Cyclistic

Harper Nguyen

5/25/2021

Case Study

Scenario I am 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, our team wants to understand how casual riders and annual members use Cyclistic bikes different. From these insights, our team will design a new marketing strategy to convert casual riders into annual members.

Phase 1: Ask

Asking 3 questions:

1. How do annual members and casual riders use Cyclistic bikes different?

2. Why would casual riders buy Cyclistic annual memberships?

3. How can Cyclistic use digital media to influence casual riders to become members?

I has been assigned the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

In particular, I will analyze some questions below:

1. What is total number of trips for members and casuals, which proportion of total trips they represent?
2. Some statistic metrics about ride length for members and casuals (min, max, mean, median)
3. Where are the most common area casuals start and end trips?
4. Which months, days of week, hours of day member and casual ride most?
5. What types of ride, casual and member use most?

Phase 2: Prepare

Where is my data located? - I use Cyclistic's historical trip data to identify how different casual and member use bikes. It's internal resource, the original , reliable source. I will use current trips data in 12 months from May 2020 to Apr 2021 which have all data I need to answer my questions above (start and end station name, start and end datetime, ridable type, member_casual).

First, download 12 ".csv" file. Using excel to filter and sort to check data mising in all columns. Found out that start_station_name and end_station_name have some blank cells. As the data is large, I will combine all file into 1 table and clean data in R.

Import required packages

```
library(rmarkdown)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyvers
e 1.3.1 ──

## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.2     ✓ dplyr   1.0.6
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse_con
flicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# install.packages("sqldf")
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Import data

```
setwd('/Users/harper/data/Data\ analytics/GOogle\ course/Case\ Study/csv/')
m05_2020 <- read.csv('202005-divvy-tripdata.csv')
m06_2020 <- read.csv('202006-divvy-tripdata.csv')
m07_2020 <- read.csv('202007-divvy-tripdata.csv')
m08_2020 <- read.csv('202008-divvy-tripdata.csv')
m09_2020 <- read.csv('202009-divvy-tripdata.csv')
m10_2020 <- read.csv('202010-divvy-tripdata.csv')
m11_2020 <- read.csv('202011-divvy-tripdata.csv')
m12_2020 <- read.csv('202012-divvy-tripdata.csv')
m01_2021 <- read.csv('202101-divvy-tripdata.csv')
m02_2021 <- read.csv('202102-divvy-tripdata.csv')
```

```
m03_2021 <- read.csv('202103-divvy-tripdata.csv')
m04_2021 <- read.csv('202104-divvy-tripdata.csv')
```

Checking the columns name and type of variables

```
glimpse(m05_2020)

## Rows: 200,274
## Columns: 13
## $ ride_id           <chr> "02668AD35674B983", "7A50CCAF1EDDB28F", "2FFCDFDB91…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike", "docke…
## $ started_at        <chr> "2020-05-27 10:03:52", "2020-05-25 10:47:11", "2020…
## $ ended_at          <chr> "2020-05-27 10:16:49", "2020-05-25 11:05:40", "2020…
## $ start_station_name <chr> "Franklin St & Jackson Blvd", "Clark St & Wrightwoo…
## $ start_station_id  <int> 36, 340, 260, 251, 261, 206, 261, 180, 331, 219, 24…
## $ end_station_name  <chr> "Wabash Ave & Grand Ave", "Clark St & Leland Ave", …
## $ end_station_id    <int> 199, 326, 260, 157, 206, 22, 261, 180, 300, 305, 14…
## $ start_lat         <dbl> 41.8777, 41.9295, 41.9296, 41.9680, 41.8715, 41.847…
## $ start_lng         <dbl> -87.6353, -87.6431, -87.7079, -87.6500, -87.6699, -…
## $ end_lat           <dbl> 41.8915, 41.9671, 41.9296, 41.9367, 41.8472, 41.869…
## $ end_lng           <dbl> -87.6268, -87.6674, -87.7079, -87.6368, -87.6468, -…
## $ member_casual     <chr> "member", "casual", "casual", "casual", "member", "…

glimpse(m06_2020)

## Rows: 343,005
## Columns: 13
## $ ride_id           <chr> "8CD5DE2C2B6C4CFC", "9A191EB2C751D85D", "F37D14B0B5…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike", "docke…
## $ started_at        <chr> "2020-06-13 23:24:48", "2020-06-26 07:26:10", "2020…
## $ ended_at          <chr> "2020-06-13 23:36:55", "2020-06-26 07:31:58", "2020…
## $ start_station_name <chr> "Wilton Ave & Belmont Ave", "Federal St & Polk St",…
## $ start_station_id  <int> 117, 41, 81, 303, 327, 327, 41, 115, 338, 84, 317, …
## $ end_station_name  <chr> "Damen Ave & Clybourn Ave", "Daley Center Plaza", "…
## $ end_station_id    <int> 163, 81, 5, 294, 117, 117, 81, 303, 164, 53, 168, 1…
## $ start_lat         <dbl> 41.94018, 41.87208, 41.88424, 41.94553, 41.92154, 4…
## $ start_lng         <dbl> -87.65304, -87.62954, -87.62963, -87.64644, -87.653…
## $ end_lat           <dbl> 41.93193, 41.88424, 41.87405, 41.97835, 41.94018, 4…
## $ end_lng           <dbl> -87.67786, -87.62963, -87.62772, -87.65975, -87.653…
## $ member_casual     <chr> "casual", "member", "member", "casual", "casual", "…

glimpse(m07_2020)

## Rows: 551,480
## Columns: 13
## $ ride_id           <chr> "762198876D69004D", "BEC9C9FBA0D4CF1B", "D2FD8EA432…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike", "docke…
## $ started_at        <chr> "2020-07-09 15:22:02", "2020-07-24 23:56:30", "2020…
## $ ended_at          <chr> "2020-07-09 15:25:52", "2020-07-25 00:20:17", "2020…
## $ start_station_name <chr> "Ritchie Ct & Banks St", "Halsted St & Roscoe St", …
## $ start_station_id  <int> 180, 299, 329, 181, 268, 635, 113, 211, 176, 31, 14…
## $ end_station_name  <chr> "Wells St & Evergreen Ave", "Broadway & Ridge Ave",…
## $ end_station_id    <int> 291, 461, 156, 94, 301, 289, 140, 31, 191, 142, 31,…
```

```
## $ start_lat          <dbl> 41.90687, 41.94367, 41.93259, 41.89076, 41.91172, 4…
## $ start_lng          <dbl> -87.62622, -87.64895, -87.63643, -87.63170, -87.626…
## $ end_lat            <dbl> 41.90672, 41.98404, 41.93650, 41.91831, 41.90799, 4…
## $ end_lng            <dbl> -87.63483, -87.66027, -87.64754, -87.63628, -87.631…
## $ member_casual      <chr> "member", "member", "casual", "casual", "member", "…
```

```
glimpse(m08_2020)
```

```
## Rows: 622,361
## Columns: 13
## $ ride_id            <chr> "322BD23D287743ED", "2A3AEF1AB9054D8B", "67DC1D133E…
## $ rideable_type      <chr> "docked_bike", "electric_bike", "electric_bike", "e…
## $ started_at         <chr> "2020-08-20 18:08:14", "2020-08-27 18:46:04", "2020…
## $ ended_at           <chr> "2020-08-20 18:17:51", "2020-08-27 19:54:51", "2020…
## $ start_station_name <chr> "Lake Shore Dr & Diversey Pkwy", "Michigan Ave & 14…
## $ start_station_id   <int> 329, 168, 195, 81, 658, 658, 196, 67, 153, 177, 313…
## $ end_station_name   <chr> "Clark St & Lincoln Ave", "Michigan Ave & 14th St",…
## $ end_station_id     <int> 141, 168, 44, 47, 658, 658, 49, 229, 225, 305, 296,…
## $ start_lat          <dbl> 41.93259, 41.86438, 41.88464, 41.88409, 41.90299, 4…
## $ start_lng          <dbl> -87.63643, -87.62368, -87.61955, -87.62964, -87.683…
## $ end_lat            <dbl> 41.91569, 41.86422, 41.88497, 41.88958, 41.90300, 4…
## $ end_lng            <dbl> -87.63460, -87.62344, -87.62757, -87.62754, -87.683…
## $ member_casual      <chr> "member", "casual", "casual", "casual", "casual", "…
```

```
glimpse(m09_2020)
```

```
## Rows: 532,958
## Columns: 13
## $ ride_id            <chr> "2B22BD5F95FB2629", "A7FB70B4AFC6CAF2", "86057FA01B…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ started_at         <chr> "2020-09-17 14:27:11", "2020-09-17 15:07:31", "2020…
## $ ended_at           <chr> "2020-09-17 14:44:24", "2020-09-17 15:07:45", "2020…
## $ start_station_name <chr> "Michigan Ave & Lake St", "W Oakdale Ave & N Broadw…
## $ start_station_id   <int> 52, NA, NA, 246, 24, 94, 291, NA, NA, NA, 273, 145,…
## $ end_station_name   <chr> "Green St & Randolph St", "W Oakdale Ave & N Broadw…
## $ end_station_id     <int> 112, NA, NA, 249, 24, NA, 256, NA, NA, NA, 273, NA,…
## $ start_lat          <dbl> 41.88669, 41.94000, 41.94000, 41.95606, 41.89186, 4…
## $ start_lng          <dbl> -87.62356, -87.64000, -87.64000, -87.66892, -87.621…
## $ end_lat            <dbl> 41.88357, 41.94000, 41.94000, 41.96398, 41.89135, 4…
## $ end_lng            <dbl> -87.64873, -87.64000, -87.64000, -87.63822, -87.620…
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "…
```

```
glimpse(m10_2020)
```

```
## Rows: 388,653
## Columns: 13
## $ ride_id            <chr> "ACB6B40CF5B9044C", "DF450C72FD109C01", "B6396B54A1…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ started_at         <chr> "2020-10-31 19:39:43", "2020-10-31 23:50:08", "2020…
## $ ended_at           <chr> "2020-10-31 19:57:12", "2020-11-01 00:04:16", "2020…
## $ start_station_name <chr> "Lakeview Ave & Fullerton Pkwy", "Southport Ave & W…
## $ start_station_id   <int> 313, 227, 102, 165, 190, 359, 313, 125, NA, 174, 11…
```

```
## $ end_station_name   <chr> "Rush St & Hubbard St", "Kedzie Ave & Milwaukee Ave…
## $ end_station_id     <int> 125, 260, 423, 256, 185, 53, 125, 313, 199, 635, 30…
## $ start_lat          <dbl> 41.92610, 41.94817, 41.77346, 41.95085, 41.92886, 4…
## $ start_lng          <dbl> -87.63898, -87.66391, -87.58537, -87.65924, -87.663…
## $ end_lat            <dbl> 41.89035, 41.92953, 41.79145, 41.95281, 41.91778, 4…
## $ end_lng            <dbl> -87.62607, -87.70782, -87.60005, -87.65010, -87.691…
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "…

glimpse(m11_2020)

## Rows: 259,716
## Columns: 13
## $ ride_id            <chr> "BD0A6FF6FFF9B921", "96A7A7A4BDE4F82D", "C61526D065…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ started_at         <chr> "2020-11-01 13:36:00", "2020-11-01 10:03:26", "2020…
## $ ended_at           <chr> "2020-11-01 13:45:40", "2020-11-01 10:14:45", "2020…
## $ start_station_name <chr> "Dearborn St & Erie St", "Franklin St & Illinois St…
## $ start_station_id   <int> 110, 672, 76, 659, 2, 72, 76, NA, 58, 394, 623, NA,…
## $ end_station_name   <chr> "St. Clair St & Erie St", "Noble St & Milwaukee Ave…
## $ end_station_id     <int> 211, 29, 41, 185, 2, 76, 72, NA, 288, 273, 2, 506, …
## $ start_lat          <dbl> 41.89418, 41.89096, 41.88098, 41.89550, 41.87650, 4…
## $ start_lng          <dbl> -87.62913, -87.63534, -87.61675, -87.68201, -87.620…
## $ end_lat            <dbl> 41.89443, 41.90067, 41.87205, 41.91774, 41.87645, 4…
## $ end_lng            <dbl> -87.62338, -87.66248, -87.62955, -87.69139, -87.620…
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "…

glimpse(m12_2020)

## Rows: 131,573
## Columns: 13
## $ ride_id            <chr> "70B6A9A437D4C30D", "158A465D4E74C54A", "5262016E0F…
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "…
## $ started_at         <chr> "2020-12-27 12:44:29", "2020-12-18 17:37:15", "2020…
## $ ended_at           <chr> "2020-12-27 12:55:06", "2020-12-18 17:44:19", "2020…
## $ start_station_name <chr> "Aberdeen St & Jackson Blvd", "", "", "", "", "", "…
## $ start_station_id   <chr> "13157", "", "", "", "", "", "", "", "", "", "", ""…
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "", "", "", "", "", ""…
## $ end_station_id     <chr> "TA1306000003", "", "", "", "", "", "", "", "", "",…
## $ start_lat          <dbl> 41.87773, 41.93000, 41.91000, 41.92000, 41.80000, 4…
## $ start_lng          <dbl> -87.65479, -87.70000, -87.69000, -87.70000, -87.590…
## $ end_lat            <dbl> 41.88872, 41.91000, 41.93000, 41.91000, 41.80000, 4…
## $ end_lng            <dbl> -87.64445, -87.70000, -87.70000, -87.70000, -87.590…
## $ member_casual      <chr> "member", "member", "member", "member", "member", "…

glimpse(m01_2021)

## Rows: 96,834
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ started_at         <chr> "2021-01-23 16:14:19", "2021-01-27 18:43:08", "2021…
## $ ended_at           <chr> "2021-01-23 16:24:44", "2021-01-27 18:47:12", "2021…
```

```
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor…
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660…
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Wood St & Augu…
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "657", "13258",…
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4…
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696…
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4…
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700…
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "…

glimpse(m02_2021)

## Rows: 49,622
## Columns: 13
## $ ride_id            <chr> "89E7AA6C29227EFF", "0FEFDE2603568365", "E6159D746B…
## $ rideable_type      <chr> "classic_bike", "classic_bike", "electric_bike", "c…
## $ started_at         <chr> "2021-02-12 16:14:56", "2021-02-14 17:52:38", "2021…
## $ ended_at           <chr> "2021-02-12 16:21:43", "2021-02-14 18:12:09", "2021…
## $ start_station_name <chr> "Glenwood Ave & Touhy Ave", "Glenwood Ave & Touhy A…
## $ start_station_id   <chr> "525", "525", "KA1503000012", "637", "13216", "1800…
## $ end_station_name   <chr> "Sheridan Rd & Columbia Ave", "Bosworth Ave & Howar…
## $ end_station_id     <chr> "660", "16806", "TA1305000029", "TA1305000034", "TA…
## $ start_lat          <dbl> 42.01270, 42.01270, 41.88579, 41.89563, 41.83473, 4…
## $ start_lng          <dbl> -87.66606, -87.66606, -87.63110, -87.67207, -87.625…
## $ end_lat            <dbl> 42.00458, 42.01954, 41.88487, 41.90312, 41.83816, 4…
## $ end_lng            <dbl> -87.66141, -87.66956, -87.62750, -87.67394, -87.645…
## $ member_casual      <chr> "member", "casual", "member", "member", "member", "…

glimpse(m03_2021)

## Rows: 228,496
## Columns: 13
## $ ride_id            <chr> "CFA86D4455AA1030", "30D9DC61227D1AF3", "846D87A156…
## $ rideable_type      <chr> "classic_bike", "classic_bike", "classic_bike", "cl…
## $ started_at         <chr> "2021-03-16 08:32:30", "2021-03-28 01:26:28", "2021…
## $ ended_at           <chr> "2021-03-16 08:36:34", "2021-03-28 01:36:55", "2021…
## $ start_station_name <chr> "Humboldt Blvd & Armitage Ave", "Humboldt Blvd & Ar…
## $ start_station_id   <chr> "15651", "15651", "15443", "TA1308000021", "525", "…
## $ end_station_name   <chr> "Stave St & Armitage Ave", "Central Park Ave & Bloo…
## $ end_station_id     <chr> "13266", "18017", "TA1308000043", "13323", "E008", …
## $ start_lat          <dbl> 41.91751, 41.91751, 41.84273, 41.96881, 42.01270, 4…
## $ start_lng          <dbl> -87.70181, -87.70181, -87.63549, -87.65766, -87.666…
## $ end_lat            <dbl> 41.91774, 41.91417, 41.83066, 41.95283, 42.05049, 4…
## $ end_lng            <dbl> -87.69139, -87.71676, -87.64717, -87.64999, -87.677…
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "…

glimpse(m04_2021)

## Rows: 337,230
## Columns: 13
## $ ride_id            <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508…
## $ rideable_type      <chr> "classic_bike", "docked_bike", "docked_bike", "clas…
```

```
## $ started_at         <chr> "2021-04-12 18:25:36", "2021-04-27 17:27:11", "2021…
## $ ended_at           <chr> "2021-04-12 18:56:55", "2021-04-27 18:31:29", "2021…
## $ start_station_name <chr> "State St & Pearson St", "Dorchester Ave & 49th St"…
## $ start_station_id   <chr> "TA1307000061", "KA1503000069", "20121", "TA1305000…
## $ end_station_name   <chr> "Southport Ave & Waveland Ave", "Dorchester Ave & 4…
## $ end_station_id     <chr> "13235", "KA1503000069", "20121", "13235", "20121",…
## $ start_lat          <dbl> 41.89745, 41.80577, 41.74149, 41.90312, 41.74149, 4…
## $ start_lng          <dbl> -87.62872, -87.59246, -87.65841, -87.67394, -87.658…
## $ end_lat            <dbl> 41.94815, 41.80577, 41.74149, 41.94815, 41.74149, 4…
## $ end_lng            <dbl> -87.66394, -87.59246, -87.65841, -87.66394, -87.658…
## $ member_casual      <chr> "member", "casual", "casual", "member", "casual", "…
```

Notice that all dataframes with the same column names and orders, so merge them into 01 table (12 months)

Merge into 01 table

```
all_trips <- rbind(m05_2020, m06_2020, m07_2020, m08_2020, m09_2020,
                   m10_2020, m11_2020, m12_2020, m01_2021, m02_2021, m03_2021,
                   m04_2021)
```

Phase 3: Process

Inspect the data to check type

```
summary(all_trips)

##    ride_id          rideable_type       started_at          ended_at
##  Length:3742202     Length:3742202     Length:3742202     Length:3742202
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:3742202     Length:3742202     Length:3742202     Length:3742202
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.64   Min.   :-87.87   Min.   :41.54   Min.   :-88.07
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.64
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.08   Max.   :-87.52   Max.   :42.16   Max.   :-87.44
##                                   NA's   :4906    NA's   :4906
##  member_casual
##  Length:3742202
```

```
##  Class :character
##  Mode  :character
##
##
##
##
```

Change type from character to datetime

```
all_trips$started_at <- ymd_hms(all_trips$started_at)
all_trips$ended_at <- ymd_hms(all_trips$ended_at)
glimpse(all_trips)
```

```
## Rows: 3,742,202
## Columns: 13
## $ ride_id           <chr> "02668AD35674B983", "7A50CCAF1EDDB28F", "2FFCDFDB91…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike", "docke…
## $ started_at        <dttm> 2020-05-27 10:03:52, 2020-05-25 10:47:11, 2020-05-…
## $ ended_at          <dttm> 2020-05-27 10:16:49, 2020-05-25 11:05:40, 2020-05-…
## $ start_station_name <chr> "Franklin St & Jackson Blvd", "Clark St & Wrightwoo…
## $ start_station_id  <chr> "36", "340", "260", "251", "261", "206", "261", "18…
## $ end_station_name  <chr> "Wabash Ave & Grand Ave", "Clark St & Leland Ave", …
## $ end_station_id    <chr> "199", "326", "260", "157", "206", "22", "261", "18…
## $ start_lat         <dbl> 41.8777, 41.9295, 41.9296, 41.9680, 41.8715, 41.847…
## $ start_lng         <dbl> -87.6353, -87.6431, -87.7079, -87.6500, -87.6699, -…
## $ end_lat           <dbl> 41.8915, 41.9671, 41.9296, 41.9367, 41.8472, 41.869…
## $ end_lng           <dbl> -87.6268, -87.6674, -87.7079, -87.6368, -87.6468, -…
## $ member_casual     <chr> "member", "casual", "casual", "casual", "member", "…
```

Add new columns: ride_length, day_of_week, hour to aggregate data

```
all_trips1 <- all_trips %>% mutate(ride_length = difftime(ended_at,
                                started_at, units = 'mins'))
all_trips1$date <- date(all_trips$started_at)
all_trips1$day_of_week <- weekdays(all_trips$started_at)
all_trips1$hour <- hour(all_trips$started_at)
all_trips1$month <- month(all_trips$started_at)
glimpse(all_trips1)
```

```
## Rows: 3,742,202
## Columns: 18
## $ ride_id           <chr> "02668AD35674B983", "7A50CCAF1EDDB28F", "2FFCDFDB91…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike", "docke…
## $ started_at        <dttm> 2020-05-27 10:03:52, 2020-05-25 10:47:11, 2020-05-…
## $ ended_at          <dttm> 2020-05-27 10:16:49, 2020-05-25 11:05:40, 2020-05-…
## $ start_station_name <chr> "Franklin St & Jackson Blvd", "Clark St & Wrightwoo…
## $ start_station_id  <chr> "36", "340", "260", "251", "261", "206", "261", "18…
## $ end_station_name  <chr> "Wabash Ave & Grand Ave", "Clark St & Leland Ave", …
## $ end_station_id    <chr> "199", "326", "260", "157", "206", "22", "261", "18…
## $ start_lat         <dbl> 41.8777, 41.9295, 41.9296, 41.9680, 41.8715, 41.847…
## $ start_lng         <dbl> -87.6353, -87.6431, -87.7079, -87.6500, -87.6699, -…
## $ end_lat           <dbl> 41.8915, 41.9671, 41.9296, 41.9367, 41.8472, 41.869…
## $ end_lng           <dbl> -87.6268, -87.6674, -87.7079, -87.6368, -87.6468, -…
```

```
## $ member_casual      <chr> "member", "casual", "casual", "casual", "member", "…
## $ ride_length        <drtn> 12.950000 mins, 18.483333 mins, 97.300000 mins, 13…
## $ date               <date> 2020-05-27, 2020-05-25, 2020-05-02, 2020-05-02, 20…
## $ day_of_week        <chr> "Wednesday", "Monday", "Saturday", "Saturday", "Fri…
## $ hour               <int> 10, 10, 14, 16, 12, 13, 12, 18, 17, 10, 14, 8, 14, …
## $ month              <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, …
```

Check type of ride_length and min, max

```
typeof(all_trips1$ride_length)

## [1] "double"

all_trips1 %>% summarise(ride_min = min(ride_length), ride_max = max(ride_length))

##         ride_min      ride_max
## 1 -29049.97 mins 54283.35 mins
```

Remove all missing values & errors

Remove all missing values in start_station_name and end_station_name Also Remove all error values for ride_length (which =<0 and >=1440 minutes or 24hours)

```
all_trips2 <- subset(all_trips1, all_trips1$ride_length > 0
                     & all_trips1$ride_length < 1440
                     & all_trips1$start_station_name!=""
                     & all_trips1$end_station_name!="")
colnames(all_trips2)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"      "ride_length"        "date"
## [16] "day_of_week"        "hour"               "month"
```

Select only some columns needed to analyze

```
all_trips_conclusion <- all_trips2 %>%
  select(ride_id, rideable_type, started_at, ended_at, member_casual,
         date, ride_length, day_of_week, month, hour)
```

Phase 4 & 5: Analyze and Share (Visualization)

What is total number of trips for members and casuals, which proportion of total trips they represent?

```
all_trips_conclusion %>%
   group_by(member_casual)%>%
   summarise(num_trips = n()) %>%
   mutate(proportion = round(num_trips / sum(num_trips)*100,0))

## # A tibble: 2 x 3
##   member_casual num_trips proportion
##   <chr>            <int>      <dbl>
```

```
## 1 casual         1443216        41
## 2 member         2053001        59
```

Calculate some statistic metrics for ride_length as min, max, mean, median

```
all_trips_conclusion %>%
  group_by(member_casual) %>%
  summarise(min_ride_length = min(ride_length),
            max_ride_length = max(ride_length),
            median_ride_length = median(ride_length),
            mean_ride_length = mean(ride_length))
```

```
## # A tibble: 2 x 5
##   member_casual min_ride_length max_ride_length median_ride_length
##   <chr>         <drtn>          <drtn>          <drtn>
## 1 casual        0.01666667 mins 1439.900 mins   21.30000 mins
## 2 member        0.01666667 mins 1439.717 mins   11.43333 mins
## # … with 1 more variable: mean_ride_length <drtn>
```

Using SQL code to query in R

Top 10 common start and end stations casual took trips by using sqldf function

```
casual_geo_start <- sqldf("SELECT member_casual, start_station_name,
count(start_station_name) AS num_trips
                FROM all_trips2
                WHERE member_casual = 'casual'
                GROUP BY start_station_name
                ORDER BY count(start_station_name) DESC
                LIMIT 10", method='auto')
casual_geo_start
```

```
##    member_casual          start_station_name num_trips
## 1         casual      Streeter Dr & Grand Ave     28202
## 2         casual    Lake Shore Dr & Monroe St     23250
## 3         casual             Millennium Park     21141
## 4         casual         Theater on the Lake     16059
## 5         casual       Michigan Ave & Oak St     15013
## 6         casual   Indiana Ave & Roosevelt Rd     14416
## 7         casual   Lake Shore Dr & North Blvd     14282
## 8         casual      Michigan Ave & Lake St     12504
## 9         casual           Clark St & Elm St     12419
## 10        casual Michigan Ave & Washington St     11445
```

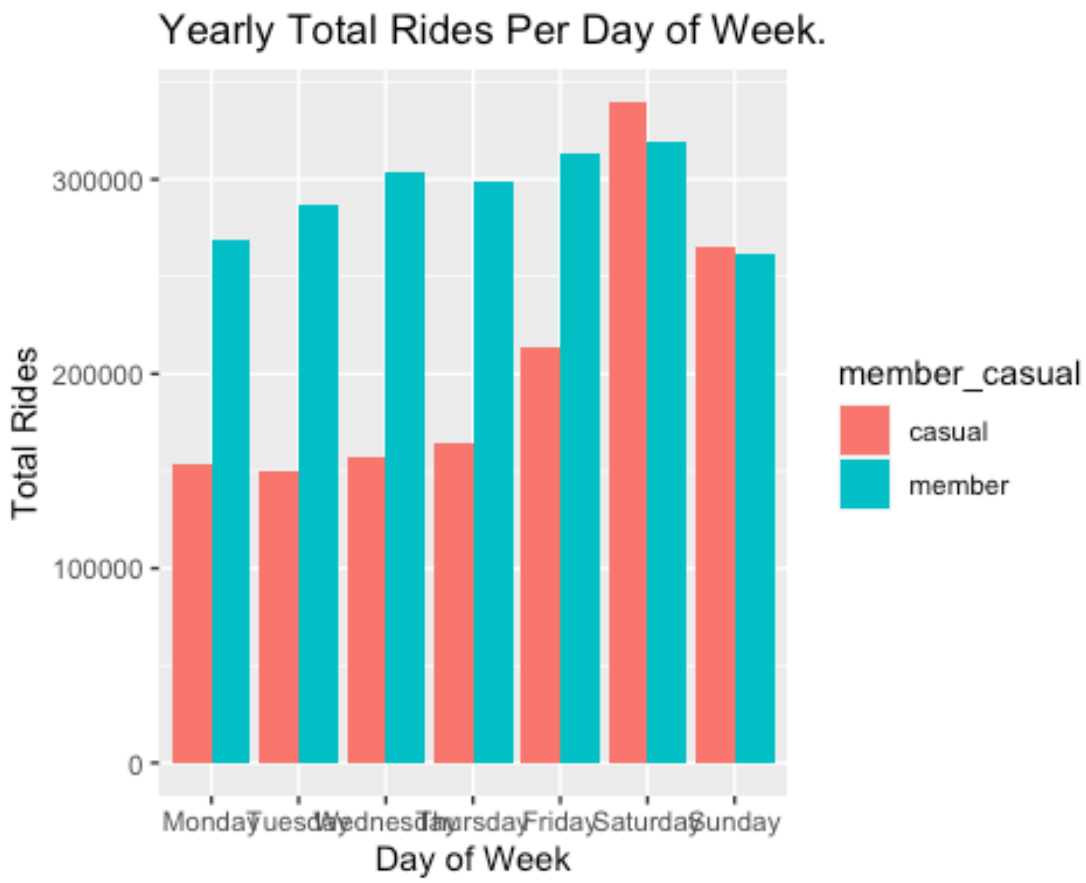
```
casual_geo_end <- sqldf("SELECT member_casual, end_station_name,
count(end_station_name) AS num_trips
                FROM all_trips2
                WHERE member_casual = 'member'
                GROUP BY end_station_name
                ORDER BY count(end_station_name) DESC
                LIMIT 10", method='auto')
casual_geo_end
```

```
##    member_casual          end_station_name num_trips
## 1         member         Clark St & Elm St     21554
## 2         member     St. Clair St & Erie St     17123
## 3         member      Broadway & Barry Ave      16536
## 4         member      Dearborn St & Erie St     16502
## 5         member      Wells St & Concord Ln     16490
## 6         member   Kingsbury St & Kinzie St     15462
## 7         member       Theater on the Lake      15282
## 8         member         Wells St & Elm St     14260
## 9         member Lake Shore Dr & North Blvd      13876
## 10         member        Wells St & Huron St     13872
```

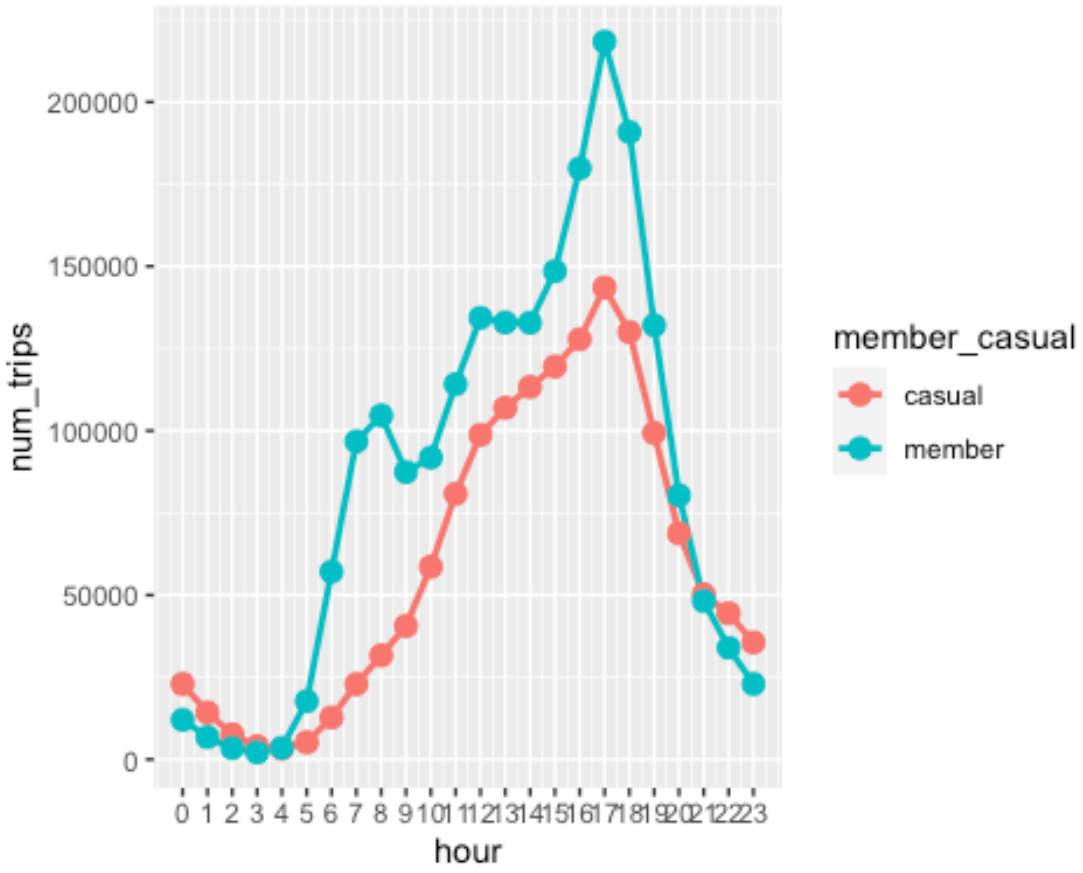
Visualization

Which days of week, hours of day member and casual ride most?

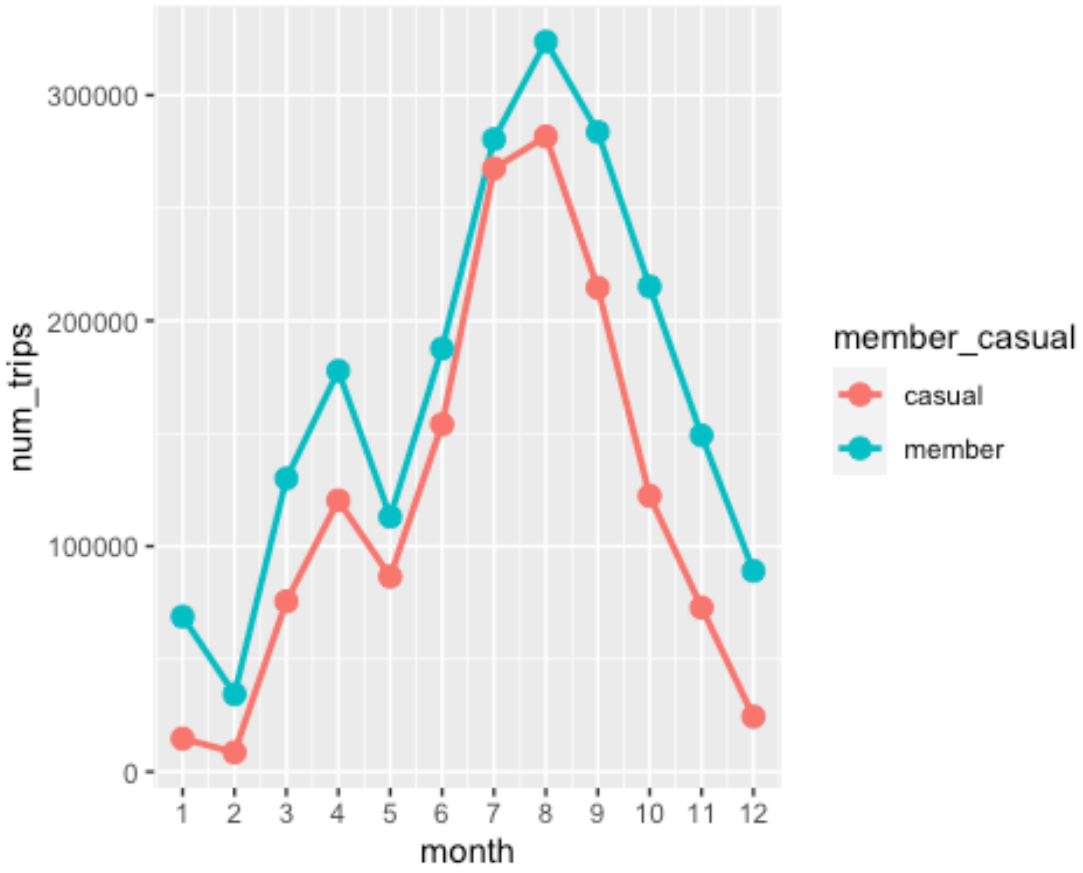
```
all_trips_conclusion %>%
  group_by(member_casual, day_of_week) %>%
  summarise(num_trips = n(), .groups = 'drop') %>%
  ggplot(aes(x = factor(day_of_week,
                        weekdays(min(all_trips_conclusion$date) + 3:9)),
             y = num_trips, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(title = "Yearly Total Rides Per Day of Week.", x = "Day of Week",
       y = "Total Rides") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



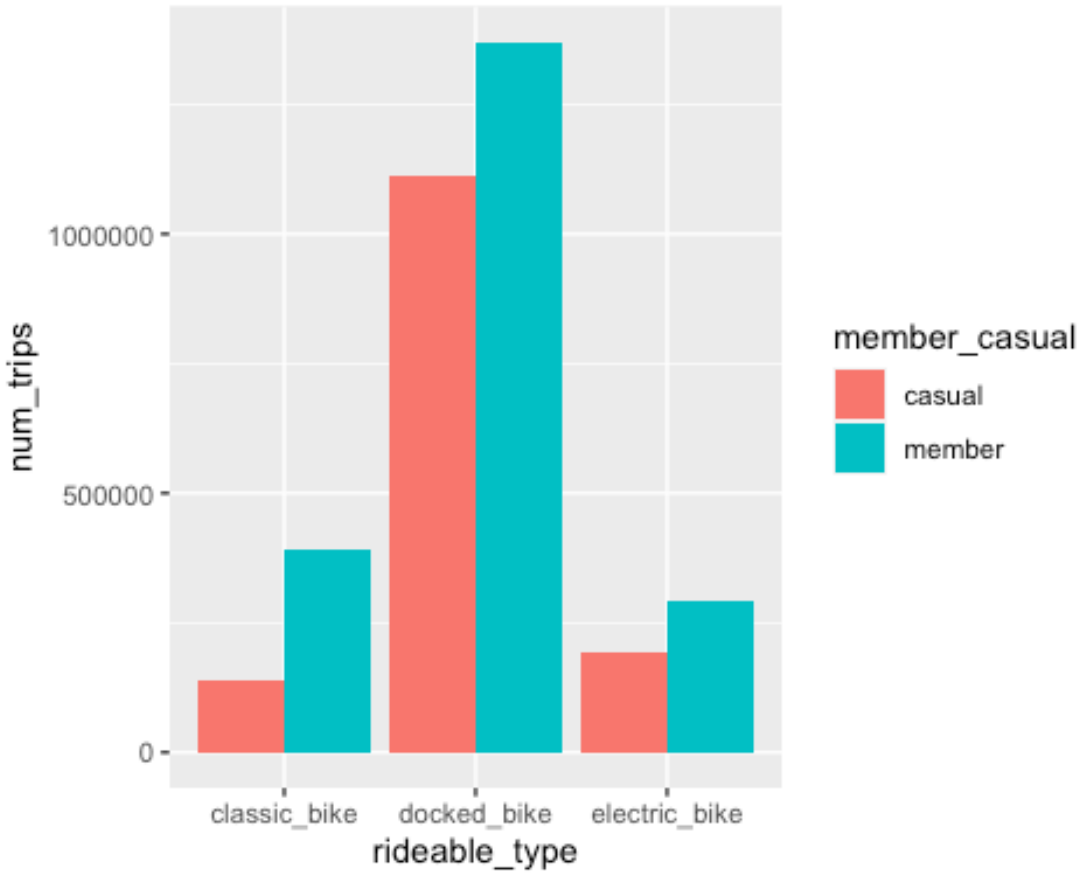

```
all_trips_conclusion %>%
  group_by(member_casual, hour) %>%
  summarise(num_trips = n(), .groups = 'drop') %>%
  ggplot(aes(x = hour, y = num_trips, fill = member_casual, colour= member_casual)) +
  geom_line(size=1) + geom_point(size=3)+
  scale_x_continuous(breaks=c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,
                              14,15,16,17,18,19,20,21,22,23))
```




```
all_trips_conclusion %>%
  group_by(member_casual, month) %>%
  summarise(num_trips = n(), .groups = 'drop') %>%
  ggplot(aes(x = month, y = num_trips, fill = member_casual, colour= member_casual)) +
  geom_line(size=1) + geom_point(size=3) +
  scale_x_continuous(breaks=c(0,1,2,3,4,5,6,7,8,9,10,11,12)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```




```
all_trips_conclusion %>%
  group_by(member_casual, rideable_type) %>%
  summarise(num_trips = n(), .groups = 'drop') %>%
  ggplot(aes(x = rideable_type, y = num_trips, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "identity")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



###

Phase 6: Act

Casual trips account for 41% total trip. If we can convert casual to member, revenue will increase significantly.

Some insights I found out: What different between members and casuals use bikes?

Ride length: Members most likely use bike for 15 minutes while casuals tend to use for longer time (30 minutes).

When they use most:

- Month: both members and casuals use most from July to September.

- Day of week: casuals most likely use in the weekend (Friday-Sunday) while members use all days of week, not much different between days of week. Maybe members use for commuting to work whereas casuals only use for hanging out.

- Hour: Peak hours are from 15PM-18PM both members and casuals. Type of ride: Both use docked bike most compared to the others.

Recommendation:

1. Run special discount for membership in some criteria that casuals most likely use so they will see the benefits if convert to be members and those who are members also enjoying their benefits of membership:

- docked_bike
- 15PM-18PM
- unlimited ride duration.

2. Run the advertising campaign to focus on benifit to use bike_share for commuting to work so casuals use more in weekday instead of weekend only. If they ride bikes to work, they tend to use much more and consider to convert to member to have more promotion.
3. Especially should run advertising campaign in peak season (July - September), so there are more chances casuals should register for membership.
4. Launch advertising campaign for membership on the 10 common start and end stations where casual took rides most.