Analysis of Kaggle Data

The objective of this assignment is to explore the survey data to understand (1) the nature of women's representation in Data Science and Machine Learning and (2) the effects of education on income level. Therefore, I perform exploratory data analysis using plots to summarize its main characteristics. First of all, mean salaries of each level of education are calculated and display in bar plot as shown in Figure 1 in Appendix. It shows that people with highest education level, Doctoral degree, have the highest average salary, between $60000 to $70000. However, people with Bachelor's degree tend to have lower average salary compared to those who without a degree. This unexpected result illustrates that education level might not be the only factor that influences wages. Therefore, I compare boxplots of salary between man and woman for each educational level. According to the Figure 4, woman's salaries are much lower than man's, except for employee with high school diploma.

In addition, the effect of age by looking at the salary distributions within each age group is shown in Figure 2. The distributions of participants at age of 18-21 and 22-24 have large difference compare to older age group: their salary range are narrower and much lower in every quantile. Overall, salary tends to increase and spreads larger as age increases, and distribution of salaries among different age group are all right skewed since the whiskers are shorter on the lower end of the box. This demonstrates a pattern that only a small portion of people can earn at relatively higher wage, and most of people's salary are below the group average. Lastly, I further investigate national composition of the participants (Figure 5) and the pattern of professional experience (Figure 3). It shows that approximately 34% of participants in the survey comes from India and the US, and coding experience tend to have positive relationship with salary by looking at 25 and 50 quantiles of the boxplots.

Next, I estimate the difference between average salary (Q25) of men vs. women (Q2). The descriptive statistics for each group are shown below in Figure 6: The number of male participants is 5 times larger than the number of female and the average salary of males is $51193.6 whereas average salary of female is only $34816.9. As their difference in salary are large, a two-sample t-test is performed in order to test if the difference between unknown population means of man and woman are statistically significant. The null hypothesis is that population means of two groups are the same with alternative hypothesis to be different. Since we have a fairly large dataset, normality assumption is not necessary. At significant level of 5%, I obtained a t-test statistic -7.77406, with p-value equals to 8.08881e-15. As p-value of t-test is very close to zero and less than 0.05, I reject the null hypothesis at 5% threshold. Thus, it can be concluded that population means of man's and woman's salary are significantly different.

Nevertheless, as mentioned above, there're only 2482 female participants in the data, and the distribution of salary for woman has heavy tail which would reduce the power of the test. We might consider increase the number of observations by resampling in order to construct good estimates of standard error and confidence intervals using central limit theorem. Therefore, by making the number of instances sampled from each group

relative to its size and conducting 1000 replications, plots of two bootstrapped distributions (for men and women) and the distribution of the difference in means are shown in Figure 7. As expected, the distributions of mean salary for men and women becomes roughly Normal. Also, the plot of difference in means indicating that we are 100% confident that difference does not equal to 0 since all samples fall on the region that is greater than zero. The reason why I choose bootstrapping is that it does not require any assumptions about the data and it is simple and straight-forward avoiding the cost of repeating the survey to get other groups of sample data. After bootstrapping, we can conduct two-sample t-test again. With a 0.05 threshold on the bootstrapped data for men and women, I obtained a t-test statistic -304.2 with p-value equals 0. Thus, I reject the null hypothesis of equal mean and conclude that the difference in mean income for men and women are statistically significant.

Furthermore, I analyze the difference among average salary (Q25) of highest level of formal education(Q4). Since professional doctoral and doctoral degree in our data are the same, I combine them naming doctoral degree. The descriptive statistics for bachelor's, master's and doctoral degree are shown below in Figure 8. Notice that mean incomes among these 3 educational levels are differ by more than $10000 USD. Thus, I perform ANOVA test to check for significance. Instead of t-test, ANOVA test is appropriate in this case because we have 3 levels rather than 2 levels. The F-statistics is 112.1446910165994 with a p-value equals to 4.786585024917043e-49. Since p-value is less than 0.05, we reject the null hypothesis that mean incomes for people with bachelor, master and doctoral degree are the same. In summary, at least one mean income is significantly different from the others. In order to estimate population mean and perform hypothesis testing checking differences among groups, I conduct bootstrapping on the data. The plots of bootstrapped distributions are in Figure 9 and they all looks normally distributed. Next, with a 0.05 threshold on the bootstrapped data for bachelor, master, doctoral degree, the ANOVA f-statistic is 112526.1826148391 with P-Value 0. Thus, I reject the null hypothesis of equal mean and conclude that the difference among above three level of education is statistically significant.

However, we might need to confirm that our original sample data is a good representative from the population, otherwise, bootstrap may fail. In our case, we should be cautious about it because almost half portion of the participants in out survey comes from India and the US. The trend and patterns about age, gender, educational level on salary might not be a good representation for the world as a whole.

Appendix

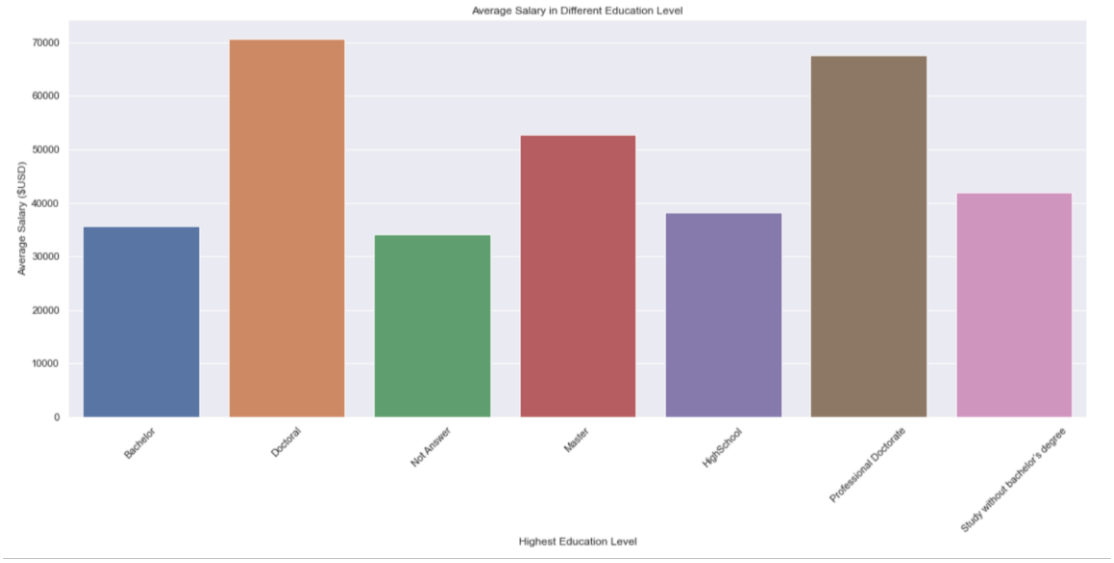


Figure 1

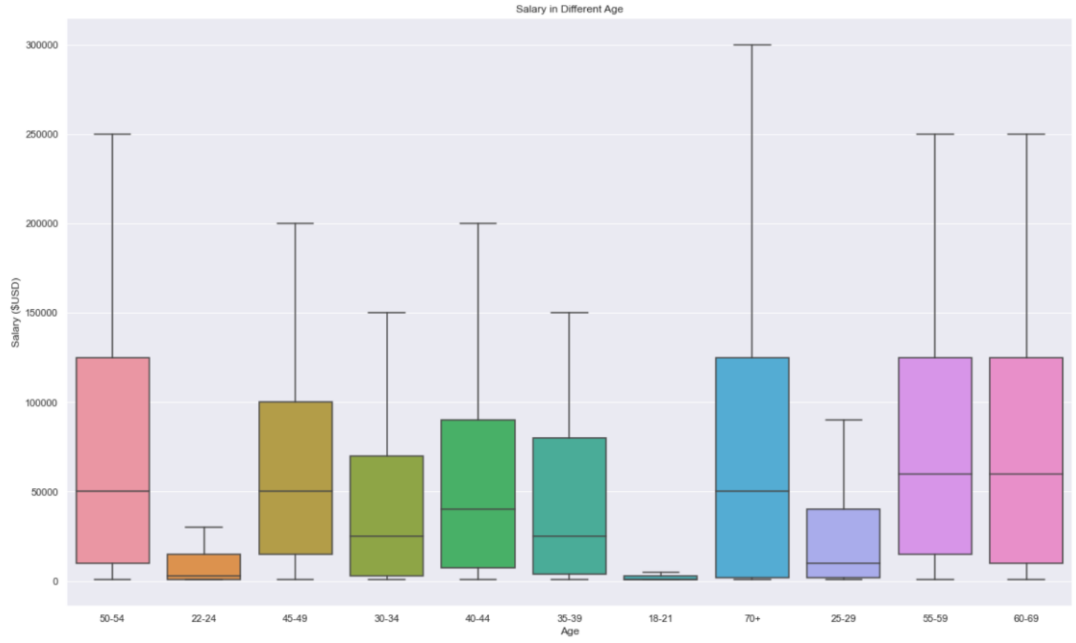


Figure 2

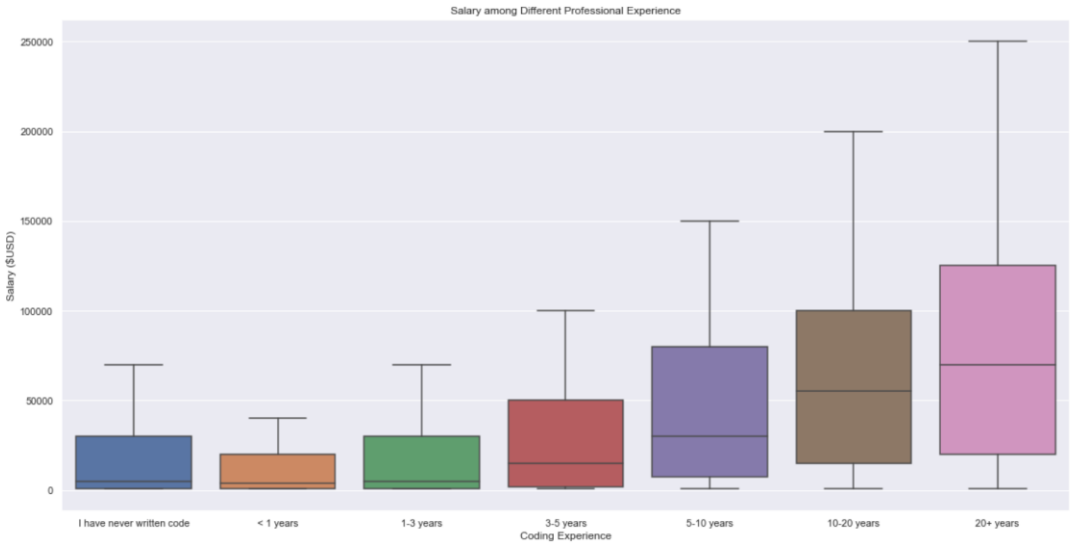


Figure 3

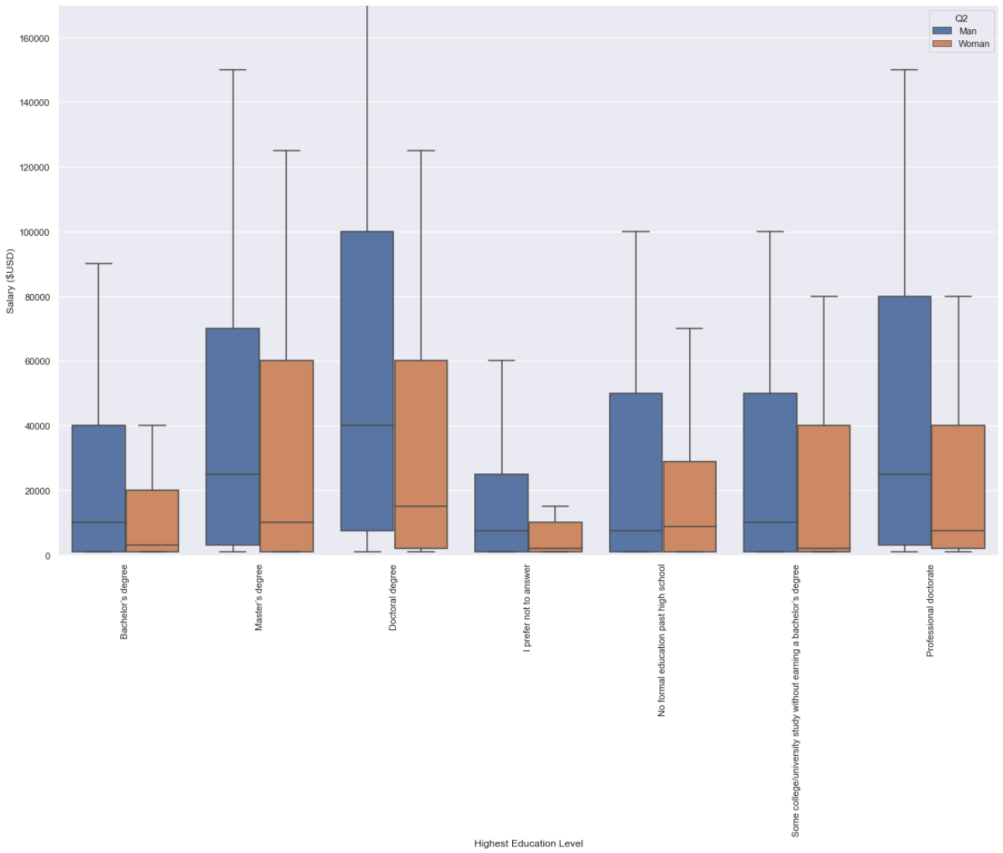


Figure 4

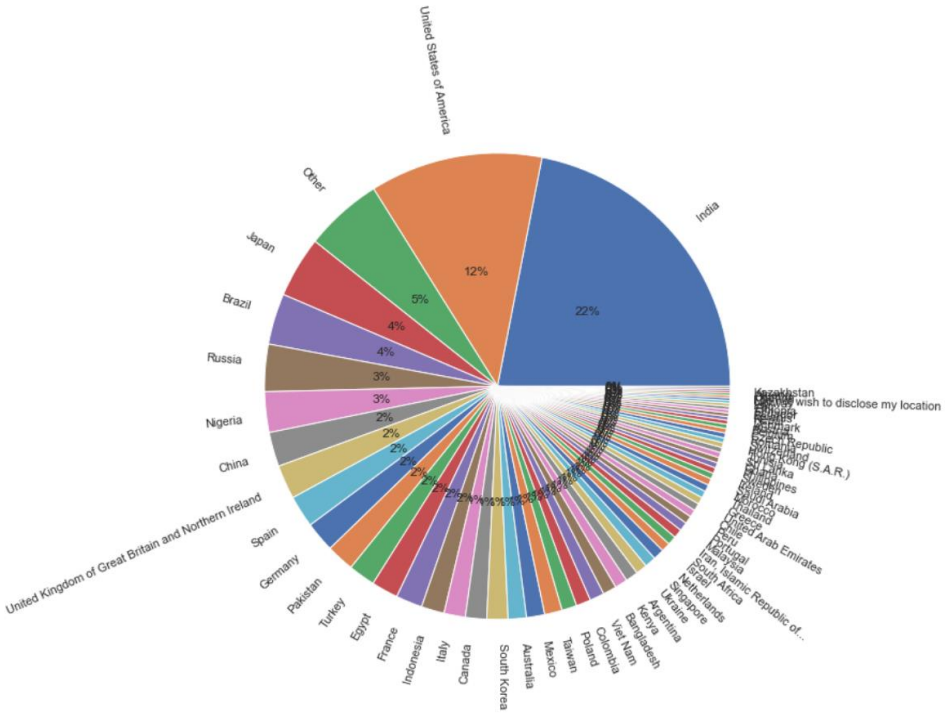
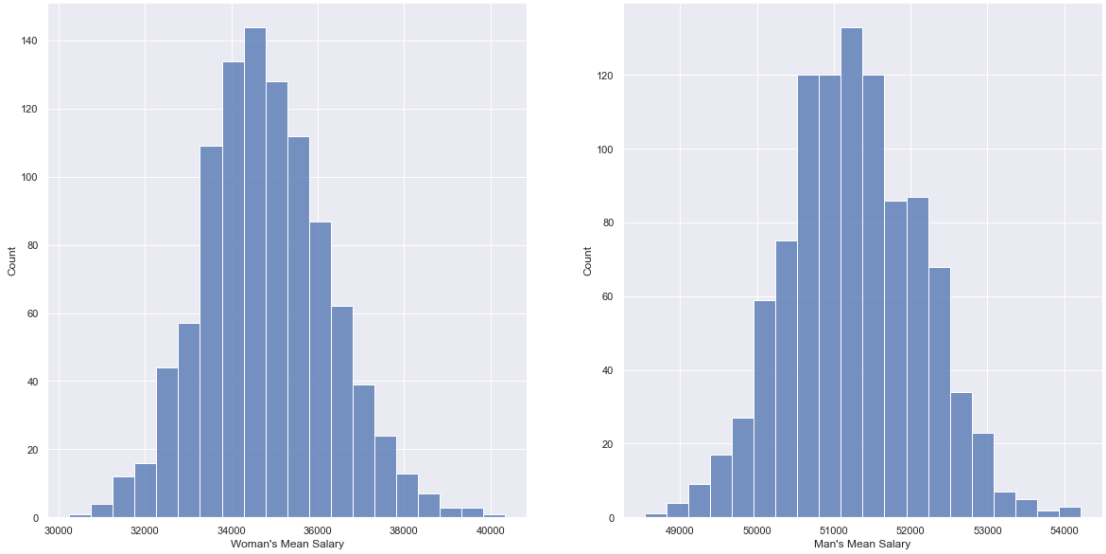


Figure 5

	count	mean	std	min	25%	50%	75%	max
Male	12642.0	51193.600696	99979.274378	1000.0	2000.0	20000.0	60000.0	1000000.0
Female	2482.0	34816.881547	72017.347888	1000.0	1000.0	7500.0	50000.0	1000000.0

Figure 6



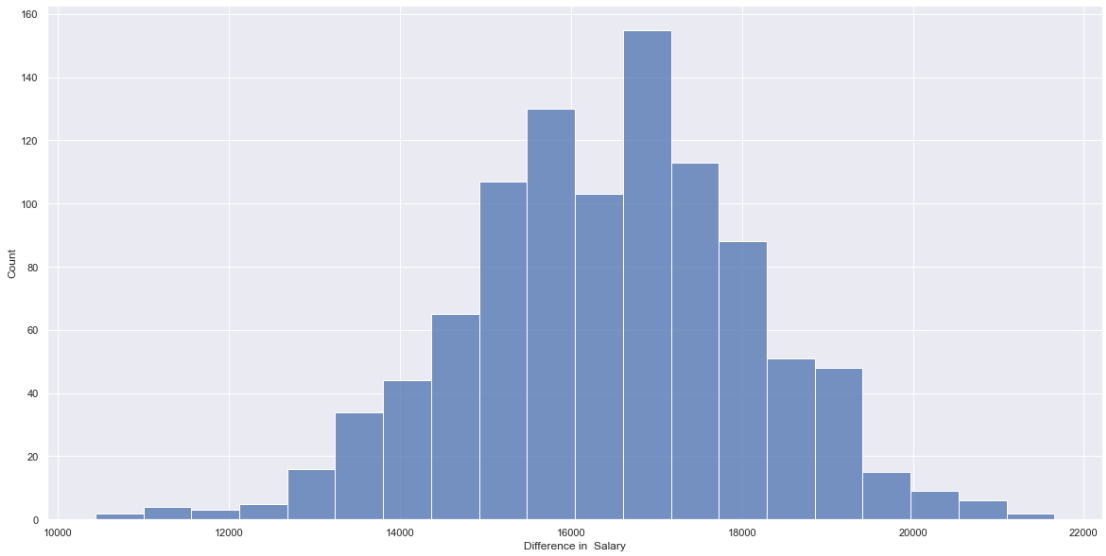
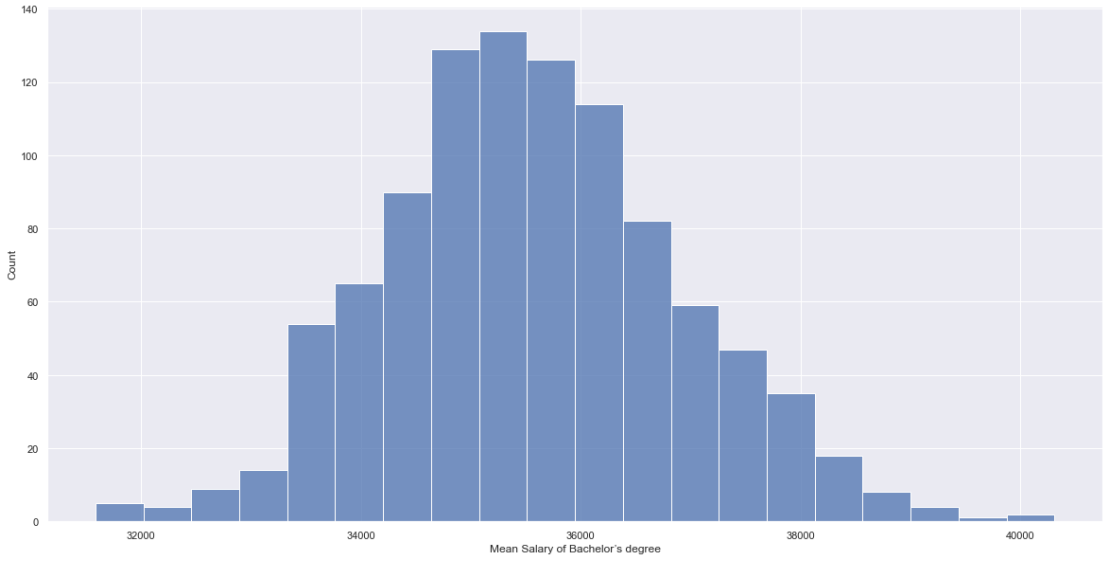
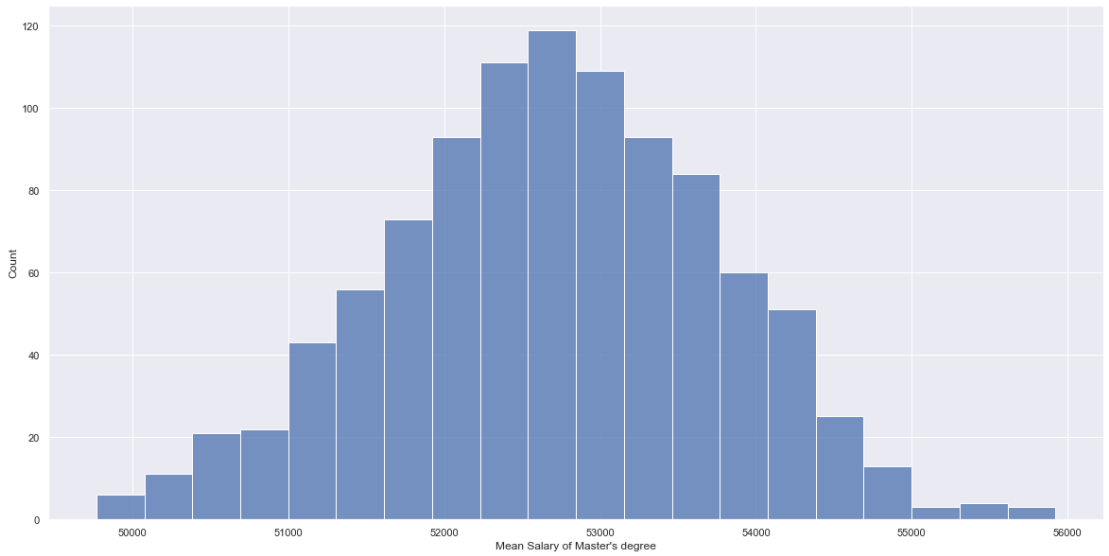


Figure 7

	count	mean	std	min	25%	50%	75%	max
Bachelor	4777.0	35578.291815	89382.060777	1000.0	1000.0	7500.0	40000.0	1000000.0
Master	6799.0	52706.868657	90928.786678	1000.0	3000.0	25000.0	70000.0	1000000.0
Doctoral	2507.0	70273.833267	119561.065858	1000.0	4000.0	30000.0	90000.0	1000000.0

Figure 8




120
100
80
60
40
20
0
Count
50000
51000
52000
53000
54000
55000
56000
Mean Salary of Master's degree

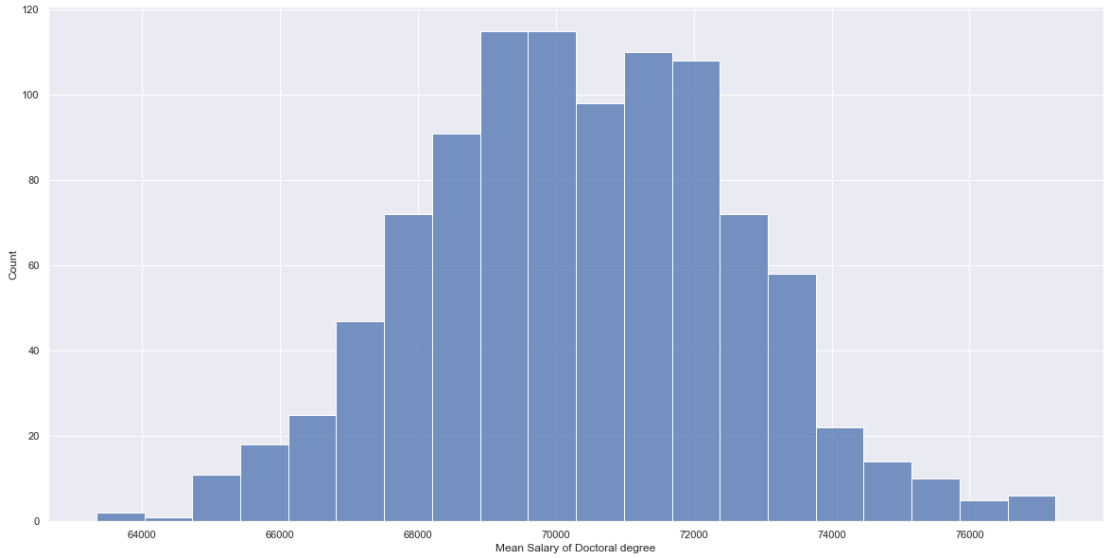

120
100
80
60
40
20
0
Count
64000
66000
68000
70000
72000
74000
76000
Mean Salary of Doctoral degree

Figure 9