Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Characters and teams

● Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

● Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

● Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.

● Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

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.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Moreno has assigned you the first question to answer: How do annual members and casual riders use Cyclistic bikes differently? You will produce a report with the following deliverables:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top three recommendations based on your analysis

Analytical Process

Step 1: Defining the problem

For this particular case, I am asked to solve the first of the three questions: How do annual members and casual members differ?

A good first step to answer this question is to get insights about what days of the week both types of cyclist prefer. Another good insight is to get an average about the distance traveled, either in distance (KMs) or in time (HH:MM:SS). This can be accompanied with a map of the town or city, to analyze behavior of both casual and member riders

These data insights can help the Executive Team get the similarities and differences between both groups of riders, and take a decision to pull more casual riders into annual membership.

Step 2: Data Source and Tools

For the purposes of this case study, the data set was taken from the public data provided by Motivate International Inc.
License: Data License Agreement | Divvy Bikes

Considering the big amount of data to handle, I considered best to put it into SQL and analyze it with either Python or R. I like Python more just because I have more experience with it.

I took the data and stored it into a local MySQL Server, which I created from zero.

```
CREATE TABLE cyclistic (

ride_id VARCHAR(50) PRIMARY KEY NOT NULL,

rideable_type VARCHAR(50),

started_at DATETIME,

ended_at DATETIME,

start_station_name VARCHAR(200),

start_station_id VARCHAR(50),

end_station_name VARCHAR(200),

end_station_id VARCHAR(50),

start_lat FLOAT,
```

```
start_lng FLOAT,

end_lat FLOAT,

end_lng FLOAT,

member_casual VARCHAR(50)

);
```

The Table created is based on the elements found in the csv archives. Then proceeded to import the CSV archives

```
LOAD DATA INFILE 'PATH/202204-divvy-tripdata.csv'

INTO TABLE cyclistic

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 ROWS

(ride_id,rideable_type,started_at,ended_at,start_station_name,start_station_id,end_station_name,end_station_id,@start_lat,@start_lng,@end_lat,@end_lng,member_casual)

SET

start_lat = NULLIF(@start_lat,''),

start_lng = NULLIF(@start_lng,''),

end_lat = NULLIF(@end_lat,''),

end_lng = NULLIF(@end_lng,'')
```

The dataset was incomplete in some rows. In particular, ‘LOAD DATA INFILE’ command fails if it encounters an empty space on FLOAT type variables. I could have cleaned data here before uploading, and it would have been the right way to proceed; but to exercise data cleaning in Python I decided to upload incomplete data into MySQL, and filter it through coding. So I use the ‘SET’ command to set null values in cases where the data was missing.

Step 3: Cleaning Data

As stated in Step 2, I decided to upload unclean data and clean it with the Python code.
In order to analyze the variables I proposed, I require the Date/Time value.

For this particular condition, I simply apply a filtered query through Python

```
SELECT *
FROM Cyclistic
WHERE started_at IS NOT NULL
```

I decided to take all columns and store them into a DataFrame, so I am able to work freely with it. I did not clean data in the DB, and only made special query with it Since I did not clean the data before uploading, there is no documentation about a cleaning process

Step 4: Analysis

I handle the analysis process with a pandas DataFrame. I first obtain the average of days. The DB has the DateTime format already. So I can use the function weekday from the datetime library

```
df["weekday"] = df["started_at"].dt.weekday
```

This is stored in a new column in the DataFrame called ‘weekday’. This function gives the days in number from 0 to 6, being 0, Sunday and 6, Saturday.

Next I create two different sub DataFrames, one for the casual members and one for the annual members. This will allow get the metrics and compare the groups

```
dfcasual = df[df.member_casual == "casual\r"]
dfmember = df[df.member_casual == "member\r"]
```

Now I want to get the AVG per day, and per member. I did this with a loop with a range option between 0 and 6. And the DataFrame will be filtered according to these ranges, which represent the days, and I calculate the arithmetic mean of each filtered DataFrame. I also sum the amount of times that these days show, to get which days are more often.

```
for i in range(0,7):
    dfcasualrange = dfcasual[dfcasual.weekday == i]
    internalvar1 = dfcasualrange["ride_length"].mean()
    internalvar2 = internalvar1.total_seconds()
    AVG_1.append(internalvar2)
    Day_1.append(dfcasualrange.duplicated(["weekday"]).sum())
    dfmemberrange = dfmember[dfmember.weekday == i]
    internalvar3 = dfmemberrange["ride_length"].mean()
    internalvar4 = internalvar3.total_seconds()
    AVG_2.append(internalvar4)
    Day_2.append(dfmemberrange.duplicated(["weekday"]).sum())
```

These four variables are set into lists. To be able to correlate the data of which day has which mean, and to difference between Casual and Member riders, I use dictionaries

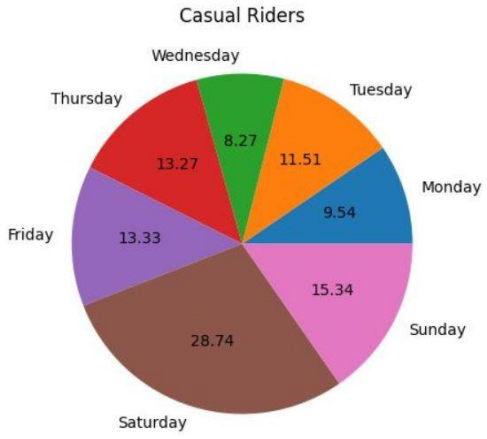
```
print("\n")
dic = {"average":AVG_1,"day":Day_1}
dfanalysis1 = pd.DataFrame(dic,columns=["average","day"])
print(dfanalysis1)

print("\n")
dic2 = {"average":AVG_2,"day":Day_2}
dfanalysis2 = pd.DataFrame(dic2,columns=["average","day"])
print(dfanalysis2)

print("\n")
dic3 = {"Casual Riders":AVG_1,"Member Riders":AVG_2}
dfanalysis3 = pd.DataFrame(dic3,columns=["Casual Riders","Member Riders"])
print(dfanalysis3)
```

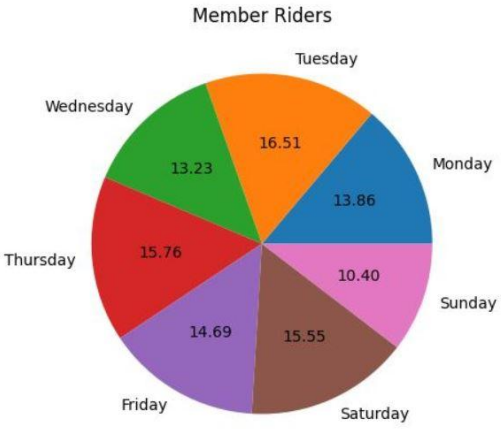
Results:

The first chart shows, on the first column, the average, per day, in total seconds, between Casual and Member riders. And in the second column a total sum of each day.

	average	day
0	1747.402801	12062
1	1579.105154	14549
2	1280.514201	10456
3	1552.069849	16778
4	1567.885756	16849
5	2085.278034	36329
6	1977.491592	19387

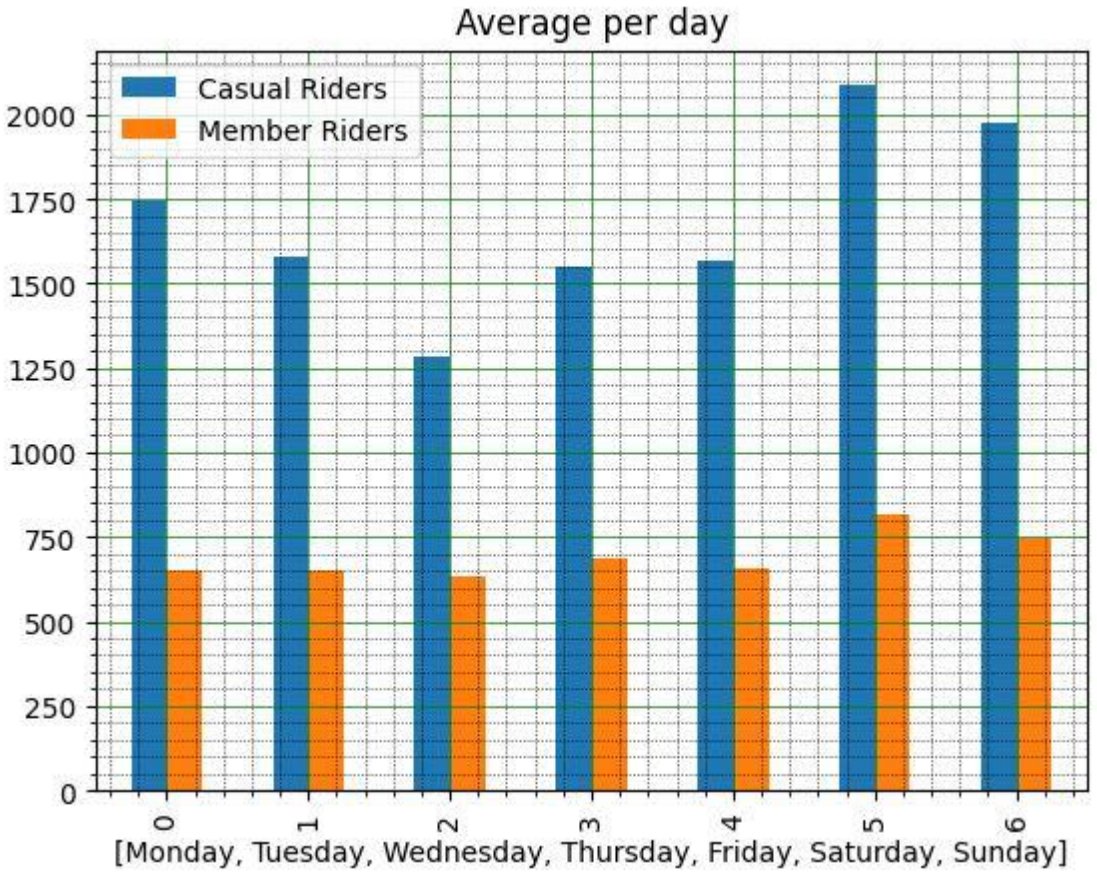
	average	day
0	648.395891	33930
1	649.915413	40431
2	632.219285	32386
3	683.779453	38594
4	658.073770	35962
5	815.279980	38066
6	745.440232	25456





The second chart shows the difference, in total seconds, between Casual and Member Riders

	Casual Riders	Member Riders
0	1747.402801	648.395891
1	1579.105154	649.915413
2	1280.514201	632.219285
3	1552.069849	683.779453
4	1567.885756	658.073770
5	2085.278034	815.279980
6	1977.491592	745.440232



This data provides an interesting design. Casual riders tend to ride for a longer time, and are a lot more frequent on Weekends (Friday, Saturday, and Sunday), whereas Member Riders ride for a shorter time, but they do so equally all days.

Step 5: Conclusions and next steps

My theory is that Casual riders use it for longer travels, usually as a recreation activity, in their free time, while Annual Members use it for day-to-day activities.

To further prove or disprove this theory, a good next step would be to draw a heat map with either the Start/End stations or Lat/long. If the theory proves right, Annual Members' heat map would show more intensity between city streets, while Casual riders' heat map would show more intensity outside the city.

If this analysis is correct, then a good marketing approach would be to inquiry about Casual Member locations and conducts. Why don't they choose riding for day-to-day activities? Maybe it is because there is not a station near their workplace, or near their homes.

Another option to look at, is maybe not about the location of the stations, but disponibility. If there are a lot of Casual and Members in a place, it can happen that a station runs out of available bikes at peak office hours.This would discourage people from subscribing.