MIE1624 Assignment1

Jiachen Pan

October 2021

1 Exploratory data analysis

In this section, I read the clean kaggle data file and save as a data frame for further process. Since this data file have no question detail, so I check the question detail with original data file. Then I pick some features and check the trend for those features. I use histogram and bar-chart to show the trend more clearly.

1.1 Age vs average compensation

I check the relationship for age of people and average salary that people earn, I can find that the mean average compensation is increasing with age and decrease when 70+. Since when 70 years old or elder people getting old and retire, therefore I can ignore this group for analysis age versus average salary. Also since higher age means more working experience, so I can conclude as the higher the age(more working experiences) the higher the salary.

1.2 Programming experience vs average compensation

Then I try to find the relationship between programming experience and average salary. For those who have programming experience I can find that longer the year that people writing code(programming), higher the salary they get. Moreover, people who having more than 3 years programming experience can have higher average compensation than those do not working with code.

1.3 Number of language used vs average compensation

I summary the number of programming language for each people used for regular basis and store the number into a new column in the data frame. For the people who knows more than 7 programming languages is much less than other people so I only conclude the trend about people who knows 0-6 programming languages. As shown in diagram, average salary in increasing with more coding languages, then decreasing for 5 and 6 languages. However, the difference between the average salary for 4,5 and 6 languages is not much. Therefore this may caused by other feature effect such as programming experiences.

2 Difference between salary of men and women

First of all, I notice that there are many options in the gender part which is not man or woman. Therefore I just keep only man and woman rows and save as a new data frame.

Gender	counts	mean	std	min	25%	50%	75%	max
Male	8872.0	50750.619928	70347.974812	1000.0	3000.0	25000.0	70000.0	500000.0
Female	1683.0	36417.112299	59442.716093	1000.0	1000.0	7500.0	50000.0	500000.0

TABLE I. Statistic information of male and female in statistic part

Then I do the stats.shapiro test to see whether the data set is normal distribution or not. But the output of p-value of both male and female is 0 which means these two data set is not normal distribution; thus those two data set is not suitable to do a two sample t-test.

2.1 Bootstrap

Then I bootstrap my data with 1000 replications. Every time, I pick random data and calculate the mean of the data then summary them in a list. After that, I combine male's bootstrap data and female's bootstrap data into one data frame and calculate the difference between those two set of data. I also give bootstrap data a shapiro test, and it give both p-value around 0.76 which mean these data are normal distribution. Then I can give bootstrap data a two sample t-test. And the p-value for the those data is 0 which means this reject the null hypothesis and I can conclude that statistically significant.

2.2 Conclusion

In this part, since t-test can be only used in normal distribution data set. Therefore we need to use bootstrap data to perform a t-test. The reason is due to the central limit theorem sufficiently large number of random data from entire population will be normal distribution around the population mean and bootstrapping is the way to do this. Then we can get a set of normal distribution data from those two group to prove or reject the null hypothesis. For the my result my model reject hypothesis which mean male and female have difference in average compensation.

3 ANOVA(analysis of variance)

This time I am going to figure out relationship between three groups(Bachelor's degree, Master's degree and Doctoral) in salary.

Degree	counts	mean	std	min	25%	50%	75%	max
Bachelor's degree	3013.0	35732.82	60247.75	1000.0	1000.0	10000.0	50000.0	500000.0
Master's degree	4879.0	52120.10	67681.57	1000.0	4000.0	25000.0	70000.0	500000.0
Doctoral degree	1718.0	68719.44	85403.65	1000.0	5000.0	40000.0	90000.0	500000.0

TABLE II. Statistic information of different degree in statistic part

For our three groups the assumption of equal variance is violated, thus we can perform Welch's ANOVA test for a preciser result. However, for the raw data of those three group is nor normally distribution(from shapiro test), thus we cannot perform ANOVA test for them. Therefore we need to bootstrap the data and test.

3.1 Bootstrap

In this section, I do the same thing with previous part. Then I plot the bootstrap data set which shows data set have already been normal distribution. Then I combine three data set to one column which is easier to perform the Welch's ANOVA test. Then I got p-value output is 0 for bootstrap data ANOVA test.

3.2 Conclusion

I got p-value equal to 0 which means my model is reject null hypothesis. Also I can conclude this as statistic significant. As a result, since my model is reject the null hypothesis, thus I can say the average salary of those three group is different.