|  |
| --- |
|  |

**Census Data Extraction, Preprocessing and Exploration**

**Question 1: Geography and variable selection (20 pts)**

Public transportation in San Francisco massively spreads across different neighborhoods. Given the convenience of buses, I want to figure out if people are taking advantage of the resources. I will use census data about people’s means of transportation to work and focus on “car”(including car, truck, or van) and “bus”(including all public transportation, excluding taxicab). I will also look at how the data vary across sex.

**Question 2: Map of Variables (10 pts)**

|  |  |  |
| --- | --- | --- |
|  | **By car** | **By bus** |
| **All** | **Graphical user interface, application  Description automatically generated** | **Graphical user interface, application  Description automatically generated** |
| **Male** | **Graphical user interface, application  Description automatically generated** | **Graphical user interface, application  Description automatically generated** |
| **Female** | **Graphical user interface, application  Description automatically generated** | **Graphical user interface, application  Description automatically generated** |

**Question 3: Data exploration (20 pts)**

1. How many census tracts your data have?
   1. 197
2. What is the Total Population in the data?
   1. Total population in the data: 495315.0
3. Calculate percentages based on your variables of interest
   1. Percentages of people going to work by private car: 0.411433128413232
   2. Percentages of people going to work by public transportation: 0.3402158222545249
   3. Percentages of male going to work by private car: 0.22857373590543392
   4. Percentages of male going to work by public transportation: 0.17095383745697182
   5. Percentages of female going to work by private car: 0.18285939250779806
   6. Percentages of female going to work by public transportation: 0.16926198479755308
4. What is the total area in square kilometers?

**Question 4: Variables Creation (20 pts)**

1. If you have population groups. Calculate and plot the probability density function of the ratio of each group to total population.

For Assignment 2, part 2 you will export the this data as a shape file and select 6 census tracts variables to do K-means Clustering.

**Question 5: Summary Statistics (10)**

Use the function describe() to summarize the statistics of the variable of interests

Table

Description automatically generated

**Question 6: Summary Statistics (10 pts)**

* Most of the pairs with the same means of transportations(car or bus) are largely correlated

**Question 7: Data Interpretation (10 pts)**

* More people commute to work by car than by bus.
* On average, more male drive to work than female.
* The differences of taking buses to work between male and female is small.

Graphical user interface, application

Description automatically generated