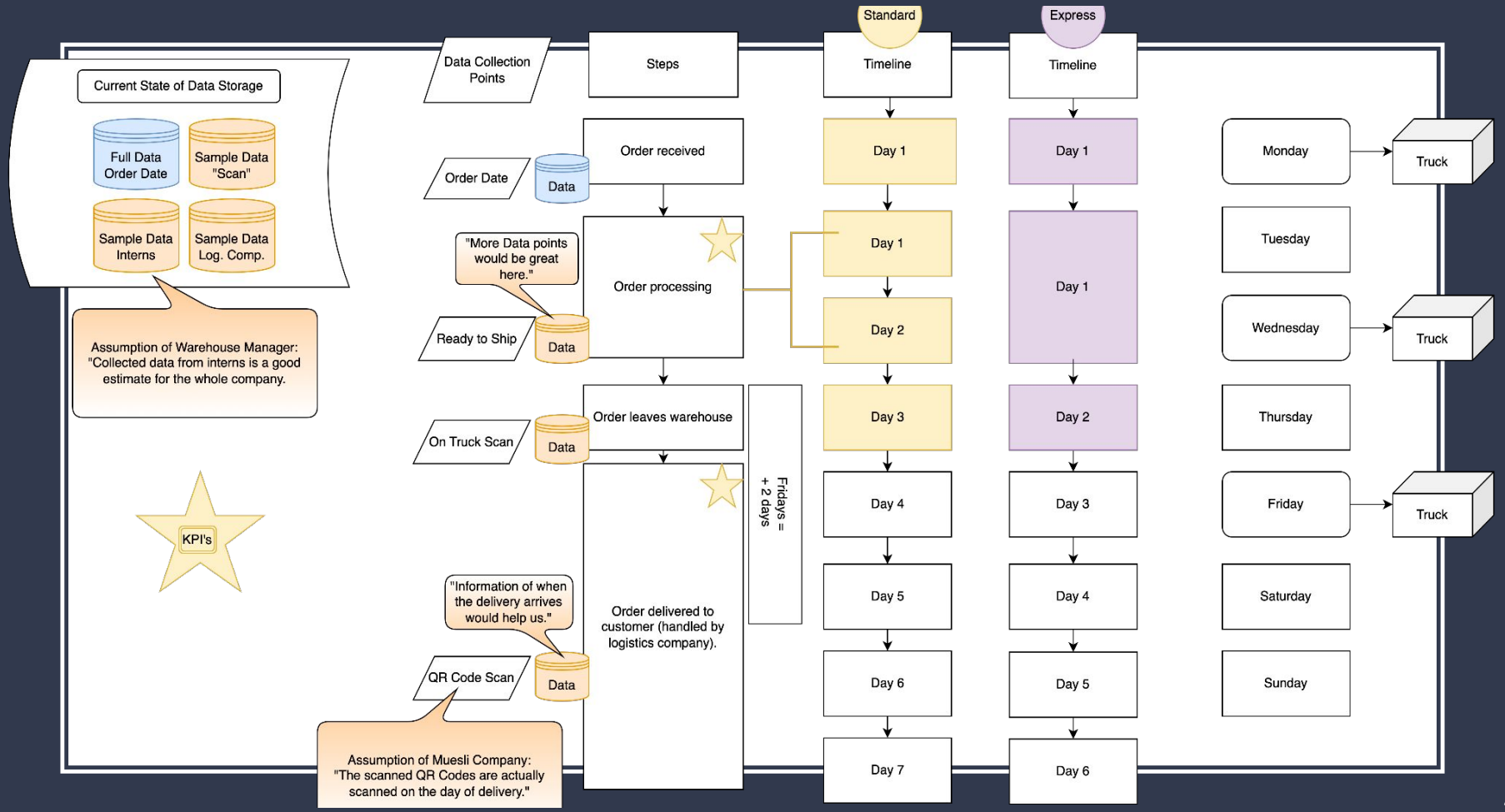
EDA- Muesli distribution company

Andreas

Christoph

Frederic

Xhina


Current State of Data Storage
Full Data Order Date
Sample Data "Scan"
Sample Data Interns
Sample Data Log. Comp.
Assumption of Warehouse Manager: "Collected data from interns is a good estimate for the whole company.
KPI's
Data Collection Points
Steps
Standard
Timeline
Express
Timeline
Order Date
Data
Order received
Day 1
Day 1
Monday
Truck
"More Data points would be great here."
Ready to Ship
Data
Order processing
Day 1
Day 2
Day 1
Tuesday
Wednesday
Truck
On Truck Scan
Data
Order leaves warehouse
Day 3
Day 2
Thursday
Fridays = + 2 days
Day 4
Day 3
Friday
Truck
"Information of when the delivery arrives would help us."
QR Code Scan
Data
Order delivered to customer (handled by logistics company).
Day 5
Day 4
Saturday
Day 6
Day 5
Sunday
Day 7
Day 6
Assumption of Muesli Company: "The scanned QR Codes are actually scanned on the day of delivery."

KPIs

Average Delivery Time

Delivery Exceptions

Delivery volume

Difference of shipping modes

Busiest times in the warehouse

Simple EDA

```
df_order = pd.read_csv("data/order_data.csv", skiprows=1)
df_scan = pd.read_csv("data/order_process_data.csv")
df_cd = pd.read_csv("data/campaign_data.csv")
df_intern = pd.read_csv("data/intern_data_study.csv")

df_order.head()
df_scan.head()
df_cd.head()
df_intern.head()

display(df_order.shape)
display(df_cd.shape)
display(df_intern.shape)
display(df_scan.shape)

df_order.info()
df_scan.info()
df_cd.info()
df_intern.info()

df_order.describe()
df_scan.describe()
df_cd.describe()
df_intern.describe()
```

```
df_order.columns = df_order.columns.str.replace(" ","_")
df_order.columns = df_order.columns.str.lower()
df_order.columns = df_order.columns.str.replace("/","_")
df_order.columns = df_order.columns.str.replace("-","_")

df_order['order_date']=pd.to_datetime(df_order['order_date'],format='%d/%m/%Y')

df_order.drop(["postal_code", "customer_name", "city", "country_region", "region",
"origin_channel", "customer_id"], axis=1, inplace=True)

df_order["order_id"].duplicated().value_counts()

display(df_order.isna().sum())
display(df_scan.isna().sum())
display(df_cd.isna().sum())
display(df_intern.isna().sum())
```

Average Delivery Time & Delivery Exceptions

- Looking at the whole process.
- KPI’s broken down by “warehouse”-process and “logistics company”-process

Processing Time

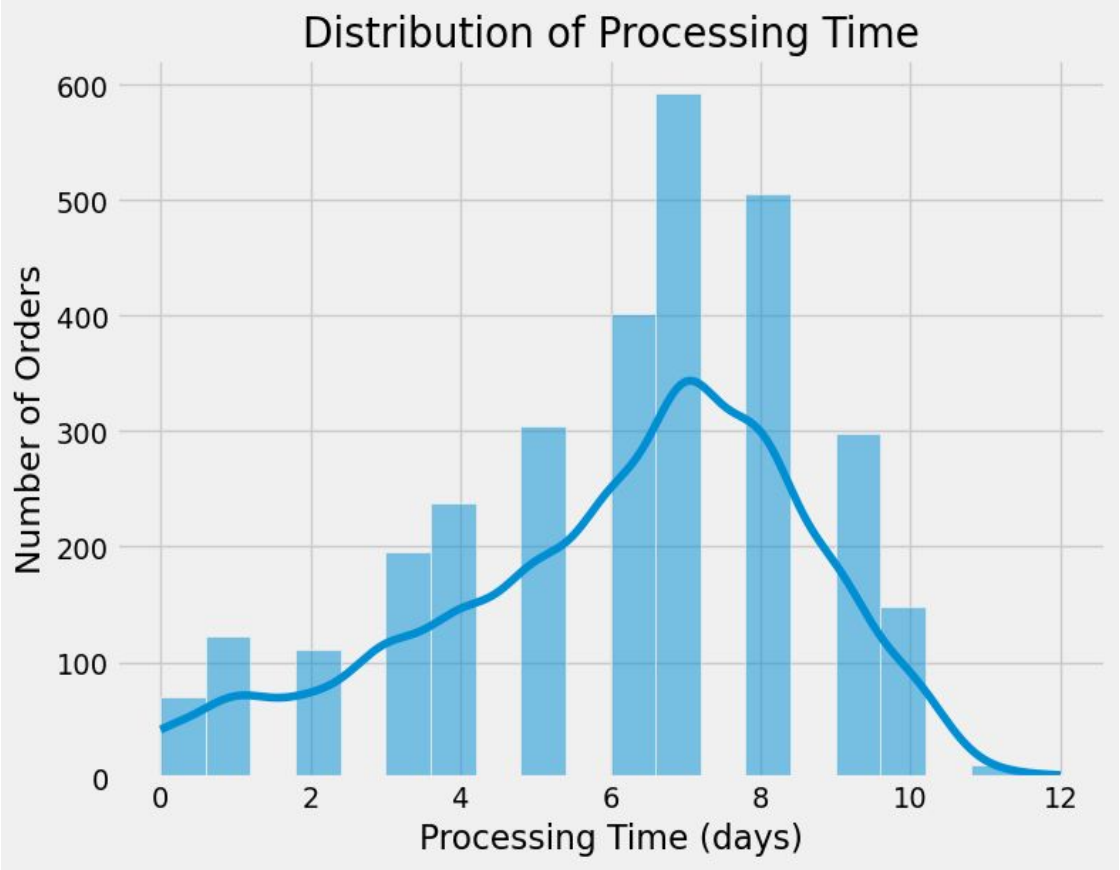
Input: “Order_Date”

Process: “Make the order ready to ship”

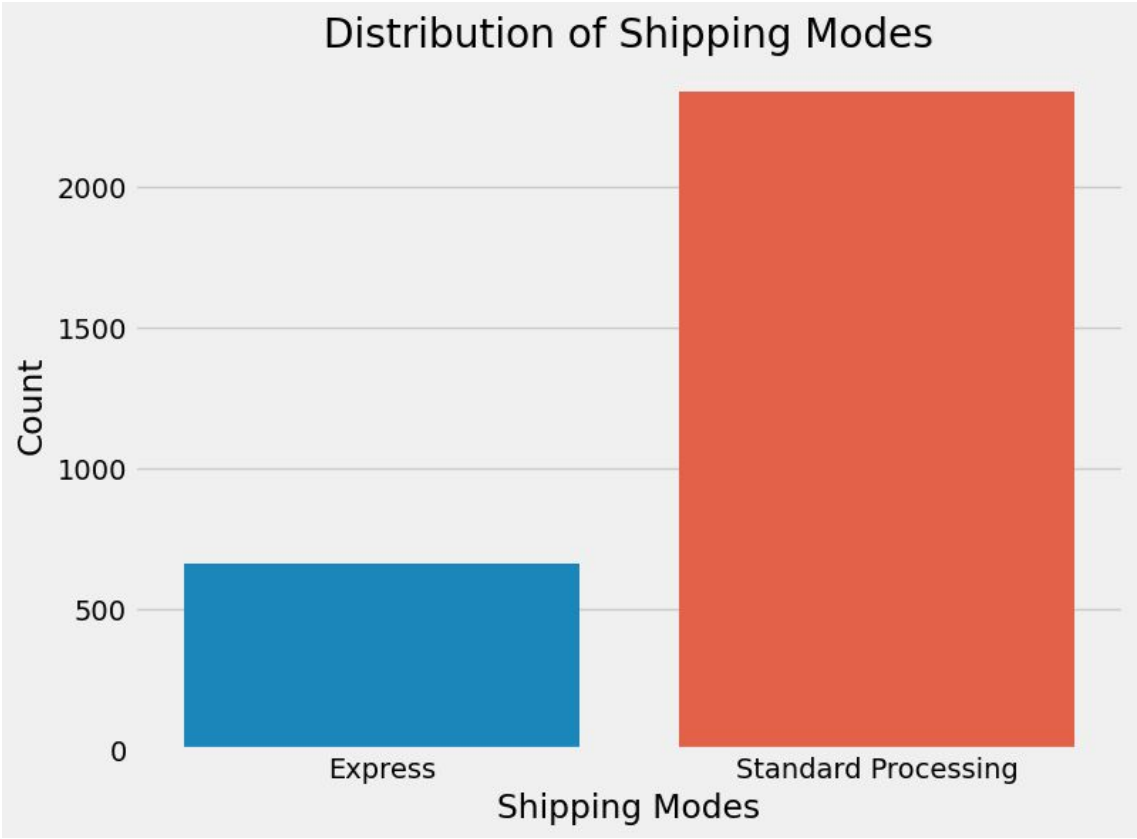
Output: “Scan_Date”

Outcome: “Package left the warehouse”

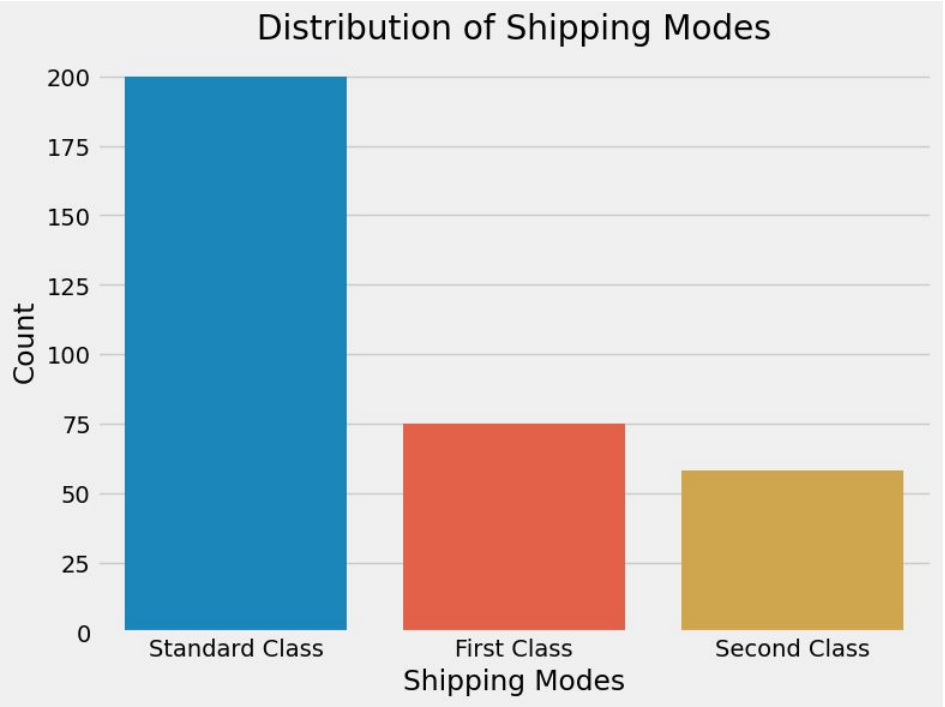
The average delivery time to logistics company is: 7.09 days

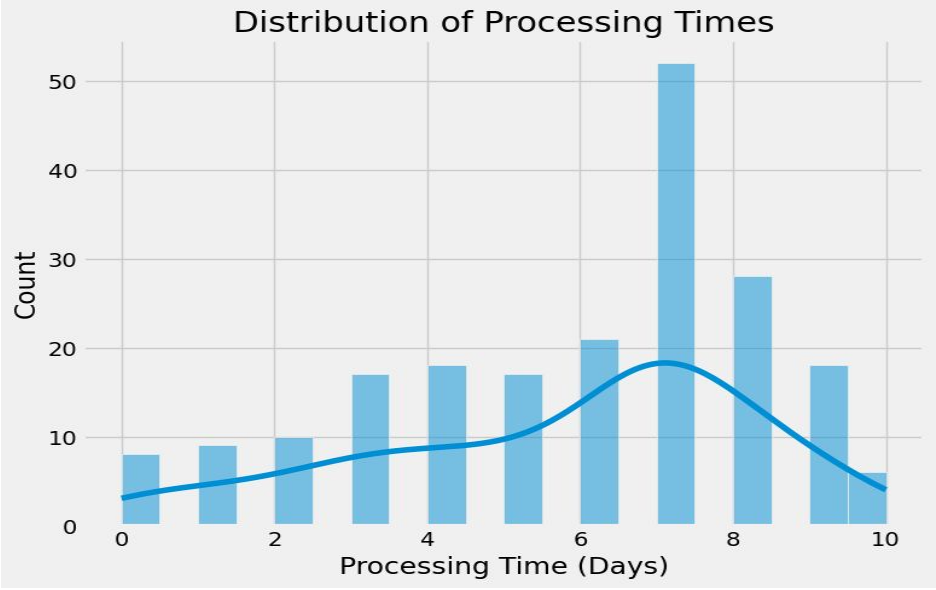


Count of orders by ship mode



The relationship between processing time and shipping mode

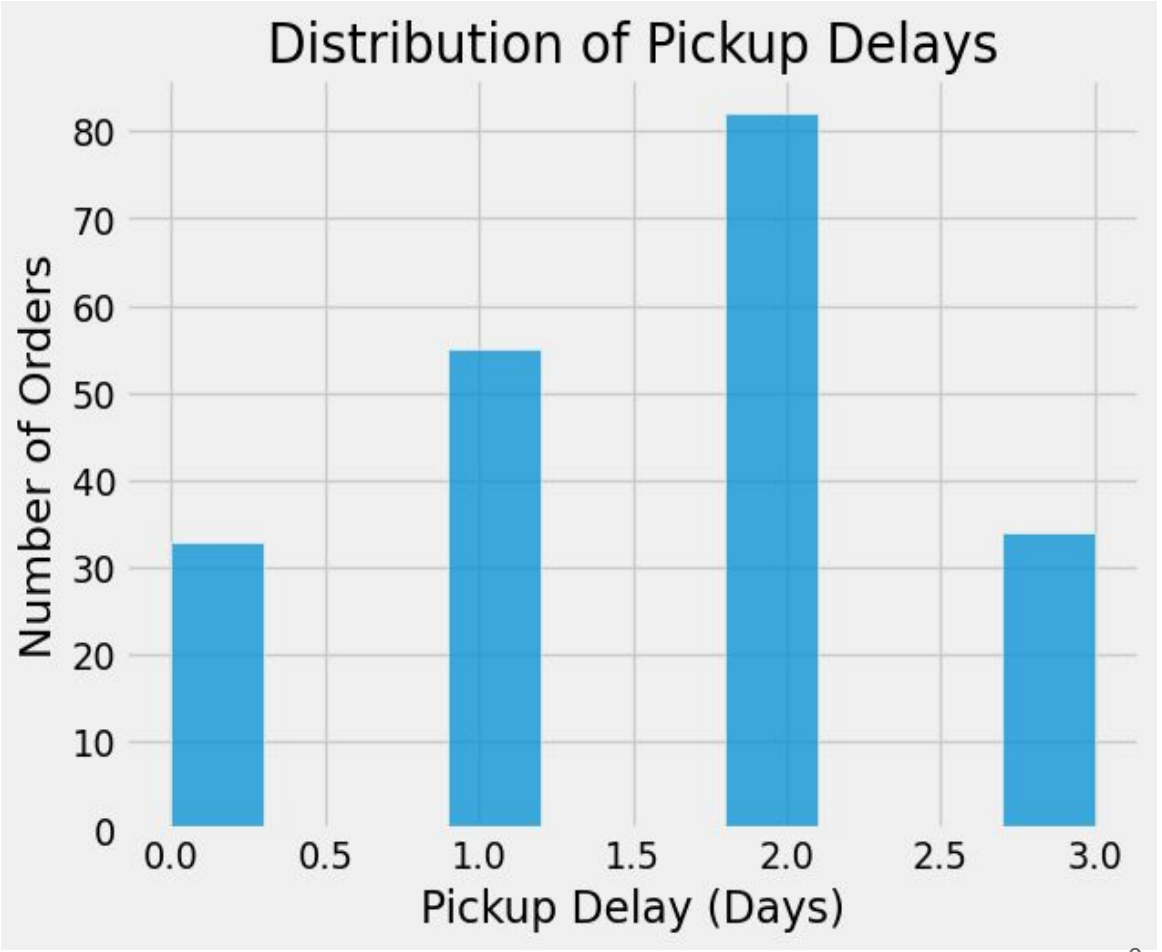




The pickup delay for each order

Summary Statistics for Pickup Delay:

count	204.000000
mean	1.573529
50%	2.000000
75%	2.000000
mode	2.000000

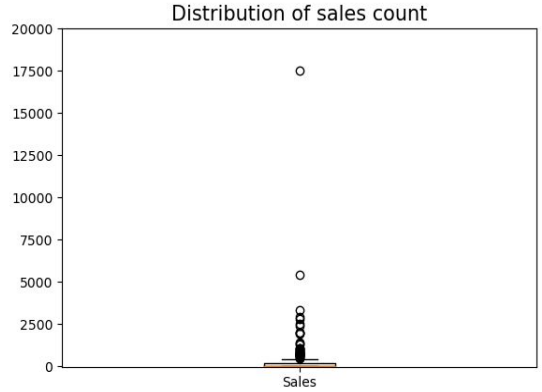


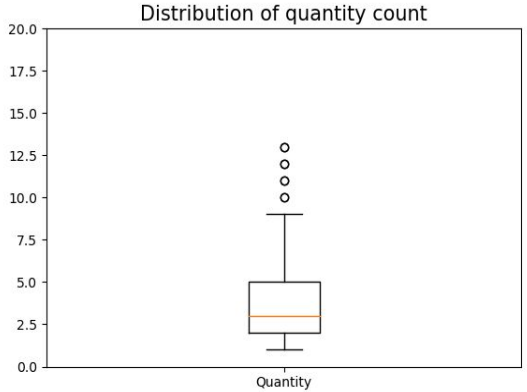
Delivery volume

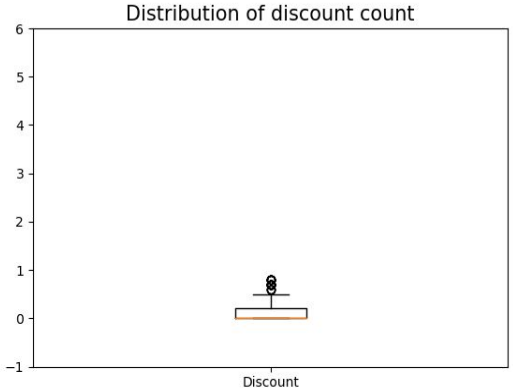
The number of orders that are delivered within a specific time period.
2019-2020

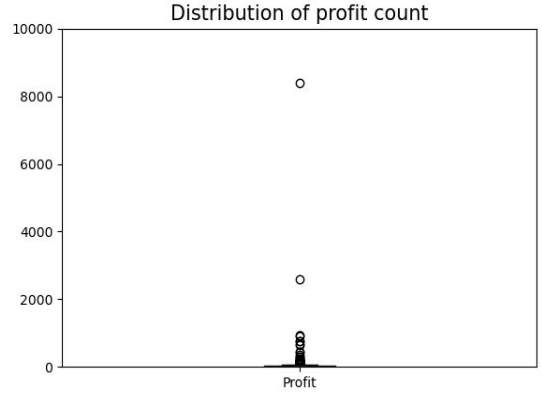
- Should we build another delivery centre because we have a lot of orders from one specific state?
- Should we do more campaigns about muesli?
- Do we need to include the data of when the order is delivered?

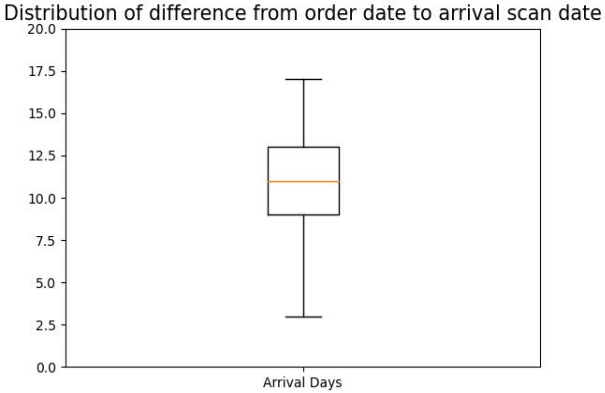
Distribution of numeric columns





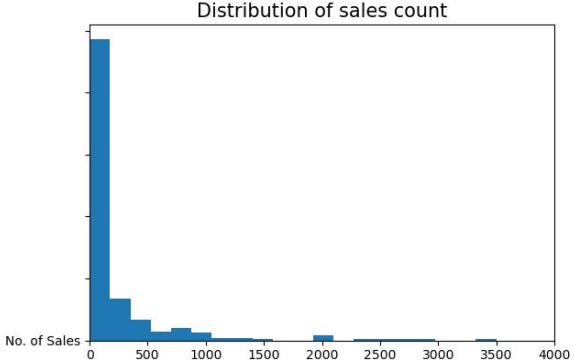


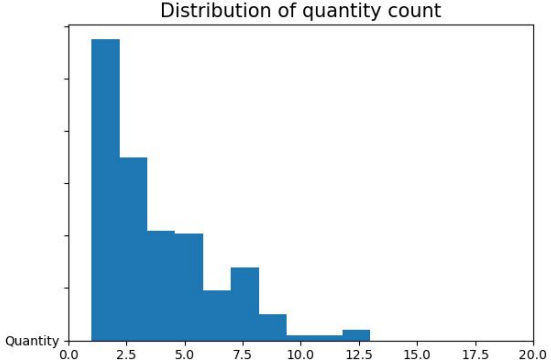


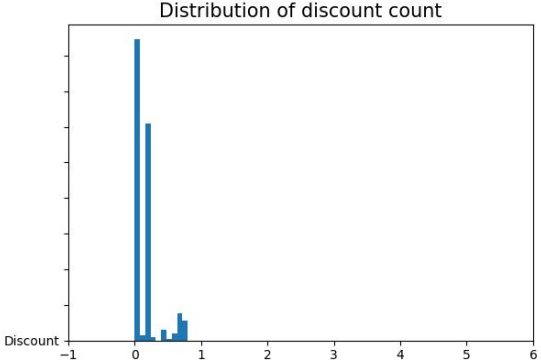


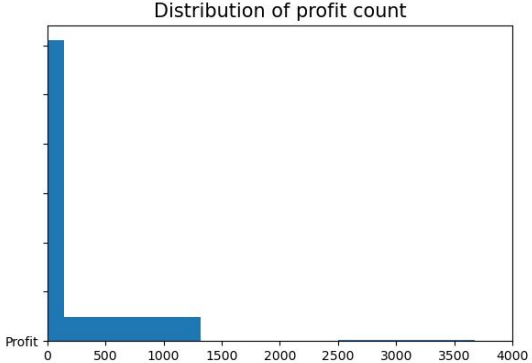
Red line represents median. The most of the values lie in the upper part of our distribution.

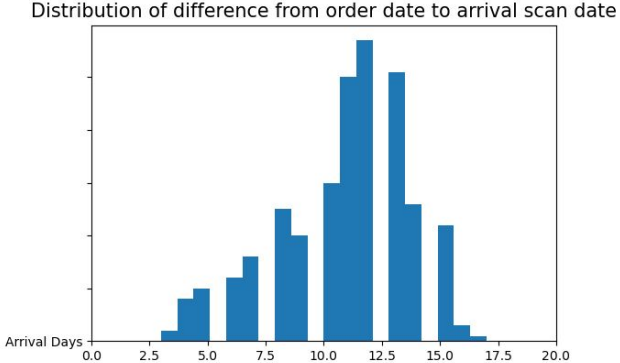
Distribution of numeric columns







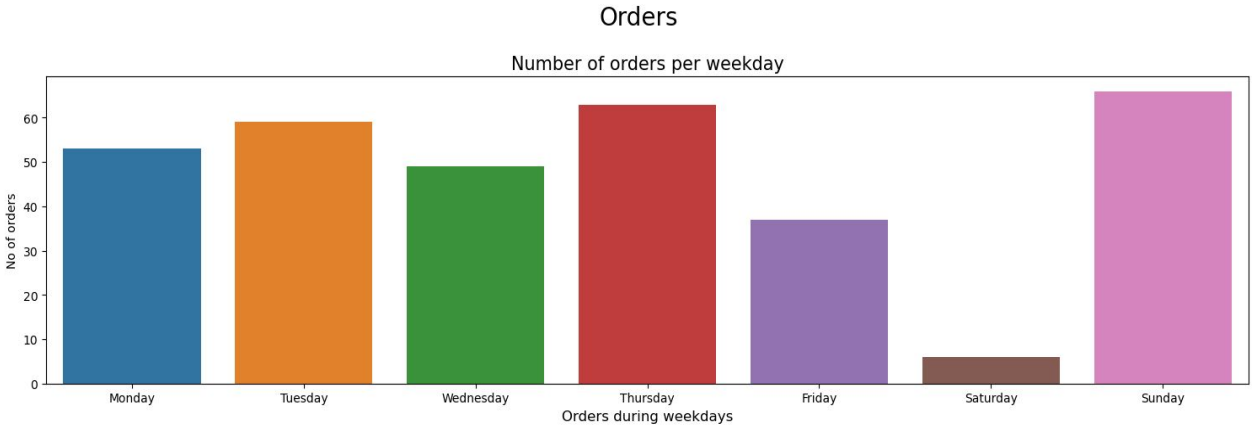


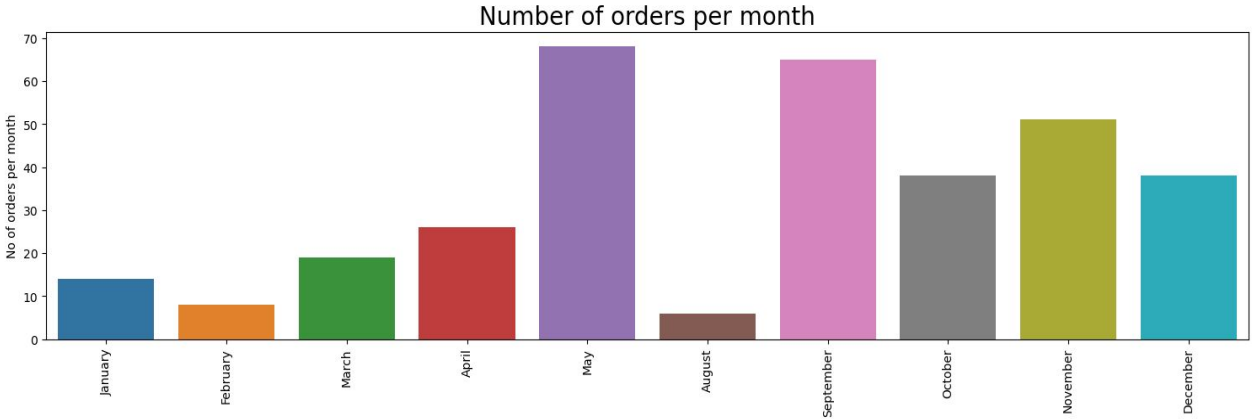


The mean, median, and mode are all different from each other.In this case, the mode is the highest point of the histogram, whereas the median and mean fall to the right of it.

No. orders from customers 2019-2020

The largest number of orders seems to occur on Sunday as well as in the month of May.

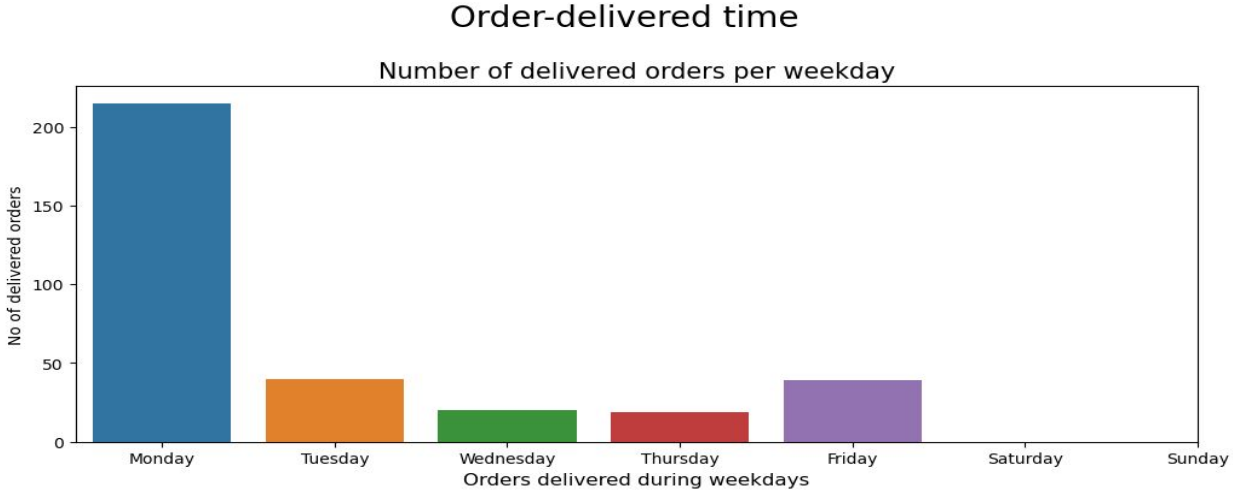


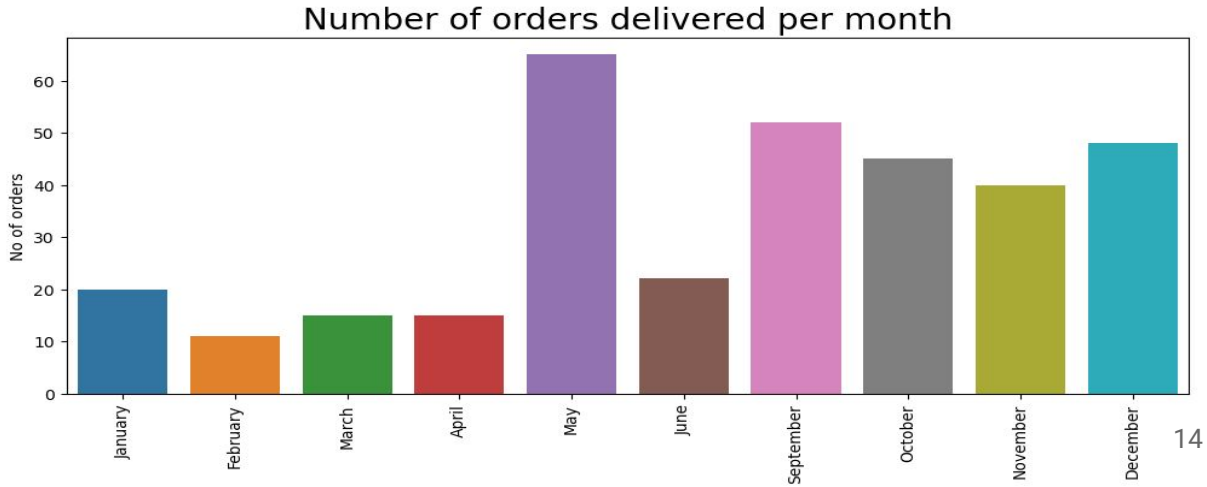


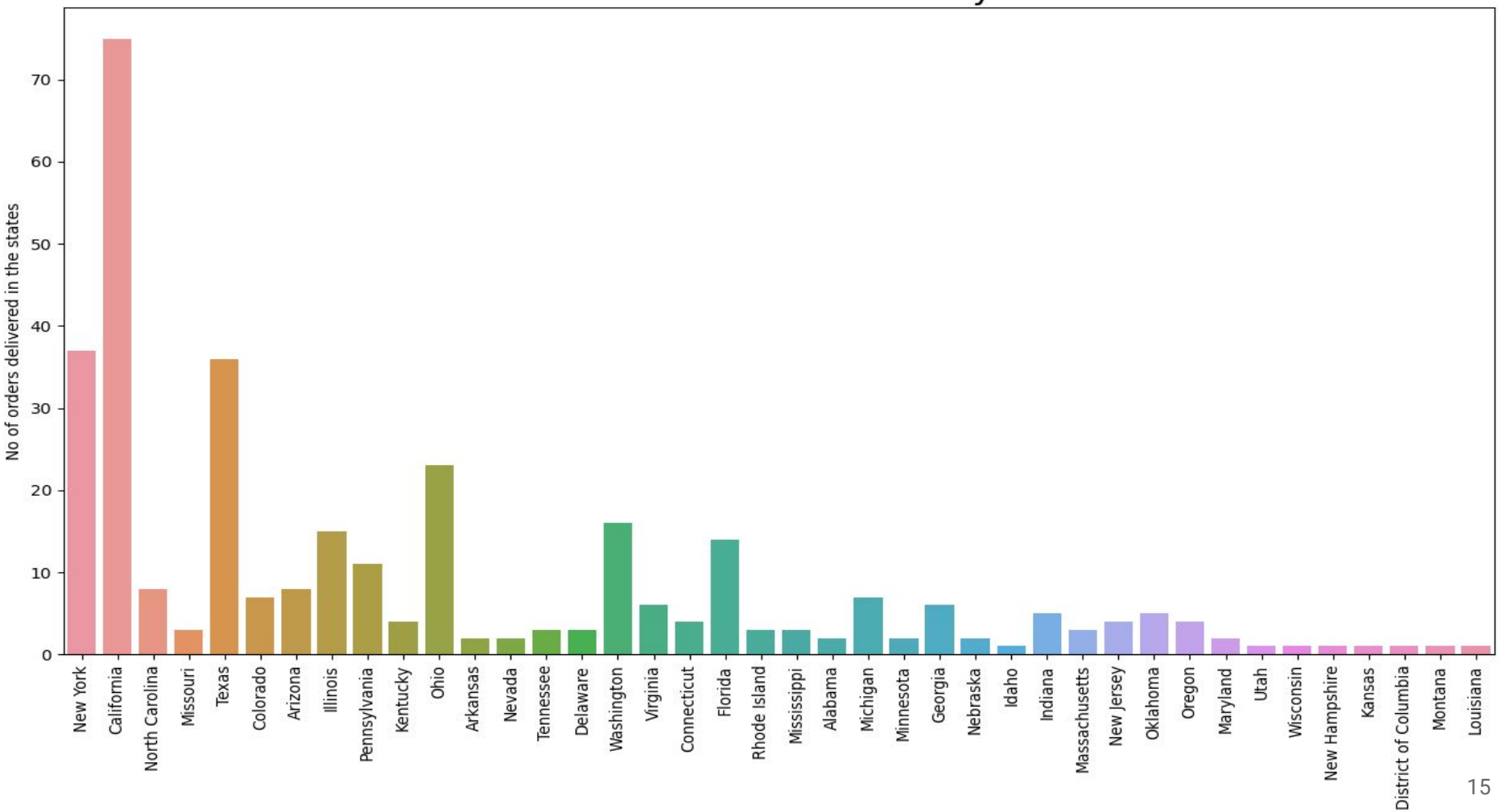
No. of orders delivered to the customer 2019-2020

The largest volume of delivery of orders seems to occur on Monday as well as in the month of May.

Min	3 days
Max	17 days
Avg Time	10.8 days
Mode	12 days






Number of orders delivered by state
No of orders delivered in the states
0
10
20
30
40
50
60
70
New York
California
North Carolina
Missouri
Texas
Colorado
Arizona
Illinois
Pennsylvania
Kentucky
Ohio
Arkansas
Nevada
Tennessee
Delaware
Washington
Virginia
Connecticut
Florida
Rhode Island
Mississippi
Alabama
Michigan
Minnesota
Georgia
Nebraska
Idaho
Indiana
Massachusetts
New Jersey
Oklahoma
Oregon
Maryland
Utah
Wisconsin
New Hampshire
Kansas
District of Columbia
Montana
Louisiana

Busiest times

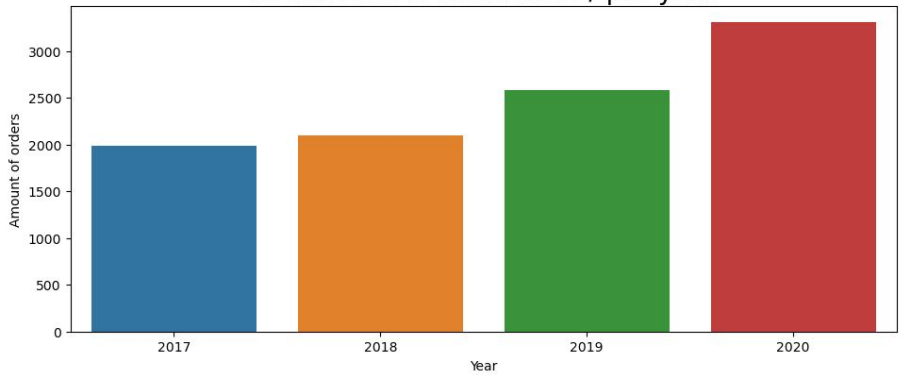
Are there days, on which we need to hire more workers?

Which times of the year we could have to deal with bottle necks?

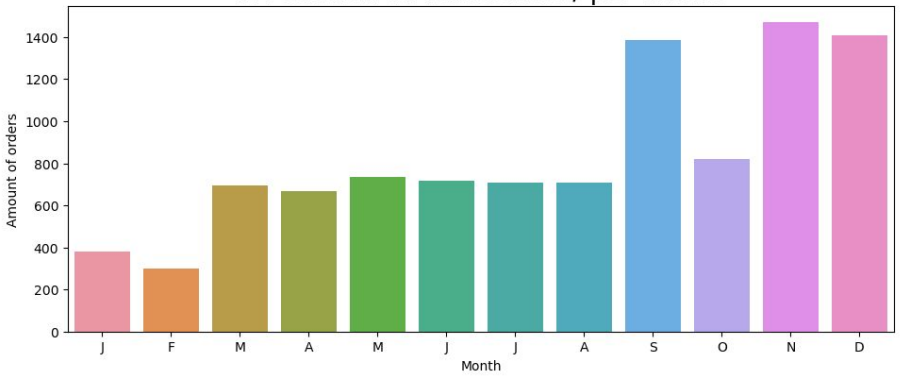
Do we have to prepare for special promotions, deals, times?

Amount of orders...

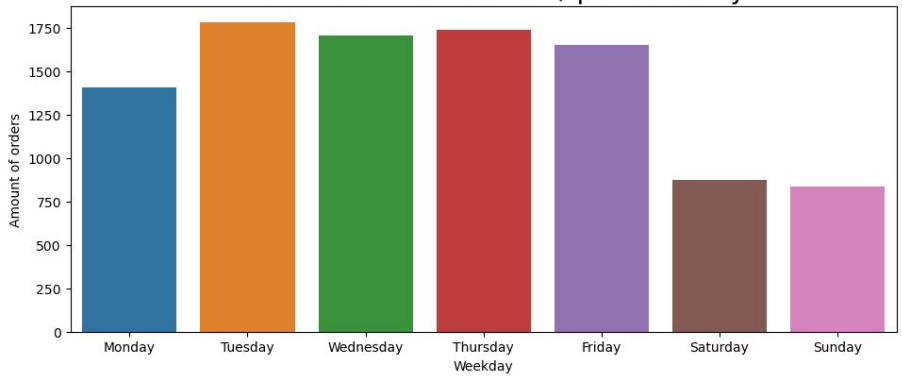
... received in warehouse / per year



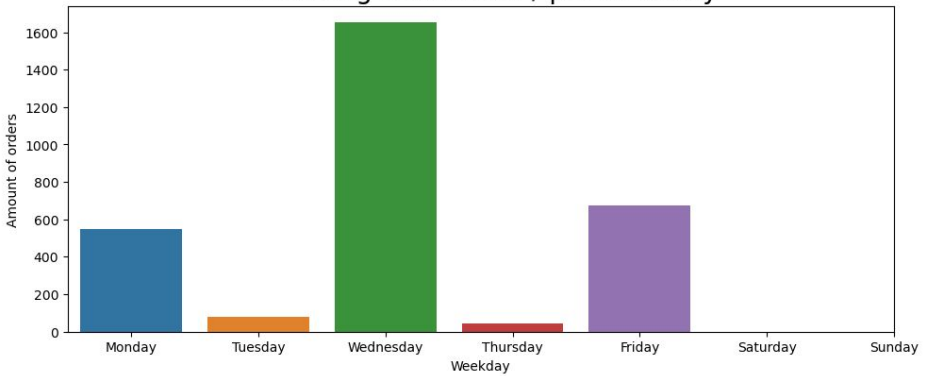
... received in warehouse / per month



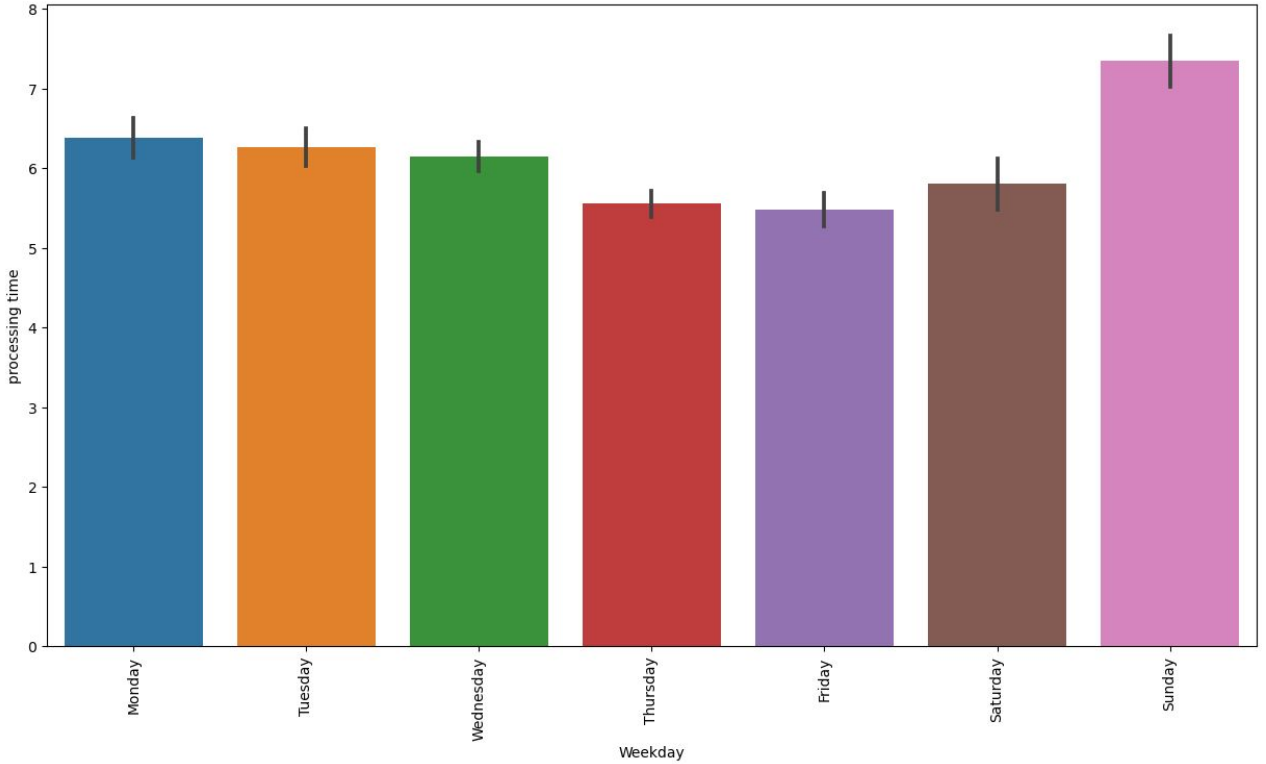
... received in warehouse / per weekday



... leaving warehouse / per weekday



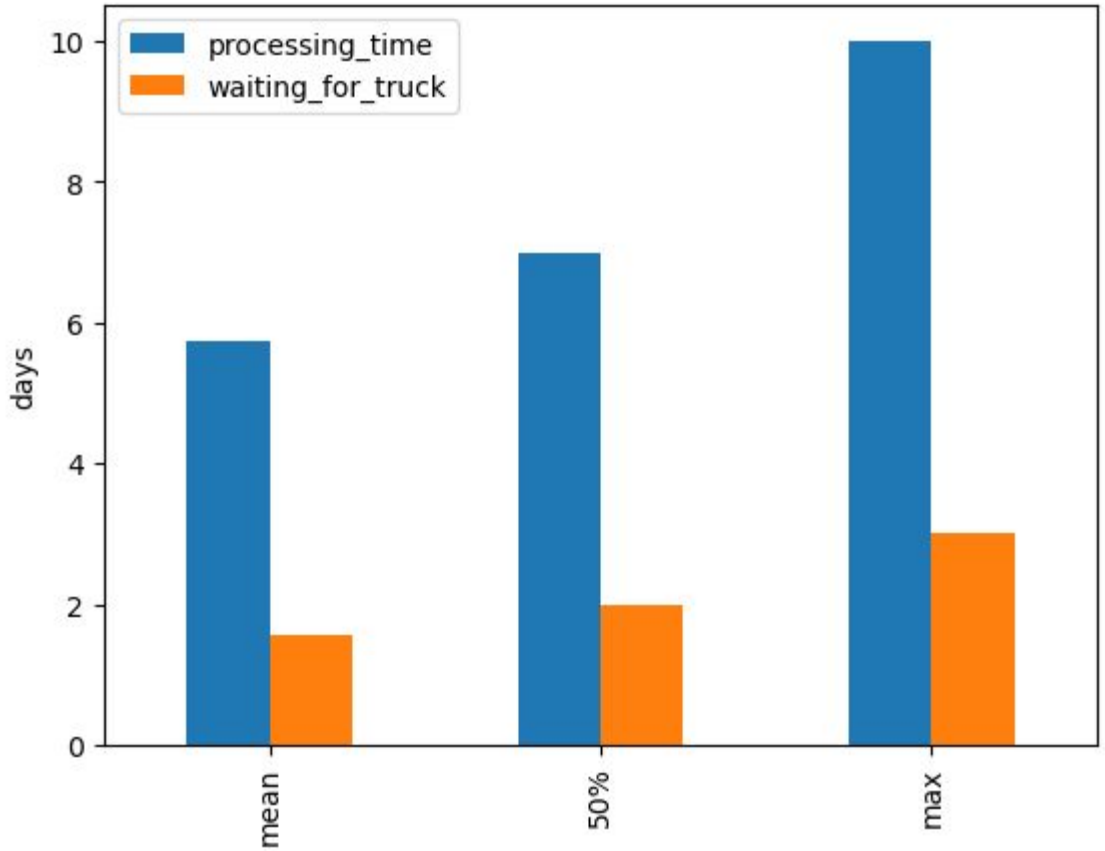
Processing time and weekday



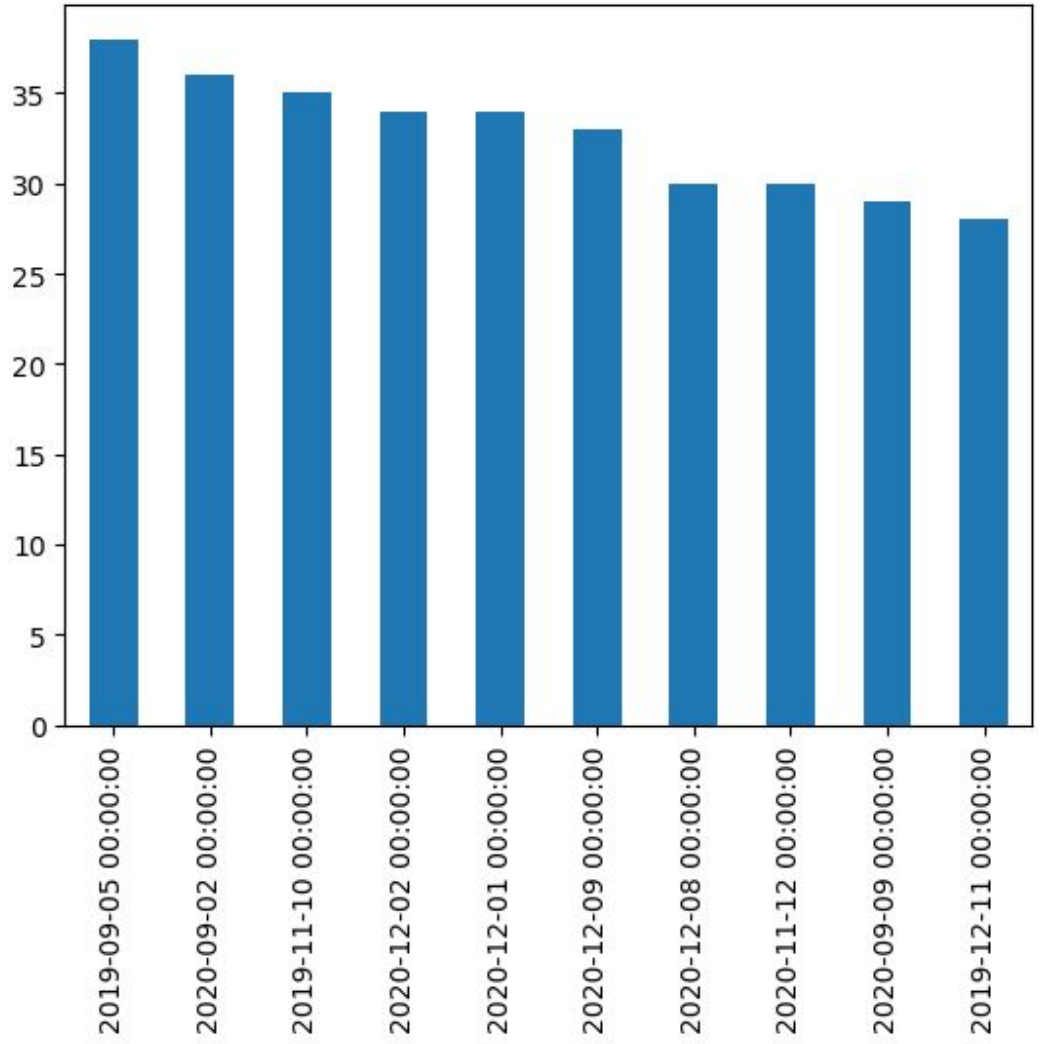
Processing / waiting time

(Based on the intern data)

In average packages are waiting one and a half days for a truck, when ready to ship.

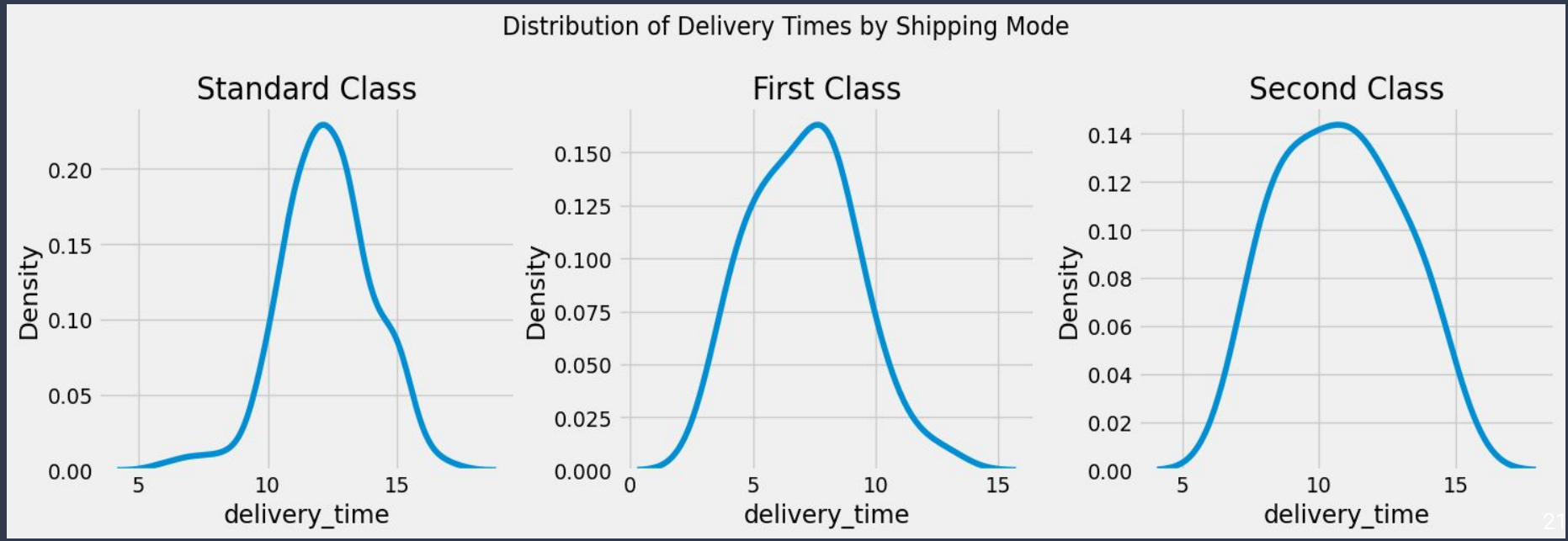


Top 10 busiest days of all time



Extra Credit

95th percentile = 15 days || In this context 95% of orders delivered within 15 days.



Thank you!