SEGMENTATION

Generating **insights**

to help **you**

understand and serve

your **customers**

better!

BAX453 Final Project | Team #4

Angie | Pallavi | Shanxing | Shruti

- Background
- EDA
- Modeling
- Business Insights
- Beyond Insights

Connecting **the dots…**

Industry Vertical, Business Domain, and Use cases

Industry Vertical:
An industry vertical (also called a vertical market) is more specific, identifying companies that offer niche products or fit into multiple industries.
In our case, the industry is specifically retail business serving business (B2B)

Business Domain:
In our case, we are focusing on the business problem which is in the domain of retail marketing function.

Use Cases:
Our analysis could be used to:

- analyze customer behaviors
- analyze sales patterns
- segment customers
- create marketing campaigns that target the specific segment of customers
- monitor metrics based on customer segments

Business Problem

- Company X is an online B2B retailer sells small widgets for home/apartment decorations to other small businesses globally
- The company was founded in December 2010 and immediately attracted many business clients within first 3 months
- The monthly total sales dropped starting in Dec 2011. Since it's the season that most people would make more purchase for gifting, the marketing and sales team were concerned about the revenue performance.
- The marketing team had analyzed all the factors and decided in order to get better conversion and retention, they need to understand the segmentation of the clients in order to create better marketing campaigns.
- **We are a team of consultants who will help Company X segment their clients to help the marketing team design the next campaign.**

Key Business Questions

- ❑ What are the criteria to define each segmentation?
- ❑ How will the segmentation be done?
- ❑ How feasible is it to “standardize” segmentation across business?
- ❑ How will Company X benefit from the segmentation?

External Research - Business

B2B Customer Segmentation Challenges:

- B2B markets have a more complex decision-making unit: In most households, even the most complex and expensive of purchases are confined to the small family unit, while the purchase of items such as food, clothes and cigarettes usually involves just one person.

- B2B buyers are more “rational”: The view that B2B buyers are more rational than consumer buyers is perhaps controversial, but we believe true.

- B2B products are often more complex: Just as the decision-making unit is often complex in business-to-business markets, so too are b2b products themselves.

- B2B target audiences are smaller than consumer target audiences: Almost all business-to-business markets exhibit a customer distribution that confirms the Pareto Principle or 80:20 rule. A small number of customers dominate the sales ledger.

External Research - Business

B2B Customer Segmentation Challenges:

- Personal relationships are more important in b2b markets: A small customer base that buys regularly from the business-to-business supplier is relatively easy to talk to. Sales and technical representatives visit the customers.

- B2B buyers are longer-term buyers: Whilst consumers do buy items such as houses and cars which are long-term purchases, these incidents are relatively rare. Long-term purchases – or at least purchases which are expected to be repeated over a long period of time – are more common in business-to-business markets, where capital machinery, components and continually used consumables are prevalent.

- B2B markets drive innovation less than consumer markets: B2B companies that innovate usually do so as a response to an innovation that has happened further upstream.

- B2B markets have fewer behavioural and needs-based segments: The small number of segments typical to b2b markets is in itself a key distinguishing factor of business-to-business markets.

(https://www.b2binternational.com/publications/b2b-segmentation-research)

External Research - Customer Segmentation

To understand the real value of customers, we can do analysis based on transaction data and cluster the customers with similar characteristics to divide customers into different strategic segments. Customer segmentation can help in the following areas and companies can use the insights from the data to make strategic decisions such as:

- Discover cross-sell and upsell opportunities
- Cut down on unnecessary costs
- Concentrate on your most valuable customers
- Prioritize new product development efforts
- Develop customized marketing programs
- Choose specific product features
- Establish appropriate service options
- Design an optimal distribution strategy
- Determine appropriate product pricing

External Research - Customer Segmentation

We will work on the following aspects to determine optimal number of customer segments:

- The segment is well definable that it should have specific features that reacts to a marketing campaign
- The segment is significant enough that could make a significant profit through a campaign since it's not possible to make a campaign for every segment
- The segment is accessible that it can be reached through some marketing channels like social media, tv, etc.
- The segment is stable that the features, indicators, and definition which are needed for analysis won't change easily
- The segment is well distinguishable that customers have different preferences, demands, problems, behavior patterns from other segments. If there are too many things in common, it might be a good idea to unify the segments in order to make the campaigns simpler and cost effective
- The segment can be activated that marketing campaigns can be made to address the problems in the segment

(https://aionhill.com/ecommerce-customer-segmentation)

Data Availability

Dataset:

- The dataset is a UK-based e-commerce retailer transactional data available through the UCI Machine Learning Repository. It contains the transactions occurring between Dec 2010 to Dec 2011
- Link to the dataset: https://archive.ics.uci.edu/ml/datasets/Online+Retail
- It contains 0.5 million instances and 8 columns which provides data on ~26K unique transactions made by ~4.3K customers on ~4K unique products.

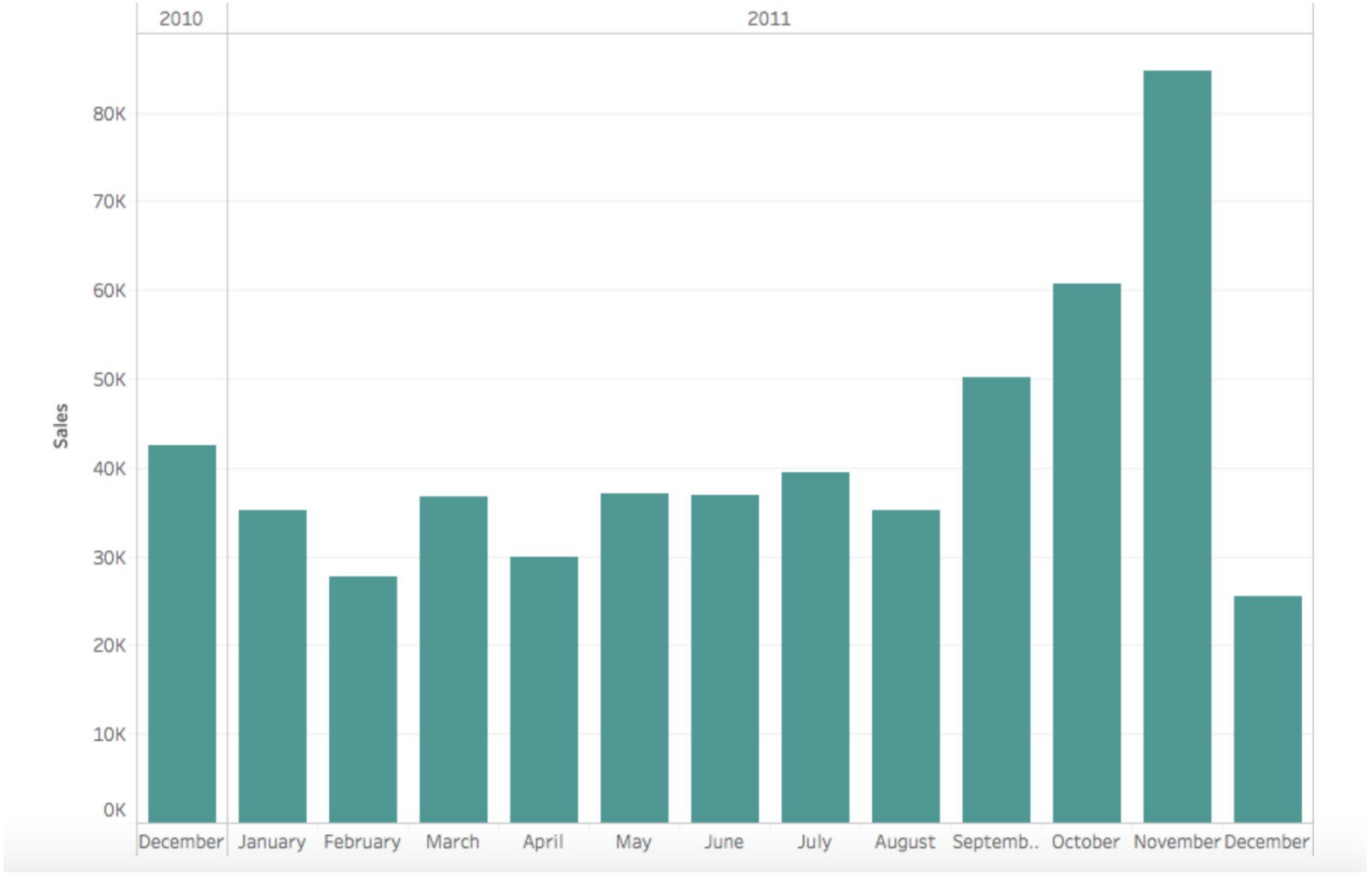
Data Dictionary :

Columns	Data Type	Description
InvoiceNo	Alphanumeric	Unique transaction number. If it starts with 'c', then it's a cancellation
StockCode	Alphanumeric	Item Code to identify each item uniquely
Description	String	Item description
Quantity	Integer	Quantity purchased of each item in a transaction
InvoiceDate	Timestamp	Date and time of the transaction
UnitPrice	Float	Price per unit of the item
CustomerID	Integer	Unique ID assigned to each customer
Country	String	Country name where each customer resides

Business Problem Exploration

Graph shows the monthly sales went down:

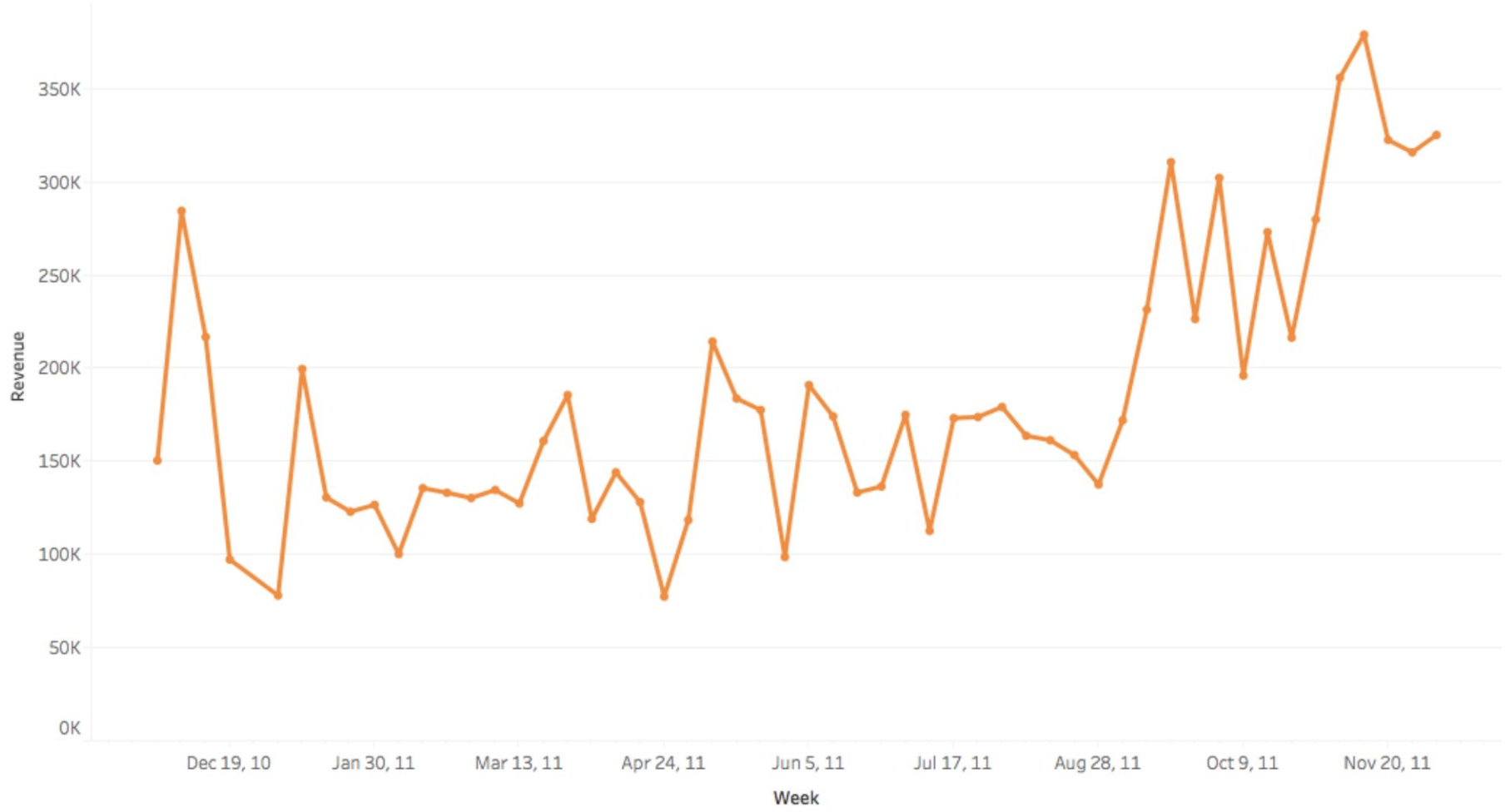
Intuitively December should be high due to gifting and decorations in the holiday season.



Business Problem Exploration - Cont'd

Graph shows Weekly revenue:

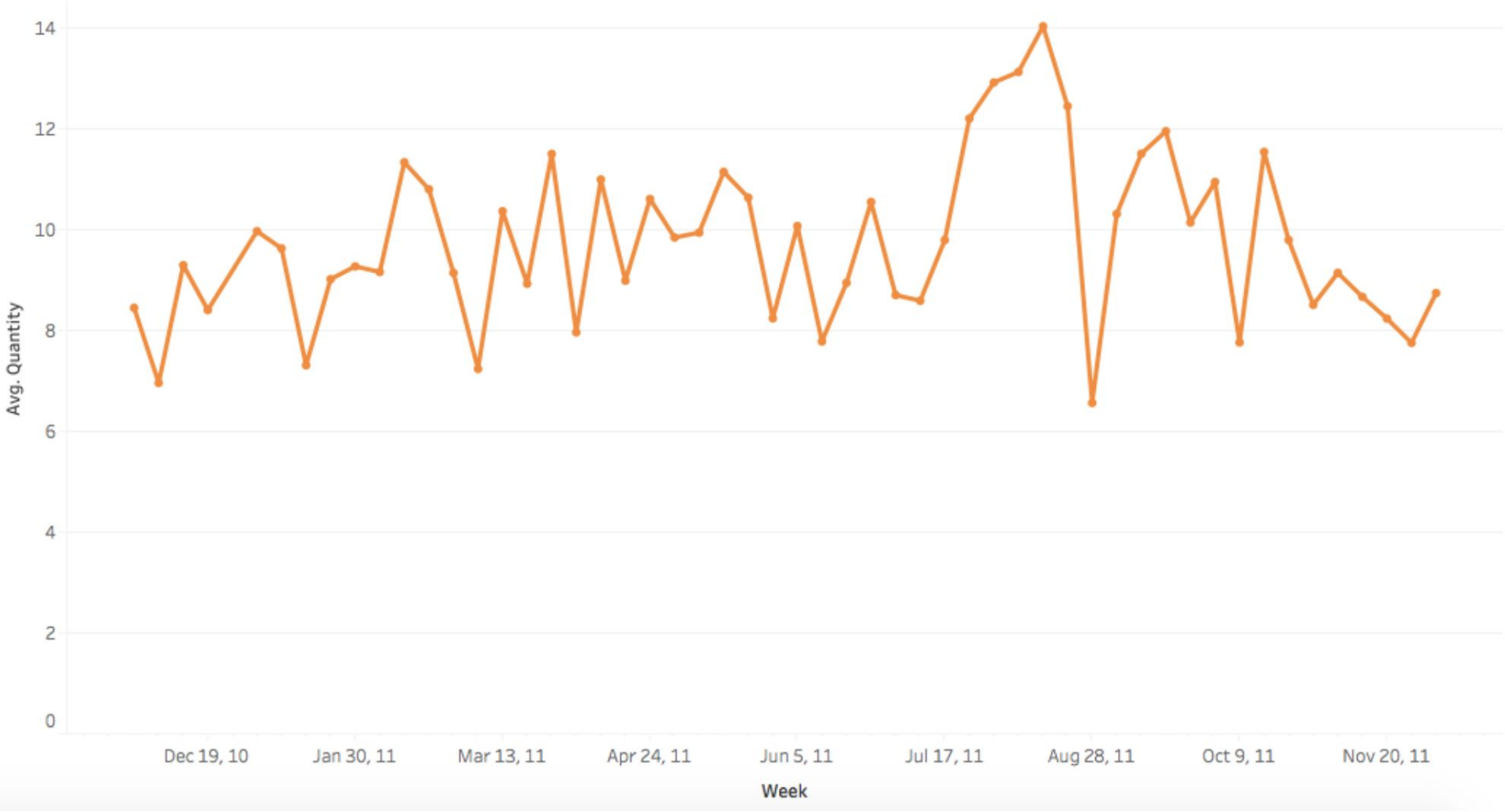
Marketing & Sales teams are concerned about maintaining the revenue performance post Nov 2011.



Business Problem Exploration - Cont'd

Graph shows Average Items Purchased:

Average number of items purchased weekly is declining.



Key Metrics

We perform RFM analysis and extract the following metrics based on their purchase history to reflect potential similarities between each client:

- **Recency**
- **Frequency**
- **Monetary value**

RFM analysis helps to answer questions like:

- Who are my best customers?
- Which customers are at the verge of churning?
- Who has the potential to be converted in more profitable customers?
- Who are lost customers that you don't need to pay much attention to?
- Which customers you must retain?
- Who are your loyal customers?
- Which group of customers is most likely to respond to your current campaign?

Exploratory **Data Analysis**

Data Quality Assessment

Check missing values

```
# check if data set contains missing values
pd.DataFrame([df.isnull().sum(),df.isnull().sum()/len(df)*100], index=['num_missing', 'percent_missing'])
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceYearMonth
num_missing	0.0	0.0	1454.000000	0.0	0.0	0.0	135080.000000	0.0	0.0
percent_missing	0.0	0.0	0.268311	0.0	0.0	0.0	24.926694	0.0	0.0

- There are around 25% of Customers missing their ID. It is impossible to impute values for the ID, hence we delete them from the analysis
- There are around 0.27% of items missing their description. And similar to customer ids, it is impossible to impute values for the description. So, we remove these from the dataset

Data Quality Assessment

Check duplicates

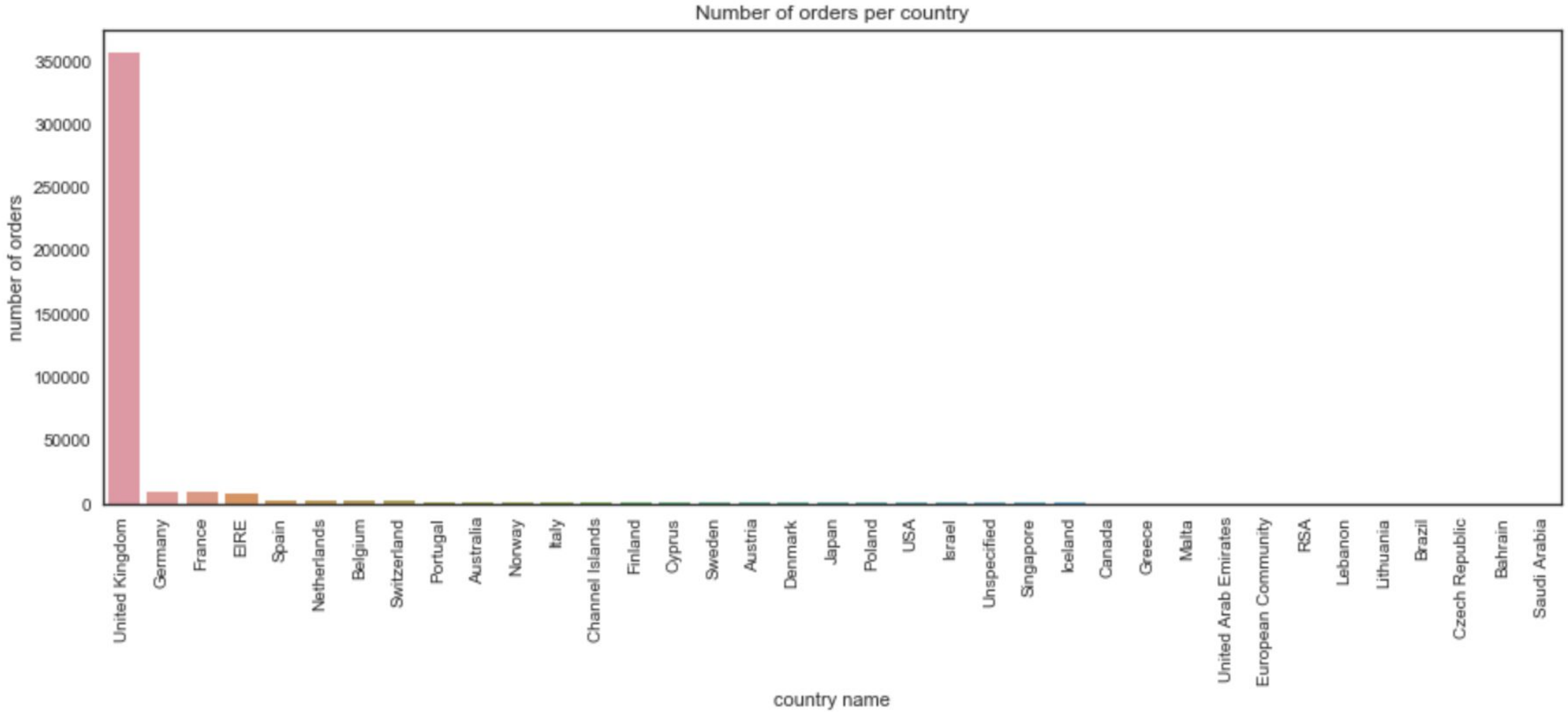
```
# check if data set contains duplicated records
print(f'The dataset contains {df.duplicated().sum()} duplicated values.')
```

```
The dataset contains 5225 duplicated values.
```

```
# remove duplicated values from dataset
df.drop_duplicates(inplace = True)
```

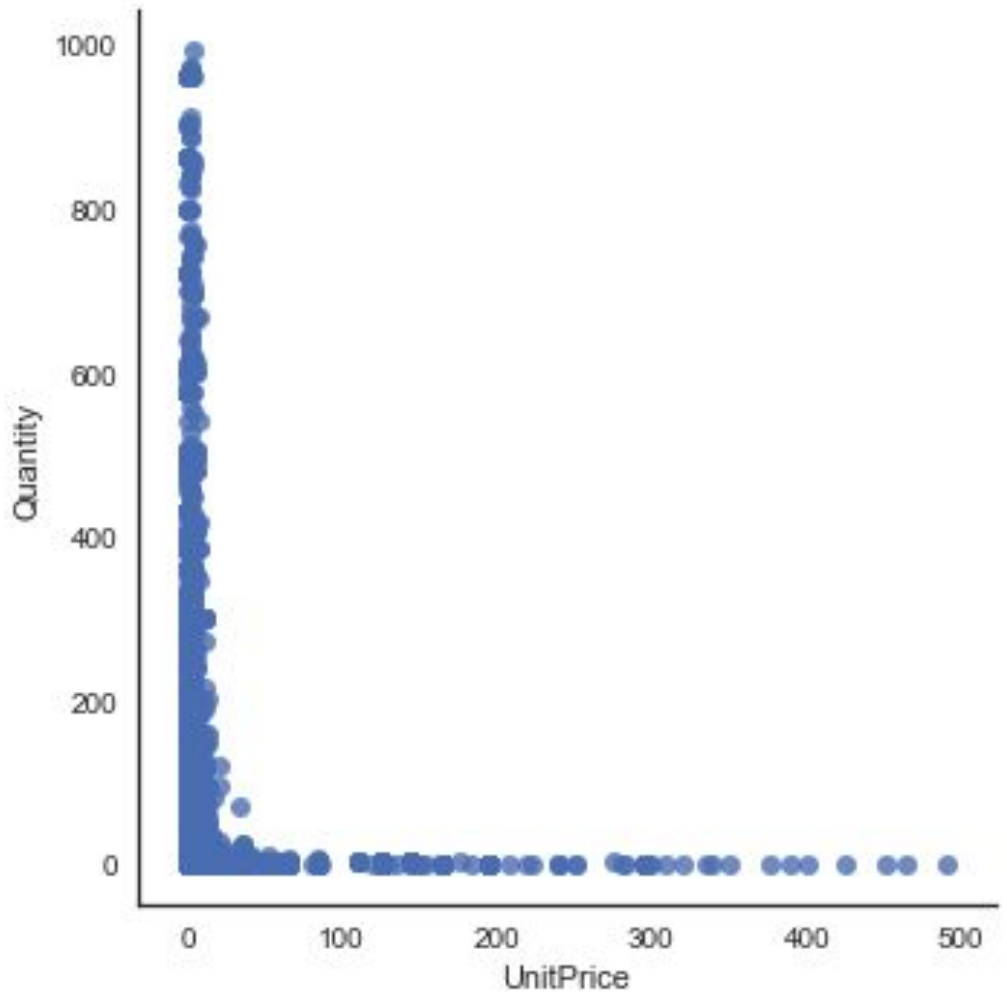
- The dataset contains 5225 duplicated values.

Country



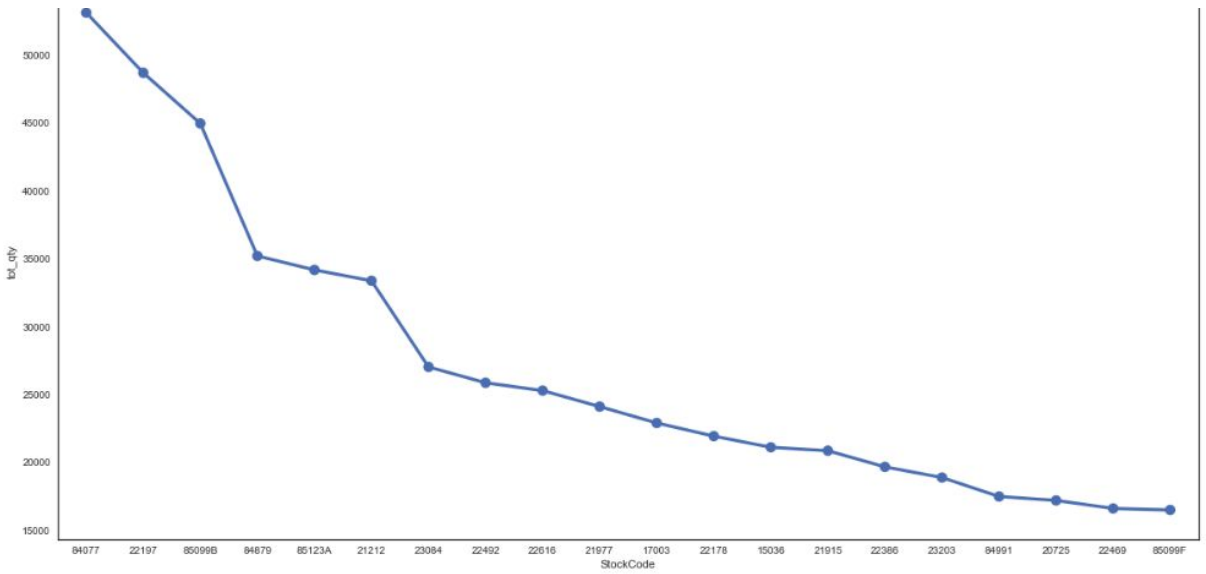
- We can see that the dataset is largely dominated by orders made from the UK.

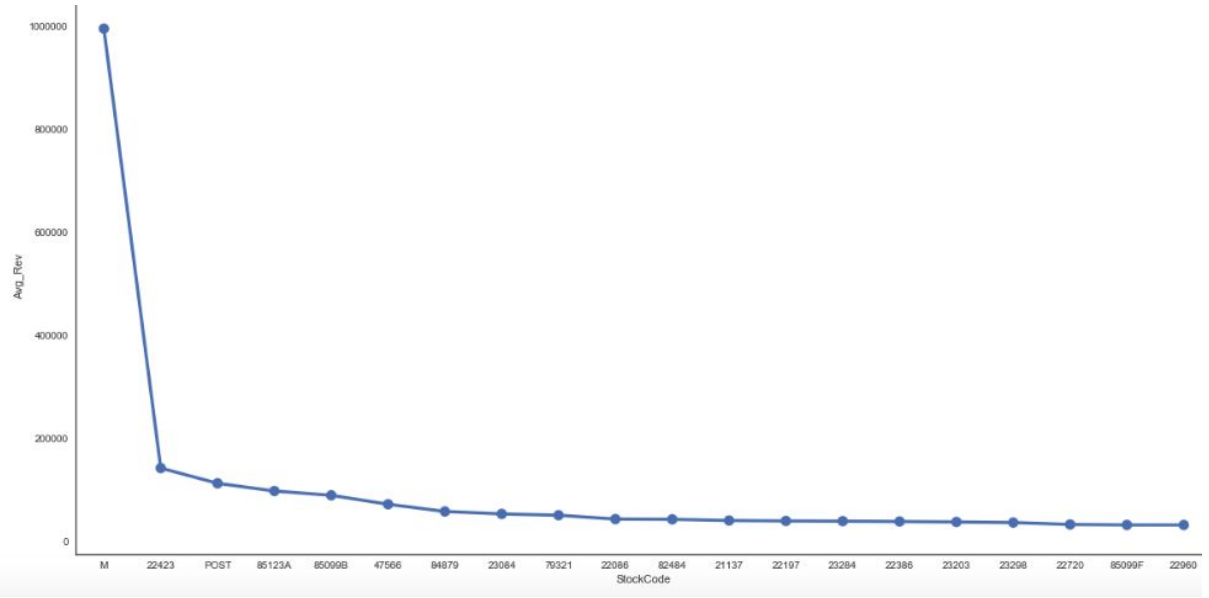
Items & transactions



- Intuitive - Customers buy more of less priced items or maybe in bulk, and very few high priced items.

Items & transactions





- Top selling products based on the stock code.
- Top selling in terms of quantity sold (first graph) and average revenue (second graph).

Customers & transactions

CustomerID	InvoiceNo	Quantity	TotalPrice
12346.0	541431	74215	77183.60
	C541433	-74215	-77183.60
12347.0	537626	319	711.79
	542237	315	475.39
	549222	483	636.25
	556201	196	382.52
	562032	277	584.91
	573511	676	1294.32
	581180	192	224.82
12348.0	539318	1254	892.80

- Create a new variable, TotalPrice = Quantity * UnitPrice.
- There are 4372 customers and 22190 transactions in the dataset.
- The existence of entries with the prefix C for the InvoiceNo variable: this indicates transactions that have been canceled.

Customers & transactions

	InvoiceNo	Quantity	UnitPrice	TotalPrice	order_canceled
141	C536379	-1	27.50	-27.50	1
154	C536383	-1	4.65	-4.65	1
235	C536391	-12	1.65	-19.80	1
236	C536391	-24	0.29	-6.96	1
237	C536391	-24	0.29	-6.96	1

- Create a new variable, order_canceled (0 not cancel, 1 canceled).
- There is a significant amount of transactions which have been canceled by customers (3654/22190 = 16.47%).
- For the orders that have been canceled, will have identical negative quantity.

Customers & transactions

	InvoiceNo	Quantity	UnitPrice	TotalPrice	TotalPaidPrice
0	536365	6	2.55	15.30	15.30
1	536365	6	3.39	20.34	20.34
2	536365	8	2.75	22.00	22.00
3	536365	6	3.39	20.34	20.34
4	536365	6	3.39	20.34	20.34

	InvoiceNo	Quantity	UnitPrice	TotalPrice	TotalPaidPrice
141	C536379	-1	27.50	-27.50	0.0
154	C536383	-1	4.65	-4.65	0.0
235	C536391	-12	1.65	-19.80	0.0
236	C536391	-24	0.29	-6.96	0.0
237	C536391	-24	0.29	-6.96	0.0

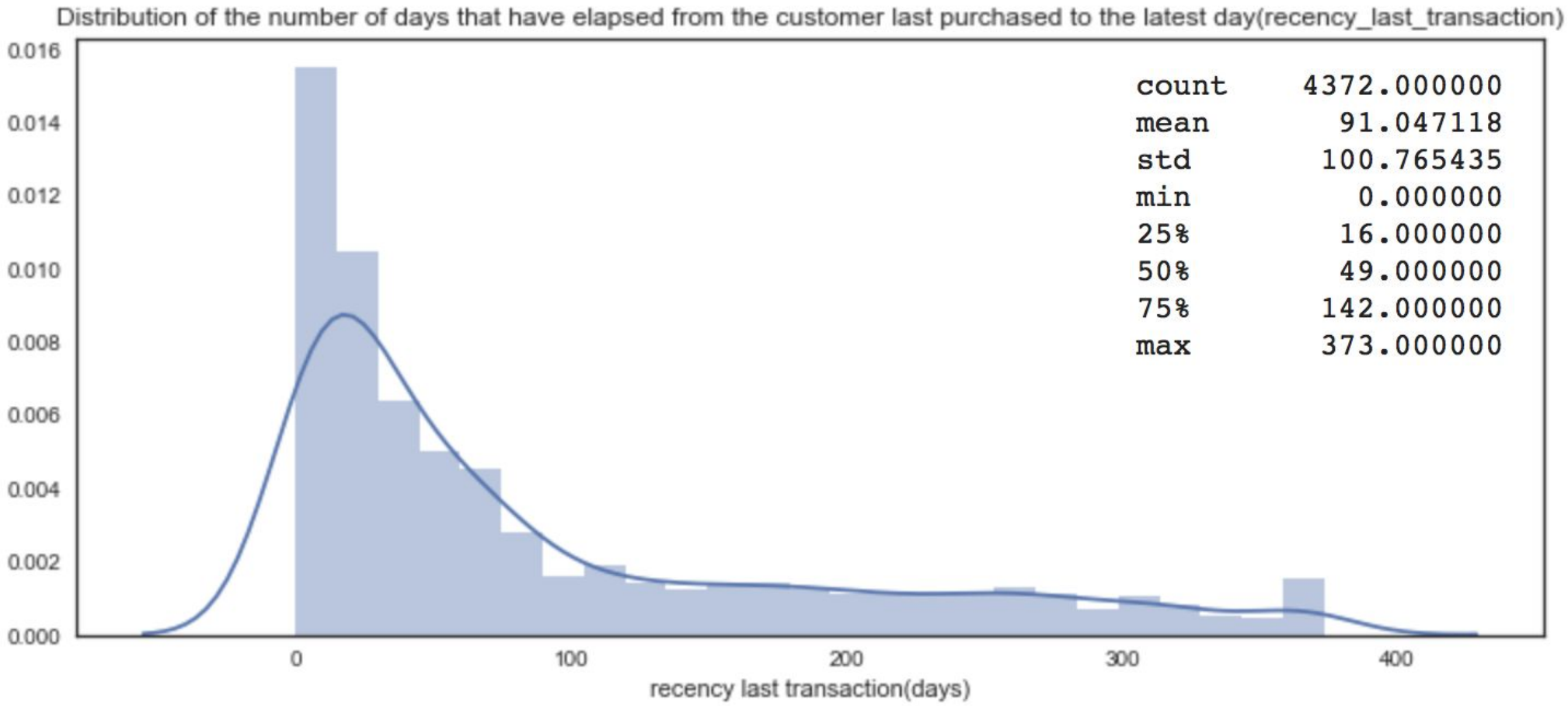
- Create a new variable, TotalPaidPrice, indicates the total amount of money that a customer paid for certain items in certain transaction(did not cancel at the end). For the items that the customer finally canceled, will assign a value of 0.
- Example of TotalPaidPrice in normal transaction (not be canceled at the end) - Left graph.
- Example of TotalPaidPrice in canceled transaction (been canceled at the end) - Right graph.

Create a customer-level dataset: Recency

	CustomerID	InvoiceNo	InvoiceDate	TotalPaidPrice	recency	recency_last_transaction
0	17850.0	536365	2010-12-01 08:26:00	15.30	373	301
1	17850.0	536365	2010-12-01 08:26:00	20.34	373	301
2	17850.0	536365	2010-12-01 08:26:00	22.00	373	301
3	17850.0	536365	2010-12-01 08:26:00	20.34	373	301
4	17850.0	536365	2010-12-01 08:26:00	20.34	373	301

- Create a new variable, "recency", defined as the number of days that have elapsed from each transaction to the latest day in the dataset.
- Create a new variable, "recency_last_transaction", defined as the number of days that have elapsed from the customer last purchased something to the latest day in the dataset.
- Hence, smaller numbers indicate more recent transaction on the customer's account.
- For "recency", the same customer with the same invoiceID will have the same value. But the same customer with different invoiceID will have different values.
- For "recency_last_transaction", the same customer with any invoiceID will have the same value.

Create a customer-level dataset: Recency



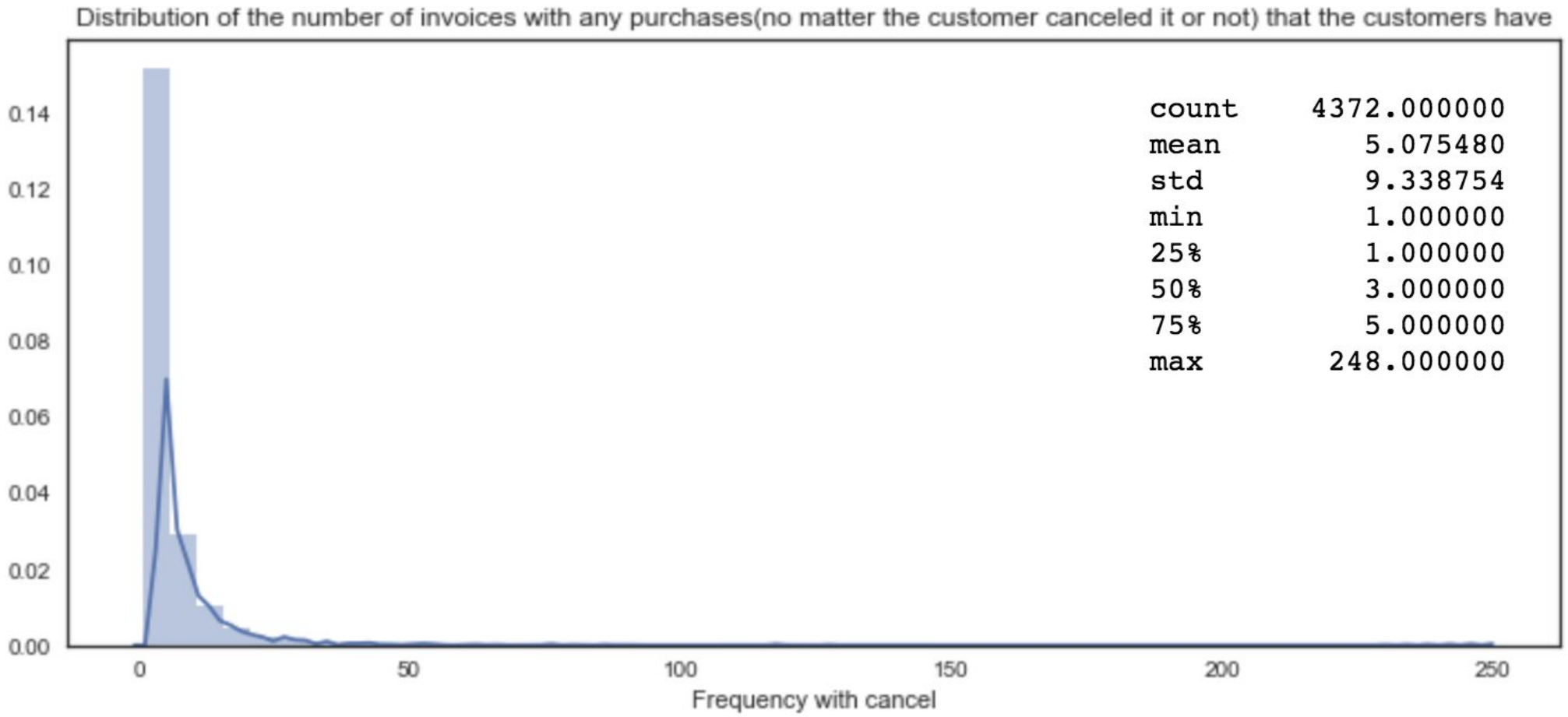
- Distribution of the number of days that have elapsed from the customer last purchased to the latest day(recency_last_transaction).

Create a customer-level dataset: Frequency

	InvoiceNo	CustomerID	recency	recency_last_transaction	Frequency_with_cancel	Frequency_exclude_cancel
0	536365	17850.0	373	301	35	34
1	536365	17850.0	373	301	35	34
2	536365	17850.0	373	301	35	34
3	536365	17850.0	373	301	35	34
4	536365	17850.0	373	301	35	34

- For every customer, the "Frequency_with_cancel" refers to the number of invoices with any purchases(no matter the customer canceled it or not) during the entire time period in this dataset (from 2010-12-01 to 2011-12-09).
- For every customer, the "Frequency_exclude_cancel" refers to the number of invoices with paid purchases(exclude the number of invoice that been canceled) during the entire time period in this dataset (from 2010-12-01 to 2011-12-09).
- The same customer with any invoiceID will have the same value.

Create a customer-level dataset: Frequency



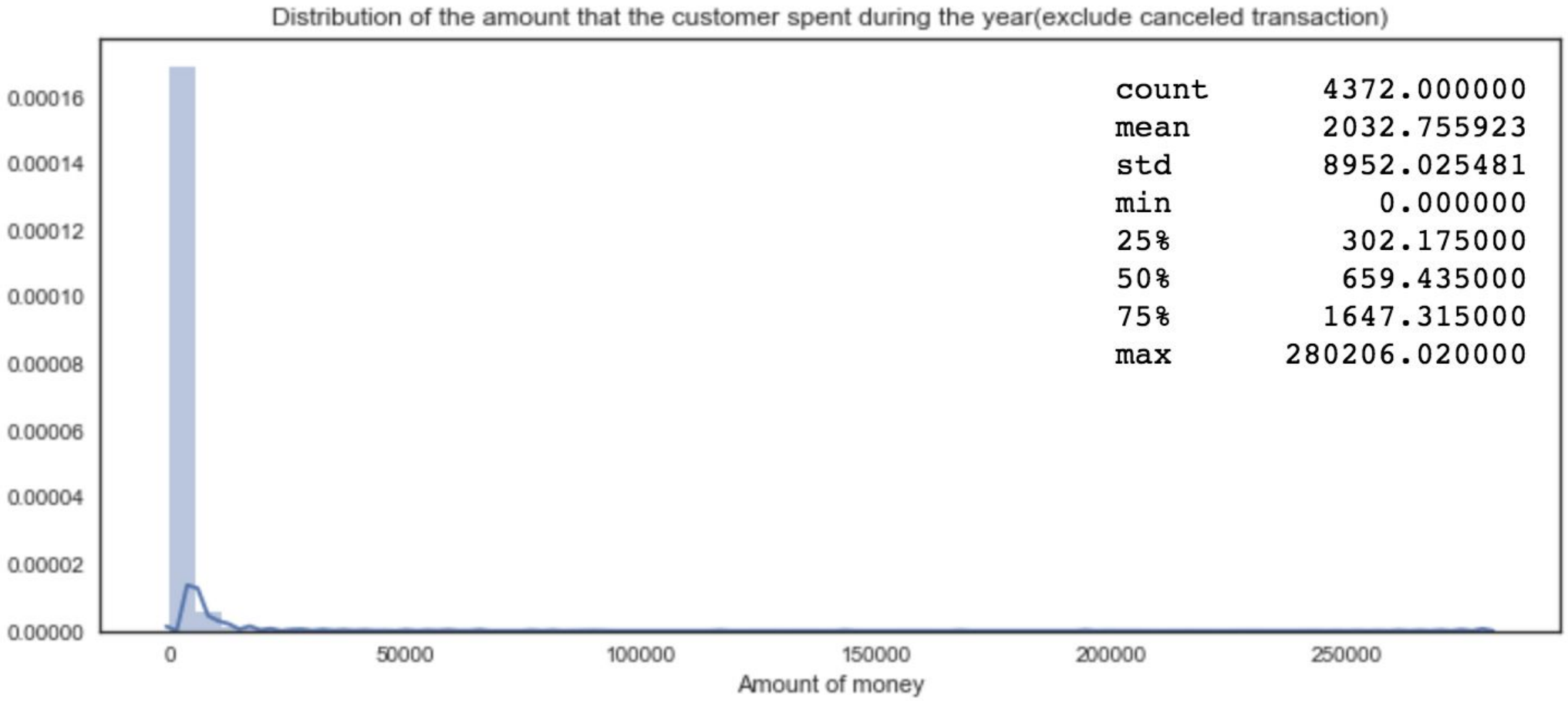
- Distribution of the number of invoices with any purchases (no matter the customer canceled it or not) that the customers have.

Create a customer-level dataset: Monetary value

	InvoiceNo	CustomerID	recency	recency_last_transaction	Frequency_with_cancel	Frequency_exclude_cancel	Monetary_value
0	536365	17850.0	373	301	35	34	5391.21
1	536365	17850.0	373	301	35	34	5391.21
2	536365	17850.0	373	301	35	34	5391.21
3	536365	17850.0	373	301	35	34	5391.21
4	536365	17850.0	373	301	35	34	5391.21

- "Monetary_value" is the amount that the customer spent during the year(exclude canceled transaction). For the customers who only have records of cancellation, will be assigned value of 0.
- The same customer with any invoiceID will have the same value.

Create a customer-level dataset: Monetary value



- Distribution of the amount that the customer spent during the year(exclude canceled transaction).

Create a customer-level dataset: Customer basket

	CustomerID	TotalPrice	Quantity	InvoiceNo	AvgBasketDollar	AvgBasketQty
0	12346.0	0.00	0	2	0.000000	0.000000
1	12347.0	4310.00	2458	7	615.714286	351.142857
2	12348.0	1797.24	2341	4	449.310000	585.250000
3	12349.0	1757.55	631	1	1757.550000	631.000000
4	12350.0	334.40	197	1	334.400000	197.000000

- Calculate average monetary value (average basket size for money spent and quantity bought)

Combining RFM metrics and treating outliers

	CustomerID	totalMonetary	totalQty	frequency	AvgBasketDollar	AvgBasketQty	recency
0	12346.0	0.00	0	2	0.000000	0.000000	325
1	12347.0	4310.00	2458	7	615.714286	351.142857	1
2	12348.0	1797.24	2341	4	449.310000	585.250000	74
3	12349.0	1757.55	631	1	1757.550000	631.000000	18
4	12350.0	334.40	197	1	334.400000	197.000000	309

- Combining RFM metrics by grouping by the customerID.
- Use 3(IQR) criterion to identify potential outliers and extreme outliers.
- We will use these variables for clustering.

Modeling

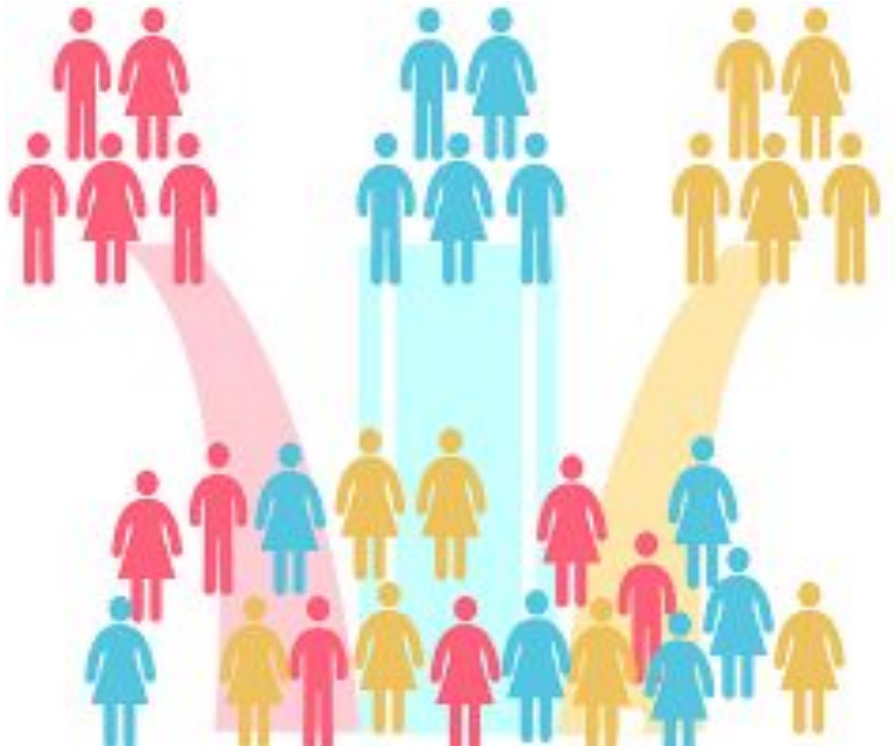
Data science and algorithms used

Clustering

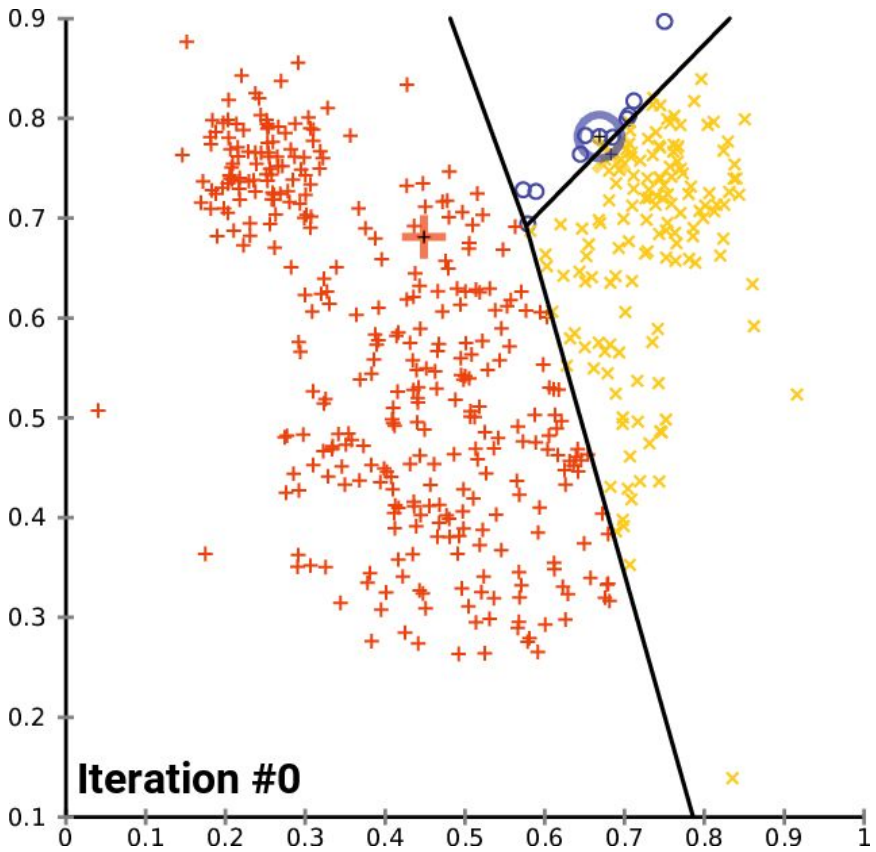
- **What** - Clustering is a category of unsupervised learning techniques that allows us to discover hidden structures in data where we do not know the right answer upfront. The goal is to find a natural grouping in data such that objects in the same cluster are more similar to each other than those from different clusters

- **Why** - Cluster analysis on customers allows companies to group customers into segments that share certain characteristics such as demographics, behavior, preferences or demand. Based on such segments, product offerings and marketing strategies can be better targeted by distinguishing certain categories of needs

- **How** - Using behavioral information from the RFM metrics - Recency of last purchase, Frequency of purchase, and Monetary value



Data science and algorithms used



K-means

- **What** - It is a widely used algorithm in cluster analysis. It is essentially an optimization problem that aims at minimizing the sum of squared errors within the cluster. “k” defines the number of clusters. Each cluster is represented by a prototype, which can either be the **centroid** of similar points with continuous features, or the **medoid** in the case of categorical features

- **Why** - It is extremely easy to implement and computationally very efficient compared to other clustering algorithms. The assignments of objects to clusters is hard assignment. Points belong explicitly to one cluster unlike other techniques like Gaussian Mixture Model where points have probabilities of belonging to clusters. This eases the understanding and actionability of the clustering results

- **How** - Leverage the off-the-shelf machine learning technique from the scikit learn in Python. It runs an iterative algorithm to find the best clusters given the input data and parameters

Key Assumptions

K-means algorithm makes the following assumptions on the clusters -

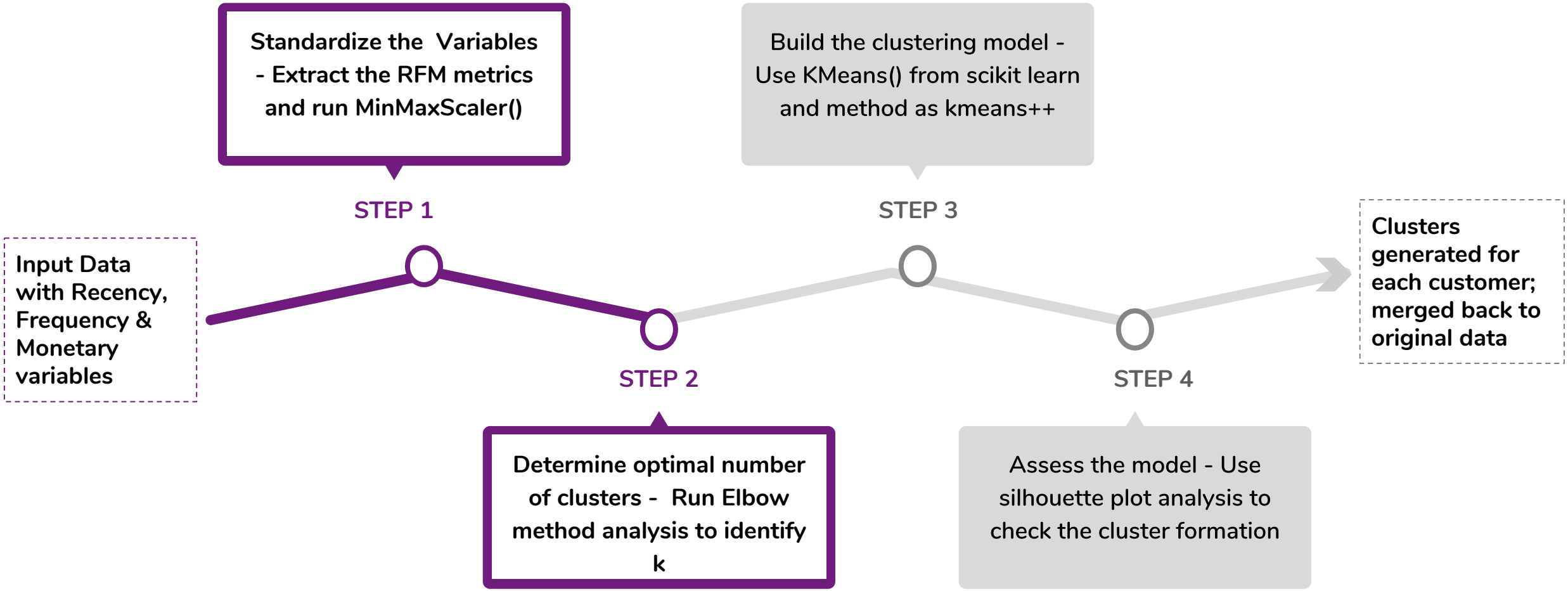
- The clusters created are ***spherical***
  - Spherical assumption helps in separating the clusters when the algorithm works on the data and forms clusters. If this assumption is violated, the clusters formed may not be what one expects.

- The clusters are of ***similar size***
  - Assumption over the size of clusters helps in deciding the boundaries of the cluster and in calculating the number of data points each cluster should have.

- There is ***at least one item*** in each cluster
  - If a cluster is empty, the algorithm will search for the sample that is farthest from the centroid of the empty cluster. Then it will reassign the centroid to be this farthest point.

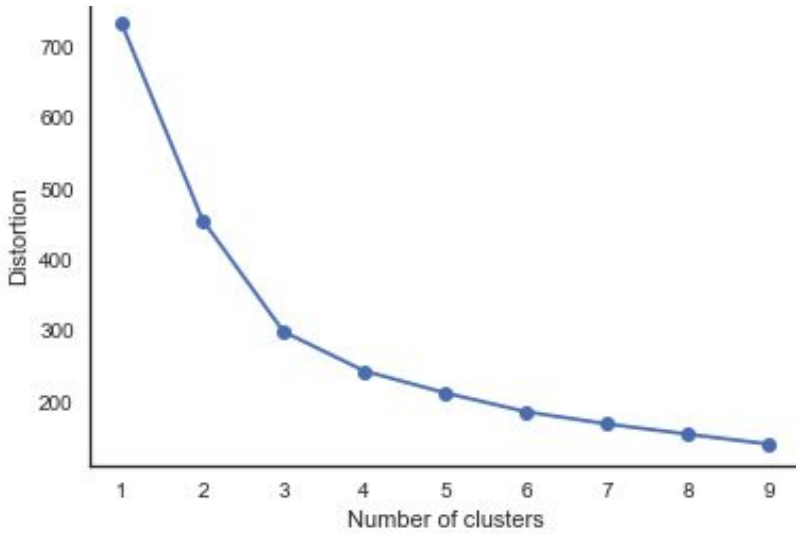
- Clusters are defined by taking the ***mean of all the data points***
  - With this assumption, one can start with the centers of clusters anywhere. Keeping the starting points of the clusters anywhere will still make the algorithm converge with the same final clusters as keeping the centers as far apart as possible.

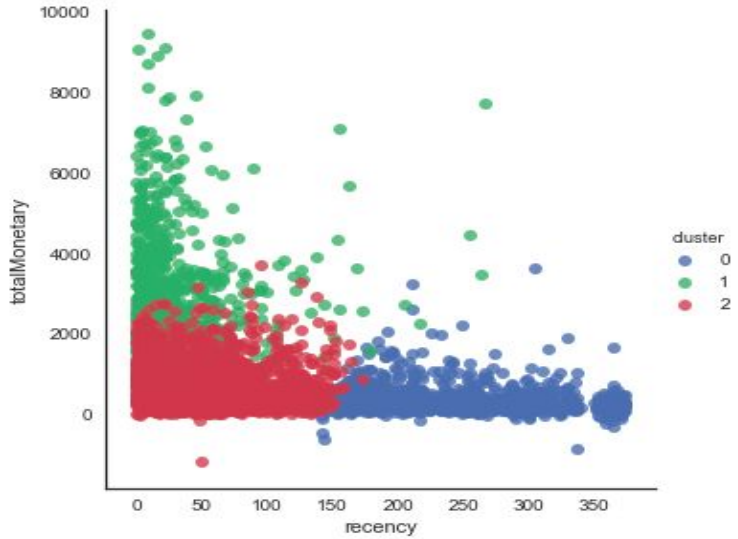
Modeling Approach

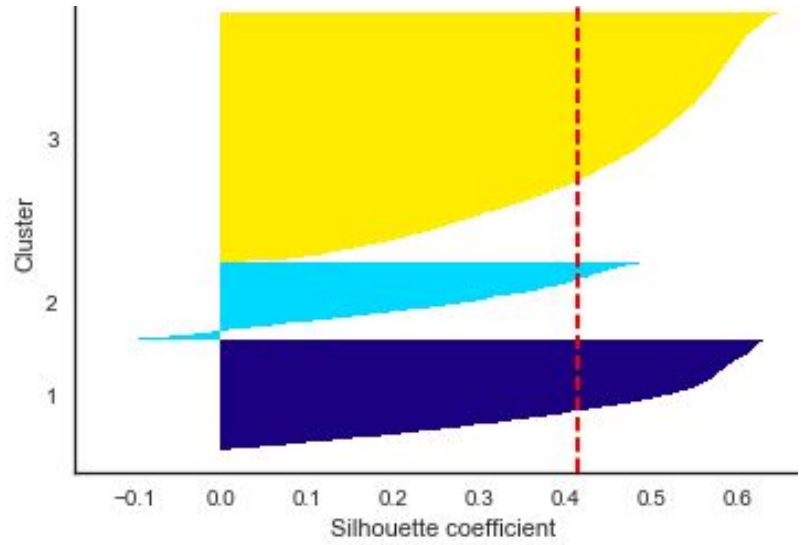


Analysis Results

Elbow method analysis → **Cluster formation using KMeans** → **Silhouette Plot Analysis for clusters**





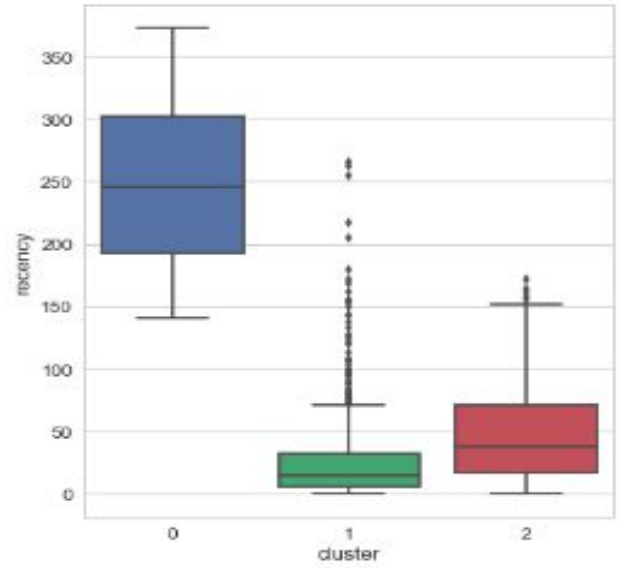


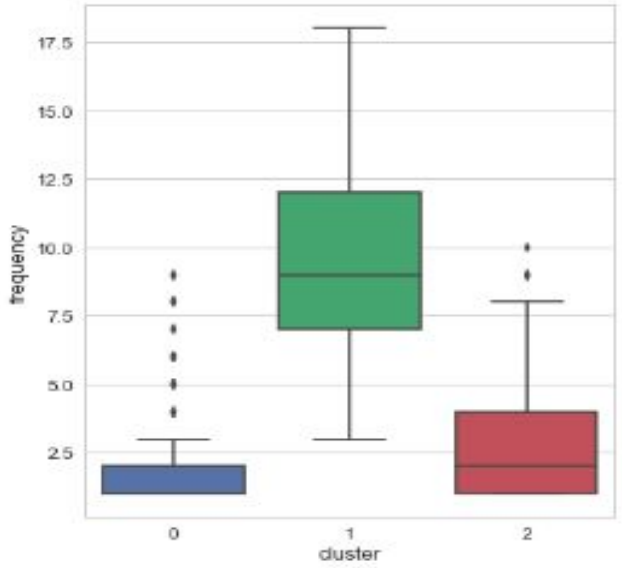
Challenges/Limitations & Improvements

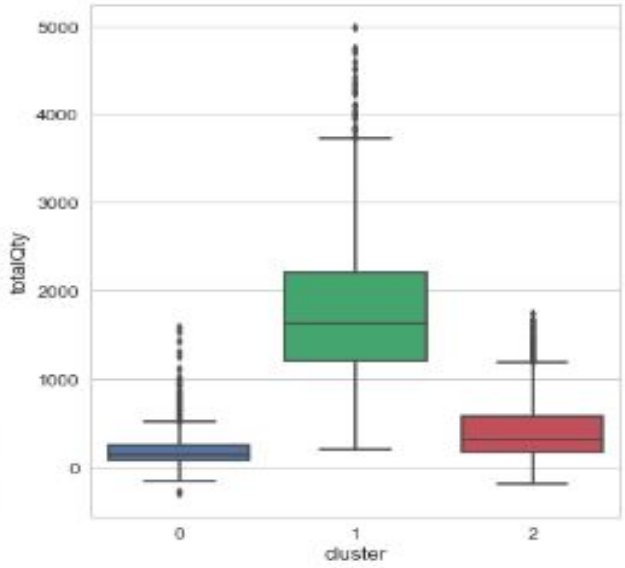
- As seen from the silhouette plot, the clusters formed are uneven. In its quest to minimize the within-cluster sum of squares, the k-means algorithm gives higher weight to larger clusters
- RFM customer segmentation only takes historical data points in consideration while there are advanced techniques like predictive analytics that use AI to predict future customer behavior.
- We have to specify the number of clusters *k* a priori, which may not always be so obvious in real-world applications, especially if we are working with a higher dimensional dataset that cannot be visualized. In such cases, hierarchical clustering works better
- Doesn't take into account the relationship between different features & clusters in the model
- Highly susceptible to outliers and extreme values. Currently, we have treated them in our code based on empirical rules but need to build a systematic approach based on business rules
- Variations of RFM metrics which include capturing digital activity - Duration, Engagement
- Including more variables - item preference analysis from the transactions, business lifecycle, number of employees, financials, industry served etc. might enhance the customer segments

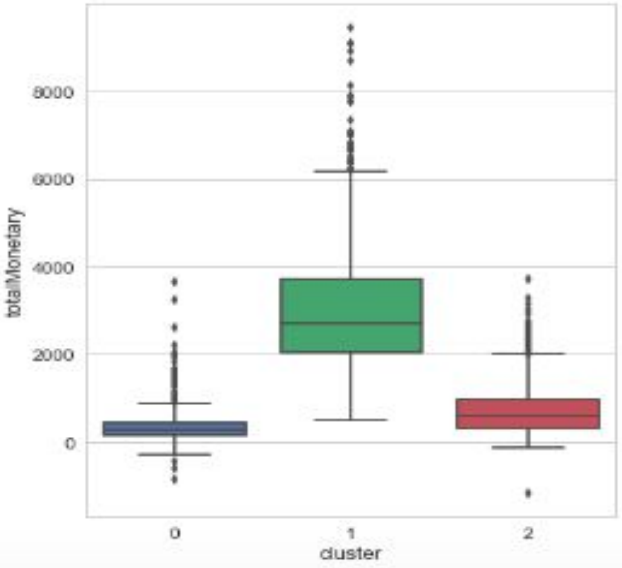
Business **Insights**

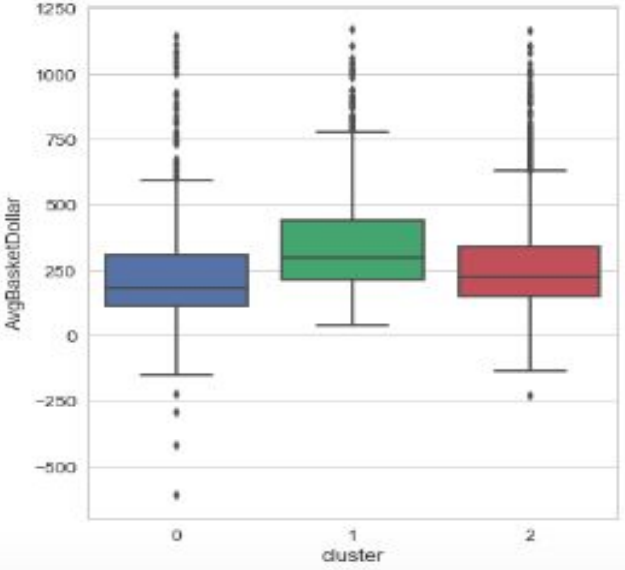
Business Insights

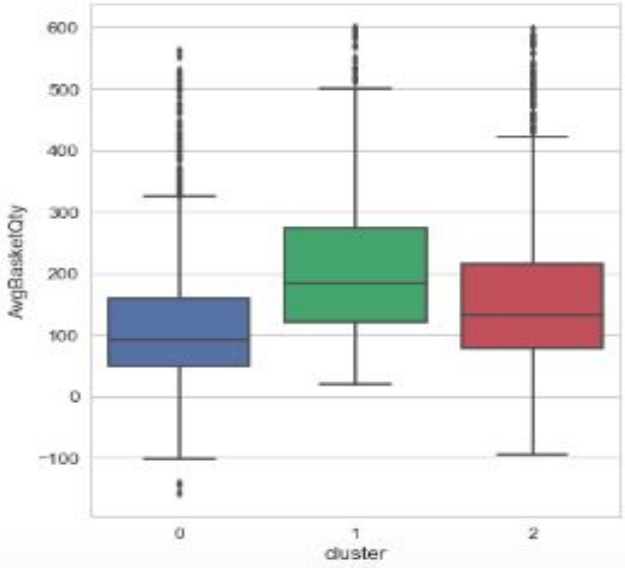












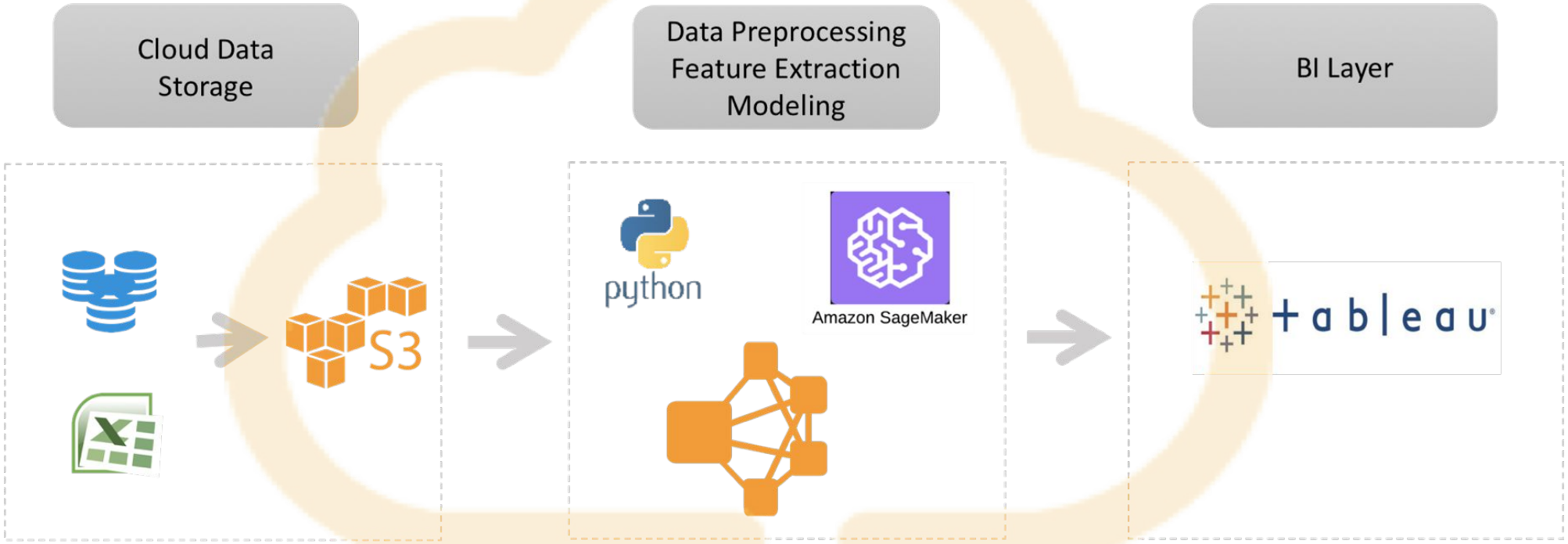
Recommendations

Cluster	Recency	Frequency	Monetary Value	Basket Size (#)
0	High	Low	Low	Low
1	Low	High	High	High
2	Medium	Medium	Medium	Medium

- Customers clearing all the three cut-offs (RFM) are the best and the most reliable customers. Business should focus on making customized promotional strategies and loyalty schemes for these customers in order to retain this valuable customer base.
- Customers failing the recency criterion only are those customers who have stopped visiting the site. Business should focus on these customers and look out for the reason why they abandoned visiting the site.
- Customers clearing the recency criterion but failing frequency criterion are the new customers. Business should provide more incentives and offers to these customers and try to retain these new customers.
- Apart from segmenting customers, business can also use RFM criterion to filter out a reliable customer base and perform analysis like Market Basket Analysis to see customer buying pattern or assess the success of marketing strategies by analyzing the response of these customers.
- Company can run A/B testing on these segments in case it's making any changes in it's services or products to see how the segments react

Beyond **Insights**

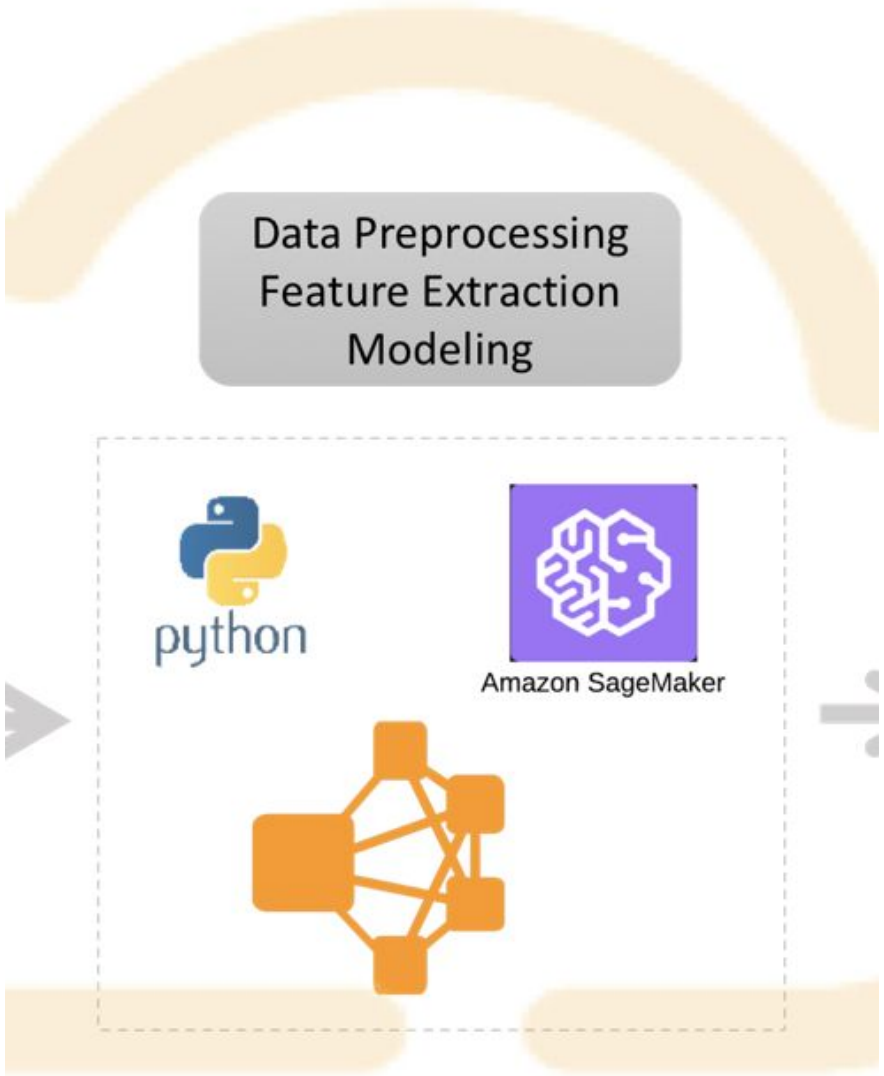
Continuous BI Extraction



Continuous BI Extraction - Contd.

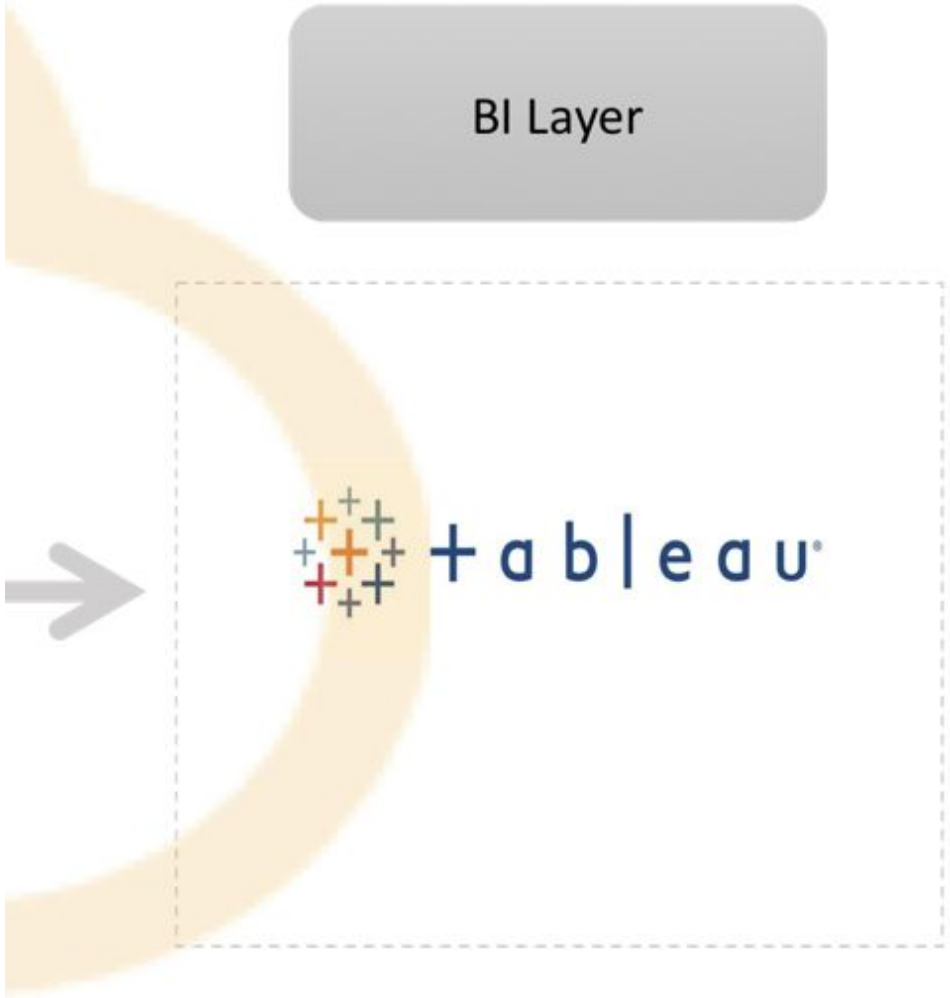
- Sales and marketing teams often require data from several sources both internal to the company as well as external market research data, competitor and customer data from third party etc. to create effective campaigns and sell more to their customers
- This part of the pipeline helps consolidate all sources to single cloud repository to serve as single source of truth
- Everyone from business leadership, sales, marketing, analytics, data science and engineering teams can have one view of this data in real time

Continuous BI Extraction - Contd.



- Heavy engineering, data science and machine learning on big data such as this require huge computational resources easily available with cloud infrastructure which comes with high scalability and reliability
- Continuous Business Intelligence requires real time data modeling and analysis through machine learning which is possible with this cloud pipeline
- Specifically, to minimize lags or delays between the time the data is available and the time a prediction is made for instance, advanced techniques such as decoupled system architecture and asynchronous API calls could be used

Continuous BI Extraction - Contd.



- This is the BI extraction layer accessible and usable to business, analysts, sales and marketing teams etc.
- Providing specific business insights and enabling self-service analytics

Decision Support - Some Use Cases

- Tiered direct marketing campaigns for different customer segments. For instance, product sampling with personalized messages to high valued customers, whereas email marketing for low valued customers
- Decide which segment to drop altogether from marketing as the cost to market and sell them may be more than their buying potential
- Coupon Optimization to maximize revenue, redemption rate, customer loyalty etc. for targeted customer segments
- Decide which marketing channels to leverage for which customer segment

- Reallocate sales support appropriate for each customer segment based on their value and potential

Decision Support - Some Use Cases - Contd.

- Enable more accurate demand forecasting for different customer segments and hence facilitate better inventory management decisions in the overall supply chain

- Second level market basket analysis followed by RFM analysis directly provides insights to make decisions for designing and building association rules to create more efficient recommendation engine algorithms

- Assess the opportunity and risk associated with different customers based on their customer value from RFM analysis and reach out for support, and cross-sell/ up-sell based on their customer profiles

Making the system autonomous, self-learning, and evolving

Deep Learning for Clustering

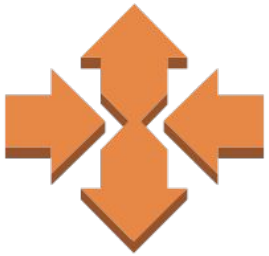
- Using advanced deep learning clustering techniques such as Deep Embedded Clustering (DEC), Deep Clustering Network (DCN)

Self-Healing Systems

- Assigning warning parameters a threshold value in self-learning algorithm and keeping a default saved training model to be applied in case of anomalies with the system can make it self-healing

- Embedding an alarming system in case of any issues within the system and a cap on failed self-healing attempts

Making the system autonomous, self-learning, and evolving



Autoscaling, Reliability, Availability

- Enabled through cloud data pipeline

Bots to the rescue for marketers

- Embedding natural language processing algorithms such as RNN-LSTM so the system can recommend more effective text to marketers for designing better campaigns such as in email marketing based on the selected customer segment

- Embedding convolutional neural network algorithm for image processing for the system to recommend more effective images to marketers for designing better campaigns

Answers to The Key Business Questions

- What are the criteria to define each segmentation?
  - We used the recency, frequency, and monetary attributes to segment customers into 3 groups.
- How will the segmentation be done?
  - We applied K-Means clustering to train and test the segmentation process.
- How feasible is it to “standardize” segmentation across business?
  - The process can be automated through a complete data pipeline from sourcing data to clustering/classification to visualization.
- How will Company X benefit from the segmentation?
  - Company X can use the segmentation information to decide which group of customers are more valuable, and create specific market campaigns to target specific group of customers.

Future is AI enabled marketing campaigns!

Thank you!

Annexures

References

- https://www.kaggle.com/esthergloriadawes/customer-segmentation/notebook
- http://www.kimberlycoffey.com/blog/2016/8/k-means-clustering-for-customer-segmentation
- http://support.sas.com/resources/papers/proceedings12/286-2012.pdf
- https://openreview.net/pdf?id=B1CEaMbR-