**PROJECT REPORT**Submitted in partial fulfilment of the requirements for the award of degree of  
**INT216 - PYTHON PROJECT**

**Dataset**: Data analysis (smart supply chain)

**SUBMITTED BY**

**Name of student: -** Manish Kumar Das  
**Registration Number: -** 12017595

**Roll Number: -** RRE267A06

**Section: -** RE267

**SUBMITTED TO**



**Lovely Professional University**

**(School of Computer Science and Engineering)**

Index

[Dataset-name: Data analysis (smart supply chain) 4](#_Toc120474123)

[What dataset am I working with the dataset? 4](#_Toc120474124)

[Why did I choose this dataset? 4](#_Toc120474125)

[Dataset dimension? 4](#_Toc120474126)

[Dataset Information 4](#_Toc120474127)

[Column’s Datatype: 6](#_Toc120474128)

[Planning 8](#_Toc120474129)

[Data Cleaning 8](#_Toc120474130)

[Univariate Analysis 9](#_Toc120474131)

[1. Which markets has most sales? 9](#_Toc120474132)

[2. Which region has most sales? 9](#_Toc120474133)

[3. Which country has made heighest orders? 10](#_Toc120474134)

[4. Which category of products has highest sales? 10](#_Toc120474135)

[5. Order Number respect to Delivery Status 12](#_Toc120474136)

[6. Numbers of orders of different customer segments 13](#_Toc120474137)

[7. Which category generates the most profit? 13](#_Toc120474138)

[8. Which payment methods customer prefer most? 14](#_Toc120474139)

[9. Which payment method is preferred the most by people in different regions? 14](#_Toc120474140)

[10. Which category products made most loss? 14](#_Toc120474141)

[11. Which region made most loss? 15](#_Toc120474142)

[12. Which region being suspected to the fraud the most? 16](#_Toc120474143)

[13. Which product is most frequently thought to be fraudulent in Western Europe? 16](#_Toc120474144)

[14. Which customers are conducting all these fraud? 17](#_Toc120474145)

[15. Which category of products are being delivered the most late? 18](#_Toc120474146)

[16. Late deliveried orders for different types of shipment method in all regions 18](#_Toc120474147)

[Bivariate Analysis 19](#_Toc120474148)

[1. Is there any relation between Sales & Product Price? 19](#_Toc120474149)

[2. Every year what average sales in each month is? 20](#_Toc120474150)

[3. Which quarter recorded highest sales? 20](#_Toc120474151)

[4. What is the average sells in years, month, weeks and hours? 21](#_Toc120474152)

[5. Check the frequency respect to Payment Type & Order Status 22](#_Toc120474153)

[6. Check whether there is any relation between Discount & Sales 22](#_Toc120474154)

[Multivariate Analysis 23](#_Toc120474155)

[1. Show relationships between two variables 23](#_Toc120474156)

[2. Histplot for all columns exclude object columns type 24](#_Toc120474157)

[Predictive Analysis 26](#_Toc120474158)

[1. Understanding Customer Needs 26](#_Toc120474159)

[Hypothesis Testing 28](#_Toc120474160)

[Conclusion 28](#_Toc120474161)

[References 29](#_Toc120474162)

# Dataset-name: Data analysis (smart supply chain)

## What dataset am I working with the dataset?

I am working with a dataset of ***supply chain by Dataco Global Company*** for my EDA project/Analysis.

**Dataset link**: <https://www.kaggle.com/code/sanketchavan5595/data-analysis-smart-supply-chain/data>

## Why did I choose this dataset?

The supply chain is an overriding part of any online business. The supply chain starts with vendors on one side followed by a manufacturer and finally distribution of goods to consumers.

After the analysis, we will understand the importance of the supply chain in the present business scenario. What are the challenges of supply chain management? It will be very easy to analyze what steps can be taken to solve those challenges.

We can do several types of analysis like, consumer purchase rates, payment methods consumers prefer, which products cost the most, country vs. purchase ratio, etc.

## Dataset dimension?

Row × Columns = **180519 × 53**

## Dataset Information

Colum’s Description:

|  |  |
| --- | --- |
| **COLUMN** | **DESCRIPTION** |
| Type | *Type of transaction made* |
| Days for shipping (real) | *Actual shipping days of the purchased product* |
| Days for shipment (scheduled) | *Days of scheduled delivery of the purchased product* |
| Benefit per order | *Earnings per order placed* |
| Sales per customer | *Total sales per customer made per customer* |
| Delivery Status | *Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time* |
| Late\_delivery\_risk | *Categorical variable that indicates if sending is late (1), it is not late (0).* |
| Category Id | *Product category code* |
| Category Name | *Description of the product category* |
| Customer City | *City where the customer made the purchase* |
| Customer Country | *Country where the customer made the purchase* |
| Customer Email | *Customer’s email* |
| Customer Fname | *Customer name* |
| Customer Id | *Customer ID* |
| Customer Lname | *Customer lastname* |
| Customer Password | *Masked customer key* |
| Customer Segment | *Types of Customers: Consumer , Corporate , Home Office* |
| Customer State | *State to which the store where the purchase is registered belongs* |
| Customer Street | *Street to which the store where the purchase is registered belongs* |
| Customer Zipcode | *Customer Zipcode* |
| Department Id | *Department code of store* |
| Department Name | *Department name of store* |
| Latitude | *Latitude corresponding to location of store* |
| Longitude | *Longitude corresponding to location of store* |
| Market | *Market to where the order is delivered : Africa, Europe, LATAM, Pacific Asia, USCA* |
| Order City | *Destination city of the order* |
| Order Country | *Destination country of the order* |
| Order Customer Id | *Customer order code* |
| order date (DateOrders) | *Date on which the order is made* |
| Order Id | *Order code* |
| Order Item Cardprod Id | *Product code generated through the RFID reader* |
| Order Item Discount | *Order item discount value* |
| Order Item Discount Rate | *Order item discount percentage* |
| Order Item Id | *Order item code* |
| Order Item Product Price | *Price of products without discount* |
| Order Item Profit Ratio | *Order Item Profit Ratio* |
| Order Item Quantity | *Number of products per order* |
| Sales | *Value in sales* |
| Order Item Total | *Total amount per order* |
| Order Profit Per Order | *Order Profit Per Order* |
| Order Region | *Region of the world where the order is delivered : Southeast Asia, South Asia, Oceania, Eastern Asia, West Asia, West of USA, US Center, West Africa, Central Africa, North Africa, Western Europe, Northern, Caribbean, South America, East Africa, Southern Europe, East of USA, Canada, Southern Africa, Central Asia, Europe, Central America, Eastern Europe, South of USA* |
| Order State | *State of the region where the order is delivered* |
| Order Status | *Order Status : COMPLETE, PENDING, CLOSED, PENDING\_PAYMENT, CANCELED, PROCESSING, SUSPECTED\_FRAUD, ON\_HOLD, PAYMENT\_REVIEW* |
| Product Card Id | *Product code* |
| Product Category Id | *Product category code* |
| Product Description | *Product Description* |
| Product Image | *Link of visit and purchase of the product* |
| Product Name | *Product Name* |
| Product Price | *Product Price* |
| Product Status | *Status of the product stock :If it is 1 not available , 0 the product is available* |
| Shipping date (DateOrders) | *Exact date and time of shipment* |
| Shipping Mode | *The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day* |

### **Column’s Datatype**:

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Type 180519 non-null object

1 Days for shipping (real) 180519 non-null int64

2 Days for shipment (scheduled) 180519 non-null int64

3 Benefit per order 180519 non-null float64

4 Sales per customer 180519 non-null float64

5 Delivery Status 180519 non-null object

6 Late\_delivery\_risk 180519 non-null int64

7 Category Id 180519 non-null int64

8 Category Name 180519 non-null object

9 Customer City 180519 non-null object

10 Customer Country 180519 non-null object

11 Customer Email 180519 non-null object

12 Customer Fname 180519 non-null object

13 Customer Id 180519 non-null int64

14 Customer Lname 180511 non-null object

15 Customer Password 180519 non-null object

16 Customer Segment 180519 non-null object

17 Customer State 180519 non-null object

18 Customer Street 180519 non-null object

19 Customer Zipcode 180516 non-null float64

20 Department Id 180519 non-null int64

21 Department Name 180519 non-null object

22 Latitude 180519 non-null float64

23 Longitude 180519 non-null float64

24 Market 180519 non-null object

25 Order City 180519 non-null object

26 Order Country 180519 non-null object

27 Order Customer Id 180519 non-null int64

28 order date (DateOrders) 180519 non-null object

29 Order Id 180519 non-null int64

30 Order Item Cardprod Id 180519 non-null int64

31 Order Item Discount 180519 non-null float64

32 Order Item Discount Rate 180519 non-null float64

33 Order Item Id 180519 non-null int64

34 Order Item Product Price 180519 non-null float64

35 Order Item Profit Ratio 180519 non-null float64

36 Order Item Quantity 180519 non-null int64

37 Sales 180519 non-null float64

38 Order Item Total 180519 non-null float64

39 Order Profit Per Order 180519 non-null float64

40 Order Region 180519 non-null object

41 Order State 180519 non-null object

42 Order Status 180519 non-null object

43 Order Zipcode 24840 non-null float64

44 Product Card Id 180519 non-null int64

45 Product Category Id 180519 non-null int64

46 Product Description 0 non-null float64

47 Product Image 180519 non-null object

48 Product Name 180519 non-null object

49 Product Price 180519 non-null float64

50 Product Status 180519 non-null int64

51 shipping date (DateOrders) 180519 non-null object

52 Shipping Mode 180519 non-null objec

## Planning

1. Analysis the quantity and distribution of data.
2. Dataset cleaning like remove duplicate/irrelevant/redundant column

* We can saw that some columns are duplicate which contain same data. So, we can drop one of the duplicate columns. i.e.: “Order Profit Per Order” & Benefit per order”

1. Handling Attributes w/o Variance

* As seen in checking for variance in the data, columns that have no variance (e.g. only one value) can also be excluded.

1. Handling NaN values

* Order\_Zipcode
* Product\_Description

1. Aggregating Department Information

* We can aggregate information available about different apartments to facilitate management decision making.

1. Mapping Supply Chain as a Bipartite Graph

* We can also map the supply chain as a bipartite graph where one node set represents department stores and the other represents customers (customer regions need to be more specific.)

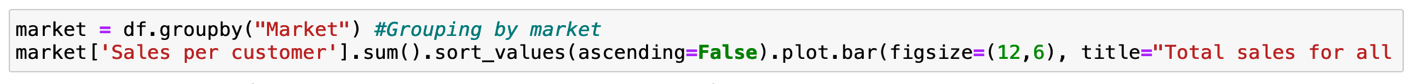
1. Heatmap for correlation matrix
2. Delvery status according to country
3. Analysis the type of customers. Like: Individual, Corporate, Home office
4. We can analyze that which type of product/which categories product consumers order most
5. We can analyze that which type of product/which categories product consumers order most by region/country wise as well
6. Analyze the numbers of order orders respect to country
7. Which year minimum/maximum sales for a particular product.
8. Various plots/Graph like Barplot, Scatterplot, PieChart, Kde plot, Boxplot, Violinplot, Bargraph, Countplot, Crosstablish, stackedbar, heatmaps
9. Hypothesis testing

## Data Cleaning

1. Two columns have Invalid Datatype I corrected it.
2. Remove the duplicate columns by writing User-defined-function called “***getDuplicateColumns***”.
3. Remove the columns which contains only single values for all of the rows like password & email column etc. For that I used “***getOneDistinctValueCols***” functions
4. I merge two column called “***Customer Fname***” & “***Customer Lname***” and Make a new Column named “**Customer Name**” & Drop those 2 columns.
5. Then I checked that two columns contains lots of null values & this column is “***Customer Zipcode***”. These columns are not necessary for out data analysis so I drop these 2 column.
6. Drop some others columns like “Product Image”, “Latitude” and “Longitude” columns as it is useless for analysis

## Univariate Analysis

### Which markets has most sales?

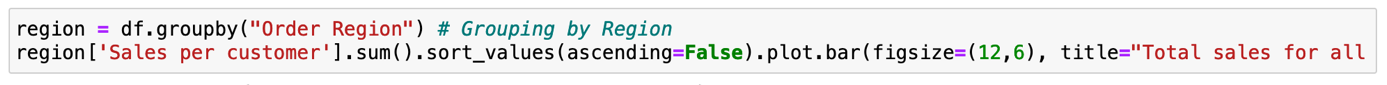


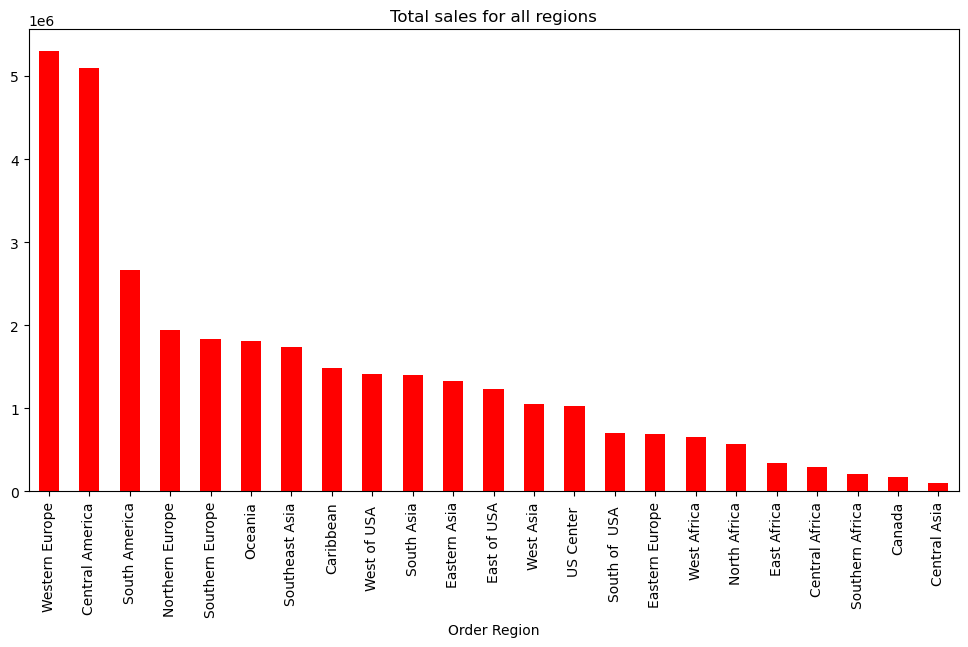
Chart, bar chart

Description automatically generated



### Which region has most sales?





### Which country has made heighest orders?

Graphical user interface, text, application, email

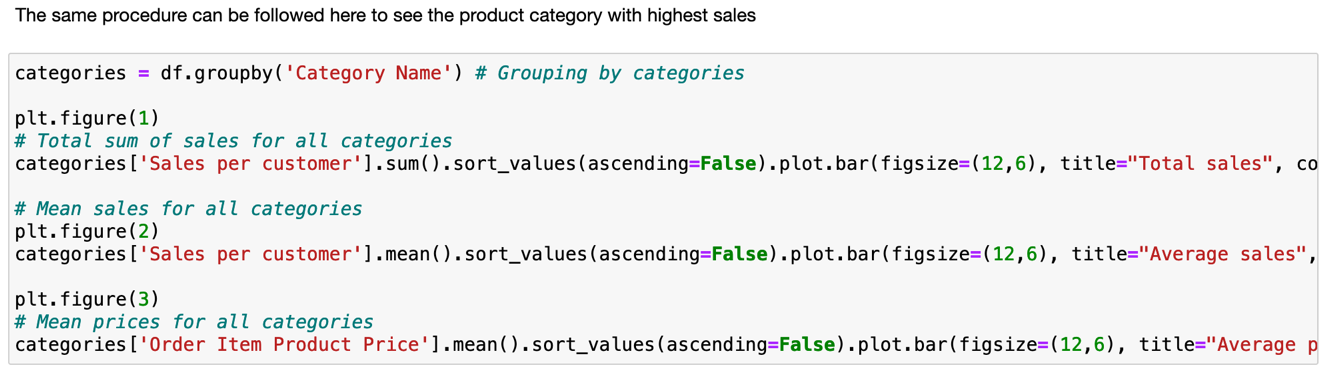
Description automatically generated

Chart, bar chart

Description automatically generated



### Which category of products has highest sales?



Chart, histogram

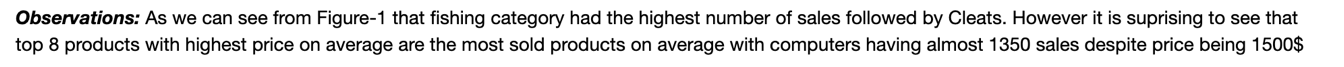
Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

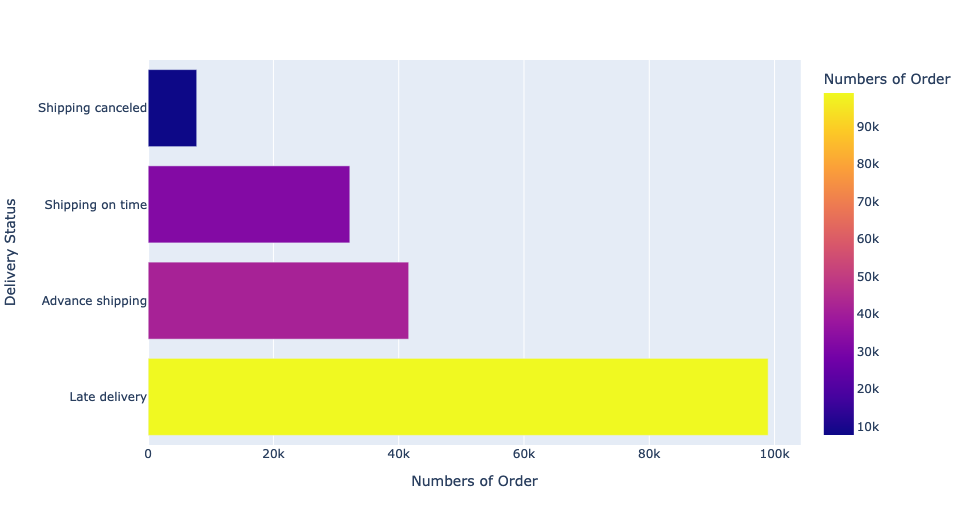
Description automatically generated



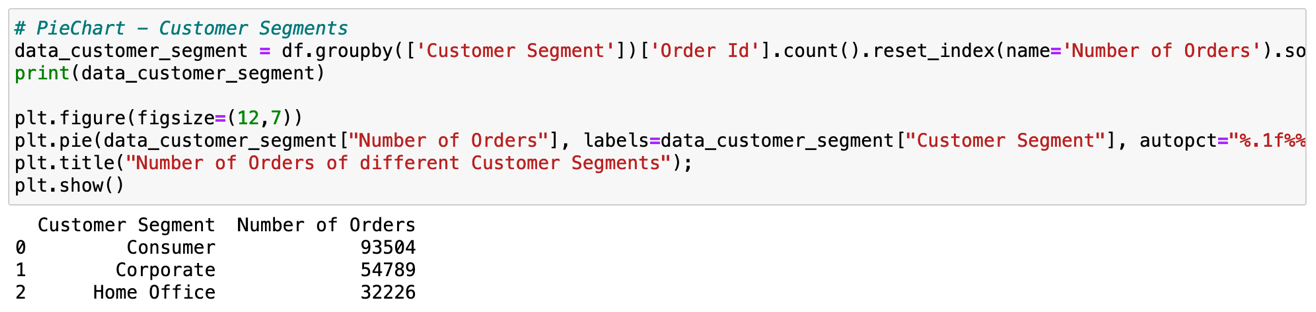
### Order Number respect to Delivery Status

Graphical user interface, text, application, email

Description automatically generated



### Numbers of orders of different customer segments



Chart, pie chart

Description automatically generated



### Which category generates the most profit?

Graphical user interface, application

Description automatically generated with medium confidence

Chart, bar chart, histogram

Description automatically generated



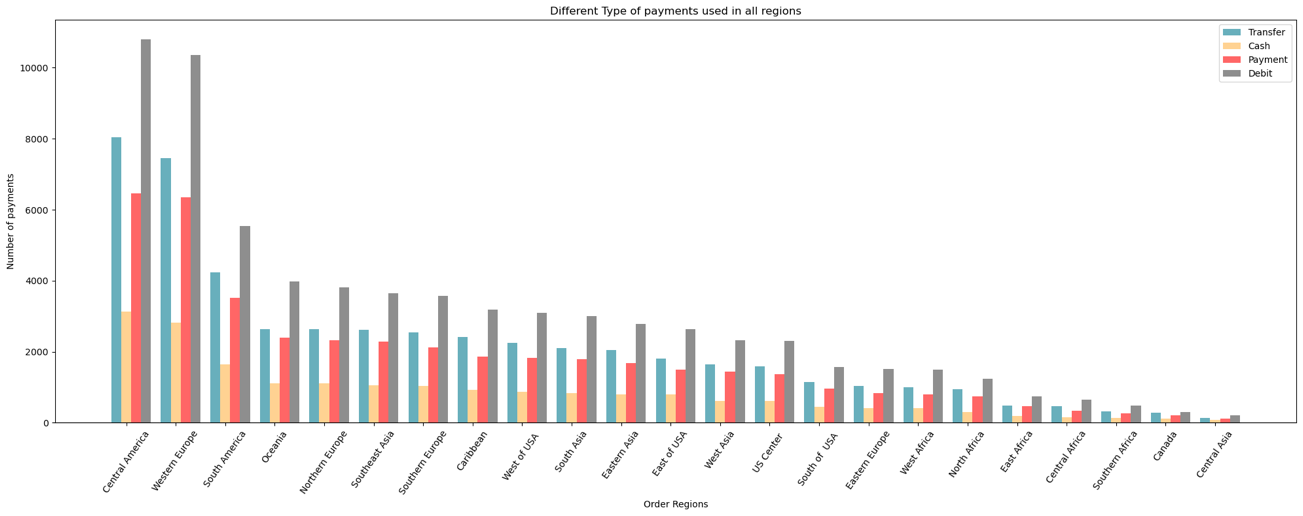
### Which payment methods customer prefer most?

Graphical user interface, application

Description automatically generated

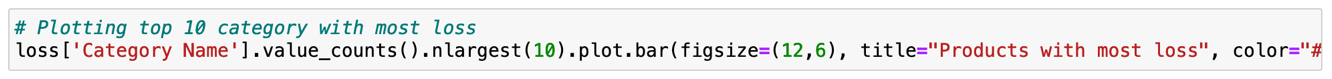
### Which payment method is preferred the most by people in different regions?







### Which category products made most loss?



Chart, bar chart

Description automatically generated



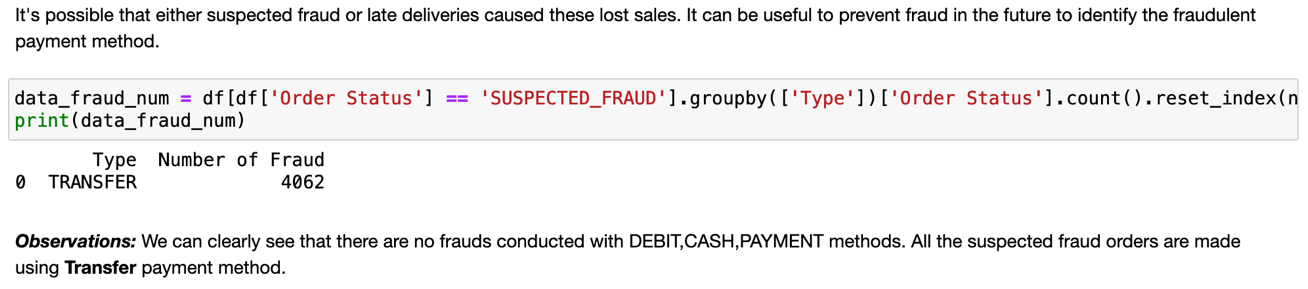
### Which region made most loss?



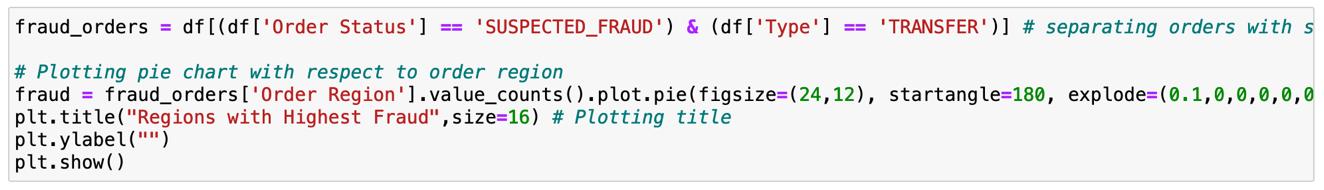
Chart, bar chart

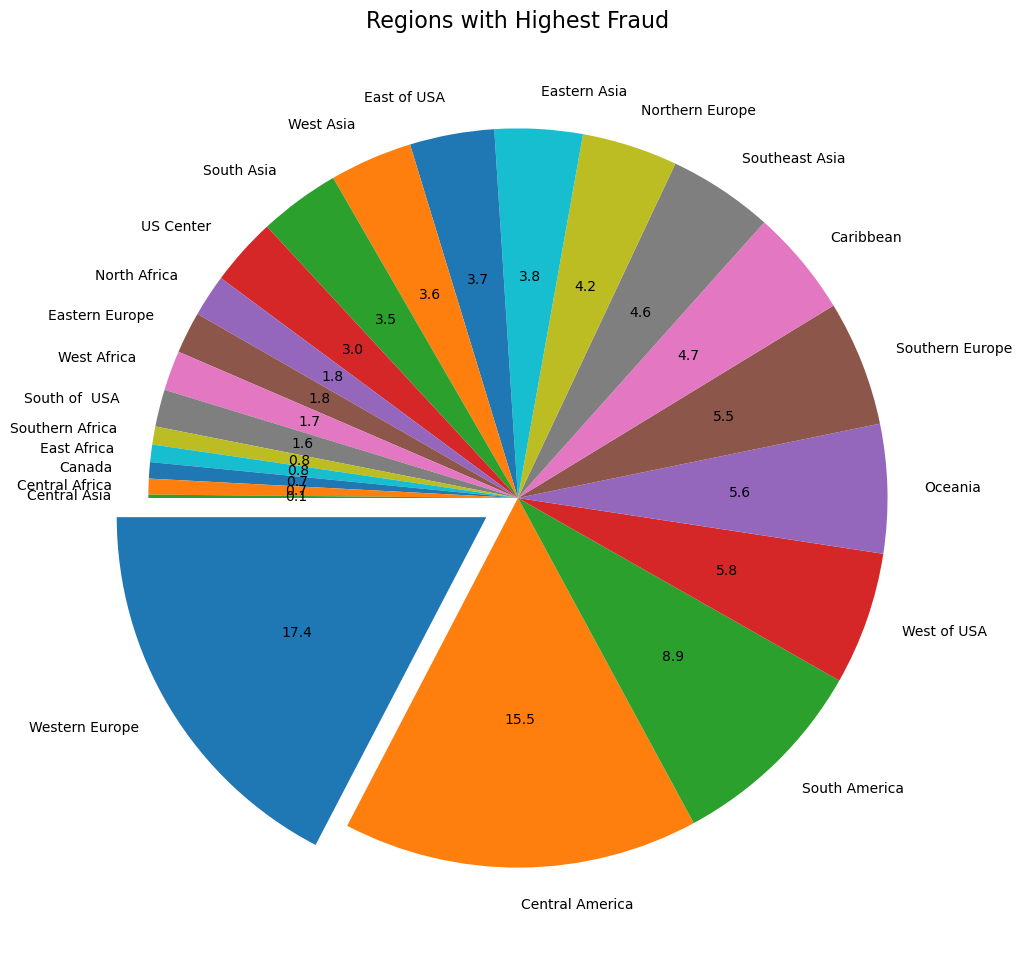
Description automatically generated

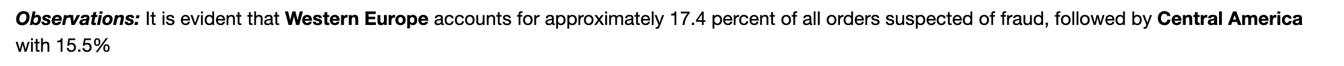




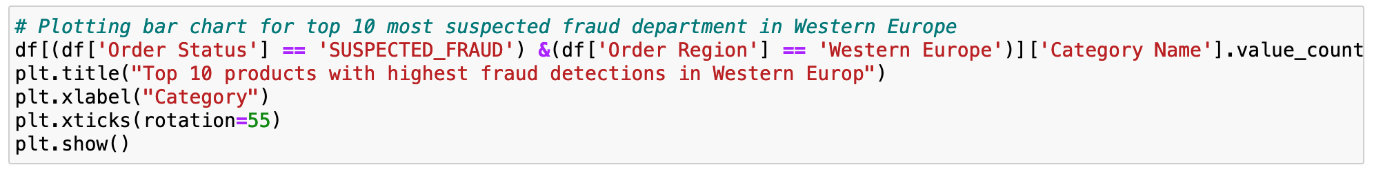
### Which region being suspected to the fraud the most?

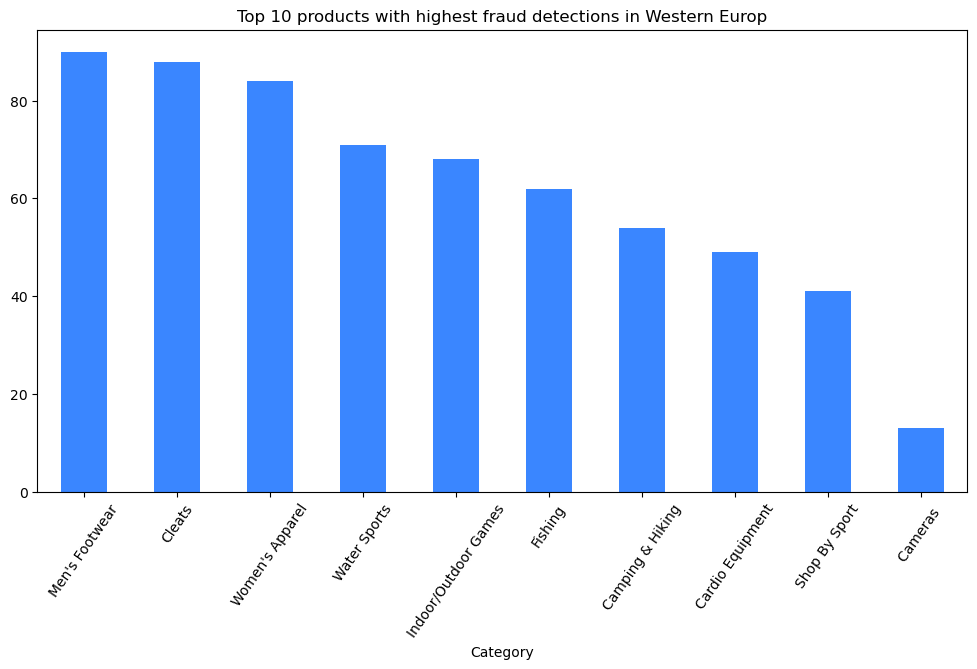


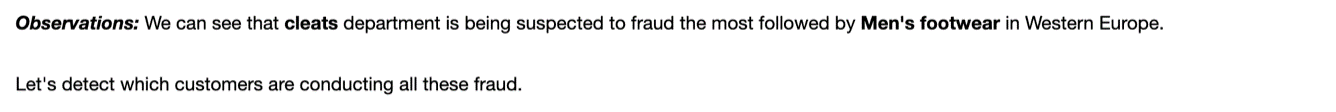




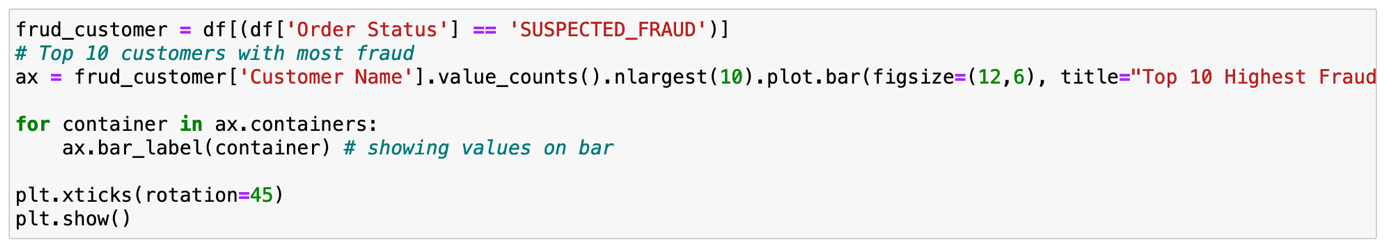
### Which product is most frequently thought to be fraudulent in Western Europe?

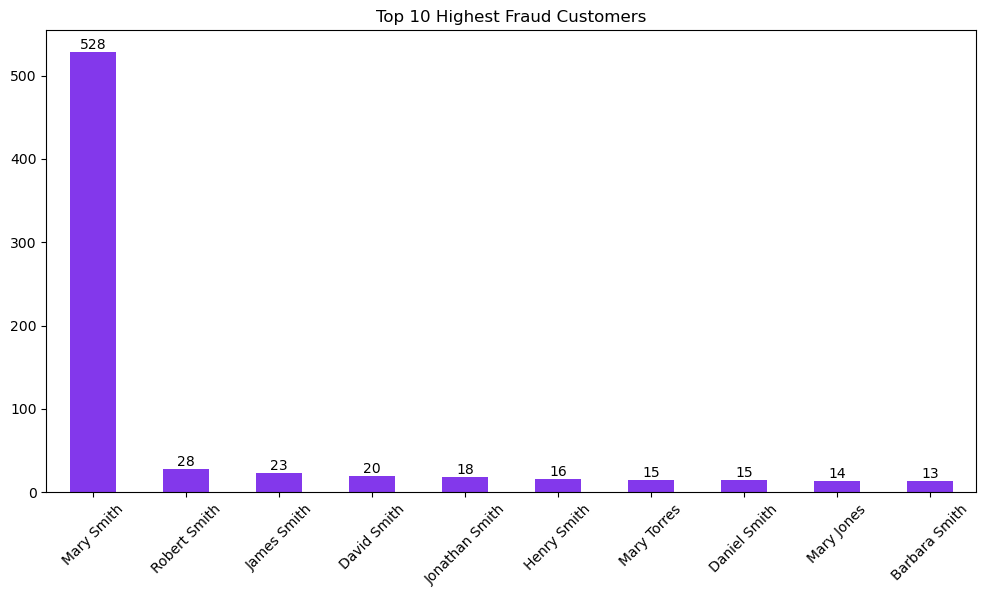


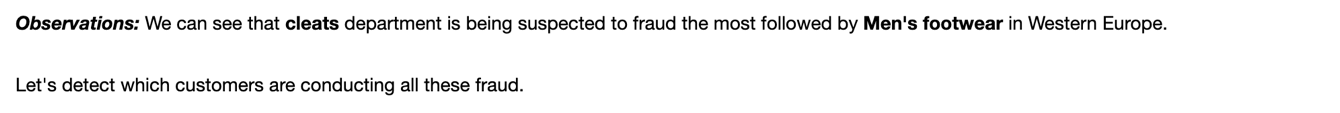




### Which customers are conducting all these fraud?







### Which category of products are being delivered the most late?

Graphical user interface, text, application

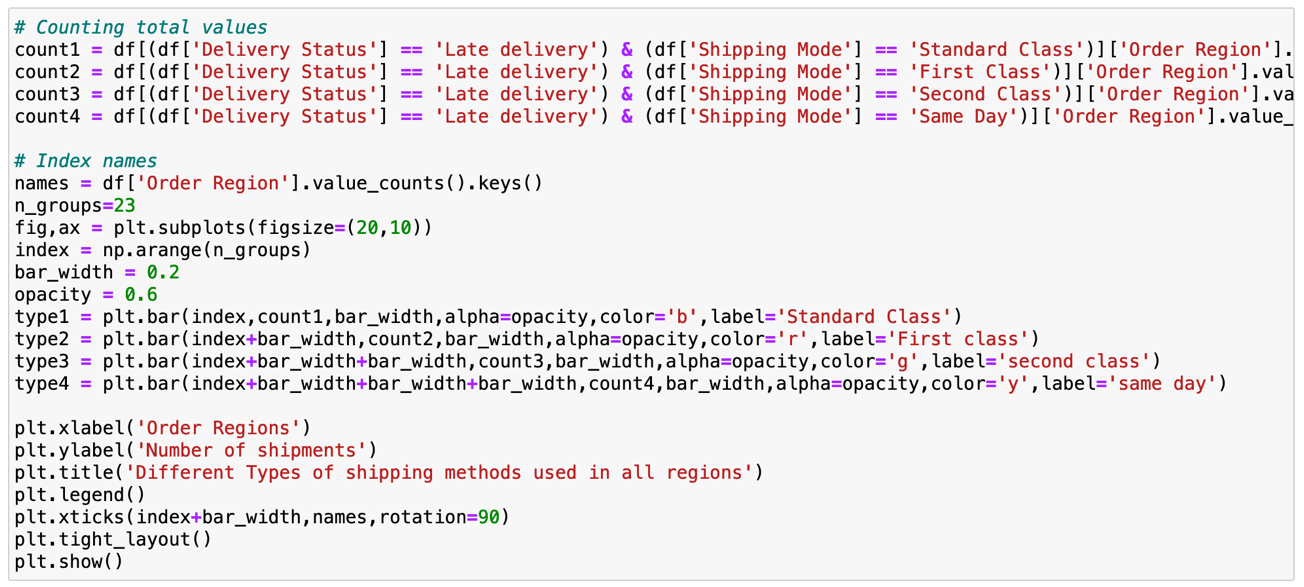
Description automatically generated

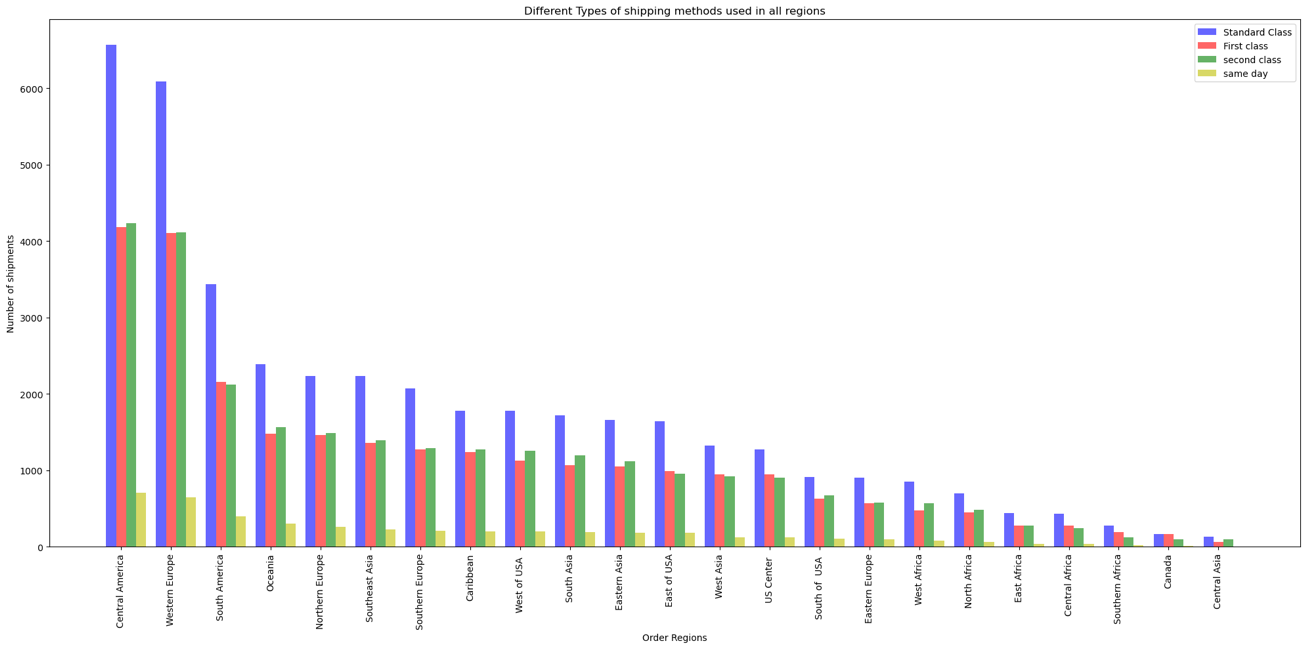
Chart, bar chart

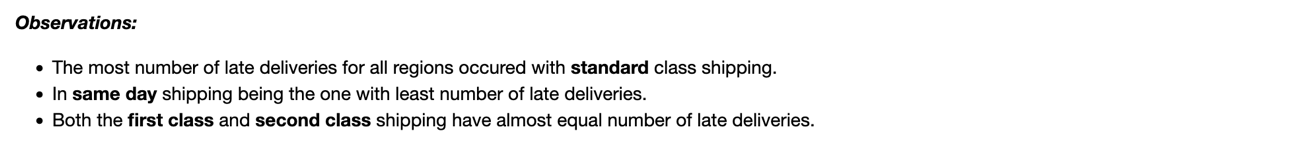
Description automatically generated



### Late deliveried orders for different types of shipment method in all regions





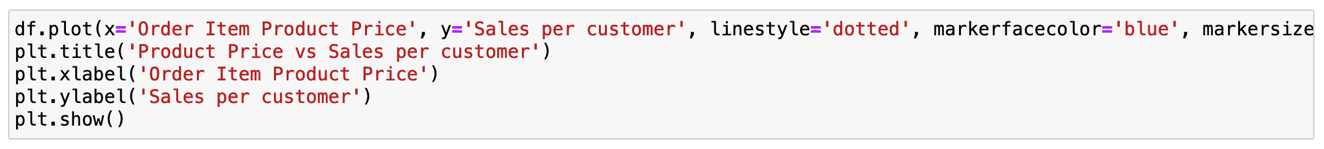


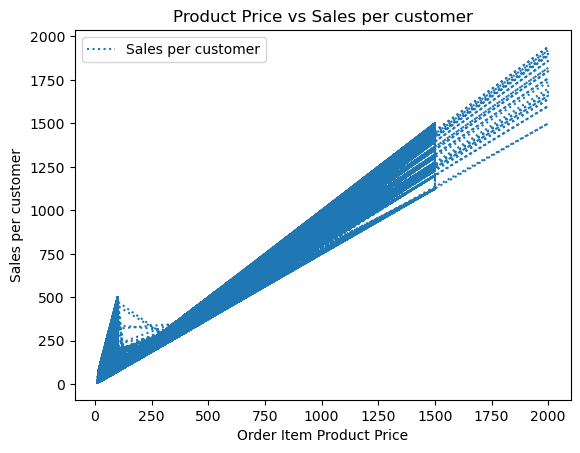
A picture containing table

Description automatically generated

## Bivariate Analysis

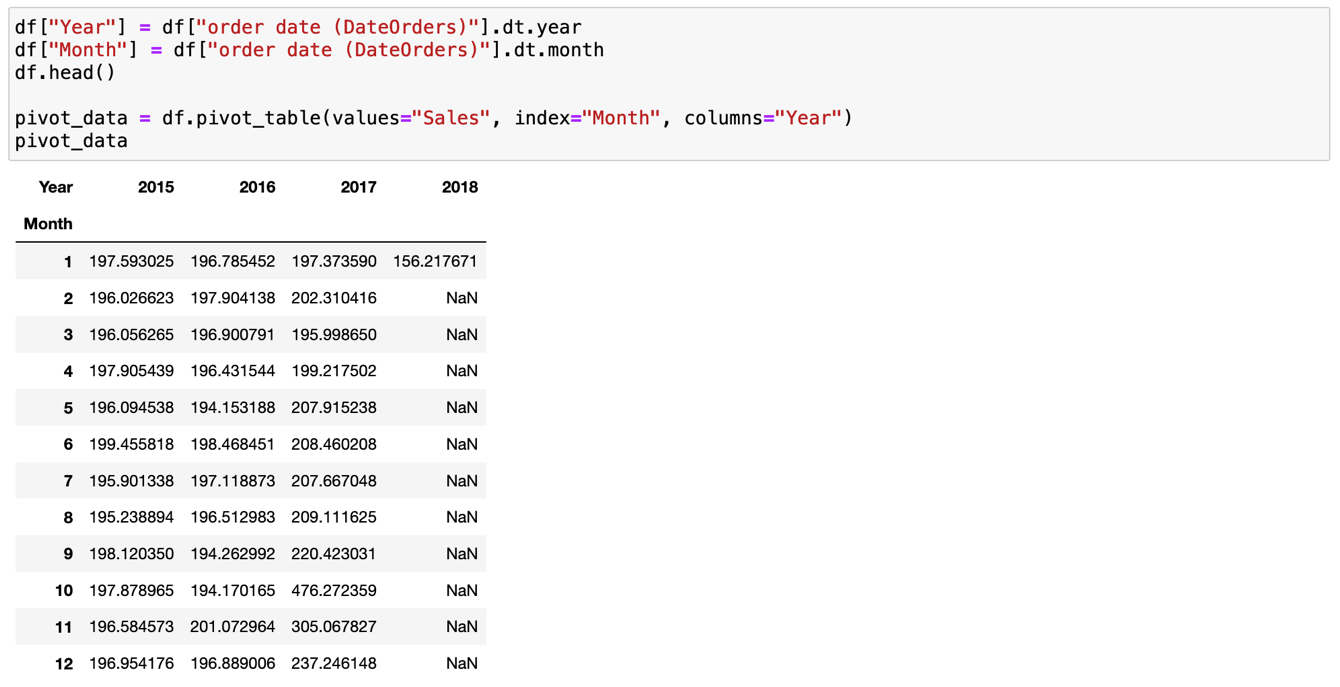
### Is there any relation between Sales & Product Price?



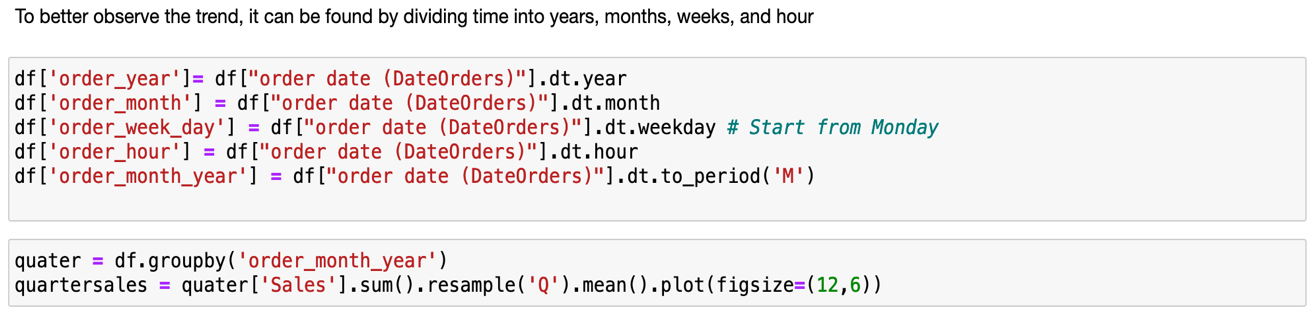




### Every year what average sales in each month is?



### Which quarter recorded highest sales?



Chart, line chart

Description automatically generated



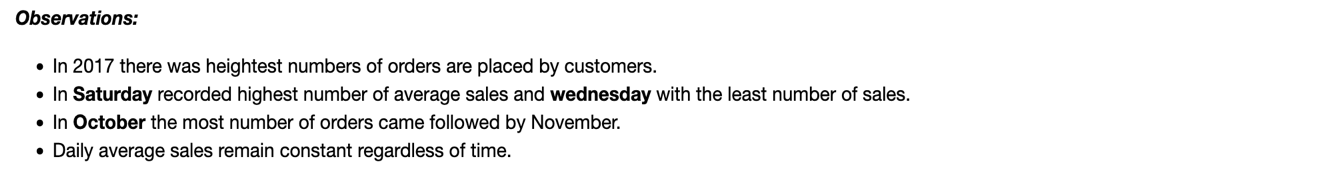
### What is the average sells in years, month, weeks and hours?

Graphical user interface

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated



### Check the frequency respect to Payment Type & Order Status

Table

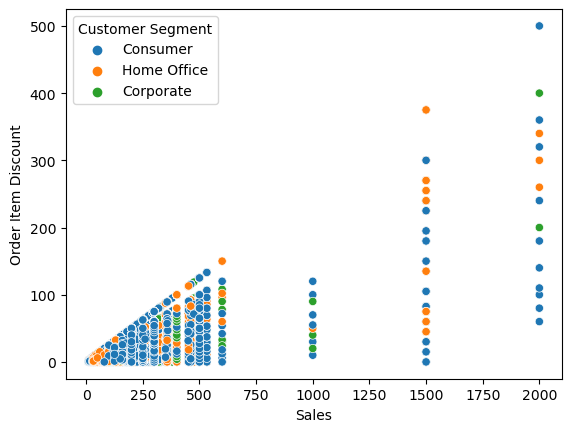
Description automatically generated

Chart, bar chart

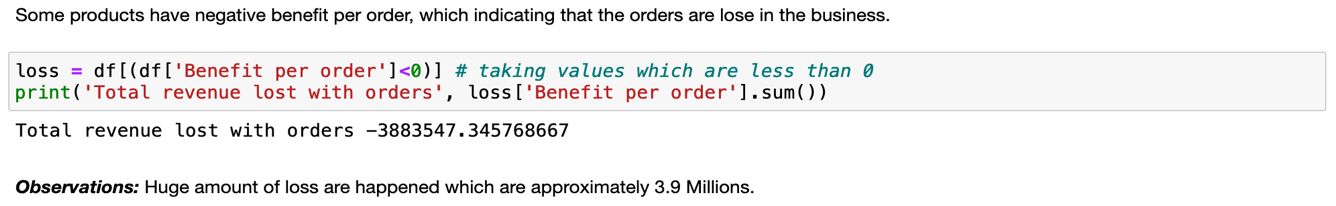
Description automatically generated

### Check whether there is any relation between Discount & Sales



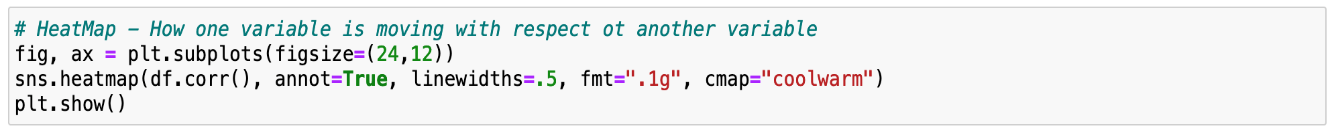


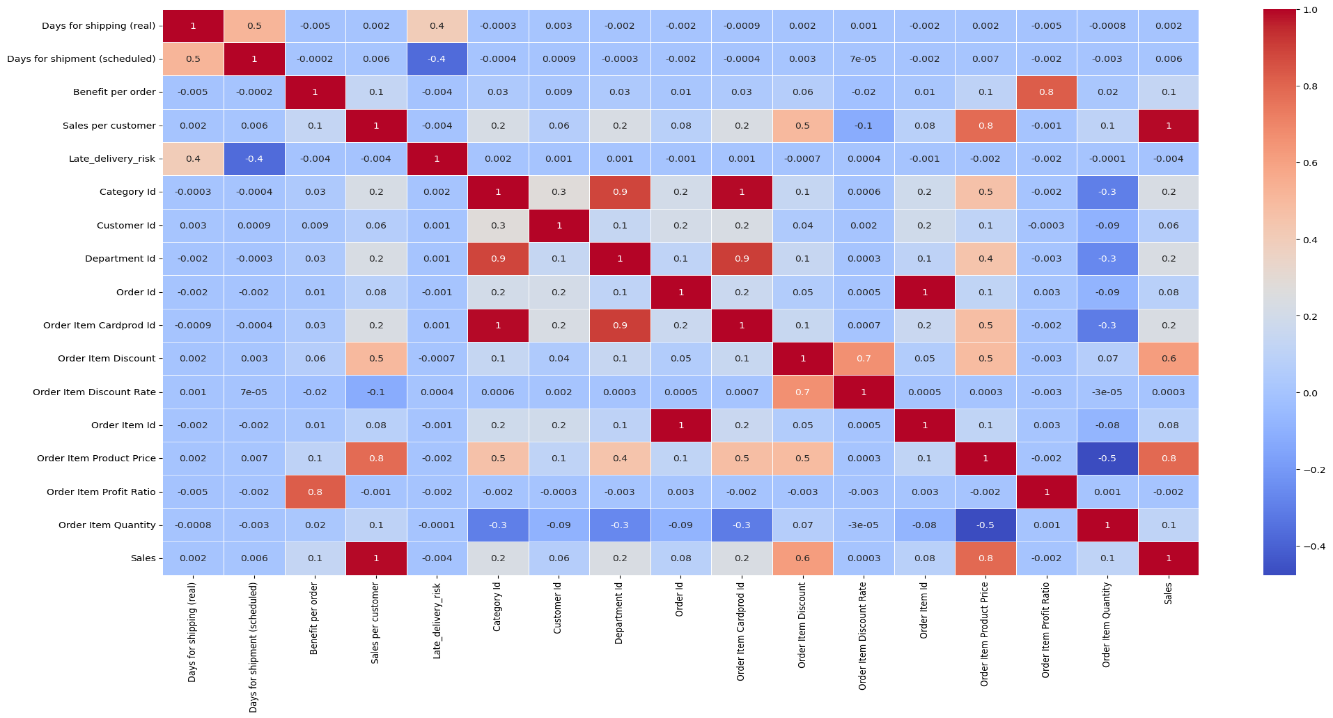




## Multivariate Analysis

### Show relationships between two variables



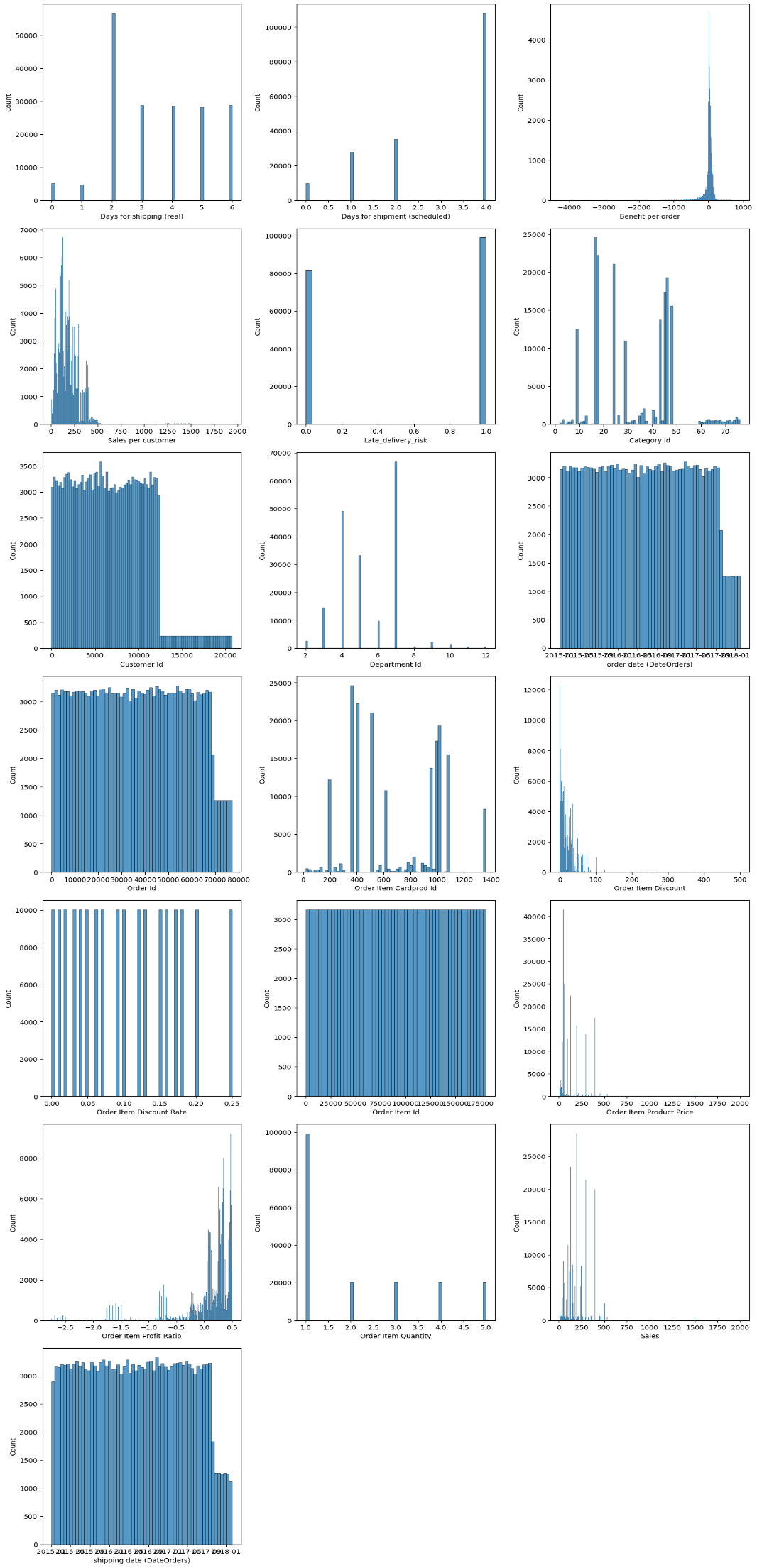


Background pattern

Description automatically generated with low confidence

### Histplot for all columns exclude object columns type



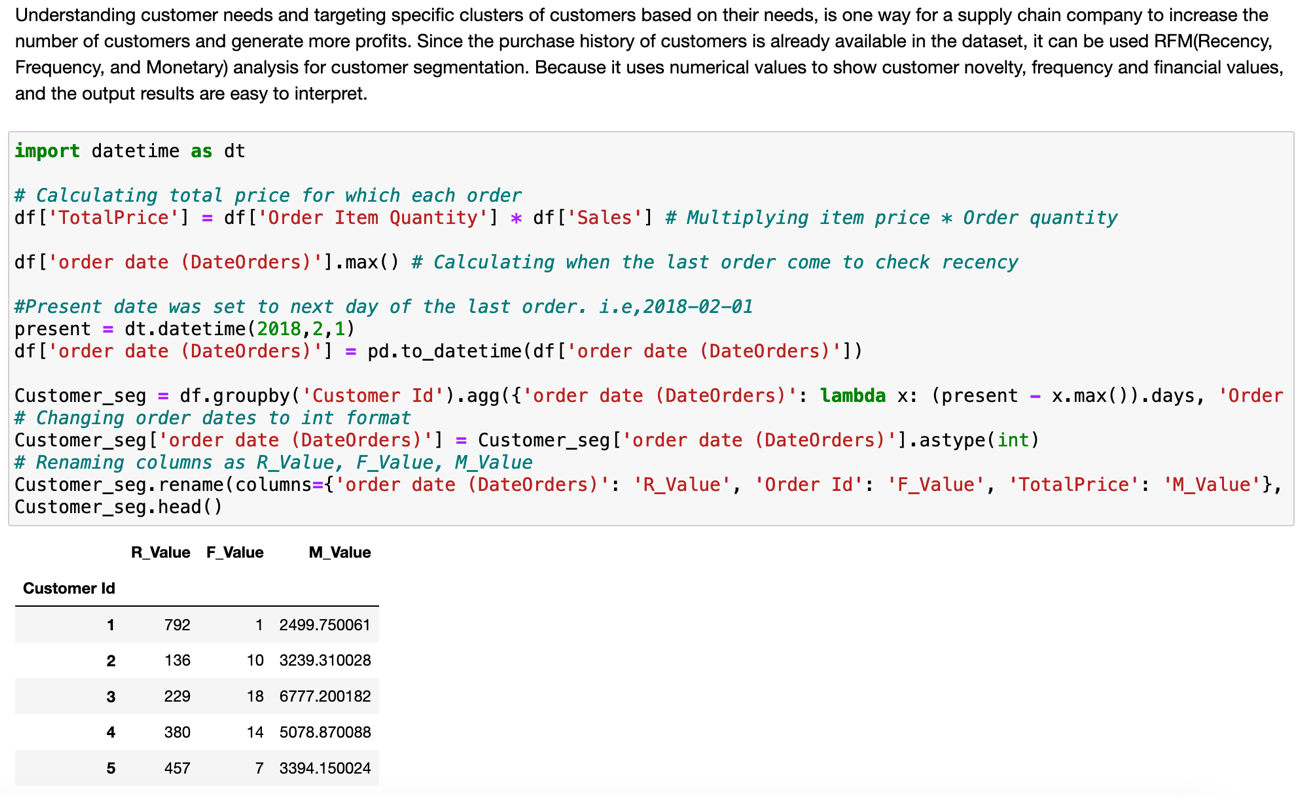


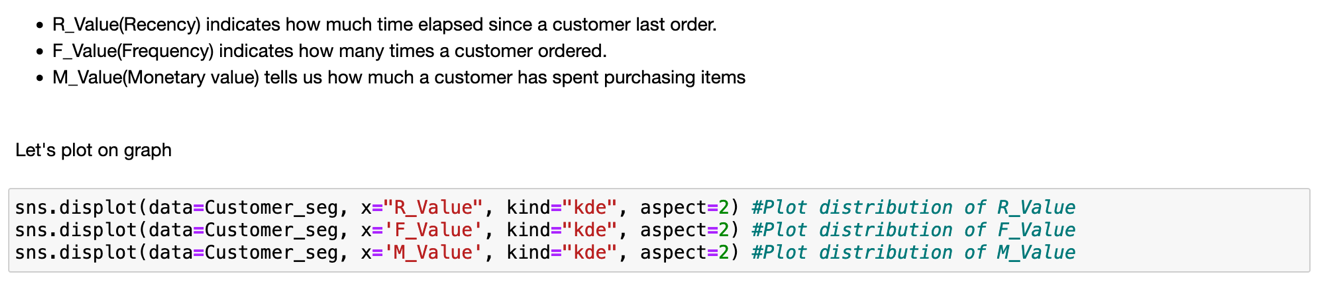
Background pattern

Description automatically generated with low confidence

## Predictive Analysis

### Understanding Customer Needs





Chart, histogram

Description automatically generated

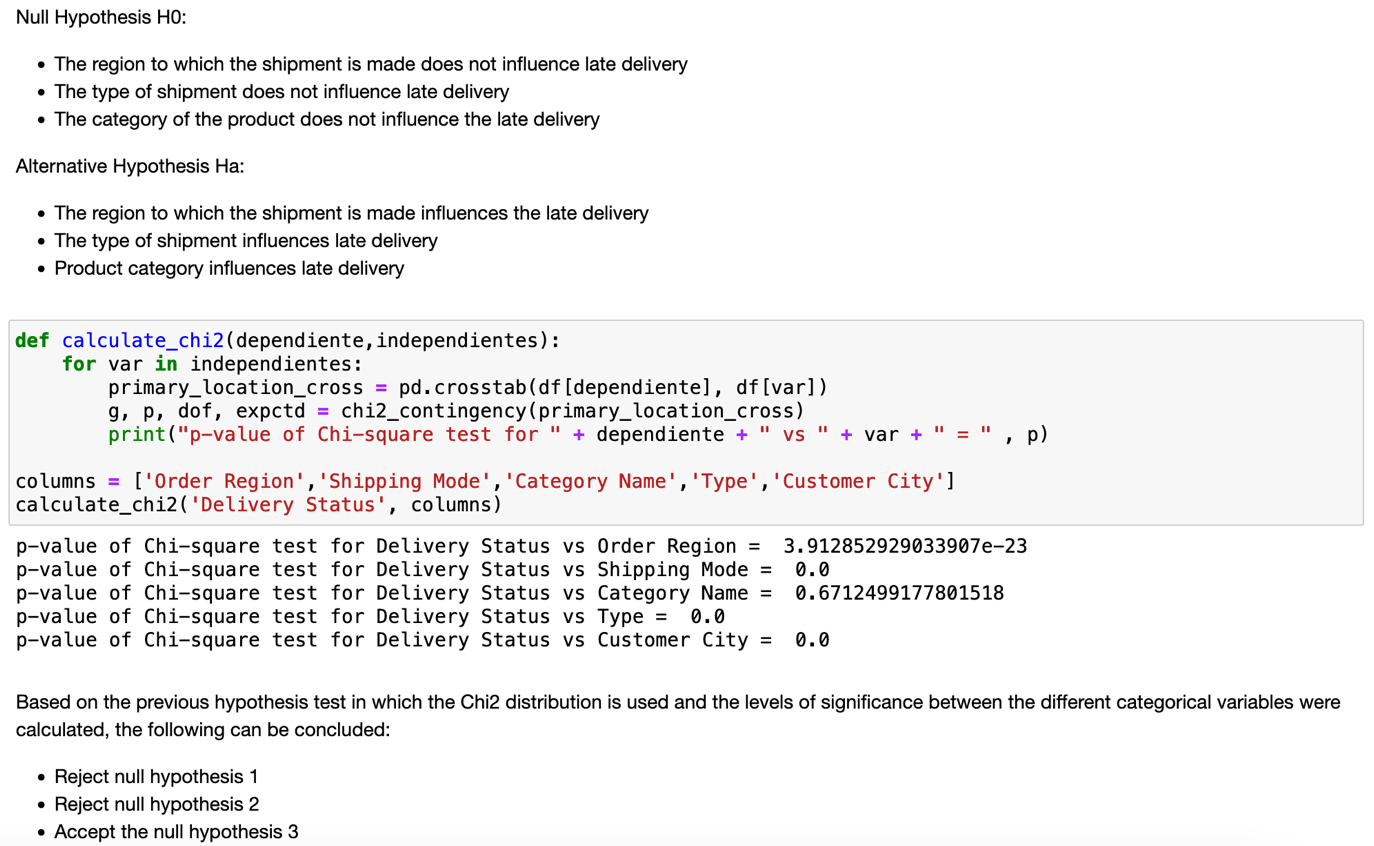
Chart, line chart, histogram

Description automatically generated

Chart

Description automatically generated

## Hypothesis Testing



## Conclusion

After DataCo analyzed the company's dataset it discovered that Western Europe and Central America are both the regions with the highest number of sales, but the company lost the most revenue from these regions only. And both these regions have the highest number of fraudulent transactions and more late delivery orders are suspected. More than half of shoppers here shop in person. Then corporate and home office. The highest profit is from the fishing category. The company's total sales were consistent till 2017 quarter 3 and total sales increased by 10% quarter over quarter and then suddenly declined by almost 65% in 2018 quarter 1. October and November are the months with the highest sales in the total year.

Saturdays are the most sold. Most people prefer to pay through debit card and all fraudulent transactions are happening with wire transfer so the company should be careful when customers are using wire transfer as the company has been scammed more than 528 times by a single customer. Products in the Cleats, Men's Shoes, and Women's Apparel categories are causing most orders to be delivered late, and these products are most likely to be suspected of fraud. In addition, hypothesis testing suggests that shipping class has a relationship with late delivery.

## References

* <https://www.geeksforgeeks.org/how-to-find-drop-duplicate-columns-in-a-pandas-dataframe/>
* <https://www.kaggle.com/code/sanketchavan5595/data-analysis-smart-supply-chain/data>
* <https://medium.com/datadriveninvestor/building-neural-network-using-keras-for-classification-3a3656c726c1>
* <https://medium.com/epfl-extension-school/advanced-exploratory-data-analysis-eda-with-python-536fa83c578a>
* <https://towardsdatascience.com/exploratory-data-analysis-eda-visualization-using-pandas-ca5a04271607>
* <https://matplotlib.org/stable/gallery/color/named_colors.html>
* <https://www.geeksforgeeks.org/how-to-find-drop-duplicate-columns-in-a-pandas-dataframe/>