Data 607 Project 3

Inna Yedzinovich, Zaneta Paulusova, Md Asaduzzaman, Md. Asadul, Md. Simon Chowdhury

2024-11-03

Introduction

For this project we focused on answering the question “What are the most valued data science skills?”.

Our approach to answering this question involved creating a survey to identify five valuable data science skills as perceived by survey respondents. We then compared these findings to the article titled “Data Science Skills Survey 2022 – By AIM and Great Learning.” In this article, we focused on a table that lists common skills desired by recruiters, categorized by years of experience. We will adjust the data to ensure the skills align with one another. Additionally, we will calculate the mean for the various years of experience columns from the website’s data. Finally, we will combine both datasets into one graph for a comparative analysis of our internal survey results alongside the findings from the website.

```
library(DBI)
library(RMySQL)

user <- Sys.getenv("MYSQL_USER")
password <- Sys.getenv("MYSQL_PASSWORD")
host <- Sys.getenv("MYSQL_HOST")
dbname <- Sys.getenv("MYSQL_DBNAME")

conn <- dbConnect(RMySQL::MySQL(), user = user, password = password, host = host, dbname = dbname)

create_experience_table <- "
CREATE TABLE experience (
  experience_id INT NOT NULL,
  respondent_id INT NULL,
  data_science_experience TINYINT(1) NULL,
  software_engineering_experience TINYINT(1) NULL,
  PRIMARY KEY (experience_id)
);"

create_respondents_table <- "
CREATE TABLE respondents (
  respondent_id INT NOT NULL,
  first_name VARCHAR(45) NULL,
  last_name VARCHAR(45) NULL,
  age INT NULL,
  PRIMARY KEY (respondent_id)
);"
```

```r
create_interestareas_table <- "
CREATE TABLE interestareas (
  interest_id INT NOT NULL,
  respondent_id INT NULL,
  interest_area VARCHAR(45) NULL,
  PRIMARY KEY (interest_id)
);"

create_softskills_table <- "
CREATE TABLE softskills (
  soft_skill_id INT NOT NULL,
  respondent_id INT NULL,
  soft_skill VARCHAR(45) NULL,
  PRIMARY KEY (soft_skill_id)
);"

create_programminglanguages_table <- "
CREATE TABLE programminglanguages (
  language_id INT NOT NULL,
  respondent_id INT NULL,
  language VARCHAR(45) NULL,
  PRIMARY KEY (language_id)
);"

create_learningresources_table <- "
CREATE TABLE learningresources (
  resource_id INT NOT NULL,
  respondent_id INT NULL,
  resource VARCHAR(45) NULL,
  PRIMARY KEY (resource_id)
);"

create_valuableskills_table <- "
CREATE TABLE valuableskills (
  valuable_skill_id INT NOT NULL,
  respondent_id INT NULL,
  skill_rank INT NULL,
  skill_name VARCHAR(45) NULL,
  PRIMARY KEY (valuable_skill_id)
);"

dbExecute(conn, create_experience_table)
```

Tables creation for our data

```
## [1] 0
```

```r
dbExecute(conn, create_respondents_table)
```

```
## [1] 0
```

```r
dbExecute(conn, create_interestareas_table)
```

```
## [1] 0
```

```r
dbExecute(conn, create_softskills_table)
```

```
## [1] 0
```

```r
dbExecute(conn, create_programminglanguages_table)
```

```
## [1] 0
```

```r
dbExecute(conn, create_learningresources_table)
```

```
## [1] 0
```

```r
dbExecute(conn, create_valuableskills_table)
```

```
## [1] 0
```

```r
dbDisconnect(conn)
```

```
## [1] TRUE
```

This process involves loading a CSV file into R, tidying and normalizing the data, and then loading the cleaned data into a database.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)

data <- read.csv("https://raw.githubusercontent.com/simonchy/DATA607/refs/heads/main/week%208/Cleaned_A
colnames(data)
```

```
##  [1] "Timestamp"
##  [2] "First.Name.or.Nickname"
##  [3] "List.the.5.most.valuable.data.science.skills..separated.by.commas."
##  [4] "Email.Address"
##  [5] "Age"
##  [6] "Any.data.science.data.analytics.experience."
##  [7] "Any.software.engineering.experience."
##  [8] "Which.programming.languages.do.you.use.most.frequently."
##  [9] "What.resources.do.you.use.for.learning.new.data.science.skills."
## [10] "What.areas.of.data.science.are.you.most.interested.in.learning.more.about."
## [11] "Name..1.most.most.valuable.data.science.skill"
## [12] "Name..2.most.most.valuable.data.science.skill"
## [13] "Name..3.most.most.valuable.data.science.skill"
## [14] "Name..4.most.most.valuable.data.science.skill"
## [15] "Name..5.most.most.valuable.data.science.skill"
## [16] "Which.soft.skill.do.you.think.is.most.important.for.a.data.scientist."
```

```r
colnames(data) <- tolower(colnames(data))

data <- data %>%
  rename(
    timestamp = timestamp,
    first_name = first.name.or.nickname,
    valuable_skills = list.the.5.most.valuable.data.science.skills..separated.by.commas.,
    email = email.address,
    age = age,
    data_science_experience = any.data.science.data.analytics.experience.,
    software_engineering_experience = any.software.engineering.experience.,
    programming_languages = which.programming.languages.do.you.use.most.frequently.,
    learning_resources = what.resources.do.you.use.for.learning.new.data.science.skills.,
    interest_areas = what.areas.of.data.science.are.you.most.interested.in.learning.more.about.,
    skill_1 = name..1.most.most.valuable.data.science.skill,
    skill_2 = name..2.most.most.valuable.data.science.skill,
    skill_3 = name..3.most.most.valuable.data.science.skill,
    skill_4 = name..4.most.most.valuable.data.science.skill,
    skill_5 = name..5.most.most.valuable.data.science.skill,
    soft_skill = which.soft.skill.do.you.think.is.most.important.for.a.data.scientist.
  ) %>%
  mutate(across(everything(), tolower))

# Create respondent_id before separating rows
data <- data %>%
  mutate(respondent_id = row_number())

# Normalize the data
respondents <- data %>%
  select(first_name, age, respondent_id)

experience <- data %>%
  select(data_science_experience, software_engineering_experience, respondent_id) %>%
  mutate(experience_id = row_number())

valuable_skills <- data %>%
  select(skill_1, skill_2, skill_3, skill_4, skill_5, respondent_id) %>%
```

```r
  pivot_longer(cols = starts_with("skill_"), names_to = "skill_rank", values_to = "skill_name") %>%
  mutate(valuable_skill_id = row_number())

programming_languages <- data %>%
  select(respondent_id, programming_languages) %>%
  separate_rows(programming_languages, sep = ",") %>%
  mutate(language_id = row_number())

learning_resources <- data %>%
  select(respondent_id, learning_resources) %>%
  separate_rows(learning_resources, sep = ",") %>%
  mutate(resource_id = row_number())

interest_areas <- data %>%
  select(respondent_id, interest_areas) %>%
  separate_rows(interest_areas, sep = ",") %>%
  mutate(interest_id = row_number())

soft_skills <- data %>%
  select(soft_skill, respondent_id) %>%
  mutate(soft_skill_id = row_number())

conn <- dbConnect(RMySQL::MySQL(), user = Sys.getenv("MYSQL_USER"), password = Sys.getenv("MYSQL_PASSWO

# Load the data into the database
dbWriteTable(conn, "respondents", respondents, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "experience", experience, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "valuableskills", valuable_skills, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "programminglanguages", programming_languages, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "learningresources", learning_resources, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "interestareas", interest_areas, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbWriteTable(conn, "softskills", soft_skills, overwrite = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```r
dbDisconnect(conn)
```

```
## [1] TRUE
```

Let's demonstrates how to connect to a MySQL database, retrieve data, and visualize it using R.

```r
library(ggplot2)

conn <- dbConnect(RMySQL::MySQL(), user = Sys.getenv("MYSQL_USER"), password = Sys.getenv("MYSQL_PASSWOR
valuable_skills <- dbReadTable(conn, "valuableskills")
dbDisconnect(conn)
```

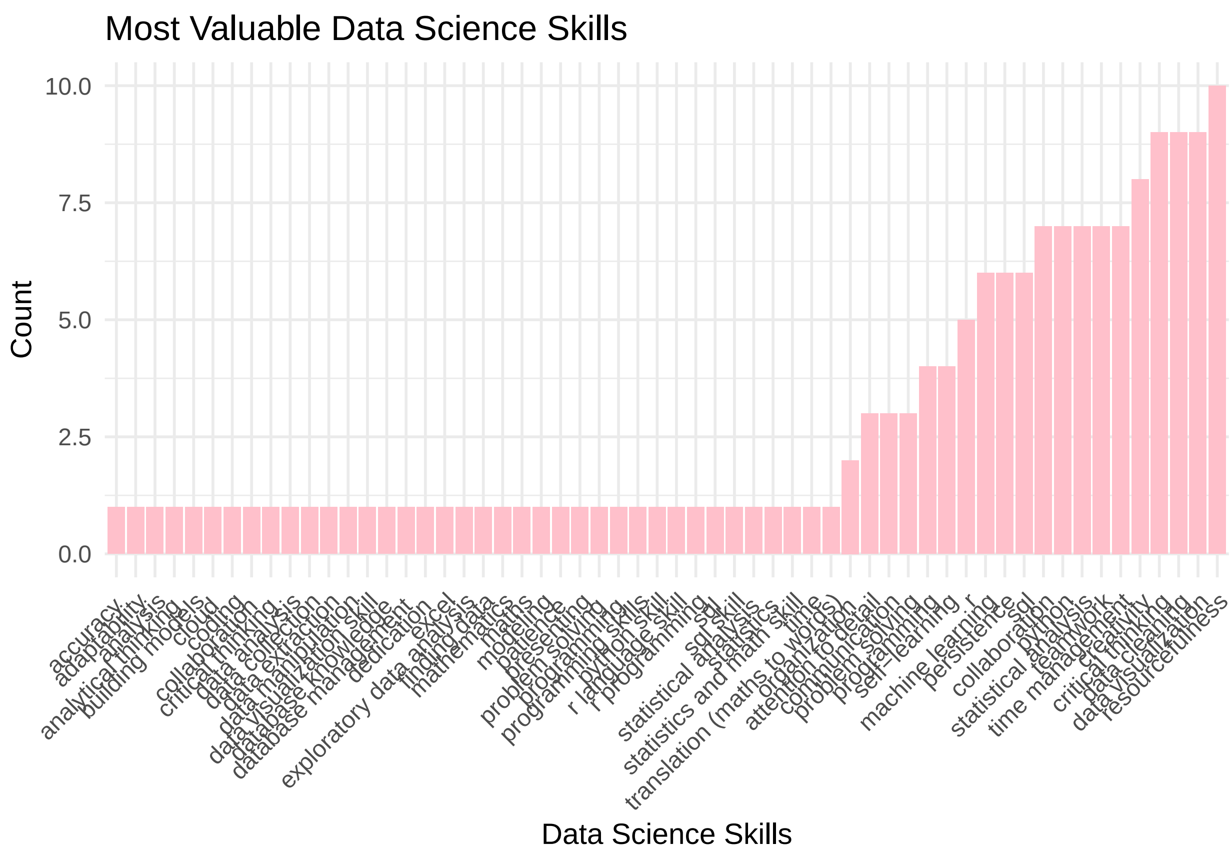
```
## [1] TRUE
```

```r
skill_counts <- valuable_skills %>%
  count(skill_name, sort = TRUE)
print(skill_counts)
```

```
##                  skill_name  n
## 1            resourcefulness 10
## 2          critical thinking  9
## 3              data cleaning  9
## 4         data visualization  9
## 5                 creativity  8
## 6              collaboration  7
## 7                     python  7
## 8       statistical analysis  7
## 9                   teamwork  7
## 10           time management  7
## 11          machine learning  6
## 12               persistence  6
## 13                       sql  6
## 14                         r  5
## 15               programming  4
## 16             self-learning  4
## 17        attention to detail  3
## 18             communication  3
## 19           problem solving  3
## 20             organization   2
## 21                 accuracy   1
## 22              adaptability  1
## 23                  analysis  1
## 24       analytical thinking  1
## 25           building models  1
## 26                    cloud   1
## 27                    coding  1
## 28            collaboration   1
```

```
## 29          critical thinking   1
## 30               data analysis  1
## 31             data collection  1
## 32             data extraction  1
## 33           data manipulation  1
## 34    data visualization skill  1
## 35         database knowledge   1
## 36        database management   1
## 37                 dedication   1
## 38                       excel  1
## 39     exploratory data analysis 1
## 40                finding data  1
## 41                 mathematics  1
## 42                       maths  1
## 43                    modeling  1
## 44                   patience   1
## 45                  presenting  1
## 46             problem solving  1
## 47                programming   1
## 48          programming skills  1
## 49                python skill  1
## 50            r language skill  1
## 51               r programming  1
## 52                        sql   1
## 53                   sql skill  1
## 54       statistical analysis   1
## 55                 statistics   1
## 56    statistics and math skill 1
## 57                        time  1
## 58 translation (maths to words) 1
```

```
ggplot(skill_counts, aes(x = reorder(skill_name, n), y = n)) +
  geom_bar(stat = "identity", fill = "pink") +
  labs(title = "Most Valuable Data Science Skills", x = "Data Science Skills", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

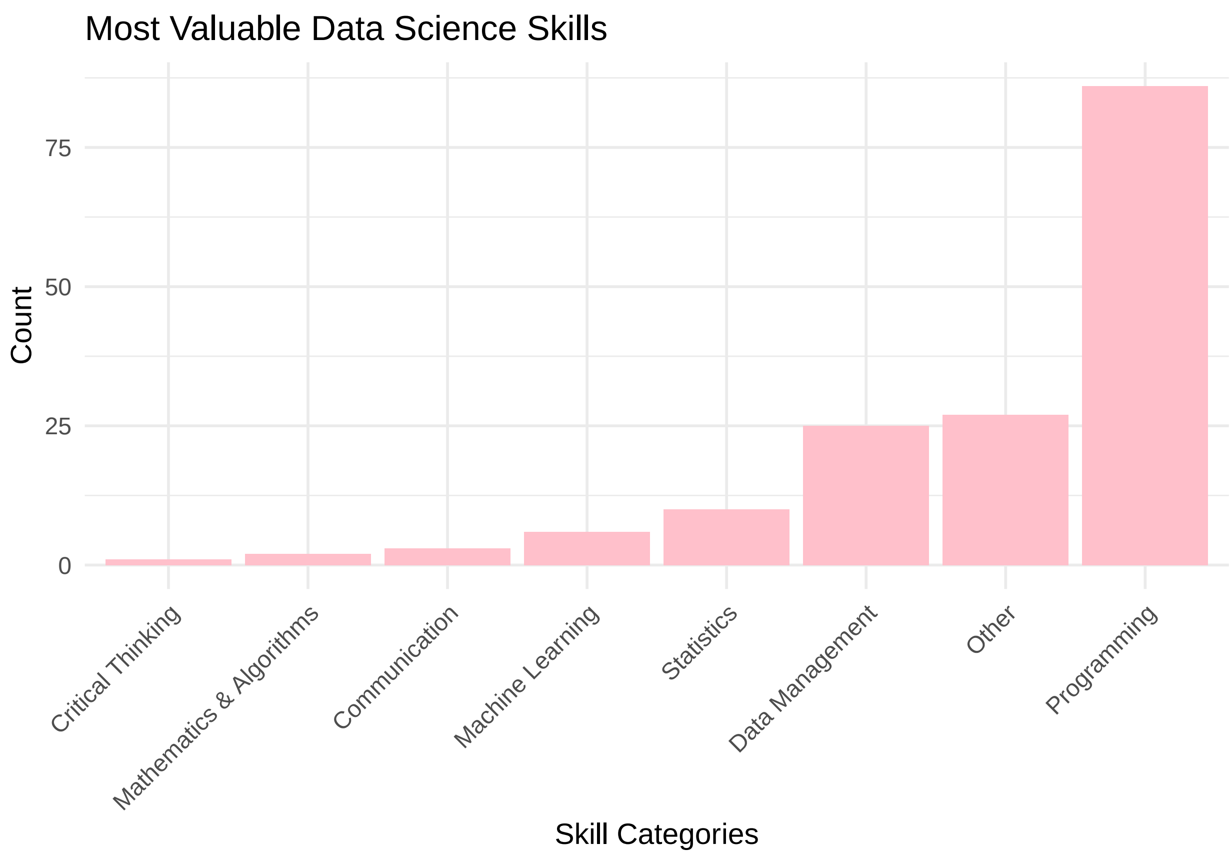


Let's make it look cleaner:

```r
# Group similar skills together using regex grepl
valuable_skills <- valuable_skills %>%
  mutate(skill_group = case_when(
    grepl("machine learning|ML", skill_name, ignore.case = TRUE) ~ "Machine Learning",
    grepl("programming|coding|software|Python|R", skill_name, ignore.case = TRUE) ~ "Programming",
    grepl("statistics|statistical", skill_name, ignore.case = TRUE) ~ "Statistics",
    grepl("data|database", skill_name, ignore.case = TRUE) ~ "Data Management",
    grepl("math|algorithm", skill_name, ignore.case = TRUE) ~ "Mathematics & Algorithms",
    grepl("communication|presentation", skill_name, ignore.case = TRUE) ~ "Communication",
    grepl("teamwork|collaborat", skill_name, ignore.case = TRUE) ~ "Teamwork",
    grepl("critical thinking|problem solving|analysis", skill_name, ignore.case = TRUE) ~ "Critical Thi
    TRUE ~ "Other"
  ))

skill_counts <- valuable_skills %>%
  count(skill_group, sort = TRUE)

ggplot(skill_counts, aes(x = reorder(skill_group, n), y = n)) +
  geom_bar(stat = "identity", fill = "pink") +
  labs(title = "Most Valuable Data Science Skills", x = "Skill Categories", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Now, we can compare our results with the findings from the Data Science Skills Survey 2022 by AIM and Great Learning, available at this link.

Below is a table from that website showing the common skills sought by recruiters across different years of experience. The aim is to determine if the skills identified in our internal survey align with those highlighted by recruiters on the website. We will be calculation the mean of the years of experience as we did not request this information on our survey.

```
common_skills_years <- read.csv("https://raw.githubusercontent.com/ZanetaP02/DATA-607/refs/heads/main/Da
common_skills_years
```

```
##                               skills less.than.3.years X3.5.years X6.10.years
## 1                   machine learning              81.9       86.8        79.3
## 2                         statistics              79.2       77.4        79.3
## 3               communication skills              68.1       73.6        75.9
## 4              programming knowledge              66.7       75.5        69.0
## 5                 data visualisation              61.1       66.0        62.1
## 6  data wrangling and pre-processing              62.5       49.1        51.7
## 7                    business acumen              37.5       45.3        41.4
## 8                      deep learning              50.0       43.4        51.7
## 9                presentation skills              34.7       37.7        41.4
## 10                  domain expertise              38.9       32.1        27.6
## 11          linear algebra & calculus              33.3       37.7        48.3
## 12                  model deployment              22.2       18.9        20.7
## 13                          big data              36.1       35.8        44.8
##    X10.years
```

```
## 1        92.3
## 2        80.8
## 3        80.8
## 4        69.2
## 5        57.7
## 6        53.8
## 7        38.5
## 8        46.2
## 9        38.5
## 10       42.3
## 11       34.6
## 12       30.8
## 13       30.8
```

To streamline data manipulation, first rename the columns for easier reference, then consolidate the different years of experience into a single column.

```r
up_csy <- common_skills_years

colnames(up_csy)[1] <- "skill_group"
colnames(up_csy)[2] <- "less than 3yrs"
colnames(up_csy)[3] <- "3-5yrs"
colnames(up_csy)[4] <- "6-10yrs"
colnames(up_csy)[5] <- "10yrs"

# Print column names to verify
print(colnames(up_csy))
```

```
## [1] "skill_group"    "less than 3yrs" "3-5yrs"         "6-10yrs"
## [5] "10yrs"
```

```r
up_csy1 <- up_csy %>%
  pivot_longer(cols = c('less than 3yrs', '3-5yrs', '6-10yrs', '10yrs'), names_to = "year_experience", 

head(up_csy1)
```

```
## # A tibble: 6 x 3
##   skill_group      year_experience percentage
##   <chr>            <chr>                <dbl>
## 1 machine learning less than 3yrs        81.9
## 2 machine learning 3-5yrs                86.8
## 3 machine learning 6-10yrs               79.3
## 4 machine learning 10yrs                 92.3
## 5 statistics       less than 3yrs        79.2
## 6 statistics       3-5yrs                77.4
```

Calculating the mean of the skills

```r
avg_csy <- up_csy1 %>% group_by(skill_group) %>%
  summarise(mean_percentage=mean(percentage),
            .groups = 'drop')
avg_csy
```

```
## # A tibble: 13 x 2
##    skill_group                        mean_percentage
##    <chr>                                        <dbl>
##  1 big data                                      36.9
##  2 business acumen                               40.7
##  3 communication skills                          74.6
##  4 data visualisation                            61.7
##  5 data wrangling and pre-processing             54.3
##  6 deep learning                                 47.8
##  7 domain expertise                              35.2
##  8 linear algebra & calculus                     38.5
##  9 machine learning                              85.1
## 10 model deployment                              23.2
## 11 presentation skills                           38.1
## 12 programming knowledge                         70.1
## 13 statistics                                    79.2
```

Grouping skills to match internal survey

```
avg_csy1 <- avg_csy %>%
 mutate(skill_group = case_when(
   grepl("communication skills", skill_group, ignore.case =TRUE) ~ "Communication",
   grepl("programming knowledge", skill_group, ignore.case =TRUE) ~ "Programming",
   grepl("statistics",  skill_group, ignore.case =TRUE) ~ "Statistics",
   grepl("machine learning",  skill_group, ignore.case =TRUE) ~ "Machine Learning",
   grepl("linear algebra & calculus",  skill_group, ignore.case =TRUE) ~ "Mathematics & Algorithms",
   grepl("big data",  skill_group, ignore.case =TRUE) ~ "Data Management",
   grepl("deep learning",  skill_group, ignore.case =TRUE) ~ "Critical Thinking",
   grepl("data visualisation|business acumen|presentation skills|model deployment|data wrangling and p
  ))
avg_csy1
```

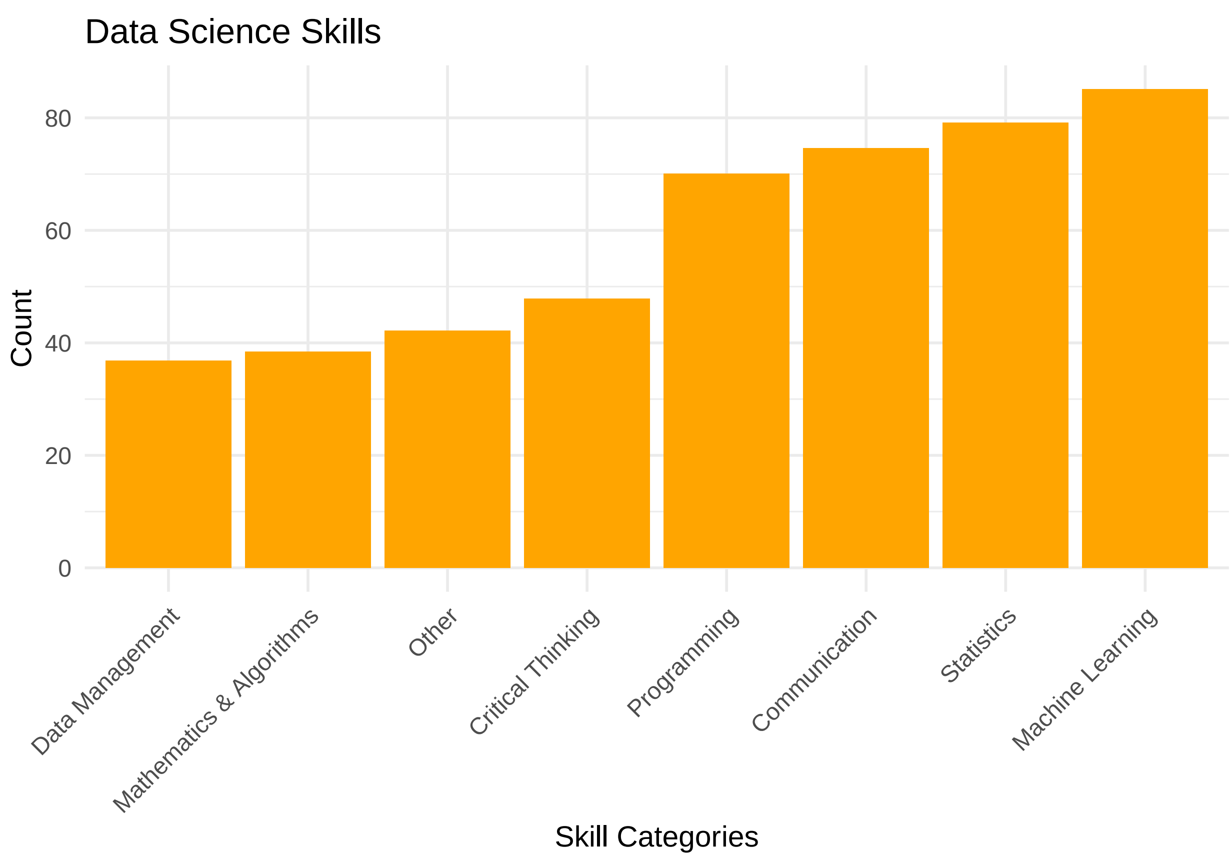
```
## # A tibble: 13 x 2
##    skill_group              mean_percentage
##    <chr>                              <dbl>
##  1 Data Management                     36.9
##  2 Other                               40.7
##  3 Communication                       74.6
##  4 Other                               61.7
##  5 Other                               54.3
##  6 Critical Thinking                   47.8
##  7 Other                               35.2
##  8 Mathematics & Algorithms            38.5
##  9 Machine Learning                    85.1
## 10 Other                               23.2
## 11 Other                               38.1
## 12 Programming                         70.1
## 13 Statistics                          79.2
```

Calculating mean of grouping skills to match internal survey skills

```r
avg_skills <- avg_csy1 %>% group_by(skill_group) %>%
  summarise(mean_percentage=mean(mean_percentage),
            .groups = 'drop')
avg_skills
```

```
## # A tibble: 8 x 2
##   skill_group               mean_percentage
##   <chr>                               <dbl>
## 1 Communication                        74.6
## 2 Critical Thinking                    47.8
## 3 Data Management                      36.9
## 4 Machine Learning                     85.1
## 5 Mathematics & Algorithms             38.5
## 6 Other                                42.2
## 7 Programming                          70.1
## 8 Statistics                           79.2
```

```r
ggplot(avg_skills, aes(x = reorder(skill_group, mean_percentage), y = mean_percentage)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Data Science Skills", x = "Skill Categories", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Now, let's align our survey skills data with the format used on the website by converting the survey results into percentages.

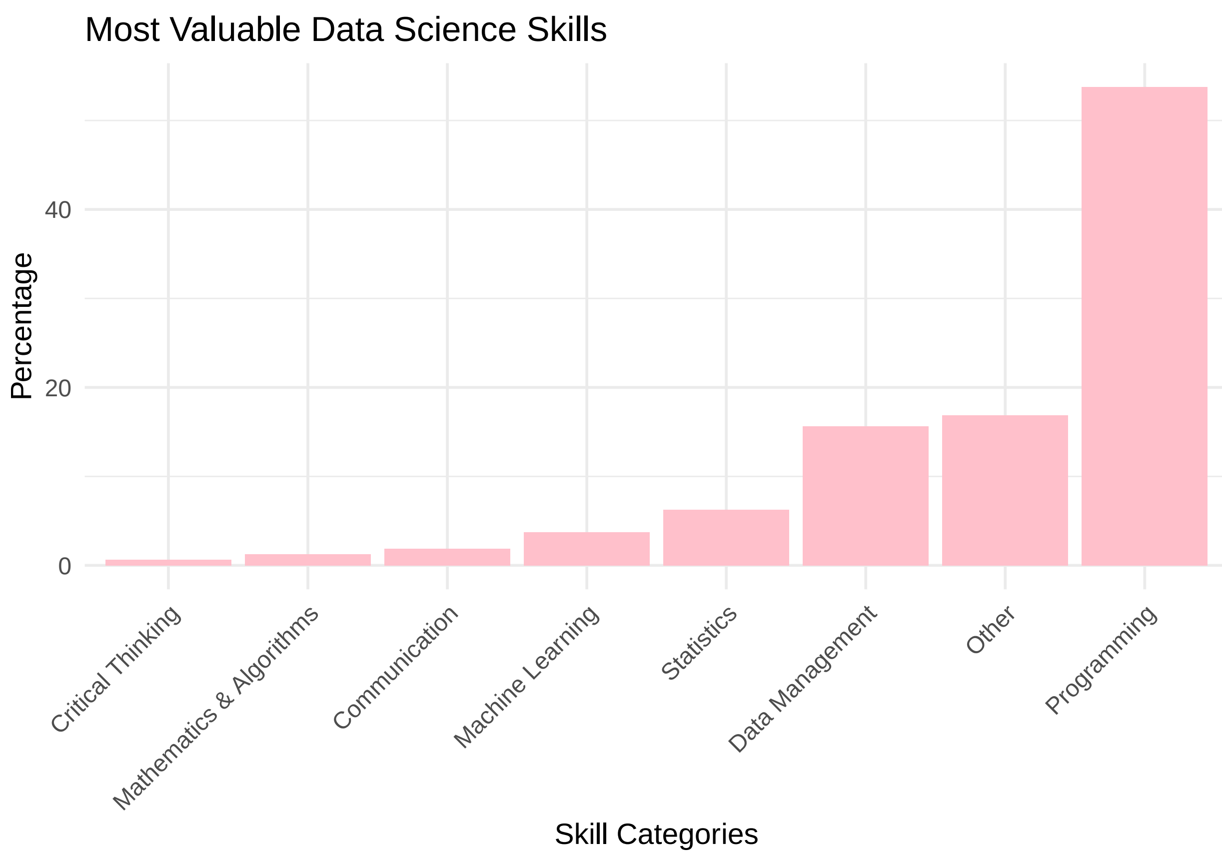
```r
total_count <- sum(skill_counts$n)
total_count
```

```
## [1] 160
```

```r
skill_counts <- skill_counts %>%
  mutate(percentage = (n / total_count) * 100)
skill_counts
```

```
##                skill_group  n percentage
## 1              Programming 86     53.750
## 2                    Other 27     16.875
## 3          Data Management 25     15.625
## 4               Statistics 10      6.250
## 5         Machine Learning  6      3.750
## 6            Communication  3      1.875
## 7 Mathematics & Algorithms  2      1.250
## 8        Critical Thinking  1      0.625
```

```r
ggplot(skill_counts, aes(x = reorder(skill_group, percentage), y = percentage)) +
  geom_bar(stat = "identity", fill = "pink") +
  labs(title = "Most Valuable Data Science Skills", x = "Skill Categories", y = "Percentage") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Let's combine the two datasets based on the percentages of the most valued skills.

```
merged_skill_counts <- avg_skills %>%
  left_join(skill_counts, by = "skill_group")
merged_skill_counts
```

```
## # A tibble: 8 x 4
##   skill_group              mean_percentage     n percentage
##   <chr>                              <dbl> <int>      <dbl>
## 1 Communication                       74.6     3      1.88
## 2 Critical Thinking                   47.8     1      0.625
## 3 Data Management                     36.9    25     15.6
## 4 Machine Learning                    85.1     6      3.75
## 5 Mathematics & Algorithms            38.5     2      1.25
## 6 Other                               42.2    27     16.9
## 7 Programming                         70.1    86     53.8
## 8 Statistics                          79.2    10      6.25
```

```
merged_skill_counts <- merged_skill_counts %>%
  select(-n) %>%
  rename(
    website = mean_percentage,
    survey = percentage
  )

merged_skill_counts
```

```
## # A tibble: 8 x 3
##   skill_group              website survey
##   <chr>                      <dbl>  <dbl>
## 1 Communication               74.6  1.88
## 2 Critical Thinking           47.8  0.625
## 3 Data Management             36.9 15.6
## 4 Machine Learning            85.1  3.75
## 5 Mathematics & Algorithms    38.5  1.25
## 6 Other                       42.2 16.9
## 7 Programming                 70.1 53.8
## 8 Statistics                  79.2  6.25
```

Convert into the long format:

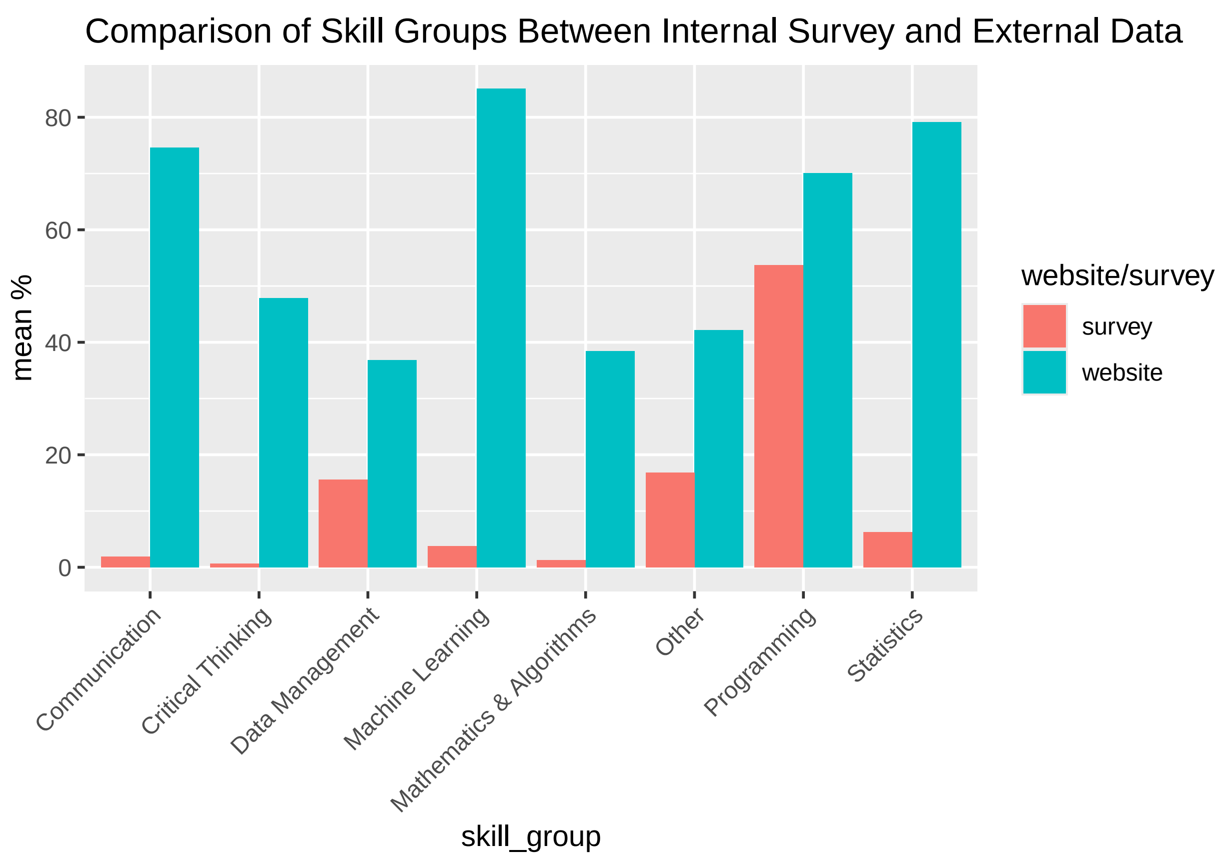
```
in_ex_skills <- merged_skill_counts %>%
  pivot_longer(cols = c('website', 'survey'), names_to = "website/survey", values_to = "mean %")
in_ex_skills
```

```
## # A tibble: 16 x 3
##    skill_group              `website/survey` `mean %`
##    <chr>                    <chr>               <dbl>
##  1 Communication            website            74.6
##  2 Communication            survey              1.88
##  3 Critical Thinking        website            47.8
##  4 Critical Thinking        survey              0.625
```

```
##   5 Data Management           website          36.9
##   6 Data Management           survey           15.6
##   7 Machine Learning          website          85.1
##   8 Machine Learning          survey            3.75
##   9 Mathematics & Algorithms website          38.5
##  10 Mathematics & Algorithms survey            1.25
##  11 Other                     website          42.2
##  12 Other                     survey           16.9
##  13 Programming               website          70.1
##  14 Programming               survey           53.8
##  15 Statistics                website          79.2
##  16 Statistics                survey            6.25
```

Plot the differences:

```
ggplot(in_ex_skills, aes(x = skill_group, y = `mean %`, fill = `website/survey`)) +
  geom_col(position = position_dodge()) +
  ggtitle("Comparison of Skill Groups Between Internal Survey and External Data") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion

Differences: When comparing our internal survey with external data, some differences stand out. For example, our respondents rated programming and statistics higher than recruiters did. This might be because our group focuses more on technical skills needed for data work. On the other hand, recruiters emphasized communication and critical thinking more. This suggests that employers want candidates who can not only handle technical tasks but also explain their insights and solve problems strategically. This difference might show that data scientists don't always realize how important these soft skills are for working with teams and explaining their work to non-technical people.

Commonalities: Despite these differences, there are skills that both our survey and the external data agree on. Skills like data management and machine learning are valued by both groups. This shows that these skills are seen as essential in the data science field. Both respondents and recruiters recognize these as key abilities because they are crucial for handling and analyzing large datasets, which is a core part of data science work. This agreement highlights that both sides understand the importance of these technical skills.

Interpretation of Differences and Commonalities: The differences likely come from different views on what a data scientist's role should be. Recruiters might prioritize communication and critical thinking because these skills help with teamwork and making strategic business decisions. In contrast, data scientists might see technical skills as more important because they focus on solving technical problems and analyzing data. However, both sides agree on the importance of data management and machine learning, showing that technical proficiency is essential for success in the field.