Question 1: Geography and variable selection (20 pts)   
Propose a data analysis project using 6 or more census variables and a region to   
motivate a question of interest. You can investigate journal articles or newspapers to   
motivate your study. Write in words what motivates your region and variables selection.

The project that I will take on is to analyze difference in demographics in areas over time. This means that I will be looking at the population (total, white, black, and Asian) as well as the population with bachelor’s degrees and then the median household income. I am curious to see how San Diego evolved over time. I am originally from San Diego, so it is certainly something that is exciting to see. I will look for the tracts that have had the biggest changes in any of the variables. Maybe I will even make a table that is the differences from 2019 and 2017. It will be interesting to cluster all three, 2019, 2017, and difference tables.

Question 2: Map of Variables

Below I will only present one set of maps. I did these maps 3 times for both years and the difference and also for percentages. Note: Percentages and real values create the same map, just the legends are different

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

Question 3: Data exploration (20 pts)   
a) How many census tracts your data have?   
b) What is the Total Population in the data?   
c) Calculate percentages based on your variables of interest   
d) What is the total area in square kilometers?   
Graphical user interface, application

Description automatically generated

Question 4: Variables Creation (20 pts) and Question 6: Summary Statistics (10 pts)  
   
a) If you have population groups. Calculate and plot the probability density function   
of the ratio of each group to total population.

Use the function grid = seaborn.PairGrid() to summarize the statistics of the variable of   
interests.   
• Which pair of values of the correlation coefficient among variables are largely   
positive (\pho > 0.5) or negative (\pho < -0.5))   
• any surprising low correlation?

Diagram

Description automatically generated with medium confidence

The correlations are not the main focus for this data analysis. San-Diego tends to be predominately white, so I would have to go at an individual tract level to really take notes on the correlations and or differences in demographic makeup.

Question 5: Summary Statistics (10)   
Use the function describe() to summarize the statistics of the variable of interests   
Table

Description automatically generated

Note: this is for the difference, negatives indicates that the value deteriorated from 2017 to 2019

Question 7: Data Interpretation (10 pts)   
Based on the results of the previous questions, summarize what can you learn from the   
data. Is there any result that you would like t to obtain from the clustering method?

Based on my findings, I notice general trends among the population that indicates that there are increases in the population in both general and in specific for each demographic. There was also an increase in median household income, but when looking at the map for it, many places did indeed go down. In further work on this project I would look at individual tracts or at least clusters of tracts as we will do in part 2 of this assignment. This will give a better understanding of the area based on demographic shifts over 2 years.