Assignment 1 - Customer Personality Analysis

Mohammad Hossein Basoulii

March 26, 2025

Abstract

In this study, we focus on analysing personality traist of customers of a business. We seek out for patterns and relationships between the different aspects of data, e.g., has our business been effective in terms of growing spending behavior of the people who have accepted to be in campaigns of the business? or is there a relationship between education level of the customers and acceptance of campaigns? On the way, we would form some hypothesis through Exploratory Data Analysis and utilize statisical testing to examine our hypotheses.

Introduction

Background: Historically, analysis of customer personality has been of great interest for businesses which look out for continuous adapted growth based on customers' ever chaging personality traits.

Objectives: We would follow an structured path; to capture hidden patterns and relationships laying between the different factors of the data, in the Exploratory Data Analysis(EDA) section, by use of Visualization, we would gain an basic understanding of the data. Then we would utilize statisical testing to test our hypotheses in Hypothesis Testing section. Then after that, we combine our understanding and the knowledge we have gained to build up insights to the current state of our business in Conclusion & Discussion section.

Initial Hypotheses:

1. Do customers from different education levels have different income levels?
2. Does the marketing campaign influence spending behavior across customer groups?
3. Do customers with children spend differently than those without children?
4. Is there a significant difference in spending on different product categories?
5. Is there a relationship between customer education level and acceptance of promotional campaigns?

Data

A Description of the Data: We will use the dataset Customer Personality Analysis from kaggle, which includes many different personality factors for a business. The dataset contains 2240 rows(customer records) and 29 columns(features).

Features: Important numerical features include Income, Kidhome, Teenhome, MntWines(Mnt at the beginning refers to 'Amount'), MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds and NumDealsPurchases(number of deal purchases), AcceptedCmp(we have this feature repeated for all of our 5 customer campaigns). And the only important categorical feature is Education(which refers to Education Level, PhD, Basic, etc.).

Initial Data Preprocessing: We have 24 rows which are missing values in *Income* column. Since it's possible for this group of customers to be a group of interest(or at least be in any of the groups of interest), the values in their other columns might be valuable later on, Therefore, we decide to preserve these rows and impute their missing values by the mean of *Income*.

Feature Engineering: Since there are quite a few of the columns, representing information related to spending behavior of the customers(namely, all of the features which start with 'Mnt'. Which actually are six of them!), we prefer working with a single feature rather than many of them. Thus we create a new column *TotalSpendings* which is just the sum of the mentioned columns.

Exploratory Data Analysis (EDA)

Visualization

Multivatiate Analysis

Relationship of Education Level and Income: We start our understanding of the data by examining the relationship between the two variables *Education* and *Income*.
By having a look at the violin plot of *Income* for different levels of *Education* in Figure 1, we extract several key insights:

- First, that the data in different groups of *Education* doesn't seem to be normally distributed.

- Second, there is a significant different between the group with **Basic** *Education*, and the others, in terms of their *Income*. The distribution seems to be highly concentrated around the middle, thus creating a great peak around the middle. Also the middle point of the data seems to be very lower than the other groups.

- Also if we examine the propotion of data in different *Education levels*(Figure 2), we figure out that the data is highly imbalanced.

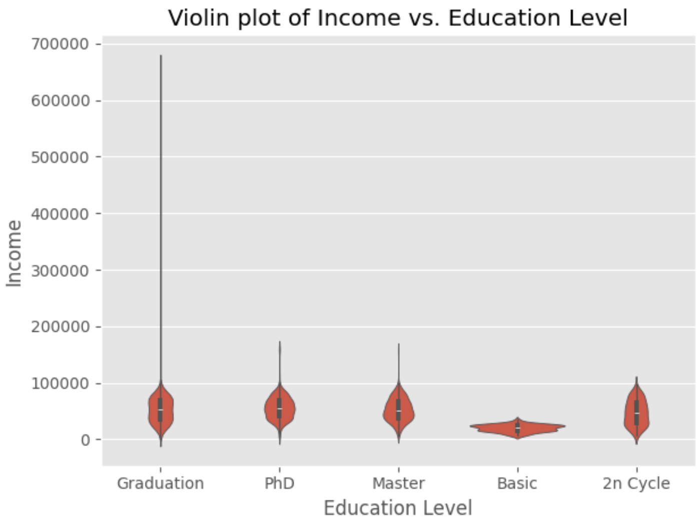
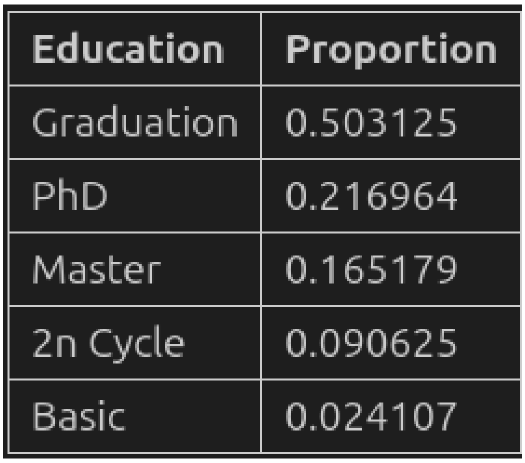


Figure 1: Violion plot of *Income* for different *Education Levels*



Education	Proportion
Graduation	0.503125
PhD	0.216964
Master	0.165179
2n Cycle	0.090625
Basic	0.024107

Figure 2: Propotions of different groups of *Education Level*

Connection of Accpeting Campaigns and Spending Behavior: KDE plot in Figure 3 shows two key points about the distribution of *TotalSpendings* in different groups of customers:

- First, the data in both of these two groups- namely, the ones who accepted customer campaigns and the ones who haven't.- is not normally distributed.
- Second, that the mean of *TotalSpendings* seems to be significantly higher in the accept group than reject group.

- Also if we examine the propotion of the data(463 accepted vs. 1777 rejected)in the these two groups, we figure out that the data is imbalanced.

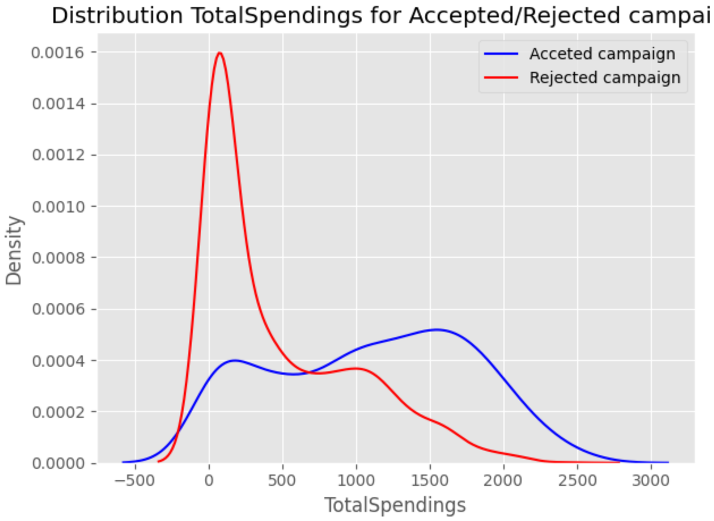


Figure 3: KDE plot of *TotalSpendings* for the group of customers who accepted *customer campaigns* and the ones who haven't.

Impact of Having Children in Spending Behavior: If we have a look at Figure 4 which shows the KDE plot of *TotalSpendings* for two different groups of customers- namely, those who have children, and the ones who have not-, we get two different insights similar to previous analyses:

- The data in both of the groups, seems not to be normally distributed.

- And also, the groups without children seem to have a significantly higher median in terms of *TotalSpendings*.

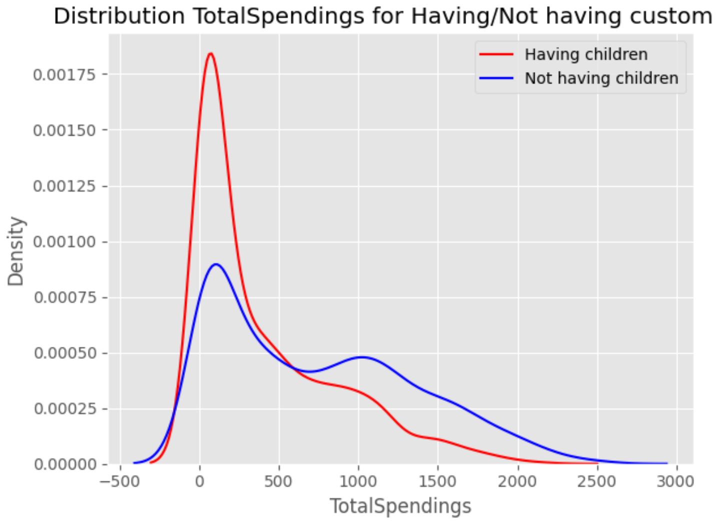


Figure 4: KDE plot of *TotalSpendings* for customers with and without children.

Relationship of Product Category and Spendings: If we have a look Figure 5, which shows the Parallel Coordinates plot of *Spendings* for different product categories, we can say that there is a significant difference between *Spendings* of different categories.

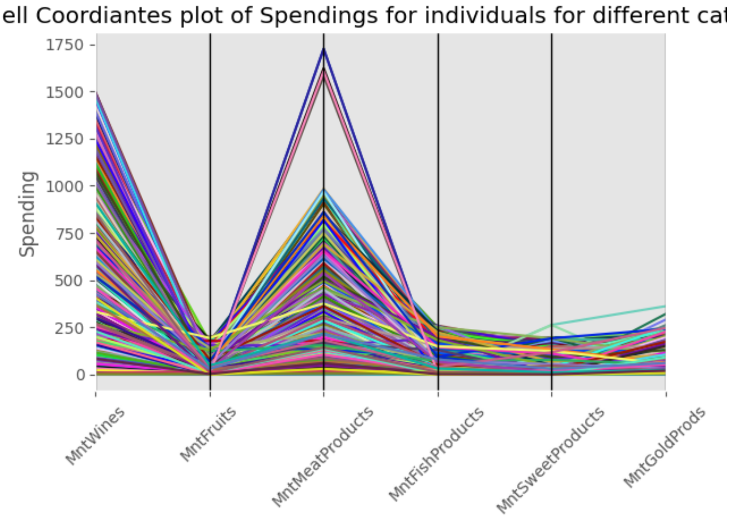


Figure 5: Parallel Coordinates plot of *Spendings* for different *product categories.*

Impact of Education Level on Acceptance of Campaigns: By having a look at Contingency Table in Figure 6, we find out that there isn't a significant in terms of *Acceptance*, for different groups of *Education Level*. We also see that all entries in the Expected Contingency Table(Figure 7) of these two variables have a expected frequency higher than 5.

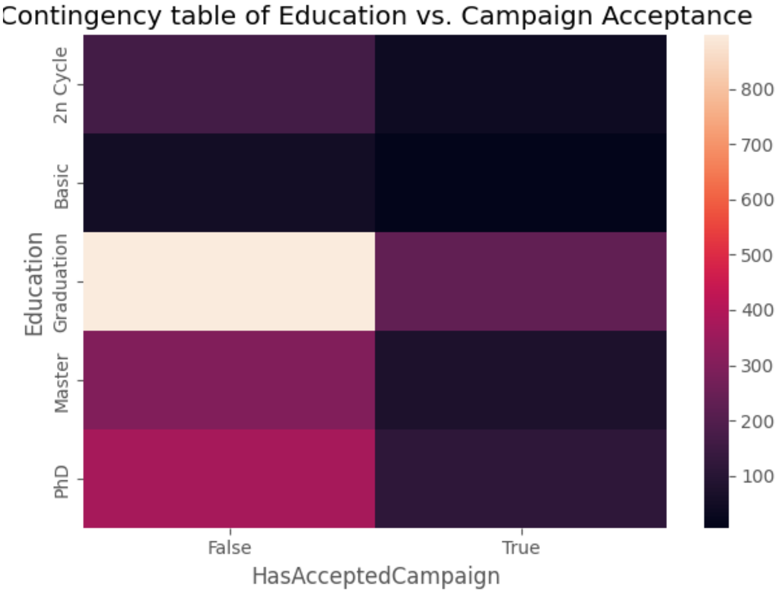
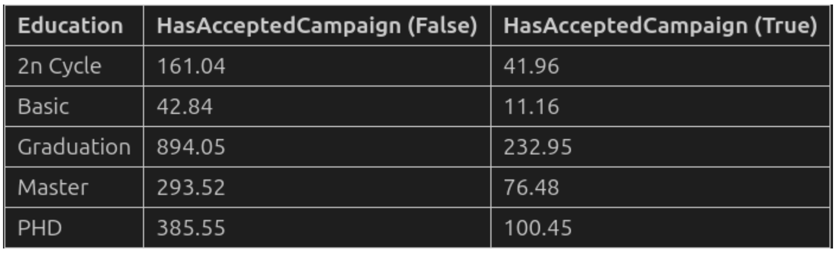


Figure 6: Contingency Table of *Education Level* and *Campaign Acceptance*



Education	HasAcceptedCampaign (False)	HasAcceptedCampaign (True)
2n Cycle	161.04	41.96
Basic	42.84	11.16
Graduation	894.05	232.95
Master	293.52	76.48
PHD	385.55	100.45

Figure 7: Expected Contingency Table of *Education Level* and *accept/reject campaign groups*

Furthur Analysis

What and Why to analyze further: Our analysis of **Connection of Accpeting Campaigns and Spending Behavior** shown us that there is an strong relationship between these two, And this could actually be an important result of our analysis which shows the effectiveness of our business in terms of growing the *Spending Bahvior* for the customers that have accepted to be in the *customer campaigns*. This motivates to do more investigations and ask questions about the measure of sucess of the business in terms of adverstisement. Therefore, we would like to add analysis of **Impact of Campaign Acceptance on Number of Deal Purchases** to our work.

Impact of Campaign Acceptance on Number of Deal Purchases: First, we have to note from our prior analysis **Connection of Accpeting Campaigns and Spending Behavior**, that the data is imbalanced here and we have to take this into consideration because it affects our interpretation here, in this analysis. Therefore, we decide to solve this problem by under sampling in the reject group. We get a better, more practial view of how the data is distributed in these two groups. By looking at the Histograms in Figure 8, we find out that the business has not been effective in terms of growing the *Number of Deals Purchases* in the *accept group*, since there is no cutting eye difference between the distributions.

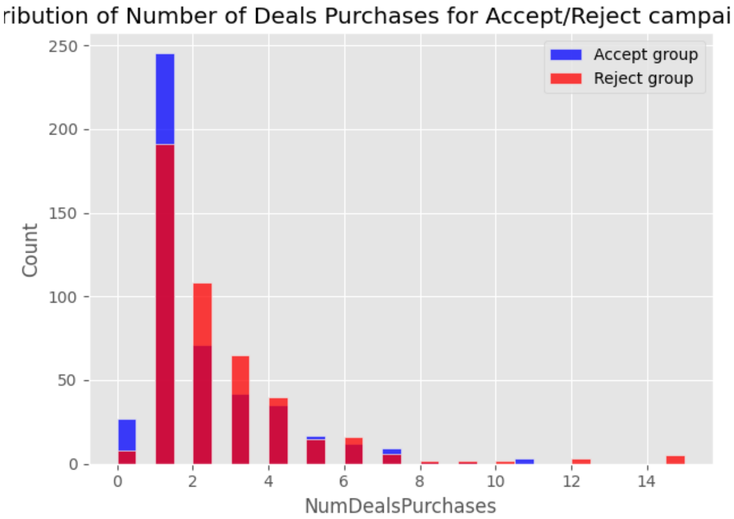


Figure 8: Distribution of *Number of Deals Purchases* for *accept/reject group*

Discussion on the Methodologies to be used:

Now, we should specify what statistical test to use for each of the hypothesis, the correct order to test the hypothesses and discuss why:

1. Does the marketing campaign influence spending behavior across customer groups? → **Mann-Whitney U test**. Based on our prior analysis in EDA section, we see that these two groups are not normally distributed in terms of *TotalSpendings*, and also the data is imbalanced. Thus the assumptions of **two sample t-test** don't hold. Therefore we should run **Mann-Whitney U test**.

2. Have our buisness been effective in terms of motivating customers in the campaigns to go towards deals purchasing? → **Mann-Whitney U test**. Based on what we saw in Figure 8, we can say that the data in these two group of customers is not normally distributed. Also the data is moderately imbalanced, Thus the assumptions of **two sample t-test** don't hold and we must perform **Mann-Whitney U test**.

3. Do customers from different education levels have different income levels? → **Kruskal-Wallis H test**. From what we saw in earlier analysis in EDA section, we figured out that the groups are not normally disributed in terms of *Income* and are imbalanced as well, Thus the assumptions of **ANOVA one-way test** don't hold. Therefore we should perform **Kruskal-Wallis H test**.

4. Is there a relationship between customer education level and acceptance of promotional campaigns? → $\chi^2 - independence\, test$. Based on what we saw in Figure 7 in our prior analysis in EDA, see that the least frequency assumption of $\chi^2 \; - independence\, test$ holds. Therefore we can perform this test.

5. Do customers with children spend differently than those without children? → **Mann-Whitney U test**. In our last analysis of this matter in EDA section, we saw that the data in these two groups of customers is not normally distributed, Thus the normality assumption of **two sample t-test** doesn't hold. Therefore we must perform **Mann-Whitney U test**.

6. Is there a significant difference in spending on different product categories? → **Friedman test**. Since the distribution of *Spendings* for different *Product Categories* might not follow a normal distribution. Thus, we must perform **Friedman test** instead of **ANOVA repeated-measures test**.

Hypothesis Testing

1. Does the marketing campaign influence spending behavior across customer groups?

Test Used: **Mann-Whitney U test** (*alternative hypothesis: customers who have accepted campaigns, have a higher spending than customers who have not accepted campaigns.*)

Results: Mann-Whitney U statistic: 628493.500, P-value: 0.000.

Interpretation and Discussion: We obtained a near-zero p-value which actually matches with our analysis in EDA section. This tells us that there is a significant difference between these two groups of customers in terms of *TotalSpendings*.

2. Have our buisness been effective in terms of motivating customers in the campaigns to go towards deals purchasing?

Test Used: **Mann-Whitney U test**. (*alternative hypotheis: customers who have accepted campaigns have a lower Number of Deals Purchases than customers who haven't accepted campaigns*)

Results: Mann-Whitney U statistic: 340665.000, P-value: 0.000.

Interpretation and Discussion: We have again obtained an extremely low p-value. That matches with our prior analysis in EDA section as well. Therefore we accept the *alternative hypotheis*.

3. Do customers from different education levels have different income levels?

Test Used: **Kruskal-Wallis H test** (*alternative hypothesis: customers from different education levels have different income levels.*)

Results: Kruskal-Wallis H statistic: 142.467, P-value: 0.000.

Interpretation and Discussion: The extremely low p-value indicates a significant difference in income levels across different education levels, which is consistent with our earlier analysis in the EDA section. We conclude that **Education Level** does have a significant effect on **Income Level**.

4. Is there a relationship between customer education level and acceptance of promotional campaigns?

Test Used: **Chi-Squared independence test** (*alternative hypothesis: there is a relationship between education level and acceptance of campaigns.*)

Results: Chi-Square Statistic: 5.865, P-value: 0.209.

Interpretation and Discussion: The p-value of 0.209 is high, indicating that there is no significant relationship between education level and campaign acceptance. This matches our previous observations in the EDA section, where we didn't find any strong correlation between these two variables.

5. Do customers with children spend differently than those without children?

Test Used: **Mann-Whitney U test** (*alternative hypothesis: customers with children spend less than customers without children.*)

Results: Mann-Whitney U statistic: 937520.000, P-value: 0.000.

Interpretation and Discussion: We obtained a very low p-value, which indicates that there is a significant difference in spending between customers with and without children. Specifically, **customers without children** tend to have **higher total spending** than those with children, as confirmed by our earlier EDA analysis.

6. Is there a significant difference in spending on different product categories?

Test Used: **Friedman Test** (*alternative hypothesis: there is a significant difference in spending across different product categories.*)

Results: Friedman Test Statistic: 5967.749, P-value: 0.000.

Interpretation and Discussion: The p-value is extremely low, indicating a significant difference in spending across different product categories. This result aligns with our EDA analysis, where we saw notable variations in spending patterns depending on the product category, suggesting that customers allocate their spending differently across various products.

Conclusion

Effectiveness of businesss in growth of *Spending Behaviors*: Based on our analysis on **Connection of Accpeting Campaigns and Spending Behavior** in EDA, and furthur investigation in Hypothesis Testing section, We can conclude that our business has been effective in the growing *Spending Behaviors* for the *customer campaigns.*

Weakness in having affect on the *Number of Deals Purchases* factor: Our findings in **Furthur Analysis** section, and the statistcal testing that we have performed for examining our hypothesis, we has shown us that our buisness has weakness in terms of motivating *customer campaigns* to increase their *Number of Deals Purchases.* Taking against in order to increase the sales of the business by adverstisement for this group of customers could be very effective, since we have just concluded that this group of customers have a significantly higher *TotalSpendings.*

Education doesn't effect Campaign Acceptance: Analysis on **Impact of Education Level on Acceptance of Campaigns**, has shown that we shouldn't pay attention to the *Education Level* of the customers while we want to advertise joining the *customer campaigns.*

It matters what product we do advertise: From our studies on **Relationship of Product Category and Spendings** in EDA section, followed by statisical testing in Hypothesis Testing section, We infer that it some *product categories* have a significantly higher *Spending.* Thus this need to be noted while advertising products.

Decission Making

Suggestion for growing *Spending Behavior* of the customers: Based on the **Conclusion** section, understanding and the knowledge that we have obtained from this study, we have several sugesstions to make:

1. We should shift our focus in advertisement of *product categories*, to a more narrow, certain range of products which have a significant higher *Spendings*.

2. We must care more about the growth of *Number of Deals Purchases* in the *customer campaigns* of the business. Since this group has a significantly higher *TotalSpendings* compared to other customers.

3. It's good to give a raise to the employees, who have been involved in our prior programs which contributed to a growth in *Spending Behaviors* inside the *customer campaigns*, since their work has been effective.