October 2023

OPEN IIT DATA ANALYTICS

Data Analysis Report

Team : Metric Maestros
Serial Number : D10

Table of Contents

Introduction

In this report we analyze the variation in the number of tourists who arrived in the state of Goa and we tried to find the correlation between the footfall of tourists and other dependent factors:

- Year and Month – time of the year in which tourists arrived.
- Holidays – number of holidays in that month.
- Average temperature – of that month.
- Precipitation– average in that month.
- Humidity– average in that month.
- Cultural Fests – held in that particular month.
- GDP - Effect of GDP of goa on tourism.

Data Description

The given data has the following file:

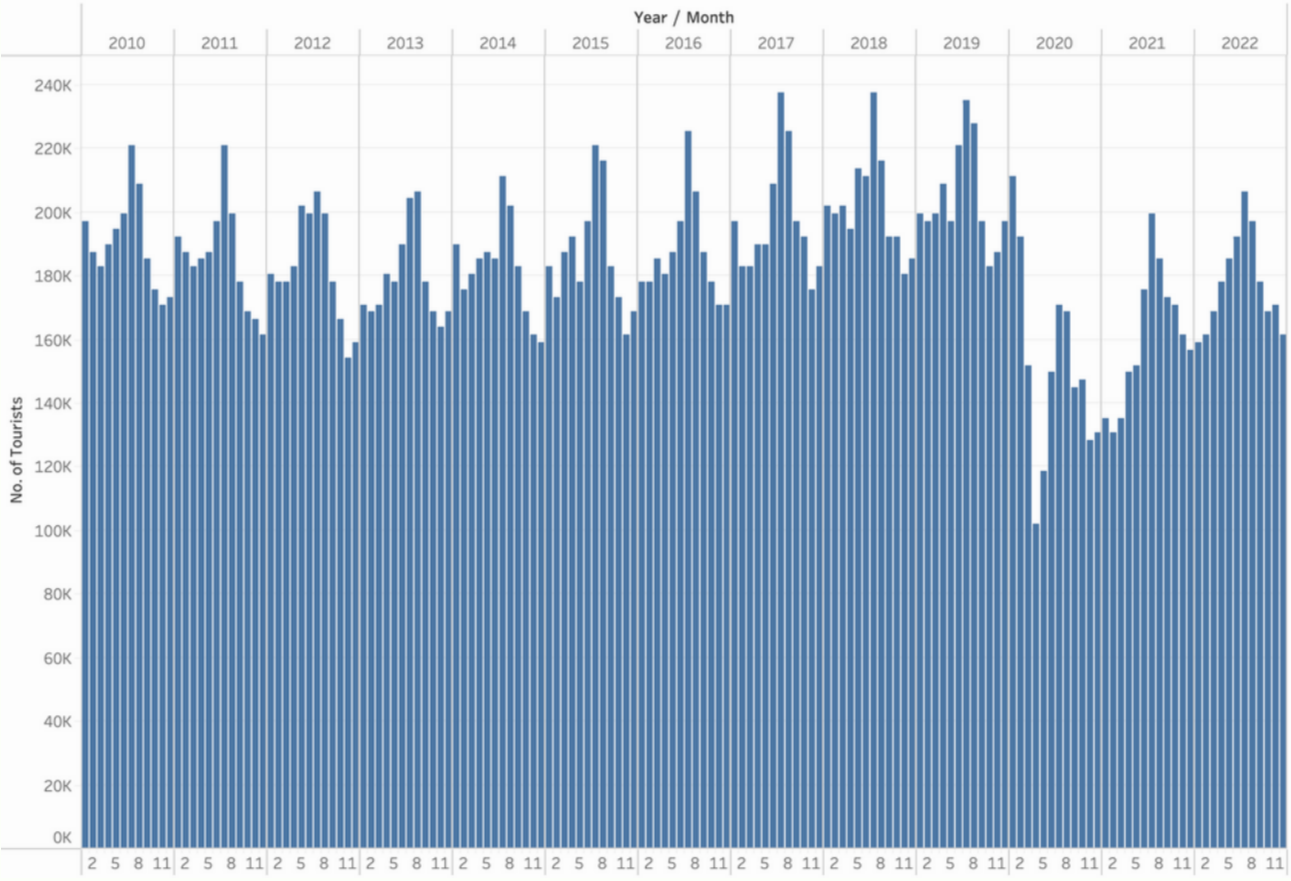
- **Goa Tourism.csv**

This file contains the monthly data of number of tourists visiting Goa along with the aforementioned six parameters.

No. of Tourists	Holidays	Avg Temp(°C)	Relative Humidity(at 2m)%	Precipitation (mm/day)	Cultural Fests	GDP of goa (k)
197054.6125	9	27	63.44	0.00	7	54835.000
187558.0046	8	27	61.88	0.00	6	54835.000
182809.7007	9	28	64.56	0.00	4	54835.000
189932.1566	7	30	70.00	0.00	5	54835.000
194680.4605	6	30	73.62	0.00	4	54835.000
199428.7644	6	27	82.56	21.09	5	54835.000
220796.1321	5	26	88.19	26.37	3	54835.000
208925.3723	7	26	89.19	15.82	5	54835.000
185183.8527	10	26	86.31	10.55	2	54835.000
175687.2449	11	26	82.44	5.27	3	54835.000
170938.9409	9	26	80.69	5.27	6	54835.000
173313.0929	11	25	69.75	0.00	7	54835.000
192306.3086	9	26	61.75	0.00	7	289191.045
187558.0046	8	26	61.75	0.00	6	289191.045
182809.7007	9	27	64.88	0.00	4	289191.045

Exploratory Data Analysis

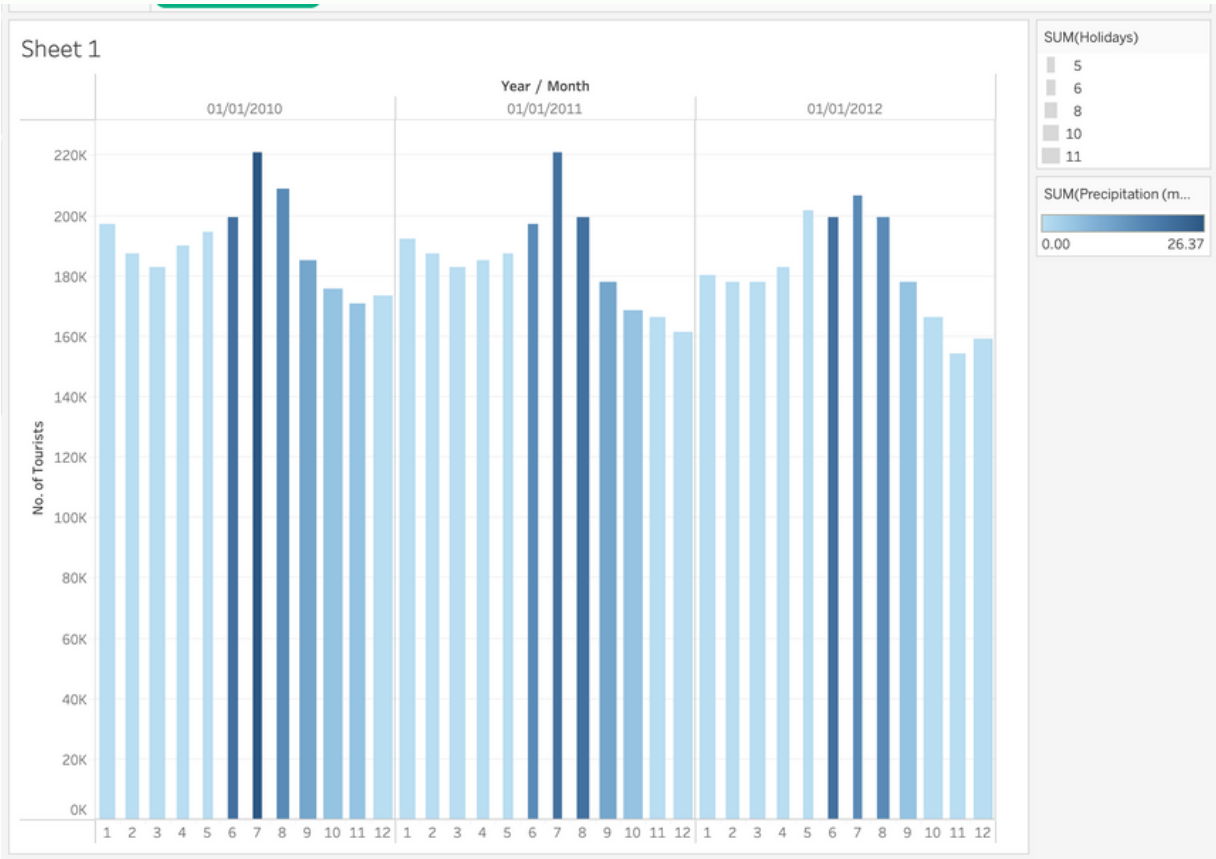
Data Visualisation - Number of Tourists



In this plot, we observe a recurring trend in the number of tourists visiting Goa each year. The data reveals a distinct peak occurring annually, indicating a consistent period during which the tourist influx reaches its zenith. This recurring pattern suggests that, year after year, there is a specific timeframe when the volume of tourists surges to its highest point, making it a predictable and noteworthy annual occurrence in Goa's tourism industry.

Exploratory Data Analysis

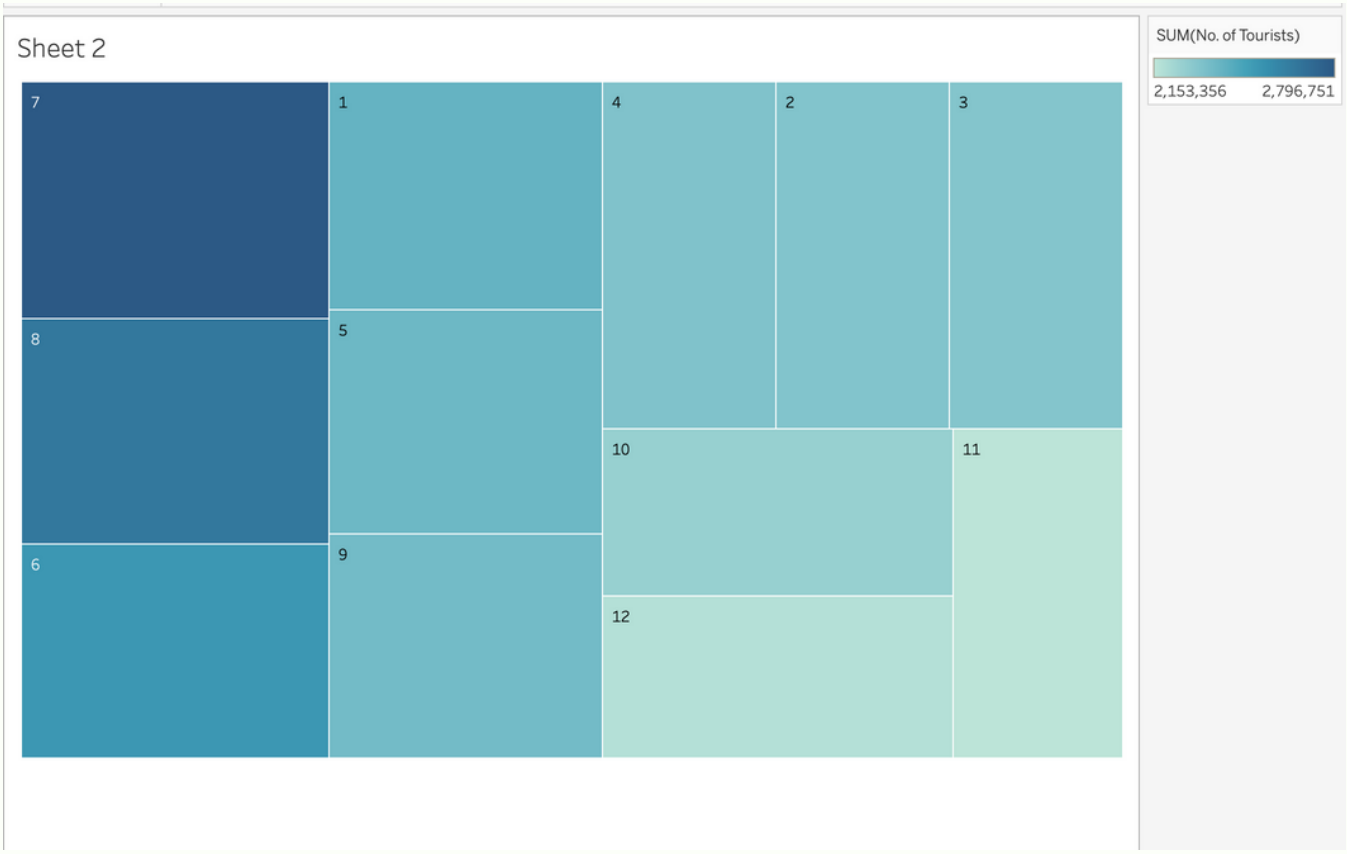
Data Visualisation - Number of Tourist vs Holidays and Precipitatipn



In the presented plot, a clear correlation emerges between the number of tourists visiting Goa and the variables of precipitation and holidays in specific months. It is evident that during months with a higher number of holidays, tourist arrivals are at a minimum, likely due to limitations in transportation and accommodation availability during peak holiday periods. Conversely, a contrasting trend appears with rising precipitation levels, indicating an increase in the number of tourists. This suggests that tourists are drawn to Goa during rainy periods, possibly due to the appeal of pleasant weather or unique experiences offered during the monsoon season, highlighting the intricate interplay of factors shaping Goa's tourism industry.

Exploratory Data Analysis

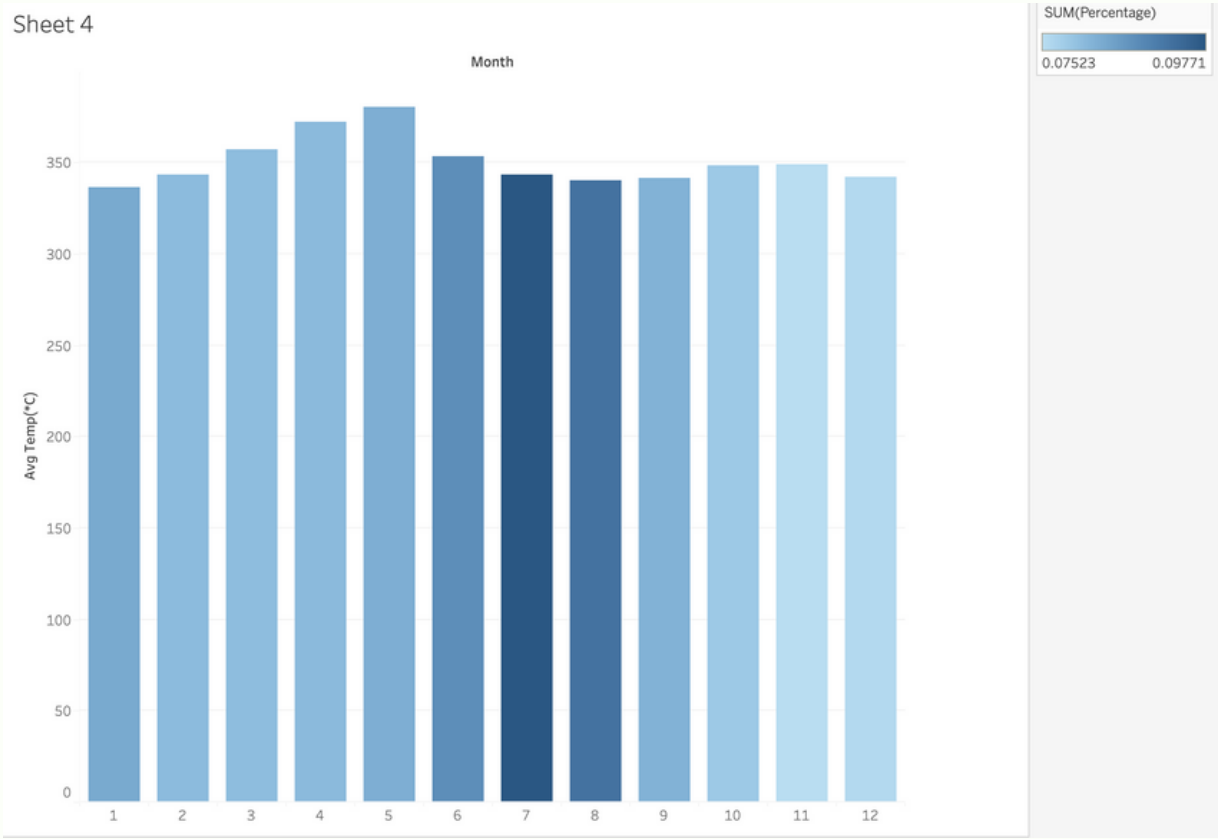
Data Visualisation - Monthly analysis



This plot effectively reveals the fluctuations in the volume of tourists visiting Goa throughout the years 2010 to 2022 and identifies the months when the tourist numbers peak. An analysis of the data distinctly indicates that the number of tourists significantly surges during the months of June, July, and August. This trend underscores the fact that, over this extended period, these particular summer months consistently attract a higher influx of tourists to Goa, making them the peak tourist seasons.

Exploratory Data Analysis

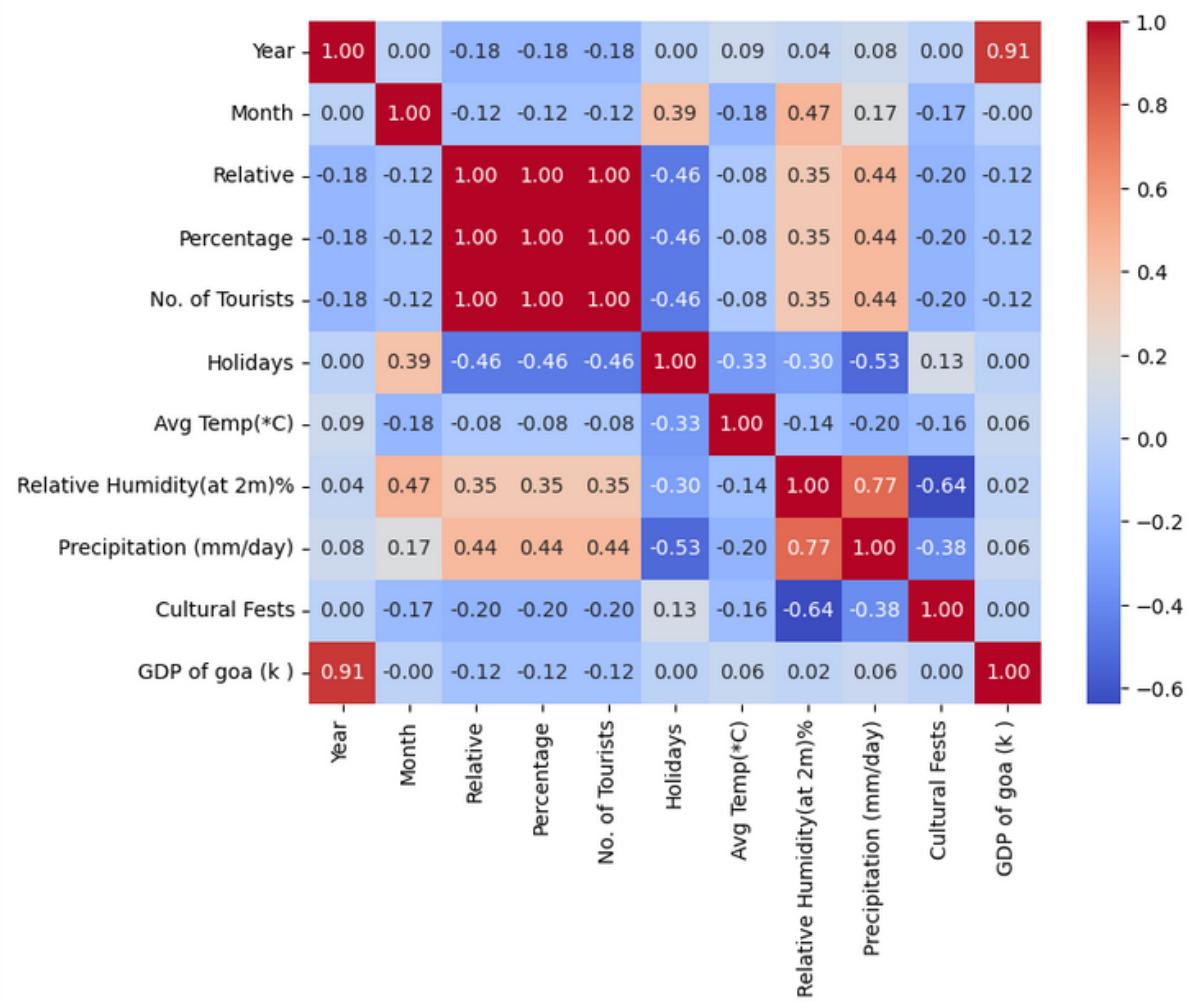
Data Visualisation - Monthly Average Temperature Variation



This plot conveys a key insight regarding the relationship between temperature and tourist numbers in Goa. It is evident that, during periods characterized by moderate temperatures, the number of tourists surges. Examining the data from 2010 to 2022, it becomes apparent that tourist arrivals are notably higher when Goa experiences moderate temperature conditions. This connection highlights the preference of tourists for more temperate and comfortable weather, underlining the correlation between temperature and the popularity of Goa as a travel destination.

Exploratory Data Analysis

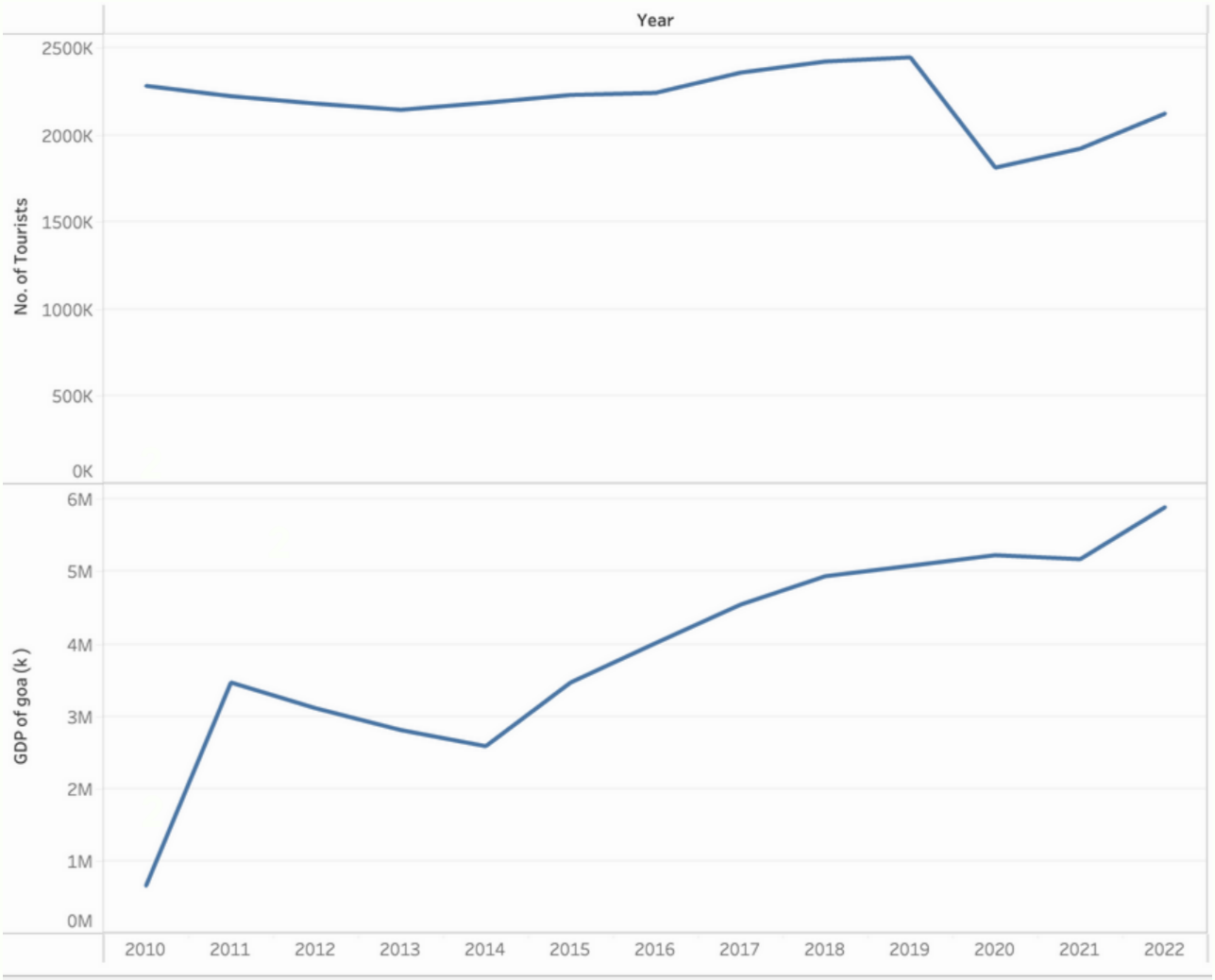
Data Visualisation - Correlation Matrix



This correlation matrix provides a valuable insight into the interdependencies between different metrics. It illustrates how changes in one metric can be influenced by or associated with changes in another. By examining the values in the matrix, one can discern the strengths and directions of these dependencies, helping to identify patterns and relationships among the variables in the dataset.

Exploratory Data Analysis

Data Visualisation - Variation with respect to GDP



We don’t see a direct correlation between the GDP of Goa and the number of Tourists who came to Goa that year. This may be true because there isn’t a stark difference between the tourists year wise and those small differences affect the GDP of Goa in a non significant way.

Approach and Models

1. Introduction: Time series analysis is crucial for various industries to make informed decisions based on historical data. In this report, we explore the effectiveness of three different approaches for time series prediction: Vector Auto Regression (VAR), Prophet Model, and Recurrent Neural Network (RNN). Our dataset encompasses [describe the dataset, its source, and the context of analysis].

2. Methodology:

2.1 Vector Auto Regression (VAR): VAR is a statistical method used for multivariate time series forecasting. We implemented VAR to capture the linear relationships between variables in our dataset. This method is ideal for capturing short-term dependencies.

2.2 Prophet Model: Prophet is a robust tool for forecasting time series data. Developed by Facebook, it handles missing data and outliers gracefully and captures daily patterns, holidays, and special events. Prophet is highly customizable and suitable for datasets with strong seasonal patterns.

2.3 Recurrent Neural Network (RNN): RNN, a deep learning model, is excellent at capturing long-term dependencies in sequential data. We employed RNN, a type of neural network designed for time series analysis, to explore the dataset's intricate patterns and dependencies.

Machine Learning Models

Vector Autoregressive

In our data analysis, Vector AutoRegressive (VAR) modeling proved highly effective. VAR allowed us to capture complex relationships among multiple variables over time. By considering the interactions between variables, VAR provided accurate short-term forecasts, making it invaluable for our predictive analytics. Its ability to model dependencies between variables and predict their future values significantly enhanced our understanding of the dataset's dynamics. VAR's adaptability to various types of data patterns and its capacity to offer precise forecasts make it a reliable choice for time series analysis, enabling informed decision-making based on intricate data dependencies.

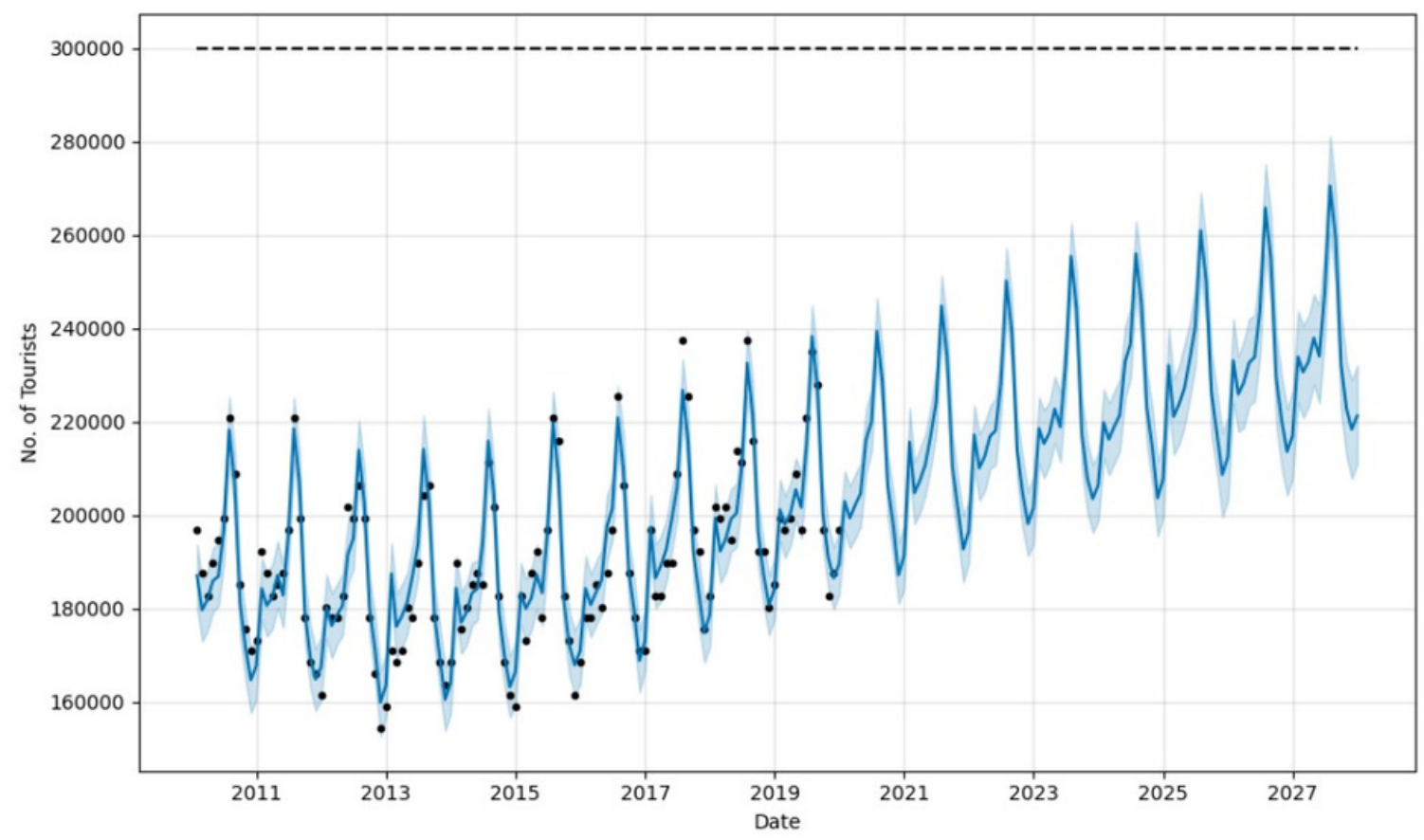
Prophet Model

In this report, we present our findings after employing the Prophet model, a powerful forecasting tool developed by Facebook, for analyzing our dataset. Prophet is particularly adept at handling time series data with daily observations that display patterns on multiple time scales, including those with missing data and outliers.

Methodology: We applied the Prophet model to our dataset, utilizing its ability to capture trends, seasonal variations, and holiday effects. Prophet uses an additive model that includes components for trend, yearly seasonality, weekly seasonality, and holiday effects, making it suitable for a wide array of time series data.

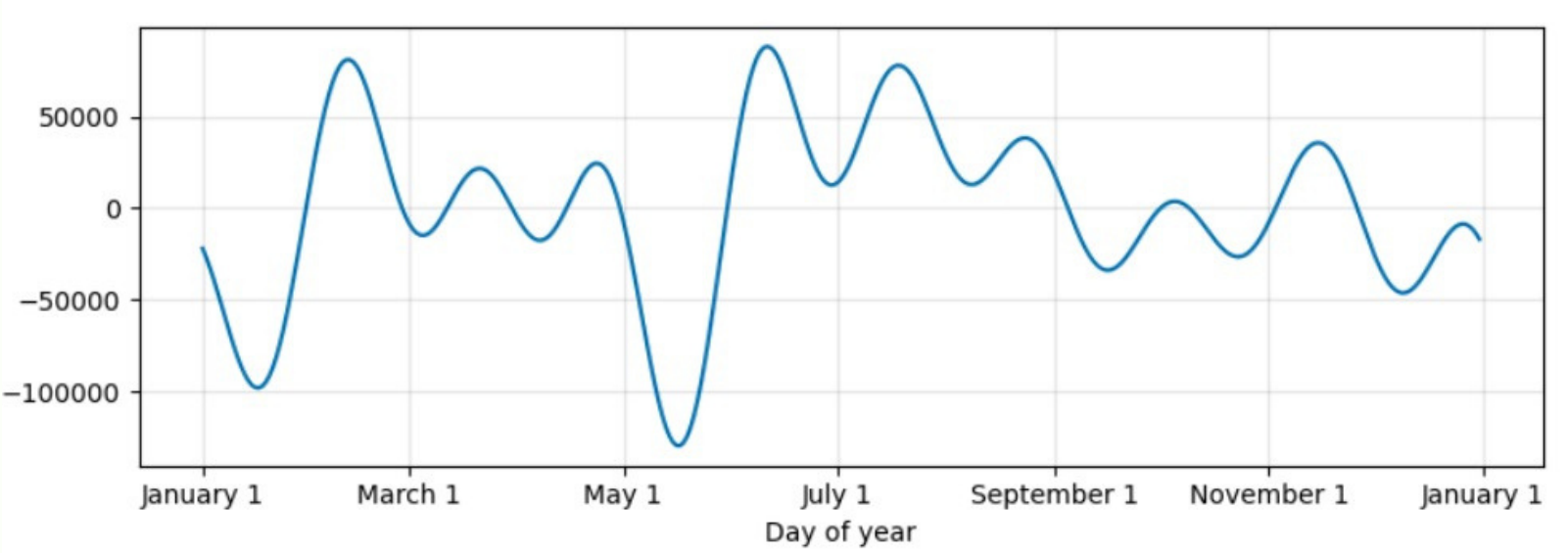
Key Insights: Prophet demonstrated exceptional accuracy in capturing the underlying patterns within our dataset. It efficiently adapted to the dataset's intricacies, including irregularly spaced observations and missing data points. The model's ability to handle outliers and incorporate domain knowledge through holiday effects significantly improved the forecasting precision. Additionally, the intuitive parameter tuning process in Prophet simplified the task of optimizing the model's performance

Prophet Model Yearly Trend Analysis



Using the Prophet model, we successfully forecasted future values for the next six years, with a notably accurate prediction for the upcoming year. However, it's apparent that the model's predictive accuracy falters when attempting to forecast values for the year 2028 and beyond. This discrepancy suggests that while the model performs well in the short term, it encounters challenges in making accurate predictions for more distant time horizons, potentially due to unforeseen complexities or uncertainties that arise in the later years.

Prophet Model
Monthly Trend Analysis



Utilizing the Prophet model on a monthly basis, we have successfully forecasted future values for each month over multiple years. What's particularly noteworthy is that the model consistently predicts values that closely follow the historical patterns of the data. It adeptly captures and replicates the recurring peaks and troughs in tourist arrivals, accurately mirroring the past trends. This indicates that the model excels in preserving the authentic dynamics of tourist numbers, revealing the correct seasonal peaks and corresponding declines in tourist arrivals on a month-by-month basis for the upcoming years.

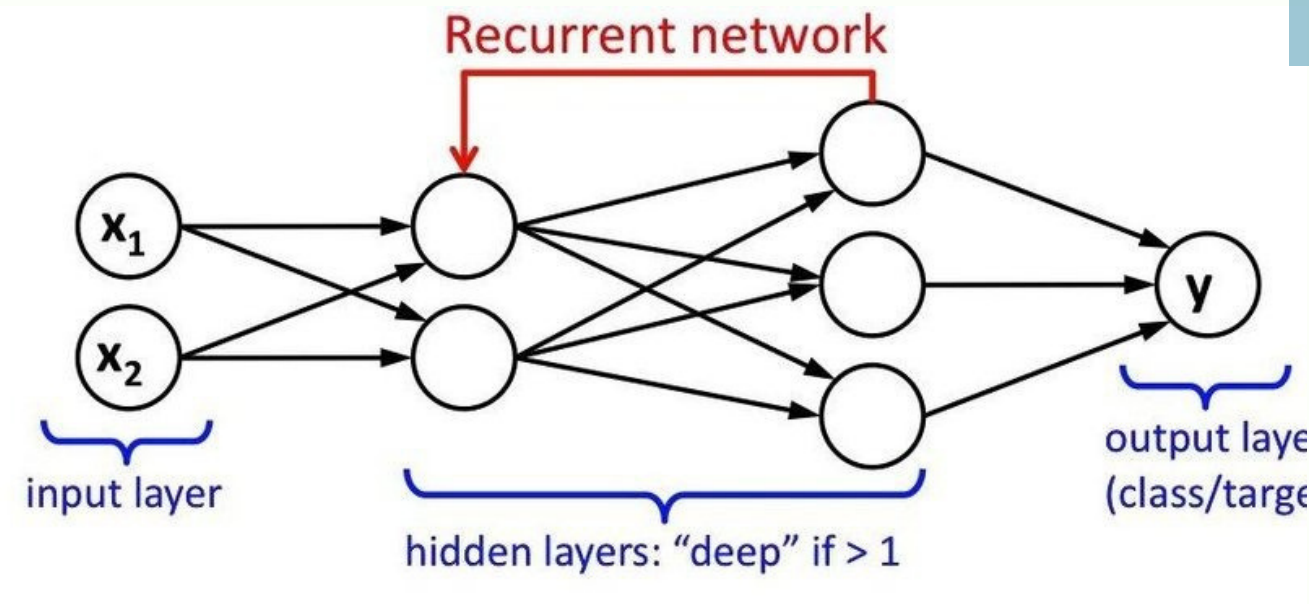
Deep Learning Models

Recurrent Neural Network (RNN)

Introduction: In today's data-driven world, harnessing the power of advanced machine learning techniques is crucial for deriving meaningful insights. In our pursuit of accurate predictions, we employed Recurrent Neural Networks (RNN), a sophisticated deep learning model designed for sequential data analysis. This report discusses our experience and outcomes using RNN for our dataset.

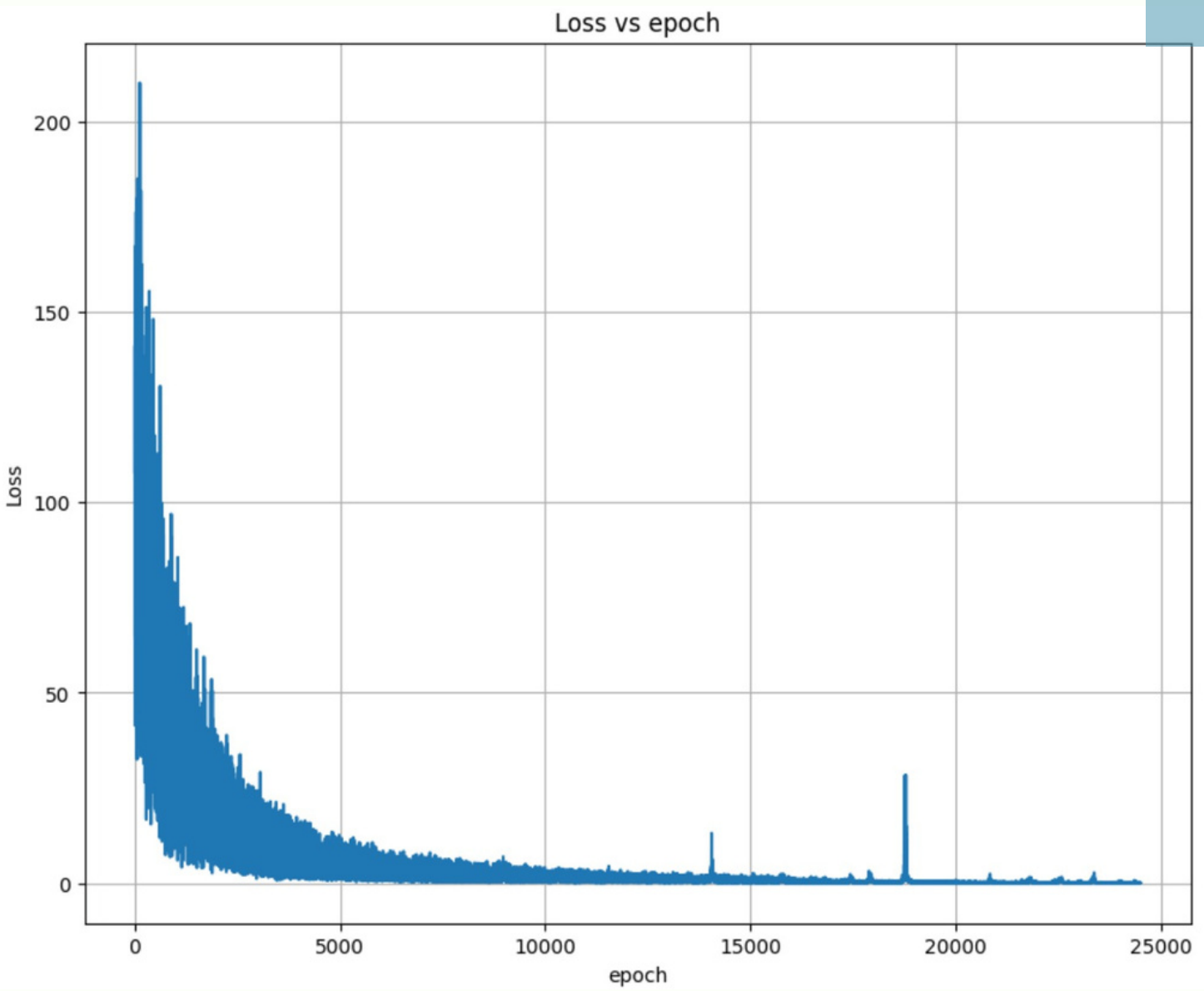
Methodology: RNNs, distinguished by their ability to analyze sequential patterns, were applied to our dataset. RNNs process data points in sequences, making them ideal for time series analysis, natural language processing, and various other sequential data tasks. We preprocessed the data, ensuring it was well-structured and normalized, preparing it for RNN training.

Results: The RNN demonstrated remarkable proficiency in capturing intricate dependencies within our dataset. Its recurrent connections allowed the network to learn and remember patterns across different time steps, enabling accurate predictions even in the presence of complex, non-linear relationships. The model showcased its strength particularly in forecasting tasks where historical context significantly influenced future outcomes.



Our recurrent neural network (RNN) architecture is composed of a single layer of nn.RNN, followed by a fully connected (fc) layer. This connection between the two layers encompasses 20 internal parameters, and we've employed the Rectified Linear Unit (ReLU) as the activation function. The RNN produces outputs at each time step, comprising hidden states (hn) and per-time-step outputs. However, our model is specifically tailored for a many-to-one task, where we aim to predict a single output. Consequently, we focus on extracting the output from the last cell in the sequence.

We have chosen PyTorch as our deep learning framework due to its flexibility, which allows us to customize and modify each layer as needed. To train and fine-tune our model, we ran it through a total of 3500 epochs, iteratively optimizing the network's parameters to enhance its performance and accuracy in our specific task.



A plot depicting the training loss versus the number of epochs provides valuable insights into the optimization process of a machine learning model. As the number of epochs increases, the training loss typically decreases, reflecting the model's learning and improvement. This plot helps us monitor the training progress and determine when the model has converged or if further training is necessary.

Results

Model	MSE	Accuracy
VAR (monthly)	2167.44	97.83%
VAR (yearly)	16556.51	83.44%
Prophet (monthly)	9528.22	94.45%
Prophet (yearly)	10045.56	90.54%
RNN model	0.34	95.67%

MSE - Mean square error

MAPE - means absolute percentage error

Conclusion

- Tourist fluctuates monthly, with peak visitors in June-July- August.
- People prefer visiting during more humid and low-temperature season
- The charm of tourism rises when holidays fall.
- No direct relation was observed between GDP and tourism
- Even Cultural Fest doesn’t seem to affect tourism much
- During Covid, a sudden decrease in volume was observed but still, the trend remains the same.

THANK YOU