**Cyclistic Bike Share Case Study**

**This project will be following the method: ASK-PREPARE-PROCESS-ANALYZE-SHARE-ACT.**

**ASK**

**Business Task:**

**Cyclistic, a bike-share company in Chicago, features more than 5,800 bicycles and 600 docking stations. The director of marketing, Lily Moreno, believes the future success of the company depends on maximizing the number of annual members. Therefore, as a member of the marketing analytics team, I will be working on the project to understand how the casual riders and member riders differ, and help design a marketing strategy to convert casual riders to annual members. The marketing analytics team will be providing recommendations to the stakeholders, the director of marketing and the executive team,with compelling data insights and data visualizations.**

**PREPARE**

**Since Cyclistic is a fictional bike-share company, the data that will be used in this case study comes from the City of Chicago's Divvy bicycle sharing service and it has been made available by Motivate International Inc. under a license. The data is public and well organized with no bias, and it's credible and ROCCC.The data contains no personal identifiable information, and it's integrated with the information that is needed for the case study. However, there are some values missing or incomplete, and they will be processed during data cleaning.**

**PROCESS**

In [1]:

```
# load packages needed
library(tidyverse) #helps wrangle data
library(lubridate) #helps wrangle date attributes
library(ggplot2)   #helps visualize data
```

In [2]:

```
# upload datasets (csv files) here
Jan_2021 <- read.csv('../input/d/haocstks/cyclistic/202101-divvy-tripdata.csv')
Feb_2021 <- read.csv('../input/d/haocstks/cyclistic/202102-divvy-tripdata.csv')
Mar_2021 <- read.csv('../input/d/haocstks/cyclistic/202103-divvy-tripdata.csv')
Apr_2021 <- read.csv('../input/d/haocstks/cyclistic/202104-divvy-tripdata.csv')
May_2021 <- read.csv('../input/d/haocstks/cyclistic/202105-divvy-tripdata.csv')
Jun_2021 <- read.csv('../input/d/haocstks/cyclistic/202106-divvy-tripdata.csv')
Jul_2021 <- read.csv('../input/d/haocstks/cyclistic/202107-divvy-tripdata.csv')
Aug_2021 <- read.csv('../input/d/haocstks/cyclistic/202108-divvy-tripdata.csv')
Sep_2021 <- read.csv('../input/d/haocstks/cyclistic/202109-divvy-tripdata.csv')
Oct_2021 <- read.csv('../input/d/haocstks/cyclistic/202110-divvy-tripdata.csv')
Nov_2021 <- read.csv('../input/d/haocstks/cyclistic/202111-divvy-tripdata.csv')
Dec_2021 <- read.csv('../input/d/haocstks/cyclistic/202112-divvy-tripdata.csv')
```

In [3]:

```
#compare column names each of the files. The names don't have to be in the exactly same o
rder, but they need to match before we can use a commend to combine them into one file.
colnames(Jan_2021)
colnames(Feb_2021)
colnames(Mar_2021)
colnames(Apr_2021)
colnames(May_2021)
colnames(Jun_2021)
colnames(Jul_2021)
colnames(Aug_2021)
colnames(Sep_2021)
colnames(Oct_2021)
colnames(Nov_2021)
colnames(Dec_2021)
```

In [4]:

```
#inspect the dataframes and look for incongruencies

str(Jan_2021)
str(Feb_2021)
str(Mar_2021)
str(Apr_2021)
str(May_2021)
str(Jun_2021)
str(Jul_2021)
str(Aug_2021)
str(Sep_2021)
str(Oct_2021)
str(Nov_2021)
str(Dec_2021)
```

In [5]:

```
# stack individual month's data frames into one big data frame
all_trips <- bind_rows(Jan_2021,Feb_2021,Mar_2021,Apr_2021,May_2021,Jun_2021,Jul_2021,Aug
_2021,Sep_2021,Oct_2021,Nov_2021,Dec_2021)
```

**Clean up data and add date**

In [6]:

```
# remove lat,long
all_trips <- all_trips %>%
 select(-c(start_lat,start_lng,end_lat,end_lng))
```

In [7]:

```
# inspect the new table that has been created
colnames(all_trips) # List of column names
nrow(all_trips) # How many rows in the data frame
dim(all_trips) # Dimensions of the data frame
head(all_trips) # See the first 6 rows of the data frame
str(all_trips) # See list of columns and data types (numeric,character etc.)
summary(all_trips) # Statistical summary of data
```

In [8]:

```
# See how many observations fall under each usertype
table(all_trips$member_casual)
```

In [9]:

```
# add colunms that list the date,month,day,and year of each ride
# this will allow us to aggregate ride data for each month,day,or year...
all_trips$date <- as.Date(all_trips$started_at) # the default format is yyyy-mm-dd
all_trips$month <- format(as.Date(all_trips$date),"%m")
all_trips$day <- format(as.Date(all_trips$date),"%d")
all_trips$year <- format(as.Date(all_trips$date),"%Y")
all_trips$day_of_week <- format(as.Date(all_trips$date),"%A")
```

In [10]:

```
# calculate ride length in minutes
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at,units="mins")
```

In [11]:

```
# inspect the structure of the columns
str(all_trips)
```

In [12]:

```
# convert "ride length" from factor to numeric so we can do calculations on the data
```

```
is.factor(all_trips$ride_length)
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

In [13]:

```
# romove ride length is zero or less than zero
# romove trips that have no start or end station names
all_trips_v2 <- all_trips[!(all_trips$ride_length <= 0|all_trips$start_station_name == "
"|all_trips$end_station_name == ""),]
```

**ANALYZE**

In [14]:

```
# create summary data frame (all figures in minutes)
mean(all_trips_v2$ride_length)
median(all_trips_v2$ride_length)
max(all_trips_v2$ride_length)
min(all_trips_v2$ride_length)
```

In [15]:

```
#compare ride length between members and casual riders
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual,FUN=mean)
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual,FUN=median)
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual,FUN=max)
aggregate(all_trips_v2$ride_length~all_trips_v2$member_casual,FUN=min)
```

In [17]:

```
# put day of week in order
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week,levels=c("Sunday","Monday",
"Tuesday","Wednesday","Thursday","Friday","Saturday"))

# average ride time by each day for members vs casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_wee
k,FUN=mean)
```

In [18]:

```
# put month in order
all_trips_v2$month <- ordered(all_trips_v2$month,levels = c("01","02","03","04","05","06
","07","08","09","10","11","12"))

# average ride length by month between members and casual riders
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$month, FUN
=mean)
```

In [19]:

```
# number of rides for each month between members and casual riders
all_trips_v2 %>%
  group_by(member_casual,month) %>%
  summarise(number_of_rides=n(),.groups='drop') %>%
  arrange(month)
```

In [20]:

```
# analyze ridership data by types and weekdays
all_trips_v2 %>%
 mutate(weekday=wday(started_at,label=TRUE)) %>% # create weekday field by using wday()
 group_by(member_casual,weekday) %>% # group by user type and weekday
 summarise(number_of_rides=n(),average_duration=mean(ride_length),.groups='drop') %>% #
calculate number of rides and average duration
 arrange(member_casual,weekday)
```

In [21]:

```
# Top 20 start stations for casual riders
```

```
# top 20 start stations for casual riders
all_trips_v2 %>%
  group_by(start_station_name, member_casual) %>%
  summarise(number_of_rides=n(),.groups='drop') %>%
  filter(start_station_name != "",member_casual != "member") %>%
  arrange(desc(number_of_rides)) %>%
  head(n=20)
```

In [22]:

```
# bike types preference between members and casual riders
all_trips_v2 %>%
  group_by(rideable_type,member_casual) %>%
  summarise(number_of_rides = n(),.groups='drop')
```

**SHARE Creating data visualizations**

In [23]:

```
# visualize the number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length),.groups='drop') %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x=weekday,y=number_of_rides,fill=member_casual))+geom_col(position="dodge")
```

In [24]:

```
# create a visualization by average duration
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length),.groups='drop') %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

In [25]:

```
# visualize number of rides for each month between members and casual riders
all_trips_v2 %>%
  mutate(month = month(started_at, label = TRUE)) %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length),.groups='drop') %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

In [26]:

```
# average ride duration between members and casual riders in general version

all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(average_duration = mean(ride_length)) %>%
  ggplot(aes(x = member_casual, y = average_duration)) +
  geom_col(position = "dodge")
```

In [27]:

```
str(all_trips_v2)
all_trips_v2$started_at_hour <- as.POSIXct(all_trips_v2$started_at, "%Y-%m-%d %H:%M:%S")

str(all_trips_v2)
```

In [28]:

```
# average number of rides for by hour between members and casual riders

all_trips_v2 %>%
  group_by(hour_of_day = hour(round_date(started_at_hour, 'hour'))) %>%
  group_by(hour_of_day, member_casual) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  arrange(-number_of_rides) %>%

  ggplot(aes(x = hour_of_day, y = number_of_rides, fill = member_casual)) +
  geom_col(position = 'dodge')
```

In [29]:

```
# T10 start stations for casual riders

all_trips_v2 %>%
  group_by(start_station_name, member_casual) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  filter(start_station_name != "", member_casual != "member") %>%
  arrange(-number_of_rides) %>%
  head(n=10) %>%

  ggplot(aes(x = reorder(start_station_name, number_of_rides), y = number_of_rides)) +
  geom_col(position = 'dodge', fill = 'lightpink') +
  scale_y_continuous(labels = scales::comma) +
  labs(title = 'Top 10 Start Stations for Casual Riders', x = '', y = "Number of Rides")
+
  coord_flip() +
  theme_minimal()
```

In [30]:

```
# usage of different bike types between members and casual riders

all_trips_v2 %>%
  group_by(rideable_type, member_casual) %>%
  summarize(number_of_rides = n(), .groups = 'drop') %>%
  drop_na() %>%

  ggplot(aes(x = member_casual, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = scales::comma) +
  facet_wrap(~rideable_type) +
  labs(fill = "Member/Casual", x = "", y = "Number of Rides",
       title = "Usage of Different Bikes: Members vs. Casual Riders")
```

**Key Findings:**

1. **In general, casual riders have longer average riding duration than members.**
2. **During the day of week from Monday to Sunday, casual riders' riding length steadily increases and reaches a peak on the weekends. While members remain almost the same amount of time in riding, the riding duration goes up slightly on the weekends too.**
3. **Also, it shows that the average numbers of rides for casual riders remain about the same level from Monday to Thursday, and then significantly increase from Friday, and reach a peak on Saturday and come down a bit on Sunday. However, the average numbers of rides for members remain quite the same throughout the week, and they reach a little peak on Wednesday and go down a little bit towards the end of week.**
4. **Analysis has also shown that there is a seasonal trend for both members and casual riders, that is, both of them prefer to use bikes more in the summertime. For casual riders, you can see that the average number of rides jumps to the top in the summertime and reaches a peak in July, and decreases over fall time and then a big jump off towards the winter time. For the members, the numbers are less dramatic, but still there is a little peak in August.**
5. **During the day, both members and casual riders have more usage of bikes around 5 or 6 pm in the afternoon than any other time during a day. But what needs to be pointed out is that members use bikes more often than casual riders in the morning around 7,8 or 9 am, while casual riders seem to use bikes a little more often around midnight or after.**
6. **The top 3 most popular stations and start stations with casual riders are Streeter Dr & Grand Ave, Millennium Park and Michigan Ave & Oak St.**

7. **The most popular bike type for both members and casual riders is classic bikes, and the least popular one is docked bikes.**

**ACT**

**Recommendations:**

1. **Based on the analysis, casual riders seem to have longer average riding duration, so Cyclistic can limit the riding length by charging casual riders extra fee if they ride over 20 minutes or do not return the bikes after full-day passes expire. Or cyclistic can inform casual riders about the benefit of getting membership if they need to ride over 20 minutes or can't return bikes at the end of the day.**
2. **In the summertime or nice weekends, Cyclistic can offer new riders or existing casual riders a discount on getting new membership or benefit of membership at local businesses or touring spots.**
3. **Cyclistic can also launch marketing campaigns or digital media marketing around the top 10 most popular stations, especially the start stations by casual riders, focusing on sharing the information of the benefit of membership.**