Prospectus: superEDA.R

Grace Lee, Emily Wang, Leo Yoon (Group 3)

The superEda.R file contains functions that carry out univariate and bivariate EDA on both continuous and categorical data. The function already contains some functions that perform univariate and bivariate EDA. This project will work on adding additional functions to perform bivariate EDA on all types of data(categorical/continuous). By default the functions that we will be adding will take in a X and Y variable. The functions will output both graphical and non-graphical EDA components. Graphical components will be done mostly using the "ggplot2" package to keep consistency. Non-graphical EDA outputs will mostly be printed out but not as a separate file. To reproduce the outputs, one would have to manually save all the graphs and tables. Largely there will be 3 functions added: biCatCont, biContCat and biContCont function.

The biCatCont function will take in 2 variables, X which is categorical and Y which is continuous.In order to better understand our variables graphically, we will use a boxplot and overlaid histogram/density plots using ggplot2. We can also look at the statistics by each category non-graphically by using the tapply function. We want to include the mean, median, 1st, standard deviation, minimum, maximum and 3rd quartile.

The biContCat function will have four arguments. One will take in the continuous independent variable, another will take in the categorical dependent variable, the third will be a user specified cut off of the continuous IVs, and the fourth will be user specified bin widths for the conditional density plot. The third argument will be used to "bin" the IV into groups to allow for cross-tabulation and conditional density plots. The non-graphical component will return summary statistics for each bin: minimum, 1st quartile, mean, standard deviation, median, 3rd quartile, and maximum. The graphical conditional density plot made with ggplot2 will be created using the fourth argument specifying bin width.

The biContCont function will take in two continuous variables for a multivariate EDA. For the non graphical eda, this function will return a correlation value between the two continuous variables(2 significant digits). The correlation value will tell the strength of the relationship between the two continuous variables. For the graphical EDA, the function will return a scatterplot with a smoothing curve(loess smoothing curve). Overall what the biContCont function will be returning a brief description of the variables names and the types of eda done (using "cat" in R) and returning the correlation value and a scatterplot with appropriate title and labels.