We are pleased to invite you to the interview process for our Finance Decision Science Team! This is a practical exercise that will test your programming and analytical skills, please **include your codes** in the submission. Following programming languages are acceptable-Python/SQL.

**Instructions**

* Answer the questions to the best of your abilities **within 24 hours BY YOURSELF**.
* 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.

**Questions:**

Attached are two sample datasets: the first one is called **FICO** and it contains customer ids and individual FICO score. 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:
2. 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.
3. Think about what you know about credit score. Segment the FICO scores into 5 groups. Give your **reasoning** for the bucketing. Display the number of customers and percentage of each segments in your answer, and create a histogram of the distribution if you are using Python.
4. Do you notice anything particular about this distribution? Do you think this reflects what’s happening in the real world?
5. 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.
6. What is your estimate of number of green T-shirts sold in the US in **2018** (provide a **number** and your **reasoning**)?
7. How many green T-shirts do you think will be sold in the US in **2020**?

Questions to consider:

1. What factors do you think will impact sales?
2. Not all of your factors have readily available data or cannot be measured numerically, how will you source this data or what can be a proxy for these factors?
3. What statistical methodology or algorithm will you use to make this forecast? Please give a brief explanation why you choose this model.
4. How would you evaluate your model or determine its accuracy?