

**CSE303: Statistics for Data Science**

**[Summer 2023]**

**Report on Assignment 01**

**Data Visualization**

**Submitted by**

**Student ID: 2020-1-60-226**

**Student Name: Sofia Noor Rafa**

# Introduction

To begin extraction of information from messy data, one needs to analyze available specific details of the data. The analysis begins with the identification of potential details. Identification of details begins when we start communicating with the data effectively and get to explore the dataset. Data visualization provides a toolkit to anybody wanting to communicate, and explore a dataset to identify potential details capable of providing valuable insights. Data visualization is a rich field of study. An array of tools are at hand to help start with the toolkit. Python has an extensive set of useful modules and libraries to get the job done, e.g. Matplotlib. pyplot module for making simple plots, Seaborn for building attractive statistical graphics, Pandas for working with tabular data, SciPy and NumPy for mathematical or computational analysis, and so on.

# Dataset Characteristics

My dataset contained **6704** entries - each of them indicating attributes of a person. All of them contained the information of the respective participant's gender status: male, female, or other, education level, age, job title, salary, and years of experience of work (**6** columns).

**Categorical columns:** job title, educational status, gender

**Numerical columns:** age, salary, years of experience.

**General characteristics: For Gender = Female,**

Total Job Titles: 132

Education Level: There are 7 categories found in the column,

Bachelor's Degree 2267

Master's Degree 1573

PhD 1368

Bachelor's 756

High School 448

Master's 288

phD 1

The dataset was cleaned by replacing Master’s Degree to Master’s, Bachelor’s Degree to Bachelor’s, and phD to PhD so that the analysis becomes more specific. The final result yielding:

Bachelor's 3023

Master's 1861

PhD 1369

High School 448

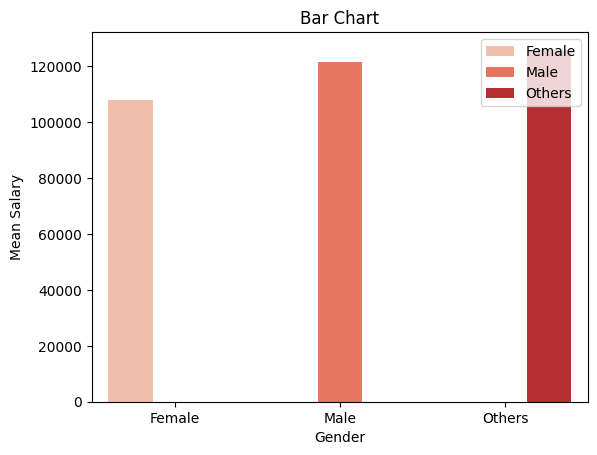
Age: Range (21, 60)

Salary: Range (500, 220000)

Years of Experience: Range (21, 60)

None of the columns contain float data points. The dataset is sourced from Kaggle. It is a stack of specified data great for applying some statistics to extract information, figure out dispersion or outlying points, play around with data, and get one’s hands dirty by doing some exploration. It is possible to find interesting insights from the dataset that may also be somewhat related to real-time situations. For example, figuring out correlations between two aspects : Female, Age and Female, Job Title. The dataset can lead to finding relation between reality and findings from the dataset.

# Exploratory Data Analysis

Figure: 01

**Explanation:** Female jobholders with the least amount of mean salary with genders other than male or female topping the list and male jobholders’ mean salary coming second.

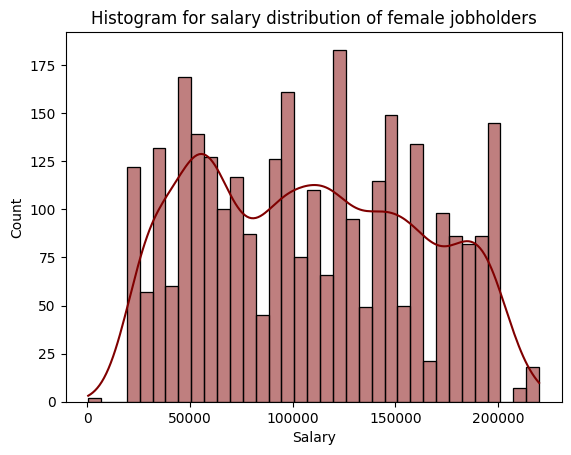


Figure 02

Explanation: Histogram for female jobholders’ salary has 35 bins. The histogram’s distribution shows that the data points don’t naturally form a bell-curve. There may be outliers or the data points may not be collected randomly.

The measures of central tendency, mean: 107888.99867241952, median: 105000.0 indicates that there are a few very high salaries that are pulling the mean up. The most common salary in the distribution is 120,000.

Skewness: 0.1453939440932321

Kurtosis: -1.129644170710776

Skewness value indicates the distribution has more concentration on the left side. More interesting results come in the displot visualization of salaries for female jobholders as the highest frequency for salary is around 50000 to 65000. Kurtosis value implies the distribution to be platykurtic - less data points lying upon the tails, more data points located near the mean.

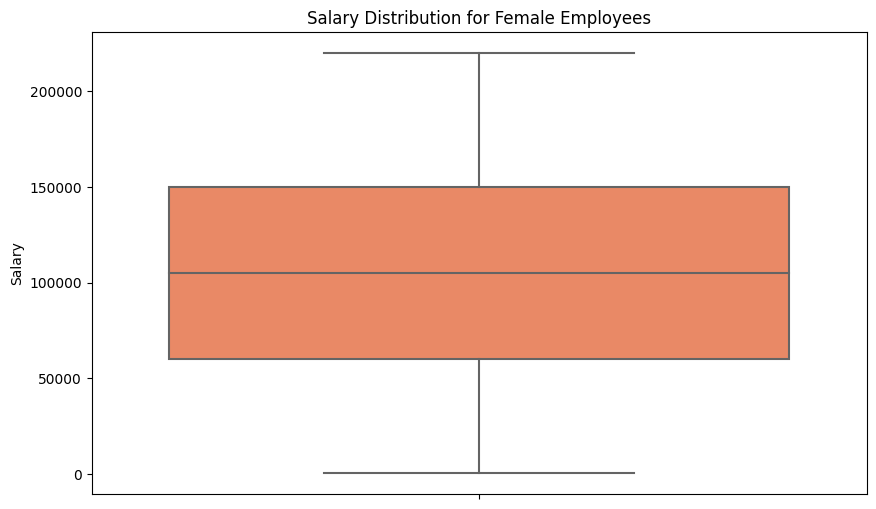


Figure 04

Explanation: No outliers after 220000 as it’s the maximum salary value. The visualization leads us to define relationships of other attributes of the dataset and salary.

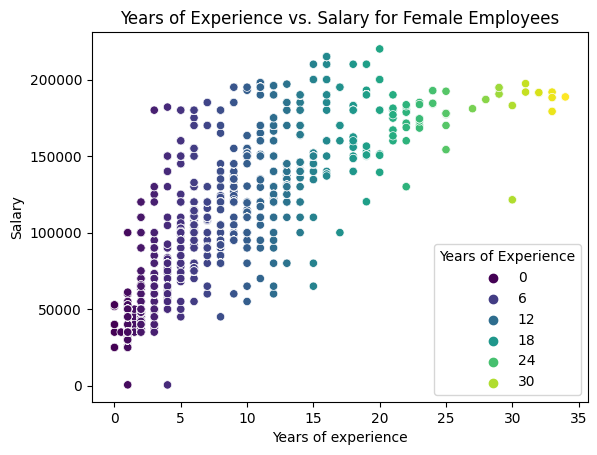


Figure: 05

Explanation: Years of Experience holds a positive relationship with salaries. As experience increases, salary increases. 0-5 years of experience led to salaries between 500 to 182000. 5 - 12: 198000, 12-18: 215000, 18 - 24: 220000, 24 - 30: 220000.

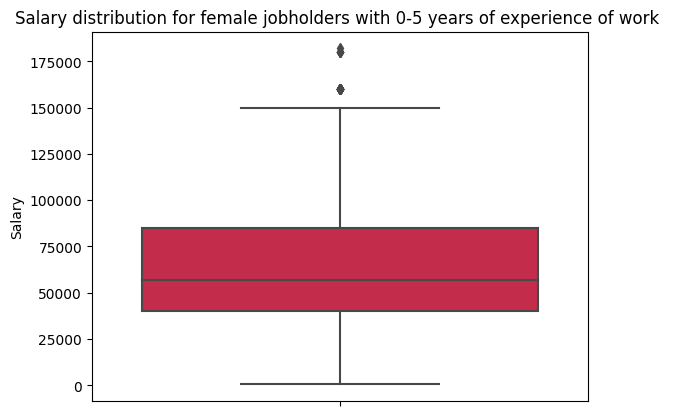


Figure 06

Explanation: 0-5 years of experience holders’ salaries gives a maximum of around 150000, 160000, thus the value 182000 is indeed an outlier, exceptional salary a 0-5 years of experience can get.

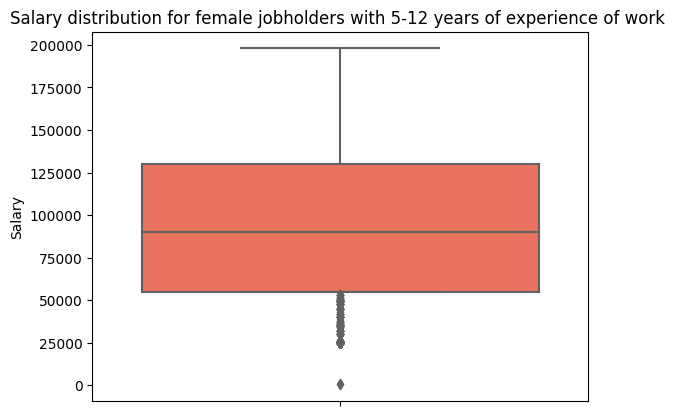


Figure 07

Explanation: Too low salaries detected as outliers for jobholders with experience of 5 to 12 years.

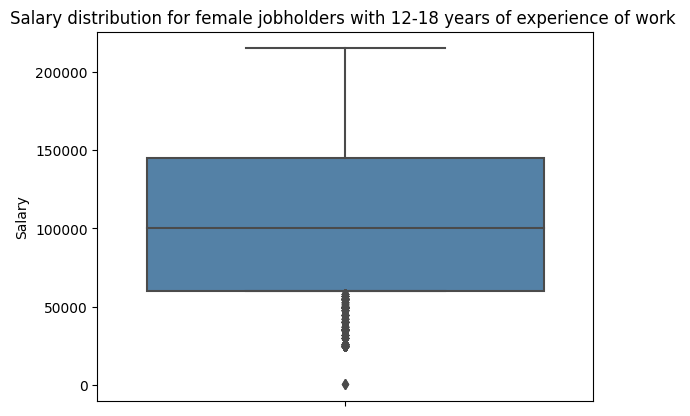


Figure 08

Too low salaries detected as outliers ranging around and below 50000.



Figure 09

Explanation: Yielding similar result like the previous one

Till 34 - the maximum value present in the dataset from the Years of Experience for female jobholders, similar results are produced. Thus, we can see for 5-34 years of working experience, there are potential outliers ranging around and below 50000 amount of salary.

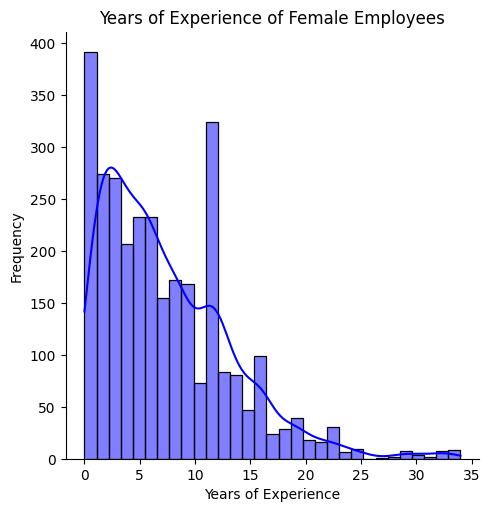


Figure 10

More female employees with fewer years of experience than there are with more years of experience. The distribution is more peaked than a normal distribution.

Moving onto the next part, correlation.

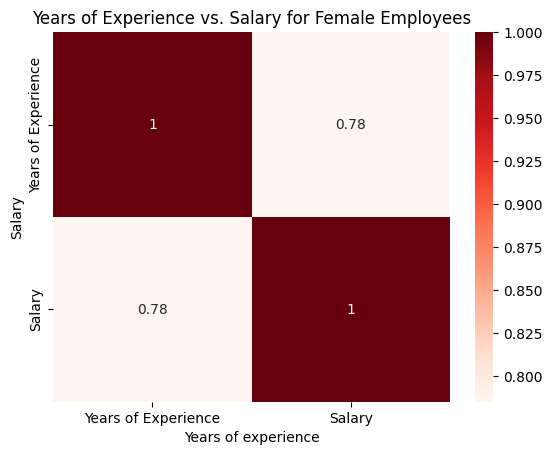


Fig 10

Explanation: strongly positive correlation

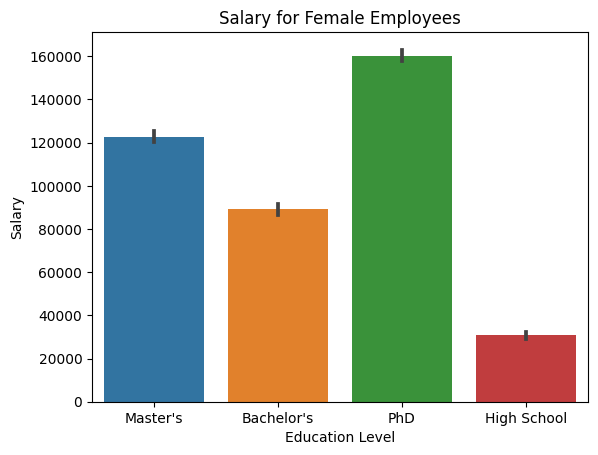


Fig 11

Explanation: Seaborn’s barplot yields the mean of plots of salaries on the Y axis based on the education levels. High School<Bachelor’s<Master’s<PhD. Mean values can become influenced by outliers. Which data points are pulling the mean value/s down or up?



Fig 12

Boxplotting verifies that for Bachelor’s the median is 75000, but the mean got pulled up and displayed more than 89000 of salary. It was influenced by some higher values, potentially outliers. For Master’s, it seems like there isn’t much difference between the median and the mean. For High School, the mean seems slightly pulled up - revealing presence of a bit upper extreme values. And for PhD, the mean is slightly pulled down - revealing presence of a bit lower extreme values.

No particular relation observed here, but salary range for each education level is seen, Female jobholders with Master’s level education form a salary range of 32000 to 198000, for Bachelor’s degree holders the range is: 500 to 215000, for PhD holders: 55000 to 220000 and for High School degree holders : 25000, 154207

Boxplots for analyzing a female jobholders’ education level and job title, and salary

The boxplots resulting for backend developers yielded interesting results: no PhD holder works as a backend developer. Jobholders with Bachelor’s degrees earn the most salary for the job.

Data scientist is the job title with the most frequency(202). Data analysts with a Bachelor's degree earn more than the ones with a Master’s degree.

Let us explore the situation in the most frequent job in the list: data scientist

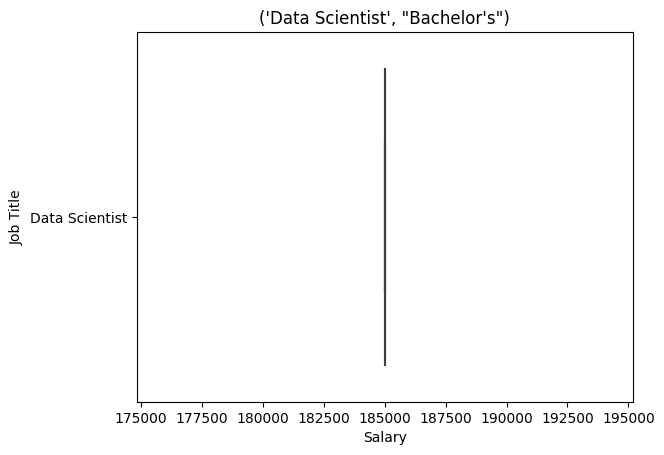


Fig 14

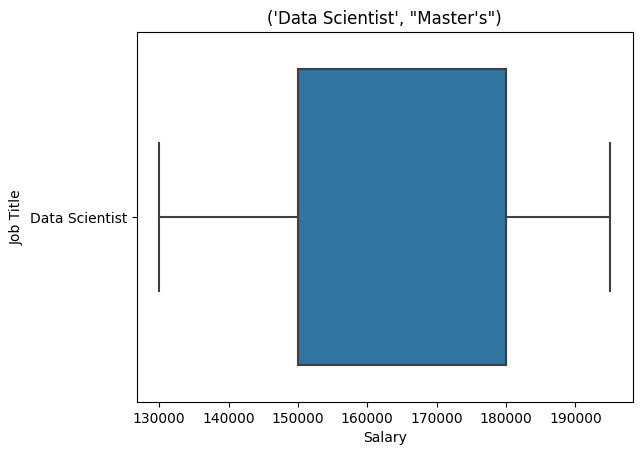


Fig 15

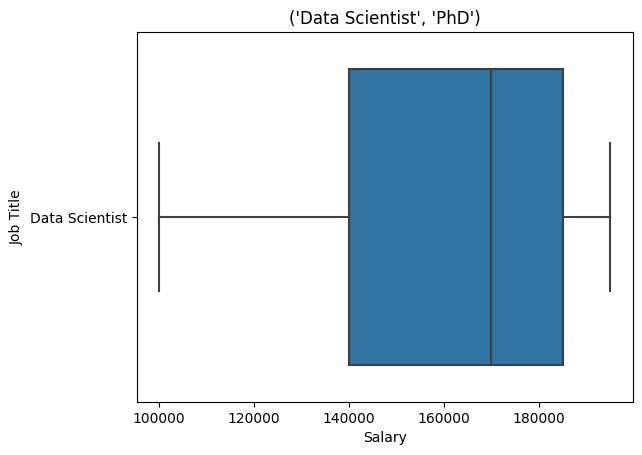


Fig 16

Data Scientists with PhD earn the most salaries and are present in the dataset the most.

Data Scientists with Master’s level education earn a little less but the presence of them is higher than those with Bachelor’s degrees in the dataset.

There are job titles where not many jobholders are available thus affecting the boxplot figure. But it’s visible that High School educated jobholders earn the least and at points, Bachelor’s degree holders earn more than Master’s degree holders. PhD holders earn the most.   
Jobs like Digital Marketing Manager pays more to Master’s degree holders. But the question remains whether these shapes are getting affected by the presence of a particular job title and Years of Experience.

A direct visualization shows:

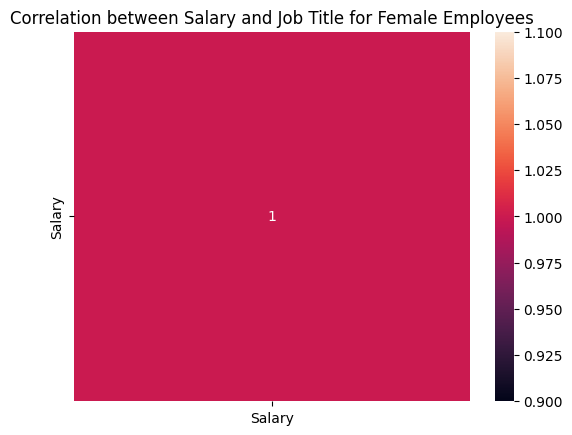


Fig 17

No correlation.

The important factor could be ‘Years of Experience’ as we see sheer randomness in the salaries based on Education Level. The randomness is again found in a job title called Director of Marketing that has Bachelor's degree holders earning the least, many of the Master’s degree holders earn more than some PhD holders but the values get little high for PhD holders in the upper extreme areas. Another example: Financial Analysts with Bachelor’s degrees earn higher than Master’s degree holders.



Fig 18

Age distribution shows that more values are present between age 24 to 36 (left-weighted)

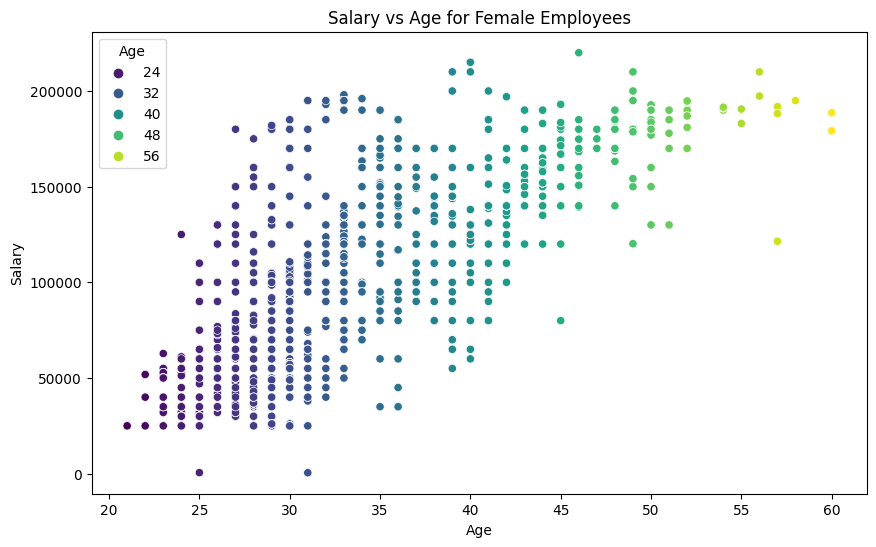


Fig 19

Positive correlation

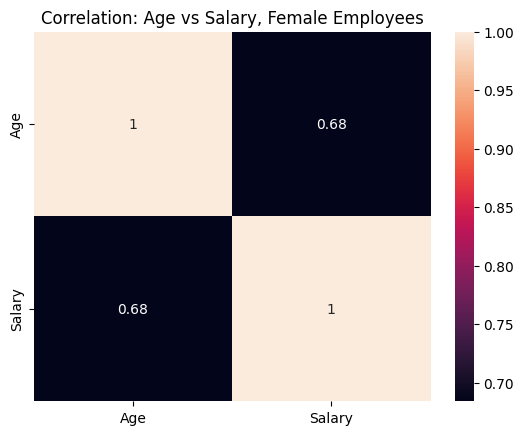


Fig 19

Positive correlation

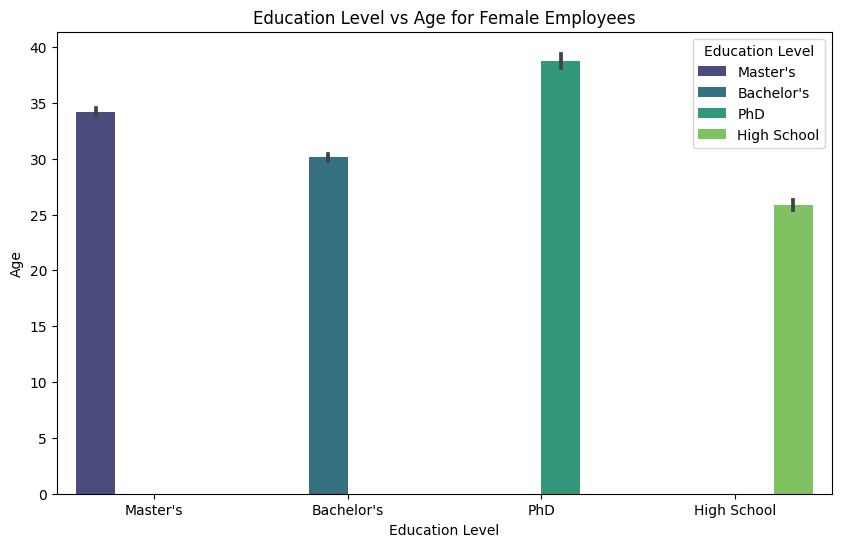


Fig 20

Lowest age group formed by people with high school education, Master’s forming a bit higher than Bachelor’s and PhD topped the list.

minimum age 21, maximum 60.

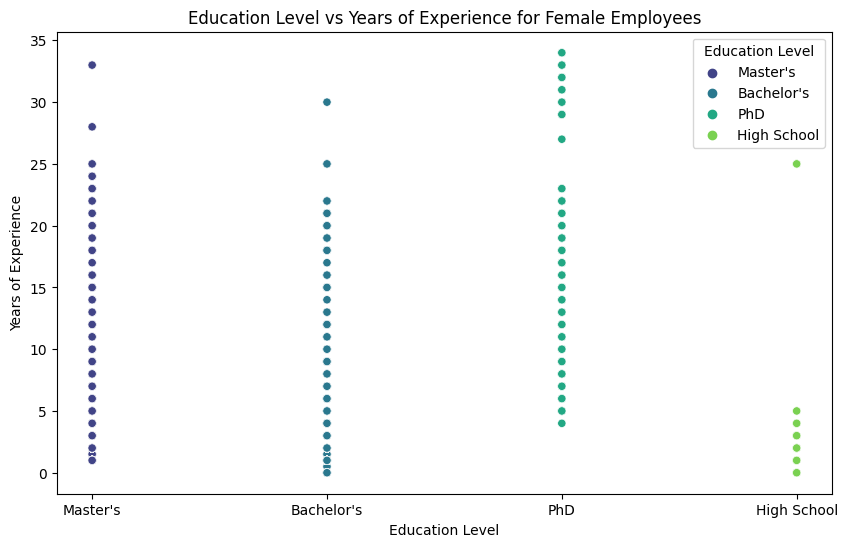


Fig 21

On most points we see Bachelor’s and Master’s and PhD educated jobholders’ years of experience collide. 5 years of experience is present among jobholders of all education levels - an interesting case.

PhD holders make the most experienced jobholders, then Master’s then Bachelor’s and High School.



Fig 22

Very positive correlation

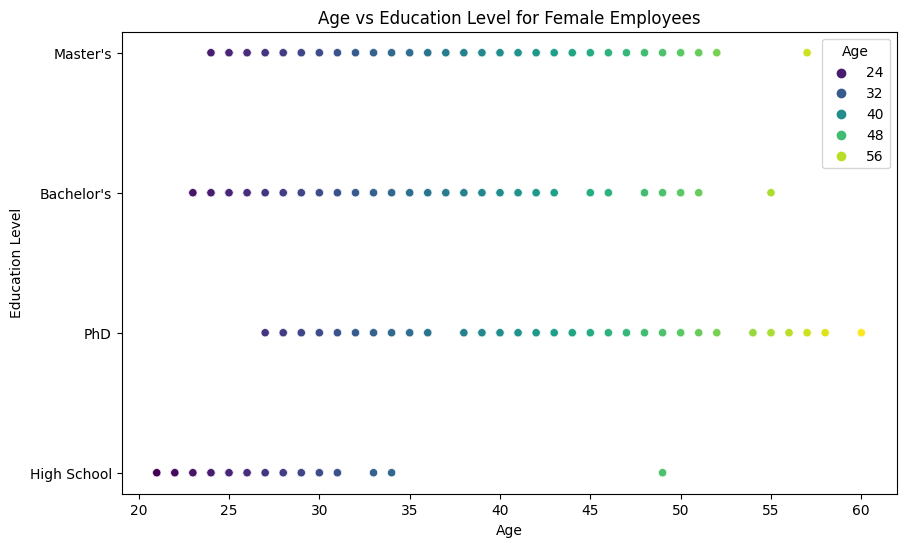


Fig 23

Realistic points. Again, most points collide for the top 3 tier of education levels forming the age groups. But PhD reaches out to become the most aged education level. Surprisingly, there is almost no difference between the age groups formed by Bachelor’s and Master’s.

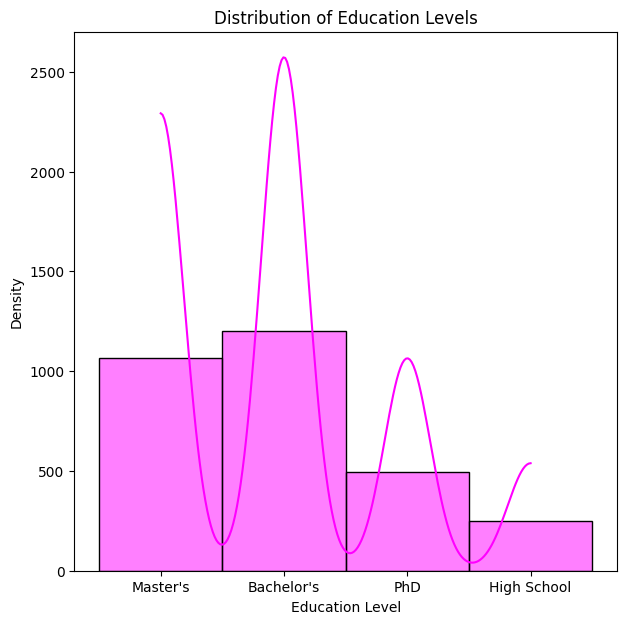


Fig 24

Bachelor's 3023

Master's 1861

PhD 1369

High School 448

Job title vs Age for female employees:

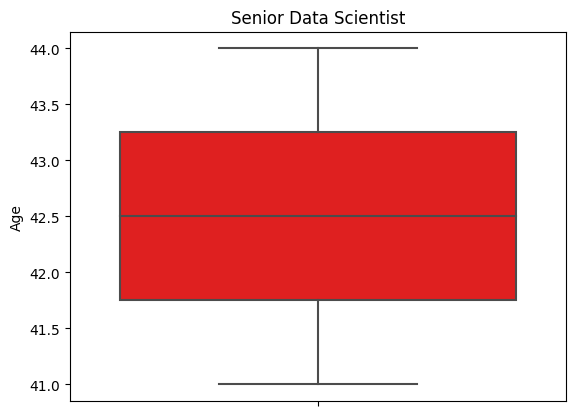


Fig 25

data analyst age: 25 - 33

product manager: above 30 and below 34, outliers above 34 and below 31

HR Manager : 42 to 46

Marketing Manager: 25 to 43

customer success representative: 26

administrative assistant 36 to 39

director of marketing : above 27 to 43

graphic designer = 26 - 27, 29 outlier

content marketing manager: 26 to 51

Junior HR analyst: 22 to 29

sales manager: 30 to 50

senior HR generalist: 26 to 42

junior marketing specialist: 25 to 29

senior data scientist: 41 to 44

concentration of data points felt between age 25 to 45

Job title vs Years of Experience

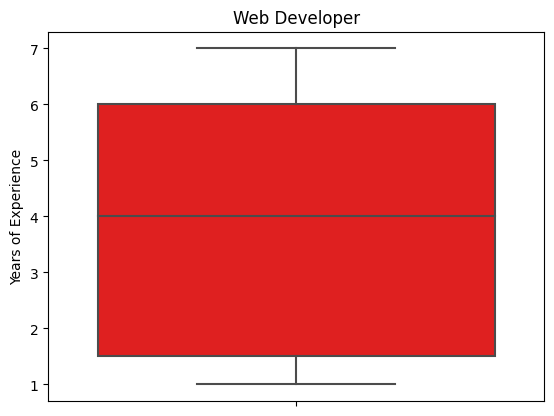


Fig 26

sales representative: 2 - 3

sales director: 5 - 7

Financial analyst: 4 - 12

Human resource manager : 6 - 15

web developer: 1 - 7

software engineer manager : 12 - 34

senior software engineer: 5 - 24

junior web developer: 1 - 7

data point concentration felt around 1 to 12

Chi-square estimation showed no relationship between the columns ‘Education Level’ and ‘Job Title’ for the female employees.

For male employees:

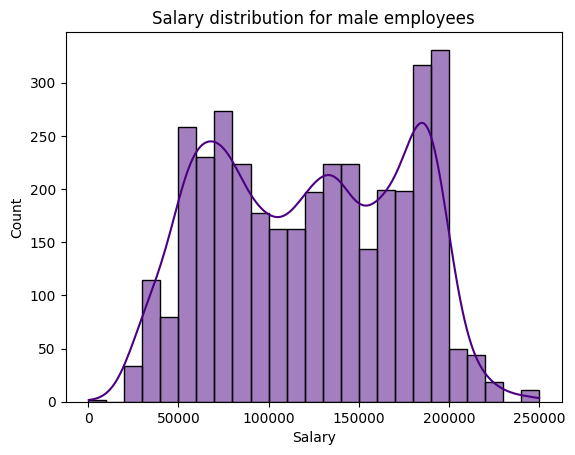


Fig 28

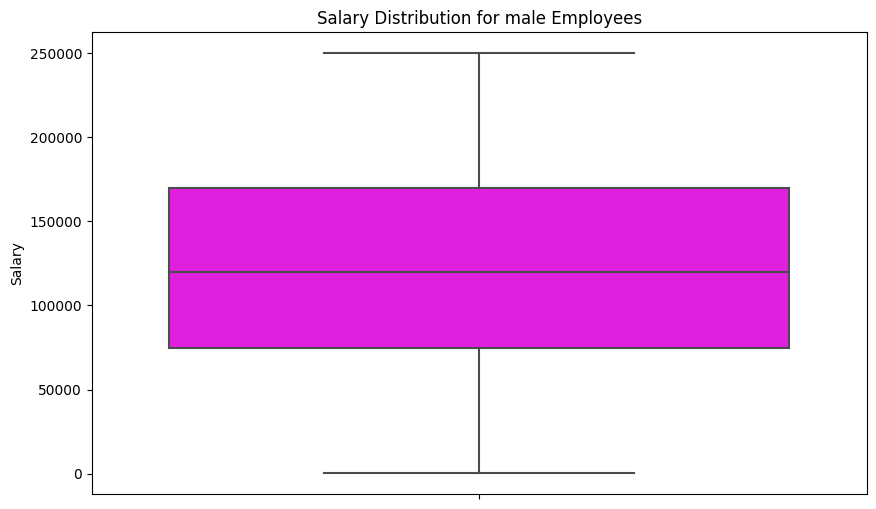


Fig 29



Fig 30

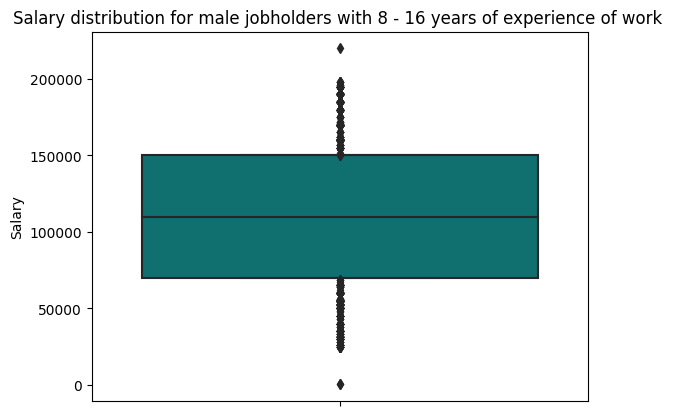


Fig 28



Fig 29



Fig 30



Fig 31



Fig 32

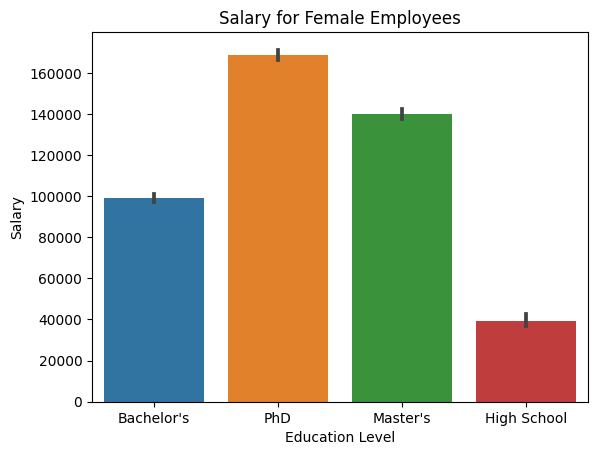


Fig 33

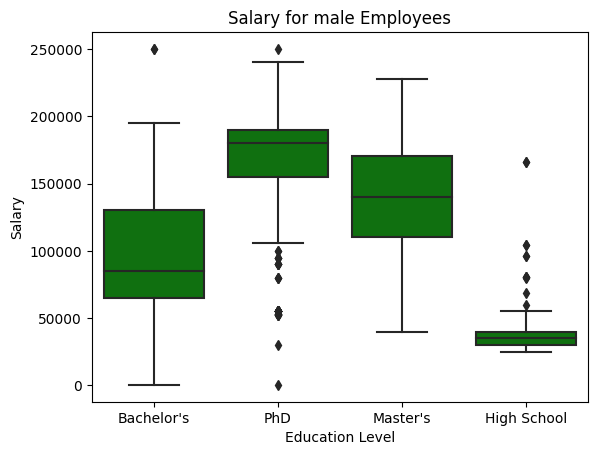


Fig 35

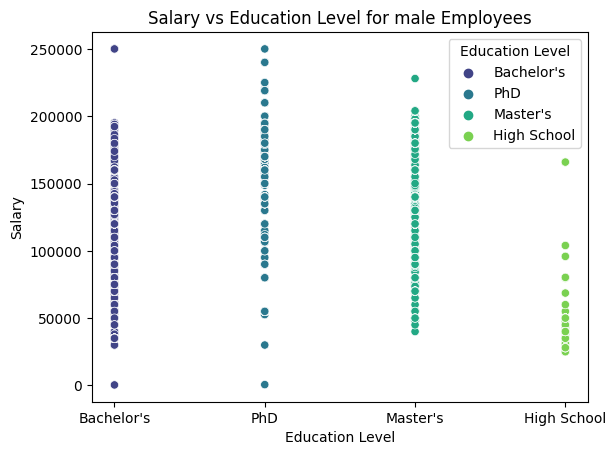


Fig 36



Fig 37



Fig 38



Fig 38

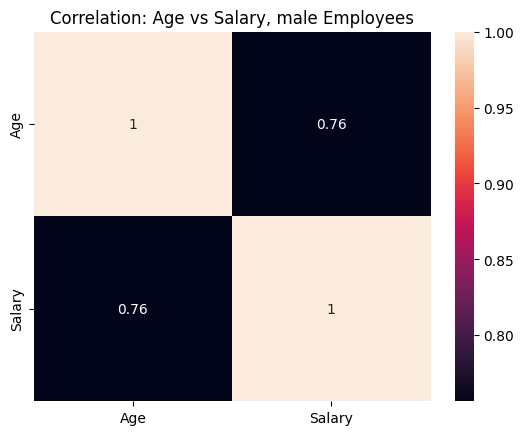


Fig 39

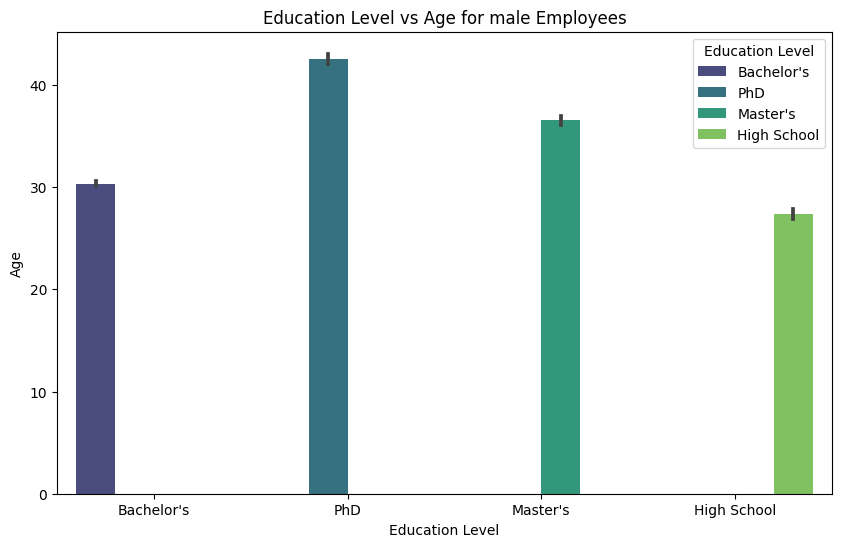


Fig 40

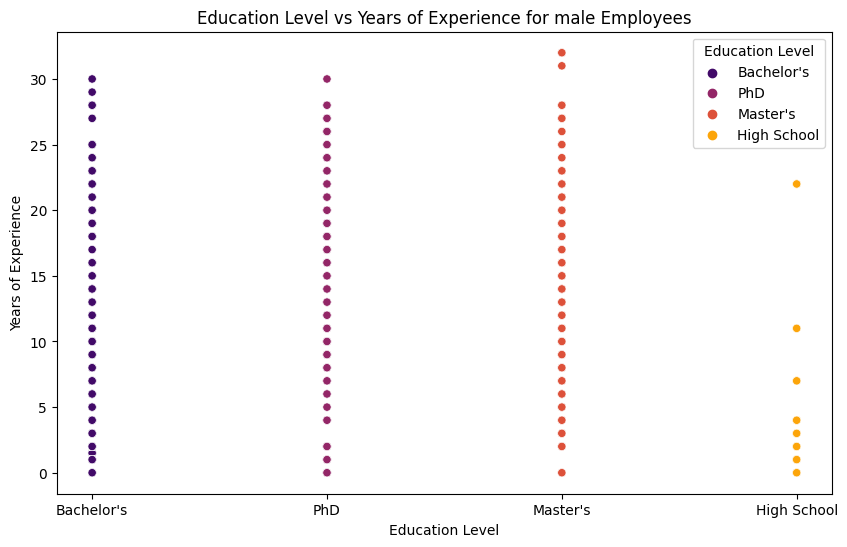


Fig 41



Fig 42



Fig 43

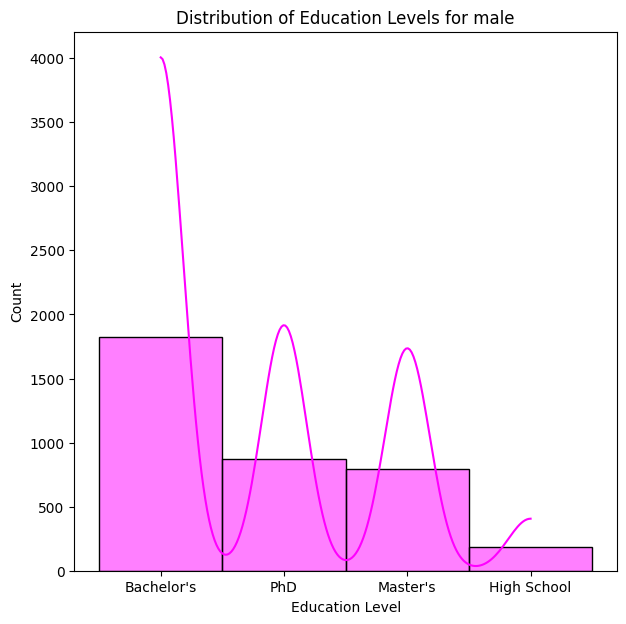


Fig 44



Fig 45

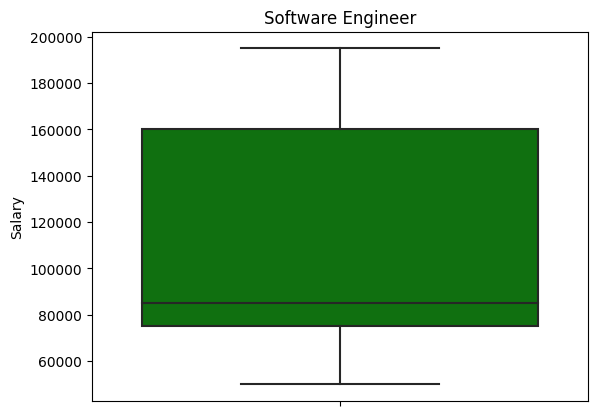


Fig 46

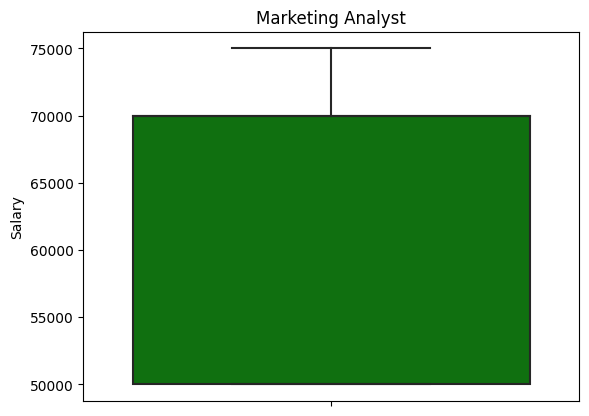


Fig 47

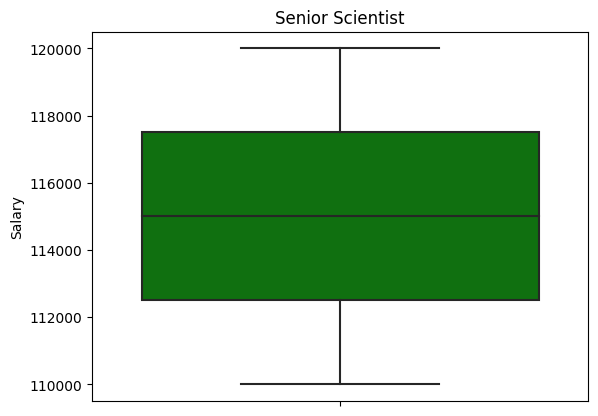


Fig 48

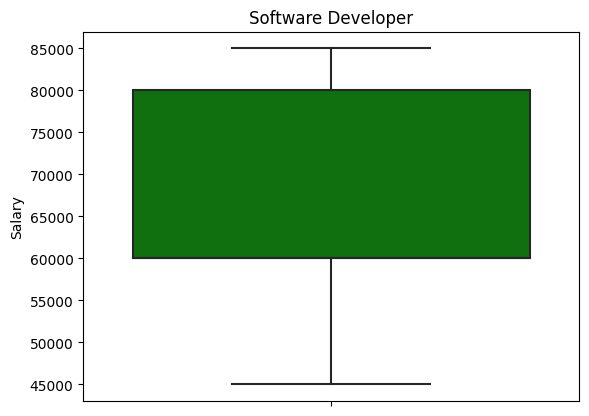


Fig 49

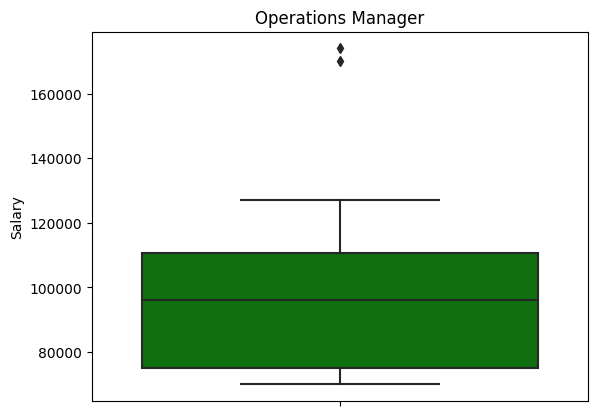


Fig 50

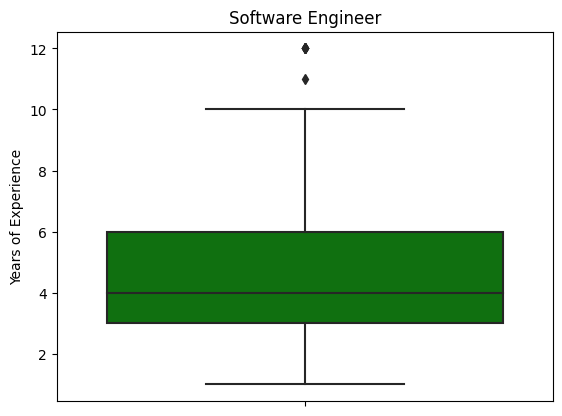


Fig 51

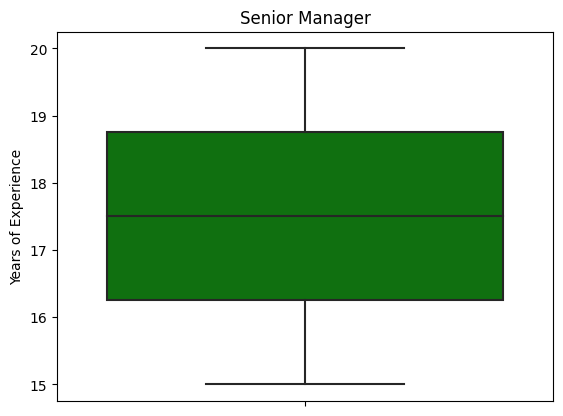


Fig 52

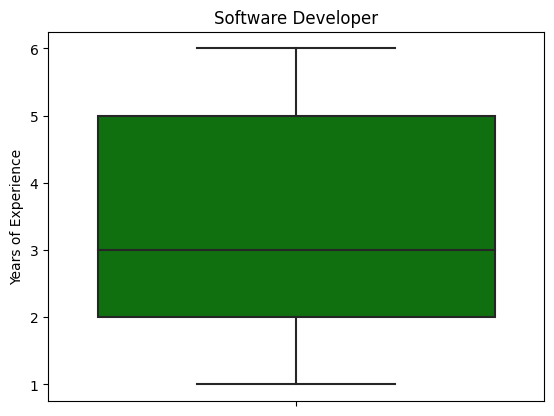


Fig 53

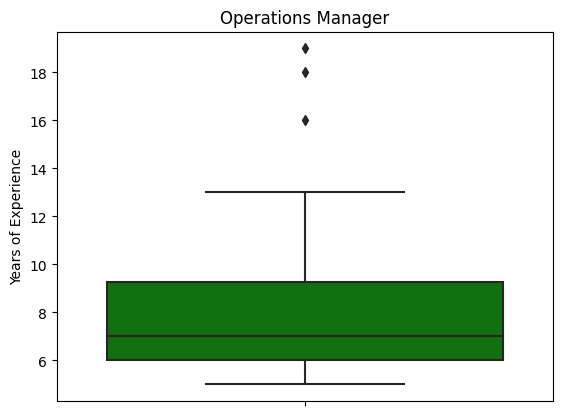


Fig 54

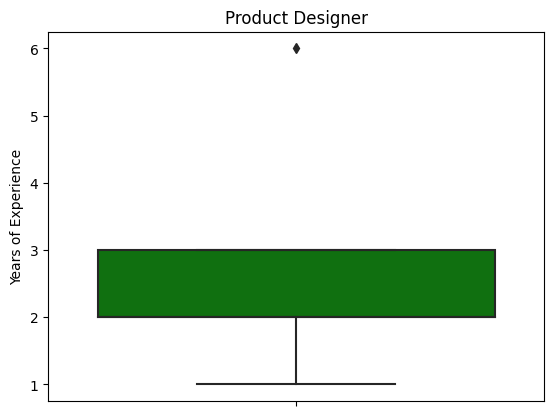


Fig 55

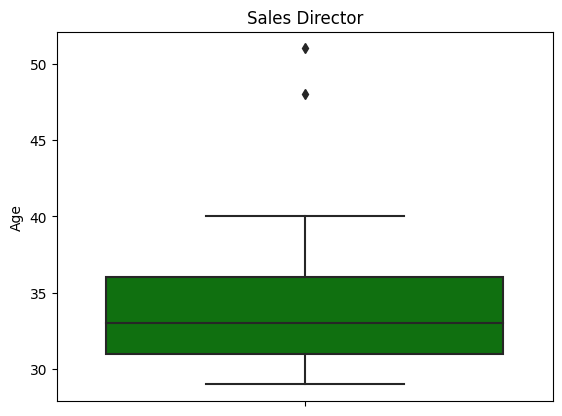


Fig 56

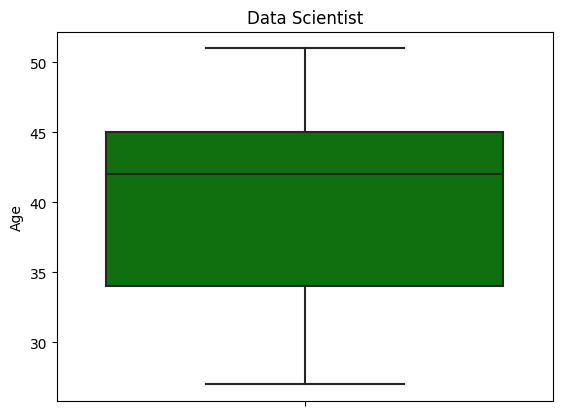


Fig 57

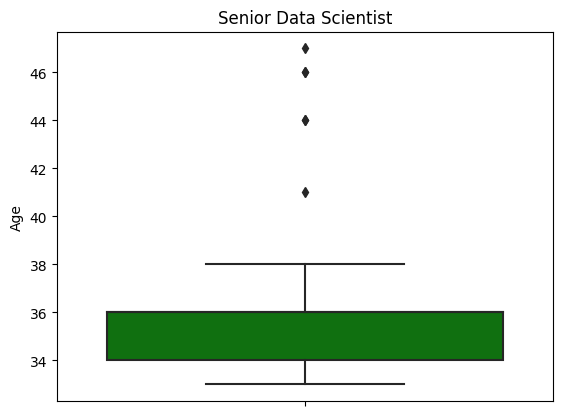


Fig 58

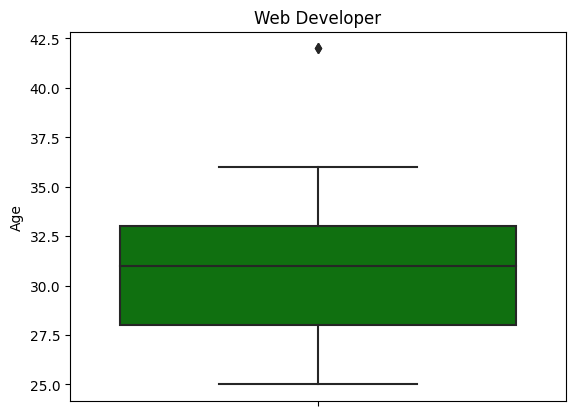


Fig 59

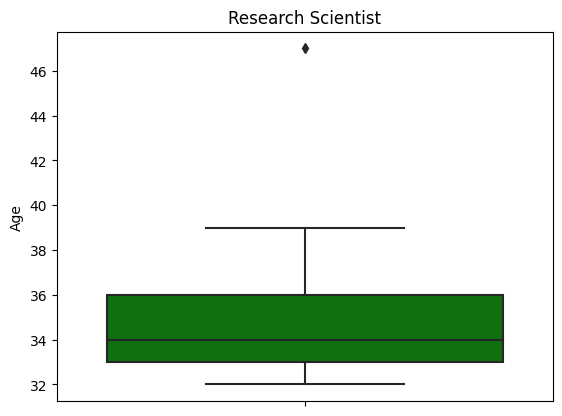


Fig 60

**For Gender=’Other’**

only 14 entries

**Education Level:**

High School 12

Master's 2

The Gender column contained an ‘Other’ category. There were no null values in the category.

**Salary range:** 62852.0 to 166109.0

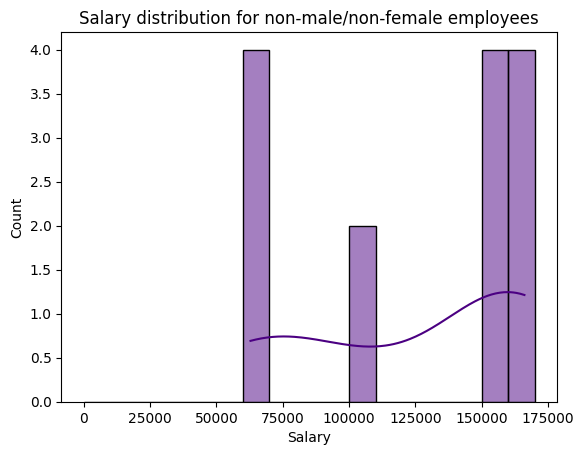


Fig 31

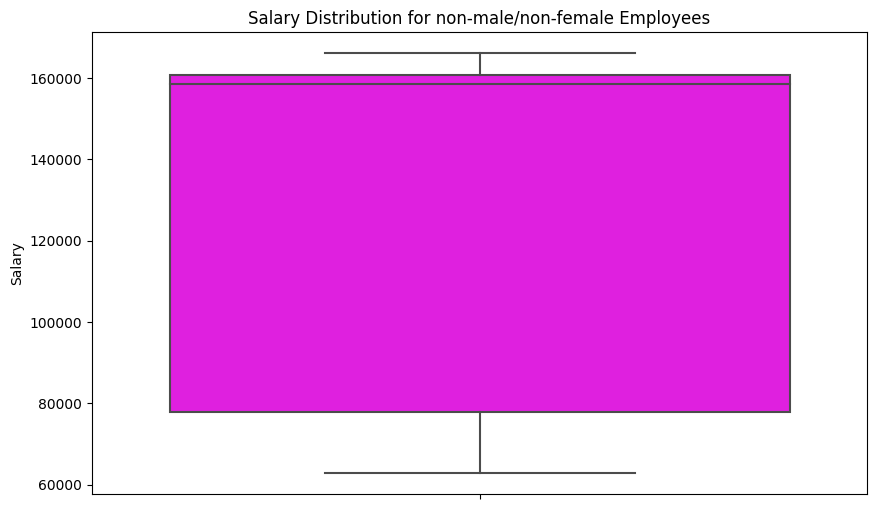


Fig 32

Skewness: -0.5447141771328795

Kurtosis: -1.7331573221175782

more data points on the right side, flatter than a normal distribution



Fig 33



Fig 34

yields same results, outliers below 70000 salary, ranges from 70000, to over 160000.



Fig 35

Yields positive relation

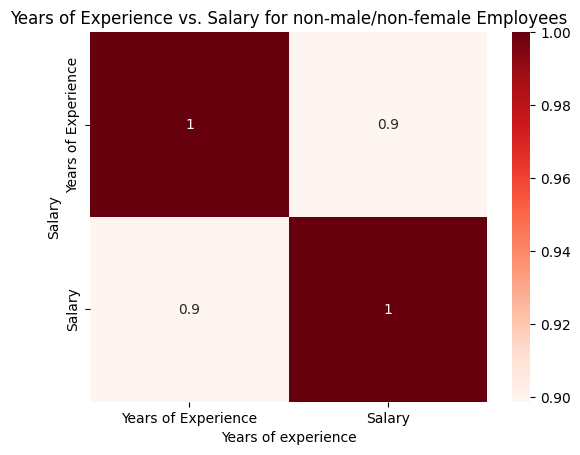


Fig 36

There are only a few entries, 14. The results show a positive relation between Salary and Years of Experience.



Fig 37

Employees with greater experience(25-30 years) exist the most.

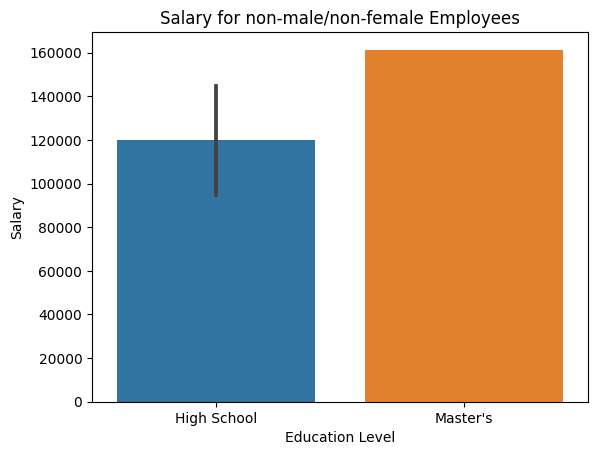


Fig 38

High school goers earn quite a lot, 120000. Master’s degree holders make a salary of 160000.

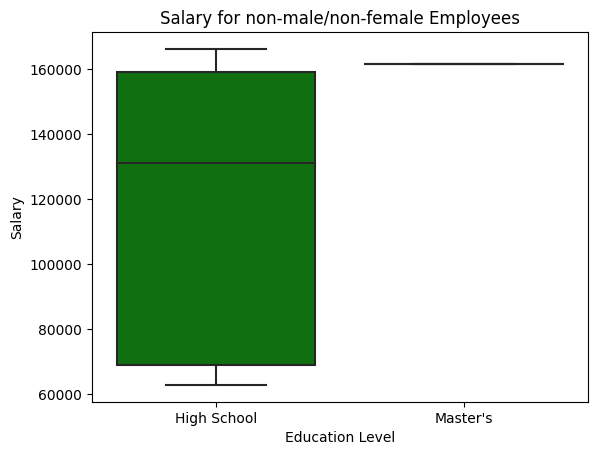


Fig 39

Most of the data points belong to high school degree holders.

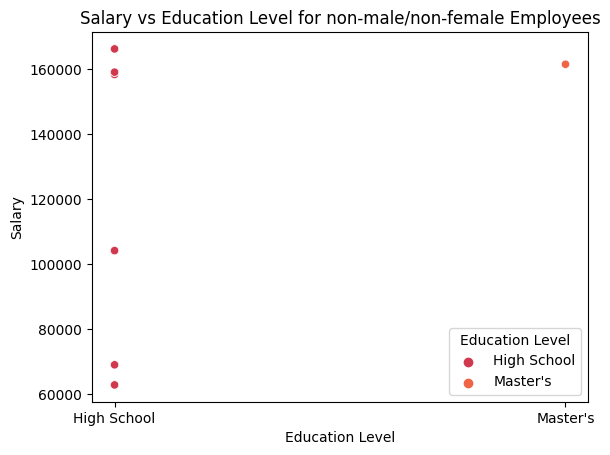


Fig 40

No relationship



Fig 41

No relationship



Fig 42

People with the highest value for age : between 53 to 55, are present in the list the most.



Fig 43

Higher the age, higher the salary.

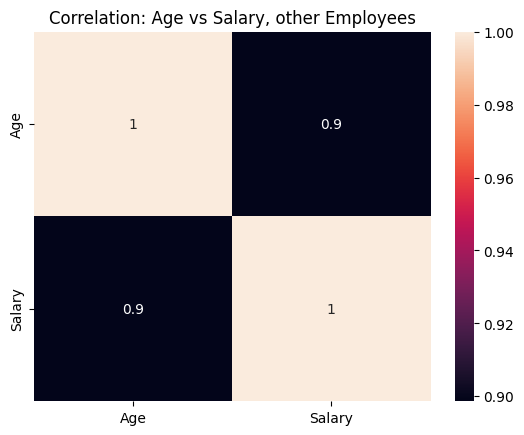


Fig 44

Almost perfect positive correlation.

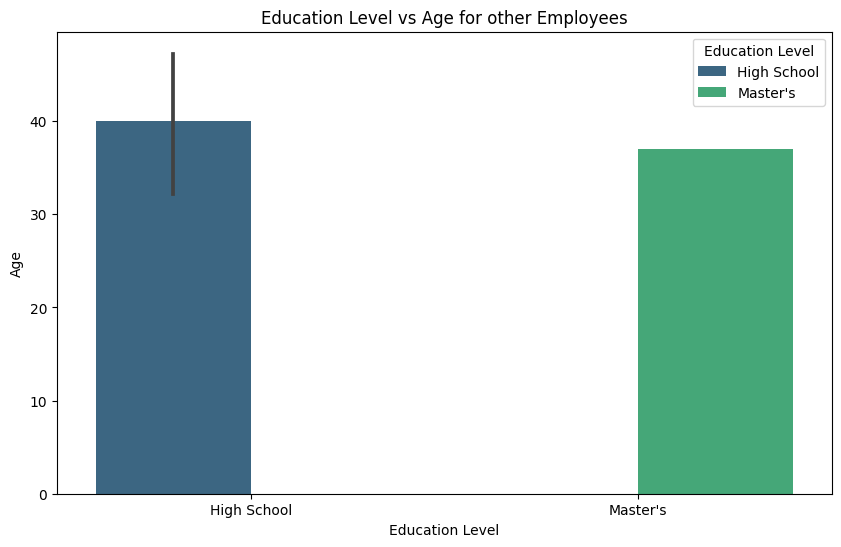


Fig 45

Only 2 values and range within 40



Fig 46

Scattered plots show no specific relationship

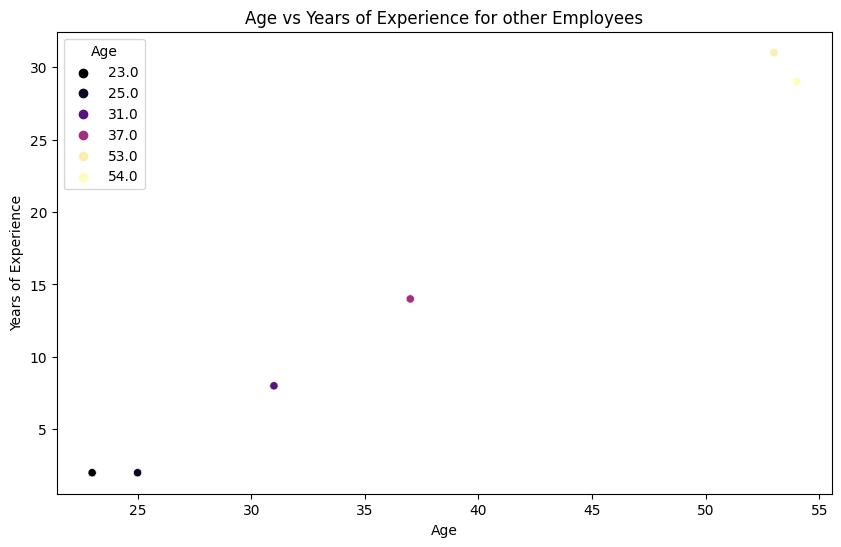


Fig 47

Positive relationship

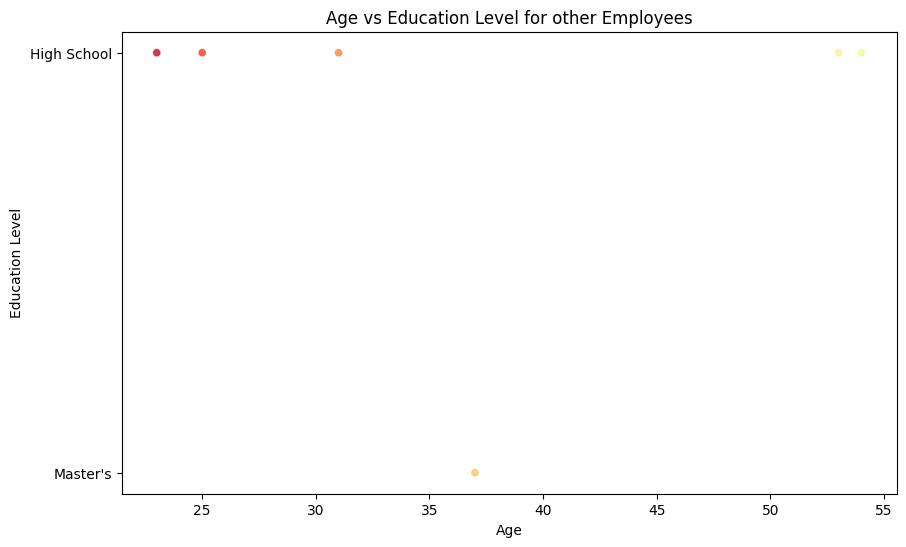
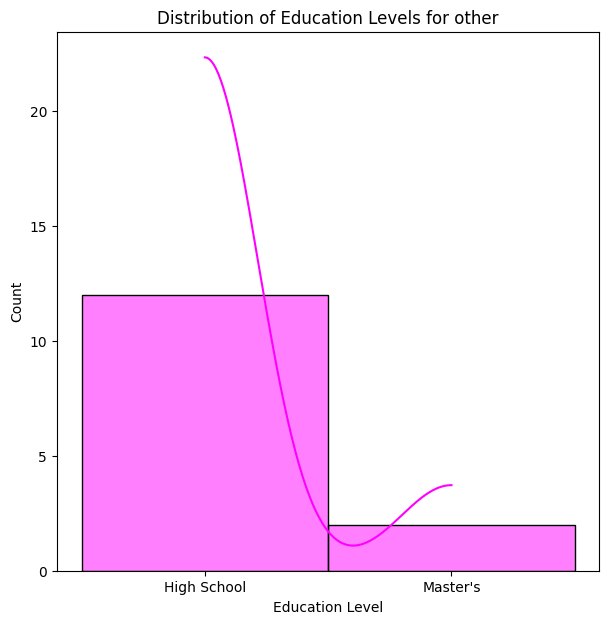


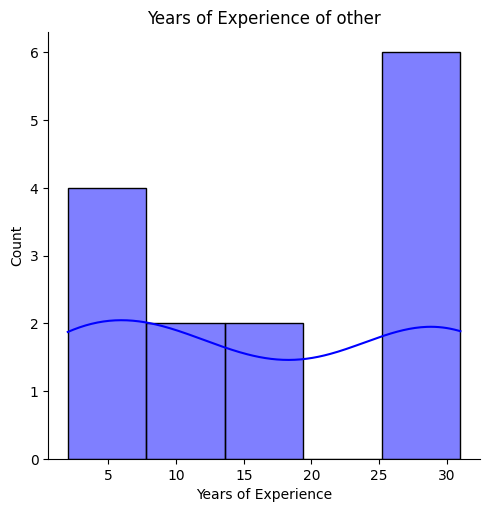
Fig 48

No particular relationship



High school with most frequnecy

Fig 49



4 to 6 years of experience distribution

Fig 50

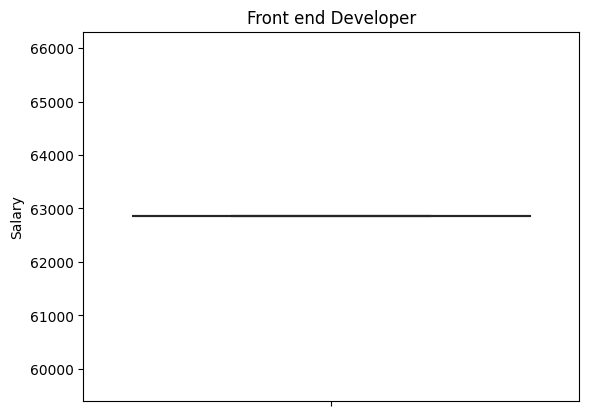


Fig 51

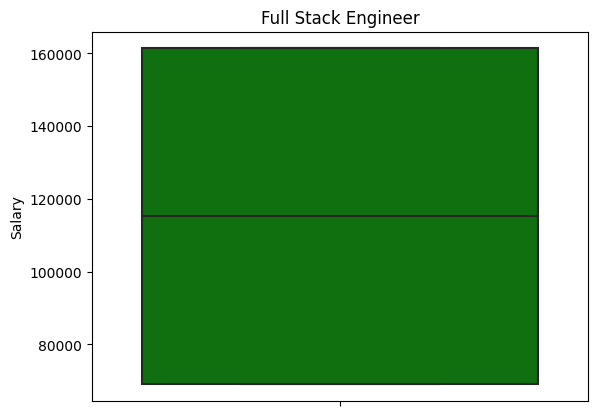


Fig 52

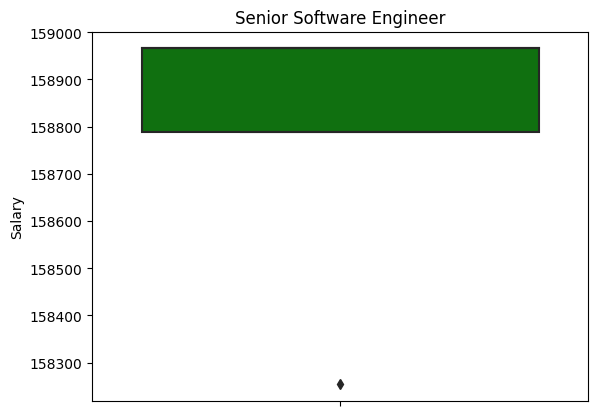


Fig 53

Outliers don’t always show up in job title figurative analysis thus not every job title is highly distributed, a lot of variation

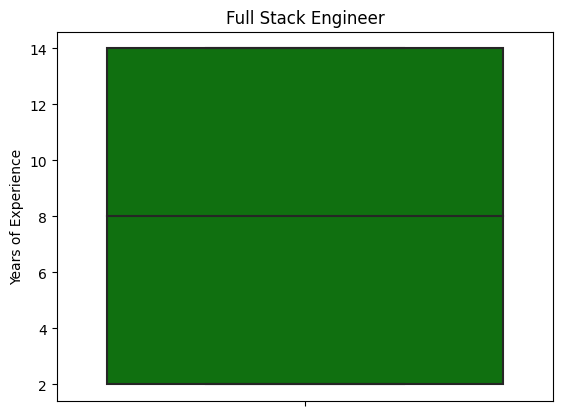


Fig 54

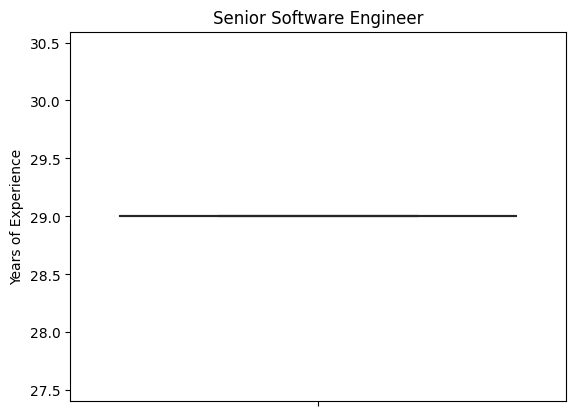


Fig 55

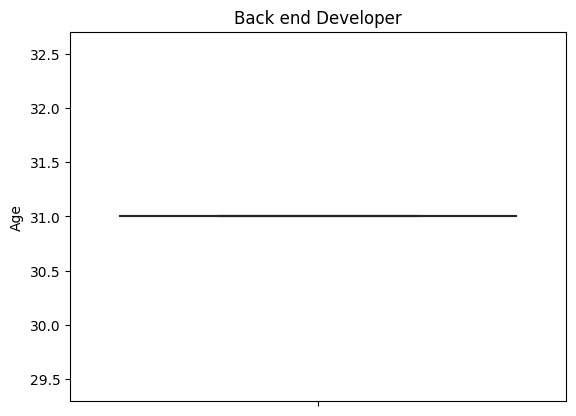


Fig 56

Age range varies within 27 to 45 the most

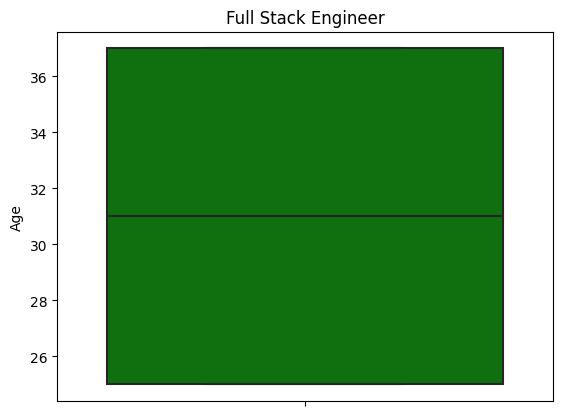
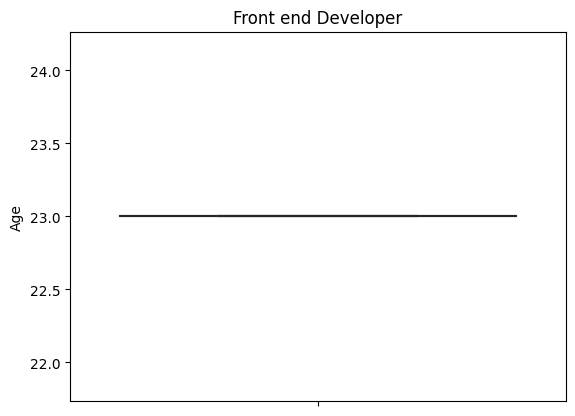


Fig 57



# Conclusion

2 columns with no data points, just null values.

For Male, Female and Other gender, age vs years of experience, age vs salary, salary vs years of experience yielded positive results.

There are points when lower educational degree holders earn more than higher educational degree holders.

Widely varied job titles, no relation with salary in particular.

But Phd, Master;s earns more overall but not when we see things from a job-specific level.

1. **Provide the Link of your Colab Notebook.**

<https://colab.research.google.com/drive/1NBi5X4G3j7_ifTMRsJhofpg_5NR_bgix#scrollTo=q_gVGmujVCZt>

<https://colab.research.google.com/drive/1b9i-BqPfTnhrEQOqZsDyRz0BauIaVATc>

<https://colab.research.google.com/drive/1RbdOsDP1Jsk6SePMki93VuYZWMw-Na9s?usp=sharing>