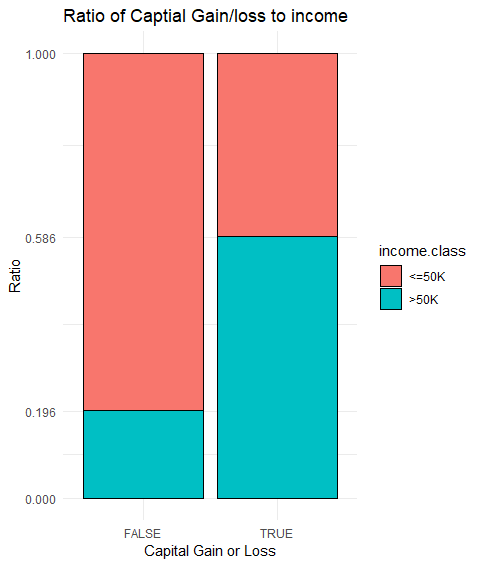
**Predicting Income Class following CRIPS-DM**

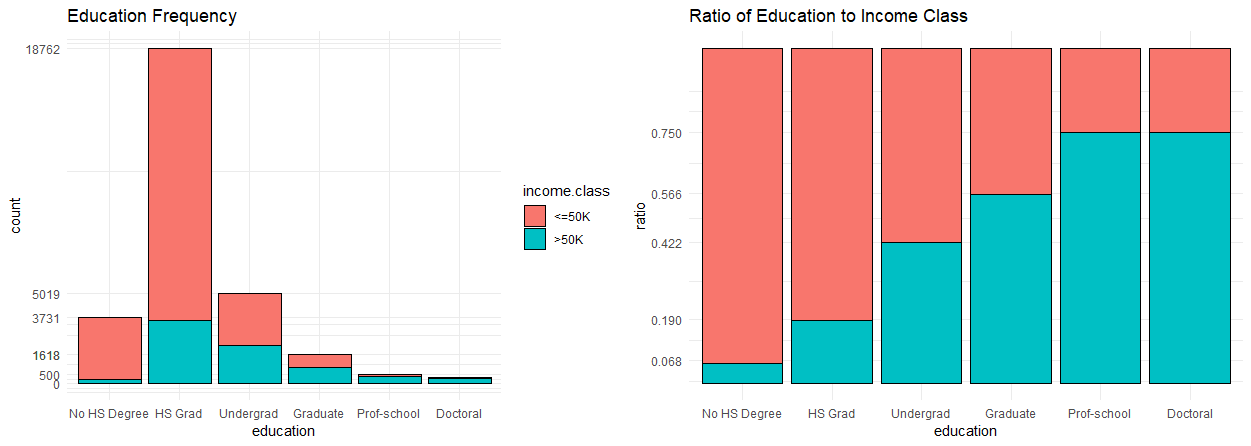
**1.Business Objective 2.Data Understanding phase**

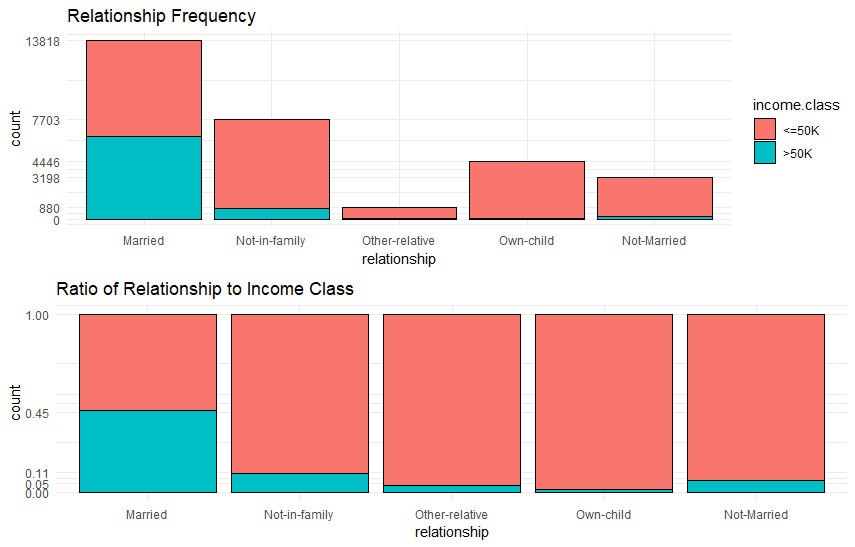
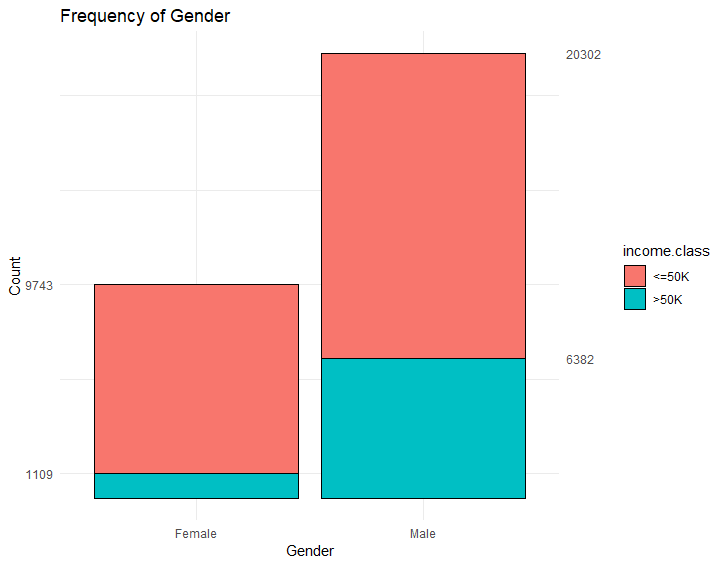
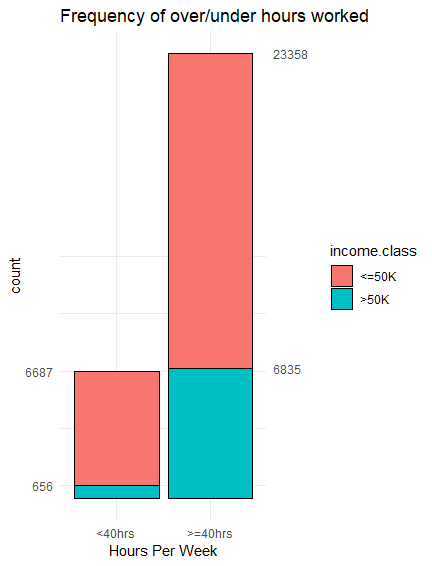
Our goal is to use be able to predict if a person can makes over/under 50K based off of basic information. We will be using the Adult data set from UCI machine learning depository and implement various classification methods after the data is pre-processed and EDA is concluded.

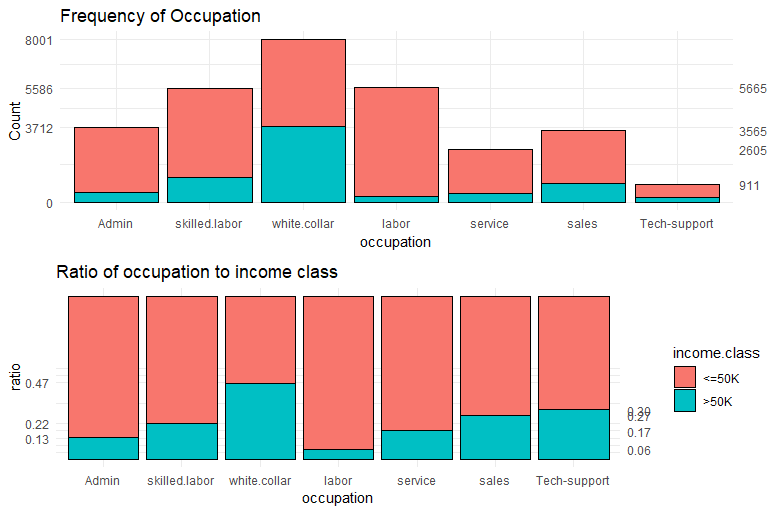
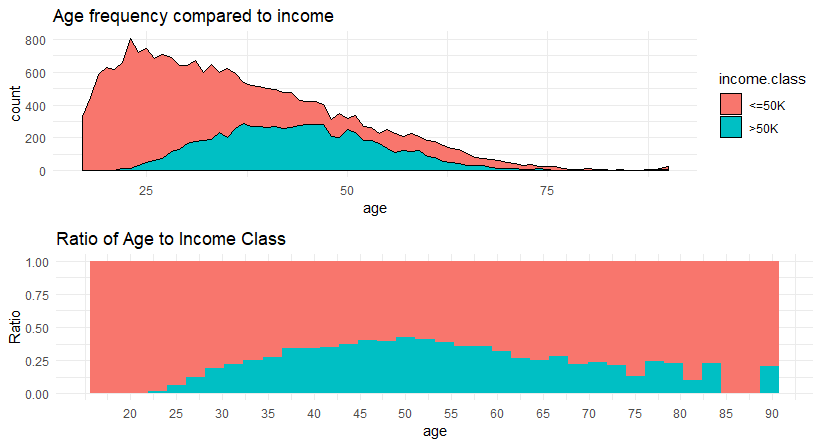
This data set contained 15 variables including the predictor. The variable types are numeric, integer, and categorical. Due to the minimal space, only the meaningful predictors will be shown. The graphs also show data that has been re-classified for the purpose of generalizability and viewing

****

**EDA**

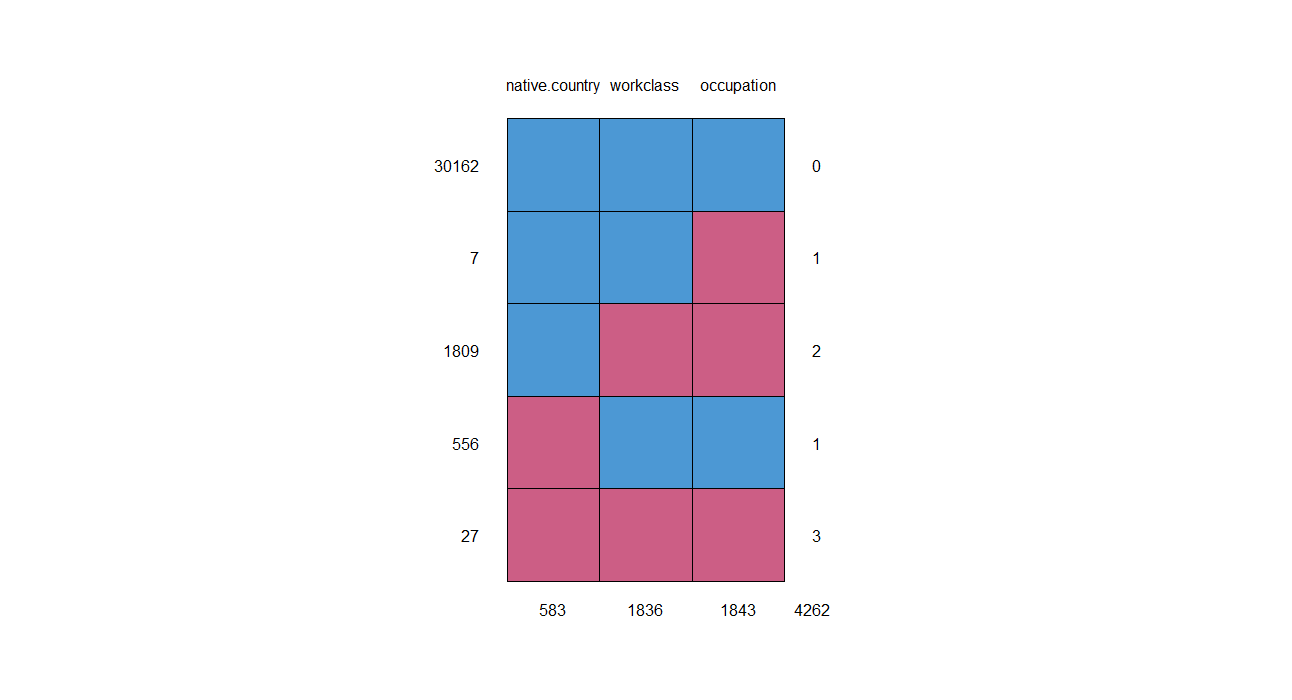
****



****

**3.Data Pre-Processing**

Normally this is where most of the time is spent in the data mining process as most raw data needs to be cleaned, tidied as well as transformed to be prepared for modeling. This data set was already clean and the main issue with it was to deal with the missing data.



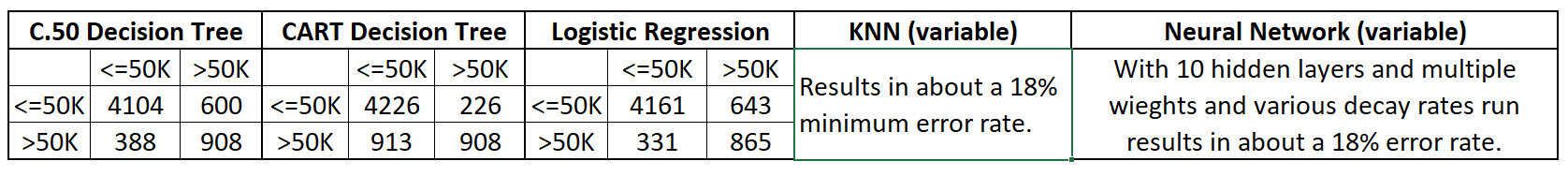
This graphic shows a nice overview of the missing data. The biggest take away is the fact that if work class is missing then occupation will missing as well and vice versa. This means when trying to run supervised classification algorithms it should not use work class to help predict occupation. Imputation was attempted using decision trees but the results where unsatisfactory. C50 and CART decision trees where used for imputation but the results where unsatisfactory. Due to the size of the complete set of observations the missing data was removed after comparisons to the complete dataset were made. The data set was also split into training and test and normalized with 80% train and 20% test.

**4. Modeling Phase**

The Models that where run were logistic regression, C50 +CART decision trees, K-nearest neighbor, and Neural Network. Each model was run with various inputs to see the effect that each these variables had on predicting income class. The selection of the final models were based on accuracy and minimizing the number of variables.

**5. Model Evaluation**

To evaluate the performance of the models we looked at their accuracy in classifying the test dataset. Since the data set had a split of around 75% <=50K and 25% 50K error rate will be looked at on how much it improves upon a base 25% error rate archived of categorizing all the data as <=50K. A lift or gain chart would be the best way to demonstrate this models effectiveness.

****