**Descriptive Analytics Data Mining Report**

**Executive Summary**

Walking through the data collection process, describing the variables of interest, and discussing the overall structure of the dataset

**Steps in Preprocess and Cleaning the Data**

The raw dataset (csv form) is extracted from the 1994 Census Bureau Databased) and obtained from Kaggle (<https://www.kaggle.com/uciml/adult-census-income>), containing relevant variables contributing to predict whether a person’s earning is greater $50K per year.

The following steps are taken to clean the data.

First, raw dataset is read and explored in R to determine necessary steps afterwards. Initial check indicates that raw dataset has both numeric and character/string data and there are also missing data in the dataset, but the missing data is not represented “correctly” in the form of “NA”. Instead, they are shown as a question mark “?”.

Next, a subset of interest of raw dataset is obtained by dropping variable “education”, “fnlsgt” and “relationship”. “Education” is dropped because there is also “education,num” (years of education) in the dataset. “fnlsgt” is dropped because it is the final weight on the Current Population Survey, which is not a variable of interest. “Relationship” is dropped because it is not a variable of interest.

Third, missing data is explored by looking at missing values per row and per column. I found out that many missing data exists in “work class” (1836 missing data) and “occupation” (1843 missing data). Besides, there are a total of 4262 missing data point. Also, based on Graph 1, work class, work occupation and naïve country have a not-so-low percentage of missing data. Since work class and occupation are important determinants in predict whether people make over $50K, I decided to delete all missing values to obtain a complete dataset for analysis.

Fourth, missing data is deleted by following the steps below. All missing data indicated by question mark “?” in the original dataset is replace with “NA”. Then a new dataset is created without missing values of data.

Finally, for the new dataset with no missing data, new coding variables are created for all categorial variables to facilitate further analysis (such as regression and prediction). Table 1 and 2 show the dataset structure before and after cleaning. After data cleaning, all strings have correspondingly numeric variables and three are no more “suspicious” marks such as “?” .

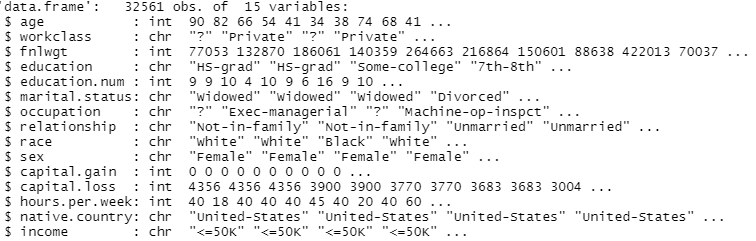


Table 1 Raw Dataset Structure

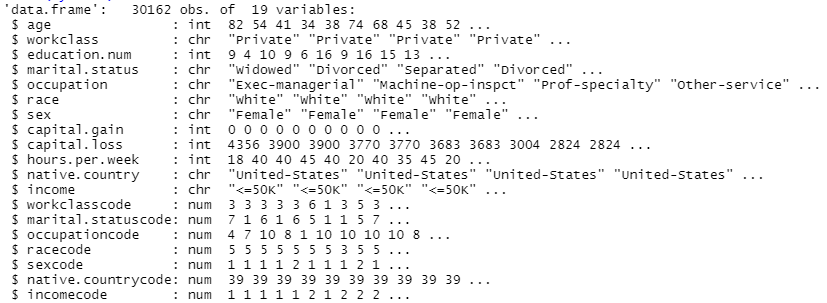
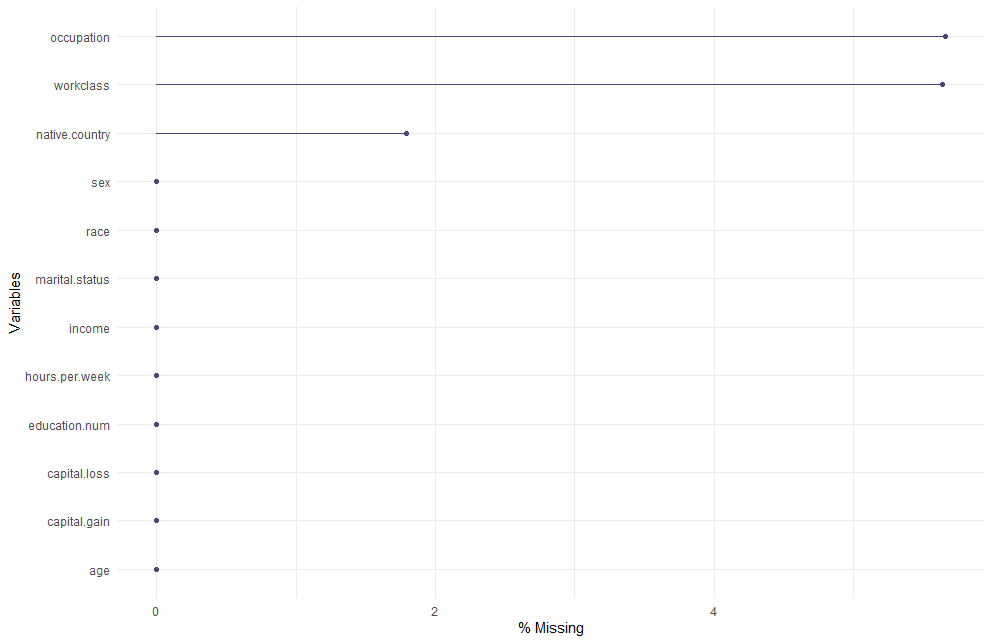


Table 2 Cleaned Dataset Structure

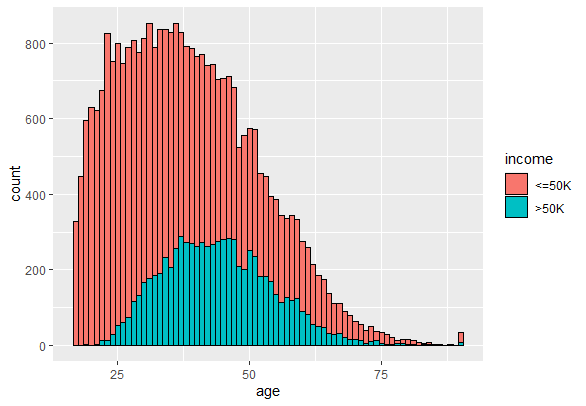


Graph 1 Missing Data Structure

**Exploratory Analysis on the Cleaned Data**

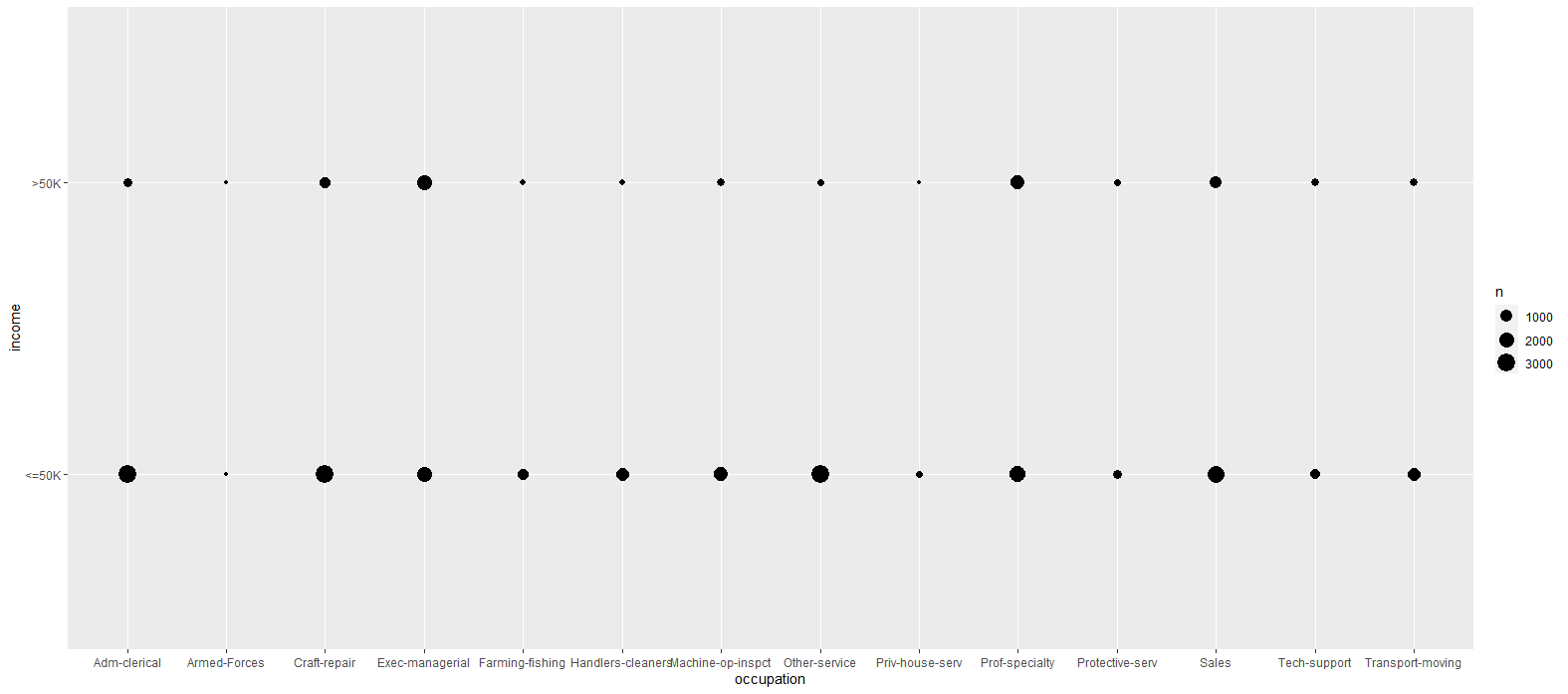
*Distribution of Income*

Graph 2 shows distribution of income by age. It illustrates that for both income groups (income <$50K and income>$50K), income distribution are not symmetrically distributed. First, there are more lower income individuals than higher income persons. This statement is also true conditional on age. Second, for lower income groups, the age distribution is “younger” and right-skewed. In contrast, for higher income groups, the age distribution is “older” with no obvious skewness.



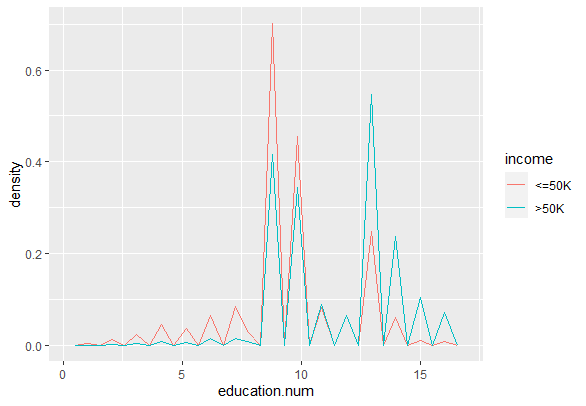
Graph 2 Income Distribution by Age

Graph 3 shows distribution of income by occupation. It illustrates that for both income groups (income <$50K and income>$50K), income distribution are not symmetrically distributed among different occupations. First, some occupations tend to be taken by lower income persons (low-income occupation): there are more lower income individuals than higher income persons for that occupation, such as craft-repair, farming-fishing, handlers-cleaners, machine-op-inspect, other service, sales and transport-moving. In contrast, some occupations tend to be taken by higher income persons (high-income occupation), such as exec-managerial. Second, for lower income groups, the difference among occupations are less severe: there are relatively less differences in the sizes of the “dots”.



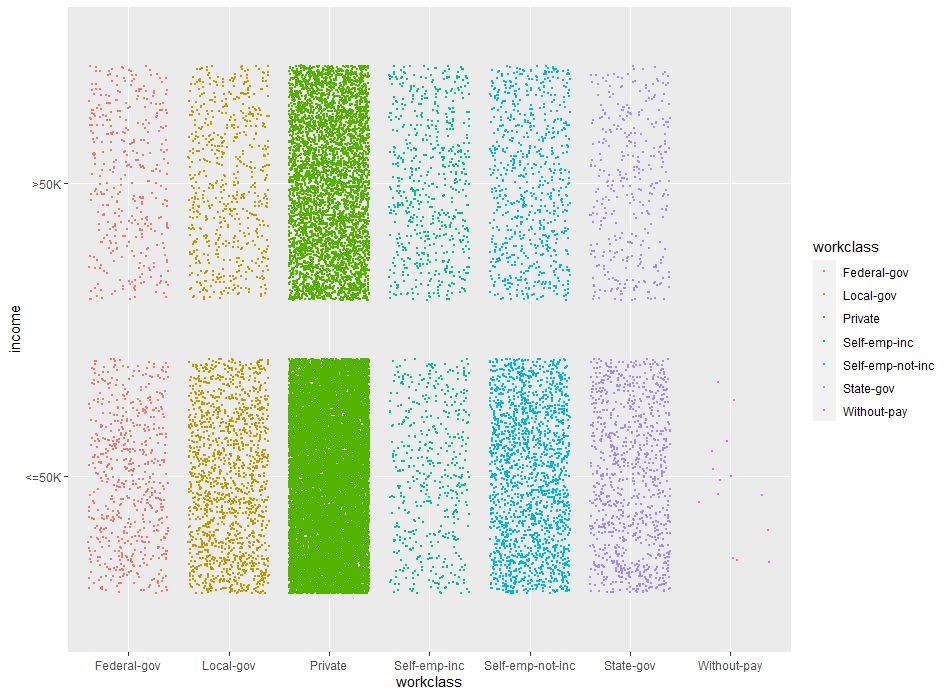
Graph 3 Income Distribution by Occupation

Graph 4 shows distribution of income by years of education. It illustrates that, first, higher income individuals show more years of education: there are more spikes in the density curve for higher income group towards the end where years of education is longer. Second, the portion of the less-educated is smaller for higher income groups: the density curve is almost flat when the years of education is less than 8. Third, the portion of the more-educated is greater for higher income groups: the density curve has more (higher) spikes when the years of education is greater than 12.



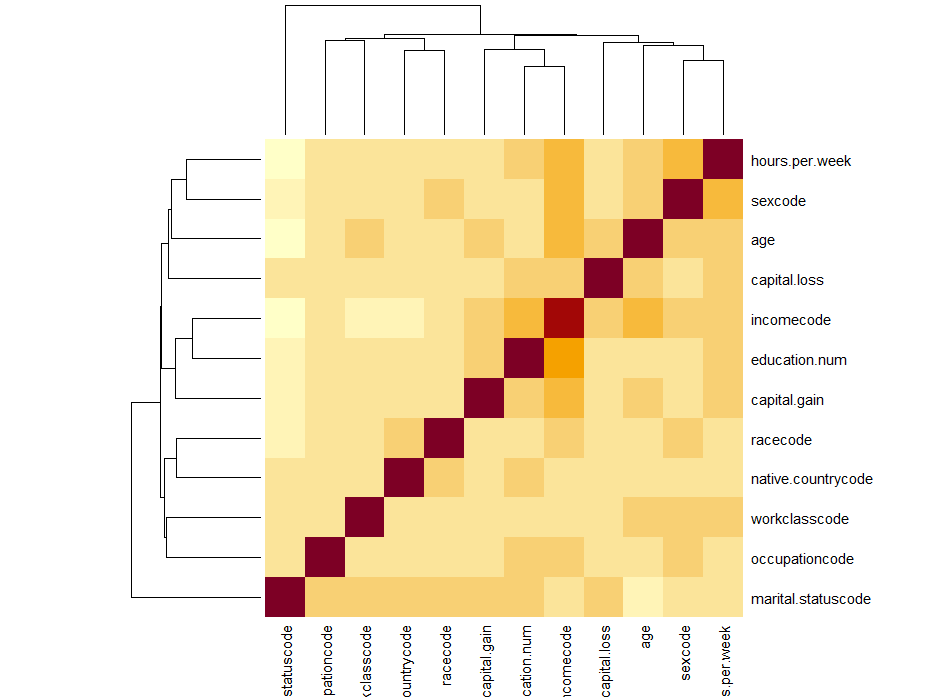
Graph 4 Income Distribution by Years of Education

Graph 5 shows distribution of income by work class. It illustrates that, first, lower income individuals mostly work in private firms and the percentage is huge as is implied by the almost-full dot graph. The statement is also true for higher income persons: most higher income persons work in private firms. Second, income distribution is more asymmetric for lower income groups: the density differences in dot graphs are greater than that for higher income groups.



Graph 5 Income Distribution by Work Class

Graph 6 and Graph 7 show correlations among variables in the dataset. The heatmap reveals no “very strong” correlation as the overall color of the map is not very “red”. Judging from the heatmap, income is mostly correlated with years of education, hours per work, gender and age. Correlation among explanatory variables are not obvious. Graph 7 confirms our observations from the heatmap and gives out the correlation coefficient. Income is mostly correlated with years of education, with a correlation coefficient of 0.3. Correlations among independent variables are small in size.



Graph 6 Correlation Heatmap



Graph 7 Correlation Matrix Map

**Conclusion and Discussion**

Based on basic exploration of the dataset, we conclude that income is affected by lots of factors, among which years of education, hours per work, gender and age are main determinants. Also income distribution between lower income group (<$50K) and higher income group (>$50K)is unbalanced in terms of age, occupation and work class.

By descriptive analytics data mining, I learned that data cleaning is the stepstone in data analysis. Proper data cleaning and pre-process can improve quality of dataset and reduce redundant work afterwards. Also, visualization is very helpful in helping us understand the basic data structure, distribution and correlation relationship in the dataset, which make us better prepared for further analysis.