Q1.

Figure 1. The number of participants taking the kaggle survey by country(top 10,sorted)

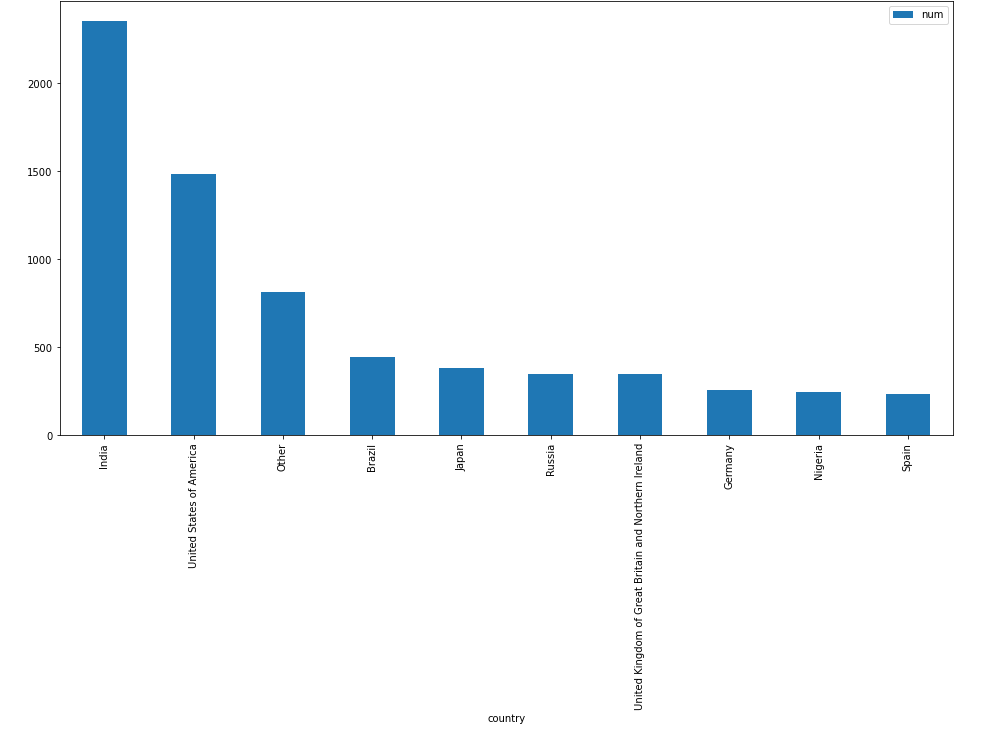


Figure 2. people taking the survey according to their gender and occupation(descending, only men and women included)

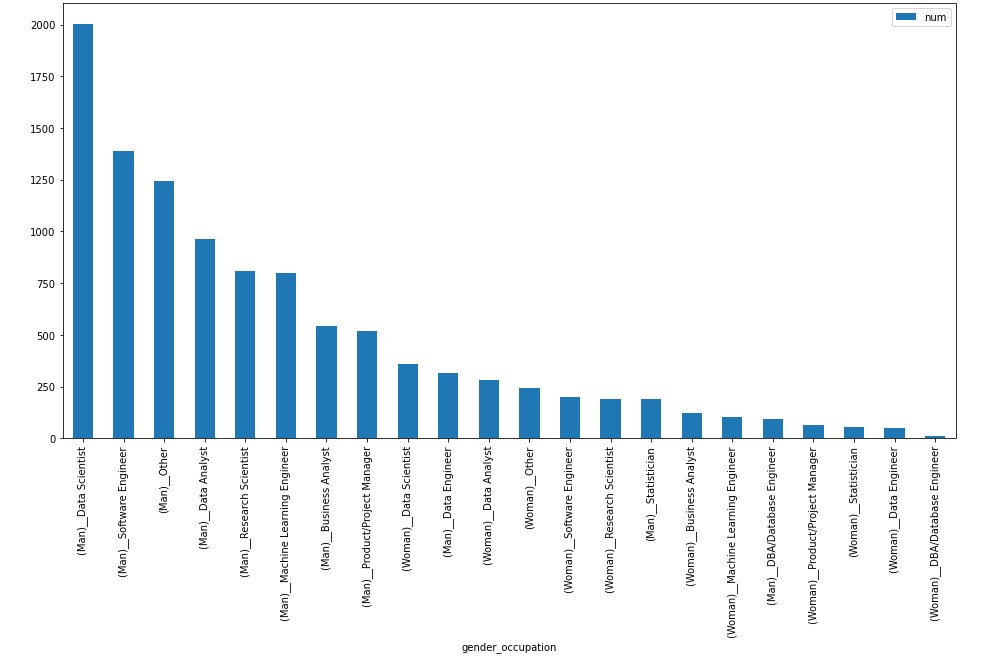
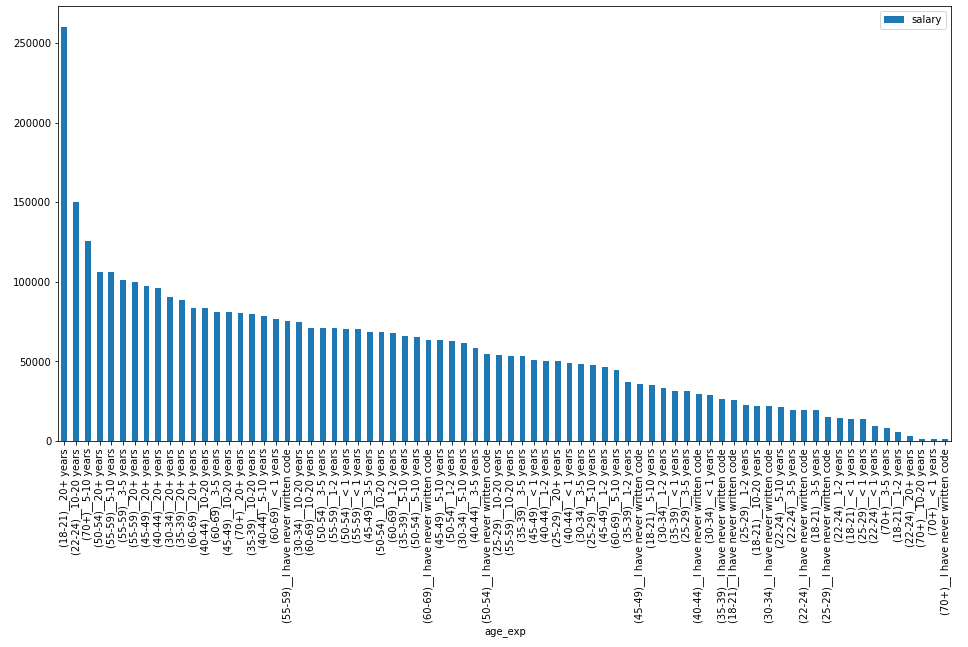


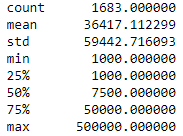
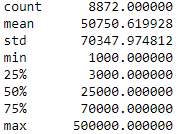
Figure 3. the relationship between [age,professional experience] and mean salary



Q2 a

Descriptive statistics for male and female salary :

Female: Male:

Q2 b

We have seen that the mean salary of men and women are different. To test if this is significant, we perform a sample t-test.

Null hypothesis: the mean salary between men and women are not different

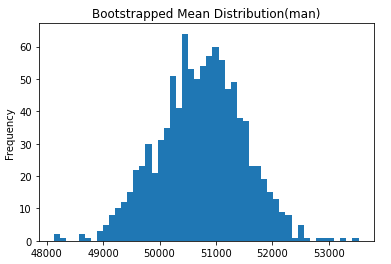
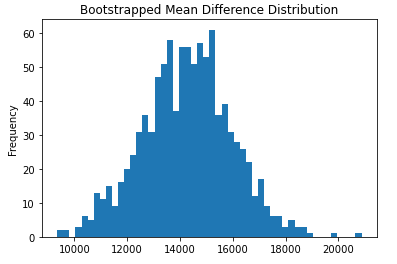
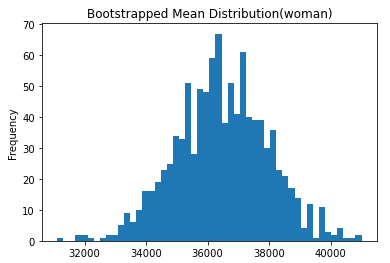
Alternative hypothesis: the mean salary of the 2 groups are different.

Our sample t-test result is as follows:

t-test: t = 7.84433 p = 4.77315e-15

In the case of salary, we reject the null hypothesis and conclude that the salary between 2 groups are different. Men tend to earn more salary (p-value = 4.8e-15<0.05, statistically significant)

Q2 c

Q2 d

We have seen that the mean salary of the bootstrapped data for men and women are quite different. To test if this is significant, we perform a sample t-test.

Null hypothesis: the mean salary between men and women are not different

Alternative hypothesis: the mean salary of the 2 groups are different.

Our sample t-test result is as follows:

t-test: t = -266.869 p = 0

In the case of salary, we reject the null hypothesis and conclude that the salary between 2 groups are different. Men tend to earn more salary (p-value <0.05, statistically significant)

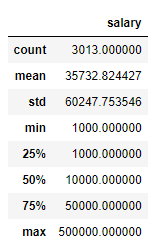
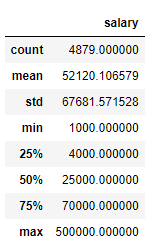
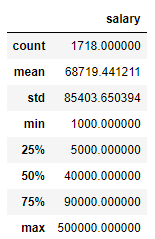
Q2 e

First, the central limit theorem holds for our bootstrapped data because if you take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed. A sample t-test is a type of inferential statistic used to see if there is a significant difference between the means of two groups.We have performed 2 sample t-test on both the original sample and the bootstrapped sample to see if the difference in the mean of the salary for men and women is significant. From the 2 test, we all reject the null hypothesis and conclude that this difference is significant. Looking at the descriptive statistics, we can manually see that this difference is actually quite obvious. The mean, 50th and 75th percentile data differs quite a lot for the 3 groups. So the test result is quite plausible. On the other hand, I observed that the gender of the participants who participated in the survey is quite imbalanced. The male participants are nearly 5.5 times the number of female participants. This imbalance may indicate that there is a lack of female participants data for our reference. And if we have a more balanced sample, we may reach a different conclusion.

\*\*\*I will call “the highest level of formal education that you have attained or plan to attain within the next 2 years” simply as ‘degree’ in the following questions to save a little space.

Q3 a

Bachelor: Master: Doctor:

Q3 b

We have seen that the mean salary between bachelor, master’s and doctoral degree holders are different. To test if this is significant, we perform a ANOVA test, which is analysis of variance.

Null hypothesis: the mean salary between bachelor, master’s and doctoral degree holders are not different

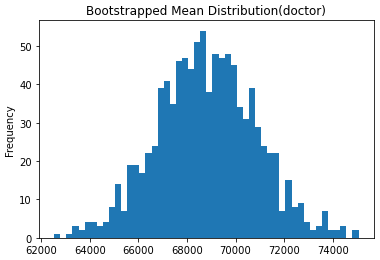
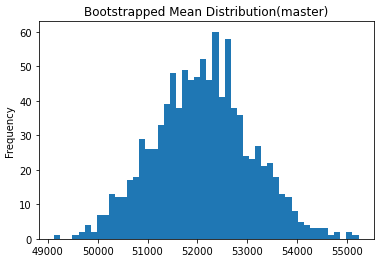
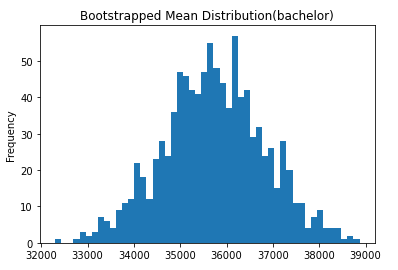
Alternative hypothesis: the mean salary of the 3 types of degree holders are different.

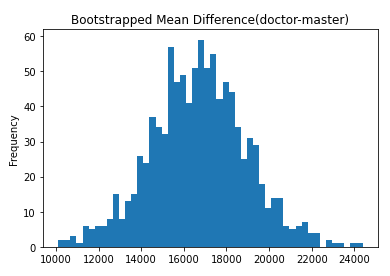
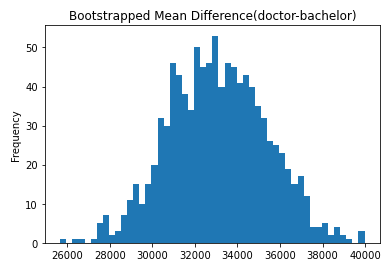
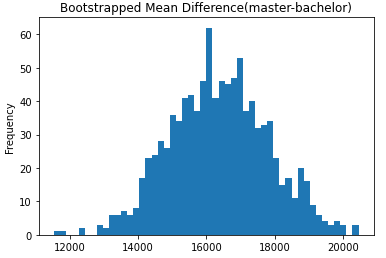
Our ANOVA result is as follows:

F\_onewayResult(statistic=129.7560112960932, pvalue=2.4852074227874282e-56)

In the case of salary, we reject the null hypothesis and conclude that the salary between the 3 type of degree holders are different. Doctoral degree holders tend to earn more salary (p-value = 2.5e-56<0.05, statistically significant)

Q3 c





Q3 d

We have seen in the the mean salary of the bootstrapped data for people are bachelor, master’s or doctoral degree holders are quite different. To test if this is significant, we perform a ANOVA test.

Null hypothesis: the mean salary between bachelor, master’s and doctoral degree holders are not different

Alternative hypothesis: the mean salary of the 3 types of degree holders are different.

Our ANOVA test result is as follows:

F\_onewayResult(statistic=130265.60474636256, pvalue=0.0)

In the case of salary, we reject the null hypothesis and conclude that the salary between the 3 groups are different. Doctoral tend to earn more salary (p-value <0.05,statistically significant)

Q2 e

First, the central limit theorem holds for our bootstrapped samples because if you take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. ANOVA test is used to check if the means of two or more groups are significantly different from each other. Specifically we choose to apply the one-way ANOVA test instead of other types like two-way ANOVA test because a one-way ANOVA test only uses one factor or independent variable while for a two-way ANOVA test, there are two independent variables. Specifically, we only have one independent variable which is the highest formal education that people attained or plan to attain in the next 2 years(degree). And this independent variable have 3 categorical groups. We do not use sample t-test here because it is a test that determines if there is a significant difference between the means of two groups and we have three groups here. We have performed 2 ANOVA test on both the original sample and the bootstrapped sample to see if the difference in the mean of the salary for the 3 groups is significant. From the 2 test, we all reject the null hypothesis and conclude that this difference is significant. Looking at the descriptive statistics, we can manually see that this difference is actually quite obvious. The mean, 50th and 75th percentile data differs quite a lot. So the test result is quite plausible. On the other hand, I observed that the doctoral degree participants are quite less than the other 2 types of participants which makes the sample imbalanced. The bachelor’s degree holders are 2 times the number of doctoral degree participants and the master’s degree participants are nearly 3 times the number of doctoral degree participants. This imbalance may indicate that there is a lack of doctoral degree participants’(or maybe bachelor’s degree participants’ as well) data for our reference. And if we have a more balanced data, we may reach a different conclusion. The so-called ‘highest level of formal degree’ actually contain the situation where people have not attained their degree yet.(Q4 says: What is the highest level of formal education that you have attained or plan to attain within the next 2 years?). This may have a impact on the salary they earned. For example, although they may declare themselves as a master’s degree holder, they haven’t got it at the time they were filling the survey. Maybe they were paid less than master’s degree holders. This may lower the mean salary for that degree holder group.