**Case Study 1**

**Owner/Stakeholder** – Cyclistic Bike-Share

**Description of the Problem / Goal**

The firm has two major categories of riders: casual riders and annual members. Casual riders are the ones who avail the single-ride passes and full-day passes. Annual members have one-year pass to use the bikes.

The business analysts have determined that the Annual members are more profitable to the the company and focus has to be on expanding this userbase. One primary solution to do this is to attract the casual riders to become annual members and this is the base for the future growth of Cyclistic.

**ASK**

*Business Goal* – To Design Marketing strategies to convert casual riders into annual members.

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?

As a Junior Data Analyst my first task is to analyse data and find the answer to the following.

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

**PREPARE**

*Data Source:* I’m using a public dataset which can be found here [click here to view the dataset](https://divvy-tripdata.s3.amazonaws.com/index.html).  
I’ll be using the data for the rides that occurred in the last 12 months i.e. September 2020 to September 2021.

*Description of the data:* There are 12 .csv files, each contains one month worth of ride data. Each ride record in these files captures attributes of the occurred ride namely:

|  |  |
| --- | --- |
| ride\_id | Unique Id of the ride record |
| rideable\_type | Types can be either docked\_bike, electric\_bike, classic\_bike |
| Started\_at | Time-stamp when the ride started |
| Ended\_at | Time-stamp when the ride ended |
| Start\_station\_name | Name of the station where the ride started |
| Start\_station\_id | Unique id of the start station |
| End\_station\_name | Name of the station where the ride ended |
| End\_station\_id | Unique id of the end station |
| Start\_lat | Latitude of the start location |
| Start\_lng | Longitude of the start location |
| End\_lat | Latitude of the end location |
| End\_lng | Longitude of the end location |
| Member\_casual | Type of rider i.e. member or casual |

**PROCESS**

**Cleaning**

To clean data, I’ve performed the following checks on each file.

1. Removed all duplicates rows.
2. Verified that “rideable\_type” and “member\_casual” columns have only categorical values and no nulls.
3. The date-time columns have been formatted as “dd-mm-yyyy hh:mm:ss”.
4. Some “ride\_id” values are formatted as text. For some of these, it results into error as scientific notation points to more than 250 characters. Such rows have been eliminated. Also, “ride\_id” values have been consistently kept in the range 14-20 characters using LEFT function.
5. Removed whitespaces for “start\_station\_name” and “end\_station\_name” using TRIM function.
6. Also, there are presence of records with end time occurring before start time. Such records have been eliminated.
7. For consistency, “start\_station\_id” and “end\_station\_id” have been formatted as text.

**Manipulation**

Additionally, I’ve created new columns in each file using Microsoft Excel. Following are the steps I’ve followed

1. Create a new column and name it “ride\_length”.
   1. Ride length is calculated by subtracting the column “started\_at” from the column “ended\_at”.
   2. Format the value in “hh:mm:ss” using the custom format “[hh]:mm:ss”.
2. Create another column “day\_of\_week”.
   1. Calculate day of week by using the spreadsheet formula WEEKDAY(C2,1)
   2. Here numerical value 1 = Sunday and 7 = Saturday
3. Repeat this for each file.

**ANALYSE**

**Descriptive Analysis**  
I’ve performed initial analysis using pivot tables for each month separately.

1. Pivot Table 1

This depicts ride numbers and average ride length for each of the weekdays.

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of ride\_id** | **Average of ride\_length** |
| 1 | 142754 | 00:26:10 |
| 2 | 110005 | 00:20:09 |
| 3 | 110921 | 00:18:25 |
| 4 | 92248 | 00:18:31 |
| 5 | 101665 | 00:19:24 |
| 6 | 111679 | 00:20:51 |
| 7 | 135051 | 00:25:10 |
| **Grand Total** | **804323** | **00:21:38** |

1. Pivot Table 2

This analysis the number of rides and average ride length for each of the rideable types namely docked bike, classic bike and electric bike.

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of ride\_id** | **Average of ride\_length** |
| classic\_bike | 503011 | 00:20:14 |
| docked\_bike | 45065 | 00:59:15 |
| electric\_bike | 256247 | 00:17:46 |
| **Grand Total** | **804323** | **00:21:38** |

1. Pivot Table 3

This compares the number of rides and average ride length for each of the member types i.e. casual and annual.

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of ride\_id** | **Average of ride\_length** |
| Casual | 412662 | 00:28:47 |
| Member | 391661 | 00:14:06 |
| **Grand Total** | **804323** | **00:21:38** |

1. Pivot Table 4

This displays the average and maximum ride length across the month.

|  |  |
| --- | --- |
| **Average of ride\_length** | **Max of ride\_length2** |
| 00:21:38 | 693:49:10 |

1. Pivot Table 5

This table records the number of rides and average ride length for each of the start stations.

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Count of ride\_id** | **Average of ride\_length** |
|  | 88455 | 00:18:03 |
| Streeter Dr & Grand Ave | 14529 | 00:37:07 |
| DuSable Lake Shore Dr & North Blvd | 9334 | 00:25:17 |
| Michigan Ave & Oak St | 8112 | 00:33:09 |
| Theater on the Lake | 7175 | 00:27:56 |
| Wells St & Concord Ln | 6903 | 00:17:04 |

I’ve collated these summaries for each month and put together in one spreadsheet do determine monthly as well as seasonal analysis.

Some interesting facts and trends that I’ve found out are as follows:

1. The longest ride occurred in June-July 2021 with a record of roughly 38 days.
2. Mid months of the year i.e. May to September recorded high number of bike rides with peak of 822k being recorded in July itself.
3. November to February witnessed low ride number. February recorded the lowest dip with just 50k rides.
4. Streeter Dr & Grand Ave is the most popular start station with as high as 15k rides in one month.

**Month-wise Comparisons**

Following table gives overall monthly analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Count of rides** | **Average ride length** | **Max Ride length** |
| 2020-09 | 532956 | 00:25:16 | 904:43:21 |
| 2020-10 | 386740 | 00:20:06 | 595:24:23 |
| 2020-11 | 259716 | 00:19:42 | 598:54:00 |
| 2020-12 | 131139 | 00:15:58 | 162:20:59 |
| 2021-01 | 96832 | 00:15:16 | 330:25:55 |
| 2021-02 | 49622 | 00:24:25 | 502:09:14 |
| 2021-03 | 228494 | 00:22:52 | 528:01:39 |
| 2021-04 | 337225 | 00:24:08 | 796:16:42 |
| 2021-05 | 531633 | 00:26:02 | 898:41:36 |
| 2021-06 | 729590 | 00:26:05 | 932:24:09 |
| 2021-07 | 822397 | 00:24:13 | 818:27:09 |
| 2021-08 | 804323 | 00:21:38 | 693:49:10 |

Some other important observations are:

1. Classic bikes are the most popular choice with 2 out of every 3 rides being with classic bike..
2. Docked bikes are a favorite for those who ride for more than 1 hour.
3. Casual riders have higher average ride length than annual member. Almost double of the later.
4. Saturdays and Sundays witnessed highest number of rides. Average ride length has also been recorded high for these two days.

**Year-view Analysis**

I will use SQL to merge all the monthly data to do an overall year analysis. I will be using Microsoft SQL Server to do all database operations.

Number of Rides on each Weekday.