This case study is a part of the [Google Data Analytics Professional Certification](https://www.coursera.org/account/accomplishments/professional-cert/HSUAKUNV6UVV). As a junior data analyst I am tasked with using historical data from Cyclistic to understand the behavior, determine trends, find similarities and differences and use the insight gained to provide recommendations for the marketing team in other to create a campaign to pull more casual riders to the annual membership plan.

**Phases of the project**

|  |  |
| --- | --- |
| **Deliverable** | **Task** |
| 1. Define and identify business task | * Consider key stakeholder and their goals. * A clear statement of the business task. |
| 1. Data preparation | * Identify where the data is located and how it is organized. * Determining data reliability, origins, comprehensiveness, current relevance, and credibility(Does the data ROCCC?). * Addressing licensing, privacy, security, and accessibility. * Verifying data’s integrity. * Identify problems in the data and does it help answer the business task. * Sort and filter the data. |
| 1. Cleaning the data | * Choosing the tools. * Check for dirty data – making sure there are no duplicates, dealing with null values, checking for errors(spelling, outdated and inaccurate data). |
| 1. Analysis | * Organize and format the data. * Perform statistical calculation. * Identify trends and relationship from calculations. |
| 1. Sharing key finds with key stakeholders | * Determine the best way to share finding and what tool to use for effective visualization. * Present findings. |
| 1. Act on key findings | * Providing three recommendations for the marketing campaign. |

**Phase 1: Define and identify business task**

Cyclistic provides a transportation service using bike-sharing with over 5000 bicycles that are geotracked and locked into a network of over 600 stations across Chicago. Cyclistic’s customers comprise of two types of users – annual members (they have yearly subscription) and casual riders (they purchase either single-rides or full-day pass). The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. With most users likely a ride for leisure, about 30% use them to commute to work each day.

|  |  |  |
| --- | --- | --- |
| **Stakeholders** | **Expectations** | **Project/Business Goals** |
| Lily Moreno (Director of Marketing) | Data evidence to back up her theory and marketing recommendations. | Convert a sizable group of casual members to annual members. |
| Cyclistic marketing analytics team | Uncovering the behavior behind different customer types. | Provide data-driven evidence for decision making. |
| Cyclistic executive team | Showing relevant and straightforward insights driven by data. | Implementing strategies to promote business goals. |

**Statement of the business task:**

In the interest of the company’s growth, Lily Moreno (Director of Marketing) believes the company should capitalize on the recent conclusion made by Cyclistic’s finance analysts which is the annual members are much more profitable than casual riders. In other to do that, the analytics team needs to analyze the company’s historical data to find differences in behavior and trends of both types of consumers.

**Phase 2: Data preparation**

The data been used was extracted from [here](https://divvy-tripdata.s3.amazonaws.com/index.html) in a comma-separated values for the previous 12 month (January 2021 – December 2021). Also The data has been made available by Motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement).

Disclaimer; Cyclistic is a fictional entity and Divvy's open data is used for the purpose of this case study.

It is also important to note that data-privacy issues prohibit you from using riders’ personally identifiable information. This means that you won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Each record consist of a bike trips with field talking about: the Ride ID, types of bikes, start & end trips by their station names, the date & time with its longitude & latitude points and lastly the types of customers.

**Data Quality:**

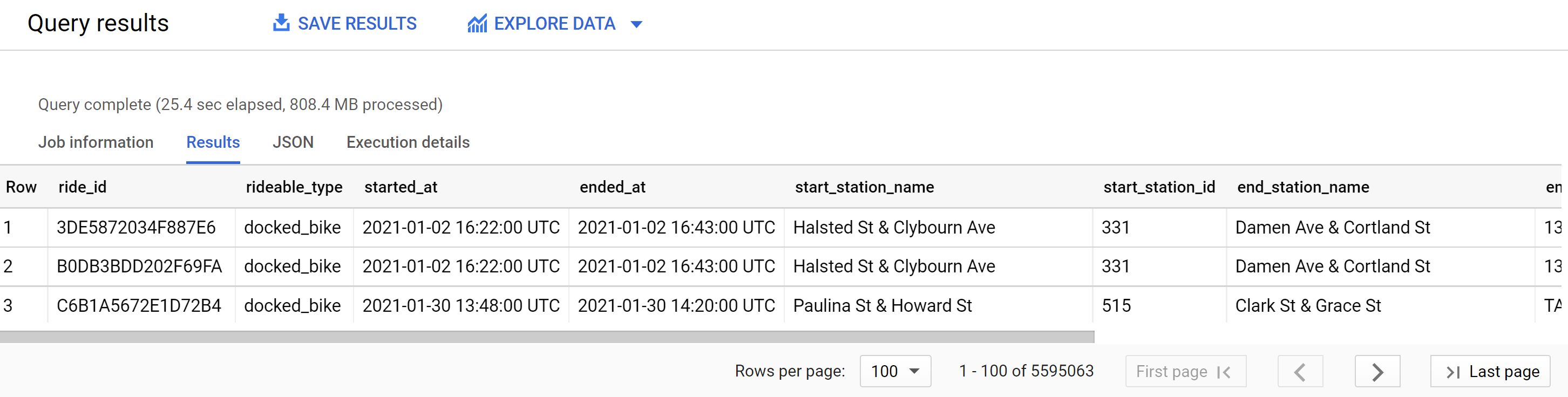
Following the ROCCC data quality assessment which deals with the Reliability, Originality, Comprehensiveness, how Current and Cited the data is. About a million plus rows had null values (Unknown values) out of five million plus rows, also there are hundreds of duplicate ride ids (which uniquely identifies the data) but asides that the data source was from the company (first-party group). In terms of how current, it is a pass for the data because the time frame been examined gives us up-to-date views and mostly free of human errors.

**Phase 3: Cleaning the data**

Due to the size of the data with over 5 million rows and 13 columns, I will be using BigQuery which is a structured query language.

Using BigQuery I uploaded the data for each month from January 2021 – December 2021 as tables into the Cyclistic database I created.

In other to have all the information in all 12 tables in one table I used the Union statement and afterwards saved the result as a table in the Cyclistic database. Click [here](https://console.cloud.google.com/bigquery?_ga=2.9145781.1063797448.1642215061-1594746148.1640790891&_gac=1.46376661.1642226010.CjwKCAiA24SPBhB0EiwAjBgkhh2lfooVHSEVB8FKgVE3MxnAW-7-iQCeiLiYEoXguAnAxyUWtz9S7xoCdNMQAvD_BwE&project=my-project-336704&ws=!1m5!1m4!1m3!1smy-project-336704!2sbquxjob_64339c9_17e83642e8a!3sUS&page=queryresults&pli=1&j=bq:US:bquxjob_64339c9_17e83642e8a) view code used.



New row number of **5,595,063**.

**The cleaning process:**

* Separated the date and time from both started\_at and ended\_at columns.
* Checked for duplicates in the ride\_id, fortunately there were no duplicates
* Checked for errors in the ride\_id and found **72** errors. Basically across the board the ride\_id is meant to have 16 characters but this **72** ride\_id did not have 16 characters. I stored them in another table and continued cleaning.
* I looked for null value in the longitude and latitude of the stations but found nothing which is a good sign, meaning I can populate the null value of station names (with **423218** rows with either null values in the station start and ending rides) using its cardinal points. I found this to be difficult and risky.
* Deleting all rows with null station\_id which totaled to **1,006,760**. Already about 18% of the data has been removed. I think the better thing I should have done is try to populate the ID or stored the null valued station\_id as a table and try populate it.
* Added new columns to indicate the specific month, date of the week and hour of the days.
* Deleted rows with rides >= 1440 minutes and <= 0 minutes **59018** rows.
* Delete columns that were not needed in the table.

Click [here](https://console.cloud.google.com/bigquery?_ga=2.9145781.1063797448.1642215061-1594746148.1640790891&_gac=1.46376661.1642226010.CjwKCAiA24SPBhB0EiwAjBgkhh2lfooVHSEVB8FKgVE3MxnAW-7-iQCeiLiYEoXguAnAxyUWtz9S7xoCdNMQAvD_BwE&project=my-project-336704&ws=!1m4!1m3!8m2!1s800076167646!2sa5beafcddb6042fd8085080d488f915b&pli=1&sq=800076167646:a5beafcddb6042fd8085080d488f915b) for the codes used in the cleaning process.

**Phase 4: Analysis**

Using the cleaned data I did a few analysis to examine the behaviors of both customers, so my first step was to do a descriptive analysis to find out the maximum, minimum and average duration of both groups in minutes.

I then did further analysis on duration and number for rides taken for each month and weekdays by both groups. Afterwards I looked into the behavior of both rides in accordance to the hours of the day for the weekends and weekdays.

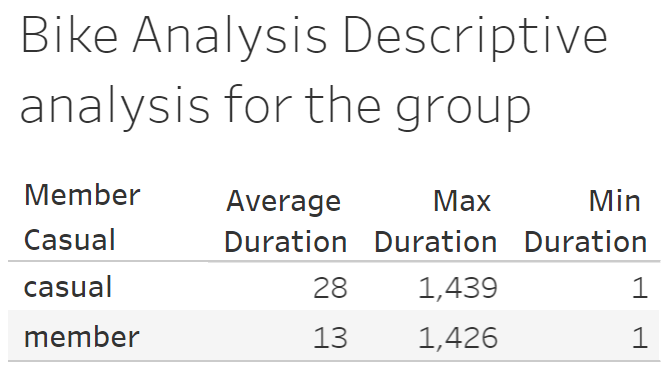
Further analysis was then on the choice of rides the riders prefer. Lastly I looked into the activities of both groups in departure stations and arrival stations.

Click [here](https://console.cloud.google.com/bigquery?_ga=2.247069380.1333351823.1642976732-1594746148.1640790891&_gac=1.188983897.1642267841.CjwKCAiA24SPBhB0EiwAjBgkhh2lfooVHSEVB8FKgVE3MxnAW-7-iQCeiLiYEoXguAnAxyUWtz9S7xoCdNMQAvD_BwE&project=my-project-336704&ws=!1m9!1m3!8m2!1s800076167646!2sc81d4f9b57c24b12840642f806d00096!1m4!4m3!1smy-project-336704!2sCyclistic_Case_Study!3sDescriptive_analysis_for_the_group&sq=800076167646:c81d4f9b57c24b12840642f806d00096) view code used.

**Phase 5: Sharing key finds with key stakeholders**

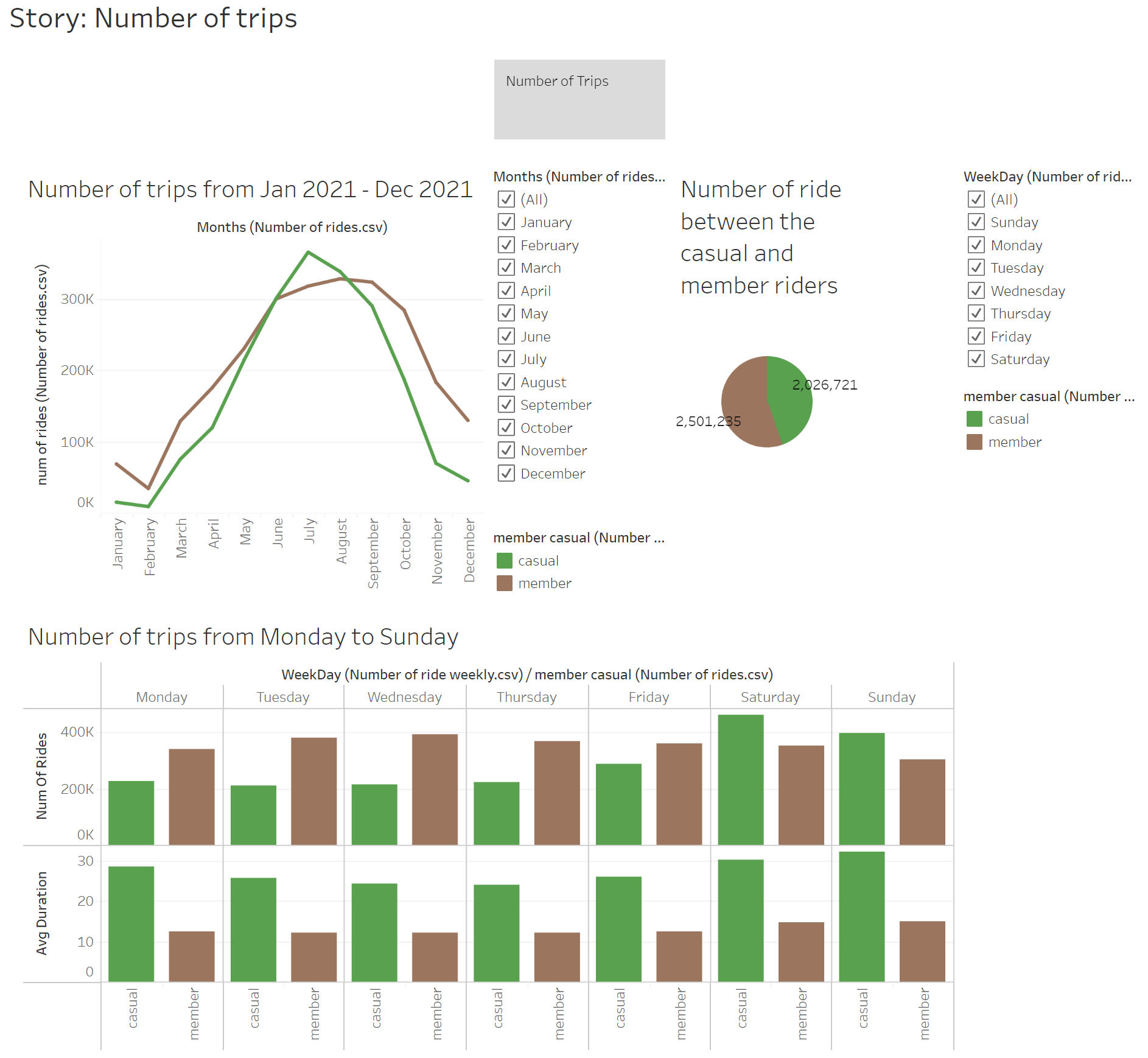
I decided to use Tableau to tell the story of what the analysis is saying. The business task is to find out how bring a number of casual rider to become annual members. Hence, the behavior of both groups were analyzed to find similarities, differences, patterns and trends. Based on the finding we can then visual, come up with hypotheses and make recommendations.

**Descriptive analysis:**



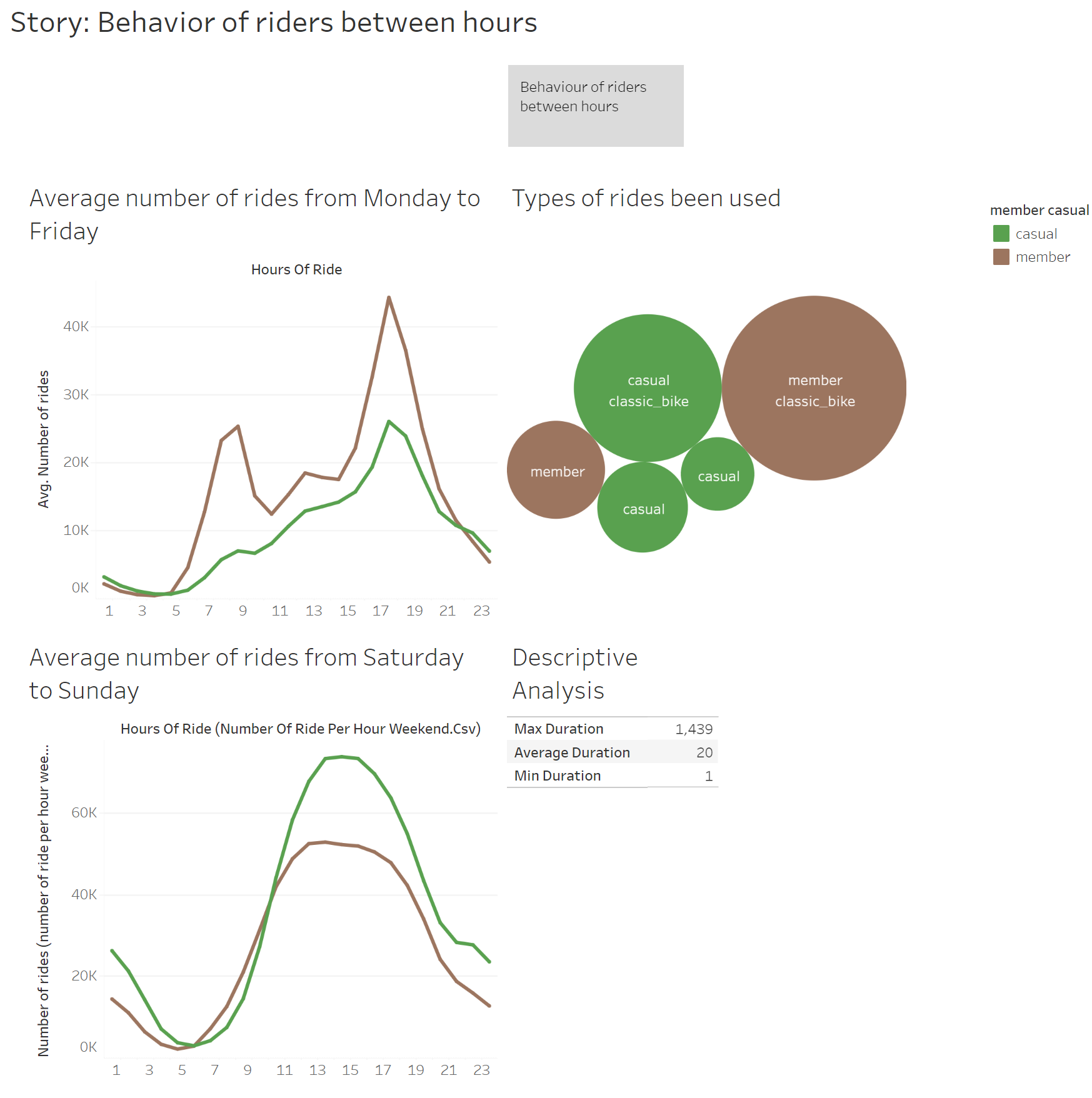
The above table shows the behavior of both groups when it comes the duration of rides in minutes and on an average we can see that the casual members spend more time using the bikes. This could be result of leisure riding from tourist etc..

**Behavior of riders over the year and week:**



From the above image we can see that we have more annual member riding with the summer time have peak number of rides. According to their weekly behavior, there are more casual riders on the weekends. This could be a result of the temperature of the weather during each season when it comes to the peak number of rides been in the summer time. Another observation is that between January and June we had more annual member riding but from June to August there were more causal riders and for the rest of the year there is a strong decline in number of riders but there more annual riders. This could be as a result of summer events or as said earlier, it could be influenced by tourist.

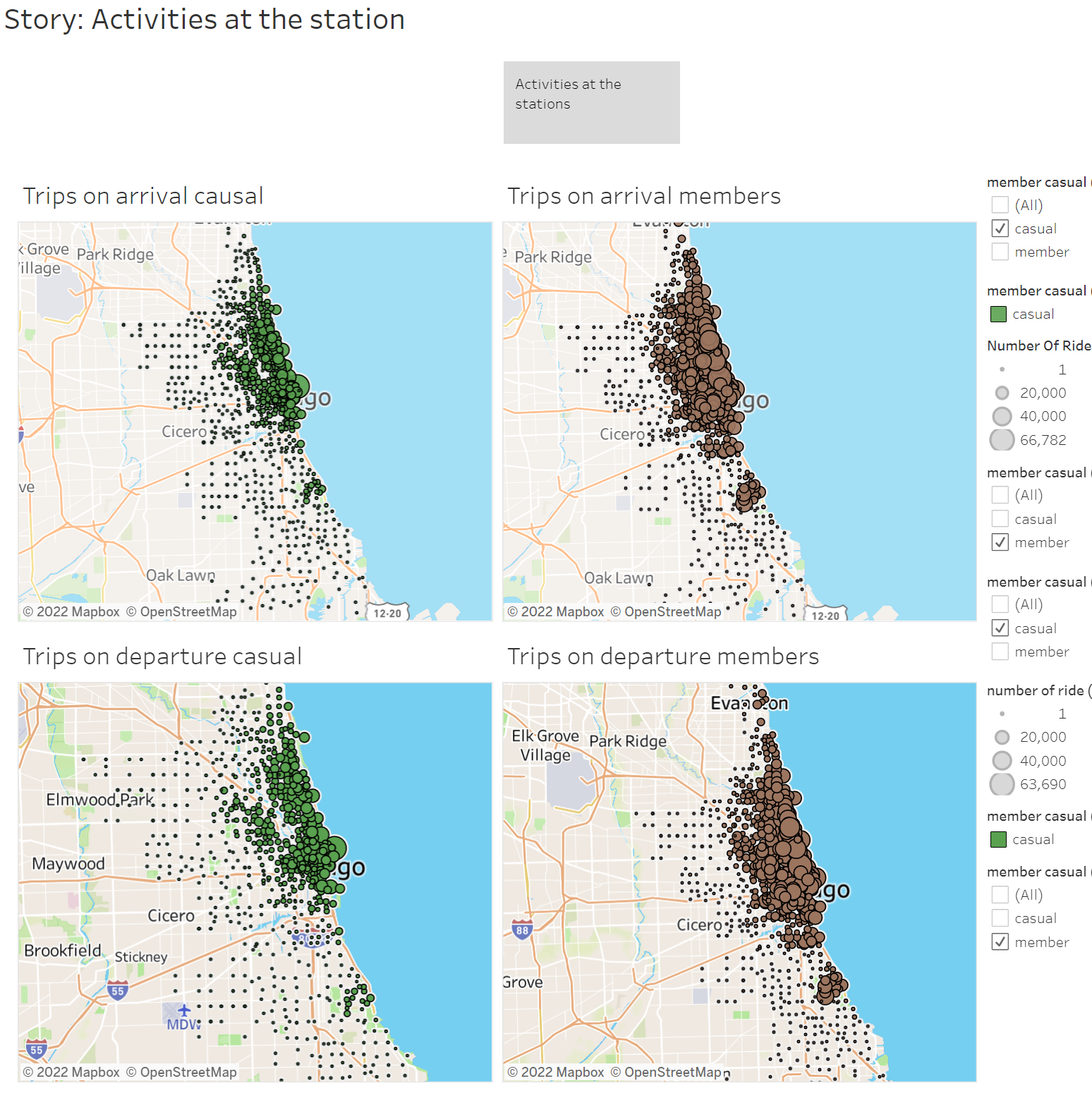
**Behavior of rides between hours over the weekends and other weekdays:**



The above images tells the story of the behavior of both riders between hours during the weekdays and weekends. During the weekdays there are two significant spikes in number of rides for the annual member, one is between 7 and 8 am and the other is at 5 pm. This could be a result of rush hour traffic (getting to work and leaving work). However, during the weekends the bike usage for both members and casual riders rises gradually peaking between 12PM and 5PM.

The most used type of bikes by both groups is still the classic bike.

**Behavior when it comes to the departure and arrival stations:**



The image shows the number of riders with respect to their geographical coordinates (stations). With tableau this maps are more interactive and we can narrow down areas in the map and draw insight from them. The density of both groups can also be reviewed.

All interactive visualization can be found [here](https://public.tableau.com/app/profile/abraham.owodunni/viz/Cyclistic_16427134255520/StoryActivitiesatthestation_1) and [here](https://public.tableau.com/app/profile/abraham.owodunni/viz/BikeAnalysisDescriptiveanalysisforthegroup/BikeAnalysisDescriptiveanalysisforthegroup)

**Phase 6: Act on key findings**

Now this is the phase where I provide recommendations for the marketing team.

**Trends and relationships**

**Speculations**

Dues to the data been provided which only answers “What” are the riders doing and not answering the “Why” questions. At this stage it is better to do more research into the behavior of the rider and tackling the “Why” questions.

With that been said here are some of the assumptions been made:

**Members:**

* They are more consistent and have a steady average ride duration all year-round.
* Also they are more active over throughout the day, but more during rush hours.

**Casuals:**

* More active in the weekends and over the summer time.
* Less consistency in comparison to actual members, when it comes to activities over the day and average duration of rides (causal member are more active in the afternoon).
* Consistently having longer ride duration over the year.

**Recommendations**

* In other to bring more casual riders to an annual membership, the marketing team campaign should be launched between February to August and in the weekends as we can see that the casual riders pick up more rides and are more active during that period.
* With casual riders usage often peaks on the weekends, a new annual membership can be introduced. Making it a weekend subscription that runs throughout the year.
* Also discounts can be introduced in the already existing annual membership for longer rides.
* Furthermore, surveys should be introduced in order to better understand the behavior of the riders. Get information on demography, their experience, why they use the bikes etc… so basically collecting more qualitative data.