# Understanding Unemployment Globally

Group 29

Riddhi Barbhaiya, Wun-Syuan Wu, Raghuram Palaniappan, Jose Luna

# 1 Introduction

The unemployment rate can can be indicative of the condition of a country's economy. Employment is an important aspect of the economy because it impacts the purchasing power of an household and overall consumption of goods. This impacts demand for goods and services triggering complicated effects in an economy[1]. The employment rate is also indicative of challenging economic times and over all satisfaction of a population. We are interested in understanding the relationship between the unemployment and other macroeconomic variables in countries using the WorldBank data.

To be more precise, unemployment rate is the percent of the label force that is unemployed. The labor force is the sum of employed and unemployed people, excluding people unable to work or those not searching for work.

# 2 Data and Methodology

## 2.1 Data

**Data Source:** World Bank data for the year 2012 for 73 countries

As mentioned in the introduction, in addition to unemployement, we are interest in a set of macroeconomic variables. The following table summarises the variables we consider.

| Notation | Abreviation | Description |
|---|---|---|
| $Y_i$ | Unemployment | Percent of the labor force that is unemployed |
| $x_1.$ | GovExp | Government GDP expenditures |
| $x_2.$ | Education | Percent of the labor force with basic education |
| $x_3.$ | GDP | GDP per capita |
| $x_4.$ | PopGrowth | Population growth rate |
| $x_5.$ | Agri | GDP added by agriculture |
| $x_6.$ | Inflation | Inflation in consumer prices |
| $x_7.$ | EmployAgri | Employment in agriculture |
| $x_8.$ | EmployInd | Employment in industry |
| $x_9.$ | EmployServ | Employment in service |

## 2.2 Methodology Overview

We start by trying to gain a better understanding of the how the variable relate to unemployment and each other through exploratory visualizations. We consider the summary statistics of the variables and the correlations they may hold. We also plot variables by their geographic location. In this, we utilize many of the methods explored in the course.

In analyzing the relationship between unemployment rate and the variables of interest we take a number of approaches. We start by splitting the data into a training and testing set. We use the training set to fit our models and use the testing set to compare the efficacy of the models and their accuracy.

In terms of modeling, we first fit an ordinary multiple linear regression model with all the variables using the statmodels library. We recognize that this is not the ideal model as a number of assumptions are not met such as multicollinearity. To deal with the violation of the multicollinearity assumption, we fit a reduced model with fewer variables. These are picked through trial and error and by evaluating the exploratory analysis. Lastly, we also get rid of the multicollinearity by performing pca on all th variables and transforming the first data using the first four principle components. We then fit a regression model with the transformed data. This makes the model hard to interpret but is done in hopes to improve the accuracy.

The analysis are described formally and in greater detail in the following section.

# 3 Results

## 3.1 Exploratory Data Analysis

We made three main plots. First we wanted to evaluate the relationship between all the variables and the distribution of the variables. To do this we created a scatter matrix plot and a correlation matrix heat-map. This allows to identify what variables are correlated with unemployment rates as well as asses multicollinearity in our explanatory variables. Next we plot a heat map of unemployment rates geographically. We do this to asses independence and to see if geographically close countries are also close in their unemployment rates since this may act as a confound. We see that there some regions of where unemployment is similar over countries but no systematic pattern emerges as to significantly challenge the independence assumption.

Additionally, there are some plots to better understand the distribution and summary statistics of our variables.

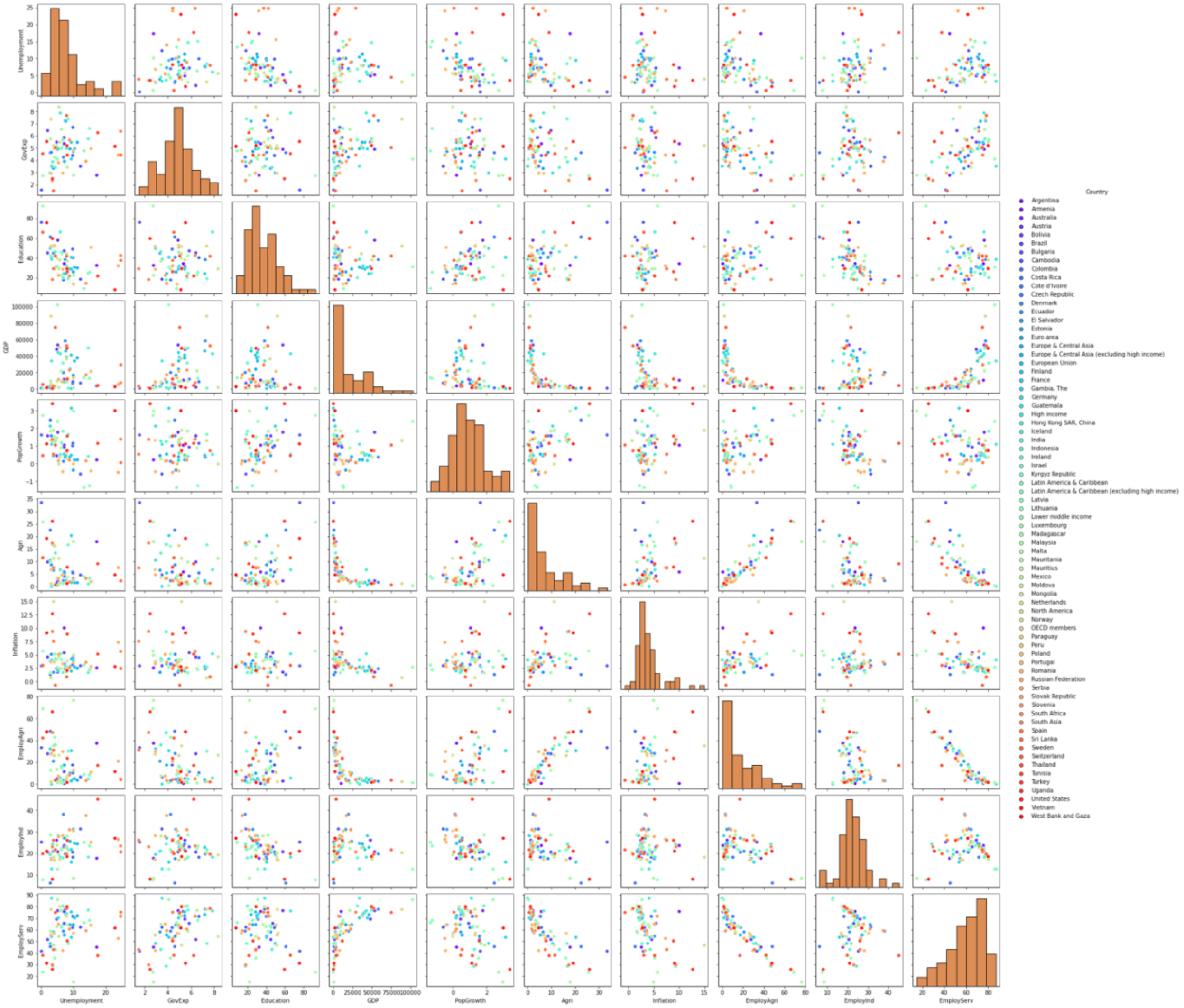
## 3.2 Full Multiple Linear Regression Model

In this model we regress unemployment rate on all the variables described in the data section. Let $x_{ji}$ be the value of jth explanatory variable for the ith sample such that $i = 1, .., n$ and $j = 1, .., 9$ where $n = 73$ is the number of countries in our data. Let $Y_i$ be the unemployment rate for the ith country. The form of the model is as follows.

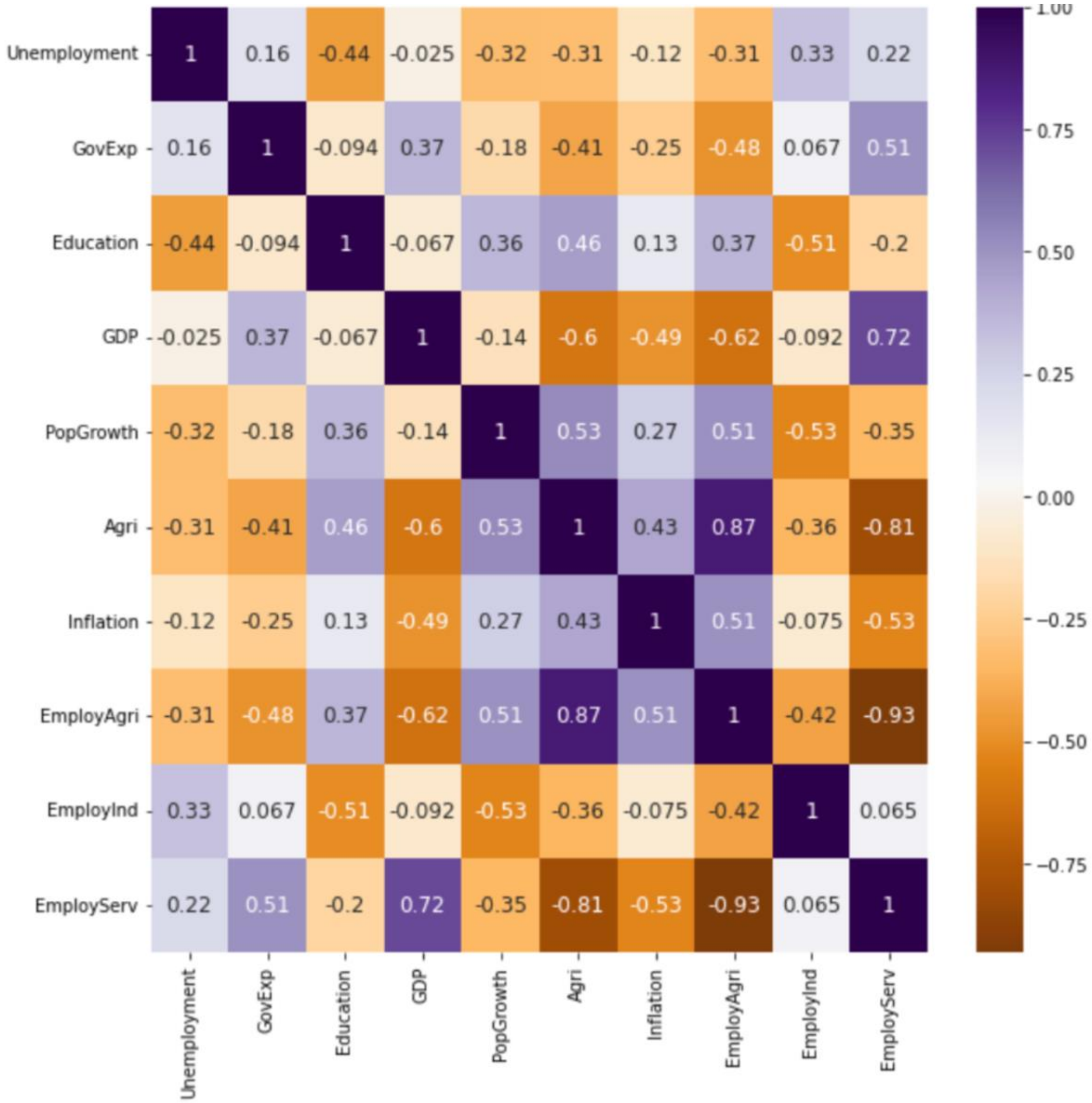
$$Y_i = \beta_1 x_{1i} + ... + \beta_9 x_{9i} + \epsilon_i$$

By the assumptions of regression, $\epsilon \sim N(0, \sigma^2)$ where sigma is a constant variance term. It then follows that $Y_i \sim N(\beta_1 x_{1i} + ... + \beta_9 x_{9i}, \sigma^2)$. We also

Plot 1

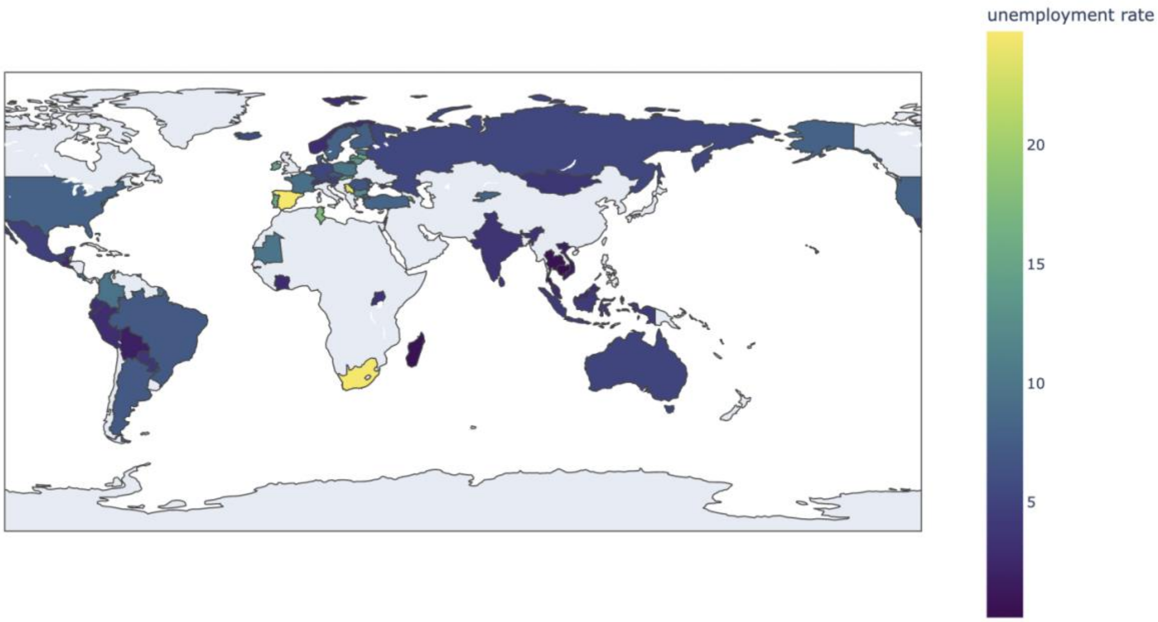

Unemployment
GovExp
Education
GDP
PopGrowth
Agri
Inflation
EmployAgri
EmployInd
EmployServ
Country
Argentina
Armenia
Australia
Austria
Bolivia
Brazil
Bulgaria
Cambodia
Colombia
Costa Rica
Cote d'Ivoire
Czech Republic
Denmark
Ecuador
El Salvador
Estonia
Euro area
Europe & Central Asia
Europe & Central Asia (excluding high income)
European Union
Finland
France
Gambia, The
Germany
Guatemala
High income
Hong Kong SAR, China
Iceland
India
Indonesia
Ireland
Israel
Kyrgyz Republic
Latin America & Caribbean
Latin America & Caribbean (excluding high income)
Latvia
Lithuania
Lower middle income
Luxembourg
Madagascar
Malaysia
Malta
Mauritania
Mauritius
Mexico
Moldova
Mongolia
Netherlands
North America
Norway
OECD members
Paraguay
Peru
Poland
Portugal
Romania
Russian Federation
Serbia
Slovak Republic
Slovenia
South Africa
South Asia
Spain
Sri Lanka
Sweden
Switzerland
Thailand
Tunisia
Turkey
Uganda
United States
Vietnam
West Bank and Gaza

Plot 2

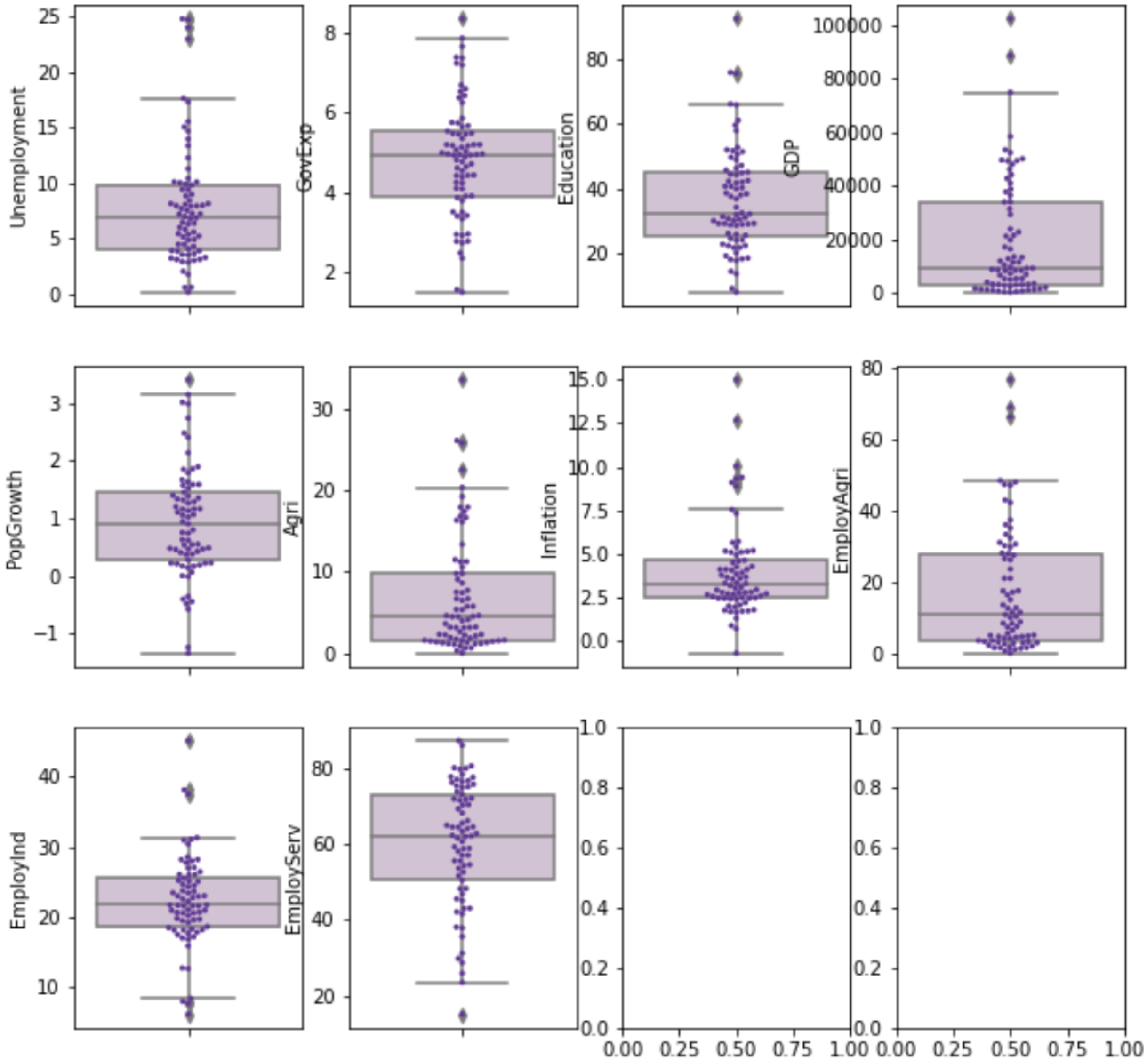


| | Unemployment | GovExp | Education | GDP | PopGrowth | Agri | Inflation | EmployAgri | EmployInd | EmployServ |
|---|---|---|---|---|---|---|---|---|---|---|
| Unemployment | 1 | 0.16 | -0.44 | -0.025 | -0.32 | -0.31 | -0.12 | -0.31 | 0.33 | 0.22 |
| GovExp | 0.16 | 1 | -0.094 | 0.37 | -0.18 | -0.41 | -0.25 | -0.48 | 0.067 | 0.51 |
| Education | -0.44 | -0.094 | 1 | -0.067 | 0.36 | 0.46 | 0.13 | 0.37 | -0.51 | -0.2 |
| GDP | -0.025 | 0.37 | -0.067 | 1 | -0.14 | -0.6 | -0.49 | -0.62 | -0.092 | 0.72 |
| PopGrowth | -0.32 | -0.18 | 0.36 | -0.14 | 1 | 0.53 | 0.27 | 0.51 | -0.53 | -0.35 |
| Agri | -0.31 | -0.41 | 0.46 | -0.6 | 0.53 | 1 | 0.43 | 0.87 | -0.36 | -0.81 |
| Inflation | -0.12 | -0.25 | 0.13 | -0.49 | 0.27 | 0.43 | 1 | 0.51 | -0.075 | -0.53 |
| EmployAgri | -0.31 | -0.48 | 0.37 | -0.62 | 0.51 | 0.87 | 0.51 | 1 | -0.42 | -0.93 |
| EmployInd | 0.33 | 0.067 | -0.51 | -0.092 | -0.53 | -0.36 | -0.075 | -0.42 | 1 | 0.065 |
| EmployServ | 0.22 | 0.51 | -0.2 | 0.72 | -0.35 | -0.81 | -0.53 | -0.93 | 0.065 | 1 |

Plot 3

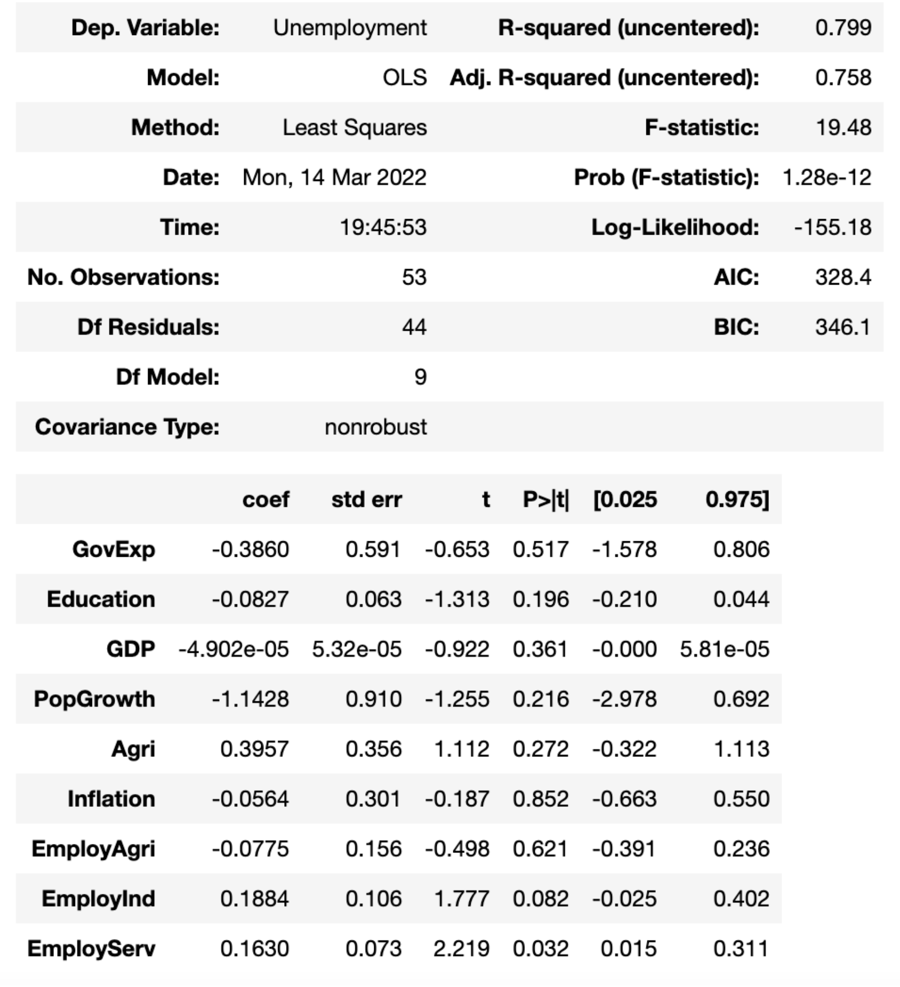


Plot 4


Unemployment
0
5
10
15
20
25
GovExp
2
4
6
8
Education
20
40
60
80
GDP
0
20000
40000
60000
80000
100000
PopGrowth
−1
0
1
2
3
Agri
0
10
20
30
Inflation
0.0
2.5
5.0
7.5
10.0
12.5
15.0
EmployAgri
0
20
40
60
80
EmployInd
10
20
30
40
EmployServ
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00

assume that $\epsilon_i$ s are iid; indicating that $Y_i$ s are independent. We also assume that $Cov(x_k., x_j.) = 0$ where $k \neq j$ or that there is no multicollinearity. This as shown by the pair plot of correlations is also not satisfied. We address this in the next iteration of the model. Lastly we assume that the relationship between unemployment and the explanatory variables is linear.

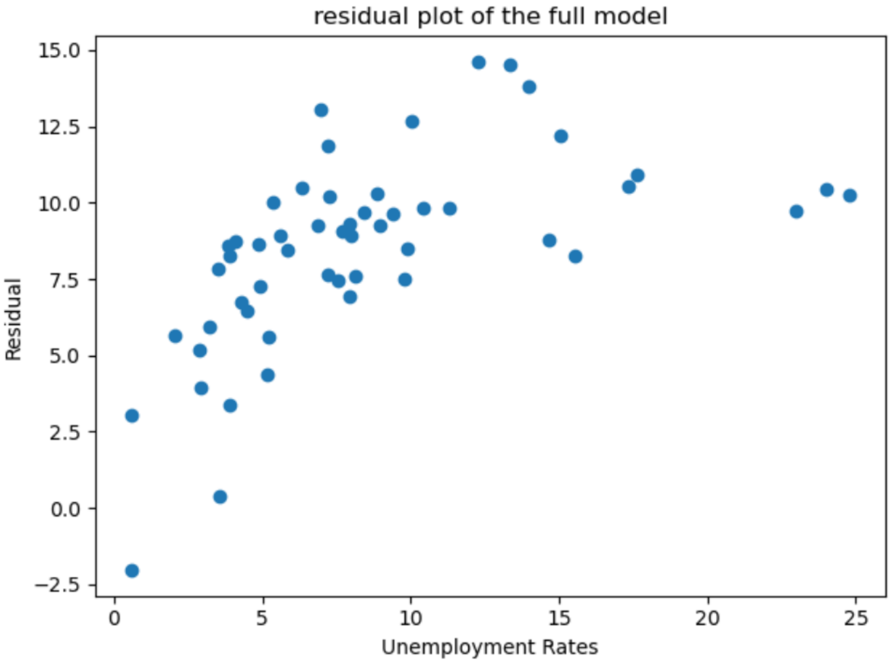
Conducting the analysis using the statsmodels library in python we find that 79.9% of the variation in unemployment can be explained by our regression analysis. The mean squared error of the test set predictions is 36.5224. Also, at $\alpha = 0.05$, only $\beta_9$ is significant.



| | | | |
|---|---|---|---|
| **Dep. Variable:** | Unemployment | **R-squared (uncentered):** | 0.799 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.758 |
| **Method:** | Least Squares | **F-statistic:** | 19.48 |
| **Date:** | Mon, 14 Mar 2022 | **Prob (F-statistic):** | 1.28e-12 |
| **Time:** | 19:45:53 | **Log-Likelihood:** | -155.18 |
| **No. Observations:** | 53 | **AIC:** | 328.4 |
| **Df Residuals:** | 44 | **BIC:** | 346.1 |
| **Df Model:** | 9 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **GovExp** | -0.3860 | 0.591 | -0.653 | 0.517 | -1.578 | 0.806 |
| **Education** | -0.0827 | 0.063 | -1.313 | 0.196 | -0.210 | 0.044 |
| **GDP** | -4.902e-05 | 5.32e-05 | -0.922 | 0.361 | -0.000 | 5.81e-05 |
| **PopGrowth** | -1.1428 | 0.910 | -1.255 | 0.216 | -2.978 | 0.692 |
| **Agri** | 0.3957 | 0.356 | 1.112 | 0.272 | -0.322 | 1.113 |
| **Inflation** | -0.0564 | 0.301 | -0.187 | 0.852 | -0.663 | 0.550 |
| **EmployAgri** | -0.0775 | 0.156 | -0.498 | 0.621 | -0.391 | 0.236 |
| **EmployInd** | 0.1884 | 0.106 | 1.777 | 0.082 | -0.025 | 0.402 |
| **EmployServ** | 0.1630 | 0.073 | 2.219 | 0.032 | 0.015 | 0.311 |

Because of the insignificant slopes, the model is hard to interpret. To evaluate the model we consider the residual plots. The plot below shows that the model often underestimates the unemployment rate and the residuals do not look evenly distributed.



Due to the assumption violations mentioned above and the residual pattern, this model may not be a good fit. For this reason we next consider a reduced model.

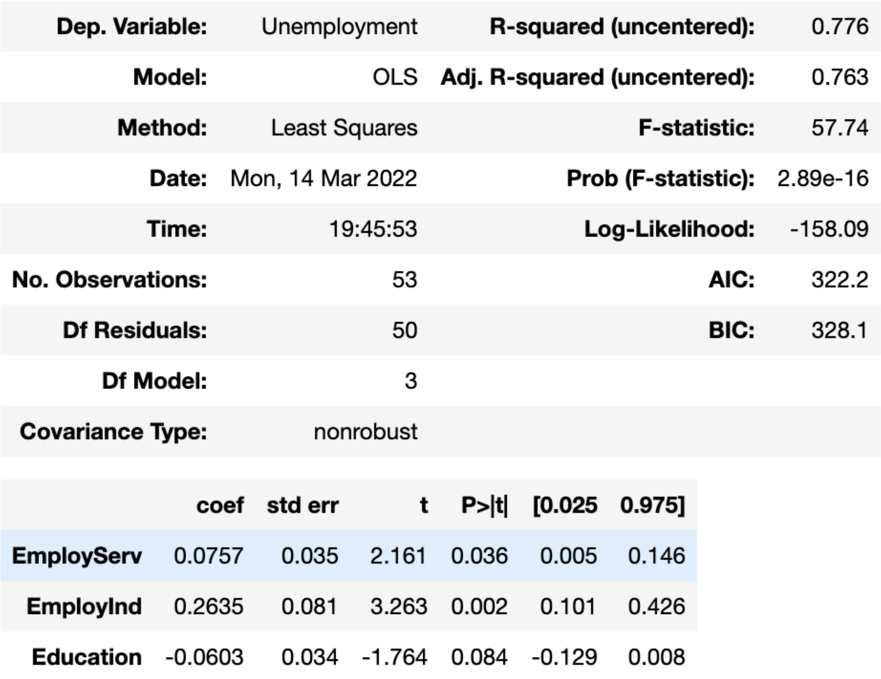
## 3.3 Reduced Multiple Linear Regression Model

In this model we regress unemployment rate on $x_2, x_8, x_9$ which are education levels, percent of the population that work in industry and the percent of the population that works in the service industry. This model has the same assumptions as the model above. Here we have picked the explanatory variables as to not encounter multicollinearity. The form of the model is as follows.

$$Y_i = \beta_1 x_{2i} + \beta_8 x_{8i} + \beta_9 x_{9i} + \epsilon_i$$

Conducting the analysis using the statsmodels library in python we find that 77.6% of the variation in unemployment can be explained by our regression analysis. The mean squared error of the test set predictions is 26.1204. While this is less than the previous model, this model is easier to interpret. Also, at $\alpha = 0.05$, $\beta_2$ and $\beta_8$ are significant.

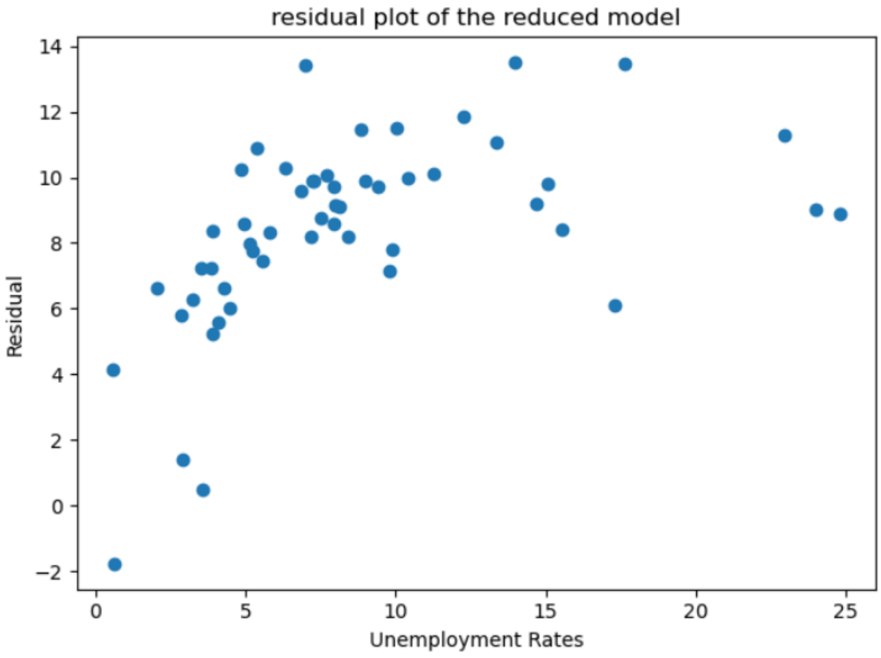
Below is the residual plot. The pattern is similar to the residual plot of the full model. This again indicates that some violation of assumptions or perhaps



| | | | |
|---|---|---|---|
| **Dep. Variable:** | Unemployment | **R-squared (uncentered):** | 0.776 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.763 |
| **Method:** | Least Squares | **F-statistic:** | 57.74 |
| **Date:** | Mon, 14 Mar 2022 | **Prob (F-statistic):** | 2.89e-16 |
| **Time:** | 19:45:53 | **Log-Likelihood:** | -158.09 |
| **No. Observations:** | 53 | **AIC:** | 322.2 |
| **Df Residuals:** | 50 | **BIC:** | 328.1 |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **EmployServ** | 0.0757 | 0.035 | 2.161 | 0.036 | 0.005 | 0.146 |
| **EmployInd** | 0.2635 | 0.081 | 3.263 | 0.002 | 0.101 | 0.426 |
| **Education** | -0.0603 | 0.034 | -1.764 | 0.084 | -0.129 | 0.008 |

the need for an intercept in the model. While this would be a worthwhile inquiry we do not pursue this further in this report.



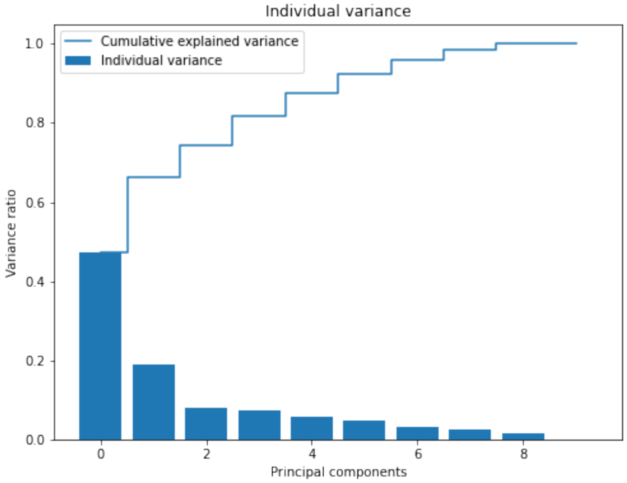
## 3.4 Principal Component Regression

In our last approach, we employ principle component analysis to project data onto an orthogonal basis to get rid of multicollinearity. A principle component is some linear combination of the variables that maximizes variation is a specific direction. We then will perform regression on the transformed data. This line of analysis is less interpretable but we hope that the accuracy of prediction improves.

To perform principle component analysis, let $X \in \mathbb{R}^{73x9}$ be the centered and scaled data matrix of the explanatory variables. We use the PCA function in the sklearn library to do pca. The plot above shows the percent of variance explained by each principle component as well as the cumulative explained variance. The first four principle components explain approximately 82% of the variance in the data matrix. We use only the data reconstructed by the first 4 principal components. This allows for dimension reduction while preserving a majority of the variation in the data. We then fit a regression model with the projected data. let $p_k$ be the data projected onto the first 4 principal components where $k = 1, ..., 4$. The model is as follows

$$Y_i = \beta_1 p_{1i} + ... + \beta_4 p_{4i} + \epsilon_i$$

Again, the assumptions of this model are as described in the full model section. Because of the data is transformed by PCA we know that the variables

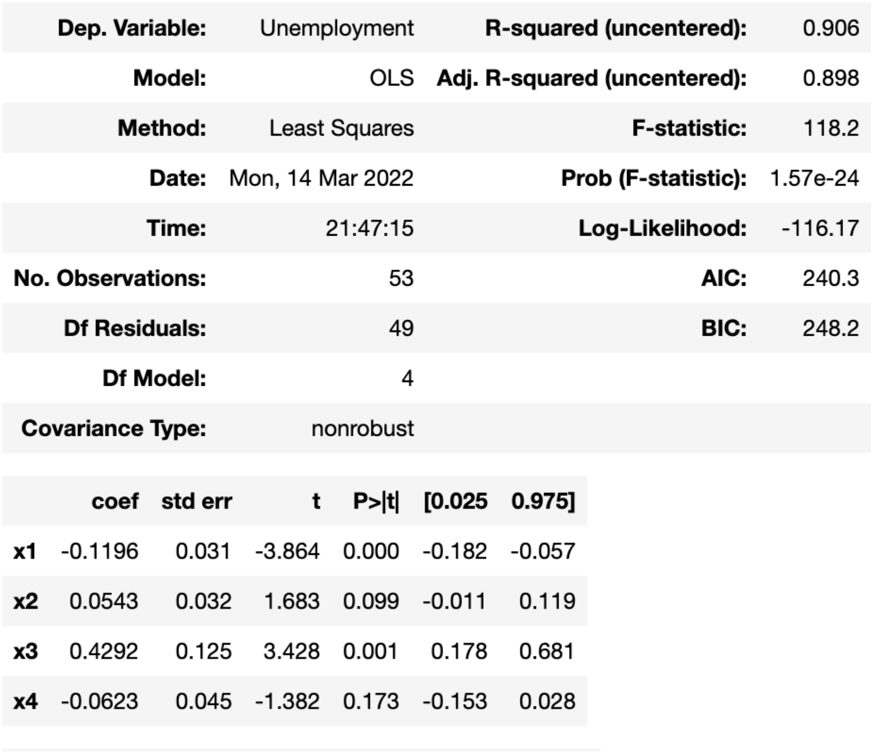


can not by multicollinear.

Fitting the regression model we find that 90.6% of the variation in unemployment can be explained by our regression analysis. At $\alpha = 0.05$, all $\beta_1, \beta_2, \beta_3$ are significant.

The mean squared error for the test set is 33.4202. This shows an improvement from previous models but is hard to interpret. In an effort to do so, we can look at the loading scores of the pca to understand what the linear combination each transformed variable contains.

The red circles show features that have a high weight in the principle component(the loadings table is at the end of the file). While this gives us a sense of what each principle component means. For example, the third (2 in the table) principle component is composed primarily of government expenditures. The third principle component has a positive slope in the regression model. Generally indicating that and increase in government expenditures is related to a slight increase in unemployment. This interpretation should be taken with a grain of salt and may not be wholly reflective since there are features with a negative weight in this loading.



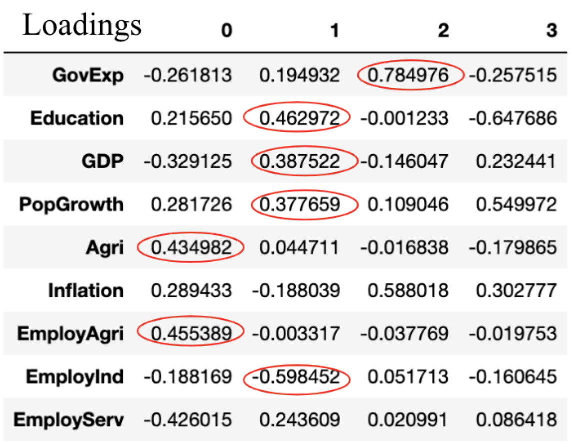
| | | | |
|---|---|---|---|
| **Dep. Variable:** | Unemployment | **R-squared (uncentered):** | 0.906 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.898 |
| **Method:** | Least Squares | **F-statistic:** | 118.2 |
| **Date:** | Mon, 14 Mar 2022 | **Prob (F-statistic):** | 1.57e-24 |
| **Time:** | 21:47:15 | **Log-Likelihood:** | -116.17 |
| **No. Observations:** | 53 | **AIC:** | 240.3 |
| **Df Residuals:** | 49 | **BIC:** | 248.2 |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **x1** | -0.1196 | 0.031 | -3.864 | 0.000 | -0.182 | -0.057 |
| **x2** | 0.0543 | 0.032 | 1.683 | 0.099 | -0.011 | 0.119 |
| **x3** | 0.4292 | 0.125 | 3.428 | 0.001 | 0.178 | 0.681 |
| **x4** | -0.0623 | 0.045 | -1.382 | 0.173 | -0.153 | 0.028 |

## 4 Conclusion

Our project analyzed 9 different macroeconomic variables and its relationship with Unemployment rates in a country. To find optimal relationships we model macroeconomic variables through regression analysis. The first full linear regression model with all variables led to issues of multicollinearity and insignificant variables. The first fix to this issue is the reduced model that only utilizes 3 out of 9 macroeconomic variables. The model uses education levels, percent of the population that work in industry, and the percent of the population that works in the service industry. This model stops multicollinearity and passes the T-test of coefficients. However, this model faces issues with the residuals. The final model is the Principal Component Regression. This model performs principal component analysis on the regression. PCR leads to the best model with almost no problems however making interpretations is difficult.

Using the reduced model, countries that had a mix or strong rates in one of the 3 main predictor variables consequently had low unemployment. If we use this model, the next project should find out how countries can increase education rates in their labor force and how to create more employment in the Service and Industry.



| Loadings | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| GovExp | -0.261813 | 0.194932 | 0.784976 | -0.257515 |
| Education | 0.215650 | 0.462972 | -0.001233 | -0.647686 |
| GDP | -0.329125 | 0.387522 | -0.146047 | 0.232441 |
| PopGrowth | 0.281726 | 0.377659 | 0.109046 | 0.549972 |
| Agri | 0.434982 | 0.044711 | -0.016838 | -0.179865 |
| Inflation | 0.289433 | -0.188039 | 0.588018 | 0.302777 |
| EmployAgri | 0.455389 | -0.003317 | -0.037769 | -0.019753 |
| EmployInd | -0.188169 | -0.598452 | 0.051713 | -0.160645 |
| EmployServ | -0.426015 | 0.243609 | 0.020991 | 0.086418 |

## 5 Sources

[1 ] https://www.investopedia.com/articles/economics/10/unemployment-rate-get-real.asp

[2 ] https://towardsdatascience.com/central-limit-theorem-70b12a0a68d8

[3 ] https://towardsdatascience.com/assumptions-of-linear-regression-fdb71ebeaa8b

# final.py

```
import pandas as pd
import statsmodels.api as sm
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import numpy as np
import os

#%%
# Reading the data

df = pd.read_csv('./Project/141b_final_project/unemploymentdata.csv')
df

#%%
# pair plots of the varibales
#off diagonal is the scatter plot for two variables
#diagnonal is the histogram od each quantitative variables
graph = sns.pairplot(df,hue = 'Country', palette = 'rainbow')
graph.map_diag(sns.histplot, hue = None, color = 'chocolate')
#graph.savefig("scatter_plot_matrix.png")

#%%
# correlation heat map
plt.figure(figsize = (10,10))
sns.heatmap(df.corr(), annot = True,annot_kws={"size": 12},cmap="PuOr")

#%%
from plotly.graph_objs import Scatter, Figure, Layout
import plotly
import plotly.graph_objs as go
import json
import numpy as np
import plotly.express as px

fig = px.choropleth(df,locationmode = 'country names' ,locations= 'Country', color='Unemployment',
                           color_continuous_scale="Viridis",
                           scope="world",
                           labels={'Unemployment':'unemployment rate'}
                          )
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()

#%%

# box plots
f,ax=plt.subplots(3,4,figsize=(10,10))
for i in range(1,5):
    sns.boxplot(x= None , y= df.columns[i], data = df, ax = ax[0,i-1],color = 'thistle')
    sns.swarmplot(x= None , y= df.columns[i], data = df,ax = ax[0,i-1], size = 3, color = 'rebeccapurple' )
for i in range(5,9):
    sns.boxplot(x= None , y= df.columns[i], data = df, ax = ax[1,i-5],color = 'thistle')
    sns.swarmplot(x= None , y= df.columns[i], data = df,ax = ax[1,i-5], size = 3, color = 'rebeccapurple' )
for i in range(9,11):
    sns.boxplot(x= None , y= df.columns[i], data = df, ax = ax[2,i-9],color = 'thistle')
    sns.swarmplot(x= None , y= df.columns[i], data = df,ax = ax[2,i-9], size = 3, color = 'rebeccapurple' )

#%%
# splitting the data into a training and testing set
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
x = df.iloc[::,2:]
y = df['Unemployment']
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.27,random_state=0)

#%%
# fitting the full model
fullmodel = sm.OLS(y_train, x_train).fit()
fullmodel.summary()

#%%
# prediction accuracy:

yhat_test = fullmodel.predict(x_test)
MSE = sum((y_test.values - yhat_test.values)**2)/len(y_test)
MSE

#%%
# Residual plots

pred_val = fullmodel.predict(x_train)
```

```
residual = y_train - pred_val
plt.scatter(y_train, pred_val)
plt.title("residual plot of the full model")
plt.xlabel("Unemployment Rates")
plt.ylabel("Residual")
plt.show()
plt.savefig("residual_fullmodel.png")


#%%
# Reduced model

reducedmodel = sm.OLS(y_train,x_train[['EmployServ','EmployInd','Education']]).fit()
reducedmodel.summary()

#%%
# prediction accuracy

yhat_test_reduced = reducedmodel.predict(x_test[['EmployServ','EmployInd','Education']])
MSE = sum((y_test.values - yhat_test_reduced.values)**2)/len(y_test)
MSE

#%%
# residual plots
pred_val = reducedmodel.predict(x_train[['EmployServ','EmployInd','Education']])
residual = y_train - pred_val
plt.scatter(y_train, pred_val)
plt.title("residual plot of the reduced model")
plt.xlabel("Unemployment Rates")
plt.ylabel("Residual")
plt.show()
plt.savefig("residual_reducedmodel.png")

#%%

#full pca to see how many PCs to pick
scaler = StandardScaler()
std_df = scaler.fit_transform(x)
pca = PCA()
pca.fit(std_df)
pca_variance =  pca.explained_variance_ratio_
cum_sum_eigenvalues = np.cumsum(pca_variance)

plt.bar(list(range(1,10)), pca_variance, align="center", label="Individual variance")
plt.step(list(range(1,len(cum_sum_eigenvalues)+1)), cum_sum_eigenvalues, where='mid',label='Cumulative explained variance')
plt.legend()
plt.ylabel("Variance ratio")
plt.xlabel("Principal components")
plt.title("Variance captured by the principle components")
plt.savefig("pca variance explained plot")
plt.show()

#%%

# the first four principle components for the analysis and transform the data
x_full = pd.concat([x_train,x_test])
scaler = StandardScaler()
std_df = scaler.fit_transform(x_full)
pca4 = PCA(n_components=4)
pca4.fit(std_df)
x_4d = pca4.transform(x_full)




#%%
# replit
p_train = x_4d[:53,::]
p_test = x_4d[53:,::]

#%%

# regression with the transformed data
pca_model = sm.OLS(y_train, p_train).fit()
#pca_model = sm.OLS(y_train, sm.add_constant(p_train)).fit()
pca_model.summary()

#%%
yhat_test_pca = pca_model.predict(p_test)
MSE = sum((y_test.values - yhat_test_pca)**2)/len(y_test)
MSE

#%%
# looking at the loadings
loadings = pd.DataFrame(pca4.components_.T, index=df.columns[2:])
```