FACEIT - DATA CHALLENGE

Thank you for showing interest in our data science opening!

We’ve prepared an assignment which will help us understand if you would be a good fit for the role, but which will also give you the opportunity to work with some of our actual data.
Don’t hesitate to ask any questions you might have and … have fun!

Q1. SQL Challenge

The following table is a snapshot of a table in our data warehouse which stores data for each match played on FACEIT (the original table has millions of rows, and the below is a snapshot). Each row has a unique combination of user_id/match_id, which together serve as the table’s primary key.

user_id	**match_id**	**game**	**created_at**	**membership**	**faction**	**winner**
ab542a	e21887	csgo	2018-01-02 18:45:25	free	faction1	faction1
ef72da	df891f	dota2	2018-01-02 08:20:01	free	faction2	faction1
f5c776	1p5c47	wot_RU	2017-12-30 15:25:25	free	faction1	faction2
5a278a	af14e8	csgo	2018-01-01 14:27:15	premium	faction2	faction2
ae346d	af14e8	csgo	2018-01-01 14:27:15	free	faction2	faction2
2b13d8	a88c44	csgo	2017-12-31 12:33:34	free	faction1	faction2
ace797	df891f	dota2	2018-01-02 08:20:01	premium	faction2	faction1
ace797	ae193a	csgo2	2018-01-03 18:18:22	premium	faction1	faction1

Using vendor neutral ANSI SQL (or if your syntax is specific to one flavour, please specify which):

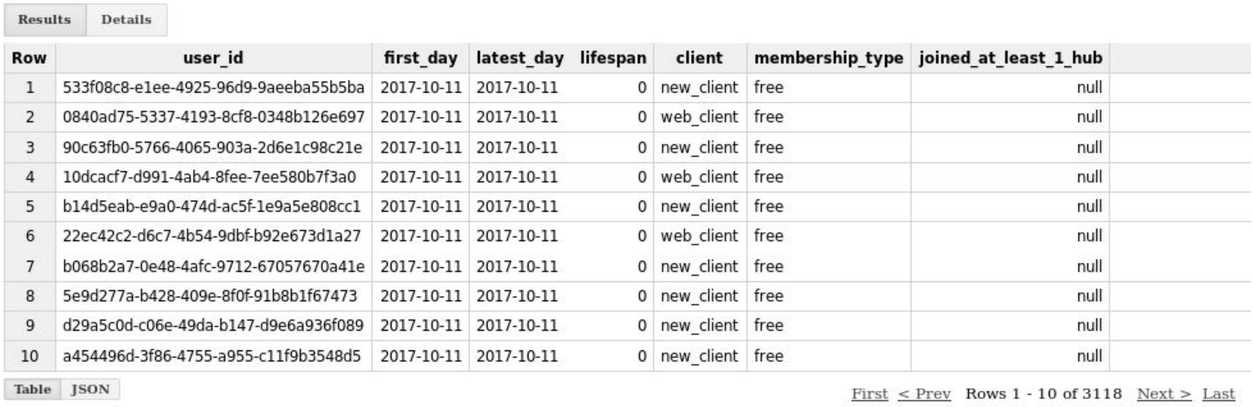
1. Write a query which counts the amount of matches which took place in 2018 and had at least one premium user participating.
2. Write a query which finds the list of all users who had at least one winning streak of 3 matches on the platform. A streak here is defined as achieving three or more consecutive wins in the same game. If a user won a match then his/her *faction* will be the same as the faction in the *winner* column.

Q2. Data Exploration & Analysis Challenge [OPTIONAL, for extra points!]

In 2017, we launched a new desktop client, which players can download and use to access Faceit. In the beta phase, the client was made available only to a subgroup of users, which was tagged accordingly.

One of our product managers is now working on releasing a mobile app and wants to understand how the new client had performed in relation to a specific competition, hubs. Supposing that the attached csv is the dataset you prepared to answer these questions, use your favourite python or R libraries to do so and ideally present your analysis in a jupyter notebook.

The image below is a snapshot of the data you will find in the CSV attached (client.csv).



Row	user_id	first_day	latest_day	lifespan	client	membership_type	joined_at_least_1_hub
1	533f08c8-e1ee-4925-96d9-9aeeba55b5ba	2017-10-11	2017-10-11	0	new_client	free	null
2	0840ad75-5337-4193-8cf8-0348b126e697	2017-10-11	2017-10-11	0	web_client	free	null
3	90c63fb0-5766-4065-903a-2d6e1c98c21e	2017-10-11	2017-10-11	0	new_client	free	null
4	10dcacf7-d991-4ab4-8fee-7ee580b7f3a0	2017-10-11	2017-10-11	0	web_client	free	null
5	b14d5eab-e9a0-474d-ac5f-1e9a5e808cc1	2017-10-11	2017-10-11	0	new_client	free	null
6	22ec42c2-d6c7-4b54-9dbf-b92e673d1a27	2017-10-11	2017-10-11	0	web_client	free	null
7	b068b2a7-0e48-4afc-9712-67057670a41e	2017-10-11	2017-10-11	0	new_client	free	null
8	5e9d277a-b428-409e-8f0f-91b8b1f67473	2017-10-11	2017-10-11	0	new_client	free	null
9	d29a5c0d-c06e-49da-b147-d9e6a936f089	2017-10-11	2017-10-11	0	new_client	free	null
10	a454496d-3f86-4755-a955-c11f9b3548d5	2017-10-11	2017-10-11	0	new_client	free	null

For the data, assume that the new client first became available to a subset of users on 11 October 2017 (experiment start date). Column *first_day* shows the first day that a user played a match after the experiment start date, column *latest_day* when they were last seen playing and column *lifespan* the difference between the two dates.

Your product manager wants to understand if the new client has increased the number of people joining hubs. They therefore asked you whether you could test if the following hypothesis is true:
H0: The *proportion* of users who used the new client and joined at least one hub is larger than the proportion of users who are using the old web client and joined at least one hub.

Conduct a test of hypothesis, and state any assumptions you have made. What would your feedback be to the product team?

Q3. Machine Learning Challenge

Preventing abusive behavior is one of the main issues that the team has been entasked to solve.

The team decides to firstly try to tackle toxicity in text messages, and to leverage external api tools, like Perspective Api.

The attached labelled dataset (toxicity.csv) is the result of the exploration and labelling of 12,000 individual comments. It has the following variables:

- Flirtation (FLOAT) pickup lines, complimenting appearance, subtle sexual innuendos, etc.
- Identity_attack (FLOAT) negative or hateful comments targeting someone because of their identity.
- Insult (FLOAT) insulting, inflammatory, or negative comment towards a person or a group of people.
- Severe_toxicity (FLOAT) a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
- Sexually_explicit (FLOAT) contains references to sexual acts, body parts, or other lewd content
- Threat (FLOAT) describes an intention to inflict pain, injury, or violence against an individual or group.
- Label (STRING) describes the categorical label that our internal team assigned to the comment.

The FLOAT values are all probability scores between 0 and 1, that the perspective api models are returning for each comment of the dataset.

Apply Machine Learning techniques to produce a model able to predict if a comment will be perceived as toxic or not toxic.