INDEX

1. INTRODUCTION

1.1 Problem Statement:

The objective of the Problem Statement is to construct a predictive model capable of forecasting the number of tourist arrivals for upcoming time periods, leveraging historical tourist arrival data and Internet search index data.
The ultimate aim is to create a precise and resilient predictive model that can offer valuable insights to tourism authorities and businesses for anticipating tourist arrivals and such insights, will facilitate improvements and strategic planning for the future.

1.2 Data Description:

The dataset encompasses a comprehensive collection of data spanning from **January 2010 to December 2022**, serving as our actual data. The dataset is structured into various columns, each providing distinct information critical for the predictive modeling task.

Detailed description of the columns present in the dataset:

Month: The specific month of the data entry.

Number of Tourists: The number of tourists recorded for the respective city during the respective month. It is further divided into two categories as **Indian Tourists** and **Foreign Tourists**, and then **Total** is also calculated.

GSDP values (Interpolated monthly) at current prices for respective State: Gross State Domestic Product (GSDP) values for the respective states are interpolated on a monthly basis and reported in current prices.

Per capita Income (State) monthly: Per capita income for the respective states are reported on a monthly basis.

Accommodation (Hotels/Restaurant) and Travelling (Flight/Train/Bus):

(Worldwide) Trends: Trends in worldwide searches to respective cities.
(India) Trends: Trends in searches to respective cities specifically from India.

Climate (Temperature/Pressure/Humidity): The recorded temperature, pressure and humidity values are taken as categorized to Low, Average and High for the respective cities for the respective months.

Worldwide Trends: General trends related to worldwide searches for respective places to visit.

The dataset has been partitioned into a **training** set comprising data from **2010 to 2021** and a **test** set specifically for the year **2022**.
Predictions for the year **2022** were made using various **Time series** and **Machine Learning** models, and the accuracy of these predictions was evaluated by comparing them with the actual test data.

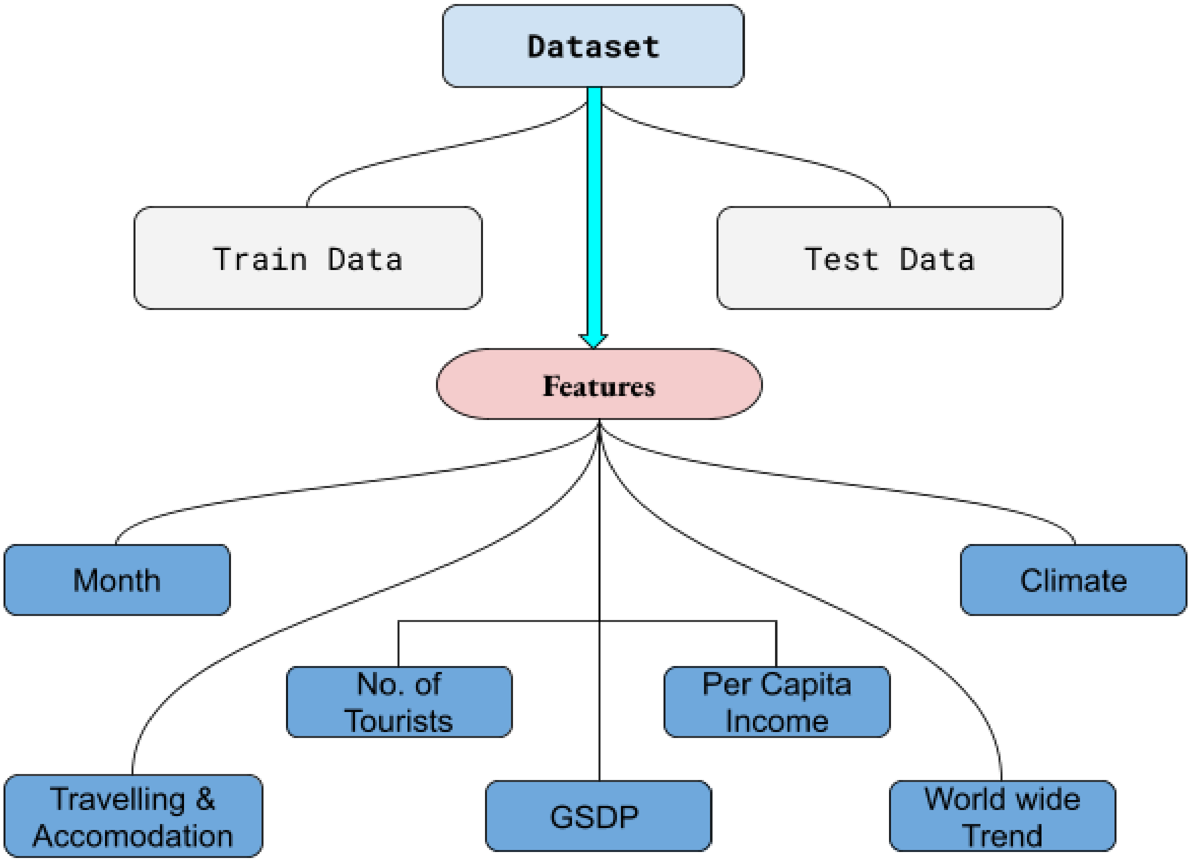


Figure 1: Flowchart depicting the data distribution

2. DATA ANALYSIS AND VISUALIZATION

2.1 Tourist Distribution:

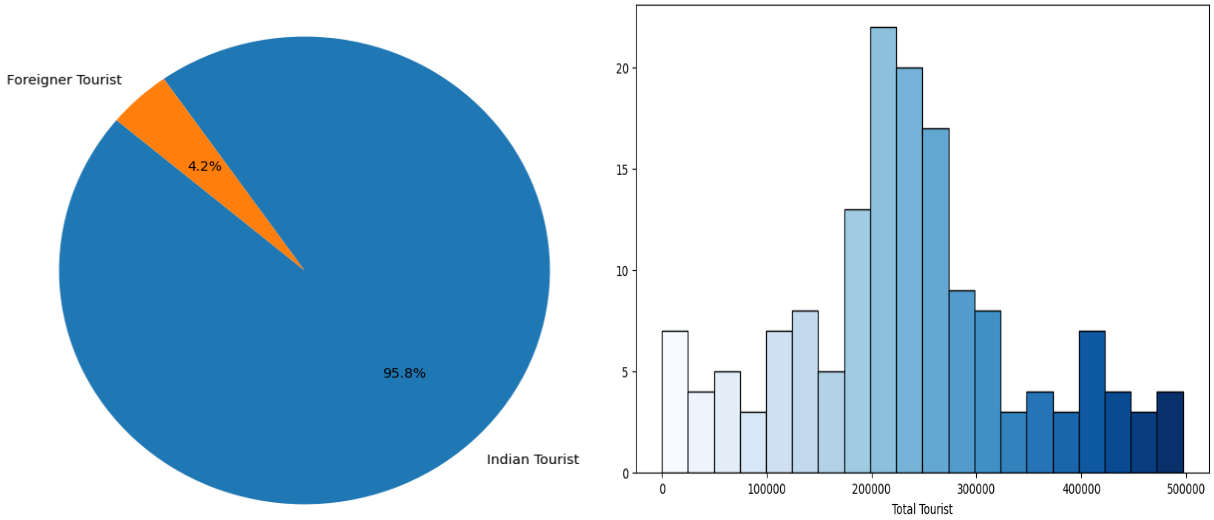


Figure 2 & 3: Flowchart depicting the Tourist distribution

The pie chart effectively illustrates the tourist distribution, highlighting a substantial contrast between Indian and foreign tourists. Specifically, Indian tourists constitute the overwhelming majority at **95.8%** higher presence, while foreign tourists represent a comparatively small proportion at **4.2%**

The generated histogram with the **KDE** curve provides valuable insights into the distribution of the 'Total Tourist' data and the frequency of tourist numbers. By examining the plot, we can observe that the 'Total Tourist' data follows a bell-shaped or roughly symmetrical distribution, which suggests that it may be approximately **normally** distributed.

2.2 Tourist Trends:

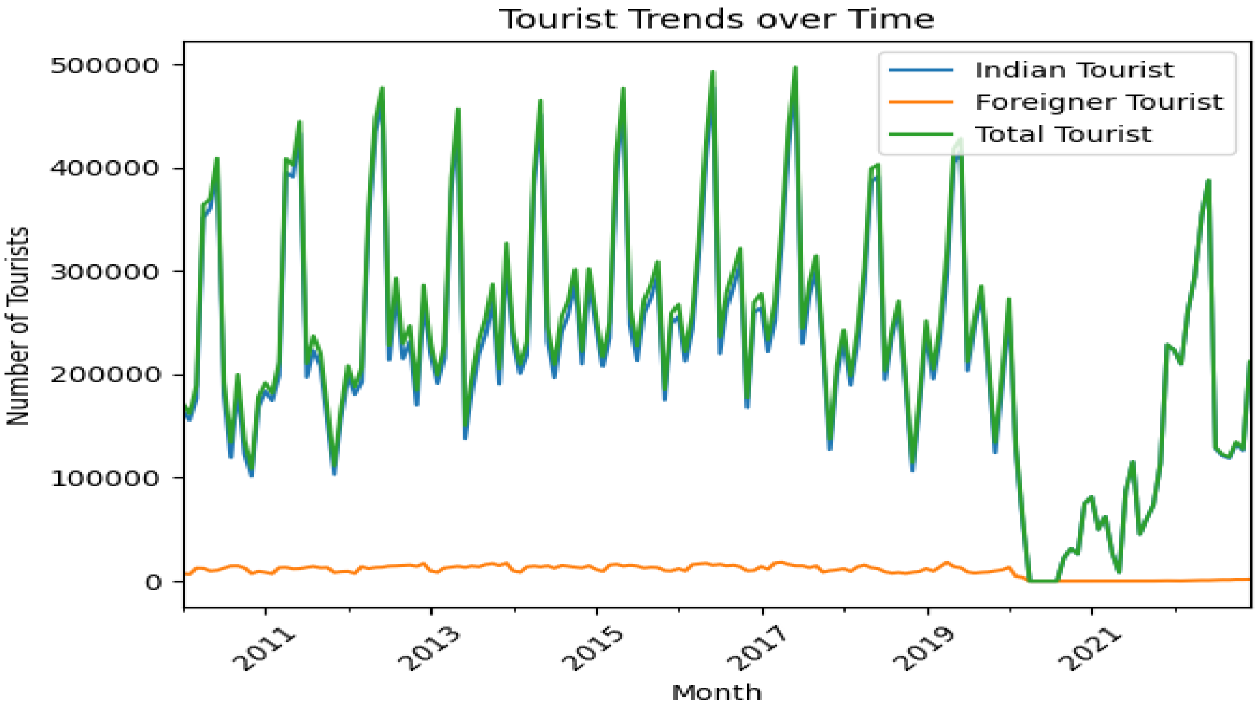


Figure 4: Tourist Trends vs *Time Plot*

The plot illustrates the tourist trends over time, with a predominant presence of Indian tourists and a smaller proportion of foreign tourists. Notably, in the year 2020 and 2021, the number of foreign tourists dropped to zero due to the impact of the **COVID-19** pandemic. This reflects a significant shift in the tourism dynamics, with Indian tourists becoming the sole contributors to the total tourist count during that period.

In summary, the plot effectively conveys the changing composition of tourists, with Indian tourists being the primary demographic and the abrupt cessation of foreign tourist arrivals in 2020 and 2021 as a consequence of the COVID-19 pandemic.

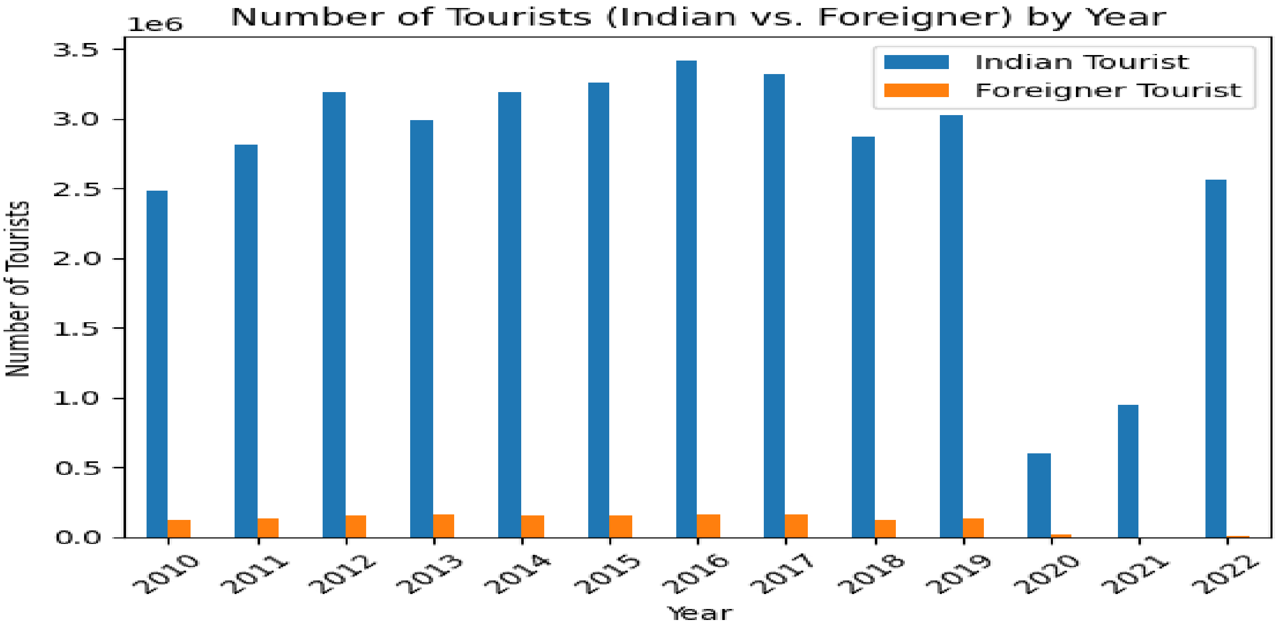


Figure 6: Indian vs Foreigner Tourists Distribution

The image shows a chart comparing the number of tourists in India versus foreign tourists over the years. The chart consists of two columns, one for Indian tourists and the other for foreign tourists.comparison of the number of tourists in India versus foreign tourists over the years, helping to understand the tourism trends in the country.

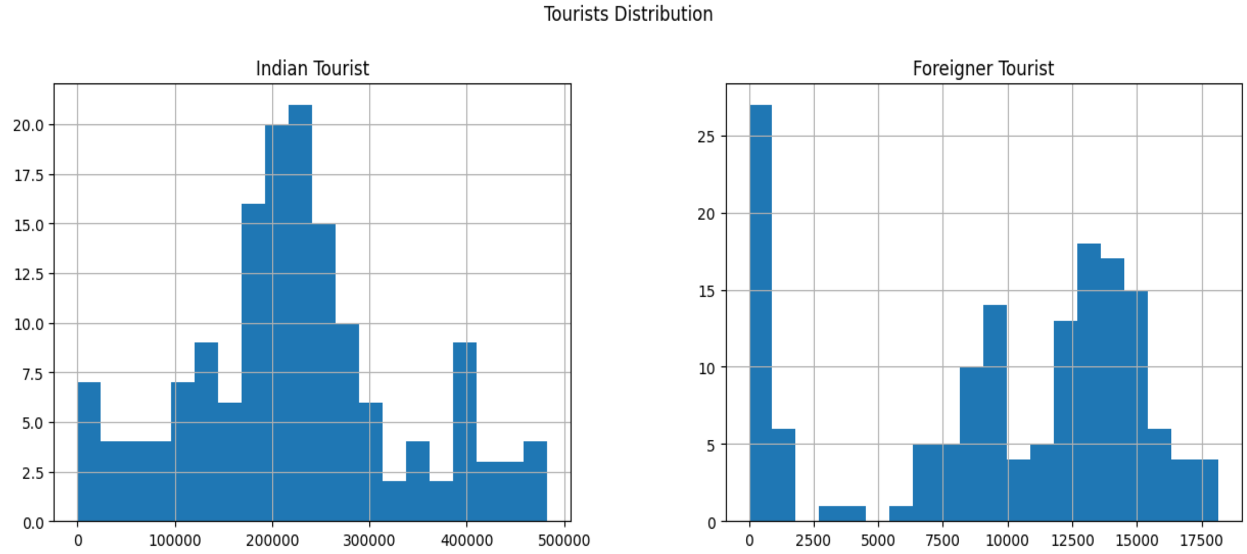


Figure 7: Flowchart depicting the Tourist count Comparison

2.3 GDP and Per Capita Trends:

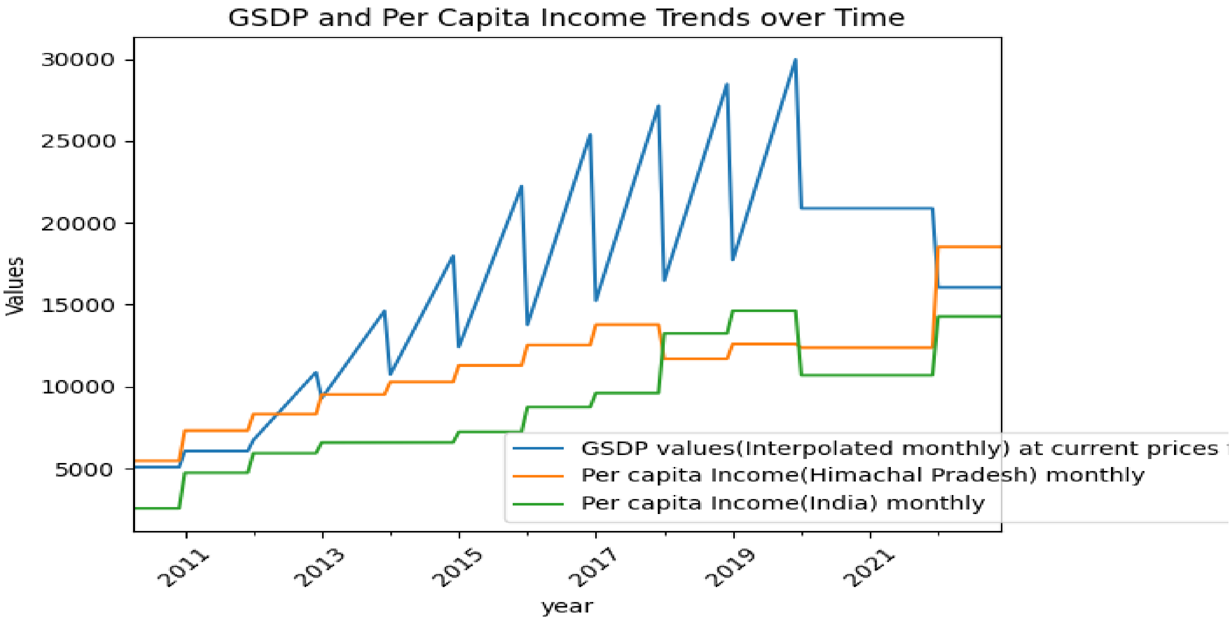


Figure 8: State GDP and Per Capita Income Distribution

The line plot offers insights into the temporal trends of economic indicators in Himachal Pradesh, specifically **GSDP** values and **Per Capita Income**, along with a comparison to the **National Per Capita Income** in India.

The graph provides an overview of the trends in **GDP** per capita income over time, specifically focusing on the value of the rupee and its impact on the price of the rupee**.**

2.4 Transportation Trends:

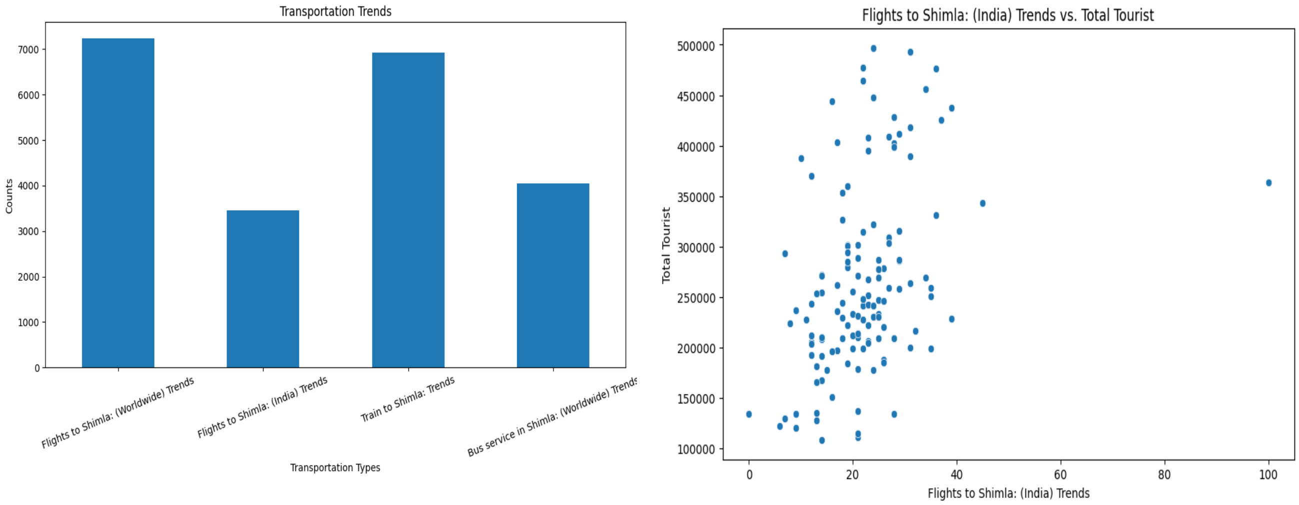


Figure 9 & 10: Types of Transportation and Flights Distribution Plots

The bar plot provides insights into transportation trends in Shimla, showcasing transportation methods more popular or heavily utilized in Shimla, as indicated by the counts for each category. Transportation Popularity, Modes of Transport

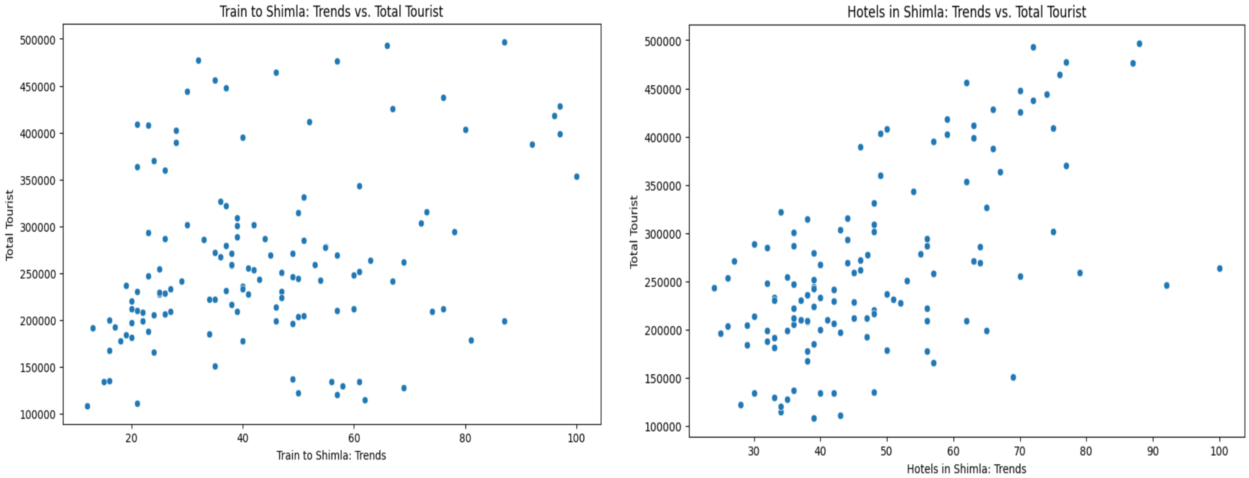
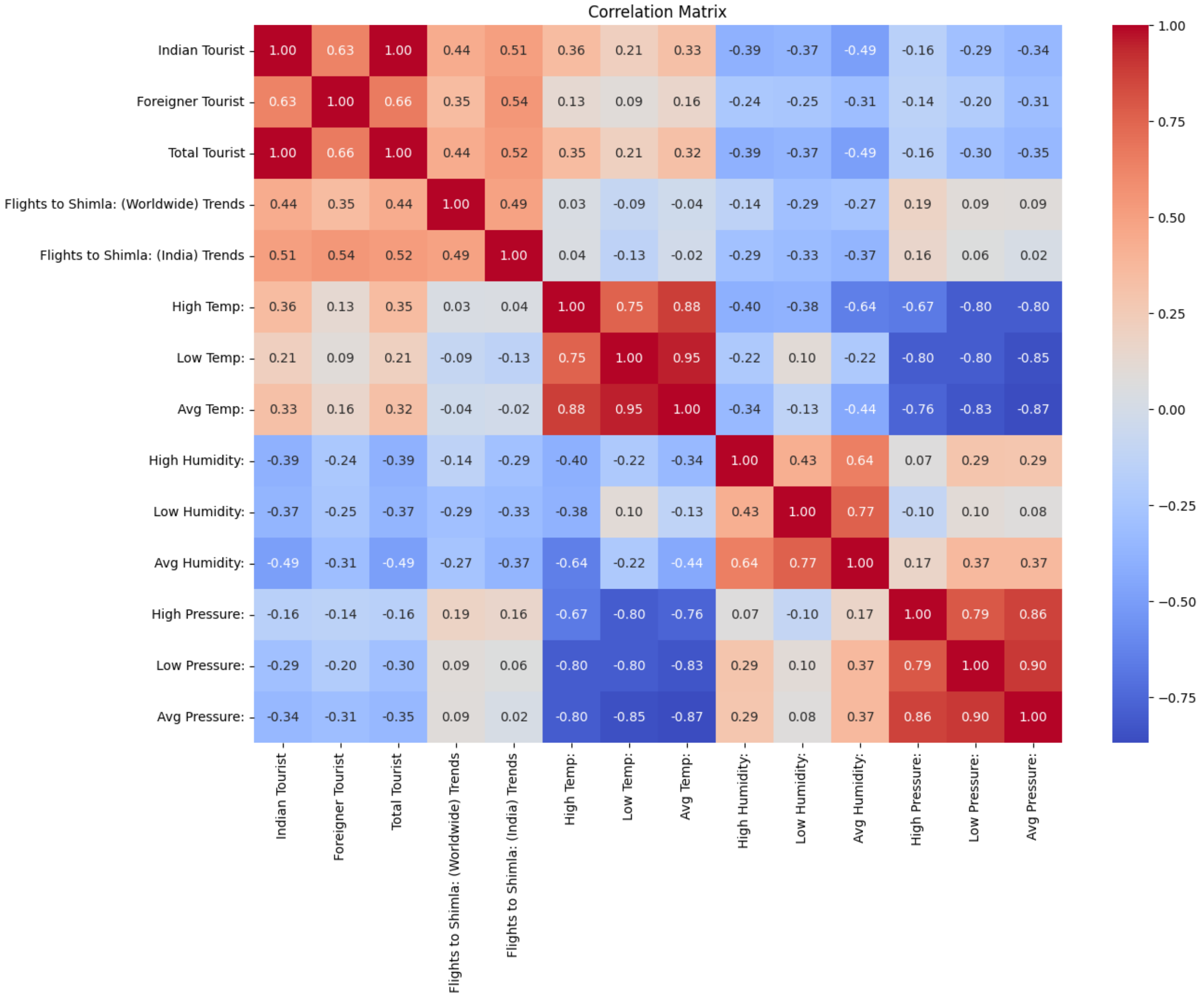


Figure 11 & 12: Travelling & Hotels Plots

2.5 Correlation Distribution:



	Indian Tourist	Foreigner Tourist	Total Tourist	Flights to Shimla: (Worldwide) Trends	Flights to Shimla: (India) Trends	High Temp:	Low Temp:	Avg Temp:	High Humidity:	Low Humidity:	Avg Humidity:	High Pressure:	Low Pressure:	Avg Pressure:
Indian Tourist	1.00	0.63	1.00	0.44	0.51	0.36	0.21	0.33	-0.39	-0.37	-0.49	-0.16	-0.29	-0.34
Foreigner Tourist	0.63	1.00	0.66	0.35	0.54	0.13	0.09	0.16	-0.24	-0.25	-0.31	-0.14	-0.20	-0.31
Total Tourist	1.00	0.66	1.00	0.44	0.52	0.35	0.21	0.32	-0.39	-0.37	-0.49	-0.16	-0.30	-0.35
Flights to Shimla: (Worldwide) Trends	0.44	0.35	0.44	1.00	0.49	0.03	-0.09	-0.04	-0.14	-0.29	-0.27	0.19	0.09	0.09
Flights to Shimla: (India) Trends	0.51	0.54	0.52	0.49	1.00	0.04	-0.13	-0.02	-0.29	-0.33	-0.37	0.16	0.06	0.02
High Temp:	0.36	0.13	0.35	0.03	0.04	1.00	0.75	0.88	-0.40	-0.38	-0.64	-0.67	-0.80	-0.80
Low Temp:	0.21	0.09	0.21	-0.09	-0.13	0.75	1.00	0.95	-0.22	0.10	-0.22	-0.80	-0.80	-0.85
Avg Temp:	0.33	0.16	0.32	-0.04	-0.02	0.88	0.95	1.00	-0.34	-0.13	-0.44	-0.76	-0.83	-0.87
High Humidity:	-0.39	-0.24	-0.39	-0.14	-0.29	-0.40	-0.22	-0.34	1.00	0.43	0.64	0.07	0.29	0.29
Low Humidity:	-0.37	-0.25	-0.37	-0.29	-0.33	-0.38	0.10	-0.13	0.43	1.00	0.77	-0.10	0.10	0.08
Avg Humidity:	-0.49	-0.31	-0.49	-0.27	-0.37	-0.64	-0.22	-0.44	0.64	0.77	1.00	0.17	0.37	0.37
High Pressure:	-0.16	-0.14	-0.16	0.19	0.16	-0.67	-0.80	-0.76	0.07	-0.10	0.17	1.00	0.79	0.86
Low Pressure:	-0.29	-0.20	-0.30	0.09	0.06	-0.80	-0.80	-0.83	0.29	0.10	0.37	0.79	1.00	0.90
Avg Pressure:	-0.34	-0.31	-0.35	0.09	0.02	-0.80	-0.85	-0.87	0.29	0.08	0.37	0.86	0.90	1.00

Figure 13 : Correlation Heatmap

This correlation heatmap is a graphical representation of a correlation matrix, which is a table that shows the correlation between pairs of variables. The correlation between two variables is a measure of how strongly they are related to each other. It can range from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation. Columns with high correlation can be problematic for machine learning models. When two columns are highly correlated, it means that they contain a lot of the same information. This can lead to overfitting, where the model learns the relationship between the correlated columns too well and is unable to generalize to new data.

To avoid this, we have dropped one of the columns in a pair with high correlation.

3. DATA RECONSTRUCTION

3.1 Impact of the COVID-19 on Tourism:

The graphical representation clearly illustrates a significant decline in tourist numbers during the years **2020** and **2021**. This abrupt drop can be directly attributed to the unprecedented challenges posed by the global COVID-19 pandemic and the subsequent implementation of strict lockdown measures. These restrictions, initiated to curb the spread of the virus, had a profound impact on the travel and tourism industry, leading to a substantial reduction in the number of visitors.

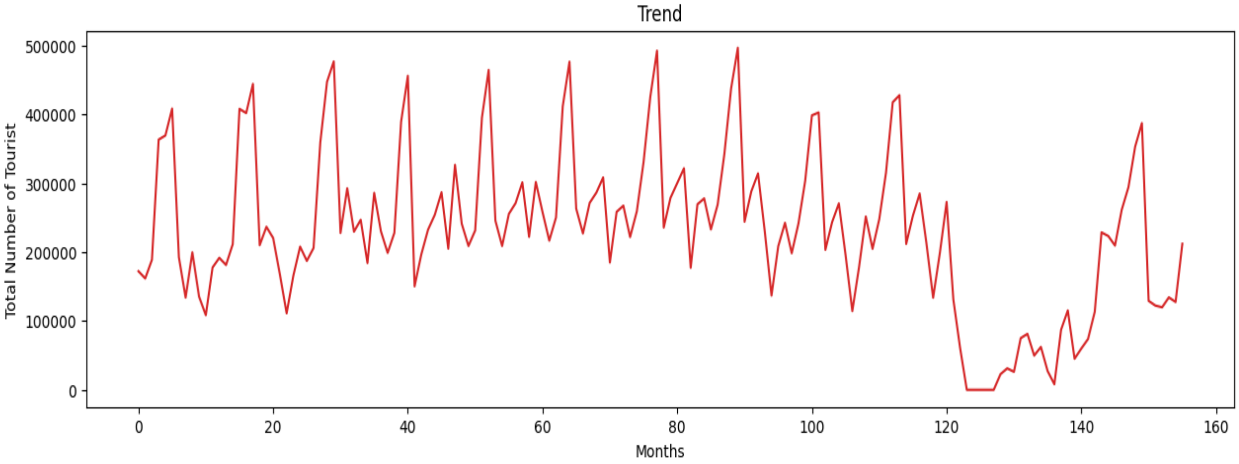


Figure 14: Number of tourists before discarding 2020 and 2021

One striking trend observation is the sharp decrease in influx of foreign tourists, particularly evident in the year 2020. This decline can be largely attributed to the severe restrictions imposed on international flights and cross-border travel. The subsequent year 2021, witnessed a near-zero presence of foreign tourists.

The tourism sector grappled with unprecedented hurdles during the COVID-19 pandemic, witnessing a sharp decline in both domestic and international tourist numbers. Given the pandemic's direct impact on these figures, the data for 2020 and 2021 was deemed unreliable. Encouragingly, 2022 showcases a promising revival in tourist arrivals, indicating a positive trajectory for the industry's recovery.

3.2 Generation of Synthetic data for years 2020 & 2021:

To ensure the integrity of the analysis, a strategic decision was made. The data from 2020 and 2021, marred by the **pandemic-induced anomalies**, has been deemed unreliable. To address this, the choice was made to discard the data for these two years entirely. In its place, synthetic data, generated from the preceding years, has been incorporated. This deliberate measure not only enhances the accuracy of our analysis but also allows for a more insightful understanding of the tourism trends post-pandemic.

4. MAKING PREDICTION

4.1 Modeling for Synthetic Data :

Two approaches are used for generation of the synthetic data

1. **Linear Regression**

   The data for 2020 and 2021 was imputed using a regression model built on data from the preceding years. This was done using **Python** and the **scikit-learn** library.

   The regression model is trained using data from years other than 2020 and 2021. It predicts missing values in the 'Indian Tourist' column for the year 2021 and fills those missing values in the 'Indian Tourist' column specifically for the year 2020 and 2021 in the prediction data. Finally, the imputed data is concatenated back to the original Data Frame.

   Results of Imputation:

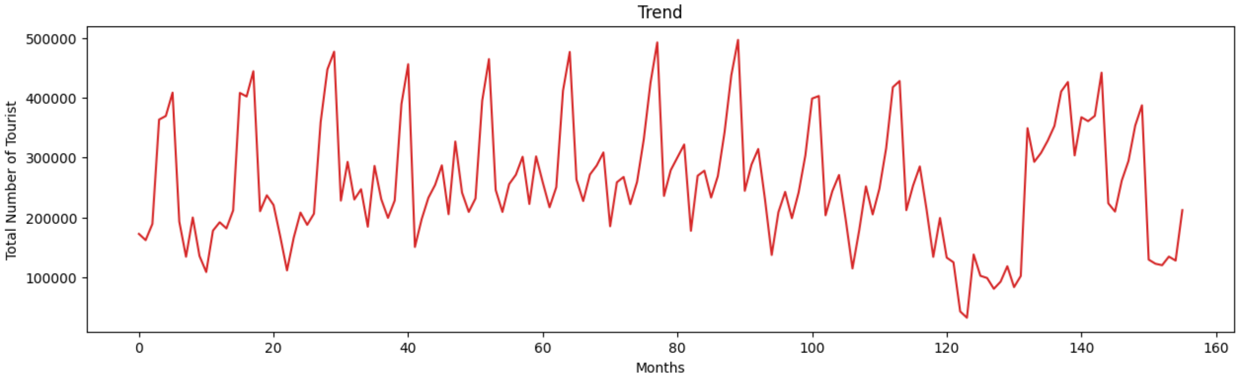


Figure 15: Number of tourists after changing data with prediction from linear regression

2. ARIMA

Autoregressive Integrated Moving Average Model is a class of statistical models for analyzing and forecasting time series data. Here the **pmdarima.arima.auto_arima** module is used for prediction.

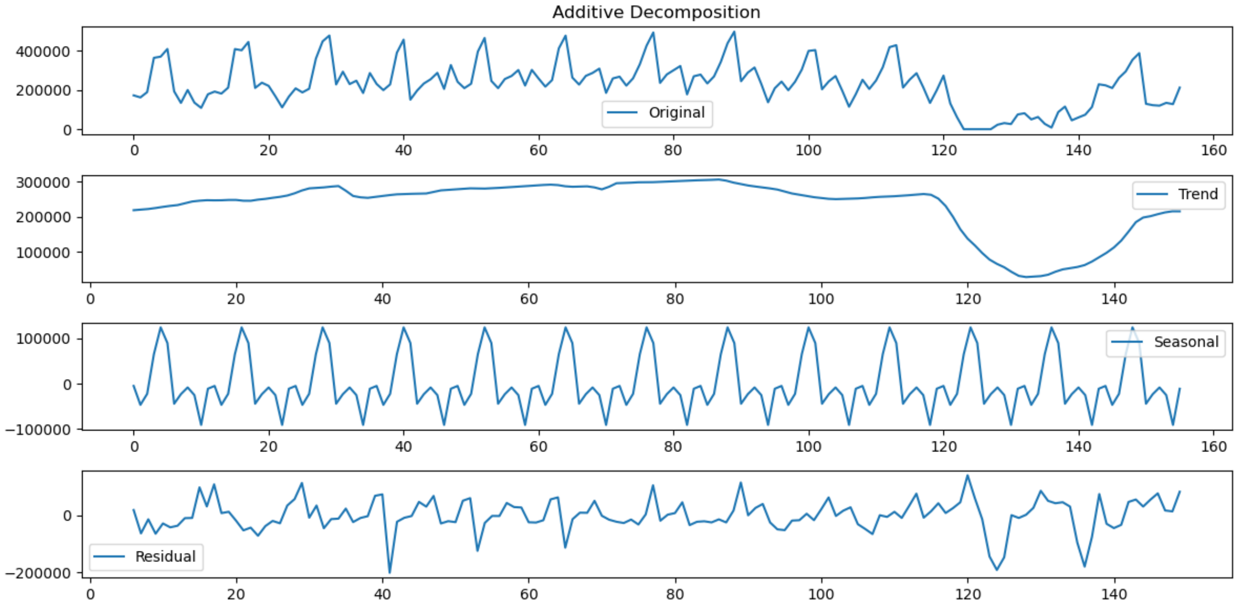


Figure 16: Additive decomposition of data for trend and seasonality identification

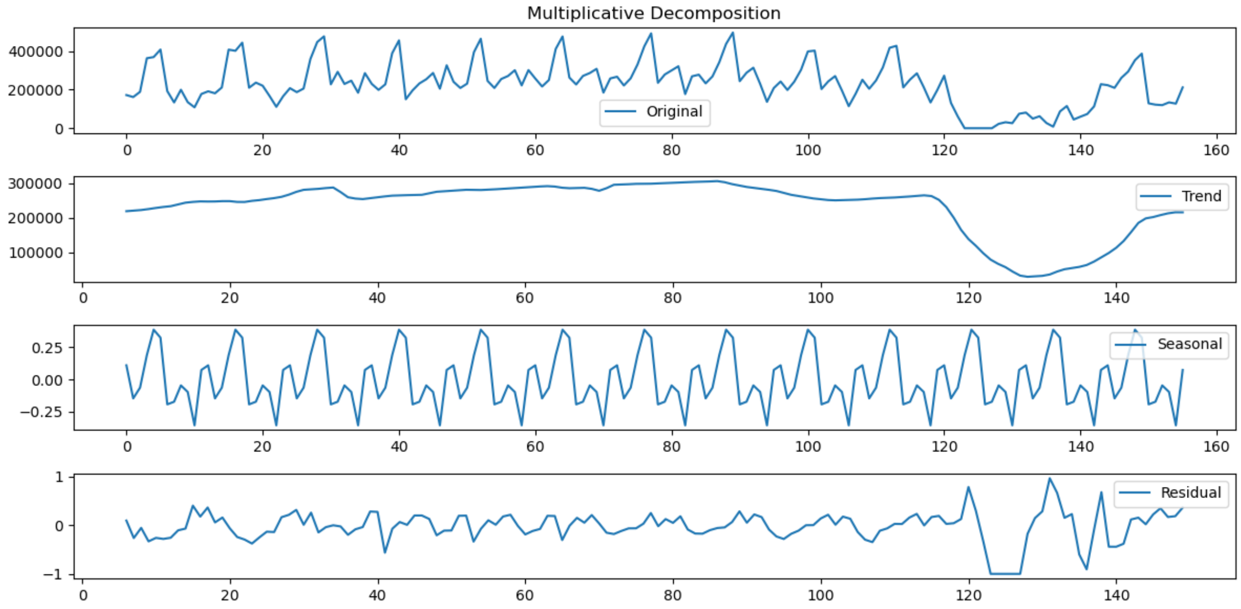


Figure 17: Additive decomposition of data for trend and seasonality identification

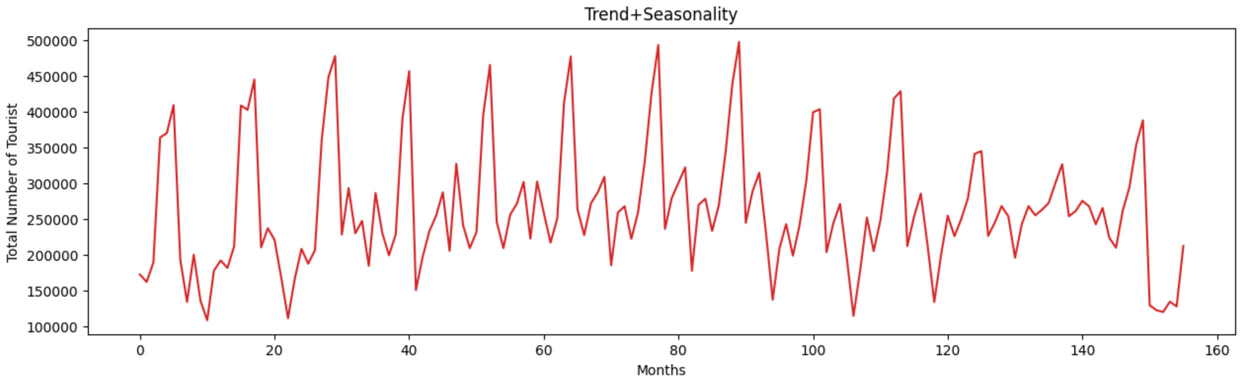


Figure 18: Plot: number of tourists after changing data with prediction from ARIMA

The evident disparity in the plots highlights the ARIMA model's superior ability to discern and replicate intricate seasonal patterns within the data, surpassing the performance of the linear regression model.

Crucially, the slight decline in numbers within the ARIMA predictions are kept for a specific purpose: mirroring the actual decrease in tourist arrivals attributed to the profound impact of the COVID-19 pandemic.

4.2 Predictive Model Performance and Analysis:

In our analysis, we employed several models to forecast tourist arrivals, each with its unique strengths and considerations.

Here's a summary of the models used and their corresponding Mean Squared Error (MSE) values:

1. **Random Forest (excluding month, with lag 1)**

   Random Forest, a robust ensemble learning method, was utilized in this model. However, the absence of month as a predictor and the lag of 1 impacted the predictive performance.
   RMSE: 82300.096

2. **XGBoost**

   XGBoost, an efficient and scalable implementation of gradient boosting, was utilized. The model exhibited a slightly higher RMSE, likely attributed to the inherent complexity of XGBoost.
   RMSE: 87273.122

3. **Random Forest (including month, no lag)**

   This model utilized Random Forest, similar to the first model, but included the month as a predictor. However, it displayed a higher RMSE, possibly due to the lack of lag incorporation.
   RMSE: 89121.7034

4. **SARIMAX (with inclusion of Month as exogenous variables)**

   Distribution: SARIMAX (Seasonal Autoregressive Integrated Moving-Average with eXogenous factors) is a time series forecasting model. Including the month as an exogenous variable, the model showed a higher RMSE, indicating its challenge in capturing the underlying patterns.
   RMSE: 91810.240

5. **SARIMAX (excluding Month as exogenous variables)**

   Similar to the previous SARIMAX model, this version excluded the month as an exogenous variable. However, it resulted in a similar RMSE, suggesting that the inclusion or exclusion of this variable did not significantly affect the forecast accuracy.
   RMSE: 91810.23

6. **Auto Arima**

   Auto Arima is an automated version of the ARIMA model. Despite its automation, it yielded a relatively higher RMSE, possibly due to the automated selection of parameters.
   RMSE: 101630.0411

7. **SARIMAX (without exogenous variables)**

   This SARIMAX model did not use any exogenous variables. The RMSE value was relatively lower, indicating a comparatively better performance compared to the previous SARIMAX models.
   RMSE: 90565.465

It is important to note that the higher RMSE values observed across models, especially during the years 2020 and 2021, can be attributed to the very low available data and unavailability of sufficient data during the COVID-19 pandemic, leading to uncertainty in the predictions. The abrupt disruptions caused by the pandemic significantly affected the tourism industry, making accurate forecasts challenging during these turbulent years.

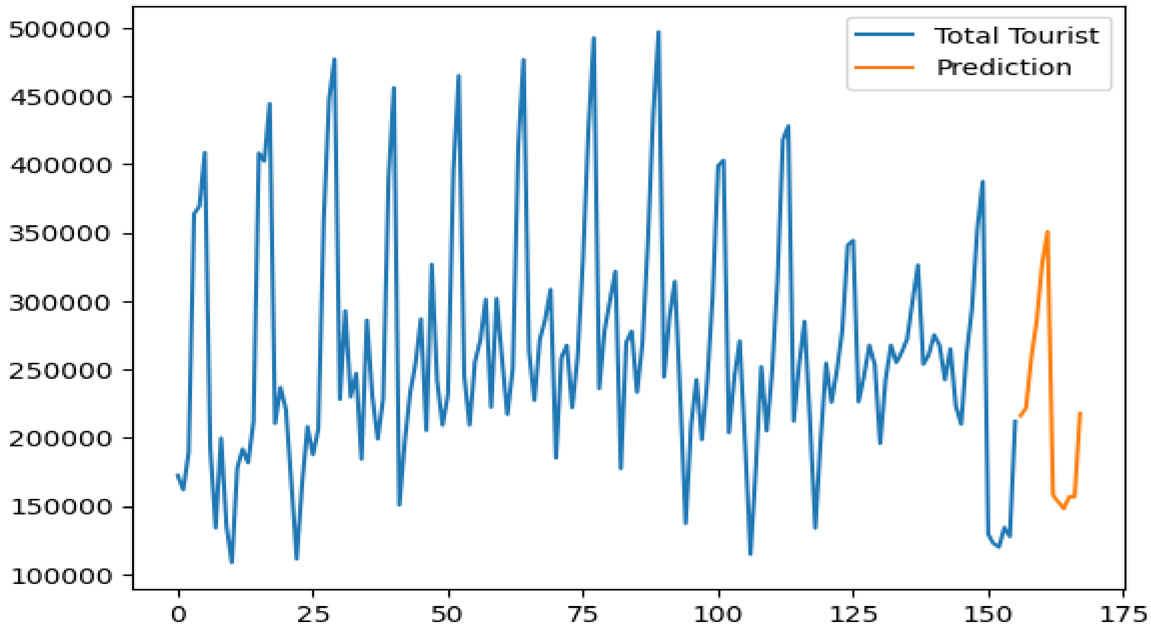
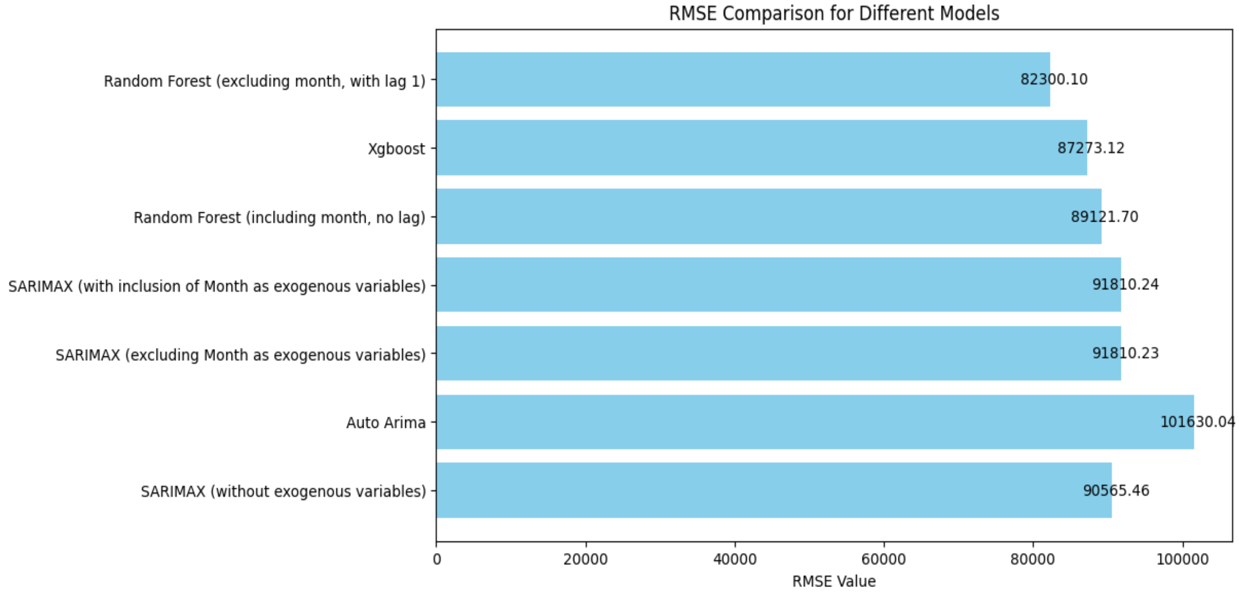


Figure 19: Plot: number of tourists after changing data with prediction for 2023

This graph plots the number of tourists predicted (for 2023), along with tourist influx from 2010 to 2022.



5. CONCLUSION

- The goal of the competition is to develop a **predictive model** that will take the data from the Internet search indices, so as to predict the future tourists arrivals which would facilitate strategic planning for the future by the authorities.

- Data-processing was done, where less-related data was either filtered or took part in the feature engineering step which created some meaningful features.

- Time series forecasting models like, ARIMA and SARIMAX, and machine learning regression models were implemented to this dataset. It was observed that the **data was seasonal** for the tourists arrivals.

ANNEXURE

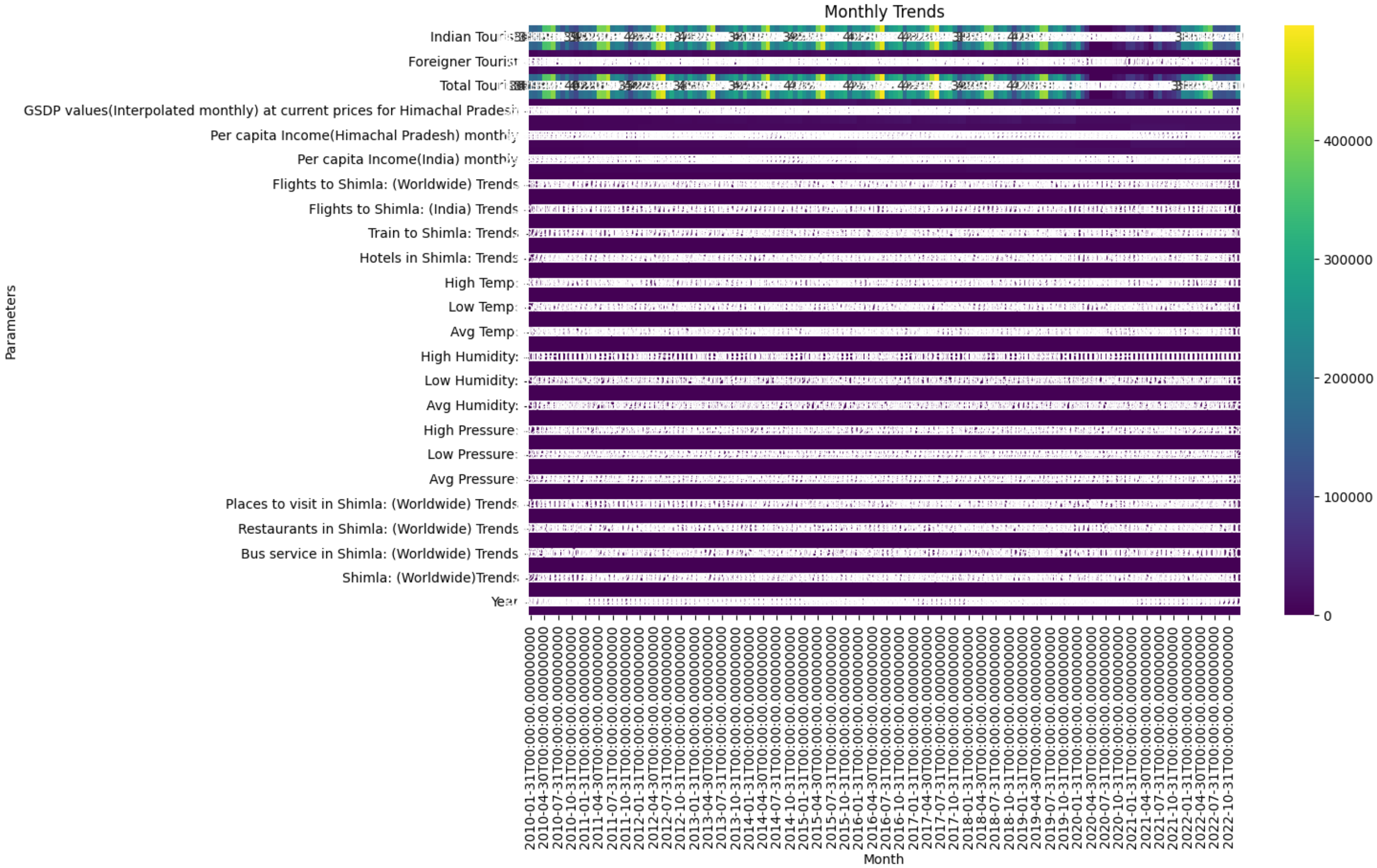
Software Stack

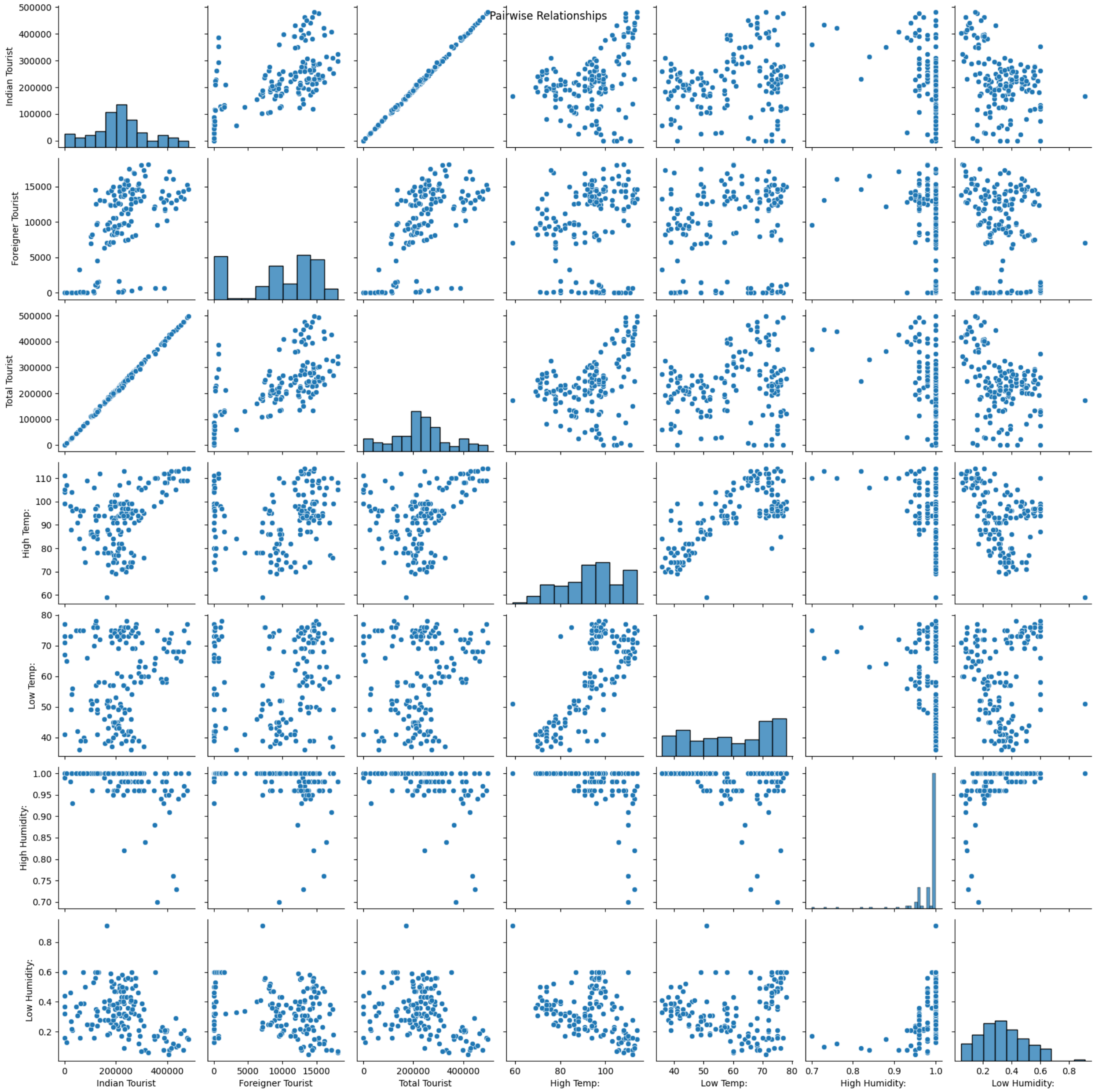
Library	Version	Purpose
Python	3.7.12	Language of choice
Numpy	1.21.5	To perform mathematical operations on arrays
Pandas	1.3.5	Data handling, manipulation and analysis
Matplotlib	3.2.2	Plotting visualizations
Seaborn	0.11.2	For making statistical graphics.
Sklearn	1.0.2	For Machine Learning and statistical modeling
pmdarima	1.8.2	time series forecasting using ARIMA and SARIMA models

REFERENCES

1) https://www.keralatourism.org/touriststatistics/
2) https://www.timeanddate.com/weather/india/shimla/historic?month=9&year=2020
3) https://mausam.imd.gov.in/shimla/mcdata/monsoon/main2018.html
4) https://www.dgca.gov.in/digigov-portal/?page=monthlyStatistics/259/4751/html&main259/4184/servicename

OTHER PLOTS




Pairwise Relationships
Indian Tourist
500000
400000
300000
200000
100000
0
Foreigner Tourist
15000
10000
5000
0
Total Tourist
500000
400000
300000
200000
100000
0
High Temp:
110
100
90
80
70
60
Low Temp:
80
70
60
50
40
High Humidity:
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Low Humidity:
0.8
0.6
0.4
0.2
0
200000
400000
Indian Tourist
0
5000
10000
15000
Foreigner Tourist
0
200000
400000
Total Tourist
60
80
100
High Temp:
40
50
60
70
80
Low Temp:
0.7
0.8
0.9
1.0
High Humidity:
0.2
0.4
0.6
0.8
Low Humidity: