Introduction

A combination of **RFM Segmentation**, **Cohort** and **Survival Analyses** is a set of robust and effective statistical techniques for customer analytics. It can be used by businesses for direct marketing, site selection, customer relationship management and prediction of customer behavior.

Instead of analyzing the entire customer base as a whole, it's better to segment them into homogenous groups, understand the traits of each group, and engage them with relevant campaigns rather than segmenting on just customer age or geography.

Data Collection

Online Retail.xlsx dataset from kaggle.com

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

RFM Analysis

RFM stands for **Recency**, **Frequency**, and **Monetary value**, each corresponding to some key customer trait. RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a **customer's lifetime value**, and recency affects **retention**, a measure of engagement.

- **Recency** (R) – refers to the interval between the time when the latest purchase happens and the present time.
- **Frequency** (F) – number of transaction that a customer made within a certain period

- **Monetary** (M) – refers to the total money spent by a customer

RFM factors illustrate these facts-

- The more recent the purchase, the more responsive the customer is
- The more frequently the customer buys, the more engaged and satisfied they are
- Monetary value differentiates heavy spenders from low-spenders

RFM Analysis Importance

1. Best customers
2. Churn rate, the customers who are dropping out
3. The most valuable customers
4. Who are most likely to respond to engagement campaigns

Process of calculating percentiles (RFM method)

- Sort customers
- Break customers into pre-defined number of groups of equal size
- Assign a label to each group

CustomerID	Recency	Frequency	MonetaryValue
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

Building RFM segments and Score

Using pandas's qcut method we can label the each attribute (Recency, Frequency, MonetaryValue) and calculate RFM score with summing the all attributes.

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_segment	RFM_score
12346.0	326	1	77183.60	1	1	4	114	6
12347.0	2	182	4310.00	4	4	4	444	12
12348.0	75	31	1797.24	2	2	4	224	8
12349.0	19	73	1757.55	3	3	4	334	10
12350.0	310	17	334.40	1	1	2	112	4

Note: it is always the best practice to investigate the size of the segments before you use them for business application.

Use RFM score to group into Gold, Silver and Bronze segments

Gold label: score 10-12

Silver label: score 6-9

Bronze label: score 1-5

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
General_segment				
Bronze	192.2	15.1	266.5	1287
Gold	20.1	225.6	5246.8	1263
Silver	72.0	49.4	1072.4	1788

Cohort Analysis

A cohort is a group of subjects who share a defining characteristics. Cohorts are used in medicine, psychology, econometrics, ecology and many other areas to perform a cross-section at intervals through time.

Types of Cohorts:

- **Time Cohorts** are customers who signed up for a product or service during particular time frame. Analyzing these cohorts shows the customers' behavior depending on the time they started using the company's products or services. The time may be monthly or quarterly even daily.
- **Behavior Cohorts** are customers who purchased a product or subscribed to a service in the past. It groups customers by the type of product or service they signed up. Understanding the needs of various cohorts can help a company design custom-made services or products for particular segments.
- **Size Cohorts** refer to the various sizes of customers who purchase company's products or services. This categorization can be based on the amount of spending in some periodic time after acquisition or the product type that the customer spent most of their order amount in some period of time.

Importance of Cohort Analysis

- Retention rate – For a subscription-based business, it is imperative to be tracking retention of customers, or when they decide to cancel service.
- Repeat rate – For an eCommerce business, repeat rate measures who frequently someone is coming back to buy.
- Lifetime Value(LTV) – Using your retention rate and multiplying it by monthly subscription rate, gives you your customer life time value.
- Customer Acquisition Cost(CAC) – Customer acquisition cost is the amount of money a business spends to acquire a customer.

Cohort Visualization

Monthly customer retention/active customers:

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	323.0	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

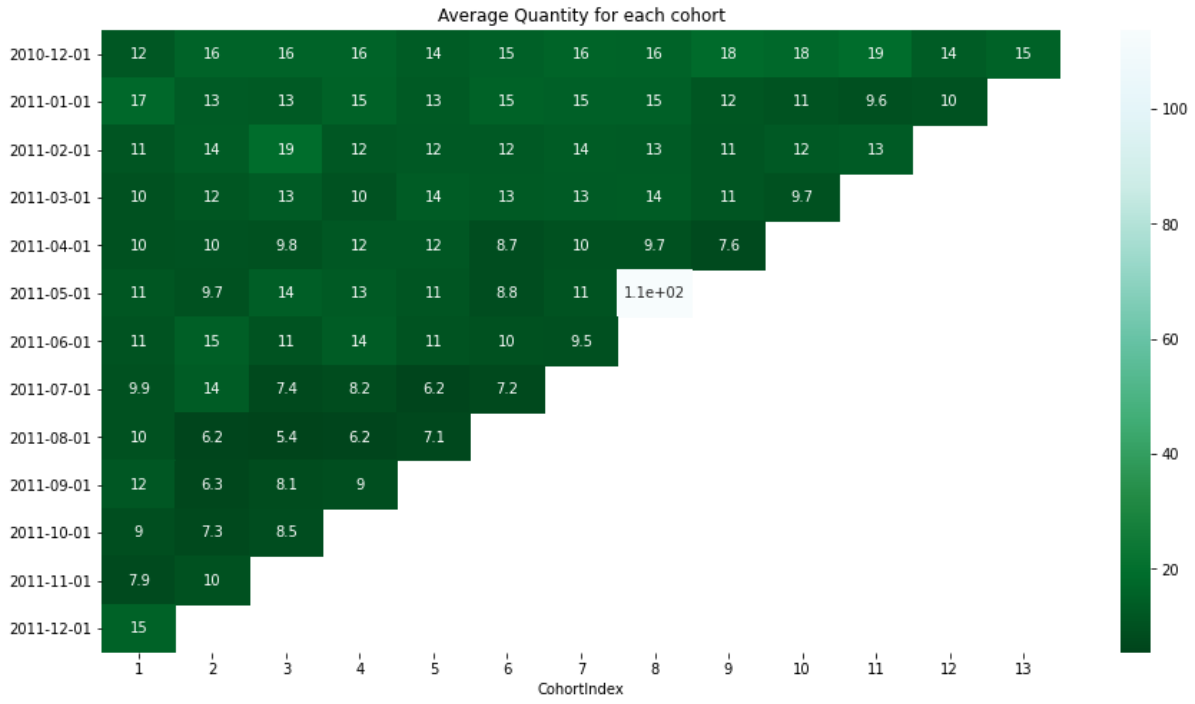
Retention Rate Table/active customer percentages:

It describes the percentages of active customer monthly. We can easily find out which groups of customers are going to cancel the service and who are the most active customers over certain period of time.

It's kind of picturization of overall business which helps to take future steps along with better understanding the customers and theirs' behaviors.

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

Heatmap for Average Product Quantity:



K-Means clustering

A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.

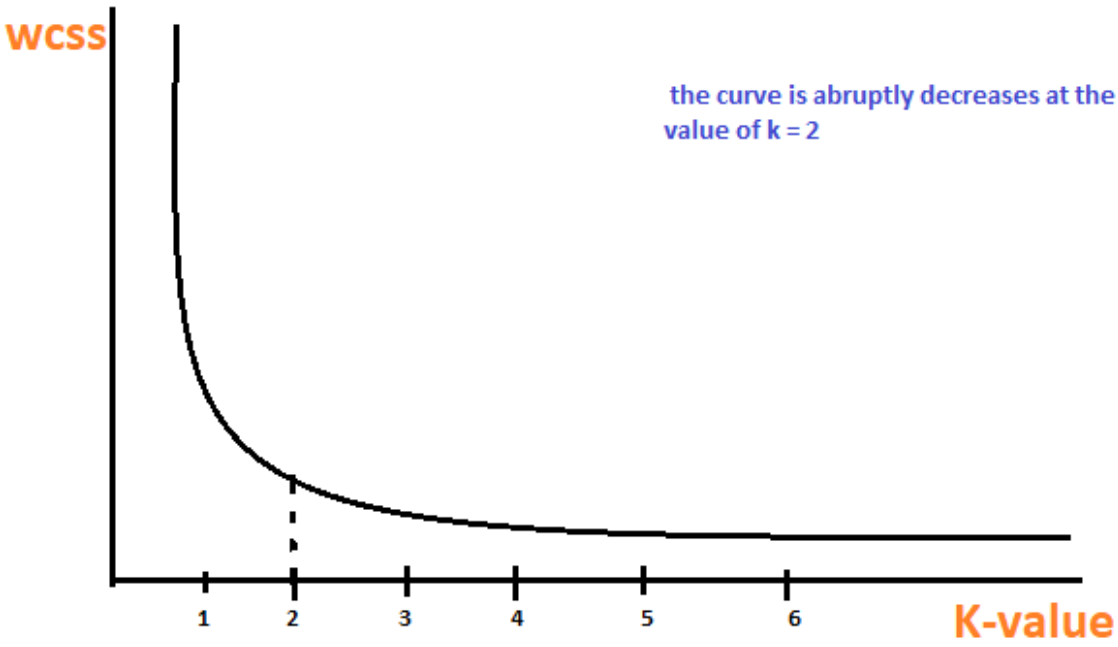
The way k-means algorithm works is as follows:

1. Specify the number of clusters ***K***
2. Initialize centroids by first shuffling the dataset and then randomly selecting ***K*** data points for the centroids.
3. Keep iterating until there is no change to centroids, assignment of data points to clusters isn't changing.

- Compute the sum of the squared between data points and all centroids.
- Assign each data point to the closest centroid(cluster)
- Compute the centroids for the cluster by averaging the all data points that belong to that cluster

Choose the value of K

<u>Elbow method:</u>



Data Pre-processing

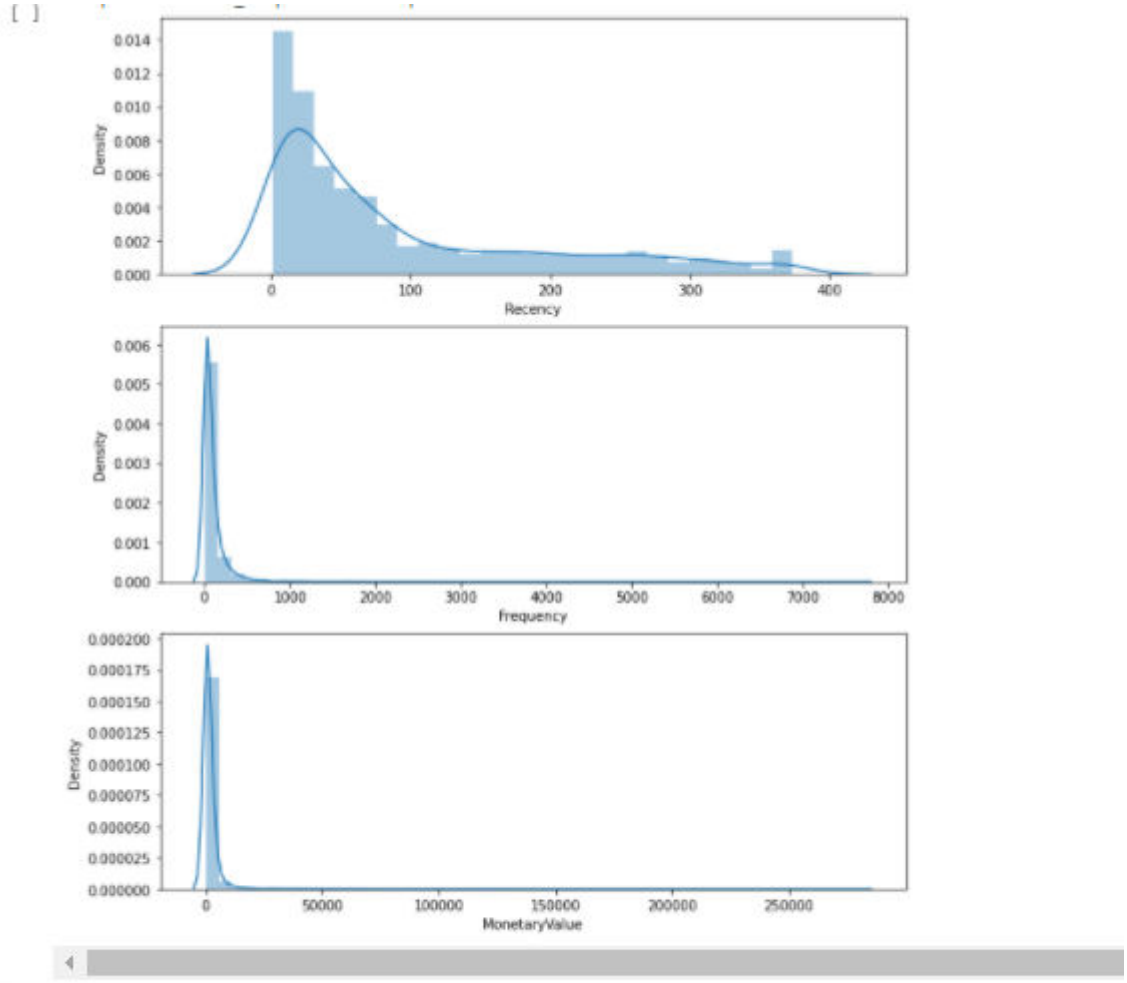
Some key k-means assumptions must be checked before step into implementing the k-means clustering mode.

- Symmetric distribution of variables (not skewed)
- Variables with same average values
- Variables with same variance

	Recency	Frequency	MonetaryValue
count	4338.000000	4338.000000	4338.000000
mean	92.536422	90.523744	2048.688081
std	100.014169	225.506968	8985.230220
min	1.000000	1.000000	3.750000
25%	18.000000	17.000000	306.482500
50%	51.000000	41.000000	668.570000
75%	142.000000	98.000000	1660.597500
max	374.000000	7676.000000	280206.020000

For this table, mean and variance are not equal

solution: Scaling variables by using scaler from scikit-library



another problem: distribution of variables are not symmetric. it's skewed data

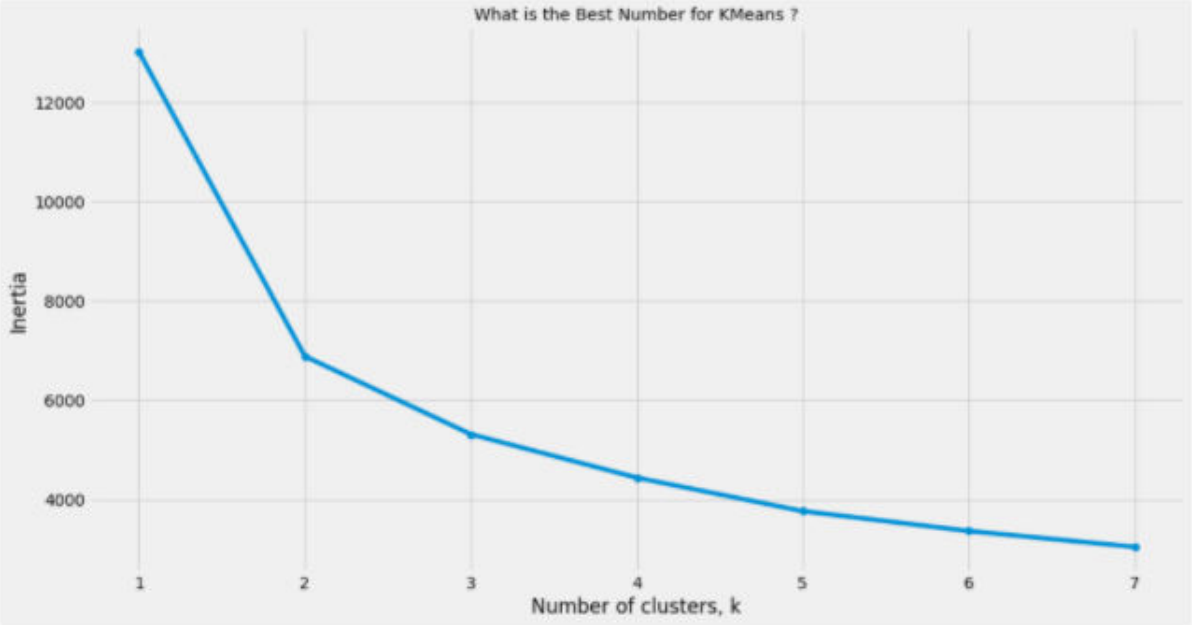
solution: logarithmic transformation(for positive values only) will manage it

Sequence of pre-processing steps:

1. Un-skew the data – log transformation
2. Standardize to the same average values
3. Scale to the same standard deviation
4. Store as a separate array to be used for clustering

Key steps of k-means clustering

1. Data pre-processing
2. Choosing a number of clusters
3. Running k-means clustering on pre-processed data
4. Analyzing average RFM values of each cluster

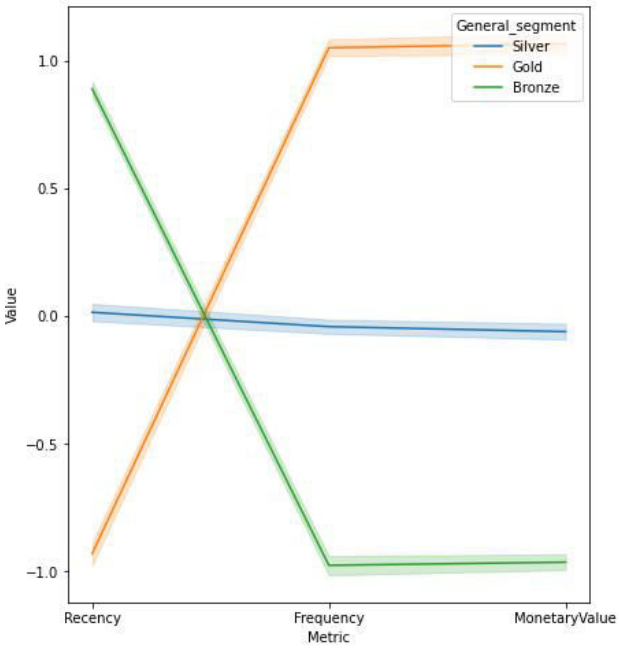


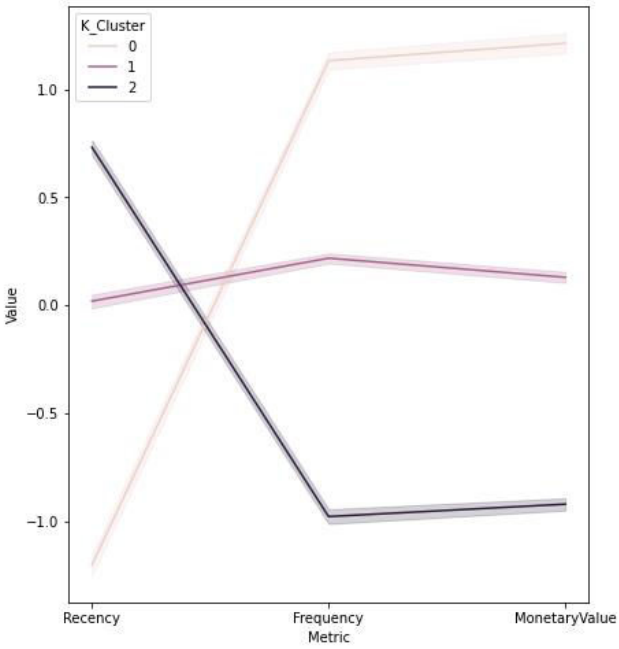
Note: here we choose the best value, k = 3

Then, we apply k-means algorithm to make K clusters.

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
K_Cluster				
0	13.0	260.0	6554.0	957
1	69.0	65.0	1167.0	1858
2	171.0	15.0	293.0	1523

Snake plot for better understanding and comparison





Heatmap for understanding Relative Importance