Customer Segmentation

Using an Extended RFM model with K-Means algorithm

BY: SHAHAD ALSHALAN

One Size Does Not Fit All
Customer Segmentation

Every buyer has individual **preferences**, **needs**, and **behavioral** patterns.

The marketing strategies have to be customized to meet the needs and requirements of each and every individual.

Customers Segmentation is the process of dividing customers into set of individuals with **distinct similarities**.

It is a part of various activities under **Customer Relationship Management(CRM)**

It is used to inform several parts of a business, including product development, marketing campaigns, direct marketing, customer retention, and process optimization

Customer Segmentation methods

Statistical Segmentations

Traditional

- Interviews
- Surveys
- Categories
- Focus groups

Supervised ML

- Traditional supervised ML techniques, artificial neural networks and complex ensemble models

Unsupervised ML

- DBSCAN, Kmeans, Hierarchical Clustering, GMM Algorithm

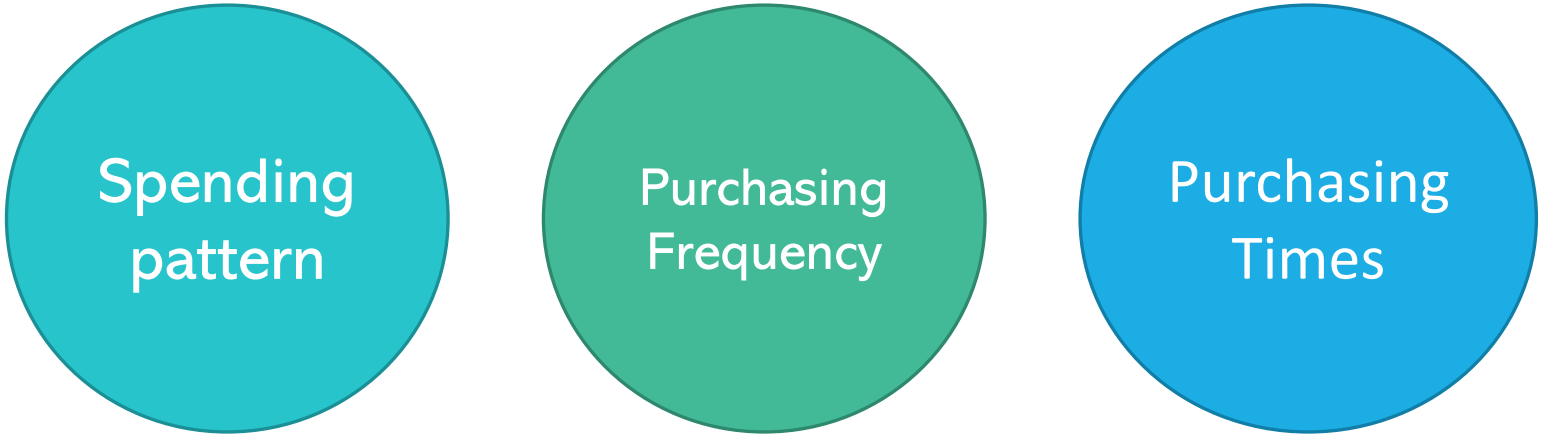
Goal

The objective of this work is to **identify different segments of customers** of Online Retail dataset on the basis of their historical purchasing behavior to run targeted marketing campaign

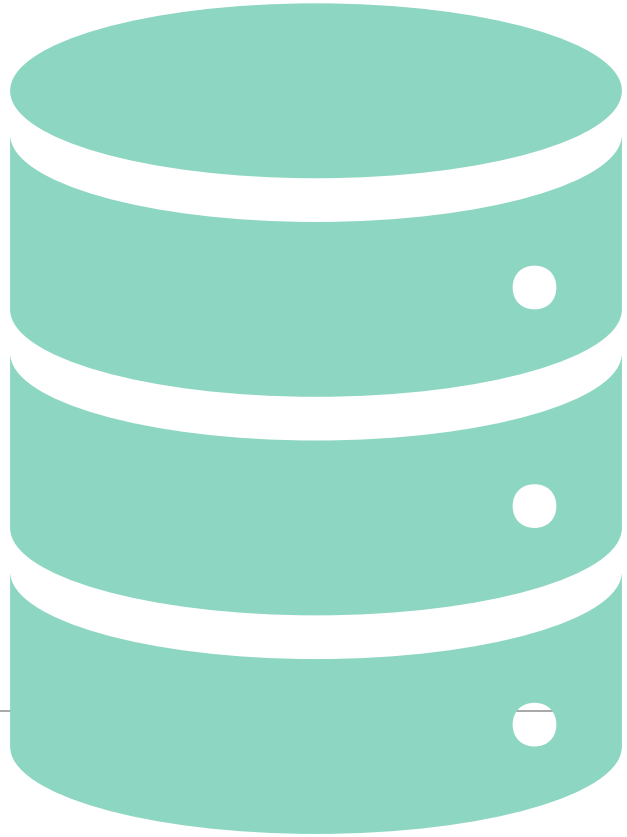
We will use **an extended RFM model with K-means** algorithm to create customer segments

Online retail customer segmentation

We will segment our customers based on:



Dataset



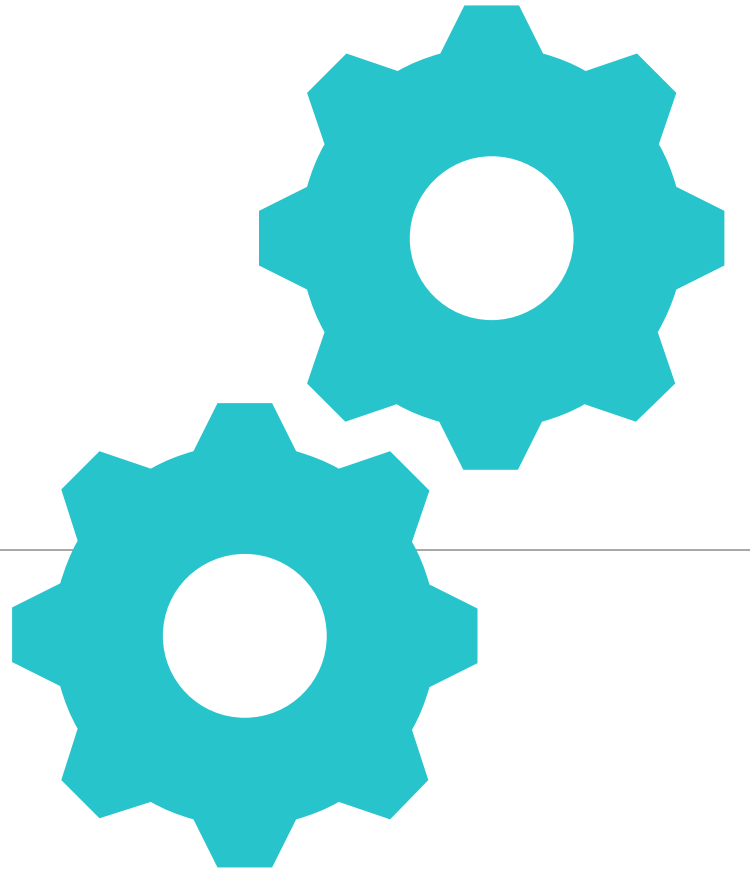
Dataset: Online retail

It is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

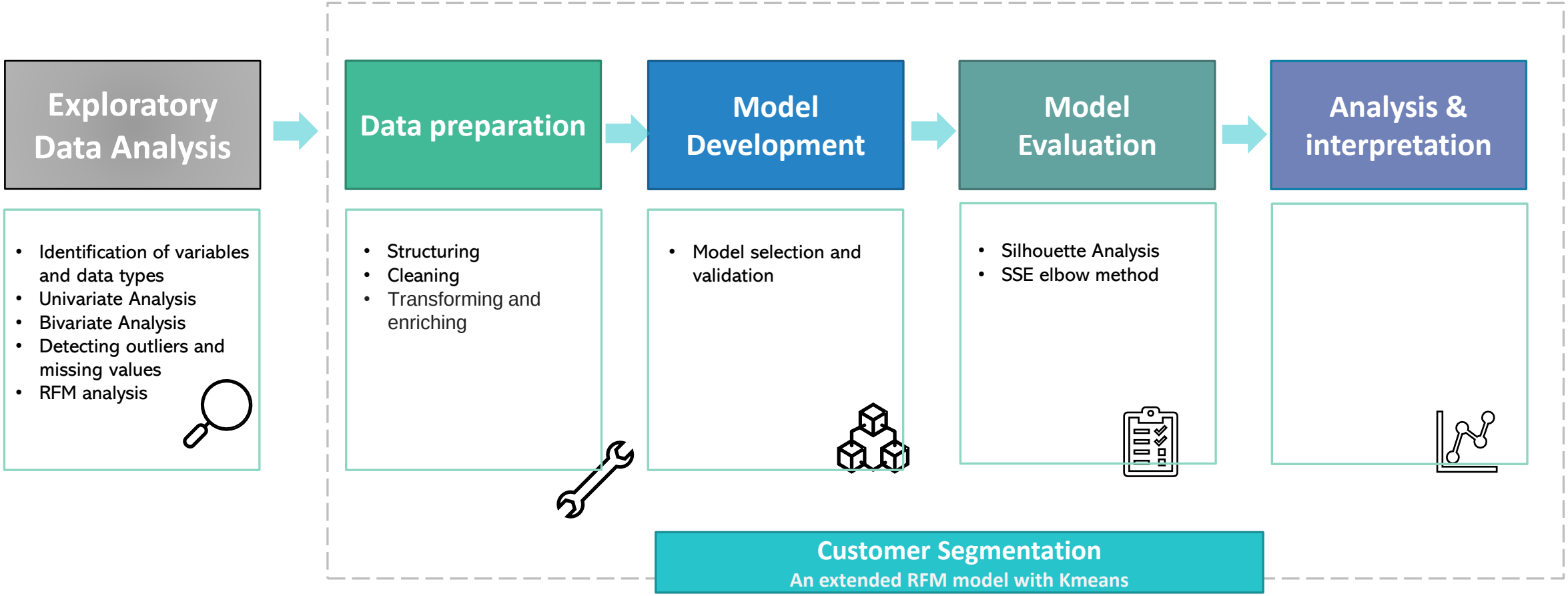
The company mainly sells unique all-occasion gifts.

There are **541909** observations with 8 variables

Methodology



Methodology



Exploratory Data Analysis

Data Description

There are 8 variables in the dataset:

Variable	Description
InvoiceNo	A unique number for each transaction/order
StockCode	A unique number for each product
Description	The product description
Quantity	Number of products purchased for each transaction
InvoiceDate	Transaction's Timestamp
UnitPrice	The price per unit
CustomerID	Unique ID for each customer
Country	Customer’s Country name

Data Key Statistics

Variable	# of nulls	% of nulls	Type
InvoiceNo	0	0	object
StockCode	0	0	object
Description	1454	0.27	object
Quantity	0	0	int64
InvoiceDate	0	0	object
UnitPrice	0	0	float64
CustomerID	135080	24.93	float64
StockCode	0	0	object

Observations:

1. There are some missing CustumerIDs (about 24% of the total)
2. Descriptions has some missing values but StockCode does not.

Data Key Statistics

	Quantity	UnitPrice
count	541909	541909
mean	9.55225	4.611114
std	218.081158	96.759853
max	80995	38970

Observations:

1. Quantity and unit prices has some negative values
2. We found that negative quantity represent about 1.9% of the dataset
3. Most of the records with negative quantity has a stock code that starts with “C”
4. The top descriptions associated with negative quantities are **regency cake stand 3 tier, manual, postage, jam making set with jars and discount**
5. We can see that most of these stock codes represent special transactions

Customers

Total Number of Customers

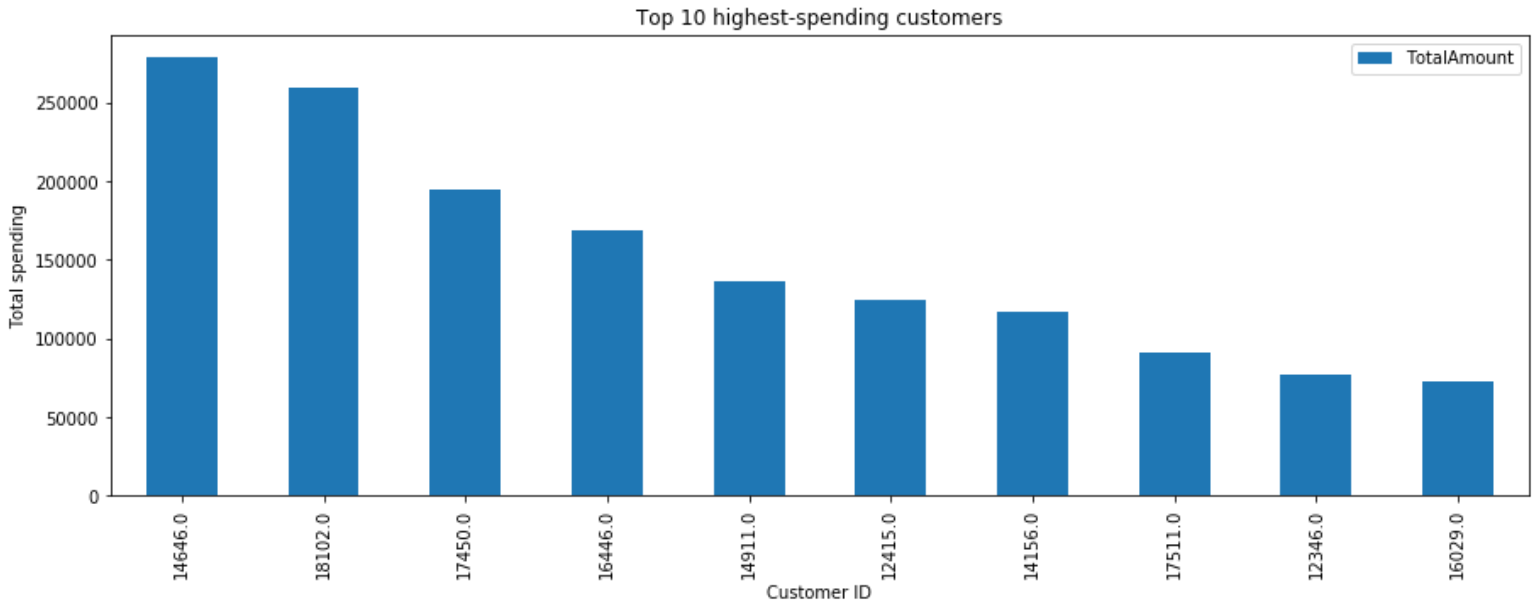
There are 4372 total number of customers in the dataset (before removing missing and incomplete data)

And 4334 after removing missing and incomplete data

Biggest Spending Customers

The biggest 5 spending customers:

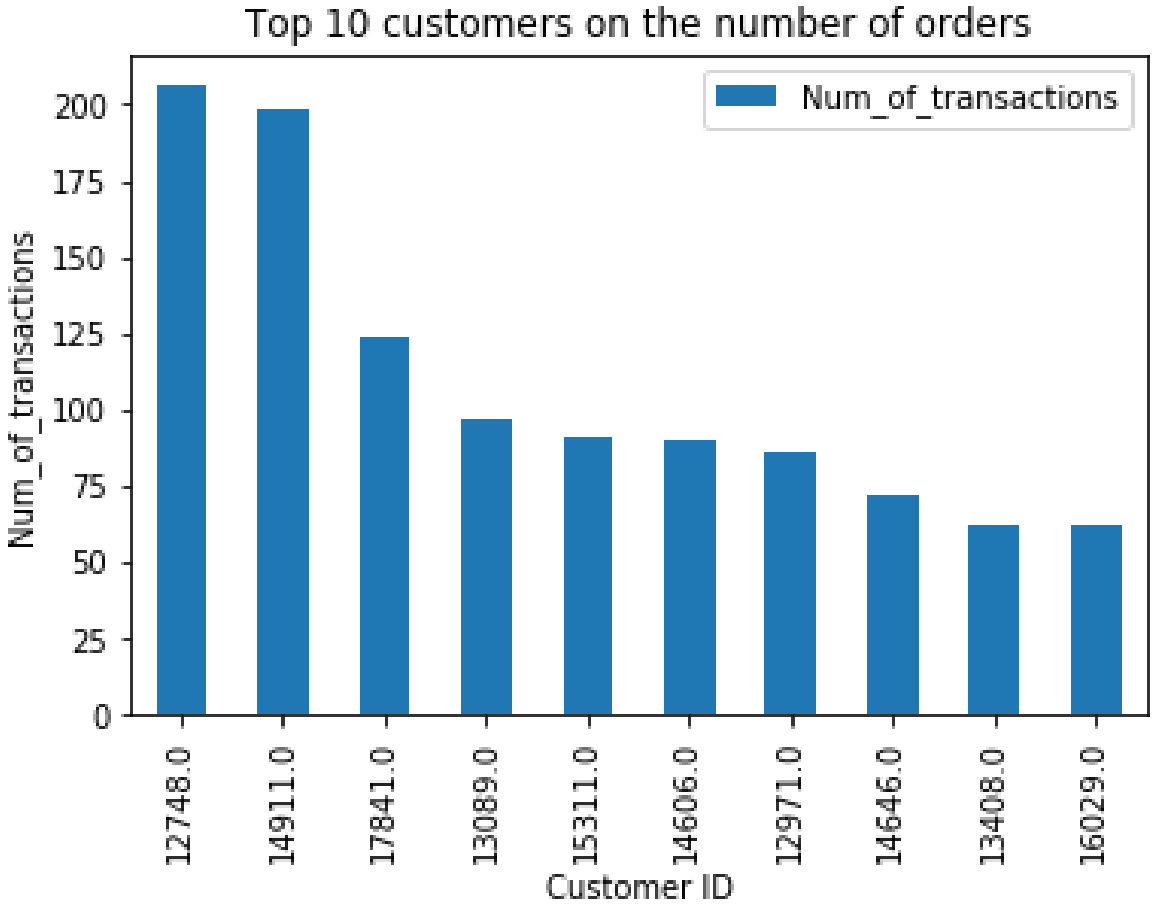
1. 14646
2. 18102
3. 17450



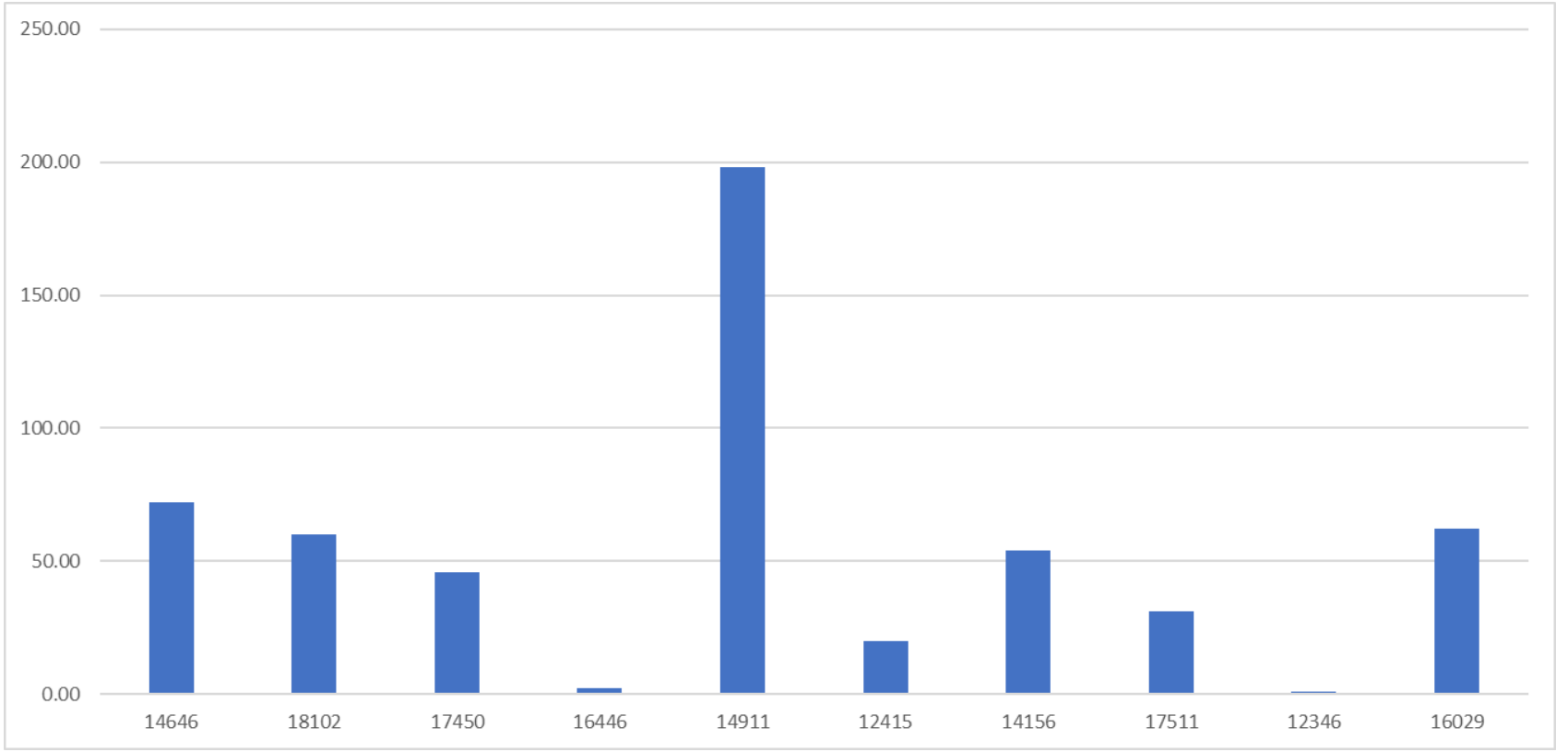
Who place the highest number of orders?

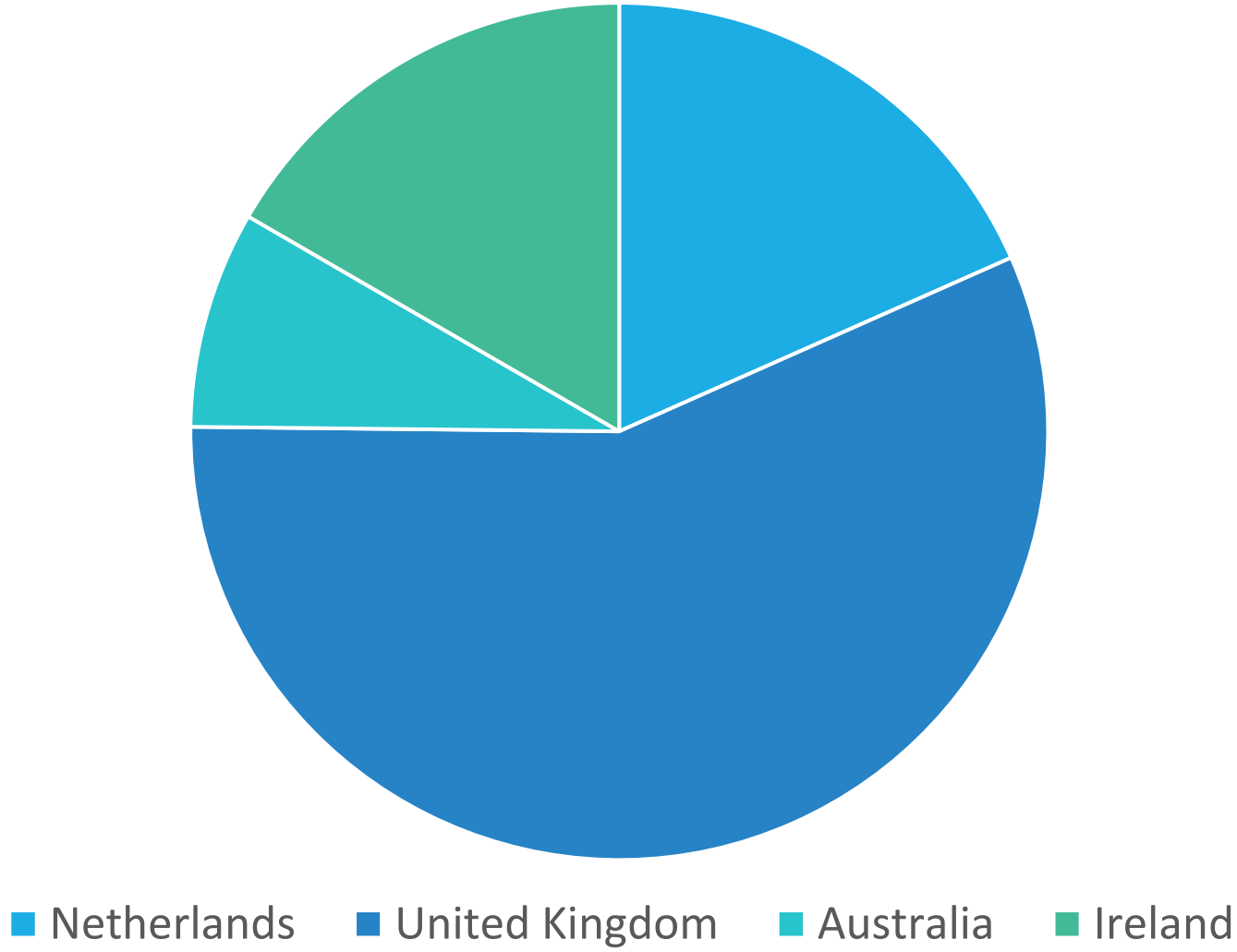
The Top 5 customers with highest number of orders:

1. 12748
2. 14911
3. 17841



How many **orders** did the biggest spending customer actually place?





Where are they form?

Do they have favorite month, day and time?

The most frequent day: Wed

The most frequent hour: 11

The most frequent month: 11

Heights spending vs Heights Number of orders
Observations

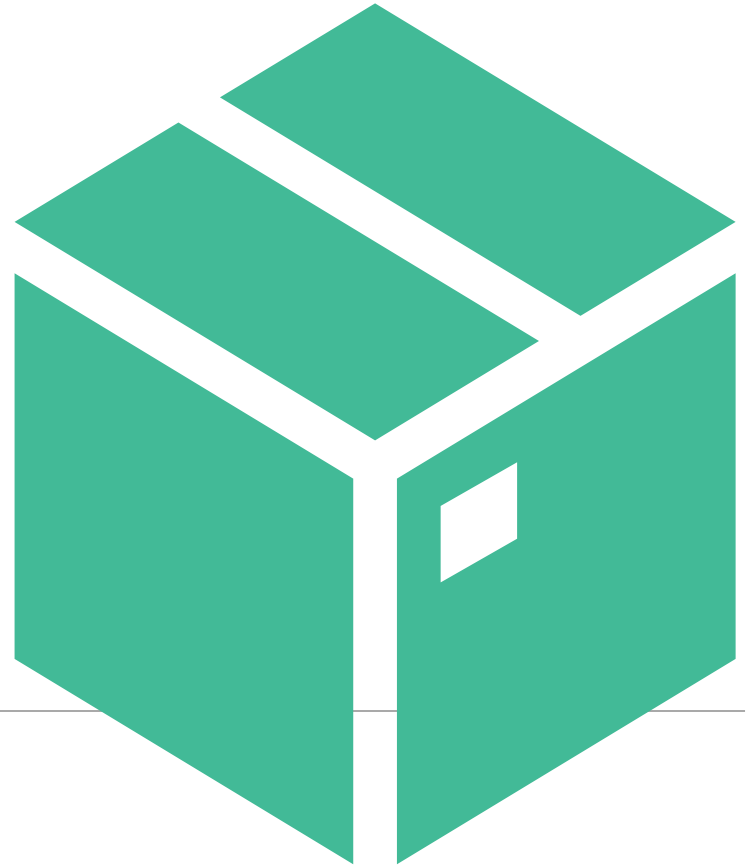
Most of the biggest spending customers are not the customers with highest number of orders

The overlap between the top 10 spending customers and those with highest number of orders are only **3** customers

Our biggest spending customer actually has only **1** order

We expect to see a small percentage of customers generating a significant percentage of the total revenue

Orders



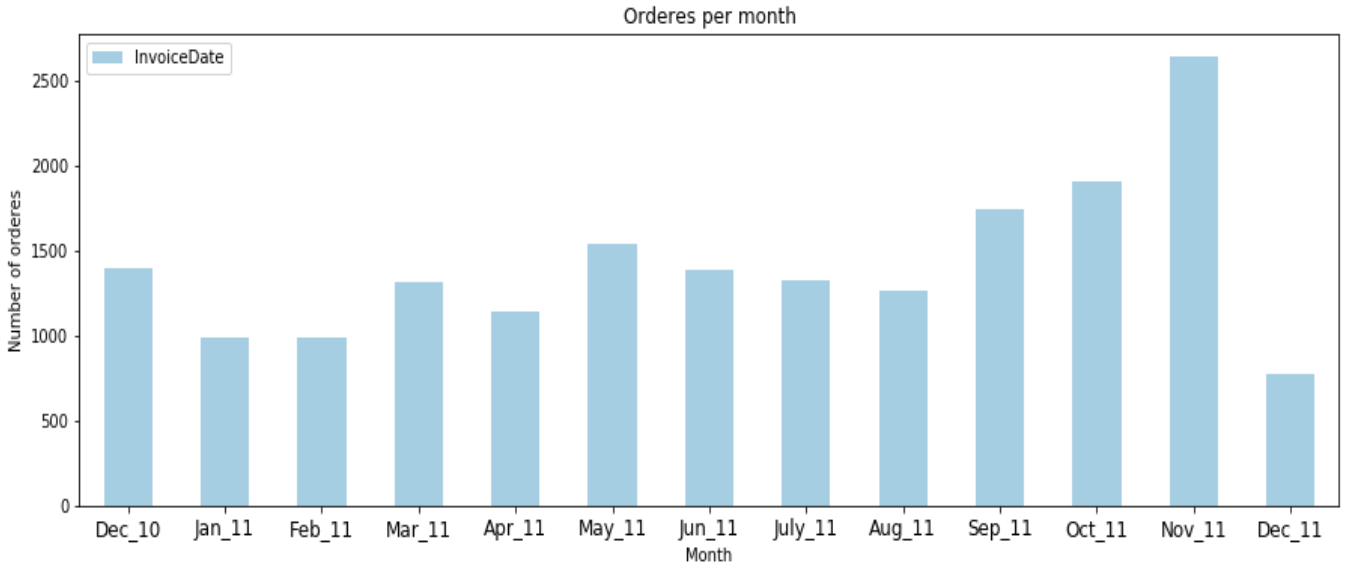
Total Number of **Orders**

There are 25900 total number of orders in the dataset (before removing missing and incomplete data)

And 18402 after removing missing and incomplete data

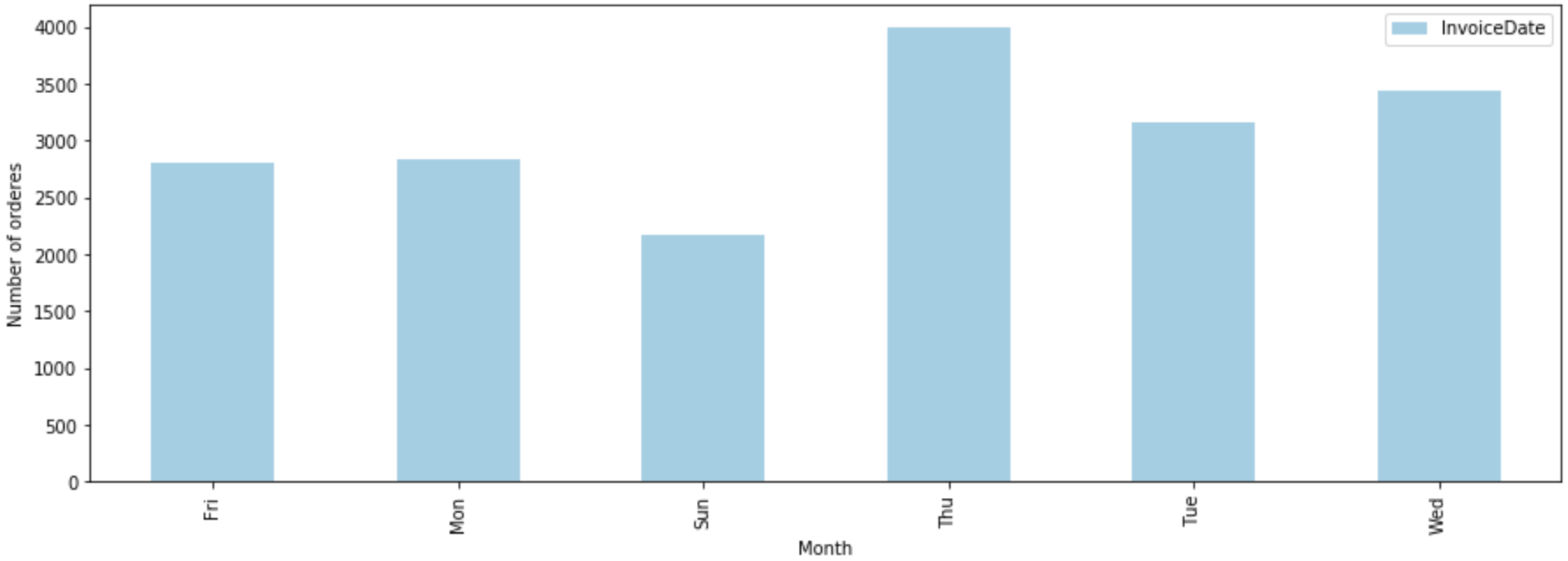
How many orders does the company has per month?

- The highest number of orders are placed in November
- The lowest number of orders are placed in Jan, Feb and Dec*
- The increase in November might be due to people doing **Christmas** shopping early in order to take advantage of **Black Friday** discounts (assuming it is offered by the company).



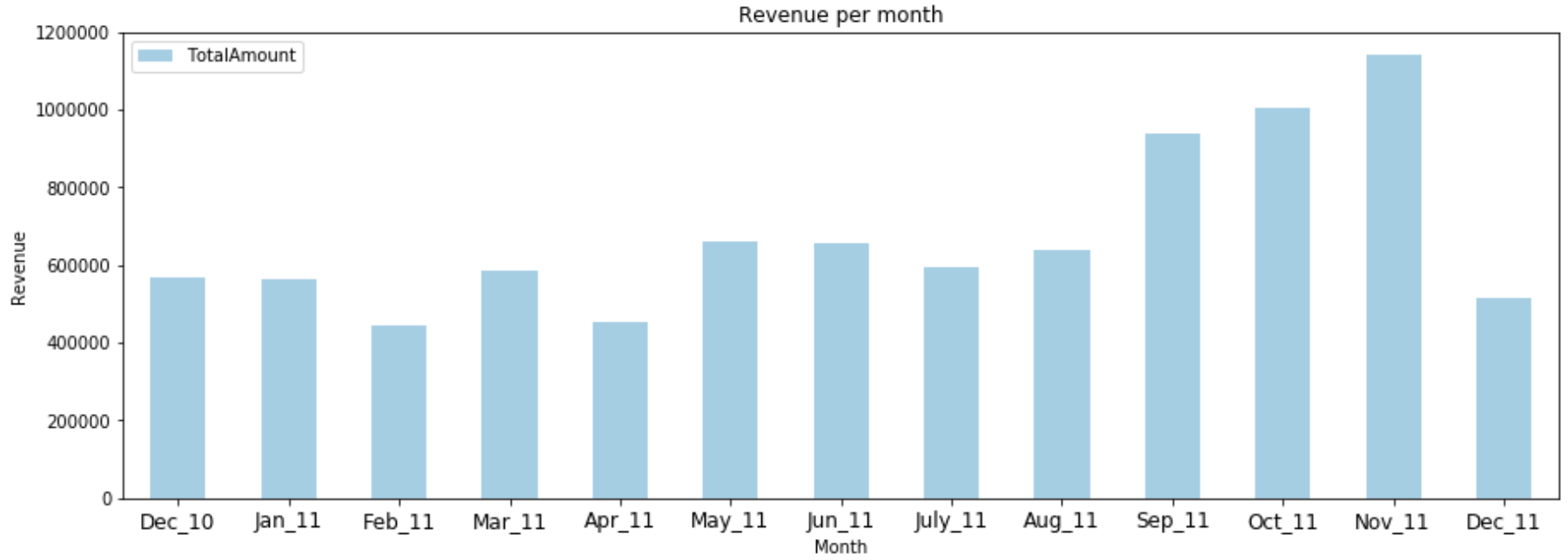
* Transactions data in December are not complete

How many orders does the company has per day?



Revenue

How much **revenue per month** does the company earn?

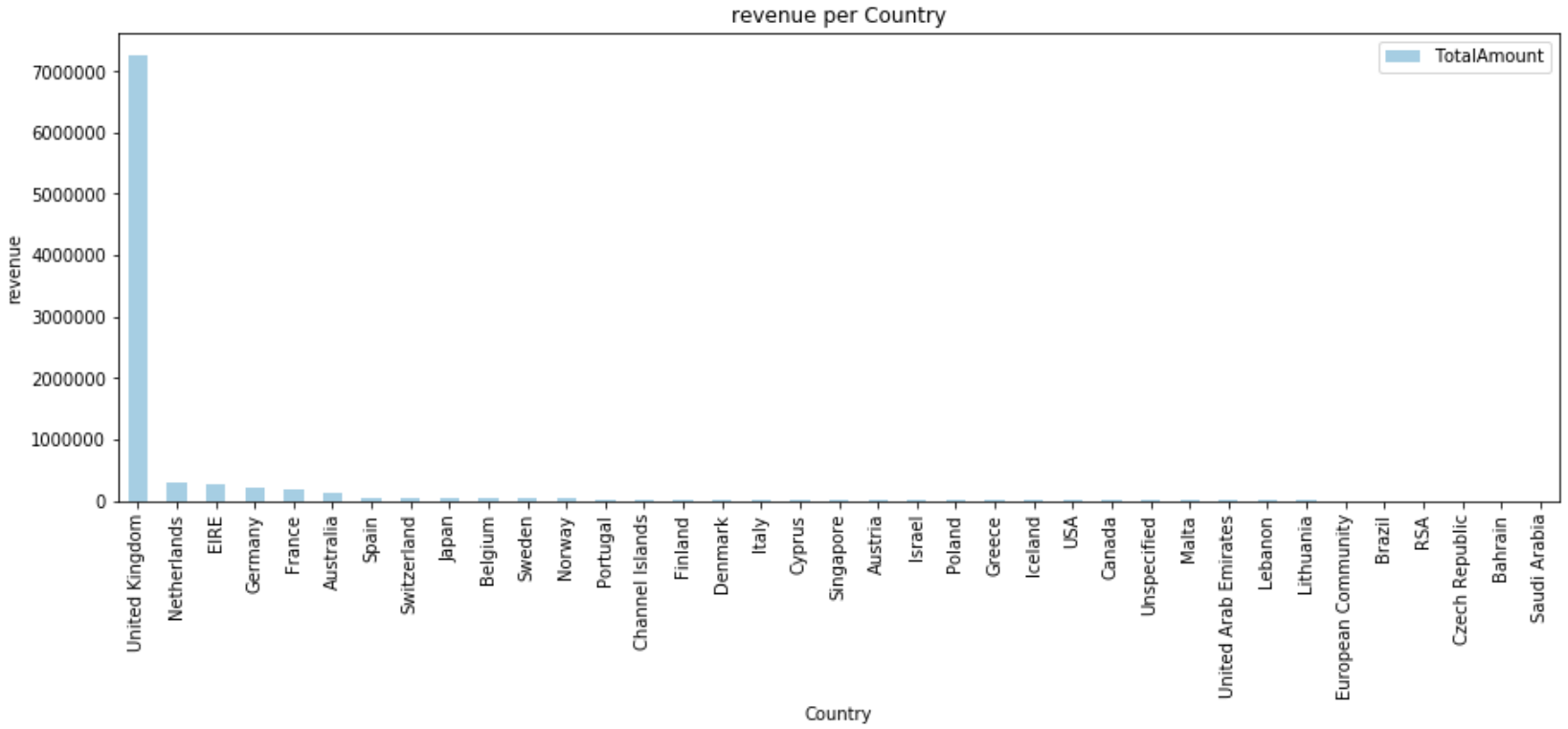


The top month on the earned revenue is November which agrees with the results we found on the total number of orders per month

Which country contribute the most on the company's **revenue**?

Top three countries:

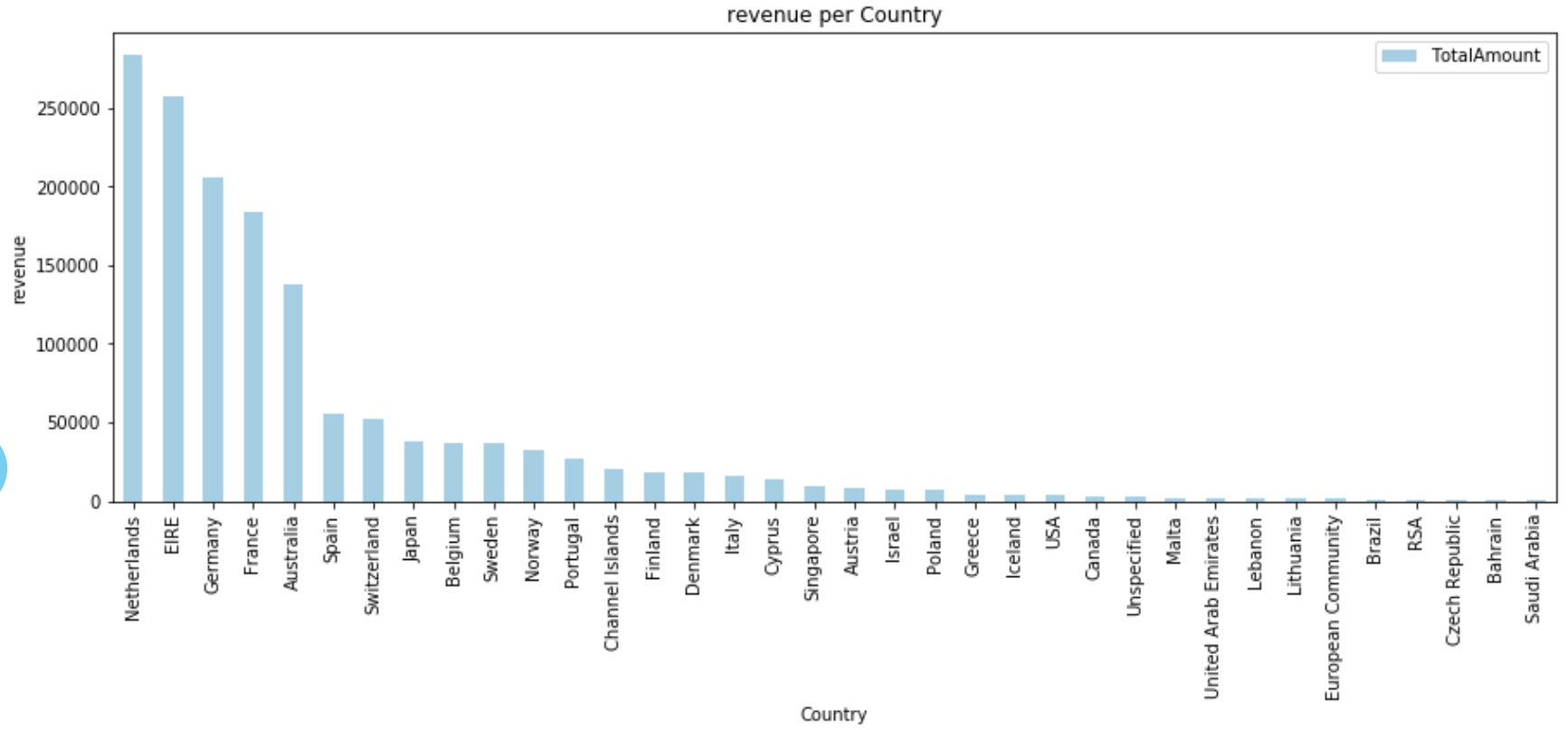
1. United kingdom **(82%)**
2. Netherland **(3.2%)**
3. Ireland **(2.3%)**



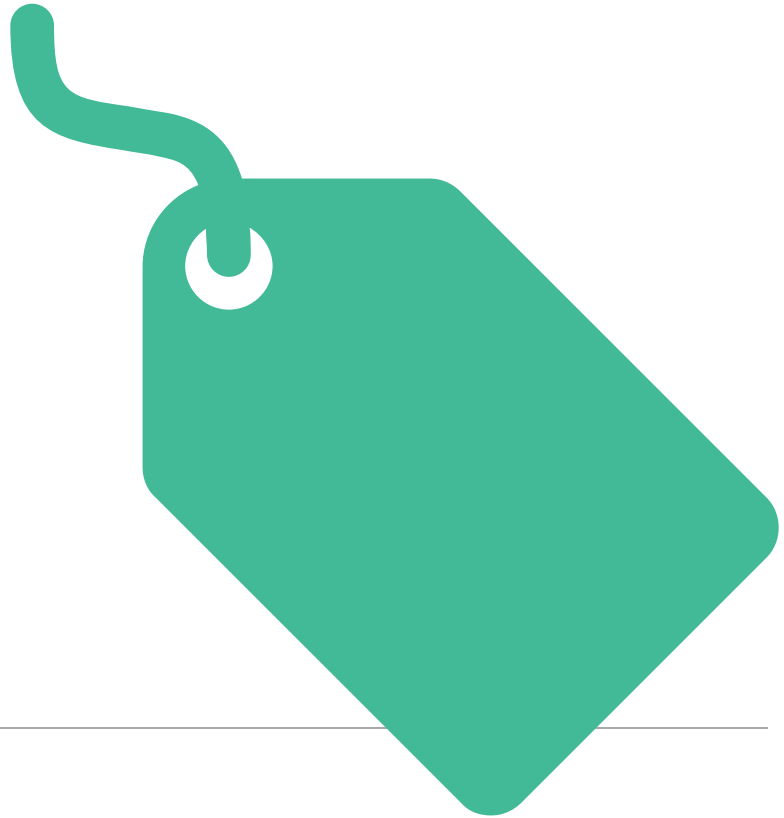
Which country contribute the most on the company's **revenue**?

Top three countries:

1. Netherland **(18%)**
2. Ireland **(17.2%)**
3. Germany (13.7%)



Products



How many **unique products** does the company has?

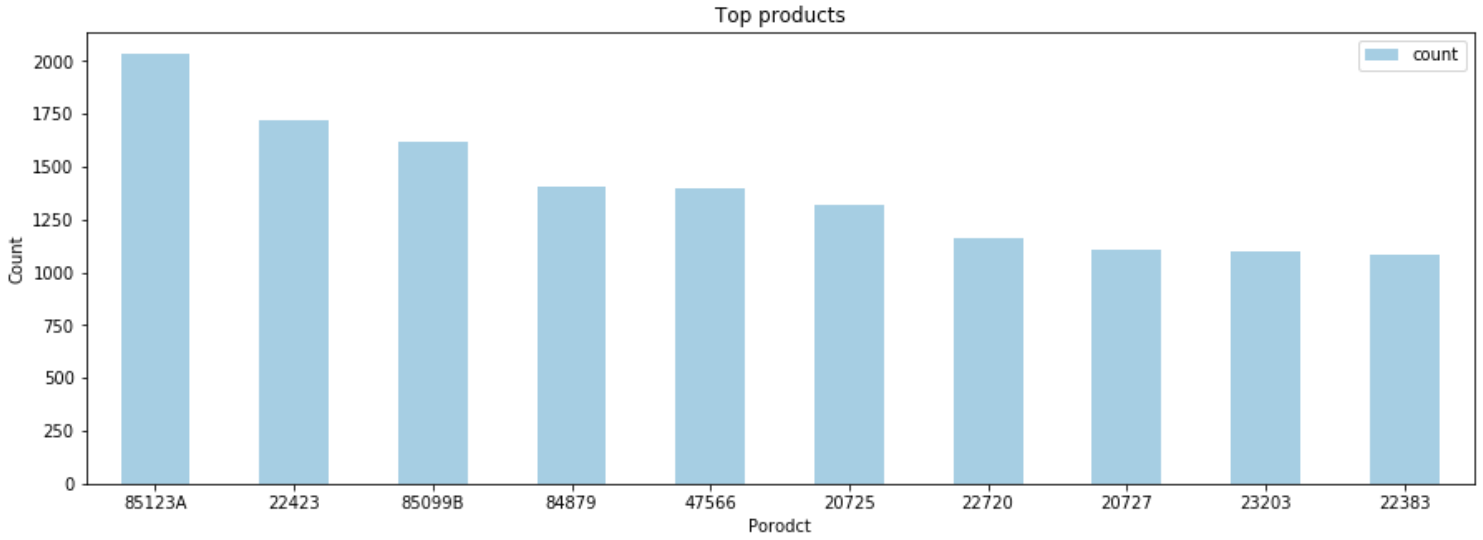
The company has different **2791** unique products

And a total of **3871** different products descriptions

The number of products is less than the number of descriptions, which means that descriptions are not unique and different descriptions can belong to the same product (due to entry errors)

Products description words cloud

What are the best-selling products?



Description of the Top 5 products:

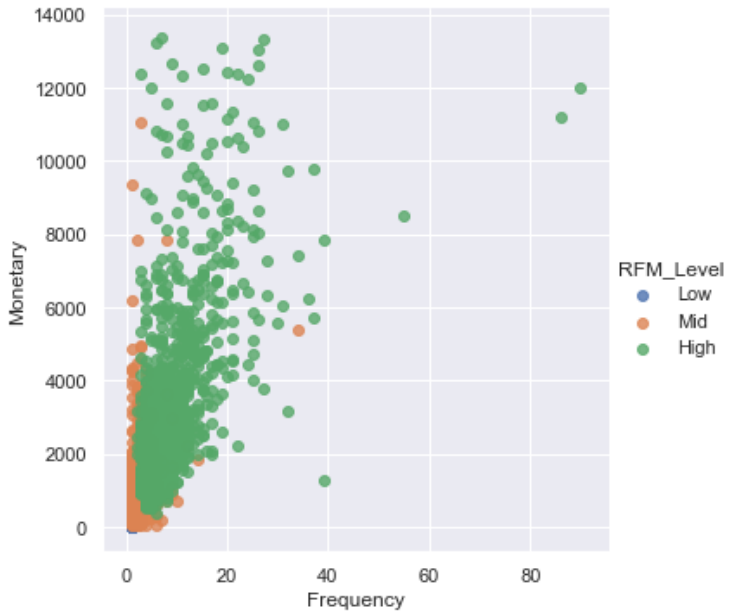
- White hanging heart t-light holder
- Regency cake stand 3 tier jumbo bag red retrospot
- Assorted colour bird ornament
- Party bunting

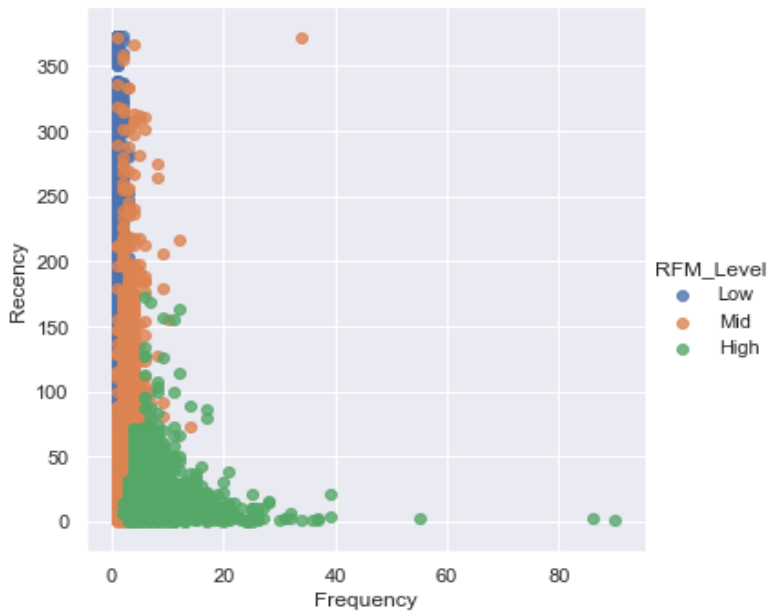
RFM analysis

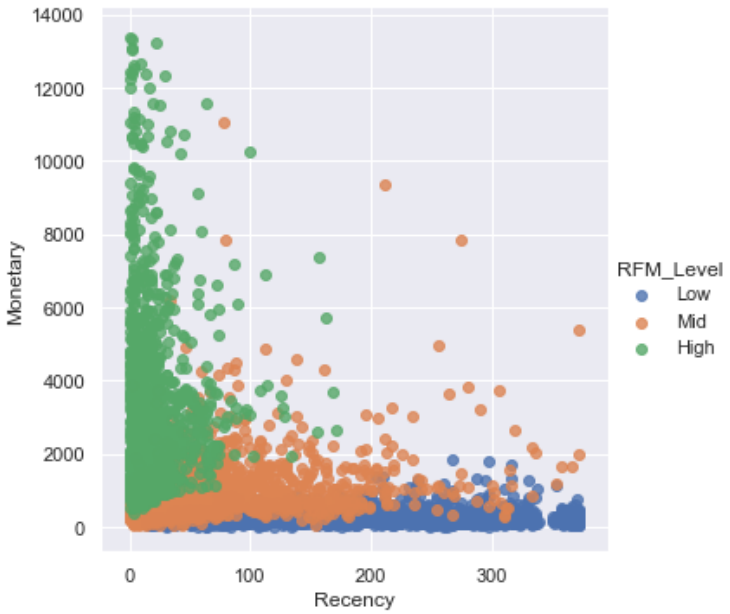
We have conducted a preliminary clustering using a simple RFM analysis as a part of our EDA in order to get insights on our customers purchasing behavior in term of: Spending (Monetary), Number of orders (Frequency) and Recency

RFM analysis

We have conducted a preliminary clustering using a simple RFM analysis as a part of our EDA. The RFM model score the customers based on the three main properties: Spending (Monetary), Number of orders (Frequency) and Recency.







From RFM to an Extended-RFM model

While RFM models can be helpful to develop marketing strategies, it also has some limitations as It uses only a limited number of variables to cluster the customers.

It does not take into account other behavioural properties at the more granular level. For example: **number of products and the average spending per order.**

Therefore, we opted to segment the customers using an extended version of RFM.

Customer Segmentation Model

An extended RFM model with K-means

Data preparation

Data Preparation **pipeline**

It is the key part of building our clustering algorithm for customer segmentations

It involves some manipulation and adaptation of the raw dataset to make it ready for applying kmeans algorithm

The main steps are:

Data Cleaning

1. Remove negative quantities (2.19% of the total)
2. Remove **free items** (with unit prices equals to zero) (40 records)
3. Remove **special transactions**: (0.7% of the total)
   - Manual
   - Postage
   - Discount
   - Bank charges
   - DOTCOM POSTAGE
   - PADS TO MATCH ALL CUSHIONS
   - CRUK Commission
4. Remove missing values:
   - Customer IDs
   - Null Description

Product categorization

Products in the dataset are **not categorized**

We believe that having product categories will help us more to understand our customers **behaviors and preferences**

Since we don’t have any labelled dataset, we use **unsupervised clustering technique** to categorize products

Product categorization

Products in the dataset are **not categorized**

We believe that having product categories will help us more to understand our customers **behaviors and preferences**

Since we don’t have any labelled dataset, we use **unsupervised clustering technique** to categorize products

Product categorization
Feature extraction

We applied two feature extractions methods:

TF IDF

- TFIDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF).
- Each word or term has its respective TF and IDF score.
- The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.
- It is intended to reflect how important a word is to a document in a collection or corpus

word2vec

- **Word2vec** is word embedding technique for natural language processing that uses a neural network model to learn word associations from a corpus of text.
- Word embeddings give us a way to use an efficient, dense representation in which similar words have a similar encoding

Product categorization
Text preprocessing

Before converting text data to vectors using TFIDF or Word2vec, text data need some preprocessing

1. **Text cleaning**
    1. Remove digits
    2. Remove punctuations
    3. Remove stop words
    4. Remove incorrect words
2. **Text normalization**
    - Lower casing
    - Stemming (optional)
    - Tokenization

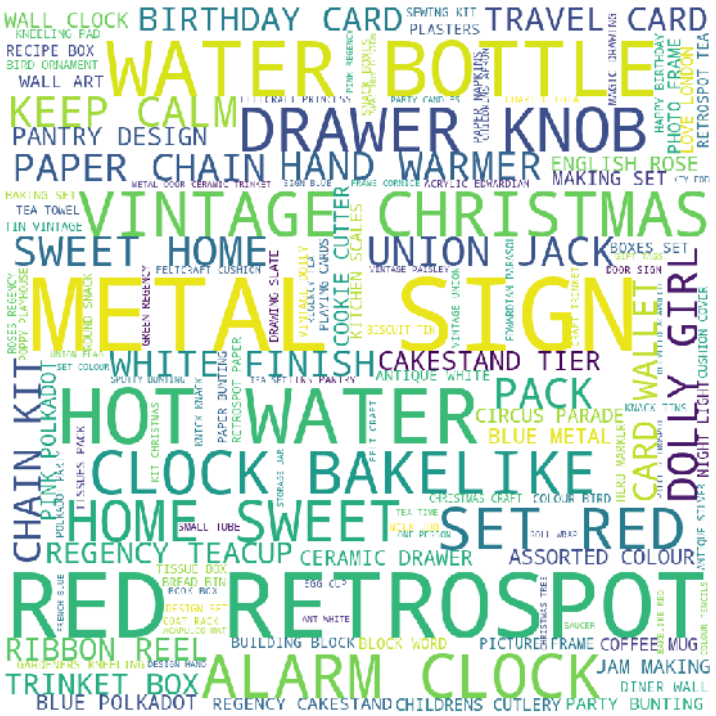
Product categorization
Apply Kmeans clustering

- We applied kmeans clustering method to cluster product categories using both TFIDF and word2vec
- Results from both methods are very similar
- We decided to choose TFIDF due to its simplicity and effectiveness
- With TFIDF, we ended up with **5** products categories

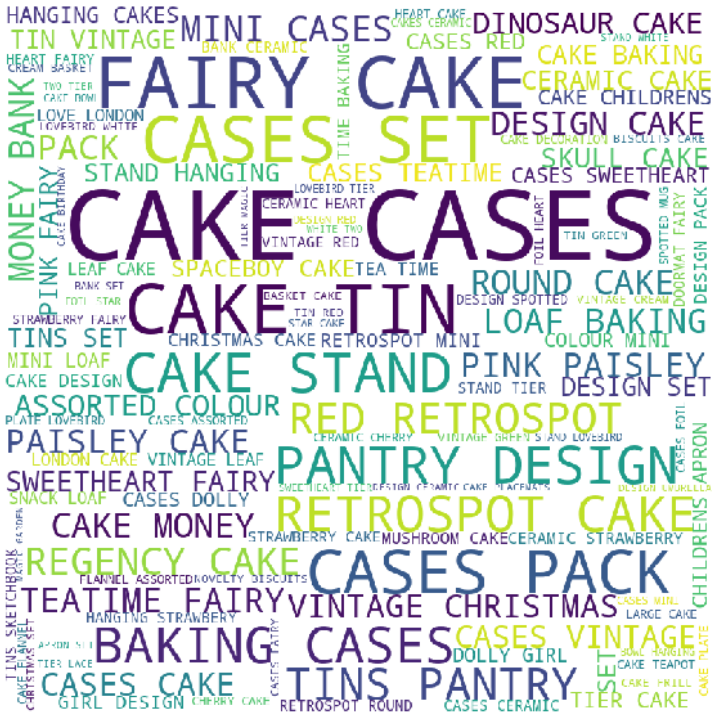
Product categorization

Top words in each product category

Category 1



Category 2



Product categorization

Top words in each product category

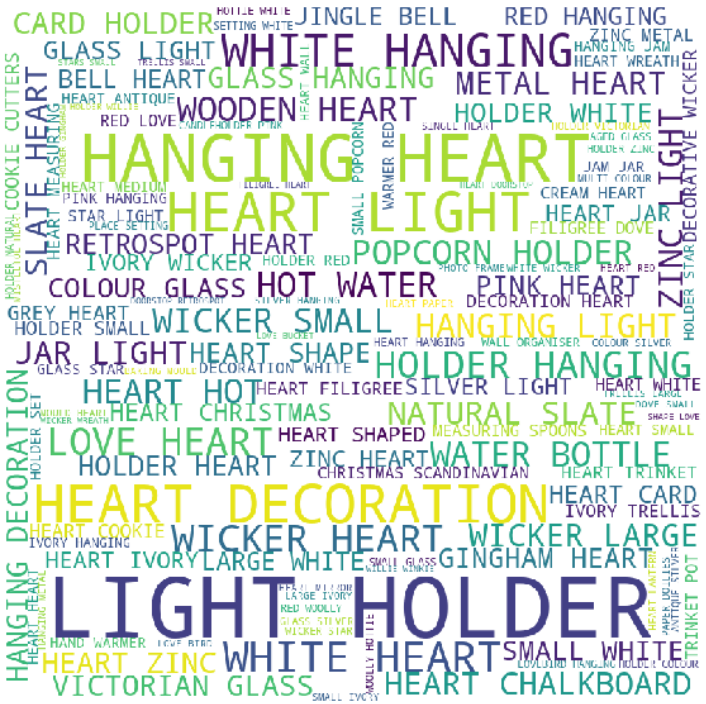
Category 3

Category 4

Product categorization

Top words in each product category

Category 5



Data Transformation

- In this step, the dataset is transformed so that each record represent unique customer purchasing history
- The amount is added for each order product: UnitPrice * Quantity
- Data aggregations performed for each customer as the following:
  - Total spending in each product category
  - Total spending in each month
  - Total spending in all orders (Monetary)
  - Average/max/min spending per order
  - Total number of orders (Frequency)
  - Total number of products (and number of unique products) in all transactions
  - Average number products in each order
  - Recency

Data Transformation

- Kmeans clustering algorithms use distance-based measurements to determine the similarity between data points
- Thus, it's recommended to standardize the data to have a mean of zero and a standard deviation of one
- Since almost all the features in our dataset have different units of measurements, we applied data standardization to our dataset

Final selected Features for clustering

Frequency
Monetary
Recency
Avg Spending Per Order
Number Of unique Products Per Order
Avg Products Per Order

Model development & evaluation

Kmeans algorithm for customer segmentation

- There are different clustering methods can be used for customer segmentation:
- K-Means cluster is one of the most used unsupervised machine learning clustering techniques.
- K-Means clustering algorithm is a is an iterative partition clustering technique that divide the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**.
- Kmeans can be used to predict the clusters for new dataset

Kmeans algorithm for customer segmentation

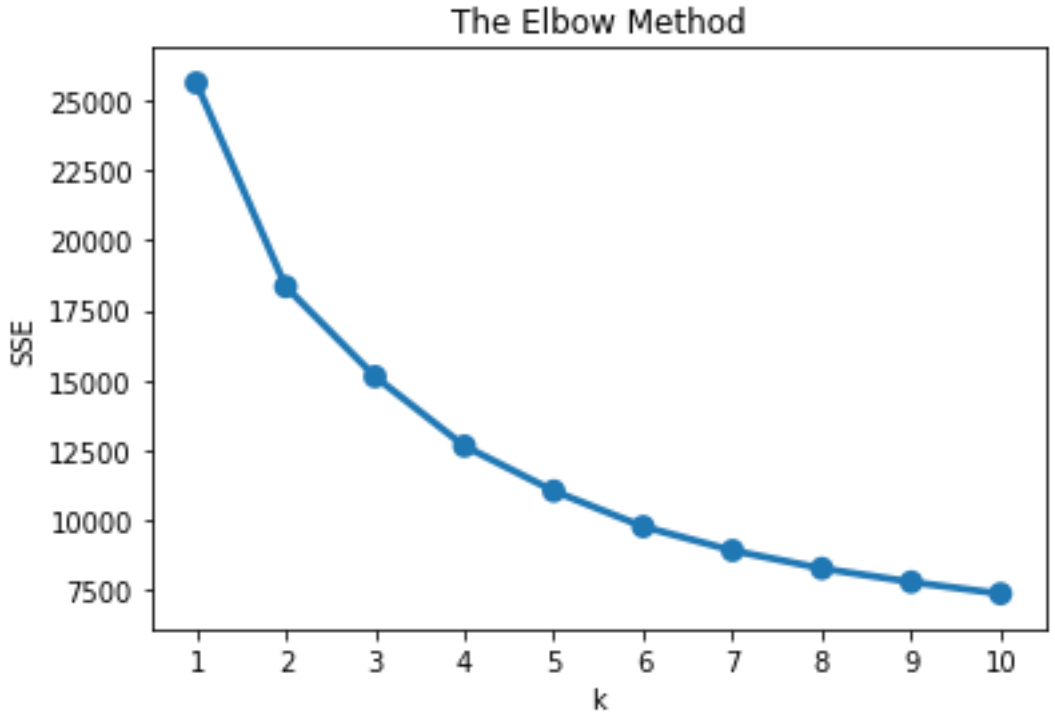
- After conducting an initial clustering with kmeans, we found that it resulted in skewed clusters where more than **60%** are fit into one cluster and only 1 customer is fit into another cluster.
- The final clusters in this way does not give useful insights about our customers behaviors (in business perspective)
- This is because kmeans does not work well with outliers.
- Thus, we decided to exclude outliers and report our analysis on this group separately.

Selecting the Optimal Number of clusters k

There are different approaches to select the optimal number of clusters for a K-Means model.

1. The Elbow Criterion
   - plots out errors and tries to find an ‘elbow’ in the data where the improvement in error versus an increase in a cluster significantly goes down.
2. Silhouette Coefficient
   - calculates how well a certain number of clusters suits a data. Low silhouette score indicate better cluster numbers)
3. Experimentation and interpretation (from business perspective)

Selecting the Optimal Number of clusters k

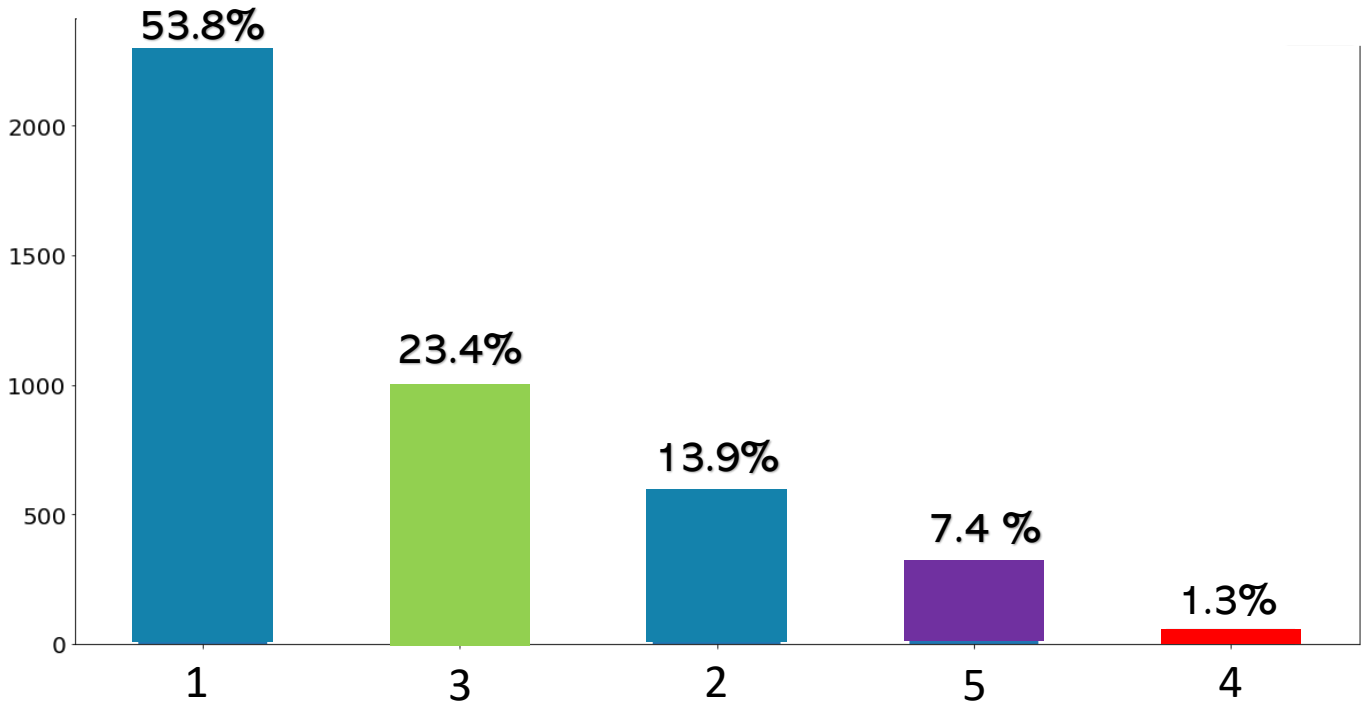



```
Silhouette Score for 3 Clusters: 0.3211
Silhouette Score for 4 Clusters: 0.3426
Silhouette Score for 5 Clusters: 0.3449
Silhouette Score for 6 Clusters: 0.3300
Silhouette Score for 8 Clusters: 0.3378
Silhouette Score for 10 Clusters: 0.3303
```

Using both metrics, we found that the optimal cluster = 5

Analysis

Total Number of Costumers in each cluster



- The highest number of customers are in **cluster 1**
- The lowest number of customers are in **cluster 4**

Where are the customers from in each cluster?

Custer 1	Custer 2	Custer 3	Custer 4	Custer 5
27 Different countries	**21** Different countries	26 Different countries	14 Different countries	9 Different countries
Top 3: **France** **Germany** **Belgium**	**Top 3:** **Germany** **France** **Spain**	**Top 3:** **Germany** **France** **Belgium**	**Top 3:** **Switzerland** **Norway** **Germany**	**Top 3:** **Switzerland** **Norway** **Germany**

Countries are sorted based on the total number of orders

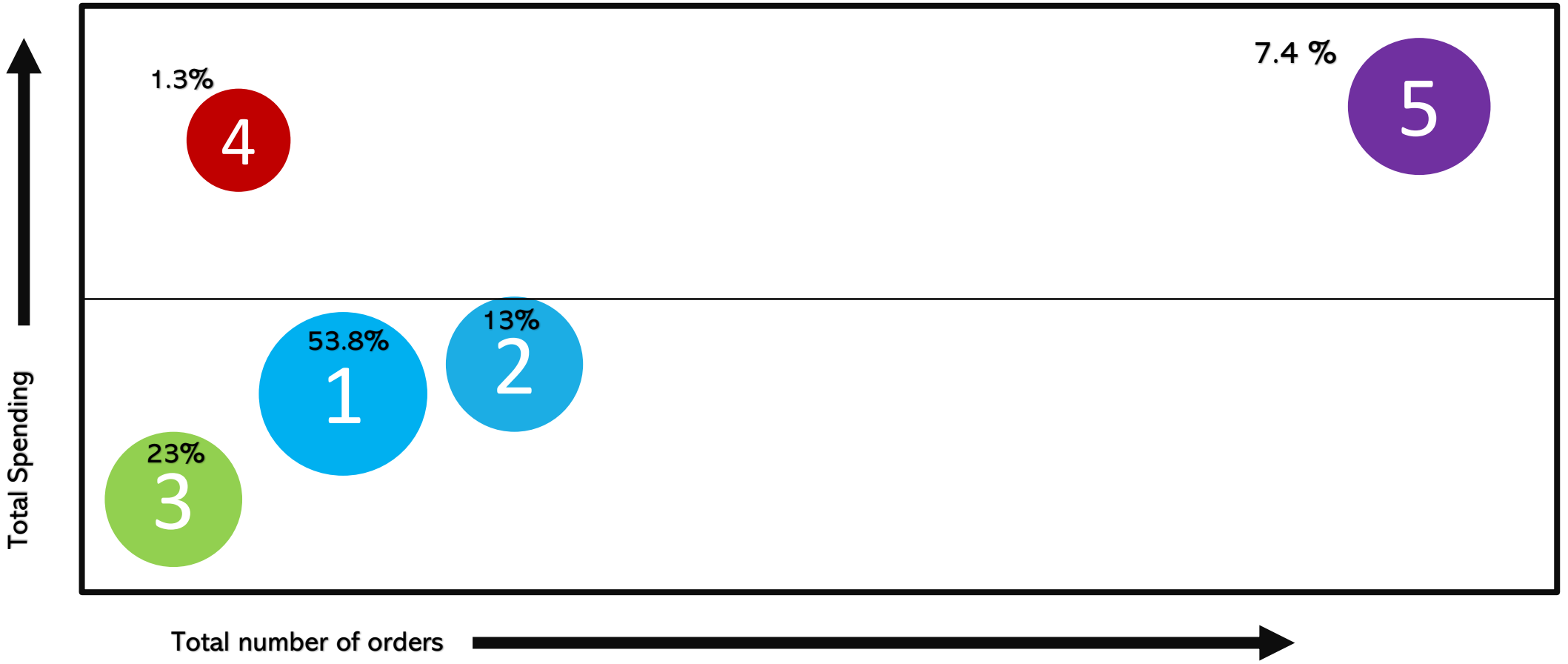
How these clusters are different?

In term of: Frequency, Monetary, Recency, spending and number of products

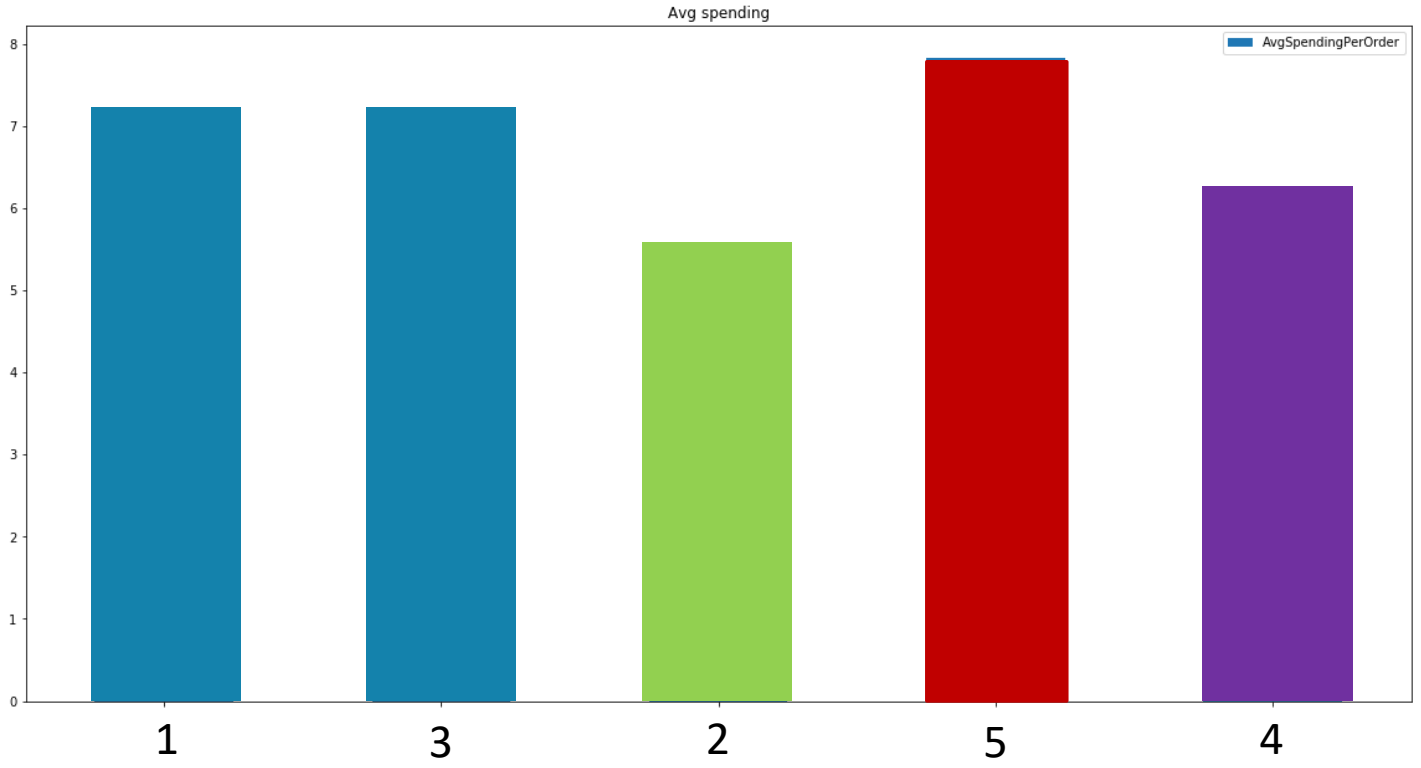
Cluster	1	2	3	4	5
Frequency	3.21	3.56	1.47	2.19	15.40
Monetary	920.75	1653.36	371.66	4885.53	5891.91
Recency	46.56	53.81	249.58	105.67	18.11
Avg Spending Per Order	292.34	521.90	261.98	2516.43	444.03
Number Of unique Products Per Order	38.81	131.23	20.37	81.34	178.26
Avg Products Per Order	16.00	54.70	15.67	49.66	23.35

How these clusters are different?

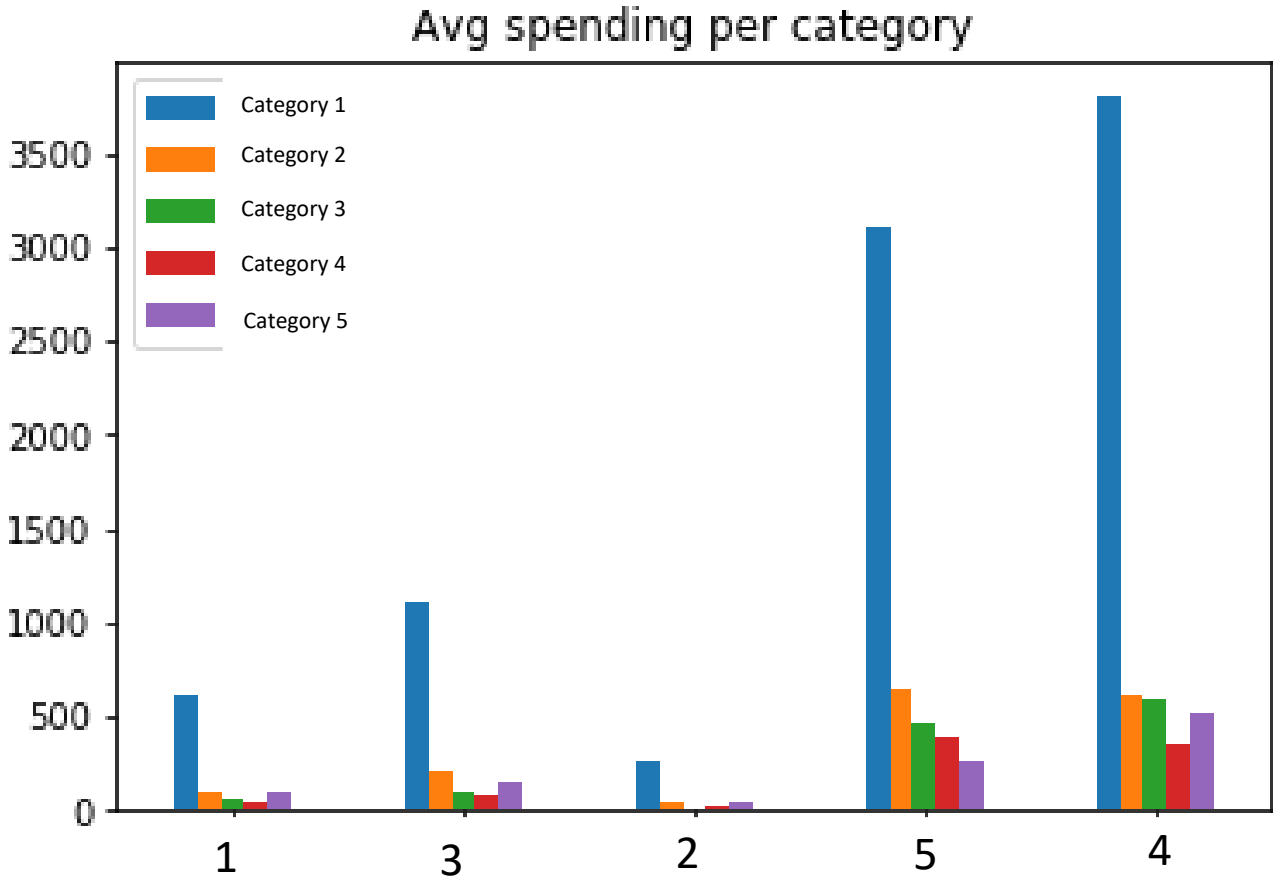
Total number of order & total spending



How these clusters are different?
Compare the average spending per order

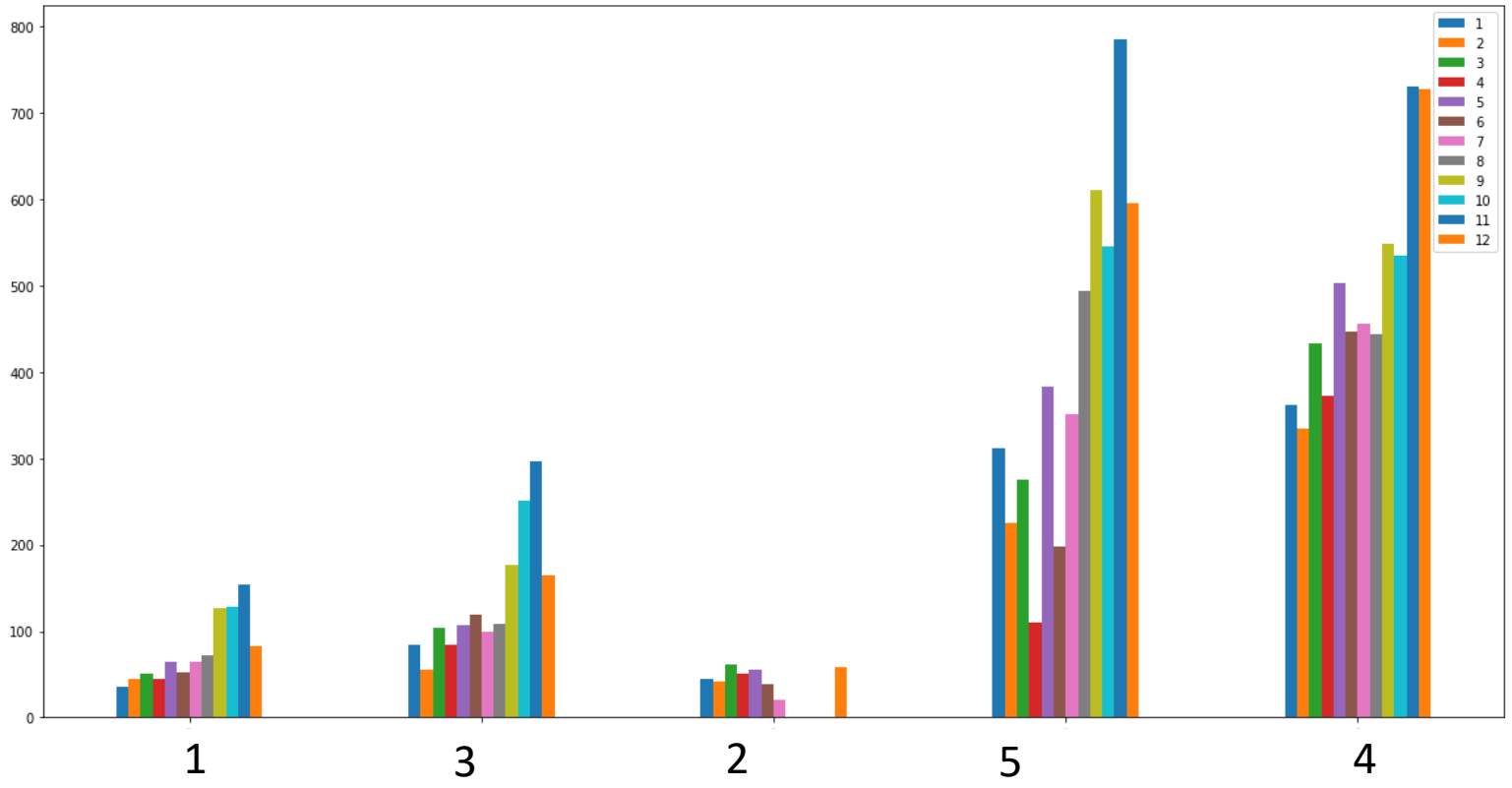


How these clusters are different?
Compare the average spending per category



All segments prefer the first category with an average spending that is significantly larger than the average spending in the second-place product category (category 2)

How these clusters are different?
Compare the average spending per month



Characterizing customers segments

Cluster 5: Loyal Customers
We cannot lose them!

They have the highest number of orders

Low average spending per order values (buying products in low prices)

However, they bought items in such high volume that this customer segment has a valuable contribution to the profit (second highest total spending)

Customers in this group shopped very frequently with stable average order value during the year.

They show a high variety of products per order

Recommended

Actions:

- Encourage them to buy more: offer a percentage discount or free shipping with
- Discount codes
- Recommend new products regularly (they showed a willingness to try different products)
- Offer loyalty program and memberships.

Cluster 3: Lost Customers

They have low order numbers and low average spending per order value.

Also, consumers in this segment rarely shop during the year and their purchased dropped significantly in three months: 9,10,11.

In average, it has been almost **a year** since the last purchase of customers in that segment (very low opportunity).

Many of them are foreigners (might be because of shipping cost?).

They have the **least** contribution to the total profit.

Returning a lost customer seems more costly than winning a new customer

Recommended

Actions:

Those customers may have experienced a problem in services, a problem with the product, or turned to different services. We need to investigate more to keep our valuable customers.

Cluster 4:
Rarely buy but spend fortune!

Have **lowest** number of orders but the highest average order value. (Has very high opportunity).

They have the highest contribution to the profit among all clusters.

In average, there has been almost a year since the last purchase from customers in this group.

They buy a lot of products per order. Probably most of them are wholesalers or suppliers.

They need more **attention** to retain them.

Recommended

Actions:

1. Marketing strategies such as email reminders or SMS push notifications targeted based on some other identifying factors.
2. Offer a discount if they return within some days (encourage them to come back soon).
3. Offer a delayed coupon (to be used in a specific time period) upon checkout.
4. Offer membership / loyalty program,

Cluster 1&2: Promising customers & Potential Loyalist

They represent high majority of our customers

They have relatively **moderate** number of orders and **moderate** average spending.

Costumers in both clusters shop frequently during the year and in average they shop very recently.

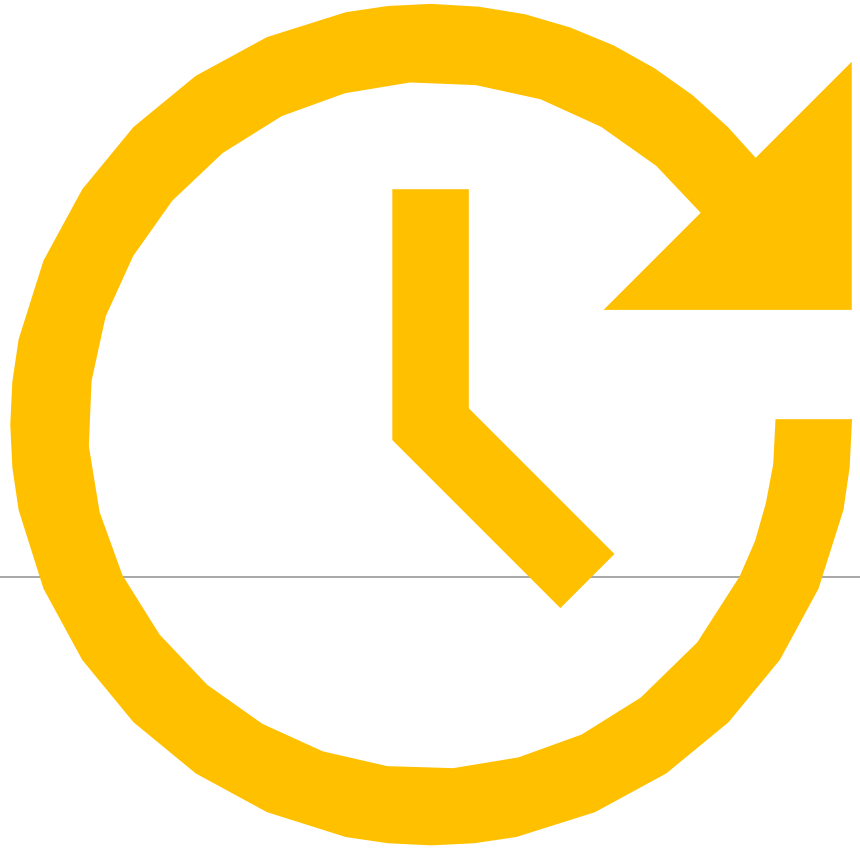
Costumers in cluster 2 show different behavior in term of products variety per order. They also have a little higher spending values.

Recommended

Actions:

1. Customers in these clusters provide a **very good opportunity**
2. With good marketing we can probably turn them into **loyal customers**
3. We can offer them membership / loyalty program and recommend other products.

Limitations & Improvements suggestions



Limitations & Improvements suggestions

Limitations

- **Limited dataset:** 1 year only.
- This customer segmentation is based on customer transactions only. If we have more data on customer demographic (i.e. gender, age, income), we might find more interesting insights.
- Not all customers are segmented due to a lack of CustomerID and purchase information.

Suggestions

Streaming clustering: estimate clusters a dynamically as new data arrive

Interactive dashboards for decision making

Predictive model / forecasting

Project Files

Project Files

All project files can be found in this GitHub repo. The repo contains the following:

1. **EDA notebook:** it shows the first step in this project where data is being explored with basic analysis and visualization

2. **Products clustering using TFIDF and W2V notebook:** it shows all steps done to apply products categorization

3. **Customers Clustering using Kmeans:** it shows all steps to apply kmeans algorithm for the final customer segmentation

* All files are provided in *.ipynb* and *HTML* (for easier reading)