**SQL - легкий уровень**

Даны 2 витрины (clnt\_aggr и clnt\_data), которые содержат:

* clnt\_data – актуальную персональную информацию о клиенте
* clnt\_aggr – агрегированную информацию о финансовом поведении клиентов

**Схема и пример данных - clnt\_aggr**

|  |  |  |  |
| --- | --- | --- | --- |
| **clnt\_aggr** | | | |
| client\_dk | Идентификатор клиента | number | Pk |
| tb\_name | Название тербанка, в котором обслуживается клиент (ТБ) | Varchar(100) |  |
| salary | Заработная плата за месяц | number |  |
| pos\_amt | Сумма POS-расходов по всем картам клиента | number |  |
| pos\_qty | Кол-во POS-операций по всем картам клиента | number |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| client\_dk | tb\_name | salary | pos\_amt | pos\_qty |
| 507 851 886 | Московский банк | 15000 | 6400 | 8 |
| 267 188 214 | Центрально-Черноземный банк | 50000 | 13200 | 10 |
| 148 849 526 | Московский банк | 30000 | 35000 | 3 |
| 613 898 474 | Юго-Западный банк | 150000 | 61000 | 26 |

**Схема и пример данных - clnt\_data**

|  |  |  |  |
| --- | --- | --- | --- |
| **clnt\_data** | | | |
| сlient\_dk | Идентификатор клиента | number | Pk |
| gender | Пол клиента | Varchar(1) |  |
| age | Возраст клиента (полных лет) | number |  |
| сhild\_qty | Кол-во детей и иждивенцев у клиента | number |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Client\_dk | gender | age | сhild\_qty |
| 507 851 886 | M | 25 | 1 |
| 267 188 214 | M | 46 | 3 |
| 148 849 526 | W | 18 | 0 |
| 613 898 474 | M | 73 | 2 |

**Задания:**

1. Вывести распределение клиентов по территориальным банкам. Учитывать только тех клиентов, у которых pos-оборот больше или равен 1000 руб.

-- 1)  
SELECT  
 ca.tb\_name AS bank\_name,  
 *COUNT*(DISTINCT client\_dk) AS cilents\_qty  
FROM clnt\_data AS cd  
LEFT JOIN clnt\_aggr AS ca ON ca.client\_dk = cd.Client\_dk  
WHERE  
 ca.pos\_amt >= 1000  
GROUP BY  
 ca.tb\_name

1. Вывести распределение клиентов по территориальным банкам и группе по возрасту (меньше 18 (включительно), от 18 до 30 (включительно), от 30 до 60 (включительно), больше 60 лет).

-- 2)  
SELECT  
 ca.tb\_name AS bank\_name,  
 *COUNT*(DISTINCT CASE WHEN cd.age < 19 THEN cd.Client\_dk END) AS 'age\_<=\_18',  
 *COUNT*(DISTINCT CASE WHEN cd.age BETWEEN 19 AND 30 THEN cd.Client\_dk END) AS 'age\_between\_19\_and\_30',  
 *COUNT*(DISTINCT CASE WHEN cd.age BETWEEN 31 AND 60 THEN cd.Client\_dk END) AS 'age\_between\_31\_and\_60',  
 *COUNT*(DISTINCT CASE WHEN cd.age > 60 THEN cd.Client\_dk END) AS 'age\_>\_60'  
FROM clnt\_data AS cd  
LEFT JOIN clnt\_aggr AS ca ON ca.client\_dk = cd.Client\_dk  
GROUP BY  
 ca.tb\_name

1. Вывести идентификаторы клиентов, у которых pos-оборот строго больше, чем в среднем по базе.

-- 3)  
SELECT  
 client\_dk AS client\_id  
FROM clnt\_aggr  
WHERE  
 pos\_amt > (  
 SELECT  
 *AVG*(pos\_amt)  
 FROM clnt\_aggr  
 )

1. Перед отпуском руководитель попросил вас: «вывести распределение клиентов по уровню заработной платы». Необходимо подготовить 1 слайд, который раскроет данную проблематику

-- 4)  
-- здесь вопрос в том как грамотно обработать информацию по банку  
-- скорее всего, лучше уточнить, по какому конкретно банку интересует данная сводная информация, чтобы не группировать по нему, а поместить в фильтр конкретный  
-- исходя из нижеописанных данных, получится выделить потенциальные категории клиентов, с которыми можно работать по-отдельности с разными подходами  
-- фильтры по возрасту и уровню з/п я взял из субъективной оценки по московским меркам, в конкретном случае это можно либо уточнить на уровне ТЗ,  
-- либо выводить эти цифры более "заумными" методами, рассчитав их, например, на имеющихся данных  
SELECT  
 tb\_name AS bank\_name,  
 gender,  
 *COUNT*(DISTINCT CASE WHEN cd.age <= 30 AND ca.salary <= 60000 THEN cd.Client\_dk END) AS 'young\_age\_and\_low\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age <= 30 AND ca.salary BETWEEN 60001 AND 200000 THEN cd.Client\_dk END) AS 'young\_age\_and\_middle\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age <= 30 AND ca.salary > 200000 THEN cd.Client\_dk END) AS 'young\_age\_and\_high\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age BETWEEN 31 AND 60 AND ca.salary <= 60000 THEN cd.Client\_dk END) AS 'middle\_age\_and\_low\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age BETWEEN 31 AND 60 AND ca.salary BETWEEN 60001 AND 200000 THEN cd.Client\_dk END) AS 'middle\_age\_and\_middle\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age BETWEEN 31 AND 60 AND ca.salary > 200000 THEN cd.Client\_dk END) AS 'middle\_age\_and\_high\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age > 60 AND ca.salary <= 60000 THEN cd.Client\_dk END) AS 'old\_age\_and\_low\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age > 60 AND ca.salary BETWEEN 60001 AND 200000 THEN cd.Client\_dk END) AS 'old\_age\_and\_middle\_income',  
 *COUNT*(DISTINCT CASE WHEN cd.age > 60 AND ca.salary > 200000 THEN cd.Client\_dk END) AS 'old\_age\_and\_high\_income'  
FROM clnt\_data AS cd  
LEFT JOIN clnt\_aggr AS ca ON ca.client\_dk = cd.Client\_dk  
GROUP BY  
 gender,  
 tb\_name

1. Вы наблюдаете среднюю з/п в течение года, и, в апреле-мае 2020 наблюдаете прирост этого показателя 20% по отношению к аналогичным периодам прошлого года. Как объяснить такую динамику статистики и какие гипотезы можно проверить чтобы сделать правильные выводы?

\*Подсказка: задание творческое. При выполнении задания можно создавать промежуточные таблицы - приложить код и описать графики, которые будут показаны на слайде.

Если рассматривать российские реалии, то рост средней зарплаты на 20% за такой промежуток времени кажется достаточно большим.

Во-первых, я бы проверил как обстоят дела с медианными значениями зарплат, возможно в данные попали выбросы, что могло сильно повлиять именно на средние значения. В целом провести краткий исследовательский анализ данных, потому что такие резкие изменения, особенно касающихся средних значений, часто обусловлены какими-либо нарушениями/аномалиями в данных.

Если делать предположение, что речь идет о России, то с этим периодом можно связать начало пандемии ковид-19 и многочисленным переходом трудящихся на удаленную работу, в том числе в сферу IT, где зарплаты выше, как один из возможных влияющих факторов.

В то время не наблюдалось высокой годовой инфляции, но если сделать предположение об резко меняющихся экономических обстоятельствах, то я бы обратил внимание на уровень инфляции, возможный процесс девальвации национальной валюты, и на какие-либо кризисные явления при их наличии. В таком случае можно сделать проверку гипотезы о соответствующем изменение расходов, что косвенно бы подтверждало эти процессы, если их рост был бы соизмерим. Также обратил бы внимание на динамику ключевой ставки ЦБ РФ в данном кейсе.

Если абстрагироваться от России или какой-либо реальной страны, то первое что имеет смысл перепроверить, что может 20% роста показателя является всё-таки нормой. Для этого можно взять исторические данные, допустим за последние 3-5 лет, разбить данные по месяцам, так как речь идет именно об изменении месячных показателей.

(

SELECT

TO\_CHAR(day, ‘YYYY-MM-01)::DATE AS month,

AVG(salary)

FROM …

GROUP BY

TO\_CHAR(day, ‘YYYY-MM-01)::DATE

)

И рассчитать, метрики дисперсии, стандартного отклонения чтобы оценить динамику изменений и сразу провести проверку гипотезы о равенстве средних, используя, например, t-тест. Возможно, по его результатам можно сделать вывод, что данное изменение средней зарплаты не является статистически значимым и находится в рамках нормальной флуктуации.