## Assignment

## Object: -

## Customer segmentation for a Retail chain

The given dataset consists of historical data of a retail chain in different countries. Your objective is to do the Data Cleaning, Data Imputation, Data Visualization, and Segment the customers with statistical method and algorithm of your choice and find out

1. Can you identify your best customers?
2. Do you know who your worst customers are?
3. Do you know which customers you just lost, and which ones you’re about to lose?
4. Can you identify loyal customers who buy often, but spend very little?
5. Can you target customers who are willing to spend the most at your store?

## Solution: -

Given “retail\_chain” Data set containing 541909 rows and 9 variables,

For this assignment, I used Python programming and Tableau Public for Visualization.

For python, I have written the command along with the output, whatever I did to achieve the object (Data Cleaning, Data Imputation, and Prediction of Amount spend by Customers) also give the answer of all the questions asked.

For Tableau, I have created 5 different sheets and one Dashboard. And attached the Link of Tableau workbook with this document.

First Part is Data Cleaning and Data Imputation

**Import required Library**

import pandas as pd # for dataframes

import matplotlib.pyplot as plt # for plotting graphs

import seaborn as sns # for plotting graphs

import datetime as dt

**Loading Dataset**

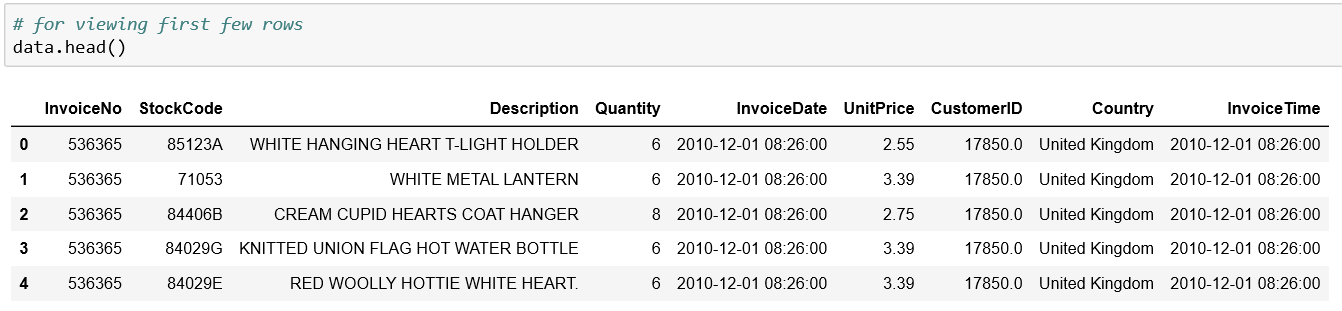
Let’s first load the given dataset using the pandas read\_excel function.

data = pd.read\_excel (\*\*\*\*\*/customer-segment/retail\_chain.xlsx')

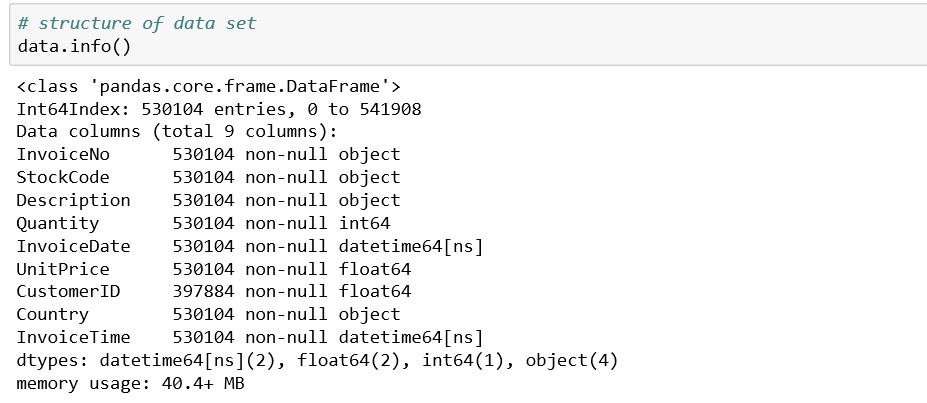
**View the First 5 rows of the dataset.**

data.head(5)

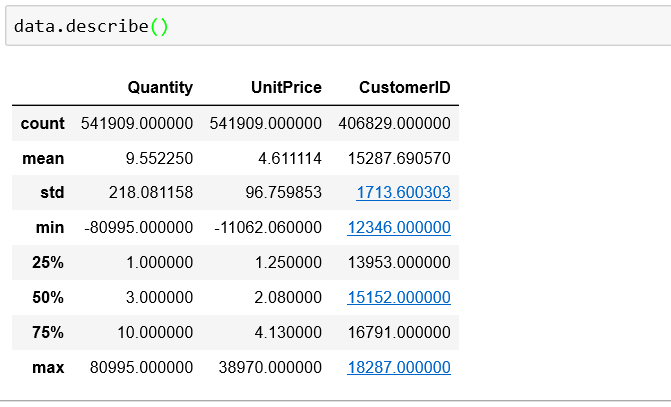
The output of this command is: -



**Structure of the dataset**



**Describe the Dataset**



**Removing Rows**

I observe some of the customers have negative values in quantity, which is not possible. Because no one can order anything in negative values or zero. So, I need to drop the “Quantity” smaller than zero.

indexNames = data[data['Quantity'] < 0].index

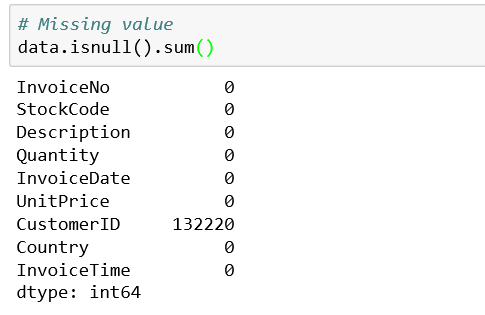
data.drop(indexNames , inplace=True)

In the above table, Can see that, there are negative values in variable “UnitPrice”, which is not possible Unit Price can’t be negative. I need to filter Unit Price smaller than equal to zero.

indexNames1 = data[data['UnitPrice'] <= 0].index

data.drop(indexNames1 , inplace=True)

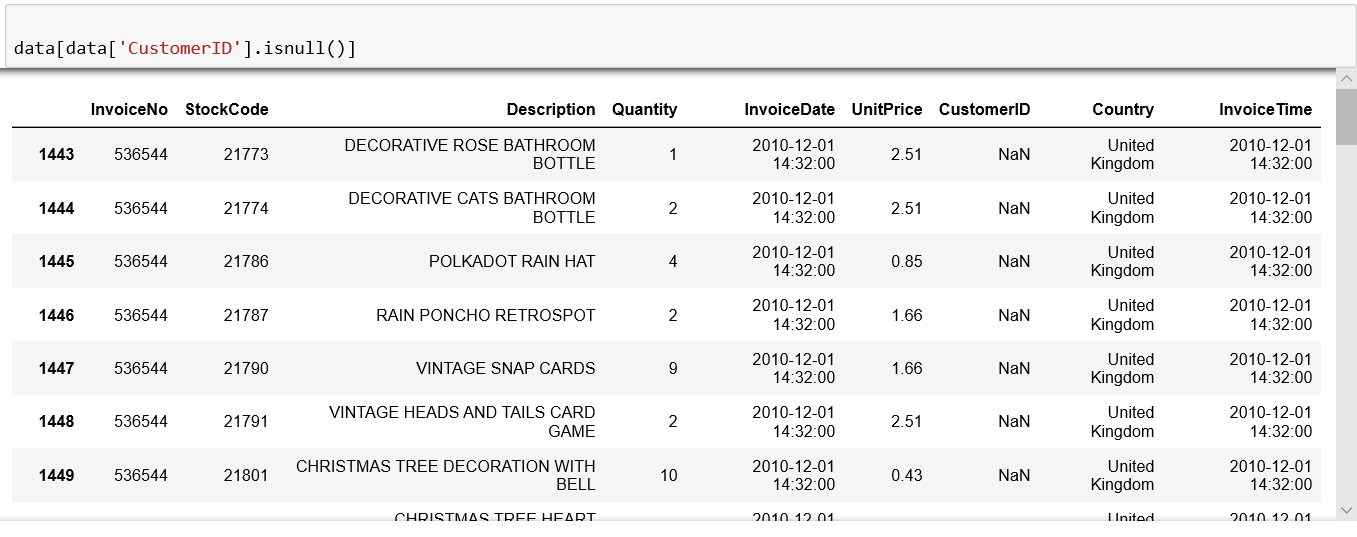
View Missing Value in the dataset by “isnull” command.



There are 132220 missing values in the Customer ID variable,

I assumed Customer ID as a primary key for this data so I removed all the rows which does not contain any value (NaN) in the Customer ID column by using the “drop” command.

I have selected only those rows which have NaN value in the Customer ID column.



Deleted all the rows which has NaN values in Customer ID column.

new\_data=data.dropna(how='any')

**Create and Add New Column**

Now I have created a new column named as “Amount”, this column contains the product of Quantity and Unit Price values.

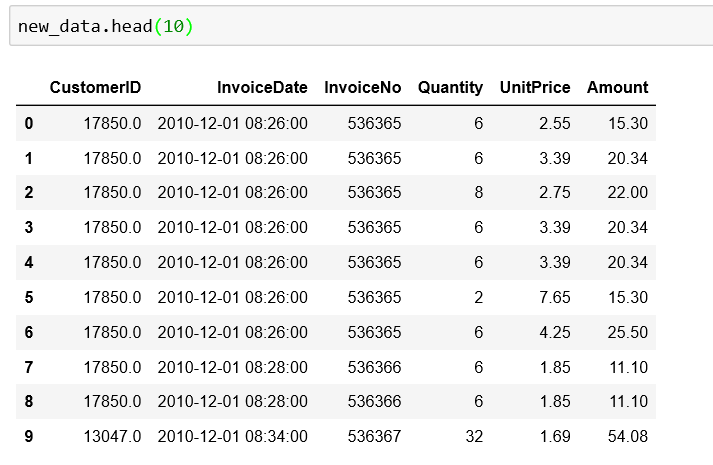
Quantity purchased in each transaction and Unit Price of each unit purchased by the customer will help me to calculate the total purchased Amount.

Amount = (new\_data.Quantity)\*(new\_data.UnitPrice)

Add column Amount into above dataset and changed name, from data to new\_data.

new\_data['Amount'] = Amount

Now print top 10 rows of new\_data



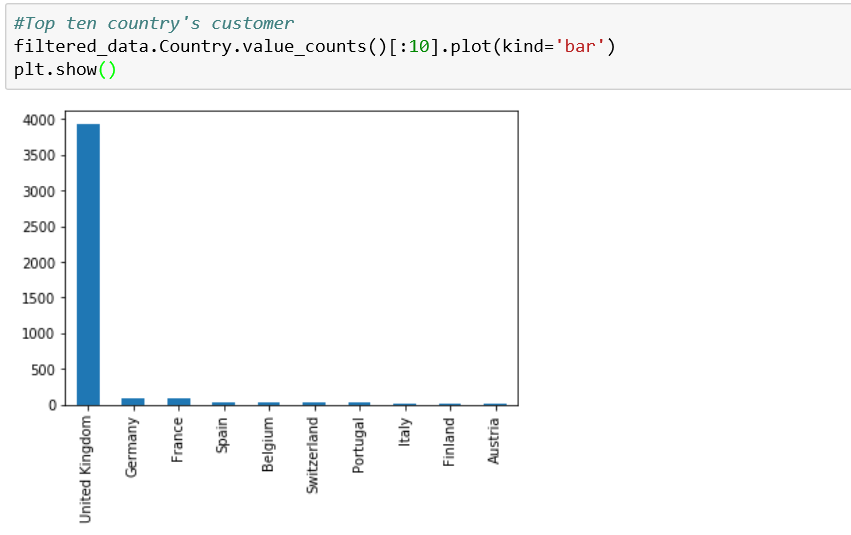
**Removing Duplicates**

I used “drop\_duplicates” command for removing duplicates.

filtered\_data=data[['Country','CustomerID']].drop\_duplicates()

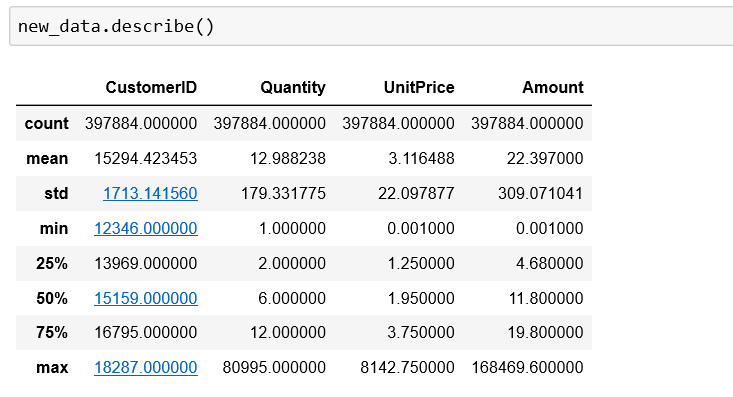
**Bar chart**

Bar chart for top ten country's customers.



In the given dataset, I observed most of the customers are from the "United Kingdom".

**Describe Data**

The describe() function in pandas is convenient in getting various summary statistics. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

**Filter required Columns**

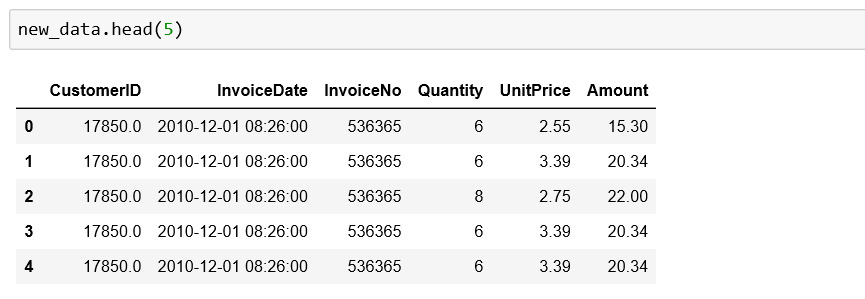
Here, I filtered the necessary columns for RFM analysis. I only need here five columns “CustomerID”, “InvoiceDate”, “InvoiceNo”, “Quantity”, and “UnitPrice”. “CustomerID” will uniquely define me, customers, “InvoiceDate” help me to calculate “recency” of purchase, “InvoiceNo” helps me to count the number of times transactions performed (frequency).

new\_data=new\_data[['CustomerID','InvoiceDate','InvoiceNo','Quantity','UnitPrice', 'Amount']]

new\_data['InvoiceDate'].min(),new\_data['InvoiceDate'].max()

PRESENT = dt.datetime(2011,12,10)

new\_data['InvoiceDate'] = pd.to\_datetime(new\_data['InvoiceDate'])



**RFM Analysis**

Then, I performed the following operations:

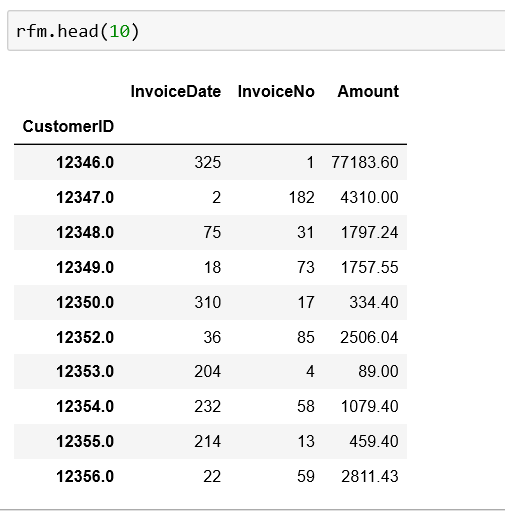
For “*Recency”*, calculate the number of days between present dates and date of last purchase each customer.

For *“Frequency”*, calculate the number of orders for each customer.

For *“Monetary”*, calculate the sum of the purchase price for each customer.



Output: -

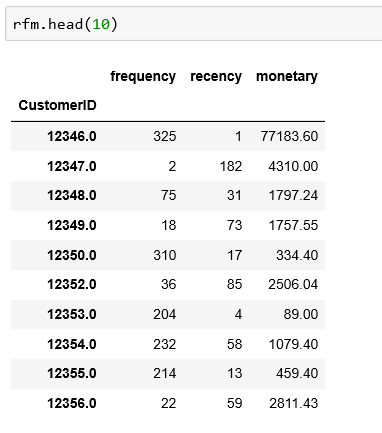


Change column’s Name

rfm.columns=['frequency','recency','monetary']

rfm['recency'] = rfm['recency'].astype(int)

Output: -



**Computing Quantile of RFM values**

Customers with the Lowest “recency”, Highest “frequency” and “monetary” amounts considered as Top Customers.

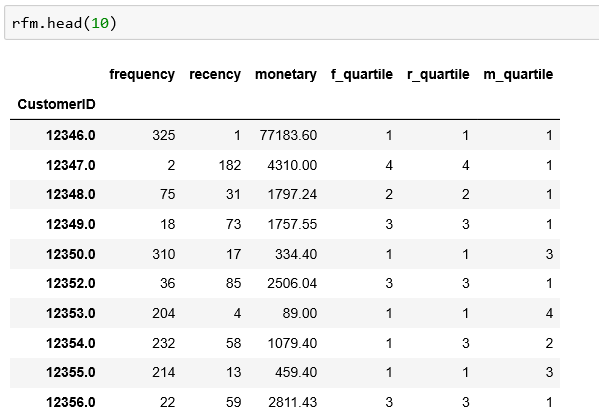
qcut() is Quantile-based discretization function. qcut bins the data based on sample quantiles. For example, 1000 values for 4 quantiles would produce a categorical object indicating quantile membership for each customer.

rfm['r\_quartile'] = pd.qcut(rfm['recency'], 4, ['1','2','3','4'])

rfm['f\_quartile'] = pd.qcut(rfm['frequency'], 4, ['4','3','2','1'])

rfm['m\_quartile'] = pd.qcut(rfm['monetary'], 4, ['4','3','2','1'])

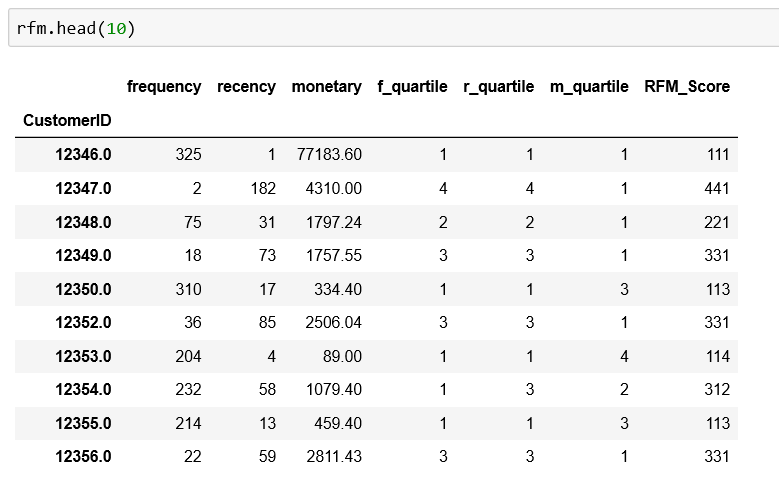
Output: -



**RFM Result Interpretation**

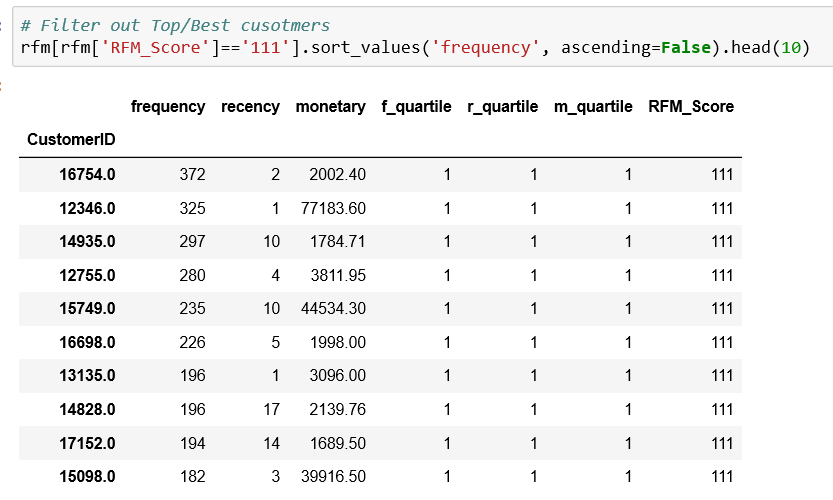
Combine all three quartiles (“r\_quartile”, ”f\_quartile”, “m\_quartile”) in a single column, this rank will help to segment the customers well group.

rfm['RFM\_Score'] = rfm.r\_quartile.astype(str)+ rfm.f\_quartile.astype(str) + rfm.m\_quartile.astype(str)



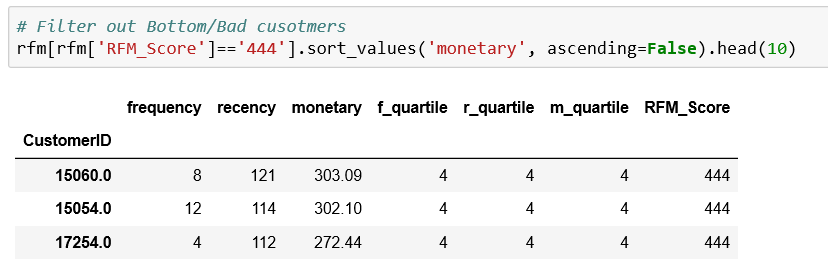
1. **Can you identify your best customers?**

**Filter out Top/Best customers**



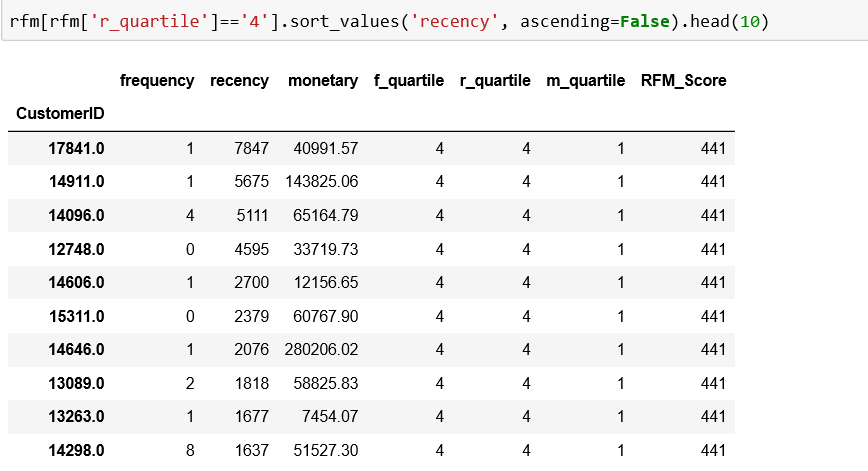
1. **Do you know who your worst customers are?**

**Filter out Bottom/worst customers**



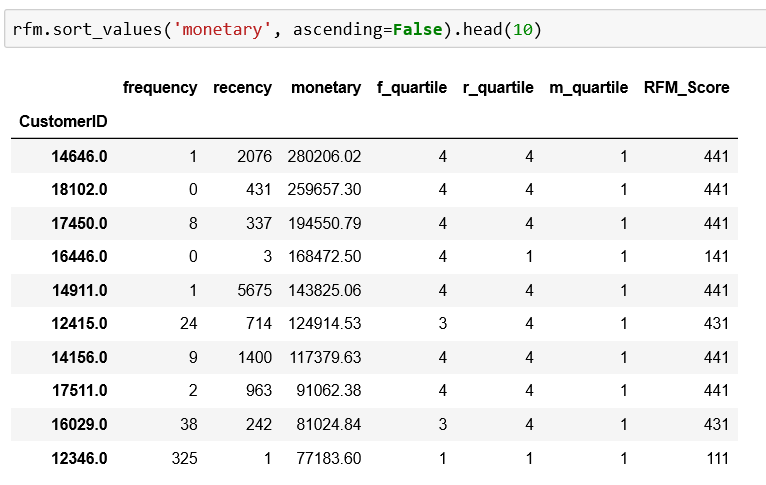
1. **Can you identify loyal customers who buy often, but spend very little?**

Top 10 Customers who buy a maximum products and visited maximum time



1. **Can you target customers who are willing to spend the most at your store?**

Top 10 customers who spend maximum amount

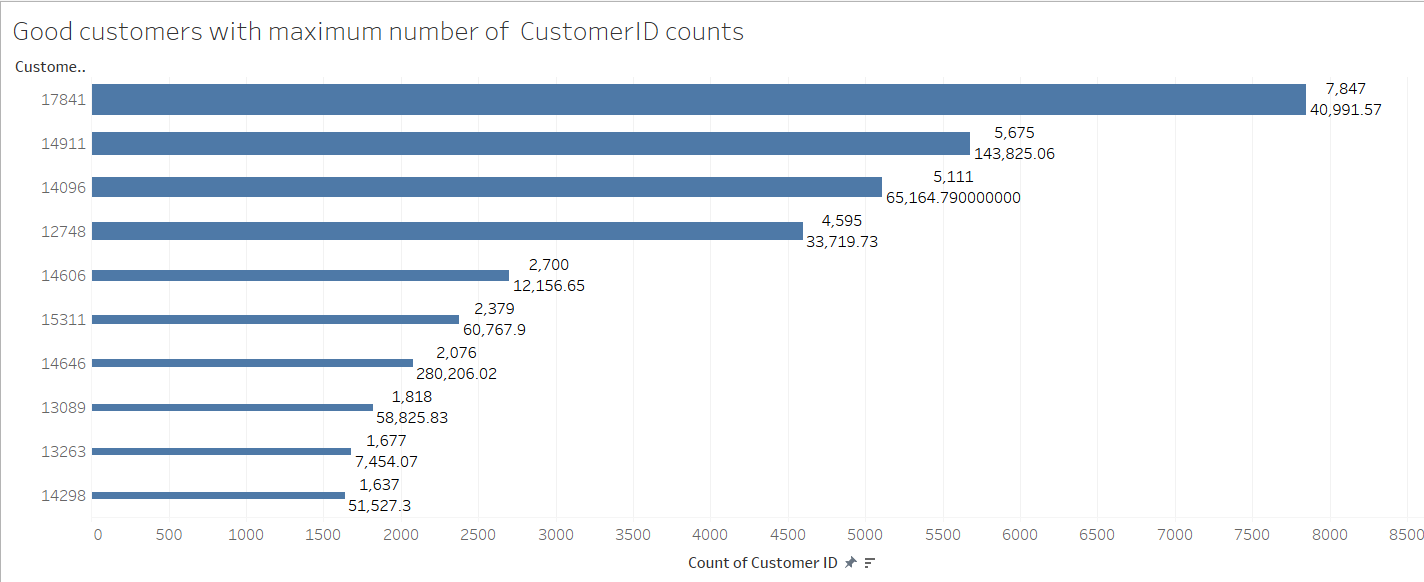


**Data visualization by Tableau Public**

This picture shows Count of Customer ID and Amount spend by customer.

This is the link of this Tableau sheet.

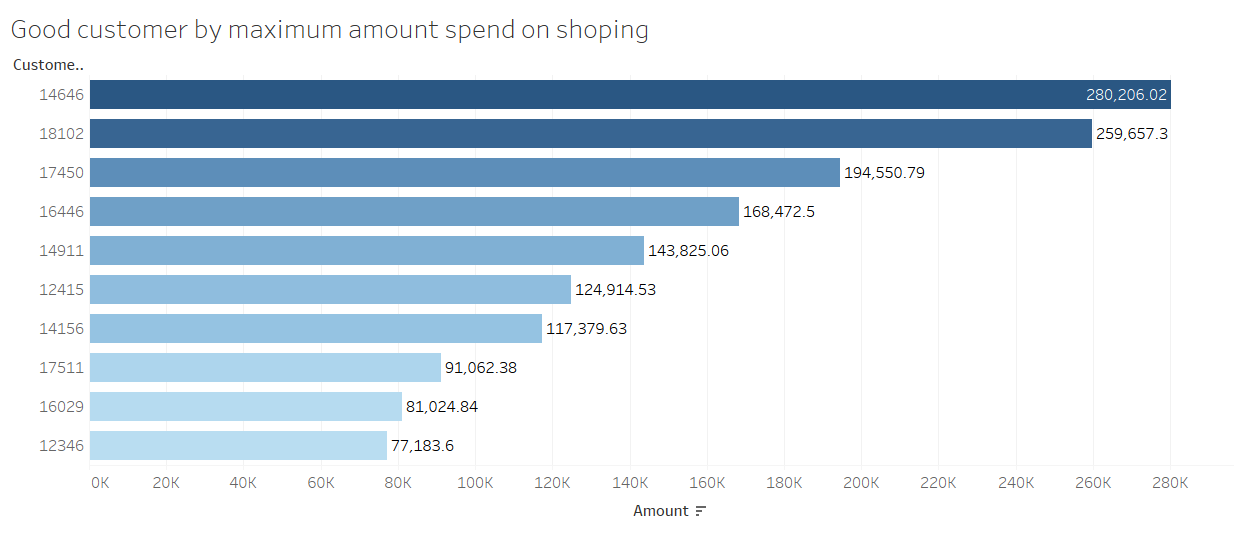
<https://public.tableau.com/profile/rohit.kumar3308#!/vizhome/retail_chain/goodbyidcount>

****

This picture shows maximum Amount spend by customer.

This is the link of this Tableau sheet.

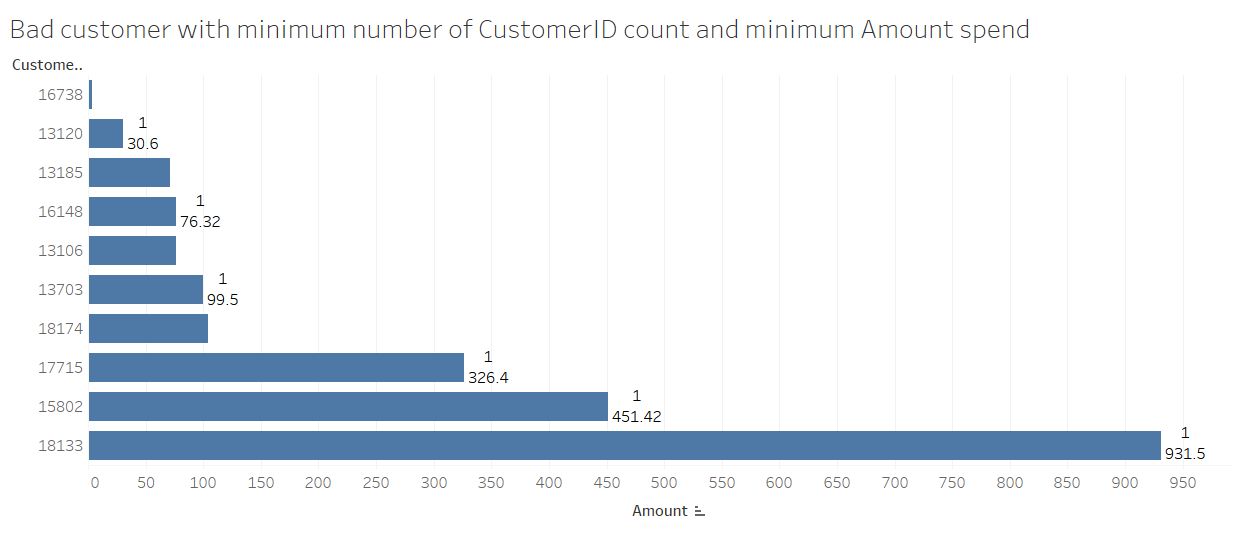
<https://public.tableau.com/profile/rohit.kumar3308#!/vizhome/retail_chain/goodbymaxamount>

****

This picture shows minimum Customer ID count and minimum Amount spend by customer.

This is the link of this Tableau sheet.

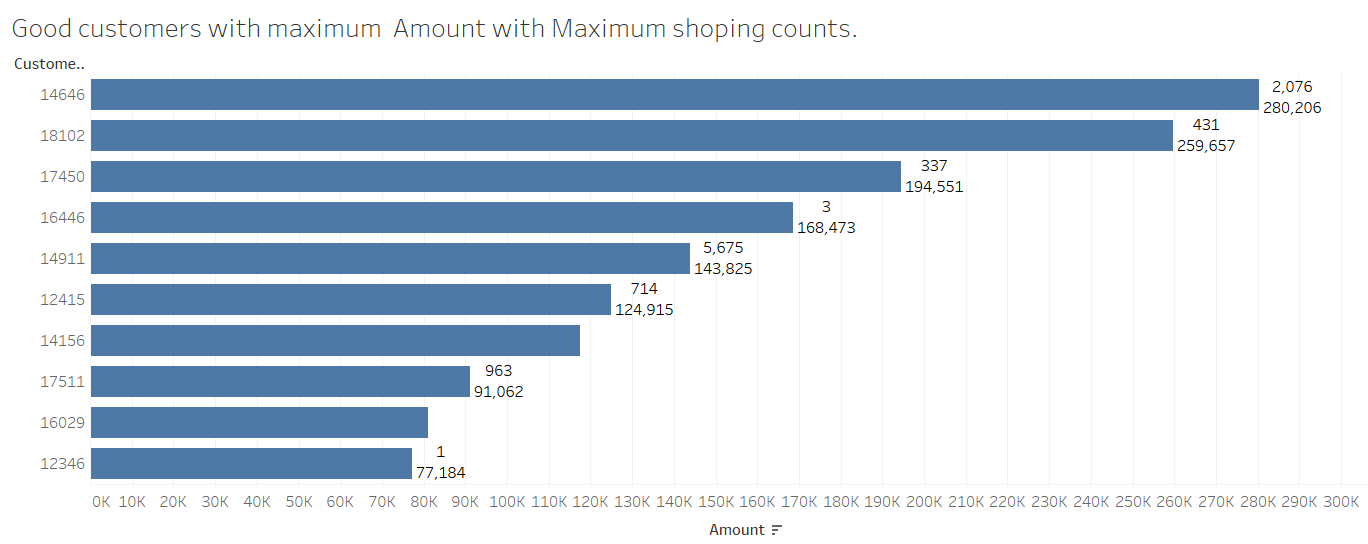
<https://public.tableau.com/profile/rohit.kumar3308#!/vizhome/retail_chain/badbyminamountandmincount>

****

This picture shows maximum Customer ID count and maximum Amount spend by customer.

This is the link of this Tableau sheet.

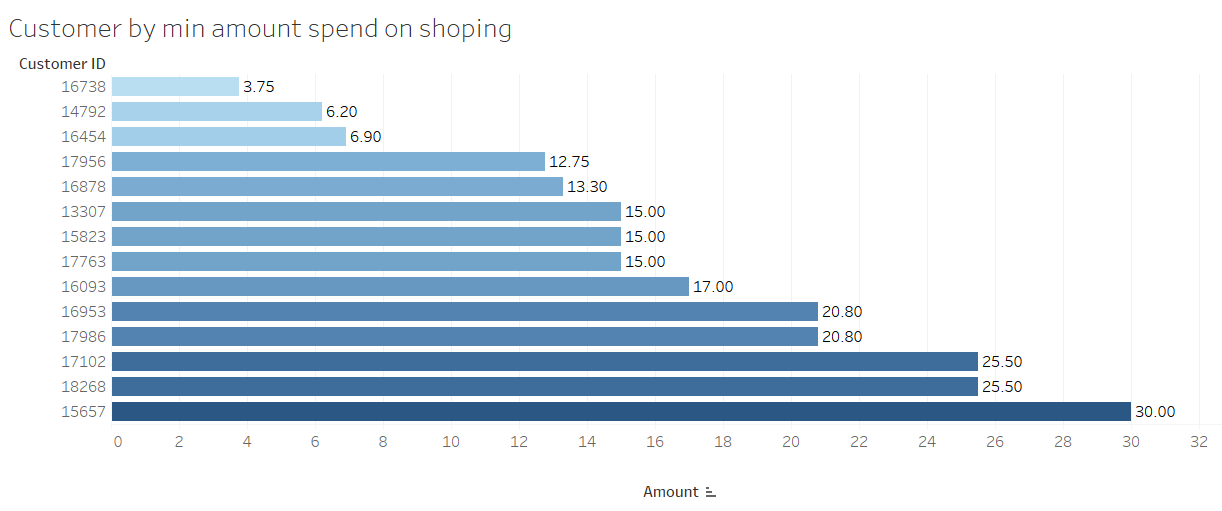
<https://public.tableau.com/profile/rohit.kumar3308#!/vizhome/retail_chain/minamountmaxcount>

****

This picture shows maximum Customer ID count and maximum Amount spend by customer.

This is the link of this Tableau sheet.

<https://public.tableau.com/profile/rohit.kumar3308#!/vizhome/retail_chain/badbyminamount>

****