Cyclistic bike-share

### Ask Phase

#### Identify the business task

**“Cyclistic”** is a bike-share company. Until now, the company marketing strategy relied on broad consumer segments. They offered to their customers several pricing plans: **single-ride pass**, **full-day pass** and **annual membership**. The customers who buy **single-ride** or **full-day** pass are referred to as **casual riders**. Customer who purchase **annual membership** are called **members**. Based on the **annual income** a conclusion is made that **annual members** are more **profitable** than **casual riders**. The head of marketing wants a new strategy to be devised for converting casual riders into annual members. In order to do that, however, the marketing team needs to better understand how annual members differ, why casual rider would buy an annual membership.

#### Question:

* How do annual members and casual riders use Cyclistic bikes differently?

### Prepare Phase

#### Guiding questions:

* Where is your data located?
* How is the data organized?
* Are there issues with bias or credibility in this data? Does your data ROCCC?
* How are you addressing licensing, privacy, security, and accessibility?
* How did you verify the data’s integrity?
* How does it help you answer your question?
* Are there any problems with the data?

The data is separate in **12 csv files** which represents the **12 months** of the year. It can be downloaded form a [server](https://divvy-tripdata.s3.amazonaws.com/index.html) and it is license under **“Data license”.** This is public data that can be used to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit using riders’ personally identifiable information. This means that it won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the **Cyclistic** service area or if they have purchased multiple single passes. The data set covers all requirements of **“ROCCC”** (reliable, original, comprehensive, current and cited). The data is organized in rows and columns.

### Process Phase

#### Loading data

df1<-read.csv("~/GoogleProject/csv/202101-divvy-tripdata.csv")  
df2<-read.csv("~/GoogleProject/csv/202102-divvy-tripdata.csv")  
df3<-read.csv("~/GoogleProject/csv/202103-divvy-tripdata.csv")  
.........

#### Cleaning data

###### **Removing any empty rows or columns**

one\_df <- remove\_empty(one\_df, which=c("cols"))  
one\_df <- remove\_empty(one\_df, which=c("rows"))

###### **Omitting NA values in the entire data frame**

new\_one\_df <- na.omit(one\_df)

###### **Checking for missing values**

count(filter(new\_one\_df, start\_station\_name==''),start\_station\_name, member\_casual,sort=TRUE)

###### **Removing duplicates**

new\_one\_df\_no\_dups <- new\_one\_df[!duplicated(new\_one\_df$ride\_id), ]  
print(nrow(new\_one\_df\_no\_dups))

### Analyze Phase

##### Riding length

riding\_len <- new\_one\_df\_no\_dups  
riding\_len <- riding\_len %>% mutate(riding\_length = as.numeric(ended\_at - started\_at)/60)

##### Year and month

riding\_len <- riding\_len %>%   
 mutate(year\_month=paste(strftime(riding\_len$started\_at, "%Y"), "-",  
 strftime(riding\_len$started\_at, "%m"),  
 paste("(", strftime(riding\_len$started\_at, "%b"),")",sep="")))

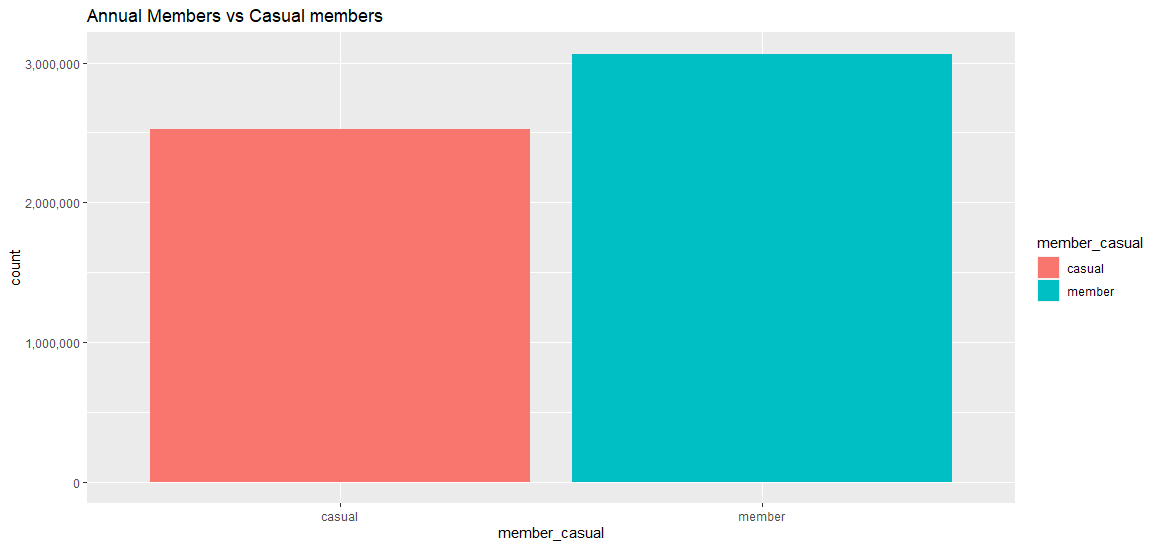
##### Weekday

riding\_len <- riding\_len %>%   
 mutate(weekday=strftime(riding\_len$ended\_at, "%a"))

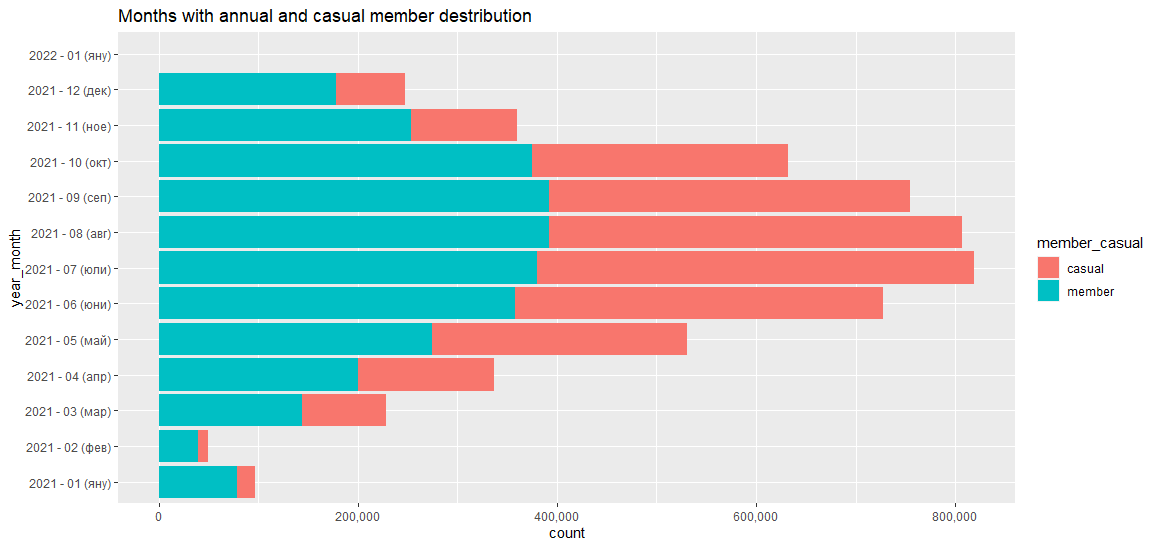
### Conduct Descriptive Analysis

**Comparing the number of members and casual riders**

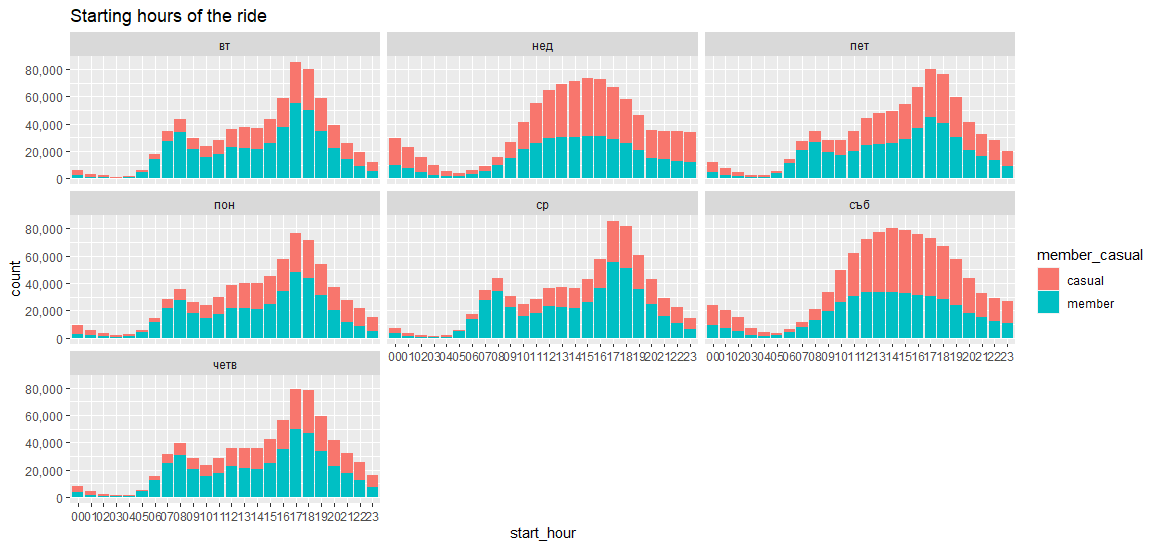
df <- riding\_len  
df %>%   
 group\_by(member\_casual) %>%   
 summarize(count=length(ride\_id),  
 percentage\_of\_total=(length(ride\_id)/nrow(df))\*100)  
  
ggplot(df, aes(member\_casual, fill=member\_casual))+  
 geom\_bar()+  
 labs(title="Chart-1 Member vs Casual distribution")+  
 scale\_y\_continuous(labels=comma)

 **Months with annual and casual members**

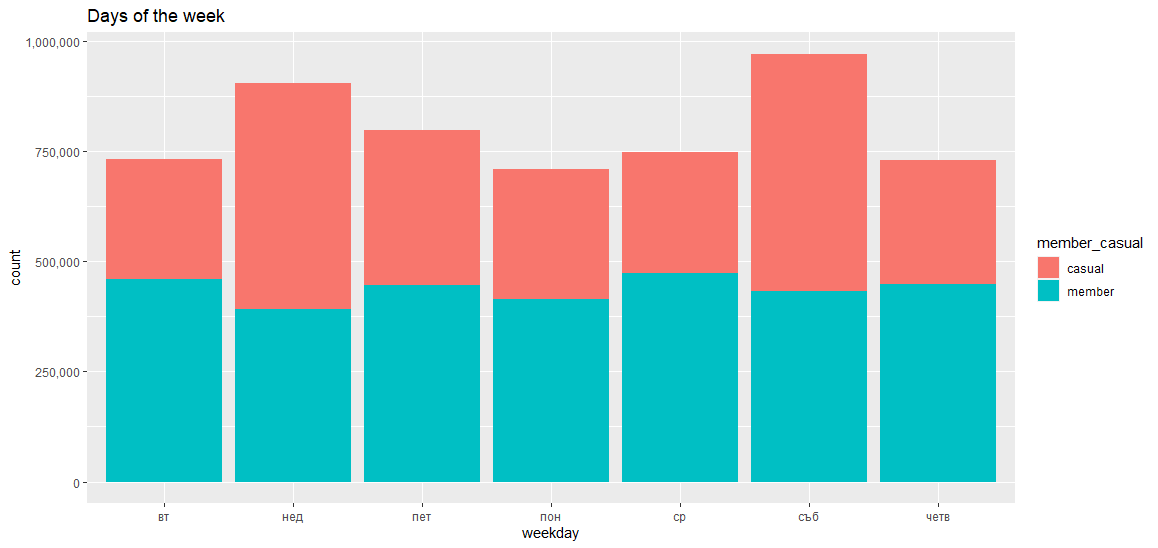
df %>%   
 group\_by(year\_month) %>%  
 summarize(count=length(ride\_id),  
 percentage\_of\_total=(length(ride\_id)/nrow(df))\*100,  
 members\_count=sum(member\_casual=="member"),  
 members\_percent=(sum(member\_casual=="member")/length(ride\_id))\*100,  
 casual\_count=sum(member\_casual=="casual"),  
 casual\_percent=(sum(member\_casual=="casual")/length(ride\_id))\*100) %>%   
 arrange(year\_month)  
  
ggplot(df, aes(year\_month, fill=member\_casual))+  
 geom\_bar()+  
 coord\_flip()+  
 scale\_y\_continuous(labels=comma)

 **Starting hours of the ride**

df %>%   
 group\_by(start\_hour) %>%  
 summarize(count=length(ride\_id),  
 percentage\_of\_total=(length(ride\_id)/nrow(df))\*100,  
 members\_count=sum(member\_casual=="member"),  
 members\_percent=(sum(member\_casual=="member")/length(ride\_id))\*100,  
 casual\_count=sum(member\_casual=="casual"),  
 casual\_percent=(sum(member\_casual=="casual")/length(ride\_id))\*100) %>%   
 arrange(start\_hour)  
  
ggplot(df, aes(start\_hour, fill=member\_casual))+  
 geom\_bar()+  
 facet\_wrap(~weekday)+  
 scale\_y\_continuous(labels=comma)

 **Days of the week**

df %>%   
 group\_by(weekday) %>%  
 summarize(count=length(ride\_id),  
 percentage\_of\_total=(length(ride\_id)/nrow(df))\*100,  
 members\_count=sum(member\_casual=="member"),  
 members\_percent=(sum(member\_casual=="member")/length(ride\_id))\*100,  
 casual\_count=sum(member\_casual=="casual"),  
 casual\_percent=(sum(member\_casual=="casual")/length(ride\_id))\*100)  
  
ggplot(df, aes(weekday, fill=member\_casual))+  
 geom\_bar()+  
 scale\_y\_continuous(labels=comma)



## Conclusion

### Observation

Annual member count in numbers are 3064777 or 54.8 percentage and casual members are 2525515 or 45.2 percentage of total members. During summer months July and August there is increase of the number of the casual users who rent a bike. Since September their number start to decrease. The lowest number of casual members is in December and January. The casual riders start to hire a bike in early morning 12:00 AM or at noon 2:00 PM. Annual members hire a bike in the morning between 8:00 AM or 10:00 AM and afternoon hours from 4:00 PM to 6:00 PM.During the weekend casual members increase in numbers 56 percentage compare to 43 percentage for annual members (Sunday). The same tendency can be seen for Saturday.

#### **Suggestions**

* Offer an annual membership plan for different part of the day (morning, afternoon, evening)
* Provide discounts for annual members base on the length ride and weekends.