df <- read.csv("result.csv") rm(df) # This removes df from the environment getwd() list.files() # This will show all files in the current working directory setwd("/cloud/project/cs") list.files() getwd("/cloud/project/cs") head(df) # Check the first few rows of the loaded data df_selected <- df %>% select(rideable_type, started_at, ended_at, member_casual) null_check <- colSums(is.na(df_selected)) print(null_check) df_selected <- df %>% distinct() head(df_selected) colnames(df) df_selected <- df %>% select(rideable_type, started_at, ended_at, member_casual)

Check the columns in df_selected

colnames(df_selected) head(df_selected) # Check the first few rows of the loaded data # Ensure that started_at and ended_at are in POSIXct format df_selected$started_at < -as.POSIXct(df_selected$started_at, format = "%Y-%m-%d %H:%M:%S") df_selected$ended_at < -as.POSIXct(df_selected$ended_at, format = "%Y-%m-%d %H:%M:%S") # Create a new column 'duration' by subtracting started_at from ended_at df_selected$duration < -as.numeric(difftime(df_selected$ended_at, df_selected$started_at, units = "secs"))

Check the first few rows of the data

head(df_selected) # Install writexl package if you haven't already # Load the writexl package library(writexl)

Write the df_selected data frame to an Excel file

write_xlsx(df_selected, path = "df_selected_data.xlsx")

Example with a specific path:

write_xlsx(df_selected, path = "C:/Users/YourUsername/Documents/df_se

Get summary statistics for 'duration'

Load the dplyr package

library(dplyr)

Calculate summary statistics for 'duration'

summary_stats <- df_selected %>% summarise(min_duration = min(duration, na.rm = TRUE), max_duration = max(duration, na.rm = TRUE), avg_duration = mean(duration, na.rm = TRUE), median_duration = median(duration, na.rm = TRUE), sd_duration = sd(duration, na.rm = TRUE))

View the summary statistics

summary_stats # Install and load required packages (if not already installed) install.packages("ggplot2") install.packages("lubridate")

library(ggplot2) library(lubridate)

Extract month from started_at column

df_selected$month < -month(df_selected$started_at, label = TRUE, abbr = TRUE) # abbr=TRUE gives abbreviated month names (e.g., Jan, Feb)

Now plot the data using ggplot2

ggplot(df_selected, aes(x = duration, y = month)) + geom_boxplot() + # You can use boxplot to show the distribution of duration per month labs(title = "Duration by Month", x = "Duration (seconds)", y = "Month") + theme_minimal()

ggplot(df_selected, aes(x = duration, y = month)) + geom_point() + # Scatter plot labs(title = "Duration by Month", x = "Duration (seconds)", y = "Month") + theme_minimal() # Load necessary libraries library(ggplot2) library(lubridate) library(dplyr)

Extract month from 'started_at'

df_selected$month < -month(df_selected$started_at, label = TRUE, abbr = TRUE)

Summarize the total duration for each month

monthly_duration <- df_selected %>% group_by(month) %>% summarise(total_duration = sum(duration, na.rm = TRUE))

Plot the data using a line graph

ggplot(monthly_duration, aes(x = month, y = total_duration, group = 1)) + geom_line(color = "blue", size = 1) + # Line graph geom_point(color = "red", size = 2) + # Points on the line labs(title = "Total Duration by Month", x = "Month", y = "Total Duration (seconds)") + theme_minimal() monthly_avg_duration <- df_selected %>% group_by(month) %>% summarise(avg_duration = mean(duration, na.rm = TRUE))

Plot the average duration by month using a line graph

ggplot(monthly_avg_duration, aes(x = month, y = avg_duration, group = 1)) + geom_line(color = "blue", size = 1) + # Line graph geom_point(color = "red", size = 2) + # Points on the line labs(title = "Average Duration by Month", x = "Month", y = "Average Duration (seconds)") + theme_minimal() ggplot(monthly_avg_duration, aes(x = month, y = avg_duration, group = 1)) + geom_line(color = "blue", size = 1) + geom_point(color = "red", size = 2) + labs(title = "Average Duration by Month", x = "Month", y = "Average Duration (seconds)") + theme_minimal() # Load necessary libraries library(ggplot2) library(lubridate) library(dplyr)

Extract month from 'started_at'

df_selected$month < -month(df_selected$started_at, label = TRUE, abbr = TRUE)

Summarize the total duration for each month

monthly_duration <- df_selected %>% group_by(month) %>% summarise(total_duration = sum(duration, na.rm = TRUE))

Plot the total duration by month using a line graph

ggplot(monthly_duration, aes(x = month, y = total_duration, group = 1)) + geom_line(color = "blue", size = 1) + # Line graph to show total duration over time geom_point(color = "red", size = 2) + # Points to highlight total duration for each month labs(title = "Total Duration by Month", x = "Month", y = "Total Duration (seconds)") + theme_minimal() # Load necessary libraries library(ggplot2) library(dplyr)

Create a bar chart showing the distribution of duration values

ggplot(df_selected, aes(x = duration)) + geom_bar(stat = “bin”, binwidth = 100) + # ‘binwidth’ controls the grouping of duration values labs(title = “Distribution of Duration Values”, x = “Duration (seconds)” theme_minimal() # Load necessary libraries library(ggplot2) library(dplyr) library(lubridate)

# Extract month from ‘started_at’ and create a new column ‘month’ df_selected$month < -month(df_selected$started_at, label = TRUE, abbr = TRUE)

# Calculate the total duration for each month (sum of durations per month) monthly_duration <- df_selected %>% group_by(month) %>% summarise(total_duration = sum(duration, na.rm = TRUE))

# Alternatively, you can calculate the average duration per month: # monthly_duration <- df_selected %>% # group_by(month) %>% # summarise(avg_duration = mean(duration, na.rm = TRUE))

# Create a simple line chart to reflect the trend in total duration across months ggplot(monthly_duration, aes(x = month, y = total_duration, group = 1)) + geom_line(color = “blue”, size = 1) + # Line graph to show total duration over months geom_point(color = “red”, size = 2) + # Add points for each month’s total duration labs(title = “Total Duration by Month”, x = “Month”, y = “Total Duration (seconds)”) + theme_minimal() # Create a bar chart to show the total duration by month ggplot(monthly_duration, aes(x = month, y = total_duration)) + geom_bar(stat = “identity”, fill = “blue”) + # Bar chart showing total duration by month labs(title = “Total Duration by Month”, x = “Month”, y = “Total Duration (seconds)”) + theme_minimal() head(df_selected) # Load necessary libraries library(ggplot2) library(dplyr)

# Group by ‘member_casual’ and count occurrences member_counts <- df_selected %>% group_by(member_casual) %>% summarise(count = n())

# Create a bar chart to show the difference between Casual and Member ggplot(member_counts, aes(x = member_casual, y = count, fill = member_casual)) + geom_bar(stat = “identity”) + # Bar chart showing counts of Casual vs Member labs(title = “Comparison of Casual vs Member Rides”, x = “Member Type”, y = “Number of Rides”) + theme_minimal() + scale_fill_manual(values = c(“blue”, “orange”)) # Optional: Customize colors