Analyzing Inclusive Growth: A Deep Dive into Boston's Economic Indicators (2013-2019)

Team Members: (GROUP - 1)

Sai Sahithi Neela -*02082013*

Venkata Sai Sandeep-*02084046*

Shrishti Sudhakar Shetty-*0208642*

ISSUES:

In this in-depth analysis, we investigate the intricate relationships between key economic indicators within Boston's market and their influence on employment levels. By examining the number of total jobs as our target variable, we delve into how elements such as air travel volume and hotel industry metrics serve as pivotal features that forecast employment trends. Our results shed light on the significant correlation between the vibrancy of Boston's tourism sector—reflected in airport passenger counts and hotel occupancy and rates—and the city's job market vitality. The insights obtained from this study provide a valuable foundation for strategic policy development aimed at bolstering job growth while synergistically fostering the growth of the tourism sector. This research emphasizes the importance of a harmonized approach to economic planning that proactively supports a flourishing job market, bolstered by a resilient and expanding tourism industry.

Examine the relationship between the number of passengers and the hotel occupancy rate, along with the average daily rate for hotels.

Analyze the dynamics between the unemployment rate and labor force participation rate. Understand if changes in one are mirrored by changes in the other or if there are counterintuitive trends that require further investigation.

Investigate the interplay between the unemployment rate and labor force participation rate to decipher the extent to which they reflect or contradict each other's trends. This analysis aims to uncover nuanced patterns or unexpected anomalies, offering deeper insights for comprehensive labor market assessments.

Use statistical methods to assess potential causal relationships between indicators, such as whether changes in unemployment rates lead to changes in labor force participation or vice versa.

Assess the impact of tourism (indicated by "logan_passengers" and "hotel_occup_rate") on local employment levels ("total_jobs"). This analysis can reveal how susceptible the job market is to fluctuations in the tourism sector.

Evaluate how different economic sectors (besides tourism) influence the overall employment rates. This can provide a broader view of the economic drivers in Boston.

Explore the relationship between hotel occupancy rates and average daily rates to understand pricing strategies in the hospitality sector and their effectiveness in different economic conditions.

Investigate how macroeconomic factors, as reflected in the unemployment rate, influence the labor force participation rate. This can shed light on potential workforce disengagement in challenging economic times.

Utilize the data to build predictive models for forecasting future employment trends based on current and past economic indicators, helping in proactive policymaking and planning.

How can K-means clustering be utilized to identify distinct groups or clusters based on the similarities in economic indicators, and how can this approach help reveal hidden patterns or relationships between different economic variables?

FINDINGS:

Analysis shows a strong correlation between Logan Airport passenger numbers and hotel occupancy rates, highlighting air travel's significant influence on the hospitality sector.

A negative correlation between unemployment rates and total jobs was observed, reflecting the typical inverse relationship between job availability and unemployment.

Variations in hotel average daily rates were found to correlate with changes in employment levels, pointing to the hospitality sector's role in job market dynamics.

Analysis identified several outliers significantly distanced from the regression line, impacting the model's accuracy. Excluding these outliers led to a slight improvement in the model's fit, underscoring their influence on predictive performance.

The model demonstrates accuracy within the range of data used for training but may lack precision for input variables at extreme high or low values. This limitation underscores the need for cautious application and interpretation of the model beyond its trained data scope.

The range of total jobs fluctuates between 320,000 to 390,000, indicating a considerable variation in employment levels, potentially driven by seasonal, economic, or industry-specific factors.

Approximately 20% of the days observed have more than 360,000 total jobs, suggesting periods of higher-than-average employment, which could be linked to seasonal hiring trends or economic growth spurts.

Regarding hotel occupancy rates, the prevalent range is observed to be between 65% to 70%. This reflects a generally healthy demand for lodging and indicates a steady influx of visitors or business travelers in the city.

In terms of hotel average daily rates, the most common rate falls between $225 to $250, which might be indicative of the city's positioning in the hospitality market and its appeal to a certain segment of travelers.

The Augmented Dickey-Fuller (ADF) test results indicate that among the examined variables, only "hotel_avg_daily_rate" is stationary. The series "total_jobs," "unemp_rate," "logan_passengers," "logan_intl_flights," and "hotel_occup_rate" are not stationary, suggesting the need for differencing or other data transformations to stabilize their mean and variance over time for effective time series analysis and forecasting models.

Upon analyzing the Autocorrelation Function (ACF) plots for time series on each Economic Indicator, significant initial lags were observed in several variables, suggesting that autoregressive models could be well-suited for these series.

The ARIMA models are performing well on the variables, accurately reflecting both the long-term trends and seasonal variations. This suggests that the models could be reliable tools for forecasting and understanding the dynamics of these economic and operational metrics.

After evaluating and comparing various models on the Economic Indicators, Linear Regression and Random Forest showed high effectiveness with R-squared scores of 95.06% and 95.87%, respectively. In contrast, Support Vector Machines underperformed significantly, indicated by a -0.09% R-squared score, suggesting its unsuitability or need for further tuning.

DISCUSSIONS:

The analysis of Boston's economic indicators offers significant insights into the city's economic dynamics, each carrying distinct implications for policymaking and strategic planning.

Firstly, the strong correlation between Logan Airport passenger traffic and hotel occupancy rates suggests a substantial reliance of the city's economy on the tourism sector. This dependency highlights the potential vulnerability of Boston's economy to fluctuations in tourism trends. Policymakers might consider this as a cue to diversify economic activities, reducing over-reliance on tourism and enhancing economic resilience.

The observed inverse relationship between unemployment rates and total jobs reaffirms fundamental economic principles. This finding indicates that job creation is a crucial lever for managing unemployment levels in the city. It suggests the need for employment policies that are responsive and adaptable to changing economic conditions, focusing on sectors that show robust employment generation.

The study also sheds light on the hospitality sector's influence on employment, as changes in hotel rates correlate with shifts in job numbers. This implies that the hospitality industry is a significant driver of the job market, and strategies aimed at bolstering this sector could positively impact overall employment levels.

However, the model's precision is limited to the range of data it was trained on, cautioning against extrapolating its findings to scenarios with extreme input values. This limitation underscores the importance of continuous model refinement and validation to ensure its applicability and reliability in evolving economic circumstances.

Moreover, the substantial variation in employment levels, influenced by factors like seasonality and economic cycles, underscores the need for economic policies that are flexible and responsive to these factors. Policymakers should consider strategies that can capitalize on seasonal trends and cushion the impact of economic downturns.

Predictive modeling for future employment trends offers a proactive approach to economic planning. By forecasting future trends based on historical data, policymakers can anticipate changes and formulate strategies to address potential challenges and opportunities in the job market.

Finally, the application of K-means clustering reveals hidden patterns and relationships between various economic indicators. This methodological approach can guide targeted economic interventions, enabling policymakers to focus on specific areas or sectors that require attention, thereby ensuring more effective and impactful economic development strategies.

Overall, the findings from this analysis provide a multi-faceted understanding of Boston's economic landscape. They highlight the interdependencies within the economy and offer a foundation for informed decision-making, enabling the development of tailored policies and strategies to foster a robust, diverse, and resilient economic environment in Boston.

APPENDIX A: METHOD

Our project is centered on the "Economic Indicators" dataset sourced from Analyze Boston, the City of Boston's open data hub. This platform provides a wealth of data reflecting various aspects of city life, governance, and operations, making it an invaluable resource for understanding the dynamics of Boston's economy.

In our analysis, we meticulously prepared the dataset for exploration and modeling. This preparation involved several crucial steps:

Began by examining the dataset for null or missing values, ensuring the integrity and completeness of the data. This step is vital in avoiding skewed analyses and conclusions.

Checked for and removed any duplicate records to maintain the uniqueness of the data, a crucial step to prevent any bias or repetition in the analysis.

Conducted a thorough exploratory analysis, diving into each variable to understand its distribution and characteristics. This process involved visualizing different aspects of the data to uncover patterns, trends, and anomalies.

Variables

logan passengers: This variable represents the number of passengers traveling through Logan Airport, providing insights into the volume of air travel and its potential impact on the city's economy.

hotel_occup_rate*:* The hotel occupancy rate is a crucial indicator of the health of the hospitality sector, reflecting the demand for lodging in the city.

total_jobs*:* This attribute signifies the total number of jobs in Boston, serving as a barometer for the city's employment landscape.

unemp_rate*:* The unemployment rate is a critical economic indicator, reflecting the percentage of the labor force that is unemployed and actively seeking employment.

Year and Month*:* These variables represent the time aspect of the dataset. They are crucial for understanding seasonal and yearly trends in the economic indicators.

Logan International Flights (logan_intl_flights)*:* This variable represents the number of international flights at Logan Airport. It can be a significant indicator of international tourism and business travel activity, potentially influencing local economic conditions.

Hotel Average Daily Rate (hotel_avg_daily_rate)*:* This metric reflects the average rate for a day's stay in hotels. It's a key indicator of the health of the hotel industry and can also signal changes in tourist inflows and economic conditions affecting leisure and business travel.

Labor Force Participation Rate (labor_force_part_rate)*:* This rate indicates the proportion of the working-age population that is either employed or actively seeking employment. It's a critical indicator of the overall health of the labor market and can provide insights into broader economic trends.

Each of these variables offers a unique lens through which we can view and analyze the economic health and trends in Boston.

ANALYTIC METHODS:

Statistical Procedures:

Linear Regression: A statistical method to model the relationship between a dependent variable (like "total_jobs") and one or more independent variables (like "hotel_occup_rate").

Random Forest Regressor: An ensemble learning method used for regression tasks. It builds multiple decision trees and merges them to get a more accurate and stable prediction.

Support Vector Machines (SVM): Used for both classification and regression challenges. In this context, it could model complex relationships between economic indicators.

Augmented Dickey–Fuller Test (ADF): Tests the null hypothesis that a unit root is present in a time series sample. It's crucial for understanding the stationarity of variables like "logan_passengers."

Differencing Method of ADF: A technique used to transform a non-stationary time series into a stationary one, by differencing the data points at lagged intervals.

Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF): These tools help in identifying the order of an autoregressive integrated moving average (ARIMA) model, crucial for time series forecasting.

Autoregressive Integrated Moving Average (ARIMA): This modeling technique is used for analyzing and forecasting time series data, essential for predicting future trends in economic indicators.

Histograms: Utilized for depicting the frequency distribution of a variable. They are instrumental in understanding the distribution pattern of each economic indicator.

Pair Plot: A matrix of scatterplots, useful for understanding pairwise relationships between all variables.

Distribution Plot: Helps in understanding the distribution of a variable, useful for variables like "hotel_avg_daily_rate" and "unemp_rate."

Box Plots: These are used to show the distribution of quantitative data and identify outliers in each economic variable.

Regression Plot: Visualizes the linear relationship between two variables, such as "logan_passengers" and "hotel_occup_rate."

Joint Plots: Combines scatter plots and histograms to offer a detailed perspective on the relationships between pairs of economic variables, along with insights into their individual distributions.

APPENDIX B: RESULTS

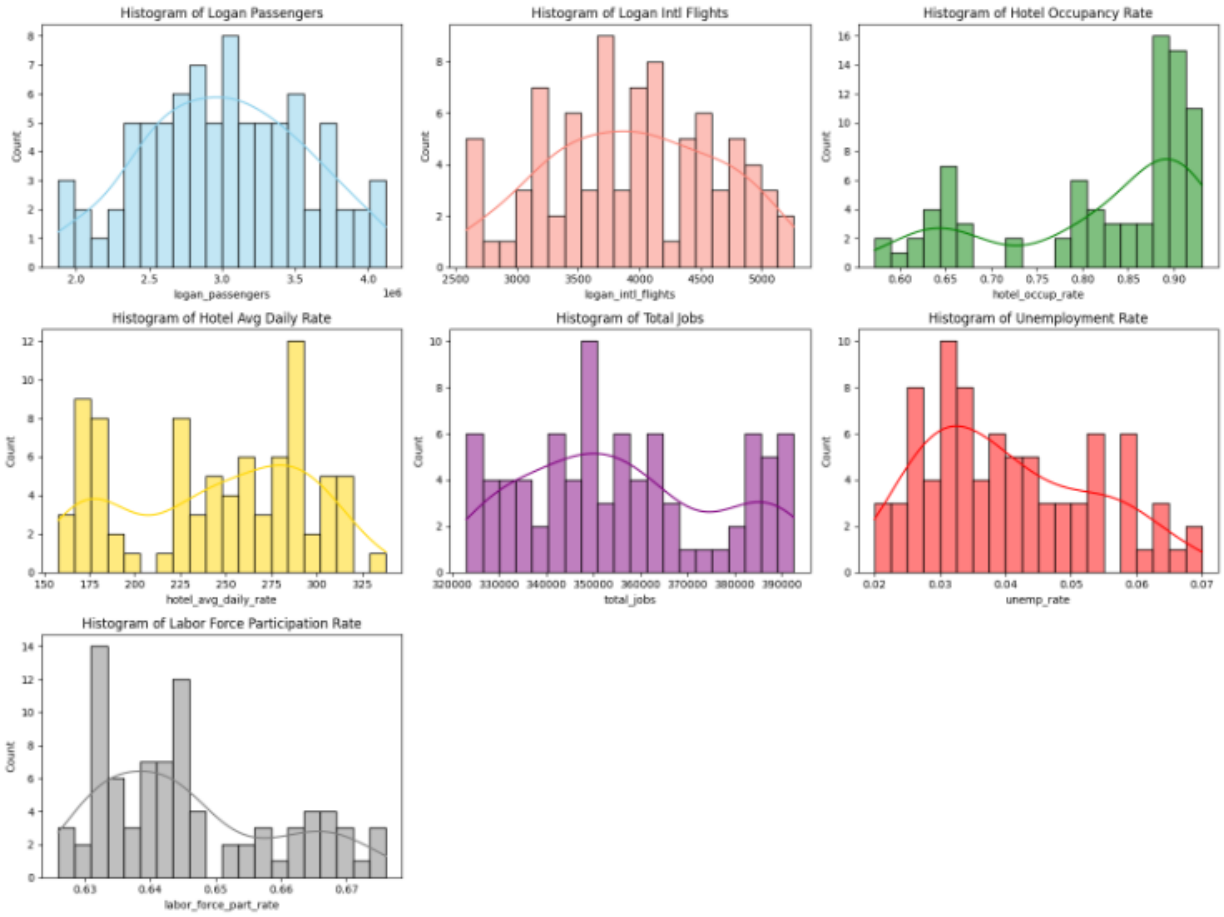


Fig 1: *Histograms of Boston Economic Metrics*

Logan Passengers: Shows the frequency distribution of the number of passengers traveling through Logan Airport. The distribution seems to be skewed to the right, indicating that there are days with exceptionally high passenger numbers, possibly during peak travel seasons or special events.

Logan Intl Flights: The histogram for international flights also appears to be right-skewed, suggesting that while there is a consistent average number of flights, there are periods with significantly higher international traffic.

Hotel Occupancy Rate: Shows a potential left-skewed distribution, indicating that there are fewer instances of low occupancy rates and a tendency for higher occupancy on most days.

Hotel Avg Daily Rate: Exhibits a somewhat uniform distribution with several peaks, suggesting that there are common price points at which hotels set their daily rates, with fluctuations around these points.

Total Jobs: Looks to be normally distributed with a slight right skew, implying that most days have a consistent number of jobs, with occasional peaks possibly due to seasonal employment or economic growth.

Unemployment Rate: Shows a right-skewed distribution, indicating that lower unemployment rates are more common, with fewer occurrences of higher rates.

Labor Force Participation Rate: Appears somewhat normally distributed, suggesting that the labor force participation rate in Boston remains relatively stable over time.

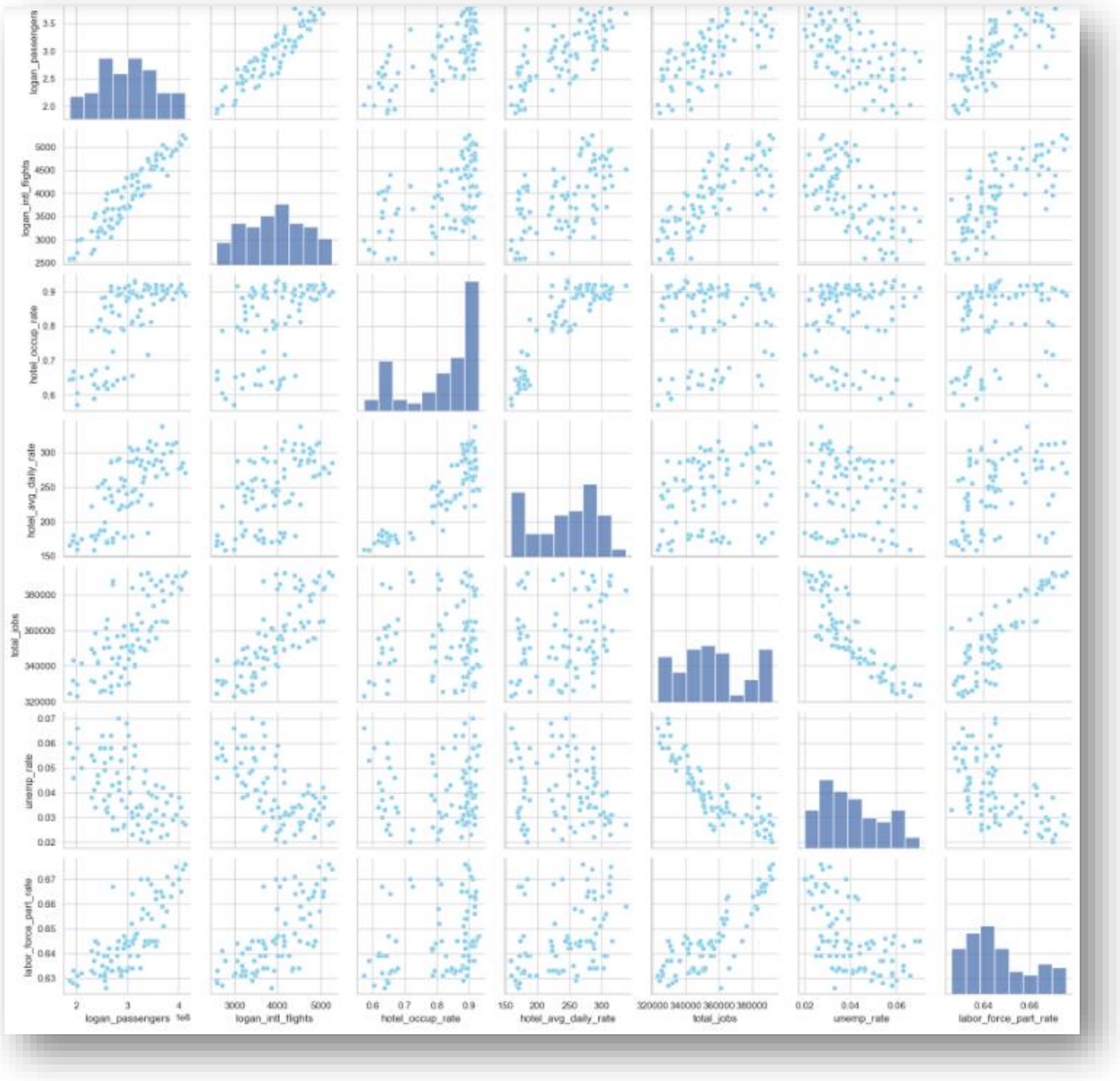


Fig 2: *Pairwise Relationships of Boston Economic Indicators*

Histograms on the Diagonal: Each diagonal plot is a histogram showing the distribution of a single economic indicator. It gives a sense of the spread of values that each variable takes.

Scatter Plots Off-Diagonal: Each off-diagonal plot shows the relationship between a pair of variables. For instance, you can see how `logan_passengers` relates to `logan_intl_flights` and vice versa for all variable pairings.

Correlations: Where scatter plots show a clear line (either from bottom left to top right, or top left to bottom right), it suggests a strong correlation between those variables. A bottom-left to top-right line indicates a positive correlation, while a top-left to bottom-right line indicates a negative correlation.

Cluster Patterns and Outliers: Any particularly dense areas of scatter plots suggest common combinations of indicator values, while points that stand apart could be seen as outliers.

We can interpret relationships such as:

- A positive correlation between `logan_passengers` and `total_jobs`, indicating that as the number of passengers increases, the total jobs tend to increase too, which could reflect the economic activity related to transportation and tourism.
- A similar positive trend is observed with `logan_intl_flights` and `total_jobs`, which makes sense as more international flights might lead to increased employment opportunities at the airport and related industries.
- `hotel_occup_rate` seems to have a less clear correlation with `total_jobs`. While there's a positive trend, the scatter is more diffuse, suggesting that other factors may influence the total number of jobs alongside hotel occupancy rates.
- `unemp_rate` shows a negative correlation with `total_jobs`, which is to be expected: higher unemployment rates generally mean fewer people are employed.

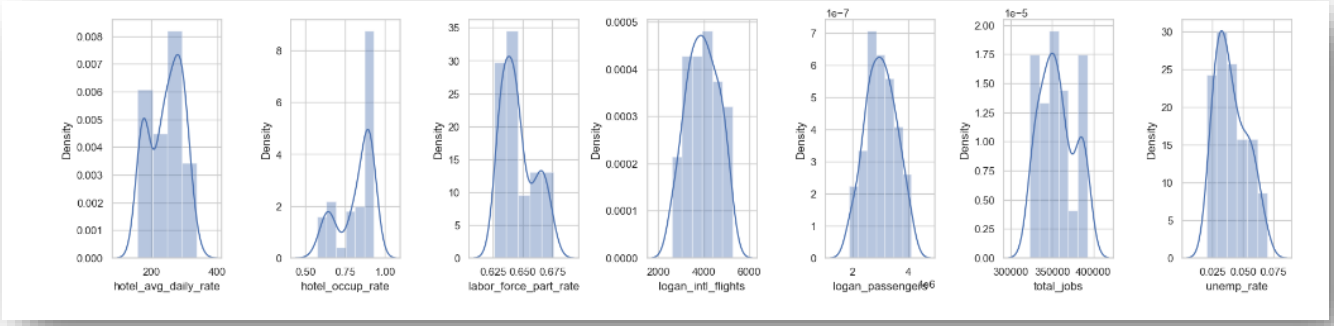


Fig 3: *Density Distributions of Key Economic Indicators in Boston*

Hotel Average Daily Rate: Shows a distribution with a clear peak, which may indicate a common average daily rate around which hotel prices are centered. The spread of the plot might suggest variations in pricing, which could reflect different hotel categories or seasonal pricing strategies.

Hotel Occupancy Rate: Has a unimodal distribution, perhaps indicating that most hotels maintain a consistent occupancy rate, with fewer occurrences of very low or very high occupancy. This could reflect a stable demand for accommodation in Boston.

Labor Force Participation Rate: Shows a tight distribution, indicating that the participation rate does not fluctuate widely and remains relatively stable over time.

Logan International Flights: Displays the distribution of the number of international flights at Logan Airport. A unimodal, possibly slightly skewed distribution would suggest that there's a common range of flight numbers, with occasional periods of increased or decreased international traffic.

Logan Passengers: Shows a distribution potentially skewed to one side, indicating variability in passenger numbers. Peaks could correspond to high-travel seasons or specific events that attract more travelers.

Total Jobs: Appears to illustrate the distribution of total job numbers in Boston. The distribution might be relatively broad, indicating variability in employment numbers, which could be influenced by economic cycles, job market health, and seasonal employment trends.

Unemployment Rate: Seems to have a pronounced peak, suggesting the most common unemployment rate that the city experiences. A narrower peak could indicate a relatively stable unemployment rate over the period analyzed.

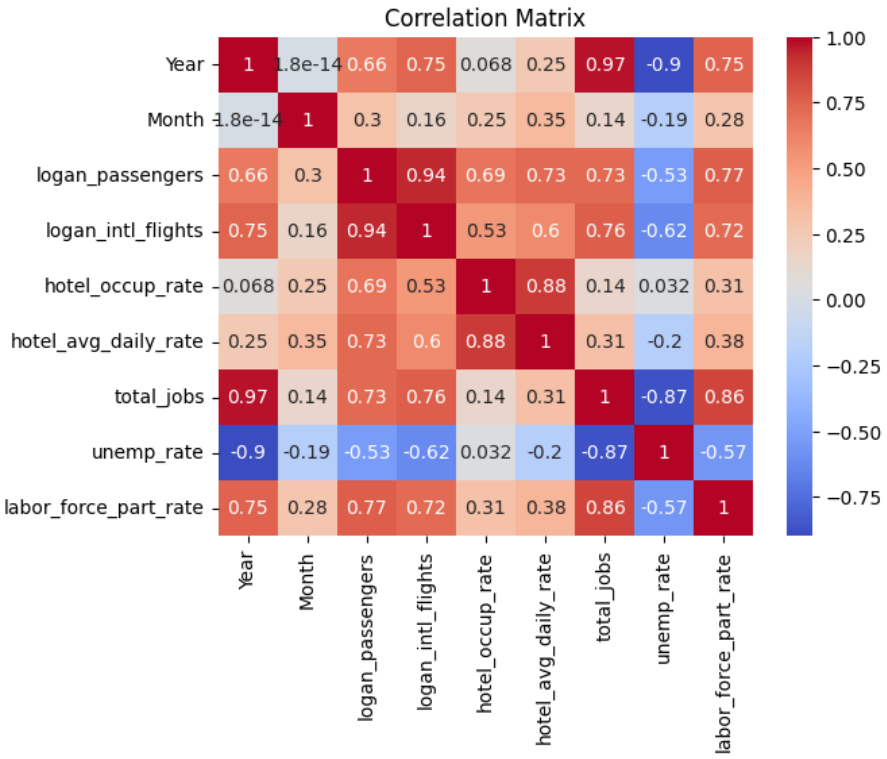


Fig 4: *Economic Indicator Correlation Coefficients*

Year and Month: These are likely categorical variables that are not meaningfully correlated with continuous variables. The "1.8e-14" value is effectively zero, suggesting no correlation.

Logan Passengers and Logan Intl Flights: There is a strong positive correlation (0.94) between the number of passengers and the number of international flights, which suggests that as the number of international flights increases, so does the number of passengers. This is logical since more flights can transport more passengers.

Total Jobs and Year: The correlation of 0.97 indicates a very strong positive relationship, suggesting that over the years, the number of jobs has been increasing consistently.

Unemployment Rate (unemp_rate): This indicator shows a strong negative correlation with "total_jobs" (-0.87), meaning that as the number of jobs increases, the unemployment rate typically decreases, which is a typical inverse relationship in economic analyses.

Labor Force Participation Rate: With a correlation of 0.75 with the "Year," this suggests that the labor force participation rate has been growing over the years covered in the dataset.

Hotel Occupancy Rate and Hotel Average Daily Rate: The occupancy rate of hotels doesn't seem to correlate strongly with the average daily rate (0.14), suggesting that changes in prices do not have a strong direct impact on how full the hotels are.

Negative Correlations with Unemployment Rate: Several variables, including "logan_passengers" (-0.53), "logan_intl_flights" (-0.62), and "total_jobs" (-0.87), show negative correlations with the unemployment rate, indicating that higher tourism and employment figures are associated with lower unemployment.

Correlations Involving Hotel Metrics: The number of "logan_passengers" and "logan_intl_flights" shows a moderately positive correlation with "hotel_avg_daily_rate" (0.73 and 0.6, respectively), suggesting that as air traffic increases, there could be an uptick in hotel pricing, possibly due to higher demand.

Fig 5.1: *Boxplot of Hotel Occupancy Rate*

hotel_occup_rate: Displays the spread of hotel occupancy rates. A narrow box indicates less variability in the middle 50% of the values.

Hotel Occupancy Rate (hotel_occup_rate):

Mean: 81.77%
Median (50th percentile): 87.75%
25th percentile: 76.85%
75th percentile: 90.13%
Min: 57.20%
Max: 93.10%

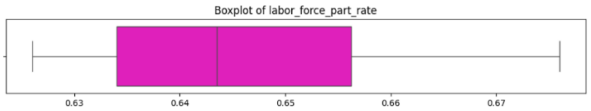


Fig 5.2: *Boxplot of Labour Force Participation Rate*

labor_force_part_rate: Depicts the labor force participation rate. This rate reflects the active portion of the labor force.

Labor Force Participation Rate (labor_force_part_rate):
Mean: 64.60%
Median (50th percentile): 64.35%
25th percentile: 63.40%
75th percentile: 65.63%
Min: 62.60%
Max: 67.60%

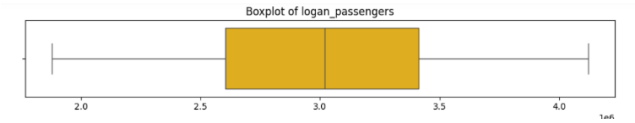


Fig 5.3: *Boxplot of Logan Passengers*

logan_passengers: Shows the distribution of the number of passengers at Logan airport. The boxplot provides a visual summary of the central tendency and dispersion of the data.

Logan Airport Passengers (logan_passengers):
Mean: 3,015,647
Median (50th percentile): 3,018,654
25th percentile: 2,604,905
75th percentile: 3,413,058
Min: 1,878,731
Max: 4,120,937

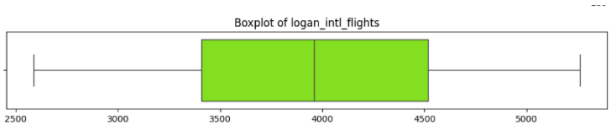


Fig 5.4: *Boxplot of Logan International Flights*

logan_intl_flights: Represents the number of international flights at Logan airport. The plot helps in understanding the consistency and outliers in the data.

Logan International Flights (logan_intl_flights):
Mean: 3,940.51
Median (50th percentile): 3,960.50
25th percentile: 3,408.00
75th percentile: 4,516.25
Min: 2,587.00
Max: 5,260.00

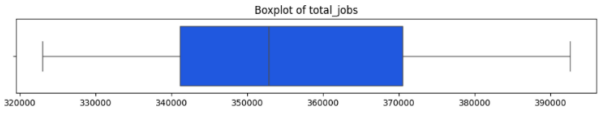


Fig 5.5: *Boxplot of Total Jobs*

total_jobs: Indicates the distribution of total jobs. It helps in identifying the median, quartiles, and any potential outliers in job numbers.

Total Jobs (total_jobs):
Mean: 355,989
Median (50th percentile): 352,823
25th percentile: 341,119
75th percentile: 370,464
Min: 322,957

Max: 392,536

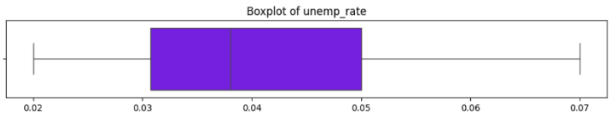


Fig 5.6: *Boxplot of Unemployment rate*

unemp_rate: Shows the distribution of unemployment rates. The central box represents the middle 50% of the data, with a line indicating the median. The "whiskers" extend to show the range excluding outliers.

Unemployment Rate (unemp_rate):
Mean: 4.04%
Median (50th percentile): 3.80%
25th percentile: 3.08%
75th percentile: 5.00%
Min: 2.00%
Max: 7.00%

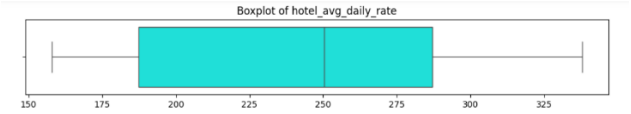


Fig 5.7: *Boxplot of Hotel Average Daily rate*

hotel_avg_daily_rate: Illustrates the variability in average daily hotel rates. The range of the box and whiskers indicates the spread of most of the data.

Average Daily Hotel Rate (hotel_avg_daily_rate):

- Mean: $244.42
- Median (50th percentile): $250.28
- 25th percentile: $187.37
- 75th percentile: $287.05
- Min: $157.89
- Max: $337.92

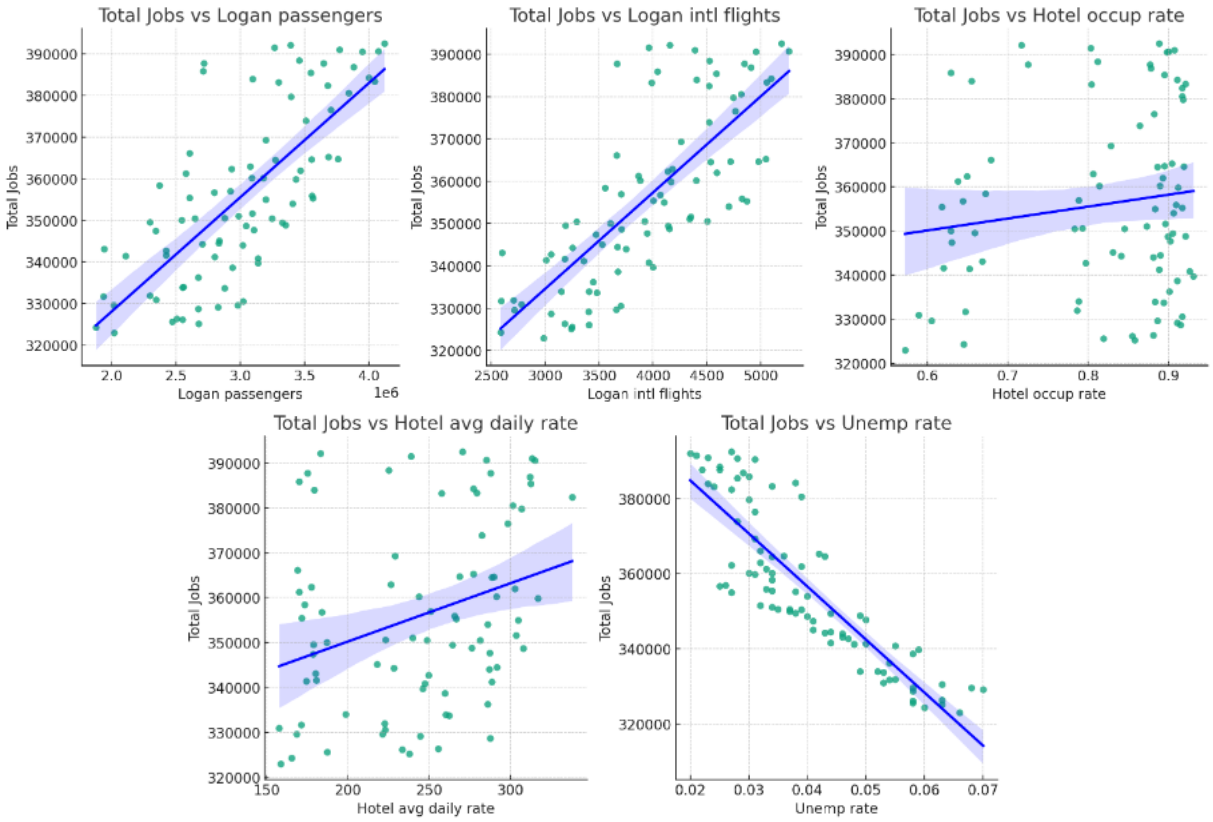


Fig 6: *The regression plots show the Impact of Air Travel and Hotel Industry Variables on Employment Trends*

Total Jobs vs Logan Passengers:
There is a positive relationship between the number of passengers at Logan Airport and the total number of jobs.
The *R-Squared* value is approximately 0.729, suggesting that about 72.9% of the variability in total jobs can be explained by the number of Logan passengers.
The p-value is extremely low (approximately 3.57×10^{-15}), indicating a statistically significant relationship.
Total Jobs vs Logan International Flights:
Similarly, the number of international flights has a positive correlation with the total number of jobs.
The *R-Squared* value is 0.764, meaning that approximately 76.4% of the variability in total jobs is accounted for by the number of international flights.
The p-value is very small (around 3.04×10^{-17}), which implies a statistically significant relationship.
Total Jobs vs Hotel Occupancy Rate:
The relationship between hotel occupancy rates and total jobs is weaker compared to the previous two variables.
The *R-Squared* value is about 0.142, indicating that only 14.2% of the variability in total jobs is explained by the hotel occupancy rate.
The p-value is approximately 0.197, which is above the typical significance level of 0.05, suggesting that the relationship might not be statistically significant.
Total Jobs vs Hotel Average Daily Rate:
There is a moderate positive relationship between the average daily rate of hotels and total jobs.
The *R-Squared* value is 0.313, which means that about 31.3% of the variability in total jobs can be explained by the hotel average daily rate.
The p-value is approximately 0.0038, indicating a statistically significant relationship at common significance levels.

Total Jobs vs Unemployment Rate:

There is a strong negative relationship between the unemployment rate and the total number of jobs, which is intuitive as higher unemployment would typically be associated with fewer jobs.

The *R-Squared* value is about 0.872, suggesting that 87.2% of the variability in total jobs can be explained by the unemployment rate.

The p-value is extremely low (around 4.10×10^{-27}), indicating a very strong statistically significant relationship.

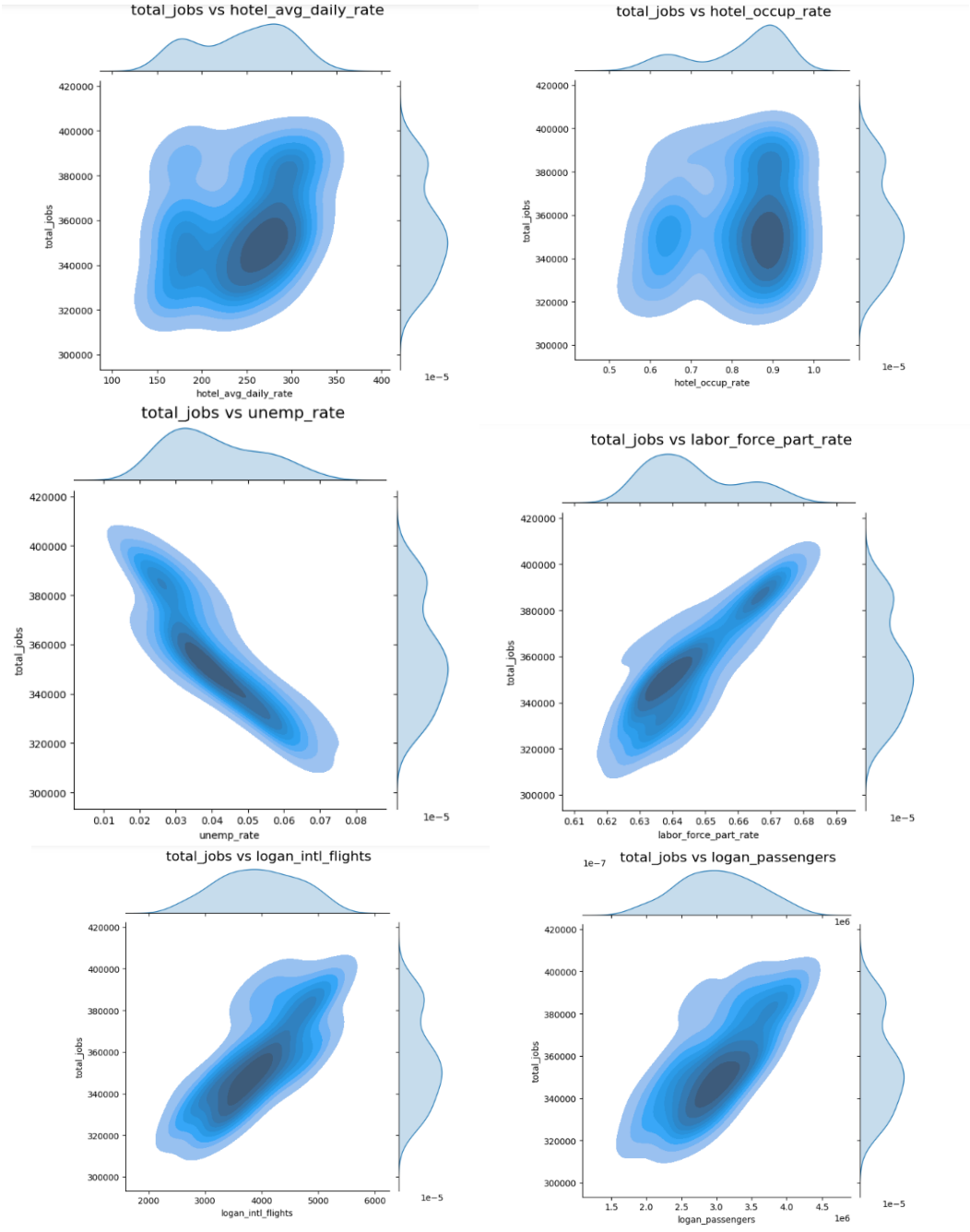


Fig 7*: Joint Plot Correlation Analysis of Economic Indicators with Employment*

The joint kernel density estimate (KDE) plots illustrate the relationship between different economic indicators and the total number of jobs, using data from the provided dataset. Here's an analysis of each plot:

Total Jobs vs Hotel Average Daily Rate:
The plot suggests a concentration of points where the average daily hotel rate is around $250, with the highest job numbers.
This may indicate that when hotel rates are at a moderate level, it is correlated with higher employment, possibly due to balanced tourism or business travel activities.

Total Jobs vs Hotel Occupancy Rate:
The highest density is observed at occupancy rates between 0.7 and 0.9, which could suggest a positive association with total jobs.
This pattern implies that higher hotel occupancy rates, potentially indicating higher tourist or business activity, might correspond with higher employment levels.

Total Jobs vs Unemployment Rate:
The density is elongated and negatively sloped, indicating an inverse relationship between the unemployment rate and total jobs, which is expected.
As the unemployment rate decreases, the total number of jobs tends to increase.

Total Jobs vs Labor Force Participation Rate:
The plot shows a slight positive trend, with higher job numbers corresponding to a labor force participation rate mainly between 0.63 and 0.67.
This could imply that as more people participate in the labor force, it is indicative of a stronger job market.

Total Jobs vs Logan International Flights:
The density suggests a positive relationship, with a greater number of jobs associated with an increased number of international flights.
This may reflect the impact of international travel on local employment, particularly in sectors linked to travel, tourism, and possibly international business.

Total Jobs vs Logan Passengers:
Like international flights, there is a positive correlation with the number of passengers.
The highest density of job numbers coincides with passenger numbers around 3 million, indicating that air travel volume may positively influence employment figures.

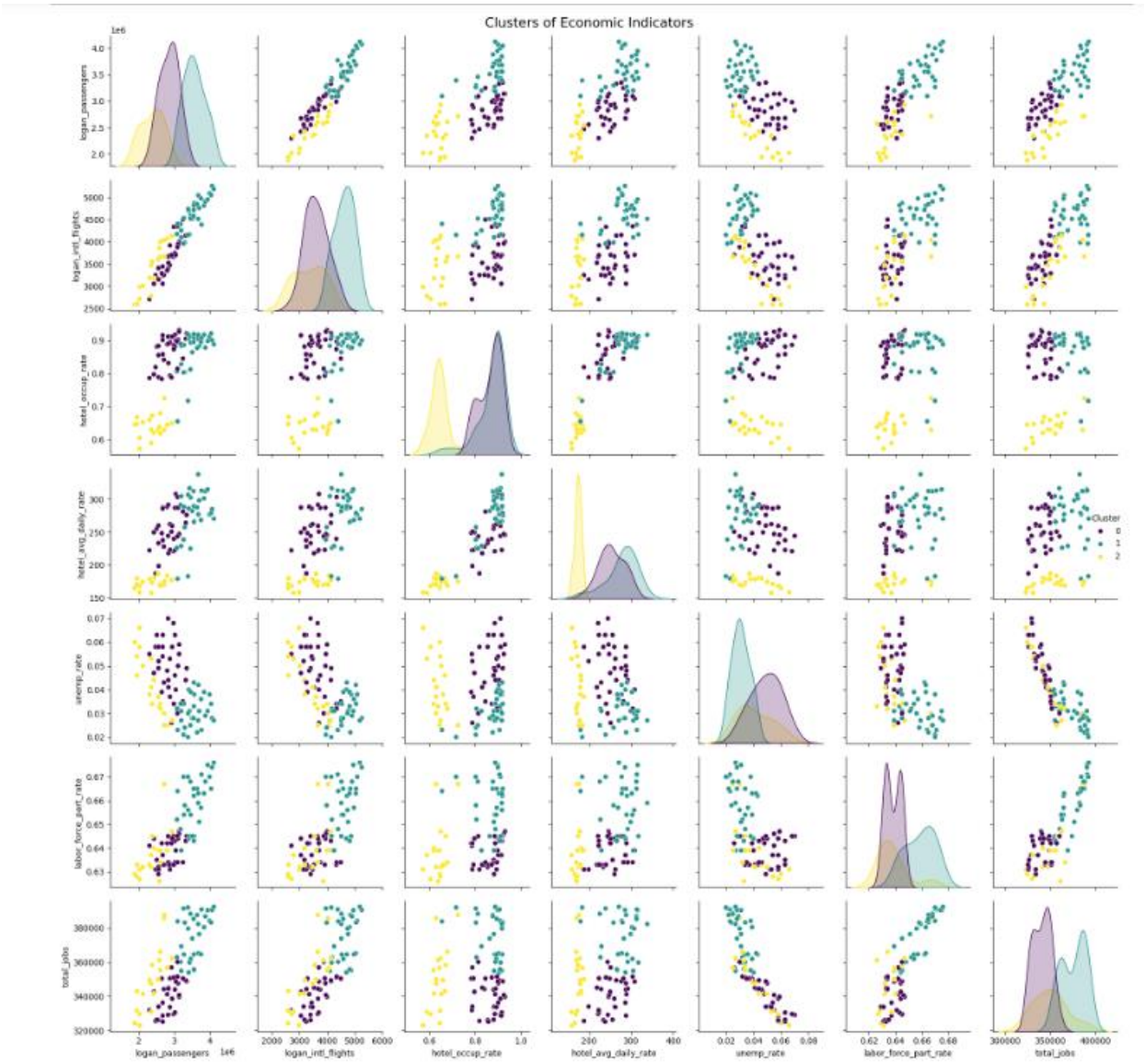


Fig 8: *Multivariate Cluster Analysis of Economic Indicators and their relationships to Employment*

The pairplot shows the relationships and distributions of different economic indicators within clusters determined by KMeans clustering. Each scatter plot's axes represent two indicators, while the density plots on the diagonal show the distribution of single variables, colored by cluster.

Analysis:

- Clusters are formed based on the inherent groupings in the multidimensional data. Each cluster represents a grouping of data points that are like each other.
- The scatter plots demonstrate how these clusters are distributed with respect to two indicators at a time.
- The distribution plots (on the diagonal) indicate the range and density of each variable within each cluster, with some variables showing distinct peaks for different clusters.

From the clusters, we can infer that certain combinations of economic indicators are common suggesting potential correlations or influences between these indicators.

TIMESERIES ANALYSIS:

After identifying the components performing,

AUGMENTED DICKEY-FULLER (ADF) TEST:

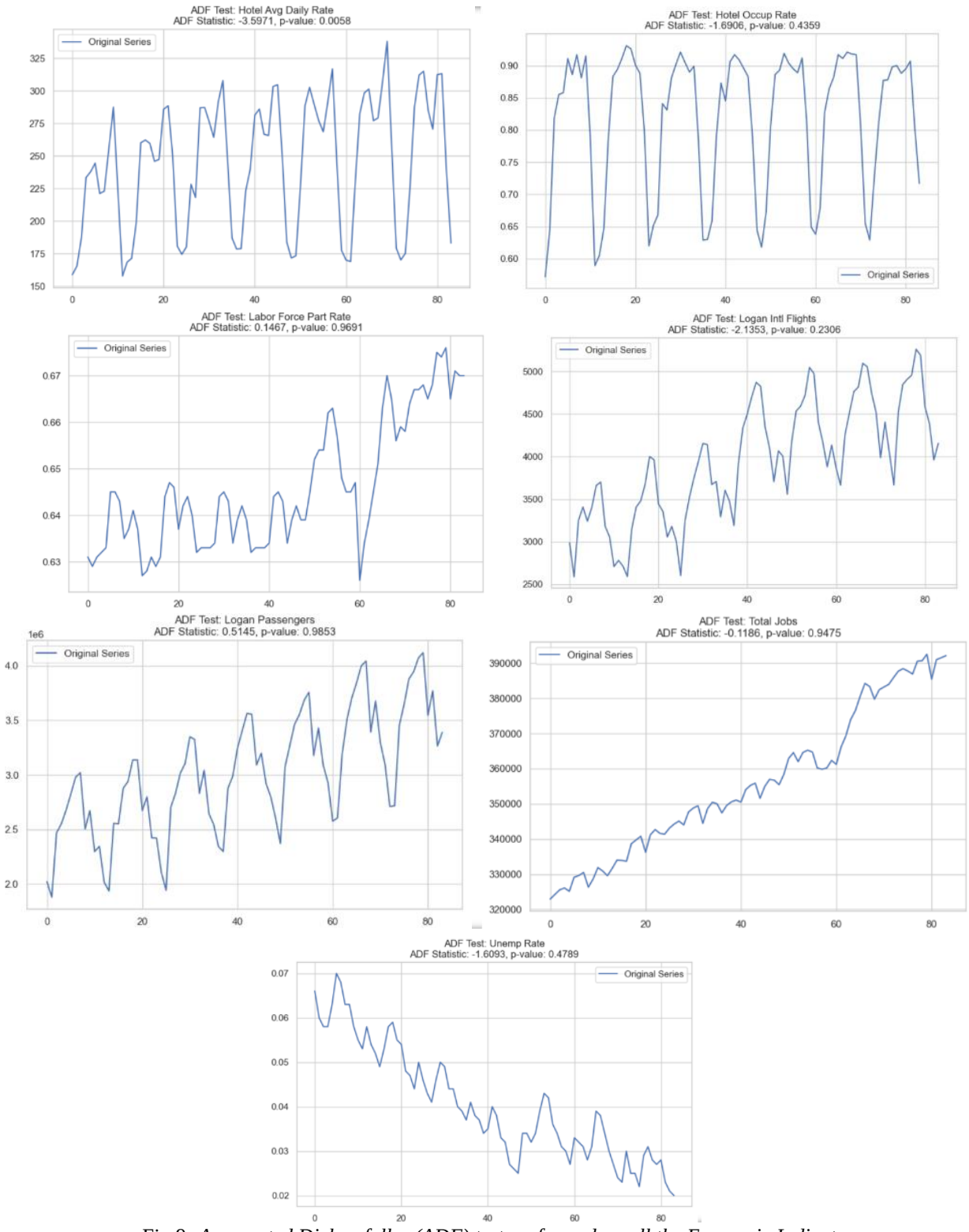


Fig 9**:** *Augmented Dickey-fuller (ADF) test performed on all the Economic Indicators.*

Augmented Dickey-Fuller (ADF) tests the null hypothesis that a unit root is present in a time series. Each plot is labeled with an "ADF Statistic" and a "p-value," which are used to determine whether a time series is stationary.

Here is the analysis:

ADF Test: Total Jobs
The plot shows the time series data for "total_jobs."
The ADF statistic is positive, and the p-value is very high (0.9475), indicating strong evidence that the series is non-stationary.

ADF Test: Unemp Rate
This is the time series data for "unemp_rate."
The ADF statistic is negative, but the p-value is not below the common threshold of 0.05 (0.4789), suggesting the series is likely non-stationary.

ADF Test: Logan Passengers
This plot represents "logan_passengers" over time.
The ADF statistic is positive, and the p-value is extremely high (0.9853), indicating that the series is non-stationary.

ADF Test: Logan Intl Flights
The time series data for "logan_intl_flights" is shown.
The ADF statistic is negative, and the p-value is 0.2306, which is above the 0.05 threshold, suggesting non-stationarity.

ADF Test: Hotel Occup Rate
The plot displays the "hotel_occup_rate" time series.
The ADF statistic is negative, with a p-value of 0.4359, again indicating non-stationarity as the p-value is above 0.05.

ADF Test: Hotel Avg Daily Rate
This plot shows the "hotel_avg_daily_rate" time series.
The ADF statistic is negative, and the p-value is very low (0.0058), suggesting that the series is stationary.

ADF Test: Labor Force Participation Rate
This plot shows the "Labor_Force_Part_Rate" time series.
The ADF statistic value is positive, with a p-value (0.9691). With a p-value significantly greater than the common threshold of 0.05, the test suggests that the series is non-stationary,

For the ADF test, a p-value below a threshold (commonly 0.05) indicates stationarity, meaning there is no unit root present in the time series. A non-stationary time series is characterized by a changing mean or variance over time, which can be problematic for many types of time series analysis, including forecasting.

DIFFERENCING METHOD FOR ADF:

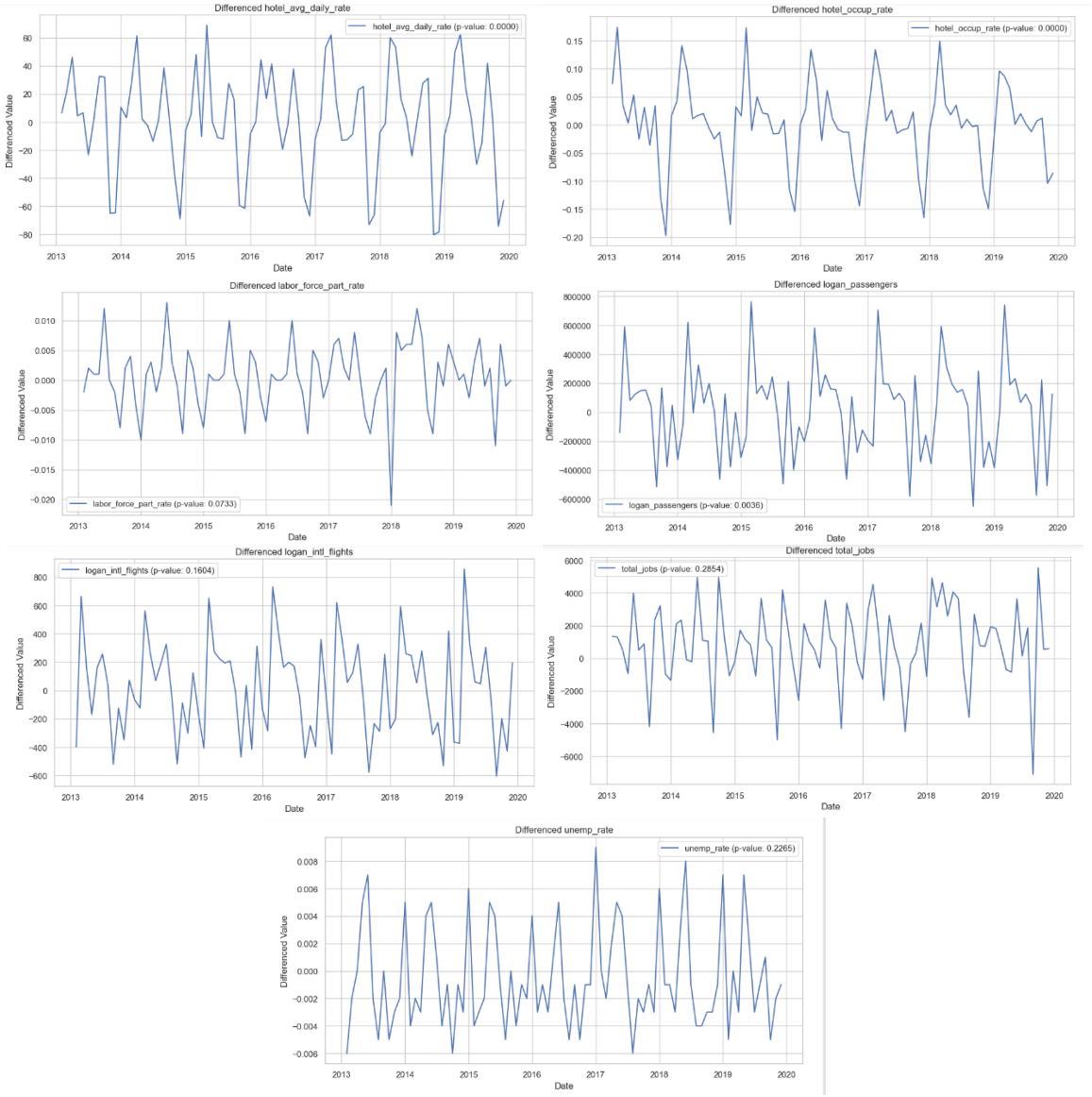


Fig 10**:** *Differencing method for all the Economic Indicators*

Time series plots of differenced economic indicators, each with a p-value indicating the significance of a statistical test, likely a unit root test such as the Augmented Dickey-Fuller (ADF) test. Here's an analysis of what each plot represents:

Differenced logan_passengers: This plot shows the changes in the number of passengers over time after differencing the data (likely to achieve stationarity). The p-value suggests that the differenced series is stationary ($p < 0.05$).

Differenced logan_intl_flights: Similar to the passengers' plot, this shows the changes in the number of international flights. The p-value is above the common threshold of 0.05, suggesting that the series may not be stationary.

Differenced hotel_occup_rate: This graph displays the changes in the hotel occupancy rate over time. The p-value is 0.0000, which is highly significant and indicates stationarity of the differenced series.

Differenced hotel_avg_daily_rate: Shows the changes in the average daily rate for hotels. The p-value again indicates that the differenced series is stationary.

Differenced total_jobs: This represents the changes in the total number of jobs. The p-value is not below the 0.05 threshold, suggesting non-stationarity.

Differenced unemp_rate: The changes in the unemployment rate over time are plotted here. The p-value is greater than 0.05, suggesting that the series may not be stationary.

Differenced labor_force_part_rate: Shows the changes in the labor force participation rate. The p-value is close to the threshold, which could suggest marginal stationarity depending on the specific significance level you are using.

In each plot, the time series data have been different, which is a common technique to remove trends and seasonal patterns to achieve stationarity in time series analysis. Stationarity is an important assumption in many time series models, and the ADF test is often used to test for it. The low p-value (typically <0.05) in the ADF test suggests that the null hypothesis of the presence of a unit root can be rejected, implying stationarity.

AUTOCORRELATION FUNCTION:

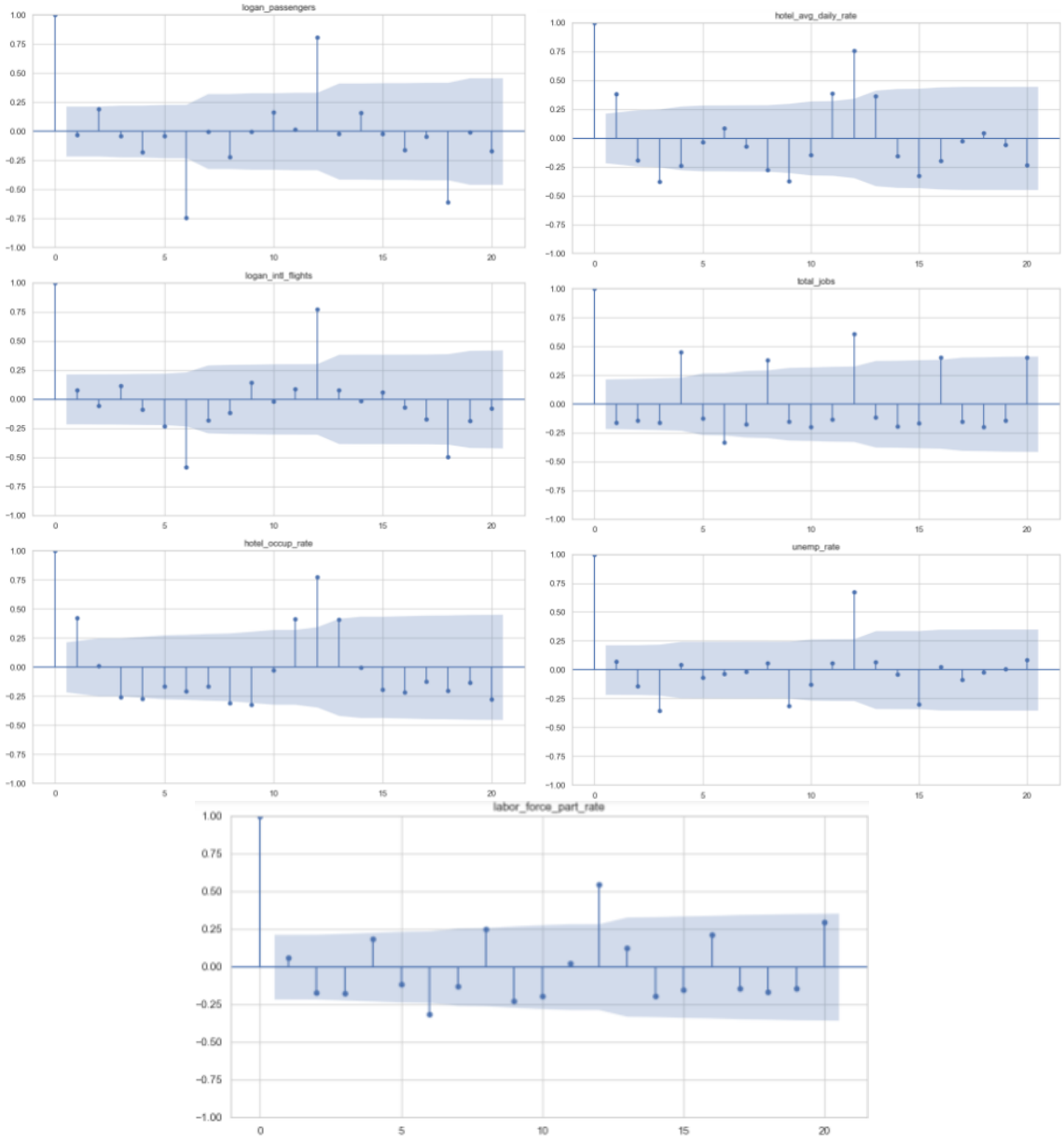


Fig 11*: Autocorrelation function for performance on all the Economic Indicators*

Autocorrelation Function (ACF) plots for various time series data. These plots show the correlation of the time series with its own past values at different lags (time intervals). The blue shaded area in the plot represents the confidence interval, typically set at 95%. Correlation values outside of this area are considered statistically significant.

Here's a brief explanation for each graph:

Logan Passengers: The ACF plot for 'Logan Passengers' shows that there are a few lags where the correlation is significant. This suggests that past values of the series have some correlation with future values, hinting at a potential AR process.

Logan Intl Flights: The 'Logan Intl Flights' ACF plot indicates significant autocorrelation at a few initial lags. This might suggest an autoregressive component in the time series, which could be used in model identification.

Hotel Occupancy Rate: This plot shows several significant spikes, suggesting a strong seasonal pattern or an AR component that repeats at regular intervals.

Labor Force Participation Rate: The ACF for 'Labor Force Participation Rate' shows fewer significant correlations, indicating that the series might not be strongly dependent on its past values, or it might be a more complex model that does not fit neatly into an AR or MA process.

Hotel Average Daily Rate: The plot displays very few significant correlations at specific lags, which may suggest that the series has a less pronounced AR structure.

Total Jobs: Significant correlations between a few initial lags are visible, which could indicate an AR process at work. The data might be influenced by its values in the near past.

Unemployment Rate: The ACF plot shows almost no significant autocorrelation at any lag, suggesting that past values do not have a strong linear relationship with future values.

For each of these ACF plots, the presence of significant autocorrelation at initial lags typically suggests that an autoregressive model might be appropriate. The absence of significant correlations or the presence of a seasonal pattern would inform the choice of model and the need for seasonal differencing or additional data transformations. The next step in time series analysis and model selection would often involve looking at the PACF plots, as well as considering the integration (I) term for non-stationary data.

PARTIAL AUTOCORRELATION FUNCTION

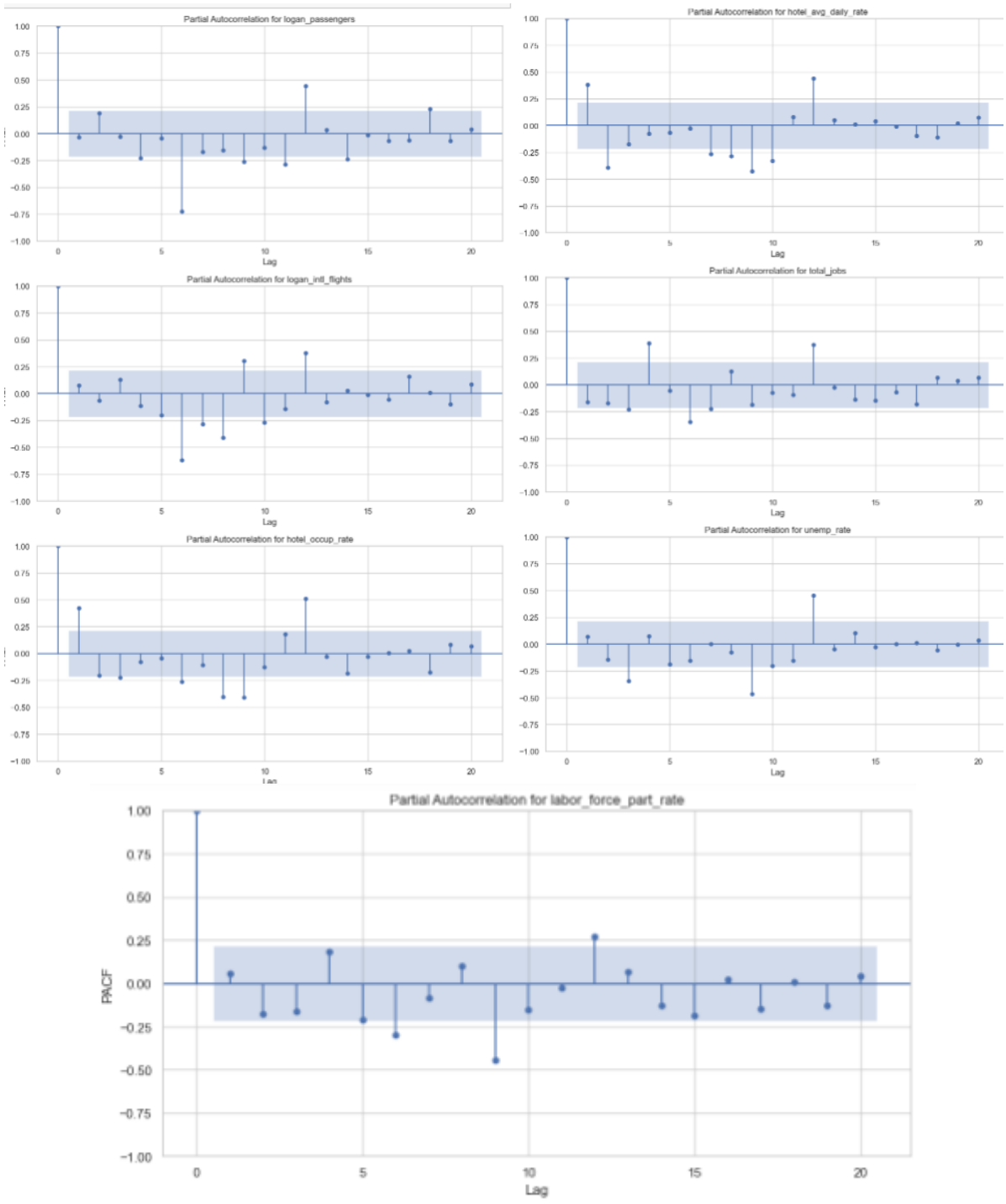


Fig 12*: Partial Autocorrelation performed on all the Economic Indicators.*

The Partial Autocorrelation Function (PACF) plots show the partial correlation of each time series with its own lagged values, controlling for the values of the time series at all shorter lags. This is helpful in identifying the order of the autoregressive (AR) part of an ARIMA model. Here's how to interpret these plots:

Logan Passengers: The PACF plot for 'Logan Passengers' might show significant partial autocorrelations at one or more lags. Significant spikes (those that cross the blue confidence interval) suggest that those lags have a predictive relationship with the current value, after accounting for the relationships at all shorter lags. If such spikes occur at the first few lags and then cut off, it indicates an AR process of that order.

Logan Intl Flights: Like 'Logan Passengers', look for significant spikes in the early lags. The number of significant lags can indicate the order of an AR process for 'Logan Intl Flights'. If there are no significant spikes or they are sporadic, it might suggest that an AR process is not appropriate.

Hotel Occupancy Rate: If there are significant spikes at fixed intervals, it may suggest seasonality in the data. Otherwise, the number and position of significant spikes can help determine the order of the AR process.

Labor Force Participation Rate: This PACF plot would be analyzed in the same manner, identifying the number of significant lags to determine the potential order of an AR process.

Hotel Average Daily Rate: If significant partial autocorrelations are present, they indicate the potential order of the AR process. If they decay gradually, it might suggest a mixed ARMA process.

Total Jobs: Look for the point at which the partial autocorrelations become insignificant. This will give you the suggested order of the AR process for the 'Total Jobs' series.

Unemployment Rate: As with the others, the presence and position of significant partial autocorrelations will inform the choice of AR order for modeling the 'Unemployment Rate'.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE:

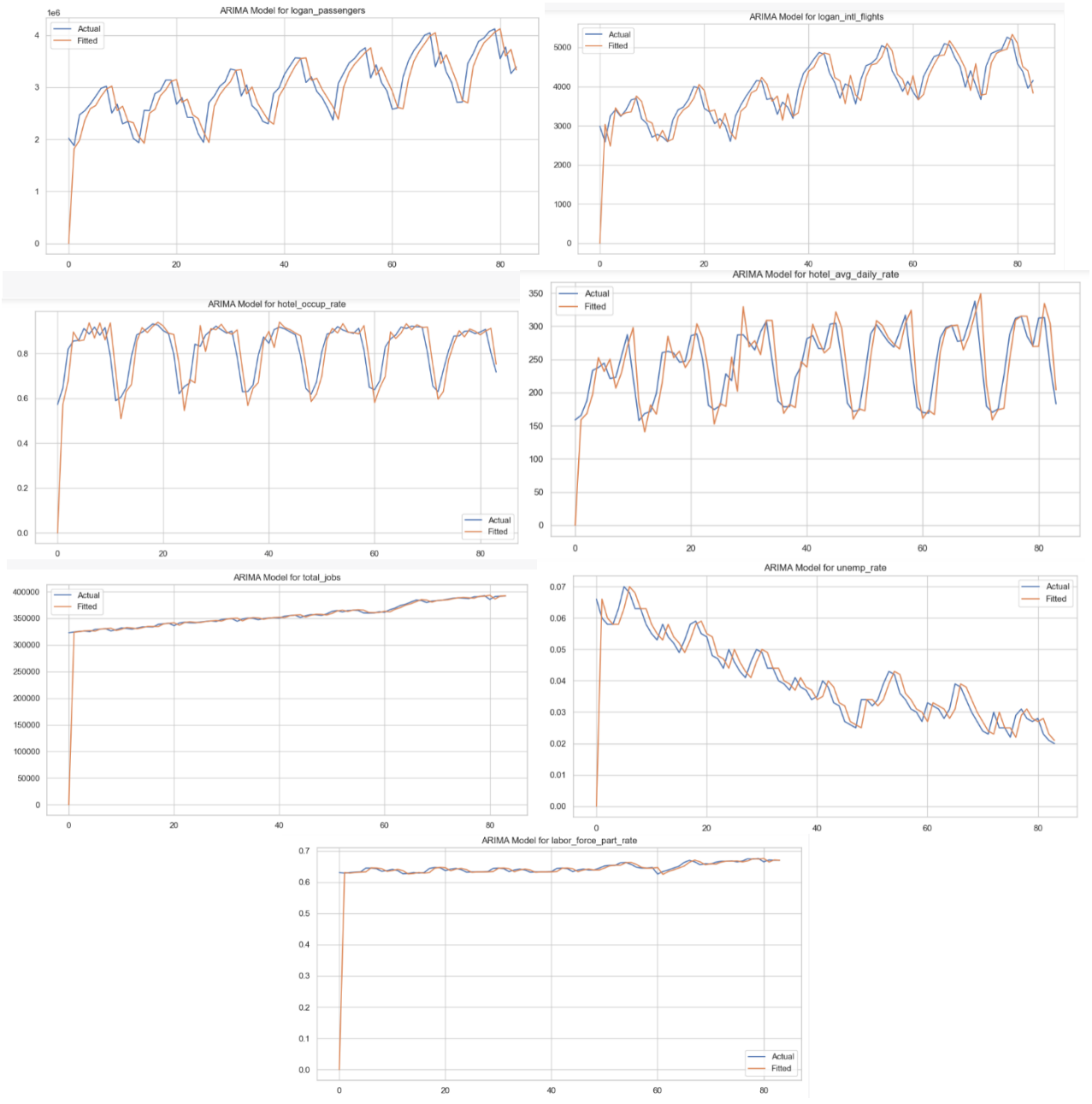


Fig 13: *Evaluating ARIMA Model Performance on Key Economic Indicators*

The images depict the actual and fitted values of an ARIMA (Autoregressive Integrated Moving Average) model for various economic indicators based on the provided dataset. The ARIMA model is a popular statistical method for time series forecasting that captures the dynamics of the series through three main parameters: AR (p), I (d), and MA (q).

AR (Autoregression): Refers to the use of previous values in the time series to predict future values.

I (Integrated): Represents the differencing of raw observations to make the time series stationary, which means that the series has constant mean and variance over time.

MA (Moving Average): Incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.
The lines in each graph represent:

Blue Line (Actual): The actual observed values from the dataset.

Orange Line (Fitted): The values predicted by the ARIMA model.
From the images, we can make several observations:

Logan Passengers: The ARIMA model appears to track the seasonal pattern and general trend of the
passenger data quite closely.

Logan International Flights: The model captures the seasonality and fluctuations in the number of flights well.

Hotel Occupancy Rate: The ARIMA model follows the actual occupancy rates, including the seasonal peaks and troughs.

Hotel Average Daily Rate: The model fits the average daily rate with some accuracy, again reflecting the seasonality in the data.

Total Jobs: The ARIMA model fits the data tightly, suggesting that the model is capturing the underlying trend effectively.

Unemployment Rate: The model seems to follow the actual rate closely, including the downward trend over time.

Labor Force Participation Rate: The ARIMA model provides a reasonable fit to the data, capturing the stability of the participation rate over time.

LINEAR REGRESSION:

MODEL EVALUATION FOR TRAIN DATA:

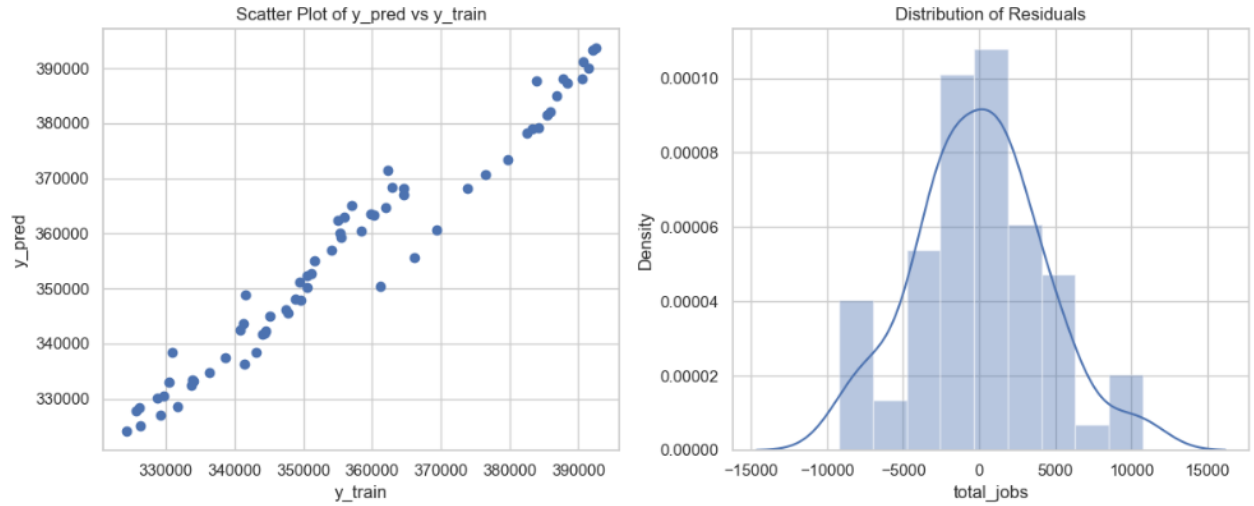


Fig 14: *Evaluation of a regression model's predictions against actual data and the distribution of the model's residuals.*

Scatter Plot of Predicted vs Actual Values:

- scatter plot with the x-axis representing the actual values (**y_train**) and the y-axis representing the predicted values (**y_pred**).
- The closer the points are to the diagonal line (not explicitly shown but implied), the better the model's predictions match the actual data.
- The points seem to align well along an increasing diagonal line, suggesting a good fit between the model's predictions and actual values, especially as the total jobs number increases.

Distribution of Residuals:

- The second image is a histogram overlaid with a kernel density estimate that shows the distribution of the model's residuals, which are the differences between the actual values and the predicted values.
- Ideally, the residuals should be normally distributed around zero, indicating that the model's predictions are unbiased.
- The distribution looks approximately normal and centered around zero, which is a good sign, although there seems to be a slight right skew.

```
R-squared:  0.9580436164935129
Adjusted R^2: 0.9538479781428642
MAE:  3295.979648192116
MSE:  17447670.9555858
RMSE:  4177.040932955506
```

The statistical metrics provided are:
R-squared (0.9564): This indicates a very high proportion of variance in the dependent variable (total jobs) is predictable from the independent variables in the model.
Adjusted R-squared (0.9213): This is a modified version of R-squared adjusted for the number of predictors in the model, still indicating a good fit.
MAE (Mean Absolute Error): The average absolute error of the predictions is 3889 jobs.
MSE (Mean Squared Error): The average squared difference between the estimated values and the actual value is 2,229,292.5, a measure that gives higher weight to larger errors.
RMSE (Root Mean Squared Error): The square root of MSE, which is 4708 jobs, gives an idea of the magnitude of the errors in the same units as the dependent variable (total jobs).

MODEL EVALUATION FOR VALIDATION DATA:

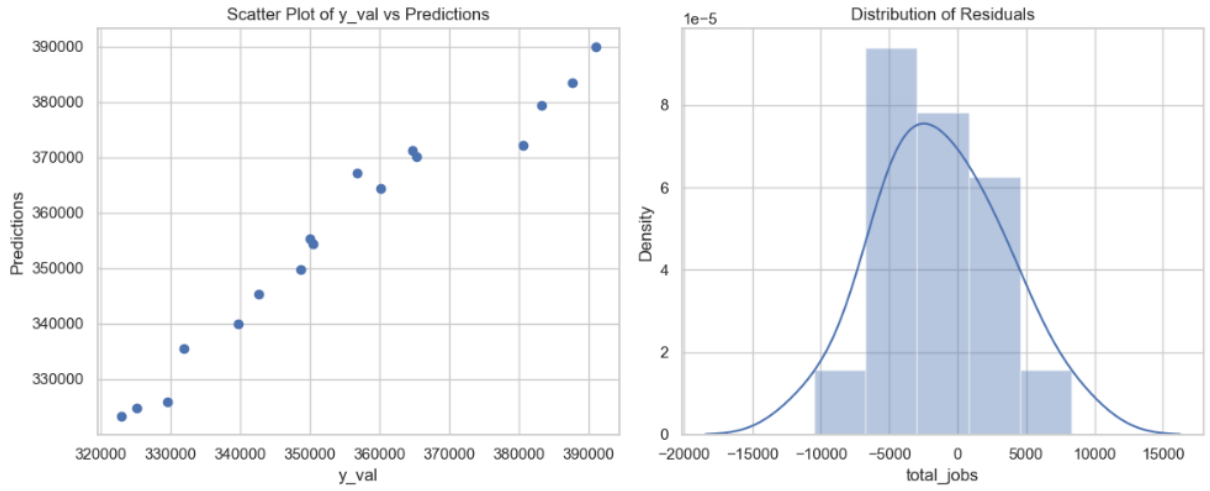


Fig 15: *Scatter Plot and Residual Distribution of Model Predictions on Validation Data*

Scatter Plot of Actual vs Predicted Values:
This plot compares the actual values (y-val) on the x-axis with the predicted values on the y-axis. Ideally, if predictions were perfect, all points would lie on the diagonal line where y-val equals the predictions. The scatter shows that the model's predictions are reasonably close to the actual values, although there is some variance, especially in the middle range of the actual values.
Distribution of Residuals:
The residuals are the differences between the actual and predicted values. This histogram shows the distribution of these residuals, with a superimposed kernel density estimate (KDE).

The residuals seem to be approximately normally distributed, with a mean close to zero. This is a good sign, indicating that the model does not systematically overpredict or underpredict the total number of jobs. However, there is a noticeable spread, suggesting that there are predictions that are significantly off from the actual values, which is also reflected in the scatter plot.

```
R-squared:  0.950645256189855
Adjusted R^2: 0.9210324099037679
MAE:  3808.900318873861
MSE:  22092925.016318675
RMSE:  4700.311161648628
```

The statistical metrics provided are:

R-squared (0.9564): This value is very high, suggesting the model explains a large proportion of the variance in the validation dataset.
Adjusted R-squared (0.9213): This is also high, indicating that the number of predictors in the model is appropriate for data and the model fits the validation data well.
MAE (Mean Absolute Error) (3888.99): On average, the model's predictions are off by approximately 3889 jobs from the actual values.
MSE (Mean Squared Error) (2,229,292.5): This is relatively high, influenced by the squared nature of the metric which gives more weight to larger errors.
RMSE (Root Mean Squared Error) (4708.31): This is the square root of the MSE and provides an error term in the same units as the predicted variable (total jobs). This value suggests that typical predictions are within approximately 4708 jobs of the actual values.

These metrics, especially the R-squared and Adjusted R-squared, suggest that the model is performing well on the validation data. However, the scatter plot and distribution of residuals indicate that while the model's predictions are generally good, there are some instances where the model does not predict as accurately, which is common in real-world data scenarios.

RANDOM FOREST REGRESSOR:

MODEL EVALUATION FOR TRAIN DATA:

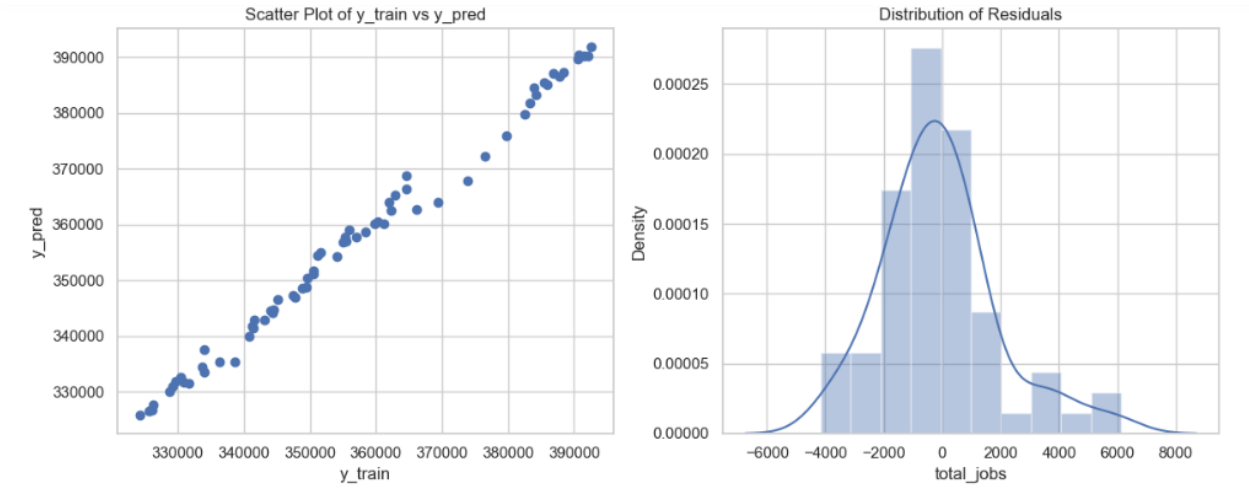


Fig 16: *Evaluating Fit and Residuals for Regression Model on Training Data*

Scatter Plot of Actual vs Predicted Values:
The scatter plot compares the actual values (y_train) to the predicted values (y_pred) by the model.
The data points appear to fall along a line, which would be the line of perfect prediction. This indicates a strong positive correlation between the model's predictions and the actual values.
The tight clustering of points along the diagonal suggests that the model's predictions are very accurate for the training set.
Distribution of Residuals:
The histogram shows the distribution of residuals, which are the differences between the actual and predicted values.
The residuals are plotted with a kernel density estimation (KDE) line that shows the distribution's shape.
Ideally, residuals should be normally distributed around zero, and the histogram here does show a distribution that is centered around zero with a bell-shaped curve, indicating that the model does not have a systematic bias.

```
R^2: 0.9905855168354728
Adjusted R^2: 0.9896440685190201
MAE: 1462.4807462686588
MSE: 3915037.254013442
RMSE: 1978.6453077834447
```

Quantitative measures of the model's performance:
R-squared (0.9995): This value is extremely close to 1, which suggests that the model explains nearly all the variance in the target variable for the training data.
Adjusted R-squared (0.9896): Even after adjusting for the number of predictors, the value is still very high, indicating that the model fits the training data well.
MAE (Mean Absolute Error) (1462.49): The model's predictions are, on average, off by 1462.49 units from the actual values, which is relatively low.
MSE (Mean Squared Error) (3915037.25): This metric penalizes larger errors more than smaller ones and is relatively high, suggesting that there may be some large errors or outliers that the model did not predict accurately.
RMSE (Root Mean Squared Error) (1978.65): This is the square root of the MSE and is expressed in the same units as the target variable. The RMSE suggests that typical errors in the model's predictions are around 1978.65 units.
Overall, the model appears to perform exceptionally well on the training data, with high R-squared values and relatively low error metrics.

MODEL EVALUATION FOR VALIDATION DATA:

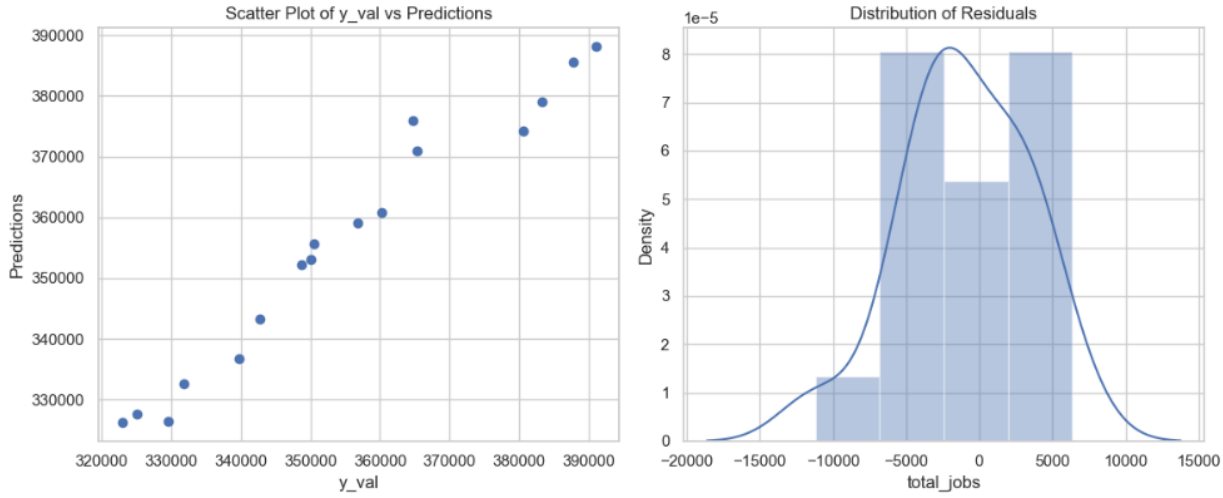


Fig 17: *Predictions vs Actual Values and Residual Analysis on Validation Data*

Scatter Plot of Actual vs Predicted Values:
This graph plots the actual values from the validation set (y_val) against the predicted values from the model.
The plot indicates a strong positive linear relationship between the predicted and actual values, with points clustering near a line that would represent perfect prediction. However, there are deviations, especially in the higher range of values, where the model appears to underestimate or overestimate the actual numbers.
Distribution of Residuals:
The histogram shows the distribution of residuals, which are the differences between the actual values and the predicted values. A perfectly accurate model would have all residuals clustered at zero.
The residuals are somewhat normally distributed around zero, but there's a noticeable spread, indicating variance in the prediction errors. The shape of the distribution suggests that there are no systematic biases in the predictions (no skewness), but there may be a few large errors (possibly outliers).

```
R^2: 0.9577880720555907
Adjusted R^2: 0.9324609152889451
MAE: 3559.506470588238
MSE: 18895548.570923522
RMSE: 4346.901030725628
```

R-squared (0.9577): This suggests that the model explains approximately 95.77% of the variability in the outcome variable, indicating a strong fit.
Adjusted R-squared (0.9324): After adjusting for the number of predictors, the model still accounts for about 93.24% of the variance, confirming the strength of the model.
MAE (Mean Absolute Error) (3559.56): On average, the model's predictions are off by about 3559.56 units from the actual values, which seems to be relatively low.
MSE (Mean Squared Error) (1,889,548.58): This value is high because MSE is sensitive to outliers and penalizes larger errors more severely.
RMSE (Root Mean Squared Error) (4346.90): This is the square root of the MSE and provides an average error in the same units as the outcome variable. It suggests that typical predictions are within about 4346.90 units of the actual values.

SUPPORT VECTOR MACHINES REGRESSOR

MODEL EVALUATION FOR TRAIN DATA:

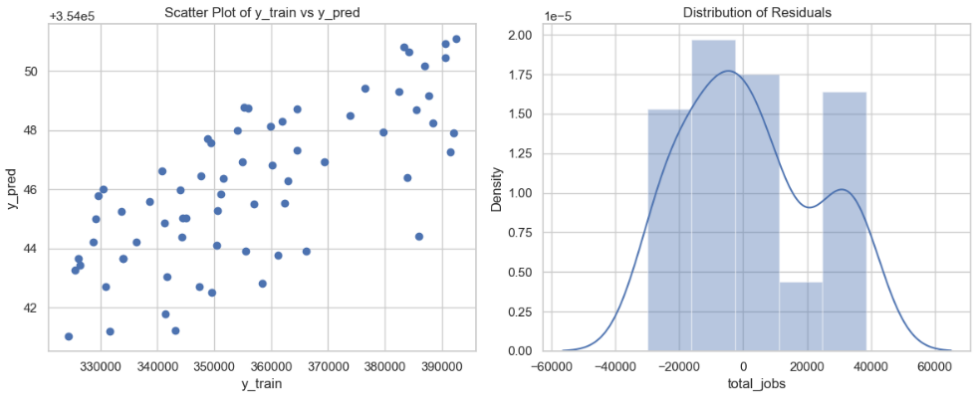


Fig 18: *Inadequacy of SVM Regression Model: Disparities in Predictions and Residuals on TrainingData*

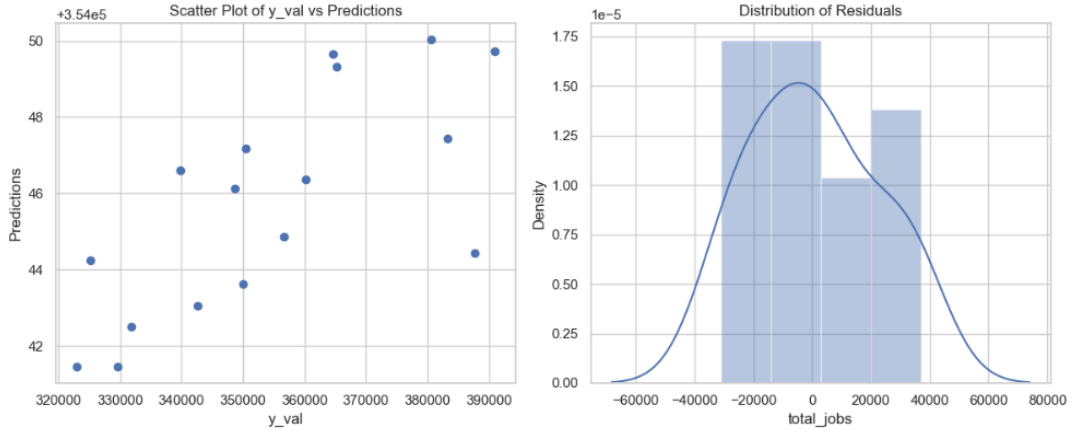
Scatter Plot of Actual vs Predicted Values:
This graph shows the actual values from the training data (y_train) plotted against the predicted values (y_pred).
The points do not align along a diagonal, which would indicate a perfect prediction. Instead, there is considerable scatter and no discernible pattern, suggesting that the model predictions are not in agreement with the actual values.
Distribution of Residuals:
The histogram with a kernel density estimate (KDE) shows the distribution of residuals—the differences between the actual and predicted values.
The distribution of residuals does not center around zero and has multiple peaks, which indicates that the model's predictions are biased and not accurate.

```
R^2: -0.012092979517875246
Adjusted R^2: -0.11330227746966282
MAE: 16921.872644358427
MSE: 420881491.8558469
RMSE: 20515.396458656287
```

The statistical metrics provided are:
R-squared (-0.0129): A negative R-squared value indicates that the model fits the data worse than a horizontal line at the mean of the dependent variable. This is a sign of a poor model fit.
Adjusted R-squared (-0.1133): This adjusted metric accounts for the number of predictors in the model and being negative further confirms that the model does not fit the data well.
MAE (Mean Absolute Error) (16921.87): The average absolute error is quite high, indicating substantial prediction errors.

MSE (Mean Squared Error) (42,838,491.86): The MSE is very large, suggesting that there are significant differences between the predicted and actual values.
RMSE (Root Mean Squared Error) (20,515.39): The RMSE is the square root of the MSE and, being very high, indicates that on average, the model predictions are over 20,000 units away from the actual values.

MODEL EVALUATION FOR VALIDATION DATA:



Scatter Plot of Actual vs Predicted Values:
The scatter plot shows a comparison between the actual validation values (y_val) and the predicted values.
The lack of alignment along a line that would indicate perfect predictions suggests that the model's predictions are not consistent with the actual values. There's considerable scatter, and no clear pattern, indicating poor model performance.
Distribution of Residuals:
This histogram shows the residuals, which are the differences between actual and predicted values. The overlaid kernel density estimation (KDE) suggests the shape of the distribution.
The residuals should ideally center around zero and follow a normal distribution if the model is accurate and unbiased. However, the plot shows a distribution that is not centered around zero, with a wide spread of residuals, which indicates poor prediction accuracy.

```
R^2: -0.0008922972425635667
Adjusted R^2: -0.6014276755881016
MAE: 17775.495095115723
MSE: 448034712.8829705
RMSE: 21166.83048741522
```

The metrics provided give a quantitative evaluation of the model's performance:
R-squared (-0.0088): A negative R-squared value implies that the model performs worse than a simple mean-based prediction. This indicates a poor fit to the data.
Adjusted R-squared (-0.6014): Being negative, it reinforces that the model has a poor fit.
MAE (Mean Absolute Error) (17775.50): The MAE is very high, suggesting significant average errors between the predicted and actual values.
MSE (Mean Squared Error) (44,834,712.89): The high MSE indicates large errors in the predictions.

RMSE (Root Mean Squared Error) (21166.39): As the square root of the MSE, the high RMSE further indicates that the model's predictions are, on average, over 21,000 units away from the actual values.

In summary, the SVM regression model seems to have a poor fit for the validation data. The scatter plot and the distribution of residuals, along with the metrics, all indicate that the model does not predict the validation data accurately.

APPENDIX C: CODE

Histogram visualizations of Economic Indicators: Distribution and Density Analysis

```
plt.figure(figsize=(16, 12))
plt.subplot(3, 3, 1)
sns.histplot(df['logan_passengers'], bins=20, kde=True, color='skyblue')
plt.title('Histogram of Logan Passengers')

plt.subplot(3, 3, 2)
sns.histplot(df['logan_intl_flights'], bins=20, kde=True, color='salmon')
plt.title('Histogram of Logan Intl Flights')

plt.subplot(3, 3, 3)
sns.histplot(df['hotel_occup_rate'], bins=20, kde=True, color='green')
plt.title('Histogram of Hotel Occupancy Rate')

plt.subplot(3, 3, 4)
sns.histplot(df['hotel_avg_daily_rate'], bins=20, kde=True, color='gold')
plt.title('Histogram of Hotel Avg Daily Rate')

plt.subplot(3, 3, 5)
sns.histplot(df['total_jobs'], bins=20, kde=True, color='purple')
plt.title('Histogram of Total Jobs')

plt.subplot(3, 3, 6)
sns.histplot(df['unemp_rate'], bins=20, kde=True, color='red')
plt.title('Histogram of Unemployment Rate')

plt.subplot(3, 3, 7)
sns.histplot(df['labor_force_part_rate'], bins=20, kde=True, color='grey')
plt.title('Histogram of Labor Force Participation Rate')
plt.tight_layout()
plt.show()
```

PAIR PLOTS:

Pairwise Relationships Amongst Economic Indicators Excluding Time Factors:

```
df_excluding_month_year = df.drop(['Month', 'Year'], axis=1)

# Generating the pairplot
sns.pairplot(df_excluding_month_year, plot_kws={'color':'skyblue'})
plt.show()
```

DENSITY PLOTS:

Distribution Analysis Across Multiple Economic Indicators:

```
rows = 2
cols = 7
fig, ax = plt.subplots(nrows=rows, ncols=cols, figsize=(18,8))

# Selecting columns excluding 'Month' and 'Year'
col = df.columns.difference(['Month', 'Year'])
index = 0

for i in range(rows):
    for j in range(cols):
        # Check to avoid index error
        if index < len(col):
            sns.distplot(df[col[index]], ax=ax[i][j])
            index += 1
        else:
            ax[i][j].set_visible(False)  # Hide empty subplots

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

CORRELATION MATRIX:

Heatmap of Correlation Matrix for Economic Indicators:

```
corr_matrix = df.corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

BOXPLOTS:

```
# Dropping 'Year' and 'Month' if present, since they are not needed for individual boxplots
df.drop(['Year', 'Month'], axis=1, errors='ignore', inplace=True)

# Determine the size of the plots (smaller height for each subplot)
plot_height = 2
num_variables = len(df.columns)
fig, axs = plt.subplots(nrows=num_variables, ncols=1, figsize=(10, plot_height * num_variables))

# Ensure axs is iterable
if num_variables == 1:
    axs = [axs]

# Plotting boxplots for each variable separately with different colors
colors = sns.color_palette('hsv', num_variables)

for i, col in enumerate(df.columns):
    sns.boxplot(x=df[col], ax=axs[i], color=colors[i])
    axs[i].set_title(f'Boxplot of {col}')
    axs[i].set_xlabel('')

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

REGRESSION PLOTS:

```
predictors = ['logan_passengers', 'logan_intl_flights', 'hotel_occup_rate', 'hotel_avg_daily_rate',
'unemp_rate']
response_variable = 'total_jobs'

# Create subplots
fig, axes = plt.subplots(3, 2, figsize=(16, 12))
axes = axes.flatten()

# Iterate through variables and create regression plots
for i, predictor in enumerate(predictors):
    sns.regplot(x=predictor, y=response_variable, data=df, ax=axes[i], scatter_kws={'s': 10})
    axes[i].set_title(f'Regression Plot of {response_variable} vs {predictor}')
    axes[i].set_xlabel(predictor)
    axes[i].set_ylabel(response_variable)

# Remove the empty subplot
fig.delaxes(axes[-1])

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

JOINT POINTS:

```
predictors = ['logan_passengers', 'logan_intl_flights', 'hotel_occup_rate', 'hotel_avg_daily_rate',
'unemp_rate', 'labor_force_part_rate']
response_variable = 'total_jobs'

plot_color = 'green'

# Create joint plots
for predictor in predictors:
    sns.jointplot(x=predictor, y=response_variable, data=df, kind='kde', fill=True)
    plt.suptitle(f'{response_variable} vs {predictor}', size=16, y=1.02)
    plt.show()
```

K-MEANS CLUSTERING METHOD:

```
selected_columns = ['logan_passengers', 'logan_intl_flights', 'hotel_occup_rate', 'hotel_avg_daily_rate',
'unemp_rate', 'labor_force_part_rate', 'total_jobs']

X = df[selected_columns]

# Standardizing the features before applying clustering
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Using KMeans clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X_scaled)

# Getting the cluster labels and adding them to the original dataframe
df['Cluster'] = kmeans.labels_

# Since we have more than two columns, we can't easily visualize clusters in a 2D scatter plot.
# Instead, we can use a pair plot to visualize the clusters in higher dimensions.
sns.pairplot(df, vars=selected_columns, hue='Cluster', palette='viridis')
plt.suptitle('Clusters of Economic Indicators', size=16)
plt.tight_layout()
plt.show()
```

LINEAR REGRESSION:

MODEL EVALUATION FOR TRAINING DATA:

```
# Prediction on train data
y_pred = reg.predict(X_train)

# Scatter plot of y_train vs y_pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
plt.subplot(1, 2, 1)
plt.scatter(y_train, y_pred)
plt.title('Scatter Plot of y_train vs y_pred')
plt.xlabel('y_train')
plt.ylabel('y_pred')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_train - y_pred)
plt.title('Distribution of Residuals')

plt.tight_layout()
plt.show()

# Model Evaluation
print('R^2:',metrics.r2_score(y_train, y_pred))
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_train, y_pred))*(len(y_train)-1)/(len(y_train)-
X_train.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_train, y_pred))
print('MSE:',metrics.mean_squared_error(y_train, y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_train, y_pred)))
```

MODEL EVALUATION FOR VALIDATION DATA:

```
pred = reg.predict(X_val)

# Scatter plot of y_val vs pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
plt.subplot(1, 2, 1)
plt.scatter(y_val, pred)
plt.title('Scatter Plot of y_val vs Predictions')
plt.xlabel('y_val')
plt.ylabel('Predictions')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_val - pred)
plt.title('Distribution of Residuals')
```

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

# Model Evaluation
r2_rf = metrics.r2_score(y_val, pred)
print('R^2:', r2_rf)
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_val, pred))*(len(y_val)-1)/(len(y_val)-X_val.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_val, pred))
print('MSE:',metrics.mean_squared_error(y_val, pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_val, pred)))
```

RANDOM FOREST REGRESSOR:

MODEL EVALUATION FOR TRAINING DATA:

```
from sklearn.ensemble import RandomForestRegressor
# Creating an instance of the regressor
reg = RandomForestRegressor()
# Training the mode
reg.fit(X_train,y_train)

# Prediction on train data
y_pred = reg.predict(X_train)

# Scatter plot of y_train vs y_pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
plt.subplot(1, 2, 1)
plt.scatter(y_train, y_pred)
plt.title('Scatter Plot of y_train vs y_pred')
plt.xlabel('y_train')
plt.ylabel('y_pred')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_train - y_pred)
plt.title('Distribution of Residuals')

plt.tight_layout()
plt.show()

# Model Evaluation
print('R^2:',metrics.r2_score(y_train, y_pred))
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_train, y_pred))*(len(y_train)-1)/(len(y_train)-
X_train.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_train, y_pred))
print('MSE:',metrics.mean_squared_error(y_train, y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_train, y_pred)))
```

MODEL EVALUATION FOR VALIDATION DATA:

```
pred = reg.predict(X_val)

# Scatter plot of y_val vs pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
plt.subplot(1, 2, 1)
plt.scatter(y_val, pred)
plt.title('Scatter Plot of y_val vs Predictions')
plt.xlabel('y_val')
plt.ylabel('Predictions')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_val - pred)
plt.title('Distribution of Residuals')

plt.tight_layout()
plt.show()

# Model Evaluation
r2_rf = metrics.r2_score(y_val, pred)
print('R^2:', r2_rf)
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_val, pred))*(len(y_val)-1)/(len(y_val)-X_val.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_val, pred))
print('MSE:',metrics.mean_squared_error(y_val, pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_val, pred)))
```

SUPPORT VECTOR MACHINES REGRESSOR:

MODEL EVALUATION FOR TRAINING DATA:

```
# Import SVM Regressor
from sklearn import svm

# Create a SVM Regressor
reg = svm.SVR()

# Train the model using the training sets
reg.fit(X_train, y_train)

# Model prediction on train data
y_pred = reg.predict(X_train)

# Scatter plot of y_train vs y_pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
```

```
plt.subplot(1, 2, 1)
plt.scatter(y_train, y_pred)
plt.title('Scatter Plot of y_train vs y_pred')
plt.xlabel('y_train')
plt.ylabel('y_pred')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_train - y_pred)
plt.title('Distribution of Residuals')

plt.tight_layout()
plt.show()

# Model Evaluation
print('R^2:',metrics.r2_score(y_train, y_pred))
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_train, y_pred))*(len(y_train)-1)/(len(y_train)-
X_train.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_train, y_pred))
print('MSE:',metrics.mean_squared_error(y_train, y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_train, y_pred)))
```

MODEL EVALUTION FOR VALUATION DATA:

```
# Predicting Test data with the model
y_val_pred = reg.predict(X_val)

# Scatter plot of y_val vs y_val_pred and Distribution of residuals
plt.figure(figsize=(12, 5))

# Scatter plot
plt.subplot(1, 2, 1)
plt.scatter(y_val, y_val_pred)
plt.title('Scatter Plot of y_val vs Predictions')
plt.xlabel('y_val')
plt.ylabel('Predictions')

# Distribution of residuals
plt.subplot(1, 2, 2)
sns.distplot(y_val - y_val_pred)
plt.title('Distribution of Residuals')

plt.tight_layout()
plt.show()

r2_svm = metrics.r2_score(y_val, y_val_pred)
print('R^2:', r2_svm)
print('Adjusted R^2:',1 - (1-metrics.r2_score(y_val, y_val_pred))*(len(y_val)-1)/(len(y_val)-
X_val.shape[1]-1))
print('MAE:',metrics.mean_absolute_error(y_val, y_val_pred))
```

```
print('MSE:',metrics.mean_squared_error(y_val, y_val_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_val, y_val_pred)))
```

EVALUATION AND COMPARISON OF ALL MODELS:

```
models = pd.DataFrame({
   'Model': ['Linear Regression', 'Random Forest', 'Support Vector Machines'],
   'R-squared Score': [r2_linreg, r2_rf, r2_svm]
})



models['R-squared Score'] = (models['R-squared Score'] * 100).round(2).astype(str) + '%'

print(models)
```

```
                  Model          R-squared Score
0         Linear Regression           95.06%
1             Random Forest           95.87%
2 Support Vector Machines           -0.09%
```

REFERENCE LINKS:

[1] : *MTH 522 (Advanced Mathematical Statistics, sections 01B & 02B)*. (n.d.). MTH 522 (Advanced Mathematical Statistics, Sections 01B & 02B). https://mth522.wordpress.com/

[2]: *Analyze Boston***.** https://data.boston.gov/dataset?q=economic+indicators