Open IIT Data Analytics 2023-2024

Team D40

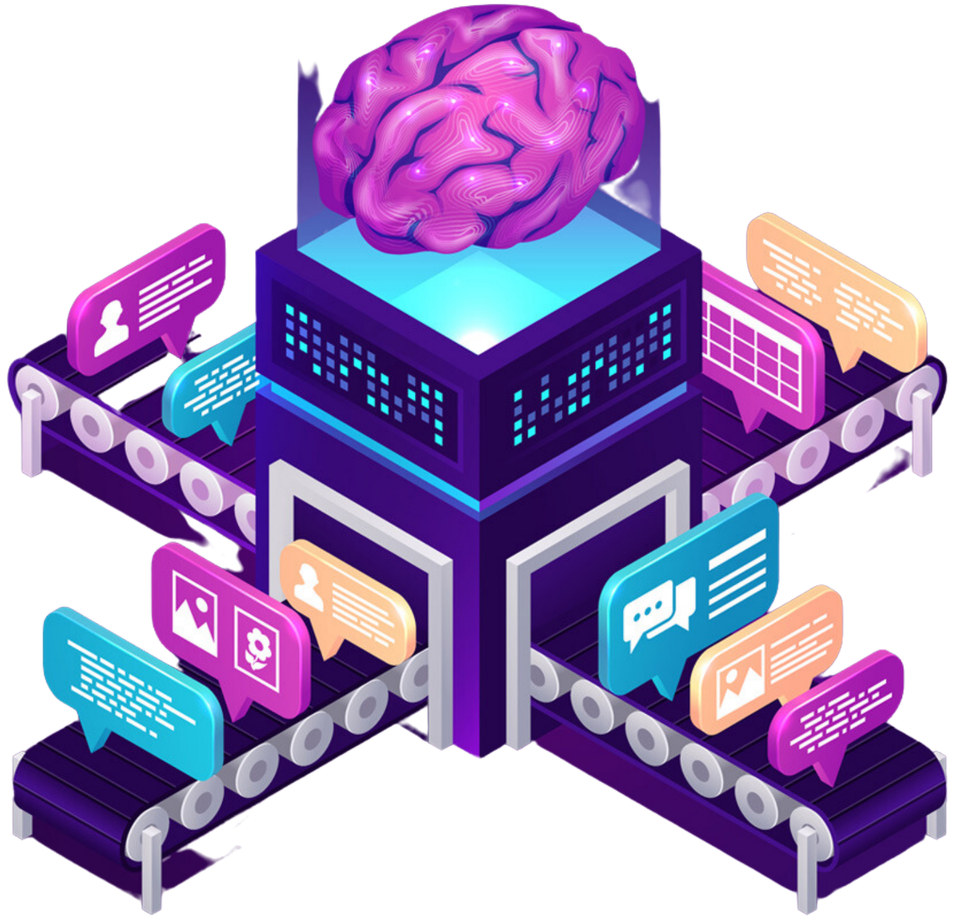


TABLE OF CONTENTS

01 INTRODUCTION

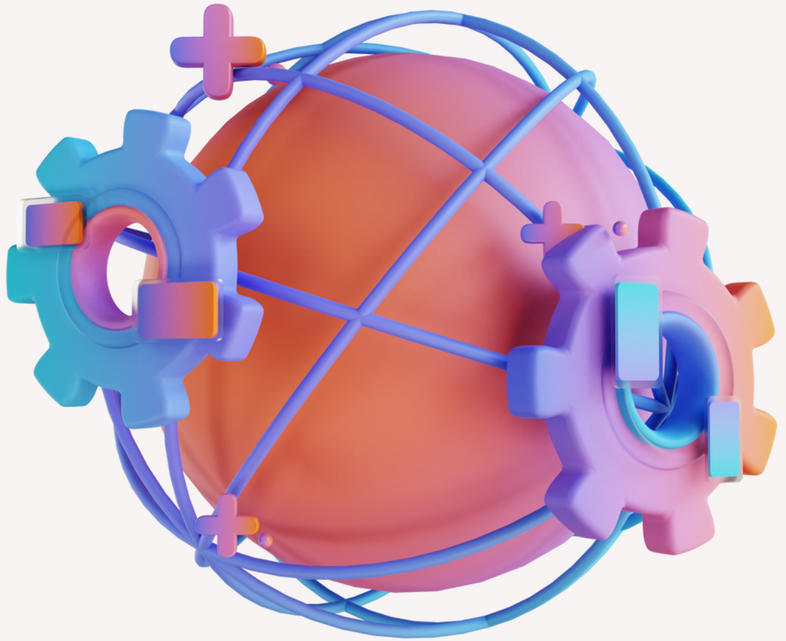
INTRODUCTION

1 **TASK:**
To create a Machine Learning Model that utilizes Internet search index data to predict tourist arrivals at Shimla and any other cities

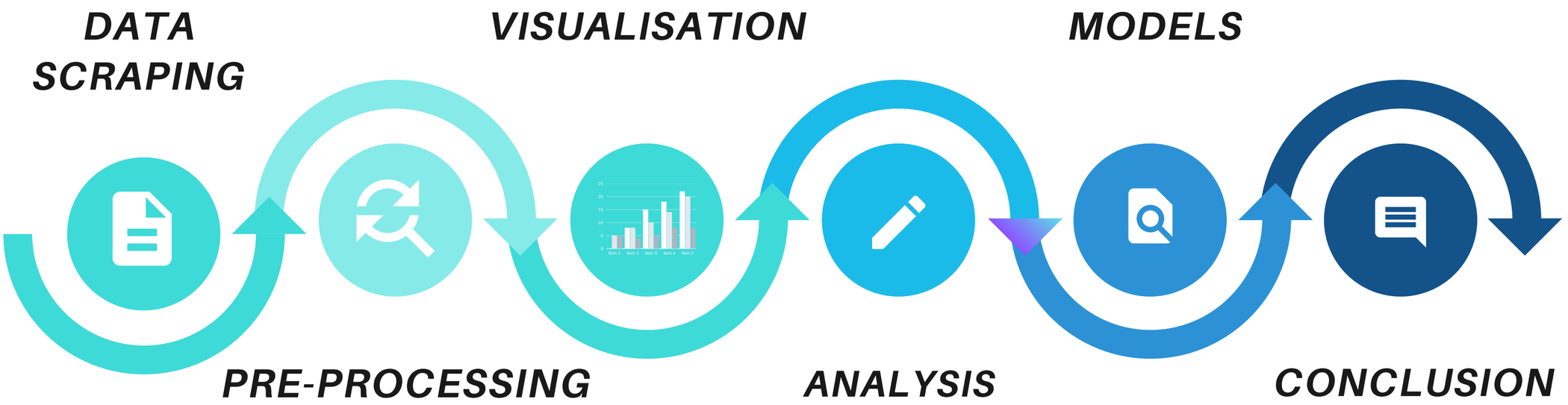
2 **FEATURES:**
Dataset features include Tourism, Traffic, Lodging, Dining, Recreation, and Shopping—comprising diverse elements crucial for comprehensive tourism analytics.

3 **TIME SERIES ANALYSIS:**
Time Series Analysis is a powerful statistical method dedicated to dissecting data points ordered chronologically. It unravels patterns, trends, and dependencies, enabling insightful predictions crucial for understanding and navigating dynamic temporal phenomena.

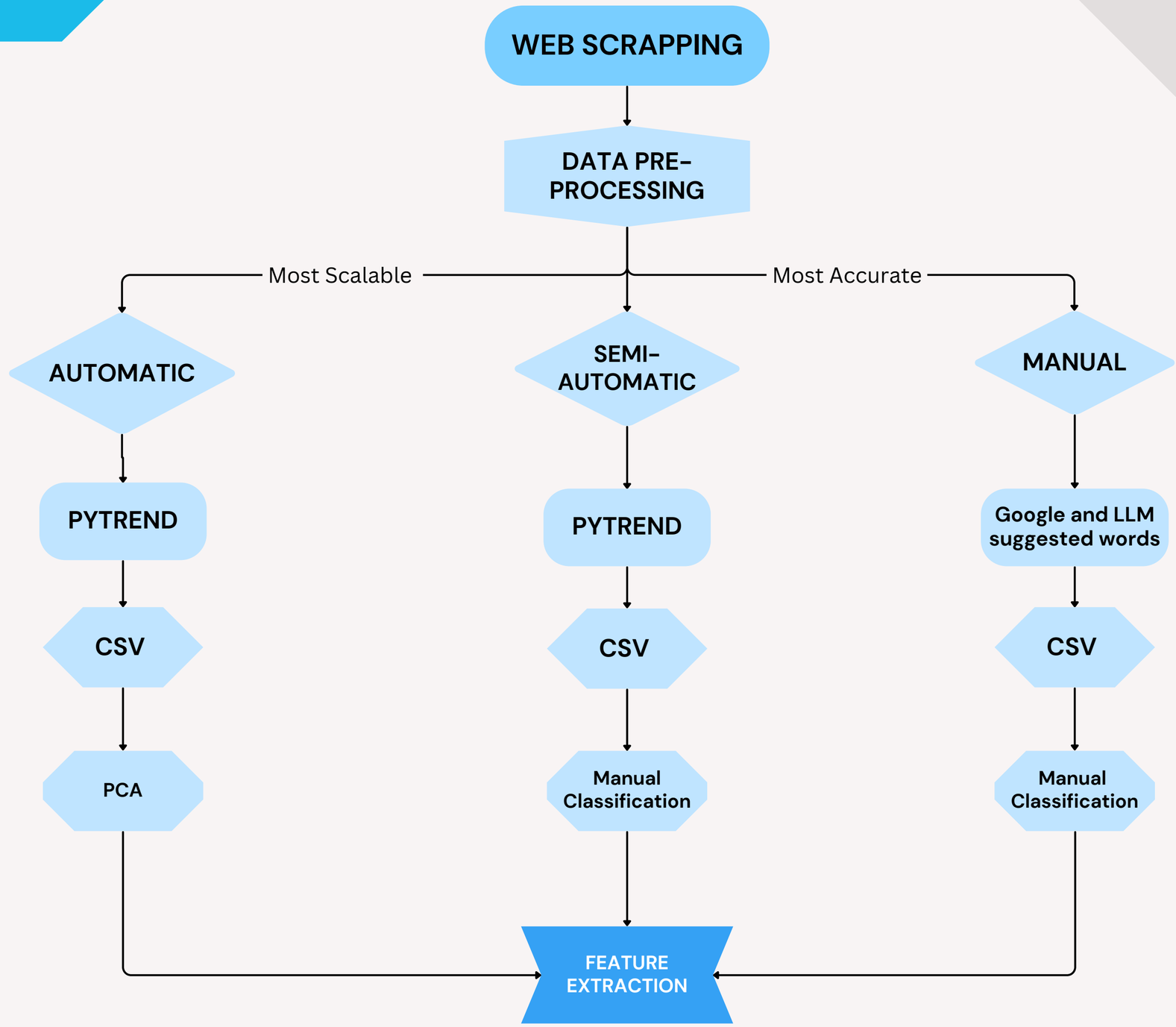
02 WORKFLOW

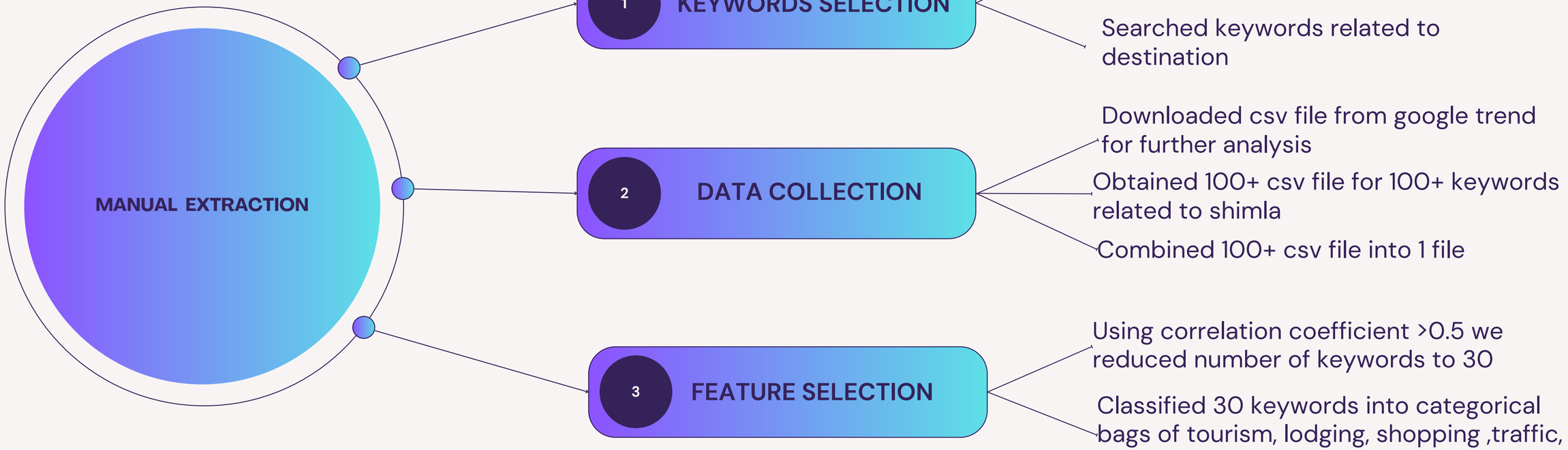


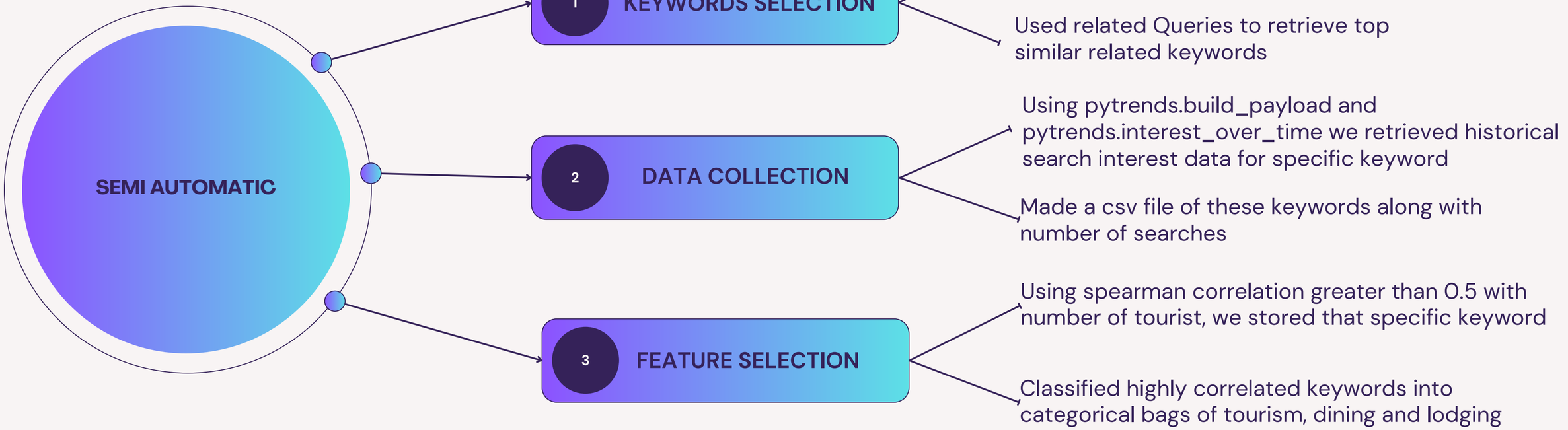
WORKFLOW

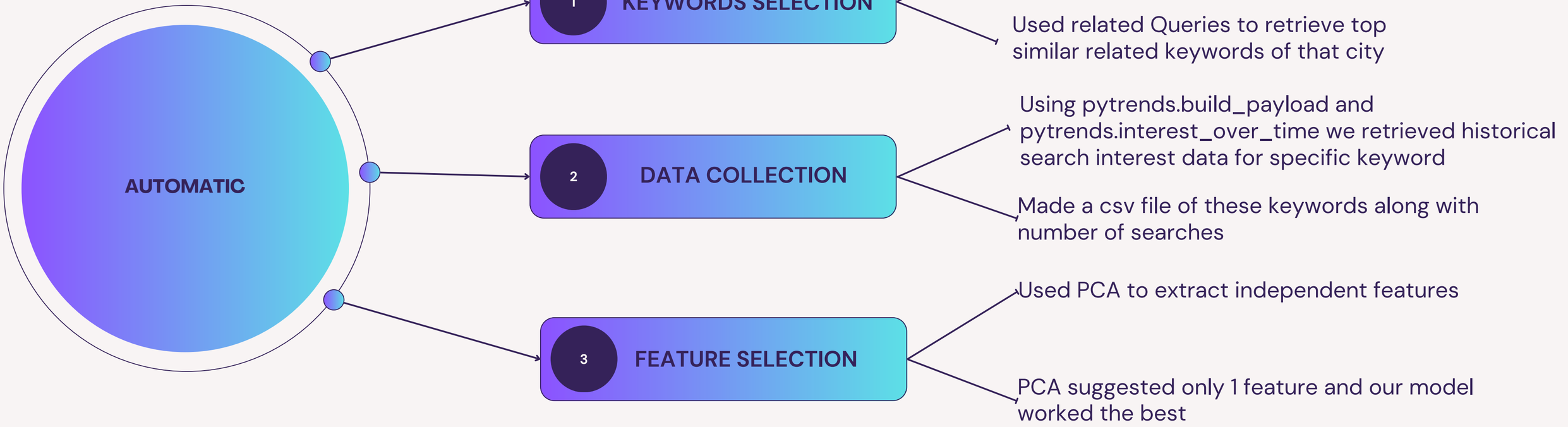

DATA SCRAPING
VISUALISATION
MODELS
PRE-PROCESSING
ANALYSIS
CONCLUSION

03 DATA PRE-PROCESSING

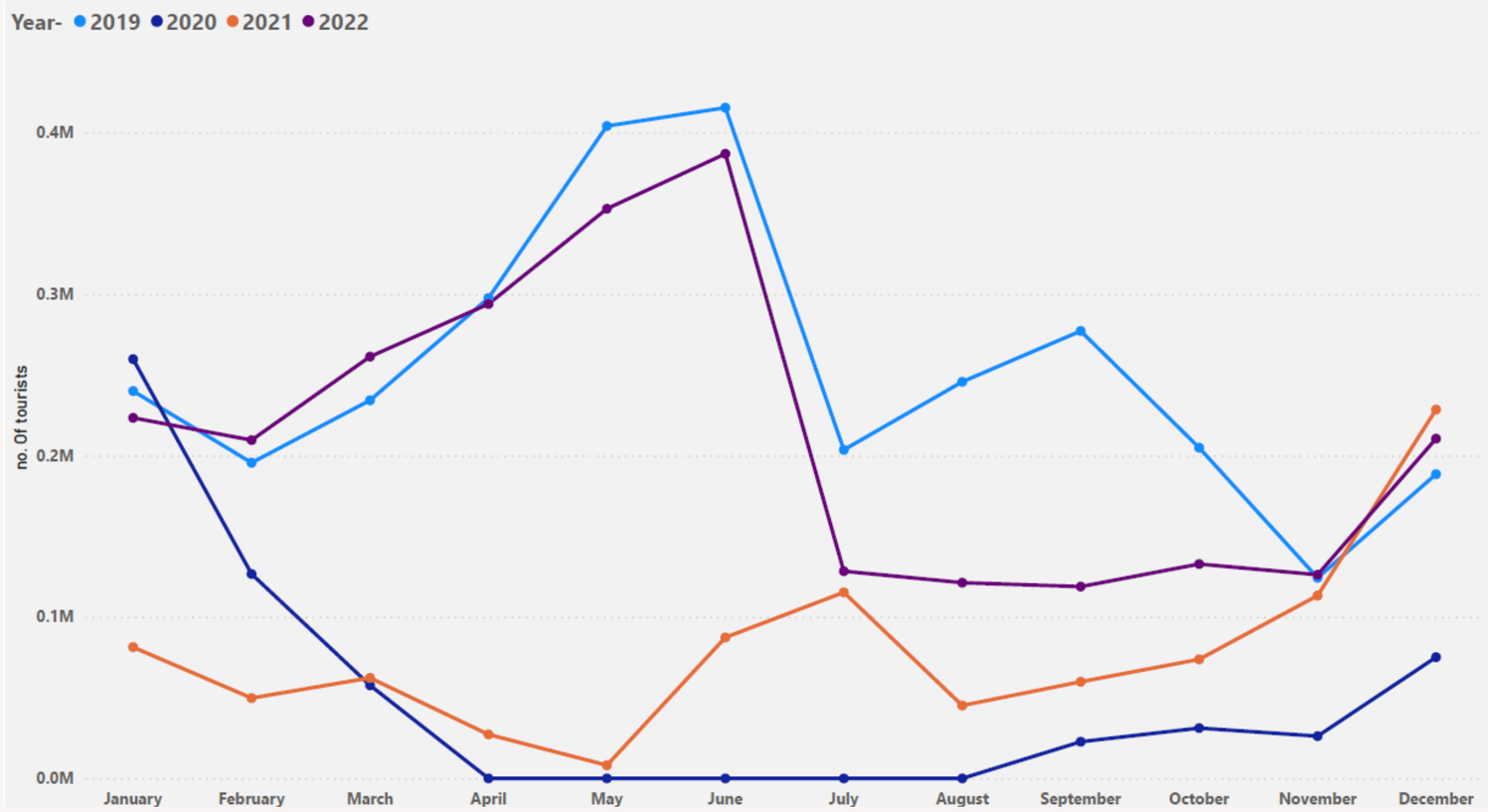

WEB SCRAPPING
DATA PRE-PROCESSING
Most Scalable
Most Accurate
AUTOMATIC
SEMI-AUTOMATIC
MANUAL
PYTREND
PYTREND
Google and LLM suggested words
CSV
CSV
CSV
PCA
Manual Classification
Manual Classification
FEATURE EXTRACTION


MANUAL EXTRACTION
1
KEYWORDS SELECTION
Visited the google trend websites.
Searched keywords related to destination
2
DATA COLLECTION
Downloaded csv file from google trend for further analysis
Obtained 100+ csv file for 100+ keywords related to shimla
Combined 100+ csv file into 1 file
3
FEATURE SELECTION
Using correlation coefficient >0.5 we reduced number of keywords to 30
Classified 30 keywords into categorical bags of tourism, lodging, shopping ,traffic, dining and recreation

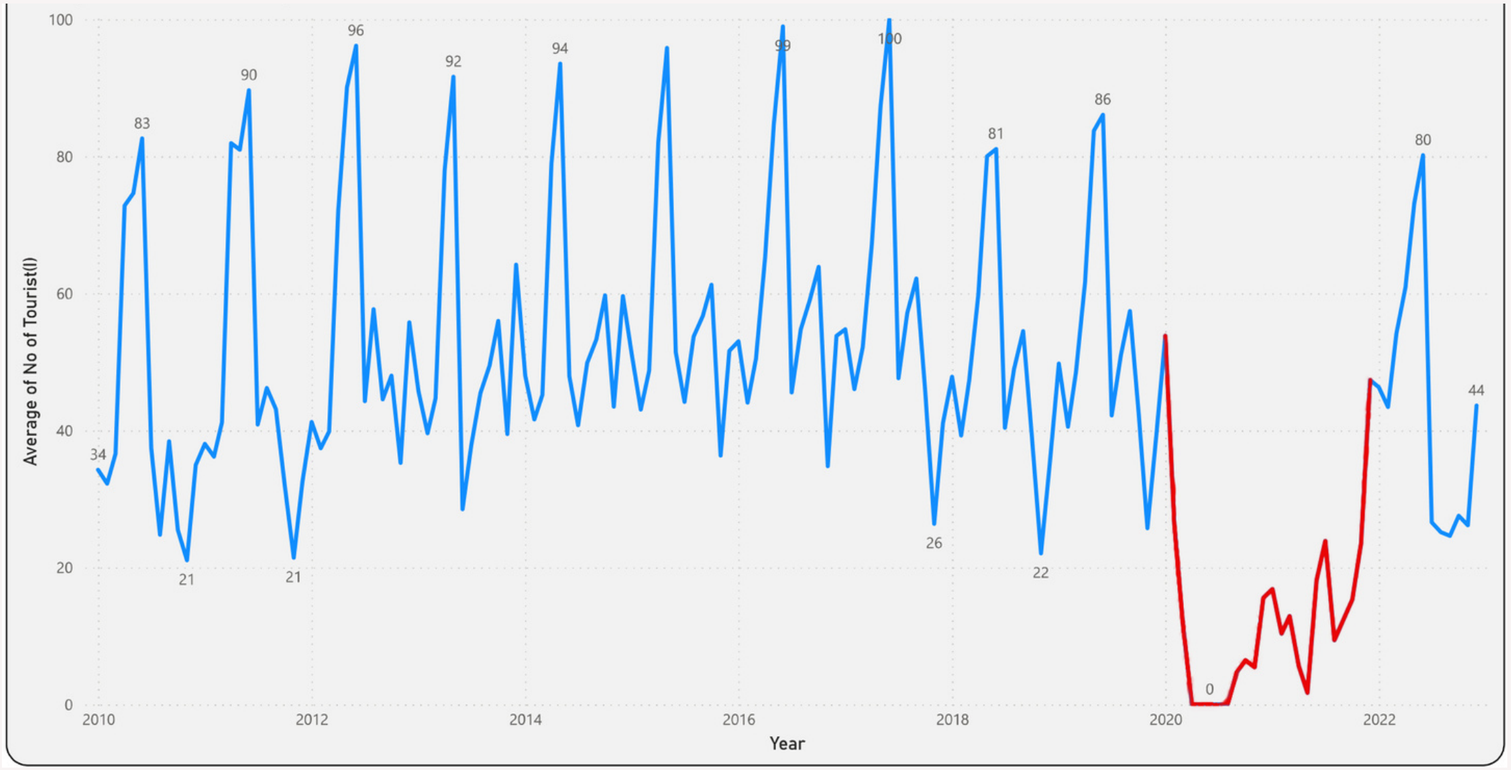

SEMI AUTOMATIC
1
KEYWORDS SELECTION
Interpreted google trends data using the pytrend library of python
Used related Queries to retrieve top similar related keywords
2
DATA COLLECTION
Using pytrends.build_payload and pytrends.interest_over_time we retrieved historical search interest data for specific keyword
Made a csv file of these keywords along with number of searches
3
FEATURE SELECTION
Using spearman correlation greater than 0.5 with number of tourist, we stored that specific keyword
Classified highly correlated keywords into categorical bags of tourism, dining and lodging


AUTOMATIC
1
KEYWORDS SELECTION
Added flexibility to enter name of any city of interest to predict number of tourist
Used related Queries to retrieve top similar related keywords of that city
2
DATA COLLECTION
Using pytrends.build_payload and pytrends.interest_over_time we retrieved historical search interest data for specific keyword
Made a csv file of these keywords along with number of searches
3
FEATURE SELECTION
Used PCA to extract independent features
PCA suggested only 1 feature and our model worked the best

04 EXPLORATORY DATA ANALYSIS



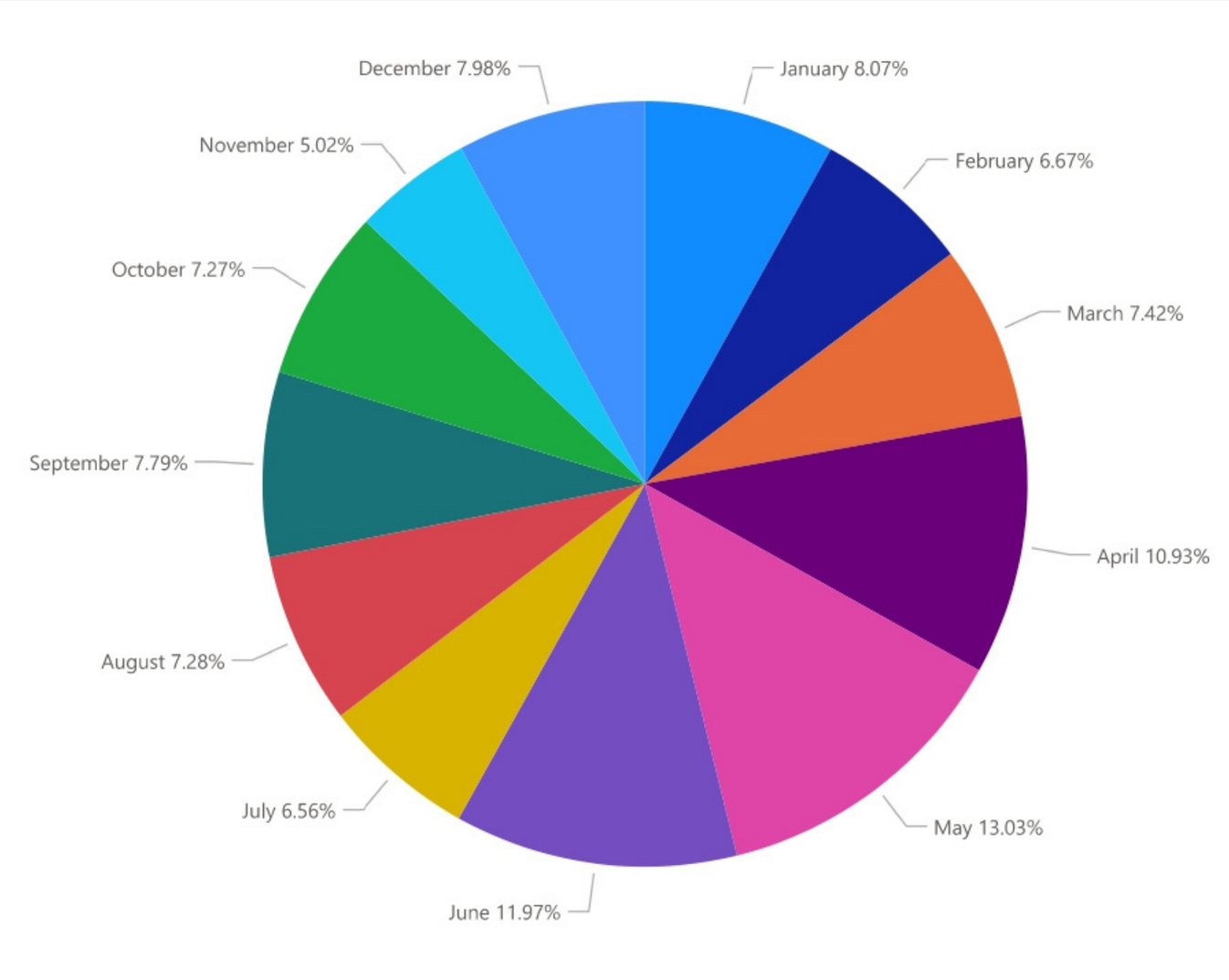
- Plot of number of tourists visits every month for the years 2019,2020,2021,2022.
- We can observe from the plot that the graph follows a similar trend for every year except 2020 and 2021 as covid spread during both years.



- Plot of number of tourists visits every month for the years 2010 to 2022

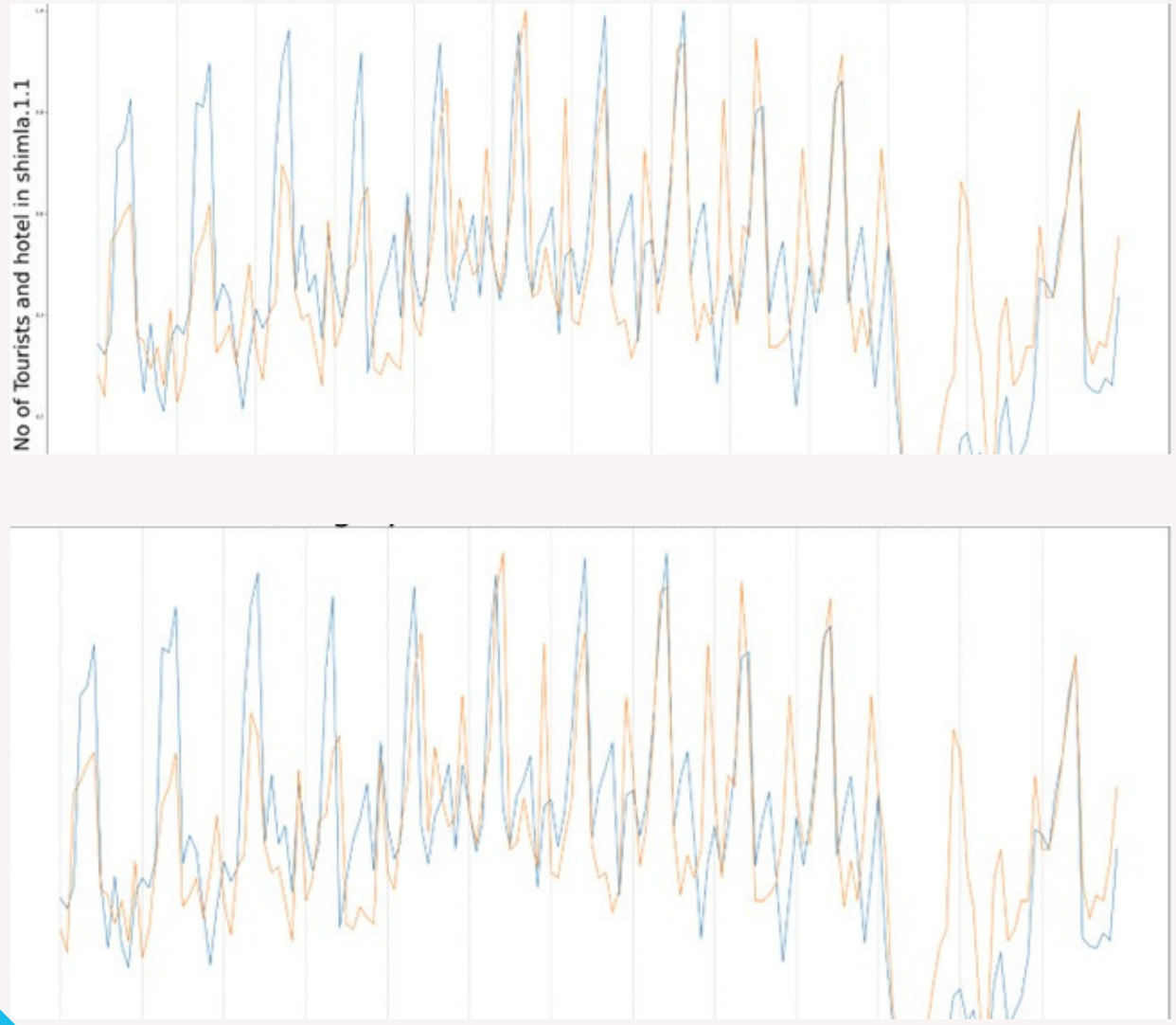
- We can see abnormally low no. of tourists visit for year 2020 and 2021 with respect to other years. So we can infer that 2020 and 2021 are outliers.

Average Number of tourist(s) per Month



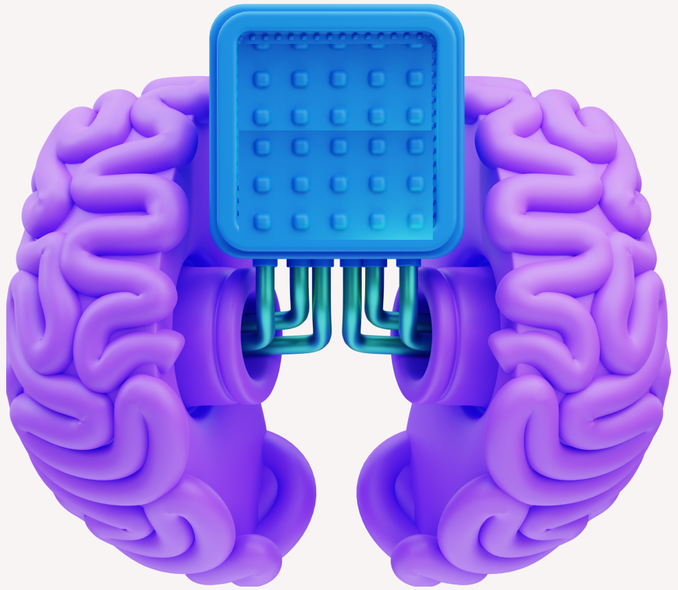
- Pie chart for average number of visits by tourists each month for years 2010-2022
- It can be observed from the pie chart that relatively more number of tourists visit shimla in months of april, may and june due to its cool climate.

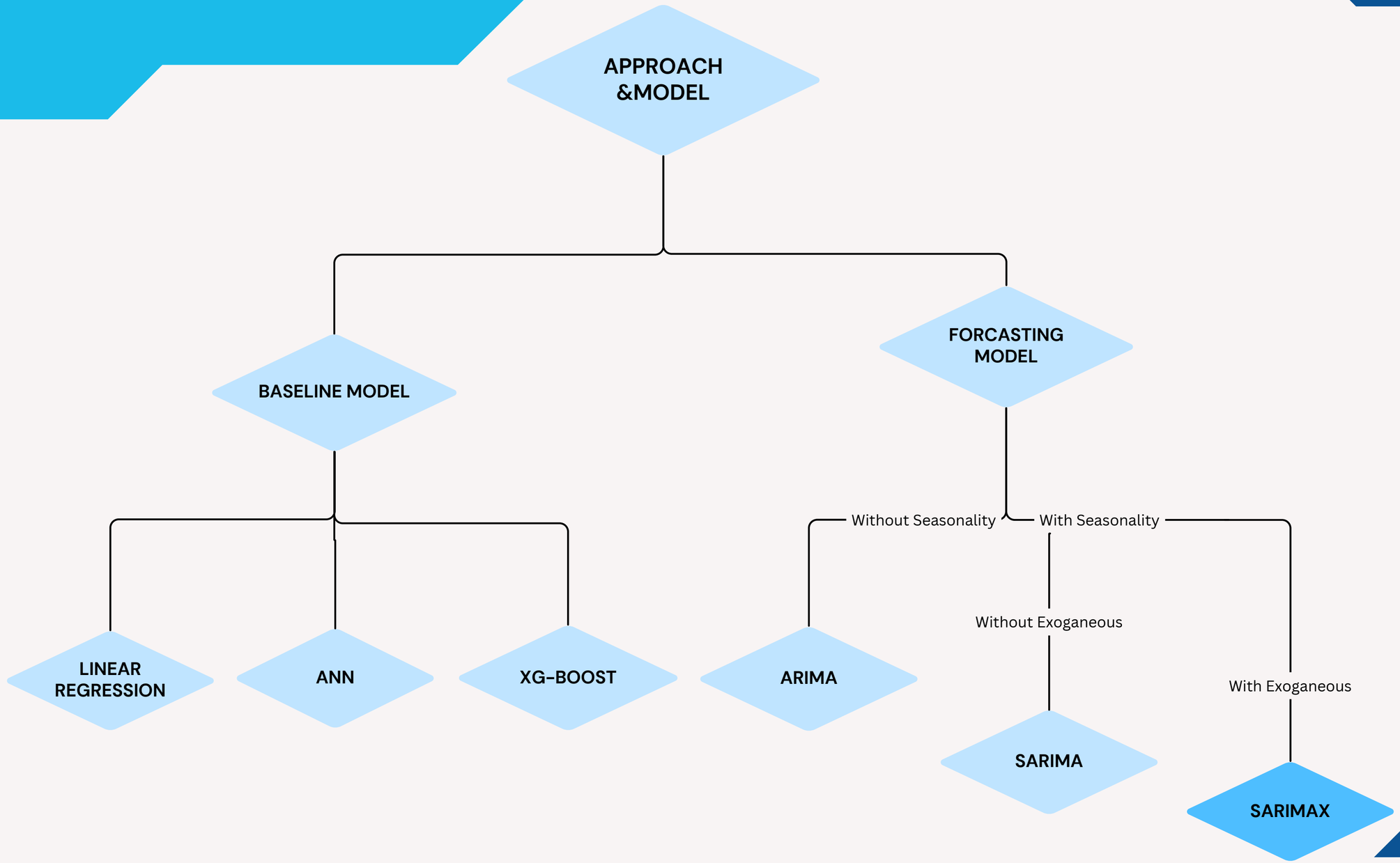
Number of Searches vs Total Number of Tourists



- It is a line chart of total number of tourists vs number of searches of words in Google Trends.
- It can be observed from the line chart that the number of search volume is proportional to the number of tourists as we can see that both lines coincides approximately.

05 APPROACH AND MODELS




APPROACH &MODEL
BASELINE MODEL
FORCASTING MODEL
LINEAR REGRESSION
ANN
XG-BOOST
Without Seasonality
With Seasonality
ARIMA
Without Exoganeous
With Exoganeous
SARIMA
SARIMAX

BASELINE MODEL

Linear Regression

- In every modeling process, there needs to be a baseline model whose results can be used to assess our primary ones. In our case, we chose to use a Linear Regression model because of its simplicity and efficiency.
- We have guessed that forecasting this time period would be very challenging due to the holiday season.

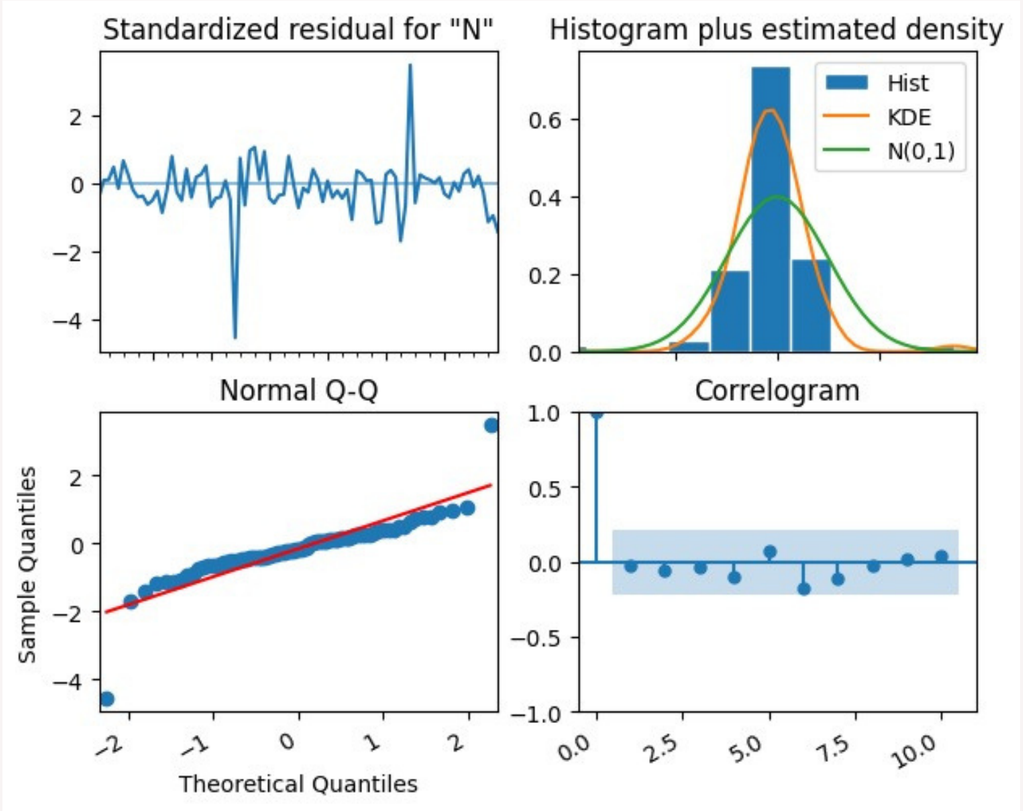
XGBoost

- XGBoost is a fast implementation of a gradient boosted tree. It has obtained good results in many domains including time series forecasting.
- It provides feature importance scores, which can help you understand the contributions of different features to the forecasts.
- It is capable of capturing complex relationships, including seasonality and trends in the data

FORECASTING MODEL

ARIMA

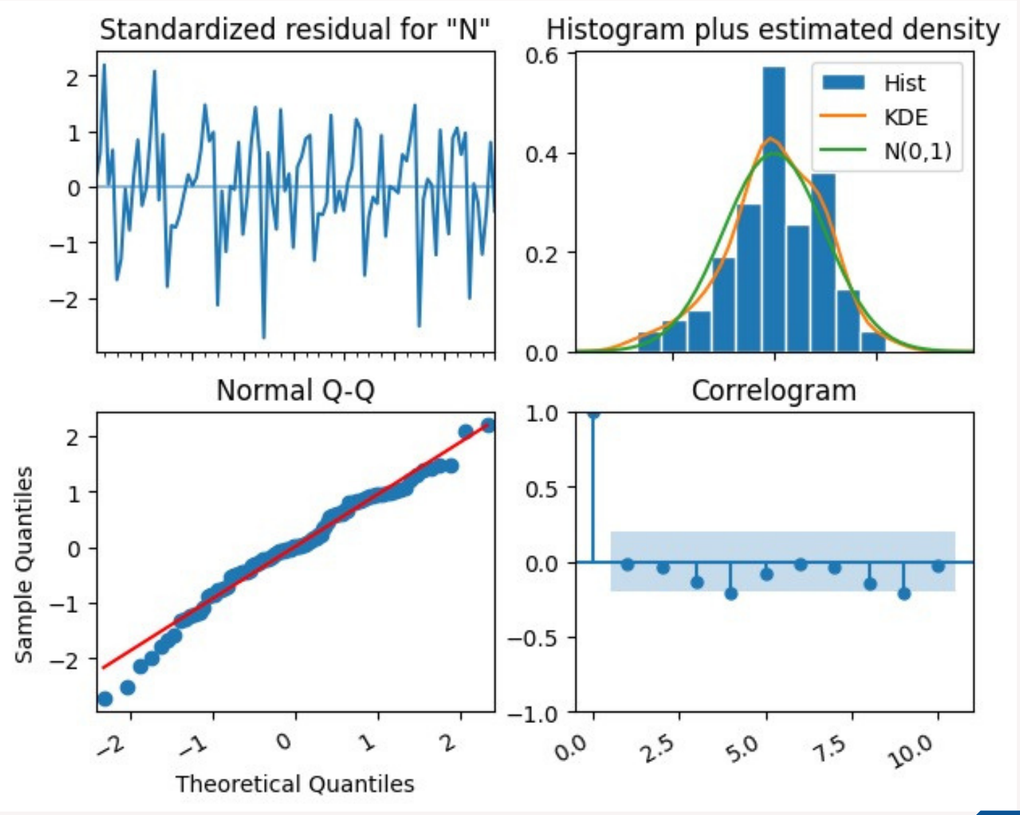
- Although our data is almost certainly not stationary (p-value = 0.34).We applied standard ARIMA model .
- Using the auto_arima() function from the pmdarima package, we can perform a parameter search for the optimal values of the model.
- As we can see from the plot below, this doesn’t seem to be a very accurate forecast.



FORECASTING MODEL

SARIMA

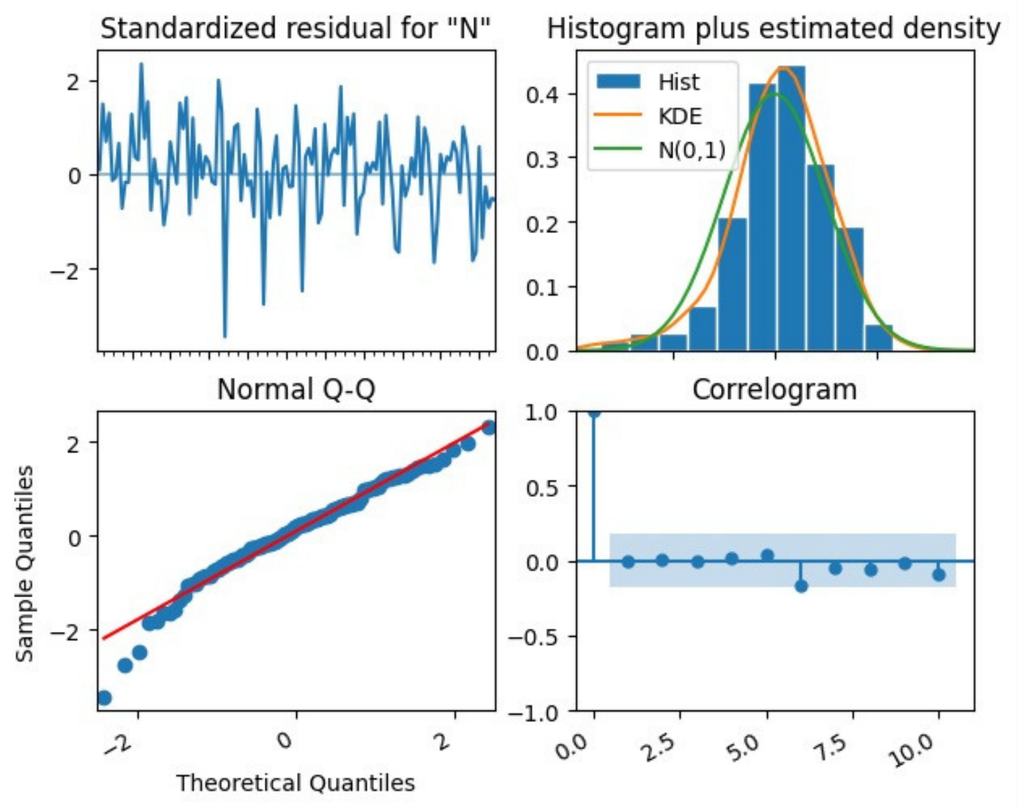
- A problem with ARIMA is that it does not support seasonal data. So we use an alternative model i.e. SARIMA.
- It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.
- As we can see from the plot , this seems to be much more accurate than the standard ARIMA model.



FORECASTING MODEL

SARIMAX

- Now we added the month number as an exogenous variable to further increase its performance .
- We can see from the following predictions that we are getting some pretty good-looking predictions and the width of the forecasted confidence interval has decreased. This means that the model is more certain of its predictions.

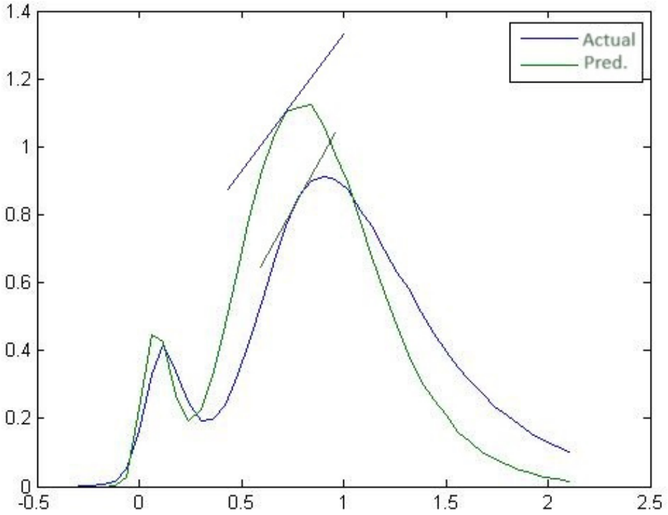


HYPERPARAMETER TUNING

1 **Analysing behaviour of model output:**

$$\text{Similarity Metric } \mathbf{e} = (\text{Slope of Actual Curve}) \times (\text{Slope of Predicted Curve})$$

$$\mathbf{e} \geq 0 \rightarrow \text{Curves have similar behaviour}$$

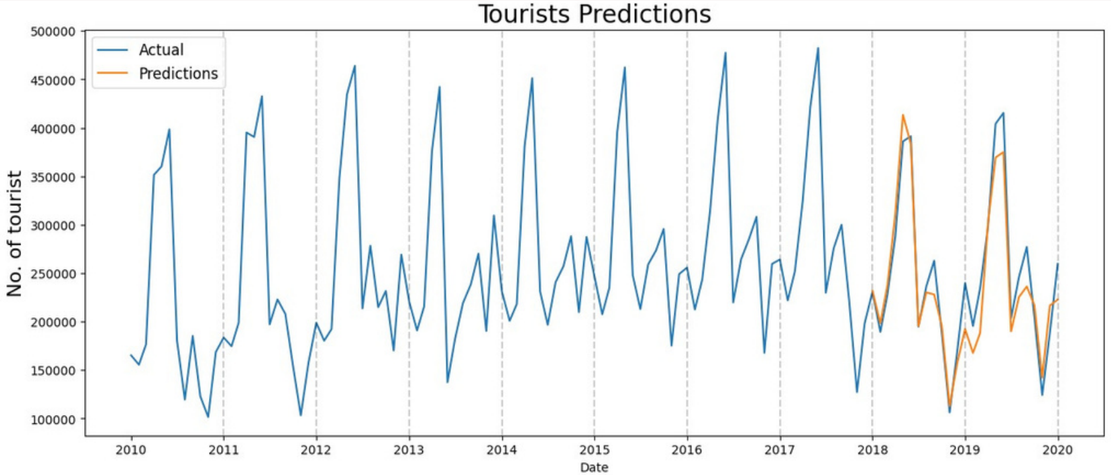


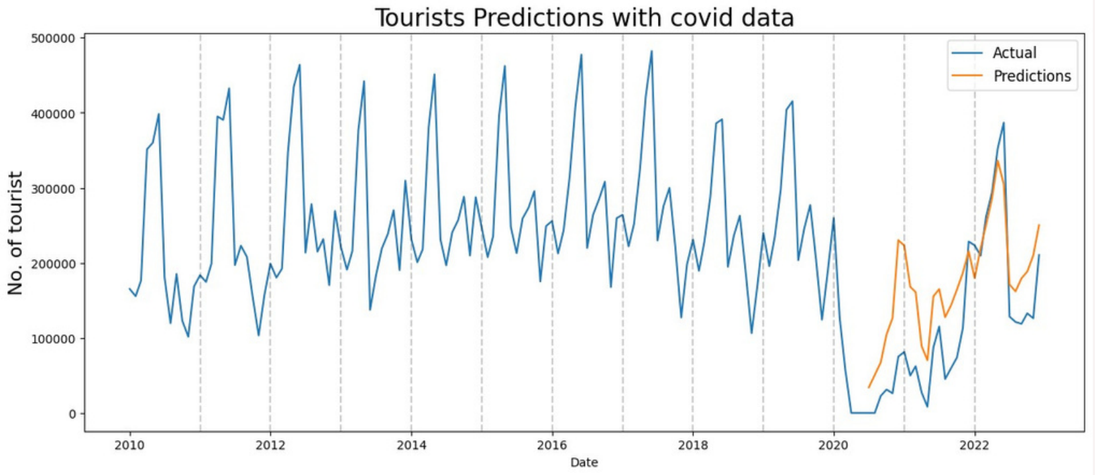
2 **Overall behaviour of predicted curve:**

- The hyperparameters were tuned using grid search
- Models parameters with highest percentage match were chosen

$$\text{Percentage Matched} = \frac{\text{Number of months where trend matched}}{\text{Total Number of Months}}$$

IMPACT OF COVID-19





Without including the covid data (i.e. 2020-2021), our model predicted the number of footfalls closer to the actual number of footfalls as the pandemic was unprecedented and could not be predicted.

Although the footfall matched as the search indices also decreased during the covid years which affect the predictions

MODEL BENCHMARKING

MODEL	RMSE	MODEL PERFORMANCE
Linear Regression	113425.62	51%
XG Boost	166695	31.2%
Catboost	88722	56.67%
LightGBM	126723	56.67%
MLP Regression	148045.38	60%
Arima	85079.39	71.875 %
Sarima	99051.99	89.28%
Sarimax	24927.41	91.66%

06 FINAL APPROACH

FINAL APPROACH

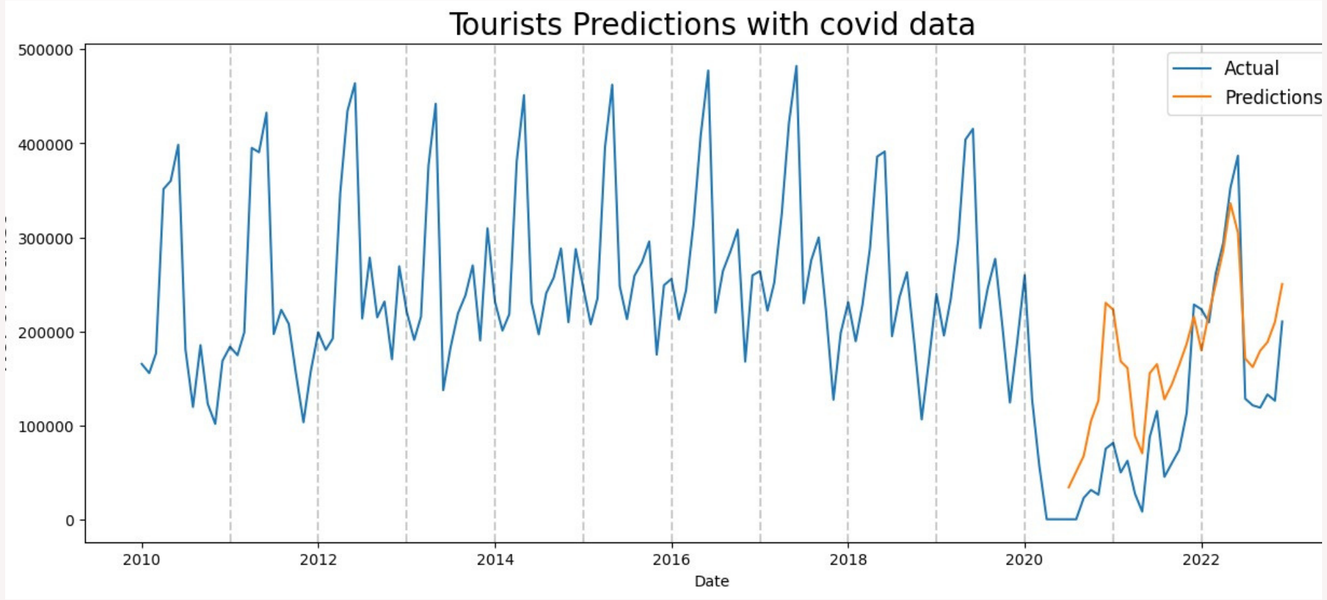
From benchmarking of various models, the model selected was **SARIMAX** with hyperparameters **(p=3,d=1,q=1)** and **(P=1,D=1,Q=1,M=12)**.

AUTOMATIC → is the most scalable

MANUAL → is the most accurate

With the removal of Covid data as outliers all models performed better.

SARIMAX MODEL PREDICTION



This plot shows no. of tourist per month and their predicted value

As we noticed that the actual values had sudden dip that is why the predicted value couldn't follow the actual values for the covid period.

Thank you