Report On

Tourist Behavior Analysis

Submitted in partial fulfillment of the requirements of the Course project in Semester VII of Fourth Year Computer Engineering

By
Vipul Bhoir(Roll No.07)
Mrudul Chaudhari(Roll No. 12)
Abhinav Desai(Roll No. 14)
Supervisor
Mrs. Sneha Mhatre

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(2024-25)

Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

CERTIFICATE

This is to certify that the project entitled “Tourist Behavior Analysis” is a bonafide work of "Vipul Bhoir(Roll No.07), Mrudul Chaudhari(Roll No. 12), Abhinav Desai(Roll No. 14)” submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in semester VII of Fourth Year Computer Engineering.

.

Supervisor
Mrs. Sneha Mhatre

Dr. Megha Trivedi
Head of Department

Abstract

Personal Tourist behavior analysis is the study of the motivations, decisions, and actions of tourists. This project aims to develop a system for tourist behavior analysis using R. The system will use a variety of data sources to identify patterns and trends in tourist behavior, and to predict tourist behavior based on past trends and current events. The system is designed as a modular system, with each module responsible for a specific task. The system is expected to be useful for a variety of stakeholders, including tourism businesses, policymakers, and researchers.

1 Introduction

1.1 Introduction

Tourist behavior analysis is the study of the motivations, decisions, and actions of tourists. It is a complex field that encompasses a wide range of factors, including psychology, sociology, economics, and geography. By understanding tourist behavior, tourism businesses and policymakers can better develop and promote products and services that meet the needs of tourists.

1.2 Problem Statement

The tourism industry is highly competitive, and businesses need to constantly innovate to stay ahead of the curve. One way to do this is to better understand the needs and wants of tourists. However, tourist behavior is complex and can be difficult to predict. This makes it challenging for tourism businesses to develop and implement effective marketing strategies

1.3 Project Scope

This project aims to develop a system for tourist behavior analysis using R. The system will use a variety of data sources, such as social media data, travel surveys, and booking records, to identify patterns and trends in tourist behavior. The system will also be able to predict tourist behavior based on past trends and current events.

2. Requirement Analysis

2.1 Software Requirements:

The system will require the following software:

- R
- RStudio
- Tidyverse libraries
- Other relevant libraries (e.g., ggplot2, caret, etc.)

2.2 Hardware Requirements

The system will require the following hardware:

A computer with at least 4GB of RAM and 100GB of free disk space

An internet connection

Recommended:

- 16 GB RAM

Minimum:

- 8 GB RAM

2.3 Functional Requirements

The system must be able to perform the following functions:

- Collect and clean data from a variety of sources
- Identify patterns and trends in tourist behavior
- Predict tourist behavior based on past trends and current events
- Visualize the results of the analysis

2.4 Nonfunctional Requirements

Performance: The system should be able to handle large datasets and complex queries efficiently. The system should be able to generate results in a reasonable amount of time. The system should be able to handle concurrent users without impacting performance.

Security: The system should be secure from unauthorized access, modification, or destruction of data. The system should protect user privacy. The system should be compliant with all relevant security regulations.

Reliability: The system should be highly available and reliable. The system should be able to recover from failures quickly and minimize downtime. The system should be regularly monitored and backed up.

Usability: The system should be easy to use and navigate. The system should be well-documented. The system should be accessible to users with disabilities.

Scalability: The system should be scalable to handle increasing data volumes and user loads. The system should be modular and designed to support future growth. The system should be deployed in a cloud environment to facilitate scalability.

Maintainability: The system should be well-designed and organized, making it easy to maintain and update. The system should be documented with code comments and documentation. The system should be tested regularly to ensure that it is working properly.

Portability: The system should be portable and able to run on a variety of platforms.

Extensibility: The system should be extensible, allowing for new features and functionality to be added easily.

Interoperability: The system should be interoperable with other systems, such as CRM and ERP systems.

3. System Design

3.1 System Design:

The system will be designed as a modular system, with each module responsible for a specific task.

The following are the main modules of the system:

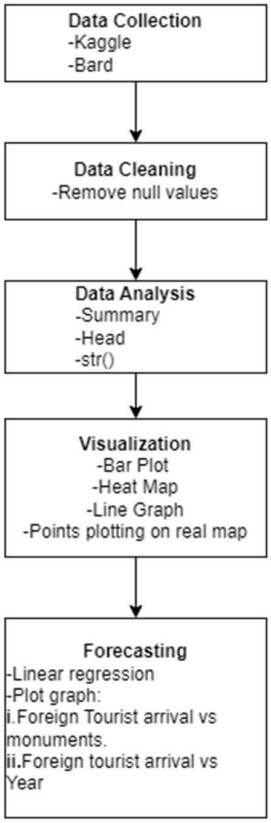
Data collection and cleaning module: This module will collect data from a variety of sources and clean it to a consistent format.

Data analysis module: This module will identify patterns and trends in the data, and predict tourist behavior based on past trends and current events.

Visualization module: This module will visualize the results of the analysis in a clear and concise way.

Prediction module: This module will analyze the pattern and forecast the tourist increase in the years for the monuments.

3.2 Diagram



3.2 Module Description:

Data collection and cleaning module: This module will collect data from a variety of sources, including social media data, travel surveys, and booking records. The data will then be cleaned to a consistent format so that it can be easily analyzed.

Data analysis module: This module will use a variety of statistical and machine learning techniques to identify patterns and trends in the data. The module will also be able to predict tourist behavior based on past trends and current events.

Visualization module: This module will visualize the results of the analysis in a clear and concise way. The module will generate a variety of charts and graphs that can be used to understand the findings of the analysis

Prediction module: This module will analyze the pattern and forecast the tourist increase in the years for the monuments.

4. Implementation

4.1 Methodology

1. Data collection

The first step is to collect data from a variety of sources. This may include social media data, travel surveys, booking records, and other relevant data sources. The data should be cleaned and preprocessed to ensure that it is in a consistent format and that any errors or missing values are addressed.

2. Data analysis

Once the data is prepared, you can begin to analyze it using a variety of statistical and machine learning techniques. This may involve identifying patterns and trends in the data, such as the most popular tourist destinations, the most popular activities, and the spending habits of tourists. You can also use data analysis to predict tourist behavior, such as the likelihood of a tourist visiting a particular destination or participating in a particular activity.

3. Visualization

Once you have analyzed the data, you can use visualization tools to present the results in a clear and concise way. This may involve creating charts, graphs, and maps that illustrate the patterns and trends that you have identified. Visualization can also be used to communicate the findings of your analysis to a variety of stakeholders, such as tourism businesses, policymakers, and researchers.

4. Deployment

Once you have developed and tested your system, you can deploy it to production. This may involve making the system available to users over the web or through a mobile app. You may also need to develop and implement maintenance and support procedures for the system.

Here are some additional details about each step:

Data collection:

When collecting data, it is important to consider the following:

Data sources: There are a variety of data sources that can be used for tourist behavior analysis.

Some common data sources include:

Social media data: Social media data can be used to track tourist movements, identify popular tourist destinations, and understand tourist sentiment.

Travel surveys: Travel surveys can be used to collect data on tourist demographics, travel motivations, and spending habits.

Booking records: Booking records can be used to track tourist itineraries, identify popular activities, and understand tourist spending.

Data sampling: Data sampling is the process of selecting a subset of data from a larger population. This can be useful for reducing the cost and complexity of data collection.

Data cleaning: Data cleaning is the process of identifying and correcting errors and inconsistencies in data. This is an important step in preparing data for analysis.

Data analysis:

There are a variety of statistical and machine learning techniques that can be used for tourist behavior analysis. Some common techniques include:

Descriptive statistics: Descriptive statistics can be used to summarize the data and identify patterns and trends.

Multivariate analysis: Multivariate analysis can be used to identify relationships between multiple variables.

Machine learning: Machine learning can be used to develop models that can predict tourist behavior.

Visualization:

There are a variety of visualization tools that can be used to present the results of your analysis. Some common visualization tools include:

Charting tools: Charting tools can be used to create a variety of charts, such as bar charts, line charts, and pie charts.

Graphing tools: Graphing tools can be used to create a variety of graphs, such as scatter plots and histograms.

Mapping tools: Mapping tools can be used to create maps that show the distribution of tourists or other relevant data.

Deployment:

When deploying your system, you need to consider the following:

System architecture: The system architecture should be designed to support the performance, security, and reliability requirements of the system.

User interface: The user interface should be designed to be easy to use and navigate.

Security: The system should be deployed in a secure environment and should be protected from unauthorized access.

Maintenance and support: You should develop and implement maintenance and support procedures for the system.

4.2 Sample Modules

1. Data collection

The first step is to collect data from a variety of sources. This may include social media data, travel surveys, booking records, and other relevant data sources. The data should be cleaned and preprocessed to ensure that it is in a consistent format and that any errors or missing values are addressed.

2. Data analysis

Once the data is prepared, you can begin to analyze it using a variety of statistical and machine learning techniques. This may involve identifying patterns and trends in the data, such as the most popular tourist destinations, the most popular activities,

and the spending habits of tourists. You can also use data analysis to predict tourist

behavior, such as the likelihood of a tourist visiting a particular destination or participating in a particular activity.

3. Visualization

Once you have analyzed the data, you can use visualization tools to present the results in a clear and concise way. This may involve creating charts, graphs, and maps that illustrate the patterns and trends that you have identified. Visualization can also be used to communicate the findings of your analysis to a variety of stakeholders, such as tourism businesses, policymakers, and researchers.

4. Deployment

Once you have developed and tested your system, you can deploy it to production. This may involve making the system available to users over the web or through a mobile app. You may also need to develop and implement maintenance and support procedures for the system.

4.3 Code

```
getwd()
#get data
data2 <- read.csv("india tour growth dataset.csv")

#data analysis
head(data2)
summary(data2)
str(data2)

#Data Cleaning
any(is.na(data2))
data2 <- na.omit(data2)
```

#Calculating Growth

```
data2$GrowthDomestic <- ((data2$Domestic.2020.21 -
data2$Domestic.2019.20) / data2$Domestic.2019.20) * 100
data2$GrowthForeign <- ((data2$Foreign.2020.21 -
data2$Foreign.2019.20) / data2$Foreign.2019.20) * 100

#Calculating Averages
average_growth_domestic <- mean(data2$GrowthDomestic, na.rm =
TRUE)
average_growth_foreign <- mean(data2$GrowthForeign, na.rm =
TRUE)
cat("Average Domestic Growth:", average_growth_domestic, "\n")
cat("Average Foreign Growth:", average_growth_foreign, "\n")

#Display of Bar-Plot
library(ggplot2)
# Create a data frame for plotting
avg_growth_data <-
data.frame( Category = c("Domestic",
"Foreign"),
  AverageGrowth = c(average_growth_domestic,
average_growth_foreign)
)

bar_plot <- ggplot(avg_growth_data, aes(x = Category, y =
AverageGrowth)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Average Growth Percentages",
       x = "Visitor Type",
```

```
y = "Average Growth (%)")
```

```
print(bar_plot)

#Heat-map

library(reshape2)
data_20 <- data2[1:20, ]

# Select the relevant columns and rename them with periods
data_subset <- data_20[, c("Name.of.the.Monument",
"Domestic.2019.20", "Foreign.2019.20", "Domestic.2020.21",
"Foreign.2020.21")]
colnames(data_subset) <- c("Monument", "Domestic.2019.20",
"Foreign.2019.20", "Domestic.2020.21", "Foreign.2020.21")

# Melt the data to create a format suitable for a heatmap
melted_data <- melt(data_subset, id.vars = "Monument")

# Create a heatmap
heatmap_plot <- ggplot(melted_data, aes(x = variable, y = Monument,
fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "blue") +
  labs(
    x = "Monument",
    y = "Year",
    fill = "Value"
  ) +
  theme_minimal() +
```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
print(heatmap_plot)

# Line Graph
ggplot(data = data_20, aes(x = Circle)) +
  geom_line(aes(y = Domestic.2019.20, color = "Domestic 2019-20"),
size = 1) +
  geom_line(aes(y = Foreign.2019.20, color = "Foreign 2019-20"),
size = 1) +
  geom_line(aes(y = Domestic.2020.21, color = "Domestic 2020-21"),
size = 1) +
  geom_line(aes(y = Foreign.2020.21, color = "Foreign 2020-21"),
size = 1) +
  xlab("City") +
  ylab("Number of Visitors") +
  labs(color = "Visitor Type") +
  theme_minimal() +
  theme(legend.position = "top") +
  scale_color_manual(values = c("Domestic 2019-20" = "blue",
                                "Foreign 2019-20" = "red",
                                "Domestic 2020-21" = "green",
                                "Foreign 2020-21" = "purple")) +
  ggtitle("Monument Visitors Growth Over the Years")

#Plot locations on map

library(leaflet)
latlong <- read.csv("lonandlat2.csv")
mymap <- leaflet(data = latlong) %>%
```

addTiles()

```
mymap <- mymap %>%
  addMarkers(
    lng = ~Longitude,
    lat = ~Latitude,
    popup = ~paste("City:", City, "<br>Monument:",
`Name.of.the.Monument`)
  )
mymap

#forecasting for domesstic growth
library(forecast)

# Create a dataset from the provided data
data_for_pred <- data.frame(
  Circle = c(data2$Circle),
  Name_of_the_Monument = c(data2$Name.of.the.Monument),
  Foreign_2019_20 = c(data2$Foreign.2019.20),
  Foreign_2020_21 = c(data2$Foreign.2020.21)
)

foreign_ts <- ts(data_for_pred$Foreign_2019_20, start = c(2019, 1),
frequency = 1)
# Assuming that foreign growth is approximately linear
# Create a linear model to predict foreign tourist growth
foreign_model <- lm(data_for_pred$Foreign_2019_20 ~
time(foreign_ts))

# Create a time series object for the next 2 years
```

```
future_years <- ts(2021:2022, frequency = 1)
```

```
# Predict foreign tourists for the next 2 years based on the model
predicted_growth <- predict(foreign_model, newdata =
data.frame(time = time(future_years)))
# Plot the historical and predicted values
# Create a sample dataset
plot(data_for_pred$Foreign_2019_20, type = "o", xlab = "Row no",
ylab = "Foreign Tourist Arrivals", col = "blue", main = "Foreign
Tourist Growth Prediction")
lines(data_for_pred$Foreign_2020_21, type = "o", col = "red")
lines(predicted_growth, type = "o", col = "green")
# Add a legend
legend("topright", legend = c("2019-20", "2020-21", "Predicted 2021-
22", "Predicted 2022-23"), col = c("blue", "red", "green"), lty = 1, cex
= 0.8)

data_transposed <- t(data_for_pred[, c("Foreign_2019_20",
"Foreign_2020_21")])

# Plot each row as a time series
matplot(data_transposed, type = "l", xlab = "Year", ylab = "Foreign
Tourist Arrivals", col = 1:nrow(data_transposed), lty = 1, main =
"Foreign Tourist Growth Prediction")

# Add a legend
legend("topright", legend = data_for_pred$Name_of_the_Monument,
col = 1:nrow(data_transposed), lty = 1, cex = 0.8)
```

4.4 Output:

1. Data Summarization:

```
> head(data2)
  Circle  Name.of.the.Monument Domestic.2019.20 Foreign.2019.20 Domestic.2020.21
1   Agra             Taj Mahal          4429710          645415          1259892
2   Agra             Agra Fort          1627154          386522           371242
3   Agra        Fatehpur Sikri           454376          184751           107835
4   Agra   Akbar Tomb Sikandra           229270           19625            99509
5   Agra  Mariam tomb Sikandra            22517             414             9765
6   Agra Itimad-ud-Daulah-Tomb           132800           82692            41016
  Foreign.2020.21 X..Growth.2021.21.2019.20.Domestic X..Growth.2021.21.2019.20.Foreign
1            9034                              -71.56                           -98.60
2            2810                              -77.18                           -99.27
3             574                              -76.27                           -99.69
4             321                              -56.60                           -98.36
5              31                              -56.63                           -92.51
6             410                              -69.11                           -99.50
>
```

```
> summary(data2)
    Circle          Name.of.the.Monument Domestic.2019.20  Foreign.2019.20  Domestic.2020.21
 Length:144         Length:144           Min.   :    530   Min.   :     0   Min.   :      0
 Class :character   Class :character     1st Qu.:  39408   1st Qu.:   140   1st Qu.:  12243
 Mode  :character   Mode  :character     Median : 118130   Median :  1065   Median :  46148
                                         Mean   : 302827   Mean   : 19143   Mean   :  91341
                                         3rd Qu.: 335418   3rd Qu.:  7759   3rd Qu.: 102457
                                         Max.   :4429710   Max.   :645415   Max.   :1259892

 Foreign.2020.21     X..Growth.2021.21.2019.20.Domestic X..Growth.2021.21.2019.20.Foreign
 Min.   :     0.00   Min.   : -99.99                    Min.   : -100.00
 1st Qu.:    11.75   1st Qu.: -77.00                    1st Qu.:  -98.98
 Median :    75.00   Median : -64.20                    Median :  -96.78
 Mean   :  2887.91   Mean   : -20.43                    Mean   :  849.58
 3rd Qu.:   233.50   3rd Qu.: -50.85                    3rd Qu.:  -82.22
 Max.   :105816.00   Max.   :4233.77                    Max.   :62078.43
                     NA's   :1                          NA's   :2
>
```

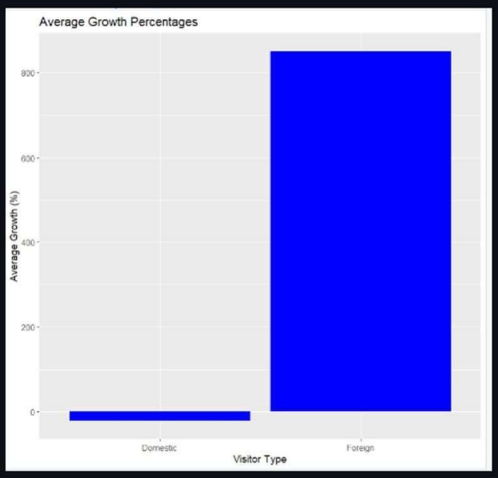
```
> str(data2)
'data.frame':   144 obs. of  8 variables:
 $ Circle                           : chr  "Agra" "Agra" "Agra" "Agra" ...
 $ Name.of.the.Monument             : chr  "Taj Mahal" "Agra Fort" "Fatehpur Sikri" "Akbar Tomb Si
kandra" ...
 $ Domestic.2019.20                 : int  4429710 1627154 454376 229270 22517 132800 84051 178574
474462 74597 ...
 $ Foreign.2019.20                  : int  645415 386522 184751 19625 414 82692 355 62325 12536 13
628 ...
 $ Domestic.2020.21                 : int  1259892 371242 107835 99509 9765 41016 18599 62652 9118
5 27201 ...
 $ Foreign.2020.21                  : int  9034 2810 574 321 31 410 54 544 321 35 ...
 $ X..Growth.2021.21.2019.20.Domestic: num  -71.6 -77.2 -76.3 -56.6 -56.6 ...
 $ X..Growth.2021.21.2019.20.Foreign : num  -98.6 -99.3 -99.7 -98.4 -92.5 ...
```

2. Data Cleaning:

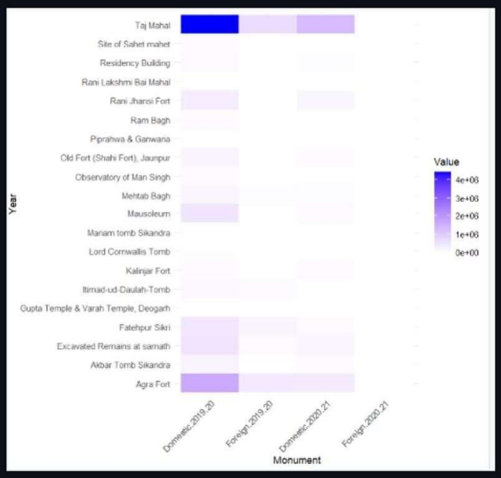
```
> #Data Cleaning
> any(is.na(data2))
[1] TRUE
> data2 <- na.omit(data2)
> #Calculating Growth
> data2$GrowthDomestic <- ((data2$Domestic.2020.21 - data2$Domestic.2019.20) / data2$Domestic.2019.
20) * 100
> data2$GrowthForeign <- ((data2$Foreign.2020.21 - data2$Foreign.2019.20) / data2$Foreign.2019.20)
* 100
> #Calculating Averages
> average_growth_domestic <- mean(data2$GrowthDomestic, na.rm = TRUE)
> average_growth_foreign <- mean(data2$GrowthForeign, na.rm = TRUE)
> cat("Average Domestic Growth:", average_growth_domestic, "\n")
Average Domestic Growth: -20.09301
> cat("Average Foreign Growth:", average_growth_foreign, "\n")
Average Foreign Growth: 849.577
```

3. Data Visualization:

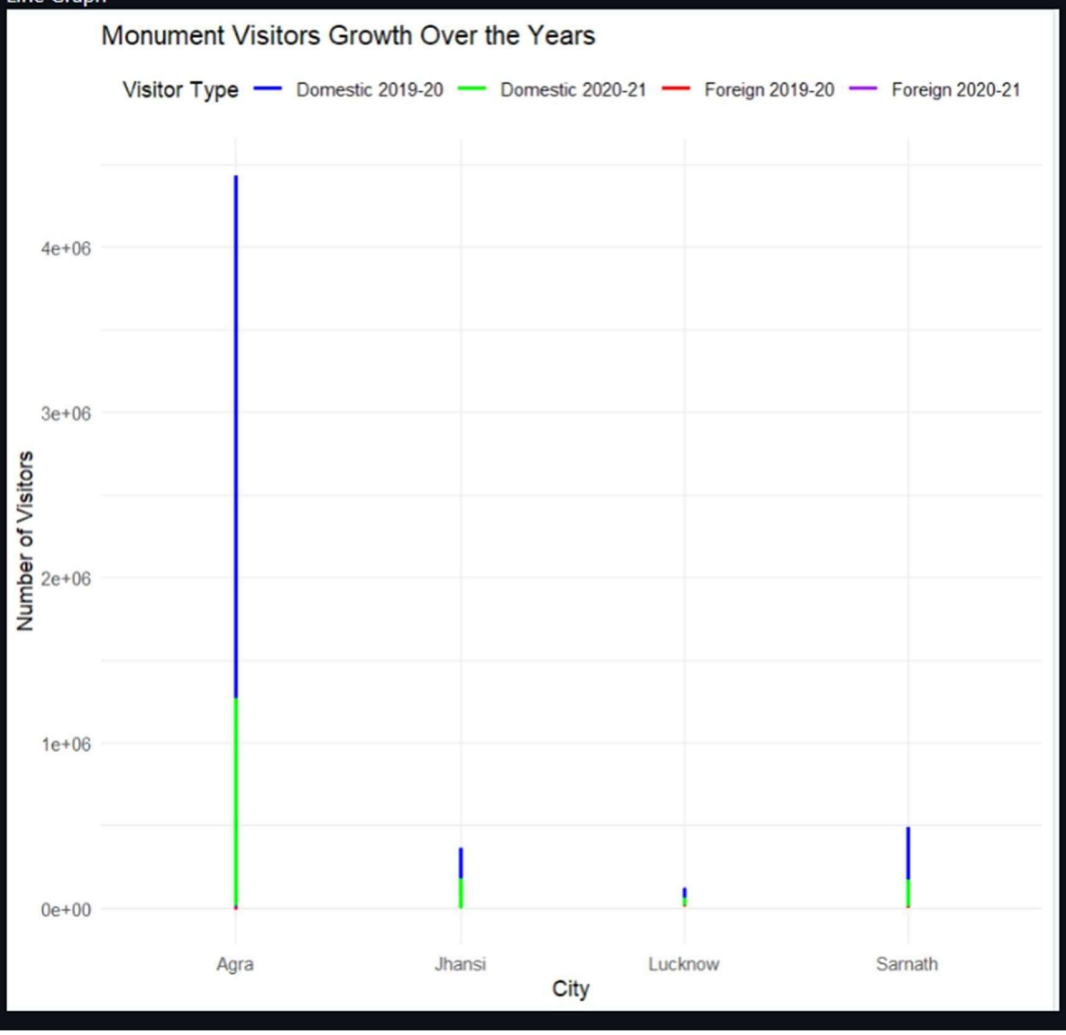
Bar Plot



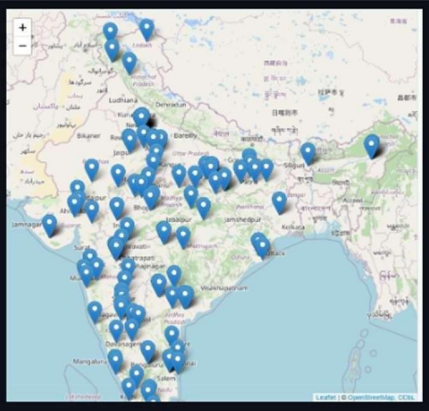
Heat Map



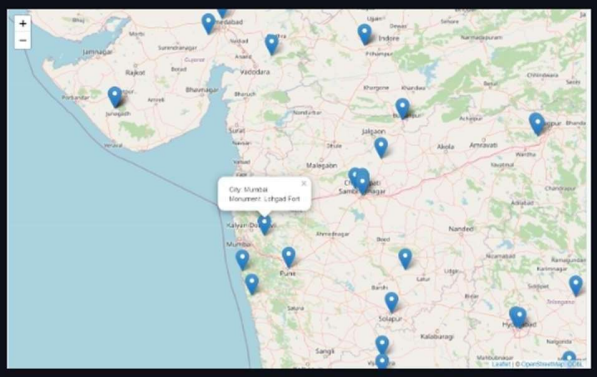
Line Graph



Point plotting on map

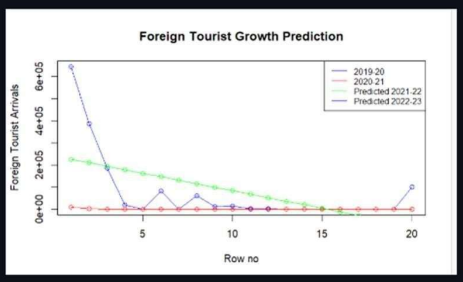


Map Pop-ups

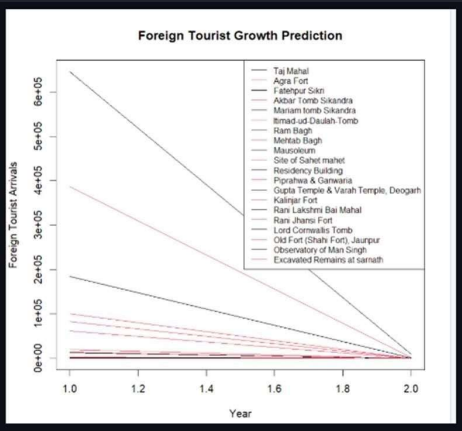


4. Forecasting for 2yrs foreign travellers using linear regression:

Foreign travels vs rows



Foreign travels vs year



5. Results:

5.1 Results:

The conclusions of the project should be based on the results of the data analysis. The conclusions should be clear, concise, and actionable. They should also be relevant to the needs of the stakeholders for whom the project is being conducted.

For example, if the project is being conducted for a tourism business, the conclusions may focus on identifying ways to attract more tourists or to increase the spending of tourists. If the project is being conducted for a policymaker, the conclusions may focus on ways to promote sustainable tourism or to mitigate the negative impacts of tourism.

The conclusions of the project should also be based on the limitations of the data and the analysis. For example, if the data is not representative of all tourists, then the conclusions should be limited accordingly.

Overall, the tourist behavior analysis project can provide valuable insights into the needs and wants of tourists. This information can be used to develop and promote products and services that meet those needs, and to develop policies that promote sustainable tourism and tourism development.

5.2 Conclusion:

The results of the tourist behavior analysis project can be used to better understand the needs and wants of tourists, to develop and promote products and services that meet those needs, and to develop policies that promote sustainable tourism and tourism development.

Some specific results that may be obtained from the project include:

Identification of the most popular tourist destinations and activities

Understanding of tourist demographics, travel motivations, and spending habits

Prediction of tourist behavior, such as the likelihood of a tourist visiting a particular destination or participating in a particular activity

Identification of trends in tourist behavior over time

References:

[1] D. -D. Lu and Y. -D. Zhong, "A tourist flows analysis system based on phone big data," 2016 IEEE International Conference on Big Data Analysis (ICBDA), Hangzhou, China, 2016, pp. 1-5, doi: 10.1109/ICBDA.2016.7509822.

[2] S. Arthan, K. Jandum and K. Tamee, "Exploring Tourist Behavior from Social Media Using Geotagged Photographs," 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, Cha-am, Thailand, 2021, pp. 285-288, doi: 10.1109/ECTIDAMTNCON51128.2021.9425761.

[3] N. Bunsaman, P. Sae-Ueng and K. Chochiang, "Analysis of the rrelationship of tourist behavior in Andaman Coast Provinces, Southern Thailand," 2021 25th International Computer Science and Engineering Conference (ICSEC), Chiang Rai, Thailand, 2021, pp. 57-62, doi: 10.1109/ICSEC53205.2021.9684582.

[4] A. Alamsyah, I. P. W. Ditya and T. Widarmanti, "Tourist Movement Analysis using Social Media Data in Indonesia," 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), Bali, Indonesia, 2021, pp. 1-6, doi: 10.1109/ICADEIS52521.2021.9701947.