

Case Study R as a Tool



Google data analytics Case Study-R as a Tool-

In this document, I showcased my proficiency in **R programming** by performing data analysis and visualization on Bellabeat's user data.

I cleaned the dataset, created summary statistics, and visualized trends using advanced R packages.

Key Skills Demonstrated:

- Data Cleaning in R:** Using packages like dplyr to clean and manipulate data.
- Statistical Analysis:** Performing summary statistics to understand user behavior.
- Data Visualization:** Creating charts and graphs with **ggplot2** to visualize key metrics.
- Advanced R Techniques:** Combining multiple datasets, analyzing relationships between variables, and creating professional visualizations that convey insights clearly.

TABLE

of contents

01. Introduction

02. Company overview

03. Scope of the analysis

04. The analysis

Introduction



Company Overview

Bellabeat is an innovative company specializing in smart devices, focusing on wellness and fitness technology. Known for its success in the niche market of health-tracking devices, Bellabeat has established itself as a leader in integrating technology with personal wellness

Current Market Position

While Bellabeat has achieved significant success as a small company, it is well-positioned to expand its influence and capture a larger share of the global smart device market. The company's innovative approach and existing customer base provide a strong foundation for scaling operations

Objective of the Analysis

Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, envisions that a detailed analysis of smart device fitness data can unlock new growth opportunities. This analysis aims to explore fitness data trends, identify potential areas for product improvement, and uncover insights that could drive strategic growth.

The business task

Analyzing data fitness App to unlock new growth opportunities for the company

Scope of the Analysis

1 Data aggregation

2 Data analyzing

3 Identify key trends and relationships

My data source

- Google data analytics capstone course



dailyActivity_merged.csv



heartrate_seconds_merged.csv



hourlyCalories_merged.csv



hourlyIntensities_merged.csv



hourlySteps_merged.csv



minuteCaloriesNarrow_merged.csv



minuteIntensitiesNarrow_merged.csv



minuteMETsNarrow_merged.csv



minuteSleep_merged.csv



minuteStepsNarrow_merged.csv



weightLogInfo_merged.csv

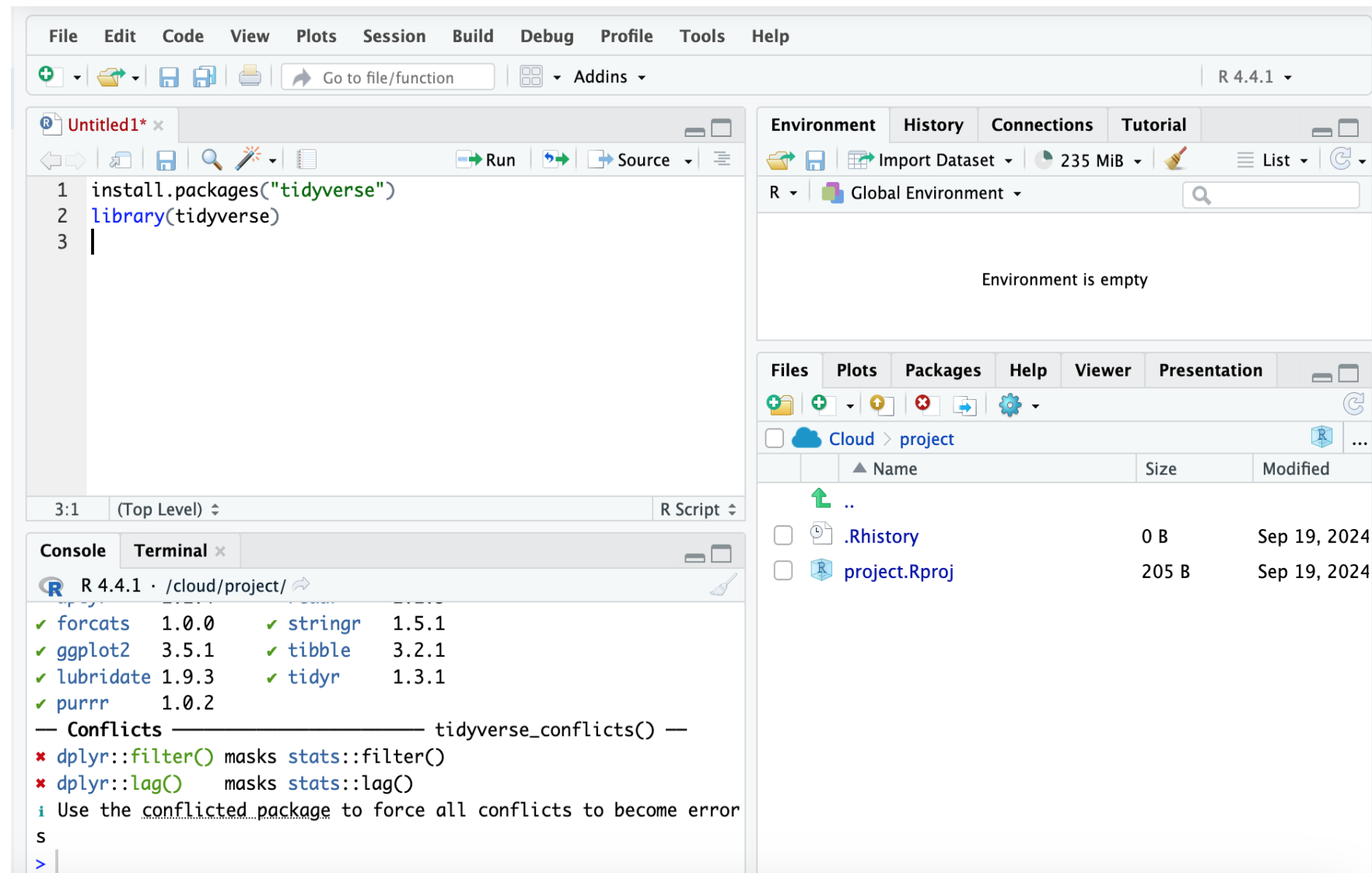
- Google data analytics capstone By Dr.Kabba

The screenshot displays the Posit Cloud web interface for a new R workspace. The browser address bar shows the URL <https://posit.cloud/content/8786555>. The interface includes a sidebar with navigation options like 'Spaces', 'Your Workspace', 'New Space', 'Learn', 'Guide', 'What's New', 'Recipes', 'Cheatsheets', 'Help', 'Current System Status', 'Posit Community', and 'Plans & Pricing'. The main workspace area features a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help) and a toolbar with icons for file operations and a search bar. The console shows the R version 4.4.1 (2024-06-14) and the R Foundation copyright notice. The environment pane on the right indicates that the environment is empty. The file pane at the bottom right shows a list of files: '..', '.Rhistory', and 'project.Rproj'.

Name	Size	Modified
..		
.Rhistory	0 B	Sep 19, 2024
project.Rproj	205 B	Sep 19, 2024

R platform
<https://posit.cloud/>.

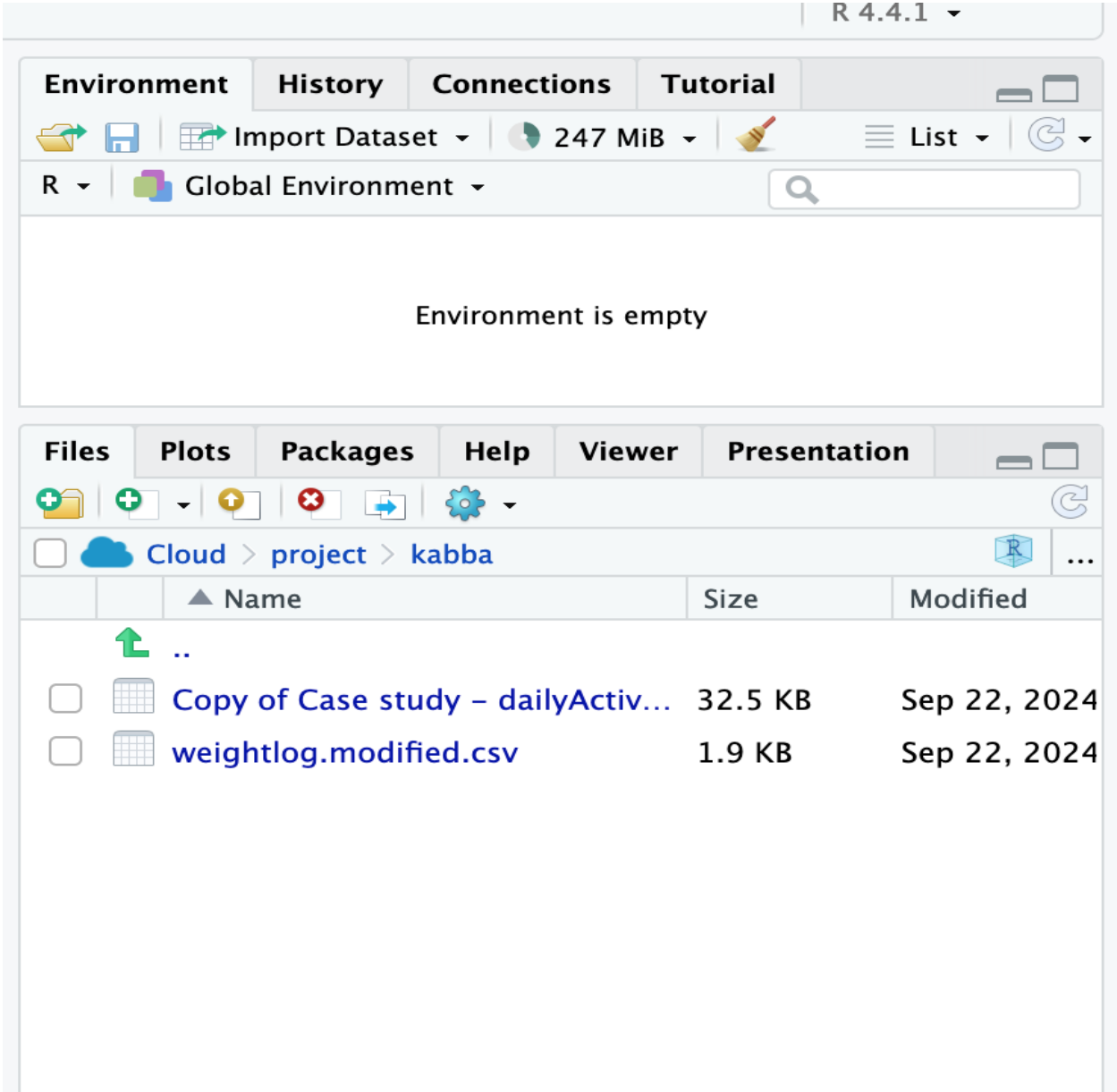
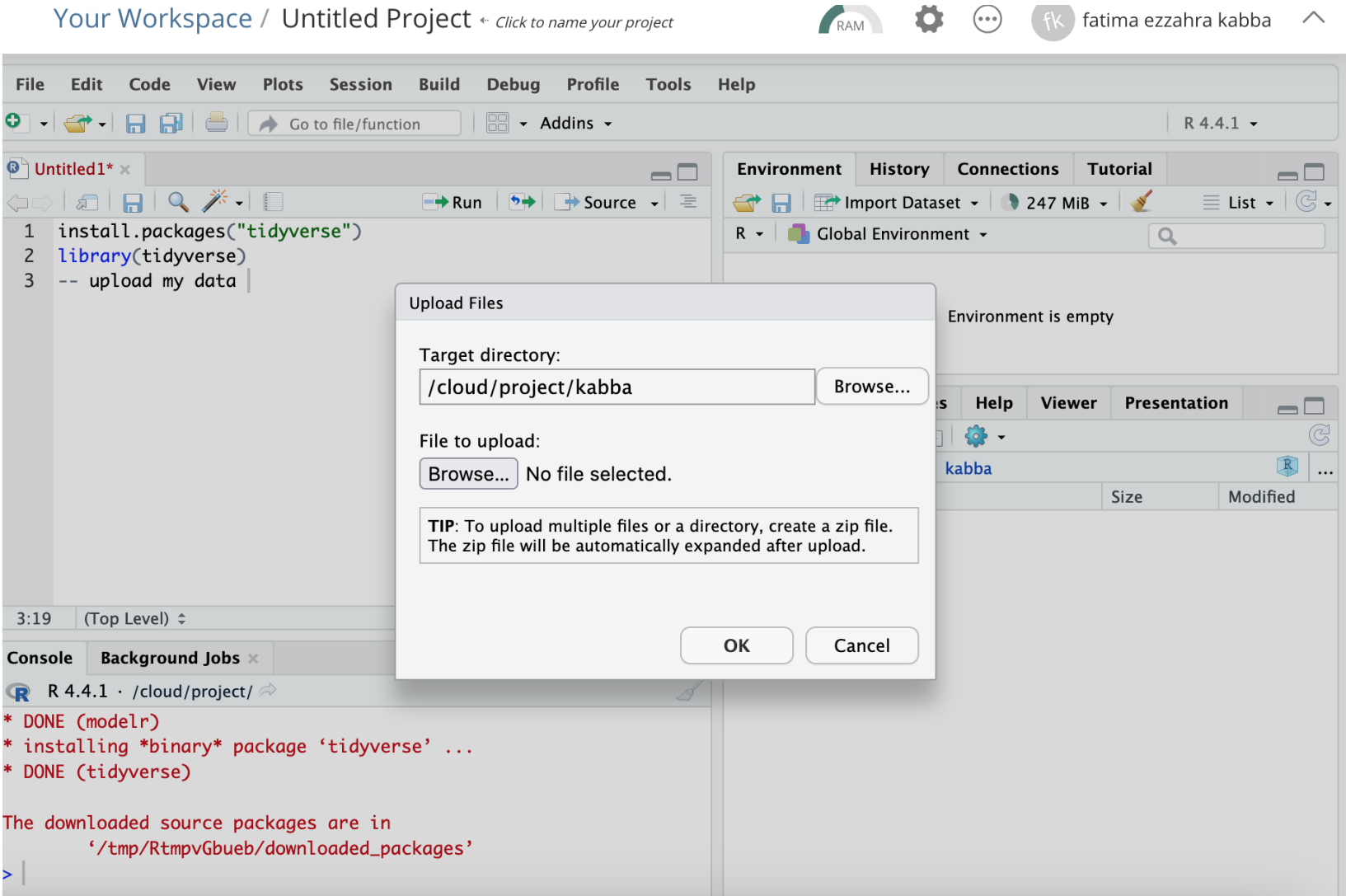
- Google data analytics capstone By Dr.Kabba



1.Install and Load Necessary Packages
Install and Load the tidyverse Package

- Google data analytics capstone

Uploading my data



- Google data analytics capstone course

Checking my data:

Display the first few rows of the data

head(data)

Check the structure of the data

str(data)

```
1 install.packages("tidyverse")
2 library(tidyverse)
3 -- upload my data
4 data <- read.csv("weightlog.modified.csv")
5 head(weightlog_modified)
6
7 |
```

7:1 (Top Level) R Script

Console Background Jobs

R 4.4.1 · /cloud/project/

A tibble: 6 × 5

	Id	Date	WeightKg	WeightPounds	BMI
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>
1	1503960366	04/05/2016	53.3	118.	23.0
2	1927972279	04/10/2016	130.	286.	46.2
3	2347167796	04/03/2016	63.4	140.	24.8
4	2873212765	04/06/2016	56.7	125.	21.5
5	2873212765	04/07/2016	57.2	126.	21.6

Untitled1* x

Run Source

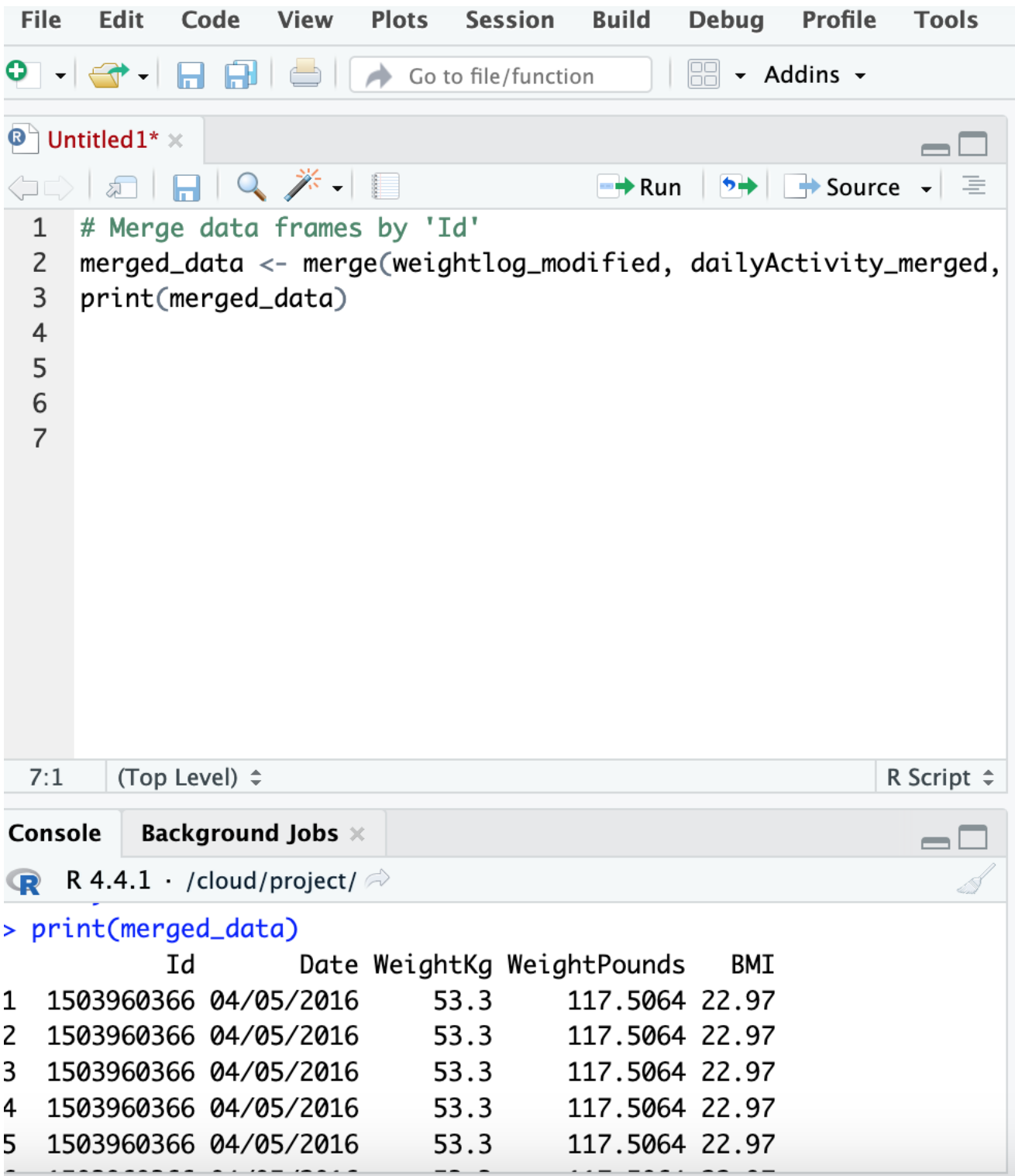
```
1 install.packages("tidyverse")
2 library(tidyverse)
3 -- upload my data
4 data <- read.csv("weightlog.modified.csv")
5 head(weightlog_modified)
5 str(weightlog_modified)
7
```

7:1 (Top Level) R Script

Console Background Jobs

R 4.4.1 · /cloud/project/

```
attr(*, "spec")=
.. cols(
..   Id = col_double(),
..   Date = col_character(),
..   WeightKg = col_double(),
..   WeightPounds = col_double(),
..   BMI = col_double()
.. )
```



- Google data analytics capstone course

Merging the data

- Google data analytics capstone

Counting the number of unique participant in the data

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for merging two data frames and counting unique participants.
- Environment:** Lists the objects in the Global Environment.
- Files:** Shows the file explorer with a list of files and folders.
- Console:** Shows the output of the R code execution.

R Code (Source Editor):

```
1 # Merge data frames by 'Id'
2 merged_data <- merge(weightlog_modified, dailyActivity_merged)
3 print(merged_data)
4 # Save the merged data frame to a CSV file
5 write.csv(merged_data, file = "merged_data.csv", row.names = FALSE)
6 num_unique_participants <- length(unique_ids)
7 unique_ids <- unique(merged_data$Id)
8 num_unique_participants <- length(unique_ids)
9 print(num_unique_participants)
10
11
12
```

Environment:

Object	Size
eightlog_modi...	33 obs. of 5 variables
merged_data	436 obs. of 20 variables
weightlog_mod...	33 obs. of 5 variables

Files:

Name	Size	Modified
..		
.Rhistory	0 B	Sep 22, 202
dailyActivity_merged.csv	42 KB	Sep 22, 202
kabba		
project.Rproj	205 B	Sep 22, 202
weightlog.modified.csv	1.9 KB	Sep 22, 202
merged_data.csv	68.6 KB	Sep 22, 202

Console:

```
R 4.4.1 · /cloud/project/
View(merged_data)
unique_ids <- unique(merged_data$Id)
unique_ids <- unique(merged_data$Id)
num_unique_participants <- length(unique_ids)
print(num_unique_participants)
] 11
```

Your Workspace / 22septmy R project

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins R 4.4.1

```
1 # Merge data frames by 'Id'
2 merged_data <- merge(weightlog_modified, dailyActivity_merged)
3 print(merged_data)
4 # Save the merged data frame to a CSV file
5 write.csv(merged_data, file = "merged_data.csv", row.names = FALSE)
6 num_unique_participants <- length(unique_ids)
7 unique_ids <- unique(merged_data$Id)
8 num_unique_participants <- length(unique_ids)
9 print(num_unique_participants)
10 num_observations <- nrow(merged_data)
11 print(num_observations)
12
13
```

Environment History Connections Tutorial

R Global Environment

eightlog_modi...	33 obs. of 5 variables
merged_data	436 obs. of 20 variables
weightlog_mod...	33 obs. of 5 variables

Values

Files Plots Packages Help Viewer Presentation

Cloud > project

	Name	Size	Modified
	..		
	.Rhistory	0 B	Sep 22, 2024
	dailyActivity_merged.csv	42 KB	Sep 22, 2024
	kabba		
	project.Rproj	205 B	Sep 22, 2024
	weightlog.modified.csv	1.9 KB	Sep 22, 2024
	merged_data.csv	68.6 KB	Sep 22, 2024

Console Background Jobs

R 4.4.1 · /cloud/project/

```
· num_unique_participants <- length(unique_ids)
· print(num_unique_participants)
[1] 11
· num_observations <- nrow(merged_data)
· print(num_observations)
[1] 436
```

- Google data analytics capstone By Dr.Kabba

determine the number of
observations (rows) in our data
`num_observations <- nrow(merged_data)`
`print(num_observations)`

- Google data analytics capstone By Dr.Kabba

```
1 # Merge data frames by 'Id'
2 merged_data <- merge(weightlog_modified, dailyActivity_merged, by = "Id")
3 print(merged_data)
4 # Save the merged data frame to a CSV file
5 write.csv(merged_data, file = "merged_data.csv", row.names = TRUE)
6 num_unique_participants <- length(unique_ids)
7 unique_ids <- unique(merged_data$Id)
8 num_unique_participants <- length(unique_ids)
9 print(num_unique_participants)
10 num_observations <- nrow(merged_data)
11 print(num_observations)
12 summary(data)
13
```

13:1 (Top Level) ↕ R Script ↕

Console Background Jobs x

R 4.4.1 · /cloud/project/ ↗

LightActiveDistance		SedentaryActiveDistance	
Max. : 21.920	Max. : 6.4000		
Min. : 0.00	Min. : 0.000000		
1st Qu.: 0.87	1st Qu.: 0.000000		
Median : 2.93	Median : 0.000000		
Mean : 2.89	Mean : 0.001904		
3rd Qu.: 4.46	3rd Qu.: 0.000000		
Max. : 12.51	Max. : 0.100000		

Calculate my summary statistics
summary(data)

```
Max.      :1440.0    Max.      :4562
Connected to your session in progress, last started 2024-Sep-22 11:13:27 UTC (34 minutes ago)
> library(knitr)
> kable(summary(data))
```

	Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistance
Min.	:1.504e+09	Length:457	: 0	: 0.000	: 0.00
1st Qu.	:2.347e+09	Class :character	: 1988	: 1.410	: 1.28
Median	:4.057e+09	Mode :character	: 5986	: 4.090	: 4.09
Mean	:4.629e+09	NA	: 6547	: 4.664	: 4.61
3rd Qu.	:6.392e+09	NA	:10198	: 7.160	: 7.11
Max.	:8.878e+09	NA	:28497	:27.530	:27.53

- print my summary statistics in a more organized and visually appealing way

- Google data analytics capstone By Dr.Kabba

merged_data

```
merged_data <- merge(weightlog_modified, dailyActivity_merged, by = "Id")
print(merged_data)
# Save the merged data frame to a CSV file
write.csv(merged_data, file = "merged_data.csv", row.names = TRUE)
num_unique_participants <- length(unique_ids)
unique_ids <- unique(merged_data$Id)
num_unique_participants <- length(unique_ids)
print(num_unique_participants)
num_observations <- nrow(merged_data)
print(num_observations)
summary(data)
library(knitr)
kable(summary(data))
# Save the summary statistics as a CSV file
write.csv(summary(data), "summary_statistics.csv")
```

Environment

History

Connect

251 MiB

Global Environment

Values

num_obs...	436L
num_uni...	11L
unique_...	num [1:11] 1.50e+09...

Files

Plots

Packages

Help

Cloud > project

Name

..

.Rhistory

dailyActivity_merged.csv

kabba

merged_data.csv

project.Rproj

weightlog.modified.csv

summary_statistics.csv

Background Jobs

1.4.1 · /cloud/project/

0.00	Min. : 0.00	Min. : 0.0	Min. : 32.0	Min. : 0
0.00	1st Qu.: 0.00	1st Qu.: 64.0	1st Qu.: 728.0	1st Qu.:1776
0.00	Median : 1.00	Median :181.0	Median :1057.0	Median :2062
6.62	Mean : 13.07	Mean :170.1	Mean : 995.3	Mean :2189
5.00	3rd Qu.: 16.00	3rd Qu.:257.0	3rd Qu.:1285.0	3rd Qu.:2667
2.00	Max. :660.00	Max. :720.0	Max. :1440.0	Max. :4562

save the summary statistics in Csv
file for later analysis

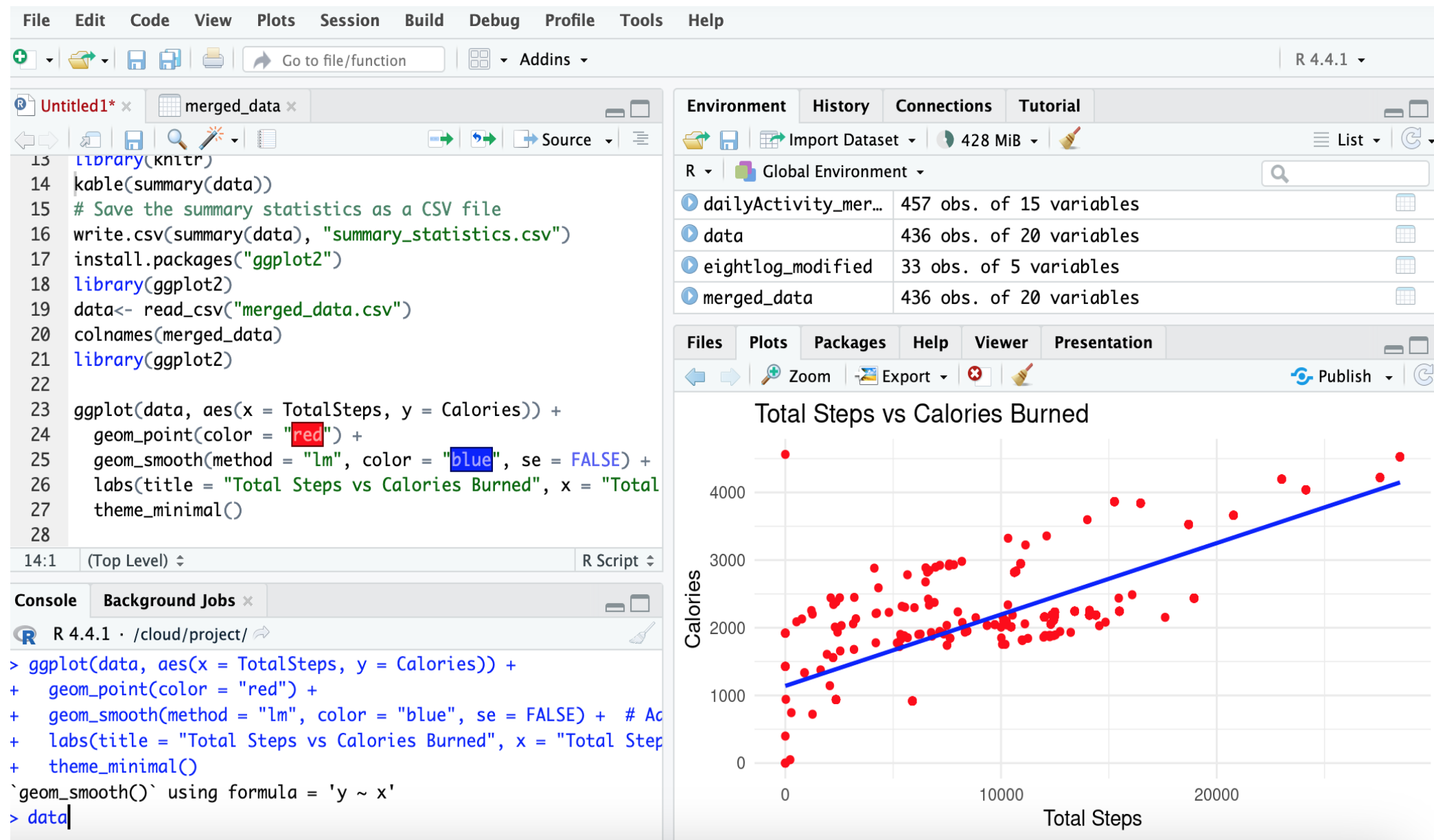
```
#write.csv(summary(data),  
"summary_statistics.csv")
```

Untitled1* × summary_statistics × merged_data ×

Filter

TotalDistance	TrackerDistance	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveD
Min. : 0.000	Min. : 0.00	Min. :0.0000	Min. : 0.000	Min. :0.0000
1st Qu.: 1.410	1st Qu.: 1.28	1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.:0.0000
Median : 4.090	Median : 4.09	Median :0.0000	Median : 0.000	Median :0.0200
Mean : 4.664	Mean : 4.61	Mean :0.1794	Mean : 1.181	Mean :0.4786
3rd Qu.: 7.160	3rd Qu.: 7.11	3rd Qu.:0.0000	3rd Qu.: 1.310	3rd Qu.:0.6700
Max. :27.530	Max. :27.53	Max. :6.7271	Max. :21.920	Max. :6.4000

- Google data analytics capstone By Dr.Kabba



Data visualization :

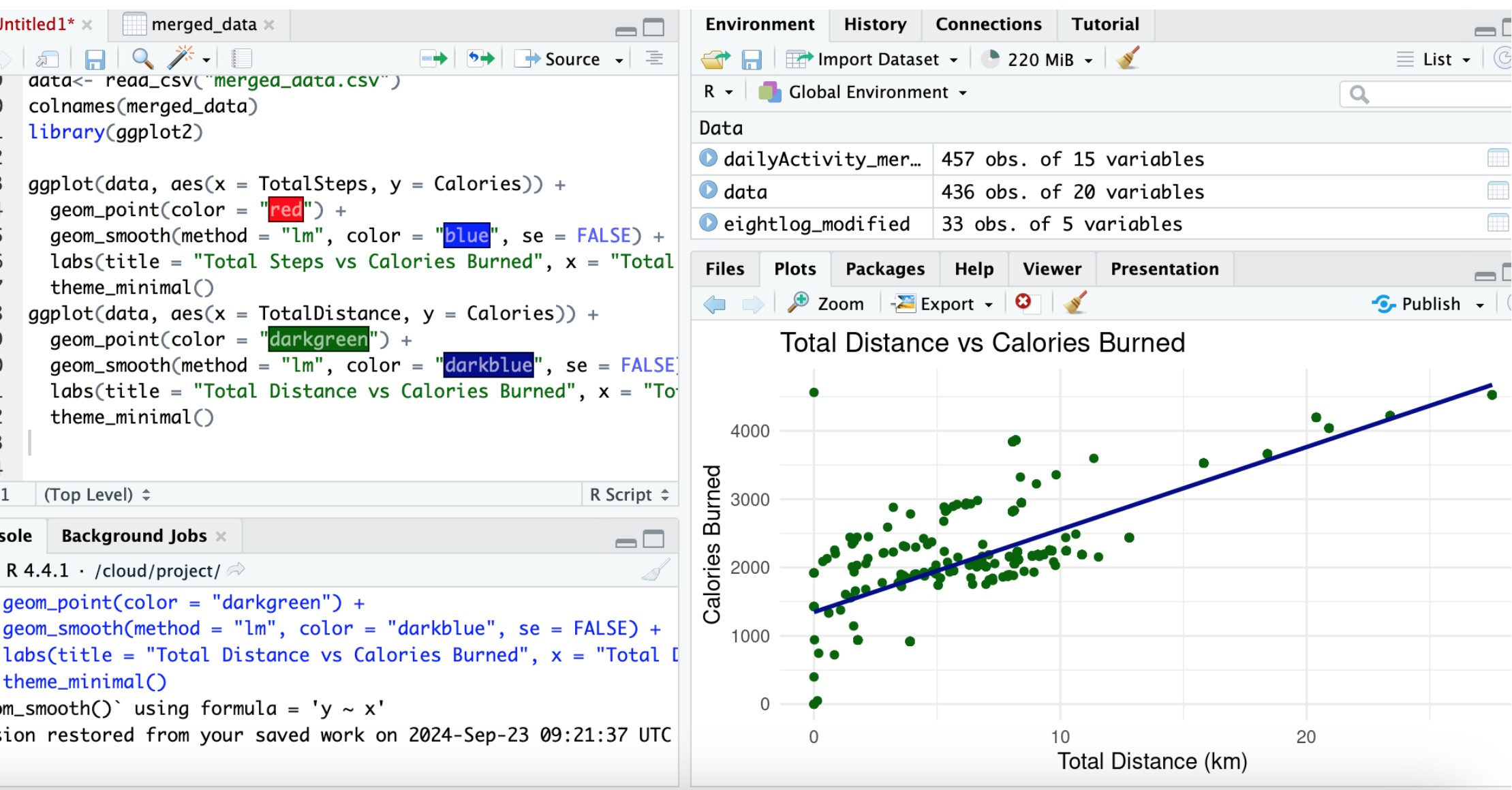
-investigate whether there's a relationship between the number of steps and calories burned.



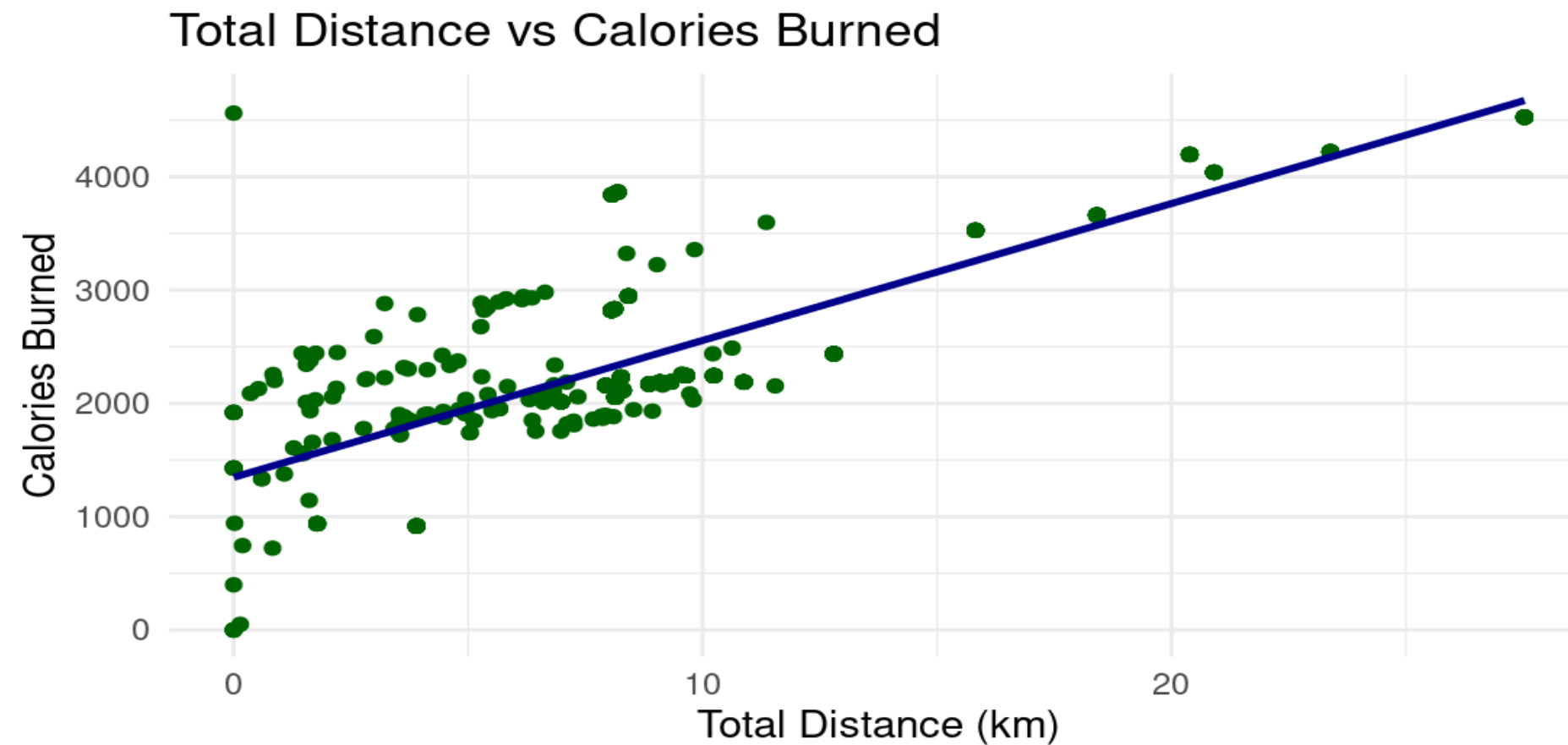
Data visualization :

-investigate whether there's a relationship between the number of steps and calories burned.

- Google data analytics capstone By Dr.Kabba



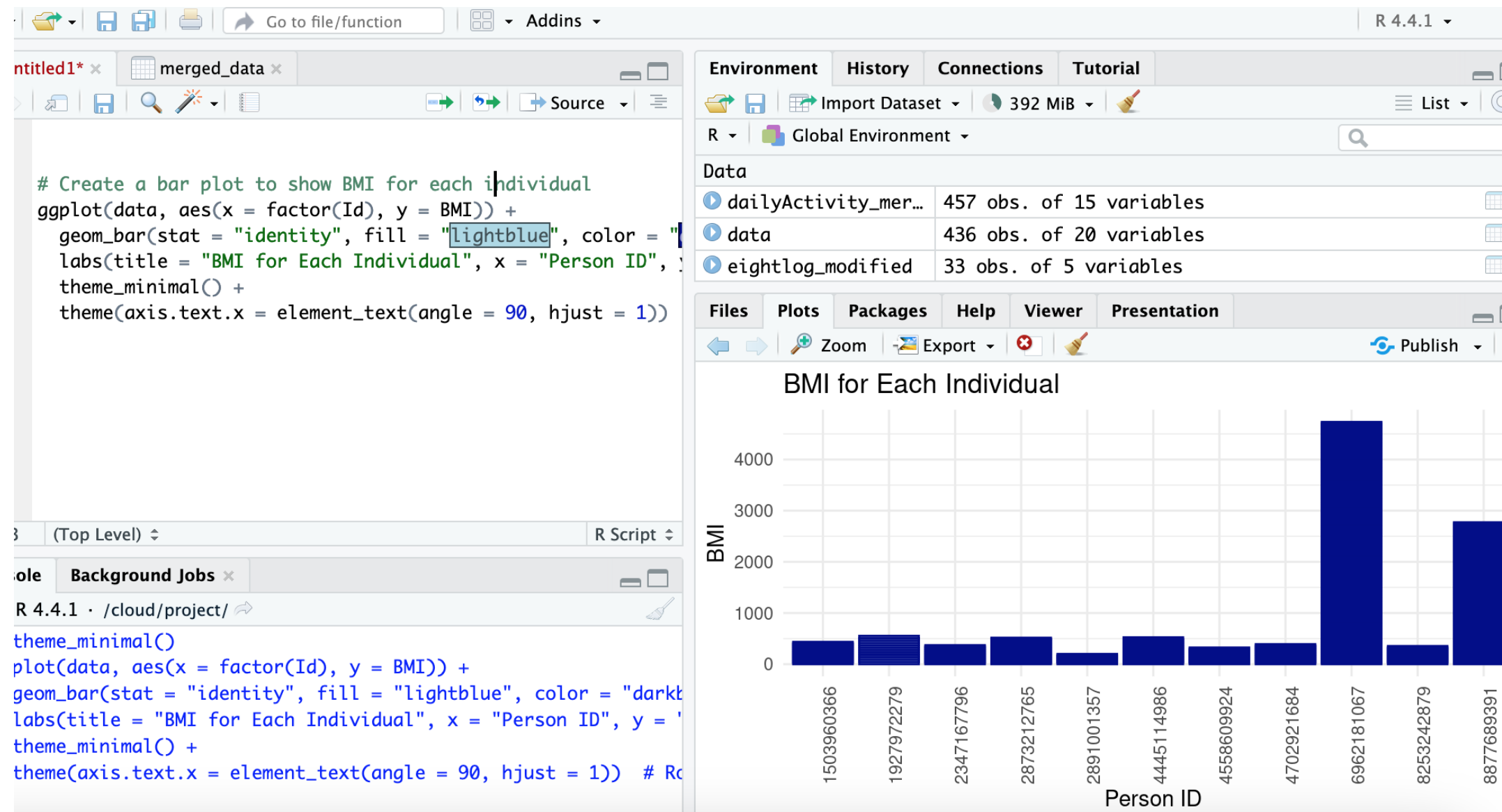
Data visualization :
-investigate whether there’s a relationship between the calories Burned and total distance



Data visualization :

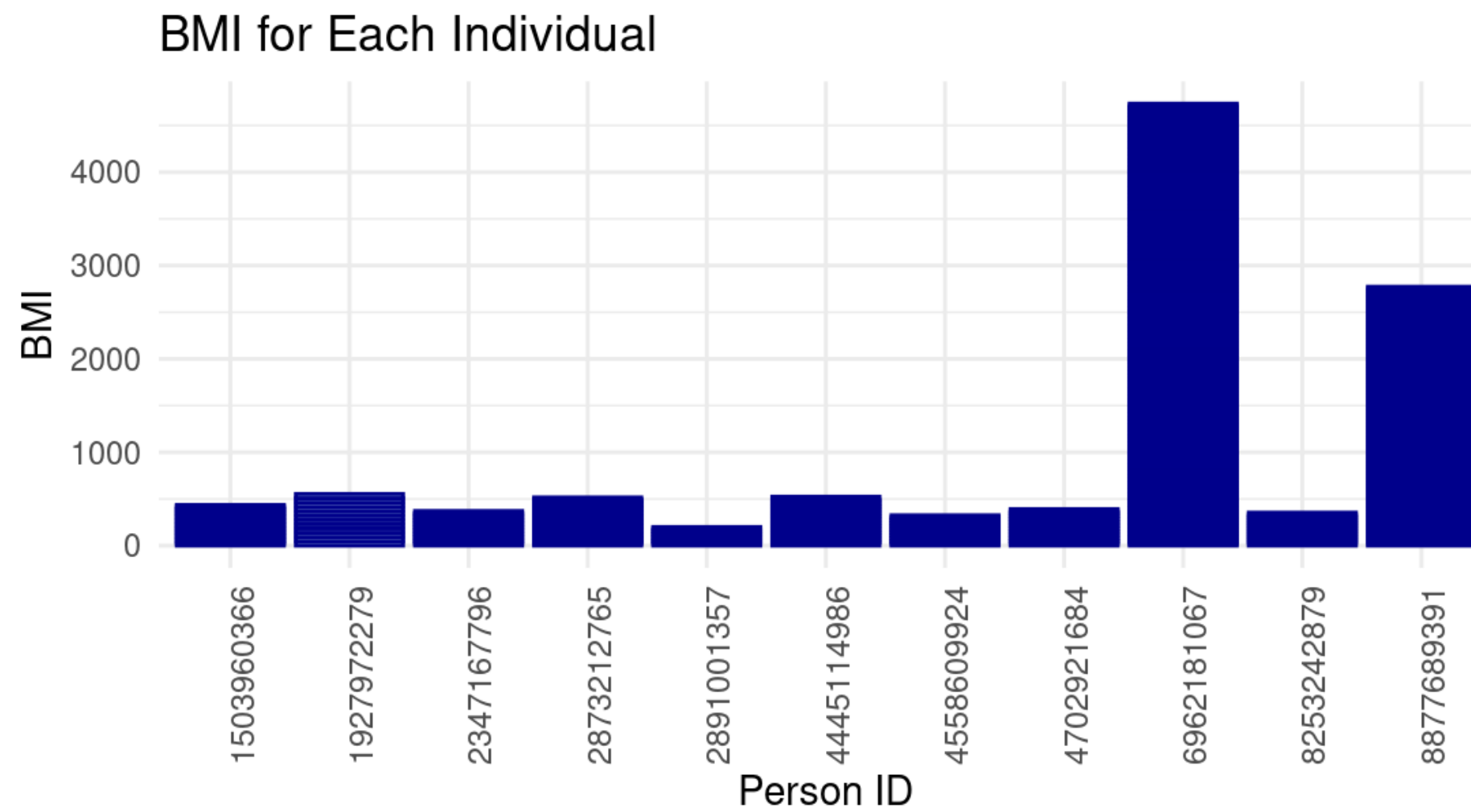
- investigate whether there's a relationship between the calories Burned and total distance

- Google data analytics capstone By Dr.Kabba



Data visualization: BMI values for each individual

```
ggplot(data, aes(x = factor(Id), y = BMI)) +
  geom_bar(stat = "identity", fill = "lightblue", color =
"darkblue") +
  labs(title = "BMI for Each Individual", x = "Person ID",
y = "BMI") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust =
1)) # Rotate x-axis labels for better readability
```



BMI values for each individual