We are pleased to invite you to the interview process for our Decision Science Team! This is a practical exercise that will test your programming and analytical skills, please **include your codes as PDF** in the submission. The programming language that is acceptable is Python/R

Instructions

- Submit your answers in pdf format
- You may not consult with any other person regarding the test.
- You may use internet searches, books or notes you have on hand.
- If you are stuck from a technical aspect, write down in words how you would go about answering this question and what other information you would need.
- The test has three parts all of which are mandatory
- For part 2 you may submit a word document.
- A thoughtful clean & commented code is expected as a submission.

Part 1

(FICO + Region)

Attached are two sample datasets: the first one is called **FICO** and it contains customer ids and individual FICO scores. The second one is named **Region**. It holds the same customer ids and regions where each customer is located.

1. You are tasked to explore the FICO dataset. Walk us through your process on the tasks below:
   1) You need to think about cleaning the data first. Common data problems include duplicates, missing, and errors in the data. Mark rows with data problems as "Missing" in the FICO column.
   2) Think about what you know about credit scores. Segment the FICO scores into 5 groups. Give your **reasoning** for the bucketing. Display the number of customers and percentage of each segment in your answer and create a histogram of the distribution if you are using Python.
   3) Do you notice anything particular about this distribution? Do you think this reflects what's happening in the real world?

2. Now that you have a clean dataset for FICO. Create a temp table to store the information of FICO score and region for each customer. Make sure the customer id is the same for each record. Display the regions which have the **second highest** and **lowest** average FICO score. The result of your query should display only **two rows** showing the region and its average FICO score. Make sure you provide all the interim steps if needed in your final submission.

Part 2

(Guesstimate questions)

3. How many **Red coloured top Honda SUV cars** do you think will be sold in India in **2022**? Questions to consider:

I. What factors do you think will impact sales? You may google about guesstimate questions to answer this question better.

II. Assume you have all the data you need, what statistical methodology or algorithm will you use to make this sales forecast? Please give a brief explanation of why you choose this model.

III. How would you evaluate your model or determine its accuracy?

Part 3

(Fyttlyf website data)

4. Write a function in Python/R
    a. To get the output that looks like the below image.
    b. x1 in the image could be read as Unique count of visitors of Level 2 / Unique count of visitors of Level 1.
    c. The function should take the rows (in this case Traffic source) as an argument i.e., we should be able to provide "devc_name" or "browser_type" as an argument to get the same table

Traffic Source	Level 2/Level1	Level3/Level2	Level4/Level3	Level5/Level4
Traffic Source 1	x1			
Traffic Source 2				
Traffic Source 3				

5. The payload column contains keys=value pairs separated by '&'. Make a function that exports a CSV that expands the data in the following format for all key-value pairs in the payload column

evnt_ts	visitor_id	payload_key	payload_val
23:00	123	isp_mobile_carr	O2 Deutschland
23:00	123	session_id_cook	00fa8a8d-b64b-
.	.	.	.
.	.	.	.
.	.	.	.

6. Can you write python/R script to answers the following questions with visualization?
    a. What is the CTR (Click-through rate) at geo_cntry level. (CTR = count distinct sessn_id where event = click / count distinct sessn_id where event = impressions)
    b. What is the trend of the distinct count of visitors on an evnt_dt level? Display your answer in the best possible graph.
    c. Which browser type get the highest click sessions & on which evnt_dt?
7. Please briefly describe the latest Data science project you did in less than 500 words.