In this project, I downloaded the dataset from Kaggle titled [D*ata Science Job salaries*](https://docs.google.com/spreadsheets/d/1IPS5dBSGtwYVbjsfbaMCYIWnOuRmJcbequohNxCyGVw/edit?resourcekey#gid=1625408792). I decided to work on this dataset because:

* There were 6 variables of the free form of text entry so cleaning had to be done.
* Dataset didn’t require any expertise to understand

The first thing I did was import the dataset into excel. This is to ensure that I know what the data is about, examine it, and do some data cleaning. This dataset contained 17 columns and 27876 rows. It includes the questions that are being asked need to be changed so I have changed them.

There is a column called salary but it had to be converted into only one currency so I choose US Dollars which I believe is for the sake of uniformity and also to aid analysis. I did it by creating a table in which given currencies were converted to USD, and VLOOKUP to call these currencies from that table.

Answers to the “What country do you live in?” segment needed some cleaning because this segment was created as a free text entry so different forms were used for representing the same country. Such as; United States, States, USA, US, U.S, u.s., etc., I used UPPER for capitalizing them all, SUBSTITUTE for getting rid off the dots, and I wanted to create shortened versions of each country. The US for the United States, the UK for the United Kingdom, and CA for Canada, I choose these countries because most of the answers were from these three countries when I counted them.

I used to shorten the United States and the United Kingdom with the formula down below:

=IFERROR(IF(SEARCH("sta",M2),"US",IF(SEARCH("king",M2),"UK",M2)),M2)

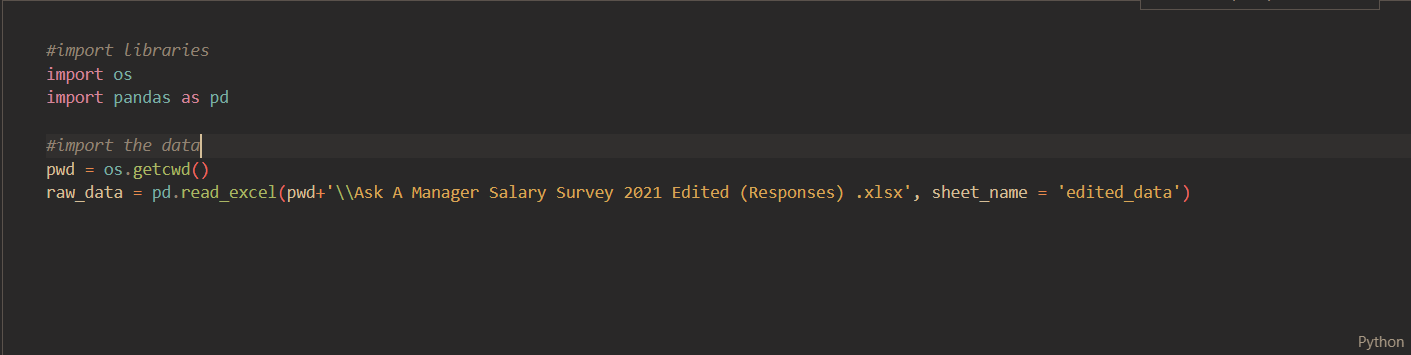
However, when I started my formula was this:

=IF(FIND(“sta”,M2),”US”,M2)

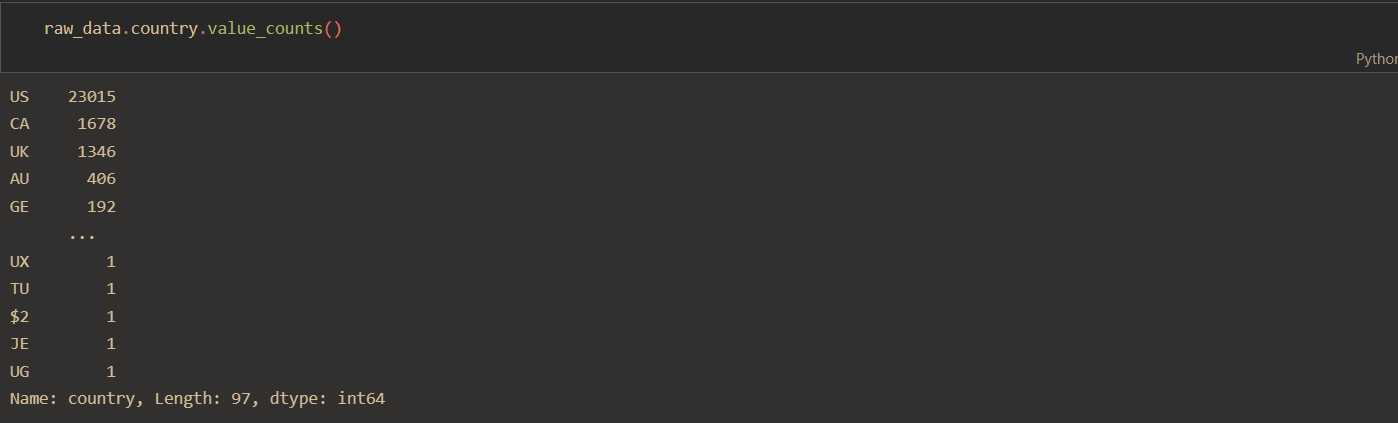
But I didn’t get the results that I got because the FIND function is case sensitive so I used SEARCH, and the formula evolved from there.

Anyhow, there were a minor amount of given different names for America that I decided not to save because it would make much difference when given data size significantly small. I have done the same with currencies.

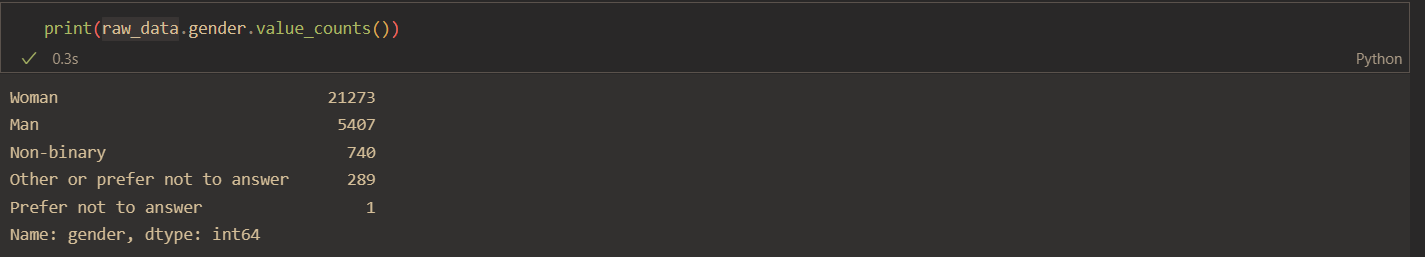
After, I finish cleaning I uploaded it to Python.



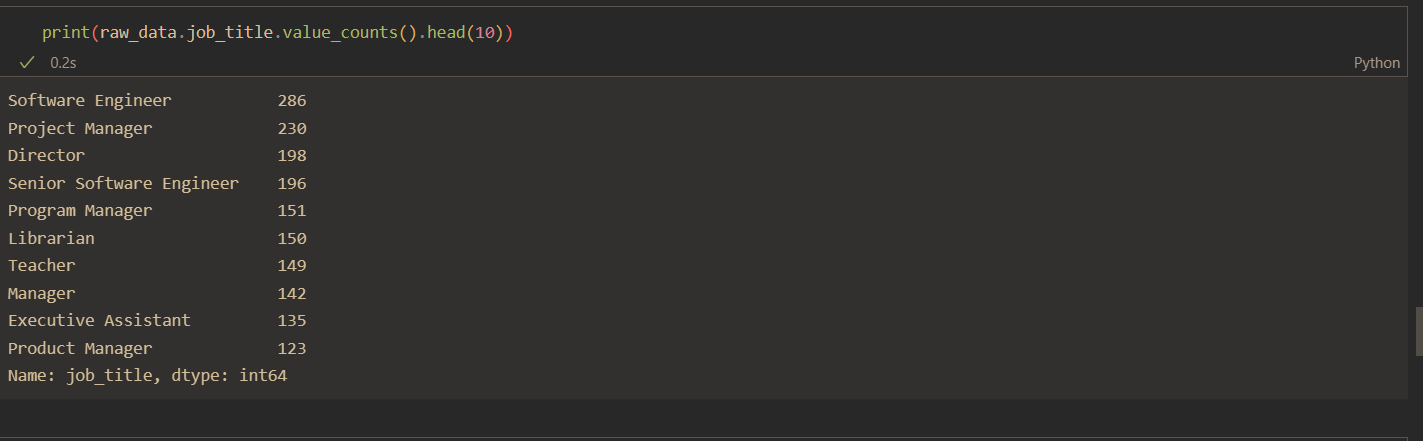
First, I started with counting the numbers of employees by country and work on the one on which we have more data. Second, I’ll count the number of employees by job title and finally, count the total of men and women.



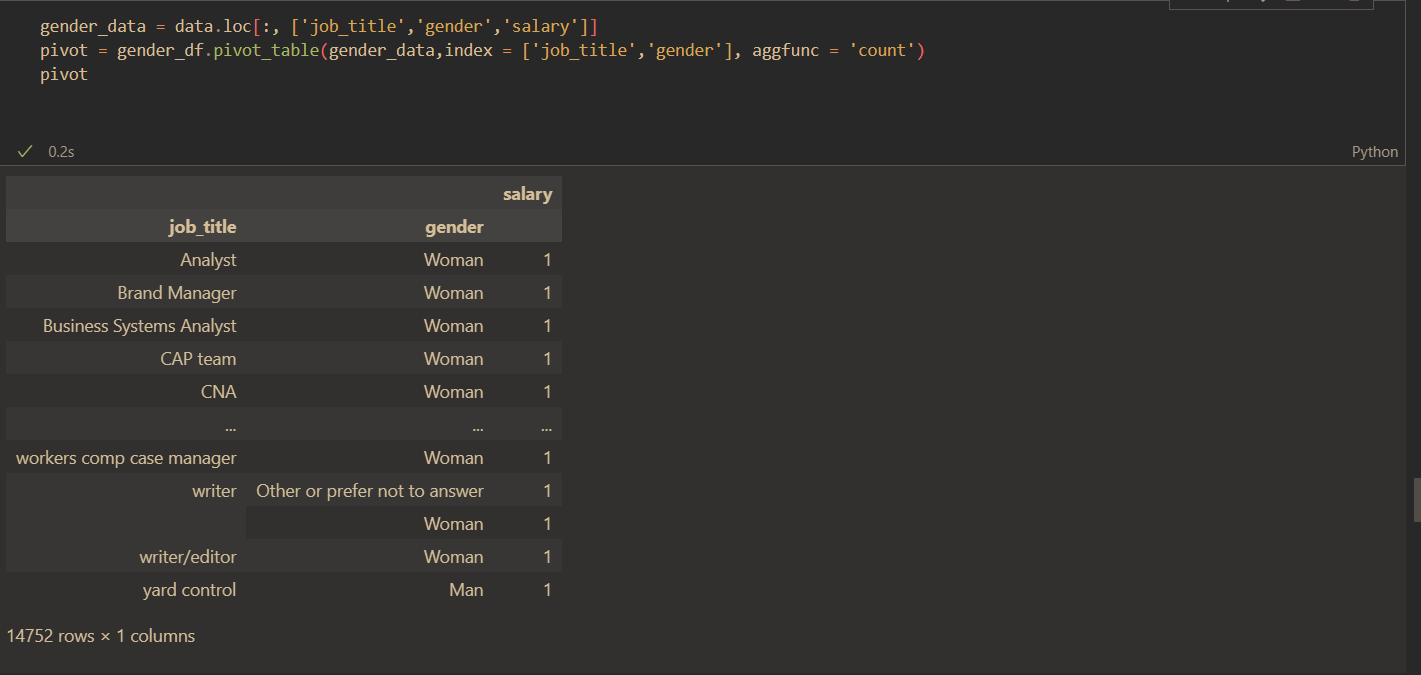
As we can see, there are way more employees from the US, Canada, and the UK in this survey and, because of that, I have filtered out the data in order to work only with these countries within the data.



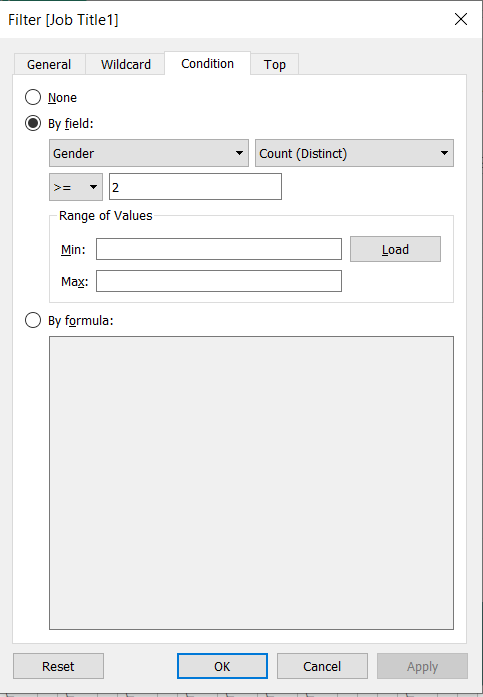
It seems that we have an issue here. Unfortunately, the total number of females is way greater than the total of males. Also, we are not able to get any statistical information from the Prefer not to answer due to the fact that only one person answered the survey in 2019. So, I dropped this specific row.

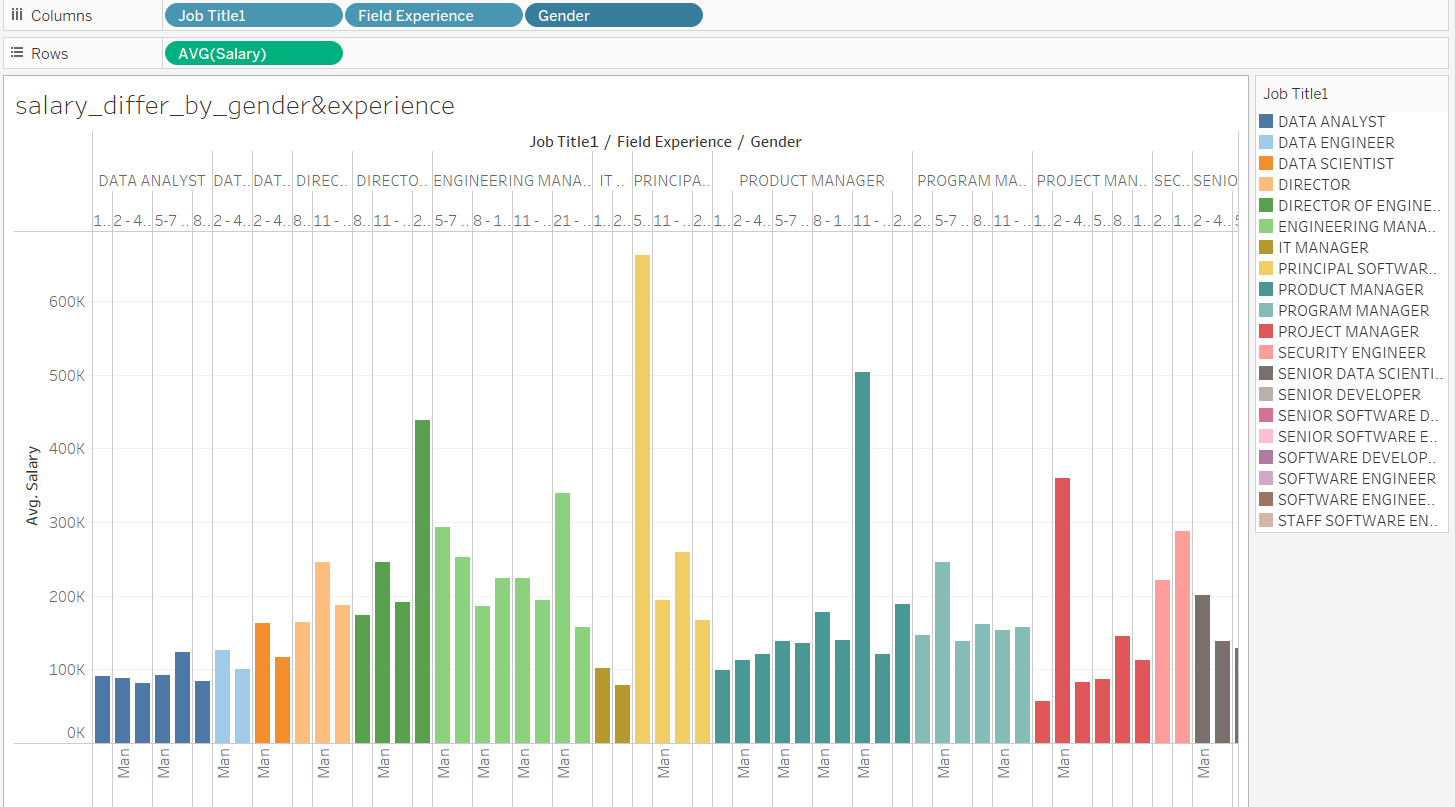


These are the most common job titles from the dataset. However, when I pivot the dataset to find out the gender ratio among the job titles, I have got surprised because there were 14752 rows.



After my failed attempts to group them under job\_title and gender columns, I have decided to move to Tableau to filter the data where there are at least two given gender for a job title. I have excluded the rows by adding a condition under the job title column.



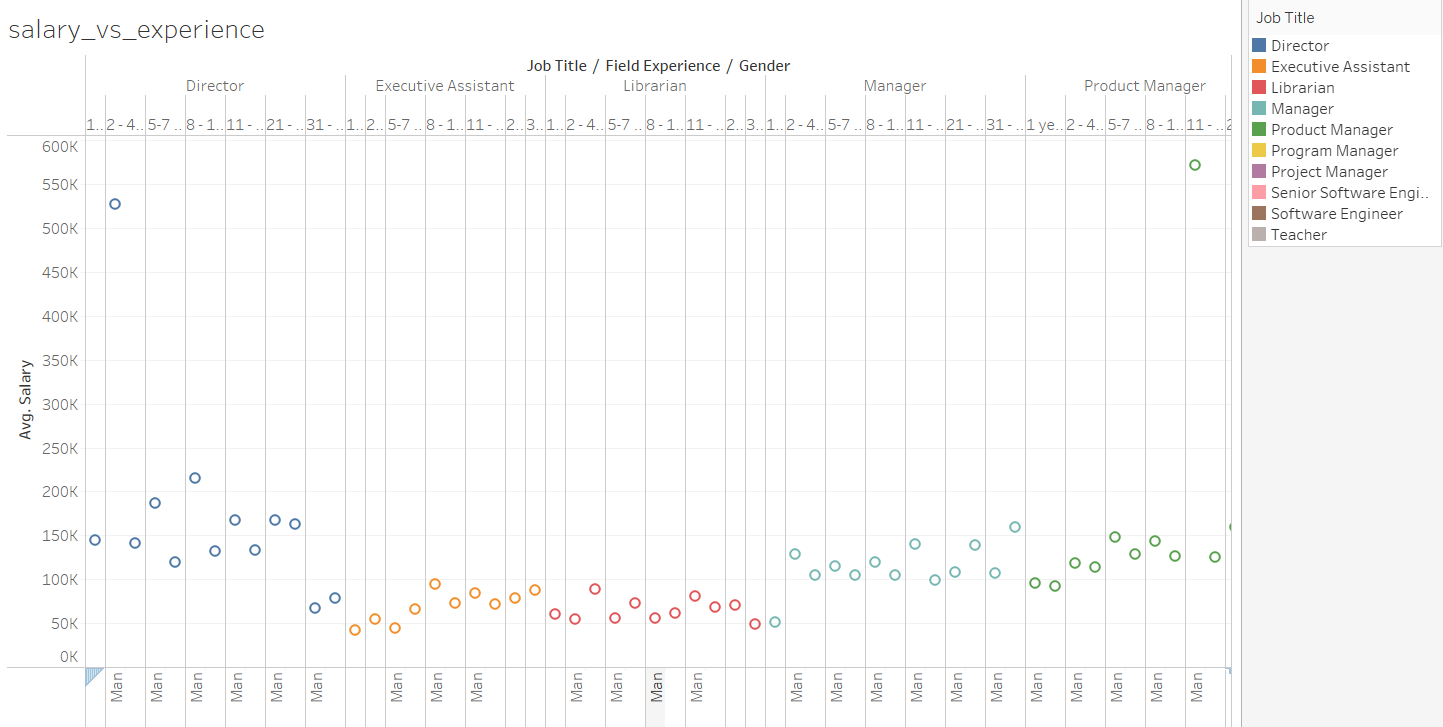


Since I finally got what I wanted, meaning tidying up the all data, now, I can answer the questions.

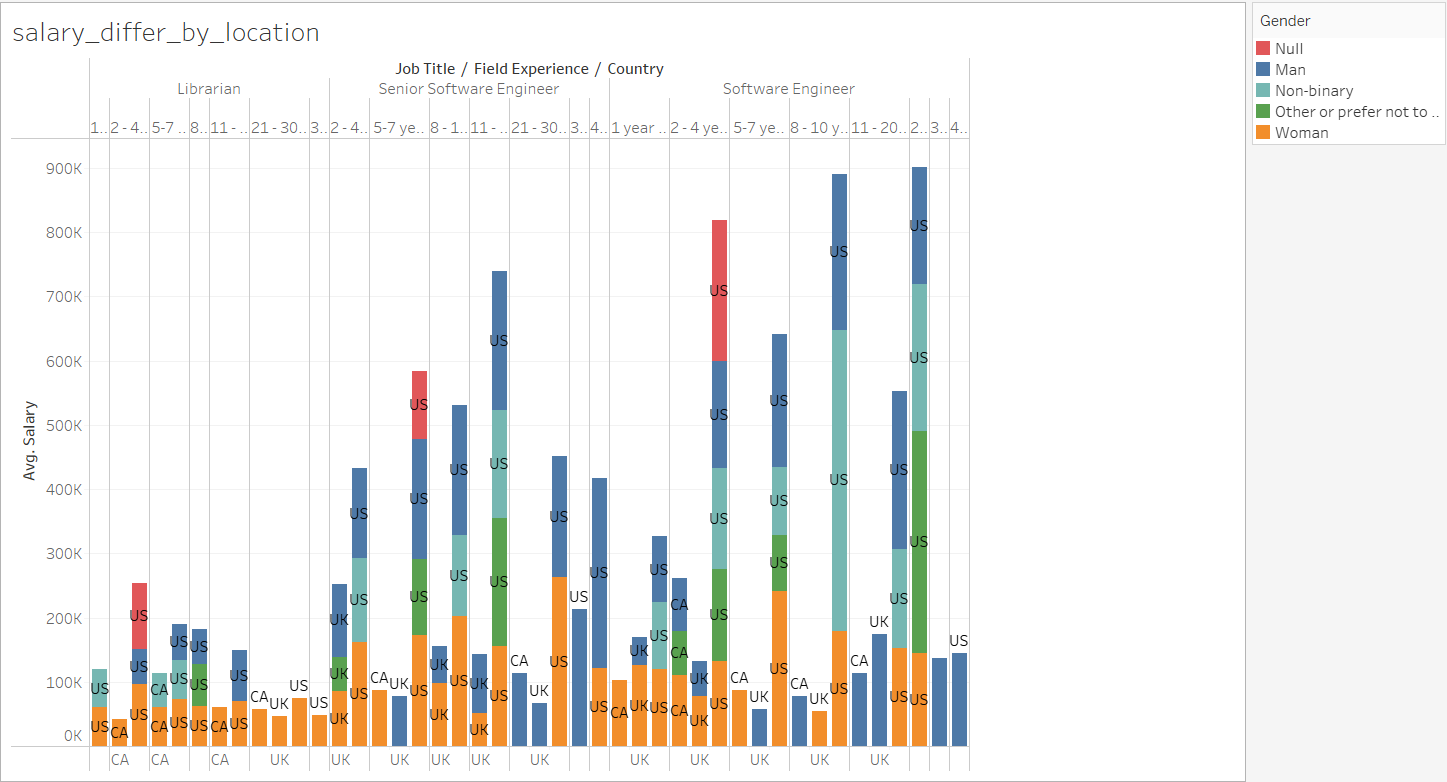
1. Which industry pays the most?
2. How does a salary increase compare to the experience?
3. How do salaries compare for the same role in different locations?
4. How much do salaries differ by gender and years of experience?

**Question:** Which industry pays the most?

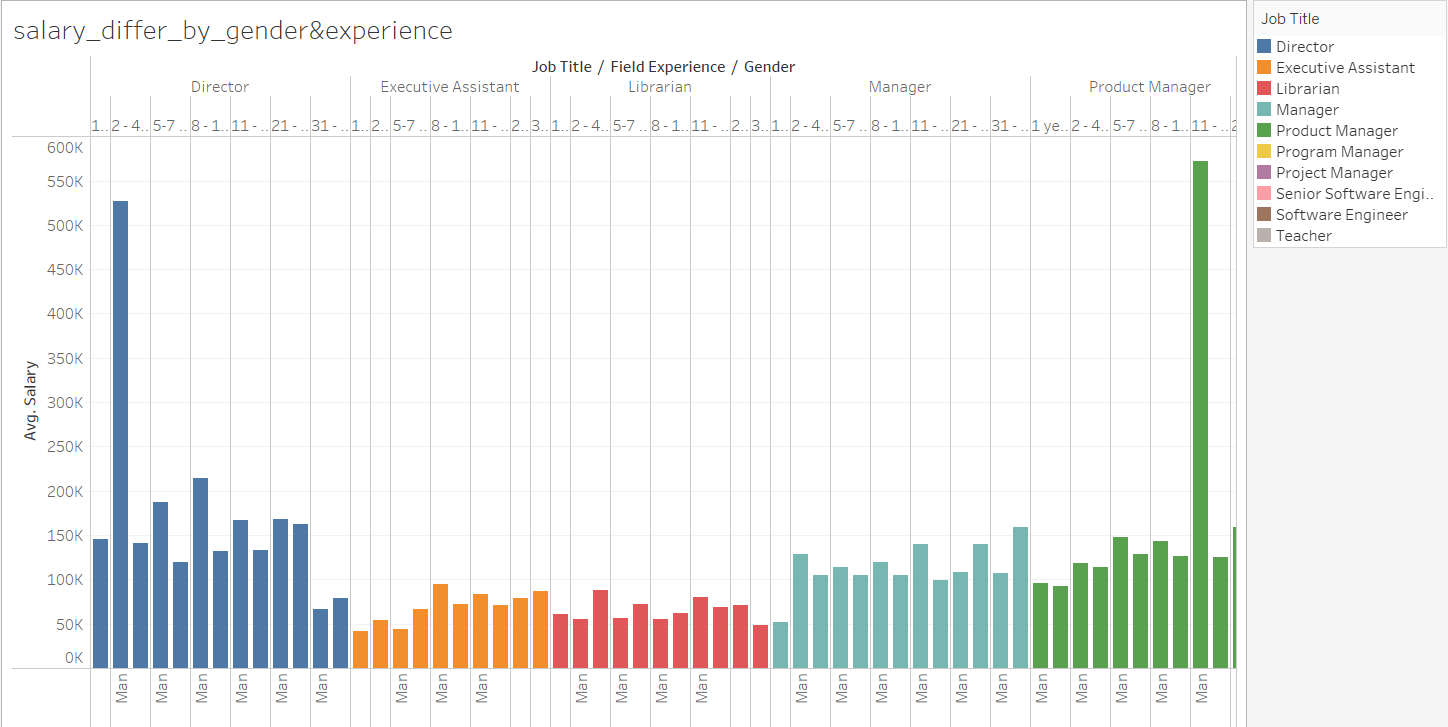


**Answer:**How does a salary increase compare to the experience?****

**Question:** How do salaries compare for the same role in different locations?

****

**Question:** How much do salaries differ by gender and years of experience?



**Conclusion:**

Despite the fact, that it is survey data, accuracy and accountability are somewhat questioned, but the data shows us some insights.

After I have done the analysis of the dataset, I would say it seems to me that both genders are getting paid equally. However, I have noticed that some men and women make significantly more money which creates anomalies. Such as, the average salary decreases over years for some professions because of anomalies either in the beginning or in the middle average salary shoots up.

The US is paying more for each profession compared to Canada or UK.

There is a direct correlation between years of experience and salary. The average salary increases as the experience years grow higher and higher.