Feynn Labs Assignment

Segmentation for Medical Market in India

Presented By

Team-Rahul(CodeBreakers)

Rahul Sharma
Jeet Kuntal Thakkar
Shubham Trivedi
Swati Prakash Mallick

“Health is wealth”

Index

Introduction

Why should Online HealthCare services be developed by any company?

With 462 million active internet users and 430 million active mobile internet users in India, the scope for ehealth services and solutions can prove to be a game-changer in the patient care space. And in the current urban milieu where internet access is far higher than the national average; the number of potential consumers especially from the corporate sector provides an untapped opportunity for healthcare providers to pilot, facilitate, regulate, and expand the span of preventive medicine, primary health care, and wellness programs beyond conventional boundaries.

“Private healthcare corporations and NGOs are also making major inroads towards offering services that enable easier online access and more responsive health-care services to a growing number of users. There is a paradigm-shift from provider-centered to patient-centered healthcare”

Various programs are focused on making the entire spectrum of medical facilities available 24*7 through the web, mobile, SMS and Call center services. The spectrum of these online interventions encompasses medical consultation, medical records, medicine supply management and Pan-India exchange of patient information.

Some of the expected outcomes are delivery of better medical amenities in terms of equitable access,

quality, affordability, lowering of disease burden, and efficient monitoring of health entitlements for citizens.

What is market segmentation?

Well to be concise and clear market segmentation is the delineation or disaggregation of the market into uniquely distinct submarkets.

Benefits of Market Segmentation in healthcare

Market segmentation is a decision-making tool for the marketing manager in the crucial task of selecting a target market. In the Healthcare industry, a company willing to offer different services to different target audiences will lead to the sustainable growth of the company. Customers would get what they are seeking, and the company would emerge as the best possible service provider. Deploying an appropriate market segmentation strategy will also provide a competitive advantage over other similar companies.

Problem Statement

In this report, our goal is to gather information about the present health care market in India and to obtain detailed knowledge about some of the specifics like online health service appointment bookings from various states and also

parts of those states and customers from which region of India are willing to use smart devices for

monitoring their vitals like Diabetes level, Blood Pressure and vitamin deficiencies. This will be done using analytical methods and market segmentation to draw out segments using the limited amount of data obtained from several trusted platforms, including government open source.

We have to analyze Medical Market in India with respect to the given problem statement using Segmentation analysis and come up with a feasible strategy to enter the market

Fermi Estimation

A Fermi estimate is one done using back-of-the-envelope calculations and rough generalizations to estimate values which would require extensive analysis or experimentation to determine exactly. Fermi estimates generally work because the estimations of the individual terms are often close to correct, and overestimates and underestimates help cancel each other out. That is, if there is no consistent bias, a Fermi calculation that involves the multiplication of several estimated factors will probably be more accurate than might be first supposed. Although Fermi calculations are often not accurate, as there may be many problems with their assumptions, this sort of analysis does tell us what to look for to get a better answer.

In our problem, we know that the population of India is in the order of billions(exact figure 1.38 billion) and health is the primary service that everyone would need. In an ideal world we would have a customer base in the order of billions. But as our services need people to have an internet connection we have to factor that in here, so we have those in the order of 100 million as 60 percent of people of India have internet connectivity as of 2021. so, our customer base dip to 100 million which is not bad, but we are not going to provide services to every person with the internet, we have been tasked to find an entry market so geographically that would be a couple of states which provide ideal situations for our start-up thus we would get our customer target down to order of 10million in good circumstances.

Data sources

1. India census data 2011
2. Internet Subscriber in India

Data collection was the hard part of the project. Since the dataset wasn't given readily, we had to explore and find datasets ourselves from various government sites like the Indian census website and data.gov.in. However, we were able to find a few datasets on the above sites and Kaggle. We had more than needed data for some factors like geography and population data because data.gov had so many datasets to exploit from but we faced the issue of selecting the right datasets to clean and make sense out of it. Then also we didn't have data on which we had our eyes on which is psychographic data of customers in a medical market which took a lot and lots of research, at last, we concluded that search because every time we researched we only came to the conclusion that this type of data is not a freely available and most efficient way of getting this type of data is survey customers directly

Data Pre-processing and Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure.
3. extract important variables.

4. detect outliers and anomalies;
5. test underlying assumptions.
6. develop parsimonious models; and
7. determine optimal factor settings.

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provide, of course, unparalleled power to carry this out.

We have carried out extensive exploration on the available data to get as much understanding of the patterns as possible. For this we have used libraries like Numpy, Pandas, matplotlib, and plotly. Express.

Firstly, we had to check for any missing values from the columns and also do a thorough check on if data values are consistent or not which was our main focus as analysis on inconsistent data is as good as garbage. Below are the steps that were taken to carry out EDA as detailed as possible by us.

Step 1 Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import plotly.express as px
```

Step 2 Reading data file into a python data frame

In [2]:
```
census_data_file_path = "C:/Users/LENOVO/Downloads/Datasets for Bio-tech problem(Feynlabs)/Literacy and population data Census
= pd.read_csv(census_data_file_path)

Census
```

Out[2]:

	District code	State name	District name	Population	Male	Female	Literate	Male_Literate	Female_Literate	SC	...	Power_Parity_Rs_90000_150000	Power_Parity_Rs_45000_150000	Power_Parity_Rs_150000_240000	Power_Parity
0	1	JAMMU AND KASHMIR	Kupwara	870354	474190	396164	439654	282823	156831	1048	...	94	588	71	
1	2	JAMMU AND KASHMIR	Badgam	753745	398041	355704	335649	207741	127908	368	...	126	562	72	
2	3	JAMMU AND KASHMIR	Leh(Ladakh)	133487	78971	54516	93770	62834	30936	488	...	46	122	15	
3	4	JAMMU AND KASHMIR	Kargil	140802	77785	63017	86236	56301	29935	18	...	27	114	12	
4	5	JAMMU AND KASHMIR	Punch	476835	251899	224936	261724	163333	98391	556	...	78	346	35	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
635	636	PONDICHERRY	Mahe	41816	19143	22673	36470	16610	19860	144	...	2316	4309	1370	
636	637	PONDICHERRY	Karaikal	200222	97809	102413	154916	79903	75013	35348	...	1063	2408	665	
637	638	ANDAMAN AND NICOBAR ISLANDS	Nicobars	36842	20727	16115	25332	15397	9935	0	...	685	1895	212	
638	639	ANDAMAN AND NICOBAR ISLANDS	North AND Middle Andaman	105597	54861	50736	78683	43186	35497	0	...	685	1895	212	
639	640	ANDAMAN AND NICOBAR ISLANDS	South Andaman	238142	127283	110859	190266	105794	84472	0	...	1371	3020	649	

640 rows × 118 columns

Step 3 Statistical Summary

In [3]:
```
Census.info()
```

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 640 entries, 0 to 639

Columns: 118 entries, District code to Total_Power_Parity
dtypes: int64(116), object(2)

memory usage: 590.1+ KB
```

In [4]:
```
Census.describe()
```

Out[4]:

	District code	Population	Male	Female	Literate	Male_Literate	Female_Literate	SC	Male_SC	Female_SC	...	Power_Parity_Rs_90000_150000	Power_Parity_Rs_45000_150000	Power_Parit
count	640.000000	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	6.400000e+02	...	640.000000	640.000000	
mean	320.500000	1.891961e+06	9.738598e+05	9.181011e+05	1.193186e+06	6.793182e+05	5.138675e+05	3.146537e+05	1.617739e+05	1.528798e+05	...	786.046875	1696.456250	
std	184.896367	1.544380e+06	8.007785e+05	7.449864e+05	1.068583e+06	5.924144e+05	4.801816e+05	3.129818e+05	1.611216e+05	1.520336e+05	...	1038.854733	1720.535151	
min	1.000000	8.004000e+03	4.414000e+03	3.590000e+03	4.436000e+03	2.614000e+03	1.822000e+03	0.000000e+00	0.000000e+00	0.000000e+00	...	0.000000	0.000000	
25%	160.750000	8.178610e+05	4.171682e+05	4.017458e+05	4.825982e+05	2.764365e+05	2.008920e+05	8.320850e+04	4.230700e+04	4.267175e+04	...	236.750000	589.000000	
50%	320.500000	1.557367e+06	7.986815e+05	7.589200e+05	9.573465e+05	5.483525e+05	4.038590e+05	2.460160e+05	1.255485e+05	1.178550e+05	...	518.000000	1220.500000	
75%	480.250000	2.583551e+06	1.338604e+06	1.264277e+06	1.602260e+06	9.188582e+05	6.641550e+05	4.477078e+05	2.284602e+05	2.140502e+05	...	941.250000	2233.250000	
max	640.000000	1.106015e+07	5.865078e+06	5.195070e+06	8.227161e+06	4.591396e+06	3.635765e+06	2.464032e+06	1.266504e+06	1.197528e+06	...	10334.000000	13819.000000	

8 rows × 116 columns

Step 4 Checking for null values

```
In [5]:Census.isnull().sum()
```

```
Out[5]:
District code 0
State name 0
District name 0
Population 0
Male 0
 ..
Power_Parity_Rs_330000_425000 0
Power_Parity_Rs_425000_545000 0
Power_Parity_Rs_330000_545000 0
Power_Parity_Above_Rs_545000 0
Total_Power_Parity 0
Length: 118, dtype: int64
```

There are no null values so carrying forward with our analysis

Step 5 Dumping unwanted columns

In [6]:
```
Census.drop(['SC','Male_SC','Female_SC','ST','Male_ST','Female_ST','Male_Workers','Female_Workers','Hindus','Muslims','Ch
 ,'Housholds_with_Electric_Lighting'], axis=1, inplace= True)
```

Here we deleted columns that were not suitable for the purpose of the analysis for our business problem

In [7]:
```
Census.info()
```

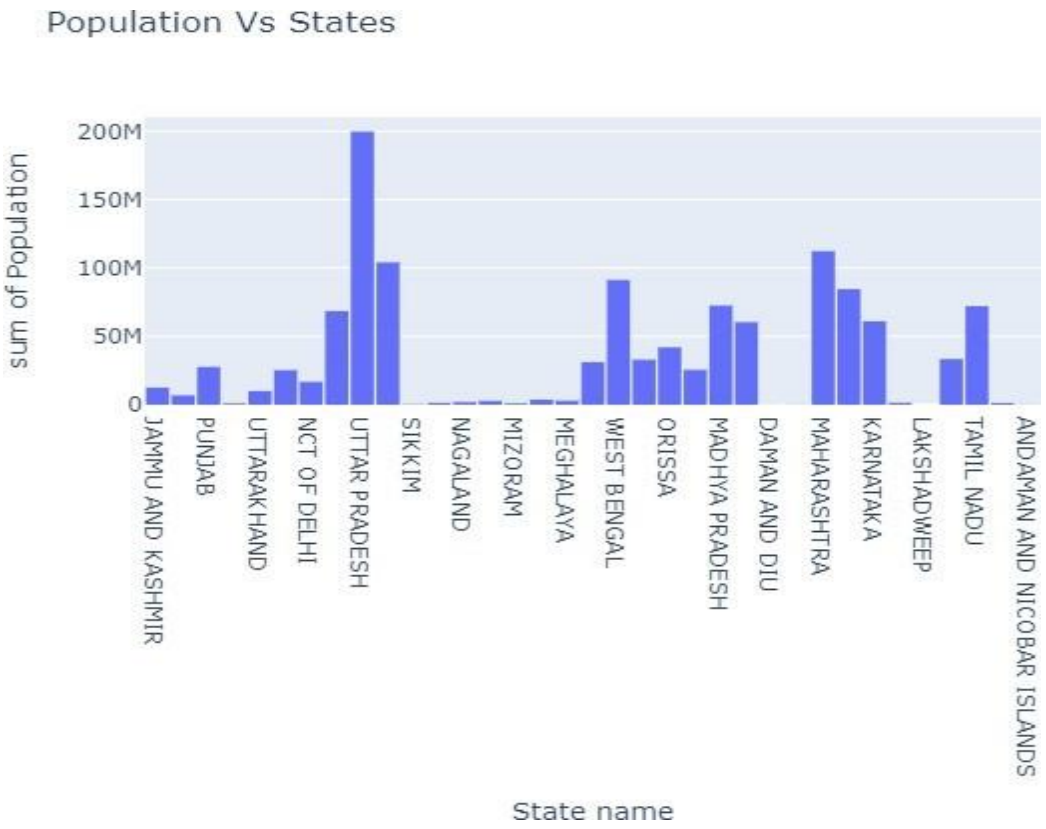
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 48 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   District code                       640 non-null    int64
 1   State name                          640 non-null    object
 2   District name                       640 non-null    object
 3   Population                          640 non-null    int64
 4   Male                                640 non-null    int64
 5   Female                              640 non-null    int64
 6   Literate                            640 non-null    int64
 7   Male_Literate                       640 non-null    int64
 8   Female_Literate                     640 non-null    int64
 9   Workers                             640 non-null    int64
 10  Main_Workers                        640 non-null    int64
 11  Marginal_Workers                    640 non-null    int64
 12  Non_Workers                         640 non-null    int64
 13  Cultivator_Workers                  640 non-null    int64
 14  Agricultural_Workers                640 non-null    int64
 15  Household_Workers                   640 non-null    int64
 16  Other_Workers                       640 non-null    int64
 17  Households_with_Internet            640 non-null    int64
 18  Households_with_Computer            640 non-null    int64
 19  Rural_Households                    640 non-null    int64
 20  Urban_Households                    640 non-null    int64
 21  Households                          640 non-null    int64
 22  Below_Primary_Education             640 non-null    int64
 23  Primary_Education                   640 non-null    int64
 24  Middle_Education                    640 non-null    int64
 25  Secondary_Education                 640 non-null    int64
 26  Higher_Education                    640 non-null    int64
 27  Graduate_Education                  640 non-null    int64
 28  Other_Education                     640 non-null    int64
 29  Literate_Education                  640 non-null    int64
 30  Illiterate_Education                640 non-null    int64
 31  Total_Education                     640 non-null    int64
 32  Age_Group_0_29                      640 non-null    int64
 33  Age_Group_30_49                     640 non-null    int64
 34  Age_Group_50                        640 non-null    int64
 35  Age not stated                      640 non-null    int64
 36  Power_Parity_Less_than_Rs_45000     640 non-null    int64
 37  Power_Parity_Rs_45000_90000         640 non-null    int64
 38  Power_Parity_Rs_90000_150000        640 non-null    int64
 39  Power_Parity_Rs_45000_150000        640 non-null    int64
 40  Power_Parity_Rs_150000_240000       640 non-null    int64
 41  Power_Parity_Rs_240000_330000       640 non-null    int64
 42  Power_Parity_Rs_150000_330000       640 non-null    int64
 43  Power_Parity_Rs_330000_425000       640 non-null    int64
 44  Power_Parity_Rs_425000_545000       640 non-null    int64
 45  Power_Parity_Rs_330000_545000       640 non-null    int64
 46  Power_Parity_Above_Rs_545000        640 non-null    int64
 47  Total_Power_Parity                  640 non-null    int64
dtypes: int64(46), object(2)
memory usage: 240.1+ KB
```

As you can see now we have left with a Data Frame that is only 49 columns down from 118 columns

Step 6 Exploring for insights at the State level

There are many valuable variables from the above list but for the starter let's select population and state columns.

In [8]:
```
fig = px.histogram(Census,
 x="State name",
 y = "Population", title='Population Vs States')
fig.update_layout(bargap=0.1)
fig.show()
```



So, there are many states which can be selected for our startup to launch their services solely based on population count. Most likely more business will be generated from states like :

Rajasthan

Uttarpradesh

Bihar

West Bengal

Madhya Pradesh

Gujarat

Maharashtra

Andhra Pradesh

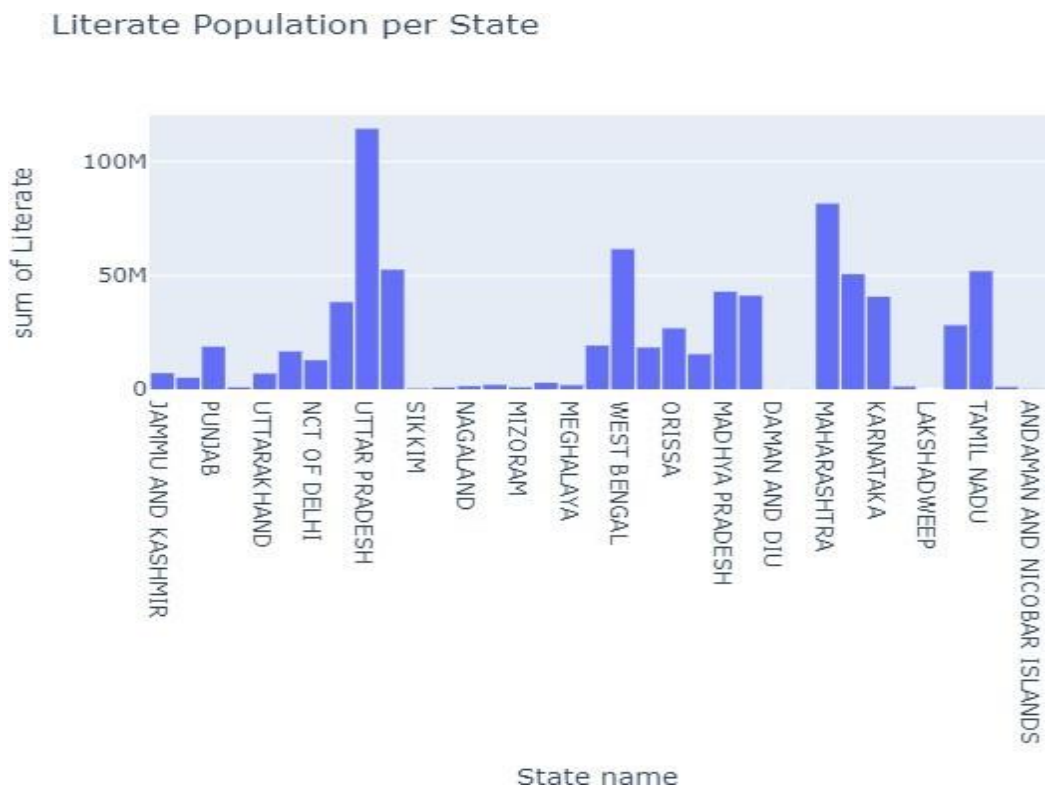
Karnataka

and Tamil Nadu

Note :- These are states with a population greater than 50 million and this does not visualize the whole scenario. It is just a speculation based on Total population count of the above given states.

Now, let's explore the number of literate people residing in every state as literacy rate is directly Correlated by regular medical checkups.

```
In [9]:fig = px. histogram(Census,
           x = "State name",
           y = "Literate",
           title = "Literate Population per State")
          fig.update_layout(bargap = 0.1)
          fig.show()
```



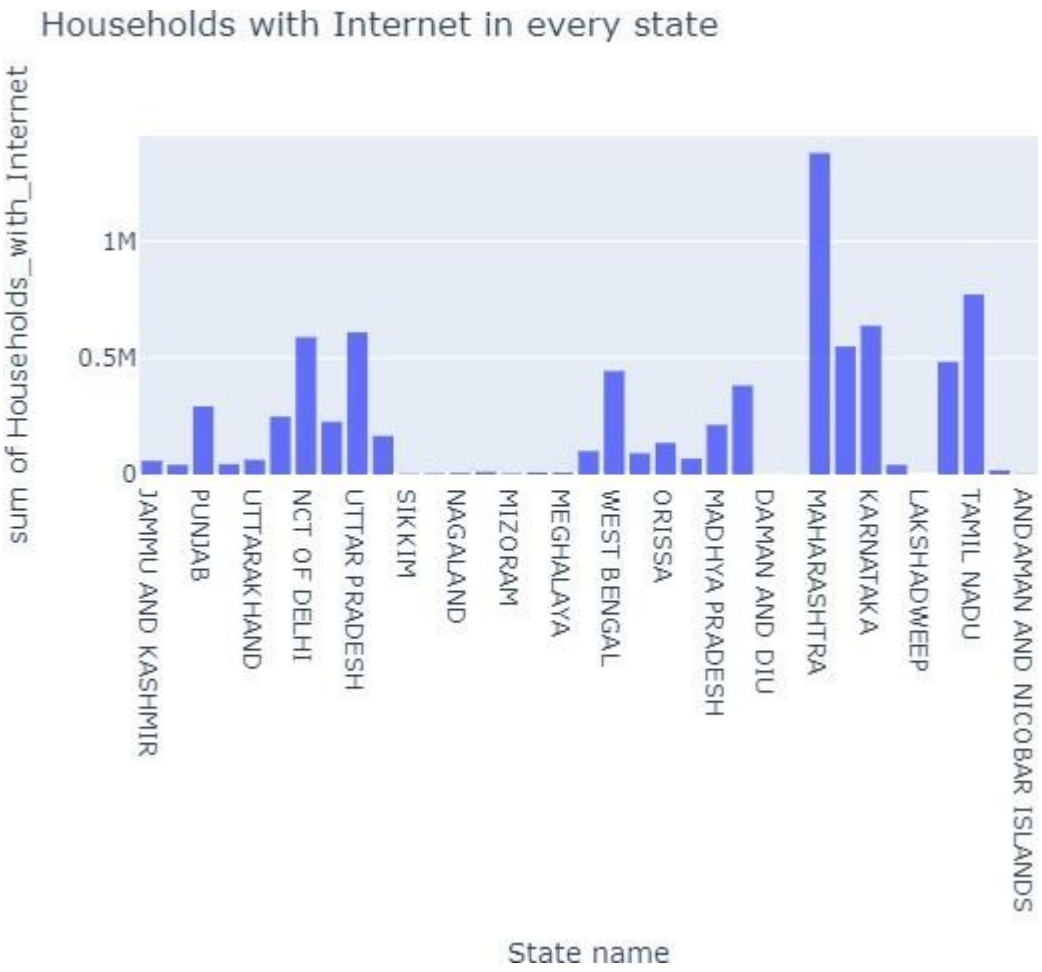
So most off the states that we selected earlier based on population has adequate amount of literate people.But to select a few i would state :

- **Rajasthan(least literate population count)**
- **Uttar Pradesh(Highest literate population count)**
- **Bihar**
- **West Bengal**
- **madhya Pradesh**
- **Gujarat**
- **Maharashtra**
- **Andhra Pradesh**
- **Karnataka**
- **and Tamil Nadu**

Our company provides services on online appointment basis and basic need for that would be an internet connection so let's

plot for that

```
In [10]:fig = px.histogram(Census,
           x = "State name",
           y = "Households_with_Internet",
           title = "Households with Internet in every state")
         fig.show()
```



from above figure it is evident that most internet users will be in Maharashtra but let's make a list of states which have households with internet over 0.5 million or close to this base limit :

NCT of Delhi

Uttar Pradesh

West Bengal

Gujarat(Least users)

Maharashtra(Maximum users)

Andhra Pradesh

Karnataka

Kerala

Tamil Nadu

previous analysis shows why it was necessary to visualize amounts of household with internet because even if state would have larger population base it does not guarantee it will generate business for the company like Biotech as it is online service provider, and which would indicate total number of internet users per state is very important variable than only focusing on totalpopulation that state has.

Step 7 -Exploring for Insights at District level

Now, we are going to explore district wise data diving deep into our previous findings for top four states for highest number of internet users

Firstly, we are going to make separate data frame for data of above listed states

```
In [11]:NCT_of_Delhi = Census[Census['State name'] == "NCT OF DELHI"]
```

```
Uttar_Pradesh = Census[Census['State name'] == "UTTAR PRADESH"]
West_Bengal = Census[Census['State name'] == "WEST BENGAL"]
Gujarat = Census[Census['State name'] == "GUJARAT"]
Maharashtra = Census[Census['State name'] == "MAHARASHTRA"]
Andra_Pradesh = Census[Census['State name'] == "ANDRA PRADESH"]
Karnataka = Census[Census['State name'] == "KARNATAKA"]
Kerala = Census[Census['State name'] == "KERALA"]
Tamil_Nadu = Census[Census['State name'] == "TAMIL NADU"]
```

It will be a very tedious task to write code for each and every state wise data frame that we made recently. So, we are going to define a function for that purpose.

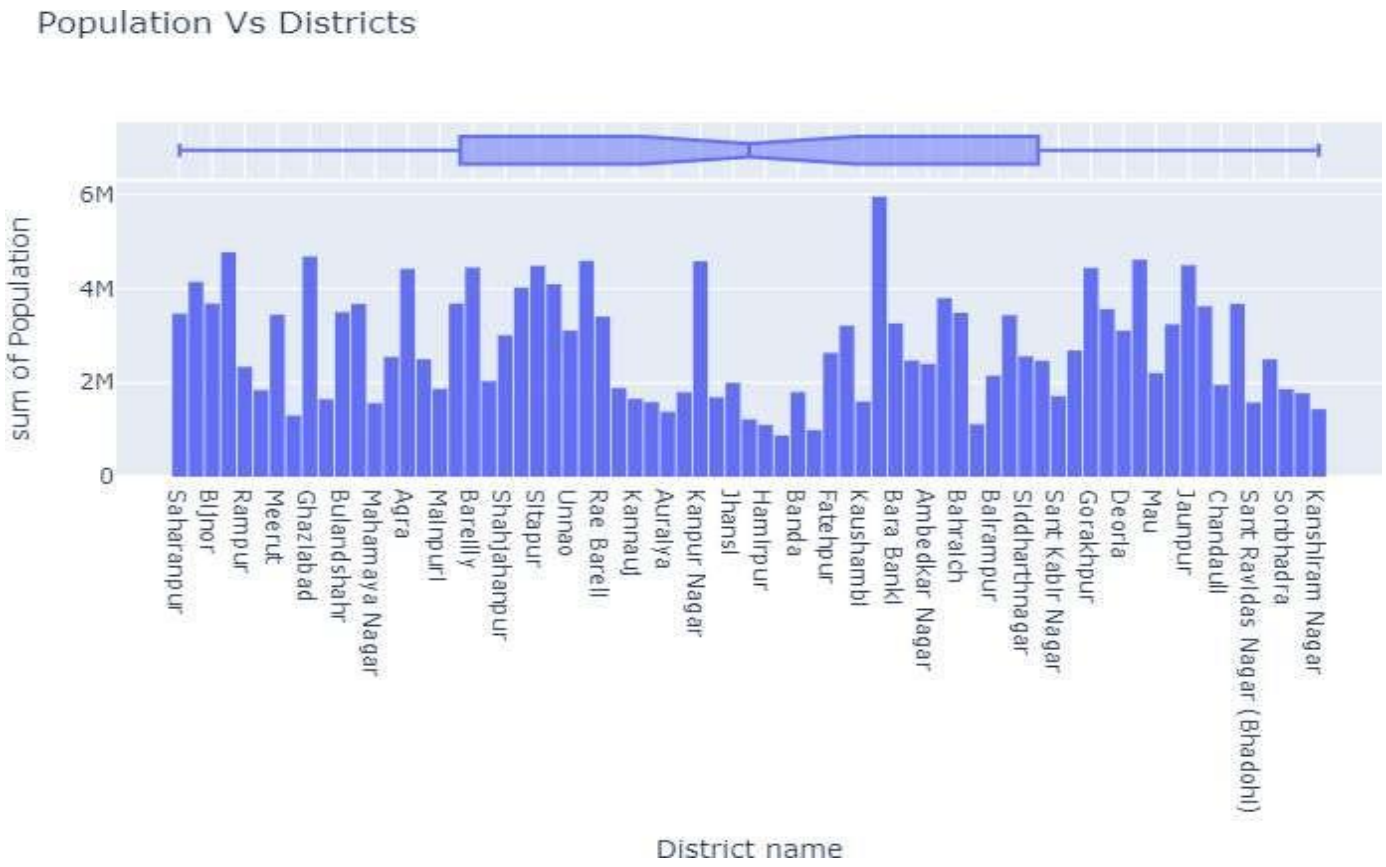
In [12]:
```
def Explore_districts_of(state):
    fig = px.histogram(state,
    marginal = 'box',
    x="District name",
    y = "Population",
    title='Population Vs Districts')
    fig.update_layout(bargap=0.1)
    fig.show()

    fig = px.histogram(state,
    marginal = 'box',
    x="District name",
    y = "Literate",
    title='Number of Literate Vs Districts')
    fig.update_layout(bargap=0.1)
    fig.show()

    fig = px.histogram(state,
    marginal = 'box',
    x = "District name",
    y = "Households_with_Internet",
    title = "Households with Internet in every District")
    fig.show()
```

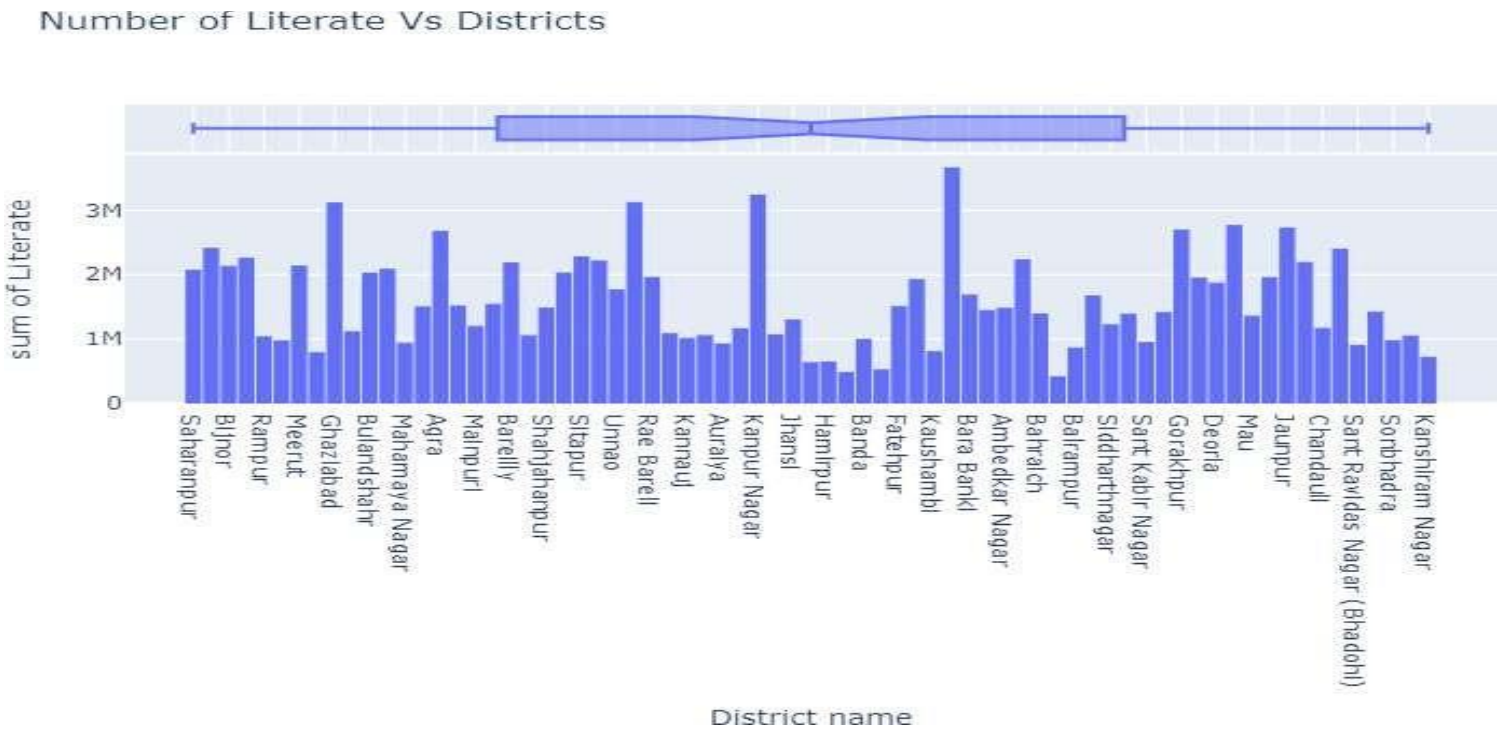
We are all set now first we are going for capital of our country after that we will choose every last state that we made a list

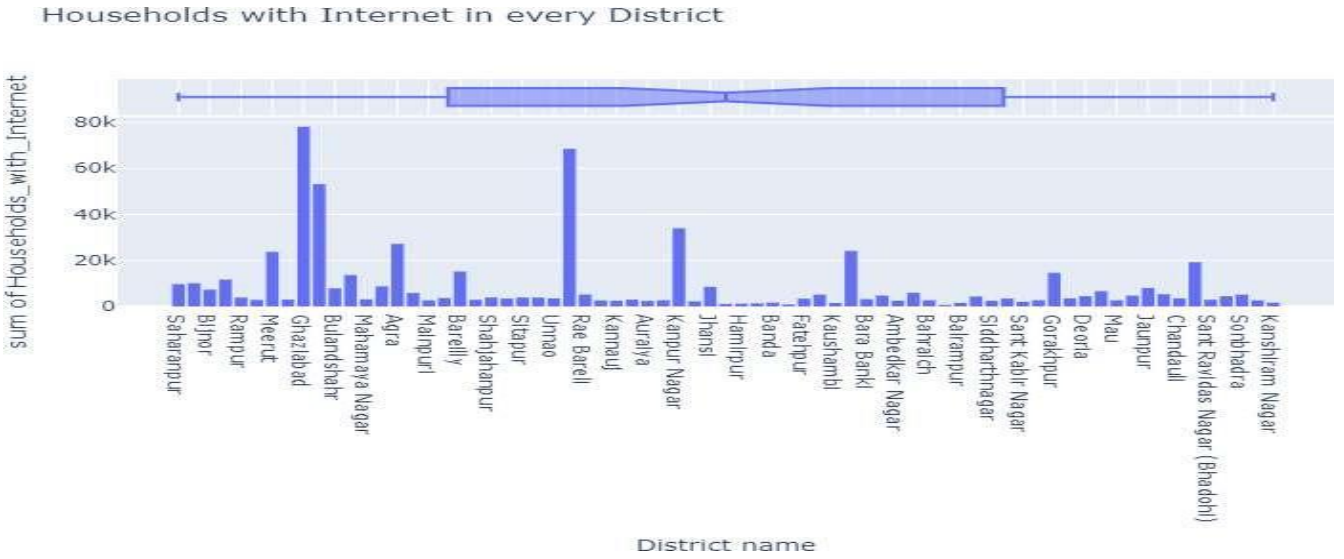
for

(1) Uttar Pradesh

In [13]:
```
Explore_districts_of(Uttar_Pradesh)
```

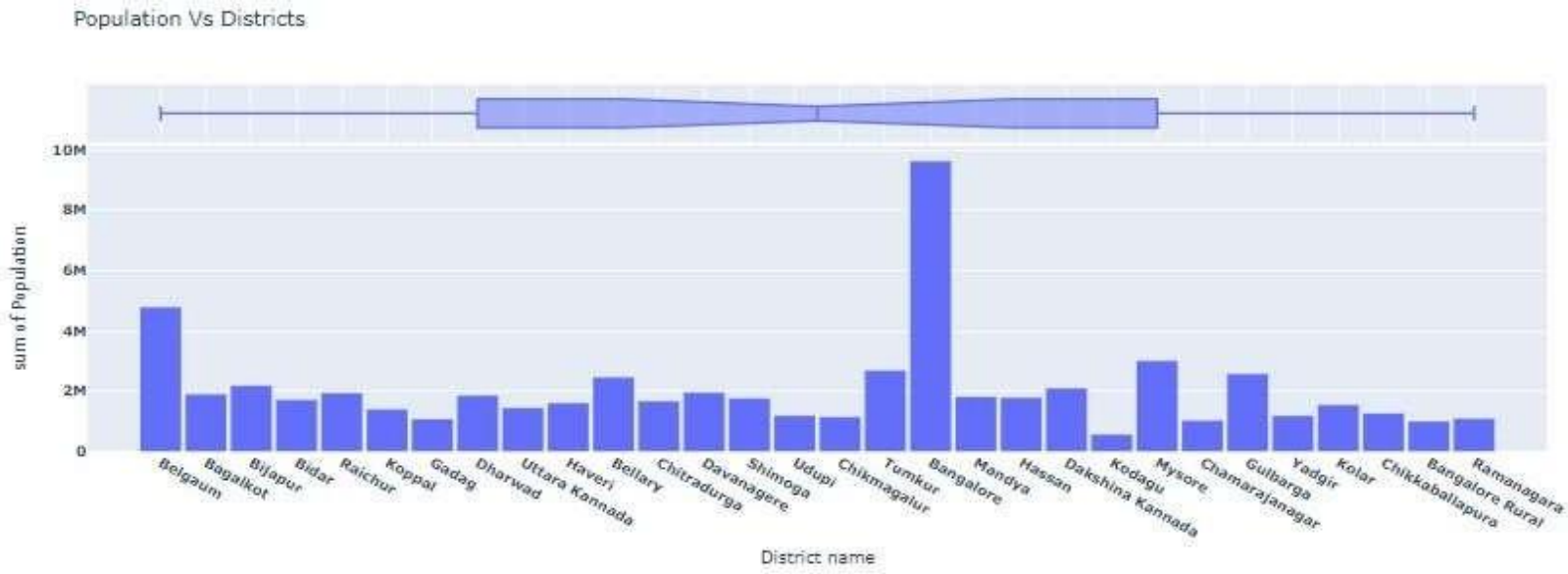


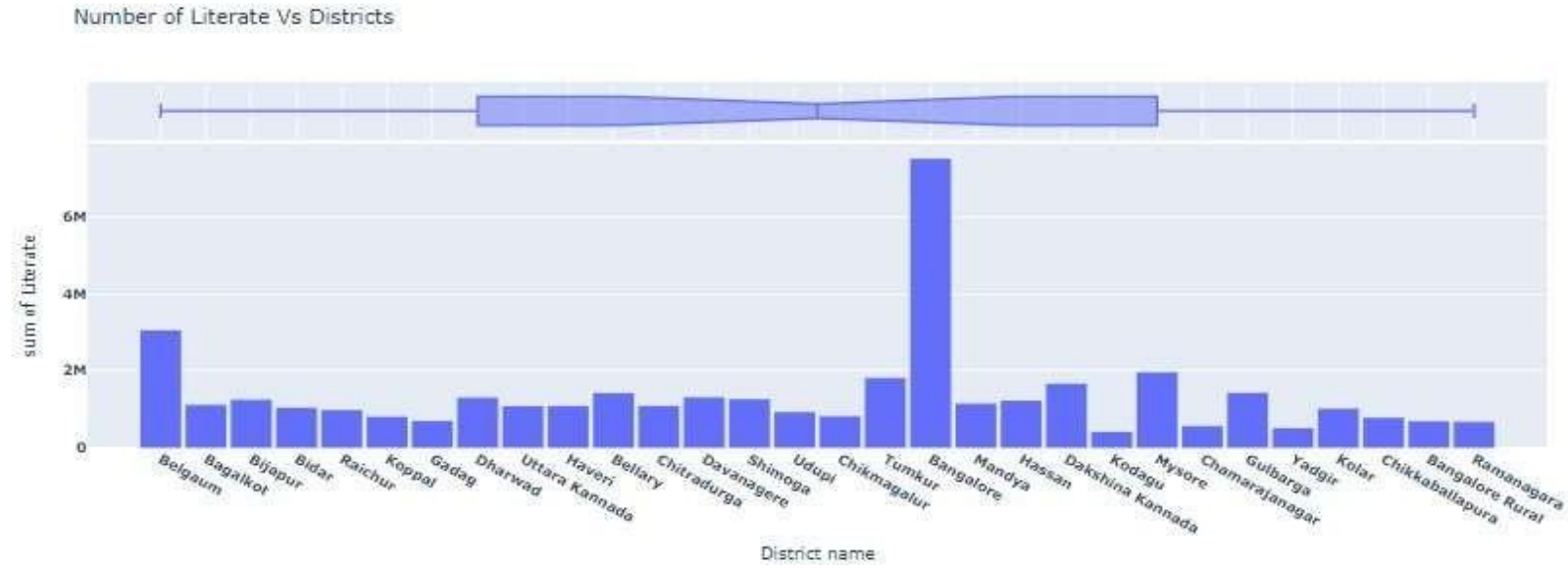

Number of Literate Vs Districts
sum of Literate
3M
2M
1M
0
Saharanpur
Bijnor
Rampur
Meerut
Ghaziabad
Bulandshahr
Mahamaya Nagar
Agra
Mainpuri
Bareilly
Shahjahanpur
Sitapur
Unnao
Rae Bareli
Kannauj
Auraiya
Kanpur Nagar
Jhansi
Hamirpur
Banda
Fatehpur
Kaushambi
Bara Banki
Ambedkar Nagar
Bahraich
Balrampur
Siddharthnagar
Sant Kabir Nagar
Gorakhpur
Deoria
Mau
Jaunpur
Chandauli
Sant Ravidas Nagar (Bhadohi)
Sonbhadra
Kanshiram Nagar
District name

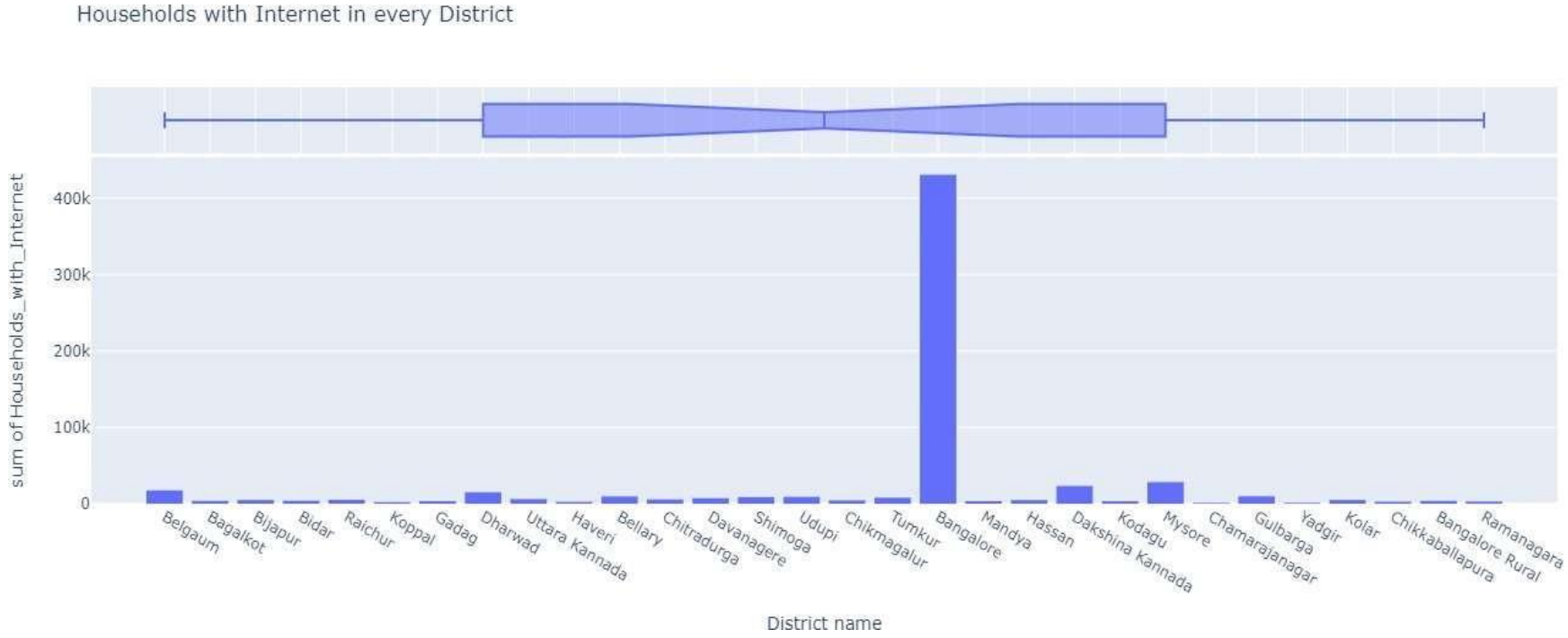

Households with Internet in every District
sum of Households_with_Internet
80k
60k
40k
20k
0
Saharanpur
Bijnor
Rampur
Meerut
Ghaziabad
Bulandshahr
Mahamaya Nagar
Agra
Mainpuri
Bareilly
Shahjahanpur
Sitapur
Unnao
Rae Bareli
Kannauj
Auraiya
Kanpur Nagar
Jhansi
Hamirpur
Banda
Fatehpur
Kaushambi
Bara Banki
Ambedkar Nagar
Bahraich
Balrampur
Siddharthnagar
Sant Kabir Nagar
Gorakhpur
Deoria
Mau
Jaunpur
Chandauli
Sant Ravidas Nagar (Bhadohi)
Sonbhadra
Kanshiram Nagar
District name

(2) Karnataka

```
In [14]:Explore_districts_of(Karnataka)
```

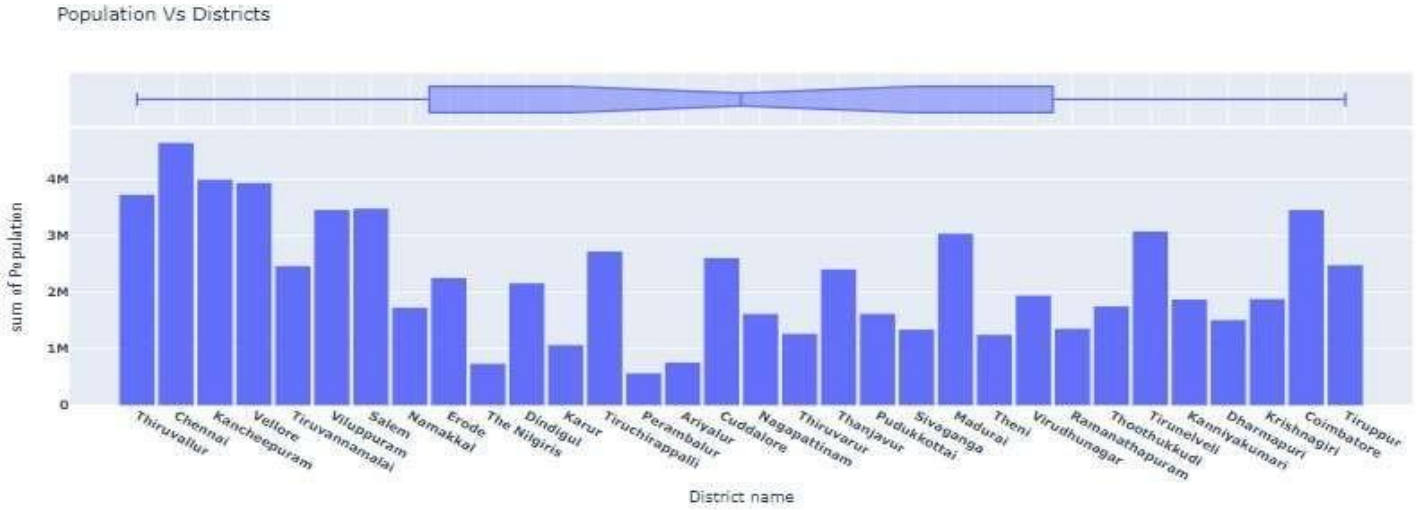


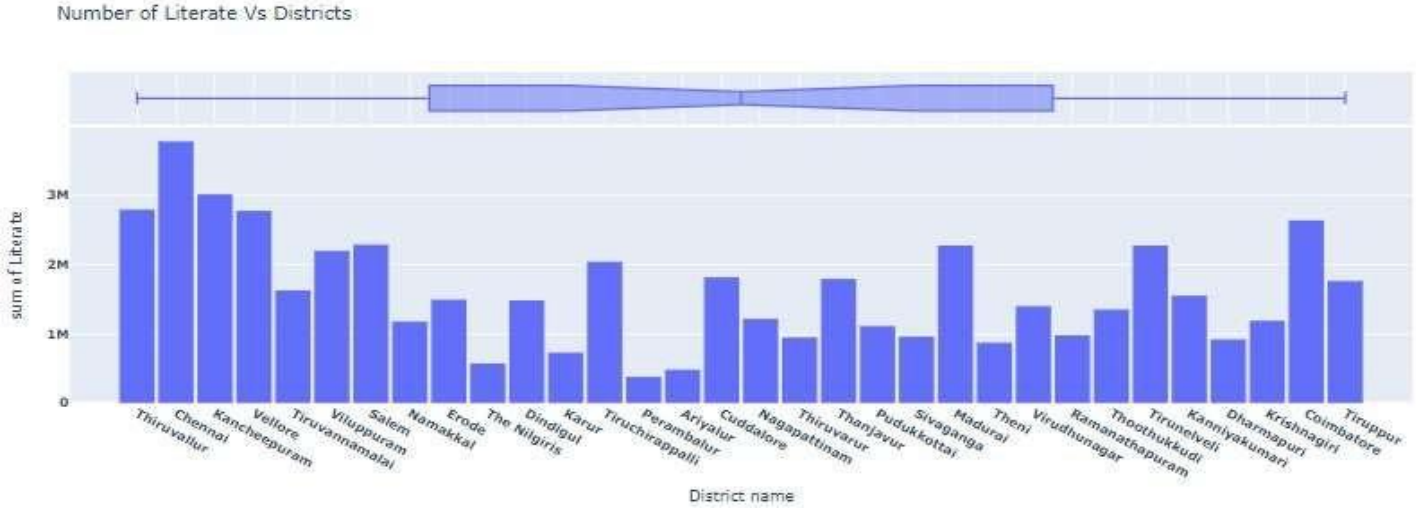


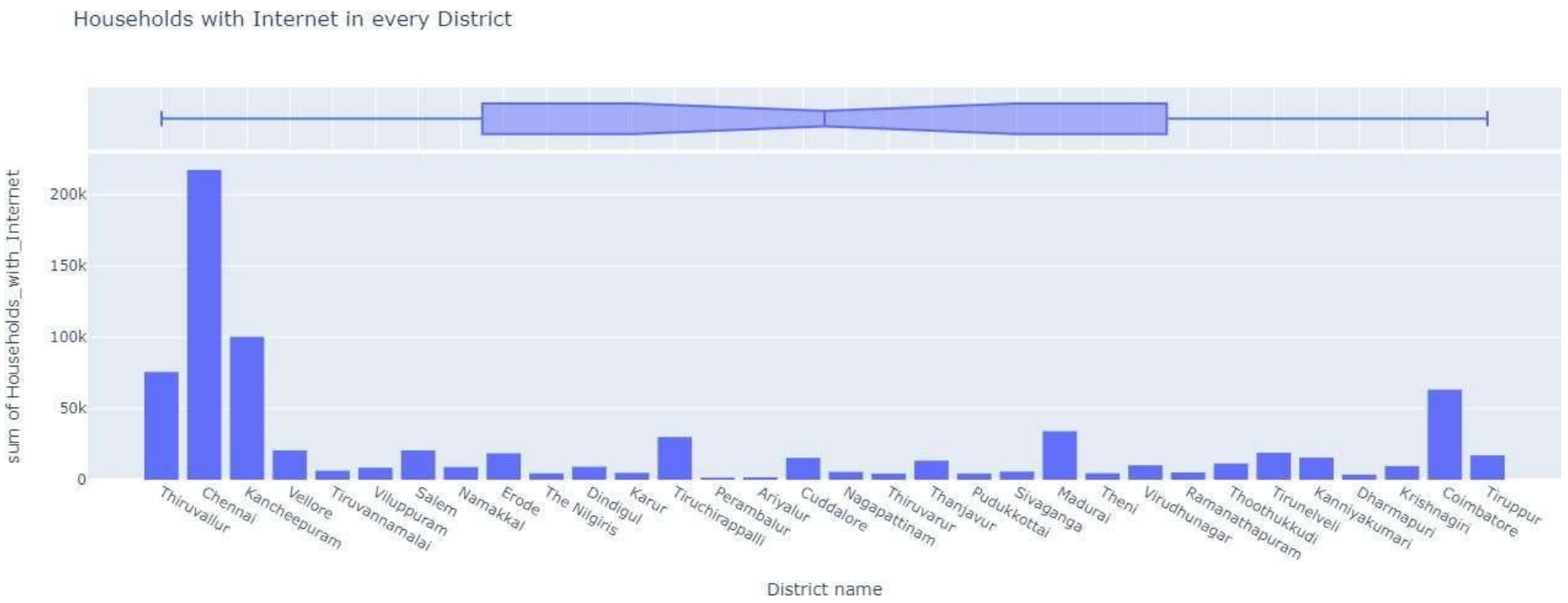

Households with Internet in every District
sum of Households_with_Internet
400k
300k
200k
100k
0
Belgaum
Bagalkot
Bijapur
Bidar
Raichur
Koppal
Gadag
Dharwad
Uttara Kannada
Haveri
Bellary
Chitradurga
Davanagere
Shimoga
Udupi
Chikmagalur
Tumkur
Bangalore
Mandya
Hassan
Dakshina Kannada
Kodagu
Mysore
Chamarajanagar
Gulbarga
Yadgir
Kolar
Chikkaballapura
Bangalore Rural
Ramanagara
District name

3) Tamil Nadu

```
Explore_districts_of(Tamil Nadu)
```

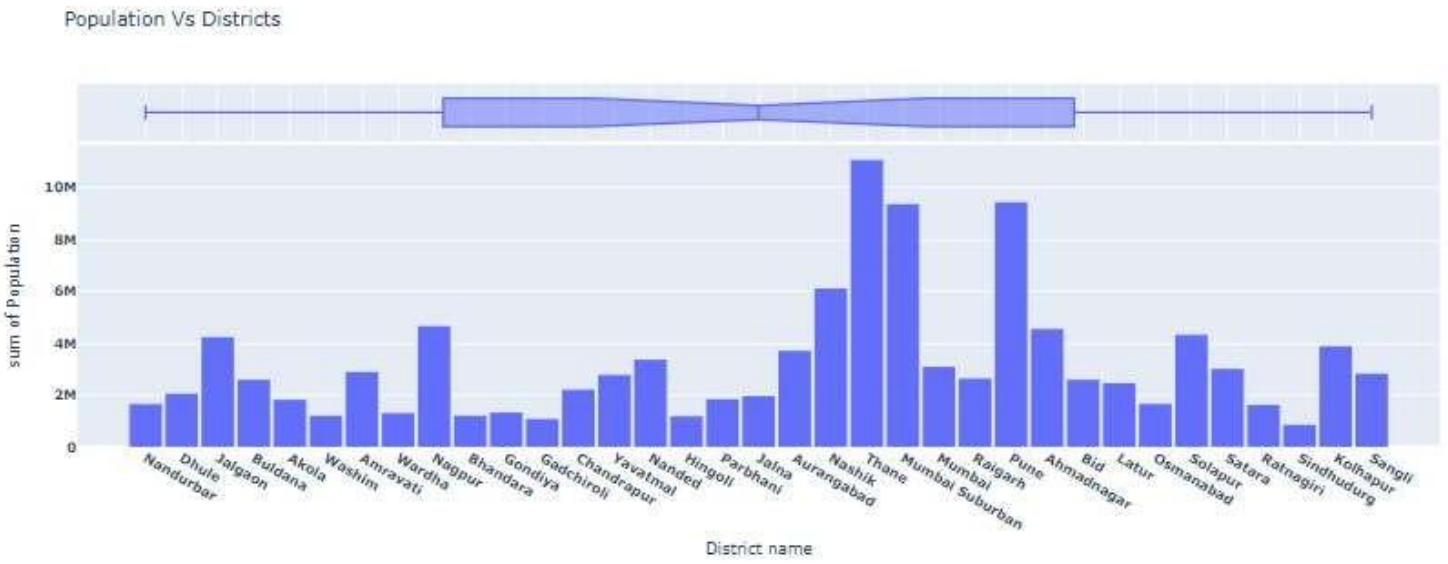


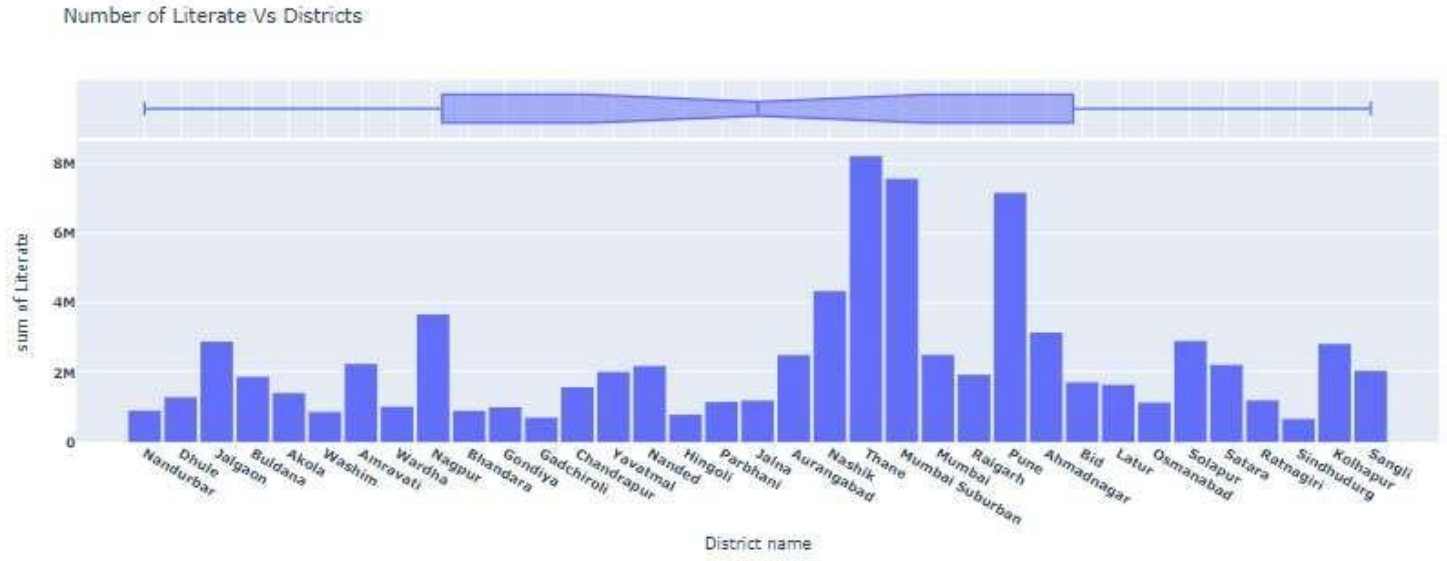


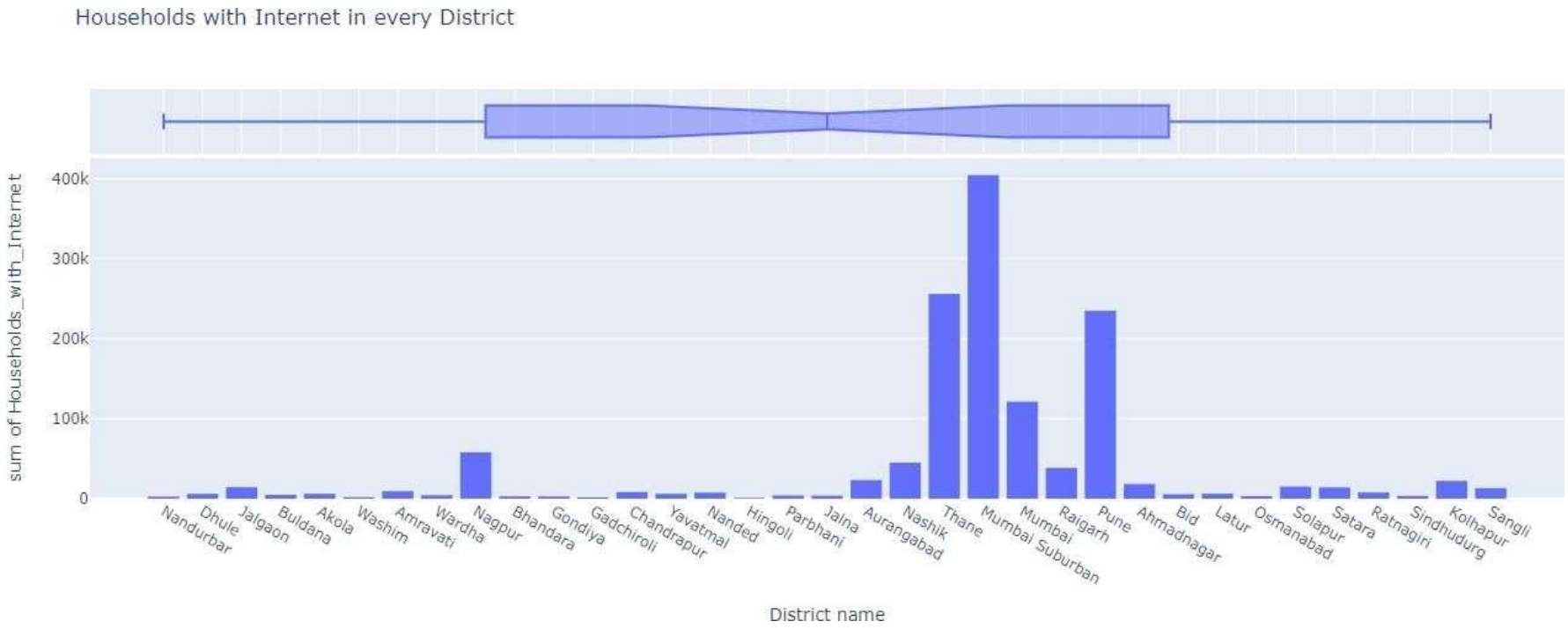

Households with Internet in every District
sum of Households_with_Internet
200k
150k
100k
50k
0
Thiruvallur
Chennai
Kancheepuram
Vellore
Tiruvannamalai
Viluppuram
Salem
Namakkal
Erode
The Nilgiris
Dindigul
Karur
Tiruchirappalli
Perambalur
Ariyalur
Cuddalore
Nagapattinam
Thiruvarur
Thanjavur
Pudukkottai
Sivaganga
Madurai
Theni
Virudhunagar
Ramanathapuram
Thoothukkudi
Tirunelveli
Kanniyakumari
Dharmapuri
Krishnagiri
Coimbatore
Tiruppur
District name

(4) Maharashtra

In [16]:
```
Explore_districts_of(Maharashtra)
```






Households with Internet in every District
sum of Households_with_Internet
400k
300k
200k
100k
0
Nandurbar
Dhule
Jalgaon
Buldana
Akola
Washim
Amravati
Wardha
Nagpur
Bhandara
Gondiya
Gadchiroli
Chandrapur
Yavatmal
Nanded
Hingoli
Parbhani
Jalna
Aurangabad
Nashik
Thane
Mumbai Suburban
Mumbai
Raigarh
Pune
Ahmadnagar
Bid
Latur
Osmanabad
Solapur
Satara
Ratnagiri
Sindhudurg
Kolhapur
Sangli
District name

Step 8 Gathering Insights for few selected states

Let's dive deep into these states before selecting one for our first market

```
In [17]:Selected_States = pd.concat([Uttar_Pradesh, Maharashtra, Tamil_Nadu, Karnataka], axis=0)
```

```
In [18]:Selected_States
```

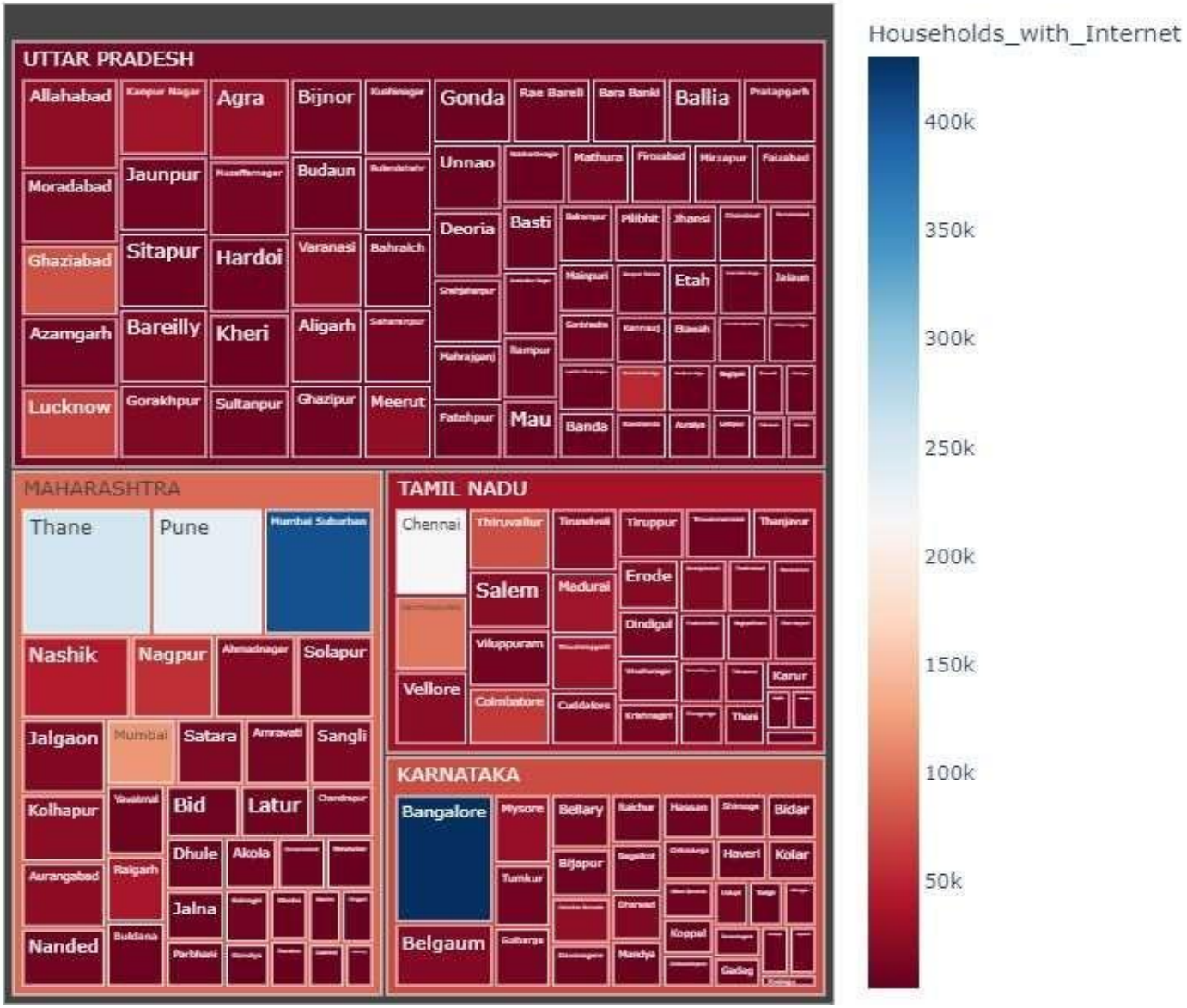
Out[18]:

	District code	State name	District name	Population	Male	Female	Literate	Male_Literate	Female_Literate	Workers	...	Power_Parity_Rs_90000_150000	Power_Parity_Rs_45000_150000	Power_Parity_Rs_150000_240000	Po
131	132	UTTAR PRADESH	Saharanpur	3466382	1834106	1632276	2077108	1220114	856994	1037344	...	974	2349	251	
132	133	UTTAR PRADESH	Muzaffarnagar	4143512	2193434	1950078	2417339	1448528	968811	1291644	...	1114	2813	316	
133	134	UTTAR PRADESH	Bijnor	3682713	1921215	1761498	2135393	1241471	893922	1088036	...	930	2481	289	
134	135	UTTAR PRADESH	Moradabad	4772006	2503186	2268820	2263848	1357435	906413	1417811	...	1250	3125	357	
135	136	UTTAR PRADESH	Rampur	2335819	1223889	1111930	1043666	630408	413258	737261	...	474	1439	186	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
579	580	KARNATAKA	Yadgir	1174271	590329	583942	510003	306751	203252	547696	...	563	1257	184	
580	581	KARNATAKA	Kolar	1536401	776396	760005	1016219	564110	452109	717872	...	799	1806	206	
581	582	KARNATAKA	Chikkaballapura	1255104	636437	618667	783222	442158	341064	639778	...	121	250	30	
582	583	KARNATAKA	Bangalore Rural	990923	509172	481751	688749	385311	303438	459891	...	368	748	113	
583	584	KARNATAKA	Ramanagara	1082636	548008	534628	674758	378461	296297	531459	...	520	1056	155	

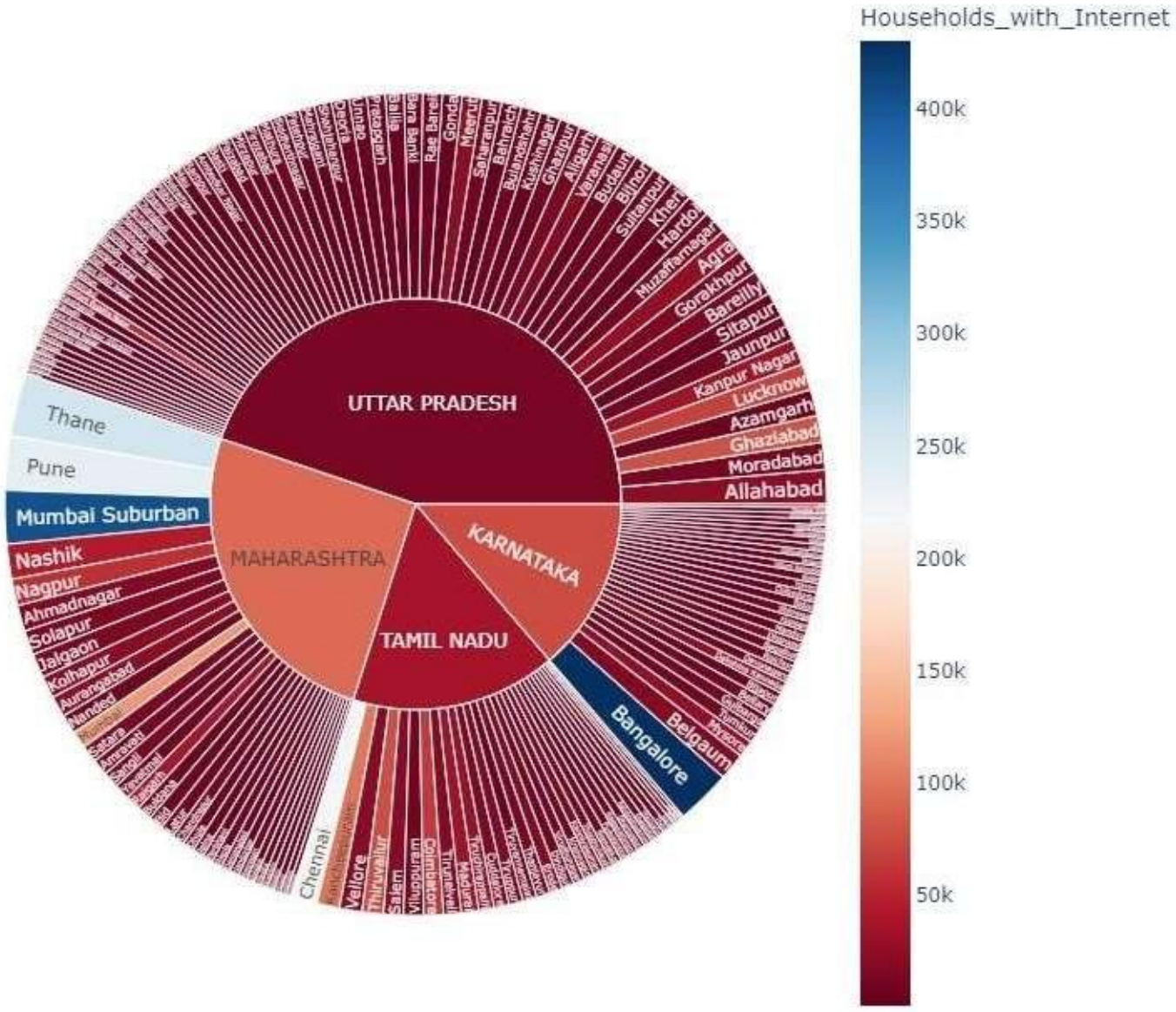
168 rows × 48 columns

In [19]:
```
fig = px.treemap(Selected_States,
 path=['State name','District name'],
 values='Population',
 color='Households_with_Internet',
 color_continuous_scale='RdBu',
 title = 'Finding out best Market')
fig.update_layout(bargap=1,autosize=False,
width=800,
 height=800,)
fig.show()

fig = px.sunburst(Selected_States,
 path=['State name','District name'],
 values='Population',
 color='Households_with_Internet',
 color_continuous_scale='RdBu',
 title = 'Finding out best Market')
fig.update_layout(
 autosize=False,
width=800,
height=800)
fig.show()
```

Finding out best Market



Finding out best Market



a larger the portion of the district in above visuals shows larger total population and colour inclination towards darker shades of blue means a larger number of households with internet connection

So, it seems there are five districts that look like promising greater business opportunities for us. especially three of which are in Maharashtra.

Step 9 Recommendations based on our EDA

- **Our company should incorporate some advanced data collection methods like focus groups and mass public surveys to these states and specially to the states corresponding to five districts that show great promise of business gains. Public surveys can be conducted online too by giving out incentives or discounts to customers that take part in it. By doing so we are also finalizing our customer base by marketing and also collecting data that can be used to create psychographic profiles of the participants which will give us enough understanding of the local community, their values and their attitude towards online health services.**
- **This type of data collection needs to be done at a large scale to get rid of bias which is very dangerous for our analysis.**
- **My personal judgment leans towards the Maharashtrian market as this state has more districts that are attractive for our profits but also has a quality population that has internet services and higher literacy rates.**
- **Marketing department should first penetrate larger cities which have denser populations because the faster the density of population will be the word-of-mouth marketing like a wildfire spreading across dense forest.**
- **After establishing concrete business there, we should move towards cities with lesser public and thentowards the rural areas as rural segment is very hard to deal with for many reasons like providing fastcustomer service is very challenging, inventory storage in nearby areas is very costly and also the possibility of people adopting this change of online healthcare services is extremelly**

low.

Segment Extraction

For extracting appropriate segments, we have used different ML techniques for clustering like k-means , hierarchical clustering, DBSCAN (density-based spatial clustering of applications with noise).

Data-driven market segmentation analysis is exploratory by nature. Consumer data sets are typically not well structured. Consumers come in all shapes and forms; a two-dimensional plot of consumers' product preferences typically does not contain clear groups of consumers. Rather, consumer preferences are spread across the entire plot. The combination of exploratory methods and unstructured consumer data means that results from any method used to extract market segments from such data will strongly depend on the assumptions made on the structure of the segments implied by the method. The result of a market segmentation analysis, therefore, is determined as much by the underlying data as it is by the extraction algorithm chosen. Segmentation methods shape the segmentation solution.

- Distance-Based Methods :
  Market segmentation aims at grouping consumers into groups with similar needs or behavior, in this example: groups of tourists with similar patterns of vacation activities.

a. Distance Measures :

Numerous approaches to measuring the distance between two vectors exist; several are used routinely in cluster analysis and market segmentation.

A distance measure has to comply with a few criteria. One criterion is symmetry,
which is: $d(x, y) = d(y, x)$.

A second criterion is that the distance of a vector to itself and only to itself is 0: $d(x, y) = 0 \Leftrightarrow x = y$.

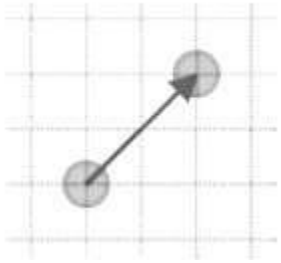
In addition, most distance measures fulfil the so-called triangle inequality :$d(x, z) \leq d(x, y) + d(y, z)$.

The triangle inequality says that if one goes from x to z with an intermediate stop in y, the combined distance is at least as long as going from x to z directly.

Euclidean distance is the most common distance measure used in market segmentation analysis. Euclidean distance corresponds to the direct "straight-line" distance between two points in two-dimensional space.

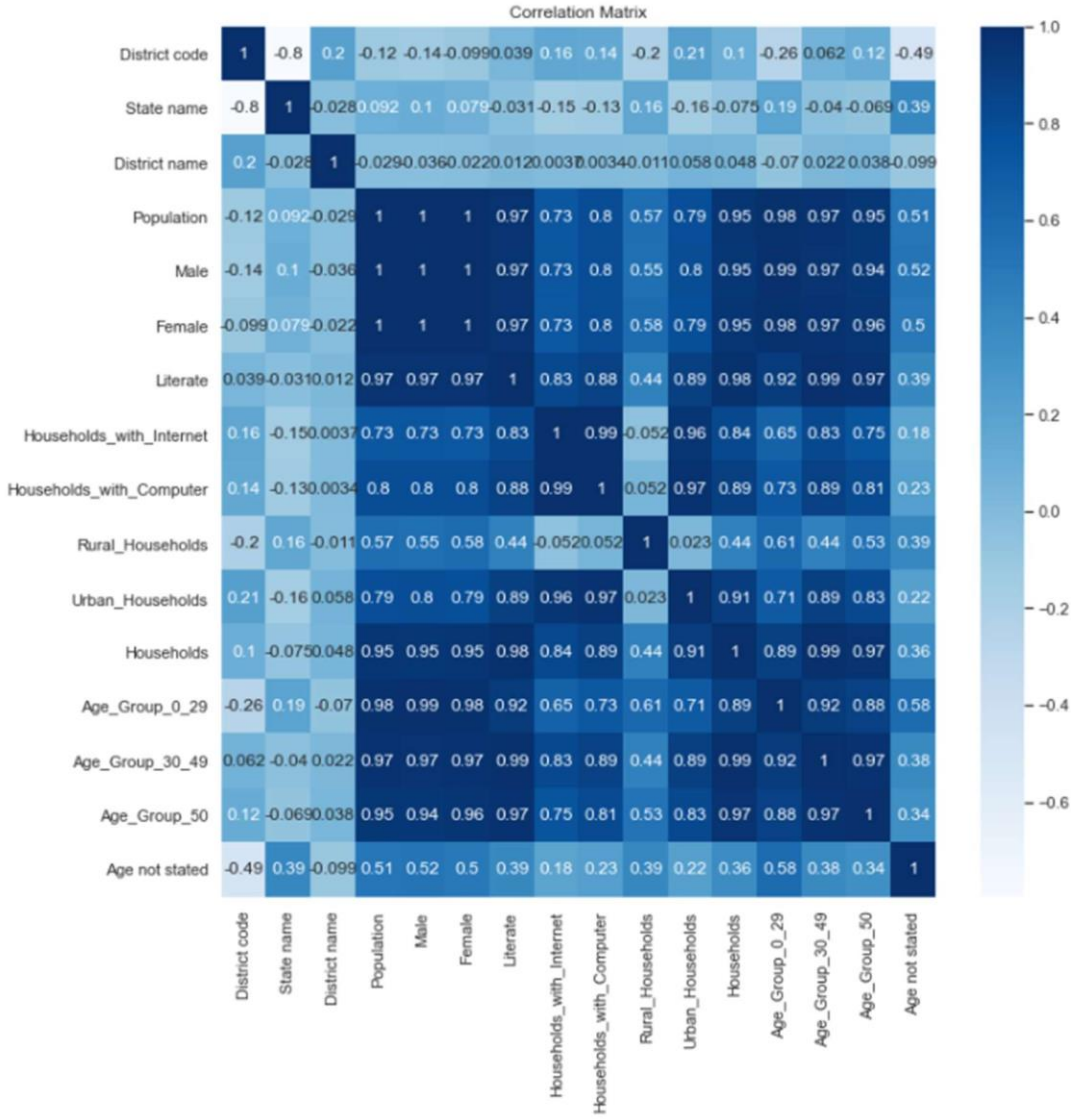
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

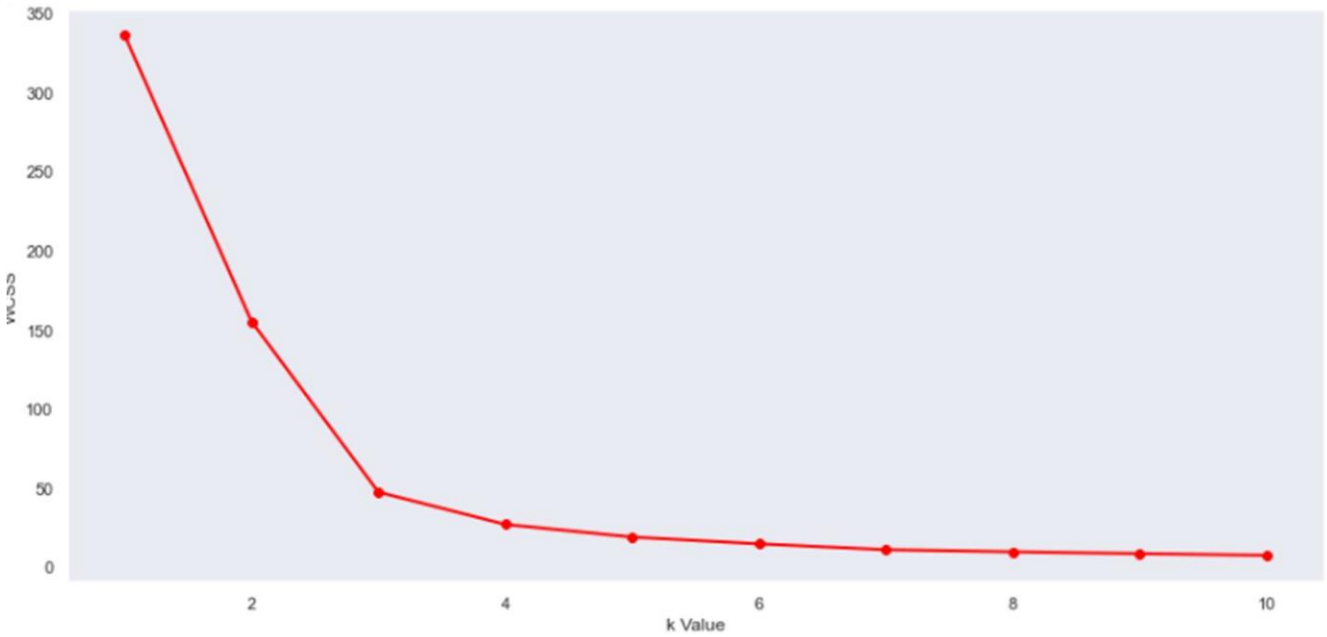
Euclidean distance



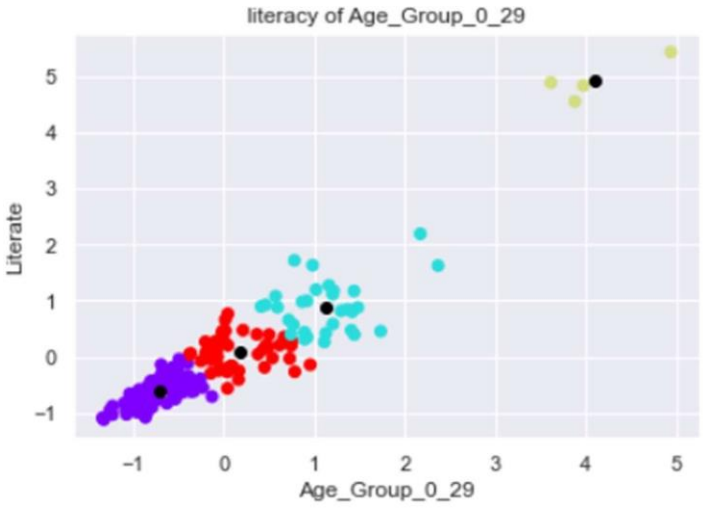
K Means Clustering Algorithm:

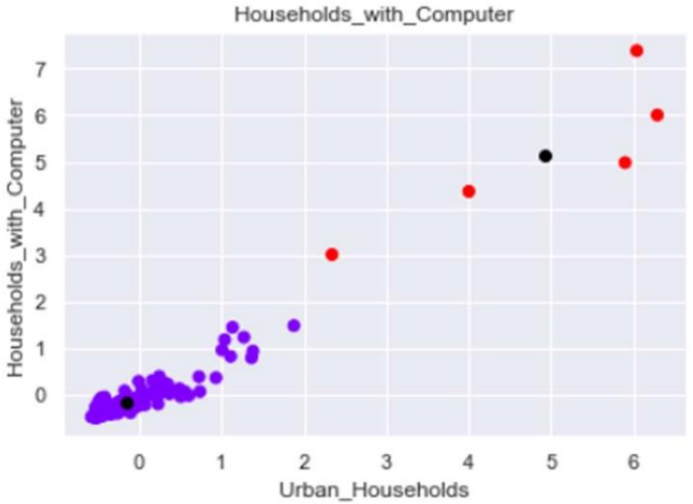
The k-means clustering algorithm is an iterative process of moving cluster centers or centroids to the mean position of their constituent points and reassigning instances to their closest clusters until there is no significant change in the number of cluster centers possible or a number of iterations are reached.

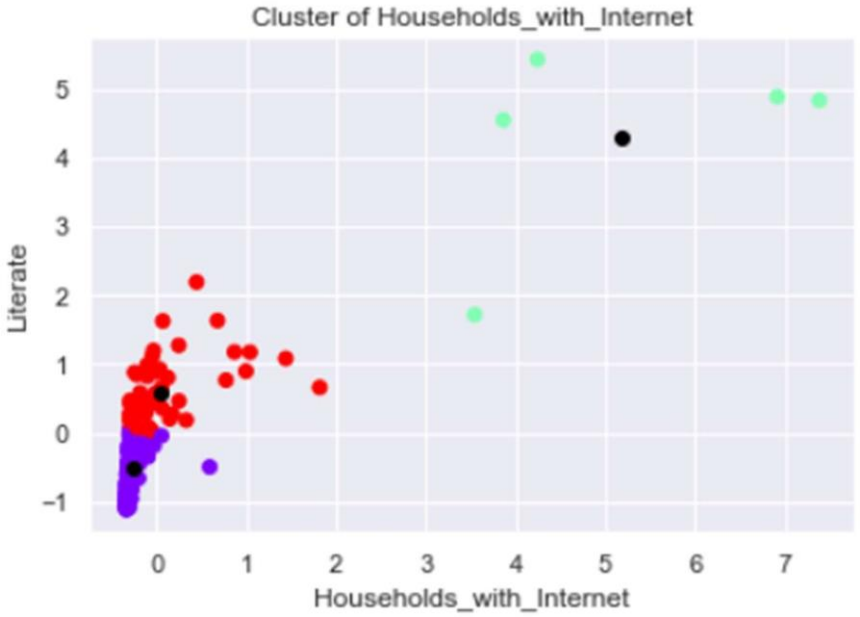

Correlation Matrix
District code
State name
District name
Population
Male
Female
Literate
Households_with_Internet
Households_with_Computer
Rural_Households
Urban_Households
Households
Age_Group_0_29
Age_Group_30_49
Age_Group_50
Age not stated
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


350
300
250
200
150
100
50
0
2
4
6
8
10
k Value


Cluster of literacy
Literate
Population
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5


literacy of Age_Group_0_29
Literate
Age_Group_0_29


Households_with_Computer
Households_with_Computer
Urban_Households


Cluster of Households_with_Internet
Literate
Households_with_Internet

Describing potential segments

1. Geographic Demographics Segmentation:

It consists of defining customers according to their location, dividing a country into regions, states, or states. Location does not mean that all consumers in a location will behave the same way, but the approach helps identify certain general patterns. In case of large companies these regions may be further subdivided into sizes – small, medium, and large.

In case of international marketing or global business different countries might be taken up as different market segments. In case of Indian Railways, they have northern railway, southern railway, eastern railway, Western Railway, North-Eastern^ Railway, Central Railway and so on so forth. Customers in different regions may have different cultures and may require marketing differently. India is a country of diversities. Another basis may be geographical density – urban, suburban, and rural.

It may be a good basis as the low-density markets require different price, promotion and distribution strategies. The next base may be locality (about which we have talked afterwards). Next basis may be climate – warm, cold, and rainy.

The other base may be urban, suburban, or rural. India's urban population may be further divided on the basis of cities – Tier I (8 cities -8% India's population), Tier II (26 cities – 4% of India's population), Tier III (33 cities -7% population), and Tier IV (5094 cities -11% population). The rest 70% is the rural population residing in India's 6,38,000 villages across India. In terms of types of commerce (Tourist, local worker, residents, businesses), retail establishments (downtown shopping districts, shopping malls), competition (underdeveloped, saturated), legislation (stringent, lax), and cost of living /Operation (low/moderate/high) are the other bases of geographical demographics).

2. Personal Demographics Segmentation:

It offers a wide variety of bases for segmentation.

i. Age:

Today virtually every age band from life to death is the focus of a marketing campaign. The requirements are different in different age groups. In case of readymade garments, it may be for new borne babies, teens, youth, middle-aged people, old people. All of them have different needs.

ii. Gender:

In case of clothes, it may be male and female, in case of fashionable clothes the two segments vary a lot. Women prefer scooties, and boys use motorcycles. By 2015, India will have 80 million working women in the age group of 18–44-year age band. The roles are changing because of womenfolk joining working groups. Now males do many jobs earlier performed by womenfolk.

iii. Family Structure:

The family life cycle concept charts the progress of family development from birth to death. A family may be in bachelor stage (young and single people), newly married couple -marriage alters the needs. Married couples need white goods and durable goods to begin with), Full Nest I (young married couple with dependent children -once a child is born, they would require baby food, baby clothes, toys, etc.), Full Nest II (older married couples with dependent children), Empty Nest (older married couples with no children living with them) and solitary survivor (older single People).

iv. Race:

The ethnic background is a good base for segmentation. Hindus celebrate Diwali, and Chinese celebrate their New Year differently and the two are good segments.

v. Political:

Different political party members have their liking for different members and commodities. For example, Congress party members in India prefer white caps, Samajwadi Party goes for red cap, BSP members want a blue cap, whereas BJP members wear a saffron colour cap.

vi. Family Size:

Two segments may be small family and the large family segments. Smaller the family small size packs would be preferred, and larger the family.

3. Socioeconomic:

i. Income:

Segmenting by income is very popular, expecially for cars, luggage, vacations and fashion goods. There may be people belonging to lower class, middle class and high net worth individuals. The housing boards offer low-income houses, middle income houses and high-income houses. The base for segmentation is income. It may be skill as well, like skilled workers, semi-skilled workers, unskilled workers, and subsistence workers (those living on state pension, casual or lowest grade workers), rich and poor.

A German statistician, Ernst Engel has made the following observations about what happens when household income increases:

1. Smaller percentage of expenditure goes for food.

2. The percentage spent on housing, household operations and clothing remain constant.

3. The percentage spent on other items (recreation, vacation, education) increases.

ii. Education:

College- going students have different demands than the people who after good higher education join the companies as executives, and those who are illiterates.

iii. Occupation:

The requirements for executives and a schoolteacher would altogether be different. The executive class would require Armani suit, whereas the other one would require a suit of any brand which is cheaper.

iv. Social Class:

Social class indicates one's social position, and is objectified through income, occupation, and location of residence. A policeman might be earning more than a college professor, off course through accepting under the table challans, but he belongs to a social class lower than that of a professor. The social class of professor will demand purchases of items and place of purchases different from that of a policeman.

4. Behavioural Segmentation:

Devid Kurtz likes to call it as production related segmentation. It takes into consideration

the purchasing behaviour as the starting point. The bases include:

i. Usage status

ii. Brand Loyalty Levels

iii. Benefit Sought

iv. Occasions for Purchase

vi. Willingness to Buy

5. Psychographic Segmentation:

Psychographic segmentation is related with similarity of values and lifestyles. People buy things because of the personality, lifestyle and the consumer values they hold.

i. Personality Characteristics:

Advertising agency, Young & Rubicam has classified customers into Mainstreamers (not to stand out of crowd), Reformers (creative and caring, many doing charities, and buying private labels), Aspirers (young, ambitious, and keen to get on, and buy latest designs and models), and success achievers (achieved in life, feel no need for status symbols or bother for what people will say). Companies marketing cigarettes, liquor, cosmetics and high priced watched create a personality for the brand to match it with the personality of the customer.

Briggs and Myers have developed four personality dimensions:

a. Extrovert/introvert

b. Sensitive/intuitive

c. Thinking/feeling

d. Judging/perceptive

ii. Lifestyle:

Lifestyle and consumption are closely related, and therefore, marketers adopt it for segmentation. AIO (Activities, interests, and opinions) reflect lifestyles of people. People are grouped on the basis of how they spend their time, the importance of things in their surroundings, beliefs about themselves and broad issues and some demographic characteristics, such as income and education.

The most popular consumer lifestyle framework is a survey from SRI Consulting Business Intelligence. It classifies customers into eight groups – Innovators, Thinkers, Achievers, Experiences, Believers, Strivers, Makers, and Survivors. A detailed profile of customers is necessary for developing effective advertising

campaigns.

iii. Values:

Values reflect the realities of life. Researchers at Survey Research Centre at University of Michigan have identified nine basic values: Self Respect, security, Excitement, Fun and enjoyment in life, having warm relationships, Self-fulfilment, Sense of belonging, Sense of accomplishment, and being well respected.

Marketing Mix

There are so many marketing ingredients in the early days of the market segmentation concept but as time goes by four tools were identified as commonly effective in this line of business which then comes to be known as the 4 Ps which are as follows

1. Product
2. Price
3. Promotion
4. Place

Market segmentation is not a marketing strategy in and of itself. Rather, it is intertwined with the other aspects of strategic marketing, the most essential of which are positioning and competition. In reality, the segmentation process is commonly viewed as part of the segmentation-targeting-positioning (STP) strategy. A sequential procedure is assumed in the segmentation-targeting-positioning strategy. Market segmentation (the extraction, profiling, and description of segments), targeting (the assessment of segments and selection of a target segment), and finally positioning (the measures an organisation can take to ensure that their product is perceived as distinct from competing products and in line with segmentation) are the

first steps in the process.

Targeting and customizing a mix that will attract the specific target segment is the main goal of this process and to make an effective mix for a specific target. Now, let’s dive into details for the 4P’s:

1. <u>Product</u>

The marketplace segments acquired for the Australian tourist through the view of Product. Imagine, for example, being a vacation spot with a completely wealthy cultural heritage. And believe having selected to goal a particular segment. The key characteristics

of this group of people have three contributors are that they have interaction much extra than the common visitor in travelling museums, monuments and gardens. They additionally love to do scenic walks and visit markets. The percentage of each of those tendencies with a number of the opposite marketplace segments.

Like maximum different segments, they prefer to relax, devour out, keep and have interaction in sightseeing. In phrases of the product-focused at this marketplace phase, viable product measures might also additionally encompass growing a brand-new product. For example, MUSEUMS and MONUMENTS enable contributors of this group to discover sports they're fascinated in, and factors to the lifestyles of those given on the vacation spot at some stage in the holiday planning process. Another possibility for focusing on this phase is that of proactively

making gardens on the vacation spot an enchantment of their personal right.

2. Price

Price is the cost consumers pay for a product. Marketers must link the price to the product's real and perceived value, but they also must consider supply costs, seasonal discounts, and competitors' prices. In some cases, business executives may raise the price to give the product the appearance of being a luxury. Alternatively, they may lower the price so more consumers can try the product.

Marketers also need to determine when and if discounting is appropriate. A discount can sometimes draw in more customers, but it can also give the impression that the product is less exclusive or less of a luxury compared to when it is was priced higher.

3. Place

When a company makes decisions regarding place, they are trying to determine where they should sell a product and how to deliver the product to the market. The goal of business executives is always to get their products in front of the consumers that are the most likely to buy them.

In some cases, this may refer to placing a product in certain stores, but it also refers to the product's placement on a specific store's display. In some cases, placement may refer to the act of including a product on television shows, in films, or on web pages in order to garner attention for the product.

4. Promotion

Promotion includes advertising, public relations, and promotional strategy. The goal of promoting a product is to reveal to consumers why they need it and why they should pay a certain price for it.

Marketers tend to tie promotion and placement elements together so they can reach their core audiences. For example, In the digital age, the "place" and "promotion" factors are as much online as they are offline. Specifically, where a product appears on a company's web page or social media, as well as which types of search functions, trigger corresponding, targeted ads forthe product.

Code Links

Github Link:-

https://github.com/RS99/Marketing-segmentation-on-Biotech