Humital Mini Project

Online Retail Customer Segmentation

Customer Segmentation

In this exercise, your aim is help an Online Retail organization help segment its customers so as they can understand their behaviors & identify approaches to target them better

What is customer segmentation

Bain defines Customer segmentation as follows:

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics.[1]

… Customer Segmentation can be a powerful means to identify unmet customer needs. Companies that identify underserved segments can then outperform the competition by developing uniquely appealing products and services. Customer Segmentation is most effective when a company tailors offerings to segments that are the most profitable and serves them with distinct competitive advantages.

1 http://www.bain.com/publications/articles/management-tools-customer-segmentation.aspx

Why is it important

- Cost & effort optimization: Lower marketing cost per prospect
- Better targeting & offering design
- Lower Spam
- Increased Customer Retention

How to segment customers

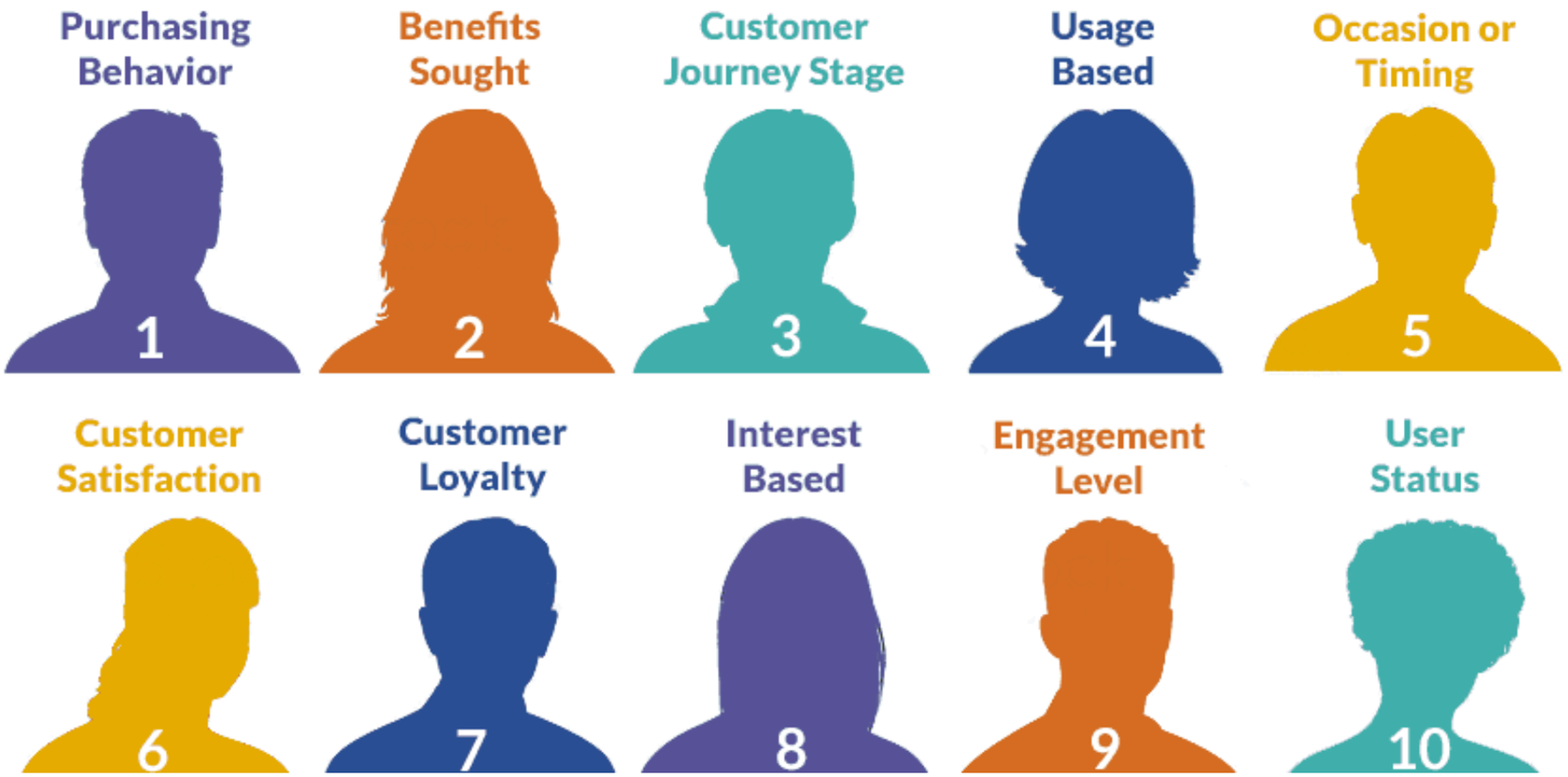
Types of Segmentation

Geographic	Demographic	Behavioral	Psychographic
• Examples: 10583, Chicago, the US, or Garth Road.	• Male, 45, never married, BS college degree	• Previous survey takers who purchase online Product X, monthly.	• Religious soccer moms focused on PTA or school activities.

The client has shared the behavioural data of his customers

http://www.thinkaroundcorners.com/2016/05/ama-2016-market-segmentation-critical-discovering-high-value-sources-growth/

Behavioral segmentation



https://www.pointillist.com/blog/behavioral-segmentation/

Our Data

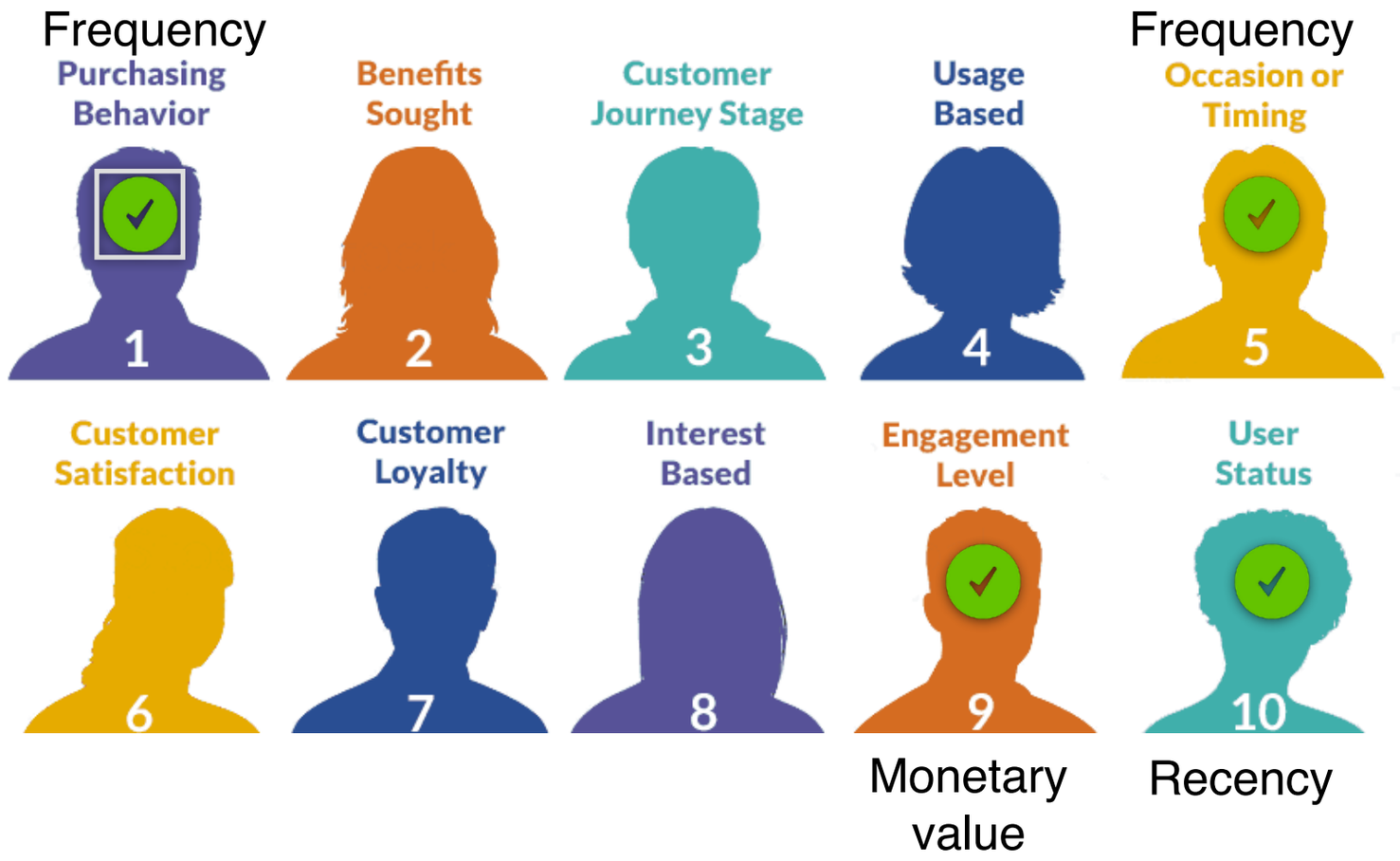
Take a pause and look at the data and see what all types of segmentation are possible

ID	Name
InvoiceNo	Invoice number
StockCode	Product (item) code
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction
InvoiceDate	Invoice Date and time
UnitPrice	Unit price
CustomerID	Customer number
Country	Country name

Our Data

Take a pause and look at the data and see what all types of segmentation are possible

ID	Name
InvoiceNo	Invoice number
StockCode	Product (item) code
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction
InvoiceDate	Invoice Date and time
UnitPrice	Unit price
CustomerID	Customer number
Country	Country name



RFM model for Customer Value

RFM stands for the three dimensions:

- **R**ecency – How recently did the customer purchase?
- **F**requency – How often do they purchase?
- **M**onetary Value – How much do they spend?

https://en.wikipedia.org/wiki/RFM_(customer_value)

RFM model for Customer Value

RFM stands for the three dimensions:

- **R**ecency – How recently did the customer purchase?
- **F**requency – How often do they purchase?
- **M**onetary Value – How much do they spend?

ID	Name
InvoiceNo	Invoice number
StockCode	Product (item) code
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction
InvoiceDate	Invoice Date and time
UnitPrice	Unit price
CustomerID	Customer number
Country	Country name

How do you think this can be applied to our data?

https://en.wikipedia.org/wiki/RFM_(customer_value)

RFM model for Customer Value

ID	Name
InvoiceNo	Invoice number
StockCode	Product (item) code
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction
InvoiceDate	Invoice Date and time
UnitPrice	Unit price
CustomerID	Customer number
Country	Country name

Frequency: Total number of unique invoices for each customer for the period

Monetary value: Unit price * quantity

Recency: Most recent invoice date for the customer

https://en.wikipedia.org/wiki/RFM_(customer_value)

RFM model for Customer Value

Now, based on this lets start to identify the clusters based on RFM for each customer in R

Data Preparation

- Consider only one year of Data
- For customers from UK only
- Remove data with no customer IDs
- Don’t forget to identify the customer returns in the dataset and consider them appropriately

Step 1: Data cleanup

1. Load the database
2. Remove the invoices with blank customer_ID
3. Convert invoice_date to date format
4. Select the one year range data (i.e 9-Dec-2010 to 9-Dec-2011)
5. Select the data for customers from UK only
6. Identify returns & mark them separately by a column

Step 2a: RFM Calculation - Recency

1. Create customer level dataset
2. Calculate recency for each invoice with difference from 2011-12-11
3. Exclude returns only consider data for most recent purchase
4. Merge recency to customer dataset

Step 2b: RFM Calculation - Frequency

1. Consider only unique invoices per customer (there are multiple rows for each invoice)
2. Calculate the frequency of orders for each customer
3. Merge frequency to customer dataset
4. Remove customers who have not made any purchases in the past year

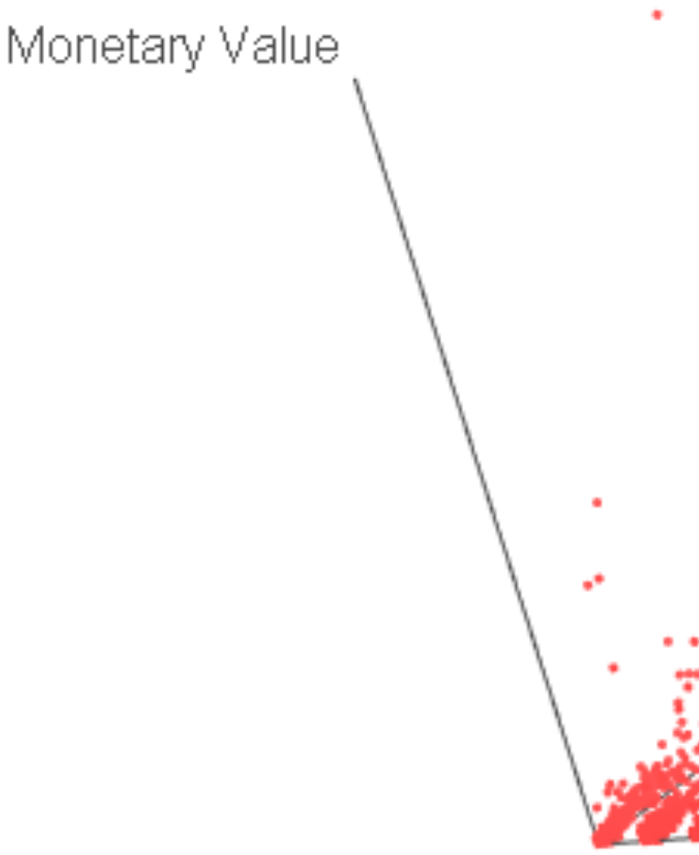
Step 2c: RFM – Monetary Value

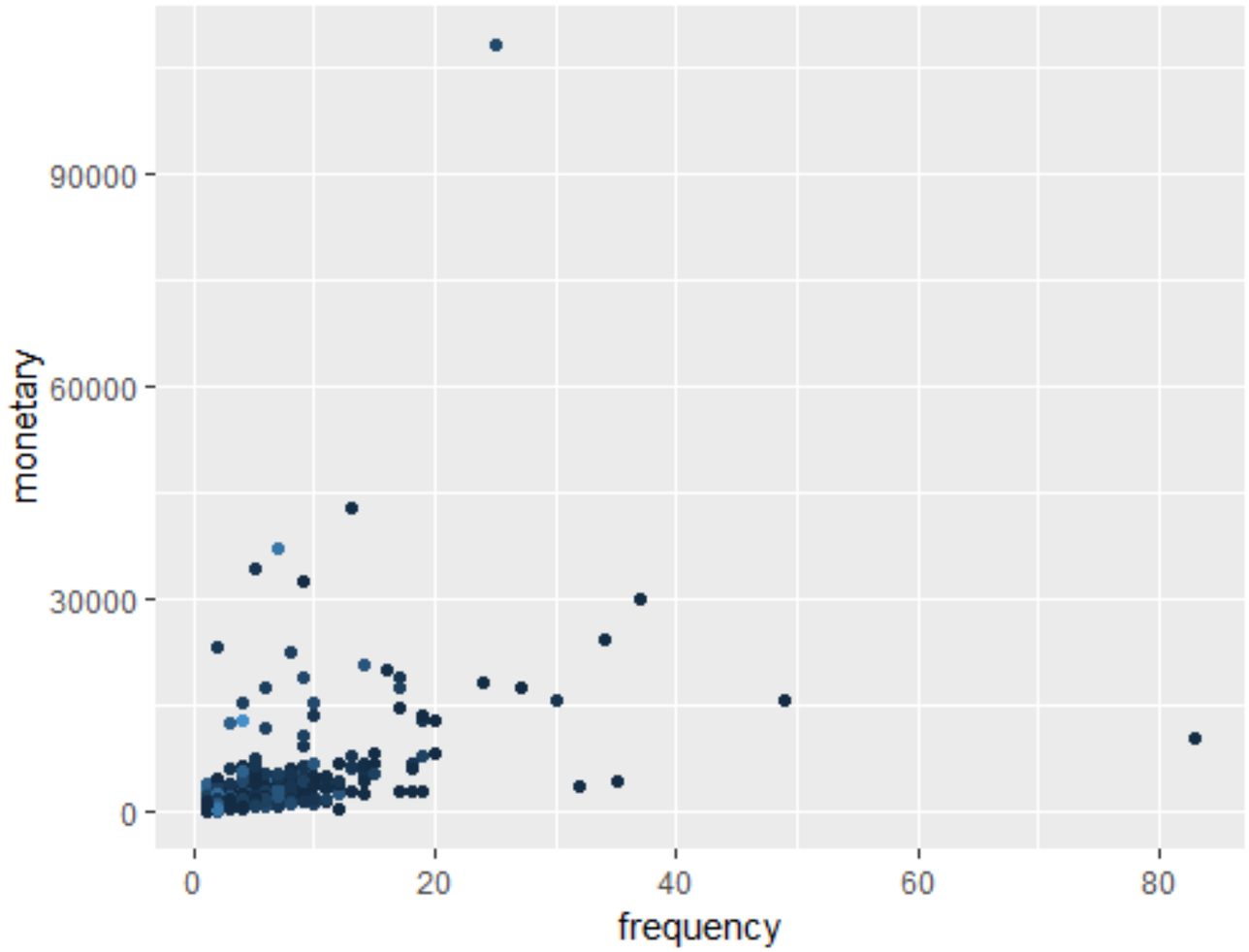
1. Calculate the total invoice value for each invoice and add to the dataset
2. Aggregate total invoice value for each customer
3. Merge monetary value to customer dataset
4. Set monetary value to 0 for customers with –ve value (returns from prior period)

Step 3: Plotting the data

1. Lets plot the data in 3D (use scatter3D)
2. Lets plot the data in 2D (use scatter)

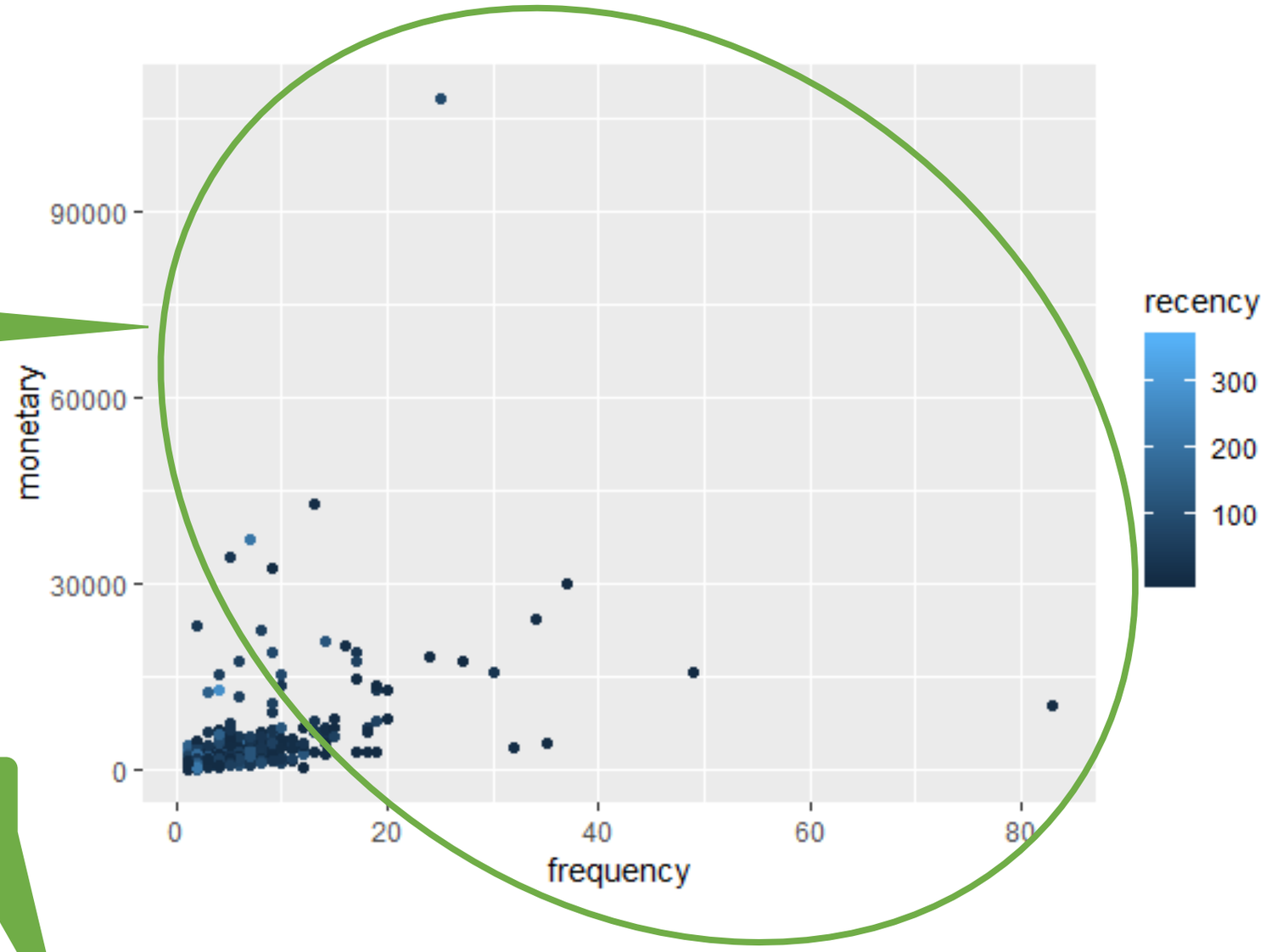
Can you determine the clusters?

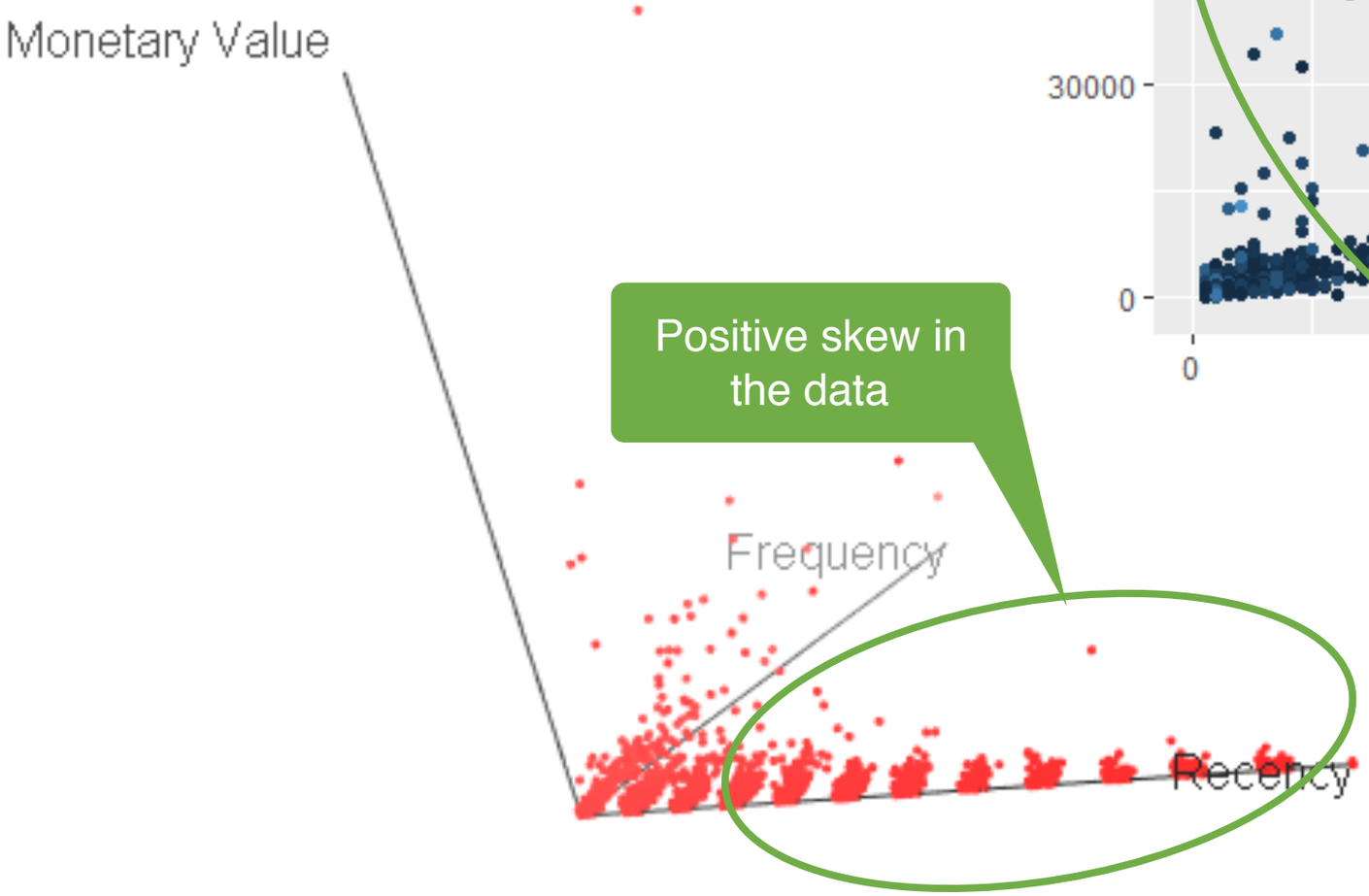




Observations

The data has a lot of spread & outliers





Positive skew in the data

Treatment: We need to transform the data to reduce spread, since KNN is very sensitive to outliers and spread

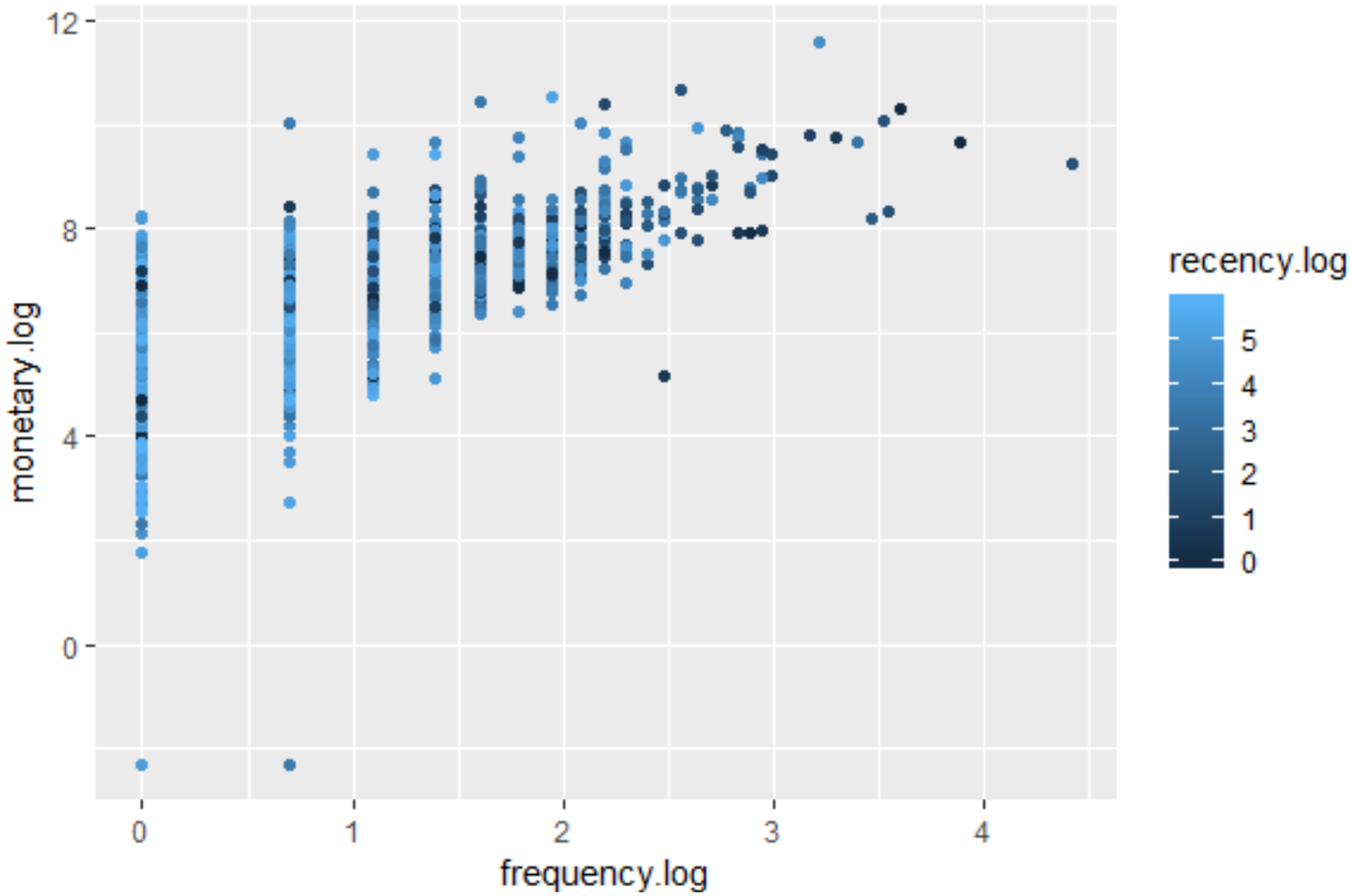
Step 4: Data Transformation

1. To reduce the impact of positive skew, lets take the log of the reach variable

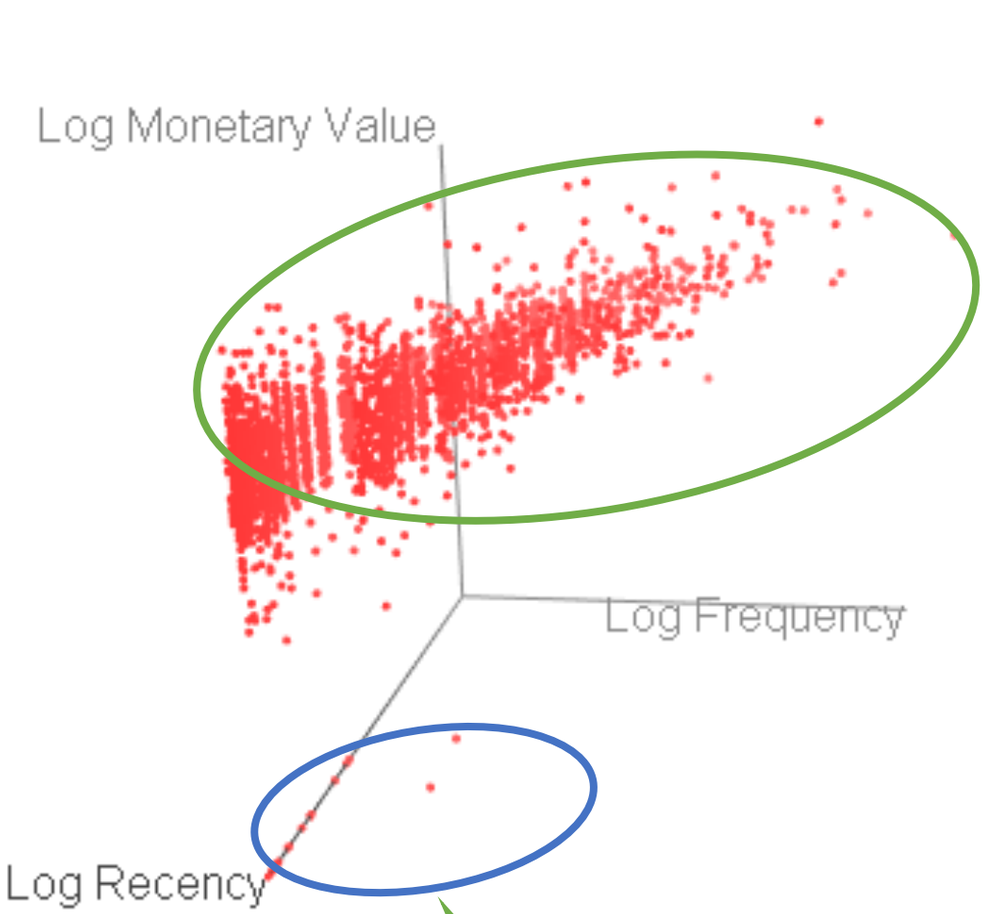
Step 5: Lets plot the transformed data

1. Lets plot the data in 3D (use scatter3D)
2. Lets plot the data in 2D (use scatter)

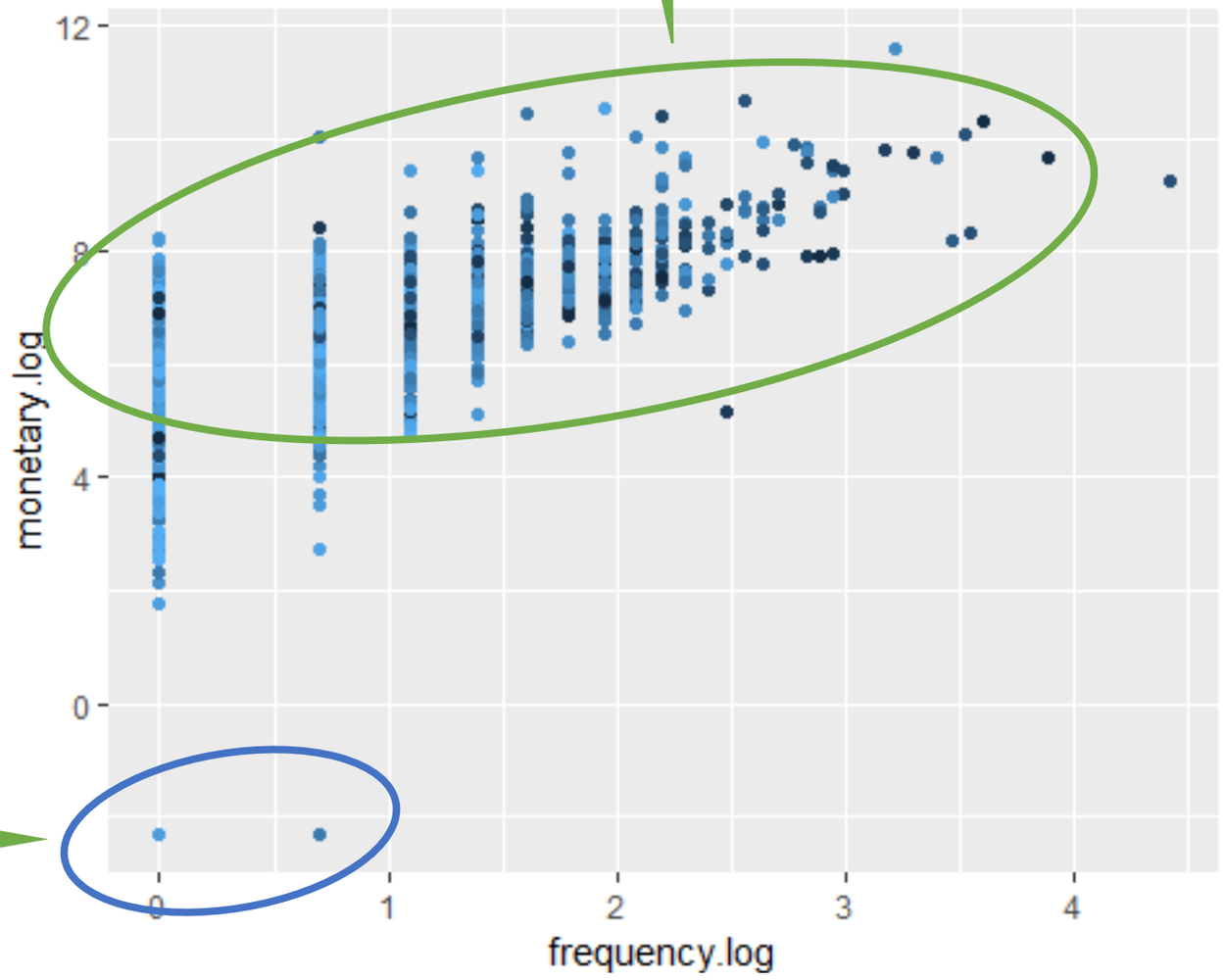
Can you determine the clusters?



Can you determine the clusters?



A continuous spread is observed



We find these small clusters

We observe that manually identifying clusters is really difficult, now let us help KNN find us clusters

Step 6: Cluster determination

Use KNN to determine the clusters with k = 3

1. Preprocess the data (store the data in a separate frame using which clustering will be done – log of R,F,M)

2. Define K & run Kmeans to identify clusters

3. Add the clusters to the customers dataset

4. Extract cluster information of each variable - recency, frequency & monetary value

Step 7: Plot the cluster output

1. Use scatter3d to plot the clusters

Step 8: Different values for k

- Run Step 6 to 7 for different values for k
- K = 2 to 6 & note down the results for cluster information & scatter plot for each

Step 9: Determine Optimum cluster k

1. Use nbclust to determine the optimum cluster

Step 10: Interpret results for cluster k

Interpret the results for each cluster k

1. Determine which cluster k is interpretable
2. Define the logic of each cluster in the chosen k
3. Publish your results (information & scatter plot)

Extended version

Hierarchical clustering

Solving using Hierarchical Clustering

1. (After Step-5) Preprocess the data (store the data in a separate frame using which clustering will be done – log of R,F,M)

2. Use hclust to identify hierarchical clusters with averaging method

3. Lets distribute the data in five clusters

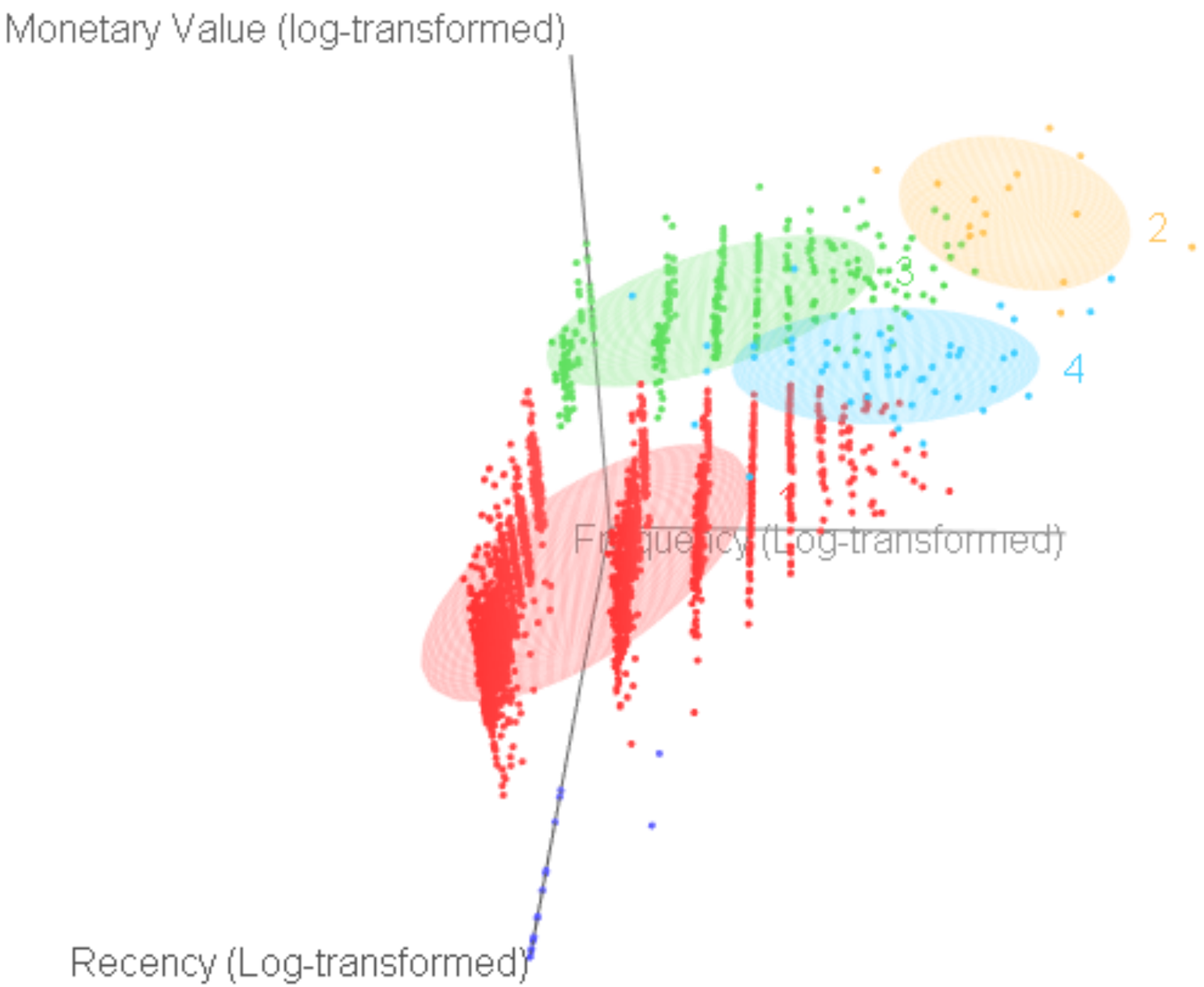
4. Merge cluster_groups to customers dataset

Solving using Hierarchical Clustering

5. Extract cluster information of each variable - recency, frequency & monetary value

6. Now lets plot the clusters using scatter3D

Solving using Hierarchical Clustering



DONE!