Parthiv Majumdar

Power BI Project

Data-Related Job Market Analysis

DATA DEFINITION

We have 2 files (Glass Door Jobs & Eda Data), we will just combine all needed data from the two files into the **Cleaned Jobs.csv** file, clean it, and work on that file.

Firstly, let me introduce to you the columns we have in the cleaned Glass Door file:

- **Job ID:** The unique identifier for the job posting (Numeric)
- **Job Title:** The title of the job (Text)
- **Min Salary:** The minimum portion of the salary range for this job title (Numeric)
- **Max Salary:** The maximum portion of the salary range for this job title (Numeric)
- **Avg Salary:** The average of the min and max salary (Numeric)
- **Hourly:** Indicates if this employee works per hour (0 & 1)
- **Job Description:** The description of the job (Text)
- **Company:** The name of the company posting the job (Text)
- **Location:** The city where the company resides (Text)
- **Rating:** The rating of the company on glass door (Numeric)
- **Size:** The size of the company speaking about employees' count (Text)
- **Size Class:** A flag from class A to class G where A is a bigger size company (Text)
- **Founded:** The foundation year of the company (Numeric)
- **Age:** The age of the company (Numeric)
- **Type Of Ownership:** Ownership type of the company (Text)
- **Competitors:** The companies that compete with this company (Text)
- **Industry:** The industry of that job (Text)
- **Sector:** The sector to which the company belongs (Text)
- **Revenue:** The average yearly revenue of the company (Text)
- **Revenue Class:** A flag from class 1 to class 13 where 1 is a company with highest revenue range (Text)
- **Same State:** A flag represents whether the employee is from the same city as the company (0 & 1)
- **Python:** A flag represents whether the employee has Python skills (0 & 1)
- **R:** A flag represents whether the employee has R skills (0 & 1)
- **Spark:** A flag represents whether the employee has Spark skills (0 & 1)
- **AWS:** A flag represents whether the employee has AWS skills (0 & 1)
- **Excel:** A flag represents whether the employee has Excel skills (0 & 1)
- **Skills Count:** The number of stated skills the employee has (Numeric)
- **Job Simp:** The job title just cleaned and reorganized (Text)
- **Comp Num:** Calculated Column holding the number of competitors (Numeric)

DATA CLEANING

Let's work column by column to achieve the target state.

Estimated Salary

- Drop rows with missing values flagged by (-1) value.
    - Being the core of our analysis there is no tolerance here, even be clean or you will be dropped.
    - There were 214 rows with missing values, all were dropped.
- 53K – 91K (Glassdoor est.) is not a pretty form, so I will convert that and split it into 2 columns:
    - Min Salary: Holding the lower limit.
    - Max Salary: Holding the upper limit.
- For people working per hour, we will estimate their salary as (1920 * Salary) as 1920 is from working 40 hours a week for a year.
    - Not a 100 % accurate estimate but I think it is the best one.
- Calculating the average salary for each row as it is the mean of the upper and lower limit of the salary range and store it in a new column.
- Job Number 741 is too dirty to keep, and has lots of missing values, so we better drop that column.

Company

Company name was stored like:

Name

Rate

We need to convert that to just hold the company's name.

Rate

This column has negative values, and as the data is scraped from the web, my guess is that those values just indicate a bad review, so we will just replace them with 0, indicating a bad rate.

Competitors

This column has tons of missing values (459), and the existing values is not organized, we may drop it, but I will leave it to get the number of competitors because I want to analyze that.

Size Class

In This Column we categorized each company according to its employees' count as the following:

- 10000+ employees — Class A
- 5001 to 10000 employees — Class B
- 1001 to 5000 employees — Class C
- 501 to 1000 employees — Class D
- 201 to 500 employees — Class E
- 51 to 200 employees — Class F
- 1 to 50 employees — Class G

Revenue Class

In This Column we categorized each company according to its revenue range as the following:

- $10+ billion (USD) — Class 1
- $5 to $10 billion (USD) — Class 2
- $2 to $5 billion (USD) — Class 3
- $1 to $2 billion (USD) — Class 4
- $500 million to $1 billion (USD) — Class 5
- $100 to $500 million (USD) — Class 6
- $50 to $100 million (USD) — Class 7
- $25 to $50 million (USD) — Class 8
- $10 to $25 million (USD) — Class 9
- $5 to $10 million (USD) — Class 10
- $1 to $5 million (USD) — Class 11
- Less than $1 million (USD) — Class 12

Age

In this column I have calculated the age of the company from its foundation's year.

Comp Num

A calculated column in which I calculated the number of competitors for the company calculated from the competitors' column, which have a lot of missing values, but we will handle that.

NOTES:

- The founded date has missing values represented by **-1**.
- Type Of Ownership has missing values represented by **unknown**.
- Size has missing values represented by **unknown**.
- Industry has missing values represented by **unknown.**
- Sector has missing values represented by **unknown**.
- Age has missing values represented by **-1**.
- I have changed all the 0/1 columns into Yes/No Columns.

The next step will be to join the 2 files.

Then we have some extra useful columns from the Eda Data file such as (Python – R – Spark -AWS – Excel – Same State), those features will help us analyze the data much deeper.

Our cleaned data had 30 columns.

That is how far my memory can recall about the cleaning I did.

Now the data is clean and ready for Cass A analysis, let's go.

Getting To know The Data Some More

There is approximately 10 Job Titles covering most of the data.

```
Data Scientist                                    131
Data Engineer                                      53
Senior Data Scientist                              34
Data Analyst                                       15
Senior Data Engineer                               14
```

Average Salaries vary from 15K to 254K.

There are only 24 people working per hour.

The job description column follows no pattern, so it will be useless.

We have 342 unique companies, and those are the most frequent companies.

Takeda Pharmaceuticals	14
MassMutual	14
Reynolds American	14
Software Engineering Institute	11
Liberty Mutual Insurance	10
PNN	10

We have 200 different locations, and those are the most frequent ones.

New York, NY	55

San Francisco, CA	49
Cambridge, MA	46
Chicago, IL	32
Boston, MA	23
San Jose, CA	13
Pittsburgh, PA	12

Ratings vary from zero to 5.

Most of the headquarters are in New York and San Francisco.

New York, NY	52
San Francisco, CA	42

We have only 10 types of ownership, mostly private/public company.

Company - Private	410
Company - Public	193

We have data for 60 different industries and 25 sectors, mostly IT (180 rows).

That is enough blind insights let’s visualize some data.

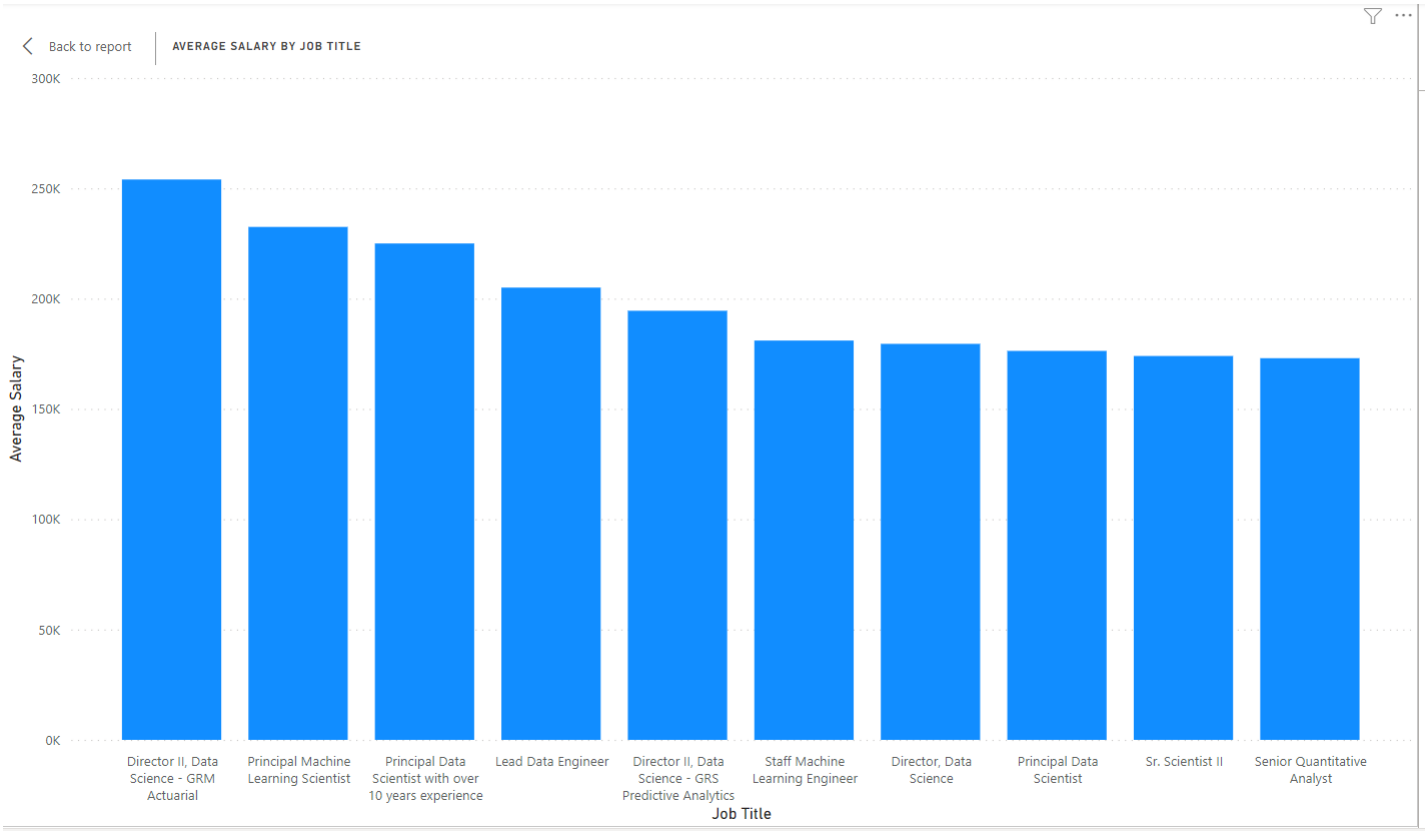
Getting Insights Exploratory DA

Let's first draw some visuals to help us deep dive into the data (Explore Page).

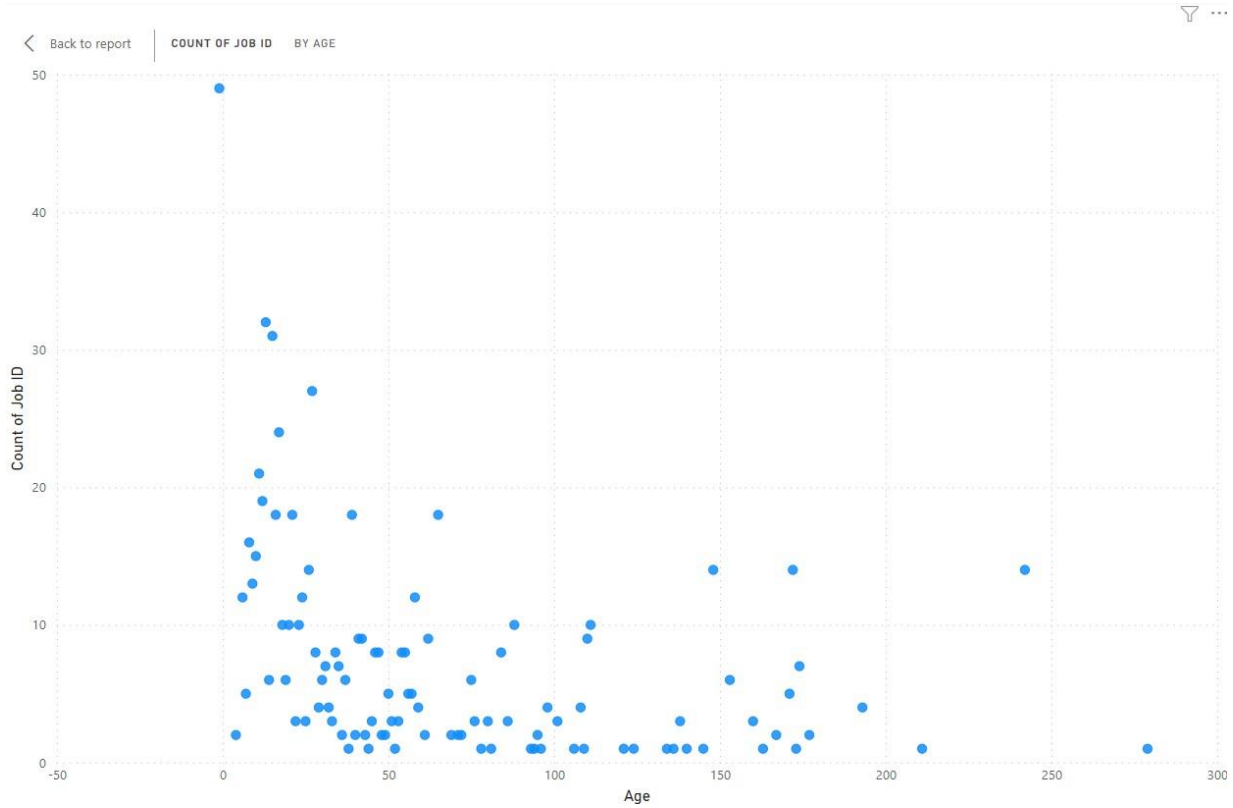
- We have tons of analysis to do, tons of stories to tell, and tons of insights to get, but for me I am mainly interested in the impact of having a specific skill, a combination of multiple skills, or even the skills counting on the average salary.
- In addition to that I want to see the variation of salary range for the same title over different companies.
- It will also be useful to see the locations having a higher salary range for every title.
- I will be insightful to see if companies with higher rates give out higher salaries.
- Mainly, we want to see the impact of the company size regarding the employees' count and the revenue range on the salary.
- Aged company offer better salaries?
- We want to see the industry or sector that has higher salaries for a specific job title.
- Finally, one of the most important ones is the impact of the number of competitors on the salaries offered by the company.

Enough talking let's deep into some visuals to understand the data more.

For me It seems that all job titles have semi-close salary ranges

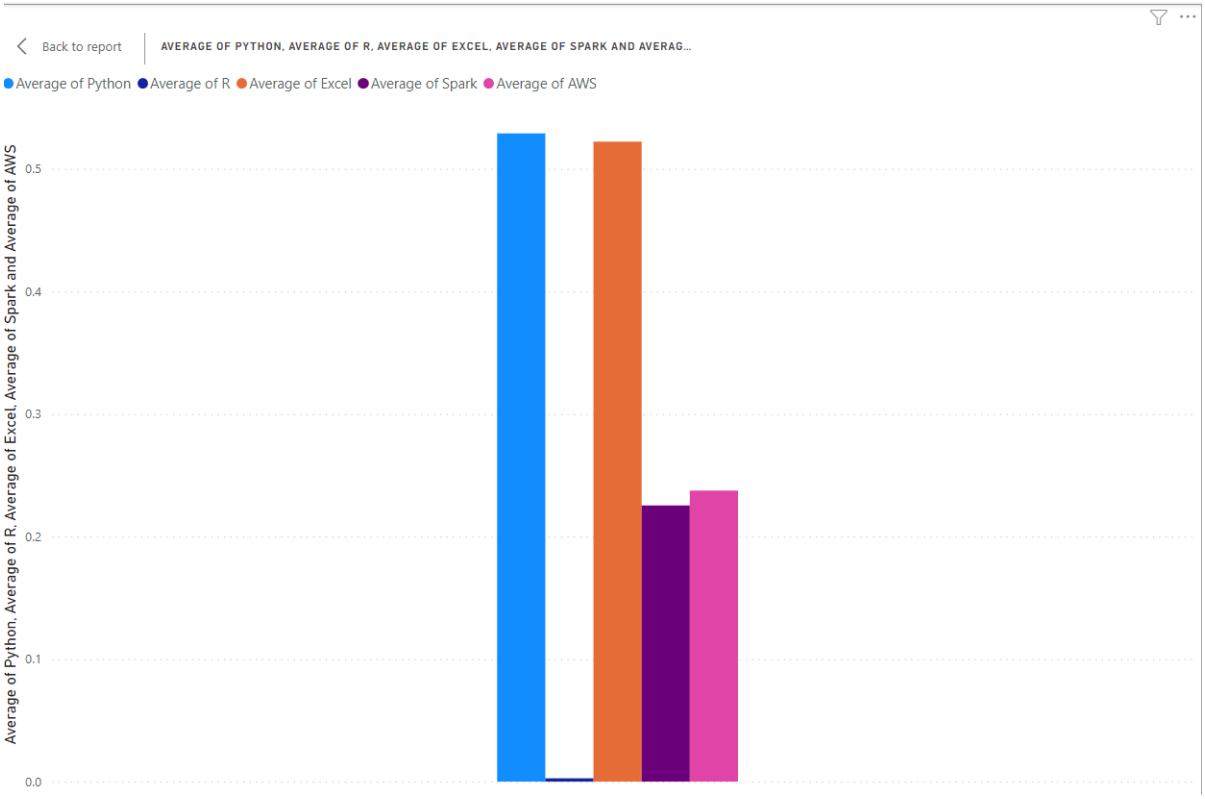

Back to report
AVERAGE SALARY BY JOB TITLE
300K
250K
200K
150K
100K
50K
0K
Average Salary
Director II, Data Science - GRM Actuarial
Principal Machine Learning Scientist
Principal Data Scientist with over 10 years experience
Lead Data Engineer
Director II, Data Science - GRS Predictive Analytics
Staff Machine Learning Engineer
Director, Data Science
Principal Data Scientist
Sr. Scientist II
Senior Quantitative Analyst
Job Title

Let's see the distribution of the companies' ages.



Looks like mostly new/middle-aged companies.

Now let's see each skill and how many applicants have this skill.



It seems like Python and Excel are viral, AWS and Spark are common, and R is Dying.

That is it for getting to know the data let the actual analysis begin.

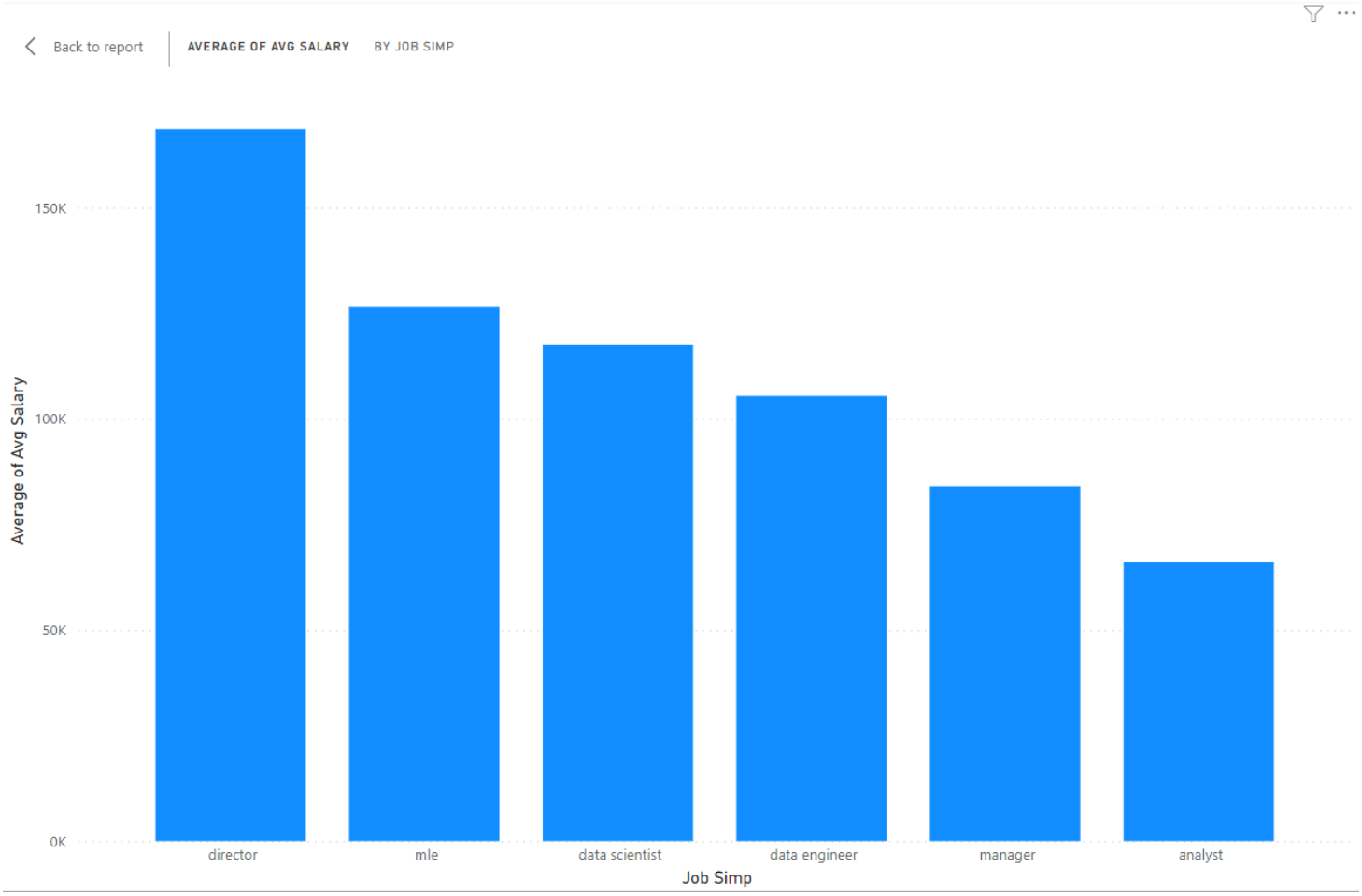
Getting Insights Analyzing the Dataset

My story will be what factors affect the expected salary of an applicant applying for a specific job, on my way I will answer all the previously raised questions.

Let's start with the basic simple question:

> What is the average salary for each job sector (Data Scientist – Data Engineer – Analyst – Manager – MLE - Director).

We will also put a filter on that, to be able to analyze each title separately.

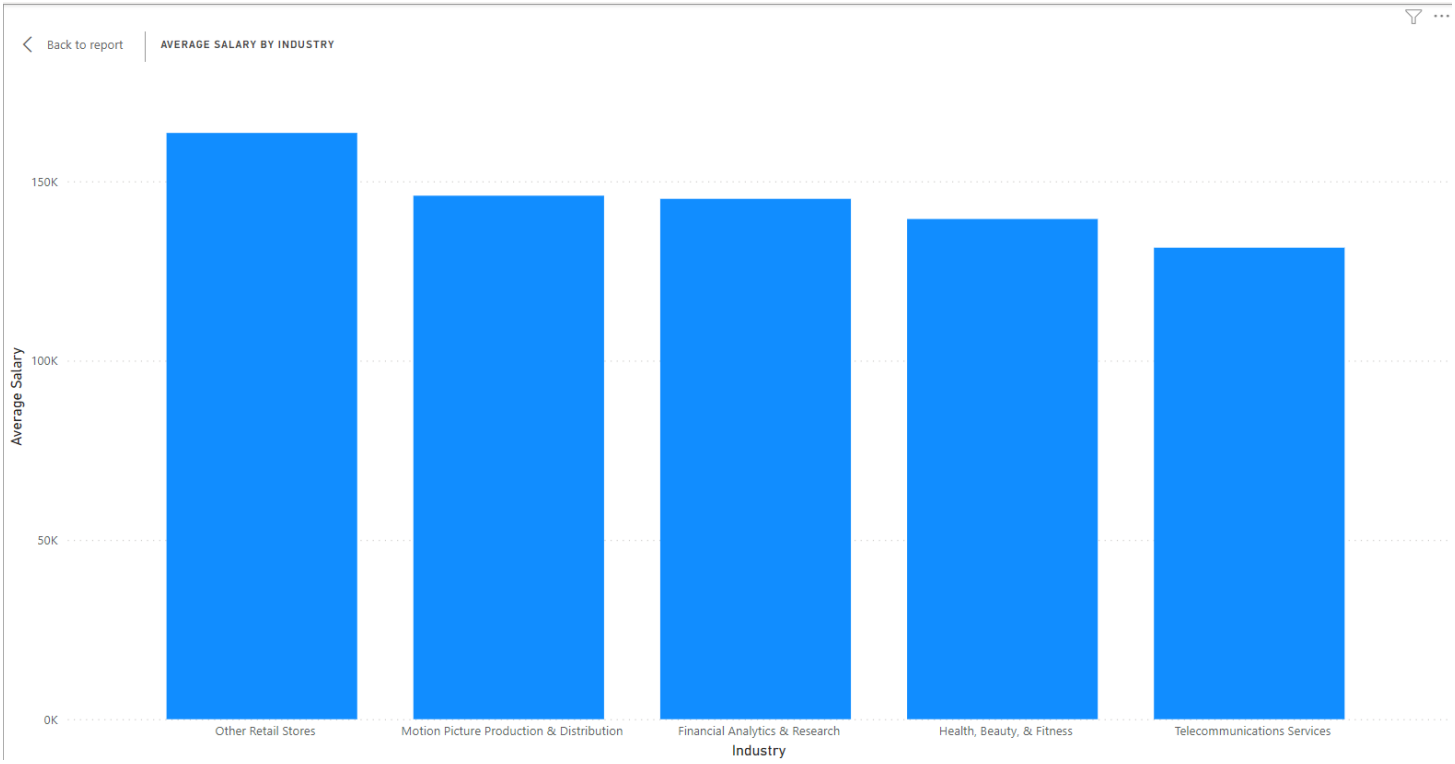


IT IS CLEAR The salary range depends on the job title, as we can see from the chart directors seem to have it all, analysts (We) are not in a good place.

As data specialists, we are not only considered with the technologies, but the business domain also and the level of knowledge we have on this domain, plays a vital role, Am I right?

> What is the average salary for a data specialist working in a specific business domain (Pharmaceuticals – Insurance – Health Care – Sports - Consulting – Energy - etc.).

Doing that for 60 industries makes no sense let's make it for the top 5 ones.

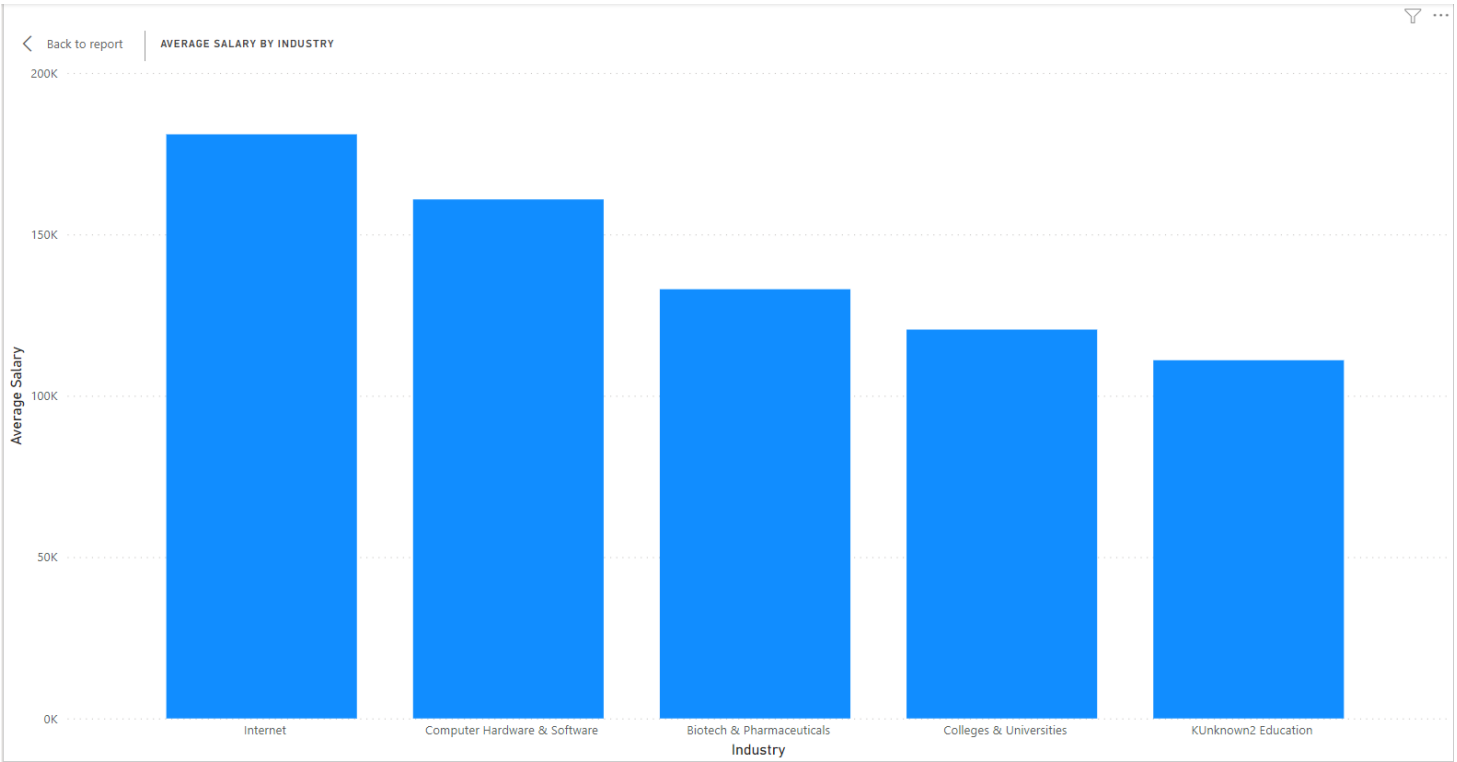


SOUNDS LIKE The connection between the average salary for an employee and the business domain he/she is working in **is not strong**, which is weird for me.

As we saw from the graph the average salary for a data specialist is higher for those who are working for a Retail Store but close for other business domains.

This cannot be right, let's investigate more.

Let's have the same look but only for Machine Learning Engineers.



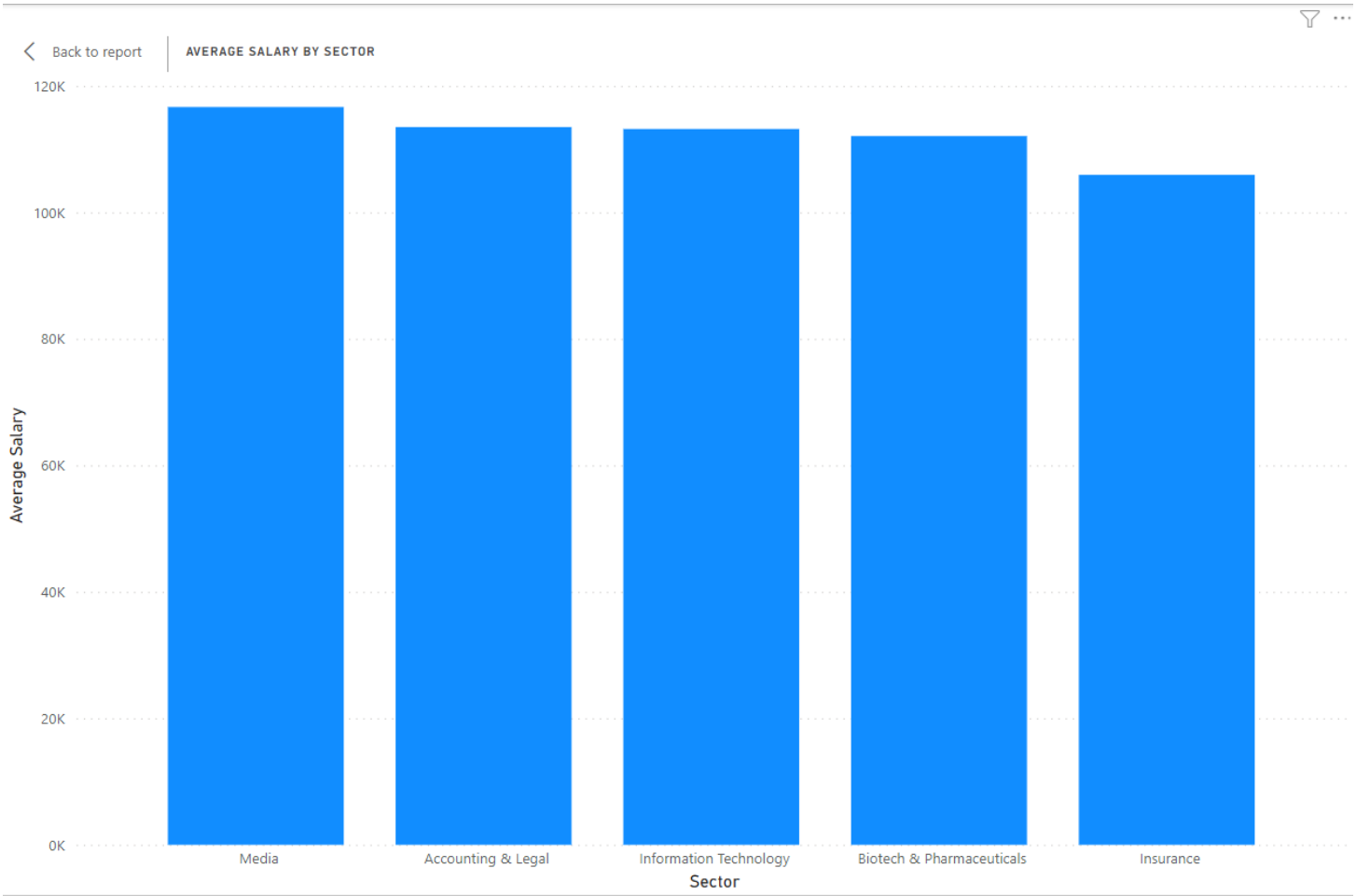
NOW IT IS CLEAR The average salary for an employee **depends** on the business domain he/she is working in.

As we saw from the graph, ML Engineers' salaries are higher for those who are working for an **Internet company**.

Let's go deeper and see if working in a specific sector inside the company affects the average salary.

In another word we want to see if a data specialist working with the media team is the same as another one working with the finance/media team.

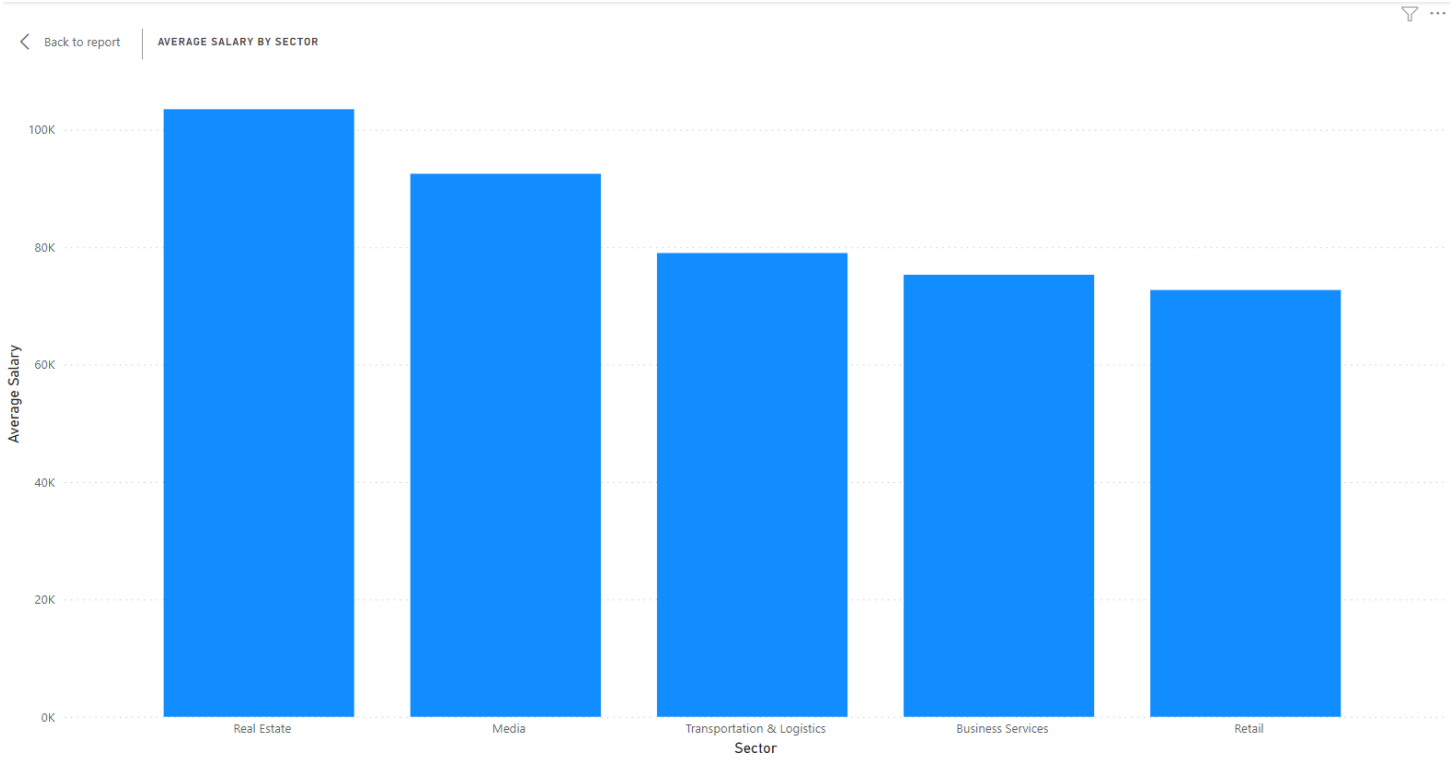
Doing that for 25 sectors makes no sense let's make it for the top 5 domains.



SOUNDS LIKE there is no connection between the salary and the team, as the average salary across each is close, to be honest I am not convinced, let's dig deeper.

This is weird, let's dig more.

Let's have the same look but only for data analysts.

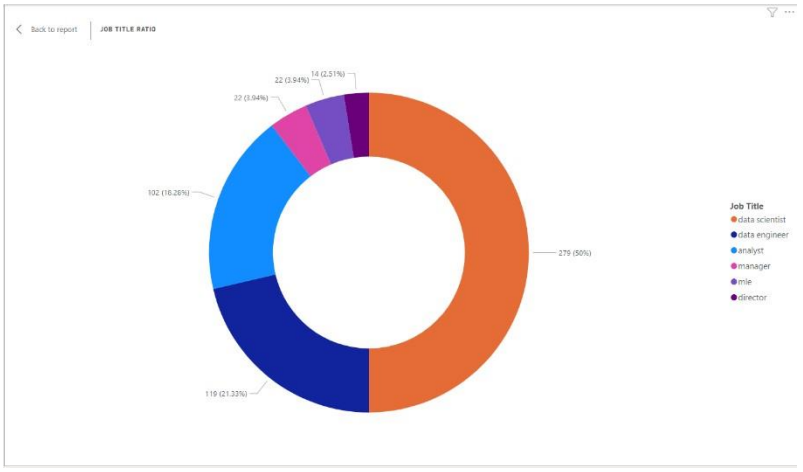


SOUNDS LIKE we found the bug here, as we can see the average salary of an employee **is connected** to the business domain.

I almost fall for this one really, but after deeper investigations, it sounds clear that the average salary **depends strongly** on the sector, but **FOR THE SAME** job title.

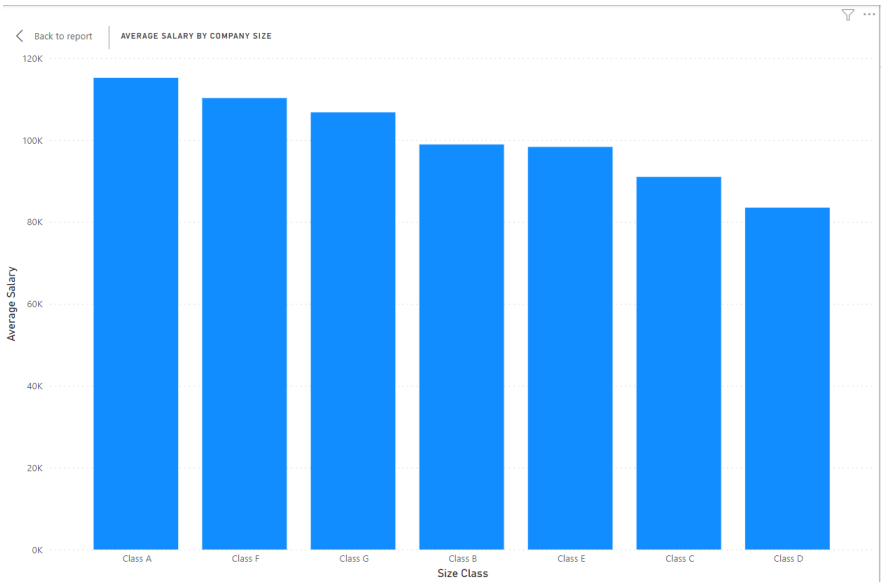
You may compare a data analyst working with the finance team with another one working with the media team, but never compare a data analyst working with the finance team with a data scientist working with the media team, **just a rookie mistake**.

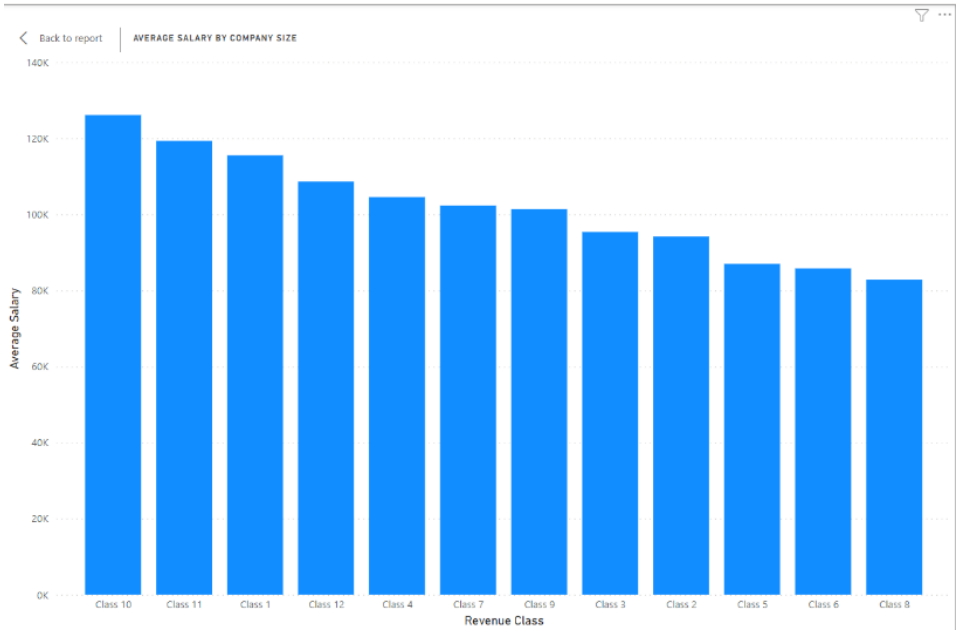
This chart shows the ratio for each job title in the dataset.



We all tend to work for a bigger company, in this dataset we have two definitions for a big company, according to its revenue range or according to number of employees.

Let's visualize each to see what change it makes.



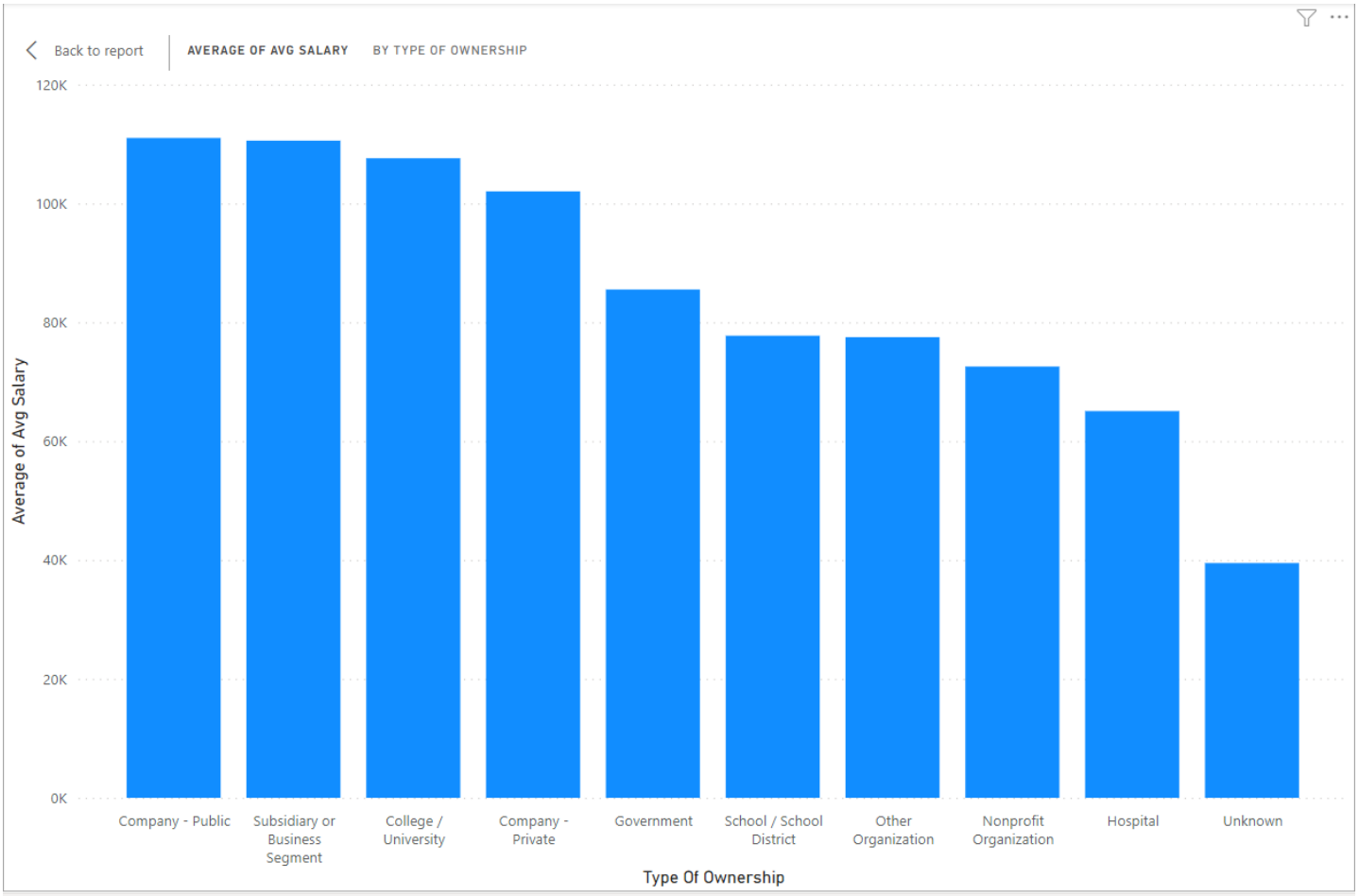


SOUNDS LIKE

There is no connection between the size of the company (according to employees' counting), and the same case according to revenue range.

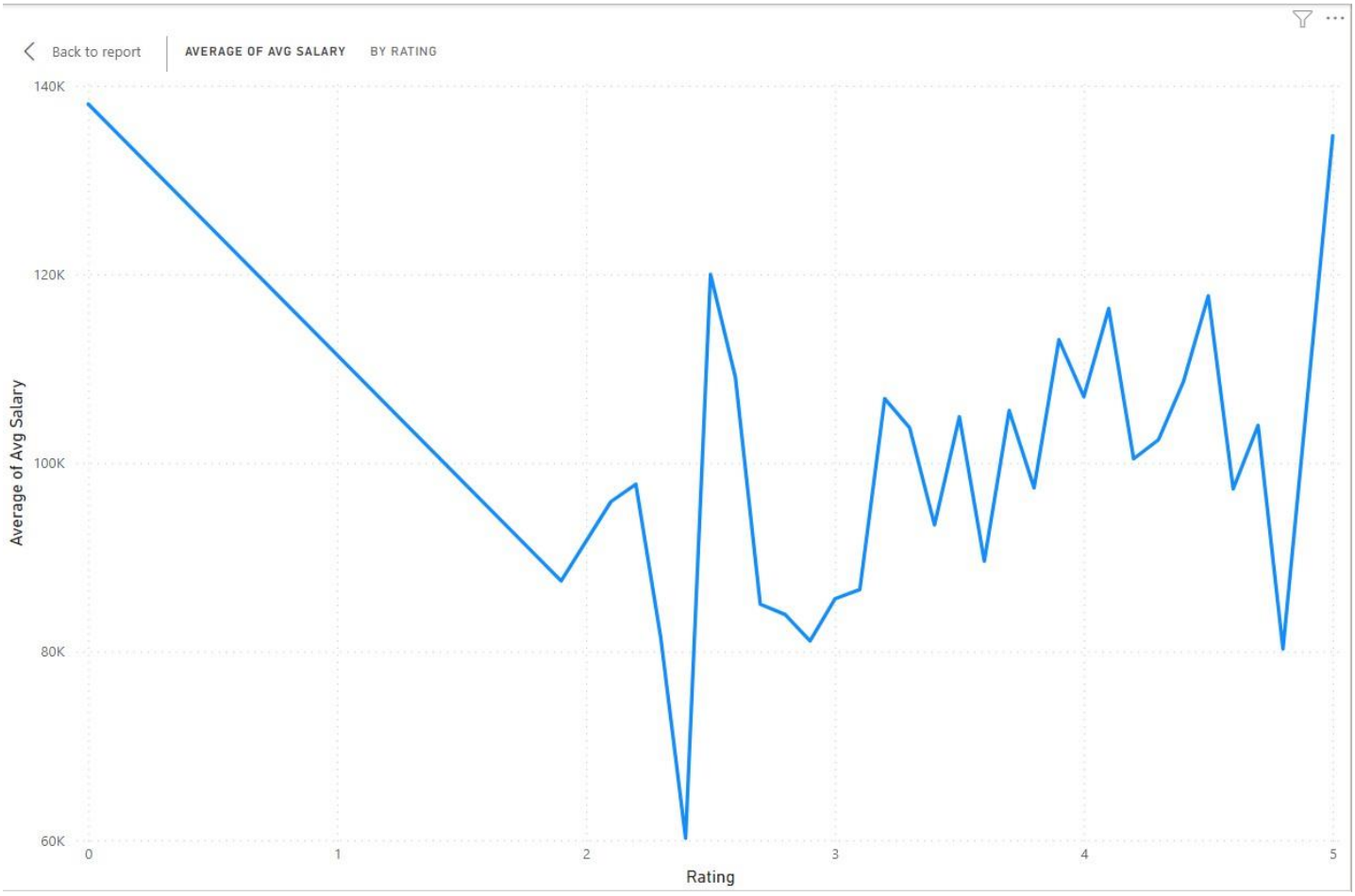
To be completely clear there is a strong connection between the company's revenue and the average salary, but it is not as clear as go on work for a bigger company, the expected salary would go something like: Class 1 has higher salaries & Class 12 the lowest.

Let's see the impact of the ownership type (Private – Public – Non-Profit – etc.) on the salary.



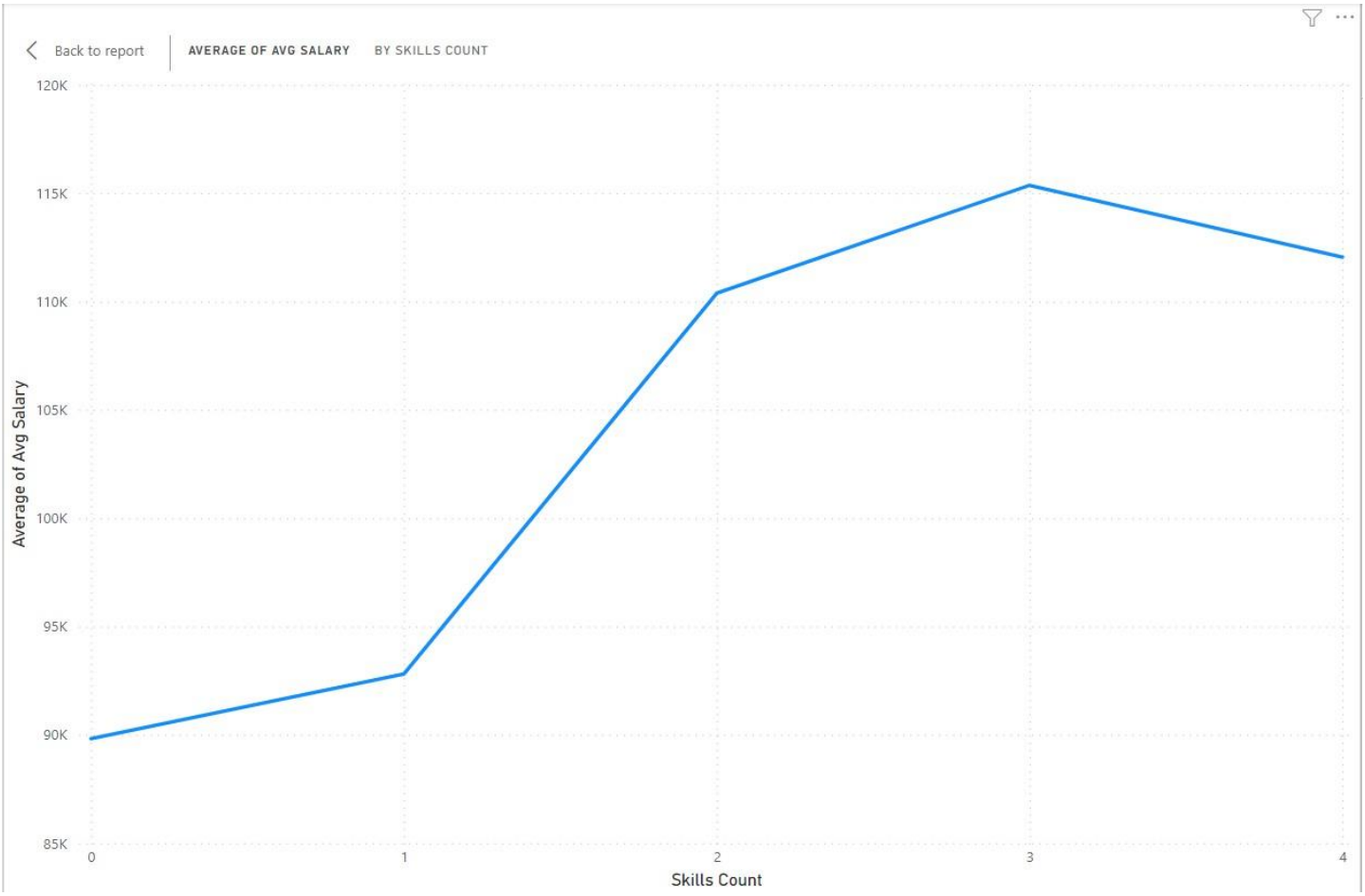
FOR ME I would not say there is a direct relationship, as even the non-profit companies are not far.

Do higher rated companies have higher salaries?



NOPE the rating of the company has no impact on the salary.

How many skills should I have to land my dream job?

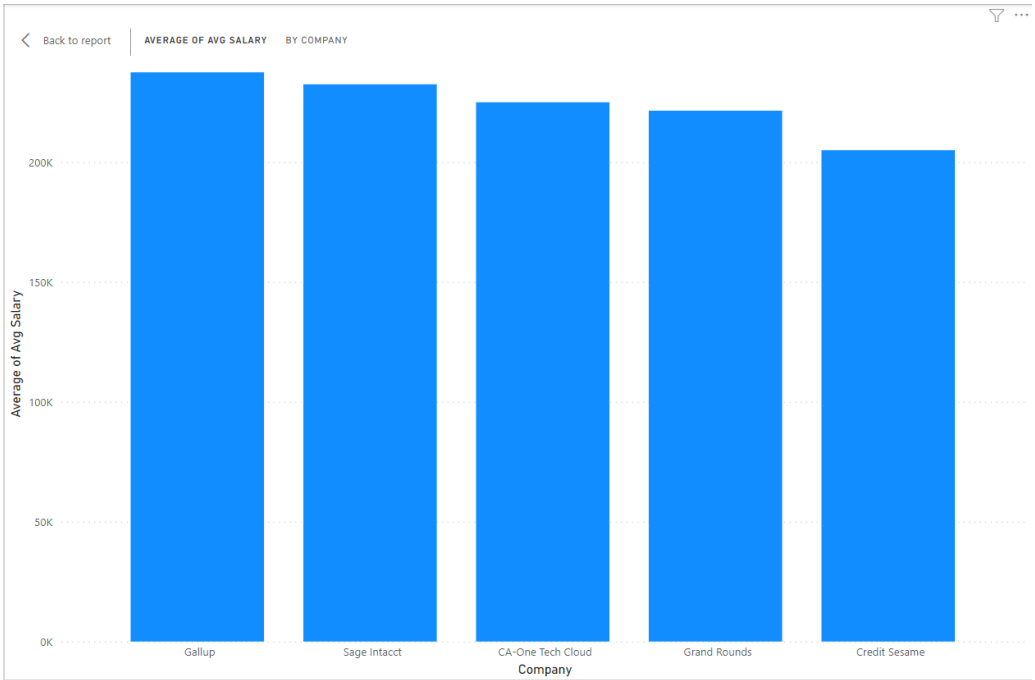


I WOULD SAY the more the better.

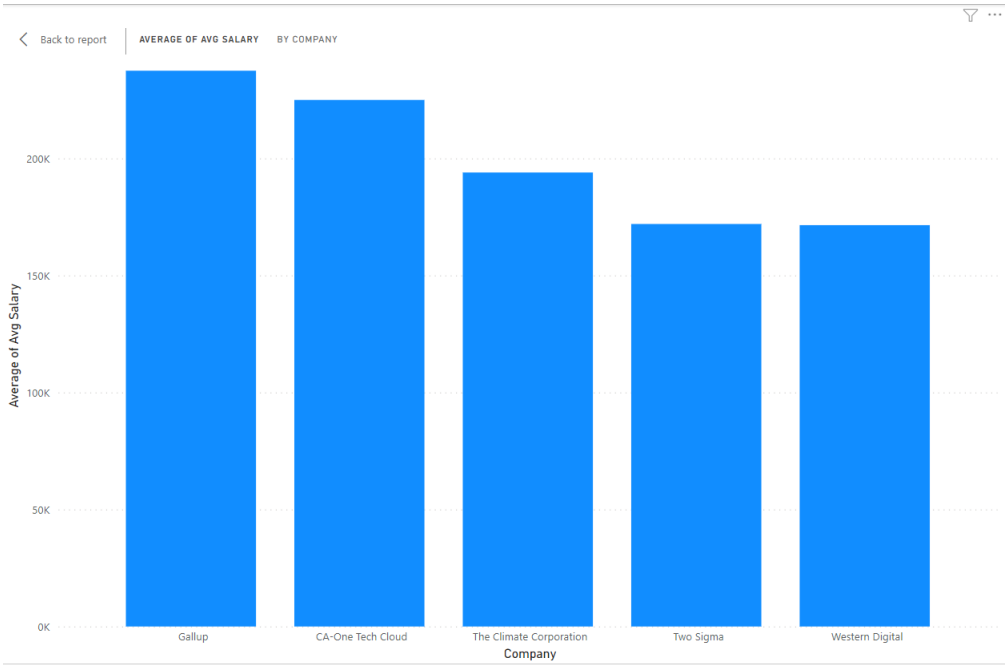
I have added a **Drill** page to the dashboard connected with the skills count, to show you the having-ratio for each skill (how many employees has each skill as a ratio).

That was some basic analysis grouped in the **Basic Analysis** page in the dashboard, whereas the **Explore** page has all the visualizations we need to understand the data, let's do some more analysis in a new page called **More Analysis**.

Average salary for each company.



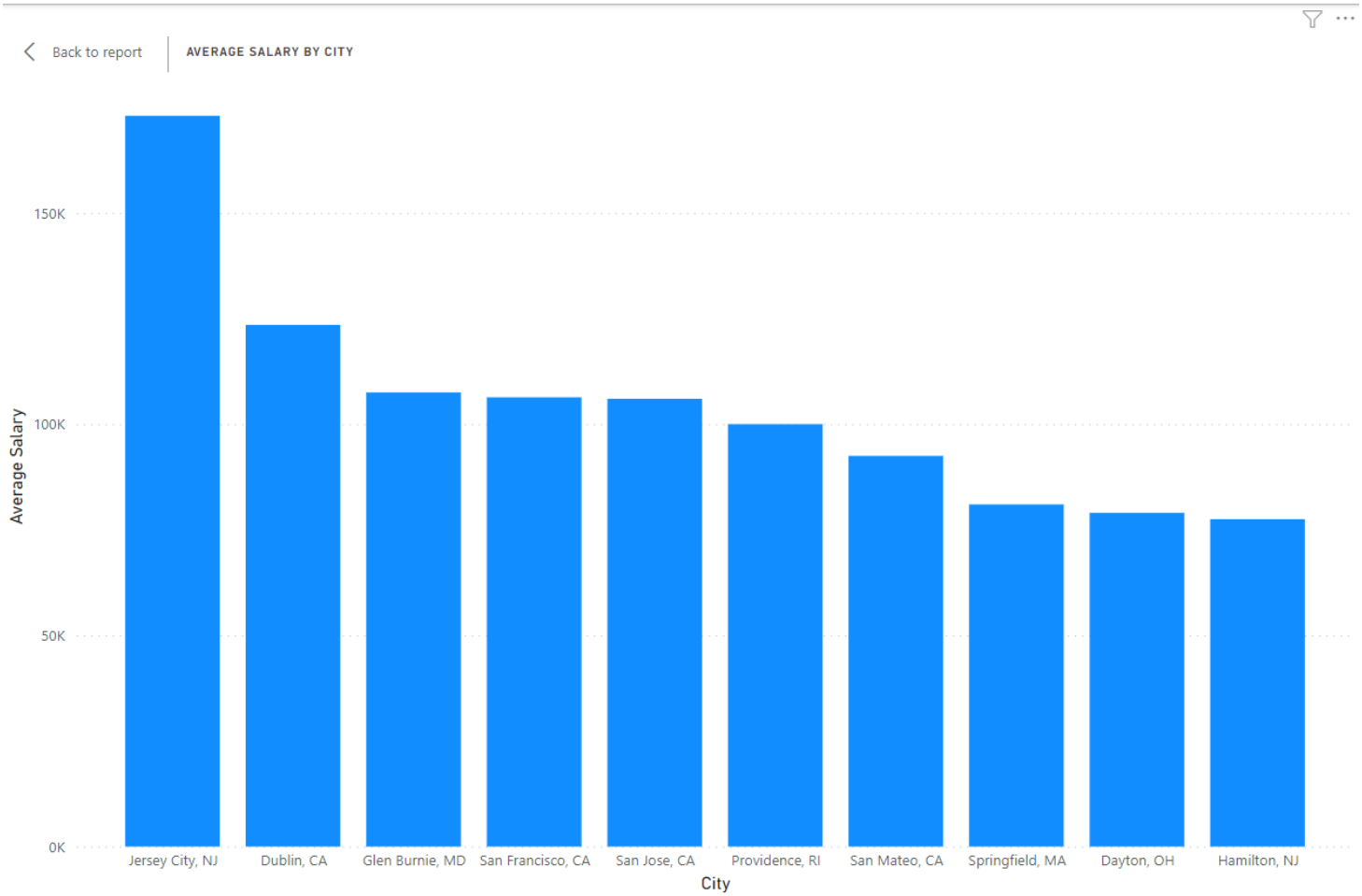
SOUNDS CLOSE let's work for a specific job title (Data Scientists).



THE CONNECTION has started to present itself, and it is clear now that if you are looking for a job you have got to pick the company you are applying for carefully.

I am looking for a job.

Does it make any difference where the company I am applying for is located?



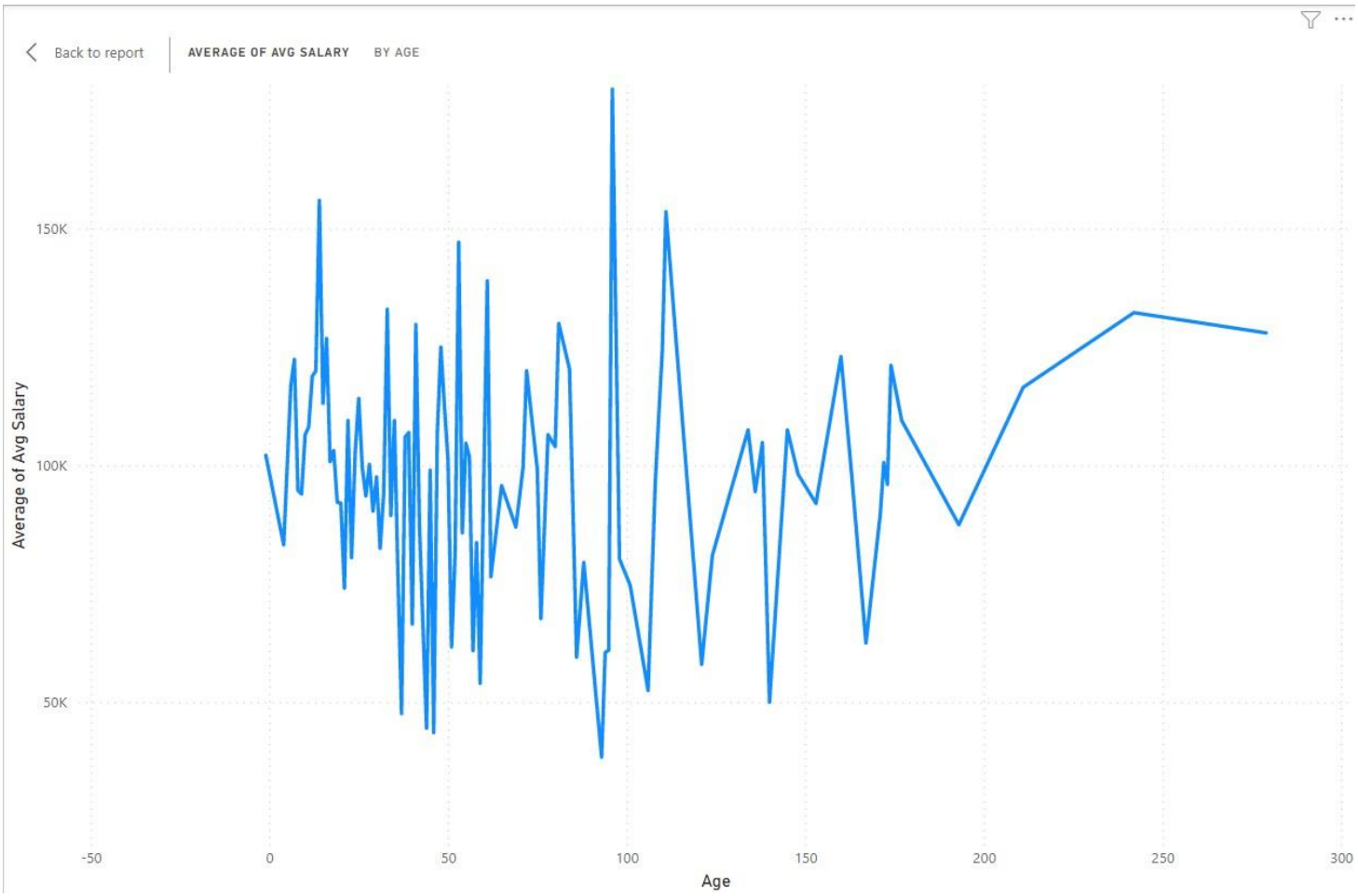
SOUNDS LIKE the city where the company is located **has a direct impact** on the average salary for a specific job title.

DEEP DIVE Into the Dataset

All the analysis we have done until this moment is a layer one, simple analysis, let's deep dive and do some complex analysis and see some hidden facts.

Does the age of the company affect the salary range?

Does it make a difference if I am applying for an old company?

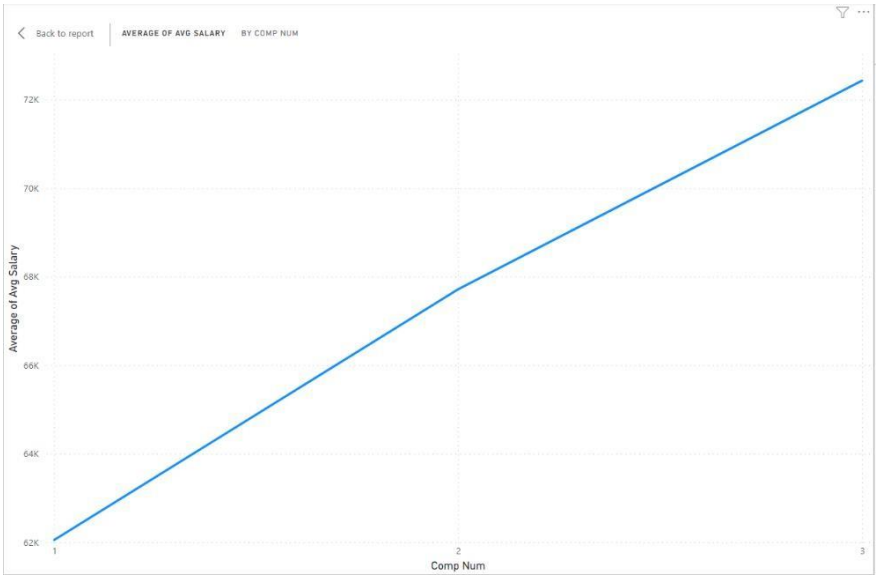


SOUNDS LIKE the foundation date of the company have no impact over the salary range.

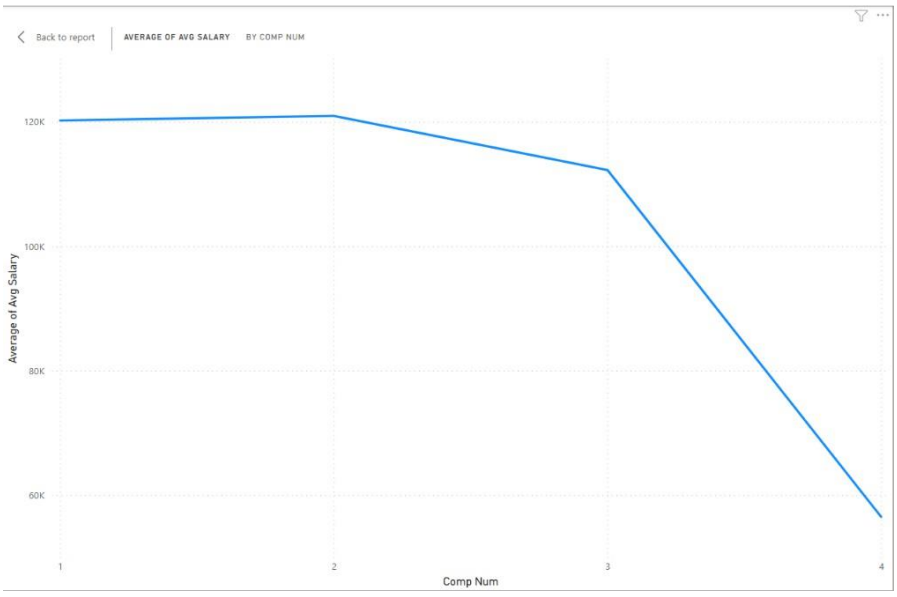
Now it is time to look at one of the most interesting factors to look at, I really was waiting to get to this moment, does the competition the company has affect the salary range for each employee?

If the company is the only provider for the service/product it is offering, does this make the company offer lower salaries?

For Data Analysts



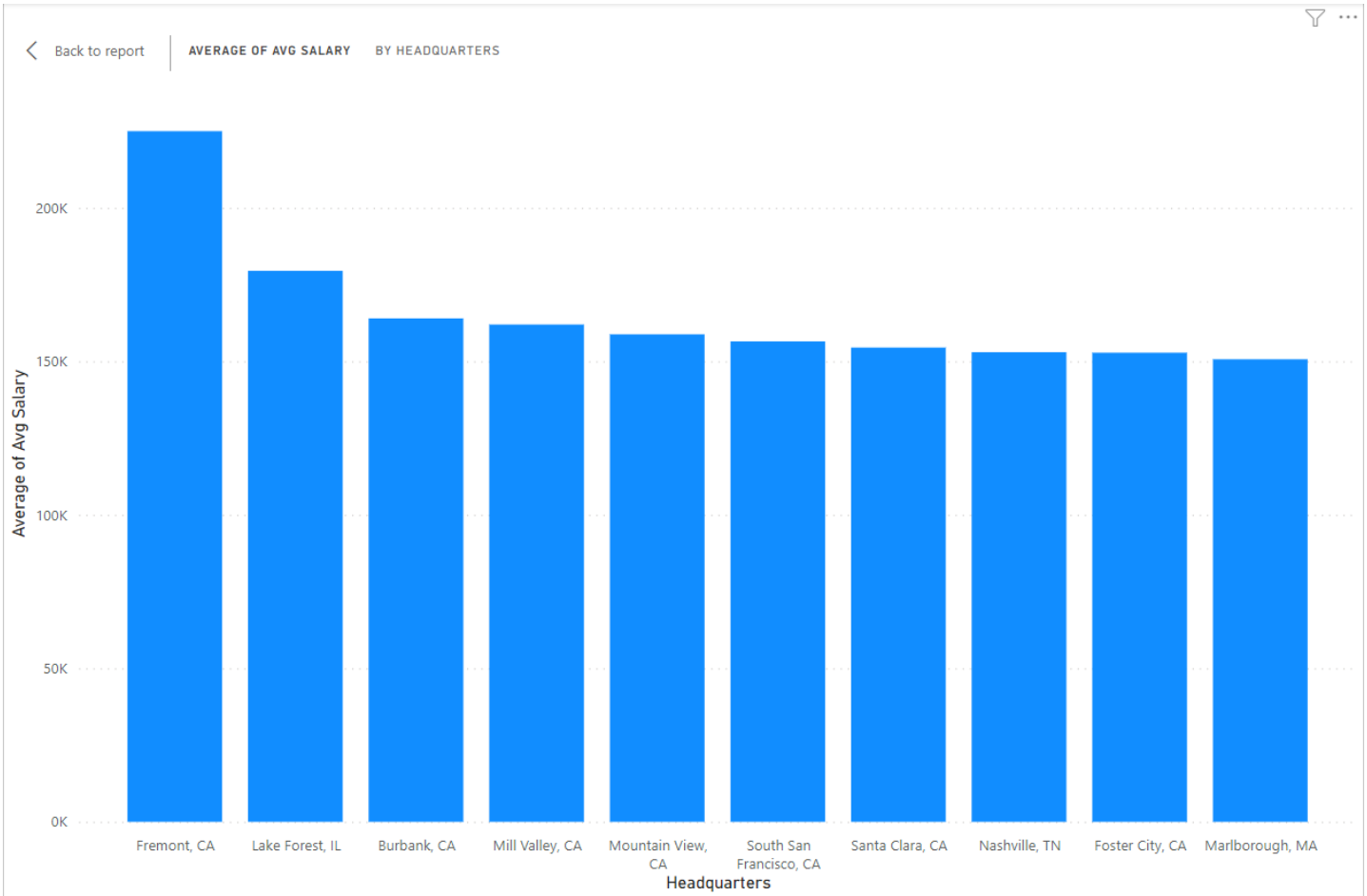
For Data Scientists



SOUNDS LIKE there is a relationship between the number of competitors and the salary range, but the direction of the relationship varies over

different jobtitles, which needs further analysis.

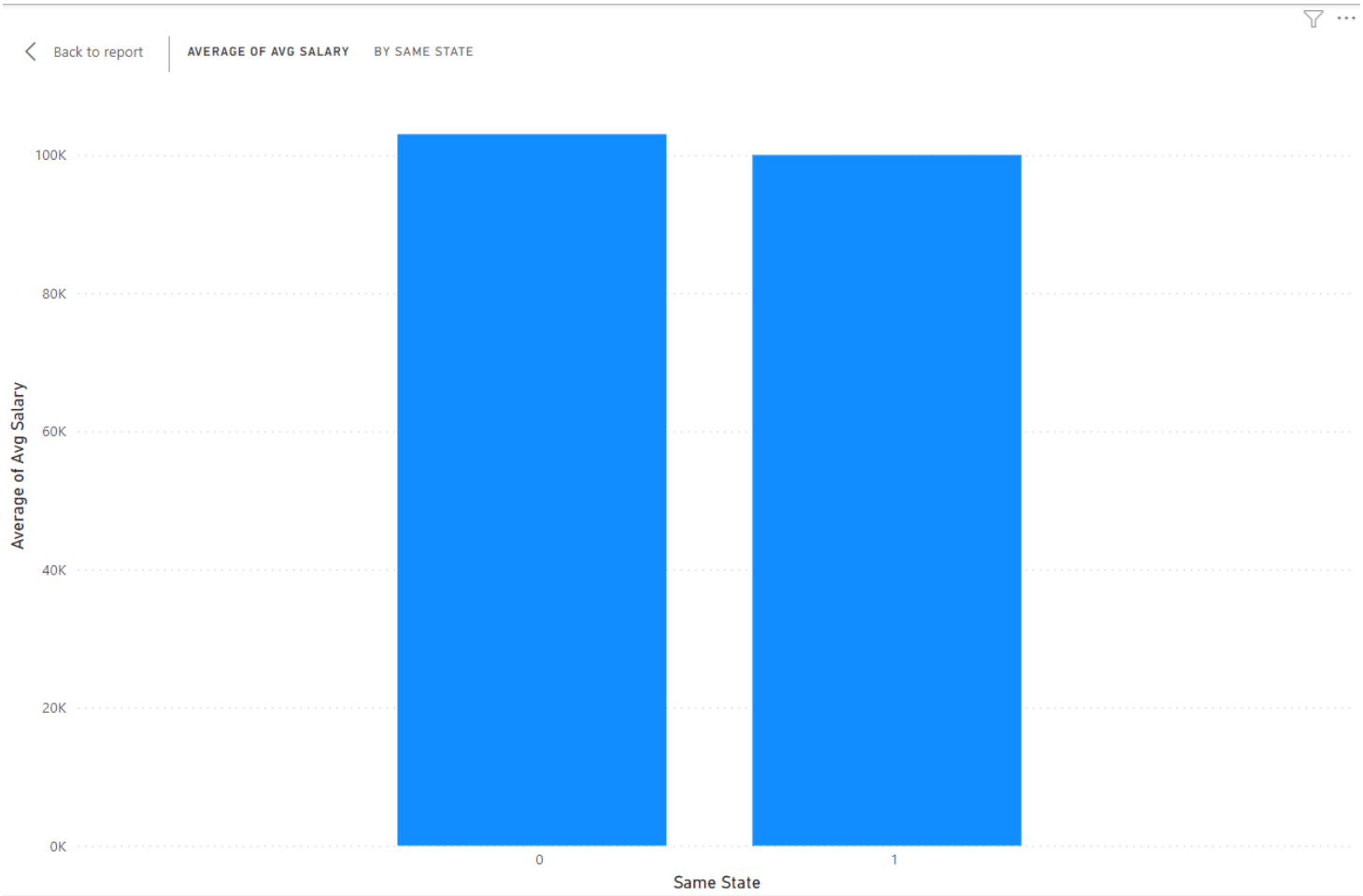
As a Software Engineer, I have always looked at Silicon Valley companies with respect, actually I have always looked at California with respect, but does it really make a difference where the company's headquarters are located?



SOUNDS LIKE the region where the headquarters are located **has no impact** on the average salary for any job title.

Does the company offer higher salaries for people from another state?

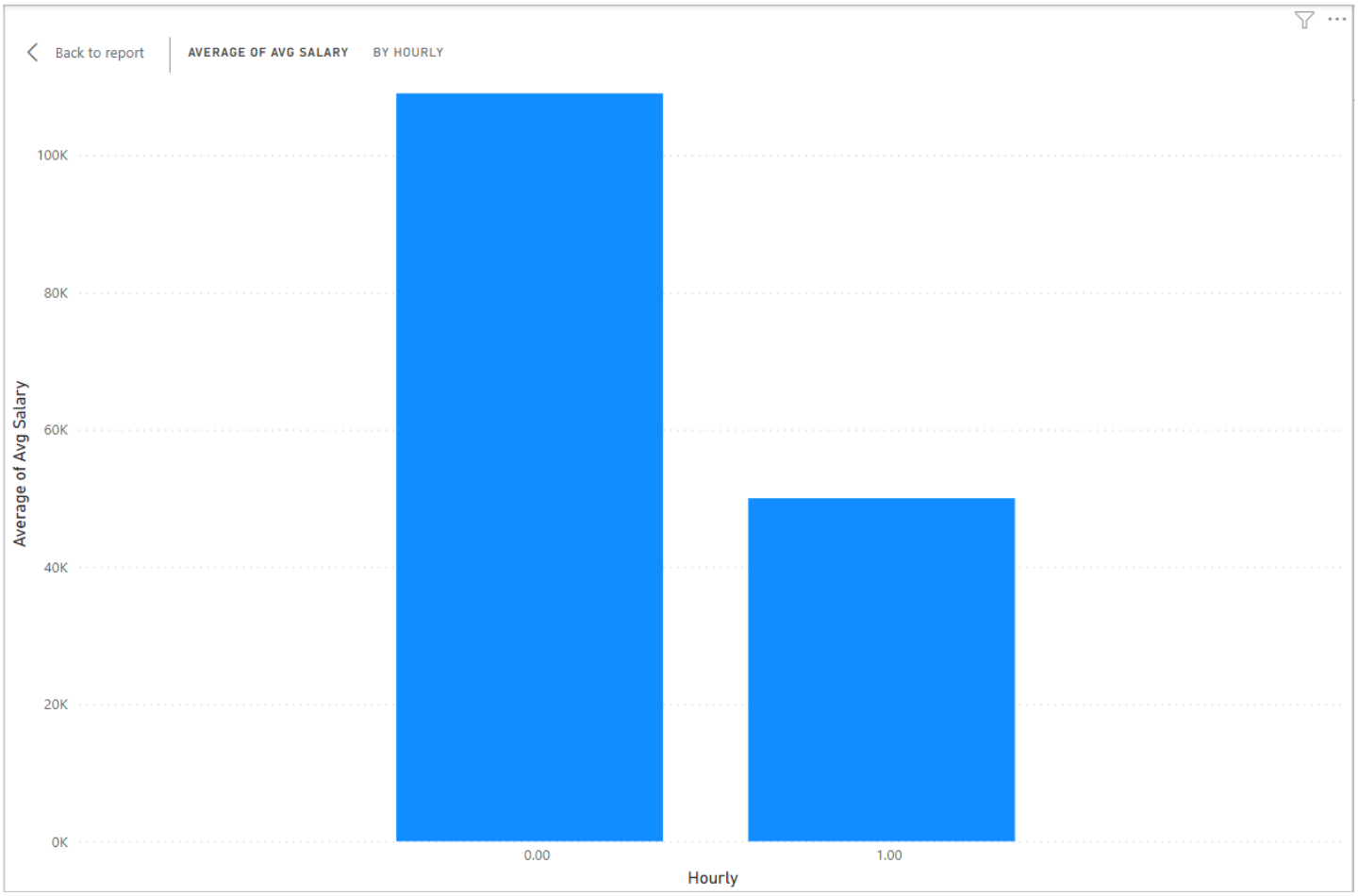
If I got a job offer from a company located in another state, I would expect more money to cover rent, travel, and being away from my family, let's investigate if this is true.



SOUNDS LIKE there is **no any connection** between the salary range and whether the employee is from the same state.

In the dataset we have people working per hour.

Do those people get higher/lower salaries than the others?



SOUNDS LIKE there is **a strong connection** between the salary range and whether the employee is working per hour.

I will give you a single hint, don't work per hour.

DISCLAIMER we only have 24 employees working per hour in the dataset.

22 Of them have missing values for the job title.

2 of them are data analysts.

so, this insight **is not** trustworthy.

COMPLEX ANALYSIS Deeper Dive into the data

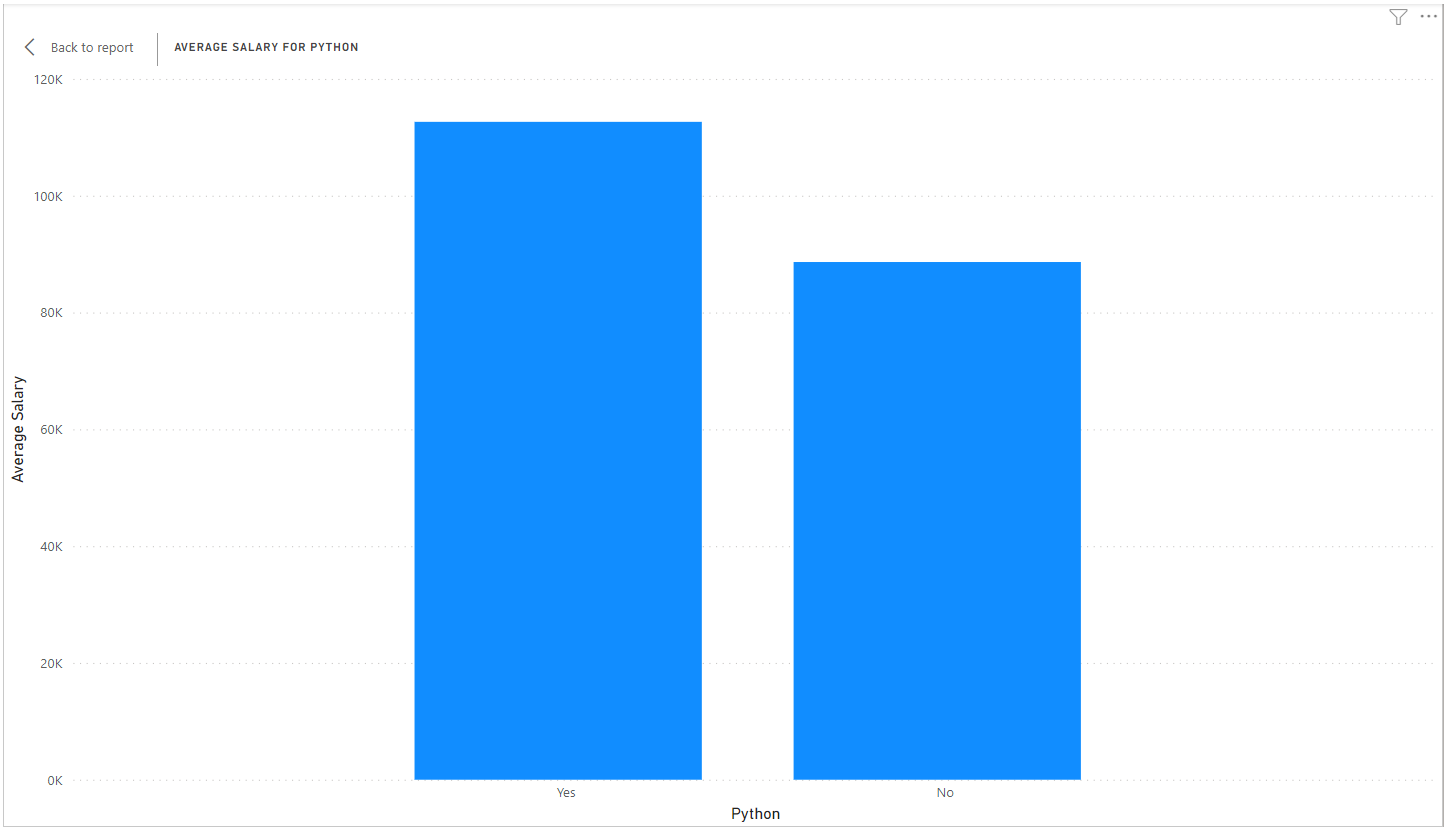
All the pervious analysis is considered a basic analysis, let's deep dive into the data, do some data mining, and extract some really hidden insights.

Here is a question?

I am eager to become a data specialist, I have the basic knowledge, but Iwant to enhance my skill set by learning one of those skills (Python – R – Excel – AWS – Spark), which one should I start with.

That is a fairly simple question, but needs a fairly complex data analysis to extract, so let's start.

The average salary for people who have/don't have Python skills.



It appears from the visuals that each skill has its own impact on each job title, but it is obvious that having Python skills strongly affects your data career.

The visuals are designed for you to be able to understand exactly the effect of having each skill on you, just play around.

In addition to the job title filter, I added a filter on skills to show you what happens if you have or don't have a specific skill to your career.

I was just doing some warmups before I dig into the actual analysis, but when I got to assess the analysis so far to start the real work, I realized there are no further features to explore in the data.

We have already looked at every little column and aspect of the dataset.

I am really frustrated by the size, capacity, and quality of the data, but that is what we have for now.

So, I will keep it up to this for now, but I will do some further work collecting some more data, doing some web scraping from web sites like Glassdoor, LinkedIn, Wuzzuf, Indeed, and so on, downloading data from Kaggle, doing some data integration, cleaning the collected data, and merge it with the dataset we already have, preserving the same data structure , just converting the data into some 500k row or something, then I will update the project.

I think with a higher quality data collection, cleaning, and modeling the same analysis we have already done will give us some great analysis, and we will add room for some extra features to analyze.

Finally, thanks for reading up to page 27, I hope you had some fun and got some insights.
