Store Analytics Report

(By Huy Pham)

1. Introduction

Company A sells fashion. They have stores in each country's capital: UK (London); FR (Paris); IT (Milan); GER (Berlin). Their customers either buy online or in the store itself.

At the same time, it maintains a website of individual product pages with write ups and images about the page. In some cases, the page might have gone up after the product was put up for sale.

Management would like you – the data analyst – to use of this data and answer these questions:

1. Is there an impact of our website traffic on revenue?
2. Which products get us pageviews and revenue?
3. What customer segments are there?

2. Report Summary

1. **The website traffic had slightly impact on revenue**
2. **Top 3 products got us most revenue**:
    - hypsrview0.321288570724117 (**11556**)
    - reeelease0.452821711209563 (**9205**)
    - audE-info0.884915261087885 (**8428**)

    Top 3 products got us pageviews:

    - sonEies-x0.139039192479287 (**232177**)
    - porQlease0.507911745358356 (**170599**)
    - pumMvideo0.837931992780922 (**128891**)
3. **There were 3 customer segmentations**:
    - Cluster 0 (Silver): **367** customers
    - Cluster 1 (Gold): **30** customers
    - Cluster 2 (Diamond): **377** customers

    Characteristics of these clusters will be analyzed in more details below.

<u>Others:</u>

4. **Germany had the most transactions and contributed the most in revenue**
5. **There was no difference in number of transactions between Online and In Store channels.**

3. Data Explanation

Source	Description	Link
Customers & Transactions	Customer info and their transactions. These transactions are a mix of in-store and online transactions.	Customers
Website Traffic	Website traffic data for individual product pages (Duration: Jan to Dec 2020)	Traffic

Table of content

1. Introduction 1

2. Report Summary 1

3. Data Explanation 2

4. Findings 3

4.1. Findings 1: The website traffic had slightly impact on revenue 3

4.2. Finding 2: Products with highest revenue were not the Products with highest pageviews 7

4.3. Finding 3: Customer Segmentation 9

4.3.1. Clusters Result: 10

5. Conclusion 13

4. Findings

4.1. Findings 1: The website traffic had slightly impact on revenue

I use two methods to determine how website traffic and income are related.

1. Plotting Revenue by Date and Number of Pageview by Date
2. Using statistics to calculate correlation coefficient between website traffic and revenue to see how closely they are related.

I look for correlations between website traffic and revenue to further analyze the impact. If the correlation coefficient is high, it suggests that website traffic is strongly impacting revenue, whereas a low correlation coefficient suggests a weak or non-existent relationship between the two variables.

We'll examine the first approach's visualization right now.

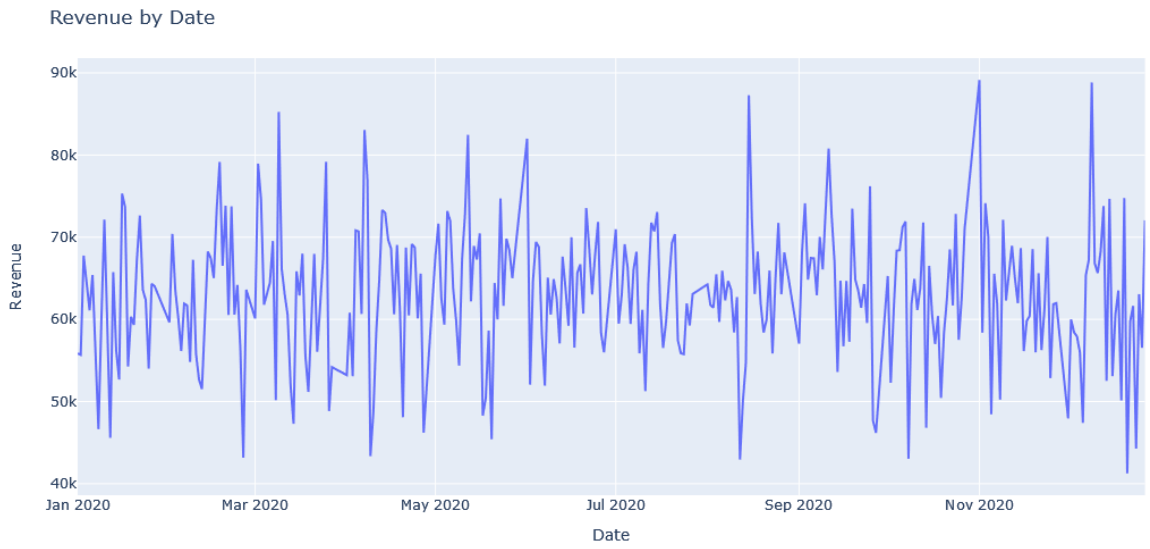

Figure 1: Revenue by Date

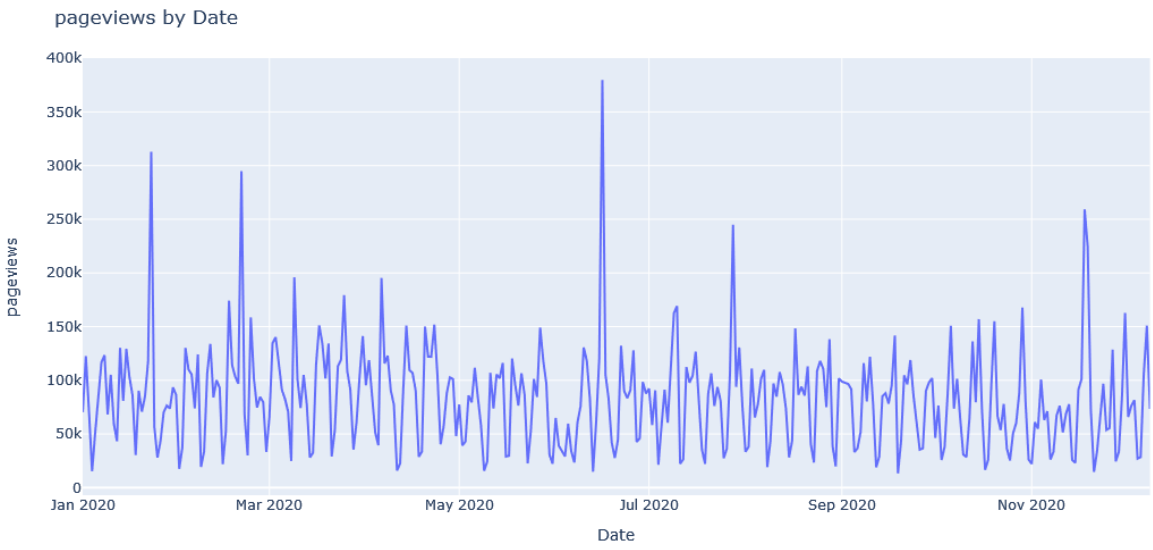

Figure 2: Number of Pageviews by Date

From Figure 1 and Figure 2 above, we can observe that there was no similarity between Revenue and Pageviews.

For example, in Figure 2, the number of Pageviews went up at peak on **Jan 23, Feb 21, Jun 16, July 28, Nov 18** for some reasons. In contrast, in Figure 1, the peak Revenue were at **Mar 9, Aug 15, Nov 1, Dec 9**.

As usual, we might want to look at the relationship between website traffic versus revenue. It makes sense when sometimes the traffic website causes slow responses which make users feel

uncomfortable or we focus too much on optimizing the performance of our website whereas it does not play a key role in generating our revenue. By looking only at the graph, we hardly can see the difference, which is why I want to apply statistical methods for better understanding.

First, we will look at the scatter plot to see the relationship between Pageviews & Revenue.

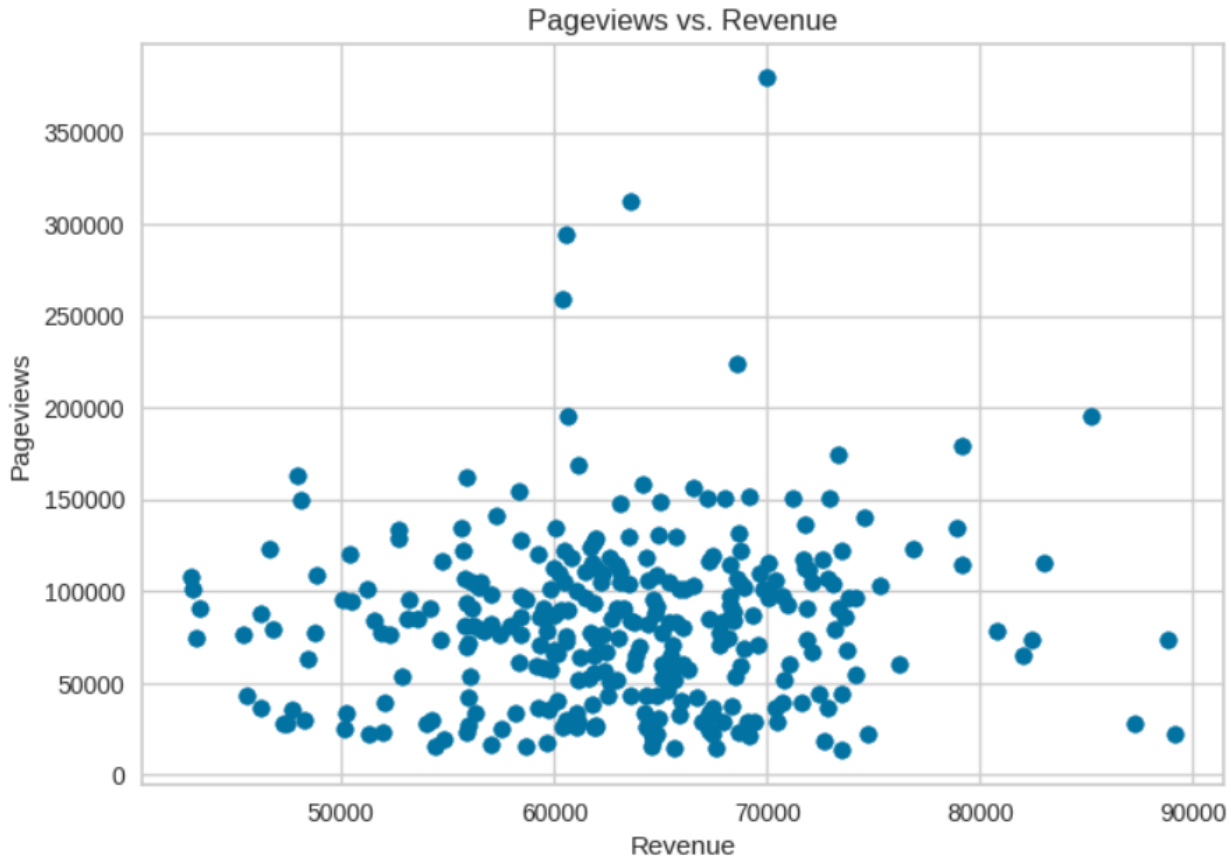

Figure 3: Correlation between Pageviews & Revenue

This method returns a value between -1 and 1, where:

- -1 indicates a perfectly negative correlation
- 0 indicates no correlation
- 1 indicates a perfectly positive correlation.

In this case, the **Correlation Coefficient** is **0.07**, which is a positive correlation coefficient, **but not significantly correlated**.

This suggests that website traffic and revenue are positively related, meaning that an increase in website traffic is associated with an increase in revenue.

Figure 4: Regression plot with Website Traffic vs. Revenue

The red line in Figure 4 is called a regression line, if it goes up from left to right, it indicates that two variables are correlated. The more slope of a red line, the higher correlation of it.

The **OLS Summary** table also tells us more about this relationship:

OLS Regression Results

Dep. Variable:	SellPrice	R-squared:	0.006
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	1.741
Date:	Fri, 19 May 2023	Prob (F-statistic):	0.188
Time:	02:17:32	Log-Likelihood:	-3177.4
No. Observations:	305	AIC:	6359.
Df Residuals:	303	BIC:	6366.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.199e+04	971.240	63.826	0.000	6.01e+04	6.39e+04
uniquePageviews	0.0150	0.011	1.320	0.188	-0.007	0.037

Omnibus:	3.980	Durbin-Watson:	1.765
Prob(Omnibus):	0.137	Jarque-Bera (JB):	4.692
Skew:	0.082	Prob(JB):	0.0958
Kurtosis:	3.585	Cond. No.	1.78e+05

Table 1: OLS Regression Results

There are many values in Table 1, but in this case, I will only focus on some values that are important (the ones in red rectangle)

The R-squared value: This represents the proportion of variance in the **dependent variable (revenue)** that can be explained by the **independent variable (unique pageviews).**

An R-squared value of 1 indicates a perfect fit between the model and the data, while a value of 0 indicates no relationship between the variables.

In this example, **the R-squared is 0.006**, which indicates that about **0.6%** of the variance in revenue can be explained by unique pageviews.

The coefficient for the independent variable: This represents the estimated impact of a one-unit increase in the independent variable (unique pageviews) on the dependent variable (revenue). A positive coefficient indicates a positive relationship between the variables, while a negative coefficient indicates a negative relationship.

The coefficient for unique pageviews is 0.015, which means that for every one-unit increase in unique pageviews, revenue is estimated to **increase by 0.015** (Currency).

In this case, **the p-value for the unique pageviews** coefficient is very small but still greater than Sig., at **0.188**, which indicates evidence against the null hypothesis (> 0.05).

This means that there is a statistically relationship between unique pageviews and revenue.

4.2. Finding 2: Products with highest revenue were not the Products with highest pageviews

Since there was a slightly correlated coefficient from what we found above in Finding 1, which leads to the fact that the Products with the highest revenue were not the Products with the highest pageviews.

For more details, let's look at the tables below:

Brand	Product	total_items	total_orders	total_customers	total_revenue
Il^(xPdB:S`#irqz	hypsrview0.321288570724117	9	9	9	11556
Relp\+KJ?D,cWw0P	reeelease0.452821711209563	7	7	7	9205
Ac8lJsKH,4xtY.Tk	audE-info0.884915261087885	7	7	7	8428
DiCo4(99zZ<nkafj	kinv-news0.29905739542661	6	6	6	7980
YMbpE$ev3qMx-h*E	ymc_orson0.319251813809483	6	6	6	7794
BMh._Fx~"+dbZy|,	bmwk-20210.31690704020655	6	6	6	7686
Ap8r`F[}sfP->SZ	ubeYanada0.100113795307302	4	4	4	7552
go@c<p.bPWb1nLrF	titM-info0.550757203954225	5	5	5	7520
Sa?9zXUH5iJbuE'S	samLement0.78565851365938	6	6	6	7518
DiL%1<G?YSFM_/TR	dioa-info0.750756927974235	6	6	6	7422

Table 2: From author's analysis

The Table 2 calculates total items, total orders, total customers, total revenue of each products.

Date	Product	Brand	users	uniquePageviews	pageviews
2020-06-16	sonEies-x0.139039192479287	So)zm5Pl^_G5_x007f_t?q	215222	221108	232177
2020-01-23	porQlease0.507911745358356	Ac8lJsKH,4xtY.Tk	155640	156789	170599
2020-11-18	pumMvideo0.837931992780922	PU]vXmX9fgf$xf!/	47779	55759	128891
2020-07-28	besfeek-50.280274313298435	BajeQM&lzKX.a8&_	53508	61120	126771
2020-02-21	rogLrecap0.644396392507623	LaX{ty9j_zZdD-\`	52381	59658	124284
2020-07-09	jefXaunch0.410067394635876	Stc\9>Cqq;Rn&/|F	43667	50536	104984
2020-12-01	chrL-york0.372865344640649	Ch6hNzXa,'=Q=)Wi	42787	49496	103863
2020-11-19	wradaunch0.370927222684073	Wrv1Qv~}Dk7@']x3	42977	49637	103850
2020-10-29	angztered0.926486710136623	An1a]{mE?nl\sM%&	34073	40323	87999
2020-03-25	goa]akers0.899516668943853	Ai/cV(XG>`JN2L_x007f_n	42587	44467	78523

Table 3: From author's analysis

The Table 3 shows top ten products that had most number of users who accessed, total unique pageviews, total pageviews on a specific day.

From the observation, we can see that products with highest number of users (users), uniquePageviews, pageviews were not the products that had the most revenue. It suggests a **potential discrepancy** between user engagement and revenue generation, it could be explained:

Conversion Rate:

The products with high user engagement might have a lower conversion rate, meaning that despite attracting a large number of users and pageviews, a smaller percentage of users actually made purchases.

Product Pricing:

The products with high user engagement might have lower sell prices, resulting in lower revenue generation compared to other products with higher sell prices.

Sales Strategies:

The products generating the most revenue could be promoted more effectively, either through targeted marketing campaigns or strategic placement on the website, resulting in higher conversion rates and revenue.

To conclude:

It would be worth investigating further to understand the factors contributing to the discrepancy between user engagement metrics and revenue generation. This analysis can help identify opportunities for improving *conversion rates*, *optimizing pricing strategies*, *enhancing product relevance*, and refining sales *and marketing efforts* to align user engagement with revenue generation

4.3. Finding 3: Customer Segmentation

In order to divide our customers into groups, I will apply a clustering algorithm named K-Means.

The main idea behind this algorithms is it categorizes our consumers based on their recent purchases, frequency of transactions, monetary value of orders, and other purchasing habits using data such as:

Demographic information: gender, age.

Geographical information: customer's city, state, or even country of residence.

Behavioral data: such as spending and consumption habits, product/service usage, and desired benefits.

Therefore, I will utilized RFM analysis - a management segmentation methodology.

We have three distinctive **features**:

- **Recency**: How recently customers made their purchase.
- **Frequency**: For simplicity, we'll count the number of times each customer made a purchase.
- **Monetary**: The total amount of money they spent

4.3.1. Clusters Result:

There were 3 customer segmentations:

- Cluster 0 (Diamond): **367** customers
- Cluster 1 (Silver): **30** customers
- Cluster 2 (Gold): **377** customers

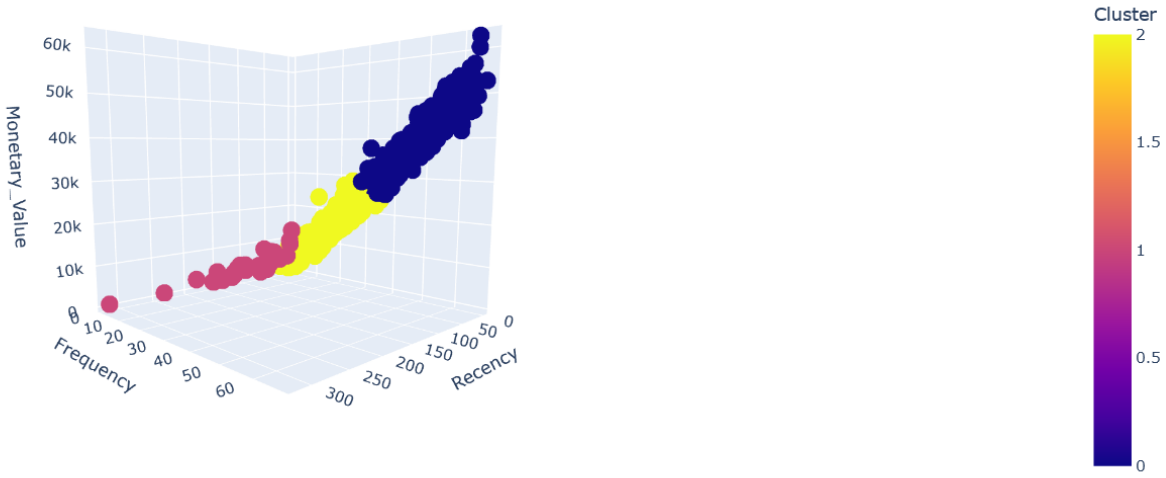

Figure 5: 3D Plot of Clusters

I will display some visualizations about behaviors, spending activities first, then I will summarize everything at the end.

4.3.1.1 Compute the average GMV for customers in Clusters

Figure 6: Monetary Value Distribution for Customers in Clusters 0, 1, and 2

Figure 7: Recency vs. Frequency for all Clusters

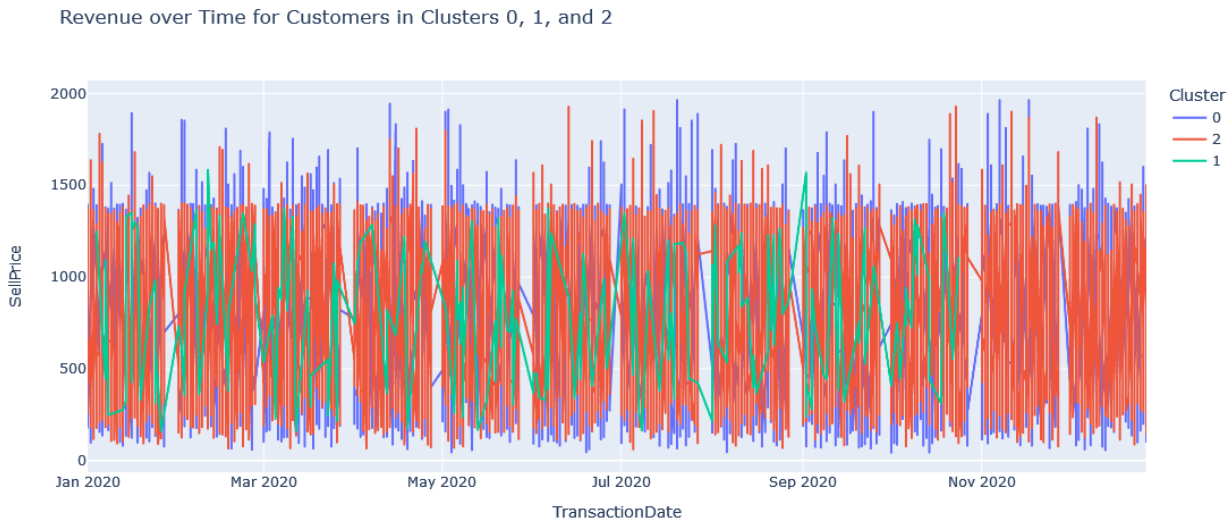

Figure 8: Revenue over Time for Customers in Clusters 0, 1, and 2

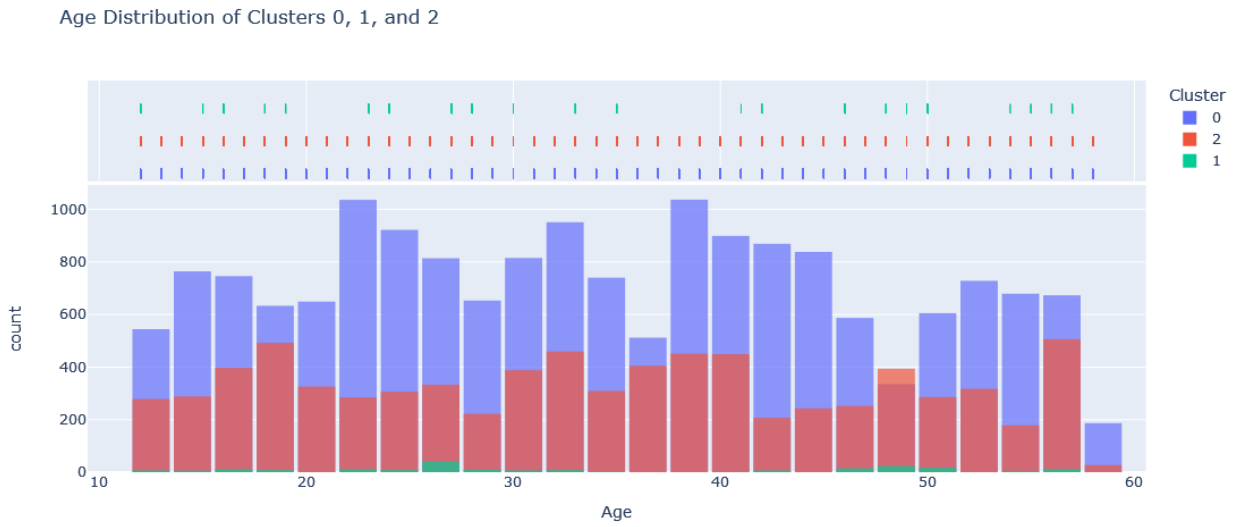

Figure 9: Age Distribution of Clusters

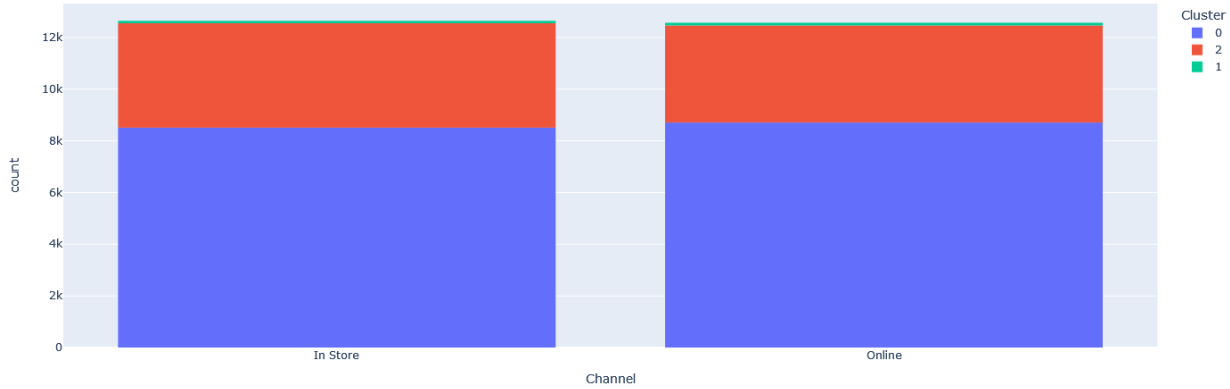

Figure 10: Number of transactions made in different channels

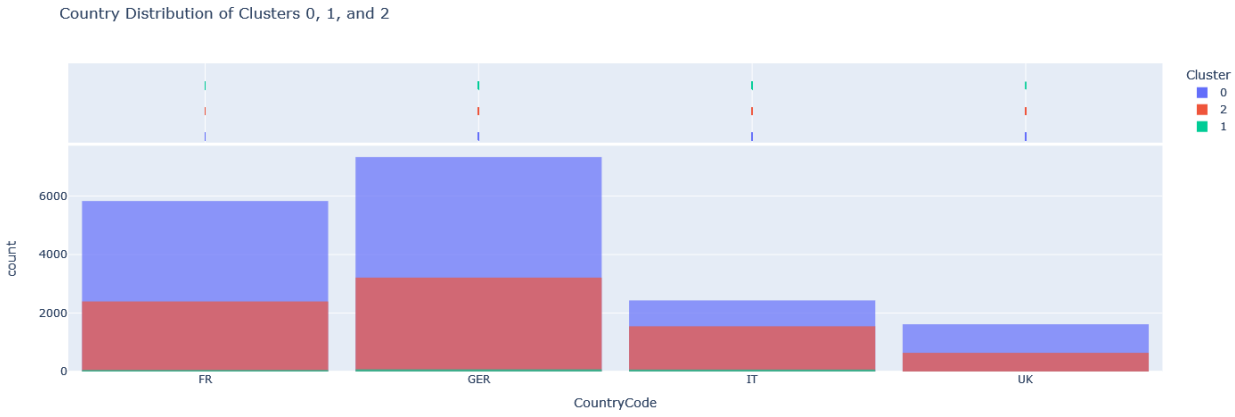

Figure 11: Number of transactions made in different countries

5. Conclusion

The characteristics of Clusters:

Cluster 0 (Diamond):

- Average Monetary Value: 39781.06
- Average Recency: 6.61
- Average Frequency: 48.87
- Mostly from Germany (Berlin), France (Paris)
- Age: Young (22 – 40)
-

Cluster 1 (Standard or Silver):

- Average Monetary Value: 7861.15
- Average Recency: 101.03
- Average Frequency: 10.20
- Mostly from Germany (Berlin), IT (Milan)
- Age: Old (46 – 50)

Cluster 2 (Gold):

- Average Monetary Value: 19253.02
- Average Recency: 13.19
- Average Frequency: 24.09
- Mostly from Germany (Berlin), France (Paris)
- Age: 18 – 57

The End

References:

Customer Segmentation Using K Means Clustering - KDnuggets

Implementing Customer Segmentation using K-Means clustering with PySpark | by Asish Biswas | Towards Data Science

How To Calculate Conversion Rate (Formula and Examples) | Indeed.com