Insights into the use of Cycltics Services by Casual and Annual Membership Riders from April 2020 -2021

Produced by: Candice Schippers

The purpose of this report is to guide future marketing planning by the Cycltics marketing team. Specifically by providing insights into how annual members and casual riders use Cycltics services to investigate marketing directed towards casual riders. To gain these insights Cyclitics historical data sets from April 2020 to March 2021 were downloaded, stored, cleaned and analysed as detailed in the Appendix.

Insights

Looking at the data for all rides taken on Cycltics bikes we have longer rides being taken in the summer months. This is shown in Figure 1 with 95% and 99% of all rides in July 2020 falling under 1:37:31 and 3:23:01 long respectively.

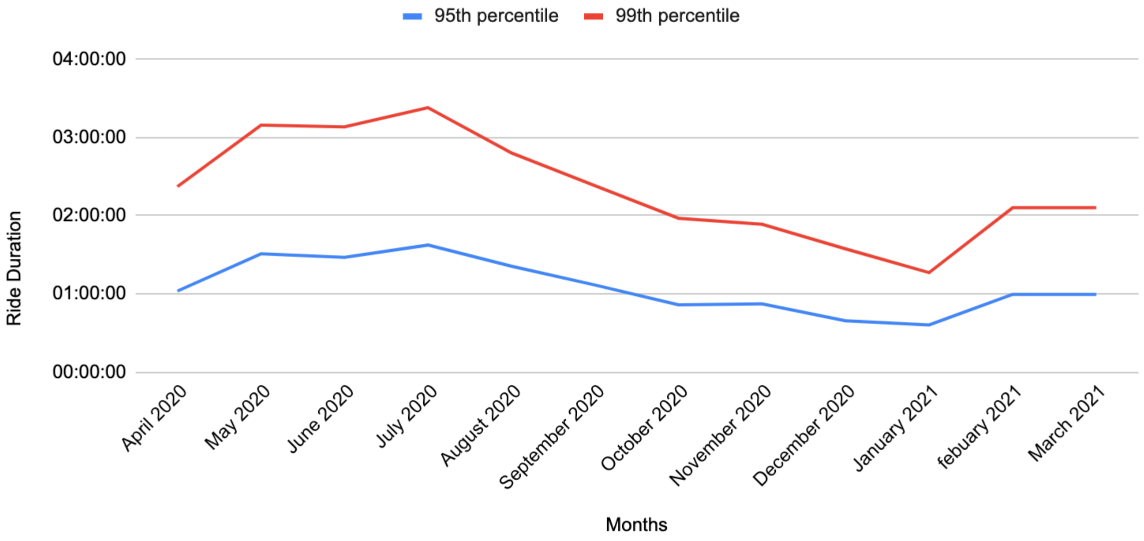


Figure 1: 95th percentile and 99th percentile's of ride duration from April 2020 - March 2021

The 95th and 99th percentile's of the duration of all Cycltic rides taken.

When we look at the total ride numbers across annual members and casual riders we find that peak numbers are occurring during these months. Annual Members make the majority of the rides taken compared to casual rides as seen in Figure 2. However, casual riders tend to take longer ride trips on the median and mean average (Figure 3 and Figure 4). Member riders tend to take short trips usually between 10 - 20 minutes on average Figure 3.

Figure 2: Total Ride Numbers April 2020 - 2021

The number of casual rides taken compared to the number of member rides taken.

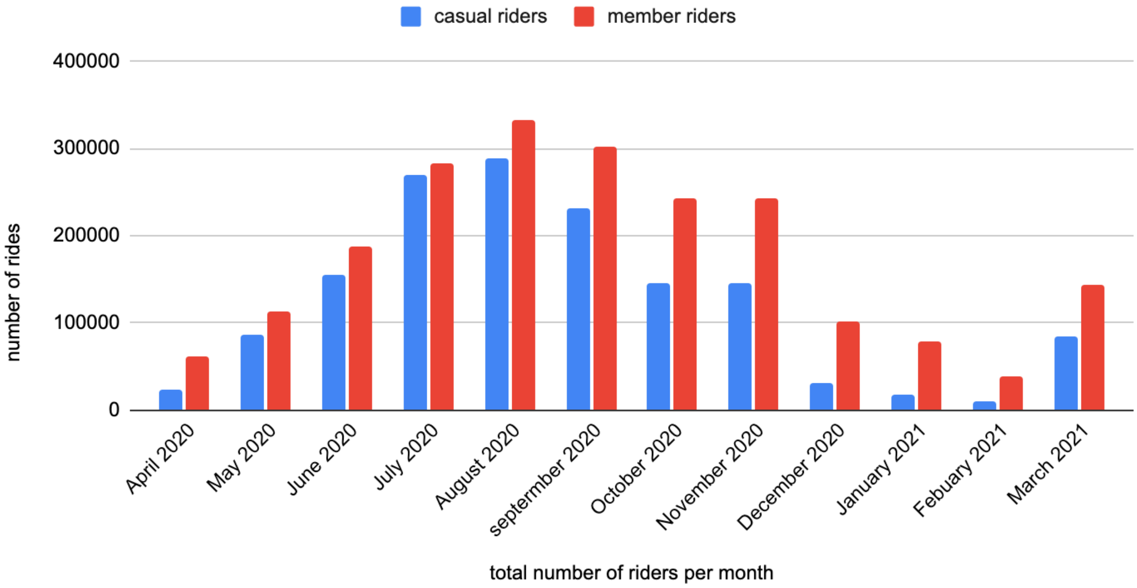


Figure 3: Mean average ride duration of all rides April 2020 - March 2021

Comparison of time time duration of annual membership riders compared to casual riders

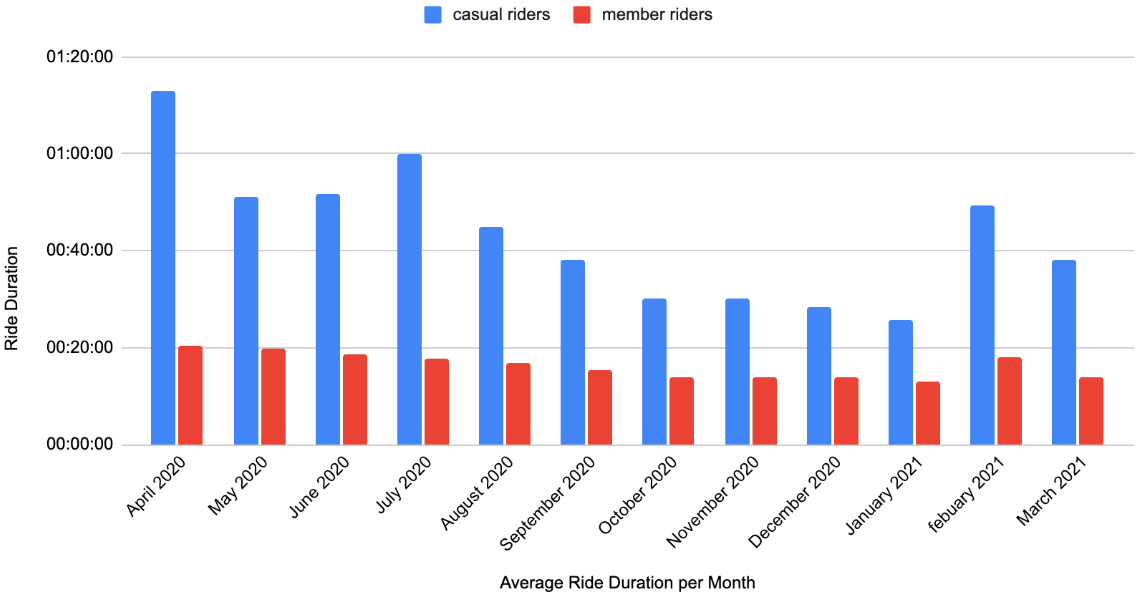
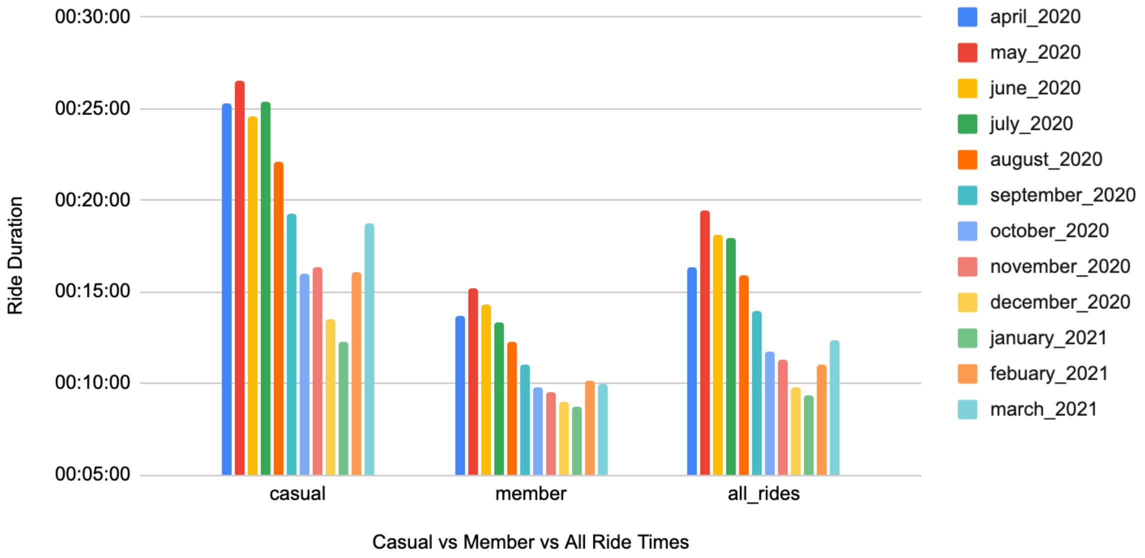


Figure 4: Median Ride Duration of Casual and Member Riders from April 2020 -March 2021

The median ride times of Cycltics casual riders compared Cycltics to member riders across a 12 month period.



When we break this down over a week we find that member riders use Cycltics bikes consistently throughout the week as shown in Figure 5. The casual riders who tend to ride more frequently over the weekend from Friday through to Monday Figures 6.

Figure 5: Total member rides taken during the week from April 2020 - March 2021

The total number of rides taken by annual membership riders.

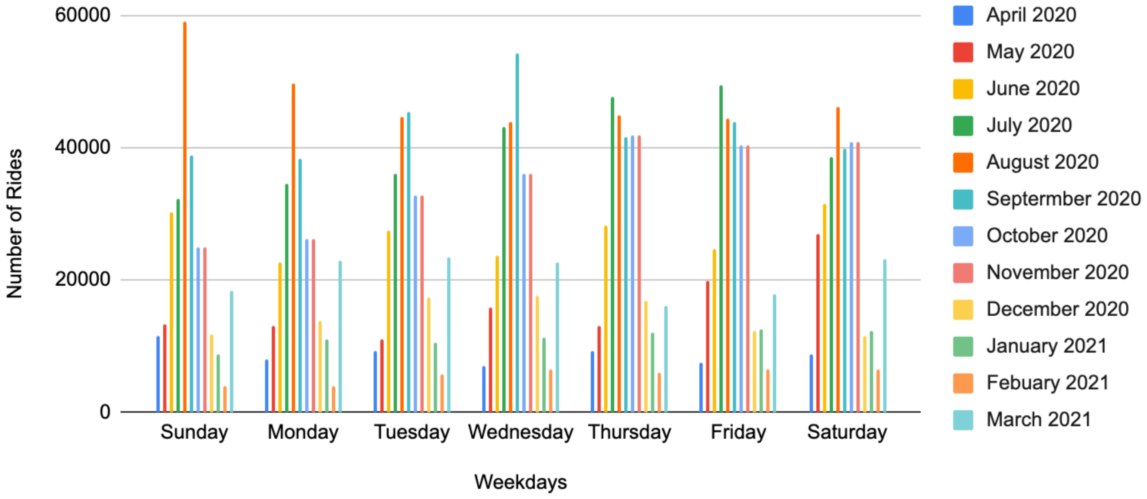
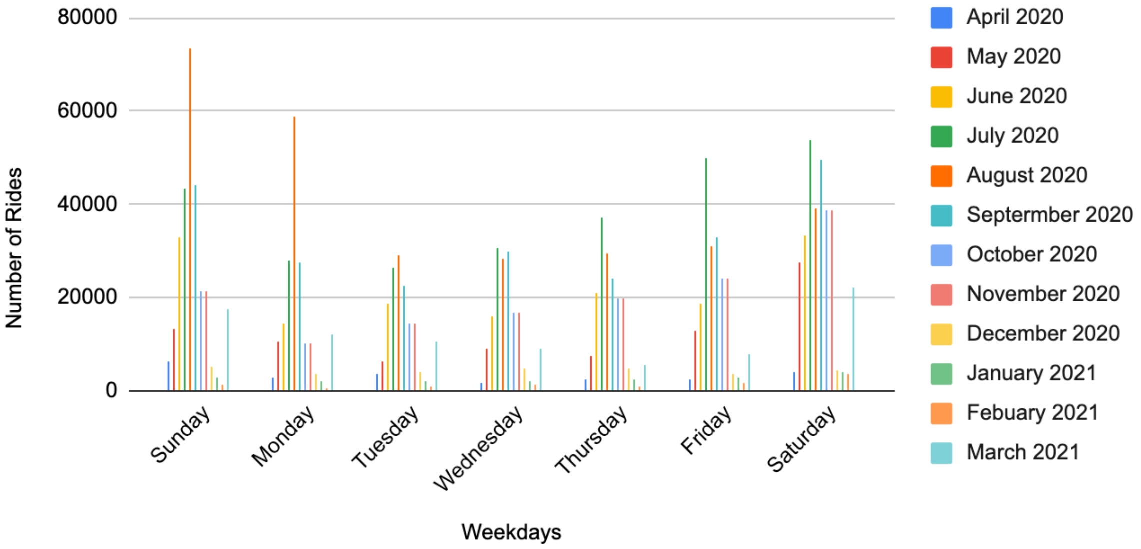


Figure 6: Total casual rides taken during the week between April 2020 - March 2021

The total number of casual rides taken throughout the year across the days of the week.



When looking at ride duration over the week we find that casual riders take longer rides in general. The longest of these casual rides occurred on the weekends Figures 7 and 8. over the weekend periods than member riders Figures 9 and 10. Member riders' ride times remain consistent throughout the week Figures 9 and 10.

Figure 7: The mean average of daily member rides during the week April 2020 - March 2021

The mean average ride duration of member riders using Cycltics services during the week

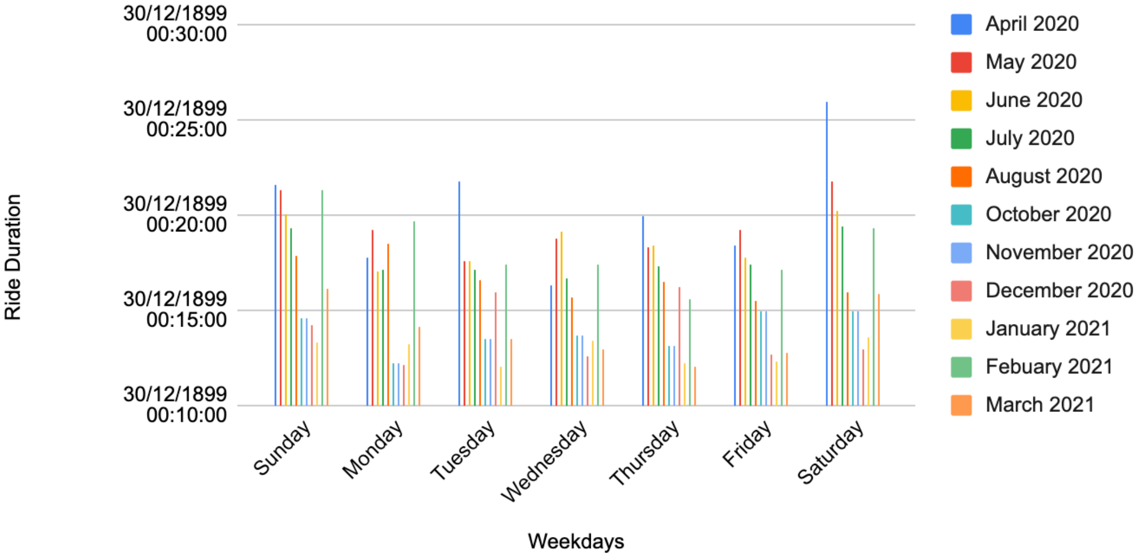


Figure 8: The median member riders daily ride duration for April 2020 - March 2021

The median ride duration of casual riders across the week throughout the year

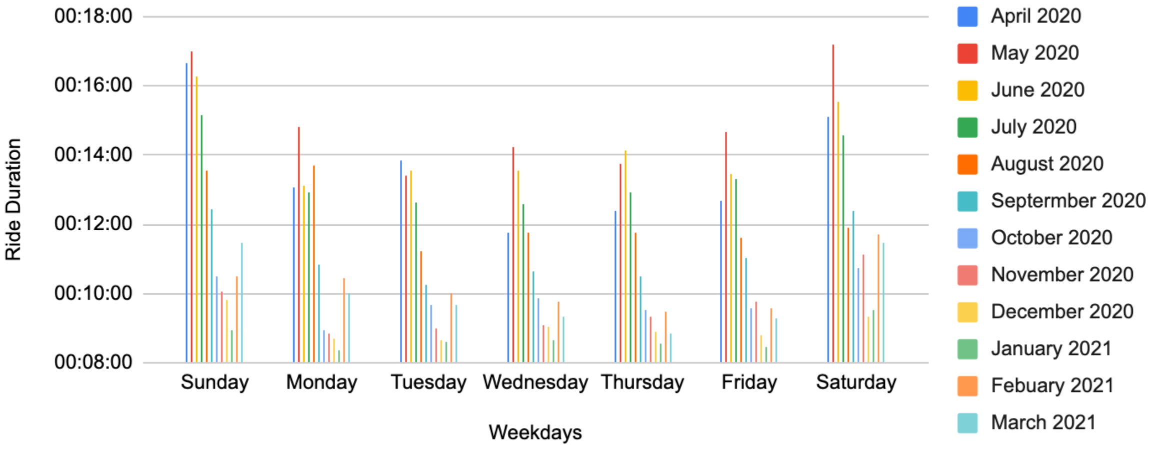
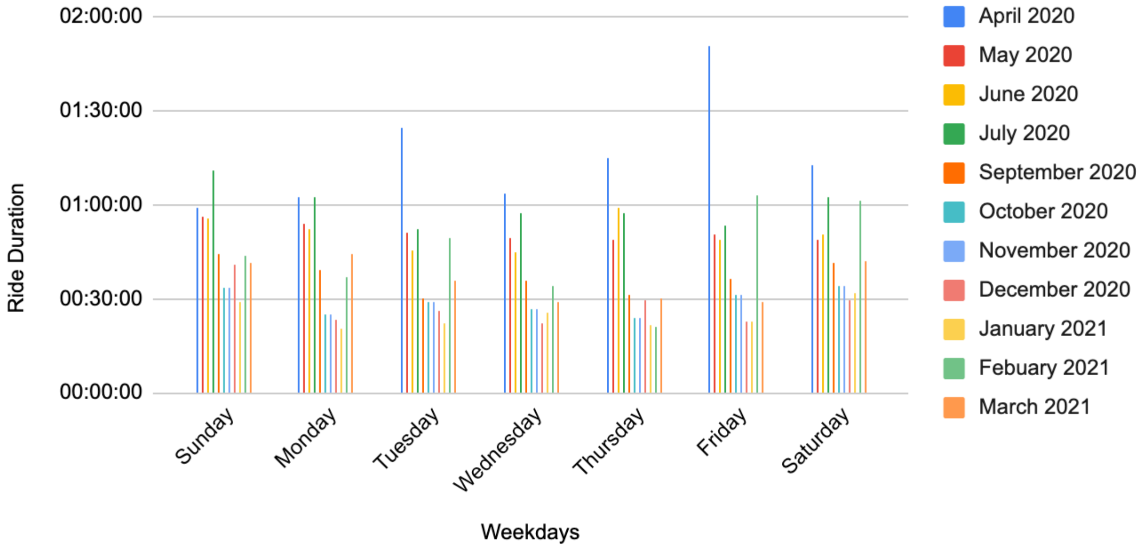


Figure 9: The mean average daily ride duration of casual riders April 2020 -March 2021

The mean average ride duration of casual riders using Cycltics services during the week.



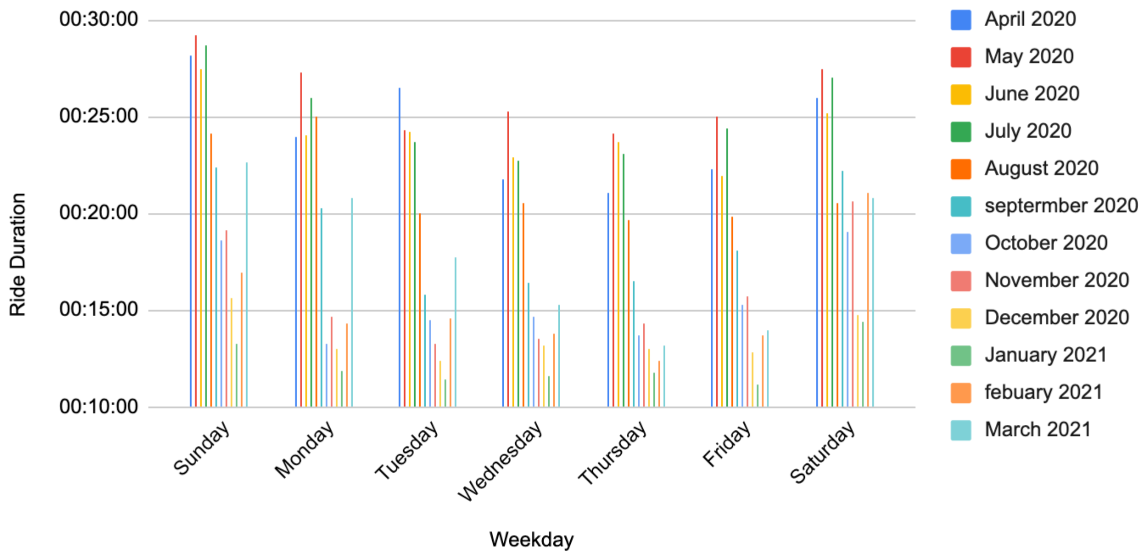


Figure 10:The median casual daily ride duration for April 2020 - March 2021

The median ride duration of casual riders across the week throughout the year

During the data cleaning and analysis of the data sets, there were inconsistencies and errors with the recording of location-based data. This meant that which locations were more popular for casual versus member riders could not be determined at this point. Due to a lack of financial data the cost implications for ride duration for casual and annual member riders could not be determined at this time. Both of these will need to be investigated to get a better picture of how casual rides versus member riders use Cycltics services for future marketing campaigns

Findings

These trends results suggest that the marketing team consider the following when developing future marketing strategies:

1. Investing in beginning new marketing campaigns in the northern hemisphere late winter and early spring in order to take advantage of the increase in ridership over the summer months in Chicago.
2. Concentrating marketing on weekdays to reduce the dip in rides taken during the week compared to weekends.
3. Advertising around these time frames will need to concentrate on how annual membership rides benefit annual riders over casual ride trips.

The first two of these steps will be acted on by the marketing team while the third recommendation will be further investigated by the Data Analytics team.

Further research to support future marketing campaigns needs to considered by the data Analytics team along with fixing the data collection issues of location based data with the relevant departments

- What locations have the most active members and casual riders?

- What is the cost per ride time of casual versus member riders?
- Why are casual riders more active over the Friday to Monday period?
- Why are casual riders travelling for longer periods?

The answers to these questions would be helpful in gaining further insights into how to develop the most effective marketing strategies possible.

Appendix 1: Acknowledgements

Cycltics is a fictional company given as a case study option as a capstone project of the Google Analytics certificate. The data is open source under a licence agreement Google has with Motivate International Inc. the company that collected and owns this data. Motivate International Inc. and Google have removed all personal identifying information related to Motivate International Inc. for their privacy. This case study is to showcase the skills I have learnt through studying the Google Analytics Course.

Appendix 2: Data Collection and Storage

I downloaded all the DIVVY_Tripdata from the following URL: https://divvy-tripdata.s3.amazonaws.com/index.html . I then created a folder system to store all the data sets as I went through the process of cleaning and analysing them. I started by creating a primary folder for the case study that I called cycltics_data_sets. Within this folder, I then made 3 sub-folders titled as follows:

- cleaned_data_sets
- orginal_data_sets
- summarised_data_sets

This was done for 3 main reasons. Firstly, maintain the integrity of the information throughout the data collection, cleaning and analysis process. To make it easy to locate data quickly and efficiently. Finally, should there be any data corruption at any stage of the case study I could return to the previous files.

Appendix 3: Data processing

Google Sheets

I began by uploading each data set individually into google sheets to conduct the initial data cleaning. I completed this for all of the DIVVY_trip data sets except for August 2020. This was because Google Sheets flagged the file as too large. I then attempted to use Excel in read-only mode to investigate this further. I noticed a lot of data was missing and/or corrupted. As google sheets had stated the file was too large for it I thought it may be the case for excel as well so I decided to try opening the .csv file using a text reader. The text file reader program showed that although there were nulls, the dataset was mostly complete. This meant that either a SQL server or R IED environment would be required to clean and analyse the DIVVY_tripdata file for August. I decided I would use R to clean and analyse this particular file as I find that it to be the better option for documentation and analysis of data sets. With this done I then moved on to cleaning each of the files.

April 2020

I started by calculating the total ride_length (=D2-C2, then formatted it to time) and added it as a column to the right of the ride_end_time(column D). Then I added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7. one being Sunday and & being Saturday). I chose to do these steps first as I would need this information to be ready to analyse the bulk of the information I would need for the report. It would also help in the identification of any errors regarding time and date-based data.

Having performed this task, I then used conditional formatting to help find any nulls within the data set by assigning empty cells a colour. I planned on using filtering to find them and use context clues from the other columns to work out what needs to be entered, into the blank cells. As these nulls were primarily associated with location data I planned on using the time-based data to assist me in filling in the blanks possible.

I then started correcting the location-based data by filtering for cells in the ride_length length column from 00:00:00 up to 00:01:00. I then used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce what the nulls were. I then used the same process for 00:02:40 and a few others, where I could find a similar ride_length time and it did not appear to have multiple options for the end_station and end_station_id used that to figure out the missing end of ride data and insert it. The rest I left as nulls as I did not have enough information at my disposal to fill them in with any level of confidence.

During the filtering process, I noticed that the ride_length column had a few outliers of 23hr and 59 minutes. I then filtered for these and noted that although the dates for these were the same and the end time listed is earlier than the start time. These entries appeared to be a data entry mistake so I corrected this by swapping the times around.

Having found nothing else to correct I used to find and replace to make the following changes:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'

I also used this function to change capitalised letters to lowercase when my case change extension timed out. I made these changes to ensure that when transferring the cleaned data set into R there would be no risk of errors due to the association of Capitalised letters and special characters within built statistical functions.

I then calculated the Mean Average ride_length, the max_ride_length and the mode of the weekdays on a new sheet I titled analysis. While calculating the max_ride_time, I noticed four large outliers, two of which had end dates well into the next month. I corrected these to be the same date as the start_date_time. I corrected these as they appeared to be a data entry error. I decided to leave the remaining two outliers as they had an end_ride_time of 2-3 days after the start_ride_time and could potentially be accurate

I finalised the processing of the April 2020 data set by renaming the file by adding cleaned to the end. This is to ensure it could be easily identified and the chain of data integrity could be maintained. I then added it to the clean_data_sets folder.

May 2020

Using the insights I gained from cleaning April 2020 dataset. I opened the May 2020 data set and for the above-listed reasoning followed the same process as listed below:

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
7. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

8. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
9Renaming the file by adding cleaned to it to denote that it

June 2020

Using the insights I gained from cleaning the previous two data sets I followed a similar process again

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
7. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', and dashes '-' as there were no parenthesis to remove

- change capitalised letters to lowercase

8. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
 9. Renaming the file by adding cleaned to it to denote that it

September 2020

Using what I learned from the previous datasets I cleaned the September 2020 data set as follows:

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls and they had increased in number. Hence I left them blank
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis
7. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
8. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', and dashes '-' as again there were no parentheses
- change capitalised letters to lowercase

9. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
 10. Renaming the file by adding cleaned to it to denote that it

October 2020

Applying what I have learned through :

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.
7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
9. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

10. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
11. Renaming the file by adding cleaned to it to denote that it

November 2020

With the November 2020 data set I once again followed these steps:

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.
7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
9. Use find and replace to:
   - ‘&’ with
   - ‘space’ between start and end stations names with underscores '_'
   - I removed full stops '.', dashes '-' and parenthesis '(' ')'
   - change capitalised letters to lowercase
10. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
11. Renaming the file by adding cleaned to it to denote that it

December 2020

I continued the data cleaning process by opening up the December data set by doing the following:

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.

7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. found some data entry dates that had ended on 2020-11-25 when the ride started on 2020-12-15. Having done this I have to swap the start time and end time around.
9. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
10. used find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

11. created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
12. Renaming the file by adding cleaned to it to denote that it

January 2021

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.
7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
9. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
Renaming the file by adding cleaned to it to denote that it

February 2021

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.

6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.
7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
9. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
Renaming the file by adding cleaned to it to denote that it

March 2021

1. calculated the total ride_length (=D2-C2, then formatted to time) and added it as a column to the right of the ride_end_time(column D). 2. added a column for weekday (=weekday(C2, 1) with the days put as a numerical value of 1 to 7.
3. used conditional formatting to help find the nulls in the data by assigning empty cells a colour.
4. filtered for cells in the ride_length 00:00:00 to 00:01:00 time and used the start_station_name, start_station_id, start_lat, start_lng data for the end_station_name, end_station_id, end_lat, end_lng to deduce the missing data points.
5. Noted that there were no obvious solutions for the other nulls, so I left them blank.
6. took note that there were inconsistencies with the data entry of start_lat, start_lng, end_lat and end_long that can not be easily rectified which would lead to errors during analysis.
7. Some of the test locations seem to be the same place entered under two different names. As I cannot confirm this I have left it as is.
8. filtered for 23 hours and 59 minutes in their previous data set via filtering and corrected them where necessary.
9. Use find and replace to:

- ‘&’ with
- ‘space’ between start and end stations names with underscores '_'
- I removed full stops '.', dashes '-' and parenthesis '(' ')'
- change capitalised letters to lowercase

created a new sheet titled Analysis where I calculated the Mean Average ride length, the max ride length and the mode of days of the week.
Renaming the file by adding cleaned to it to denote that it

RStudio

August 2020

I uploaded the August 2020 file into R. Once that was completed I went to import it into my environment and allowed all the packages that were required to do so to be installed and added to the library. Once this process was completed I began setting up the environment for data cleaning and analysis by running the following lines of code. I then set up the environment for data cleaning and analysis by installing and adding the following packages: the 'tidyverse' package for data cleaning and analysis; The 'lubridate' package for managing the date information and 'ggplot' for analysis purposes. The code for this is shown below.

~~~
```{install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")}```
~~~

~~~
```{library(tidyverse)
library(lubridate)
library(ggplot2)}```
~~~

These packages were installed and loaded into the library as they are useful for data cleaning and analysing data sets in R. Specifically the 'tidyverse' package is used primarily for data cleaning and analysis. The 'lubridate' package is used for managing the date information and 'ggplot' for analysis purposes.

I then made the same calculations I did in Google sheets for obtaining the ride_length and weekdays. I also noted that unlike in google sheets where 1 is Sunday; in R it is Monday. I did this so I could account for it later during my analysis of the combined datasets. The following lines of code denote the creation of the columns so they match those added in google sheets. I also noted the same issues I saw in the other data sets in regards to Nulls associated with location data. The following lines of code were adapted from a link given as part of the case study option 1 from the google data analytics professional certificate course (Divvy Exercise R Script)

~~~
```{X202008_divvy_tripdata <- mutate(X202008_divvy_tripdata, ride_time =(ended_at-started_at))

X202008_divvy_tripdata <-mutate(X202008_divvy_tripdata, weekday
=wday(X202008_divvy_tripdata$started_at, week_start = getOption("lubridate.week.start", 7)))
}```
~~~

Appendix 4: Analysis

RStudio

I had planned to run further analysis in the cloud version of R studio as I prefer the documentation and report creation features. I also wanted to combine all the data into one large data set to run the analysis. I wished to do this to easily process a large amount of data and reduce the risk of errors. So I uploaded all the cleaned data files into the cloud version of RStudio as my current laptop was having issues with installing a local version of the IED.

As I began to import each of the files into the environment to begin merging the data sets into one, the program became unstable. It continued to crash as I was using too much RAM due to the large size of the datasets in the small cloud environment. I got some help setting up a docker environment to get around the lack of RAM for processing. Unfortunately while running the code to analyse the datasets I ran into caching issues. This instability in the Docker environment led me to make the call to just use R to get the information I would need to match pivot tables I could create in Google Sheets. I am certain there was a way around these issues however a case study should be deliverable in a timely fashion hence my decision.

I used the following code to obtain the mean average ride_length of the entire month of August 2020, the mean average ride time of all member and casual riders for the month and finally, the mean averages of member and casual rides over the week where 1 is Monday and 7 are Sunday(the series of code listed with the exception of the percentile calsulations code was adapted from Divvy Exercise R Script):

```
'''r{mean(X202008_divvy_tripdata$ride_time)

X202008_divvy_tripdata %>%
group_by(member_casual) %>%
summarize(avg_value = mean(ride_time))

X202008_divvy_tripdata %>%
group_by(member_casual, weekday) %>%
summarize(avg_value = mean(ride_time))}'''
```

The next line of code is to get the total ride time for the month, the total member and casual rides and the member and casual rides over the days of the week

```
'''r{count(X202008_divvy_tripdata, vars = weekday, wt_var = member_casual)}'''
```

The next lot of code to obtain the median values for August 2020:

```
'''r{median(X202008_divvy_tripdata$ride_time)

X202008_divvy_tripdata %>%
group_by(member_casual) %>%
summarize(med_value = median(ride_time))

X202008_divvy_tripdata %>%
group_by(member_casual, weekday) %>%
summarize(med_value = median(ride_time))}'''
```

Finally, I used this last piece of code to get the 95th and 99th percentile values for all the ride duration data to see how long 95% and 99% of all the following ride categories are (the code was developed based on the code shown in the following tutorial URL: http://www.r-tutor.com/elementary-statistics/numerical-measures/percentile) :

```
'''r{quantile(X202008_divvy_tripdata$ride_time, c(.95, .99))

X202008_divvy_tripdata %>%
group_by(member_casual) %>%
summarize(perc_value = quantile(ride_time, c(.95, .99)))

percenticles <- X202008_divvy_tripdata %>%
group_by(member_casual, weekday) %>%
summarize(perc_value = quantile(ride_time, c(.95, .99)))
print(percentiles, n = nrow(percentiles))
}'''
```

I then went to add these to a summarised DIVVY_tripdata spreadsheet. I realised that the results from converting the seconds R data to hours, minutes, and seconds data with formatting were incorrect. To resolve this I used the following online seconds-to-time calculator:

https://www.inchcalculator.com/seconds-to-time-calculator/

It was a time-consuming process but the accuracy of the analysis of the data is more important. I had attempted a few different methods to convert the second's information to time in R. However I was running into errors and decided to use this method to ensure I had an accurate conversion.

Google Sheets

I created pivot tables concentrating on the average ride times, median ride times and total count of riders for:
each month
causal versus members
and casual versus member riders per day of the week

I then went to the Analysis sheet to work out the 95th and 99th percentile of the ride_legth data. I then combined these into one spreadsheet with the results I got in R from August 2020. I titled DIVVY_Tripdata_2020_2021_summarydata. I then imported this into RStudio.

RStudio

I uploaded and imported the DIVVY_Tripdata_2020_2021_summarydata file into the environment allowing for the packages to open and read the file to be installed and added to the library. I started by setting up the environment to conduct the analysis focusing on charts to better visualise and understand any trends. I started by installing and adding the RStudio library the ‘tidyverse’, ‘Lubridate’ and ‘ggplot’.

```
'''r{install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")}'''
```

```
'''r{library(tidyverse)
library(lubridate)
library(ggplot2)}'''
```

I then began making some bar graphs and scatter plots using the ggplot package but ran into issues of the data points being clustered into small areas making them difficult to read. I went to stack overflow and other places to look for a few solutions to adjust the x and y manually to see the data visualisation more clearly. These sadly didn’t pan out so I decided to go back to the

DIVVY_Tripdata_2020_2021_summarydata to see if I could create a better visualisation of the dataset there.

Google Sheets

I went back into the DIVVY_Tripdata_2020_2021_summarydata spreadsheet and opened 7 new sheets. I named the sheets grand_total_average_ride_time_2020_2021 and grand_total_riders_per_month_2020_2021, average_ride_time_2020_2021 and total_number_of_riders_per_day_2020_2021. I then reorganised the relevant data points into these.

I then turned the following sheets into their separate spreadsheets to create the charts:

- grand_total_average_ride_time_2020_2021
- grand_total_riders_per_month_2020_2021
- median_ride_time_ per_week_per_month_2020_2021(with the data separated into member and casual sheets)
- 95_and_99_perc_ride_time_per_month_2020_2021
- average_ride_time_2020_2021 (with the data separated into the member and casual sheets)
- total_number_of_riders_per_day_2020_2021 (with the data separated into member and casual sheets)

This was done so that I could change the names to ensure I had clear, easy-to-understand visuals and maintain the integrity of my analysis results. I added DIVVY_tripdata in front of these names to maintain naming conventions so it would be easier to track relevant files later if necessary. I then decided to use the insert chart function and created bar charts for all the datasets except for the 95_and_99_perc_ride_time_per_month_2020_2021 which I did as a line graph. I then customised each of these so they would be easy to understand at a glance and then added them to my report.