**SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Aneete Rose K.S**

**Megha S.S**

**Kavitha K**

**Nitheesh R.S**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

**List of Figures**

* Correlation heatmap
* Total sales for all markets
* Total sales for all regions
* Average sales
* Average price
* Product Price vs Sales per customer
* Type of payment method
* Products with most loss
* Markets with most loss
* Regions with Highest Fraud
* Fraud product in Western Europe
* Top 10 Highest Fraud Customers
* Top 10 products with most late deliveries
* Days for shipping (real) VS Days for shipment (scheduled)
* Boxplot

**List of Abbreviations**

corrmatrix - correlation matrix

cus - customer

df - data frame

fraud - fraud orders

late - late delivery

sf - suspected fraud

numcols - numerical columns

new\_df - new data frame

le - label encoder

xf - x value for fraud detection

yf - y value for fraud detection

xl - x value for late delivery

yl - y value for late delivery

std - standard scaler

xs - x value for sales

ys - y value for sales

xq - x value for quantity

yq - y value for quantity

logit model - logistic regression

rf - random forest classifier

dt\_model - decision tree classifier

lr - linear regression

ls - lasso regression

dtr - decision tree regressor

**Table of Contents**

**Abstract**

In recent years, the technologies have been unbelievable and more companies are moving towards the adoption of these technologies 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. Here we are working on a project named ‘SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS’. Since the dataset used is related to supply chain, important parameters are identified and the machine learning models are trained with the dataset for detection of fraud transactions, late delivery of orders, sales revenue and quantity of products which customer orders. This project aims to compare different machine learning models and measure their performance to find out which machine learning model performs better.

**1. Problem Definition**

**1.1 Overview**

A Dataset of Supply Chains used by the company Data Co Global was used for the analysis. Dataset of Supply Chain, which allows the use of Machine Learning Algorithms and Python software.

**1.2 Problem Statement**

# Sales prediction and Fraud detection on Supply Chain Dataset

**2. Introduction**

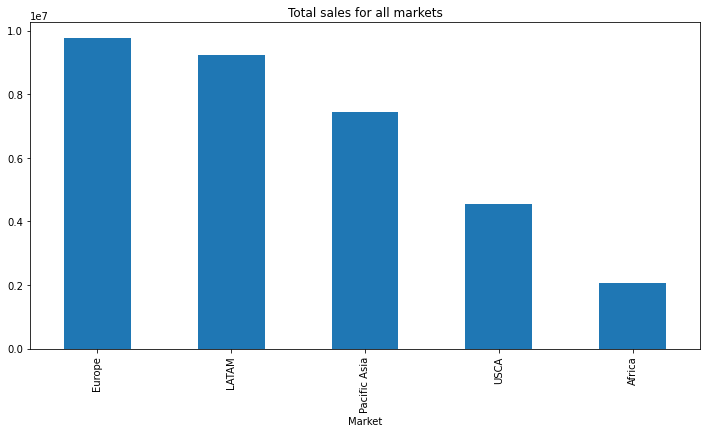
This is about the analysis made on the dataset obtained from, DataCo supply chain.

The variables in this dataset can be related to the supply chain around an e-commerce company. In specific, the variables contain Customer, Production, Sales, Stores, and Logistics information.

Analysis is carried out by using the data visualization techniques to uncover patterns and generate insights.

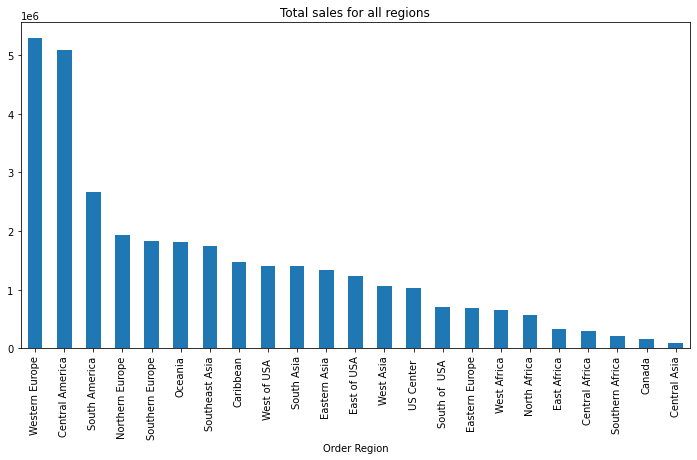
**7. Result**

Fig 7.1



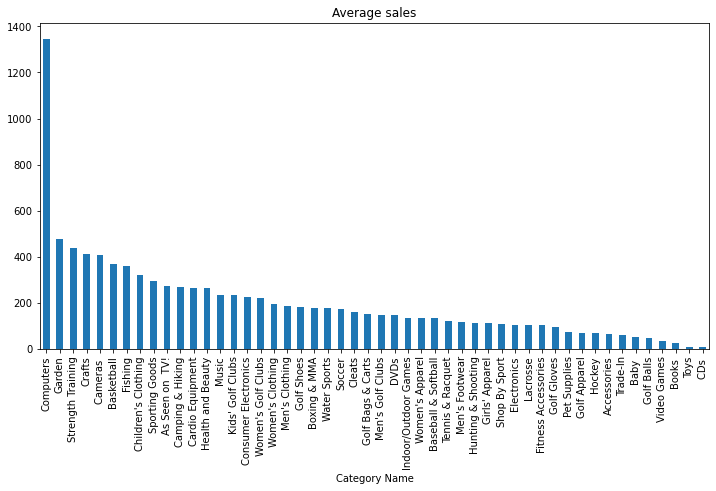
By the analysis of sales in each market and region using bar chart, we understood that European market and western European region have the most sales and Africa and central Asia has least sales.

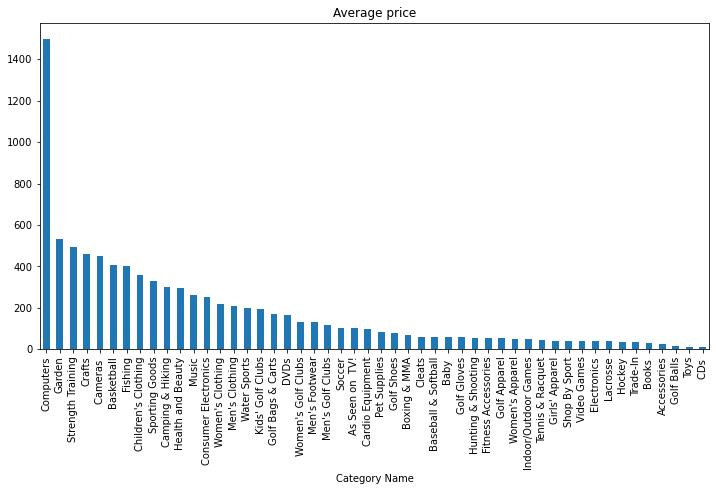
Fig 7.2



In region wise, Western Europe came first and Central Asia came last in sales.

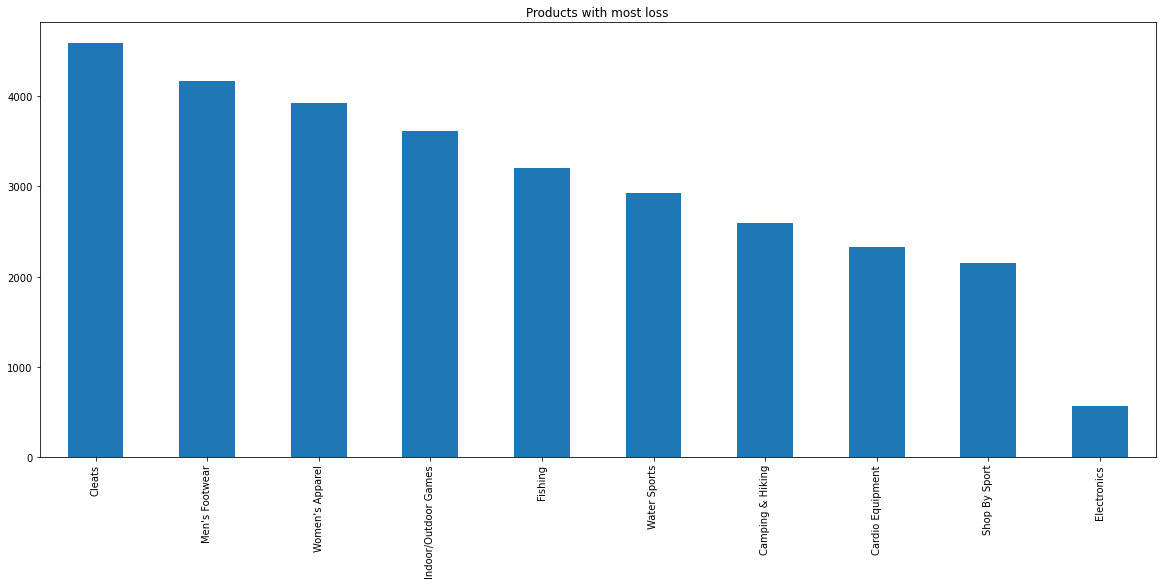
Fig 7.3

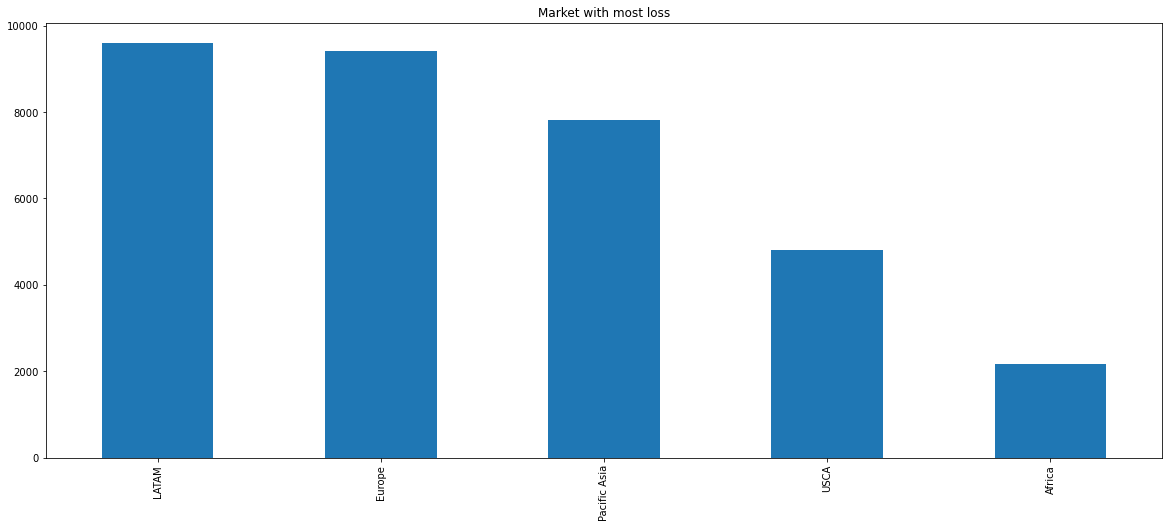
****

****

Some of the top products with highest price on average are the most sold products on average.

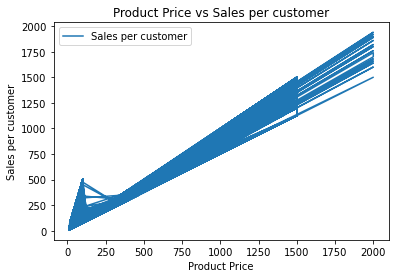
Fig 7.4

****

****

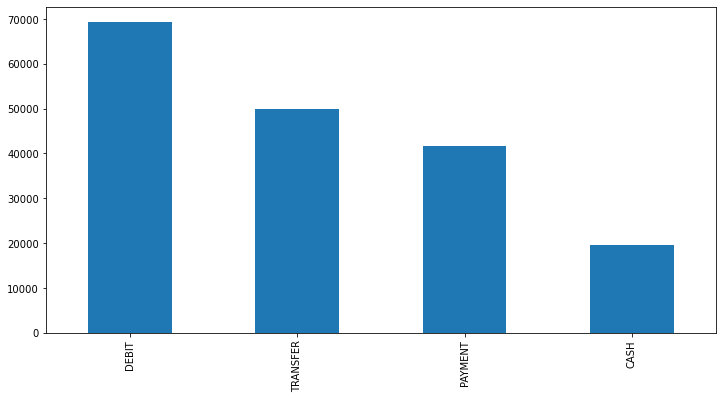
Product with most loss is cleats and Market with most loss is Latin America

Fig 7.5

****

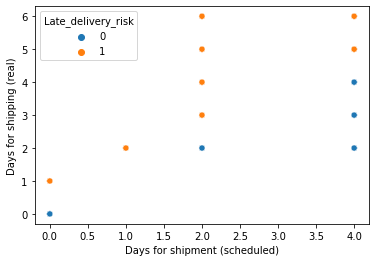
sales and product price have a linear relationship with each other.

Fig 7.6

****

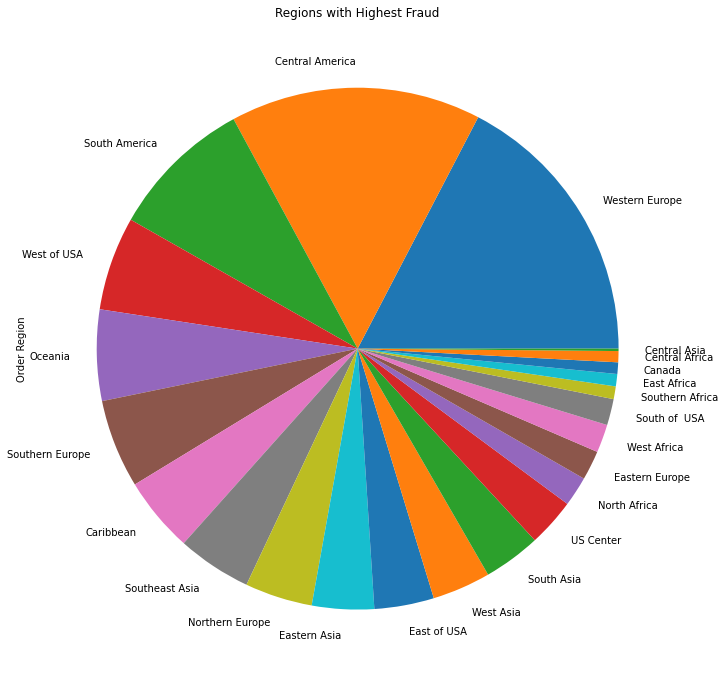
Debit type is the most preferred payment method and cash payment is the least preferred method.

Fig 7.7

****

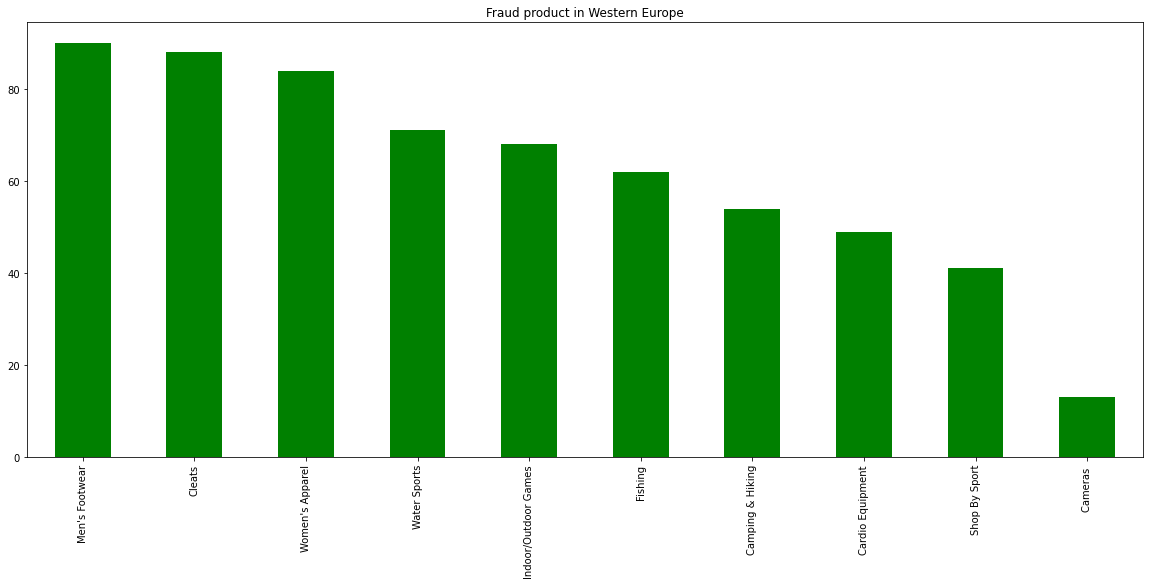
From this we can understand the risk of late delivery related to days for scheduled shipment and real.

Fig 7.8

****

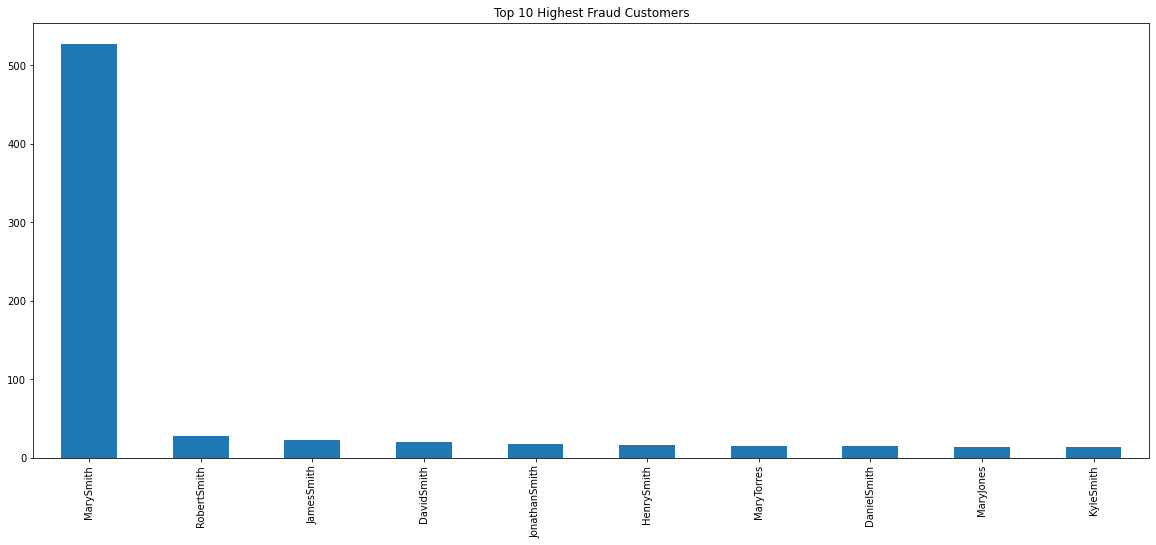
Highest number of suspected fraud orders are from Western Europe and Central America.

Fig 7.9



Men's Footwear and Cleats products are the most suspected fraud orders in Western Europe.

Fig 7.10



The customer named Mary Smith was responsible for most of the fraud transactions.

**Result**

Market with the most number of sales - European

Market with the least number of sales -Africa

Region with the greatest number of sales - Western Europe

Region with the greatest number of sales - Central Asia

Most preferred payment method - Debit

Least preferred payment method - Cash

Product with most loss sales - Cleats

Region with highest no of fraud orders- Western Europe

Columns with missing values - Customer zip code, Order zip code, Product description

Columns with outliers - Benefit per order, Sales per

customer, Order item

discount, Order item product

price, Order item profit ratio,

Sales, Order item total,

Order profit per order,

Product price

Best fit model for Classification - Decision Tree Classifier

Best fit model for Regression - Decision Tree Regressor

**Conclusion**

After analyzing the dataset, it has been discovered that both Western Europe and Central America are the regions with the highest number of sales. And both these regions are suspected to the highest number of fraud transactions and orders with more late deliveries. Most people preferred debit payment method. All the orders with the risk of late delivery are delivered late every time. Decision Tree model did a good job of identifying orders with later delivery and detecting fraudulent transactions. For regression type data Decision Tree Regressor is the best fit model to predict the sales and quantity.