

**Project Report**

**In-depth Analysis of  
 Data Professions**

Subject: DTA301 - Data Analysis

Lecturer: TienNQ27 - Nguyễn Quốc Tiến

Class: IS1803

Student Name: Đinh Gia Bảo

Student ID: SE183741

# **I - Introduction**

This report analyzes key trends in the data profession based on a survey of 630 professionals from various regions, including the U.S., U.K., Canada, and India. It focuses on average salaries by job title, favorite programming languages, and the challenges of entering the field. Additionally, the study highlights respondents' satisfaction with work-life balance and salary.

The goal is to provide insights into the data profession landscape, helping professionals and organizations understand career trends and challenges.

# **II - Methodology**

**Data Collection Methods:**

● The data for this analysis was sourced from [Data.gov Home - Data.gov](https://data.gov/) - **The Home of the U.S. Government's Open Data**. This platform provides access to a wide range of datasets, tools, and resources to support research and development across various fields.

● For this project, relevant datasets related to the data profession, including job roles, salary information, and industry challenges, were extracted and analyzed.

● The datasets were processed to provide insights into trends in salary, programming language preferences, and the difficulty of breaking into the data profession.

**Data Analysis Methods:**

● Key statistical techniques and visualizations, such as bar charts, pie charts, and treemaps, were utilized to illustrate insights from the survey

**Tools and Software Used:**

● R programming language: Used for data processing, statistical analysis, and visualization.

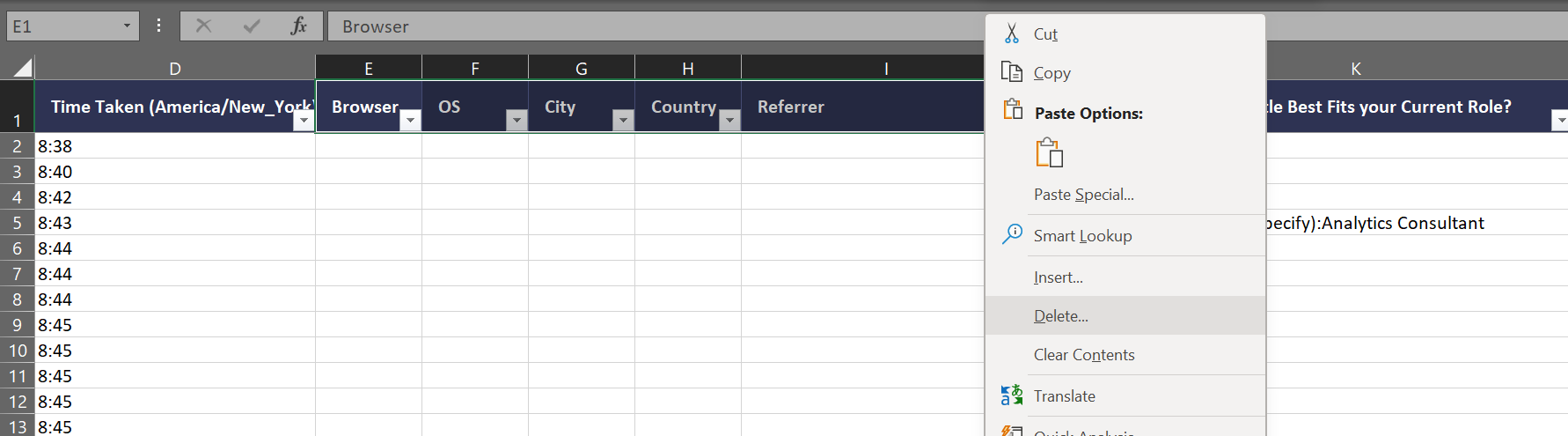
● R packages: Dplyr, ggplot2, reshape2 were utilized for data manipulation and visualization tasks.

● Microsoft Power BI: Used to create interactive visualizations of data.

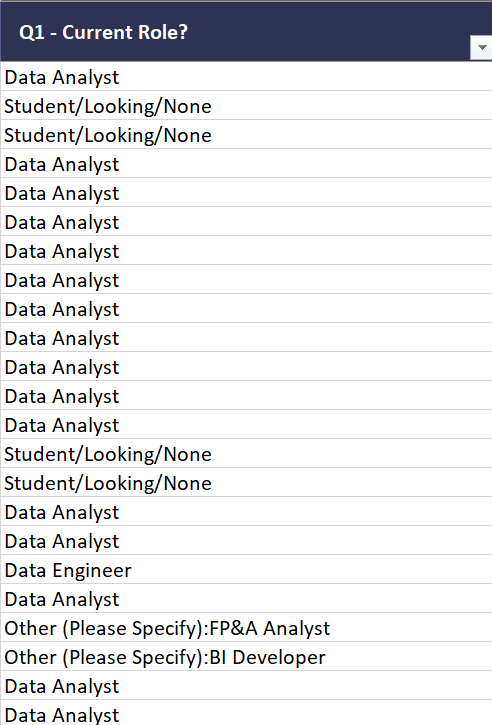
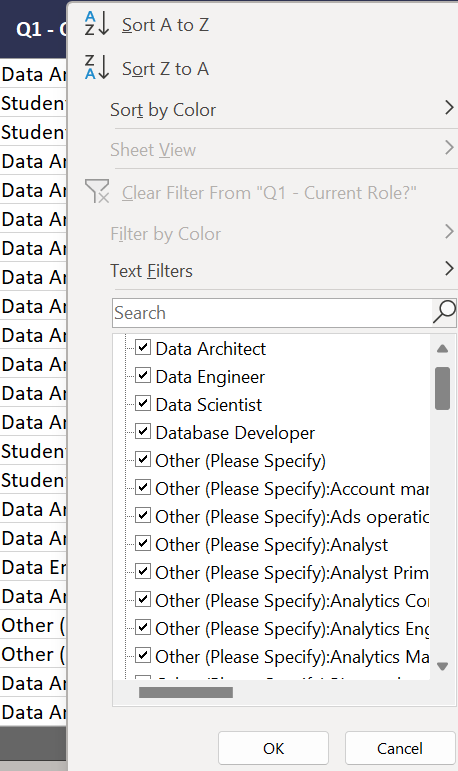
● The combination of R and Power BI allowed for a detailed breakdown of survey results, helping to identify trends and satisfaction levels across different job roles in the data profession.

# **III - Data Transformation**

Delete columns whose data was empty (**Cleaning** step)



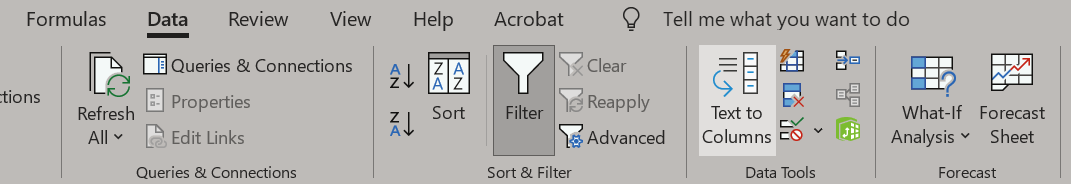
Split Other in to a specific column

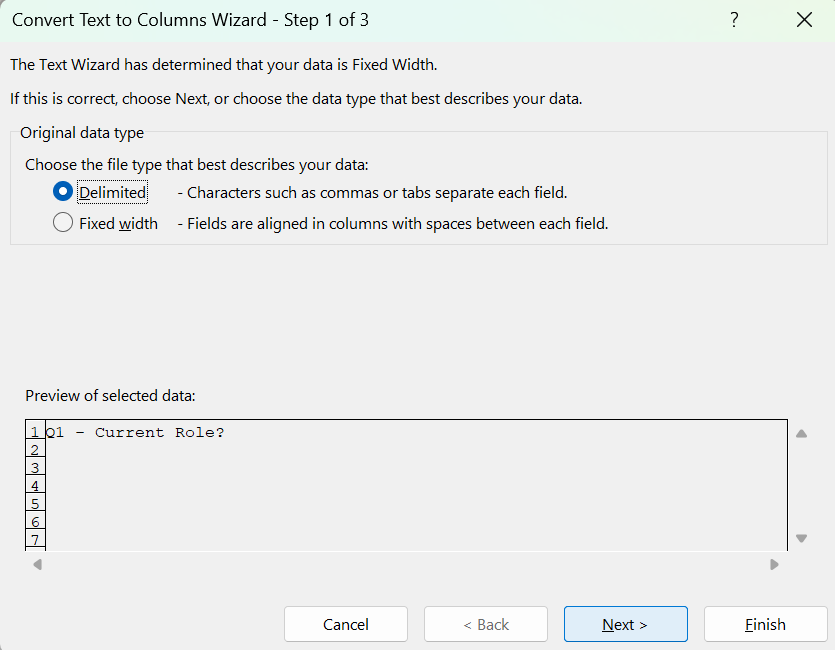
Choose a column

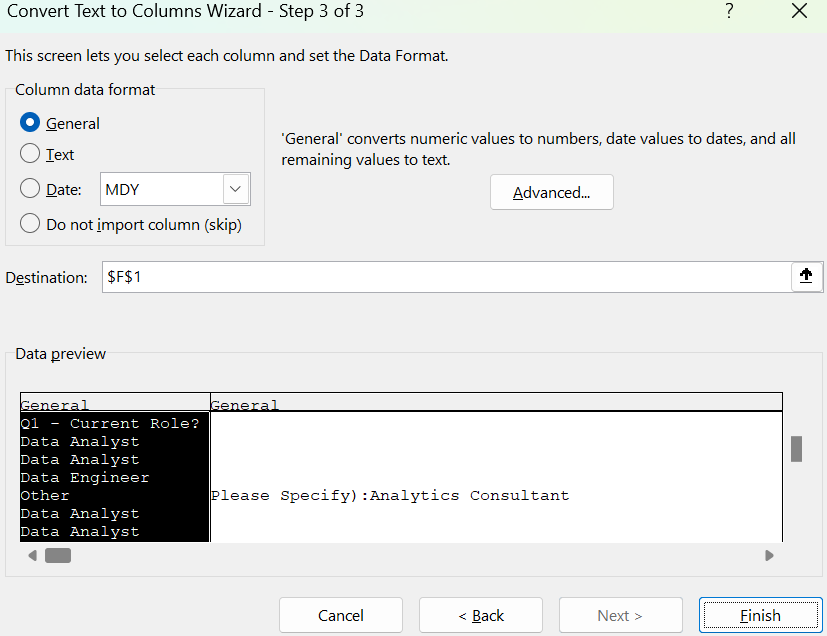


Choose Data on Ribbon. In Data Tools choose Text to Columns

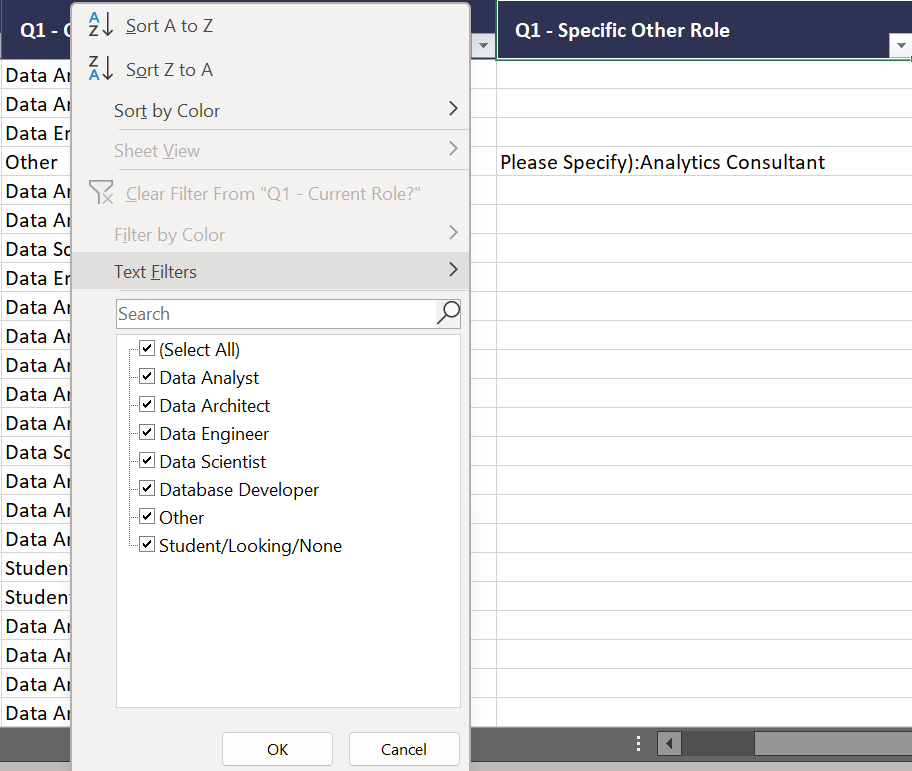


Choose delimited and Next

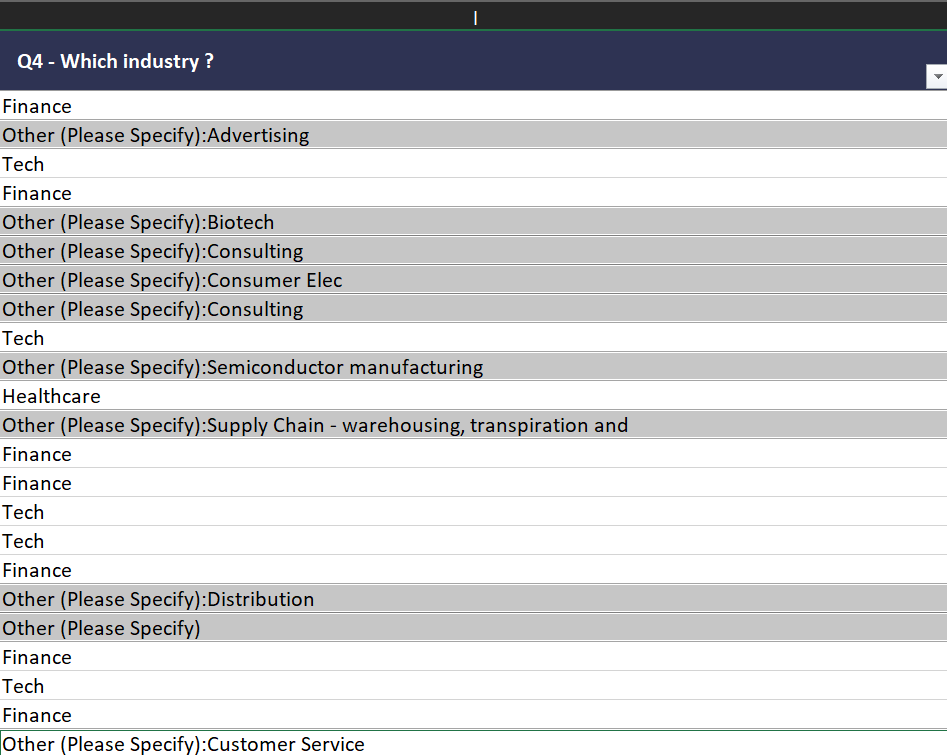
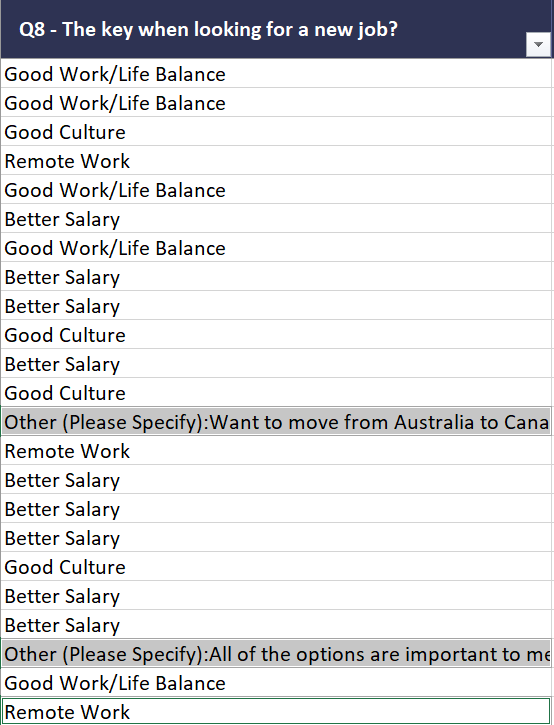
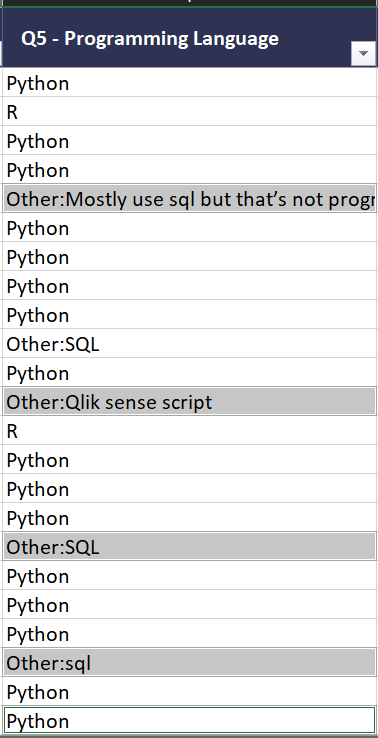
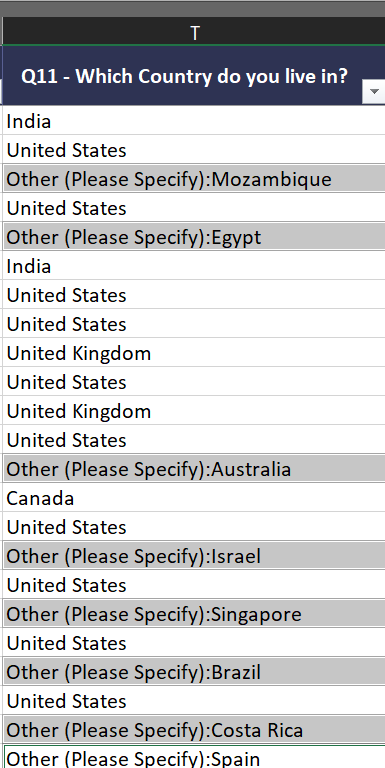


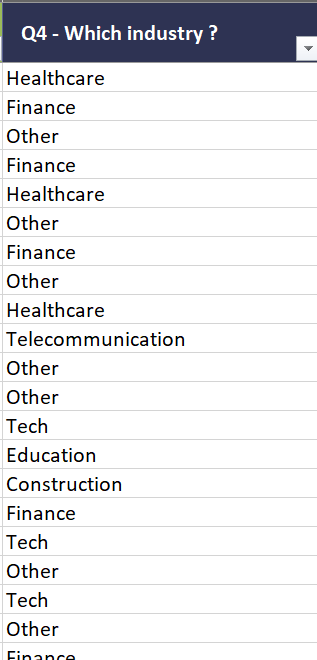
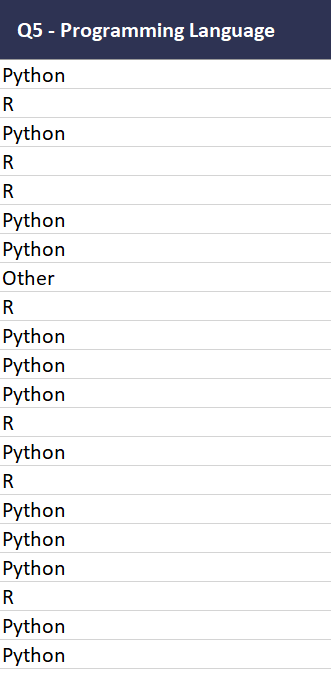


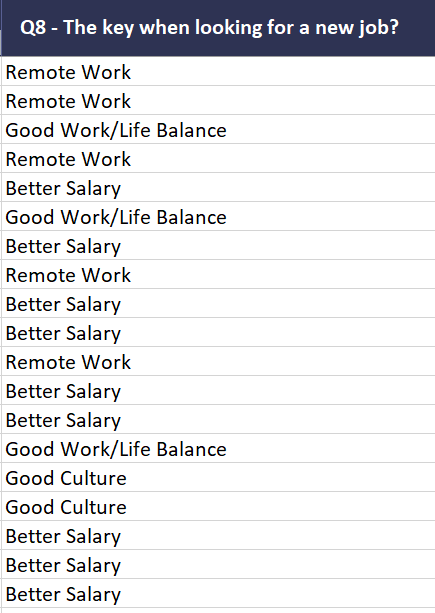
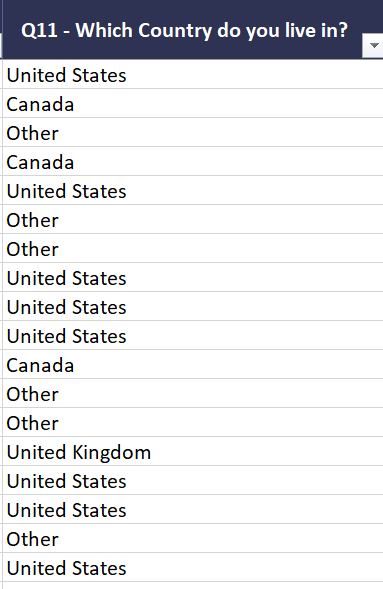
Re-check and complete. After that I will deleted the separated Columns



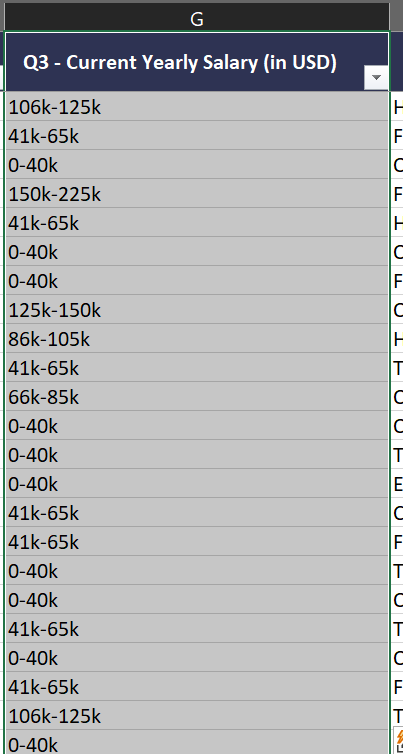
Doing the same techniques for the “Q4.Industry, Q5.Programming, Q8.Key finding, and Q11.Country” columns

Data after modified:  
 

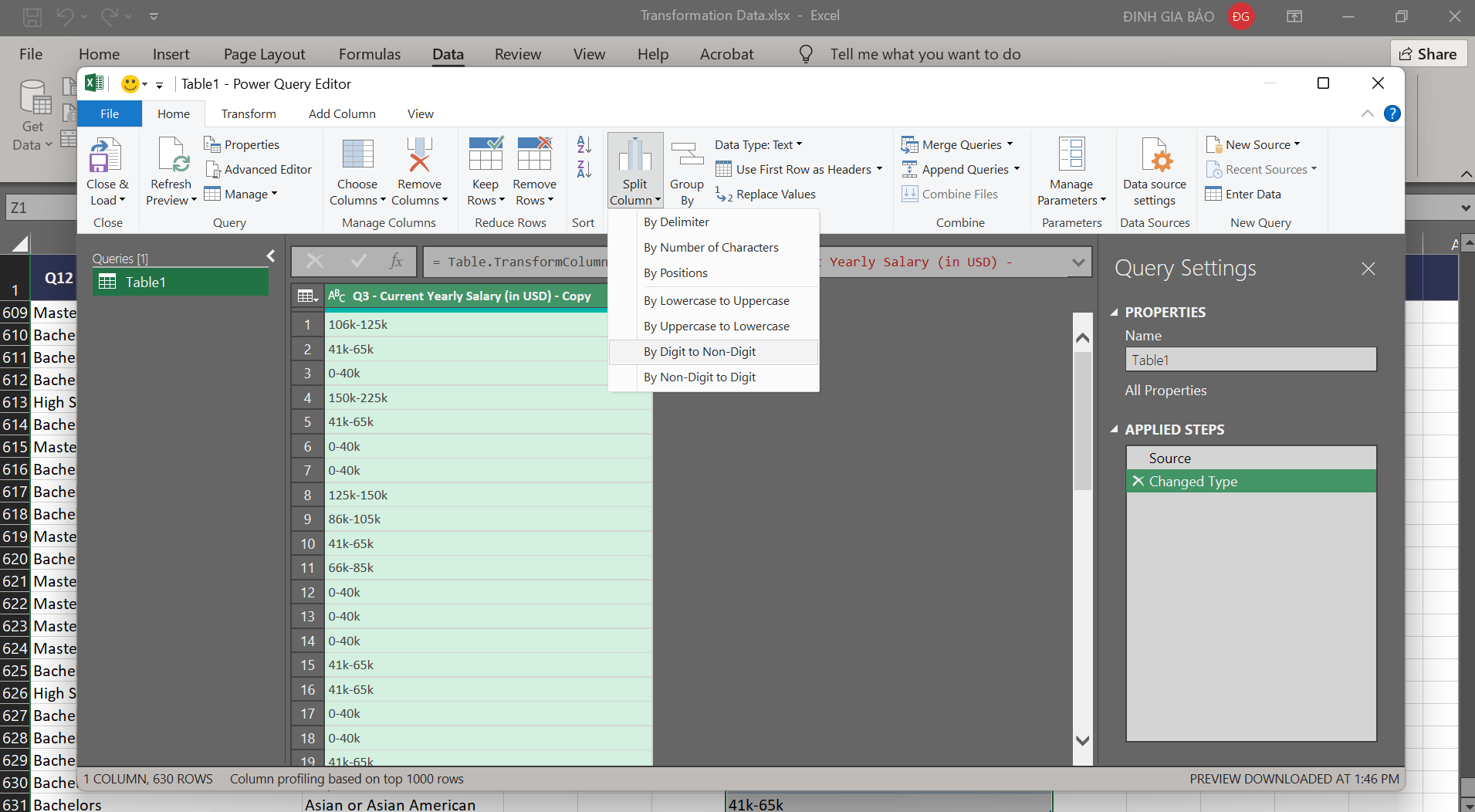
 

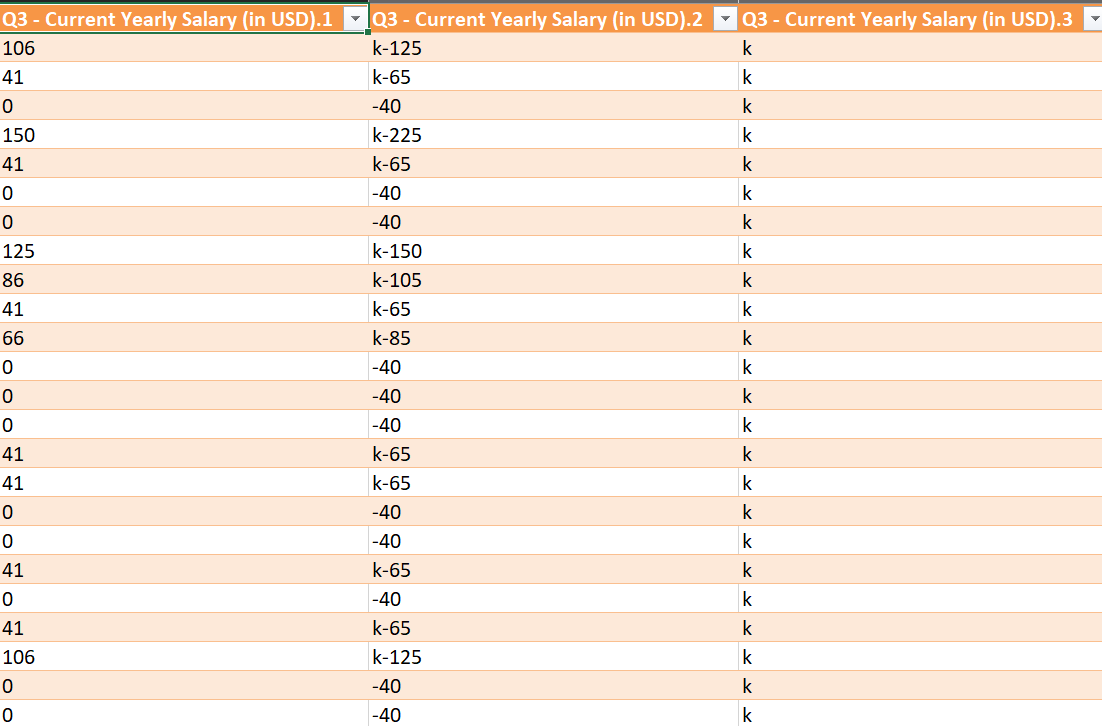
**Modified “ranging data” to specific average data**



**1. First step:**

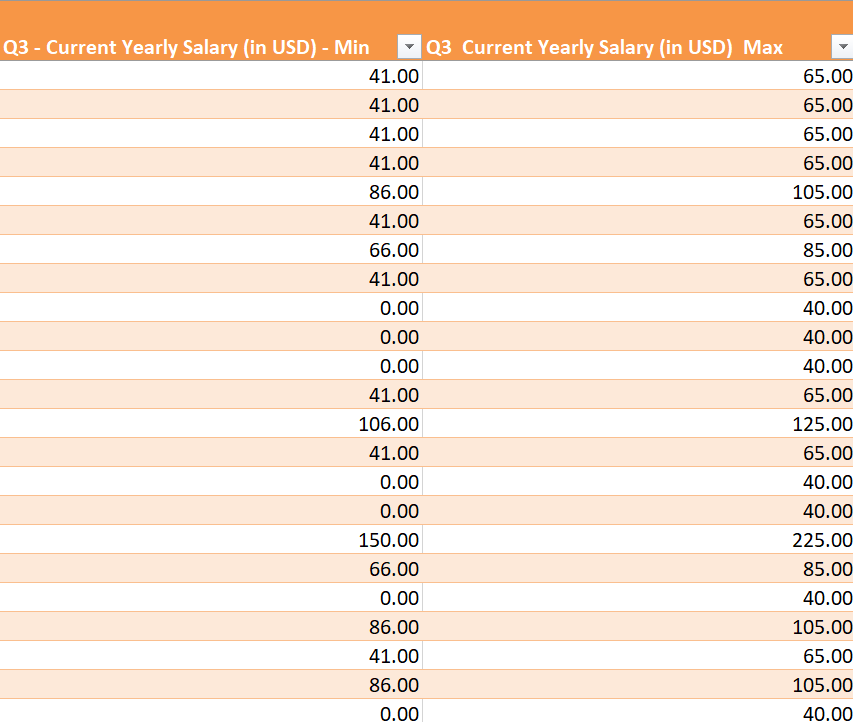
Using power query editor to change digit to non digit





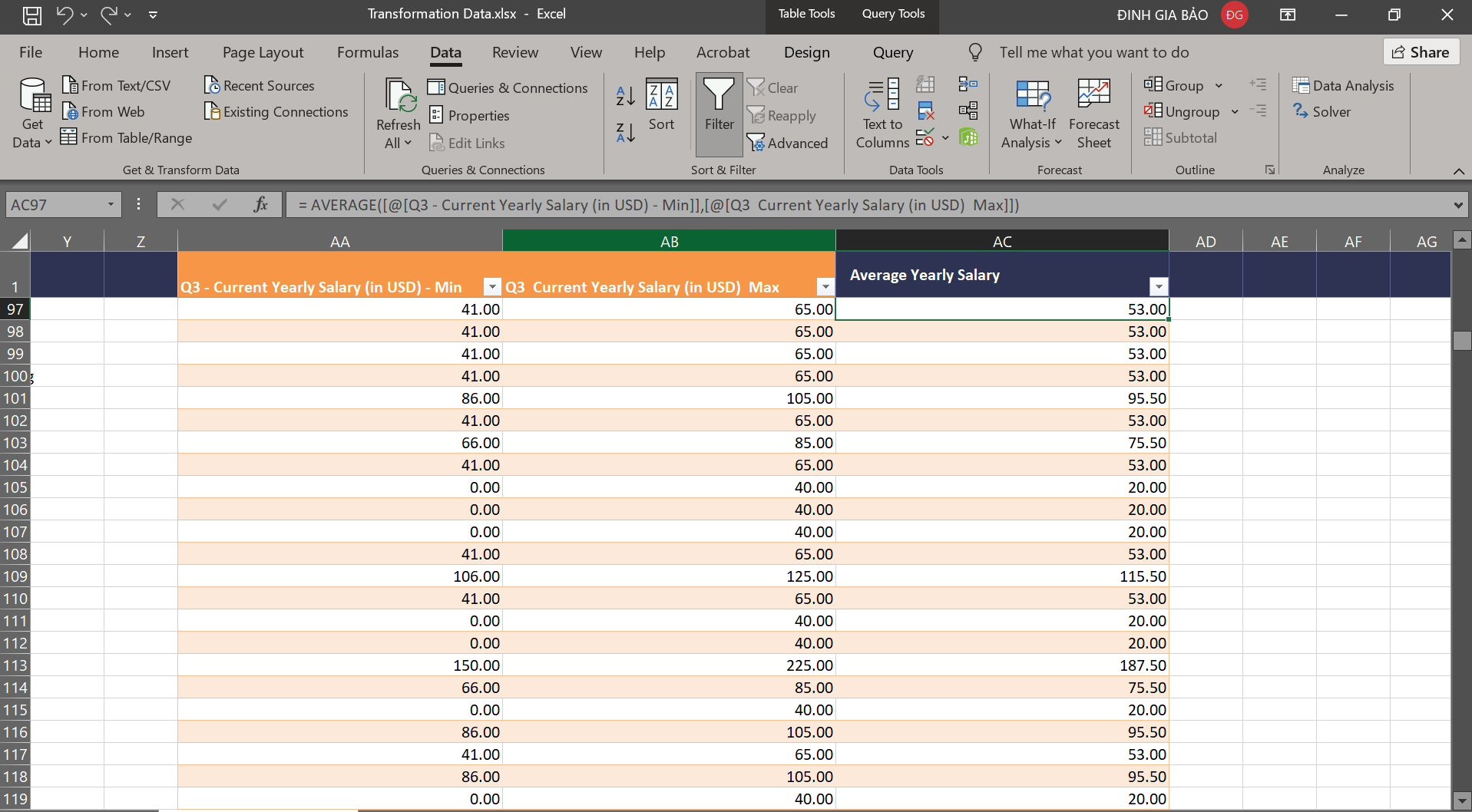
**2. Second step**

Delete column Q3.3 , and modify data in Q3.1 and Q3.2 to meaning data



**3. Final step**

Calculate average yearly salary based on Minimum and Maximum value.



# 

# **IV - ANALYSIS WITH R**

## GENERAL ANALYSIS

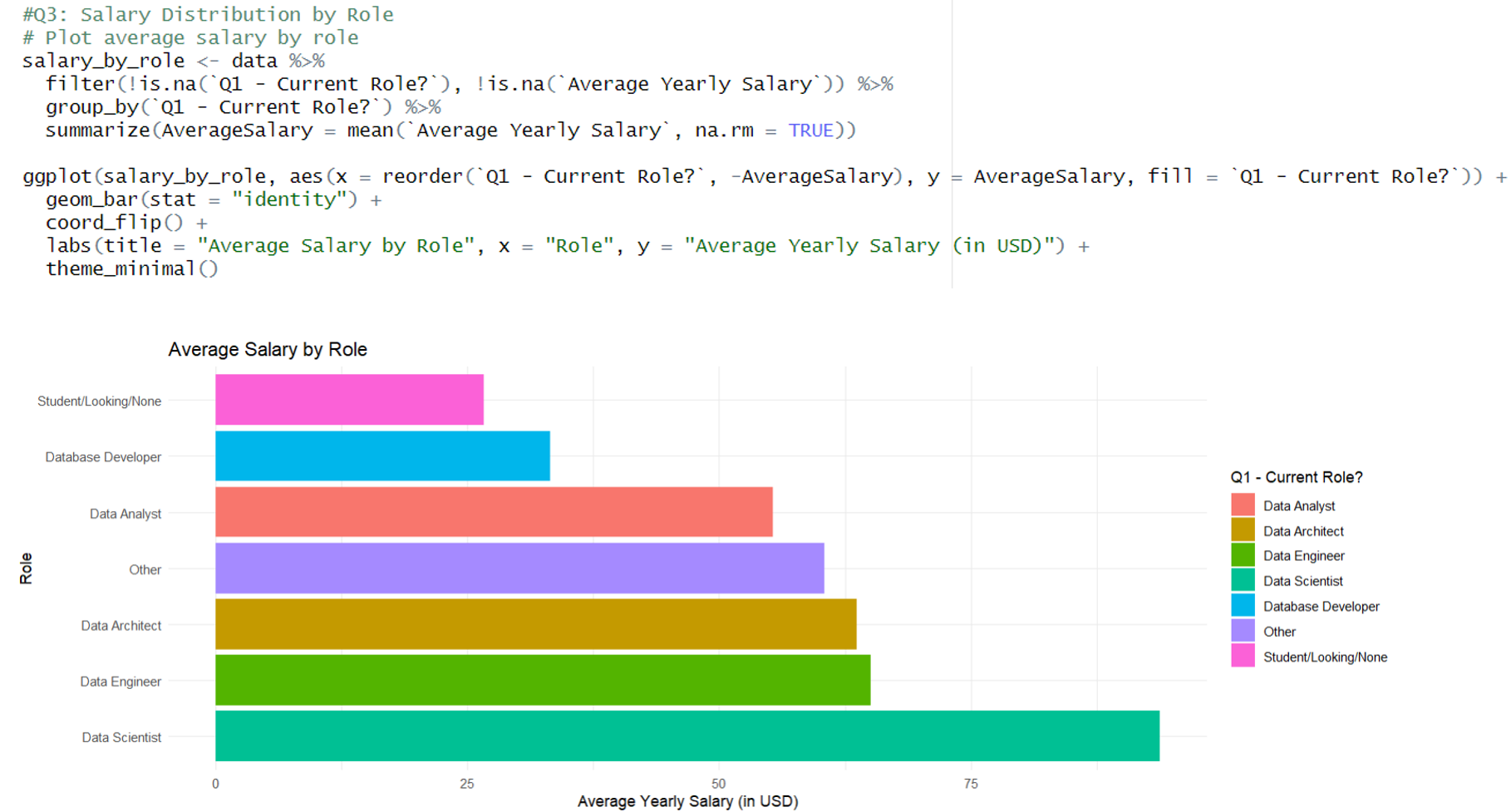
1. **Distribution of Current Roles**

****

* **Data Analysts** and **Data Scientists** dominate: These two roles are in high demand in the labor market, reflecting the increasing need for data analysis and machine learning applications.
* **Data Engineers** also play a crucial role: This role is essential for building and managing data infrastructure, supporting analytical activities.
* The diversity of other roles: Besides the three key roles mentioned, there are many other specialized roles such as **Data Architect** and **Database Developer**, highlighting the high level of specialization in this field.
* Growth potential: The large number of people studying and seeking jobs in the data field indicates a promising future.

=> This presents an overview of the job market in the data industry, with high demand for data analysis and data science professionals.

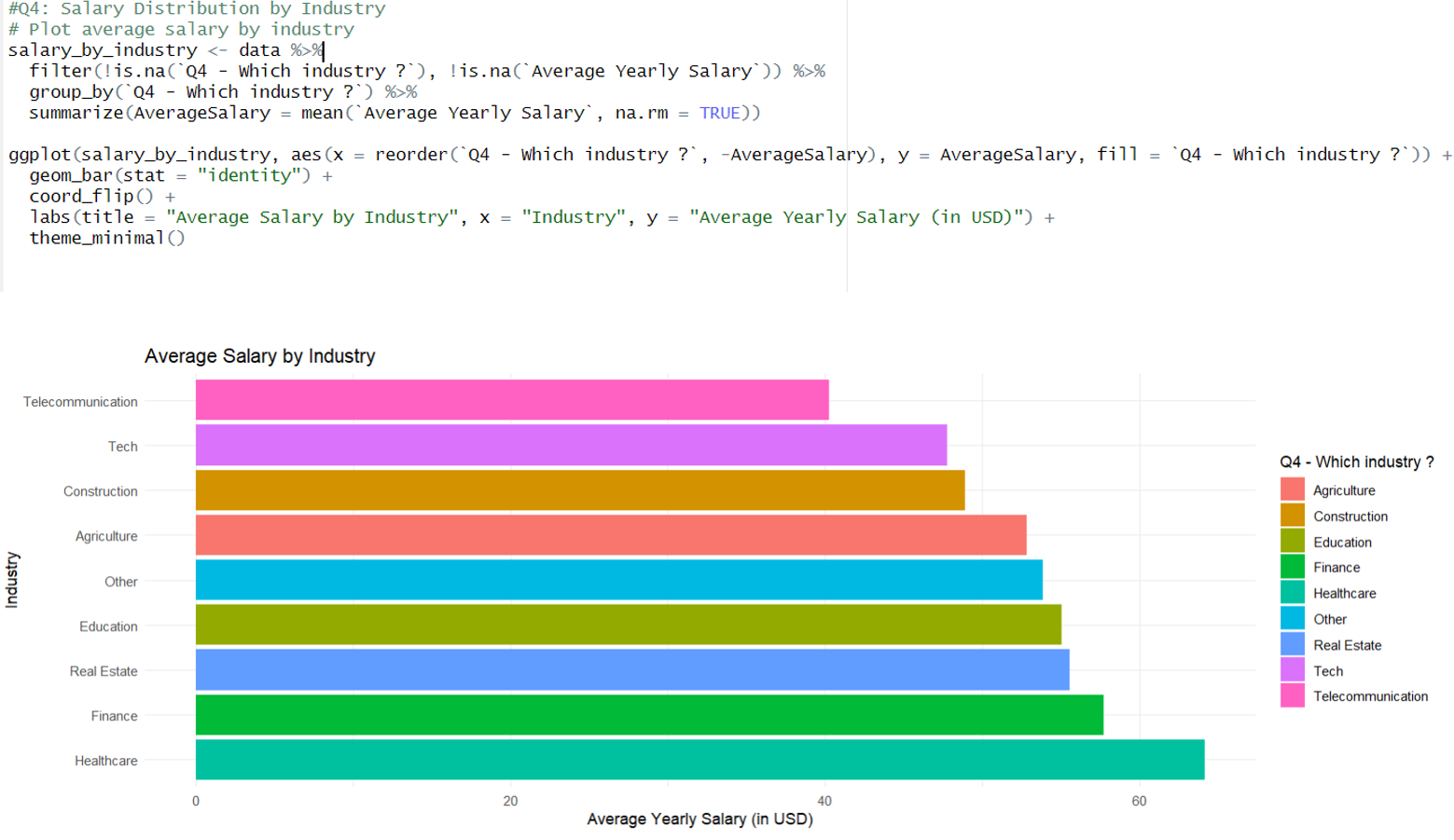
1. **Salary Distribution by Role**



* **Machine Learning Engineer** (Data Scientist): Has the highest salary, reflecting the high demand for AI/ML skills and the scarcity of talent in this field.
* **Data Engineer**: Second highest in salary, showing the importance of building and managing data infrastructure.
* **Data Analyst**: Has a lower salary, often considered an entry-level position.
* Other roles: Salaries vary depending on the level of expertise and responsibility of each role.

=> Clearly illustrates the relationship between salary and roles in the data field. Specialized skills and market demand are the key factors influencing the salary for each role.

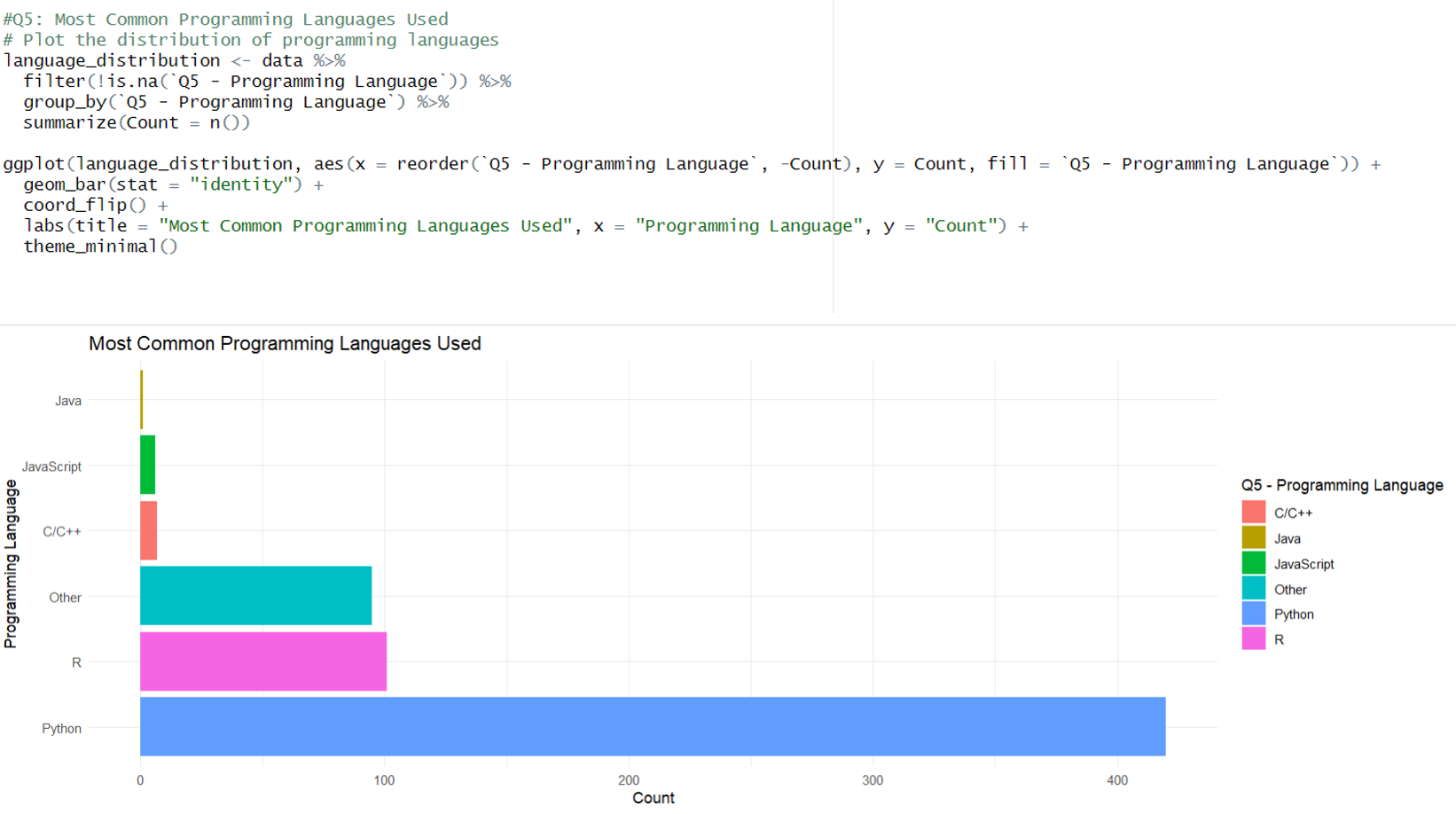
1. **Salary Distribution by Industry**



* **The tech industry** leads in **salary**: This reflects the financial strength and heavy reliance on data of tech companies.
* **Finance** and **Banking** rank second: This shows the strong investment of this sector in data analytics.
* Significant differences between industries: **Education** and **Nonprofit organizations** tend to offer the lowest salaries.
* **Consulting** is highly competitive: This reflects the trend of outsourcing data analysis services.
* The salary gap may lead to a concentration of experts: Top professionals may concentrate in certain industries.
* Salary also depends on the complexity of the data and the importance of analysis in each industry: These factors influence the salary levels.

=> The chart illustrates the salary disparity across industries, with the tech sector leading. Factors such as market demand, data complexity, and the importance of analytics in each industry affect salaries. The salary gap could motivate data professionals to further develop their careers.

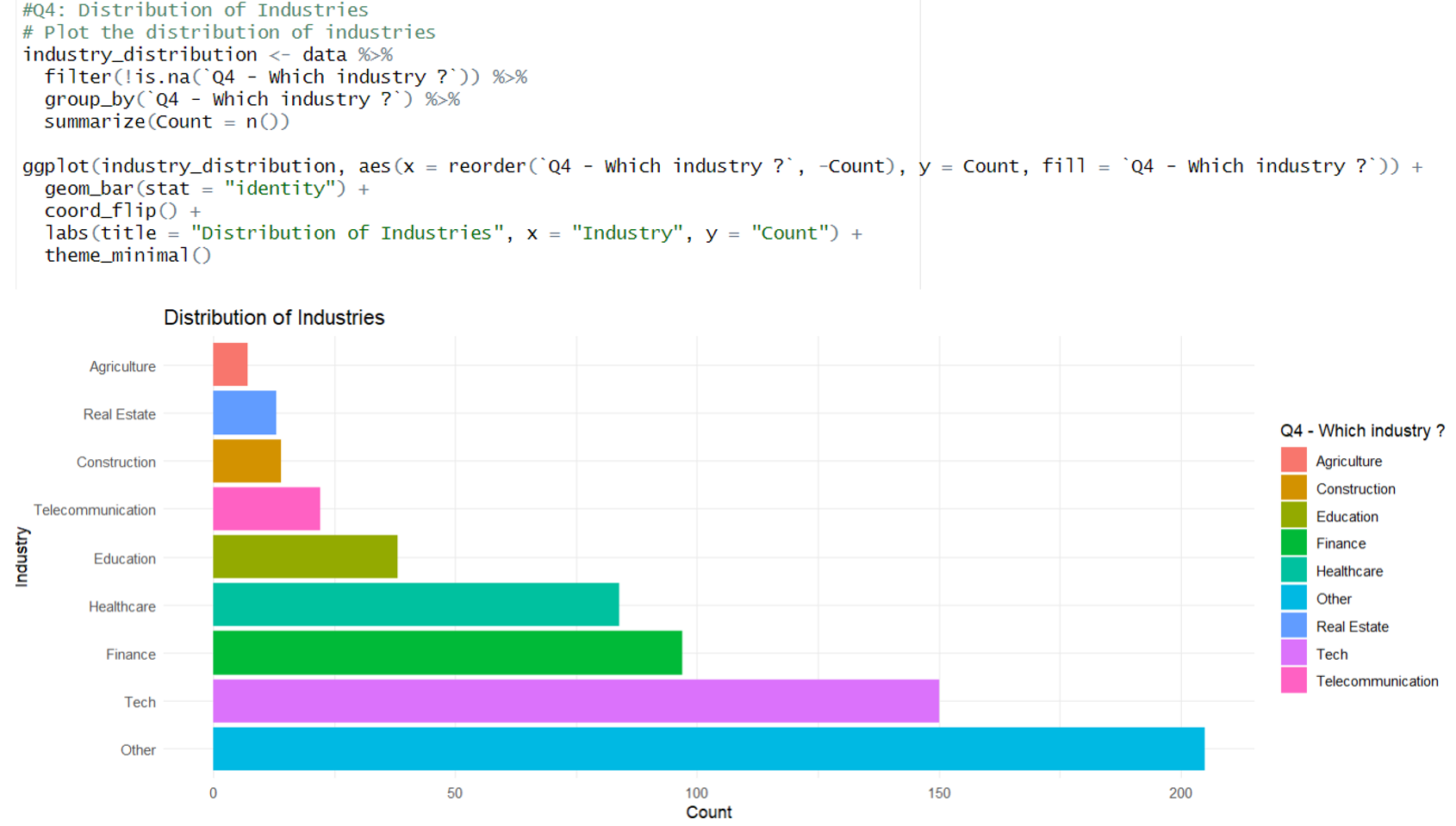
1. **Common Programming Languages**



* Python dominates the market with overwhelming usage, thanks to its rich ecosystem for data science and machine learning
* SQL ranks second, underscoring the importance of relational data processing and querying skills
* R remains a key player important, especially in statistics and academic research
* JavaScript is starting to appear more, reflecting the trend of data visualization and web applications
* Scala and Java are used mainly in big data projects
* This trend affects training programs and recruitment requirements in the industry

=> The chart shows Python's dominance in the data space, but other languages ​​like SQL, R, JavaScript, Scala, and Java still have important roles to play. Data professionals need to master the right languages ​​to meet market demands and project requirements.

1. **Distribution of Industries**



· Technology industry has the largest proportion, reflecting the pioneering role in data analysis applications

· Finance-Banking is the second industry, showing the importance of data in risk management and investment

· The diversity of industries shows the trend of digitization and dataization is spreading

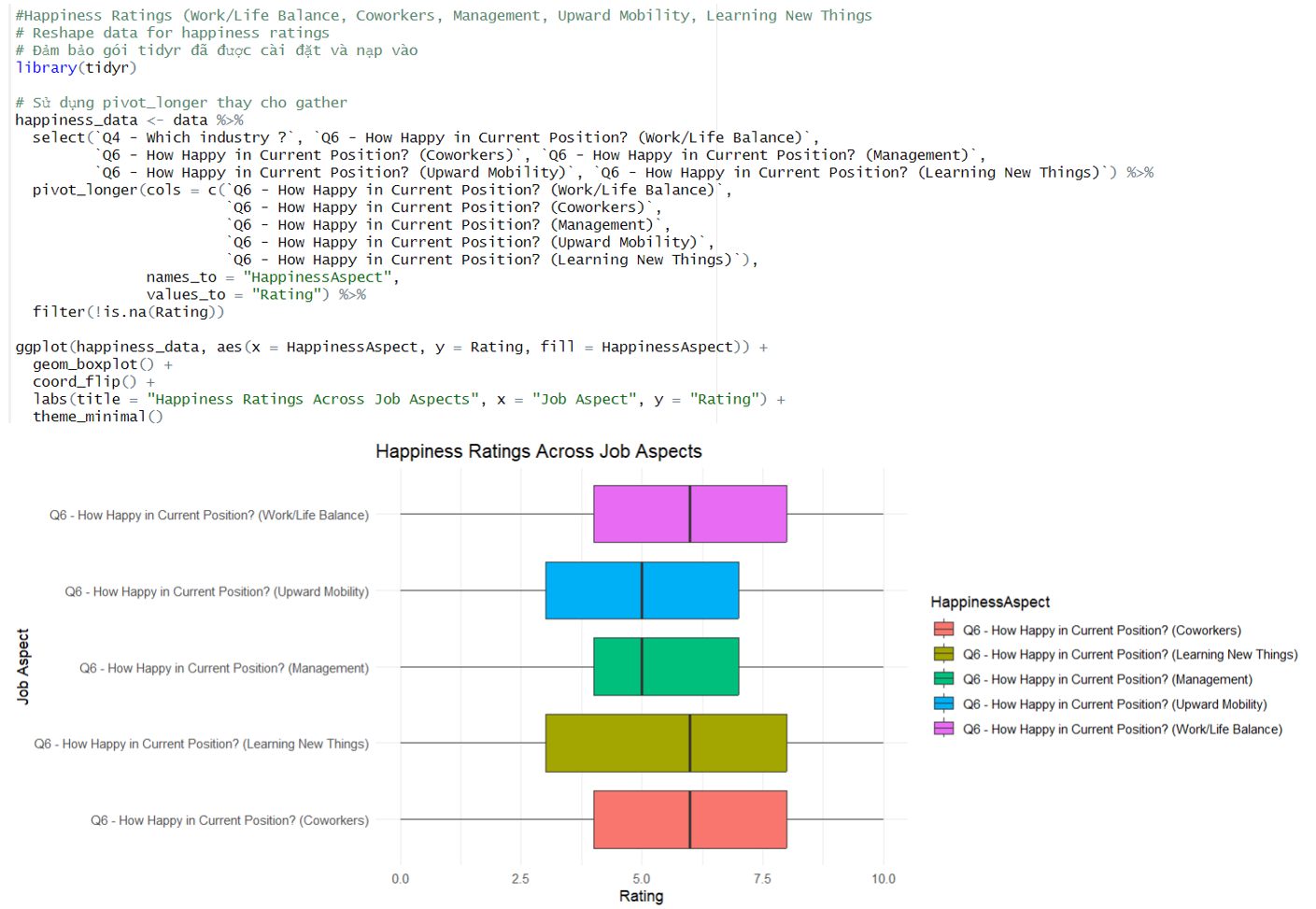
· Healthcare and Retail are on the rise, reflecting the need for data analysis in service improvement

· Consulting industry has a significant proportion, showing the trend of outsourcing data analysis

· This distribution may change in the future as traditional industries accelerate digital transformation

=> The chart shows the diversity of industries using data, with technology and banking and finance leading the way. Traditional industries are gradually transforming digitally and increasing their use of data. This distribution may change in the future due to technological developments and market demands.

1. **Happiness Ratings (Work/Life Balance, Coworkers, Management, Upward Mobility, Learning New Things)**



· Work/Life Balance is rated highest, indicating that the industry has a good policy on work-life balance

· Positive relationships with colleagues, reflecting a professional and cooperative working environment

· Learning opportunities are rated quite high, consistent with the characteristics of an ever-evolving industry

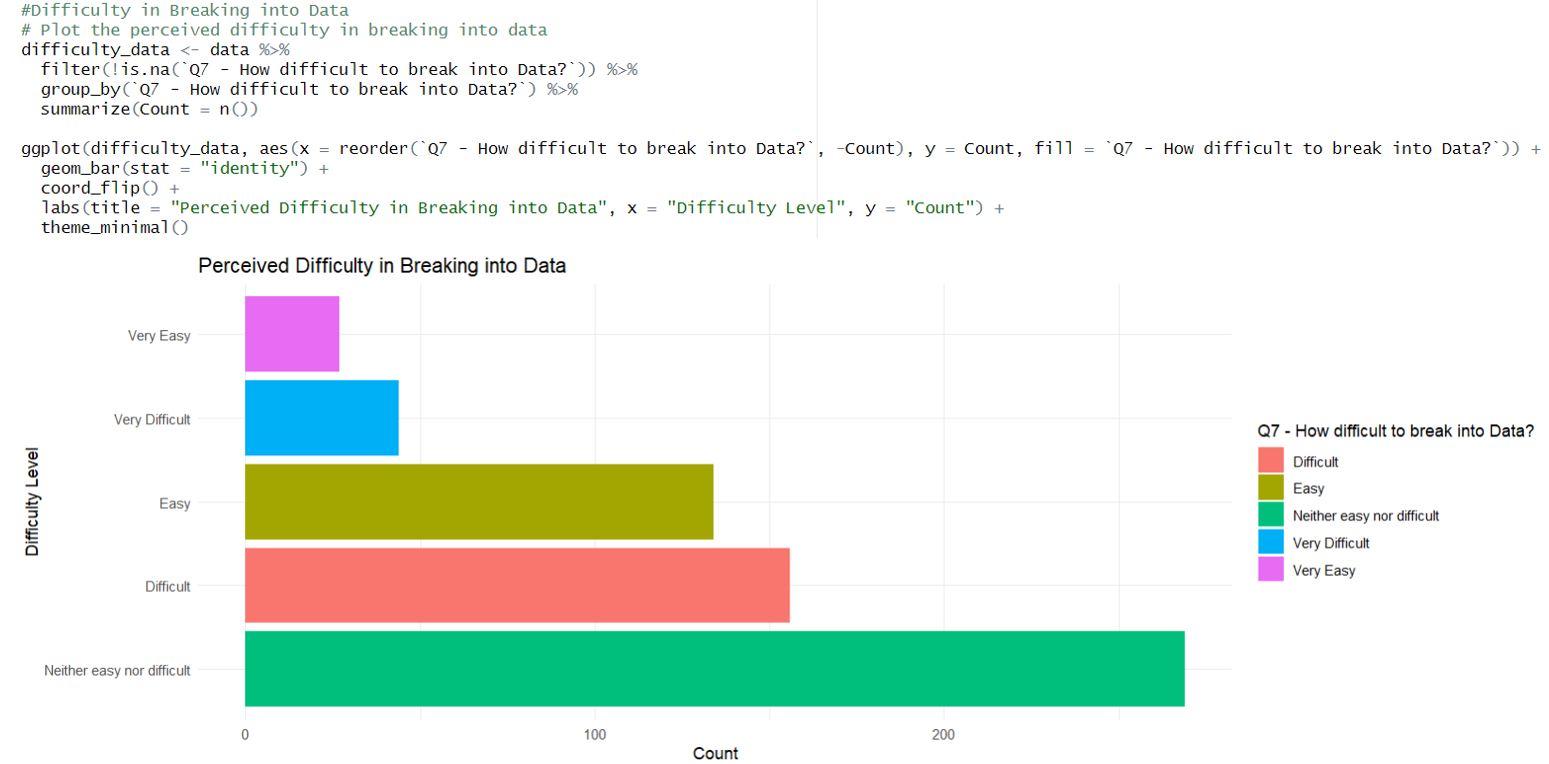
· Management is rated at an average level, possibly needing improvement in leadership skills in the technical field

· Promotion opportunities have the lowest score, indicating the need for a clearer career development roadmap

· Overall, satisfaction levels are quite positive, contributing to retaining talent in the industry

=> Employee satisfaction ratings show that the data industry has a positive work environment and good growth opportunities. However, management needs to be improved and career paths need to be clearer to increase satisfaction and retain talent.

*Difficulty in Breaking into Data*



· Most rated at medium to high difficulty, reflecting barriers to entry

· High requirements for technical skills and mathematical background are a big challenge

· High competition due to many people switching to data

· Difficulty in accumulating practical experience when starting out

· Degree and certification requirements can be a barrier for newcomers

· Need to have a strategy for skill development and networking effective to enter the industry

=> The chart shows the difficulty of entering the data industry, due to the skill requirements, high competition, and difficulty in gaining experience. Those who want to enter the industry need to have an effective skills development and networking strategy.

1. **Key Factors When Looking for a New Job**



· Salary and benefits are the top factor, reflecting the expectation of high income in the industry

· Work/Life Balance is the second priority, showing the lifestyle trend of the new generation

· Learning and development opportunities are valued due to the constantly updated nature of the industry

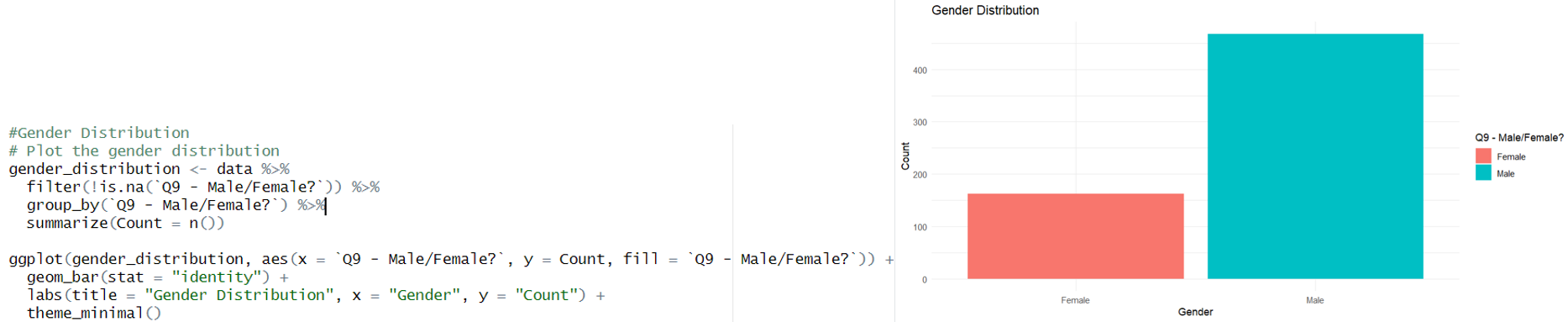
· Company culture and working environment are also important factors

· The ability to work remotely/hybrid is more concerned after the pandemic

· Technology and tools used also influence the decision

=> The chart shows important factors when choosing a company in the data industry, including salary and benefits, work-life balance, learning and development opportunities, company culture, remote work capabilities, and technology used. Companies need to meet these factors to attract and retain talent.

1. **Gender Distribution**



· The industry is overwhelmingly male, reflecting a serious gender imbalance

· The low proportion of women may be due to social prejudice and barriers in STEM education

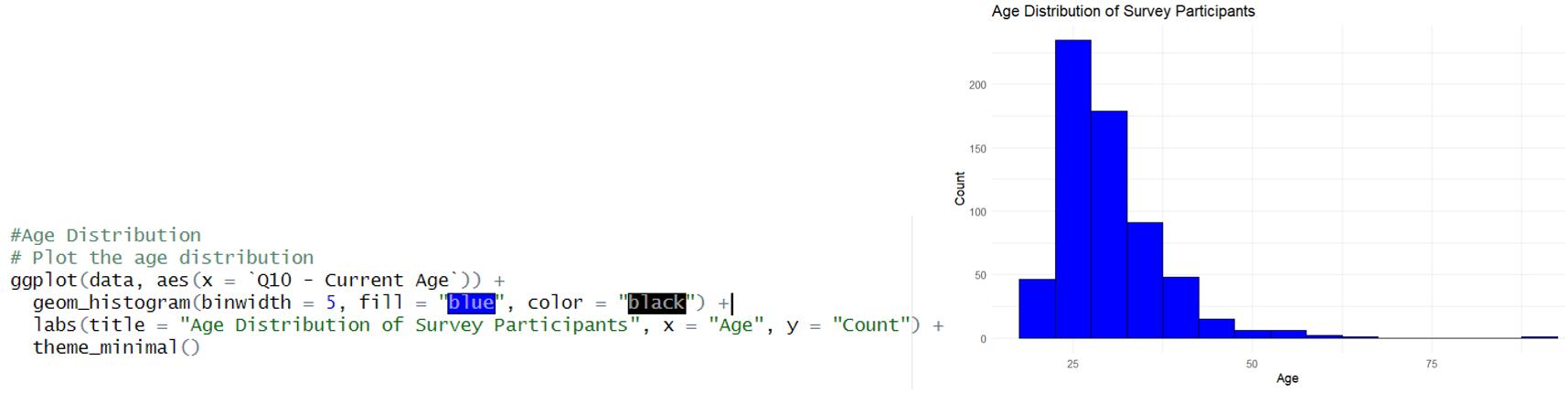
· Programs and policies are needed to attract women into the industry

· Gender diversity issues may affect creativity and team efficiency

· This trend is improving but more efforts are needed

· The The company is taking initiatives to increase the proportion of women in data roles

=> The chart shows a serious gender imbalance in the data industry. Policies and initiatives are needed to attract women to the industry and increase gender diversity. Gender diversity can bring many benefits to the industry, such as increased creativity and team effectiveness.

1. **Age Distribution**  
   

· The 25-34 age group is the majority, indicating that this is a young industry

· The low proportion of the over-45 age group may be due to the industry being relatively new

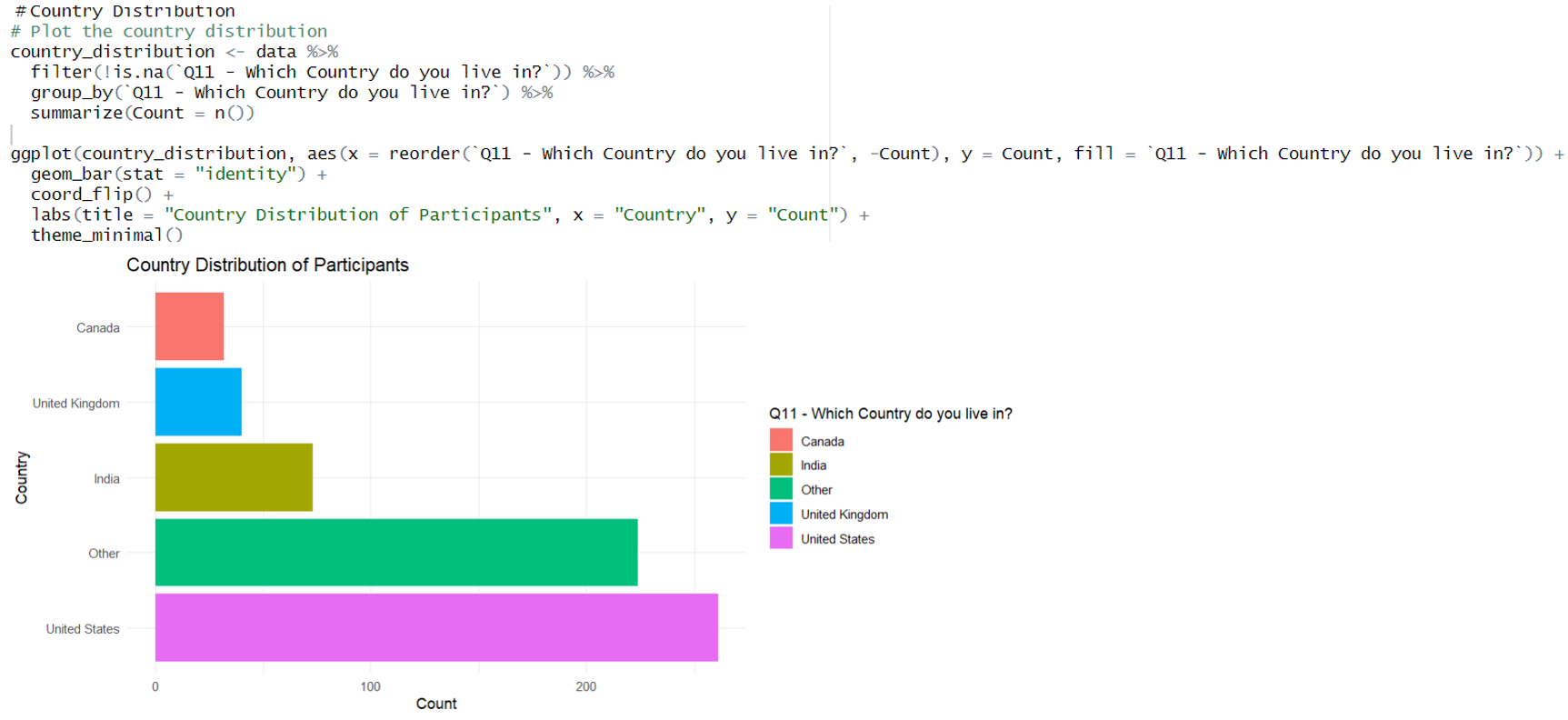
· There may be issues with age discrimination in recruitment

· Age imbalance may affect the integration of experience

· There is a need for policies to support older people in career transitions

This trend may change as the industry matures

=> The chart shows an age imbalance in the data industry, with younger generations in the majority. Supportive policies are needed to help older adults transition successfully and integrate experience. The age distribution may change as the industry matures.

1. **Country Distribution**

· The US leads in the number of professionals, reflecting a mature technology market

· India ranks second, reflecting the strong growth of the outsourcing industry

· Geographic diversity but mainly concentrated in developed countries

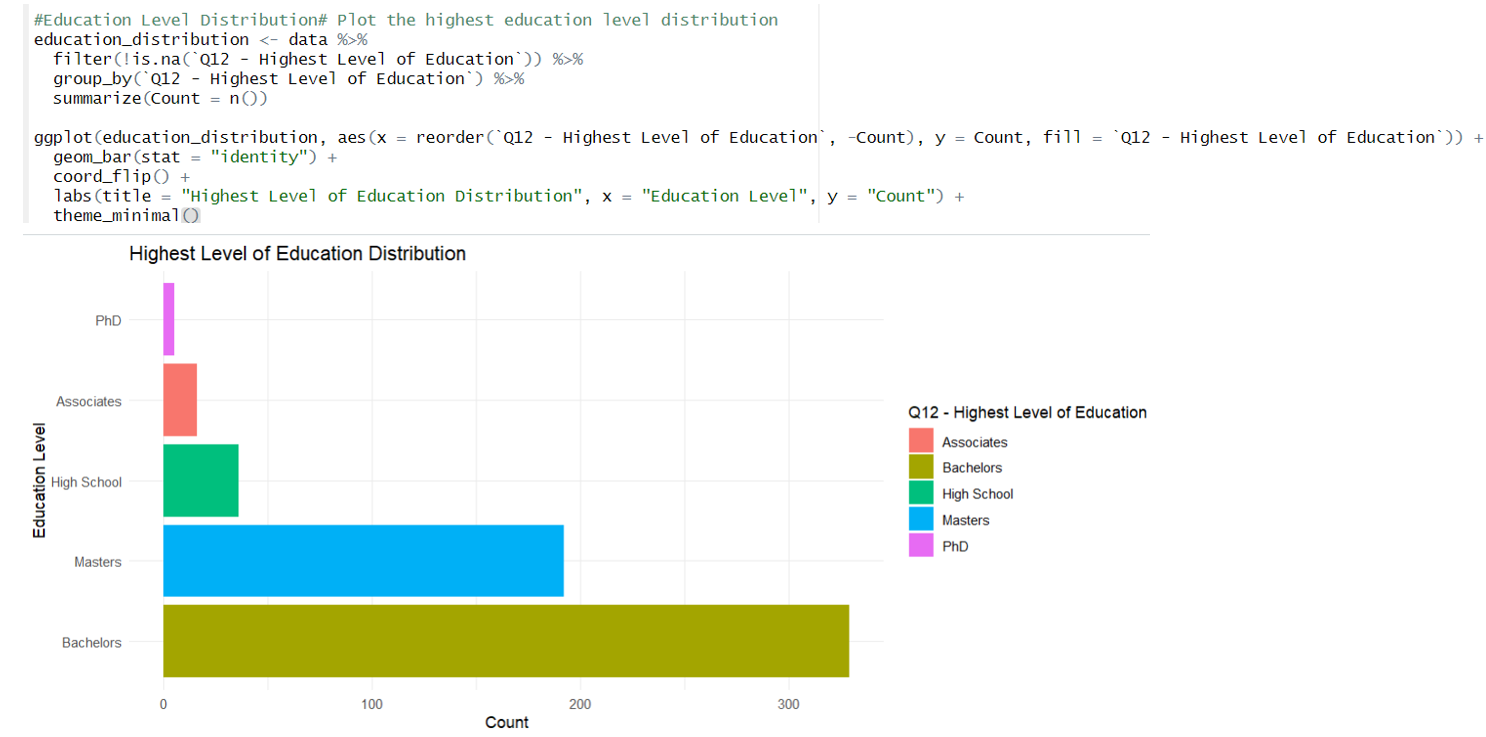
· Asia is emerging as a potential market

· Regional disparities may create opportunities for remote work

· This trend affects companies’ global recruiting strategies

=> The chart shows the uneven distribution of data professionals globally, with the US and India leading the way. Other regions such as Asia are emerging as potential markets. Regional differences could create opportunities for remote work and influence companies’ global hiring strategies.

1. **Education Level Distribution**



· Masters are the most prevalent, reflecting the high educational demands of the industry

· Significant proportion with Bachelors degrees suggests opportunities for those with less education

· Small numbers with PhDs but often in research positions

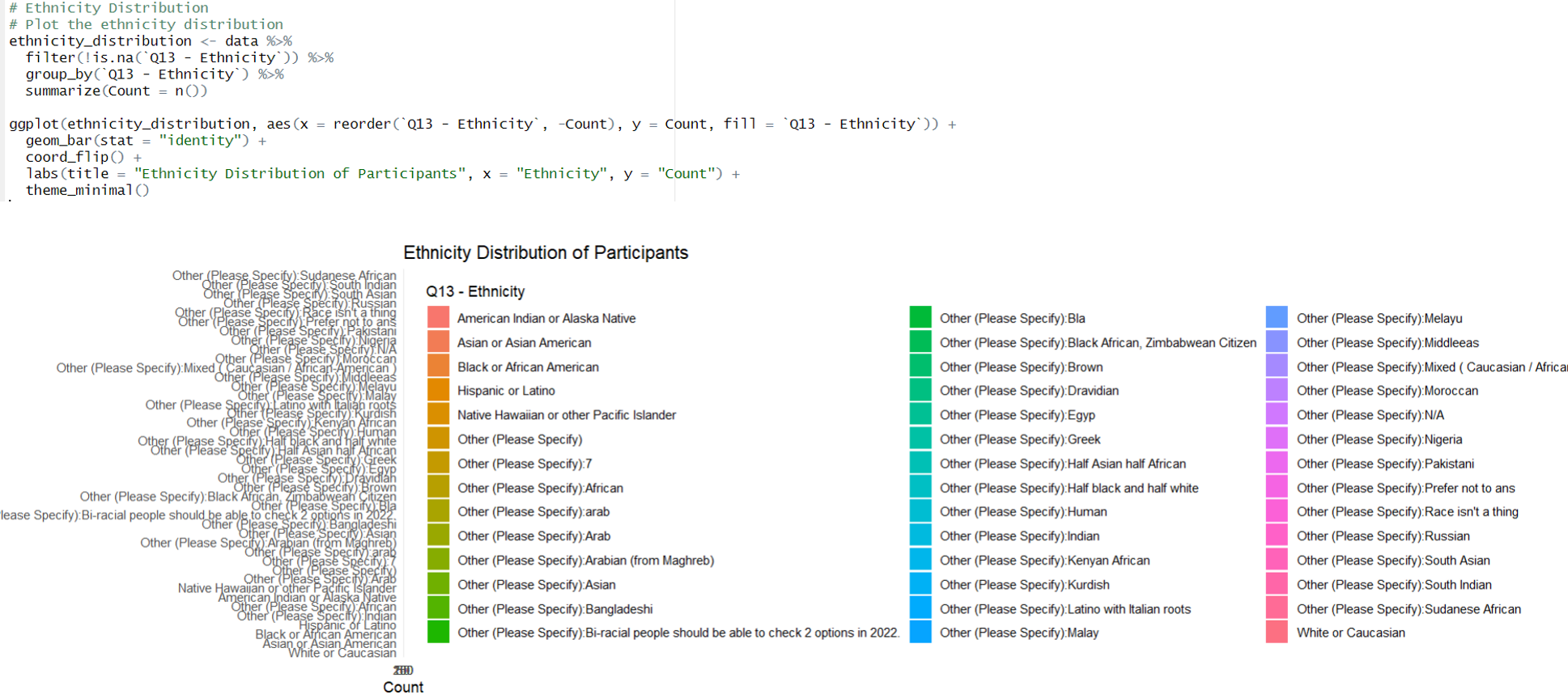
· This trend impacts the cost and time investment to enter the industry

· There is a difference in educational requirements between different roles

· Bootcamps and self-learning are emerging as alternatives to traditional education

=> The chart shows that the education requirements for the data industry are relatively high, with a master’s degree being the most common. However, bootcamps and self-paced programs are emerging as alternatives to traditional education, opening up many opportunities for those without a master’s or doctorate.

1. **Ethnicity Distribution**



· Whites are the most prevalent, reflecting a diversity issue

· Asians are significantly overrepresented, particularly in technical roles

· Low proportions of other minorities reflect inequalities in access

· Need for training and recruitment programs targeting minorities

· Ethnic diversity issues are a concern for large companies

· This trend could improve with policies DEI (Diversity, Equity, Inclusion)

=> The chart shows the racial imbalance in the data industry. Policies and initiatives are needed to improve diversity and equity. Racial diversity can bring many benefits to the industry, such as increased creativity and team effectiveness.

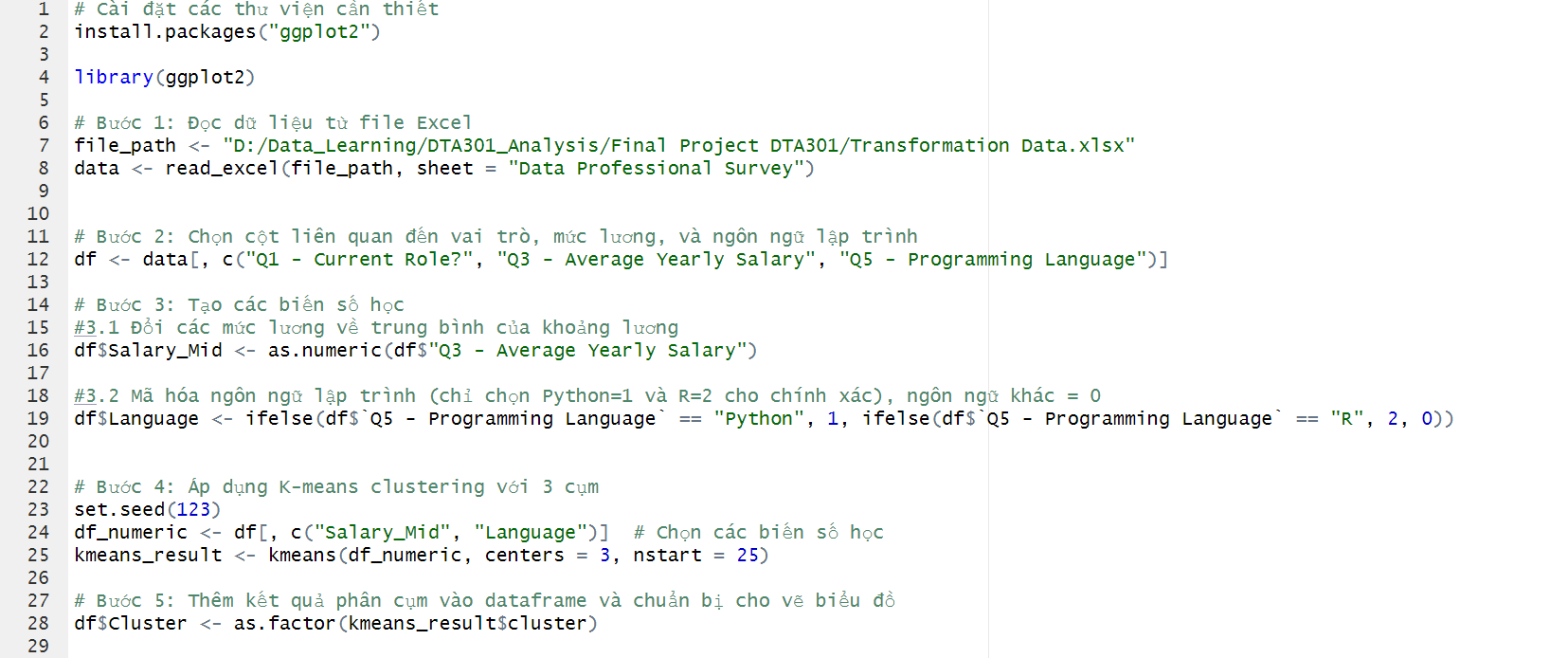
## 

## B. IN-DEPTH ANALYSIS USING ADVANCED ALGORITHMS AND METHODS.

### **Clustering Algorithm (k-means)**

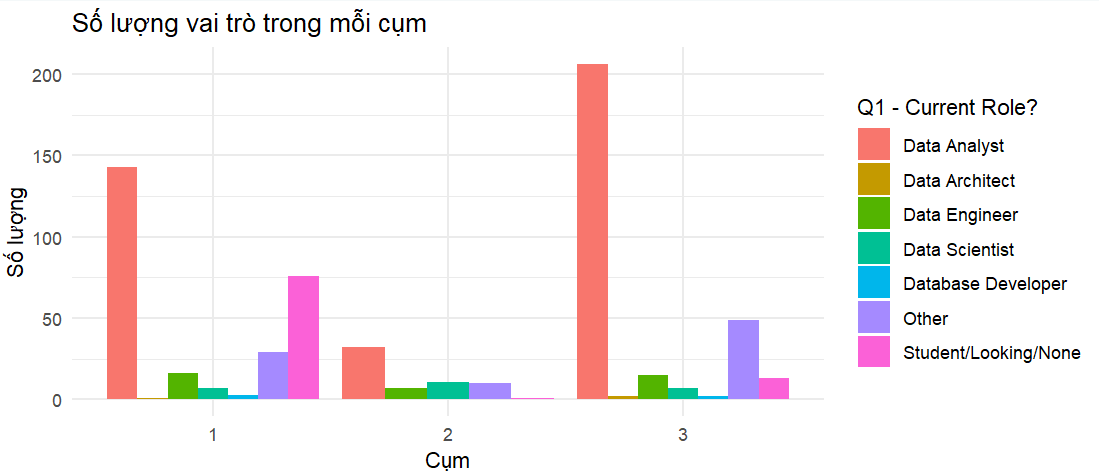
K-means is a clustering algorithm used to divide a set of data into *k* groups (clusters) based on the features of the data. The main goal of the algorithm is to partition the data points into clusters such that points within the same cluster are more similar to each other than to those in other clusters.

**Application: Clustering job roles based on average salary   
 and programming languages.**

****

1. **Visualization using a Bar Plot.**

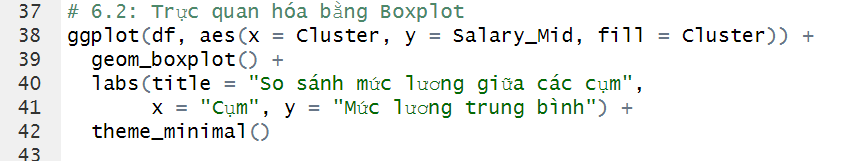
****

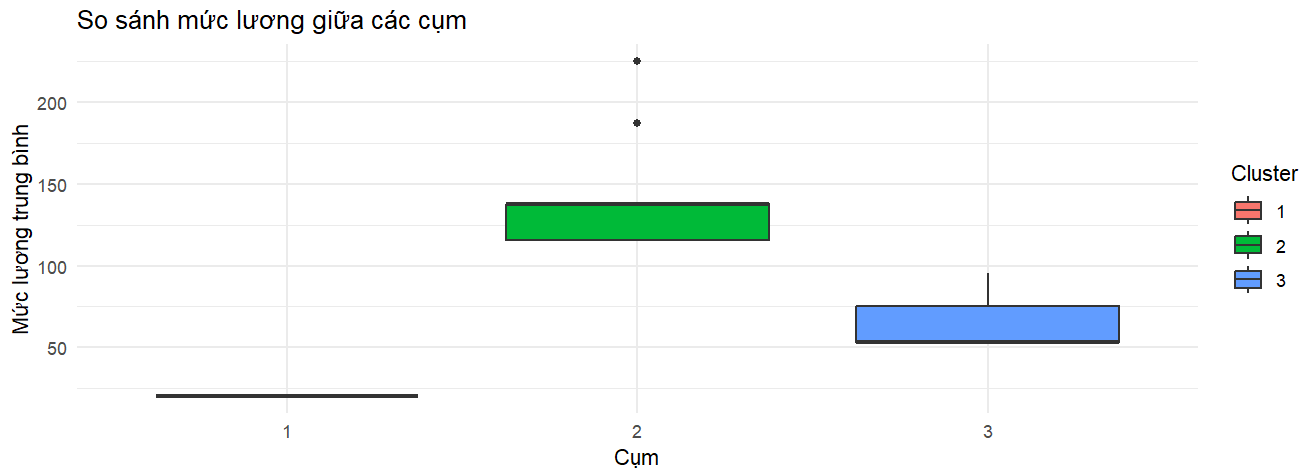
****

**Insight:**

**Cluster 1** and **Cluster 3** primarily consist of Data Analyst roles, with more than 150 people in each cluster. **Cluster 2** has a more balanced distribution of roles, with a significant number of roles such as Student/Looking/None, Data Engineer, and Data Architect.

1. **Visualization using a Box Plot.**

****

****

**Insight:**

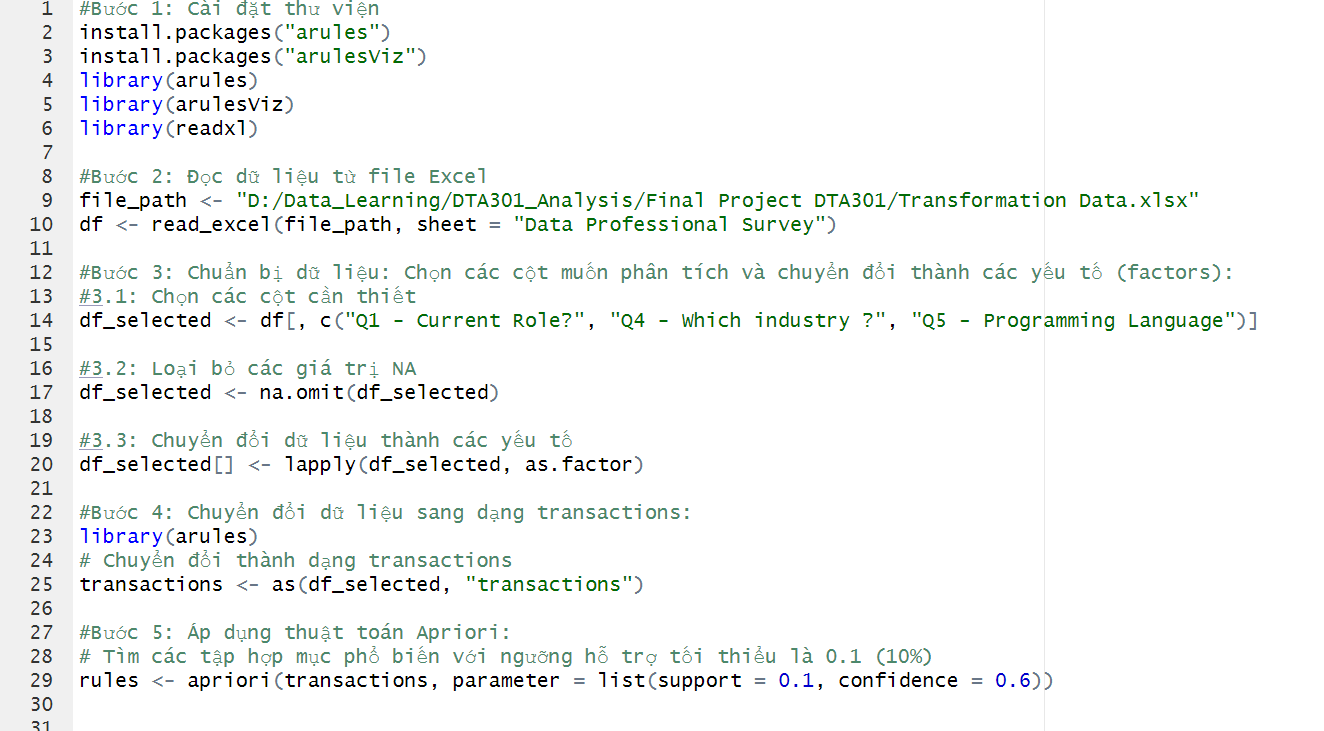
**Cluster 1** has the lowest average salary, with a narrow salary range and a very low mean value, almost showing no variation. Meanwhile, **Cluster 2** has a higher average salary and a wider salary range. **Cluster 3** has an average salary lower than **Cluster 2** but higher than **Cluster 1**, and also has a relatively narrow salary range.

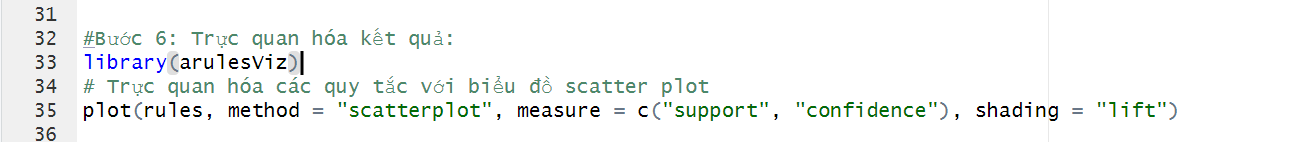
## 2. **Association Rules (Apriori)**

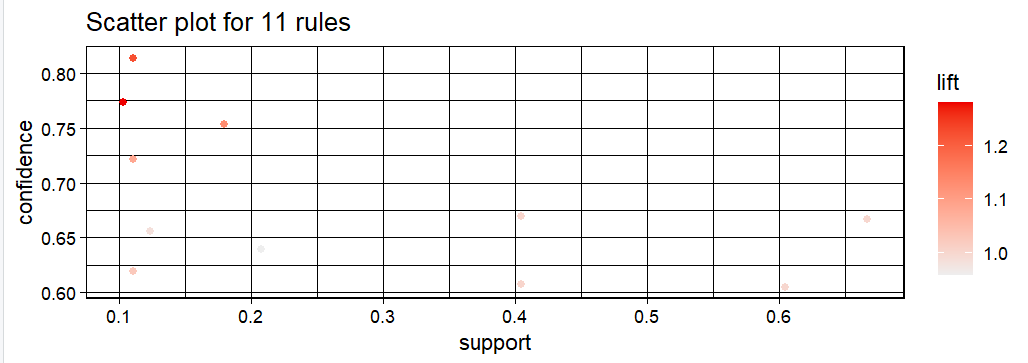
The main purpose of **Association Rule Analysis** is to discover relationships or hidden patterns between attributes in the data.

**Application:**

**Analyzing the association of three attributes - Discovering rules or relationships such as "People working in industry A with role B often use programming language C."**

****

****

****

**Comments from the chart:**

1. **X-axis - Support:** Represents the support, which is the frequency of occurrence of item sets in the data. From the chart, it can be seen that the rules have support values ranging from 0.1 to 0.7. This means that the frequency of occurrence of these item sets is between 10% and 70% in the entire dataset.
2. **Y-axis - Confidence:** Represents the confidence, which indicates the probability that the right-hand side of the rule will occur when the left-hand side occurs. The rules here have confidence values from about 0.6 to 0.8, indicating that these rules have a relatively high level of confidence (from 60% to 80%).
3. **Color - Lift:** The hue (color) of the points represents the lift value of each rule. Darker colors indicate higher lift values. In this chart, the lift values range from 1.0 to 1.2. A lift value greater than 1 indicates a positive (or significant) relationship between the item sets in the rule.

**Insights:**

1. **Rules with low support but high confidence:** Some rules on the left side of the chart have low support values (around 0.1) but high confidence values (around 0.75-0.80). This means that these rules, while infrequent, have high reliability, indicating that when one item set occurs, it is likely that the other item set will also appear.
2. **Lift not too high:** The lift values range from 1.0 to 1.2, indicating that these association rules are not very strong. This may suggest that the relationships between the factors are not yet significant, or these factors do not strongly influence each other.
3. **Dispersion of rules:** There are a few rules concentrated in the area of low support (around 0.1 - 0.2) but with high confidence levels (above 0.75). These rules are often important because, despite being rare, they are highly accurate when they do occur.

### **3. Linear Regression Model**

Linear Regression is a statistical method used to model the relationship between a **dependent** variable (target variable) and one or more **independent** variables (predictor factors). Linear regression aims to find the best-fitting straight line (regression line) that describes the relationship between these variables.

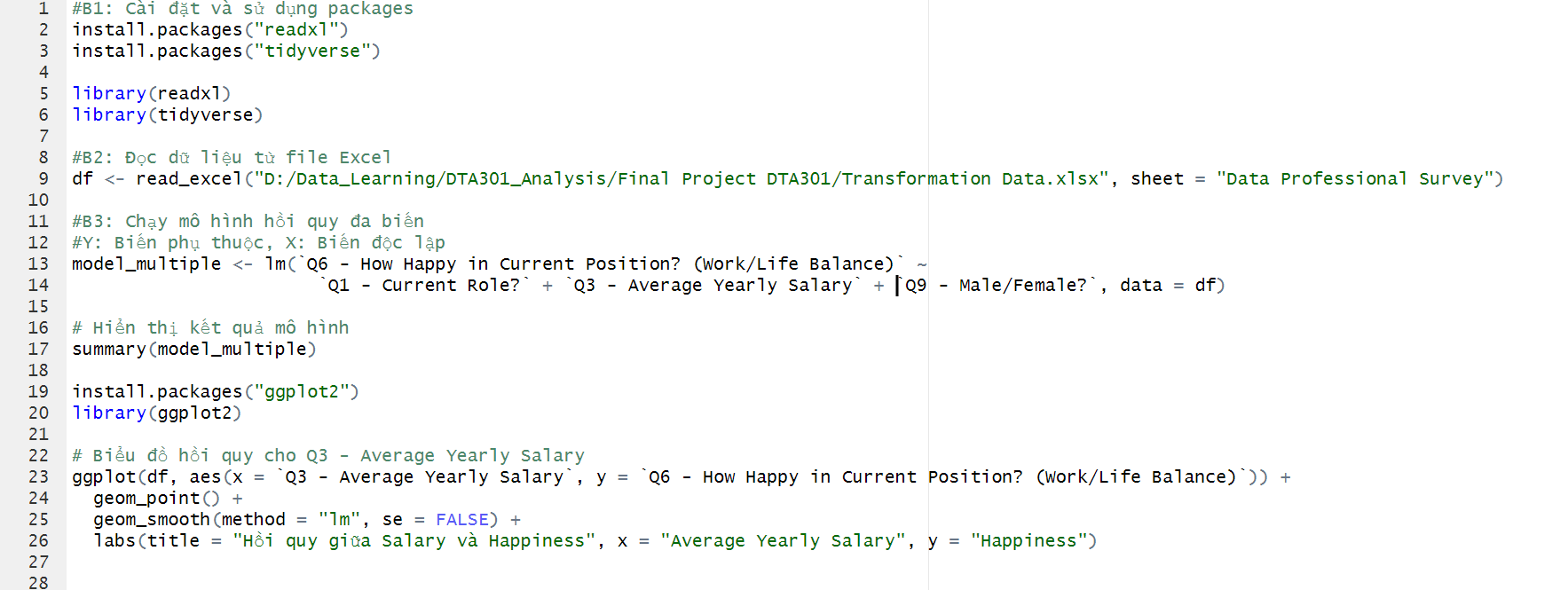
**Dependent Variable:**

* **Job Satisfaction Level:** This is an important target variable that you may want to predict based on other factors. If this variable is measured on a scale (e.g., from 1 to 10), it can be considered a continuous variable.

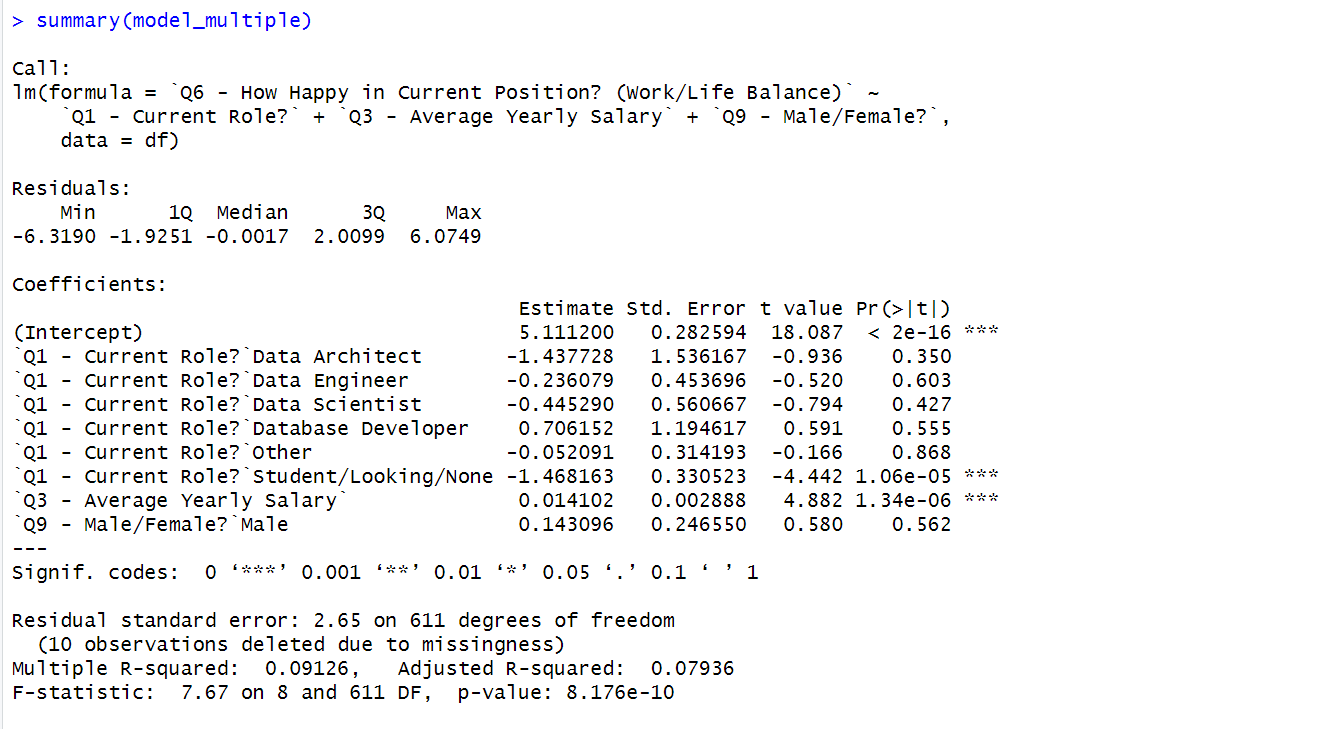
**Independent Variables:**

* **Job Role:** Encoding job roles into categorical variables.
* **Industry:** Transforming industries into categorical variables
* **Years of Experience:** If applicable, this can be a continuous variable for analysis.
* **Programming Language:** You can encode programming languages to examine whether they influence job satisfaction levels.

**Application:   
Utilize multiple independent variables to assess their combined impact on the dependent variable. This helps identify the factors that strongly affect the target.**

****

**Model summary**

****

**Comments:**

1. **Regression Coefficients:**

**Intercept:** 5.1112, with high statistical significance (p < 2e-16).

**Q1 - Current Role?**

**SStudent/Looking/None: Coefficient -1.4682, statistically significant (p < 0.001).**

**Q3 - Average Yearly Salary:** Coefficient 0.0141, statistically significant (p < 0.001).

1. **Model Fit:**

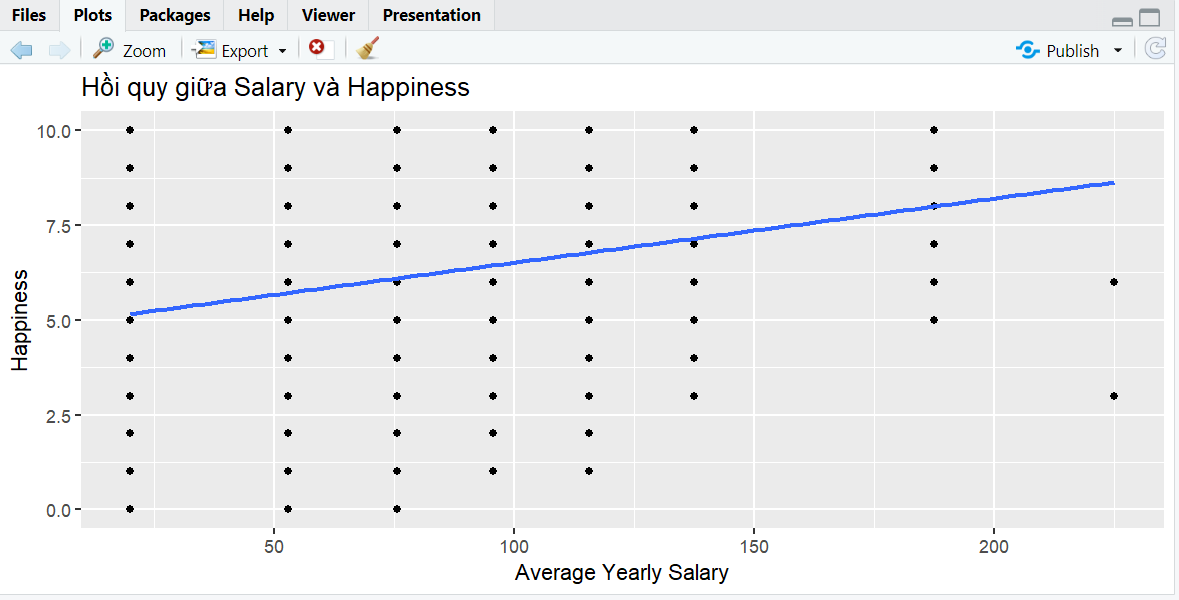
**Multiple R-squared:** 0.09126, indicating that the model explains approximately 9.1% of the variability in the dependent variable.

**Adjusted R-squared:** 0.07936, adjusted for the number of independent variables in the model.

1. **Overall Statistical Significance:**

**F-statistic:** 7.67 with a p-value of 8.176e-10, indicating that the model is statistically significant overall.

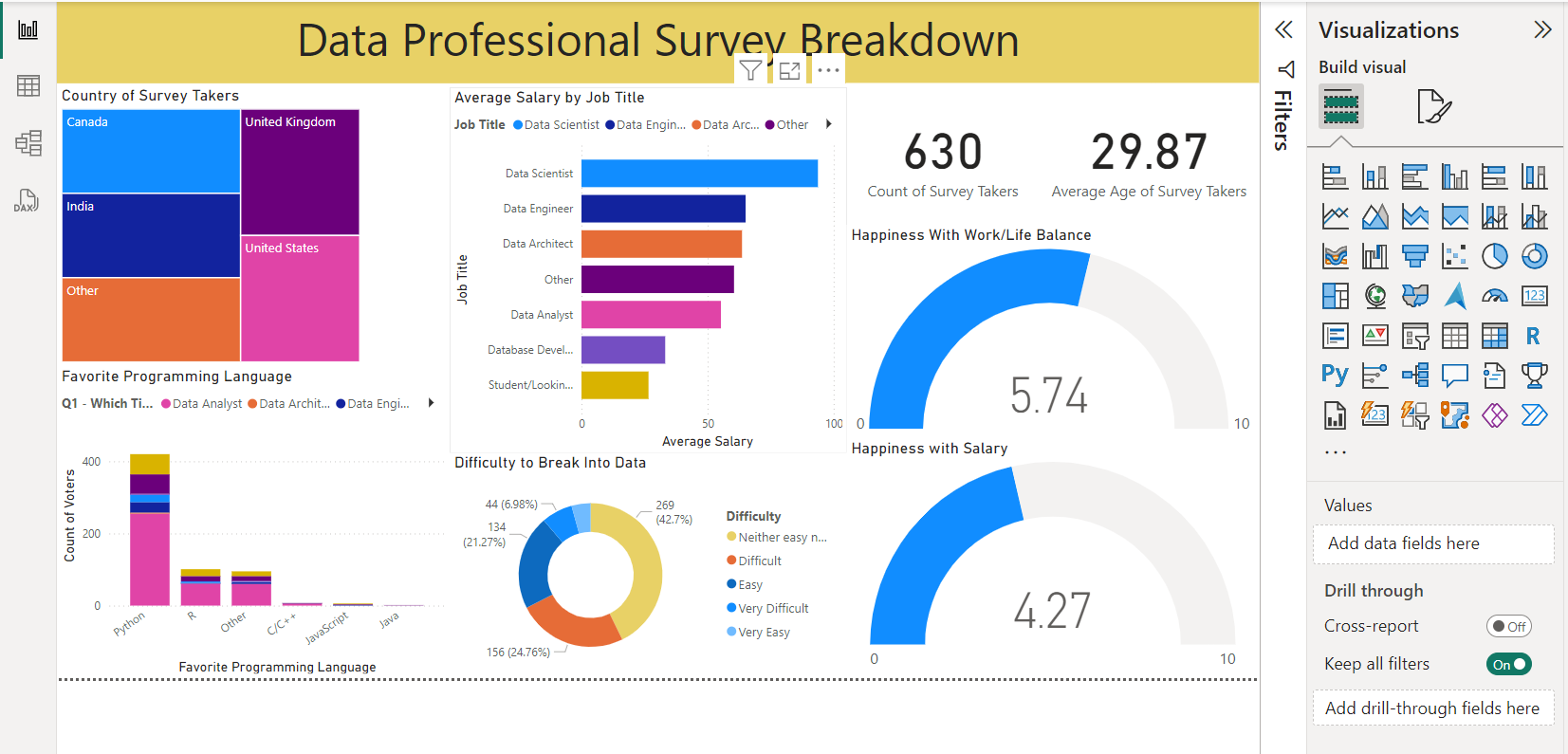
=> The model shows that several independent variables significantly affect job satisfaction, particularly the current role and average yearly salary.



**Positive Relationship:** The chart shows a positive relationship between average yearly salary and job satisfaction levels. This means that as salary increases, job satisfaction also tends to increase.

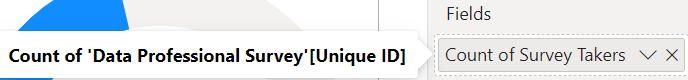
# **V - VISUALIZATION WITH POWER-BI**

Power BI provides interactive data visualization capabilities, enabling users to easily analyze, make data-driven decisions, and collaborate effectively through customized reports and dashboards.

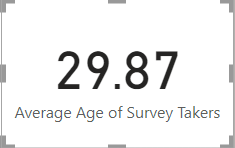


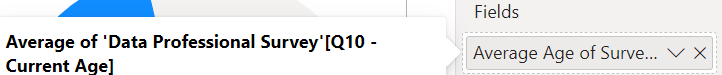
Go in depth to each visualization:

* **Total number of survey takers** (How many people took this survey ?)

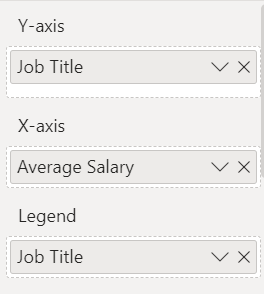
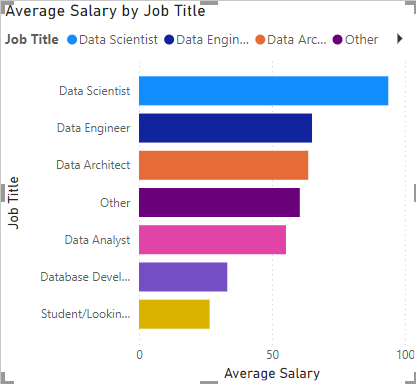


* **Average age of survey takers**

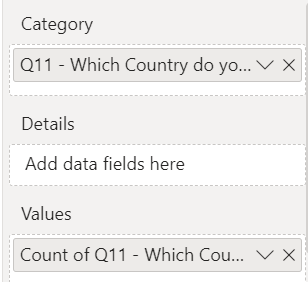
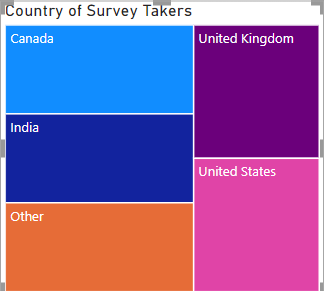




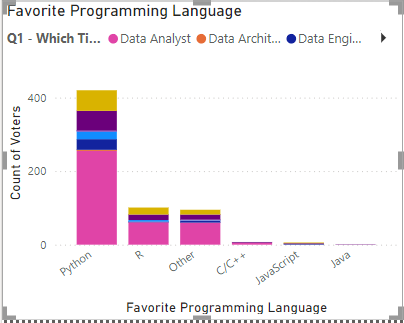
* **Average salary by Job title**

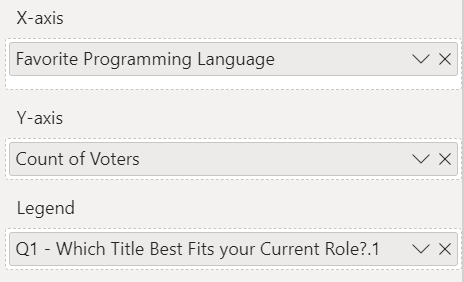


* **Country of survey takers**

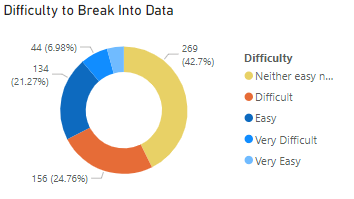


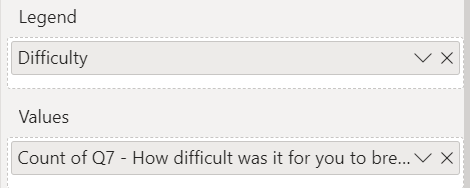
* **Favorite programming Language**



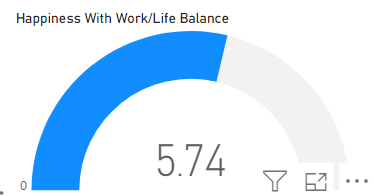


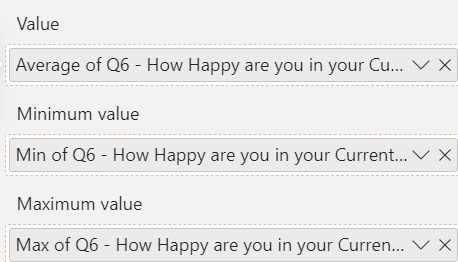
* **Difficulty to break into data profession**





* **Happiness with Work/Life Balance**





* **Happiness with Salary**

