**MRA Milestone 2**

**Project**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Table of Contents:

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

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

Sample data ……..............................................................................................................................................................3 Data Summary. ………………………………………………………………………..……………………………………………………………………….….....4

Data Description. ……………………………………………………………………………………………………………………………………………..........5

Trend Analysis ………………………………………………………………………………………………………………………………………………………….6

Yearly Trend Analysis ………………………………………………………………………......…………………………………………………………........6

Quarterly Trend. …………………………………………………….....................................................................................................7

Monthly Trend. .............................................................................................................................................................7

Product Count. ……………………………...............................................................................................................................8

Summary of analysis. …...............................................................................................................................................9

Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

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 20642 entries. We are provided with 3 features. 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.

Sample data

|  | **Date** |  | **Order\_id** | **Product** |
| --- | --- | --- | --- | --- |
| **0** | 2018-01-01 |  | 1 | yogurt |
| **1** | 2018-01-01 |  | 1 | pork |
| **2** | 2018-01-01 |  | 1 | sandwich bags |
| **3** | 2018-01-01 |  | 1 | lunch meat |
| **4** | 2018-01-01 |  | 1 | all- purpose |
| **...** | ... |  | ... | ... |
| **20636** | 2020-02-25 |  | 1138 | soda |
| **20637** | 2020-02-25 |  | 1138 | paper towels |
| **20638** | 2020-02-26 |  | 1139 | soda |
| **20639** | 2020-02-26 |  | 1139 | laundry detergent |
| **20640** | 2020-02-26 |  | 1139 | shampoo |

20641 rows × 3 columns

We can observe that there are 20641 entries and only 3 columns.

Data Summary

RangeIndex: 20641 entries, 0 to 20640

Data columns (total 3 columns):

# Column Non-Null Count Dtype

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

0 Date 20641 non-null datetime64[ns]

1 Order\_id 20641 non-null int64

2 Product 20641 non-null object

dtypes: datetime64[ns](1), int64(1), object(1)

memory usage: 483.9+ KB

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

Total columns are 3 and total 20641 entries.

Number of duplicate entries:

Number of Duplicated Rows are 4730

Duplicate entries have been deleted.

**Data description:**

|  | **count** | **unique** | **top** | **freq** | **first** | **last** |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | 15911 | 603 | 2019-02-08 00:00:00 | 138 | 2018-01-01 | 2020-02-26 |
| **Order\_id** | 15911 | 1139 | 311 | 26 | NaT | NaT |
| **Product** | 15911 | 37 | poultry | 480 | NaT | NaT |

We now only have 15911 entries

There is total 37 products.

Most popular product is poultry.

Start date is 01-01-2018 and end date is 26-02-2020.

Dataset statistics

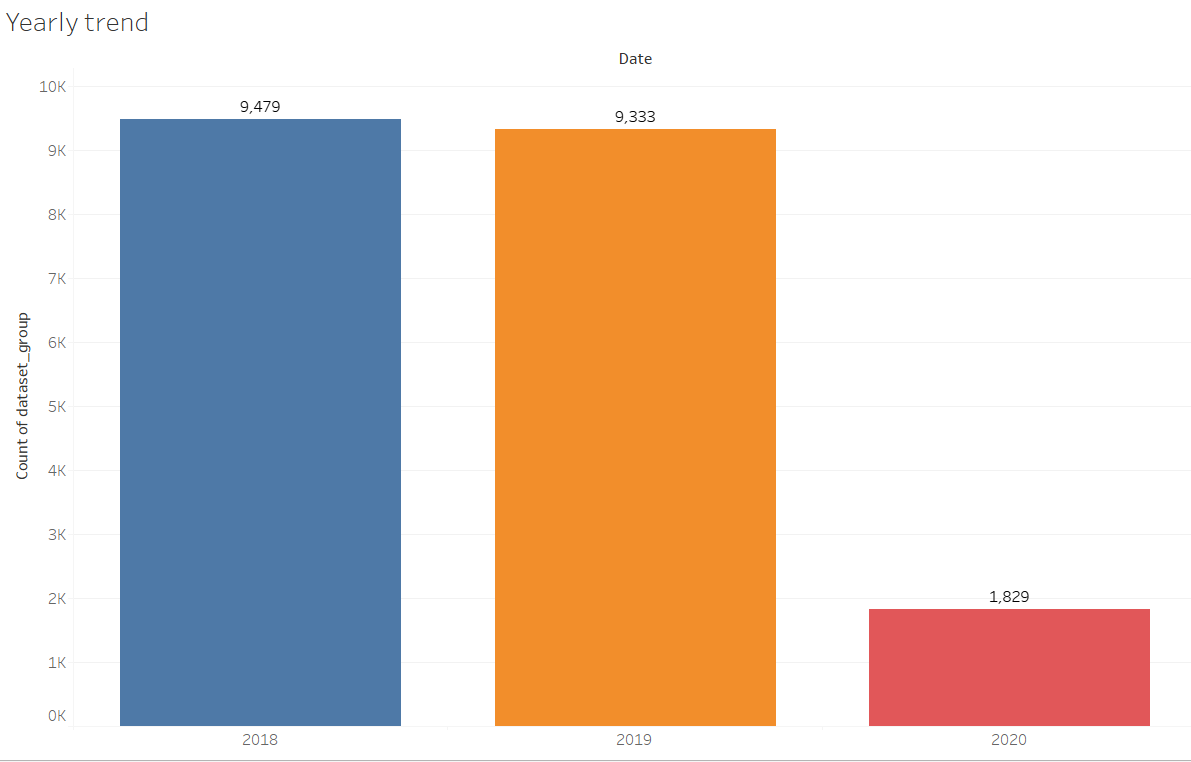
|  |  |
| --- | --- |
| **Number of variables** | 4 |
| **Number of observations** | 15911 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |
| **Total size in memory** | 497.3 KiB |
| **Average record size in memory** | 32.0 B |

Variable types

|  |  |
| --- | --- |
| **Numeric** | 2 |
| **DateTime** | 1 |
| **Categorical** | 1 |
|  |  |

**Let us check trends:**

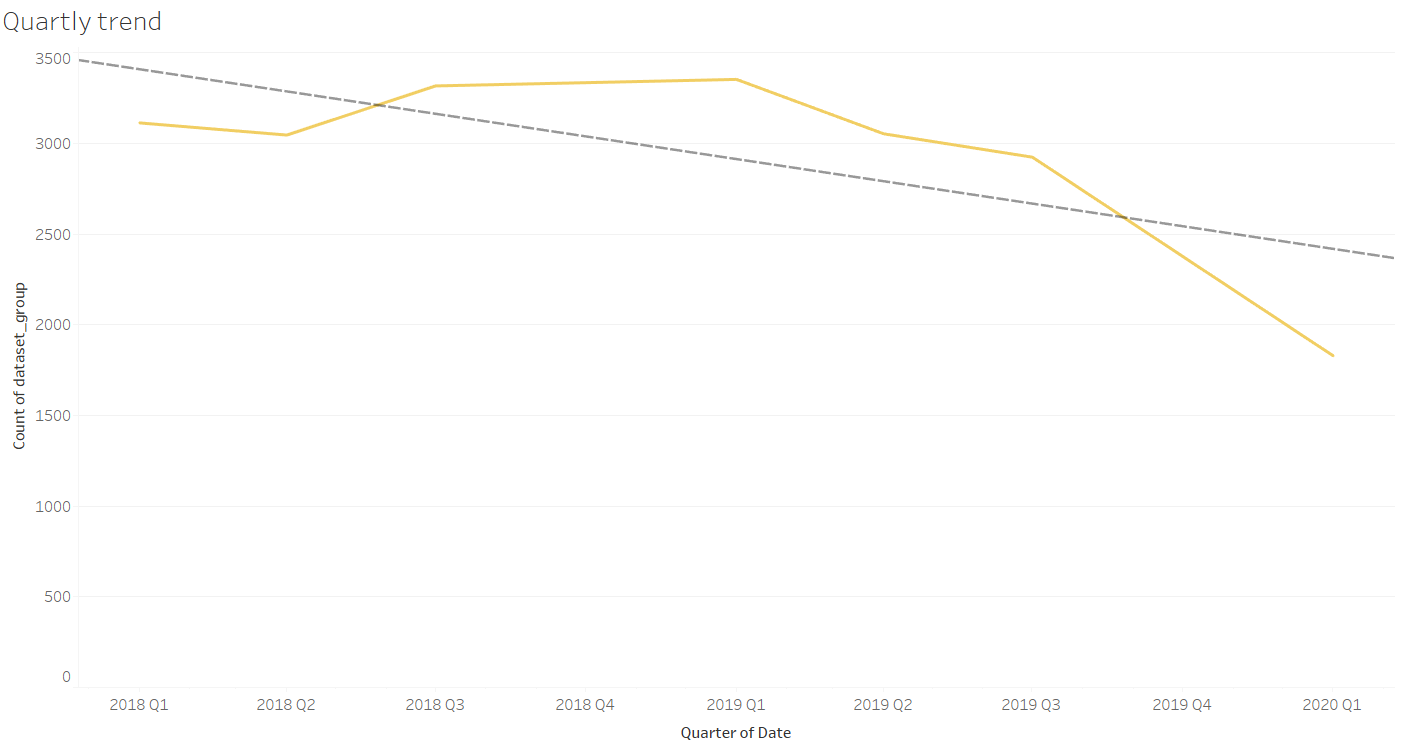
**Yearly trend:**



We can see that there was only a slight decrease in orders in 2019 as compared to 2018.

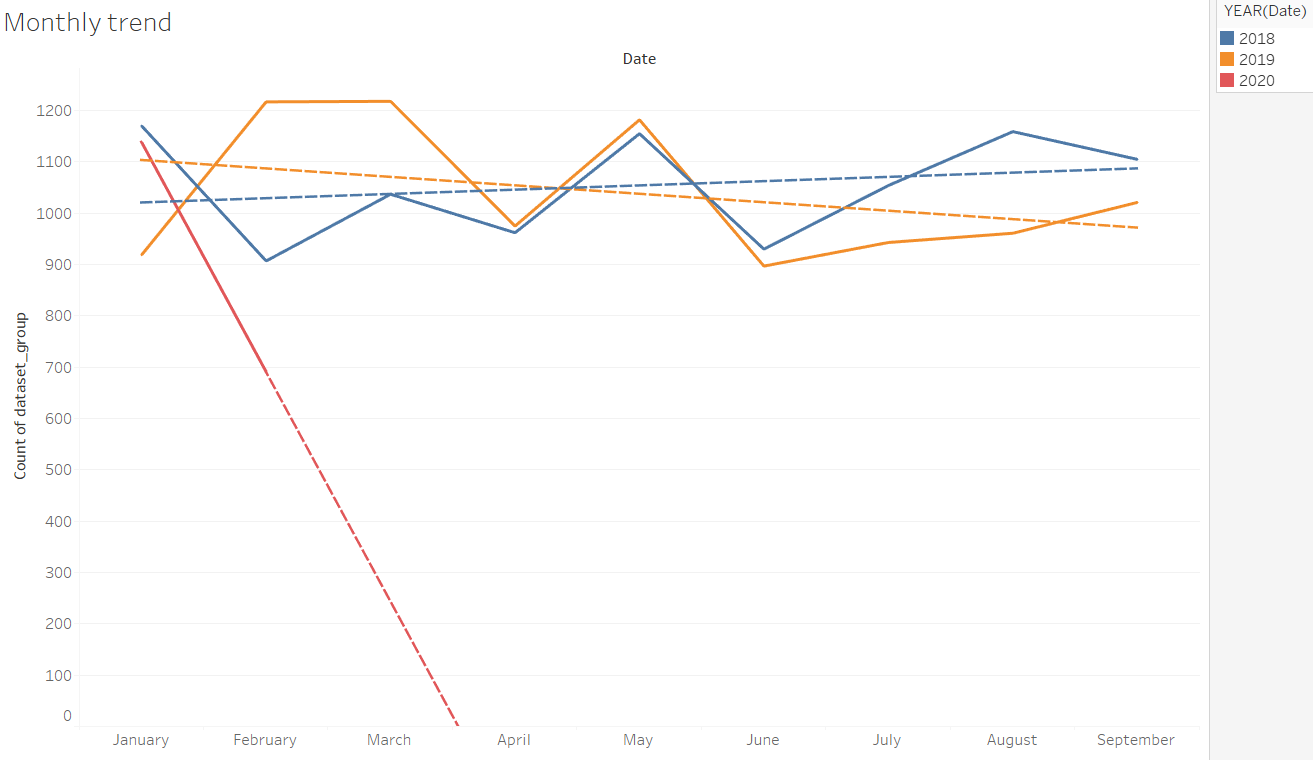
We do not have full data of the year of 2020.

**Quarterly trend:**



We see here that the trend line is showing a decline in the in the upcoming quarter. But we should remember that data for 2020 is only till February and thus the sales for Q1 of 2020 will increase though we see that there is a decline since Q1 2019 and trend may be correct but it may not be as sharp as shown here. We would require more data to study this.

**Monthly trend:**

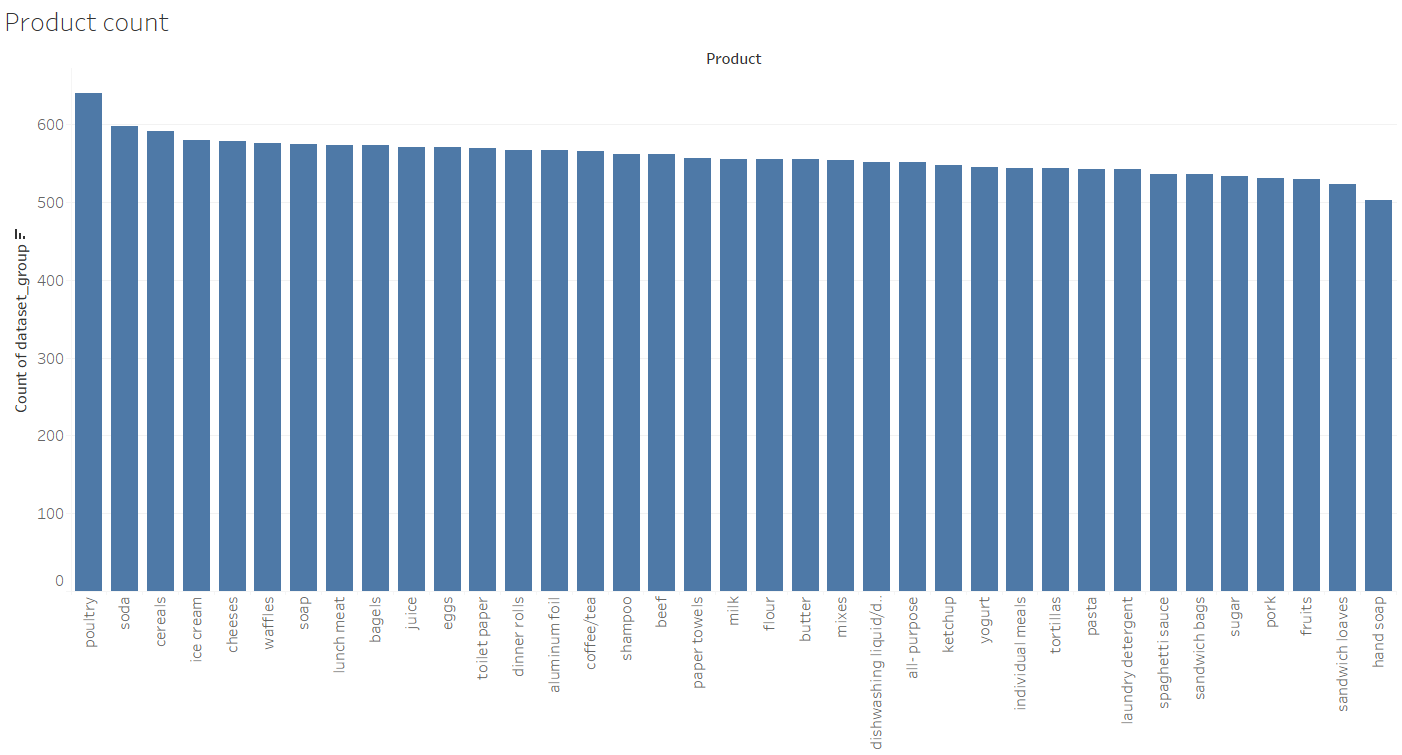


There was an upward trend in 2018 but in 2019 a downward trend has started and it is steeper than the incline in 2018, which is a very worrisome sign for the business.

When we look at the monthly trend, we are able to see that the month of April, May and June have some similarity. On other months the orders are not regular to show any seasonality. For example, month of February in 2018 there was a less sale than February 2019 but in February 2020 it declined even further than was in 2018.

This shows that sales are sporadic and regional factors should be studied probably to understand this better.

**Product Count:**



We can see that poultry orders are the largest followed by soda. Least ordered products are hand soap and sandwich loves.

**Summary of analysis:**

