King Saud University
College of Computer and Information Sciences
Information Technology department

IT 326: Data Mining

Course Project

Online Retail

Project Report

Group#:	***Section#:***	***Name***		***ID***
#1	***56548***	***Leader***	*Ruwayda Alshowiman*	*###*
		2	*Ayan alqahtani*	*###*
	Email:	***3***	*Amjad alotibi*	*###*
	###@student.ksu.edu.sa	***4***	*Reema al Qahtani*	*###*

12/2/19

Table of Contents

1 Problem

Our idea is to analyze the Online Retail Data Set. By analyzing them, will we use a massive amount of data for predictive analytics that lead to increased sales.
With an increasing competition and a business' motivation to improve profits, retail has become one of the early adopters of data analytics. One of the primary goals why predictive analytics is crucial in a business is to optimize customer relationships that would lead to customer retention, improve revenues, and increase values on their products.
Based on our findings, we'll be able to find the best way to understand customer behavior and help managers improve their sales initiatives.

2 Data Mining Task

We plan to do Clustering task on our data set. Basically, it is used to identify data objects that are similar to one another and group them together. in our data set, we plan to perform clustering on the country column to group the countries based on where the customer resides.

3 Data

It's a multivariate, Sequential, Time-Series data for non-store Online Retail. We get the data from Center for Machine Learning and Intelligent Systems by Dr Daqing Chen, Director: Public Analytics group. School of Engineering, London South Bank University, London SE1 0AA, UK.
The data set made up from 8 attributes and 541910 observation.

- ***InvoiceNo (Nominal)***
  Missing value:0
  number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation

- ***StockCode (Nominal)***
Missing value:0
  number uniquely assigned to each distinct product.

- ***Description (Nominal Data)***
Missing value:0
  Product (item) name

- ***Quantity (Numeric Data)***

Missing value:0
The quantities of each product (item) per transaction.

- ***Unit Price (Numeric Data)***

Missing value:0
Product price per unit in sterling.

- ***InvoiceDate (Numeric Data)***

Missing value:0
The day and time when each transaction was generated.

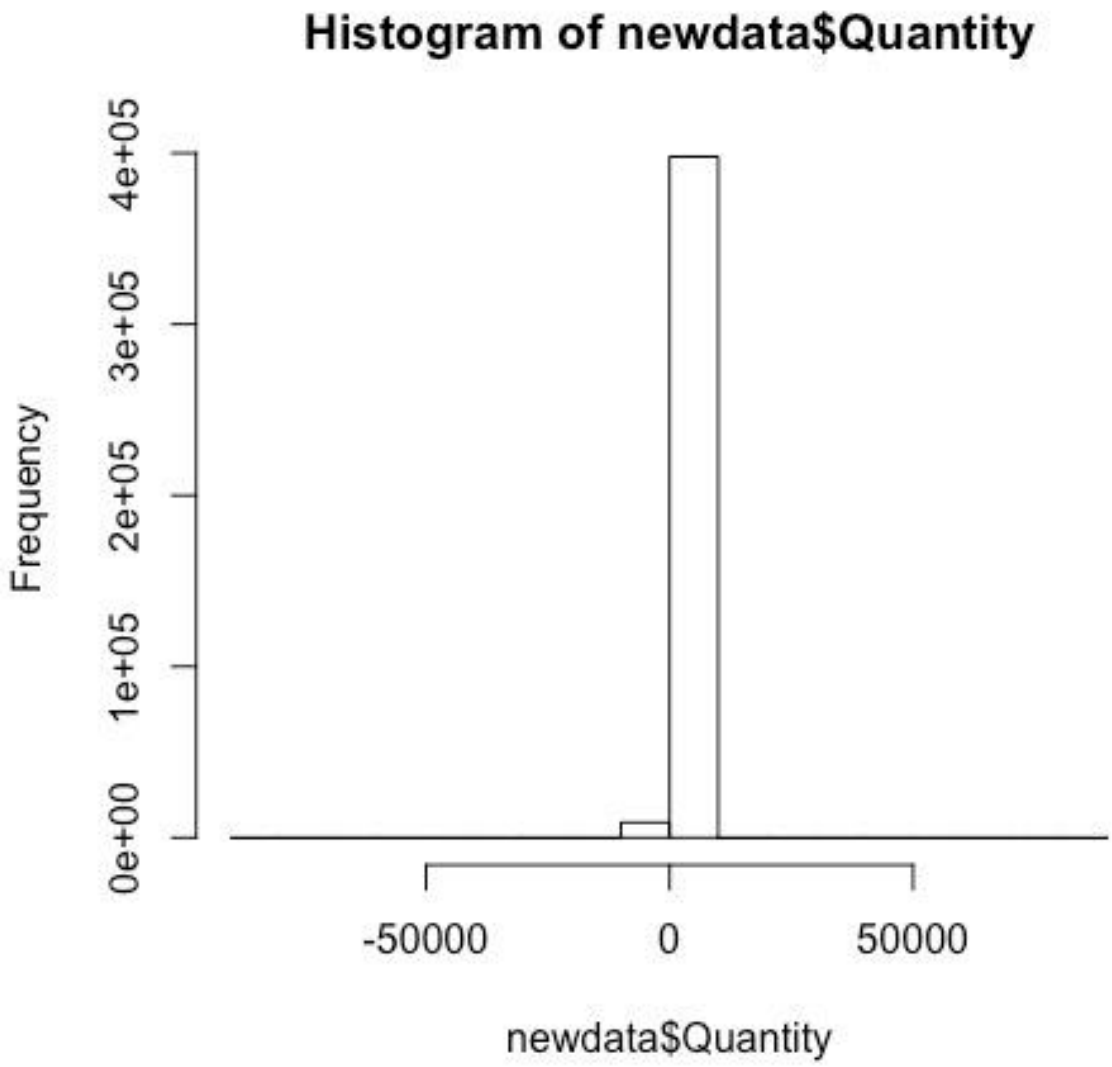
- ***CustomerID (Nominal Data)***

Missing value:0
The number uniquely assigned to each customer.

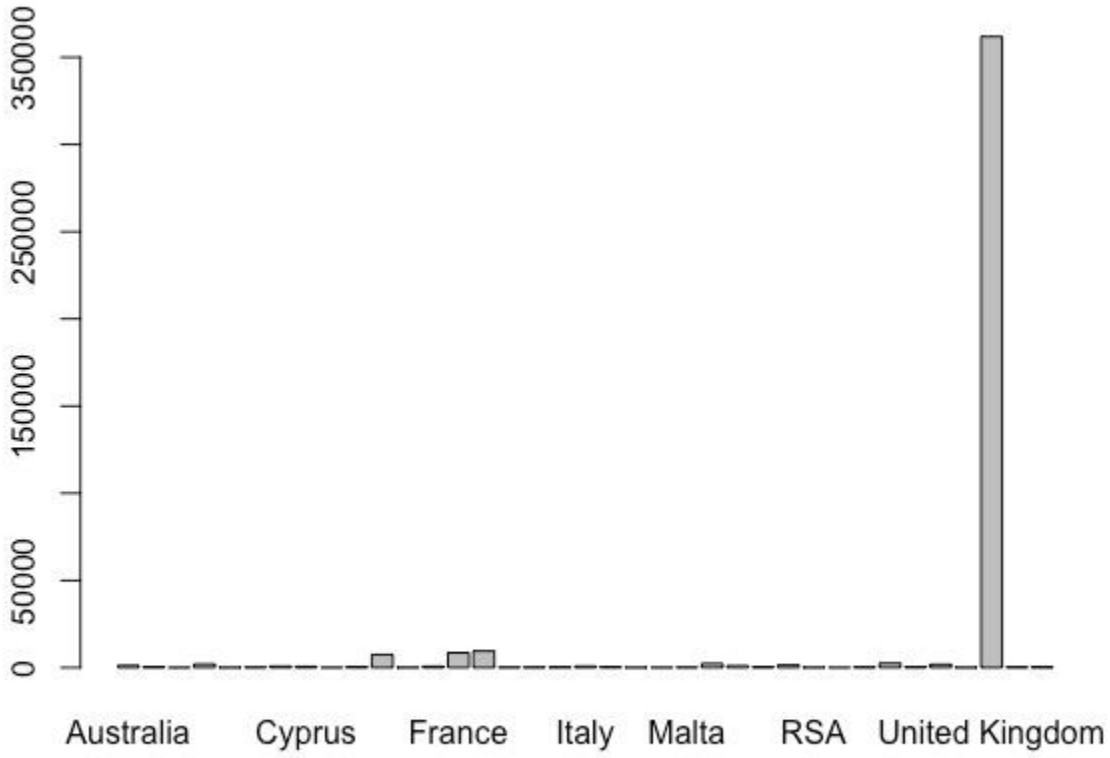
- ***Country (Nominal Data)***

Missing value:0
The name of the country where each customer resides.

- ***Histogram:*** *in x axis (Quantity) in y axis (Frequency), In this diagram the frequency of the quantity is very high and also has a very high period for other periods and also have a few values, but the frequency is large.*



- ***Bar Chart:*** *in x axis (Country) in y axis (Frequency), I have deduced from the chart that most customers are from the UK and less customer than Cyprus, Italy and Malta.*



4 Data preprocessing

Data processing is required to handle with incomplete data, like missing attribute values, And noisy data containing errors or outlier, also to deal with inconsistent data.

1- Cleaning Data

First step in data preprocessing, we will

a- Identify N\A attribute

b- remove all the NA rows from the dataset.

a-

raw data:

```
Data
Online_Ret… 541909 obs. of 8 variables
```

Identify N\A attribute using the following code:

```
sum(is.na(Online_Retail))
```

preprocessed data:

```
> sum(is.na(Online_Retail))
[1] 136534
>
```

b- remove all the NA rows from the dataset using the following code:

```
newdata <- na.omit(Online_Retail)
```

preprocessed data:

```
Data
   newdata    406829 obs. of 8 variables
```

c- ***Identify outliers*** *using (outliers) package in R remove all outliers from the dataset. We identified the outliers in the country column in the new cleaned data (newdata), which can be subsisted and using newdata$UnitPrice command:*

```
outlier_tf = outlier(newdata$UnitPrice,logical=TRUE)
```

Result:

outlier_tf	Large logical (406829 elements, 1.6 Mb)

Then:

```
sum(outlier_tf)
```

Result:

```
> sum(outlier_tf)
[1] 1
>
```

Then specify were the outlier was (it's location):

```
find_outlier = which(outlier_tf==TRUE,arr.ind=TRUE)
```

Result:

find_outlier	159241L

Then Remove the outlier:

```
newdata = newdata[-find_outlier,]
nrow(newdata)
```

preprocessed data:

newdata	406827 obs. of 8 variables

2- Data integration

This step needed if we Integration of multiple databases, data cubes, or files.
In our case, we use a data set from single source, so we don't have to apply it.

3- Data reduction

In predictive modeling, dimensionality reduction or dimension reduction is the process of reducing the number of irrelevant variables. It is a very important step of predictive modeling.

we make sure that our data does not need reduction.

4- Data transforming

Data transformation is a fancy term for changing the values of observations through some mathematical operation.

We applied tow type of transformation,

First, we normalize (Quantity) attribute by using the following code:

```
newdata$Quantity<-normalize(newdata$Quantity)
```

Quantity column before normalization:

	V1
1	6
2	6
3	8
4	6
5	6
6	2
7	6
8	6
9	6
10	32
11	6

Quantity column after normalization:

	V1
1	0.5000370
2	0.5000370
3	0.5000494
4	0.5000370
5	0.5000370
6	0.5000123
7	0.5000370
8	0.5000370
9	0.5000370
10	0.5001975
11	0.5000370

Second, we converted the attribute (Country) from nominal type to numeric to simplify clustering process

Country column before transformation (Nominal)

	V1
1	United Kingdom
2	United Kingdom
3	United Kingdom
4	United Kingdom
5	United Kingdom
6	United Kingdom
7	United Kingdom
8	United Kingdom
9	United Kingdom
10	United Kingdom

We applied the following code

```
newdata$Country <- as.factor(newdata$Country)
is.factor(newdata$Country)
newdata$Country <- as.numeric(newdata$Country)
```

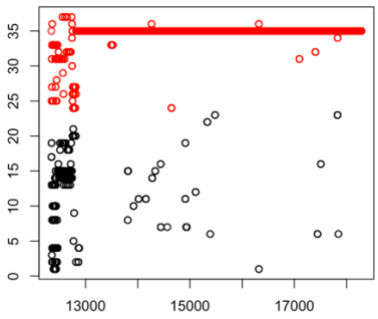
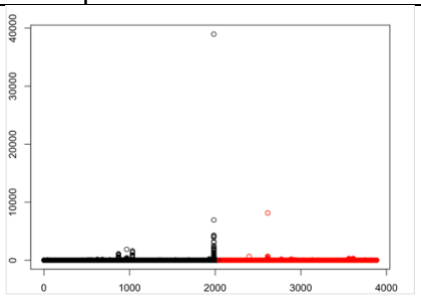
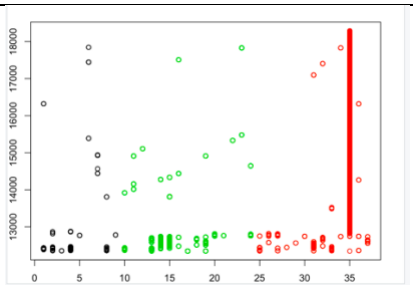
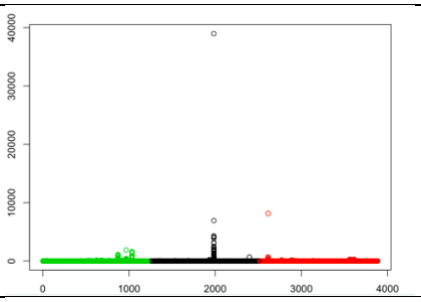
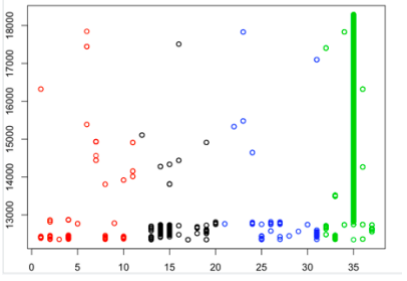
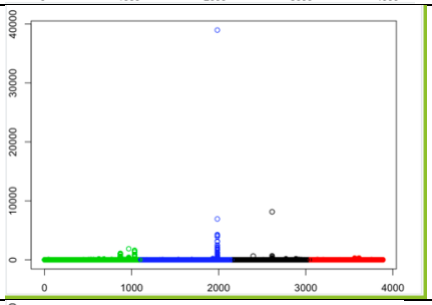
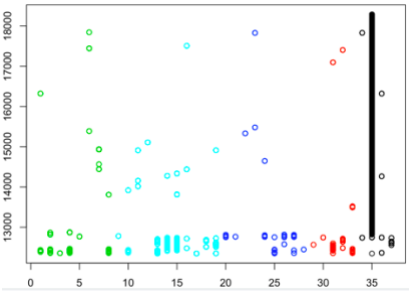
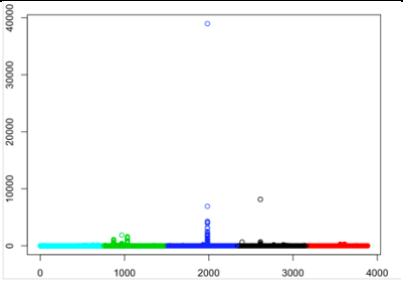
Country column after transformation (Numeric)

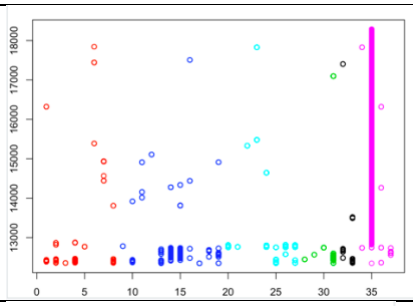
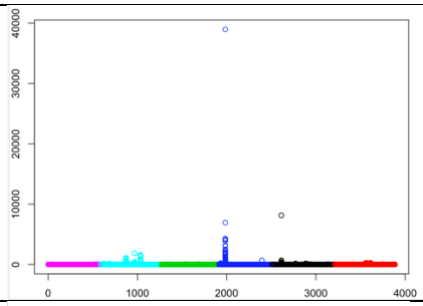
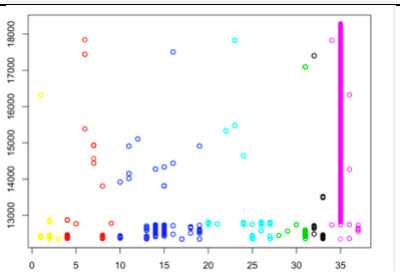
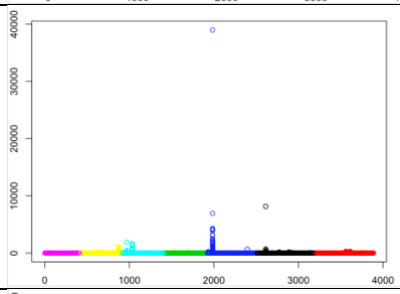
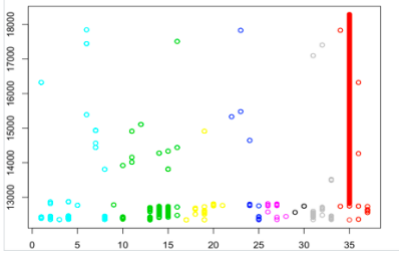
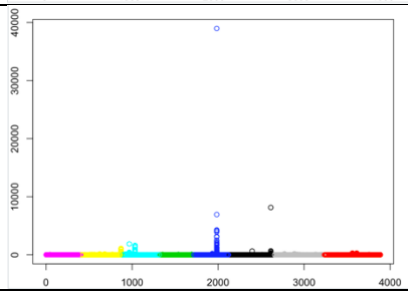
	V1
1	35
2	35
3	35
4	35
5	35
6	35
7	35
8	35
9	35
10	35

5 Data Mining Technique

We choose Clustering technique to group a set of objects in such a way that objects in the same cluster .We used it on country column to see how our dataset will be group based on where the customer resides .They are different types of clustering methods and we will use K-means clustering because it is the most popular partitioning method . The K-means requires to specify the number of clusters. After we prepare the dataset in preprocessing section, we will use the kmeans () function that indicate the number of clusters (k). As well as this we will import "ggmap" library to visualize the results of the k-means clustering, also we will import "cluster" library to perform clustering algorithm.

6 Evaluation and Comparison

K	Country	description	Sum of squares
2			Country=93.2% Description=73%
3			Country=96% Description=88.7%
4			Country=98.2% Description=93.3%
5			Country=99% Description=95.7%

6			Country=99.1% Description=97.2%
7			Country=99.2% Description=97.8%
8			Country=99.3% Description=98.3%

Depending on the table above we perform the technique on two attributes such as country and description and we see the most accurate attribute is country because it has **99.3%** Sum of squares.

7 Findings

we execute Evaluation and Comparison on two attributes such as country and description and we find the country the best attributes to help solve our problem is by knowing the country with the largest number of customers, the country attribute is interesting because we can increase the number of advertisement and sales in the country with the largest customers.

8 Code

```
#Read the data
        library("readxl")
        myData <- read_excel ("/Users/ruwayda/Desktop/Online Retail.xlsx”)

# clean the data
        myData <- na.omit(myData)

# View the first 3 rows of the data
        head(myData)

#Identify N\A attribute
        sum(is.na(myData))

#print the sum of the null in specific attribute
        sum(is.na(myData$Description))

#Identify outliers
        outlier_tf = outlier (myData $UnitPrice, logical=TRUE)
        sum(outlier_tf)

#specify outlier location
        find_outlier=which (outlier_tf==TRUE, arr.ind = TRUE)

#Remove the outlier
        myData = myData[-find_outlier,]
        nrow(myData)

#normalize (Quantity) attribuite
        myData$Quantity<-normalize(myData$Quantity)

#converted the attribute (Country) from nominal type to numeric
        newdata$Country <- as.factor(myData$Country)
        is.factor(myData$Country)
        myData$Country <- as.numeric(myData$Country)

#converted the attribute (Description) from nominal type to numeric
        myData$Description <- as.factor(myData$Description)
        is.factor(myData$Description)
        myData$Description <- as.numeric(myData$Description)
```

```
# View the first 3 rows of the data

        head(myData)
        str(myData)


# Compute k-means with k = N
        set.seed(8953)
        myData2 <- myData
        kmeans.result <- kmeans(myData2$Description,8)

#represent the size of the cluster and cluster means and cluster vectors also sum of squares by
cluster
        kmeans.result
        table(myData$Description,kmeans.result$cluster)

#Getting the descriptives of the data

        summary(myData)

#visualize the result

        par(mar = rep(2, 4))
        plot(myData2[,c("Description","UnitPrice")],col=kmeans.result$cluster)
        op <- par(oma=c(5,7,1,1))
        par(op)
        png(filename="myfile.png", res=150, width = 1000, height = 1000)
```

9 References

[1] Unh.edu, 2019. [Online]. Available: http://www.unh.edu/halelab/BIOL933/Labs/Lab6.pdf. [Accessed: 21- Nov- 2019].

[2] D. Bhalla, "Dimensionality Reduction with R", ListenData, 2019. [Online]. Available: https://www.listendata.com/2015/06/simplest-dimensionality-reduction-with-r.html. [Accessed: 21- Nov- 2019].

[3] W. Preprocessing, "Why is Data Preprocessing required? Explain the different steps involved in Data Preprocessing", Ques10.com, 2019. [Online]. Available : https://www.ques10.com/p/9224/why-is-data-preprocessing-required-explain-the-dif/. [Accessed: 21- Nov- 2019].

10 Tasks Distribution

ID	Name	Responsibilities
437200102	Ayan AL Qahtani	Data and Data preprocessing
437925223	Amjad AL Otaibi	Problem
437202002	Ruwayda Alshowiman	Data Mining Technique, Data Mining Task , Code, evaluation and comparison
437200969	Reema AL Qahtani	Histogram, bar chart, miss value,finding