**TDS 3301 Data Mining**

**Assignment Part 1**

**Exploratory Data Analysis**

|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Email** |
| **Khoo Huai Yeng** | **1142702184** | **yengkhoo3981@gmail.com** |
| **Delvine Lee Sow Hui** | **1141328205** | **delvinelsh@gmail.com** |
| **Chong Kai Yun** | **1142702629** | **starryskycloud@gmail.com** |

**Dataset Description**

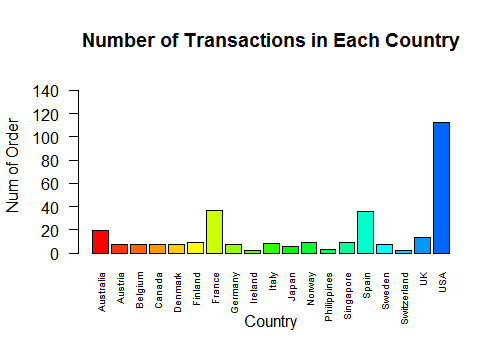
The source of this dataset is from: <https://www.kaggle.com/kyanyoga/sample-sales-data>

This dataset is a sample sales invoice of a toy factory outlet and has 2823 instances and 25 attributes. Each row corresponds to an order or transaction.

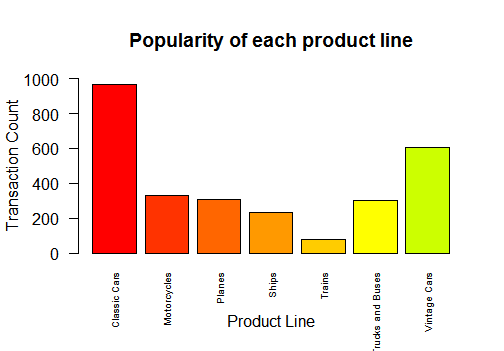
|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| ORDERNUMBER | Order number |
| QUANTITYORDERED | Quantity of product ordered |
| PRICEEACH | Unit price for the product |
| ODERLINENUMBER | Based on *ORDERNUMBER* and *PRODUCTNUMBER*. A customer (*ORDERNUMBER*) can purchase multiple products (*PRODUCTCODE*), first product ordered will be numbered ‘1’ and so on. |
| SALES | Product of QUANTITYORDERED and PRICEEACH |
| ORDERDATE | Date when order was made |
| STATUS | Order status (Shipped, In Process, Resolved, Cancelled, On Hold, Disputed) |
| QTR\_ID | Quarter of the year order was made (integer) |
| MONTH\_ID | Month when order was made (integer) |
| YEAR\_ID | Year when order was made |
| PRODUCTLINE | Type of product ordered |
| MSRP | Manufacturer suggested retail price (Price recommended by manufacturer to retailer to sell the product) |
| PRODUCTCODE | Code of product |
| CUSTOMERNAME | Name of customer (Company name) |
| PHONE | Contact number of customer |
| ADDRESSLINE1 | Customer’s street address |
| ADDRESSLINE2 | Customer’s unit number |
| CITY | Customer’s city |
| STATE | Customer’s state |
| POSTALCODE | Customer’s postal code |
| COUNTRY | Customer’s country |
| TERRITORY | Area of country (NA=North America, APAC=Asia Pacific Region, EMEA = Europe, the Middle East and Africa |
| CONTACTLASTNAME | Contact’s last name |
| CONTACTFIRSTNAME | Contact’s first name |
| DEALSIZE | Based on SALES (Small if *SALES* < 3000, Medium if 3000 < *SALES* < 7000, Large if *SALES* > 7000) |

**Possible Insights from Dataset Sales.csv**

1. Identify the country with the highest number of purchase from this company.

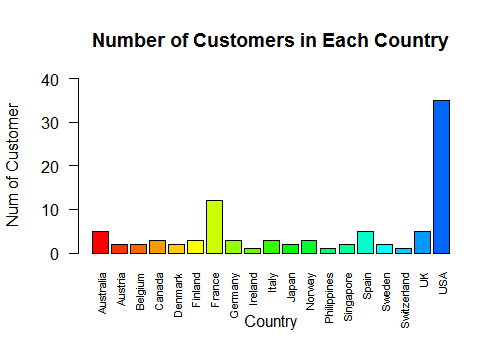


From the bar chart above, it is clear that USA has purchased the most from this company.

1. Determine the product line with the highest probability to be bought in each transaction.

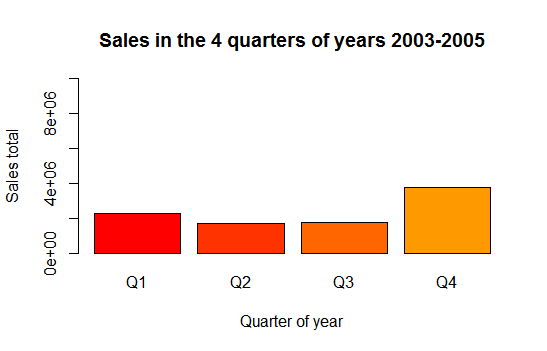
The bar chart above shows that the product line ‘Classic Car’ is the most popular among the customers as it has the highest rate to appear in transactions.

1. Identify the country which the customer that has made purchase with this company resides in.



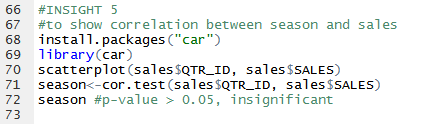
This bar chart shows that USA has the highest number of customers purchasing from this company. By comparing the charts from *insight 1* and *insight 3*, we can see the clear relation of number of customers in a country affects the sales in that country. From this view, the company can make a further decision on whether to focus their sale in USA or increase their advertising in the other countries.

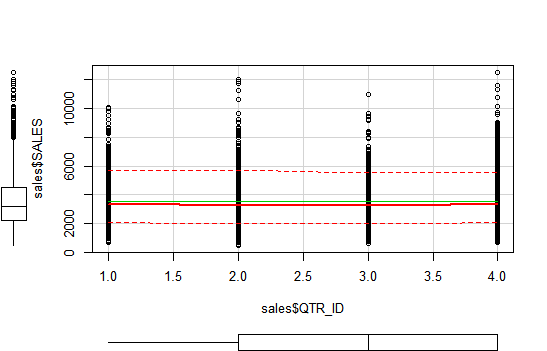
1. Determine which quarter or month of the year has the highest sales rating.

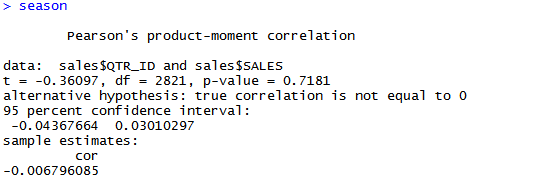


The company has the highest sales in the fourth quarter (October, November, December) of the years.

1. Examine if seasonal holidays have effect on the sales.

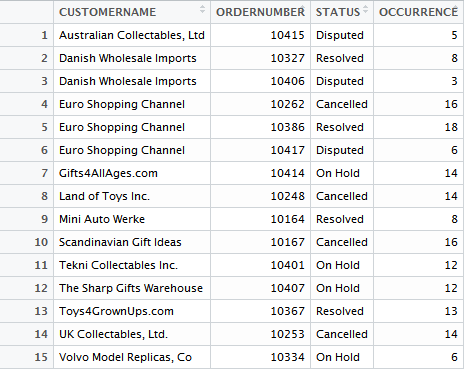






A p-value of 0.7181 is considered statistically insignificant. A negative value of cor, -0.006796085 indicates that it is a weak linear correlation. So, we can assume that seasonal holiday does not affect the sales or has a very little influence to the sales that can be ignored.

1. From the records, identify which customer has the highest probability to call off the purchase.



The above table shows list of customer (name) which purchases were either disputed, resolved, on hold or cancelled. The company should identify these customers and diagnose the cause behind each of these call offs.

**Data Mining Technique**

1. **Association Rule** – Identify the product lines or products that are frequently bought together and predict the sales. This can be done by referring to the *ORDERNUMBER* column. From each distinct *ORDERNUMBER* identify the set of products that are bought. From these sets then we can deduce and pinpoint the pairs or sets of products that are frequently bought together.
2. **Coverage** – Identify the products that are popular among the customers. This product line should be suitable to all the customers. The result can be used to design a product catalog for the company.

**Data Quality Issues**

1. Missing values in columns (*STATE, ADDRESSLINE2*)
2. Data Inconsistency (*PHONE*, *POSTALCODE*)
3. Naming Issues, in *TERRITORY* column, R reads “NA” as “Not Available”, but in this column of the dataset, “NA” is used as an abbreviation for “North America”.

**Pre-processing Task**

1. Correcting naming convention in *TERRITORY* column. Rename “NA” to “N.America”, because in the original dataset, “NA” was meant for “North America”. However, the “NA” was read as ‘not available’ in R.
2. Missing values in the *STATE* column are replaced with”NA”, not available.
3. Convert classes of some columns to the correct ones. The *ORDERNUMBER, QTR\_ID, MONTH\_ID, YEAR\_ID* from integer class to character. The values of these columns should be categorical and are not used for calculations. The *ORDERDATE* was read as character class, the lubridate package was used to transform the column’s values into date type. On the other hand, the *DEALSIZE* column’s class is transformed to factor, with levels “Small”, “Medium”, “Large”.
4. Dropping insignificant columns from the dataset. These columns are dropped because they are inessential to the analysis and contains too many missing and wild values. The columns that are considered ignorable and dropped are *PHONE*, *ADDRESSLINE1*, *ADDRESSLINE2*, *POSTALCODE*, *CONTACTLASTNAME* and *CONTACTFIRSTNAME*.