Final Assignment – Population in Tamil Nadu

Introduction / Business Problem

This project is divided into 3 phases

- Identifying the city with the highest population
- Create a model to predict the population in this particular city
- Use Foursquare location data to search for a few places in this city

This project is majorly aimed at building a model to predict the population of a particular city based on parameters such as:

- Birth Rate
- Death Rate
- Migration into Chennai

These parameters are most vital when determining the population of a particular place. This project will be helpful in approximating the census of a particular place. Census takes places every 10 years in India. The next census is due to take place next year(in the year 2021).

Target Audience

The government and the people involved in determining the census will benefit from this project.

Data

- Population of different districts in Tamil Nadu according to the latest census.
  Using this data we plot a bar graph and find out the district having maximum population. We find out that the Chennai district has the highest population.
  Example : This data consists of population in different districts of Tamil Nadu such as Ariyalur, Kancheepuram, Thiruvallur, Chennai, etc.
- Latitude and Longitude of Tamil Nadu(used for mapping)
  Using this data, we visualise the map of Tamil Nadu and see how the different districts of Tamil Nadu are spread out.
  Example : Latitude and Longitude of Tamil Nadu are 10.9094344 and 78.3665347
- Population in the Chennai district for the past Century(1911-2011)

This is the main data that will be used to build a model in order to predict the population of Chennai.
Example: Population recorded in each census that is in 1911,1921,1931,…….,2011 is available in this dataset.

- Birth Rates, Death Rates and Migration-in data for Chennai from 1911-2011
  These are the parameters that will be used to predict the population of Chennai.
  Example: Birth Rate, Death Rate and Migration Data recorded in each census that is in 1911,1921,1931,…….,2011 is available in this dataset.
- Foursquare location data to search for places in Chennai.
  We find out the names, categories and location data of 30 places in Chennai.
  Example: Details of a particular place in Chennai
  Name: Chennai Fort Railway Station
  Category: Train Station
  Latitude: 13.083362
  Longitude: 80.283049

Methodology

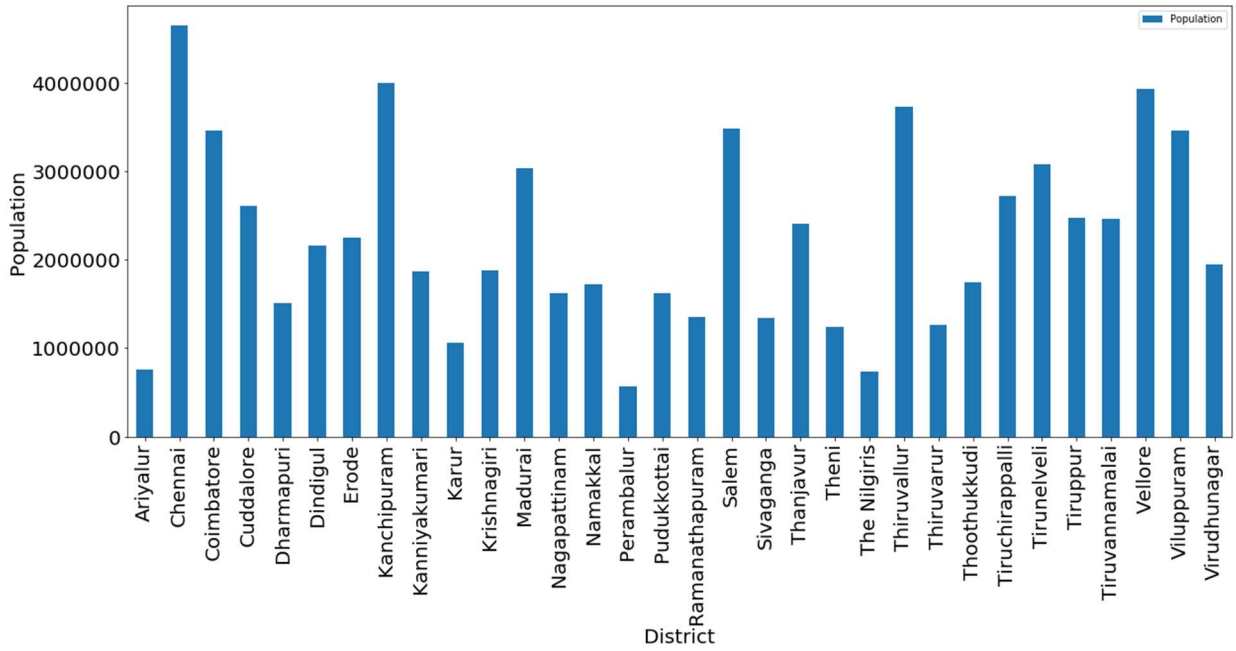
Identifying the city with the highest population

Scrape the following page 'https://www.citypopulation.de/php/india-tamilnadu.php'.

This page contains the population data for every district in Tamil Nadu during the 1991,2001 and 2011 census. We require the population only for the 2011 census for now since we are looking for the most populous district in Tamil Nadu.

Cleaning Data : Rename the column containing population count. Remove the row containing the total population count for Tamil Nadu. Also some districts contain multiple names. Remove the names in brackets.

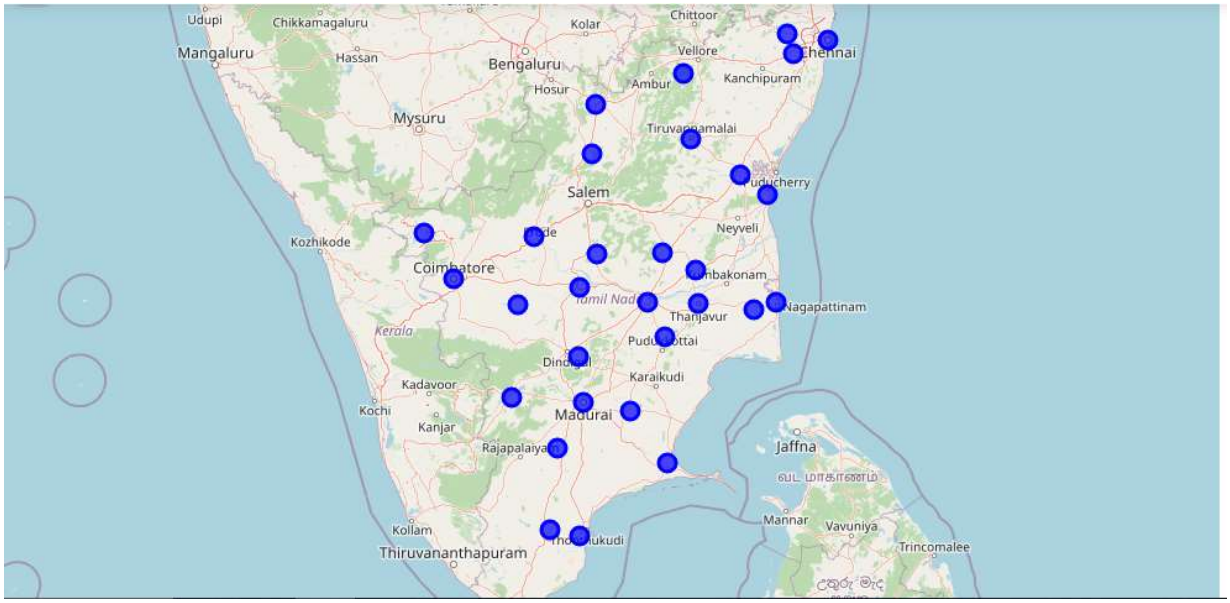
Plot a bar chart and identify the most populous district using matplotlib.



We can clearly see that the ‘Chennai’ district has the highest population.

Using geopy.geocoders, let us visualise the map of Tamil Nadu and see how the different districts are spread out.

Using the above library, the location data for Tamil Nadu and for each of its districts can be acquired by passing the address accordingly. A for loop is used for this purpose.



On hovering over the markers, we get the name of that particular district. We can see that the districts of ‘Thiruvallur’ and ‘Kanchipuram’ are also very close to the most populous district ‘Chennai’.

Create a model to predict the population in Chennai

Let us now look for population data in Chennai from the past census (1911-2011).

Scrape 'https://en.wikipedia.org/wiki/Demographics_of_Chennai#cite_note-6' to get the population count for Chennai in the past census

Clean the data: Remove the unwanted columns and rows. Rename the column ‘Census’ as ‘Year’.

Load the birth and death rates from the 1901 till the 1991 census using the following csv file : 'birth&death.csv'. This csv file is available in my Github Repository.

Link to my repository : https://github.com/varshapraburam/Coursera_Capstone

Clean the data. Change the format of the column ‘Year’.

	Year	Birth Rate	Death Rate
0	1901-11	34.0	38.9
1	1911-21	33.6	37.5
2	1921-31	34.0	33.9
3	1931-41	36.5	30.7
4	1941-51	32.8	25.5
5	1951-61	41.3	23.6
6	1961-71	38.6	12.3
7	1971-81	31.5	10.2
8	1981-91	25.7	9.4

	Year	Birth Rate	Death Rate
0	1911	34.0	38.9
1	1921	33.6	37.5
2	1931	34.0	33.9
3	1941	36.5	30.7
4	1951	32.8	25.5
5	1961	41.3	23.6
6	1971	38.6	12.3
7	1981	31.5	10.2
8	1991	25.7	9.4

Load another dataframe ‘birth&death(2001).csv’ containing birth and death rates for each year from 1991-2001.

Take mean of all the data to get the average birth and death rates during the 2001 census.

Append the data to the above dataframe.

Get the birth and death rate data for the 2011 census from the following websites.

http://www.spc.tn.gov.in/DHDR/Chennai.pdf

https://censusindia.gov.in/vital_statistics/SRS_Report/12SRS%20Statistical%20Report%20Table%20-%2020111.pdf

Join the columns birth and death rates to the dataframe containing the population of ‘Chennai’.

Scrape the html file 'chapter 4.htm' to get the migration data for Chennai city between 1911-2011.

Since 2011 does not have a migration data, we take the mean of the migration data of all the other years as the value for 2011.

Predicting the Model

Firstly split the whole dataset into train and test datasets.

‘Birth Rate’, ‘Death Rate’ and ‘Migration into Chennai’ are the values of x.

‘Population’ is the y value.

Fit the training data using **multiple regression**. Predict the data using the testing data(x_test).

Find the accuracy score using variance(regr.score). The variance is found to be 0.70.

Print the coefficients and intercept for this model.

Using Foursquare location API

Load the client_id and client_secret of your Foursquare account.

Type out the url to view a few places in Chennai. Set the radius as 19500(radius of Chennai district).

url : `'https://api.foursquare.com/v2/venues/search?client_id=3FME4APJ21KKXL3TEFBUF2PF5VE2IOTSS3VEPUSL33CW24EG&client_secret=LVY3BYRQDCDZC5SSBSVRNUFMW2IL1IA4WU3QKBVVUE2WRSFG&ll=13.0801721,80.2838331&v=20180604&radius=19500'`

Get a json file contain details of places in Chennai. Use **json_normalize** to view it in a dataframe.

Results

Most populous district in Tamil Nadu: Chennai.
Variance of the multiple linear regression model is **0.70.**
R2 score is found to be 0.42.
We were able to view 30 places in Chennai along with its category and location data.

Discussions

The variance of the model is decent enough but the R2 score value is slightly les s.The model can be enhanced further to make sure that the R2 score is better. Having more parameters for evaluation such as number of people migrating out of Chennai can help make a better model.

Conclusion

In this project, the most populated district in Tamil Nadu has been evaluated usi ng a bar plot and is found to be ‘Chennai’.

Equation of the model :

```
6336312.63772044-43774.97211650234x-113134.85515103952y-2.5335652771489
94z

X= Birth Rate
Y=Death Rate
Z=Migration into Chennai
```

The **Foursquare Location** data was used to check out a few places in Chennai.