Salary Analysis: Exploration of data

About:

In this analysis i will show what matters when exploring this dataset of about 6700 entries. I have split this analysis in to three parts showcasing different data from one dataset.

- Who participated in the survey?
- Relationships between a single factor and salary.
- In-depth charts looking at salary, years of experience, gender & more.

All data analysis is built on this one dataset which might not reflect the real world.

Data reading and Preprocessing:

Taking a quick look at the data i found some things that needed to be fixed. Biggest thing that needed to be fixed were inconsistent values for all entres based on their Education Level.

In [1]:

```
salary_data['Education Level'].value_counts()
```

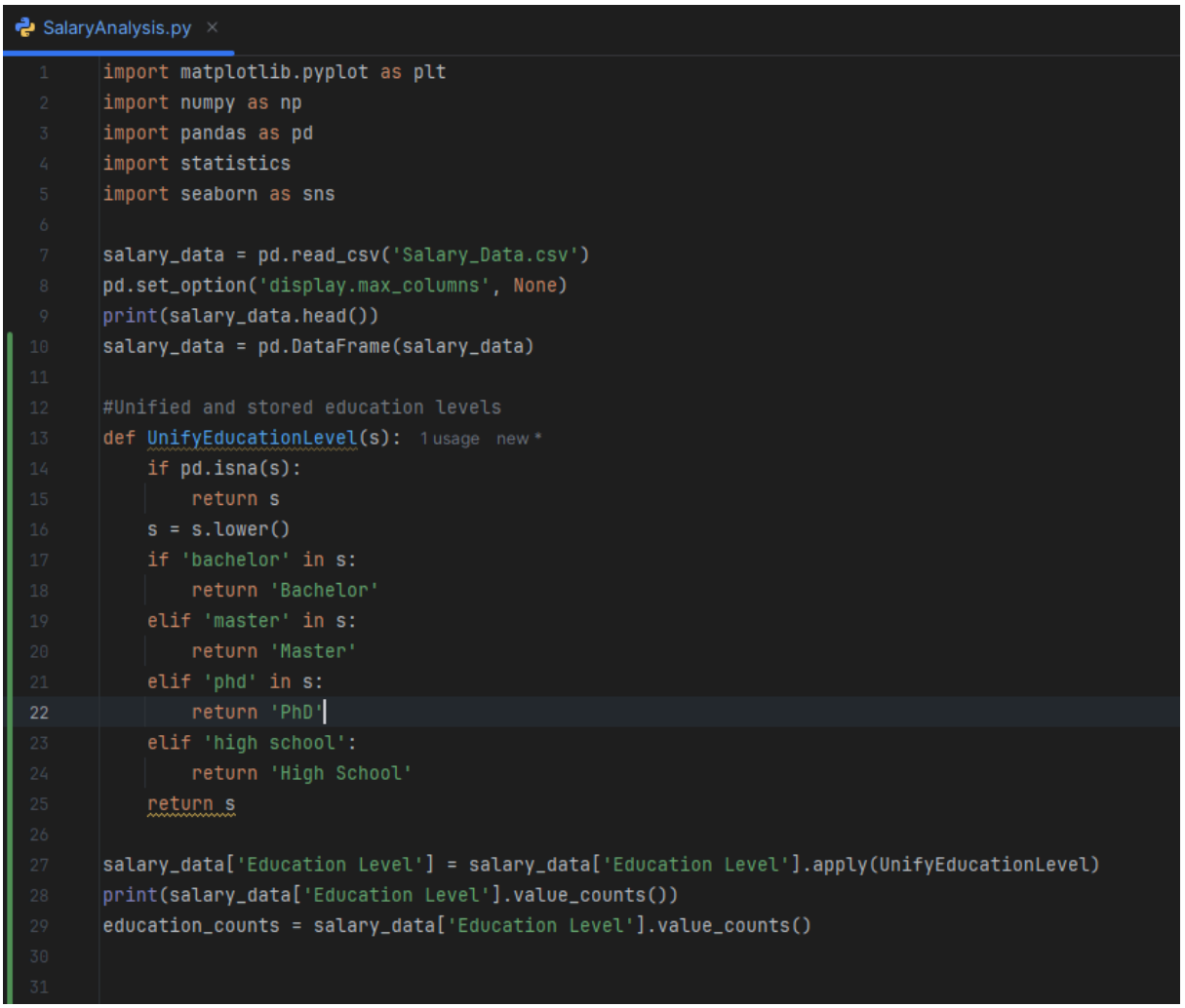
Out [1]:

```
Bachelor's Degree    2265
Master's Degree      1572
PhD                  1368
Bachelor's            756
High School           448
```

```
Master's          288
phD               1
Name: Education Level, dtype: int64
```

As we easily can notice here we have incosistent values for the same degree just entered differently with all the entries. To solve this problem you can unify entries

In [2] :



Out [2]:

```
C:\Users\xioni\PycharmProjects\PythonProject2\.venv\Scripts\python.exe C:\Users\xioni\PycharmProjects\PythonProject2\SalaryAnalysis.py
    Age  Gender Education Level         Job Title  Years of Experience  \
0  32.0    Male      Bachelor's  Software Engineer                 5.0
1  28.0  Female        Master's       Data Analyst                 3.0
2  45.0    Male             PhD     Senior Manager                15.0
3  36.0  Female      Bachelor's    Sales Associate                 7.0
4  52.0    Male        Master's           Director                20.0

     Salary
0   90000.0
1   65000.0
2  150000.0
3   60000.0
4  200000.0
Education Level
Bachelor       3023
Master         1861
PhD            1369
High School     448
Name: count, dtype: int64

Process finished with exit code 0
```

Using duplicated sum i found entries 4912 entries that were duplicates. Is it possible that they are duplicates? Of course but i kept them because i didn't want to lose around 70% of the survey and i couldn't garantee that they were duplicates.

In [3]:

```
salary_data.duplicated().sum()
salary_data.pivot_table(index = ['Age', 'Gender', 'Education Level',  'Job Titl
e', 'Years of Experience', 'Salary'], aggfunc ='size').sort_values().tail(10)
#salary_data.drop_duplicates()  # choose to keep them since it is possible t
o have duplicate dat
```

Out [3]:

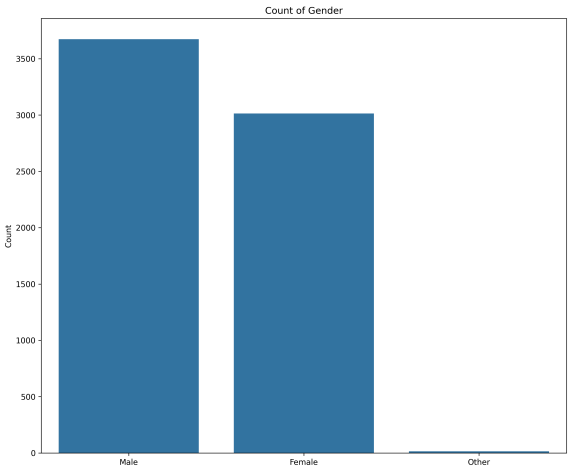
Age	Gender	Education Level	Job Title	Years of Experience	Salary	
29.0	Male	Bachelor	Marketing Analyst	4.0	70000.0	33
26.0	Male	Bachelor	Data Analyst	3.0	130000.0	35
25.0	Male	Bachelor	Product Manager	1.0	60000.0	36
27.0	Male	Bachelor	Software Engineer	4.0	140000.0	37
29.0	Female	Master	Data Scientist	6.0	180000.0	37
33.0	Female	Master	Product Manager	11.0	198000.0	38
32.0	Male	Bachelor	Software Engineer	8.0	190000.0	39

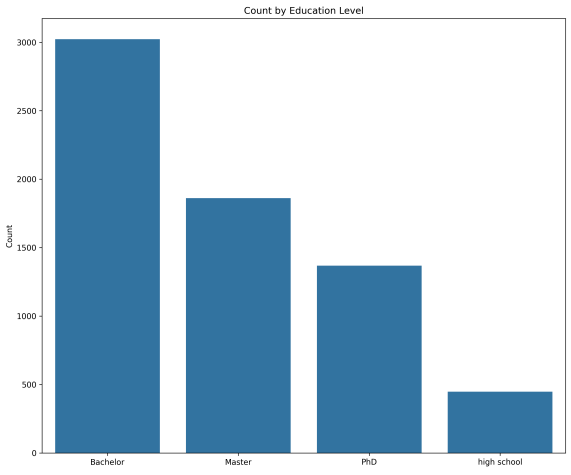
```
27.0  Male    Bachelor          Software Engineer  3.0          80000.0    45
24.0  Female  High School       Receptionist       0.0          25000.0    45
32.0  Male    Bachelor          Product Manager    7.0          120000.0   45
dtype: int64
```

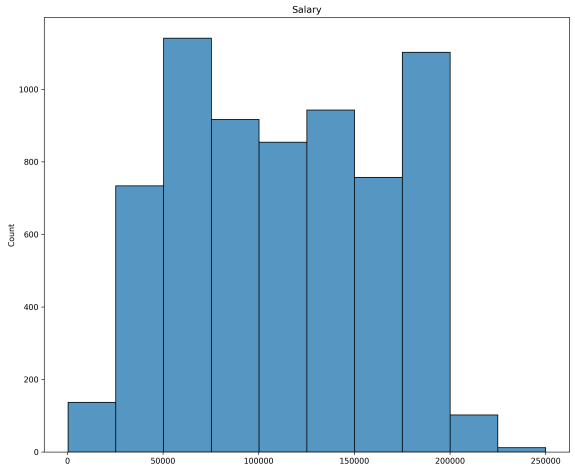
Part 1. Who participated in the survey?

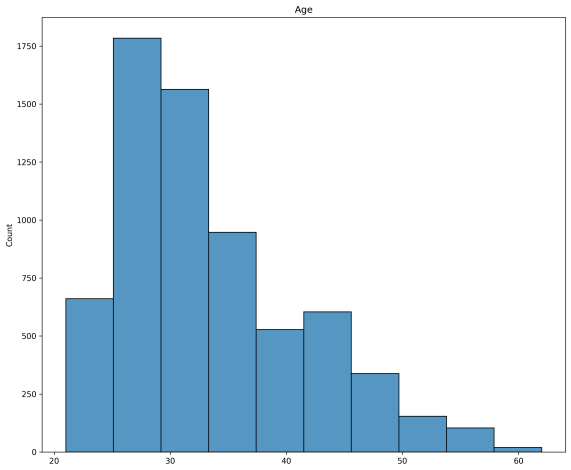
Data points we can confirm from the analysis.

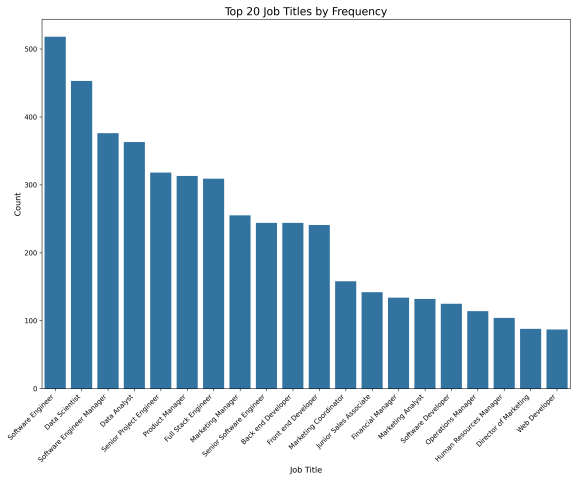
- Most of the jobs seem to be in the IT sector.
- Mot frequent age seem to be between 25 - 35.
- Male count of participants is a bit higher than female.
- Majority of participants hold a Bachelor's degree.
- Salary seem to range between 50 000 and 200 000 for the majority.
- Majority of participants years of experience seem to range between 0 - 10. After 10 years it's a fall off.

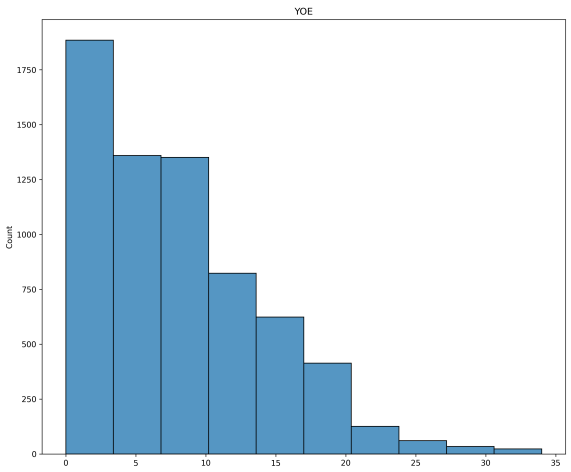








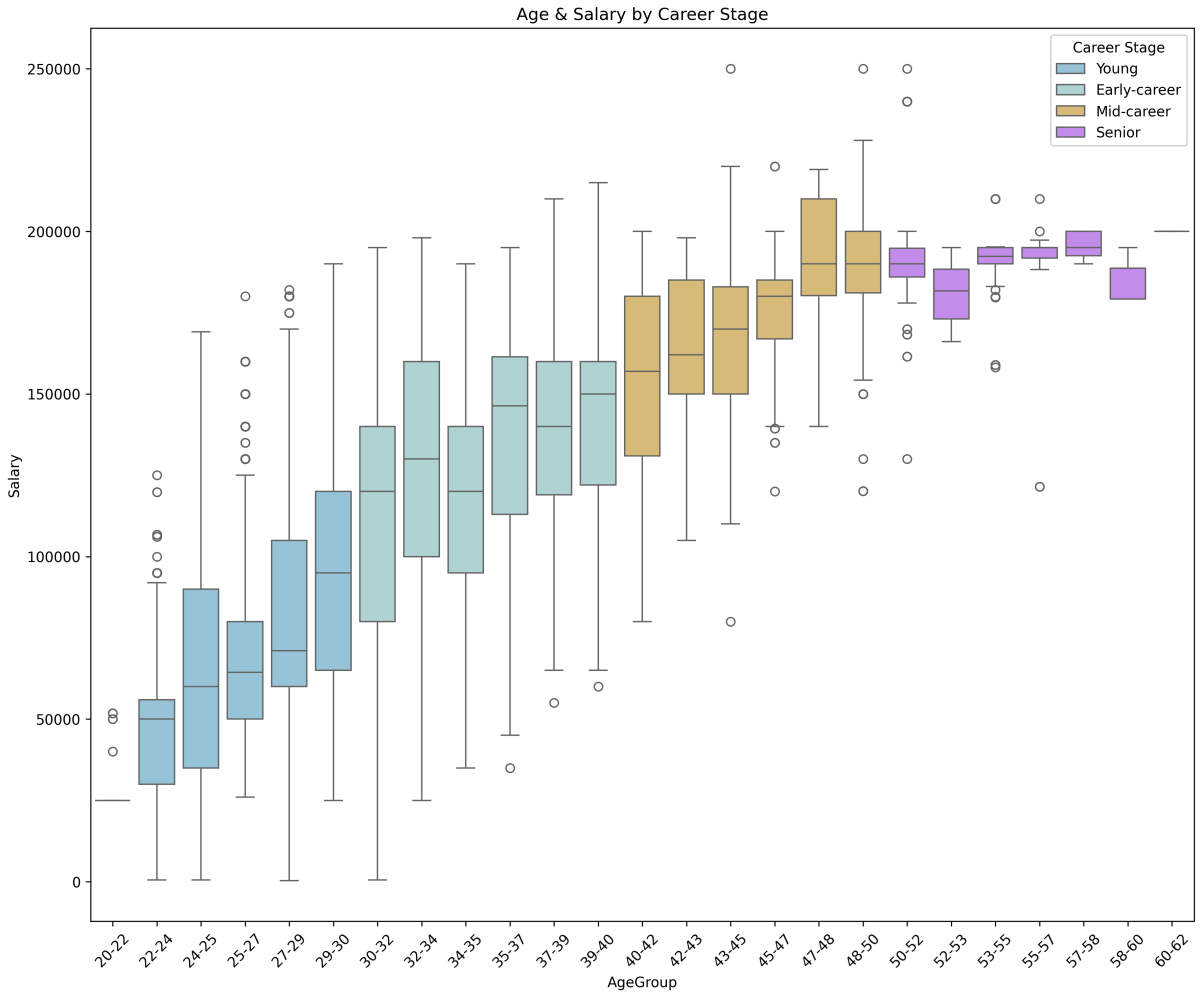


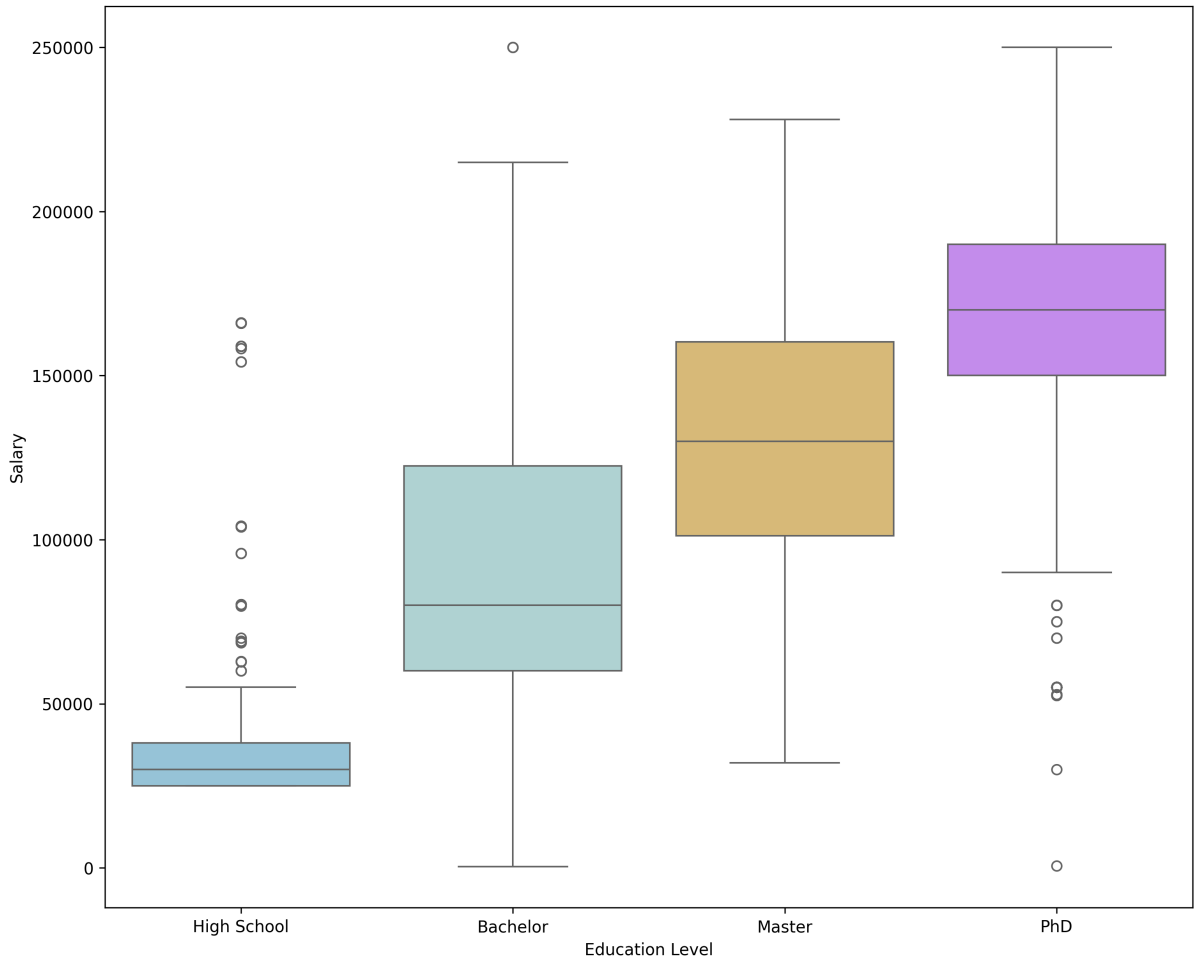


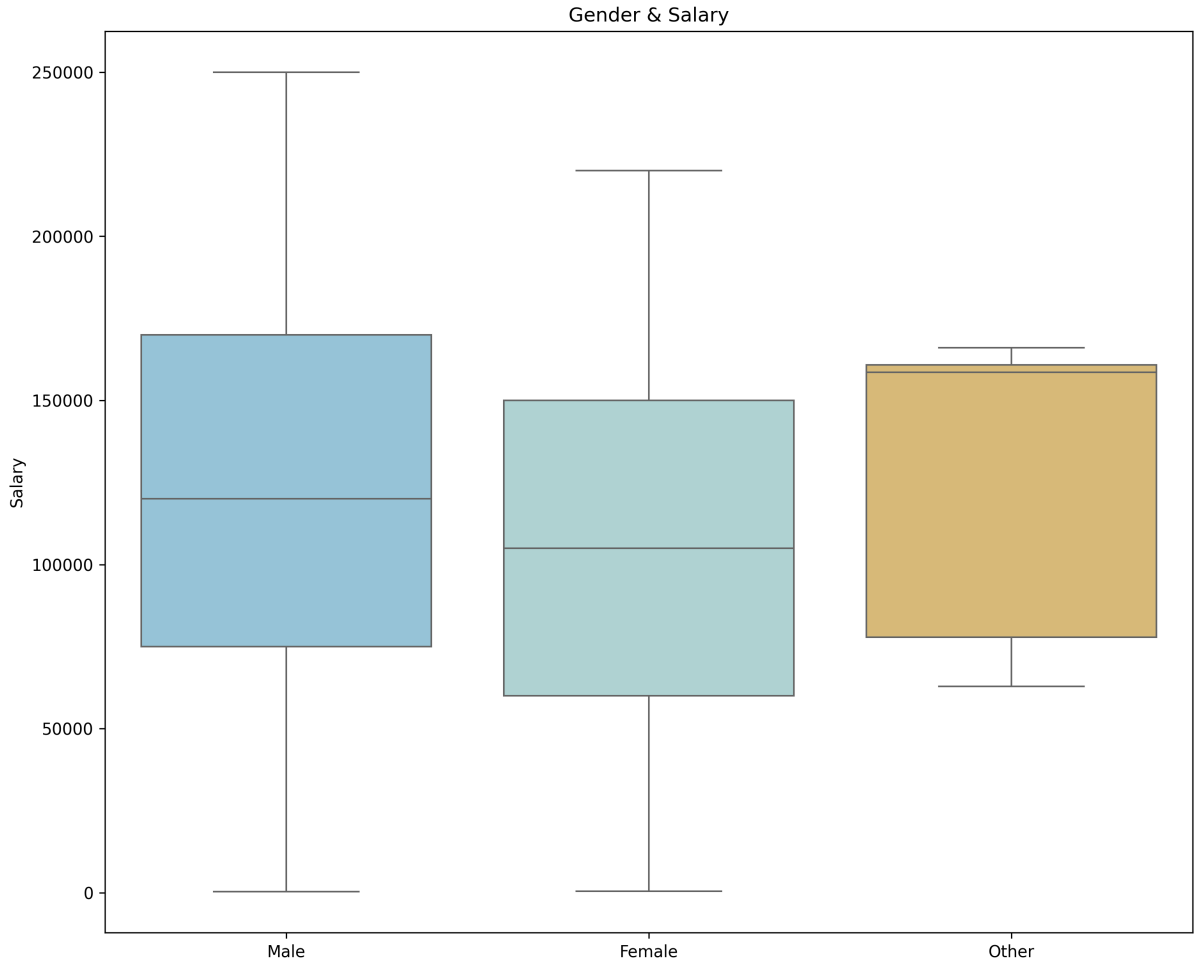
Part 2. Relationships between a single factor and salary.

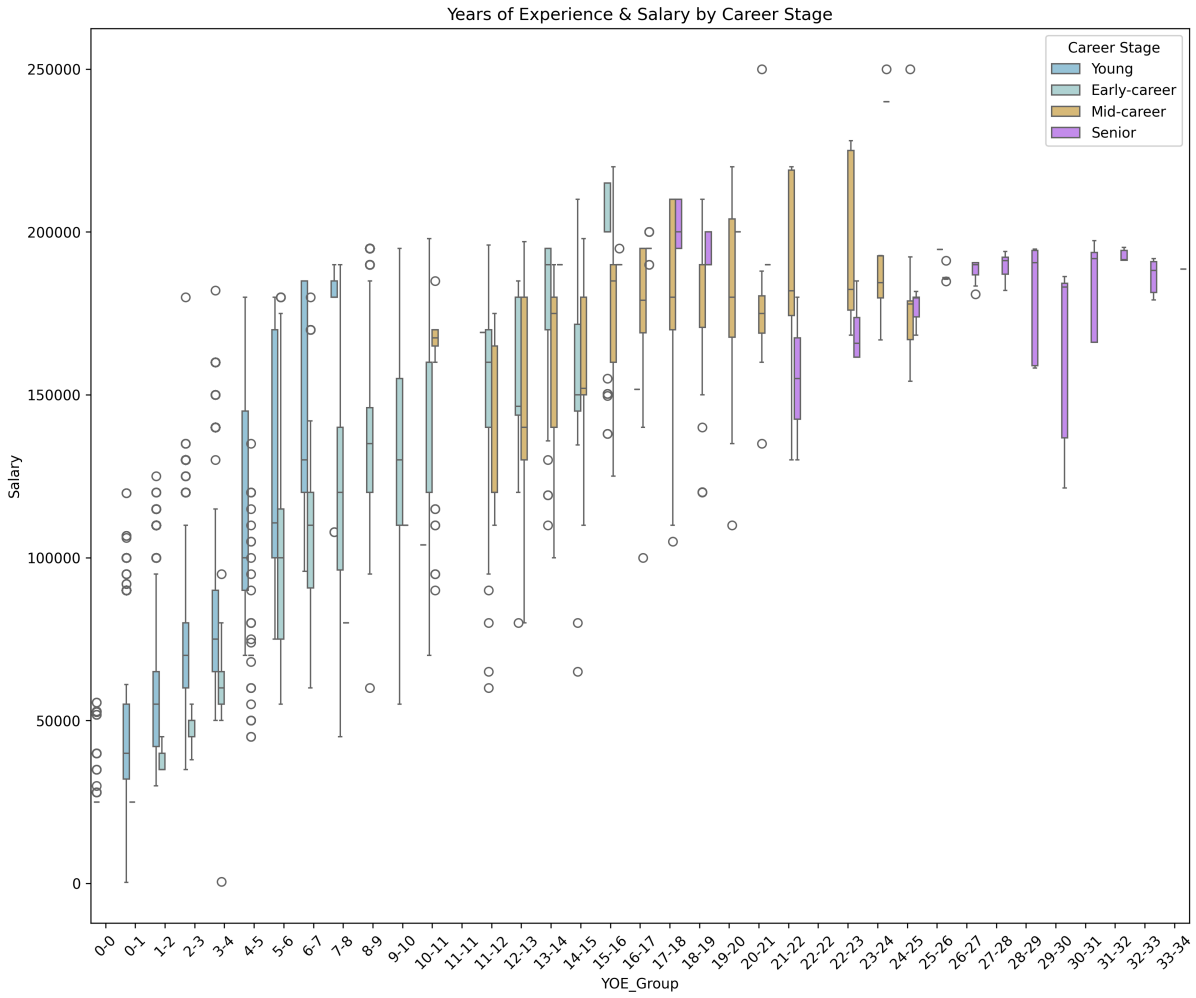
Data points we can confirm from the analysis.

- Salary increases with age and years of experience.
- Education matters in deciding your salary.
- The salary of male participants is higher than that of female participants


Age & Salary by Career Stage
Career Stage
Young
Early-career
Mid-career
Senior
Salary
250000
200000
150000
100000
50000
0
20-22
22-24
24-25
25-27
27-29
29-30
30-32
32-34
34-35
35-37
37-39
39-40
40-42
42-43
43-45
45-47
47-48
48-50
50-52
52-53
53-55
55-57
57-58
58-60
60-62
AgeGroup


250000
200000
150000
100000
50000
0
Salary
High School
Bachelor
Master
PhD
Education Level


Gender & Salary
250000
200000
150000
100000
50000
0
Salary
Male
Female
Other



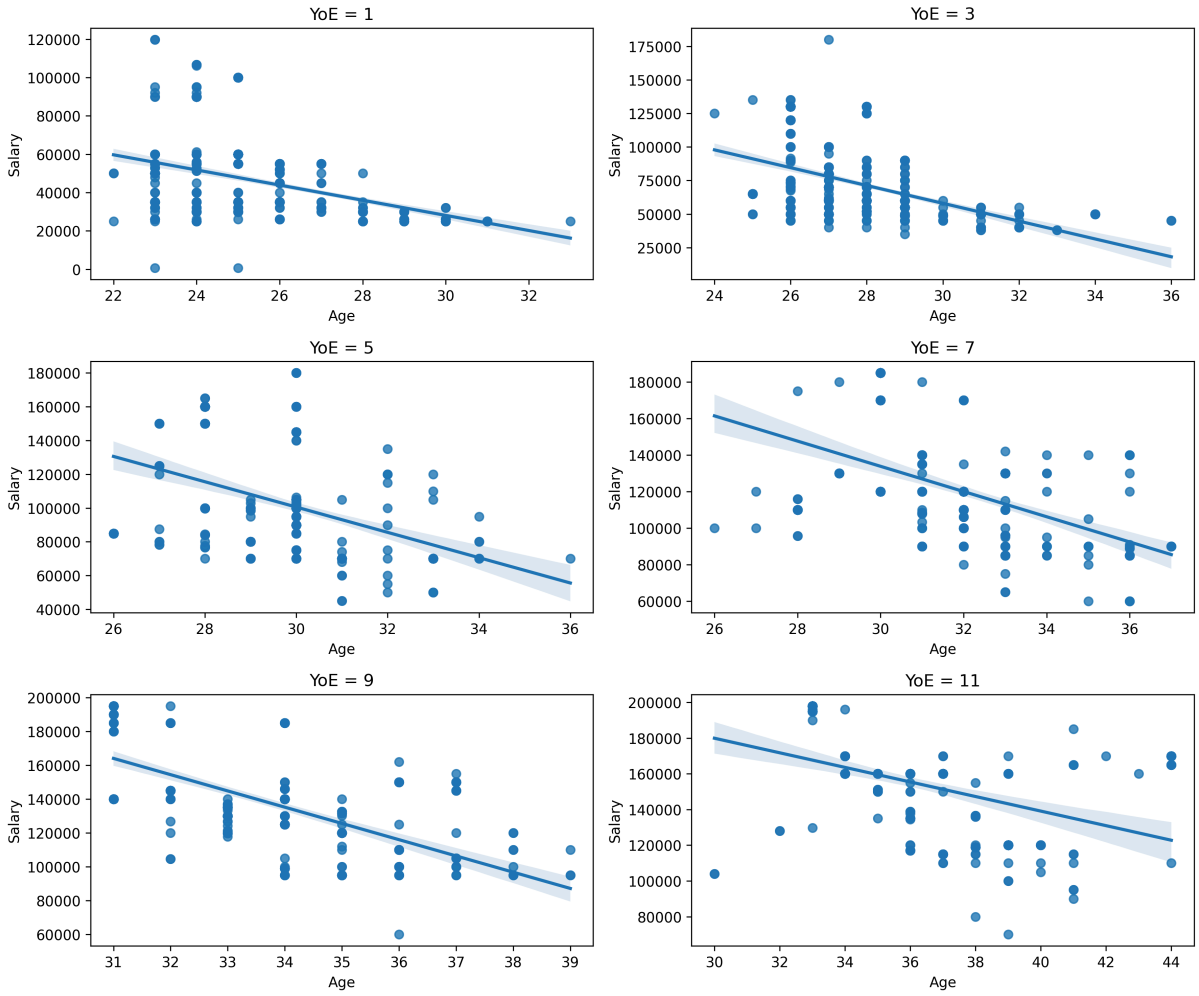
Part 3. In-depth questions

Data points we can confirm from the analysis.

- Does Age really positively affect the salary?
- Gender of participants from 10 biggest jobs. Difference in salary?
- Education of both genders and is there a difference?

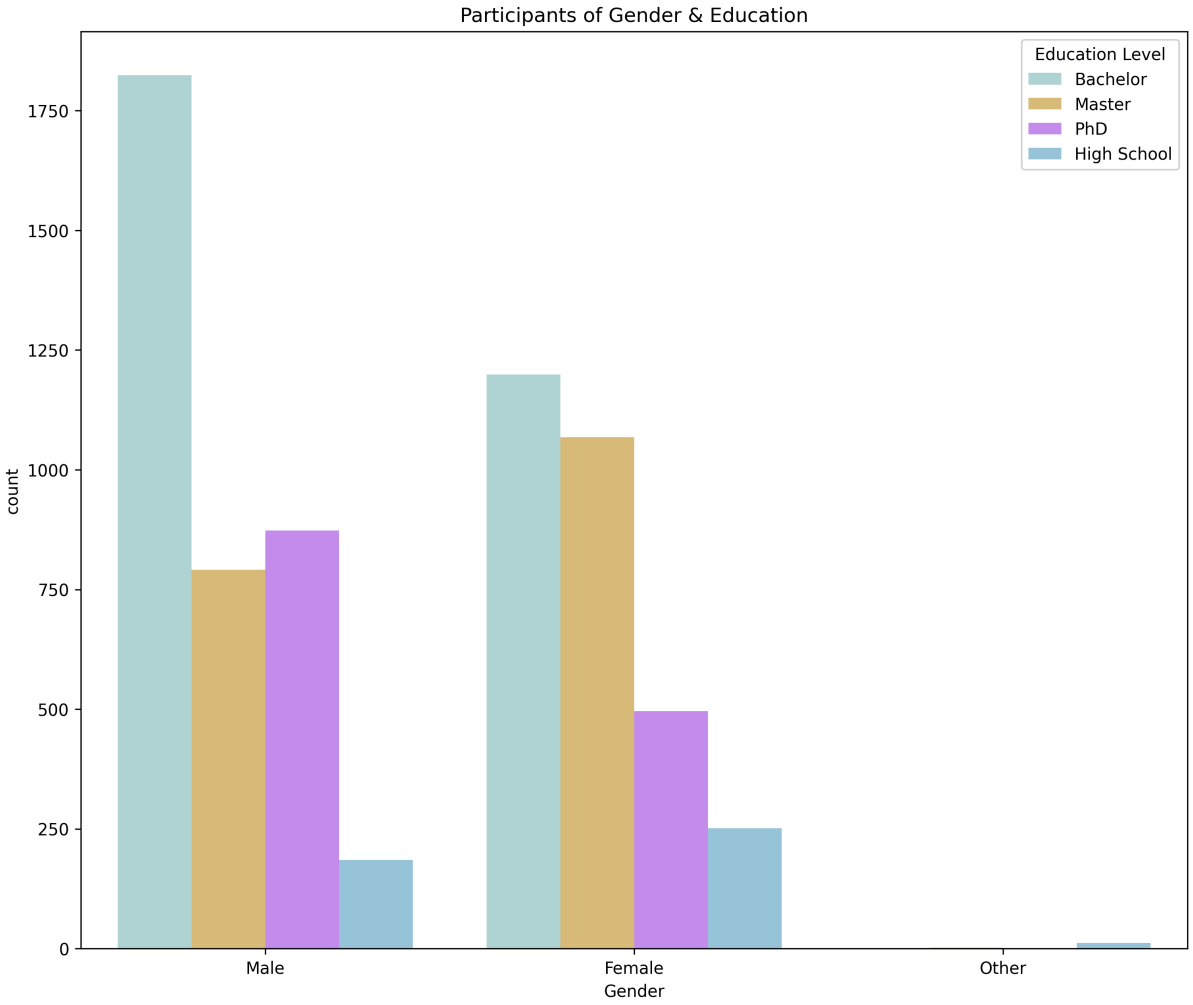
Age and salary

We can expect that salary and YOE (years of experience) have a positive correlation but is it the same case for age? The truth is that when the YOE is the same, the age and the salary have a negative correlation.



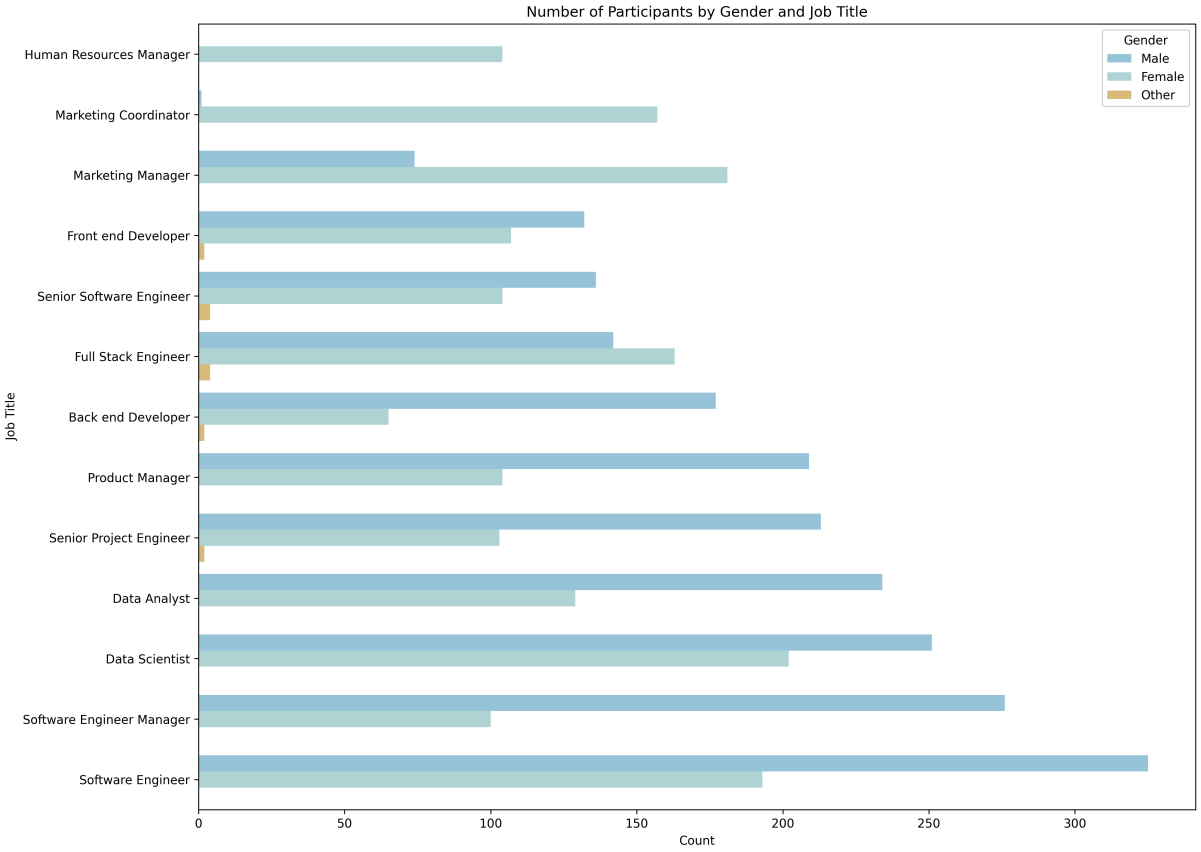
Participants gender and education

Diving in to the data as we visualize earlier we found that on average male participants earned more than the average female. Now we're diving in to how the degrees are divided by the genders. Looking at the data we can se that the men on average hold more bachelor's degrees and PhD's while the women hold more Master's degrees.



How are the jobs distributed between the genders?

I analyzed gender distribution among the most popular job titles and found that roles like "marketing coordinator," "marketing manager," "human resource manager," and "full stack engineer" have more female than male participants. A box plot of salary ranges shows that the median salary for three of these female-dominated jobs is relatively lower. Additionally, in most roles, the median salary for females is lower than for males, except for "project manager." Looking at the most popular jobs, males hold the majority of these positions, which may explain why the average male salary is higher.



Final thoughts:

Dataset can be found in the github repository that contains all of the code for data processing and making of the charts. This has been a really good learning project for learning how to visualize data and understanding the factors that make a data analysis good.