****

**Guidelines for PGPDSE FT Capstone Project – Interim Report**

**Project Group Info:**

|  |  |
| --- | --- |
| **Batch details** | Batch 1 |
| **Team members** | 1.Fayiq Ahmed  2.Supriyo Dey  3.Sai Charan  4.Akhil Raj  5.Bharath |
| **Domain of Project** | Market Analysis |
| **Proposed project title** | DataCo Smart Supply Chain Data Analysis |
| **Group Number** | 1 |
| **Team Leader** | Fayiq Ahmed |
| **Mentor Name** | Mr. Animesh Tiwari |

**Date**: 28/04/2021

**Signature of the Mentor Signature of the Team Leader**

**Industry Review**

A supply chain is an entire system of producing and delivering a product or service, from the very beginning stage of sourcing the raw materials to the final delivery of the product or service to end-users. The supply chain lays out all aspects of the production process, including the activities involved at each stage, information that is being communicated, natural resources that are transformed into useful materials, human resources, and other components that go into the finished product or service.



**Business Problem Understanding**

In recent years, the rise of the Internet of things (IoT) as an emerging technology has been unbelievable, more companies are moving towards the adoption of these technologies and many IoT sensors are being deployed to share information in real-time which leads to the generation of a huge amount of data. This data when used correctly, will be very helpful to the company to discover hidden patterns for better decision making in the future. For example, with the DataCo company, dataset customer segmentation analysis that will be performed in this project will help the company to better understand its customers and target them to increase customer responsiveness and the company's revenue. Along with that Suspected Fraudulent product category, products, regions and customers will also be identified through proper Data Science Techniques. With a lot of options available to analyse data, it is very difficult to decide which method and machine learning model to use since the performance of the model vary on the parameters available in the data. We will be using the below mentioned corners of Data science and accordingly will fetch the required outputs.

**Business Objective**

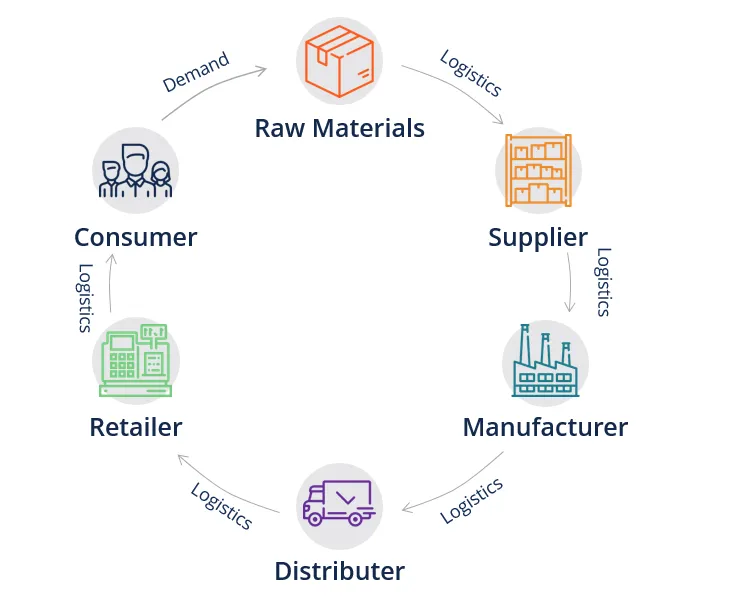
* Detection of high risk (fraudulent) product category, products along with demographic characteristics and personal attributes of customer who are susceptible / contributing towards fraudulent activities.
* High-risk payment modes (transactions) of DataCo Company.
* Customer Segment through RFM Analysis.

**Background Research:**

### **Why Should A Company Understand Its Supply Chain?**

Mapping out a supply chain is one of the critical steps in performing an external analysis in a strategic planning process. The importance of clearly laying out the supply chain is that it helps a company define its own market and decide where it wants to be in the future. In developing corporate-level strategies, a company often needs to make decisions on whether to operate a single line of business or enter into other related or unrelated industries.

Each stage of a supply chain is essentially a different industry, for example, raw material extraction and manufacturing. The supply chain enables a company to understand others that are involved in each of the stages, and thereby provides some insights on the attractiveness or competitiveness in industries the company might want to enter in the future.



**Normal Supply Chain E – Commerce Supply Chain**

**Applications of Supply Chain Analytics:**

**By Solution**

Visualization & Reporting, Sales & Operations Analytics, Manufacturing Analytics,

Other (Planning & Procurement)

**By Deployment Model**

On-Premises, Cloud-Based

**By Industry Vertical**

Manufacturing, Retail, Healthcare, Transportation & Logistics, Others (Aerospace &Defense)

**Literature Survey**

1. <https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf>
2. <https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining>
3. <https://www.analyticsvidhya.com/blog/2016/02/bigmart-sales-solution-top-20/>
4. <https://data.mendeley.com/datasets/8gx2fvg2k6/5>
5. <http://dx.doi.org/10.17632/8gx2fvg2k6.5#file-5046ef5f-6df4-4ee7-9eb8-b33456b0d49e>
6. <https://towardsdatascience.com/find-your-best-customers-with-customer-segmentation-in-python-61d602f9eee6>
7. <https://www.grandviewresearch.com/industry-analysis/the-global-supply-chain-analytics-market>

**Dataset and Domain**

**Data Dictionary**

There are 180519 rows and 53 columns in the original dataset. Columns are as follows -

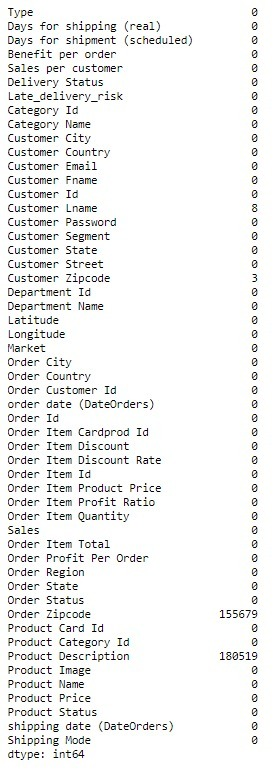
|  |  |
| --- | --- |
| Type | This is the feature for the payment modes available to the company |
| Days for shipping (real) | Actual shipping duration up to the product is delivered |
| Days for shipment (scheduled) | Estimated shipping time up to the product is delivered |
| Benefit per order | Benefit received by the company for each order irrespective of qty |
| Sales per customer | Sales done by the company to all the unique customers |
| Delivery Status | Shipment Tracking status at the time of delivery |
| Late delivery risk | Binary field (0/1), 0 – No risk, 1 – Yes there is risk |
| Category Id | Product domain / category unique ID |
| Category Name | Names of Product Class / category |
| Customer City | Residential city of customer |
| Customer Country | Residential country of customer |
| Customer Email | Customer mail registered with the company |
| Customer Fname | Customer’s First Name |
| Customer Id | Unique ID assigned to every customer |
| Customer Lname | Customer’s Last Name |
| Customer Password | Password for accessing the respective accounts |
| Customer Segment | Category of customers |
| Customer State | State of customer coded by the company |
| Customer Street | Residential address of customers |
| Customer Zip code | Pin code of the corresponding address |
| Department Id | Supply Chain Storage department Id |
| Department Name  Latitude | Storage department Name  Customer Location dimension |
| Longitude | Customer location dimension |
| Market | Continent of residence of customer |
| Order City | Destination city where order has to be delivered |
| Order Country | Destination country where order has to be delivered |
| Order Customer Id | Same as Customer ID |
| order date (Date Orders) | Date when the order was placed |
| Order Id | Unique Order ID |
| Order Item Card prod Id | Based on storage department it is being assigned |
| Order Item Discount | Amount of Discount given on a product. |
| Order Item Discount Rate | Discount Percent |
| Order Item Id | Unique Item ID in an order |
| Order Item Product Price | Item Price of an order |
| Order Item Profit Ratio | Ratio of SP and CP for that product |
| Order Item Quantity | Quantity of an item that’s placed as an order |
| Sales | Sales happened in a particular domain of product |
| Order Item Total | CP of each item \* Qty ordered |
| Order Profit Per Order | Profit received by the company per order  (Same as Benefit per order) |
| Order Region | Region from a specific continent from where the order was placed |
| Order State | Residential state of customer |
| Order Status | Status of supply chain order cycle |
| Order Zip-code | Coded by the company for an order |
| Product Card Id | Depending on department it is being assigned |
| Product Category Id | Based on domain, it is being assigned |
| Product Description | Details of the ordered product |
| Product Image | URL to the image of the ordered product |
| Product Name | Name of the selling product |
| Product Price | Selling price of the product which the customer  has to pay |
| Product Status | Status of Product filled with zeroes only |
| shipping date (Date Orders) | Date when the product was shipped |
| Shipping Mode – Delivery | Transportation modes used by the company |

**Variable categorization (count of numeric and categorical)**

No.of Numerical Columns: 29

No.of Categorical Columns: 24

**Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)**

We have dropped all the columns that were containing null values as those were not important for our targeted analysis.

* Customer Email and Customer Password was having encoded characters.
* Order Zip code and Product Description were blank columns
* Product Image was having hyperlink.
* Order Profit Per Order was same as Benefit per order.
* Customer Street, Customer Names were not needed for analysis as we are having customer Ids
* Latitude and Longitude were also not needed for analysis.

**Project Justification -**

**Project Statement**

A Data Set of Supply Chains used by the company DataCo Global was used for the analysis. Areas of important registered activities: Provisioning, Production, Sales, Commercial Distribution. Here, through this supply chain analysis, we are trying to detect the fraudulent product categories, segments of customers involved in that and risky transaction mode to safe guard the DataCo company from further losses.

**Complexity involved**

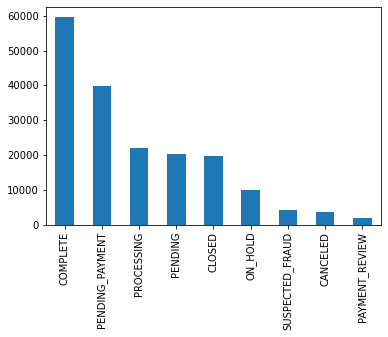
Supply chain complexity does not mean complicated, but rather it describes a condition of inter-connectedness and inter-dependencies across a network where a change in one element can have an effect on other elements. This dataset being a big dataset brings a lot of complexities along with it. Above we have mentioned about the complexities of the columns that we dropped after null value imputation analysis.

**Project Outcome –Commercial, Academic or Social value**

Supply chain analytics can identify known risks and help to predict future risks by spotting patterns and trends throughout the supply chain. Increase accuracy in planning. By analysing customer data, supply chain analytics can help a business better predict fraudulent categories of customers, products or transaction modes.

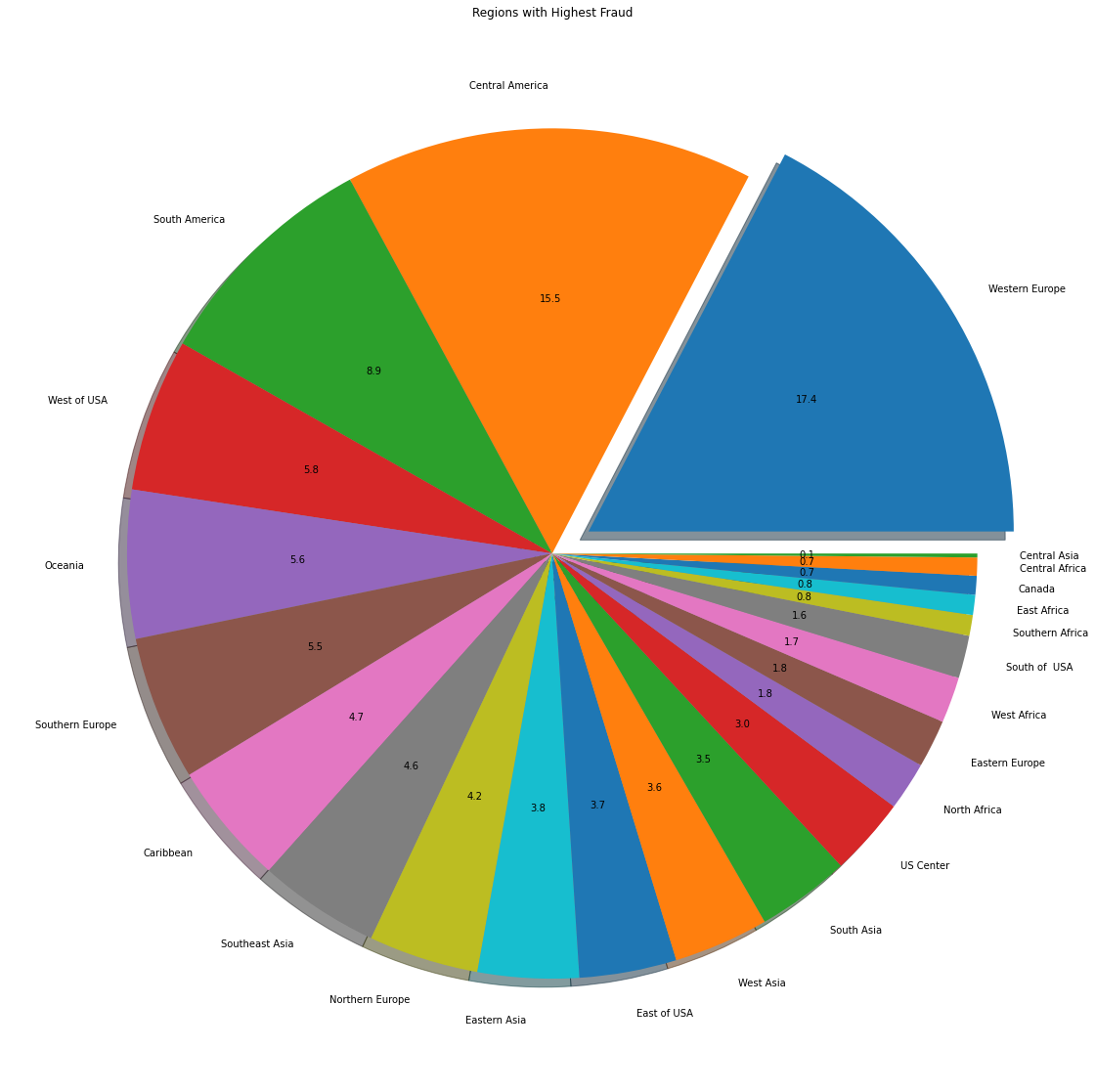
**Data Exploration (EDA)**

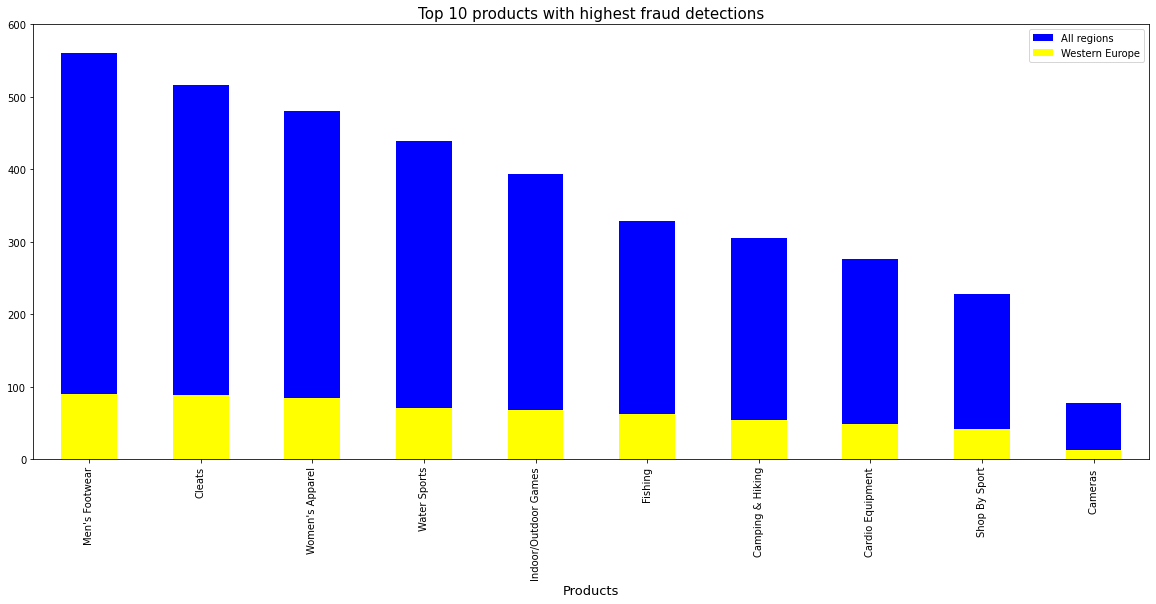
**Relationship between variables**

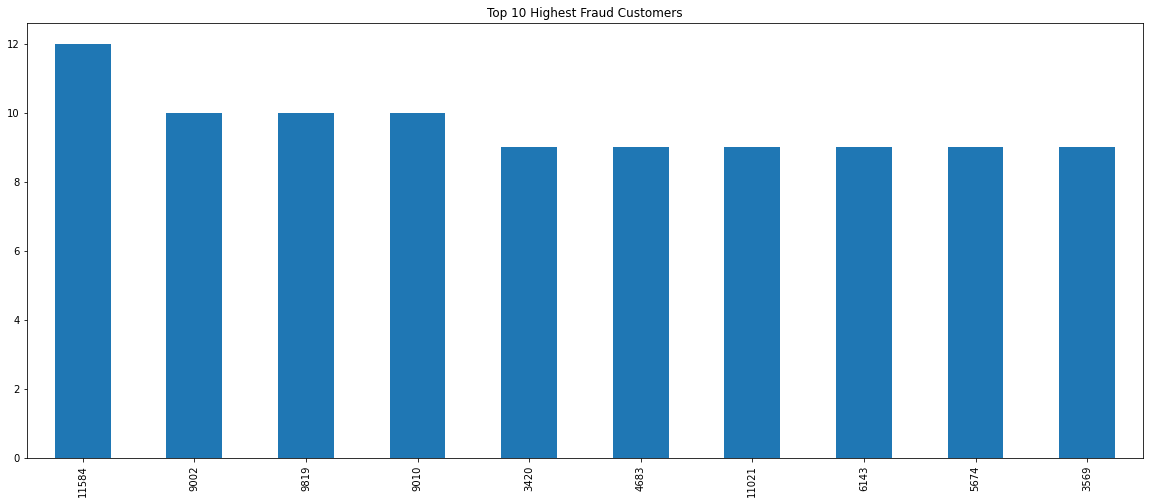


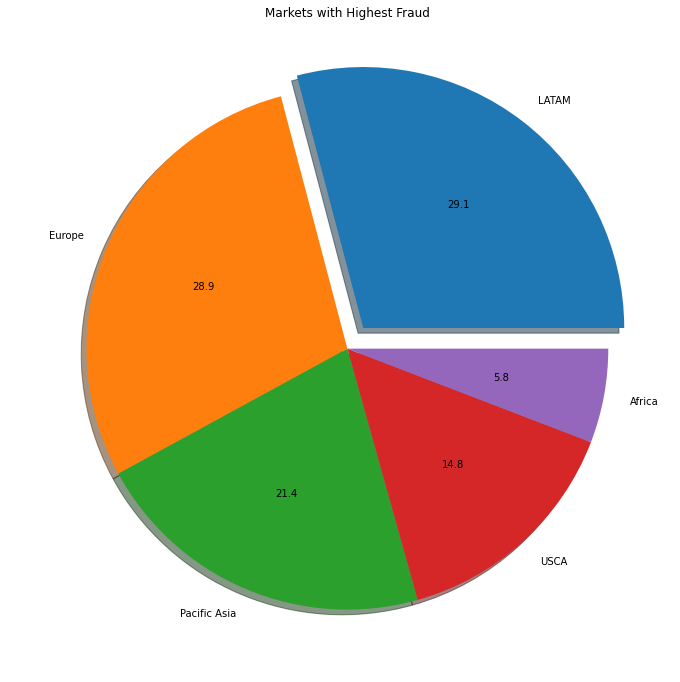
**Univariate Analysis of Target**

* Below are some results of Bivariate analysis with respect to our target column.

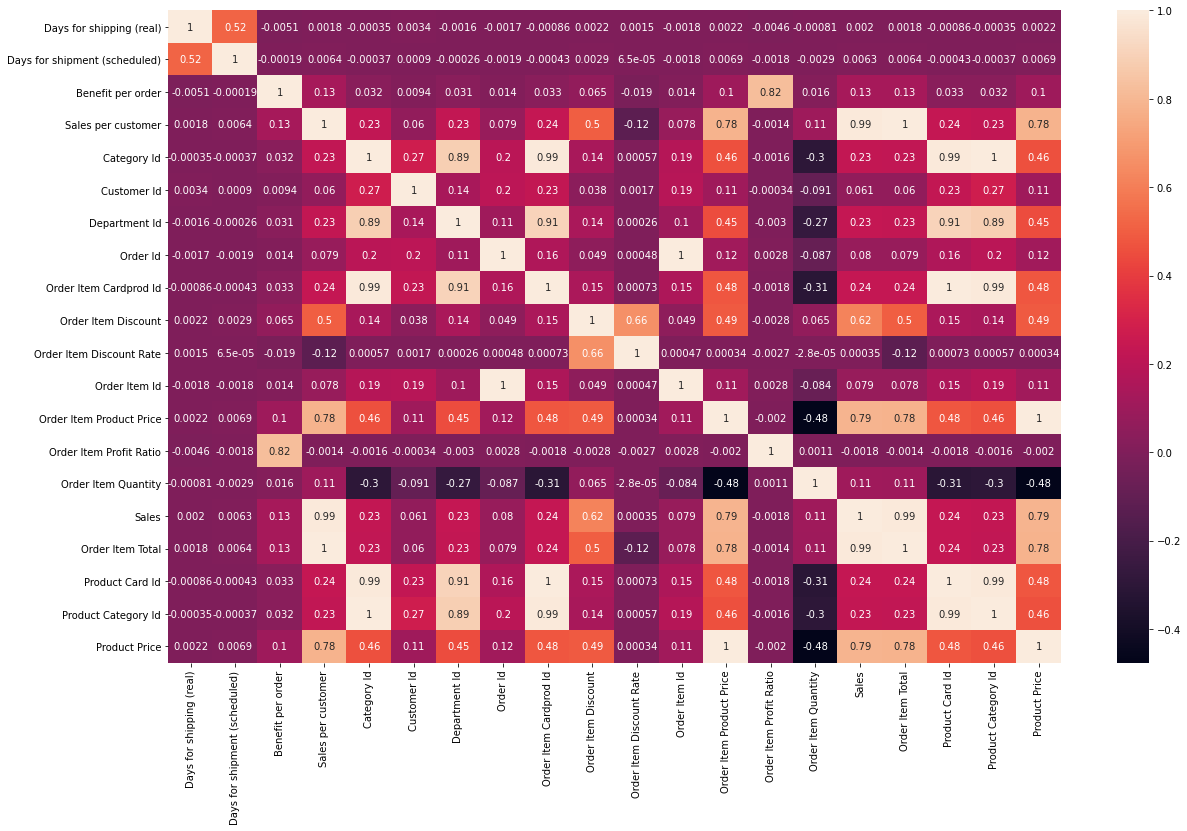








**Multi-collinearity**

****

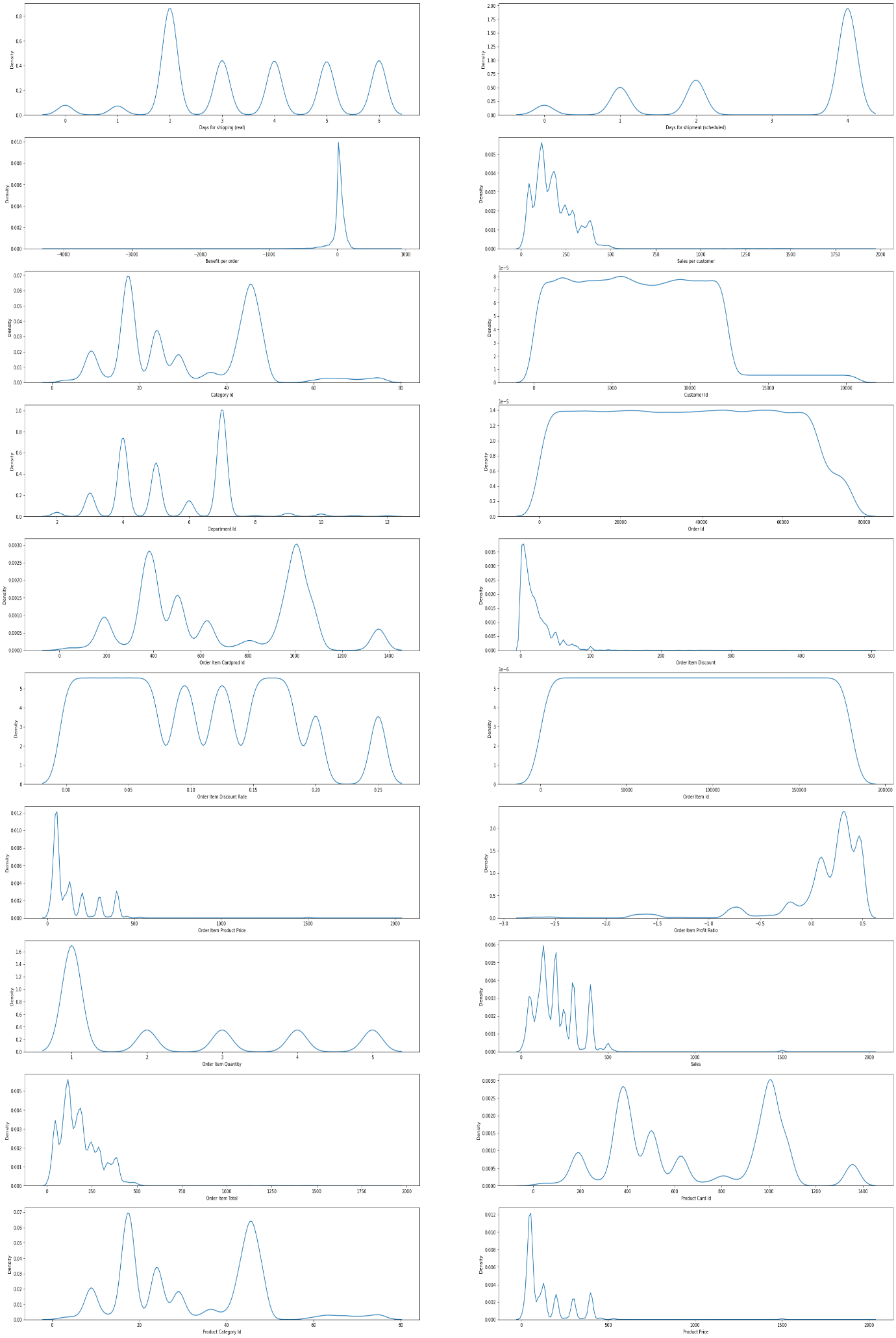
There is a presence of Multicollinearity in the Dataset.

* Product Status was the column with zeroes only
* Late delivery risk was an encoded categorical column stored as numerical column.
* Order Customer Id was dropped as it was same as Customer Id

After further analysis, we were left with 20 numerical & 18 categorical columns.

* This is the final dimension that went for creating base model.
* Label Encoding to be used instead of dummy encoding so as to reduce the model complexity by keeping the training dimension intact.

**Distribution of variables**

****

* This distribution was achieved by plotting Kernel Density Estimation plot for numerical columns.

**Presence of outliers and its treatment**

Benefit per order, Sales per Customer, Customer Id, Order Customer Id, Order Item Discount, Order Item Product Price, Order Item Profit Ratio, Sales, Order Item Total, Product Price are having outliers.

After performing IQR analysis, we incurred information loss. So, didn’t prefer to do treatment. More than 70% data loss was there.

**Statistical significance of variables**

**For Categorical parameters**

p value for column: Type: 0.0

p value for column: Delivery Status: 0.0

p value for column: Category Name: 0.722756766656929

p value for column: Customer City: 3.84005955363135e-157

p value for column: Customer Country: 0.0043149927839307295

p value for column: Customer Segment: 0.0006840787359301104

p value for column: Customer State: 8.582620498747587e-11

p value for column: Department Name: 0.9687137361897139

p value for column: Market: 0.062336109666202545

p value for column: Order City: 0.0

p value for column: Order Country: 2.1962143921929054e-36

p value for column: order date (DateOrders): 0.0

p value for column: Order Region: 5.097832591336198e-09

p value for column: Order State: 0.0

p value for column: Order Status: 0.0

p value for column: Product Name: 0.8136553665199319

p value for column: shipping date (DateOrders): 0.0

p value for column: Shipping Mode: 0.03737290804174167

**For Numerical Parameters**

p value for column: Days for shipping (real): 0.8875301449393683

p value for column: Days for shipment (scheduled): 0.28208032353745677

p value for column: Benefit per order: 0.5306970894045245

p value for column: Sales per customer: 0.7620103443655895

p value for column: Category Id: 0.49237954210421364

p value for column: Customer Id: 6.65326461943043e-05

p value for column: Department Id: 0.5181855014730359

p value for column: Order Id: 0.3406781823347087

p value for column: Order Item Cardprod Id: 0.5619027377252022

p value for column: Order Item Discount: 0.7139450676307028

p value for column: Order Item Discount Rate: 0.30580772812974344

p value for column: Order Item Id: 0.3260184531064966

p value for column: Order Item Product Price: 0.8573626313851704

p value for column: Order Item Profit Ratio: 0.31008474821160636

p value for column: Order Item Quantity: 0.7476769178204016

p value for column: Sales: 0.8302077566837589

p value for column: Order Item Total: 0.7620103443655895

p value for column: Product Card Id: 0.5619027377252022

p value for column: Product Category Id: 0.49237954210421364

p value for column: Product Price: 0.8573626313851704

Customer City, Customer State, Order Country, Order Region, Customer Id

Had to be dropped for further process **as** it can be seen the p-values are exponential which stands out and shows its insignificance.

**Class imbalance and its treatment**

0 176457

1 4062



We will have to use resampling technique – SMOTE while modelling for the same.

**Feature Engineering**

**Whether any transformations required**

Some of the columns were containing -inf and zero as their values. So, log and sqrt transformations fail in such cases. So, Power transformation is being applied and skew ness were totally under control.

**Scaling the data**

After checking the distributions and skewness scaling, we decided not to do scaling. But improvement scope is always there which will be evaluated after seeing model performance.

**Feature selection**

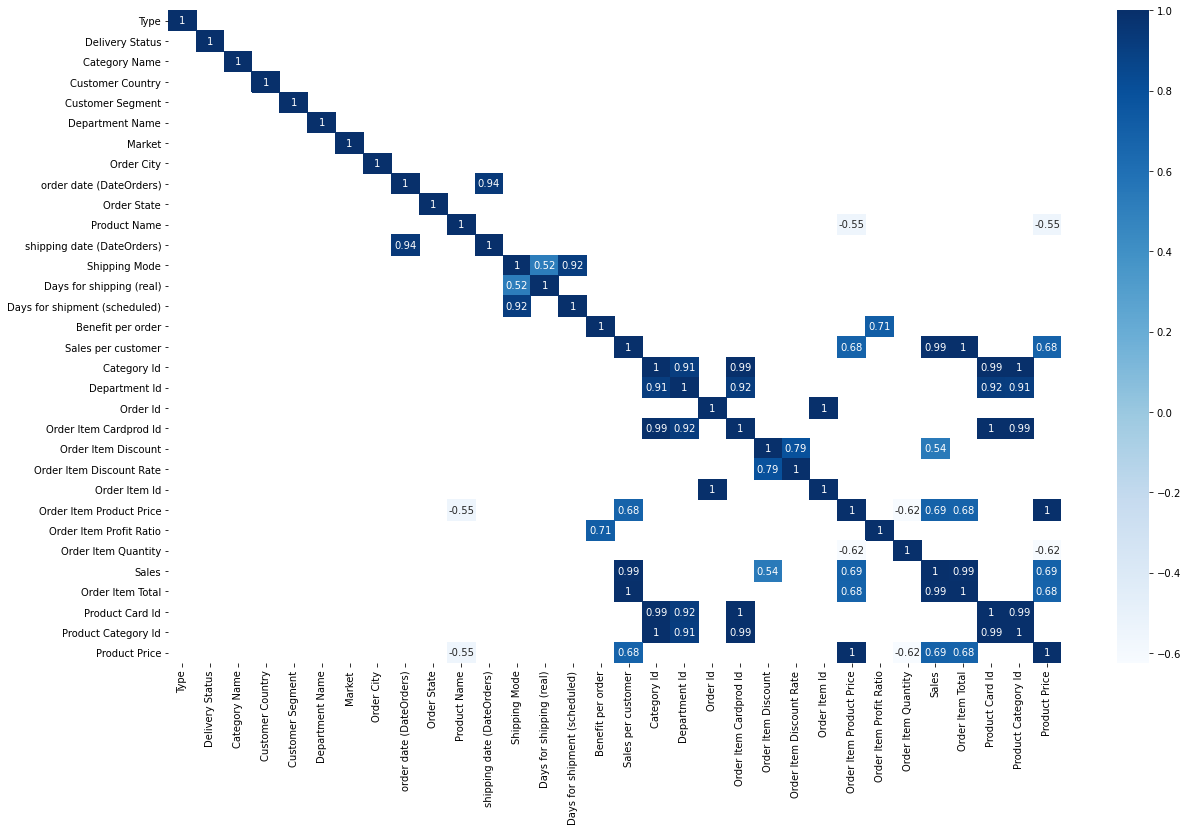
This is done step wise in each and every mentioned process. We have mentioned the feature selections above in which ever stage it was done.

**Dimensionality reduction**

This is the result of feature engineering. As we have mentioned from 53 columns, we concluded at 33 columns are doing several analyses as a part of feature engineering before proceeding for base model.

**Assumptions for Classification Base Model**

* Target variable should be numeric – Passed.
* There should be less multi – collinearity of the features ---- Passed



* Ascertain through statistical test that train & test data

represent the overall data

Ttest\_indResult (statistic=0.2980565264343673,

pvalue=0.7656603642388153)

**Base Model Performance**

confusion matrix - train:

[[123511 0]

[ 2852 0]]

Accuracy of the train data: 0.9774301021659821

Area under the curve AUC on the train: 0.9461652960415097

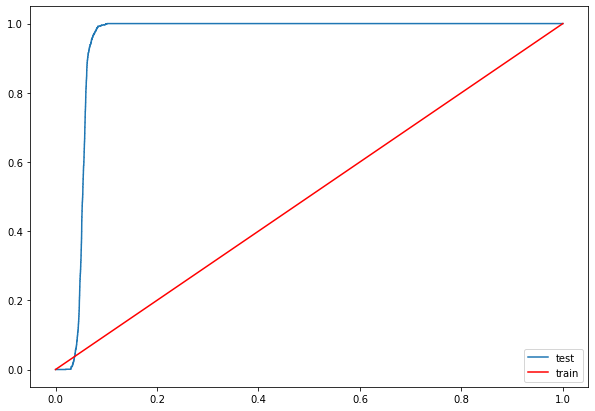
confusion matrix - test:

[[52946 0]

[ 1210 0]]

Accuracy of the test data: 0.9776571386365315

Area under the curve AUC on the test: 0.946080647271054

****

In this stage of Base model, our model seems to be overfitting. Improvements going further with SMOTE analysis and application of advanced classification algorithms.

**Customer Segmentation**

Understanding customer needs and targeting specific clusters of customers based on their need is one way for a supply chain company to increase number of customers and also to gain more profits. RFM analysis is being used because it utilizes numerical values to show Customer recency, frequency and monetary values.

R\_Value (Recency) indicates how much time elapsed since a customer last order.

F\_Value (Frequency) indicates how many times a customer ordered.

M\_Value (Monetary value) tells us how much a customer has spent purchasing items.

We have started working on it. Soon in the next stage we will come up with our outcome.

**---------------------------- Interim Presentation Checkpoint----------------------------------------------------------**