TOURIST BEHAVIOUR ANALYSIS

MINI PROJECT

By

Sarvesh Navare	**60004180096**
Saurav Tiwari	**60004180097**
Shreerang Taparia	**60004180101**
Shruti Sawant	**60004180103**

Guide:

Pankaj Sonawane

Asst. Professor

Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)

University of Mumbai

2020-2021

CERTIFICATE

This is to certify that the mini project entitled **“Tourist Behaviour Analysis”** is a bonafide work of **Sarvesh Navare (60004180096), Saurav Tiwari (60004180097), Shreerang Taparia (60004180101), Shruti Sawant (60004180103)** submitted to the University of Mumbai in partial fulfilment of the requirement for the award of the degree of B.E. in Computer Engineering.

Prof. Pankaj Sonawane
Guide

Dr. Meera Narvekar
Head of Department

Dr. Hari Vasudevan
Principal

Mini Project Report Approval

This mini project report entitled ***Tourist Behaviour Analysis*** by **Sarvesh Navare (60004180096), Saurav Tiwari (60004180097), Shreerang Taparia (60004180101), Shruti Sawant (60004180103)** is approved for the partial fulfilment of the degree of ***B.E. in Computer Engineering.***

Examiners

1.---

2.---

Date:
Place: Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

S.R.Navare

Sarvesh Navare

Stiwari

Saurav Tiwari

Taparia

Shreerang Taparia

Sawant

Shruti Sawant

Date: 01/05/2020

Abstract

The tourism industry in India generated about 16.91 Lakh Crore or US$240 billion in the year 2018 which was 9.2% of India’s GDP. It was responsible for generating 42.673 million jobs or 8.1% of India's total employment. With such a large contribution, it is important that the government and the local authorities focus on improving this sector on the economy with the right research and the investments in the right places. The conventional method of analyzing tourist locations and numbers was with obtaining figures from the airports and train stations along with surveys conducted on the ground. However, this was a very cumbersome process. The dearth of technological advances in this sector marks the genesis of this project. The integral part of our project is extraction of large amounts of geotagged data from a very famous social media platform, ‘Instagram’. For the sake of this project, we have taken the example of ‘Goa’, a state in India famous for its beaches. For further analysis and comparison, we took the example of the culturally famous state, ‘Rajasthan’ and the hilly state ‘Himachal Pradesh’. The analysis of spatial data allowed us to form clusters and know about the top areas visited in a particular state and the access to time variant data ensured us information about the changes in the tourism demographics of a state. Also after the CoVid-19 pandemic, the tourism industry suffered a lot due diminishing number of tourists[1]. With our research, the authorities can now put focus on areas that are less visited by the tourists and try to understand the reasons for the same. The three large groups of people can benefit from the results of analysis: tourists themselves, public sector managers and businessmen[6]. Further scope in this project includes designing an application that allows any user to access the most visited locations in a state. While this information is available online on various websites, getting it from social media involves a sense of human touch which makes it more reliable. A comprehensive review system can also be developed from social media analysis.

Contents

List of Figures

List of Tables

CHAPTER 1
INTRODUCTION

1.1. Description

The behaviour of tourists is the most important indicator or predictor of future tourist behaviour. Taking into account the social role of the tourist, the behaviour of an individual tourist can also be an indicator of the behaviour of others. With their behaviour, tourists set the social norms of behaviour in the context of tourism. These norms are also followed by other consumers; those who do not yet engage in travel or tourist behaviours, as well as those who do.

Tourist behaviour occurs in the planning and implementation stages of the holidays, and also after they return home. In order for the tour operator or destination to assess the relevance of its marketing and operational approaches to the development, marketing and implementation of tourism activities, it is necessary to recognise the different forms of behaviour in each stage.

Only by knowing the fundamentals of tourist behaviour, as well as knowing how to observe and measure them, can we effectively plan offers and other sales activities in tourism. Theoretical foundations are crucial in empirical research/the measurements of tourist behaviour, as they reveal the concepts that should be measured, and usually also the ways to measure them.

Data analytics is now the technique that is growing in importance in the industry. In the tourism area, behaviour analysis can provide a meaningful difference in the way in which business is traditionally done. With modern solutions, the work can be done quickly and with better outcomes. Therefore, it would be beneficial for the tourists as well as the people involved in the tourism industry by facilitating quicker services to the tourists, thus augmenting contentment and fidelity.

1.2. Problem Formulation

After studying a few papers, we realised that major research has been done on tourist recommendation systems rather than tourist behaviour analysis.

Some physical factors like geographical and climatic conditions, facilities and amenities available at the destination, advertising and marketing conducted by tourism businesses alter the decision making of the tourists. A few social factors such as a person's social network, which provide first hand information that can alter a person's decision of visiting or not visiting a particular place. The more educated the tourist is, the wider range of choices, curiosity, and the knowledge of places he would have. This drives the decision making when it comes to choosing a destination. There can be a broad spectrum of tourist behavior depending upon the place they belong to. North Americans like to follow their own cultural framework. Japanese and Korean tourists like to visit places in groups.Also, tourism destinations are a major contributing factor altering tourist behavior. If a destination has all basic provisions such as electricity, water, clean surroundings, proper accessibility, amenities, and has its own significance, it largely attracts tourists.

Many businesses in the tourist industry face massive losses every year, which can be either ascribed to errors on the tourist sides or the businesses themselves . For instance, if a tourist visits a particular place in an off-season without having prior knowledge about which time of the year is the best to visit that place. This leads to a necessity to provide a complete analysis of tourists who have previously visited that place thus providing information about which time period is the best to visit .

1.3. Motivation

The project is based on a real life problem, faced by several tourists and businesses involved in tourism. If the problem is solved and their actions are clubbed collectively, it can produce phenomenal results and lead to enormous profits. Also after studying a few papers, we realised that major research has been done on tourist recommendation systems rather than tourist behaviour analysis.Also, research done on tourist behaviour analysis includes more information regarding the tourist attraction than the tourists themselves.. So, we aim to provide a crystal clear analysis about the behaviour of the tourists . Due to the covid crisis, the tourism industry of our country has been facing a lot of repercussions. The pandemic has a huge impact on the tourism industry due to the resulting travel restrictions as well as slump in demand among travellers. The collapse in travel will bring long-term changes.

1.4. Proposed Solution

We aim to analyse tourists behaviour based on the locations and places they have visited so far, to identify tourist interests,tourism demographics and to plan future tourism demands. It supports strategic decision-making in tourism destination management. We are going to make use of geotagged images available on social media sites like Instagram for data extraction. The extracted data would include the name, the location visited, the coordinates of the location(the latitude and longitude). Also we would structure the tourist demographic data for all the locations in the vicinity. Make geographical clusters to identify popular tourist locations from tourist interests. Construct a time series data to show the number of tourists at a particular spot throughout the year etc. Predict tourism demands for various locations with the help of time variant data and develop a comprehensive review system with the help of image and text processing which can be passed on to relevant authorities

1.5. Scope of the project

For this project, we propose to create a more efficient and holistic system that will encompass the learnings and fill in the gaps of the other research works. Research in the section of tourist behaviour has not been explored extensively and thus going through the works we noticed the need to research in this specific domain. Moreover, there have been minimal efforts to optimise the losses incurred by the businesses involved in the tourism industry which gives us an incentive to further navigate this area. Extremely large amounts of data can be collected from our social media sites about people who have visited a particular place. While this data will not be in a presentable form, with help of certain data processing techniques, we can make use of this data and provide it to the government or the local authorities and inform them about various tourism interests in their areas. Hotel Chains and Restaurants can also use this data for knowing which periods have the maximum tourist footfall and plan accordingly. By comparing tourism traffic data over the years, we can try and identify strategic decisions which led to the boom of tourism at a particular place and also some logistical shortcomings which if corrected will lead to more tourists at that place. This can also work as a 'Places to Visit' guide for tourists who know nothing about a particular place.e.

CHAPTER 2
REVIEW OF LITERATURE

2.1. Previous work

The conventional method of analysing tourist information was obtaining the figures from airport and train stations. Surveys were also conducted on ground level to understand which places are famous, which are less visited and which places require attention from the government [2]. Other methods included obtaining check in information from hotels[3]. However, with the increasing number of hotels, this would also become a very cumbersome task. Also, major hotel chains would not give away the details of their guests very happily.

Through big data analysis, we aim to reduce the time and workforce required to conduct such surveys. Extraction and analysis of such data can give us a quick and brief overview of the tourism condition in an area. Using machine learning models, predictions can be made about the number of tourists visiting the next year and so on.

A research done by Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. in 2017 [4] made use of the social sharing site 'Flickr' to extract images and then plot them on a map. While we have followed a loosely similar approach, usage of 'Instagram' has helped us in a way due to the fact that 'Instagram', being a social media site, has a much richer data set because users actually want to share their experiences with their followers and friends.

More research done in this field involved the determination of the country of origin of the photographer from the geo tagged images obtained from 'Flick'[5]. While it was not much related to our topic, it helped us for the extraction part. Reference [8] presents a tourist travel analysing framework based on the location-based data from social media. Data from 67,000 Twitter users in Florida was gathered. The authors have used the following clustering methods: K-Means, DBSCAN and Mean-Shift and several classification approaches for the data analysing

A case study done by Mikhailov, S., & Kashevnik, A. in 2020[6] involved the use of building neural networks to analyze and predict the behaviour of tourists. This case study involved the usage of digital patterns of life (sensors, route, content and plan) to make a model and predict tourism demands.

2.2. Research Gap

1. Big data generated across social media sites has created numerous opportunities for bringing more insights to decision-makers. Few studies on big data analytics, however, have demonstrated the support for strategic decision-making. Moreover, a formal method for analysing social media-generated big data for decision support is yet to be developed, particularly in the tourism sector[4].

2. Scientists use Big Data methods and techniques to handle a large amount of open information. Different sources, such as user generated content (photo/videos/attractions reviews), social networks and different sensors from smart gadgets, can be used as the basis for tourist behaviour prediction models. However there is no one single source to obtain the information in a neat format.[6]

3. The authors of Reference [7] aim to integrate multiple data sources to analyze tourists' spatial-temporal behaviour patterns on micro scale distances. Information about tourists' temporal-spatial behaviour was gathered using handheld GPS tracking devices, and questionnaires were distributed to assess tourists' socio-psychological characteristics. As mentioned earlier, the usage of questionnaires was the most cumbersome process.

4. Data from 67,000 Twitter users in Florida was gathered. The authors have used the following clustering methods: K-Means, DBSCAN and Mean-Shift and several classification approaches for the data analysing. However, Twitter poses the issue where a person can click a picture at one location while displaying it to be at another location; hence, the dataset generated is not reliable.
5. Insufficient use of various device sensors used by tourists when traveling - The extracted data will help to more accurately track visits to attractions, modes of travel,

intersections with other tourists, etc. Some of the reviewed articles[6] work with GPS devices, but did not consider in their work the values of the sensors throughout the entire route—only visiting certain attractions.

6. One of the biggest challenges while trying to access social media data is the blocking by various sites. Due to advancement in technologies, the various social media sites have become invulnerable to web bots and manual scraping of such large amounts of data is not possible[9].

CHAPTER 3
SYSTEM ANALYSIS

3.1. Functional Requirements

Since it's a research based project, only a fast processing CPU is required to extract data from social media at high speed. Also, the social media sites should have geotagged images from where the location can be extracted. Further these locations can be used to find the latitudes and longitudes which would help in formation of clusters and further processing.

3.2. Non Functional Requirements

The analysis should be crystal clear. It should provide the desired locations and must be available when required. It should be able to extract data from instagram as and when desired and also store it in the database so it could be retrieved in the future.The analysis should be of tourist places. This is to ensure that the analysis is not used for inappropriate purposes. The database should be secured against attacks of SQL injections and should be accessible only to the authorised person. Should have sufficient accuracy that the users can rely on it. It should meet client satisfaction standards and be able to gain and maintain their trust. Whenever the need is there for the analysis it should be available. It should not provide misleading information when it is required the most and function seamlessly. It should have security to ensure it is not tampered with and is not used for illegal purposes. The analysis should be easily maintainable, the users should be able to add delete locations and should be able to update the report with ease at a later time.

3.3. Specific Requirements

1. There are no specific requirements for the except for having the reports in the specified format.

2. There are no browser specific requirements but a latest browser would be preferred to ensure an infallible process of application.

3.4. Use-Case Diagrams and description

A UML use case diagram is the primary form of system/software requirements for a new software program underdeveloped. Use cases specify the expected behaviour (what), and not the exact method of making it happen (how). Use cases once specified can be denoted both textual and visual representation. A key concept of use case modelling is that it helps us design a system from the end user's perspective. It is an effective technique for communicating system behaviour in the user's terms by specifying all externally visible system behaviour.

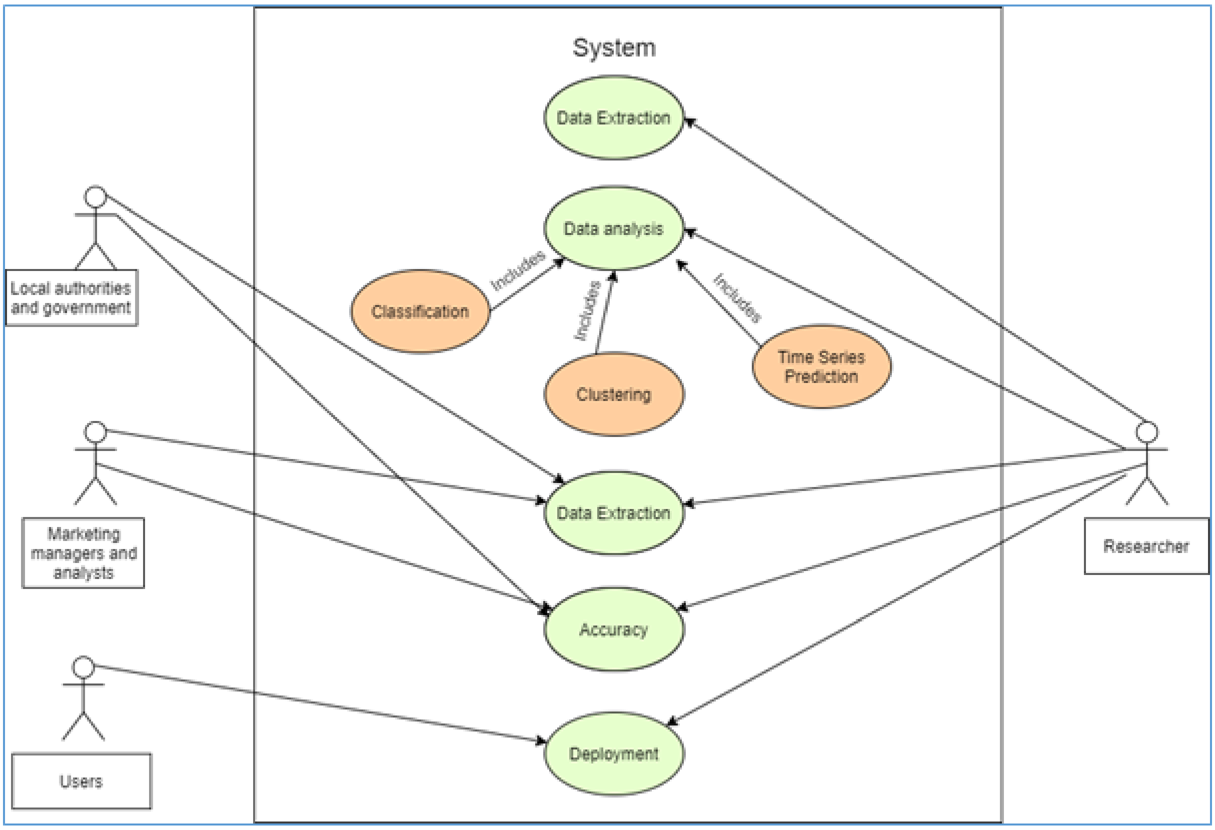


Fig. 3-1 Use-case diagram of the proposed system

The process is described below:

1. Researchers will research and extract data and relevant information from various sources like the internet, past surveys and from local government authorities and hotels.
2. Researchers will then carry out a detailed analysis on the gathered data using classification tools, clustering algos etc.
3. Researcher, Marketing managers, Analysts will check out the accuracy of analyzed data, form training and testing sets for models and then use results for visualization and drawing conclusions.
4. Users and everyone else will have the access to deployed software and give their choice of input to get the result.

CHAPTER 4
ANALYSIS MODELLING

4.1 Class diagram

The project begins with the extraction of the data from social media. This is extremely raw data with large amounts of noise and outliers. This data is cleaned, parsed, standardized and matched so that only those attributes are left which are of utmost importance. Matching helps in the removal of duplication. Further, various operations are performed on this data like clustering and time series prediction due to the presence of spatial and time variant data. Finally, the results obtained are published and passed on to the relevant authorities so that suitable action can be taken.

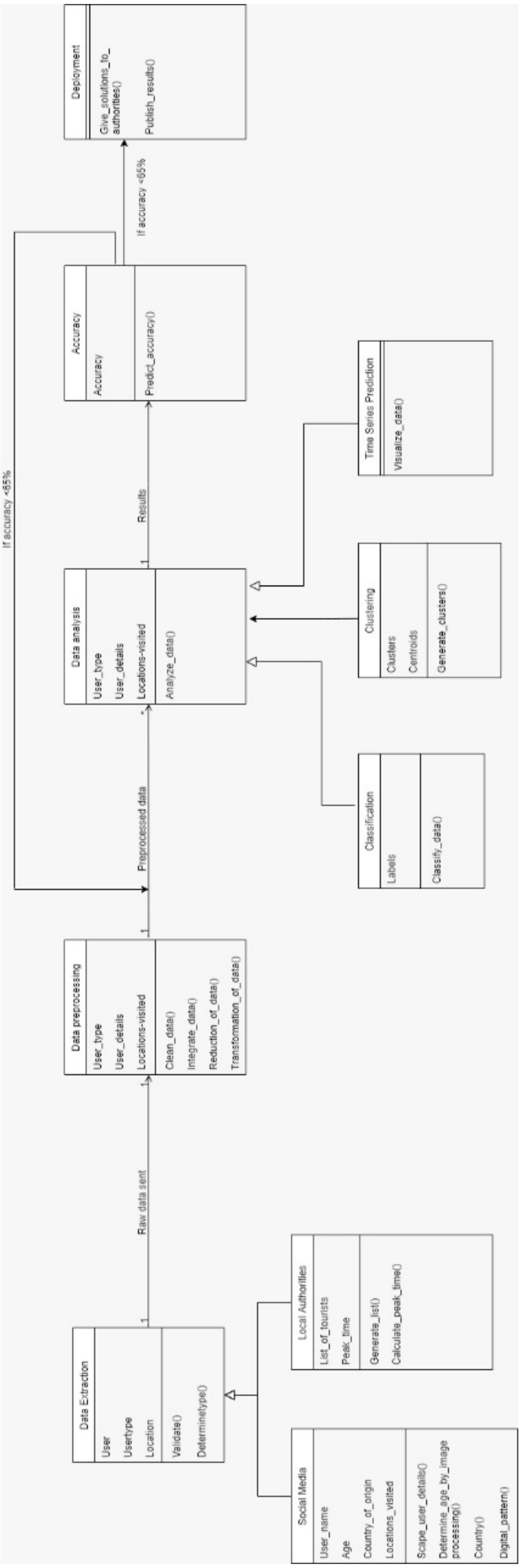


Fig. 4-1 Class diagram of the proposed system

4.2 Sequence diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

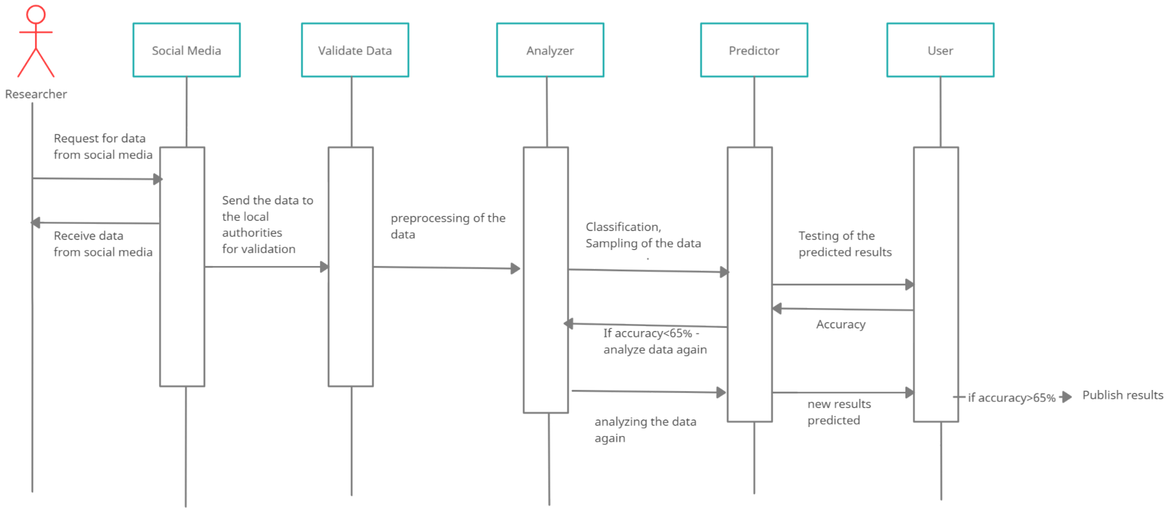

Researcher
Social Media
Validate Data
Analyzer
Predictor
User
Request for data from social media
Receive data from social media
Send the data to the local authorities for validation
preprocessing of the data
Classification, Sampling of the data
Testing of the predicted results
Accuracy
If accuracy<65% - analyze data again
analyzing the data again
new results predicted
if accuracy>65%
Publish results

Fig. 4-2 Sequence diagram of the proposed system

4.3 State Diagram

A state diagram is used to represent the condition of the system or part of the system at finite instances of time. It's a behavioural diagram and it represents the behaviour using finite state transitions. State diagrams are also referred to as State machines and State-chart Diagrams. These terms are often used interchangeably. So simply, a state diagram is used to model the dynamic behaviour of a class in response to time and changing external stimuli. We can say that each and every class has a state but we don't model every class using State diagrams. We prefer to model the states with three or more states.

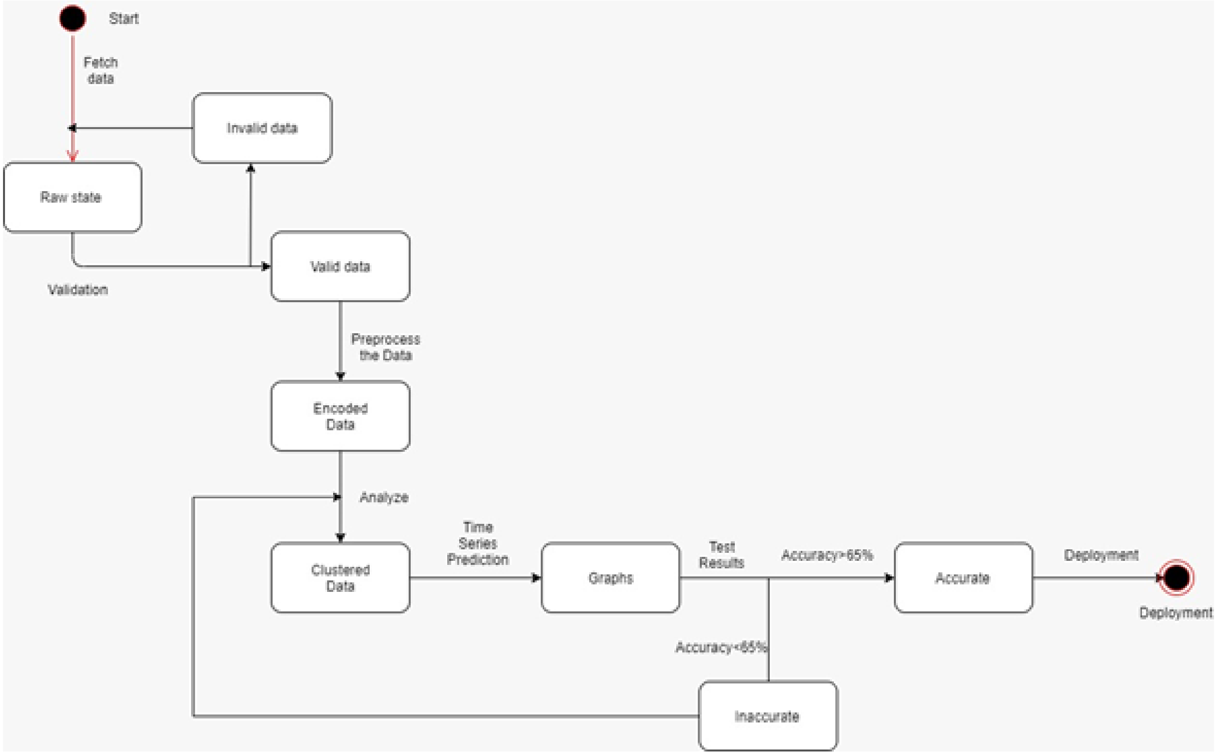


Fig 4-3 State Diagram

CHAPTER 5

DESIGN

5.1. Architectural Design for proposed system

Our project needs the architectural design to represent the design of software. IEEE defines architectural design as "the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system." The software that is built for computer-based systems can exhibit many architectural styles.

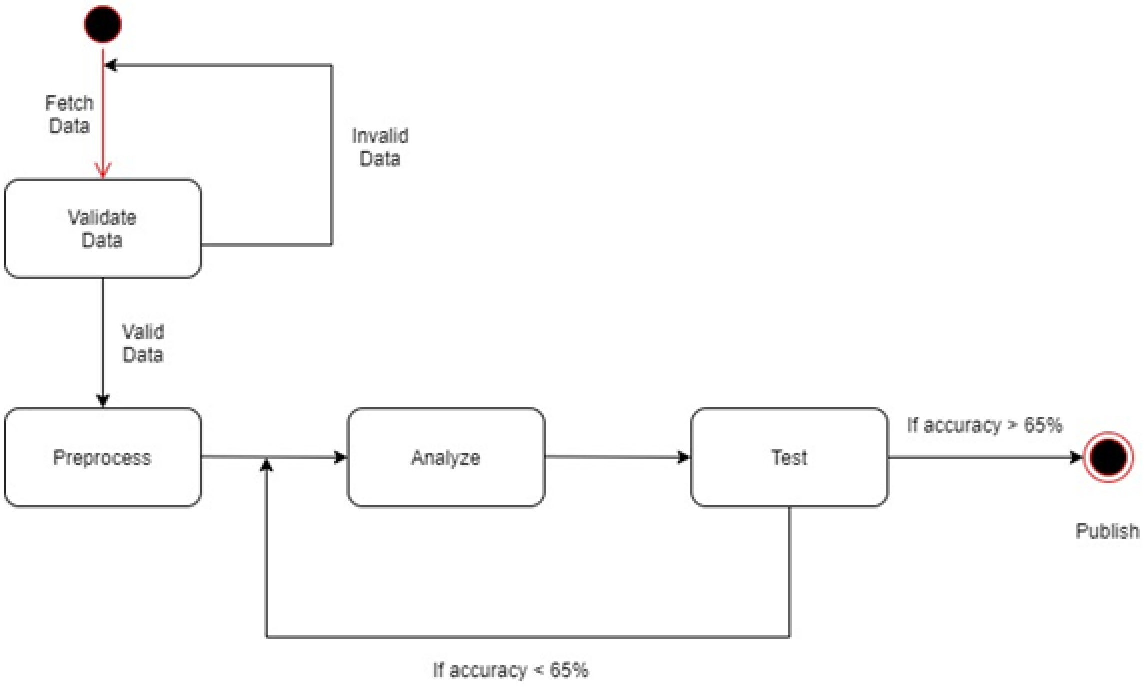


Fig 5-1 Architectural Design for proposed system

The working of the project is described as follows:

1. **Validate Data**

- Fetch data from the social media site.
- If the data is invalid, then fetch again.

- Pass on the data for preprocessing.

2. Preprocess

- Clean and remove inconsistent data.
- The clean and consistent data is fed to the analyzer.

3. Analyze

- Data is analyzed using several data mining tools.
- Clusters are made and used for determining various graphs in the testing process.

4. Test

- Accuracy is determined.
- If the accuracy is more than 65 percent, the analysis is published.
- If the accuracy obtained is less than 65 percent, the data is analyzed again using some different methods.

CHAPTER 6

IMPLEMENTATION

6.1. Algorithms / Methods Used

6.1.1 Description of Algorithms Used

KMeans Clustering:

For this mini project, we have used the KMeans clustering algorithm. It is an extensively used technique for data cluster analysis. We extracted time variant data of tourists visiting various locations (Goa, Rajasthan, Himachal Pradesh).

Our objective was to group frequently visited locations according to their latitude and longitude and discover the underlying patterns. The KMeans algorithm was used to create 10 clusters from the time variant data for a particular location. Libraries like numpy, pandas, matplotlib were used along with sklearn to import KMeans. We selected the columns named as ‘Location’, ‘Latitude’ and ‘Longitude’ from the dataset generated. We have considered k=10, therefore KMeans has identified 10 centroid locations, then it allocated every other location to the nearest cluster while keeping the centroids as small as possible. It aims at minimizing the objective function known as the squared error function.

$$\boxed{J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2}$$

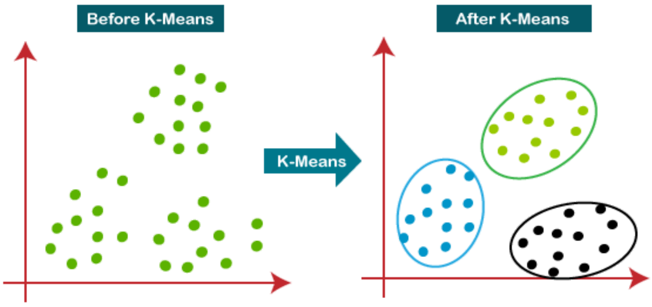
$||x_i - v_j||$’ is the Euclidean distance between x_i and v_j.

c_i’ is the number of data points in *ith* cluster.

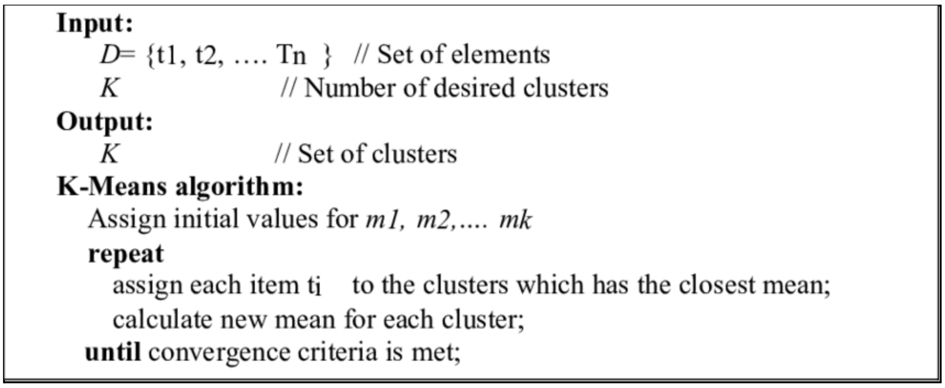
‘c’ is the number of cluster centers.

It performs iterative calculations to optimize the positions of the centroid locations. The process of creation and optimization of the clusters halts when the centroids have stabilized i.e. , there is no change in their values and hence the clustering is successful.

Initially, the various locations are plotted on a graph of latitude vs longitude represented by different colors corresponding to the clusters they belong to. These 10 clusters are then plotted on the map of their respective states which gives an overview of the distinct popular tourist locations along with the strength of the tourists that have visited that particular cluster location.



6.1 KMeans Clustering

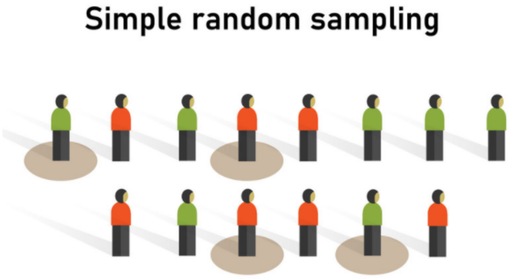


```
Input:
    D= {t1, t2, …. Tn  }   // Set of elements
    K                      // Number of desired clusters
Output:
    K                      // Set of clusters
K-Means algorithm:
   Assign initial values for m1, m2,…. mk
   repeat
      assign each item ti    to the clusters which has the closest mean;
      calculate new mean for each cluster;
   until convergence criteria is met;
```

Description of Methods Used:

Simple Random Sampling:

Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample. Here the selection of items entirely depends on luck or probability, and therefore this sampling technique is also sometimes known as a method of chances.



6.2 Simple random sampling

Since a larger volume of data is extracted, we have performed simple random sampling for the analysis purpose so as to make generalizations about a larger group of data. Every selection made had equal chances of being selected.

We ensure that the results obtained from the sampled data approximates what would have been obtained if the original data had been measured.

6.1.2 Data Extraction

Web Scraping was performed in python to extract data of users and their geotagged locations from the social media platform - Instagram. Instagram does not provide an API for extracting the details required. Hence a code was written in python to extract the data using various libraries like Beautiful Soup, Selenium, Pandas, Urllib etc.

Beautiful Soup is a python library for pulling data out of HTML and XML files. It creates a parse tree for parsed pages that can be used to extract data from HTML.
Selenium is an open-source web based powerful automation tool for controlling web browsers through programs and performing browser automation. Selenium webdriver uses ChromeDriver to control Chrome which is a separate executable. Therefore ChromeDriver installation was carried out.

1. We found out several hashtags from Instagram related to the states that we have taken into consideration. These hashtags were given as an input to our code.
2. Selenium webdriver opens the hashtag page in the browser using the ChromeDriver.
3. We find the source code of that page and parse it through beautiful soup which returns the HTML. The body and script tags are then found in this HTML code.
4. The script tag which is a string is converted into a Python Dictionary by using json_loads(). We get a list of the urls to individual posts present on that hashtag page. '?__a=1' was appended to the urls in the list.
5. A data frame is initialized with the respective column names. For every link in the list received we again find the source code, parse it through beautiful soup html parser and find the dictionary required which contains the data of users by using data_json().
6. Instagram posts may or may not have the location tagged. We consider only those posts which have the location tagged by the user and traverse through the dictionary in step 5 in order to retrieve the user details like timestamp, user id, username, full name, location id, location name, latitude, longitude.
7. It uses the datetime module to find the date and the time of post upload from the timestamp extracted.
8. This data is appended in the dataframe initialized in step 5 and then converted into a csv file as seen in our output.

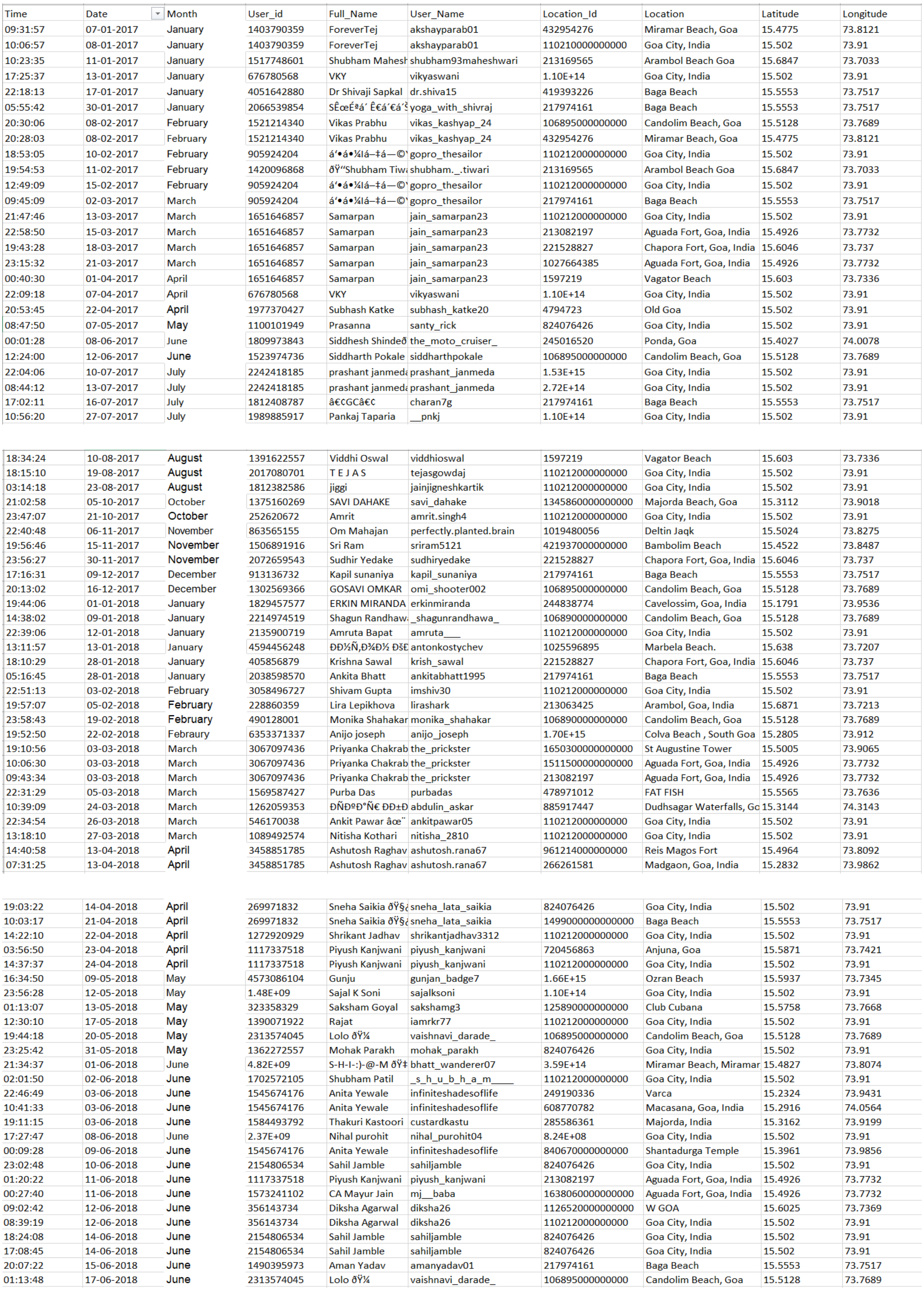
6.1.3 Dataset Description

In this mini project, we have created the dataset from the data extracted by scraping information of users and their geotagged locations from a social media platform - Instagram. Data extraction as well as data preprocessing were the prior steps for generating a clean dataset. The various data fields that we have in our dataset include the time, date, month when the location was tagged in a post, user details like user id, username, full name along with the locations and their latitude and longitude.

Metadata of the dataset used:

Parameter	**Description**
Time	Time of post upload.
Date	Date of post upload.
Month	Month of post upload.
User_id	Unique user id of the user account by whom the location was tagged.
Username	Username of the user account by whom the location was tagged.
Full Name	Full Name of the user by whom the location was tagged.
Location_id	Unique location id that is assigned to various locations.
Location	Location name tagged by the user.
Latitude	Latitude of the location tagged.
Longitude	Longitude of the location tagged.

Table 6-1 Table describing the dataset used



Time	Date	Month	User_id	Full_Name	User_Name	Location_Id	Location	Latitude	Longitude
09:31:57	07-01-2017	January	1403790359	ForeverTej	akshayparab01	432954276	Miramar Beach, Goa	15.4775	73.8121
10:06:57	08-01-2017	January	1403790359	ForeverTej	akshayparab01	110210000000000	Goa City, India	15.502	73.91
10:23:35	11-01-2017	January	1517748601	Shubham Mahesh	shubham93maheshwari	213169565	Arambol Beach Goa	15.6847	73.7033
17:25:37	13-01-2017	January	676780568	VKY	vikyaswani	1.10E+14	Goa City, India	15.502	73.91
22:18:13	17-01-2017	January	4051642880	Dr Shivaji Sapkal	dr.shiva15	419393226	Baga Beach	15.5553	73.7517
05:55:42	30-01-2017	January	2066539854	SÊœÉªá´ Ê€á´€á´Š	yoga_with_shivraj	217974161	Baga Beach	15.5553	73.7517
20:30:06	08-02-2017	February	1521214340	Vikas Prabhu	vikas_kashyap_24	106895000000000	Candolim Beach, Goa	15.5128	73.7689
20:28:03	08-02-2017	February	1521214340	Vikas Prabhu	vikas_kashyap_24	432954276	Miramar Beach, Goa	15.4775	73.8121
18:53:05	10-02-2017	February	905924204	á‘•á•¼Iá–‡á—©	gopro_thesailor	110212000000000	Goa City, India	15.502	73.91
19:54:53	11-02-2017	February	1420096868	ðŸ“Shubham Tiwa	shubham._.tiwari	213169565	Arambol Beach Goa	15.6847	73.7033
12:49:09	15-02-2017	February	905924204	á‘•á•¼Iá–‡á—©	gopro_thesailor	110212000000000	Goa City, India	15.502	73.91
09:45:09	02-03-2017	March	905924204	á‘•á•¼Iá–‡á—©	gopro_thesailor	217974161	Baga Beach	15.5553	73.7517
21:47:46	13-03-2017	March	1651646857	Samarpan	jain_samarpan23	110212000000000	Goa City, India	15.502	73.91
22:58:50	15-03-2017	March	1651646857	Samarpan	jain_samarpan23	213082197	Aguada Fort, Goa, India	15.4926	73.7732
19:43:28	18-03-2017	March	1651646857	Samarpan	jain_samarpan23	221528827	Chapora Fort, Goa, India	15.6046	73.737
23:15:32	21-03-2017	March	1651646857	Samarpan	jain_samarpan23	1027664385	Aguada Fort, Goa, India	15.4926	73.7732
00:40:30	01-04-2017	April	1651646857	Samarpan	jain_samarpan23	1597219	Vagator Beach	15.603	73.7336
22:09:18	07-04-2017	April	676780568	VKY	vikyaswani	1.10E+14	Goa City, India	15.502	73.91
20:53:45	22-04-2017	April	1977370427	Subhash Katke	subhash_katke20	4794723	Old Goa	15.502	73.91
08:47:50	07-05-2017	May	1100101949	Prasanna	santy_rick	824076426	Goa City, India	15.502	73.91
00:01:28	08-06-2017	June	1809973843	Siddhesh Shindeð	the_moto_cruiser_	245016520	Ponda, Goa	15.4027	74.0078
12:24:00	12-06-2017	June	1523974736	Siddharth Pokale	siddharthpokale	106895000000000	Candolim Beach, Goa	15.5128	73.7689
22:04:06	10-07-2017	July	2242418185	prashant janmeda	prashant_janmeda	1.53E+15	Goa City, India	15.502	73.91
08:44:12	13-07-2017	July	2242418185	prashant janmeda	prashant_janmeda	2.72E+14	Goa City, India	15.502	73.91
17:02:11	16-07-2017	July	1812408787	â€¢GCâ€¢	charan7g	217974161	Baga Beach	15.5553	73.7517
10:56:20	27-07-2017	July	1989885917	Pankaj Taparia	__pnkj	1.10E+14	Goa City, India	15.502	73.91
18:34:24	10-08-2017	August	1391622557	Viddhi Oswal	viddhioswal	1597219	Vagator Beach	15.603	73.7336
18:15:10	19-08-2017	August	2017080701	T E J A S	tejasgowdaj	110212000000000	Goa City, India	15.502	73.91
03:14:18	23-08-2017	August	1812382586	jiggi	jainjigneshkartik	110212000000000	Goa City, India	15.502	73.91
21:02:58	05-10-2017	October	1375160269	SAVI DAHAKE	savi_dahake	134586000000000	Majorda Beach, Goa	15.3112	73.9018
23:47:07	21-10-2017	October	252620672	Amrit	amrit.singh4	110212000000000	Goa City, India	15.502	73.91
22:40:48	06-11-2017	November	863565155	Om Mahajan	perfectly.planted.brain	1019480056	Deltin Jaqk	15.5024	73.8275
19:56:46	15-11-2017	November	1506891916	Sri Ram	sriram5121	421937000000000	Bambolim Beach	15.4522	73.8487
23:56:27	30-11-2017	November	2072659543	Sudhir Yedake	sudhiryedake	221528827	Chapora Fort, Goa, India	15.6046	73.737
17:16:31	09-12-2017	December	913136732	Kapil sunaniya	kapil_sunaniya	217974161	Baga Beach	15.5553	73.7517
20:13:02	16-12-2017	December	1302569366	GOSAVI OMKAR	omi_shooter002	106895000000000	Candolim Beach, Goa	15.5128	73.7689
19:44:06	01-01-2018	January	1829457577	ERKIN MIRANDA	erkinmiranda	244838774	Cavelossim, Goa, India	15.1791	73.9536
14:38:02	09-01-2018	January	2214974519	Shagun Randhawa	_shagunrandhawa_	106890000000000	Candolim Beach, Goa	15.5128	73.7689
22:39:06	12-01-2018	January	2135900719	Amruta Bapat	amruta___	110212000000000	Goa City, India	15.502	73.91
13:11:57	13-01-2018	January	4594456248	ÐÐ½Ñ‚Ð¾Ð½ ÐšÐ	antonkostychev	1025596895	Marbela Beach.	15.638	73.7207
18:10:29	28-01-2018	January	405856879	Krishna Sawal	krish_sawal	221528827	Chapora Fort, Goa, India	15.6046	73.737
05:16:45	28-01-2018	January	2038598570	Ankita Bhatt	ankitabhatt1995	217974161	Baga Beach	15.5553	73.7517
22:51:13	03-02-2018	February	3058496727	Shivam Gupta	imshiv30	110212000000000	Goa City, India	15.502	73.91
19:57:07	05-02-2018	February	228860359	Lira Lepikhova	lirashark	213063425	Arambol, Goa, India	15.6871	73.7213
23:58:43	19-02-2018	February	490128001	Monika Shahakar	monika_shahakar	106890000000000	Candolim Beach, Goa	15.5128	73.7689
19:52:50	22-02-2018	Febraury	6353371337	Anijo joseph	anijo_joseph	1.70E+15	Colva Beach , South Goa	15.2805	73.912
19:10:56	03-03-2018	March	3067097436	Priyanka Chakrab	the_prickster	1650300000000000	St Augustine Tower	15.5005	73.9065
10:06:30	03-03-2018	March	3067097436	Priyanka Chakrab	the_prickster	1511500000000000	Aguada Fort, Goa, India	15.4926	73.7732
09:43:34	03-03-2018	March	3067097436	Priyanka Chakrab	the_prickster	213082197	Aguada Fort, Goa, India	15.4926	73.7732
22:31:29	05-03-2018	March	1569587427	Purba Das	purbadas	478971012	FAT FISH	15.5565	73.7636
10:39:09	24-03-2018	March	1262059353	ÐÑÐºÐ°Ñ€ ÐÐ±Ð	abdulin_askar	885917447	Dudhsagar Waterfalls, Go	15.3144	74.3143
22:34:54	26-03-2018	March	546170038	Ankit Pawar âœ¨	ankitpawar05	110212000000000	Goa City, India	15.502	73.91
13:18:10	27-03-2018	March	1089492574	Nitisha Kothari	nitisha_2810	110212000000000	Goa City, India	15.502	73.91
14:40:58	13-04-2018	April	3458851785	Ashutosh Raghav	ashutosh.rana67	961214000000000	Reis Magos Fort	15.4964	73.8092
07:31:25	13-04-2018	April	3458851785	Ashutosh Raghav	ashutosh.rana67	266261581	Madgaon, Goa, India	15.2832	73.9862
19:03:22	14-04-2018	April	269971832	Sneha Saikia ðŸ§¿	sneha_lata_saikia	824076426	Goa City, India	15.502	73.91
10:03:17	21-04-2018	April	269971832	Sneha Saikia ðŸ§¿	sneha_lata_saikia	1499000000000000	Baga Beach	15.5553	73.7517
14:22:10	22-04-2018	April	1272920929	Shrikant Jadhav	shrikantjadhav3312	110212000000000	Goa City, India	15.502	73.91
03:56:50	23-04-2018	April	1117337518	Piyush Kanjwani	piyush_kanjwani	720456863	Anjuna, Goa	15.5871	73.7421
14:37:37	24-04-2018	April	1117337518	Piyush Kanjwani	piyush_kanjwani	110212000000000	Goa City, India	15.502	73.91
16:34:50	09-05-2018	May	4573086104	Gunju	gunjan_badge7	1.66E+15	Ozran Beach	15.5937	73.7345
23:56:28	12-05-2018	May	1.48E+09	Sajal K Soni	sajalksoni	1.10E+14	Goa City, India	15.502	73.91
01:13:07	13-05-2018	May	323358329	Saksham Goyal	sakshamg3	125890000000000	Club Cubana	15.5758	73.7668
12:30:10	17-05-2018	May	1390071922	Rajat	iamrkr77	110212000000000	Goa City, India	15.502	73.91
19:44:18	20-05-2018	May	2313574045	Lolo ðŸ¼	vaishnavi_darade_	106895000000000	Candolim Beach, Goa	15.5128	73.7689
23:25:42	31-05-2018	May	1362272557	Mohak Parakh	mohak_parakh	824076426	Goa City, India	15.502	73.91
21:34:37	01-06-2018	June	4.82E+09	S-H-I-:)-@-M ðŸ‡	bhatt_wanderer07	3.59E+14	Miramar Beach, Miramar	15.4827	73.8074
02:01:50	02-06-2018	June	1702572105	Shubham Patil	_s_h_u_b_h_a_m____	110212000000000	Goa City, India	15.502	73.91
22:46:49	03-06-2018	June	1545674176	Anita Yewale	infiniteshadesoflife	249190336	Varca	15.2324	73.9431
10:41:33	03-06-2018	June	1545674176	Anita Yewale	infiniteshadesoflife	608770782	Macasana, Goa, India	15.2916	74.0564
19:11:15	03-06-2018	June	1584493792	Thakuri Kastoori	custardkastu	285586361	Majorda, India	15.3162	73.9199
17:27:47	08-06-2018	June	2.37E+09	Nihal purohit	nihal_purohit04	8.24E+08	Goa City, India	15.502	73.91
00:09:28	09-06-2018	June	1545674176	Anita Yewale	infiniteshadesoflife	840670000000000	Shantadurga Temple	15.3961	73.9856
23:02:48	10-06-2018	June	2154806534	Sahil Jamble	sahiljamble	824076426	Goa City, India	15.502	73.91
01:20:22	11-06-2018	June	1117337518	Piyush Kanjwani	piyush_kanjwani	213082197	Aguada Fort, Goa, India	15.4926	73.7732
00:27:40	11-06-2018	June	1573241102	CA Mayur Jain	mj__baba	1638060000000000	Aguada Fort, Goa, India	15.4926	73.7732
09:02:42	12-06-2018	June	356143734	Diksha Agarwal	diksha26	1126520000000000	W GOA	15.6025	73.7369
08:39:19	12-06-2018	June	356143734	Diksha Agarwal	diksha26	110212000000000	Goa City, India	15.502	73.91
18:24:08	14-06-2018	June	2154806534	Sahil Jamble	sahiljamble	824076426	Goa City, India	15.502	73.91
17:08:45	14-06-2018	June	2154806534	Sahil Jamble	sahiljamble	824076426	Goa City, India	15.502	73.91
20:07:22	15-06-2018	June	1490395973	Aman Yadav	amanyadav01	217974161	Baga Beach	15.5553	73.7517
01:13:48	17-06-2018	June	2313574045	Lolo ðŸ¼	vaishnavi_darade_	106895000000000	Candolim Beach, Goa	15.5128	73.7689

Fig 6-3 Overview of the dataset

6.1.4 Data Preprocessing

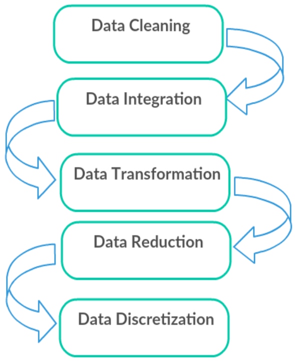


Fig 6-4 Steps of data preprocessing

Several steps of data preprocessing were carried out on the dataset generated by data extraction:

Data Cleaning

Data cleaning is a process of preparing the raw data for analysis by removing the dirty data, organizing the raw data and filling the null values. Our dataset required data cleaning as :

- The data entries with null username had to be removed.
- Location names were incorrectly spelled by users thereby resulting in different names for the same location. For example: Chittorgarh Fort, Chittaurgarh Fort, Chittor Fort - these names were given by users which refer to the same location. Such discrepancies were removed by replacing them with the correct names.
- Users had incorrectly tagged the locations. For example: The data of Rajasthan had Birla Mandir as one of the locations. But the user had tagged Birla Mandir in Maharashtra instead of the one in Jaipur. This was a dirty data which was cleaned by replacing the coordinates with the appropriate ones.

Data Transformation

Data transformation is the process of converting data from one format or structure into another format or structure.

Attribute construction has been performed, where new attributes are created from an existing set of attributes. This method of reconstruction made mining more efficient and helped us in creating new datasets quickly. We reconstructed 3 different datasets based on their locations (Goa, Rajasthan, Himachal Pradesh) from the original dataset consisting of all the data.

Data Reduction

Data Reduction is a process that reduces the volume of original data and represents it in much smaller volume. It ensures integrity of the data while reducing the data. The extracted data consisted of several data tuples that were reduced by a certain factor. Proportionate distribution of data was ensured during the process of reduction.

6.2. Working of the project

We have considered 3 states (Goa, Rajasthan, Himachal Pradesh) as our target states for the analysis purpose. The different stages involved in this project are:

- **Data extraction by web scraping:** Data of users and their geotagged locations were extracted by scraping data from Instagram. A python code was run to scrape the data, since instagram does not provide an API, whose output was a csv file consisting of the entire data of users. Different libraries like Beautiful Soup, Selenium, pandas etc were used.

- **Data Preprocessing:** The raw data extracted was converted into an efficient and a useful format.
  - **Data Cleaning:** Missing values, Null values and the dirty data was cleaned by either replacing them with the correct data or removing those tuples.
  - **Data Transformation:** State wise dataset was reconstructed from the original dataset of the extracted data.
  - **Data Reduction:** The original data was reduced to smaller volume in a proportionate manner so as to ensure that the results obtained approximate the ones that would have been obtained from the larger volume of data.

- **Plotting:** Unique locations visited by tourists during several time periods were plotted on their respective state maps. Yearly plots were also made for the no of tourists that visit a particular state.

- **Clustering:** KMeans clustering algorithm was used to cluster the locations based on their latitude and longitudes along with the density of tourists.

- **Remapping the clusters:** Improvised clusters were mapped.

- **Comparisons:** Comparisons were made depending on the number of people visiting the states in different quarters of the year.

- **Conclusions:** Conclusion was derived from the various graphs plotted.

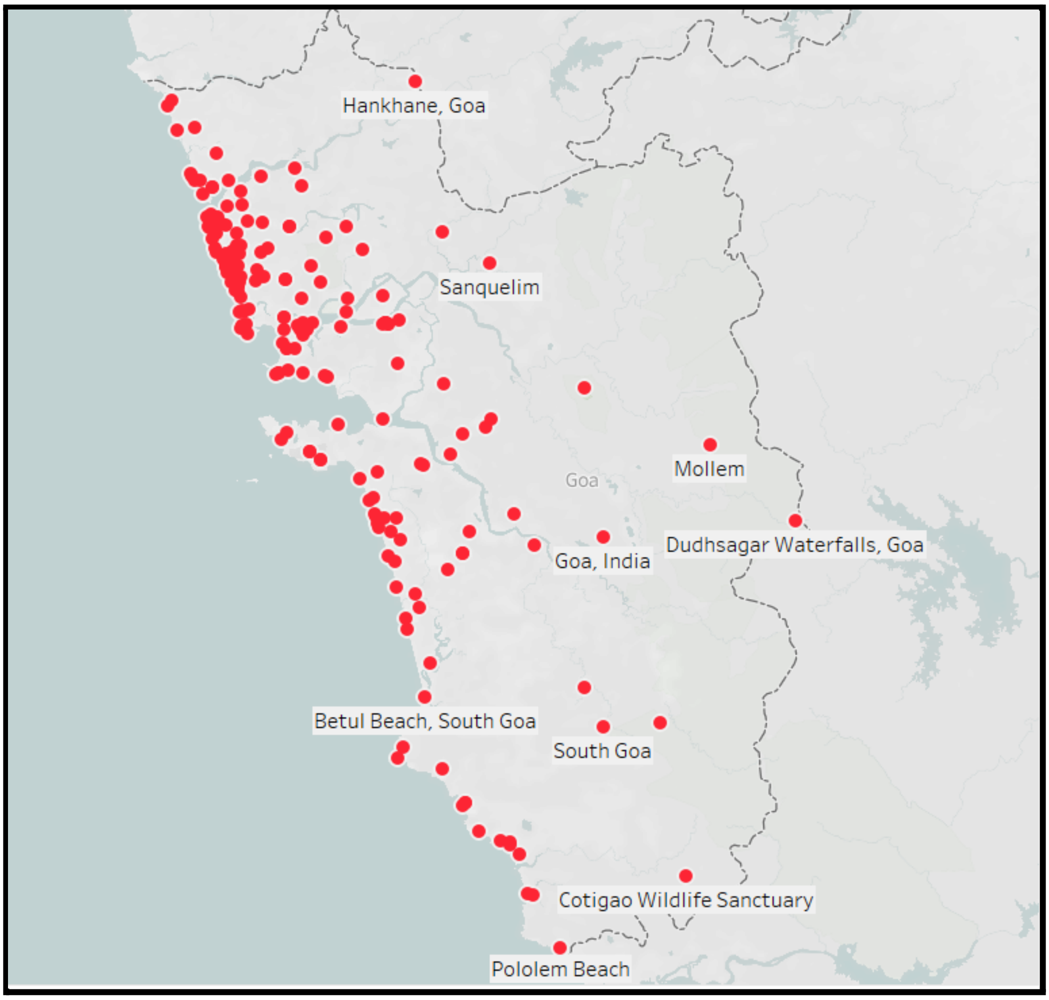
CHAPTER 7

EXPERIMENTATION AND ANALYSIS

For experimentation and analysis purposes, we shortlisted 3 different states having varied geographical features and decided to analyse the tourism in those states. We considered Goa, Rajasthan and Himachal Pradesh. Below are various visualizations that we generated using the data we had extracted and cleaned of these places which give us the rough overview of tourism demography in these states.

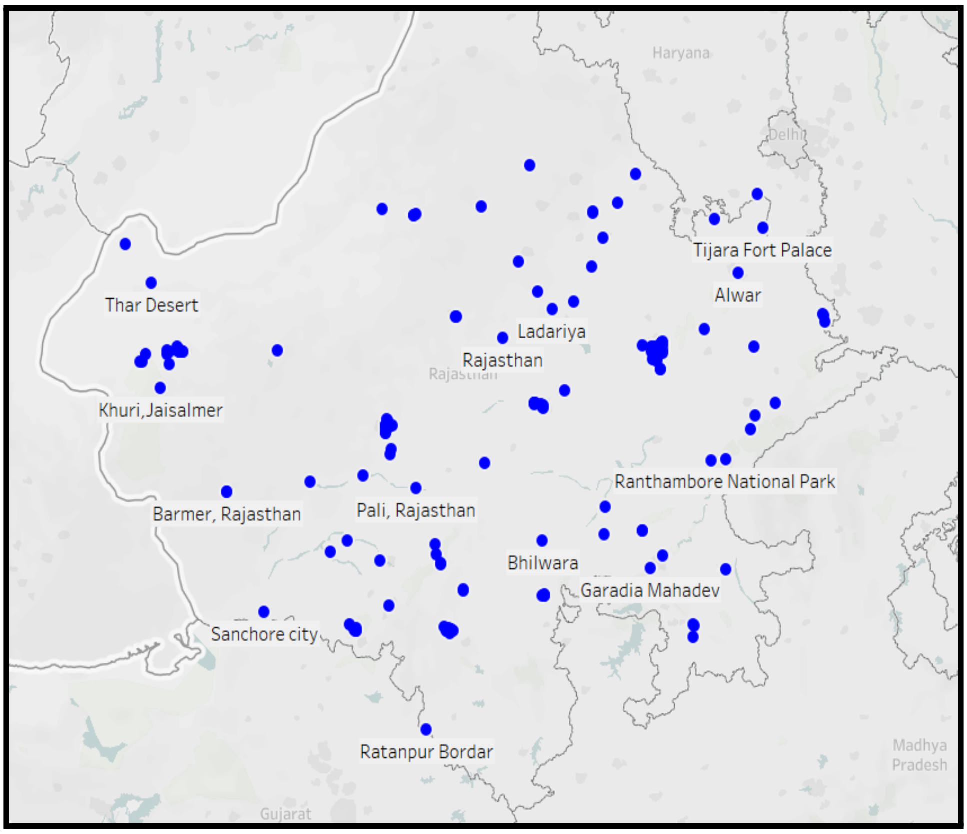
1. Unique Locations

Shown below are the unique locations visited by the sample of people considered by us during the period 2017-2021.

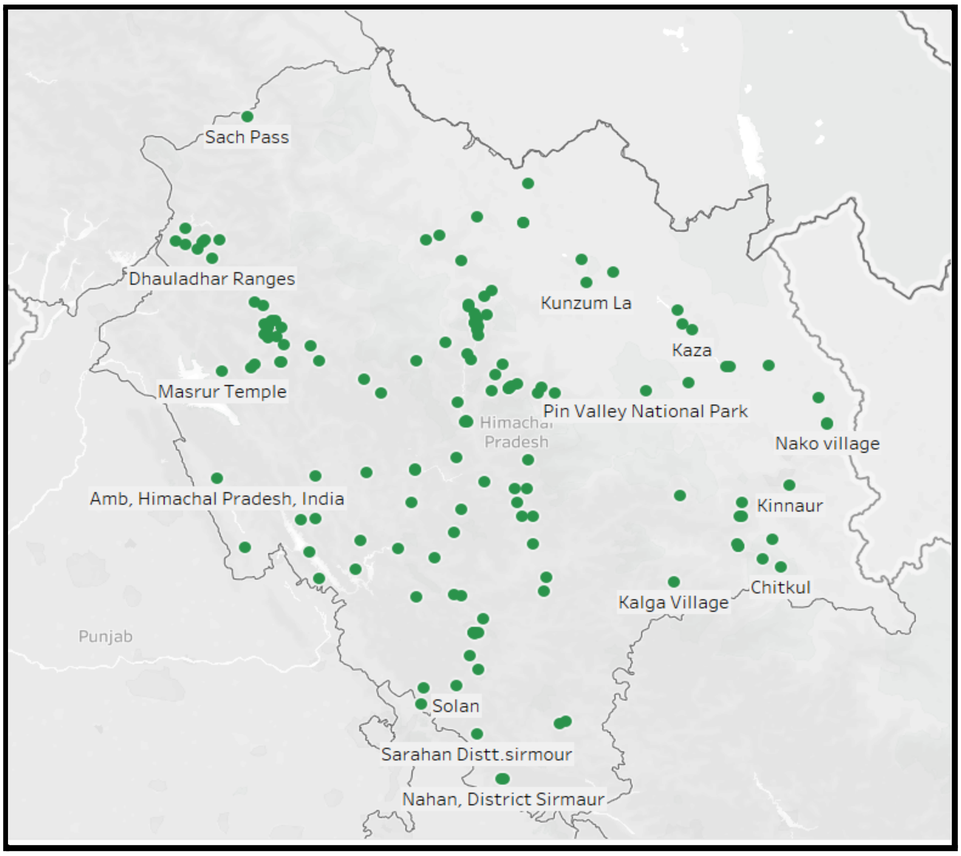


7.1 Goa - unique locations

7.2. Rajasthan - unique locations



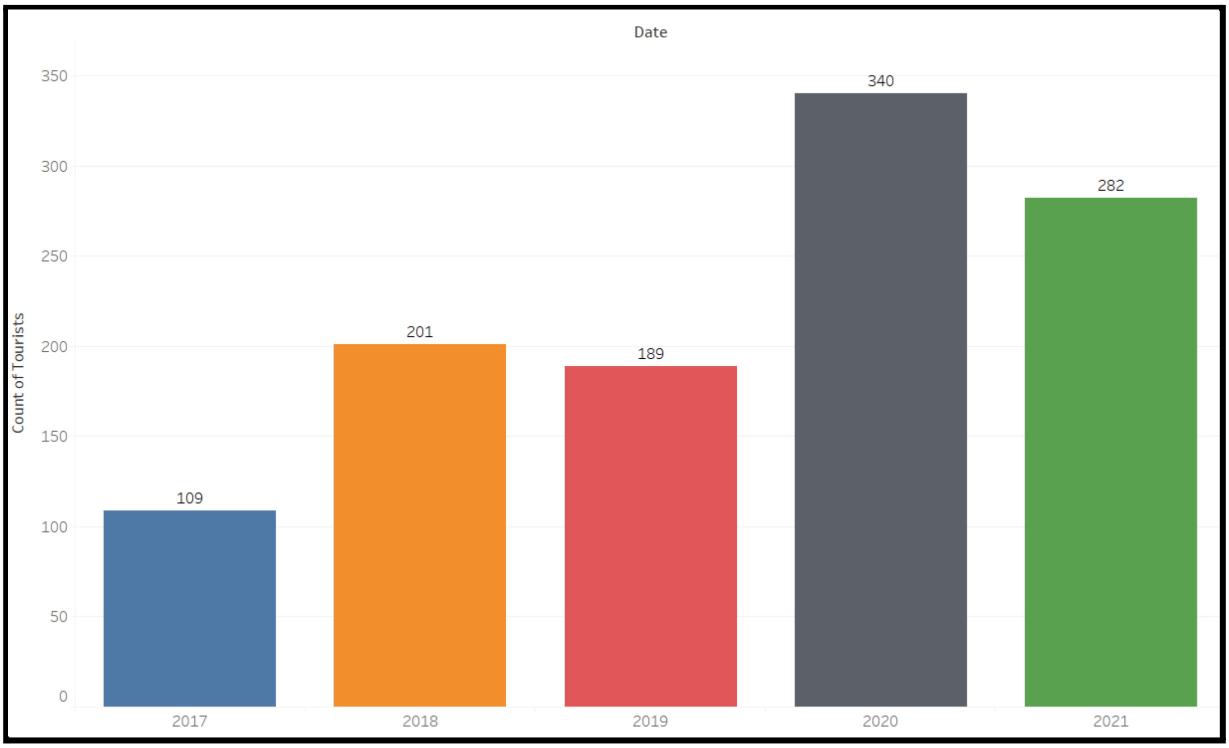
7.3. Himachal Pradesh - unique locations



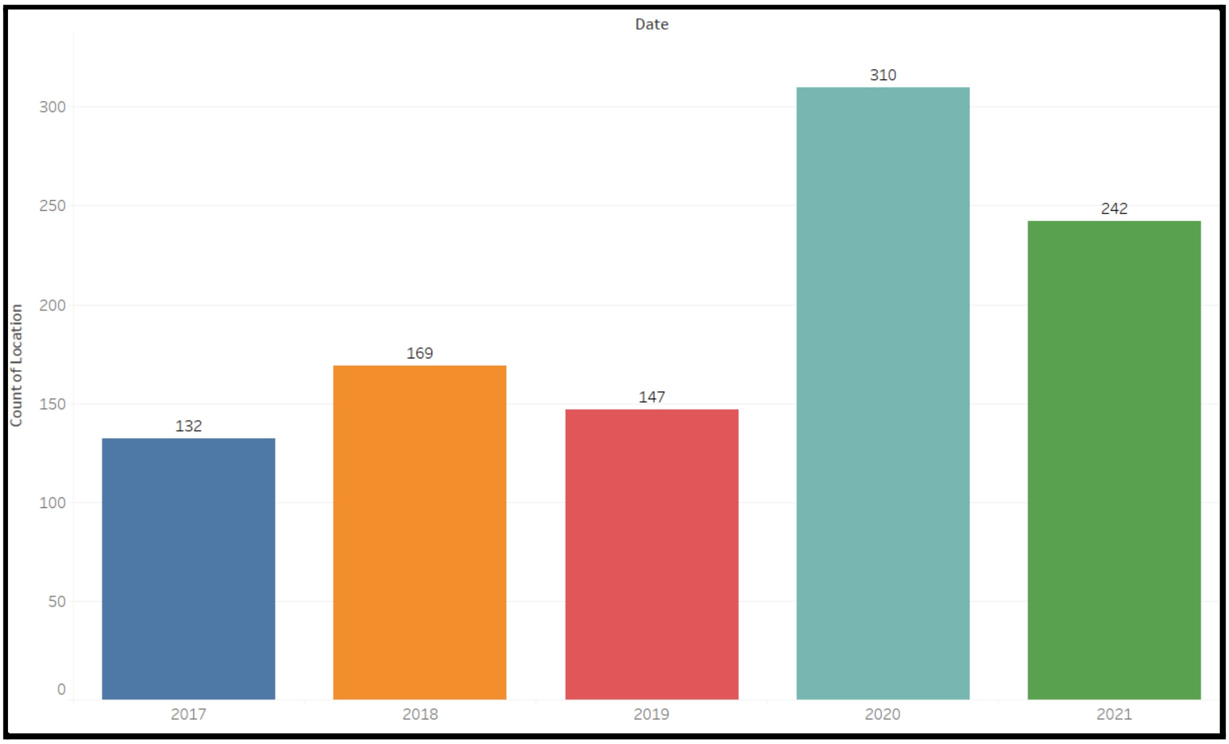
2. **People touring these states per year**

Shown below are the graphs which help us see the distribution of our sample over the spread of 5 years(our selected timeline).

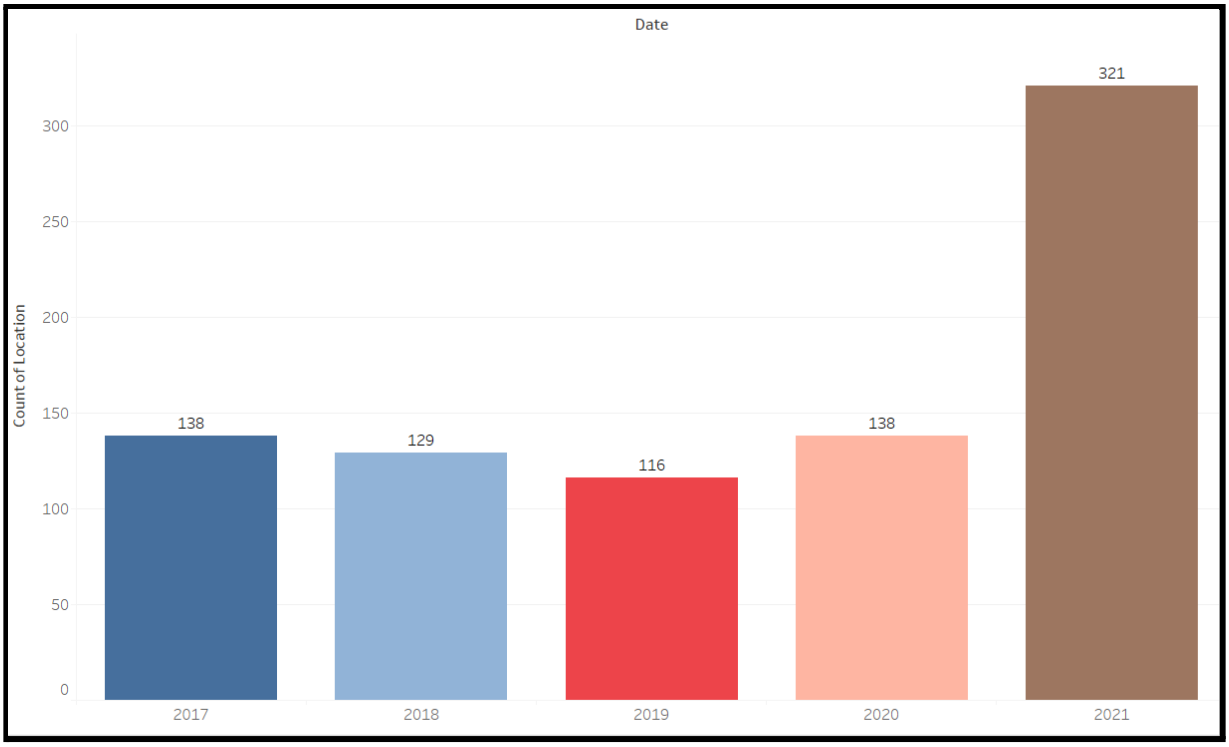
7.4 GOA



7.5 RAJASTHAN

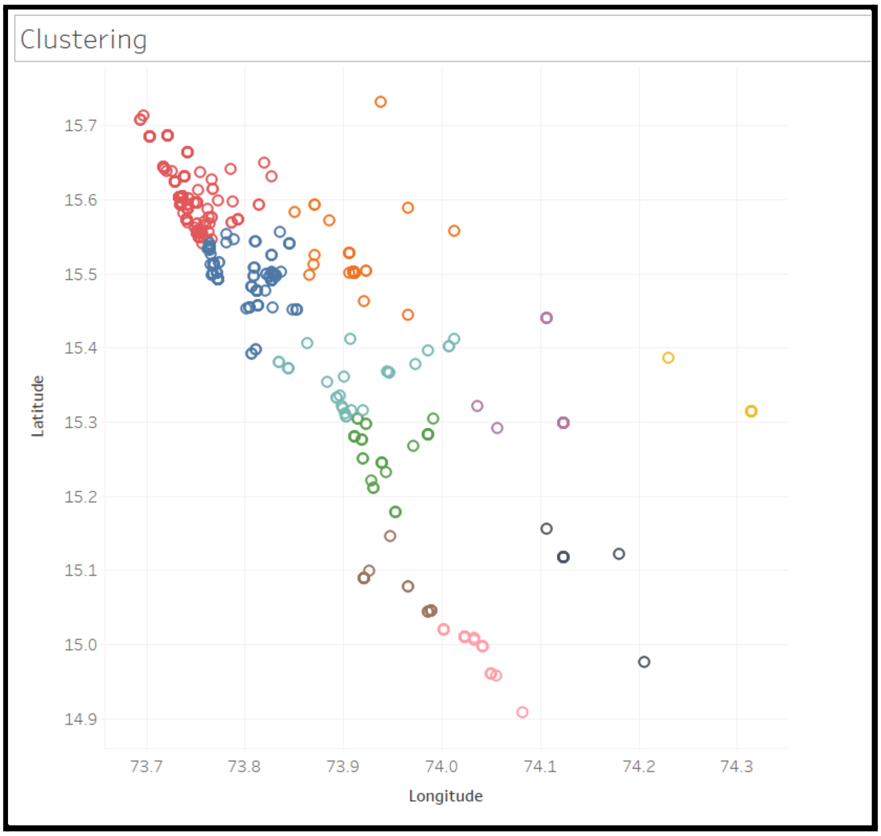


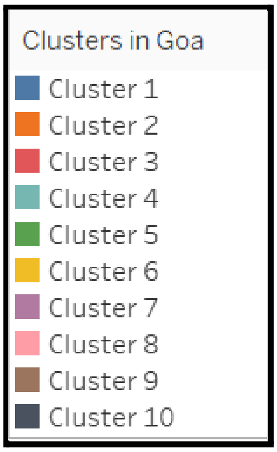
7.6 HIMACHAL



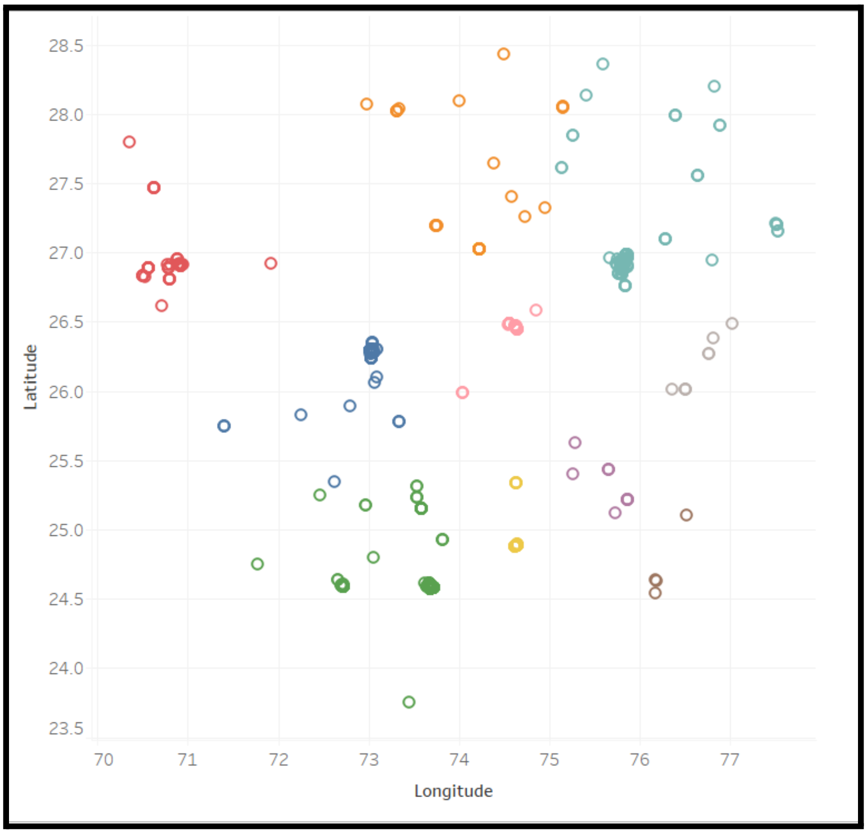
3. Clustering

We applied the K-Means Clustering algorithm to the locations we got from the data to spot high tourism density locations/areas in the states. For each state we made 10 clusters based on the latitudes and longitudes of the locations.

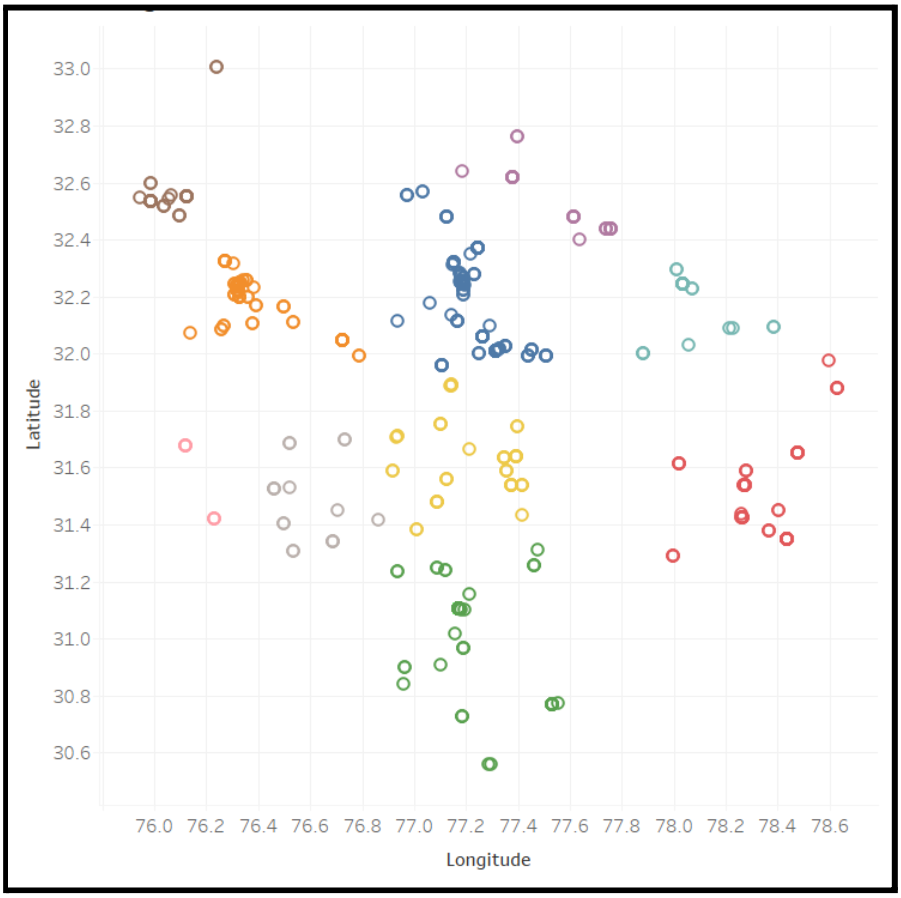




7.8 Goa



7.9 Rajasthan

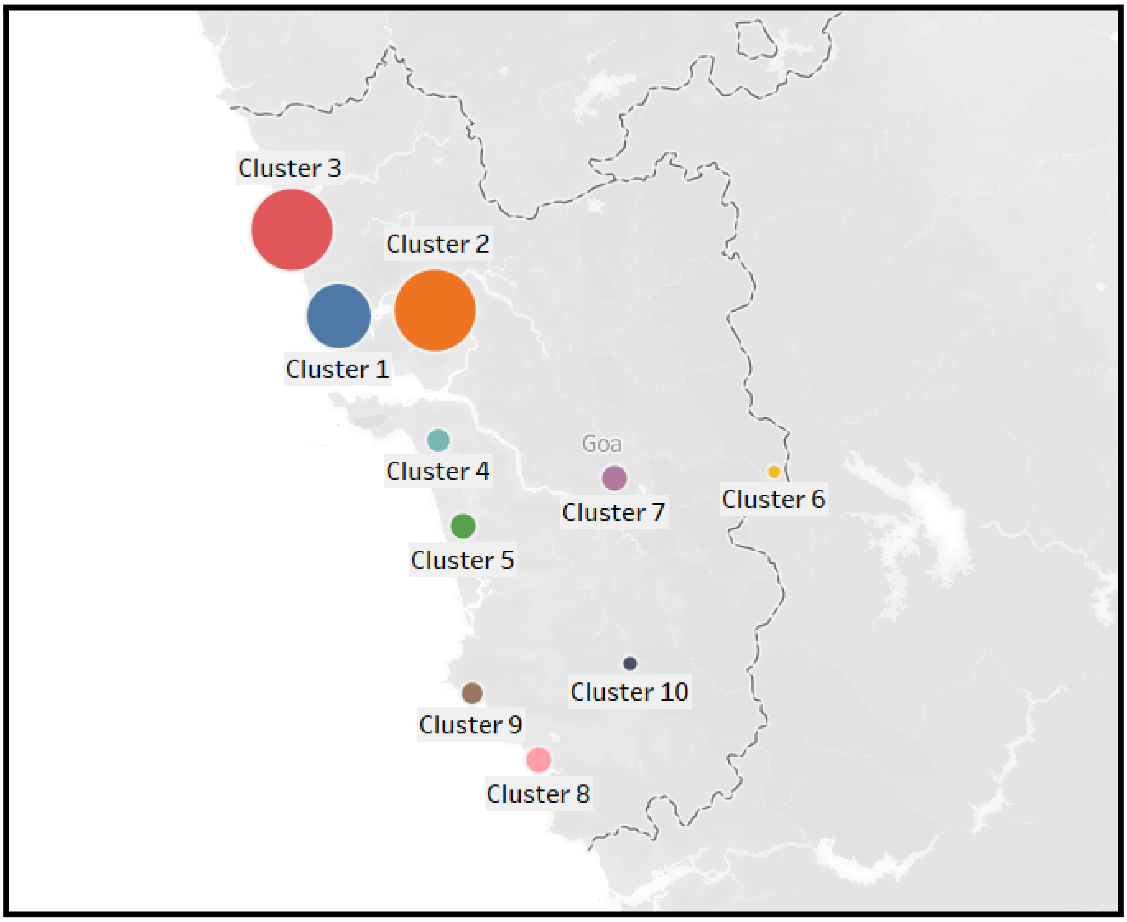


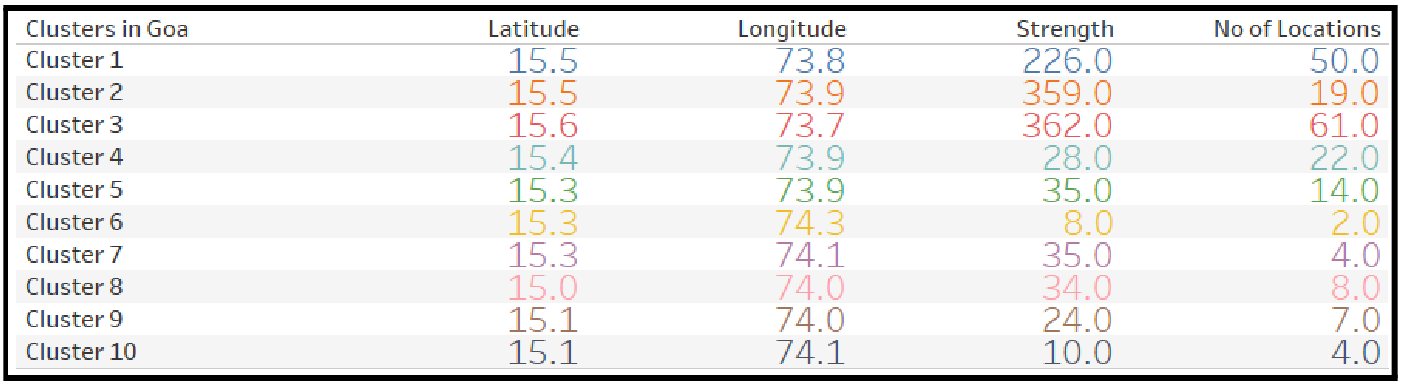
7.10Himachal

4. Plotting of Clusters and Cluster details

We plotted the clusters made on the respective state maps by their cluster centers to identify on a general level the areas of the state that experience the maximum number of tourists. We also found the number of unique locations included in that cluster and tourist density of that cluster out of our sample population.

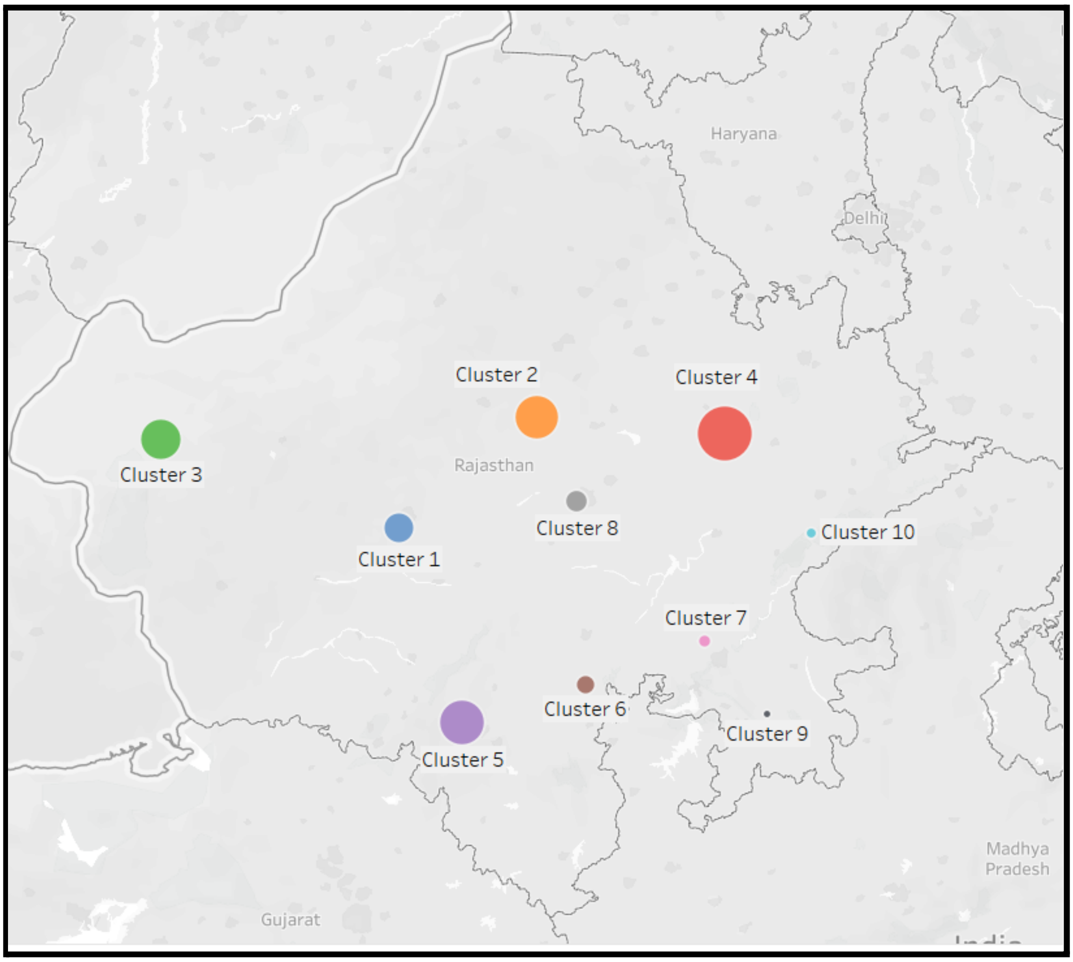
7-10 .GOA

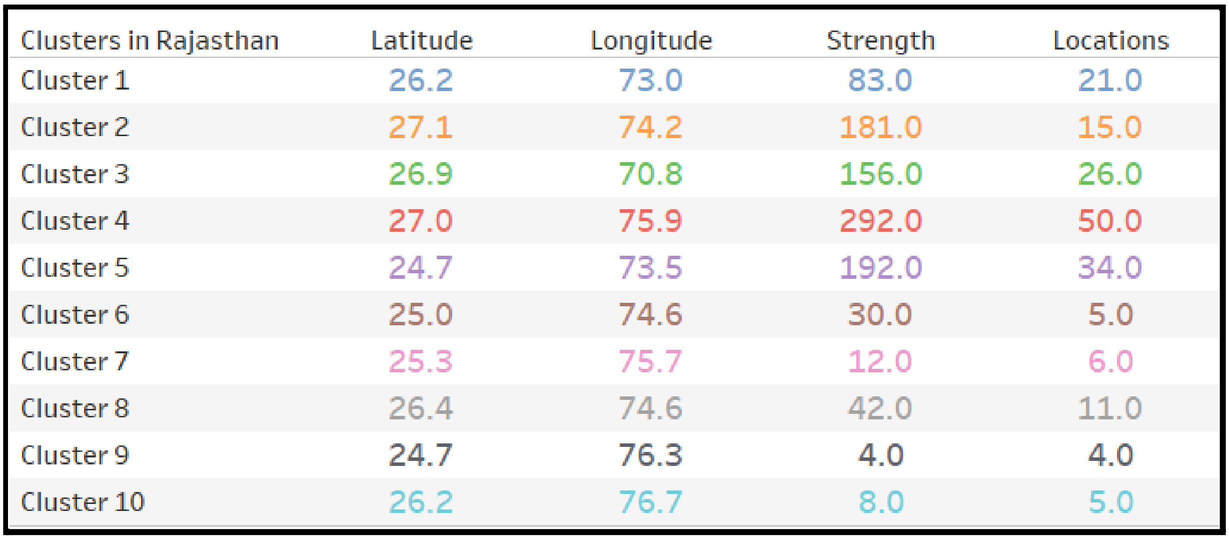




Clusters in Goa	Latitude	Longitude	Strength	No of Locations
Cluster 1	15.5	73.8	226.0	50.0
Cluster 2	15.5	73.9	359.0	19.0
Cluster 3	15.6	73.7	362.0	61.0
Cluster 4	15.4	73.9	28.0	22.0
Cluster 5	15.3	73.9	35.0	14.0
Cluster 6	15.3	74.3	8.0	2.0
Cluster 7	15.3	74.1	35.0	4.0
Cluster 8	15.0	74.0	34.0	8.0
Cluster 9	15.1	74.0	24.0	7.0
Cluster 10	15.1	74.1	10.0	4.0

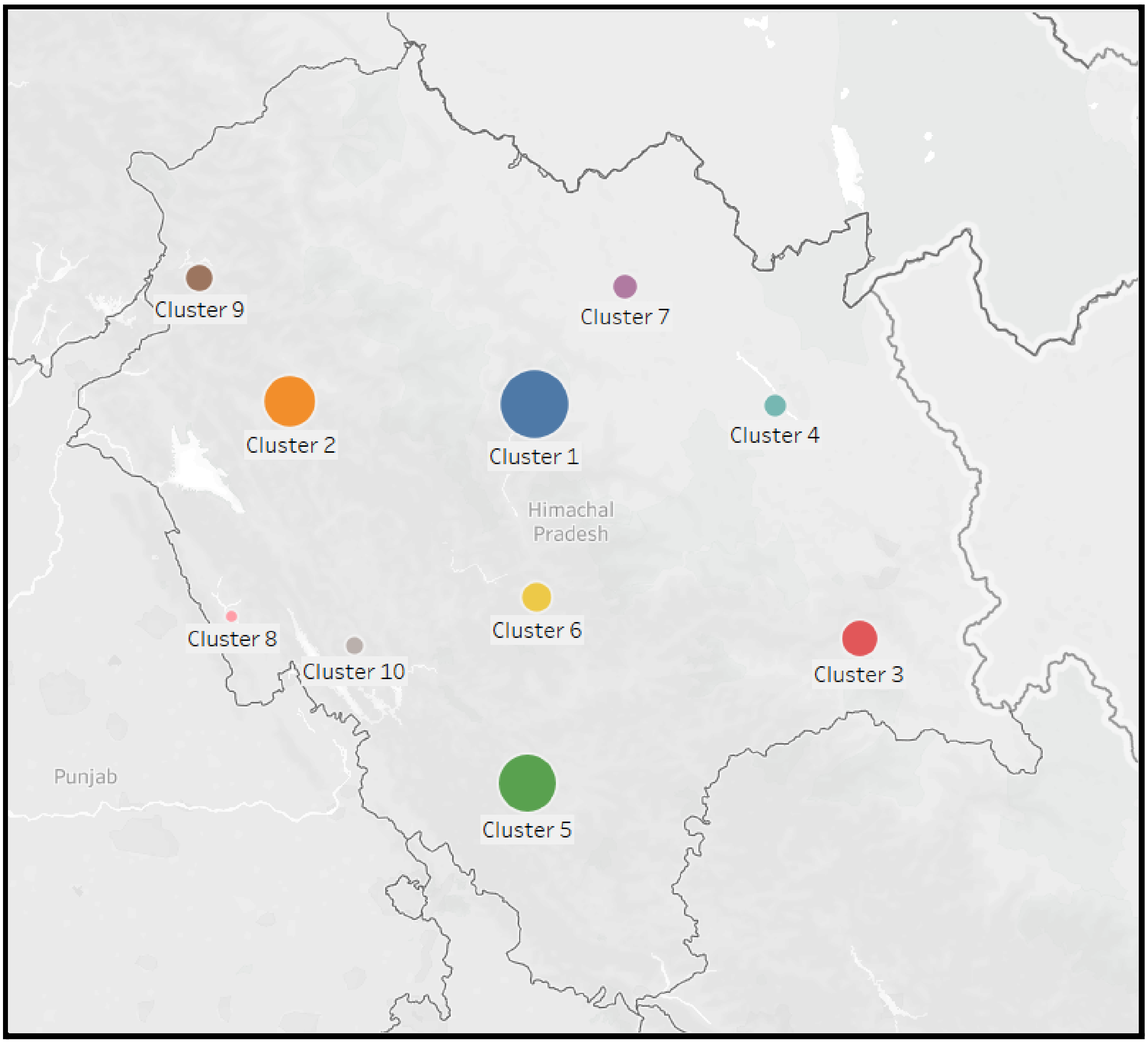
7-11. RAJASTHAN

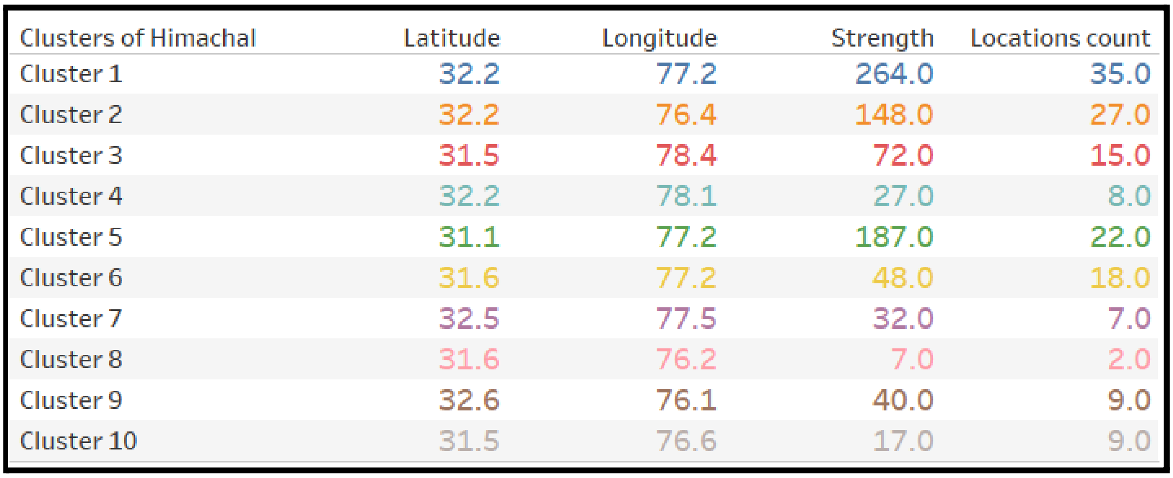




Clusters in Rajasthan	Latitude	Longitude	Strength	Locations
Cluster 1	26.2	73.0	83.0	21.0
Cluster 2	27.1	74.2	181.0	15.0
Cluster 3	26.9	70.8	156.0	26.0
Cluster 4	27.0	75.9	292.0	50.0
Cluster 5	24.7	73.5	192.0	34.0
Cluster 6	25.0	74.6	30.0	5.0
Cluster 7	25.3	75.7	12.0	6.0
Cluster 8	26.4	74.6	42.0	11.0
Cluster 9	24.7	76.3	4.0	4.0
Cluster 10	26.2	76.7	8.0	5.0

7-12 . HIMACHAL



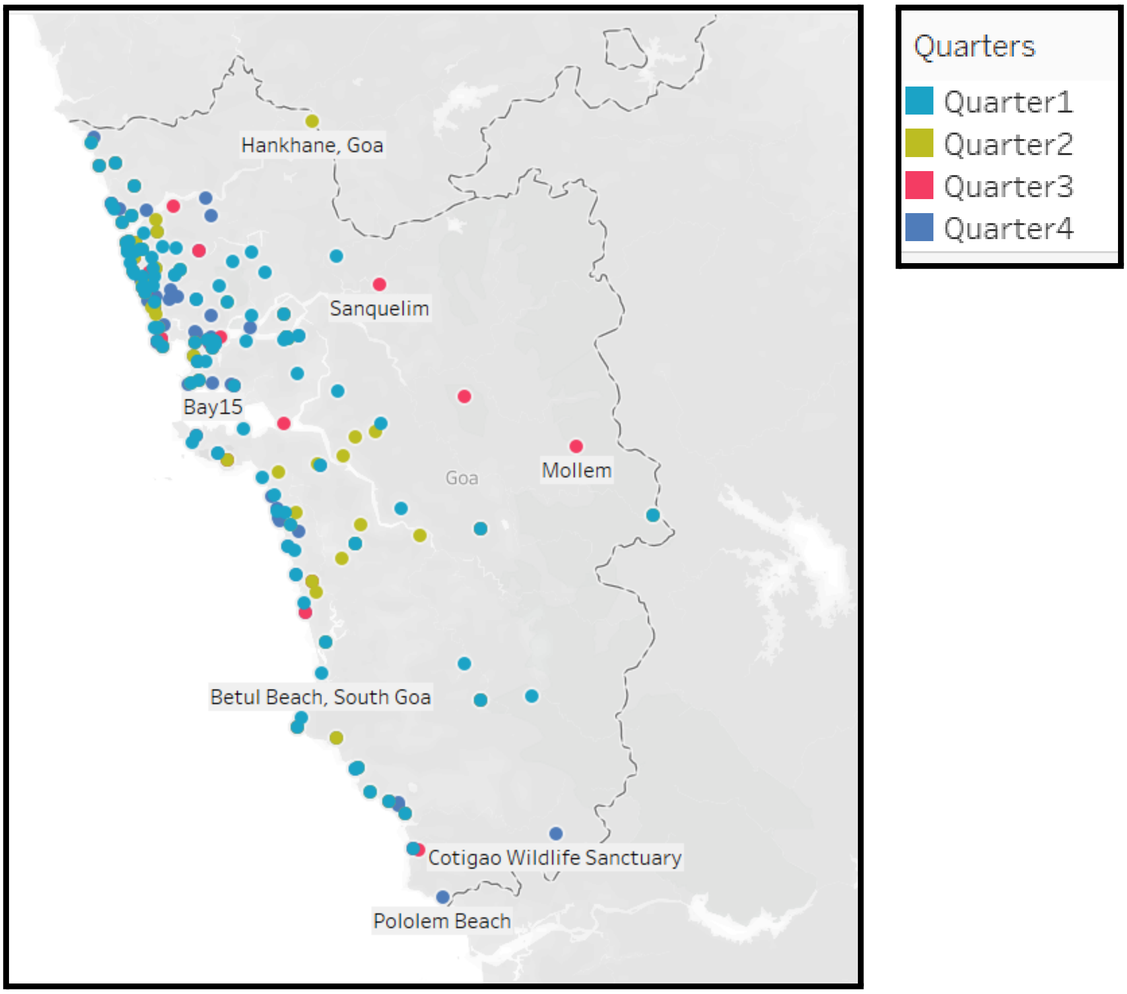


Clusters of Himachal	Latitude	Longitude	Strength	Locations count
Cluster 1	32.2	77.2	264.0	35.0
Cluster 2	32.2	76.4	148.0	27.0
Cluster 3	31.5	78.4	72.0	15.0
Cluster 4	32.2	78.1	27.0	8.0
Cluster 5	31.1	77.2	187.0	22.0
Cluster 6	31.6	77.2	48.0	18.0
Cluster 7	32.5	77.5	32.0	7.0
Cluster 8	31.6	76.2	7.0	2.0
Cluster 9	32.6	76.1	40.0	9.0
Cluster 10	31.5	76.6	17.0	9.0

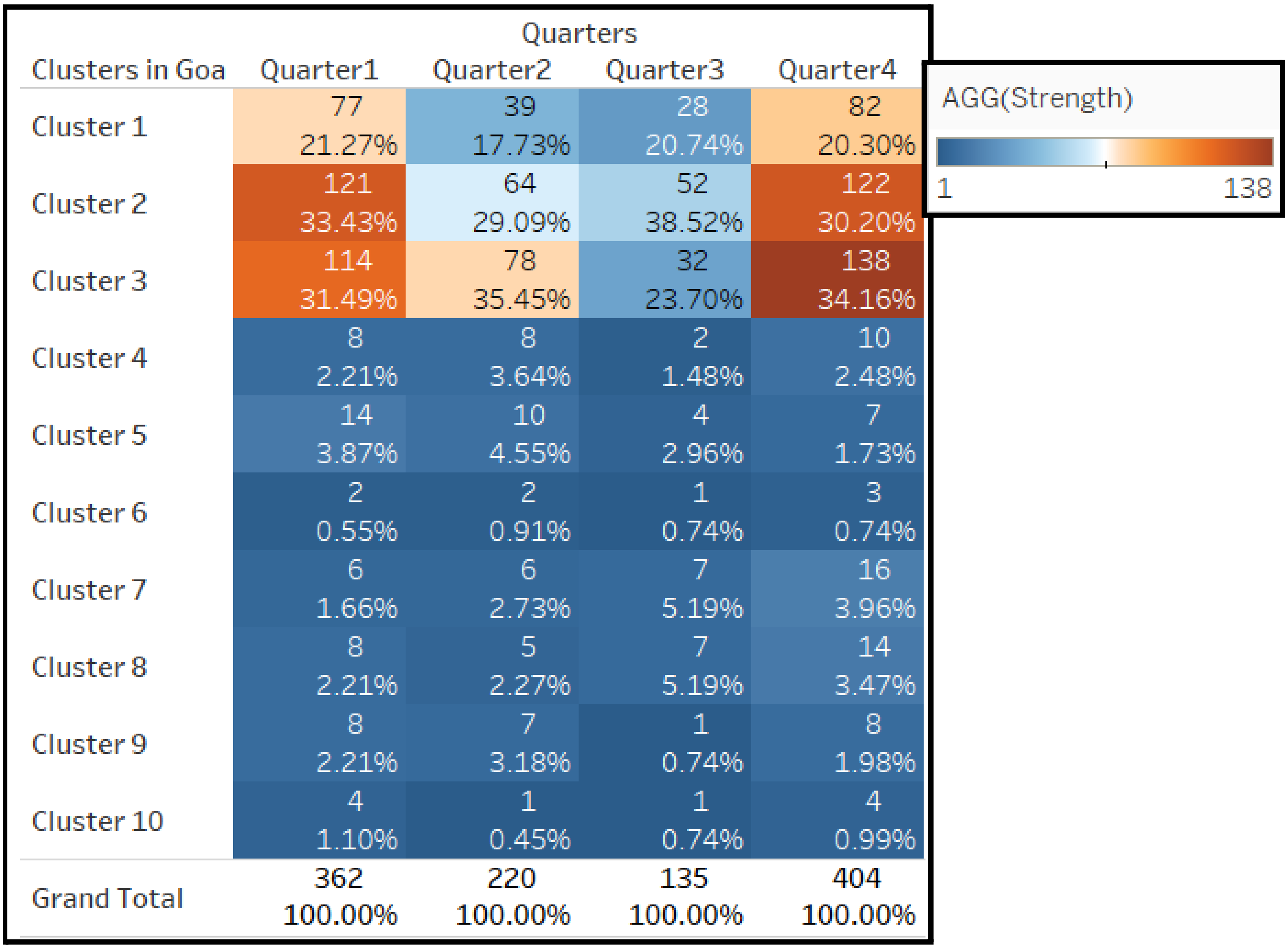
CHAPTER 8
RESULTS AND DISCUSSIONS

Based on our experimentation and analysis, we came to a few conclusions regarding the tourism trends in the respective states. We tried to implement a quarterly analysis on the cluster centers of each state and have tried to support our conclusions with related visualizations. We will be discussing each state one by one below.

8.1 GOA

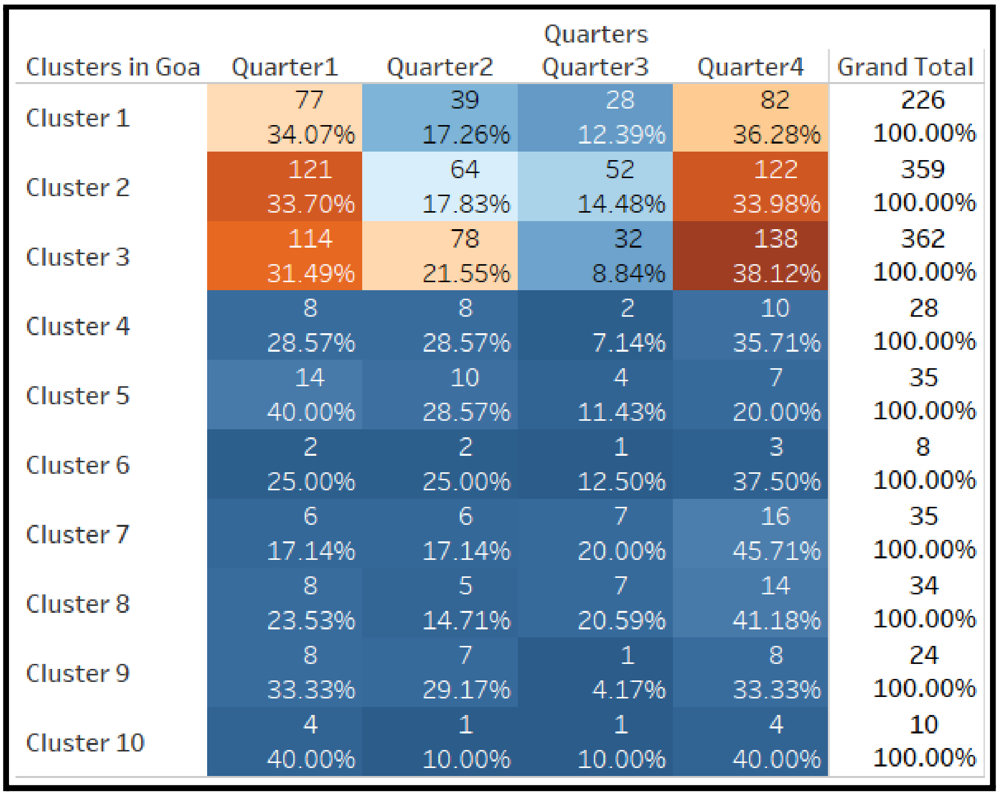


In the above image, we can see the spread of tourists of our sample over the state, classified by the quarter of the year in which they visited the state of Goa.

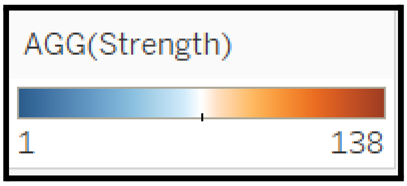


	Quarters			
Clusters in Goa	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 1	77 21.27%	39 17.73%	28 20.74%	82 20.30%
Cluster 2	121 33.43%	64 29.09%	52 38.52%	122 30.20%
Cluster 3	114 31.49%	78 35.45%	32 23.70%	138 34.16%
Cluster 4	8 2.21%	8 3.64%	2 1.48%	10 2.48%
Cluster 5	14 3.87%	10 4.55%	4 2.96%	7 1.73%
Cluster 6	2 0.55%	2 0.91%	1 0.74%	3 0.74%
Cluster 7	6 1.66%	6 2.73%	7 5.19%	16 3.96%
Cluster 8	8 2.21%	5 2.27%	7 5.19%	14 3.47%
Cluster 9	8 2.21%	7 3.18%	1 0.74%	8 1.98%
Cluster 10	4 1.10%	1 0.45%	1 0.74%	4 0.99%
Grand Total	362 100.00%	220 100.00%	135 100.00%	404 100.00%

The above highlight table gives us an insight about the clusterwise analysis in each quarter and also indicates which quarter of the year has been most popular for tourism in the state. Here we can see that Quarter1 and Quarter4 are the most popular quarters for touring Goa and in Goa, the places which fall under cluster 1, cluster 2 and cluster 3 are the popular tourist places in the state. These clusters include beaches of north Goa, party locations etc. which are extremely popular among tourists. This highlight table also shows the column wise percent totals that help in comparing clusters with each other. Thus, Goa is always expected to have more number of people touring North Goa than South Goa which can be clearly seen from the analysis of the sample population we have taken.



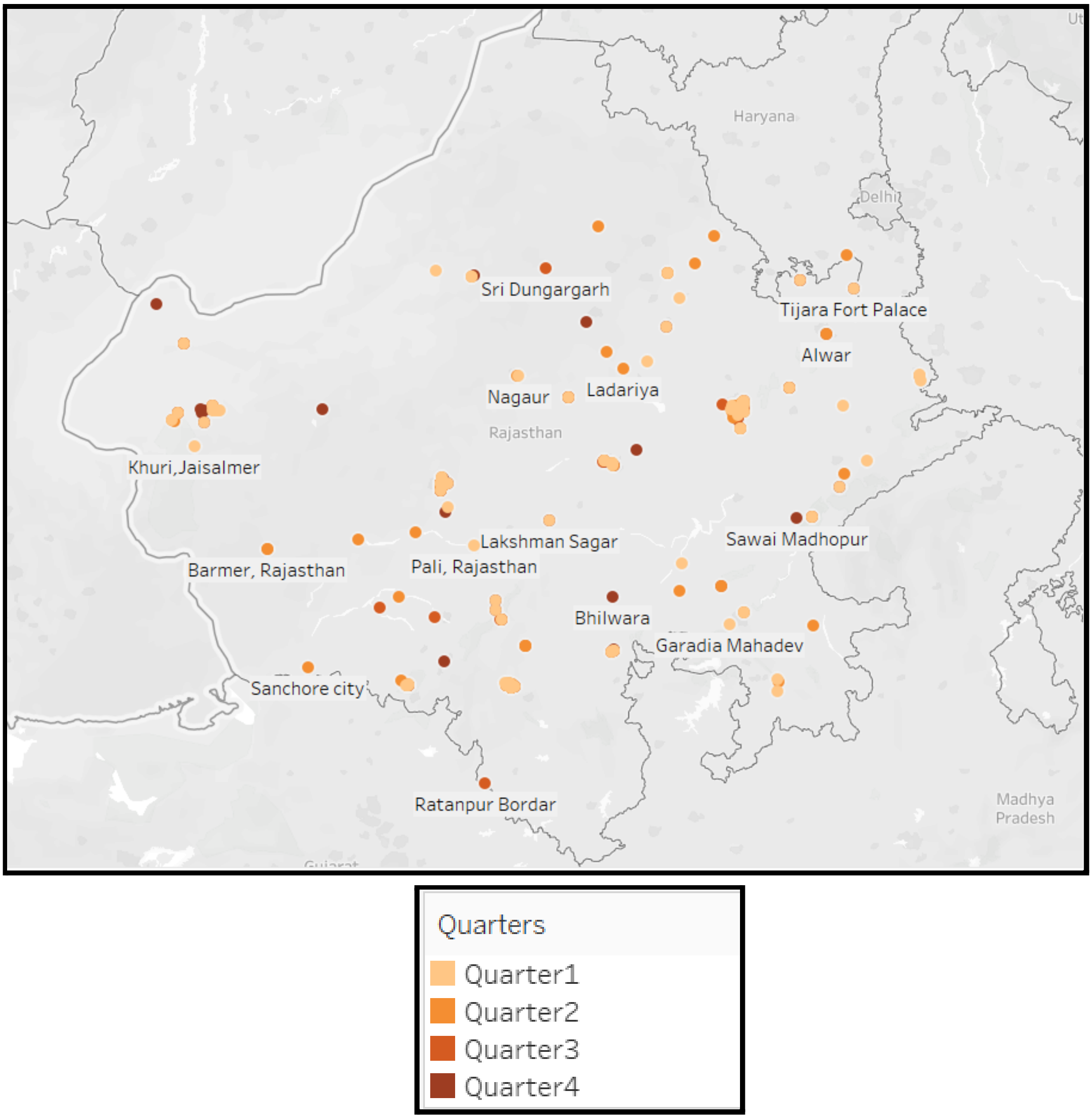
	Quarters				
Clusters in Goa	Quarter1	Quarter2	Quarter3	Quarter4	Grand Total
Cluster 1	77 34.07%	39 17.26%	28 12.39%	82 36.28%	226 100.00%
Cluster 2	121 33.70%	64 17.83%	52 14.48%	122 33.98%	359 100.00%
Cluster 3	114 31.49%	78 21.55%	32 8.84%	138 38.12%	362 100.00%
Cluster 4	8 28.57%	8 28.57%	2 7.14%	10 35.71%	28 100.00%
Cluster 5	14 40.00%	10 28.57%	4 11.43%	7 20.00%	35 100.00%
Cluster 6	2 25.00%	2 25.00%	1 12.50%	3 37.50%	8 100.00%
Cluster 7	6 17.14%	6 17.14%	7 20.00%	16 45.71%	35 100.00%
Cluster 8	8 23.53%	5 14.71%	7 20.59%	14 41.18%	34 100.00%
Cluster 9	8 33.33%	7 29.17%	1 4.17%	8 33.33%	24 100.00%
Cluster 10	4 40.00%	1 10.00%	1 10.00%	4 40.00%	10 100.00%



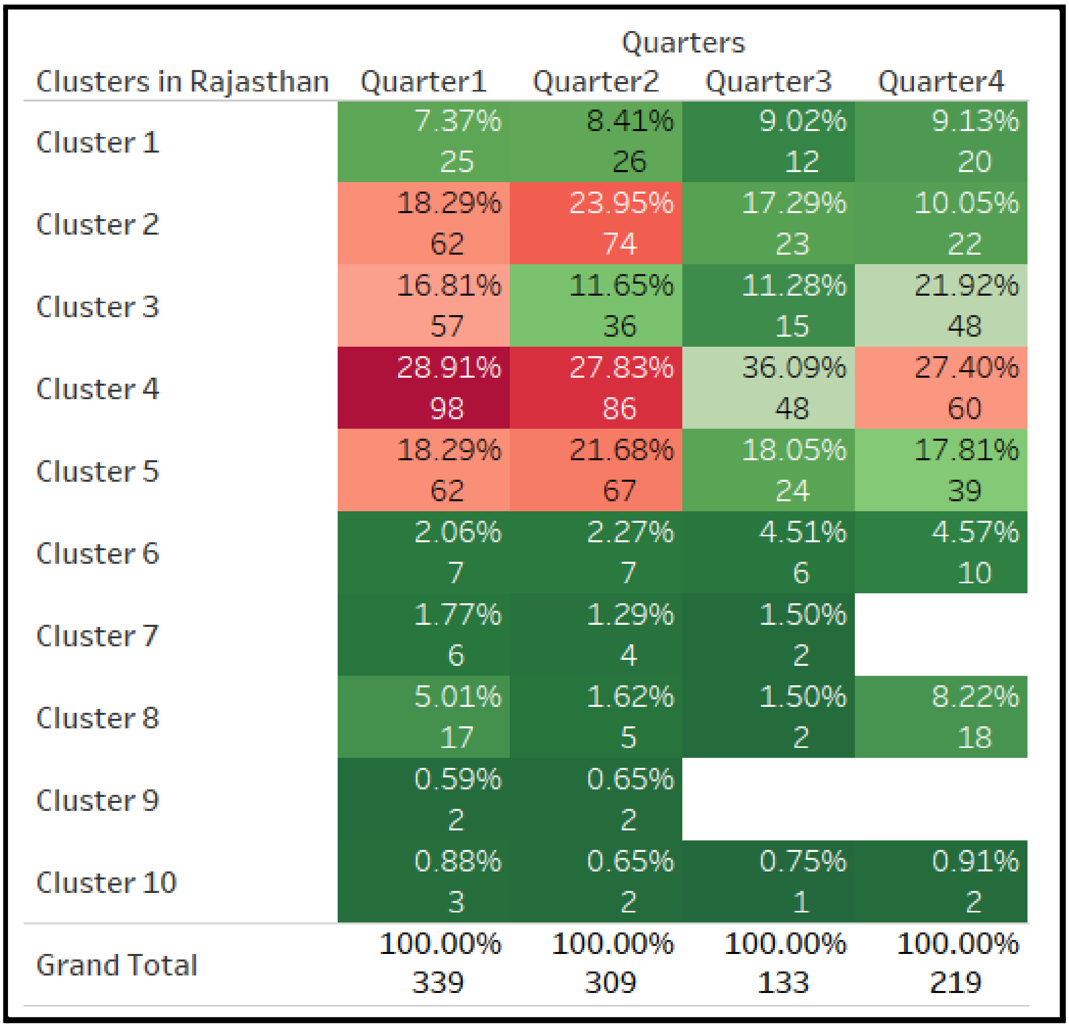
The above highlight table gives us an insight about the quarterwise analysis in each cluster This highlight table also shows the row wise percent totals that help in comparing quarters with each other. It supports all the results which we got using the previous highlight table but is important as it helps us look at the data from another perspective, like a pivot operation on the datacube.

Thus, Goa is always expected to have more number of people touring North Goa than South Goa which can be clearly seen from the analysis of the sample population we have taken.

8.2 RAJASTHAN



In the above image, we can see the spread of tourists of our sample over the state, classified by the quarter of the year in which they visited the state of Rajasthan.

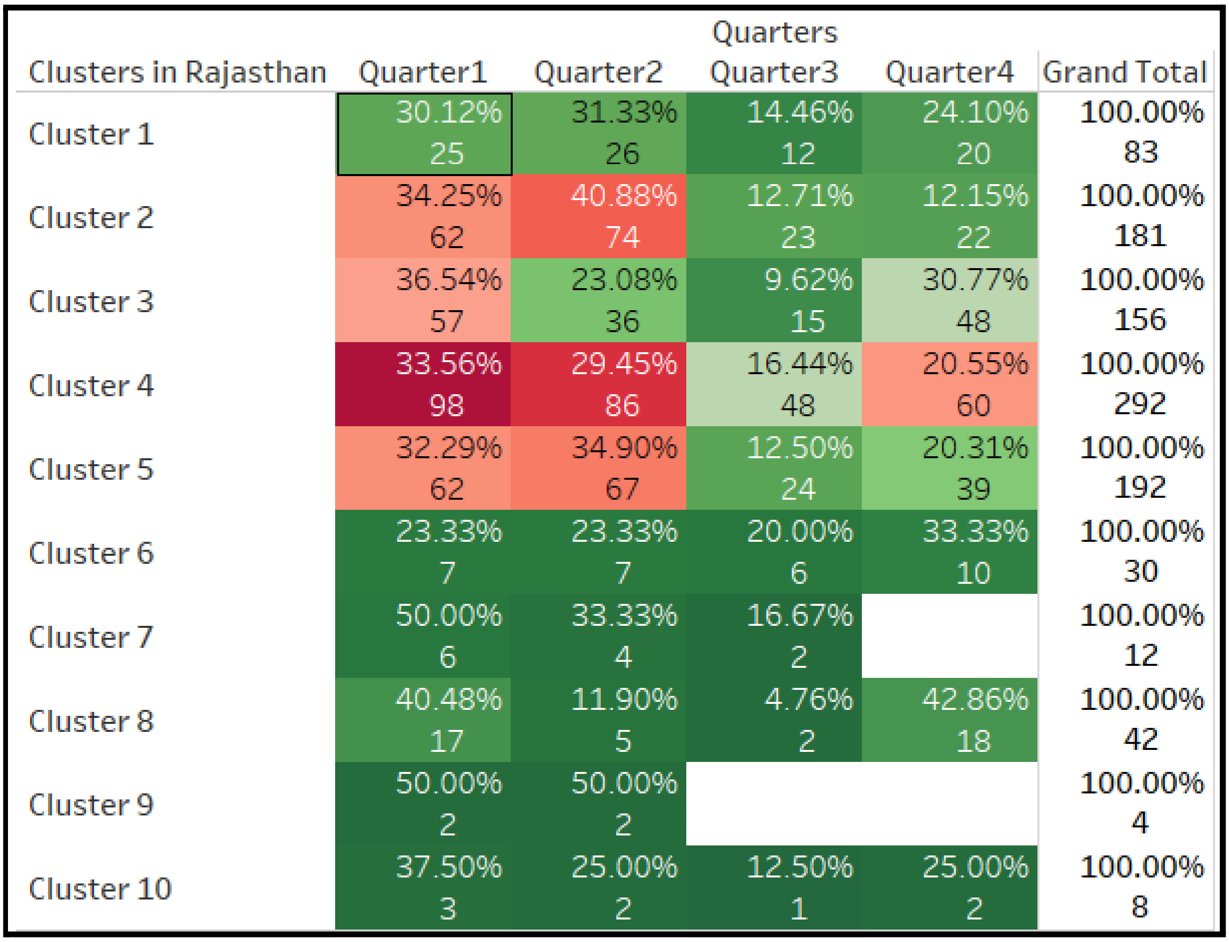


	Quarters			
Clusters in Rajasthan	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 1	7.37% 25	8.41% 26	9.02% 12	9.13% 20
Cluster 2	18.29% 62	23.95% 74	17.29% 23	10.05% 22
Cluster 3	16.81% 57	11.65% 36	11.28% 15	21.92% 48
Cluster 4	28.91% 98	27.83% 86	36.09% 48	27.40% 60
Cluster 5	18.29% 62	21.68% 67	18.05% 24	17.81% 39
Cluster 6	2.06% 7	2.27% 7	4.51% 6	4.57% 10
Cluster 7	1.77% 6	1.29% 4	1.50% 2	
Cluster 8	5.01% 17	1.62% 5	1.50% 2	8.22% 18
Cluster 9	0.59% 2	0.65% 2		
Cluster 10	0.88% 3	0.65% 2	0.75% 1	0.91% 2
Grand Total	100.00% 339	100.00% 309	100.00% 133	100.00% 219

The above highlight table gives us an insight about the clusterwise analysis in each quarter and also indicates which quarter of the year has been most popular for tourism in the state. Here we can see that Quarter1 and Quarter2 are the most popular quarters for touring Rajasthan and in Rajasthan, the places which fall under cluster 2, cluster 4 and cluster 5 are the popular tourist places in the state. These clusters include udaipur, amer fort, jaipur, chawki dhani etc. which are extremely popular among tourists. This highlight table also shows the column wise percent totals that help in comparing clusters with each other.

There are certain blanks in the table that indicates that nobody visited locations in that cluster during that quarter from the sample of people we have taken.

Thus, Rajasthan has lots of tourists touring jaipur, udaipur etc which can be clearly seen from the analysis of the sample population we have taken.

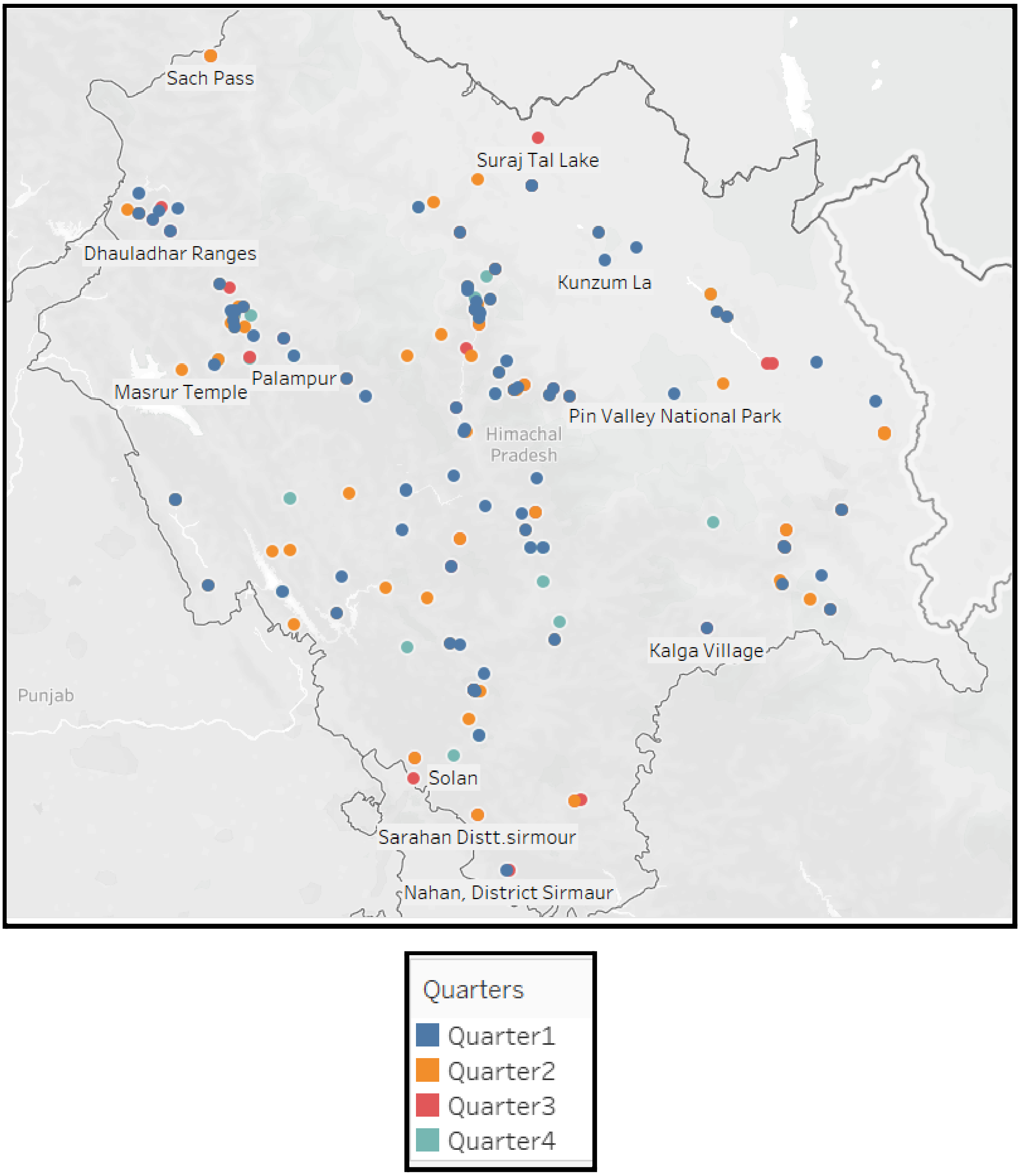


	Quarters				
Clusters in Rajasthan	Quarter1	Quarter2	Quarter3	Quarter4	Grand Total
Cluster 1	30.12% 25	31.33% 26	14.46% 12	24.10% 20	100.00% 83
Cluster 2	34.25% 62	40.88% 74	12.71% 23	12.15% 22	100.00% 181
Cluster 3	36.54% 57	23.08% 36	9.62% 15	30.77% 48	100.00% 156
Cluster 4	33.56% 98	29.45% 86	16.44% 48	20.55% 60	100.00% 292
Cluster 5	32.29% 62	34.90% 67	12.50% 24	20.31% 39	100.00% 192
Cluster 6	23.33% 7	23.33% 7	20.00% 6	33.33% 10	100.00% 30
Cluster 7	50.00% 6	33.33% 4	16.67% 2		100.00% 12
Cluster 8	40.48% 17	11.90% 5	4.76% 2	42.86% 18	100.00% 42
Cluster 9	50.00% 2	50.00% 2			100.00% 4
Cluster 10	37.50% 3	25.00% 2	12.50% 1	25.00% 2	100.00% 8

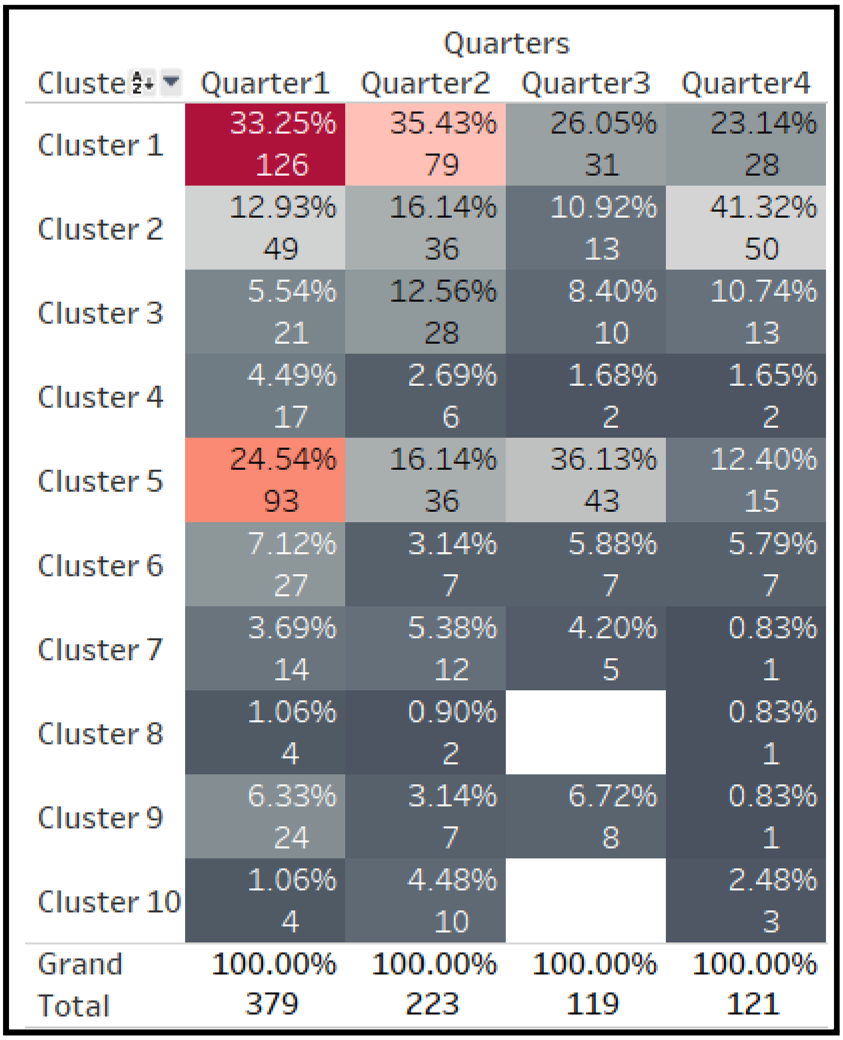
The above highlight table gives us an insight about the quarterwise analysis in each cluster This highlight table also shows the row wise percent totals that help in comparing quarters with each other. It supports all the results which we got using the previous highlight table but is important as it helps us look at the data from another perspective, like a pivot operation on the datacube.

Rajasthan has lots of people touring central and southern Rajasthan which has places like jaipur, udaipur etc. which can be clearly seen from the analysis of the sample population we have taken.

8.3 HIMACHAL PRADESH



In the above image, we can see the spread of tourists of our sample over the state, classified by the quarter of the year in which they visited the state of Himachal Pradesh.

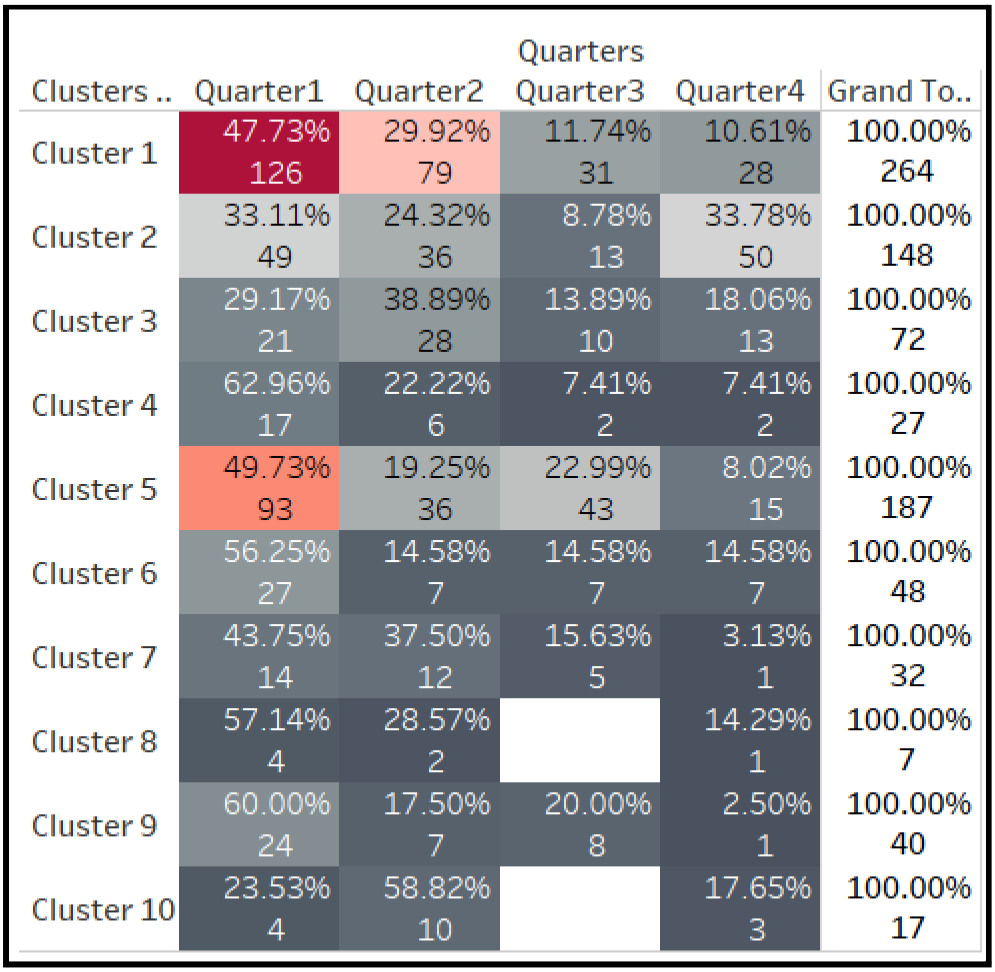


	Quarters			
Cluste	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 1	33.25% 126	35.43% 79	26.05% 31	23.14% 28
Cluster 2	12.93% 49	16.14% 36	10.92% 13	41.32% 50
Cluster 3	5.54% 21	12.56% 28	8.40% 10	10.74% 13
Cluster 4	4.49% 17	2.69% 6	1.68% 2	1.65% 2
Cluster 5	24.54% 93	16.14% 36	36.13% 43	12.40% 15
Cluster 6	7.12% 27	3.14% 7	5.88% 7	5.79% 7
Cluster 7	3.69% 14	5.38% 12	4.20% 5	0.83% 1
Cluster 8	1.06% 4	0.90% 2		0.83% 1
Cluster 9	6.33% 24	3.14% 7	6.72% 8	0.83% 1
Cluster 10	1.06% 4	4.48% 10		2.48% 3
Grand Total	100.00% 379	100.00% 223	100.00% 119	100.00% 121

The above highlight table gives us an insight about the clusterwise analysis in each quarter and also indicates which quarter of the year has been most popular for tourism in the state. Here we can see that Quarter1 and Quarter2 are the most popular quarters for touring Himachal Pradesh and in Himachal Pradesh, the places which fall under cluster 1, cluster 2 and cluster 5 are the popular tourist places in the state. These clusters include shimla, kulu, manali, dharamshala, mcleodganj etc. which are extremely popular among tourists. This highlight table also shows the column wise percent totals that help in comparing clusters with each other.

There are certain blanks in the table that indicates that nobody visited locations in that cluster during that quarter from the sample of people we have taken.

When says he is going to Himachal Pradesh, the top tourist places that come to our mind are the ones which can be clearly seen from the analysis of the sample population we have taken.



			Quarters		
Clusters ..	Quarter1	Quarter2	Quarter3	Quarter4	Grand To..
Cluster 1	47.73% 126	29.92% 79	11.74% 31	10.61% 28	100.00% 264
Cluster 2	33.11% 49	24.32% 36	8.78% 13	33.78% 50	100.00% 148
Cluster 3	29.17% 21	38.89% 28	13.89% 10	18.06% 13	100.00% 72
Cluster 4	62.96% 17	22.22% 6	7.41% 2	7.41% 2	100.00% 27
Cluster 5	49.73% 93	19.25% 36	22.99% 43	8.02% 15	100.00% 187
Cluster 6	56.25% 27	14.58% 7	14.58% 7	14.58% 7	100.00% 48
Cluster 7	43.75% 14	37.50% 12	15.63% 5	3.13% 1	100.00% 32
Cluster 8	57.14% 4	28.57% 2		14.29% 1	100.00% 7
Cluster 9	60.00% 24	17.50% 7	20.00% 8	2.50% 1	100.00% 40
Cluster 10	23.53% 4	58.82% 10		17.65% 3	100.00% 17

The above highlight table gives us an insight about the quarterwise analysis in each cluster. This highlight table also shows the row wise percent totals that help in comparing quarters with each other. It supports all the results which we got using the previous highlight table but is important as it helps us look at the data from another perspective, like a pivot operation on the datacube.

Himachal Pradesh has lots of people touring central, western and southern Himachal Pradesh which has places like shimla, kullu, manali, dharamsala, mcleodganj etc. which can be clearly seen from the analysis of the sample population we have taken.

CHAPTER 9

CONCLUSIONS AND FUTURE SCOPE

This project is based on the behaviour analysis of tourists which focuses on the frequent patterns in their visits. Our aim to help the tourism industry revive, which suffered huge losses in the pandemic last year, has been fulfilled. The analysis presented is constructed from every minute data available of the previously visited tourists.

For this mini project, we extracted the data of users and their geotagged locations from Instagram. A dataset was created from this extracted data. Simple Random Sampling was carried out on the available data. Data preprocessing was performed in order to get clean data for analysis. We performed several visualizations to understand the distinct locations visited by tourists in a state, frequency of visiting a location, and popular destinations. KMeans clustering was applied on the data which showed 10 clusters of the most popular locations along with their strength. In addition, we scrutinized various research papers published in this domain.

Our analysis presents that

- Goa is always expected to have more number of people touring North Goa than South Goa and can expect maximum crowds in Quarter 4.
- Rajasthan has lots of people touring central and southern Rajasthan which has places like jaipur, udaipur and can expect maximum crowds in Quarter 1.
- Himachal Pradesh has lots of people touring central, western and southern Himachal Pradesh and can expect maximum crowds in Quarter 1 and 2.

The future scope of this project includes construction of a comprehensive review system with the help of social media using image and text processing which can be forwarded to authorities to improve the tourism in that area. We would also like to develop an application where a user can enter the name of an area and similar data can be provided for the same like the top locations and the best time to visit with all the data extracted from social media.

REFERENCES

[1]Bindiya Chari, “Goa tourism to conduct survey to assess losses of hospitality industry”, May 7, 2020 https://timesofindia.indiatimes.com/city/goa/goa-tourism-to-conduct-survey-to-assess-losses-of-hospitality-industry/articleshow/75589187.cms

[2]“Collection of Domestic Tourism Statistics for the State of Goa”, Datamation Consulatants Private Limited, https://tourism.gov.in/sites/default/files/2020-04/03%20goa.pdf

[3] “Tourist Arrival Statistics”, http://goatourism.gov.in/wp-content/uploads/2019/09/tourist_arrivals_statistics_apr_2019.pdf

[4] Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. Information & Management, 54(6), 771–785. doi:10.1016/j.im.2016.11.011

[5]Da Rugna, J., Chareyron, G., & Branchet, B. (2012). *Tourist behavior analysis through geotagged photographies: A method to identify the country of origin. 2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI).* doi:10.1109/cinti.2012.6496788

[6] Mikhailov, S., & Kashevnik, A. (2020). *Tourist Behaviour Analysis Based on Digital Pattern of Life—An Approach and Case Study. Future Internet, 12(10), 165.* doi:10.3390/fi12100165

[7] Huang, X.; Li, M.; Zhang, J.; Zhang, L.; Zhang, H.; Yan, S. Tourists’ spatial-temporal behavior patterns in theme parks: A case study of Ocean Park Hong Kong. J. Destin. Mark. Manag. 2020, 15, 100411.

[8] Hasnat, M.M.; Hasan, S. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. Transp. Res. Part C Emerg. Technol. 2018, 96, 38–54. [

[9] The remaining Instagram Legacy API permission ("Basic Permission") was disabled on June 29, 2020. As of June 29, third-party apps no longer have access to the Legacy API. To avoid disruption of service to your app and business, developers previously using the Legacy API should instead rely on Instagram Basic Display API and Instagram Graph API. Please request approval for required permissions through the App Review process. *https://www.instagram.com/developer/*

ACKNOWLEDGMENTS

We would like to express our special thanks of gratitude to our project guide Asst. Prof. Pankaj Sonawane for their able guidance, support and suggestions which helped us in completing this project. We would also like to extend our gratitude to the Head of the Computer Engineering Department, Dr. Meera Narvekar for providing us with all the facility that was required.