Сначала создаем дата фрейм через pd.read\_csv, потом заливаем его в greenplum.

C помощью sql преобразуем две таблицы в одну, также сразу выведем некоторые дополнительные поля.

drop table if exists orders\_x\_customers;

create table orders\_x\_customers as

select

    a.id as order\_id

    ,a.order\_date::date

    ,a.ship\_mode

    ,a.customer\_id

    ,a.sales

    ,b.segment

    ,b.state

    ,b.city

    ,min(order\_date::date) over (partition by customer\_id) as first\_order\_date

    ,row\_number() over (partition by customer\_id order by order\_date) as order\_number

    ,max(order\_date::date) over (partition by customer\_id order by order\_date rows between current row and 1 following) - order\_date::date as days\_before\_next\_order

from orders a

    inner join customers b

        on a.customer\_id = b.id

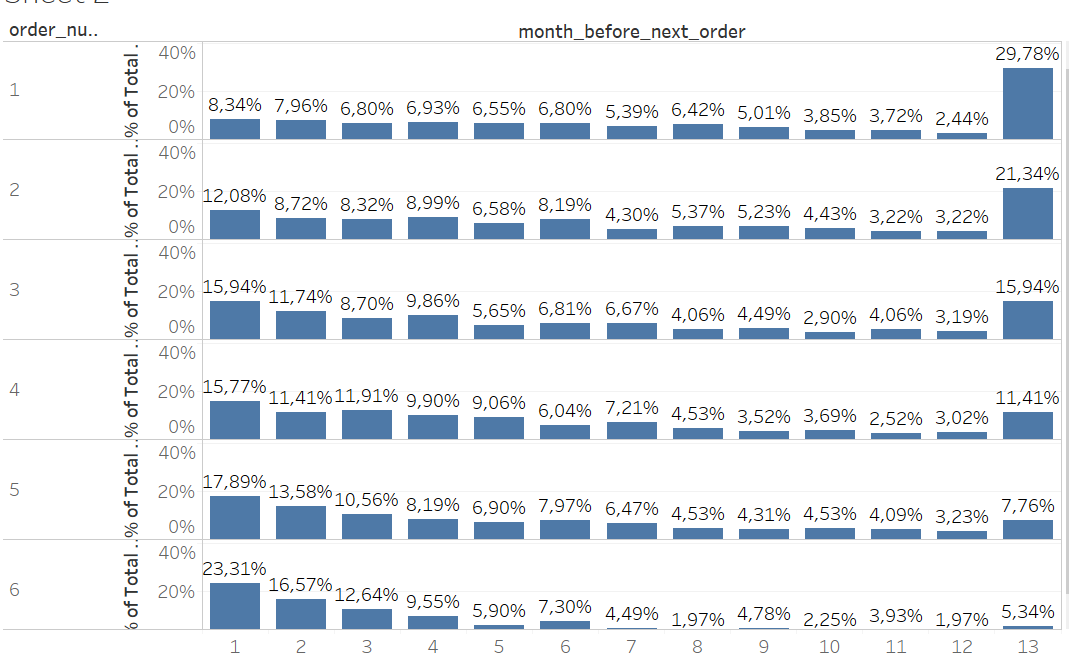
Графики я строила по полученной таблице orders\_x\_customers.

В данных есть покупки за период 01.2014-12.2017. Я подумала, что сравнить нужно ситуацию «сейчас» (то есть конец 2017г) и год назад (конец 2016г).

Сначала я посмотрела распределение кол-ва дней между покупками у тех, кто возвращается, чтобы оценить, как часто клиенты совершают покупки и какой промежуток времени нужно отложить. Например, если взять данные за ноябрь 2017 и смотреть, как возвращались эти клиенты, можно получить низкие возвраты, но не потому что стало хуже, а просто потому что прошел только месяц и многие могли не успеть купить еще раз. То есть сравнивать возвраты тех, кто сделал покупку год назад и прямо сейчас (декабрь 2017) некорректно, так как мы не имеем будущих данных.

Я посмотрела распределение – кол-во заказов в зависимости от кол-ва месяцев между текущим и следующим заказом (в расчет брались только заказы, после которых обязательно был еще заказ).

Так же время между заказами может меняться в зависимости от порядкового номера заказа, поэтому распределения представлены также в зависимости от order\_number.



На этом графике числом 13 обозначены заказы, у которых разница между текущим и следующим составила более 12 месяцев.

На основании этих распределений можно увидеть, что время между покупками уменьшается, с увеличением порядкового номера покупки, все больше людей покупают в течение первого месяца после прошлого заказа и все меньше возвращаются только через 12+ месяцев.

(Здесь не совсем корректно делать такой вывод на всех имеющихся данных, так как если кто-то сделал первый заказ в 2016, второй в 2017, то он просто не успел сделать третий заказ через год, чтоб попасть в наши данные, поэтому в третьем заказе он не учтется, то есть доля тех, кто покупает раньше может увеличиться только за счет того, что те, кто покупают редко не успели сделать покупку. Но я фильтровала даты первых заказов, распределение практически не меняется, даже если оставить только тех, кто совершил первый заказ в 2015, поэтому, можно считать, что это приближено к реальности)

Таким образом, после первого заказа, только 70% (из тех кто возвращается) делают второй в течение года. Поэтому если мы рассмотрим тех, кто сделал первый заказ в октябре-ноябре 2017 года, у них не будет достаточно времени совершить второй.

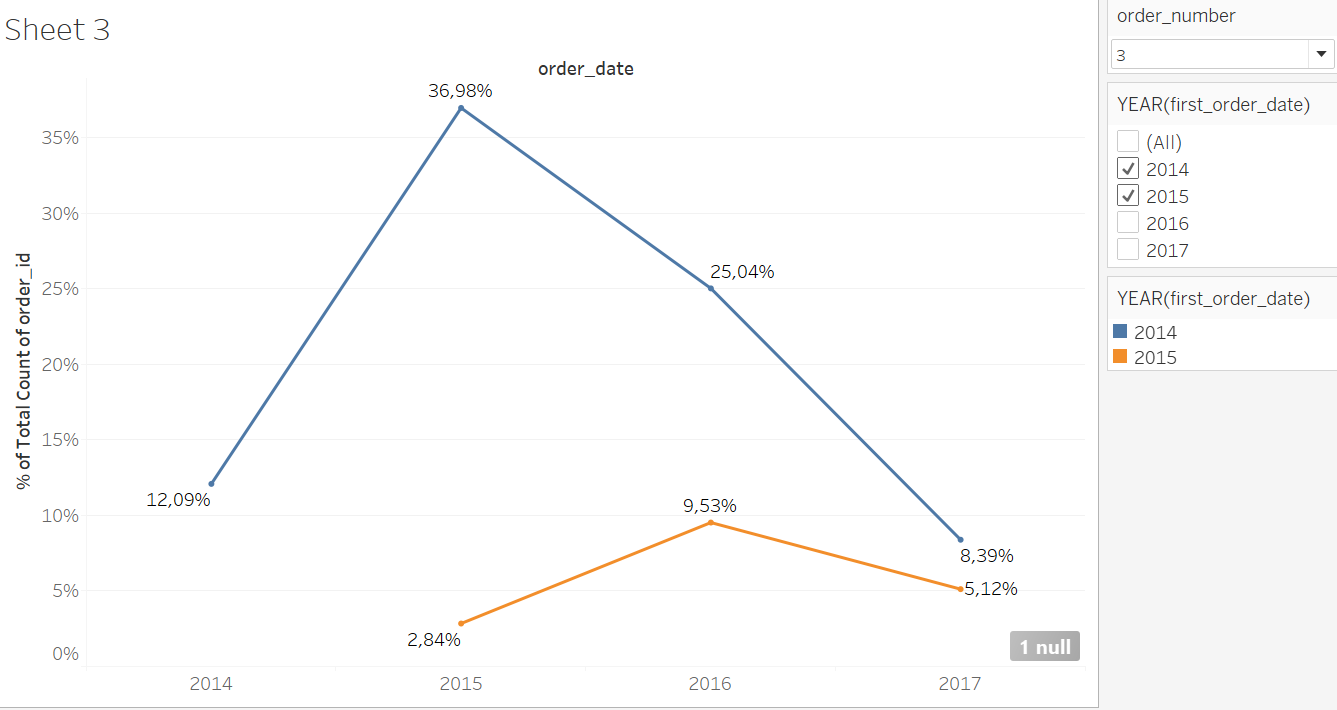
Я хотела сравнивать тех, кто сделал первый заказ в первой половине 2016 года с теми, кто сделал первый заказ в первой половине 2017 года. И посмотреть, какая доля из них совершила 2-3 покупку. Но новых клиентов в 2016 и 2017 (в 2017) году настолько мало, что эти данные точно не будут стат значимыми и на их основании нельзя сделать выводы, поэтому будем смотреть старых клиентов.



На этих графиках видно, какая доля людей от сделавших первую покупку делает потом вторую, третью и тд.

На этом графике видно, что те, кто пришел в 2014 году больше возвращаются, чем те, кто пришел в 2015, но это не обязательно связано с ухудшениями, на самом деле вторую покупку совершили все 100% в обоих случаях, а дальше разница в третьей и дальнейших покупках может заключаться в том, что пользователи 2015 года прожили на год меньше и могли просто не успеть совершить следующую покупку.

На графике ниже видно, что 9% тех, кто совершил 1 покупку в 2014 и совершил 3 покупку, сделали это только спустя 3 года – в 2017. Если среди тех, кто совершил первую покупку в 2015, тоже 9% совершат 3 покупку в 2018, то и доля совершивших третью покупку увеличилась бы с 90% до 99% и стала бы больше, чем доля тех, кто совершил 3 покупку из 2014 года.



Таким образом, имеющихся данных недостаточно, чтобы подтвердить гипотезу, ухудшения в возвратах и следующих покупках заметны, но они вполне могут быть связаны с тем, что люди пришедшие позже, еще просто не успели сделать новые заказы.

Инструментом для принятия решения является sql-запрос и воркбук в табло, когда будут новые данные, руководитель сможет самостоятельно обновить графики и проанализировать ситуацию.