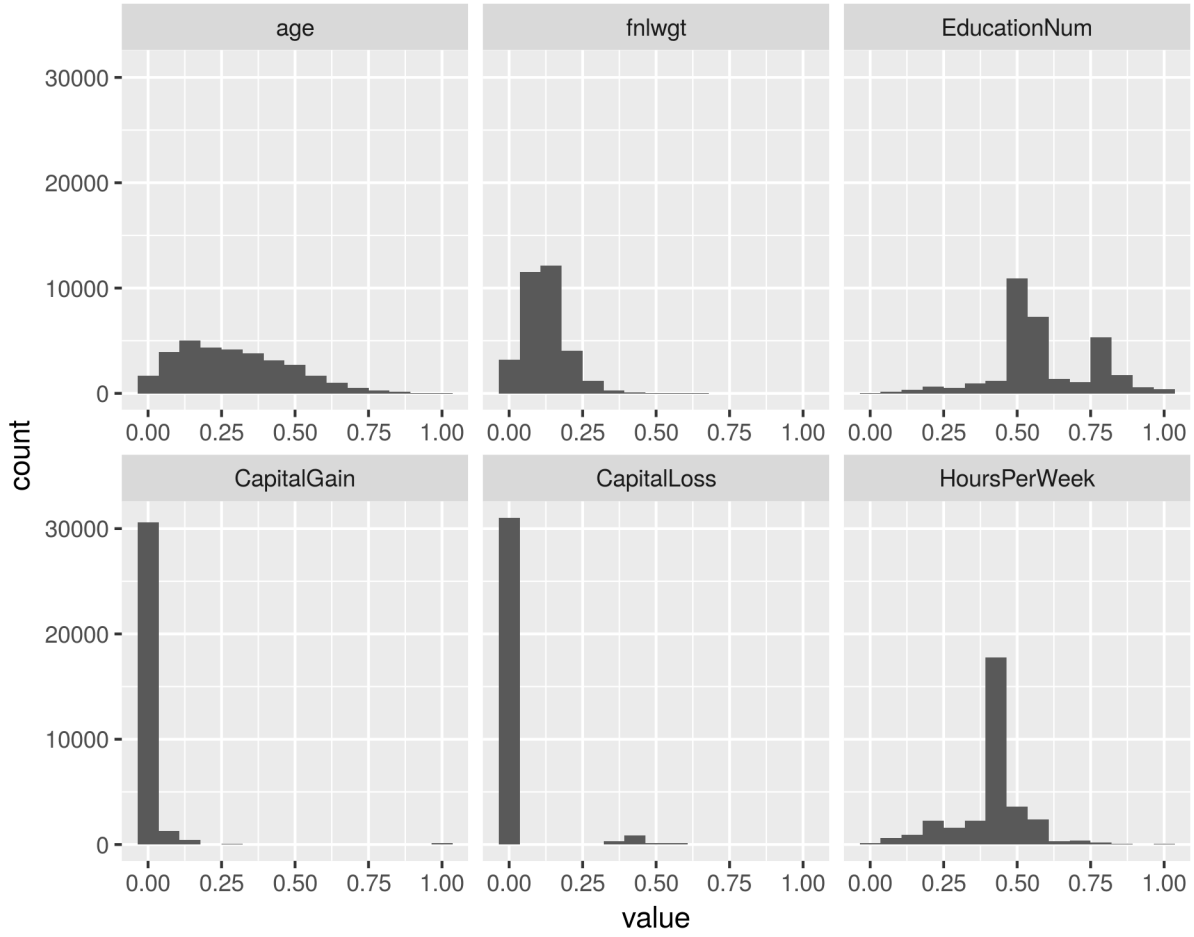
3. Analyzing a Data Set

Data exploration
the adult data contains both categorical and continues variables. To explore it, we use 2 level of data explorations; univariate and bivariate analysis.

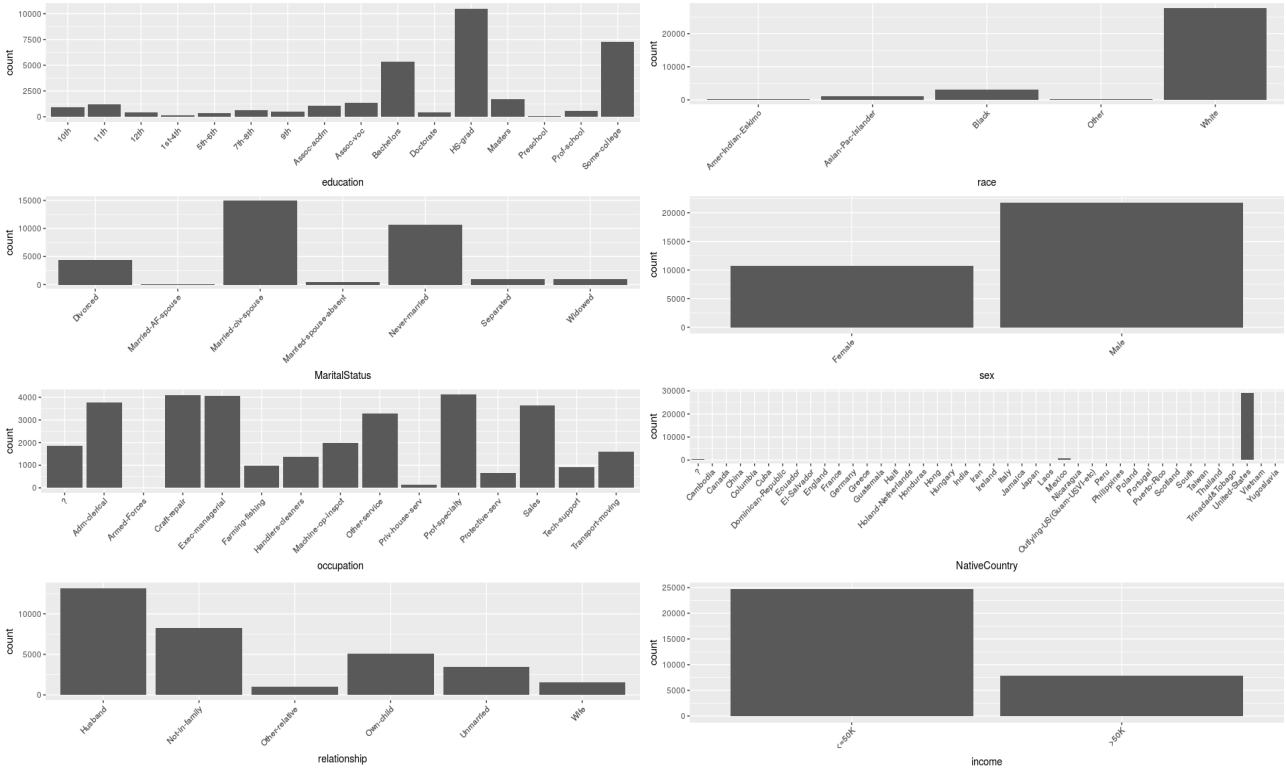
A - Univariate Analysis: for univariate analysis we use histograms for continues data

1. Continuous data histograms



Findings:
non of the continues variables is symmetric. The **Age** variable tends to be positively skewed, where most of the objects are located before 50 years old. The **fnlwgt** variable is mainly ranges between 125k and 250k, along with some outliers larger than 500k. The **Education** variable has 2 significant peaks at “HS-grad” and “Some-college” values. These peaks can act as means of 2 separate normal distributions. For both **Capital Gain** and **Capital Loss** variables, majority of data is centralized at 0, which means adults have not gain or loss money by selling or buying assets, and statistically both variables are useless in any model learning. The **Hours per Week** variable tends to fit a left skewed density function, with obvious mean and mode values of 40 Hours.

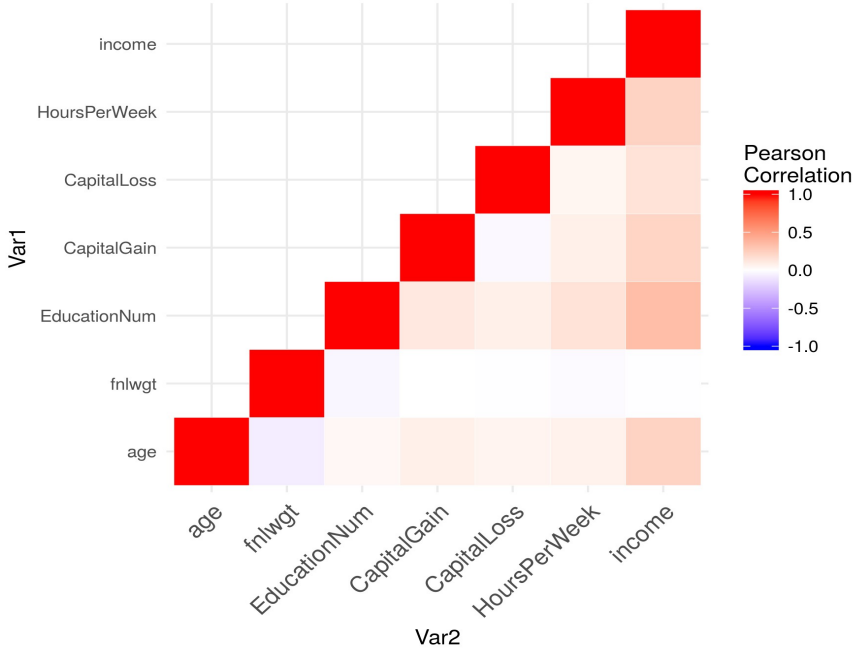
2. Categorical data histograms



Findings:

1. the majority of objects represents white people.
2. male objects are fast double of female objects
3. nearly half of objects represent "**Married-civ-spouse**", and third of data have never been married
4. occupations of "Prof-specialty", "Craft-repair", and "Exec-managerial" represent third of the data, followed by "Adm-clerical" and "Sales"

B – Bi-ivariate Analysis: for this analysis we use correlation heatmap between continues variables and the numeric representation of "**income**" variable (i.e. set '<=50K' to 0 and '>50K' to 1).



The above heatmap shows there is generally no significant correlation between any of the variables.

However, regarding the “**income**” variable, we notice slight positive correlation between “**income**” and “**EducationNum**”.