Case Study: How Does a Bike-Share Navigate Speedy Success?

Background

Scenario: You are a junior data analyst 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, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Studi kasus ini merupakan bagian dari Google Data Analytics Specialization pada platform Coursera. Dalam mengerjakan studi kasus ini, saya menggunakan framework analisis data yang telah diajarkan pada program sertifikasi teresebut, dimana proses data analysis akan dibagi menjadi 6 tahapan, yaitu Ask, Prepare, Process, Analyze, Share, Act.

Report

Tools: PostgreSQL and JupyterLab (R Programming)

Business Task

Mengidentifikasi perbedaan dari casual riders dan annual members dalam menggunakan sepeda Cyclistic. Casual riders merupakan pelanggan yang membeli single-ride atau full-day pass di Cyclistic, sedangkan annual members merupakan pelanggan yang membeli annual memberships di Cyclistic.

Data Description

Saya akan menggunakan data perjalanan dari bulan Juni tahun 2022. Data ini disusun dalam format tabel dan disimpan menggunakan format csv yang kemudian akan diimport pada database menggunakan postgreSQL. Data yang digunakan dapat diakses pada link berikut, 202206-divvy-tripdata.zip.

Data ini berisikan informasi yang direkam dari setiap perjalanan yang terjadi pada penggunaan sepeda Cyclistic mulai dari tanggal 1 Juni 2022 hingga 31 Juni 2022. Informasi ini mencakup:

1) *Ride ID*: unique identifier yang disematkan pada setiap perjalanan yang terjadi
2) *Rideable type*: jenis sepeda yang digunakan
3) *Started at*: waktu dan tanggal perjalanan dimulai

4) *Ended at*: waktu dan tanggal perjalanan selesai
5) *Start station name:* nama stasiun keberangkatan
6) *Start station ID:* ID stasiun keberangkatan
7) *End station name*: nama stasiun akhir
8) *End station ID:* ID stasiun akhir
9) *Start latitude*: nilai latitude keberangkatan
10) *Start longitude*: nilai longitude keberangkatan
11) *End latitude*: nilai latitude akhir
12) *End longitude*: nilai longitude akhir

Untuk kredibilitas dari data ini, saya akan menggunakan sistem “ROCCC” untuk menilai hal tersebut:

- **Reliability:** Data ini dapat diandalkan karena jumlah data yang cukup besar
- **Original:** Data ini original karena berasal dari primary source dimana perusahaan mengumpulkan data secara langsung.
- **Comprehensive:** Data ini cukup komprehensif dan mencakup fitur-fitur yang cukup banyak untuk proses analisis.
- **Current:** Data ini masih sangat relevan dengan kondisi saat ini, mengingat data berasal dari bulan Juni tahun 2022.
- **Cited**: Data diketahui berasal dari sistem internal perusahaan.

Berdasarkan analisa ROCCC diatas, dapat disimpulkan bahwa data ini sangat kredibel untuk memberikan insight yang valid bagi perusahaan.

Data Cleaning Documentation

Saya menggunakan postgreSQL untuk melakukan data cleaning dan data manipulation pada data. Beberapa tindakan yang saya lakukan adalah:

a. Melakukan sort pada table berdasarkan kolom started_at dan menyimpan hasil query pada tabel trip_data

```
SELECT *
INTO trip_data
FROM tripdata_jun_2022
ORDER BY started_at ;
```

Saya melakukan sort pada dataset awal dan kemudian hasilnya disimpan pada tabel baru bernama trip_data. Hal ini dilakukan agar data pada tabel lebih terorganisir dan mudah dibaca.

b. Menghapus 4 kolom titik koordinat pada tabel trip_data

```
ALTER TABLE trip_data
    DROP COLUMN start_lat,
    DROP COLUMN start_lng,
    DROP COLUMN end_lat,
    DROP COLUMN end_lng;
```

Saya memutuskan untuk menghapus 4 kolom titik kordinat pada tabel trip_data dikarenakan variabel ini tidak relevan untuk proses analisis nantinya. Saya baru akan menggunakan variabel ini pada saat visualisasi data.

c. Memeriksa apakah ada duplikasi data yang terjadi pada tabel trip_data

```
SELECT ride_id, count(ride_id)
FROM trip_data td
GROUP BY ride_id
HAVING count(ride_id) > 1
```

Tidak ditemukan adanya data duplikat pada tabel ini.

d. Memeriksa apakah ada data yang berada diluar cakupan waktu

```
SELECT min(started_at), max(started_at)
FROM trip_data td
```

Data yang ada sepenuhnya hanya mencakup perjalanan yang dimulai pada bulan 6 tahun 2022, lebih tepatnya antara 1 Juni 2022 pukul 00.00 hingga 30 Juni 2022 pukul 23.59.

e. Memeriksa apakah ada nilai yang salah pada kolom rideable_type

```
SELECT DISTINCT rideable_type
FROM trip_data td
```

Tidak ditemukan adanya kesalahan nilai pada kolom ini. Nilai yang ada hanya mencakup docked bike, classic bike, dan electric bike.

f. Memeriksa apakah ada nilai yang salah pada kolom member_casual

```
SELECT DISTINCT member_casual
FROM trip_data td
```

Tidak ditemukan adanya kesalahan nilai pada kolom member_casual. Nilai yang ada hanya mencakup member dan casual, sesuai dengan jenis pengguna.

g. Menghitung jumlah missing values pada masing-masing kolom

```
SELECT
    sum(CASE WHEN ride_id IS NULL THEN 1 ELSE 0 END) AS ride_id,
    sum(CASE WHEN rideable_type IS NULL THEN 1 ELSE 0 END) AS rideable_type,
    sum(CASE WHEN started_at IS NULL THEN 1 ELSE 0 END) AS started_at,
    sum(CASE WHEN ended_at IS NULL THEN 1 ELSE 0 END) AS ended_at,
    sum(CASE WHEN start_station_name IS NULL THEN 1 ELSE 0 END) AS start_station_name ,
    sum(CASE WHEN start_station_id IS NULL THEN 1 ELSE 0 END) AS start_station_id ,
    sum(CASE WHEN end_station_name IS NULL THEN 1 ELSE 0 END) AS end_station_name ,
    sum(CASE WHEN end_station_id IS NULL THEN 1 ELSE 0 END) AS end_station_id ,
    sum(CASE WHEN member_casual IS NULL THEN 1 ELSE 0 END) AS member_casual
FROM trip_data td ;
```

Terdapat banyak missing values pada kolom start_station_name, start_station_id, end_station_name, dan end_station_id. Untuk mengatasi hal tersebut saya akan mengisi missing values dengan catatan “unrecorded”, yang menandakan tidak tercatatnya nilai pada kolom tersebut pada saat pemakaian sepeda.

h. Mengisi missing values pada 4 kolom dengan catatan 'unrecorded'

```
UPDATE trip_data
SET
    start_station_name = 'unrecorded',
    start_station_id = 'unrecorded',
    end_station_name = 'unrecorded',
    end_station_id = 'unrecorded'
WHERE (start_station_name ISNULL AND start_station_id ISNULL) OR (end_station_name IS NULL AND end_station_id ISNULL);
```

Disini saya mengisi missing values dengan nilai “unrecorded” seperti yang telah saya jelaskan sebelumnya.

i. Menambahkan kolom lama penggunaan, hari waktu penggunaan, dan rute penggunaaan sepeda yang kemudian disimpan pada tabel trip_data_update

```
SELECT *,
    ended_at - started_at AS ride_length,
    EXTRACT("isodow" FROM started_at) AS day_of_week,
    concat(start_station_name, ' to ', end_station_name) AS track
INTO trip_data_updated
FROM trip_data td;
```

- **ride_length,** berisikan data lama perjalanan saat penggunaan sepeda
- **day_of_week,** berisikan data hari dimana perjalanan dimulai
- **track,** berisikan rute perjalanan yang ditempuh yang merupakan kombinasi dari tempat keberangkatan dan tempat berhenti.

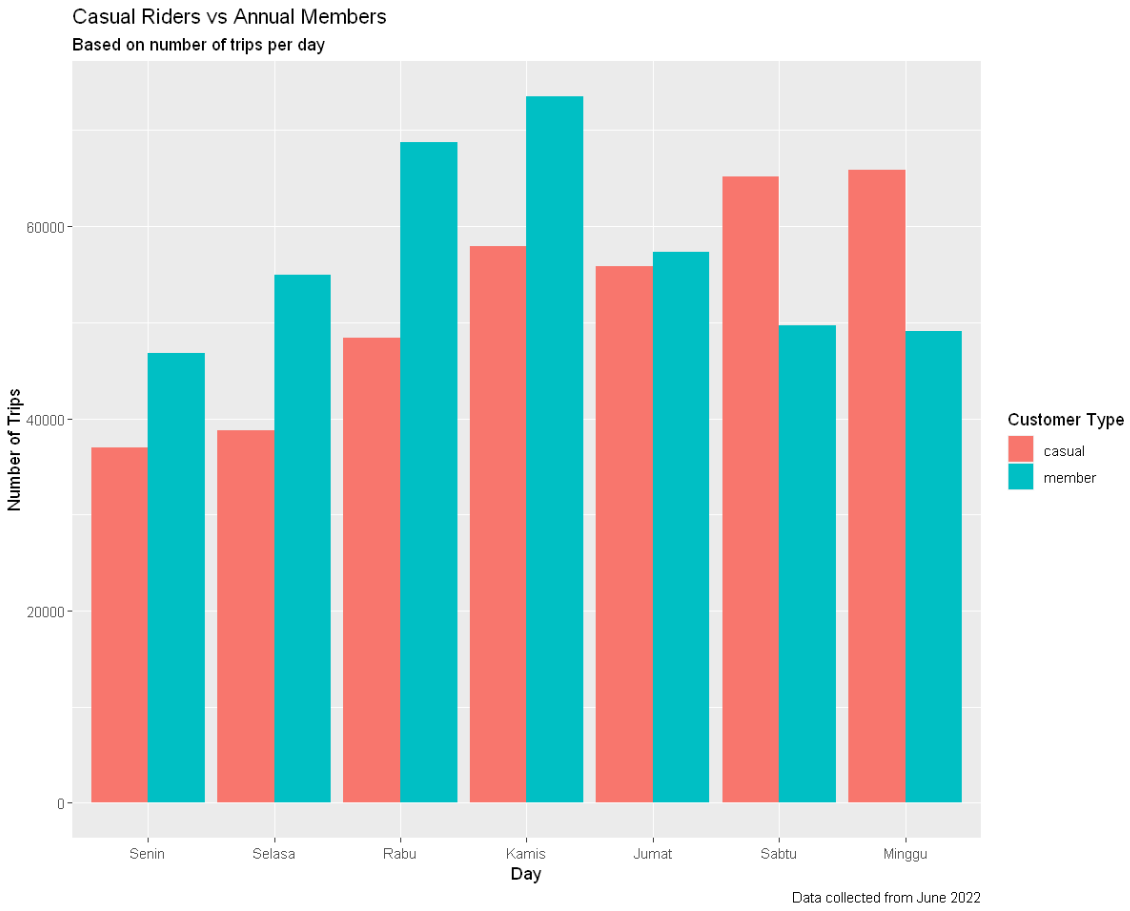
Hasil query tersebut juga saya masukkan ke dalam tabel baru yang bernama trip_data_update untuk mempermudah akses.

j. Mengekspor data menjadi csv untuk proses analisis menggunakan R

Findings

Berdasarkan hasil transformasi dataset yang telah saya buat, ada 3 fitur yang dapat dijadikan perbedaan karakteristik dari casual riders dan annual member dalam menggunakan sepeda Cyclistic, yaitu jumlah perjalanan per hari, rata-rata waktu perjanalan per hari, dan rute populer.

1) Jumlah perjalanan per hari



Berdasarkan hasil visualisasi di atas, ada beberapa insight yang saya dapatkan, yaitu:

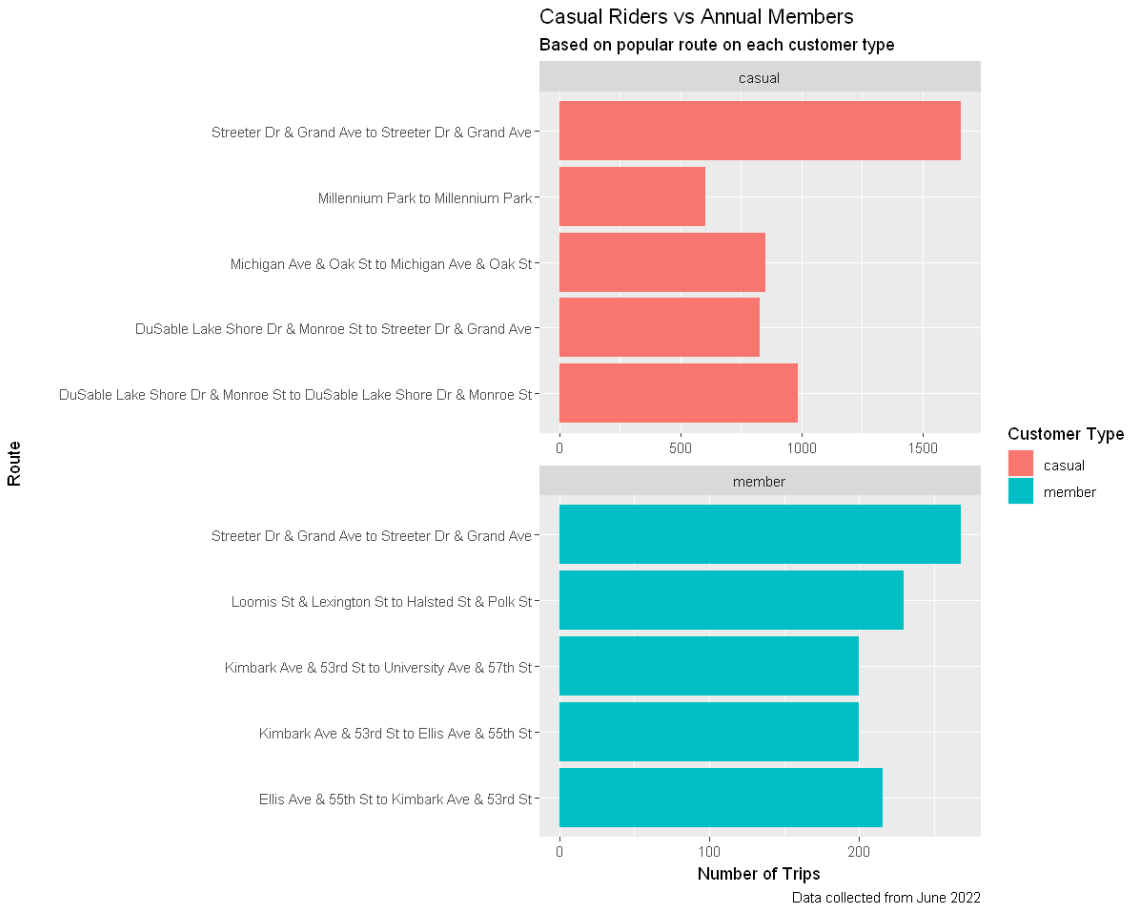
- Jumlah perjalanan pada jenis pengguna annual members lebih banyak dibandingkan dengan casual riders pada hari Senin sampai Jumat, namun lebih sedikit pada pada hari Sabtu dan Minggu
- Jumlah perjalanan pada jenis pengguna annual members meningkat mendekati tengah pekan dengan puncak jumlah penggunaan pada hari Kamis dan menurun mendekati akhir pekan mulai dari hari Jumat sampai Minggu
- Jumlah perjalanan pada jenis pengguna casual riders memuncak pada akhir pekan dengan jumlah penggunaan terbanyak yaitu pada hari minggu dengan sedikit penurunan terjadi pada hari Jumat

2) Rata-rata waktu perjalanan per hari

Berdasarkan hasil visualisasi di atas, ada beberapa insight yang saya dapatkan, yaitu:

- Rata-rata waktu perjalanan pada jenis pengguna casual riders jauh lebih lama dibandingkan annual members setiap harinya
- Rata-rata waktu perjalanan pada jenis pengguna casual riders berada di antara 30-35 menit setiap harinya, dengan pengecualian pada hari Selasa yang lebih sedikit dari 30 menit dan hari Minggu yang lebih banyak dari 35 menit
- Rata-rata waktu perjalanan pada jenis pengguna annual members berada di antara 10-15 menit setiap harinya, dengan pengecualian pada hari Sabtu dan Minggu yang melebihi 15 menit.

3) Rute perjalanan populer



Berdasarkan hasil visualisasi di atas, ada beberapa insight yang saya dapatkan, yaitu:

- Empat dari 5 rute populer pada pengguna tipe casual riders merupakan rute loop, dimana stasiun awal sama dengan stasiun akhir. Disini dapat disimpulkan bahwa kebanyakan dari pengguna tipe casual riders tidak menggunakan sepeda Cyclistic sebagai sarana transportasi untuk kegiatan sehari-hari, melainkan mungkin sebagai sarana olahraga atau relaksasi.
- Empat dari 5 rute populer pada pengguna tipe annual members merupakan rute linier, dimana stasiun awal bukan merupakan stasiun akhir. Disini dapat disimpulkan bahwa kebanyakan dari pengguna tipe annual members menggunakan sepeda Cyclistic sebagai sarana transportasi untuk kegiatan sehari-hari.

Reccomendations

Berdasarkan insight yang didapatkan dari proses analisis diatas, ada 3 rekomendasi yang dapat saya berikan kepada para stakehorlder. Rekomendasi saya adalah sebagai berikut:

a) Melakulan targeted advertising terhadap segmen pengguna tipe casual untuk konversi menjadi annual members dengan kriteria sebagai berikut:

    - Rutin menggunakan sepeda Cyclistic pada hari Senin sampai Jumat dengan minimum penggunaan selama 3 hari dari 5 hari tersebut
    - Rata-rata waktu penggunaan berada jauh di bawah 30 menit dengan rute yang dilalui merupakan rute linier, dimana stasiun awal bukan merupakan stasiun akhir
    - Jarang melakukan perjalanan lebih dari 30 menit dan melalui rute loop.
2) Melakukan mass marketing di sekitar stasiun yang kurang populer. Hal ini dapat dilakukan dengan memasang billboard, poster, atau pamflet.
3) Melakukan marketing campaign mengenai hidup sehat dengan sepeda untuk meningkatkan jumlah pengguna tipe casual riders yang menggunakan sepeda Cyclistic sebagai sarana transportasi sehari-hari.

Diluar dari business task pada proyek analisis ini, saya juga ingin merekomendasikan agar dilaksanan pemeriksaan terhadap perangkat GPS yang ada pada sepeda Cyclistic. Hal ini dikarenakan banyaknya data stasiun awal dan stasiun akhir yang tidak terekam. Ada 92994 nama dan id stasiun awal dan 100152 nama dan id stasiun akhir yang tidak terekam. Oleh karena itu, saya ingin merekomendasikan hal ini untuk memperbaiki kualitas analisis yang akan dilakukan kedepannya.