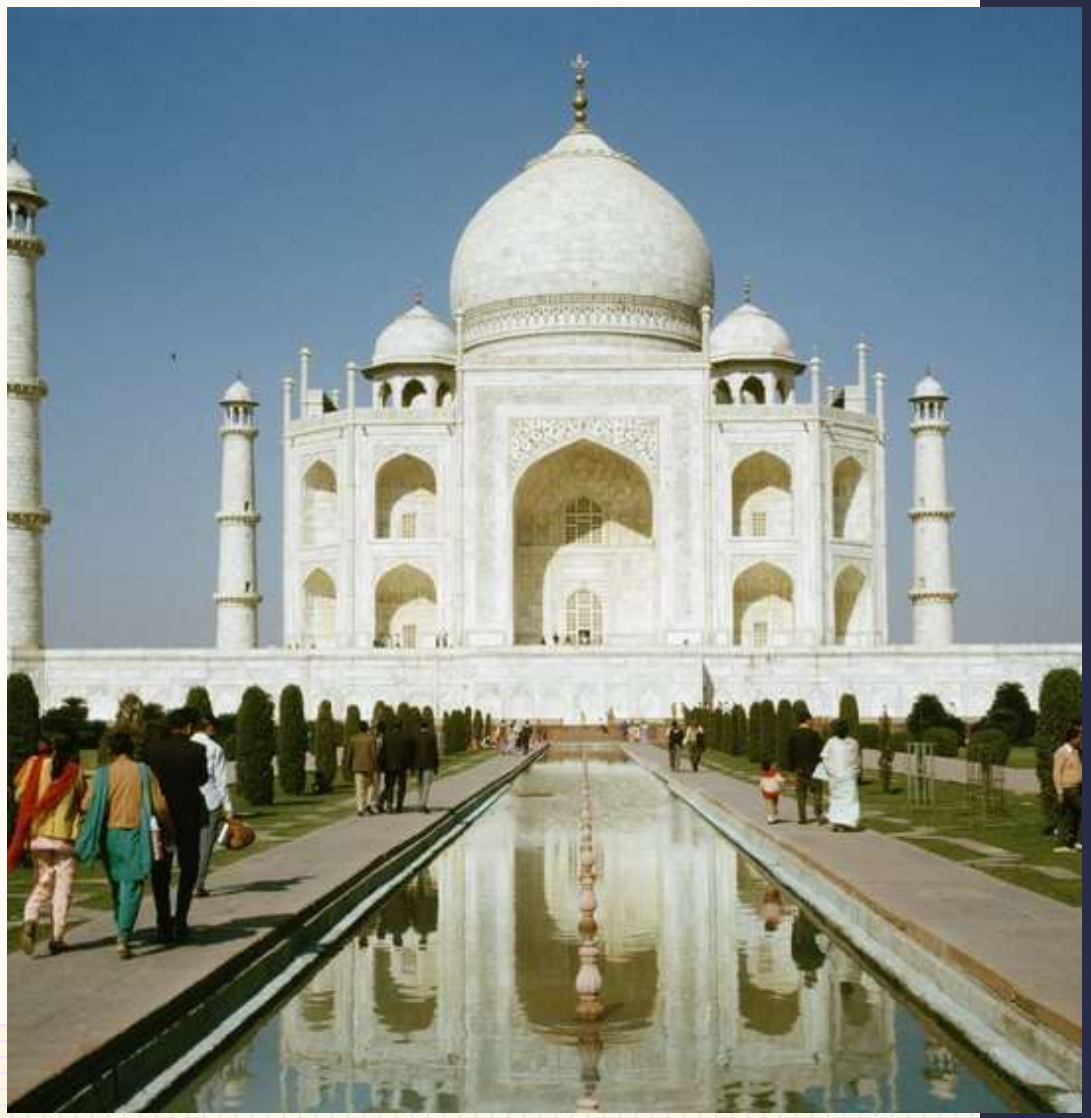
# TOURISM STATISTICS

Group members:
Snehal Nazirkar
Shradha Katkar
Anjali kadre



# INTRODUCTION

- Tourism is an activity of great socioeconomic importance, as it facilitates economic growth and development of the region with social and cultural intermixing.
- India, one of the oldest civilizations of the world with its socio-cultural uniqueness, vast physiographical theatre, highest levels of natural biodiversity, is the obvious and evident destination of tourists.
- Historical monuments, great cultural diversities and marvelous cuisines, etc. also attract peoples towards India from all around the world.
- Tourist influx in India is in abundance. Total tourists were 2.65 million in the year 2000 which increased to 10.93 million in the year 2019. The total increase was about 4.12 times with 7.742 per cent of compound annual growth rate (CAGR)(Annualizied average rate of revenue growth between two years, assuming growth takes palce at exponentially compounded rate) from the year 2000 to 2019.
- Presently, the country is cultivating the benefits of rising tourism industry with severe socioeconomic disparities, and environmental and infrastructural constraints.
- The project evaluates and investigates the trends of foreign tourist arrivals in India and analyses it's effect on Indian Economy. This project also highlights the effect of covid-19 on the tourism industry in India.

# OBJECTIVE

- To fit a suitable Time series model for monthly foreign tourist arrivals in India.
- Forecasting the future arrivals
- Recommendation of hotels based on your budget
- Emphasizing the effect of covid-19 on tourism and hence on Indian Economy.

# Methodology

The study is essentially empirical and has utilized the secondary source of Information. Secondary data is used to find out the historical and recent trends of foreign tourist arrivals, purpose for travelling to India, GDP and foreign expenditure. The data used for all the computational work has been taken from trusted government websites and reports.

Source of data:

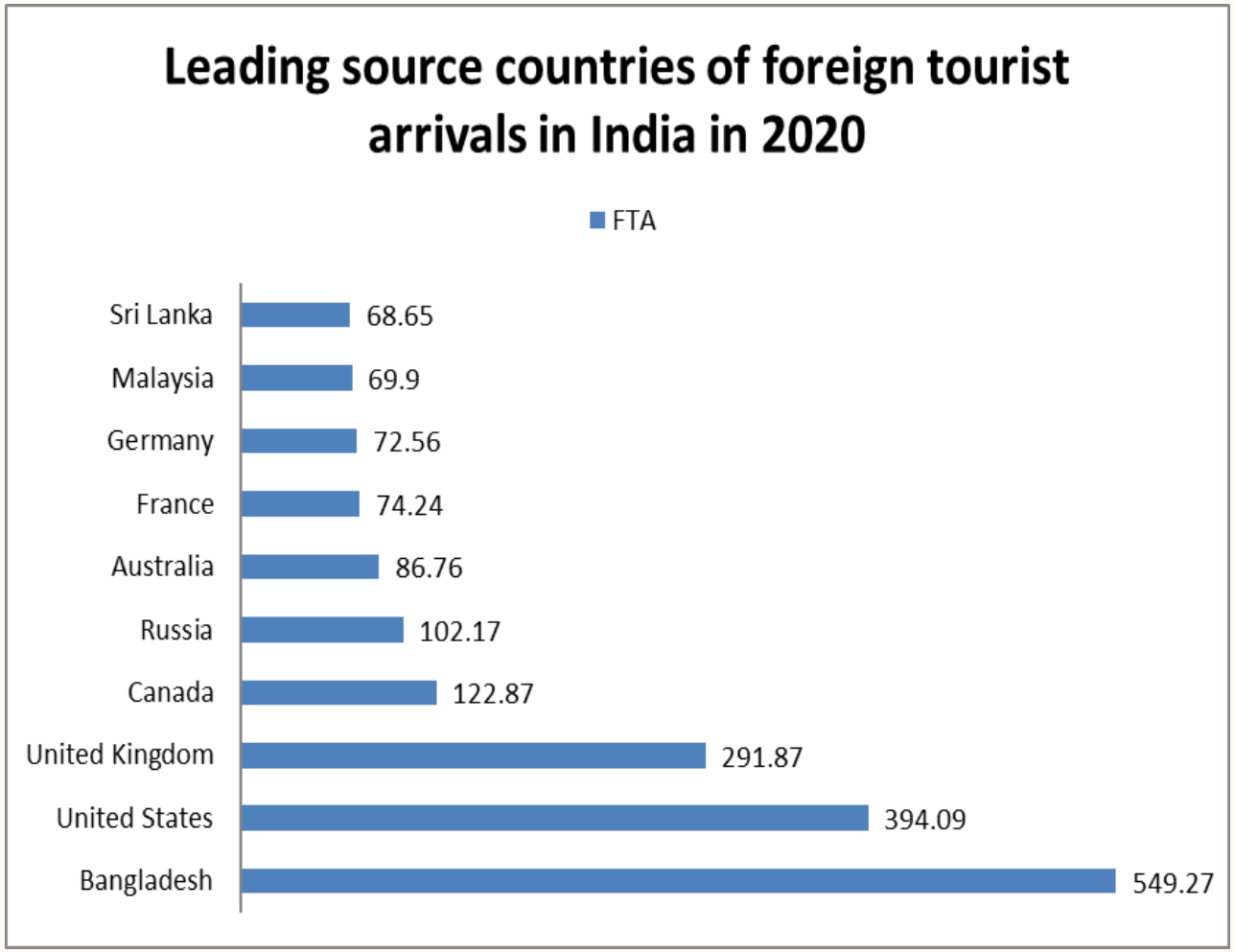
https://tourism.gov.in/
https://dpiit.gov.in/

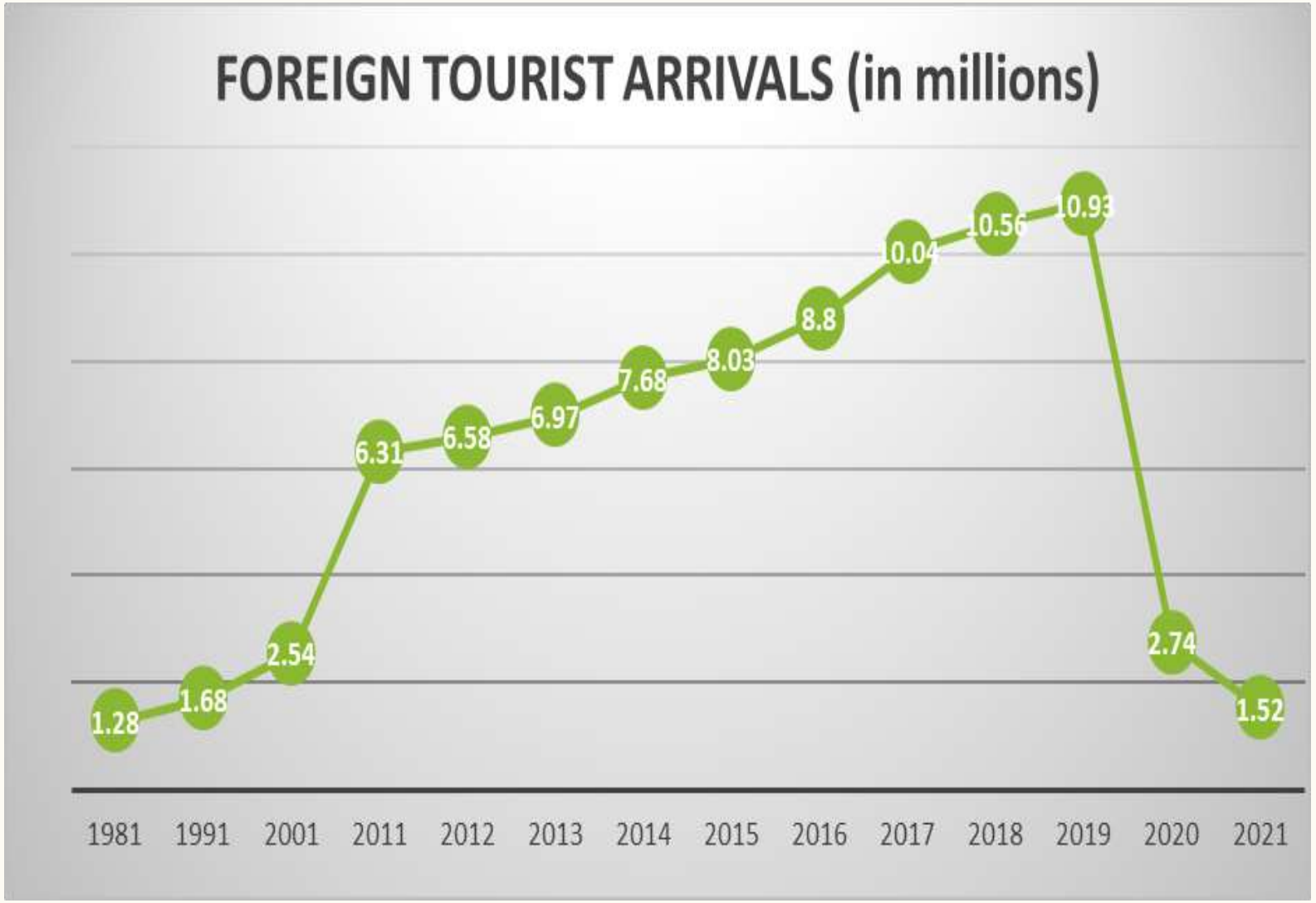
# EXPLORATORY DATA ANALYSIS

Bangladesh was seen to be the leading country with respect to the foreign tourist arrivals in India.

The reason is the cultural ties of the past leading to majority travelling for pilgrimage purpose.

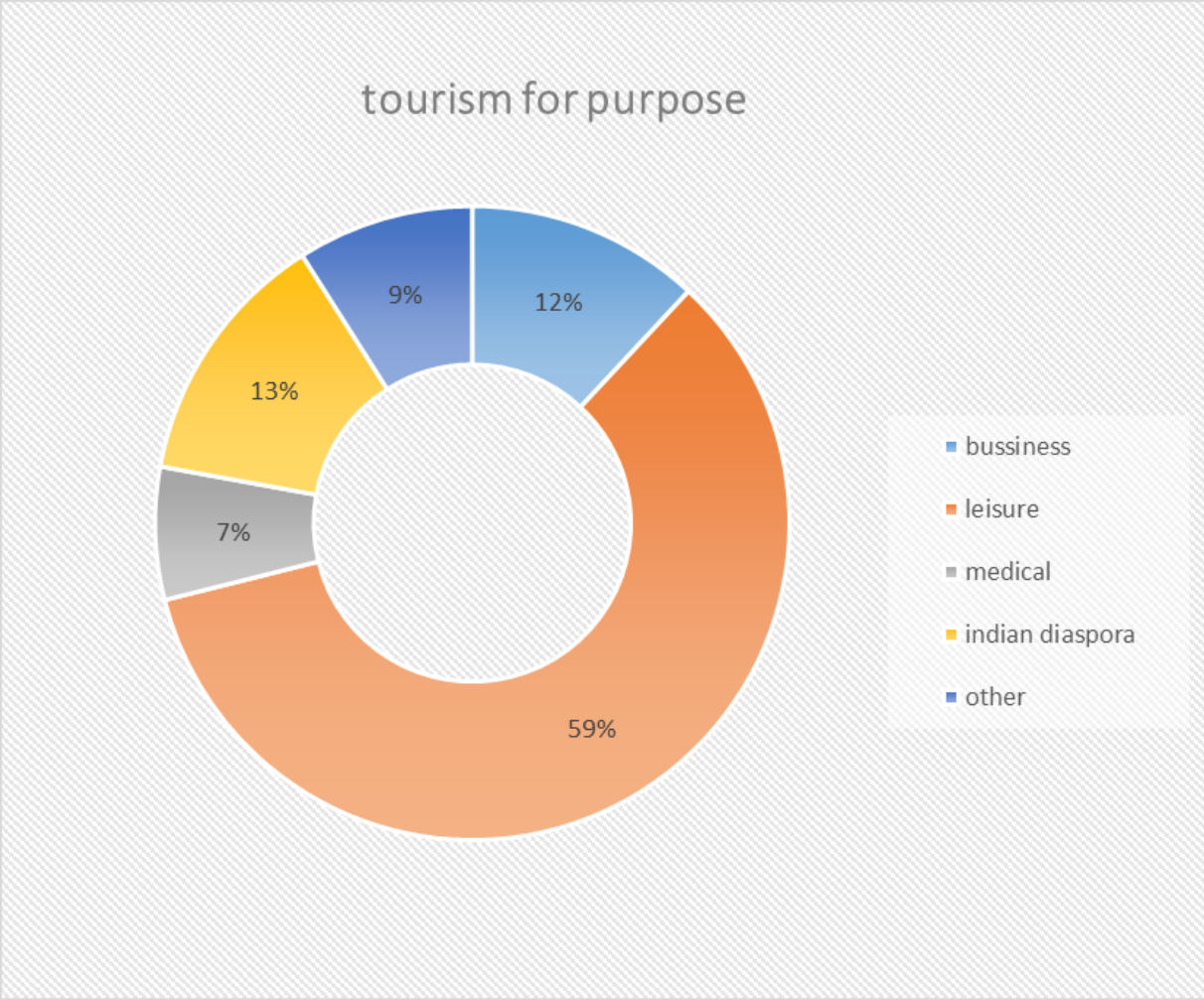
Also the various MoUs signed by the govt of the two countries for training of various diplomats and army personnels which leads to arrival.



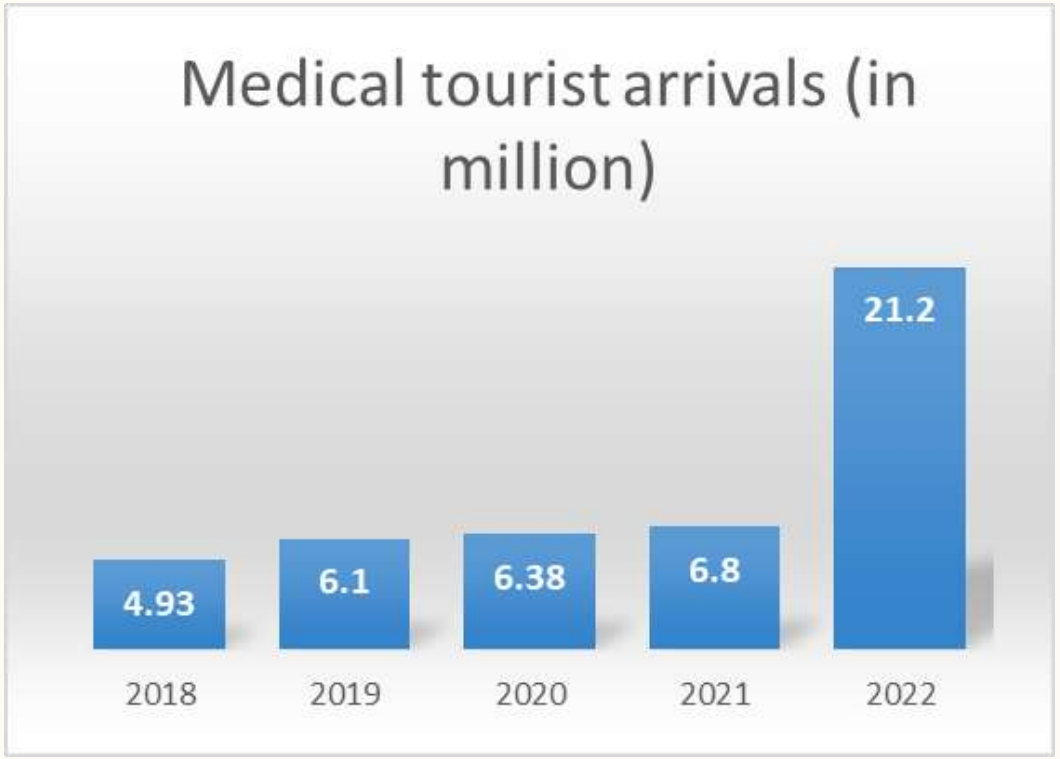


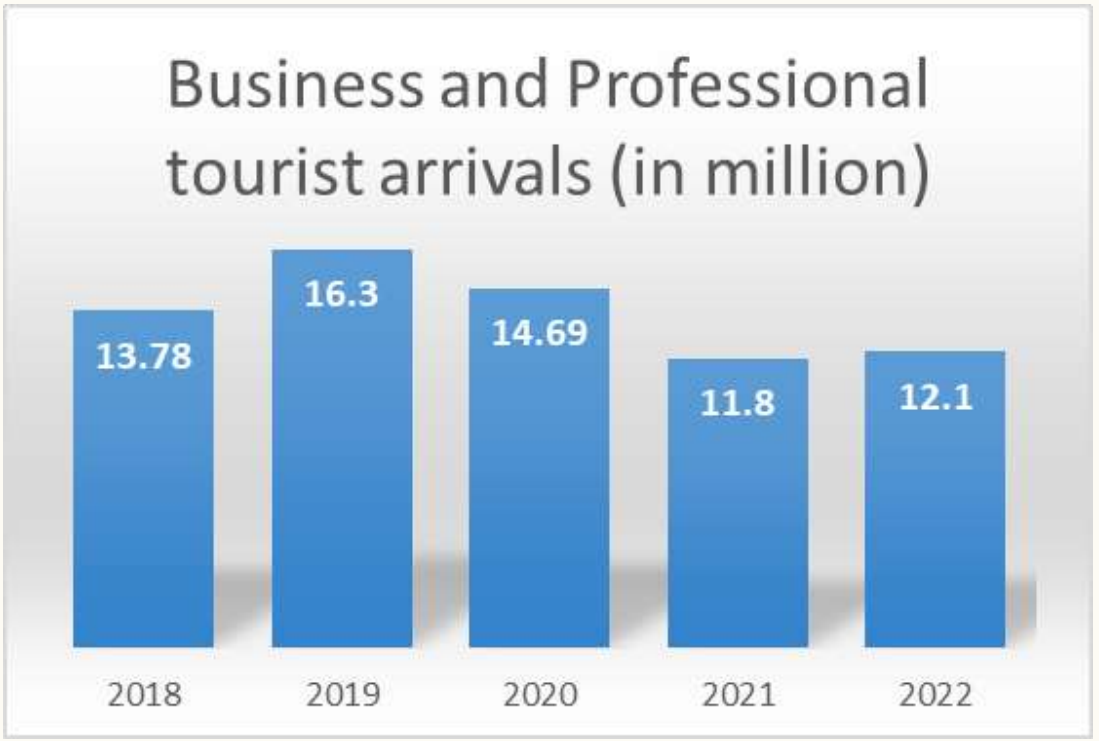

Foreign tourist arrival has seen an increasing trend over the last few decade. The reason found out are the many new policies initiated by the GoI like the "Incredible India","Atithi Devo Bhava" and the on-going "Swacha Bharat Mission" that has helped to change India's image in the world.

# PURPOSE OF TRAVEL




The major reason of foreign tourist arrival in India is found to be for leisure purpose which is obvious owing to rich bio-diversity, more that 30 wetlands, vast coastline having numerous beaches and pleasant climate..!!

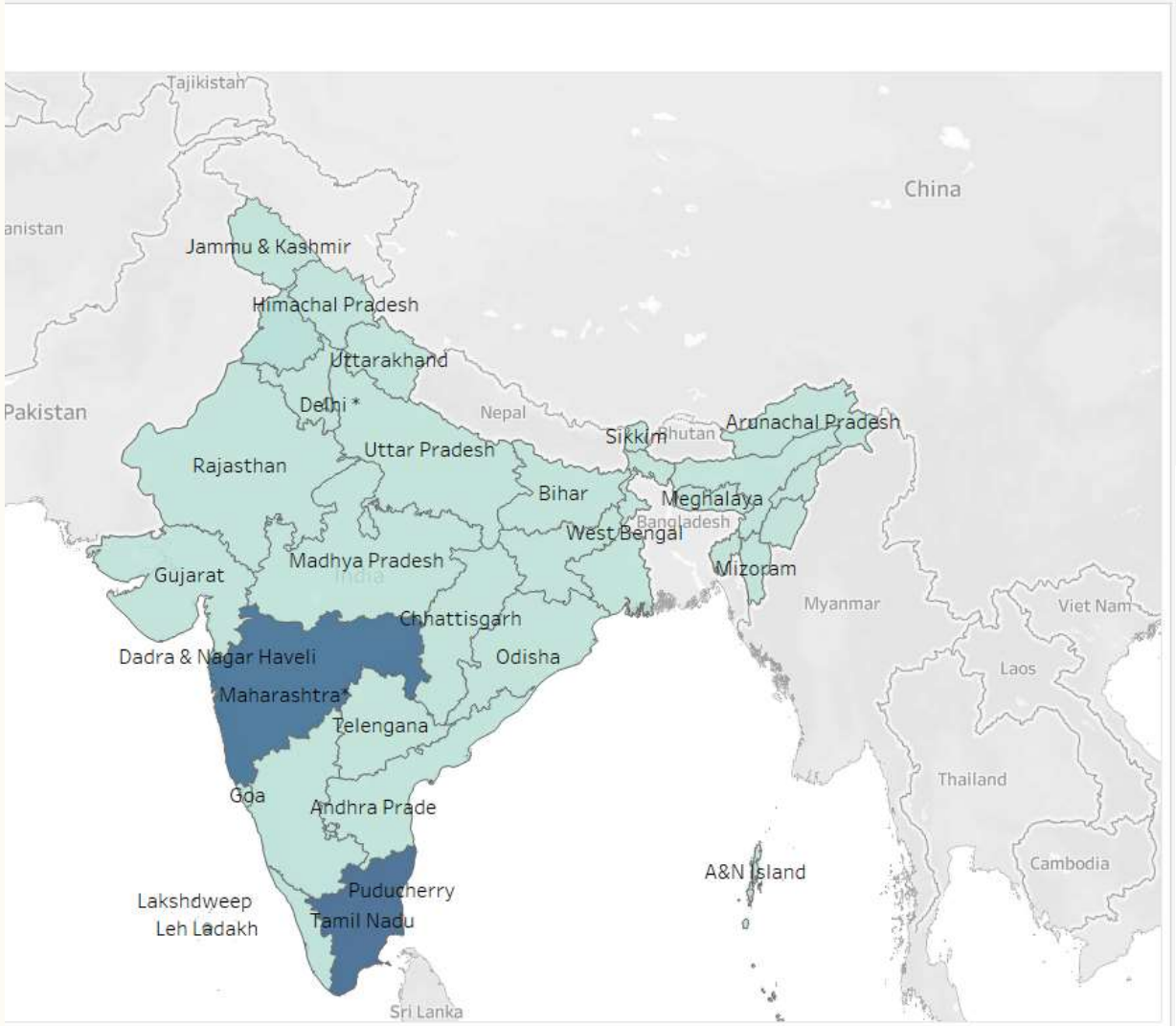


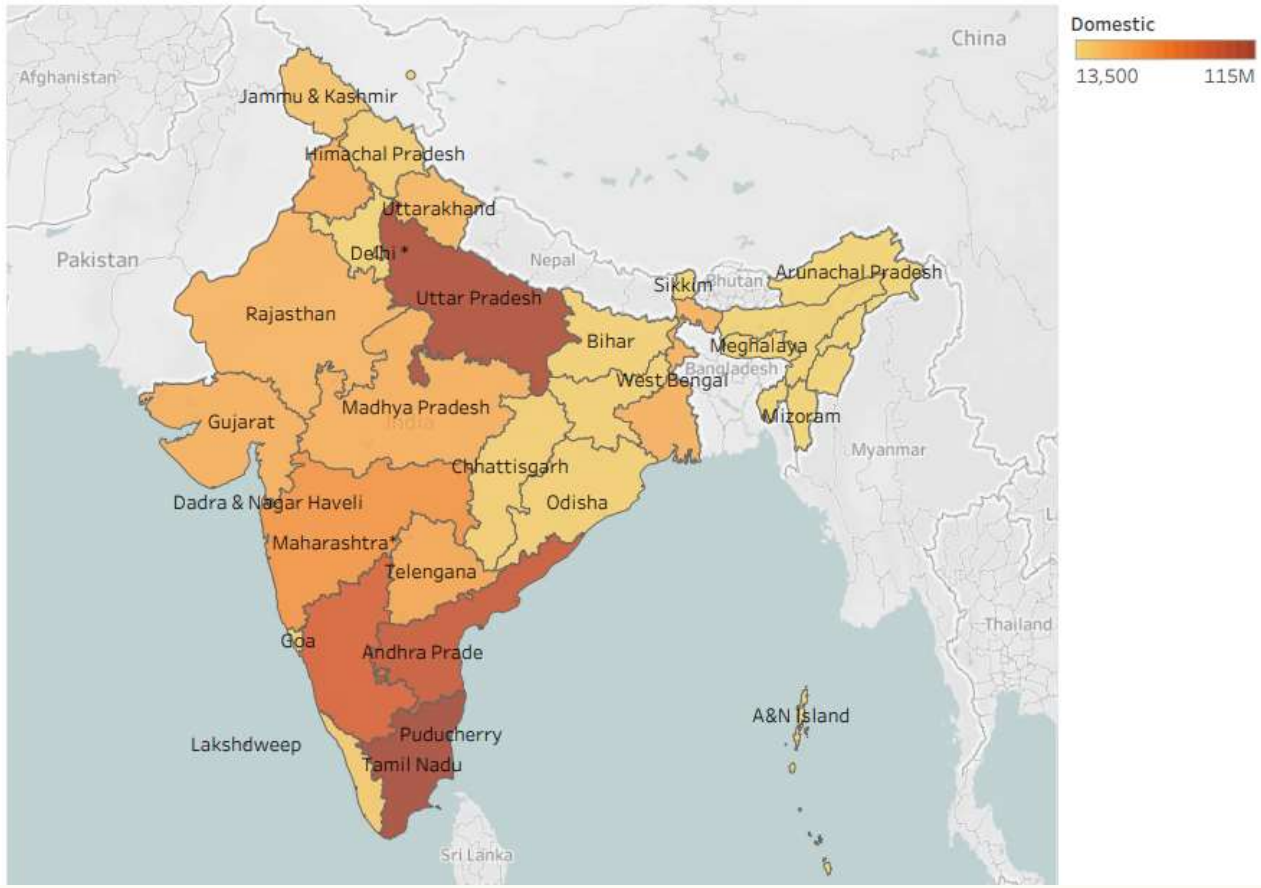




Due to the recent progress and India now having many state of the art infrastructure hospitals providing best medical facilities at affordable rate compared to the western countries the trend in medical tourism is seen increasing.

According to the World Competitive Index India now ranks 37$^{th}$ out of 63 nations in ease of doing business..!! Thus attracting many people for business purpose in India.

# Leading States visited by foreign tourists and domestic tourists



**With so many foreign tourist arrivals each year a logical question that follows is the accommodation facility in India.**

# Recommendation of Hotels based on your Budget

## Importing libraries

In [1]:
```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
```

## Importing Dataset

In [2]:
```
hotels = pd.read_csv('/content/Hotels.csv')
```

## Data Preprocessing

In [3]:
```
hotels.isna().sum()
```

Out[3]:
```
HID                 0
HOTEL               0
PRICE_RUPEES        0
NUMBER_OF_REVIEWS   0
Lat                 0
Lng                 0
dtype: int64
```

Interpretation:There are no missing values in our dataset

In [4]:
```
hotels.info()
```

```
RangeIndex: 254 entries, 0 to 253
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   HID                254 non-null    object 
 1   HOTEL              254 non-null    object 
 2   PRICE_RUPEES       254 non-null    float64
 3   NUMBER_OF_REVIEWS  254 non-null    int64  
 4   Lat                254 non-null    float64
 5   Lng                254 non-null    float64
dtypes: float64(3), int64(1), object(2)
memory usage: 12.0+ KB
```

In [5]:
```
hotels.describe()
```

Out[5]:

| | PRICE_RUPEES | NUMBER_OF_REVIEWS | Lat | Lng |
|---|---|---|---|---|
| **count** | 254.000000 | 254.000000 | 254.000000 | 254.000000 |
| **mean** | 2983.149803 | 540.582677 | 26.914314 | 75.800445 |
| **std** | 7420.243135 | 863.444151 | 0.043156 | 0.042395 |
| **min** | 150.000000 | 1.000000 | 26.767485 | 75.590469 |
| **25%** | 1123.650000 | 88.500000 | 26.902572 | 75.789400 |
| **50%** | 1797.840000 | 212.000000 | 26.919133 | 75.796085 |
| **75%** | 2846.580000 | 549.250000 | 26.928681 | 75.810282 |
| **max** | 112365.000000 | 6189.000000 | 27.066656 | 76.146076 |

# Recommendation Based on your Budget

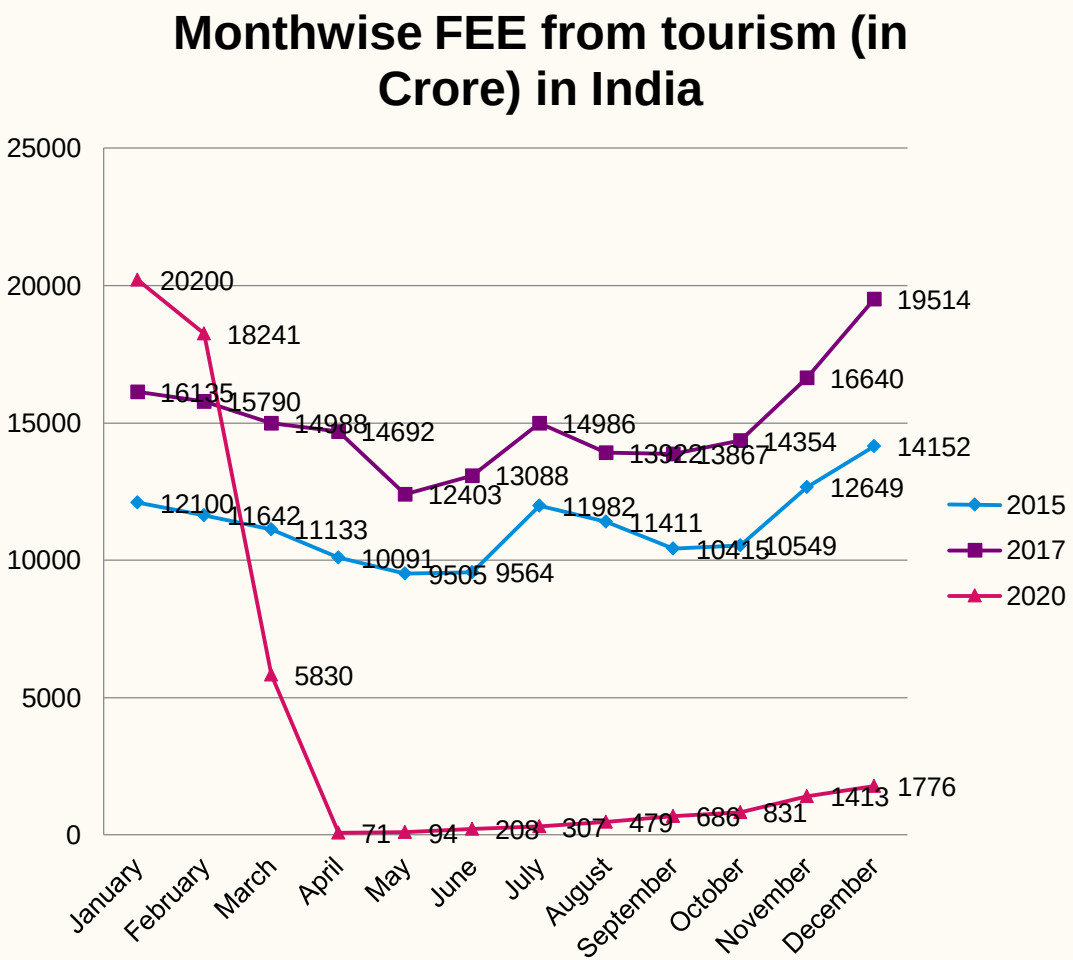
In [9]:
```
hotels.loc[hotels['PRICE_RUPEES'] ==749.1, 'HOTEL']
```

Out[9]:
```
77                          KRISHNA PALACE JAIPUR
89                        HOTEL REEVE INN JAIPUR
123         KALYAN HERITAGE AND PARADISE JAIPUR
167                   HOTEL HAYAT RABBANI JAIPUR
193                 HOTEL ABHINEET PALACE JAIPUR
196                              ABHI ROOMS JAIPUR
202    JAIPUR CM A LUXURY BOUTIQUE HOTEL JAIPUR
207                        THE AMER VALLEY JAIPUR
Name: HOTEL, dtype: object
```

In [10]:
```
hotels.loc[hotels['PRICE_RUPEES'] >=1000, 'HOTEL']
```

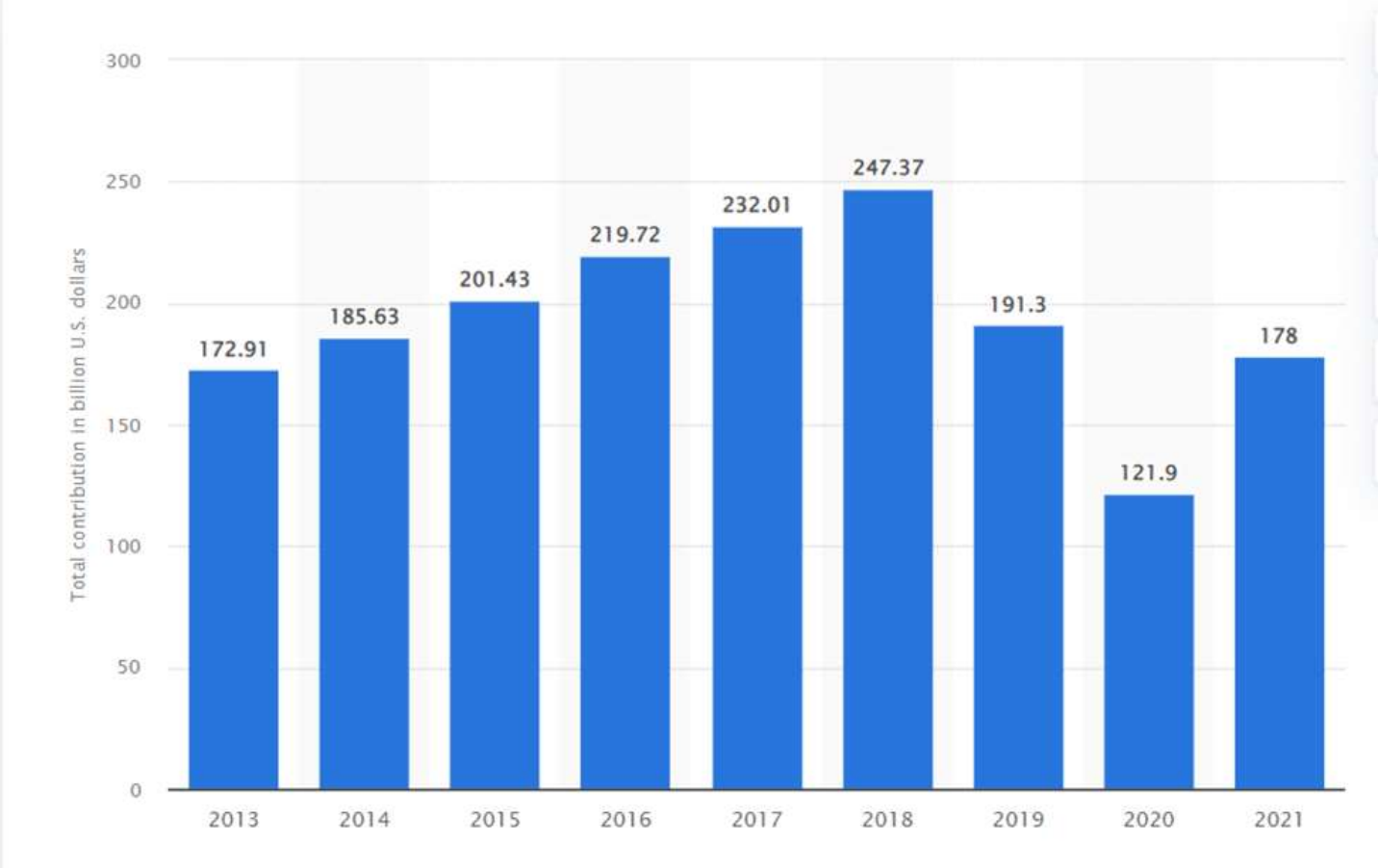
Out[10]:
```
1                                        HILTON JAIPUR
2                                      TRIDENT, JAIPUR
3                                THE FERN RESIDENCY JAIPUR
4      ITC RAJPUTANA, JAIPUR A LUXURY COLLECTION HOTE...
5                  HOLIDAY INN JAIPUR CITY CENTRE JAIPUR
                              ...
246                                  HOTEL SAVOY JAIPUR
247                           HOTEL JAI MAA PALACE JAIPUR
248                               THE PRIME HOTEL JAIPUR
249                           COMFORT INN SAPPHIRE JAIPUR
251                             GIRISADAN HOMESTAY JAIPUR
Name: HOTEL, Length: 202, dtype: object
```

# IMPACT OF COVID-19




The foreign exchange earnings of the country was considerably stable in the year 2015 and the same trend can be seen for the year 2017 just with increased in the scale. But for the year 2020, the foreign exchange earning was highest with 20200 (Cr) in month of January and showed a sudden drop in month of April with least foreign exchange earning of 71 (Cr) and further increased at slower rate .Hence , It proves that the covid-19 crises had drastic impact on the economy.

# Total Contribution Of Tourism To GDP (2013-2022)




Tourism sectors contribution is seen increasing year on year basis since 2013.

The trend is seen disrupted in 2019 due to Covid-19 pandemic that caused travel ban world wide.

The sector is seen regaining its strength in 2021

# Regression Analysis

- **Regression 1:**
- **Independent variable**: Foreign Direct Investment (FDI)
- **Dependent Variable** : Gross Domestic Product
- Hypothesis to be tested:
- **Ho**: FDI do not have significant effect on GDP
- **H1**: FDI do have significant effect on GDP
- The data taken of GDP and FDI is yearly from 2000-2020

```
>reg=lm(y~x);reg
```

- Call:

  lm(formula = y ~ x)

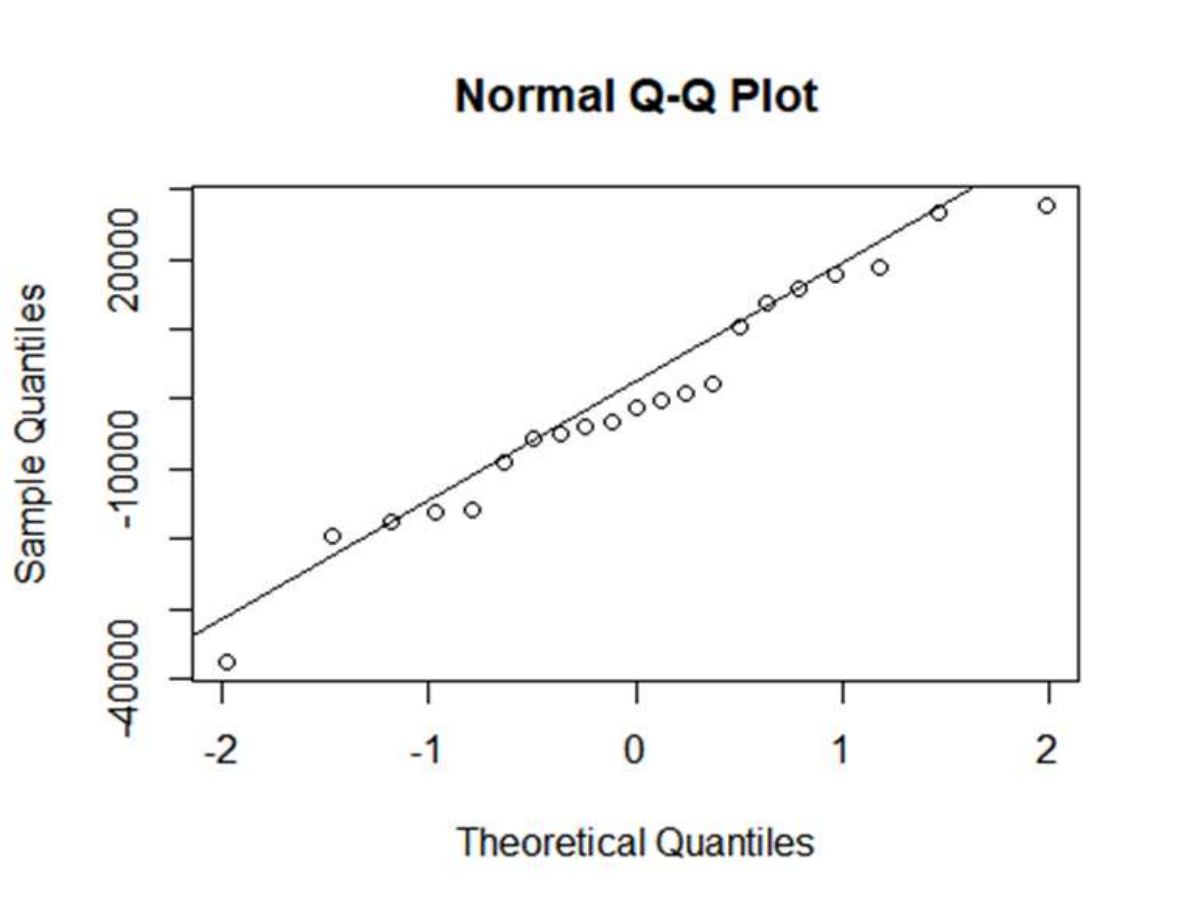
Coefficients:

| (Intercept) | x |
|---|---|
| 4.957e+04 | 3.222e-01 |

**Regression Equation:**
**Y= 4.957e+04 + 3.222e-01 x**

Checking Normality Assumptions

>qqnorm(res)
>qqline(res)




>shapiro.test(res)

p-value = 0.7383

As p-value> 0.5 ----→Normality Satisfied

>summary(reg)

Multiple R-squared:  0.8624, Adjusted R-squared:  0.8552

Conclusion:

We, therefore conclude that FDI has significant impact on GDP.

Coefficient of determination (R2) is 0.8624. i.e 86.24% variation in GDP is explained by FDI.

Thereby indicating positive correlation between the two.

## Regression 2:

**Independent variable**: Foreign Exchange Earnings(FEE)
**Dependent Variable** : Gross Domestic Product

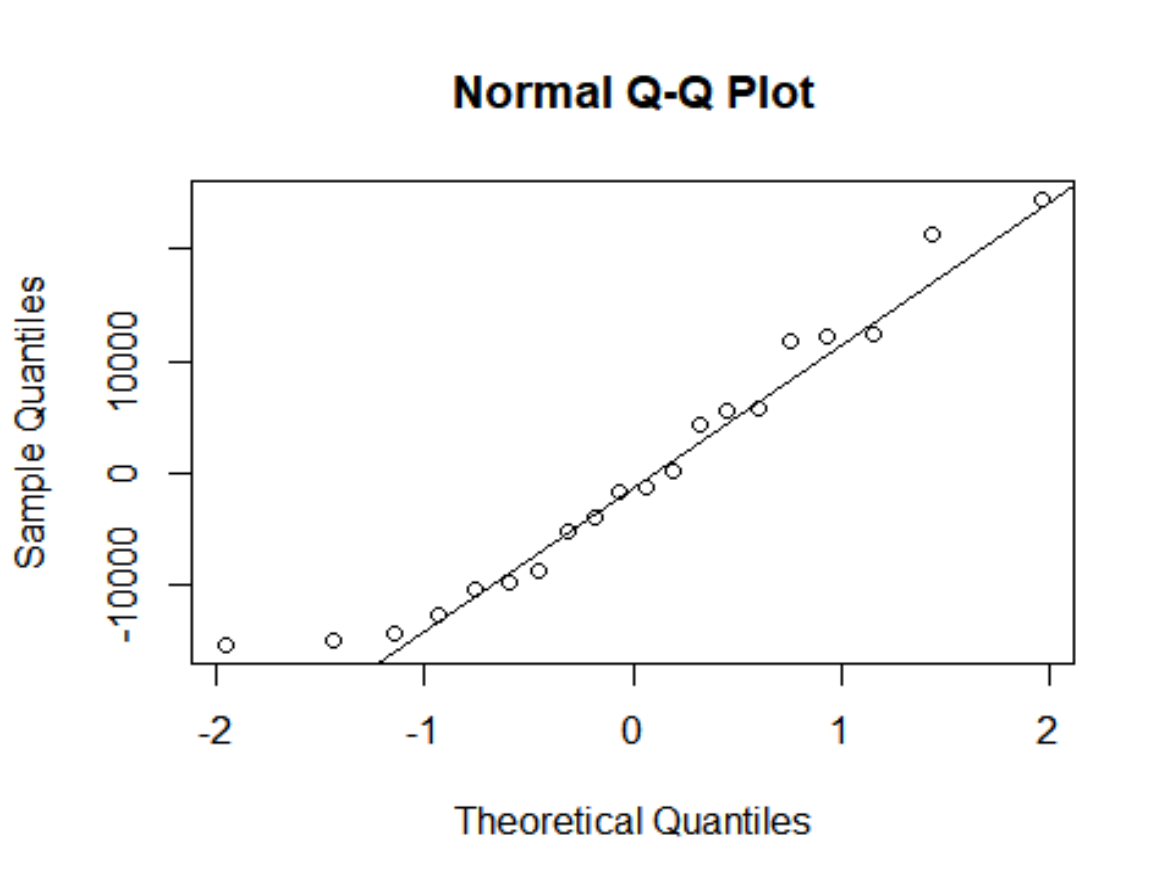
Hypothesis to be tested:
**Ho**: FEE do not have significant effect on GDP
**H1**: FEE do have significant effect on GDP

```
>reg=lm(y~x);reg
Call:
 lm(formula = y ~ x)

 Coefficients:
(Intercept)           x
  4.035e+04    6.632e-01
```

**Regression Model is**
**Y= 4.035e+04 +6.632e-01 x**

Checking Normality Assumptions
>qqnorm(res)
>qqline(res)



>shapiro.test(res)
p-value = 0.244
Since p-value>0.05 we accept Ho ie  Normality Satisfied

>summary(reg)

Multiple R-squared:  0.9256, Adjusted R-squared:  0.9215

Conclusion:

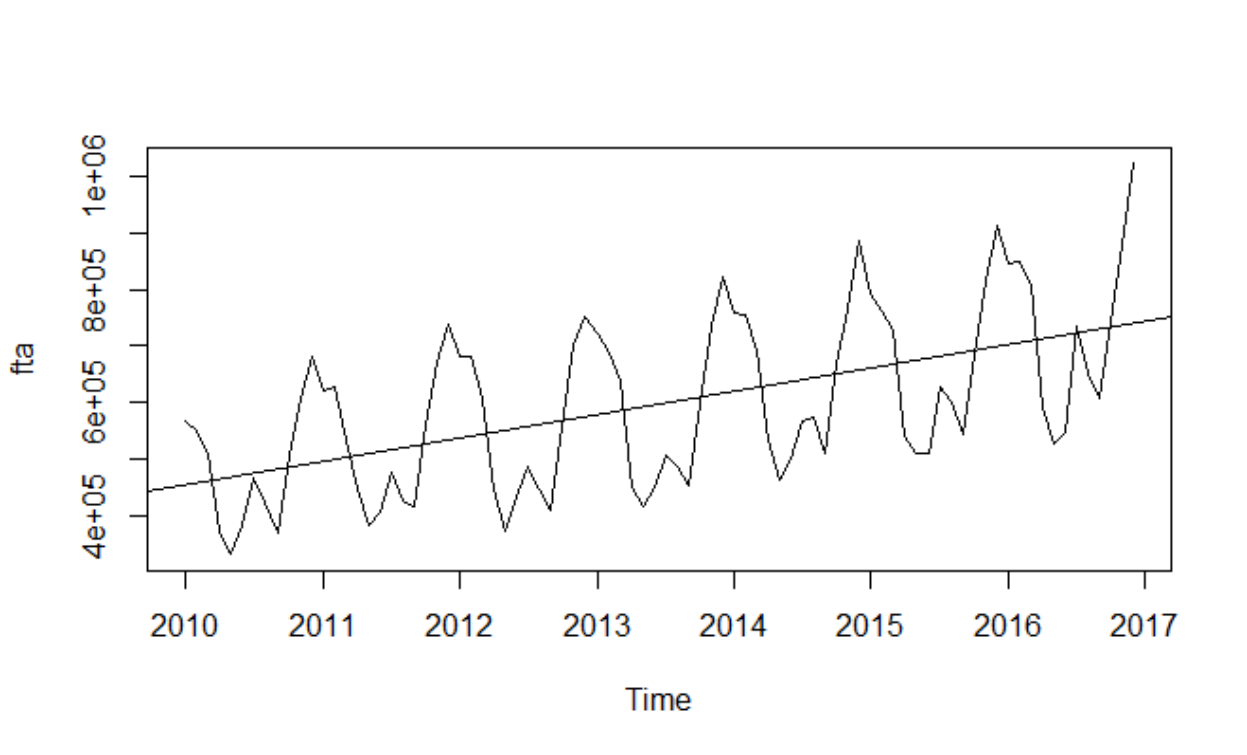
We, therefore conclude that FEE has significant impact on GDP.

Coefficient of determination (R2) is 0.9256 i.e 92.56% variation in GDP is explained by FEE. Thereby indicating positive correlation between the two.

# TIME SERIES ANALYSIS

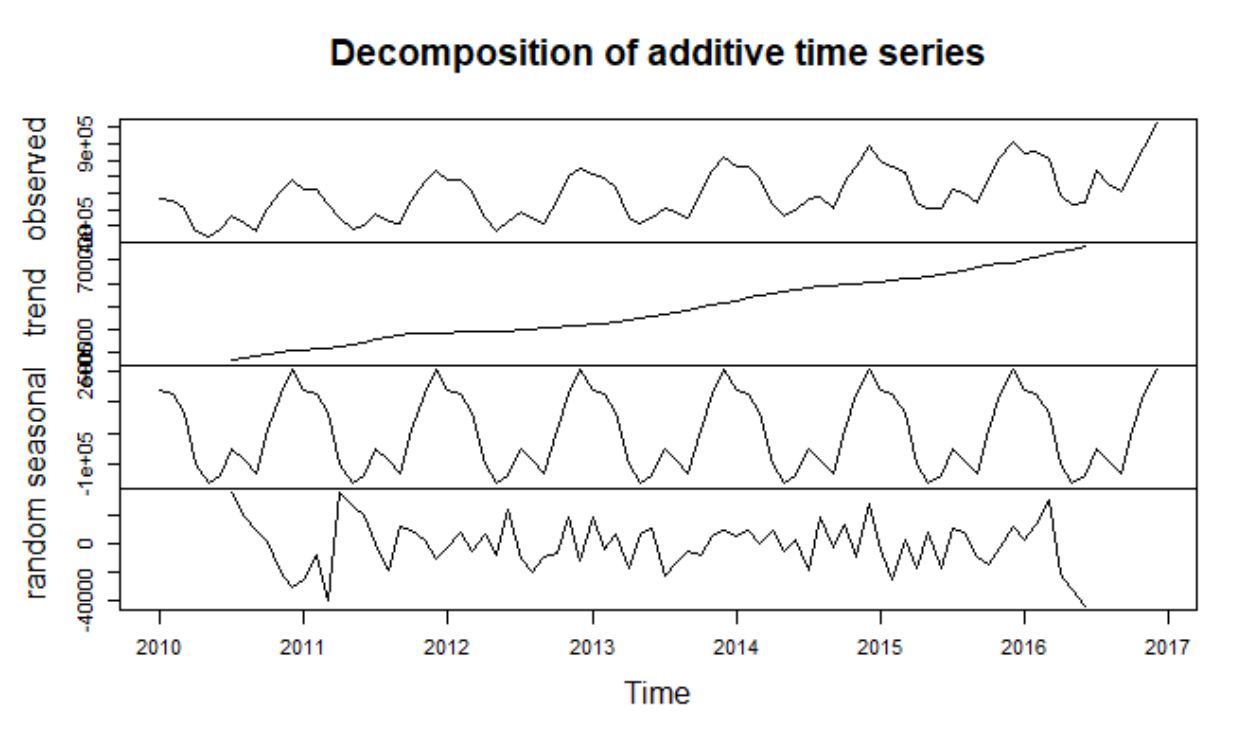
- Time series arise as recordings of processes which vary over time.
- Time series plot displays the values of the process output in the order in which the values occur.
- A key for analyzing a time series is to understand the form of any underlying pattern of the data ordered over time.
- We may wish to develop a simple mathematical model which explains the observed pattern of Y1, Y2, . . . , YT.
- A key idea in time series is that of stationarity. Roughly speaking, a time series is stationary if its behaviour does not change over time. This means, for example, that the values always tend to vary about the same level and that their variability is constant over time.
- The variable taken under study is : Foreign Tourist arrivals
- Data taken under study is monthly FTA form year 2010 -2016

```
plot.ts(fta)
abline(reg=lm(fta~time(fta)))
```




**Comment:** An increasing trend, seasonal component and time dependent variance is observed.

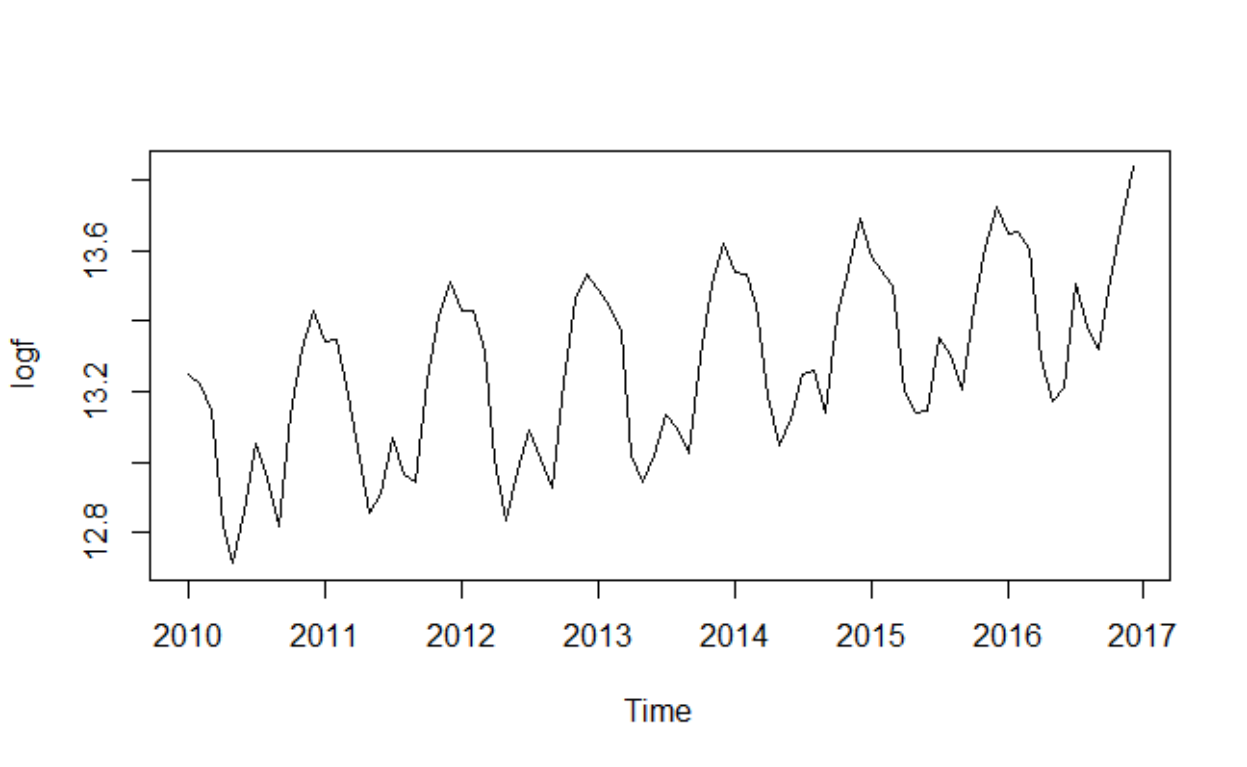
```
decompose=decompose(fta)
plot(decompose)
```




As, the variance is increasing with respect to time, to make it constant (time independent) we apply transformation (log transformation used)

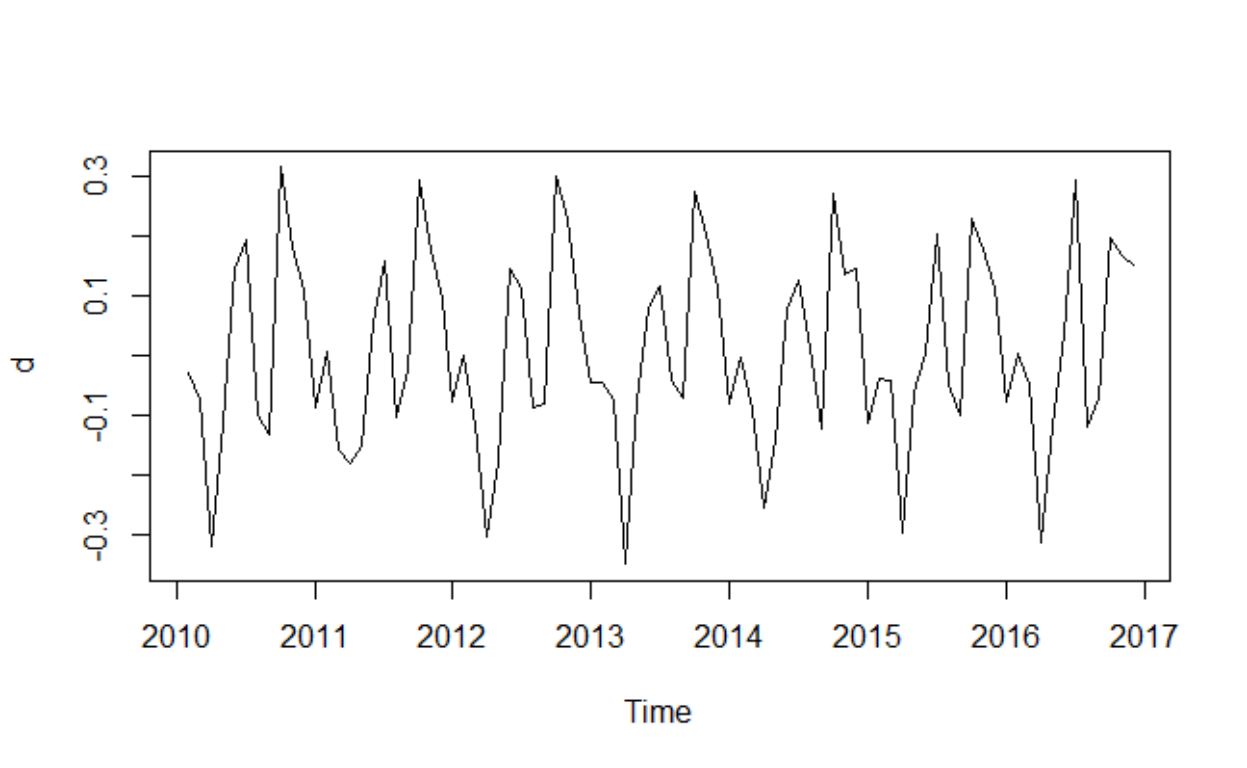
```
logf=log(fta)

plot.ts(logf)
```




**Comment**: The variance is now constant over the entire series. (Not changing wrt time).But trend is still present in the series. So, to remove trend , we difference the original data

```
d=diff(logf)
plot.ts(d)
```




Comment: The trend component is now removed.

So, the time variance and mean is independent of time.
The factors determining Stationarity of series are adjusted.

We go for checking Stationarity of series by,
Augmented Dickey Fuller Test

Ho: The series is not Stationary
H1: The series is Stationary

> adf.test(d)

Augmented Dickey-Fuller Test

data: d
Dickey-Fuller = -5.1382, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(d) : p-value smaller than printed p-value

Comment:Series is Stationary.
The vital step(Stationarity) to model the Time series is Satisfied.
We go for fitting the model

```
> ars=auto.arima(d,seasonal = TRUE);ars
Series: d
ARIMA(0,0,1)(0,1,1)[12]


Coefficients:
          ma1     sma1
      -0.6886  -0.4637
s.e.   0.1088   0.1572

sigma^2 estimated as 0.001624:  log likelihood=126.49
AIC=-246.98   AICc=-246.63   BIC=-240.2
```

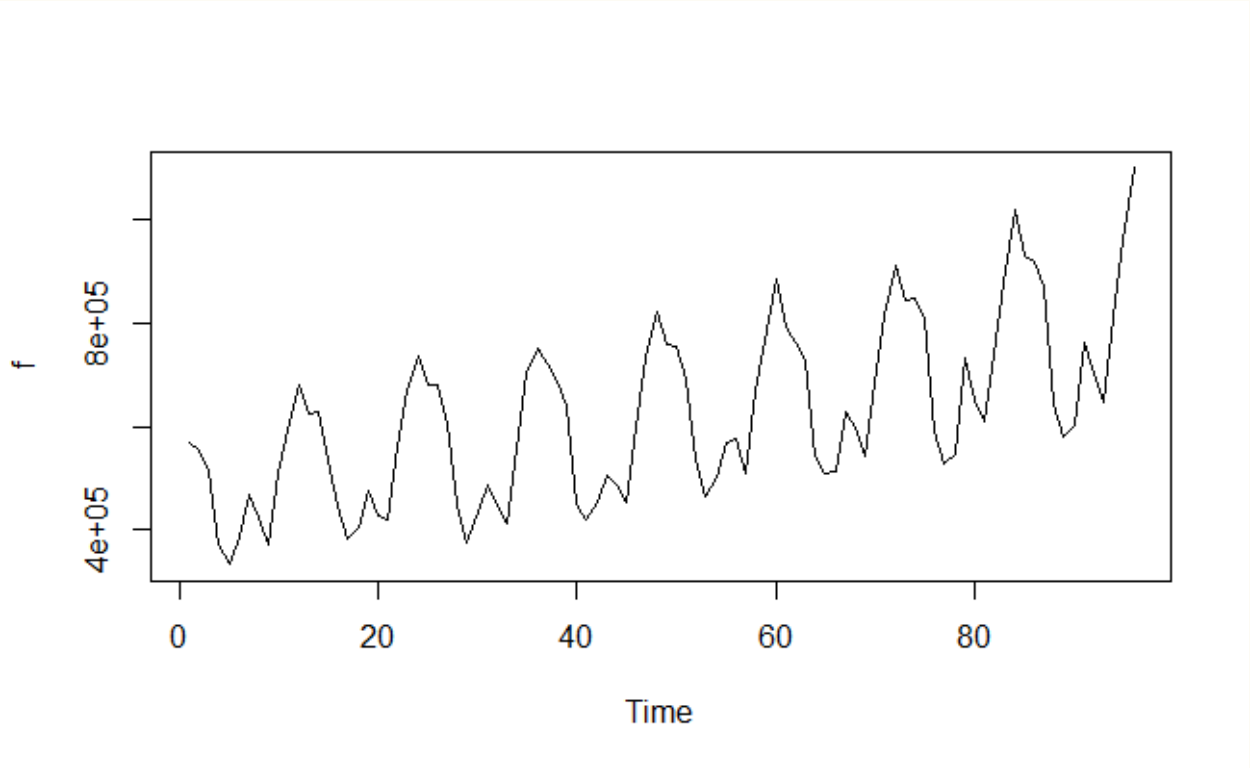
Fitted Model is :
Seasonal ARIMA (0,0,1)(0,1,1)[12]

After fitting the model, next step is to forecast

```
> x=predict(ars,n.ahead=1*12)
>x
[1]  929710.2  920696.2  870348.1  643337.5  579457.2  602649.3  762622.8
[8]  703927.6  647522.9  810040.5  961039.1 1100540.3
```

Hence we have predicted values Of Foreign Tourist arrival for next one year (2017)

```
> plot.ts(f)
```

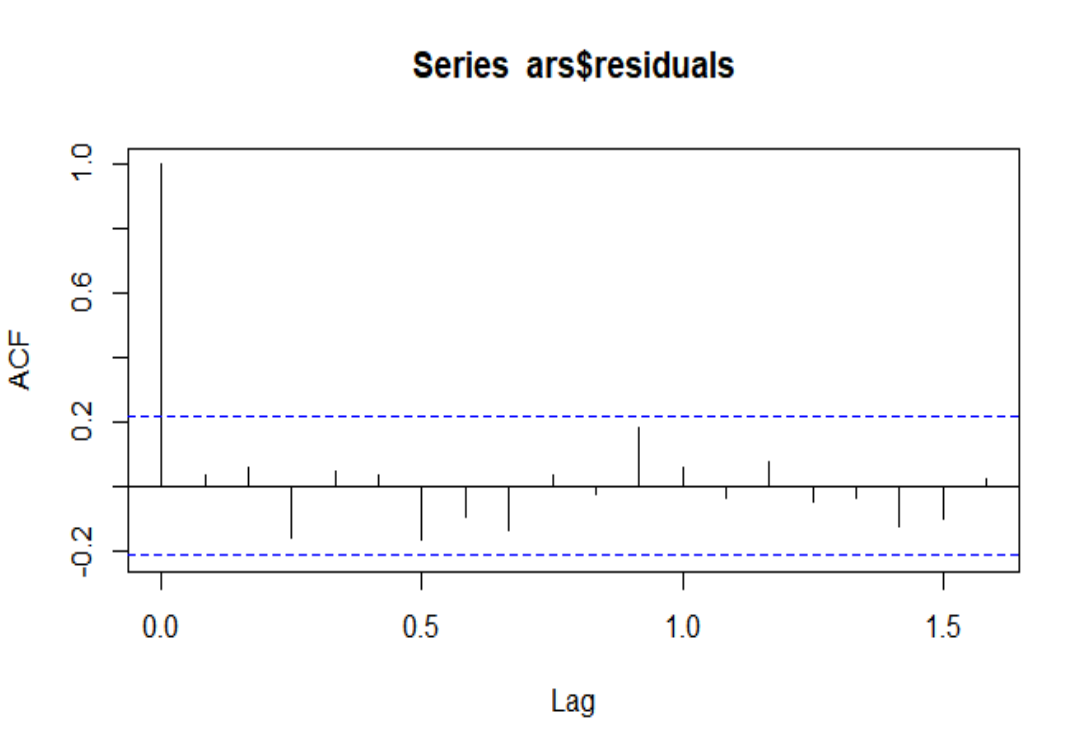


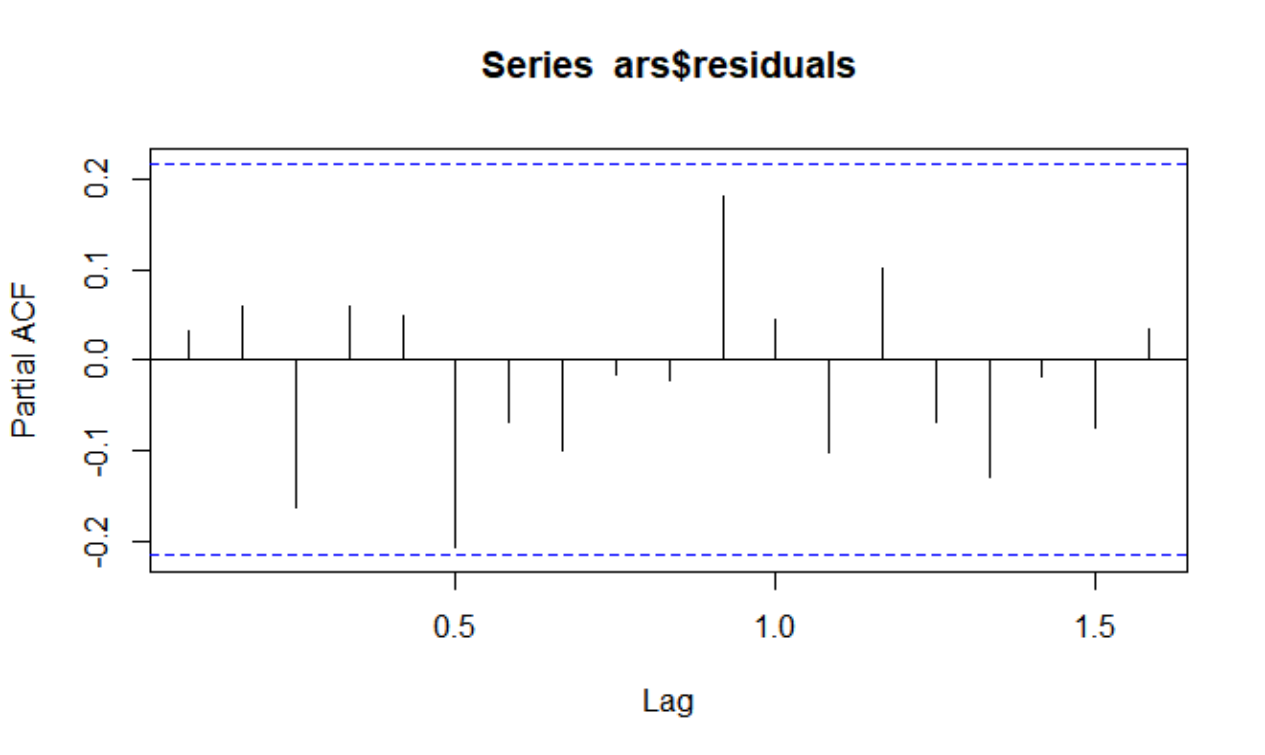

Graph of series with forecasted values of next 1 year

The important step of any analysis is its Validation.
(Basically we check the errors / residuals of the model)

>acf(ars$residuals)

> pacf(ars$residuals)



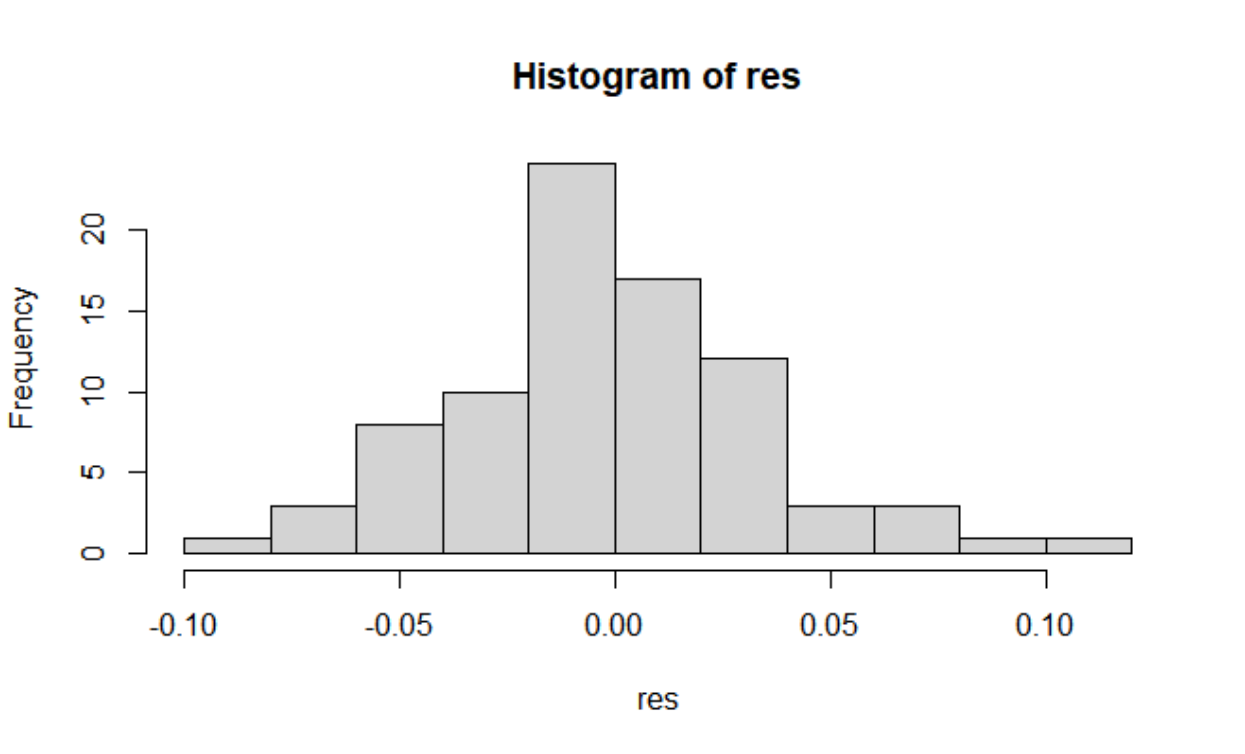



> Box.test(res,type="Ljung-Box")

Box-Ljung test

data:  res
X-squared = 0.092712, df = 1, p-value = 0.7608
Comment: No autocorrelation present within the residuals.
Residuals are random.

> hist(res)




Comment : Normality (behaviour) of the residuals can be observed

By using accuracy () function, we get the value of RMSE of the forecast as 0.03674856, since the value is significantly small, our forecast is accurate.

# Conclusion

- The objective of this project was to study the tourism statistics in India and impact of Covid-19 on tourism industry using various techniques like-Machine Learning and Time Series.
- Tourism industry contributes 5.8% to the total GDP of the country.
- From the exploratory analysis it is observed that

1. Bangladesh is the leading contributor of foreign tourist arrival in India
2. The main reason of foreign tourist arrival in india is for leisure purpose followed by indian diaspora and business purpose respectively.
3. Foreign tourist arrivals for medical purpose has seen increasing trend from 2018-2022.

- Hotel recommendation model done with the help of machine learning helps find a budget friendly accommodation facility in a range as low as Rs 150/- per day to as high as Rs 112365/- per day
- A SLR model for FDI from tourism sector and GDP is fitted. It is found that the two variables are positively related with Karl Pearson Correlation coefficient 0.8624.
  i.e FDI in tourism industry plays a significant role on the GDP of the country.
- A SLR model for FEE and GDP is fitted and found that the two variables are positively related with Karl Pearson Correlation coefficient 0.9256.
  i.e FEE plays a significant role on the GDP of the country.

ARIMA(0,0,1)(0,1,1)[12] model Time series is fitted and future foreign tourist arrivals are predicted

# Limitation and scope for future study

- As tourism sector is affected by many other factors like global geo-political tensions, connectivity facilities, strength of visa etc so it is not feasible to get the complete picture

- Inadequacy of data in required form and time span.

- The geographical area under study was limited to India which can be further expanded on the global level.