

# **INDIAN CITIES DASHBOARD**

## **A PROJECT REPORT**

**TEAM – TDS014**

**Submitted by**

**VIJAYAKUMARAN S (21ADR061)**

**ROHITH S J (21ADR040)**

**VIBEESH N (21ADR059)**

**RISHI RAGHAV G (21ADR038)**

**VIGNESH T (21ADR060)**

*for*

**20ADC33 DATA ANALYSIS**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE**



**KONGU ENGINEERING COLLEGE  
(Autonomous)**

**PERUNDURAI ERODE – 638 060**

**DECEMBER 2022**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE**

**KONGU ENGINEERING COLLEGE**

**(Autonomous)**

**PERUNDURAI ERODE – 638 060**

**DECEMBER 2022**

Department of Artificial Intelligence

**20ADC33 – Data Analysis Project Report**

Signature of course in-charge

Signature of the HOD

Submitted for the continuous Assessment viva voice examination held on \_\_\_\_

**EXAMINER I**

**EXAMINER II**

## **ABSTRACT**

Cities are very important part of the states and the country. Because the development of the country majorly depends on the cities. Cities are the Large human settlements. More effective service and product delivery are happens in cities.

On the account of this busy cities we have many datas(like population, sex ratio, literacy rate etc...), so it is difficult to analyze the overall condition of the city. This project is to solve the analysis of those data and to easily get insights. Here analysis made on the cities of India and then a dashboard is created. At first the dataset collection, Four datasets were collected and taken to make the analysis. Two datasets named cities\_r2 and startup funding were taken from the Kaggle, One dataset which includes the tax details was taken from the official website of RBI(Reserve Bank of India). And another dataset which contains the language details was created. A tax is a mandatory financial charge or other sort of levy that is levied on a taxpayer by an institution of government to pay for government expenses and other public expenditures(regional, local, or national). This tax analysis on each city is also done in our project.

Here the analysis and the charts are drawn using PowerBI desktop with required DAX Formulas and the Dashboard is created and published using PowerBI service.

Then after making those complex analysis on the different datasets, the PowerBI file's report is uploaded in the PowerBI service. Then the new dashboard is created using the drawn chart and published. This Dashboard will give more details about the population, literacy, sex ratio, fundings, languages, tax etc., of the cities in India which will be more useful to make a study about the city and also solve many problems. The resources needed for the particular city or location can be predicted by analysis of population. Many relationship. The analysis of graduates in cities will helps the big companies to build their new branch over the required location.

## TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE NO.
	<b>ABSTRACT</b>	1
1.	<b>INTRODUCTION</b>	
	1.1 INTRODUCTION	2
	1.2 DATA COLLECTION	4
	1.3 PROBLEM STATEMENT	7
	1.4 BUSINESS OBJECTIVE	7
2.	<b>DATA PREPARATION AND MODELING</b>	
	2.1 DATA CLEANING	8
	2.2 DATA TRANSFORMATION	12
	2.3 DATA MODELLING	18
3.	<b>DATA ANALYSIS AND INTERPRETATION</b>	
	3.1 DATA ANALYSIS	21
	3.2 PUBLISHING DASBOARDS	35
	3.3 INFERENCE	36
4.	<b>CONCLUSION</b>	
	4.1 RECOMMENDATIONS	38
5.	<b>REFERENCES</b>	39

# CHAPTER 1

## INTRODUCTION

### 1.1 INTRODUCTION

Large human settlements include cities. It can be described as a permanent, heavily inhabited area with clearly defined administrative boundaries, whose residents largely participate in non-agricultural activities. For housing, transportation, sanitation, utilities, land use, commodity production, and communications, cities typically have vast systems. Their density makes it easier for people, organisations, and enterprises to interact, which may be advantageous to a number of parties.

More effective service and product delivery are happens in cities. More than half of the world's population currently lives in cities, a dramatic increase from the historically small percentage of people who called cities home. This has a tremendous impact on the sustainability of the planet.

The majority of commuters in today's cities travel to city centres for work, play, and education, forming metropolitan areas and metropolitan centres in the process. All cities are, however, to some extent globally networked outside of these regions in a world that is becoming more and more globalised. Due to their expanding importance, cities are also having a bigger impact on global challenges like global warming, sustainable development, and public health. The international community has emphasised investment in sustainable cities through Sustainable Development Goals because of the considerable impact on these global challenges.

Due to less efficient transportation and less land use, densely crowded cities may have a lower ecological footprint per resident than less densely inhabited locations. As a result, the importance of compact cities in the fight against climate change is frequently noted. Toxic side effects from this concentration include, but are not limited to Straining water supplies and other resources, concentrating pollution, and creating urban heat islands.

A tax is a mandatory financial charge or other sort of levy that is levied on a taxpayer by an institution of government to pay for government expenses and other public expenditures(regional, local, or national). A tax is an obligatory financial charge or other sort of levy that is levied on a taxpayer by a government entity to pay for public services and other expenses. This tax analysis on each city is also done in our project. Thus finally our project output is an Dashboard which consists minimum of 20 chart. This Dashboard gives an overall analysis of an City which will be useful to some Business Objectives in which they need data like analysing the City population, Tax, graduates, literates etc.,

The dataset that is given for analysis and visualize it using dashboard were “Indian cities Dashboard”.

- The dataset is collected from Kaggle and it contains mostly all the attributes essential for creating visuals and dash boards.
- The dataset is available in Excel format.
- The dataset was directly download data's from Kaggle and import them into power bi file for further processing and visualisations.
- Then the collected data's are subjected for pre-processing.
- In data pre-processing it involves filtering, cleansing, de-duplicating, validating and finally authenticating data.
- Formatting the data into tables or joined tables to match target schema.
- Performing calculations, translations, summarizations, changing rows and columns datatype, change null values, apply DAX measures, etc..

## 1.2 DATA COLLECTION

### Indian cities – Dataset

Source <https://www.kaggle.com/datasets/zed9941/top-500-indian-cities>

### Dataset description

This dataset is created and uploaded in KAGGLE by ARIJIT MUKHERJEE. This Indian Cities Dataset contains 493 rows and 22 columns. Each and every single row in this Dataset represents a city which gives 22 set of information of that city. This Dataset gives an enough data's to analyze a city and to get a very good insight about that city.

Any group of individuals will use this analysis the most frequently. This dataset may take further action based on our understanding of the correlations between male and female literacy rates, such as the areas with the greatest rates of literacy and the greater representation of women in the education and workforce. Despite a more than six-fold increase, the level still falls short of the global average literacy rate of 84%. The 2011 census showed a decadal gain in literacy of 9.2% from 2001 to 2011, which is less than the growth experienced over the preceding decade.

And there are more information like 0-6 age population (it is classified into total population, male population , female population), literacy(it is classified into to (male literates , female literates, total literates) and sex-ratio (child-sex-ratio), location(longitude and latitude).

India has a 74.04% literacy rate in 2011. According to Census 2011, the literacy rates for men are 82.14% and for women are 65.46%. Kerala has the highest literacy rate among the Indian states (93.91%), followed by Mizoram (91.58%). Lakshadweep has the highest percentage of literacy (92.28%) among the Union Territories. India's lowest literacy percentage is 63.82% in Bihar.

Lakshadweep (96.11%) and Kerala (96.02%) have the highest male literacy rates. Kerala (91.98%) and Mizoram (89.40%) have the highest rates of female literacy. Bihar 73.39% has the lowest male literacy rate. In Rajasthan, just 52.66% of women are literate.

The dataset comprises of latitude and longitude of most of the important cities in India. The dataset consisted of city name, state to which it belongs & its latitude and longitude. Latitude and longitude are angles that specifically identify places on a sphere, if such were the case. The angles make up a coordinate system that may be used to find or identify certain geographic locations on the surfaces of planets like the earth.

So to locate any the city on the earth, the city's longitude and latitude need to be known. Well longitude and latitude are basically measured in degrees but it will be considered as float for the sake of simplicity.

Literacy is defined as the capacity to read and write at least a few simple lines or messages in any language in both national and international usage. The absence or lack of this capacity is referred to as illiteracy. In other words, someone is said to be literate if they have both the ability to read and write. This dataset that contains and comprises of literates in the particular city.

## Variables in the dataset and what it represents

'name_of_city'	:	Name of the City
'state_code'	:	State Code of the City
'state_name'	:	State Name of the City
'dist_code'	:	District Code where the city belongs (99 means multiple district)
'population_total'	:	Total Population
'population_male'	:	Male Population
'population_female'	:	Female Population
'0-6_population_total'	:	0-6 Age Total Population
'0-6_population_male'	:	0-6 Age Male Population
'0-6_population_female'	:	0-6 Age Female Population
'litrates_total'	:	Total Literates
'litrates_male'	:	Male Literates
'litrates_female'	:	Female Literates
'sex_ratio'	:	Sex Ratio
'child_sex_ratio'	:	Sex ratio in 0-6
'effective_literacy_rate_total'	:	Literacy rate over Age 7
'effective_literacy_rate_male'	:	Male Literacy rate over Age 7
'effective_literacy_rate_female'	:	Female Literacy rate over Age 7
'location'	:	Latitude and Longitude
'total_graduates'	:	Total Number of Graduates
'male_graduates'	:	Male Graduates
'female_graduates'	:	Female Graduates

More than 90 percentage of the people in the cities were educated and graduates. From this dataset a good insight about the graduates in the cities can be obtained. And also it can be used to perform some analysis on graduates from our dataset.

### **TAX – DATASET :**

**Source** <https://m.rbi.org.in/scripts/PublicationsView.aspx?id=20809>

### **Dataset description**

This dataset from the official site of RBI(Reserve Bank of India).

This dataset contains two tables. The both tables contains the tax details of all the states of India. In the Tax dataset, Table 1 contains 7 columns and 32 rows, then Table 2 contains 10 columns and 32 rows. In one table the tax details from the year 2004 to 2012 and on the other table the tax details from the year 2012 to 2021 of the all states of India. This dataset will be very useful to analyze the money and financial details of cities and also the states of India. Since the tax details year wise, the analysis made will be more precise. In this dataset the table 1 contains fields which includes state name, 2004-05, 2005-06, 2006-07, 2007-08, 2008-09, 2009-10, 2010-11, 2011-12 and the table 2 contains fields which includes state name, 2012-13, 2013-14, 2014-15, 2015-16, 2016-17, 2017-18, 2018-19, 2019-20, 2020-21.

### **Startup funding dataset**

#### **Source**

<https://www.kaggle.com/code/codename007/top-funding-startups-in-india/data>

### **Dataset description**

The Startup-funding-dataset is taken from the Kaggle. The Startup-funding-dataset contains 7 columns and 1385 rows of data. This Startup-funding-dataset gives the details like what the startup name, from which city it was started, when it was started, Investment Type and how much money of funding is sanctioned for the startup which will be very useful for us to get a very good analysis on the startups based on the cities of India.

The Fields in this Startup-funding-dataset are serial number, Date, Startup name, Investor name, Investment Type, Amount in USD. This dataset contains details of 1385 startups which is very useful to get a analysis how our country is growing in this startup and Entrepreneur Domains which is the major thing to make our country from developing stage to developed stage which will also one of the main in increase in the financial status of each cities and states of India.



## **States languages dataset**

This dataset is created which contains state name and the languages spoken in the states. This States Languages dataset contains serial number, state name, first language and second language spoken in the states. By using this data a good analysis on the languages spoken in cities and states of India.

### **1.3 PROBLEM STATEMENT**

There are many cities in India, so there will be many Data's to store(for example total population ,male population , female population , literature, male-literacy-rate, female-literacy-rate , child-sex-ratio , sex-ratio , tax , language spoken etc..). On the view this whole data there will be more rows(which is independent of priority) so it will be difficult to view top values in data.

Sex ratio is higher in many cities for example Bhiwandi sex ratio is 700 so there 700 women's for every 1000 men. It is major problems now a days the men population is dominant than female population. This will lose of many job opportunities for one gender in the cities.

In top cities population is major problem it leads to crowding, pollution etc.. for example Delhi is the second most populated in India it has 75% of air polluted because of over population and usage of many vehicle. It also an emerging problem in many leading cities.

### **1.4 BUSINESS OBJECTIVE**

- To visualize the data using dashboard so the data of top cities can be viewed. The dashboard is used for education of cities in school, college etc.
- To analyze the better performing cities based on educational and other fields.
- To analyze of graduates in cities which will helps for big companies to make their branch office in that particular city.
- To Analyze the population of cities, which is used to predict the resources needed for the particular city or location.

## **CHAPTER 2**

### **DATA PREPARATION AND MODELING**

#### **2.1 DATA CLEANING**

Editing, fixing, and organising data inside a data collection to make it more consistent and ready for analysis is known as data cleaning. This requires removing inaccurate or unneeded data and arranging it in a way that computers can comprehend in order to produce the best analysis.

In data analysis, the phrase "garbage in, garbage out" is frequently used to suggest that if you start with faulty data (trash), all of your outputs will be junk.

Although it might be a time-consuming procedure, data cleansing is vitally necessary to extract the best findings and most insightful conclusions from your data.

The 1-10-100 principle effectively explains this: It costs \$1 to prevent faulty data, \$10 to correct bad data, and \$100 to fix a problem that bad data caused downstream. Therefore, it's crucial that you carry out comprehensive data cleansing to guarantee that you obtain the greatest outcomes.

Data Cleaning Process includes :-

Step-I : Removal of irrelevant-datas

Step-II : Deduplicating our data

Step-III: Fixing structural-errors

Step-IV: Dealing with missing-datas

Step-V : Filtering out data-outliers

Step-VI : Validating our data

STEP 1 : Loading the Cities\_r2 dataset into the PowerBI Desktop.

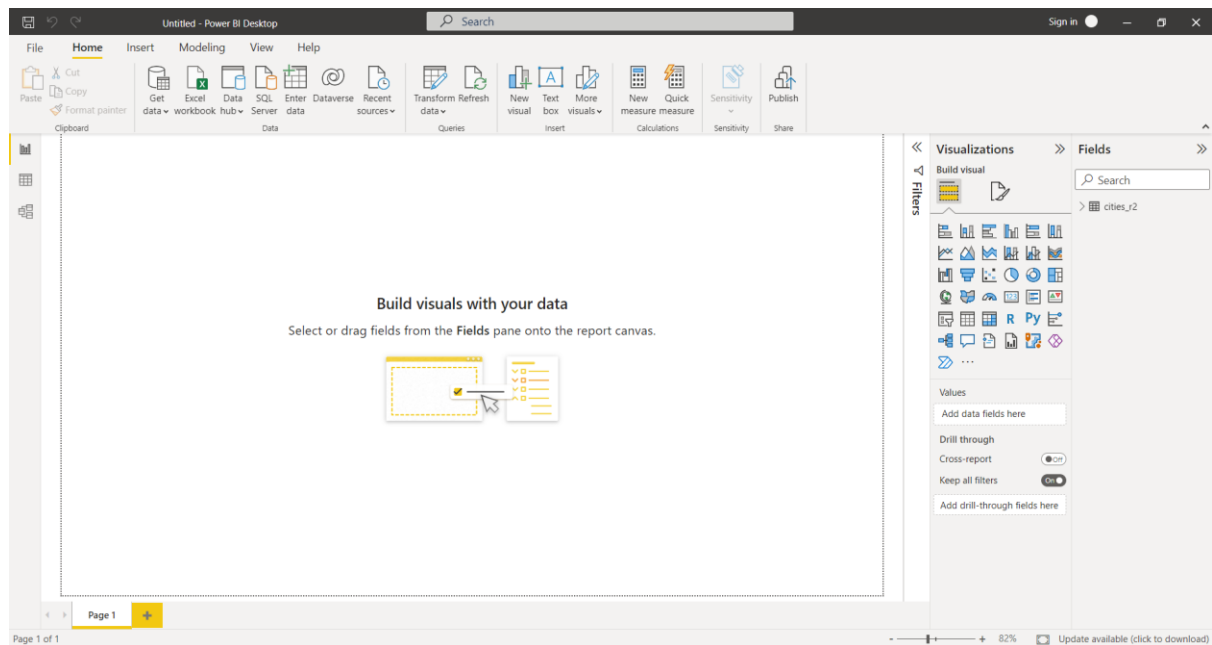


Figure : 2.1 Loading the Cities\_r2 Dataset into PowerBI

STEP 2 : Power Query Editor window is opened by clicking on the Transform data.

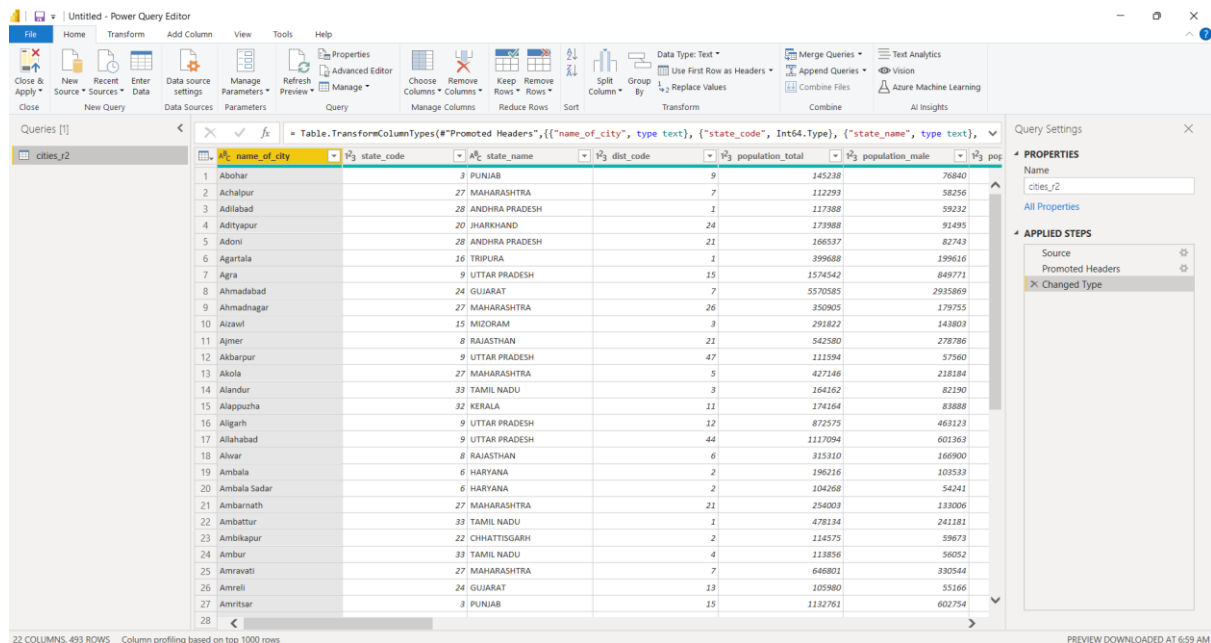


Figure : 2.2 Opening the Query editor window

STEP 3 : Change the table name from cities\_r2 to Indian\_cities using table tools.

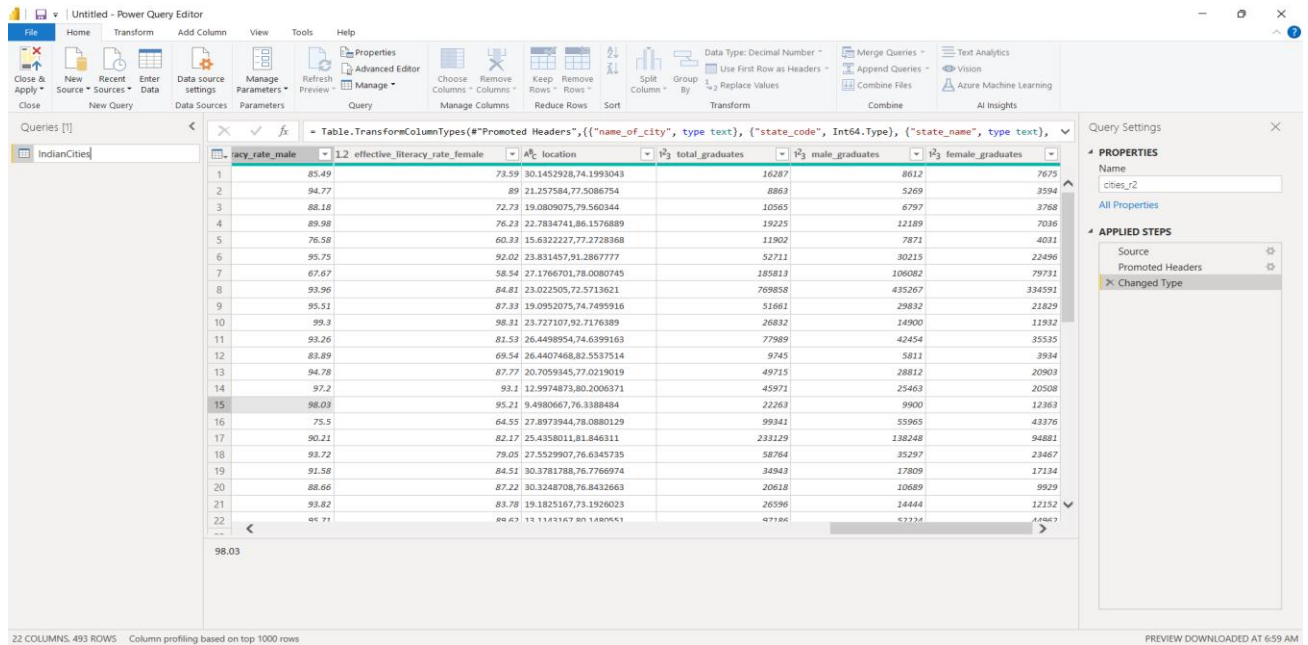


Figure : 2.3 Changing the table name as IndianCities

STEP 4: The column name is changed from dist\_code to district\_code using column tools.

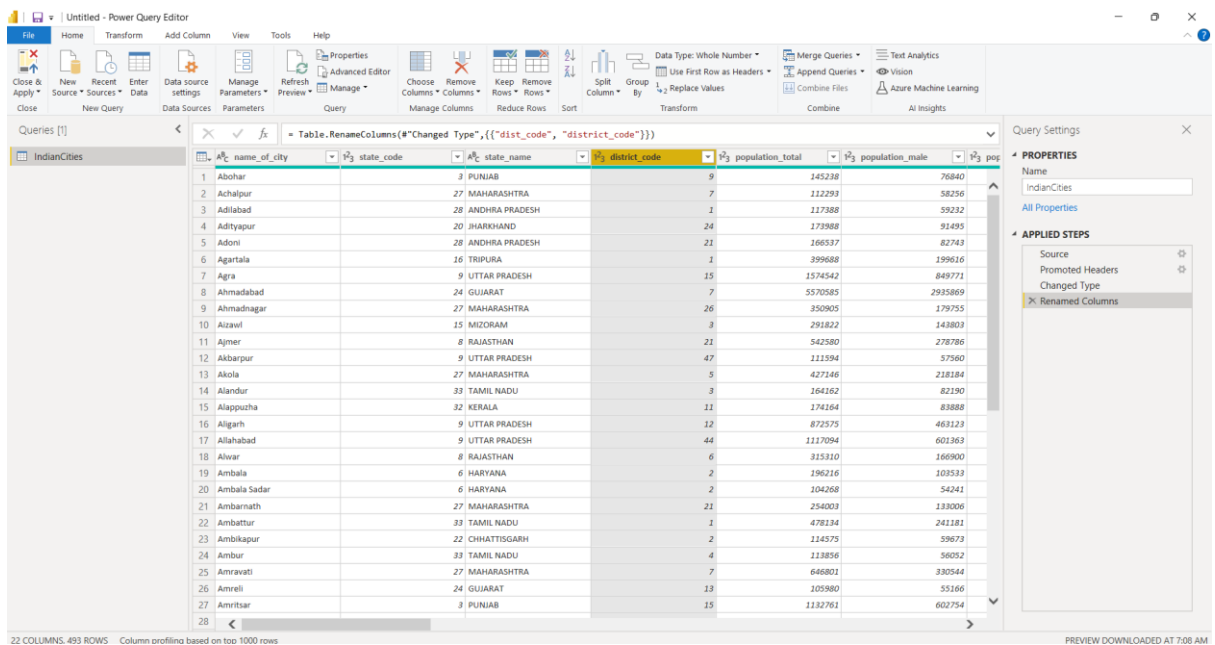


Figure : 2.4 Changing the Column name as district\_code

STEP 5: Then the Tax dataset which has two tables (1. Tax data of year 2004-2012 and 2. Tax data of year 2012 - 2021) was imported by clicking the New Source on the home tab

Table: TransformColumnTypes(\*Promoted Headers", ({"TABLE 150: STATE-WISE OWN TAX REVENUE (ConcId.)", type text}, {"Column2", type

State/Union Territory	2012-13	2013-14	2014-15	2015-16
Andhra Pradesh	59875	64124	42618	39907
Arunachal Pradesh	316	435	462	535
Assam	8250	8995	9450	10107
Bihar	16253	19961	20750	25449
Chhattisgarh	13034	14343	15707	17075
Goa	2940	3582	3896	3975
Gujarat	53897	56372	61340	62649
Haryana	23559	25567	27635	30929
Himachal Pradesh	4626	5121	5940	6696
Jammu and Kashmir	5832	6273	6334	7326
Jharkhand	8224	9380	10350	11479
Karnataka	53754	62604	70180	75550
Kerala	30077	31995	35233	38995
Madhya Pradesh	30582	33552	36567	40214
Maharashtra	103449	108598	115064	126608
Manipur	333	473	517	552
Meghalaya	848	949	939	1057
Mizoram	223	230	267	358
Nagaland	340	333	389	427
Odisha	15034	16892	19828	22527
Punjab	22588	24079	25570	26690
Rajasthan	30503	33478	38673	42713
Sikkim	435	525	528	567
Tamil Nadu	71254	73718	78657	80476
Telangana	-	-	29288	39975
Tripura	1005	1074	1174	1332
Uttar Pradesh	58098	66582	74172	81106

Figure : 2.5 Changing the Tax dataset tables names as TAX\_2004\_12 & TAX\_2012\_21

STEP 6 : The column name State/Union Territory in both Tax table is changed to state\_name.

Table: RenameColumns(\*Changed Type", ({"State/Union Territory", "state\_name"}))

state_name	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18
Andhra Pradesh	59875	64124	42618	39907	44181	44181
Arunachal Pradesh	316	435	462	535	709	709
Assam	8250	8995	9450	10107	12080	12080
Bihar	16253	19961	20750	25449	23742	23742
Chhattisgarh	13034	14343	15707	17075	18945	18945
Goa	2940	3582	3896	3975	4261	4261
Gujarat	53897	56372	61340	62649	64443	64443
Haryana	23559	25567	27635	30929	34026	34026
Himachal Pradesh	4626	5121	5940	6696	7039	7039
Jammu and Kashmir	5832	6273	6334	7326	7819	7819
Jharkhand	8224	9380	10350	11479	13299	13299
Karnataka	53754	62604	70180	75550	82956	82956
Kerala	30077	31995	35233	38995	42176	42176
Madhya Pradesh	30582	33552	36567	40214	44194	44194
Maharashtra	103449	108598	115064	126608	136592	136592
Manipur	333	473	517	552	587	587
Meghalaya	848	949	939	1057	1186	1186
Mizoram	223	230	267	358	442	442
Nagaland	340	333	389	427	511	511
Odisha	15034	16892	19828	22527	22852	22852
Punjab	22588	24079	25570	26690	27747	27747
Rajasthan	30503	33478	38673	42713	44372	44372
Sikkim	435	525	528	567	653	653
Tamil Nadu	71254	73718	78657	80476	85941	85941
Telangana	-	-	29288	39975	48408	48408
Tripura	1005	1074	1174	1332	1422	1422
Uttar Pradesh	58098	66582	74172	81106	85966	85966

Figure : 2.6 Renaming the column name as state\_name

**STEP 7 :** The States\_languages dataset which contains states, its languages and its Second languages was imported by clicking on the New Source on home tab.

S.No.	States	Languages	Secondary_Languages
1	Andhra Pradesh	Telugu	English
2	Arunachal Pradesh	English	English
3	Assam	Assamese	Bengali, Bodo
4	Bihar	Hindi	Urdu
5	Chhattisgarh	Hindi	Chhattisgarhi
6	Goa	Konkani, English	Marathi
7	Gujarat	Gujarati	Hindi
8	Haryana	Hindi	English, Punjabi
9	Himachal Pradesh	Hindi	Sanskrit
10	Jharkhand	Hindi	Angika, Bengali, Bhojpuri, Ho, Kharia, Khortha, Kurmali, Kurukh, Maga...
11	Karnataka	Kannada	English
12	Kerala	Malayalam	English
13	Madhya Pradesh	Hindi	English
14	Maharashtra	Marathi	English
15	Manipur	Manipuri	English
16	Meghalaya	English	Khasi and Garo
17	Mizoram	Mizo	English, Hindi
18	Nagaland	English	English
19	Odisha	Odia	English
20	Punjab	Punjabi	English
21	Rajasthan	Hindi	English
22	Sikkim	English, Nepali, Sikkimese, Lepcha	Gurung, Limbu, Magar, Mukhia, Newari, Rai, Sherpa and Tamang
23	Tamil Nadu	Tamil	English
24	Telangana	Telugu	Urdu
25	Tripura	Bengali, English, Kokborok	English
26	Uttar Pradesh	Hindi	Urdu
27	Uttarakhand	Hindi	Sanskrit
28	West Bengal	Bengali, English	Nepali, Urdu, Hindi, Odia, Santali, Punjabi, Kamtapuri, Rajbanshi, Kurm...

Figure : 2.7 State\_languages dataset is imported

## 2.2 DATA TRANSFORMATION

The process of changing data from one format to another, usually from that of a source system into that needed by a destination system, is known as data transformation. Most data integration and management operations, including data wrangling and data warehousing, include some type of data transformation.

Data transformation, a phase in the ELT/ETL process, can be categorised as "simple" or "complicated" depending on the kind of adjustments that must be made to the data before it is sent to its intended destination. The data transformation procedure can be carried out automatically, manually, or by combining the two.

Data transformation is now more crucial for organisations than ever because of the realities of big data. Numerous gadgets, apps, and algorithms continuously generate enormous amounts of data. Additionally, data compatibility is constantly under danger due to the vast amount of inconsistent data coming from many sources. To solve this problem, businesses and organisations may turn data from any source into a format that can be integrated, saved, examined, and eventually mined for useful business insight.

STEP 1 : In the IndianCities dataset, There are two values given in the location column which can be separated by comma “,”.

Figure 2.8 shows the Power Query Editor interface. The 'Queries' pane on the left lists 'Indian\_cities'. The main area displays a table with columns: 'racy\_rate\_male', '1.2 effective\_iteracy\_rate\_female', '1.2 location', '1.2 total\_graduates', '1.2 male\_graduates', and '1.2 female\_graduates'. The 'location' column is highlighted. The formula bar at the top shows the M code: `= Table.RenameColumns(#"Changed Type",{{"dist_code", "district_code"}})`. The right-hand 'Query Settings' pane shows the 'APPLIED STEPS' list, which includes 'Source', 'Promoted Headers', 'Changed Type', and 'Renamed Columns'.

Figure : 2.8 IndianCities dataset

Figure 2.9 shows the Power Query Editor interface after renaming the location column. The 'Queries' pane on the left lists 'Indian\_cities'. The main area displays a table with columns: 'active\_iteracy\_rate\_female', '1.2 longitude', '1.2 latitude', '1.2 total\_graduates', '1.2 male\_graduates', and '1.2 female\_graduates'. The 'longitude' and 'latitude' columns are highlighted. The formula bar at the top shows the M code: `= Table.RenameColumns(#"Changed Type1",{{"location.1", "longitude"}, {"location.2", "latitude"}})`. The right-hand 'Query Settings' pane shows the 'APPLIED STEPS' list, which includes 'Source', 'Promoted Headers', 'Changed Type', 'Renamed Columns', 'Split Column by Delimiter', 'Changed Type1', and 'Renamed Columns1'.

Figure : 2.9 Renaming the separated columns into Latitude and Longitude

STEP 2 : Since the column in the Tax table are the first row of data. So the first row is changed to headers by clicking the use first row as header in the two table also.

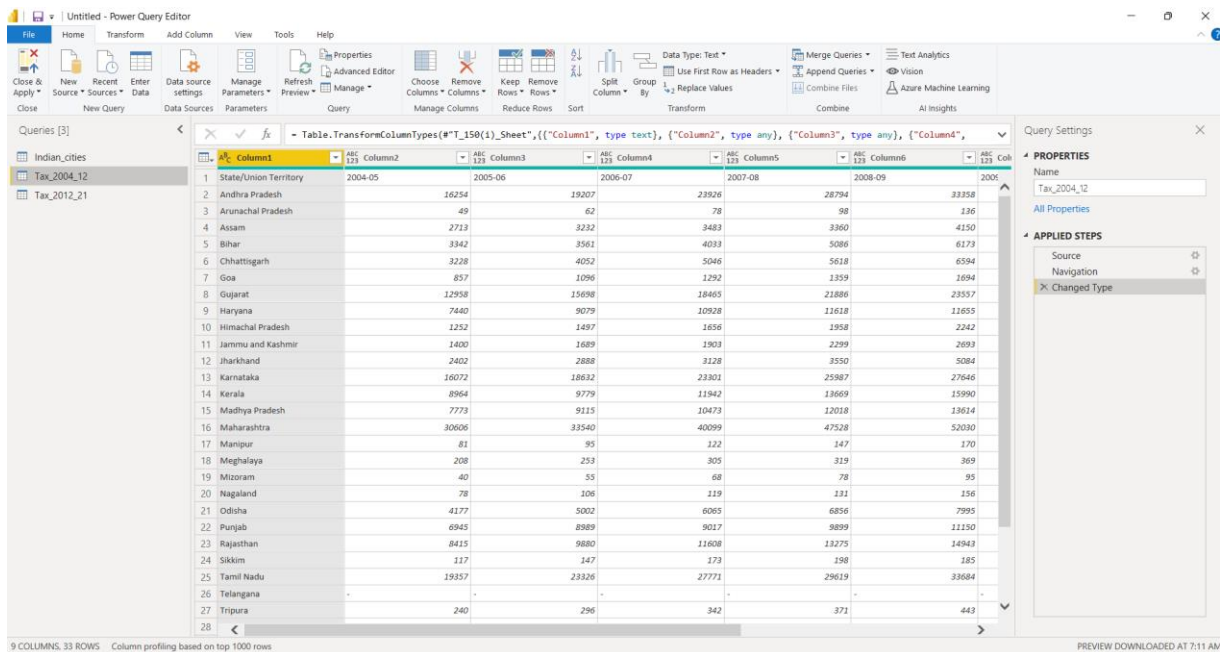


Figure : 2.10 Changing the first row of the Tax tables as headers

STEP 3 : The startup\_funding dataset which has attributes like startups name, City name, Amount of funding, Date , etc., This dataset has more null values, so the null values are found, treated and removed by Python programming using pandas package.

- To model the Dataset, first read the dataset and explored its shape and info.

```
[1] import pandas as pd

[2] data = pd.read_csv("/content/startup_funding.csv")

[3] data.shape

(2372, 10)

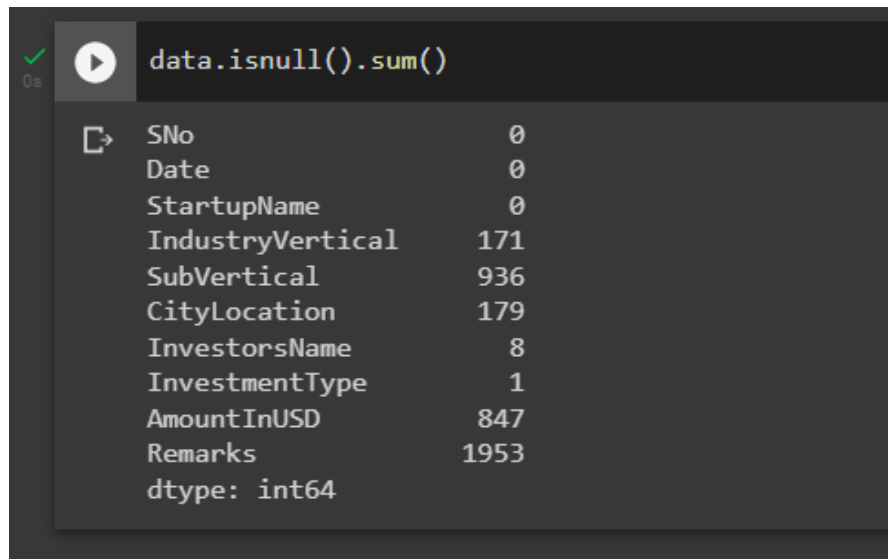
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2372 entries, 0 to 2371
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   SNo                    2372 non-null  int64
1   Date                   2372 non-null  object
2   StartupName            2372 non-null  object
3   IndustryVertical        2201 non-null  object
4   SubVertical            1436 non-null  object
5   CityLocation           2193 non-null  object
6   InvestorsName          2364 non-null  object
7   InvestmentType         2371 non-null  object
8   AmountInUSD            1525 non-null  object
9   Remarks                419 non-null   object
dtypes: int64(1), object(9)
memory usage: 185.4+ KB
```

Figure : 2.11 Startup dataset was imported and explored in G-Colab



- Then found the null values in each attributes.

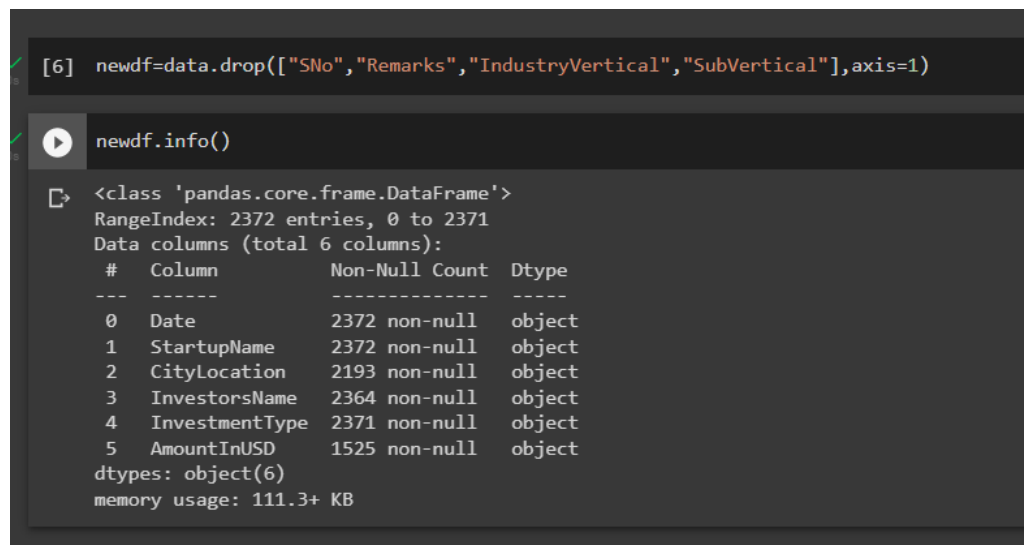


```
data.isnull().sum()
```

SNo	0
Date	0
StartupName	0
IndustryVertical	171
SubVertical	936
CityLocation	179
InvestorsName	8
InvestmentType	1
AmountInUSD	847
Remarks	1953
dtype:	int64

Figure : 2.12 Found null values in Startup dataset

- Then remove the unwanted attributes like SNo, Remarks, IndustryVertical, SubVertical.



```
[6] newdf=data.drop(["SNo","Remarks","IndustryVertical","SubVertical"],axis=1)
```

```
newdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2372 entries, 0 to 2371
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Date             2372 non-null   object
1   StartupName      2372 non-null   object
2   CityLocation     2193 non-null   object
3   InvestorsName    2364 non-null   object
4   InvestmentType   2371 non-null   object
5   AmountInUSD      1525 non-null   object
dtypes: object(6)
memory usage: 111.3+ KB
```

Figure : 2.13 Removing unwanted attributes in Startup dataset

- Then remove rows which contains atleast one null values and rechecked for null values in each attributes/columns.

```
[8] newdf.isnull().sum()
Date      0
StartupName 0
CityLocation 179
InvestorsName 8
InvestmentType 1
AmountInUSD 847
dtype: int64

[9] newdf.shape
(2372, 6)

[10] final_dataset=newdf.dropna()

[11] final_dataset.isnull().sum()
Date      0
StartupName 0
CityLocation 0
InvestorsName 0
InvestmentType 0
AmountInUSD 0
dtype: int64

[12] final_dataset.shape
(1385, 6)
```

Figure : 2.14 Removing rows containing null values in Startup dataset

- Finally the dataset is saved in csv file.

```
[13] final_dataset.index = range(0,len(final_dataset),1)

[14] final_dataset
```

	Date	StartupName	CityLocation	InvestorsName	InvestmentType	AmountInUSD
0	01/08/2017	TouchKin	Bangalore	Kae Capital	Private Equity	1,300,000
1	02/08/2017	Zepo	Mumbai	Kunal Shah, LetsVenture, Anupam Mittal, Hetal ...	Seed Funding	500,000
2	02/08/2017	Click2Clinic	Hyderabad	Narottam Thudi, Shireesh Palle	Seed Funding	850,000
3	01/07/2017	Billion Loans	Bangalore	Reliance Corporate Advisory Services Ltd	Seed Funding	1,000,000
4	03/07/2017	Ecolibriumenergy	Ahmedabad	Infuse Ventures, JLL	Private Equity	2,600,000
...	...	...	...	...	...	...
1380	29/04/2015	Icertis	Pune / US	Greycroft Partners, Fidelity Growth Partners	Private Equity	6,000,000
1381	29/04/2015	Tracxn	Bangalore	SAIF Partners	Private Equity	3,500,000
1382	29/04/2015	Tradelab	Bangalore	Rainmatter	Seed Funding	400,000
1383	29/04/2015	PIQube	Chennai	The HR Fund	Seed Funding	500,000
1384	29/04/2015	Travel Triangle	Noida	Bessemer Venture Partners, SAIF Partners	Private Equity	8,000,000

1385 rows x 6 columns

```
final_dataset.to_csv("/content/Startup_funding_dataset.csv")
```

Figure : 2.15 New Startup dataset is Saved

STEP 4 : Then the preprocessed Startup\_funding\_dataset is imported into the Power Query Editor by clicking on the New Source in Power Query Editor.

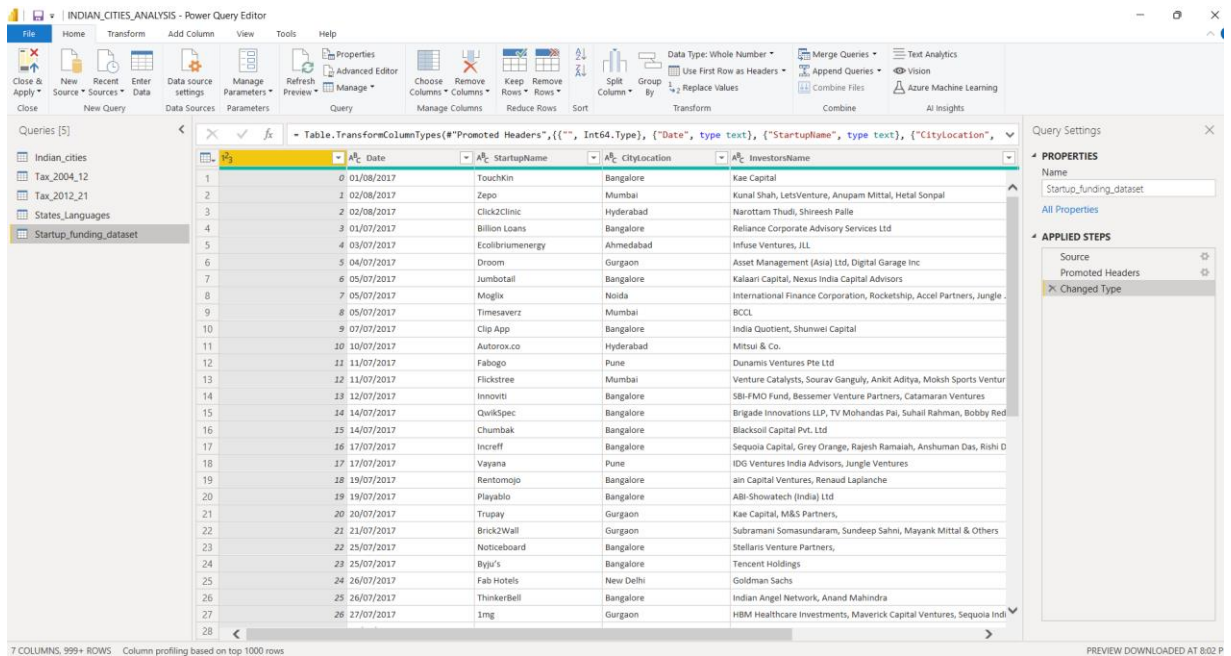


Figure : 2.16 The new Startup Funding dataset is loaded into PowerBI

STEP 5 : The first column is renamed as SNo.

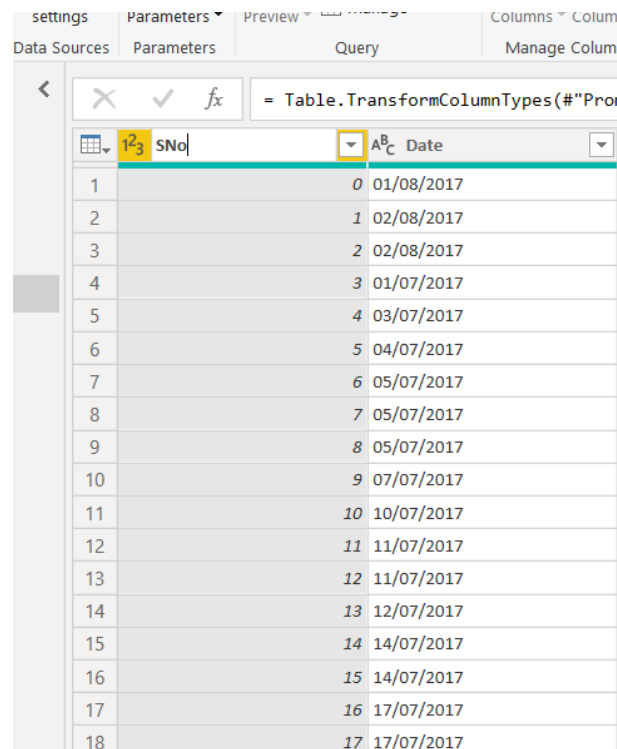


Figure : 2.17 Changing the Column name as SNo

## 2.3 DATA MODELLING

Data modelling is the act of describing and evaluating all the many types of data that your company creates or gathers, as well as the connections between those data. Data Modelling is the process of creating relationship between the tables to make use the dataset in many different angles in the analysis part.

There are four types of relationships can be made between two datasets they are,

- Many to Many relationship
- One to One relationship
- One to Many relationship
- Many to One relationship

By making this relationship, a very good analysis on different fields from different tables can be made on the same chart. To make or create a relationship between two tables (or Entity) first a clear analysis on the attributes in both the entities should be done and then the common attributes which has similar values in the both the entities should be identified. Then this attributes are used to create a relationship between the tables. This attributes which is responsible to make relationship are known as keys. They may be either primary or foreign key.

**STEP 1 :** To create relationship between the table click on the model and by drag and drop the attributes which is needed to make relationships

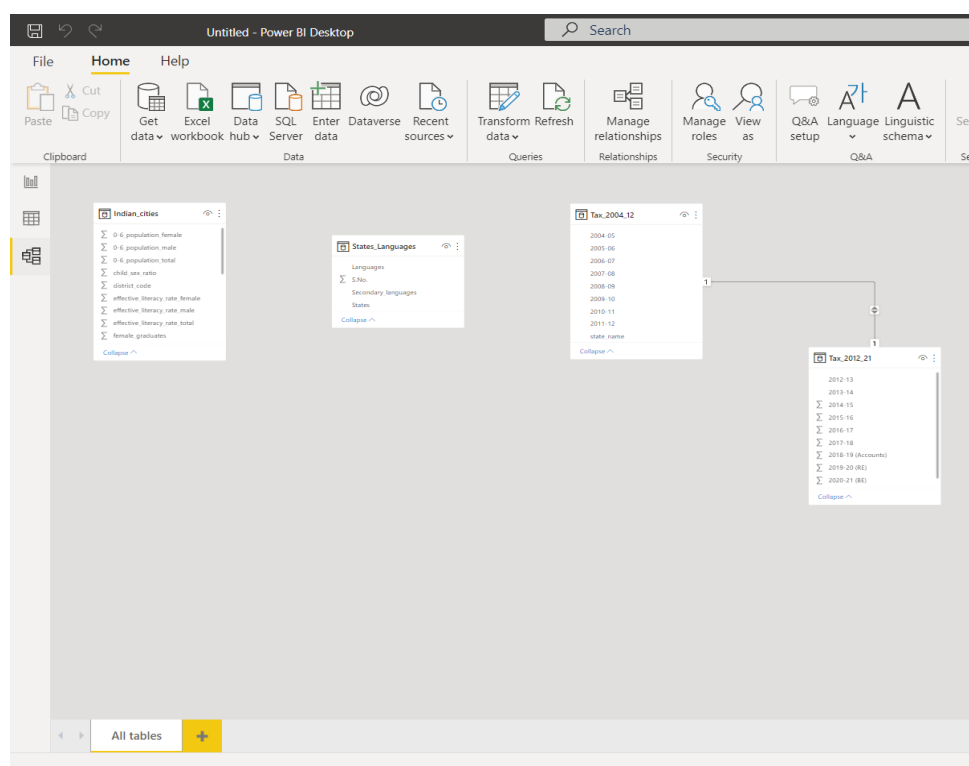


Figure : 2.18 Model view

STEP 2 : The relationship between the Indian\_cities table and the States\_language table by drag and dropping the States from the States\_language table to the state\_name of Indian\_cities table. A many to one relationship is created between Indian\_cities table to States\_language table.

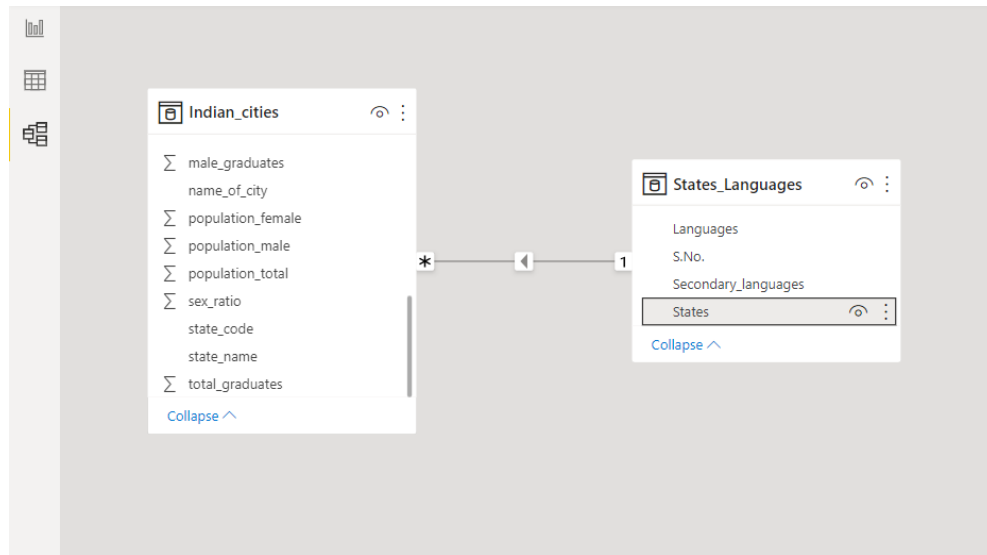


Figure : 2.19 Relationship between Indian\_cities and States\_languages [many to one]

STEP 3 : The relationship between the Indian\_cities table and the Tax\_2004\_12 and Tax\_2012\_21 tables by drag and dropping the state\_name from the Tax\_2004\_12 and Tax\_2012\_21 tables to the state\_name of Indian\_cities table. A many to one relationship is created between Indian\_cities table to Tax\_2004\_12 and Tax\_2012\_21 tables.

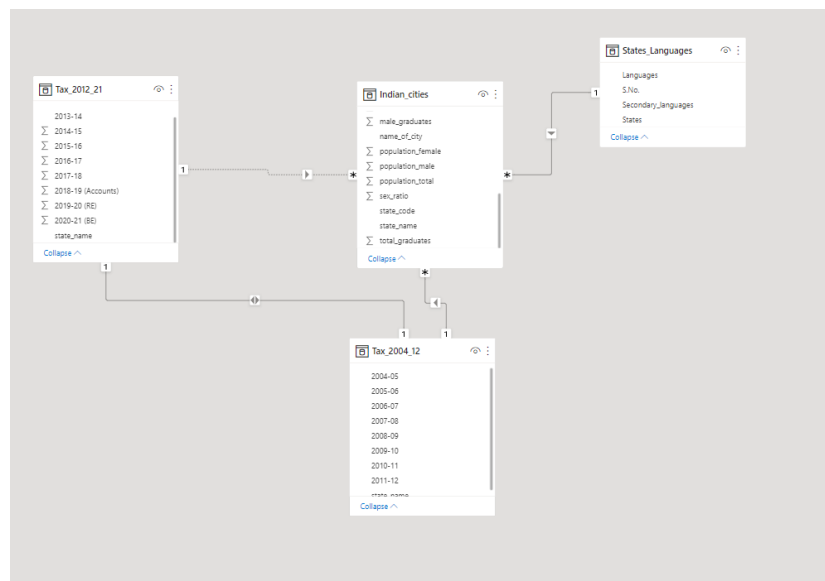


Figure : 2.21 The relationship between the Indian\_cities table and the two Tax tables [many to one]

STEP 4 : The relationship between the Indian\_cities table and the Startup\_funding\_dataset table by drag and dropping the CityLocation from the Startup\_funding\_dataset tables to the name\_of\_city of Indian\_cities table. A many to many relationship is created between Indian\_cities table to Startup\_funding\_dataset table.

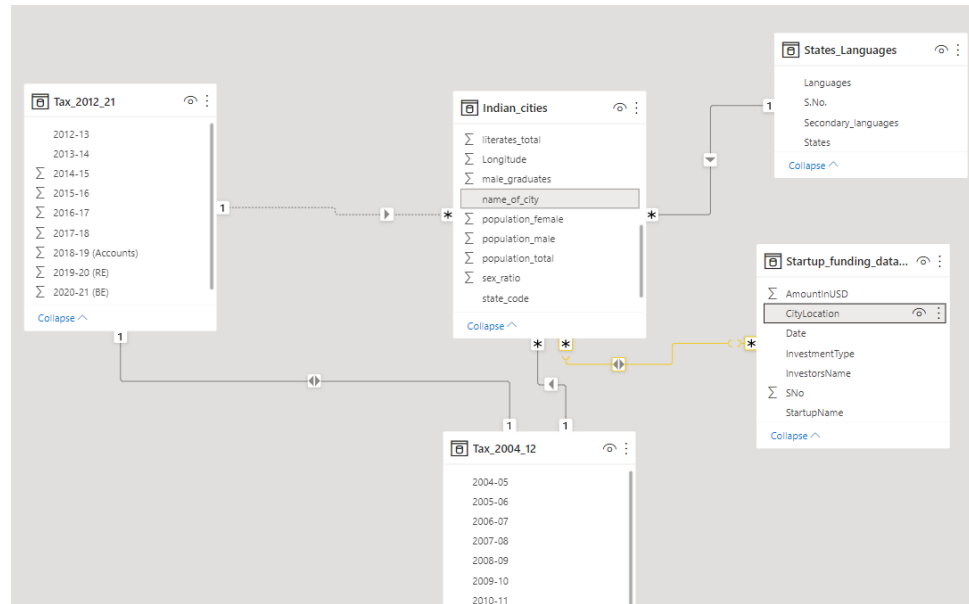


Figure : 2.22 The relationship between the Indian\_cities table and startup\_funding [many to many]

## **CHAPTER 3**

### **DATA ANALYSIS AND INTERPRETATION**

#### **3.1 DATA ANALYSIS**

Data Analysis is the main process to get the insights from the data. After cleaning, transforming and modelling the data, the data will be in form which is easy to make analysis. From this Through data analysis, valuable information can be extracted, draw conclusions, and help decision-making. It is a process of dividing the whole data into smaller individual components and get a very good insight from the data.

#### **Analysis made**

- Male graduates rate by each city In india and which city has most number of male graduates.
- What are the least populated cities in India ?
- Population of both males and female in each cities.
- Child population form age 0-6 in each cities?
- In which cities most of the people speak Tamil language ?
- Population of female child from age 0-6 in each cities.
- In which cities most number of female graduates are present?
- Which city has most Population of people speaking Hindi ?
- In which city most of people speak English as a second language?
- In which city has large amount of investment in its start up companies?
- Total start-up amount invested in each states.
- Literacy rate of male and female by each cities.
- In which city has least amount of investment in its start up companies?
- Which city has highest number of child sex ratio?
- Which city has highest number of child sex ratio?
- Total tax form all states and Tamil Nadu tax form 2017 to 2021?
- Total of people who all are speaking tamil , hindi and english languages ?
- Maximum ,minimum and average population of cities.
- Which city has most literacy rate in India?

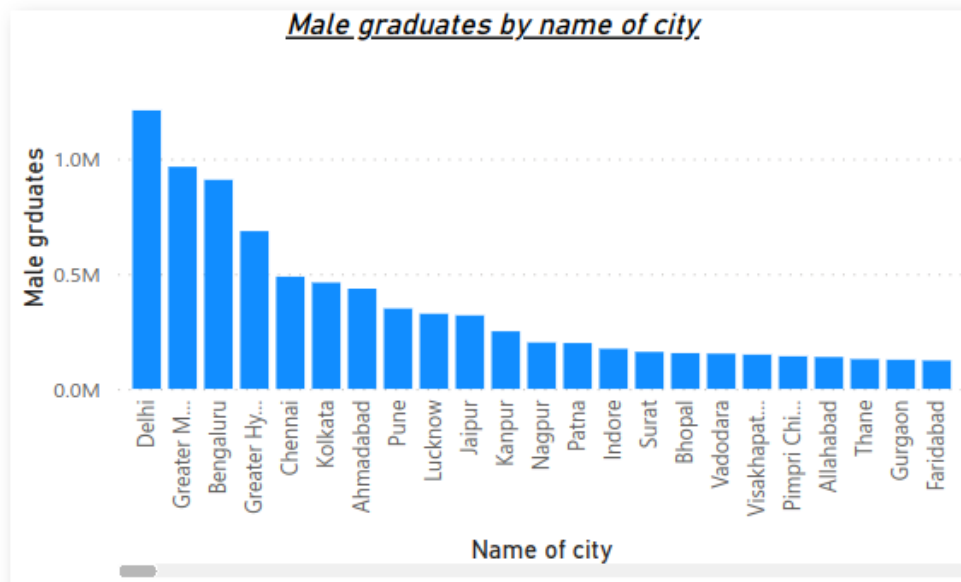


Figure : 3.1 Male Graduates by City

From the above bar chart, it can infer that the population of male graduates in each city. The Delhi city has the most male graduates around 1210040 and Mumbai has 964964 it is in second position.

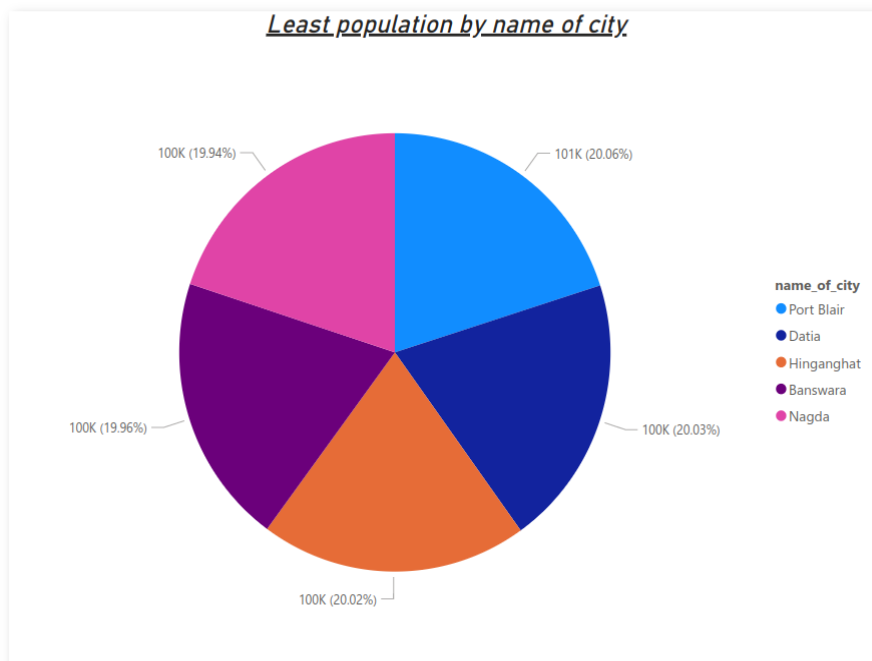


Figure : 3.2 Least population by City

The above Bar chart infer that the information about last 5 least population by city. Port blair, Datia, Hinganghat, Banswara and Nagda are the least populated cities in India which has population less than 100k.



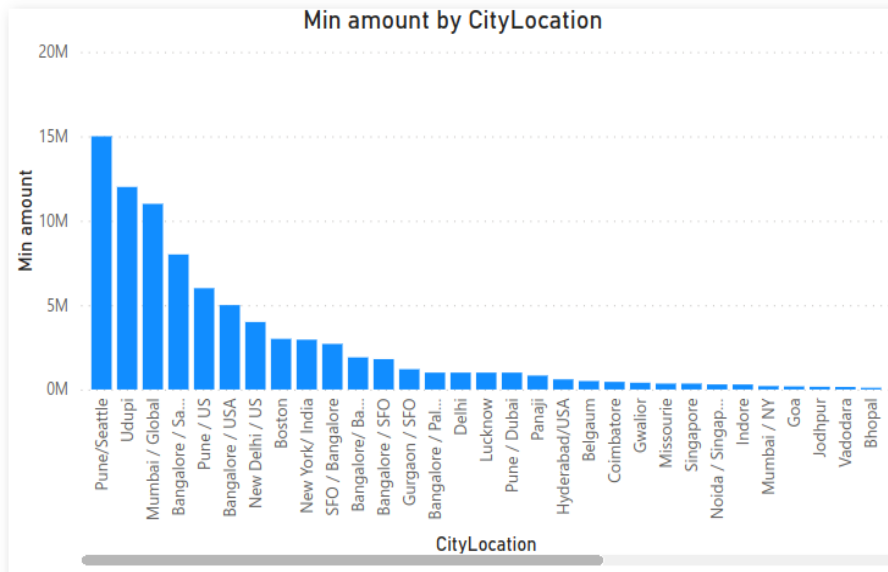


Figure : 3.3 Minimum amount invested in startup by City

From the above bar chart the minimum investment amount in each city is analysed. In that Pune has the largest least amount invested for startups.

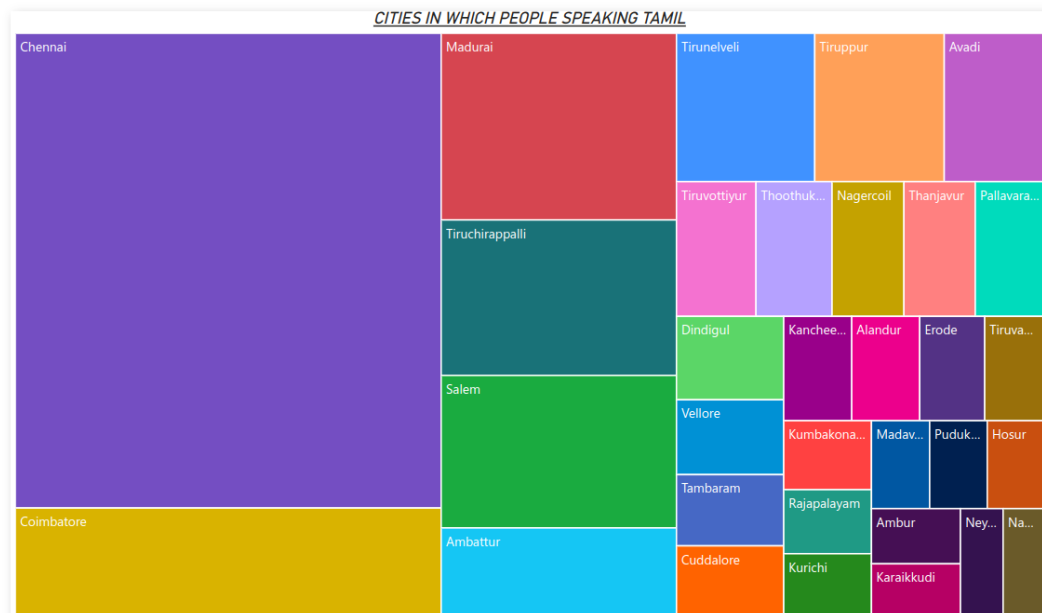


Figure : 3.4 Cities of Tamil speaking people

The above Tree chart infer the information about the cities in which people speak tamil language. The Highest number of people speaking tamil language is in Chennai and then Coimbatore.

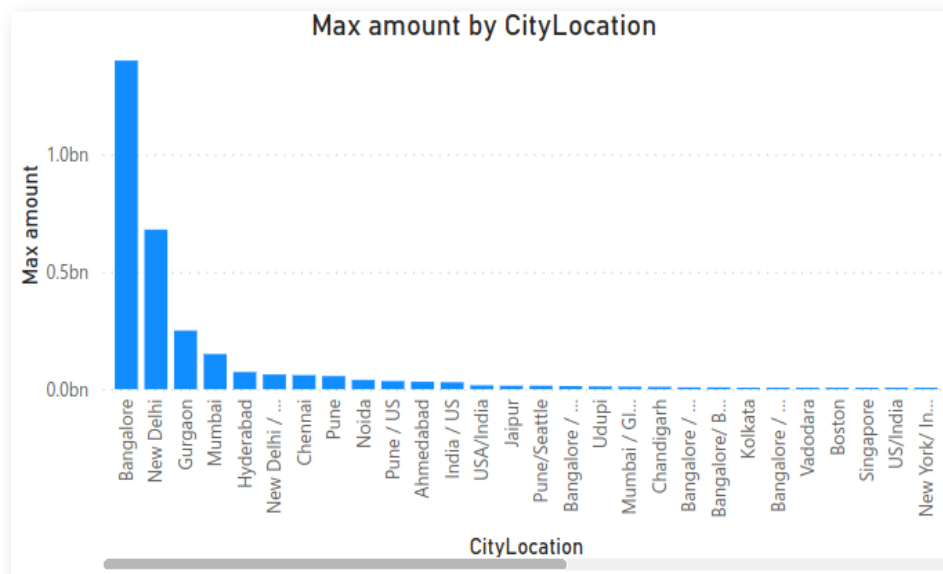


Figure : 3.5 Maximum amount invested fro startups by City

From the above Bar chart it can infer that maximum investment amount of each city. Bangalore has the largest amount around 1.4 billion invested in its start up companies. And next to Bangalore New Delhi has the second largest amount of around 0.7 billion.

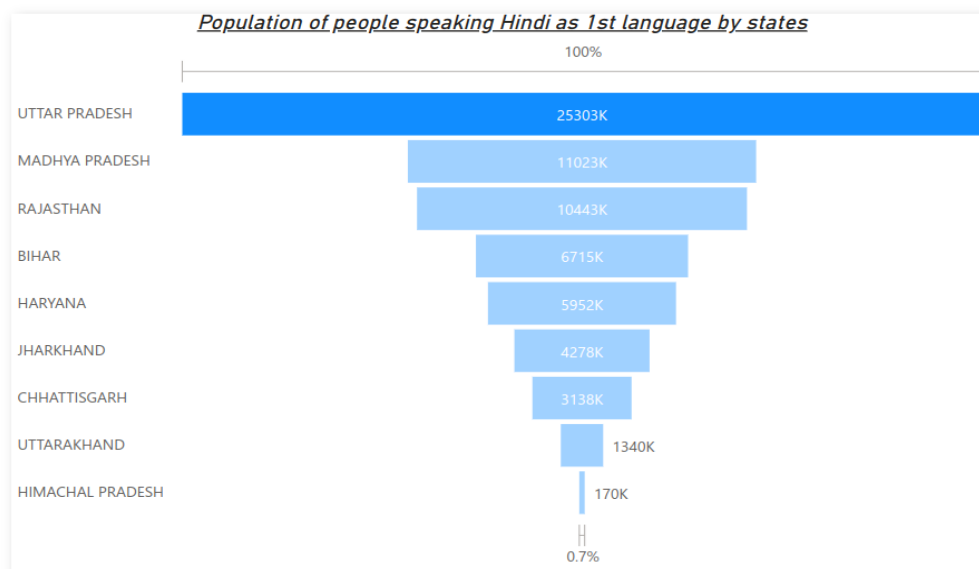


Figure : 3.6 Population of people speaking Hindi as 1<sup>st</sup> Language

The above chart represent the analysis on Population of people speaking Hindi as 1<sup>st</sup> Language. Uttar Pradesh has the most number of people around 25303k who are all speaking Hindi and followed by Madhya Pradesh which has nearly 11023k people speaking Hindi.

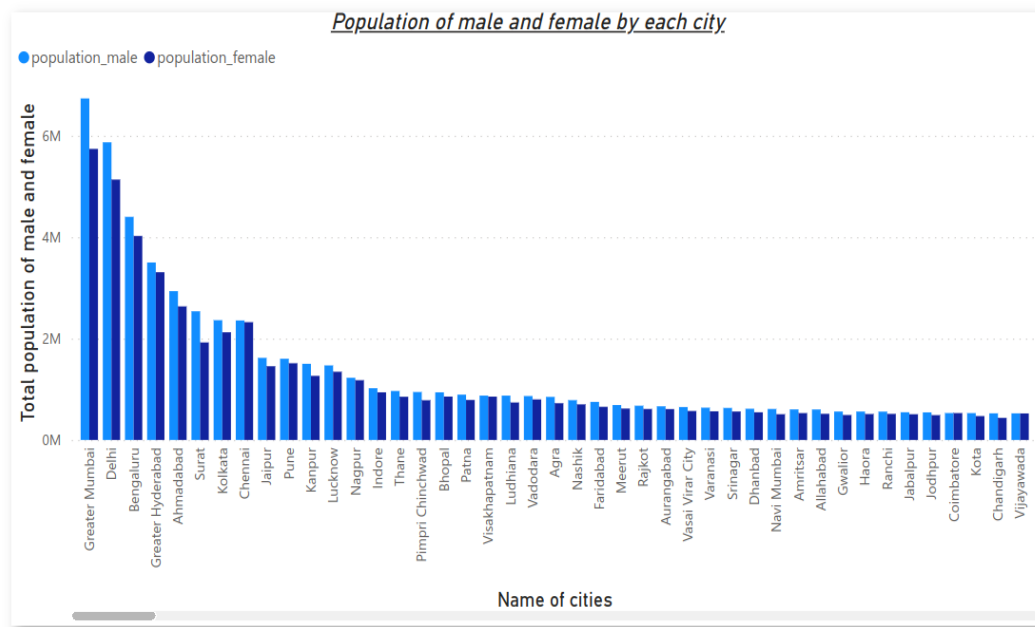


Figure : 3.7 Population of male and Female By City

The above chart represent the analysis on Population of male and Female of each City. The Mumbai has around 6736815 males and 5741632 female population and it is followed by Delhi which has 5871362 males and 5136473 female population.

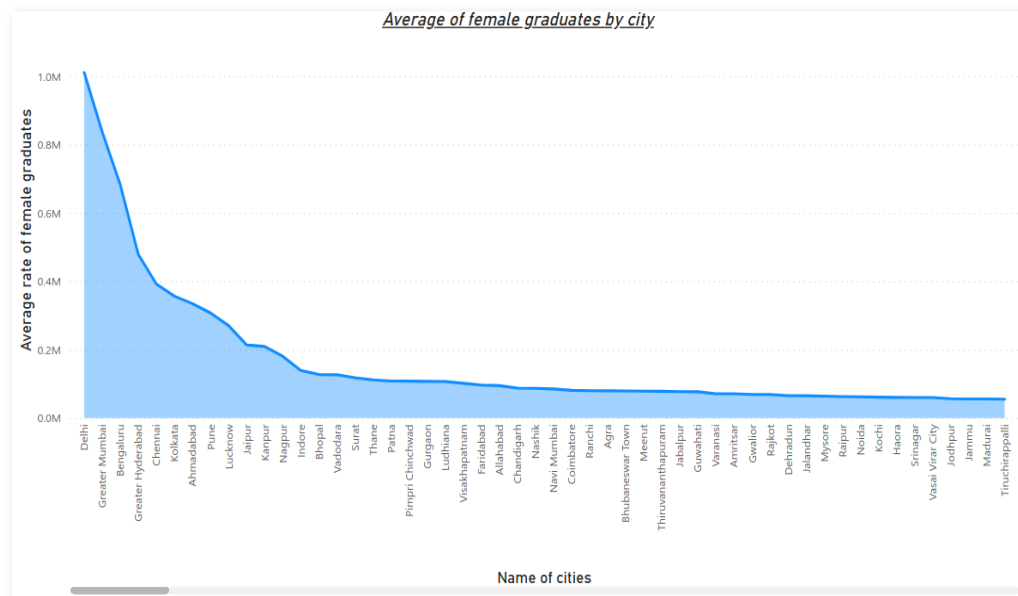


Figure : 3.8 Average of Female graduates by city

From the above area chart, the analysis of Average of Female graduates of each city. The average graduates female graduates are present in Delhi nearly 1 million and followed by Mumbai has nearly 0.8 million.

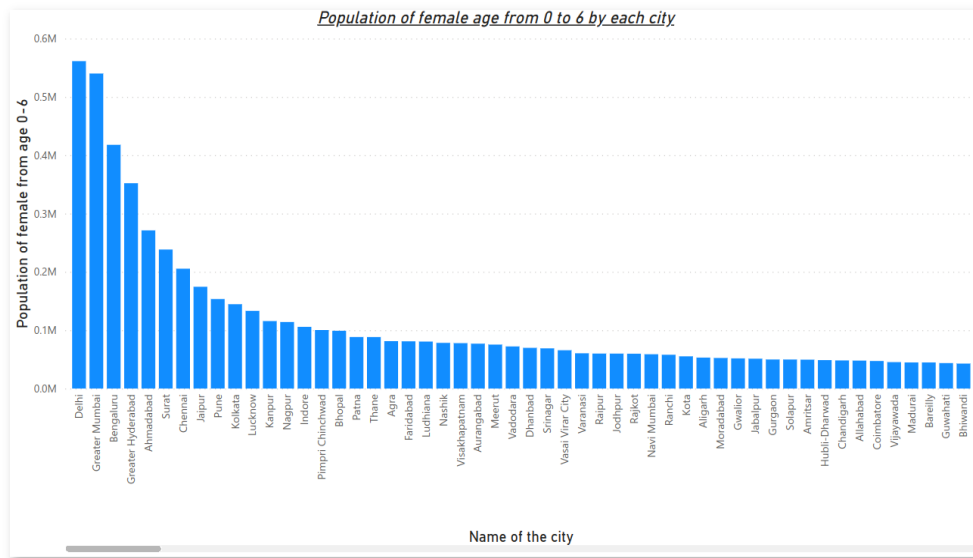


Figure : 3.9 Population of Female age from 0 to 6 by City

This chart represents about the female population from age 0 to 6, The female children population form age 0 to 6 are higher in Delhi(561337) and second highest in Mumbai(540139).

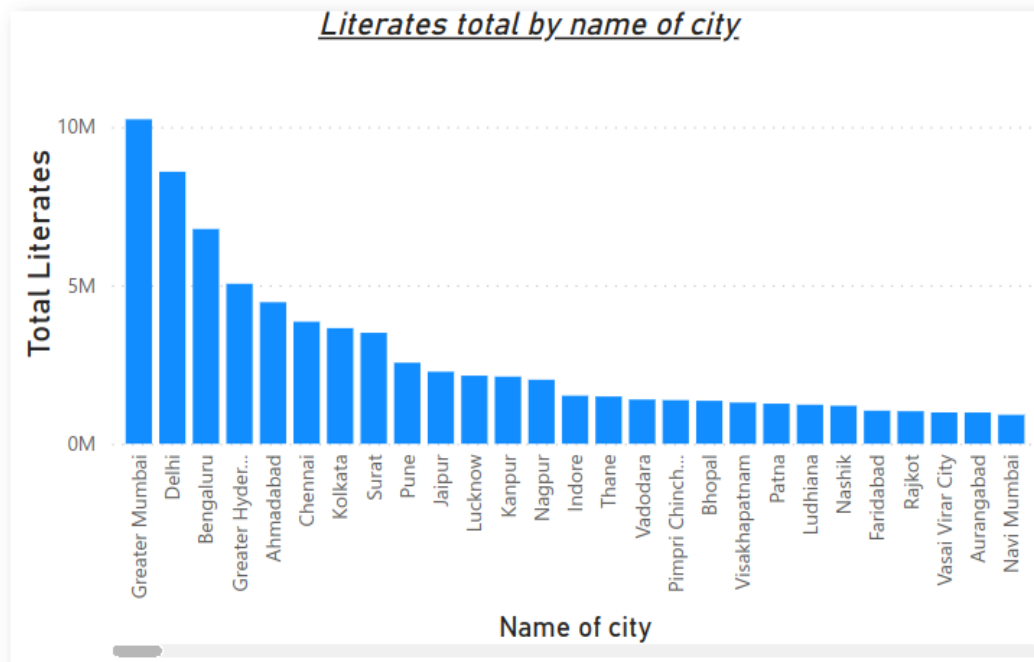


Figure : 3.10 Literates total by City

The above chart is about total literacy rate, in which mumbai has highest number of lietrates around 10 million literates and followed by delhi.

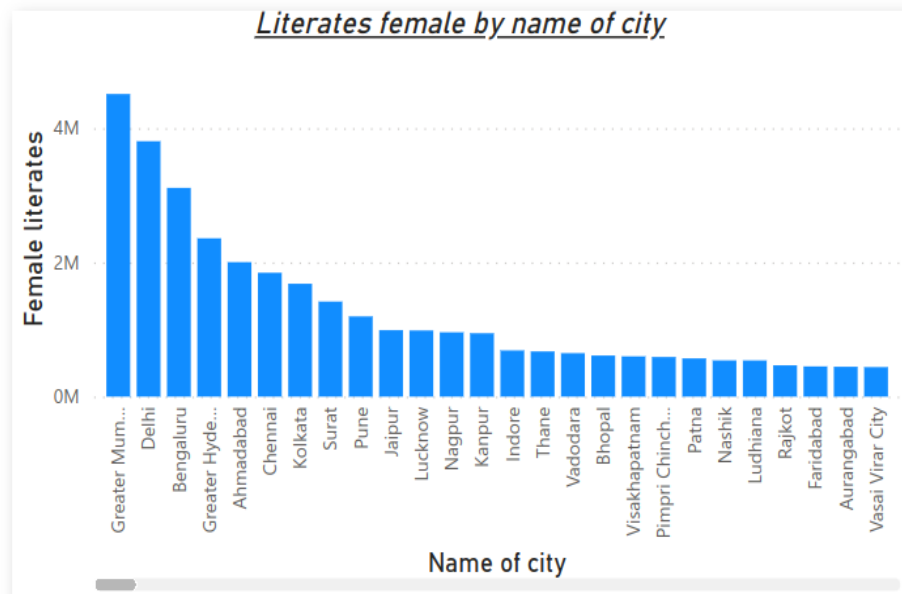


Figure : 3.11 Literates Female by City

The above chart explains about female literacy rate in which Mumbai has the most highest literacy rate when compared to other cities.

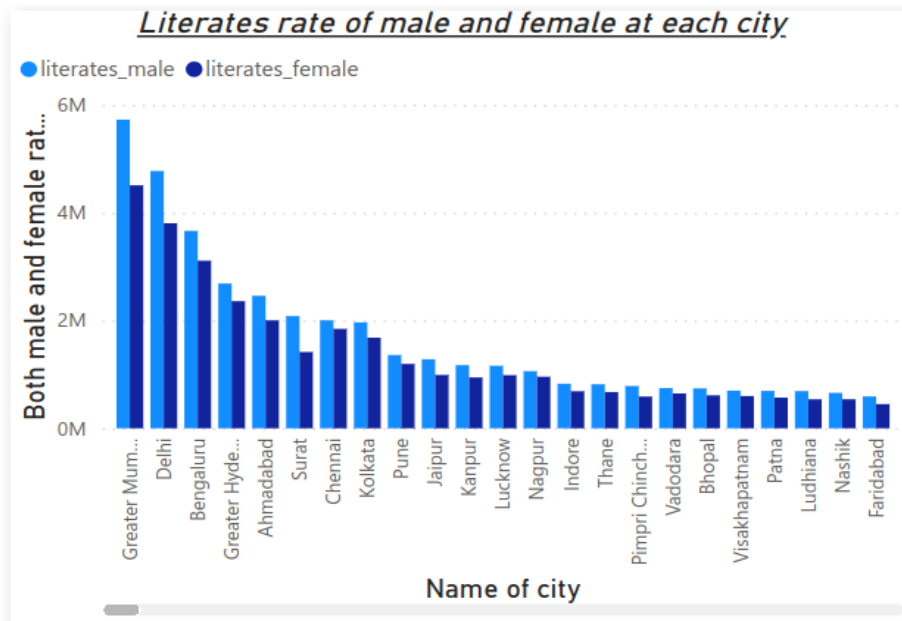


Figure : 3.12 Literates rateof Male and Female by City

The above chart explains about male and female literacy rate , compared to other cities mumbai has the most male and female literacy rate.

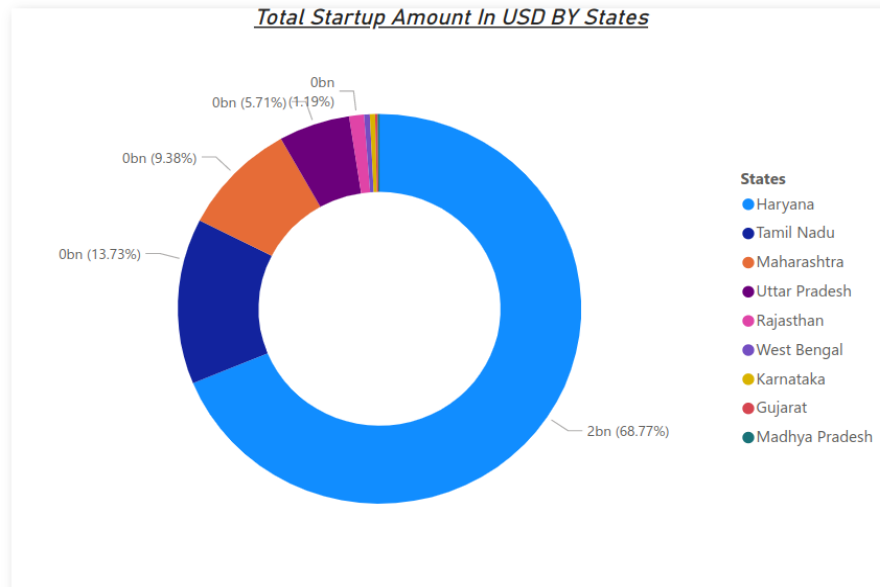


Figure : 3.13 Total Startup Amount in USD by Sates

This chart is about amount invested in the start up compaines , in that hayana has the most amount invested compared to other states.

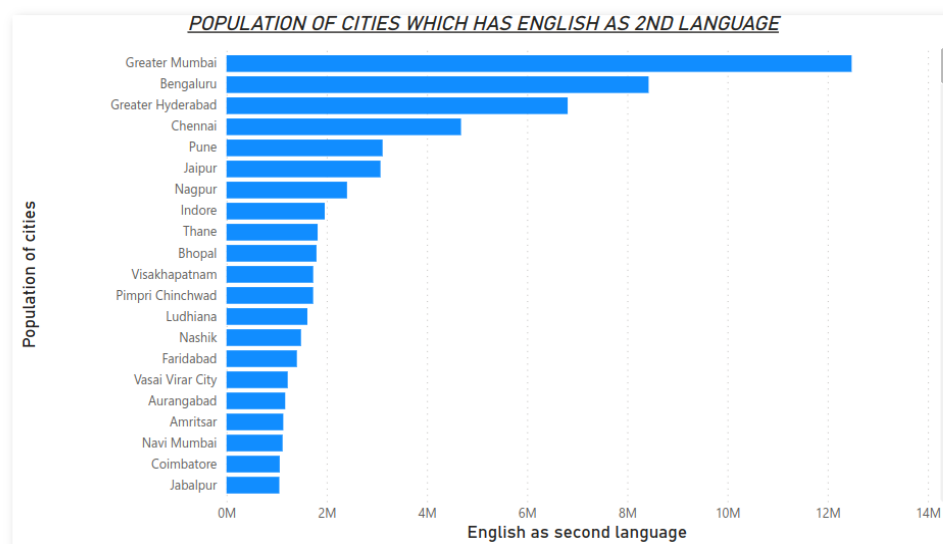


Figure : 3.14 Population of Cities which has English as 2<sup>nd</sup> Language

This chart explains about second language English , Mumbai has around 13 million people who speak English as the second language.

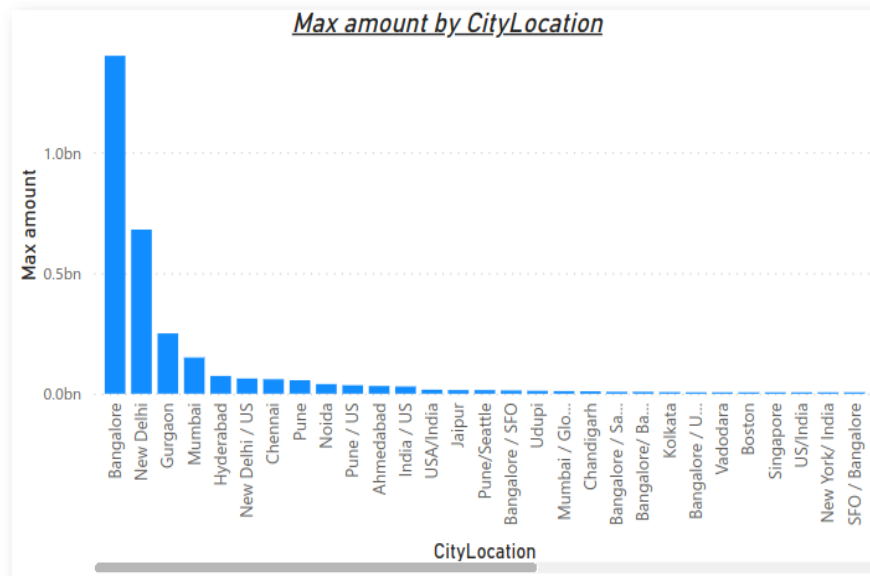


Figure : 3.15 Max amount by City

This chart is about amount invested in the start up compaines , in that Bangalore has the most amount invested compared to other cities.

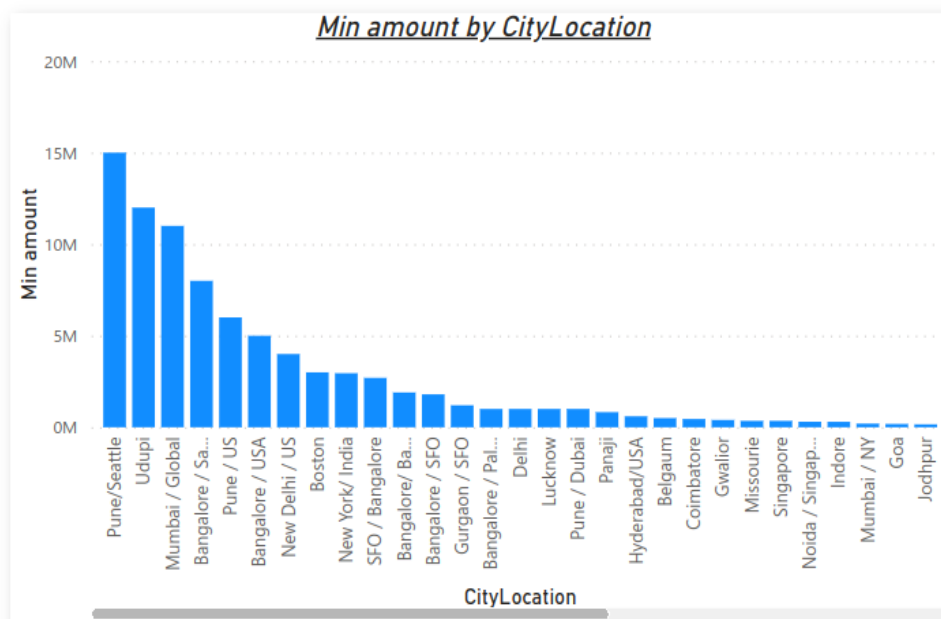


Figure : 3.16 Minimum amount By City

This chart is about amount invested in the start up compaines , in that Pune has the most amount invested compared to other cities which is about 15 Million.

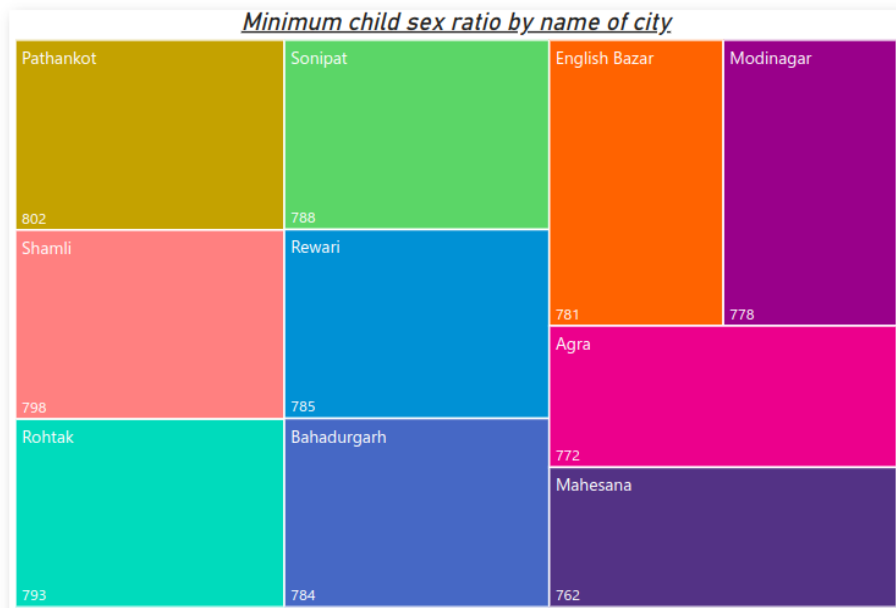


Figure : 3.17 Minimum child sex ratio by city

This chart is about minimum child sex ratio, Pathankot has the least child sex ratio of 762 and then shamili has 798.

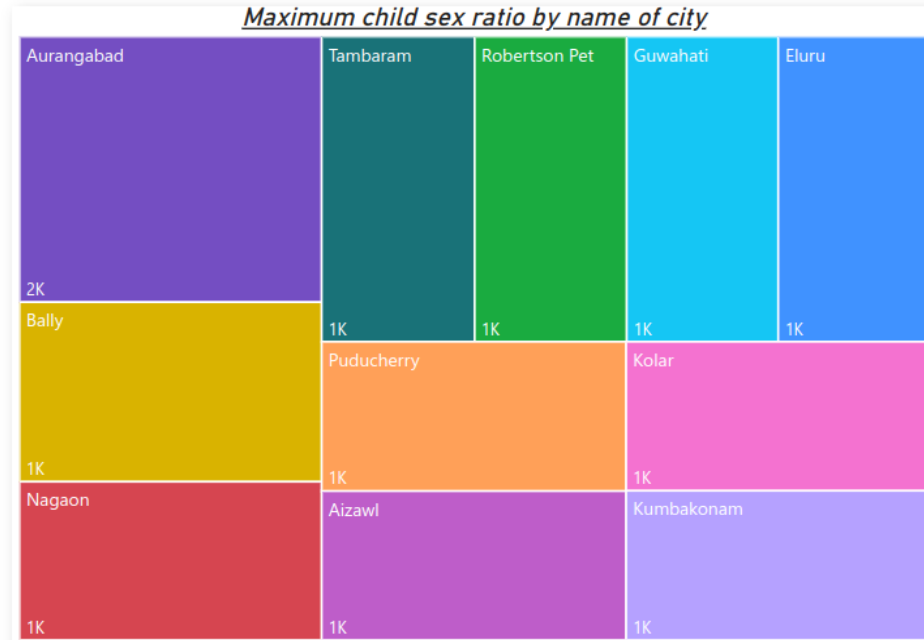


Figure : 3.18 Maximum child sex ratio by City

This chart is about child Maximum sex ratio, in that Aurangabad has the highest child sex ratio of 1185 and followed by Bally it has 1043.





Figure : 3.19 Minimum population of City

The measure,  $\text{MIN\_POL\_CITY} = \text{MIN}(\text{Indian\_cities}[\text{population\_total}])$  is used to create the above chart which gives the information about the minimum populated city, the population is equal to 100k.

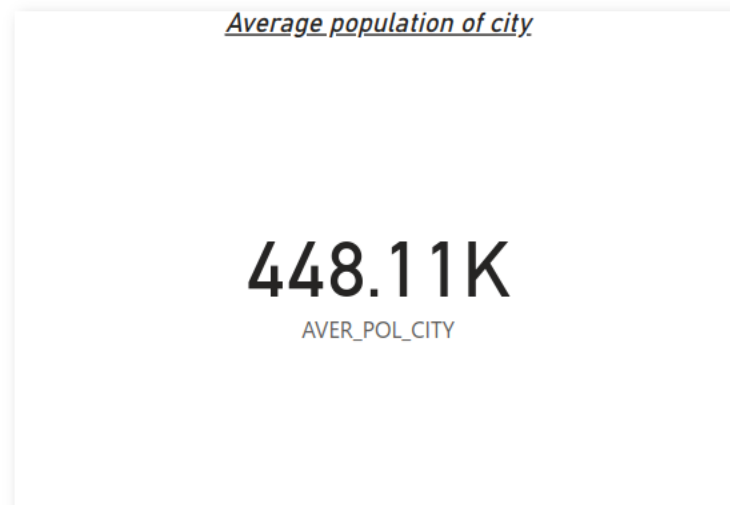


Figure : 3.20 Average population of city

The measure,  $\text{AVER\_POL\_CITY} = \text{AVERAGE}(\text{Indian\_cities}[\text{population\_total}])$  is used to create the above chart. This chart gives the insight about the average population of all cities and the population is equal to 488.11k.

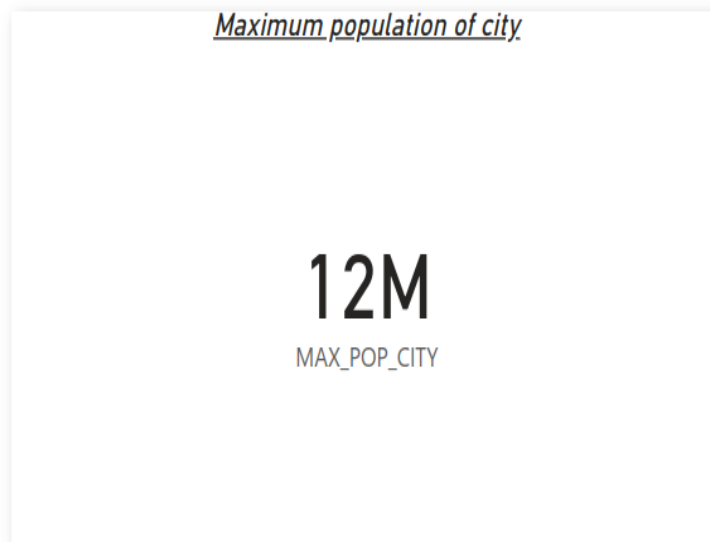


Figure : 3.21 Maximum Population of City

The measure, MAX\_POP\_CITY =  $\text{MAX}(\text{Indian\_cities}[\text{population\_total}])$  is used to create the above chart. This chart gives the insight about the maximum populated city and the population is equal to 12Million.

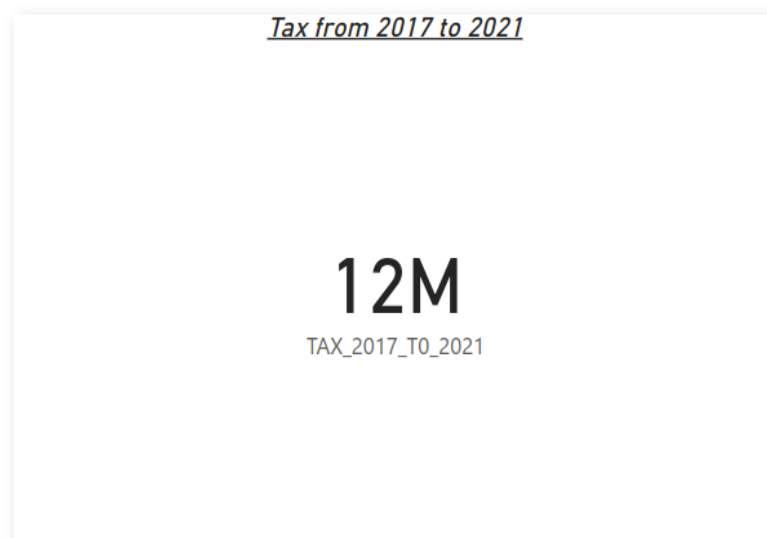


Figure : 3.22 Tax from 2017 to 2021

The measure, TAX\_2017\_T0\_2021 =  $\text{SUMX}(\text{Tax\_2012\_21}, \text{Tax\_2012\_21}[2016-17] + \text{Tax\_2012\_21}[2017-18] + \text{Tax\_2012\_21}[2018-19(\text{Accounts})] + \text{Tax\_2012\_21}[2019-20(\text{RE})] + \text{Tax\_2012\_21}[2020-21(\text{BE})])$  is used to create the above chart which gives the Sum of total Tax from 2017 to 2021 and is equal to 12 Million.

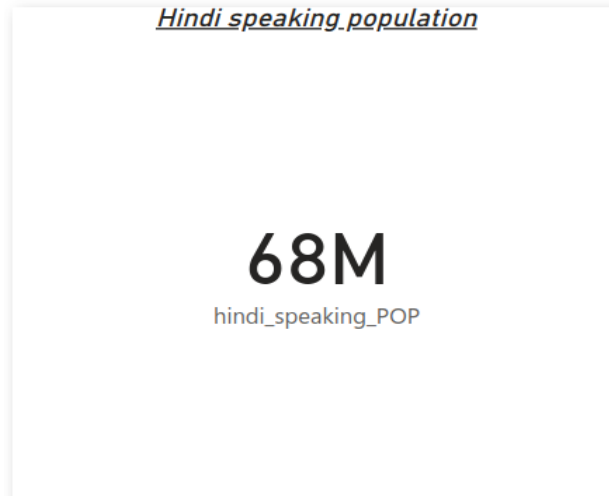


Figure : 3.23 Hindi speaking Population

The measure, hindi\_speaking\_POP = `CALCULATE(SUM(Indian_cities[population_total]), FILTER(States_Languages,States_Languages[hindlan]="Yes"))` which is used to create the above chart. This chart gives the information about the Hindi speaking population which is equal to 68 Million.



Figure : 3.24 Tamilnadu Tax from 2017 to 2021

The measure, TN\_TAX\_2017\_T0\_2021 = `CALCULATE(SUMX(Tax_2012_21,Tax_2012_21[2016-17] + Tax_2012_21[2017-18]+Tax_2012_21[2018-19 (Accounts)] + Tax_2012_21[2019-20 (RE)] + Tax_2012_21[2020-21 (BE)]),FILTER(Tax_2012_21,Tax_2012_21[state_name]="Tamil Nadu"))` is used to create the above chart which gives the information about the tax details from 2017 to 2021 in Tamilnadu which is equal to 543k.



Figure : 3.25 People speaking English as second Language

The measure, `ENG_AS_SEC_LAN` = `CALCULATE(SUM(Indian_cities[population_total]),FILTER(States_Languages,States_Languages[Secondary_languages]="English"))`. This measure is used to create the above chart which gives the analysis of People speaking English as second Language which is almost equal to 123 Million.



Figure : 3.26 Tamil language speaking Population

The measure, `tam_lanSpeaking_Pop` = `CALCULATE(SUM(Indian_cities[population_total]),FILTER(States_Languages,States_Languages[Languages]="Tamil"))` which is used to create the above chart and gives the Tamil language speaking population which is equal to 14 Million.

## 3.2 PUBLISHING DASHBOARD

### INDIAN CITIES ANALYSIS

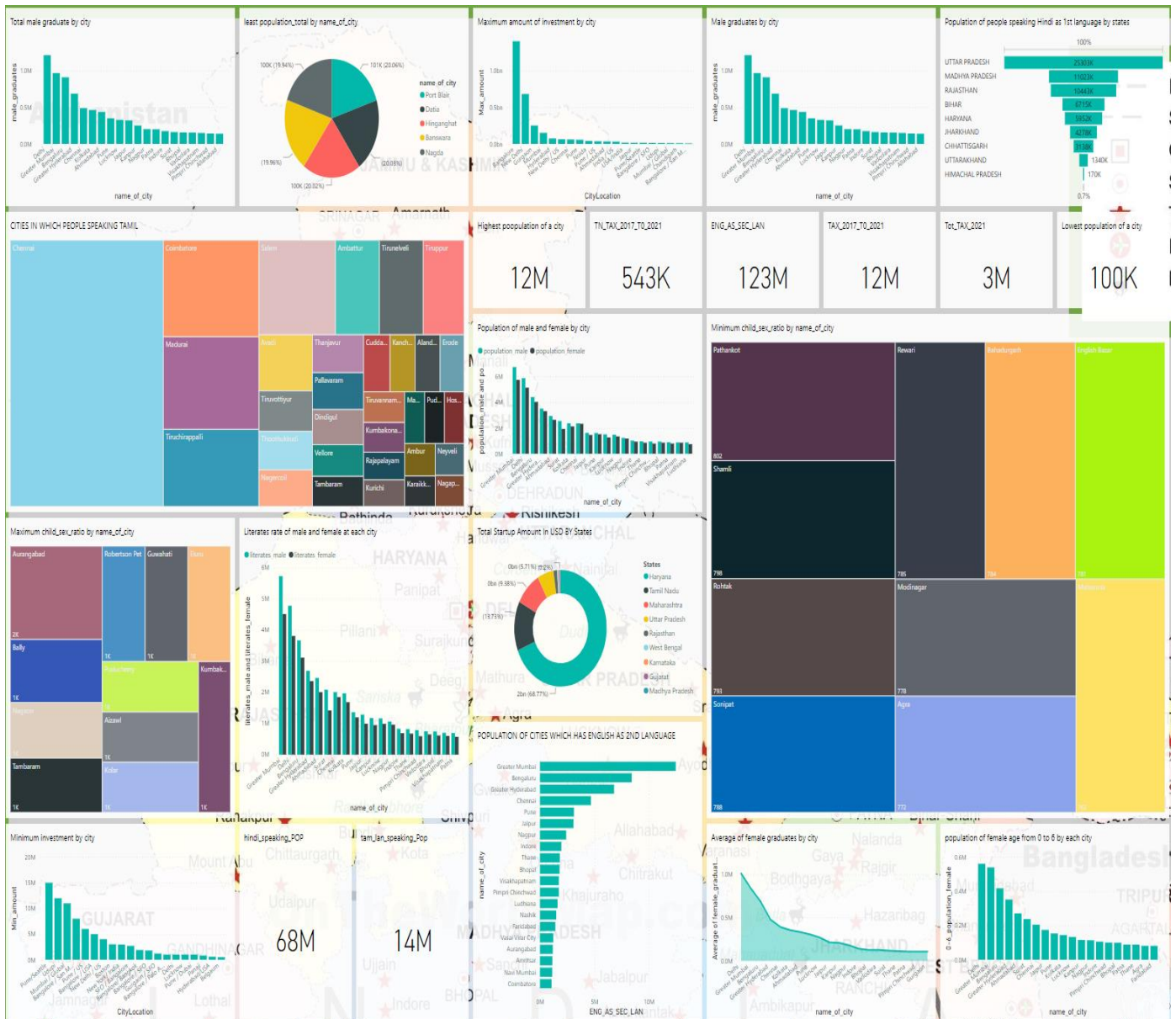


Figure : 3.1 Indian Cities Analysis Dashboard

### 3.3 INFERENCES

- The Delhi city has the most male graduates around 1210040 and Mumbai has 964964 it is in second position.
- Port blair, Datia, Hinganghat, Banswara and Nagda are the least populated cities in India which has population less than 100k.
- Mumbai has around 6736815 males and 5741632 female population and it is followed by Delhi which has 5871362 males and 5136473 female population.
- Delhi has the most child population around 1209275 and It is followed by Mumbai it has an child population of 1139146 .
- Highest number of people speaking tamil language is in Chennai and then Coimbatore.
- The female children population form age 0 to 6 are higher in Delhi(561337) and second highest in Mumbai(540139).
- The most graduates female graduates are present in Delhi and followed by Mumbai.
- Uttar Pradesh has the most number of people around 25303k who are all speaking Hindi and followed by Madhya Pradesh.
- Mumbai has around 13 million people who speak English as the second language.
- Bangalore has the largest amount around 1.4bn invested in its start up companies.
- Pune has the least amount 15million invested in its start up companies.
- Total invested amount in startup companies are given by states, in that Haryana has 2bn and it is the highest in India.
- Literacy of both male and women in each cities, in that Mumbai has the highest no of literacy rate in that 5727774 are males and 4509812 are female.
- Aurangabad has the highest child sex ratio of 1185 and followed by Bally it has 1043.
- Pathankot has the least child sex ratio of 762 and then shamili.
- All the state has paid tax form 2017 to 2021 is 12 million ,in that Tamil Nadu has paid 543K tax amount form 2017 to 2021.

- Mumbai has the most literacy rate and it is followed by Delhi.
- Around 123M people speak English as second language .
- Nearly 68M people are speaking Hindi.
- And 14M people are speaking Tamil.
- The Maximum population in all cities is 12M
- The Average population of Indian cities is 448.11k
- And the Minimum population of all cities is 100k.

## **CHAPTER 4**

### **CONCLUSION AND FUTURE WORK**

#### **4.1 RECOMMENDATIONS**

The Indian cities analysis has been done using Power BI tool and various insights are performed and visualised. It can identify the least and highest population in cities and it can be used for business objective. The analysis of population will help in many ways like how many buses are needed for public transportation?, how much food supplement is needed when in flood or cyclone period.

The analysis of graduates in cities will help for big companies to make their branch office. The visualization of Indian cities in Power BI charts will be helpful for teaching or education purpose, because it is more effective by studying in diagram than of theory. And it will be helpful for children they can easily remember like which city has most population or which city has least population? Etc.,

It will be helpful for many foreign countries to make their business in cities, like if it is a school they need more population and if it is a chemical factory they need less population.

Using this it can be able to understand about Indian cities for population, male population, female population, male population, s-x-ratio, child-sex-ratio, literacy-rate, male-literacy-rate, female-literacy-rate, tax etc..



## REFERENCES

- [1] Anuj Tiwari and Dr. Kamal Jain, “GIS Steering Smart Future for Smart Indian Cities.” International Journal of Scientific and Research Publications, Volume 4, Issue 8, August 2014.
- [2] Sejal S. Bhagat, Palak S. Shah and Manoj L. Patel, “Smart cities in context to Urban Development.” International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development, Volume 4, Issue 1, February 2014, 41-48.
- [3] Charbel Aoun, “The Smart city Cornerstone: Urban Efficiency.”Schneider Electric White Paper, 2013.
- [4] R. R. Widner, “Physical Renewal of the Industrial City,” Annals of the American Academy of Political and Social Science, vol. 488, pp. 47–57, Nov. 1986.
- [5] Wikipedia contributors, “Adaptive reuse,” Wikipedia, the free encyclopedia. Wikimedia Foundation,Inc., 21-May-2012.
- [6] Ellie Cosgrave and Theo Tryfonas, “Exploring the Relationship between Smart City Policy and Implementation.”The First International Conference on Smart Systems, Devices and Technologies, 2012.”
- [7] TERI (2002): “Performance Measurement of Pilot Cities” Tata Energy Research Institute, New Delhi, India.
- [8] W. H. Frey, “Black In-Migration, White Flight, and the Changing Economic Base of the Central City,” American Journal of Sociology, vol. 85, no. 6, pp. 1396–1417, May 1980.