Business Tasks

- Problems
    1. How annual members and casual riders differ?
    2. Why casual riders would buy a membership?
    3. How digital media could affect their marketing tactics?
- Business Tasks
    1. Is maximizing the number of annual memberships the key to company's future success?
    2. Design marketing strategies aimed at converting casual riders into annual members.
- Stakeholders
    1. Cyclistic marketing analytics team
    2. Cyclistic executive team
    3. Lily Moreno

Data Collection

- Data source
  1. Data is from Motivate International Inc. (reliable, original, cited)
  2. License URL: https://www.divvybikes.com/data-license-agreement
  3. Data is public data.
- Original data organization
  1. CSV file
  2. Data is named by year and season or month
  3. 4 datasets for 2019 (4 seasons)
  4. 10 datasets for 2020 (1 seasons and 9 months)
  5. 10 datasets for 2021 (10 months) (current, not comprehensive)
- Issues and bias
  1. 2019 datasets: Do not contain latitude and longitude which means distance cannot be calculated. (Not comprehensive)
  2. 2019 datasets: Do not contain rideable_type only bikeid. (Not comprehensive)
  3. 2021 datasets: Covers from Jan to Oct only. (Not comprehensive)
- Licensing, privacy, security and accessibility
  1. License by Motivate International Inc (https://www.divvybikes.com/data-license-agreement).
  2. Personal information about birthday year and gender will be removed.
  3. Currently original data is stored only in local. Cleaned data will be stored both in local and BigQuery.
- Data integrity: Data integrity will be considered for all process. Currently only look at original data.
  1. Accuracy
     - All data was collected by computer so it should be accurate. However, inaccurate data will be cleaned later.
  2. Completeness
     - Some rows contain NULL which might affect calculation.
  3. Consistency
     - Column names are not consistent for 2019 and 2020, 2021.
     - Station IDs are not consistent for 2019 and 2020, 2021.
     - Rideable_type and bikeid is not consistent.
     - Type values are not consistent for 2019 and 2020, 2021.
  4. Trustworthiness
- How does the data help to answer the questions?
  1. Duration (needed to be calculated) vs Type
  2. Distance (needed to be calculated) (not comprehensive) vs Type
  3. Weekday (needed to be converted) vs Type
  4. Rideable type vs Type
- Sort and filter: Data will be sorted and filtered in the next stage (process)

Process

- Tools
  1. BigQuery: cleaning data
  2. Import data to BigQuery. Some datasets needed to be split to reduce file size.
- Cleaning data process
  1. Step 1: Cleaning 2019 datasets
     - Rename columns
     - CAST() changes data types
     - ROUND() round decimals to the second decimal
     - Change all values to NULL for rideable_type and distance_km
     - Change values (Subscriber, Customer) for member_casual to match 2020 and 2021 (member, casual)
     - Repeat the process to clean 2019 Q1 to Q4.

```
SELECT
    # Rename trip_id to ride_id
    CAST(trip_id AS STRING) AS ride_id,
    # Change all values under bikeid to NULL and change data type to STRING, then rename
    CAST(CASE
            WHEN bikeid is not NULL THEN NULL ELSE NULL
        END
        AS STRING) AS rideable_type,
    # Change start_time and end_time data type to DATETIME
    CAST(start_time AS DATETIME) AS start_time_local,
    CAST(end_time AS DATETIME) AS end_time_local,
    # Get the weekday
    CAST(FORMAT_DATE('%u', start_time) AS INT64) AS day_of_week,
    # Transfer duration unit from seconds to minute
    CAST(ROUND(tripduration/60, 0) AS INT64) AS trip_duration_minute,
    # Rename from_station_name and to_station_name
    from_station_name AS start_station_name,
    to_station_name AS end_station_name,
    # Set all values of distance_km to NULL
    CAST(CASE
        WHEN birthyear is not NULL THEN NULL ELSE NULL
    END AS FLOAT64) AS distance_km,
    # Rename usertype, change values into member and casual
    CASE
        WHEN usertype='Subscriber' THEN 'member'
        WHEN usertype='Customer' THEN 'casual'
        ELSE usertype
    END AS member_casual,
    # Get year
    EXTRACT(YEAR FROM start_time) AS year
FROM `coursera-case-study20211109.Case_study1.2019_Q1`
```

  2. Step 2: TRIM() and ROUND()
     - TRIM() and ROUND() 2020 Q1, other 9 months datasets and all 2021 datasets
     - TRIM() 2019 Q1 ~ Q4 datasets

```
# TRIM and ROUND
SELECT
    TRIM(ride_id) AS ride_id,
    TRIM(rideable_type) AS rideable_type,
    started_at,
    ended_at,
    TRIM(start_station_name) AS start_station_name,
    TRIM(end_station_name) AS end_station_name,
    TRIM(member_casual) AS member_casual,
    ROUND(start_lat, 2) AS start_lat,
    ROUND(start_lng, 2) AS start_lng,
    ROUND(end_lat, 2) AS end_lat,
    ROUND(end_lng, 2) AS end_lng
FROM `coursera-case-study20211109.Case_study1.divvy_trips`
```

  3. Step 3: Combine all datasets for 2020 and 2021 using UNION ALL

4. Step 4: Calculate distance_km for 2020 and 2021 datasets
   - Calculate distance_km for 2020 and 2021

```
SELECT
    *,
    DATE_DIFF(td.ended_at, td.started_at, MINUTE) AS trips_duration_minute,
    ROUND(ST_DISTANCE(ST_GEOGPOINT(td.start_lng, td.start_lat), ST_GEOGPOINT(td.end_lng, td.end_lat))/1000 , 2) AS distance_km,
FROM
    `coursera-case-study20211109.Case_study1.divvy_trips` AS td
```

   - Remove start_lat, start_lng, end_lat and end_lng for 2020 and 2021 datasets
5. Step 5: Switch started_at and ended_at for 2020
   - Some ended_at values are before started_at for 2020. So switch them by using CASE WHEN ELSE END.

```
SELECT
    ride_id,
    rideable_type,
    CASE
        WHEN start_time_temp > end_time_temp THEN end_time_temp ELSE start_time_temp
    END AS start_time_local,
    CASE
        WHEN end_time_temp < start_time_temp THEN start_time_temp ELSE end_time_temp
    END AS end_time_local,
    day_of_week,
    trips_duration_minute,
    start_station_name,
    end_station_name,
    distance_km,
    member_casual
FROM `coursera-case-study20211109.Case_study1.temp`
```

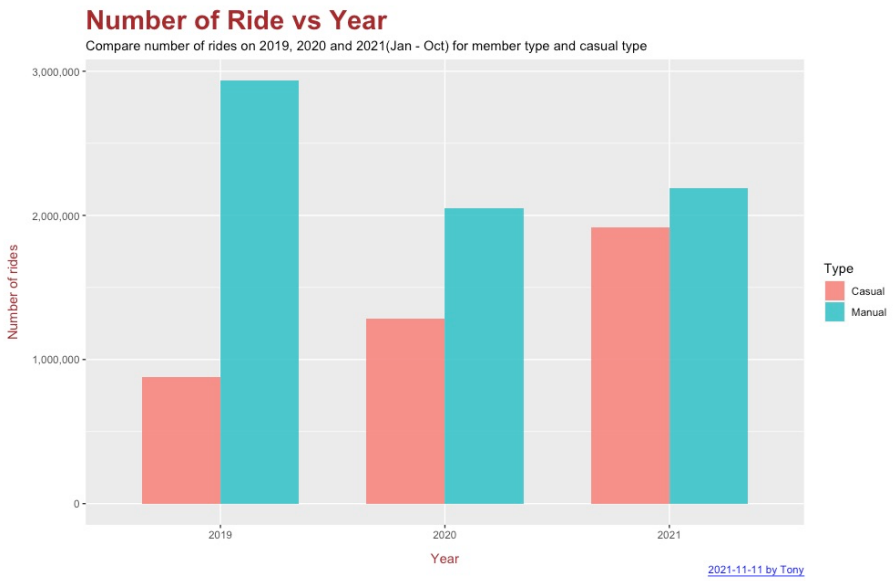
6. Step 6: Calculate duration for 2020 and 2021
   - Use DATE_DIFF function to calculate duration

```
SELECT
    ride_id,
    rideable_type,
    start_time_local,
    end_time_local,
    day_of_week,
    DATE_DIFF(end_time_local, start_time_local, MINUTE) AS trips_duration_minute,
    start_station_name,
    end_station_name,
    distance_km,
    member_casual
FROM `coursera-case-study20211109.Case_study1.temp2`
ORDER BY
    trips_duration_minute
```

- Verify data
  1. Verify data to check whether it is clean and ready to analyze
     - Use LENGTH() to check start_time_local and end_time_local
     - Use GROUP BY to check if there is any bad data about rideable type
     - Use GROUP BY to check if there is any bad data about weekday
     - Use ORDER BY to check if there is any duration lower than and equal to 0.
     - Use WHERE to check if there is any rows' end_time_local is before start_time_local.
     - Use WHERE to check if there is any distance_km is lower than 0.
  2. Verify data and cleaning data process are always be conducted during the whole data analysis process.
- Data transformation
  1. After data is cleaned by using BigQuery, data is stored into Google Drive in csv file.
  2. Download csv file from Google Drive into local.
  3. Import csv file to R and use str() function to simply check data completeness.

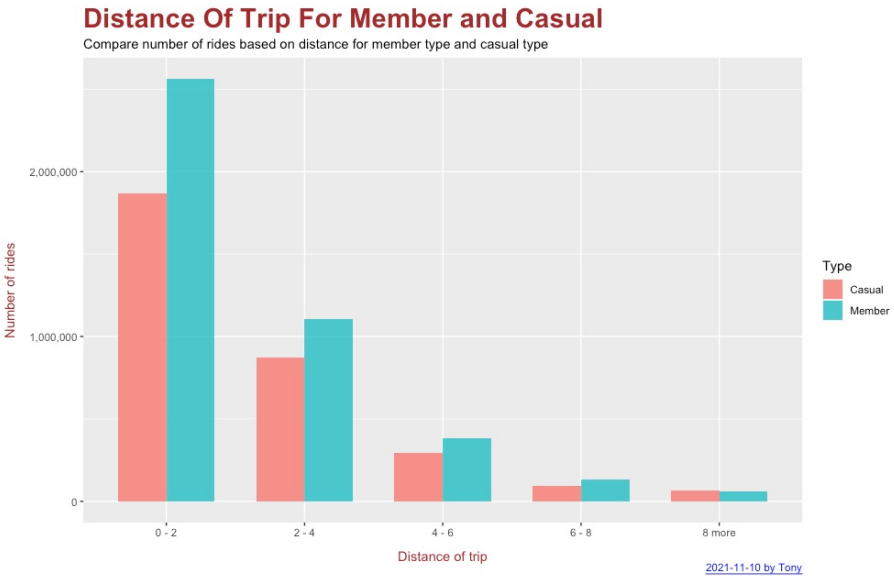
Analyze

- Tools
  1. BigQuery to calculate, filter, sort and find trends initially.
  2. After using BigQuery to initially explore data, use R to analyze calculate and create data visualization.
  3. Excel ???
- Format data
  1. BigQuery
     - All columns are properly formatted including data type and values.
     - rideable_id and ride_id are STRING with some missing value indicated as NULL.
     - start_time_local and end_time_locat are DATETIME.
     - distance_km is FLOAT.
     - day_of_week, year and trips_duration_minute are INTEGER.
  2. R
     - Column named member_casual is renamed as type. Other columns are the same as BigQuery.
     - day_of_week data type is changed to CHAR. Other data types are the same as BigQuery.
- Calculations
  1. All calculations are saved in BigQuery and R.
  2. BigQuery local file: google_case_study1_sql_code.sql
  3. R local file: divvy_trips_code.R
- Trends and data visualization
  1. Year vs type: causal increases over time and member decreases over time.



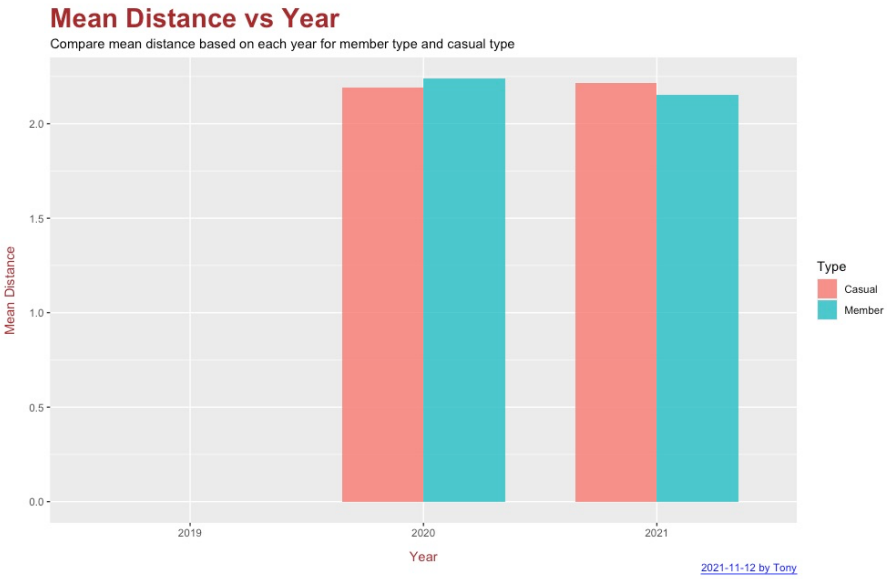
User Type and Year	Number of ride
Casual	4075951
2019	880637
2020	1281328
2021	1913986
Member	7178941
2019	2937367
2020	2052821
2021	2188753
Total	11254892

2. Distance vs type: both decreases over distance



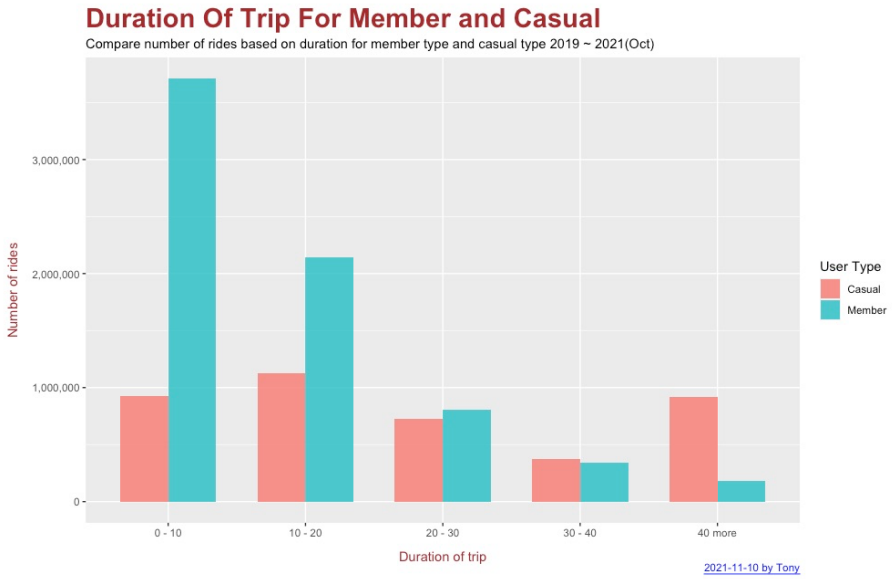
User Type / Distance	0 ~ 2 km	2 ~ 4 km	4 ~ 6 km	6 ~ 8 km	8 km more	Total
Casual	1866482	873042	295328	93851	66611	3195314
Member	2564788	1103569	380275	131008	61934	4241574

3. Mean distance vs type by year: no significate changes



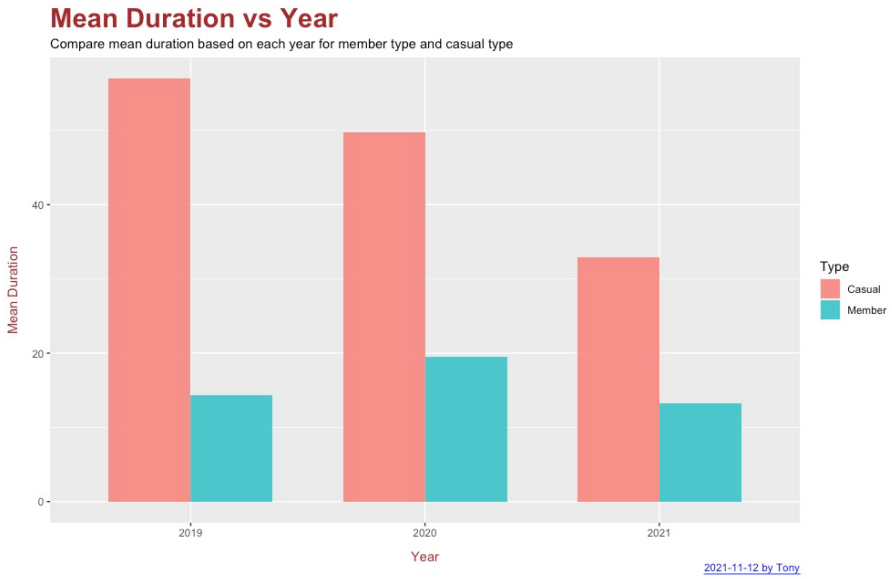
User Type and Year	Mean Distance (KM)
Casual	
2019	N/A
2020	2.19
2021	2.22
Member	
2019	N/A
2020	2.24
2021	2.15

4. Duration vs type: member decreases over duration, casual increases, decreases then increases again (WHY?).



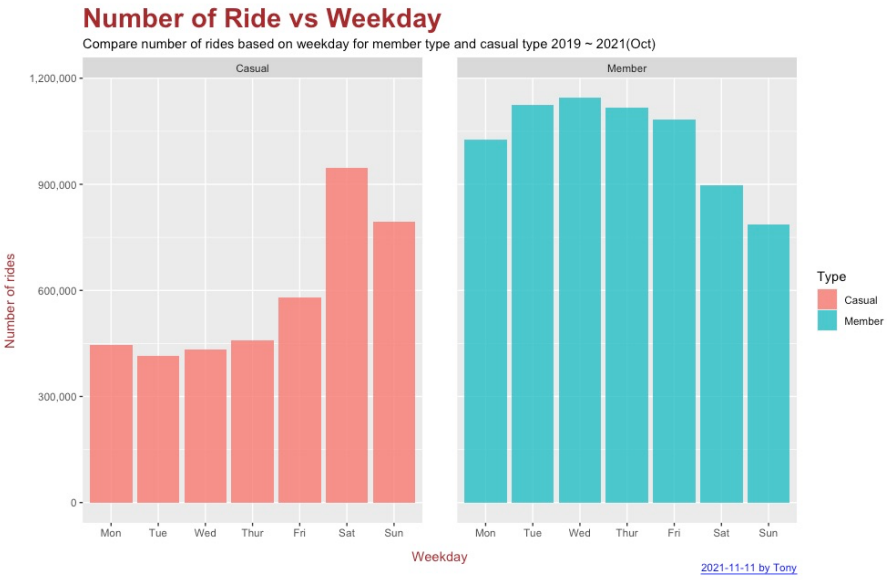
User Type / Duration	0 ~ 10 min	10 ~ 20 min	20 ~ 30 min	30 ~ 40 min	40 min more	total
Casual	924688	1129420	727114	373878	920851	4075951
Member	3712620	2139374	809044	338100	179803	7178941

5. Mean duration vs type: casual decreases, member increases then decreases (WHY?)
   - Number of casual rides increases but why mean duration decreases?
   - Number of member rides decreases from 2019 to 2020 but why mean duration increases?



User Type and Year	Mean Duration (Minute)
Casual	
2019	57.02
2020	49.8
2021	32.89
Member	
2019	14.33
2020	19.46
2021	13.24

6. Weekday vs type: most interesting one. Causal increases when nearly end of week. Member decreases nearly end of week (WHY?).



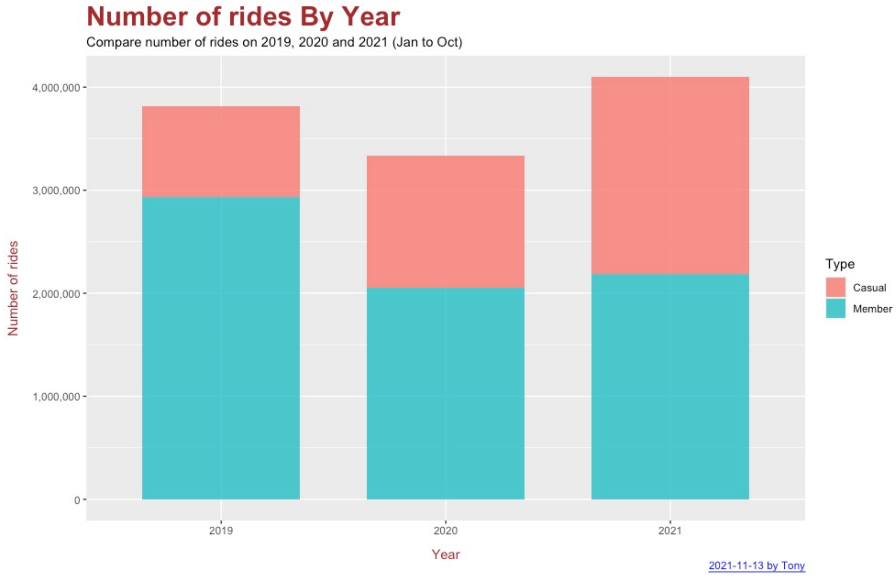
User Type / Weekday	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Casual	445254	415341	432156	459722	580809	947291	795378
Member	1026350	1123916	1144274	1117698	1082610	897060	787033

- Analysis
  1. Membership
     - $9/month
     - First 45 min free. After 45 min, $0.15/min
     - Unlimited times.
  2. Casual
     - Single ride
       - $3.3/trip
       - First 30 min free. After 30 min, $0.15/min.

- Full day pass
  - $15/day
  - First 3h min free. After 3h min, $ 0.15/min.
  - Unlimited times.

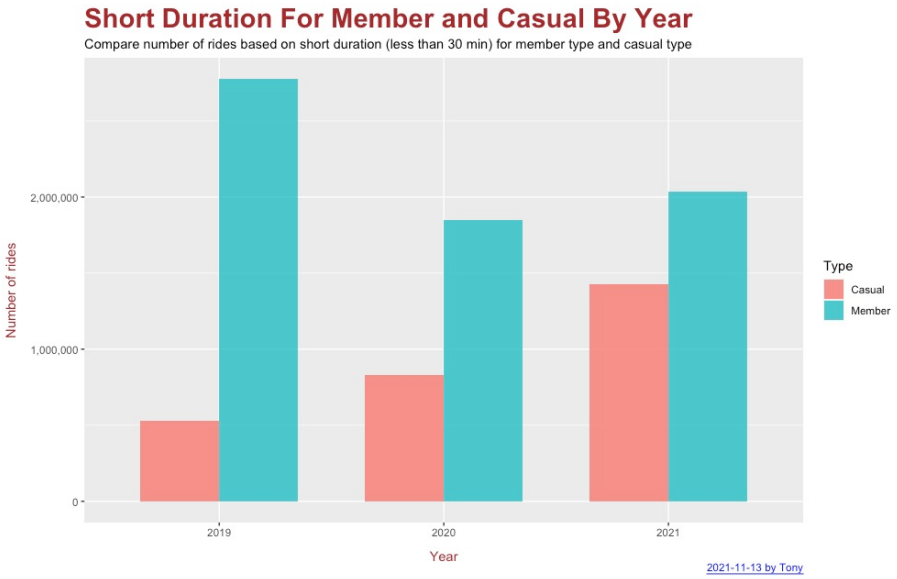
3. Users' habits of using bike has been changing
   - As we can see number of ride less than 30 min decrease then increase again from 2019 to 2020. And note than the data of 2021 only covers from Jan to Oct.
     - Decrease in 2020 from 2019 because of COVID-19. Movement inside the city decreased due to locked down or strict epidemic prevention policies causing the use of bike decreases. From number 6 above, we can conclude that most members are commutors for works or schools.

Year	Total
2019	3818004
2020	3334149
2021	4102739



     - However, surprisingly, although number of rides and total number of rides less than 30 min decreases from 2019 to 2020, the number of rides for casual and number of rides less than 30 min for casual increases.

Year	Number of ride less than 30 min
2019	3305914
2020	2677056
2021 (Jan to Oct)	3459290



- Then in 2021, casual rides has higher increase rates than member rides.
- Less movement changes people's life style. The change of life style changes the way people use bikes. People don't want to pay monthly fees because they don't ride bikes as often as before. They just want to pay for the trips or time they ride.
- Let's say people rides bikes less than 3 times a month, which costs $9.9 the most and $0 the least. The cost for member per month is $9. This is probably the reason why people change from member to casual. However, if we reduce monthly fee for member to $6 or $6.6. We might be able receive more member.

Share

- Tools
    1. Excel: build readable tables from the results queried from BigQuery
    2. R: create data visualization charts
    3. Powerpoint: tell stories and create reports
- Findings and stories
    1. Findings: the number of rides and durations decreases on 2020 and increase again on 2021. However, casual riders increase more than member riders.
    2. Stories:
        - Why does this happen? Because COVID-19 changes our life styles. Data and visualization can support this theory.
        - We can use social media, internet and TV to advertise instead of physical advertisement.
- Conclusion:
    1. COVID-19 changes our life style. Our life style changes the way people use bikes.
    2. Reduce prices for member monthly fee to around $6/month to $6.6/month.
    3. Social media advertisement is more efficient than physical advertisement.
- Audience:
    1. Cyclistic marketing analytics team (in person or email)
    2. Cyclistic executive team (in person PPT)
    3. Lily Moreno (in person PPT)