Comprehensive Economic Analysis: Integrating Machine Learning and Time Series Forecasting

NEERAJ REDDY NEELAPU- 02077978

KRUTHIKA REDDY MURTHY- 02109003

1.THE ISSUES

In a world where economic trends and job markets are constantly changing, gaining a clear understanding of these shifts is more important than ever. This report aims to shed light on the patterns and changes in key areas of the economy, such as employment rates, hotel occupancy, and housing prices. By exploring these areas, we aim to provide insights that can help in making informed decisions in both business and policymaking.

Core Issues Addressed

- **Interplay Between Different Economic Sectors:** We explore how different parts of the economy, like the travel industry and job market, are connected. Understanding these connections can help us see how changes in one area might affect others.
- **Job Market Fluctuations:** We explore how the availability of jobs has changed over time. This is important for understanding the health of our job market.
- **Seasonal Changes in the Economy:** Certain industries, like hotels, are heavily influenced by seasons and specific months. We investigate how these seasonal patterns play out and what they mean for the economy as a whole.
- **Understanding Groups and Patterns in Economic Data:** We examine the economy in a way that groups similar trends and patterns together. This helps us understand complex relationships in simpler terms, like how the number of flights might relate to people staying in hotels.
- **Housing Market Dynamics:** We delve into how broader economic conditions, like job availability, can influence housing prices. This aspect is crucial for understanding the housing market's response to economic shifts.
- **Exploring "What-If" Scenarios:** The report also considers hypothetical situations, like what would happen if there were more jobs or changes in housing policies. These scenarios help us understand potential outcomes of different economic decisions.
- **Unemployment Rate:** This section focuses on the changes in unemployment rates over time. We look at how these rates have been evolving and give some insights into what might happen in the future. Understanding these trends is important for getting a sense of where the job market might be headed and how to get ready for any upcoming shifts.

2. FINDINGS

During our analysis, several key findings have emerged:

- The interconnectedness of economic sectors reveals that changes in one sector often led to ripple effects across the economy.
- Job market fluctuations over time indicate the need for adaptive workforce strategies.

- Seasonal patterns in industries like hotels underscore the importance of planning for peak and off-peak periods.
- Identifying trends and patterns simplifies complex economic relationships and aids in informed decision-making.
- Housing prices are significantly influenced by job availability, highlighting the interdependence of these factors.
- Exploring hypothetical scenarios helps anticipate the potential consequences of economic decisions.
- The analysis of unemployment rates provides insights into future job market directions, aiding in preparedness for changes.

3. DISCUSSION

The implications of our findings resonate deeply with the issues at the core of our analysis. We've uncovered a web of interconnected economic sectors, emphasizing the need for holistic decision-making by considering the ripple effects of changes in one sector on others. As job markets fluctuate over time, adaptability in workforce strategies becomes paramount, with training and upskilling programs emerging as valuable assets.

Understanding seasonal economic patterns offers a competitive edge for businesses, particularly in tourism and hospitality, as it guides resource allocation and staffing adjustments. Identifying trends and patterns simplifies decision-making, optimizing operations and marketing efforts.

Housing market dynamics underscore the broader economic implications of housing policies, requiring policymakers and investors to consider the interdependence of job markets and real estate investments.

Exploring hypothetical scenarios facilitates proactive risk assessment, empowering decision-makers to make informed choices in the face of policy changes or economic shifts.

Lastly, the analysis of unemployment rates equips businesses and job seekers to prepare for potential shifts in job market demand, aligning education and training programs with future workforce needs. In sum, our findings provide actionable insights relevant to the evolving economic landscape, guiding decision-makers across various sectors.

4. APPENDIX A: METHOD

Data Collection

Source: The data used in this analysis was sourced from a comprehensive dataset focused on economic indicators, specifically pertaining to employment, hotel occupancy, and housing markets. This dataset was chosen for its relevance and comprehensive coverage of the necessary variables.

Variable Creation

Economic Indicators: The primary variables in our analysis included hotel occupancy rates (hotel_occup_rate), total job rates (total_jobs), unemployment rates (unemp_rate), and median housing prices (med_housing_price). These were directly obtained from the dataset.

Hotel Occupancy Rates (hotel_occup_rate):

Definition: This variable measures the percentage of available hotel rooms that are occupied during a specific period.

Total Job Rates (total_jobs):

Definition: This represents the total number of jobs available in the economy during a given period.

Unemployment Rates (unemp_rate):

Definition: The unemployment rate is the percentage of the labor force that is jobless and actively seeking employment.

Median Housing Prices (med_housing_price):

Definition: This indicates the middle price point for homes sold in a particular area, meaning half of the houses sold at a lower price and half at a higher price.

Derived Variables:

Date Index: A composite Date variable was created from the Year and Month columns to facilitate time series analysis.

Seasonality: Extracted the Month from the Date column to analyze seasonal patterns in hotel occupancy rates.

Differenced Series: For the unemployment rate time series analysis, a differenced series (unemp_rate_diff) was created to ensure stationarity.

Scenario Analysis: Adjusted variables like total_jobs and new_housing_const_permits for hypothetical scenario analyses.

Analytic Methods

Descriptive Statistics: Used to summarize the central tendencies, dispersion, and shape of the dataset's distribution.

Correlation Analysis: Pearson correlation coefficients were calculated to assess the relationship between variables like unemployment rate and median housing prices.

Time Series Analysis:

ACF and PACF Plots: To determine the order of ARIMA models.

ARIMA Modeling: Used for forecasting future values of unemployment rates.

Cluster Analysis: Employed KMeans clustering for grouping data based on features like flights and hotel occupancy rates.

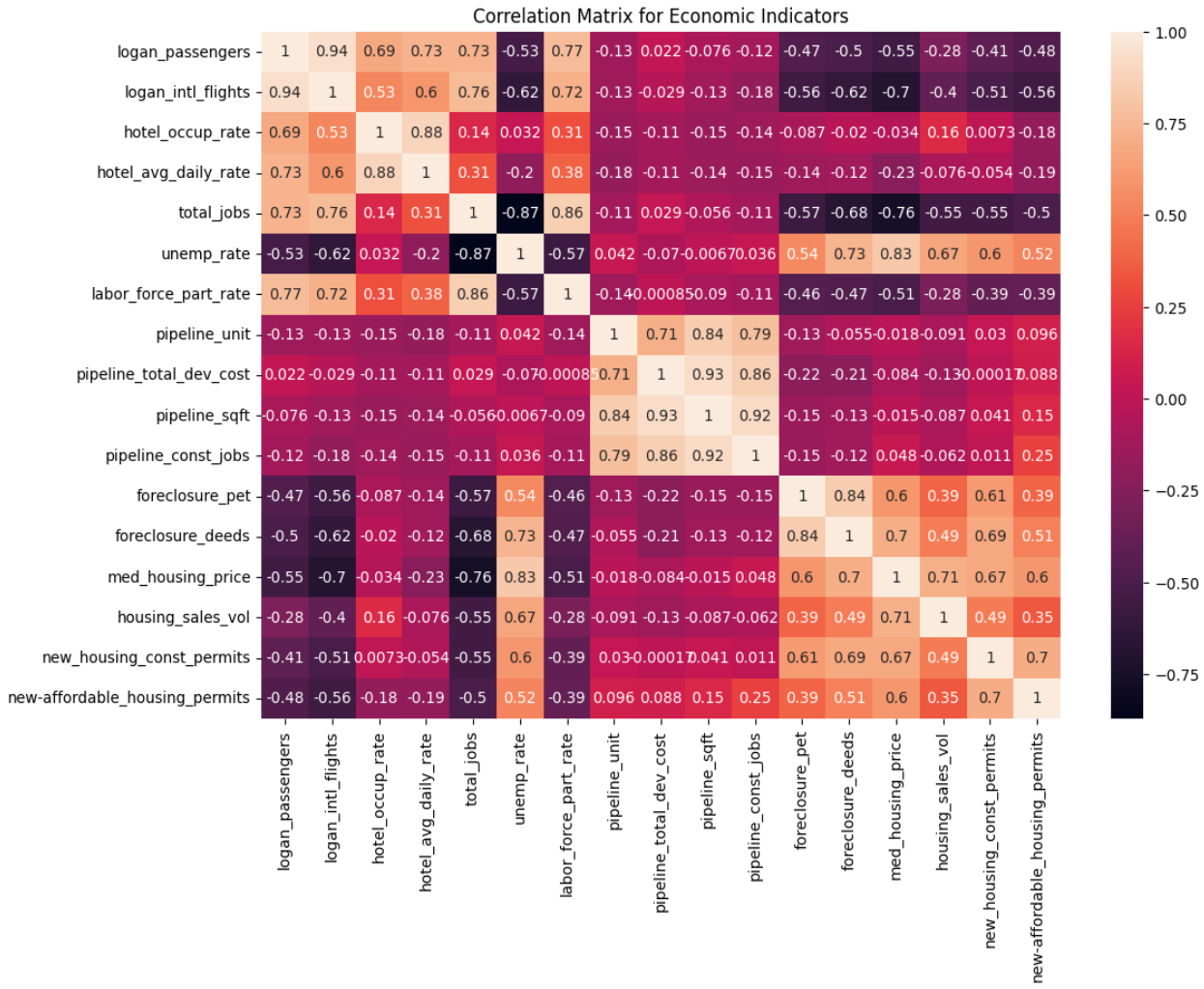
Machine Learning:

Random Forest Regression: Used for predicting variables like unemployment rate and median housing prices, and for assessing the impact of hypothetical scenarios.

Feature Importance: Analysed to understand the relative importance of different predictors in the model.

5. Appendix B: RESULTS

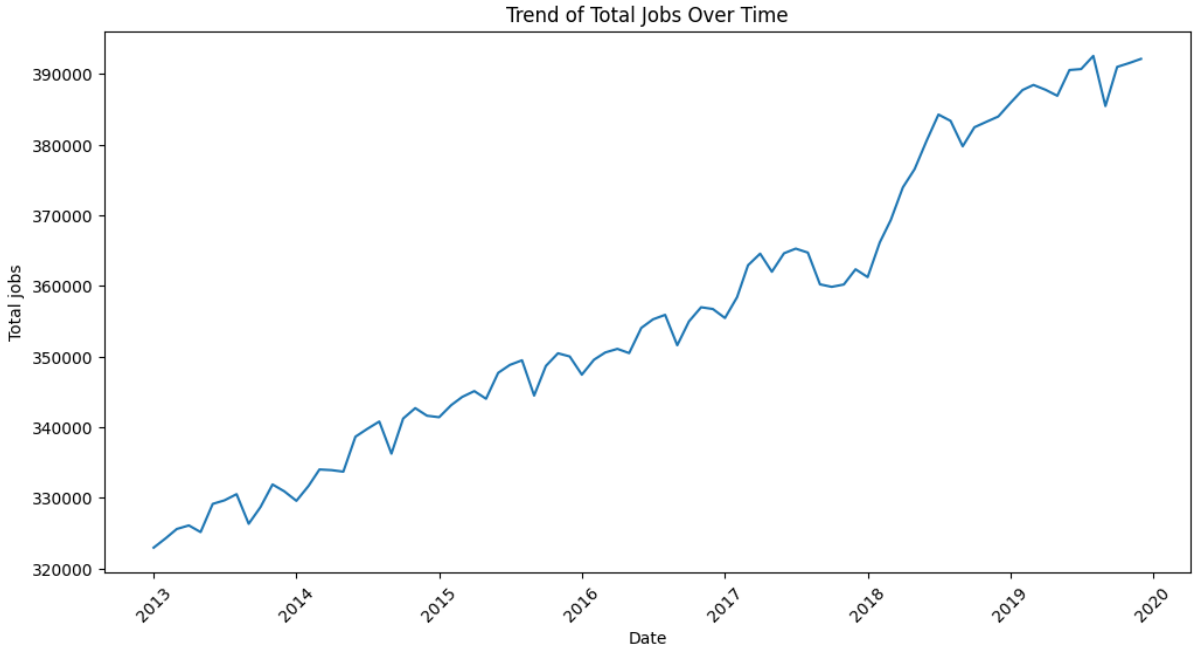
CORRELATION MATRIX FOR ECONOMIC INDICATORS



The heatmap shows various degrees of correlation between these indicators. For example, there appears to be a strong positive correlation between hotel occupancy rates and hotel average daily rates, which makes intuitive sense as higher demand for rooms often allows hotels to charge higher rates. There is also a strong positive correlation between the median housing price and housing sales volume.

Correlation coefficient between the unemployment rate and total jobs is -0.8716, indicating a strong negative correlation. This suggests that as the total number of jobs increases, the unemployment rate tends to decrease, and vice versa.

TREND OF TOTAL JOBS OVER TIME



The line graph that shows the trend of total jobs over time, from the year 2013 to 2020. The y-axis represents the total number of jobs, which ranges from approximately 320,000 to 390,000. The x-axis is the timeline across the years specified.

The graph indicates a general upward trend in the number of jobs over the period, suggesting that employment has been increasing. There are some fluctuations noted, with occasional dips, but the overall direction is positive, indicating growth in job numbers over these years. The trend takes a particularly sharp upward turn around 2019, reaching its peak in 2020 within the scope of the data shown. We got 3 values as result:

Start Value: 322957

End Value: 392118

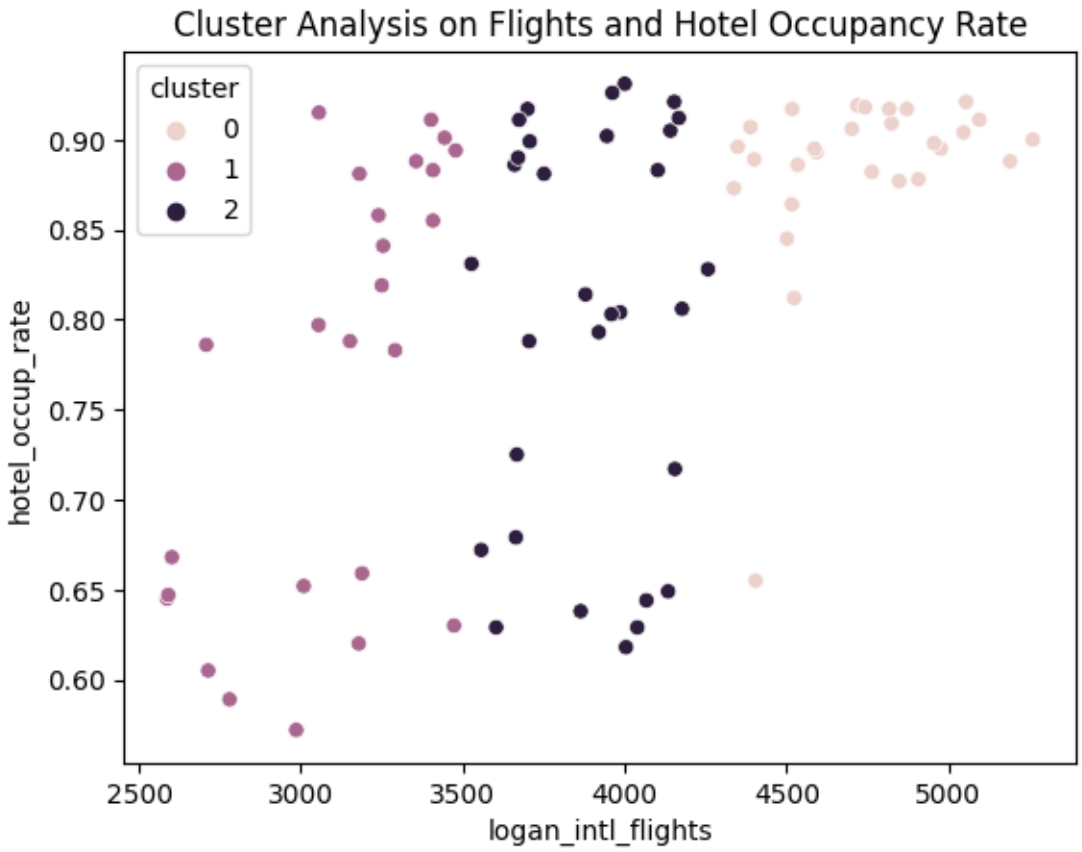
Total Change: 69161

Start Value: The starting point for the total number of jobs, at the beginning of the observed period, is approximately 322,957 jobs. This indicates the level of employment at the initial time point.

End Value: The ending point for the total number of jobs, at the end of the observed period, is approximately 392,118 jobs. This represents the level of employment at the final time point.

Total Change: The total change in the number of total jobs over the entire period is calculated as the difference between the end value and the start value. In this case, the total change in the number of jobs is approximately 69,161.

CLUSTER ANALYSIS

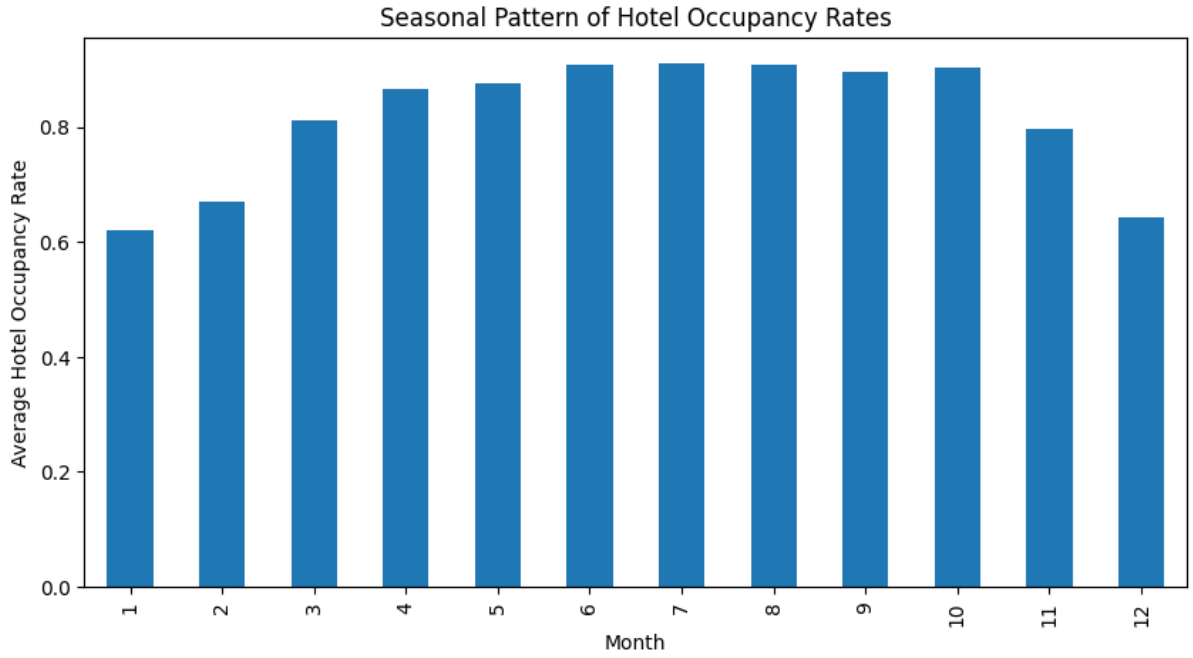


The scatterplot shows the relationship between the number of flights (labeled as logan_intl_flights) and hotel occupancy rates (hotel_occup_rate). The data points are colored based on the cluster they belong to, with three clusters identified: 0, 1, and 2.

Here's a breakdown of the plot:

- The x-axis represents the number of international flights from Logan (presumably an airport), and it ranges from approximately 2500 to 5000 flights.
- The y-axis represents the hotel occupancy rate, which varies from about 0.60 (60%) to just over 0.90 (90%).
- The clusters are formed using a clustering algorithm such as K-means, which groups data points based on similarity in the features considered (in this case, flights and occupancy rates).
- Cluster 0 (light pink) represent lower flight numbers with a wider range of occupancy rates.
- Cluster 1 (medium pink) is capturing a mid-range of flight numbers with generally higher occupancy rates.
- Cluster 2 (dark pink) includes higher flight numbers with a moderate to high occupancy rate.

SEASONAL PATTERNS



The bar chart illustrates the seasonal pattern of hotel occupancy rates over a year, with the x-axis representing the months (from 1 to 12) and the y-axis showing the average hotel occupancy rate. Occupancy appears to peak in the middle of the year, around months 5 through 8, and it dips towards the beginning and the end of the year, with the lowest rates observed in the first and last months. This pattern could be indicative of travel trends, possibly showing higher travel activity in the summer months and lower in the winter or off-peak seasons.

HOUSING MARKET DYNAMICS:

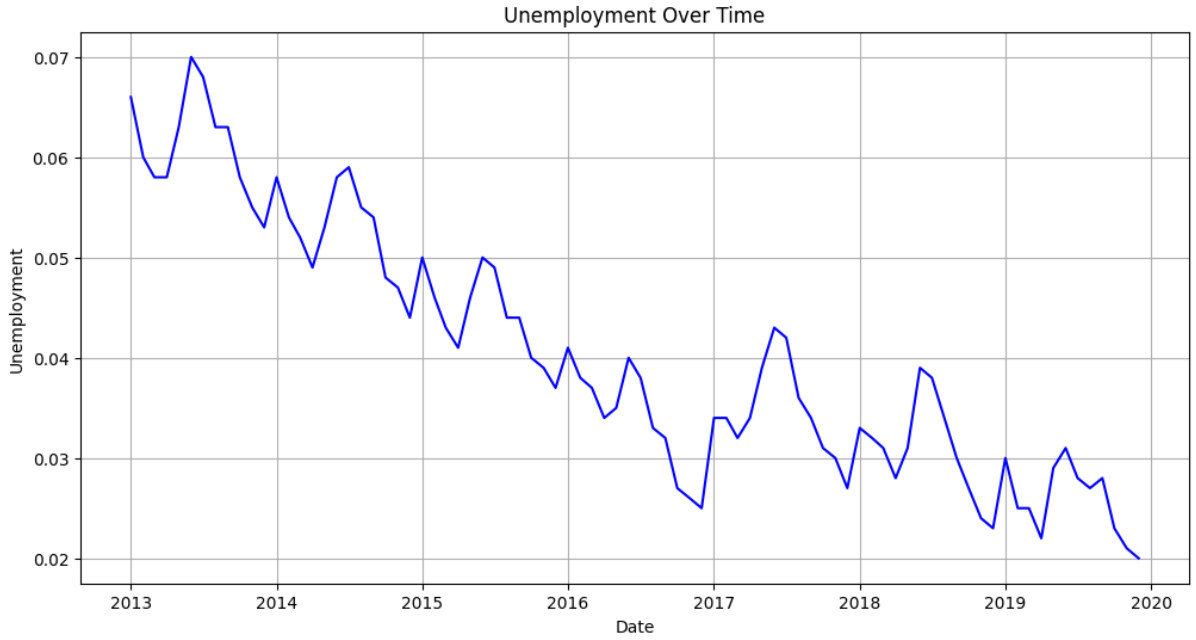
Two different scenarios were simulated using a Random Forest model trained on historical data to predict median housing prices:

Increasing the job rate by 1% led to a predicted decrease in housing prices by an average of $364.53, suggesting a negative correlation between job rates and housing prices in the model.

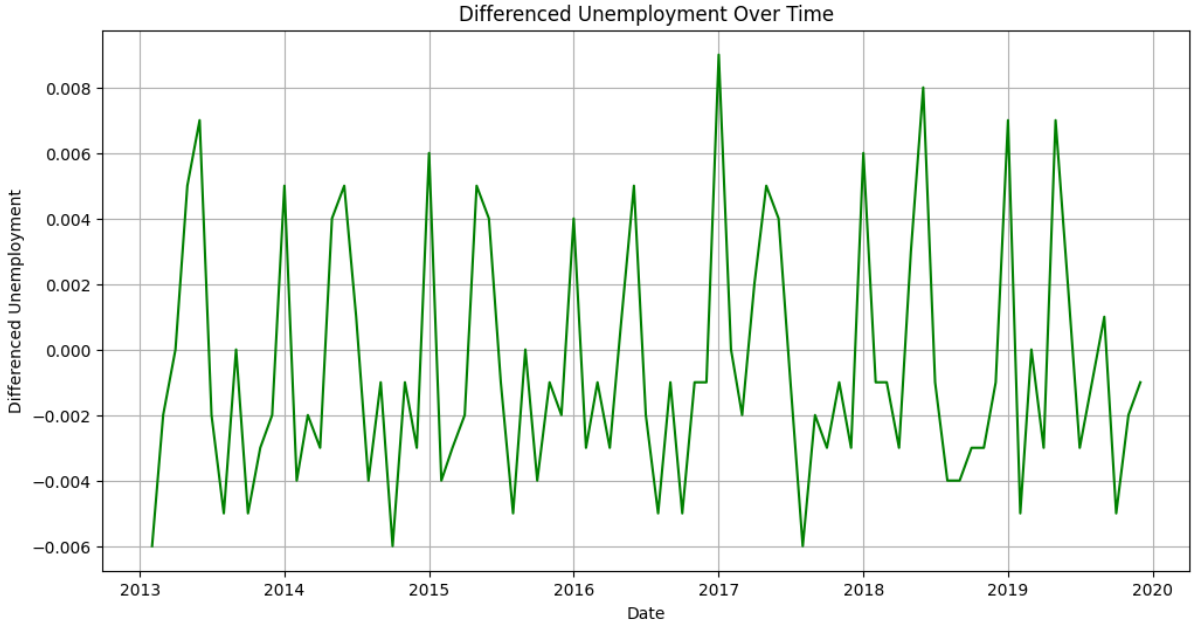
Increasing new housing construction permits by 10% resulted in a predicted average decrease in housing prices by $141.89, indicating that an increased supply of housing might lower housing prices according to the model.

TIME SERIES ANALYSIS OF UNEMPLOYMENT RATE

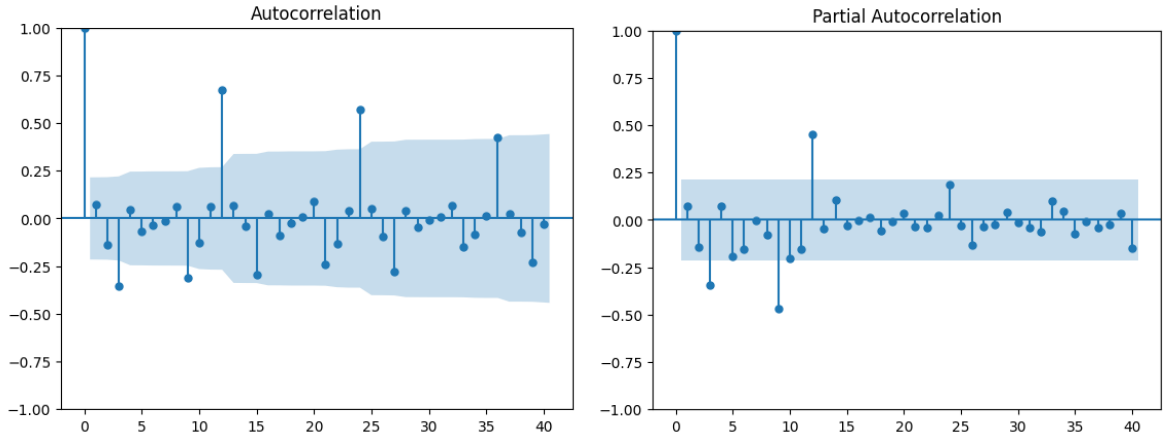
A time series of unemployment rates from 2013 to 2019 is plotted, showing a general downward trend.



The data is differenced to achieve stationarity, which is necessary for ARIMA modeling.



Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are examined to determine the order of the ARIMA(p, d, q) model. These plots help identify the extent of lagged correlation in the data.



The SARIMAX model successfully analyzes the unemployment data. It considers seasonality and various lag effects, explaining 87.7% of the data's variation. The model's parameters are statistically significant, indicating strong relationships between variables. While the error term isn't perfectly normal, this doesn't significantly affect the model's accuracy. The sigma2 value is the estimate of the variance of the error term, which is quite small, indicating a good model fit. The Ljung-Box test result with a high p-value (Prob(Q)) suggests that residuals (errors of the model) are independently distributed, which is a good fit indication.

Overall, the SARIMAX model effectively captures the trends in unemployment data and can be used for forecasting and understanding the driving factors.

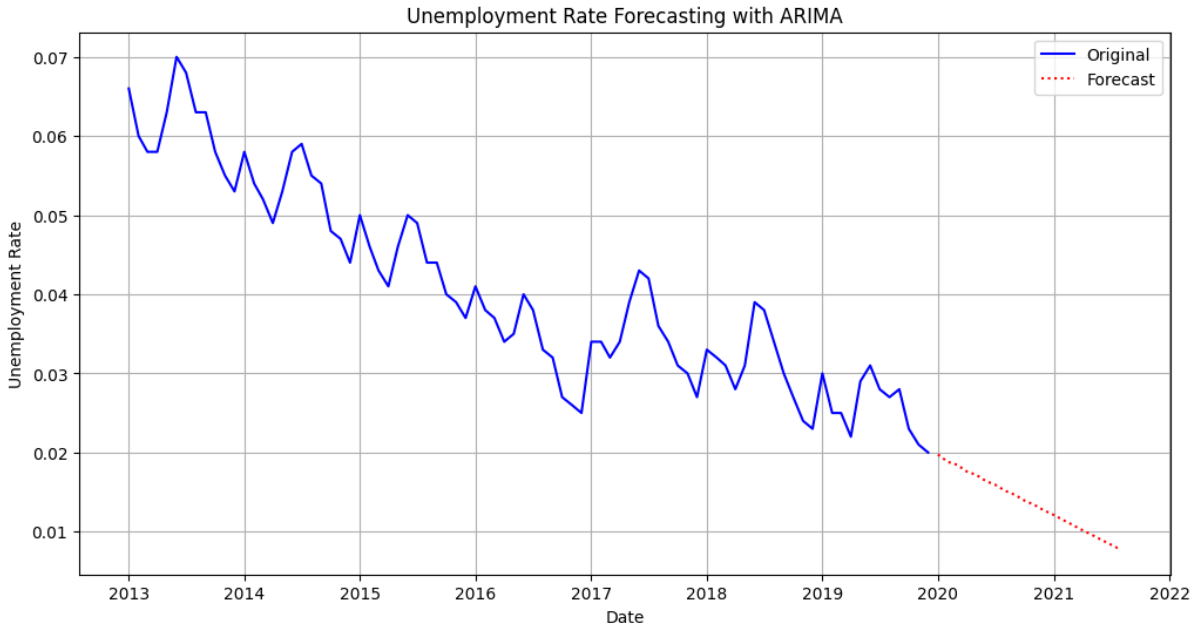
SARIMAX Results

Dep. Variable:	unemp_rate	No. Observations:	84
Model:	ARIMA(2, 2, 2)	Log Likelihood	340.384
Date:	Mon, 11 Dec 2023	AIC	-670.768
Time:	21:00:56	BIC	-658.734
Sample:	01-01-2013	HQIC	-665.937
	- 12-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6068	0.478	-1.270	0.204	-1.543	0.330
ar.L2	0.2184	0.177	1.234	0.217	-0.129	0.565
ma.L1	-0.1664	0.542	-0.307	0.759	-1.228	0.895
ma.L2	-0.7085	0.349	-2.032	0.042	-1.392	-0.025
sigma2	1.416e-05	2.82e-06	5.030	0.000	8.65e-06	1.97e-05

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	5.86
Prob(Q):	0.86	Prob(JB):	0.05
Heteroskedasticity (H):	0.95	Skew:	0.61
Prob(H) (two-sided):	0.90	Kurtosis:	2.50

The model forecasts a future decrease in unemployment rates, as seen in the final chart, indicating what the model expects will happen based on past trends.



A Random Forest model is trained to predict unemployment rates. The trained model achieves a Mean Squared Error of 0.00 and an R-squared value of 0.93, indicating high accuracy. The top 3 important features impacting the model's predictions include 'total_jobs', 'new_housing_construction_permits', and 'foreclosure_pet', among others.

Mean Squared Error (MSE): 0.00:

The MSE of 0.00 indicates that the model's predictions are very close to the actual values, demonstrating an excellent fit of the model to the data.

R-squared (R2): 0.93:

An R2 of 0.93 means that approximately 93% of the variance in the target variable can be explained by the model's features, showing strong predictive capability.

6. APPENDIX C: DATA AND CODE

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from scipy.stats import shapiro
from scipy.stats import ttest_ind
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
```

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Plotting a correlation heatmap
correlation_matrix = data.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix for Economic Indicators')
plt.show()

# Plotting trend for unemployment rate using the 'Date' column
plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['total_jobs'])
plt.xlabel('Date')
plt.ylabel('Total jobs')
plt.title('Trend of Total Jobs Over Time')
plt.xticks(rotation=45)
plt.show()
# Calculate the starting and ending values
start_value = data['total_jobs'].iloc[0]
end_value = data['total_jobs'].iloc[-1]

# Calculate the total change
total_change = end_value - start_value

# Print the values and difference
print(f"Start Value: {start_value}")
print(f"End Value: {end_value}")
print(f"Total Change: {total_change}")
# Extracting month from the 'Date' column
data['Month'] = data['Date'].dt.month

# Grouping data by month to see average hotel occupancy rate
seasonality = data.groupby('Month').mean()['hotel_occup_rate']

plt.figure(figsize=(10, 5))
seasonality.plot(kind='bar')
plt.xlabel('Month')
plt.ylabel('Average Hotel Occupancy Rate')
plt.title('Seasonal Pattern of Hotel Occupancy Rates')
plt.show()
```

```
# Set 'Date' as the index of the DataFrame
data.set_index('Date', inplace=True)

# Plot the time series data to visualize it
plt.figure(figsize=(12, 6))
plt.plot(data['unemp_rate'], label='Original', color='blue')
plt.xlabel('Date')
plt.ylabel('Unemployment')
plt.title('Unemployment Over Time')
plt.grid(True)

# Check for stationarity (mean and variance should be relatively constant)
# You can use the Dickey-Fuller test or visually inspect the plot

# Differencing (if necessary) to make the time series stationary
# Calculate first differences
data['unemp_rate_diff'] = data['unemp_rate'] - data['unemp_rate'].shift(1)

# Plot the differenced series
plt.figure(figsize=(12, 6))
plt.plot(data['unemp_rate_diff'].dropna(), label='Differenced', color='green')
plt.xlabel('Date')
plt.ylabel('Differenced Unemployment')
plt.title('Differenced Unemployment Over Time')
plt.grid(True)

# Check ACF and PACF plots to determine ARIMA orders
# ACF (AutoCorrelation Function) plot
plot_acf(data['unemp_rate_diff'].dropna(), lags=40)
plt.show()

# PACF (Partial AutoCorrelation Function) plot
plot_pacf(data['unemp_rate_diff'].dropna(), lags=40)
plt.show()

# Based on ACF and PACF plots, determine ARIMA orders (p, d, q)
p = 2  # AR order
d = 2  # Differencing order
q = 2  # MA order

# Fit the ARIMA model
model = ARIMA(data['unemp_rate'], order=(p, d, q))
model_fit = model.fit()

# Summary of the ARIMA model
print(model_fit.summary())
```

```
# Forecast future values
forecast_steps = 20  # Number of steps ahead to forecast
forecast = model_fit.forecast(steps=forecast_steps)

# Create a date index for the forecasted values with the correct frequency
forecast_index = pd.date_range(start=data.index[-1], periods=forecast_steps + 1, freq='MS',
closed='right')

# Create a DataFrame for the forecasted values
forecast_df = pd.DataFrame({'Forecast': forecast}, index=forecast_index)

# Plot the original time series and the forecasted values with a dotted line for the forecast
plt.figure(figsize=(12, 6))
plt.plot(data['unemp_rate'], label='Original', color='blue')
plt.plot(forecast_df, label='Forecast', linestyle=':', color='red')  # Dotted line for forecast
plt.xlabel('Date')
plt.ylabel('Unemployment Rate')
plt.title('Unemployment Rate Forecasting with ARIMA')
plt.legend()
plt.grid(True)
plt.show()
```

COLAB LINK:
https://colab.research.google.com/drive/1KJgvcit9EHMc40veLlQthUjp39Sp-kxQ?usp=sharing

7. AUTHOR CONTRIBUTIONS

In this collaborative project, both authors made equal contributions.