**Case study: how does a bike share navigate speedy success**

Deze case study is onderdeel van het Google Data Analytics Certificate. Ik ga een dataset analyseren van een fictief bedrijf. In deze case study zal ik de stappen van het data analyse process doorlopen: ask, prepare, process, analyze, share en act.

**Scenario**

Ik werk als junior data analist voor Cyclistic. Cyclistic is een bedrijf dat deelfietsen aanbiedt in Chicago. Ze verhuren zowel elektrische fietsen als standaard fietsen die tussen 692 docking stations gebruikt kunnen worden. Cyclistic heeft twee soorten klanten: casual riders en abonnees. Casual riders zijn klanten die een enkele rit of dagkaart hebben gekocht en abonnees hebben een jaarabonnement aangeschaft. De marketing directeur gelooft dat het maximaliseren van het aantal jaarabonnementen belangrijk is voor de toekomst van het bedrijf. Als onderdeel van het marketing analyse team is het mijn taak om te begrijpen wat het verschil is tussen casual riders en klanten die een jaarabonnement hebben. Mijn team ontwikkelt een strategie om casual riders om te zetten naar abonnees.

**Ask**

De vraag die beantwoord moet worden is:

In welk opzicht verschillen casual riders van klanten met een jaarabonnement?

Het antwoord op deze vraag moet ondersteund worden door data inzichten en data visualisaties die aan het marketing team worden gepresenteerd.

**Prepare**

De dataset die ik ga gebruiken is publieke data die beschikbaar is gemaakt door Motivate International Inc onder deze licentie: [Licentie](https://ride.divvybikes.com/data-license-agreement).

Link naar dataset: [Dataset](https://divvy-tripdata.s3.amazonaws.com/index.html). De data is in een CSV format. De data die ik gebruik is van mei 2022 tot april 2023 en is als laatst bijgewerkt op 4-05-2023.

De dataset voldoet aan de ROCCC analyse volgens de volgende criteria:

R - Reliable: De data komt van een betrouwbare bron die de data zonder vooroordelen heeft verzameld.

O – Original: De originele data is publiek.

C – Comprehensive: De data is compleet en bevat geen gaten.

C – Current: De data wordt maandelijks vernieuwd en is dus up to date.

C – Cited: Motivate International Inc heeft de data verzameld en dit is een betrouwbare bron.

**Process**

Ik heb gekozen om de data op te schonen in R studio. Hiervoor heb ik een aantal redenen:

* De dataset is erg groot en hiervoor is Excel niet geschikt.
* R studio kan grote hoeveelheden data in korte tijd bewerken.
* Door een programmeertaal te gebruiken kan het proces gereproduceerd worden voor een ander project.

1. laden van R studio packages. Om de data op te schonen maak ik gebruik van de tidyverse en lubridate packages.

# instaleren van de tidyverse en lubridate.  
install.packages('tidyverse')  
install.packages('lubridate')  
library(tidyverse)  
library(lubridate)

1. Importing data in R

# data laden in R studio  
mei\_2022 <- read.csv("Google bike sharing data/CSV files/202205-divvy-tripdata.csv")  
juni\_2022 <- read.csv("Google bike sharing data/CSV files/202206-divvy-tripdata.csv")  
juli\_2022 <- read.csv("Google bike sharing data/CSV files/202207-divvy-tripdata.csv")  
augustus\_2022 <- read.csv("Google bike sharing data/CSV files/202208-divvy-tripdata.csv")  
september\_2022 <- read.csv("Google bike sharing data/CSV files/202209-divvy-publictripdata.csv")  
oktober\_2022 <- read.csv("Google bike sharing data/CSV files/202210-divvy-tripdata.csv")  
november\_2022 <- read.csv("Google bike sharing data/CSV files/202211-divvy-tripdata.csv")  
december\_2022 <- read.csv("Google bike sharing data/CSV files/202212-divvy-tripdata.csv")  
januari\_2023 <- read.csv("Google bike sharing data/CSV files/202301-divvy-tripdata.csv")  
februari\_2023 <- read.csv("Google bike sharing data/CSV files/202302-divvy-tripdata.csv")  
maart\_2023 <- read.csv("Google bike sharing data/CSV files/202303-divvy-tripdata.csv")  
april\_2023 <- read.csv("Google bike sharing data/CSV files/202304-divvy-tripdata.csv")

1. Data validatie om te zien of de kolomnamen overeenkomen.

# ophalen kolomnamen om te checken of deze met elkaar overeenkomen  
colnames(mei\_2022)  
colnames(juni\_2022)  
colnames(juli\_2022)  
colnames(september\_2022)  
colnames(oktober\_2022)  
colnames(november\_2022)  
colnames(december\_2022)  
colnames(januari\_2023)  
colnames(februari\_2023)  
colnames(maart\_2023)  
colnames(april\_2023)

1. Data samenvoegen in 1 dataframe.

# Samenvoegen data in data frame ritten\_totaal  
ritten\_totaal <- rbind(mei\_2022,juni\_2022,juli\_2022,augustus\_2022,september\_2022,oktober\_2022,november\_2022,  
 december\_2022,januari\_2023,februari\_2023,maart\_2023,april\_2023)

1. Geef korte samenvatting van dataframe ritten\_totaal.

# samenvatting dataframe met behulp van de glimpse functie  
glimpse(ritten\_totaal) # 13 kolommen en 5,859,061 rijen aan data.

1. Sla het nieuwe dataframe op als nieuw bestand.

# opslaan gecombineerde dataframe  
write.csv(ritten\_totaal, file = 'Google bike sharing data/ritten\_totaal.csv',row.names = FALSE)

Nu de data samengevoegd is in 1 dataframe moet de data opgeschoond worden. Er zijn een aantal taken die hiervoor gedaan moeten worden.

1. Verwijderen ontbrekende data

# Ontbrekende waarden verwijderen  
ritten\_totaal\_schoon <-ritten\_totaal[complete.cases(ritten\_totaal),]

1. Verwijderen dubbele waarden.

# Dubbele waarden verwijderen  
ritten\_totaal\_schoon <- unique(ritten\_totaal\_schoon)

1. Voeg kolommen toe met datum, dag, maand en jaar. Dit doe ik om de data op deze kolommen samen te kunnen voegen.

# voeg kolommen toe voor datum  
ritten\_totaal\_schoon$datum <- as.Date(ritten\_totaal\_schoon$started\_at)  
ritten\_totaal\_schoon$maand <- format(as.Date(ritten\_totaal\_schoon$datum), "%m")   
ritten\_totaal\_schoon$dag <- format(as.Date(ritten\_totaal\_schoon$datum), "%d")  
ritten\_totaal\_schoon$jaar <- format(as.Date(ritten\_totaal\_schoon$datum), "%Y")  
ritten\_totaal\_schoon$dag\_van\_de\_week <-format(as.Date(ritten\_totaal\_schoon$datum), "%A")

1. Voeg een kolom toe die berekent hoe lang de rit duurt en maak van deze kolom een numerieke waarde.

# kolom trip\_duur toevoegen  
ritten\_totaal\_schoon$trip\_duur <- difftime(ritten\_totaal\_schoon$ended\_at, ritten\_totaal\_schoon$started\_at)  
ritten\_totaal\_schoon$trip\_duur <- as.numeric(as.character(ritten\_totaal\_schoon$trip\_duur))

1. Verwijder data waarbij einddatum eerder is dan startdatum.

# einddatum voor startdatum verwijderen  
ritten\_totaal\_schoon <- ritten\_totaal\_schoon[!ritten\_totaal\_schoon$trip\_duur<0,]

1. Verwijder ritten die meer dan 24 uur duren. Deze fietsen zijn niet op tijd ingeleverd en maken de data minder betrouwbaar.

# verwijderen ritten langer dan 24 uur  
ritten\_totaal\_schoon <- ritten\_totaal\_schoon[!ritten\_totaal\_schoon$trip\_duur>86400,]

1. Geef samenvatting van de schone dataset.

# samenvatting schone data  
colSums(is.na(ritten\_totaal\_schoon))  
summary(ritten\_totaal\_schoon)  
View(head(ritten\_totaal\_schoon))

1. Sla opgeschoonde dataset op voor analyse.

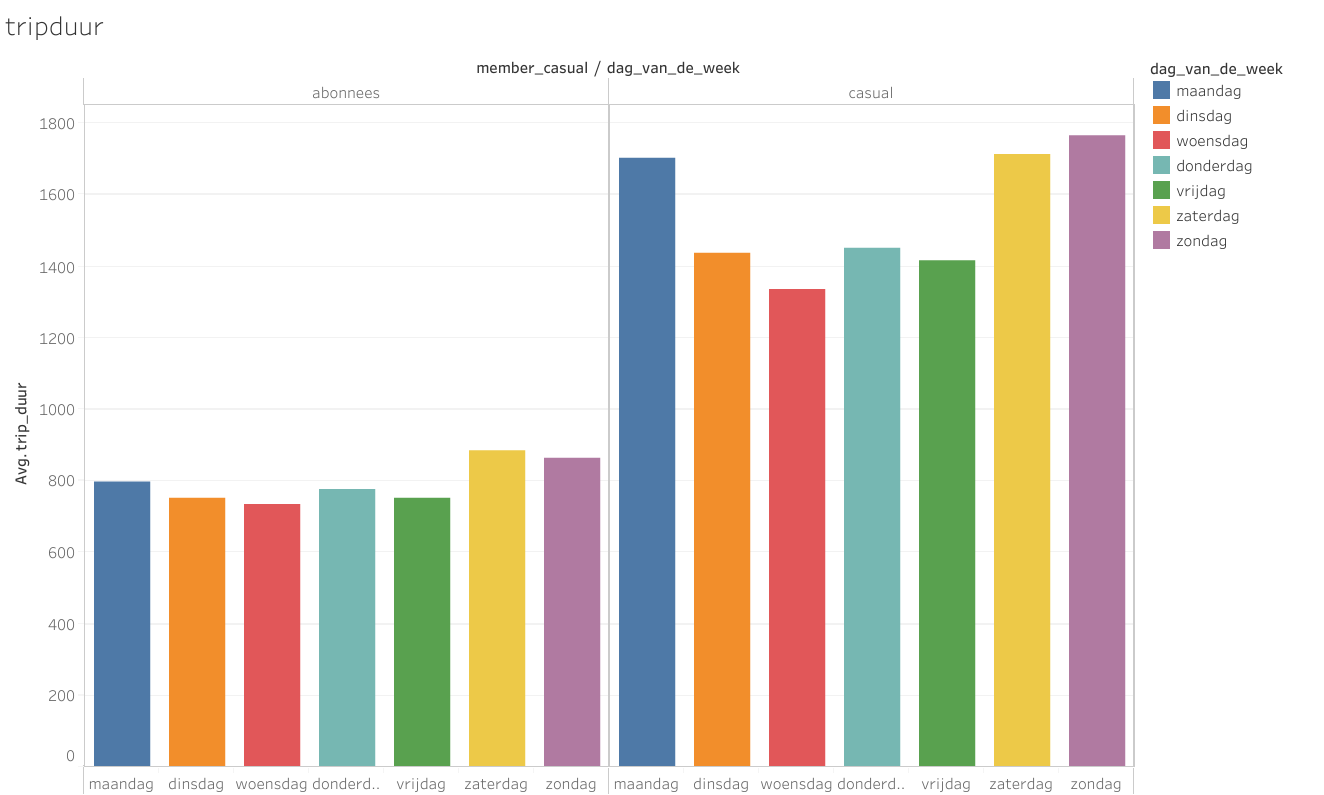
# opslaan data  
write.csv(ritten\_totaal\_schoon,file = 'Google bike sharing data/ritten\_totaal\_schoon.csv',row.names = FALSE)

**Analyse**

*Gemiddelde triplengte*

Om de verschillen tussen casual riders en abonnees te bepalen vergelijk ik eerst het gemiddelde maximum en minimum tussen de twee groepen. Uit deze analyse blijkt dat casual riders gemiddeld langere ritten hebben (mean = 1591 seconden en mediaan =945 seconden) dan abonnees (mean = 794 seconden en mediaan = 577 seconden).

# samenvatting data tussen casual riders en abonnees  
aggregate(ritten\_totaal\_schoon$trip\_duur~ritten\_totaal\_schoon$member\_casual, FUN = mean)  
aggregate(ritten\_totaal\_schoon$trip\_duur~ritten\_totaal\_schoon$member\_casual, FUN = median)  
aggregate(ritten\_totaal\_schoon$trip\_duur~ritten\_totaal\_schoon$member\_casual, FUN = max)  
aggregate(ritten\_totaal\_schoon$trip\_duur~ritten\_totaal\_schoon$member\_casual, FUN = min)

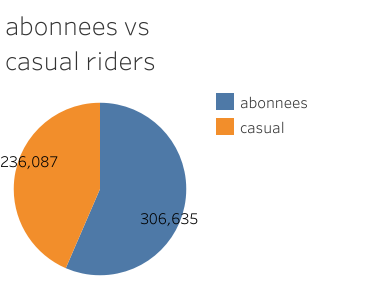


In bovenstaande grafiek is te zien dat casual riders gemiddeld langere ritten maken dan abonnees. Ook is te zien dat weekendtrips gemiddeld langer duren dan doordeweekse ritten bij casual riders. Bij de abonnees is het verschil in triplengte op de verschillende dagen verwaarloosbaar.

*Abonnees vs casual riders*

Vervolgens vergelijk ik het aantal casual riders met het aantal abonnees met behulp van een piechart.

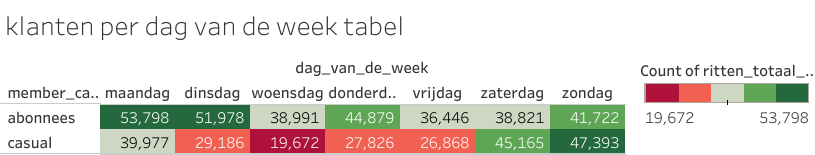
# aantal abbonees vs caual riders  
View(table(ritten\_totaal\_schoon$member\_casual))

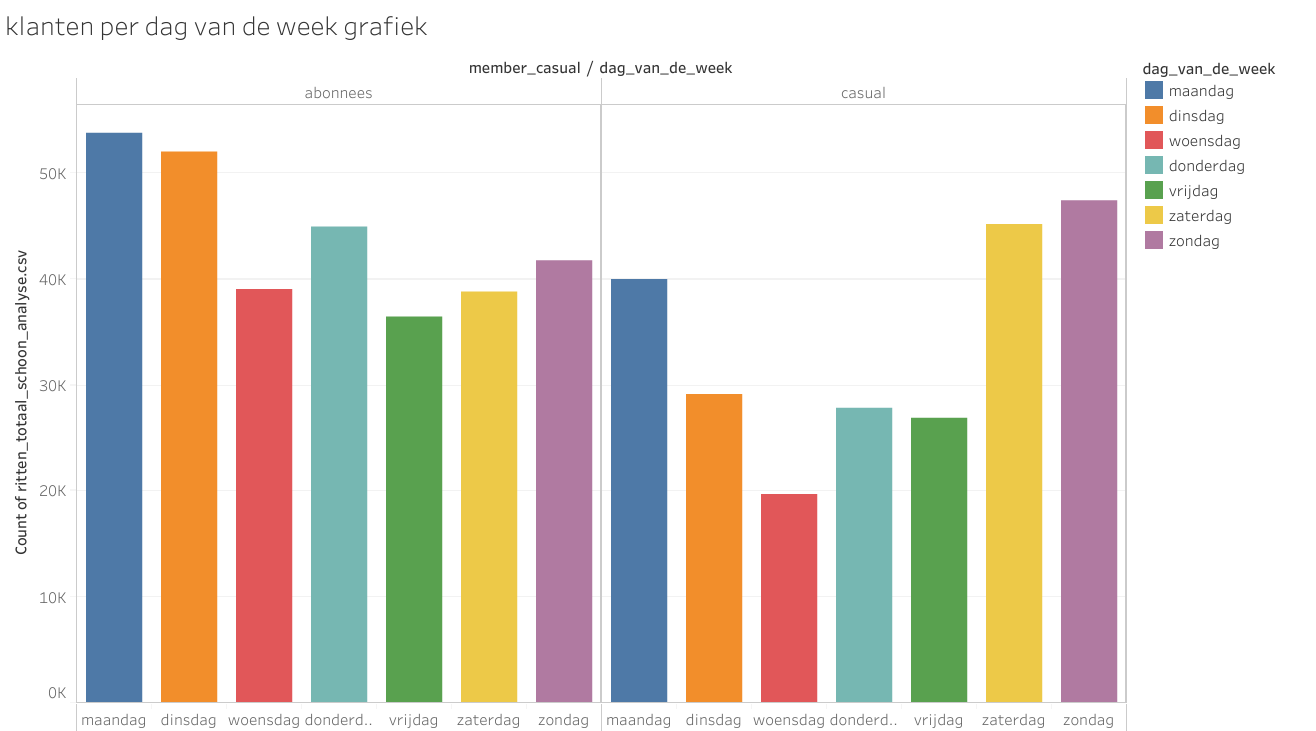
*+*

*Klanten per dag van de week*

Vervolgens analyseer ik het aantal klanten per categorie per dag van de week.

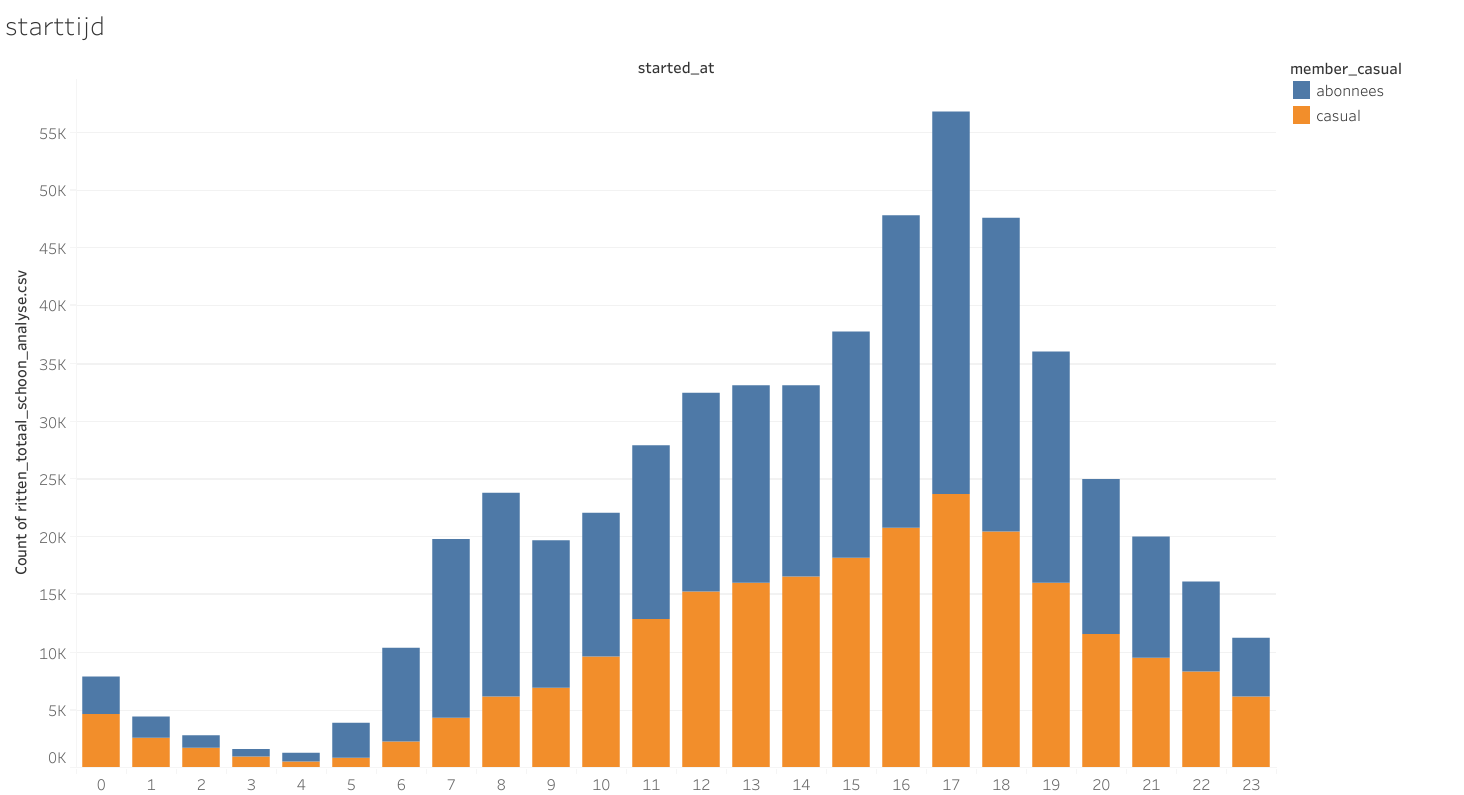
ritten\_totaal\_schoon$dag\_van\_de\_week <- ordered(ritten\_totaal\_schoon$dag\_van\_de\_week, levels=c("zondag","maandag","dinsdag","woensdag",   
 "donderdag", "vrijdag", "zaterdag"))  
aggregate(ritten\_totaal\_schoon$trip\_duur~ritten\_totaal\_schoon$member\_casual + ritten\_totaal\_schoon$dag\_van\_de\_week, FUN = mean)





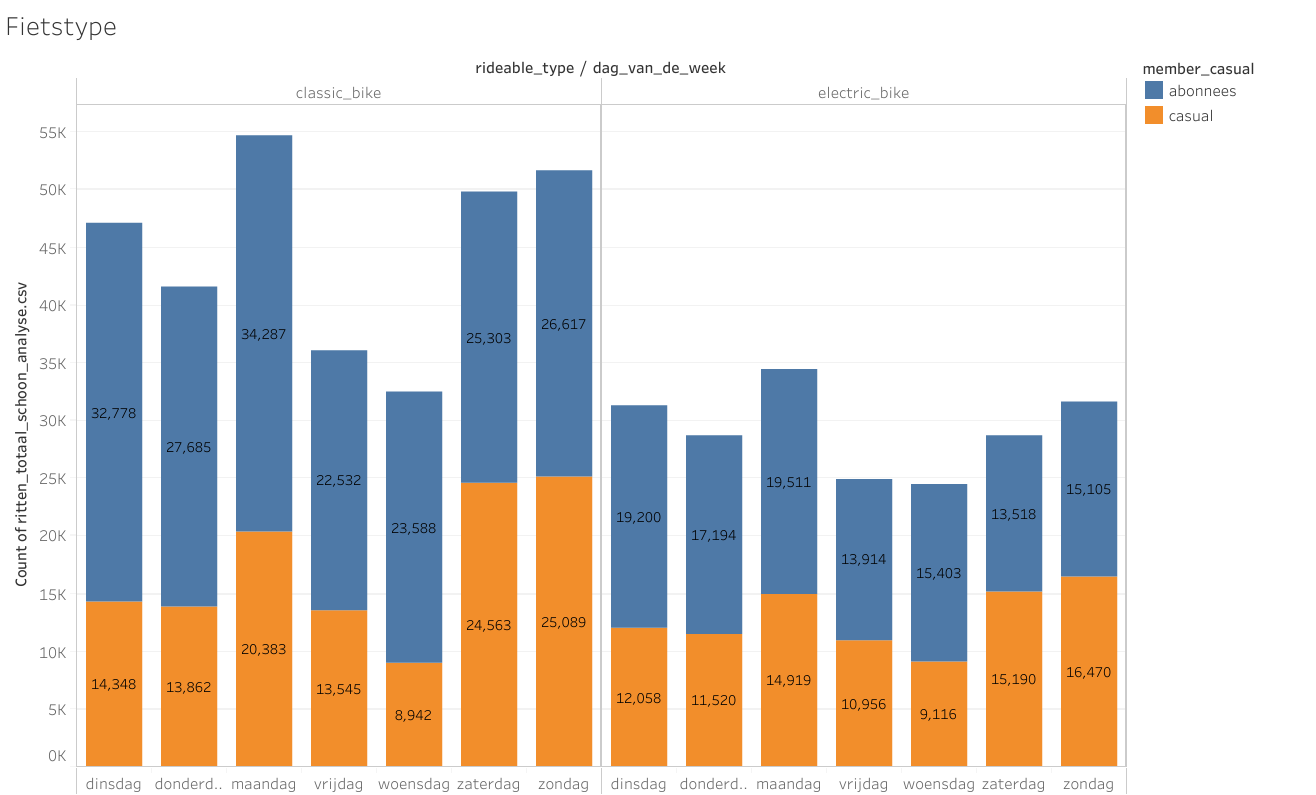
Zoals in bovenstaande figuren te zien is, zijn er een aantal belangrijke verschillen tussen de twee groepen. Er is een groot verschil tussen werkdagen en het weekend bij de casual riders, terwijl het aantal ritten bij abonnees redelijk constant verdeeld is over de week. Dit kan betekenen dat abonnees de fietsen gebruiken als woon-werk verkeer, terwijl de casual riders de fietsen gebruiken voor ontspanning.

*Klanten vs tijd*



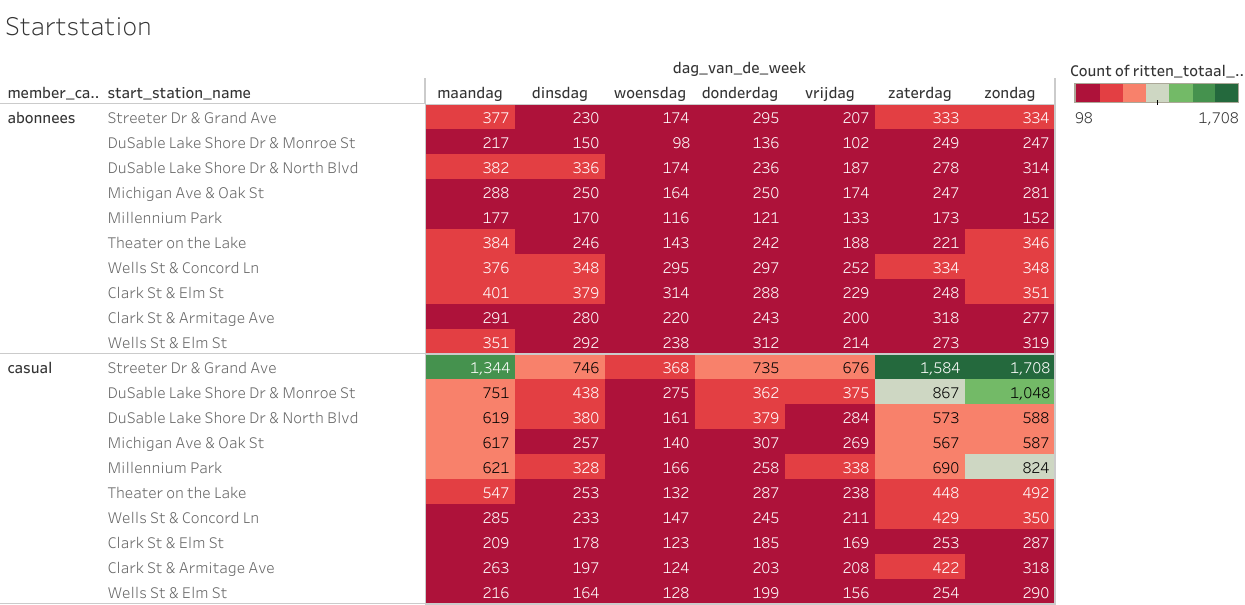
In bovenstaande grafiek is te zien dat abonnees vaak in de ochtend- en avondspits fietsen huren, terwijl casual riders meer verspreid over de middag fietsen huren. Een mogelijke verklaring hiervoor is dat abonnees fietsen gebruiken voor woon-werk verkeer, terwijl casual riders fietsen voor ontspanning.

*Fietstype*

**

Zoals in bovenstaande grafiek te zien is, is de standaard fiets populairder dan de elektrische fiets . Dit verschil is groter bij abonnees dan bij casual riders. Uit de grafiek is op te maken dat abonnees een duidelijke voorkeur hebben voor standaard fietsen en dat casual riders een minder sterke voorkeur hebben voor een fietstype.

*Locatie*



In bovenstaande tabel zijn de 10 meest bezochte startstations weergegeven per dag van de week. Hieruit blijkt dat er voornamelijk fietsen geleend worden vanuit Streeter Dr & Grand Ave en DuSable Lake Shore Dr & Monroe St gedurende het weekend.



Een vergelijkbaar fenomeen is te zien bij de eindstations.

**Share**

*Belangrijkste leerpunten*

Door de data te analyseren heb ik 4 leerpunten geïdentificeerd die een belangrijke rol spelen om de volgende vraag te kunnen beantwoorden: In welk opzicht verschillen casual riders van klanten met een jaarabonnement?

* Casual riders gebruiken de fietsen voornamelijk in het weekend, terwijl abonnees een constant gebruikerspatroon tonen.
* Casual riders hebben een ongeveer 50% langere ritduur ten opzichte van abonnees.
* Casual riders huren fietsen vooral gedurende de middag en het begin van de avond, terwijl abonnees meer fietsen huren tijdens woon-werk verkeer uren.
* Casual riders tonen een kleine voorkeur voor standaard fietsen, terwijl abonnees een significante voorkeur voor standaard fietsen tonen. Dit is vooral te zien gedurende de werkweek.

*Conclusie*

Uit de data en de leerpunten blijkt dat abonnees de fietsen voornamelijk gebruiken voor woon-werk verkeer, terwijl de casual riders ze gebruiken voor ontspanning en vrije tijdsbesteding. Om casual riders om te zetten naar abonnees moet de fiets een dagelijks vervoermiddel worden om mee op en neer naar het werk te gaan.

**Act**

*Aanbevelingen*

Ik heb drie aanbevelingen gevonden die het marketing team kan doen om meer abonnees te krijgen.

* Introduceer weekpassen. Door het voordelig maken van fietsverhuur gedurende de werkweek wordt de fiets een goed alternatief voor de auto als transportmiddel naar het werk.
* Promoot abonnementen voornamelijk tijdens het weekend.
* Maak gebruik van advertenties rond de populairste stations in het bijzonder Streeter Dr & Grand Ave en DuSable Lake Shore Dr & Monroe. Als dit succesvol is kan dit uitgebreid worden naar andere populaire stations.