EDA
-
A Muesli distribution company

Julian, Markus, Yixin, Johannes

Overview

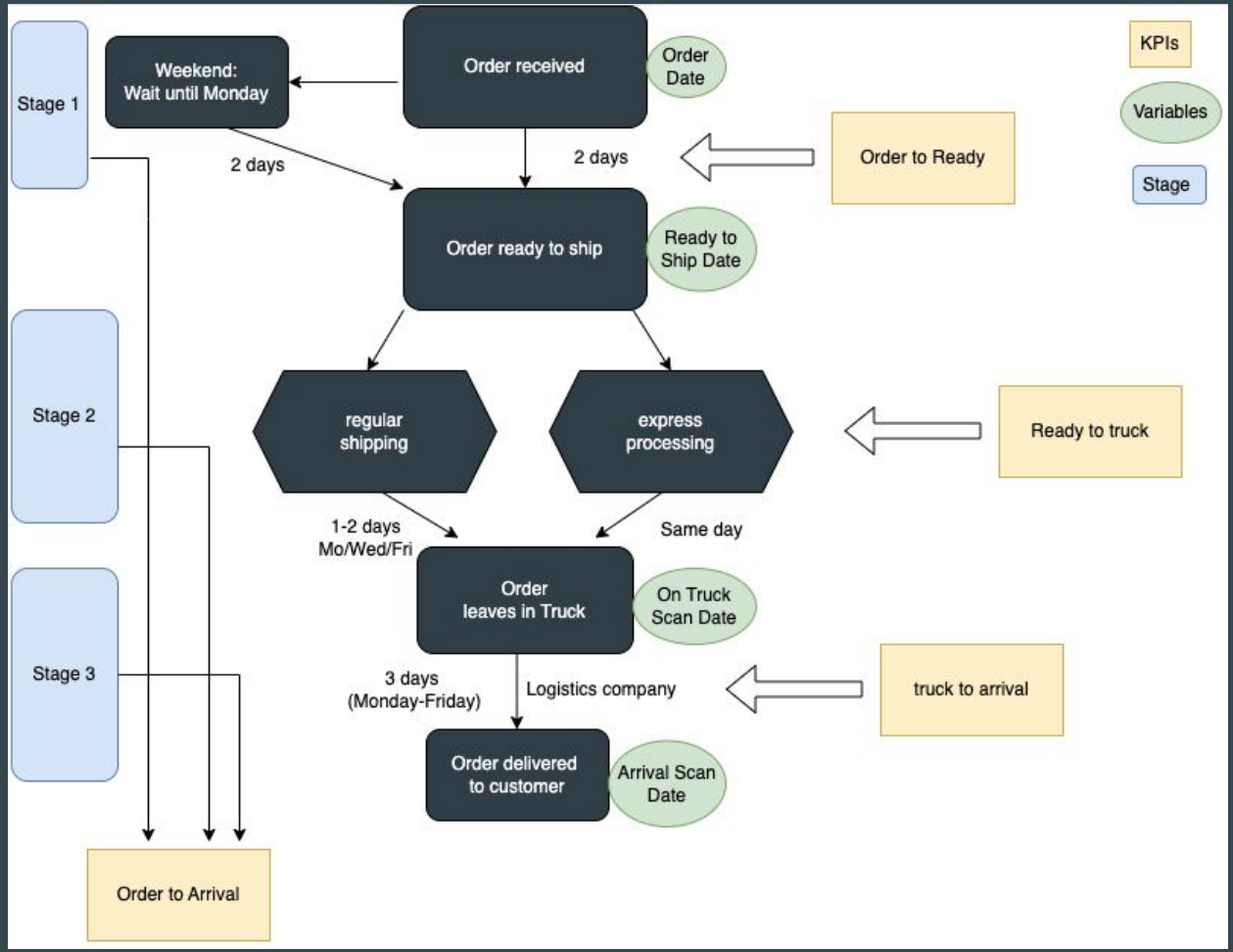
1. Introduction
2. Flowchart
3. EDA explanation and findings
4. Summary and recommendation for action

Introduction

A Muesli distribution company

- Data: Overall order data including delivery process
- KPIs: To keep track of the health of their business
- Purpose: Improving the service

Flowchart


Stage 1
Stage 2
Stage 3
Weekend:
Wait until Monday
Order received
Order Date
KPIs
Variables
Stage
2 days
2 days
Order to Ready
Order ready to ship
Ready to Ship Date
regular shipping
express processing
Ready to truck
1-2 days
Mo/Wed/Fri
Same day
Order leaves in Truck
On Truck Scan Date
3 days
(Monday-Friday)
Logistics company
truck to arrival
Order delivered to customer
Arrival Scan Date
Order to Arrival

EDA

1. Reading and checking Dataframes
2. Removing duplicates/checking for missing values
3. Merging Dataframes
4. Finding KPIs
5. Matplotlib

EDA - Checking the different Dataframes

```
display(
    df_orders.head(1),
    df_camp.head(1),
    df_intern.head(1),
    df_process.head(1)
    )
hh
```

	Index	Order ID	Order Date	Ship Mode	Customer ID	Customer Name	Origin Channel	Country/Region	City	State	Postal Code	Region	Category	Sub-Category	Product ID	Sales	Quantity
0	27	CA-2019-121755	2019-01-16	Second Class	EH-13945	Eric Hoffmann	Email	United States	Los Angeles	California	90049.0	West	Special Projects Muesil	Gluten Free	TEC-AC-10003027	90.57	3

	Order ID	Arrival Scan Date	Customer Name
0	CA-2019-109666	2019-05-03	Kunst Miller

	Order ID	Ready to Ship Date	Pickup Date
0	CA-2019-116540	2019-09-02	2019-09-03

	Row ID	Order ID	Order Date	On Truck Scan Date	Ship Mode
0	3074	CA-2019-125206	2019-01-03	2019-01-07	Express

```
### cleaning column names
df_list = [df_orders, df_camp, df_intern, df_process]

for df in df_list:
    df.columns = df.columns.str.lower()
    df.columns = df.columns.str.replace(" ", "_")
```

Column names:

- lower case
- replace ‘ ’ with ‘_’

EDA - Checking the different Dataframes

```
for df in df_list:
    print(df.columns)
```

```
Index(['index', 'order_id', 'order_date', 'ship_mode', 'customer_id',
       'customer_name', 'origin_channel', 'country/region', 'city', 'state',
       'postal_code', 'region', 'category', 'sub-category', 'product_id',
       'sales', 'quantity', 'discount', 'profit'],
      dtype='object')
Index(['order_id', 'arrival_scan_date', 'customer_name'], dtype='object')
Index(['order_id', 'ready_to_ship_date', 'pickup_date'], dtype='object')
Index(['row_id', 'order_id', 'order_date', 'on_truck_scan_date', 'ship_mode'], dtype='object')
```

```
### checking for null values
for df in df_list:
    print(df.isnull().sum())
```

```
index               0
order_id            0
order_date          0
ship_mode           0
customer_id         0
customer_name       0
origin_channel      0
country/region      0
city                0
state               0
postal_code        11
region              0
category            0
sub-category        0
product_id          0
sales               0
quantity            0
discount            0
profit              0
dtype: int64
order_id            0
arrival_scan_date   0
customer_name       0
dtype: int64
order_id            0
ready_to_ship_date  0
pickup_date         0
dtype: int64
row_id              0
order_id            0
order_date          0
on_truck_scan_date  0
ship_mode           0
dtype: int64
```

```
1  # drop rows with duplicates in order_id column
2  df_order_no_dups = df_orders.drop_duplicates(subset="order_id")
```

- Finding useful column names
- Checking for null values
- Dropping duplicate/group orders so as not to distort statistical values

EDA - Merging Dataframes on Order ID

Dataframes:
df_orders
df_camp
df_intern
df_process

	Index	Order ID	Order Date	Ship Mode	Customer ID	Customer Name	Origin Channel	Country/Region	City
0	27	CA-2019-121755	2019-01-16	Second Class	EH-13945	Eric Hoffmann	Email	United States	Los Angeles

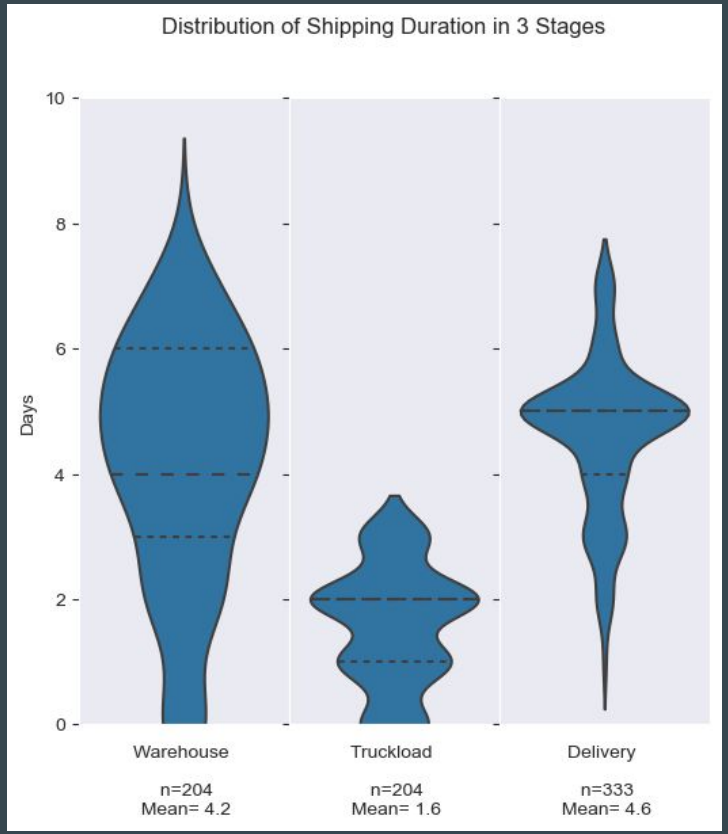
	Order ID	Arrival Scan Date	Customer Name
0	CA-2019-109666	2019-05-03	Kunst Miller

	Order ID	Ready to Ship Date	Pickup Date
0	CA-2019-116540	2019-09-02	2019-09-03

	Row ID	Order ID	Order Date	On Truck Scan Date	Ship Mode
0	3074	CA-2019-125206	2019-01-03	2019-01-07	Express

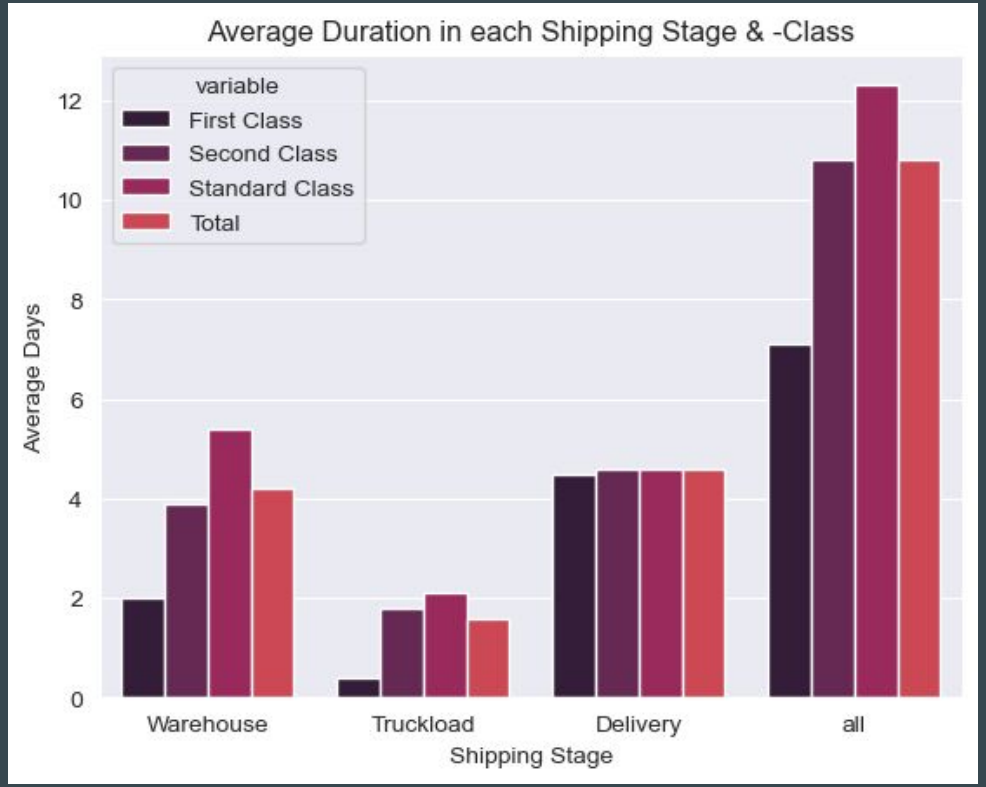
EDA - The Three Shipping Stages

- Big Variance in Stage 1
- Waiting-time in Stage 2
- Overall long delivery in Stage 3



EDA - The Three Shipping Stages

- Differences between Shipping-Classes in Stage 1 & 2
- It takes to long



EDA - KPIs

Order to Ready (Stage 1)	• To evaluate the efficiency of preparing the order • Calculation: Ready to Ship Date - Order Date
Ready to Truck Scan (Stage 2)	• Looking for inefficiencies in order pick up • Calculation: On Truck Scan Date - Ready to Ship date
Truck Scan to Arrival (Stage 2 + 3)	• To evaluate the efficiency of logistics company • Calculation: Arrival Scan Date - Truck Scan Date
Order to Arrival (Stage 1 - 3)	• To evaluate the efficiency of whole process • Calculation: Arrival scan date - Order date

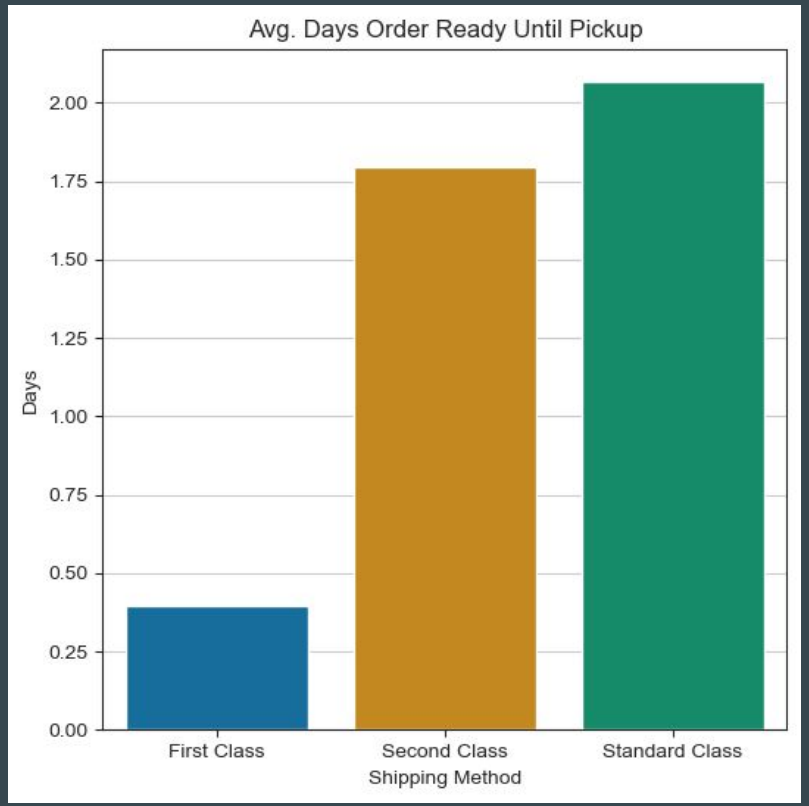
EDA - KPI: Order to Ready Days Frequency

count	204.000000
mean	4.176471
std	1.969780
min	0.000000
50%	4.000000
max	8.000000

Company assumption: 0 to 2 days Mo-Fr

Findings: Overall too slow

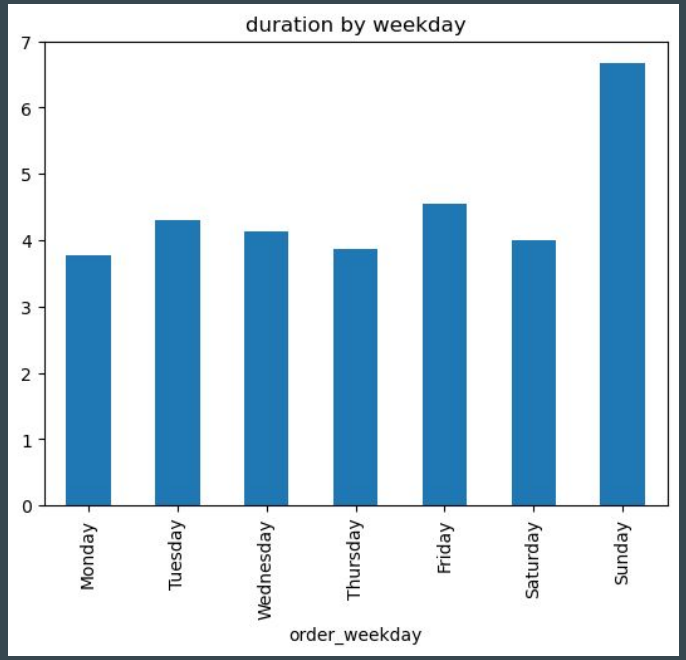
EDA - KPI: Order to Ready Days by shipping method



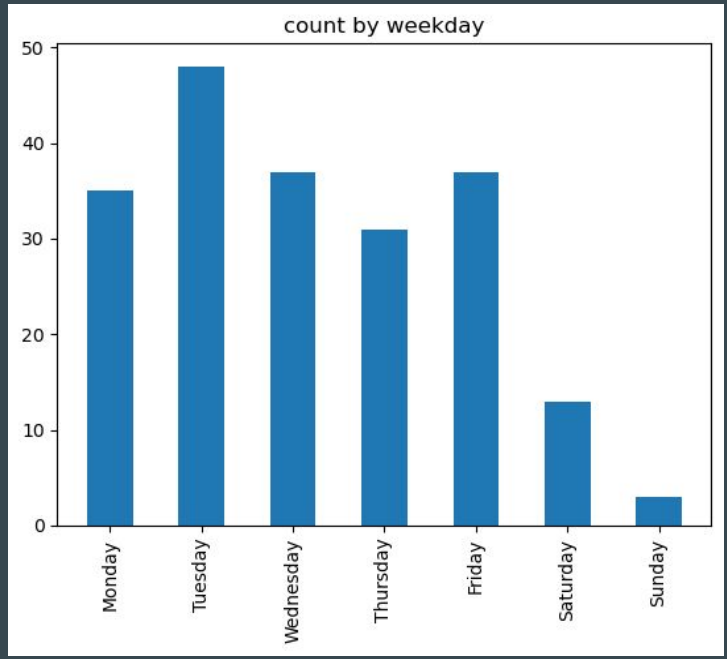
Findings:

- First class is as fast as promised.
- No real difference between “Second Class” and “Standard Class” and they are too slow.

EDA - KPI: Order to Ready by weekday



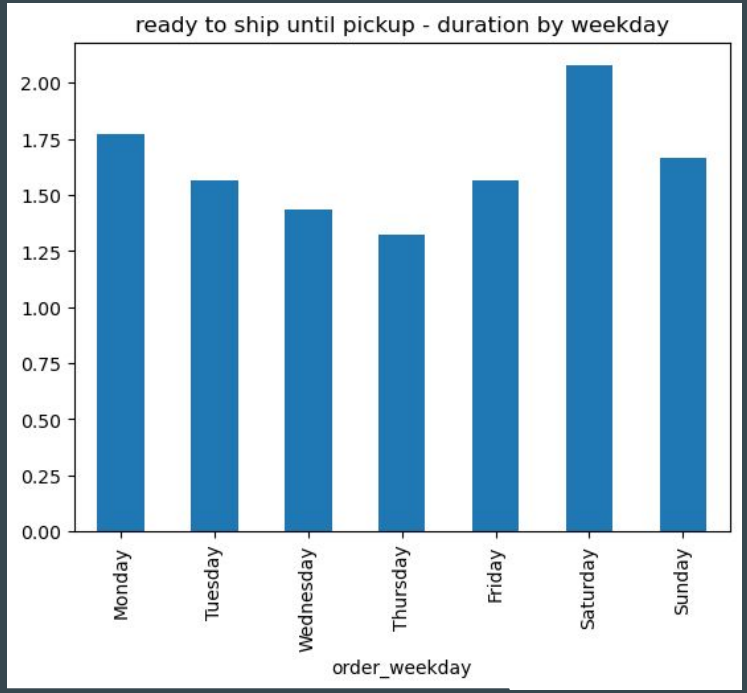
There is not enough data for orders on Saturday and Sunday.



No conclusive result due to sample size problem
Get more information about processing time from different products and categories.

EDA - KPI: Ready to Truck Scan

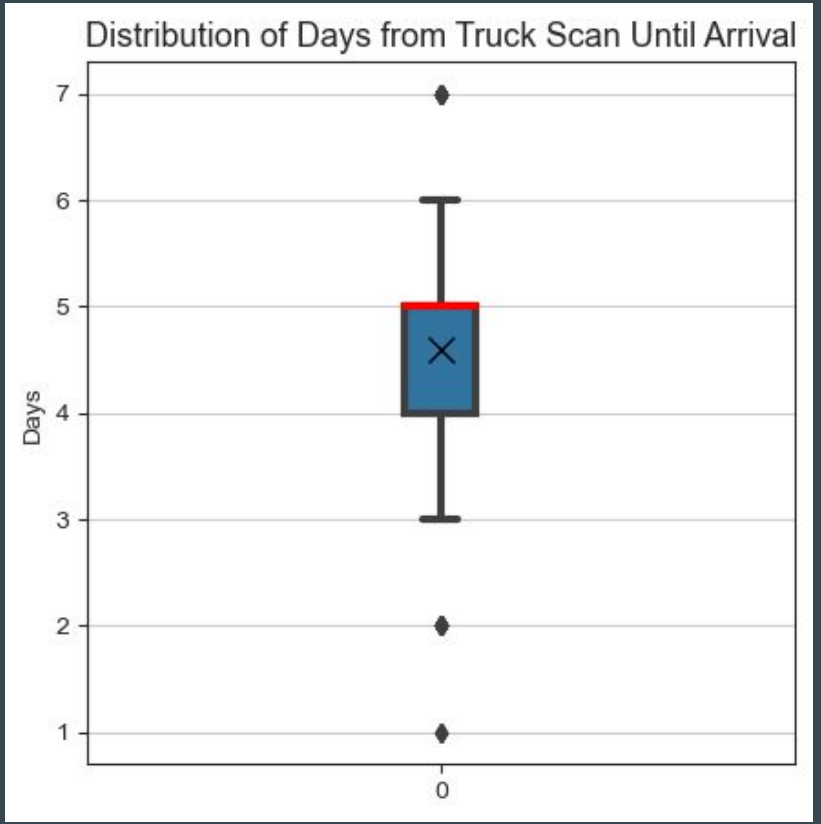
count	204
mean	1 days 13:45:52
std	0 days 22:50:34
min	0 days
50%	2 days
max	3 days



Company assumption: 0 to 2 days
Findings: This area is not too problematic.

EDA - KPI: Truck Scan to Arrival

count	333
mean	4 days 14:29:11
std	1 days 04:47:16
min	1 day
25%	4 days
50%	5 days
75%	5 days
max	7 days



Assumption: Average of 3 days (given by logistics company)

Findings:
The logistics company is not keeping its promise.

EDA - KPI: Order to Arrival

To evaluate the whole process holistically.

Assumption: 3 to 7 days

Weekday	Average time
Monday	11 days 01:48:40
Tuesday	10 days 23:35:35
Wednesday	11 days 04:53:52
Thursday	10 days 06:51:25
Friday	10 days 05:50:16
Saturday	9 days 08:00:00
Sunday	11 days 05:27:16

Summary

- Orders are taking a lot longer than assumed.
- Orders in non-express shipping are taking too long to get ready.
- The logistics company is not fulfilling its promise and taking much longer for delivery.

Recommendation for action

- Management should reevaluate the internal warehouse processes.
  - Looking at processing times for different products and categories.
  - Overload with many orders in a short timeframe?
  - Rerunning data collection more thoroughly or implement a system for that?
- The logistics company should be talked to.
  - Possibly the Mo/We/Fr rule could be eliminated.
  - Express option?