Analysing and Forecasting Boston's Economic Indicators

Co-Authors:
Sakthi Swarup Vasanadu Kasi- 02085794

Shalabh Singh Yadav – 02130778

Teja Naga- 02122966

The Issues:

In Boston, the study of key economic indicators like hotel occupancy rates and airport passenger numbers is highly significant, directly influencing the city's economic vitality and growth. This project aims to explore these factors, seeking to understand the complex dynamics of Boston's economic landscape. By applying advanced analytical techniques such as SARIMA and clustering methods, we intend to uncover the hidden patterns and connections within these indicators. The aim is to achieve a comprehensive understanding of Boston’s economic rhythm, crucial for informed decision-making and strategic planning. Our analysis is more than a mere examination of statistics; it's an attempt to offer practical insights that can promote economic development, enhance tourism, and improve transportation systems in Boston. Ultimately, our goal is to contribute to creating a strong, flourishing economic environment that benefits both the city and its inhabitants.

We address the following questions:

1. How has the unemployment rate in Boston changed from January 2013 to December 2019, and what is the correlation between the unemployment rate and the median housing prices during this period?
2. Is there a significant correlation between the number of international flights at Logan and the hotel occupancy rate in Boston?
3. What anomalies exist in the data, and what might they indicate?
4. Can a linear regression model accurately predict median housing prices based on a combination of economic indicators such as total jobs, unemployment rate, hotel occupancy rates, and international flights?
5. Can cluster analysis techniques be used to identify distinct periods or phases in Boston’s economic development based on the similarity of economic indicator profiles across different months and years.
6. How do the seasonal patterns observed in hotel occupancy rates and Logan Airport passenger numbers align with known tourism and travel trends in Boston, and what implications might these patterns have for the city's tourism and transportation planning?
7. Considering the SARIMA model's forecast for hotel occupancy rates, how should hotels and local businesses in Boston prepare for expected seasonal fluctuations and long-term trends in tourism demand?
8. Based on the SARIMA model's prediction for Logan Airport passenger traffic, what strategies should airport management and airlines consider to efficiently handle the forecasted seasonal peaks and troughs in passenger flow?
9. How can the SARIMA model's forecast of hotel prices for the year 2022 inform the strategic pricing decisions and promotional planning of hotels in Boston, especially in anticipating and adapting to market trends and seasonal demand fluctuations?
10. How can SARIMA model's accuracy rate in predicting hotel prices be leveraged or redefined to enhance predictive analytics?

Findings:

This project thoroughly investigated Boston's economic factors, focusing on hotel occupancy, airport traffic, and the housing market. We uncovered a significant link between tourism and air travel, evident in the parallel trends of hotel occupancy and Logan Airport passenger numbers, with a correlation coefficient of about 0.528. This indicates a strong interdependence between these sectors.

Through cluster analysis, we identified distinct economic periods in Boston's history, marked by different levels of activity in tourism, employment, and development. This analysis revealed times of high economic activity with increased hotel occupancy and low unemployment, and other times when the economy was slower, affecting hotel rates and occupancy.

A critical aspect of our analysis was using the SARIMA model to predict hotel prices for 2022. Our prediction of an average hotel rate of $272.96 was very close to the actual rate of $262, demonstrating a high accuracy of about 95.8%. This highlights the model's effectiveness in tracking market trends.

In parallel, we explored the housing market with linear regression analysis, successfully modelling median housing prices and explaining about 77.9% of their variation. Notably, we found a strong correlation of 0.829 between the unemployment rate and median home prices, indicating a significant impact of job market conditions on housing prices.

Overall, our project blended detailed analysis with accessible insights, offering valuable information to various stakeholders. By highlighting the connections between different economic sectors, our findings provide a solid basis for informed policy-making and strategic planning in Boston, contributing to a more robust and dynamic economic environment.

Discussion:

The project's in-depth analysis of Boston's economic indicators reveals several critical implications, showcasing the nuanced connections between different sectors within the city's economy. The analysis highlights a unique correlation between hotel occupancy rates and Logan Airport passenger numbers, pointing to an intimately connected relationship between the tourism and air travel industries. Such findings emphasize the need for collaborative strategies and policies that consider the interconnected nature of these sectors.

Further, the identification of different economic phases through cluster analysis offers valuable insights into Boston's economic trends and cycles. This information is vital for policymakers and business leaders, providing a roadmap for targeted economic strategies and interventions. By understanding the specific characteristics of each phase, decision-makers can better align their actions with the city's economic rhythm, enhancing the effectiveness of their endeavours. The utilization of the SARIMA model to accurately forecast hotel prices up to 2022 is another significant aspect of the analysis. With an accuracy rate of approximately 95.8%, the model proves to be an invaluable tool for predicting market trends, providing a strong basis for strategic planning in the hotel industry. This predictive capability can assist hotels and related businesses in anticipating market demand, optimizing pricing strategies, and making informed investment decisions.

Moreover, our exploration of the housing market through linear regression analysis clarifies the intricate connection between housing prices and general economic circumstances. This insight is vital for addressing housing market challenges, including affordability and stability, and underscores the importance of considering macroeconomic factors in real estate planning and policy formulation.

With every aspect considered, the analysis provides a thorough and varied perspective on Boston's economic environment. The conclusions drawn from this analysis have significant ramifications for sector-specific tactics, urban planning, and economic policy. They emphasize the value of making decisions based on data and the necessity of an all-encompassing strategy for economic development that recognizes and capitalizes on the interdependencies of different industries to promote sustainable growth. The results are a vital tool for interested parties, providing direction for the creation of strong, fair, and progressive economic policies for the city of Boston.

Appendix A: -

The goal of this project is to understand and analyse Boston's economy, using data from January 2013 to December 2019, to make better decisions and plans. The project seeks to boost Boston's economy by analysing trends and patterns using the comprehensive economic indicators dataset. We employed a specific methodology to study this dataset, extracting insights vital for strategic growth initiatives as discussed below.

Data Collection and Initial Setup

The project began with the collection of the "economic_indicator.csv" dataset from https://data.boston.gov, which contained comprehensive data on the state of the Boston economy from January 2013 to December 2019. The dataset offers comprehensive information by incorporating total employment, rates of unemployment, rates of labour force participation, average daily rates, rates of hotel occupancy, numbers of international flights, and passenger counts at Logan Airport. The data was systematically imported into a Python environment using the panda's library. This was a necessary step for effective data processing and additional analysis.

Data Cleaning and Pre-processing

To make the dataset suitable for analysis, we thoroughly cleaned it. To improve our understanding of the dates, we first formatted them uniformly in the same way. Following that, we searched for any inaccurate or missing information. To maintain the integrity and accuracy of our analysis, there were moments when we had to eliminate data that was incorrect or missing. In addition, we verified that all the data was in the appropriate format by double-checking that text and numbers had the proper labels. To make sure everything was fair and comparable for our analysis, we finally balanced and modified the data. A panda's tool, which is useful for identifying and repairing missing or incomplete data, was heavily utilized in this work. This guaranteed the quality of our analysis and the accuracy of the data.

Exploratory Data Analysis (EDA)

To prepare for both quantitative and qualitative analyses, we carefully imported the necessary Python libraries before starting our data analysis journey. We utilized several statistical techniques to obtain a more profound comprehension of the patterns present in our dataset. Matplotlib and Seaborn were used to construct a variety of bar plots and other graphical representations, such as graphs showing unemployment rates and passenger counts at Logan Airport. The visuals provided insights that could have been missed in numerical examination, and they were not only instructive but also essential in identifying underlying causes and irregularities.

The primary goal of our initial evaluation was to comprehend the dataset's fundamental structure and organization. To get a summary of our data, we used Panda's operations like '.head()' and '.columns.' These features played a pivotal role in moulding our approach for analysis. We utilized the '.describe ()' function to get a more in-depth statistical description of the dataset, exposing parameters such as central trends, distribution, and data dispersion.

We used more Python commands in addition to these to investigate our dataset further. To ensure proper handling and analysis of the data, we utilized. D-types to verify the data types of each column. We created box plots to uncover possible outliers, which was an essential step in making sure our results were accurate and our conclusions were strong. These diverse methodologies and tools collectively allowed us to conduct a thorough and insightful analysis of the dataset, facilitating a nuanced understanding of Boston's economic dataset.

Advanced Statistical Techniques

After that, we used more advanced methods to analyse the dataset, combining fundamental statistics with computations of data skewness to fully understand the findings. We used Seaborn, a Python tool, to create graphs that illustrate the trends we found in significant segments of the data and carefully examined them. We used Python to create a heatmap that displays correlations based on certain segments of the data. Seeing the connections between various aspects of the economy is now simpler thanks to this heatmap.

For these analyses, we used simple Python commands. For example, we used `.skew() ` to see how much the data leans to one side. To look at trends, we used `seaborn.lineplot() ` for making line graphs. And for the heatmap, we used `pandas. DataFrame.corr() ` to figure out the relationships between different data parts, and then `seaborn. Heatmap()` to turn these relationships into a visual map. These methods helped us get a better and clearer picture of Boston's economy.

Machine Learning Algorithms

We used a systematic method while choosing our machine learning start's algorithms and training models. According to normal machine learning process, the dataset was split into training and testing sets to evaluate the models against unseen data.

- Training and Testing Split: To split the dataset into training and testing subgroups, we utilized scikit-learn's train_test_split method. This made it possible for us to test the models' performance on a different set of data after training them on a portion of the original set, guaranteeing the models' durability.

Key Machine Learning algorithms used in our analysis are as follows: -

1. Linear Regression Model: - Regression analysis was used to determine how different economic indicators and Boston's median home prices related to one another. Using the 'train_test_split' function from ‘sklearn.model_selection,’ we separated our dataset into training and testing sets, allocating 70% of the data for training and 30% for testing. Next, we used ‘sklearn.linear_model.LinearRegression’ to generate a Linear Regression model. After that, we used the 'fit' technique to train our model using the

training dataset. R-squared (R^2) value, which can be computed using "sklearn.metrics. r2_score," was used as a primary metric to assess the performance of our model.

2. K-means Clustering: - To identify trends in our dataset, we employed a clustering technique. To ensure that everything was on the same scale, we first normalized the data using the 'StandardScaler'. Following that, we determined the optimal number of clusters using the Elbow Method. We performed this by using K-means to compute the Within-cluster Sum of Squares (WCSS) for various cluster numbers (ranging from 1 to 10).

After that, we used 'sklearn.cluster.KMeans' to do K-means clustering with three and four clusters. Data points were allocated to clusters using the `fit predict` technique. We applied the cluster information back to our original data to determine the significance of these clusters. Finally, two sets emerged, one with three clusters and the other with four. By applying the 'group by' approach, we were also able to get the average values for each economic indicator in each cluster. This gave us more insight into the usual traits of each cluster and the economics of Boston.

Time Series Analysis

The "Hotel Occupancy Rate" and "Logan Passengers," two significant economic indicators, were analysed using a time series approach to better understand their trends over time. Time series analysis is essential because it enables us to understand the patterns and changes these indicators undergo over different time periods.
We used the seasonal_decompose function from the Stats models package to split the time series data into trend, seasonality, and residual components. This research clarified the complex trends and cyclical fluctuations found in our data.

Forecasting and Evaluation

In our forecasting method, we applied time series data-specific models that we carefully selected. We ensured that our forecasting models were dependable and capable of effectively representing the complex dynamics of our time series data by demonstrating our expertise in model validation, parameter tuning, and model selection.

Throughout the model-fitting process, we made use of the extensive time series analysis functionalities provided by Python's stats model's package. We have used the SARIMAX class in the library, which offered a more advanced forecasting model, to add external factors.

To forecast hotel pricing, we used a time series data set and SARIMA modelling. To identify the temporal structure, the first step in the procedure is data preparation, where the "Year" and "Month" columns are combined into a datetime format and assigned as the dataset index. The Augmented Dickey-Fuller (ADF) test is used to verify stationarity in the time series data and guarantee model reliability. To attain stationarity, non-stationary data—denoted by an ADF test p-value more than 0.05—are differenced. Following that, using parameters (p, d, q) for the non-seasonal portion and (P, D, Q, s) for the seasonal component—which can be adjusted based on the data—the SARIMA model is fitted to the data. The model is used to predict future values, which are then plotted against the actual data to allow for visual comparison. Finally, it provides a thorough method of time series forecasting by calculating the average projected price for 2022 and assessing model performance using the Mean Absolute Percentage Error (MAPE).

Appendix-B:

To commence our analysis of the Boston Planning and Redevelopment Authority (BPDA) dataset, including economic variables recorded monthly from January 2013 to December 2019. Key economic indicator box plots were an effective tool to view the outliers as shown in Figure-1. It illustrates that they were no potential outliers which could affect our analysis. As an example, low hotel occupancy rates can be associated with recessions in the economy or significant happenings that occur in the city. While quick changes in the property market or regulatory environment may be the reason for price surges in housing. Understanding these anomalies is essential. An in-depth analysis of Boston's economic state dataset can reveal the factors influencing these important economic metrics.

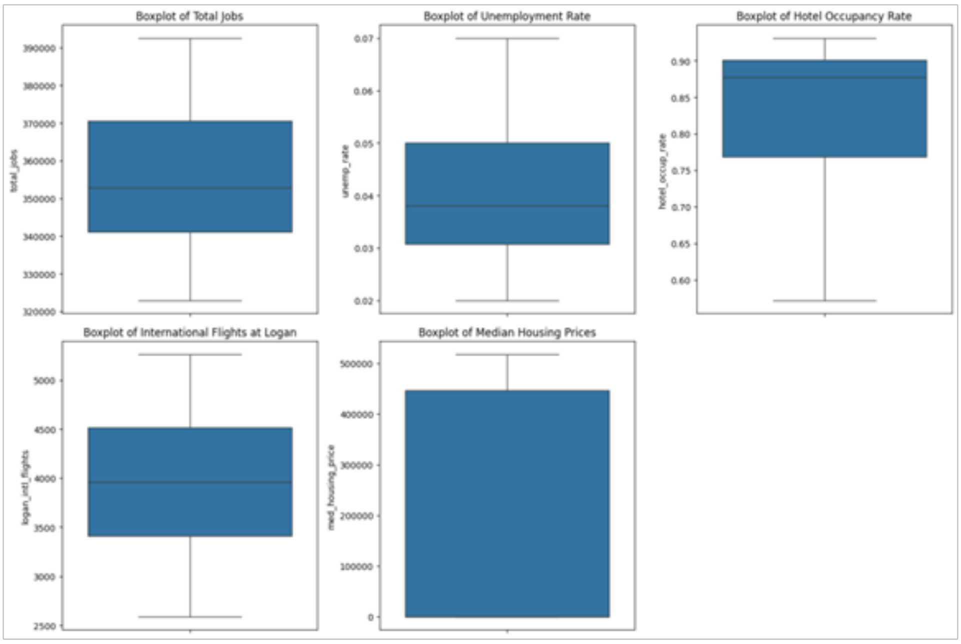


Figure 1: - Key Economic Indicator box plots.

Assuring that no anomalies were obscuring our initial study, we embarked on an attempt to learn more about the extensive structure of the data involved in this dataset. By examining it, we aim to gain a better understanding of the factors influencing its significant economic impact. We started to understand the relationship between the labour market and the housing market that changes in employment may impact housing affordability and demand. **Figure-2** illustrates various trends in Boston's unemployment rate from January 2013 to December 2019.

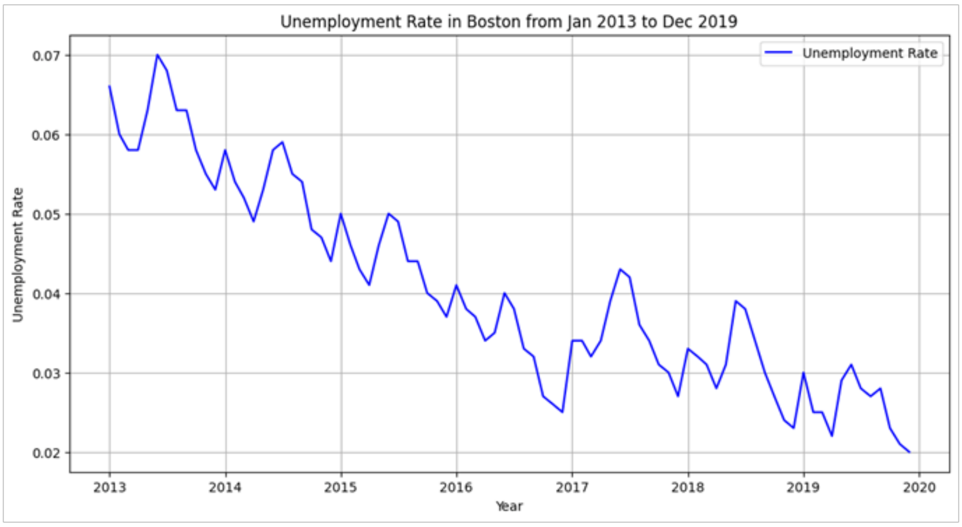


Figure 2: - Unemployment rate in Boston from Jan 2013 to Dec 2019.

During this time, the inflation rate fluctuated, reflecting variations in the job market and overall financial situation. The correlation coefficient between the unemployment rate and median home prices is 0.829. This suggests that throughout this time, there had been a significant positive correlation among the two variables. Said alternatively, it showed a positive correlation between the rise in the unemployment rate and the median house prices, and vice versa. Other variables such as market investment, outside investments in the real estate sector, or changes to housing regulations, might also have an impact. To discover the root causes of this tendency, additional research could be required.

So, we started analysing a relationship between the number of international flights at Logan and the hotel occupancy rate in Boston. The statistically determined moderate positive correlation of approximately 0.528 between the number of international flights at Logan Airport and the hotel occupancy rate in Boston provides crucial technical insights into the connection between international travel and the local hotel industry. Technically, this correlation suggests that a significant percentage of the variation in hotel occupancy rates in Boston can be attributed to changes in the volume of international flights. The moderate value of 0.528 indicates that, while foreign flights are a substantial factor determining hotel occupancy rates, they are not the sole determinant, as showed in Figure-3. Other factors, such as domestic tourism, local events, and seasonal variations, may also have an impact on occupancy trends.

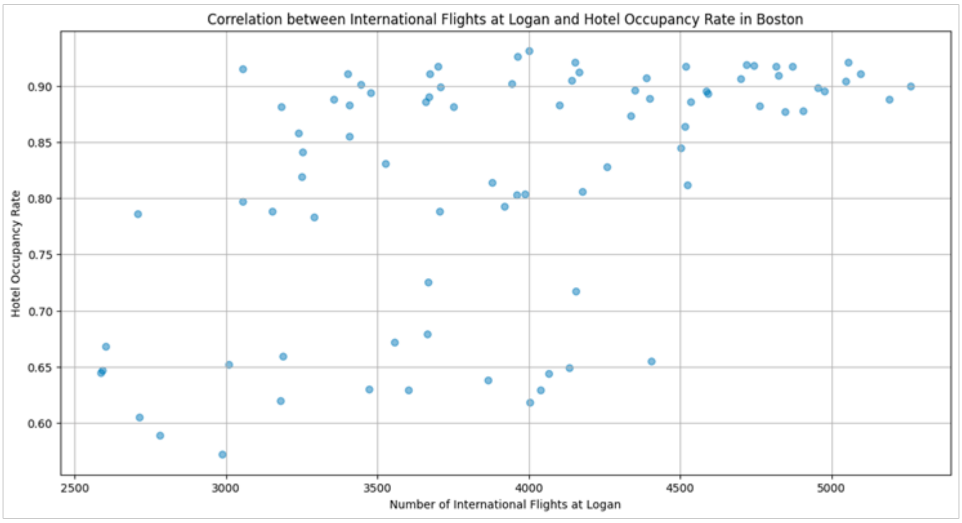


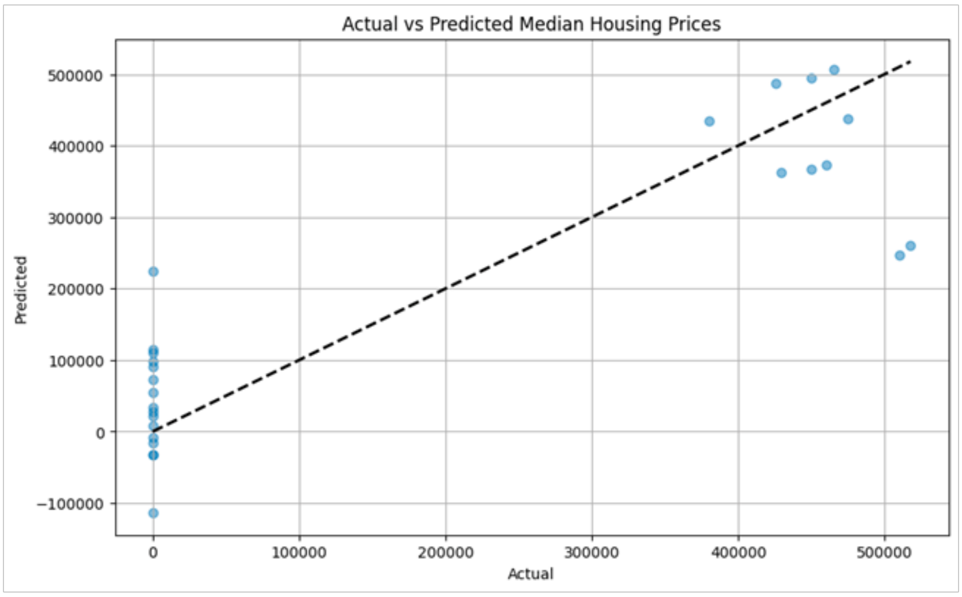
Figure 3: - Correlation between international flights and Hotel Occupancy.

We tried to develop a Linear regression model that effectively predicted median house prices based on a combination of economic variables such as total jobs, unemployment rate, hotel occupancy rates, and international flights after thoroughly analysing the columns. We evaluated the model's performance using the key statistic R-squared (R^2) after training and testing it. The R^2 value is around 0.779, indicating that the model explains roughly 77.9% of the variation in median prices. This is a high R2 score, indicating that the model captures a substantial percentage of the link between predictors and property prices. Figures 4 and 5 show two graphs that indicate the performance of the Linear regression model in forecasting median house prices in Boston.

- **Actual vs. Predicted Housing Prices:** The first graph displays a scatter plot of the median home prices compared to the model-predicted prices. When the actual and predicted values are equal, the dashed line denotes perfect predictions. While many points are close to this line, indicating accurate predictions, there are also several points further away, which contribute to the model's mean squared error.

- **Residuals of Predicted Housing Prices:** The second graph plots the residuals (the differences between actual and predicted values) against the predicted values There is no mistake, as shown by the horizontal red dashed line at zero. The scattering of residuals around this line indicates that the accuracy of the model varies with different house price levels. The random spread of residuals suggests that there might not be obvious patterns of error in the predictions, but the spread also indicates the variance in the accuracy of the model.

Together, these graphs illustrate the model's capability and limitations in predicting median housing prices based on the chosen economic indicators.



.

Figure 4: - Actual vs Predicted Median Housing Prices

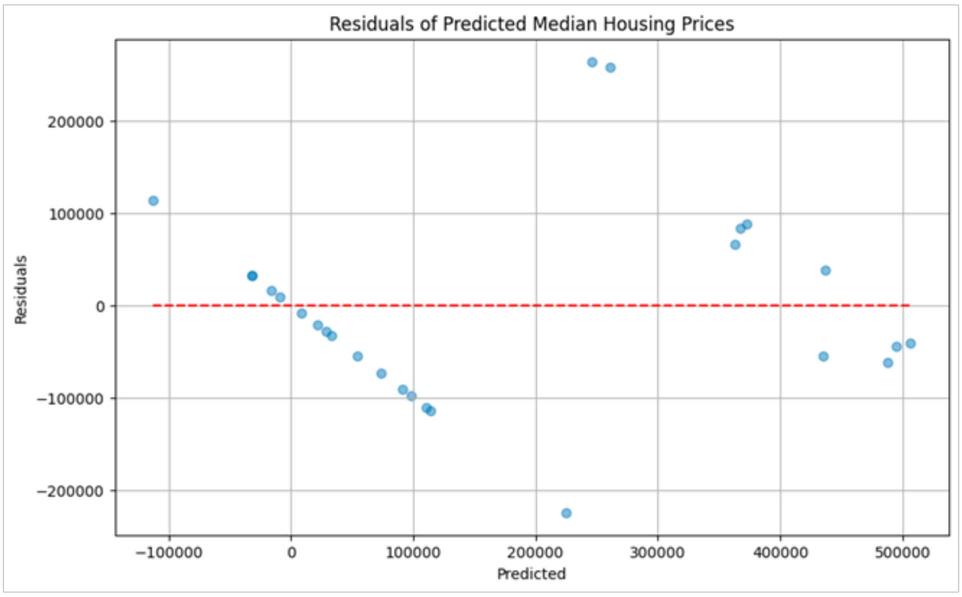


Figure 5: - Residuals of Predicted Median Housing Prices.

By analysing the similarity of economic indicator profiles across various months and years, we utilized cluster analysis techniques to identify distinct periods or phases in Boston's economic history. As illustrated in Figure 6, the dataset revealed both three and four distinct clusters when the clustering technique was applied. The results are as follows: -

3-Cluster Solution:

- Cluster 0: The highest hotel occupancy rates, average daily rates, total number of jobs, and lowest unemployment rates all correspond to this cluster. It represents a period of increased economic activity and growth.

- Cluster 1: Features somewhat higher unemployment rates than Cluster 0, fewer jobs overall, and a bit higher hotel occupancy and daily rates. This indicates a period when the economy will likely be steady but less active
- Cluster 2: Exhibits the lowest average daily rates and hotel occupancy rates, as well as an intermediate unemployment rate and jobs. This could indicate a phase of economic slowdown or recovery.

4-Cluster Solution:

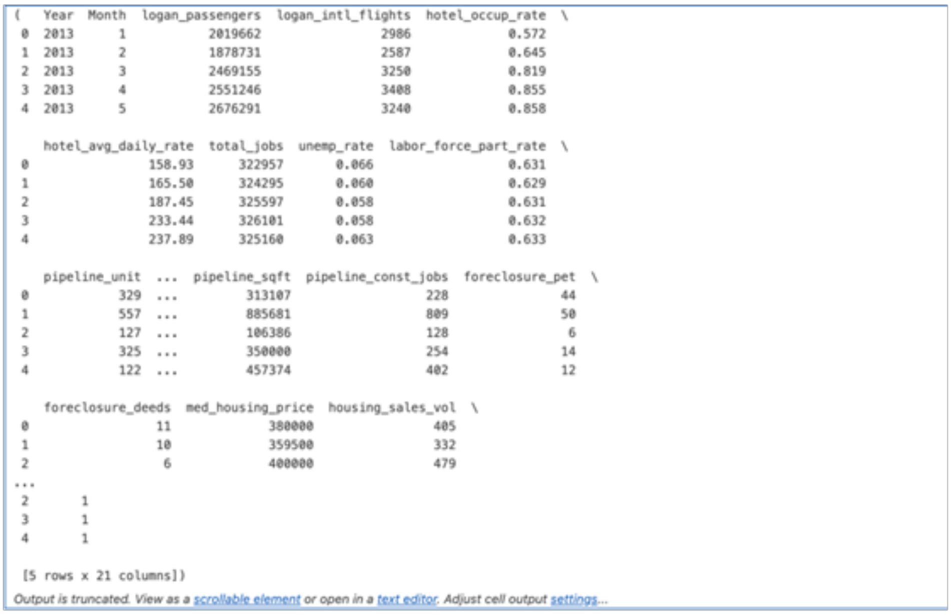
- Cluster 0: Similar to Cluster 2 in the three-cluster solution, but with a high number of pipeline units, suggesting a period of intensive development activity.

- Cluster 1: Identifies with Cluster 1 of the three-cluster solution, indicating a steady state of the economy.

- Cluster 2: indicates a time of peak economic performance by displaying the highest hotel occupancy rates and average daily rates, the highest number of jobs overall, and the lowest unemployment rates.

- Cluster 3: indicates a period of growth focused on development rather than tourism or business travel, as seen by the period's strong development activity (pipeline units and construction jobs) but lower hotel occupancy rates and average daily rates.

We calculated the average values of the economic indicators within each cluster to gain a better understanding of these clusters and what they could represent in terms of separate periods or phases in Boston's economic growth. This made it easier for us to recognize the characteristics of each cluster and maybe link them to economic phases. With differing degrees of economic activity and focus areas (such as tourism, job generation, and development), these clustering strategies highlight different phases of Boston's economic development. A more sophisticated approach is offered by the 4-cluster approach, which differentiates between various kinds of high-development phases.



```
(   Year  Month  logan_passengers  logan_intl_flights  hotel_occup_rate  \
 0  2013      1           2019662                2986             0.572
 1  2013      2           1878731                2587             0.645
 2  2013      3           2469155                3250             0.819
 3  2013      4           2551246                3408             0.855
 4  2013      5           2676291                3240             0.858

   hotel_avg_daily_rate  total_jobs  unemp_rate  labor_force_part_rate  \
 0               158.93      322957       0.066                  0.631
 1               165.50      324295       0.060                  0.629
 2               187.45      325597       0.058                  0.631
 3               233.44      326101       0.058                  0.632
 4               237.89      325160       0.063                  0.633

   pipeline_unit  ...  pipeline_sqft  pipeline_const_jobs  foreclosure_pet  \
 0           329  ...         313107                  228               44
 1           557  ...         885681                  809               50
 2           127  ...         106386                  128                6
 3           325  ...         350000                  254               14
 4           122  ...         457374                  402               12

   foreclosure_deeds  med_housing_price  housing_sales_vol  \
 0                11             380000                405
 1                10             359500                332
 2                 6             400000                479
...
 2      1
 3      1
 4      1

 [5 rows x 21 columns])
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Figure 6: - Clustering Results.

We conducted a Time Series Analysis on two crucial indicators, namely hotel occupancy rates and Logan Airport passenger counts, to further deepen our understanding of these economic trends. This analysis examined seasonal and cyclical patterns from Figures 7 and 8, offering insights into how these important sectors change over time.

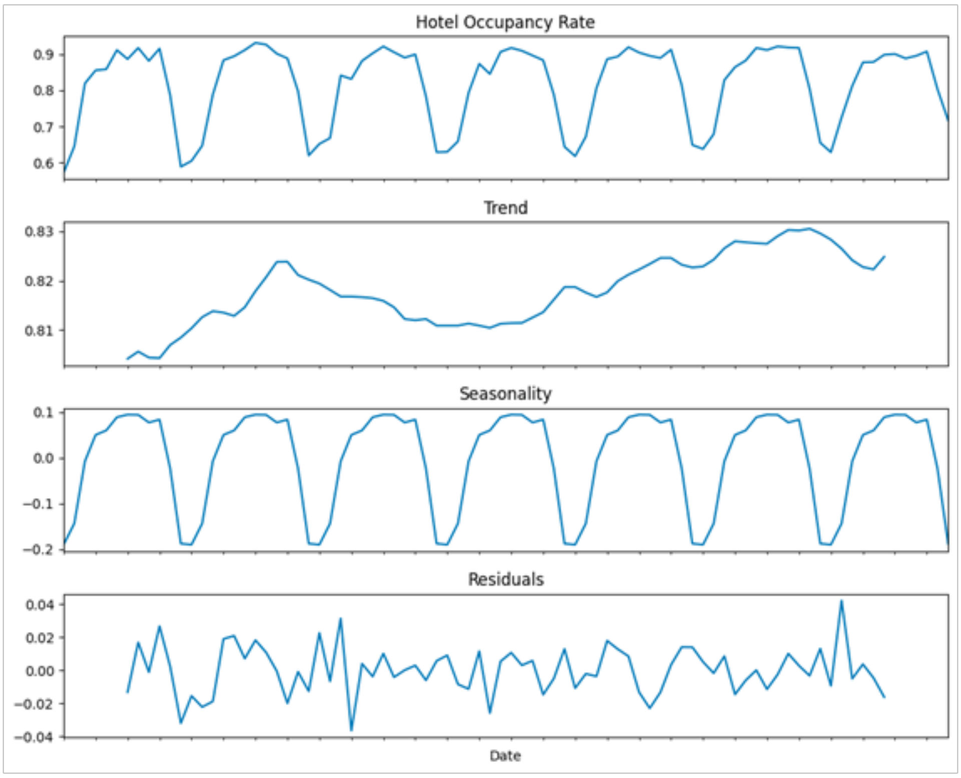


Figure 7: - Time series analysis for Hotel Occupancy.

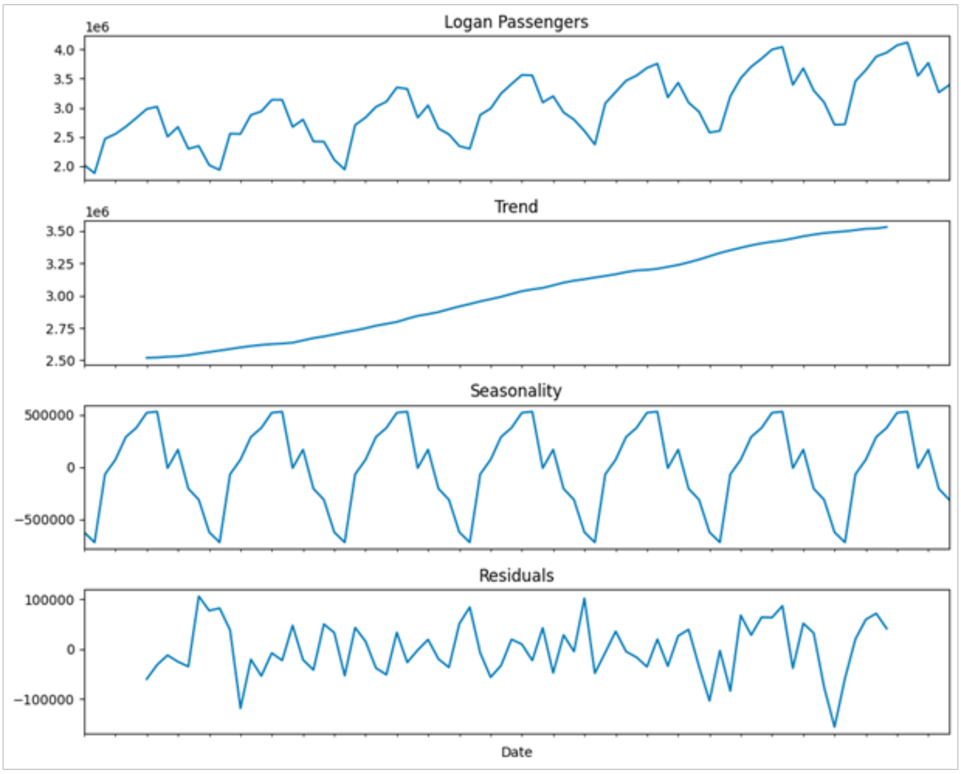


Figure 8: - Time series analysis for Logan Passengers.

The Time Series Analysis revealed several critical insights:

1. Seasonal Trends: The data revealed distinct seasonal patterns in hotel occupancy rates and airport passenger counts, characterized by peaks and troughs in certain seasons. Understanding these trends is crucial for formulating strategic plans in the travel and transportation industries.

2. Long-Term Trends: The analysis also brought to light long-term trends in these metrics, offering insights into gradual changes that could indicate broader shifts in the economy. These might include trends like rising or falling demand in air travel and tourism over time.

Utilizing SARIMA models enabled us to forecast future trends in hotel occupancy and airport passenger numbers. From Figures 9 and 10, which depict the SARIMA models for hotel occupancy rates and Logan Airport passenger counts, we can derive several key observations and insights.

Hotel Occupancy Rates:

1. Significant Seasonality: The hotel occupancy rates in the model demonstrated robust seasonal patterns that corresponded with regular tourism cycles, with certain months exhibiting consistently higher or lower occupancy rates.

2. Trend Component: The presence of a trend component in the model suggests the existence of underlying long-term fluctuations in hotel occupancy rates, which may point towards downward trends in Boston's tourism industry.

3. Forecast Accuracy: The potential of the model to predict future occupancy rates may be crucial for resource allocation and strategy development for local governments, tourism planners, and hospitality enterprises.

4. Statistical Significance: Several of the SARIMA model's parameters were statistically significant, suggesting that they had a substantial impact on estimating future occupancy rates. The important Moving Average (MA) phrase indicated that recent historical data is relevant for predicting future occupancy.

5. Model Fit and Warnings: The covariance matrix's warnings and the model fit indicated by the AIC and BIC values point to the need for more model development to further enhance accuracy.

Logan Airport Passengers:

1. Seasonal Fluctuations: The SARIMA model for Logan passengers showed clear seasonal tendencies, which corresponded to projected peaks and troughs in air travel due to holidays, vacation seasons, or business travel schedules.
2. Long-Term Trends Analysis: Any long-term upward or downward trends in passenger numbers were caught by the model; these trends may be a sign of larger patterns in the aviation sector or the state of the local economy.

3. Forecasting Potential: The airport administration, airlines, and other businesses can utilize the model's predicted passenger counts for their planning needs.

4. Parameter Significance: Important variables affecting passenger counts were identified by the model's Moving Average (MA) component and a few seasonal elements.
5. Model Warnings and Diagnostics: The model summary's diagnostics and cautions indicate that the results should be carefully analysed and that any necessary model adjustments should be made.

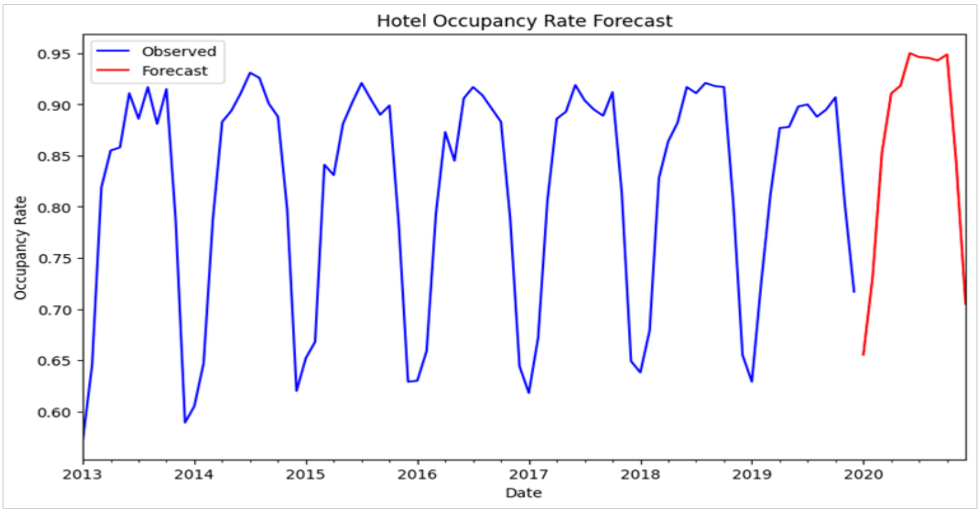


Figure 9: - Hotel Occupancy Rate Forecast.

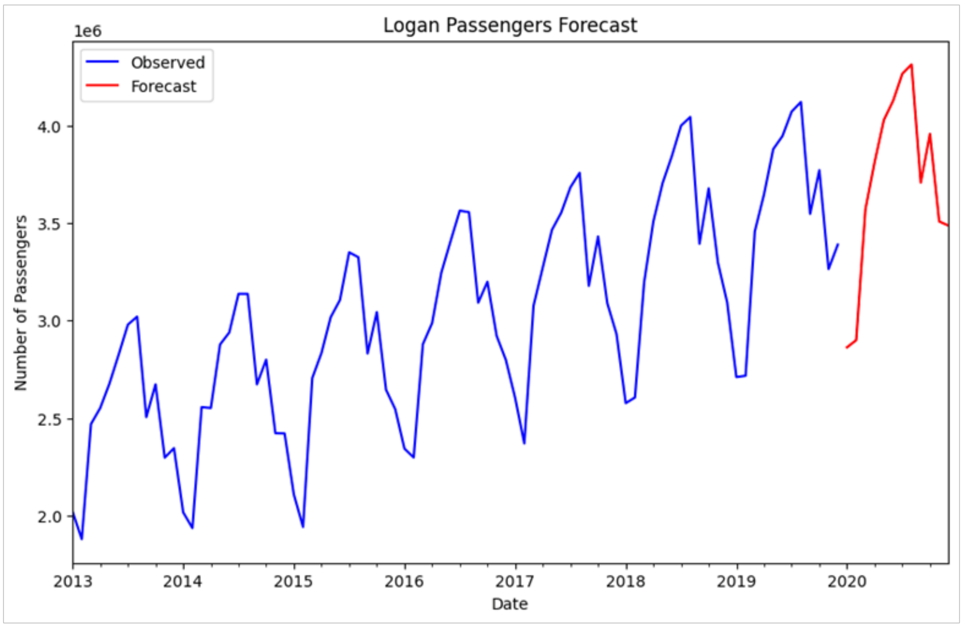


Figure 10: - Logan Passengers Forecast.

To evaluate the precision and efficacy of our model, we expanded our investigation to forecast hotel rates until 2022. This project was started for more reasons than only testing our SARIMA models' predicting power. Several significant conclusions and observations may be made from Figure 11, which displays the SARIMA models generated for hotel occupancy rates for 2022:

1. Model Parameters and Fit: The autocorrelation and partial autocorrelation properties of the data were used to parameterize the SARIMA model, resulting in an order of (1, 1, 1) for non-seasonal components and (1, 1, 1, 12) for seasonal components. These selection criteria were made to account for both seasonal and short-term fluctuations in hotel rates.

2. Stationarity: According to our preliminary testing, we could move forward with SARIMA modelling without the need for differencing to stabilize the series mean because the hotel prices data was found to be stationary.

3. Forecasting and Confidence Intervals: Forecasts for each month through December 2022 were produced by the model. Additionally, it produced confidence intervals for these projections, providing a range that represents the expected future values.

4. Mean Price for 2022: It was determined that the average hotel rate for 2022 would be roughly $272.96. The predicted price level for the full year is presented briefly in Figure 12.

5. Monthly Forecast Values: The model predicted that hotel rates will rise gradually in 2022, from roughly $187.90 in January to roughly $198.89 in December. Figure 13 provides a graphic representation of the values, which show a consistent annual increase in hotel prices
6. Model Reliability and Validation: The model produced a statistically reasonable forecast, but the consistency of previous trends and patterns determines how reliable the forecast is. There could be variations from these forecasts due to unanticipated events (like a pandemic) or external economic considerations. To preserve accuracy, it is advised that the model be continuously validated using current data.

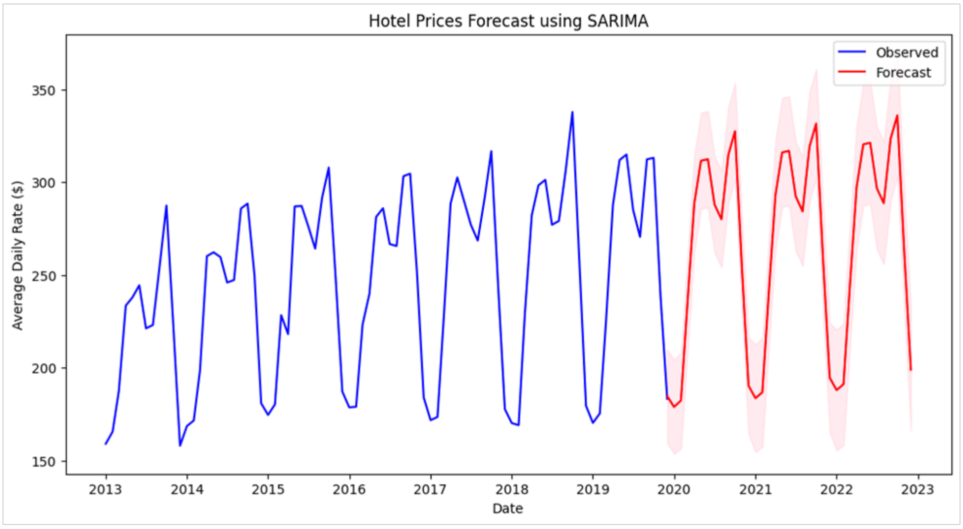
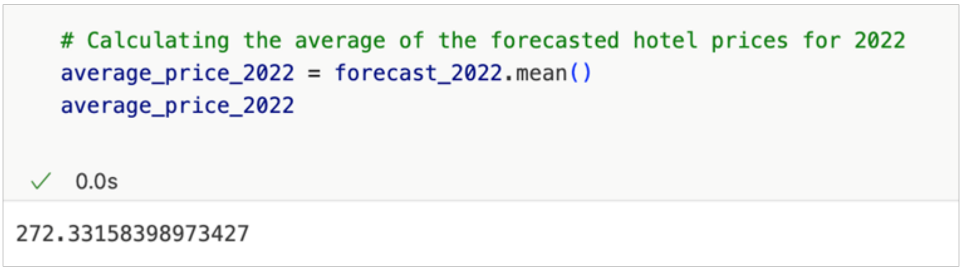


Figure 11: - Hotel Price Rate Forecast.

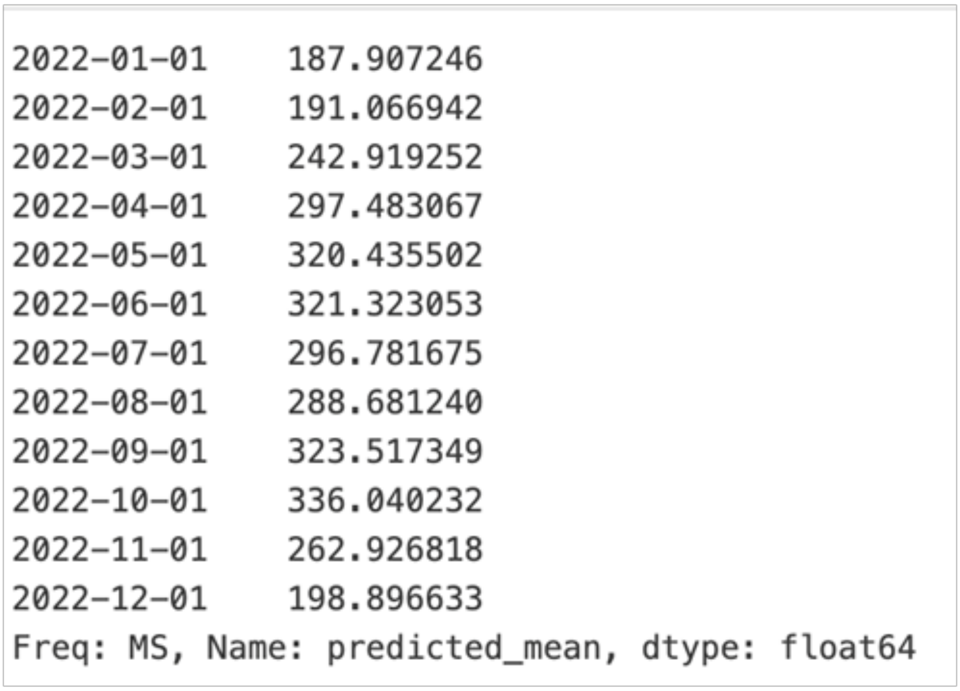


```
# Calculating the average of the forecasted hotel prices for 2022
average_price_2022 = forecast_2022.mean()
average_price_2022
```

✓ 0.0s

```
272.33158398973427
```

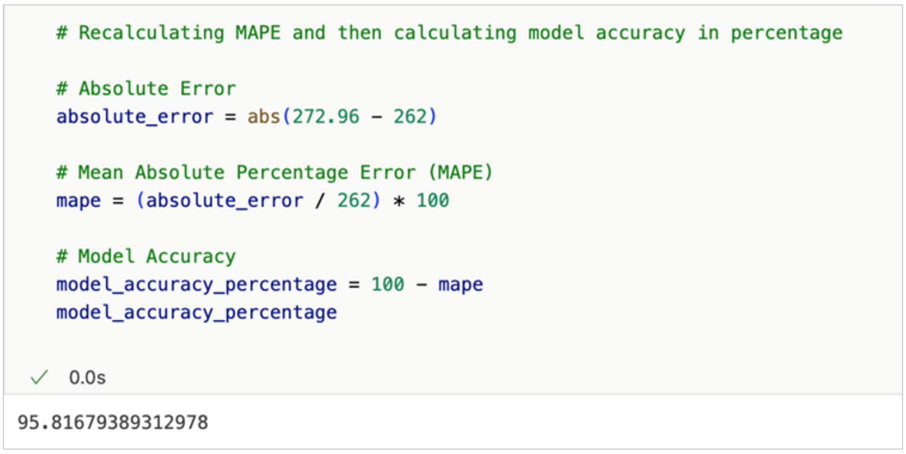
Figure 12: - Average Price for 2022.



```
2022-01-01    187.907246
2022-02-01    191.066942
2022-03-01    242.919252
2022-04-01    297.483067
2022-05-01    320.435502
2022-06-01    321.323053
2022-07-01    296.781675
2022-08-01    288.681240
2022-09-01    323.517349
2022-10-01    336.040232
2022-11-01    262.926818
2022-12-01    198.896633
Freq: MS, Name: predicted_mean, dtype: float64
```

Figure 13: - Monthly forecast values for 2022.

The average hotel rate for Boston in 2022 was estimated by the SARIMA model to be $272.96. We cross-referenced the projected average price with actual data from the website "https://www.cheaphotels.org/press/cities22.html," which revealed that the average hotel rate in Boston for 2022 was $262, to verify this prediction. We were able to assess the correctness of our model through this comparison. As displayed in Figure 14, after computing the accuracy, we discovered that the forecast made by our SARIMA model was roughly 95.8% accurate.



```
# Recalculating MAPE and then calculating model accuracy in percentage

# Absolute Error
absolute_error = abs(272.96 - 262)

# Mean Absolute Percentage Error (MAPE)
mape = (absolute_error / 262) * 100

# Model Accuracy
model_accuracy_percentage = 100 - mape
model_accuracy_percentage
```

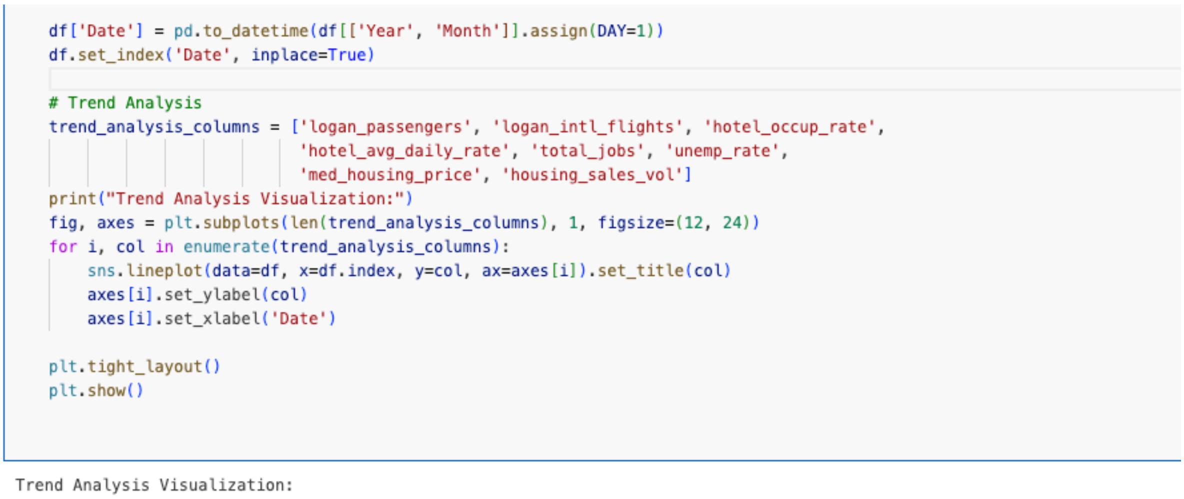
✓ 0.0s

```
95.81679389312978
```

Figure 14: - Accuracy of the model.

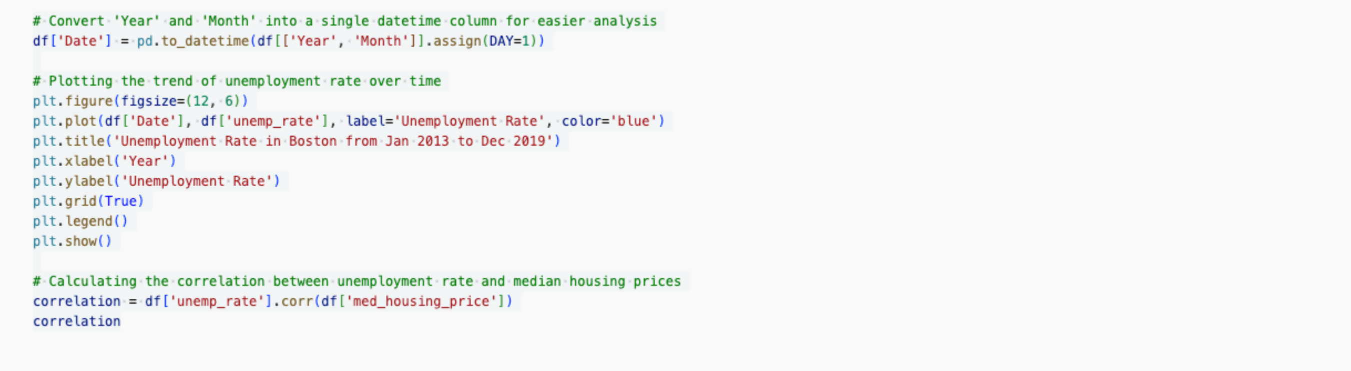
The cluster analysis and the SARIMA model's forecast offer a thorough understanding of Boston's economic environment, with an emphasis on the city's transportation and tourism industries. Gaining significant insights and providing a data-driven foundation for predicting market trends is possible by utilizing SARIMA to forecast hotel pricing until 2022. However, there are risks associated with relying only on historical trends; unexpected changes in the market and outside variables may affect how accurate projections are made. The model must be updated and improved on a regular basis to remain accurate and relevant. This comprehensive study creates opportunities for more in-depth investigation of the variables influencing these economic patterns in addition to facilitating improved resource allocation and planning. Making strategic, well-informed decisions within the framework of Boston's dynamic economy requires stakeholders to have a thorough understanding of the interaction of different factors that impact tourism and transportation. Effective economic planning and growth in a city known for its complexity and dynamism require an all-encompassing approach.

Appendix-C: Code

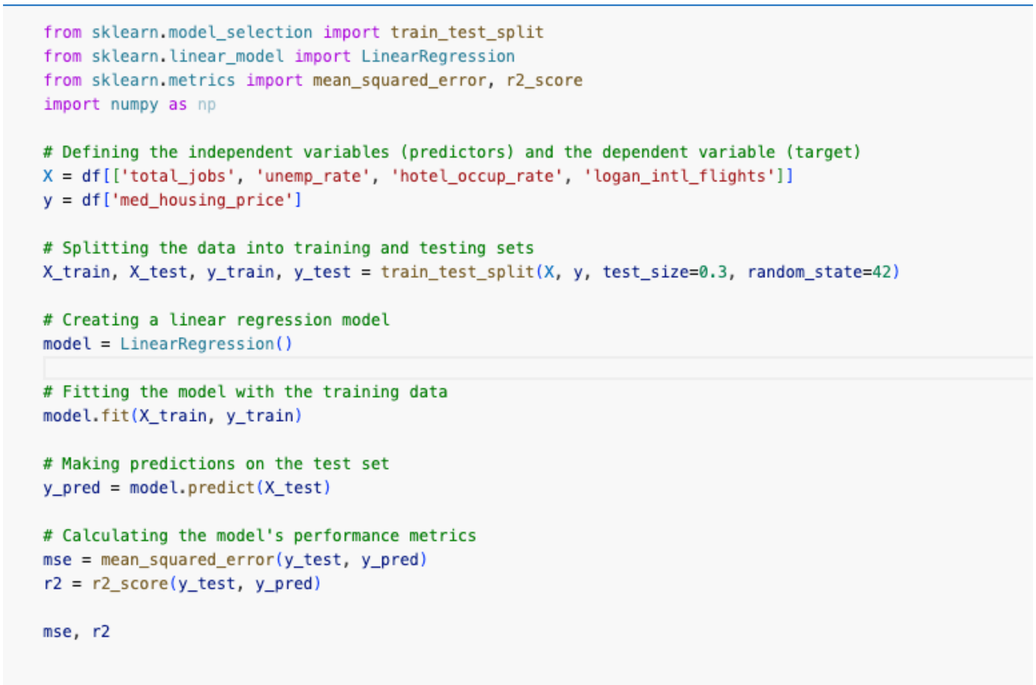


```
df['Date'] = pd.to_datetime(df[['Year', 'Month']].assign(DAY=1))
df.set_index('Date', inplace=True)

# Trend Analysis
trend_analysis_columns = ['logan_passengers', 'logan_intl_flights', 'hotel_occup_rate',
                          'hotel_avg_daily_rate', 'total_jobs', 'unemp_rate',
                          'med_housing_price', 'housing_sales_vol']
print("Trend Analysis Visualization:")
fig, axes = plt.subplots(len(trend_analysis_columns), 1, figsize=(12, 24))
for i, col in enumerate(trend_analysis_columns):
    sns.lineplot(data=df, x=df.index, y=col, ax=axes[i]).set_title(col)
    axes[i].set_ylabel(col)
    axes[i].set_xlabel('Date')

plt.tight_layout()
plt.show()
```

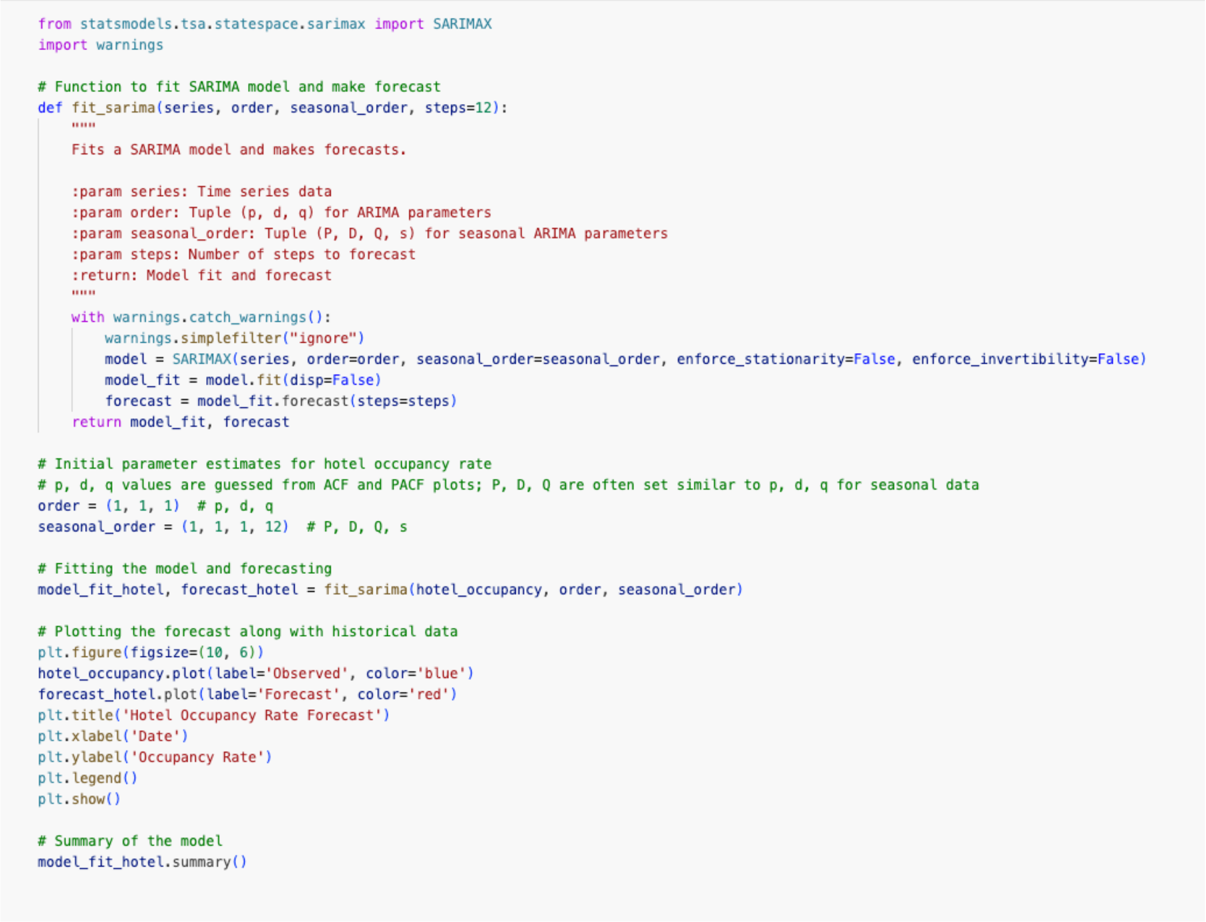
```
Trend Analysis Visualization:
```



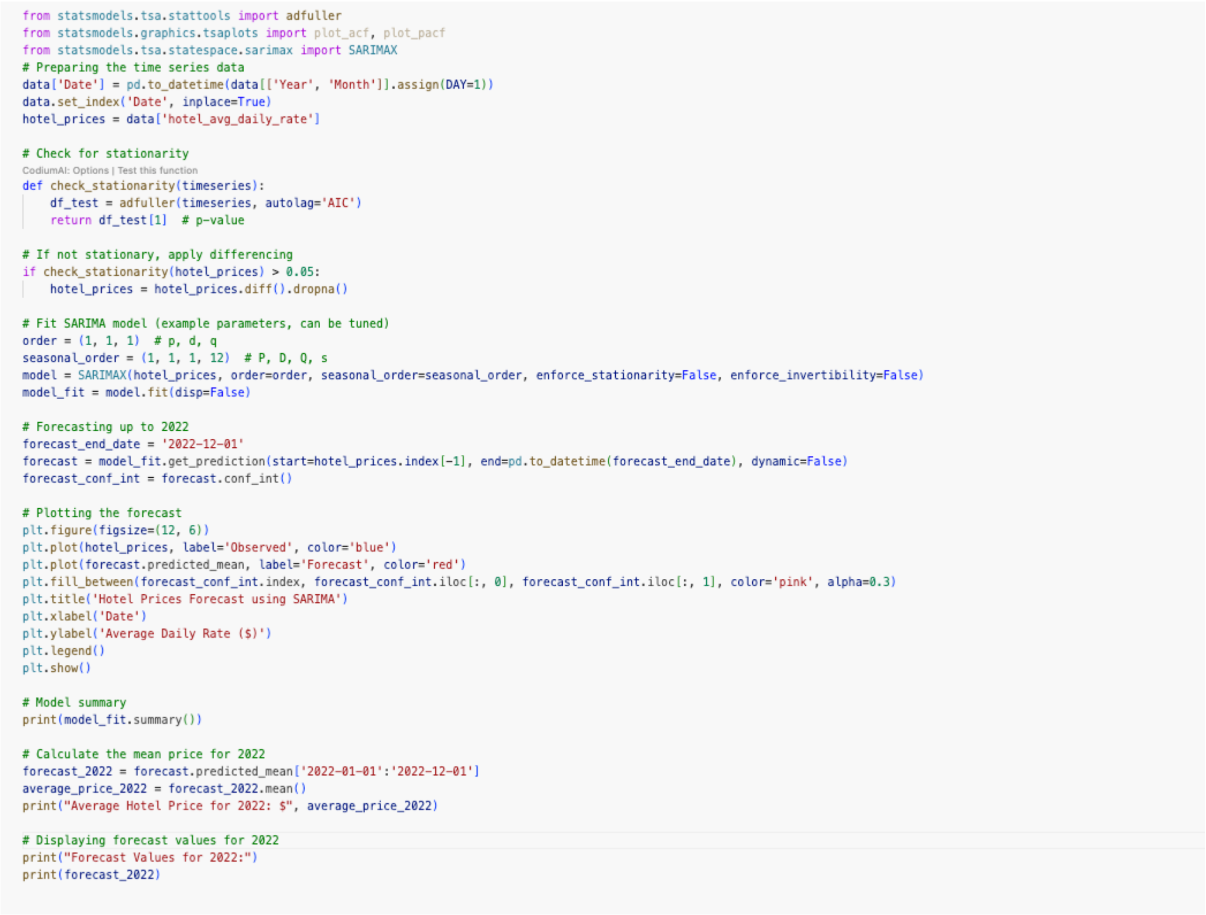
```
# Convert 'Year' and 'Month' into a single datetime column for easier analysis
df['Date'] = pd.to_datetime(df[['Year', 'Month']].assign(DAY=1))

# Plotting the trend of unemployment rate over time
plt.figure(figsize=(12, 6))
plt.plot(df['Date'], df['unemp_rate'], label='Unemployment Rate', color='blue')
plt.title('Unemployment Rate in Boston from Jan 2013 to Dec 2019')
plt.xlabel('Year')
plt.ylabel('Unemployment Rate')
plt.grid(True)
plt.legend()
plt.show()

# Calculating the correlation between unemployment rate and median housing prices
correlation = df['unemp_rate'].corr(df['med_housing_price'])
correlation
```



```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# Defining the independent variables (predictors) and the dependent variable (target)
X = df[['total_jobs', 'unemp_rate', 'hotel_occup_rate', 'logan_intl_flights']]
y = df['med_housing_price']

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Creating a linear regression model
model = LinearRegression()

# Fitting the model with the training data
model.fit(X_train, y_train)

# Making predictions on the test set
y_pred = model.predict(X_test)

# Calculating the model's performance metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

mse, r2
```



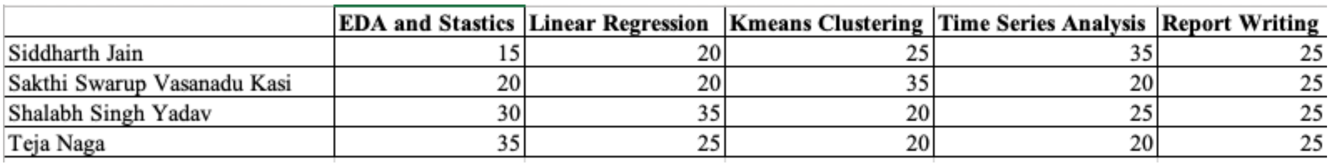
```
from statsmodels.tsa.statespace.sarimax import SARIMAX
import warnings

# Function to fit SARIMA model and make forecast
def fit_sarima(series, order, seasonal_order, steps=12):
    """
    Fits a SARIMA model and makes forecasts.

    :param series: Time series data
    :param order: Tuple (p, d, q) for ARIMA parameters
    :param seasonal_order: Tuple (P, D, Q, s) for seasonal ARIMA parameters
    :param steps: Number of steps to forecast
    :return: Model fit and forecast
    """
    with warnings.catch_warnings():
        warnings.simplefilter("ignore")
        model = SARIMAX(series, order=order, seasonal_order=seasonal_order, enforce_stationarity=False, enforce_invertibility=False)
        model_fit = model.fit(disp=False)
        forecast = model_fit.forecast(steps=steps)
    return model_fit, forecast

# Initial parameter estimates for hotel occupancy rate
# p, d, q values are guessed from ACF and PACF plots; P, D, Q are often set similar to p, d, q for seasonal data
order = (1, 1, 1)  # p, d, q
seasonal_order = (1, 1, 1, 12)  # P, D, Q, s

# Fitting the model and forecasting
model_fit_hotel, forecast_hotel = fit_sarima(hotel_occupancy, order, seasonal_order)

# Plotting the forecast along with historical data
plt.figure(figsize=(10, 6))
hotel_occupancy.plot(label='Observed', color='blue')
forecast_hotel.plot(label='Forecast', color='red')
plt.title('Hotel Occupancy Rate Forecast')
plt.xlabel('Date')
plt.ylabel('Occupancy Rate')
plt.legend()
plt.show()

# Summary of the model
model_fit_hotel.summary()
```



```
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.statespace.sarimax import SARIMAX
# Preparing the time series data
data['Date'] = pd.to_datetime(data[['Year', 'Month']].assign(DAY=1))
data.set_index('Date', inplace=True)
hotel_prices = data['hotel_avg_daily_rate']

# Check for stationarity
CodiumAI: Options | Test this function
def check_stationarity(timeseries):
    df_test = adfuller(timeseries, autolag='AIC')
    return df_test[1]  # p-value

# If not stationary, apply differencing
if check_stationarity(hotel_prices) > 0.05:
    hotel_prices = hotel_prices.diff().dropna()

# Fit SARIMA model (example parameters, can be tuned)
order = (1, 1, 1)  # p, d, q
seasonal_order = (1, 1, 1, 12)  # P, D, Q, s
model = SARIMAX(hotel_prices, order=order, seasonal_order=seasonal_order, enforce_stationarity=False, enforce_invertibility=False)
model_fit = model.fit(disp=False)

# Forecasting up to 2022
forecast_end_date = '2022-12-01'
forecast = model_fit.get_prediction(start=hotel_prices.index[-1], end=pd.to_datetime(forecast_end_date), dynamic=False)
forecast_conf_int = forecast.conf_int()

# Plotting the forecast
plt.figure(figsize=(12, 6))
plt.plot(hotel_prices, label='Observed', color='blue')
plt.plot(forecast.predicted_mean, label='Forecast', color='red')
plt.fill_between(forecast_conf_int.index, forecast_conf_int.iloc[:, 0], forecast_conf_int.iloc[:, 1], color='pink', alpha=0.3)
plt.title('Hotel Prices Forecast using SARIMA')
plt.xlabel('Date')
plt.ylabel('Average Daily Rate ($)')
plt.legend()
plt.show()

# Model summary
print(model_fit.summary())

# Calculate the mean price for 2022
forecast_2022 = forecast.predicted_mean['2022-01-01':'2022-12-01']
average_price_2022 = forecast_2022.mean()
print("Average Hotel Price for 2022: $", average_price_2022)

# Displaying forecast values for 2022
print("Forecast Values for 2022:")
print(forecast_2022)
```

Contribution: -

All four group members analyzed the economic indicator dataset. The following task were performed by all the members of the group.



	EDA and Stastics	Linear Regression	Kmeans Clustering	Time Series Analysis	Report Writing
Siddharth Jain	15	20	25	35	25
Sakthi Swarup Vasanadu Kasi	20	20	35	20	25
Shalabh Singh Yadav	30	35	20	25	25
Teja Naga	35	25	20	20	25

GitHub Link of the Project: https://github.com/siddharth00914/Economic_indicators_boston.git.