Project 3

MD SIMON CHOWDHURY

2024-10-30

Loading and Exploring the Data

Let’s load the dataset and take a quick look at the first few rows to understand its structure.

```
# Load data
data <- read_csv("https://raw.githubusercontent.com/simonchy/DATA607/refs/heads/main/week%20
8/Most%20Valuable%20Data%20Science%20Skills%20(Responses)%20-%20Form%20Responses%201.csv")
```

```
## Rows: 11 Columns: 17
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (15): Timestamp, First Name or Nickname, List the 5 most valuable data s...
## dbl  (1): Age
## lgl  (1): Name One Most Useful Data Sc
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Display the first few rows of the data
head(data)
```

```
## # A tibble: 6 × 17
##   Timestamp  First Name or Nickna…¹ List the 5 most valu…² `Email Address`   Age
##   <chr>      <chr>                  <chr>                  <chr>           <dbl>
## 1 10/22/202… Inna                   Programming, Data Man… <NA>               NA
## 2 10/23/202… Md Asaduzzaman         <NA>                   m.zaman3201@gm…    42
## 3 10/23/202… Alex                   <NA>                   alexander.ptac…    27
## 4 10/23/202… Inna                   <NA>                   innayedzinovic…    29
## 5 10/23/202… Cindy                  <NA>                   cindylin90@gma…    34
## 6 10/23/202… Sarika                 <NA>                   ssgupta.phd@gm…    46
## # ℹ abbreviated names: ¹`First Name or Nickname`,
## #   ²`List the 5 most valuable data science skills (separated by commas)`
## # ℹ 12 more variables: `Any data science/data analytics experience?` <chr>,
## #   `Any software engineering experience?` <chr>,
## #   `Which programming languages do you use most frequently?` <chr>,
## #   `What resources do you use for learning new data science skills?` <chr>,
## #   `What areas of data science are you most interested in learning more about?` <chr>, …
```

Data Cleaning

We’ll clean up the data by renaming columns and removing rows with missing values in skill columns.

```
# Standardize column names to avoid spaces and special characters
colnames(data) <- make.names(colnames(data))

# Select only the columns with the top skills
skills_data <- data %>%
  select(Name..1.most.most.valuable.data.science.skill,
         Name..2.most.most.valuable.data.science.skill,
         Name..3.most.most.valuable.data.science.skill,
         Name..4.most.most.valuable.data.science.skill,
         Name..5.most.most.valuable.data.science.skill) %>%
  pivot_longer(cols = everything(),
               names_to = "skill_rank",
               values_to = "skill") %>%
  filter(!is.na(skill)) # Remove rows with NA skills

# Check cleaned and reshaped data
head(skills_data)
```

```
## # A tibble: 6 × 2
##   skill_rank                                    skill
##   <chr>                                         <chr>
## 1 Name..1.most.most.valuable.data.science.skill R language skill
## 2 Name..2.most.most.valuable.data.science.skill Python skill
## 3 Name..3.most.most.valuable.data.science.skill Statistics and math skill
## 4 Name..4.most.most.valuable.data.science.skill Data Visualization skill
## 5 Name..5.most.most.valuable.data.science.skill SQL skill
## 6 Name..1.most.most.valuable.data.science.skill data cleaning
```

Analysis: Most Valued Data Science Skills

Let’s calculate the frequency of each skill to determine which are the most valued.

```
# Count the occurrences of each skill
skill_counts <- skills_data %>%
  group_by(skill) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

# Display the top skills
head(skill_counts, 10)
```

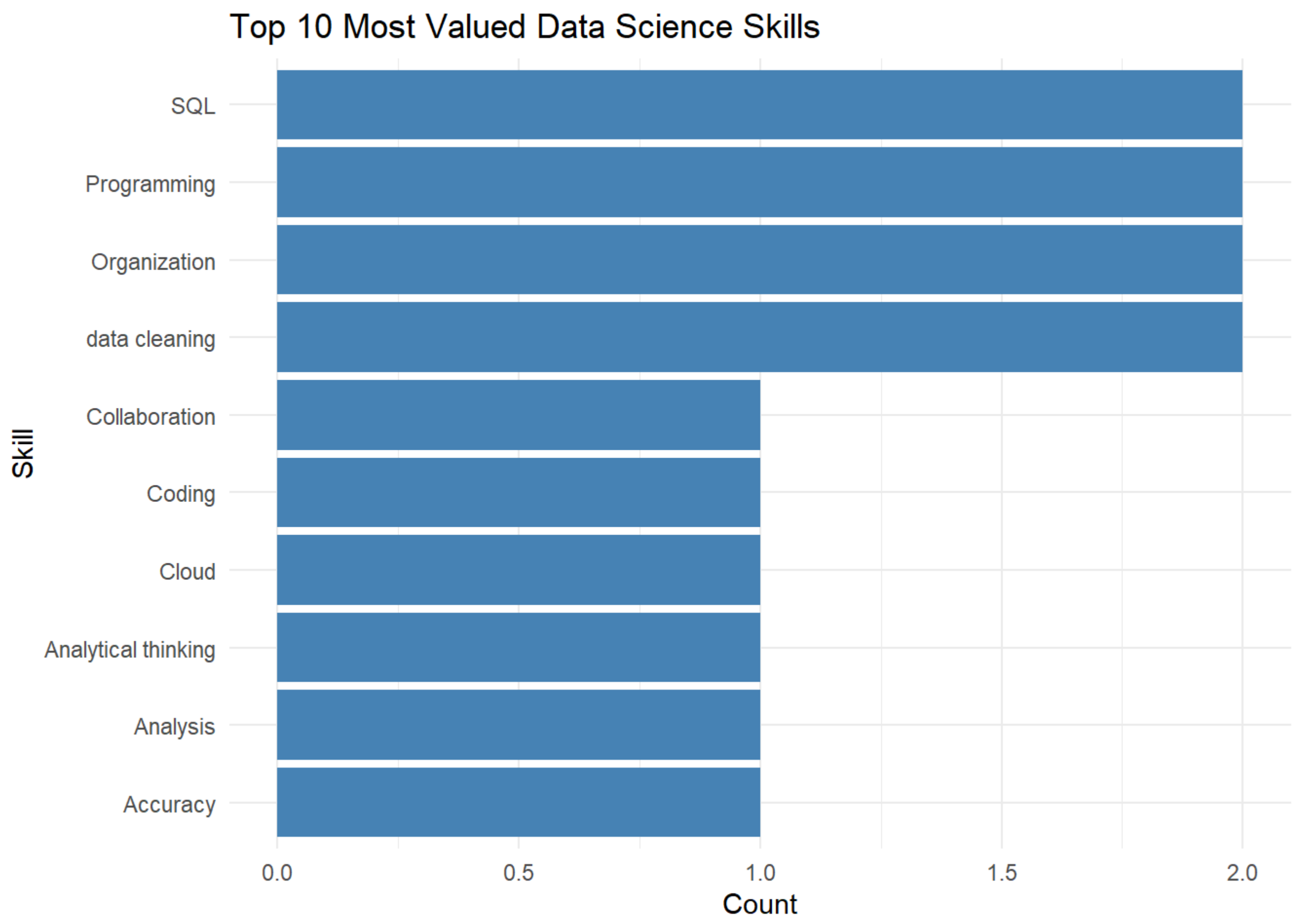
```
## # A tibble: 10 × 2
##    skill               count
##    <chr>               <int>
##  1 Organization            2
##  2 Programming             2
##  3 SQL                     2
##  4 data cleaning           2
##  5 Accuracy                1
##  6 Analysis                1
##  7 Analytical thinking     1
##  8 Cloud                   1
##  9 Coding                  1
## 10 Collaboration           1
```

Visualization

Finally, we visualize the most valued data science skills using a bar plot.

```
# Plot the top 10 most valued skills
library(ggplot2)

ggplot(skill_counts[1:10, ], aes(x = reorder(skill, count), y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Most Valued Data Science Skills",
       x = "Skill",
       y = "Count") +
  theme_minimal()
```



Conclusion

Based on our analysis, the top skills identified in the dataset are:

```
# Display top skills as a list
skill_counts[1:10, ]
```

```
## # A tibble: 10 × 2
##    skill               count
##    <chr>               <int>
##  1 Organization            2
##  2 Programming             2
##  3 SQL                     2
##  4 data cleaning           2
##  5 Accuracy                1
##  6 Analysis                1
##  7 Analytical thinking     1
##  8 Cloud                   1
##  9 Coding                  1
## 10 Collaboration           1
```

These skills represent the most valued abilities in the data science field according to the responses provided and Resourcefulness is in top as most valued ability in the data science field.