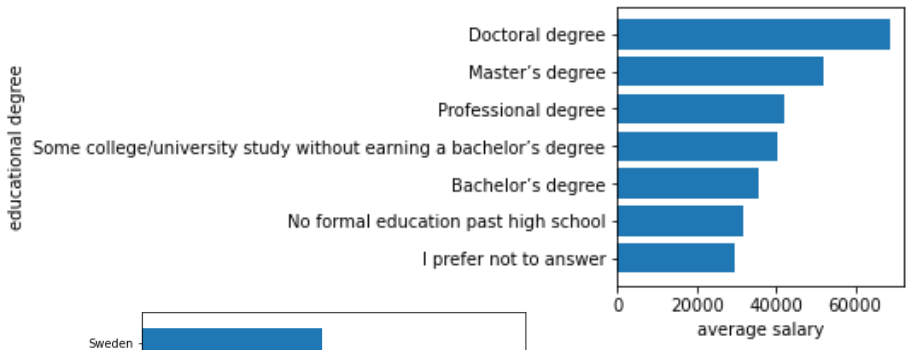
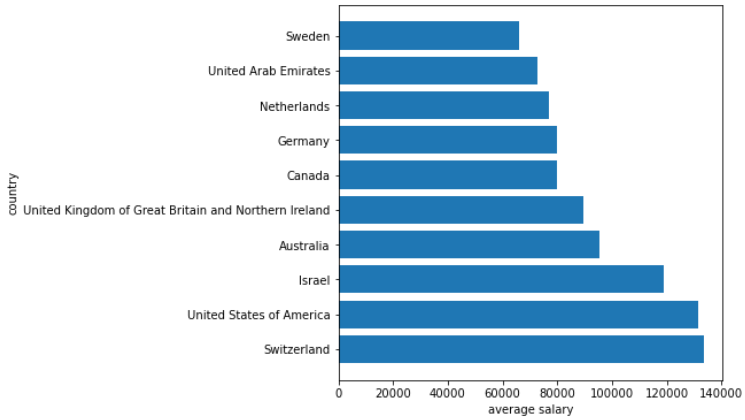
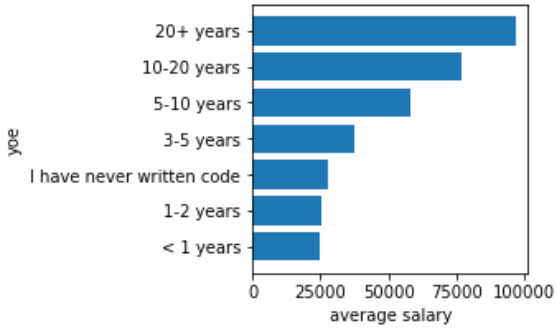
Question 1:
(one feature plot not included in pdf)
We can see geographic distribution of DS and SDE from the pie figure in notebook(not included in pdf because of space). India has the most programmers and USA follows it. We can find out the top six countries: India, US, Brazil, Japan, Russia, UK have 50% data scientists of the world. But we also find some limitation of the survey. China has the most population in the world and highly developed internet industial but only has less than 2% DS positions in the survey. We can say there are some reasons that chinese are not that willing to or able to finish the survey than people in other countries. However, we can cursely find that the geographic distribution is highly centralized which means the internet and data science development is not balanced all over the world.
From the age range figure, we can know the age distribution. Most data scientists are in the age range 25-29. We can find out that data science and software are really novel field and most of the workers are young people. Number of workers over 40 years old has a relative small proportion.

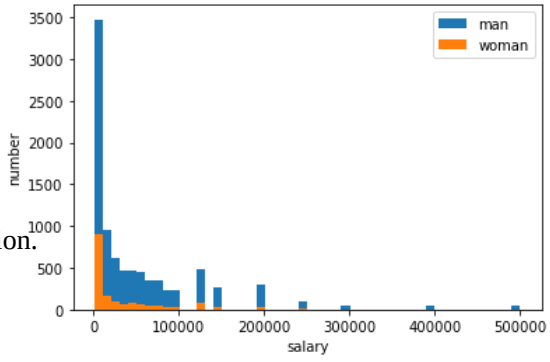
From the average salary and country figure, we can know about the data science development and treatment in different countries. We can figure out the treatment does correlate with the overall economy of the country. Countries with high salary in data science field are all highly developed countries. So the data and ML industry is a field with high technology and high profits. From the average salary and degree figure, we can find out that doctors earn the most but professional degree has higher salary than bachelor degree which means data science may requires more professional and practical knowledge. The yoe and salary figure shows that longer yoe always leads to higher salary. The experience in data science is really important.



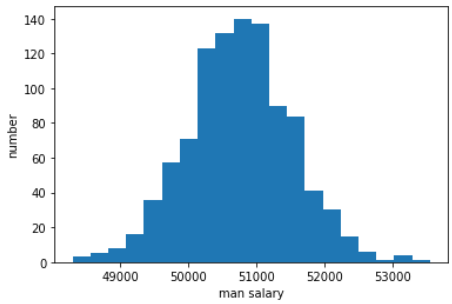


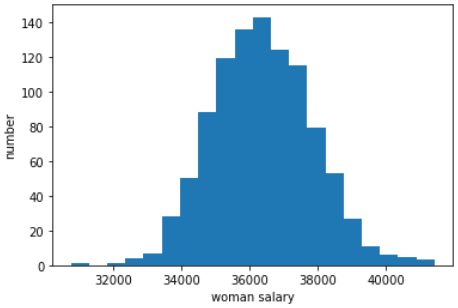


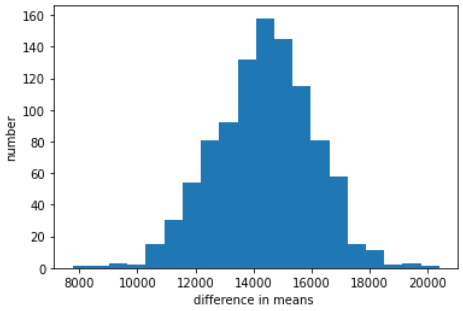
Question 2:
a: man salary: mean:50750.6 median:25000.0 mode:1000 sdv:70344.0 variance: 4948837560
woman salary: mean:36417.1 median:7500.0 mode: 1000 sdv:59425.1 variance: 3533436496
We can find that the deviation of the data is very large. Mean value or median value cannot describe the dataset well.



b: From the distribution plot we know that both man and woman salary are not normal distribution and their variances have huge differences. Hence we cannot use t-test here.
c: After bootstrapping, the distribution turn to be normal distribution or similar to normal distribution. Also, the distribution of the difference of means also obeys normal distribution.







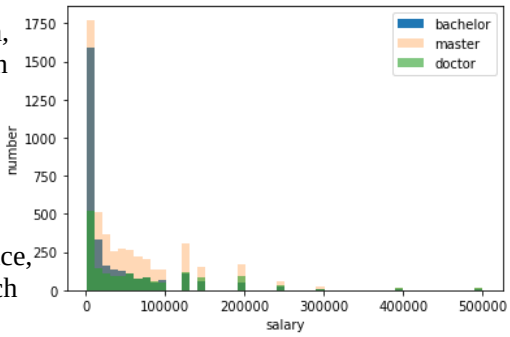
d: We already know the bootstrapped data obeys normal distribution, now we can check the variance. After computing, man salary variance is 587723 and woman salary variance is 2150368. Using levene test here to judge if there's huge difference between the variance and the p-value is 8.33e-66 which is lower than 0.05. There is a significant difference between the two variances so we still can't use normal t-test. However, we can use Welch's t-test which do not assume the equal variance. The result of t-test is 274.62 and p-value is similar to 0.0. We conclude that there's significant difference between the means of man and woman salary.

e: The analysis shows that the distributions of man and woman salary are not normal distribution and the fluctuation of data is significant which indicates there maybe other factors influencing the data. From the welch t-test, there is huge difference between the means of two set and we should refuse the null hypothesis.

Question 3:

a: mean of bachelor's, master's, doctor's salary: 35732 52120 68719
median of bachelor's, master's, doctor's salary: 10000 25000 40000.0
sdv of bachelor's, master's, doctor's salary: 60247.75 67681.57 85403.65
mode of bachelor's, master's, doctor's salary: 1000 1000 1000
variance of bachelor's, master's, doctor's salary: 3629791807 4580795124 7293783500

From those basic statistics we know that there is also large deviation in the distributions of salary of different education degree.



b: From the plot we know that the distributions are not normal distribution, and there are also significant differences between their variances for which we can't use ANOVA here to evaluate the different between their means.

c: After bootstrapping, the distributions turn to normal distributions or at least similar normal distributions. The difference of means also obey the normal distribution and difference between doctor-master and master-bachelor is similar.

d: Because ANOVA also assumes the normal distribution and equal variance, levene can also be used here. Getting the p-value equals to 7.50e-111 which is much lower than 0.05 and nearly 0. There are significant differences between the variances of bootstrapped data so ANOVA still can not be used here. Using Welch ANOVA and get the result F=120260 and p=0. There are significant differences between the means and we should refuse null hypothesis.

e: The analysis shows there are huge differences between the means of three groups and we can find it only observing the data. Levene test shows large variance differences between the data too, so we cannot use ANOVA here. Welch ANOVA told us huge differences between means still exist. Bootstrapping help the distribution but didn't help with the variance.