**Python Project**

**Project: Design a Web Scraper to Scrape Data from a Website. Analyse the Data and Make a Report on the Analysis**

**Submitted By: Mohit Vaidya**

**Roll No.: 045032**

**Submitted to: Prof. Amarnath Mitra**

**Subject: DEVP - Data Exploration & Visualization using Python-I**

**GitHub Link:** <https://github.com/vaidyamohit/045032>

## TOPIC: List of Most Populated Indian Cities

**Project Objective:**

* The Primary objective of this project was to design a web scraper to extract data from a website containing information about the most populated cities in India.
* We aimed to analyze this data and create comprehensive report on the population trends of these cities over 3 year (2022, 2011 and 2001)
* We will divide these cities with their respective states or union territories and compare the population with each other.
* Check the correlation between populations of cities, statement.
* Make pictorial charts, graphs to analyze the population on various factors.
* Ultimately, the project aims to provide valuable insights through a report based on the analysed data of populations.

**General Descriptions of Data:**

**Data Source:**

Website URL: <https://www.indiaconvey.com/population/cities/>

**Data Type:**

Tabular form data mainly containing city names, population data for the years 2022, 2011 and 2001, and the respective states or union territories.

**Data Structure:**

The dataset used in this analysis consists of several columns, each serving a specific purpose in understanding the population trends of the most populated Indian cities. Here's an explanation of each column:

1. **Rank**: This column represents the ranking of each city based on its population. The city with the highest population is assigned Rank 1, the second-highest Rank 2, and so on. The ranking allows us to identify the most populous cities within India.

2. **City**: The "City" column contains the names of the cities under consideration. Each row corresponds to a specific city, and this column is essential for identifying and distinguishing between different urban centers.

3. **Population in Year (2022)**: This column displays the estimated population of each city as of the year 2022. The population figures are presented in numeric format and are represented in millions. This data allows us to assess the current population size of these cities, which is crucial for various analyses and decision-making processes.

4. **Population in Year (2011)**: The "Population in Year (2011)" column provides the population count for each city as of the year 2011. This data point is valuable for understanding population growth or decline over the past decade and assessing the rate of urbanization.

5. **Population in Year (2001)**: Similar to the 2011 population data, this column displays the population count for each city in the year 2001. It serves as a historical reference point and allows for the examination of long-term population trends.

6. **State or Union Territory**: The "State or Union Territory" column specifies the administrative region to which each city belongs. This information is crucial for understanding the geographical context and governance structure of each city. It helps identify the state or union territory responsible for urban planning and development.

**Data Collection Process:**

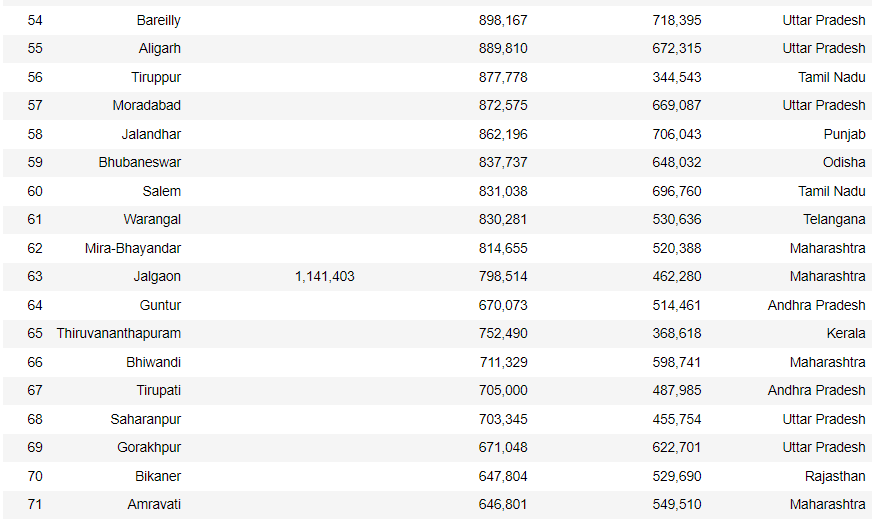
We developed a web scrapping script using Python and the beautiful Soup library to extract data from the website. The script navigated through the web pages, and collected all the tabular data and stored it in a structured way.

**Data Sample:**

Here are the snippets added of the data table which we have taken from the website.







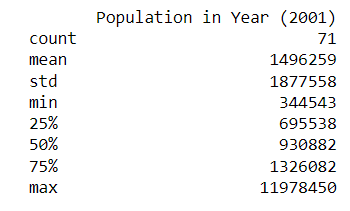
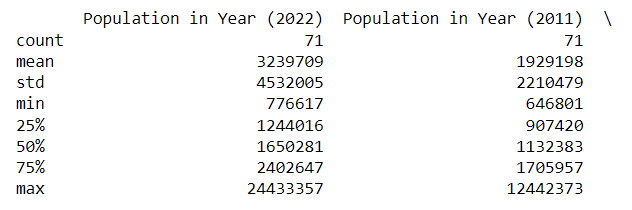
**Analysis:**

**Data Preprocessing**

* Before starting the analysis, we handled the missing values and added all the values manually by doing research and finding population of each city from different sources.
* Transformed population data into numeric format and also removed unnecessary characters from population numbers.

Below I have added how I added the data manually into the data table using python, by this we can start off with our analysis part and as we have complete data now the results will be better.

I have also found the descriptive statistics of each population where we even found mean, median, mode, standard deviation, max and all other factors.





**Top 15 Populated Cities in India (2022)**

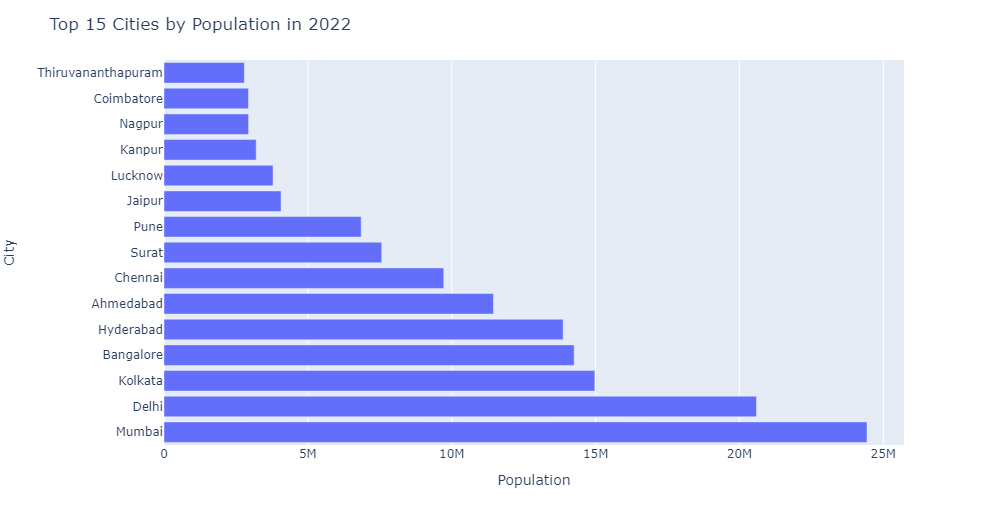
Here I have tried to find the top populated cities of India in 2022 and find out which are those top cities where most of the people live in India.

We found out that Mumbai is the most populated city to live in.

Followed by Delhi, Kolkata, Bangalore, Hyderabad.

All the cities in the list are considered as Tier-1 cities of India.

Below is the chart which represent all the 15 cities with their population.

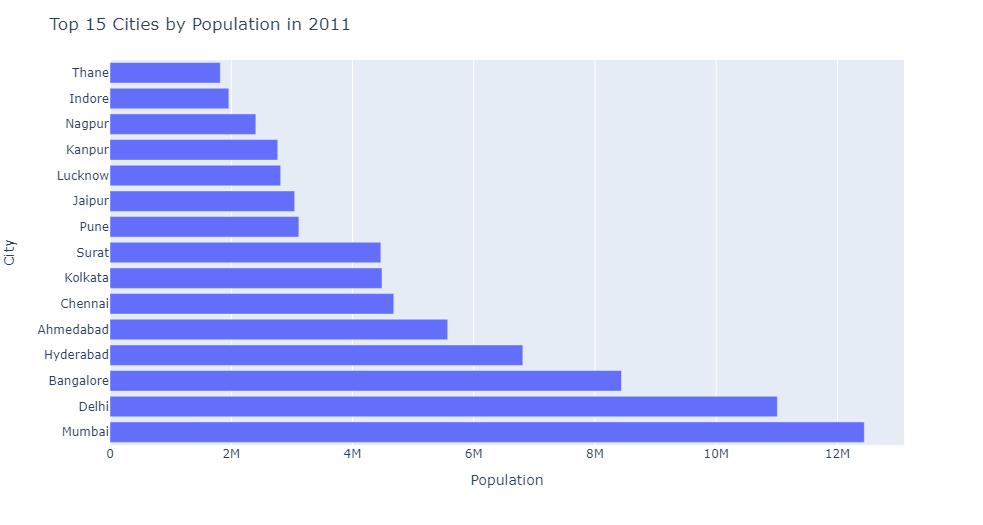


**Top 15 Populated Cities in India (2011)**

Here we compare the chart of population of year 2011 where we again find top 15 cities, we see in this also Mumbai is the leading one and followed by Delhi, Bangalore.

We see that Kolkata is not in the top 5 of the lists where as in 2022 it is in top 5.

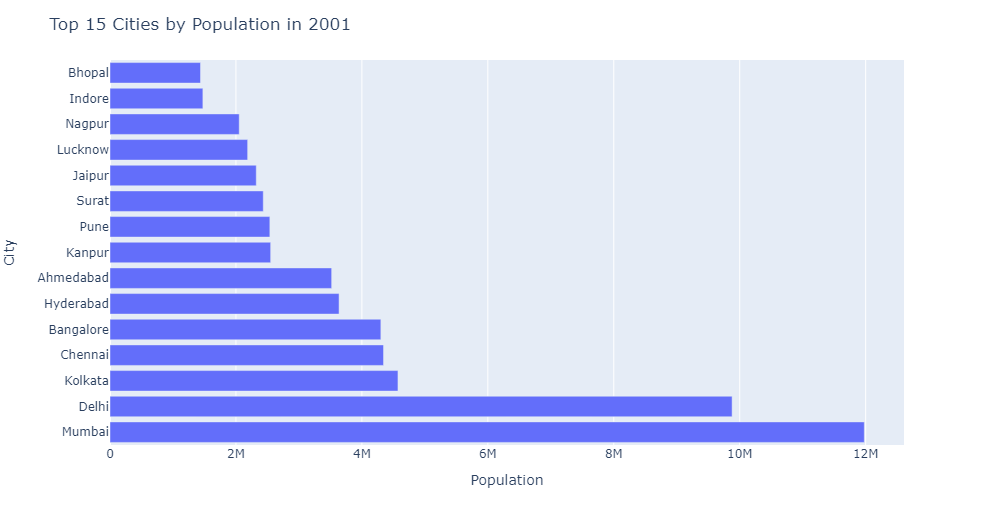
There are many cities which are in this list and were not in 2022 top 15 graph so we can say that urbanization has been quick in those places compared to in this place.



**Top 15 Populated Cities in India (2001)**

In this chart we compare the populations in the year 2001. In this list also, Mumbai is the leading city with highest population.

We have created bar graph to compare and check the cities of all the 3 years of populations we got. And we can see Mumbai, Delhi have been the top 2 in all the 3 cases.

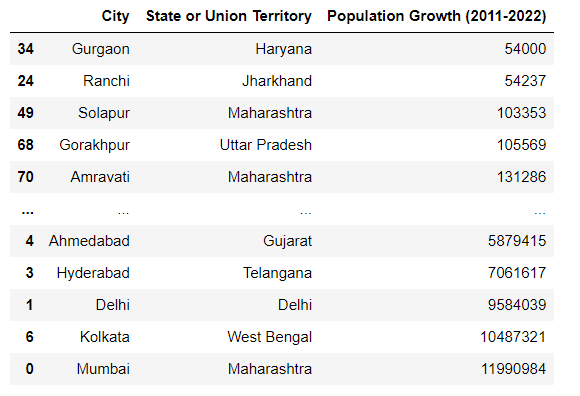
****

**Population Growth Between 2022 to 2011**

Here we will check the growth of population from 2011 to 2022 and see where has been the growth highest.

We see that highest growth is again in Mumbai and Delhi whereas the lowest has been in Gurgaon and Ranchi.

Below I have attached the table where we see all the cities names, which state they belong to and how much has been the growth in their population,

****

Now let us check the growth of population in a graphical view and see which are the top 15 cities with highest growth in population from 2011 to 2022.

We see that Patna is placed at 15 and Mumbai has topped the list, along with Kolkata and Delhi

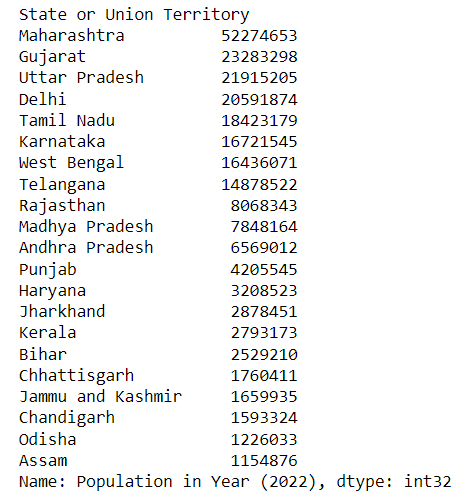


**State wise Population in 2022**

I have here added the data of each city and union territory with their respective states and we have got the total population of states given in the table.   
Below is the table which we got is arranged in alphabetical order and we can see the total population of each state in it.

****

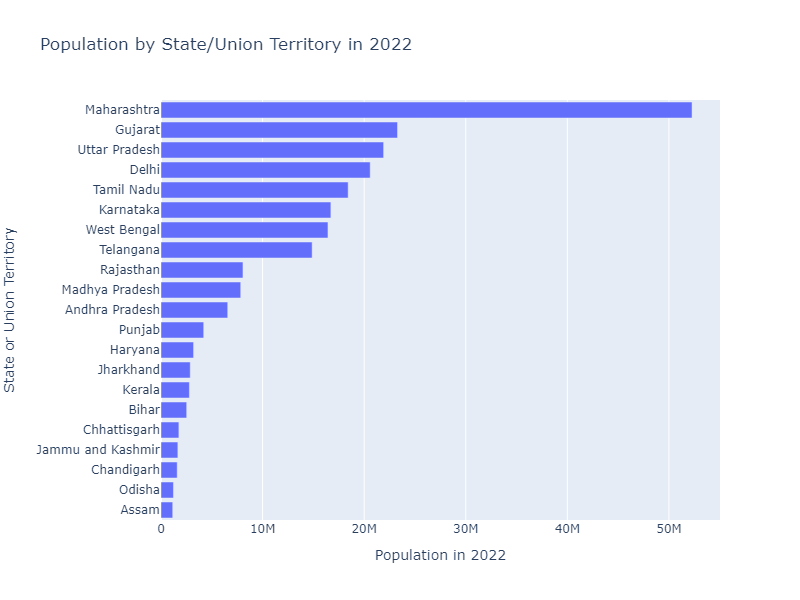
I have then arranged the data of states in decreasing order so we can see which is the state with highest population and which is with lowest.



We see that Maharashtra has the highest population in the given data, which is followed by Gujarat and then Uttar Pradesh.

Lowest population is at Assam, Odisha and Chandigarh.

Below is the graphical view of each state, so we can compare and see the difference between each state accordingly.



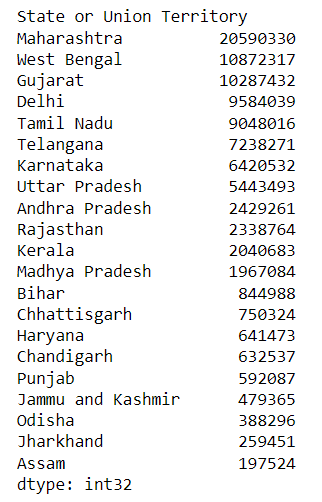
Similarly, we can see for each year of 2011 and 2001 and get insightful outcomes.

**Population Growth With respect to States/Union Territories**

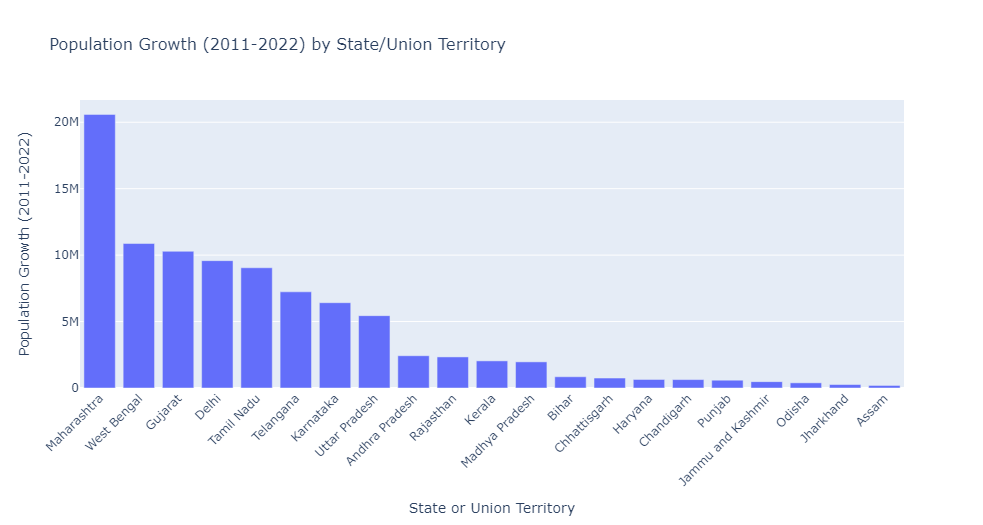
In this analysis we will see the growth of population from 2011 to 2022, We get to know which is the state in which the population grew the highest.

Below is the table and we see that In Maharashtra the population growth is been the highest in the years from 2011 to 2022.

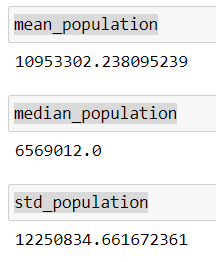
Followed by West Bengal and Gujarat, The growth of every state has been increasing at a very fast rate because of urbanization.



We see below the growth of each states in graphical format.



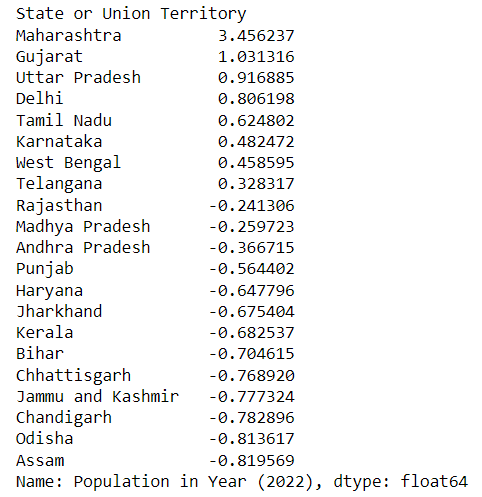
Below we get the mean, median and standard deviation of growth of population from 2011 to 2022 and see the output



**Outliers**

I here try to find the outliers, which means the growth at that states have been very high or very low and that value is away from the mean either greater or lower.

In our data, "Maharashtra" has the highest positive z-score, indicating that it has a significantly larger population compared to the mean. Conversely, "Chhattisgarh" and "Assam" have the lowest negative z-scores, indicating significantly smaller populations compared to the mean.



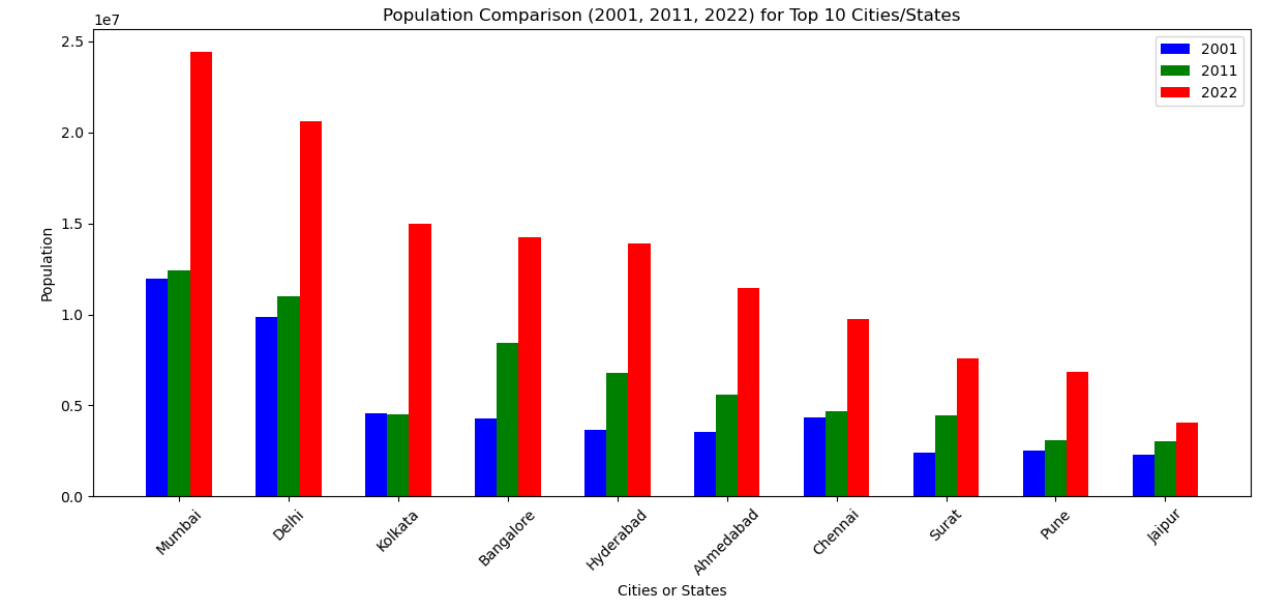
**Yearly Comparison of Population Growth in Top 10 Cities**

We will compare yearly population of cities with each other.

I have taken top 15 cities and we see that there is drastic change in population from 2011 to 2022 in each of the cities.

With this we get to compare growth of population in each cities year wise as well as compare other cities growth year wise.

Below is the chart and see the changes yearly.

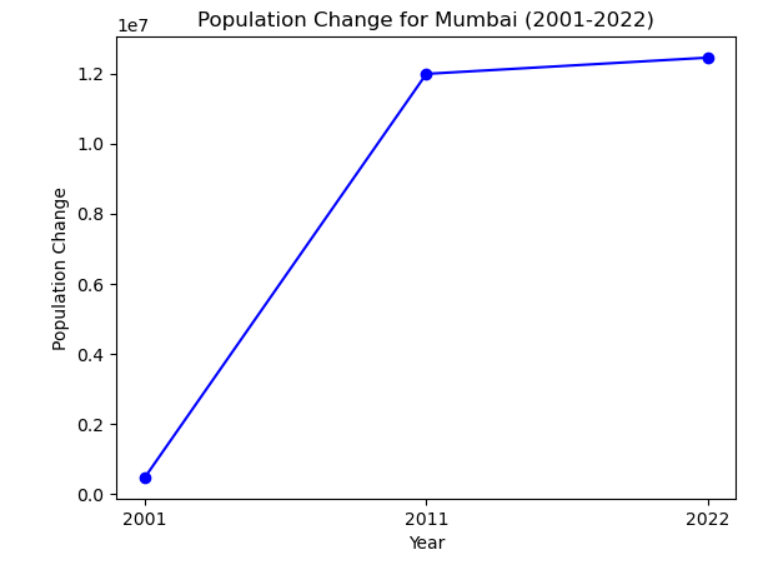


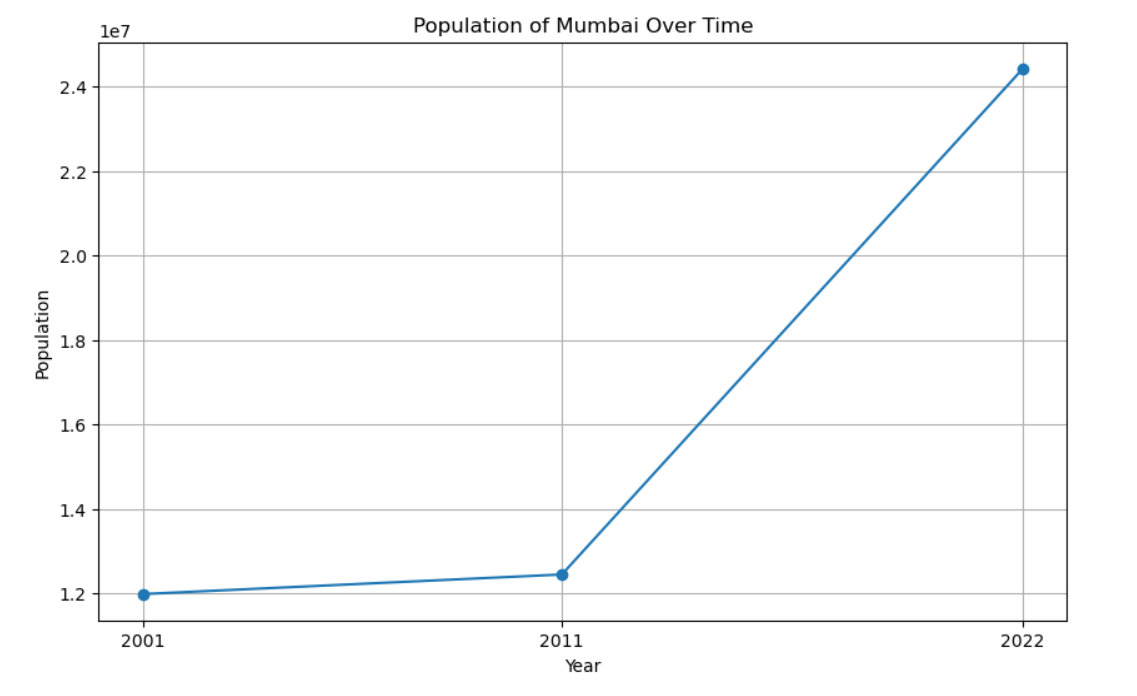
**Mumbai Population Growth**

As Mumbai has been the highest populated city each year let us see the trend of the population growth of Mumbai from 2001 to 2022.

We know Mumbai is the finance capital of India and Urbanization in Mumbai has come early and so we see the trend that the growth between 2001 to 2011 has been the highest and then the growth from 2011 to 2022 has been increasing as well.

This graph tells us the trend of population change in Mumbai and we can also check for other cities similarly.





**Ranking of Population with respect to Year 2022**

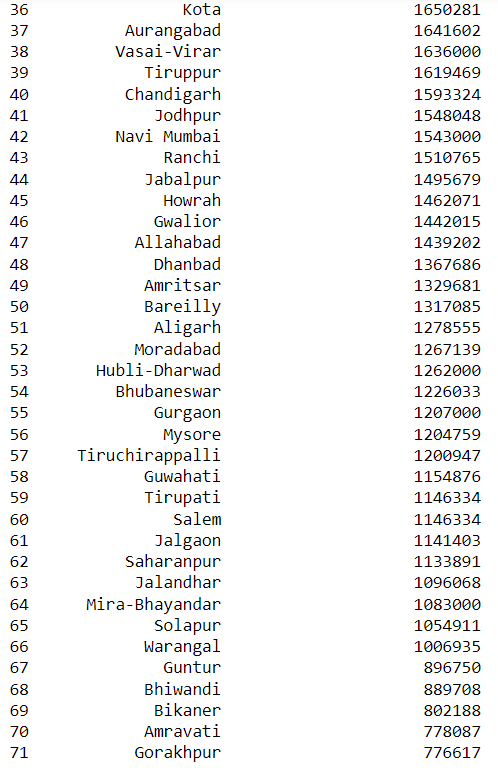
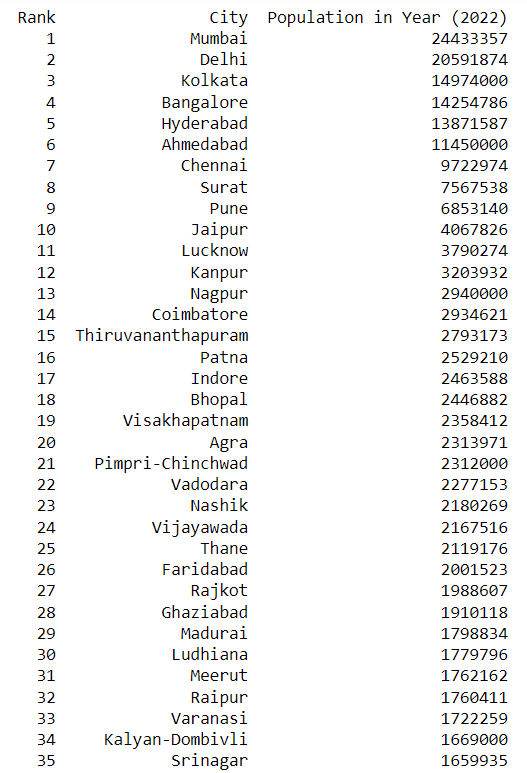
I have here ranked the population again with respect to their population in the year 2022.

As the ranking were from 2011 Sensex and also as data table, we web scrapped was empty, and so I manually entered population in 2022 in many cities as it was not mentioned.

So, we have ranked them and see that again Mumbai has been in the Rank 1. Followed by Delhi, Kolkata, Bangalore and Hyderabad in the top 5

We see that Gunter, Bhiwandi, Bikaner and Gorakhpur have been in the least 5 of the data table.

These population are not exact as Official Sensex has not yet happened. I have taken the population from various trustable websites and added according I have ranked them.

****

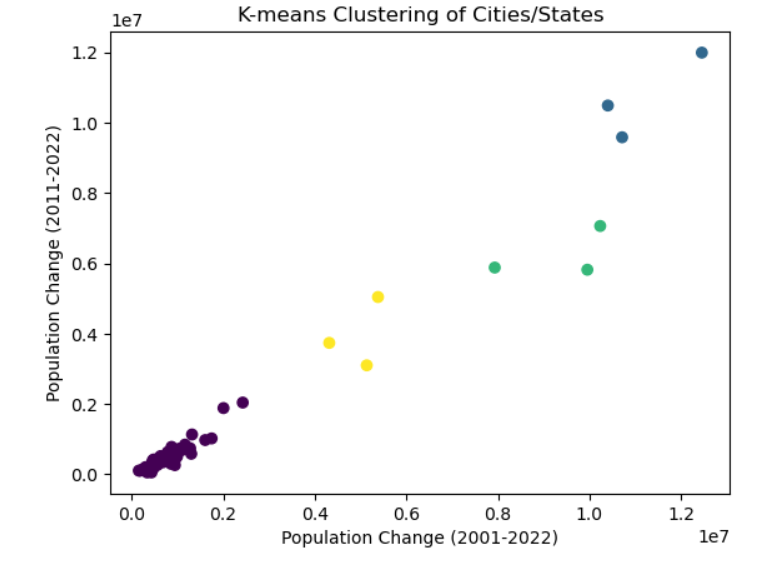
**Clustering of Data**

Here I try to group the similar data into one that is I make similar growth of cities into cluster 1 and so on

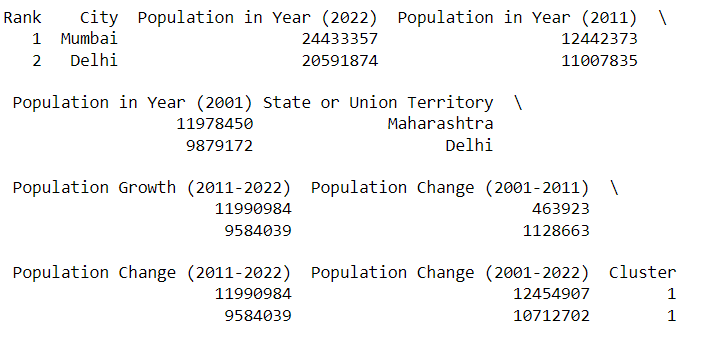
By seeing the graph, we will get to know it more easily.

So, we have considered the population trends from 2001 to 2022, We see that only a few clusters are there in the top (Mumbai, Delhi) so we can say that Mumbai and Delhi form Cluster 1 and so on we see there are many cities in the bottom which are grouped together as their growth from 2001 to 2022 has been similar to each other.

So, we have Totally made 4 clusters and see that in cluster 1, 2 , 3 only 3 cities are present and in the cluster 4 all the rest cities are present.



Below is the example how Mumbai and Delhi come under cluster 1



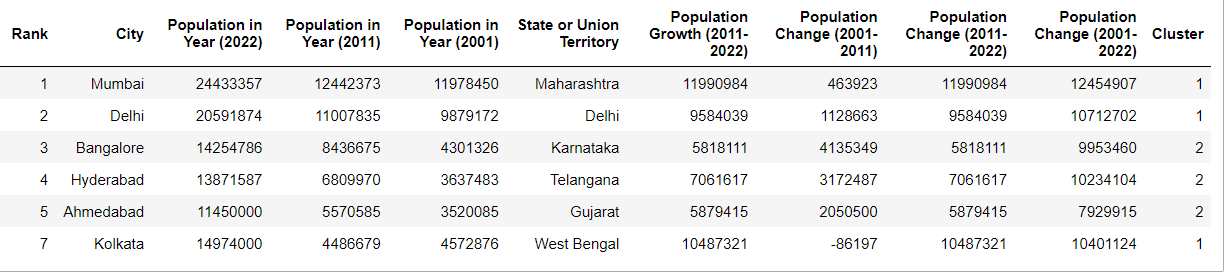
As we can see the growth of these 2 cities has been very similar from 2001 to 2022.

Hence these 2 cities come under Cluster 1 or have k-value come to each other.

**Country with Population More Than 10 million in 2022**

This analysis will help us to tell the top tier cities of India.

We see that all the cities appearing here are the IT-hubs of India and are rapidly moving to becoming the top cities of the World.



**Correlation Between Population Growth**

By Calculating the correlation, we get to know how much similar is the growth of population is from each year.

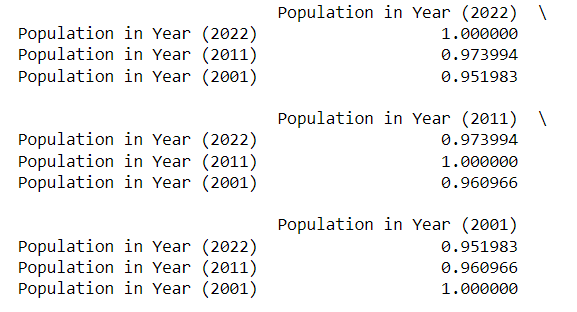
Similarly let’s take the correlation of the years 2011 and 2022.



Above is the value what we are getting, so if this was closer to 1 then the data is same as this is less than 1 then the population growth of 2022 is this much different from the population of 2011.

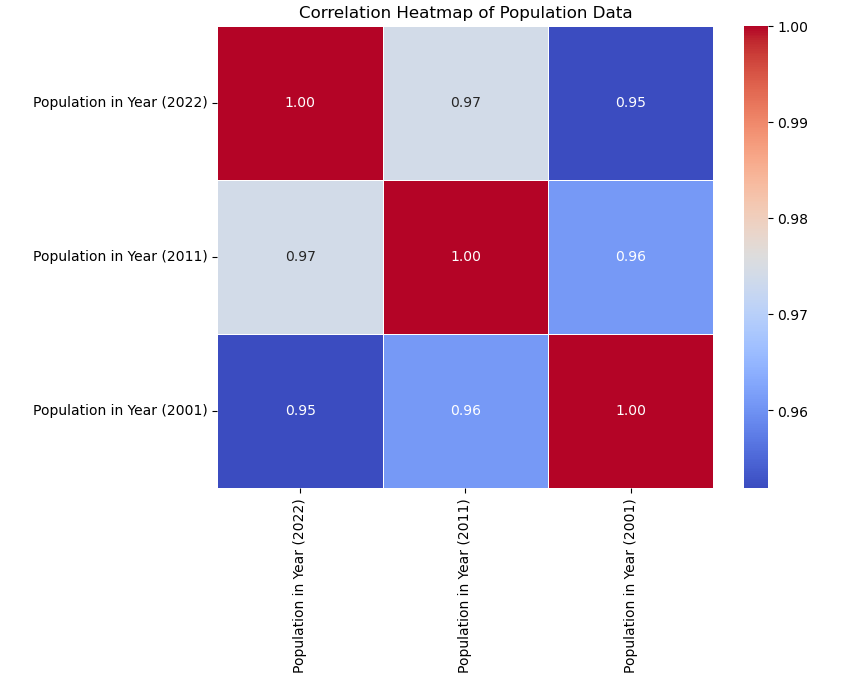
Lets us now compare correlation from each of the 3 years.

Below is the output which we got.



Let us make this into a heatmap as it is the best way to compare correlations with each other.

Red color say that the data is same and grey tells the data is not that similar and blue tells that data is very different to each other. We can understand it better by seeing the scale below in the heatmap.



**Null Hypothesis**

I tried to find the null hypothesis but as the data are not significantly different from each other, I could find it. Below is the output which I got.



**Findings & Inferences**

**Key Findings**

From our analysis, we found the following key findings:

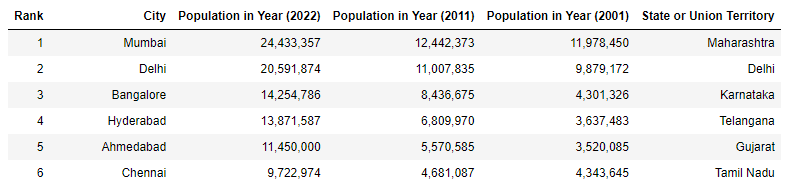
1. *Mumbai Dominates*: Mumbai is the most populated city in India in 2022 and also in 2011 and 2001, followed by Delhi and Bangalore.



2. *Decadal Growth*: Many cities experienced significant population growth between 2001 and 2011, indicating urbanization trends and also digitalization has brought in more changes from 2011 to 2022.

3. *Diverse State Performances*: Population growth varies significantly among states, emphasizing the importance of state-specific policies. States having company friendly policy will see rapid development.

4. *Southern Boom*: Cities in South India, such as Bangalore, Chennai, and Hyderabad, have shown remarkable population growth, potentially due to their thriving IT sectors.



5. *Population Disparities*: While some cities exhibited rapid growth, others faced stagnation or even decline, highlighting regional disparities in development. There may be various reasons for it to happen but it is important to changes those factors and work towards development.

6. *Emergence of NCR*: The National Capital Region (NCR), including Delhi and neighboring cities, has seen substantial population growth, indicating its significance as an economic and cultural hub. The surrounding areas like Gurgaon, Noida also saw significant changes in their population.

7. *Maharashtra's Urban Centers*: Maharashtra houses several highly populated cities, making it a key state for urban development initiatives.



8. *Urban Sprawl*: Several cities have expanded their boundaries over the years, leading to challenges related to infrastructure and services and also exploitation of forest and oceans.

**Inferences**

* *Economic Opportunities*: Cities with higher population growth may present increased economic opportunities for businesses, such as retail, real estate, and technology services.
* *Infrastructure Strain*: Rapid urbanization implies greater demands on infrastructure, necessitating immediate attention to transportation, housing, and public services which is not easy.
* *Policy Prioritization*: States and union territories should prioritize policies that address urban challenges, considering the diverse population trends in their regions and implementing strategy to benefit both customers and companies.
* *Migration Dynamics*: The influx of people into cities suggests internal migration patterns that can inform labor market policies and housing development strategies which tells us that there has been increase in Infrastructure work and rapid urbanization is taking place.
* *Regional Development*: There is a need for balanced regional development to prevent over-concentration of population and resources in select urban centers.

**Recommendations**

**For Businesses**

1. *Market Segmentation*: Businesses should segment their target markets based on population trends, considering both high-growth and stable cities and also target cities which have potential to grow.

2. *Localized Marketing*: Tailor marketing strategies to meet the specific needs and preferences of consumers in each city and also each state/union territory.

3. *Expansion Strategies and Opportunities:* Evaluate expansion opportunities in rapidly growing cities while considering factors like competition, consumer demand and consumer incomes.

4*. Infrastructure Partnerships*: Collaborate with local governments to address infrastructure gaps that may impact business operations in various cities.

**For Policymakers**

1. *Population Forecasting*: Develop robust population forecasting models to anticipate future urbanization trends and plan resources accordingly and see which cities is growing at what speed.

2. *Sustainable Urban Growth*: Prioritize sustainable urban planning, focusing on public transportation, affordable housing, green spaces and infrastructure.

3. *Inclusive Growth of all Classes*: Implement policies that ensure inclusive growth, reducing income and social disparities among urban residents and treat each urban classes equally.

4. *Investment Attraction*: Create an investor-friendly environment to attract businesses and promote job creation in underdeveloped urban areas and also rural areas so that growth in that area is fast and hence population growth will happen when there is increase in businesses.

**For Urban Planners**

1. *Mixed-Use Zoning*: Encourage mixed-use zoning in city planning to reduce commute times and improve overall quality of life in cities.

2. *Public Transportation*: Invest in efficient and sustainable public transportation systems to ease traffic congestion and reduce pollution. For Example, In most of the tier-1 cities electric buses are present which are sustainable as well as help in transportation.

3*. Smart City Initiatives*: Embrace smart city technologies to enhance urban services and improve quality of life of citizens of cities.

4. *Community Engagement*: Involve local communities in the urban planning process to ensure their needs are addressed and fulfilled.

**Managerial Insights | Implications**

**Business Implications**

The insights from this analysis have several business implications:

* *Market Expansion Strategy*: Companies can expand strategically into high-growth cities while optimizing their existing operations in mature markets. They can also invest in tier 2 cities and find a potential to grew businesses.
* *Consumer-Centric Strategies*: Tailoring products and services to meet the specific demands of diverse urban populations can improve customer satisfaction and loyalty. But the companies should also note to focus on rural consumers as well.
* *Supply Chain Management and Operations*: Efficient supply chain management is crucial, considering the varying logistics challenges posed by different cities. New ways can be found to improve it and hence increase customer base.
* *Competition Analysis*: Understanding the competitive landscape in each city is vital for market entry and sustainability of every business.

**Policy and Planning Implications**

The analysis also has implications for urban policy and planning:

* *Resource Allocation*: Governments should allocate resources efficiently, focusing on areas with the highest population growth and addressing infrastructure gaps.
* *Sustainability Initiatives*: Encouraging sustainable practices in construction, transportation, and waste management can mitigate urban challenges and help ease problems.
* *Inclusivity*: Policymakers must ensure that urban development benefits all residents, reducing income inequalities and enhancing social cohesion among themselves.
* *Data-Driven Decision-Making*: Embracing data analytics can empower policymakers to make informed decisions on urban planning and efficient resource allocation.

**Conclusion**

In conclusion, this project successfully extracted and analyzed data on the most populated Indian cities, revealing critical population trends and their implications. This data-driven approach can guide businesses, policymakers, and urban planners in making informed decisions to promote balanced urban development and economic growth. Hence benefitting customers with various choices in market.