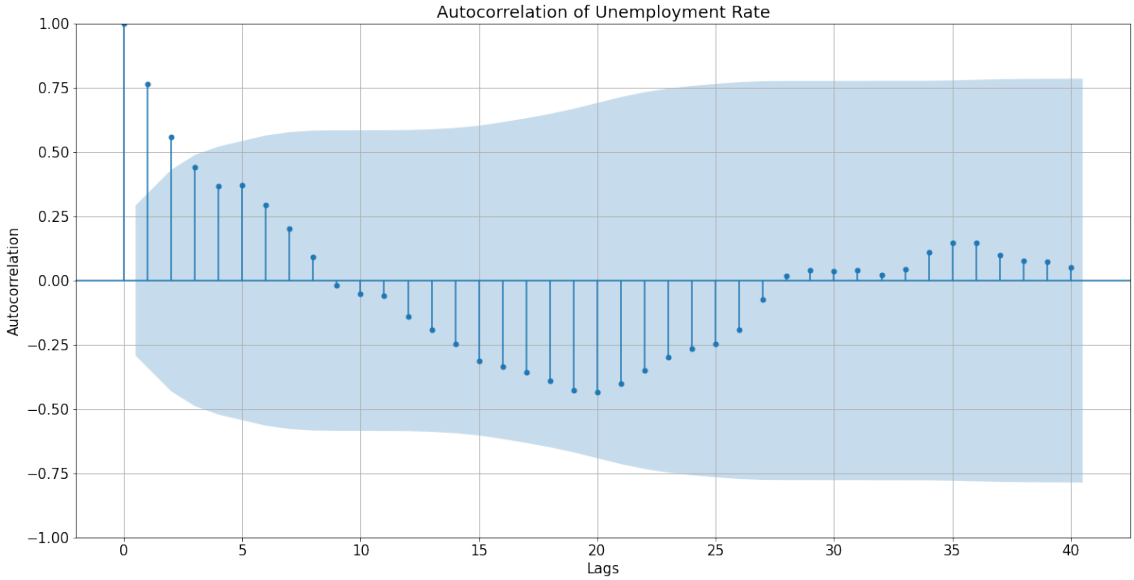

Figure 7. Autocorrelation of Unemployment Rate

Although the graph shows a strong seasonality at first, it tends to diminish at later lags which indicates no useful seasonality to be used for predicting values in 2020.

Using linear regression to predict the values in 2020, the graph in Figure 8 shows the linearly regressed line and the actual unemployment rate.

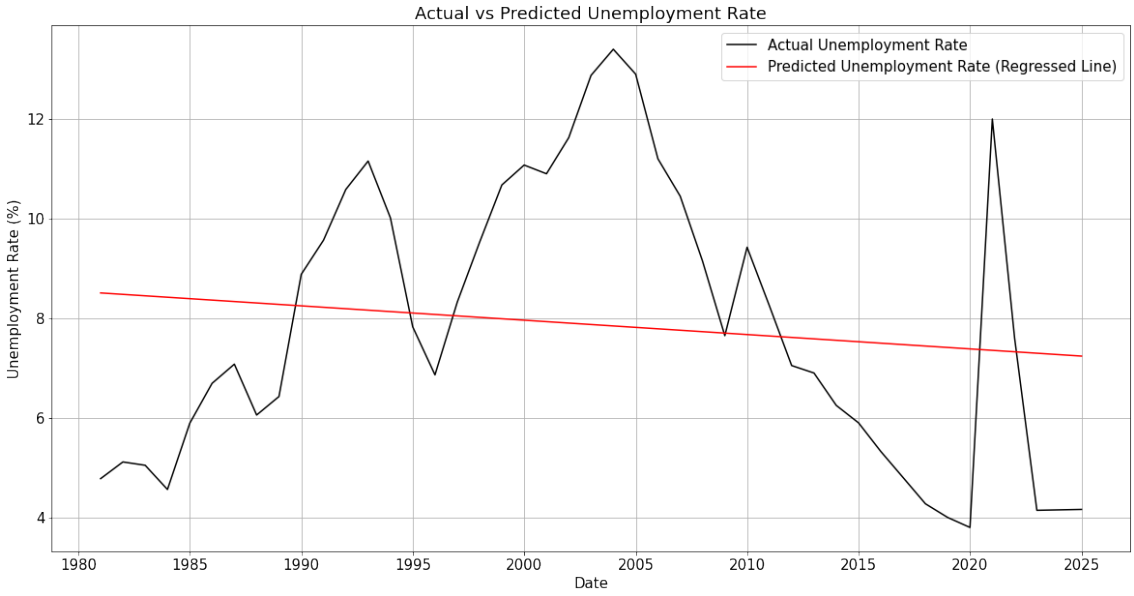

Figure 8. Actual vs Predicted Unemployment Rate

MAPE	Actual 2020 Unemployment Rate	Predicted 2020 Unemployment Rate
23.09%	12%	7.36%

Table 11. Model MAPE and Predicted Unemployment Rate for 2020

The MAPE for the unemployment rate of 2020 is not sufficient to accurately predict because $7.36\% \pm 23.09\% = \{9.06\%, 5.66\%\}$. Worthy of noting is that this prediction was made soley on time being the independent variable, which could be useful if the unemployment rate is seasonal.