Challenge

Develop a "Recommendation Engine" AI solution for a website using an Unsupervised segmentation/clustering approach complete with economic analysis and presentation to management to convince the CMO and Webmaster to implement the solution.

Description:

An ecommerce website currently recommends "most popular" items to all users with the hope that users will add items to their shopping cart and increase the size of each purchase. This approach is successful by increasing sales for the website and the current CMO/Webmaster do not believe there is a more intelligent solution that is more relevant to each user based on their prior purchases and relationship with the ecommerce website.

INSTRUCTIONS

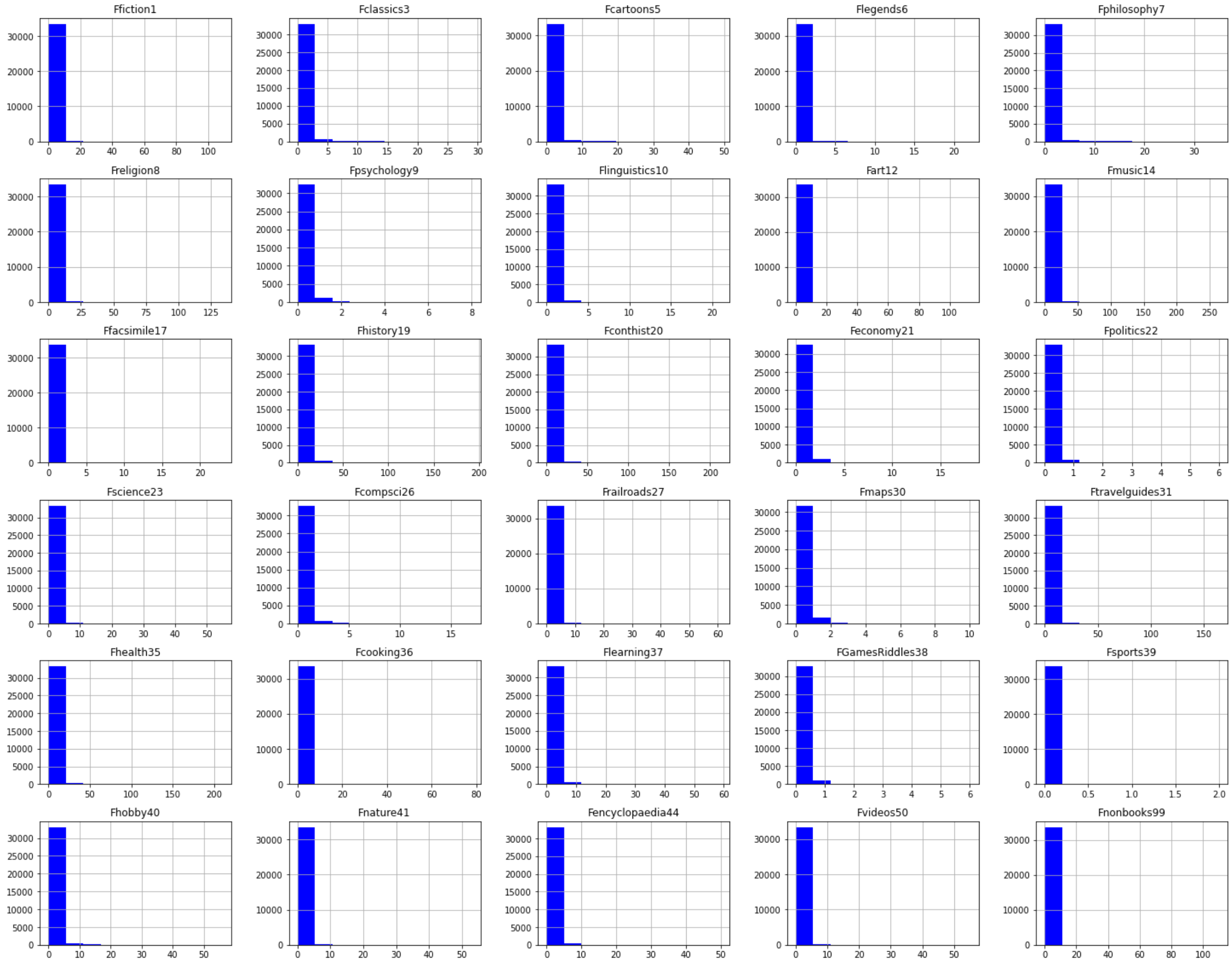
Perform the following tasks to help create a recommendation engine and AI solution to help improve the performance of the ecommerce website:

1. Perform descriptive statistics on all variables to help understand the data, the distributions and basic info you have to work with on this challenge. (Mean, Standard Deviation, Median, Min, Max and Histogram)
2. Create a new data set with more descriptive labels and the data needed for analysis
3. Document the problem you will address using the SMART framework – use hypothetical data and goals for the improvement your Recommendation Engine will deliver
4. Select from a pool of titles and roles within the company to create a core team (Maximum 6-8 members) to perform the analysis and develop the pitch to your CMO. (Pool: VP Finance, Data Engineer, SVP Data Scientist, Financial Analyst, Customer Satisfaction Manager, Website Analyst, Webmaster, VP Marketing, Performance Marketing SEM/SEO Analyst, Customer Retention Manager, Marketing Manager, Customer Research Analyst, Data Visualization Specialist, Sr. Data Scientist, Marketing Analytics Manager, IT Manager for Ecommerce Data Storage, Marketing Messaging/Creative Designer, Ecommerce Financial Manager)
5. Describe the team strengths and discipline focus areas that justify your need for each person and why this talent will be necessary for your success. This justification is intended to both motivate the team members and gain support from their managers to join your team.
6. Choose either a "supervised" or an "unsupervised" approach to segment/cluster current customers using the data provided. You may develop a rules base algorithm to ID a customer on each new transaction and assign them to one of your segments/clusters. Create sample data (5-10 records) with synthetic data to show how a new transaction will be scored and assigned to a segment/cluster. (You can use RFM<Recency, Frequency, Monetary Value>, K-Means or other rules based approached. Your segmentation must be MECE (Mutually Exclusive, Comprehensively Exhaustive) and any record of any incoming customer must receive one score and become assigned to a single segment – even if the transaction is incomplete or abandoned prior to purchase.

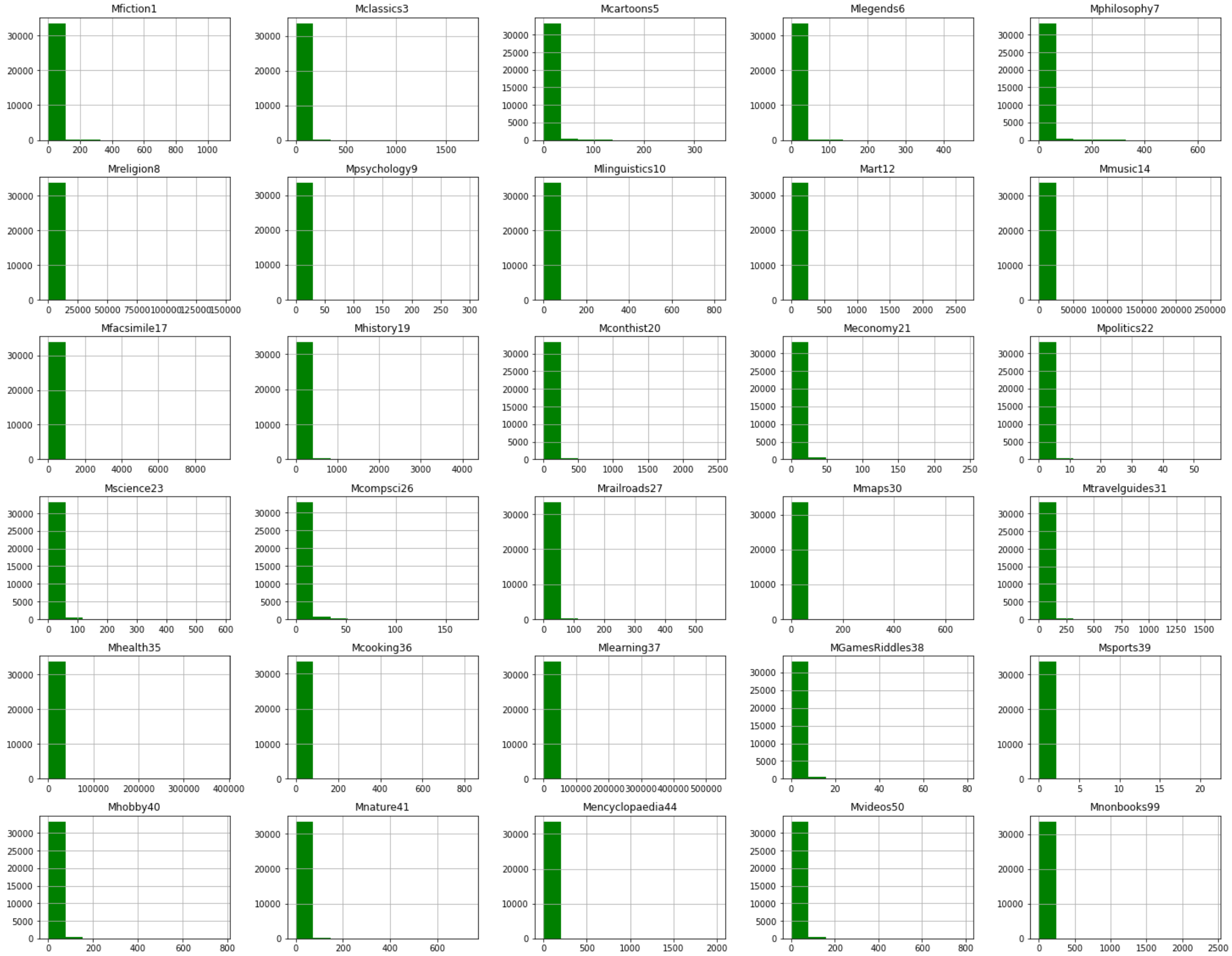
1. **Perform descriptive statistics on all variables to help understand the data, the distributions and basic info you have to work with on this challenge. (Mean, Standard Deviation, Median, Min, Max and Histogram)**

I segregated the 65 columns into three subsets for ease of analysis:

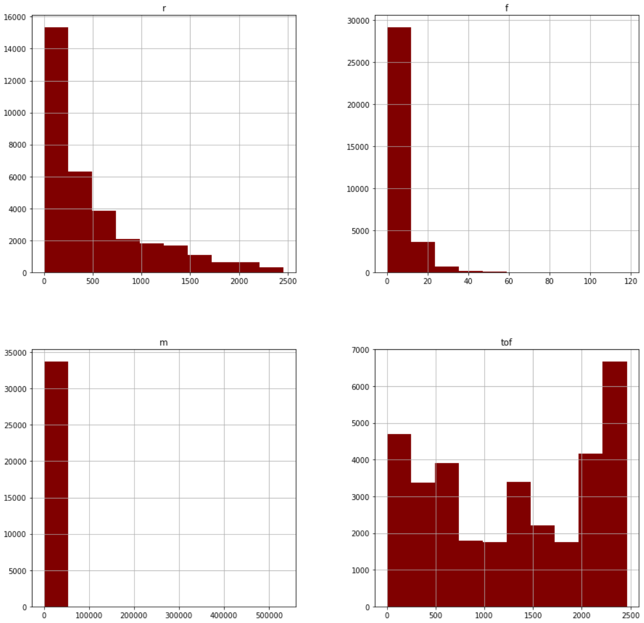
- Frequency: Columns starting with an F, denoting the # of transactions for each product category



- Monetary: Columns starting with an M, denoting the monetary value of transactions for each product category



- System calculated R,F,M categories



2. **Create a new data set with more descriptive labels and the data needed for analysis**

Understanding RFM Variables

Recency:
The more recently a customer has interacted or transacted with a brand. How long has it been since a customer engaged in an activity or made a purchase with the brand? The most common activity is a purchase for an online book store, though other examples include the most recent visit to a website.

Frequency:
During a given time period, how many times has a consumer transacted or interacted with the brand? Customers who participate in activities regularly are clearly more involved and loyal than those who do so infrequently.
It answers the question, how often?

Monetary:
This factor, also known as "monetary value," reflects how much a customer has spent with the brand over a given period of time. Those who spend a lot of money should be handled differently from customers who spend a little.
The average purchase amount is calculated by dividing monetary by frequency, which is a significant secondary element to consider when segmenting customers.

- I am creating two new monetary and frequency variables by adding all monetary and frequency product variables.
- This is done as a safeguard against system generated F, M values, and also take into account product level data metrics

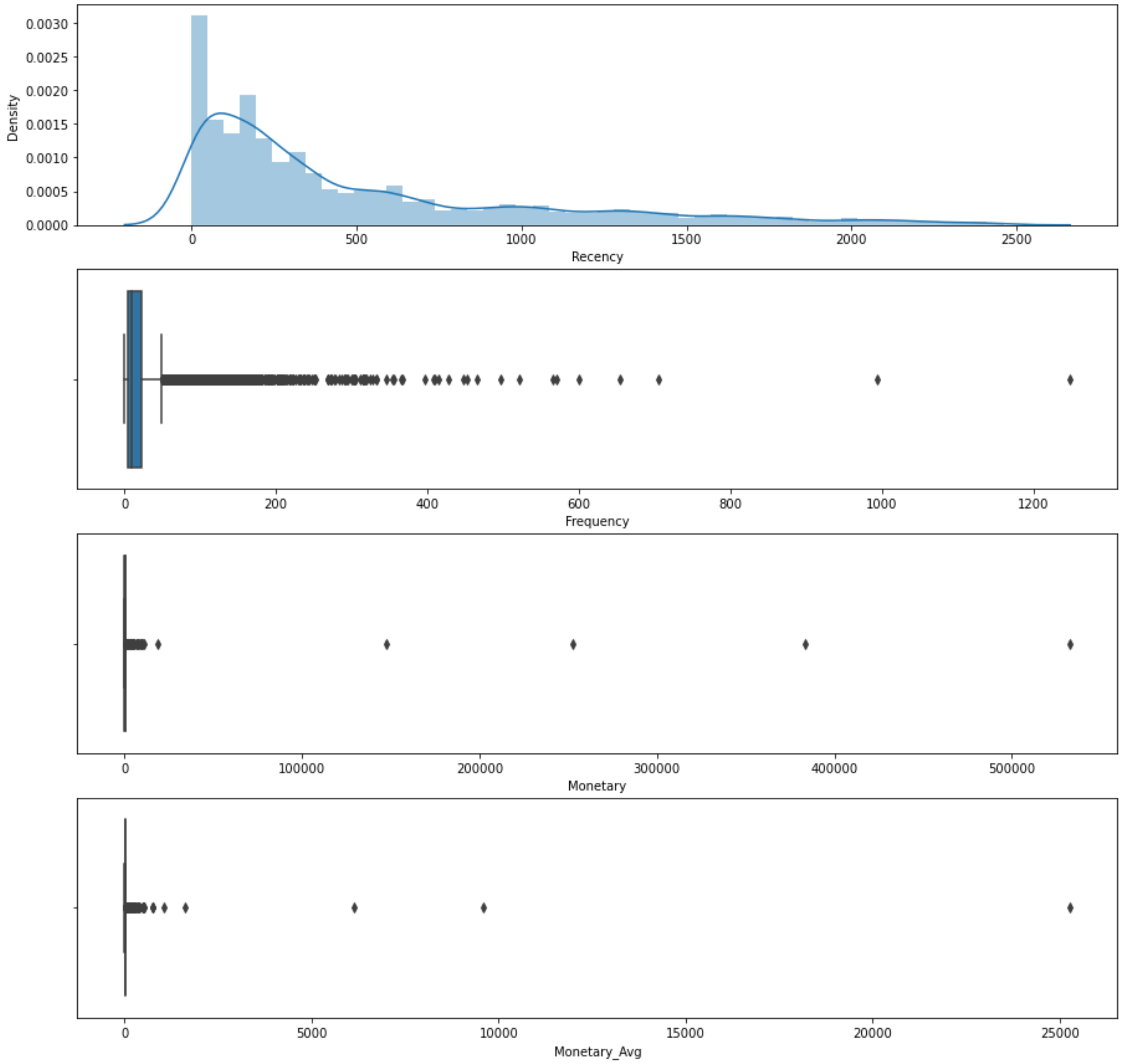
Column Reduction

- Once new R, F, M variables are calculated, then all product level variables are removed from the modelling dataset
- Final variables in the model include:

1. ID (rownames)
2. Recency (System Generated)
3. Frequency (Calculated)
4. Monetary (Calculated)
5. System generated Monetary
6. System generated Frequency
7. Monetary_Avg (which is equal to Monetary/Frequency to calculate the average price paid per purchase)

#2,3,4 will be used for RFM calculations

Row Reduction



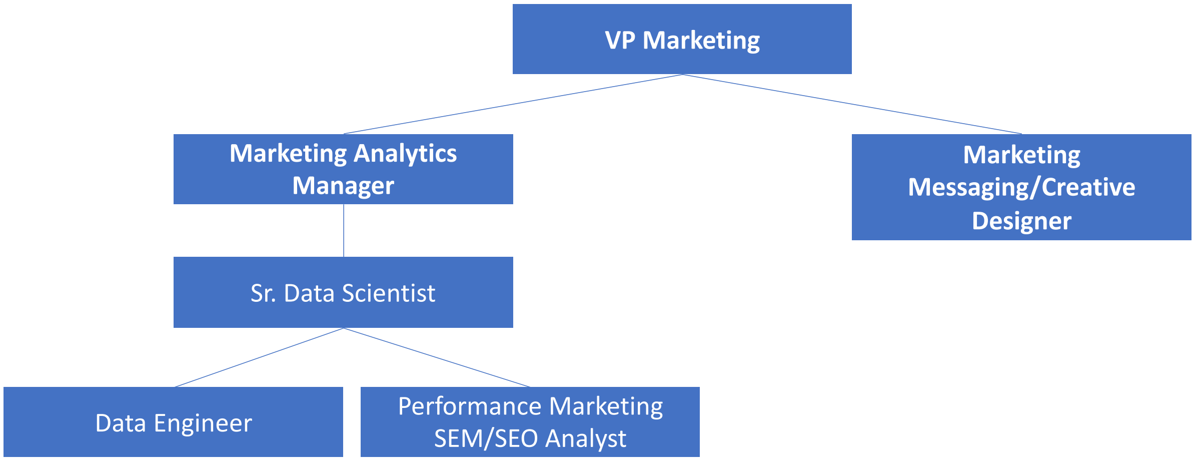
- After analyzing the spread for R,F,M variables, we can see that all three variables are highly right skewed.
- To take care of this, we are removing observations that are in the top 1% percentile for these three variables to analyze relevant trends better
- After the outliers are removed, the row count decreases from 33K to around 32K

3. **Document the problem you will address using the SMART framework – use hypothetical data and goals for the improvement your Recommendation Engine will deliver**

Specific • What are we trying to accomplish with this goal? • Will others easily understand this goal? • Is this goal specific enough, too specific**?**	Increase annual sales by 8-10% from the last fiscal year
Measurable • How will we know when the goal is accomplished? • How will we measure progress & success?	Increase consumer activity by a delta of 30% on the online platform. Refocus attention on 20% high value customers to provide a better user experience
Actionable • What resources do we need to accomplish the goal?	Analyze consumer preferences and product affinity trends using a team of 6 data scientists, analysts and consultants
Relevant • Is this goal really important right now? Why? • Why is this goal relevant to the company & mission	The intent is to test new methods to understand consumer insights better, and check if these can be translated to higher sales for the online bookstore
Time Bound • When is the start and due date for this goal? • What could push out the achievement date?	Roll out first set of updates within 60 days to a select few users Monitor KPIs for these specific users, and plan for a broader launch if we see an uptake on sales metrics

4. **Select from a pool of titles and roles within the company to create a core team (Maximum 6-8 members) to perform the analysis and develop the pitch to your CMO.**

Core Team – 7 members


VP Marketing
Marketing Analytics Manager
Marketing Messaging/Creative Designer
Sr. Data Scientist
Data Engineer
Performance Marketing SEM/SEO Analyst

5. **Describe the team strengths and discipline focus areas that justify your need for each person and why this talent will be necessary for your success. This justification is intended to both motivate the team members and gain support from their managers to join your team.**

- **Data Engineer**
  Strengths: Data Engineering and cleaning, Data Modelling
  Focus Areas: Responsible for ensuring an adequate data engineering pipeline has been set up to get high quality data for analysis

- **SEO Analyst**
  Strengths: Product analytics, Search Engine Optimization, A/B Testing
  Focus Area: Responsible for analyzing product affinities at a consumer level, and proposing strategies for content personalization

- **Sr. Data Scientist**
  Strengths: Data Science Strategy, Technical Communication
  Focus Area: Planning and executing the analytics along with data engineer and SEO analyst

- **Marketing Analytics Manager**
  Strengths: Technical and Non-technical communication, Marketing Strategy
  Focus Area: Translating Data Science insights and analytics into actionable marketing strategies, middle layer management

- **Marketing Creative Designer**
  Strengths: Content creation and delivery

Focus Area: Working with the analytics manager to come up with content refinements, and how different products should be positioned for a particular consumer segment

- **VP Marketing**
  Strengths: Project Management, Team Management
  Focus Area: Responsible for overall execution of the project, and providing insights and direction to the team wherever required

6. **Choose either a "supervised" or an "unsupervised" approach to segment/cluster current customers using the data provided. You may develop a rules base algorithm to ID a customer on each new transaction and assign them to one of your segments/clusters. Create sample data (5-10 records) with synthetic data to show how a new transaction will be scored and assigned to a segment/cluster. (You can use RFM<Recency, Frequency, Monetary Value>, K-Means or other rules based approached. Your segmentation must be MECE (Mutually Exclusive, Comprehensively Exhaustive) and any record of any incoming customer must receive one score and become assigned to a single segment – even if the transaction is incomplete or abandoned prior to purchase.**

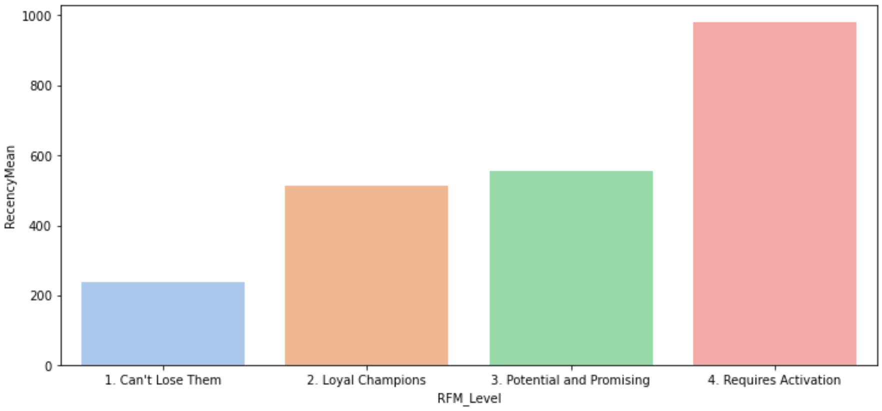
RFM unsupervised methodology was used to build 4 consumer segments:

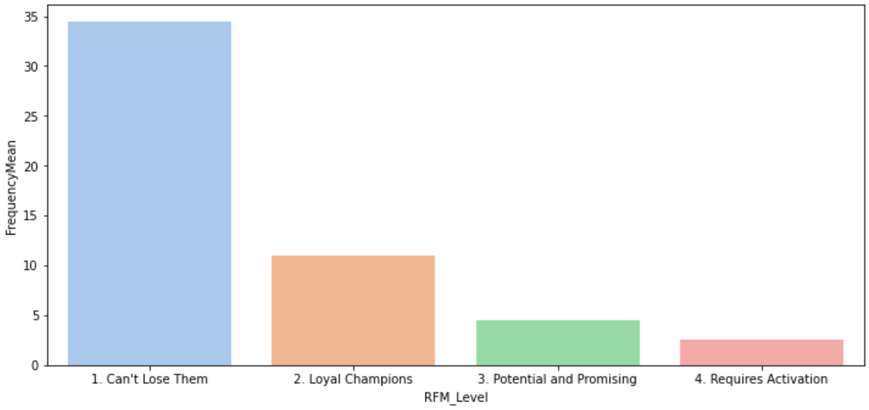
- Recency, Frequency, and Monetary variables are divided into their respective quantiles using the training data set
  - The quantile levels will stay constant for the initial transactions as per the training data set, these can be later made dynamic
- Each RFM Quantile was assigned a label:
  - 4,3,2,1 for Recency in decreasing order of values
  - 1,2,3,4 for Frequency and Monetary in increasing order of values
- A composite RFM tag was given to each particular customer ID using these 3 quantiles
- Using this tag, a composite RFM score was given to each customer
  - For example: A respondent with '444' RFM tag will be allotted a score of 12 (4+4+4)
- Using this score, four segment levels can be defined:
  - Can't Lose Them
    - **Score >=9**
    - Highest value segments with least recency, and highest frequency and monetary
    - These select customers account for majority of the platform sales

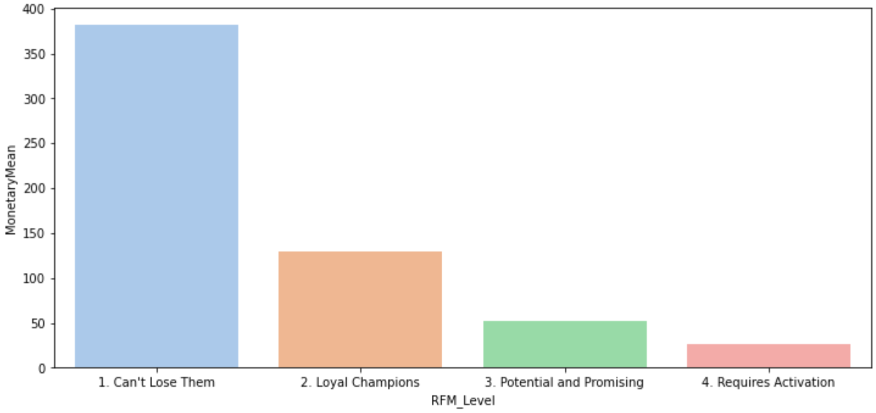
  - Loyal Champions
    - **Score >= 7 and <9**
    - These are the loyal set of customers who might not have high monetary purchases, but they are frequent shoppers

- Potential and Promising
  - **Score >=5 and <7**
  - These customers can be pushed into the loyal champions segments if activated correctly

- Require Activation
  - **Score <5**
  - These are low value customers who don’t have either good Recencies, Frequencies or Monetary value for the platform
    Pushing them towards high value segments will be a hard task for the company


RecencyMean
1. Can't Lose Them
2. Loyal Champions
3. Potential and Promising
4. Requires Activation
RFM_Level


FrequencyMean
1. Can't Lose Them
2. Loyal Champions
3. Potential and Promising
4. Requires Activation
RFM_Level


MonetaryMean
1. Can't Lose Them
2. Loyal Champions
3. Potential and Promising
4. Requires Activation
RFM_Level