1.Perform a preliminary data inspection and data cleaning.
a. Check for missing data and formulate an apt strategy to treat them.
b. Remove duplicate data records.
c. Perform descriptive analytics on the given data.

There were columns Description and Customer ID having null values. We will remove those records.
There were 3124 duplicate records. We will remove them as well.
We will also remove the orders which were reversed.

Descriptive analysis:

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

2. Perform cohort analysis. Observe how a cohort behaves across time and compare it to other cohorts.
a. Create month cohorts and analyze active customers for each cohort.
b. Analyze the retention rate of customers.

For cohort analysis, there are a few labels that we have to create:

- Invoice period: A string representation of the year and month of a single transaction/invoice.
- Cohort group: A string representation of the year and month of a customer's first purchase. This label is common across all invoices for a particular customer.
- Cohort period / Cohort Index: A integer representation a customer's stage in its "lifetime". The number represents the number of months passed since the first purchase.

- Monthly active customers from each cohort

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12-01	885.0	324.0	286.0	340.0	321.0	352.0	321.0	309.0	313.0	350.0	331.0	445.0	235.0
2011-01-01	417.0	92.0	111.0	96.0	134.0	120.0	103.0	101.0	125.0	136.0	152.0	49.0	NaN
2011-02-01	380.0	71.0	71.0	108.0	103.0	94.0	96.0	106.0	94.0	116.0	26.0	NaN	NaN
2011-03-01	452.0	68.0	114.0	90.0	101.0	76.0	121.0	104.0	126.0	39.0	NaN	NaN	NaN
2011-04-01	300.0	64.0	61.0	63.0	59.0	68.0	65.0	78.0	22.0	NaN	NaN	NaN	NaN
2011-05-01	284.0	54.0	49.0	49.0	59.0	66.0	75.0	27.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	242.0	42.0	38.0	64.0	56.0	81.0	23.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	188.0	34.0	39.0	42.0	51.0	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	169.0	35.0	42.0	41.0	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	299.0	70.0	90.0	34.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	358.0	86.0	41.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	324.0	36.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	41.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- **Retention Rate**

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12-01	100.0	36.6	32.3	38.4	36.3	39.8	36.3	34.9	35.4	39.5	37.4	50.3	26.6
2011-01-01	100.0	22.1	26.6	23.0	32.1	28.8	24.7	24.2	30.0	32.6	36.5	11.8	NaN
2011-02-01	100.0	18.7	18.7	28.4	27.1	24.7	25.3	27.9	24.7	30.5	6.8	NaN	NaN
2011-03-01	100.0	15.0	25.2	19.9	22.3	16.8	26.8	23.0	27.9	8.6	NaN	NaN	NaN
2011-04-01	100.0	21.3	20.3	21.0	19.7	22.7	21.7	26.0	7.3	NaN	NaN	NaN	NaN
2011-05-01	100.0	19.0	17.3	17.3	20.8	23.2	26.4	9.5	NaN	NaN	NaN	NaN	NaN
2011-06-01	100.0	17.4	15.7	26.4	23.1	33.5	9.5	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	100.0	18.1	20.7	22.3	27.1	11.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	100.0	20.7	24.9	24.3	12.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	100.0	23.4	30.1	11.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	100.0	24.0	11.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	100.0	11.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Build a RFM (Recency Frequency Monetary) model. *Recency* means the number of days sin ce a customer made the last purchase. *Frequency* is the number of purchase in a given

period. It could be 3 months, 6 months or 1 year. *Monetary* is the total amount of money a customer spent in that given period. Therefore, big spenders will be differentiated amon g other customers such as MVP (Minimum Viable Product) or VIP.

In the RFM analysis we are going to group the data based on percentiles and calculate indivi dual recency, frequency, monetary and merge them to a table.

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	366	182	4310.00
2	12348.0	357	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	309	17	334.40

2. Calculate RFM metrics.
3. Build RFM Segments. Give recency, frequency, and monetary scores individually by dividing them into quartiles.
b1. Combine three ratings to get a RFM segment (as strings).
b2. Get the RFM score by adding up the three ratings.
b3. Analyze the RFM segments by summarizing them and comment on the findings.

	CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMSegment	RFMScore
1609	14547.0	53	361	3989.79	4	4	4	444	12
3752	17528.0	70	253	3628.50	4	4	4	444	12
1298	14121.0	59	159	2780.15	4	4	4	444	12
3438	17084.0	34	152	2791.28	4	4	4	444	12
398	12856.0	63	312	2175.73	4	4	4	444	12

From the RFM metrics it was found that:
Best Customers: 38
Frequent Customers: 1060
Money spending Customers: 1072
Lost Customers: 96

1. Create clusters using k-means clustering algorithm.
a. Prepare the data for the algorithm. If the data is asymmetrically distributed, manage the skewness with appropriate transformation. Standardize the data.
b. Decide the optimum number of clusters to be formed.
c. Analyze these clusters and comment on the results.

The skewness for Frequency and Monetary was found out to be right skewed.
We will try to make the curve as normally distributed as possible by using appropriate Techniques
So, we will first remove the outliers.

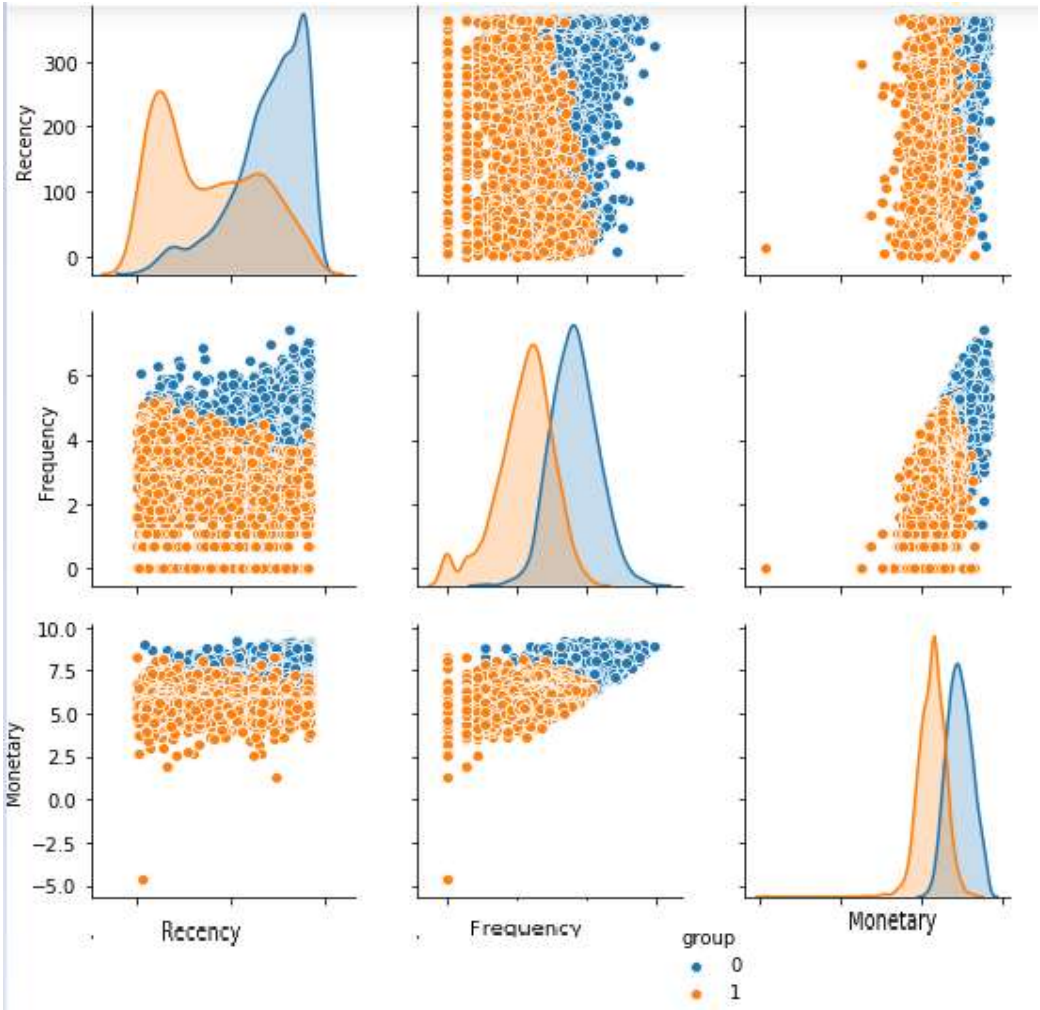
We will normalize the data using log transformation

We will create cluster analysis by taking optimum number of clusters to target the right cust omer and identify groups
To create cluster using kmeans we need to find the number of clusters to be used.

We will use silhouette coefficient technique to decide the optimal value of K
For n_clusters = 2 The average silhouette_score is : 0.38013836380490273
For n_clusters = 3 The average silhouette_score is : 0.3742839206122452
For n_clusters = 4 The average silhouette_score is : 0.33274966031463526
For n_clusters = 5 The average silhouette_score is : 0.31561443144249934
For n_clusters = 6 The average silhouette_score is : 0.3094804251916038

We will take the value as 2.

Analysis:



From the above curve it is found that:
Group 1- Dormant, spend less and made less number of transactions.
Group -0 - Active, frequent transactions and high monetary

1. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

a. Country-wise analysis to demonstrate average spend. Use a bar chart to show the monthly figures

b. Bar graph of top 15 products which are mostly ordered by the users to show the number of products sold

c. Bar graph to show the count of orders vs. hours throughout the day

d. Plot the distribution of RFM values using histogram and frequency charts

e. Plot error (cost) vs. number of clusters selected

f. Visualize to compare the RFM values of the clusters using heatmap

https://public.tableau.com/profile/bedant8177#!/vizhome/retail_1_16103085412550/Dashboard1?publish=yes