Annual Cyclistic Bike Share Data Sets 2020

Samad Ali

11/17/2021

Introduction to the Cyclistic Bike Share Data Set

Welcome to the Cyclistic bike-share analysis case study! In this case study, you will perform many real-world tasks of a junior data analyst. You will work for a fictional company, Cyclistic, and meet different characters and team members. In order to answer the key business questions, you will follow the steps of the data analysis process: ask, prepare, process, analyze, share, and act. Along the way, the Case Study Roadmap tables — including guiding questions and key tasks — will help you stay on the right path.

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.

1) PHASE(Ask)

According to the director of marketing team from Cyclistic Bikes assigned me to work on.

Guiding Question answer

- How do annual members and casual riders use Cyclistic Bikes differently?
- By doing Data analyzing and Data visualization I figure out How do annual members and casual riders use Cyclistic Bikes differently?

Key Tasks

- Identity and analyzing the Data sets from Cyclist company related to members and causal riders.
- Key stakeholders are the Cyclistic marketing analytic team and executive team, for making final decisions.

Deliverables

How do annual members and casual riders use Cyclist bikes differently, and to create new marketing strategy to convert casual bike rider to a member.

2) PHASE(Prepare)

The data that we will be using is Cyclistic's historical trip data from last 12 months (Jan-2020 to Dec-2020).

Guiding Question answer

- The data sets are located online in the Cyclistic Bike share AWS Cloud server database https://divvy-tripdata.s3.amazonaws.com/index.html.
- The data sets are available in the .zip folders containing the .csv files for each months from Jan-Dec 2020 i stored the original data sets into the separate folder for the backup and create the copy of the data sets to another folder with .xlsx file extension for cleaning and analyzing the data sets.
- The data set is coming directly from the Cyclists company, so it consider as a internal data and it is unbiased and authentic. During the analysis in spreadsheet I figured out that some of the data sets have missing cells for missing start and end station names, and also end of the months data set not have correct date format I also figured out that rider_id is unique for each rider for data Integrity.

Key Tasks

- I downloaded the original data sets in my local drive and working on the copy of the data sets.
- Its organized in the .zip folder for each month containing the .csv file, I change the naming convention of file and extension to .xslx to more readable and editable in EXCEL.
- I use R Studio to sort and filter the large data sets for each months.
- Data sets are unbiased and authentic it coming from the internal data from the Cyclists company.

Deliverables

Data is comming from the Cyclists internal server, and can only be used by Cyclistic marketing analytic team and executive teams following is the link for the Cyclist data source https://divvy-tripdata.s3.amazonaws.com/index.html

3) PHASE(Process)

Before I start analyzing, I uses to analyze data to make sure data is clean, free of error and in the right format.

Guiding Question answer

Tools

- I used R studio for data cleaning, manipulation and analysis and to combine the each month data into one data frame then combining all data_frame into full_yrs data

frame for analysis and remove station name, station id, end station name and end station id which are not needed for this data set according to Question in the ASK phase.

Key Tasks

- I used R studio for combining all the monthly files into each month data frame and combining all the monthly data frame into one single year data frame.

```
library(tidyverse)
library(lubridate)

# use library readxl for reading xlsx files

library("readxl")
month_Jan_Mar_2020 <- read_xlsx("Divvy_Trips_Jan_2020_March_2020.XLSX")
month_Apr_2020 <- read_xlsx("Divvy-Tripdata_April_2020.xlsx")
month_May_2020 <- read_xlsx("Divvy-tripdata_May_2020.xlsx")
month_June_2020 <- read_xlsx("Divvy-tripdata_June_2020.xlsx")
month_July_2020 <- read_xlsx("Divvy-tripdata_July_2020.xlsx")
month_August_2020 <- read_xlsx("Divvy-tripdata_Aug_2020.xlsx")
month_September_2020 <- read_xlsx("Divvy-tripdata_Sept_2020.xlsx")
month_October_2020 <- read_xlsx("Divvy-tripdata_Oct_2020.xlsx")
month_November_2020 <- read_xlsx("Divvy_Trips_Nov_2020.xlsx")
month_December_2020 <- read_xlsx("Divvy_Trips_Dec_2020.xlsx")


# use library dplyr for piping to remove station name, station id, end
station name and end station id which are not needed for this data set
according to Question in the ASK phase.

library(dplyr)
mth_Jan_Mar_2020 <- month_Jan_Mar_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Apr_2020 <- month_Apr_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_May_2020 <- month_May_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Jun_2020 <- month_June_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Jul_2020 <- month_July_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Aug_2020 <- month_August_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Sep_2020 <- month_September_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Oct_2020 <- month_October_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Nov_2020 <- month_November_2020 %>% select(-c(start_station_name,
start_station_id, end_station_name, end_station_id))
mth_Dec_2020 <- month_December_2020 %>% select(-c(start_station_name,
```

```
start_station_id, end_station_name, end_station_id))


# moving all the data frame into one full_yrs_trips data frame
full_yrs_trips <-
bind_rows(mth_Jan_Mar_2020,mth_Apr_2020,mth_May_2020,mth_Jun_2020,mth_Jul_202
0,mth_Aug_2020,mth_Sep_2020,mth_Oct_2020,mth_Nov_2020,mth_Dec_2020)
```

Then I formatted the date into separate fields in R data frame and create the columns(fields) for month, date, and year in order to calculate day of the week and ride length, and create filed for days_of_week and ride_length, then i remove the "docked bikes" from the rideable_type field because of negative time intervals in last months of the full_yrs_trips data frame after removing docked bikes I move the full_yrs_trips data frame to the full_yrs_trips_v2 data frame.

```
full_yrs_trips$date <- as.Date(full_yrs_trips$started_at)
full_yrs_trips$month <- format(as.Date(full_yrs_trips$date), "%m")
full_yrs_trips$day <- format(as.Date(full_yrs_trips$date), "%d")
full_yrs_trips$year <- format(as.Date(full_yrs_trips$date), "%Y")

full_yrs_trips$day_of_week <- format(as.Date(full_yrs_trips$date), "%A")
full_yrs_trips$ride_length <- difftime(full_yrs_trips$ended_at,
full_yrs_trips$started_at)


# ommiting all the docked bikes in a data frame in many various trips for the
docked_bike where ride_length is negative durration
full_yrs_trips_v2 <- full_yrs_trips[!(full_yrs_trips$rideable_type ==
"docked_bike" | full_yrs_trips$ride_length<0),]
View(full_yrs_trips_v2)
```

Deliverables

I used R_studio for creating reports and notebooks for documentation.

PHASE(Analyze)

In this phase I analyze the data frame into tibble to get the annual average ride length, day of the week and rideable type between members and casual riders.

Guiding Question answer

The data is organized in full_yrs_trip_v2 data frame to perform analysis on it, the data is properly formatted into fields and the in the new required fields(days_of_week, ride_length), for finding average ride_length, of User types(member, casual) and I also added date, month, day, and year for properly analysis the day of the week for each User types.

Key Tasks

To summarize and analyze the annual number of rides for members and casual I using piping and group_by function and using mean for average ride_length between user types to create a tibble in R studio.

```
full_yrs_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length))
```

To summarize and analyze the annual number of rides between User Types and also with rideable_type(classic and electric bikes) rides annualy.

```
full_yrs_trips_v2 %>%
  group_by(member_casual, rideable_type) %>%
  summarise(number_of_rides = n())
```

Deliverables

By using the the above R script the data frame gives the output in the tibble for the annual number of rides, average ride length, and the each day of the week for the rides by user types and rideable_types.

PHASE(Share)

In this Phase I converting the full_yrs_trips_v2 data frame into .xlsx format data set for visualizing and analyzing the data more clearly using Tableau.

Guiding Question answer

To answer the business questions I transfer the data frame for the full_yrs_trips_v2 into full_yrs_trips.xlsx file using library(writexl) package to visualize the data set into tableau its helps to clearly answer the business question by visualization.

Key Tasks

converting the data frame to data set in .xlsx format.

```
library(writexl)
write_xlsx(full_yrs_trips_v2,"C:/Users/samad/Desktop/R_studio_Cyclist_data_sets/Cyclist_Data_Sets_R/full_yrs_tripdata.xlsx" )
```

Dashboard link for visualization of the data set in Tableau.

https://public.tableau.com/views/VisualizationStory_Annual_Cyclists_Dataset_2020/Anual_Cyclist_Data_Set

Story link for visualization, key finding and recommendation of the data set in Tableau.

https://public.tableau.com/views/VisualizationStory_Annual_Cyclists_Dataset_2020/Annual_Cyclist_Data_Set

Deliverables

After visualizing the data set in tableau i figure out the following key findings in the data sets.

Key Findings

- Member riders, uses Cyclists bikes more then the casual riders.
- In Weekends riders uses Cyclists Bikes more rather than the week days.
- Average ride length time duration of Casual riders are more than the Member riders.

PHASE(ACT)

In this phase I figure out the recommended steps by the help of my data visualization. ## Guiding Question Answer

By the help of the data visualization I concluded some steps should be taken by the executive team for the business question that asked in the Ask phase.

Key Tasks:

I created my portfolio and add my case study with key findings and recommendations to share with executive team at Cyclistic Bike share company.

Dileverables

Following are my top three recommendations after visualizing the data sets for the Cyclisitic bike share company data sets.

- Giving memberships discount to members and price discount to casual bike riders for classic bikes to increase rides.
- Promote marketing strategies on weekends to increase the bike rides on week days because of more rides on weekends for both user types.
- Proper track of the docked bike time interval because of start at and end at timings have negative values.