

Задание 1. SQL.

Напиши sql-запросы, позволяющие ответить на следующие вопросы:

Ссылка на файл sql: https://github.com/xssokolova/profi/blob/master/football_stat.sql

1. В каких командах есть вратари ('GK'), получавшие карточки на турнире?
(если таблица пустая, значит ни один из вратарей не получил желтую карточку за время турнира)

```
SELECT teams.name AS 'name of team' FROM teams JOIN players_stat
ON teams.id = players_stat.team_id
WHERE players_stat.position='GK' and (players_stat.yellow_cards +
players_stat.red_cards) > 0
GROUP BY teams.id;
```

2. Каков 90-ый перцентиль возраста игроков, сыгравших на турнире не менее 270 минут?
(используется метод ближайшего ранга)

```
WITH play AS (SELECT age FROM players_stat WHERE min_played >= 270 )
SELECT
CAST(SUBSTRING_INDEX(SUBSTRING_INDEX( GROUP_CONCAT(age ORDER BY
age SEPARATOR ','), ',', 90/100 * COUNT(*) + 1), ',', -1) AS
DECIMAL)
AS '90th Per'
FROM play;
```

3. В каких командах голы на турнире забивали только нападающие ('AT')?
(если таблица пустая, значит нет команд, в которых голы на турнире забивали только нападающие)

```
SELECT teams.name FROM teams
WHERE teams.goals > 0 AND NOT EXISTS
(SELECT * FROM players_stat
WHERE (players_stat.team_id = teams.id AND players_stat.position <>
'AT' AND players_stat.goals > 0));
```

4. Представители какой команды чаще других получали желтые карточки во время игр?
Используя таблицу "teams"

```
SELECT name, yellow_cards
FROM teams
WHERE teams.yellow_cards >= ALL(SELECT teams.yellow_cards FROM teams);
```

Или используя данные из таблицы "players_stat"

```
WITH temp AS
(SELECT players_stat.team_id, teams.name, players_stat.yellow_cards
FROM players_stat
JOIN teams
ON players_stat.team_id = teams.id)
SELECT name, sum(yellow_cards)
FROM temp
GROUP BY team_id
HAVING sum(yellow_cards) >= ALL(SELECT sum(yellow_cards) FROM temp GROUP
BY team_id);
```

Тестовые данные:

```
DROP TABLE IF EXISTS teams, players_stat;
```

```
CREATE TABLE teams (  
    id                int,  
    name              varchar(32),  
    games_played      smallint,  
    goals             smallint,  
    goals_minus       smallint,  
    yellow_cards      smallint,  
    red_cards         smallint,  
    PRIMARY KEY(id)  
);
```

```
CREATE TABLE players_stat (  
    id                int,  
    name              varchar(32),  
    team_id           int,  
    position          varchar(8),  
    age               smallint,  
    min_played        int,  
    goals             smallint,  
    yellow_cards      smallint,  
    red_cards         smallint,  
    PRIMARY KEY(id)  
);
```

```
INSERT INTO  
    players_stat(id, name, team_id, position, age, min_played, goals,  
    yellow_cards, red_cards)  
VALUES  
    (1, 'Ivan Ivanov', 1, 'GK', 18, 270, 0, 1, 0),  
    (2, 'Max Maximov', 1, 'GK', 32, 270, 0, 0, 1),  
    (3, 'Petr Petrov', 1, 'DF', 24, 180, 0, 0, 0),  
    (4, 'Vasya Vasyev', 1, 'MD', 17, 220, 0, 0, 0),  
    (5, 'Fedor Fedorov', 1, 'AT', 18, 60, 4, 2, 0),  
  
    (6, 'Johny John', 2, 'GK', 26, 360, 0, 1, 0),  
    (7, 'Peter Peter', 2, 'DF', 30, 280, 0, 1, 0),  
    (8, 'Basil Basil', 2, 'MD', 28, 300, 0, 1, 0),  
    (9, 'Smith Smith', 2, 'AT', 21, 20, 1, 0, 1),  
  
    (10, 'Ernst Ernst', 3, 'GK', 28, 270, 0, 0, 0),  
    (11, 'Hans Hans', 3, 'DF', 28, 270, 0, 1, 1),  
    (12, 'Frank Frank', 3, 'MD', 23, 180, 0, 0, 1),  
    (13, 'Klaus Klaus', 3, 'AT', 36, 90, 0, 0, 0);
```

```
-- Задаём непараметризуемые поля teams
```

```
INSERT INTO  
    teams(id, name, games_played, goals_minus)  
VALUES  
    (1, 'Russia', 3, 2),  
    (2, 'England', 4, 1),  
    (3, 'Germany', 3, 3);
```

```
-- Задаём поля teams, параметризуемые от players_stat
```

```
UPDATE teams SET  
    goals = (SELECT sum(goals) FROM players_stat WHERE team_id = teams.id GROUP  
    BY team_id),  
    yellow_cards = (SELECT sum(yellow_cards) FROM players_stat WHERE team_id =  
    teams.id GROUP BY team_id),  
    red_cards = (SELECT sum(red_cards) FROM players_stat WHERE team_id =  
    teams.id GROUP BY team_id);
```

Задание 2. Python.

Ссылка на файл Jupiter Notebook:

<https://github.com/xssokolova/profi/blob/master/Task2%20Notebook.ipynb>

Задание 3.

1.

Какова примерная вероятность победы Франции по прогнозам букмекеров?

$P(\text{победа Франции}) = 100 / 1.85 = 54 \%$

Каков прогнозируемый доход букмекерской конторы в процентах от суммы ставок на данный матч?

$D = ((1/1.85 + 1/4.35 + 1/3.45) - 1) * 100\% = 6\%$ от общего объема ставок

Как поменяются коэффициенты на матч, если букмекеры не будут закладывать в них свой предполагаемый доход (оценка вероятности исходов сохраняется)?

Если букмекеры не будут закладывать в коэффициенты свой предполагаемый доход, то коэффициенты увеличатся.

Например: коэффициент выигрыша Франции с включенным предполагаемым доходом равен 1.85 ($1/0.54=1.85$), допустим доход букмекера при ставке на Францию составляет 2%, тогда коэффициент станет равным 1.92 ($1/0.52=1.92$).

2. Верно ли утверждение, что жители Санкт-Петербурга в среднем более довольны инфраструктурой города, чем жители Москвы? Почему?

Для ответа на этот вопрос недостаточно данных, необходимы стандартные отклонения выборок или ранги и оценки города по критериям.

Сравнить средние значения двух независимых выборок между собой можно с помощью t-критерия Стьюдента (при нормальном распределении признаков в каждой выборке) или с помощью U-критерия Манна-Уитни (если распределение признаков не соответствует нормальному). После определения критерия проводим проверку гипотез, на основании которой выявляем есть ли различие между средними оценками рейтинга городов.

H_0 : Среднее_{мск} = Среднее_{спб}

H_1 : Среднее_{спб} > Среднее_{мск}

Проверяемая статистическая гипотеза (различия между группами не значимы/отсутствуют) - H_0 .
Альтернативная статистическая гипотеза (различия между группами свидетельствует о большем Среднем_{спб}) - H_1 .

Задание 4.

Есть ли какие-то ограничения и ошибки в таком подходе? Изменил ли бы ты что-то в проектировании АБ-теста? Аргументируй свой ответ.

При распределении трафика во время АБ-теста надо учесть его равномерность распределения не только между 2 группами (экспериментальной и контрольной), но и по срезам: между городами, операционными системами мобильных устройств, новыми и лояльными пользователями и т.п.

Правильно выбрать время проведения теста и его длительность. Например, в канун праздников может снизиться спрос на репетиторов, но повыситься спрос на услуги мастеров красоты.

Результаты АБ-теста за этот период нельзя считать объективными. Длительность проведения теста должна быть основана на наблюдении за временем, которое пользователи тратят от момента скачивания приложения до момента размещения заказа. Если присутствует значительное число пользователей, тратящих на оформление заказа более 1 недели, то следует скорректировать время исследования.

Проверить правильность работы инструмента распределения трафика, чтобы в тест попало 100% аудитории приложения. Провести АБ-тест, ничего не меняя в приложении и убедиться, что в группах с одинаковыми версиями мобильного приложения нет лидирующей версии, т.е. не должно быть разницы по ключевым показателям.

Цель новой версии – повысить конверсию из открытия приложения в оставление заказа. Замерить эффективность теста предлагается одним показателем - сравнением конверсии в заказах в группах. Возможно, не стоит ограничиваться одним показателем, а внедрить несколько. Например, добавить показатель отказов, когда пользователь скачивает приложение, но не регистрируется в нем и не пользуется им.

Для принятия решения по результатам АБ-теста нужно отсеять экстремальные значения по показателям эффективности теста, проверить их по разным срезам и убедиться, что нет крайних случаев (если есть, то понять их причину), сравнить конверсии на основании интервалов распределения.

Задание 5.

Ты – продуктовый аналитик в Profi. Твоя первая задача - разобраться с причинами того, что метрика %клиентов, успешно нашедших специалистов после оставления заказа, которая стабильно находилась на уровне 50% последние 2 года, в августе 2020 года упала до 40%. Расскажи, как ты подойдешь к решению этой задачи.

С какой целью используется метрика? Какие метрики зависят от метрики клиентов, успешно нашедших специалистов после оставления заказа (далее метрика К), и от каких зависит она?

Проверю как посчитана метрика К:

- осталась ли формула для расчета метрики К прежней или в нее были внесены изменения;
- исключу возможность ошибки в коде при подсчете метрики К;
- достаточно ли данных использовалось для подсчета (например, данные могли не загрузиться, потеряться или быть записаны с искажениями).

Полезным инструментом для понимания контекста могут быть старые исследования метрики К, с помощью них можно узнать свойственны ли данной метрике колебания раз в период, например, раз в 5 лет или нет? Что было характерно для того периода?

Проведу самостоятельное исследование данных и обращу внимание на то, что могло привести к изменениям:

- Проверю эту метрику по категориям (например, репетиторы, мастера по ремонту и т.п.) Возможно, конверсия уменьшилась для какого-то одно или несколько направлений;
- Определю взаимосвязь других метрик с этим показателем (например, средняя стоимость сделки могла возрасти, но количество сделок уменьшиться);
- Узнаю были ли внедрены новшества в сервис (например, новый интерфейс, новые условия взаимодействия между клиентом и профессионалом);
- Рассмотрю внешние условия (например, карантинные меры и как следствие ограниченный контакт между людьми);
- Посмотрю были ли изменения на рынке сервиса по поиску и подбору профессионалов (новый участник рынка или новые возможности прежних участников, что отвлекло часть профессионалов и/или часть клиентов на другие ресурсы).

Задание 6.

Почему тебя привлекла данная вакансия?

Мне нравится работать с данными, погружаться в новые темы, задавать вопросы и искать ответы на них. Поэтому я прошла полное обучение в John Hopkins University по специализации Data Science и приступила к поиску работы.

В вакансии младшего продуктового аналитика в компании PROFI.RU меня привлекает возможность работать над продуктами, которые могут качественно повлиять на повседневную жизнь. Кроме того, Тарас Торконяк рекомендовал компанию как открытую к диалогу, с теплой атмосферой в коллективе, регулярной обратной связью от руководителей, отлично организованном обучающем процессе для начальных позиций.

Какие у тебя зарплатные ожидания?

50-70 тыс. руб. net