ALY_6010 Module 2 Report

Junchen Yi

College Professional Studies in Analytics, Northeastern University

ALY_6010: Probability Theory and Introductory Statistics

Na Yu

Submission Date: 12/11/2023

Introduction:

In this study, we explore a comprehensive dataset that encompasses various occupational attributes in the computer industry. It contains five key variables: age, experience, job role, education level and salary. These elements are critical to understanding job market dynamics and labor force trends. Age and experience reflect how much work experience the labor force has, while job roles and education level reflect professional diversity and academic background. Salary links these elements together. This analysis aims to reveal the relationships between these variables, providing a nuanced understanding of how different factors contribute to career development and pay.

Data Analysis:

1. Data import and clean

```
salary_data <- read.csv("Salary_Data.csv")
# Clean column name
salary_data <- salary_data %>%
  clean_names()
#Check whether the data frame contains null values and clean
any(is.na(salary_data))
salary_data <- na.omit(salary_data)
```

2. Descriptive Statistics and Grouped Descriptive Statistics

```
> describe(salary_data)
                     vars    n      mean       sd median   trimmed      mad min    max  range  skew
age                     1 6699     33.62     7.62     32     32.82     7.41  21     62     41  0.90
gender*                 2 6699      1.55     0.50      2      1.56     0.00   1      3      2 -0.16
education_level*        3 6699      4.77     2.11      4      4.71     2.97   1      8      7  0.33
job_title*              4 6699     95.92    58.71     98     95.90    83.03   1    192    191  0.06
years_of_experience     5 6699      8.10     6.06      7      7.44     5.93   0     34     34  0.98
salary                  6 6699 115326.96 52786.18 115000 115167.45 66717.00 350 250000 249650  0.06
                     kurtosis     se
age                     -0.18   0.09
gender*                 -1.84   0.01
education_level*        -1.35   0.03
job_title*              -1.35   0.72
years_of_experience      0.76   0.07
salary                  -1.17 644.93
```

```
# Grouped Descriptive Statistics
sub_salary_data <- salary_data %>%
  select(-c(gender, job_title, education_level)) #Remove some columns containing strings
grouped_describe <- describe(sub_salary_data)
selected_describe <- grouped_describe[c("n", "mean", "median", "sd", "se", "min", "max")]
```

	n	mean	median	sd	se	min	max
age	6698	3.362302e+01	32	7.615784	0.09305550	21	62
years_of_experience	6698	8.095178e+00	7	6.060291	0.07404929	0	34
salary	6698	1.153263e+05	115000	52790.093907	645.02992432	350	250000

3. Visualization

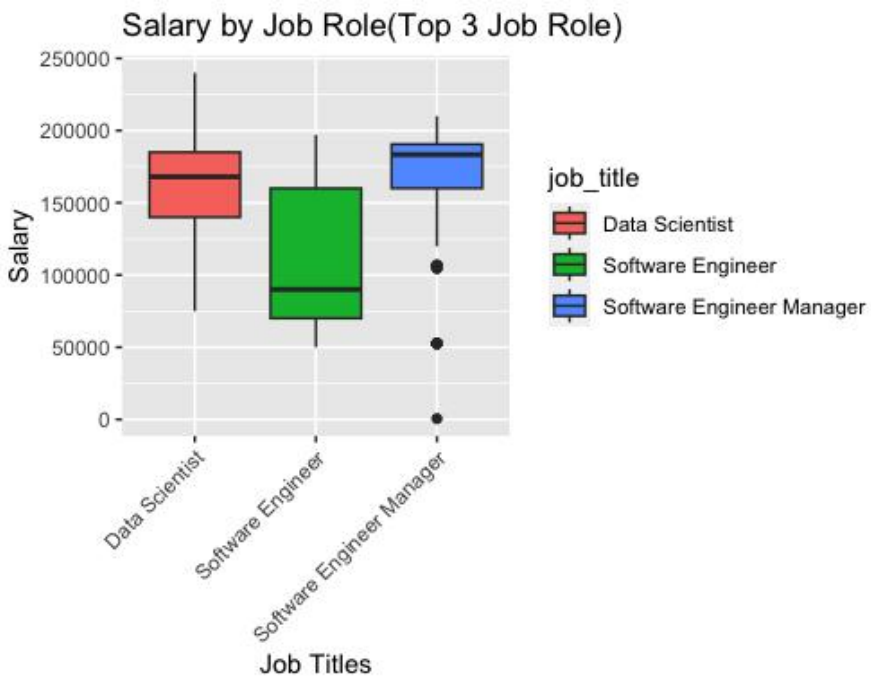
Scatter plot:

```
# Scatter plot for experience and salary
# Set graphics parameters
par(mfrow = c(1, 1), mar = c(4, 4, 2, 2))
# Create a scatter plot
plot(salary_data$years_of_experience, salary_data$salary, main="Experience vs Salary",
     xlab="Years of Experience",
     ylab="Salary",
     pch=19,
     col="blue")
# Calculate the regression model
model <- lm(salary ~ years_of_experience, data=salary_data)
# Add a regression line
abline(model, col="red")
```

The graph shows that as work experience increases, salaries generally increase too. However, there are many other factors, like job level, region, and industry, that affect salary in addition to experience. Therefore, even if two people have the same amount of experience, they may not have the same salary due to these variables.

Box plot:

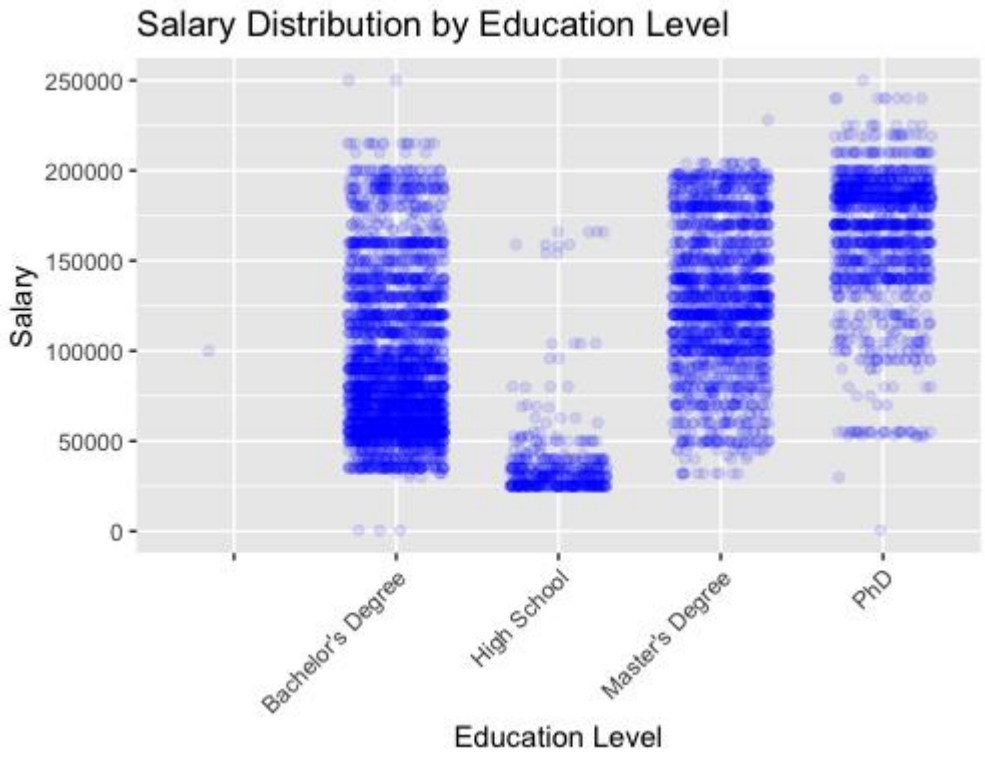
```
#Create a box plot showing salary distribution of different job roles
#Select the top three positions with the largest sample size
top_roles <- names(sort(table(salary_data$job_title), decreasing = TRUE))[1:3]
# filter data
filtered_data <- subset(salary_data, job_title %in% top_roles)
# Create box plot
ggplot(filtered_data, aes(x=job_title, y=salary, fill = job_title)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title="Salary by Job Role(Top 3 Job Role)", x="Job Titles", y="Salary")
```



In this box plot, we see the salary distribution for the three most common technical jobs in the dataset: data scientist, software engineer, and software engineer manager. Data scientists have the lowest median salary, but the distribution is more compact, suggesting that salaries are relatively concentrated and less volatile. Software Engineers have a median salary in the middle of the pack with a wider quartile range, reflecting a more volatile salary distribution. Software Engineer Managers not only have the highest median salary, but also the widest interquartile range, indicating the greatest salary variability within this position.

Jitter chart:

```
# Use ggplot2 to draw a jitter chart
salary_data <- subset(salary_data, education_level != "phD")
#Unify row names
salary_data$education_level <- gsub("\\bMaster's\\b", "Master's Degree", salary_data$education_level)
salary_data$education_level <- gsub("\\bMaster's Degree Degree\\b", "Master's Degree", salary_data$education_level)
salary_data$education_level <- gsub("\\bBachelor's\\b", "Bachelor's Degree", salary_data$education_level)
salary_data$education_level <- gsub("\\bBachelor's Degree Degree\\b", "Bachelor's Degree", salary_data$education_level)
#Draw jitter chart
ggplot(salary_data, aes(x=education_level, y=salary)) +
  geom_jitter(width=0.3, alpha=0.1, color="blue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title="Salary Distribution by Education Level", x="Education Level", y="Salary")
```



The jitter plot provides a visual representation of the relationship between education level and the distribution of salaries, illustrating that higher levels of education may be associated with higher salaries, but also suggesting that even among individuals with the same level of education, the differences in salaries may be large. Furthermore, given the presence of outliers, average salaries may be inflated by these higher values.

Conclusion：

Through a thorough analysis of more than six thousand data points, including age, experience, job role, education level, and salary, we reveal the interrelationships and effects of these variables. We learned that experience and age do not always directly correspond to higher salaries, and that while education level is usually associated with higher earnings, this relationship is significantly affected by job role. By using scatter plots, jitter charts, and box plots, we obtained a better understanding of data patterns and outliers. These findings offer important perspectives for making informed decisions in human resource management and policy creation.

Reference:

Reddy, M. S. R. (2023, May 18). *Salary_data*. Kaggle. https://www.kaggle.com/datasets/mohithsairamreddy/salary-data?select=Salary_Data.csv