Assignment 1

Sanduni Silva

2023-05-13

Introduction

The objective of this data-wrangling project is to gather relevant information from the Ministry of Higher Education website and apply data modification to generate a data frame that shows the universities and the courses they offer together with the enrollment figures for each degree program.

We can recognize patterns and trends in the enrollment statistics for various courses and colleges by manipulating the data. This can aid in our understanding which university degree programs are the most desired by students, which colleges are most effective at attracting learners, and how enrollment trends have evolved over time. Additionally, this information might provide light on Sri Lanka's job market and demand for certain course specialties.

Steps of this process

1. Collect data
2. Clean and Organize collected data
3. Data Manipulation
4. Analyse data

Step 1 - Collect data

The website of Ministry of Education and excel files provided by the lecturer were used for this task.

Importing Libraries

```
library(rvest)
library(dplyr)
```

- `library(rvest)` package mainly used for web scrapping. Since we need to extract data from websites this package is needed.
- `library(dplyr)` package is mostly used for data manipulation & data wrangling. After clearing the collected data we need to this package to manipulate/ wrangle data.

Defining website link as a character variable

```
URL <- "https://www.mohe.gov.lk/index.php?option=com_courses&view=course_details&Itemid=225&lang=en#"

moe_link <- read_html(URL)
```

- `read_html()` function used to read the html file of the provided URL of Ministry of Education and save it as "moe_link"

Get data from relevant columns of the table in the website

```
Course <- moe_link %>%
  html_nodes("td:nth-child(1)") %>%
  html_text()

Institute <- moe_link %>%
  html_nodes("td:nth-child(3)") %>%
  html_text()
```

- These two codes are used to extract data from the 1st row to the last row of both Course & Institute columns.

Create a data frame for the data extracted from the MOE wbsite

```
df_uni <- data.frame(Institute,
                     Course,
                     stringsAsFactors = FALSE)

View(df_uni)
```

- `data.frame()` function used to create a data frame with two variables (Institute & Course).

Extract data from Excel

- The provided Book1 excel file were used for this data wrangling assignment. Since some of the data in that excel file were included as images first I used `tesseract::ocr()` and `cat()` functions to get the texts in those images. These two functions are R functions which do OCR. Since those images are blur/ low in quality, it didn't work. So, i used online OCR website (https://www.onlineocr.net/) to convert the image data into text.
- At the same time I did some manual edits in the excel file by creating one single table for with 3 columns (University, Course of Study, No - Intake) for provided Universities. This couldn't be done through R since some of the texts needed to be corrected such as spellings, etc. The new excel file named as Assignment 1 (https://docs.google.com/spreadsheets/d/1hBLO7Xd5ttcPTrA8o4pkIELWDjcftzGw/edit?usp=share_link&ouid=117499314796445237154&rtpof=true&sd=true)

```
library(readxl)
excel_data <- read_excel("D:\\Uni\\5 Semester\\Data Wrangling\\Assignmnet 1\\excel files\\Assignment 1.

View(excel_data)
```

- `library(readxl)` package provide `read_excel()` function which convert excel files into data frames. Using this new data frame called "excel_data" was created.

Step 2 - Clean and Organize collected data

Sort the data frame

```
df_uni_sorted <- df_uni[order(df_uni$Institute),]
head(df_uni_sorted)
```

```
##                          Institute
## 282 Eastern University, Sri Lanka
## 283 Eastern University, Sri Lanka
## 284 Eastern University, Sri Lanka
## 285 Eastern University, Sri Lanka
## 286 Eastern University, Sri Lanka
## 287 Eastern University, Sri Lanka
##                                                    Course
## 282    Degree of Bachelor of the Science of Agriculture \t
## 283                   Degree of Bachelor of Arts (General)
## 284  Degree of Bachelor of Arts Special in Drama & Theatre
## 285        Degree of Bachelor of Arts Special in Economics
## 286        Degree of Bachelor of Arts Special in Education
## 287        Degree of Bachelor of Arts Special in Fine Arts
```

- `order()` function in the above code is used to sort the data frame by institute in ascending order.

Create CSV file

```
write.csv(df_uni_sorted, "df_uni.csv")
```

`write.csv()` function is used to create csv files to reuse if necessary in future.

Remove NULL values

```
df_uni_clean <- data.frame(lapply(df_uni_sorted, function(x) {gsub("\n", "", x)}))

write.csv(df_uni_clean, "df_uni_clean.csv")
```

- A new data frame "df_uni_clean" with modified columns after removing null characters was created by using this code.
- After this step the new data frame is a corrected & organized data frame that can be used for further tasks.

Step 3 - Data Manipulation

Combine the data frames

- Previously created data frames (excel_data & df_uni_clean) are different from each other. "excel_data" data frame has 250 rows and "df_uni_clean" data frame has 544 rows.

```
nrow(excel_data)
```

```
## [1] 250
```

```
nrow(df_uni_clean)
```

```
## [1] 544
```

- And also these two data frames does not have a common column so we can't use `join()` , `cbind()` or `rbind()` so, using `merge()` function by merging the row names, the "final_df" data frame was created.

```
final_df <- merge(data.frame(df_uni_clean, row.names = NULL), data.frame(excel_data, row.names = NULL),

View(final_df)
```

Step 4 - Analyse data

Summerize data

```
summary(final_df$No...Intake)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.51   75.25  111.00  165.55  180.75 1200.00     294
```

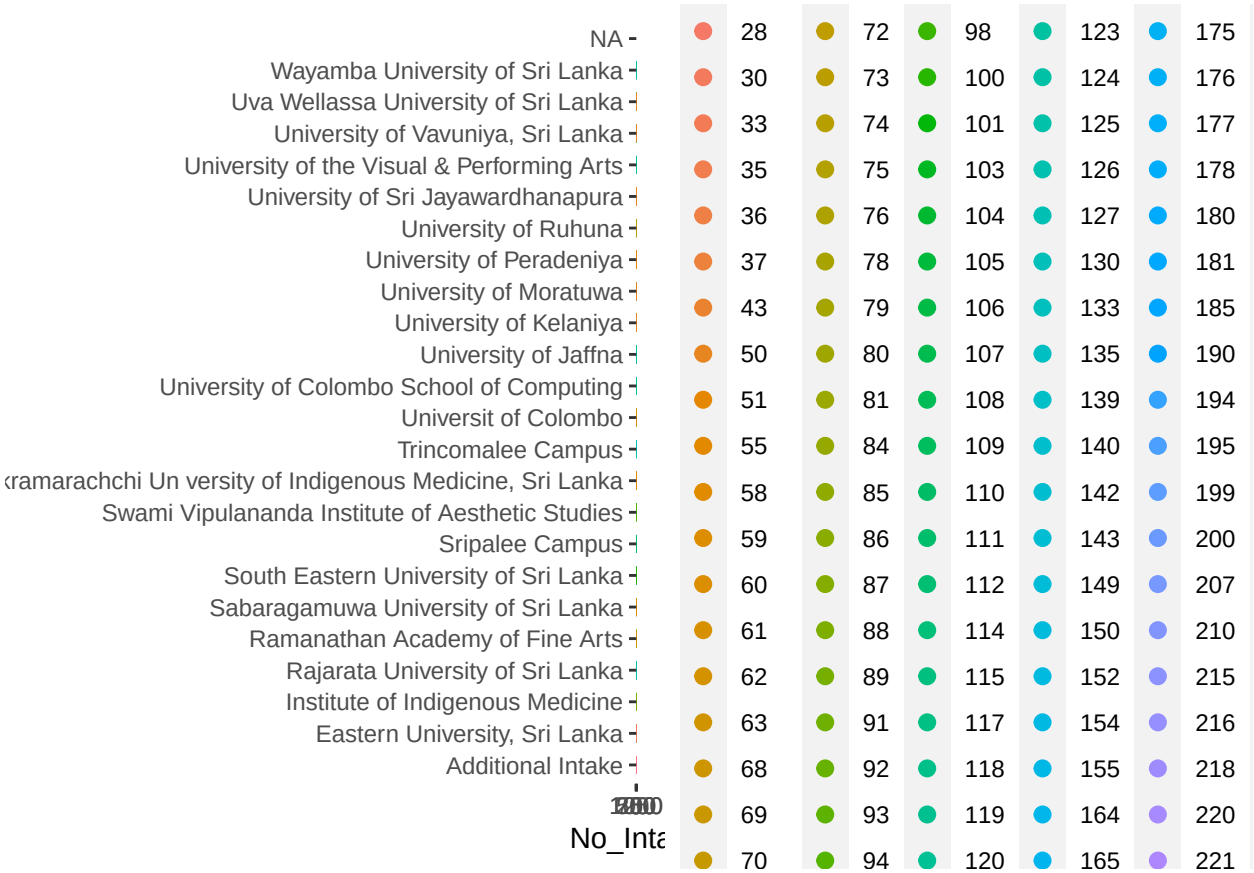
- Using `summary()` function we can take the summary statistics of variables. According to this output we can conclude that one average 166 students were taken to most of the causes.

Data Visualizations

```
library(ggplot2)

No_Intake <- final_df$No...Intake
uni <- final_df$University

Scatterplot <- ggplot(final_df, aes(x = No_Intake, y = uni, colour = factor(No_Intake)))+
  geom_point(size=2.5)
Scatterplot
```



- This code is to create a scatter plot between no. of students in an intake & universities. `ggplot()` function used to create this chart. Each dot in this chart represts a course and as it shows University of Kelaniya has the highest no of student per intake for a course.

Conclusion

This data-wrangling task involved gathering pertinent data from the Ministry of Higher Education website, manipulating the data to generate a data frame that indicates the institutions of higher learning and the courses they offer and the number of intakes for each degree program, and then understanding the data to get insights into the universities and courses offered.

Based on our data, we discovered that most courses took 166 students on average, with the University of Kelaniya having the highest intake rates. Additionally, the majority of university courses enroll between 10 and 250 students each intake.