Customer Segmentation Analysis

An Exploration of iFood Marketing Data in Python

Elizabeth McHugh

Weatherhead School of Management, Case Western Reserve University

DESN 210: Introduction to Programming for Business

Professor Miyeon Jung

April 25, 2025

Table of Contents

Introduction

Marketing teams divide their customer base into distinct segments based on shared characteristics and behavior. Through segmentation, organizations can efficiently target multiple audiences, improve the relevance of their messaging, and prioritize high-value segments.

This project aims to identify high-value customer segments for iFood, a Brazil-based food ordering and delivery platform, using customers' demographic and behavioral data. Exploratory data analysis and visualization methods identified relationships between demographic factors—marital status, education level, and household income—and customer behavior, including spending trends and responsiveness to marketing campaigns. Statistical testing evaluated whether these relationships were statistically significant, and findings were used to develop strategic recommendations focused on high-value segmentation.

The analysis is guided by the following research question: Do customer spending and campaign acceptance rates significantly vary across demographic groups?

Significance of the Research Question

Customer segmentation allows marketing teams to tailor messaging and promotions to the audiences most likely to respond. McKinsey & Company reports that personalization can increase revenue by up to 15% while reducing acquisition costs by as much as 50% (McKinsey, 2023). Harvard Business Review also emphasizes that data-driven decision-making improves marketing efficiency and customer satisfaction (Davenport, 2018). Without effective segmentation, companies risk allocating marketing resources toward customers who generate low engagement or minimal long-term value. Conversely, awareness of high-value customer groups enables targeting strategies and supports efficient budget allocation, stronger customer engagement, and higher ROI. This project demonstrates an approach to identifying statistically significant demographic segmentation variables using customer data.

Data

This project uses the iFood Data Classification dataset sourced from GitHub. iFood is an online food ordering and delivery service in Brazil, similar to Uber Eats and DoorDash. The dataset includes demographic and behavioral information for 2,204 randomly selected customers collected over a two-year period. Customers were contacted using the phone number provided during registration. All monetary values are reported in Brazilian Real (BRL/R$).

Observations

Sample Demographics
n = 2,204 customers

Marital Status Distribution

- Married: 854
- In a relationship: 568
- Single: 477
- Divorced: 230
- Widowed: 76

Education Level Distribution

- Basic education: 54
- College graduate: 1,113
- Second-cycle education: 198
- Master’s degree: 364
- PhD: 476

Summary Statistics

- Average age: 51 years
- Average household income: 51,622 BRL

Variables

To improve clarity, several variables from the original dataset were renamed.

Original Variable Name	Description	New Variable Name
ID	Customer ID	
Year_Birth	Birth year	

Income	Annual household income	Household_Income
Kidhome	Number of children (< 13 years old) in customer’s household	Number_Kids
Teenhome	Number of children (13 – 19 years old) in customer’s household	Number_Teens
Customer_Days	Length (days) of customer’s enrollment with iFood	Enrollment_Length
Recency	Days since last purchase	Days_Since_Purchase
MntWines	R$ spent on wine over 2 years	AmtWines
MntFruits	R$ spent on fruit products	AmtFruits
MntMeatProducts	R$ spent on meat products	AmtMeatProds
MntFishProducts	R$ spent on fish products	AmtFishProds
MntSweetProducts	R$ spent on sweet products	AmtSweetProds
MntGoldProds	R$ spent on gold products	AmtGoldProds
MntRegularProds	R$ spent on all non-gold products	AmtRegularProds
NumDealsPurchases	Number of purchases made with a discount	
NumWebPurchases	Number of purchases made through iFood’s website	
NumCatalogPurchases	Number of purchases made using the catalog	
NumStorePurchases	Number of purchases made directly in-store	
NumWebVisitsMonth	Number of visits to the iFood website each month	

AcceptedCmp1	Boolean: 1 = customer accepted the offer in the 1st campaign; 0 = otherwise	
AcceptedCmp2	Boolean: 1 = customer accepted the offer in the 2nd campaign; 0 = otherwise	
AcceptedCmp3	Boolean: 1 = customer accepted the offer in the 3rd campaign; 0 = otherwise	
AcceptedCmp4	Boolean: 1 = customer accepted the offer in the 4th campaign; 0 = otherwise	
AcceptedCmp5	Boolean: 1 = customer accepted the offer in the 5th campaign; 0 = otherwise	
Complain	Boolean: 1 = customer complained at least once; 0 = otherwise	
Response	Boolean: 1 if customer accepted the offer in the last campaign, 0 otherwise	
Z_CostContact	Unavailable	
Z_Revenue	Unavailable	

Education Level Indicators *(Boolean: 1 = True; 0 = False)*

Original Variable Name	Description	New Variable Name
education_Basic	Basic education	Basic
education_Graduation	Graduated from college	Graduated
education_2n Cycle	Completing a second-cycle education	2nCycle
education_Master	Earned a master’s degree	Masters

education_PhD	Earned a PhD	PhD

Marital Status Indicators *(Boolean: 1 = True; 0 = False)*

Original Variable Name	Description	New Variable Name
marital_Married	Married	Married
marital_Together	In a relationship	Together
marital_Single	Single	Single
marital_Divorced	Divorced	Divorced
marital_Widow	Widowed	Widowed

New Columns

To measure the total number of children in each household, the variables Number_Kids and Number_Teens were combined into a new variable titled Total_Children. To measure average spending on non-gold products, AmtRegularProds was divided by five (the number of non-gold product categories), and results were stored in a new column titled AvgRegularProds.

In the original iFood dataset, boolean variables indicated marital status and education level. To improve clarity and simplify analysis, marital status and education level values were consolidated into two separate categorical variables titled marital_status and education_level.

Methodology

Data Cleaning

The pd.isna() function returned zero missing or duplicate values; however, dropna() and drop_duplicates() were applied to ensure the data were fully cleaned.

Exploratory Data Analysis

Exploratory analysis began by summarizing demographic distributions across marital status and education level groups. The .sum() function returned the number of customers in each demographic category, while .describe() calculated summary statistics, including average customer age, average household income, and average spending on regular products.

The .query() function counted the number of customers in each demographic group who purchased wine, fruit, meat, fish, sweet, and gold products. The same approach identified customers with at least one child and calculated how many customers accepted at least one marketing campaign.

The .groupby() function returned mean spending by product category across demographic groups. Customers spending above the sample average were identified using .query() and compared across marital status and education level groups. Campaign acceptance rates were calculated by counting how many customers accepted each campaign. Acceptance overlap across campaigns was also evaluated using .query().

Data Visualization

Customer segmentation was explored using three demographic factors: marital status, education level, and household income. To support visualization and modeling, multiple summary DataFrames were created: marital_regularprods displayed average spending on regular products across marital status groups; education_regularprods displayed average spending across education levels; householdincome_regularprods displayed household income and average spending on regular products.

Campaign acceptance rates were analyzed across marital status and education level groups using additional DataFrames titled marital_campaignX and education_campaignX, where X represents the campaign number. Each DataFrame contained the number of customers in each demographic group who accepted the campaign.

Bar charts were used to visualize differences in spending and campaign acceptance across categorical demographics. A scatter plot was used to visualize the relationship between household income and average spending on regular products.

Statistical Modeling

Two statistical techniques were used to evaluate how customer demographics relate to spending behavior and campaign responsiveness: one-way Analysis of Variance (ANOVA) and correlation analysis with linear regression.

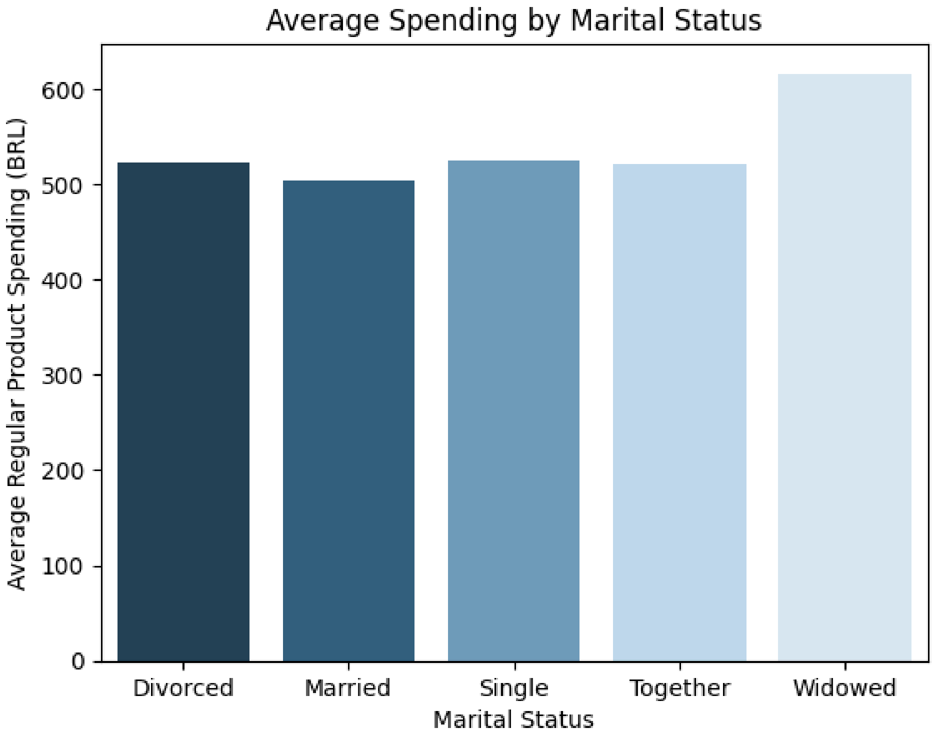
One-way ANOVA compared mean spending levels across categorical demographic variables, specifically marital_status and education_level. ANOVA was selected because it compares average outcomes across multiple demographic groups in a single test. This approach avoids running multiple separate comparisons, which increases the likelihood of falsely identifying a difference that does not actually exist (Type I error). Each ANOVA test used a significance level of 0.05. A p-value below this threshold indicates statistically significant differences among group means.

Correlation analysis and linear regression were used to evaluate the relationship between household income and average product spending. Both variables are continuous, making these methods appropriate for measuring relationship strength and direction. Linear regression also quantified the predictive relationship between income and spending and provided model fit metrics, including R^2 and the F-statistic. An F-statistic compares the variance between group means to the variance within groups, with a value > 1 indicating possible significant difference. R^2 measures how well the independent variable explains variation in the dependent variable.

Results

Spending by Marital Status

Figure I



Hypothesis

H^1_0: The difference in average spending by widowed customers compared to other marital status groups is not statistically significant.
H^1_1: The difference in average spending by marital status is statistically significant.

Findings

Widowed customers spent an average of 616.25 R$, while other marital status groups spent between 503.66 R$ and 525.32 R$.

Results

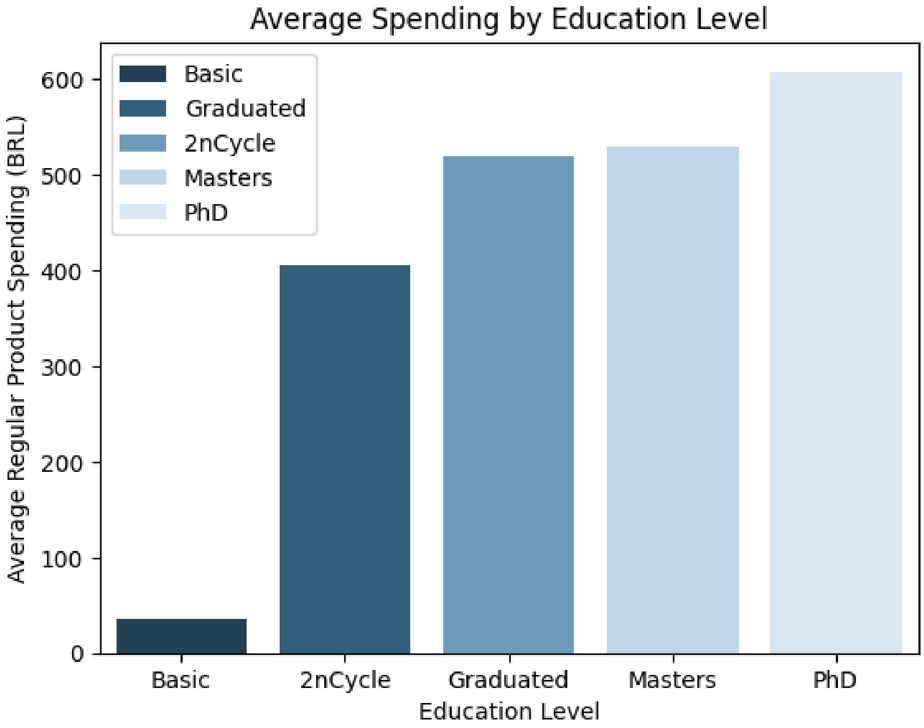
Failure to reject the null (F = 0.77, p = 0.55).

Conclusion

Marital status does not have a significant impact on regular product spending and does not provide a meaningful basis for segmentation.

Spending by Education Level

Figure II



Hypothesis

H^2_0: Education level does not significantly affect spending on regular products.

H^2_1: Education level significantly affects spending on regular products.

Findings

Customers with a PhD spend an average of 608.21 R$ on regular products, which is 89.50 R$ greater than the entire sample average. The range of the data is 572.08 R$.

Statistical Results

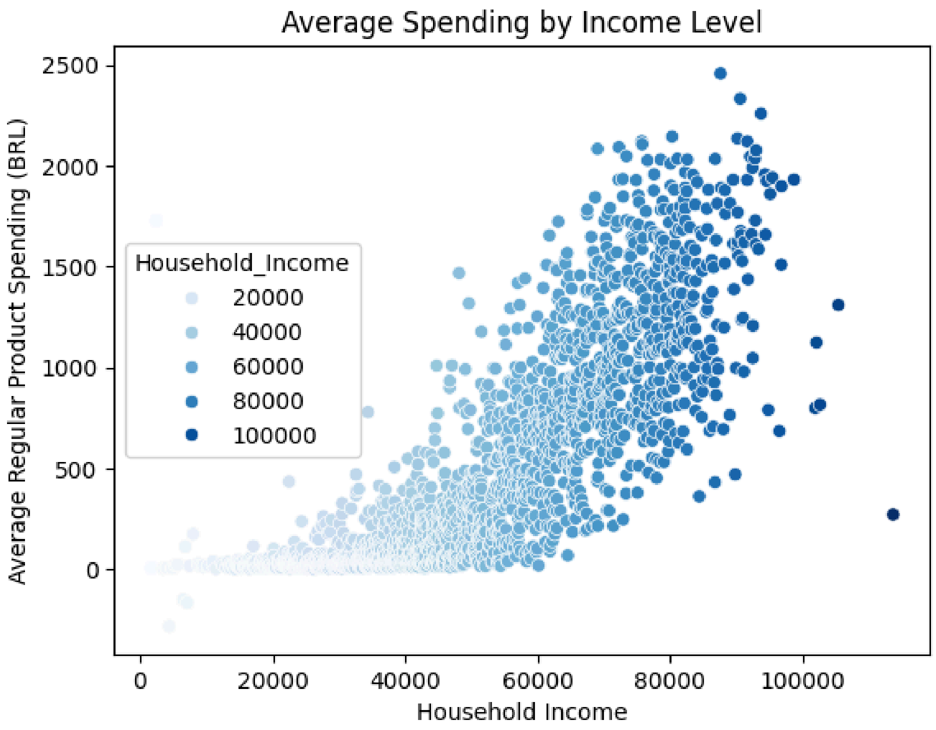
Reject the null and accept the alternative ($F = 15.91$, $p < 0.05$).

Conclusion

Education level significantly affects spending on regular products, with higher education levels associated with greater spending. Education level is a meaningful basis for segmentation within regular product categories.

Spending by Household Income

Figure III



Hypothesis

H^3_0: Household income does not have a strong positive relationship with average spending on regular products.

H^3_1: Household income has a strong positive relationship with average spending on regular products.

Findings

Spending on regular products increases as household income increases.

Statistical Results

R^2 = 0.667	F = 4418	Intercept = -608.85	Coefficient = 0.0281

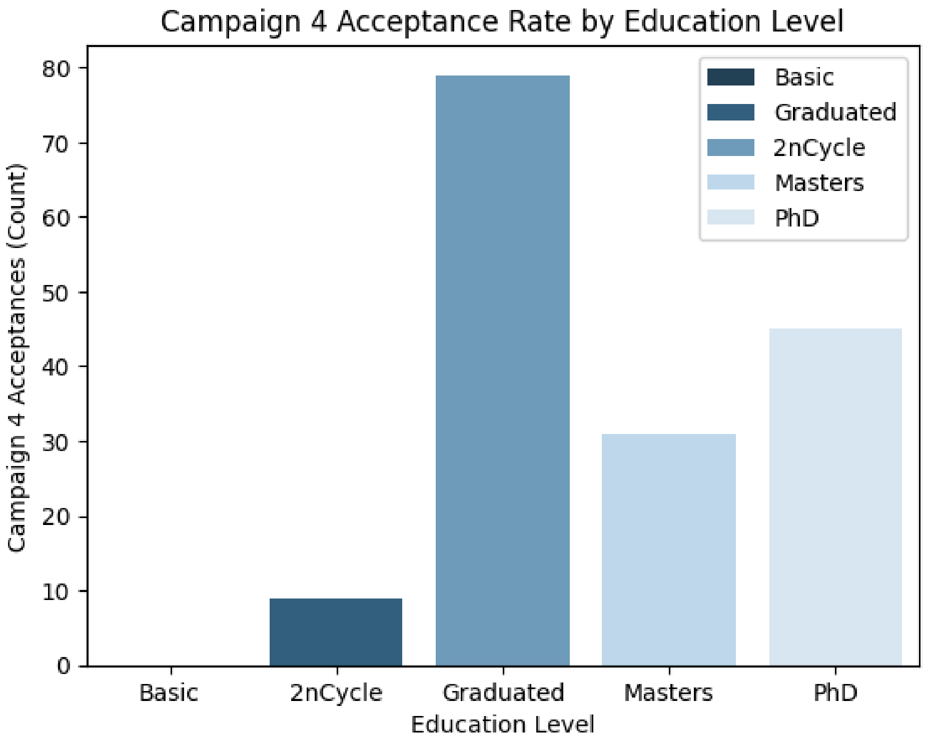
Reject the null and accept the alternative ($p < 0.05$). Approximately 66.7% of the variance in average spending on regular products is explained by household income. For every 1 R$ increase in household income, average regular product spending increases by approximately 0.03 R$.

Conclusion

Household income has a strong positive relationship with regular product spending. Income level supports effective segmentation across these product categories.

Campaign Effectiveness by Marital Status and Education Level

Figure IV



Hypothesis

H^4_0 : Marital status and education level do not influence the acceptance rate of Campaign X.
H^4_1: Marital status and education level influence the acceptance rate of Campaign X.

Findings

One-way ANOVA between marital status, education level, and campaign acceptance did not yield statistically significant results—with the exception of Campaign 4. Customers in the "Graduated" education class were significantly more likely to accept Campaign 4.

Statistical Results

Between education level and Campaign 4, the null hypothesis was rejected ($F = 2.60$, $p = 0.035$). For all other campaigns, ANOVA results were not statistically significant across marital status or education level ($p > 0.05$).

Conclusion

Education level influenced the acceptance rate of Campaign 4, making it a meaningful basis for targeted marketing strategies. A/B testing can evaluate which campaign elements contributed to higher acceptance rates among the "Graduated" group.

Limitations

1. **Causal Inference Limitations**
   One-way ANOVA and correlation analysis identify associations but do not establish causation. For example, household income is positively correlated with spending on regular products, but higher income does not necessarily cause higher spending.

2. **Uneven Sample Sizes Across Demographic Groups**
   Some demographic groups contained significantly larger sample sizes than others, which may introduce bias. For example, the "Widowed" group included only 76 customers, while the "Married" group included 854 customers. Smaller sample sizes increase variability and reduce confidence in conclusions drawn for smaller demographic groups.

3. **Data and Context Limitations**
   The dataset lacked important contextual information, including item-level purchase details and campaign characteristics such as channel strategy, campaign duration, creative content, targeting criteria, and whether campaigns were paid or organic.

   The dataset did not clarify whether spending values were self-reported or collected directly from iFood's electronic database, which may affect data reliability.

   Interaction effects between demographic variables were not evaluated. These limitations restrict interpretation of why differences in spending behavior and campaign acceptance occurred across demographic groups.

Conclusion

Customer spending and campaign acceptance patterns varied across demographic groups, revealing supporting demographic-based targeting strategies. Based on the findings, the following recommendations are made:

1. **Segment by Education Level**
   The significant relationship between education level and spending on regular products suggests education can serve as a key factor for segmentation. iFood should develop targeted marketing strategies for customers with higher education levels, particularly those with bachelor's degrees or higher, as these groups are associated with higher spending on regular products.

2. **Segment by Household Income Levels**
   The strong positive correlation between household income and spending on regular products supports income-based segmentation. iFood should create targeted campaigns for higher-income segments.

3. **Refine Campaign Strategies**
   Campaign 4 performed significantly better among customers with a bachelor's degree, indicating education level may influence campaign responsiveness. Campaign design and targeting should be refined based on education level, supported by A/B testing to identify which campaign elements generate higher acceptance rates.

4. **Address Uneven Sample Sizes in Future Research**
   Uneven sample sizes across demographic groups, particularly the small "Widowed" segment, may reduce reliability. Future analyses should use more balanced samples to strengthen the validity of group comparisons.

5. **Future Analysis & Data Collection**
   Future analysis should explore interaction effects between key demographic variables (e.g., income × education), and identify relationships between behavioral indicators (e.g., recency, web visits, discount-driven purchasing) and outcomes. Future data collection should include product-level purchase data and campaign metadata (e.g., channel strategy, creative content, timing).

These recommendations support segmentation strategies to efficiently allocate resources across high-value customer groups.

References

[1] iFood. (2020). *iFood data business analyst test.*

https://github.com/nailson/ifood-data-business-analyst-test

[2] Davenport, T. H. (2006). *Competing on Analytics.* Harvard Business Review.

https://www.researchgate.net/publication/7327312_Competing_on_Analytics

[3] McKinsey & Company. (2023). *What is personalization?*

https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-personalization

[4] Jung, M. (2025). DESN 210, Case Western Reserve University.