delivery status of a supply chain

Marius Gören
Alejandra Cárdenas
DSML JAN2026

Data Set

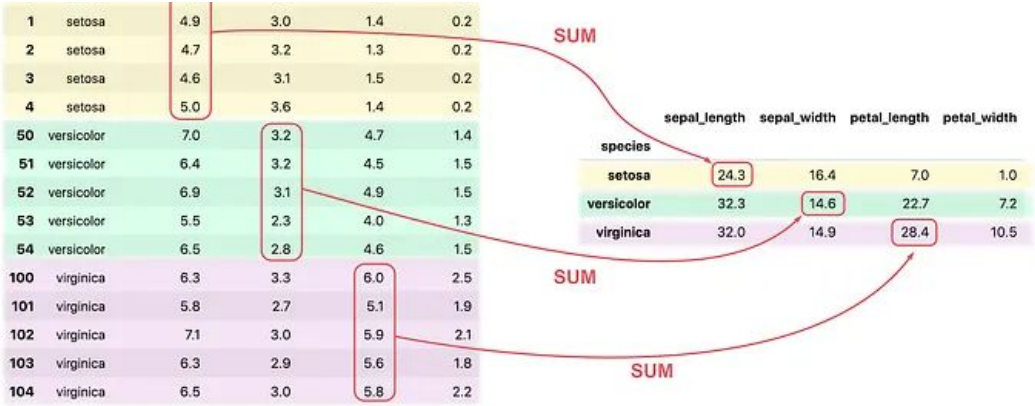
- **53 columns**
- **Difficult to understand relation**
- **Some repeated**
- **180519 rows**
- **Each 1 item -> row**

Goals

- **What affect shipment time?**
- **Origin Continent vs Delivery Status**
- **Quantity Items vs Delivery Status**

Data Cleaning

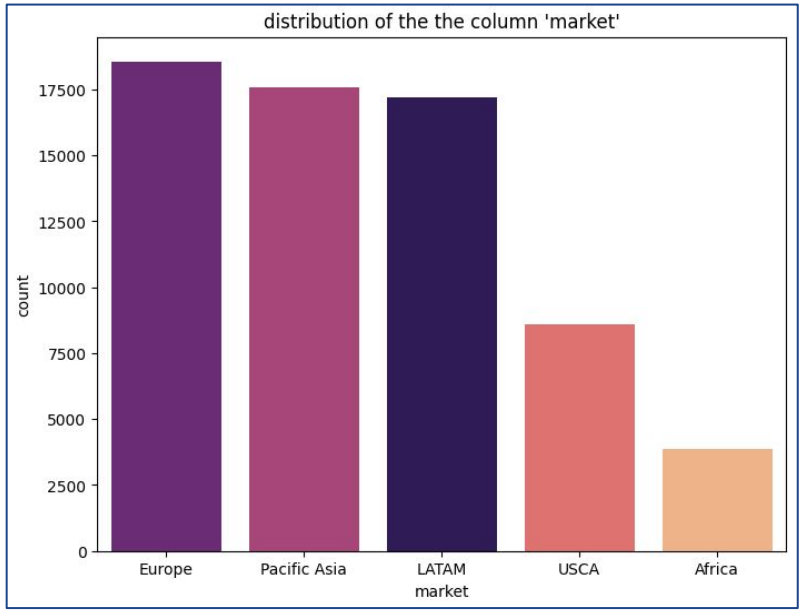
- Chose 9 columns
  - related to “delivery_status”
- Grouped by ‘Order Id’
  - 6 columns not affected
  - Sum Item Quantity
  - ΣDiscount/ΣPrice
- No Null, No duplicated, No empty spaces

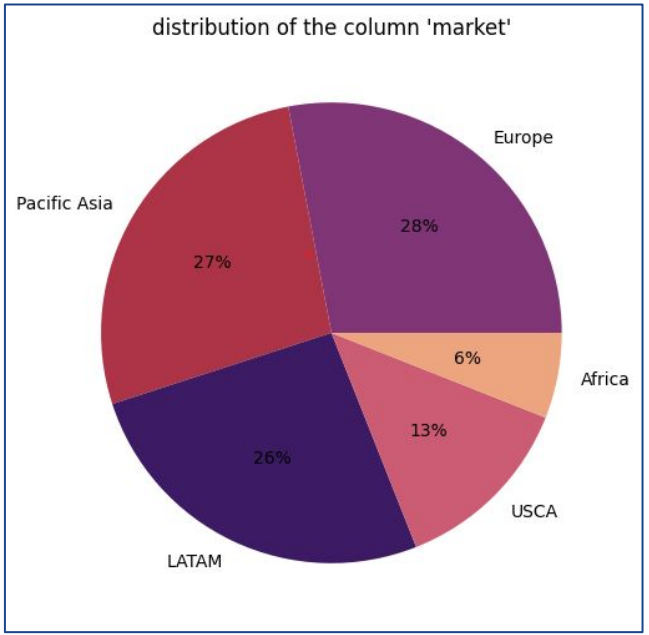


Order Id	days_for_shipping_real	days_for_shipment_scheduled	delivery_status	market	order_country	order_item_quantity	discount_ratio
1	2	4	Advance shipping	LATAM	México	1	0.200013
2	3	4	Advance shipping	LATAM	Colombia	7	0.087244
4	5	4	Late delivery	LATAM	Colombia	14	0.112853
5	6	4	Late delivery	LATAM	Colombia	10	0.126378
7	3	2	Late delivery	LATAM	Brasil	7	0.093806

univariate analysis: **“market”**

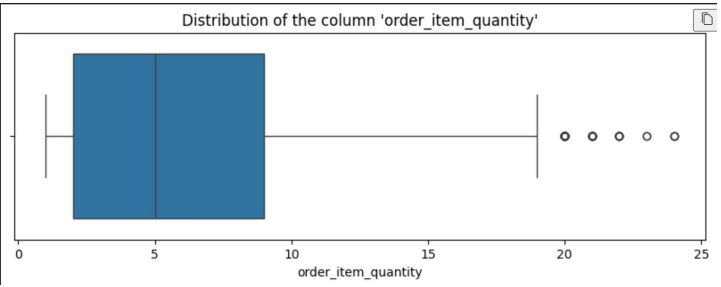
- 5 categorical nominal values
- ‘Europe’, ‘Pacific Asia’ and ‘LATAM’ are the most common markets

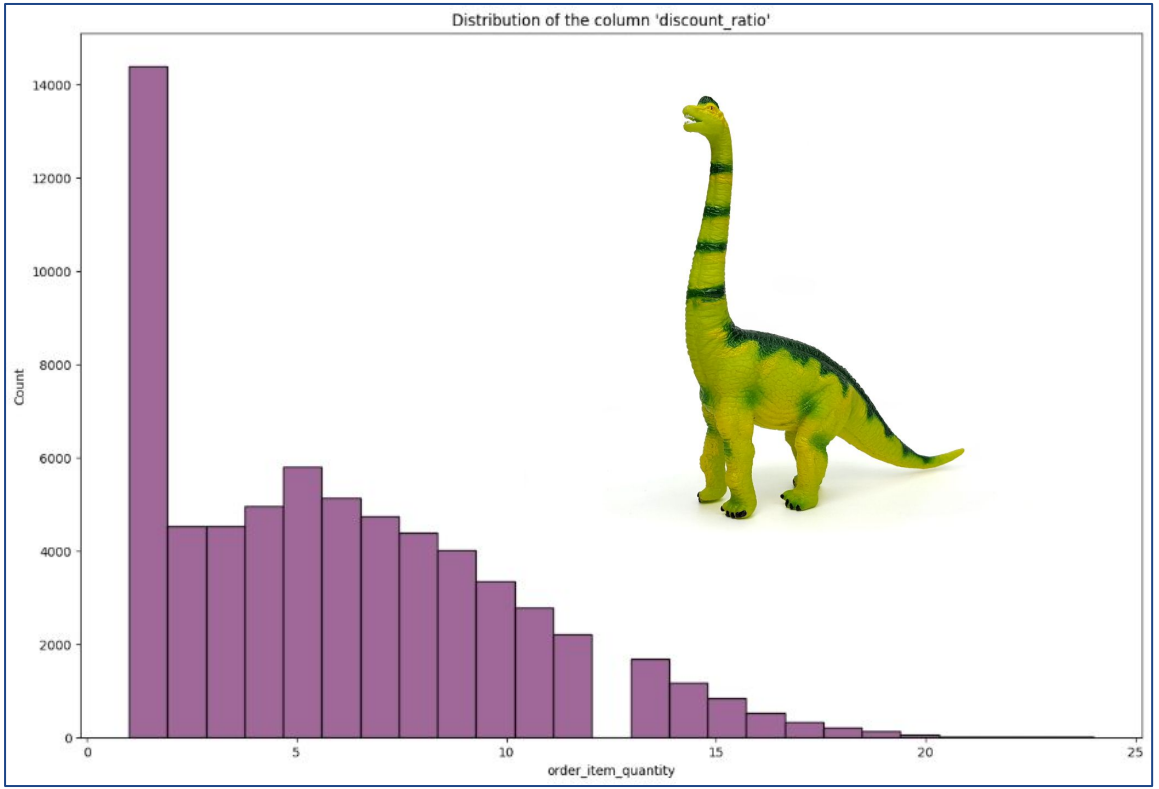




univariate analysis: **“items_per_order”**

- 24 numerical discrete values
- Most common: 1 item per order 22%
- Skewed right, platykurtic





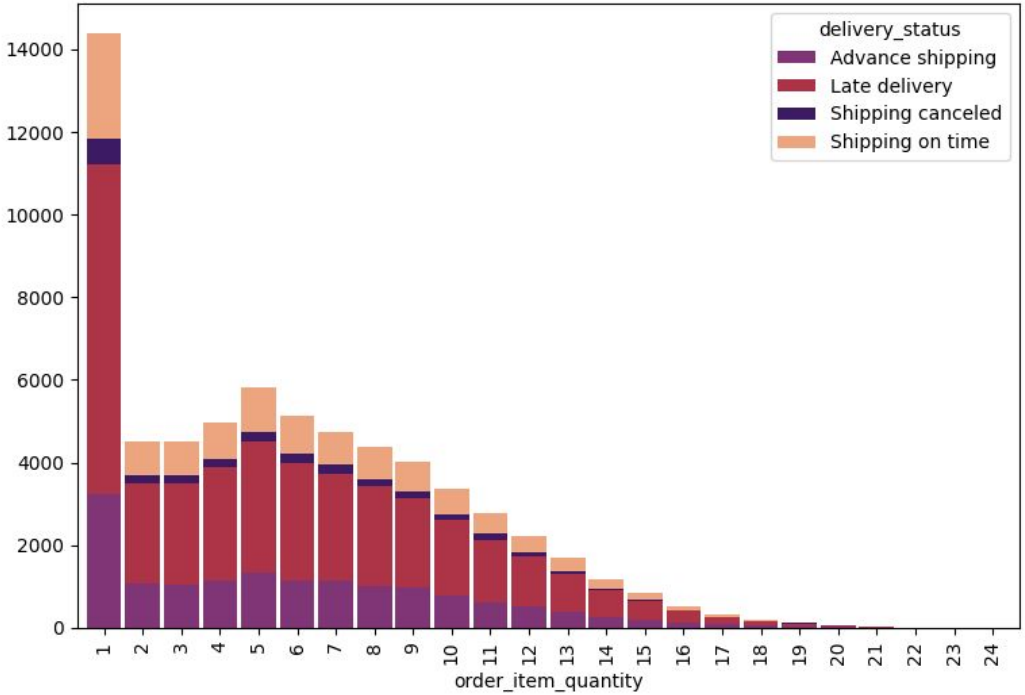
bivariate analysis: **“delivery_status” vs “items_per_order”**

The amount of items per order do not affect the delivery status

order_item_quantity	1	2	3	4	5	6	7	8	9	10
delivery_status										
Advance shipping	3247	1077	1050	1125	1332	1147	1126	995	974	794
Late delivery	7968	2401	2444	2750	3181	2830	2609	2423	2169	1805
Shipping canceled	633	220	200	210	233	226	203	187	161	137
Shipping on time	2543	819	825	868	1060	937	797	786	716	612

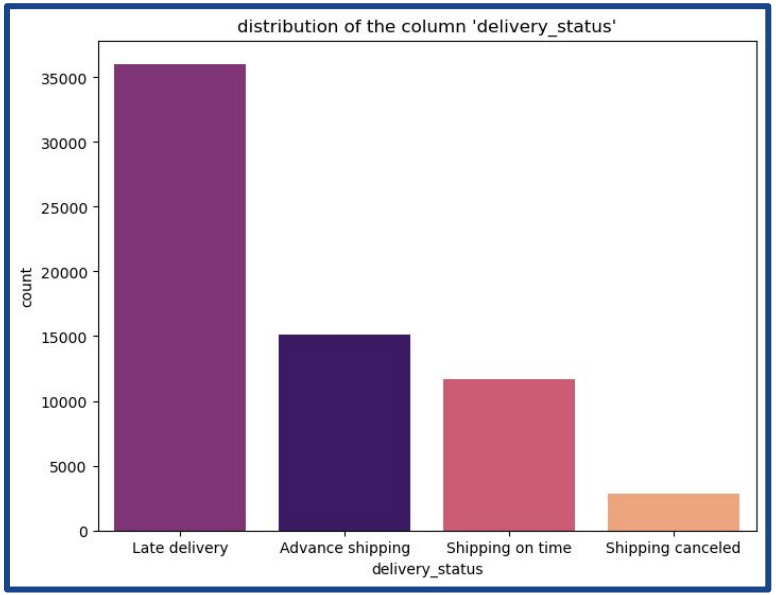
bivariate analysis: **“delivery_status” vs “items_per_order”**

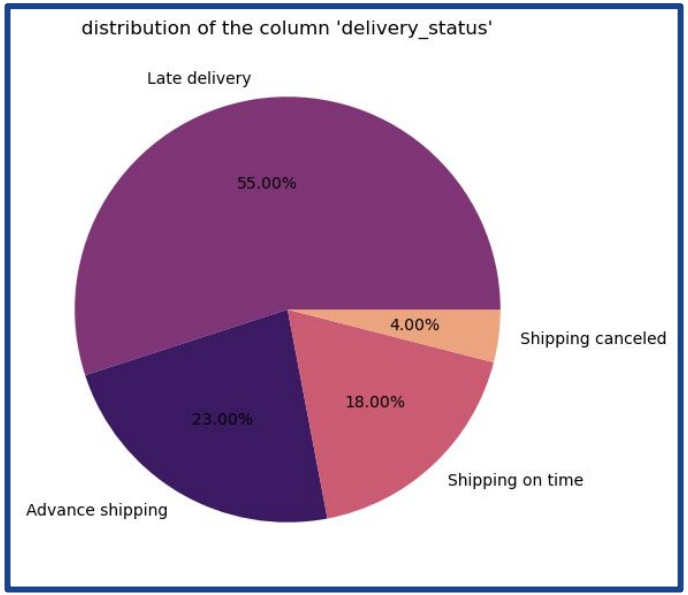
Chi = 0.69 > 0.05
Cramer = 0.018 -> [0.0-.01]



univariate analysis: **“delivery_status“**

- categorical/nominal value (4 possible values)
- the most common is “Late delivery” -> more than 50%
  - the specified delivery time is not realistic

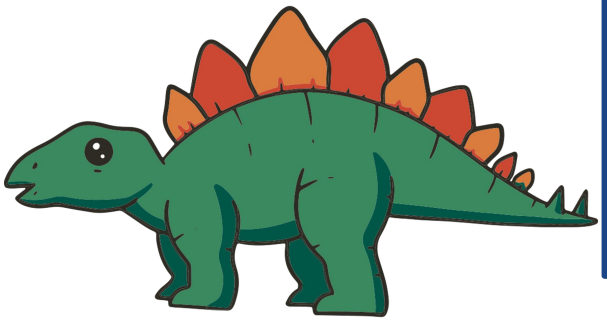


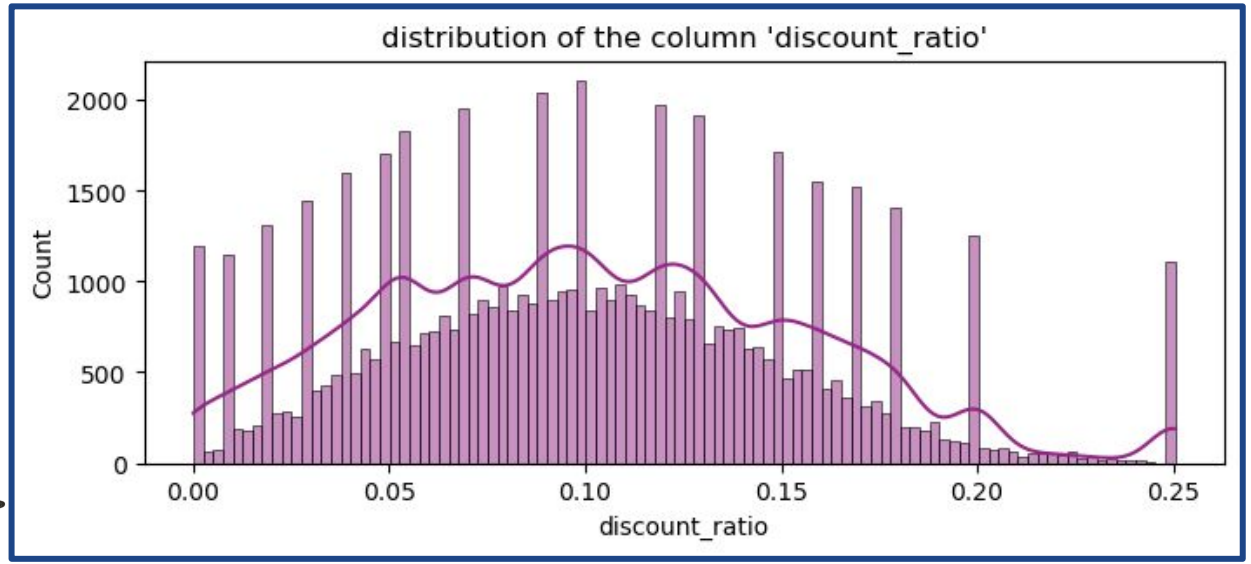


univariate analysis: **“discount_ratio“**

1

- **numerical/continuous** value (44653 values in dataset -> 0.0 - 1.0)
- **calculated** value -> sum(“order_item_discount) / sum(“sales”) -> for each order
- the **histplot** looks like:
  - a normal distribution
    (not completely symmetrical)
  - it’s platykurtic
    (not narrow but wide)

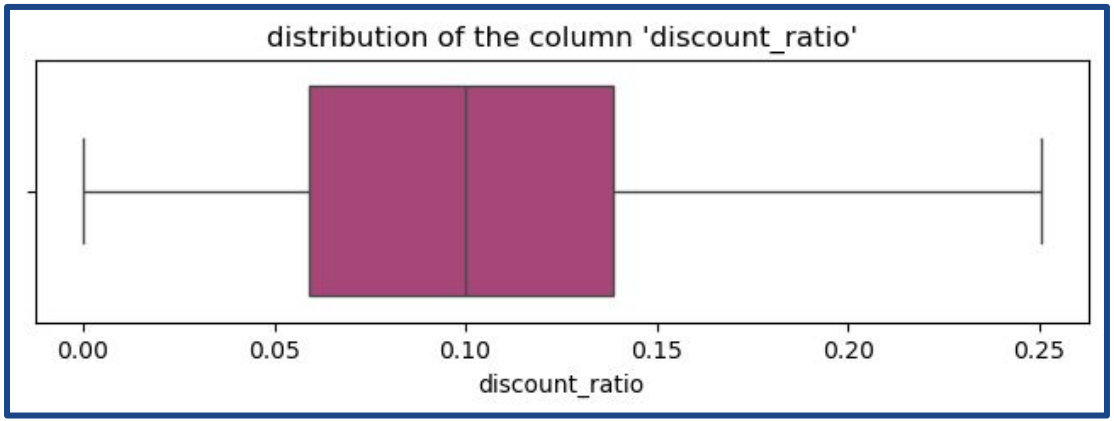




univariate analysis: **“discount_ratio“**

2

- **skewness** = 0.32 -> it’s a bit **right skewed**
- **kurtosis** = -0.3 -> it’s **platykurtic** (not narrow but **wide**)
- **outliers** = we don’t have any outliers



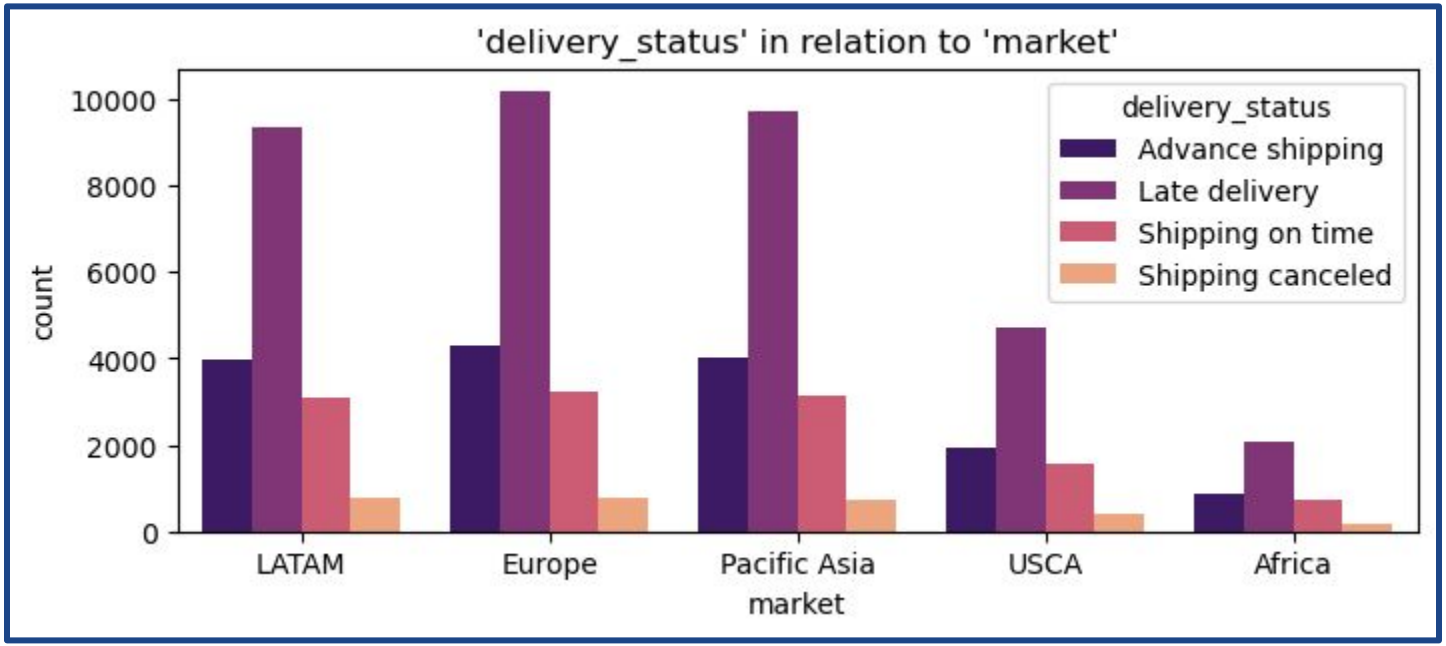
bivariate analysis: **“delivery_status” vs “market”** 1

- **“Late delivery”** is the **most common** -> **no matter** which **continent** (>54 %)
- **proportion** for each delivery_status is **almost the same** for every continent
- **check** that there is **no correlation**:
  - **chi quare test** = 0.49 -> **no association**
  - **cramer’s v** = 0.0076 -> **strength** is almost **0**
    - it is not needed to calculate cramer’s v because of chi square test

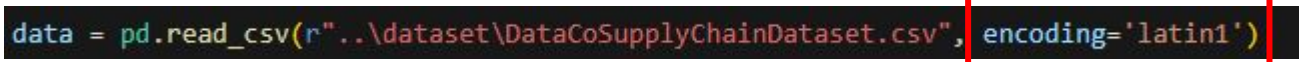
delivery_status market	Advance shipping	Late delivery	Shipping canceled	Shipping on time
Africa	0.230929	0.541256	0.042813	0.185003
Europe	0.232423	0.549485	0.042993	0.175098
LATAM	0.232233	0.543566	0.044642	0.179559
Pacific Asia	0.227968	0.552995	0.041190	0.177846
USCA	0.224502	0.548316	0.046742	0.180441
	≈ 23 %	≈ 54 %	≈ 4 %	≈ 17 %

bivariate analysis: **“delivery_status” vs “market”** 2

- “Late delivery” is the most common -> no matter which continent



Lessons learned

- Importing a CSV file sometimes results in an error
  - check for the correct encoding type
    - standard is “UTF-8”
    - 

- Processing the data can take long
- Result is not always like expected
  - our result was the complete opposite
  - studying some variables and found out not relevant info
  - they should improve the delivery time
- Dataset should be understandable
  - ask the customer if something is not clear/logic


delivery status
of a supply chain
Q&A