# Case Study - Cyclist New Marketing Strategy

## Libraries & Data

In [1]:
```
import pandas as pd
```

In [2]:
```
data_19_q1 = pd.read_csv('../data/Divvy_Trips_2019_Q1.csv')
```

## Data Cleaning

***Performing data cleaning on first dataset to standardize a format***

> **Dataset - 2019 Q1**

In [3]:
```
data_19_q1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 365069 entries, 0 to 365068
Data columns (total 12 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   trip_id            365069 non-null  int64  
 1   start_time         365069 non-null  object 
 2   end_time           365069 non-null  object 
 3   bikeid             365069 non-null  int64  
 4   tripduration       365069 non-null  object 
 5   from_station_id    365069 non-null  int64  
 6   from_station_name  365069 non-null  object 
 7   to_station_id      365069 non-null  int64  
 8   to_station_name    365069 non-null  object 
 9   usertype           365069 non-null  object 
 10  gender             345358 non-null  object 
 11  birthyear          347046 non-null  float64
dtypes: float64(1), int64(4), object(7)
memory usage: 33.4+ MB
```

In [4]:
```
data_19_q1.head(3)
```

Out[4]:

| | trip_id | start_time | end_time | bikeid | tripduration | from_station_id | from_station_name | to_static |
|---|---|---|---|---|---|---|---|---|
| **0** | 21742443 | 2019-01-01 00:04:37 | 2019-01-01 00:11:07 | 2167 | 390.0 | 199 | Wabash Ave & Grand Ave | |
| **1** | 21742444 | 2019-01-01 00:08:13 | 2019-01-01 00:15:34 | 4386 | 441.0 | 44 | State St & Randolph St | |
| **2** | 21742445 | 2019-01-01 00:13:23 | 2019-01-01 00:27:12 | 1524 | 829.0 | 15 | Racine Ave & 18th St | |

**1. Removing null values**

In [5]:
```
data_19_q1 = data_19_q1[~data_19_q1['gender'].isnull() & ~data_19_q1['birthyear'].isnull()]
```

**2. Trip Duration --> Converting to integer**

In [6]:
```
data_19_q1['tripduration'] = data_19_q1['tripduration'].apply(lambda x: float(x.replace(',',
data_19_q1['tripduration'] = data_19_q1['tripduration'].apply(lambda x: int(x))
```

**3. Birth Year --> Converting to integer & getting age (as in 2020)**

In [7]:
```
data_19_q1['birthyear'] = data_19_q1['birthyear'].apply(lambda x: int(x))
data_19_q1['age'] = data_19_q1['birthyear'].apply(lambda x: 2020-x)
```

**4. Setting user type --> Member/ Casual**

In [8]:
```
data_19_q1['usertype'].value_counts()
```

Out[8]:
```
usertype
Subscriber    339423
Customer        5934
Name: count, dtype: int64
```

In [9]:
```
data_19_q1['usertype'] = data_19_q1['usertype'].replace({'Subscriber':'Member', 'Customer':'
```

**5. Start/ End time --> datetime**

In [10]:
```
data_19_q1['start_time'] = pd.to_datetime(data_19_q1['start_time'])
data_19_q1['end_time'] = pd.to_datetime(data_19_q1['end_time'])
```

***Creating columns with standard names***

In [11]:
```
data_19_q1['start_id'] = data_19_q1['from_station_id']
data_19_q1['end_id'] = data_19_q1['to_station_id']

data_19_q1['start_name'] = data_19_q1['from_station_name']
data_19_q1['end_name'] = data_19_q1['to_station_name']

data_19_q1['bike_id'] = data_19_q1['bikeid']
data_19_q1['trip_duration'] = data_19_q1['tripduration']

data_19_q1['user_type'] = data_19_q1['usertype']
```

***Dropping older columns***

In [12]:
```
data_19_q1 = data_19_q1.drop(['trip_id', 'bikeid', 'tripduration', 'from_station_id', 'to_st
```

***New Column order***

In [13]:
```
column_order = ['start_time', 'end_time', 'trip_duration', 'start_id', 'end_id', 'start_name
data_19_q1 = data_19_q1[column_order]
```

In [14]:
```
data_19_q1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 345357 entries, 0 to 365068
Data columns (total 11 columns):
 #   Column         Non-Null Count   Dtype         
---  ------         --------------   -----         
 0   start_time     345357 non-null  datetime64[ns]
 1   end_time       345357 non-null  datetime64[ns]
 2   trip_duration  345357 non-null  int64         
 3   start_id       345357 non-null  int64         
 4   end_id         345357 non-null  int64         
 5   start_name     345357 non-null  object        
 6   end_name       345357 non-null  object        
 7   bike_id        345357 non-null  int64         
 8   gender         345357 non-null  object        
 9   age            345357 non-null  int64         
 10  user_type      345357 non-null  object        
dtypes: datetime64[ns](2), int64(5), object(4)
memory usage: 31.6+ MB
```

In [15]:
```
data_19_q1.head(3)
```

Out[15]:

| | start_time | end_time | trip_duration | start_id | end_id | start_name | end_name | bike_id | gender | age |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2019-01-01 00:04:37 | 2019-01-01 00:11:07 | 390 | 199 | 84 | Wabash Ave & Grand Ave | Milwaukee Ave & Grand Ave | 2167 | Male | 31 |
| **1** | 2019-01-01 00:08:13 | 2019-01-01 00:15:34 | 441 | 44 | 624 | State St & Randolph St | Dearborn St & Van Buren St (*) | 4386 | Female | 30 |
| **2** | 2019-01-01 00:13:23 | 2019-01-01 00:27:12 | 829 | 15 | 644 | Racine Ave & 18th St | Western Ave & Fillmore St (*) | 1524 | Female | 26 |

**Standardizing for other datasets**

> **Dataset - 2019 Q2**

In [16]:
```
data_19_q2 = pd.read_csv('../data/Divvy_Trips_2019_Q2.csv')
```

In [17]:
```
data_19_q2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1108163 entries, 0 to 1108162
Data columns (total 12 columns):
 #   Column                                          Non-Null Count    Dtype
---  ------                                          --------------    -----
 0   01 - Rental Details Rental ID                   1108163 non-null  int64
 1   01 - Rental Details Local Start Time            1108163 non-null  object
 2   01 - Rental Details Local End Time              1108163 non-null  object
 3   01 - Rental Details Bike ID                     1108163 non-null  int64
 4   01 - Rental Details Duration In Seconds Uncapped  1108163 non-null  object
 5   03 - Rental Start Station ID                    1108163 non-null  int64
 6   03 - Rental Start Station Name                  1108163 non-null  object
 7   02 - Rental End Station ID                      1108163 non-null  int64
 8   02 - Rental End Station Name                    1108163 non-null  object
 9   User Type                                       1108163 non-null  object
 10  Member Gender                                   922609 non-null   object
 11  05 - Member Details Member Birthday Year        927210 non-null   float64
dtypes: float64(1), int64(4), object(7)
memory usage: 101.5+ MB
```

In [18]:
```
data_19_q2.head(3)
```

Out[18]:

| | 01 - Rental Details Rental ID | 01 - Rental Details Local Start Time | 01 - Rental Details Local End Time | 01 - Rental Details Bike ID | 01 - Rental Details Duration In Seconds Uncapped | 03 - Rental Start Station ID | 03 - Rental Start Station Name | 02 - Rental End Station ID | 02 - Rental End Station Name | User Type |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 22178529 | 2019-04-01 00:02:22 | 2019-04-01 00:09:48 | 6251 | 446.0 | 81 | Daley Center Plaza | 56 | Desplaines St & Kinzie St | Subscriber |
| **1** | 22178530 | 2019-04-01 00:03:02 | 2019-04-01 00:20:30 | 6226 | 1,048.0 | 317 | Wood St & Taylor St | 59 | Wabash Ave & Roosevelt Rd | Subscriber |
| **2** | 22178531 | 2019-04-01 00:11:07 | 2019-04-01 00:15:19 | 5649 | 252.0 | 283 | LaSalle St & Jackson Blvd | 174 | Canal St & Madison St | Subscriber |

In [19]:
```
# Removing Null Values
data_19_q2 = data_19_q2[~data_19_q2['Member Gender'].isnull() & ~data_19_q2['05 - Member Det

# Trip duration --> interger
data_19_q2['01 - Rental Details Duration In Seconds Uncapped'] = data_19_q2['01 - Rental Det
data_19_q2['01 - Rental Details Duration In Seconds Uncapped'] = data_19_q2['01 - Rental Det

# User Type --> Member/ Casual
data_19_q2['User Type'] = data_19_q2['User Type'].replace({'Subscriber': 'Member', 'Customer

# Age (as in 2020) --> from birthyear
data_19_q2['05 - Member Details Member Birthday Year'] = data_19_q2['05 - Member Details Mem
data_19_q2['age'] = data_19_q2['05 - Member Details Member Birthday Year'].apply(lambda x: 2

# Obj --> datetime
data_19_q2['01 - Rental Details Local Start Time'] = pd.to_datetime(data_19_q2['01 - Rental
data_19_q2['01 - Rental Details Local End Time'] = pd.to_datetime(data_19_q2['01 - Rental De

# Renaming
```

```
data_19_q2['start_time'] = data_19_q2['01 - Rental Details Local Start Time']
data_19_q2['end_time'] = data_19_q2['01 - Rental Details Local End Time']

data_19_q2['start_id'] = data_19_q2['03 - Rental Start Station ID']
data_19_q2['end_id'] = data_19_q2['02 - Rental End Station ID']

data_19_q2['start_name'] = data_19_q2['03 - Rental Start Station Name']
data_19_q2['end_name'] = data_19_q2['02 - Rental End Station Name']

data_19_q2['trip_duration'] = data_19_q2['01 - Rental Details Duration In Seconds Uncapped']
data_19_q2['bike_id'] = data_19_q2['01 - Rental Details Bike ID']

data_19_q2['gender'] = data_19_q2['Member Gender']
data_19_q2['user_type'] = data_19_q2['User Type']

# Column Order
data_19_q2 = data_19_q2[column_order]
```

In [20]:
```
data_19_q2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 922608 entries, 0 to 1108162
Data columns (total 11 columns):
 #   Column         Non-Null Count   Dtype         
---  ------         --------------   -----         
 0   start_time     922608 non-null  datetime64[ns]
 1   end_time       922608 non-null  datetime64[ns]
 2   trip_duration  922608 non-null  int64         
 3   start_id       922608 non-null  int64         
 4   end_id         922608 non-null  int64         
 5   start_name     922608 non-null  object        
 6   end_name       922608 non-null  object        
 7   bike_id        922608 non-null  int64         
 8   gender         922608 non-null  object        
 9   age            922608 non-null  int64         
 10  user_type      922608 non-null  object        
dtypes: datetime64[ns](2), int64(5), object(4)
memory usage: 84.5+ MB
```

In [21]:
```
data_19_q2.head(3)
```

Out[21]:

| | start_time | end_time | trip_duration | start_id | end_id | start_name | end_name | bike_id | gender | age |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2019-04-01 00:02:22 | 2019-04-01 00:09:48 | 446 | 81 | 56 | Daley Center Plaza | Desplaines St & Kinzie St | 6251 | Male | 45 |
| **1** | 2019-04-01 00:03:02 | 2019-04-01 00:20:30 | 1048 | 317 | 59 | Wood St & Taylor St | Wabash Ave & Roosevelt Rd | 6226 | Female | 36 |
| **2** | 2019-04-01 00:11:07 | 2019-04-01 00:15:19 | 252 | 283 | 174 | LaSalle St & Jackson Blvd | Canal St & Madison St | 5649 | Male | 30 |

> **Dataset - 2019 Q3**

In [22]:
```
data_19_q3 = pd.read_csv('../data/Divvy_Trips_2019_Q3.csv')
```

In [23]:
```
data_19_q3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1640718 entries, 0 to 1640717
Data columns (total 12 columns):
 #   Column             Non-Null Count    Dtype  
---  ------             --------------    -----  
 0   trip_id            1640718 non-null  int64  
 1   start_time         1640718 non-null  object 
 2   end_time           1640718 non-null  object 
 3   bikeid             1640718 non-null  int64  
 4   tripduration       1640718 non-null  object 
 5   from_station_id    1640718 non-null  int64  
 6   from_station_name  1640718 non-null  object 
 7   to_station_id      1640718 non-null  int64  
 8   to_station_name    1640718 non-null  object 
 9   usertype           1640718 non-null  object 
 10  gender             1353368 non-null  object 
 11  birthyear          1362624 non-null  float64
dtypes: float64(1), int64(4), object(7)
memory usage: 150.2+ MB
```

In [24]:
```
data_19_q3.head(3)
```

Out[24]:

| | trip_id | start_time | end_time | bikeid | tripduration | from_station_id | from_station_name | to_static |
|---|---|---|---|---|---|---|---|---|
| **0** | 23479388 | 2019-07-01 00:00:27 | 2019-07-01 00:20:41 | 3591 | 1,214.0 | 117 | Wilton Ave & Belmont Ave | |
| **1** | 23479389 | 2019-07-01 00:01:16 | 2019-07-01 00:18:44 | 5353 | 1,048.0 | 381 | Western Ave & Monroe St | |
| **2** | 23479390 | 2019-07-01 00:01:48 | 2019-07-01 00:27:42 | 6180 | 1,554.0 | 313 | Lakeview Ave & Fullerton Pkwy | |

In [25]:
```
# Removing Null Values
data_19_q3 = data_19_q3[~data_19_q3['gender'].isnull() & ~data_19_q3['birthyear'].isnull()]

# Trip duration --> interger
data_19_q3['tripduration'] = data_19_q3['tripduration'].apply(lambda x: float(x.replace(',',
data_19_q3['tripduration'] = data_19_q3['tripduration'].apply(lambda x: int(x))

# User Type --> Member/ Casual
data_19_q3['usertype'] = data_19_q3['usertype'].replace({'Subscriber': 'Member', 'Customer':

# Age (as in 2020) --> from birthyear
data_19_q3['birthyear'] = data_19_q3['birthyear'].apply(lambda x: int(x))
data_19_q3['age'] = data_19_q3['birthyear'].apply(lambda x: 2020-x)

# Obj --> datetime
data_19_q3['start_time'] = pd.to_datetime(data_19_q3['start_time'])
data_19_q3['end_time'] = pd.to_datetime(data_19_q3['end_time'])

# Renaming
data_19_q3['start_id'] = data_19_q3['from_station_id']
data_19_q3['end_id'] = data_19_q3['to_station_id']
```

```
data_19_q3['start_name'] = data_19_q3['from_station_name']
data_19_q3['end_name'] = data_19_q3['to_station_name']

data_19_q3['bike_id'] = data_19_q3['bikeid']
data_19_q3['trip_duration'] = data_19_q3['tripduration']

data_19_q3['user_type'] = data_19_q3['usertype']

# Column Order
data_19_q3 = data_19_q3[column_order]
```

In [26]:
```
data_19_q3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1353368 entries, 0 to 1640717
Data columns (total 11 columns):
 #   Column         Non-Null Count    Dtype         
---  ------         --------------    -----         
 0   start_time     1353368 non-null  datetime64[ns]
 1   end_time       1353368 non-null  datetime64[ns]
 2   trip_duration  1353368 non-null  int64         
 3   start_id       1353368 non-null  int64         
 4   end_id         1353368 non-null  int64         
 5   start_name     1353368 non-null  object        
 6   end_name       1353368 non-null  object        
 7   bike_id        1353368 non-null  int64         
 8   gender         1353368 non-null  object        
 9   age            1353368 non-null  int64         
 10  user_type      1353368 non-null  object        
dtypes: datetime64[ns](2), int64(5), object(4)
memory usage: 123.9+ MB
```

In [27]:
```
data_19_q3.head(3)
```

Out[27]:

| | start_time | end_time | trip_duration | start_id | end_id | start_name | end_name | bike_id | gender | ag |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2019-07-01 00:00:27 | 2019-07-01 00:20:41 | 1214 | 117 | 497 | Wilton Ave & Belmont Ave | Kimball Ave & Belmont Ave | 3591 | Male | 2 |
| **5** | 2019-07-01 00:02:21 | 2019-07-01 00:07:31 | 310 | 300 | 232 | Broadway & Barry Ave | Pine Grove Ave & Waveland Ave | 4941 | Male | 3 |
| **18** | 2019-07-01 00:06:51 | 2019-07-01 00:26:22 | 1171 | 624 | 237 | Dearborn St & Van Buren St | MLK Jr Dr & 29th St | 2758 | Male | 2 |

> **Dataset - 2019 Q4**

In [28]:
```
data_19_q4 = pd.read_csv('../data/Divvy_Trips_2019_Q4.csv')
```

In [29]:
```
data_19_q4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 704054 entries, 0 to 704053
Data columns (total 12 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   trip_id            704054 non-null  int64
 1   start_time         704054 non-null  object
 2   end_time           704054 non-null  object
 3   bikeid             704054 non-null  int64
 4   tripduration       704054 non-null  object
 5   from_station_id    704054 non-null  int64
 6   from_station_name  704054 non-null  object
 7   to_station_id      704054 non-null  int64
 8   to_station_name    704054 non-null  object
 9   usertype           704054 non-null  object
 10  gender             637463 non-null  object
 11  birthyear          642373 non-null  float64
dtypes: float64(1), int64(4), object(7)
memory usage: 64.5+ MB
```

In [30]:
```
data_19_q4.head(3)
```

Out[30]:

| | trip_id | start_time | end_time | bikeid | tripduration | from_station_id | from_station_name | to_static |
|---|---|---|---|---|---|---|---|---|
| **0** | 25223640 | 2019-10-01 00:01:39 | 2019-10-01 00:17:20 | 2215 | 940.0 | 20 | Sheffield Ave & Kingsbury St | |
| **1** | 25223641 | 2019-10-01 00:02:16 | 2019-10-01 00:06:34 | 6328 | 258.0 | 19 | Throop (Loomis) St & Taylor St | |
| **2** | 25223642 | 2019-10-01 00:04:32 | 2019-10-01 00:18:43 | 3003 | 850.0 | 84 | Milwaukee Ave & Grand Ave | |

In [31]:
```
# Removing Null Values
data_19_q4 = data_19_q4[~data_19_q4['gender'].isnull() & ~data_19_q4['birthyear'].isnull()]

# Trip duration --> interger
data_19_q4['tripduration'] = data_19_q4['tripduration'].apply(lambda x: float(x.replace(',',
data_19_q4['tripduration'] = data_19_q4['tripduration'].apply(lambda x: int(x))

# User Type --> Member/ Casual
data_19_q4['usertype'] = data_19_q4['usertype'].replace({'Subscriber': 'Member', 'Customer':

# Age (as in 2020) --> from birthyear
data_19_q4['birthyear'] = data_19_q4['birthyear'].apply(lambda x: int(x))
data_19_q4['age'] = data_19_q4['birthyear'].apply(lambda x: 2020-x)

# Obj --> datetime
data_19_q4['start_time'] = pd.to_datetime(data_19_q4['start_time'])
data_19_q4['end_time'] = pd.to_datetime(data_19_q4['end_time'])

# Renaming
data_19_q4['start_id'] = data_19_q4['from_station_id']
data_19_q4['end_id'] = data_19_q4['to_station_id']

data_19_q4['start_name'] = data_19_q4['from_station_name']
data_19_q4['end_name'] = data_19_q4['to_station_name']

data_19_q4['bike_id'] = data_19_q4['bikeid']
data_19_q4['trip_duration'] = data_19_q4['tripduration']
```

```
data_19_q4['user_type'] = data_19_q4['usertype']

# Column Order
data_19_q4 = data_19_q4[column_order]
```

In [32]:
```
data_19_q4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 637463 entries, 0 to 704053
Data columns (total 11 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   start_time     637463 non-null  datetime64[ns]
 1   end_time       637463 non-null  datetime64[ns]
 2   trip_duration  637463 non-null  int64
 3   start_id       637463 non-null  int64
 4   end_id         637463 non-null  int64
 5   start_name     637463 non-null  object
 6   end_name       637463 non-null  object
 7   bike_id        637463 non-null  int64
 8   gender         637463 non-null  object
 9   age            637463 non-null  int64
 10  user_type      637463 non-null  object
dtypes: datetime64[ns](2), int64(5), object(4)
memory usage: 58.4+ MB
```

In [33]:
```
data_19_q4.head(3)
```

Out[33]:

| | start_time | end_time | trip_duration | start_id | end_id | start_name | end_name | bike_id | gender | age |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2019-10-01 00:01:39 | 2019-10-01 00:17:20 | 940 | 20 | 309 | Sheffield Ave & Kingsbury St | Leavitt St & Armitage Ave | 2215 | Male | 33 |
| **1** | 2019-10-01 00:02:16 | 2019-10-01 00:06:34 | 258 | 19 | 241 | Throop (Loomis) St & Taylor St | Morgan St & Polk St | 6328 | Male | 22 |
| **2** | 2019-10-01 00:04:32 | 2019-10-01 00:18:43 | 850 | 84 | 199 | Milwaukee Ave & Grand Ave | Wabash Ave & Grand Ave | 3003 | Female | 29 |

In [34]:
```
data = pd.concat([data_19_q1, data_19_q2, data_19_q3, data_19_q4])
```

In [35]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 3258796 entries, 0 to 704053
Data columns (total 11 columns):
 #   Column         Dtype         
---  ------         -----         
 0   start_time     datetime64[ns]
 1   end_time       datetime64[ns]
 2   trip_duration  int64         
 3   start_id       int64         
 4   end_id         int64         
 5   start_name     object        
 6   end_name       object        
 7   bike_id        int64         
 8   gender         object        
 9   age            int64         
 10  user_type      object        
dtypes: datetime64[ns](2), int64(5), object(4)
memory usage: 298.4+ MB
```

In [37]:
```
data.to_csv('../data/data.csv', index=False)
```