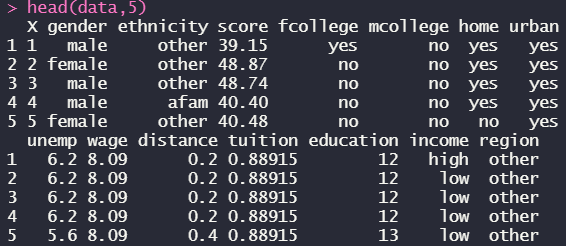
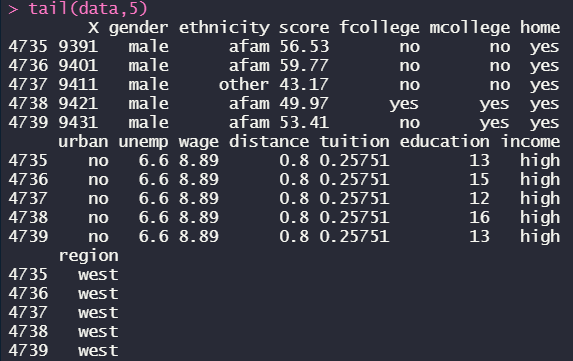
**Using R to Analyze the Data related to the College Distance conducted by the Department of Education in 1980.**

Statistical Computing – Project 1

**Objective**: The purpose of this project is to make some analysis upon the dataset related to the College Distance from the High School and Beyond survey conducted by the Department of Education in 1980. There are 4739 observations and 14 variables included in this dataset, the variables are “**gender**”, “**ethnicity**”, “**score**”, “**fcollege**(is father a college graduate)”, “**mcollege**(is mother a college graduate)”, “**home**”, “**urban**”, “**unemp**”, “**wage**”, “**distance**”, “**tuition**”, “**education**”, “**income**” and “**region**”. Given the dataset, I’ll try to apply all of the related knowledge and techniques that I’ve learned so far in this course for digging deeply into the dataset and trying to extract some valuable information from it.





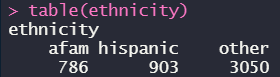




First, I import the dataset from the directory where it stores, since the type of the dataset is.csv(comma separated values) that we use the “**read.csv()**” and specify the header is already included, then display the first five lines of data to make sure everything is fine, and use the “**dim()**” to determine the dimension of data is exactly as we expected it to be.

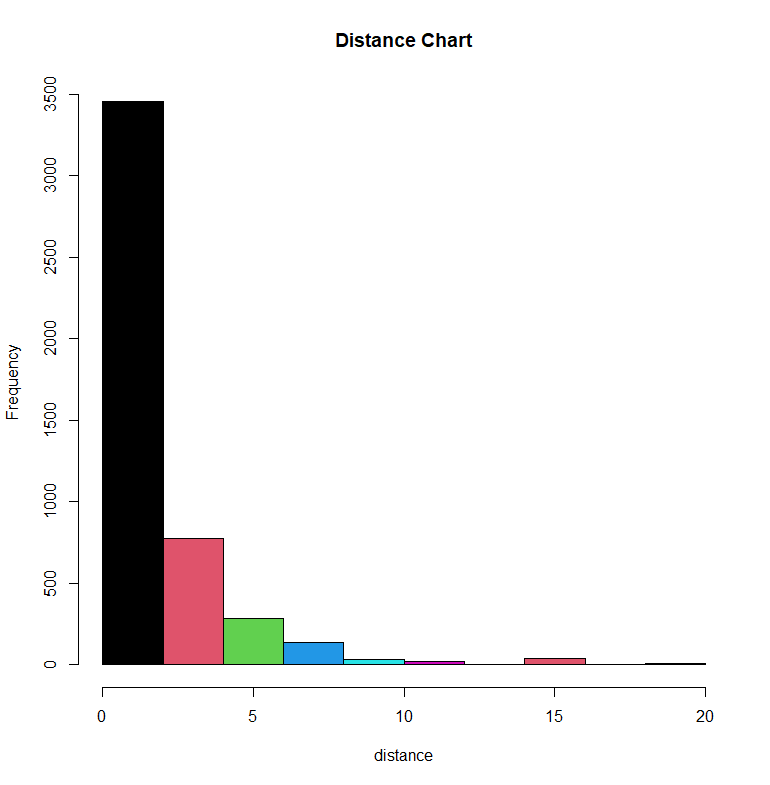




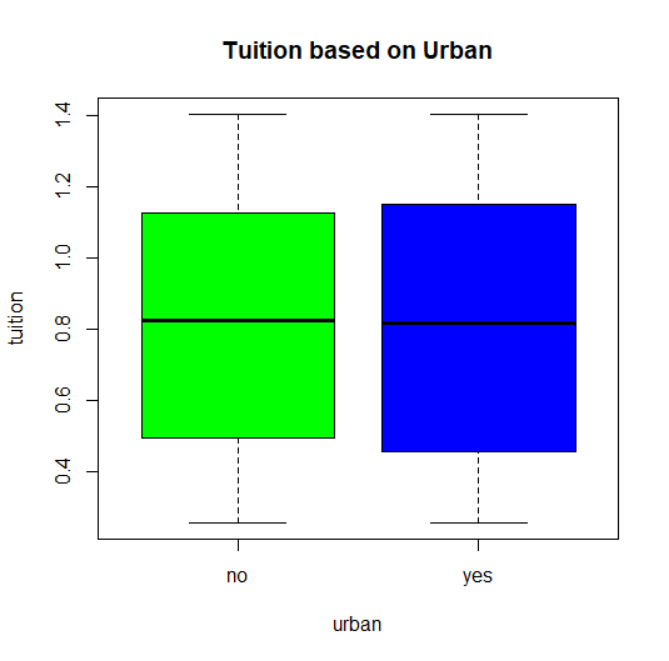


The function of the “**attach()**” is to be able to access the variables without having to use the ‘$’ sign in front of them which is more convenient for us to handle such many variables, and I tried to test if there’s any missing values in the dataset and it ended up with “FLASE(NO)”, meaning that we don’t need to clean the dataset, in order to ensure that the ‘NA’ was not replaced by some other signs like ‘**\***’ or ‘**?**’, “**sum(!complete.cases((data))**” is to determine how many observations contain at least one special signs, and the result means that the dataset is complete. For the variable of “ethnicity”, I used “**table()**” to view the contingency table of the counts and result returns statistical numbers of each type of this variable.

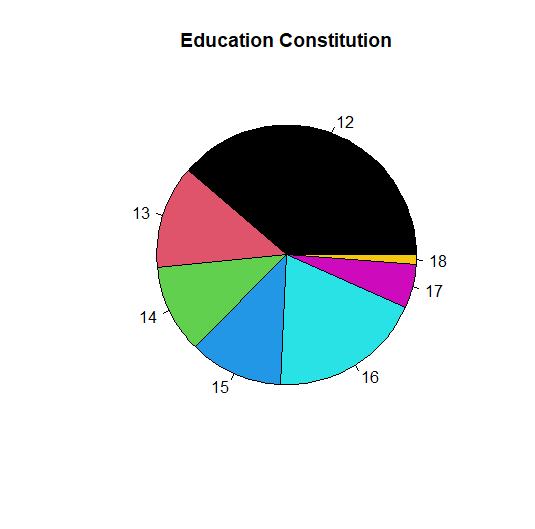






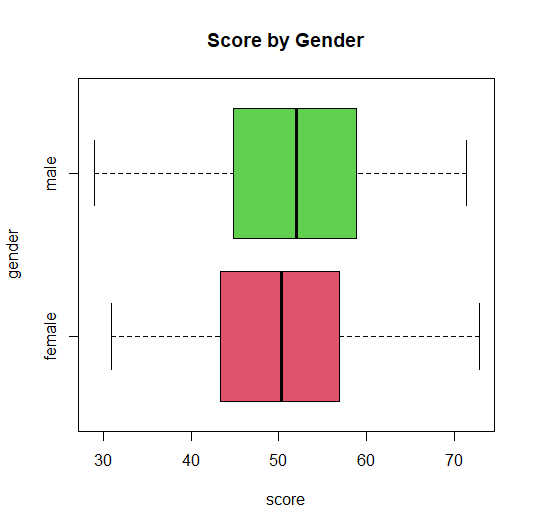




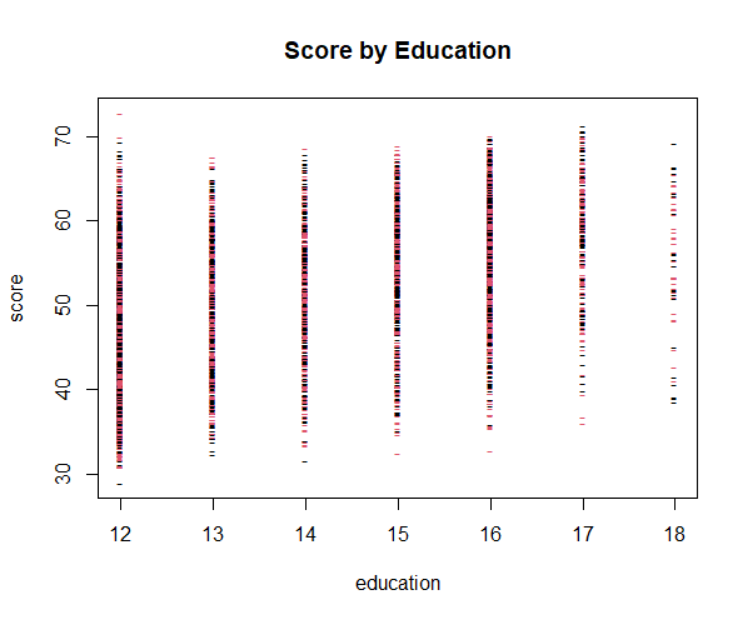




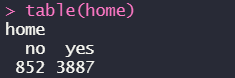


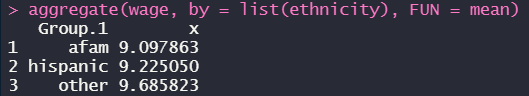






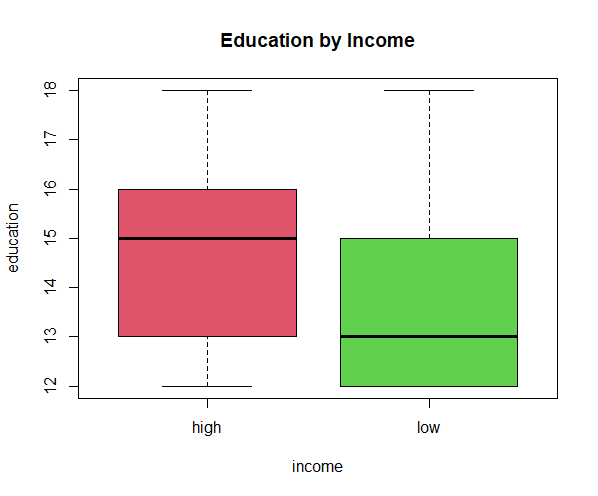
I plotted five graphs based on different variables ,the first one is the histogram of distance and from which we can see that the distance between 0 and 3 takes up the most proportion of the whole dataset, the second one is a boxplot of tuition by urban, obviously the tuition doesn’t change much with urban of not, the third one is a piechart of education and amongst it, the 12 years of educations take up the most and I used the “**sum()**” to get the number of people whose education is 12 years which is 1832, the fourth one is also a boxplot but I set it as horizontal, we can see that the overall trend of male is relatively higher than female’s, and the last one is a scatterplot of Score by Education, the mark was changed by ‘-‘ and the color is red and black, from the picture we can see that the score distribution of 17 and 18 is relatively higher than others.



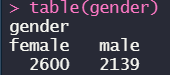


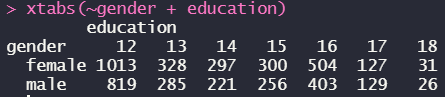
From the “table()” function that we used we can see that most of the people in the dataset owned a home, the “aggregate()” function can be used to compute summary statistics for each group and the function will be based on the one that we specified, for example, I applied “mean” to the FUN which means that I wanna get the subset of wage by the list of ethnicity.





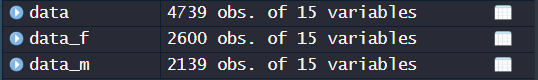
The boxplot of Education by Income indicated that people with high education could earn higher income than people with a plain education



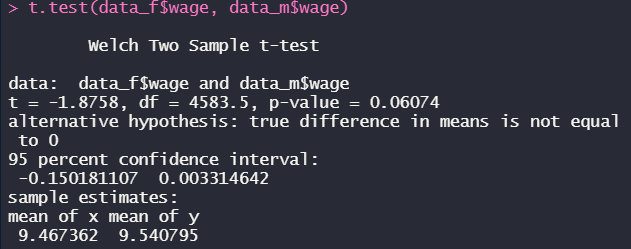


By using the “xtabs()” function, I created a cross-tabulation for male and female based on the years of education they got.

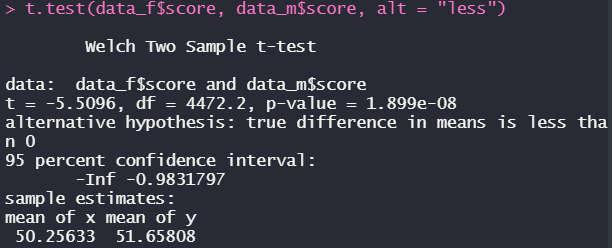




In order to prepare for the following tests by Male and Female, I made different subsets for them and from above, we can see that the sum of dimensional is the same as the original dataset.



I was trying to use “t.test()” to determine if there’s a significance difference between the wage of male and female, and from the p-value we can see that it’s greater than significance level which means that we can’t say there’s a significance difference between them.



Then I used another “t.test()” to find out if the female’s mean score is less than that of the male’s, and this time is p-value is extremely tiny which is in favor of the alternative hypothesis that the mean of female is exactly relatively less.

*Conclusions:*

For the midterm project I applied almost all of the key points that I’ve learned through this courser as of now, the dataset is pretty valuable and contains multiple variable information of boolean and numeric which is pretty convenient for me to plot, make some subsets and do some comparison tests. As far as it goes, there usually aren’t a lot of variables in a dataset we need to deal with, in this case then some other variables will be wasted or invaluable for the data analysis, so I do hope that as we learn more and more in this field, we’ll have the chance to deal with overall variables in the dataset or to at least can set up more connections or relations between them, to make all of the variables have correlations with each other and provide us more valuable and essential information for our investigation and that is exactly what I expect to happen in the future.