Fpt University

TRƯỜNG ĐẠI HỌC FPT

Project Report

In-depth Analysis of Data Professions

Subject: DTA301 - Data Analysis

Lecturer: TienNQ27 - Nguyễn Quốc Tiến

Class: IS1803

GROUP 6		
Group members	ĐINH GIA BẢO	SE183741
	NGUYỄN VŨ HIẾU	SE184611
	PHẠM NGỌC DUY MINH	SE180043

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 - **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:

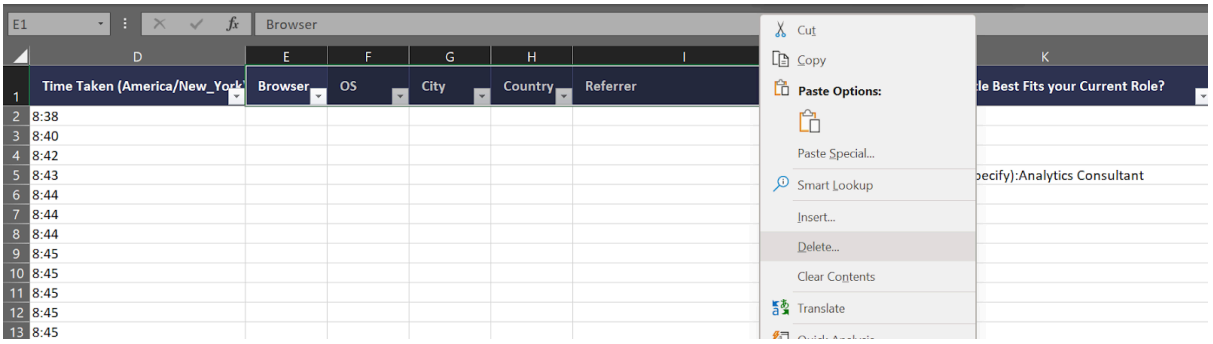
- R programming language: Used for data processing, statistical analysis, and visualization.
- R packages: Dplyr, ggplot2, reshape2 were utilized for data manipulation and visualization tasks.

- 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.

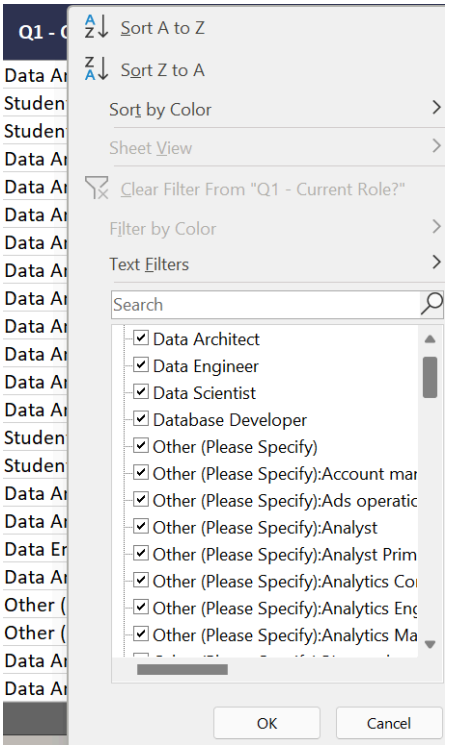
III - Data Transformation

Delete columns whose data was empty

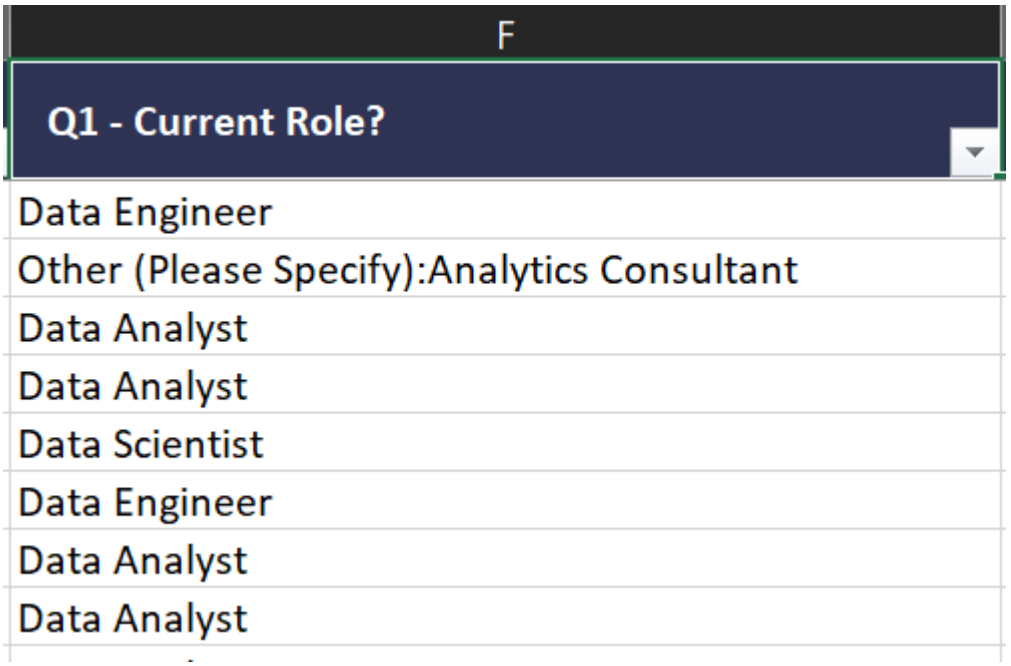


Split Other in to a specific column

Q1 - Current Role?
Data Analyst
Student/Looking/None
Student/Looking/None
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Data Analyst
Student/Looking/None
Student/Looking/None
Data Analyst
Data Analyst
Data Engineer
Data Analyst
Other (Please Specify):FP&A Analyst
Other (Please Specify):BI Developer
Data Analyst
Data Analyst

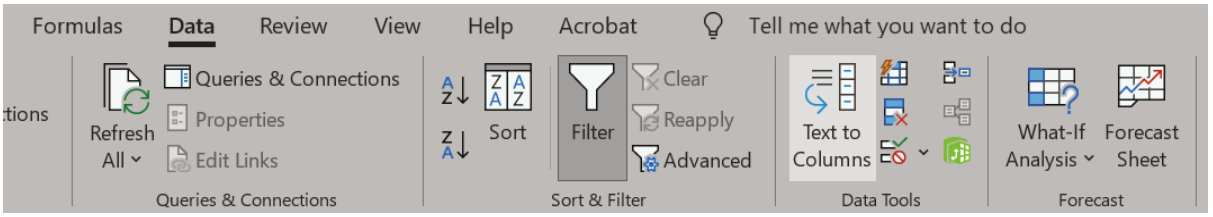


Choose a column

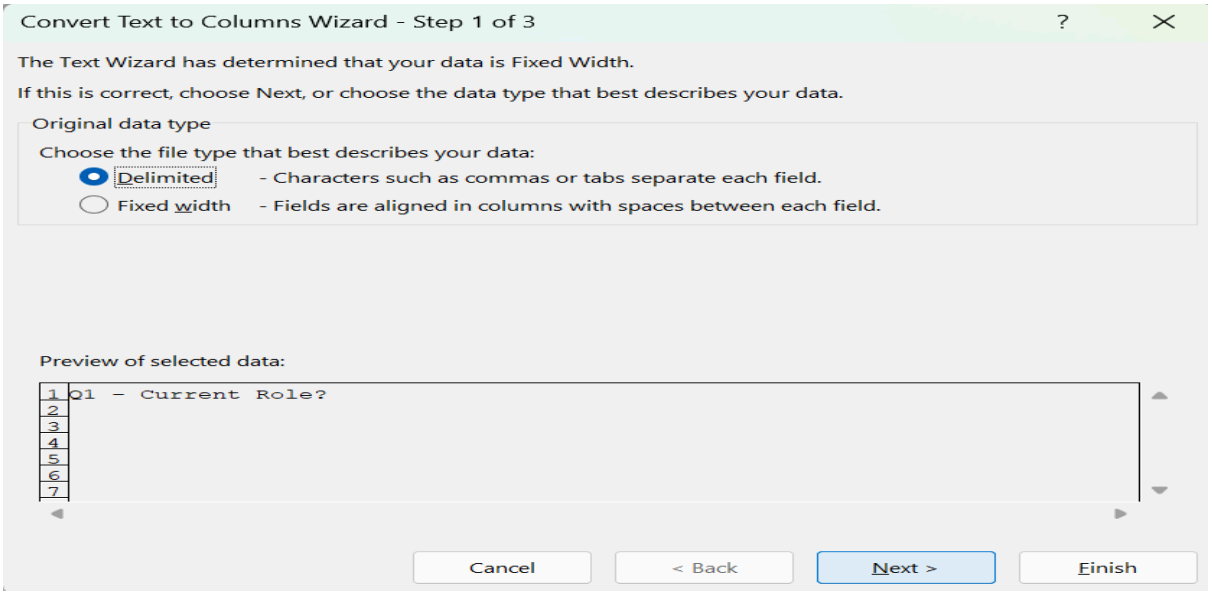


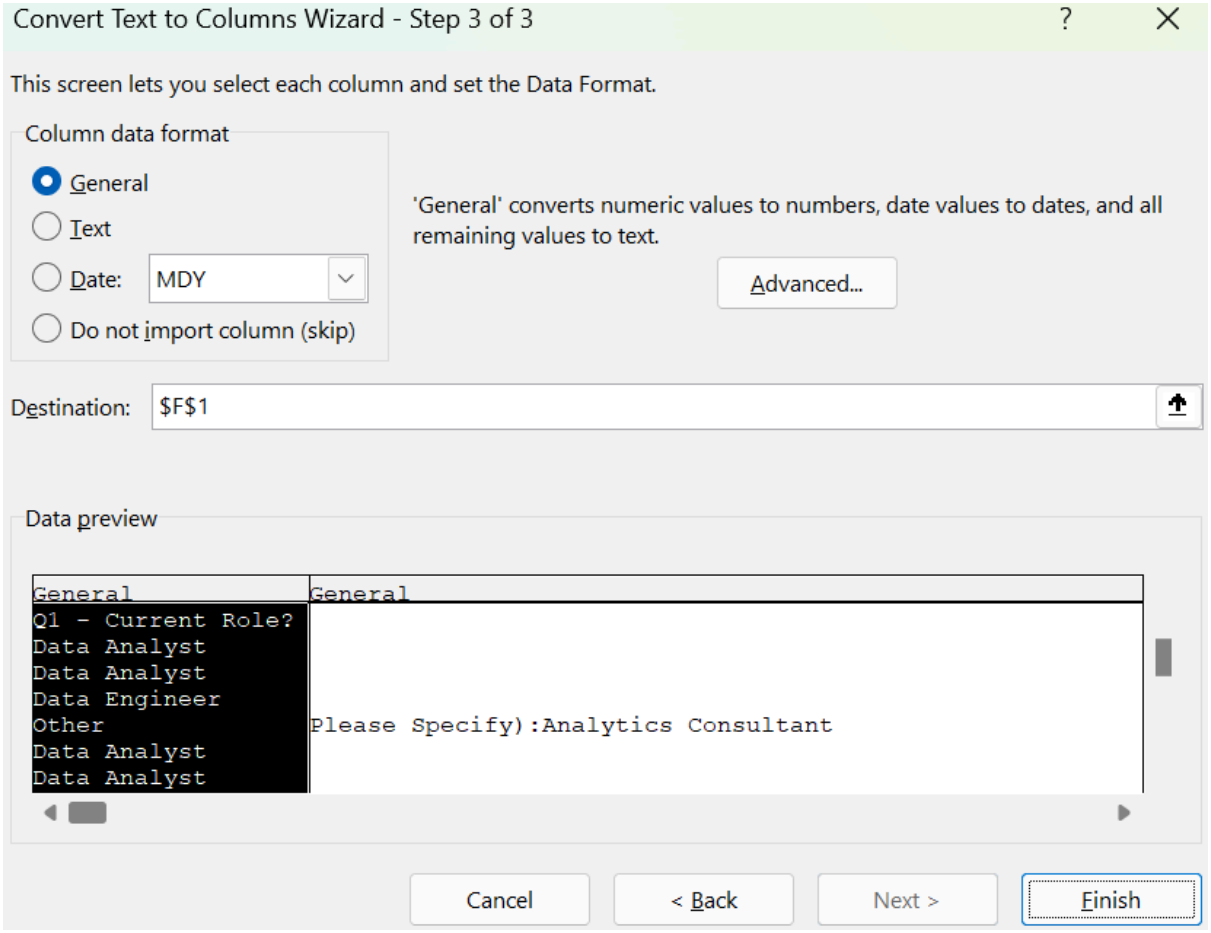
F
Q1 - Current Role?
Data Engineer
Other (Please Specify):Analytics Consultant
Data Analyst
Data Analyst
Data Scientist
Data Engineer
Data Analyst
Data Analyst

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

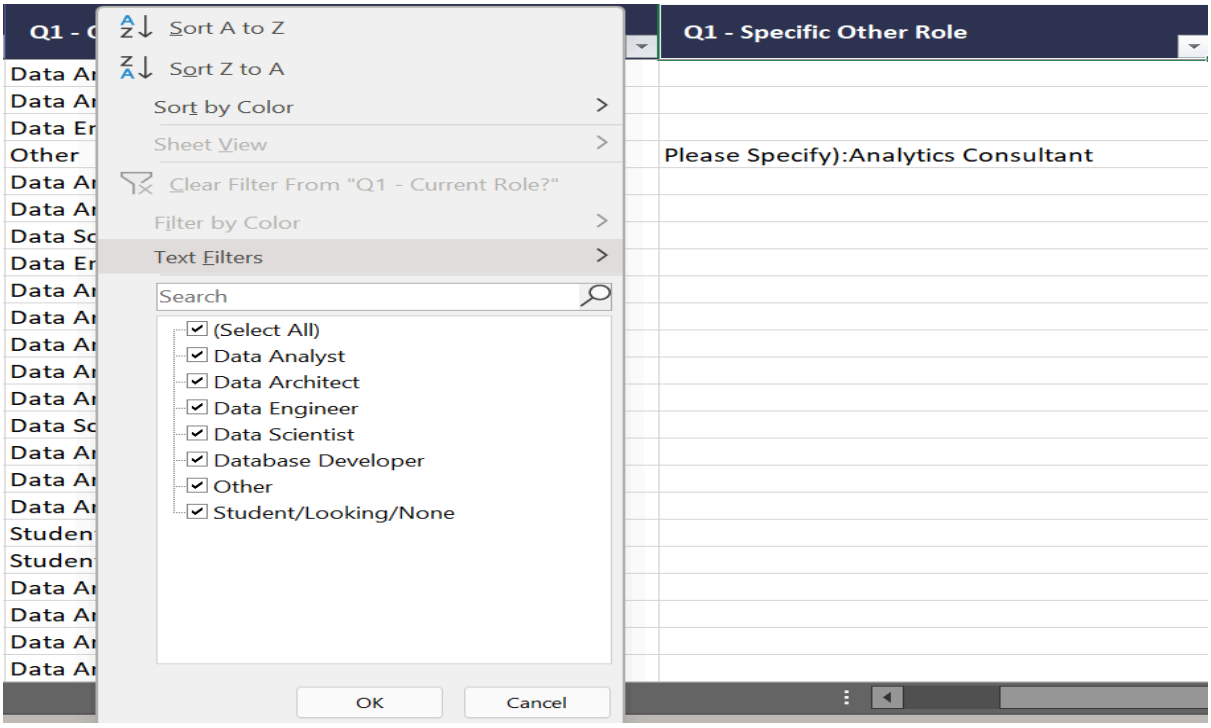


Choose delimited and Next





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

Q4 - Which industry ?
Finance
Other (Please Specify):Advertising
Tech
Finance
Other (Please Specify):Biotech
Other (Please Specify):Consulting
Other (Please Specify):Consumer Elec
Other (Please Specify):Consulting
Tech
Other (Please Specify):Semiconductor manufacturing
Healthcare
Other (Please Specify):Supply Chain - warehousing, transpiration and
Finance
Finance
Tech
Tech
Finance
Other (Please Specify):Distribution
Other (Please Specify)
Finance
Tech
Finance
Other (Please Specify):Customer Service

Q5 - Programming Language
Python
R
Python
Python
Other:Mostly use sql but that’s not progr
Python
Python
Python
Python
Other:SQL
Python
Other:Qlik sense script
R
Python
Python
Python
Other:SQL
Python
Python
Python
Other:sql
Python
Python

Q8 - The key when looking for a new job?
Good Work/Life Balance
Good Work/Life Balance
Good Culture
Remote Work
Good Work/Life Balance
Better Salary
Good Work/Life Balance
Better Salary
Better Salary
Good Culture
Better Salary
Good Culture
Other (Please Specify):Want to move from Australia to Cana
Remote Work
Better Salary
Better Salary
Better Salary
Good Culture
Better Salary
Better Salary
Other (Please Specify):All of the options are important to me
Good Work/Life Balance
Remote Work

T
Q11 - Which Country do you live in?
India
United States
Other (Please Specify):Mozambique
United States
Other (Please Specify):Egypt
India
United States
United States
United Kingdom
United States
United Kingdom
United States
Other (Please Specify):Australia
Canada
United States
Other (Please Specify):Israel
United States
Other (Please Specify):Singapore
United States
Other (Please Specify):Brazil
United States
Other (Please Specify):Costa Rica
Other (Please Specify):Spain

Data after modified:

Q4 - Which industry ?
Healthcare
Finance
Other
Finance
Healthcare
Other
Finance
Other
Healthcare
Telecommunication
Other
Other
Tech
Education
Construction
Finance
Tech
Other
Tech
Other
Finance

Q5 - Programming Language
Python
R
Python
R
R
Python
Python
Other
R
Python
Python
Python
R
Python
R
Python
Python
Python
R
Python
Python

Q8 - The key when looking for a new job?
Remote Work
Remote Work
Good Work/Life Balance
Remote Work
Better Salary
Good Work/Life Balance
Better Salary
Remote Work
Better Salary
Better Salary
Remote Work
Better Salary
Better Salary
Good Work/Life Balance
Good Culture
Good Culture
Better Salary
Better Salary
Better Salary

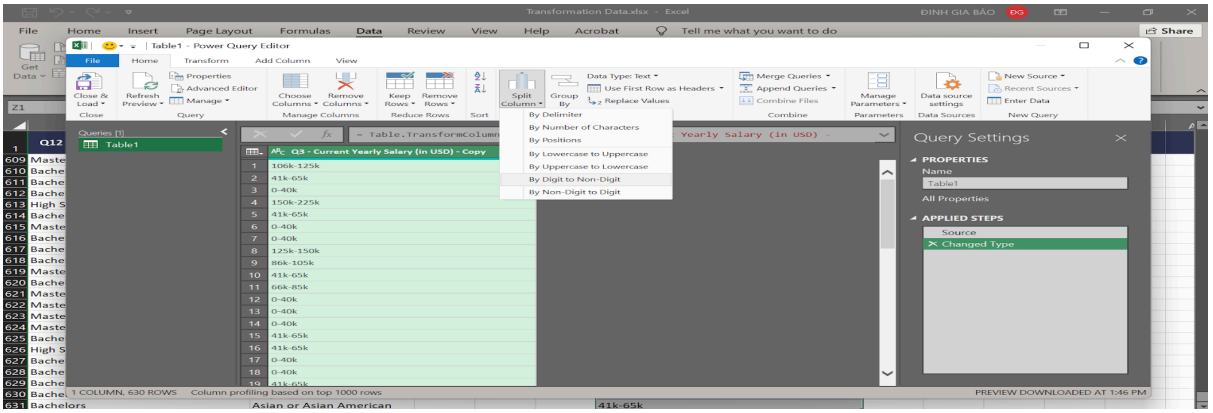
Q11 - Which Country do you live in?
United States
Canada
Other
Canada
United States
Other
Other
United States
United States
United States
Canada
Other
Other
United Kingdom
United States
United States
Other
United States

Modified “ranging data” to specific average data

G
Q3 - Current Yearly Salary (in USD)
106k-125k
41k-65k
0-40k
150k-225k
41k-65k
0-40k
0-40k
125k-150k
86k-105k
41k-65k
66k-85k
0-40k
0-40k
0-40k
41k-65k
41k-65k
0-40k
0-40k
41k-65k
0-40k
41k-65k
106k-125k
0-40k

1. First step:

Using power query editor to change digit to non digit



Q3 - Current Yearly Salary (in USD).1	Q3 - Current Yearly Salary (in USD).2	Q3 - Current Yearly Salary (in USD).3
106	k-125	k
41	k-65	k
0	-40	k
150	k-225	k
41	k-65	k
0	-40	k
0	-40	k
125	k-150	k
86	k-105	k
41	k-65	k
66	k-85	k
0	-40	k
0	-40	k
0	-40	k
41	k-65	k
41	k-65	k
0	-40	k
0	-40	k
41	k-65	k
0	-40	k
41	k-65	k
106	k-125	k
0	-40	k
0	-40	k

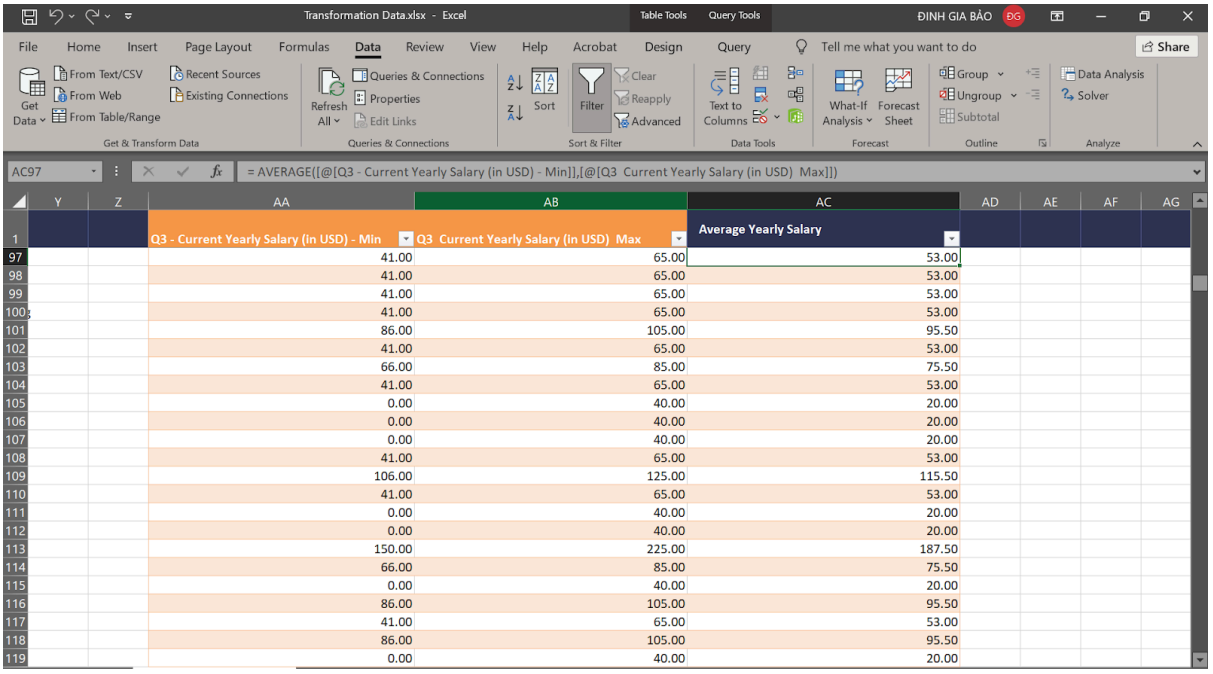
2. Second step

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

Q3 - Current Yearly Salary (in USD) - Min	Q3 Current Yearly Salary (in USD) Max
41.00	65.00
41.00	65.00
41.00	65.00
41.00	65.00
86.00	105.00
41.00	65.00
66.00	85.00
41.00	65.00
0.00	40.00
0.00	40.00
0.00	40.00
41.00	65.00
106.00	125.00
41.00	65.00
0.00	40.00
0.00	40.00
150.00	225.00
66.00	85.00
0.00	40.00
86.00	105.00
41.00	65.00
86.00	105.00
0.00	40.00

3. Final step

Calculate average yearly salary based on Minimum and Maximum value.



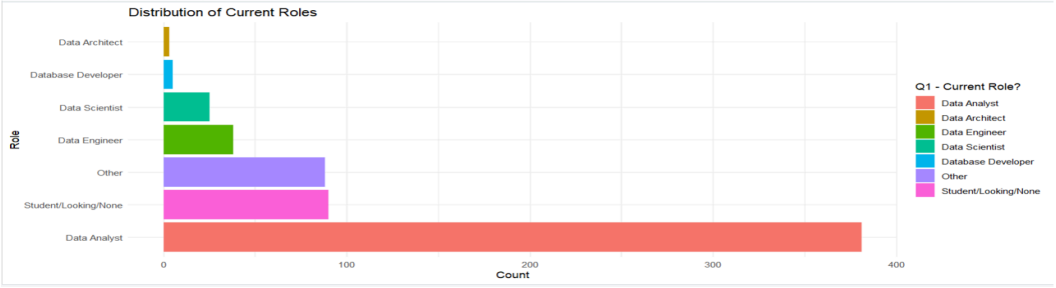
	Y	Z	AA	AB	AC	AD	AE	AF	AG
1			Q3 - Current Yearly Salary (in USD) - Min	Q3 Current Yearly Salary (in USD) Max	Average Yearly Salary				
97			41.00	65.00	53.00				
98			41.00	65.00	53.00				
99			41.00	65.00	53.00				
100			41.00	65.00	53.00				
101			86.00	105.00	95.50				
102			41.00	65.00	53.00				
103			66.00	85.00	75.50				
104			41.00	65.00	53.00				
105			0.00	40.00	20.00				
106			0.00	40.00	20.00				
107			0.00	40.00	20.00				
108			41.00	65.00	53.00				
109			106.00	125.00	115.50				
110			41.00	65.00	53.00				
111			0.00	40.00	20.00				
112			0.00	40.00	20.00				
113			150.00	225.00	187.50				
114			66.00	85.00	75.50				
115			0.00	40.00	20.00				
116			86.00	105.00	95.50				
117			41.00	65.00	53.00				
118			86.00	105.00	95.50				
119			0.00	40.00	20.00				

IV - ANALYSIS WITH R

A. GENERAL ANALYSIS

1. Distribution of Current Roles

```r
#Q1: Distribution of Current Roles (Q1 - Current Role)  ERROR
# Plot the distribution of roles
role_distribution <- data %>%
  filter(!is.na(`Q1 - Current Role?`)) %>%
  group_by(`Q1 - Current Role?`) %>%
  summarize(Count = n())

ggplot(role_distribution, aes(x = reorder(`Q1 - Current Role?`, -Count), y = Count, fill = `Q1 - Current Role?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Distribution of Current Roles", x = "Role", y = "Count") +
  theme_minimal()
```

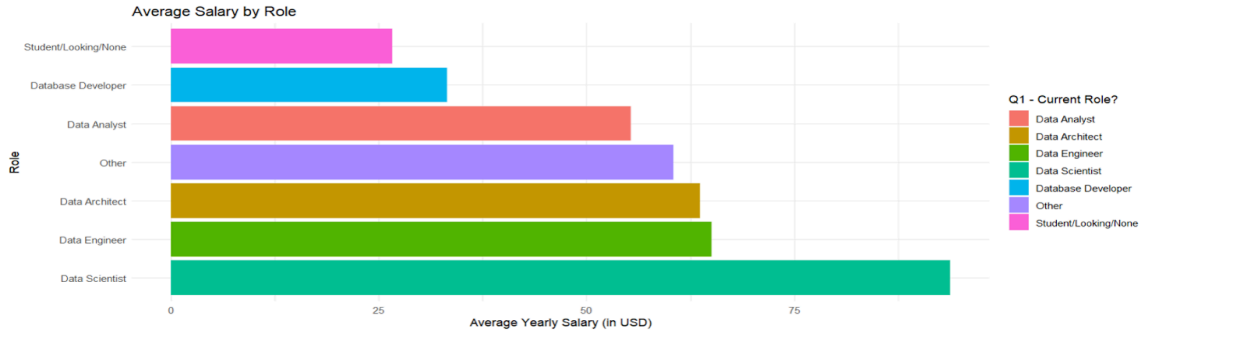


- **Data Analysts** and **Data Scientists** dominate: These two roles are in high demand in the labor market, reflecting the increasing need for data analysis and machine learning applications.
- **Data Engineers** also play a crucial role: This role is essential for building and managing data infrastructure, 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.

2. Salary Distribution by Role

```
#Q3: Salary Distribution by Role
# Plot average salary by role
salary_by_role <- data %>%
  filter(!is.na(`Q1 - Current Role?`), !is.na(`Average Yearly Salary`)) %>%
  group_by(`Q1 - Current Role?`) %>%
  summarize(AverageSalary = mean(`Average Yearly Salary`, na.rm = TRUE))

ggplot(salary_by_role, aes(x = reorder(`Q1 - Current Role?`, -AverageSalary), y = AverageSalary, fill = `Q1 - Current Role?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Average Salary by Role", x = "Role", y = "Average Yearly Salary (in USD)") +
  theme_minimal()
```

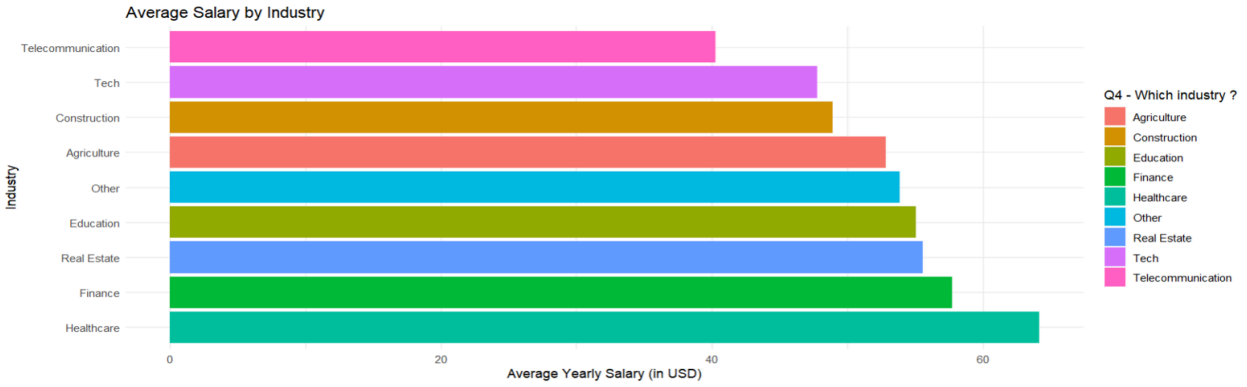


- **Machine Learning Engineer** (Data Scientist): Has the highest salary, reflecting the high demand for AI/ML skills and the scarcity of talent in this field.
- **Data Engineer**: Second highest in salary, showing the importance of building and managing data infrastructure.
- **Data Analyst**: Has a 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.

3. Salary Distribution by Industry

```r
#Q4: Salary Distribution by Industry
# Plot average salary by industry
salary_by_industry <- data %>%
  filter(!is.na(`Q4 - Which industry ?`), !is.na(`Average Yearly Salary`)) %>%
  group_by(`Q4 - Which industry ?`) %>%
  summarize(AverageSalary = mean(`Average Yearly Salary`, na.rm = TRUE))

ggplot(salary_by_industry, aes(x = reorder(`Q4 - Which industry ?`, -AverageSalary), y = AverageSalary, fill = `Q4 - Which industry ?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Average Salary by Industry", x = "Industry", y = "Average Yearly Salary (in USD)") +
  theme_minimal()
```

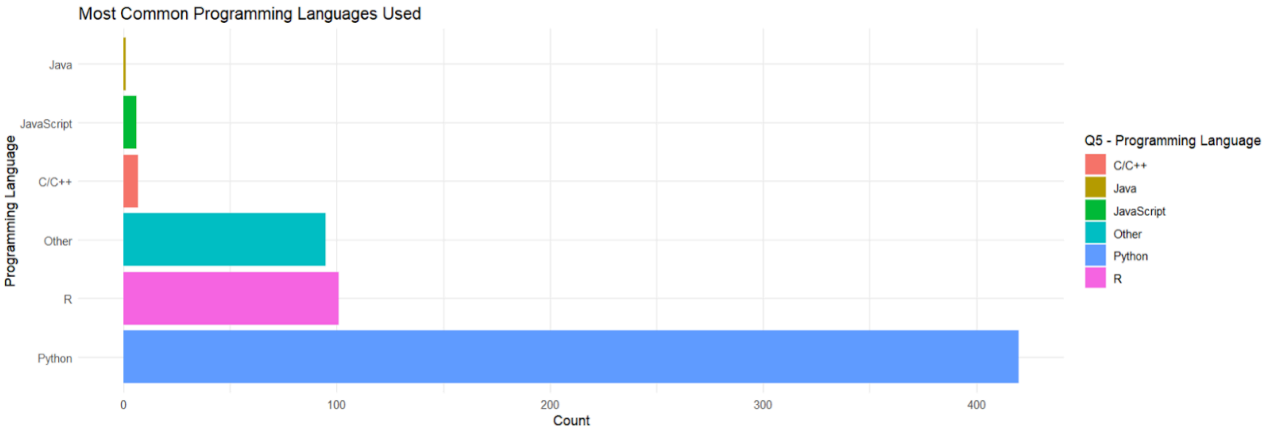


- **The tech industry** leads in **salary**: This reflects the financial strength and heavy reliance on data of tech companies.
- **Finance** and **Banking** rank second: This shows the strong investment of this sector in data analytics.
- Significant differences between industries: **Education** and **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.

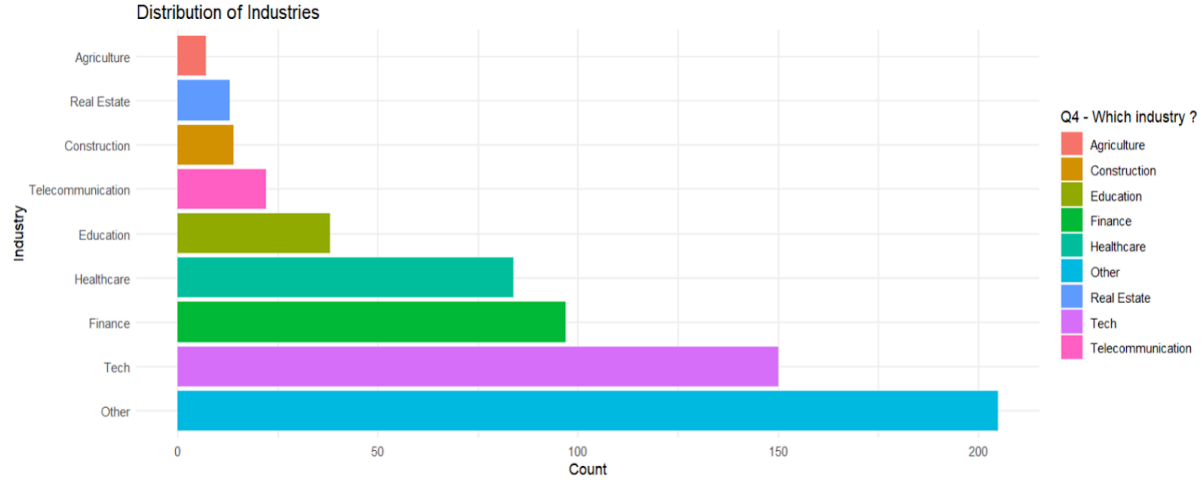
4. Common Programming Languages

```
#Q5: Most Common Programming Languages Used
# Plot the distribution of programming languages
language_distribution <- data %>%
  filter(!is.na(`Q5 - Programming Language`)) %>%
  group_by(`Q5 - Programming Language`) %>%
  summarize(Count = n())

ggplot(language_distribution, aes(x = reorder(`Q5 - Programming Language`, -Count), y = Count, fill = `Q5 - Programming Language`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Most Common Programming Languages Used", x = "Programming Language", y = "Count") +
  theme_minimal()
```



5. Distribution of Industries

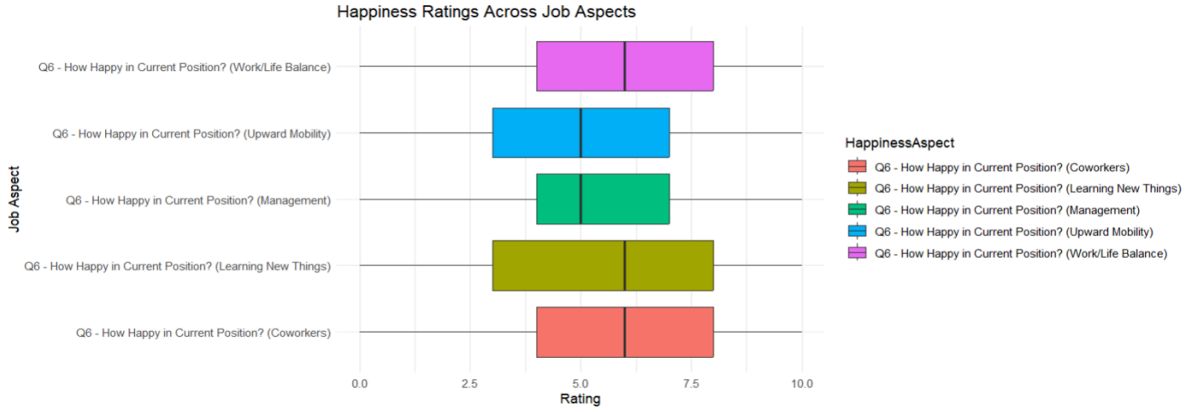
```
#Q4: Distribution of Industries
# Plot the distribution of industries
industry_distribution <- data %>%
  filter(!is.na(`Q4 - Which industry ?`)) %>%
  group_by(`Q4 - Which industry ?`) %>%
  summarize(Count = n())

ggplot(industry_distribution, aes(x = reorder(`Q4 - Which industry ?`, -Count), y = Count, fill = `Q4 - Which industry ?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Distribution of Industries", x = "Industry", y = "Count") +
  theme_minimal()
```



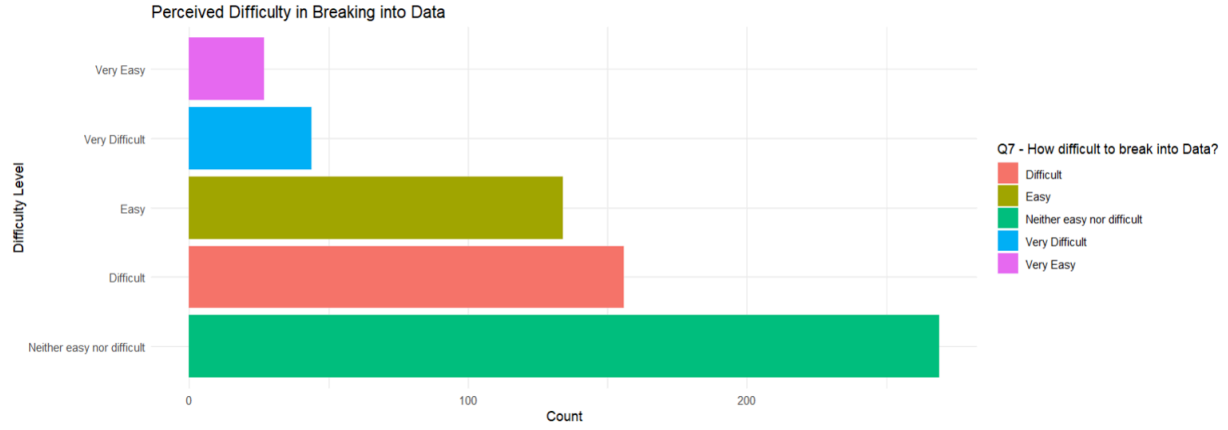
6. Happiness Ratings (Work/Life Balance, Coworkers, Management, Upward Mobility, Learning New Things)

Difficulty in Breaking into Data

```r
#Happiness Ratings (Work/Life Balance, Coworkers, Management, Upward Mobility, Learning New Things
# Reshape data for happiness ratings
# Đảm bảo gói tidyr đã được cài đặt và nạp vào
library(tidyr)

# Sử dụng pivot_longer thay cho gather
happiness_data <- data %>%
  select(`Q4 - Which industry ?`, `Q6 - How Happy in Current Position? (Work/Life Balance)`,
         `Q6 - How Happy in Current Position? (Coworkers)`, `Q6 - How Happy in Current Position? (Management)`,
         `Q6 - How Happy in Current Position? (Upward Mobility)`, `Q6 - How Happy in Current Position? (Learning New Things)`) %>%
  pivot_longer(cols = c(`Q6 - How Happy in Current Position? (Work/Life Balance)`,
                        `Q6 - How Happy in Current Position? (Coworkers)`,
                        `Q6 - How Happy in Current Position? (Management)`,
                        `Q6 - How Happy in Current Position? (Upward Mobility)`,
                        `Q6 - How Happy in Current Position? (Learning New Things)`),
               names_to = "HappinessAspect",
               values_to = "Rating") %>%
  filter(!is.na(Rating))

ggplot(happiness_data, aes(x = HappinessAspect, y = Rating, fill = HappinessAspect)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Happiness Ratings Across Job Aspects", x = "Job Aspect", y = "Rating") +
  theme_minimal()
```

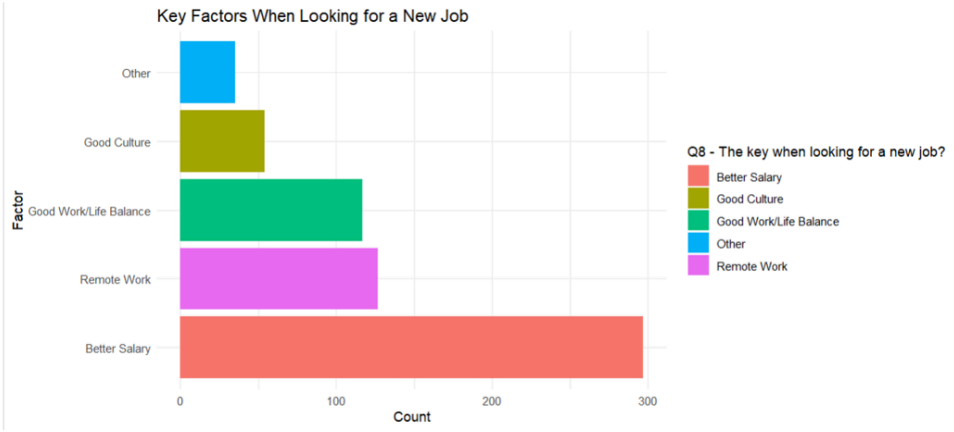



```r
#Difficulty in Breaking into Data
# Plot the perceived difficulty in breaking into data
difficulty_data <- data %>%
  filter(!is.na(`Q7 - How difficult to break into Data?`)) %>%
  group_by(`Q7 - How difficult to break into Data?`) %>%
  summarize(Count = n())

ggplot(difficulty_data, aes(x = reorder(`Q7 - How difficult to break into Data?`, -Count), y = Count, fill = `Q7 - How difficult to break into Data?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Perceived Difficulty in Breaking into Data", x = "Difficulty Level", y = "Count") +
  theme_minimal()
```



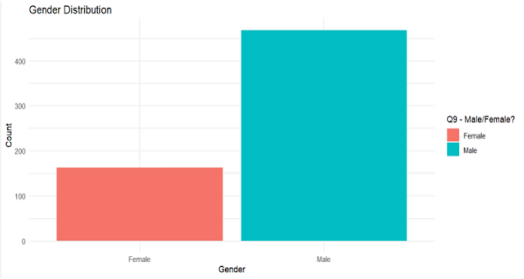
7. Key Factors When Looking for a New Job

```r
#Key Factors When Looking for a New Job
# Plot the key factors when looking for a new job
job_factors <- data %>%
  filter(!is.na(`Q8 - The key when looking for a new job?`)) %>%
  group_by(`Q8 - The key when looking for a new job?`) %>%
  summarize(Count = n())

ggplot(job_factors, aes(x = reorder(`Q8 - The key when looking for a new job?`, -Count), y = Count, fill = `Q8 - The key when looking for a new job?`)
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Key Factors When Looking for a New Job", x = "Factor", y = "Count") +
  theme_minimal()
```



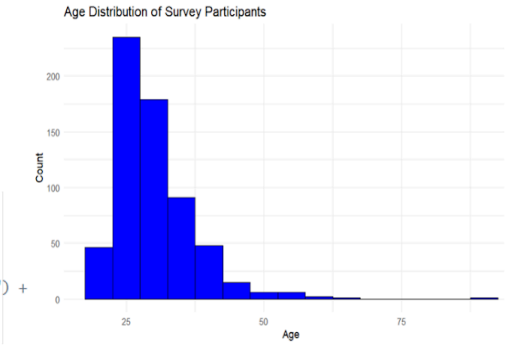
8. Gender Distribution

```r
#Gender Distribution
# Plot the gender distribution
gender_distribution <- data %>%
  filter(!is.na(`Q9 - Male/Female?`)) %>%
  group_by(`Q9 - Male/Female?`) %>%
  summarize(Count = n())

ggplot(gender_distribution, aes(x = `Q9 - Male/Female?`, y = Count, fill = `Q9 - Male/Female?`)) +
  geom_bar(stat = "identity") +
  labs(title = "Gender Distribution", x = "Gender", y = "Count") +
  theme_minimal()
```



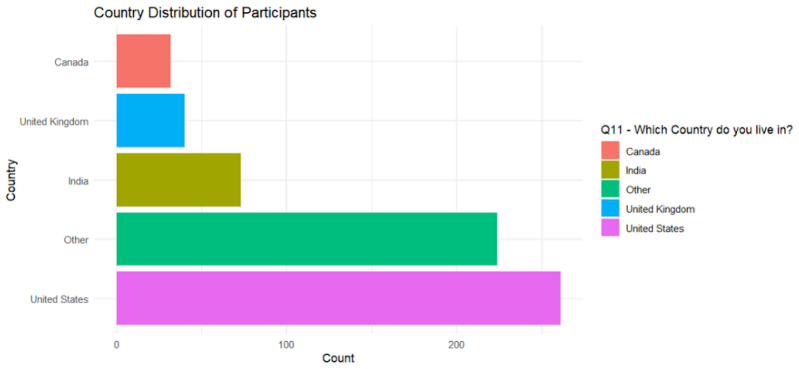
9. Age distribution

```r
#Age Distribution
# Plot the age distribution
ggplot(data, aes(x = `Q10 - Current Age`)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Age Distribution of Survey Participants", x = "Age", y = "Count") +
  theme_minimal()
```



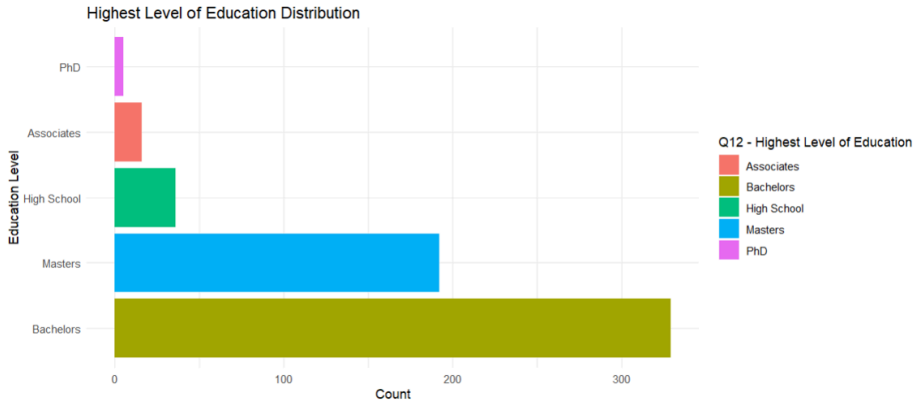
10.Country Distribution

```
#Country Distribution
# Plot the country distribution
country_distribution <- data %>%
  filter(!is.na(`Q11 - Which Country do you live in?`)) %>%
  group_by(`Q11 - Which Country do you live in?`) %>%
  summarize(Count = n())

ggplot(country_distribution, aes(x = reorder(`Q11 - Which Country do you live in?`, -Count), y = Count, fill = `Q11 - Which Country do you live in?`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Country Distribution of Participants", x = "Country", y = "Count") +
  theme_minimal()
```



11. Education Level Distribution

```
#Education Level Distribution# Plot the highest education level distribution
education_distribution <- data %>%
  filter(!is.na(`Q12 - Highest Level of Education`)) %>%
  group_by(`Q12 - Highest Level of Education`) %>%
  summarize(Count = n())

ggplot(education_distribution, aes(x = reorder(`Q12 - Highest Level of Education`, -Count), y = Count, fill = `Q12 - Highest Level of Education`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Highest Level of Education Distribution", x = "Education Level", y = "Count") +
  theme_minimal()
```



12. Ethnicity Distribution

```
# Ethnicity Distribution
# Plot the ethnicity distribution
ethnicity_distribution <- data %>%
  filter(!is.na(`Q13 - Ethnicity`)) %>%
  group_by(`Q13 - Ethnicity`) %>%
  summarize(Count = n())

ggplot(ethnicity_distribution, aes(x = reorder(`Q13 - Ethnicity`, -Count), y = Count, fill = `Q13 - Ethnicity`)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Ethnicity Distribution of Participants", x = "Ethnicity", y = "Count") +
  theme_minimal()
```

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

1. Clustering Algorithm (k-means)

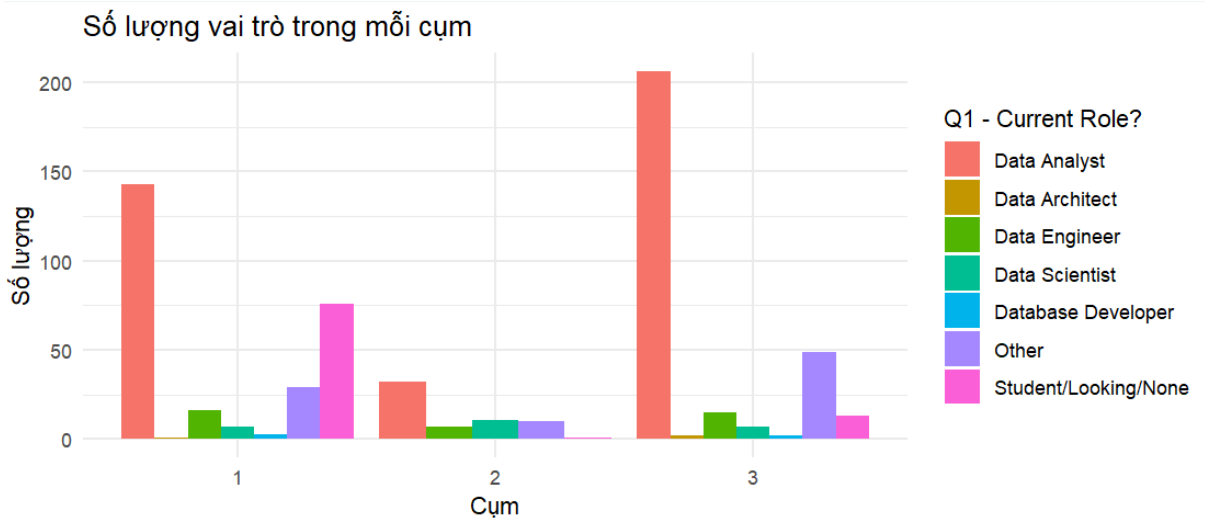
K-means is a clustering algorithm used to divide a set of data into *k* groups (clusters) based on the features of the data. The main goal of the algorithm is to partition the data points into clusters such that points within the same cluster are more similar to each other than to those in other clusters.

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

```r
# Cài đặt các thư viện cần thiết
install.packages("ggplot2")

library(ggplot2)

# Bước 1: Đọc dữ liệu từ file Excel
file_path <- "D:/Data_Learning/DTA301_Analysis/Final Project DTA301/Transformation Data.xlsx"
data <- read_excel(file_path, sheet = "Data Professional Survey")


# Bước 2: Chọn cột liên quan đến vai trò, mức lương, và ngôn ngữ lập trình
df <- data[, c("Q1 - Current Role?", "Q3 - Average Yearly Salary", "Q5 - Programming Language")]

# Bước 3: Tạo các biến số học
#3.1 Đổi các mức lương về trung bình của khoảng lương
df$Salary_Mid <- as.numeric(df$"Q3 - Average Yearly Salary")

#3.2 Mã hóa ngôn ngữ lập trình (chỉ chọn Python=1 và R=2 cho chính xác), ngôn ngữ khác = 0
df$Language <- ifelse(df$`Q5 - Programming Language` == "Python", 1, ifelse(df$`Q5 - Programming Language` == "R", 2, 0))


# Bước 4: Áp dụng K-means clustering với 3 cụm
set.seed(123)
df_numeric <- df[, c("Salary_Mid", "Language")]  # Chọn các biến số học
kmeans_result <- kmeans(df_numeric, centers = 3, nstart = 25)

# Bước 5: Thêm kết quả phân cụm vào dataframe và chuẩn bị cho vẽ biểu đồ
df$Cluster <- as.factor(kmeans_result$cluster)

```

a. Visualization using a Bar Plot.

```r
# 6.1: Trực quan hóa bằng Bar Plot
ggplot(df, aes(x = Cluster, fill = `Q1 - Current Role?`)) +
  geom_bar(position = "dodge") +
  labs(title = "Phân cụm K-means: Số lượng vai trò trong mỗi cụm",
       x = "Cụm", y = "Số lượng") +
  theme_minimal()
```

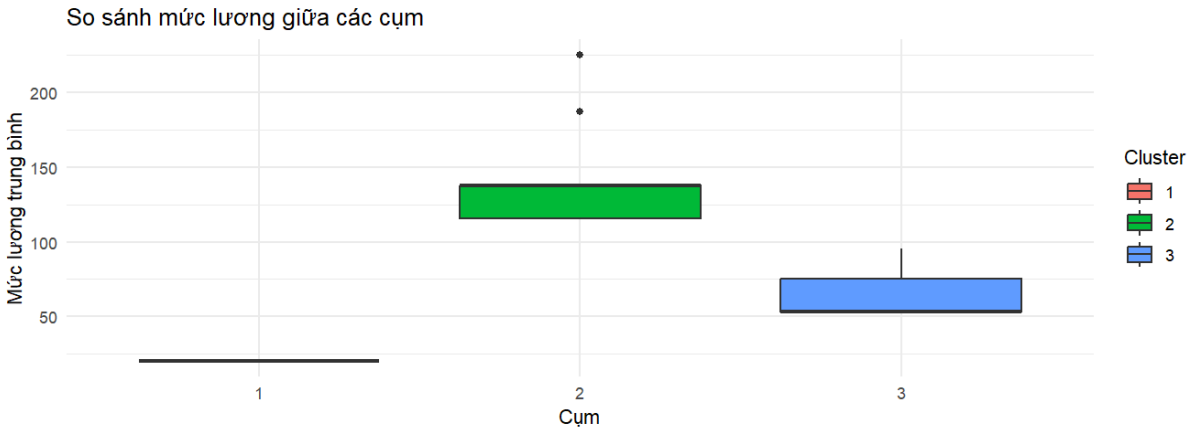


Insight:

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.

b. Visualization using a Box Plot.

```
# 6.2: Trực quan hóa bằng Boxplot
ggplot(df, aes(x = Cluster, y = Salary_Mid, fill = Cluster)) +
  geom_boxplot() +
  labs(title = "So sánh mức lương giữa các cụm",
       x = "Cụm", y = "Mức lương trung bình") +
  theme_minimal()

```



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)

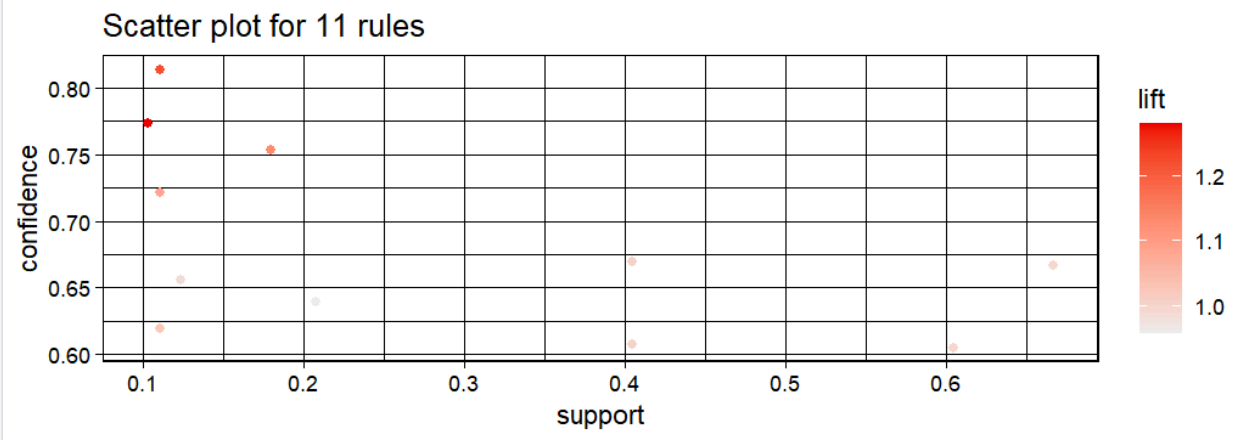
The main purpose of **Association Rule Analysis** is to discover relationships or hidden patterns between attributes in the data.

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."

```r
#Bước 1: Cài đặt thư viện
install.packages("arules")
install.packages("arulesViz")
library(arules)
library(arulesViz)
library(readxl)

#Bước 2: Đọc dữ liệu từ file Excel
file_path <- "D:/Data_Learning/DTA301_Analysis/Final Project DTA301/Transformation Data.xlsx"
df <- read_excel(file_path, sheet = "Data Professional Survey")

#Bước 3: Chuẩn bị dữ liệu: Chọn các cột muốn phân tích và chuyển đổi thành các yếu tố (factors):
#3.1: Chọn các cột cần thiết
df_selected <- df[, c("Q1 - Current Role?", "Q4 - Which industry ?", "Q5 - Programming Language")]

#3.2: Loại bỏ các giá trị NA
df_selected <- na.omit(df_selected)

#3.3: Chuyển đổi dữ liệu thành các yếu tố
df_selected[] <- lapply(df_selected, as.factor)

#Bước 4: Chuyển đổi dữ liệu sang dạng transactions:
library(arules)
# Chuyển đổi thành dạng transactions
transactions <- as(df_selected, "transactions")

#Bước 5: Áp dụng thuật toán Apriori:
# Tìm các tập hợp mục phổ biến với ngưỡng hỗ trợ tối thiểu là 0.1 (10%)
rules <- apriori(transactions, parameter = list(support = 0.1, confidence = 0.6))

```

```r
#Bước 6: Trực quan hóa kết quả:
library(arulesViz)
# Trực quan hóa các quy tắc với biểu đồ scatter plot
plot(rules, method = "scatterplot", measure = c("support", "confidence"), shading = "lift")
```



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.

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

```r
#B1: Cài đặt và sử dụng packages
install.packages("readxl")
install.packages("tidyverse")

library(readxl)
library(tidyverse)

#B2: Đọc dữ liệu từ file Excel
df <- read_excel("D:/Data_Learning/DTA301_Analysis/Final Project DTA301/Transformation Data.xlsx", sheet = "Data Professional Survey")

#B3: Chạy mô hình hồi quy đa biến
#Y: Biến phụ thuộc, X: Biến độc lập
model_multiple <- lm(`Q6 - How Happy in Current Position? (Work/Life Balance)` ~
                       `Q1 - Current Role?` + `Q3 - Average Yearly Salary` + `Q9 - Male/Female?`, data = df)

# Hiển thị kết quả mô hình
summary(model_multiple)

install.packages("ggplot2")
library(ggplot2)

# Biểu đồ hồi quy cho Q3 - Average Yearly Salary
ggplot(df, aes(x = `Q3 - Average Yearly Salary`, y = `Q6 - How Happy in Current Position? (Work/Life Balance)`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Hồi quy giữa Salary và Happiness", x = "Average Yearly Salary", y = "Happiness")

```

<u>Model summary</u>

```
> summary(model_multiple)

Call:
lm(formula = `Q6 - How Happy in Current Position? (Work/Life Balance)` ~
    `Q1 - Current Role?` + `Q3 - Average Yearly Salary` + `Q9 - Male/Female?`,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3190 -1.9251 -0.0017  2.0099  6.0749

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             5.111200   0.282594  18.087  < 2e-16 ***
`Q1 - Current Role?`Data Architect     -1.437728   1.536167  -0.936    0.350
`Q1 - Current Role?`Data Engineer      -0.236079   0.453696  -0.520    0.603
`Q1 - Current Role?`Data Scientist     -0.445290   0.560667  -0.794    0.427
`Q1 - Current Role?`Database Developer  0.706152   1.194617   0.591    0.555
`Q1 - Current Role?`Other              -0.052091   0.314193  -0.166    0.868
`Q1 - Current Role?`Student/Looking/None -1.468163   0.330523  -4.442 1.06e-05 ***
`Q3 - Average Yearly Salary`            0.014102   0.002888   4.882 1.34e-06 ***
`Q9 - Male/Female?`Male                 0.143096   0.246550   0.580    0.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 611 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.09126,   Adjusted R-squared:  0.07936
F-statistic:  7.67 on 8 and 611 DF,  p-value: 8.176e-10
```

<u>Comments</u>:

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).

2. **Model Fit:**

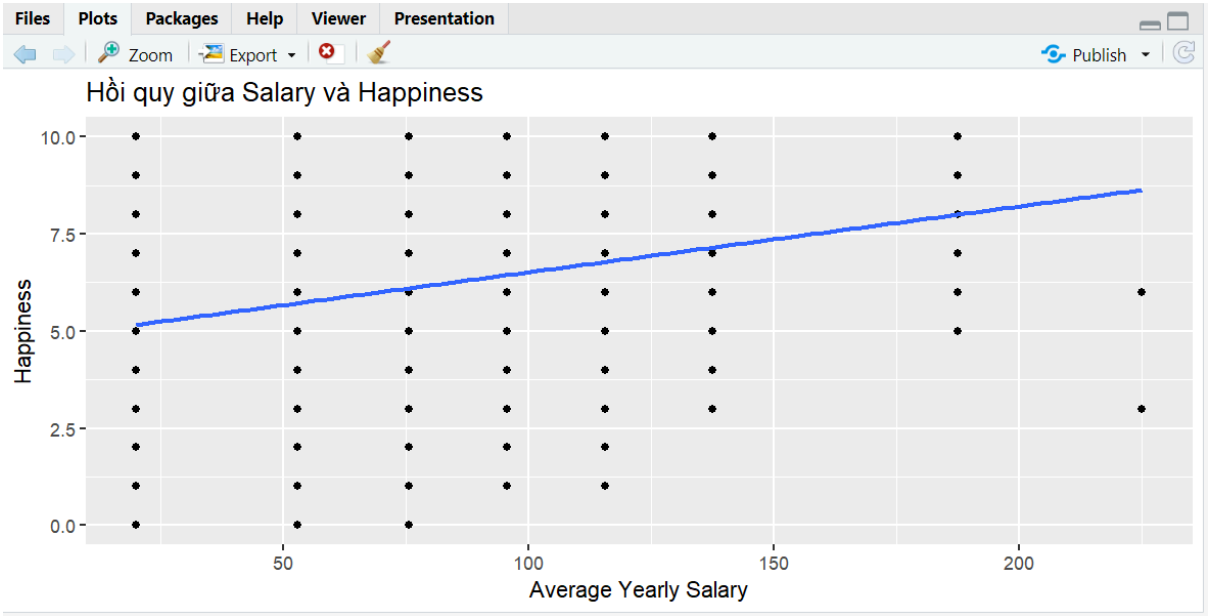
   Multiple R-squared: 0.09126, indicating that the model explains approximately 9.1% of the variability in the dependent variable.

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

3. **Overall Statistical Significance:**

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

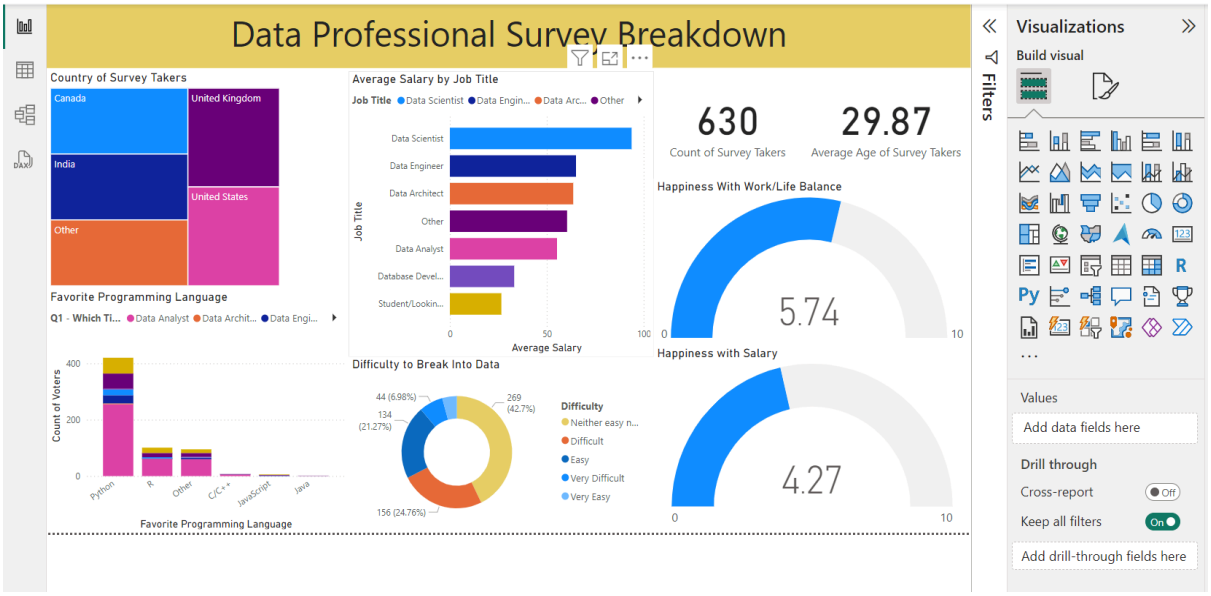
=> The model shows that several independent variables significantly affect job satisfaction, particularly the current role and average yearly salary.



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

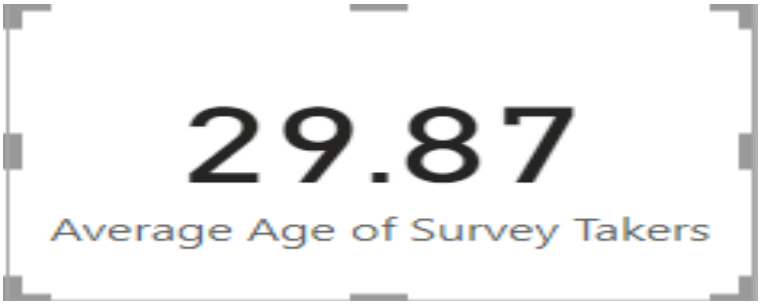
Power BI provides interactive data visualization capabilities, enabling users to easily analyze, make data-driven decisions, and collaborate effectively through customized reports and dashboards.

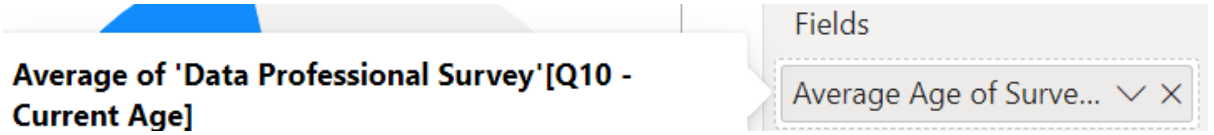


Go in depth to each visualization:

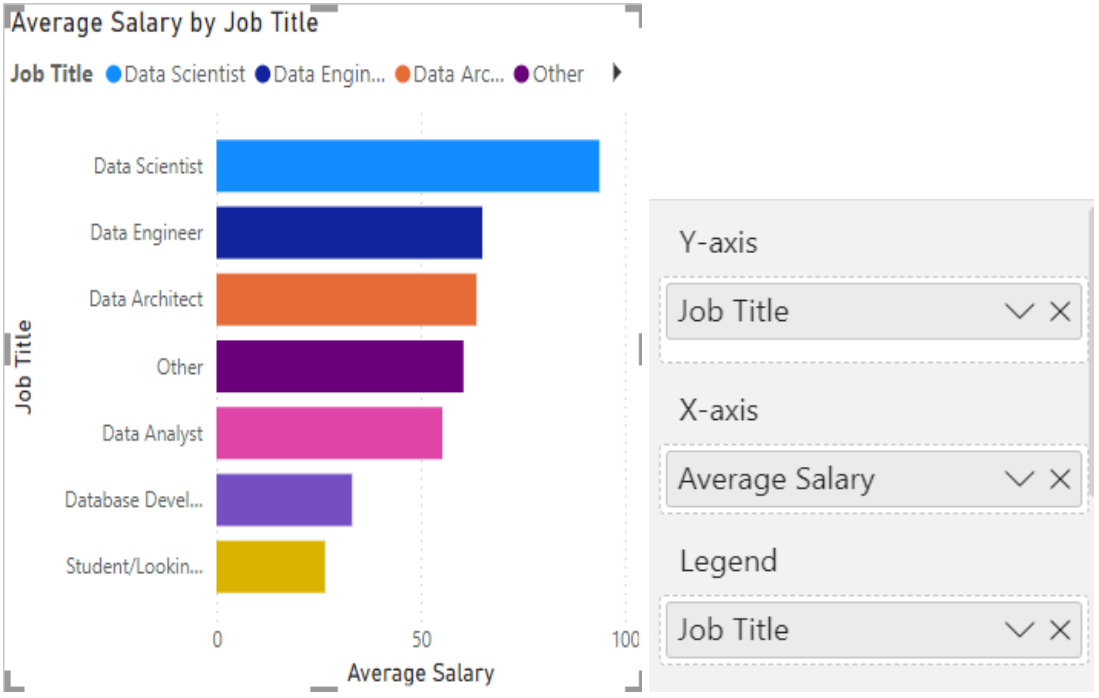
- **Total number of survey takers** (How many people took this survey ?)

- **Average age of survey takers**

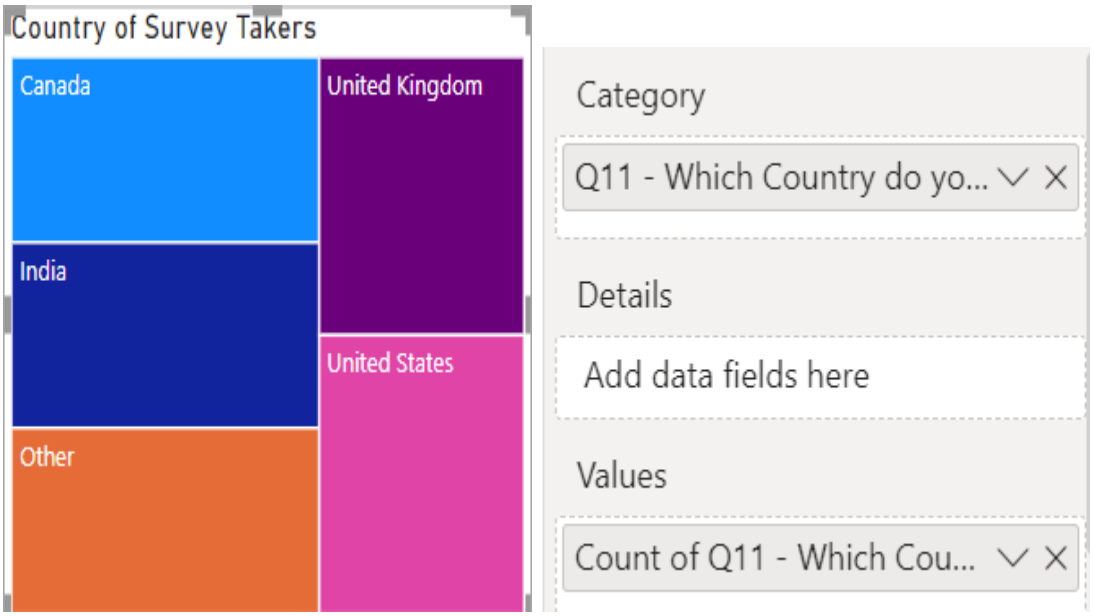




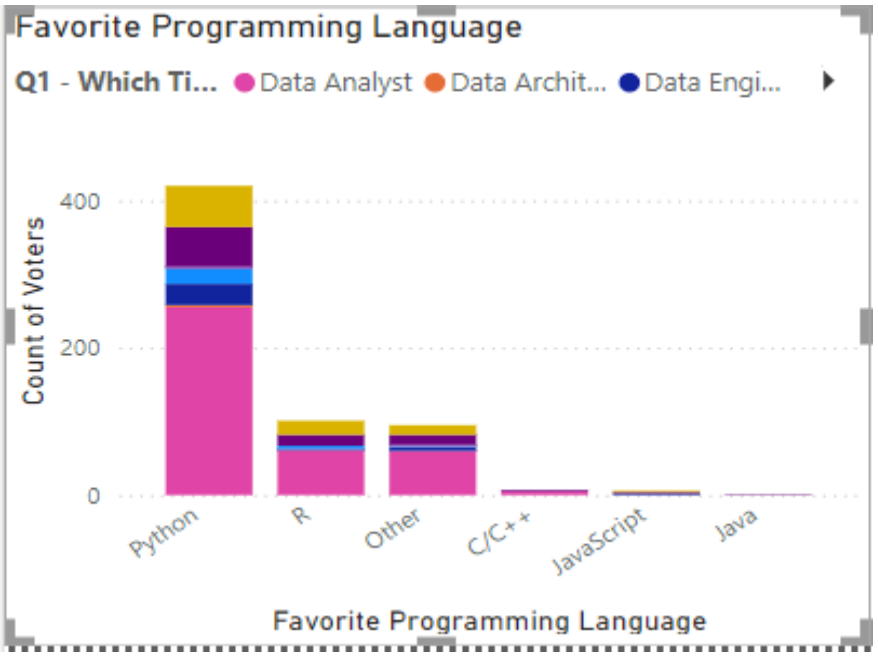
- **Average salary by Job title**



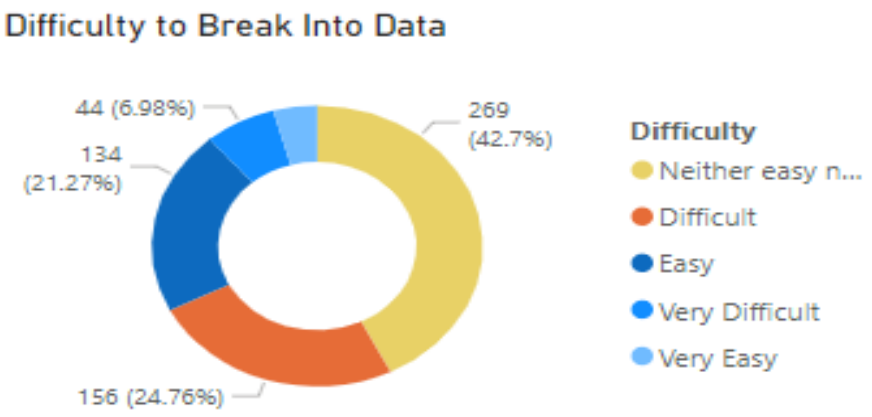
- **Country of survey takers**

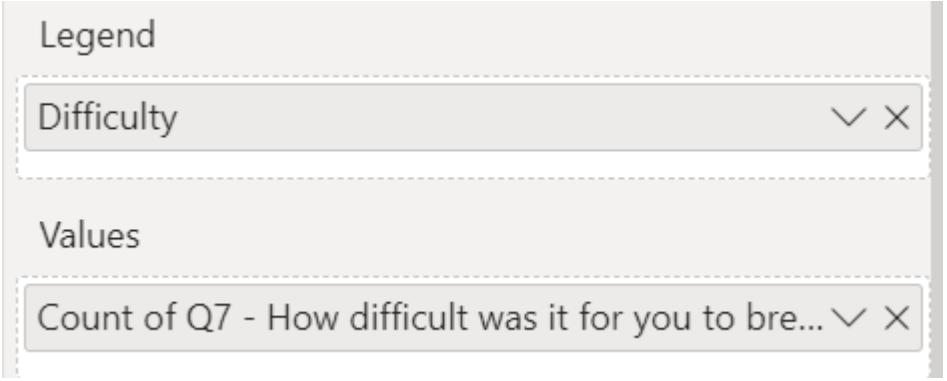


- **Favorite programming Language**

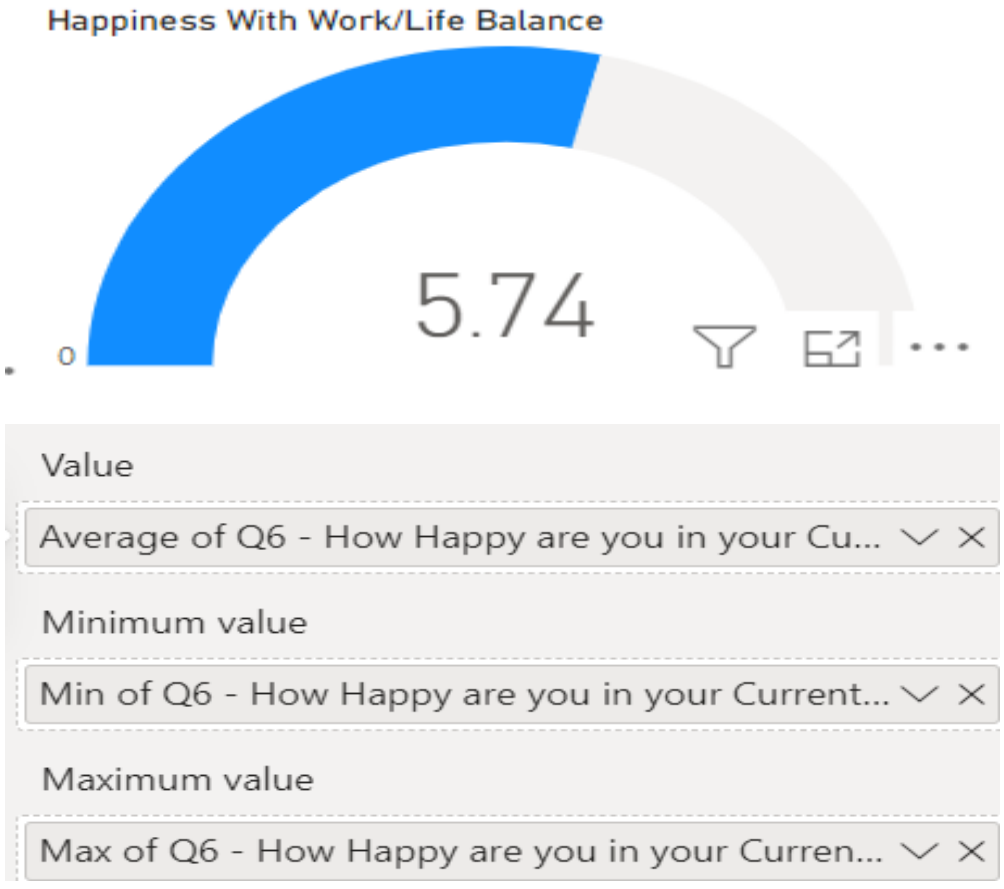


- **Difficulty to break into data profession**





- **Happiness with Work/Life Balance**



- **Happiness with Salary**