Assignment\_AP

Phan Minh Trung

2025-06-03

Report for assignment

Author: Phan Minh Trung

Student id: 22117538.

By including this statement, we the authors of this work, verify that:

• We hold a copy of this assignment that we can produce if the original is lost or damaged.

• We hereby certify that no part of this assignment/product has been copied from any other student’s work or from any other source except where due acknowledgement is made in the assignment.

• No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

• We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).

• We hereby certify that we have read and understand what the School of Computing, Engineering and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

# Introduction

In this day and age, understanding user behavior and engagement patterns is significant in developing a digital platforms and constructing a meaningful user experiences. By tasks provided below, this report provides an analytical perspective of user activity within an online community, based on real-world data comprising user profiles, business listings, and their reviews. By applying data science workflow, the primary objective of this project is to gather insights from these datasets- users, business, and reviews using RStudio as a major software and language for analysis .

# Pre-process data procedures

Before getting into analysis, the first and foremost step in analysis is loading, getting sense and cleaning datasets.

## Load the data

The provided datasets are formatted as csv files. Therefore we use read.csv built in the Rstudio

library(readr)  
#Load data  
data\_users=read.csv("users.csv")  
data\_bs=read.csv("businesses.csv")  
data\_rv=read\_csv("reviews.csv",show\_col\_types = FALSE)

#First ten rows of users dataset  
head(data\_users,10)

## user\_id name review\_count average\_stars member\_since  
## 1 u\_0 Alan 32 2.08 2019-04-05  
## 2 u\_1 Joel 90 1.97 2015-11-15  
## 3 u\_2 Claire 93 1.10 2021-10-05  
## 4 u\_3 Samantha 59 3.01 2017-05-15  
## 5 u\_4 Monique 42 4.44 2021-04-05  
## 6 u\_5 Lucas 62 1.63 2023-09-16  
## 7 u\_6 Cole 19 3.37 2015-07-25  
## 8 u\_7 Christopher 93 3.88   
## 9 u\_8 Randy 35 2.47   
## 10 u\_9 Latoya 76 3.81 2015-05-15

We witnessed data\_users with 38801 observations, 5 fields namely user\_id, name, review\_count, average\_stars, and member\_since

#Structure and summary of the dataset  
summary(data\_users)

## user\_id name review\_count average\_stars   
## Length:38801 Length:38801 Min. : 1.00 Min. :1.000   
## Class :character Class :character 1st Qu.:25.00 1st Qu.:2.000   
## Mode :character Mode :character Median :50.00 Median :3.000   
## Mean :49.93 Mean :2.998   
## 3rd Qu.:75.00 3rd Qu.:3.990   
## Max. :99.00 Max. :5.000   
## member\_since   
## Length:38801   
## Class :character   
## Mode :character   
##   
##   
##

dim(data\_users)

## [1] 38801 5

#First ten rows of users dataset  
head(data\_bs,10)

## business\_id name city state  
## 1 b\_0 Steele, Hampton and Odonnell Michaelbury NV  
## 2 b\_1 Kim, Andrews and Joyce East Susan KY  
## 3 b\_2 Simmons PLC New Tamara PA  
## 4 b\_3 Noble-Murphy Patrickhaven CA  
## 5 b\_4 East Thomasshire GA  
## 6 b\_5 Dean, Martin and Grant Bakerberg DC  
## 7 b\_6 Simpson LLC Gomezmouth OR  
## 8 b\_7 Lee PLC Jenniferchester MD  
## 9 b\_8 Griffin Inc Vargasfurt WI  
## 10 b\_9 Harper, Faulkner and Smith West Michael IA  
## business.avg.stars review\_count categories business\_group  
## 1 2.5 351 anything, week, if A  
## 2 4.8 267 right A  
## 3 3.9 397 establish A  
## 4 3.4 54 right, campaign, so A  
## 5 1.6 278 hour, rest   
## 6 1.6 320 success B  
## 7 1.2 169 apply B  
## 8 4.5 287 always   
## 9 3.4 354 join, could, statement A  
## 10 3.8 484 executive B

We witnessed data\_bs with 19401 observations, 8 fields namely business\_id, name, city, state, business.avg.stars, and business\_group.

#Structure and summary of the dataset  
summary(data\_bs)

## business\_id name city state   
## Length:19401 Length:19401 Length:19401 Length:19401   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## business.avg.stars review\_count categories business\_group   
## Min. :1.000 Min. : 10.0 Length:19401 Length:19401   
## 1st Qu.:2.000 1st Qu.:130.0 Class :character Class :character   
## Median :3.000 Median :253.0 Mode :character Mode :character   
## Mean :2.994 Mean :253.9   
## 3rd Qu.:4.000 3rd Qu.:378.0   
## Max. :5.000 Max. :499.0

dim(data\_bs)

## [1] 19401 8

#First ten rows of users dataset  
head(data\_rv,10)

## # A tibble: 10 × 6  
## review\_id user\_id business\_id stars date text   
## <chr> <chr> <chr> <dbl> <date> <chr>   
## 1 r\_0 u\_11073 b\_4559 5 2023-02-01 Audience hour west television…  
## 2 r\_1 u\_35221 b\_10665 3 2023-03-12 Summer ability art beat race …  
## 3 r\_2 u\_3710 b\_7683 5 2025-02-19 Reason range future the chair…  
## 4 r\_3 u\_23891 b\_9113 3 2023-01-10 Up change final prepare area …  
## 5 r\_4 u\_10374 b\_7612 4 2023-01-02 Size pass including performan…  
## 6 r\_5 u\_30798 b\_5793 2 2022-08-21 Pm yeah laugh necessary else …  
## 7 r\_6 u\_24924 b\_8921 3 2025-01-23 Today loss experience account…  
## 8 r\_7 u\_4847 b\_16018 2 2025-04-10 Pull military color piece mat…  
## 9 r\_8 <NA> b\_6377 1 2024-04-24 Forward store Democrat turn a…  
## 10 r\_9 u\_21910 b\_9549 4 NA Day for participant increase …

We witnessed data\_rv with 19401 observations, 6 fields namely review\_id, user\_id, business\_id, stars, date, and text.

#Structure and summary of the dataset  
summary(data\_rv)

## review\_id user\_id business\_id stars   
## Length:194001 Length:194001 Length:194001 Min. :1   
## Class :character Class :character Class :character 1st Qu.:2   
## Mode :character Mode :character Mode :character Median :3   
## Mean :3   
## 3rd Qu.:4   
## Max. :5   
##   
## date text   
## Min. :2022-04-15 Length:194001   
## 1st Qu.:2023-01-15 Class :character   
## Median :2023-10-15 Mode :character   
## Mean :2023-10-15   
## 3rd Qu.:2024-07-15   
## Max. :2025-04-14   
## NA's :5819

dim(data\_rv)

## [1] 194001 6

str(data\_rv$date)

## Date[1:194001], format: "2023-02-01" "2023-03-12" "2025-02-19" "2023-01-10" "2023-01-02" ...

# Load libraries

After understanding tasks included in this reports which are mentioned later, we needs to load necessary librabries for analysis including tidyverse,ggplot2, and kableExtra

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ purrr 1.0.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ── 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 become errors

library(ggplot2)  
library(kableExtra)

##   
## Attaching package: 'kableExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## group\_rows

# Clean datasets

There are steps in cleaning including reformating, removing duplicated observations and handling N/A. After having first glance of summary dataset. We can see that:

* Date in data\_user is misformated as character instead of date
* There are some N/A values in collunms of 3 of datasets- namely Name (formated as “” instead of N/A) in data\_bs, or Date in data\_rv and data\_users.
* Removing duplicated rows in datasets with its primary key (id).

One thing to notice that, the author after cleaning data still uses na.rm in this analysis to ensure droping N/A in the analysis.

#Assign NA value for "" in name for data\_bs  
data\_bs[data\_bs == ""] <- NA  
#Assign NA value for "" in name for data\_user  
data\_users[data\_users == ""] <- NA  
#Assign Na value for "" in text  
data\_rv[data\_rv == ""] <- NA  
#Format date for date in data\_users  
data\_users$member\_since=as.Date(data\_users$member\_since)  
  
# Remove all duplicated rows  
data\_users=data\_users[!duplicated(data\_users$user\_id), ]  
data\_bs=data\_bs[!duplicated(data\_bs$business\_id), ]  
data\_rv=data\_rv[!duplicated(data\_rv$review\_id), ]  
  
#Remove all N/A values in 3 datasets  
data\_users <- na.omit(data\_users)  
data\_bs <- na.omit(data\_bs)  
data\_rv <- na.omit(data\_rv)  
  
#Remove 2 collumns as the annoucement   
data\_users<- data\_users %>% select(-review\_count, -average\_stars)

At this stage, we check the dimension after cleaning:

#Check dimensions again for NA rows  
dim(data\_users)

## [1] 36518 3

dim(data\_bs)

## [1] 16661 8

dim(data\_rv)

## [1] 171767 6

After filtering out the N/A values in collumns of three datasets, it significant decreased numbers of rows for

* Data\_users from 38801 to 36518
* Data\_bs 194001 to 16661
* Data\_rv 194001 to 171767

#Analysis tasks

1. Write the code to analyse the review behaviour across user groups. The users should be grouped into 3 group: Veteran, Intermediate and New (based on their member since date) before 2017, between 2017-2022, and after 2022 respectively. Calculate the numbers of users, their average review stars and average number of reviews per user. Tabulate the data using kable or kableextra. Visualise the Average Review Stars by User Age Group. You are required to make sure you handle the NA value in your analysis. Explain your findings.

Solution:

Task 1.1: The users are grouped into 3 group: Veteran, Intermediate and New (based on their member since date) before 2017, between 2017-2022, and after 2022 respectively.

For this task, the author created a new column for the category which each user\_id belong to. Case\_when is used as conditional function for grouping.

#Create new dataframe from data of data\_users  
  
data\_users <- data\_users %>%  
  
 #Create a new column for data\_users for Veteran, Intermediate, and New by mutate  
 mutate(   
   
#The column is defined as users\_group  
 users\_group = case\_when(  
   
 #Case\_when to categorise based on member\_since of data\_users  
   
 #Case 1: Before 2017-01-01  
 data\_users$member\_since < as.Date("2017-01-01") ~ "Veteran",  
   
 #Case 2: Before 2022-12-31 and after 2017-01-01  
 data\_users$member\_since >= as.Date("2017-01-01") &data\_users$member\_since <= as.Date("2022-12-31") ~ "Intermediate",  
   
 #Case 3: After 2022-12-31  
 data\_users$member\_since > as.Date("2022-12-31") ~ "New",  
 #If data can not be categorised, assign NA  
 TRUE ~ NA\_character\_  
 )  
 )

Task 1.2: Calculate the numbers of users, their average review stars and average number of reviews per user.

For this task, since the count of reviews for each users is not available (as requested in the annoucement). Therefore, firstly, data\_rv and data\_users are joined to get the dataframe for counting number of reviews (review\_id) by users\_id, and assign that to the new dataframe. Then it can be joined back to the original user dataset. This dataset once is joined again with the data\_review to update the counts as a new column. Then we can use that for summarising the numbers of users, their average review stars and average number of reviews per user.

#Inner join valid user\_id in both tables   
joined\_data= data\_rv %>% inner\_join(data\_users, by="user\_id")  
head(joined\_data,10)

## # A tibble: 10 × 9  
## review\_id user\_id business\_id stars date text name member\_since  
## <chr> <chr> <chr> <dbl> <date> <chr> <chr> <date>   
## 1 r\_1 u\_35221 b\_10665 3 2023-03-12 Summer abi… Chri… 2020-10-18   
## 2 r\_2 u\_3710 b\_7683 5 2025-02-19 Reason ran… Rhon… 2020-01-08   
## 3 r\_3 u\_23891 b\_9113 3 2023-01-10 Up change … Erik 2021-11-27   
## 4 r\_4 u\_10374 b\_7612 4 2023-01-02 Size pass … Chri… 2018-01-02   
## 5 r\_5 u\_30798 b\_5793 2 2022-08-21 Pm yeah la… Dani… 2021-01-24   
## 6 r\_6 u\_24924 b\_8921 3 2025-01-23 Today loss… Rona… 2017-08-24   
## 7 r\_7 u\_4847 b\_16018 2 2025-04-10 Pull milit… Bren… 2016-03-08   
## 8 r\_11 u\_11140 b\_3606 5 2023-04-04 Leg head d… Mary 2016-06-04   
## 9 r\_12 u\_18297 b\_12747 2 2023-09-12 President … John 2020-09-12   
## 10 r\_13 u\_7012 b\_1571 3 2024-07-21 Ever view … Karen 2017-11-20   
## # ℹ 1 more variable: users\_group <chr>

#Create new dataframe for review counts  
user\_review\_counts <- joined\_data %>%  
 group\_by(user\_id) %>%  
 summarise(review\_count = n\_distinct(review\_id))  
  
data\_users <- data\_users %>%  
 left\_join(user\_review\_counts, by = "user\_id")  
head(data\_users,10)

## user\_id name member\_since users\_group review\_count  
## 1 u\_0 Alan 2019-04-05 Intermediate 3  
## 2 u\_1 Joel 2015-11-15 Veteran 7  
## 3 u\_2 Claire 2021-10-05 Intermediate 4  
## 4 u\_3 Samantha 2017-05-15 Intermediate 10  
## 5 u\_4 Monique 2021-04-05 Intermediate 3  
## 6 u\_5 Lucas 2023-09-16 New 8  
## 7 u\_6 Cole 2015-07-25 Veteran 6  
## 8 u\_9 Latoya 2015-05-15 Veteran 4  
## 9 u\_10 Robin 2019-02-25 Intermediate 10  
## 10 u\_11 Curtis 2020-01-08 Intermediate 5

#Rejoin data with the new collumn  
joined\_data= data\_rv %>% inner\_join(data\_users, by="user\_id")  
head(joined\_data,10)

## # A tibble: 10 × 10  
## review\_id user\_id business\_id stars date text name member\_since  
## <chr> <chr> <chr> <dbl> <date> <chr> <chr> <date>   
## 1 r\_1 u\_35221 b\_10665 3 2023-03-12 Summer abi… Chri… 2020-10-18   
## 2 r\_2 u\_3710 b\_7683 5 2025-02-19 Reason ran… Rhon… 2020-01-08   
## 3 r\_3 u\_23891 b\_9113 3 2023-01-10 Up change … Erik 2021-11-27   
## 4 r\_4 u\_10374 b\_7612 4 2023-01-02 Size pass … Chri… 2018-01-02   
## 5 r\_5 u\_30798 b\_5793 2 2022-08-21 Pm yeah la… Dani… 2021-01-24   
## 6 r\_6 u\_24924 b\_8921 3 2025-01-23 Today loss… Rona… 2017-08-24   
## 7 r\_7 u\_4847 b\_16018 2 2025-04-10 Pull milit… Bren… 2016-03-08   
## 8 r\_11 u\_11140 b\_3606 5 2023-04-04 Leg head d… Mary 2016-06-04   
## 9 r\_12 u\_18297 b\_12747 2 2023-09-12 President … John 2020-09-12   
## 10 r\_13 u\_7012 b\_1571 3 2024-07-21 Ever view … Karen 2017-11-20   
## # ℹ 2 more variables: users\_group <chr>, review\_count <int>

user\_stats <- joined\_data %>%group\_by(users\_group) %>%   
 #Use summarise to reduce data of data\_users into 1 row then return neeeded statistics as defined names like total\_users,avg\_review\_stars...  
 summarise(  
 # Total number of users in the first collumn  
 total\_users = n\_distinct(user\_id),  
   
 # Average review stars per user in the second collumn  
 avg\_review\_stars = mean(stars, na.rm = TRUE),  
   
 # Average number of reviews per user in the third collumn  
 avg\_reviews\_per\_user = mean(review\_count, na.rm = TRUE)  
 )

Task 1.3 Tabulate the data using kable or kableextra

As a personal preference, this author change the colnames of the dataframe, then choosing font of Camberia with the classic theme. However, due to unexpected kniting problems(using kbl to produce image as HTML instead of kable), the comments using kable for below tables shows the HTML version as an output.

#Change the name of columns for easy-reading  
  
colnames(user\_stats)=c("Users Types","Number of users", "Average review stars per user","Average number of reviews per user")  
user\_stats %>%   
 #Set the title   
 kable(caption = "Table for numbers of users and their reviews")

Table for numbers of users and their reviews

| Users Types | Number of users | Average review stars per user | Average number of reviews per user |
| --- | --- | --- | --- |
| Intermediate | 21700 | 2.997970 | 5.308745 |
| New | 8052 | 3.005696 | 5.305847 |
| Veteran | 6277 | 3.000623 | 5.265513 |

#Format the table, hover indicate the change of color in html,   
 #Full\_width=F expanding across paper,   
 #Hover using to change color in html/ webpage  
 #kable\_classic("hover",full\_width=F,html\_font = "Camberia")

As the summary of user\_stats, we can see that there is a big number of users from Intermediate with 21700, followed by New and Veteran users. There is no major difference in review behaviour in 3 groups considering average review stars per user, and average number of reviews

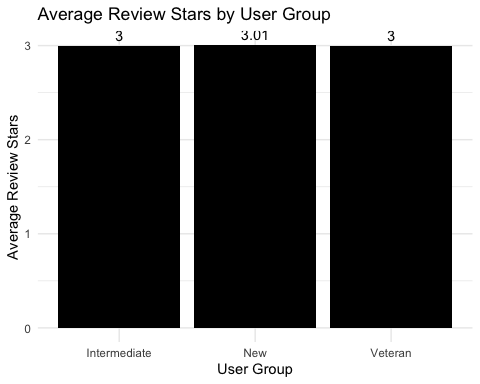
Task 1.4 Visualise the Average Review Stars by User Age Group.

We use ggplot as our visualization function. After choosing x and y axis, by using aes() or mapping it in geom\_col, we fill with black, and adjust title,lables, then choose theme.

#Draw graph by using ggplot  
summary(user\_stats)

## Users Types Number of users Average review stars per user  
## Length:3 Min. : 6277 Min. :2.998   
## Class :character 1st Qu.: 7164 1st Qu.:2.999   
## Mode :character Median : 8052 Median :3.001   
## Mean :12010 Mean :3.001   
## 3rd Qu.:14876 3rd Qu.:3.003   
## Max. :21700 Max. :3.006   
## Average number of reviews per user  
## Min. :5.266   
## 1st Qu.:5.286   
## Median :5.306   
## Mean :5.293   
## 3rd Qu.:5.307   
## Max. :5.309

ggplot(user\_stats, aes(x = `Users Types`, y =`Average review stars per user`)) +  
 geom\_col(fill = "black") +  
 geom\_text(aes(label = round(`Average review stars per user`, 2)), vjust = -0.5)+  
 labs(  
 title = "Average Review Stars by User Group",  
 x = "User Group",  
 y = "Average Review Stars"  
 ) +  
 theme\_minimal()



1. Write the code to analyse the average reviews star by State. Calculate the average review star, the number of reviews and the number of unique users. Visualise the Average Review Stars by State. You are required to make sure you take care of the NA value in your analysis. Elaborate on the findings.

Solution:

Task 2.1: The code to analyse the average reviews star by State

Since we are interested in analyzing the average reviews stars by State. There are 2 ways to doing that

1. We use 2 columns of available business.avg.stars and review\_count grouped by state for analyzing. As below, we can get sense by looking by the first 10 rows.

business\_stats <- data\_bs %>% group\_by(state) %>%   
 summarise(  
 # Total number of businesses  
 total\_business = n\_distinct(business\_id),  
   
 # Average review stars per business  
 avg\_review\_stars = mean(business.avg.stars, na.rm = TRUE),  
   
 # Number of reviews of all businesses  
 num\_reviews = sum(review\_count, na.rm = TRUE)  
 )  
head(business\_stats,10)

## # A tibble: 10 × 4  
## state total\_business avg\_review\_stars num\_reviews  
## <chr> <int> <dbl> <int>  
## 1 AK 336 3.00 87059  
## 2 AL 318 2.95 83434  
## 3 AR 342 2.90 84126  
## 4 AZ 332 3.01 85702  
## 5 CA 346 3.02 81643  
## 6 CO 326 3.00 86857  
## 7 CT 317 2.92 81857  
## 8 DC 341 3.06 87296  
## 9 DE 315 3.07 78763  
## 10 FL 316 2.98 78880

1. Join two table together then calculate the number\_reviews by distinct(review\_id). There are difference in dimension due to the N/A appeared in matching methods between these two tables. As we can see 3338 NA in the second methods

joined\_data= joined\_data %>% inner\_join(data\_bs, by="business\_id")  
business\_stats <- joined\_data %>% group\_by(state) %>%   
 summarise(  
 # Total number of businesses  
 total\_business = n\_distinct(business\_id),  
   
 # Average review stars per business  
 avg\_review\_stars = mean(stars, na.rm = TRUE),  
   
 # Number of reviews of all businesses  
 num\_reviews = n\_distinct(review\_id)  
 )  
business\_stats

## # A tibble: 51 × 4  
## state total\_business avg\_review\_stars num\_reviews  
## <chr> <int> <dbl> <int>  
## 1 AK 336 3.00 2628  
## 2 AL 318 3.03 2537  
## 3 AR 342 3.03 2695  
## 4 AZ 332 2.96 2578  
## 5 CA 346 3.00 2677  
## 6 CO 326 2.97 2533  
## 7 CT 316 3.03 2479  
## 8 DC 341 3.04 2715  
## 9 DE 315 2.99 2496  
## 10 FL 316 2.96 2463  
## # ℹ 41 more rows

Task 2.2 Calculate the average review star, the number of reviews and the number of unique users.

For this task, we join data\_bs with the reviews then calculate the distinct number of users with valid reviews. We can obtain the number of review with number of review\_id as well

# Add user count from a separate data\_users dataframe  
stat\_join\_table <- joined\_data %>%  
 summarise(  
 user\_count = n\_distinct(user\_id),  
 averge\_count= mean(stars,na.rm = TRUE),  
 number\_of\_review=n\_distinct(review\_id)  
 )  
  
# Rename columns  
colnames(stat\_join\_table)= c(  
 "Total number of unique users",  
 "Average review stars ",  
 "Total number of reviews"  
)  
# Create formatted table  
stat\_join\_table %>%  
 kable(caption = "Table: Summary of Total and Average Review Stars, and Total Reviews")

Table: Summary of Total and Average Review Stars, and Total Reviews

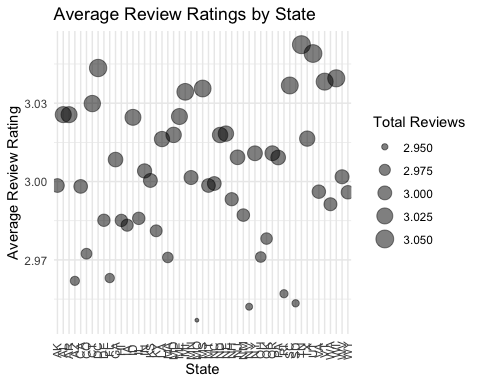
| Total number of unique users | Average review stars | Total number of reviews |
| --- | --- | --- |
| 35552 | 3.002 | 130985 |

#%>%  
 #kable\_classic("hover", full\_width = F, html\_font = "Cambria")

Task 2.3 Visualise the Average Review Stars by State

Once again, ggplot2 is used. However, by using both size of the points and its heights(y), the author uses that to demonstrate the average star of reviews by each state.

#Load dataset and choose x,y variables  
plot\_state =ggplot(business\_stats, aes(x = state, y = avg\_review\_stars,size = avg\_review\_stars))+  
 #Use scaterplot for this visualization with 0.5 in transparency  
 geom\_point(alpha=0.5)+  
 #Add title and labels for x and y  
 labs(  
 title = "Average Review Ratings by State",  
 x = "State",  
 y = "Average Review Rating",  
 color = "Avg Rating",  
 size = "Total Reviews"  
 ) +  
 #Select theme  
 theme\_minimal()+   
 #Adjust labels  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
print(plot\_state)



1. Write the code to analyse the top users and their behaviours. First, identify the top 10 users by the review count. For those top 10 users, calculate their average review stars. Tabulate the summary of the data (kable/kableextra). You are required to make sure you handle the NA value in your analysis. Visualise their rating distrubtion using ggplot2 - boxplot. Discuss your findings.

Task 3.1 First, identify the top 10 users by the review count.

By grouping user\_id and using joined\_data to calculate average stars as new column, we sorted dataset in descending order by review\_count , we then split the dataframe for top 10 by using slice and update column that average stars on data\_users.

top\_10= joined\_data %>% group\_by(user\_id) %>%summarise(  
 average\_stars=mean(stars,na.rm=TRUE),  
) %>% inner\_join(data\_users,by="user\_id") %>% arrange(desc(review\_count)) %>% slice(1:10)  
top\_10

## # A tibble: 10 × 6  
## user\_id average\_stars name member\_since users\_group review\_count  
## <chr> <dbl> <chr> <date> <chr> <int>  
## 1 u\_27070 3.08 Rebecca 2023-07-22 New 16  
## 2 u\_11551 2.8 Christopher 2021-12-17 Intermediate 14  
## 3 u\_22933 3 Stephanie 2018-10-03 Intermediate 14  
## 4 u\_27907 3.7 Jesse 2018-03-06 Intermediate 14  
## 5 u\_4617 2.8 Joel 2023-03-07 New 14  
## 6 u\_5497 2.25 Abigail 2017-11-04 Intermediate 14  
## 7 u\_6766 3.25 Tracy 2021-08-04 Intermediate 14  
## 8 u\_11137 3.22 Austin 2024-04-06 New 13  
## 9 u\_11229 3.25 Benjamin 2023-11-20 New 13  
## 10 u\_13638 3.56 Melanie 2018-12-04 Intermediate 13

Task 3.2 Tabulate the summary of the data (kable/kableextra).

top\_10 %>%   
 kable(caption = "Top 10 Most Active Users: Review Summary")

Top 10 Most Active Users: Review Summary

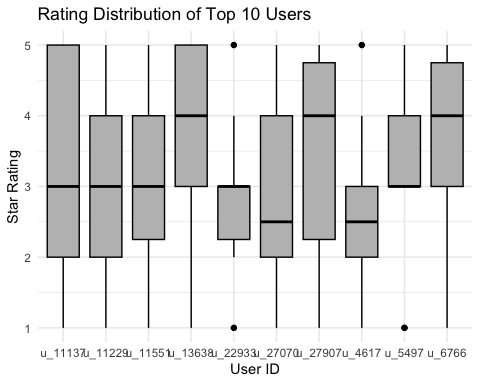
| user\_id | average\_stars | name | member\_since | users\_group | review\_count |
| --- | --- | --- | --- | --- | --- |
| u\_27070 | 3.083333 | Rebecca | 2023-07-22 | New | 16 |
| u\_11551 | 2.800000 | Christopher | 2021-12-17 | Intermediate | 14 |
| u\_22933 | 3.000000 | Stephanie | 2018-10-03 | Intermediate | 14 |
| u\_27907 | 3.700000 | Jesse | 2018-03-06 | Intermediate | 14 |
| u\_4617 | 2.800000 | Joel | 2023-03-07 | New | 14 |
| u\_5497 | 2.250000 | Abigail | 2017-11-04 | Intermediate | 14 |
| u\_6766 | 3.250000 | Tracy | 2021-08-04 | Intermediate | 14 |
| u\_11137 | 3.222222 | Austin | 2024-04-06 | New | 13 |
| u\_11229 | 3.250000 | Benjamin | 2023-11-20 | New | 13 |
| u\_13638 | 3.555556 | Melanie | 2018-12-04 | Intermediate | 13 |

#kable\_classic("hover", full\_width = F, html\_font = "Cambria")

Task 3.3 Visualise their rating distrubtion using ggplot2 - boxplot

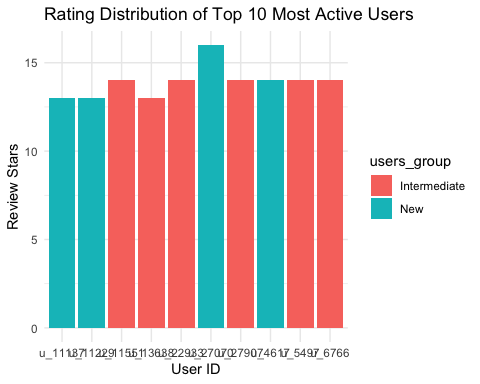
First, we filter out the user\_id in the top\_10 in the data\_rv. Then we can visualise the distrubtion of stars (in the data\_rv) by using box plot.

top\_10\_reviews <- data\_rv %>%  
 filter(user\_id %in% top\_10$user\_id)   
  
ggplot(top\_10\_reviews, aes(x = user\_id, y = stars)) +  
 geom\_boxplot(fill = "grey", color = "black") +  
 labs(  
 title = "Rating Distribution of Top 10 Users",  
 x = "User ID",  
 y = "Star Rating"  
 ) +  
 theme\_minimal()



In case, there is further interest in their review behavior, we can look at the group which they are in and the number of reviews they posted.

top\_10$users\_group <- as.factor(top\_10$users\_group)  
plot\_3=ggplot(data=top\_10, aes(x = user\_id, y = review\_count,fill = users\_group)) +  
 geom\_bar(stat = "identity") +   
 labs(  
 title = "Rating Distribution of Top 10 Most Active Users",  
 x = "User ID",  
 y = "Review Stars"  
 ) +  
 theme\_minimal()   
print(plot\_3)



It is evidence that the number of users who are Intermediate made up the largest proportion. This is followed by New users, suggesting that there are the more recent users actively responding in comparision to the older group (Veteran).

1. Write the code to analyse if there is a major difference between the review behavior of users who joined before and after 2020. For these 2 groups of users, compare their star rating behaviour and the length of the reviews (number of charaters in the review text). You are required to make sure you handle the NA value in your analysis. Visualise the average review length by the two groups. Discuss your findings.

Task 4.1: The code to analyse if there is a major difference between the review behavior of users who joined before and after 2020.

To explore the difference between these two groups, the dataset can add another column with TRUE/FALSE value for whom joined before 2020. This dataframe (user\_cat) can joining the data\_rv to give the insights/details about their behaviours of reviewing. By using summary to look at average of stars, length of review text, and their users in both groups, the result is

user\_cat=data\_users %>%  
 #Create a new collumn for a new category  
 mutate(member\_since=as.Date(member\_since),  
 Before\_2020=if\_else(  
 member\_since < as.Date("2020-01-01"),   
 TRUE,   
 FALSE  
 )  
) %>%   
 #Inner join with data\_review for new collumn lenght  
 inner\_join(data\_rv,by="user\_id") %>%  
 #Create new column for review length  
 mutate(review\_length=nchar(text))  
  
head(user\_cat)

## user\_id name member\_since users\_group review\_count Before\_2020 review\_id  
## 1 u\_0 Alan 2019-04-05 Intermediate 3 TRUE r\_127724  
## 2 u\_0 Alan 2019-04-05 Intermediate 3 TRUE r\_154264  
## 3 u\_0 Alan 2019-04-05 Intermediate 3 TRUE r\_182927  
## 4 u\_1 Joel 2015-11-15 Veteran 7 TRUE r\_20756  
## 5 u\_1 Joel 2015-11-15 Veteran 7 TRUE r\_55585  
## 6 u\_1 Joel 2015-11-15 Veteran 7 TRUE r\_75016  
## business\_id stars date  
## 1 b\_2170 5 2022-06-14  
## 2 b\_5349 4 2024-09-03  
## 3 b\_7604 4 2024-02-06  
## 4 b\_10212 1 2023-02-01  
## 5 b\_758 1 2025-01-04  
## 6 b\_18188 4 2023-05-17  
## text  
## 1 National former fact of. Decide kid sure fly treatment avoid staff perform.  
## 2 Leg window though gas majority. Administration require artist whole.  
## 3 Student activity commercial throw know hair development outside. Her improve about difference.  
## 4 Such history position forget. Person no effect pass clear public sure.  
## 5 Event edge movement national. Two report occur better time because.  
## 6 Feel argue special discuss. Husband when season in have bed.  
## review\_length  
## 1 75  
## 2 68  
## 3 94  
## 4 70  
## 5 67  
## 6 60

review\_by\_group=user\_cat %>%   
 #Group by TRUE/False value  
 group\_by(Before\_2020) %>%   
 #Summary  
 summarise(  
 #Mean  
 average\_stars\_rating=mean(stars,na.rm=TRUE),  
 #Average length of reviews  
 avg\_review\_length = mean(review\_length, na.rm = TRUE),   
 #Number of reviews  
 n\_reviews = n(),  
 #Number of valid users  
 n\_unique\_users = n\_distinct(user\_id))  
review\_by\_group

## # A tibble: 2 × 5  
## Before\_2020 average\_stars\_rating avg\_review\_length n\_reviews n\_unique\_users  
## <lgl> <dbl> <dbl> <int> <int>  
## 1 FALSE 3.00 60.7 82764 19035  
## 2 TRUE 3.00 61.0 74196 16994

Task 4.2 Visualise the average review length by the two groups

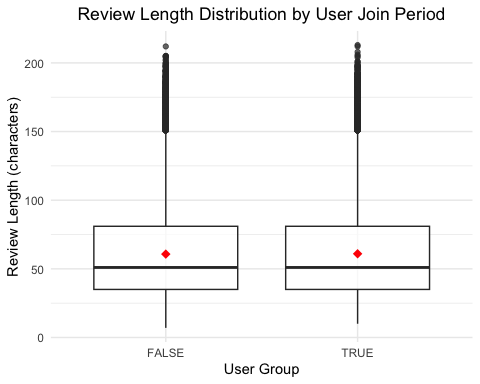
To visualise the datasets or the summary of 2 groups, the box plot is used to show the mean, quartile, and outliners between groups for comparing the length of reviews.

#Rename column names for the dataset  
colnames(review\_by\_group)=c("Join before 2020","Average star rating","Average length of review ", "Number of reviews","Number of users")  
  
#Draw table for the analysis  
  
review\_by\_group %>% kable(caption = "Summary by Group") #%>% kable\_paper("hover",full\_width=F,html\_font = "Camberia")

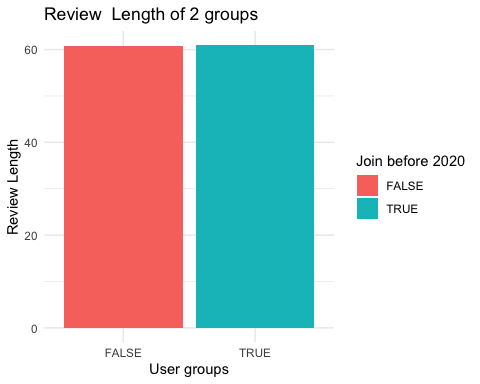
Summary by Group

| Join before 2020 | Average star rating | Average length of review | Number of reviews | Number of users |
| --- | --- | --- | --- | --- |
| FALSE | 3.002791 | 60.73166 | 82764 | 19035 |
| TRUE | 2.997224 | 60.96706 | 74196 | 16994 |

#Draw boxplot graph for analysis and with x, and y axis  
ggplot(user\_cat, aes(x = Before\_2020, y = review\_length)) +  
 #Choose type of graph and colors  
 geom\_boxplot(fill = "white", alpha = 0.7) +  
 #Choose statistic for x value   
 stat\_summary(fun = mean, geom = "point", shape = 18, size = 3, color = "red") +  
 #Set title and x, and y labels  
 labs(title = "Review Length Distribution by User Join Period",  
 x = "User Group",  
 y = "Review Length (characters)") +  
 theme\_minimal() +  
   
 #Adjust title  
 theme(plot.title = element\_text(hjust = 0.5))



plot\_3=ggplot(data=review\_by\_group, aes(x = `Join before 2020`, y = `Average length of review `,fill = `Join before 2020`)) +  
 geom\_bar(stat = "identity") +   
 labs(  
 title = "Review Length of 2 groups",  
 x = "User groups",  
 y = "Review Length"  
 ) +  
 theme\_minimal()   
print(plot\_3)

 As the result is obtain, there is no major difference between these groups. If the interest is whether there chances for difference, hypothesis testing is needed for further investigation. While there is more numbers of users, average stars, and reviews joining after 2020, the mean of review text is shorter approximately 0.2.

# Conclusion

This reports provided an overal analysis of three datasets including information for users, businesses, and reviews. Through data cleaning and reformating, this report identified some trends, behavior of users in each defined groups, and state of the businesses.By using knowledge introduced in the lectures and from reference list, this report utilize some common libraries for data analysis.

# Reference list

GeeksforGeeks (2023) How to Use na.omit in R?, GeeksforGeeks. GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/remove-unnecessary-values-from-an-object-in-r-programming-na-omit-function/> (Accessed: 5 June 2025).

Hao, Z. (2024). Awesome HTML table with knitr::kable and kableExtra. Available at: <https://haozhu233.github.io/kableExtra/awesome_table_in_html.html> [Accessed 5 June 2025].

Kassambara, A. (no date) Identify and remove duplicate data in R. Datanovia. Available at: <https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/> (Accessed: 5 June 2025).