SVKM's Narsee Monjee Institute of Management Studies

Mukesh Patel School of Technology Management and Engineering

FINAL REPORT

ON

FORECASTING FOREIGN TOURIST ARRIVALS IN INDIA

BY

ADITYA GAVANKAR

ROLL NUMBER: J072

SAP ID: 70041019023

FACULTY MENTOR

PROF. SIBA PANDA

INDUSTRY MENTOR

MR. SUBHASH BHOLA

A Report

ON

FORECASTING FOREIGN TOURIST ARRIVALS IN INDIA

BY

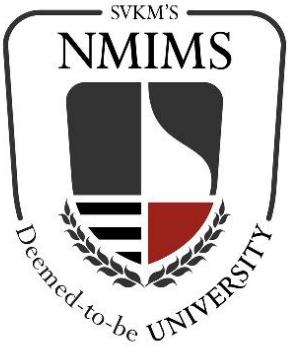
Aditya Gavankar

(Roll No: J072)

SAP ID: 70041019023

A report submitted in partial fulfillment of the requirements

of 4 years B. Tech Data Science Program



Department of Data Science

Mukesh Patel School of Technology, Management and Engineering

SVKM's NMIMS University, Mumbai

April 2022

Unit 601, 6th Floor,
A Wing, One BKC,
Plot C-66, Bandra East,
Mumbai - 400 051
Maharashtra | India

Colliers

Colliers International (India) Property Services Private Limited
CIN: U74140MH1995PTC087914

Aditya Gavankar

Intern

Sub: Management Internship

Dear **Aditya,**

We are pleased to engage you as an Intern in our Organization as per the details given below:

- Date of Commencement of Training :08.03.2023
- Date of Completion of Training :31.05.2023
- Department :Advisory Services
- Location : Mumbai BKC Office
- Stipend Amount / Month : INR 8000

During your assignment as an intern, you will be required to follow all instructions given by your mentor from time to time. In addition to the learning, the Internship Program is also meant for familiarization with the corporate culture and norms, and it is expected that you will abide by the rules & procedures of the company.

You shall abide by all protocols for COVID 19 given by local Health Authorities and safety measures that may be required, during your internship with us.

We wish you success in your assignment.

For Colliers International (India) Property Services Pvt. Ltd.

Appaya Chenanda

National Director |People & Performance

CERTIFICATE

This is to certify that the thesis entitled "**Forecasting Foreign Tourist Arrivals in India**" is a bonafide work of "**Aditya Gavankar (SAP ID: 70041019023)**" submitted to the NMIMS University in partial fulfilment of the requirement for the award of the degree of "**Bachelor of Technology**" in "**Data Science**".

_____

Prof. Siba Panda, Ph.D.

HOD – Data Science

SVKM's NMIMS University, Mumbai

ACKOWLEDGEMENT

It is with a feeling of great pleasure that I would like to express my most sincere heartfelt gratitude to **Mr. Subhash Bhola** for their steady and able guidance throughout my internship. I am greatly indebted to them for providing me an opportunity to be a part of the team at **Colliers International** (March 2022 to Present).

I express my sincere thanks to **Prof. Sarada Samantaray (HOD, Data Science)** for his guidance and for providing the necessary facilities in the department. I would like to thank **Prof. Siba Panda** for their inspiration and guidance. I am also thankful to all the staff members of the department of Data Science.

TABLE OF CONTENT

1. ABSTRACT

India is become one of the most popular tourist destinations in the world thanks to its unique climatic conditions and rich cultural heritage. It has a dramatic growth of tourism over the last 25 years and it is one of the most remarkable economic changes. India is the only nation that provides tourists with a variety of tourism options each year. The Indian government made an attempt to promote various forms of tourism there. They used a variety of strategies to promote tourism there. India intends to change its visa regulations in 2014 by allowing tourists from the majority of nations to apply for an electronic visa online. The Indian government made the decision to promote India as the "ultimate tourist spot" in an effort to boost tourism-related earnings. So, there is no doubt that the development of tourism industry in India will be very strong in the next five to ten years.

Unfortunately, due to the sphere of pandemics and infectious diseases, known as COVID-19, the tourism industry in India had become very unstable. As the contagious were emerging, the authorities had implemented travel restrictions, social distancing, lockdown which completely brought the tourism sector of India to recession. As resulting the threat of virus, foreign tourists revoked their tour and agitation tumbled the tourist graph significantly.

So, in this project, the major objective is to understand the general trends in the data, gain some quick insights, and then predict and forecast the foreign tourist arrivals in India of the given time series data. Moreover, this project also attempts to understand the impact of COVID-19 on Indian tourism sector. We will first explore and understand the given time series data by applying the statistical method and visualizations. Then, we will check whether the given time series data is stationary or non-stationary. This method is very much important so that we can assume that future statistical properties are the same or proportional to the current statistical properties. To this end, we will design the ARIMA model (Auto-Regressive Integrated Moving Average) to predict and forecast the given time series data. This model explains a given time series based on its own past values so that it can be used to forecast future values.

2. COLLIERS INTERNATIONAL

Colliers is a Canada-based diversified professional service and investment management company.

The firm provides services to commercial real estate users, owners, investors, and developers; they include consulting, corporate facilities, investment services, landlord and tenant representation, project management, urban planning, property and asset management, and valuation and advisory services.

The organization serves the hotel, industrial, mixed-use, office, retail, and residential property sectors.

It has approximately 18,000 employees in more than 400 offices in 65 countries.

Colliers is a full-service real estate brokerage firm that operates in 67 countries and is traded on the NASDAQ stock market exchange under the symbol CIGI.

The firm has headquarters in Toronto, Ontario. Annual revenues were $4.09 billion in 2021.

In 2010, Colliers consolidated its franchises under a single name, hoping to increase their market share, according to the New York Times.

On 1 June 2015, it was announced that Jay S. Hennick was appointed Chairman and Chief Executive Officer. First Services and Colliers split into two independent publicly traded entities.

3. INTRODUCTION

India's tourism industry is a big economic multiplier and is growing in significance as the nation aims for rapid economic growth and the creation of new jobs. India is filled with natural beauty in every nook and cranny. The vast landscape of this country is dotted with the most different populations, cultures, and topographies. India also offers a wide range of geographical areas, top-notch tourist destinations, and specialized travel services, such as eco-tourism, heritage tourism, adventure tourism, medical tourism, etc.

After the nation's independence, planning for Indian tourism began. Sir John Sargent led the committee in 1945. He was serving as the Indian government's educational advisor at the time. Then, India's tourism industry began to grow steadily. The methodology for tourism planning evolved in the second and third five-year plans. The sixth five-year plan places a lot of emphasis on using tourism as a tool for building social cohesion, economic development, and maintaining peace.

After 1980, the tourism industry expanded as a source of employment, income, foreign currency, and leisure time. Through a variety of crucial initiatives, the government has supported the tourism sector. But the growth of tourism didn't start until around 1980. The government then launched a number of significant projects. In 1982, the Indian government published its initial tourism policy. The goal of the First Tourism Policy was to enhance India's image as a country with a proud past, a thriving present, and a bright future by promoting sustainable tourism as a tool for social and economic inclusion. The policies established to do this would focus on six major categories: Swagat (welcome), suchana (information), suvidha (facilitation), suraksha (safety), sahyog (cooperation), and Samrachana (infrastructure development). Additionally given priority in this policy are the creation and promotion of tourism-related products, environmental preservation, and cultural heritage preservation. India's planning commission recognized tourism as an industry in 1982. The tourist sector will be added to the concurrent list in accordance with the new policy because doing so would grant it constitutional validity and allow the central government to pass legislation that will control the activities of various service providers in the tourism sector. When the tourism sector was included in the Concurrent List of the Indian Constitution, the first tourism policy underwent a significant change.

In order to develop a long-term strategy for the tourism sector, the Indian Planning Commission established the National Committee on Tourism in 1986. Among the concessions, the central government granted for the sector were additional tax breaks on foreign exchange revenues from tourism. The Tourism Development Finance Corporation was founded in 1987 with a capital pool of Rs. 100 crores to provide commercial financing for the tourism industry. In 1992, a new National Action Plan for Tourism was announced.

In order to advance tourism planning in India, the 8th Five-Year Plan (1992-1997) placed a major emphasis on the private sector's growing involvement in the industry. The Indian government announced its national tourism policy in 2002. This policy is built on a multidimensional approach that includes cutting-edge marketing techniques, strengthening hospitality sector capacity, accelerating the development of tourism projects, and integrated

tourism circuits. The National Tourism Policy of 2002's main objective is to position tourism as a vital engine of economic growth.

The government aims to achieve this goal by fostering domestic and international inbound travel, establishing new tourist routes, constructing tourist infrastructure, fostering agro-rural tourism, and promoting new tourist sites.

Following the 2002 National Tourism Policy, the 10th Five-Year Plan (2002-2007) encouraged skill development by encouraging training programs in the hospitality and catering industry sectors. The eleventh five-year strategy promoted beach and Himalayan adventure travel. Wellness tourism encompasses practices like Ayurveda, conventional craft shops, and pilgrimage destinations. As a result, the 11th Five-Year Plan has increased spending on tourism development. (2007-2012). The national tourism policy from 2002 is being expanded as part of the 11th five-year plan to promote cooperation between the federal, state, and private sectors. Later, the Ministry of Tourism launched other campaigns, including the "Incredible India Campaign," the "Atithi Devo Bhavah program," and the "Visit India Campaign," to encourage tourism in the nation. (2009).

The 12th Five-Year Plan introduced a new aspect of tourism participation. (2012–2017). To increase the net benefits of tourism to the poor and ensure that tourism expansion aids in the fight against poverty, the plan emphasizes the need to adopt a "pro-poor tourism" approach.

However, the worldwide coronavirus epidemic had a devastating impact on the travel and tourism sector. The tourism industry has suffered a detrimental impact in every area, including aviation, hotels, transportation, tour guides, and restaurants. The situation in India is not all that different. As early as February 2020, there was a decrease in traveler arrivals or movements. Additionally, job losses in both the formal and informal sectors were inevitable. Due to the three COVID-19 waves, 21.5 million people in the tourism sector have lost their jobs.

The lockdown, which disrupted the lives of billions of people, led to an economic catastrophe. Only 24.80% of the half-year's revenue for the Indian enterprises remained from the half-year. Arunachal Pradesh or Manipur's nominal GDP for the 2017–18 fiscal year, respectively, are essentially similar to the loss of income over six months, which comes to Rs. 23636.27 Crore. The loss suffered in the first half of the year after Covid is equal to 75.20% of the half-year before. Companies' half-year net loss is equal to 74.70% of their present half-year income. In India, there are hundreds of tiny, unorganized travel and tourism service providers. Many people who work in this industry were negatively impacted, and many now face financial difficulties.

The pandemic is currently renewing global interest in travel. Worldwide visitors to Asia are expected to increase by 100% between 2022 and 2023. Flight reservations to the region, especially to India, are rising as well, showing that after the COVID-19 outbreak, people are once again confident in air travel. Hotels have developed all-inclusive packages and offer to capitalize on the expanding travel trend. These initiatives will entice clients to pack their bags and take overseas trips like they formerly did.

"The outlook for the next decade is looking very positive with India accounting for one in five of all new Travel & Tourism jobs globally," stated Julia Simpson, President & CEO of the WTTC (World Travel and Tourism Council). Before the outbreak, India's travel and tourist industry made about 7% of the nation's GDP ($15.7 trillion, or $212 billion), but that percentage decreased to just 4.3% ($9.2 trillion, or $124 billion), a staggering 41.7% decline, in 2020. The sector's potential to boost the nation's economy by 1% from 2019 to close to $15.9 trillion (about $215 billion) in 2022. There will be about 35 million jobs in the travel and tourism sector this year, an 8.3% growth in employment. Additionally, the forecast indicates that over the next 10 years, the sector is expected to create more than 24 million jobs, or more than 2.4 million new positions yearly. In 2019, the sector supported more than 40 million jobs, but as the COVID-19 epidemic ravaged the sector in 2020, that number fell to a little over 29 million. In the most recent year, it contributed $178 billion (or 13.2 trillion) to GDP, up 43.6% from the year before.

Over the next ten years, India's travel and tourism industry is expected to grow at an average annual rate of 7.8%, outperforming the nation's 6.7% pace of overall economic growth, and eventually makeup 7.2% of the GDP, or more than $33.8 trillion (U.S. $457 billion). Even though the travel and tourism industry added just under 3 million workers in 2021 (bringing the total number of employees up by 10.2% to more than 32 million), there were still less than 8 million more positions available than in 2019. Imagine that the consequences of the Omicron variation hadn't led to a global slowdown in this industry's recovery and the reinstatement of harsh travel restrictions by many nations. In that situation, the sector's contribution to job creation and economic growth could have been greater.

India's tourist industry is undoubtedly growing, but this should not cause authorities to get careless about the new developments that are happening quickly in this industry. For the new sorts of tourism, proper planning and infrastructure development are necessary.

4. LITERATURE REVIEW

Here are some literatures review in which it has become a highlight in my project.

4.1. Forecasting the demand factors in the tourism industry

Tourism makes the ideas and opinions held by people who form their decisions about going on a trip, where to go or where not to go, what to do or not to do, and how to relate to other visitors, locals, and service personnel [1]. Tourism can cover all kinds of trips, as long as it is both sightseeing and recreation. Tourism is a trip that takes place for a while, organized from one place to another, not to do business or to make a living in the areas visited but solely to enjoy a trip for sightseeing and recreation or to fulfill diverse desires [2]. Almost every organization, large and small, private and public, make explicit and implicit assumptions because an organization should plan to meet the potential conditions that have imperfect knowledge. Furthermore, modern tools for forecasting, together with computer capabilities, have become essential for organizations operating in the modern world [3]. Demand, in this case, more generally referred to as market demand, implies there is a demand on the market for certain goods at a specific price and at some time in any event [4]. Forecasting is necessary for the tourism industry. Accurate forecasting provides help to government and industry stakeholders to help them make important decisions and prevent waste and inefficiency of tourism capital while reducing risk and uncertainty [5][6]. The determinant of macroeconomic factors such as GDP, bilateral trade volumes, fuel prices, and exchange rate can be used as a guideline if it becomes a supporting variable for coming to a specific country [7].

4.2. Forecasting tourist arrival with autoregression integrated moving average (Part 1)

Existing quantitative methods for tourism forecasting can be classified into three categories: time series, econometric, and AI [8][9]. Time series models provide simplicity by employing a lag of Internet data as an explanatory variable [9]. This model can provide accurate predictions, notably for short-term forecasting horizons [10][11]. The most commonly used time series models include autoregressive, autoregressive integrated moving averages, and seasonal autoregressive integrated moving averages [8][9]. The econometric models are concerned with the causality of various explanatory variables [12][13]. The previous studies demonstrated that econometric models can improve accuracy in more extended time horizons [14][10]. However, all variables included in these models should be stationary to avoid spurious results [15][13][8]. The autoregressive distributed lag model, time-varying parameter, and vector autoregression are among the most popular econometric models [8][9].

4.3 Forecasting tourist arrival with autoregression integrated moving average (Part 2)

There is another vast body of research on modeling and forecasting tourism arrivals, using different econometric techniques. Among the plethora of publications on this subject, many authors supported the autoregressive integrated moving average (ARIMA) type of models, first proposed in Box and Jenkins in 1976 [16]; In 2013, Stellwagen E. & L. Tashman provide an excellent tutorial to their method [17]. In 2003, Goh and Law found that multivariate ARIMA (ARIMAX) is one of the most accurate among eight types of time-series models projecting tourism arrivals to Hong Kong [18]. Preez & Witt demonstrated the advantage of simple ARIMA over univariate and multivariate state space modeling for forecasting tourism arrivals to Seychelles [19]. In 2011, Athanasopoulos et al concluded that pure ARIMA was among the most accurate models, and even out-performed models with explanatory variables (although recognizing that the latter could be due to the models' misspecification's) [20]. Kim et al. emphasized that in tourism forecasting, point predictions should be complemented with the confidence intervals, which can enable the authorities and tourism promotion agencies to plan with higher flexibility; they evaluated several time-series models for tourism forecasting and found that SARIMA produced accurate point forecasts and narrow prediction intervals [21]. In 2015, Gunter and Önder concluded that the univariate models of ARMA (1,1) were the most accurate in forecasting tourism arrivals to Paris from the USA and the United Kingdom [10]. In 2014, Claveria and Torra favored ARIMA models for predicting tourism demand in Catalonia [22].

5. METHODOLOGY

This section outlines the plan and method that how the data is collected, the pre-processing and cleaning of data, checking the trend and seasonality of the time series data, checking whether the time series is stationary using the **Augmented Dickey-Fuller (ADF)** test, and finally forecasting the time series using **Autoregressive Integrated Moving Average (ARIMA)** model.

5.1 Data collection

For this project, I have created and used the monthly time series data that contains the total number of international tourists that arrived in India from the year 1981 to 2022, along with the total number of passengers based on their continents, ages, gender, and travel accommodation. This time series data is taken from the annual reports of the "**Indian Tourism statistics**" for the duration till 2022 which are available at the Indian **Ministry of Tourism** [23].

Every year, the **Ministry of Tourism** releases "**India Tourism Statistics**", a report that provides statistics on both domestic and international travel, as well as information on hotels that have been categorized. A brief booklet titled "**Tourism Statistics at a Glance**" is also released, providing the most recent and updated vital statistical data. The Ministry also projects month-by-month data for FTAs (foreign tourist arrivals) and FEE (foreign exchange earnings) from tourism.

Here is the list of variables in this dataset:

- Date
- FTA
- North America
- C&S America
- Western Europe
- Eastern Europe
- Africa
- West Asia
- South Asia
- South East Asia
- East Asia
- Australasia
- Male passenger
- Female passenger
- Air Transport
- Land Transport
- Sea Transport
- 0-14 age passenger
- 15-24 age passenger

- 25-34 age passenger
- 35-44 age passenger
- 45-54 age passenger
- 55-64 age passenger
- 65 and above age passenger

5.2 Exploratory Data Analysis

After cleaning and pre-processing the given data using the methods in Python, here are some data visualizations that I have applied to understand and gain some quick insights into the given data before forecasting. These visualizations are performed in Power BI and Python.

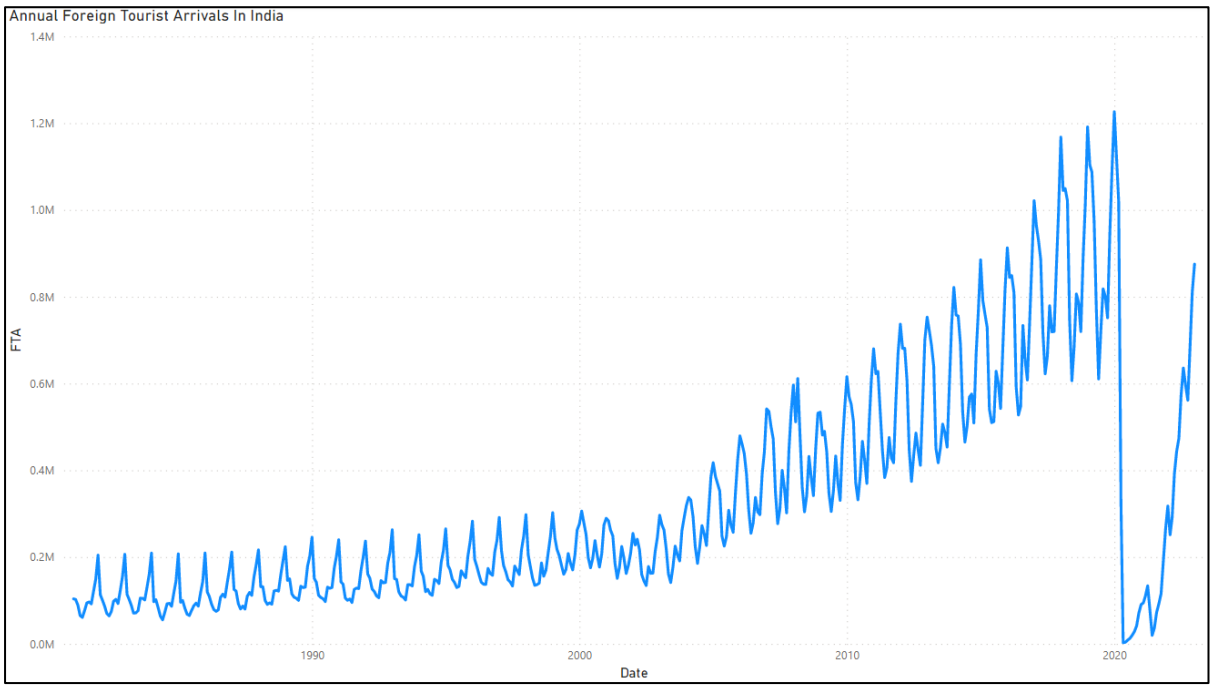


Figure 1: - Annual Foreign Tourist Arrivals in India

The above figure shows how foreign tourists are arriving in India periodically. From 1981 to 2001, the growth of foreign tourist arrivals had to go consistently. From 2003 onwards, the growth of foreign tourist arrivals has been increasing exponentially, more compared to the last 20 years of data. Over 17.9 million foreign tourists arrived in India in 2019 compared to 17.4 million in 2018, representing a growth of 3.5%, out of the maximum number of foreign tourist arrivals in any year. However, in 2020, we see a sudden downfall in the graph, due to COVID-19. The number of foreign tourist arrivals has fallen to 2.74 million in 2020, around a 44.5% decline from the last year, and also to 1.52 million in 2021. The data reached an all-time high of 1,226,398 foreign tourists in December 2019 and a record low of 2,820 foreign tourists in April 2020.

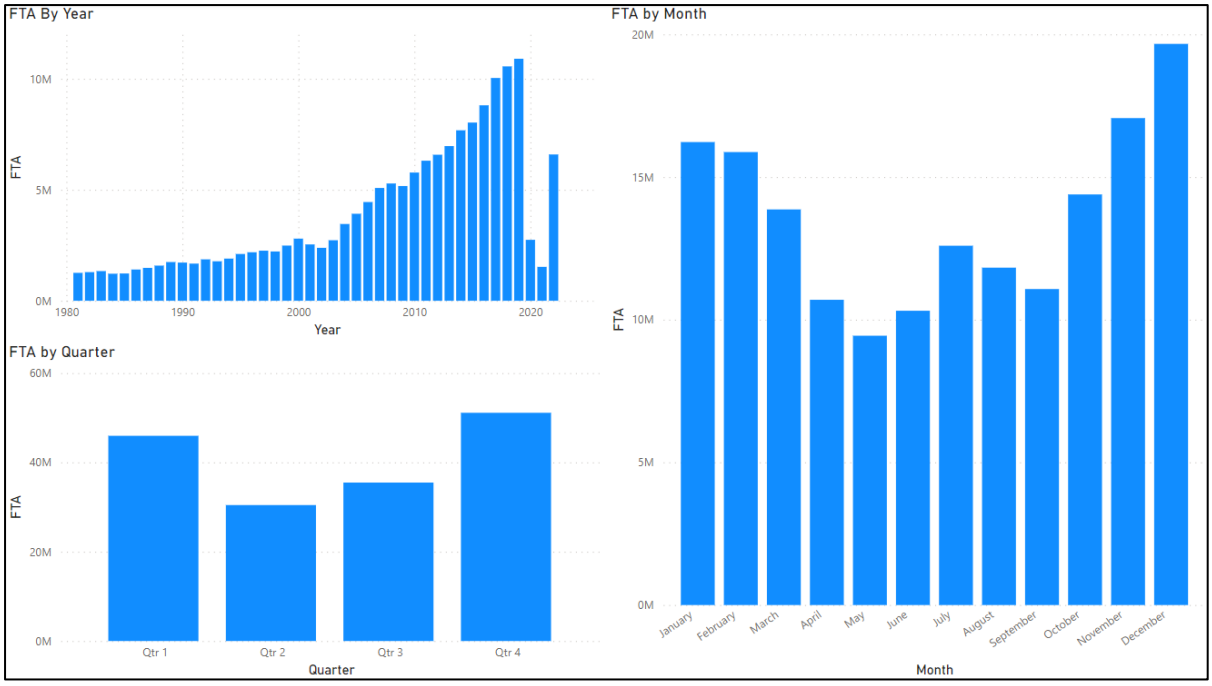


Figure 2: - Foreign Tourist Arrivals by Year, Quarter and Month

Every year, over 1 million foreign tourist arrives in India. Most of them are seen to arrive in fourth quarter or in month of November or December due to the popularity of winter seasons. Blessed with a tropical climate, India is one of the most popular winter holiday destinations in the world.

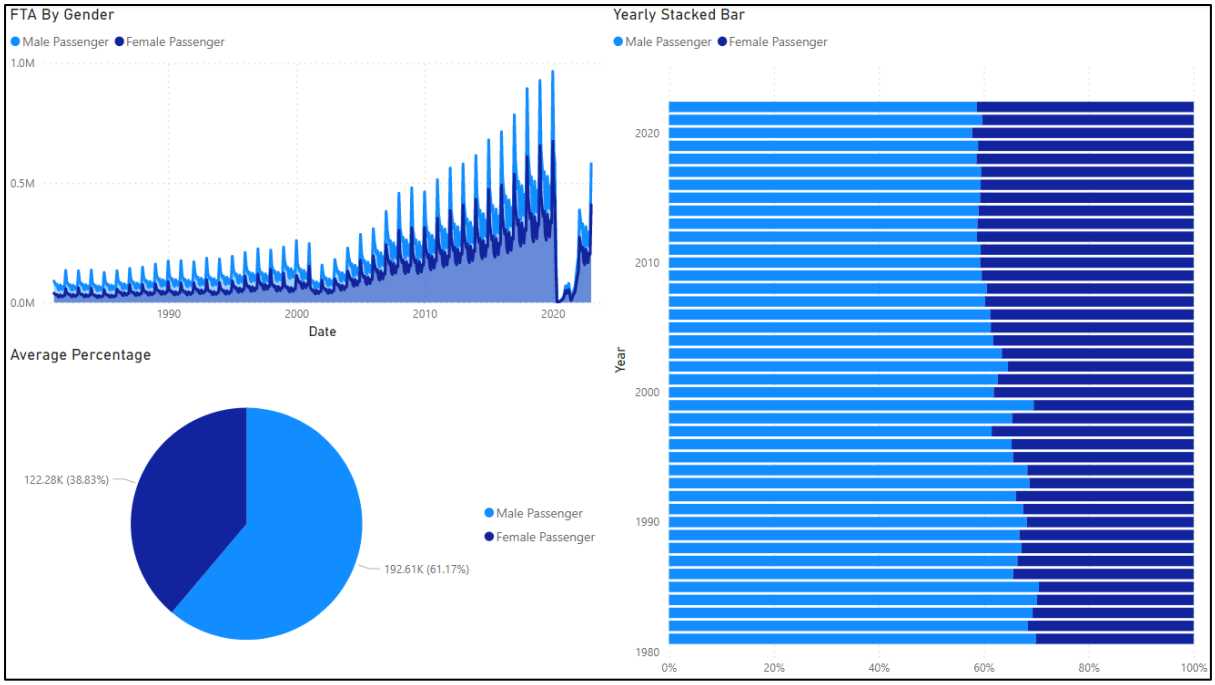


Figure 4: - Data visualization on Foreign Tourist Arrivals by Gender

Most of the foreign tourist that arrived in India are male, with around 61% of the total international tourist, compared to female, with around 38% of the total international tourist.

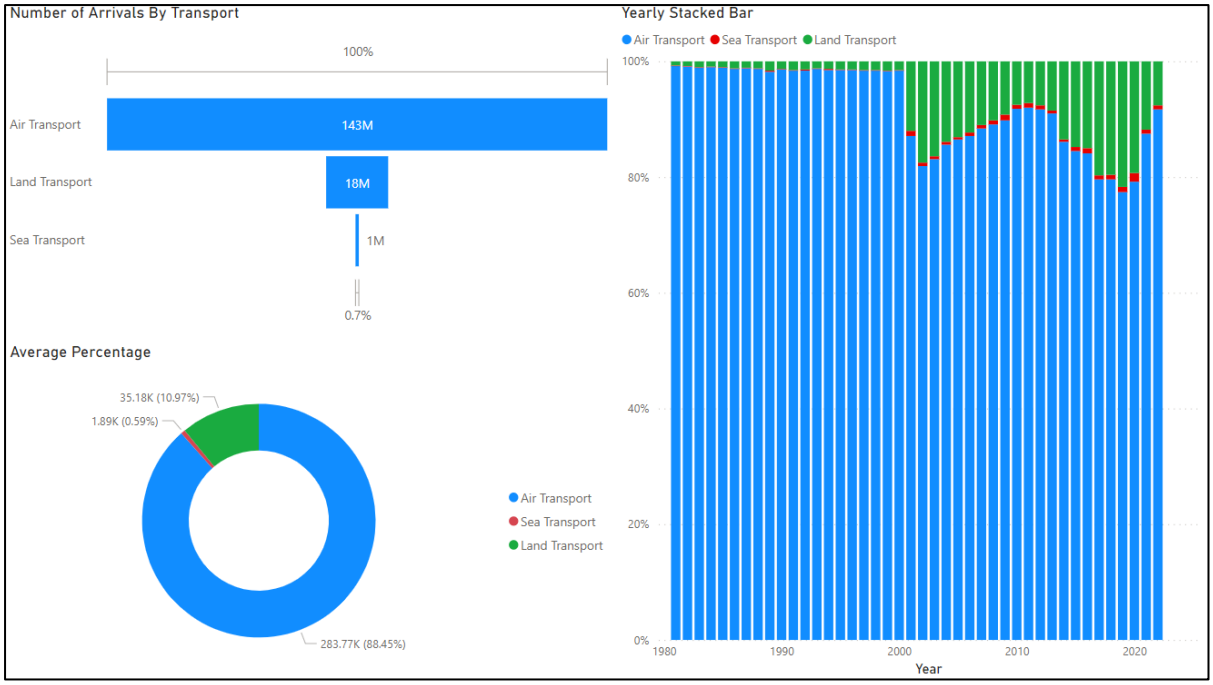


Figure 6: - Data visualization on Foreign Tourist Arrivals by Transport

Most of the foreign tourist prefers to travel in India by air (88.45% of total tourist). From 2001 onwards, it is shown that more few foreign tourists prefer to travel by land (10% of total tourist), and very few to none of the foreign tourists travel by sea.

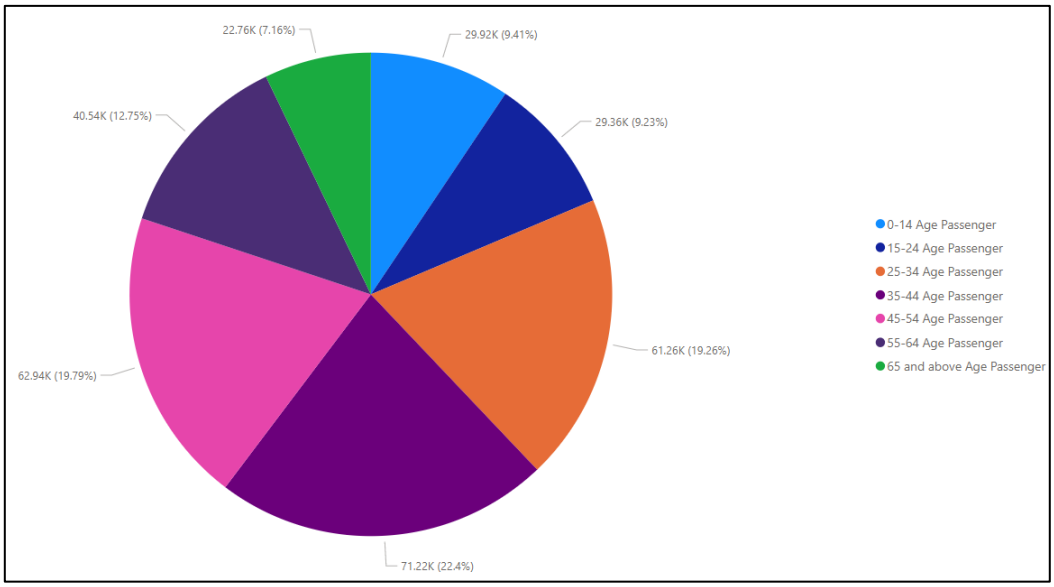


Figure 5: - Average Percentage of Foreign Tourist Arrivals by Age

The foreign tourist of the age ranging from 25 to 54 are shown to be the one arriving in India mostly, due to their job, vacation, or stay. 65 or above age foreign tourist are shown to be the minimum passengers, before foreign tourist of the age below 14 years.

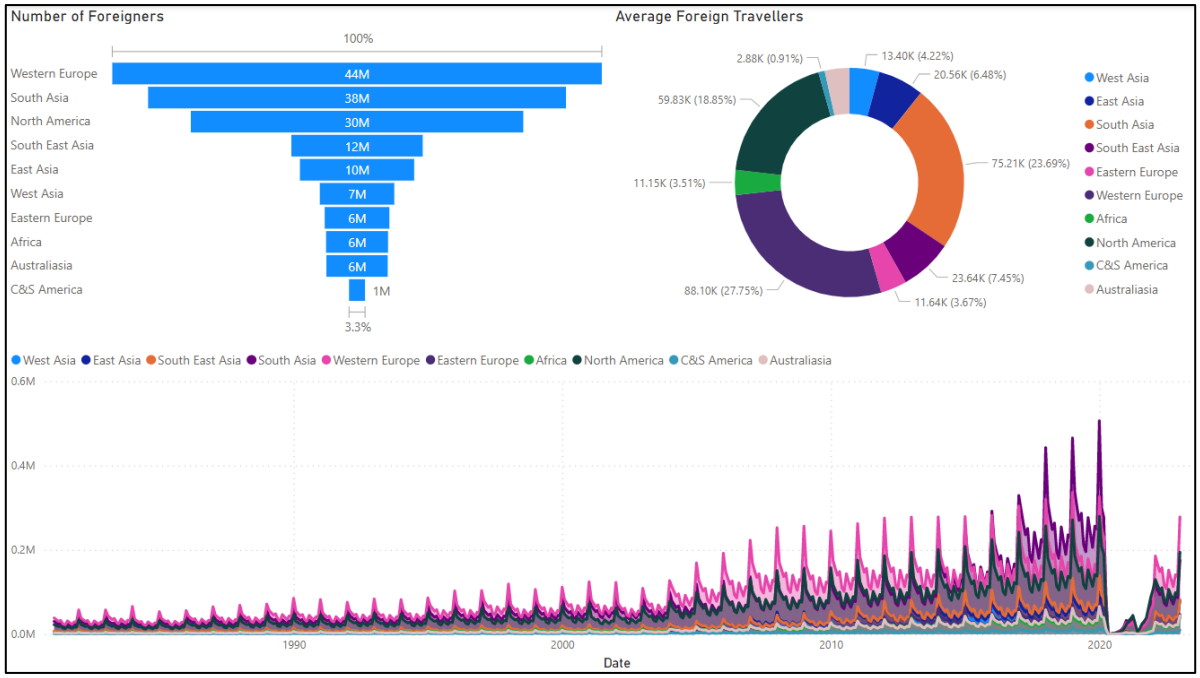


Figure 3: - Data visualization on Foreign Tourist Arrivals by Continents

It is seen that foreign tourist from Western Europe have the highest number of arrivals in India with a total of 44 million foreigners, around 27% of the total international arrivals. People from UK, France, Germany and Spain are probably the most well-known foreign tourists. The second highest number of arrivals are the foreign tourist from South Asia (23% of the total arrivals), followed by foreign tourist from North America (18% of the total arrivals). The lowest number of arrivals is Central and State America with a total of around 1 million foreigners (0.91% of the total arrivals).

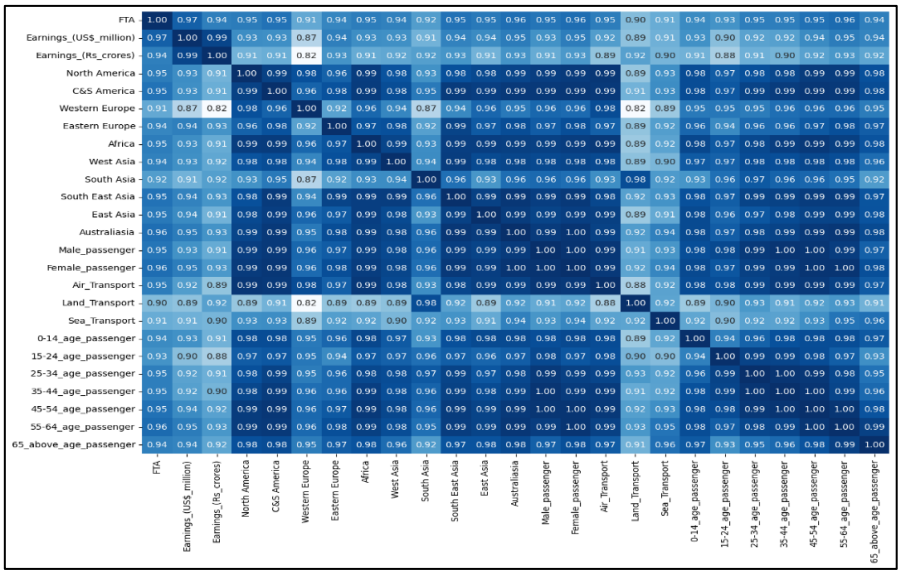


Figure 7: - Correlation Plot

Here, I used the correlation plot to show the relationship between all the variables present in the data, to summarize the amount of data and its pattern, and to check whether the data is ready to be analysed or not.

5.3 Time Series Decomposition

After exploring the data, I used time series decomposition technique to extract multiple types of variation from the time series data. There are three important components in the temporal data of a time series:

- **Seasonality** is a recurring movement that is present in our time series variable.
- **Trend** can be a long-term upward or downward pattern.
- **Noise** is the part of the variability in a time series that can neither be explained by seasonality nor by a trend. When building models, we would end up combining different components into a mathematical formula. Two parts of such a formula can be seasonality and trend. A model that combines both will never represent the values of data perfectly: an error will always remain. This is represented by the noise factor.

Here, I did a classical decomposition of a time series by considering the series as an additive or multiplicative combination of the base level, trend, seasonal index and the noise. The **seasonal_decompose** in **statsmodels** implements this conveniently. It has visualized the data using a method called time-series decomposition that allows us to decompose our time series into 3 distinct components: - trend, seasonality, and noise.

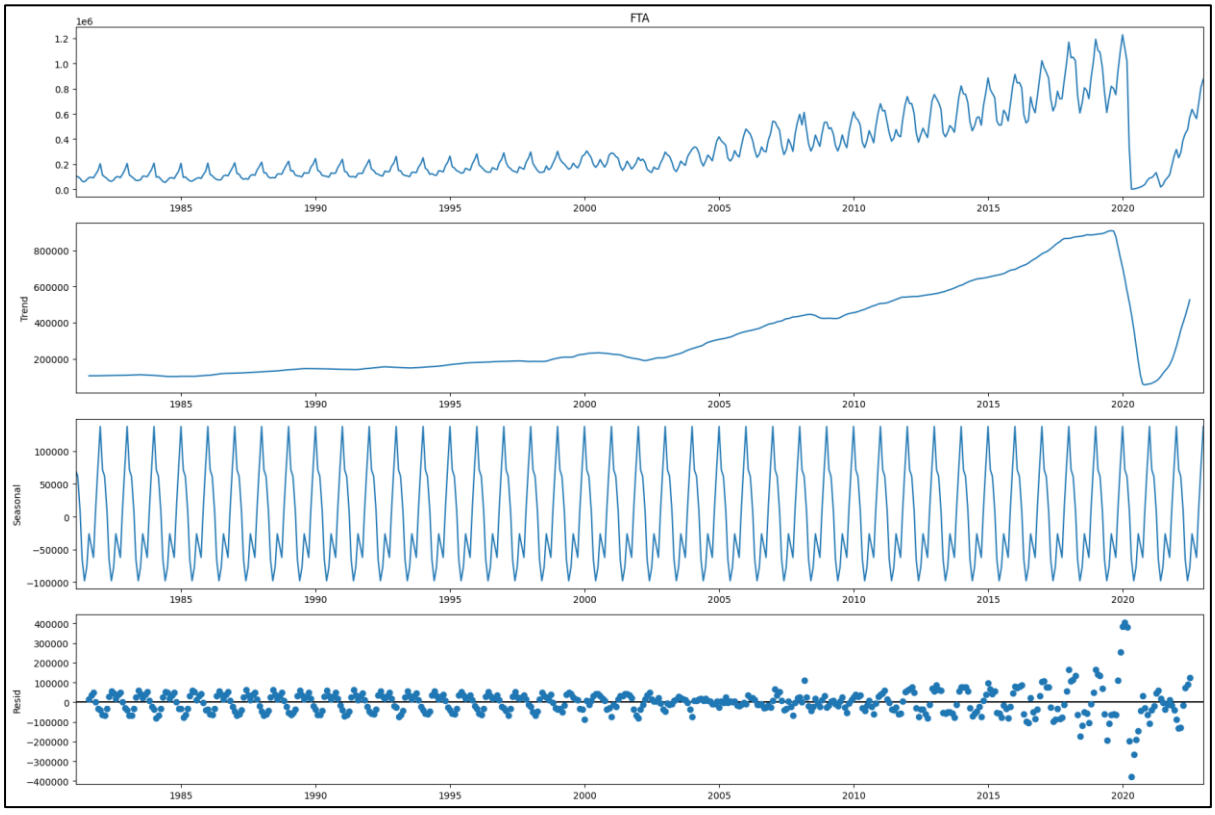


Figure 8: - Time Series Decomposition

The plot below clearly shows that number of foreign tourist arrivals are slightly unstable, along with its obvious seasonality. We can still see some noises.

5.4 Stationarity

Now that we have checked the trends and seasonality, we will see how good or bad our time series data is in terms of **'stationarity'**.

'Stationarity' is one of the most important concepts while working time series data.

Data points are often **non-stationary** or have means, variances and covariances that change over time. Non-stationary behaviours can be trends, cycles, random walks or combinations of the three. Non-stationary data, as a rule, are unpredictable and cannot be modelled or forecasted. The results obtained by using non-stationary time series may be spurious in that they may indicate a relationship between two variables where one does not exist. In order to receive consistent, reliable results, the non-stationary data needs to be transformed into stationary data.

A time series is said to be **stationary** if its statistical properties such as mean, variance remain constant over time. Though stationarity assumption is taken in many time series models, almost none of practical time series are stationary. Most of the time series models work on the assumption that the time series is stationary. Intuitively, we can say that if a time series has a particular behaviour over time, there is a very high probability that it will follow the same in the future.

Advantages:

- It is simple to predict as we can assume that future statistical properties are the same or proportional to current statistical properties.
- Most of the models we use in time series analysis assume covariance-stationarity. This means the descriptive statistics these models predict (e.g., means, variances, and correlations) are only reliable if the time series is stationary and invalid otherwise.
- "For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. And if the mean and variance of a series are not well-defined, then neither are its correlations with other variables."
- With that said, most time series we encounter is NOT stationary. Therefore, a large part of time series analysis involves identifying if the series we want to predict is stationary, and if it is non-stationary, we must find ways to transform it such that it is stationary.

Here is the method we used to check stationarity:

- **Statistical Tests**: We will use Augmented Dickey-Fuller test to check if the expectations of stationarity are met or have been violated.
- **Rolling Plots**: We will review a time series plot (moving average or moving variance) of the data and visually check if the time series is stationary.

5.4.1 Augmented Dickey-Fuller (ADF) Test

The ADF test belongs to a category of tests called **'Unit Root Test'**, which is the proper method for testing the stationarity of a time series. **Unit root** is a characteristic of a time series that makes it non-stationary. Technically speaking, it is said to exist in a time series of the value of alpha = 1 in the below equation.

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

where,

- Y_t is the value of the time series at time 't'
- X_e is an exogenous variable (a separate explanatory variable, which is also a time series).

It means that the presence of a unit root means the time series is non-stationary. Besides, the number of unit roots contained in the series corresponds to the number of differencing operations required to make the series stationary.

A **Dickey-Fuller** test is a unit root test that tests the null hypothesis that α=1 in the following model equation. The alpha is the coefficient of the first lag on Y.

Null Hypothesis (H_0): alpha=1

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

where,

- $y_{(t-1)}$ = lag 1 of time series
- delta $Y_{(t-1)}$ = first difference of the series at time (t-1)

Fundamentally, it has a similar null hypothesis as the unit root test. That is, the coefficient of $Y_{(t-1)}$ is 1, implying the presence of a unit root. If not rejected, the series is taken to be non-stationary.

The **Augmented Dickey-Fuller** test evolved based on the above equation and is one of the most common forms of Unit Root test. It expands the Dickey-Fuller test equation to include high order regressive process in the model.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p} + e_t$$

We have only added more differencing terms, while the rest of the equation remains the same. This adds more thoroughness to the test.

The null hypothesis however is still the same as the Dickey Fuller test. Since the null hypothesis assumes the presence of unit root, that is alpha=1, the p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis. Thereby, inferring that the time series is stationary.

The **statsmodel** package provides a reliable implementation of the ADF test via the **adfuller()** function in **statsmodels.tsa.stattools**. It returns the following outputs:

- The p-value
- The value of the test statistic
- Number of lags considered for the test
- The critical value cut-offs.

When the test statistic is lower than the critical value shown, we can reject the null hypothesis and infer that the time series is stationary. So, I have used this method and found out that the p-value is greater than its significant value, so our time series is non-stationary. But that does not mean we could not model our time series data; we can still convert it to stationary by these following methods:

- Differencing Transformation
- Smoothing Technique

Differencing is a method of transforming a time series dataset. It can be used to remove the series dependence on time, so-called temporal dependence, including trends and seasonality. It is performed by subtracting the previous observation from the current observation. Inverting the process is required when a prediction must be converted back into the original scale. This process can be reversed by adding the observation at the prior time step to the difference value. In this way, a series of differences and inverted differences can be calculated.

Smoothing techniques are kinds of data pre-processing techniques to remove noise from a data set. This allows important patterns to stand out. In some analysis, smoothed data is preferred because it generally identifies changes in the values compared to unsmoothed data. The idea behind data smoothing is that it can identify simplified changes to help predict different trends and patterns. It acts as an aid for statisticians or traders who need to look at a lot of data.

For this project, I have used two major smoothing techniques:

- Moving Average Smoothing
- Exponential smoothing

Moving average smoothing: It is a simple and common type of smoothing used in time series analysis and forecasting. Here time series derived from the average of last kth elements of the series.

$$S_t = \frac{(X_{t-k} + X_{t-k+1} + X_{t-k+2} + \ldots\ldots + X_t)}{k}$$

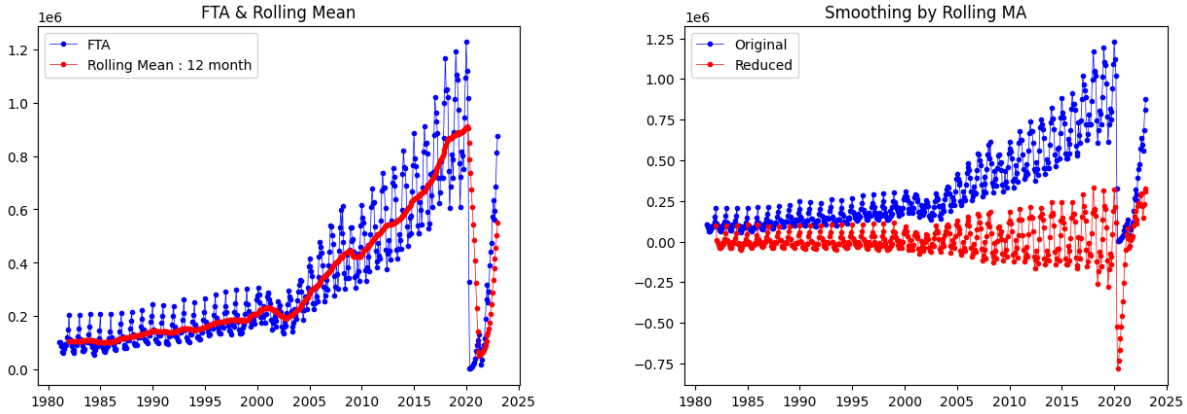


Figure 9: - Moving Average Smoothing

Here, the rolling values appear to be varying slightly. The p-value was 0.0001 is which is less than 0.05 so we reject the null hypothesis. Also, the test statistic is smaller than the 4% critical values so we can say with 96% confidence that this is a stationary series.

Exponential smoothing: Exponential smoothing is a weighted moving average technique. In the moving average smoothing the past observations are weighted equally. In this case, smoothing is done by assigning exponentially decreasing weights to the past observations.

$$S_0 = X_0$$
$$S_t = \alpha * X_t + (1 - \alpha) * S_{t-1}$$
$$\{t > 0, 0 < \alpha < 1\}$$

In the above equation, we can see that (1−α) is multiplied by the previously expected value St-1 which is derived using the same formula, which makes the expression recursive, and if we were to write it all out on paper, we would quickly see that (1−α) is multiplied by itself again and again. And this is why this method is called exponential.

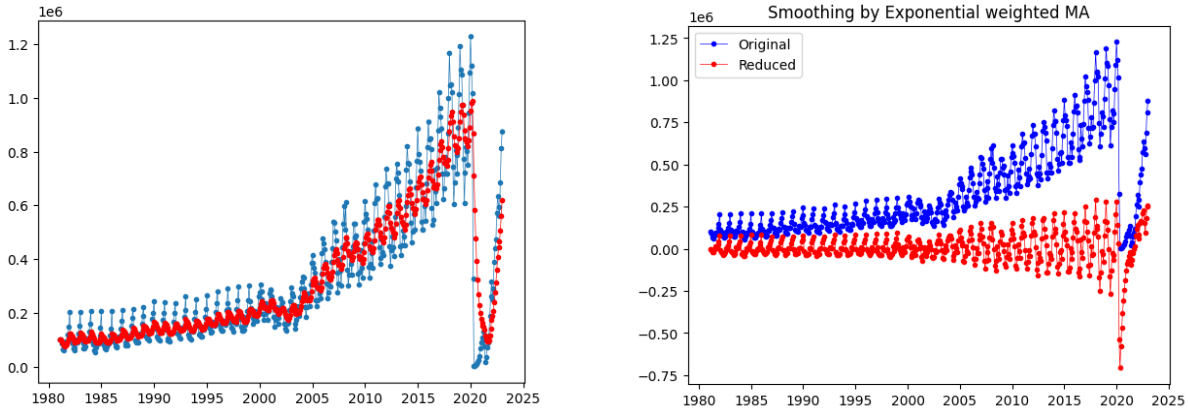


Figure 10: - Exponential Smoothing

Here, the p-value was 0.0000007 is which is less than 0.05 so we reject the null hypothesis. Also, the test statistic is smaller than the 5% critical values so we can say with 95% confidence

that this is a stationary series. So, we can now proceed to the time series forecasting using ARIMA model.

5.5 ARIMA model

An **Autoregressive Integrated Moving Average (ARIMA)** model is a statistical analysis model that uses time series data to either better understand the data or to predict future trends. It is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The goal is to predict future trending data by examining the differences between values in the series instead of through actual values. An ARIMA model can be understood by outlining each of its components as follows:

- **Autoregression (AR)**: It refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- **Integrated (I):** It represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- **Moving Average (MA)**: It incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Advantages:

1. It is good for short term forecasting
2. It only needs historical data
3. It also models non-stationary data

Disadvantages:

1. It is not used for long term forecasting
2. It is poor at predicting turning points
3. It is computationally expensive
4. The parameters are subjective

5.5.1 ARIMA Parameters

Each component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with **p, d, and q**, where the values substitute for the parameters to indicate the type of ARIMA model used. The parameters are defined as:

- **p**: the number of lag observations in the model, also known as the lag order.
- **d**: the number of times the raw observations are differenced
- **q**: the size of the moving average window, also known as the order of the moving average.

ARIMA models are denoted with the notation **ARIMA (p, d, q)**. These three parameters account for seasonality, trend, and noise in data. If a time series has seasonal patterns, then we need to add seasonal terms and it becomes **SARIMA (Seasonal ARIMA)**. SARIMA notation is quite a bit more complex than ARIMA, as each of the components receives a seasonal parameter on top of the regular parameter. They denoted with notation **SARIMA (p, d, q, m)**, where 'm' term is the amount of seasonality in data, (for example, m=12 for monthly data).

'd' term

- The first step to build an ARIMA model is to make the time series stationary. Because, term 'Auto Regressive' in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models work best when the predictors are not correlated and are independent of each other.
- The most common approach is to **difference** it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.
- The value of **d**, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

'p' and 'q' terms

- **'p'** is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors.
- And **'q'** is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

5.5.2 How to find the parameters of ARIMA model?

- For '**d'** term, the purpose of differencing it to make the time series stationary. But we need to be careful to not over-difference the series. Because, an over differenced series may still be non-stationary, which in turn will affect the model parameters.
- The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick. If the autocorrelations are positive for many numbers of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced.
- For **'p'** term, after differencing the time series, we need to ensure there is no autocorrelation in the differenced time series. We need to determine the AR and MA terms. We can find out the required number of AR terms by inspecting the Partial Autocorrelation (PACF) plot.
- For **'q'** term, the ACF tells how many MA terms are required to remove any autocorrelation in the stationary's series.

So, I used two famous graphs can helps to detect autocorrelation in the time series data: the **ACF (Autocorrelation function) plot** and the **PACF (Partial Autocorrelation function) plot**.

- **ACF plot**: This function is a tool that helps identify whether autocorrelation exists in our time series. On the x-axis, there are time steps (back in time) which is also called the number of lags. On the y-axis, there are the amount of correlation of every time step with 'present' time.
- **PACF plot**: The PACF is an alternative to the ACF. Rather than giving the autocorrelations, it gives us the partial autocorrelation. This autocorrelation is called partial, because with each step back in the past, only additional autocorrelation is listed. This is different from the ACF, as the ACF contains duplicate correlations when variability can be explained by multiple points in time.

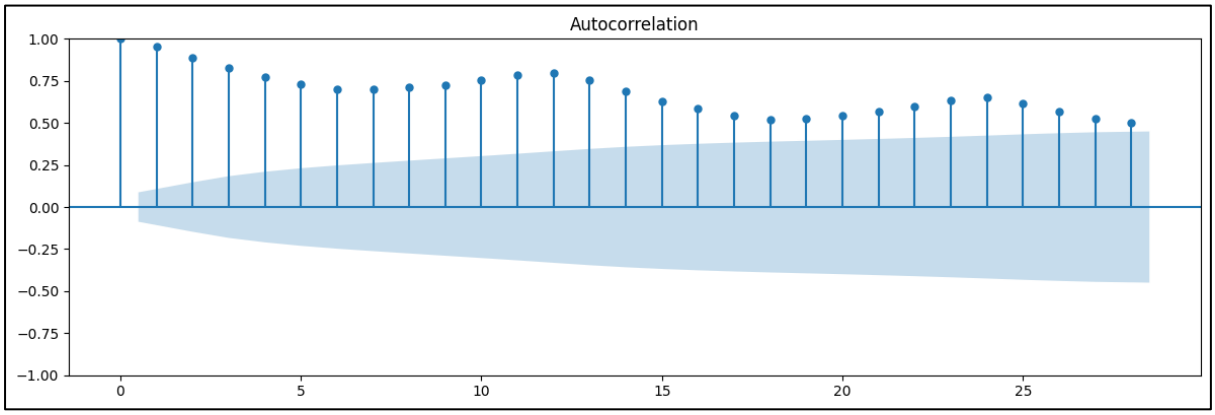


Figure 11: - ACF plot of the Original Time Series

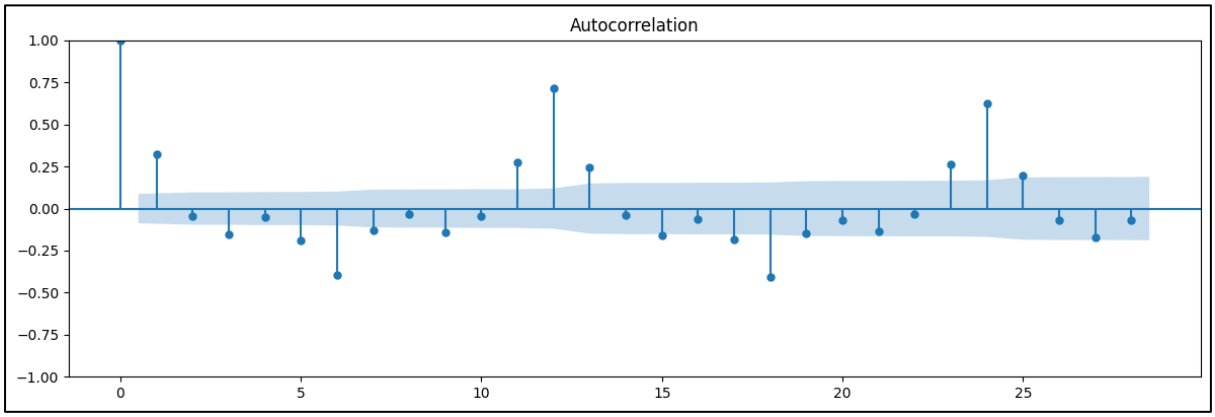


Figure 12: - ACF plot of the 1st Order Difference Time Series

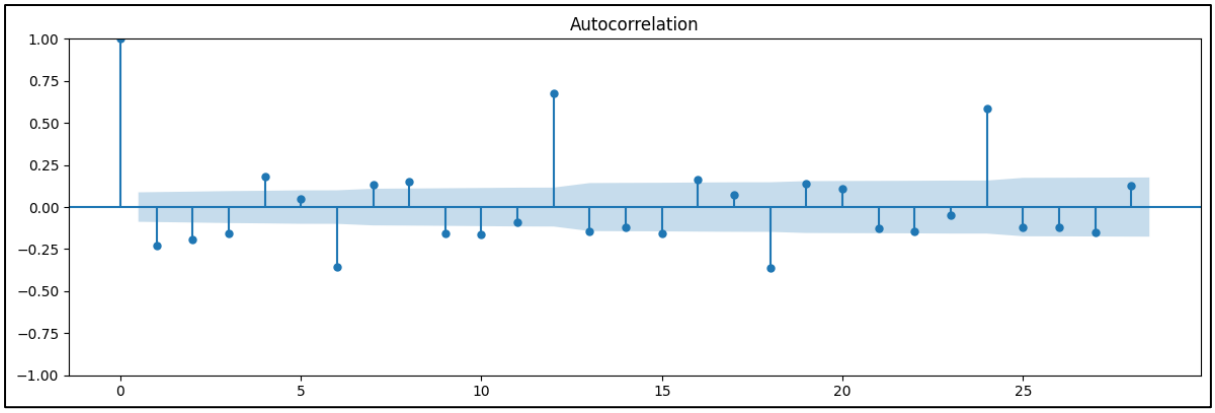


Figure 13: - ACF plot of the 2nd Order Difference Time Series

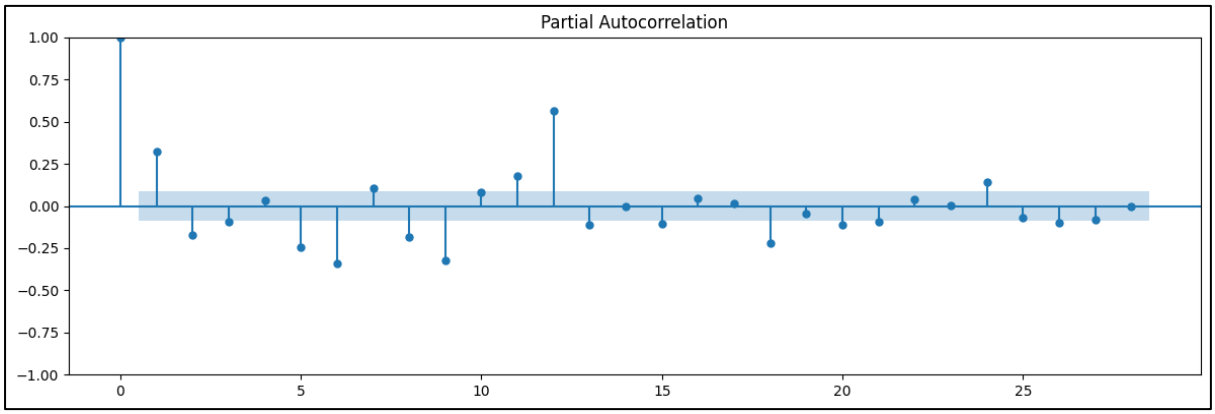


Figure 14: - PACF plot of the 1st Order Difference Time Series

- Here, we can see that in second-order differencing ACF plot, the immediate lag has gone on the negative side, representing that in the second-order the series has become over the difference. So, we can select value of d as 1.
- Here, we can see that in first-order differencing PACF plot, the first lag is significantly out of the limit and the second one is also out of the significant limit but it is not that far so we can select the order of the p as 1.
- Here, we can see that first-order differencing ACF plot, only one of the lags are out of the significance limit so we can say that the optimal value of our q is 1.

5.5.3 Experiment Setup and Model Diagnostic

Now that I have determined the values of p, d and q, we have everything needed to fit the ARIMA model. We implement ARIMA model from **statsmodels** package.

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.2472	0.038	6.545	0.000	0.173	0.321
ar.S.L12	0.1534	0.032	4.868	0.000	0.092	0.215
ma.S.L12	-0.7909	0.029	-27.670	0.000	-0.847	-0.735
sigma2	2.706e+09	2.05e-11	1.32e+20	0.000	2.71e+09	2.71e+09

Figure 15: - SARIMAX Summary Table

The summary attribute that results from the output of SARIMAX returns a significant amount of information. The **coef** column shows the weight (i.e., importance) of each feature and how each one impacts the time series. The **P>|z|** column informs us of the significance of each feature weight. Here, each weight has a p-value lower or close to 0.05, so it is reasonable to retain all of them in our model.

When fitting seasonal ARIMA models, it is important to run model diagnostics to ensure that none of the assumptions made by the model have been violated. So, I have quickly generated the model diagnostics and investigated for any unusual behaviour.

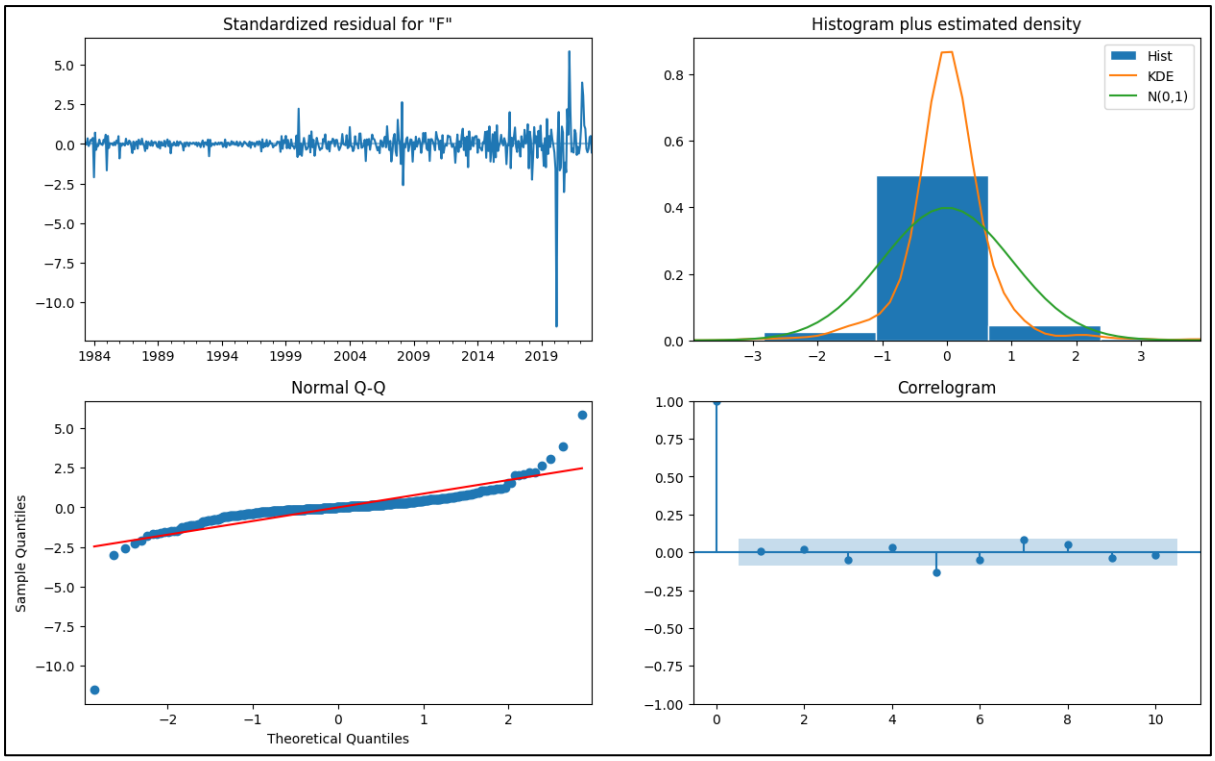


Figure 16: - Model Diagnostics

Our primary concern is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved. In this case, our model diagnostics suggests that the model residuals are normally distributed. Plus, more observations as follow: -

- In the top right plot, we see that the **red KDE line** really does not follow closely with the N(0,1) line (where N(0,1)) is the standard notation for a **normal distribution** with mean 0 and standard deviation of 1). Hence the residual plots are not perfectly normally distributed.
- The **QQ-plot** on the bottom left shows that the moderately ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1). Again, this is not really a strong indication that the residuals are normally distributed.

Those observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values.

6. RESULTS

This section outlines all the results and findings of the research methodology I have used in my project.

I have obtained a model for our time series that can now be used to produce forecasts. I start by comparing predicted values to real values of the time series, which will help us understand the accuracy of our forecasts.

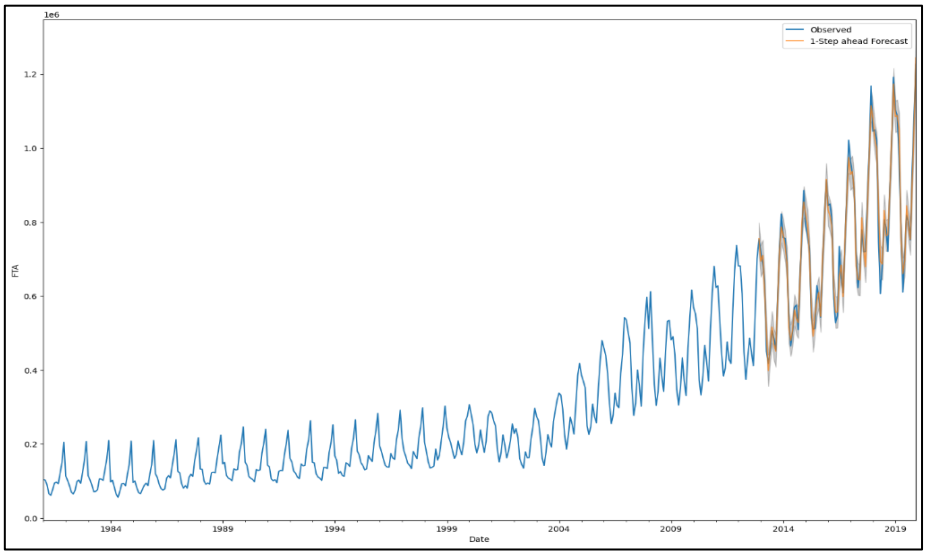


Figure 17: - Time Series Prediction of FTA before COVID-19

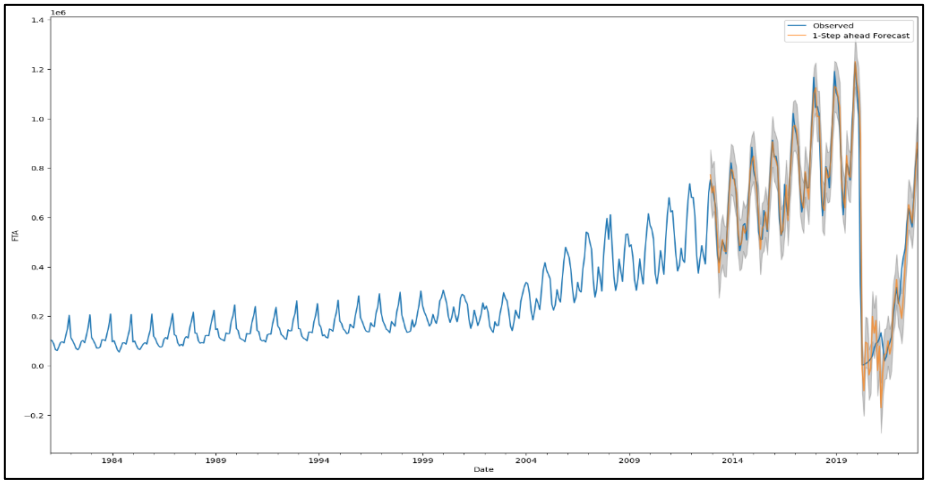


Figure 18: - Time Series Prediction of FTA after COVID-19

Overall, our forecasts slightly differ with the true values, but shows an upward trend starts from the beginning of the year and captured the seasonality toward the end of the year. Till the year 2010, we tried to find the forecast using **root mean square error** so that it can tell us how our model was able to forecast within the real set. Here is the table below to show the root mean square error of all the variables.

Variables	Root Mean Squared Error	
	Before COVID-19	After COVID-19
FTA	29447.28	79882.34
North America	1239.65	22660.64
C&S America	352.45	1306.45
Western Europe	3915.82	30201.40
Eastern Europe	2765.18	7366.54
Africa	592.84	4475.28
West Asia	1386.44	5612.68
South Asia	12131.47	38267.01
South East Asia	1668.77	11985.95
East Asia	1880.77	10319.26
Australasia	463.84	5496.37
Male passenger	11270.17	77428.50
Female passenger	6435.40	54650.07
Air Transport	14618.71	106809.37
Land Transport	11597.70	27546.85
Sea Transport	703.10	1432.71
0-14 age passenger	2867.22	12137.38
15-24 age passenger	2598.84	10722.28
25-34 age passenger	5001.82	24406.50
35-44 age passenger	3370.17	27594.35
45-54 age passenger	4490.73	25897.22
55-64 age passenger	2427.46	19097.19
65 and above age passenger	4059.80	12761.46

Now finally, here is the figure below to show and examine the forecast of the next 5 years.

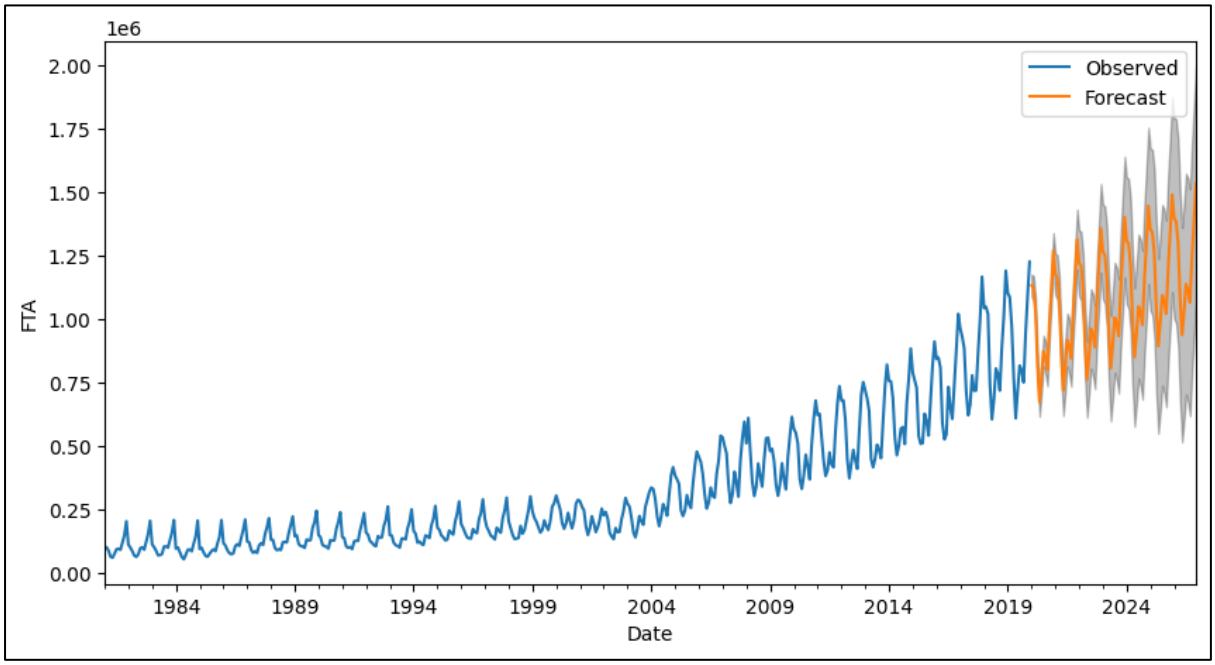


Figure 18: - Time Series Forecasting of FTA before COVID-19

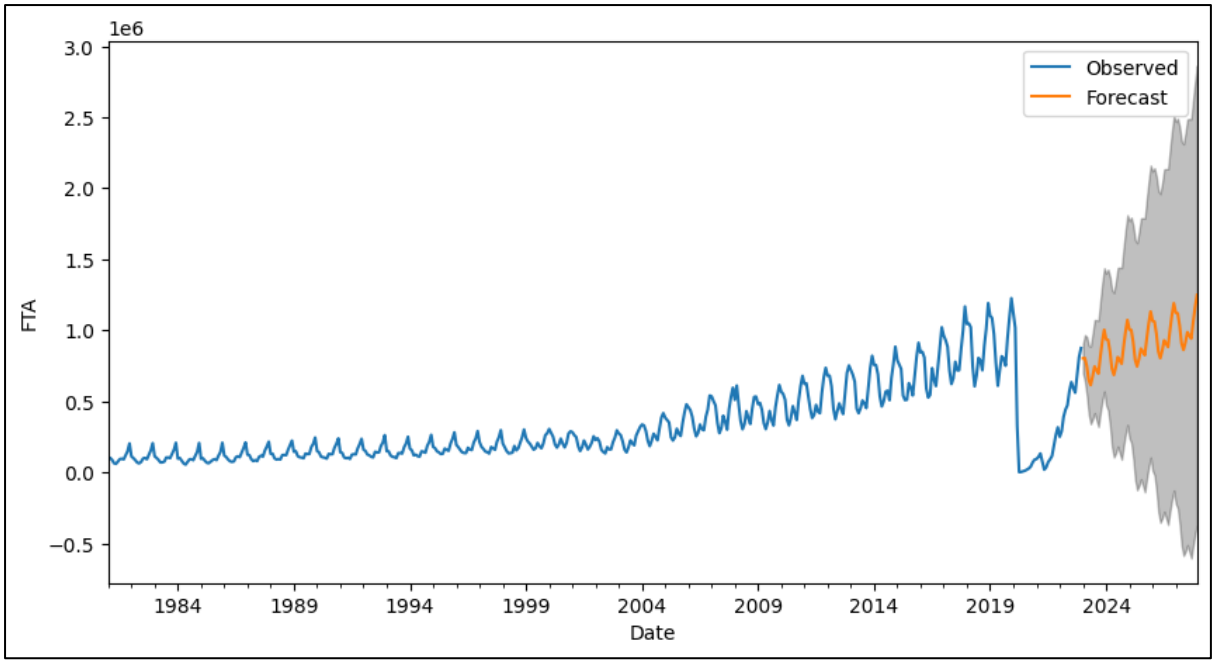


Figure 19: - Time Series Forecasting of FTA after COVID-19

Both the forecasts and associated confidence interval that I have generated for the next five years can now be used to further understand the time series and foresee what to expect. Our forecasts show that the time series is expected to increase linearly. As we forecast further out into the future, it is natural for us to become less confident in our values, especially the second plot above. This is reflected by the confidence intervals generated by our model, which grow larger as we move further out into the future.

7. CONCLUSION

After checking the accuracy score and forecasting two different time series data, we could see how much time series forecasting was necessary for this project. It also makes sense as why many industries rely on time series forecasting for many reasons. Despite the potential of time series forecasting to transform business models and improve bottom lines, many companies have yet to adopt its technologies and reap the benefits.

This project also helps us to understand how COVID-19 had really impacted the tourism sector in India. It has also enabled the tourism industry to face the huge thread as well as the slowdown of economy and tourist arrivals are seen.

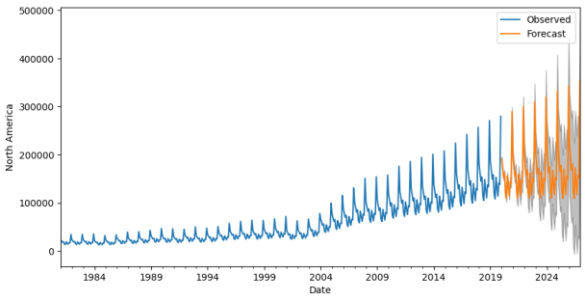
For future improvements of our projects, here are a few other things that we should try:

- We could change the start date of the prediction to see how this affects the overall quality of our forecasts.
- We could try more combinations of parameters to see if we can improve the goodness-of-fit of the model.
- We could select a different metric to select the best model. We used the ACF and PACF plot to find the best model, but we could seek to optimize the out-of-sample mean square error instead. We also could have used the AIC measure to find the best model.

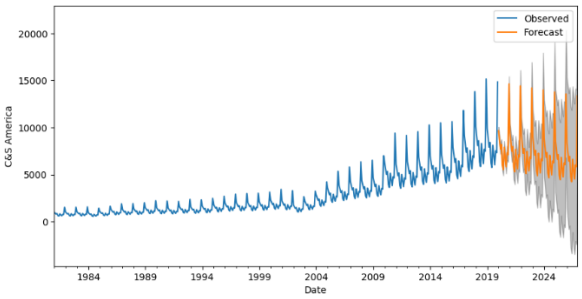
8. APPENDIX

8.1 Forecasting the different foreign tourist arrivals before COVID-19

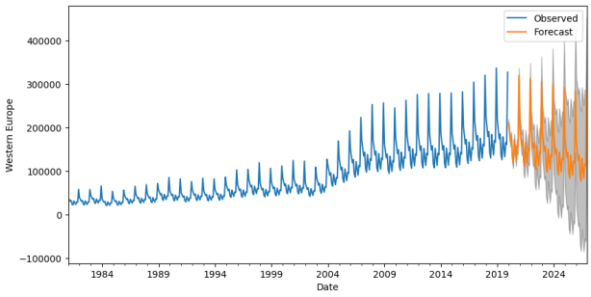
Here are some of the time series forecast I have implemented for all the variables of the time series data before the year of 2020, i.e., before COVID-19.



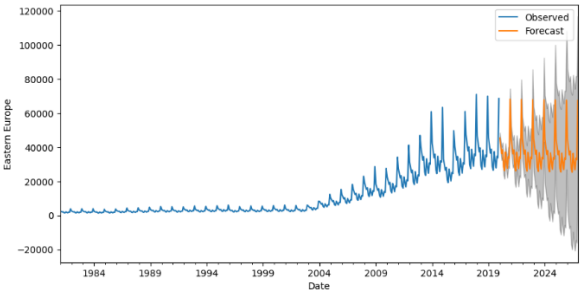
(1) North America



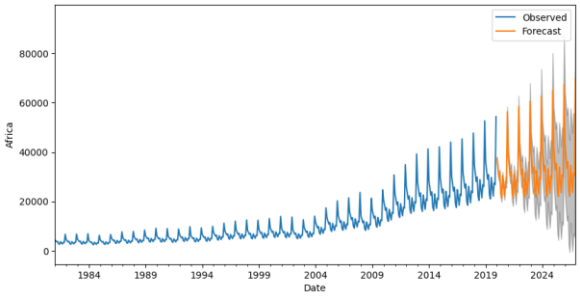
(2) C&S America



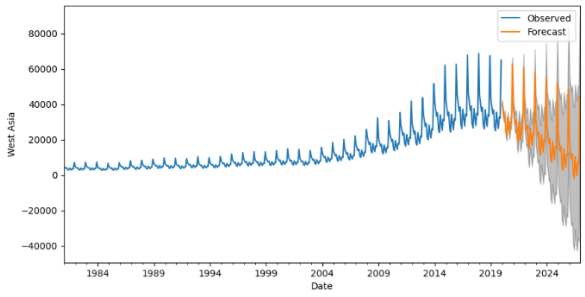
(3) Western Europe



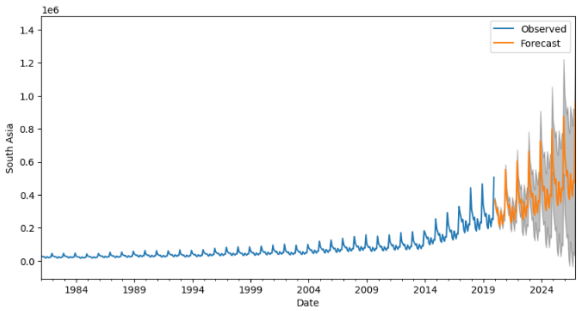
(4) Eastern Europe



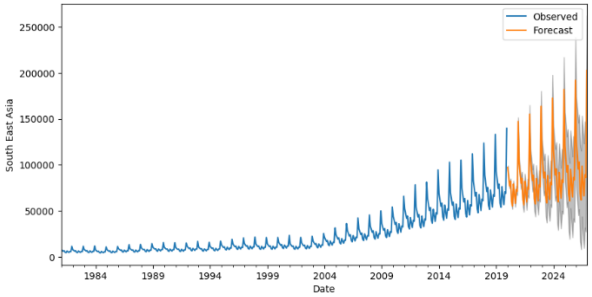
(5) Africa



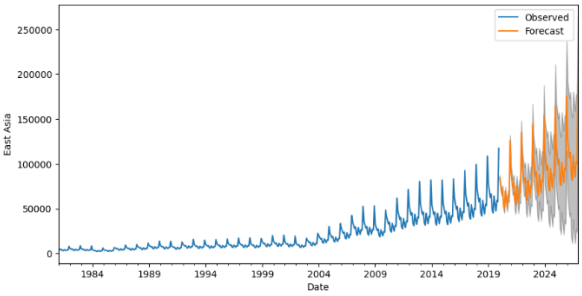
(6) West Asia



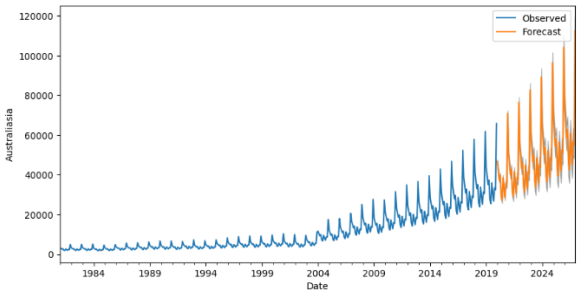
(7) South Asia



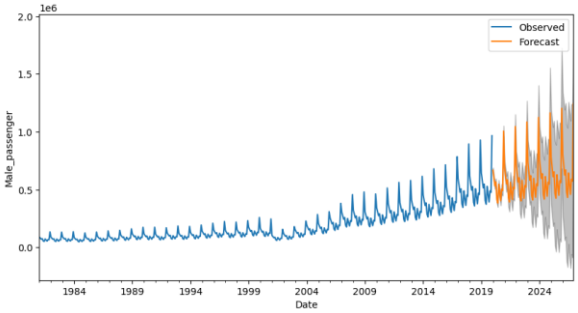
(8) South East Asia



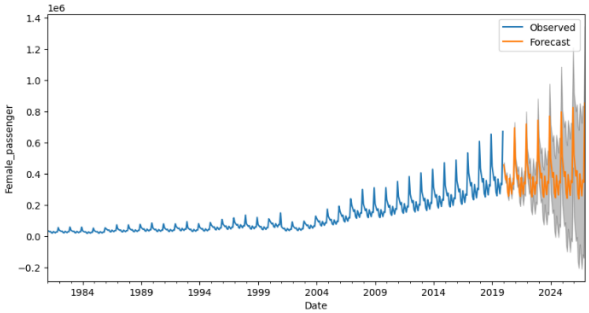
(9) East Asia



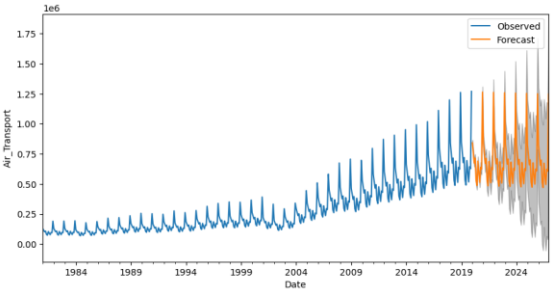
(10) Australasia



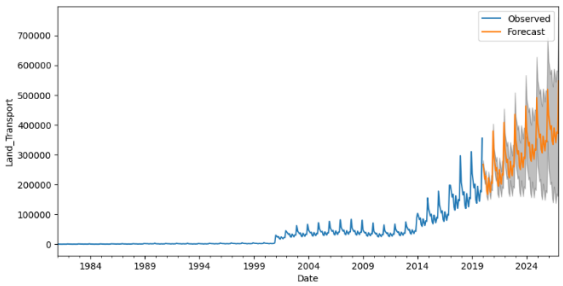
(11) Male passenger



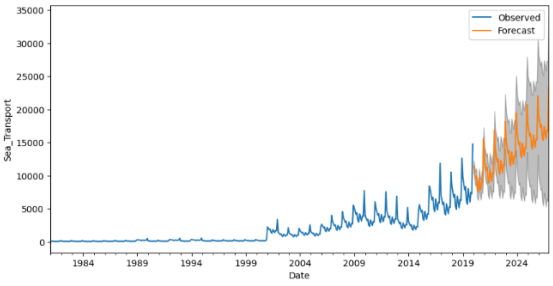
(12) Female passenger



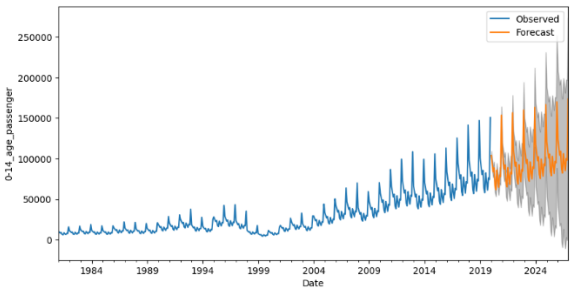
(13) Air Transport



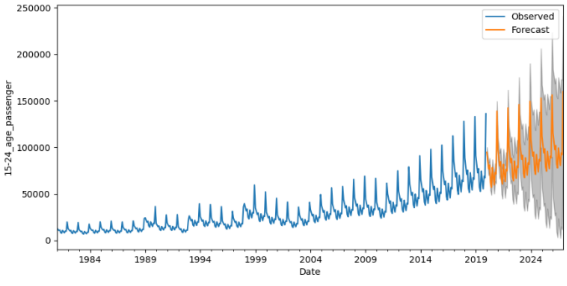
(14) Land Transport



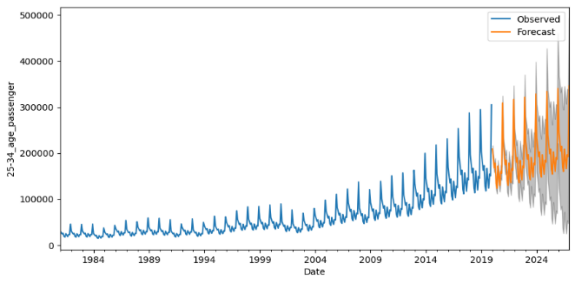
(15) Sea Transport



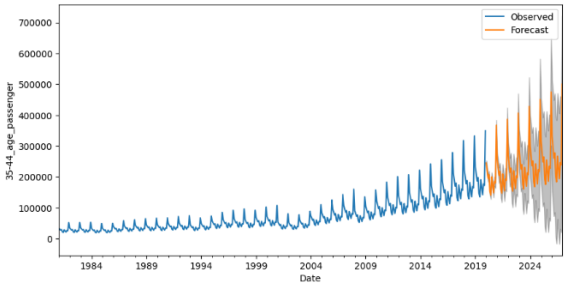
(16) 0-14 age passenger



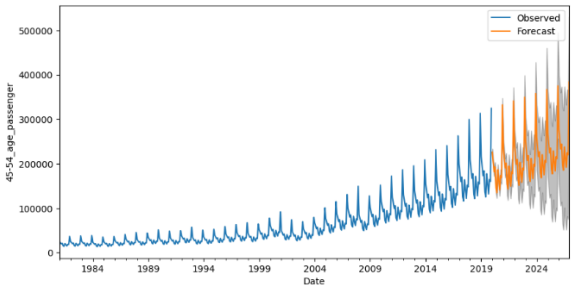
(17) 15-24 age passenger



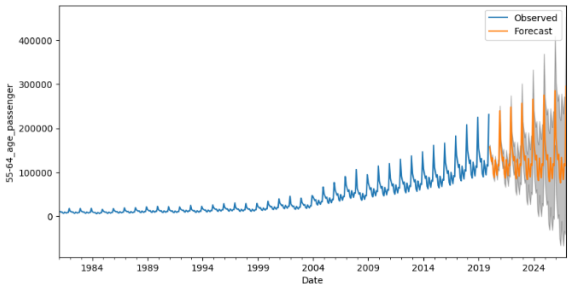
(18) 25-34 age passenger



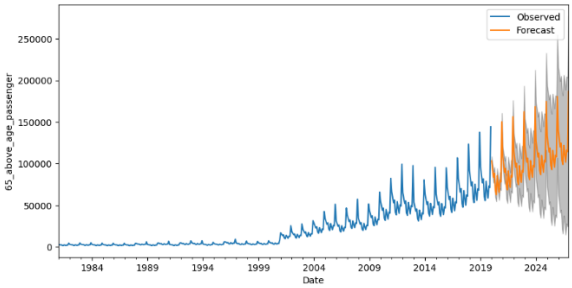
(19) 35-44 age passenger



(20) 45-54 age passenger



(21) 55-64 age passenger

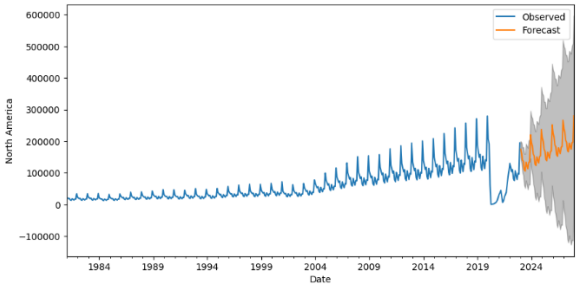


(22) 65 and above age passenger

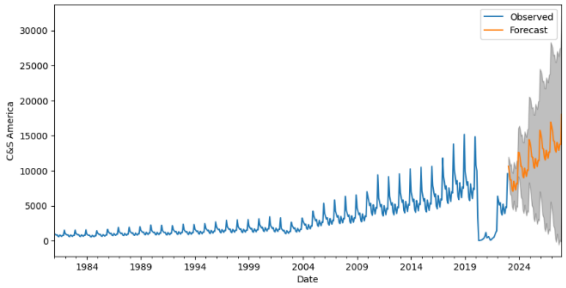
Figure 20: - Time Series Forecasting before COVID-19

8.2 Forecasting the different foreign tourist arrivals after COVID-19

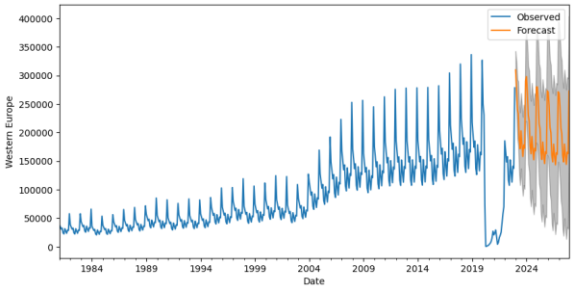
Here are some of the time series forecast I have implemented for all the variables of the time series data, i.e., after COVID-19.



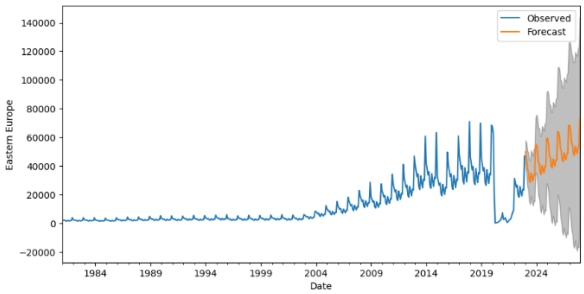
(1) North America



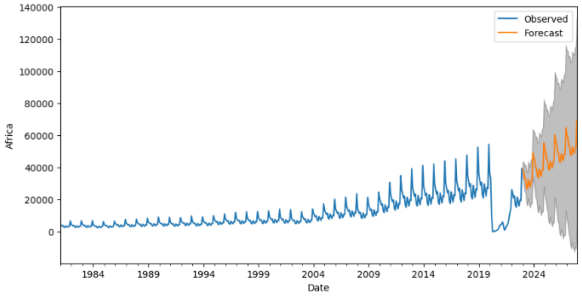
(2) C&S America



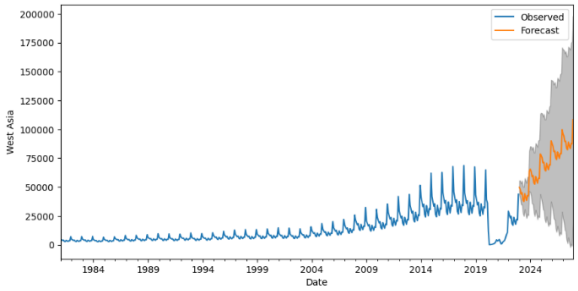
(3) Western Europe



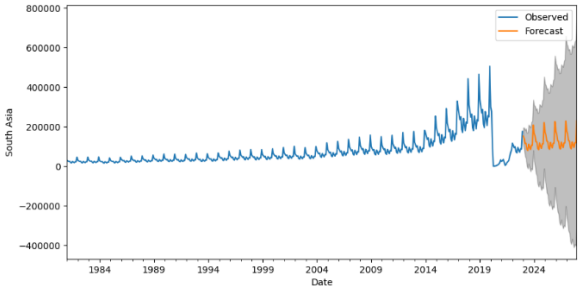
(4) Eastern Europe



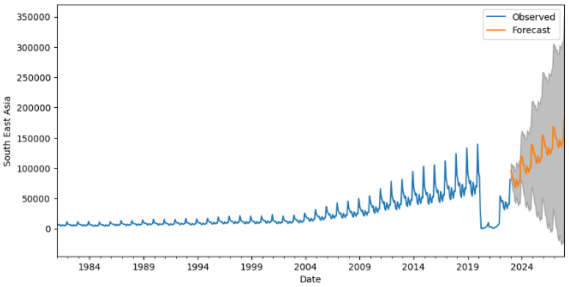
(5) Africa



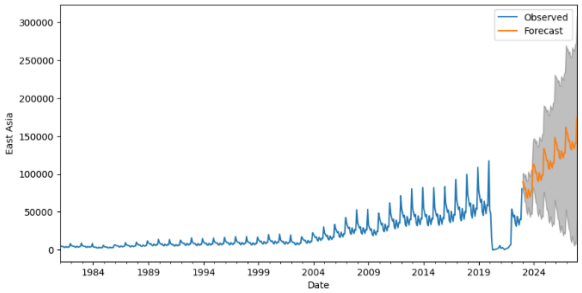
(6) West Asia



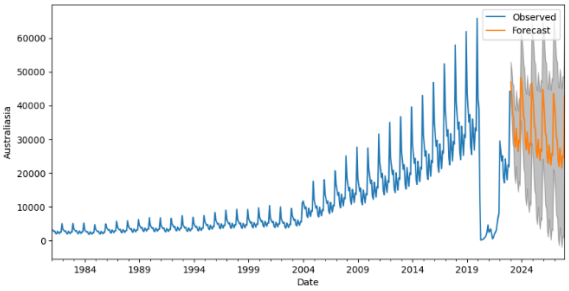
(7) South Asia



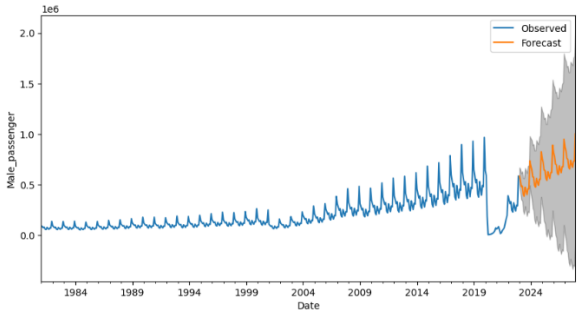
(8) South East Asia



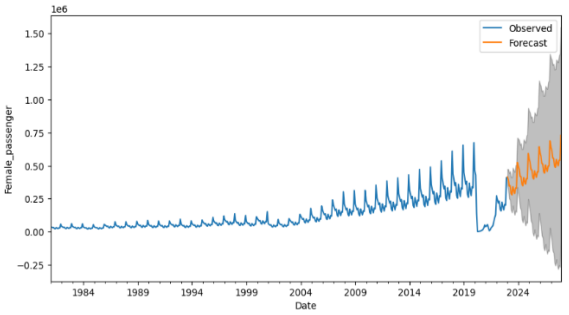
(9) East Asia



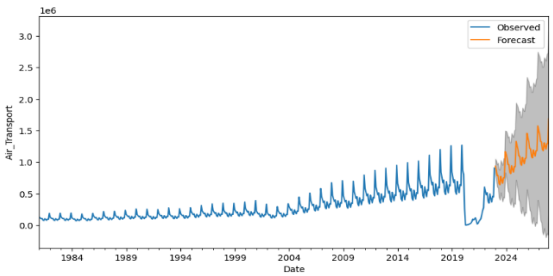
(10) Australasia



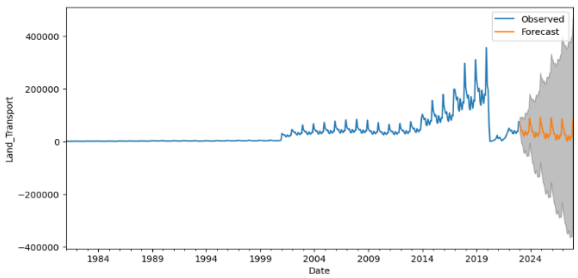
(11) Male passenger



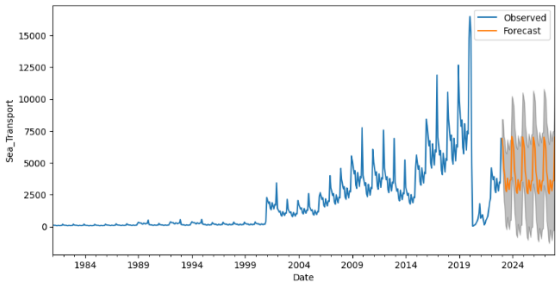
(12) Female passenger



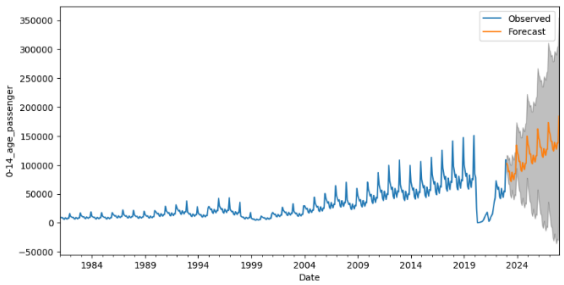
(13) Air Transport



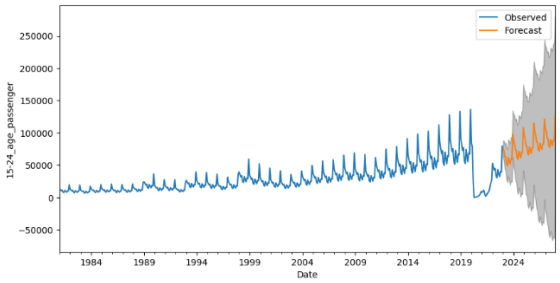
(14) Land Transport



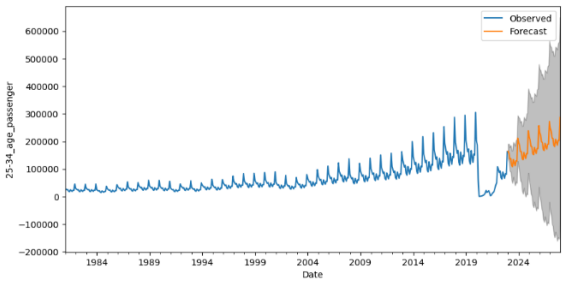
(15) Sea Transport



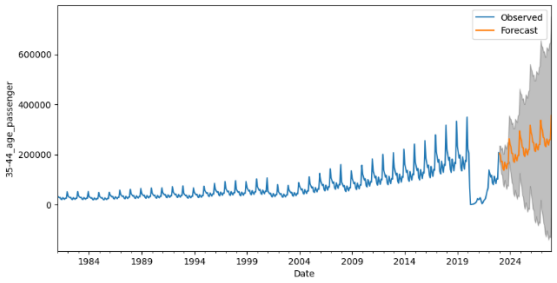
(16) 0-14 age passenger



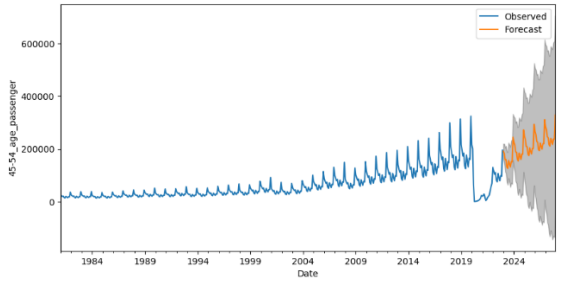
(17) 15-24 age passenger



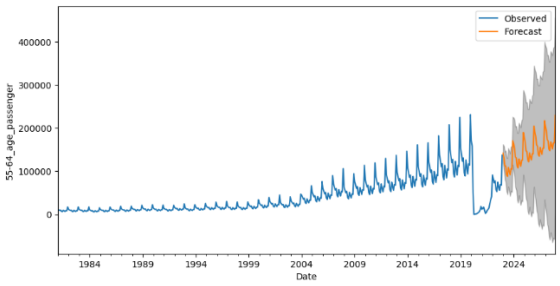
(18) 25-34 age passenger



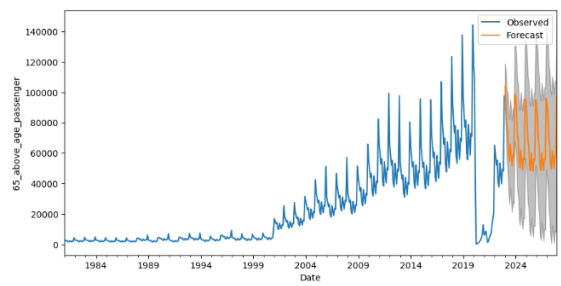
(19) 35-44 age passenger



(20) 45-54 age passenger



(21) 55-64 age passenger



(22) 65 and above age passenger

Figure 21: - Time Series Forecasting after COVID-19

9. REFERENCE

[1] Leiper N 1995 Tourism Management (Collingwood: TAFE Publications)

[2] Yoeti O A 1993 Pengantar Ilmu Pariwisata (Bandung: Angkasa)

[3] Sarjono H and Abbas B S 2017 Forecasting: Aplikasi Penelitian Bisnis QM for Windows vs Minitab vs Manual (Jakarta: Mitra Wacana Media)

[4] Yoeti O A 2003 Tours and Travel Marketing (Jakarta: Pradnya Paramita)

[5] Chen C F, Lai M C and Yeh C C 2012 Forecasting tourism demand based on empirical mode decomposition and neural network Knowledge-Based Systems 26 pp 281–7 DOI: 10.1016/j.knosys.2011.09.002

[6] Chen R, Liang C, Hong W C and Gu D X 2015 Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm App. Soft Computing 26 pp 435–43 DOI: 10.1016/j.asoc.2014.10.022

[7] Balli H O, Tsui W H and Balli F 2018 Modelling the volatility of international visitor arrivals to New Zealand J. of Air Transport Management 75 pp 204–14 DOI: 10.1016/j.jairtraman.2018.10.002

[8] Song, H., Qiu, R.T.R. and Park, J. (2019), "*A review of research on tourism demand forecasting*", Annals of Tourism Research, Vol. 75, pp. 338-362, doi: 10.1016/j.annals.2018.12.001.

[9] Li, X., Law, R., Xie, G. and Wang, S. (2021), "*Review of tourism forecasting research with internet data*", Tourism Management, Vol. 83, p. 104245, doi: 10.1016/j.tourman.2020.104245.

[10] Gunter, U. and Önder, I. (2016), "*Forecasting city arrivals with Google Analytics*", Annals of Tourism Research, Vol. 61, pp. 199-212, doi: 10.1016/j.annals.2016.10.007.

[11] Park, S., Lee, J. and Song, W. (2017), "*Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data*", Journal of Travel and Tourism Marketing, Vol. 34 No. 3, pp. 357-368, doi: 10.1080/10548408.2016.1170651.

[12] Zhou-grundy, Y. and Turner, L.W. (2015), "*The challenge of regional tourism demand forecasting: the case of China*", Journal of Travel Research, Vol. 53 No. 6, pp. 747-759, doi: 10.1177/0047287513516197.

[13] Dergiades, T., Mavragani, E. and Pan, B. (2018), "*'Google Trends and tourists' arrivals: emerging biases and proposed corrections*", Tourism Management, Vol. 66, pp. 108-120, doi: 10.1016/j.tourman.2017.10.014.

[14] Bangwayo-Skeete, P.F. and Skeete, R.W. (2015), "*Can Google data improve the forecasting performance of tourist arrivals?*", Mixed-data Sampling approach' Tourism Management, Vol. 46, pp. 454-464, doi: 10.1016/j.tourman.2014.07.014.

[15] Huang, X., Zhang, L. and Ding, Y. (2017), "*The Baidu Index: uses in predicting tourism flows – a case study of the Forbidden City*", Tourism Management, Vol. 58, pp. 301-306, doi: 10.1016/j.tourman.2016.03.015.

[16] Box, G.E.P. & G. M. Jenkins (1976), Time Series Analysis: Forecasting and Control, Revised Edition, San Francisco: Holden-Day.

[17] Stellwagen E. & L. Tashman (2013): ARIMA: The Models of Box and Jenkins, Summer 2013, 28-33.

[18] Goh, C., & R. Law (2002), Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention, Tourism Management, Vol. 23 (Issue 5), 499-510. https://doi.org/10.1016/S0261-5177(02)00009-2

[19] Preez, J., Witt, S.F. (2003), Univariate versus multivariate time series forecasting: an application to international tourism demand, International Journal of Forecasting, Vol. 19 (Issue 3), 435-451, https://doi.org/10.1016/S0169-2070(02)00057-2.

[20] Athanasopoulos G., Hyndman, R. J., Song, H., & Wu, D. C. (2011), The tourism forecasting competition, International Journal of Forecasting, Vol. 27 (Issue 3), 822-844, https://doi.org/10.1016/j.ijforecast.2010.04.009

[21] Kim, J.H., Wong, K., Athanasopoulos, G., & Liu, S. (2011), Beyond point forecasting: Evaluation of alternative prediction intervals for tourist arrivals, International Journal of Forecasting, Vol. 27 (Issue 3), 887-901, https://doi.org/10.1016/j.ijforecast.2010.02.014

[22] Claveria, O. & S. Torra (2014), Forecasting Tourism Demand to Catalonia: Neural Networks vs. Time Series Models, Economic Modelling, Vol. 36, 220-228. https://doi.org/10.1016/j.econmod.2013.09.024

[23] Ministry of Tourism, Market Research and Statistics, https://tourism.gov.in/market-research-and-statistics