UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Student ID Number(s): 2655175

Programme: MSc Business Analytics

Module: Data Analytics and Predictive Modelling

Name of Tutor: Dr Hannan Amoozad Mahdiraji

Assignment Title: Population Glut in India

Date and Time of Submission: Time: 15th January, 04:24 AM

Actual Word Count: 3242

Extension: N **Extension Due Date:**

I do wish my assignment to be considered for including as an exemplar in the **School Bank of Assessed Work.**

The purpose of this template is to ensure you receive targeted feedback that will support your learning. It is a requirement to complete to complete all 3 sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).
Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution* • I should add precisely the information required in the assignment • I should take care of the word limit • Understanding the idea of the project and making the project accordingly.
Section Two: In this assignment, I have attempted to act on previous feedback in the following ways • I have added precisely the necessary information that was required in this assignment • I have taken care of the word limit

• I understood the idea of the project and made the project accordingly.
Section Three: Feedback on the following aspects of this assignment (i.e. content/style/approach) would be particularly helpful to me: (3 bullet points) • Relevancy of data • Inclusion of data structure framework and the table explaining variables • Clustering, Classification and Statistical Modelling

Abstract

As discussed in Assignment 1, I have introduced the problem of **Population Glut in India**. Various aspects of Indian Population such as literacy rate, poverty, urban-rural migration, healthcare were being reflected upon and the adverse effects of Population Glut on these factors will be shown. Different datasets from distinct sources will be used to evaluate the population effect on variables.

This purpose of the report is to evaluate the data sets that are used to address the questions and objectives of the report. Dataset on the population will be clustered into Densely Populated, Averagely populated and Sparsely Populated by K Means and Hierarchial method. Dataset will be classified into **28 States and 8 Union Territories**.

Later this report will focus on how Population Glut will affect different variables and features. I will further be predicting the future affect of Population Glut on differnet variables including literacy rate, poverty, urban-rural migration, healthcare etc.

Table of Contents

Introduction

In this report, we will be studying the affect of Population Glut on different variables and attributes.

The research questions that I will be addressing are:

1. To determine at what rate the population is increasing in India and define if it is alarming situation for India.
2. To determine and analyze the **fertility rate** of Indian population.
3. To determine how **Population Glut** and **Poverty** have negative affects on **Unemployment rate**.
4. To determine the population distribution within the country and address the issue of intense **urban-rural migration**.

The variables that I will be commenting upon will be **Population Growth Rate, Population Density, Fertility Rate, Unemployment Rate, Poverty Rate, Literacy rate**. We will be seeing the affect of population growth on these variables. I am briefing the variables below.

Population Growth Rate	This variable shows the population's percentage change over a given time period. Formula: (Current - Previous /Previous)*100

- It is useful to comprehend how the size of a population varies over time.

Population Density	The number of people residing in a given area, such as a square mile or kilometer, is known as the population density.

- It sheds light on how dispersed or concentrated a population is in a specific region.

Fertility Rate	The average number of children born to a woman during her reproductive years is referred to as her fertility rate.

- It is a significant demographic indicator that has the power to affect trends in population growth.

Unemployment Rate	The percentage of the workforce that is unemployed but nonetheless looking for work is measured by the unemployment rate.

- It has an impact on community health and is a crucial economic indicator.

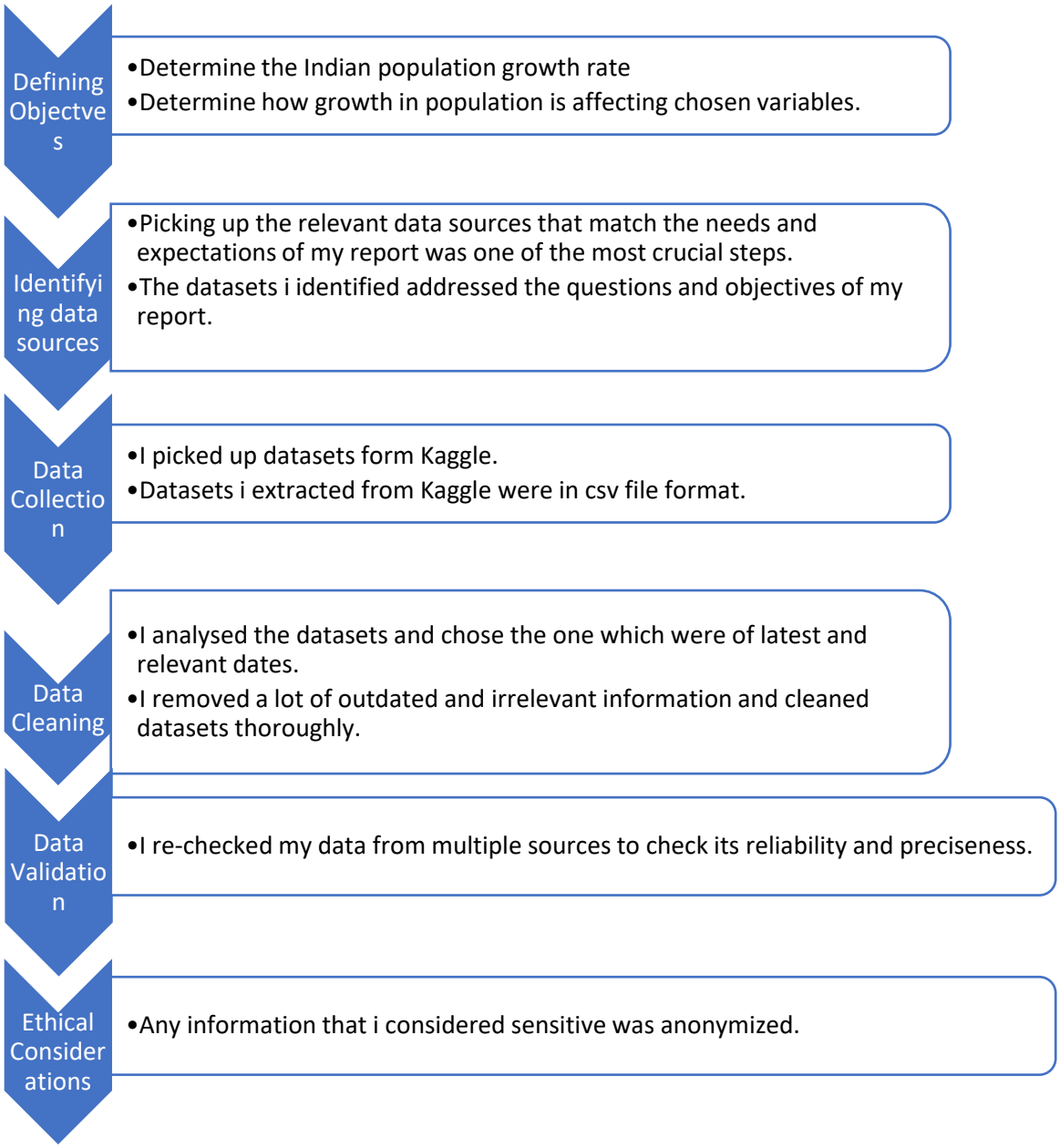
Poverty Rate	The share of people who are listed below the poverty line is known as the poverty rate.

- It can draw attention to differences in the allocation of income and aid in evaluating a society's financial condition.

Literacy Rate	The percentage of people who are capable of reading and writing is known as the literacy rate.

- It is an essential measure of the educational status of society and can impact both economic and social growth.

Data Gathering approach that I followed was:

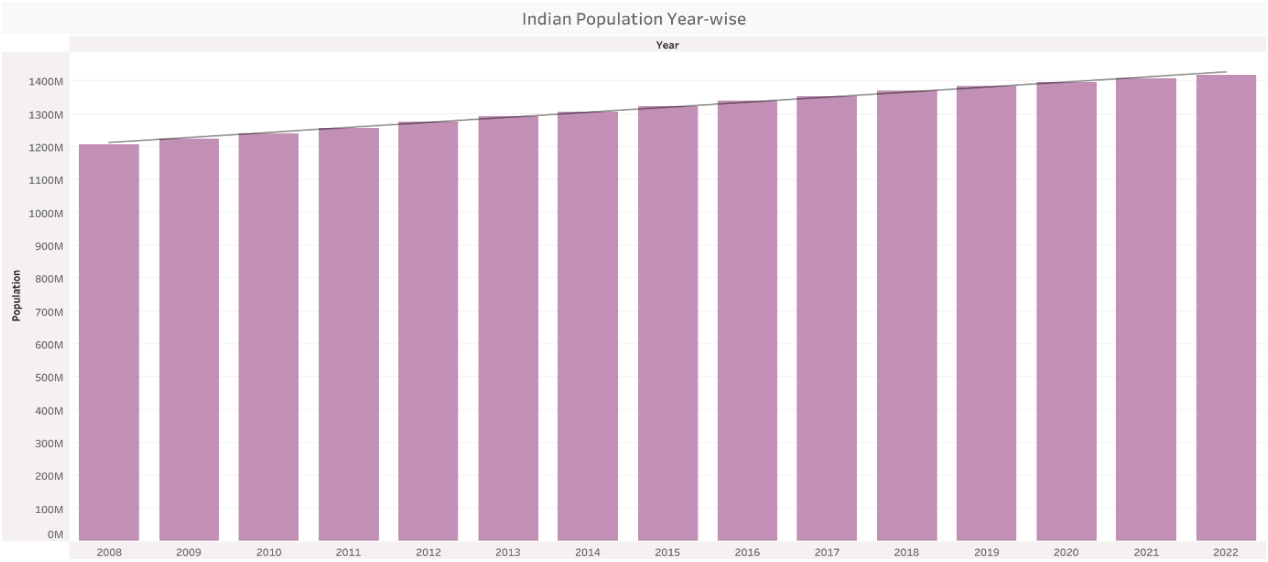


Data Analysis approach that I followed was:

EDA	• I listed the primary features of the data and represented it through graphs and plots.
Statistical Methods	• I chose appropriate statistical methods for analysis. • Used Descriptive statistics for summarizing data.
Data Visualization	• I made visual aids (charts, graphs, etc.) to effectively convey my findings. • I used visual tools like Tableau, Minitab and SPSS to create visual representations.
Interpretation	• I interpreted the results that i got from graphs. Decisions for futher plan of action are made after this.
Conclusion	• I summarized the results that i got by data analysis and concluded the interpretation.

Data Pre-processing

Population (2008-2022)



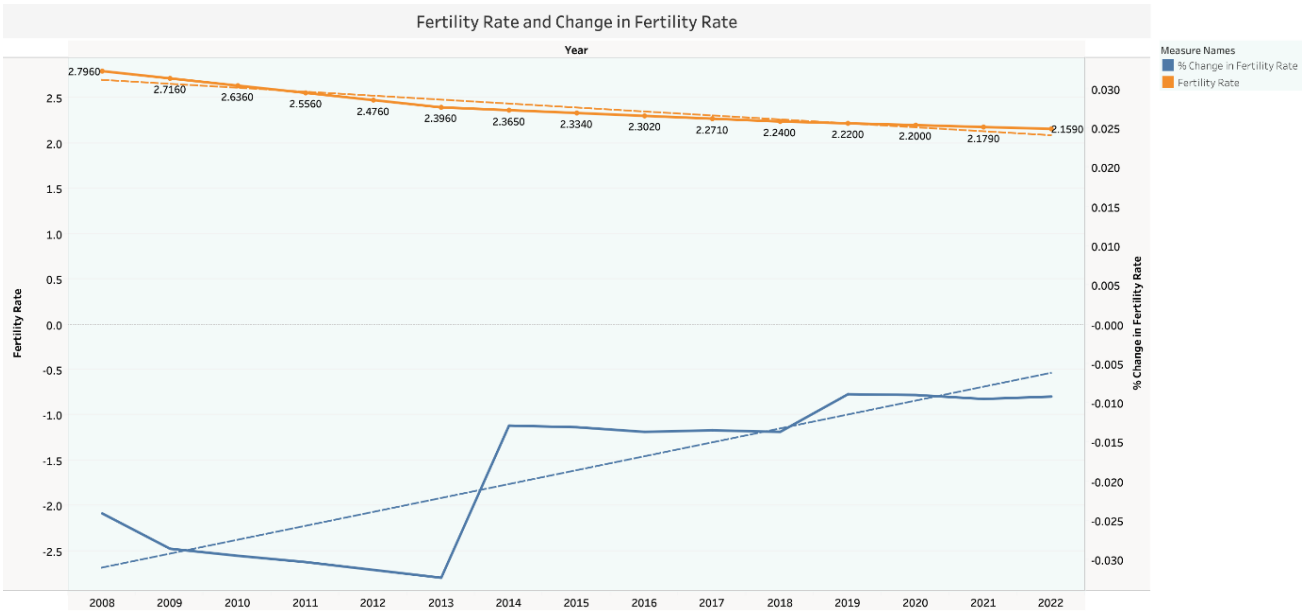
This graph represents the population in millions from **2008 to 2022**. We can see a gradual rise of population with the following years.

Variable	Mean	StDev	Variance	CoefVar	Minimum	Q1	Median	Q3
Population(In Millions)	1351.1	48.2	2322.7	3.57	1275.0	1307.0	1354.0	1396.0

Variable	Maximum	IQR	Mode	N for Mode	Skewness	Kurtosis
Population(In Millions)	1417.0	89.0	*	0	-0.18	-1.26

- This table shows Descriptive Satistics of the Indian Population from 1960 to 2023. It has a high Standard Deviation of **48.2**.
- In statistics, the **variance** serves as a gauge for how widely distributed a set of variables is.
- IQR is **89**. It is important to mention that this value is in between **Q1 and Q3**. The **outliers** are not considered here.

Fertility Rate

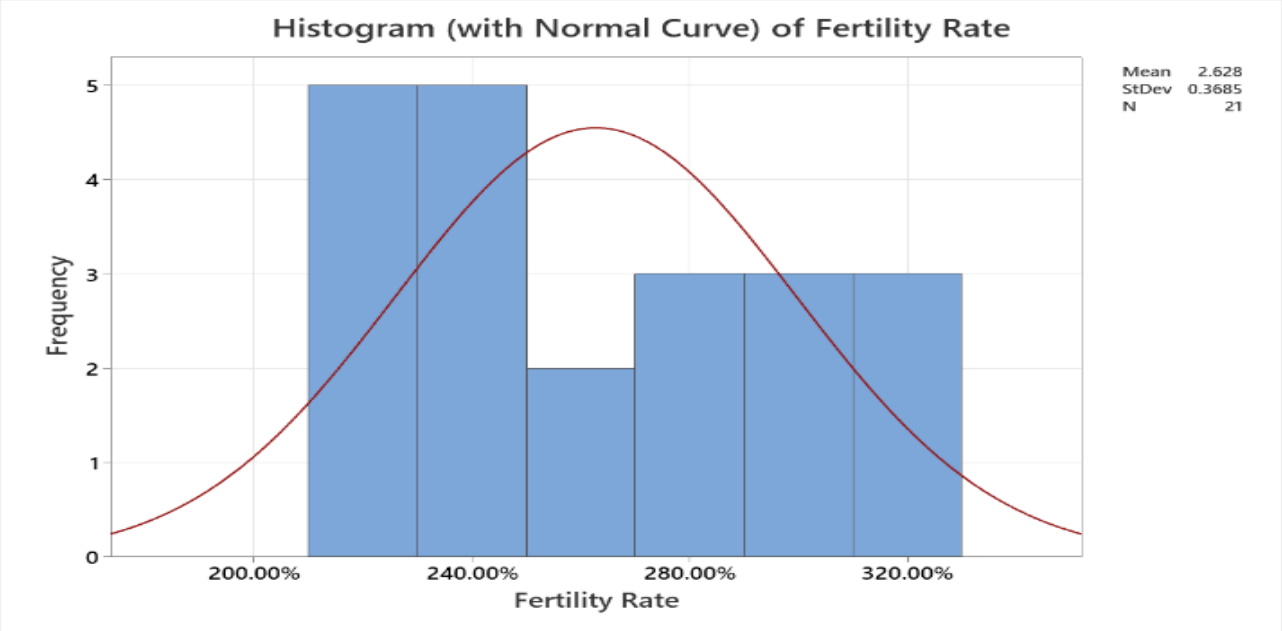


I used Tableau to make this above graph. The above graph used illustrates the **Fertility rate Year-wise**. It's a positive sign for India that fertility rate is continuously declining. As all the values are in negative in lower side of the above graph), this signifies that the fertilty rate is decreasing with following years as (Final)-(Initial) = Negative

Variable	Mean	SE Mean	StDev	Minimum	Median	Maximum	IQR	N for Mode
Fertility Rate	2.6279	0.0804	0.3685	2.1790	2.5560	3.2770	0.6815	0

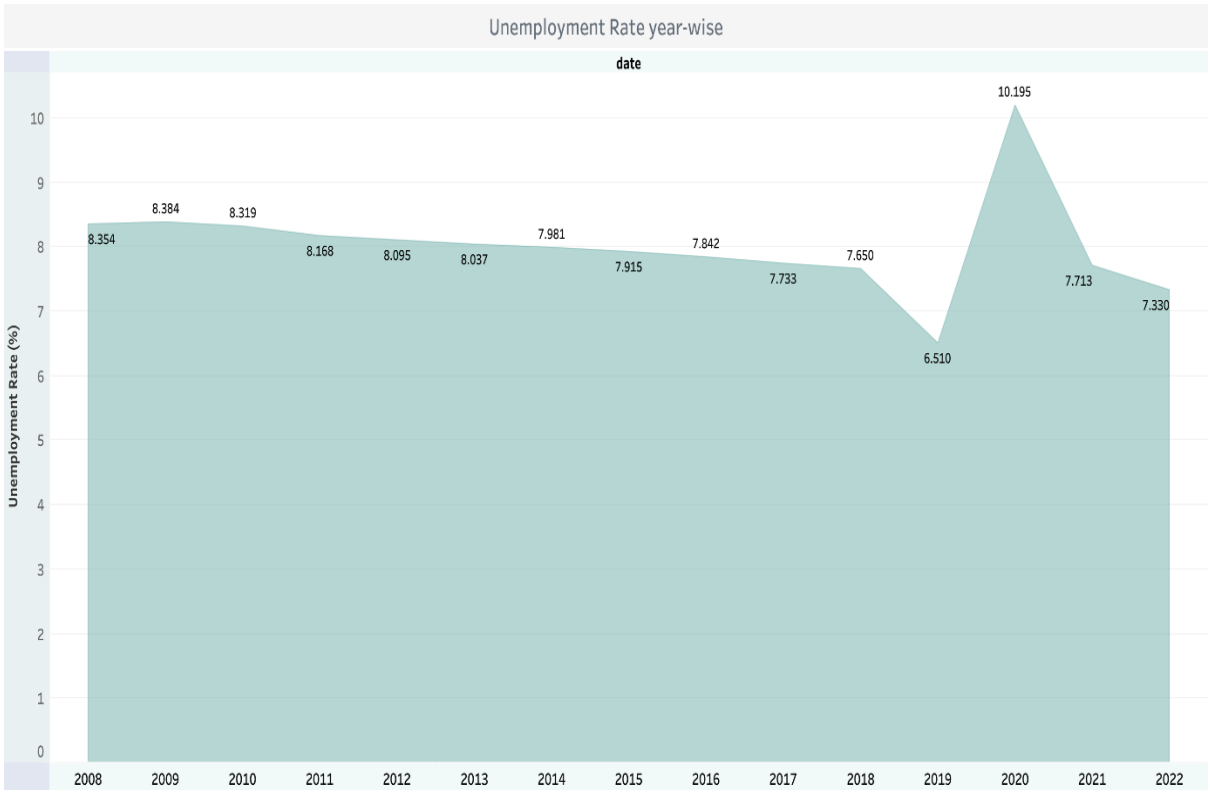
Variable	Skewness	Kurtosis
Fertility Rate	0.39	-1.32

- Mean, Median, Minimum and Maximum are very close to each other.
- A **standard deviation** of **0.37** denotes a very low degree of dataset variability or dispersion.
- **Skewness** of **0.39** indicates that the distribution is somewhat positively skewed. This implies that the mean might be shifted to the right by some values at the higher end of the distribution.
- The distribution appears to be **platykurtic** with a **kurtosis** of **-1.32**, indicating negative kurtosis.



This Frequency vs Fertility rate graph illustrates the frequency of the 21 values used to make this graph. We can observe that frequency is quite higher for values less than **2.4%.**

Unemployment (2011-2022)



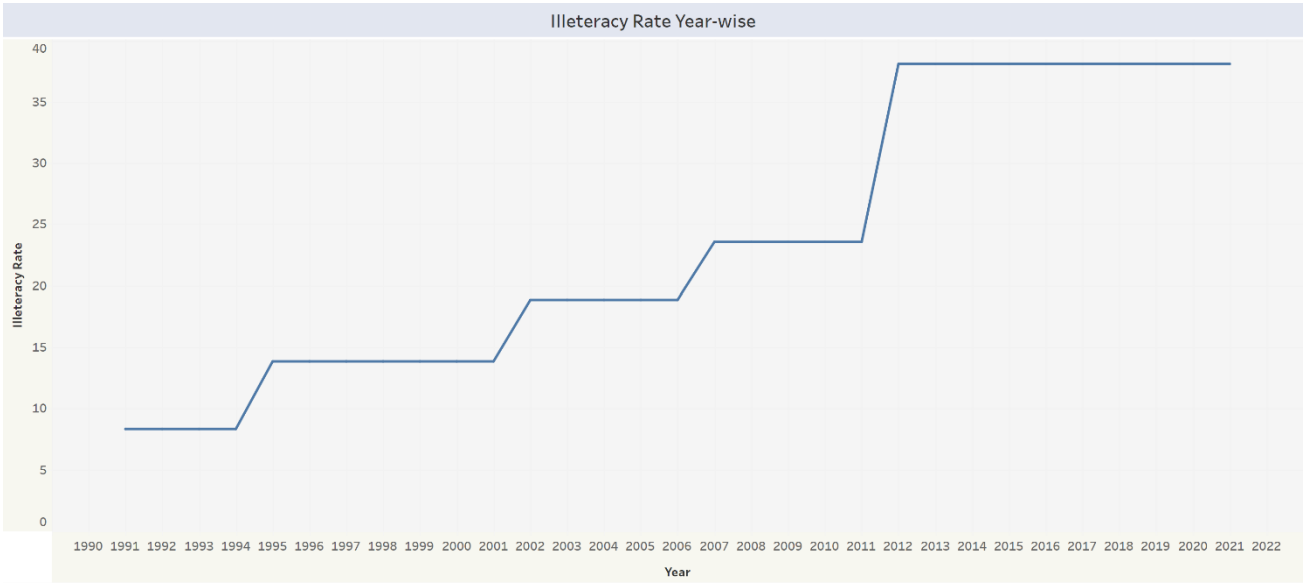
This graph shows the unemployment rate in India over the years. The values against each year has been labelled on the graph.

Variable	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum	Range
Unemployment Rate (%)	7.931	0.841	0.707	10.60	6.510	7.878	10.195	3.685

Variable	IQR	Mode	N for Mode	Skewness	Kurtosis
Unemployment Rate (%)	0.415	*	0	1.56	5.58

- In a country with 1.4 billion plus population. The mean of approx **8 %** over the past 10 years itself states the grim situation of employment in India.
- With a standard deviation of **0.84** and a variance of **0.7**, its a dataset with a moderate level of variability.
- A distribution with a moderately right skewness of **1.56** is indicated. This implies that the mean might be being pulled to the right by some values at the higher end of the distribution.
- A kurtosis of **5.58** shows that the distribution has positive kurtosis, which means it is leptokurtic.

Illeteracy Rate (1991-2021)

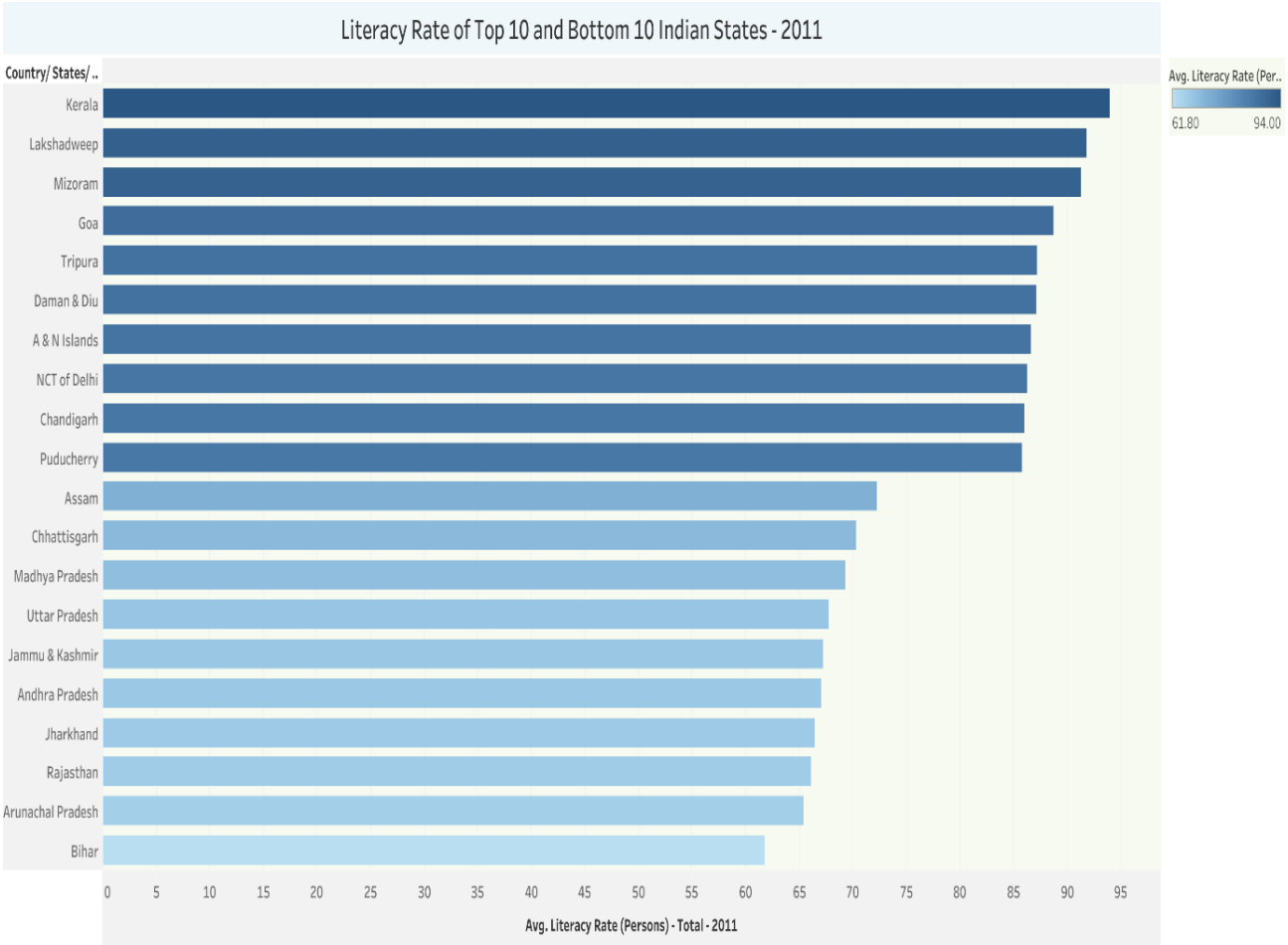


I used Tableau to make this graph. This graph clearly indicates that with increasing population, illeteracy rate is **increasing** drastically.

Variable	Mean	SE Mean	StDev	Minimum	Median	Maximum	IQR	Mode	N for Mode
Illeteracy Rate	23.34	2.02	11.27	8.34	18.87	38.10	24.24	38.1	10

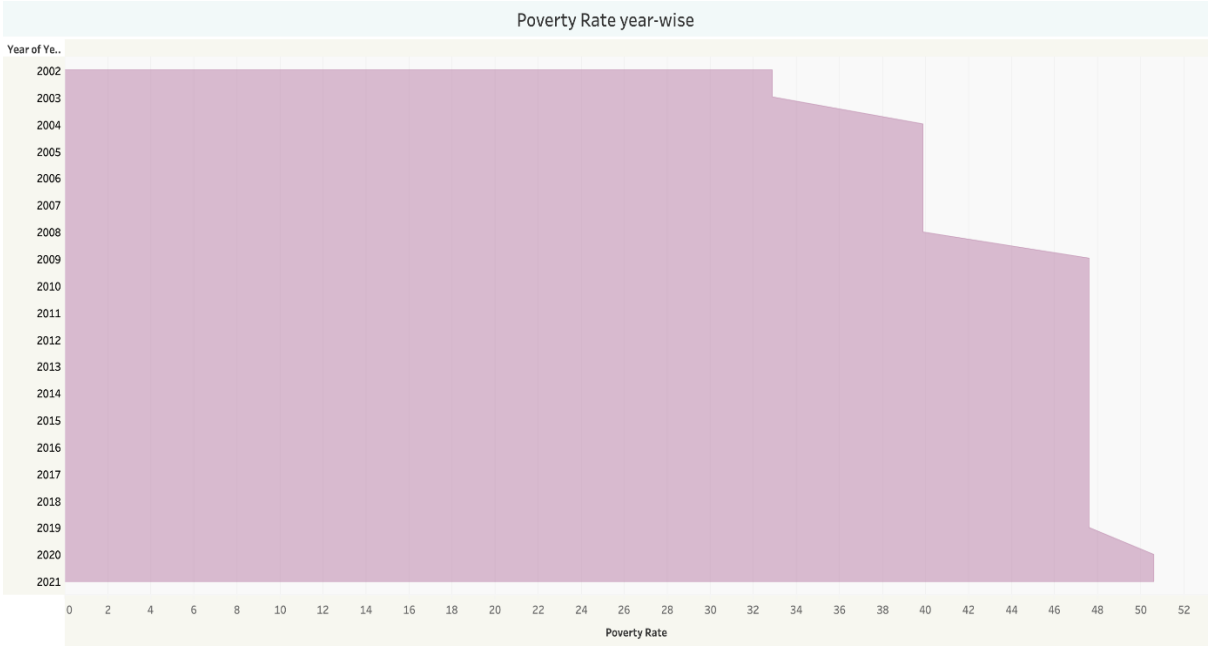
Variable	Skewness	Kurtosis
Illeteracy Rate	0.30	-1.47

- Standard Deviation of **11.27** means that the it has high deviation.
- The range varies from **8.34% to 38.10**% and illiteracy rate increases continously year by year as the population keeps on increasing.
- A distribution with a skewness of **0.3** has a little rightward skewness.
- A kurtosis of **-1.47** indicates that the distribution has **negative** kurtosis, suggesting that it is platykurtic. It has fewer extreme values or outliers than one with a normal distribution.



I used '**Create Parameter**' and '**Create Group**' features of Tableau to create this visualization. This graph shows the states and union territories against Literacy Rate within India. I filtered the States to Top 10 and Bottom 10 states by percentage of their Literate population.

Poverty Rate (2002-2021)



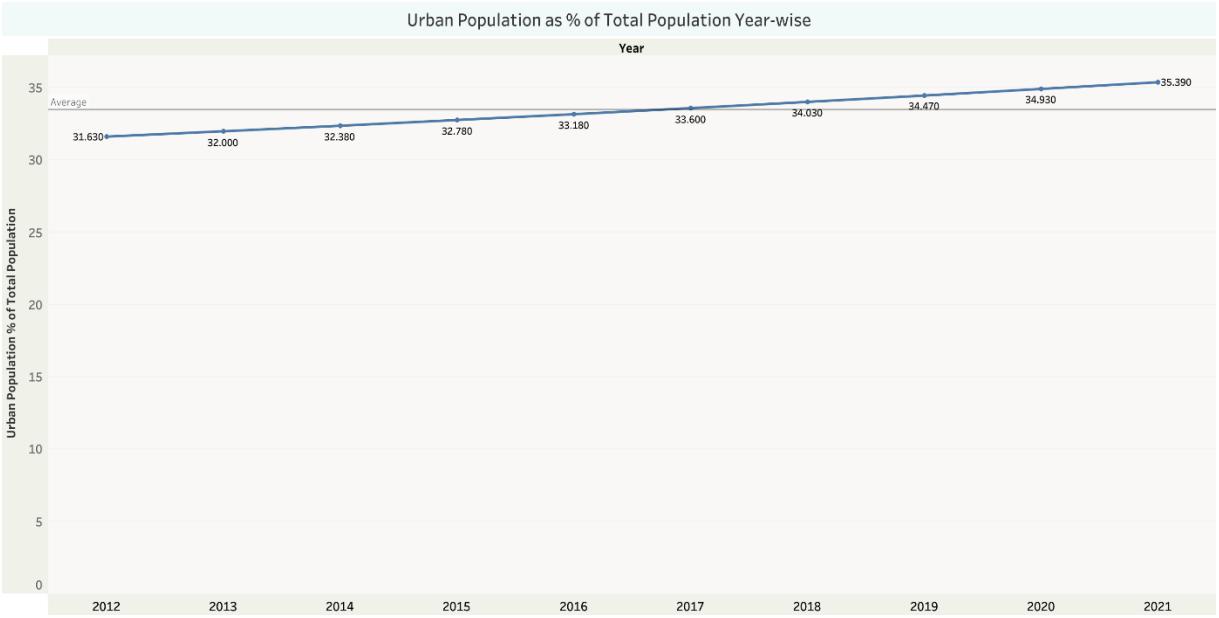
This graph clearly indicates how poverty rate has significantly increased from 2002(approx **33%**) to 2021(**50.6%**).

Variable	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum	Range	IQR	Mode
Poverty Rate	44.51	5.40	29.21	12.14	32.90	47.60	50.60	17.70	7.70	47.6

Variable	N for Mode	Skewness	Kurtosis
Poverty Rate	11	-1.05	0.04

- A dataset's moderate to significant level of variability is indicated by a standard deviation of **5.4**.
- Variance of **29.21** is quite high seeing the range of values therefore the data is **spread out and dispersed**.
- A distribution that is substantially left-skewed is indicated by a skewness of **-1.05**. This implies that the mean might be being pulled to the left by a few values around the bottom of the distribution. **(Kim, 2013)**
- Positive kurtosis in the distribution is indicated by a kurtosis of **0.04**. This suggests that there is a small leptokurtic distribution. **(Senger, 2013)**

Urban Population(% of Total) (2012-2022)



This graph made in Tableau indicates the trend of migation from rural to urban areas in India. The Urban population is continuously **increasing** year by year.

Variable	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum	Range
Urban Population % of Total Pop	33.439	1.266	1.604	3.79	31.630	33.390	35.390	3.760

Variable	IQR	Mode	N for Mode	Skewness	Kurtosis
Urban Population % of Total Pop	2.300	*	0	0.11	-1.19

- This graph has a standard deviation of **1.2**. Therefore the data is not dispersed much. This is also visible through the range it represents with a minimum of 31.63 and maximum of approx **35.4**.
- A distribution with a very minor right skew is indicated by a skewness of **0.11**.
- The distribution appears to have **negative** kurtosis when the kurtosis is **-1.19**. This suggests that platykurtic distributed.

Proximity Analysis

Proximity Matrix

Case	1:Uttar Pradesh	2: Bihar	3: Maharashtra	4:West Bengal	5: Madhya Pradesh	6: Rajasthan	7:Tamil Nadu	8:Gujarat	9: Karnataka	10: Andhra Pradesh	11:Odisha	12: Jharkhand
	Squared Euclidean Distance											
1:Uttar Pradesh	.000	1.E...	1.2E+16	1.87E+16	2.22E+16	2.392E+16	2.52E+16	2.696E+16	2.822E+16	3.33E+16	3.588E+16	3.850E+16
2:Bihar	1.19E+16	.000	1.4E+11	7.66E+14	1.61E+15	2.091E+15	2.49E+15	3.052E+15	3.489E+15	5.42E+15	6.477E+15	7.620E+15
3:Maharashtra	1.19E+16	1.E...	.000	7.45E+14	1.58E+15	2.058E+15	2.45E+15	3.012E+15	3.445E+15	5.36E+15	6.417E+15	7.555E+15
4:West Bengal	1.87E+16	8.E...	7.5E+14	.000	1.56E+14	3.261E+14	4.94E+14	7.605E+14	9.855E+14	2.11E+15	2.789E+15	3.554E+15
5:Madhya Pradesh	2.22E+16	2.E...	1.6E+15	1.56E+14	.000	3.085E+13	9.45E+13	2.272E+14	3.567E+14	1.12E+15	1.624E+15	2.220E+15
6:Rajasthan	2.39E+16	2.E...	2.1E+15	3.26E+14	3.08E+13	.000	1.73E+13	9.059E+13	1.778E+14	7.77E+14	1.207E+15	1.727E+15
7:Tamil Nadu	2.52E+16	2.E...	2.5E+15	4.94E+14	9.45E+13	1.735E+13	.000	2.865E+13	8.405E+13	5.62E+14	9.354E+14	1.398E+15
8:Gujarat	2.70E+16	3.E...	3.0E+15	7.60E+14	2.27E+14	9.059E+13	2.87E+13	.000	1.455E+13	3.37E+14	6.366E+14	1.027E+15
9:Karnataka	2.82E+16	3.E...	3.4E+15	9.85E+14	3.57E+14	1.778E+14	8.41E+13	1.455E+13	.000	2.11E+14	4.586E+14	7.967E+14
10:Andhra Pradesh	3.33E+16	5.E...	5.4E+15	2.11E+15	1.12E+15	7.767E+14	5.62E+14	3.368E+14	2.113E+14	.000	4.733E+13	1.874E+14
11:Odisha	3.59E+16	6.E...	6.4E+15	2.79E+15	1.62E+15	1.207E+15	9.35E+14	6.366E+14	4.586E+14	4.73E+13	.000	4.638E+13
12:Jharkhand	3.85E+16	8.E...	7.6E+15	3.55E+15	2.22E+15	1.727E+15	1.40E+15	1.027E+15	7.967E+14	1.87E+14	4.638E+13	.000

Proximity analysis is the study of spatial correlations, separations, or resemblances among data points. This graph represents the **similarity** in population among **12 Indian States** through distances. The smaller the distance, more is the similarity between them. The factor on which similarity and dissimilarity in these states depends upon is the 'Population'. This can also be called as **'Similarity Matrix' or 'Dissimilarity Matrix'**. **(Emamgholian, 2020)**

Lets discuss some important Data Pre-processing stages are important to maintain the uniformatity and proper flow of information:

Data Cleaning

Data cleaning involves identifying and rectifying errors in a database. I removed unnecessary rows, corrected dates, and eliminated irrelevant columns and null values to ensure reliable descriptive analysis.

Data integration

Data integration, merging information from diverse sources for a cohesive view, was crucial for my project's objectives. Careful consideration of data authenticity and viability was paramount.

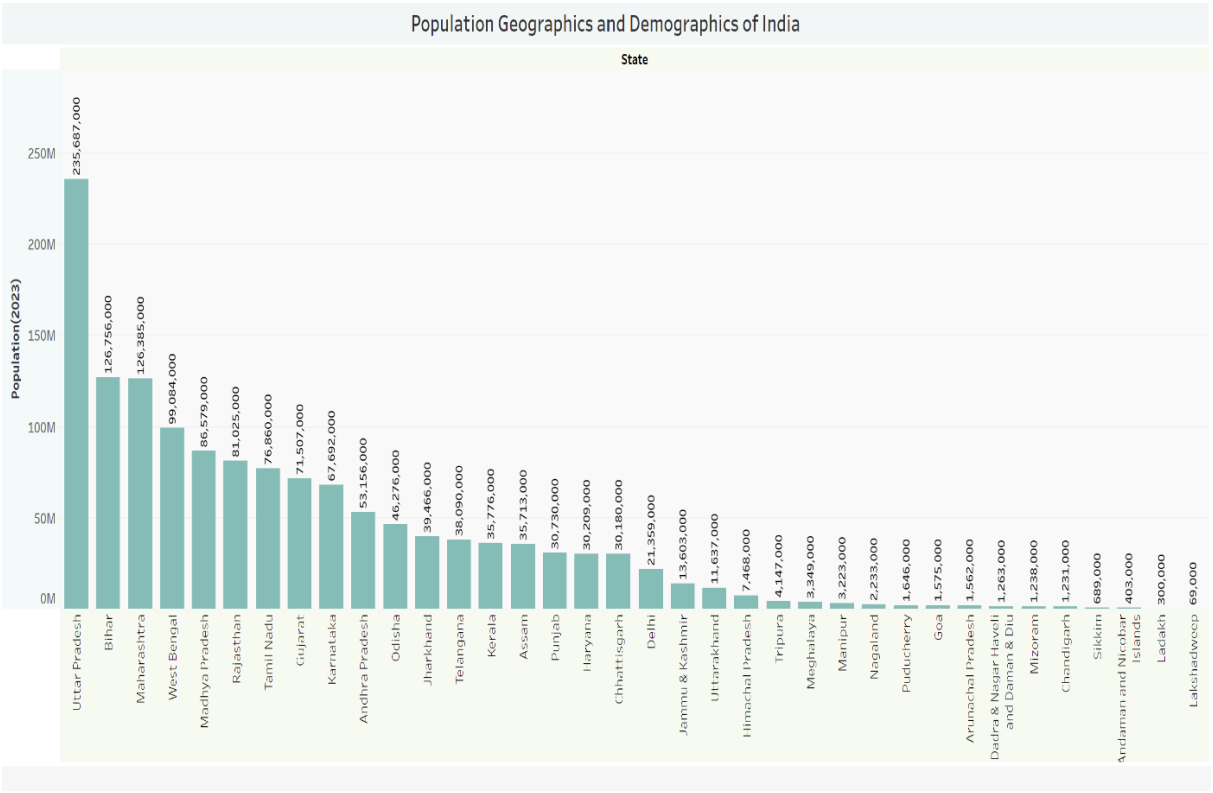
Data transformation

Data transformation occurred during Tableau visualizations, requiring changes in data types and the creation of new columns for population unit conversion. The project's success hinged on meticulous data management, encompassing cleaning, integration, and transformation, ultimately facilitating meaningful insights and conclusions from multiple datasets.

Clustering Analysis

I will be doing in depth clustering analysis for the dataset of Indian Population. **This dataset contains 4 columns namely States, Population, Male Population and Female Population**.

There are states with large population and states with minimilastic population too as indicated in the visualization below. The dataset I am using signifies the same. **(Setyaningsih, 2012)**



I have planned to cluster data on the basis of Population of States within India. **I have created categories 'Highly Populated', 'Averagely Populated' and 'Sparsely Populated'. 36 states and Union Territories within India is divided into one of these 3 categories.**

Therefore, I have formed 3 clusters with all the 36 States and Union Territories falling within one of these 3 clusters.

K Means Clustering

Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	35	99.9979	0.00014	31	32	31	2
2	34	99.9961	0.00026	28	29	28	2
3	33	99.9904	0.00064	30	31	30	3
4	32	99.9575	0.00283	17	18	17	2
5	31	99.9543	0.00304	14	15	14	2
6	30	99.9513	0.00324	27	28	27	3
7	29	99.9481	0.00345	34	35	34	2
8	28	99.9437	0.00374	24	25	24	2
9	27	99.8398	0.01066	34	36	34	3
10	26	99.8327	0.01114	2	3	2	2
11	25	99.8236	0.01174	27	30	27	6
12	24	99.7655	0.01561	16	17	16	3
13	23	99.7208	0.01859	33	34	33	4
14	22	99.5977	0.02678	23	24	23	3
15	21	99.5812	0.02788	26	27	26	7
16	20	99.4135	0.03904	12	13	12	2
17	19	99.1705	0.05521	20	21	20	2
18	18	99.0696	0.06193	26	33	26	11
19	17	98.4085	0.10594	12	14	12	4
20	16	98.3953	0.10682	8	9	8	2
21	15	98.2511	0.11641	23	26	23	14
22	14	98.2328	0.11764	6	7	6	2
23	13	97.4025	0.17290	20	22	20	3
24	12	97.0677	0.19519	10	11	10	2
25	11	96.0565	0.26250	12	16	12	7
26	10	95.8763	0.27449	5	6	5	3
27	9	94.2422	0.38326	20	23	20	17
28	8	92.3295	0.51058	12	19	12	8
29	7	91.9873	0.53336	5	8	5	5
30	6	88.2679	0.78094	2	4	2	3
31	5	86.5092	0.89801	10	12	10	10
32	4	77.4565	1.50059	10	20	10	27
33	3	74.9514	1.66735	2	5	2	8
34	2	46.2314	3.57908	2	10	2	35
35	1	0.0000	6.65645	1	2	1	36

In Step 1 of cluster analysis, **2 observations (31 and 32)** formed **35** new clusters with high similarity **(99.99%)**. As steps progress, the number of clusters decreases, reaching **3** clusters at Step **33** with a lower similarity of **74.95%**. The new cluster is labeled '**2**' and comprises **8** observations.

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	.000	34770000.000	15360296.296
2	.000	.000	.000

We achieved results quickly with just **2 steps** of Iteration. Therefore it suggests that the convergence to form final clusters was achieved quickly.

Cluster Membership

Case Number	State	Cluster	Distance
1	Uttar Pradesh	1	.000
2	Bihar	2	34770000.000
3	Maharashtra	2	34399000.000
4	West Bengal	2	7098000.000
5	Madhya Pradesh	2	5407000.000
6	Rajasthan	2	10961000.000
7	Tamil Nadu	2	15126000.000
8	Gujarat	2	20479000.000
9	Karnataka	2	24294000.000
10	Andhra Pradesh	3	37726703.704
11	Odisha	3	30846703.704
12	Jharkhand	3	24036703.704
13	Telangana	3	22660703.704
14	Kerala	3	20346703.704
15	Assam	3	20283703.704
16	Punjab	3	15300703.704
17	Haryana	3	14779703.704
18	Chhattisgarh	3	14750703.704
19	Delhi	3	5929703.704

20	Jammu & Kashmir	3	1826296.296
21	Uttarakhand	3	3792296.296
22	Himachal Pradesh	3	7961296.296
23	Tripura	3	11282296.296
24	Meghalaya	3	12080296.296
25	Manipur	3	12206296.296
26	Nagaland	3	13196296.296
27	Puducherry	3	13783296.296
28	Goa	3	13854296.296
29	Arunachal Pradesh	3	13867296.296
30	Dadra & Nagar Haveli and Daman & Diu	3	14166296.296
31	Mizoram	3	14191296.296
32	Chandigarh	3	14198296.296
33	Sikkim	3	14740296.296
34	Andaman and Nicobar Islands	3	15026296.296
35	Ladakh	3	15129296.296
36	Lakshadweep	3	15360296.296

Cluster membership table shows the clusters assigned to each and every State and Union Territory within India. States and Union Territories that are sparsely populated are assigned the cluster 3, states that are averagely populated are assigned the cluster 2, states that are highly populated are assigned the cluster 1.

Number of Cases in each Cluster

Cluster	1	1.000
	2	8.000
	3	27.000
Valid		36.000
Missing		.000

As shown **cluster 1 just contains 1 observation**. This observation is the State that has the highest population namely, Uttar Pradesh.

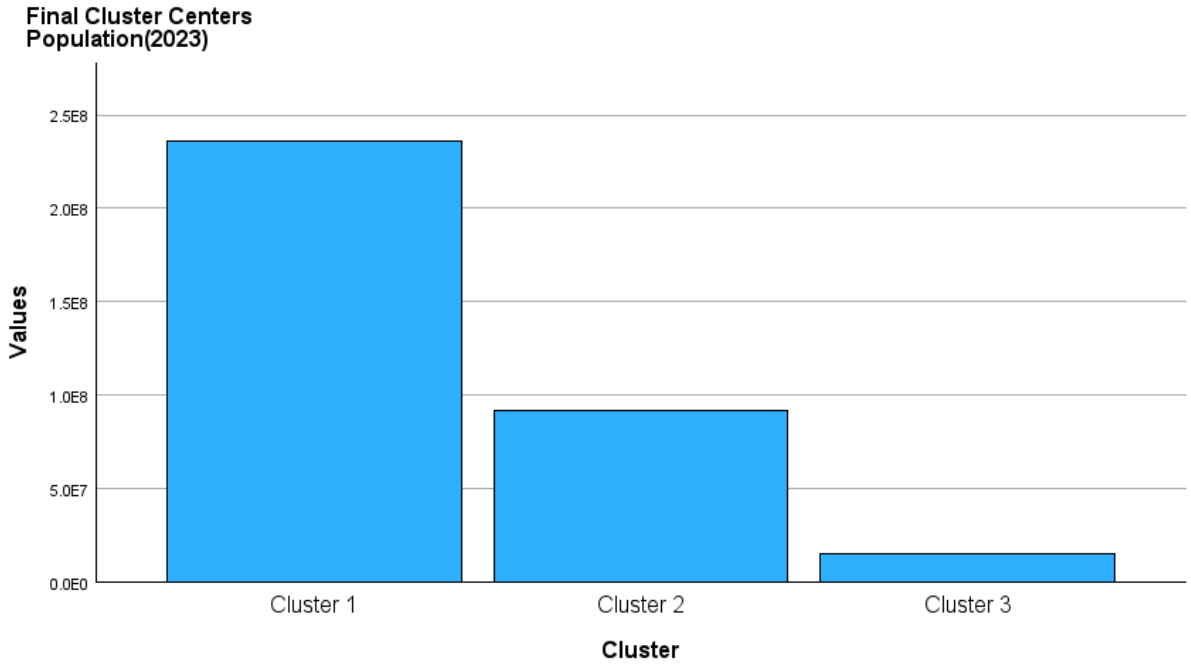
Cluster 2 has 8 observations, these states are the one that are 'Averagely Populated'.

Cluster 3 has 27 observations, these states are 'Sparsely Populated'.

Final Cluster Centers

	Cluster		
	1	2	3
Population(2023)	235687000	91986000	15429296

This table repesents the centeral value of different clusters.



This is graph showcasing centers of different clusters and their distance from the center.

Distances between Final Cluster Centers

Cluster	1	2	3
1		143701000.000	220257703.704
2	143701000.000		76556703.704
3	220257703.704	76556703.704	

This table showcases the distances between different cluster centers. This information provides us insights into how well separated the clusters are.

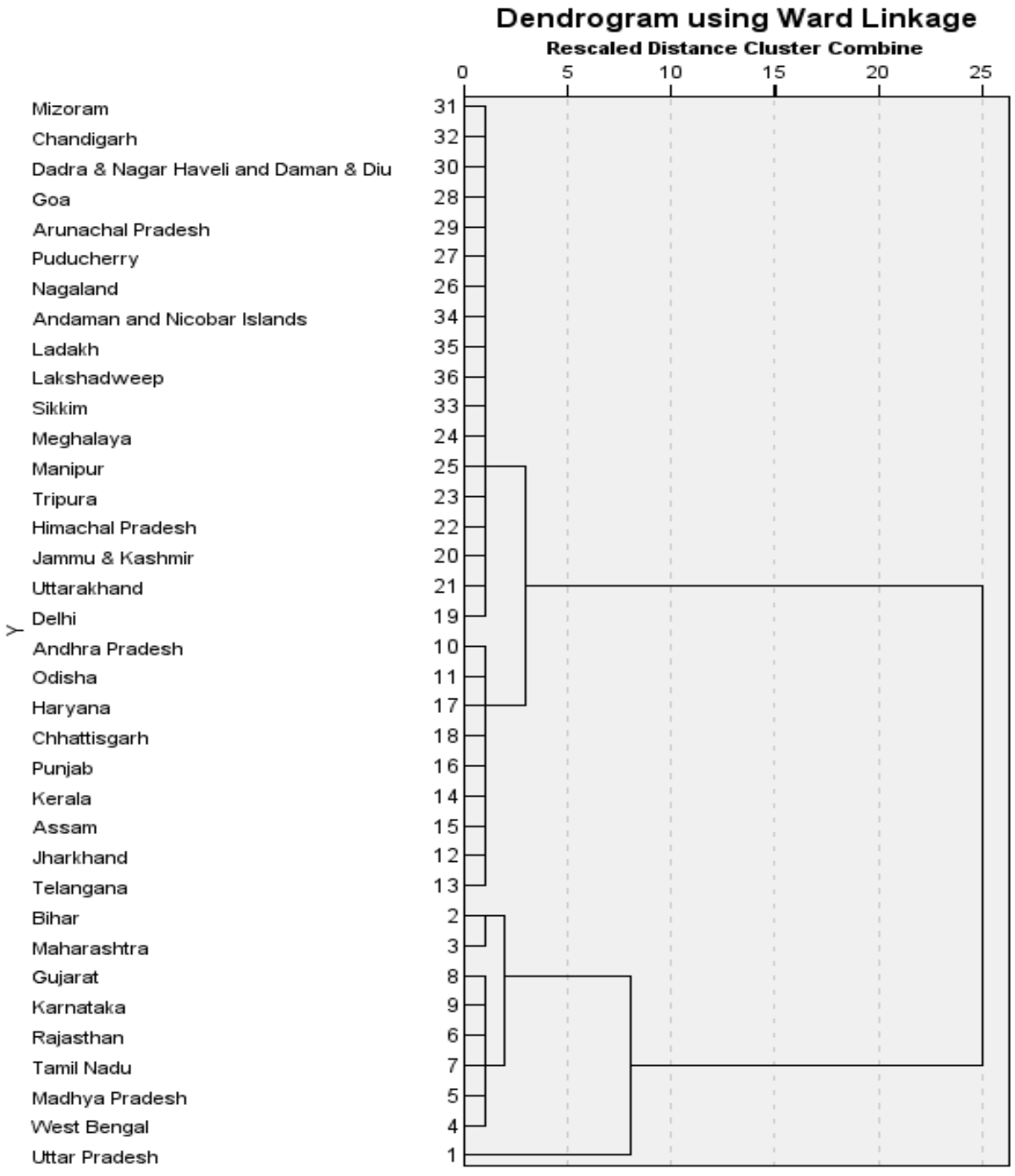
ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Population(2023)	3806977459 4685184.000	2	3512256215 03928.250	33	108.391	<.00

The last column stating **'Sig.'** tells us the Significance level that a particular variable holds in discriminating clusters. Value less than **0.001** states that the variable is really **significant**.

Hierarchial Clustering (SPSS)

Population of Indian States is clustered into 3 groups using **Ward** Linkage. Cluster 1 has 1 state, Cluster 2 has 8 states, and Cluster 3 has 27 states, as depicted in the Dendrogram.



The dendrogram in hierarchical clustering reveals **relationships** between data points. At the bottom are leaf nodes, representing individual data points. Intermediate nodes depict clusters merging, with branch height indicating **dissimilarity**; shorter branches signify **higher** similarity. The horizontal axis guides the order of merging. Clusters proximity on the dendrogram reflects their similarity, with the height of branches illustrating the **separation**. To identify clusters, one can cut the dendrogram at a specific height, determining the number of clusters desired. Understanding dissimilarity at different heights is crucial for interpreting the dendrogram and extracting meaningful clusters from the hierarchical structure.

Cluster Validity

Cluster Silhouettes

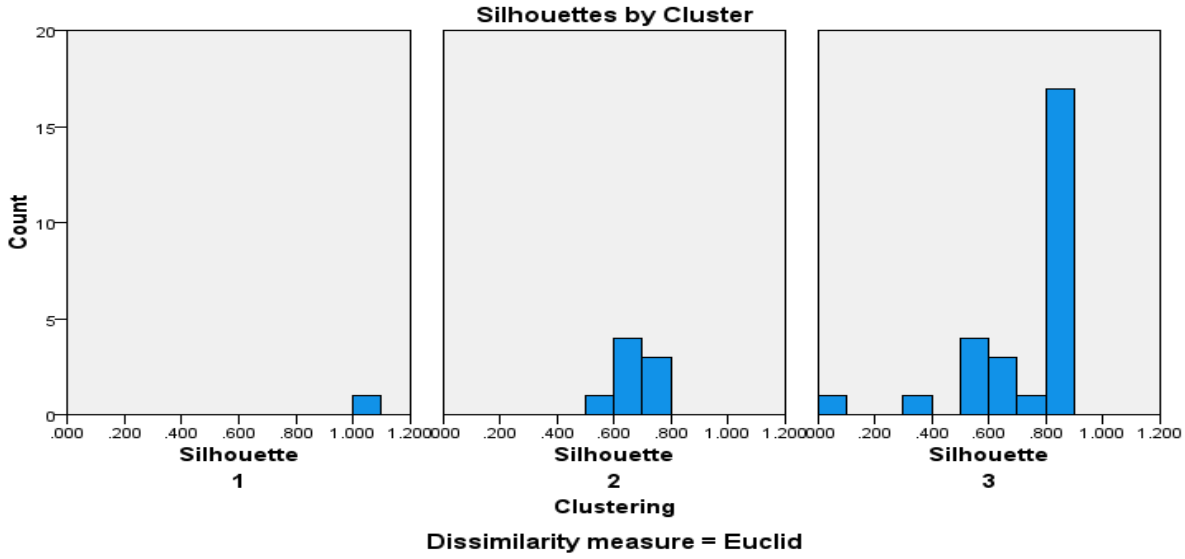
Silhouette Statistics

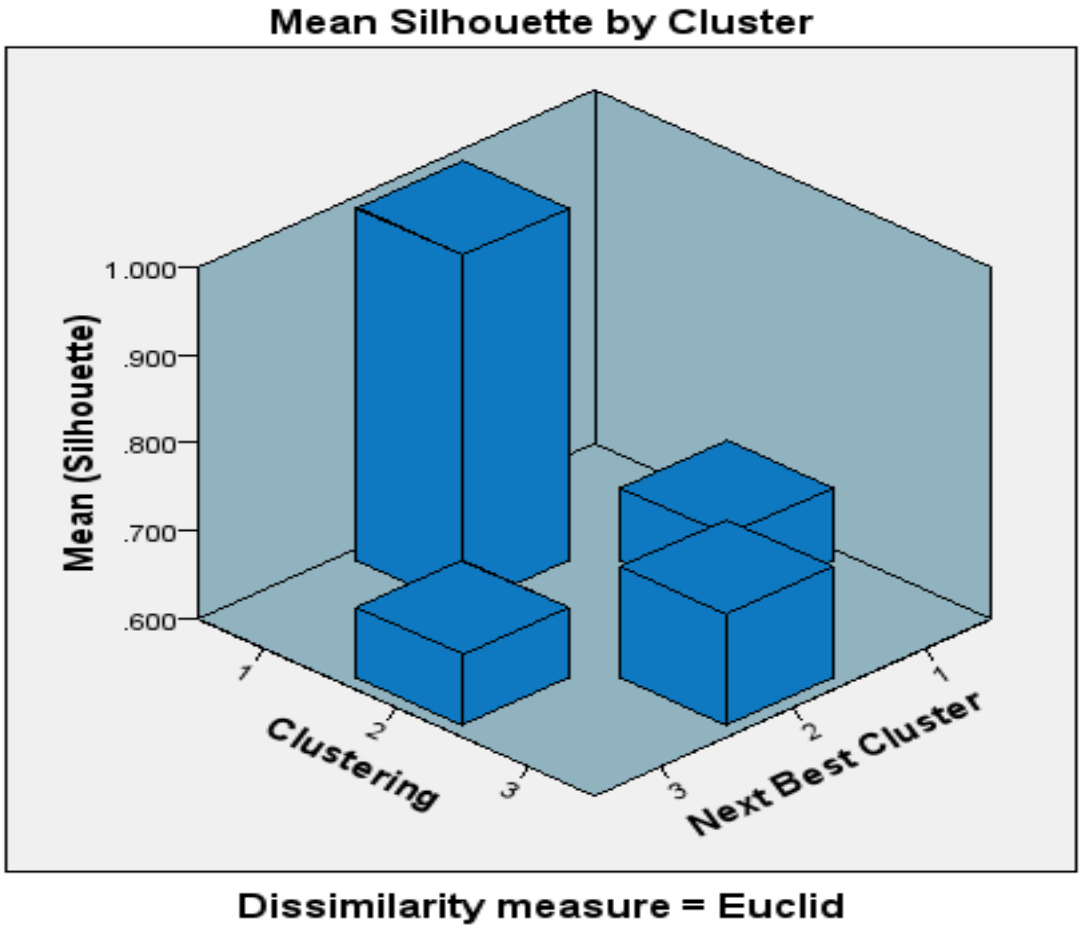
Statistics

Cluster	Case Count	Mean	Minimum	Maximum
1	1.000	1.000	1.000	1.000
2	8.000	.681	.535	.751
3	27.000	.726	.028	.843
Total	36.000	.724	.028	1.000

Dissimilarity measure = Euclid

Silhouette Graphs





A high silhouette score means that the object is well matched to its own cluster and poorly matched to nearby clusters. The silhouette score goes from **-1 to 1**. Above, cluster 1 has a flawless silhouette score of **1**, whereas clusters **2 and 3** show decent clustering despite having somewhat **lower** silhouette values.

Classification Analysis

Naive Bayes Classification (Population Density)

Population Density is defined as number of people living per square Km. Many states with high population can have low population density and vice versa. **(Chowdhury, 2023)**
Therefore, **Population Density= Population/Area(Sq Km)**

Case Processing Summary			
		N	Percent
Population Density		1	2.8%
	Average	13	36.1%
	High	10	27.8%
	Low	12	33.3%
Valid		36	100.0%
Excluded		0	
Total		36	

'Case Processing Summary' Table summarizes that there are **10 states with High Population Density, 13 with Average and 12 with Low**.

Subset Summary

Subset	Predictor Added	Rank	Pseudo-BIC	Average Log-Likelihood
0	(Initial Subset)[a]			
1	State or Union Territory	1	.118	-.069
2	Density	2	.161	-.061

To take out this result, I added 2 Predictors, namely **'State and Union Territory'** and **'Density'**.

Selected Predictors

	Predictors
Categorical	State or UnionTerritory

The categorical predictor i added is **State or Union Territory** that becomes the basis of its classification.

Classification Validity

Classification

Observed		Predicted			Percent Correct
		Average	High	Low	
	1	0	0	0	100.0%
Average	0	13	0	0	100.0%
High	0	0	10	0	100.0%
Low	0	0	0	12	100.0%
Overall Percent	2.8%	36.1%	27.8%	33.3%	100.0%

Dependent Variable: Population Density

This table shows the number of observations that falls under each category, namely (High, Average, Low).

One thing we need to keep in mind is that these categories show High, Mediuma and Low in comparison to each other and not the entire world. For instance, numbers with 'Low' Population Density in India will be 'High' for some other country.

As the classification results from '**High', 'Average','Low'** and an unclassified value sums upto 100%. Therefore the cassification is valid.

Tree'based classification

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Polpulation
	Independent Variables	Population(2023)
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	No Independent Variable Included
	Number of Nodes	1
	Number of Terminal Nodes	1
	Depth	0

Polpulation

Node 0

Category	%	n
Averagely Populated	22.2	8
Highly Populated	2.8	1
Sparsely Populated	75.0	27
Total	100.0	36

Classification Validity

Classification

Observed	Predicted: Averagely Populated	Predicted: Highly Populated	Predicted: Sparsely Populated	Percent Correct
Averagely Populated	0	0	8	0.0%
Highly Populated	0	0	1	0.0%
Sparsely Populated	0	0	27	100.0%
Overall Percentage	0.0%	0.0%	100.0%	75.0%

Growing Method: CHAID
Dependent Variable: Polpulation

Here **Averagely Populated(22.2%), Highly Populated(2.8%) and Sparsely Populated(75%)** makes up 100%. Therefore the Classification is **100%** valid

Statistiscal Modelling and analysis of Unemployment Rate through Regression Analysis

Here we will be discussing the effect of increasing Population and Poverty on **Unemployment Rate** in India. **(Kanyongo, 2006)**

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Poverety Rate, Population[b]	.	Enter

a. Dependent Variable: Unemployment Rate

This table shows us the variables that I used to determine how much of effect those have on Population Growth in India.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.696[a]	.485	.448	.559530

a. Predictors: (Constant), Poverety Rate, Population
b. Dependent Variable: Unemployment Rate

The **R Square** value of **0.485** in the Model Summary Table indicates that **48.5%** of the variation in the Unemployment Rate can be explained by changes in Population and Poverty Rate.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F
1	Regression	8.258	2	4.129	13.188
	Residual	8.766	28	.313	
	Total	17.024	30		

a. Dependent Variable: Unemployment Rate
b. Predictors: (Constant), Poverety Rate, Population

We can notice the **Sig.(Significance value)** to be less than **0.001**. This means the variables chosen**(Population, Povert Rate)** are quite **significant** in determining the Rate of Unemployment in India.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t
1	(Constant)	11.220	1.854		6.053
	Population	-5.517E-9	.000	-1.174	-2.422
	Poverety Rate	.088	.025	1.741	3.592

a. Dependent Variable: Unemployment Rate

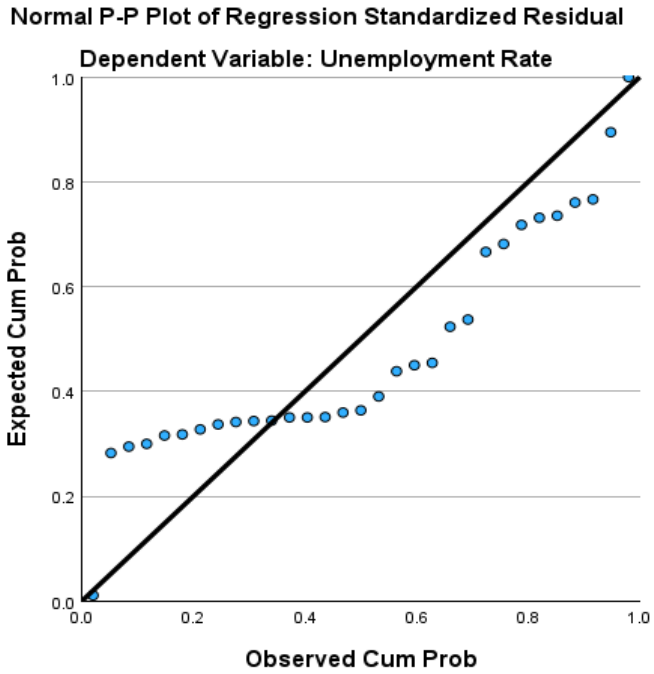
B coefficients signify how the dependent variable changes with a one-unit shift in the predictor, holding other variables constant. Beta coefficients, standardized in standard deviations, assess predictors' relative importance in explaining dependent variable variation. The **t-value**, obtained by dividing the coefficient by its standard error, gauges significance. A significant constant term **(11.220)** is supported by a **t-value of 6.053 and Sig. < 0.001.** With a Sig. Value of **0.022 < 0.05**, a **negative beta** suggests statistical significance in estimating the variable, indicating a potential **negative** association in the population.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	6.98578	8.67761	7.84926	.524655	31
Residual	-1.287736	2.205212	.000000	.540557	31
Std. Predicted Value	-1.646	1.579	.000	1.000	31
Std. Residual	-2.301	3.941	.000	.966	31

a. Dependent Variable: Unemployment Rate

Predicted values in regression represent the outcomes computed by the model for each observation, while residuals depict the distances between observed and predicted values. Well-fitted models show residuals with a mean close to zero. Examining the range and spread of residuals, including standardized residuals and their standard deviation, provides insights into the accuracy and variability of predictions.



This plot represents, the expected vs observed cum probability of the Regeression Standardized Residual. We can see the expected and observed values lie close to each other indicating that the plot made is correct.

Conclusion

This project helped me to explore all the hidden secrets within a dataset. Starting off form giving intro of the problem to slowly moving through discussions of project objectives, variables selected, my approach towards data gathering and data analysis. **Data pre-processing** was an important part where I cleaned the dataset by removing empty rows and columns. I removed all the null values and removed duplicate rows and columns. Modifying the data types of various columns in order to match the needs of the project. The **descriptve statistics** on the selected variables provided good insights on the data and helped me to analye the deviations, variances, mean, median and mode of the data provides. **Visualizations** I created using Tableau further helped me to discover the trends, pattterns and hidden facts within the dataset. It helped me to efficiently analyze the data and draw important statements and conclusions from it. **Proximity Analysis** helped me combine related data points according to pre-determined criteria. I was able to judge the degree of similarity and dissimilarity based on Proximity analysis.

Then I did clustering of the 'Population'. I did clustering using 2 methods **'K-Means Clustering'** and **'Hierarchial Clustering'**. I explained 'Amalgamation Steps','K-cluster Centers','Cluster Membership' etc. under 'K-means Clustering'. Under **Hierarchial Clustering**, i drew and explained the dendrogram diagram of the clusters formed. For both the clustering models, I made total of 3 clusters. All three clusters were later classified as 'Highly Populated', 'Averagely Populated' and 'Sparsely Populated'. I used SPSS and Minitab to so data clustering. While doing classification of the dataset, the two models that I picked were **'Naïve Bayes Classification'** and **'Tree Based Classification'**. I picked 'Population Density' as a dependent variable for Naïve Bayes and 'Population' as a dependent variable for 'Tree' based classification. Different subheadings like Model Summary, Subset Summary, Predictors etc helped me to properly evaluate and write discussions on the results.

For Statistical Modelling and Analysis, I used **Regression Analysis** to run analysis as well as do predictive modelling. I also compared the existing values with the predicted values and draw the similarity matrix which further validated the results I got. Understanding the type and strength of correlations between variables is made easier with the use of regression. It sheds light on the relationships between changes in one variable and those in another.

There were many limitations while moving forward with the project. I had to connect and correlate data from many datasets to address and answer all the objectives and research problems. I had to make new columns and fill up information in to address research objectives. Here I would recommend to pick up research objectives carefully so that they can be researched upon easily. Descriptive Statistics often relies upon quality of data. Outliers, values that are absent or inaccuracies might affect how trustworthy and valid the results are. Make data quality a top priority

by preprocessing and filtering the data to address data gaps, outliers, and inconsistencies. While doing custering **SPSS** and **Minitab** didn't offer option to make graph showcasing different clusters. I would recommend to introduce this option to visually represent the clusters. Another lmitation comes with Statistical Modelling and Analysis of data. These kind of Predictive modelling are often based on assumptions based on the underlying data points and association between the variables. These cannot be 100 percent relied upon as some real life factors which can't be measured on statistics for example, government policies that can drastically affect the results. We can run **sensitivity analysis test** and analyse the impact of each variable on the final outcome.

References

Cassen, R., 1999. *India: looking ahead to one and a half billion people.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1116811/

Chowdhury, S., 2023. *Research Paper Classification using Supervised Machine Learning Techniques.* [Online]
Available at:
https://www.researchgate.net/publication/346853360_Research_Paper_Classification_using_Supervised_Machine_Learning_Techniques

Emamgholian, S., 2020. *Exploring the applications of 3D proximity analysis in a 3D digital cadastre.* [Online]
Available at: https://www.tandfonline.com/doi/full/10.1080/10095020.2020.1780956

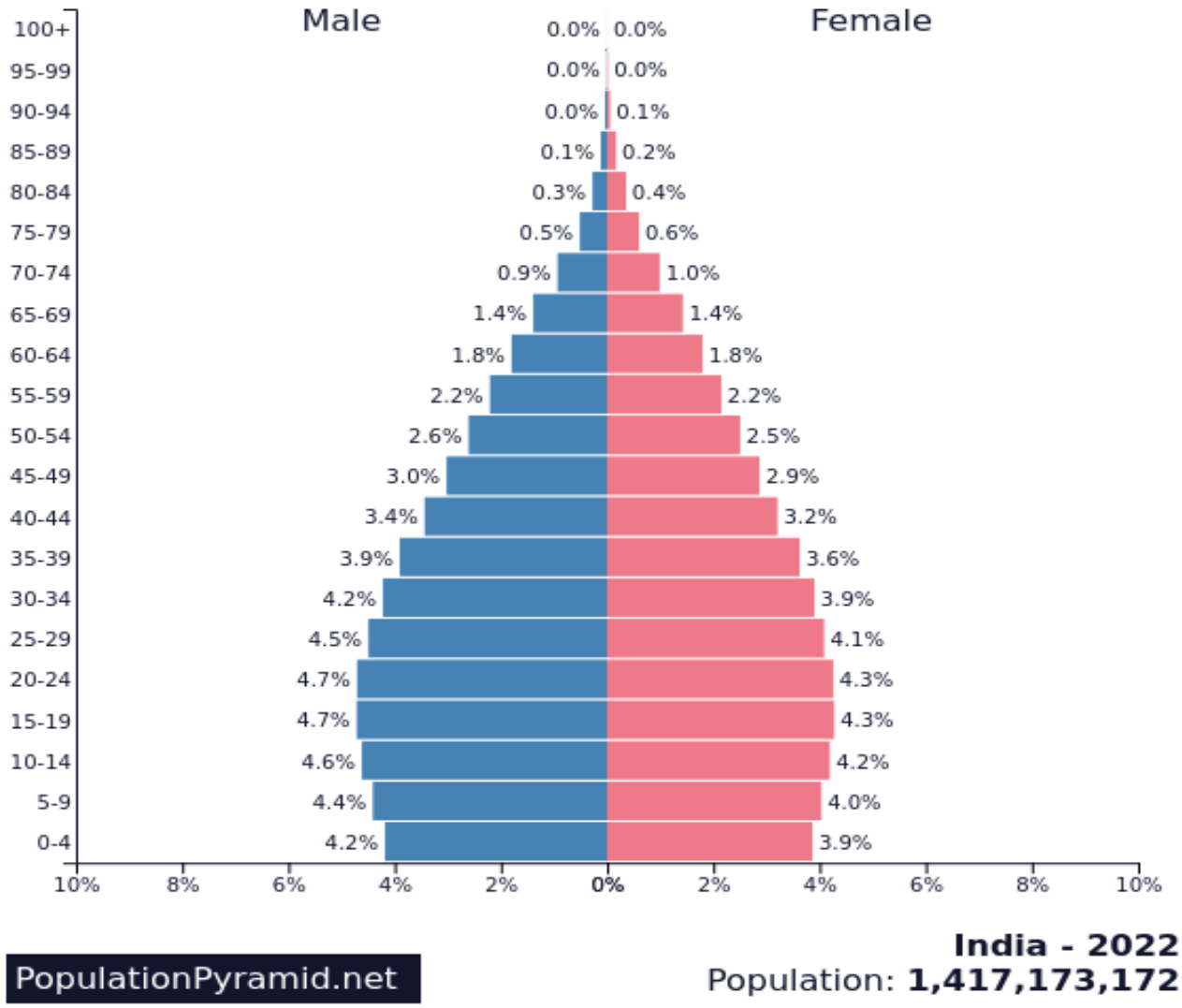
Kanyongo, G. Y., 2006. *Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe.* [Online]
Available at: https://files.eric.ed.gov/fulltext/EJ854316.pdf

Kim, H.-Y., 2013. *Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3591587/

Senger, Ö., 2013. *Impact of Skewness on Statistical Power.* [Online]
Available at:
https://www.researchgate.net/publication/269672002_Impact_of_Skewness_on_Statistical_Power

Setyaningsih, S., 2012. *Using Cluster Analysis Study to Examine the Successful Performance Entrepreneur in Indonesia.* [Online]
Available at:
https://www.researchgate.net/publication/271580351_Using_Cluster_Analysis_Study_to_Examine_the_Successful_Performance_Entrepreneur_in_Indonesia

Appendices



I was studying age demographics while studying the Indian Populaion.