Cyclistic Analysis

Sharon Makunura

2022-03-24

Introduction

This report summarizes the analysis process I followed in completing my capstone project for the Google Data Analytics Certification. For this project, I chose track one which analyzed data for Cyclistic Bikeshare.

1. Ask

The goal of my analysis is to identify patterns and trends that are unique to casual riders, and differentiate them from annual members. The analysis will help inform a marketing strategy that targets casual riders to convert them to annual members. The results will be communicated to the marketing executives of Cyclistic.

2: Prepare

A reliable public dataset was used to ensure that the data provided was good data. 9 CSV files were downloaded to compile the data for the year 2020. I scanned through the files in Excel to confirm each contained the same number of variables.

I then uploaded the data into RStudio Desktop, which . I installed four libraries for use: readr, tidyverse,dplyr and tidyr. I loaded the 9 csv files and saved them as dataframe objects.

NB: To optimize performance the R code for all analysis and output shown here is available for download from GitHub

3: Process

I began the process by using glimpse function to look through the datasets.

All the datasets contain the same number of variables. Next, I combined some of the dataframes to compress the number of dataframes I have to look through, without compromising performance. I combined dataframes for the following periods:

Dataframe 1	+ Dataframe 2
Q1	Apr
May	June
Nov	Dec

This resulted in 7 dataframes to work with. The next step was to check for duplicates. By using the duplicated function I was able to determine there were no duplicated records in the datasets.

The next step was to clean the data. Because of the number of dataframes and the potential for repeating code, I created custom functions to perform the cleaning operations. The functions performed the following operations:

- Replace missing values: some rows were missing data for end stations. Working on the assumption that these were round trips, the missing data was coalesced with the start station data.
- Format the started_at and ended_at columns as datetime objects.
- Use the formatted started_at columns to compute and create new columns for day of week and month
- Calculate trip duration by subtracting starting time from ending time.
- Delete all records with duration of less than 0 minutes.
- Delete all coordinates from the data frames.

NB: A comprehensive cleaning log is also available from GitHub

When completed the structure of the dataframes had the following structure:

```
Observations: 622,191
Variables: 12
$ ride_id            <fct> 322BD23D287743ED, 2A3AEF1AB9054D8B, 67DC1D133E8B5816, C79...
$ rideable_type      <fct> docked_bike, electric_bike, electric_bike, electric_bike,...
$ started_at         <dttm> 2020-08-20 18:08:14, 2020-08-27 18:46:04, 2020-08-26 19:...
$ ended_at           <dttm> 2020-08-20 18:17:51, 2020-08-27 19:54:51, 2020-08-26 21:...
$ start_station_name <fct> Lake Shore Dr & Diversey Pkwy, Michigan Ave & 14th St, Co...
$ start_station_id   <int> 329, 168, 195, 81, 658, 658, 196, 67, 153, 177, 313, 71, ...
$ member_casual      <fct> member, casual, casual, casual, casual, casual, casual, c...
$ end_station_name   <fct> Clark St & Lincoln Ave, Michigan Ave & 14th St, State St ...
$ end_station_id     <int> 141, 168, 44, 47, 658, 658, 49, 229, 225, 305, 296, 283, ...
$ duration_mins      <dbl> 10, 69, 129, 48, 11, 41, 20, 13, 15, 28, 8, 10, 22, 9, 54...
$ start_day          <chr> "Thu", "Thu", "Wed", "Thu", "Thu", "Thu", "Wed", "Wed", "...
$ start_month        <chr> "Aug", "Aug", "Aug", "Aug", "Aug", "Aug", "Aug", "Aug", "...
>
```

Data frame structure after cleaning

Finally I combined all the dataframes into one.

4: Analyze

I began the analysis by calculating the average and max length of all the rides using the calculated field duration. I also created a function to calculate the most common day, which turned out to be Saturday. The most frequent month was August.

Next, I created variations of the above analysis by organizing the average and max length of rides by membership type, day and month.

	member_casual	mean_duration	max_duration
	<fct>	<dbl>	<dbl>
1	casual	47.3	156450
2	member	15.7	93794

Ave and Max by membership type

	start_day	mean_duration	max_duration
	<chr>	<dbl>	<dbl>
1	Fri	26.9	117323
2	Mon	25.2	93794
3	Sat	33.2	79218
4	Sun	35.6	143937
5	Thu	25.2	156450
6	Tue	23.0	69505
7	Wed	23.5	74703

Ave and Max by day

	start_month	mean_duration	max_duration
	<chr>	<dbl>	<dbl>
1	Apr	35.9	58720
2	Aug	29.8	40846
3	Dec	16.0	9741
4	Feb	23.4	143937
5	Jan	19.3	156450
6	Jul	38.3	49965
7	Jun	33.5	41271
8	Mar	23.7	93794
9	May	33.4	28897
10	Nov	19.7	35934
11	Oct	20.0	35724
12	Sep	25.3	54283

Ave and Max by month

Next, I looked at the actual number of rides, grouping them by membership type, day of week, and month of the year.

	start_month	n
1	Apr	84768
2	Aug	622191
3	Dec	131179
4	Feb	139585
5	Jan	143884
6	Jul	551271
7	Jun	342980
8	Mar	143415
9	May	200262
10	Nov	259538
11	Oct	387858
12	Sep	532808

Number of rides by month

	member_casual	start_month	n
	<fct>	<chr>	<int>
1	casual	Apr	23627
2	member	Apr	61141
3	casual	Aug	289599
4	member	Aug	332592
5	casual	Dec	30001
6	member	Dec	101178
7	casual	Feb	12870
8	member	Feb	126715
9	casual	Jan	7785
10	member	Jan	136099

... with 14 more rows

Number of rides by month and member type

	start_day	n
1	Fri	527670
2	Mon	433565
3	Sat	625966
4	Sun	527045
5	Thu	491193
6	Tue	452461
7	Wed	481839

rides by day

	member_casual	start_day	n
	<fct>	<chr>	<int>
1	casual	Fri	201998
2	member	Fri	325672
3	casual	Mon	141603
4	member	Mon	291962
5	casual	Sat	313734
6	member	Sat	312232
7	casual	Sun	256080
8	member	Sun	270965
9	casual	Thu	162671
10	member	Thu	328522
11	casual	Tue	137441
12	member	Tue	315020
13	casual	Wed	152498
14	member	Wed	329341

Number of rides by day and member type

I then split the dataframe into two by member type. The casual riders are the fewer with around 1,3 million records to the 2,2 million records of members. By running the mode function again I determined that they shared the same frequent month. However the most common day for casuals is Saturday, while for members it is Wednesday.

I also looked at the different rideable types to determine if there was any distinct pattern, which surprisingly showed none.

Finally, I merged my summaries logically to create two dataframes that I exported to CSV files. I also exported my formatted complete data set in CSV format.

5: Share

The results suggest that the most significant aspect of casual riders is the trend towards longer bike rides. The data implies casual riders ride longer, despite taking fewer and more seasonal rides overall. This is significant for the business if membership is predicated on the number of rides, or if membership rewards have traditionally focused on number of rides.

I needed to share my findings with the Cyclistic, who are detail oriented and sophisticated. I decided to created my visualizations in Tableau using the combined dataset I exported from R. I also created some summary tables in Excel. Finally, I downloaded all my Tableau sheets as a PowerPoint presentation, which I then edited to create a presentation for the executives. I then created a separate report detailing my entire analysis process for the marketing analytics team at Cyclistic using RMarkdown and Word.

6: Act

The aim in this final process was to ensure that I shared the results of my analysis. My conclusion was that a marketing strategy that aims to convert casual riders to members should appeal to seasonal/weekend riders. It should also have benefits that accrue from longer rides rather than more frequent rides. Lastly, I noted that there is opportunity for further analysis particularly related to geographical locations.

To share these insights I utilized two main approached. I created a GitHub repository here to store all the pieces of the project I had accumulated. Then, I created a portfolio on Google Sites to display my work.

Conclusion

This approach to the project allowed me to structure my process and record it for replication. It also enabled me to most of the tools from the course.