|  |
| --- |
|  |
| IST 652 |
| Mini Project 2 |
|  |
| **ARYAN KAKADE** |
|  |

**Professor- Benjamin Nichols**

|  |
| --- |
|  |

* **The Data and its Source:** The dataset was taken from the Data.gov in URL format and converted in dataset using pandas. This dataset contains information about the census population by zip code from the year 2010.

(<https://catalog.data.gov/dataset/2010-census-populations-by-zip-code>)

* **About Dataset:** The dataset initially had 319 rows and 13 columns but after altering the unnecessary columns the final count was 319 rows and 9 columns.

1) Id- Each state is assigned a unique id.

2) Updated attributes- Number of updated count.

3) Zip Code- Unique identifying code for each city in the state.

4) Total Population- Total number of population in the city.

5) Median Age- Median age of male and female per city.

6) Total Males- Total number of males in the city.

7) Total Females- Total number of females in the city.

8) Total households- Total number of houses in the city.

9) Average Household Size- Average number of people living in the house.

* **Data Exploration:**

1) Improved dataset analysis using the head and tail function.

2) Using the describe function; I checked the unique values, total count, top, frequency, standard deviation, mean, min, and maximum.

3) Verified the data types of each column before manipulating the data.

4) A summary of the total number of rows and columns using the shape function to gain more detailed insights into the data.

* **Data Cleaning Techniques:**

1) After examining the data, I decided to remove unnecessary columns from the datasets and rename them further to facilitate analysis.

2) Using the df.duplicated function, count the number of duplicate rows, and remove these numbers from the dataset.

3) Totaled the rows to see if the duplicate rows had been eliminated or not.

4) Used the isnull function to look for null values and removed them from the dataset.

* **Comparative Analysis:**

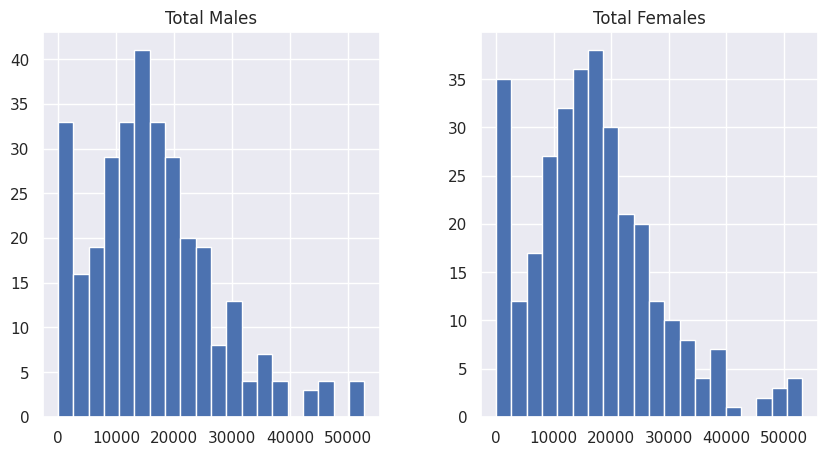
1) Comparison of data based on total population and zip code using pivot tables : We compared the mean of total population against the zip code which was the only unique variable in the dataset. This comparison uses 'Total Population’ as the unit of analysis. Computed the average total population and arranged it in descending order using the sored function, yielding the zip codes with the highest population.

2) Comparison of data based on total males and females using groupby() : We compared the total male and female population with the zip code. We used the first () function to return the first entry of males and females after grouping the zip code and population as columns. The population and gender breakdown for each individual zip code are shown in the result.

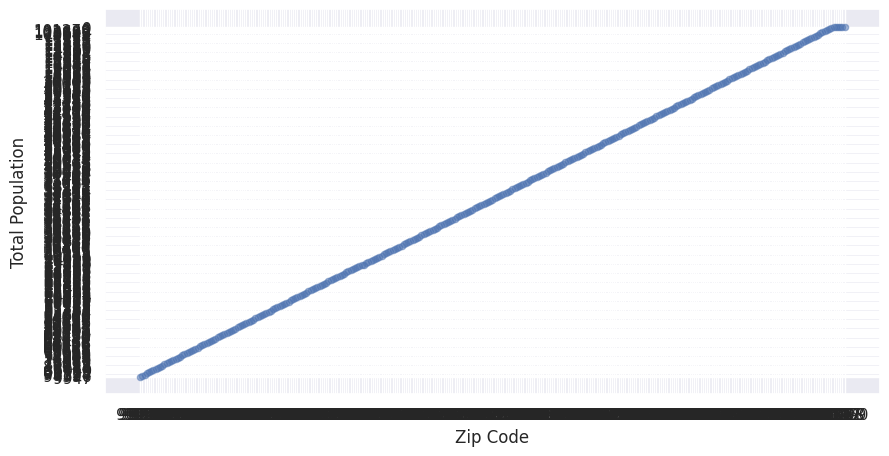
3) Comparison using statistical description: On the population, zip code, and males and females columns, we used describe () and the following metrics: count, mean, standard deviation, maximum value, and minimum value. The population and gender distribution of the various zip codes are compared using these statistics.

* **Exploratory Analysis:**

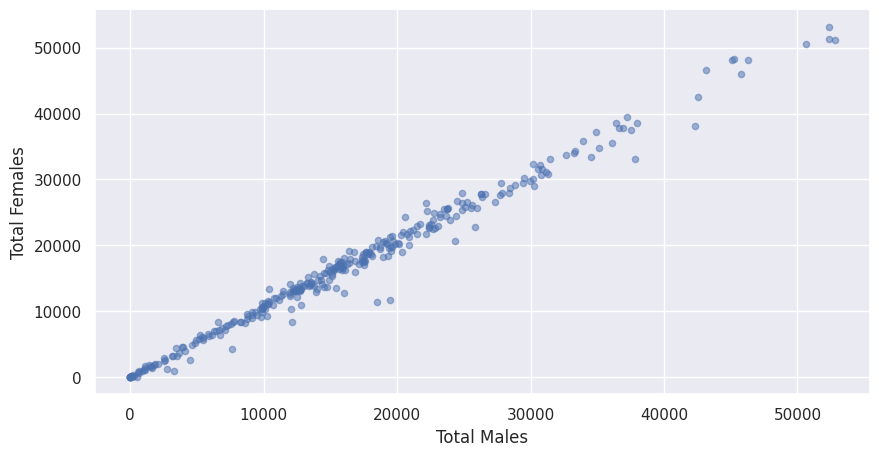
1) First analysis was done using histogram showing males and females against the population.



2) Second analysis was done using scatterplot showing total population against the zip code.



3) Final analysis was done using scatterplot showing total males against the total number of females.



* **Program Summary:**

1) Imported necessary libraries.

2) Connected Google colab with the Google drive.

3) Returning the HTTP status code for the response object.

4)The first 500 characters of the JSON string are pretty printed using the pprint function after reading the response to an HTTP request, decoding the response's content as a UTF-8 encoded JSON string.

5) Loading the json file in python object and converting it to dictionary using json.loads() function.

6) Checking for the keys in dictionary using keys() function.

7) Reading both the keys to get insights on the data to be extracted.

8) Checking the length and type of the keys.

9) Reading in pandas

10 Calling the dataset df to check the dataset for basic overview or analysis.

11) Using head and tail function on dataset to get insights about the data.

12) Checked the data types of all the columns using df.datatypes function and checked all the column names using df.columns to further rename them.

13) Use of df.describe function to check all the attributes of the data like, unique values, min, max, mean and, standard deviation.

14) Dropping of irrelevant columns using df.drop function.

15) Counting the total number of rows and columns and checking for the duplicate values.

16) Use of count function to verify the data cleaning process.

17) Finding the null values using isnull function and dropping them from the dataset.

18) Comparison of the data using pivot tables, groupby, aggregate function, and exploratory analysis on various attributes.

19) Sorted the comparison result in descending order to give better analysis result.