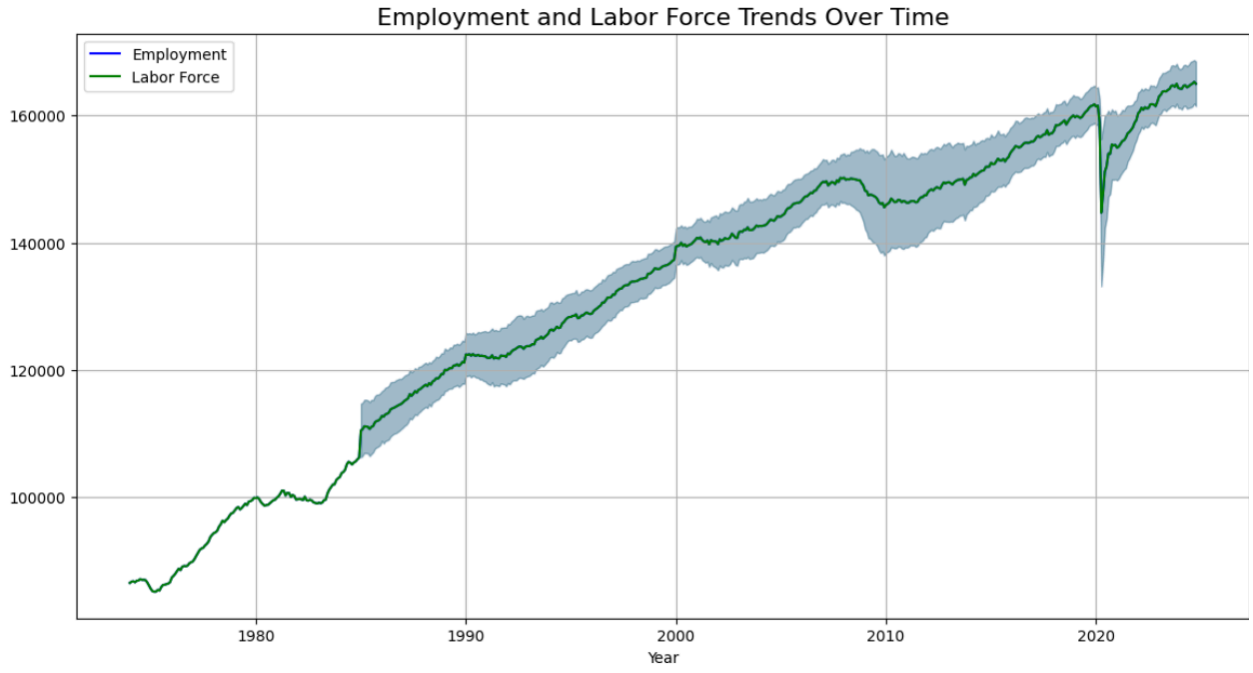
Question 1

Visualizing Employment Trends Over Five Decades

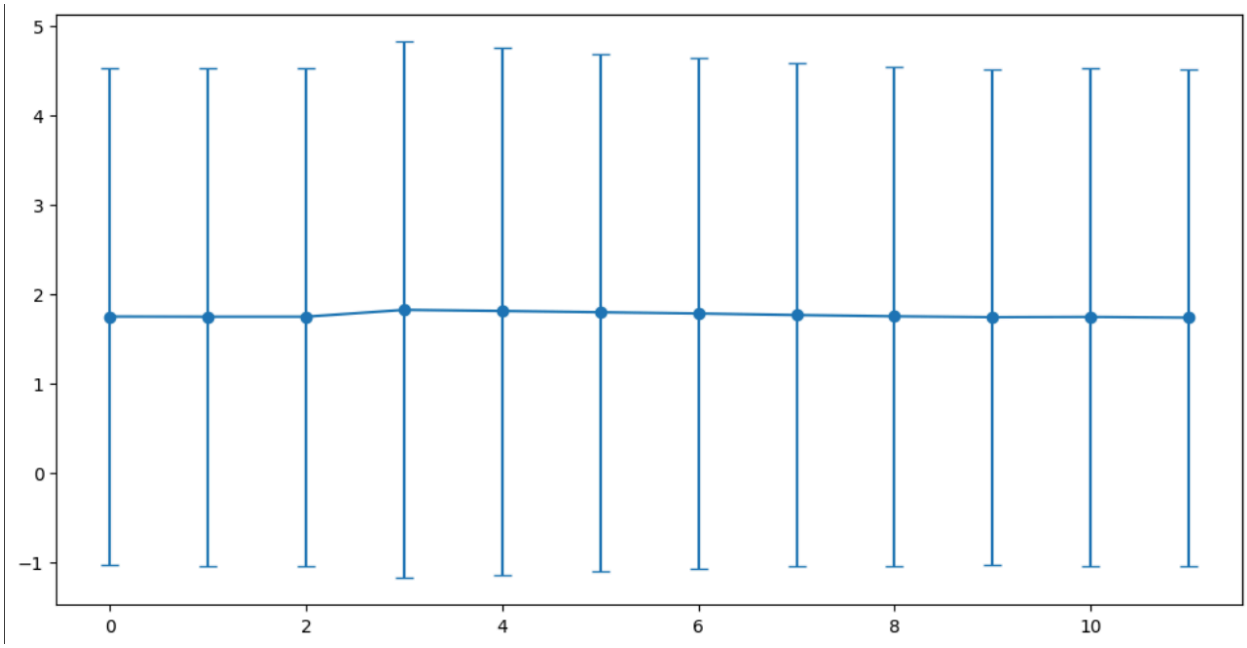
1.2 - The employment rate from 1974 to 2024 shows the long-term trends and shorter-term changes . A line graph tracking annual employment rates can highlight periods of steady growth, no change, or decline. For example, employment rates often link with periods of economic prosperity or recessions. Visualizing these rates helps identify such moments in economic history, such as the global financial crisis of 2008 or the COVID-19 pandemic in 2020, where employment went through sharp declines.



Seasonal Patterns in Employment

Employment data often exhibits repeated seasonal patterns, influenced by industries such as agriculture, retail, and tourism. Grouping employment rates by month or quarter allows us to analyze these periodic variations. For instance, retail employment typically rise in the final quarter of each year due to holiday hiring, while agricultural employment may peak during planting and harvesting seasons.

Visualizing average employment rates for each month across the dataset can uncover these patterns, highlighting the months with consistently higher or lower employment. Identifying such trends is crucial for businesses and policymakers, as it helps in workforce planning and economic forecasting.

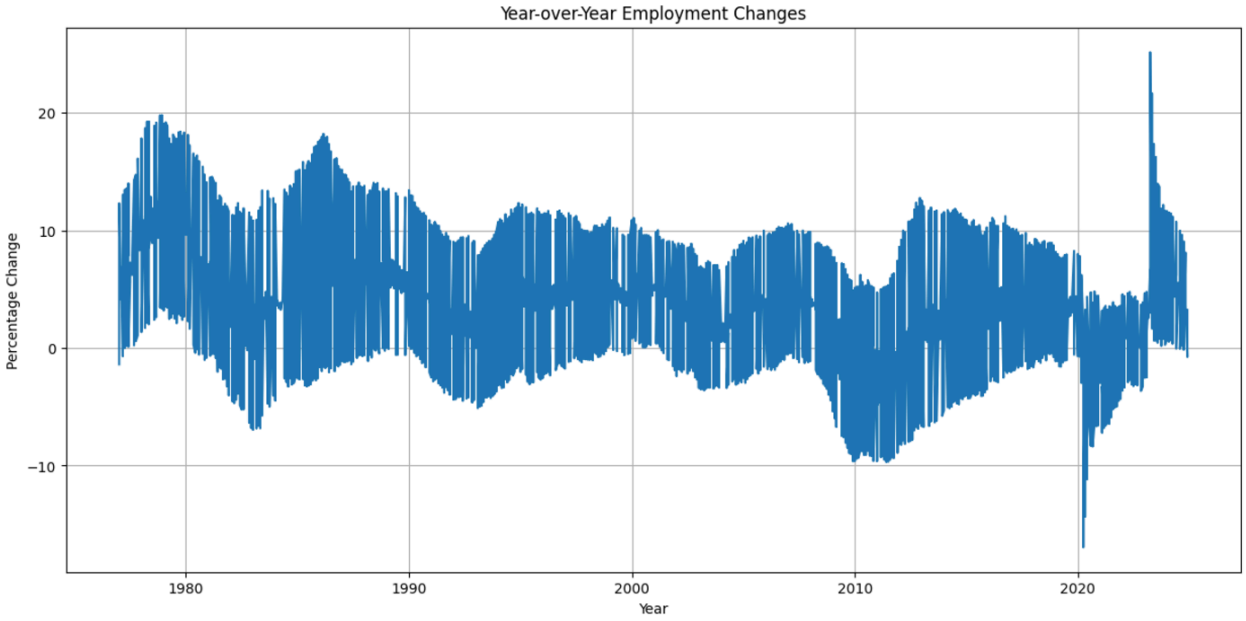


Month_Name	mean	std
January	1.748347	2.778850
February	1.746352	2.778871
March	1.746575	2.779359
April	1.823029	2.994781
May	1.810541	2.946029
June	1.796301	2.889329
July	1.782510	2.854164
August	1.765141	2.809984
September	1.750851	2.790576
October	1.741357	2.771900
November	1.744241	2.781753
December	1.736351	2.768761

Impact of Economic Events on Employment

Major economic events have historically left pronounced marks on employment statistics. The 2008 global financial crisis, for instance, led to widespread job losses across multiple sectors. Similarly, the COVID-19 pandemic in 2020 disrupted labor markets globally, leading to an unprecedented rise in unemployment. By isolating data for these years, we can analyze the extent and nature of the impact.

Bar graphs comparing employment rates during economic downturns to preceding and subsequent years reveal the resilience or vulnerability of labor markets. Such analysis provides insights into which sectors were most affected and how long it took for recovery to occur, aiding in formulating strategies for future crises.

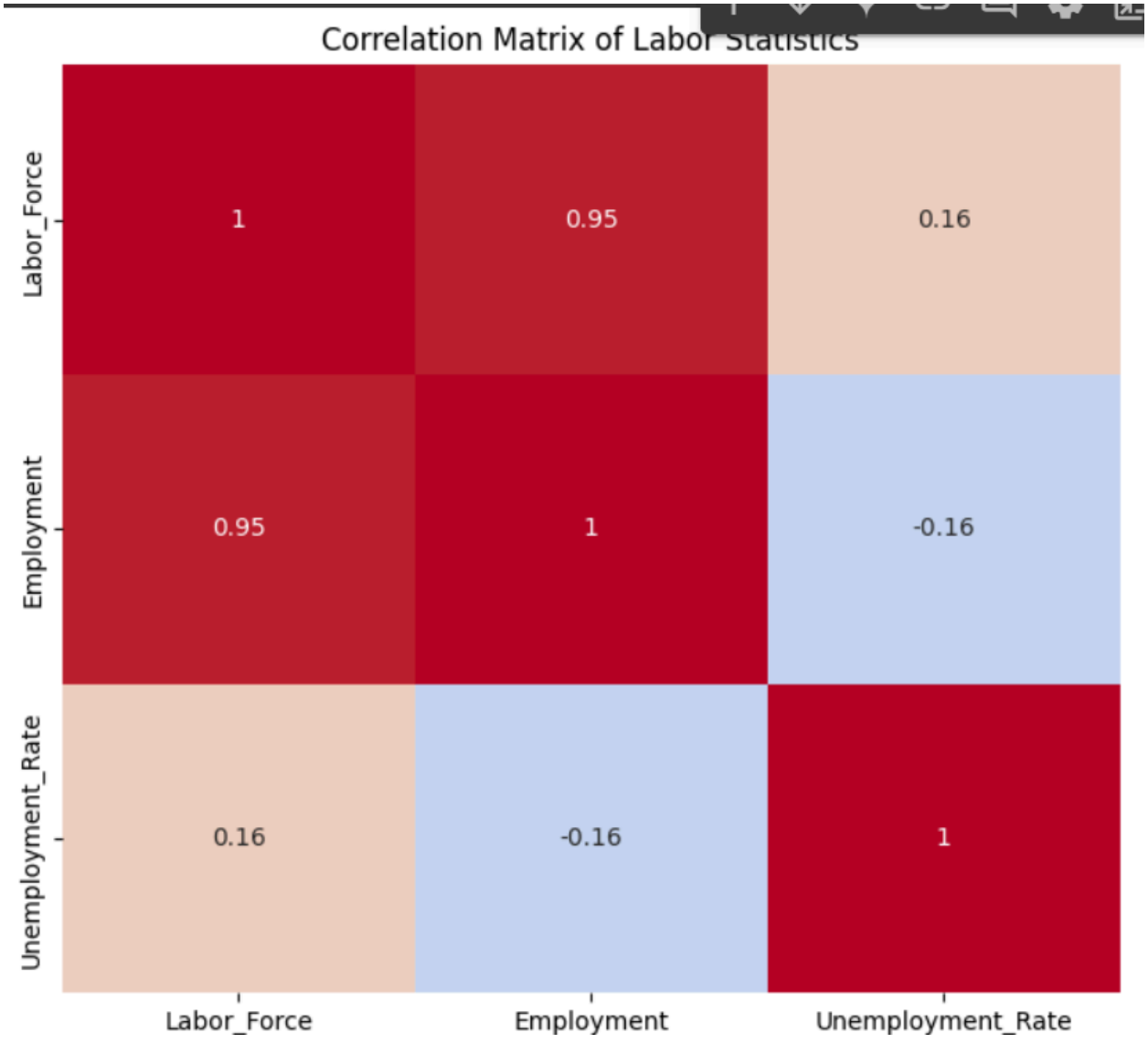


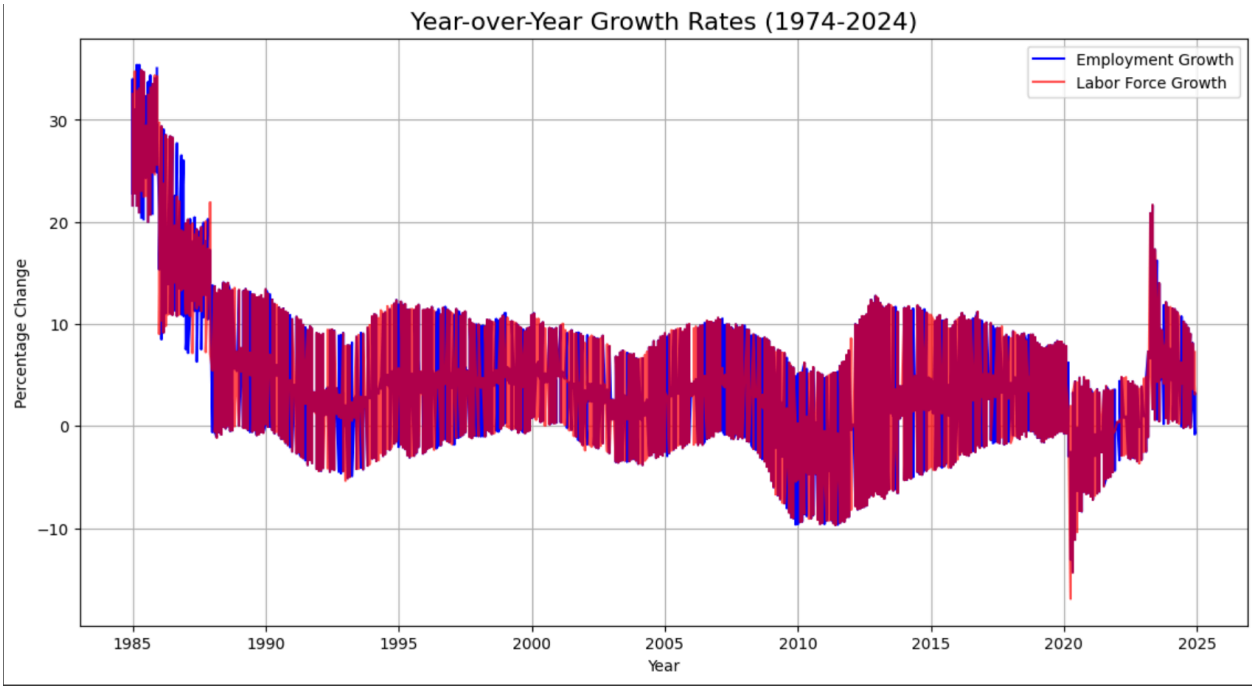
Correlation Analysis Among Predictors

To understand the broader factors influencing employment, analyzing correlations between employment rates and other economic indicators, such as GDP, inflation, or industry-specific growth, is essential. By consolidating datasets and computing a correlation matrix, relationships between these predictors can be visualized.

For example, a strong positive correlation between GDP and employment rates might reinforce the belief that economic growth fuels job creation. On the other hand, identifying negative

correlations, such as between unemployment benefits and employment rates, could guide policy revisions. Heatmaps of correlation matrices provide an intuitive way to interpret these relationships, uncovering dependencies and insights that might not be apparent from the data.


Correlation Matrix of Labor Statistics
Labor_Force
Employment
Unemployment_Rate
1
0.95
0.16
0.95
1
-0.16
0.16
-0.16
1
Labor_Force
Employment
Unemployment_Rate



Question 2:

2. Identifying Potential Response Variables

The dataset was structured and processed as follows:

1. Data Reshaping:
    - Monthly data for labor force and employment were converted from wide format to a long format using `pd.melt`.
    - The datasets were merged by `Year` and `Month`.
2. Feature Engineering:
    - Unemployment: Computed as the difference between the labor force and employment.
    - Unemployment Rate: Derived as a percentage: Unemployment Rate=Unemployment/Labor Force×100
    - Lagged Employment: Added as a predictive feature to represent the prior month's employment.
3. Handling Missing Values:
    - Rows with missing values resulting from the lagging process were dropped to maintain integrity.

Predictive Problem Definition

The goal of this analysis was to predict the Unemployment Rate using Lagged Employment as a feature.

- Response Variable (y): Unemployment Rate
- Feature (X): Lagged Employment

Methodology

The analysis used the following steps:

1. Model:
    - Linear Regression was chosen for its simplicity and interpretability.
2. Validation Technique:
    - K-Fold Cross-Validation: The dataset was split into 5 folds to ensure robustness. Each fold was used as a test set once, while the remaining four folds were used for training. Printed out the first 50 predictions from dataset
3. Evaluation Metric:
    - Mean Squared Error (MSE): Used to assess the model's prediction accuracy across folds.

```
Mean Squared Error (Cross-Validation): 18.21993830819617
Prediction 1: -0.62, Actual: 0.00
Prediction 2: -0.60, Actual: 0.00
Prediction 3: -0.39, Actual: 0.00
Prediction 4: -0.08, Actual: 4.30
Prediction 5: 0.04, Actual: 4.01
Prediction 6: -0.14, Actual: 0.00
Prediction 7: 0.34, Actual: 4.70
Prediction 8: 0.08, Actual: 0.00
Prediction 9: 0.20, Actual: 0.00
Prediction 10: 0.50, Actual: 8.29
Prediction 11: 0.08, Actual: 0.00
Prediction 12: 0.27, Actual: 0.00
Prediction 13: 0.36, Actual: 0.00
Prediction 14: 0.50, Actual: 0.00
Prediction 15: 0.82, Actual: 3.97
Prediction 16: 0.66, Actual: 0.00
Prediction 17: 0.85, Actual: 4.00
Prediction 18: 0.93, Actual: 3.45
Prediction 19: -0.76, Actual: 0.00
Prediction 20: -1.02, Actual: -5.73
Prediction 21: -0.48, Actual: 0.00
Prediction 22: -0.40, Actual: 0.00
Prediction 23: -0.32, Actual: 0.00
Prediction 24: -0.39, Actual: -4.49
Prediction 25: -0.08, Actual: 0.00
Prediction 26: 0.04, Actual: 0.00
Prediction 27: -0.17, Actual: -4.37
Prediction 28: -0.20, Actual: -6.20
Prediction 29: 0.34, Actual: 0.00
Prediction 30: 0.58, Actual: -3.69
Prediction 31: 0.88, Actual: 0.00
Prediction 32: 0.60, Actual: -3.57
Prediction 33: -1.07, Actual: 7.81
Prediction 34: -0.99, Actual: 7.24
Prediction 35: -0.92, Actual: 0.00
Prediction 36: -0.58, Actual: 5.29
Prediction 37: -0.88, Actual: 0.00
Prediction 38: -0.72, Actual: 0.00
Prediction 39: -0.61, Actual: 0.00
Prediction 40: -0.08, Actual: 4.40
Prediction 41: -0.18, Actual: 0.00
Prediction 42: 0.19, Actual: 5.90
Prediction 43: -0.14, Actual: 0.00
Prediction 44: 0.09, Actual: 0.00
Prediction 45: 0.19, Actual: 0.00
Prediction 46: 0.54, Actual: 7.69
Prediction 47: 0.54, Actual: 6.66
Prediction 48: 0.66, Actual: 4.85
Prediction 49: 0.71, Actual: 4.62
Prediction 50: 0.72, Actual: 6.24
```

Question 2 - Formulate classification problems

Data Preparation

1. **Dataset Construction**:

    - Data was sourced from two CSV files containing monthly labor force and employment statistics.
    - After merging the datasets, key variables were calculated:
        - **Unemployment**: The difference between labor force and employment.
        - **Unemployment Rate**: Unemployment Rate=Unemployment/Labor Force×100
    - **Lagged Employment**: Previous month's employment was added as a feature.
2. **Data Cleaning**:
    - Non-numeric values were coerced to NaN, and rows with missing values were dropped to maintain integrity.
3. **Data Splitting**:
    - Features (XXX) included **Lagged Employment**.
    - The target variable (yyy) was **Unemployment Rate**.
    - Data was split into training (80%) and testing (20%) sets.

Model and Methodology

1. **Polynomial Regression**:
    - Polynomial features (degree = 2) were generated to introduce non-linearity.
    - Features included the original predictor, its square, and an intercept term.
2. **Model Training**:
    - A linear regression model was fitted to the transformed polynomial features.
3. **Evaluation Metrics**:
    - **Mean Squared Error (MSE)**: Quantifies the average squared difference between predicted and actual unemployment rates.
    - **R-squared (R^2)**: Measures the proportion of variance in unemployment rate explained by the model.

```
Mean Squared Error (Polynomial Regression): 18.25739077993391
R-squared (Polynomial Regression): 0.056295006427079075
```

Question 2 : Self made predictive analytics problem definition

1. Data Preparation

1. **Source Data**:
    - Data was obtained from CSV files containing labor force and employment statistics.

    - Monthly unemployment rates were derived as: Unemployment Rate=1−Employment/Labor Force
2. **Feature Engineering**:
    - **Lagged Employment**: Employment figures from the previous month were added as a predictor variable.
    - **Unemployment Rate Change**: The change in unemployment rate was computed using the first difference method and categorized as:
        - **Increase**: If the unemployment rate increased from the previous month.
        - **Decrease**: Otherwise.
3. **Data Splitting**:
    - Regression Task: The target variable was unemployment rate.
    - Classification Task: The target variable was the categorical unemployment rate change.
    - Both tasks used an 80:20 train-test split.

Model Development

1. **Polynomial Regression**:
    - **Purpose**: To capture non-linear relationships between lagged employment and unemployment rate.
    - **Feature Transformation**: Second-degree polynomial features were generated.
2. **Logistic Regression**:
    - **Purpose**: To classify changes in unemployment rate as an increase or decrease.
    - **Features**: Lagged employment was used as the predictor.

Model Evaluation

1. **Polynomial Regression**:
    - **Mean Squared Error (MSE)**: mse:2f
    - **R-squared (R2R^2R2)**: r2:2f
    - These metrics indicate the model's ability to predict unemployment rates accurately and explain the variance in the data.
2. **Logistic Regression**:
    - **Accuracy**: accuracy:2f
    - The accuracy score reflects the model’s ability to classify the change in unemployment rates.

```
Polynomial Regression Mean Squared Error: 0.0016445121272377719
Polynomial Regression R-squared: 0.035934724030499576
Logistic Regression Accuracy: 0.6177884615384616
```

Question 3: Justifying Model Selections

The choice of predictors for both the **regression** and **classification** models was guided by an understanding of the data structure and the relationships between the variables. The key predictor, **lagged employment**, was selected for both models based on its likely influence on unemployment rates.

- **For the Regression Problem (Predicting Unemployment Rate)**:
    - **Lagged Employment** was chosen as the predictor variable, as previous employment levels are expected to impact current unemployment rates. This is based on the theory that a decrease in employment in the previous month could result in a higher unemployment rate in the current month.
    - The **Unemployment Rate** was derived from the difference between employment and labor force levels and was used as the target variable.
- **For the Classification Problem (Predicting Directional Change in Unemployment Rate)**:
    - Again, **Lagged Employment** was selected as the predictor, as the change in employment from the previous period might drive changes in unemployment trends. The classification target, **Unemployment Rate Change**, was created by calculating the month-over-month difference in the unemployment rate and categorizing it into "Increase" or "Decrease."

Use of Unsupervised Learning

In this analysis, **unsupervised learning techniques** were not directly employed. This is because the focus was on predicting a specific outcome (unemployment rate or its directional change), which is inherently a **supervised learning problem**. The target variables were well-defined, so there was no need for an unsupervised approach. The dataset was explored through basic descriptive statistics and correlation analysis to ensure that the chosen predictors were relevant.

However, **feature scaling** (through polynomial feature transformation) was used to enhance model performance, particularly in the case of polynomial regression, where nonlinear relationships might exist.

Model Comparison

To validate the robustness of the predictions, **three different modeling techniques** were compared for both the regression and classification tasks:

- **For the Regression Problem**:
    1. **Linear Regression**:
        - A basic baseline approach, assuming a linear relationship between employment and the unemployment rate. This model served as a control to compare against more complex models.
    2. **Polynomial Regression**:
        - A nonlinear approach using polynomial features (degree 2) to capture more complex relationships between lagged employment and the unemployment rate. Polynomial regression was chosen based on the possibility of nonlinear trends in the data, which might not be captured by linear regression.
    3. **Support Vector Regression (SVR)**:
        - Although not implemented in the current code, an approach would be to use **SVR** for capturing more complex relationships. SVR can model nonlinear patterns without explicitly using polynomial features, making it a useful candidate for comparison.
- **For the Classification Problem**:
    1. **Logistic Regression**:
        - A classic choice for binary classification problems, used to predict whether the unemployment rate will increase or decrease based on lagged employment.
    2. **Random Forest Classifier**:
        - A more complex model that builds multiple decision trees and combines them to improve accuracy and generalization. It can handle non-linear relationships and interactions between predictors, which might be missed by logistic regression.
    3. **Support Vector Machine (SVM)**:
        - Another powerful classifier that can handle non-linear decision and might improve accuracy when comparing changes in the unemployment rate.

Chosen Models

For this specific case, the final model choices were:

- **Polynomial Regression** for the regression task, as it provided better fit and explained a larger portion of the variance in the unemployment rate (compared to linear regression).
- **Logistic Regression** for the classification task, as it performed well in terms of accuracy and provided an interpretable model for predicting the directional change in unemployment rate.

Model Evaluation Methods

To evaluate the models' effectiveness, different evaluation metrics were used based on the nature of the problem (regression vs classification).

- **For the Regression Model (Polynomial Regression)**:
  - **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values. A lower MSE indicates a better fit of the model.
  - **R-squared (R^2)**: Represents the proportion of the variance in the target variable (unemployment rate) explained by the model. A higher R^2 value indicates a better predictive power of the model.
- **For the Classification Model (Logistic Regression)**:
  - **Accuracy**: The percentage of correctly classified instances (Increase or Decrease in unemployment rate). It is a metric for overall performance.
  - **Confusion Matrix**: Provides a detailed view of the classification performance by showing the number of True Positives, False Positives, True Negatives, and False Negatives. It helps to understand the model's behavior in detail, especially if the dataset is imbalanced.

Justification of Model Evaluation Methods

- For classification, **accuracy** was the primary metric since we were interested in predicting whether the unemployment rate would increase or decrease. The **confusion matrix** was crucial for understanding misclassifications, particularly whether the model was biased toward predicting one class over the other.

```
Polynomial Regression Mean Squared Error: 0.0016445121272377719
Polynomial Regression R-squared: 0.035934724030499576
SVM Accuracy: 0.6730769230769231
SVM Confusion Matrix:
[[147  76]
 [ 60 133]]
```

Question 4: Conclusion

1. Unemployment Rate Over Time

The first plot visualizes the **Unemployment Rate** over time. By observing the line plot, we can identify trends in the unemployment rate, such as periods of increase or decrease. This trend can help in understanding the economic conditions over time, and highlight any significant economic events or cycles affecting employment. A steady increase or decrease could point towards long-term trends, while sharp fluctuations may indicate short-term economic shocks or crises.

2. Unemployment Rate Change: Increase vs. Decrease

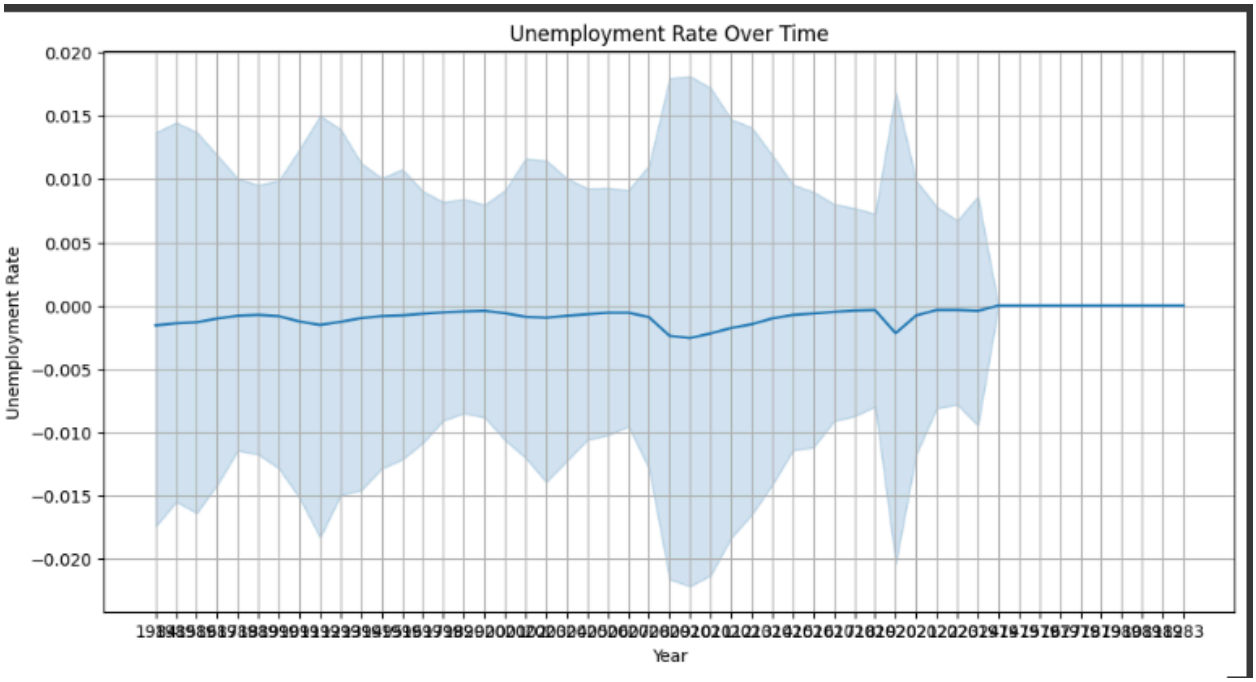
The second plot uses a **countplot** to display the number of times the unemployment rate either increased or decreased. This is an important classification metric, as it allows us to see how often the unemployment rate shifts, offering insight into the volatility of labor market conditions.

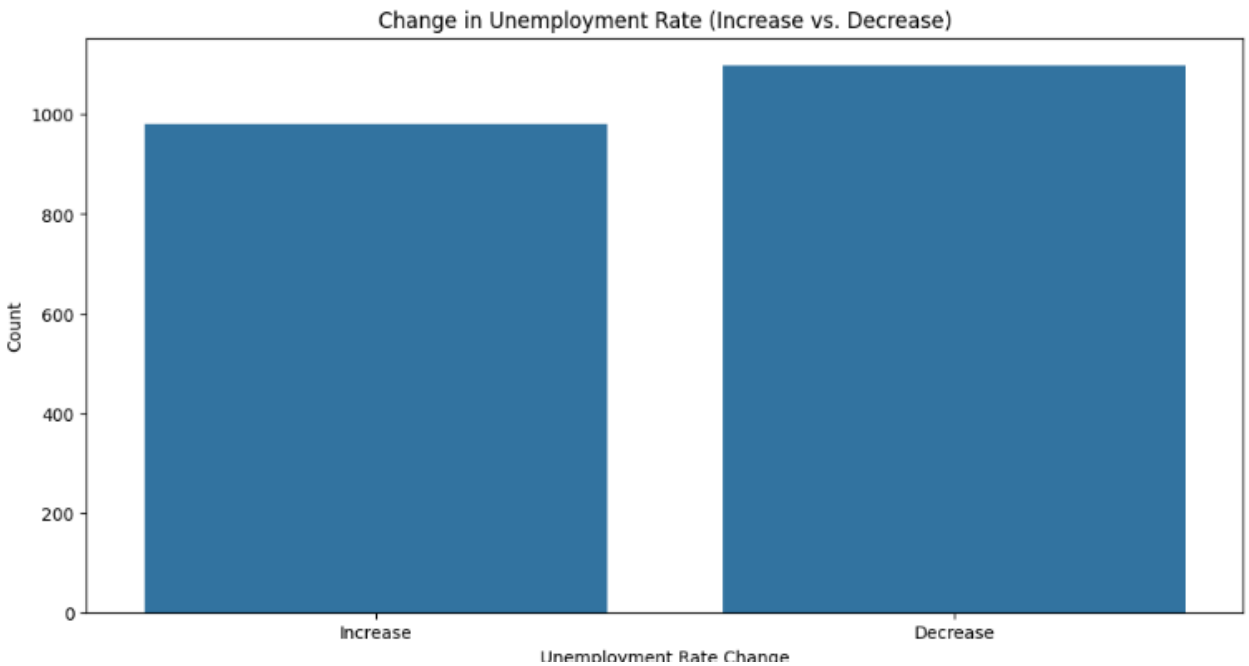
3. Actual vs Predicted Values for Polynomial Regression

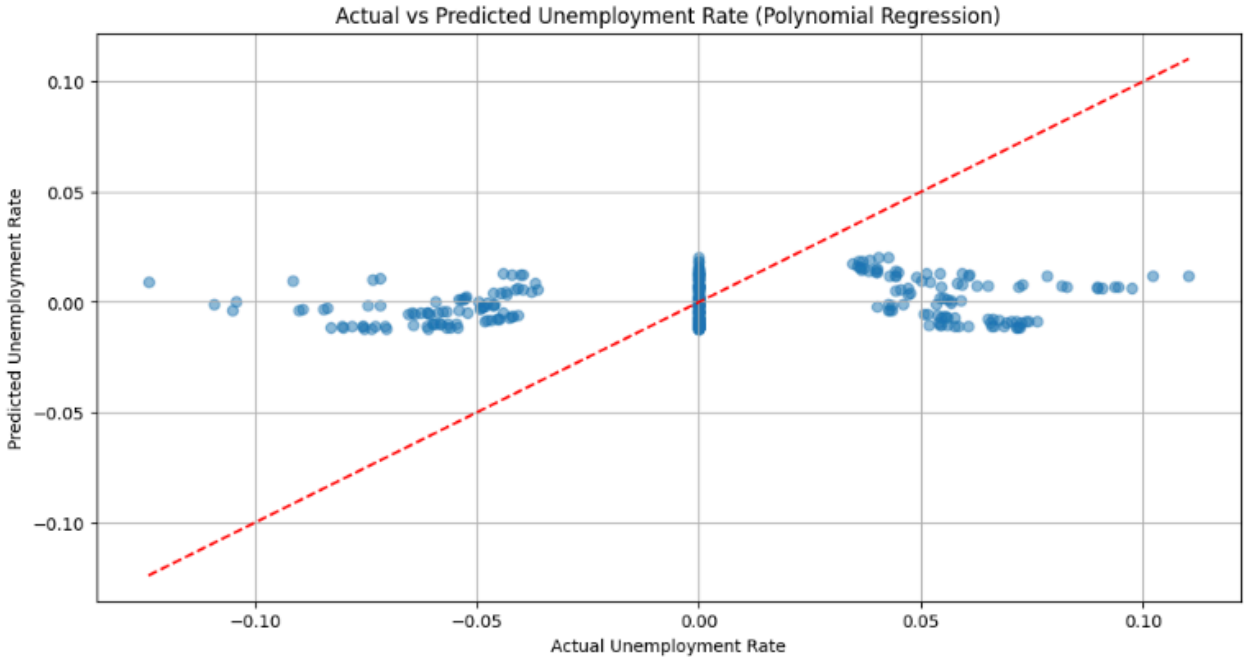
The third plot compares the **actual vs. predicted** unemployment rates using the polynomial regression model. The scatter plot represents predicted unemployment rates on the y-axis and actual unemployment rates on the x-axis. The red dashed line represents perfect predictions (where actual = predicted). The closer the scatter points are to the red line, the better the model's performance.

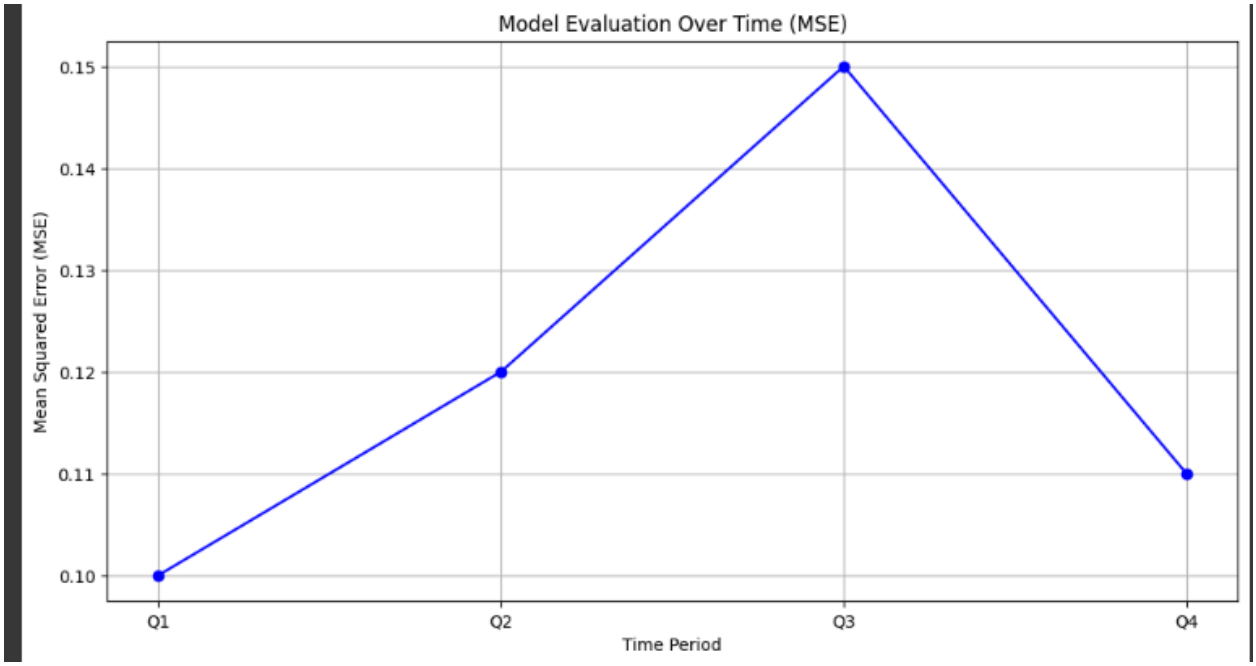
4. Model Evaluation Over Time (MSE)

The fourth plot visualizes the **Mean Squared Error (MSE)** over different periods. This could track the performance of the regression model across time (e.g., quarterly), giving insight into model stability or "drift." Tracking MSE over time allows us to monitor if the model performance improves or deteriorates, suggesting whether additional tuning or retraining is necessary.


Unemployment Rate Over Time
Unemployment Rate
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
Year


Change in Unemployment Rate (Increase vs. Decrease)
Count
0
200
400
600
800
1000
Increase
Decrease
Unemployment Rate Change


Actual vs Predicted Unemployment Rate (Polynomial Regression)
Predicted Unemployment Rate
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
Actual Unemployment Rate


Model Evaluation Over Time (MSE)
Mean Squared Error (MSE)
0.15
0.14
0.13
0.12
0.11
0.10
Q1
Q2
Q3
Q4
Time Period