Data Scientist – technical test

These questions are designed to test your logical thinking, data modelling, stats and programming ability.

Some tips:

- There is no time limit to finish the test but we would like to receive your answer during a week since you receive the technical test (Some guidance is given below about how long we expect you to take on each question, but this will vary a lot depending on your skillset)
- An important skill for data scientist is to validate your work, so please check your answers (and where appropriate show how you have checked them)

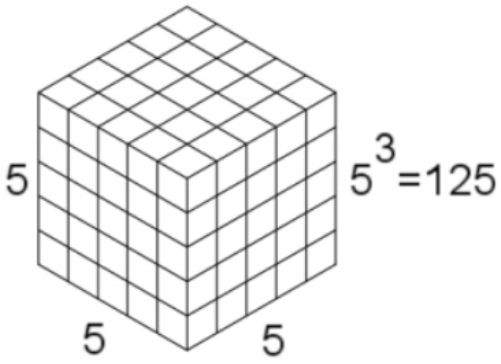
Once you have completed your test, please email the answer (you can type out answers on a word document or github repo).

Part 1. Logic (10-20 minutes)

A cube is painted Blue on all six sides. It is divided into 125 (=5 x 5 x 5) equal smaller cubes.

Find:

1. The number of smaller cubes having
    a) 3 faces coloured?
    b) Exactly 2 faces coloured?
    c) Exactly 1 face coloured?
    d) 0 faces coloured?



2. All 125 cubes are put into a bag. If a single cube is selected at random from the bag, find probability of picking a cube having 1 or more Blue faces

3. What is the average number of Blue faces on a small cube?

In the above situation N=5, (with N^3 =125)

4. For general N, give a formula for (1.2) the number with exactly 2 faces coloured

5. For what values of N is this formula correct?

Part 2. SQL (15-30 minutes)

We have the following tables:

user - table with information about registered students

Column name	Datetype	Description
id	integer	Unique id of user
date_joined	timestamp	Date and time of registration
country_code	varchar(2)	country of user (2-letter country abbreviation)

payment - table with information about payments

Column name	Datetype	Description
id	integer	Unique id of payment ((one student can have from 0 to X payments)
user_id	integer	Id from table user
payment_amount	float	Paid amount in USD
created_at	timestamp	Date and time of the payment

lesson - table with list of lessons.

Possible statuses of lessons:

CONFIRMED - lesson happened successfully

SCHEDULED - for future lessons

CANCELED - for lessons that were canceled

Each lesson appears on the table only one time. Lesson status is being updated.

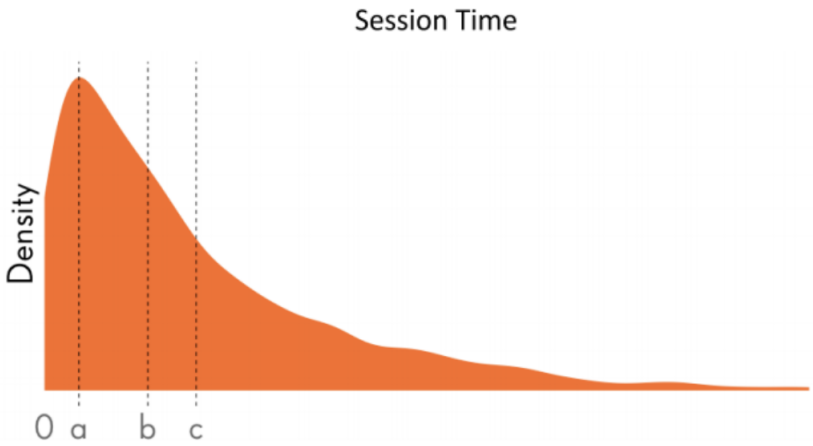
Column name	Datetype	Description
id	integer	Unique id of lesson ((one student can have from 0 to X lessons)

user_id	integer	Id from table user
status	varchar(255)	Current status of lesson (can be updated)
created_at	timestamp	Date and time of the payment
hours	integer	Duration in terms of hours of the lesson

Write SQL query that returns:

1. number of registered users by country
2. % of users, who made their first payment in 3 days after registration by country
3. % of users, who made their first payment in 3 days after registration and had 2 confirmed lessons in 7 days after registration by country
4. % of weekly new users that never have done a payment
5. Advanced level (Extra point): Write the SQL that returns how many hours of confirmed lessons a specific user (for example user_id=1) has taken between payments.

Part 3. Statistics (5-10 minutes)



1. What standard measures are a,b and c most likely denoting? Please explain why.

2. If you drew n samples from this distribution and measured their mean, then repeated that many times, how would you expect the distribution of those sample means to differ from the distribution?

3. Would its standard deviation be bigger, smaller or the same as this distribution's standard deviation and why?

Part 4. Storytelling (10-20 minutes)

Describe a technological abstract concept of your choice (for example: internet, electricity, credit card, email, slack, …) to a 4 years old child. Please make sure you don't use other complex concepts while describing it and use a plain language in order to maximize the chances that the child will be able to understand it. We are going to evaluate your capacity to simplify and explain complex subjects.

(Mínimum 100 words, maximum 400 words)

Part 5. Recommendation Engine (+60 minutes)

Preply wants to include a small recommendation engine to suggest potential good tutors based on a collaborative filtering. We want to show "users that viewed this tutor also viewed…" section within the tutors profile page.

Given a Preply profile views dataset, make a small recommendation system that given one profile_id returns the 10 most potential tutors to be viewed based on a collaborative filtering. We want to consider potential good tutors those tutors that have been viewed by users that also viewed our input tutor_id. The return should include the top 10 potential tutor_id's (ordered by popularity).

The dataset includes the following columns:

- id: sequential id of the profile views table (int)
- user_id: unique code that identifies the users viewing the page (varchar(255))
- tutor_id: unique code that identifies our tutors (varchar(255))

Each row of the dataset represents one tutor profile view. A tutor profile view is triggered when a user visits a tutor profile page to view more information about their experience teaching.

Dataset: https://docs.google.com/spreadsheets/d/e/2PACX-1vTyQAs46R3D8qJcvrXhl0E03fEGb8kN1kMLx1KNA_HbAwBcHOOx_eJBXU8pnplPOTdQgiQOsDEBzX5L/pub?gid=1102523268&single=true&output=csv

Questions:

1. Which tutors will your recommendation engine return given the tutor_id "ff0d3fb21c00bc33f71187a2beec389e9eff5332"? Will it work for any tutor_id of the dataset?
2. What you will suggest to do to avoid a cold start of a tutor (a tutor that has not yet received any view)?
3. What will you do if given a tutor_id the recommendation system returns less than 10 potential tutor_ids?
4. If the dataset was including the possibility of a user to view a tutor_id more than once, how it will influence the recommendation system? Which dataset do you think will perform better and why?
5. Try to reason why do you agree or disagree on removing from the dataset, the rows where tutors are visiting their own tutors page?
6. Do you have any improving proposition that could maximize Preply's tutors page views? (different algorithm? Different dataset? …)