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 skills, please **include your codes** in the submission. Any programming language is acceptable, Python/R/SQL/SAS preferred.

**Instructions**

* Answer the questions to the best of your abilities **within 4 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.
* Please find two data sets for questions 1-2. No data is provided for questions 3-4

**Pre-Test Questions (Optional):**

What was your undergraduate studies GPA?

What was your graduate program GPA if available?

**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 **2021** (provide a **number** and your **reasoning**)?
7. How many green T-shirts do you think will be sold in the US in **2023**?

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?

Answer 1

1.3 refer to csv q1\_1.csv

1.2

Chart, histogram

Description automatically generated

Histogram with FICO scores of all customers. We would expect it to follow Normal Distribution but in reality it is right skewed. This could be because on average the customers generally have a higher FICO score, which is greater than 700.

Chart, bar chart

Description automatically generated

By taking 20% quantile intervals, we get another histogram in which customers are grouped by these intervals –

Interval(419.999, 758.0, closed='right'),

Interval(758.0, 785.0, closed='right'),

Interval(785.0, 791.0, closed='right'),

Interval(791.0, 836.0, closed='right'),

Interval(836.0, 850.0, closed='right')

Not all customers will have scores at lower range of 300-400. Hence it makes sense to bucket according to quantile ranges in which the scores of these customers fall. By using a 20% quantile result, I have created this bucket, i.e. [0.2, 0.4, 0.6, 0.8, 1]. So there is nothing weird in the distribution and it is a correct representation of what happens in the real world.

Answer 2

Cleaned data in q1\_3.csv

Queries in Sqlite –

1. CREATE TEMP TABLE Amex(  
   AcctId,  
   Fico,  
   Region  
   );
2. .mod csv
3. .import /Users/charvigupta/Desktop/q1\_3.csv Amex
4. CREATE TABLE AmexFinal(  
   AcctId,  
   Fico,  
   Region  
   );
5. INSERT INTO AmexFinal  
   SELECT \*  
   FROM Amex;
6. SELECT Region, AVG(Fico) as avg FROM AmexFinal GROUP BY Region ORDER BY avg DESC LIMIT 1,1;  
   Phoenix,788.120314389359
7. SELECT Region, AVG(Fico) as avg FROM AmexFinal GROUP BY Region ORDER BY avg ASC LIMIT 1,1;

Dallas,783.397524071527

Answer 3

For estimating number of Green tshirts sold in USA in 2021, we can follow these steps –

1. Population of USA in 2021 let say 400M
2. Lets assume 200M men and 200M female
3. Then we estimate how many men fall under categories for infants 10M, children < 5 years 60M, children < 18 yrs, 18 yrs < adult < 65 120M (working population), and elders > 65 10M
4. Then we estimate how many tshirts each age segment **buys** per month. Suppose infants need 4 tshirts a month, children need 3 tshirts a month, adults 3 tshirts per month and elders 1 per month
5. Then we get an estimate of how many green tshirts are produced at retail shops (by seeing their catalog) This gives us a chance that how many of tshirts are green of the total produced. Say 10 percent
6. Then we assume how many each age segment would buy each month. (Assumption, they don’t discard tshirts, also all tshirts produced are sold) 4\*0.1, 3\*0.1, etc….
7. We add the values for each age segment and repeat this for female segment. 4\*0.1\*10M…
8. Assume that males need more tshirts than females as females have other clothing options as well.

Answer 4

1. Sales would be impacted by various factors like –
   1. manufacturing capacity of brands - suppose if there is less demand of tshirts, then manufacturers would slow down production.
   2. Purchasing capacity of customers – high inflation could mean people can reduce spending on non-essential items like clothes and increase it on food.
   3. Do people need to go out that often ? Eg. If a pandemic like covid strikes, a lot of people shift to WFH and they don’t need as many tshirts as they would in a regular year. This could be a categorical feature.
2. We can source the data from manufacturing points. But for other factors, economic indices like inflation, food prices, etc could be a substitute for measuring the impact created by that factor on the sales of tshirt
3. We can use a time series model to forecast the number of tshirts that could sell. Assuming we have past data of tshirts sold in the previous months, along with production data and purchase data
4. To determine its accuracy we can do a market survey and estimate accuracies of our factors.