Report for Assignment 1

Angela Hsu

1003328874

Feb 16

Angela HSU
1003328874

1624 Assignment Report

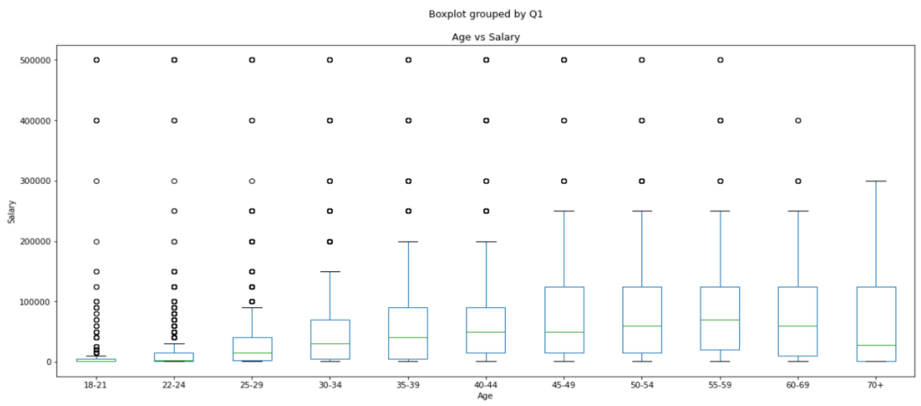
It was the purpose of the report to understand the nature of women's representation in Data Science and Machine Learning, and the effects of education on income level. Using the data analyzed in the study, the report will show that the salary for women differed significantly from that for men, and the salary for Bachelor's, Master's, and Doctoral's degree also differed significantly.

Problem 1

For the exploratory data analysis, three boxplot figures were drawn (Age vs Salary, Education vs Salary, Country vs Salary). Boxplot is a standardized way of displaying the distribution of data based on minimum (Q1-1.5IQR), first quartile (Q1), median, third quartile (Q3), and maximum (Q3+1.5IQR). It can illustrate how data is grouped and how data is skewed.

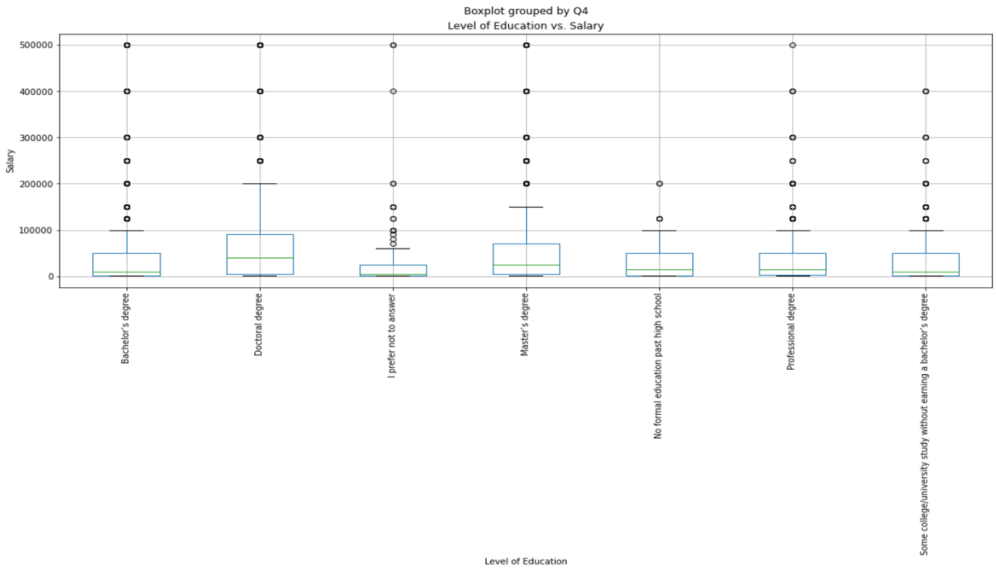
The first graph shows that the median of salary gradually increases as the age moves from 18 to 59 years and decreases as the age moves from 60 to over 70 years. Moreover, all distribution skewed to the right, and the number of outliers decreases as the age moves from 18 to over 70 years. In the sample, the highest salary $500,000 was achieved in all age group except people who aged 60 years and older. Thus, it can be concluded that age ranging from 18 to 59 years is positively correlated with salary and age ranging from 60 to over 70 years is negatively correlated with salary.

Fig1. Age vs. Salary



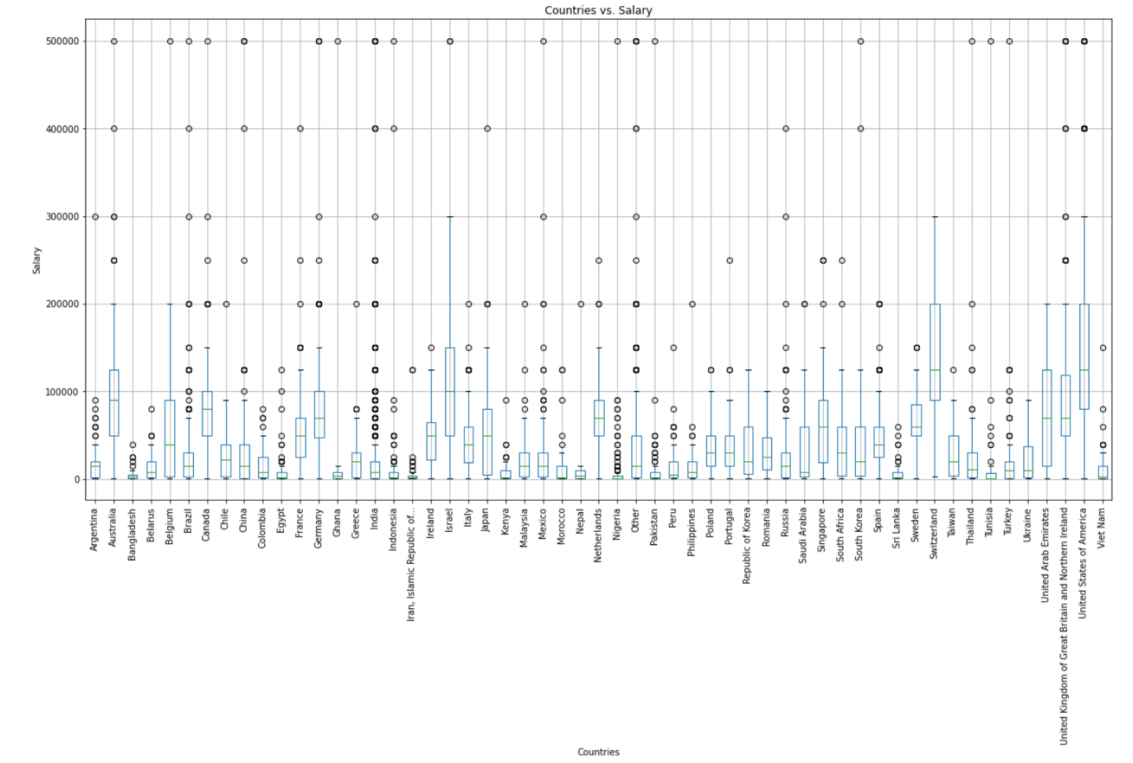
The second graph shows that the median of salary is positively correlated with education level. More specifically, the median of salary for Doctoral degree is higher than that for Master's degree, while the median of salary for Master's degree is higher than that for Bachelor's degree. Surprisingly, the median of salary for professional degree is nearly the same as that for the group which has no formal education after high school. However, the former group has more outliers than the latter one, saying that educated people are more likely earning higher salary. Moreover, the highest salary $500,000 was achieved in all groups except people who have no formal education after high school or people who haven't earn Bachelor's degree. Therefore, it can be concluded that level of education is positively correlated with salary despite of some special case in which people have no Bachelor's degree can perform well (salary >= $200,000).

Fig.2 Level of Education vs. Salary



The third graph shows that the median of salary for US, Switzerland, Israel, and Australia are the highest while that for Nigeria, Pakistan, Kenya, and Egypt are the lowest. Thus, people in developed countries are more likely to earn higher salary, and people in developing countries tend to have lower salary.

Fig.3 Countries vs. Salary



Angela HSU
1003328874

Problem 2

There are 8872 men with average salary $50,750 and standard deviation in salary $70,347. The minimum of salary is $1000 and the maximum of salary is $500,000. 1st quantile is $3,000; median is $25,000; 3rd quantile is $70,000.

```
count      8872.000000
mean      50750.619928
std       70347.974812
min        1000.000000
25%        3000.000000
50%       25000.000000
75%       70000.000000
max      500000.000000
Name: Q24, dtype: float64
```

There are 1683 women with average salary $36,417 and standard deviation in salary $59,442. The minimum of salary is $1000 and the maximum of salary is $500,000. 1st quantile is $1000; median is $7,500; 3rd quantile is 50,000.

```
count      1683.000000
mean      36417.112299
std       59442.716093
min        1000.000000
25%        1000.000000
50%        7500.000000
75%       50000.000000
max      500000.000000
Name: Q24, dtype: float64
```

As shown above, both the median and mean for women's salary is lower than those for men's salary. To determine whether there is a difference between the average salary of men and women, two-sample t-test with significance-level of 95% is performed. T-test is appropriate in this case since (1) the values of salary are independent with each other; (2) assume the sample is randomly sampled from the population; (3) assume the distribution of salary is normal due to large sample size; (4) the data values are continuous; (5) two samples have unequal variances. Since t-value is 8.79 and p-value is less than 5%, null hypothesis (H0: mean of men's salary = mean of women's salary) is rejected and it is 95% confident to conclude that the average salary of men is different from that of women.

```
import scipy.stats as stats

stats.ttest_ind(Man['Q24'],Woman['Q24'],equal_var=False)
```

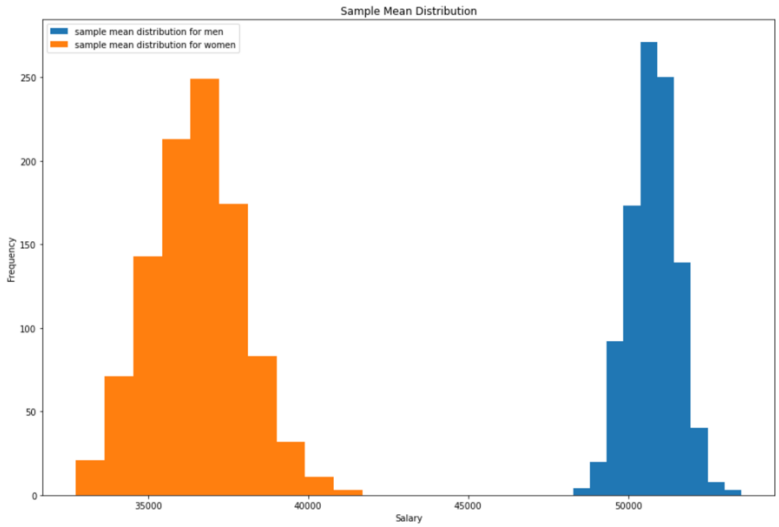
```
Ttest_indResult(statistic=8.792916020018406, pvalue=2.5655984594105773e-18)
```

Another t-test is conducted over the bootstrapped sample, which is a random sample taken with replacement from the original sample and with same size as original sample. Bootstrapping is good for its simplicity and it is more accurate than the standard intervals obtained using sample variance and assumptions of normality. However, bootstrapping requires representative sample and is time-consuming.

To compute bootstrapped sample, the original sample is resampled 1000 times with same size. The sample means for the salary of men and women are also computed for each bootstrapped

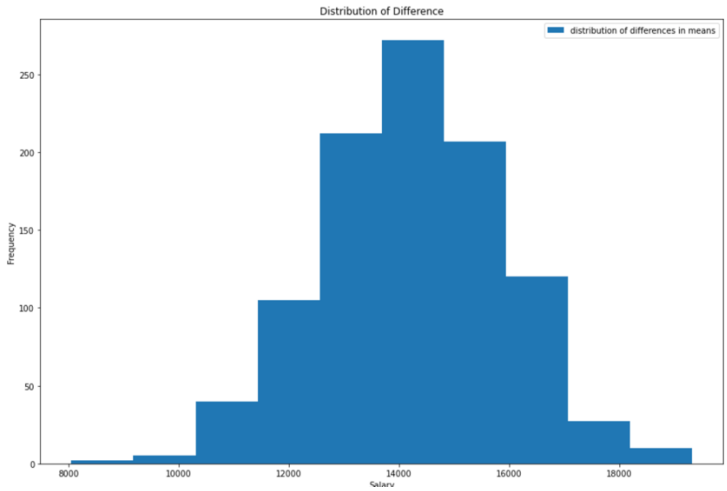
sample and stored in a list. Later on, the distribution is drawn as following. The spread of women's salary is a little larger than that of men's salary; the range of women's salary is a lot lower than that of men's salary.

Fig.4 Distribution of Sample Mean



Moreover, the difference between men and women's average salary is computed for 1000 bootstrapped samples and the distribution is shown as following. Salary of men is higher than that of women in the range from about $8,000 to nearly $20,000.

Fig.5 Distribution of Difference in Mean



Another two-sample t-test is conducted over the bootstrapped sample and the result shows that t-score is very large and p-value is approximately zero. Therefore, it is 95% confident to conclude that the average salary of men is different from that of women.

```
: stats.ttest_ind(man_avg_salary,woman_avg_salary,equal_var=False)
: Ttest_indResult(statistic=282.3593988314961, pvalue=0.0)
```

Both two-sample t-test and bootstrapped two-sample t-test says that the difference between men's and women's salary is statistically significant. Therefore, sex is a significant factor in Data Science and Machine Learning industry. The advantage of the bootstrap is that it can estimate the sampling distribution without many of the assumptions needed by parametric methods. The disadvantage of the bootstrap is that it is very dependent on the sample representing population. The bootstrapped t-test is actually more accurate than original t-test because bootstrapped sample are simulating the population while the original sample is only empirical. Since bootstrapped t-test produces larger t-score than original t-test, it is confident to conclude that the difference is statistically significant.

Angela HSU
1003328874

Problem 3

There are 3013 Bachelor's degree earning average salary $35,732 with standard deviation $60,247. The median of salary is 10,000; 1st quantile is $1000; 3rd quantile is $50,000.

```
count      3013.000000
mean      35732.824427
std       60247.753546
min        1000.000000
25%        1000.000000
50%       10000.000000
75%       50000.000000
max      500000.000000
Name: Q24, dtype: float64
```

There are 1718 Doctoral degree earning average salary $68,719 with standard deviation $85,403. The median of salary is 40,000; 1st quantile is $5,000; 3rd quantile is $90,000.

```
count      1718.000000
mean      68719.441211
std       85403.650394
min        1000.000000
25%        5000.000000
50%       40000.000000
75%       90000.000000
max      500000.000000
Name: Q24, dtype: float64
```

There are 4879 Master's degree earning average salary $52,120 with standard deviation $67,681. The median of salary is 25,000; 1st quantile is $4,000; 3rd quantile is $70,000.

```
count      4879.000000
mean      52120.106579
std       67681.571528
min        1000.000000
25%        4000.000000
50%       25000.000000
75%       70000.000000
max      500000.000000
Name: Q24, dtype: float64
```

As shown above, Bachelor's degree has the lowest mean and median in salary while Doctoral degree has the highest mean and median in salary. ANOVA is conducted to compare the means of salary for Bachelor's, Doctoral's, and Master's degree. ANOVA is a useful and economical method of parametric testing since it can analyze more than 2 groups of data and can also analyze multidimensional data. However, it requires strict assumption and may require post-ANOVA t-test for further testing.

In this assignment, One-way ANOVA is used because only one categorical variable (salary) is considered. There are five assumptions: (1) samples are drawn from normal distribution of the population; (2) dependent variable is expressed in interval or ratio; (3) independence of samples; (4) more than two groups of data; (5) homogeneity of variance. The null hypothesis (H0) is the equity in all population means while an alternative hypothesis is a difference in at least one mean.

As a result, the F value is 129 and p value is 0. Null hypothesis is rejected and the differences between group means are statistically significant.

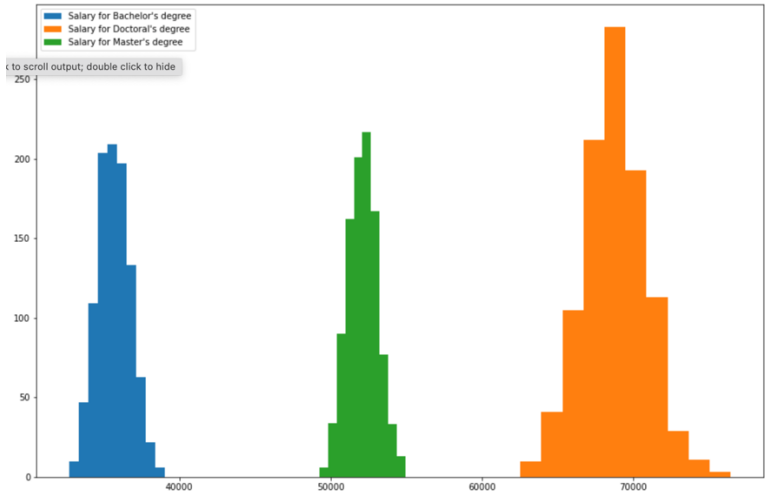
```
from scipy.stats import f_oneway

F, p = f_oneway(Bachelor['Q24'],Doc['Q24'],Master['Q24'])
print('ANOVA result: F =', F, ' p =', p)
```

```
ANOVA result: F = 129.7560112960932  p = 2.4852074227874282e-56
```

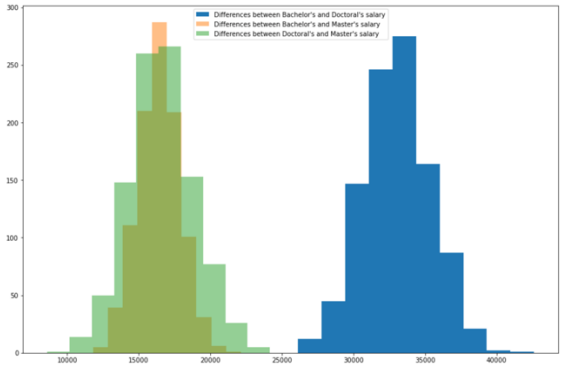
Another ANOVA is conducted over the bootstrapped sample, which is a random sample taken with replacement from the original sample and with same size as original sample. To compute bootstrapped sample, the original sample is resampled 1000 times with same size. The sample means of salary for Bachelor's, Doctoral's, and Master's degree are also computed for each bootstrapped sample and stored in a list. Later on, the distribution is drawn as following. The spread of Doctoral salary is a little larger than that of Master's and Bachelor's salary; the range of Bachelor's salary is a lot lower than that of Master's salary, while the range of Master's salary is a lot lower than that of Doctoral salary.

Distribution of Sample Mean



Moreover, the difference between Bachelor's, Doctoral's, and Master's salary is computed for 1000 bootstrapped samples and the distribution is shown as following. The difference between Doctoral and Bachelor's salary is the biggest while the difference between Master's and Doctoral salary and the difference between Master's and Bachelor's salary are about the same. It is evident that education level is positively correlated with salary.

Angela HSU
1003328874



As a result of ANOVA, the F value is 127416 and p value is 0. Null hypothesis is rejected and the differences between group means are statistically significant.

```
F, p = f_oneway(Bac_avg_salary,Doc_avg_salary,Mas_avg_salary)
print('ANOVA result: F =', F, ' p =', p)
```

```
ANOVA result: F = 127416.90131321909  p = 0.0
```

Both the ANOVA and bootstrapped ANOVA says that the difference between Bachelor's, Doctoral's, and Master's salary is statistically significant. Therefore, education is a significant factor for income level.