Visualizing and predicting the population demography of India

Authors: Nikhil Srinivasan and Ranjith Kumar

Coauthor's: Prawin
and Manohar

ABSTRACT

In India, there is a platform which shows the exact population, growth rate this platform helps to make plans and decisions using the population information. There may be many problems when it comes to manipulating the data and also projecting/predicting the future population and at a certain time like this where we can't get the data manually because of the Pandemic situation. To clear these problems and to get the data without placing anyone in harm's way we have developed a population Prediction algorithm using a machine learning. We have adopted a Linear Regression Model and also Polynomial Regression Model. Moreover, the population is difficult to predict because of some circumstances that can alter a location's population and displace them to different places. Events like Virus outbreaks, Natural calamities, Natality rate, Mortality rates and migration. Natural Calamities can alter population demographic very-fast, especially when it is a result of life-changing outcomes like earthquakes, tsunamis, volcanic eruptions. In this event, one area's demographic population is reduced while another place's population Grows. The results have shown the linear regression has higher percentage error when compared to Polynomial regression. We have also visualized the population data of the years 2001 and 2011 for a better understanding of what the future may hold for us.

Keywords – Population Prediction, Population Visualization, Machine Learning

LIST OF FIGURES

TABLE OF CONTENTS

I.INTRODUCTION

An Essential part of most successful countries is that they have to be ready to face any challenges that the future may hold for them. This requires planning for both short as well as long term. The data with the help of which these planning’s are done should be reliable as well as based on the present situation. However, the future will not always be a straight line, sometimes it may go up or sometimes even down but the data should always adjust along with the events so that it can be very much reliable. In our case the data of the population must very accurate in order for us to predict with future with low error rate’s. The information obtained from our project may help the country prepare itself for a stable and positive population growth. The visualized data will help us understand how the population was distributed throughout the country and will help us get a clear understanding how the population ratio was for many number of things like education, working population, literates.

II. Understanding the Problem

Predicting the population trend of humans is a very complicated process. There many numbers of uncertainties with a demography of a country. Many traditional approaches can be used in predicting a countries population but these methods are not suitable for this purpose because these methods assume many things like the population growth will only be linear and will not take into account the factors such as Natality rates and Mortality rates. These things could lead to a prediction where there are many errors. Secondly population growth is difficult to predict because some events like earthquakes, tsunamis, volcanic eruptions could alter a person’s location in a short period of time. Migration can rapidly alter a country’s demography especially when it is caused by events which cause high mortality rates like war. In case of events like war the population of that particular place will decrease while the population of another place will increase due to the exodus. Some model like polynomial regression models will have higher accuracy when compared to models like linear regression models because they consider the factors of demography while the latter does not. Optimizing the polynomial model will give us a prediction with high accuracy rate.

III. METHODOLOGY

We have adopted a Polynomial regression model. Polynomial regression is a special case of linear regression where we fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables.

In a curvilinear relationship, the value of the target variable changes in a non-uniform manner with respect to the predictor (s).

In Linear Regression, with a single predictor, we have the following equation:

$$Y = \theta_0 + \theta_1 x$$

where,

Y is the target,

x is the predictor,

$\boldsymbol{\theta 0}$ is the bias,

and $\boldsymbol{\theta 1}$ is the weight in the regression equation

This linear equation can be used to represent a linear relationship. But, in polynomial regression, we have a polynomial equation of degree ***n*** represented as:

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_n x^n$$

Here:

θ0 is the bias,

θ1, θ2, …, θn are the weights in the equation of the polynomial regression,

and ***n*** is the degree of the polynomial

The number of higher-order terms increases with the increasing value of ***n,*** and hence the equation becomes more complicated.

We have used three different datasets from Kaggle for both visualization and prediction. For visualization we have used the population data of the years 2001 and 2011 which includes data such as population of male, population of female, population of male literates, population of female literates, population of people who have received primary, middle and secondary education, total number of graduates and population of working class for each state of India. For the Prediction of population, we have used a dataset which has the actual population of India from the year 1950 to 2011, using this dataset we have predicted the population of the Indian country from 2012 to 2030. The Visualization and prediction were done using data science and machine learning through python programming language.

IV. RESULTS AND DISCUSSIONS

A. Visualization

1) 2001

1) Population

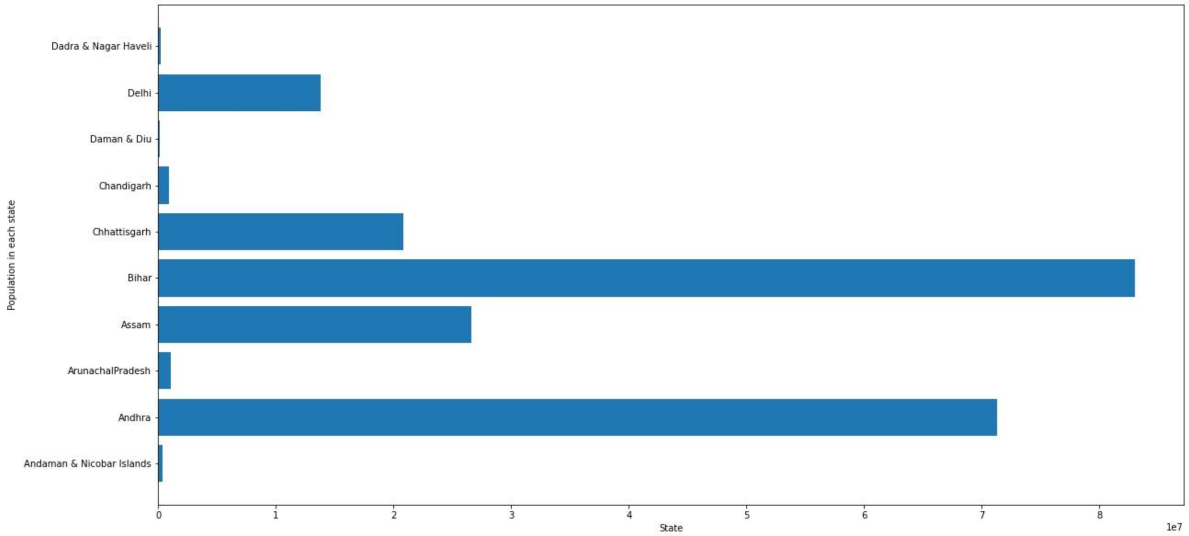

Figure-1

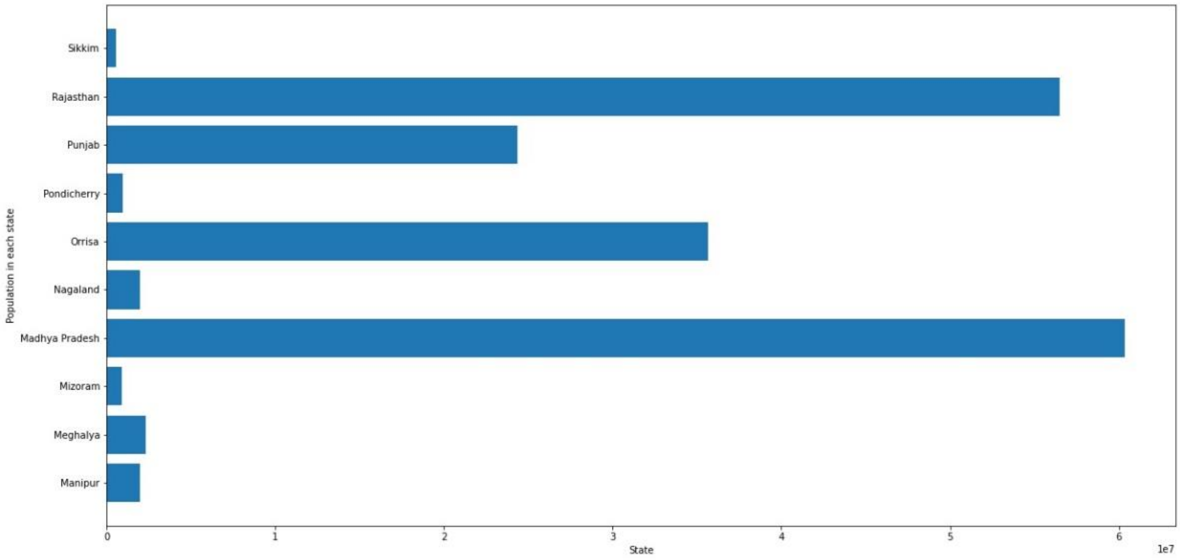

Figure-2

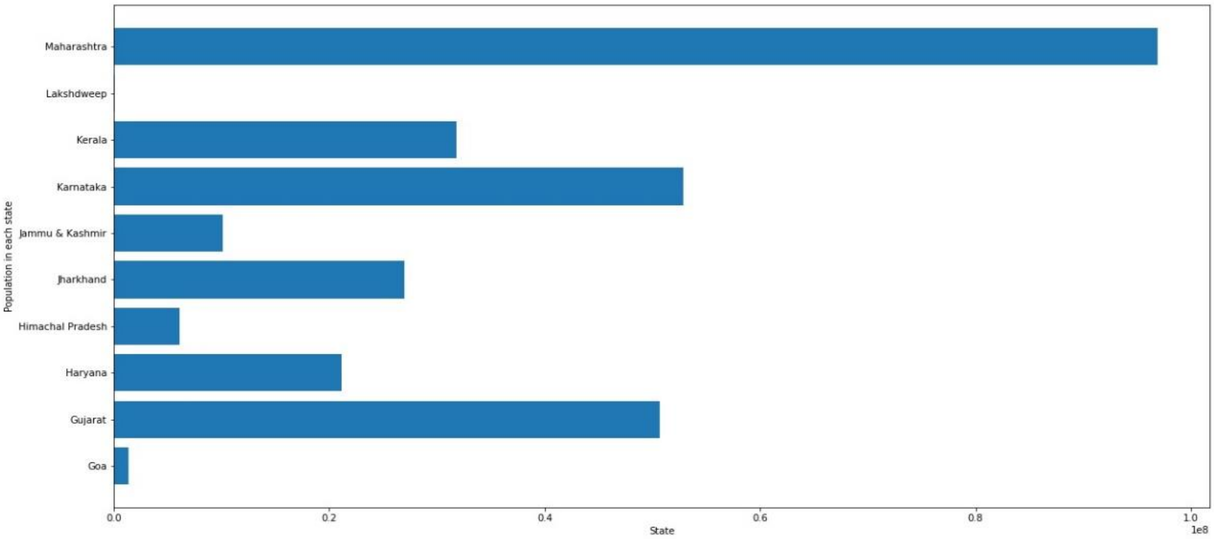

Figure-3

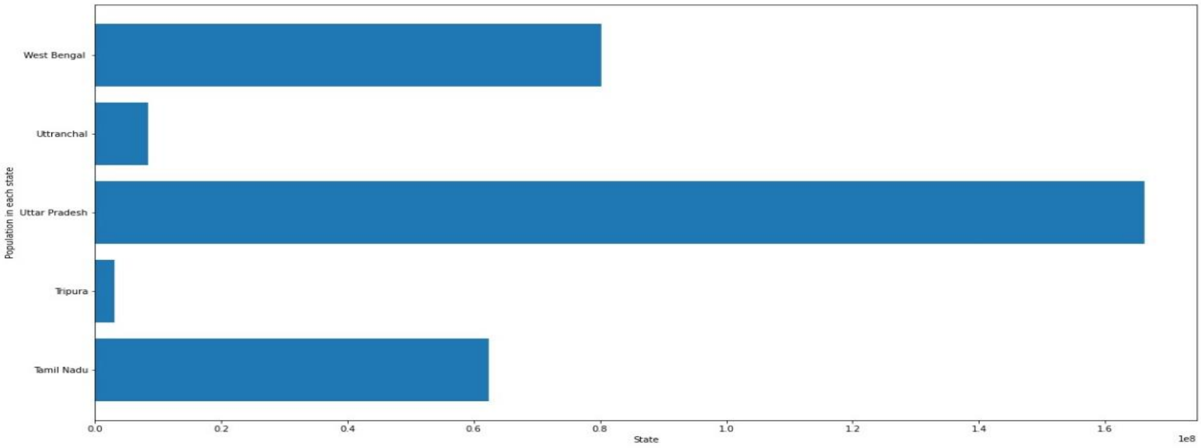

Figure-4

The above four images depict the visualized data of the population of India in the year 2001. In the images 1 and 2 the scale of the graph is 10^7 while, in the images 3 and 4 the scale of the graph is 10^8. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7 or 1e8. From the above images we can clearly understand that the most populated state of our country in the year 2001 was Utter Pradesh and the Least population state was Lakshadweep during the year 2001. In the figure 3 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 3. The above image is the visual representation of population of all the states and union territories of India during 2001.

2) **Literates**

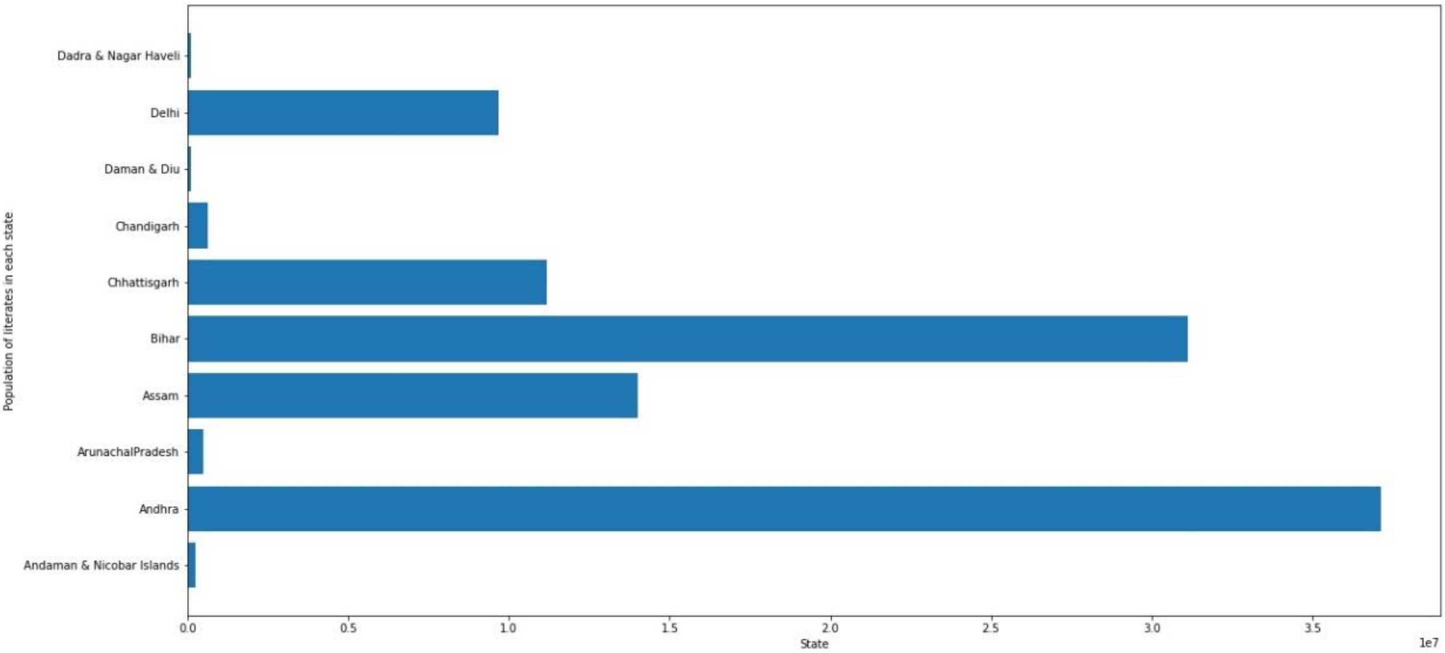

Figure-5

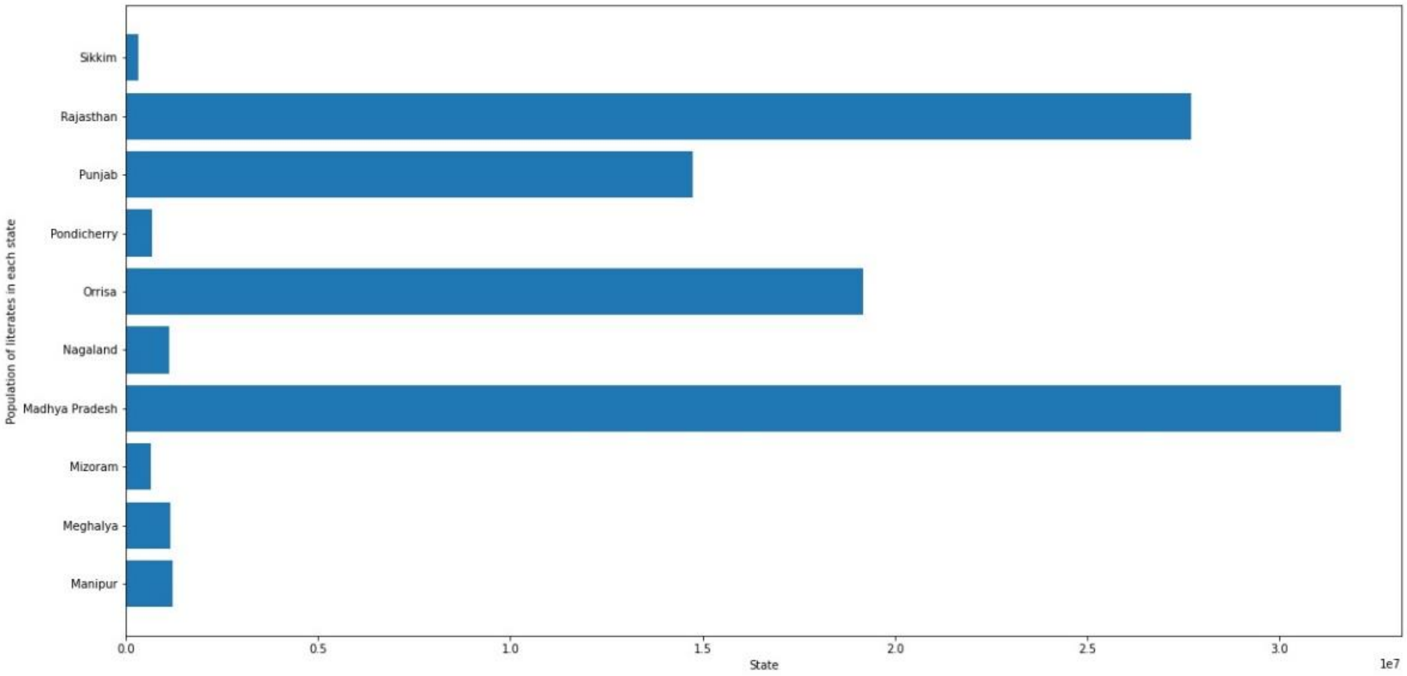

Figure-6

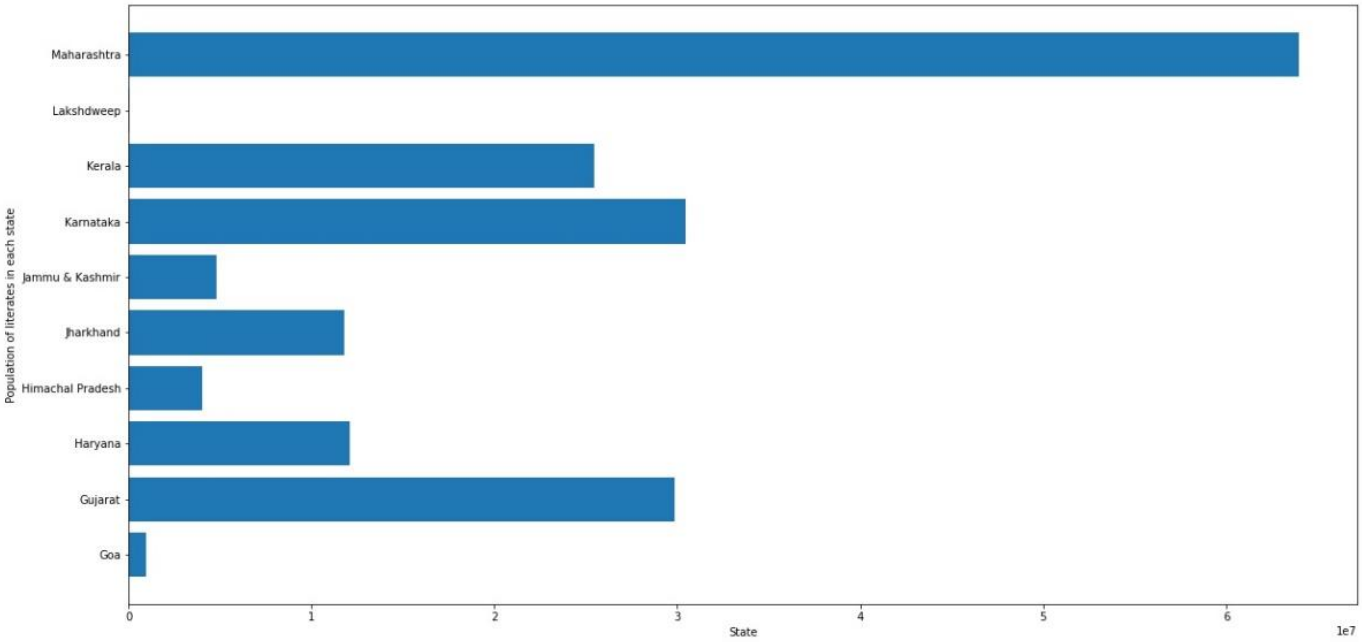

Figure-7

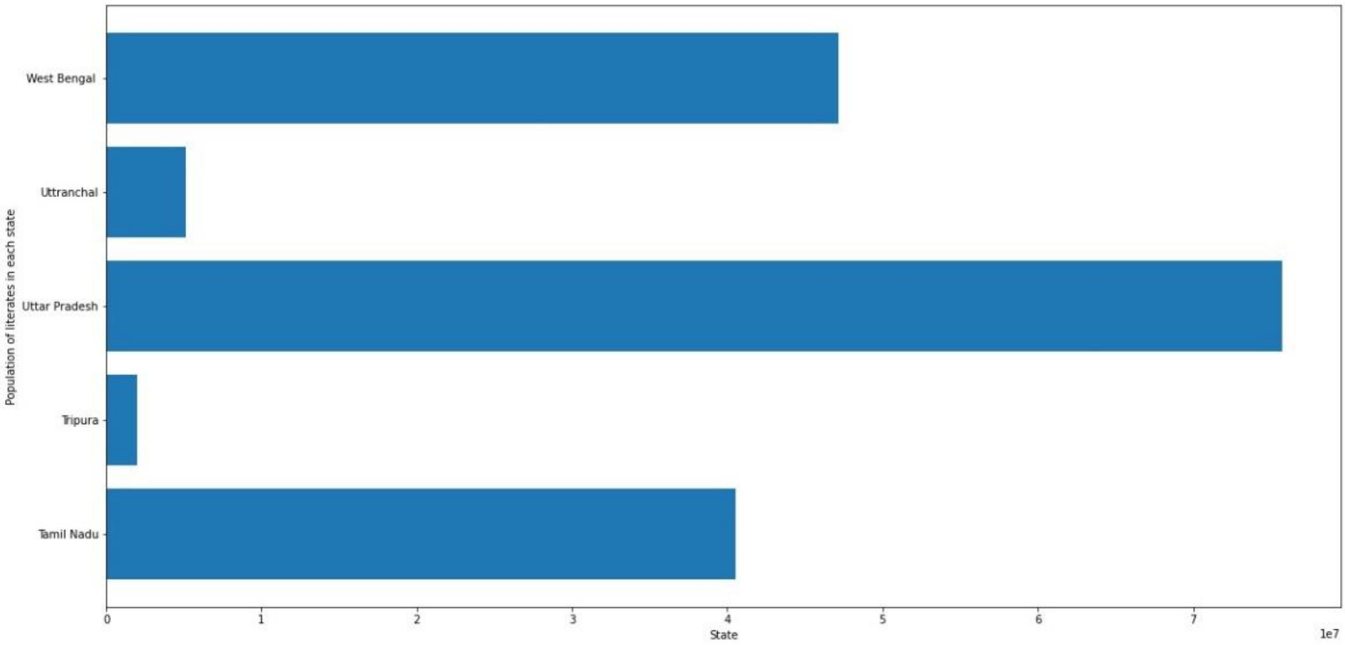

Figure-8

The above four images depict the visualized data of the Literate population of India in the year 2001. In the images all the images the scale of the graph is 10^7. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7. From the above images we can clearly understand that the most literate state of our country in the year 2001 was Utter Pradesh and the Least literate state was Lakshadweep during the year 2001. In the figure 7 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 7. The above image is the visual representation of literate population of all the states and union territories of India during 2001.

3) **Working Population**

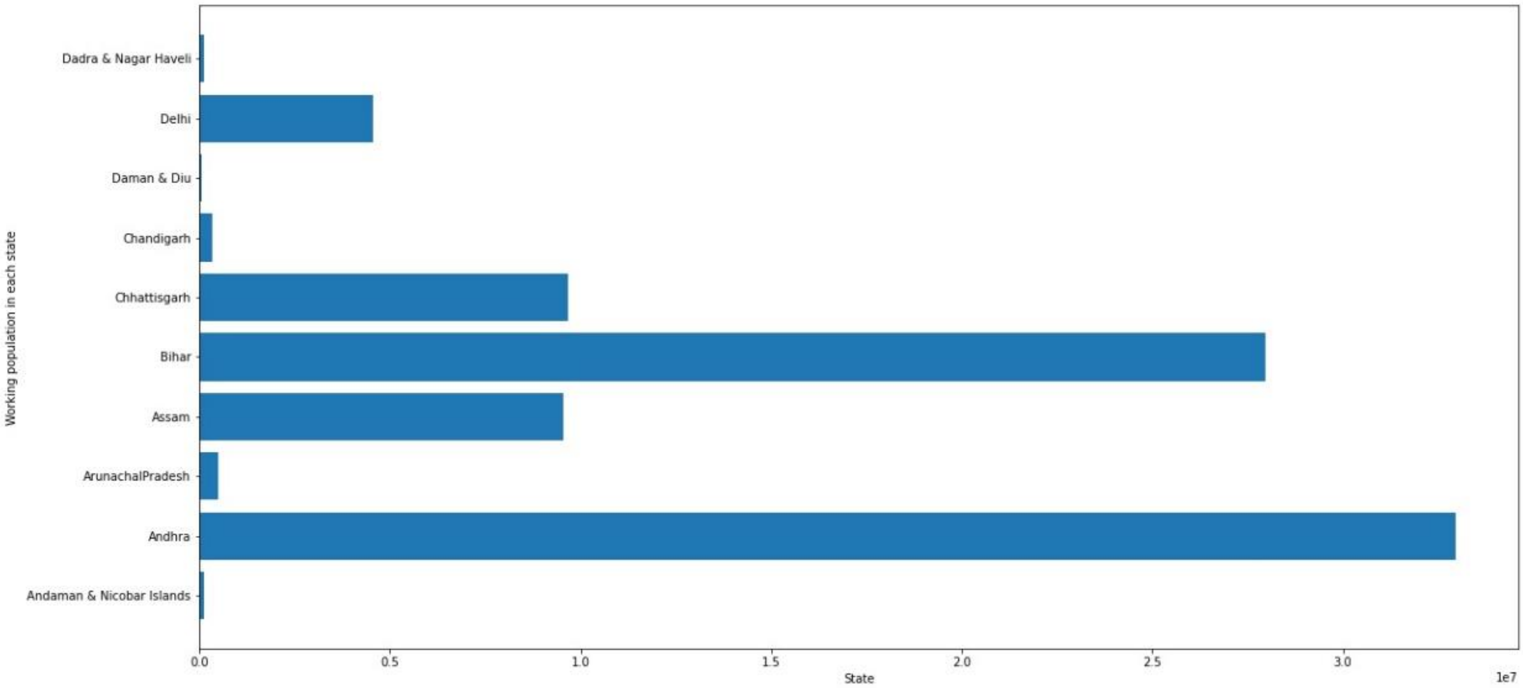

Figure-9

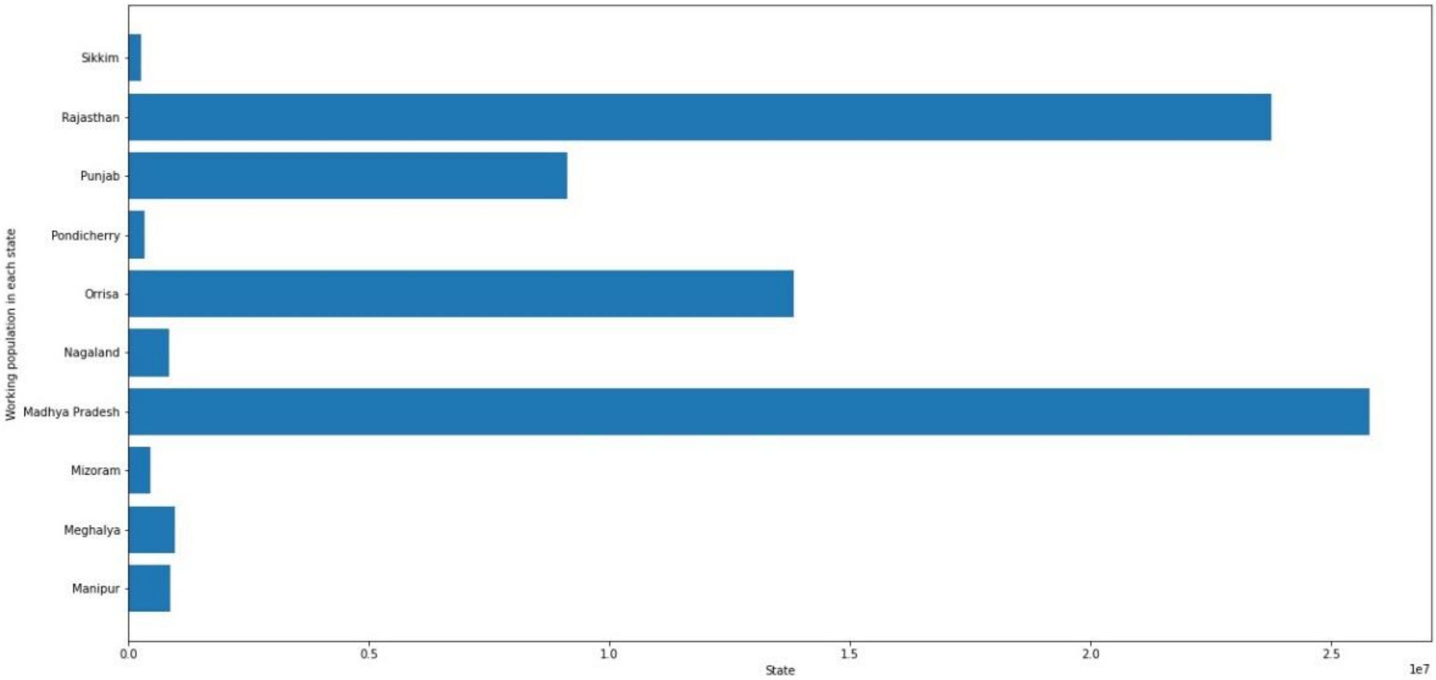

Figure-10

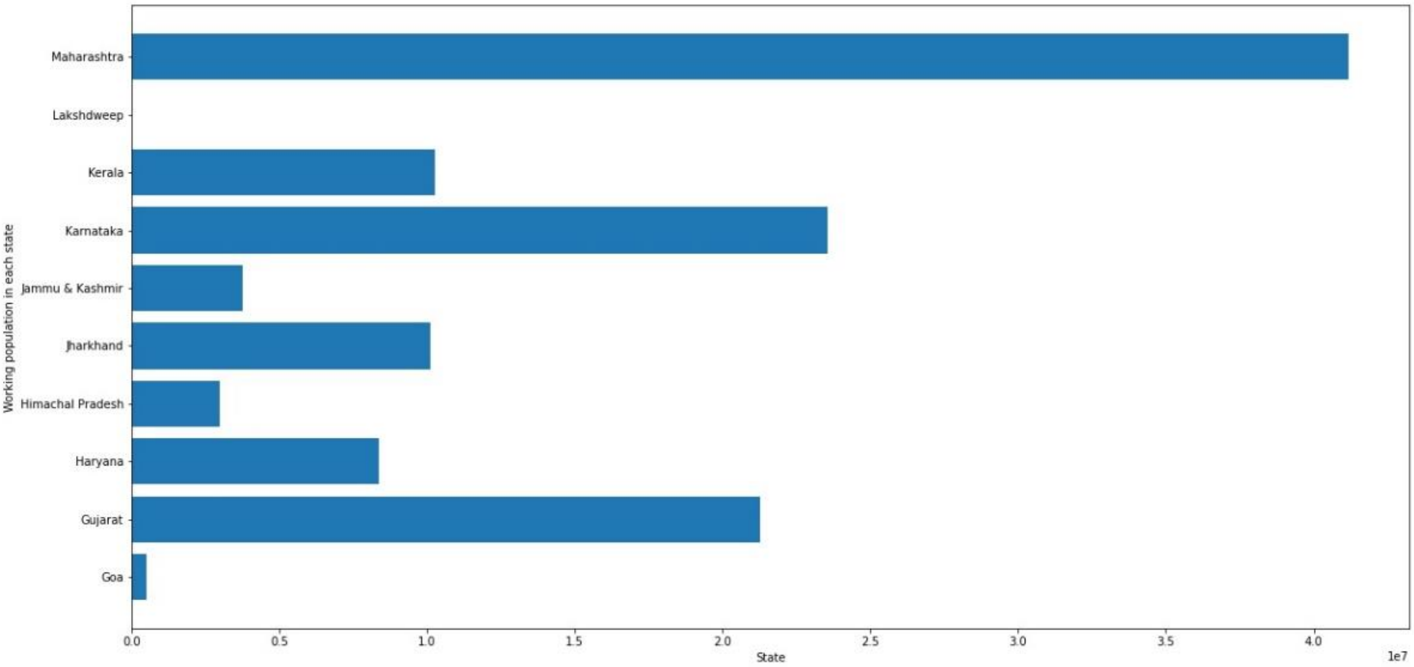

Figure-11

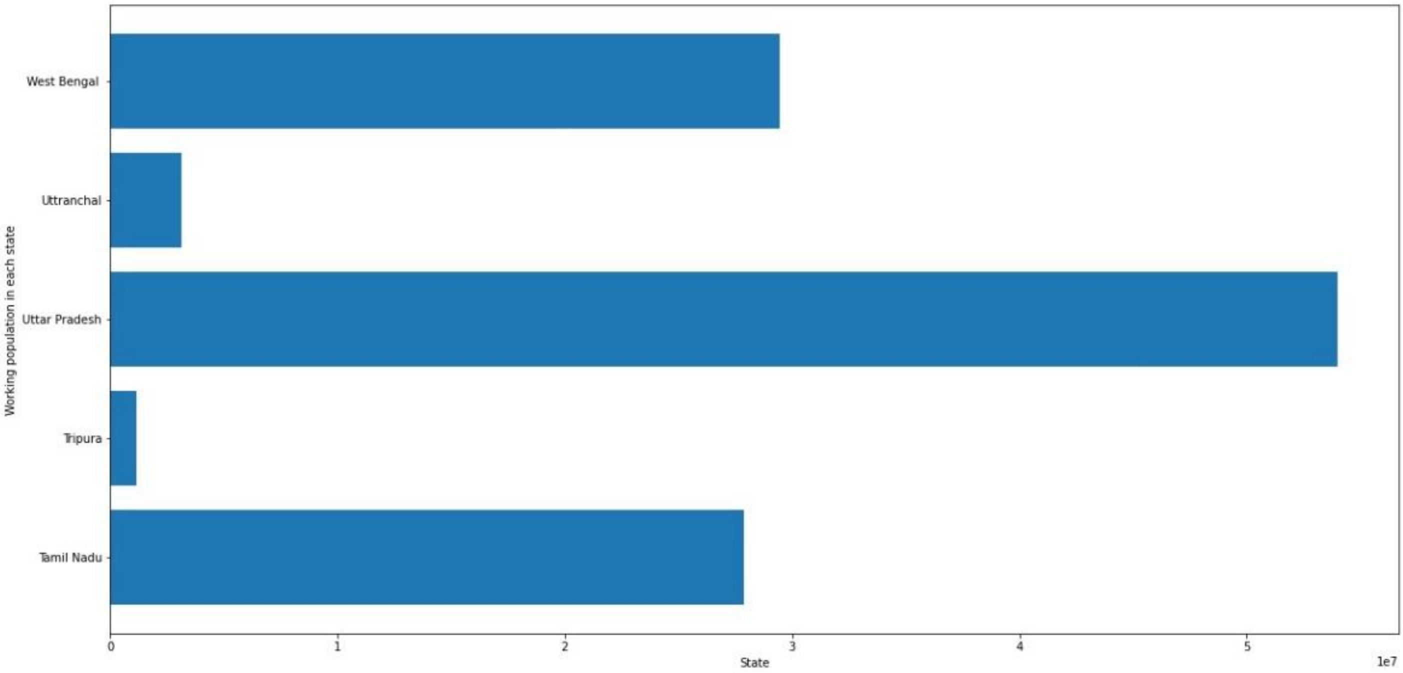

Figure-12

The above four images depict the visualized data of the Working population of India in the year 2001. In the images all the images the scale of the graph is 10^7. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7. From the above images we can clearly understand that the State with most working population of our country in the year 2001 was Utter Pradesh and the Least state was Lakshadweep during the year 2001. In the figure 11 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 11. The above image is the visual representation of Working population of all the states and union territories of India during 2001.

1) 2011

1) Population

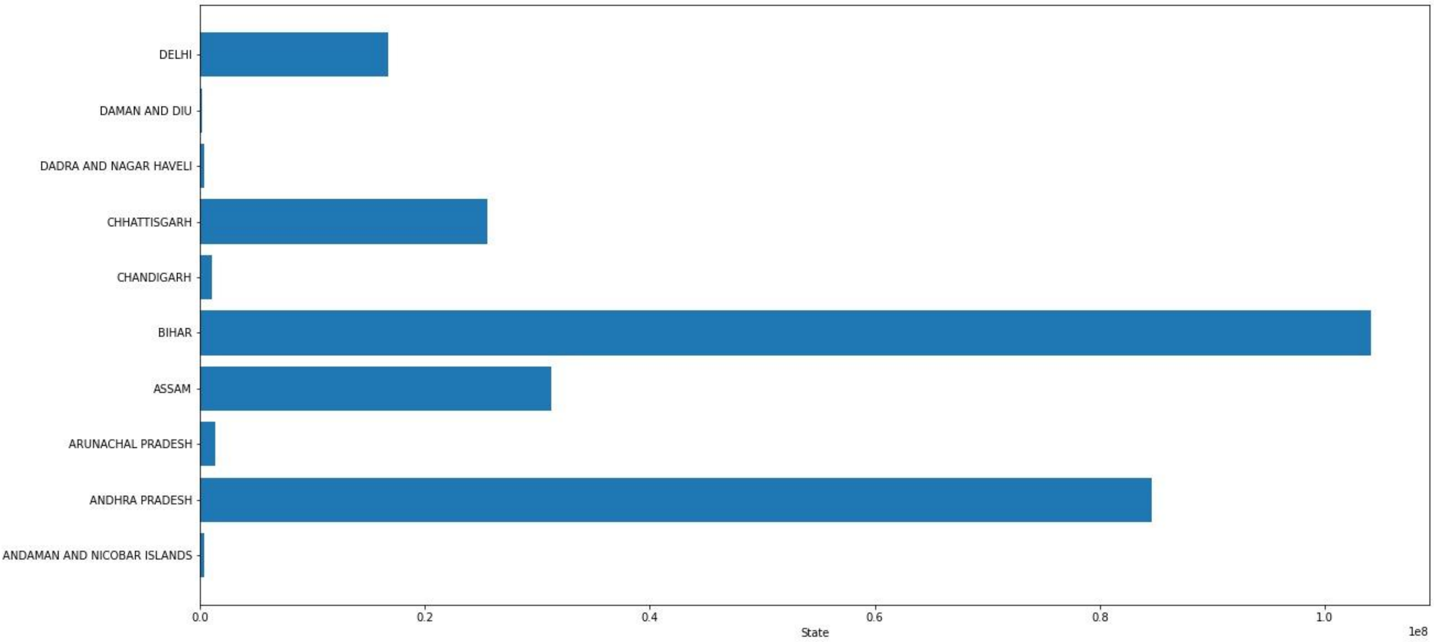

Figure-13

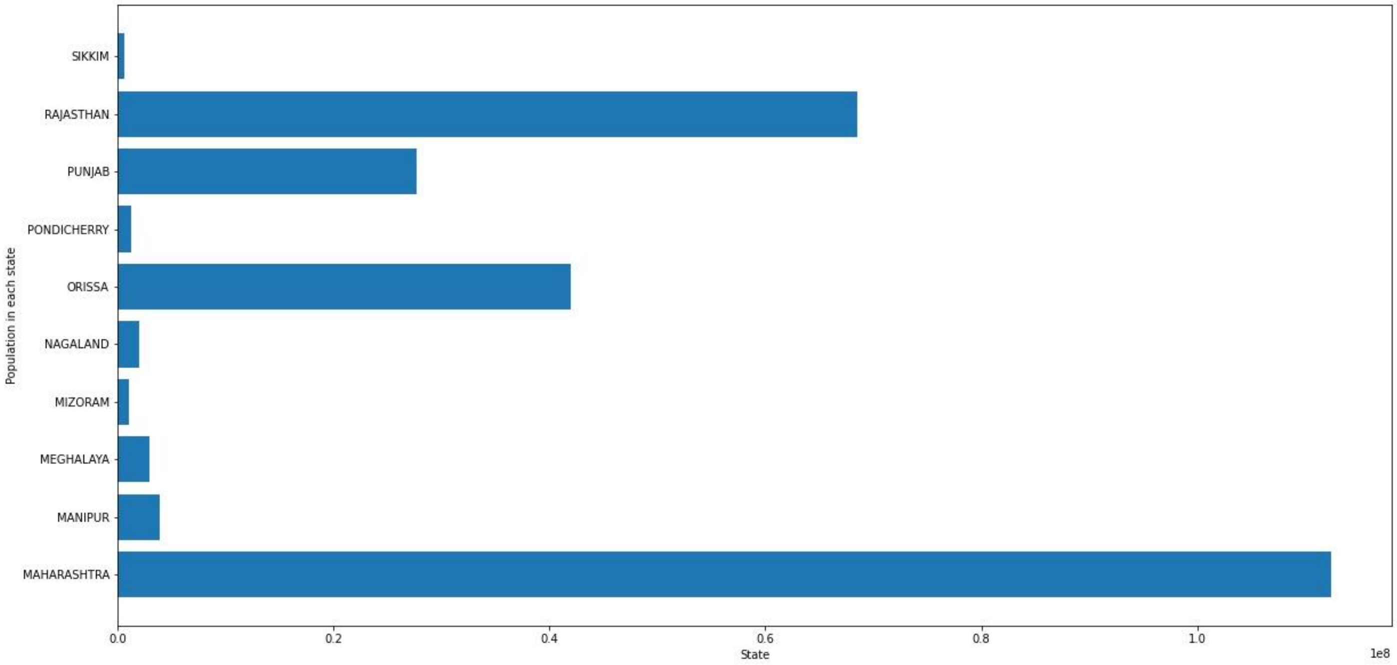

Figure-15

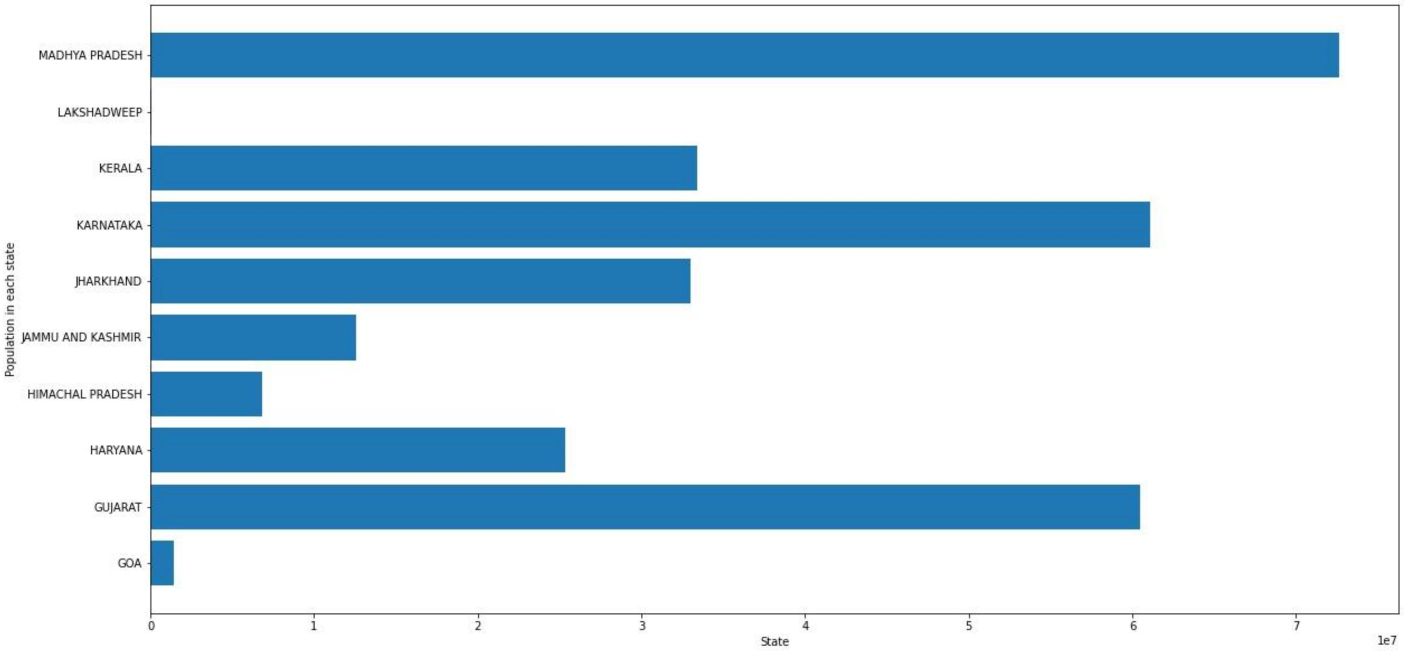

Figure-14

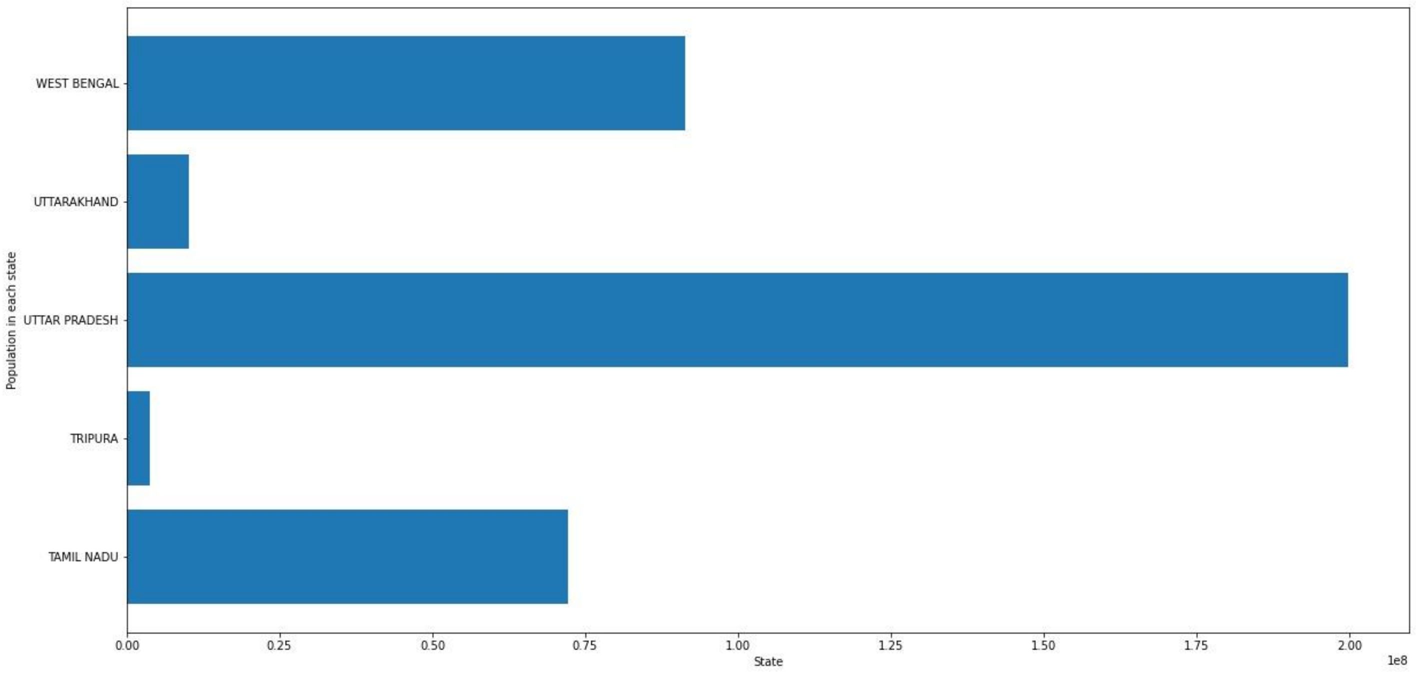

Figure-16

The above four images depict the visualized data of the population of India in the year 2011. In the images 1,2,4 the scale of the graph is 10^8 while, in the images 3 the scale of the graph is 10^7. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7 or 1e8. From the above images we can clearly understand that the most populated state of our country in the year 2011 was Utter Pradesh and the Least population state was Lakshadweep during the year 2011. In the figure 14 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 14. The above image is the visual representation of population of all the states and union territories of India during 2011

2) Literates

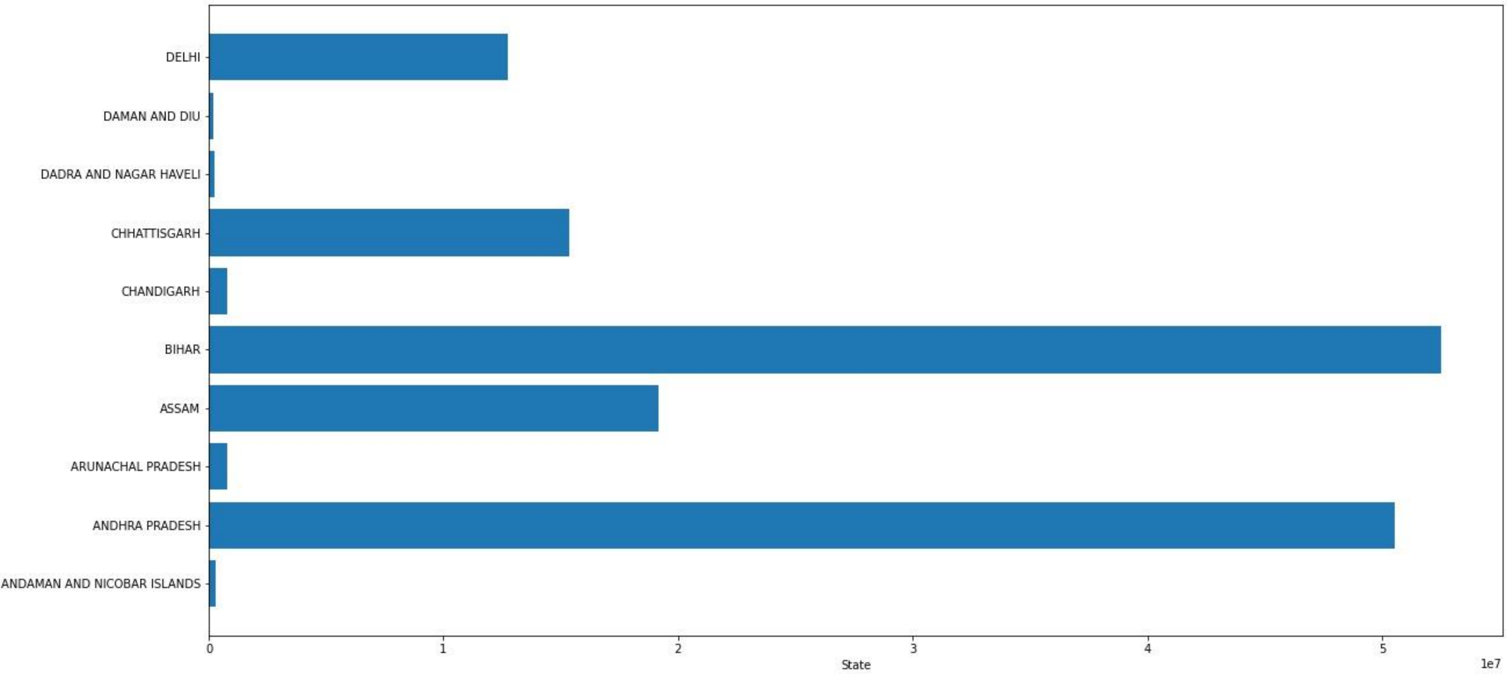

Figure-17

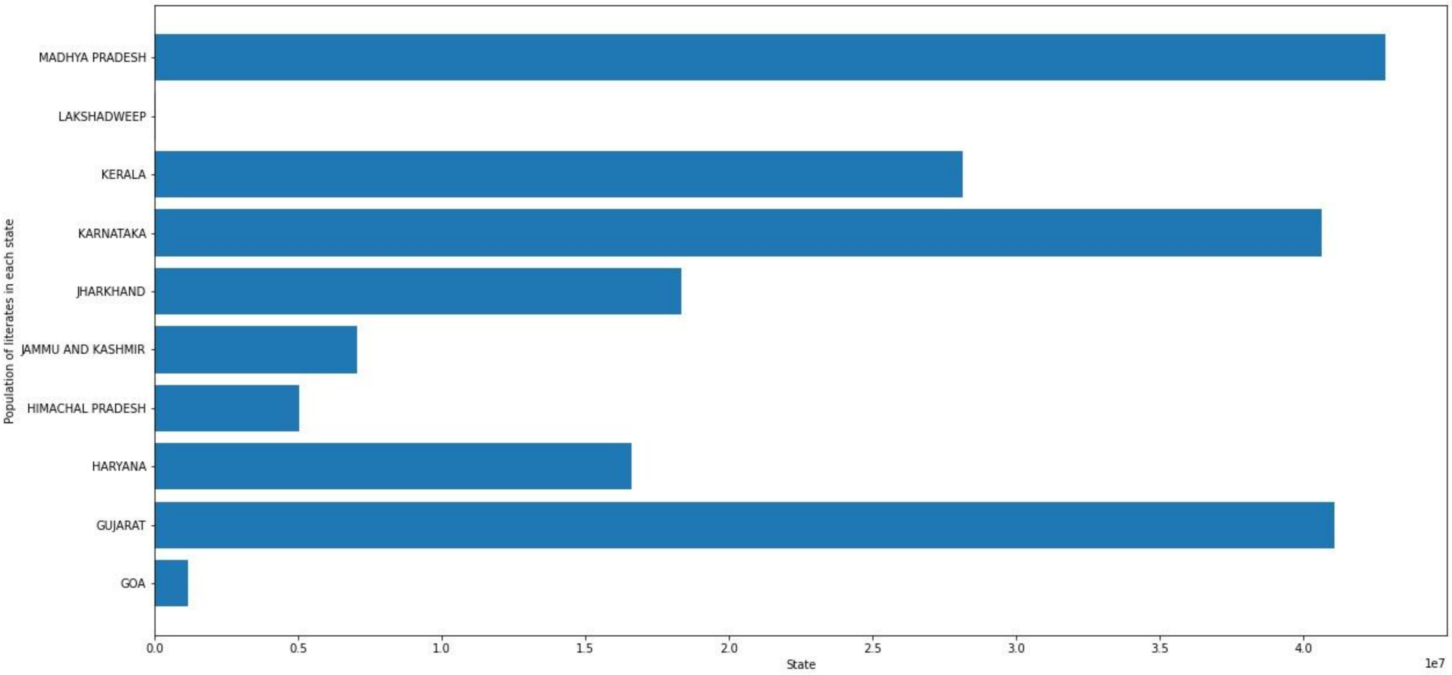

Figure-18

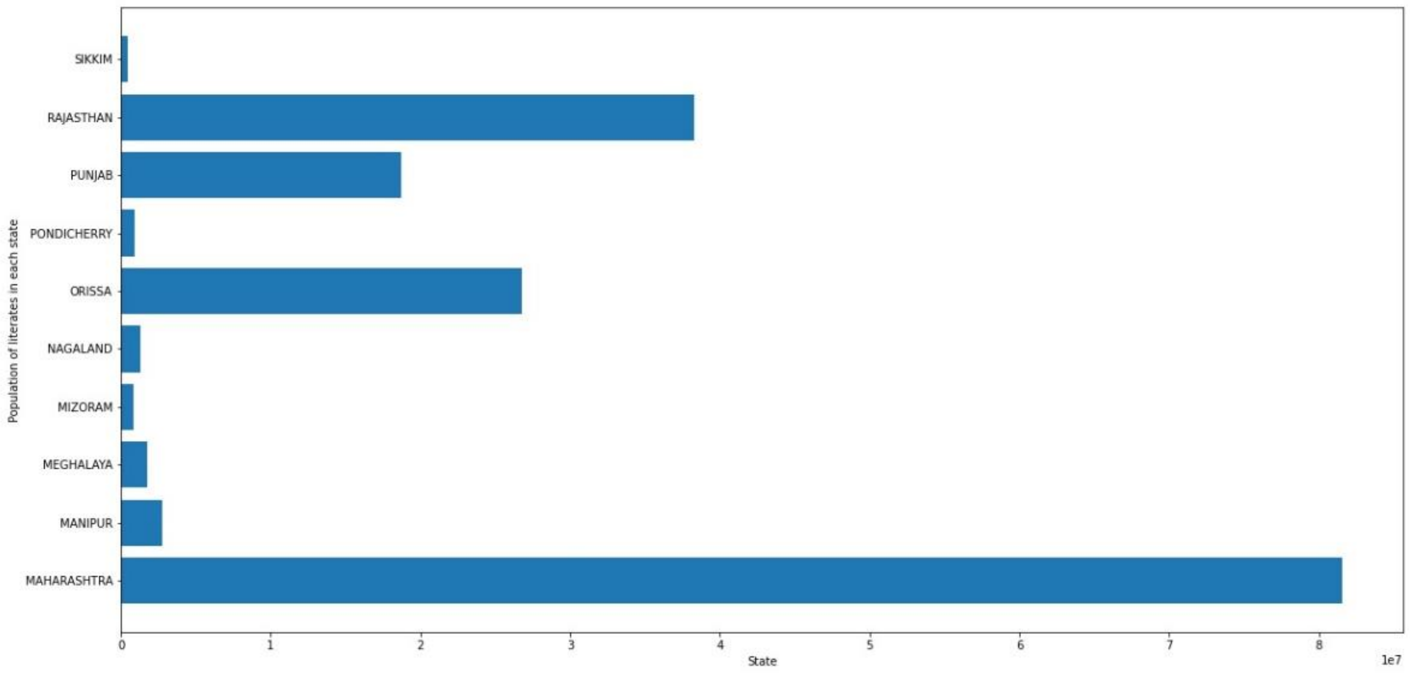

Figure-19

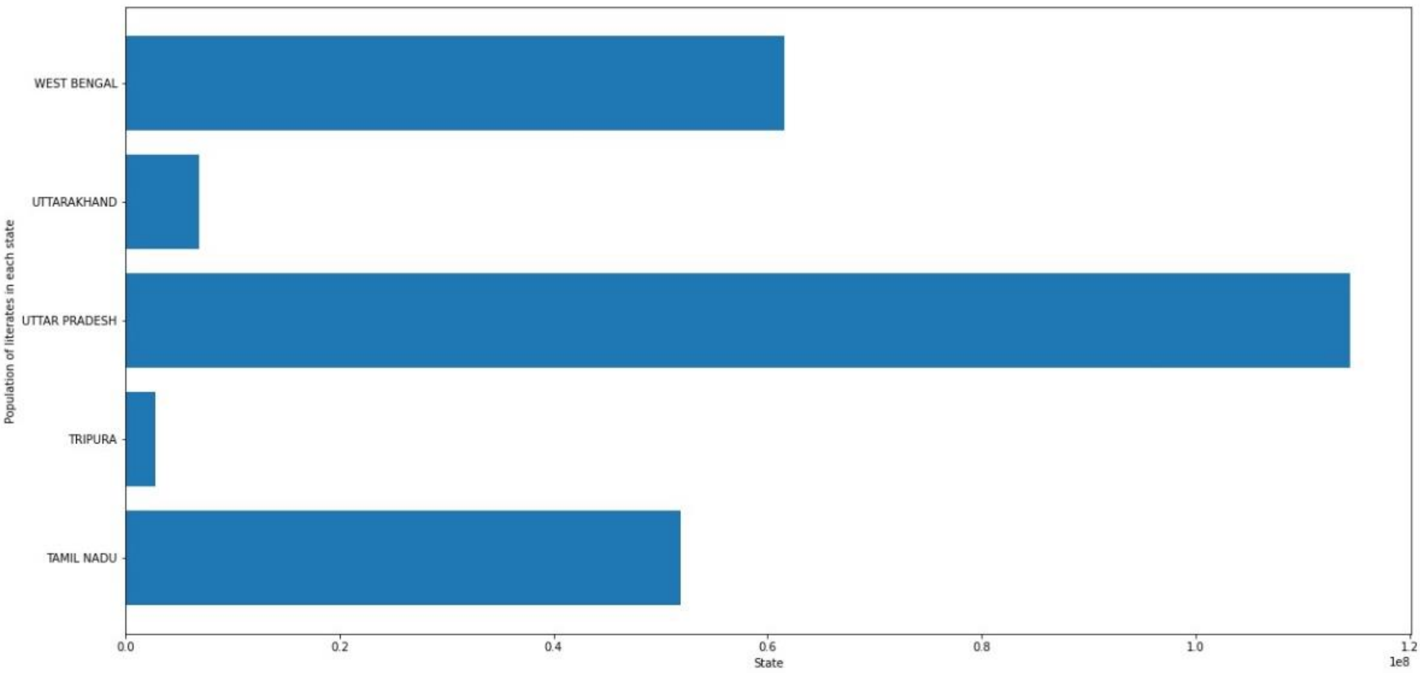

Figure-20

The above four images depict the visualized data of the Literate population of India in the year 2011. In the images except figure 20 the scale of the graph is 10^7. In figure 20 the scale is 10^8. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7 or 1e8. From the above images we can clearly understand that the most literate state of our country in the year 2011 was Utter Pradesh and the Least literate state was Lakshadweep during the year 2011. In the figure 18 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 18. The above image is the visual representation of literate population of all the states and union territories of India during 2011.

3) <u>Working Population</u>

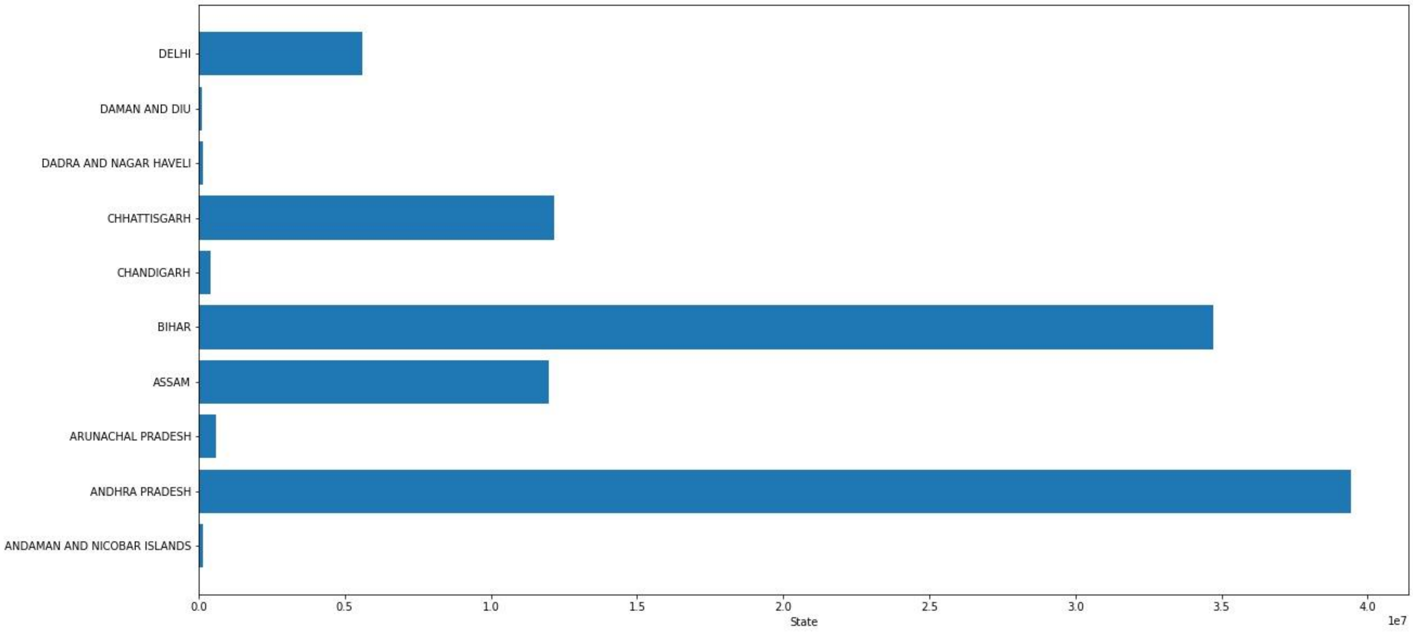

Figure-21

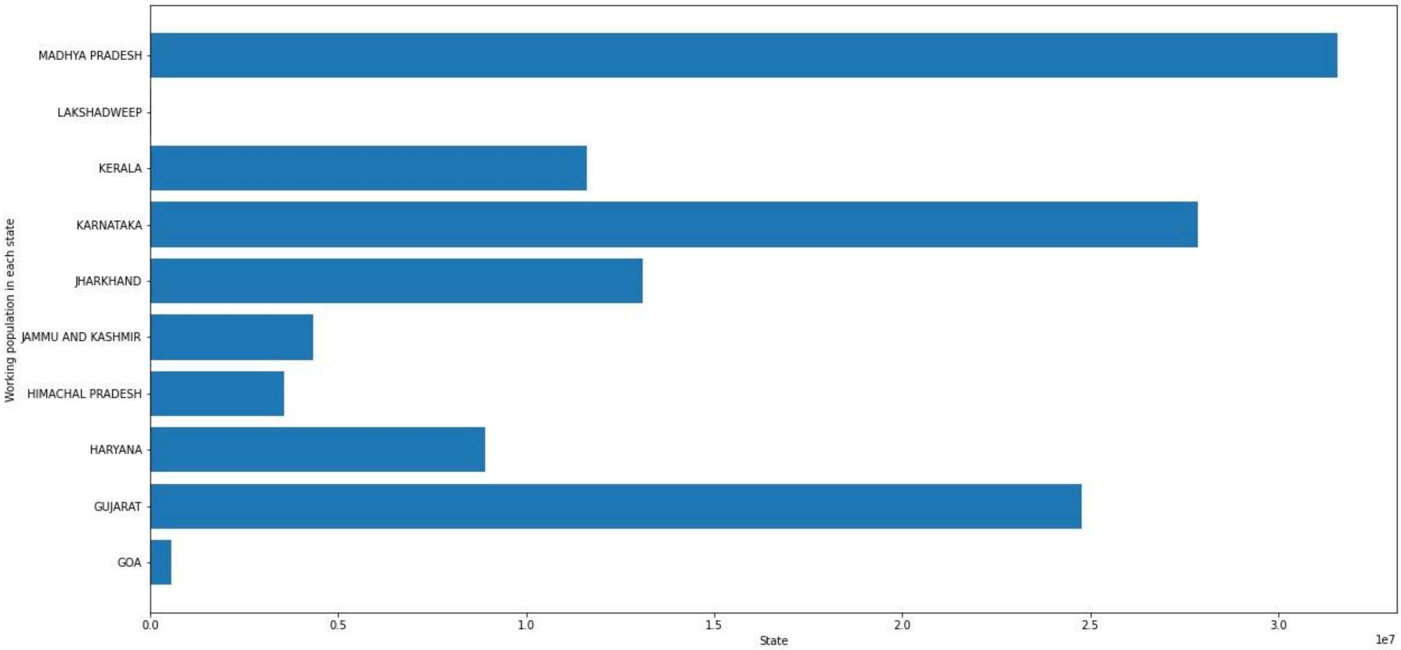

Figure-22

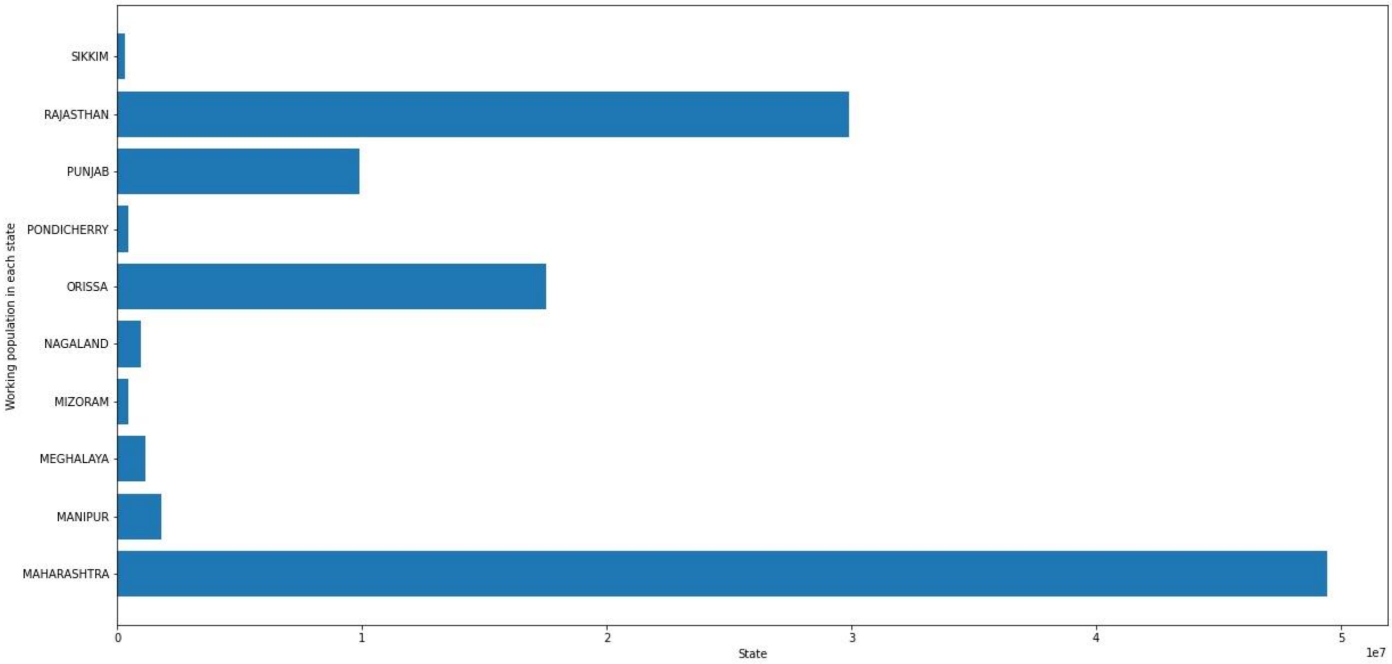

Figure-23

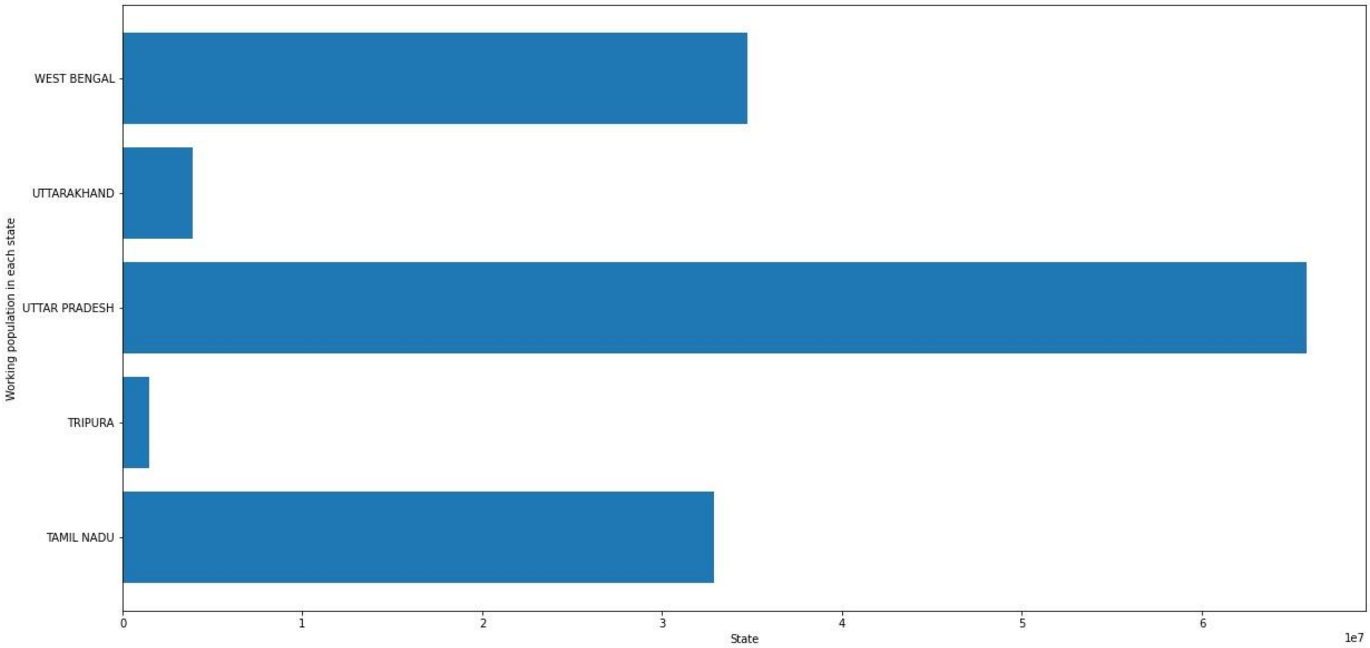

Figure-24

The above four images depict the visualized data of the Working population of India in the year 2011. In the images all the images the scale of the graph is 10^7. This is nothing but one unit of the graph is equal to 1 multiplied by the appropriate scale. The same this is shown in the images as 1e7. From the above images we can clearly understand that the State with most working population of our country in the year 2011 was Utter Pradesh and the Least state was Lakshadweep during the year 2011. In the figure 23 it is show as if there is no population in Lakshadweep islands this is because while comparing the population of other states to Lakshadweep has very low population. So, the scale is very low for the population of Lakshadweep this is why it is seen as if there is no population in the figure 23. The above image is the visual representation of Working population of all the states and union territories of India during 2011.

V) Prediction

T1. Efficiency of Linear Regression Model

Year	Actual Population	Predicted Population	Error	Error-Percentage (%)
1950	376325200	383403034.5	-7077834.51	0.018460
1951	382376948	387287219	-4910271.03	0.012678
1952	388799073	391741368.8	-2942295.84	0.007510
1953	395544369	396754401.8	-1210032.84	0.003049
1954	402578596	402315218	263377.96	0.000654
1955	409880595	408412699.5	1467895.50	0.003594
…………………………….	……………………………….	………………………………….	………………………………	……………………………..
2001	1075000085	1072939130	2060955.11	0.001920
2002	1093317189	1091388665	1928523.79	0.001767
2003	1111523144	1109820129	1703014.81	0.001534
2004	1129623456	1128221526	1401930.26	0.001242
2005	1147609927	1146580841	1029086.17	0.000897
2006	1165486291	1164886043	600248.48	0.000515
2007	1183209472	1183125081	84391.05	0.000071
2008	1200669765	1201285888	-616123.31	0.000512
2009	1217726215	1219356379	-1630163.92	0.001336
2010	1234281170	1237324449	-3043279.14	0.002459
2011	1250287943	1255177977	-4890034.43	0.003895
			Average	**0.018460**

T2. Efficiency of Polynomial Regression Model

Year	Actual Population	Predicted Population	Error	Error-Percentage (%)
1950	376325200	383403034.51	-7077834.51	0.018460
1951	382376948	387287219.04	-4910271.03	0.012676
1952	388799073	391741368.84	-2942295.84	0.007510
1953	395544369	396754401.84	-1210032.84	0.003049
1954	402578596	402315218.04	263377.96	0.000654
1955	409880595	408412699.50	1467895.5	0.003594
…………………………	……………………..	…………………………	………………………	……………………………
2001	1075000085	1072939129.88	2060955.11	0.001920
2002	1093317189	1091388665.21	1928523.79	0.001767
2003	1111523144	1109820129.19	1703014.81	0.001534
2004	1129623456	1128221525.74	1401930.26	0.001242
2005	1147609927	1146580840.83	1029086.17	0.000897
2006	1165486291	1164886042.52	600248.48	0.000515
2007	1183209472	1183125080.94	84391.058	0.000071
2008	1200669765	1201285888.31	-616123.31	0.000510
2009	1217726215	1219356378.93	-1630163.92	0.001336
2010	1234281170	1237324449.15	-3043279.14	0.002459
2011	1250287943	1255177977.43	-4890034.43	0.003895
			Average	**0.003621**

T3. Predicted Population for the year’s 2011 to 2030

Year	Prediction
2011	1,25,51,77,977.43115
2012	1,27,29,04,824.29785
2013	1,29,04,92,832.34228
2014	1,30,79,29,826.26269
2015	1,32,52,03,612.81103
2016	1,34,23,01,980.81835
2017	1,35,92,12,701.21923
2018	1,37,59,23,526.99072
2019	1,39,24,22,193.21142
2020	1,40,86,96,417.03320
2021	1,42,47,33,897.68359
2022	1,44,05,22,316.47119
2023	1,45,60,49,336.77880
2024	1,47,13,02,604.06835
2025	1,48,62,69,745.87841
2026	1,50,09,38,371.83398

2027	1,51,52,96,073.63476
2028	1,52,93,30,425.04638
2029	1,54,30,28,981.92480
2030	1,55,63,79,282.20703

Table 1 is Efficiency table of Liner Regression Model. In this model the average error percentage is 0.018460. The contents of the table show us how accurate the linear regression model is and how much error percentage it has. From the finding’s we can say that the linear regression model is not as accurate as it does not take into account many factors that helps in population prediction.

Table 2 is Efficiency table of Polynomial Regression Model. In this model the average error percentage is 0.003621. The contents of the table show us how accurate the polynomial regression model is and how much error percentage it has.

Table 3 is the Prediction of Indian population for the years 2011 to 2030. The population of India during 2030 would be 1,55,63,79,282 i.e. One Hundred Crore Fifty-Five Crore Sixty-Three Lakh Seventy-Nine Thousand Two hundred eighty-Two.

VI) Conclusion

By the year 2030 Indian population would be 1,55,63,79,282 i.e. One Hundred Crore Fifty-Five Crore Sixty-Three Lakh Seventy-Nine Thousand Two hundred eighty-Two. Our Country should be ready for this population and we should be able to produce much needed jobs and increase medical facilities and also school facilities. This data would be really helpful our government to be ready for the Future.

VII) Reference

- https://www.kaggle.com/bazuka/census2001
- https://www.kaggle.com/danofer/india-census
- https://www.kaggle.com/sansuthi/indian-population
- https://learn.datacamp.com/courses/machine-learning-for-time-series-data-in-python
- https://learn.datacamp.com/courses/introduction-to-data-science-in-python
- https://learn.datacamp.com/courses/introduction-to-data-visualization-with-matplotlib
- **N. Keilman, D. Pham, and A. Hetland. Why population forecasts should be probabilistic – illustrated by the case of Norway. Demographic Research, 6, (2002) 409–453.**