**Case study: How does a bike-share navigate speedy success?**

**ASK**

Guiding questions

● What is the problem you are trying to solve?

*Understand how casual riders and annual members use Cyclistic bikes differently.*

● How can your insights drive business decisions?

*From these insights, my team will design a new marketing strategy to convert casual riders into annual members.*

Key tasks

● Identify the business task

● Consider key stakeholders

* *Lily Moreno: The director of marketing and manager*
* *Marketing analyst team at Cyclistic*
* *Cyclistic executive team*

● A clear statement of the business task.

*Design marketing strategies aimed at converting casual riders into annual members, with the purpose of maximizing the number of annual members and therefore, company’s profits.*

**PREPARE**

Guiding questions

● Where is your data located?

[*https://divvy-tripdata.s3.amazonaws.com/index.html*](https://divvy-tripdata.s3.amazonaws.com/index.html)

● How is the data organized?

*The data is organized by month and year (or quarter and year).*

● Are there issues with bias or credibility in this data? Does your data ROCCC?

*Data is not updated (2019 and 2020). Both tables share some columns but not all. Some of the shared columns have different names.*

● How are you addressing licensing, privacy, security, and accessibility?

*The data has been made available by Motivate International Inc.*

● How did you verify the data’s integrity?

*By applying the ROCCC concept.*

● How does it help you answer your question?

*I can identify patterns from the two customer kinds (casual riders and members).*

● Are there any problems with the data?

*There are some missing data values in the “gender” and “birthyear” columns.*

Key tasks

● Download data and store it appropriately.

● Identify how it’s organized.

● Sort and filter the data.

● Determine the credibility of the data.

Deliverable

● A description of all data sources used

**PROCESS**

Guiding questions

● What tools are you choosing and why?  
*I chose Google Sheets to process and clean data.*

● Have you ensured your data’s integrity?

*Explained in the next question.*

● What steps have you taken to ensure that your data is clean?

*First of all, check for empty cells, whitespaces and duplicates. The whole row will be removed if it contains an empty cell (filtering). Then, I calculated the ride\_lengh for each trip and converted that outcome into duration format. I also created the weekday column, which each day of the week is represented with a number.*

*Then, tables columns (Cyclist\_2019Q1 and Cyclist\_2020Q1) were renamed with the same name to keep concistensy.*

● How can you verify that your data is clean and ready to analyze?  
*Using tools as conditional formatting, cleans ups and filtering.*

● Have you documented your cleaning process so you can review and share those results?

*Explained above.*

Key tasks

● Check the data for errors.

● Choose your tools.

● Transform the data so you can work with it effectively.

● Document the cleaning process.

Deliverable

● Documentation of any cleaning or manipulation of data

**ANALYZE**

Guiding questions

● How should you organize your data to perform analysis on it?

*Explained above.*

● Has your data been properly formatted?

*Explained above.*

● What surprises did you discover in the data?

*It looks either a customer retained a bike for 4 months (2953:20:22) or was a data entry mistake.*

● What trends or relationships did you find in the data?

*Basic calculations performed with Google Sheets:*

*2019 Q1*

| *ride\_lengh\_mean* | *0:16:40* |
| --- | --- |
| *ride\_lengh\_max* | *2953:20:22* |
| *ride\_lengh\_min* | *0:01:01* |
| *weekday\_mode* | *5* |

*2020 Q1*

| *ride\_lengh\_mean* | *0:22:07* |
| --- | --- |
| *ride\_lengh\_max* | *2607:30:24* |
| *ride\_lengh\_min* | *-0:09:12* |
| *weekday\_mode* | *3* |

*Pivot tables*

*2019 Q1*

|  | Grand Total |  |
| --- | --- | --- |
| *usertype* | AVERAGE of ride\_length | COUNT of trip\_id |
|  |  |  |
| Customer | 1:01:57 | 23163 |
| Subscriber | 0:13:36 | 341905 |
| **Grand Total** | **0:16:40** | **365068** |

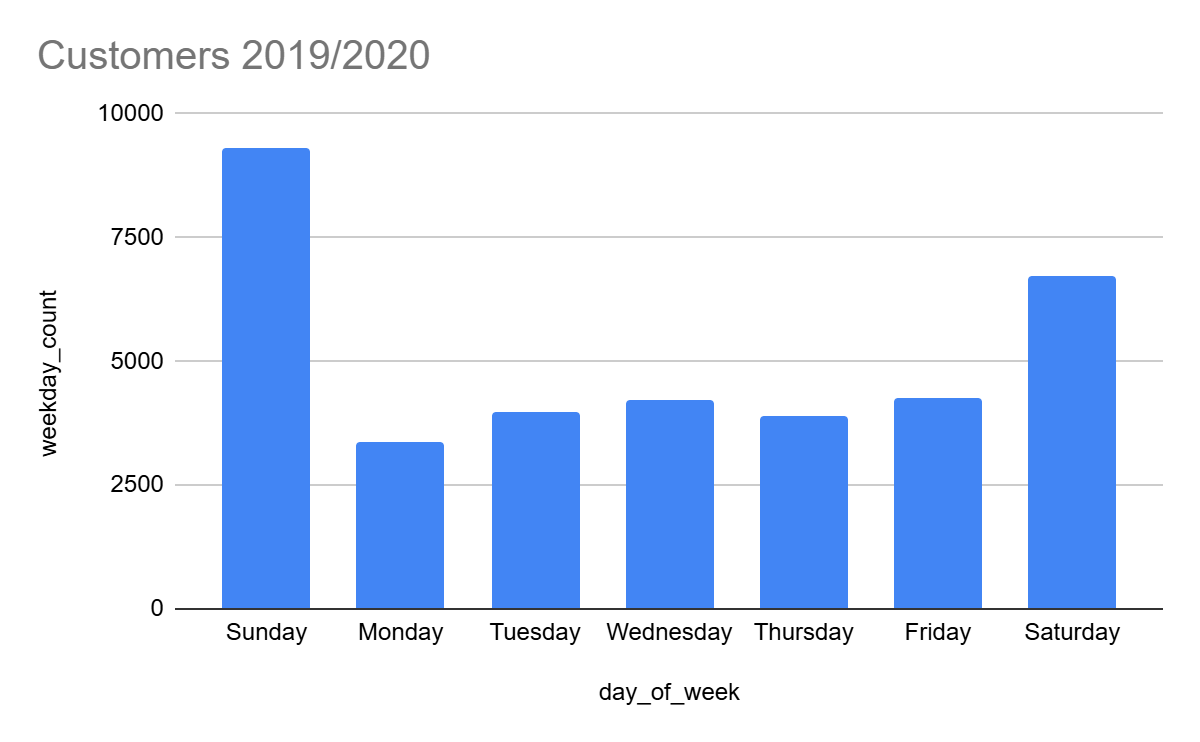
*2020 Q1*

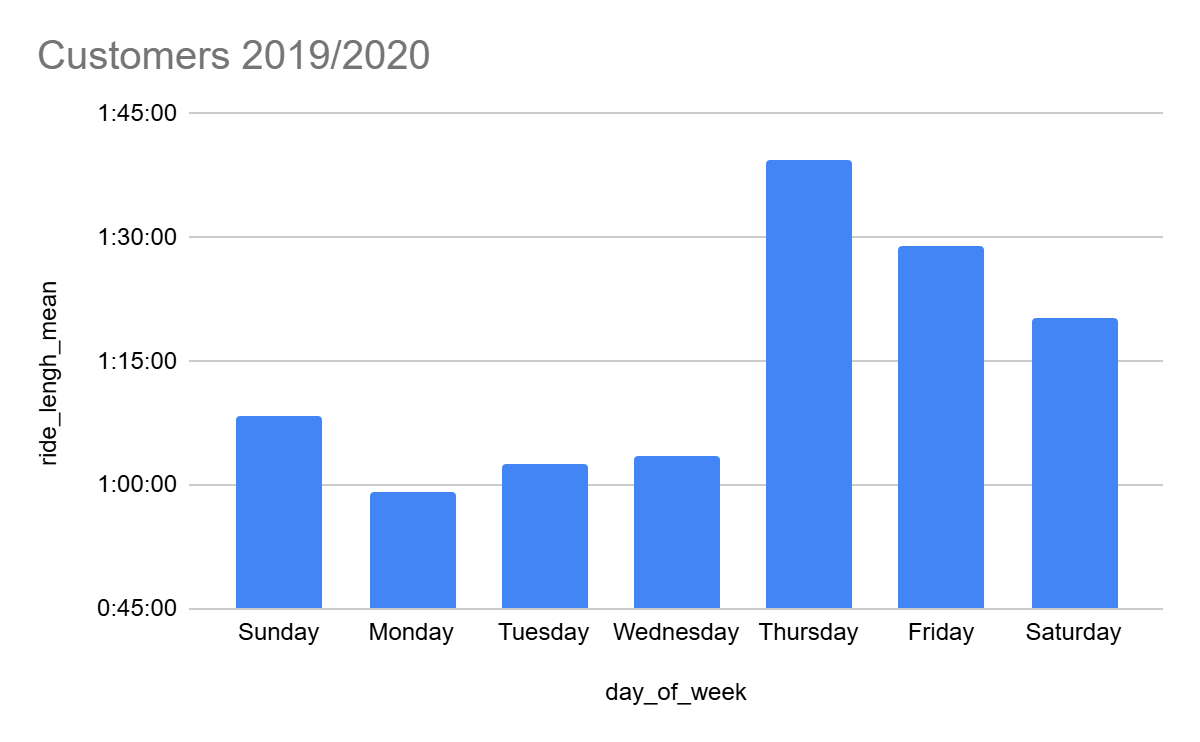
|  |  |  |
| --- | --- | --- |
|  | Grand Total |  |
| *usertype* | AVERAGE of ride\_length | COUNT of day\_of\_week |
|  |  |  |
| casual | 1:35:47 | 48480 |
| member | 0:12:41 | 378407 |
| **Grand Total** | **0:22:07** | **426887** |

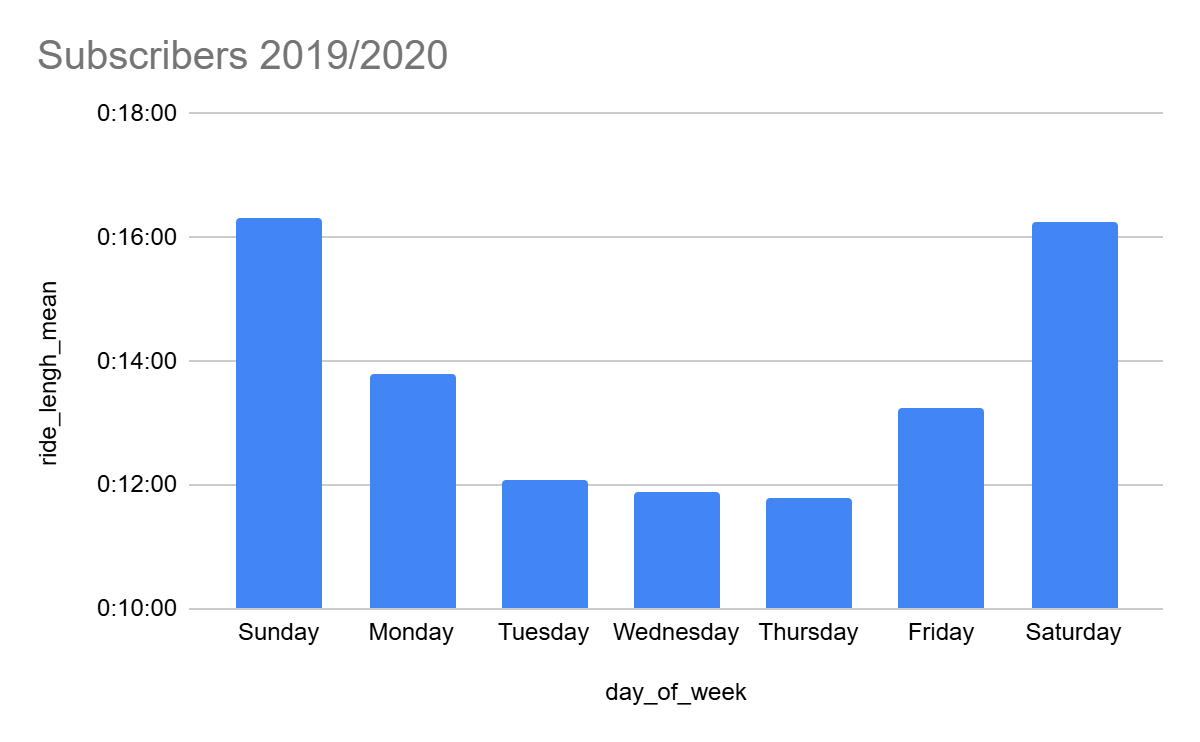
Customers 2019/2020 most busiest days

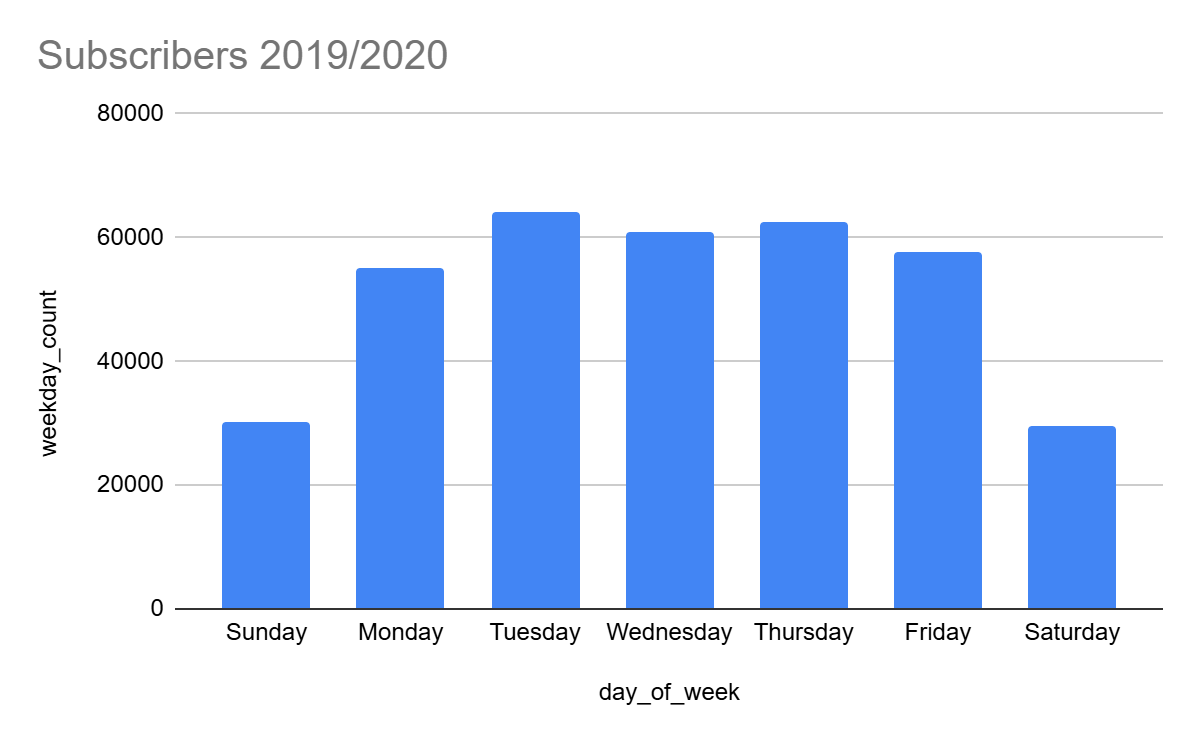
| Customers |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| day\_of\_week | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
| ride\_lengh\_mean | 1:08:23 | 0:59:11 | 1:02:33 | 1:03:29 | 1:39:30 | 1:29:06 | 1:20:18 |
| weekday\_count | 9326 | 3373.5 | 3996 | 4211 | 3906.5 | 4271 | 6736.5 |

| Subscribers |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| day\_of\_week | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
| ride\_lengh\_mean | 0:16:18 | 0:13:49 | 0:12:05 | 0:11:53 | 0:11:47 | 0:13:15 | 0:16:15 |
| weekday\_count | 30098.5 | 55215 | 63986.5 | 60951.5 | 62614 | 57584 | 29706.5 |









● How will these insights help answer your business questions?

*These insights help understand about the time period that each rider class retains the bike and which weekday is most eligible by each rider class.*

Key tasks

● Aggregate your data so it’s useful and accessible.

● Organize and format your data.

● Perform calculations.

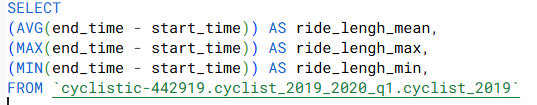
● Identify trends and relationships.

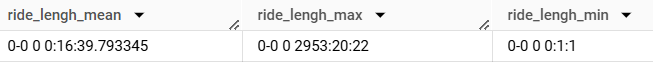
Deliverable

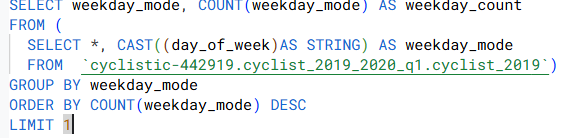
● A summary of your analysis

**SQL**

***cyclist\_2019***

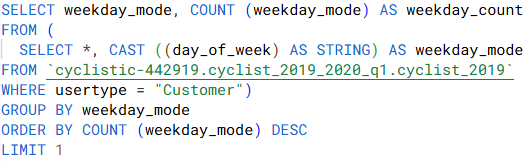
****

****

****

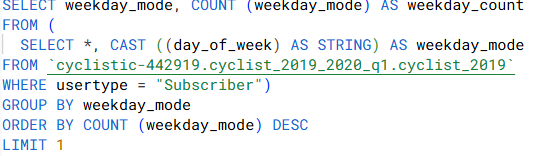
****

*Customers*

****

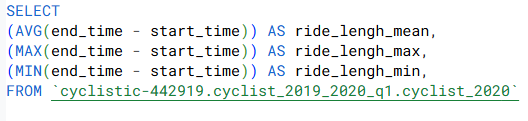
****

*Subscribers*

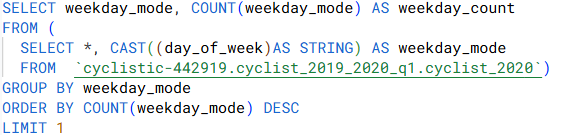
**

**

**cyclist\_2020**

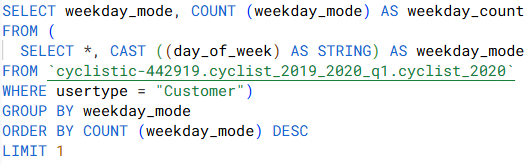






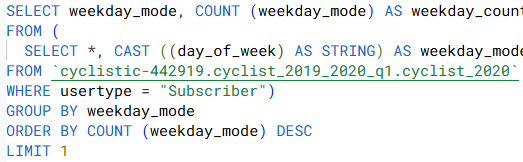


*Customers*

**

**

*Subscribers*





**RStudio**

#Join tables

CK <- rbind(Cyclistic\_2019\_RStudio, Cyclistic\_2020\_RStudio)

Studio

*Ride length time mean group by “usertype” and “day\_of\_week”*

# Convert start\_time and end\_time (as.POSIXct)

CK$start\_time <- as.POSIXct(CK$start\_time, format="%Y-%m-%d %H:%M:%S")

CK$end\_time <- as.POSIXct(CK$end\_time, format="%Y-%m-%d %H:%M:%S")

# Difference between end\_time and start time

CK$ride\_length\_seconds <- as.numeric(difftime(CK$end\_time, CK$start\_time, units = "secs"))

----

# Create a vector of day names corresponding to the numbers 1-7

day\_names <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")

# Convert the day\_of\_week column to a factor with the corresponding day names

CK$day\_of\_week <- factor(CK$day\_of\_week, levels = 1:7, labels = day\_names)

----

# Load dplyr package for easy data manipulation

library(dplyr)

# Calculate mean of ride\_lengh\_seconds for each "usertype" and "day\_of\_week".

mean\_ride\_length\_by\_day <- CK %>%

group\_by(usertype, day\_of\_week) %>%

summarise(mean\_ride\_length = mean(ride\_length\_seconds\_revised, na.rm = TRUE))

# Convert the mean ride length (in seconds) back to HH:MM:SS

# Function to convert seconds back to HH:MM:SS format

convert\_seconds\_to\_time <- function(seconds) {

hours <- floor(seconds / 3600)

minutes <- floor((seconds %% 3600) / 60)

remaining\_seconds <- seconds %% 60

# Ensure all values are integers and format them properly with leading zeros

return(sprintf("%02d:%02d:%02d", as.integer(hours), as.integer(minutes), as.integer(remaining\_seconds)))

}

mean\_ride\_length\_by\_day$mean\_ride\_length\_time <- sapply(mean\_ride\_length\_by\_day$mean\_ride\_length, convert\_seconds\_to\_time)

print(mean\_ride\_length\_by\_day)

| **usertype day\_of\_week mean\_ride\_length mean\_ride\_length\_time**  **<chr> <fct> <dbl> <chr>**  **1 Customer Sunday 5061. 01:24:21**  **2 Customer Monday 3938. 01:05:38**  **3 Customer Tuesday 4173. 01:09:33**  **4 Customer Wednesday 4091. 01:08:11**  **5 Customer Thursday 7729. 02:08:49**  **6 Customer Friday 5714. 01:35:13**  **7 Customer Saturday 4951. 01:22:30**  **8 Subscriber Sunday 973. 00:16:12**  **9 Subscriber Monday 822. 00:13:42**  **10 Subscriber Tuesday 722. 00:12:01**  **11 Subscriber Wednesday 712. 00:11:51**  **12 Subscriber Thursday 707. 00:11:47**  **13 Subscriber Friday 797. 00:13:16**  **14 Subscriber Saturday 974. 00:16:14**  *Weekday count per usertype and day*  #Calculate weekday\_count for "usertype" and "day\_of\_week"  weekday\_count\_per\_day <- CK %>%  group\_by(usertype, day\_of\_week) %>%  count (usertype, day\_of\_week)  **usertype day\_of\_week n**  **<chr> <fct> <int>**  **1 Customer Sunday 18652**  **2 Customer Monday 6747**  **3 Customer Tuesday 7992**  **4 Customer Wednesday 8422**  **5 Customer Thursday 7815**  **6 Customer Friday 8542**  **7 Customer Saturday 13473**  **8 Subscriber Sunday 60197**  **9 Subscriber Monday 110430**  **10 Subscriber Tuesday 127973**  **11 Subscriber Wednesday 121903**  **12 Subscriber Thursday 125228**  **13 Subscriber Friday 115168**  **14 Subscriber Saturday 59413**  # Bar chart mean\_ride\_length\_time vs. day\_of\_week, grouped by usertype  ggplot(mean\_ride\_length\_by\_day, aes(fill=usertype, y=mean\_ride\_length\_time, x=day\_of\_week)) +  geom\_bar(position="dodge", stat="identity")    # Bar chart mean\_ride\_length\_time vs. day\_of\_week, grouped by usertype  ggplot(weekday\_count\_per\_day, aes(fill=usertype, y=n, x=day\_of\_week)) +  geom\_bar(position="dodge", stat="identity") + scale\_y\_continuous(  name = "Number of trips per day")    #Combine data frames mean\_ride\_length\_by\_day and weekday\_count\_per\_day  correlation\_RN <- merge(mean\_ride\_length\_by\_day, weekday\_count\_per\_day)  #Correlations  correlation\_customer <- subset(correlation\_RN, usertype == "Customer")  correlation\_subscriber<- subset(correlation\_RN, usertype == "Subscriber")  correlation1 <- cor(correlation\_customer$mean\_ride\_length, correlation\_customer$n)  correlation2 <- cor(correlation\_subscriber$mean\_ride\_length, correlation\_subscriber$n)  ggplot(correlation\_customer, aes(x = correlation\_customer$n, y = correlation\_customer$mean\_ride\_length)) +  geom\_point(color = "blue", size = 3, alpha = 0.7) + # Scatter plot  geom\_smooth(method = "lm", color = "red", linetype = "dashed") + # Trend line  annotate("text", x = min(correlation\_customer$n), y = max(correlation\_customer$mean\_ride\_length),  label = paste("Correlation:", round(correlation1, 2)),  hjust = 0, size = 5, color = "darkred") + # Add correlation text  labs(  title = "Correlation between ride length mean and daily number of trips for Customers",  x = "number\_of\_trips\_per\_day",  y = "mean\_ride\_length (seconds)"  ) +  theme\_minimal()  ggplot(correlation\_customer, aes(x = correlation\_customer$n, y = correlation\_customer$mean\_ride\_length)) +  geom\_point(color = "blue", size = 3, alpha = 0.7) + # Scatter plot  geom\_smooth(method = "lm", color = "red", linetype = "dashed") + # Trend line  annotate("text", x = min(correlation\_customer$n), y = max(correlation\_customer$mean\_ride\_length),  label = paste("Correlation:", round(correlation1, 2)),  hjust = 0, size = 5, color = "darkred") + # Add correlation text  labs(  title = "Correlation between ride length mean and daily number of trips for Customers",  x = "number\_of\_trips\_per\_day",  y = "mean\_ride\_length (seconds)"  ) +  theme\_minimal()  ggplot(correlation\_customer, aes(x = correlation\_subscriber$n, y = correlation\_subscriber$mean\_ride\_length)) +  geom\_point(color = "blue", size = 3, alpha = 0.7) + # Scatter plot  geom\_smooth(method = "lm", color = "red", linetype = "dashed") + # Trend line  annotate("text", x = min(correlation\_subscriber$n), y = max(correlation\_subscriber$mean\_ride\_length),  label = paste("Correlation:", round(correlation2, 2)),  hjust = 0, size = 5, color = "darkred") + # Add correlation text  labs(  title = "Correlation between ride length mean and daily number of trips for Subscribers",  x = "number\_of\_trips\_per\_day",  y = "mean\_ride\_length (seconds)"  ) +  theme\_minimal()      Guiding questions  ● Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?  Yes  ● What story does your data tell?  Ride length time for Customers is, on average, higher than Subscribers time.  Number of trips per day for Subscribers is, on average, higher than Customers time. The most preferred days for Customers are the weekend days, whereas the least preferred days for Subscribers is during the weekend, too.  There’s a strong relationship between ride length time average and number of trips for Subscribers. When the number of trips is higher, ride length time decreases.  ● How do your findings relate to your original question?  These findings help to understand the Customer group behaviour (which will be used for marketing strategies to attract more potential subscribers.  ● Who is your audience? What is the best way to communicate with them?   * *Lily Moreno: The director of marketing and manager* * *Marketing analyst team at Cyclistic* * *Cyclistic executive team*   The best way to communicate is a slides presentation.  ● Can data visualization help you share your findings?  Yes.  ● Is your presentation accessible to your audience?  Yes. |
| --- |
|  |
| |  | | --- | |