MIE 1624 Introduction to Data Science and Analytics – Fall 2021
Assignment 1 - “Kaggle ML & DS Survey Challenge _ 2020 Responses”

Student Name: Jingwen Zhu
Student ID: 1003008229
Due Date: 11:59 pm, Oct 13, 2021

This report is prepared as a summary of the analysis findings of the 2020 Kaggle Machine Learning & Data Science Survey Responses dataset. For this assignment, the clean_kaggle_data.csv file was provided, which contains survey results in 355 columns from 10729 participants (rows with null responses in the salary column were dropped). Column Q24 “What is your current yearly compensation (approximate $USD)?” was selected as a target variable for the analysis. The analysis followed a workflow starting with understanding the spreadability and variance in the dataset, then towards the focus of gender differences on salaries, and the effect of education level on salaries.

Question 1 - Exploratory analysis of survey data and the visualization of its characteristics

Country, Age, and Education were the focus of my exploratory analysis. The top 10 countries for which the participants residing were shown in Figure 1. From the plot, we can see that India has the most number of participants, with 2553 participants, followed by the US contributing to1484 participants. This is probably related to the large size of the population of those two countries. Also, I have seen that lots of data analysts, researchers, and data scientists are from India, indicating that the country has a great focus on mathematics and science.

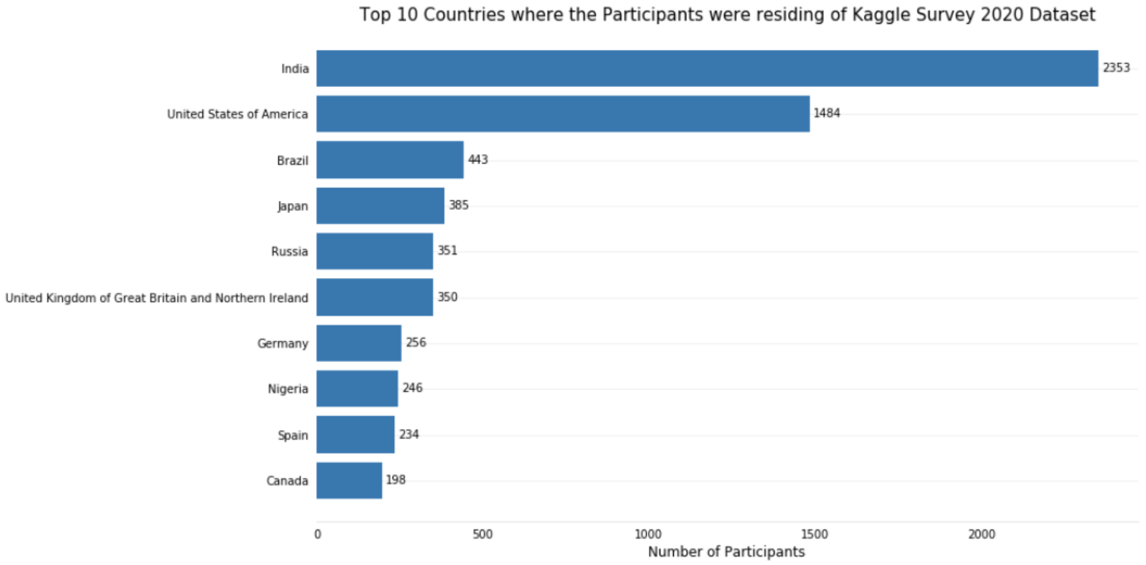


Figure 1. Top 10 countries where the participants were residing

Since one of the objectives of the analysis is to understand the nature of women’s representation in DS & ML, I would like to start with knowing how the dataset was distributed among different gender groups. As shown by Figure 2, 82.7% of the participants are male, the female group only contributed to 15.7% of the dataset; meanwhile, participants who chose to neither identify themselves as male nor female are grouped as “Other”.

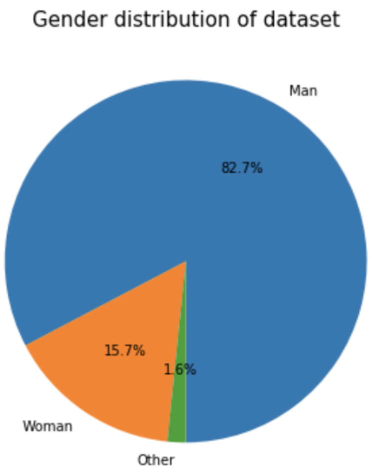


Figure 2. Gender distribution of the dataset

The number of participants over a variety of age groups and the highest level of education groups is plotted, as shown in Figures 3 & 4 below. Taking into consideration the objective of this assignment, the gender distribution in each age and education group is provided, in order to understand the nature of women's representation in DS & ML.

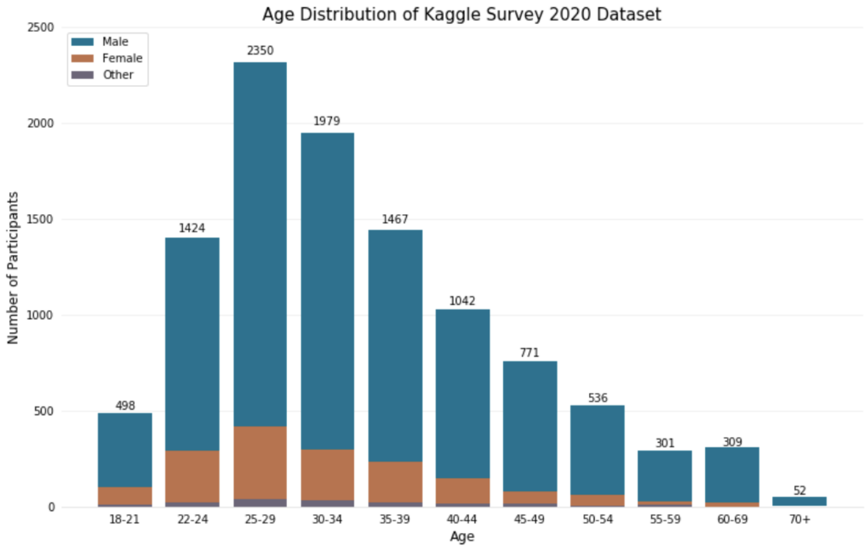


Figure 3. Number of participants for different age groups

As shown by the age plot, most of the participants were in their 20s to 40s, with 1424 participants in the age range of 22 to 24, 2350 participants in the age range of 25 to 29, and 1979 participants in the age range of 30-34. The male group represents the majority of the participants, and the proportionality difference between male and female groups is significant.

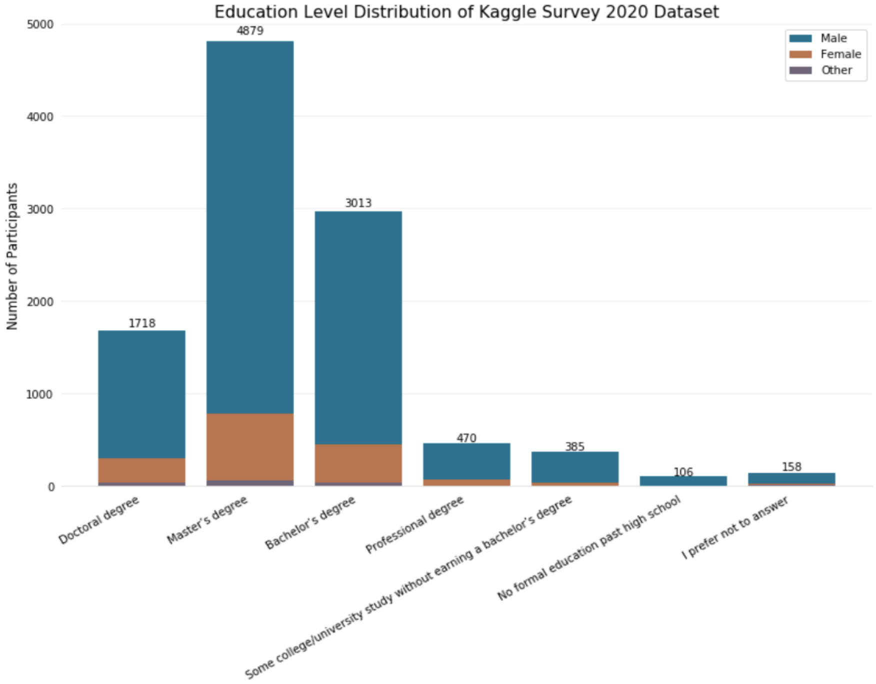


Figure 4. Number of participants for different groups of highest education level

With respect to education level, 4879 participants had a Master's Degree - which is approximately half the number of the total participants, while approximately 3000 participants had a Bachelor's degree, and around 1800 people had a Doctoral degree. It shows that the data science community, in general, requires a higher level of education. The proportion of male and female groups with respect to each education level does not vary much from the overall gender distribution with other characteristics in the dataset.

Question 2 - Estimating the difference between the average salary (Q24) of men vs. women (Q2)

a) Statistic analysis of the average salary with respect to the gender group

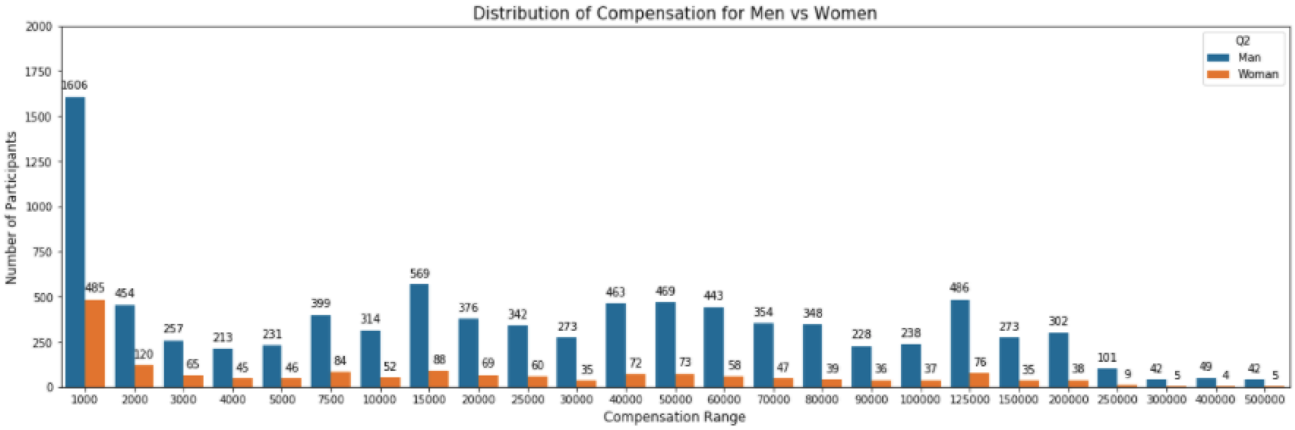
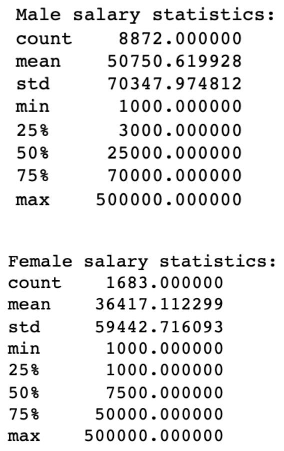


Figure 5. Number of males and females in each average salary range

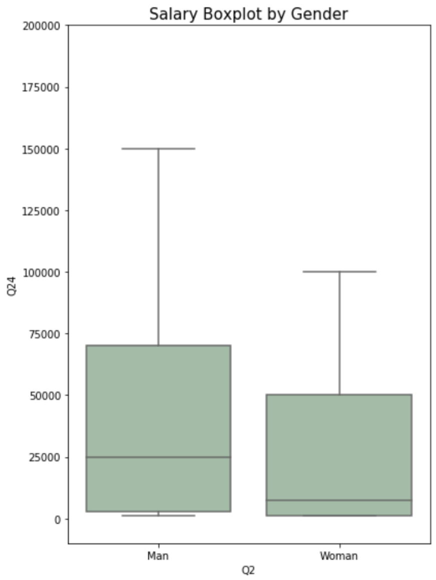
The figure above shows the number of males and females in each average salary entry. Approximately one-fourth (2606 out of 8872 male participants) of the male group had an annual salary below $1,000 USD, while few peaks were seen at $15,000, $40,000 to $60,000, and $125,000 sections. In terms of the female group, most participants had an annual salary below $1,000, and a decreasing trend was shown with the increased salaries.

The descriptive statistics for male and female salaries were computed and a separate boxplot was generated to visualize the median salary and variance of salaries for each gender group. The average salary among 8,872 male participants was 50,751 $USD, with a standard deviation of 70,348 $USD; while the average salary for the female group was 36,417 $USD with a standard deviation of 59,443 $USD. The significant SDs were resulting from the heavily skewed salary ranges, therefore looking at the medians should provide more representative information. From the boxplot, we noticed that there is greater variability in the male group salaries, and the salary median (25,000 $USD) for the male group is noticeably higher than that of the female group (7,500 $USD).



```
Male salary statistics:
count      8872.000000
mean      50750.619928
std       70347.974812
min        1000.000000
25%        3000.000000
50%       25000.000000
75%       70000.000000
max      500000.000000

Female salary statistics:
count     1683.000000
mean     36417.112299
std      59442.716093
min       1000.000000
25%       1000.000000
50%       7500.000000
75%      50000.000000
max     500000.000000
```



b) Perform t-test for hypothesis analysis (if applicable)

Since the entire population is unknown, we could not make the assumption that the population is normally distributed and the male and female samples were collected from a representative, randomly selected portion of the total population, therefore t-test is not applicable here to assess the difference between male and female average salary.

c) Men vs. Women Bootstrap Means

The male and female salaries were saved into separate dataframes and bootstrap sampling was performed with 1000 replications using *pandas.DataFrame.sample()*. The parameters were specified as *n* = *len(df)* and *replace* = *True*. In each iteration, the salary mean was computed and the difference between two means was calculated. Bootstrapping treats the sample as if it was the original population and simulates the random sampling processes to quantify the estimate of means. The distribution plot of bootstrapped sample means and the mean differences are shown in Figure 6 & 7 below.

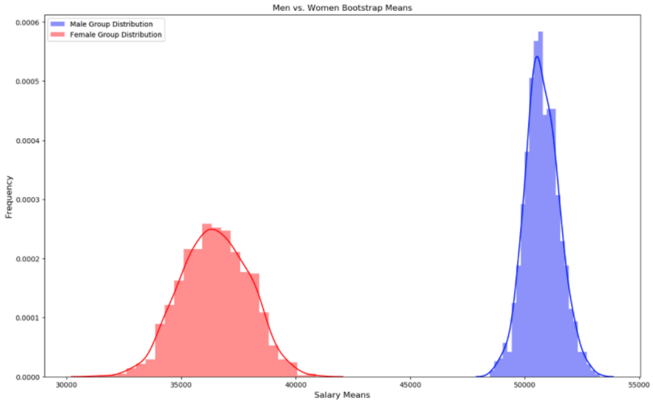


Figure 6. Bootstrapped Salary Means

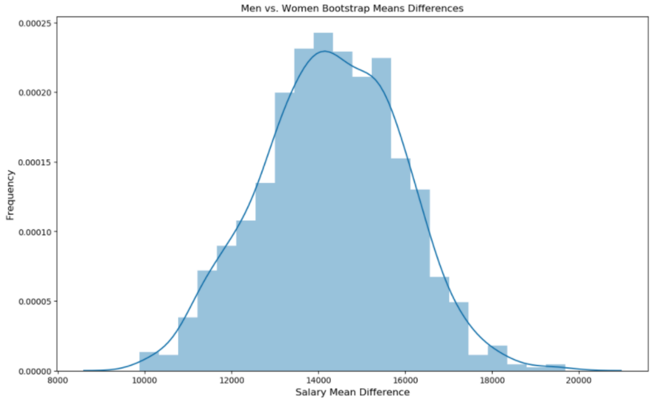


Figure 7. Bootstrapped Salary Mean Differences

d) Hypothesis testing:

According to the distribution plot, the sampling distribution of the means and mean differences approach a normal distribution, and this was in accordance with the Central Limit Theorem. The t-test was then applied to the bootstrap sampling means with a threshold significance level of 0.05 to test the null hypothesis. The test result was obtained as t = 275.031, p = 0. With a p-value equal to 0 (<0.05), it was concluded that the difference between the average salary of males and females is statistically significant, and the null hypothesis is thus rejected.

e) Comments on findings:

By computing the statistics of the sampling means, the point estimate of difference of male and female average salary was obtained as 14376.89 $USD, and the 95.0 % confidence interval for the difference was found as (11395.34, 17409.66) in $USD. Overall, the yearly salary in the data science community according to the male and female gender groups, is estimated as 50740.30 $USD and 36363.4 $USD, respectively - higher yearly compensation is observed with the male group according to 2020 Kaggle ML&DS survey.

Question 3 - Compare the means of salary between Bachelor’s degree, Doctoral degree, and Master’s degree participants

a) Statistic analysis of the average salary with respect to the education group

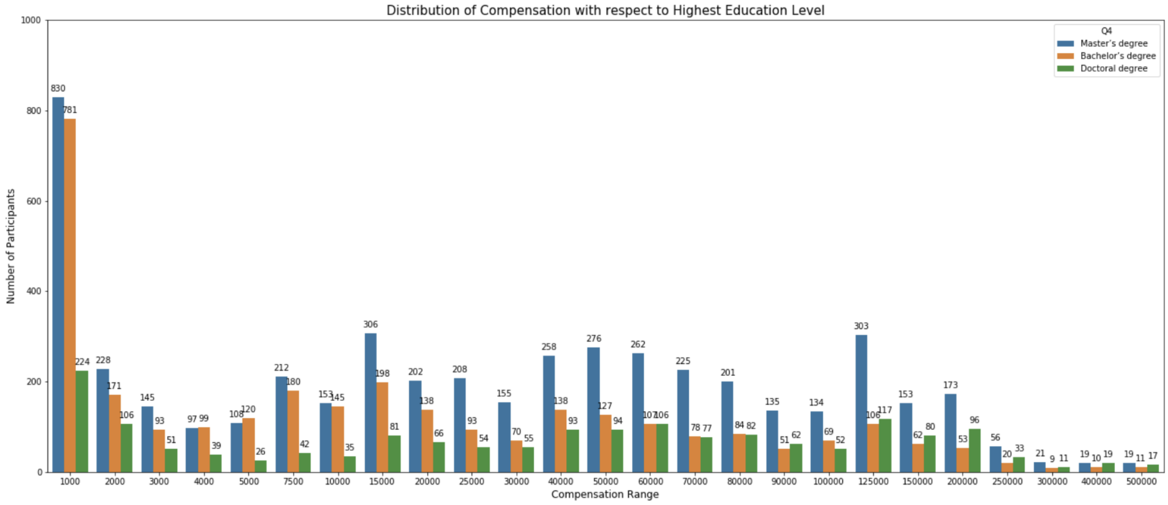
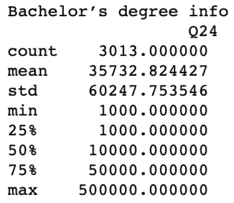


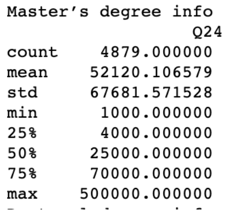
Figure 8. Number of participants in each average salary range according to the education level group

The figure above shows the number of participants in different salary ranges with the highest formal education of Bachelor’s degree, Master’s degree, and Doctoral degree. Similar to the gender analysis, a significant number of participants’ yearly salaries fell below 1000$USD, and peaks were seen at 15,000, 40,000 - 60,000, and 125,000 sections.

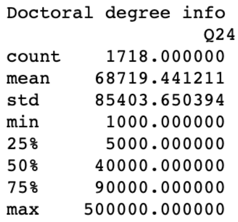
The descriptive statistics for each education group were computed and the boxplot was generated to study the means and medians, and the results are shown below. Among 3013 participants with a Bachelor’s degree, the average salary was 35,733 $USD, while the average salary of 4879 Master’s degree participants was 52,120 $USD, and the average salary of 1718 Doctoral degree participants was 68,719 $USD. Again, the large std resulting from skewed dataset with outliers of the very high salaries made the mean values not representative. Looking at the salary box plot for three groups, the Doctoral degree group had a higher median salary and a greater variance in the salary range, and the lower income level of the Bachelor’s group, compared to the other two groups, was evident. It is not surprising that having a higher formal education degree would allow the candidate to have a better income at work.



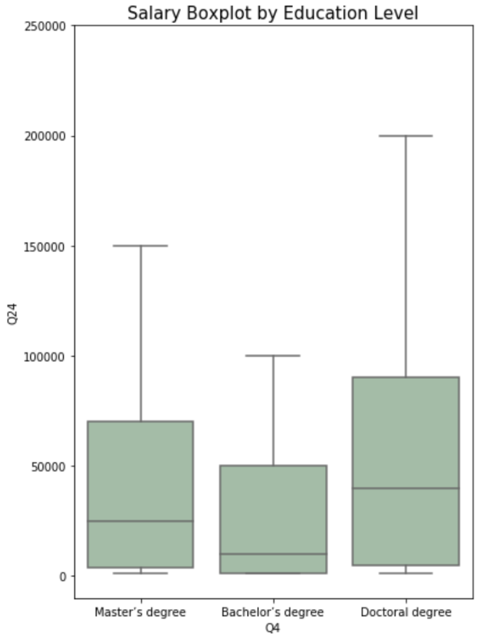
```
Bachelor's degree info
                   Q24
count     3013.000000
mean     35732.824427
std      60247.753546
min       1000.000000
25%       1000.000000
50%      10000.000000
75%      50000.000000
max     500000.000000
```



```
Master's degree info
                   Q24
count     4879.000000
mean     52120.106579
std      67681.571528
min       1000.000000
25%       4000.000000
50%      25000.000000
75%      70000.000000
max     500000.000000
```



```
Doctoral degree info
                   Q24
count     1718.000000
mean     68719.441211
std      85403.650394
min       1000.000000
25%       5000.000000
50%      40000.000000
75%      90000.000000
max     500000.000000
```



b) Perform ANOVA for hypothesis analysis (if applicable)

ANOVA test was used to compare the means of more than 2 groups. One of the ANOVA assumptions is that the residuals are approximately normally distributed. To see whether the ANOVA test could be applied to the dataset, the normality assumption was checked with the Shapiro-Wilks test with a threshold value of 0.05. If the test results in a p-value greater than the threshold, we can deduce that the dataset is very likely drawn from a Gaussian distribution and then apply the ANOVA test to compare the means. Upon applying the Shapiro-Wilks tests, a p-value of 0 was obtained for all three datasets, hence the ANOVA test is not applicable here.

c) Bootstrap sampling of Bachelor's/Master's/Doctoral dataset

Similar to section 2c), bootstrap sampling was applied to the three dataset for 1000 replications. Sampling means were computed in each iteration and differences between the means were calculated pair-wisely for the three groups. The plot of sampling means distribution and the sampling mean differences are shown below.

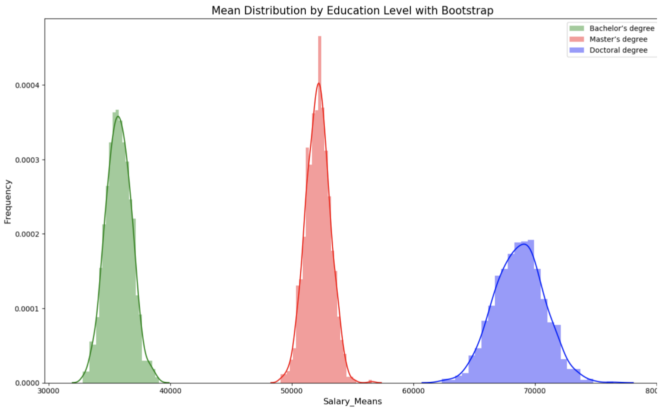


Figure 9. Bootstrapped Salary Means

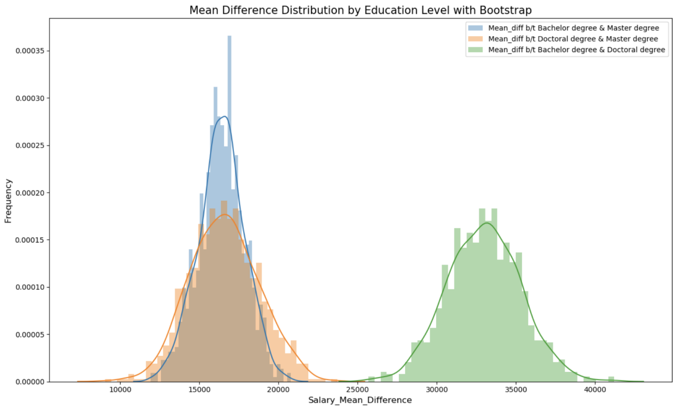


Figure 10. Bootstrapped Salary Mean Differences

The sampling distribution of the means and the pair-wise mean differences approach a normal distribution, and the ANOVA test was applied to the sampling means to estimate the variability in means.

d) ANOVA test with bootstrap sampling mean

One-way ANOVA test was performed on the bootstrapped means of the Bachelor's, Master's, and Doctoral salaries, and the result was obtained as f-value = 129236 , p-value = 0. With a threshold level of significance of

0.05, we can conclude that there are significant differences between the yearly average salaries of groups with different formal education levels.

e) Comment on the findings

With the bootstrap sampling means, the estimate of average yearly salary for data science community with a Bachelor's, Master's, or a Doctoral degree were summarized as below:

Table 1. Estimate of average yearly salary for data science community by Highest Level of Education

Highest Level of Education	Estimated Average Yealy Salary ($USD)	Standard Deviation ($USD)
Bachelor's degree	35740.00	1055.43
Master's degree	52128.60	990.17
Doctoral degree	68702.50	2050.58

The 95% confidence interval of the difference in average salaries with pair-wise analysis are shown in the table below.

Table 2. 95.0 % Confidence interval for difference in the average salary

Comparing pair	Difference in Salaries ($USD)	95.0 % Confidence interval for Mean Difference ($USD)
Bachelor vs. Master	16388.63	(13398.78, 19160.18)
Master vs. Doctoral	16573.92	(12501.84, 20898.76)
Bachelor vs. Doctoral	32962.55	(28555.86, 37413.69)

It is interesting to know that a person with a Doctoral degree in the datascience community can possibily makes a doubling amount of yearly income comparing to a person with a Bachelor's degree.