Cyclistic Case Study

mahmoud hassan

November 30th, 2022

### Background

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked and locked into a network of 692 stations across Chicago.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

### What Are We Talking About?

* How Do **Annual Members** And **Casual Riders** Use Cyclistic bikes Differently?
* Why would casual riders buy Cyclistic annual memberships?
* I used Cyclistic’s historical trip data to analyze and identify trends.to visit data source click here [link](https://divvy-tripdata.s3.amazonaws.com/index.html)

#### first install.packages()

library("tidyverse")

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library("lubridate")

## Loading required package: timechange  
##   
## Attaching package: 'lubridate'  
##   
## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library("janitor")

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library("dplyr")  
library("scales")

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

library("ggplot2")

#### importing and cleaning data

* Import csv files without blank data which does not read as (NA)

setwd("D:/data analysis/capstone project/case study 1/2022")  
f12<-read.csv("202112-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f1<-read.csv("202201-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f2<-read.csv("202202-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f3<-read.csv("202203-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f4<-read.csv("202204-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f5<-read.csv("202205-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f6<-read.csv("202206-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f7<-read.csv("202207-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f8<-read.csv("202208-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f9<-read.csv("202209-divvy-publictripdata.csv",stringsAsFactors =F,na.strings ="")  
f10<-read.csv("202210-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")  
f11<-read.csv("202211-divvy-tripdata.csv",stringsAsFactors = F,na.strings = "")

#### combining the files into one file

total\_12\_month<-rbind(f12,f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11)

#### remove duplicated data

total\_12\_month<-total\_12\_month[!duplicated(total\_12\_month$ride\_id),]

#### remove incompleted data

total\_12\_month<-total\_12\_month[complete.cases(total\_12\_month),]

#### convert data as character to date and time

total\_12\_month$started\_date<-as.Date(total\_12\_month$started\_at)  
total\_12\_month$ended\_date<-as.Date(total\_12\_month$ended\_at)

#### Create columns to represent month, day and hour of service usage

total\_12\_month$month<-format(as.Date(total\_12\_month$started\_at),"%B")  
total\_12\_month$day<-format(as.Date(total\_12\_month$started\_at),"%A")  
total\_12\_month$hour<-format(as.POSIXct(total\_12\_month$started\_at),format=("%H"))

#### create a column to Calculate the cycling trip time

total\_12\_month$ride\_length<-  
 difftime(total\_12\_month$ended\_at,total\_12\_month$started\_at,units = c('mins'))  
  
total\_12\_month<-total\_12\_month %>% filter(total\_12\_month$ride\_length>0)

#### create a column to represent work day and weekend

total\_12\_month$weekdays<-  
 ifelse(total\_12\_month$day=='Sunday'|total\_12\_month$day=='Saturday','weekend','work')

#### convert data as character to factor

total\_12\_month$month<-  
 factor(total\_12\_month$month,  
 levels =c("December","January","February",  
 "March","April","May","June","July",  
 "August","September","October","November"))  
  
total\_12\_month$day<-  
 factor(total\_12\_month$day,  
 levels =c("Monday","Tuesday","Wednesday",  
 "Thursday", "Friday","Saturday","Sunday"))

#### Saving the result as a CSV

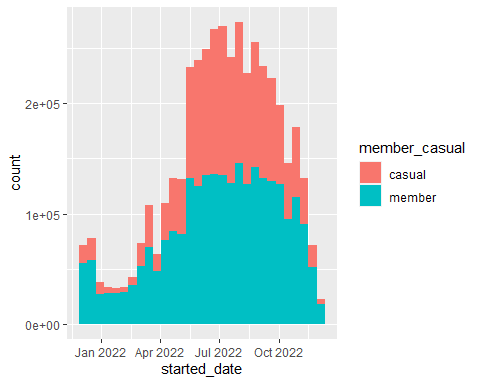
total\_12\_month %>% write.csv("cyclistic\_clean.csv")

#### Data distribution

Here we want to try to answer the most basic questions about how the data is distributed. \* there is increasing in trips in summer for annual members and casual riders

ggplot(aes(x=started\_date,fill=member\_casual),data = total\_12\_month)+  
 geom\_histogram()

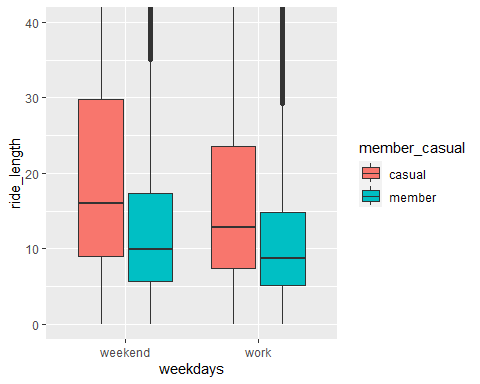
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



* there is an increasing in count of the trips at the weekend and decreasing on working days for casual riders
* there is an increasing in trips on working days and decreasing at the weekends for annual members

ggplot(aes(x=weekdays,y=ride\_length,fill=member\_casual),data=total\_12\_month)+  
 geom\_boxplot()+  
 coord\_cartesian(ylim =c(0,40))

## Don't know how to automatically pick scale for object of type <difftime>.  
## Defaulting to continuous.



* there is an increasing in average ride length of trip for casual riders more than annual members per day #### group subset data by day and calculate mean of (ride\_length)

subset\_cyclistic\_bicke<-total\_12\_month %>% group\_by(day,member\_casual) %>%  
 summarise(ride\_length\_mean=mean(ride\_length))

## `summarise()` has grouped output by 'day'. You can override using the `.groups`  
## argument.

#### separate member\_casual column into to column

subset\_cyclistic\_bicke <- subset\_cyclistic\_bicke%>%   
 mutate( casual= ifelse(member\_casual=='casual', ride\_length\_mean, NA))  
  
subset\_cyclistic\_bicke<-subset\_cyclistic\_bicke%>%   
 mutate( member= ifelse(member\_casual=='member', ride\_length\_mean, NA))

#### separate member\_casual column into to column

subset\_cyclistic\_bicke<-subset\_cyclistic\_bicke%>%   
 group\_by(day) %>% summarise(casual,member)

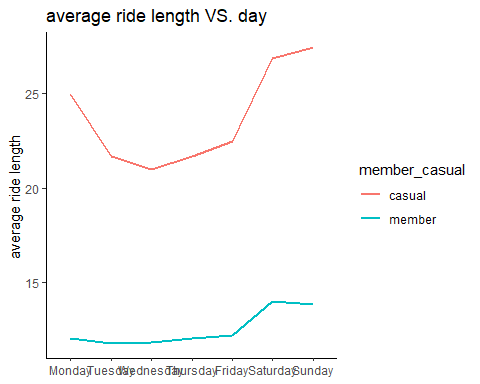
## `summarise()` has grouped output by 'day'. You can override using the `.groups`  
## argument.

subset\_cyclistic\_bicke<-subset\_cyclistic\_bicke %>%  
 summarise(member = max(member, na.rm=TRUE),casual = max(casual, na.rm=TRUE))

#### create line chart for member\_casual and average ride length

ggplot(data =subset\_cyclistic\_bicke ,aes(x=day,group=1))+  
 geom\_line(aes(y=casual,color='casual' ),size=1)+  
 geom\_line(aes(y=member,color='member' ),size=1)+  
 theme\_classic()+theme(axis.title.x = element\_blank())+ylab("average ride length ")+  
 labs(title ='average ride length VS. day ' ,color='member\_casual')

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.



#### Conclusion

* **Casual riders** have many trips on weekends.
* **annual members** riders have many trips on working days.
* average ride length for **casual riders** more than **annual members**
* The evidence is inconclusive
* There may be more than meets the eye.
* Moving forward as is with a conversion marketing strategy is risky

#### Recommendations

* Yearly subscription providing an un limited pass for every weekend
* Focus on the benefits of cycling on health and the environment
* Preference for annual members bike reservations during peak cycling seas