Dr.Fatima Ezzahra Kabba

SEPTEMBRE 2024

Case Study R as a Tool

bellabeat

Prepared by Dr.Fatima Ezzahra Kabba

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

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

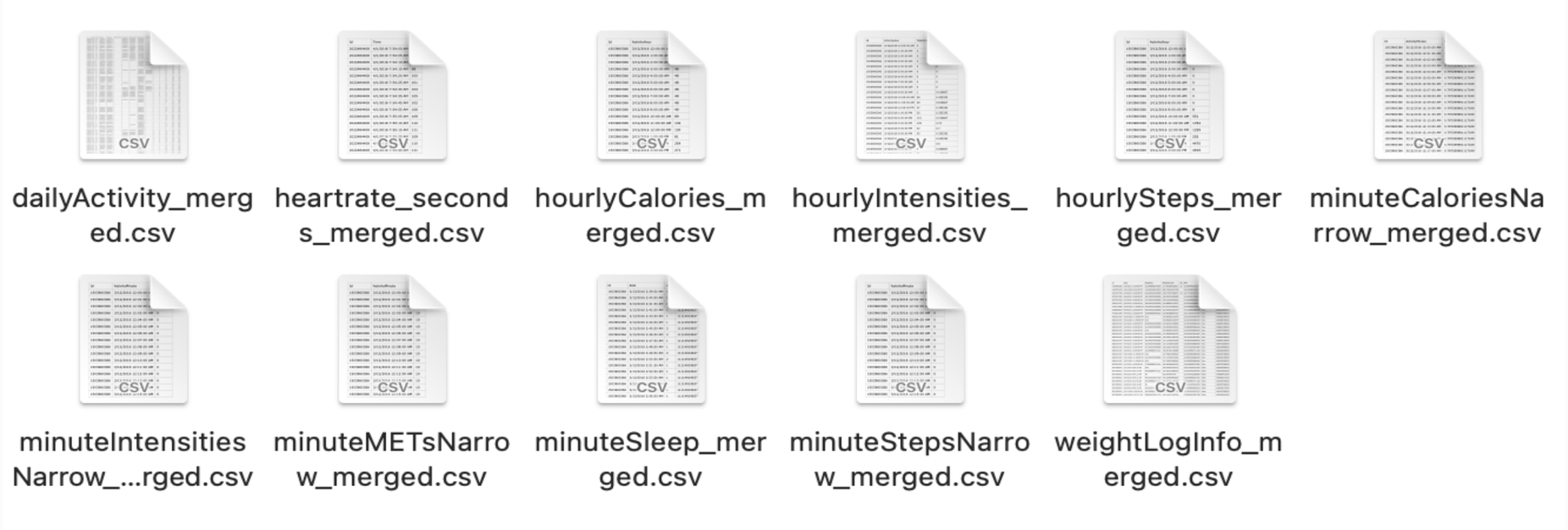
Analyzing data fitness App to unlock new growth opportunities for the company

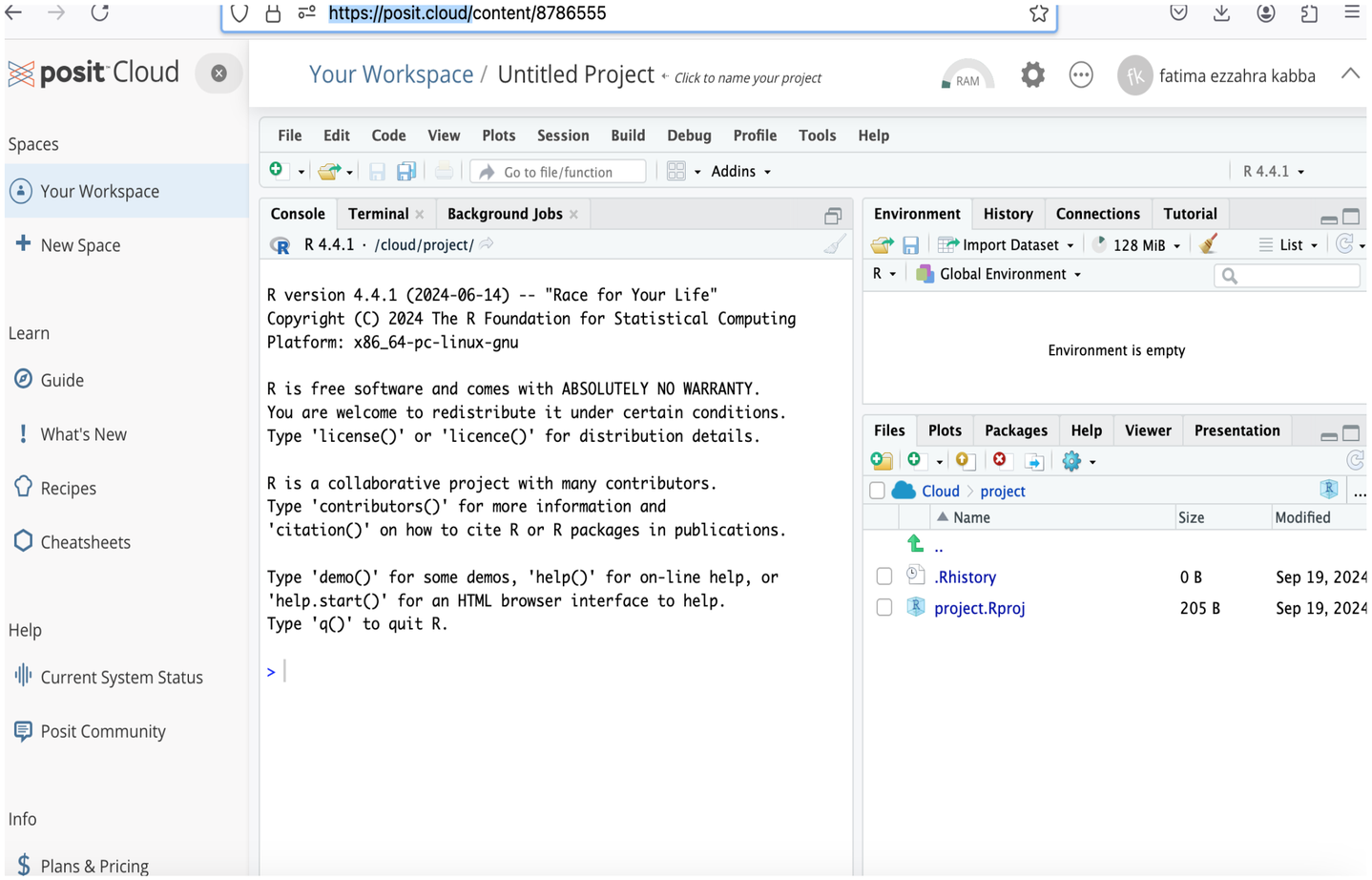
Scope of the Analysis

1. Data aggregation
2. Data analyzing
3. Identify key trends and relationships

My data source

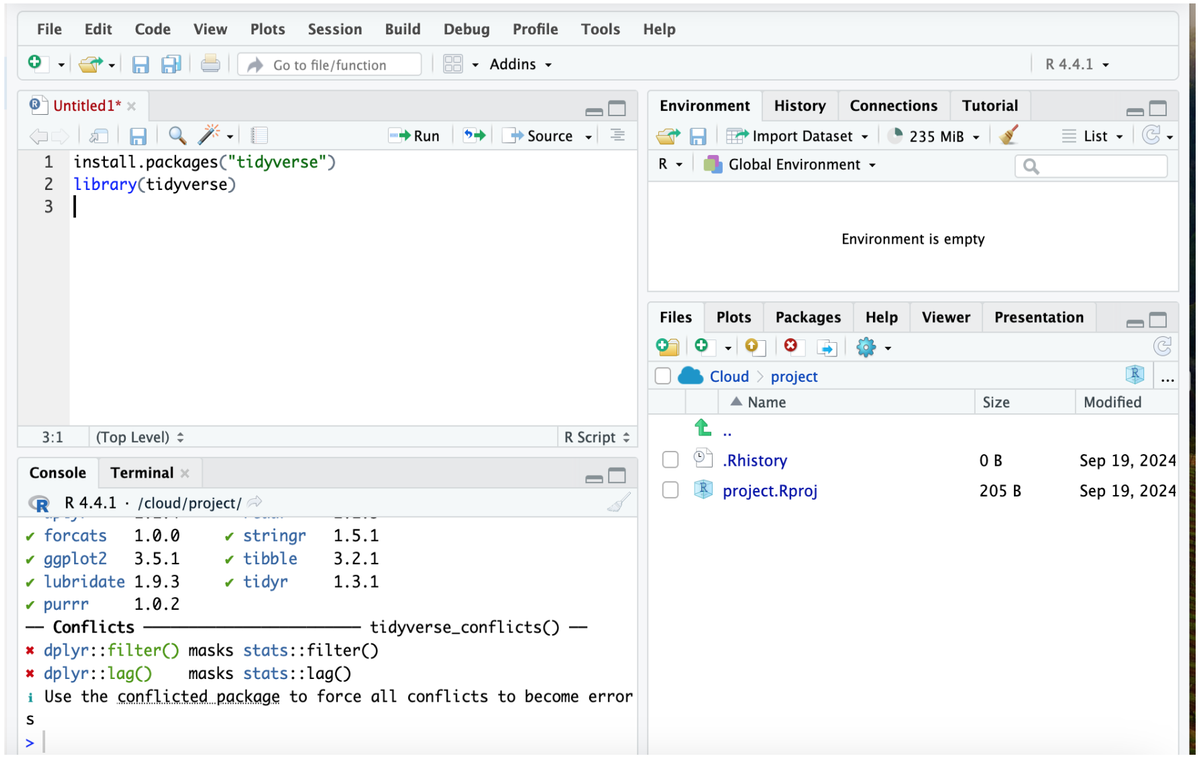
- Google data analytics capstone course





R platform
https://posit.cloud/.

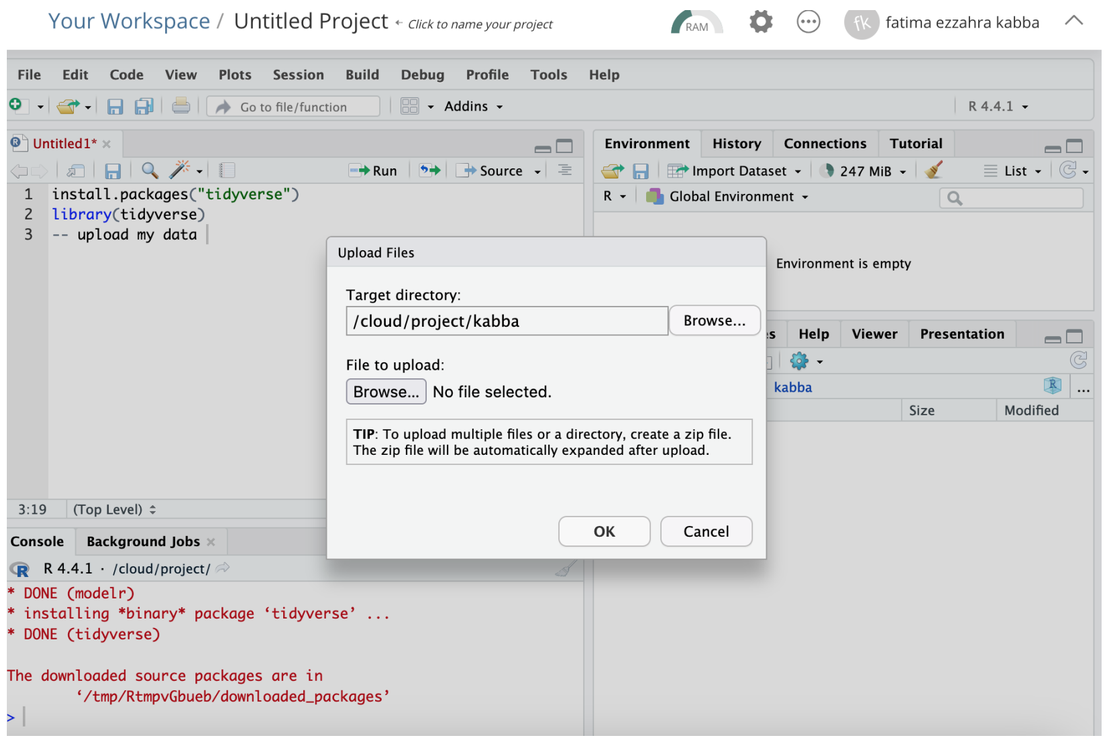
- Google data analytics capstone By Dr.Kabba

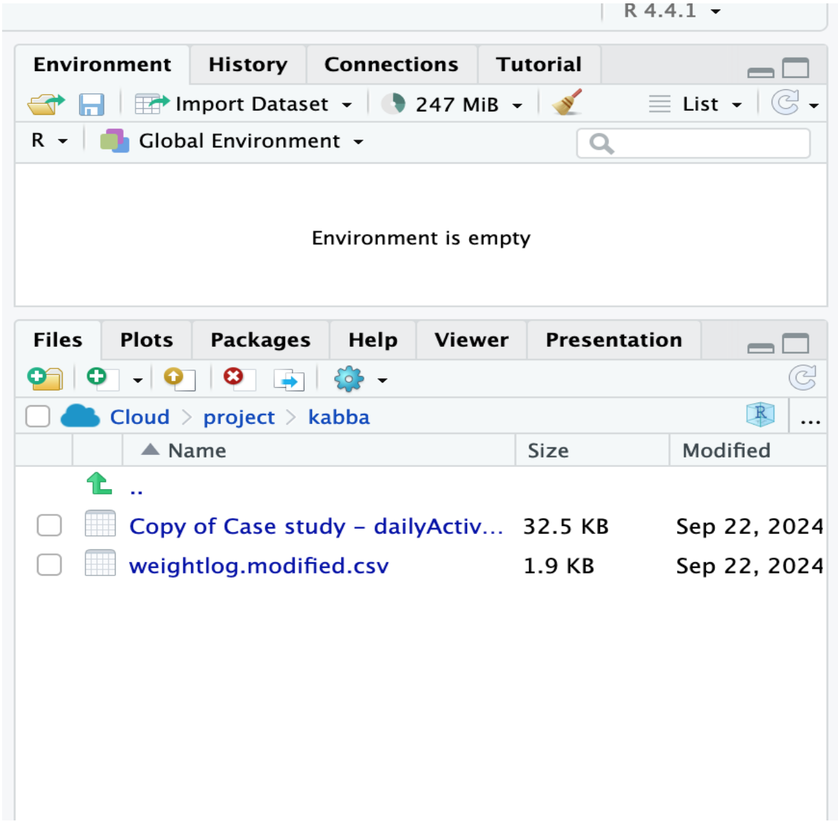


1.Install and Load Necessary Packages
Install and Load the tidyverse Package

- Google data analytics capstone

Uploading my data





- Google data analytics capstone course

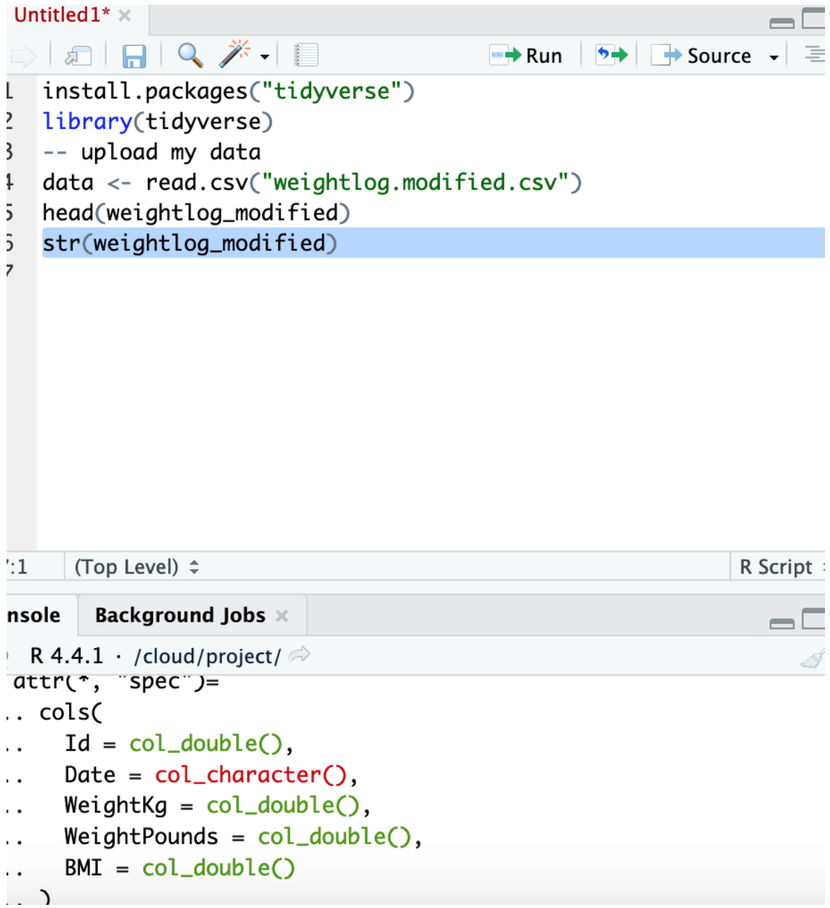
```
install.packages("tidyverse")
library(tidyverse)
-- upload my data
data <- read.csv("weightlog.modified.csv")
head(weightlog_modified)
```

(Top Level) R Script

Background Jobs

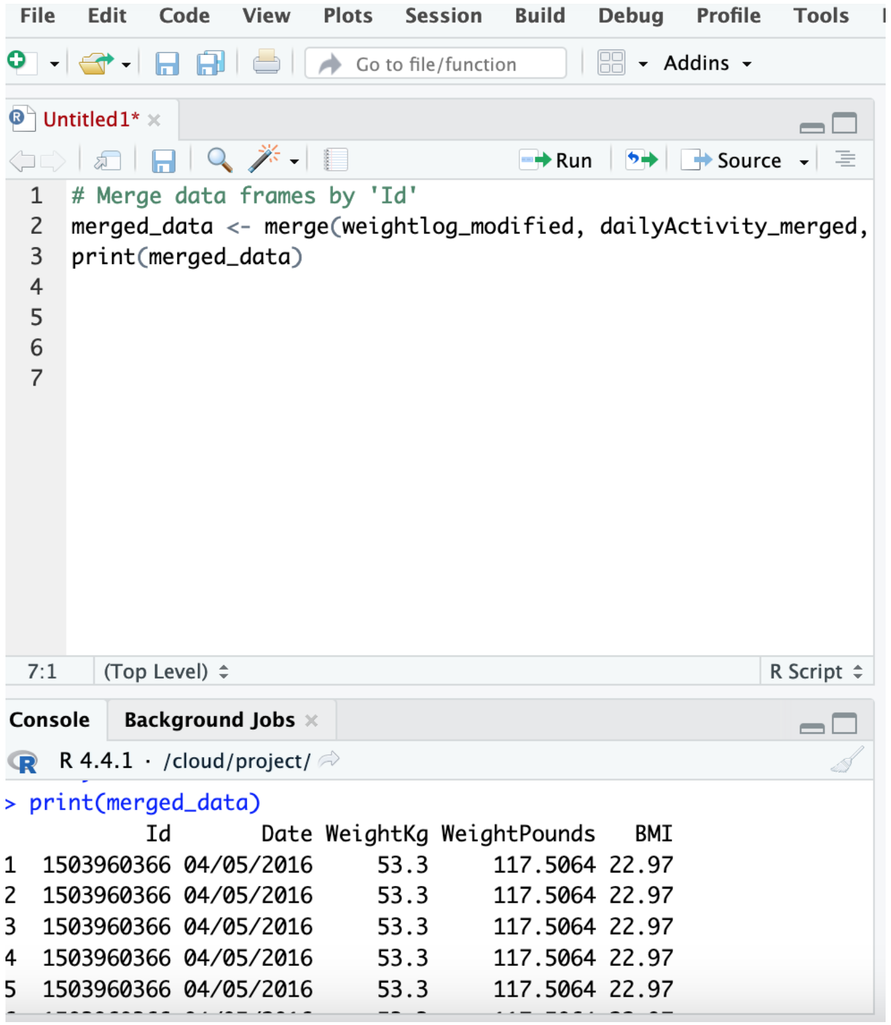
R 4.4.1 · /cloud/project/

```
A tibble: 6 × 5
        Id Date       WeightKg WeightPounds   BMI
     <dbl> <chr>         <dbl>        <dbl> <dbl>
1503960366 04/05/2016     53.3         118.  23.0
1927972279 04/10/2016    130.          286.  46.2
2347167796 04/03/2016     63.4         140.  24.8
2873212765 04/06/2016     56.7         125.  21.5
2873212765 04/07/2016     57.2         126.  21.6
```



Checking my data:

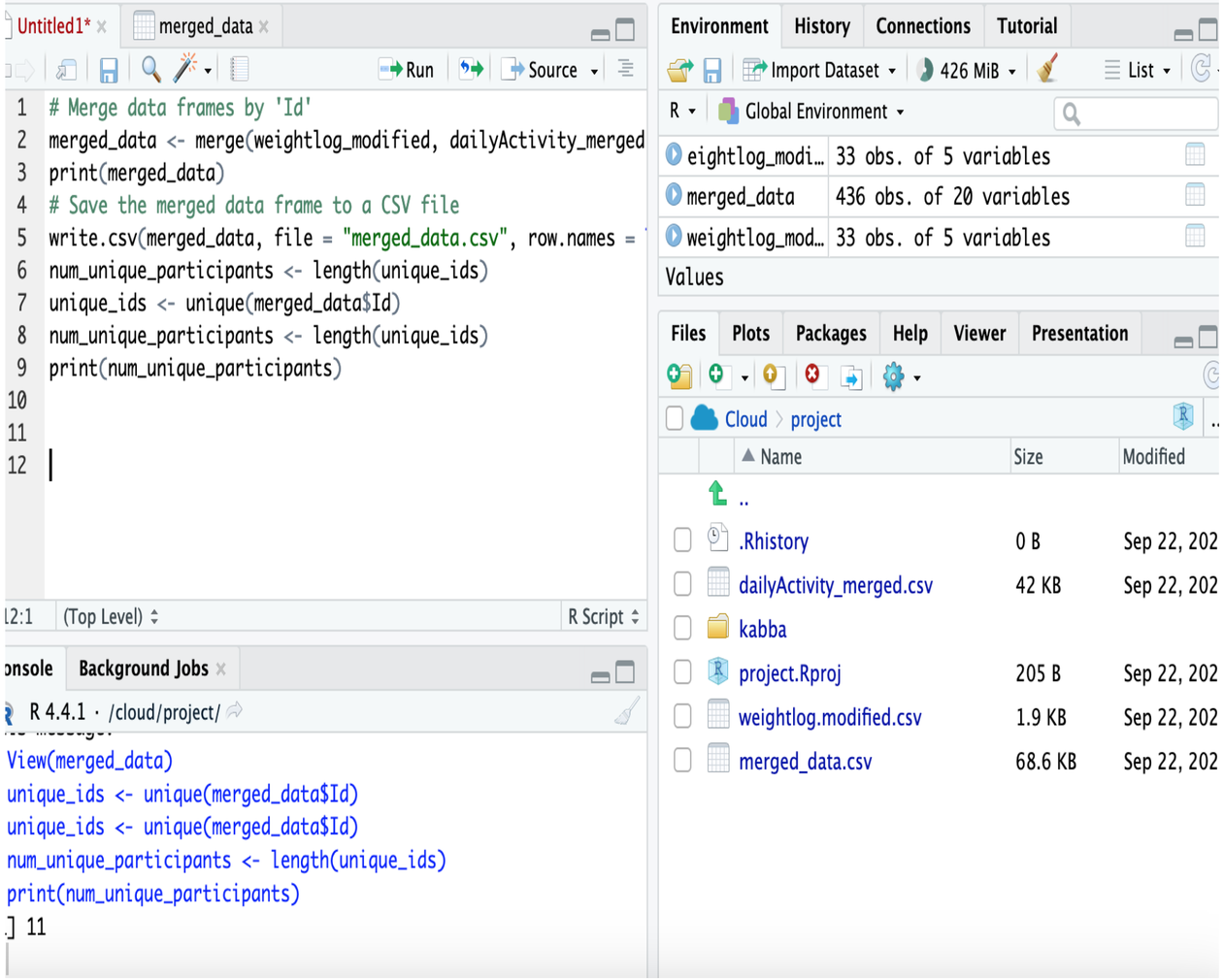
# Display the first few rows of the data
head(data)
# Check the structure of the data
str(data)



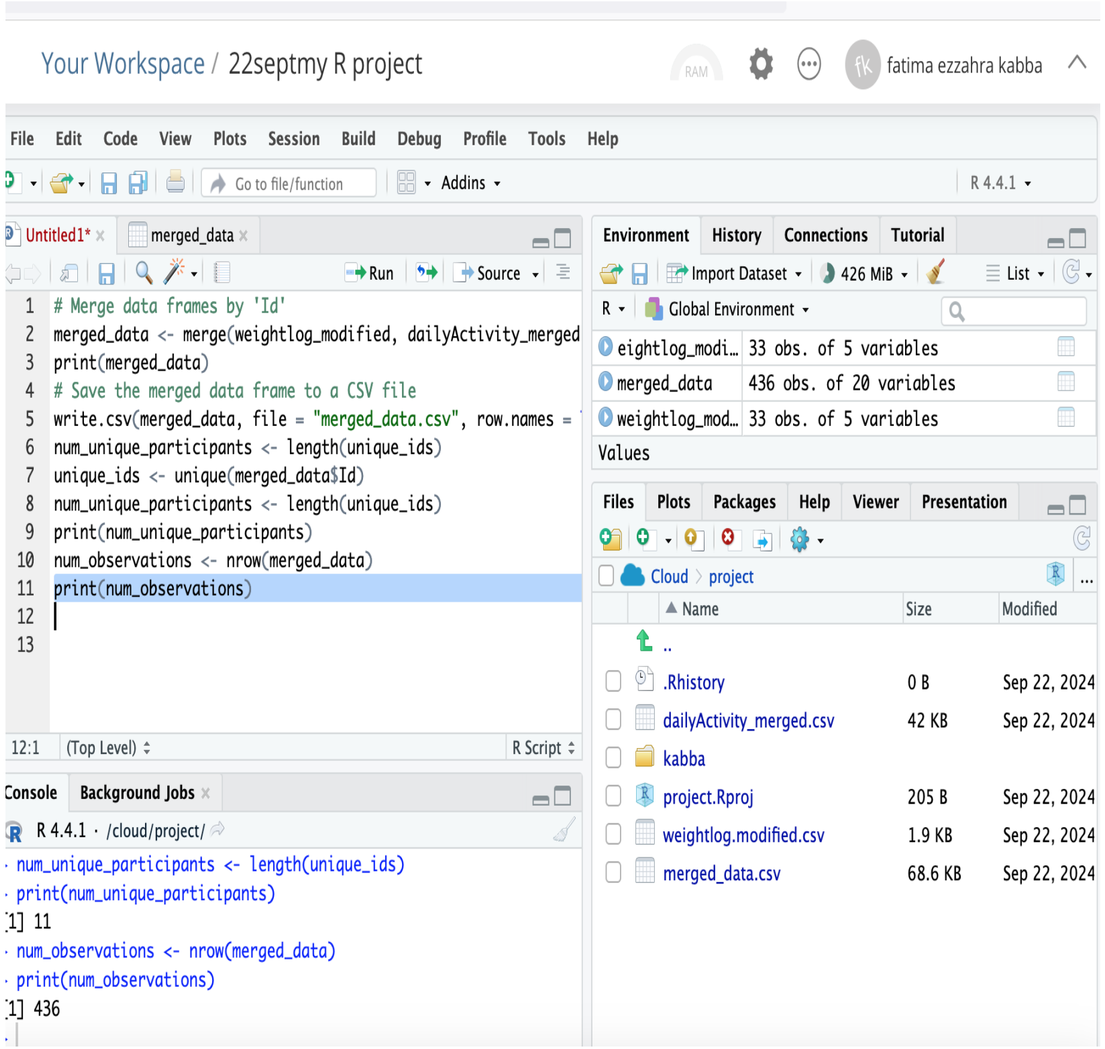
- Google data analytics capstone course

Merging the data

- Google data analytics capstone



Counting the number of unique participant in the data

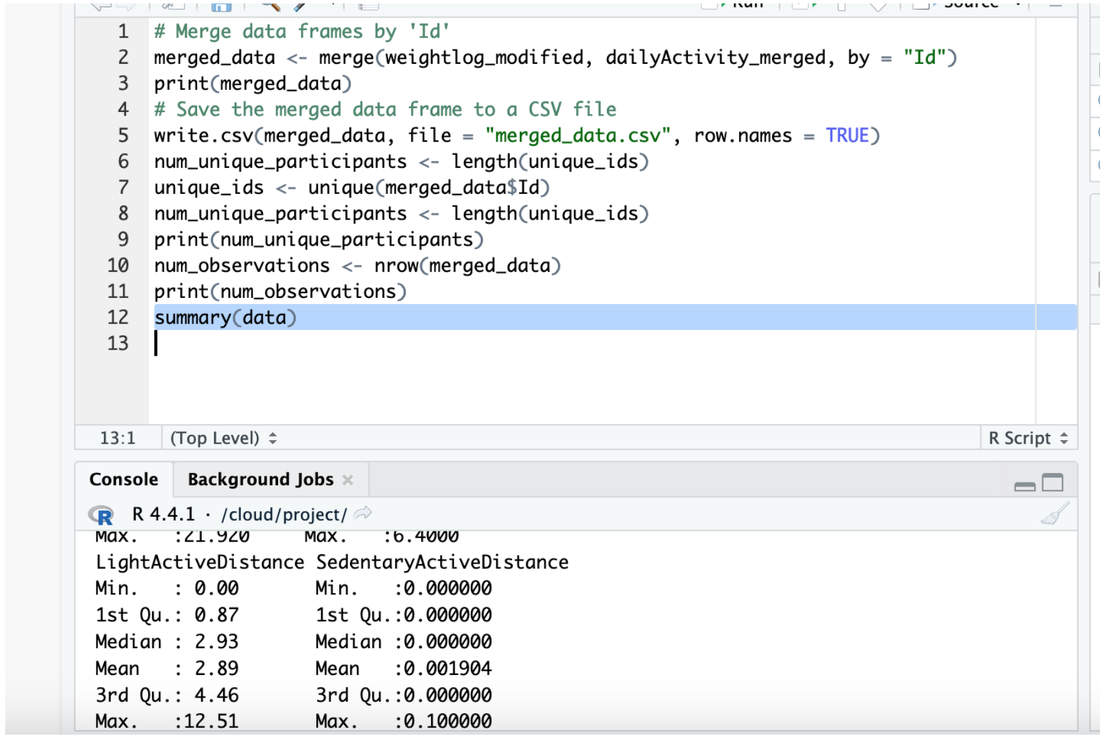


- Google data analytics capstone By Dr.Kabba

determine the number of observations (rows) in our data

```
num_observations <- nrow(merged_data)
print(num_observations)
```

- Google data analytics capstone By Dr.Kabba



Calculate my summary statistics

summary(data)

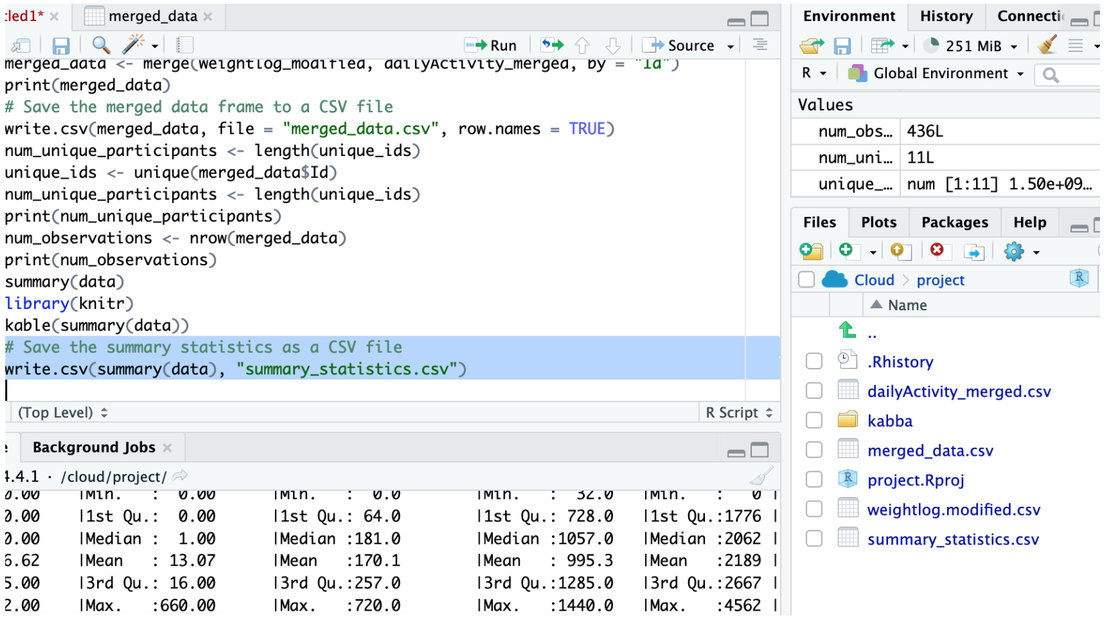
```
 Max.   :1440.0   Max.   :4562
Connected to your session in progress, last started 2024-Sep-22 11:13:27 UTC (34 minutes a
> library(knitr)
> kable(summary(data))
```

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

```
>
```

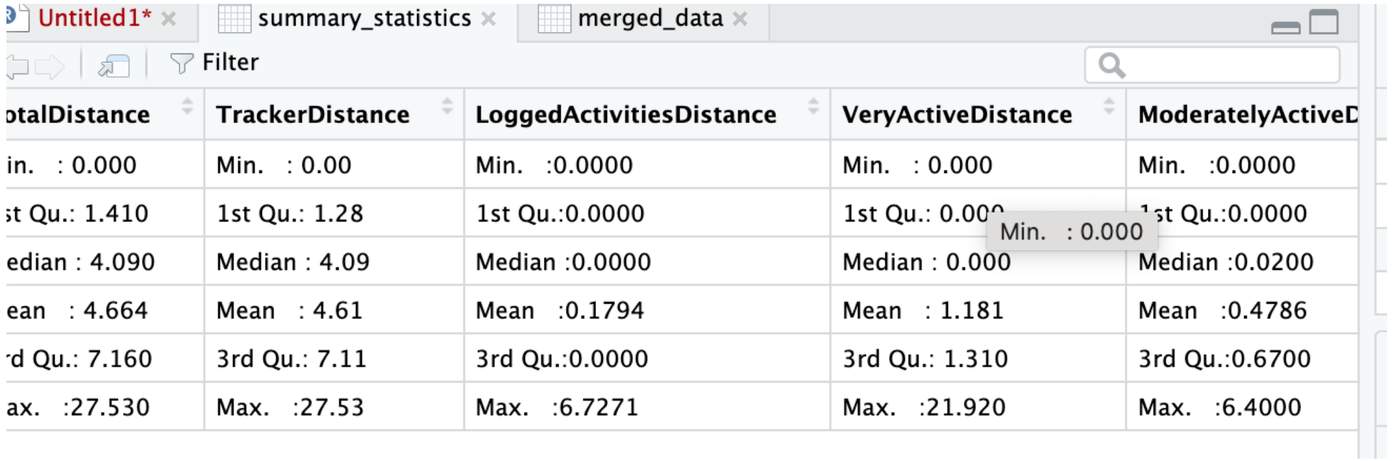
- print my summary statistics in a more organized and visually appealing way

- Google data analytics capstone By Dr.Kabba

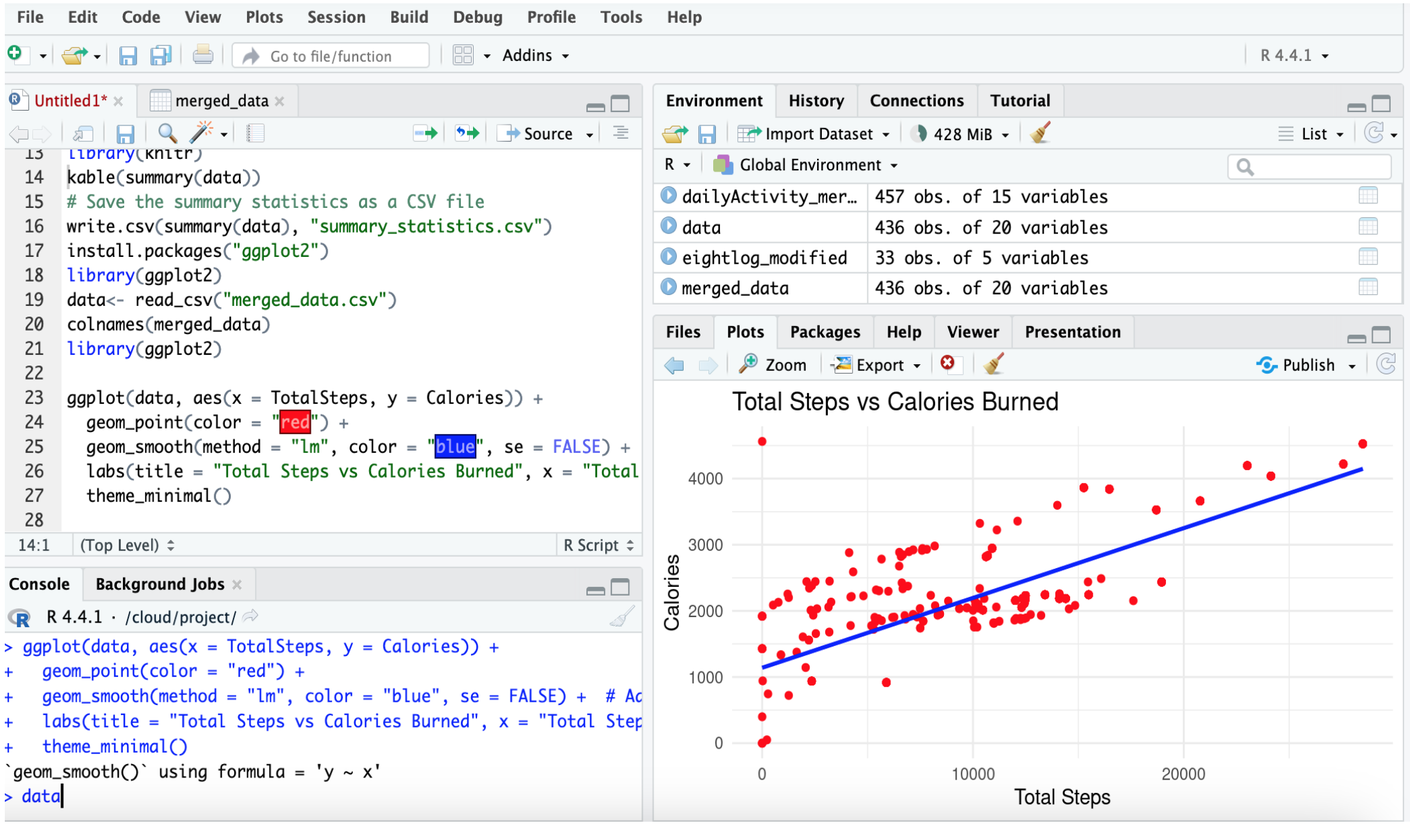


save the summary statistics in Csv file for later analysis

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



- Google data analytics capstone By Dr.Kabba

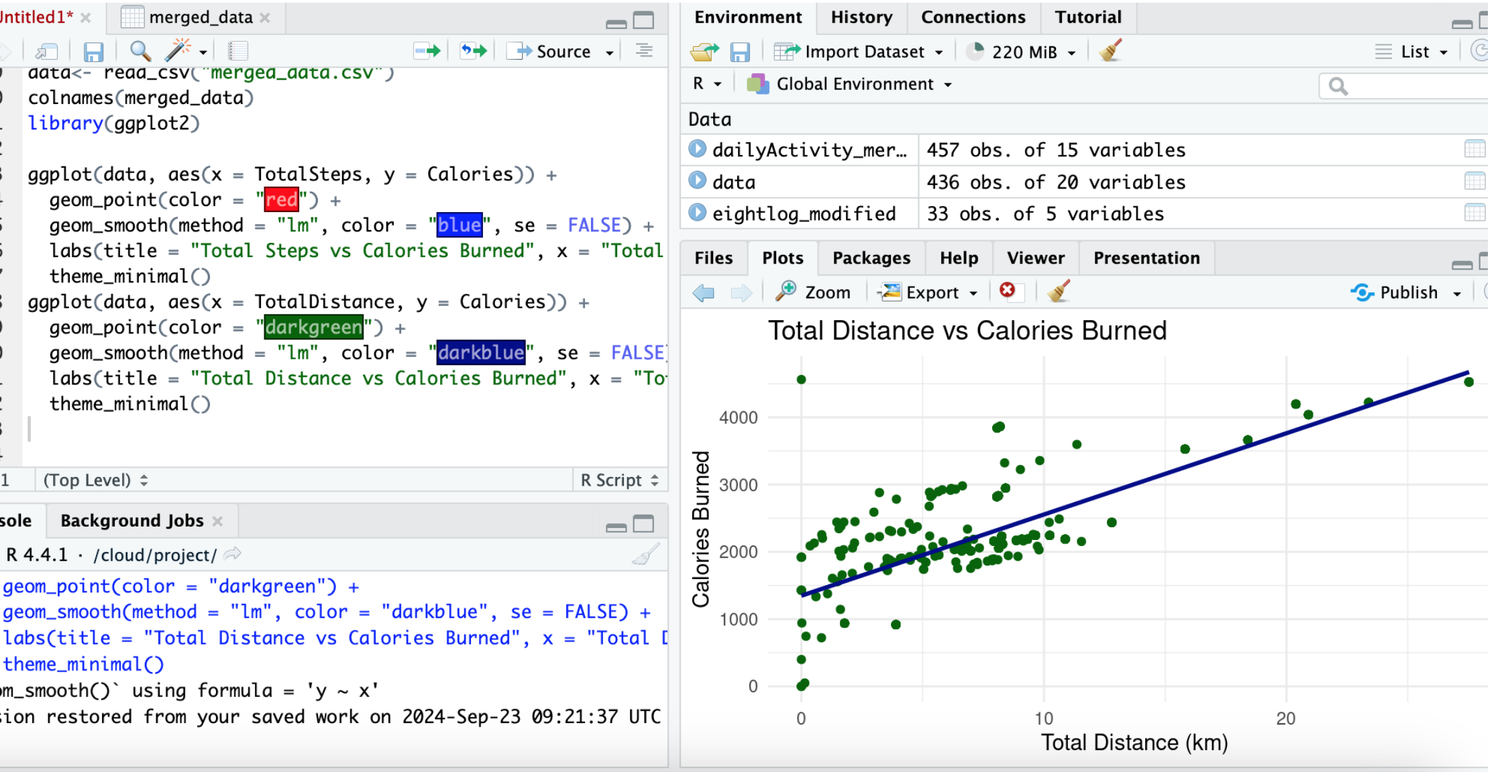


Data visualization :
-investigate whether there's a relationship between the number of steps and calories burned.

- Google data analytics capstone By Dr.Kabba

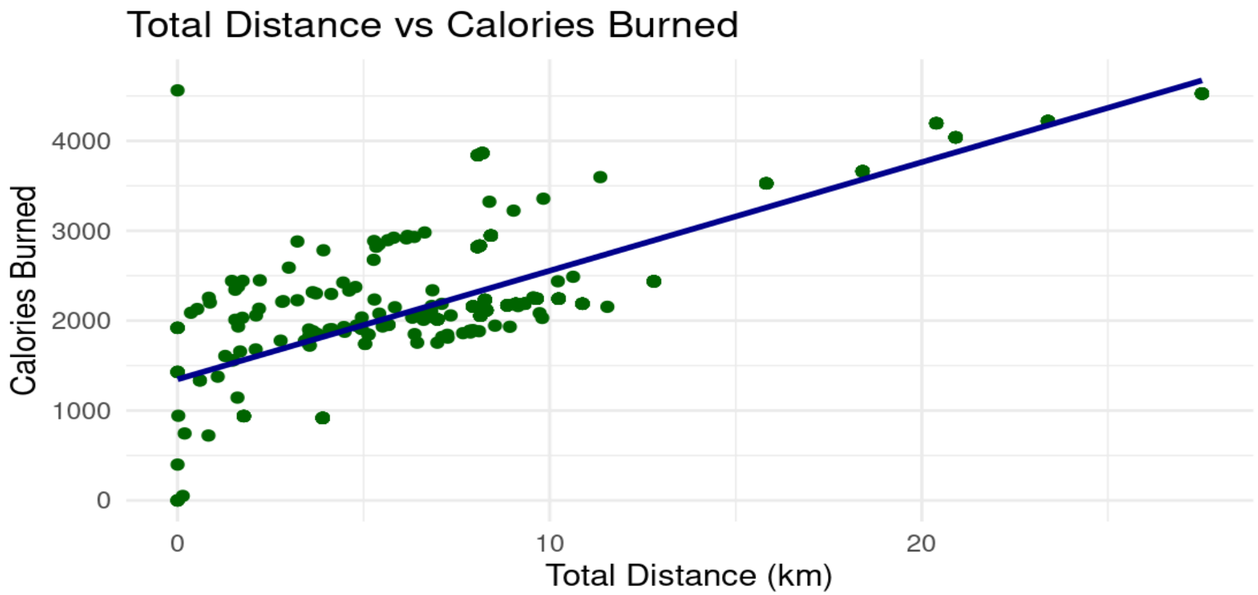
Data visualization :
-investigate whether there’s a relationship between the number of steps and calories burned.

- Google data analytics capstone By Dr.Kabba



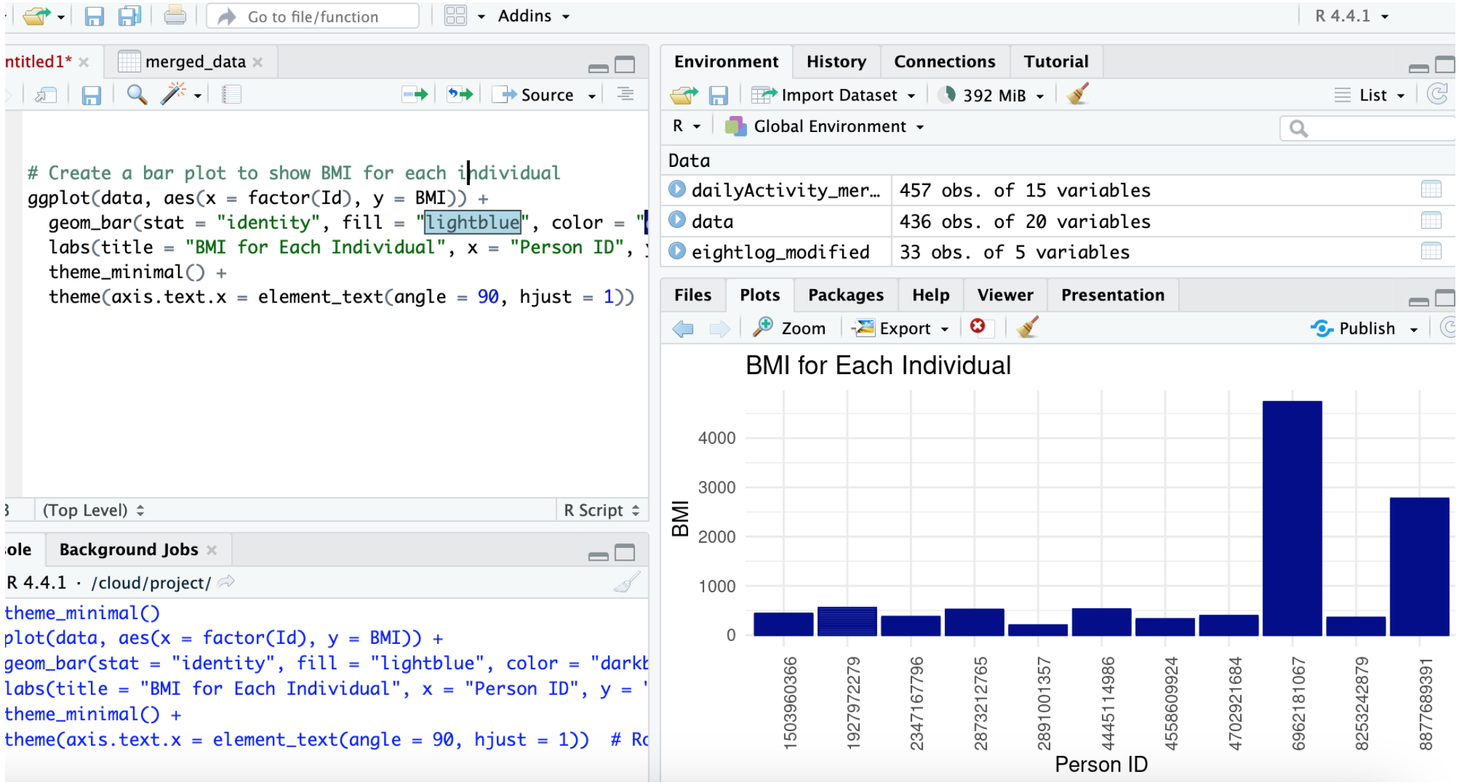
Data visualization :
-investigate whether there's a relationship between the calories Burned and total distance

- Google data analytics capstone By Dr.Kabba



Data visualization :

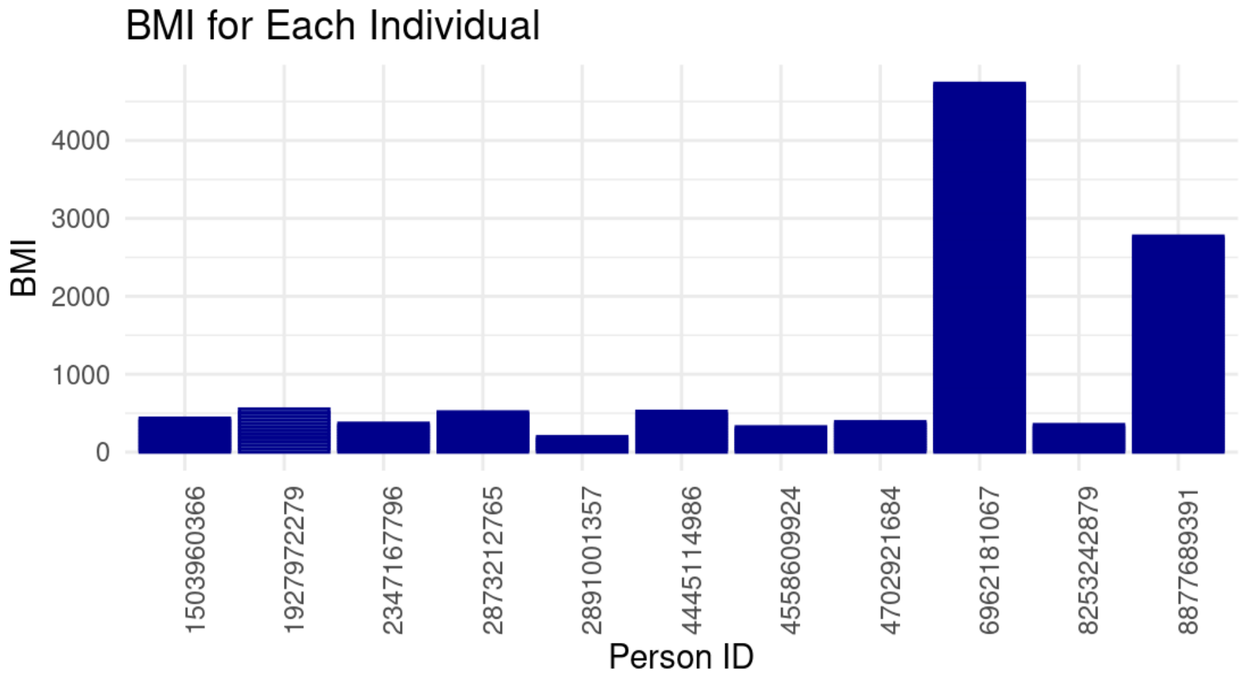
- investigate whether there’s a relationship between the calories Burned and total distance



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
```

- Google data analytics capstone By Dr.Kabba



BMI values for each individual