The final project

**The SQL block**

Using data from the tables [customer\_info.xlsx](https://docs.google.com/spreadsheets/d/1jZjcHXxFHyXf74MpTBLbHyKi-Flfx6ph/edit?gid=1066990326#gid=1066990326) (customer information) and [transactions\_info.xlsx](https://docs.google.com/spreadsheets/d/1hyHea5QGhM70FDbTCLgzVJnus558VBD1/edit?gid=173543560#gid=173543560) (information about transactions for the period from 06/01/2015 to 06/01/2016), you need to withdraw:

CREATE TABLE sales\_data (

date\_new DATE,

id\_check BIGINT,

id\_client BIGINT,

count\_products NUMERIC,

sum\_payment NUMERIC

);

CREATE TABLE customer\_info (

id\_client BIGINT,

total\_amount NUMERIC,

gender CHAR(1),

age INT,

count\_city INT,

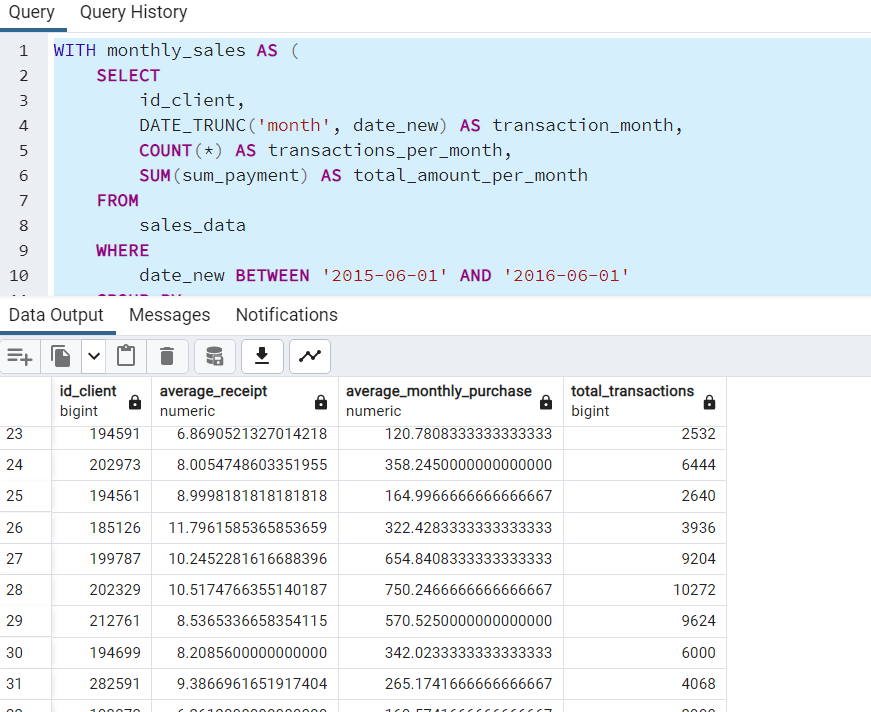
response\_communication INT,

communication\_3month INT,

tenure INT

);

1. a list of clients with a continuous history for the year, that is, every month on a regular basis without omissions for the specified annual period, the average receipt for the period from 06/01/2015 to 06/01/2016, the average amount of purchases per month, the number of all transactions per client for the period;



WITH monthly\_sales AS (

SELECT

id\_client,

DATE\_TRUNC('month', date\_new) AS transaction\_month,

COUNT(\*) AS transactions\_per\_month,

SUM(sum\_payment) AS total\_amount\_per\_month

FROM

sales\_data

WHERE

date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

id\_client, DATE\_TRUNC('month', date\_new)

),

continuous\_clients AS (

SELECT

id\_client

FROM

monthly\_sales

GROUP BY

id\_client

HAVING

COUNT(DISTINCT transaction\_month) = 12 -- Ensures every month is covered

)

SELECT

c.id\_client,

AVG(s.sum\_payment) AS average\_receipt, -- Average receipt per transaction

AVG(monthly.total\_amount\_per\_month) AS average\_monthly\_purchase, -- Average monthly amount

COUNT(s.id\_check) AS total\_transactions -- Total transactions in the period

FROM

continuous\_clients cc

JOIN

customer\_info c ON cc.id\_client = c.id\_client

JOIN

sales\_data s ON c.id\_client = s.id\_client

JOIN

monthly\_sales monthly ON monthly.id\_client = c.id\_client

WHERE

s.date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

c.id\_client;

2. information by month:

the average amount of the check per month;

A screenshot of a computer

Description automatically generated

average number of operations per month;

A screenshot of a computer

Description automatically generated

the average number of clients who performed transactions;

A screenshot of a computer

Description automatically generated

the share of the total number of transactions for the year and the share per month of the total amount of transactions;

A screenshot of a computer

Description automatically generated

print the % ratio of M/F/NA in each month with their share of costs;

A screenshot of a computer

Description automatically generated

WITH gender\_totals AS (

SELECT

DATE\_TRUNC('month', s.date\_new) AS transaction\_month,

COUNT(s.id\_check) AS total\_transactions,

SUM(s.sum\_payment) AS total\_payment

FROM

sales\_data s

WHERE

s.date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

transaction\_month

),

gender\_data AS (

SELECT

DATE\_TRUNC('month', s.date\_new) AS transaction\_month,

c.gender,

COUNT(s.id\_check) AS gender\_count,

SUM(s.sum\_payment) AS gender\_total

FROM

sales\_data s

JOIN

customer\_info c ON s.id\_client = c.id\_client

WHERE

s.date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

transaction\_month, c.gender

)

SELECT

g.transaction\_month,

SUM(CASE WHEN g.gender = 'M' THEN g.gender\_count ELSE 0 END)::DECIMAL / NULLIF(SUM(g.gender\_count), 0) \* 100 AS male\_percentage,

SUM(CASE WHEN g.gender = 'F' THEN g.gender\_count ELSE 0 END)::DECIMAL / NULLIF(SUM(g.gender\_count), 0) \* 100 AS female\_percentage,

SUM(CASE WHEN g.gender IS NULL THEN g.gender\_count ELSE 0 END)::DECIMAL / NULLIF(SUM(g.gender\_count), 0) \* 100 AS na\_percentage,

SUM(CASE WHEN g.gender = 'M' THEN g.gender\_total ELSE 0 END)::DECIMAL / NULLIF(SUM(t.total\_payment), 0) \* 100 AS male\_cost\_share,

SUM(CASE WHEN g.gender = 'F' THEN g.gender\_total ELSE 0 END)::DECIMAL / NULLIF(SUM(t.total\_payment), 0) \* 100 AS female\_cost\_share,

SUM(CASE WHEN g.gender IS NULL THEN g.gender\_total ELSE 0 END)::DECIMAL / NULLIF(SUM(t.total\_payment), 0) \* 100 AS na\_cost\_share

FROM

gender\_data g

JOIN

gender\_totals t ON g.transaction\_month = t.transaction\_month

GROUP BY

g.transaction\_month

ORDER BY

g.transaction\_month;

3. age groups of clients in increments of 10 years and separately clients who do not have this information, with the parameters amount and number of transactions for the entire period, and quarterly - averages and %.

A screenshot of a computer

Description automatically generated

WITH age\_groups AS (

SELECT

CASE

WHEN c.age IS NULL THEN 'Unknown'

WHEN c.age < 20 THEN '0-19'

WHEN c.age < 30 THEN '20-29'

WHEN c.age < 40 THEN '30-39'

WHEN c.age < 50 THEN '40-49'

WHEN c.age < 60 THEN '50-59'

WHEN c.age < 70 THEN '60-69'

WHEN c.age < 80 THEN '70-79'

ELSE '80+'

END AS age\_group,

COUNT(s.id\_check) AS total\_transactions,

SUM(s.sum\_payment) AS total\_amount

FROM

sales\_data s

JOIN

customer\_info c ON s.id\_client = c.id\_client

WHERE

s.date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

age\_group

)

SELECT

age\_group,

total\_transactions,

total\_amount,

ROUND(total\_amount::DECIMAL / NULLIF(total\_transactions, 0), 2) AS avg\_amount\_per\_transaction

FROM

age\_groups

ORDER BY

age\_group;

-- Quarterly averages and percentages

WITH quarterly\_data AS (

SELECT

DATE\_TRUNC('quarter', s.date\_new) AS transaction\_quarter,

CASE

WHEN c.age IS NULL THEN 'Unknown'

WHEN c.age < 20 THEN '0-19'

WHEN c.age < 30 THEN '20-29'

WHEN c.age < 40 THEN '30-39'

WHEN c.age < 50 THEN '40-49'

WHEN c.age < 60 THEN '50-59'

WHEN c.age < 70 THEN '60-69'

WHEN c.age < 80 THEN '70-79'

ELSE '80+'

END AS age\_group,

COUNT(s.id\_check) AS quarterly\_transactions,

SUM(s.sum\_payment) AS quarterly\_amount

FROM

sales\_data s

JOIN

customer\_info c ON s.id\_client = c.id\_client

WHERE

s.date\_new BETWEEN '2015-06-01' AND '2016-06-01'

GROUP BY

transaction\_quarter, age\_group

)

SELECT

transaction\_quarter,

age\_group,

AVG(quarterly\_transactions) AS avg\_transactions,

AVG(quarterly\_amount) AS avg\_amount,

(SUM(quarterly\_transactions) \* 100.0 / SUM(SUM(quarterly\_transactions)) OVER ()) AS percentage\_of\_total\_transactions,

(SUM(quarterly\_amount) \* 100.0 / SUM(SUM(quarterly\_amount)) OVER ()) AS percentage\_of\_total\_amount

FROM

quarterly\_data

GROUP BY

transaction\_quarter, age\_group

ORDER BY

transaction\_quarter, age\_group;

**Блок SQL**

Используя данные таблиц [customer\_info.xlsx](https://docs.google.com/spreadsheets/d/1nTJNrbLNE4XmIwHkKQhqp-7nUf9Y7qIl/edit?usp=sharing&ouid=117439300426549762698&rtpof=true&sd=true) (информация о клиентах) и [transactions\_info.xlsx](https://docs.google.com/spreadsheets/d/13UHT4zRitnFgVWn8nQGAVHYXui5tJu9m/edit?usp=sharing&ouid=117439300426549762698&rtpof=true&sd=true) (информация о транзакциях за период с 01.06.2015 по 01.06.2016), нужно вывести:

1. список клиентов с непрерывной историей за год, то есть каждый месяц на регулярной основе без пропусков за указанный годовой период, средний чек за период с 01.06.2015 по 01.06.2016, средняя сумма покупок за месяц, количество всех операций по клиенту за период;
2. информацию в разрезе месяцев:
3. средняя сумма чека в месяц;
4. среднее количество операций в месяц;
5. среднее количество клиентов, которые совершали операции;
6. долю от общего количества операций за год и долю в месяц от общей суммы операций;
7. вывести % соотношение M/F/NA в каждом месяце с их долей затрат;
8. возрастные группы клиентов с шагом 10 лет и отдельно клиентов, у которых нет данной информации, с параметрами сумма и количество операций за весь период, и поквартально - средние показатели и %.

**The Excel block**

1. Combine the data of 2 sheets in the [Тестирование Аналитик](https://docs.google.com/spreadsheets/d/1tXftgaNbYwQu-Tx6qvdYNdn5Iz1IJMA8/edit?gid=2099745552#gid=2099745552) - List1 and List2.
2. Using the resulting array, create a summary table in the context of weeks and Territories.
3. Find the top 3 territories by share in total turnover and the top 3 territories by turnover to the warehouse for the last week.
4. Calculate the margin in % and the profitability in% of the turnover in all territories.
5. Create a schedule with the daily dynamics of turnover and profitability in % in general for all territories.

**Блок Excel**

1. Совместить данные 2-х листов в файле [Тестирование Аналитик](https://docs.google.com/spreadsheets/d/1EASuUO61AlLsWE-bdFCHuE8A2Ic9teYu/edit?usp=sharing&ouid=117439300426549762698&rtpof=true&sd=true)  - Лист1 и Лист2.
2. Используя полученный массив сформировать сводную таблицу в разрезе недель и Территорий.
3. Найти топ-3 территорий по доле в общем товарообороте и топ-3 территорий по товарообороту на склад за последнюю неделю.
4. Рассчитать наценку в % и доходность в % от товарооборота по всем территориям.
5. Сформировать график с понедельной динамикой товарооборота и доходности в % в целом по всем территориям.

**Power BI block**

* Task 1

You need to make a report in Power BI based on the attached tables and send the .pbix file. Link to the tables:

[Выгрузка для отчета Power BI](https://drive.google.com/drive/u/0/folders/1Vr45uuIuaSt1akSQBxTsdKuzbnnzKcvj)

Tables:

- warehouses

- products

- orders

- order\_lines

\*The tables contain random data, but it is worth taking them as actual.

All processing needs to be done inside Power BI - leave the source files as they are.

Requirements for the report - imagine that this is all the data that exists at all and you are making the first report in the company. Your task is to make the most of this data. Which indicators to calculate, which visualizations to choose and how to organize the report are **your choice.**

**Bonus:**

**If you find something interesting (anomalies, trends, dependencies, etc.), formalize your thoughts in the form of a comment on a separate sheet of the report.**

* Task 2

Write an SQL query on the database from task 1, which will display a list of those users who bought any animal feed for the period August 1-15, except "Kitekat cat food, with rabbit in sauce, 85 g". Attach it in a text document.

* Task 3

Write an SQL query that will display a list of the top 5 most common products in user orders in St. Petersburg for the period August 15-30. Attach it in the same text document where you wrote the request from the previous paragraph.

**Блок Power BI**

* Задание 1

Нужно на основе приложенных таблиц сделать отчет в Power BI и отправить файл *.pbix*. Ссылка на таблицы:

<https://drive.google.com/drive/u/1/folders/1NKr6UriAvwIedyKspHhNdyHyAmfmYEJY>

Таблицы:

- warehouses

- products

- orders

- order\_lines

\*В таблицах рандомные данные, но стоит воспринимать их как фактические.

Всю обработку нужно делать внутри Power BI - исходные файлы оставьте как есть.

**Требования к отчету - представьте, что это все данные, которые вообще есть и вы делаете первый в компании отчет. Ваша задача выжать из этих данных максимум пользы. Какие показатели посчитать, какие визуализации выбрать и как организовать отчет - ваш выбор.**

**Бонус:**

**Если вы найдете что-то интересное (аномалии, тренды, зависимости, etc.) - оформите свои мысли в виде комментария на отдельном листе отчета.**

* Задание 2

Напишите SQL запрос по базе данных из задания 1, который выведет список тех пользователей, которые купили за период 1-15 августа 2 любых корма для животных, кроме "Корм Kitekat для кошек, с кроликом в соусе, 85 г". Приложите его в текстовом документе.

* Задание 3

Напишите SQL запрос, который выведет список топ 5 самых часто встречающихся товаров в заказах пользователей в СПб за период 15-30 августа. Приложите его в том же текстовом документе, где вы написали запрос из предыдущего пункта.

**The Python block**

You have been given the [data.csv](https://drive.google.com/file/d/1VHr3q9qMvVzJDxC-6iB-KPf885-lrU3p/view).

Answer all the questions in the file:

[Блок 3](https://drive.google.com/drive/u/0/folders/1JqA-2QkspYAI01p3lwYu7S9z-VRDLO5d)

**Блок Python**

Вам даны данные [data.csv](https://drive.google.com/file/d/1VHr3q9qMvVzJDxC-6iB-KPf885-lrU3p/view).

Ответьте на все вопросы в файле:

[Блок 3](https://drive.google.com/drive/u/0/folders/1JqA-2QkspYAI01p3lwYu7S9z-VRDLO5d)

**The test task block**

Download the [Тестовое задание](https://docs.google.com/document/d/15ygIJFoP5EAeaUo-O3zQxKS7CVzwoXl2/edit#heading=h.gjdgxs) file in the format.docx and answer the questions (in the file itself).

**Блок тестовое задание**

Скачайте файл [Тестовое задание](https://docs.google.com/document/d/1kNtb_vzmey29bd3e6Fw2I3vMmNOQUe80/edit?usp=drive_link&ouid=117439300426549762698&rtpof=true&sd=true) в формате .docx и ответьте на вопросы (в самом файле).

**Design of the project**

**Create a repository for each block on**  [github](https://github.com)**, where you will upload all your files that you will have at the end of each block. Attach links to the repositories in a text document or in the comments to the lesson on the moodle. As a result, you will have 5 links: SQL, Power BI, Excel, Python, and a test task.**

**Оформление проекта**

Создайте репозиторий для каждого блока на [github](https://github.com), куда вы зальете все ваши файлы, которые у вас будут по завершению каждого блока. Приложите ссылки на репозитории текстовым документом или в комментарии к уроку на мудле. По итогу у вас будет 5 ссылок: SQL, Power BI, Excel, Python, Тестовое задание.

**Total:**

**After completing all the blocks, you need to attach a single zip file on the moodle:**

**- SQL file with your queries**

**- Excel file with completed tasks**

**- PBIX report file, as well as a PDF version of the report and a text document with two SQL queries.**

**- IPYNB file with your code for the third block**

**- A word document with answers to the test questions**

**Also, don't forget to send links to the repositories to github.**

**Итого:**

**По завершению всех блоков, вам необходимо прикрепить одним zip файлом на в мудле:**

**- SQL файл с вашими запросами**

**- Excel файл с выполненными заданиями**

**- PBIX файл отчета, а также PDF версию отчета и текстовый документ с двумя SQL запросами.**

**- IPYNB файл с вашим кодом для третьего блока**

**- Word документ с ответами на вопросы теста**

**Также не забудьте отправить ссылки на репозитории в github.**