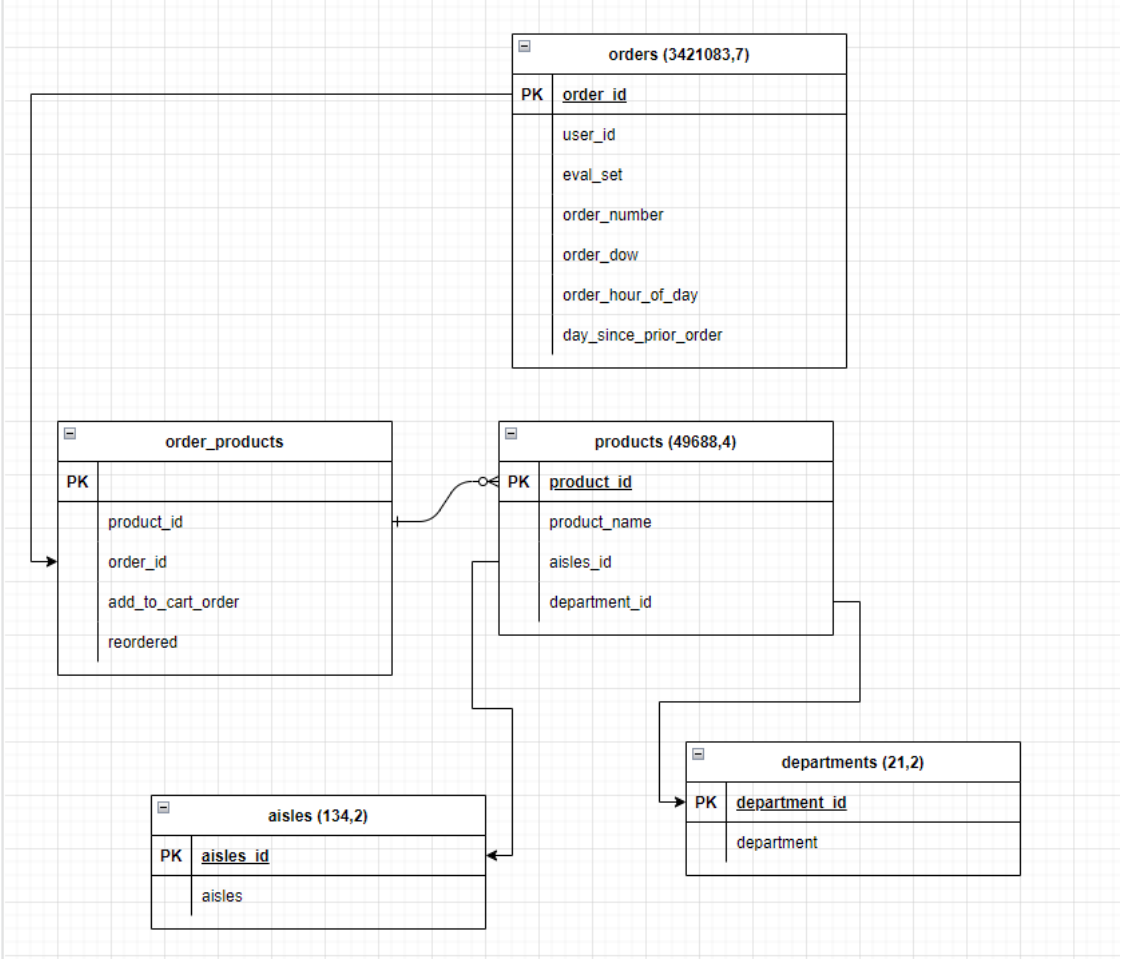
```
aisles_df = pd.read_csv('aisles.csv')
departments_df = pd.read_csv('departments.csv')
orders_df = pd.read_csv('orders.csv')
products_df = pd.read_csv('products.csv')
order_products_train = pd.read_csv('order_products__train.csv')
order_products_prior = pd.read_csv('order_products__prior.csv')
```



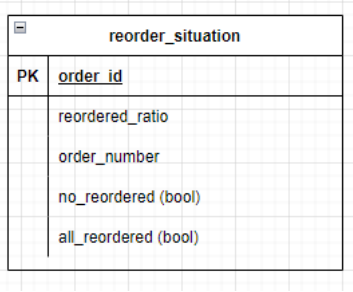
1.四张表leftjoin 通过各自的id，利用order的unique值求出总order数

```
order_products = order_products.merge(products_df, on='product_id', how='left').merge(orders_df, on='order_id', how='left').merge(depa
```

2.把order_products_train 和proir 进行concat 得到order_products表，通过提取reordered value数量，除以总order数，就是reorder的占比

3.通过标记有reorder的order为1，没有的为0，计算reorder占比，以及标记是否为存在reorder的order，以及是否全部reorder，形成新的表格如下图：

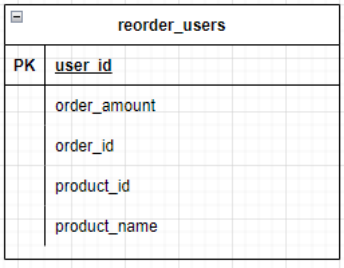
	order_id	reordered_ratio	order_number	no_reordered	all_reordered
0	1	0.500000	4	False	False
1	2	0.666667	3	False	False
2	3	1.000000	16	False	True
3	4	0.923077	36	False	False
4	5	0.807692	42	False	False



4.利用order_product的表，按照order_id进行groupby，之后count每个order中的product_id的数量，就能得到每个order的product种类数量，进而统计出购买不同数量product的order分布情况，得出结论是购买5件商品之内的订单占大多数

5.分析用户行为

1）首先在order_products表中找到所有order_number不为1的用户id，这些就是至少买过一次以上的用户。对order_id进行groupby操作，对该表中每个unique的user 的reorder列求平均值，对order_number列进行count。生成新的表格：




```
Count of users who always order the same products every time: 685
Orders of user 48242:
Count of his orders: 16
Order number473900:
['Natural Artesian Bottled Water']
————————
Order number2569948:
['Natural Artesian Bottled Water']
————————
Order number2884277:
['Natural Artesian Bottled Water']
————————
Order number304497:
['Natural Artesian Bottled Water']
————————
Order number643763:
['Natural Artesian Bottled Water']
————————
Order number1453178:
['Natural Artesian Bottled Water']
————————
Order number3377207:
```

2）分析购物时间的影响

从表orders分析：

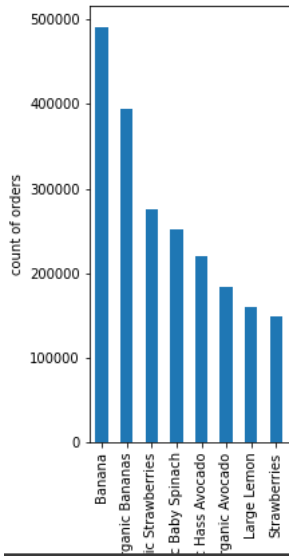
对unique order_id 按照order_dow进行count聚合操作，得到一周七天每天的订单数量

同理对order_hour_of_day, day_since_prior_order进行相同操作

- Most orders are ordered on Day 0 and Day 1.
- Orders are mostly ordered during day, from 9:00 AM to 4:00 PM.
- Peak orders happens at Saturday afternoon (1:00PM), and Sunday morning (10:00AM)
- By more than 65%, People usually buy previously ordered products from 6:00AM to 8:00AM
- Most users make orders after a week from their last order. or from a month of their last order.
- After a week from the last order, the probability of reordering within the same month is small
- The Next order has higher probability to be during 10 days from the current order.

6.分析产品购买情况

1）从order_product表中对每一个order_id的product_name进行count操作，并且降序排列



2）**How often a product is the first item purchased?**

- 5 Most Add to Cart First Products
  - Banana
  - Bag of Organic Bananas
  - Organic Whole Milk
  - Organic Strawberries
  - Organic Hass Avocado
- 3.4% of the orders, Banana is being the first product added to cart.
- Products contianing milk have very high probability to be reordered.
- Organic products have very high probability to be reordered.

从order_product表中找到所有add_to_ocart_order=1的记录，对product_name进行count操作并且降序排列，可以得到放在购物车第一位的时候购买的次数。

3）**Probability of reordering a product**