**PAIR PROJECT**

**Sampada Dhungana (207033)**

**Ezee Adhikary (207004)**

We the students of Kathmandu University School of Management currently pursuing Bachelors in Business Administration ( BBA) 7th semester were assigned with a pair project required for the completion of the course Evidence based Business Analytics under the guidance of our faculty Mr. Devindra Katuwal. For the project, we as a pair were required to perform visualization through various applications which we learnt during the course.

The operation of our project consisted majorly of four steps :

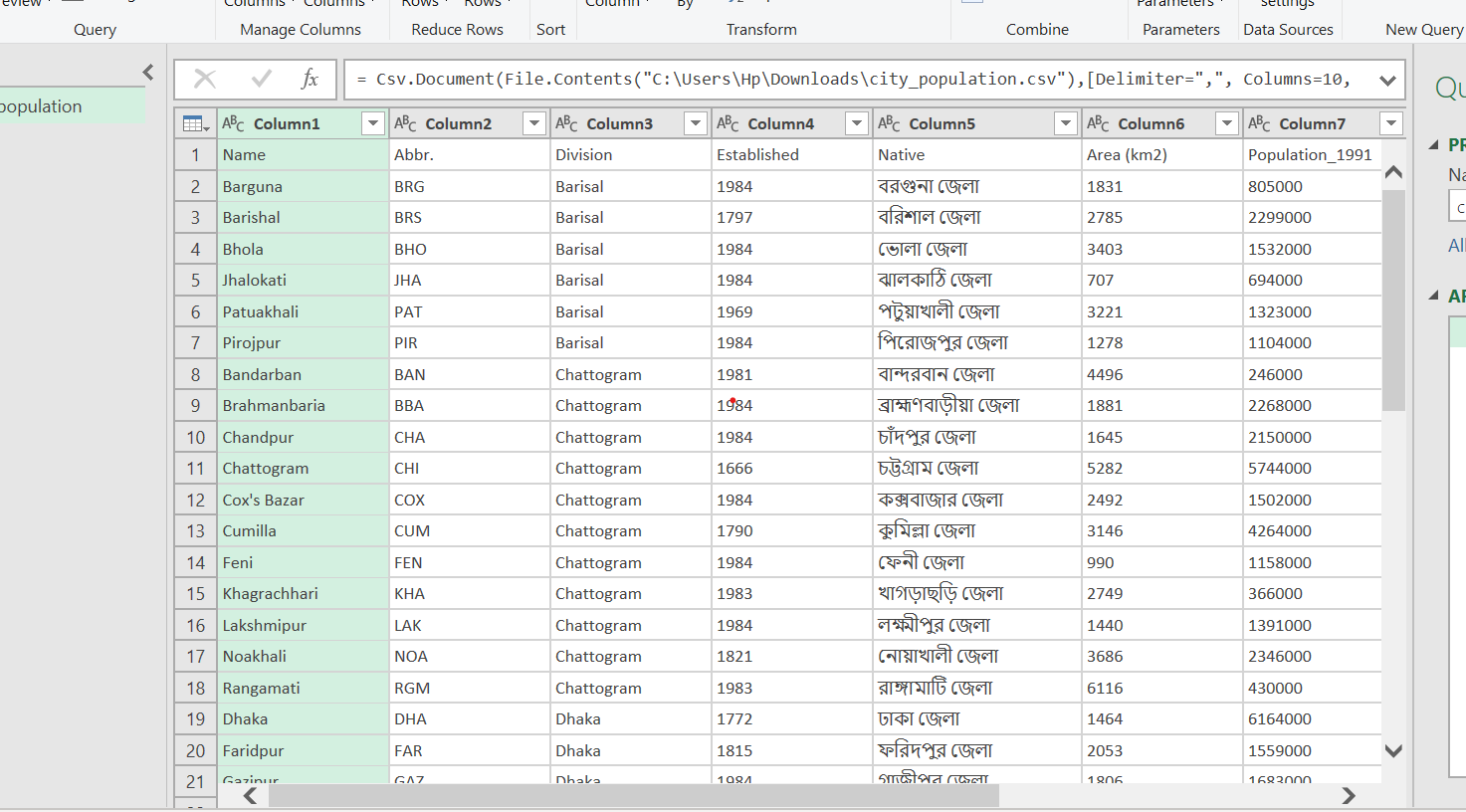
* Data collection
* Data cleaning
* Analysis and Visualization
* Publishing the work through any social media platform

1. **Data Collection**

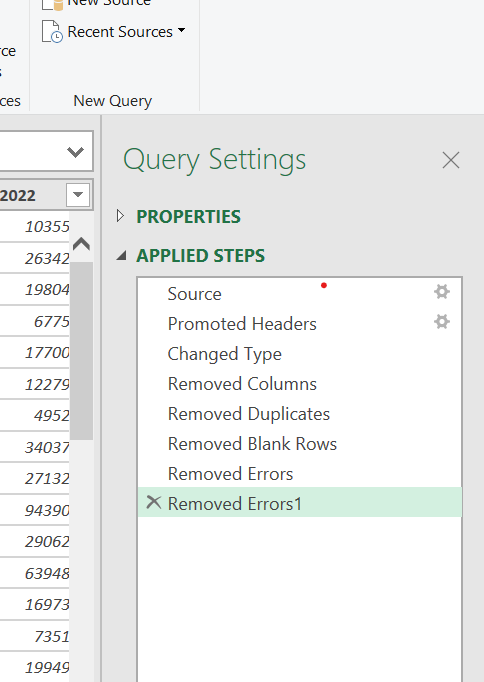
For the data collection, we collected our data through Kaggle which is the world’s largest data science community where we can find datasets related to any industry for analysis. Through the website we obtained data related to the city population of bangladesh. The data consists of multiple columns which comprise bangladesh city names along with their abbreviations, their division and the population of those cities in the year 1991, 2001, 2011 and 2022. The link for our data is given below:

<https://www.kaggle.com/datasets/msjahid/bangladesh-districts-wise-population>

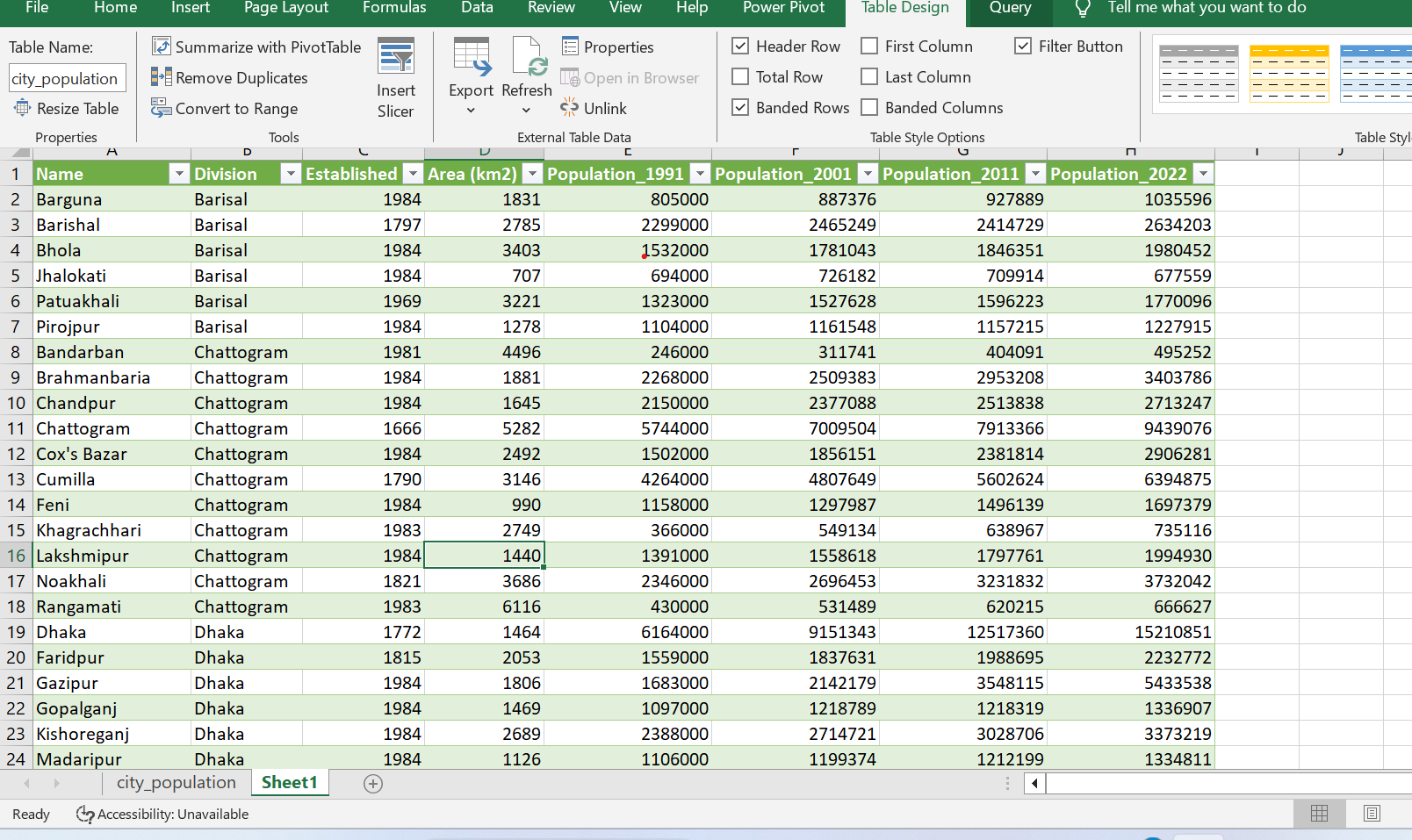
1. **Data Cleaning**

For data cleaning, we used the Power Query function of excel. Initially we downloaded the data in CSV form and when opened it was as the picture below : 

For the cleaning of data we followed multiple steps such as promoting the first row as headers, removing duplicates, removing unwanted rows, columns and alo errors in the rows. The steps can be seen in visual form below which is extracted from the excel sheet query.



After following the steps mentioned above we obtained a cleaned for of data which can be seen in the picture below :



1. **Analysis and Visualization**

For the analysis and visualization we used two applications Power BI and JupyterLab. Through the jupyter lab we performed Exploratory Data Analysis and through Power BI we tried to visualize the data and further beautify it. For the analysis we have set up our hypothesis as follow :

**Hypothesis**

H1 : Dhaka , capital of Bangladesh is the most populated city of the country.

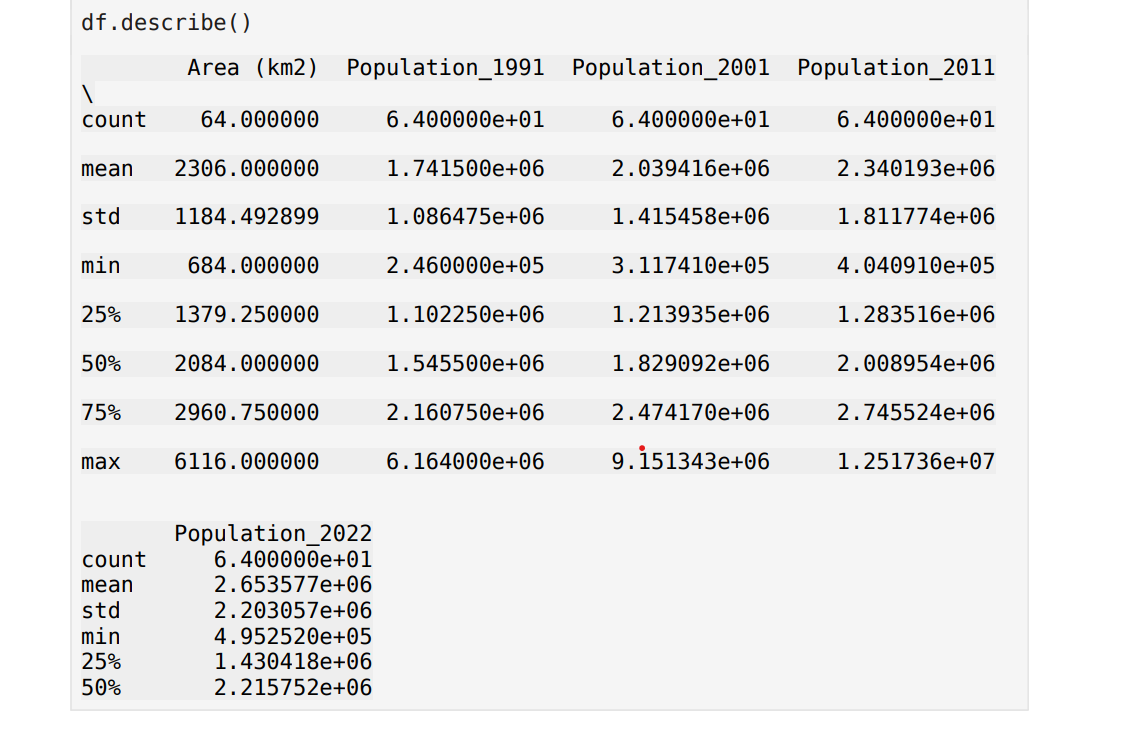
H0 : Dhaka, the capital of Bangladesh is not the most populated city of the country.

**Activities Performed in Jupyterlab**

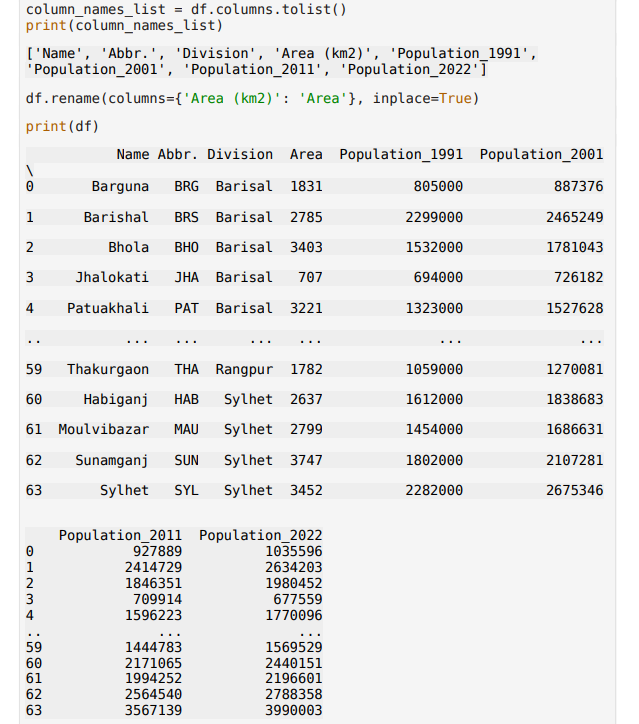
In jupyterlab, we performed activities such as renaming columns for easy analysis, creating histograms, scatter plots, testing hypotheses and calculating correlation between two variables using python tools . The analysis done is divided into two sections . First we have minor analysis and secondly we have major analysis.

* **Minor Analysis**

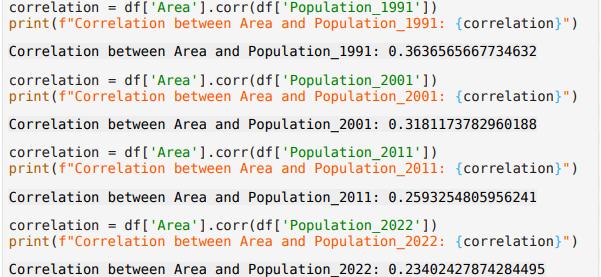
For the minor analysis, we have calculated the mean, standard deviation , the maximum values and also calculated the correlation between the area and population of the four different years. The initial analysis part also comprises miscellaneous works such as importing the python tools, displaying a certain number of rows and columns. renaming columns, printing them etc.



The picture above shows the mean , standard deviation, the minimum value along with 25% and 50% quartile population of the four years along with the area. The data above shows that the standard deviation of area in square km is 1184.492 which means that the area of cities of bangladesh deviate from the mean by 1184.92 square kilometers.



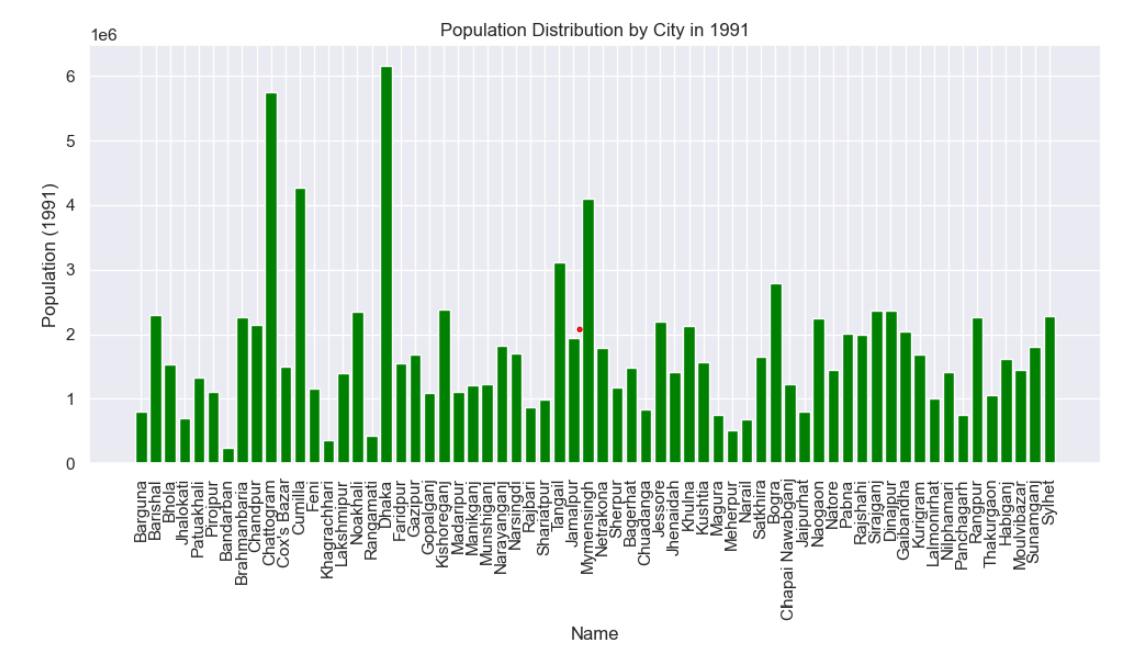
Similarly, the picture above shows the renaming and displaying of the renamed columns from Area (km2) to Area for easier process .

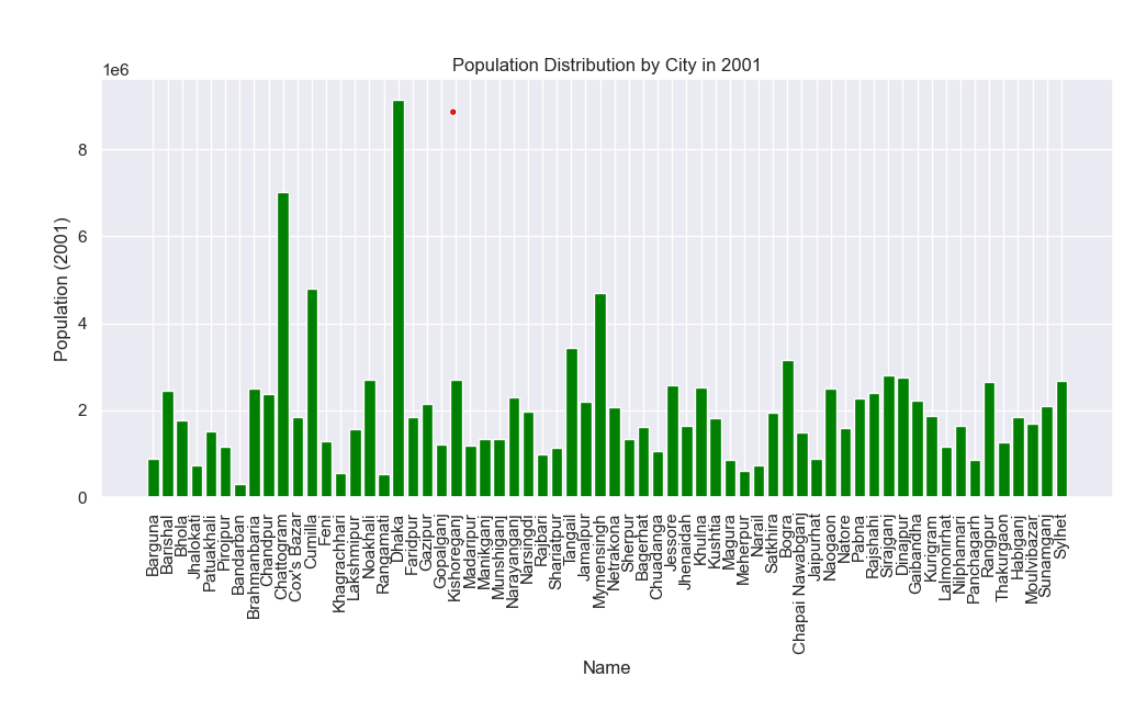


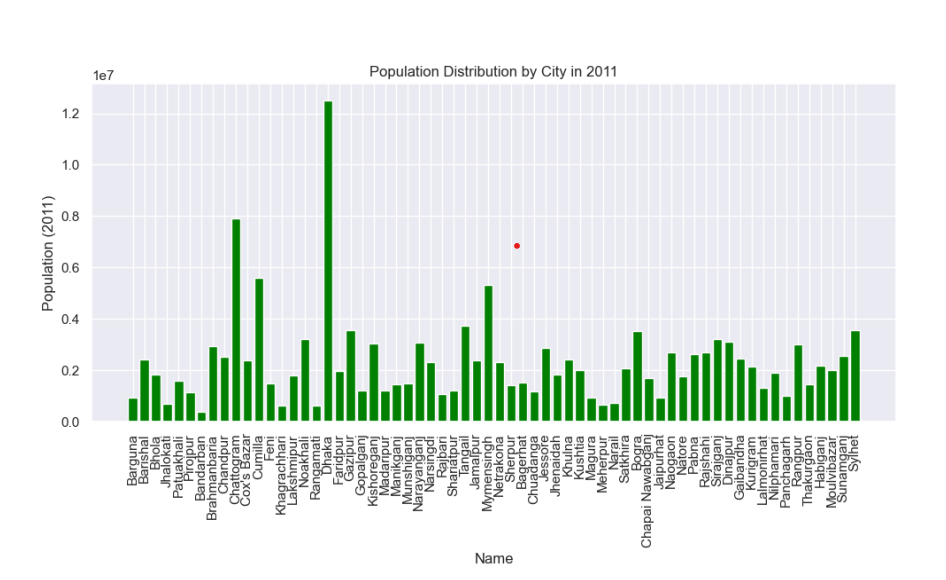
For the analysis of relationship between area and population, calculation of correlation between the two variables area and population of the years were done The picture above shows positive correlation between these two variables throughout the years with values standing ast 0.36, 0.31,0.25 and 0.23 which shows that with increment of one variable, the other variable also tends to increase. The correlation between the variables are not very strong however they do show a moderate degree of correlation.

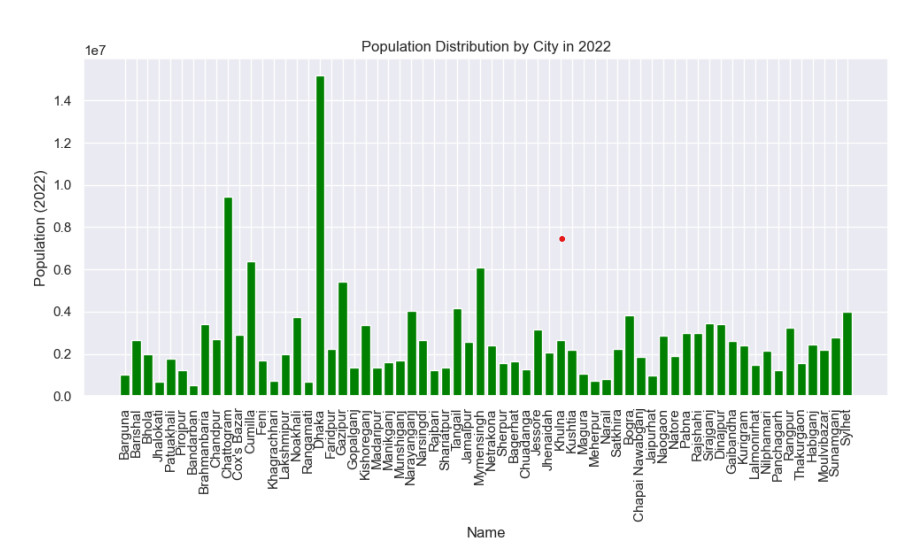
* **Major Analysis**

For the major analysis we have created histograms, scatterplot along with testing the hypothesis which works towards our hypothesis and justifies it that like majority of the capital cities across the world is Dhaka, the capital city of bangladesh also densely populated as compared to other cities. Below are the bar charts of population of the year 1991, 2001 , 2011 and also 2022 which is distributed among the towns of bangladesh.



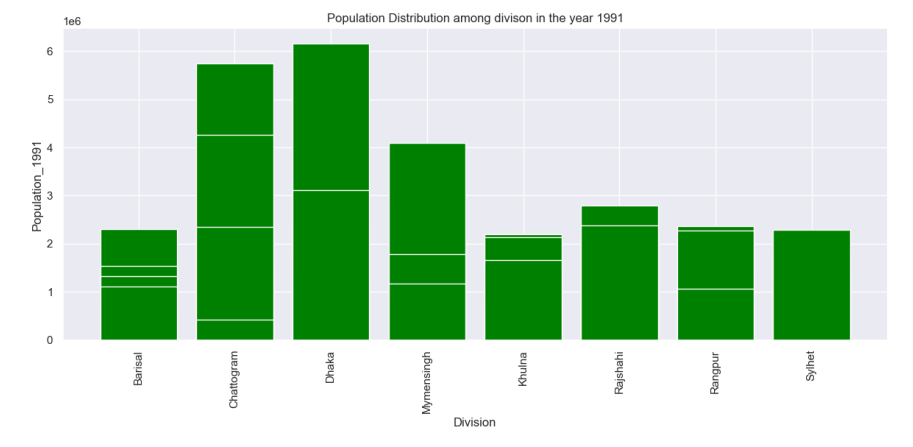


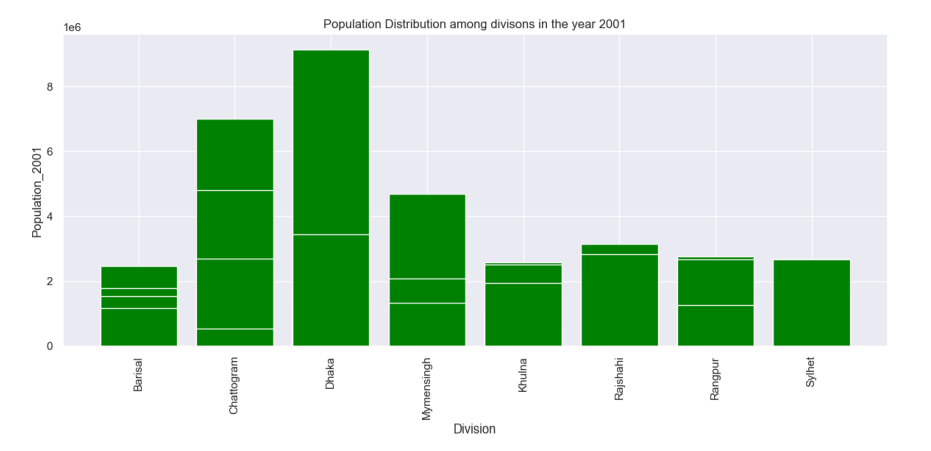


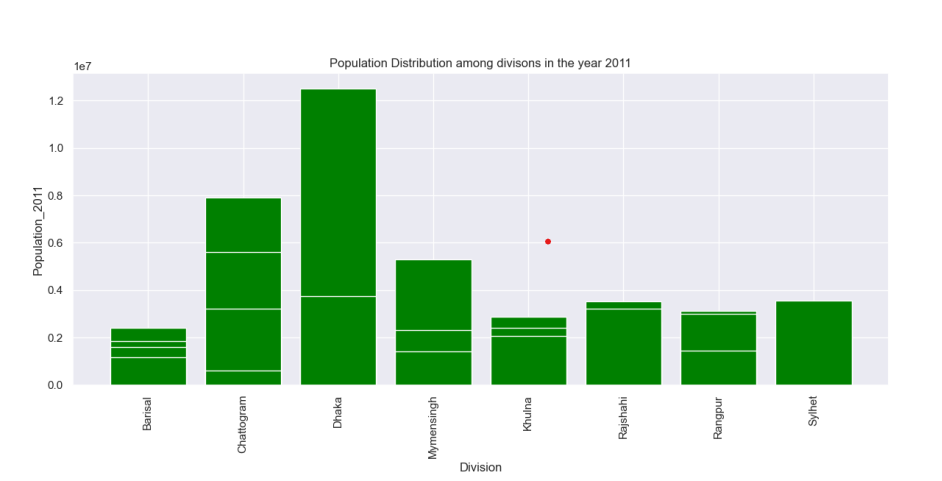


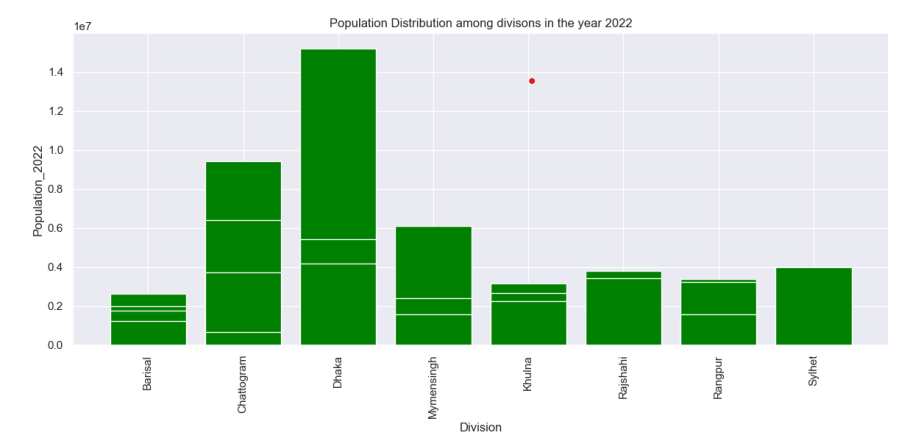
The four barcharts above show that the most populated town in Bangladesh is Dhaka, the capital which is further followed by Chattogram and Cumilia. The trend of Dhaka being more densely populated as compared to other towns has been followed throughout the four decades.

Similarly we also have barcharts which distribute population by the division of bangladesh cities. Bangladesh is divided into multiple divisions, the divisions are the highest level of administrative units below the national government. The divisions are headed by the Division Commissioner. The division comprises of multiple towns which varies according to the size and number

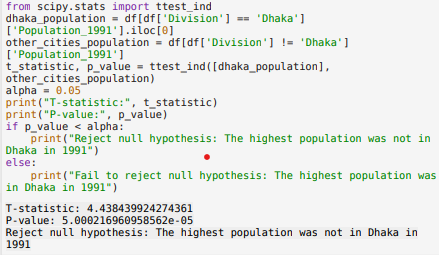


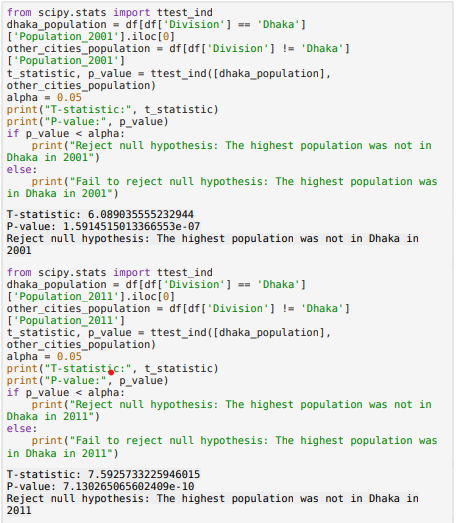


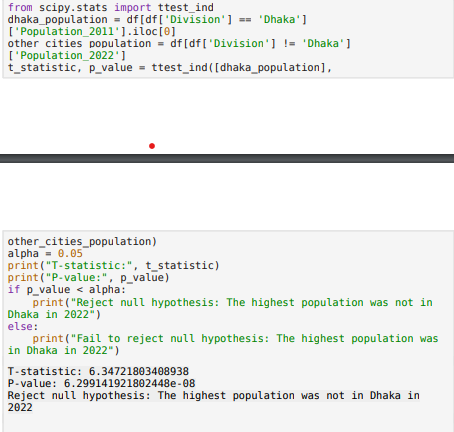




The barcharts above also support the analysis provided by the analysis initially done. The division Dhaka which comprises Dhaka town is the most populated division in all four years. Through all these four decades the most populated division has been Dhaka which is further followed by Chattogram. Dhaka division comprises ten towns such as savar, tongi, gazipur etc.







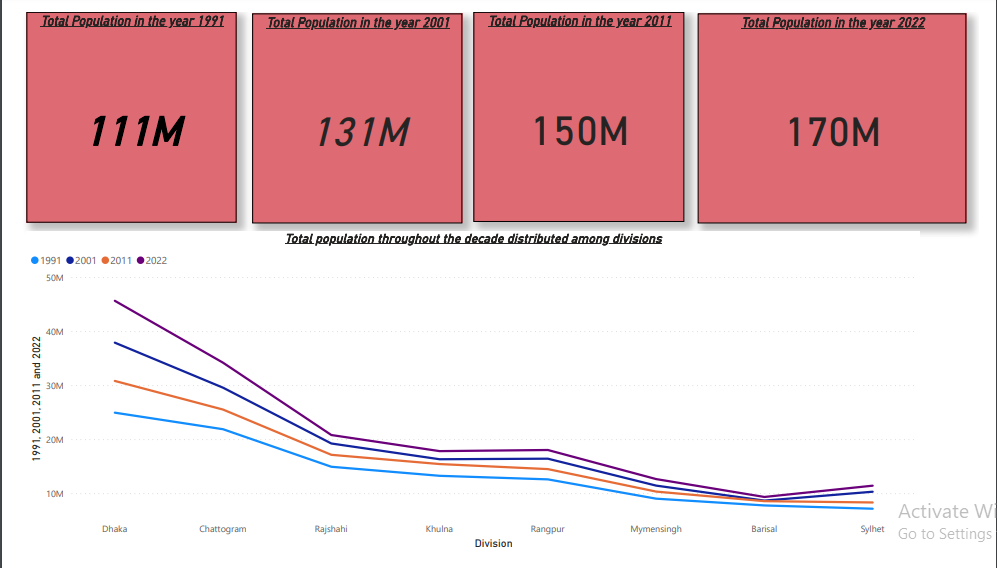
Furthermore, for explaining the problem statement we also performed hypothesis testing. For the testing, we conducted four tests for the population in division and the population throughout the years. All the screenshots attached above shows that the hypothesis on the basis of p value and t statistics rejects the null hypothesis that states

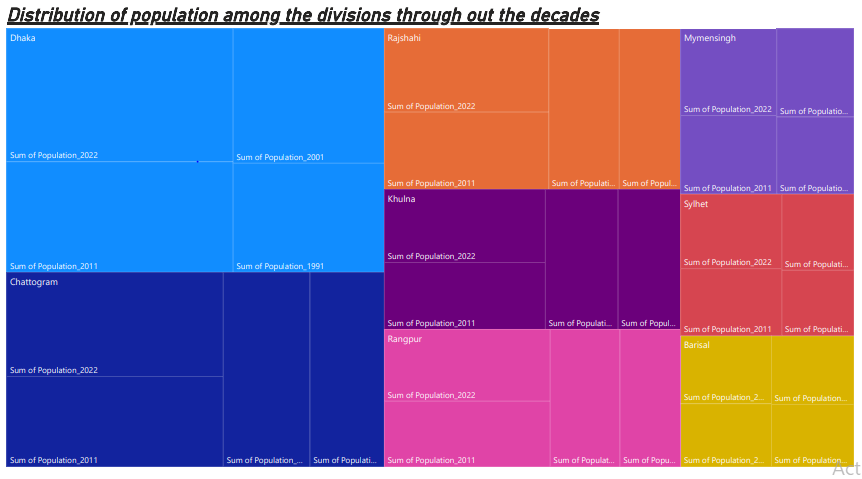
‘ The highest population was not in Dhaka”

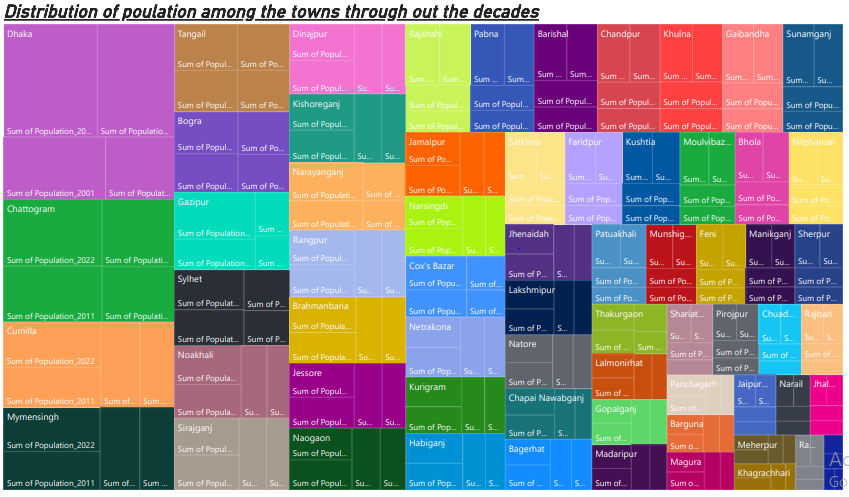
Which means that through the testing, it is true that the data set supports the highest population in Dhaka division.

**Activities performed in Power BI**

For the visualization through power BI , we used tools such as maps, table charts etc. The visualization obtained from Power BI also presented similar facts provided by jupyterlab which is that Dhaka is the most populated city among other cities of Bangladesh from the starting of the decade. The pictures obtained through visualization are attached below :









1. **Publishing through social media platform**

For the publishing work, we did the publication through,

.

**Conclusion**

Through the analysis and visualization performed through Power BI and jupyterlab and the entire work we performed throughout our pair project through the data sheet, we can conclude that Dhaka is the most densely populated city in Bangladesh. The major reasons behind Dhaka being the most populated city are :

* As a capital city, Dhaka acts as an administrative hub for the country. It accommodates national and international organizations, financial corporations and business houses.
* Dhaka provides a wide range of employment opportunities, and the Bangladeshi citizens from all over the country immigrate to Dhaka for it.
* Dhaka offers standard healthcare and educational services and the rapid urbanization also has contributed to an increase in population.
* Dhaka also serves as a gateway for international trade. The airports, ports and roads facilitate the movement of people along with goods and services both nationally and internationally.