Introduction

Time series data is a series of data points collected at regular intervals of time. In the context of the economy, time series data can be used to track various economic indicators such as GDP, inflation, employment, and trade. These indicators are often measured quarterly or monthly and can be used to understand the overall health and trend of the economy over time. Time series data can also be used to make forecasts about future economic performance and to identify trends and patterns in economic activity. Following are the term used in dataset that need to be defined in order to understand them rightly.

Unemployment rate: The unemployment rate is the percentage of the labor force that is unemployed but actively seeking employment and willing to work. It is a measure of the health of the job market and is closely watched by policymakers, businesses, and consumers. A low unemployment rate is generally seen as a sign of a strong economy, while a high unemployment rate can indicate a weak economy and difficulty for workers to find jobs.

GDP growth: GDP, or gross domestic product, is a measure of the total value of goods and services produced in an economy. GDP growth refers to the increase in GDP over a certain period of time, usually measured quarter-by-quarter or year-by-year. GDP growth is often used as a key indicator of the overall health and growth of an economy.

Inflation: Inflation is an increase in the overall price level of goods and services in an economy over a period of time. It is usually measured as an annual percentage increase. Central banks attempt to maintain a target inflation rate, as too high or too low of an inflation rate can have negative effects on an economy.

Job vacancies: A job vacancy is a position in a company or organization that is available and ready to be filled. The number of job vacancies can be an indicator of the strength of the job market and the demand for labor. A high number of job vacancies may indicate a strong job market with many opportunities for workers, while a low number of job vacancies may indicate a weak job market with few available positions.

Population: The population of a place is the number of people living there. It is often used to refer to the number of people living in a city, country, or other geographic area. The population of a place can be determined by a census, which is a count of the number of people living in an area. The size of a population can be affected by many factors, including birth rates, death rates, immigration, and emigration. Understanding population trends can be important for a variety of reasons, including planning for housing, education, and healthcare needs.

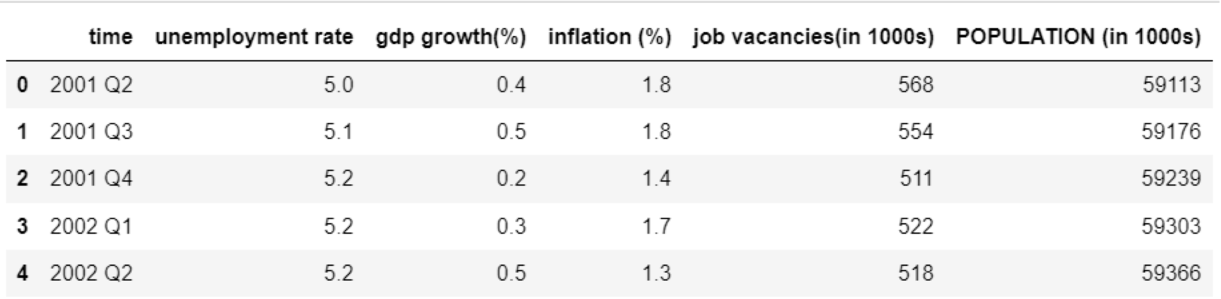
A dataset that contains GDP, population, inflation, and job vacancies as columns and has quarterly data would likely be used to analyze economic trends and conditions in a particular place or region. By analyzing this data on a quarterly basis, it is possible to see how these economic indicators are changing over time and to identify trends and patterns. For example, if GDP is increasing but population and job vacancies are decreasing, it could indicate that the economy is growing but there is a lack of job opportunities. On the other hand, if population and job vacancies are increasing while GDP remains stable or decreases, it could indicate that the economy is struggling to keep up with the demands of a growing population. This type of dataset could be used by governments, businesses, and other organizations to make informed decisions and to plan for the future. This report is about calculating economic trends of countries of quarterly basis using the dataset given (described in dataset column).

Dataset

The dataset used in the study includes historical economic data as well as some of the other factors that may have affected them. It includes data for the 22-year time period from 2001 to September, 2022. The data were collected quarterly. The raw dataset has 86 rows and 6 columns that contain unwanted rows and columns with missing values. It has 232 rows and 8 columns after cleaning and removing those unwanted values. The 6 columns are time, Unemployment rate, gdp growth, inflation, job vacancies, and Population.

- **Time:** It contain the data about year with specifying quarter of the year
- **Unemployment rate:** The unemployment rate is the percentage of the labor force that is unemployed but actively seeking employment and willing to work.
- **Gdp growth:** GDP, or gross domestic product, is a measure of the total value of goods and services produced in an economy.
- **Inflation:** Inflation is an increase in the overall price level of goods and services in an economy over a period of time.
- **Job vacancies:** A job vacancy is a position in a company or organization that is available and ready to be filled.
- **Population:** The population of a place is the number of people living there. It is often used to refer to the number of people living in a city, country, or other geographic area.

The head is the top part of the DataFrame. It is used to view a small sample of the data in the DataFrame, typically the first few rows. You can use the head method to view the head of a DataFrame. Figure 1 shows the head of the dataset and showing top 5 rows of all the columns. It can be seen that gdp and inflation are given in percentage while job vacancies and population are given in 1000s.



	time	unemployment rate	gdp growth(%)	inflation (%)	job vacancies(in 1000s)	POPULATION (in 1000s)
0	2001 Q2	5.0	0.4	1.8	568	59113
1	2001 Q3	5.1	0.5	1.8	554	59176
2	2001 Q4	5.2	0.2	1.4	511	59239
3	2002 Q1	5.2	0.3	1.7	522	59303
4	2002 Q2	5.2	0.5	1.3	518	59366

Figure 1 Head of the dataset

Statistical Description

A statistical description of a dataset is a summary of its main characteristics using statistical measures and graphs. Some common statistical measures used to describe a dataset include the mean, median, mode, standard deviation, and range. These measures can help you understand the overall pattern and distribution of the data, and identify any outliers or unusual values.

There are also various types of graphs that can be used to visualize and summarize a dataset, such as histograms, box plots, and scatter plots. These graphs can help you see the shape and spread of the data, and understand how the data is distributed.

Overall, a statistical description of a dataset can provide important insights into the data and help you understand the relationships and patterns within it. The important term used in statistical analysis are as follow.

Mean: The mean of a dataset is the sum of all the values in the dataset divided by the number of values in the dataset. It is a measure of the central tendency of the data.

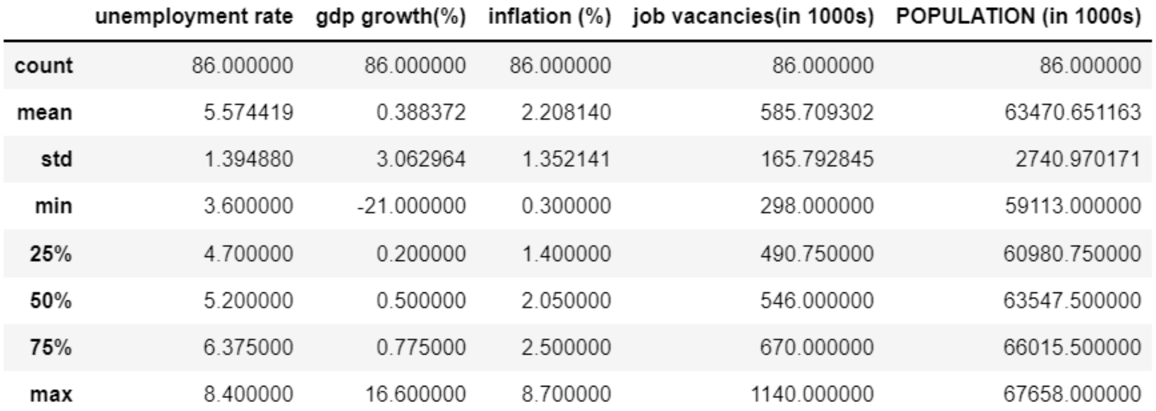
Minimum: The minimum value of a dataset is the smallest value in the dataset.

Maximum: The maximum value is the largest value in the dataset.

Count: The count of a dataset is the number of values in the dataset.

Standard deviation: The standard deviation of a dataset is a measure of the spread or dispersion of the data. It is calculated by taking the square root of the variance, which is the average of the squared differences of the values from the mean. A small standard deviation indicates that the values in the dataset are close to the mean, while a large standard deviation indicates that the values are spread out over a wider range.

The figure 2 shows the statistical analysis of the dataset and it was found that there were no outlier (unusual activity) found on it.



	unemployment rate	gdp growth(%)	inflation (%)	job vacancies(in 1000s)	POPULATION (in 1000s)
count	86.000000	86.000000	86.000000	86.000000	86.000000
mean	5.574419	0.388372	2.208140	585.709302	63470.651163
std	1.394880	3.062964	1.352141	165.792845	2740.970171
min	3.600000	-21.000000	0.300000	298.000000	59113.000000
25%	4.700000	0.200000	1.400000	490.750000	60980.750000
50%	5.200000	0.500000	2.050000	546.000000	63547.500000
75%	6.375000	0.775000	2.500000	670.000000	66015.500000
max	8.400000	16.600000	8.700000	1140.000000	67658.000000

Figure 2 Statistical description of dataset

Analysis

The analysis of a dataset refers to the process of examining, cleaning, transforming, and modeling the data in order to discover useful insights and answer relevant questions. There are many different approaches and techniques that can be used to analyze a dataset, depending on the nature of the data and the research question being addressed.

Some common steps involved in the analysis of a dataset include:

Importing and cleaning the data: This involves reading the data into a software program and checking for errors, missing values, and other issues that need to be addressed before the data can be analyzed.

Exploratory data analysis: This involves visualizing and summarizing the data in order to get a better understanding of its characteristics and identify any patterns or trends.

Modeling and analysis: This involve using statistical or machine learning techniques to build models of the data and test hypotheses or make predictions.

Interpreting and communicating results: This involves presenting the findings of the analysis in a clear and concise way, and discussing their implications and limitations.

Overall, the goal of the analysis of a dataset is to extract useful insights and knowledge from the data that can inform decision-making or further research.

Univariant Analysis

Univariate analysis is a statistical technique that deals with the analysis of data that consists of a single variable. It involves summarizing and describing the data, and identifying patterns and trends in the data. Univariate analysis is often used to understand the characteristics of a single variable, and to identify any unusual or extreme values that may be present in the data. It is a useful tool for exploring and understanding data, and for identifying relationships between variables.

Bar Graph: A bar graph is a graphical display of data using bars of different heights. The bars can be plotted vertically or horizontally. A bar graph is used to compare the sizes of different quantities or to compare the sizes of quantities over time**.**

Line Plot: A line plot is a graphical display of data using a number line. It is used to display the distribution of a single variable. Each value is plotted as a point on the number line, and then a line is drawn to connect the points. Line plots are often used to show trends over time or to compare the distribution of a variable between different groups.

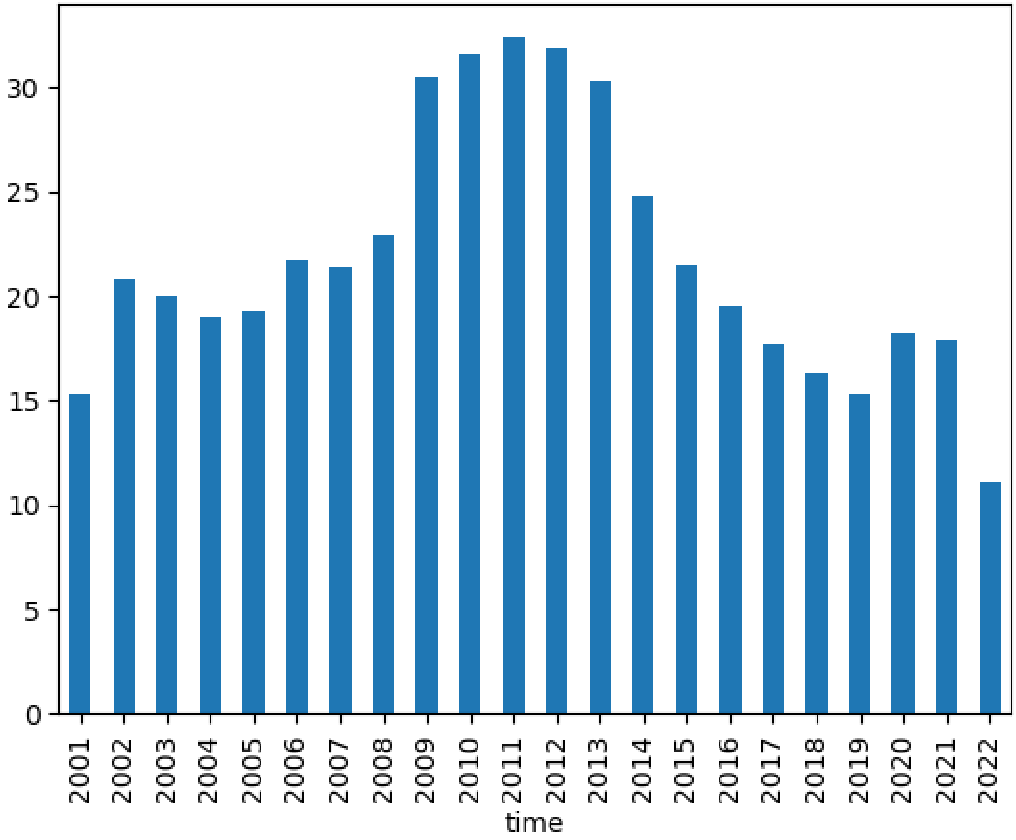


Figure 3 Unemployment rate during the period of 2001 to 2022

The figure 3 shows the bar plot of the unemployment rate and it can be seen that it was very high during the year of 2009 to 2014. It was found that highest value is record in 2011 and lowest was recorded in 2022 which mean in 2011 most labor force are searching for job and they were unemployed at that time and while in 2022 there are very low people who are searching for job because most of them were already doing job. Figure 4 shows the GDP growth which was seen to be very low during the year of 2020 that was most probably because of corona outbreaks which affects the life of people and also their business and after that as can be seen a drastic increase is recorded goes to 8+ which is highest of all the time and then again fall to 4 in next year. Before it a low value was found in 2007 that was approximately -4.

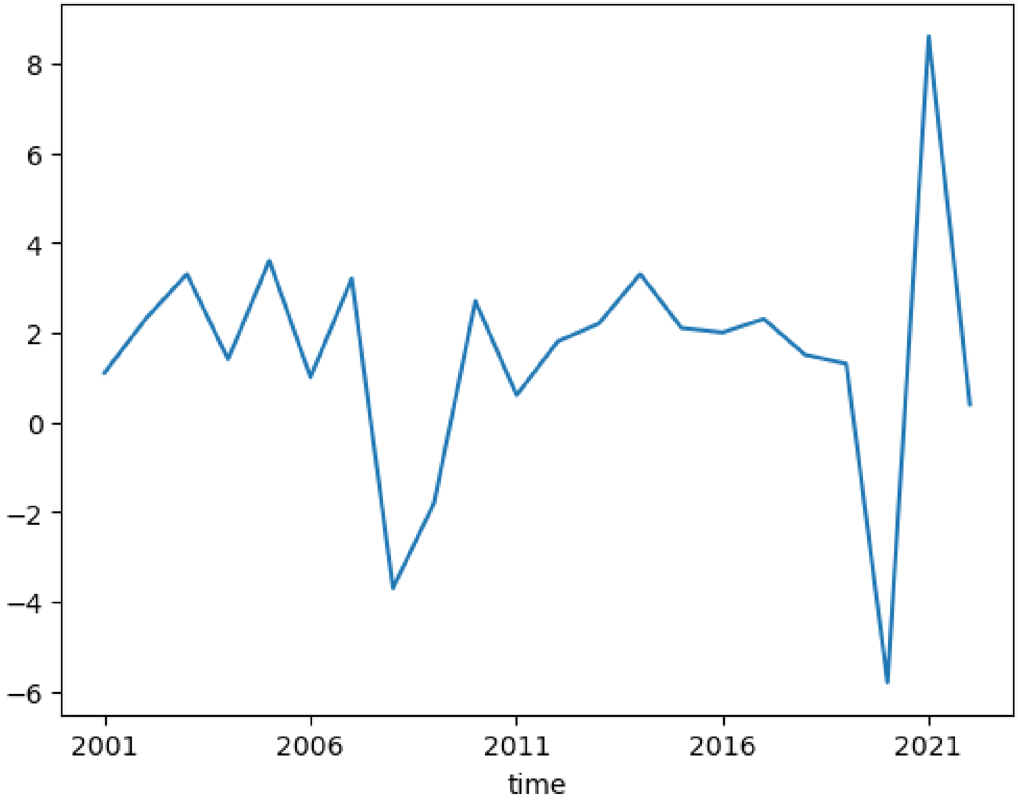


Figure 4 GDP growth during the period of 2001 to 2022

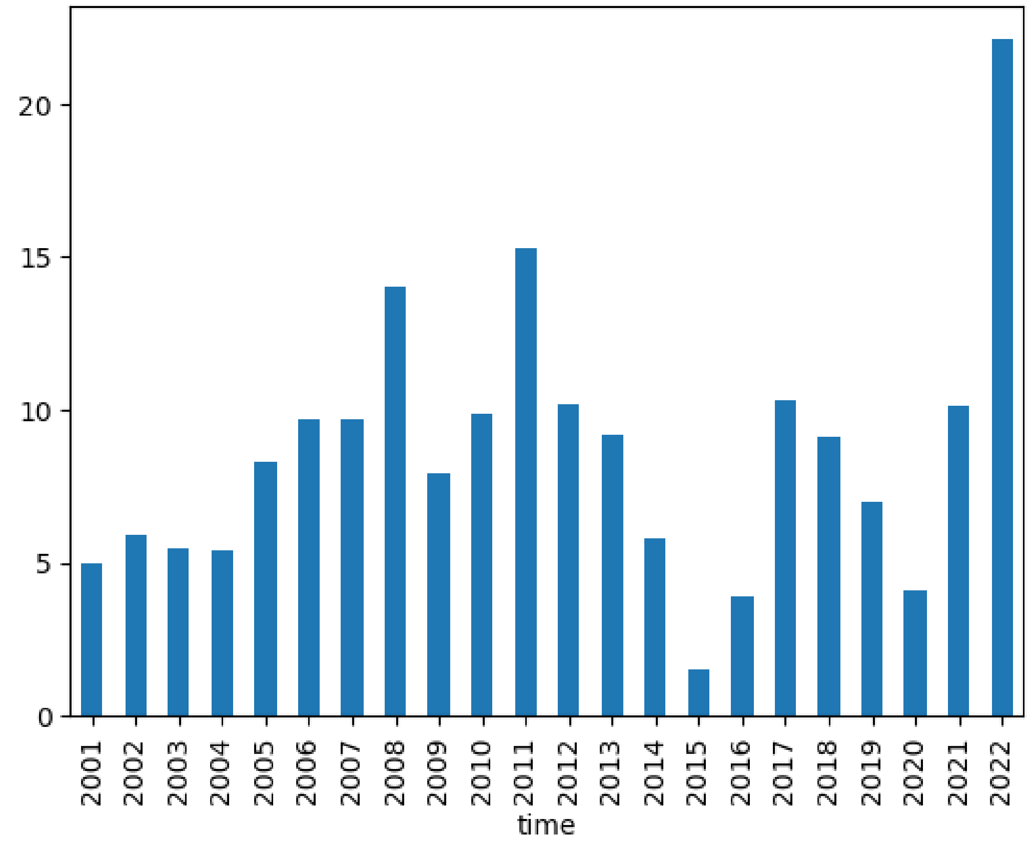


Figure 5 Inflation during the period of 2001 to 2022

Figure 5 shows the inflation of whole duration and it can be seen that inflation was very low in 2015 while it is highest during the year 2022 which i think is due to the after effect of the corona which has impact the market very badly and production and preparation would be stop due to it and now after two year we are affected by these issues. If we look at graph above it can be seen that it was high or low during different time periods. It continually raises during 3 year after 2005 and it fallen down during year 2015 from 5+ to 2 value. Job vacancies data shows that how many vacancies have in each year and by seeing figure 6 it can be seen that there are around 3000+ job vacancies during year 2021 and 2022.Sudden decrease in vacancy number is found during 2005 to 2006 and considerable increase was shown during 2020 to 2021.Figure 7 shows the population record which from 2002 to 2021 has increase of fixed amount every year while in 2001 to 2002 the rate increase is high and while 2021 to 2022 it decrease by number which previously increasing from about previous 20 years.

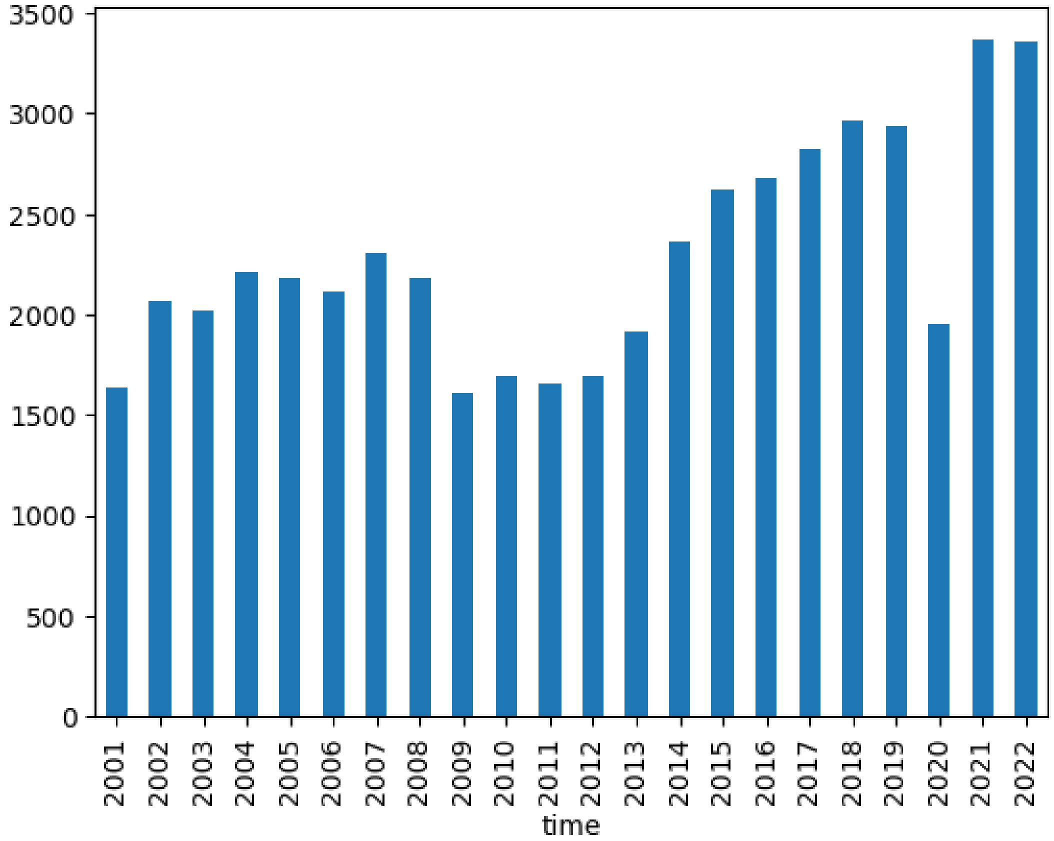


Figure 6 Job Vacancies during 2001 to 2022

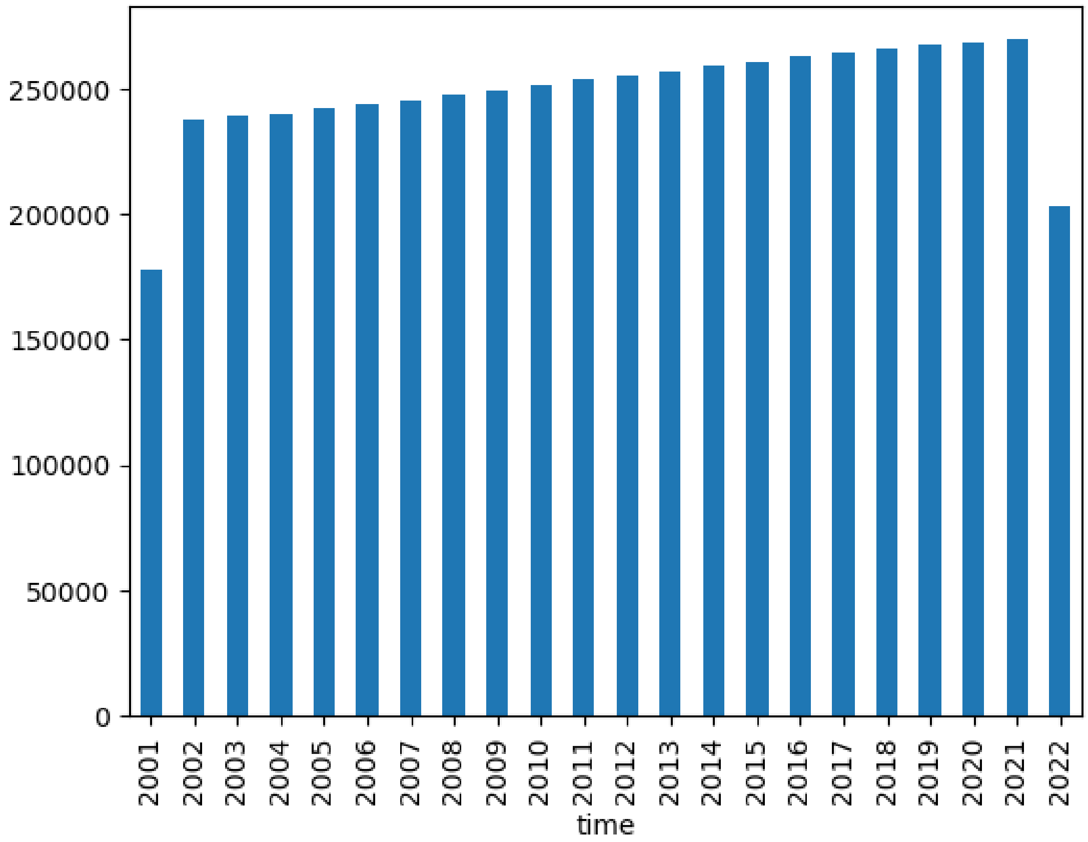


Figure 7 Population counts during the duration 2001 to 2022

After doing univariant analysis of every given column or variable which is on yearly record now we look it in quarterly record. Firstly, unemployment rate quarterly record is shown in figure 8 and it is shown that Q2 and Q3 has most unemployment rate then other twos. Moreover, figure 9 shows GDP growth of quarterly and it shows that most GDP growth was during Q3 and it was negative during Q1 and Q2. If we look at quarterly record of inflation Q3 has most value for quarter while Q1 and Q4 has nearly same value for inflation the quarterly graph of inflation can be seen in figure 10.

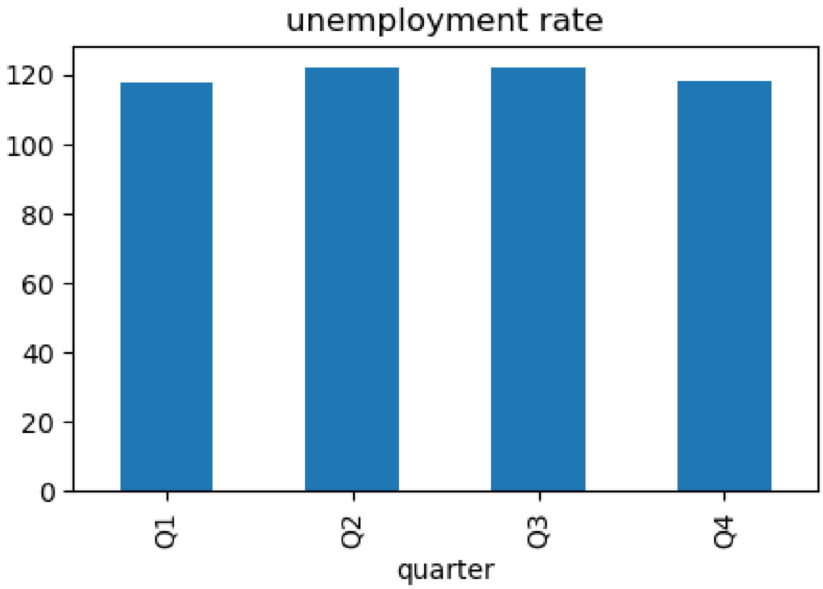


Figure 8 Quarterly record of the unemployment rate

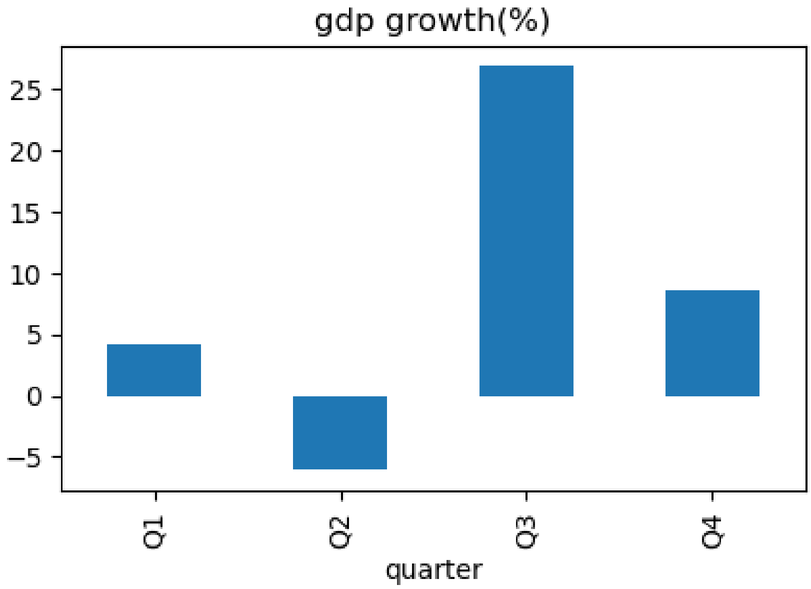


Figure 9 GDP growth of Quarterly

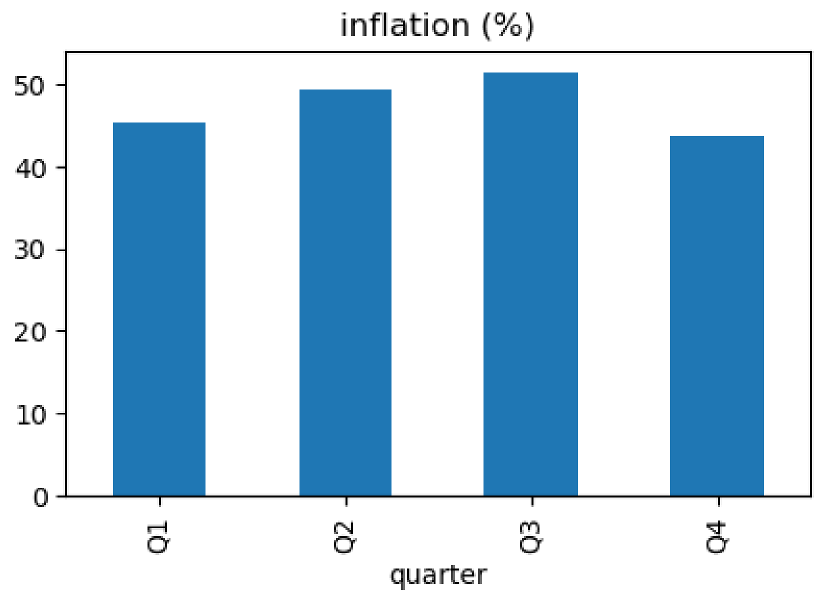


Figure 10 Inflation quarterly record

Job vacancies quarterly record is shown in figure 11 that shows that Q2 and Q3 has most number of job vacancies than other two quarter while the difference is not as much significant as it can be. Figure 12 shows the population record of Q2 and Q3 which as high population then other two which is quite common to upper graph.

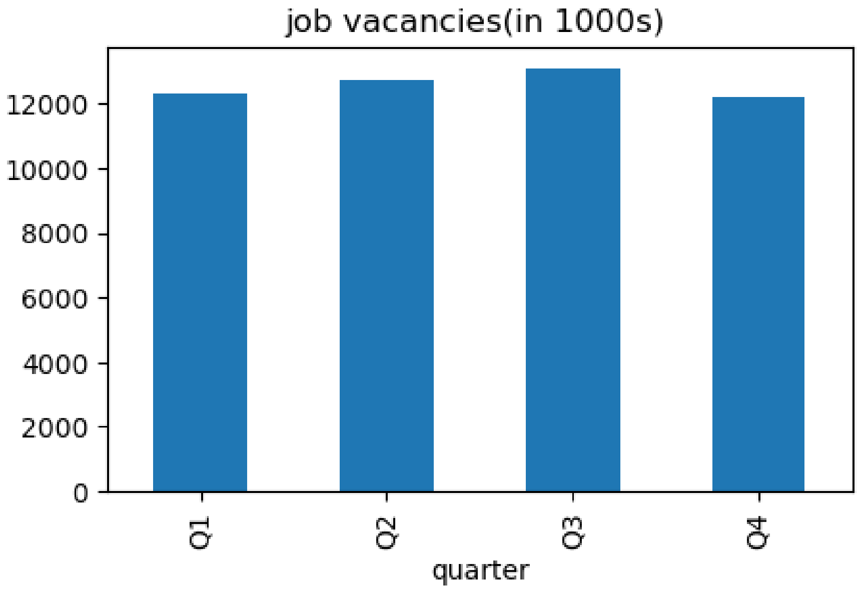


Figure 11 Job vacancies quarterly record

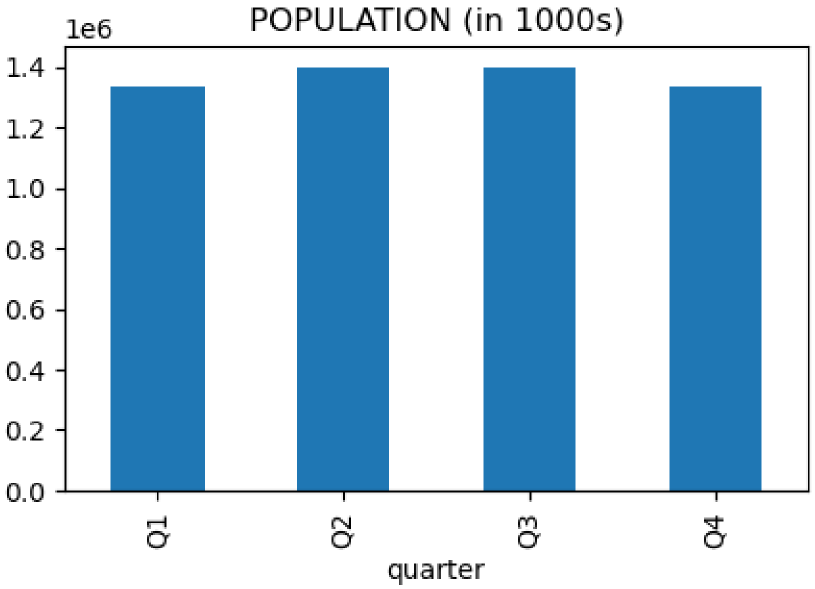


Figure 12Population quarterly record

Bivariate Analysis

Bivariate analysis is a statistical method that is used to examine the relationship between two variables. It involves analyzing the data to determine if there is a correlation between the two variables, and to understand the nature of that relationship. There are several different techniques that can be used in bivariate analysis, including:

Scatter plots: These are plots of the data on a graph, with one variable on the x-axis and the other on the y-axis. A scatter plot can show the relationship between the two variables, and can help to identify trends or patterns in the data.

Correlation coefficients: These are statistical measures that describe the strength and direction of the relationship between two variables. The most commonly used correlation coefficient is the Pearson correlation coefficient, which ranges from -1 to 1. A value of 1 indicates a strong positive correlation, a value of -1 indicates a strong negative correlation, and a value of 0 indicates no correlation.

Regression analysis: This is a statistical method that is used to model the relationship between two variables. It involves fitting a line to the data that represents the relationship between the two variables, and can be used to make predictions about one variable based on the value of the other.

Bivariate analysis is a useful tool for understanding the relationship between different variables, and can be used to make predictions about future trends or patterns. It is often used in fields such as economics, psychology, and sociology to understand the factors that influence behavior or outcomes.

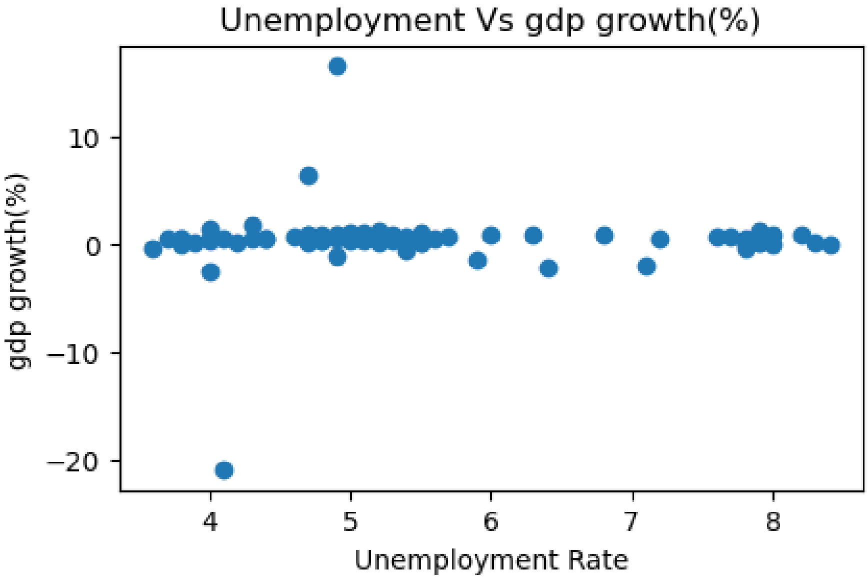


Figure 13 GDP and Unemployment Rate

As seen in figure 13 the unemployment rate is increase during different GDP growth when GDP is around 10 the unemployment reaches to 8 point and it can be seen that during GDP growth near 10 the unemployment is low compared to other. Overall, unemployment and GDP growth has no significant correlation between them. Figure 14 shows that there is no significant correlation between inflation and unemployment.

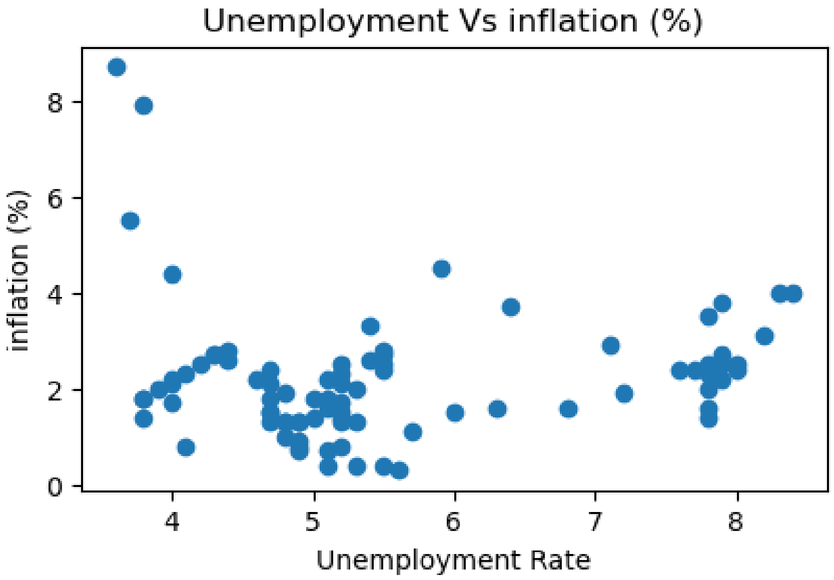


Figure 14 Inflation and Unemployment rate

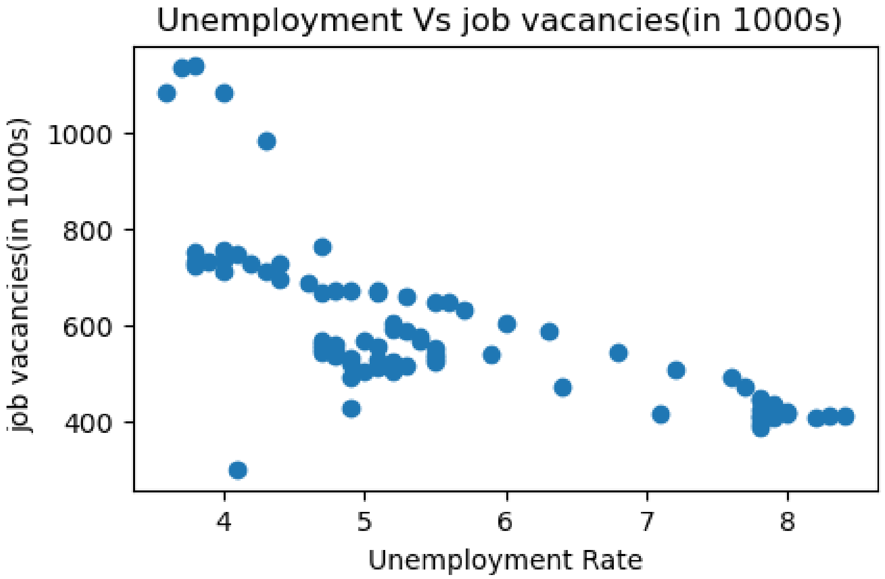


Figure 15 Unemployment and Job vacancies

Figure 15 shows that the job vacancies and unemployment have inverse relationship as job vacancies decreases the unemployment rate of the country increases and which seem that if there is no job vacancy than most of graduate seek for job and then they are unemployed at that time. While if there is job vacancy then the unemployment rate is decreases which shows that job in market will attract peoples. Figure 16 shows that there is no relation between unemployment and population.

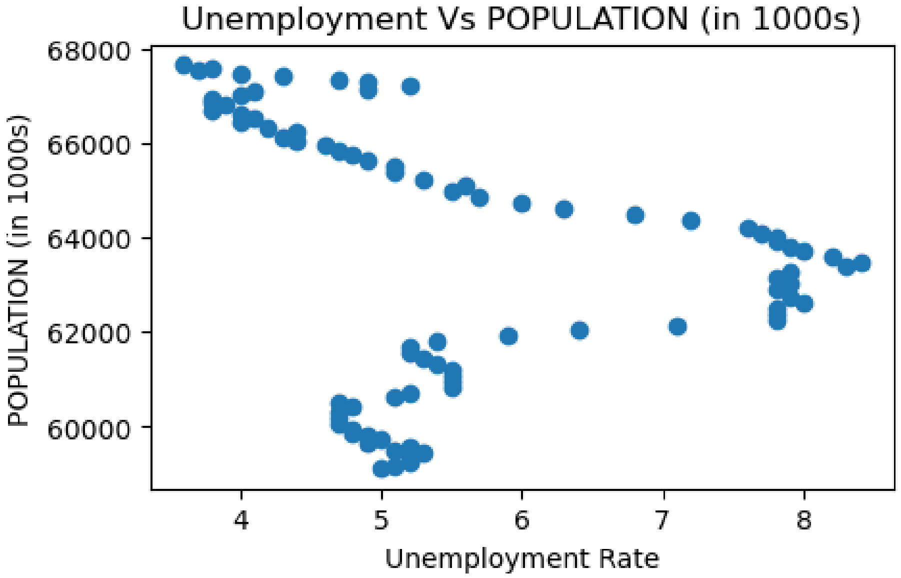


Figure 16 Unemployment and Population

Correlation refers to the relationship between two variables. When we talk about the relationship between two variables, we are trying to understand how the values of one variable might be related to the values of another variable.

There are three types of relationships that can occur between two variables: positive correlation, negative correlation, and no correlation.

Positive Correlation: A positive correlation occurs when an increase in one variable is associated with an increase in the other variable. For example, there might be a positive correlation between the number of hours someone studies and their test scores. This means that if someone studies more, they might be more likely to get a higher test score.

Negative Correlation: A negative correlation occurs when an increase in one variable is associated with a decrease in the other variable. For example, there might be a negative correlation between the number of hours someone watches TV and their test scores. This means that if someone watches more TV, they might be less likely to get a high-test score.

No Correlation: No correlation occurs when there is no relationship between the two variables. This means that the value of one variable does not have any effect on the value of the other variable.

A correlation graph is a graphical representation of the relationship between two variables. It can be used to visualize the strength and direction of the relationship between the variables. In a scatter plot, the position of each point on the graph indicates the values of the two variables for a particular data point. If there is a positive correlation between the variables, the points on the scatter plot will tend to slope upward from left to right. If there is a negative correlation, the points will tend to slope downward from left to right. If there is no correlation, the points will be scattered randomly and there will be no clear pattern.

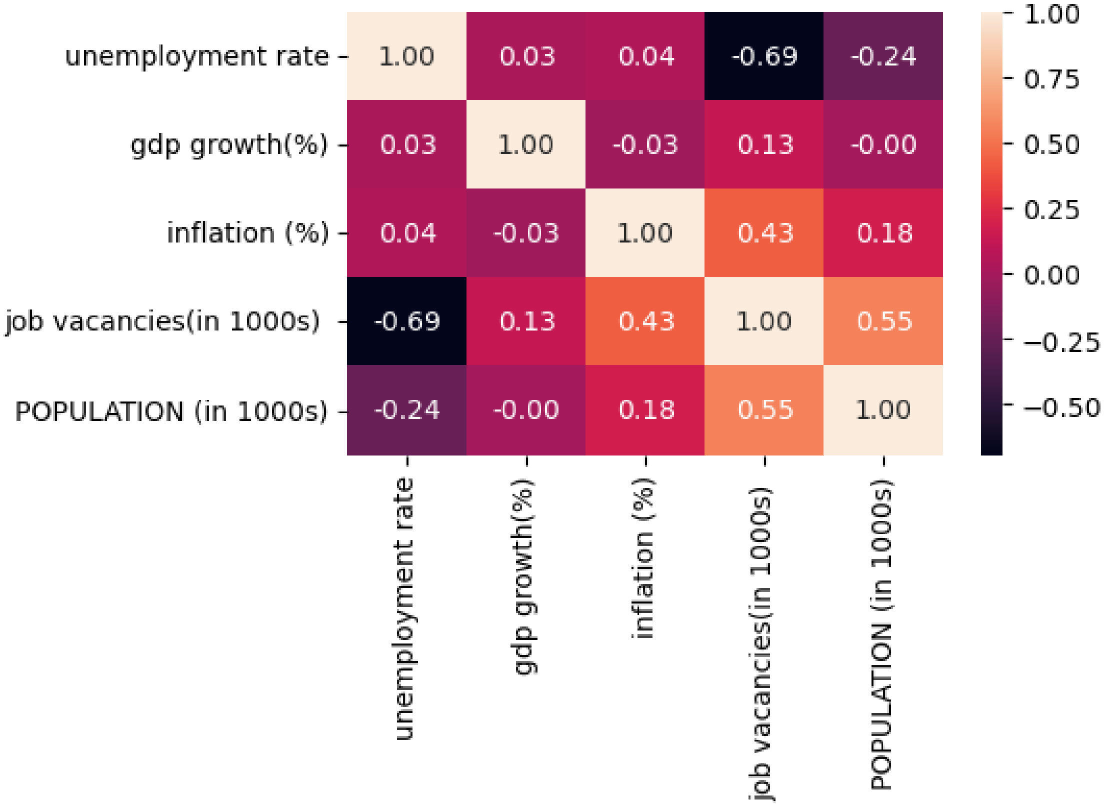


Figure 17 Correlation Graph

So from figure 17, we can see that these correlation relates with our previous analysis which we have done based on the scatter plots. We can see that Job Vacancies and Unemployment have Negative Correlation. There is also positive correlation of 55% between Job Vacancies and Population and 1 thing which is surprised that there is a positive correlation between Job Vacancies and inflation. By seeing the table, it can be seen that most of the column don’t have correlation that is significant.

Algorithm

An algorithm is a set of steps or instructions that are followed in order to solve a problem or complete a task. Algorithms are used in many different areas, including mathematics, computer science, and other fields. In computer science, algorithms are used to perform tasks such as searching, sorting, and operating on data. Algorithms are typically designed to be efficient and to solve problems in a logical and systematic way.

Machine learning

In machine learning, algorithms are trained on a data set, which is a collection of examples that includes input data and the corresponding correct output. For example, a machine learning algorithm might be trained on a data set of images and their associated labels (e.g., "cat", "dog", "car", etc.). The algorithm uses this training data to learn how to map the input data (the images) to the correct output (the labels).

For example, imagine you have a dataset of customer transactions at a store, and you want to use machine learning to predict which customers are likely to purchase a particular item. You could train a machine learning model on this dataset, and then use the model to predict which customers are likely to buy the item when it is next available. By continually updating the model with new data as it becomes available, the model can learn and improve its predictions over time.

Once the algorithm has been trained, it can be used to make predictions on new, unseen data. For example, if the algorithm was trained to recognize images of cats and dogs, it could then be given a new image and predict whether it is a cat or a dog.

There are many different types of machine learning algorithms, including supervised learning algorithms, which learn from labeled training data, and unsupervised learning algorithms, which learn from unlabeled data. There are also semi-supervised and reinforcement learning algorithms, which are variations of the other two types.

Machine learning is used in a wide range of applications, including image and speech recognition, natural language processing, recommendation systems, and predictive analytics. It has the potential to revolutionize many industries by automating tasks and making them more efficient.

Types of Machine learning

There are three main types of machine learning:

Supervised learning: In supervised learning, the algorithm is trained on labeled data, which includes both input data and the corresponding correct output. The algorithm learns to make predictions based on this input-output mapping. For example, a supervised learning algorithm might be trained on a data set of customer data, including information about their age, income, and whether they purchased a particular product. The algorithm would learn to predict whether a new customer is likely to purchase the product based on their age and income.

Unsupervised learning: In unsupervised learning, the algorithm is not given any labeled training data. Instead, it must find patterns and relationships in the data on its own. One common application of unsupervised learning is clustering, in which the algorithm groups similar data points together. For example, an unsupervised learning algorithm might be used to group customers into different segments based on their purchasing habits.

Reinforcement learning: In reinforcement learning, the algorithm learns through trial and error, receiving rewards or penalties for certain actions. The goal is to learn the best action to take in a given situation in order to maximize the reward. Reinforcement learning is often used in control systems and games.

There are also variations on these three types of machine learning, such as semi-supervised learning, in which the algorithm is given some labeled training data and some unlabeled data, and active learning, in which the algorithm can request additional labeled data to improve its performance.

Algorithm in Machine learning

In machine learning, an algorithm is a set of instructions that is used to learn from data and make predictions or take actions. There are many different types of machines learning algorithms, including:

Linear regression: A linear regression algorithm is used for supervised learning tasks, and is used to predict a continuous outcome. It works by fitting a line to the data, and can be used for tasks such as predicting prices based on historical data.

Logistic regression: A logistic regression algorithm is also used for supervised learning tasks, but is used to predict a binary outcome (e.g., 0 or 1). It works by fitting a curve to the data and can be used for tasks such as predicting whether a customer will churn.

Decision trees: A decision tree is a flowchart-like structure in which an internal node represents a feature (e.g., age), and the branches represent the decisions based on that feature (e.g., age < 18 or age >= 18). The leaves of the tree represent the final prediction. Decision trees can be used for tasks such as predicting whether a customer will default on a loan.

Random forests: A random forest is an ensemble learning method that combines multiple decision trees to make a prediction. It works by training multiple decision trees on different subsets of the data, and then averaging the predictions of the individual trees to make a final prediction.

K-means clustering: K-means clustering is an unsupervised learning algorithm that is used to group similar data points together. It works by selecting K centroids, and then assigning each data point to the closest centroid. The algorithm then iteratively updates the centroids until the data points are optimally grouped.

Neural networks: A neural network is a machine learning algorithm that is inspired by the structure of the brain. It consists of multiple layers of interconnect

ted "neurons," which process the input data and produce an output. Neural networks are often used for tasks such as image and speech recognition.

These are just a few examples of the many different types of machine learning algorithms that exist.

Evaluation criteria

Evaluation criteria are the standards or metrics used to measure the performance of a model or system. They are used to determine how well a model is able to solve a problem or complete a task, and can help to identify areas for improvement.

Different types of evaluation criteria are used depending on the nature of the task and the type of model being used. For example, common evaluation criteria for classification tasks (in which the model predicts a class label for an input) include accuracy, precision, and recall. Common evaluation criteria for regression tasks (in which the model predicts a continuous outcome) include mean absolute error, mean squared error, and root mean squared error.

It is important to choose appropriate evaluation criteria for a task, as different criteria may emphasize different aspects of model performance. For example, accuracy is a good evaluation criteria for a model that needs to make a high number of correct predictions, while precision is a better choice for a model that needs to minimize false positives.

Evaluation criteria can also be used to compare the performance of different models on the same task. This can be helpful when trying to determine which model is the best fit for a particular problem.

Accuracy: Accuracy is the number of correct predictions made by the model, divided by the total number of predictions. It is a simple and intuitive metric, but can be misleading if the classes in the data are imbalanced (e.g., there are many more negative examples than positive examples).

Precision: Precision is the number of true positive predictions made by the model, divided by the total number of positive predictions made. It is a measure of the model's ability to avoid false positives.

Recall: Recall is the number of true positive predictions made by the model, divided by the total number of actual positive examples in the data. It is a measure of the model's ability to find all of the positive examples.

R Squared Value: The R-squared value is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, where 0 means that the model explains none of the variance in the dependent variable, and 1 means the model explains all of the variance.

An R-squared value of 1 indicates that the model fits the data perfectly, while a value of 0 indicates that the model does not fit the data at all. Generally, a higher R-squared value indicates a better fit, as it means that more of the variance in the dependent variable is explained by the independent variable(s). However, a high R-squared value does not necessarily mean that the model is a good model, since it can also be the result of overfitting.

R-squared is commonly used in linear regression, where the goal is to predict the value of a continuous dependent variable from one or more independent variables. It is also used in other types of regression analysis, such as logistic regression.

Mean squared error (MSE): Mean Squared Error is a measure of how well a model predicts the target variable. It calculates the average of the squared differences between the predicted values and the actual values. The predicted values are obtained by applying the model to the input data, while the actual values are the known target values.

The MSE is a scalar value, which means it has no direction. It is always positive and the smaller the value, the better the model is at predicting the target variable. In other words, a small MSE indicates that the model's predictions are close to the actual values. The MSE is commonly used to evaluate the performance of a model, and it is also used as a loss function for training models using optimization algorithms such as gradient descent.

For example, if you have a model that predicts housing prices, the MSE would be a measure of how accurate the model's predictions are in comparison to the actual prices. A low MSE would indicate that the model's predictions are close to the actual prices, while a high MSE would indicate that the model's predictions are far from the actual prices.**Root mean squared error (RMSE):** RMSE is the square root of the MSE. It is also a popular choice for regression tasks and is often used to compare the performance of different models.

These are just a few examples of the many evaluation criteria that are used in machine learning and statistical analysis. It is important to choose the appropriate evaluation criteria for a particular task and to consider the trade-offs between different metrics.

Machine learning Algorithms evaluated

Ordinary least square

Ordinary least squares (OLS) is a method used to find the line of best fit for a set of data. It is a common method used in linear regression, which is a statistical technique used to predict a continuous outcome.

In OLS, the line of best fit is found by minimizing the sum of the squared differences between the predicted values and the actual values. This is done by finding the values of the coefficients (a and b in the equation $Y = aX + b$) that minimize this sum.

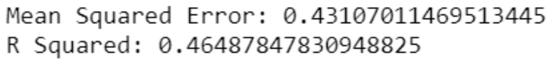
OLS can be used to find the line of best fit for both simple linear regressions, in which there is only one input variable, and multiple linear regression, in which there are multiple input variables. It is a widely used method in machine learning and statistical analysis, as it is relatively easy to implement and can be used with a variety of data types.

In addition to finding the line of best fit, OLS can also be used to make predictions on new data, test hypotheses about the data, and identify relationships between variables.

By applying ordinal least square method, we have summarized following things

- R^2 is computed without centering (uncentered) since the model does not contain a constant.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.
- The condition number is large, 5e+04. This might indicate that there are strong multicollinearity or other numerical problems.
- From the results we can see that our model worked best on the training data. giving 98% r_squared value. which is quite good.

If we look into evaluation measure of OLS as shown in figure 18 it can be seen that both measure mean square error and R_squared error are very low for this algorithm .Both values are below 50% which show that how well algorithm has performed.



```
Mean Squared Error: 0.43107011469513445
R Squared: 0.46487847830948825
```

Figure 18 Evaluation metric of OLS

Linear Regression

Linear regression is a statistical method used for supervised learning tasks. It is used to predict a continuous outcome, such as the price of a stock or the weight of a person.

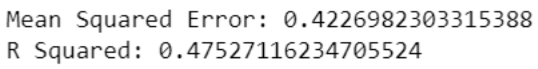
The goal of linear regression is to find the line of best fit that minimizes the difference between the predicted values and the actual values. This line is represented by an equation of the form Y = aX + b, where X is the input data, Y is the predicted output, a is the slope of the line, and b is the y-intercept (the point where the line crosses the y-axis).

To find the line of best fit, the linear regression algorithm tries to minimize the sum of the squared differences between the predicted values and the actual values. This is known as the "least squares" method.

Once the line of best fit has been determined, the linear regression algorithm can be used to make predictions on new data. For example, if the algorithm was trained on data about the relationship between a person's age and their weight, it could be used to predict a person's weight based on their age.

Linear regression can be used for both simple linear regression, in which there is only one input variable, and multiple linear regression, in which there are multiple input variables

From here we can see that we received 68% r_squared on the training data. which is lower than the OLS stats model. Figure 19 shows the error rate of the algorithm i.e. linear regression and that was recorded that is low than the OLS in mean square error which it R_squared error is low. The evaluation is shows in figure 19



```
Mean Squared Error: 0.4226982303315388
R Squared: 0.47527116234705524
```

Figure 19 Evaluation of linear regression

Ridge Regression

Ridge regression is a variant of linear regression, a method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. Ridge regression is similar to linear regression, but it includes an additional term called the "penalty" or "shrinkage term" that helps to prevent overfitting.

The goal of ridge regression is to find the coefficients of the linear equation that minimize the residual sum of squares (RSS) subject to a constraint on the L2 norm of the coefficients. The L2 norm of a vector is the square root of the sum of the squares of the elements in the vector. The constraint on the L2 norm helps to shrink the coefficients towards zero, which can reduce the variance of the model and improve the generalization performance of the model.

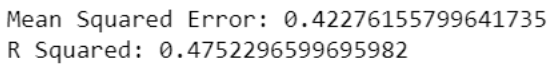
The optimization problem for ridge regression can be written as:

minimize RSS + α * (L2 norm of coefficients)

where α is a hyperparameter that controls the strength of the penalty. A higher value of α results in a stronger penalty, which leads to a greater shrinkage of the coefficients.

Ridge regression can be used to improve the performance of a linear regression model by reducing the variance of the model. It is particularly useful when the number of independent variables is large, or when there is multicollinearity (a high correlation between the independent variables).

Ridge regression has accuracy nearly above 68% almost the same as linear regression. The error shows that as in figure 20 that mean square error is low then previous one while R-Squared error is quite high than previous two (OLS and linear) algorithm.



```
Mean Squared Error: 0.42276155799641735
R Squared: 0.4752296599695982
```

Figure 20 Evaluation of Ridge Regression

Lasso Regression

Lasso regression is a type of linear regression that is used to prevent overfitting. It is a supervised learning algorithm that is used to predict a continuous outcome.

Like linear regression, the goal of lasso regression is to find the line of best fit that minimizes the sum of the squared differences between the predicted values and the actual values. However, lasso regression also includes a penalty term in the objective function that is being minimized. This penalty term, called the "L1 regularization term," shrinks the coefficients of the model towards zero, which can help to reduce overfitting.

The strength of the regularization term is controlled by a hyperparameter called alpha, which determines the amount of shrinkage that is applied to the coefficients. A larger alpha results in greater shrinkage, and a smaller alpha results in less shrinkage.

One key difference between lasso regression and ridge regression (another type of regularization method) is that lasso regression can completely eliminate some features by setting their coefficients to zero. This makes lasso regression useful for feature selection, as it can automatically select the most important features in the data.

Lasso regression is typically used when there are a large number of features in the data and it is necessary to prevent overfitting. It is one of a number of techniques known as "regularization" methods that are used to prevent overfitting in linear models.

Lasso regression is a type of linear regression that uses an L1 regularization term to shrink the coefficients of the model towards zero. The regularization term is controlled by a hyperparameter called alpha, which determines the amount of shrinkage that is applied to the coefficients. A larger alpha results in greater shrinkage, and a smaller alpha results in less shrinkage.

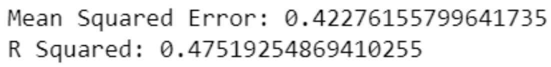
One advantage of lasso regression is that it can automatically perform feature selection by setting the coefficients of some features to zero. This is because the L1 regularization term has a "absolute value" form, which means that it tends to shrink the coefficients of less important features more aggressively than those of more important features. As a result, the coefficients of some less important features may be set to zero, effectively eliminating them from the model.

Lasso regression is often used when there are a large number of features in the data and it is necessary to prevent overfitting. It is one of a number of techniques known as "regularization" methods that are used to prevent overfitting in linear models.

One limitation of lasso regression is that it is sensitive to the scaling of the features. It is generally recommended to scale the features (e.g., by standardizing them) before applying lasso regression.

Lasso regression can be computationally expensive to fit, especially when there are a large number of features in the data. There are several methods that can be used to speed up the fitting process, such as coordinate descent and subgradient descent.

Lasso regression has also accuracy 68% which is similar to both above algorithm (ridge and linear regression).While figure 21 shows the error msg of the lasso algorithm which shows that it has nearly same error rate than above algorithm.

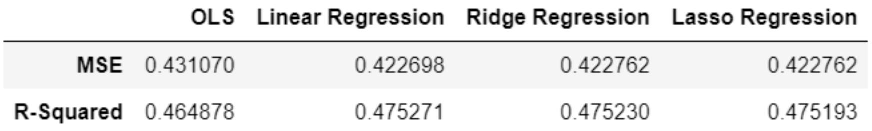


```
Mean Squared Error: 0.42276155799641735
R Squared: 0.47519254869410255
```

Figure 21 Evaluation of Lasso regression

Comparison

In this section, comparison of different algorithm applied is discuss. The table below in figure 22 shows the metric i.e., MSE (mean square error) and R-Squared of different algorithms which shows that OLS has more MSE than other algorithms which shows that it has less accuracy than others. While in ‘R-Squared’ value the OLS has least value compare to other algorithms. Moreover, it can be clearly seen in figure that different in values is not as significant this is due to the small dataset and more importantly due to variable are highly unrelated mean they are not correlated to the each other because of which model are not training good as in other case.



	OLS	Linear Regression	Ridge Regression	Lasso Regression
MSE	0.431070	0.422698	0.422762	0.422762
R-Squared	0.464878	0.475271	0.475230	0.475193

Figure 22 Comparison of different algorithms

In conclusion, machine learning is a powerful tool for making predictions and decisions based on data. It has a wide range of applications, from optimizing business processes and improving healthcare outcomes to building self-driving cars and developing intelligent personal assistants. While machine learning algorithms can be complex, there are many resources available to help you get started, including online courses, textbooks, and open-source libraries. The report is about the evaluation of dataset about economic record about unemployment, inflation, job vacancies and population. Firstly, the analysis is done on the variable of dataset which shows that dataset is very small and its variables aren’t related to each other, they have low correlation between them. Lately, Different model of machine learning and variant of regression are applied on the dataset i.e. OLS, Linear regression, Lasso and ridge are applied on it. Accuracy of all algorithm is nearly same this is due to small dataset and less variable to be evaluated while same case is in error evaluation.

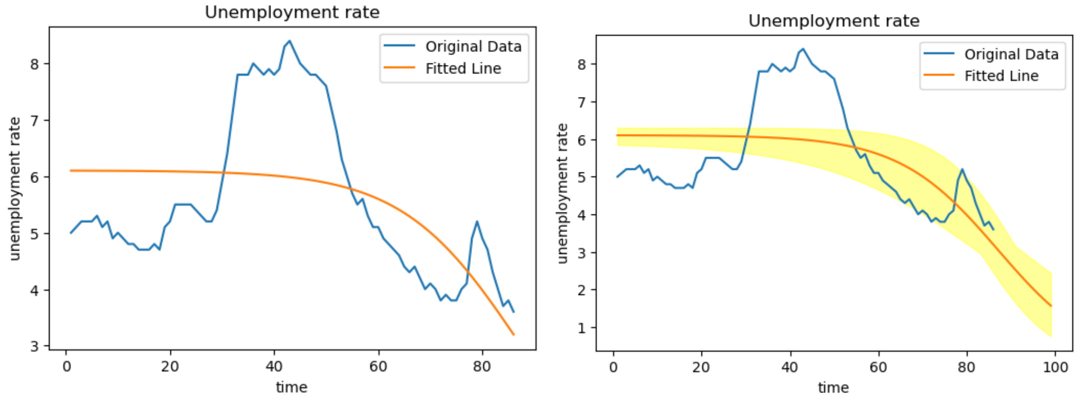


Figure 23 Curve fitting of Unemployment rate using logistic function

Figure 23 shows the curve of logistic function that is used to map the unemployment rate of the economic and shows its results. It can be seen that logistic could map the unemployment accurately in first figure and in second figure the feasibility region with curve is also shown to give maximum and minimum value of the curve and it can be seen that these value are show performance better than first one.