Stat 6021: Homework Set 2

1. For this question, we will work on the dataset `PoliceKillings.csv`. This dataset was the basis for this article on Police Killings in the year 2015. You may read more about the data and the variable descriptions here.

   (a) Using the `raceethnicity` variable, create a table and a bar chart that displays the proportions of victims in each race / ethnic level. Also, use your table and bar chart in conjunction with the US Census Bureau July 1 2022 estimates to explain what your data reveal.

   (b) Convert the variable `age`, the age of the victim, to be numeric, and call this new variable `age.num`. Use the `is.numeric()` function to confirm that the newly created variable is numeric (and output the result), and add this new variable to your data frame.

   (c) Create a density plot of the variable `age.num`. Comment on this density plot.

   (d) Create a visualization to compare the ages of victims across the different race / ethnicity levels. Comment on the visualization.

   (e) Create a visualization to compare the different causes of death (variable `cause`) across the different race / ethnicity levels. Comment on this visualization, specifically on whether the cause of death appears to be independent of the victim's race / ethnicity.

   (f) Pick at least two variables from the dataset and create a suitable visualization of the variables. Comment on what the visualization reveals. You may create new variables based on existing variables, and decribe how you created the new variables.

2. For this question, use the .csv data file that you created at the end of the previous homework set, `stateCovid.csv`. The dataset should contain 4 columns:

   - the name of the state (55 "states", the 50 states, plus DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands)
   - the number of cases
   - the number of deaths

- the case fatality rate, defined as the number of deaths divided by the number of cases

You may realize that when you exported the data file as a .csv file, an extra column was added to the dataframe. Remove this column.

(a) There is a dataset on Canvas, called `State_pop_election.csv`. The data contain the population of the states from the 2020 census (50 states plus DC and Puerto Rico), as well as whether the state voted for Biden or Trump in the 2020 presidential elections. Merge these two datasets, `stateCovid.csv` and `State_pop_election.csv`. Use the `head()` function to display the first 6 rows after merging these two datasets.

(b) Pick at least two variables from the dataset and create a suitable visualization of the variables. Comment on what the visualization reveals. You may create new variables based on existing variables, and decribe how you created the new variables.