Factors which affect unemployment rates amongst Boroughs in London

Abstract

The study aims to investigate and predict variables which effect unemployment rates within London boroughs. Machine learning techniques were used in the method and analysis of the study using the CRISP-DM process. The supervised technique used was Bivariate Linear Regression method whilst the unsupervised technique used was the K-Means Clustering method. Overall the use of machine learning helped answer the research question and found levels of correlation between certain independent variables and the dependent variable. Further research should be carried out to get an update and test with different variables.

I. Introduction

Unemployment is present all around the world and is one of the biggest macroeconomic problems that all countries want to solve [14]. Economic growth is said to be essential for a country's development in which one study concluded that a 1% rise in Gross Domestic Product (GDP) will drop the country's unemployment rate by 0.08%. This suggests the link between employment and economic success [15]. Unemployment rates differ across labour markets worldwide, which impacts communities socially as well as it does economically [14].

There are many reasons to why high rates of unemployment may be present in a particular country, city or region. The COVID-19 pandemic had a significant impact on unemployment rates across the World. This agrees with a study from Caballero and Valdés, who concluded that unemployment increased persistently after all previous pandemics in the world's history [16]. Other research speak about variables which unemployment may influence, such as an increased crime rate, higher government spending and decreased consumer spending, all having a negative effect economically.

London is divided into 32 Boroughs which make up the administrative area of Greater London. Each Borough is managed by a council responsible for many local services such as education and housing. The city is one of the most culturally mixed cities in Europe and is ranked in the top 10 for the highest percentage of foreign-born residents, revealing its diverse cultures and communities [17].

In 2020 the Labour Market statistics showed that London had the second highest unemployment rate in the UK behind the Northeast region [18]. Therefore, this journal will explore the London Boroughs which have the highest levels of unemployment and aim to predict the different attributes which correlate with unemployment. The data ranges between 2012 and 2020 including factors such as demographics, education attainment and job skills which will be looked at through machine learning to predict why certain boroughs may have a higher level of unemployment rates. The data has been gathered from Gov.UK which provides public datasets.

II. Literature Review

Employment can be investigated by focusing on many different target variables and attributes. Different analysis has been done on many different factors which have created valuable information used to understand reasons for unemployment and help create ways to confront it.

Much existing research has been done within the field of unemployment across the World. Trends such as unemployment and crime rates have been found through research analysis. One study found that unemployment has a direct lead to increased criminality in the UK and France [19]. The authors concluded that long-term unemployment drives criminality, which is in according with Cantor and Land's criminal motivational effect theory (1985). Other studies have investigated factors such as inflation and population size on the affect of unemployment, concluding that these are factors with a negative correlation on employment rates [20].

When considering the job market, researchers have examined whether there is a distinction between the need for high academic qualifications to obtain jobs in certain fields [21]. Education has been researched in many studies to have a significant affect on peoples job prospects. Yearly income has been specifically linked to education levels, showing a higher income for those completing higher education [22]. The Department of Education in the UK presented findings that men who attend higher education earn around 25% more than those who do not, and women who attend high education earn 50% more. Therefor it is important to consider the effects of education on unemployment rates as there is clear link between a similar aspect regarding salary.

Many studies on unemployment have focused their data in a specific country or city. Few research have focused on differences in unemployment rates across areas within a city. Previous research has been carried out stating that there are employment gaps in London Boroughs [23], therefor this project aims to determine which factors affect employment rates in the different London Boroughs with the view on tackling key challenges which certain communities may face and provide an evaluation in helping with this problem.

III. Methodology

This project journal will be produced following the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. CRISP-DM is an industry-independent process model for data mining. The process is broken down into six phases; business understanding, data understanding, data preparation, modelling, evaluation and deployment [2]. The aim is to implement and achieve each of these stages of CRISP-DM within the journal so that all the objectives are achieved.

During the business understanding phase, the aim is to determine the objectives of the research by carrying out a situational assessment and determining the goals [1]. This action should assess the situation to get an overview of the available and required resources such as the determination of the data mining goal [3]. The main objective of this research is to find the Boroughs with the highest unemployment rates, and then measure unemployment with the independent variables in the data to see which correlate the most. Supervised and unsupervised machine learning methods in data mining will be implemented through Python on Google Colab.

The data understanding phase is the process of grasping what the data is about by collecting, describing and verifying the quality of the data. The description of the data can be produced by using statistical analysis and determining attributes [3]. The UK Labour Market dataset has been taken from a verified website which collects public datasets in the UK and allows access to the general public [4]. This dataset fits the target population and contains the variables needed to successfully carry out this piece of research.

The third phase is data preparation. It is essential to make sure the data is fully explored, cleansed and transformed before carrying out any analysis. The selection of the data should be carried out by defining inclusion and exclusion criteria [2]. The UK Labour Market dataset will be checked for anomalies, null values, errors and inconsistencies to ensure it is at its most suitable condition, allowing the next stages to be smooth and reliable.

The modelling stage involves selecting and building the modelling techniques. The extraction and analysis of the data can take part in this phase in which data mining techniques can be used depending on the business objective and the data [3]. The Bivariate Linear Regression supervised model and K-Means Clustering unsupervised model processes will be used to extract data so that the most accurate results can be obtained. A Bivariate Regression technique in machine learning is used to understand the relationship between a single dependent variable and multiple independent variables. This technique is considered one of the most common and comprehensive learning algorithms [5]. The coding produces heat-maps to look for correlations between the variables in which the strongest will be visualised using scatter plots [6]. K-means clustering involves the process of splitting data points and partitioning them into groups based on similarities. This is an unsupervised learning model as it does not require a training dataset to learn the model parameters.

The Bivariate Linear Regression machine learning model is the most suitable technique to use for this dataset due to there being multiple explanatory variables which are to be tested against unemployment rates. The outcome of this analysis will be able to present if there are any variables within the dataset which correlate to unemployment rates in London boroughs and therefore predict changes in unemployment rates.

The evaluation stage is carried out to assess the results against the defined business objectives. The process in its entirety should be evaluated and interpreted so that further actions can be defined [1]. Overall, this stage compares the degree to which the model meets the criteria defined at the start of the project. It can be defined through the equation RESULTS = MODELS + FINDINGS, meaning the total output of the data mining project is not solely the model but the factors it contributes to[3]. The results should be evaluated and reviewed to see if the model shows a satisfactory answer to the business question. This section will be carried out in the discussion and conclusion part of the journal.

The final stage is the deployment phase where solutions will be created and established based on the results to make changes based on the objective of the research. This will be done outside of the research journal itself where the results will influence what is to be done next [1]. As well as creating solutions, ethical situations should also be addressed during this phase along with privacy, consent, and the responsible handling of sensitive information [3]. A summary of the project and its results should be produced, including a reinforced reason for the chosen methods during the machine learning process. When considering this project it can be said that there were no privacy issues as no personal details were present in the raw data. No consent was needed as the data was collected from a public data source allowing free access. The chosen method for analysis was appropriate for addressing the research question due to the multiple variables in the dataset and the objective of finding correlations.

IV. Results and Discussion

The Bivariate Linear Regression supervised model and K-Means Clustering unsupervised model processes are suitable methods to use due to their ability to analyse multiple variables and find relationships between them. The theory behind a linear relationship is that for every one-unit change in the independent variable, there will be a consistent and uniform change in the dependent variable [7]. The difference between the dependent and independent variable is called the residual. This can help with predicting the outcome of future events and is used for predictive analysis [11].

The k-means algorithm uses a clustering method to recognise patterns in the data by finding similarities and dissimilarities between data points [9]. This algorithm is considered unsupervised as it does not require a training dataset to learn the model parameters and all variables are considered independent. For each cluster produced in the analysis there is a representative point at the mean and can be denoted as a centroid. Although there are many advantages to this method of machine learning, disadvantages may include the difficulty in choosing the best distance measure and the resulting clustering being sometimes hard to interpret [10].

Key variables were included to analyse based on whether they link to unemployment rates. The original dataset contained 15 columns in which seven were dropped as they were not considered necessary to answering the research question.

The raw data sourced from a public website was converted and imported to Google Colab as a CSV file. All string values were changed to numerical when necessary and then normalised due to each variable having different measurements and scales. Without normalisation, certain variables will outweigh others, creating biases in the data [8]. No errors, null values or missing values were present in the data when checked in Google Colab. Each London borough was given an ID number in a new separate column to allow each borough to be identified through a numerical value. These cleaning steps were important to ensure the data was as reliable and as least bias as possible.

Figure 1 shows which London boroughs have the highest unemployment rates in 2020. It can be seen that the boroughs with the highest unemployment rates are Brent and Hackney, with Figure 2 showing that the figures are significantly higher than the rest of the Boroughs through as box plot.

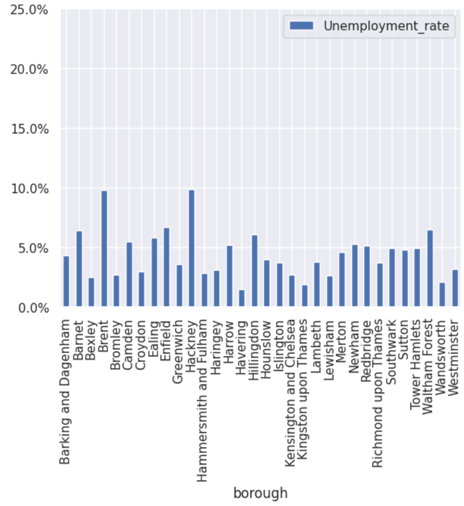


Figure 1: Unemployment Rates in London Boroughs

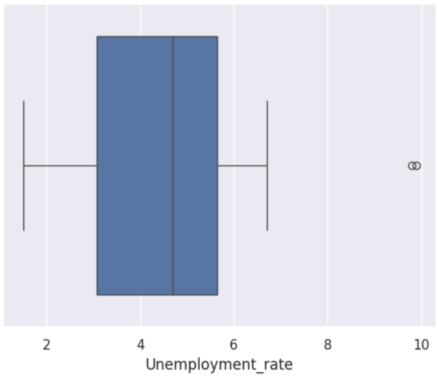


Figure 2: Box Plot of Unemployment Rates in London Boroughs

A correlation matrix figure was then produced to look at the relationship strength between variables. It is generally seen that a correlation result of 0.3 or higher with the dependent variable is considered high enough for predicting the outcome. The p-value is also a fundamental in testing if there is a significant difference between variables. If the p-value is smaller than 0.05 ($p < 0.05$) then the results are statistically different [12].

Figure 3 shows that the variables which correlate the most with unemployment rates are 5_A*_C_grades, year and black ethnicity. 5_A*_C_grades represents the percentage of students passing at least five GCSE's, black ethnicity represents the amount of black people living in the borough and the year is the unemployment rate percentage in a certain year.

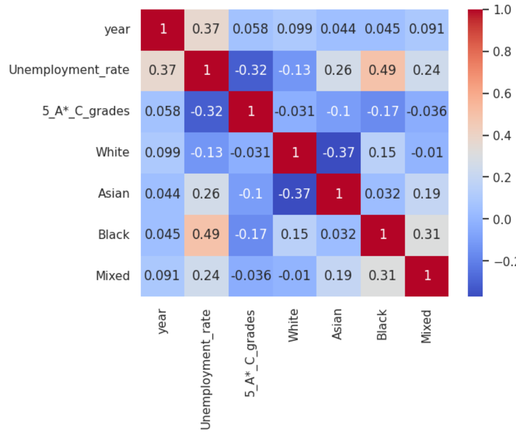


Figure 3: Correlation Metrics between variables

A pair plot was then made in to get a visual representation of the relationship between the independent and dependent variable. Figure 4 shows three scatter plots for the variables with the highest relationship with unemployment rates. It can be seen that there is a weak negative correlation between unemployment rates and the variable 5_A*_C_grades. There is a weak positive correlation with the black ethnicity variable. It can be seen in Figure 4 unemployment rates generally increased from 2012 onwards.

Figure 4: Pair Plot of unemployment rates against the three variables with the highest correlation

Scatter plots were then made including a line of best fit by minimising the Residual Sum of Squares. The Mean Squared Error is also used which is the average of squared error between the predicted and actual value. The train size was set to 0.3 and the test set to 0.7. A constant term was added to the independent variable. The Root Mean Square Error (RMSE) and R-squared value were checked to evaluate the performance and goodness-of-fit of the model. The RSME measures the average deviation of the predicted values from the actual values. A value of zero would indicate the model perfectly predict all data points. The R-squared value ranges from 0 to 1 where 1 indicates that the model explains all the variability of the response data around its mean [13].

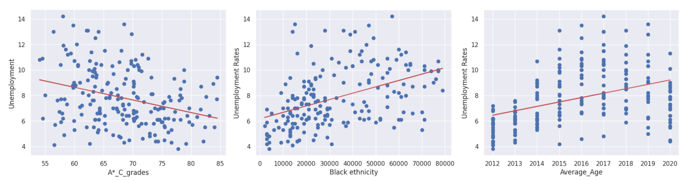


Figure 5: Scatter Plot of unemployment rates fitted against GCSE results, black ethnic population, and year including line of best fit and predicted values.

The results from the machine learning analysis show that boroughs with a higher level than others in London may have a higher amount of black ethnic residents. This suggests that when it comes to considering ethnic backgrounds, residents of black ethnicity are more likely to be unemployed compared to other groups in the London boroughs. GCSE results showed a weak correlation with unemployment rates therefore it cannot be seen as direct causation but still considered to maybe influencing it. Further research is needed to confirm if eduction is directly linked with unemployment in London.

V. Conclusion and Recommendation

A few key insights were discovered with the use of machine learning on finding reasons for unemployment rates in London boroughs. Based on the results it can be suggested that unemployment is more permanent when GCSE results are worse and that adults of black ethnicity may be more likely to be unemployed than others. The results cannot be seen as a perfect prediction of cause due to the correlation values not being significant.

It must also be considered that the dataset included data points between 2012 and 2020, and a more recent dataset may produce results and predictions more accurate to the current period. Future research should aim to look at unemployment rates from 2020 onwards and also introduce more dependent variables such as crime rates, population density and others. This could aid find different variables which may influence unemployment rates. The same variables could be used again with different machine learning techniques to see if similar results are discovered.

Based on this studies results, schools with the lowest GCSE results could look to providing different ways to prepare studies for work and employment even when results are low. Schemes could be put into place by local councils to aim to tackle the potential problem with ethic minorities getting work.

The peer feedback provided for this study helped change the certain areas which needed improvement. The insights I gained from the feedback was to consider changing the method to allow the results to be as accurate as possible and to elaborate more within the methodology section. This enabled the development of a more thought out plan, better preparation and a better look into previous research to greatly improve the quality of the journal.

VI. Link To The Google Colab

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sb
import matplotlib.ticker as mtick
```

```
unemployment = pd.read_csv("/content/ML Data possibility 2 - London_Borough_dataset_1.6..csv")
```

```
### DATA UNDERSTANDING ###
```

```
unemployment.tail()
```

	borough	year	Unemployment_rate	5_A*_C_grades	id	White	Asian	Black	Mixed
283	Westminster	2016	7.4	75.2	32	136000	31000	18000	54000
284	Westminster	2017	7.8	78.4	32	131000	34000	19000	50000
285	Westminster	2018	8.2	65.0	32	140000	33000	17000	51000
286	Westminster	2019	7.7	58.3	32	140000	31000	15000	53000
287	Westminster	2020	7.7	57.7	32	159000	18000	19000	44000

```
unemployment.describe()
```

	year	Unemployment_rate	5_A*_C_grades	id	White	Asian	Black	Mixed
count	288.000000	288.000000	288.000000	288.000000	288.000000	288.000000	288.000000	288.000000
mean	2016.000000	7.851389	67.986458	16.500000	160989.583333	49687.500000	32788.194444	26878.472222
std	2.586483	2.272069	6.973478	9.249164	43898.956229	39911.132772	20370.614884	12582.376166
min	2012.000000	3.700000	51.900000	1.000000	76000.000000	7000.000000	2000.000000	5000.000000
25%	2014.000000	6.100000	62.675000	8.750000	126000.000000	20000.000000	16750.000000	18000.000000
50%	2016.000000	7.550000	68.000000	16.500000	156000.000000	32000.000000	26000.000000	24000.000000
75%	2018.000000	9.400000	73.125000	24.250000	192250.000000	74000.000000	49000.000000	35000.000000
max	2020.000000	14.200000	84.800000	32.000000	272000.000000	166000.000000	81000.000000	61000.000000

```
print(unemployment.columns)
```

```
Index(['borough', 'year', 'Unemployment_rate', '5_A*_C_grades', 'id', 'White',
       'Asian', 'Black', 'Mixed'],
      dtype='object')
```

```
unemployment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 288 entries, 0 to 287
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   borough            288 non-null    object
 1   year               288 non-null    int64
 2   Unemployment_rate  288 non-null    float64
 3   5_A*_C_grades      288 non-null    float64
 4   id                 288 non-null    int64
 5   White              288 non-null    int64
 6   Asian              288 non-null    int64
 7   Black              288 non-null    int64
 8   Mixed              288 non-null    int64
dtypes: float64(2), int64(6), object(1)
memory usage: 20.4+ KB
```

```
unemployment.duplicated().sum()
```

```
0
```

```
unemployment.isnull().sum()
```

```
borough              0
year                 0
Unemployment_rate    0
5_A*_C_grades        0
id                   0
White                0
Asian                0
Black                0
Mixed                0
dtype: int64
```

```
### DATA CLEANING ###
```

```
unemployment_rates = unemployment.drop(columns=['borough'])
```

```
unemployment_rates.head()
```

	year	Unemployment_rate	5_A*_C_grades	id	White	Asian	Black	Mixed
0	2012	6.8	64.3	1	106000	27000	44000	12000
1	2013	9.7	73.2	1	106000	32000	43000	12000
2	2014	10.7	60.0	1	99000	33000	44000	22000
3	2015	13.6	62.1	1	95000	41000	52000	14000
4	2016	13.5	59.8	1	106000	41000	47000	13000

Next steps: View recommended plots

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled = scaler.fit_transform(unemployment_rates)
```

```
print(data_scaled.mean(axis=0))
print(data_scaled.std(axis=0))
```

```
[ 0.00000000e+00  4.50257116e-16  2.71387850e-16  0.00000000e+00
 -2.46716228e-16  2.46716228e-17 -4.93432455e-17 -3.70074342e-17]
[1. 1. 1. 1. 1. 1. 1. 1.]
```

```
unemployment_rates_final = unemployment_rates.drop(columns=['Population', 'Two-year_business_survival_rates', 'Fires_per_thousand', 'Gross_Annual_Pay', 'Median_House_Price',
```

```
unemployment_rates_final = unemployment_rates.drop(columns=['id'])
```

```
### DATA ANALYSIS ###
```

```
unemployment.plot.bar (x='borough', y='Unemployment_rate', rot=90)
plt.ylim(0, 20)
plt.gca().yaxis.set_major_formatter(mtick.PercentFormatter())
plt.show()
```

```
sb.set_theme()
sb.boxplot(x=unemployment['Unemployment_rate'])
```

```
sb.heatmap(unemployment_rates_final.corr(), cmap='coolwarm', annot = True)
```

```
sb.pairplot(unemployment_rates_final, x_vars=['Black', '5_A*_C_grades', 'year'], y_vars='Unemployment_rate', height=6, aspect=1, kind='scatter')
```

```
from sklearn.model_selection import train_test_split
```

```
x = unemployment_rates_final[ '5_A*_C_grades']
y = unemployment_rates_final[ 'Unemployment_rate']
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.7, test_size = 0.3, random_state = 100)
```

```
print( x_train.shape)
print( x_test.shape)
print( y_train.shape)
print( y_test.shape)
```

```
(201,)
(87,)
(201,)
(87,)
```

```
import statsmodels.api as sm
```

```
x_train_sm = sm.add_constant(x_train)
```

```
lr = sm.OLS(y_train, x_train_sm).fit()
```

```
lr.params
```

```
const            14.571015
5_A*_C_grades    -0.098996
dtype: float64
```

```
print(lr.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Unemployment_rate   R-squared:                       0.098
Model:                            OLS   Adj. R-squared:                  0.094
Method:                 Least Squares   F-statistic:                     21.73
Date:                Sun, 28 Apr 2024   Prob (F-statistic):           5.76e-06
Time:                        13:24:58   Log-Likelihood:                -438.42
No. Observations:                 201   AIC:                             880.8
Df Residuals:                     199   BIC:                             887.5
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            14.5710      1.459      9.985      0.000      11.693      17.449
5_A*_C_grades    -0.0990      0.021     -4.661      0.000      -0.141      -0.057
==============================================================================
Omnibus:                        6.725   Durbin-Watson:                   2.269
Prob(Omnibus):                  0.035   Jarque-Bera (JB):                6.419
Skew:                           0.386   Prob(JB):                       0.0404
Kurtosis:                       2.586   Cond. No.                         660.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
x_test_sm = sm.add_constant(x_test)
```

```
y_pred = lr.predict(x_test_sm)
```

```
from sklearn.metrics import mean_squared_error
```

```
from sklearn.metrics import r2_score
```

```
print('RSME:', np.sqrt(mean_squared_error(y_test, y_pred)))
print('r-squared:', r2_score(y_test, y_pred))
```

```
RSME: 2.1721802932726764
r-squared: 0.10028866843912854
```

```
plt.scatter(x_train, y_train)
plt.plot(x_train, 14.571015 + -0.098996*x_train, 'r')
plt.xlabel('A*_C_grades'), plt.ylabel( 'Unemployment')
plt.show()
```

```
# Code for new independent variable
```

```
x1 = unemployment_rates_final[ 'Black']
y1 = unemployment_rates_final[ 'Unemployment_rate']
```

```
x1_train, x1_test, y1_train, y1_test = train_test_split(x1, y1, train_size=0.7, test_size = 0.3, random_state = 100)
```

```
print( x1_train.shape)
print( x1_test.shape)
print( y1_train.shape)
print( y1_test.shape)
```

```
(201,)
(87,)
(201,)
(87,)
```

```
x1_train_sm = sm.add_constant(x1_train)
```

```
lr1 = sm.OLS(y1_train, x1_train_sm).fit()
```

```
lr1.params
```

```
const    6.175948
Black    0.000050
dtype: float64
```

```
[ ] print(lr1.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:      Unemployment_rate   R-squared:                       0.200
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     49.80
Date:                Sun, 28 Apr 2024   Prob (F-statistic):           2.77e-11
Time:                        13:26:20   Log-Likelihood:                -426.39
No. Observations:                 201   AIC:                             856.8
Df Residuals:                     199   BIC:                             863.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.1759      0.272     22.724      0.000       5.640       6.712
Black       5.009e-05    7.1e-06      7.057      0.000    3.61e-05    6.41e-05
==============================================================================
Omnibus:                       17.455   Durbin-Watson:                   2.114
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.321
Skew:                           0.709   Prob(JB):                     6.37e-05
Kurtosis:                       3.543   Cond. No.                     7.27e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.27e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
[ ] x1_test_sm = sm.add_constant(x1_test)
```

```
[ ] y1_pred = lr1.predict(x1_test_sm)
```

```
[ ] print('RSME:', np.sqrt(mean_squared_error(y1_test, y1_pred)))
    print('r-squared:', r2_score(y1_test, y1_pred))

    RSME: 1.870005562486234
    r-squared: 0.3331974554612953
```

```
plt.scatter(x1_train, y1_train)
plt.plot(x1_train, 6.175948 + 0.000050*x1_train, 'r')
plt.xlabel('Black ethnicity'), plt.ylabel( 'Unemployment Rates')
plt.show()
```

```
# Code for new independent variable no.3
```

```
[ ] x2 = unemployment_rates_final[ 'year']
    y2 = unemployment_rates_final[ 'Unemployment_rate']
```

```
[ ] x2_train, x2_test, y2_train, y2_test = train_test_split(x2, y2, train_size=0.7, test_size = 0.3, random_state = 100)
```

```
[ ] print( x2_train.shape)
    print( x2_test.shape)
    print( y2_train.shape)
    print( y2_test.shape)

    (25,)
    (7,)
    (25,)
    (7,)
```

```
[ ] x2_train_sm = sm.add_constant(x2_train)
```

```
[ ] lr2 = sm.OLS(y2_train, x2_train_sm).fit()
```

```
[ ] lr2.params

    const   -689.092957
    year       0.345698
    dtype: float64
```

```
print(lr2.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:      Unemployment_rate   R-squared:                       0.156
Model:                            OLS   Adj. R-squared:                  0.152
Method:                 Least Squares   F-statistic:                     36.78
Date:                Sun, 28 Apr 2024   Prob (F-statistic):           6.53e-09
Time:                        13:27:57   Log-Likelihood:                -431.79
No. Observations:                 201   AIC:                             867.6
Df Residuals:                     199   BIC:                             874.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -689.0930    114.906     -5.997      0.000    -915.682    -462.504
year           0.3457      0.057      6.065      0.000       0.233       0.458
==============================================================================
Omnibus:                        9.374   Durbin-Watson:                   2.214
Prob(Omnibus):                  0.009   Jarque-Bera (JB):                9.621
Skew:                           0.534   Prob(JB):                      0.00814
Kurtosis:                       3.086   Cond. No.                     1.58e+06
==============================================================================
```

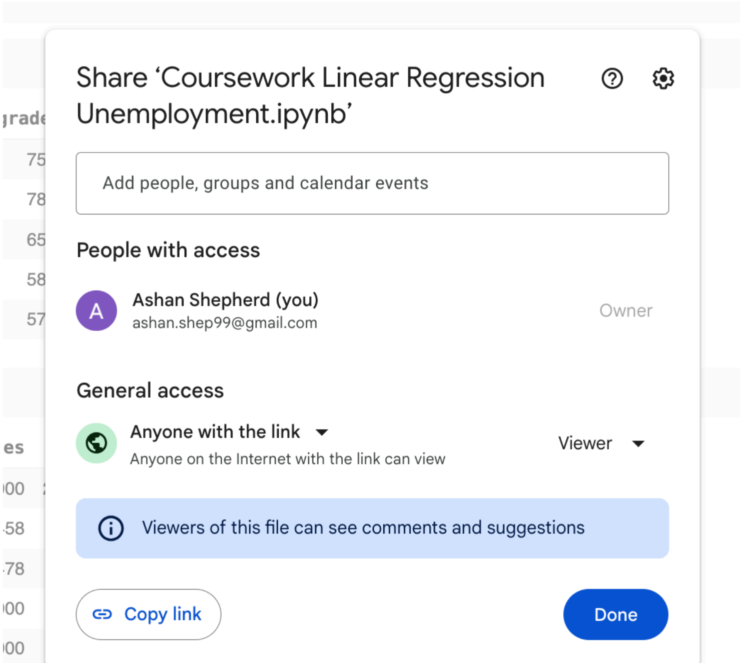
```
[ ] x2_test_sm = sm.add_constant(x2_test)
```

```
[ ] y2_pred = lr2.predict(x2_test_sm)
```

```
[ ] print('RSME:', np.sqrt(mean_squared_error(y2_test, y2_pred)))
    print('r-squared:', r2_score(y2_test, y2_pred))

    RSME: 2.1770604647514635
    r-squared: 0.09624141888954496
```

```
plt.scatter(x2_train, y2_train)
plt.plot(x2_train, -689.092957 + 0.345698*x2_train, 'r')
plt.xlabel('Average_Age'), plt.ylabel( 'Unemployment Rates')
plt.show()
```



https://colab.research.google.com/drive/1zycO7VkEq0Tj0Mv_KmjIzRN3QqcJ38Wk?usp=sharing

VII. References

[1] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," Procedia *Computer Science, https://www.sciencedirect.com/science/article/pii/S1877050921002416* (accessed Mar. 1, 2024).

[2] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining." *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (4), pp. 29–39. 2000

[3] Chapman, Pete, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R Wirth. "CRISP-DM 1.0. Step-by-step data mining guide." 2000

[4] "London labour market, skills and employment indicators," London Datastore News, https://data.london.gov.uk/dataset/london-labour-market-indicators (accessed Mar. 1, 2024).

[5] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", *JASTT*, vol. 1, no. 2, pp. 140-147, 2020.

[6] T. M. H. Hope, "Chapter 4 - Linear regression, Machine Learning" *Academic Press*, 2020

[7] Chapter 7 Modeling Relationships of Multiple Variables with Linear Regression

[8] A. Peshawa, J. Muhammad, R. Hassan, Faraj, E. Koya, Peshawa J. M. Ali, and R. H. Faraj. "Data normalization and standardization: a technical report." *Mach Learn Tech Rep 1*, 1, 1-6, 2014.

[9] K. P. Sinaga and M. S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE*, (8), pp. 80716-80727, 2020

[10] T. M. Ghazal, "Performances of k-means clustering algorithm with different distance metrics." *Intelligent Automation & Soft Computing,* 30(2), 735-742, 2021

[11] A. A. Suleiman, U. A. Abdullahi and U. A. Ahmad, "An Analysis of Residuals in Multiple Regressions." *International Journal of Advanced Technology in Engineering and Science, 3*(1), pp.563-570. 2015

[12] M. Zaki and W. Meira, "Data Mining and Machine Learning (2nd ed.)." *Cambridge University Press*, 2020

[13] D. C. Montgomery, A. E. Peck, and G. G. Vining. "Introduction to Linear Regression Analysis. 5th ed" *Hoboken, NJ, USA: Wiley,* 2012.

[14] A. Yunusova, H. Budak, A. Daval, N. Degerli, D. Balkaya, K. Uzboyali, E. Turgut, R. Aykoc, E. D. Aktekin, Z. Yildirim and G. Coskun, "Discussions Between Economic Agents: Panel Data Analysis" *IKSAD Publishing House*. 2021

[15] Ã. B. Soylu, Ä. Ãakmak, and F. Okur, "Economic growth and unemployment issue: Panel data analysis in Eastern European countries," Journal of International Studies, 11(1), pp. 93-107, 2018. doi:10.14254/2071-8330.2018/11-1/7

[16] C. V. Rodríguez-Caballero and J. E. Vera-Valdés, "Long-lasting economic effects of pandemics: evidence on growth and unemployment," *Econometrics,* 8(3), p. 37, 2020. doi:10.3390/econometrics8030037

[17] J. Eade, "Cultural and Ethnic Diversity in Cities: Challenges and Chances" *Networking European Citizenship Education,* 2010

[18] A. Powell, " Labour market statistics: UK regions and countries" *House of Commons Library*, 7950, 2021

[19] F. Jawadi, S. K. Mallick, A. Idi Cheffou, and A. Augustine, "Does higher unemployment lead to greater criminality? revisiting the debate over the business cycle" *Journal of Economic Behavior & amp; Organization*, vol. 182, pp. 448-471, 2021. doi:10.1016j.jebo.2019.03.025

[20] J. Alam, Q. N. Alam and T. Hoque, "Impact of GDP, inflation, population growth and FDI on unemployment: A study on Bangladesh economy," *African Journal of Economics and Sustainable Development,* 3(3), pp. 67–79, 2020. doi:10.52589/ajesd/cah2iyqj

[21] H. Lauder, and K. Mayhew, "Higher education and the labour market: an introduction." *Oxford Review of Education,* 46(1), 1–9. 2020

[22] C. Belfield *et al.,* "The impact of undergraduate degrees on early-career earnings," *Institution for Fiscal Studies,* 2018. doi:10.1920/re.ifs.2019.0808

[23] D. A. Smith *et al.*, "A Compact City for the wealthy? employment accessibility inequalities between occupational classes in the london metropolitan region 2011," *Journal of Transport Geography,* vol. 86, p. 102767, 2020. doi:10.1016/j.jtrangeo.2020.102767