MIE1624 Assignment 1 Report

Ziqingqing Ye 1009167544

October 15, 2022

The objective of this assignment is to explore the survey data to understand the nature of women's representation in Data Science and Machine Learning and the effects of education on income level.

Question 1

In question 1, an exploratory data analysis was presented to analyze the data set. So here I chose three factors: age (Q1), gender (Q2), the country that participants currently reside (Q3) and the highest level of formal education the participants have attained or plan to attain within the next 2 years (Q4) to examine the relationships between these characteristics with the annual allowance (Q25).

Firstly, from the summary table salaries of different education backgrounds, in general, the mean salary increases with the improvement of education background. From this histogram (Figure 1), the mean annual allowance of people with doctorate degrees is about 70,000 USD, higher than that of participants with other degrees.

Then, from the salary summary table for different age groups, the mean salary increases with age. A boxplot between annual salary and different age groups is plotted (Figure 2), the standard deviation is very high which means situations are different from case to case. Besides, figure 2 is right-skewed, most people have annual income less than 200 thousand, but there are a few people that can earn more than a million a year.

I also plotted a bar plot between the women's average annual salary and the country they currently reside. As shown in Figure 3, the average annual salary of female data scientists working in Vietnam is the highest, which is $125,000, and that of female data scientists working in Algeria is the least, which is $1000.

Question 2

a. The second part focuses on estimating the difference between the mean of males and females. Because there are five gender categories in the survey, only male and female categories are selected. Descriptive statistics are presented; the male participants are much more than the female participants. And according to statistics, the average salary of male participants is $51,193 and it is much higher than the average salary of woman $34,816.

b. Before performing the test, an assumption of fitness, mainly normality, is validated. The normality assumption can be tested with *the scipy.stats.shapiro()* function. The null hypothesis that the data are normally distributed is rejected for both datasets. The assumption of equal variance can be checked with *scipy.stats.levene()* and the result is: homogeneity of variance is rejected. Therefore, the assumptions of normality do not apply. A 2-sample t-test cannot be performed at this time.

c. The bootstrap method is a way of resampling the original sample and creating dummy samples. Both male and female datasets are plucked 1000 times. The *np.random.choice()* method is used. The sample mean is recorded for each replicate, and the mean distribution for males and females and the distribution of the mean difference are shown in Figure 4 and 5. From the figures it can be seen that both the male and female have a normal mean distribution and their mean difference also has a normal distribution.

d. A two-sample t-test is performed on bootstrapped data. First, we need if the assumptions are met. The method used is the same as before. The assumption of normality is not rejected, but the third assumption of equal variance is rejected. The t-test can still be performed due to incomplete statistics. The t-test can still be performed because *the scipy.stats.ttest_ind()* function, we can perform Welch's t-test, which does not assume equal variance. After running the test, the p-value is equal to 0, which is less than 0.05, the null hypothesis is rejected. The result of the Welch's t-test shows that there is a statistically significant difference in the average salary between men and women.

e. The bootstrapped mean distribution plots show that the pay gap between men and women is quite large. The mean salary gap graph shows that the salary difference between the two groups is approximately $16,000. The average salary for women is around $35,000, the difference between the two groups is almost 40% of the average salary for women.

Question 3

a. The third part focuses on the impact of education on income levels, three groups (bachelor, master and PhD) are selected for analysis. First, descriptive statistics are presented for each group. Participants with a higher educational background tend to have a higher median income.

b. Since there are now three groups, ANOVA is used instead of the two-sample t-test. Before running the test, the assumptions must be verified. The same methods are used to test assumptions. It is based on the assumption of independence. All three datasets failed the normality assumption and the homogeneity of variance assumption because the p-value for both tests of the assumptions is less than the threshold of 0.05. Therefore, the ANOVA test is not performed.

c. Three groups are selected, resampling the three datasets for bachelor's, master's, and doctoral earnings relative to their size 1000 times and averaging each. In addition, the difference between two groups is calculated. Three histograms are plotted for the bootstrap data. The plots show a bell shape, suggesting that the data are now approximately normally distributed. A similar distribution also occurred with bootstrapped difference data.

d. ANOVA test performed on the three formation dates with boot. First, the assumptions of normality and homogeneous variance must be tested. The assumption of independence is assumed. The normality assumption is not rejected, but the third equal variance assumption is rejected. Since the variances are in the same order, the homogeneity of the variances is retained. A simple ANOVA test can be performed. The *scipy.stats.f_oneway()* function is used to perform the one-way ANOVA test. The ANOVA result shows that there is a statistically significant difference in mean salary between the three groups

e. From the bootstrapped average distribution plots and the descriptive statistics, it is clear that the participants with higher education level will have higher average salary if they working in data science and machine learning area.