**Tourist-Behaviour-Analysis**

**Analysis and Prediction of future tourism demands**

**Student Name**

**Student ID**

Table of Contents

[1 Introduction 2](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094346)

[2 Methodology 3](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094347)

[3 Work Tutorial 4](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094348)

3.1 Requirements………………………………………………………………………….4

3.2 Setup Instructions…………………………………………………………………….. 4

3.3 Dataset………………………………………………………………………………... 4

3.4 Taks Folder…………………………………………………………………………… 5

3.5 Implementation………………………………………………………………………. 5

[4 Result/Outcome…………………………………………………………………………. 5](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094349)

4.1 Challenges Faced……………………………………………………………………. 12

4.2 Limitations and Future work……………………………………………………….. 13

[5 Conclusion](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094354) 14

[6 References](file:///C:\Users\HP\Downloads\Readme.docx#_Toc140094355) 15

# Introduction

**Problem**

* The tourism sector is quite dynamic and is impacted by a number of variables, including seasonality, the state of the economy, and cultural events.
* Effective tourism planning, resource allocation, and marketing strategies depend on an understanding of visitor behavior patterns and the ability to anticipate future trends in visitor arrivals.
* This project's goal is to evaluate a dataset of visitor arrivals in order to pinpoint important trends, seasonal patterns, and variables affecting visitor behavior.
* The project's objectives are to analyze the tourist behavior, to identify tourists’ interests and most visited locations and accordingly, and to predict future tourism demands by using the regression models that can precisely forecast visitor arrivals based on historical data.

**Introduction:**

**Solution:**

The Tourist-activity-Analysis project is an in-depth study that uses big data analytics to explore and identify patterns and trends in tourist activity. The tourism sector has seen a great increase in the availability of information about travellers, their tastes, and their interactions with diverse places in an era of digital transformation and data proliferation.

This project uses big data analytics to evaluate vast amounts of data from many sources, including social media platforms, online travel firms, booking systems, and tourist surveys. We hope to gain significant insights regarding tourist behaviour, preferences, decision-making processes, and overall satisfaction levels by leveraging the power of big data.

By analysing massive volumes of data, this initiative aims to give tourism stakeholders with actionable insights to improve tourist experiences, optimize tourism offers, and boost destination competitiveness. These insights can be used to adjust marketing efforts, create personalized recommendations, and design more efficient infrastructure and services to match tourists' changing requirements and expectations.

**2. Methodology**

To extract insights from huge data and analyse tourist behaviour, the Tourist-Behaviour-Analysis project adopts a rigorous and methodical methodology. The methodology utilized in this project is outlined in the following steps:

1. **Data Collection**: The first stage is to collect information from a variety of sources, such as social media platforms, online travel companies, booking systems, and visitor surveys. These resources offer useful information regarding tourist demographics, preferences, reviews, ratings, and travel patterns. Web scraping, API integration, or collaboration with data providers may be used to acquire data.
2. **Data Preprocessing**: Once acquired, data is pre-processed to ensure its quality and consistency. Cleaning the data entails deleting duplicate records, dealing with missing values, and standardizing formats. Text preprocessing techniques like as tokenization, stop-word removal, and stemming may be applied to textual data such as reviews and social media posts.
3. **EDA (Exploratory Data Analysis):** EDA is used to get a basic understanding of the data and detect patterns and trends. To identify insights and linkages within the data, descriptive statistics, data visualization tools, and exploratory approaches like as clustering and dimensionality reduction are used.
4. **Model Development**: A variety of analytical techniques are used to construct predictive models and find hidden patterns in tourist behaviour. Based on the given characteristics, machine learning algorithms such as regression, classification, clustering, and recommendation systems are used to anticipate and classify tourist behaviour. Natural language processing (NLP) and sentiment analysis are advanced techniques used to derive insights from textual data.
5. **Forecasting**: Based on past data, time series forecasting models are constructed to estimate future values of tourist behaviour. Forecasting methods such as autoregressive integrated moving average (ARIMA), exponential smoothing methods, and machine learning algorithms such as recurrent neural networks (RNNs) can all be used. To provide reliable forecasts, these models take into account the detected patterns, trends, and seasonality in the data.

**3 Work Tutorial**

**3.1 Requirements**

Python 3.5 was used to complete this project. This project requires the following Python packages to run. To get a first look at the dataset, use Free File Viewer (or any equivalent program). You can also use Anaconda to make installing the essential packages easier.

**Natural Language Tool Kit (NLTK)**

**Folium**

**Pandas**

**matplotlib**

**NumPy**

**Hdbscan**

**basemap**

**jupyter notebook**

**3.2 Setup Instructions**

In comparison to the Python Environment, We used the Anaconda Environment for a lot easier user experience. Anaconda or Python (https://www.python.org/downloads/) can be used.

1. Install Anaconda Python Environment Manager by visiting <https://conda.io/docs/user-guide/install/index.html>.
2. Create a new Python environment and use the pip/conda install command to install the necessary packages.

**3.3 Dataset**

* The dataset utilized in this study is a subset of the YFCC dataset, which may be downloaded at "Yahoo Webscope" (<http://webscope.sandbox.yahoo.com>).
* Setting their respective geographical coordinates in RefineDataset.ipynb yields records for the San Francisco region and other locations such as India, the United States, the United Kingdom, and Europe.

**3.4 Tasks:**

This folder houses the ipython notebooks for each task:

1. **Refinement of the dataset**

* RefineDataset.ipynb notebook is used to set geographical coordinates in order to extract data from a certain region which can be used for further processing or analysing the trends and for estimation.

1. **Processing of textual metadata**

* Textual metadata processing is a crucial step in analysing and extracting meaningful information from text data.
* It involves techniques to preprocess, analyze, and extract relevant information from textual metadata.
* Tokenization involves breaking down the text into individual words or tokens.
* Common techniques include word tokenization and regular expression-based tokenization.
* Stop words are commonly used words that do not carry significant meaning and can be removed from the text.
* Examples of stop words include "the," "is," "and," etc.
* Removing stop words helps reduce noise and focus on more meaningful terms.
* Text\_Metadata\_Proc.ipynb notebook is used to remove extraneous characters and stop words, as well as to stem words into their root form. The output data is then used to generate the geo clustering data.
* After preprocessing, the textual metadata can be analyzed using various techniques such as clustering.
* Clustering algorithms group similar documents or words together based on their semantic similarity.
* Clustering can uncover hidden patterns, identify topics, or assist in organizing textual data.
* Visualizing the processed textual metadata can provide insights and facilitate understanding.
* Techniques like word clouds, topic modelling, and network graphs can be used for visualization.

1. **Clustering of geographical data:**

* Geographical data clustering is a technique used to group similar data points based on their geographic attributes.
* It helps in discovering spatial patterns, identifying clusters or regions with similar characteristics, and gaining insights from spatial data.
* The output data generated by the textual metadata processing is fed as an input to this notebook. The HDBSCAN algorithm is used by GeoCluster\_data. Ipynb notebook to group popular sites together. (<https://hdbscan.readthedocs.io/en/latest/index.html>).
* Various clustering algorithms can be applied to geographical data, such as K-means, DBSCAN, HDBSCAN, and OPTICS.
* These algorithms partition the data points into clusters based on proximity or density criteria.
* After clustering, analyzing the clusters can reveal meaningful information about spatial trends, hotspots, or regions of interest.
* Statistical analysis, density estimation, or spatial autocorrelation analysis can be performed to study cluster properties.

1. **Prediction of regional trends:**

* Trend estimation plays a crucial role in analyzing time series data, such as tourist arrivals over different time periods.
* Regression techniques, including linear regression, ridge regression, and lasso regression, can be used to estimate trends and make predictions.
* Trend-Master File-Any Region. Ipynb notebook contains a tool for estimating the trend of visitor arrivals in a specific region for a specific location.
* Polynomial regression is used to capture nonlinear trends in the data by adding polynomial terms to the linear regression model.
* The code demonstrates polynomial regression of degree 1 and degree 3.
* Ridge regression is a regularization technique that adds a penalty term to the loss function, helping to prevent overfitting.
* The code presents ridge regression using polynomial features of degree 1 and degree 3.
* Mean absolute error (MAE) is used as the evaluation metric to assess the performance of the regression models.
* The MAE between the predicted tourist arrivals (Y\_pred) and the actual values (Ct) is computed for each model.

1. **Time Series Modelling:**

* Time series modelling is a statistical technique used to analyze and forecast data that varies over time.
* Time series data consists of observations collected at regular intervals over time.
* It exhibits temporal dependence, where values at one time point are influenced by previous values.
* Components of time series include trend, seasonality, cyclicality, and random variation.
* Time\_series. Ipynb notebook is a Python script that analyses the seasonal trend of visitor arrivals in a specific region and estimated the future patterns based on the past analysis.

**3.5 Implementation**

* 1. **Textual Metadata Processing:**

The large amount of data is used representing the various countries and time zones**.**

* In the first phase "Textual metadata processing" is done in Text\_Metadata\_Proc.ipynb, which takes as input the filtered San Francisco country records stored as Filtered1M.csv file.

(We can take any country data by modifying the geographical coordinates in the RefineDataset.ipynb notebook to generate the data for that country).

* The filtering threshold value must be defined within the notebook, and the filtered entries are saved in the file TP\_op1M.csv. This output file is then fed as an input to the next processing.

**b. Geographical Data Clustering:**

Now with the above obtained data for the particular country we will process that data for geographical clustering.

* The TP\_op1M.csv file is now given as an input to the second stage present in GeoCluster\_Data.ipynb notebook for "Geographical data clustering."
* The clustering parameters can be changed in the ipython notebook, and the results are displayed as images within the notebook. Within the notebook, you can also see the geographical values for the country.
* We get the geographical clustering image formed before and after the cluster formation. The resulting images are stored in Results/GeographicalDataClustering folder.

**c. Trend Analysis:**

Now to estimate the trend analysis of tourists in different years we process the data of any country to get the desired output.

* We are using the Trend-Master File-fr any Region.ipynb notebook. It performs the "region wise trend estimation" for many regions. (As per our choice).
* The implementation outcomes are recorded and saved in the file Trend Analysis.pdf in Results -> TrendEstimation.
* The estimation graph is also obtained while running the code for a specific country.

**d. Time Series Model:**

Now to construct a time series data in order to estimate the tourism demands and tourists behaviour we used the time series model to generate the graph depicting the future estimation.

* We used the same San Francisco records that was used in the first phase for the Time series modelling.
* We can use any other region also which can be fed as an input into the Seasonal\_SF. ipynb notebook and the results can be viewed as estimation graph in the notebook.
* The implementation's results are captured and saved in the Results -> Trend Estimation -> Time Series folder.

The following image shows the architecture of the proposed system.

Metadata of geo tagged photo

Textual Metadata Textual Refined Support Interested

Processing Metadata Photo Tags Threshold Tags

Geographical Geographical Density Based Clusters Denoting Exemplar

Data Clustering Data Clustering Tourists Spot of cluster

Trend Temporal Regression Estimate

Estimation Info Function Trend

Time Series Estimated Linear Regression Predictions Plotting

Modelling Data (Model Prediction) Against Actual Tourists

**Fig 3a**

These steps are followed one after the other (fig 3a) in a notebook in order to get the desired results.

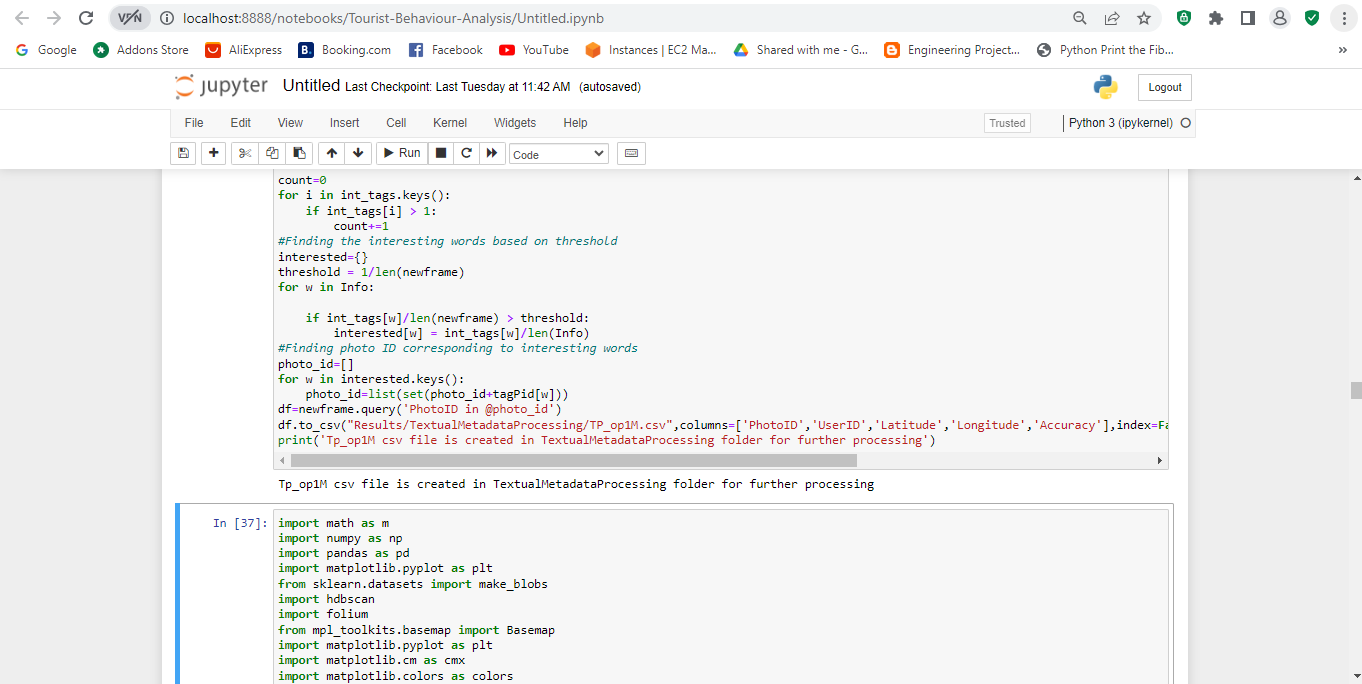
**4) Outcome/Result**

This folder contains the outcomes of each stage. This can be utilized to improve the visualisation of the output and acquire a better knowledge of the output graphs and data we got.

1. **Input Records** – In this folder we generated the filtered data from the large data set in the Filtered1M.csv file, which is used as input for the Textual metadata processing stage.

­­Filtered1M.csv is created by using a Python script to extract 1 million entries from the YFCC Dataset.

1. **Textual Metadata Processing** – This step generates the output file TP\_op1M.csv as a result of Textual metadata processing and store it in the mentioned location as shown in fig 4a.

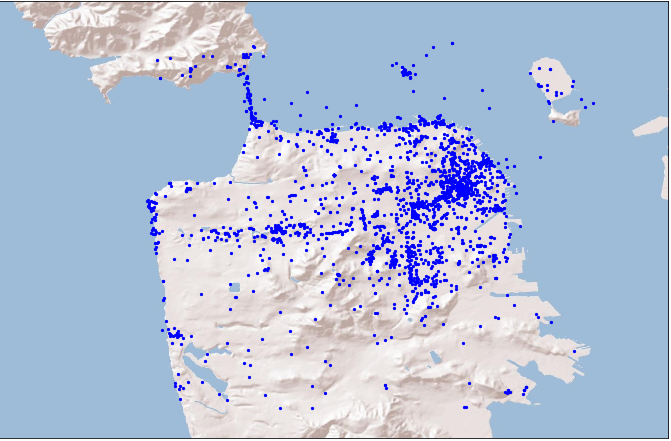


**Fig 4a**

1. **Geographical Data Clustering** – This folder consists of the Geographical clustering stage output. After running the notebook depicting the code to generate the geographical data clustering, we get the image representing the clusters.

Below are the Images depicting cluster formation.

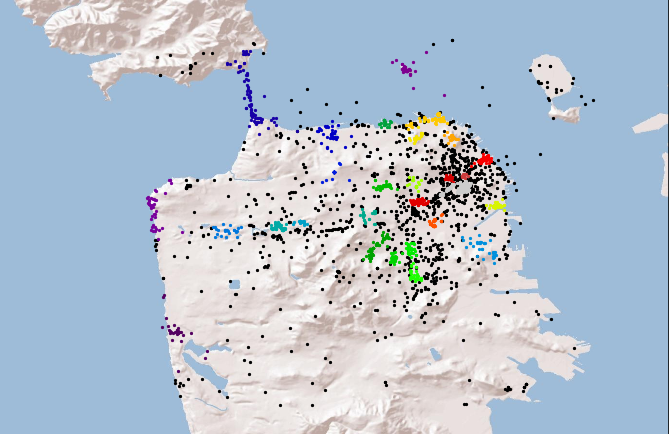
**Image Before Cluster Formation:**



**Fig 4b**

The first plot Fig 4b shows all the data points as blue dots on a map. The map is centered around a specific region defined by the latitude and longitude boundaries. It provides a general overview of the data distribution in that area.

**Image After Cluster Formation:**



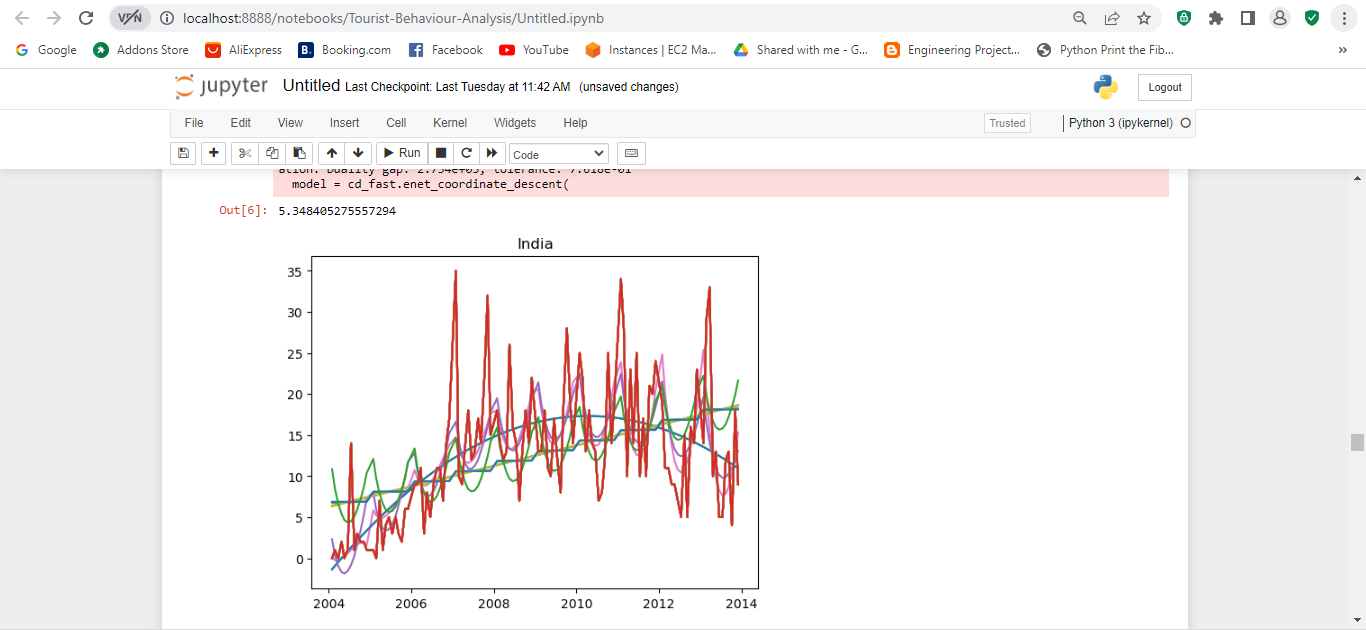
**Fig 4c**

The second plot Fig 4c represents the clusters identified by the HDBSCAN algorithm. Each cluster is assigned a different color. The data points belonging to each cluster are plotted using the corresponding color. Outliers that do not belong to any cluster are shown as black dots. This plot helps visualize the spatial clustering patterns in the dataset.

The graph output helps us understand the spatial concentration and dispersion of the data points and how they are grouped into different clusters based on their geographic proximity.

**4)** **Trend Estimation** – This folder stores the CSV files for various regions where the trend needs to be estimated. After using the file for particular country, we can generate the graph which shows how the predicted values (Y\_pred) and the actual number of tourists (Ct) vary over time for particular country.

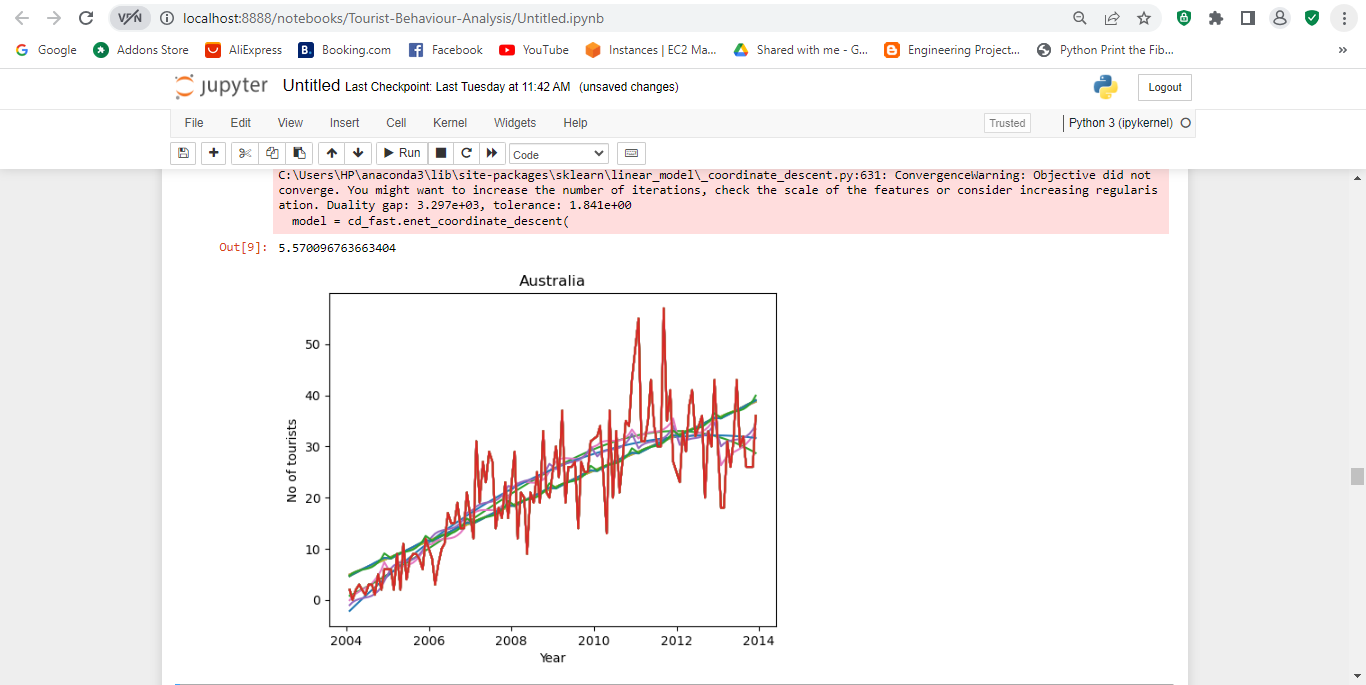
The below graph Fig 4d predicts the number of tourists in a specific region, particularly in India. The mean absolute error is also calculated to assess the accuracy of the predictions.



**Fig 4d**

The below image Fig 4e predicts the number of tourists in a specific region, particularly in Australia.

The mean absolute error is also calculated to assess the accuracy of the predictions.



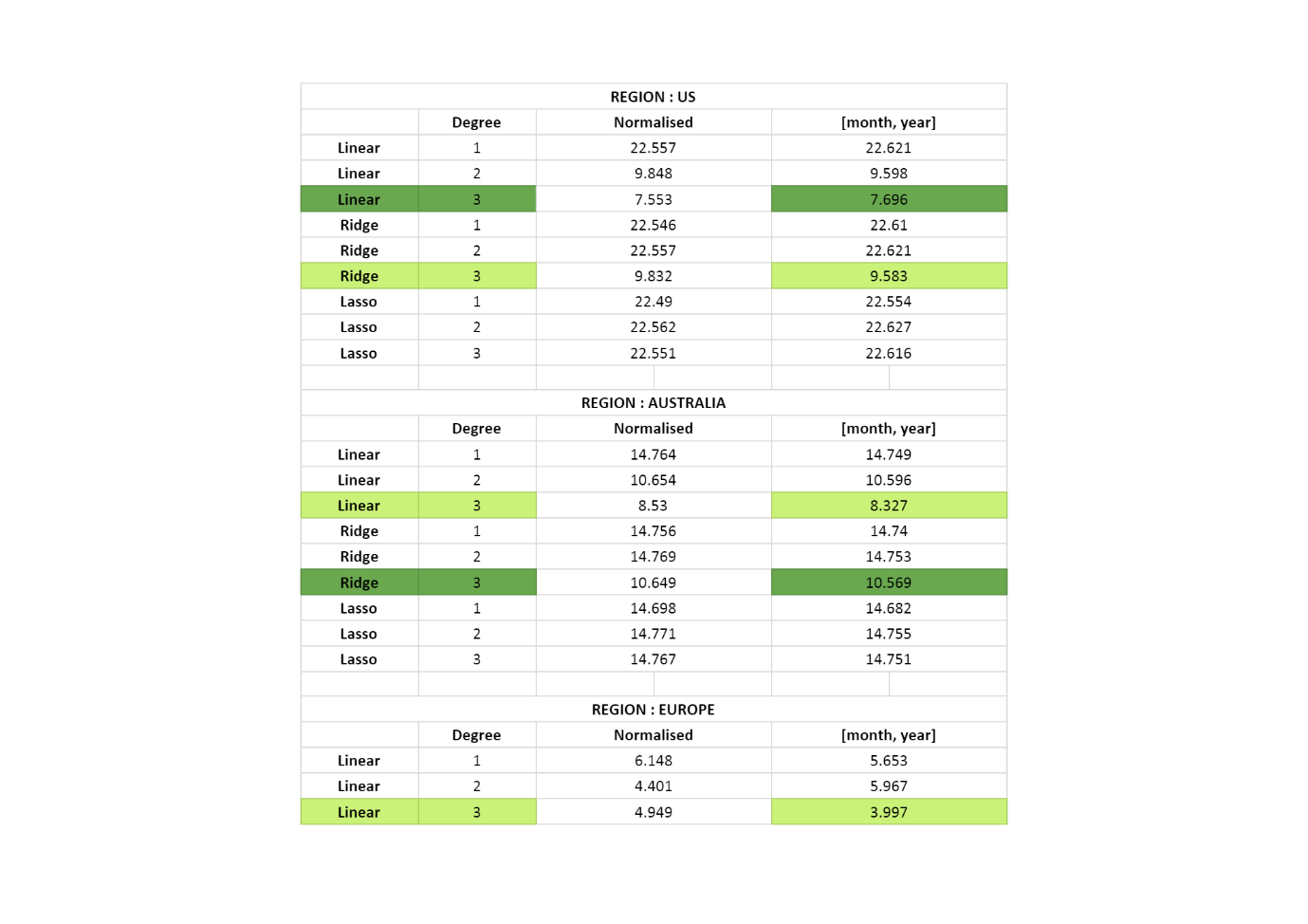
**Fig 4e**

In the above graphs, the different colors of the curves represent different regression models used to fit the data. Each regression model is associated with a specific degree of polynomial features.

* Blue: Linear regression with polynomial degree 1.
* Orange: Polynomial regression with polynomial degree 3.
* Green: Ridge regression with polynomial degree 1.
* Red: Ridge regression with polynomial degree 3.
* Purple: Lasso regression with polynomial degree 1.
* Cyan: Lasso regression with polynomial degree 3.

Each color represents a different combination of regression technique and polynomial degree used to fit the data. The curves show the predicted values by the corresponding regression models against the actual values (represented by the dots).

The colors help differentiate between the different regression models and visually compare their performance in approximating the relationship between the month-year values and the number of tourists.

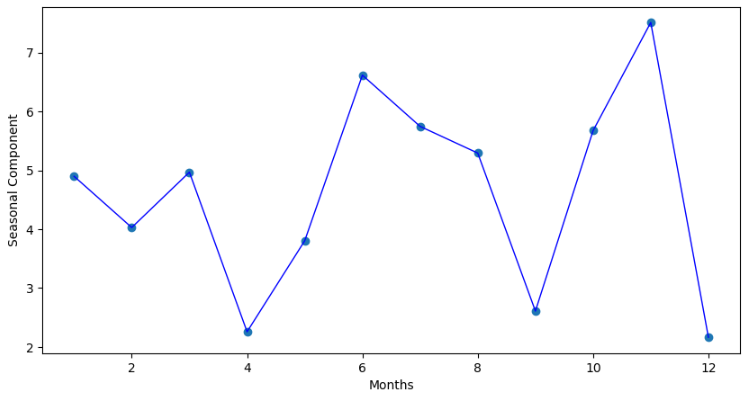
There is one Trend Analysis.pdf which contains the results of an analysis performed on these regions as shown below in fig 4

**Fig 4f**

Another folder titled 'Time Series' provides the region's seasonal trend analysis. It consists of multiple Graphs as described below:

**Time Series Component**

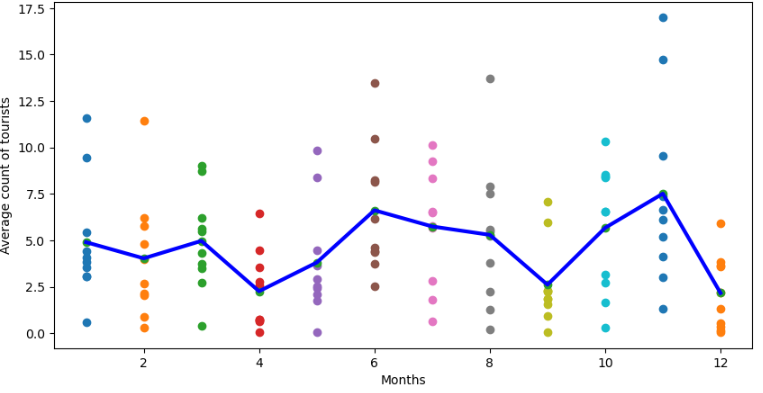
The graph below fig 4g shows the average absolute difference between the actual count of tourists and the predicted count based on the regression models. The y-axis represents the seasonal component, which indicates the deviation of the actual count from the predicted trend.



**Fig 4g**

**Time Series Output**

The final figure fig 4h combines both the average count of tourists and the seasonal component in a single plot. The blue line represents the average count of tourists for each month, while the scatter points represent the seasonal component for each month.

****

**Fig 4h**

We can see that Time Series Component analysis and Time Series output estimation is fairly aligned to each other.

**4.1 Challenges Faced:**

**1) Data Quality:**

One of the most diffic­­ult challenges we faced is guaranteeing the accuracy of the input data. The code assumes that the data is clean, appropriately formatted, and has all necessary information. However, missing values, outliers, inconsistent formats, and other data quality concerns are common in real-world datasets. We took care of such issues by using the data preprocessing techniques such as data cleansing, imputation, and outlier detectionto make the data close to accurate.

**2) Data Size and Performance:**

The dataset we used is very huge, having millions of records or significant temporal coverage, so the execution time and memory use of the code was becoming a problem. To deal with it we optimized the code and explore utilizing more efficient algorithms or data structures to effectively manage the large data.

**3) Model Selection**:

In our code linear, ridge, and lasso regression models of simple regression were used to estimate trends. It was difficult to choose the right model for a particular dataset, though. After careful consideration and testing we chose the ideal model complexity, take into account overfitting or underfitting, and investigated different modelling techniques (such as time series models).

**4.2 Limitations and future work**

**Including External Factors**:

The code presently assesses the number of tourists based purely on the dataset provided. External elements such as holidays, events, economic indicators, or weather conditions that may influence tourism patterns should be considered in future studies. This can provide a more complete picture of the factors influencing tourist numbers.

**Conclusion**:

In conclusion, the project utilized the various methods to analyse the tourist behaviour based on the data provided for different regions, identified different tourists interests and most visited locations and accordingly predicted future tourism demands.

The project used python and big data technology and various algorithms such as HDBSCAN to perform spatial clustering on a dataset of latitude and longitude coordinates. Apart from that Time series modelling was used to generate the time series data to calculate the number of tourists on a monthly basis. Overall, the project gave excellent insights into the analysis of tourist behaviour, including the ability to spot trends, identify seasonal patterns, and make predictions. These results can help the tourism sector make better planning and management decisions through data-driven decision-making procedures.

**References:**

* The dataset utilized in this study is a subset of the YFCC dataset, which may be downloaded at "Yahoo Webscope" (<http://webscope.sandbox.yahoo.com>).
* Research and usage of HDBSCAN algorithm (<https://hdbscan.readthedocs.io/en/latest/index.html>)
* Anaconda Python environment manager  <https://conda.io/docs/user-guide/install/index.html>
* Research and Usage of Matplotlib library https://matplotlib.org/