70% of your grade is programming accuracy and elegance (appropriate utilization of SQL and other functions) and 30% of your grade is on professional presentation and clarity of your answers (no screenshots).

Consider the CITIBIKE 1, 2 and 3 datasets

a. Combine files into a single dataframe.

b. Report the dimensions, structure and overall summary statistics from your new dataset.

c. Create a new variable called “age” based upon birthyear. Impute any outliers (provide a definition) with the median.

d. Create appropriate visualizations for gender, age, trip duration.

e. Report the top 10 start stations and end stations by frequency.

f. Create three categories for trip duration - <600 seconds, 600 – 800 seconds and >800 seconds. Report the distribution of rentals across these categories.

g. Answer the following questions, including appropriate visualizations
   • Is there a statistically significant difference in trip duration of rental by gender?
   • Is there a statistically significant difference in category of trip duration by gender?
   • Is there a statistically significant difference in category of trip duration by age?

h. Please provide appropriate tables, graphics and formatted output in your paper. You are suggested to use MLA or APA for all formatting.

i. Provide all of your annotated code as an appendix. Your code must match your output.