Data Professional Survey

Aly Eltokatly

2023-12-15

Data Profissional Survey

Introduction

This project aims to conduct a thorough analysis Using **R Programming Language** of various aspects of the data profession, taking into account market rates based on categorical factors such as industry, education, role, gender, age, race, and country. The dataset used in this project is a data frame comprising results from a survey administered to professionals in the field of Data Science worldwide. The survey successfully obtained responses from **630** participants. Through this analysis, the project endeavors to provide valuable insights into the prevailing trends and preferences within the data profession.

Report Deliverable:

- Clear statement of the business task.
- Description of all data sources used.
- Documentation for processing of cleaning or manipulation of data.
- Visualizations and Key Findings: Supporting visualizations and key findings.
- Recommendations: Top three recommendations based on the analysis.

Business Task

Since the the Data professional career is trending right now, the stakeholder need to determine the following.

- What is the average salary per each data professional?
- Which programming language is most commonly used by data professionals?
- What is the average salary for data analysts in each country that took part in this survey?

Preparing and Data Source.

Will preparing and setting up my **R** environment by loading the ‘tidyverse’ and ‘data_profissional_survey’ packages.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(treemapify)
data_profissinal_survey <- read.csv("data_profissinal_survey.csv")
```

Processing and Cleaning Data.

I used Rstudio to clean and manipulate the data. Upon exploring the data. I backup of the original dataset before proceeding with further restructuring. And I noticed the following.

- Given the diverse input in the survey data, I will need to standardize it by identifying keywords and categorizing responses based on shared attributes.
- I excluded certain columns from consideration as they contained minimal data and will not be relevant to my analysis.
- I renamed some variables headers that were too long and inefficient.
- Interested in determining the number of participants who took part in this survey.

```
length(data_profissinal_survey$ID)
```

```
## [1] 630
```

- Convert the data type of variable [current_.age] from integer to numeric.

```
data_profissinal_survey$current_.age <- as.numeric(data_profissinal_survey$current_.age)
```

- Calculate the mean of [current_.age].

```
mean(data_profissinal_survey$current_.age)
```

```
## [1] 29.86667
```

- Clean the variable [salary] by removing the extra symbol and blank space, then will separate into 2 variables [low_salary] and [high_salary], and after that will change the type of data for new variables to numeric in order to calculate the mean,

```
data_profissinal_survey$salery <- gsub("k","",data_profissinal_survey$salery)
```

```
data_profissinal_survey <- separate(data_profissinal_survey, col = salery,
                                      into = c("low_salary","high_salary",
                                               sep = "-"))
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 630 rows [1, 2, 3, 4, 5,
## 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
data_profissinal_survey$low_salary <- as.numeric(data_profissinal_survey$low_salary)
  data_profissinal_survey$high_salary <- as.numeric(data_profissinal_survey$high_salary)
```

- change the value of the low_salary which equal “0” to be “40” in order to get the right average salary.

```
data_profissinal_survey$low_salary[data_profissinal_survey$low_salary == 0]<- 40
```

-In [high-salary] replace NA value with new value “225” in order to keep consistency.

```
data_profissinal_survey<- data_profissinal_survey %>%
  mutate(high_salary = replace_na(high_salary,225)) %>%
  mutate(average_salary = (low_salary + high_salary)/2) %>%
  view()
```

- Clean the variable [title] by exclude the text after ’other” to a new column.

```
data_profissinal_survey_v1<-  data_profissinal_survey %>%
 separate(col = titel, into = c("title", "dep", "other"), sep = " ") %>%
 view()
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 78 rows [4, 36, 37, 50, 56,
## 58, 62, 69, 74, 76, 90, 94, 97, 103, 115, 127, 129, 143, 146, 150, ...].
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 542 rows [1, 2, 3, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ...].
```

- replace the observation in variable [title].

```
data_profissinal_survey_v1$title <- recode(data_profissinal_survey_v1$title,
                                        "Student/Looking/None" = "Student")
```

- Remove the text string in variable [dep] in order to unite with variable [title]

```
data_profissinal_survey_v1<- data_profissinal_survey_v1 %>%
  mutate(dep = if_else(title == "Other"," ", (dep))) %>%
  unite(col = position, title, dep, sep = " ") %>%
  view()
```

- clean the variable [position] by replace NA.

```
data_profissinal_survey_v1$position <- recode(data_profissinal_survey_v1$position,
                                        "Student NA" = "Student")

data_profissinal_survey_v1$position <- recode(data_profissinal_survey_v1$position,
                                           "Other  " = "Other")
```

- Clean [lang_1,lang_2, lang_3].

```
data_profissinal_survey_v4<- data_profissinal_survey_v1 %>%
  separate(col= programming_language,into = c("lang_1","lang_2","lang_3")
           ,sep = ":") %>%
  view()
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): input string 22 is invalid UTF-8
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): input string 59 is invalid UTF-8
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 630 rows [1, 2, 3, 4, 5,
## 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
data_profissinal_survey_v4$lang_2 <- substr(data_profissinal_survey_v4$lang_2,1,3)

data_profissinal_survey_v4 %>%
 mutate(lang_2 = str_to_upper(lang_2)) %>%
  View()
data_profissinal_survey_v4 %>%
  mutate(lang_1 = if_else(lang_1 == "Other",
                          if_else(lang_2 == "SQL","SQL",(lang_1)),lang_1)) %>%
  view()
data_profissinal_survey_v4<- data_profissinal_survey_v4 %>%
    mutate(lang_1 = if_else
           (lang_1 == "Other:Mostly use sql but that\x92s not programming language..",
             "Other",(lang_1))) %>%
    mutate(lang_1 = if_else(lang_1 == "Other:I don\x92t know any",
                             "Other", (lang_1))) %>%
    view()
```

- Clean variable [country]

```
data_profissinal_survey_v4$country[data_profissinal_survey_v4$country == "United States"]<- "US
A"
 data_profissinal_survey_v4$country[data_profissinal_survey_v4$country == "United Kingdom"]<-"U
K"
 data_profissinal_survey_v4$country[data_profissinal_survey_v4$country ==
                                   "Other (Please Specify):Per\xfa "]<- "Other"
 data_profissinal_survey_v4<- data_profissinal_survey_v4 %>%
   separate(col = country, into = c("country", "coun_2"),sep = " ") %>%
    view()
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 223 rows [3, 6, 7, 12, 13,
## 17, 20, 21, 23, 24, 25, 30, 36, 37, 41, 43, 44, 45, 52, 54, ...].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 407 rows [1, 2, 4, 5, 8,
## 9, 10, 11, 14, 15, 16, 18, 19, 22, 26, 27, 28, 29, 31, 32, ...].
```

Data Analysis

Overview

In this analysis, I examine the contribution of 630 data professional from different countries with **median age** 29 year.

Average Salary Per Titel

```
ave_salary_per_position <-data_profissinal_survey_v4 %>%
  group_by(position) %>%
  summarise("mean_salary" = round(mean(average_salary),2)) %>%
  View()
```

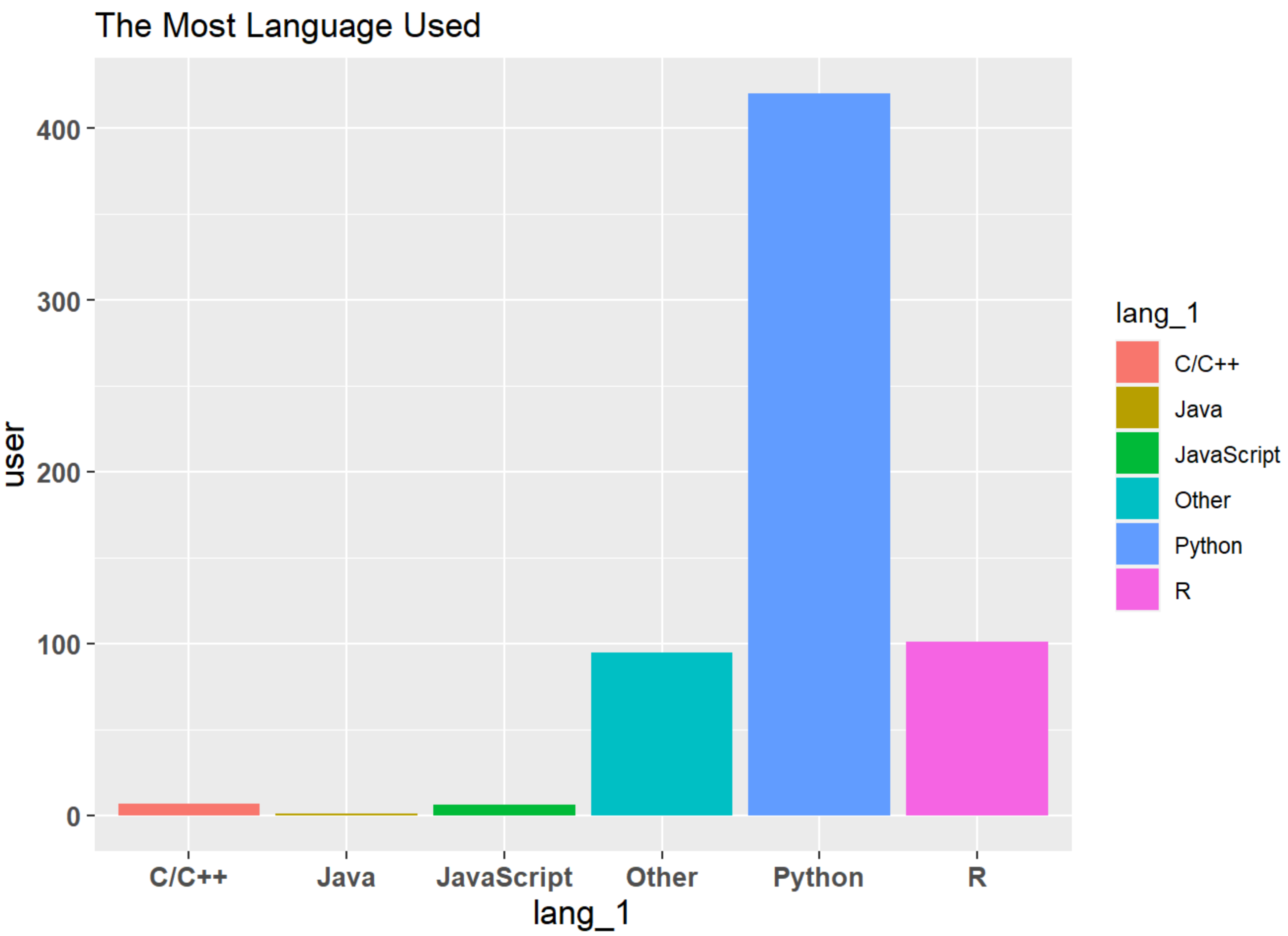
programming language is most commonly used by data professionals

I this analysis I found the most commonly uses language is “Python” followed by “R” language.

```
used_language <-data_profissinal_survey_v4 %>%
  group_by(lang_1) %>%
    summarise(user = n()) %>%
    view()

used_language<- used_language %>%
  arrange(user) %>%
  view()

ggplot(data = used_language)+
  geom_col(mapping = aes(x = lang_1, y = user, fill = lang_1))+
  labs(title = "The Most Language Used")+
  theme(axis.text = element_text(size = 10, face = "bold"))+
  theme(axis.title = element_text(size = 13))
```



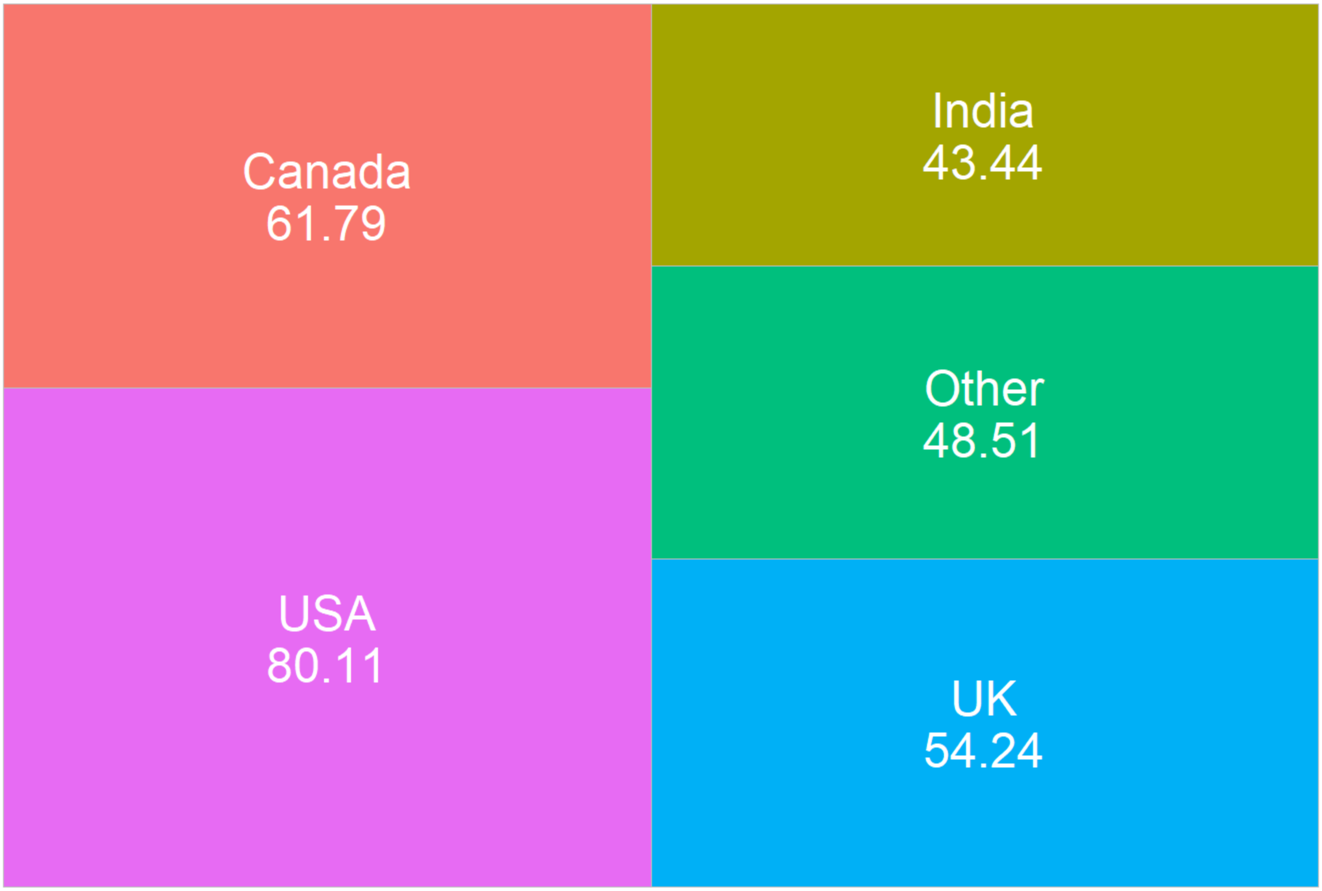
the average salary for Data Analysts in each country that took part in this survey.

I found that USA pay more than other countries followed By Canada then UK, and I think the reason behind that could be. - Size of the USA market and and huge volume of sportiness there. - living cost in the western countries is higher more than others.

```
data_analyst_salary <- data_profissinal_survey_v4 %>%
  select(position,country,average_salary) %>%
  filter(position == "Data Analyst") %>%
  group_by(country) %>%
  summarise(mean_salary = round(mean(average_salary),2)) %>%
    arrange(-mean_salary) %>%
  view()

data_analyst_salary %>%
  ggplot(aes(fill = country,area = mean_salary,
             label = paste0(country,"\n",mean_salary)))+
  geom_treemap()+
  geom_treemap_text(color = "white", place = "centre")+
  theme(legend.position = "none")+
  labs(title = "Data Analyst Ave Salary per Country")
```

Data Analyst Ave Salary per Country


Canada
USA
India
Other
UK