**MRA Milestone 1**

**Project**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Table of Contents:

Problem Statement……………………...................................................................................................................................3

Introduction ...................................................................................................................................................................3

Data Description ............................................................................................................................................................3 About Data (Summary). ………………………………………………………………………..………………………………………………………….….....4

Data Description. …………………………………………………………………………………………………………………………………………….........4

Description Categorical variables …………………………………………………………………………………………………………………………….5

Exploratory Data Analysis ………………………………………………………………......…………………………………………………………........6

Checking for outliers. …………………………………………….....................................................................................................7

Sales Trend. ..................................................................................................................................................................8

Product line and deal size analysis. ..............................................................................................................................8

Country based Analysis. ...............................................................................................................................................9

Parameters used for RFM analysis ……………………………………………………........................................................................10

KNIME workflow. ……………...........................................................................................................................................11

Customer Segmentation ………………………………………………………………………………………………………………………………………..12

Top 5 customers …………………………………………………………………………………………………………………………………………………….13

Customer about to churn ……………………………………………………………………………………………………………………………………….13

Lost customer ………………………………………………………………………………………………………………………………………………………..13

Loyal customer ………………………………………………………………………………………………………………………………………………………14

Problem Statement

An automobile parts manufacturing company has collected data of transactions for 3 years. They do not have any in-house data science team, thus they have hired you as their consultant. Your job is to use your magical data science skills to provide them with suitable insights about their data and their customers.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and give insights in business. The data consists of 2747 entries of accidents. We are provided with 20 features mentioned below in the Data Description. Our mission is to analyse the data given by the company and provide insights about their business and also present it in a way that is easy to understand. For this we will be using KNIME, Tablue, Excel and jupyter notebook.

Data Description

1. ORDERNUMBER: Order Number

2. QUANTITYORDERED: Quantity ordered

3. PRICEEACH: Price of each item

4. ORDERLINENUMBER: Order line

5. SALES: Sale amount

6. ORDERDATE: Order date

7. DAYS\_SINCE\_LASTORDER: Days\_since\_last\_Order

8. STATUS: Status of order like shipped or not

9. PRODUCTLINE: Product Line – Category

10. MSRP: Manufacturers Suggested Retail Price

11. PRODUCTCODE: Code of Product

12. CUSTOMERNAME: Customer

13. PHONE: Phone of Customer

14. ADDRESSLINE1: Address of Customer

15. CITY: City of Customer

16. POSTALCODE: Postal code of Customer

17. COUNTRY: Country customer

18. CONTACTLASTNAME: Contact person customer

19. CONTACTFIRSTNAME: Contact person customer

20. DEALSIZE: Size of the deal based on Quantity and Item Price

1. About Data (Summary)

RangeIndex: 2747 entries, 0 to 2746

Data columns (total 20 columns):

# Column Non-Null Count Dtype

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

0 ORDERNUMBER 2747 non-null int64

1 QUANTITYORDERED 2747 non-null int64

2 PRICEEACH 2747 non-null float64

3 ORDERLINENUMBER 2747 non-null int64

4 SALES 2747 non-null float64

5 ORDERDATE 2747 non-null datetime64[ns]

6 DAYS\_SINCE\_LASTORDER 2747 non-null int64

7 STATUS 2747 non-null object

8 PRODUCTLINE 2747 non-null object

9 MSRP 2747 non-null int64

10 PRODUCTCODE 2747 non-null object

11 CUSTOMERNAME 2747 non-null object

12 PHONE 2747 non-null object

13 ADDRESSLINE1 2747 non-null object

14 CITY 2747 non-null object

15 POSTALCODE 2747 non-null object

16 COUNTRY 2747 non-null object

17 CONTACTLASTNAME 2747 non-null object

18 CONTACTFIRSTNAME 2747 non-null object

19 DEALSIZE 2747 non-null object

dtypes: datetime64[ns](1), float64(2), int64(5), object(12)

We can see with the table above that there are that there are 12 object type features, 7 Numeric and 1 date time type of feature. Also which column is labelled as which datatype.

Total columns are 20 and total 2747 entries.

Data Description

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ORDERNUMBER** | 2747.0 | 10259.76 | 91.88 | 10100.00 | 10181.00 | 10264.00 | 10334.50 | 10425.00 |
| **QUANTITYORDERED** | 2747.0 | 35.10 | 9.76 | 6.00 | 27.00 | 35.00 | 43.00 | 97.00 |
| **PRICEEACH** | 2747.0 | 101.10 | 42.04 | 26.88 | 68.74 | 95.55 | 127.10 | 252.87 |
| **ORDERLINENUMBER** | 2747.0 | 6.49 | 4.23 | 1.00 | 3.00 | 6.00 | 9.00 | 18.00 |
| **SALES** | 2747.0 | 3553.05 | 1838.95 | 482.13 | 2204.35 | 3184.80 | 4503.09 | 14082.80 |
| **DAYS\_SINCE\_LASTORDER** | 2747.0 | 1757.09 | 819.28 | 42.00 | 1077.00 | 1761.00 | 2436.50 | 3562.00 |
| **MSRP** | 2747.0 | 100.69 | 40.11 | 33.00 | 68.00 | 99.00 | 124.00 | 214.00 |

Inferences:

1. Average sales is about 3553 with average quantity ordered is about 35.
2. MSRP is very close to price of each item. Which may be good as manufacturer may be in touch with the market but it may also be studied more to see if suggested prices could be pushed a little more to increase the profits.
3. Days since last order can vary from 42 to 3562. With average of 1757.

Description categorical variables

STATUS

count 2747

unique 6

top Shipped

freq 2541

PRODUCTLINE

count 2747

unique 7

top Classic Cars

freq 949

COUNTRY

count 2747

unique 19

top USA

freq 928

DEALSIZE

count 2747

unique 3

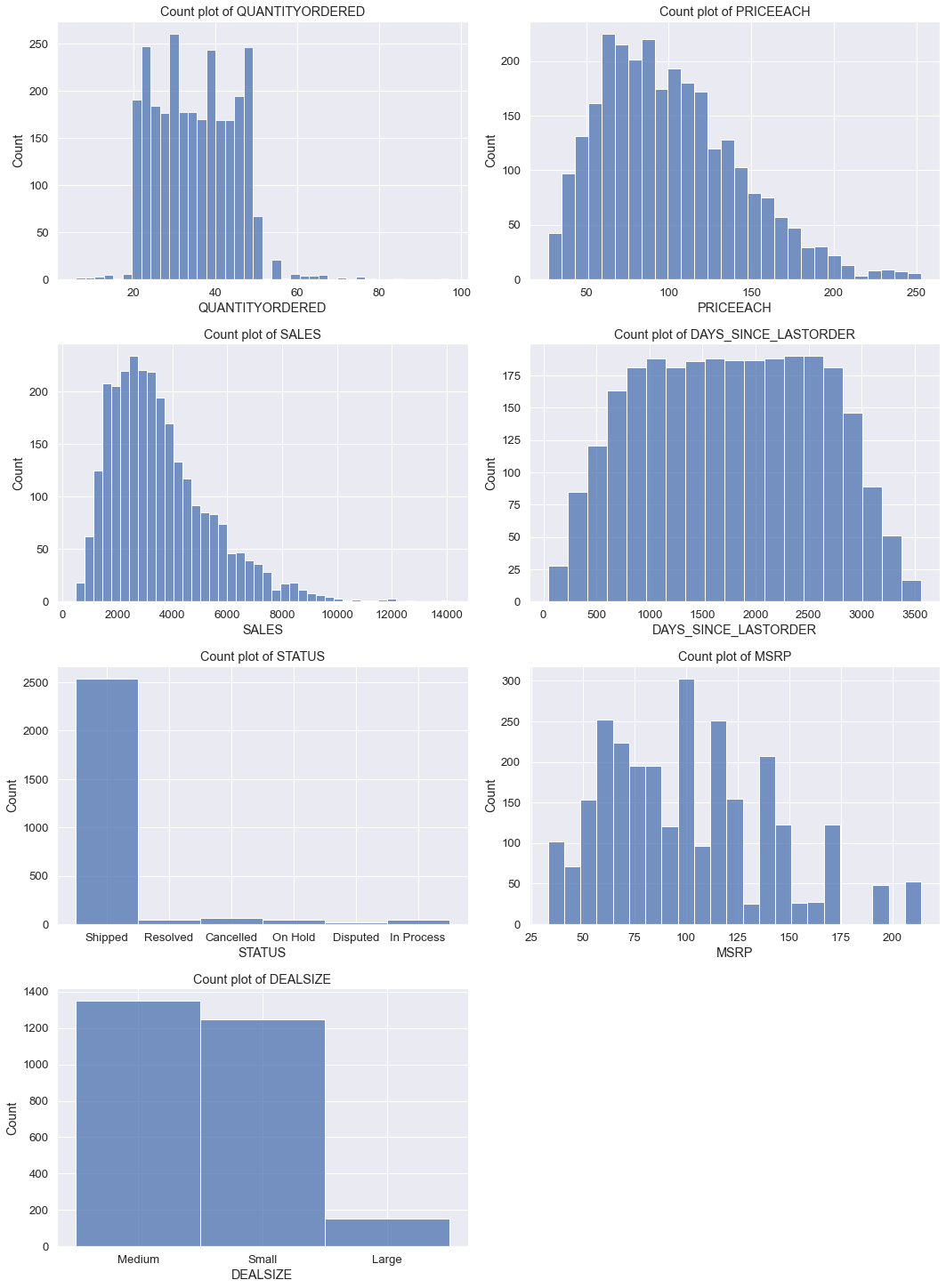
top Medium

freq 1349

Inferences:

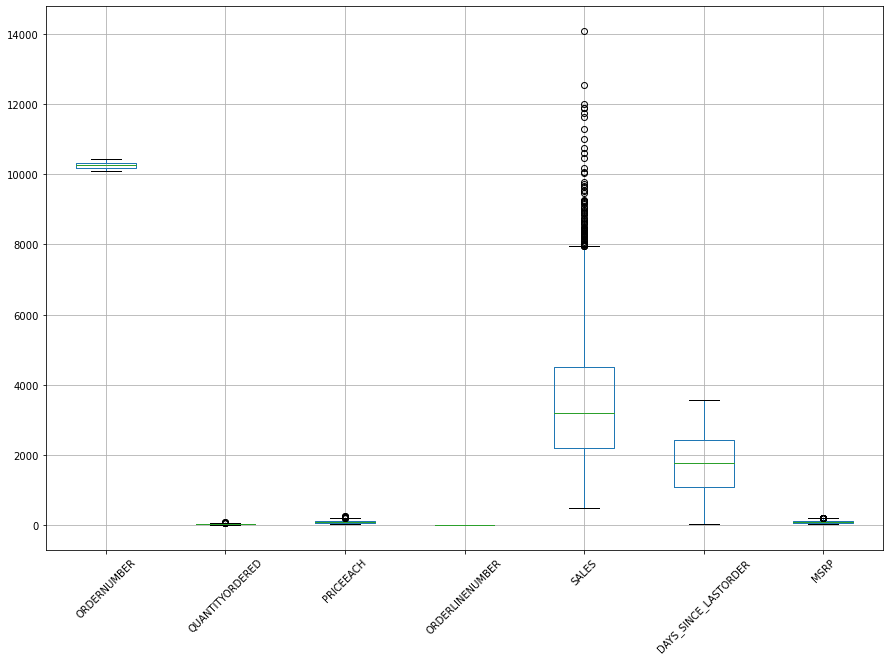
1. There are a total of 19 countries from where the data has been collected and most orders are from USA
2. 3 Deal size Small, Medium and Large. Medium is most ordered automobile.
3. There are 7 product lines that the company offer to its customers.

Exploratory data analysis



* We can see from above charts that sales data is skewed. Very few sales are above 8000.
* Maximum orders are placed for medium size and there are very few orders for large sized automobiles.
* There are very few orders in disputed status which is a good sign.
* Mostly quantity ordered are between 20 to 50

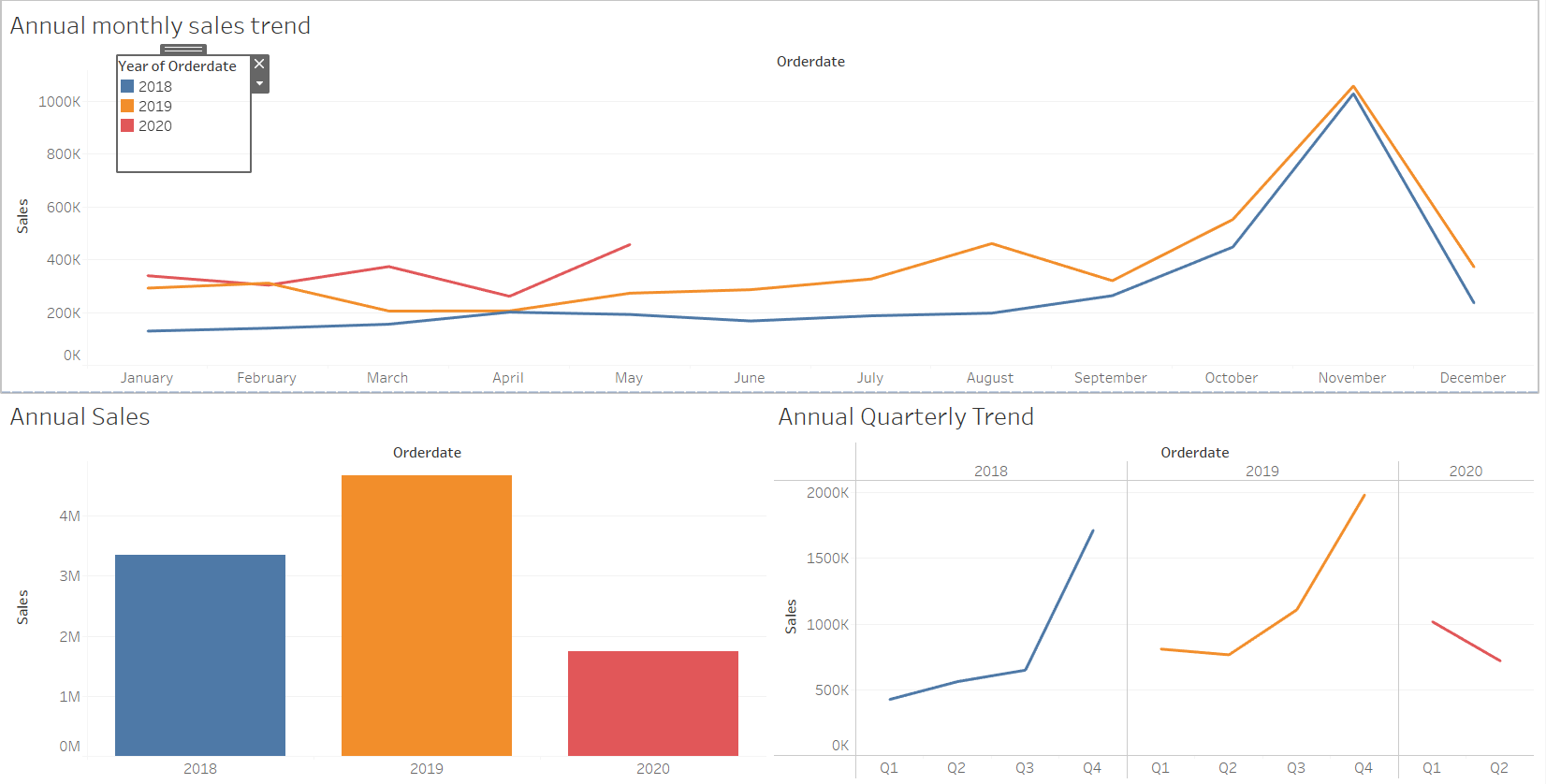
Let us now check for outliers:



Days since last order has no outlier. Sales has many outliers and rest have a couple of outliers.

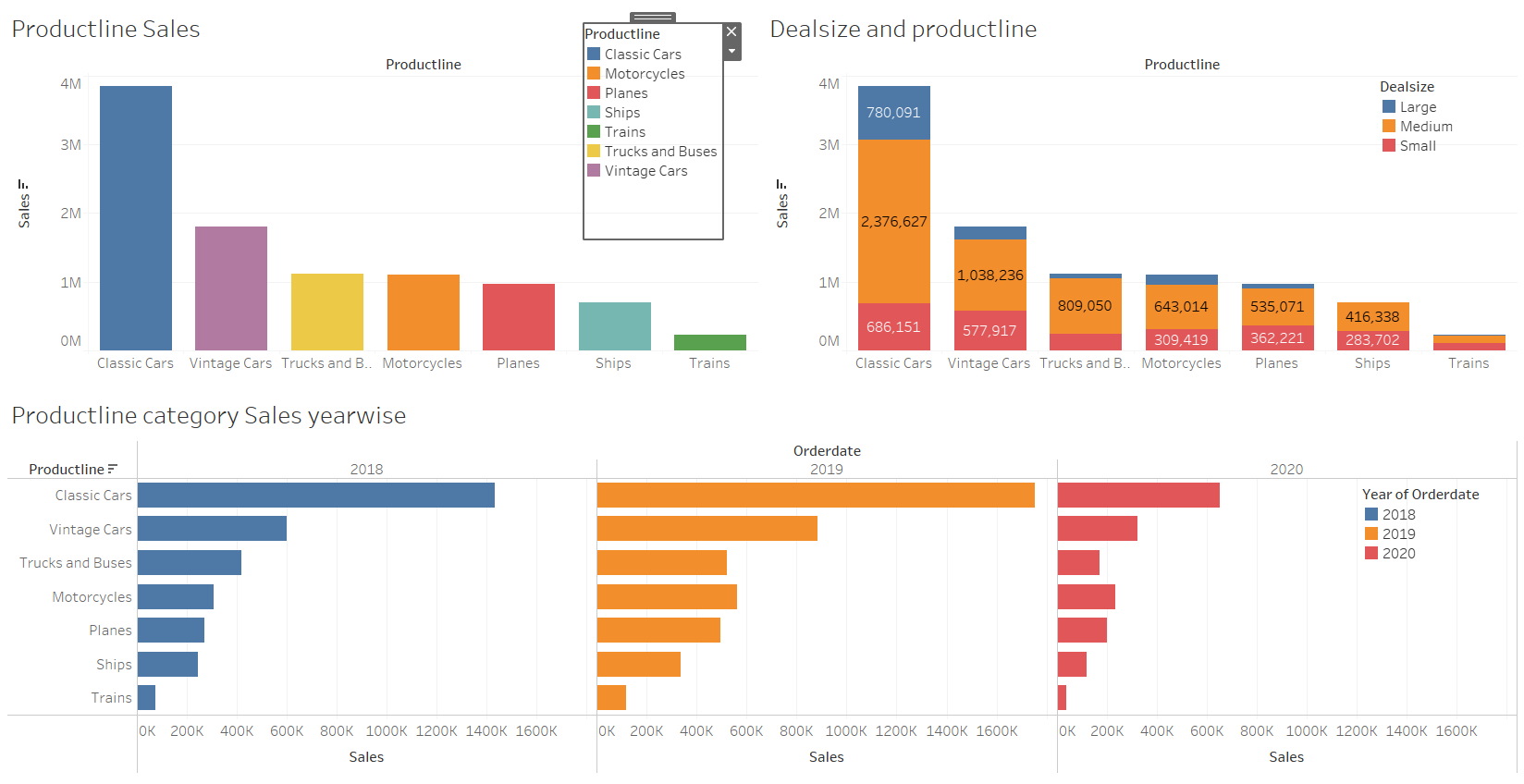
Thus, company has been able to break a few big deals maximum of 14000.

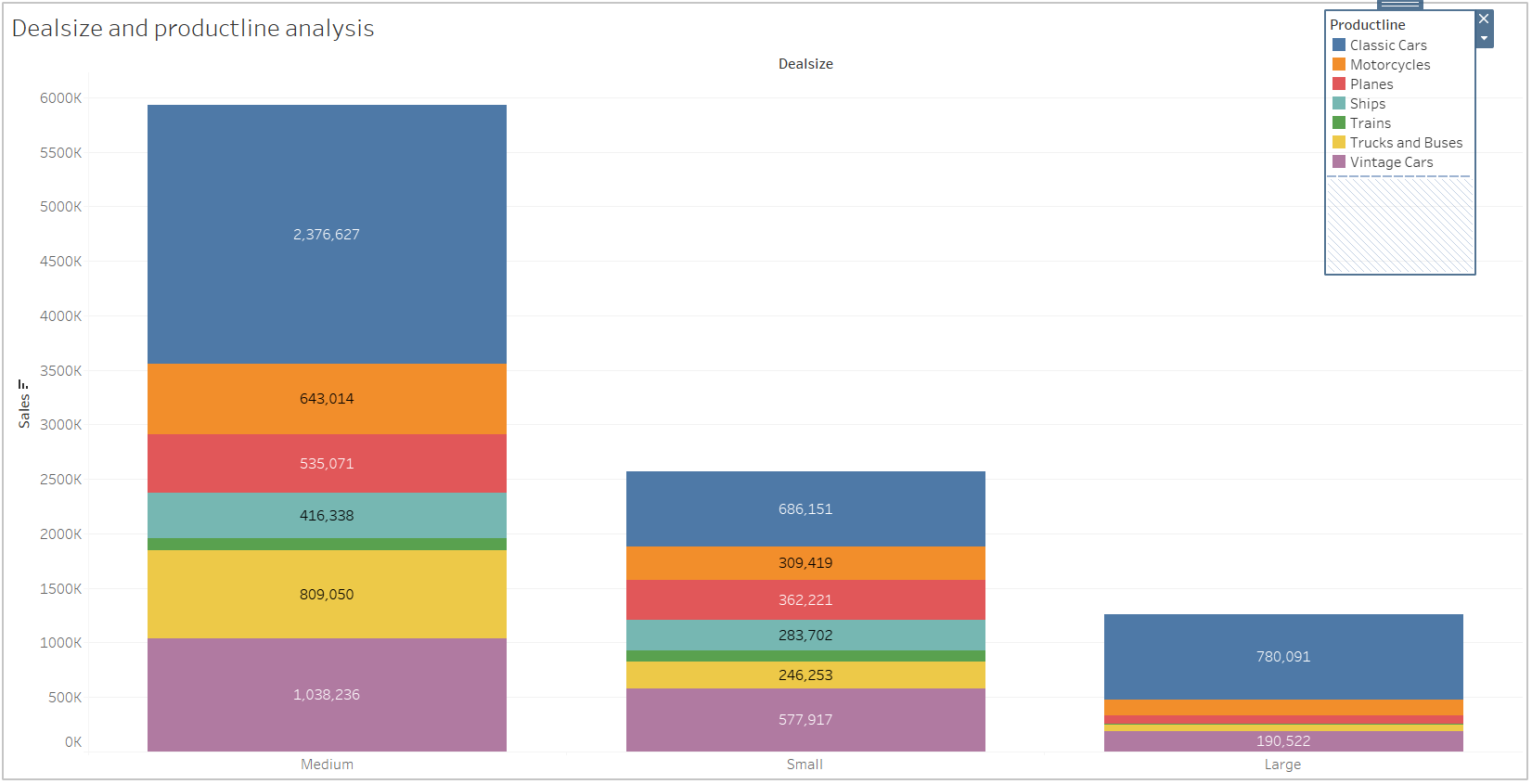
**Sales Trend:**



* We can observe in the above chart that there is an upward trend as sales have been increasing for the second year of 2019. Also, in 2020 sales have been better than the previous years.
* November has the highest sales in the year but dip again in the month of December.
* Because we do not have complete data for Q2 for 2020 the graph has a downward movement but if the we receive more data for Q2 it can be expected that sales for Q2 for 2020 will be better than previous 2 years.

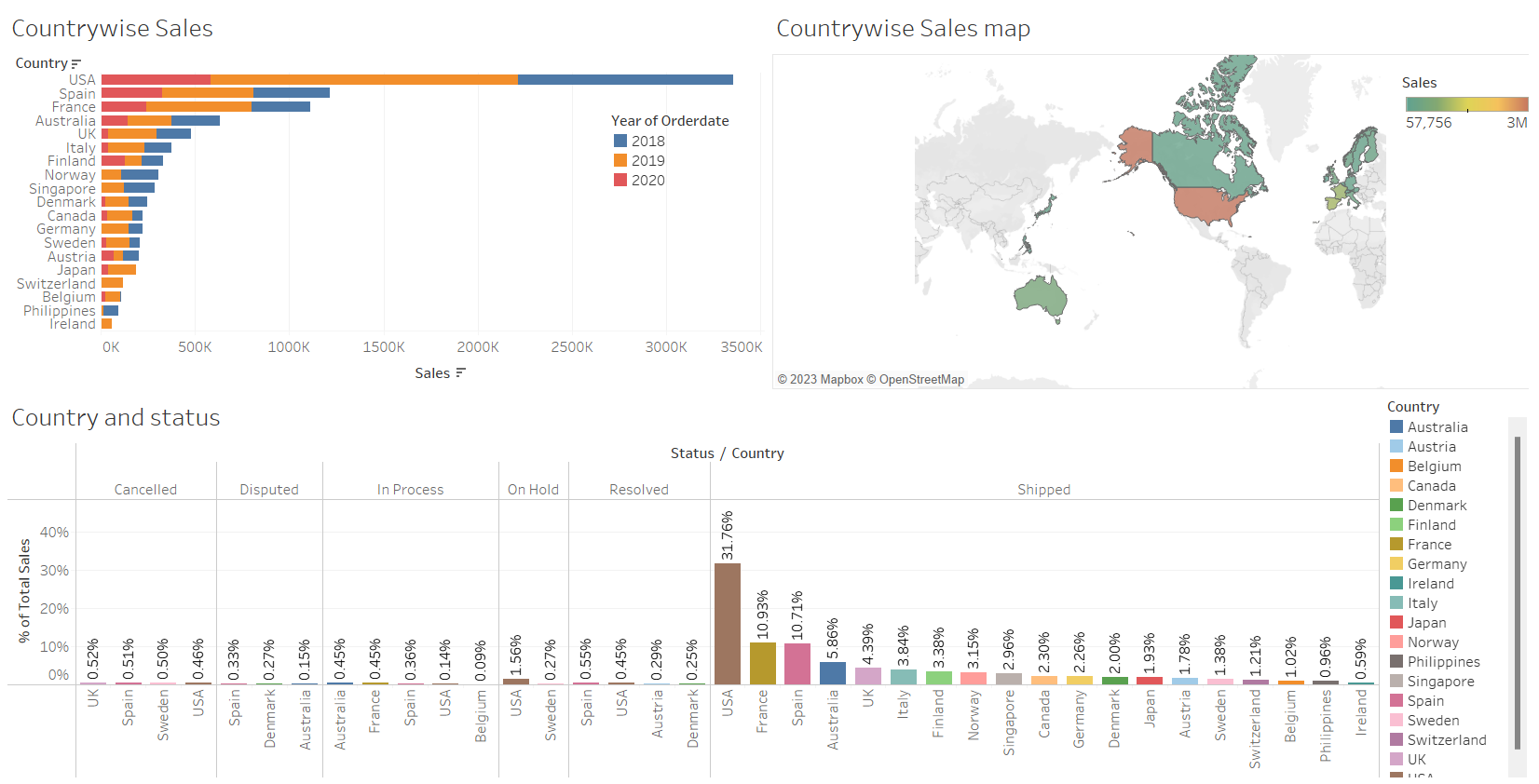
**Product line and deal size analysis:**





* From the above graphs we can say that Classic cars are the best sellers and trains are the least that are sold.
* Sales of every product has increased compared to past year.
* We can also see the sales of different categories in respect to deal size.

**Country based analysis:**



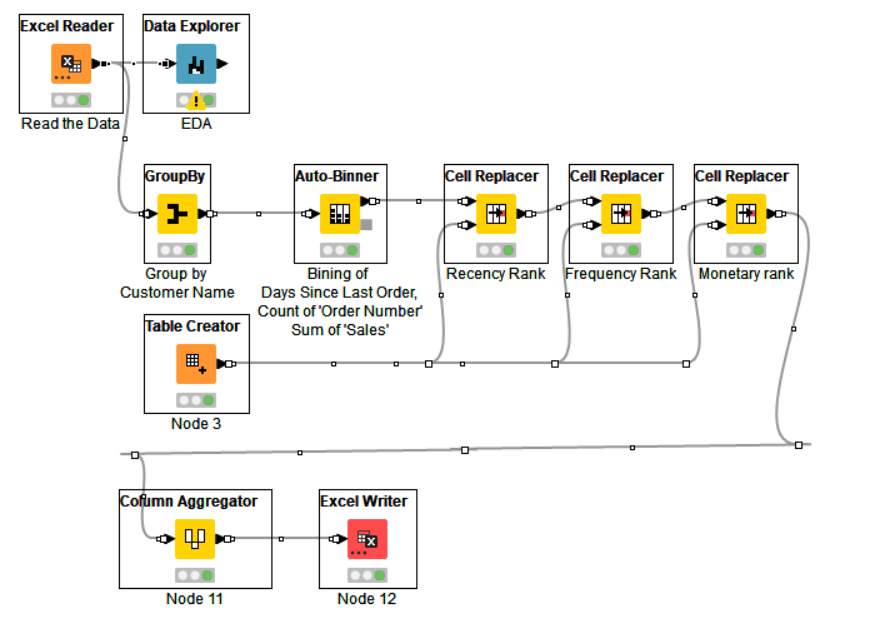
* Referring to above chart we can say that the share of USA in the sales is the largest.
* USA has the maximum sales but has the least cancelled status. In respect to sales, it is very minuscule.

Parameters used for RFM analysis

* Recency - Minimum of Days since last order, after grouping the data by customer name.
* Frequency - Count of order numbers after grouping the data by customer name.
* Monetary - Sum of sales after grouping the data by customer name.
* 4 bins are created considering above features:
  + - Bin 1: 0 - 0.25 Bin 2: 0.25 - 0.5
    - Bin 3: 0.5 – 0.75 Bin 4: 0.75 – 1.0
* Bins are labelled 0 to 3 for for count of order quantity (Frequency) and Sum of Sales (Monetary) from Bin 1 to Bin 4. Any value in the top 25 percentile (Bin4) would get the highest score of 3 and Bin 1 would get the lowest score of 0.
* For Recency metric, Bin 1 would get the highest of 3 and Bin 4 would get 0.
* After obtaining the scores, the same are added to obtain the RFM score.
* Thus, we have 4 segments eof customers as per RFM scores.

|  |  |
| --- | --- |
| Segment Name | RFM Score Criteria |
| Platinum | 8 to 9 |
| Gold | 6 to 7 |
| Silver | 4 to 5 |
| Bronze | 1 to 3 |

KNIME Workflow



**Customer segmentation:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer Segmentation** | **RFM Score** | **RFM\_Label** | **Customer Count** |
| ***Platinum*** | **9** | 333 | 11 |
|  | **9 Total** |  | **11** |
|  | **8** | 233 | 2 |
|  |  | 323 | 1 |
|  | **8 Total** |  | **3** |
| **Platinum Total** |  |  | **14** |
| ***Gold*** | **7** | 133 | 5 |
|  |  | 223 | 1 |
|  |  | 322 | 4 |
|  | **7 Total** |  | **10** |
|  | **6** | 033 | 2 |
|  |  | 132 | 2 |
|  |  | 222 | 5 |
|  |  | 312 | 2 |
|  | **6 Total** |  | **11** |
| **Gold Total** |  |  | **21** |
| ***Silver*** | **5** | 122 | 2 |
|  |  | 212 | 3 |
|  |  | 311 | 3 |
|  | **5 Total** |  | **8** |
|  | **4** | 022 | 1 |
|  |  | 112 | 1 |
|  |  | 211 | 5 |
|  |  | 301 | 1 |
|  | **4 Total** |  | **8** |
|  | **3** | 012 | 2 |
|  |  | 021 | 1 |
|  |  | 111 | 3 |
|  |  | 201 | 2 |
|  |  | 210 | 1 |
|  |  | 300 | 1 |
|  | **3 Total** |  | **10** |
| **Silver Total** |  |  | **26** |
| ***Bronze*** | **2** | 011 | 4 |
|  |  | 101 | 2 |
|  |  | 200 | 3 |
|  | **2 Total** |  | **9** |
|  | **1** | 001 | 1 |
|  |  | 010 | 1 |
|  |  | 100 | 7 |
|  | **1 Total** |  | **9** |
|  | **0** | 000 | 10 |
|  | **0 Total** |  | **10** |
| **Bronze Total** |  |  | **28** |
| **Grand Total** |  |  | **89** |

* Highest RFM rating is of 333and only 14 customers are included in it.
* 28 customers have the lowest rating of 000

**Top 5 customers:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CUSTOMERNAME | COUNTRY | R | F | M | RFM \_ Label | Sum of SALES |
| Euro Shopping Channel | Spain | 3 | 3 | 3 | 333 | $9,12,294 |
| Mini Gifts Distributors Ltd. | USA | 3 | 3 | 3 | 333 | $6,54,858 |
| Australian Collectors, Co. | Australia | 3 | 3 | 3 | 333 | $2,00,995 |
| La Rochelle Gifts | France | 3 | 3 | 3 | 333 | $1,80,125 |
| Land of Toys Inc. | USA | 3 | 3 | 3 | 333 | $1,64,069 |

**Customers about to churn:**

|  |  |  |
| --- | --- | --- |
| **CUSTOMERNAME** | **RFM \_ Label** | **Sum of Sum(SALES)** |
| **Dragon Souveniers, Ltd.** | 033 | $1,72,990 |
| **Rovelli Gifts** | 033 | $1,37,956 |
| **Toys4GrownUps.com** | 022 | $1,04,562 |
| **Cruz & Sons Co.** | 012 | $94,016 |
| **Stylish Desk Decors, Co.** | 012 | $88,805 |

**Lost customers:**

|  |  |  |
| --- | --- | --- |
| **CUSTOMERNAME** | **RFM \_ Label** | **Sum of Sum(SALES)** |
| **Alpha Cognac** | 000 | $70,488 |
| **Clover Collections, Co.** | 000 | $57,756 |
| **Gift Ideas Corp.** | 000 | $57,294 |
| **Iberia Gift Imports, Corp.** | 000 | $54,724 |
| **Mini Auto Werke** | 000 | $52,264 |

**Loyal customers:**

|  |  |  |
| --- | --- | --- |
| **CUSTOMERNAME** | **RFM \_ Label** | **Sum of Sum(SALES)** |
| **Euro Shopping Channel** | 333 | $9,12,294 |
| **Mini Gifts Distributors Ltd.** | 333 | $6,54,858 |
| **Australian Collectors, Co.** | 333 | $2,00,995 |
| **Muscle Machine Inc** | 133 | $1,97,737 |
| **La Rochelle Gifts** | 333 | $1,80,125 |

We can see that top 5 customers and most loyal customers are common.