Section 1 - Data wrangling and Visualization

Instructions:

1. Using R/Python and the dataset provided, complete the challenges below
2. You can use any visualization tools to create the dashboard
3. Final code along with results, graphs etc should be shared. If using Rmd, please knit into HTML. If using Python, please submit the ipynb file.

Data source:

Download data from this GDrive link - Assignment Data

Data set description

customer_id (integer): customer identifier, which is unique at customer level
driver_id (integer): customer identifier, which is unique at customer level
order_no (string): unique identifier at order level
booking_time (timestamp): the timestamp when booking completed
service_type (string): type of service that customer booked
actual_gmv(string): the amount of money that customer spent on particular booking

Context:

Gojek's aim is to increase daily transaction volumes (sum of total actual_gmv per day) in the next 6 months by 5X. Marketing team wants some preliminary data analysis to understand how to achieve this objective.

Questions:

1. Perform any cleaning, exploratory analysis necessary to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the drivers deliver only food orders on 2021-01-20?
2. Using the data that you had processed, perform an RFM analysis and create a dashboard for Marketing team to visualize key patterns in the data. Which customer cohorts should the Marketing team prioritize first?
3. Hypothesize how Gojek might leverage the insights gained from the data to generate higher transaction volumes.

Section 2 - SQL

Instructions:

1. This part must be solved strictly in SQL, use any SQL flavour that you are most familiar with

2. Please consider the readability of the code (e.g. adding comments along the query) and share only final executable SQL query in a .pdf or .txt format

Data source:

Use the cleaned dataset from the data wrangling section. For this section only, assume the dataset provided has data from 2016 onwards stored in the same format.

Context:

1. “Platform hard churn” is defined as someone who has not made any transaction (any service) in the past 6 months or more (6 months <= inactive)
2. “Platform soft churn” is defined as someone who has not made any transaction in more than 1 month but below 6 months (1 month <= inactive < 6 months)
3. “Product hard churn” and “Product soft churn” are defined very similarly, but instead of any transaction, it is on a specific service_type. Meaning GoFood hard churn users are those who have not completed GoFood transactions in the past 6 months

Questions:

1. Calculate average spending of GO-SEND users throughout every month of 2020
2. Calculate number of “platform hard churn” users and “platform soft churn” users for each month in 2020.
   For example:
   platform hard churn in February 2020 means how many users who have not made a booking in the past 6 months or more, i.e those who last transacted july 2019 or later (jan 2020, dec 2019, nov 2019, oct 2019, sept 2019, aug 2019 → no transaction in 6 months)
3. What is the reactivation rate (those who transacted after labelled churn) of “platform hard churn” users throughout each month of 2020? Identify which product helps in reactivation the most in each month.

Section 3 - Experimental Design

Instructions:

1. Explain your answers using a few sentences
2. Final answers can be submitted in a pdf file

Questions:

1. The original version of an ecommerce website is quite basic, featuring only text on the site. They plan to add some product images, and intend to run an A/B test on their website to see if it helps sales. You are the analyst for this test.

   a. What would your hypothesis be for running this test?
   b. What would be your primary metric to measure the success and why?
   c. Tell us what secondary metrics you might look at to help you make a call on if a
   test is performing correctly and why?

d. You are asked to filter down the results further to see if the primary metric is performing well on different breakdowns. What potential problems could occur with your analysis if you layer different filters on top of the other?

e. Below are the results for an A/B test that you have analysed. Based off the
results displayed below, what would be your recommendation to your product manager? Why did you make this recommendation?

Cohort	Visitors	Converters	Traffic split
Variant 1	8000	3000	10%
Control	72000	25000	90%

f. In what situations would you choose not to run an experiment, favouring rolling out a new feature immediately? When this occurs, how should you aim to handle this as an
Analyst