**Case Study – Data Scientist**

**Problem Statement:**

1. Describe the dataset and eventual anomalies you find.
2. Which pattern do you find in the purchasing behaviour of the customers?
3. What are the categories and genres which customer are mostly interested in?
4. Split customers in different groups based on their purchase behaviour.
5. Justify your choice for your adopted method(s) and model(s).
6. Describe the defined customer groups. What are the features which are driving the differentiation amongst the groups?
7. Give suggestions on how the business should treat these clusters differently.
8. (Optional) Assuming that the ‘Category\_Reporting’ tells you the category of all the items in that order, predict:
9. The number of items per category which will be ordered on a monthly basis for the rest of May 2021.
10. The number of returns for the rest of May 2021.
11. (Optional) As, at this point in your analysis, you are the dataset expert, suggest any ideas (initiatives, further analysis) you might have in mind which can be helpful for the business.

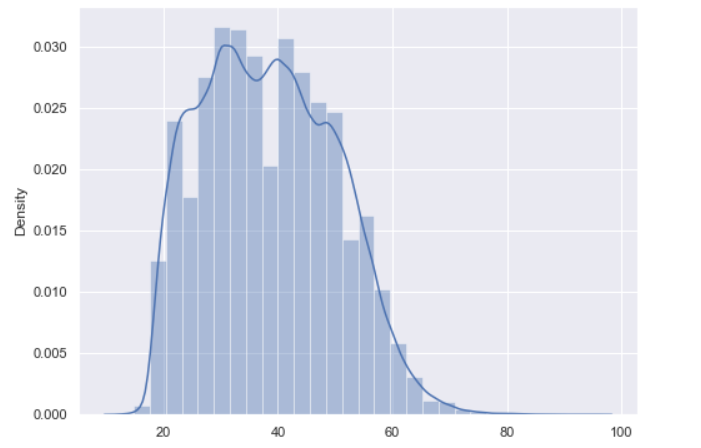
**Tools and Technology Used:**

The size of the dataset provided to me is 250 MB which is not very big data but applying ML models will be complex with this big data. So, I will be using **Spark and Python** to process this data and apply machine learning model to it.

**Data Analysis:**

The dataset was provided for the sales data produced by customers visiting and purchasing items from website emp.de in Germany between 2019-11-10 and 2021-05-10.

1. There is total 1,119,023 rows in the data set and 24 features.
2. Total number of unique customers in the data set are: 700549
3. Total number of order ids are: 1,119,023.
4. There is total 14 categories of items which is mentioned in Category\_Reporting column which states, category of the ‘OrderStarter’ item, that is, the category of the first product in the basket that was bought.
5. There is total 9 Genre of the items which is mentioned in Genre column which states, Genre of the ‘OrderStarter’ item, that is, the genre of the first product in the basket that was bought.
6. Minimum age of the customer is 12 and maximum age is 96 years. Mean age of the customers is 38 years.



1. Maximum tenure of the customers with the company is 16 years.

The dataset consists of details of the order such as cost of items, total quantity, returned items and revenue generated from the order.

**Patterns between customers in purchase behaviour and categories and genres customers are mostly interested based on the age of the customer.**

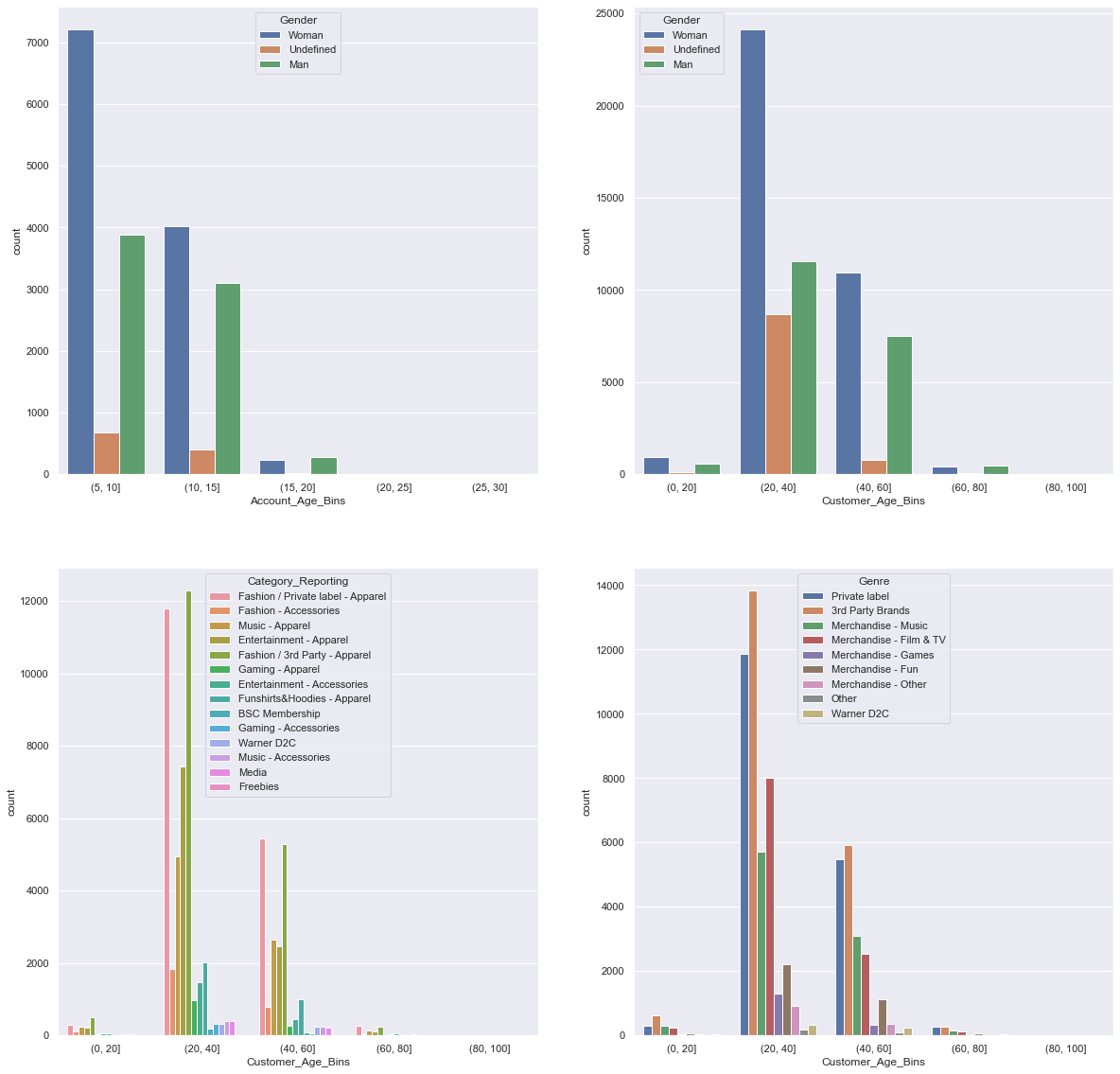
We can see from the below plot that woman tends to be the more loyal customers w.r.t man in shopping with EMP. Woman in the age bracket of 20-40 are shopping more with the company and from account age also we can see that who is new customers of the company in this also female customer are doing more shopping.

From Category Reporting we can see that customers in their young (20-40) age are more attracted towards Fashion / 3rd Party- Apparel followed by Fashion / Private Label- Apparel. From the plots we can say that youngsters are doing more online shopping.

From Genre we can say that customers are more interested in 3rd Party Brands followed by Private Labels and Merchandise – Film & TV all the customers in this are also in the age bracket of 20-40 years.

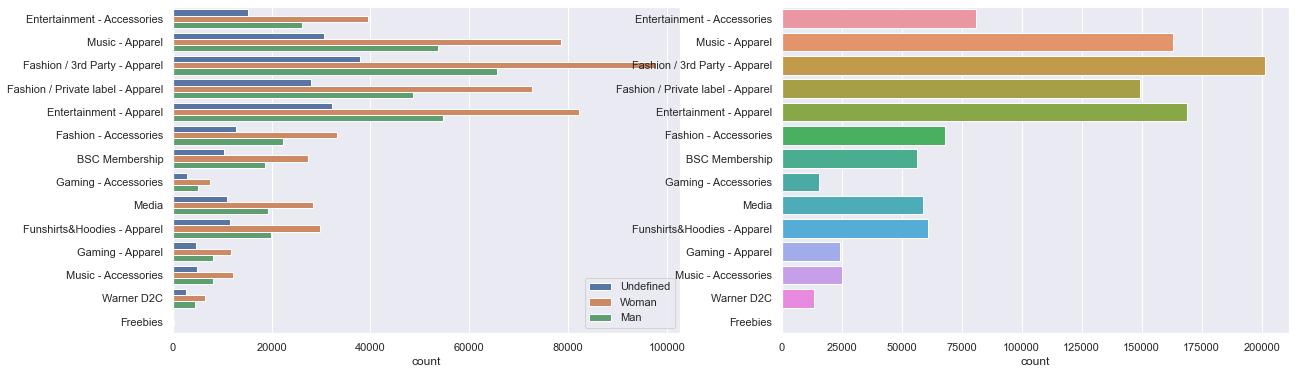
With respect to customer base, we can conclude that most of customers who are shopping with EMP are youngsters and new customers (tenure in the range of 5-10 years) of the company.

From the below plots it is also evident that EMP has mostly female customer base and it should focus more on females for marketing.

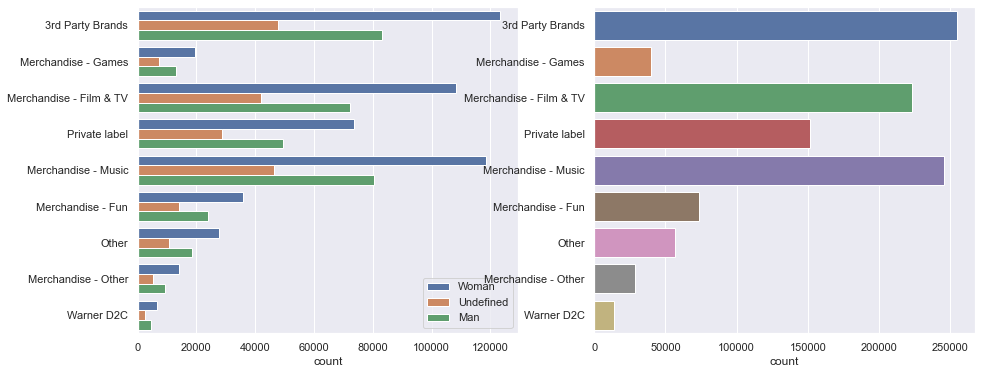


**Categories and Genre in which customers are mostly interested in**

**Category**

****

From the category plot above we can see that customers are more interested in Fashion / 3rd Party- Apparel followed by Entertainment Apparel and Music Apparel. Also, Female are best buyers of these Categories.



From Genre plot we can see that item with genre 3rd Party Brands followed by Merchandise – Music is the genre in which customers are more interested and like to buy more stuffs from these genre.

**Data Pre-Processing / Cleaning:**

There are no duplicate rows in the dataset, but there are few missing values which needs to be either imputed or dropped from the dataset since most of the machine learning models are sensitive to null values.

1) We can see that almost 80% of data is missing in Pieces\_Returns and similarly, 80% in Return\_Value.

2) There is 29.2% data is missing in customer age column to impute these values we will use mean of the age.

3) ClientType 1% of data is missing maybe we can drop the rows where the client type is missing since we have big data set.

4) We have 2% missing data in Pieces\_Outbound.

5) We have 2% missing data in Pieces\_Fulfilled.

6) We have 2% missing data in Revenue\_Goods.

7) We have 2% missing data in Delivery\_Value.

8) We have 2% missing data in OrderProfit.

**Assumption:**

1) I assume that for missing values in these fields (Pieces\_Returns and Return\_Value) meaning that no item has been returned so if no item is returned than no return value. In this case I will mark all the missing values with zero.

2) For Pieces\_Outbound, this column represents the total number of items shipped to the customer so missing values means no item are shipped to the customer. We will drop these rows where Pieces\_Outbound is missing since some of the items were ordered in 2019. I think other features (Pieces\_Fulfilled, Revenue\_Godds, Deliver\_Value and OrderProfit) are also dependent on this so may be after droping rows for Pieces\_Outbound we may not have to impute these features so we will first drop the rows and then see again if we need any imputation.

After handling missing value, we must change the categorical values to the numeric values since for our ML models we cannot use character it has to be numeric values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Imputation | Gender | AffinityProductGender | Category\_Reporting | Genre | Marketing\_Channel |
| 1 | Man | Male | Fashion / Private label - Apparel | Private label | direct |
| 2 | Woman | Female | Warner D2C | 'Warner D2C | Seo |
| 3 | Undefined | Unisex | BSC Membership | Other | Newsletter |
| 4 | - | - | Entertainment-Apparel | Merchandise - Film & TV | sea nonbrand |
| 5 | - | - | Media | Merchandise - Music | Pla |
| 6 | - | - | Funshirts&Hoodies - Apparel | Merchandise - Fun | sea brand |
| 7 | - | - | Music - Apparel | 3rd Party Brands | Retargeting |
| 8 | - | - | Fashion– Accessories | Merchandise - Other | social media |
| 9 | - | - | Fashion / 3rd Party - Apparel | Merchandise - Games | Marketplaces |
| 10 | - | - | Entertainment- Accessories | - | Affiliating |
| 11 | - | - | Gaming- Accessories | - | social referrer |
| 12 | - | - | Gaming - Apparel | - | social ads |
| Imputation | **Gender** | **AffinityProductGender** | **Category\_Reporting** | **Genre** | **Marketing\_Channel** |
| 13 | - | - | Music- Accessories | - | Unknown |
| 14 | - | - | Freebies | - | Referrer |
| 15 | - | - | - | - | Mc |
| 16 | - | - | - | - | partner program |
| 17 | - | - | - | - | Other om campaigns |
| 18 | - | - | - | - | display |
| 19 | - | - | - | - | Payback |

After imputing values, we must change data types of some of the feature from string to numeric as ML can only work with numbers.

**Data Modelling:**

Since there is no target provided, I will use un-supervised learning approach to split the customers in different groups based on their purchase behaviour.

I will be using K-Means Clustering approach to make a cluster of customers based on their purchase behaviour.

Method that I will be using for this clustering is **RFM (Recency, Frequency and Monetary Value)** approach for creating the clusters.

For using RFM approach I must do some feature engineering before applying model to the dataset.

**Feature Engineering:**

I would be creating duration column for calculating the recency of the customer. Recency basically tells that how recently the customer has shopped anything from the company.

I would also be creating a frequency column which gives the frequency with which the customer is doing online shopping with the EMP.

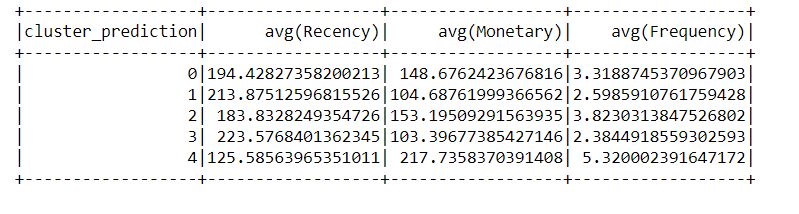
I would also be creating Monetary column for this I would be using revenue generated by the customer after return of item if there is any, considering the customer will not generate any revenue from the returned item, as the money will also be returned to the customer after return of item.

***Recency****:* It refers to the number of days before the reference date when a customer made the last purchase. Lesser the value of recency, higher is the customer visit to a store.

***Frequency****:* It is the period between two subsequent purchases of a customer. Higher the value of Frequency, more is the customer visit to the company.

***Monetary****:* This refers to the amount of money spent by a customer during a specific period of time. Higher the value, more is the profit generated to the company.

Along with these 3 features I would be using other features also such as category of product, genre, and demographic details of the customer for creating a better cluster.



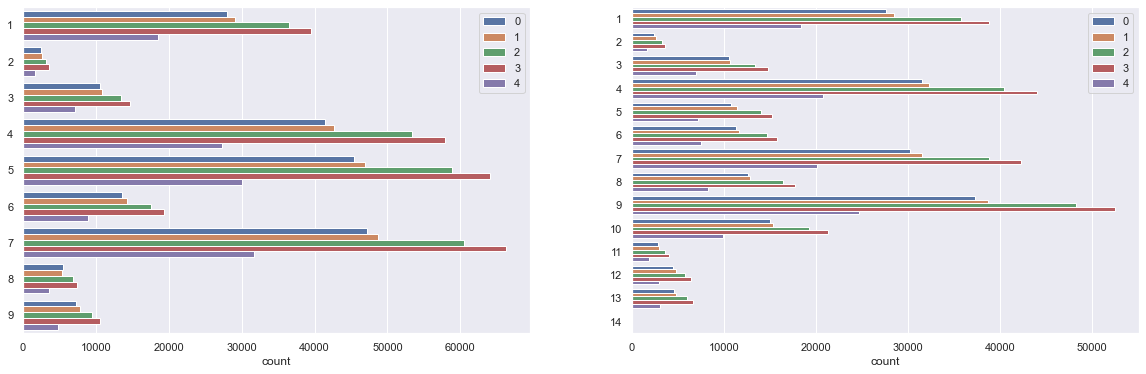
**Cluster 0:** Average Recency of the customers in this cluster is 194 and monetary value is 149 and Frequency of shopping is 3. Customers in this cluster are more recent buyers of the EMP products and spending decent amount.

**Cluster 1:** Average Recency of the customers in this cluster is 213 which shows that they are not more recent buyers and average monetary value is 105 and frequency is also 3.

**Cluster 2:** Average Recency is 184 and monetary value is 153 whereas frequency is 4 in terms of frequency and monetary value, frequency, and recency the customers could be more profitable customers of the company.

**Cluster 3:** Average Recency is 224 and monetary value is 103 whereas frequency is only 2. Based on the RFM we can say that customers in this cluster needs more attention may be with special offers to make them more frequent buyers of EMP.

**Cluster 4:** Average Recency is 124 and monetary value is 218 whereas frequency is 5. These statistics suggest that the customers in this cluster are more recent customers and spending more and very frequent buyers of company.



Based on the above plot we can see for cluster 0 the business should promote items from 3rd Party Brans in Genre and if we consider category of item then they should promote Fashion / 3rd Party – Apparel.

As we know the customers in cluster 4 are more frequent and recent buyers of the company and spending more money so for customers in cluster 4 business can give some offers to the items from genre in Merchandise – Other and Merchandise – Games as we can see there is very less sale in these genres and if somehow, we can make customers from cluster 4 buy these products we may get more profit.

**Future Work / Further Analysis:**

For Future work and Further analysis, we can do few things such as:

1. We can create a propensity model for items that will be sold to customers in cluster, with this, we can make the direct marketing of the exact items that have high propensity of being sold in that clusters.
2. We can create a recommendation of items based on the ratings given by the customer to the item. For example: items in the same cluster can be recommended to the customer based on our classification model which will tells us that whether the item should be recommended to the customer or not. With classification model we will check if the item not rated bad by customer in past.