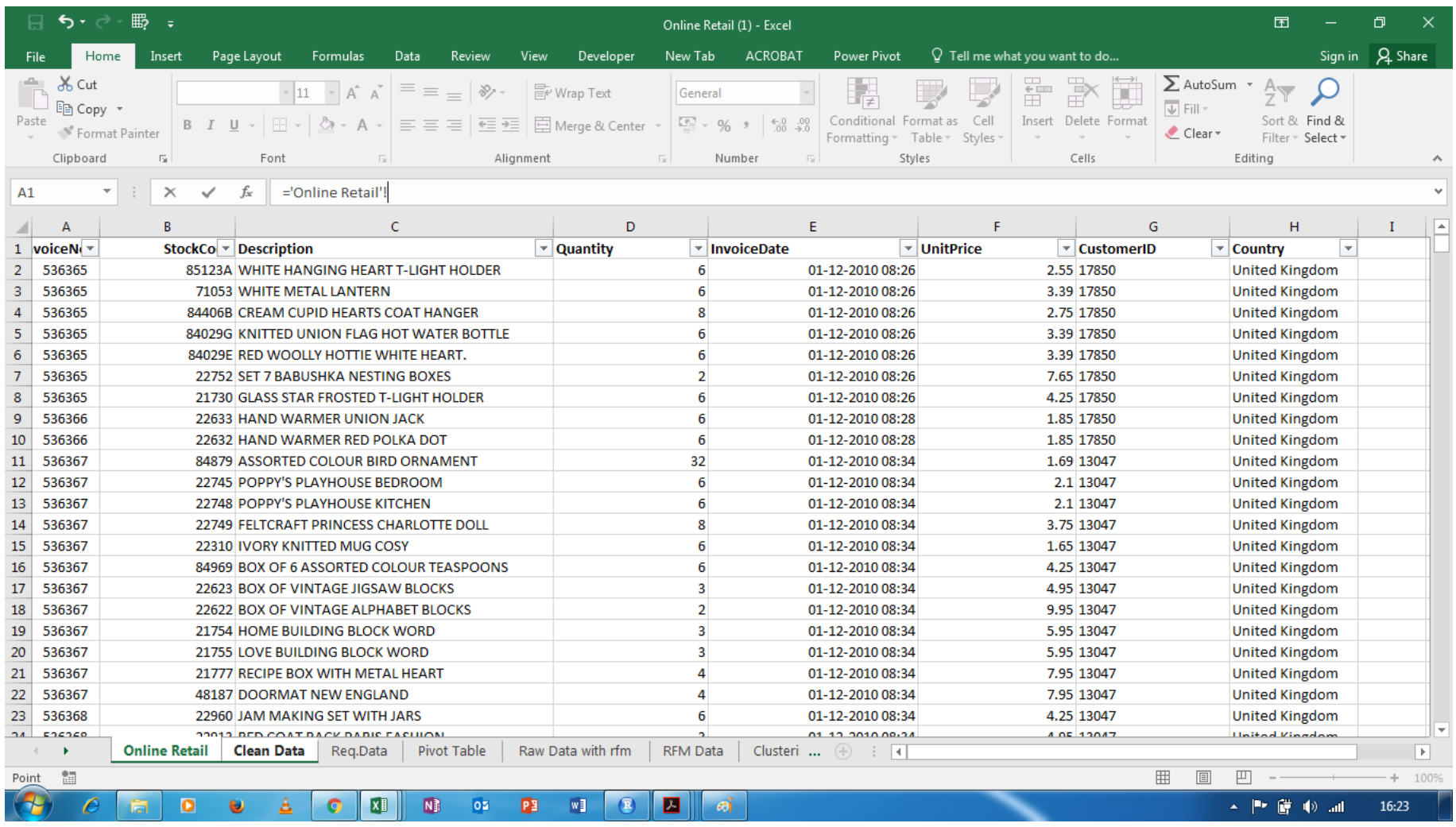
RFM Analysis in Excel and R

Group Members Ayush , Sonal and Vaibhav

Description of Dataset

Online_Retail.csv Dataset is Stock Exchange dataset from **01-12-2010** to **09-12-2011** ,This dataset have 541910 rows and eight columns.



	A	B	C	D	E	F	G	H
1	voiceN	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010 08:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010 08:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	01-12-2010 08:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01-12-2010 08:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	01-12-2010 08:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	01-12-2010 08:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	01-12-2010 08:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	01-12-2010 08:34	4.25	13047	United Kingdom


```
retail <- read.csv("online_retail.csv")
```

We will view the dataset and check the NA values in data

```
head(retail)
```

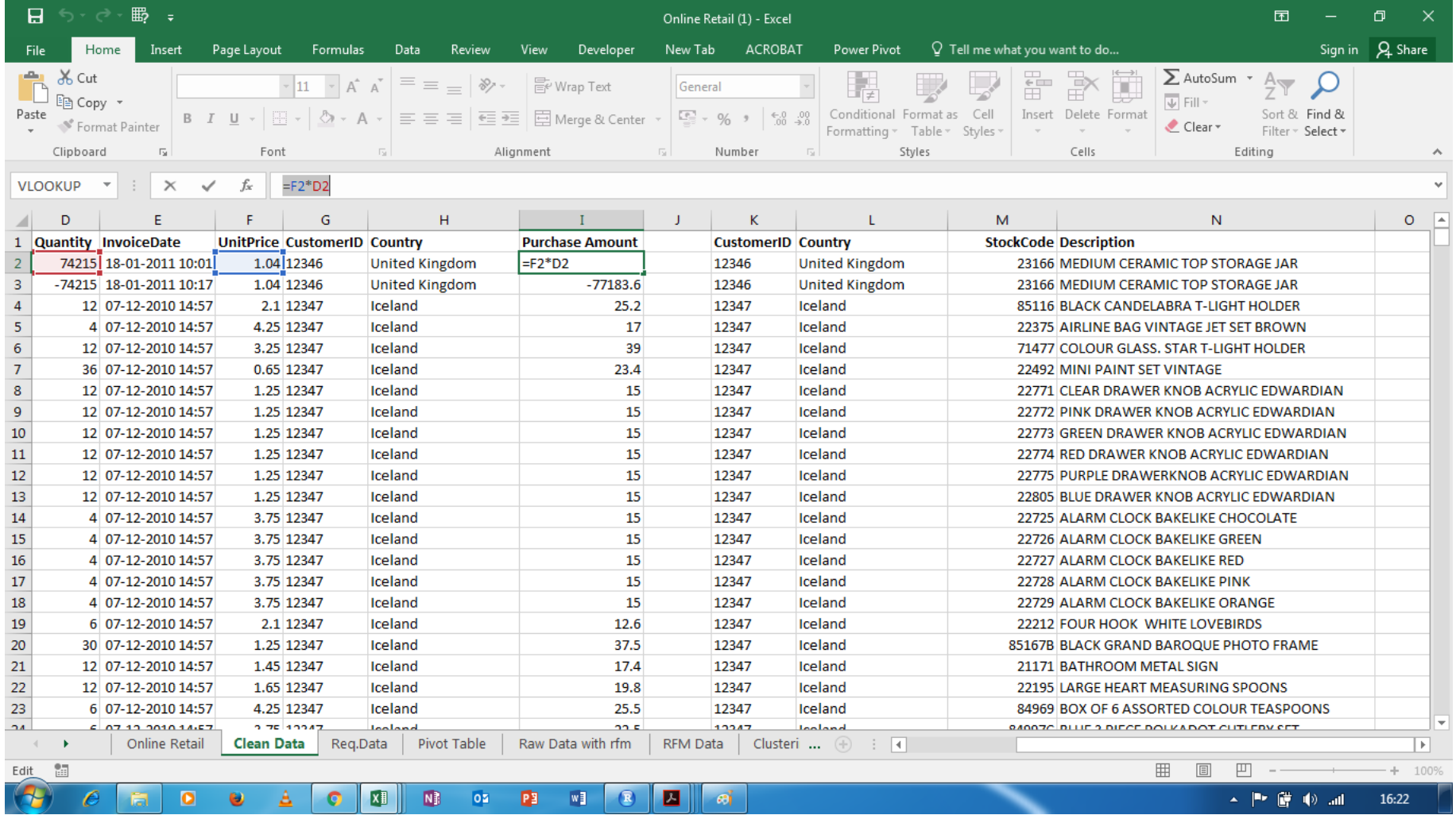
```
##   InvoiceNo StockCode                          Description Quantity
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2    536365     71053                 WHITE METAL LANTERN        6
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2
##        InvoiceDate UnitPrice CustomerID        Country
## 1 01-12-2010 08:26      2.55      17850 United Kingdom
## 2 01-12-2010 08:26      3.39      17850 United Kingdom
## 3 01-12-2010 08:26      2.75      17850 United Kingdom
## 4 01-12-2010 08:26      3.39      17850 United Kingdom
## 5 01-12-2010 08:26      3.39      17850 United Kingdom
## 6 01-12-2010 08:26      7.65      17850 United Kingdom
```

```
summary(retail$CustomerID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12350   13950   15150   15290   16790   18290  135080
```

Clean Data

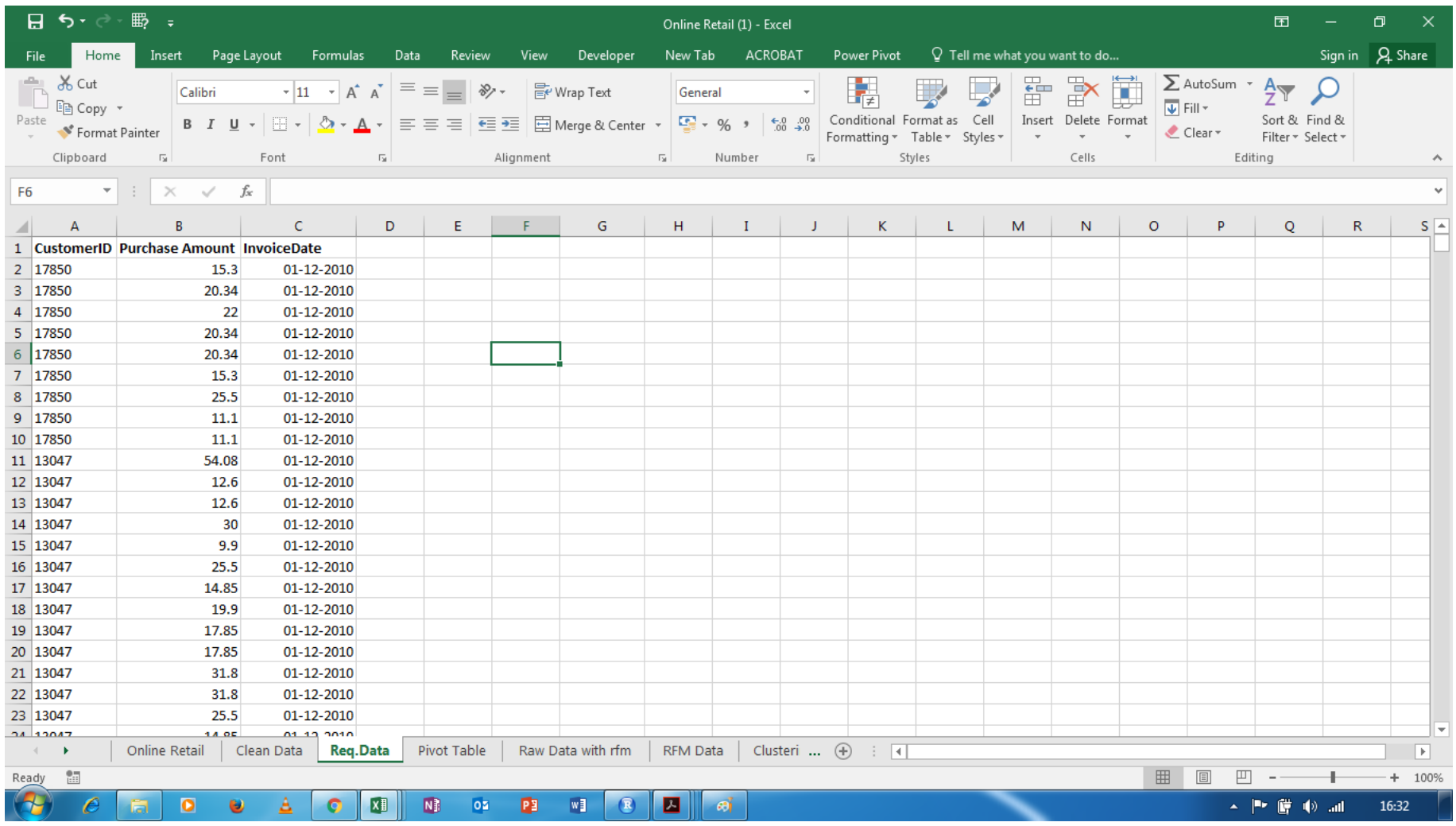
There are 135080 NA Customer ID in data.We will Remove the NA values And add Purchase Amount Colunm by Multiply UnitPrice Quantity i.e =F2 x D2



	D	E	F	G	H	I	J	K	L	M	N	O
1	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Purchase Amount		CustomerID	Country	StockCode	Description	
2	74215	18-01-2011 10:01	1.04	12346	United Kingdom	=F2*D2		12346	United Kingdom	23166	MEDIUM CERAMIC TOP STORAGE JAR	
3	-74215	18-01-2011 10:17	1.04	12346	United Kingdom	-77183.6		12346	United Kingdom	23166	MEDIUM CERAMIC TOP STORAGE JAR	
4	12	07-12-2010 14:57	2.1	12347	Iceland	25.2		12347	Iceland	85116	BLACK CANDELABRA T-LIGHT HOLDER	
5	4	07-12-2010 14:57	4.25	12347	Iceland	17		12347	Iceland	22375	AIRLINE BAG VINTAGE JET SET BROWN	
6	12	07-12-2010 14:57	3.25	12347	Iceland	39		12347	Iceland	71477	COLOUR GLASS. STAR T-LIGHT HOLDER	
7	36	07-12-2010 14:57	0.65	12347	Iceland	23.4		12347	Iceland	22492	MINI PAINT SET VINTAGE	
8	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22771	CLEAR DRAWER KNOB ACRYLIC EDWARDIAN	
9	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22772	PINK DRAWER KNOB ACRYLIC EDWARDIAN	
10	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22773	GREEN DRAWER KNOB ACRYLIC EDWARDIAN	
11	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22774	RED DRAWER KNOB ACRYLIC EDWARDIAN	
12	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22775	PURPLE DRAWERKNOB ACRYLIC EDWARDIAN	
13	12	07-12-2010 14:57	1.25	12347	Iceland	15		12347	Iceland	22805	BLUE DRAWER KNOB ACRYLIC EDWARDIAN	
14	4	07-12-2010 14:57	3.75	12347	Iceland	15		12347	Iceland	22725	ALARM CLOCK BAKELIKE CHOCOLATE	
15	4	07-12-2010 14:57	3.75	12347	Iceland	15		12347	Iceland	22726	ALARM CLOCK BAKELIKE GREEN	
16	4	07-12-2010 14:57	3.75	12347	Iceland	15		12347	Iceland	22727	ALARM CLOCK BAKELIKE RED	
17	4	07-12-2010 14:57	3.75	12347	Iceland	15		12347	Iceland	22728	ALARM CLOCK BAKELIKE PINK	
18	4	07-12-2010 14:57	3.75	12347	Iceland	15		12347	Iceland	22729	ALARM CLOCK BAKELIKE ORANGE	
19	6	07-12-2010 14:57	2.1	12347	Iceland	12.6		12347	Iceland	22212	FOUR HOOK WHITE LOVEBIRDS	
20	30	07-12-2010 14:57	1.25	12347	Iceland	37.5		12347	Iceland	85167B	BLACK GRAND BAROQUE PHOTO FRAME	
21	12	07-12-2010 14:57	1.45	12347	Iceland	17.4		12347	Iceland	21171	BATHROOM METAL SIGN	
22	12	07-12-2010 14:57	1.65	12347	Iceland	19.8		12347	Iceland	22195	LARGE HEART MEASURING SPOONS	
23	6	07-12-2010 14:57	4.25	12347	Iceland	25.5		12347	Iceland	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	

Required Data for RFM analysis

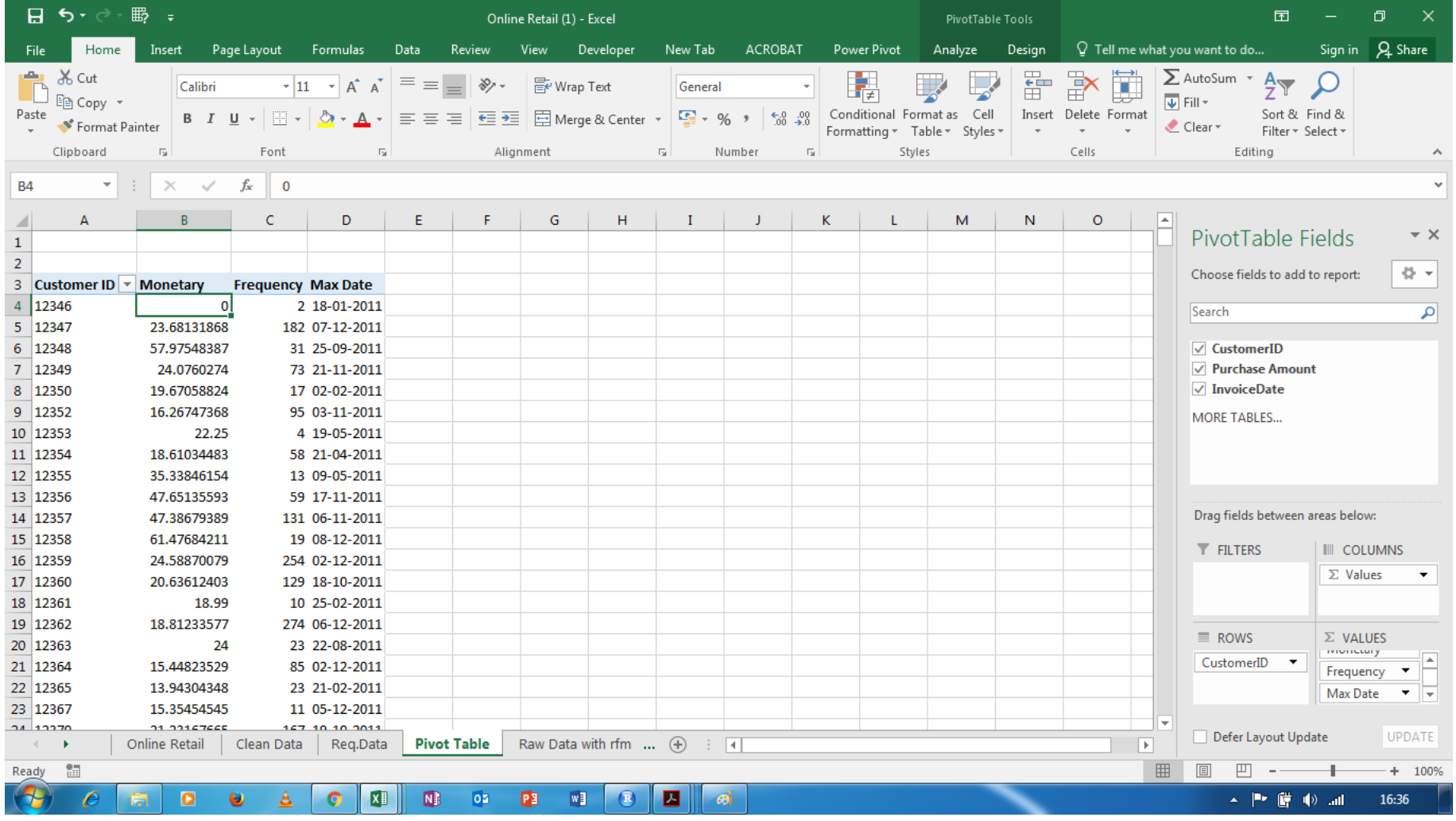
Take Customer ID , Purchase Amount and Invocie Date Column for analysis There are 406829 Customer ID



	A	B	C
1	CustomerID	Purchase Amount	InvoiceDate
2	17850	15.3	01-12-2010
3	17850	20.34	01-12-2010
4	17850	22	01-12-2010
5	17850	20.34	01-12-2010
6	17850	20.34	01-12-2010
7	17850	15.3	01-12-2010
8	17850	25.5	01-12-2010
9	17850	11.1	01-12-2010
10	17850	11.1	01-12-2010
11	13047	54.08	01-12-2010
12	13047	12.6	01-12-2010
13	13047	12.6	01-12-2010
14	13047	30	01-12-2010
15	13047	9.9	01-12-2010
16	13047	25.5	01-12-2010
17	13047	14.85	01-12-2010
18	13047	19.9	01-12-2010
19	13047	17.85	01-12-2010
20	13047	17.85	01-12-2010
21	13047	31.8	01-12-2010
22	13047	31.8	01-12-2010
23	13047	25.5	01-12-2010

By using Pivot Table remove Frequency, Monetary Value and Max Date

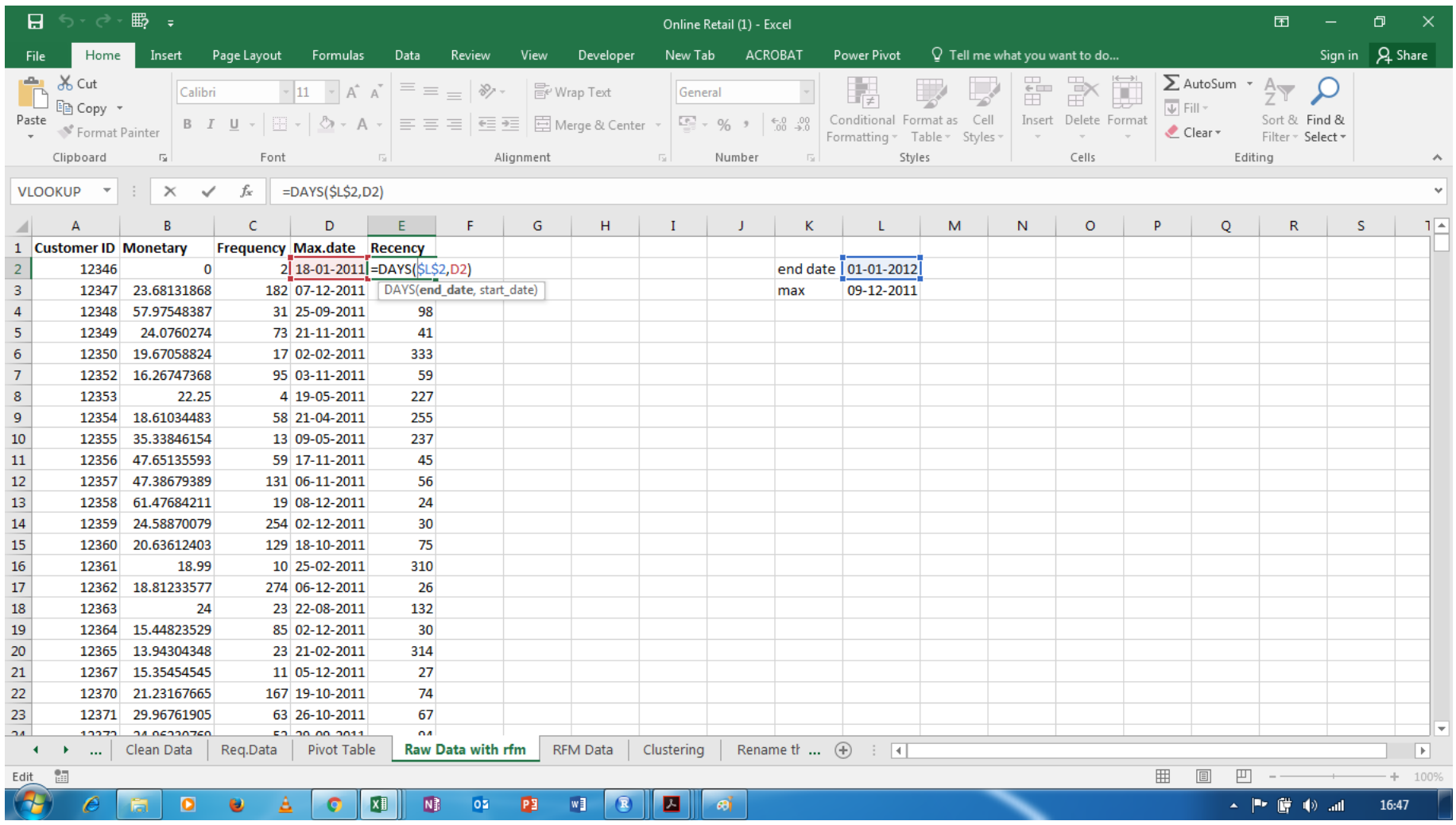
By using Pivot Table we took Customer ID in Rows and Customer ID in Values and in field value setting we took Count,We got Frequency of Customer.We done same for removing the Monetary value by taking Purchase amount in Values and in field value setting as average.For Recency we need the Max Date of a Customer so we took Invoice date in Values in field Value as Max so we got a maximum date of Customer.



Customer ID	Monetary	Frequency	Max Date
12346	0	2	18-01-2011
12347	23.68131868	182	07-12-2011
12348	57.97548387	31	25-09-2011
12349	24.0760274	73	21-11-2011
12350	19.67058824	17	02-02-2011
12352	16.26747368	95	03-11-2011
12353	22.25	4	19-05-2011
12354	18.61034483	58	21-04-2011
12355	35.33846154	13	09-05-2011
12356	47.65135593	59	17-11-2011
12357	47.38679389	131	06-11-2011
12358	61.47684211	19	08-12-2011
12359	24.58870079	254	02-12-2011
12360	20.63612403	129	18-10-2011
12361	18.99	10	25-02-2011
12362	18.81233577	274	06-12-2011
12363	24	23	22-08-2011
12364	15.44823529	85	02-12-2011
12365	13.94304348	23	21-02-2011
12367	15.35454545	11	05-12-2011

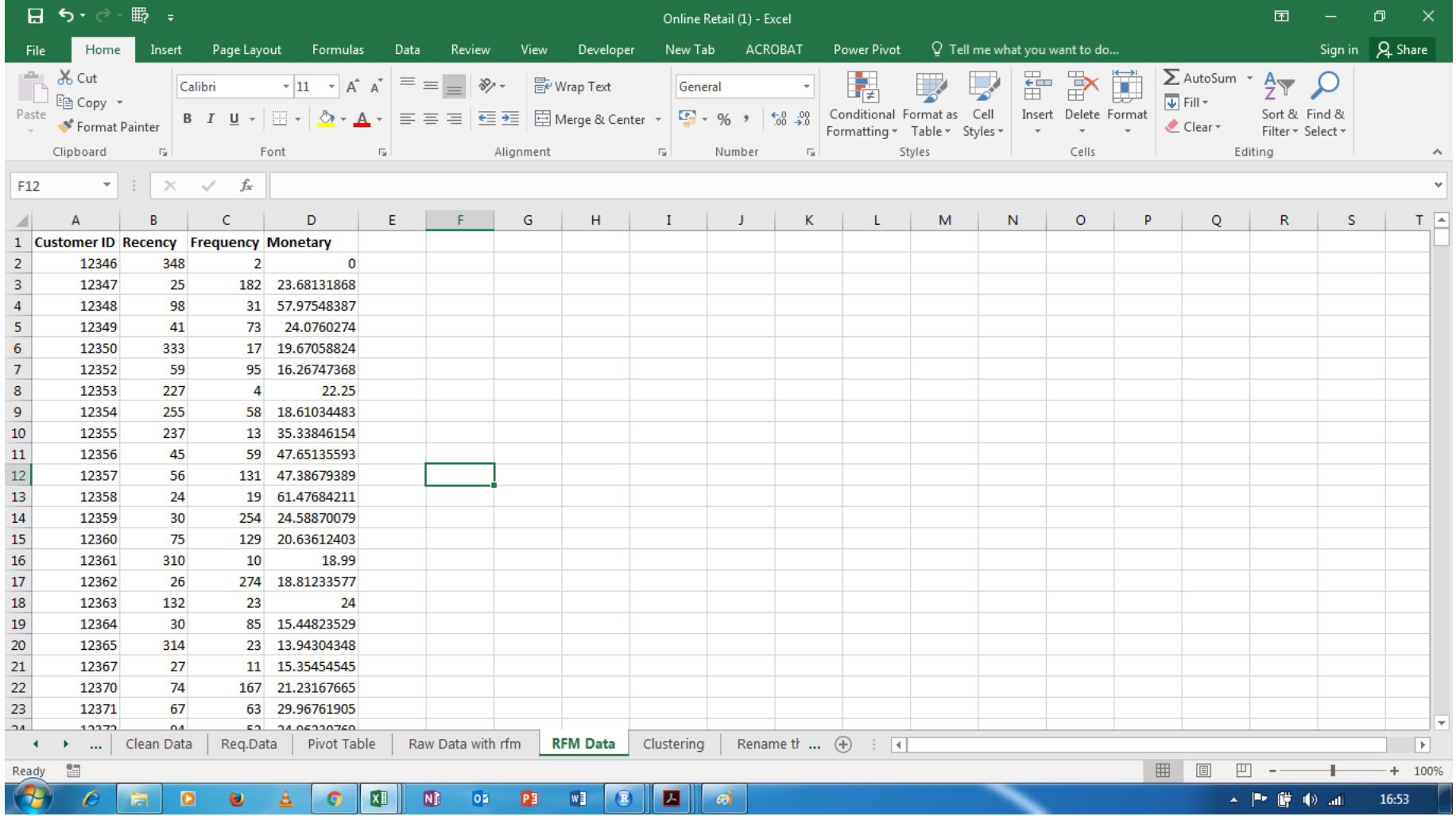
Extracted Recency From Max Date

By using Max Date Column we Extracted Recency by taking End date as 01-01-2012 Recency = Days((L2,D2)) i.e End date - Max Date



Customer ID	Monetary	Frequency	Max.date	Recency
12346	0	2	18-01-2011	=DAYS(L2,D2)
12347	23.68131868	182	07-12-2011	98
12348	57.97548387	31	25-09-2011	41
12349	24.0760274	73	21-11-2011	333
12350	19.67058824	17	02-02-2011	59
12352	16.26747368	95	03-11-2011	227
12353	22.25	4	19-05-2011	255
12354	18.61034483	58	21-04-2011	237
12355	35.33846154	13	09-05-2011	45
12356	47.65135593	59	17-11-2011	56
12357	47.38679389	131	06-11-2011	24
12358	61.47684211	19	08-12-2011	30
12359	24.58870079	254	02-12-2011	75
12360	20.63612403	129	18-10-2011	310
12361	18.99	10	25-02-2011	26
12362	18.81233577	274	06-12-2011	132
12363	24	23	22-08-2011	30
12364	15.44823529	85	02-12-2011	314
12365	13.94304348	23	21-02-2011	27
12367	15.35454545	11	05-12-2011	74
12370	21.23167665	167	19-10-2011	67
12371	29.96761905	63	26-10-2011	

We got Recency, Frequency and Monetary Value of a Customer



Customer ID	Recency	Frequency	Monetary
12346	348	2	0
12347	25	182	23.68131868
12348	98	31	57.97548387
12349	41	73	24.0760274
12350	333	17	19.67058824
12352	59	95	16.26747368
12353	227	4	22.25
12354	255	58	18.61034483
12355	237	13	35.33846154
12356	45	59	47.65135593
12357	56	131	47.38679389
12358	24	19	61.47684211
12359	30	254	24.58870079
12360	75	129	20.63612403
12361	310	10	18.99
12362	26	274	18.81233577
12363	132	23	24
12364	30	85	15.44823529
12365	314	23	13.94304348
12367	27	11	15.35454545
12370	74	167	21.23167665
12371	67	63	29.96761905

Know we need to give Score to each Recency , Frequency and Monetary Value Scoring can be done by Subjective Approach or by doing Clustering. We will do Clustering in R because of data is large

Clustering in Recency,Frequency and Monetary Value

```
Scoring <- read.csv("Retail_RFM.csv")

recency <- Scoring$Recency   ##Assign Recency from data to new Varibale recency

frequency <- Scoring$Frequency ##Assign Freq from data to new Varibale frequency

monetary <- Scoring$Monetary   ##Assign Monetary from data to new Varibale
```

We will use **Kmean algorithem** for Clustering,we will form **five cluster** for Recency ,Frequency and Monetary Value

```
recencyScoring <- kmeans(recency,centers = 5) ## 5 cluster (Recency)

freqScoring <- kmeans(frequency,centers = 5) ## 5 cluster (Frequency)

monScoring <- kmeans(monetary,centers = 5) ## 5 cluster (Monetary)
```

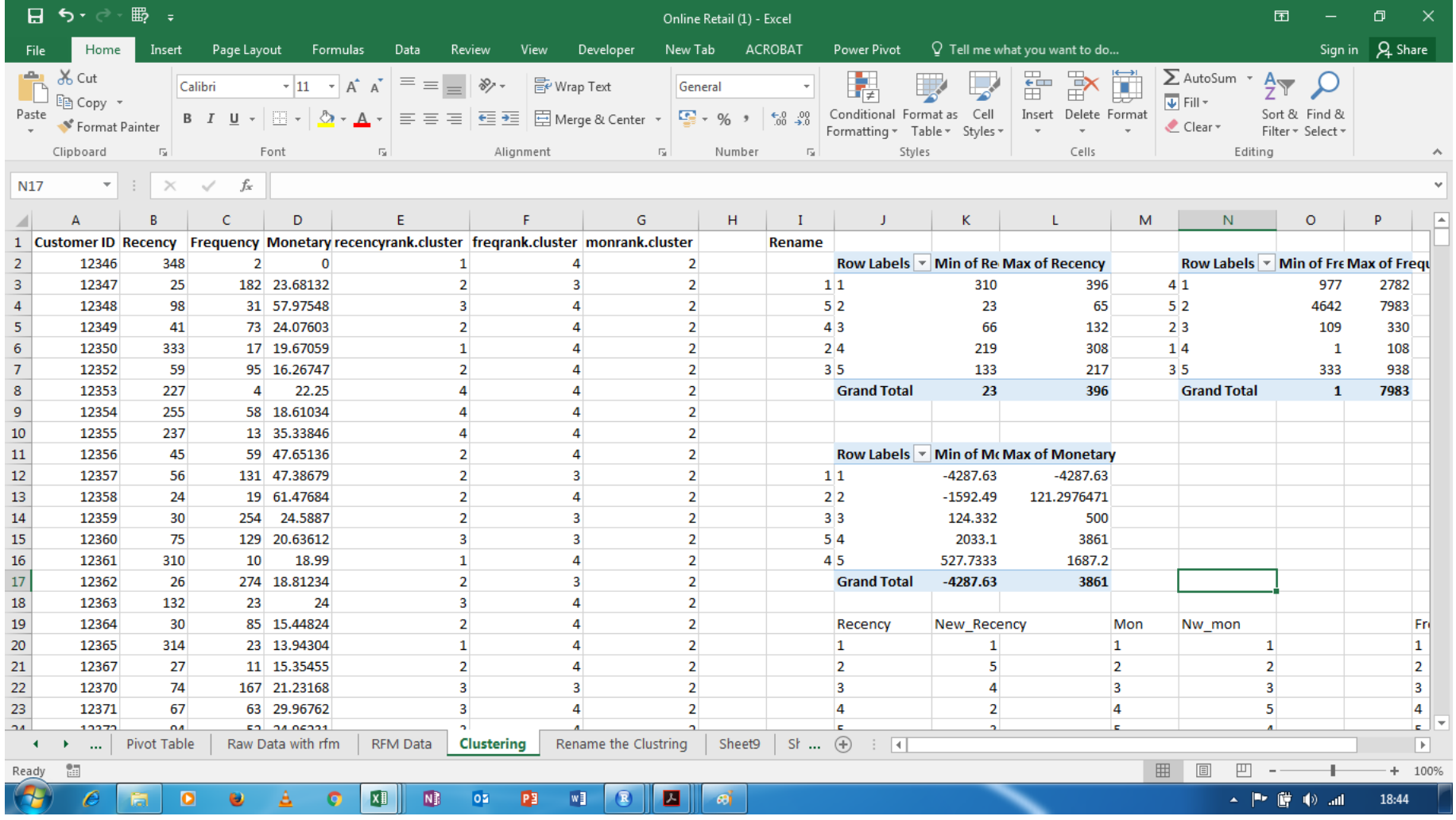
```
Rfm_Scoring <- data.frame(Scoring$Customer.ID,recencyScoring$cluster,freqScoring$cluster,monScoring$cluster)

head(Rfm_Scoring) # View the fist 6 rows of Rfm_rank dataset
```

```
##   Scoring.Customer.ID recencyScoring.cluster freqScoring.cluster
## 1               12346                      2                   2
## 2               12347                      3                   5
## 3               12348                      4                   2
## 4               12349                      3                   2
## 5               12350                      2                   2
## 6               12352                      3                   2
##   monScoring.cluster
## 1                  1
## 2                  1
## 3                  1
## 4                  1
## 5                  1
## 6                  1
```

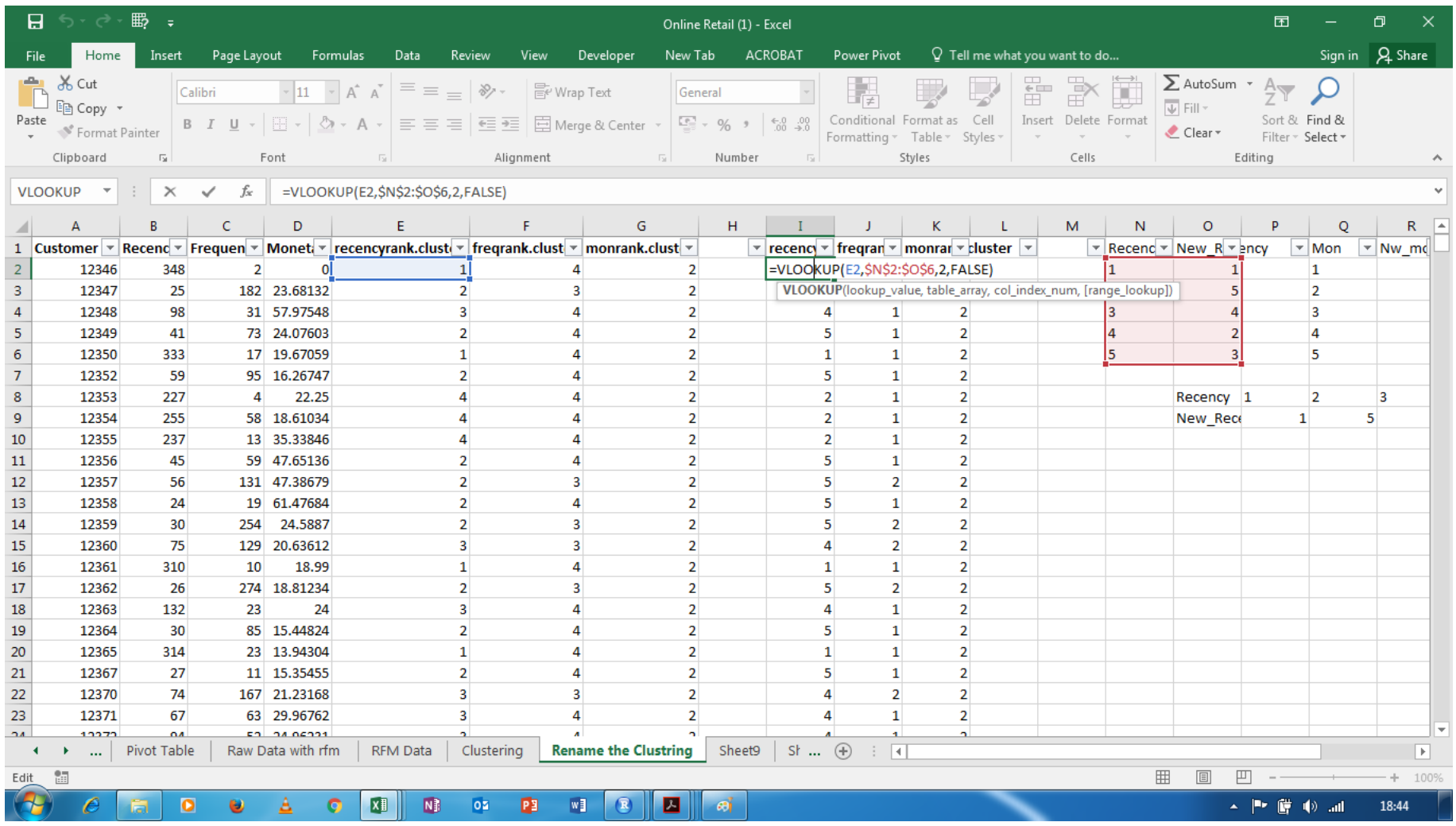
```
write.csv(Rfm_Scoring,"Scoring.csv") ##Save the Rank in csv
```

Renaming the Score (criteria)

We will Copy paste the Score into Online Retail file for renaming the cluster. It begins with Scoring customers based on recency, i.e. period since last purchase, in order of lowest to highest (most recent purchasers at the top).Cluster 2 we will rename to 5 and so on base on below pivot table. Customers Ranked for frequency – from the most to least frequent and same for Monetary value- from Highest to Lowest i.e from 5 to 1.

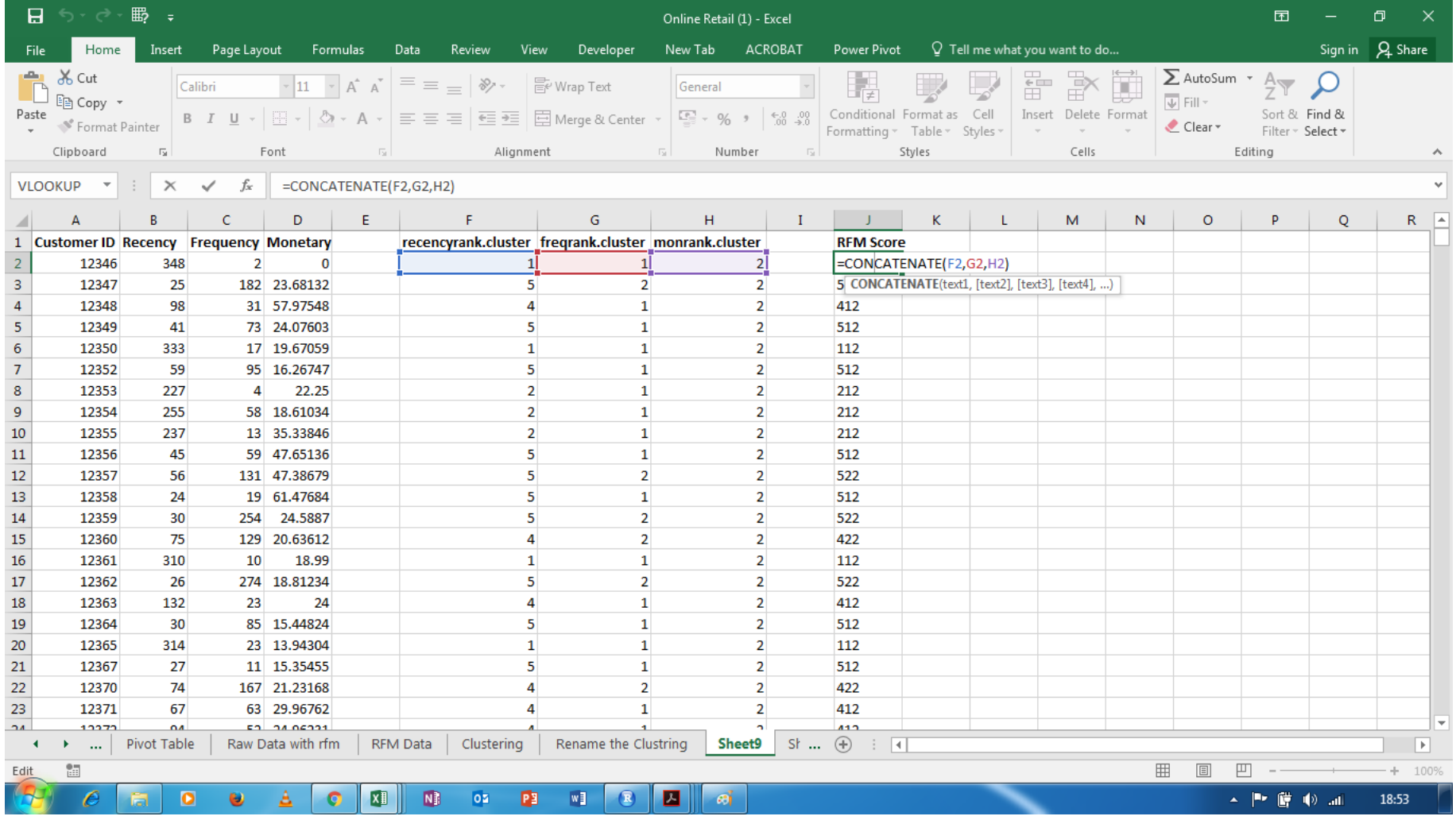


Renaming using VLook Up



RFM Score

We will Extract RFM Score by using Concenate Function from Excel. We got RFM Score for each Customer in dataset



Segmentation of Customers

Kmean Clustering

```
RFM_score <- read.csv("RFM_Score.csv")

head(RFM_score)
```

```
##   Customer.ID Recency Frequency Monetary RFM_Score
## 1       12347      25       182       24       522
## 2       12348      98        31       58       412
## 3       12349      41        73       24       512
## 4       12352      59        95       16       512
## 5       12356      45        59       48       512
## 6       12357      56       131       47       522
```

```
## For Clustering we will take only Recency,Frequency and Monetary colunm

clusterRFM <- RFM_score[c(2,3,4)]
head(clusterRFM)
```

```
##   Recency Frequency Monetary
## 1      25       182       24
## 2      98        31       58
## 3      41        73       24
## 4      59        95       16
## 5      45        59       48
## 6      56       131       47
```

```
summary(clusterRFM)
```

```
##     Recency        Frequency          Monetary
##  Min.   : 23.0   Min.   :   1.00   Min.   :-4288.00
##  1st Qu.: 39.0   1st Qu.:  17.00   1st Qu.:   11.00
##  Median : 73.0   Median :  42.00   Median :   17.00
##  Mean   :114.6   Mean   :  93.05   Mean   :   28.84
##  3rd Qu.:166.0   3rd Qu.: 102.00   3rd Qu.:   24.00
##  Max.   :396.0   Max.   :7983.00   Max.   : 3861.00
```

We need to do Scaling for the data set beacuse they are having different units

```
ScaleRFM <- scale(clusterRFM) ## Scaling

summary(ScaleRFM)
```

```
##     Recency          Frequency           Monetary
##  Min.   :-0.9088   Min.   :-0.39598   Min.   :-33.90392
##  1st Qu.:-0.7500   1st Qu.:-0.32715   1st Qu.: -0.14016
##  Median :-0.4126   Median :-0.21961   Median : -0.09303
##  Mean   : 0.0000   Mean   : 0.00000   Mean   :  0.00000
##  3rd Qu.: 0.5102   3rd Qu.: 0.03849   3rd Qu.: -0.03805
##  Max.   : 2.7926   Max.   :33.93940   Max.   : 30.09723
```

Now After Scaling we will take Distances

```
RFM_dist <- dist(ScaleRFM) ## Distances

kmean_result <- kmeans(RFM_dist,centers = 3)

o <- order(kmean_result$cluster)

KfinalResult <- data.frame(RFM_score$Customer.ID,RFM_score$RFM_Score,kmean_result$cluster)

head(KfinalResult)
```

```
##   RFM_score.Customer.ID RFM_score.RFM_Score kmean_result.cluster
## 1                 12347                 522                    1
## 2                 12348                 412                    1
## 3                 12349                 512                    1
## 4                 12352                 512                    1
## 5                 12356                 512                    1
## 6                 12357                 522                    1
```

```
tail(KfinalResult)
```

```
##      RFM_score.Customer.ID RFM_score.RFM_Score kmean_result.cluster
## 4367                 18280                 212                    2
## 4368                 18281                 312                    1
## 4369                 12748                 552                    3
## 4370                 14096                 552                    3
## 4371                 14911                 552                    3
## 4372                 17841                 552                    3
```

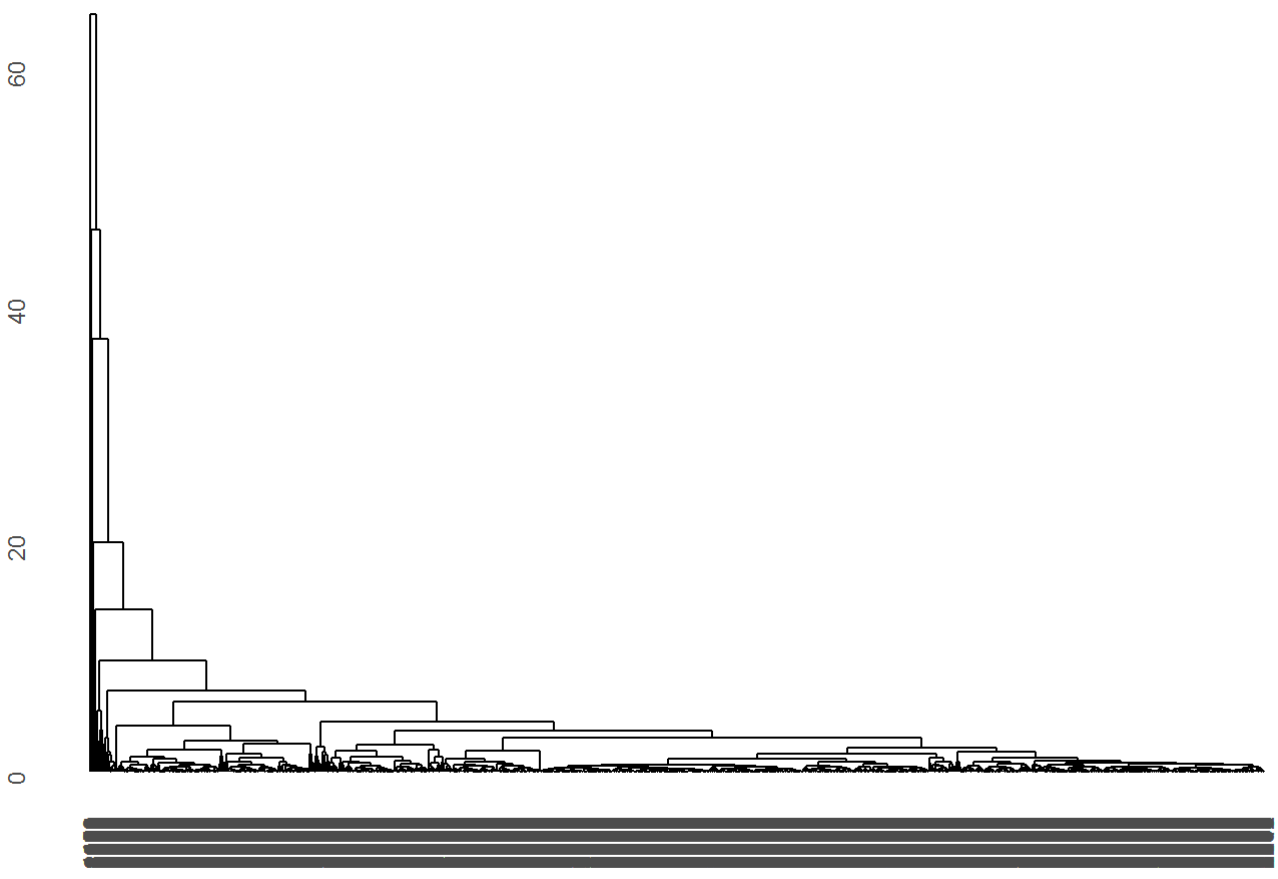
```
write.csv(KfinalResult,"KfinalResult.csv")
```

Hierarchical Clustering

```
hclust_result<- hclust(RFM_dist,method = "complete")

library(ggdendro)
```

```
## Warning: package 'ggdendro' was built under R version 3.2.5
```

```
ggdendrogram(hclust_result) ## dendrogram
```



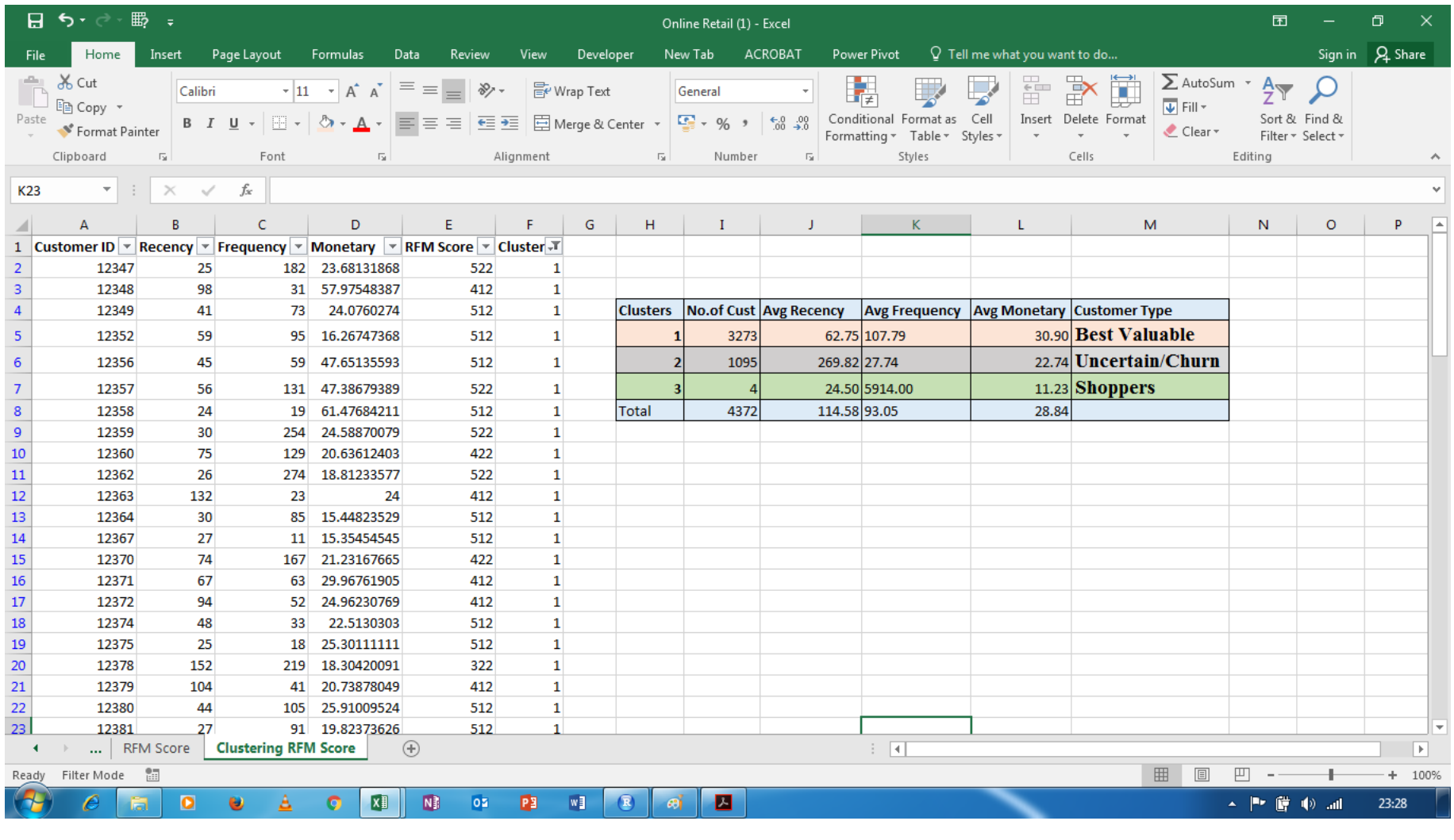
```
group <- cutree(hclust_result,k=3)

hclustResult <- data.frame(RFM_score$Customer.ID,RFM_score$RFM_Score,group)

head(hclustResult)
```

```
##   RFM_score.Customer.ID RFM_score.RFM_Score group
## 1                 12347                 522     1
## 2                 12348                 412     1
## 3                 12349                 512     1
## 4                 12352                 512     1
## 5                 12356                 512     1
## 6                 12357                 522     1
```

```
tail(hclustResult)
```

```
##      RFM_score.Customer.ID RFM_score.RFM_Score group
## 4367                 18280                 212     1
## 4368                 18281                 312     1
## 4369                 12748                 552     1
## 4370                 14096                 552     1
## 4371                 14911                 552     1
## 4372                 17841                 552     1
```

Both Kmean and Hierarchical Clustering have given same clusters We will take Kmean Result for Segmentation of Customer



Customer ID	Recency	Frequency	Monetary	RFM Score	Cluster
12347	25	182	23.68131868	522	1
12348	98	31	57.97548387	412	1
12349	41	73	24.0760274	512	1
12352	59	95	16.26747368	512	1
12356	45	59	47.65135593	512	1
12357	56	131	47.38679389	522	1
12358	24	19	61.47684211	512	1
12359	30	254	24.58870079	522	1
12360	75	129	20.63612403	422	1
12362	26	274	18.81233577	522	1
12363	132	23	24	412	1
12364	30	85	15.44823529	512	1
12367	27	11	15.35454545	512	1
12370	74	167	21.23167665	422	1
12371	67	63	29.96761905	412	1
12372	94	52	24.96230769	412	1
12374	48	33	22.5130303	512	1
12375	25	18	25.30111111	512	1
12378	152	219	18.30420091	322	1
12379	104	41	20.73878049	412	1
12380	44	105	25.91009524	512	1
12381	27	91	19.82373626	512	1

Clusters	No.of Cust	Avg Recency	Avg Frequency	Avg Monetary	Customer Type
1	3273	62.75	107.79	30.90	Best Valuable
2	1095	269.82	27.74	22.74	Uncertain/Churn
3	4	24.50	5914.00	11.23	Shoppers
Total	4372	114.58	93.05	28.84	

In **Customers segment 1** have most best valuable customers, because its consists of customer who have regularly purchased and have high purchase frequency and purchase amount and number of customers are also more.

In **Customers segment 2** have least likly to buy customer because they have purchase very long ago with very less Frequency and Monetary value.we named them as Uncertain or Churn customer type.

In **Customers segment 3** have very less customers who have done more purchase and are regularly purchasing with very less purchase amount.for this type of customer we have named them as Shoppers who purchase regularly with less amount.

Next

2)Customer behavior prediction: We can predict RFM Score of customers base on demographic varibales(Country)

3)Product Recommedation: We can recommed product(i.e Stock) on the basis of purchase behavior and extract frequent product purchase in particular segment with particular RFM Score.

Reference

1)Derya Birant (2011). Data Mining Using RFM Analysis, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.)

2)Segmentation and Lifetime Value Models Using SAS,Edward C.Malthouse