



Prajwal Patil

Power BI Project

Data-Related Job Market Analysis

Business Problem

The business problem focuses on understanding salary trends in data-related jobs, specifically for job titles like Data Scientist, Data Engineer, and Data Analyst. The key objective is to identify what factors influence the salaries in these positions across different companies.

The analysis aims to uncover:

- The impact of specific skills (e.g., Python, R, Spark, AWS, Excel) on salary.
- How job titles, company size, and industry sectors correlate with salary ranges.
- The influence of company ratings, age, and ownership type on salary.
- Location-specific salary trends.

The goal is to provide actionable insights for job seekers and companies by understanding the salary landscape across different variables such as company revenue, size, and industry type.

DATA REQUIREMENT

The data required for this analysis is sourced from two files:

1. **Glass Door Jobs Data:** Contains detailed information on job postings such as:
 - Job Title, Salary Range (Min, Max, Avg)
 - Company details (Name, Size, Industry, Revenue, Rating)
 - Skills required (Python, R, AWS, Excel, etc.)
 - Company's location and size classification
 - Age of the company and competitors
2. **Eda Data:** Provides additional insights and useful features, like the same state flag (whether the employee lives in the same city as the company), skills count, and more.

DATA COLLECTION AND DATA UNDERSTANDING

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 VALIDATION

Data validation plays a crucial role in ensuring the accuracy and reliability of the dataset used for analyzing salary trends in data-related job roles. In the context of this report, data validation ensures that the data we work with is consistent, complete, and properly formatted, which is essential for drawing meaningful insights and making informed decisions.

Here's how data validation is tailored for this report:

1. Consistency Across Columns:

- The dataset includes columns with various types of data, such as salary ranges, company details, job titles, and skills. During the validation process, it is crucial to ensure that the values are consistent across these columns. For example:
 - Salary Ranges should not contain any non-numeric or erroneous entries (e.g., text or symbols).
 - Company Ratings should fall within the accepted range (e.g., between 0 and 5), ensuring no invalid or out-of-range values are included.
 - Categorical columns such as Job Titles, Company Type, and Location need to be validated for spelling consistency, ensuring that variations are removed.

2. Handling Missing Data:

- Missing data can significantly impact the outcome of any analysis. In this dataset, missing values are identified and addressed through careful handling. For example:
 - Salary data may be missing in some instances, requiring imputation or removal depending on the amount of missing data. Rows with missing salary values are either cleaned up with reasonable imputation (e.g., replacing with average salary by job title) or discarded if the data is too sparse.
 - Columns like Company Age, Competitors, and Type of Ownership have missing values represented by placeholders (e.g., "-1" for age or "unknown" for ownership type), which are dealt with during the data cleaning phase.
- The strategy for missing data validation ensures the completeness of the dataset while avoiding skewed or unreliable results.

3. Verification of Range and Format:

- For numerical data, such as salary ranges and company age, data validation ensures that the values fall within logical and acceptable ranges:
 - Salary Ranges (Min, Max, Avg) must fall within realistic figures, ensuring there are no extreme or unreasonable values (e.g., salaries with negative values or abnormally high amounts).
 - Company Age is calculated from the Foundation Year. Any invalid dates (like a future year or negative values) are flagged and corrected.
- The Revenue Class and Company Size Class columns are validated to ensure they fall within the predefined categories based on the company's revenue and employee count.

4. Correct Data Types:

- It is important that each column in the dataset contains the correct data type for accurate analysis. For example:
 - Columns such as Job ID, Salary, and Skills Count must contain numeric values.
 - Company Name, Job Title, and Location should be in text format.
 - Skill Flags (e.g., Python, R, AWS) should be binary, represented by "Yes" or "No" values instead of numeric values or inconsistencies.

5. Categorical Data Consistency:

- The dataset includes several categorical columns, such as Industry, Sector, Location, and Company Ownership Type. These values need to be consistent throughout:
 - The Industry column should be consistent and limited to specific industries, like IT, Pharmaceuticals, and Insurance, to avoid variations or misspellings.
 - The Ownership Type should be categorized correctly as Private, Public, or Non-profit, without inconsistent or unrecognized values.
- Ensuring consistency in categorical data ensures that the analysis can be grouped and aggregated correctly.

6. Outlier Detection:

- Outliers in salary data or other numerical columns are detected and addressed during the validation process. For instance:
 - Salary values that fall outside a reasonable range (such as excessively low or high salaries) are flagged for review.
 - Extreme outliers might indicate data entry errors or unusual

circumstances that could distort the analysis. These are either corrected or excluded based on their impact.

7. Validation of Derived Columns:

- Derived columns, such as Avg Salary (calculated from the min and max salary) and Company Age (calculated from the foundation year), require validation to ensure that their values are accurate:
 - Avg Salary is validated by recalculating the average of the min and max salary values.
 - Company Age is checked to ensure that it is correctly derived from the Foundation Year of each company.

Through a rigorous data validation process, we ensure that the dataset used in this job market analysis is clean, reliable, and ready for meaningful insights. Validated data is essential to draw accurate conclusions regarding factors influencing salaries, such as job title, company size, industry, and required skills. This approach guarantees that the analysis provides trustworthy insights to stakeholders in the job market.

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 $(1920 * \text{Salary})$ 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.

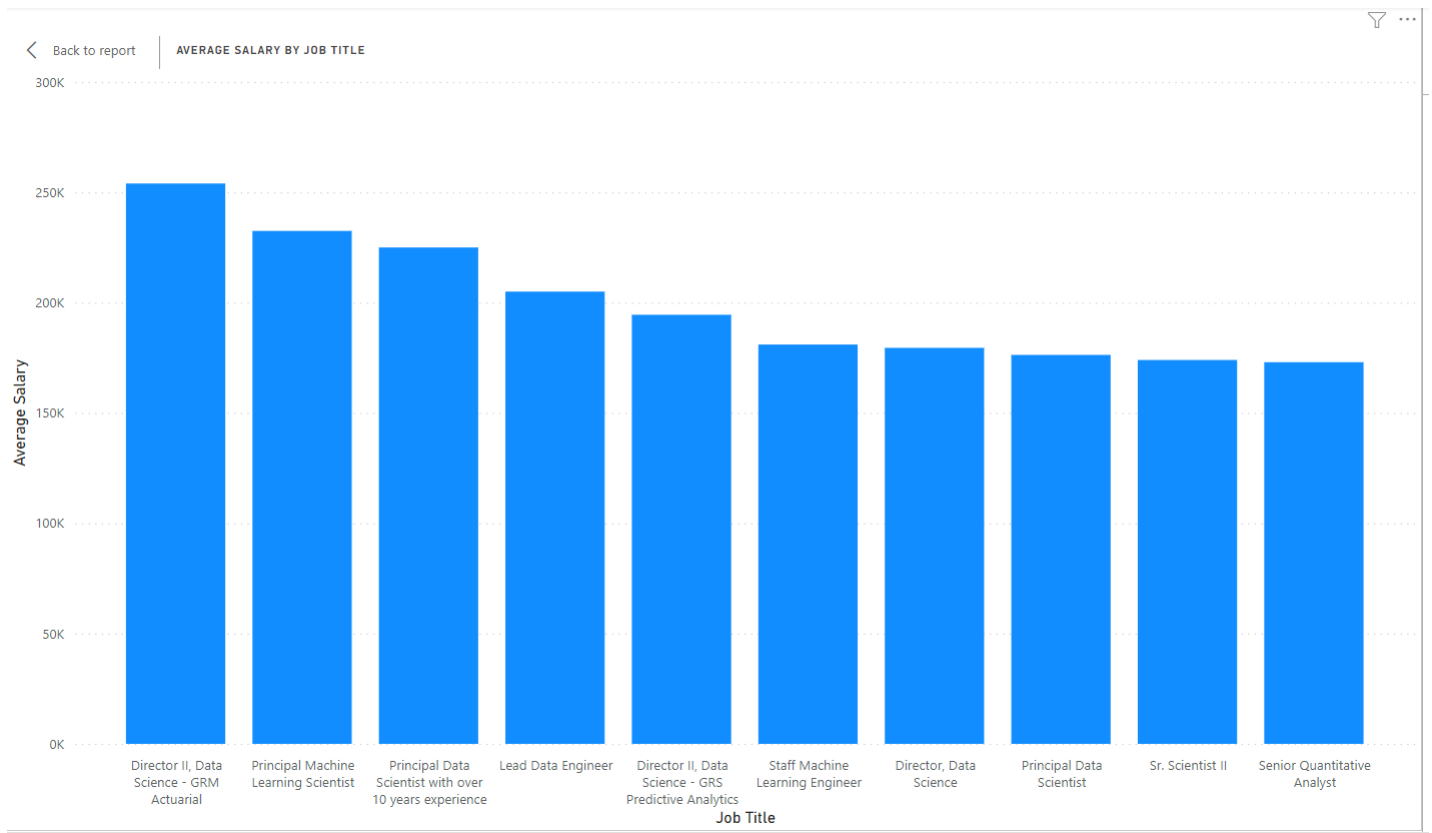
Getting Insights Exploratory DA

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

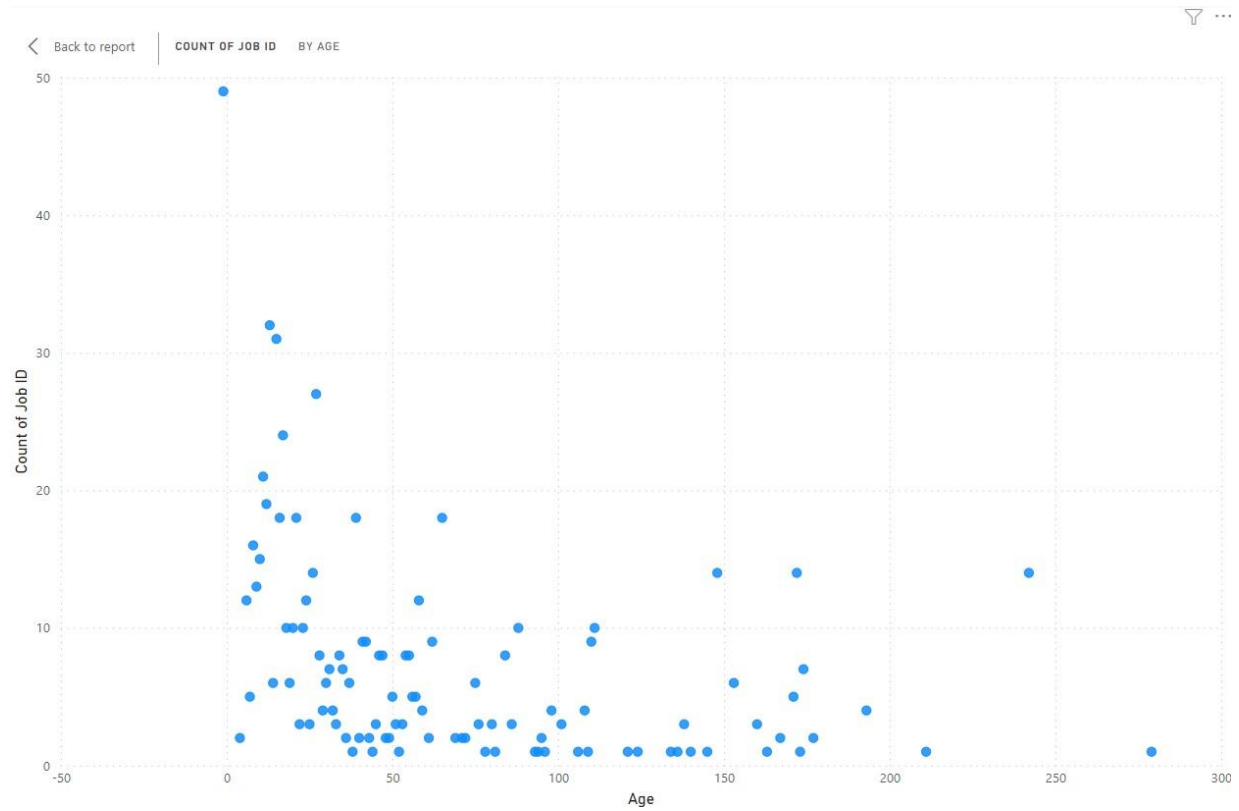
- We have tons of analysis to do, tons of stories to tell, and tons of insights to get, but for me I am mainly interested in the impact of having a specific skill, a combination of multiple skills, or even the skills counting on the average salary.
- In addition to that I want to see the variation of 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

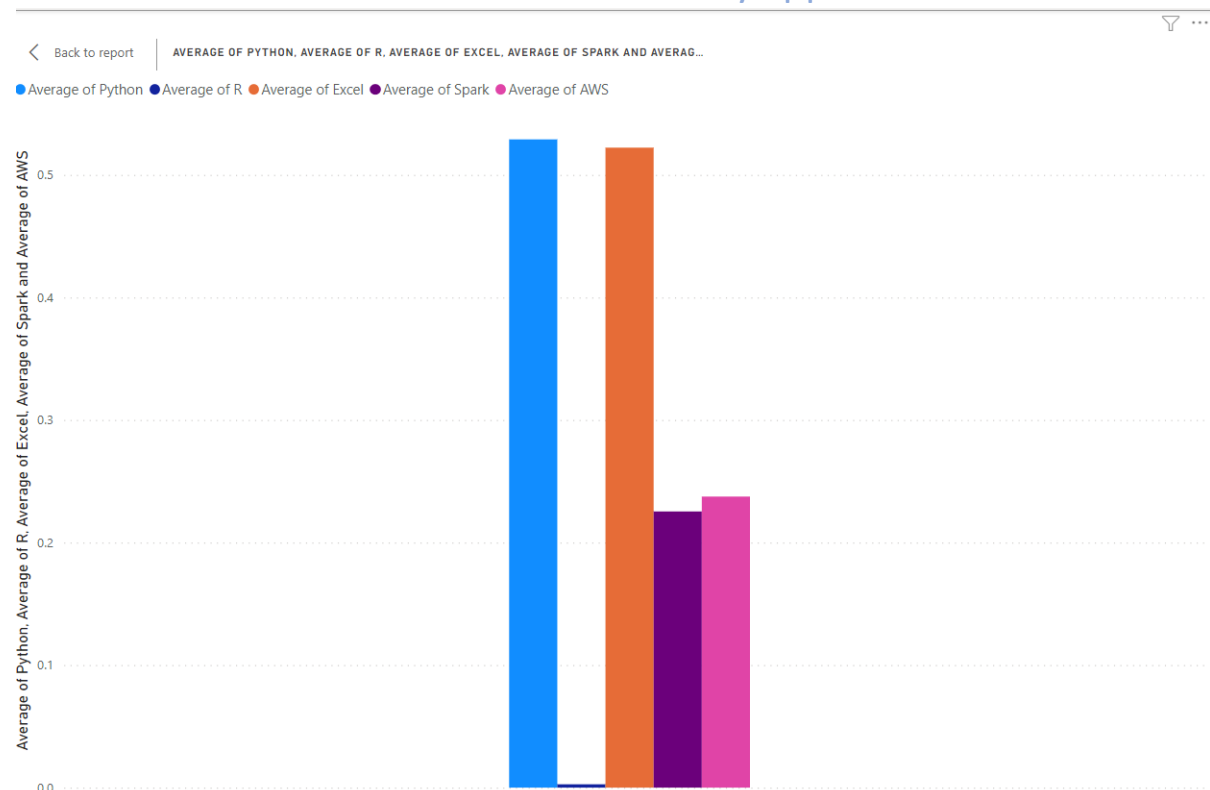


Let's see the distribution of the companies' ages.



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.

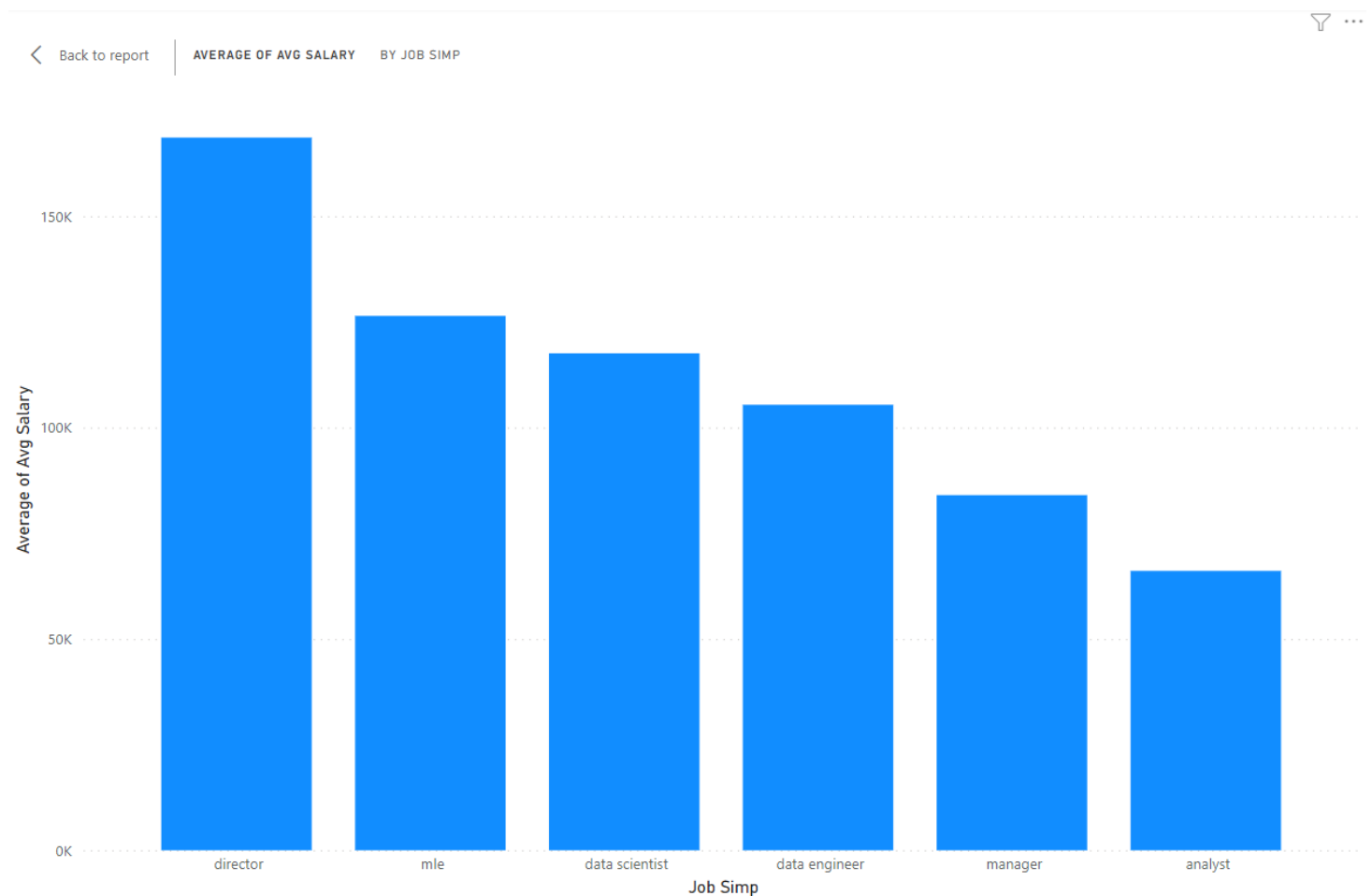
Getting Insights Analyzing the Dataset

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.

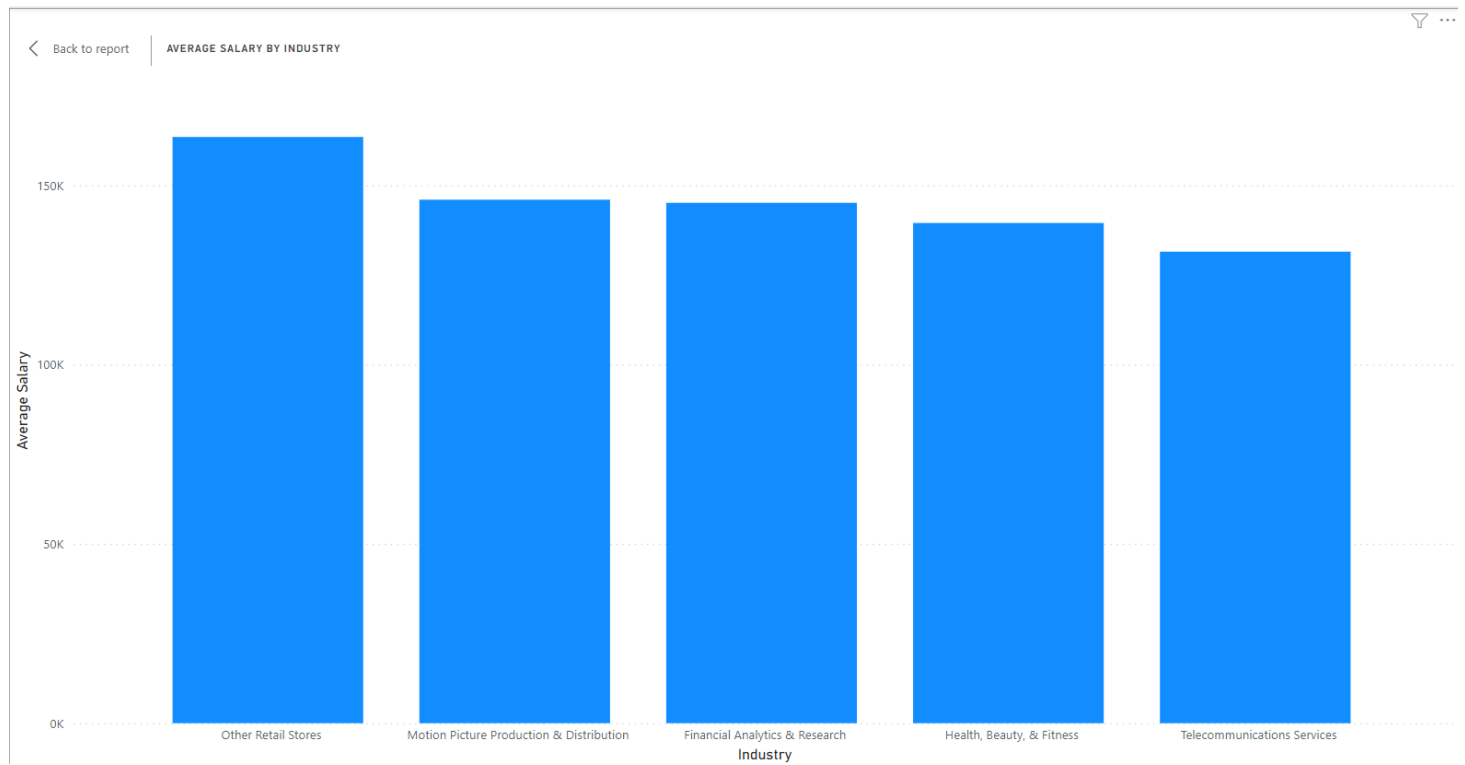


IT IS CLEAR The salary range **depends** on the job title, as we can see from the chart directors seem to have it all, analysts **(We)** are not in a good place.

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.

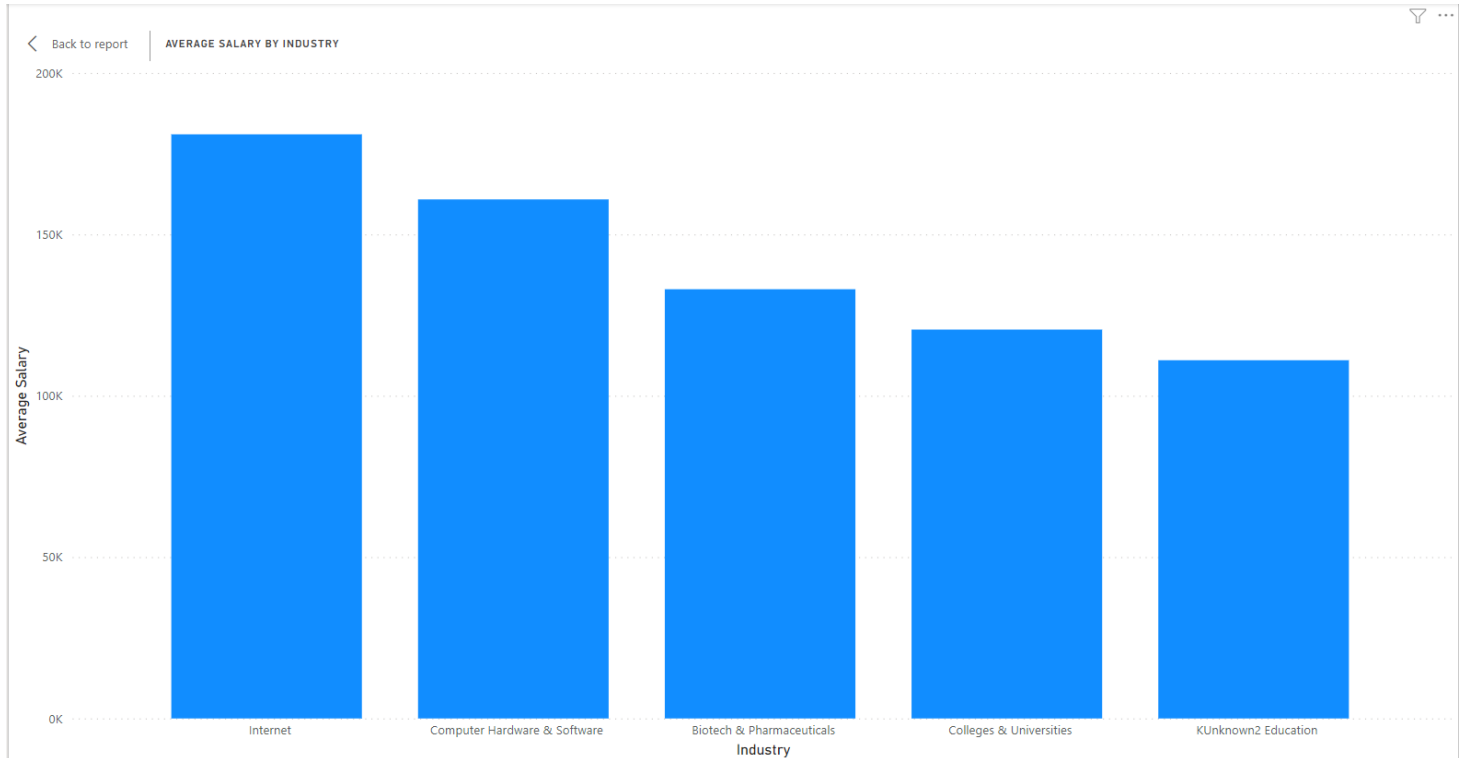


SOUNDS LIKE The connection between the average salary for an employee and the business domain he/she is working in **is not strong**, which is weird for me.

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.



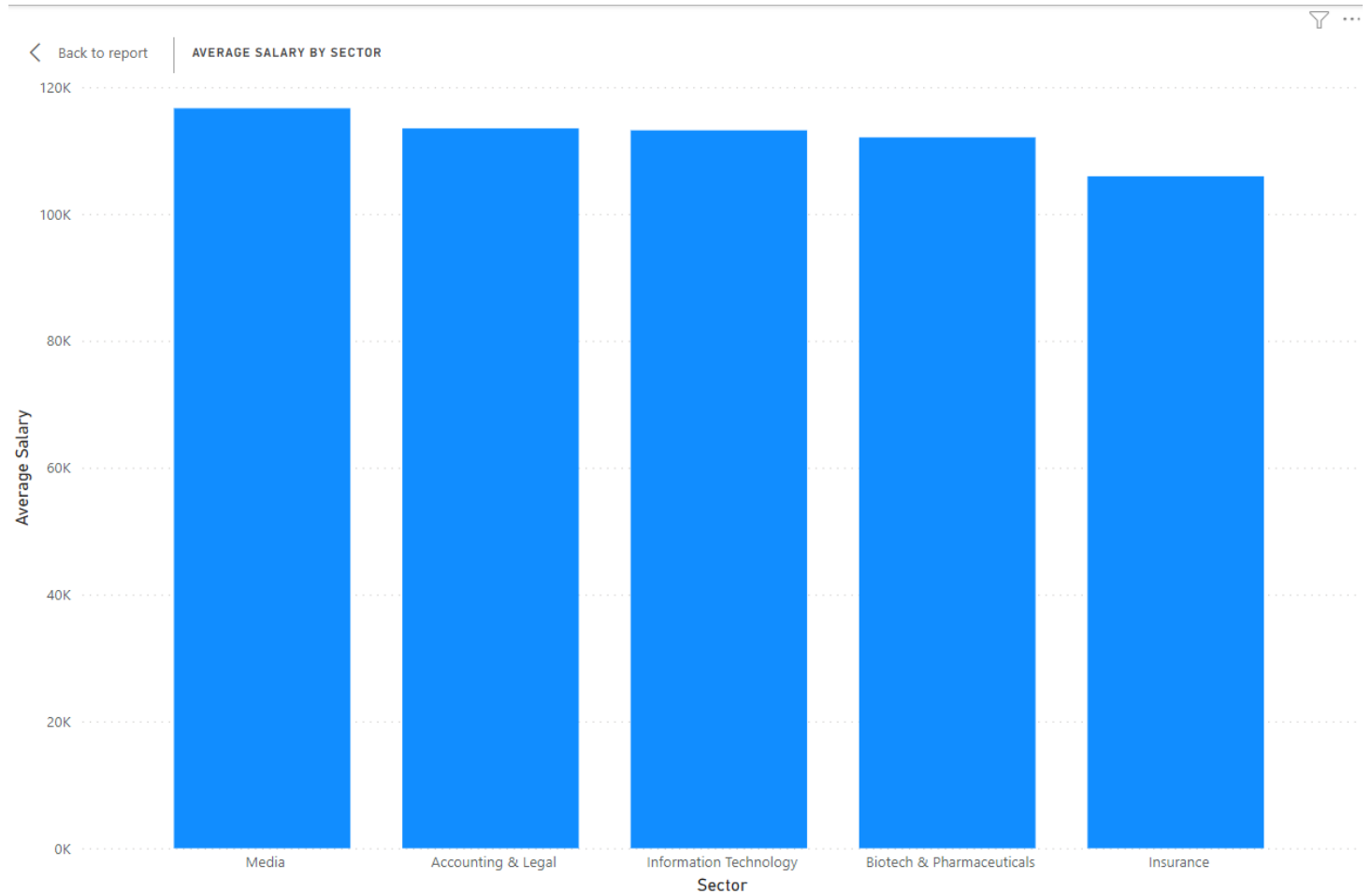
NOW IT IS CLEAR The average salary for an employee **depends** on the business domain he/she is working in.

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.

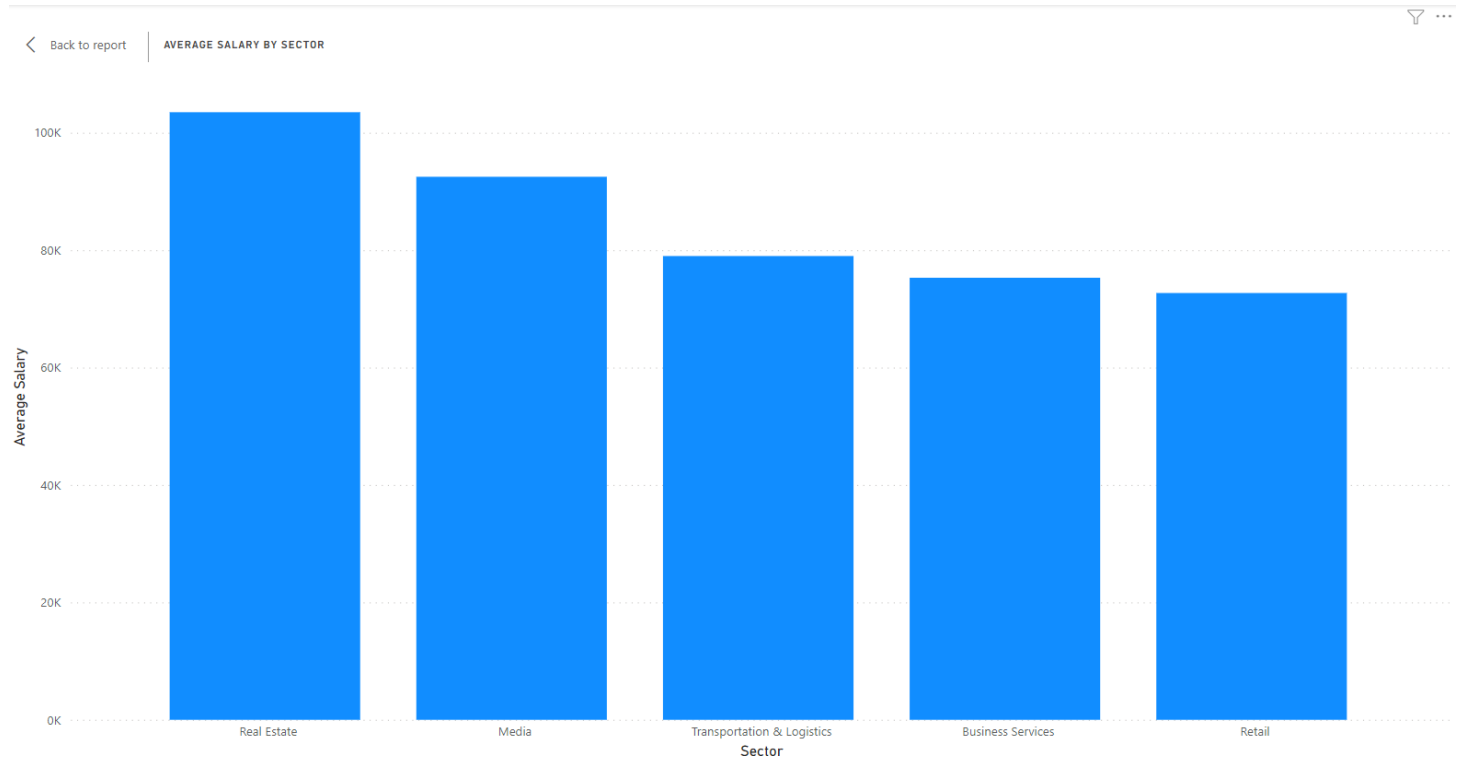
Doing that for 25 sectors makes no sense let's make it for the top 5 domains.



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.

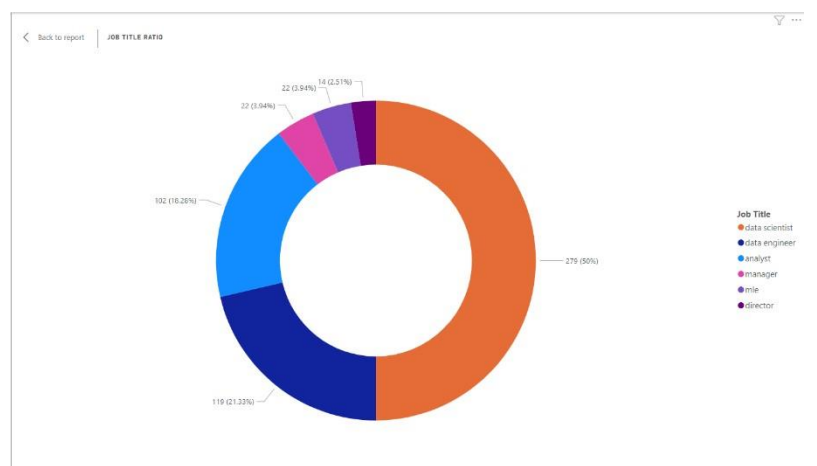


SOUNDS LIKE we found the bug here, as we can see the average salary of an employee **is connected** to the business domain.

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.

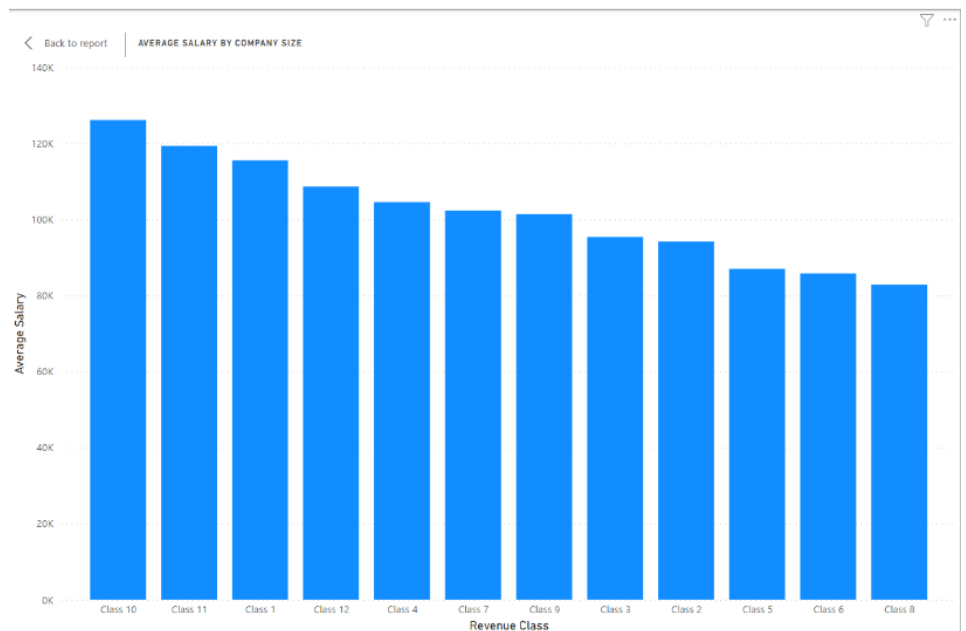
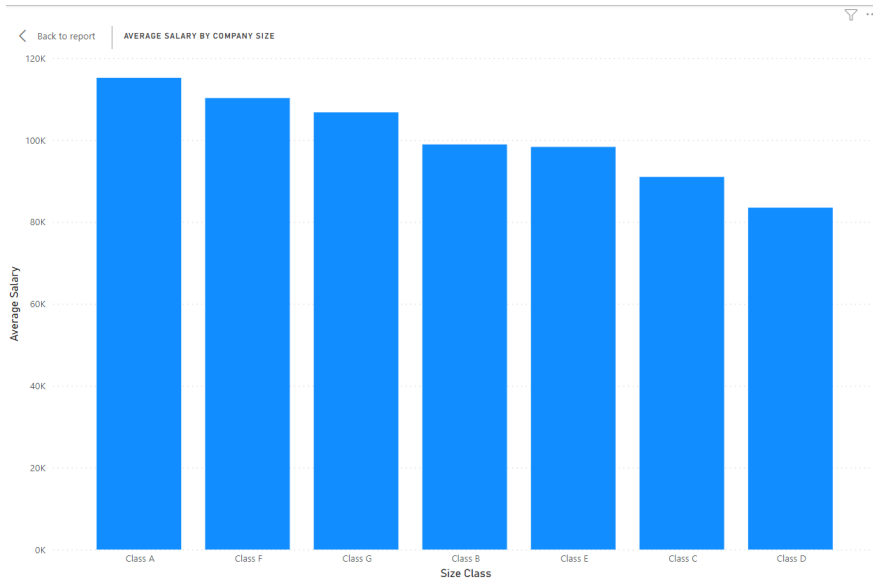
You may compare a data analyst working with the finance team with another one working with the media team, but never compare a data analyst working with the finance team with a data scientist working with the media team, **just a rookie mistake**.

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.

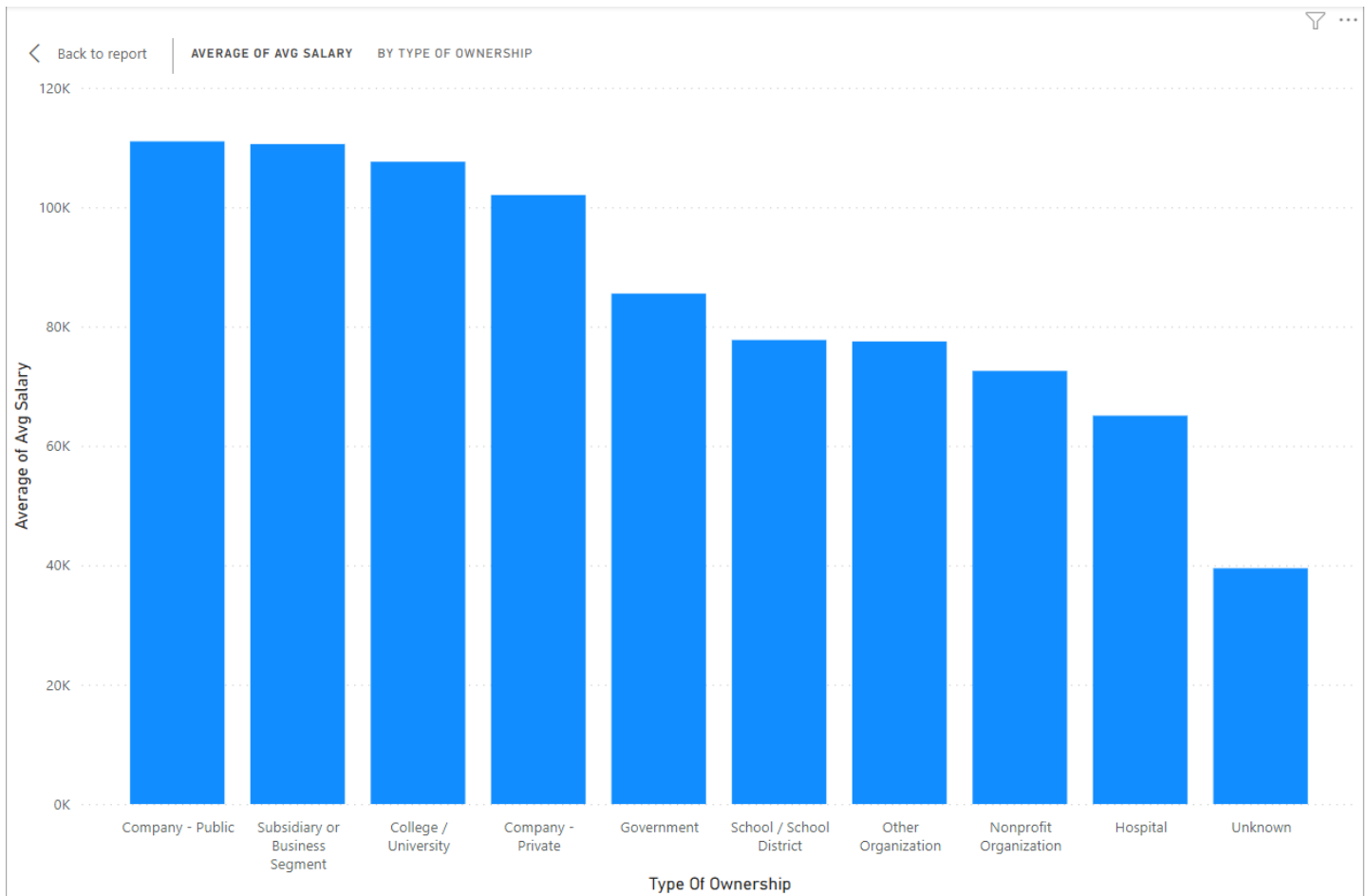


SOUNDS LIKE

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

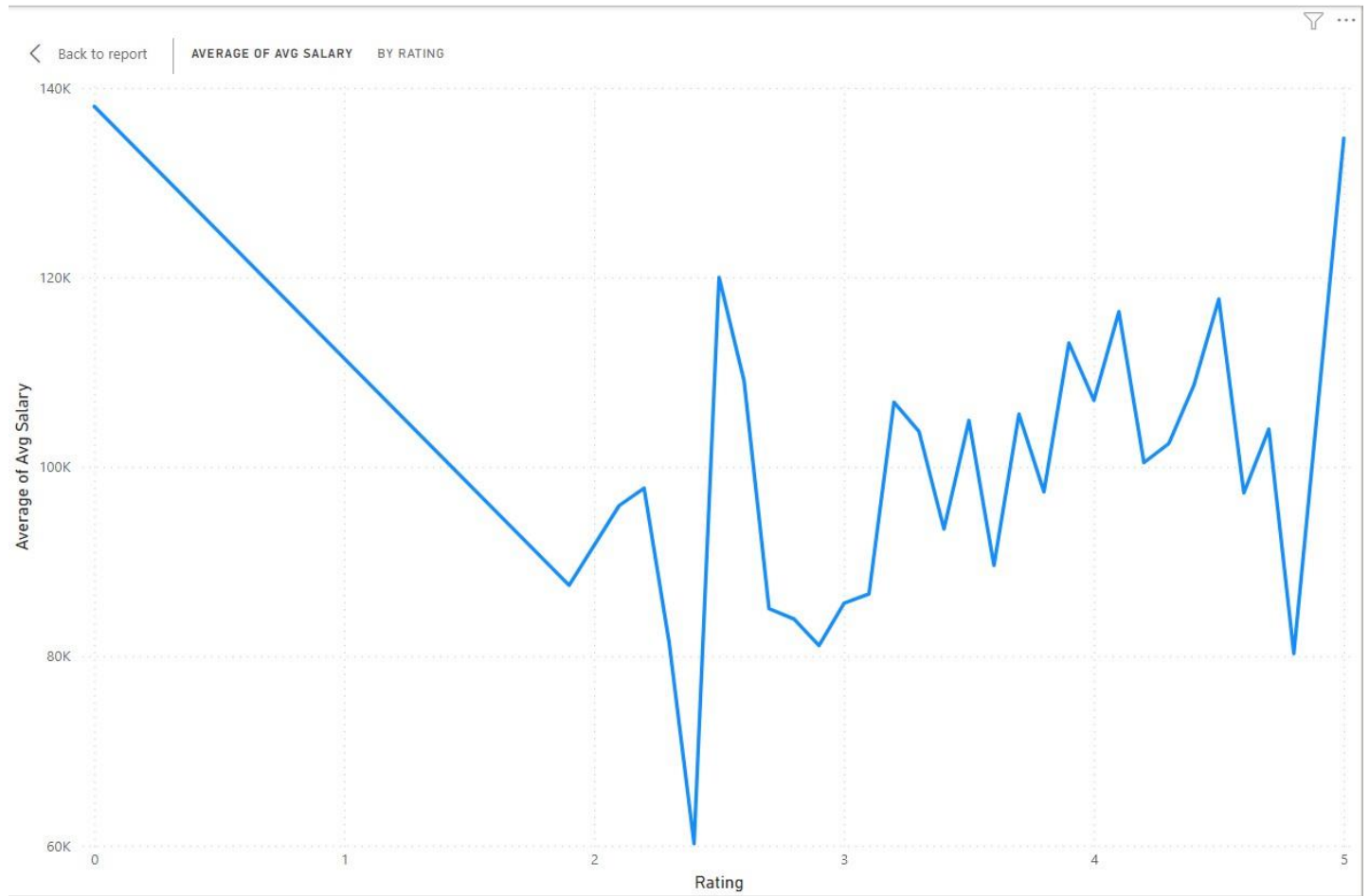
To be completely clear there is a strong connection between the company's revenue and the average salary, but it is not as clear as go on work for a bigger company, the expected salary would go something like: Class 1 has higher salaries & Class 12 the lowest.

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



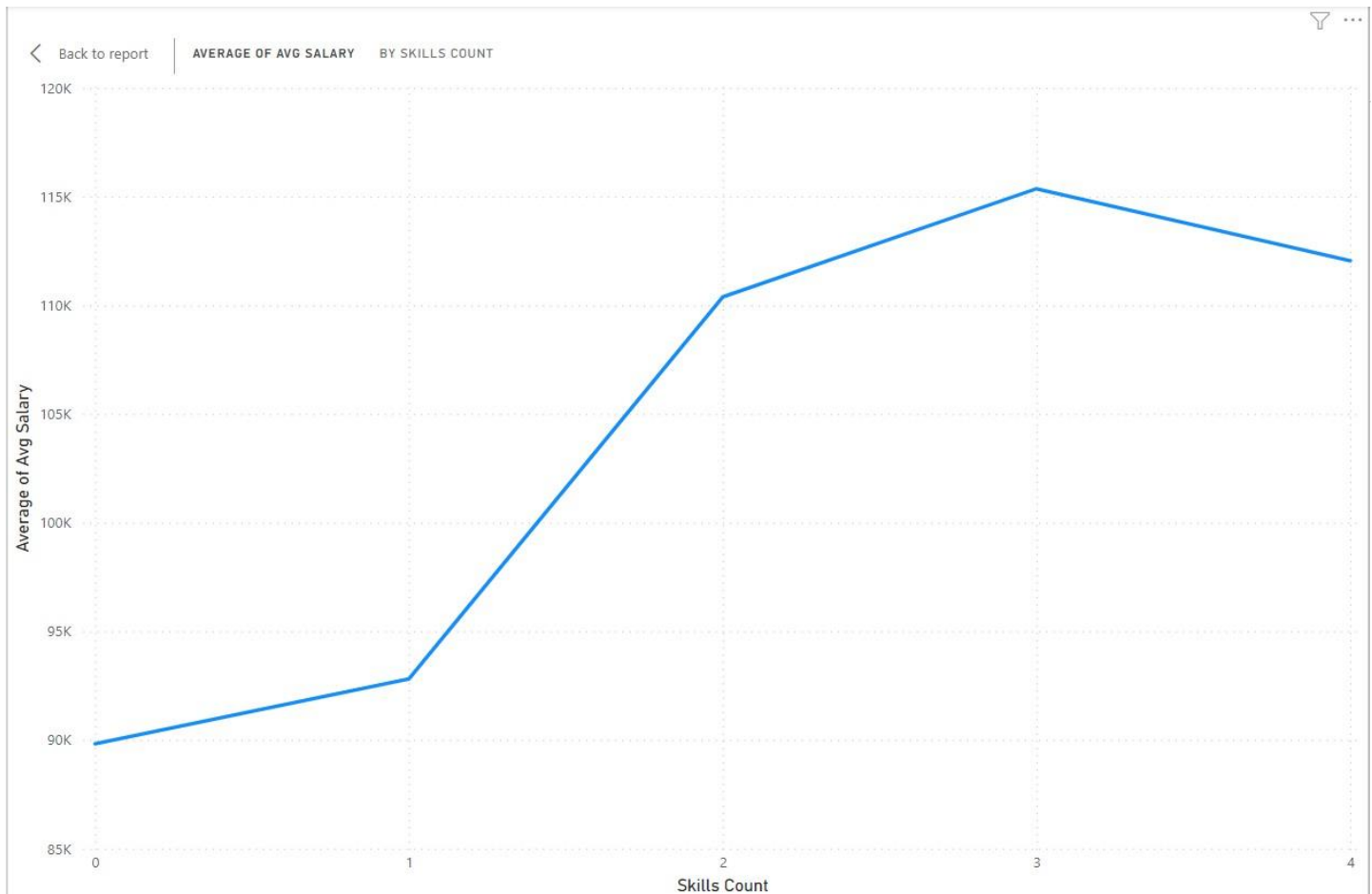
FOR ME I **would not** say there is a direct relationship, as even the non-profit companies are not far.

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?

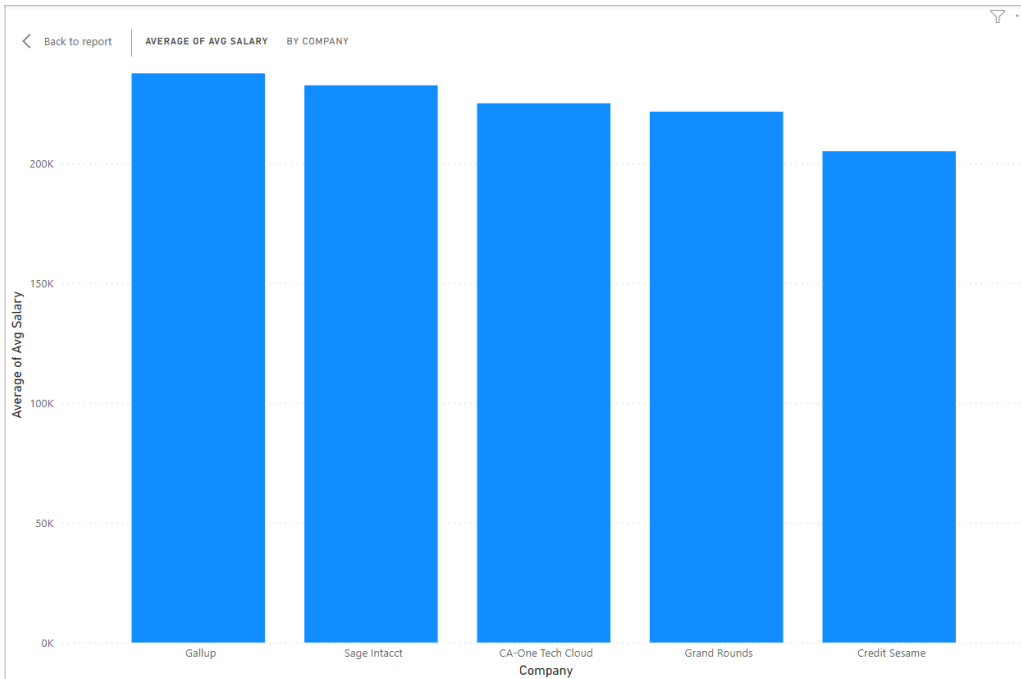


I WOULD SAY the more the better.

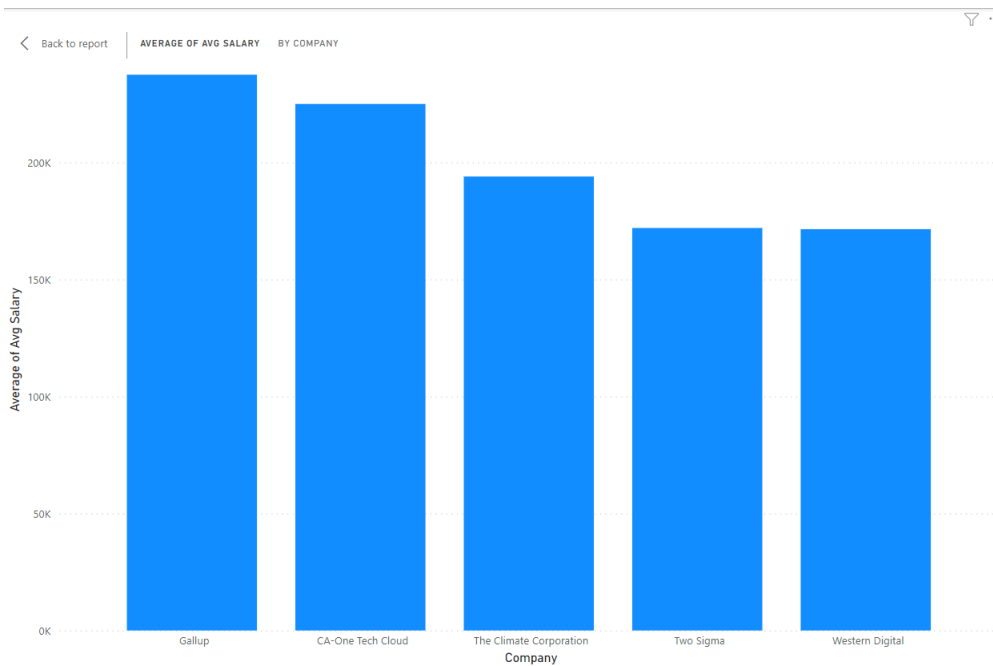
I have added a **Drill** page to the dashboard connected with the skills count, to show you the having-ratio for each skill (how many employees has each skill as a ratio).

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.



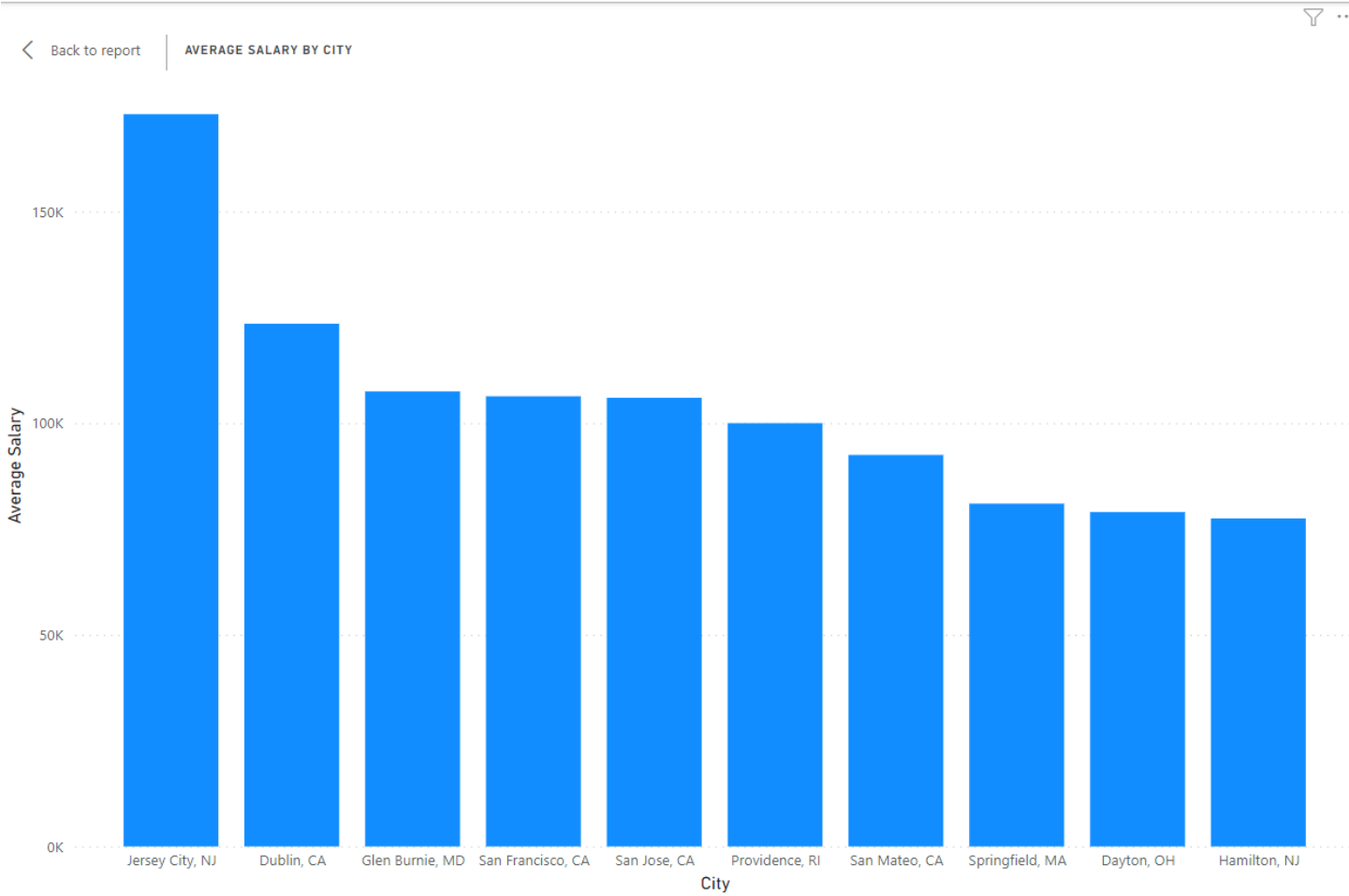
SOUNDS CLOSE let's work for a specific job title (**Data Scientists**).



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?



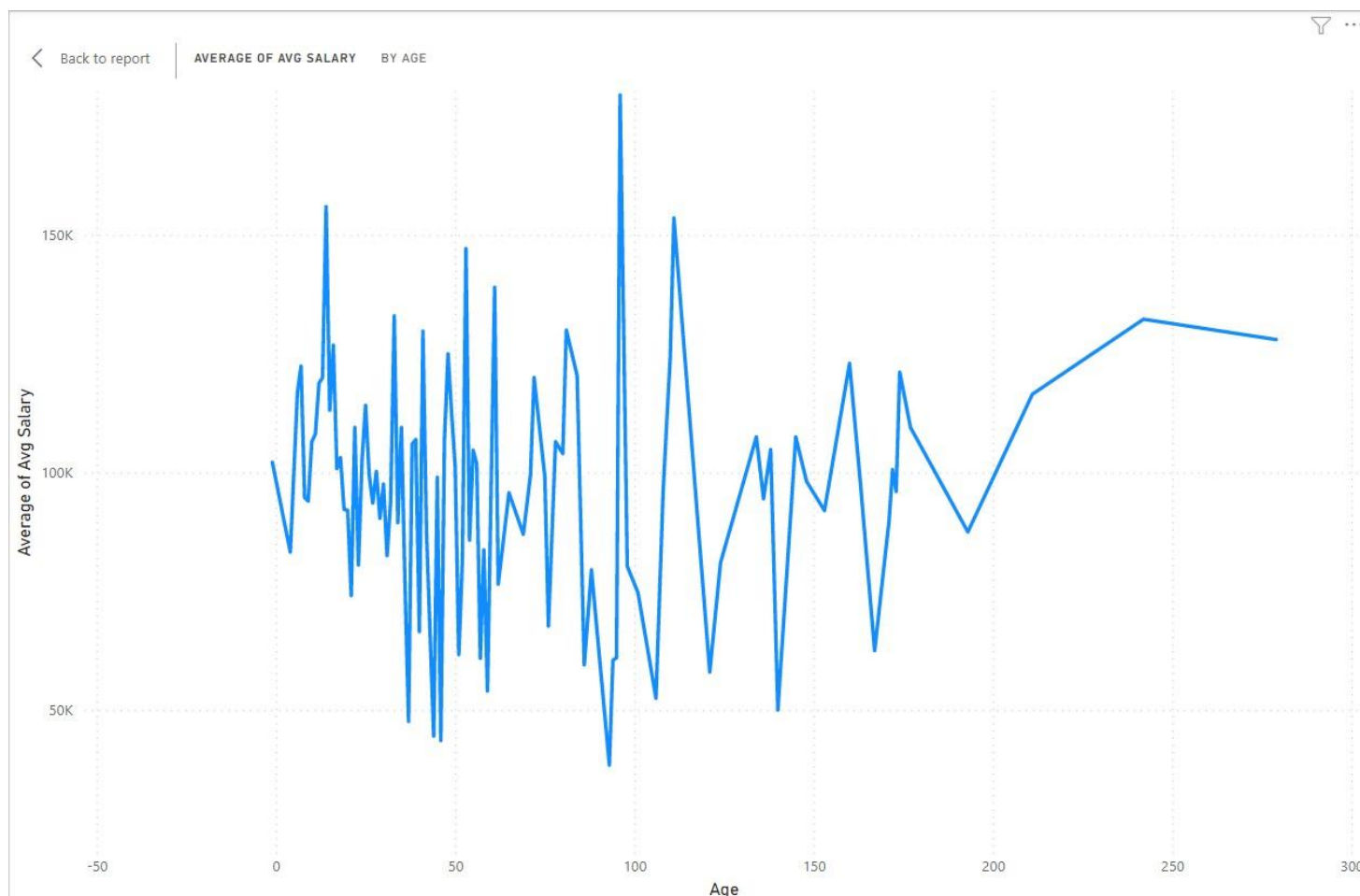
SOUNDS LIKE the city where the company is located **has a direct impact** on the average salary for a specific job title.

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?

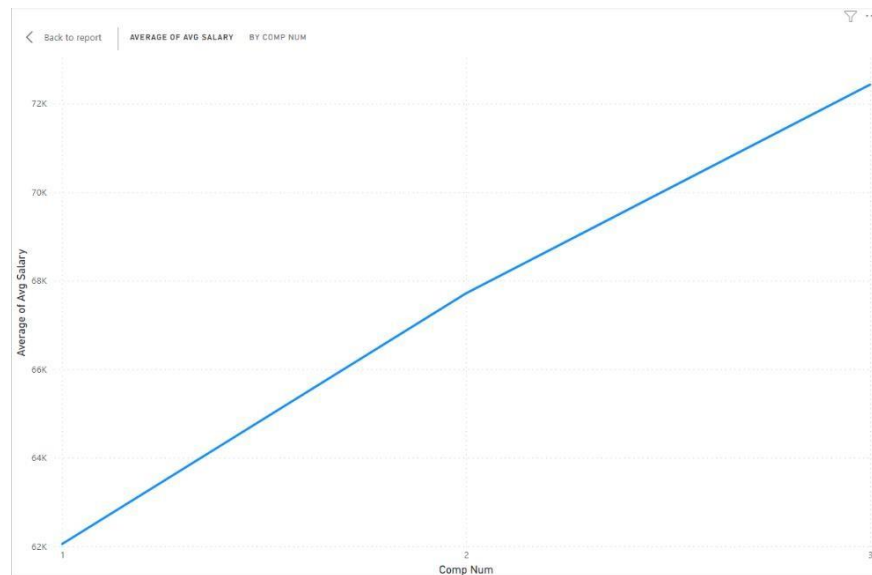


SOUNDS LIKE the foundation date of the company **have no impact** over the salaryrange.

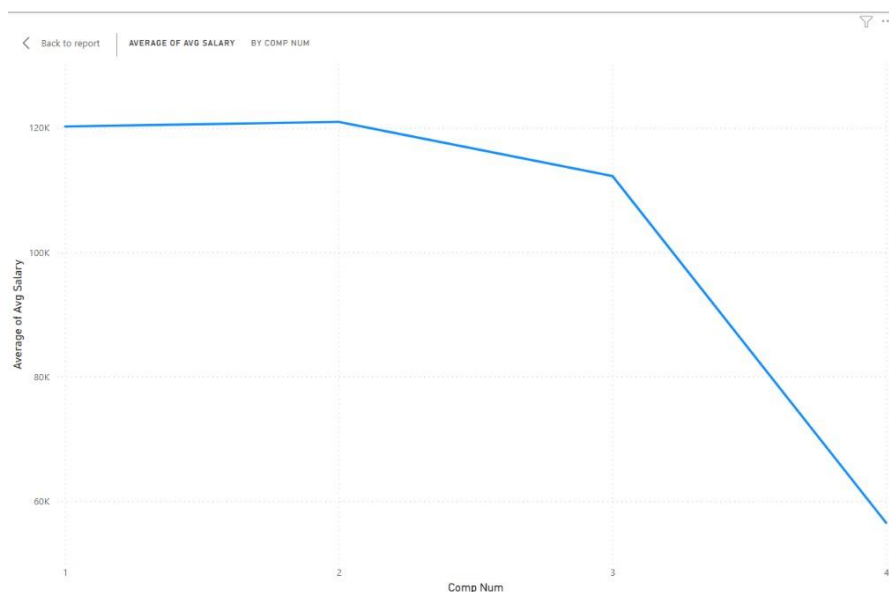
Now it is time to look at one of the most interesting factors to look at, I really was waiting to get to this moment, does the competition the company has affect the salary range for each employee?

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



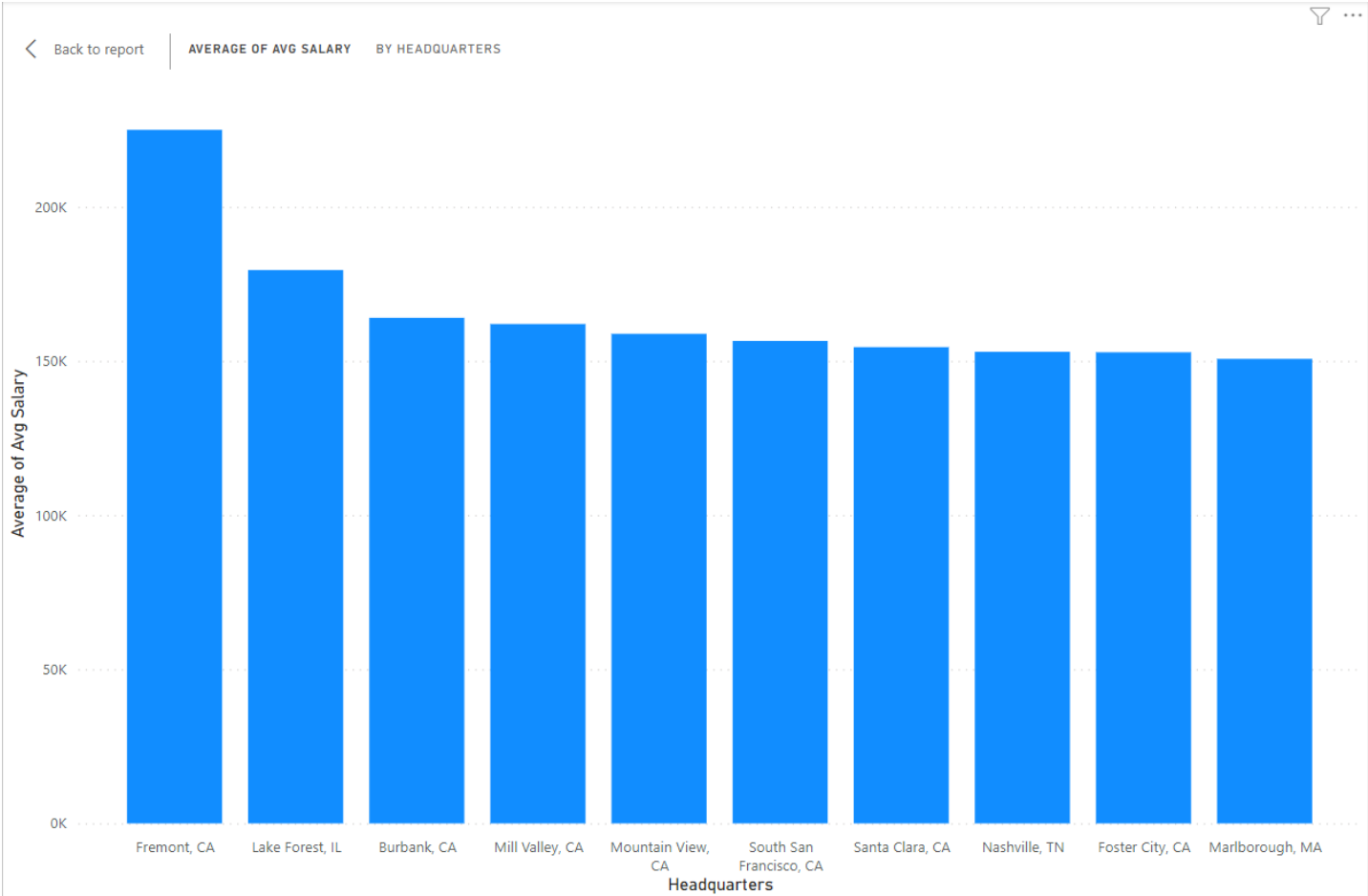
For Data Scientists



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.

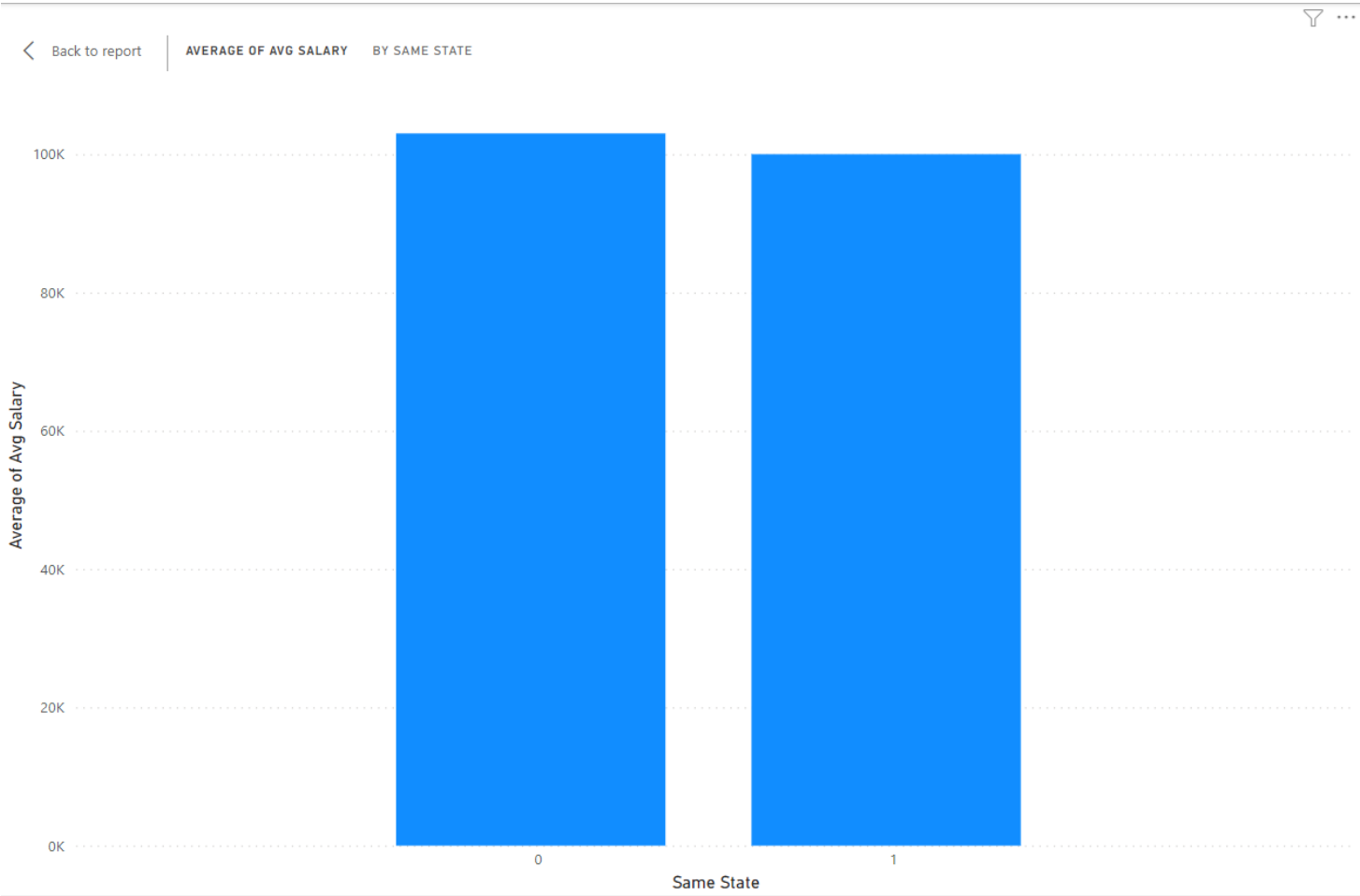
As a Software Engineer, I have always looked at Silicon Valley companies with respect, actually I have always looked at California with respect, but does it really make a difference where the company's headquarters are located?



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?

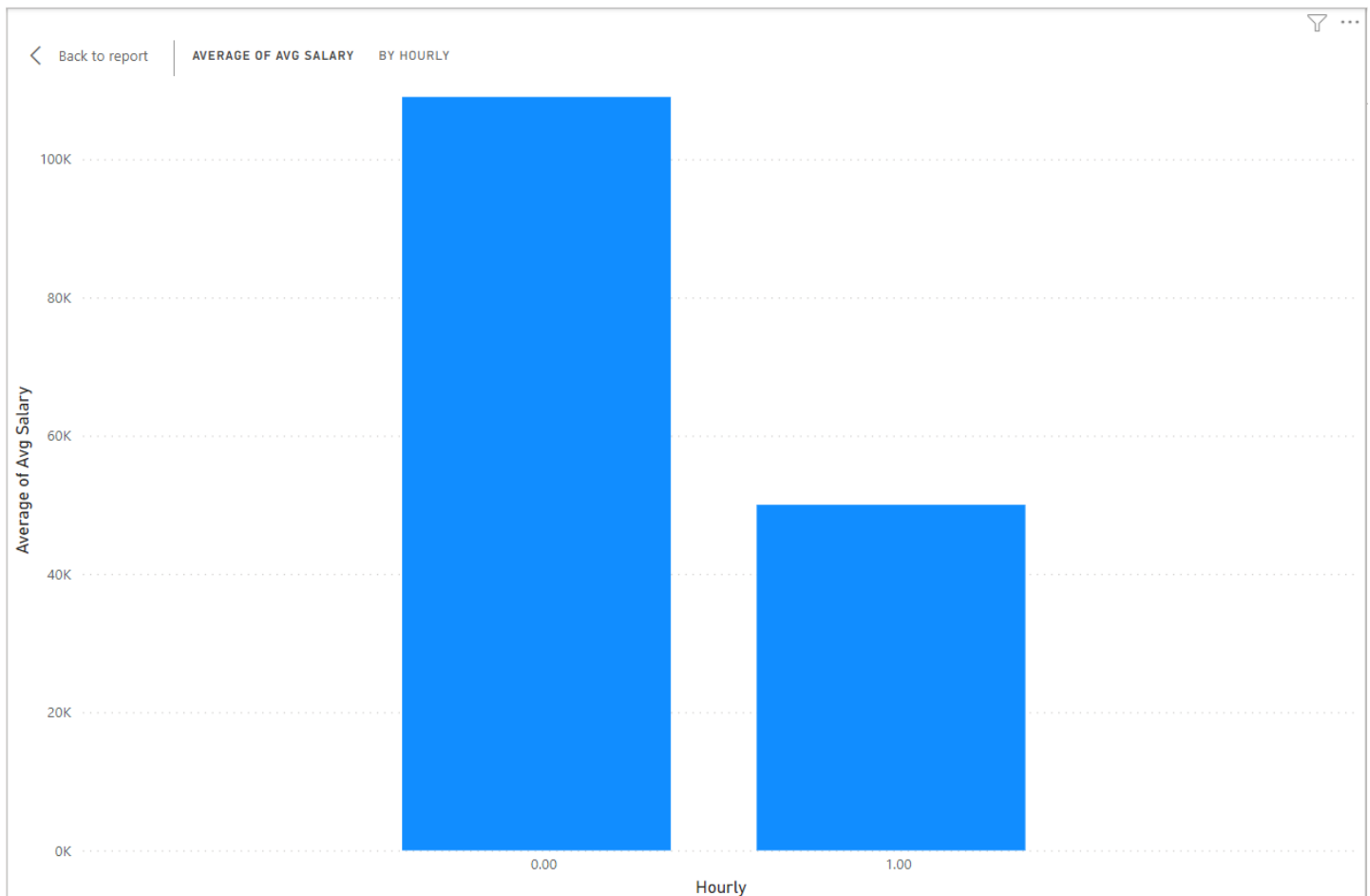
If I got a job offer from a company located in another state, I would expect more money to cover rent, travel, and being away from my family, let's investigate if this is true.



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

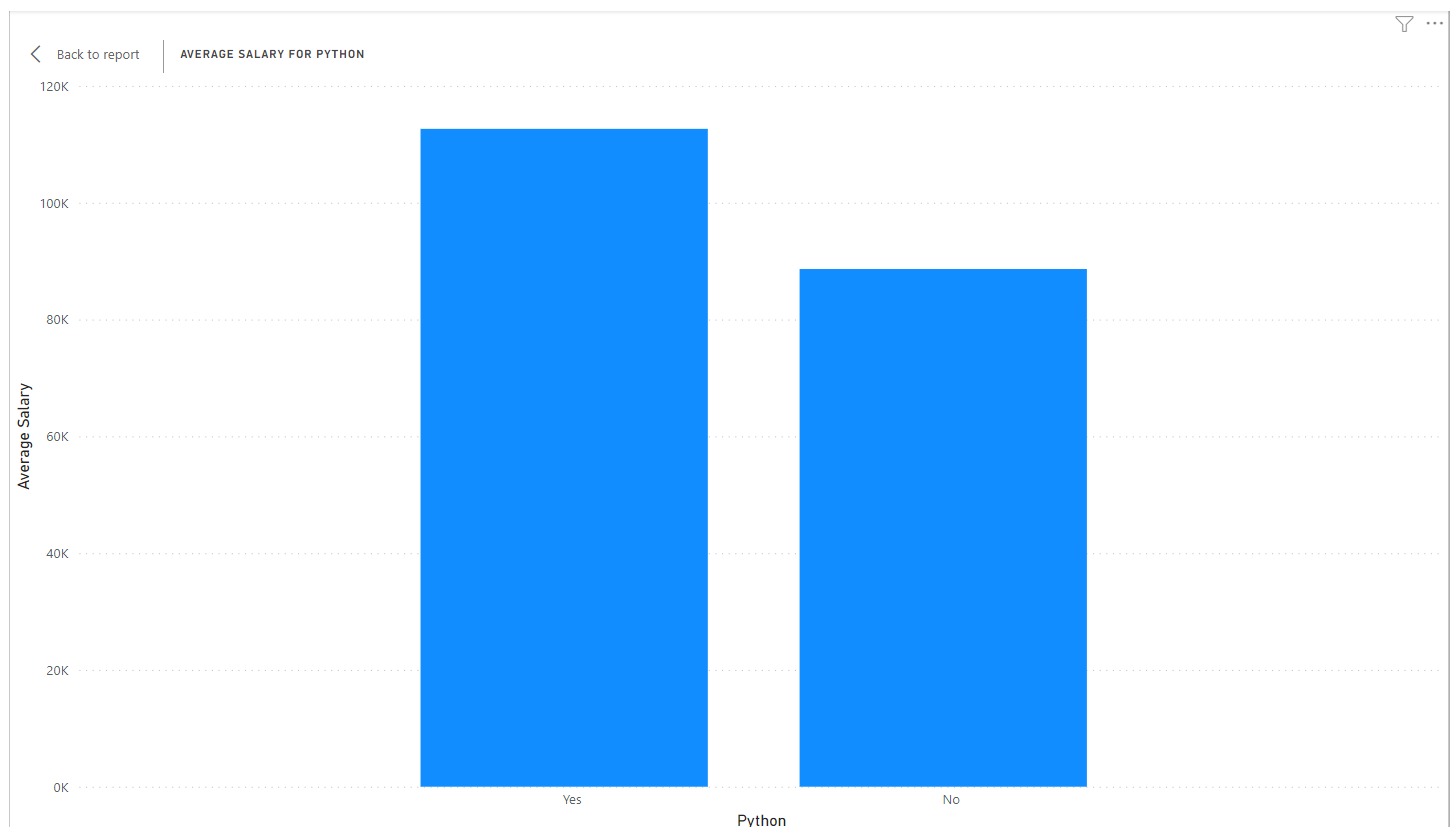
All the pervious analysis is considered a basic analysis, let's deep dive into the data, do some data mining, and extract some really hidden insights.

Here is a question?

I am eager to become a data specialist, I have the basic knowledge, but I want 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.

DASHBOARD

Salary Analysis for All Jobs



Salary Analysis for All Jobs



STORYTELLING

Introduction

In today's competitive job market, companies are striving to attract the best talent, particularly in high-demand fields like Data Science, Data Engineering, and Analytics. However, understanding what factors drive salary expectations for these roles is a complex task. That's where our in-depth analysis of the data-related job market can provide a game-changing advantage for your business.

By leveraging real-time data insights and exploring key variables such as skills, industry, company size, and location, we are able to offer tailored recommendations that can help businesses optimize their recruitment strategies, better compensate their talent, and stay ahead of competitors.

The Challenge

In the fast-paced world of data-related jobs, companies often struggle with:

- **Attracting the Right Talent:** Identifying what skills and job titles are in demand and paying competitive salaries.
- **Optimizing Compensation:** Understanding how company size, industry, and location influence salary expectations.
- **Strategic Workforce Planning:** Determining how to position their company to attract top talent in a highly competitive market.

The Solution: In-depth Job Market Data Analysis

We have conducted a comprehensive analysis of job postings, focusing on roles like Data Scientist, Data Engineer, Data Analyst, and more. Our dataset, derived from Glassdoor and Eda Data files, covers over 342 companies and more than 200 locations, offering a deep dive into the factors that impact salaries.

Here's how our analysis helps solve the challenges:

1. **Understanding Salary Trends:**
 - **Impact of Skills:** By analyzing the correlation between skill sets (Python, R, AWS, Spark, etc.) and salary levels, we help companies identify which technical skills are most highly valued in the market. This empowers businesses to focus on acquiring and retaining talent with the most sought-after skills.
 - **Industry-Specific Insights:** Different industries offer varying salary packages. Our analysis reveals the industries that pay the highest salaries for specific job roles, allowing companies to strategically align their compensation offerings with industry standards.
 - **Company Size & Revenue Influence:** Larger companies with higher revenues tend to offer better compensation packages. We provide detailed insights on how company size and revenue impact salary expectations, guiding companies to adjust their offerings accordingly.

2. Location-Based Salary Insights:

- Geographical Trends: Locations like New York and San Francisco dominate as high-paying hubs for data-related jobs. Understanding regional salary trends helps companies position themselves effectively in specific locations or determine if remote work options can attract top talent.
- Competitor Analysis: Our research also includes competitor insights, showing how the number of competitors in the job market impacts salary levels. Companies can leverage this data to stay competitive and attract top talent without overextending their budgets.

3. Optimizing Recruitment Strategy:

- Tailored Compensation Packages: With insights into the average salary ranges for specific job titles, companies can optimize their compensation packages to stay competitive and avoid overpaying or underpaying candidates.
- Targeting the Right Talent: By understanding which skills are most in demand, businesses can streamline their hiring processes, focusing on candidates with the right skill set to drive success.

4. Building a Data-Driven Workforce:

- Skill Gap Analysis: Companies can identify skills that are underrepresented in their current workforce and prioritize upskilling efforts or recruit for these high-demand skills.
- Long-Term Strategy: Our insights provide businesses with actionable data to craft long-term workforce strategies, helping them stay ahead in a rapidly evolving job market.

Why This Matters for Your Business

- Increase Recruitment Efficiency: By understanding the market's salary trends and candidate expectations, your recruitment team can focus on attracting and retaining top talent with competitive offers.
- Enhance Talent Retention: Offering salaries that align with market expectations and skills can significantly improve employee satisfaction and retention, reducing turnover and associated costs.
- Maximize ROI on Talent: With accurate salary benchmarking, you can ensure that your company invests the right amount in attracting and retaining high-performing employees, driving business growth.
- Strategic Market Positioning: Stay competitive in the job market by offering attractive compensation packages that align with both industry standards and competitor practices.

Conclusion

With data-driven insights, companies can optimize their hiring and compensation strategies to remain competitive and attract the best talent. Our analysis not only provides a snapshot of the current job market but also offers actionable recommendations to ensure your business is

ahead of the curve.

By partnering with us, you'll gain the tools needed to navigate the evolving job market, ensuring that you remain an employer of choice in the data-driven economy.

Let's unlock the power of data to transform your recruitment strategy and secure your place as a leader in your industry.