**Проект №27 “Решение бизнес-задач с помощью SQL”**

**Работу выполнил студент группы SDA160 Николай Голомбовский**

**ЛЕГЕНДА**

Слухи о ваших успехах на аналитическом поприще вышли за пределы вашей гейм-студии.

Благодаря этому вы получили возможность поучаствовать в стороннем проекте для одной e-commerce-компании. Пока коллеги погрузились в решение выявленных вами проблем, вы решили воспользоваться временным затишьем и взяться за этот проект — не сидеть же без дела!

Итак, в услугах аналитика нуждается молодая, но перспективная компания Pens and Pencils. Компания базируется в США и там же осуществляет свою деятельность. С марта 2017 года её профиль — B2B и B2C-продажи офисных товаров (оргтехника, канцтовары и мебель) онлайн.

Вы отправили резюме и сопроводительное письмо, и вас уже готовы допустить к проектной работе, осталось только выполнить тестовое задание. Что ж, для вас это не проблема.

**ЗАДАЧА**

В рамках тестового задания вас просят проанализировать работу компании с точки зрения её эффективности и дать рекомендации по масштабированию бизнеса, а именно в каком штате лучше открыть офлайн-магазин.

**КОНКРЕТНЫЕ ШАГИ (ФОРМАЛИЗОВАННАЯ ЗАДАЧА)**

1. Оценить динамику продаж и распределение выручки по товарам.
2. Составить портрет клиента, а для этого — выяснить, какие клиенты приносят больше всего выручки.
3. Проконтролировать логистику компании (определить, все ли заказы доставляются в срок и в каком штате лучше открыть офлайн-магазин).

**ПОЕХАЛИ…**

**Задание 27.5.1**

**Описание задачи:**

Напишите запрос, который выведет сумму выручки по месяцам:

date (месяц заказа) ― тип date;

revenue (объём выручки) ― округлите значения до целых с помощью round. Отсортируйте запрос по дате заказа.

**SQL-запрос для выгрузки нужных данных:**

– Здесь достаточно простой запрос и, честно сказать, я не знаю что нужно комментировать

select

to\_char(sd.order\_date, 'YYYY-MM-01') date, –определение месяца

round(sum(sc.quantity \* sp.price \* (1 - sc.discount))) revenue – формула расчёта дохода

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

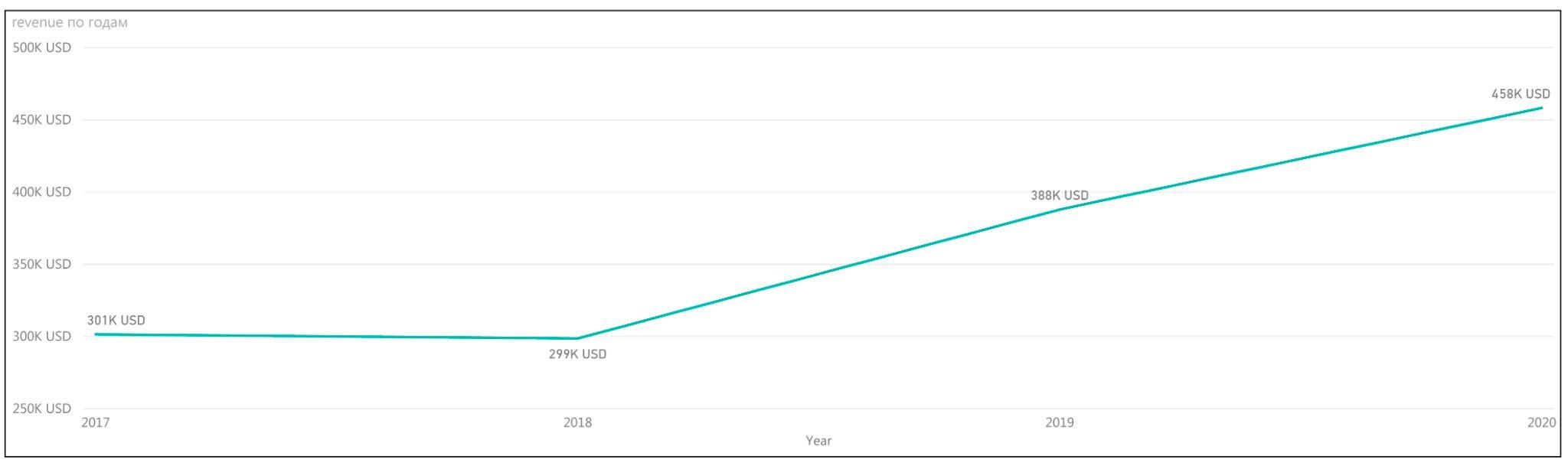
group by 1

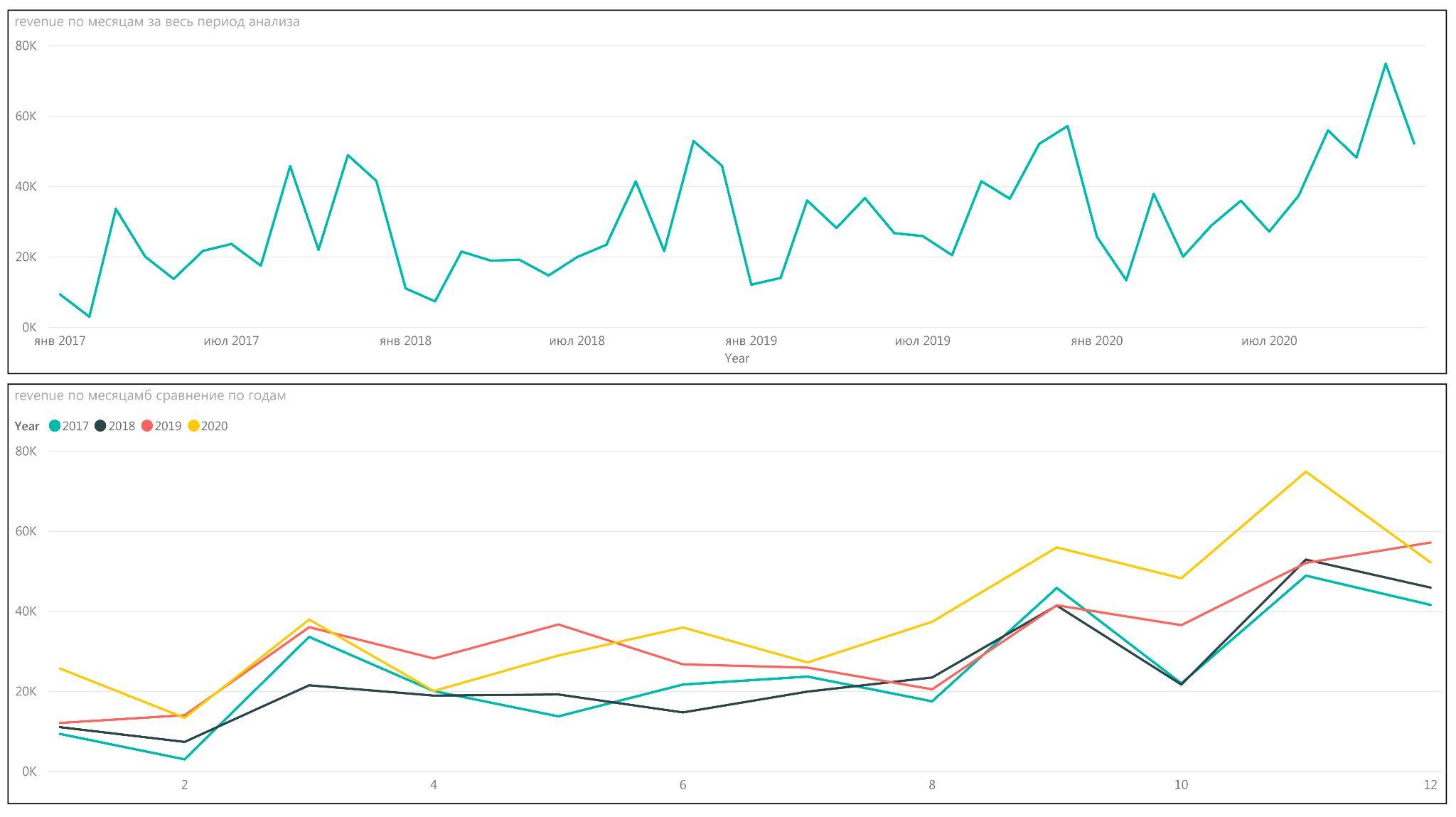
order by 1

**Результат работы запроса (данные упорядочены по иному в таблицу для удобства восприятия):**

| **Revenue** | **Янв** | **Фев** | **Мар** | **Апр** | **Май** | **Июн** | **Июл** | **Авг** | **Сен** | **Окт** | **Ноя** | **Дек** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2017** | $9 394 | $3 040 | $33 638 | $20 108 | $13 784 | $21 739 | $23 720 | $17 536 | $45 852 | $22 021 | $48 931 | $41 632 |
| **2018** | $11 088 | $7 400 | $21 547 | $18 971 | $19 253 | $14 764 | $19 969 | $23 499 | $41 462 | $21 726 | $52 948 | $45 949 |
| **2019** | $12 141 | $14 067 | $36 057 | $28 246 | $36 745 | $26 777 | $25 959 | $20 535 | $41 531 | $36 558 | $52 114 | $57 208 |
| **2020** | $25 729 | $13 416 | $37 962 | $20 114 | $28 979 | $35 985 | $27 247 | $37 409 | $55 987 | $48 273 | $74 906 | $52 241 |

**Визуализация результатов с помощью Power BI:**

****



**Выводы:**

1. Годовая выручка компании “Pens and Pencils” последние 2 года растёт - на 30% в 2019 году и на 18% в 2020.
2. Сравнение месячной выручки разных лет показывает явное наличие сезонности в продажах.

**Задание 27.5.5**

**Описание задачи:**

Напишите запрос, который выведет сумму выручки по различным категориям и подкатегориям:

* category (категория товара);
* subcategory (подкатегория товара);
* revenue (объём выручки) ― округлите до целых с помощью round.

Отсортируйте запрос по убыванию выручки.

**SQL-запрос для выгрузки нужных данных:**

– Так же весьма простой запрос, не требующий, на мой взгляд, каких то пояснений

select

sp.category,

sp.subcategory,

round(sum(sc.quantity \* sp.price \* (1 - sc.discount))) revenue

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

group by 1, 2

order by 3 desc

**Результат работы запроса**

| **category** | **subcategory** | **revenue** |
| --- | --- | --- |
| Furniture | Chairs | $235 318 |
| Technology | Phones | $221 110 |
| Office Supplies | Storage | $179 736 |
| Technology | Accessories | $125 442 |
| Furniture | Tables | $114 532 |
| Technology | Machines | $113 978 |
| Technology | Copiers | $95 844 |
| Office Supplies | Paper | $62 148 |
| Furniture | Bookcases | $59 271 |
| Office Supplies | Appliances | $55 550 |
| Furniture | Furnishings | $51 695 |
| Office Supplies | Binders | $45 312 |
| Office Supplies | Supplies | $37 457 |
| Office Supplies | Art | $22 148 |
| Office Supplies | Envelopes | $13 809 |
| Office Supplies | Labels | $10 345 |
| Office Supplies | Fasteners | $2 459 |

**Дополнительные результаты**

| **category** | **revenue** |
| --- | --- |
| Technology | $556 375 |
| Furniture | $460 817 |
| Office Supplies | $428 964 |

**Выводы:**

1. Ну то, что товарная категория “Technology” вышла на 1 место по продажам - в этом ничего удивительного и нет.
2. А вто что на самом деле сильно удивляет, так это то что в подкатегориях с очень большим отрывом выиграли “офисные кресла” и обычные “телефоны”! И это сегодня - в век компьютерных технологий! Что бы это могло значить - таков рынок или таковы клиенты именно нашей компании?

**Задание 27.5.6**

**Описание задачи:**

Напишите запрос, который выведет данные топ-25 товаров по объёму выручки в следующем формате:

* product\_nm — наименование товара;
* revenue — объём выручки;
* quantity — количество проданных товаров; Не забудьте просуммировать количество!
* percent\_from\_total — доля от общей выручки в процентах. Не забудьте умножить на 100!
* cost\_per\_unit - стоимость единицы продукта. Это я добавил от себя - мне кажется, там можно будет найти интересную особенность.

Примечание. Значения revenue и percent\_from\_total необходимо округлить до двух знаков после разделителя. Включите решение этого задания в итоговый документ. При необходимости снабдите дополнительными комментариями.

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который включаем поля всех нужных таблиц.

– Плюс создаём в данном CTE дополнительную колонку с общей суммой всех продаж для удобства последующего расчёта доли каждого продукта

with add\_total\_revenue\_column as

(

select

\*,

sum(sc.quantity \* sp.price \* (1 - sc.discount)) over() as total\_revenue

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

)

– Подсчитываем нужные нам параметры

select

atr.product\_nm,

round(sum(atr.quantity \* atr.price \* (1 - atr.discount)), 2) revenue,

sum(atr.quantity) quantity,

round((sum(atr.quantity \* atr.price \* (1 - atr.discount)) / avg(atr.total\_revenue)) \* 100, 2) as percent\_from\_total,

round(sum(atr.quantity \* atr.price \* (1 - atr.discount)) / sum(atr.quantity), 2) cost\_per\_unit

from add\_total\_revenue\_column atr

group by 1

order by 2 desc

limit 25

**Результат работы запроса**

| **product\_nm** | **revenue, USD** | **quantity, шт** | **percent\_from\_total, %** | **cost\_per\_unit, USD** |
| --- | --- | --- | --- | --- |
| Canon imageCLASS 2200 Advanced Copier | 36959.82 | 20 | 2.56 | 1847.99 |
| HON 5400 Series Task Chairs for Big and Tall | 15309.53 | 39 | 1.06% | 392.55 |
| 3D Systems Cube Printer, 2nd Generation, Magenta | 14299.89 | 11 | 0.99 | 1299.99 |
| High Speed Automatic Electric Letter Opener | 13624.21 | 11 | 0.94 | 1238.56 |
| Martin Yale Chadless Opener Electric Letter Opener | 13325.00 | 22 | 0.92 | 605.68 |
| Cisco TelePresence System EX90 Videoconferencing Unit | 11319.24 | 6 | 0.78 | 1886.54 |
| Hewlett Packard LaserJet 3310 Copier | 11303.69 | 38 | 0.78 | 297.47 |
| Samsung Galaxy Mega 6.3 | 11154.87 | 39 | 0.77 | 286.02 |
| Canon PC1060 Personal Laser Copier | 9295.83 | 19 | 0.64 | 489.25 |
| HP Designjet T520 Inkjet Large Format Printer - 24" Color | 9187.50 | 12 | 0.64 | 765.63 |
| Global Troy Executive Leather Low-Back Tilter | 9082.87 | 32 | 0.63 | 283.84 |
| Honeywell Enviracaire Portable HEPA Air Cleaner for 17' x 22' Room | 9043.55 | 40 | 0.63 | 226.09 |
| Tennsco 6- and 18-Compartment Lockers | 8740.17 | 43 | 0.60 | 203.26 |
| Plantronics CS510 - Over-the-Head monaural Wireless Headset System | 8657.89 | 35 | 0.60 | 247.37 |
| SAFCO Arco Folding Chair | 8100.95 | 53 | 0.56 | 152.85 |
| Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish | 7805.48 | 24 | 0.54 | 325.23 |
| Apple iPhone 5 | 7798.00 | 24 | 0.54 | 324.92 |
| Tennsco Double-Tier Lockers | 7704.86 | 47 | 0.53 | 163.93 |
| Ativa V4110MDD Micro-Cut Shredder | 7699.89 | 11 | 0.53 | 699.99 |
| Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back | 7446.44 | 51 | 0.51 | 146.01 |
| Tennsco Single-Tier Lockers | 7326.59 | 28 | 0.51 | 261.66 |
| Office Star - Professional Matrix Back Chair with 2-to-1 Synchro Tilt and Mesh Fabric Seat | 7019.50 | 28 | 0.49 | 250.70 |
| Zebra ZM400 Thermal Label Printer | 6965.70 | 6 | 0.48 | 1160.95 |
| Hot File 7-Pocket, Floor Stand | 6510.77 | 49 | 0.45 | 132.87 |
| Bretford Rectangular Conference Table Tops | 6497.82 | 46 | 0.45 | 141.26 |

**Выводы:**

1. Как и ожидалось, в TOP25 продуктов попали техника и кресла.
2. Кажется, что имеется корреляции между стоимостью единицы товара и долей продаж данного товара..Было бы не плохо провести анализ данной корреляции - ведь гипотеза подразумевает, что больше покупают самые дорогие товары. Что врят ли может иметь место в реальной жизни.

**Задание 27.6.1**

**Описание задачи:**

Напишите запрос, который выведет количество клиентов и выручку по категориям клиента:

* category (категория клиента);
* cust\_cnt (количество клиентов);
* revenue (объём выручки) ― округлите до целых с помощью round.

Отсортируйте запрос по убыванию выручки.

**SQL-запрос для выгрузки нужных данных:**

– Вначале соединяем все нужные таблицы с помощью join и проводим на основе полученной объединённой таблицы вычисление нужных данных.

select

scu.category,

count (distinct scu.cust\_id) cust\_cnt,

round(sum(sc.quantity \* sp.price \* (1 - sc.discount))) revenue

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

group by 1

order by 1 desc

**Результат работы запроса   
(три последние колонки добавлены с помощью средств Google таблицы для облегчения анализа полученных данных)**

| **category** | **cust\_cnt** | **revenue** | **Доля от количества** | **Доля revenue** | **revenue на 1 клиента** |
| --- | --- | --- | --- | --- | --- |
| Corporate | 645 | $1 172 009 | 81% | 81% | $1 817 |
| Consumer | 148 | $274 148 | 19% | 19% | $1 852 |

**Выводы:**

1. Категория B2B “Corporate” занимает 81% как по количеству, так и по объёму закупок от всех клиентов и всех закупок.
2. Гораздо более поразителен тот факт, объём закупок на одного клиента в категории B2C на 2% больше, чем в категории B2B. Как такое может быть в реальной жизни - я затрудняюсь ответить.На основе данного факта склонен сделать вывод об искусственности анализируемых данных.

**Задание 27.6.2**

**Описание задачи:**

Напишите запрос, который выведет количество новых корпоративных клиентов по месяцам:

* month (месяц) ― тип date;
* new\_custs (количество новых клиентов);
* revenue (суммарные покупки данных новых клиентов, добавил сам данную колонку для более интересного анализа);
* revenue\_per\_customer (покупки на 1 клиента, добавил сам данную колонку для более интересного анализа).

Отсортируйте запрос по первому столбцу в порядке возрастания.

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который для каждого клиента считаем месяц первой закупки и сумму всех его закупок.Конечно же, только для корпоративных клиентов.

– На основе полученного CTE проводим расчёт нужных параметров, группируя по месяцам

with first\_purchase\_per\_customer as

(

select

sd.cust\_id,

to\_char(min(sd.order\_date), 'YYYY-MM-01') date,

sum(sc.quantity \* sp.price \* (1 - sc.discount)) revenue

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

where scu.category = 'Corporate'

group by 1

order by 3 desc

)

select

date,

count(distinct(cust\_id)) new\_custs,

round(sum(revenue)) revenue,

round(sum(revenue) / count(distinct(cust\_id))) revenue\_per\_customer

from first\_purchase\_per\_customer

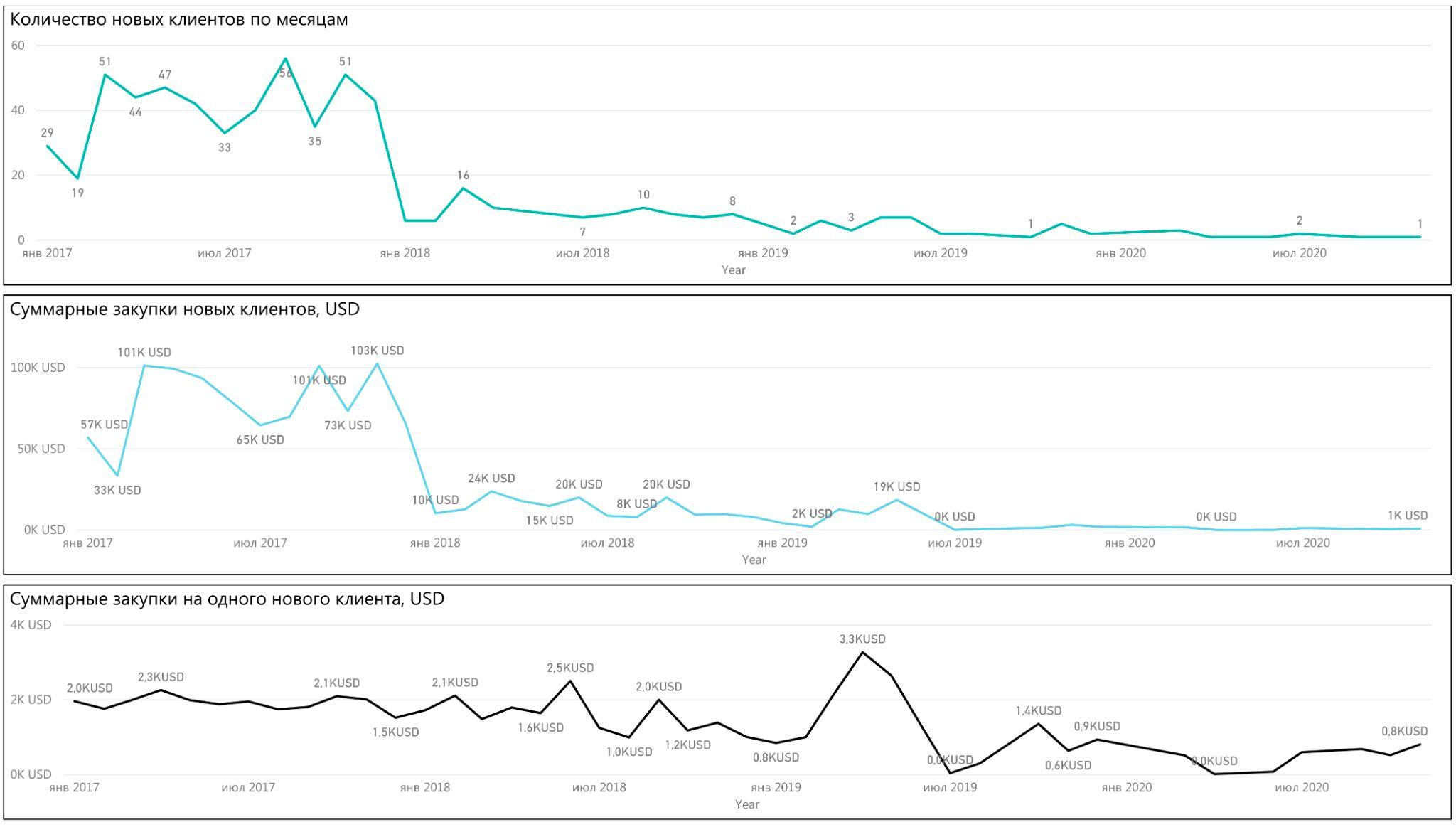
group by 1

order by 1

**Результат работы запроса   
(последняя колонка добавлена по желанию автора данного документа)**

| **date** | **new\_custs** | **revenue** | **revenue\_per\_customer** |
| --- | --- | --- | --- |
| 2017-01-01 | 29 | $56 912 | $1 962 |
| 2017-02-01 | 19 | $33 426 | $1 759 |
| 2017-03-01 | 51 | $101 341 | $1 987 |
| 2017-04-01 | 44 | $99 429 | $2 260 |
| 2017-05-01 | 47 | $93 530 | $1 990 |
| 2017-06-01 | 42 | $78 946 | $1 880 |
| 2017-07-01 | 33 | $64 529 | $1 955 |
| 2017-08-01 | 40 | $69 816 | $1 745 |
| 2017-09-01 | 56 | $101 208 | $1 807 |
| 2017-10-01 | 35 | $73 365 | $2 096 |
| 2017-11-01 | 51 | $102 505 | $2 010 |
| 2017-12-01 | 43 | $65 379 | $1 520 |
| 2018-01-01 | 6 | $10 329 | $1 721 |
| 2018-02-01 | 6 | $12 664 | $2 111 |
| 2018-03-01 | 16 | $23 755 | $1 485 |
| 2018-04-01 | 10 | $17 930 | $1 793 |
| 2018-05-01 | 9 | $14 795 | $1 644 |
| 2018-06-01 | 8 | $20 029 | $2 504 |
| 2018-07-01 | 7 | $8 733 | $1 248 |
| 2018-08-01 | 8 | $7 933 | $992 |
| 2018-09-01 | 10 | $19 992 | $1 999 |
| 2018-10-01 | 8 | $9 434 | $1 179 |
| 2018-11-01 | 7 | $9 710 | $1 387 |
| 2018-12-01 | 8 | $8 061 | $1 008 |
| 2019-01-01 | 5 | $4 223 | $845 |
| 2019-02-01 | 2 | $2 002 | $1 001 |
| 2019-03-01 | 6 | $12 657 | $2 110 |
| 2019-04-01 | 3 | $9 819 | $3 273 |
| 2019-05-01 | 7 | $18 516 | $2 645 |
| 2019-06-01 | 7 | $9 139 | $1 306 |
| 2019-07-01 | 2 | $76 | $38 |
| 2019-08-01 | 2 | $600 | $300 |
| 2019-10-01 | 1 | $1 356 | $1 356 |
| 2019-11-01 | 5 | $3 174 | $635 |
| 2019-12-01 | 2 | $1 872 | $936 |
| 2020-03-01 | 3 | $1 537 | $512 |
| 2020-04-01 | 1 | $11 | $11 |
| 2020-06-01 | 1 | $75 | $75 |
| 2020-07-01 | 2 | $1 193 | $596 |
| 2020-09-01 | 1 | $681 | $681 |
| 2020-10-01 | 1 | $518 | $518 |
| 2020-11-01 | 1 | $807 | $807 |

**Визуализация результатов с помощью Power BI:**

****

**Выводы:**

1. Имеющиеся данные однозначно показывают, что новые клиенты в категория B2B “Corporate” появлялись практически только в 2017 году (76% от всех клиентов в анализе).   
   Доля новых клиентов в 2018-2020 годах в данной категории составляет всего 24% от общего количества клиентов и 20% от суммарных закупок всех клиентов.
2. Из этого можно вновь сделать вывод об искусственности предоставленных к анализу данных. Согласитесь, трудно допустить, что хорошо поработав в 2017 году по привлечению новых клиентов, компания “Pens and Pencils” вдруг в последующие 3 года не озаботилась столь драматичным падением новых клиентов и ничего не предприняла для исправления ситуация в течении аж 3 лет!  
   Хотя с другой стороны, мы ведь не знаем ТОЧНО когда именно появились эти новых клиенты - просто они сделали первые закупки в рамках имеющихся для анализа данных.  
   Теоретически можно же предположить, что новые клиенты 2017 года на самом деле становились клиентами “Pens and Pencils” в течение предыдущих нескольких лет.
3. Если сравнивать средний объём закупок у новых клиентов на ОДНОГО клиента, то за небольшим отличием в период март-июнь 2019 года, он плавно понижается со временем.   
   Можно сказать, мельчать стали новые клиенты.

**Задание 27.6.4.1**

**Описание задачи:**

Сколько в среднем различных товаров в заказах у корпоративных клиентов?:

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который для каждого клиента и каждого отдельного заказа данного клиента подсчитываем количество продуктов в заказе.Конечно же, только для корпоративных клиентов.

– На основе полученной таблицы рассчитываем среднее количество продуктов на 1 корзину. Для более лучшего анализа я так же добавил минимальное и максимальное количество продуктов на один заказ

with number\_of\_products\_per\_order as

(

select

sd.cust\_id,

sd.order\_id,

count(sc.product\_id) quantity\_per\_order

from sql.store\_carts sc

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

where scu.category = 'Corporate'

group by 1, 2

order by 1

)

select

round(avg(quantity\_per\_order), 1) average\_products,

min(quantity\_per\_order) min\_products,

max(quantity\_per\_order) max\_products

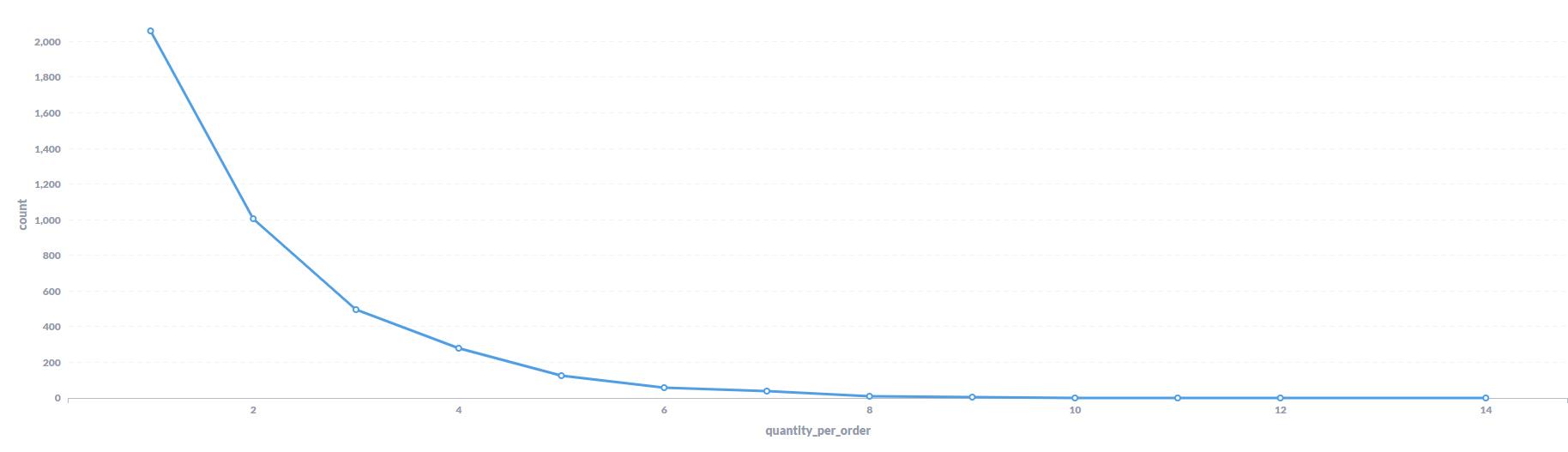
from number\_of\_products\_per\_order

**Результат работы запроса**

| **average\_products** | **min\_products** | **max\_products** |
| --- | --- | --- |
| 2.0 | 1 | 14 |

**Выводы:**

1. Анализ показывает, что корпоративные клиенты в среднем заказывают 2 продукта на один заказ. Очевидно возникает следующий вопрос - это много или мало? Вот и попытаемся хоть как то на него ответить.
2. Расчёт аналогичных показателей для B2C клиентов даёт следующие цифры - 1.96, 1, 11. Цифры весь и весьма похожи, особенно среднее количество продуктов на заказ. И в очередной раз закрадывается уверенность в искусственном происхождении анализируемых данных.
3. Помимо этого представляет большой интерес гистограмма количества заказов в зависимости от количества продуктов в корзине корпоративных клиентов.  
   Для этого во второй части вставим предыдущего SQL-запроса поменяем последнюю часть на:  
     
   select  
    quantity\_per\_order,  
    count(order\_id)  
   from number\_of\_products\_per\_order  
   group by 1  
   order by 1  
     
   Если результат его работы вывести в графическом виде, то получится весьма интересная визуализация, показывающая экспоненциальное падение количество заказов в зависимости от количества продуктов в заказе.

****

**Задание 27.6.4.2**

**Описание задачи:**

Какая в среднем сумма заказов у корпоративных клиентов?

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который для каждого клиента и каждого отдельного заказа данного клиента подсчитываем сумма заказа.Конечно же, только для корпоративных клиентов.

– Затем на основе полученной CTE рассчитываем цену среднего чека. От себя для более глубокого анализа я добавил цену минимального и максимального чеков.

with revenue\_per\_order\_per\_customer as

(

select

sd.cust\_id,

sd.order\_id,

sum(sc.quantity \* sp.price \* (1 - sc.discount)) revenue\_per\_order

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

where scu.category = 'Corporate'

group by 1, 2

order by 3 desc

)

select

round(avg(revenue\_per\_order), 1) average\_check,

round(min(revenue\_per\_order), 1) min\_check,

round(max(revenue\_per\_order), 1) max\_check

from revenue\_per\_order\_per\_customer

**Результат работы запроса**

| **average\_check, USD** | **min\_check, USD** | **max\_check, USD** |
| --- | --- | --- |
| 285.9 | 0.1 | 10 780.8 |

**Выводы:**

1. Вот получили мы цифру 285.9 USD. И вновь вопрос - это много или мало?
2. Если посчитать средний чек для B2C клиентов, но каким то удивительным образом он окажется на 5,5% выше - 301.6 USD.
3. Более удивительным является то, что для B2C клиентов размер максимального чека на 10% выше аналогичного показателя для B2B клиентов и составляет 11 820.1 USD! Как такое может быть? Что же за такие B2C клиенты у нашей компании?!

**Задание 27.6.4.3**

**Описание задачи:**

Сколько в среднем различных офисов у корпоративных клиентов?

(Поскольку данных о полном адресе у нас нет, разные адреса будем считать по различным индексам (zip\_code).)

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который для каждого клиента определяем количество различных ZIP кодов, по которым ему производилась доставка. Конечно же, только для корпоративных клиентов.

– Затем на основе полученной CTE рассчитываем среднее количество офисов у корпоративных клиентов.

with number\_offices\_per\_customer as

(

select

sd.cust\_id,

count(distinct sd.zip\_code) number\_offices

from sql.store\_delivery sd

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

where scu.category = 'Corporate'

group by 1

order by 2 desc

)

select

round(avg(number\_offices), 1) average\_offices,

min(number\_offices) min\_offices, – сам добавил расчёт данного показателя

max(number\_offices) max\_offices – сам добавил расчёт данного показателя

from number\_offices\_per\_customer

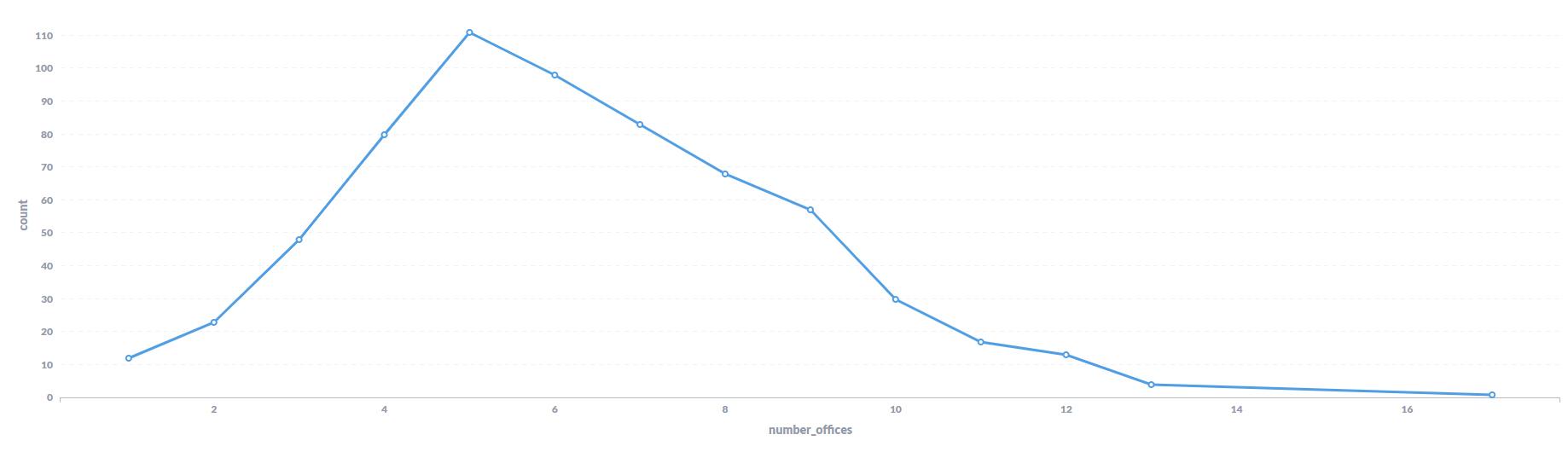
**Результат работы запроса**

**(последние две колонки добавил автор доклада в попытке найти интересные особенности в данных)**

| **average\_offices** | **min\_offices** | **max\_offices** |
| --- | --- | --- |
| 6.2 | 1 | 17 |

**Выводы:**

1. Если среднее количество офисов среди корпоративных клиентов чуть-чуть превышает 6, то почему они так мало покупают?
2. Количество вопросов только множится, особенно если вычислить тьак же аналогичный показатель для B2C клиентов, который составляет 6.0.  
   Но ведь можно сравнить максимальное количество офисов у B2C и B2B клиентов - 12 против 17. Вот мне просто интересно, что это за B2C клиент, для которого доставку проводят по 12 различным адресам? Это его разные квартиры, разные члены семьи, но при этом закупками для них всех занимается один человек? Вопросы, вопросы, вопросы…
3. Довольно интересно смотрится гистограмма распределения количества корпоративных клиентов в зависимости от количества офисов.Кривая гистограммы имеет явный куполообразный контор, что говорит о случайном характере данного распределения. Ну хоть что-то в предоставленных нам данных выглядит естественно!



В задании так же просят сделать выводы и рекомендации на основе всех данных в рамках вопросов Задания 27.6.4. Я не совсем понимаю, что именно от нас ожидается, так как данных с гулькин нос и у нас нет маркетинговой информации о рынке, чтобы адекватно соотнести полученные данные с реальностью.

Могу сказать одно - либо у нас B2B клиенты слабые по сравнению с B2C клиентами, либо последние у нас какие то супер крутые, что выглядят на уровне B2B клиентов. Либо нам нам есть куда расти по объёмам закупок среди B2B клиентов, либо мы можем гордиться отделом продаж по работе с B2C клиентами.

**Задание 27.7.1**

**Описание задачи:**

Напишите запрос, который выведет:

* тип доставки;
* общее количество заказов (orders\_cnt);
* количество заказов, которые не были доставлены вовремя (late\_orders\_cnt);
* долю выполненных вовремя заказов, в процентах (% success), округлённую до двух знаков после запятой.

Отсортируйте запрос по четвёртому столбцу в порядке возрастания.

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который для каждого типа доставки через оконные функции определяем количество доставок и количество НЕ удачных доставок. Количество записей CTE равно количеству доставок.

– Затем на основе полученной CTE рассчитываем требуемые в задании показатели. Используем GROUP BY чтобы схлопнуть количество записей в таблице к количеству разных типов доставки.

with prom\_table as

(

select

sd.ship\_mode,

count(sd.order\_id) over(partition by sd.ship\_mode) orders\_cnt,

count(sd.order\_id) filter

(

where ((sd.ship\_mode = 'Standard Class' and (sd.ship\_date - sd.order\_date) > 6) or

(sd.ship\_mode = 'Second Class' and (sd.ship\_date - sd.order\_date) > 4) or

(sd.ship\_mode = 'First Class' and (sd.ship\_date - sd.order\_date) > 3) or

(sd.ship\_mode = 'Same Day' and (sd.ship\_date - sd.order\_date) > 0))

) over(partition by sd.ship\_mode) late\_orders\_cnt

from sql.store\_delivery sd

)

select

ship\_mode,

orders\_cnt,

late\_orders\_cnt,

round((avg(orders\_cnt) - avg(late\_orders\_cnt)) / avg(orders\_cnt) \* 100, 2) "% success"

from prom\_table

group by 1, 2, 3

order by 4

**Результат работы запроса**

| **ship\_mode** | **orders\_cnt** | **late\_orders\_cnt** | **% success** |
| --- | --- | --- | --- |
| Second Class | 964 | 202 | 79.05 |
| Standard Class | 2994 | 309 | 89.68 |
| Same Day | 264 | 12 | 95.45 |
| First Class | 787 | 1 | 99.87 |

**Выводы:**

1. В силу того, что у меня нет цифр для сравнения, то могу сделать только общие замечания по количеству заказов для каждого типа доставки:
   1. 60% доставок приходится на стандартный тип (в течение 6 дней).
   2. Чем более жёсткие условия доставки по количеству дней - скорее всего, тем больше цена доставки. Данные показывают уменьшение количества доставок с ужесточением условий доставки.
2. Так же мне трудно что-то определённое сказать о невыполненных заказах:
   1. Почему “Стандарный Класс” доставки имеет большую долю успешных чем “Второй класс” на 10.6 пунктов? Лично для меня это загадка - куда смотрит служба доставки.
   2. Я ещё могу понять почему “Первый Класс” доставки имеет больше успешных доставок по сравнению с доставками “день-в-день” - всякое в жизни бывает.

**Задание 27.7.2 и Задание 27.7.3**

**Описание задачи:**

Напишите запрос, чтобы вывести долю заказов, отправленных вторым классом, которые были доставлены с опозданием, по кварталам.

По получившимся данным постройте график.

Какой вывод можно сделать: заказы доставляются с опозданием постоянно или только на протяжении ограниченного периода времени?

**SQL-запрос для выгрузки нужных данных:**

– Вначале создаём CTE, в который только для заказов “Second Class” по каждому кварталу (квартал определяется по дате приёма заказа) через оконные функции определяем количество доставок и количество НЕ удачных доставок.

– Затем на основе полученной CTE рассчитываем для каждого квартала требуемые в задании долю заказов, доставленных с опозданием. Используем GROUP BY чтобы схлопнуть количество записей в таблице к количеству кварталов.

with prom\_table as

(

select

to\_char(sd.order\_date, 'YYYY Q кв') quarter\_of\_shipments,

count(sd.order\_id) over(partition by sd.ship\_mode, to\_char(sd.order\_date, 'YYYY Q кв')) cnt\_order,

count(sd.order\_id) filter

( where (sd.ship\_date - sd.order\_date) > 4 )

over(partition by sd.ship\_mode, to\_char(sd.order\_date, 'YYYY Q кв')) cnt\_NOTsuccess\_order

from sql.store\_delivery sd

where sd.ship\_mode = 'Second Class'

order by 1

)

select

quarter\_of\_shipments,

round(avg(cnt\_NOTsuccess\_order) / avg(cnt\_order) \* 100, 2) "%\_NOTsuccess"

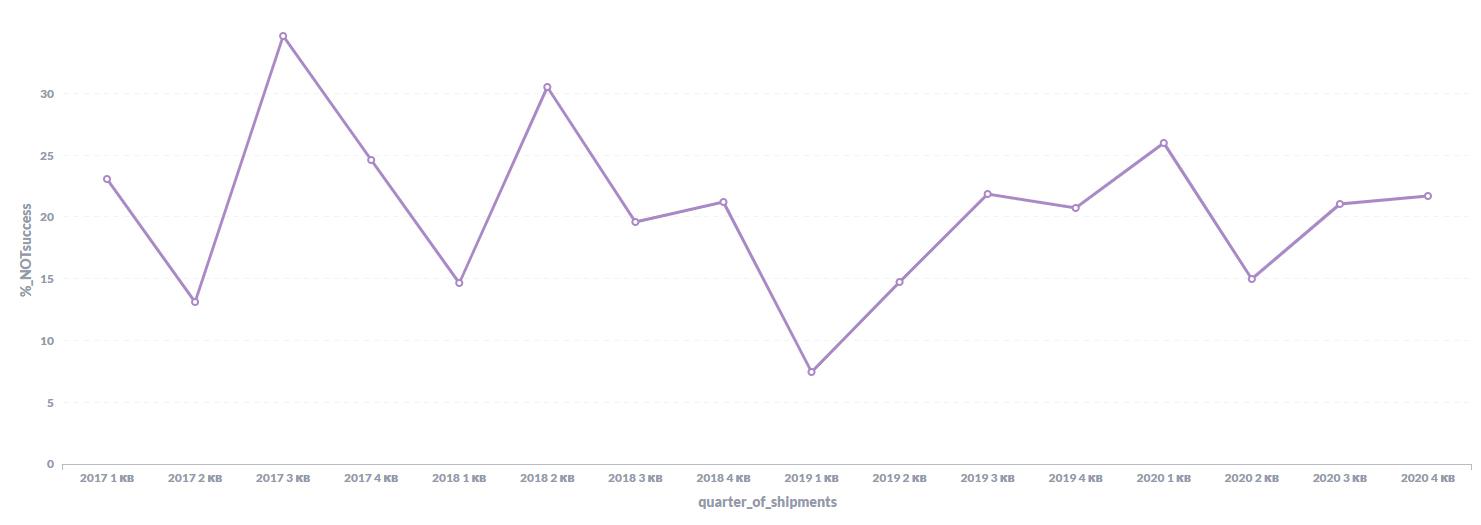
from prom\_table

group by 1

order by 1

**Результат работы запроса**

| **quarter\_of\_shipments** | **%\_NOTsuccess** |
| --- | --- |
| 2017 1 кв | 23,08 |
| 2017 2 кв | 13,16 |
| 2017 3 кв | 34,69 |
| 2017 4 кв | 24,68 |
| 2018 1 кв | 14,71 |
| 2018 2 кв | 30,56 |
| 2018 3 кв | 19,64 |
| 2018 4 кв | 21,25 |
| 2019 1 кв | 7,5 |
| 2019 2 кв | 14,81 |
| 2019 3 кв | 21,92 |
| 2019 4 кв | 20,78 |
| 2020 1 кв | 26 |
| 2020 2 кв | 15 |
| 2020 3 кв | 21,11 |
| 2020 4 кв | 21,77 |



**Выводы:**

1. Не знаю, что ожидают услышать от меня составители данного задания относительно данного графика, но могу сказать следующее:
   1. Задержки с доставкой по типу второго класса происходят ежеквартально при минимальном значении за период анализа в 7,5% и максимальным 34,7%.
   2. График показывает периодические скачки в доле задержек как вверх, так и вниз. Имеющиеся данные не позволяют однозначно сказать с чем именно связаны данные скачки.
2. Аналогичный график для “стандартного типа” доставки показывает, что подобные заказы доставляются так же постоянно с опозданием. При этом доля “опоздавших” заказов имеет незначительную тенденцию к увеличению с годами.



**Задание 27.7.6**

**Описание задачи:**

Напишите запрос, который выведет количество доставок по штатам и визуализируйте результат с помощью карты в Metabase.

Какой штат является наиболее перспективным для открытия офлайн-магазина? Почему?

**SQL-запрос для выгрузки нужных данных:**

– Даже не знаю, что по данному запросу можно прокомментировать - самый простой из возможных запросов с группировкой по штату.

select

sd.state,

count(sd.order\_id)

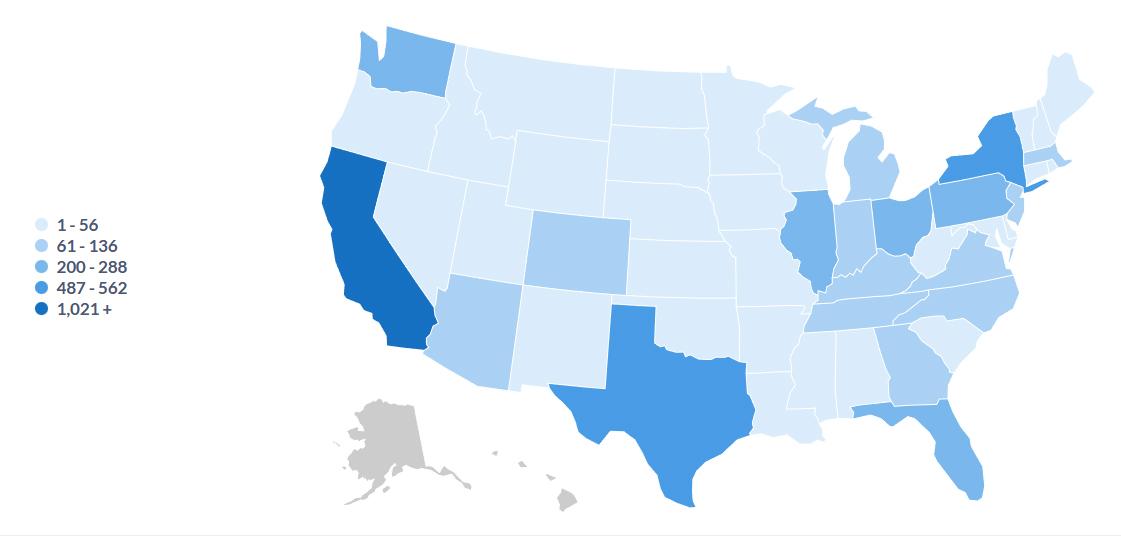
from sql.store\_delivery sd

group by 1

order by 2 desc

**Результат работы запроса**

| **state** | **count** |
| --- | --- |
| California | 1021 |
| New York | 562 |
| Texas | 487 |
| Pennsylvania | 288 |
| Illinois | 276 |
| Washington | 256 |
| Ohio | 236 |
| Florida | 200 |
| North Carolina | 136 |
| Michigan | 117 |
| Virginia | 115 |
| Arizona | 108 |
| Georgia | 91 |
| Tennessee | 91 |
| Colorado | 79 |
| Indiana | 73 |
| Massachusetts | 62 |
| New Jersey | 61 |
| Kentucky | 61 |
| Oregon | 56 |
| Wisconsin | 53 |
| Maryland | 46 |
| Connecticut | 45 |
| Minnesota | 44 |
| Delaware | 44 |
| Alabama | 34 |
| Oklahoma | 34 |
| Missouri | 30 |
| Arkansas | 27 |
| Mississippi | 26 |
| Utah | 26 |
| Rhode Island | 25 |
| Nebraska | 23 |
| Nevada | 23 |
| New Mexico | 22 |
| Louisiana | 21 |
| South Carolina | 20 |
| New Hampshire | 17 |
| Iowa | 17 |
| Kansas | 14 |
| Idaho | 11 |
| Montana | 8 |
| Vermont | 6 |
| South Dakota | 5 |
| District of Columbia | 4 |
| Maine | 3 |
| West Virginia | 2 |
| North Dakota | 2 |
| Wyoming | 1 |



**Выводы:**

1. Если тупо смотреть на полученные данные, то офлайн магазин вроде бы как нужно открывать в California, New York, Texas и так далее по списку и уменьшению количества доставок нашей компании.
2. Однако очевидно, что нужно учесть и финансовую составляющую - оценить доход, который мы получаем по каждому штату. Вышеприведённые 3 штата в таком же порядке являются и лидерами по продажам. А вот начиная с 4 штата и далее порядок по количеству доставок и объёму продаж не совпадает.
3. Так же не маловажным является показатель прибыли. К сожалению, имеющиеся данные не предоставляют нам такую информацию. Вполне может оказаться, что порядок штатов даже среди тройки лидеров может быть иной по получаемой прибыли.
4. Следующими факторами несомненно нужно учесть стоимость содержания офлайн магазина, престижность района, логистические цепочки.
5. В общем выбор города и штата для открытия офлайн магазина - это комплексный учёт различных показателей.
6. Дополнительно я хотел бы привлечь внимание к следующему. Беглый взгляд на вышеприведённую карту показывает, что мы в большей степени проводим доставки по краям страны, чем в центре и на севере. Вот лично мне было бы интересно понять с чем это связано - так исторически сложилось, по краям страны меньше офлайн магазинов, там менее активны наши конкуренты или что-то иное.

**ABC-анализ**

**Описание задачи:**

Возникло желание как то ещё по иному использовать новые знания по SQL. Я недавно ознакомился с таким анализом как ABC-анализ.

Суть его в том, что к категории A относятся самые крупные клиенты, приносящие 70% оборота, категории B - следующие 20% оборота и категории C - оставшиеся 10%.

Естественно, нужно отдельно провести данное ABC-разделение для B2B и B2C клиентов.

**SQL-запрос для выгрузки нужных данных:**

– Код SQL не оптимизирован **исключительно** с целью лучшего понимания проводимых вычислений

– Я просто пошаговая провожу анализ с помощью последовательного создания CTE

– Сначала из таблицы всех клиентов создаю таблицу с номером клиента, его категорией и суммой его закупок за весь период анализа

with prom\_table\_1 as

(

select

sd.cust\_id,

sum(sc.quantity \* sp.price \* (1 - sc.discount)) revenue\_per\_customer,

scu.category

from sql.store\_carts sc

join sql.store\_products sp on sc.product\_id = sp.product\_id

join sql.store\_delivery sd on sd.order\_id = sc.order\_id

join sql.store\_customers scu on scu.cust\_id = sd.cust\_id

group by 1, 3

order by 2 desc

),

– Сейчас к таблице добавляю колонки, рассчитанные с помощью оконных функций, с количеством клиентов в каждой категории, задаю порядковый номер клиента в каждой категории по уменьшению объёма закупок и общую сумму закупок каждой категории

prom\_table\_2 as

(

select

\*,

count(cust\_id) over(partition by category) number\_of\_customers,

row\_number() over(partition by category order by revenue\_per\_customer desc) customer\_number,

sum(revenue\_per\_customer) over(partition by category) total\_revenue

from prom\_table\_1 pt1

),

– Добавляю к таблице накопительную сумму закупок по порядковому номеру клиента в каждой категории

prom\_table\_3 as

(

select

\*,

sum(revenue\_per\_customer)

over(partition by category order by customer\_number) up\_to\_revenue

from prom\_table\_2 pt2

),

– Расставляю ранги каждому клиенту (A, B или C) по указанным в задании выше критериям

prom\_table\_4 as

(

select

\*,

case when up\_to\_revenue / total\_revenue <= 0.7 then 'A'

when up\_to\_revenue / total\_revenue <= 0.9 then 'B'

else 'C'

end abc\_number

from prom\_table\_3 pt3

)

– Группирую для каждой категории клиентов по A, B и C, вывожу для них количество клиентов и их доля от всех клиентов в категории

select

category,

abc\_number,

count(cust\_id),

count(cust\_id) \* 100 / min(number\_of\_customers) part\_of\_count

from prom\_table\_4 pt4

group by 1, 2

order by 1, 2

**Результат работы запроса**

| **category** | **abc\_number** | **count, к-во** | **part\_of\_count, %** |
| --- | --- | --- | --- |
| Consumer | A | 53 | 35 |
| Consumer | B | 41 | 27 |
| Consumer | C | 54 | 36 |
| Corporate | A | 254 | 39 |
| Corporate | B | 170 | 26 |
| Corporate | C | 221 | 34 |

**Выводы:**

1. Скажу честно, я ожидал увидеть несколько иные цифры. Что-то типа, 20% клиентов дают 80% оборота и так далее.   
   Здесь же получается, что для B2B 39% клиентов дают 80% закупок и для B2C 35% клиентов дают те же 80% закупок.
2. Так же не понятно почему количество клиентов A и C в каждой категории примерно одинаково.
3. В общем, чем больше я анализирую эти данные, тем больше я склоняюсь к мысли что данный датасет создан искусственно по определённым параметрам. В результате если анализировать по несколько иным параметрам, то появляются неожиданные результаты, которые исходя из общей логики датасета, не должны существовать в реальности.