MIE1624H
Introduction to Data Science and Analytics

Assignment 1

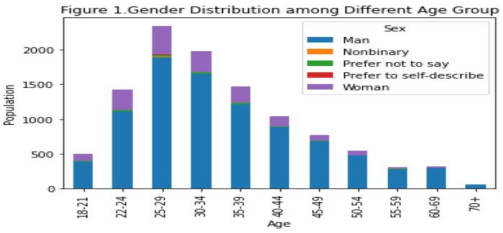
Zhi Xin Zhu
1002117112
16/02/2021

1. Exploratory Data Analysis(EDA)

The dataset for this assignment is a set of pre-processed survey results exploring employment in the data science community. The purpose of this assignment is to analyze the difference in average salary between different gender groups and different educational level groups. Before that, an EDA will be done to explore some of the main characteristics of this dataset.

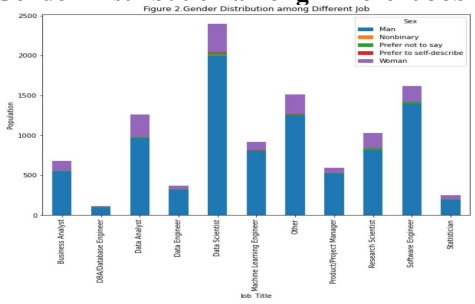
The dataset has 10729 entries and 356 features. Many of those features include NaN values because some of the features are broken down into multiple parts. For the purpose of this assignment, we will only look at the features that are represented in only one column. These features are Q1: Age Q2: Sex Q3: Country Q4: Education Q5: Job Title Q6: Work Experience (in years) Q24: Salary. To explore the relationship between features, 3 distribution figures are constructed.

1) Gender Distribution among Different Age Group



From this histogram, it can be seen that the distribution is right-skewed. The majority of the respondents of the surveys are male and between the age of 22- 44.

2) Gender Distribution among Different Job



From this histogram, it can be seen that data scientists and software engineers are the most popular jobs among the respondents. Very few of the respondents are DBA or statisticians. Similar to above, the majority of the population for each job is male.

3) Boxplot of Salary Distribution among Different Education Level

From figure 3[1], it can be seen that the education level of individuals does affect the salary. In general, individuals with higher education levels have higher salaries. Among all, individuals with Doctoral degrees have the highest 25 percentile, median, and 75 percentile salary. Individuals who preferred not to answer rarely earned over $75,000. It can also be noted that individuals with Bachelor's or Professional degrees have a very similar distribution.

2. Average Salary Difference Between Man and Woman Using T-test

a. Other than men and women, the dataset includes other gender groups. For the purpose of the assignment, the data for these gender groups will be dropped. The detailed descriptive statistics of salary between man and woman are shown in figure 4. It can be noted that the dataset is very unbalanced, the size of man samples is more than 5 times of the woman. From the statistics below, it can be seen that the mean salary of men is $50,751 and it is much higher than the average salary of women $36,417.

Figure 4. Descriptive Statistics of Man and Woman

Descriptive Statistic of Man	Salary
count	8872.000000
mean	50750.619928
std	70347.974812
min	1000.000000
25%	3000.000000
50%	25000.000000
75%	70000.000000
max	500000.000000

Descriptive Statistic of Woman	Salary
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. To perform a two-sample t-test, the following assumptions must be met.

1) The two samples are independent.
2) The data must follow a normal distribution
3) Homogeneity of Variance: The group variance is equal.

---

1 Due to page limitation, all figures mentioned but not shown in this report can be found on the *zhu_1002117112_assignment1.ipynb* Jupyter Notebook file.

The first assumption is tested when the survey is conducted. The normality assumption can be tested using the *scipy.stats.shapiro()* function. The null hypothesis that the data are normally distributed is rejected for both datasets. The equal variance assumption can be checked using *scipy.stats.levene()* function and the result is: homogeneity of variance is rejected. Since both assumptions 2 and 3 are not satisfied, the test cannot be performed.

c. Bootstrapping Data
The bootstrap method is a way to resample the original sample and create any simulated samples. Both the man and woman dataset is bootstrapped 1000 times. The method *data.sample()* is used. The sample mean is recorded for each replication and the mean distribution for men and women and mean difference distribution is plotted in figure 5,6,7[1] respectively. From the figures, it can be seen that both man and woman have a normal mean distribution and their mean difference also has a normal distribution.

d. To perform the t-test on the bootstrapped data, the assumptions mentioned above must be checked. The method used is the same as before. The independence assumption is assumed. The normality assumption is not rejected but the third assumption of equal variance rejected. The t-test can still be performed because the *scipy.stats.ttest_ind()* function allows us to perform Welch's t-test where an equal variance is not assumed.
The result of Welch's t-test shows that there is a statistically significant difference in the average salary between man and the woman

e. The result of Welch's t-test is not surprising. From the bootstrapped mean distribution plots, it is obvious that the salary difference between men and women is quite significant. The salary means difference graph shows that the salary difference between the two groups is around $14,000. The average salary for women is around $36,000, the difference between the two groups is nearly 40% of the woman's average salary.

3. Average Salary Difference Between Different Education Level Using ANOVA

a. Other than Bachelor's Degree, Master's Degree and Doctoral degree, the dataset includes education levels. For the purpose of this analysis, the data for these other educational level groups will be dropped. The detailed descriptive statistics of Salary between the three degrees are shown in figure 8. The statistics showed that the Master's Degree group has the largest sample size among the three. The mean of each group proves that the higher the education level, the higher the average salary.

Figure 8. Descriptive Statistics of 3 Different Education Levels

Descriptive Statistic of Bachelor' Salary

	Salary
count	3013.000000
mean	35732.824427
std	60247.753546
min	1000.000000
25%	1000.000000
50%	10000.000000
75%	50000.000000
max	500000.000000

Descriptive Statistic of Master's Salary

	Salary
count	4879.000000
mean	52120.106579
std	67681.571528
min	1000.000000
25%	4000.000000
50%	25000.000000
75%	70000.000000
max	500000.000000

Descriptive Statistic of Doctor's Salary

	Salary
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. The assumption for the one-way ANOVA test is the same as the t-test. The same methods are used to check the assumptions. The independence assumption is assumed. All three datasets did not satisfy the normality assumption and the homogeneity of variance assumption since the p-value for both assumption tests is smaller than the threshold of 0.05. Therefore, the ANOVA test will not be performed.

c. The process of bootstrapping the data is similar to the above, except three mean differences will be recorded. They are differences between 1) Bachelor's and Master's; 2) Bachelor's and Doctoral; 3) Master and Doctoral. The distribution plots are shown in figure 9 to figure 14[1]. The mean salary From the figures, it can be seen that all three groups have a normal mean distribution and their mean differences also have a normal distribution. The mean differences between Bachelor's vs Master's degree and Master's vs Doctoral degree is very similar, both around $16,000. The mean difference between Bachelor's and Doctoral degree is the greatest, it is around $32,000.

d. To perform the ANOVA test on the bootstrapped data, the assumptions mentioned above must be checked. The method used is the same as before. The independence assumption is assumed. The normality assumption is not rejected but the third assumption of equal variance rejected. However, for the purpose of showing the implementation of the ANOVA test, the equal variance assumption is assumed to be satisfied.
The *scipy.stats.f_oneway()* function will be used to perform the one-way ANOVA test. The result of ANOVA shows that there is a statistically significant difference in the average salary between the three groups

e. The result of Welch's t-test is not surprising. From the bootstrapped mean distribution plots and descriptive statistics, it is obvious that the salary difference of the three groups is significant. The average salary for Bachelor's, Master's, and Doctoral degree is $36,000, $52,000, $68,000 respectively. The average salary of individuals with Doctoral degrees is nearly double compared to those with Bachelor's degrees.