Statement of the business task:

To find out what differs between annual riders and casual riders bike usage?

Ask

Im trying to solve the main reason(s) that answer the question of what is the difference between casual and annual riders in bike users.

When I find out what the reasons are, the company can use this information to their advantage, and convert/steer the casual riders in to annual riders.

Business task is to convert the casual riders in to annual riders to get more riders overall and increase revenue.

The key stakeholders are Lily Moreno, Cyclist excutive team and Cyclist.

My task is to fdind out the difference in bike usage between these two groups. hence the title of this page.

Prepare:

the data is located on this cyclist database site: <https://divvy-tripdata.s3.amazonaws.com/index.html>

And the files I chose to use are the following:  Divvy\_Trips\_2019\_Q1.zip -  Divvy\_Trips\_2019\_Q4.zip

The data is organized monthly on years 2020-2023 and by quarters on year 2013-2020(q1)

The files are .csv documents and are in columns and rows.

data looks unbiased and the results seem to be aquired in a result oriented way.

I trust that the Cyclist company has made agreements with the customers when aquireing this data and we all have an understanding in what matter itsa going to be used. As far as the datasets go, there are no private information about clients that would put their privacy at risk.

Only me and the trusted personel and data analytics team have access to these files, and are accessable if customers want to see data collected about them on a setup meeting.

I compared the data sets to each other and inspected the datas stucture and methods used to store it.

The data gives information about what geographical locations the reiders are riding their bikes in, and how long for and which time of the year. Also it identifies the different categories of riders (casual, subscriber etc). So we could see some universal trends or patterns on how the bikes are being used.

also specifiec infromation about what types of people are subscribers, and who are not. And what kind of bikes are they riding(bikeid).

Possible problems are formatting issues with trip duration, and possibly start time end time date formatting. I am still deciding wether im going to use the whole year of data or just a quarter. The data set joining might be a problem so we will see how that goes.

keytasks

data is downloaded and stored in a spreadsheet as a copy of the original

oragnized in rows and columns with variables

sorted and filtered

the data is reliable for the reasons that there were no anomalies or uncategorized data

Sources used:

<https://divvy-tripdata.s3.amazonaws.com/index.html>

Divvy\_Trips\_2019\_Q1

Divvy\_Trips\_2019\_Q2 if needed

Divvy\_Trips\_2019\_Q3 if needed

Divvy\_Trips\_2019\_Q4 if needed

Process:

Im starting of with Divvy\_Trips\_2019\_Q1 in excel just get used to cleaning and manipulating the data with easy tools. I know it is not the most efficient way, but I plan on using R studio cloud for the whole analysis after. The big query SQL tool is not going to work because importing a dataset larger than 100 thousand kilobytes was not allowed. At the end I will try to join q1-q4 in r to get the analysis on a scale of a year.

I have made a copy, and only I can access it. I have verified that my data indeed is the one that I need and it has valuable information.

checked for duplicates, filtered for blank values or typos

There is no data on the spreadsheet that can surprise me as an error, or anomaly, and I am able to manipulate the data as I want to.

deliverable:

Cleaning process:

text to columns from a .csv file

fix column widths

checked for duplicates on the whole spreadsheet

filtered the data see if there are any missing values, blanks or n/a values

created extra columns such as “ride\_length” and “weekday”

formatted every column to either general, date, time or number depending on the need.

analyze:  
data should be using same naming convetions. Columns should be made so they can be aggregated an joined together. All data that is not necessary for the analysis should be delete. Extras columns should be made for more comprehensive analysis.

all data has been formatted properly from the right palce, to the columns and datypes they should have.

I was surprised that there were so called maintenance runs done on bikes which skewed the first results I did in excel. And after deleting them and the ride\_lengths that were negative values, I got proper results. Also from the analysis stand point I was surprised that casual riders ride for longer periods of time in one go than the subscribers on average.

key tasks:

Data has been aggregated to all\_trips dataset

Columns have been removed and added, and formatted to the latest most accurate form

Calculations have been done to find out:

average ride length

average ride length on a given weekday

Casual riders ride length median, max, min

And how the average ride length comapres to members

members like to ride more times on average on monthly and weekly basis. Casuals like to ride longer on monthly and weekly basis.

both groups have the same tendency to ride more at certain time of the year.

Summary:

casuals ride longer on average than members

member ride more often than casuals

members ride more often than average on weekdays, casuals on weekends.

everyones ride time on average doesn’t change significantly depending on weekday

Casuals and members both do most num of rides on spring-autumn period

Casuals ride time average is significantly higher on january and february than the rest of the year. members are even throughout the year.

Share:

yes, there are main differences pointed out in analyze summary

It tells that bot customers groups have different usages of bikes for riding, In some cases they do things alike and in others they are mirroring each other.

My findings cover that there are many key differences in usage and I believe we can utilize the findings to out advantage

My audience are my stake holders and the best way to communicate is sharing a presentation where I present my findings and analytical/process of work.

They will be verey important in sharing this data

Yes, I have kept notes through out and have everything avaiable for inspection if needed.

presentation with graphs and findings

In tableau

presentation ready

everything is stored in one place and shared to stakeholders