Question 1

Exploratory data analysis

As the first graph shows, it is a boxplot of Age vs Salary. It is important to see how the salary changes in the career of data science. Before the age of 45-49, we can see that the median of salary has a relatively obvious increase. In comparison, after the age of 45-49, the median of salary doesn't change much and just fluctuate in a very small scale. However, as the age goes bigger, the maximum salary keeps increasing in the data science industry.

The second graph is also a boxplot, which is the Country vs Salary. It is important to see how different can data science career be paid in different countries. The box plot is sorted in a descending order by median of salary in each country. Hence, we can see that there are 6 European countries out of the 10 countries which have the highest median of salary. Also, Switzerland, USA happens to have the highest median paying around 130,000 among all the other countries. In contrast, after the middle of the graph, the paying in all countries are around 20,000 or lower. Among all those countries, one thing we need to mention for later reference is that India's median of salary is in the lower middle zone.

The third graph is a count plot which counts the number of participants in each country. It is important which country has the most participants in this survey. The graph is sorted in a descending order. Since the data is a subset of data science community, we can see that India, USA and Japan acts as the 3 biggest suppliers for this data science community. Especially for the number of participants in India, is not only is the largest among all countries, but also is almost two times bigger than the participants in the second place. However, according to the second graph, we know that the India's median of salary is in the lower middle zone among all countries.

Question 2

Descriptive statistics

For the descriptive statistics, I use the "describe()" function and histogram on different gender's salaries to analyze some basic characteristics of the data. These methods are simple, but very useful to give us profound information and a general view on the data we are dealing with.

First, according to the "describe()" function, we can the number of male participants is 12642 while the number of female participants is only 2482, which is almost one-sixth of the number of male participants. When other measurements of salary show to be close in two genders, the mean and median of the salary is a noticeable gap of 15,000. However, we cannot just rush to the conclusion of two genders have significant difference in the mean of salary by only using these two measurements. It can be solved in the later test.

Secondly, in the histograms of two gender's salaries, we can discover the same pattern, is that over 50% of participants, whether is man or woman, are paid under 125,000.

Two-sample t-test

Since the threshold is 0.05 for this two-sample t-test, the p-value is far smaller than the threshold. Thus, we can reject the null hypothesis, and say that the mean of woman's salary is significantly different from the mean of man's salary.

Bootstrap your data

Since the dataset is small, I use the modified Cochran formula to determine the bootstrapped sample size for each gender. After doing the 1000 replications, I draw the histograms above. As the first two graphs show, the highest frequency of bootstrapped mean of man's salary is around 50,000 and is around 35,000 for woman. Both are very close to the original mean of salary in both genders (51,194 and 34,817). The same thing applies on the bootstrapped difference of mean of salary between man and woman. The highest frequency is around 16,000 and is very close to original difference of mean (51,194 – 34,817 = 16377).

Two-sample t-test on bootstrapped data

Since the threshold is 0.05 for this two-sample t-test, the p-value is 0 and definitely smaller than the threshold. Thus, we can reject the null hypothesis, and say that the mean of woman's salary is significantly different from the mean of man's salary in the bootstrapped data.

Comment on findings

From the descriptive statistics to two-sample t-test on bootstrapped data, we can see the number of female participants is far smaller than the man participants. More importantly, we can see the difference in mean of salary in two genders at the beginning. As we progress, this assumption is being proved by the t-test. Also, we can find out that proper bootstrapped data can really save us time for giving the same test results while facing the large datasets.

Question 3

Descriptive statistics

For the descriptive statistics, I use the "describe()" function and histogram on different degree's salaries to analyze some basic characteristics of the data.

First, according to the "describe()" function, we can the number of participants is 4,777 for Bachelor's degree, 6,799 for Master's Degree, and 2,217 for Doctoral Degree. Each degree has an approximated difference of 2000 from the other degrees while Master's degree being the highest. Like genders, all 3 degrees shows very different number in mean of salaries with similar difference around 20,000 in the descending order of Doctoral, Master and Bachelor. However, we cannot just rush to the conclusion of 3 degrees all have significant difference in the mean of salary by only using this one measurement. It can be solved in the later test.

Secondly, in the histograms of 3 degrees salaries, we can discover the same pattern, is that over 50% of participants, no matter has which degree, are paid under 125,000. Nevertheless, the portion of people who are paid around 150,000 increases as the degree levels up.

ANOVA test

Since the threshold is 0.05 for this ANOVA test, the p-value is far smaller than the threshold. Thus, we can reject the null hypothesis, and say that the means of 3 degrees' salaries are all significantly different from each other.

Bootstrap your data

Since the dataset is small, I use the modified Cochran formula to determine the bootstrapped sample size for each degree. After doing the 1000 replications, I draw the histograms above. As the three graphs show above, the highest frequency of bootstrapped mean of salary is around 34,000 for Bachelor, around 50,000 for Mater and around 69,000 for Doctoral. All of them are very close to the original mean of salary in 3 degrees (35,578, 52,707 and 70,641).

The same thing applies on the bootstrapped difference of mean of salary among 3 degrees. The highest frequency is around 15,000 for Master and Bachelor, around 34,000 for Doctoral and Bachelor and around 15,000 for Doctoral and Bachelor. They are all very close to the difference of original means.

(Master and Bachelor: 52,707 – 35,578 = 17,129

Doctoral and Bachelor: 70,641 – 35,578 = 35063

Doctoral and Master: 70,641 – 52,707 = 17,934)

ANOVA test on bootstrapped data

Since the threshold is 0.05 for this ANOVA test, the p-value is definitely smaller than the threshold. Thus, we can reject the null hypothesis, and say that the means of 3 degrees' salaries are all significantly different from each other.

Comment on findings

From the descriptive statistics to ANOVA test on bootstrapped data, we can see the number of different degree's participants is different in a similar gap around 2,000 and ranks in the descending order of Master, Bachelor, Doctoral while the number of Master participants is 6,799 and being the mainstream in the data science industry. More importantly, we can see the difference in mean of salary in 3 different degrees at the beginning. As we progress, this assumption is being proved by the ANOVA test. Also, we can find out that proper bootstrapped data can really save us time for giving the same test results while facing the large datasets.