**VIETNAM NATIONAL UNIVERSITY OF HO CHI MINH CITY**

**INTERNATIONAL UNIVERSITY**





**STATISTICAL METHOD**

**FINAL PROJECT**

**“CUSTOMER SEGMENTATION ANALYSIS”**



**ABOUT THE PROJECT**

We would like to express our deepest gratitude to all those who have contributed to the successful completion of this project for the Statistical Methods course at International University. First and foremost, we are sincerely thankful to Ms. Uyên, our esteemed lecturer, for her invaluable guidance, insightful lectures, and constant support throughout the duration of this project. Her profound knowledge and enthusiasm for the subject have been a significant source of inspiration for us.

We also wish to acknowledge the help and support provided by our fellow classmates, whose collaborative spirit and constructive feedback have greatly enriched our learning experience. Their cooperation and encouragement have been instrumental in overcoming the various challenges encountered during the project.

Furthermore, we would like to extend our gratitude to our teammates for providing the necessary resources and a conducive environment for research and study.

Thank you all for your immense contributions to this project.

**PROJECT GOAL**

Customer Personality Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. In this report, I’m going to introduce you to a data science project on customer personality analysis with Python and Statistical Method.

**TABLE CONTENTS**

1. **Introduction**
2. **Members**
3. **Preprocessing**
4. **Explanatory Data Analysis**
5. **Statistics and Visualization**
   1. **Age / Wines**
   2. **Education / Wines**
   3. **Marital Status / Wines**
   4. **Has Child / Wines**
   5. **Income, Spending , Seniority / Wines**
6. **Conclusions**
7. **References**
8. **INTRODUCTION**

In today's competitive retail landscape, understanding customer behavior and preferences is crucial for businesses to stay ahead. Customer segmentation, the process of dividing customers into distinct groups based on shared characteristics, has become a valuable tool for convenience stores to optimize their marketing strategies, product assortment, and overall customer experience.

This project aims to analyze the customer segmentation of a local convenience store using the store's transactional data. By exploring the different customer segments and their purchasing patterns, the store can gain valuable insights to better serve its clientele and drive business growth.

The key objectives of this study are:

1. To identify distinct customer segments within the convenience store's customer base based on their demographic and purchasing characteristics.
2. To analyze the most popular products purchased by each customer segment, revealing their preferences and consumption habits.
3. To provide strategic recommendations for the convenience store to effectively target and engage with these customer segments, ultimately enhancing customer satisfaction and loyalty.

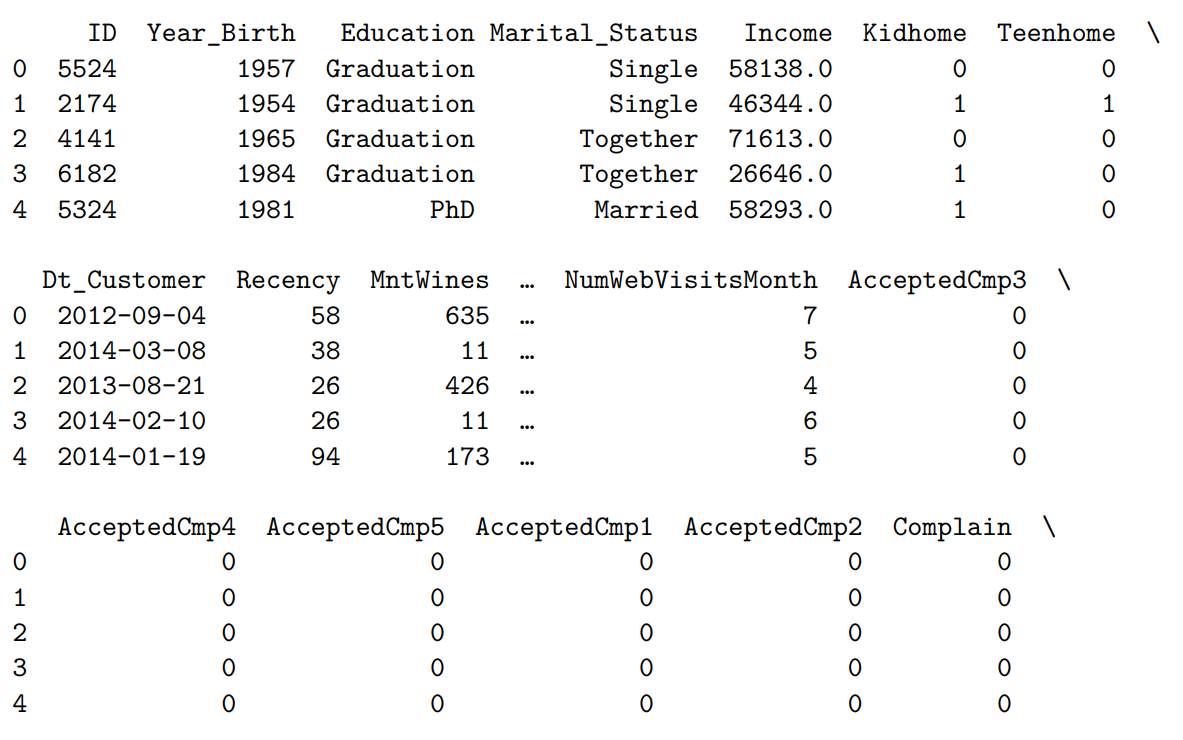
The findings of this customer segmentation analysis will enable the convenience store to make data-driven decisions, optimize its product assortment, and implement tailored marketing and promotional strategies for each identified customer segment. This, in turn, will help the store strengthen its competitive position, increase sales, and foster long-term relationships with its valued customers.

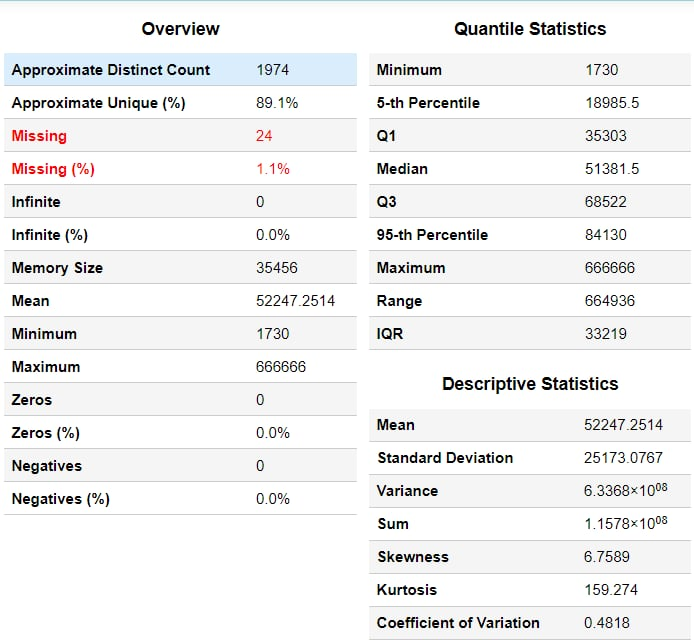
1. **MEMBERS**

|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Contribution** |
| Hoàng Văn Mạnh | ITDSIU21099 | 100% |

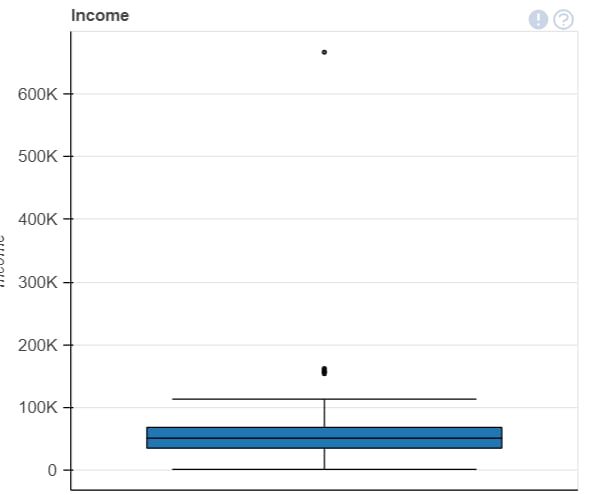
1. **PREPROCESSING**

**Raw Data**

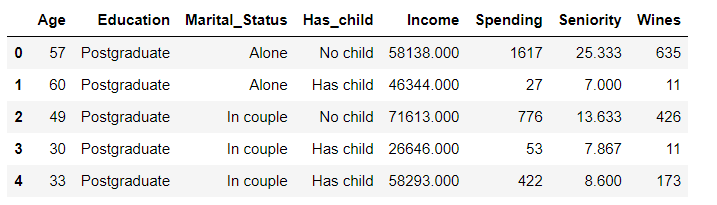


**Before Remove Missing Values**  
 

**Before Remove Outlier**



**New Data**



**Data Evaluation**

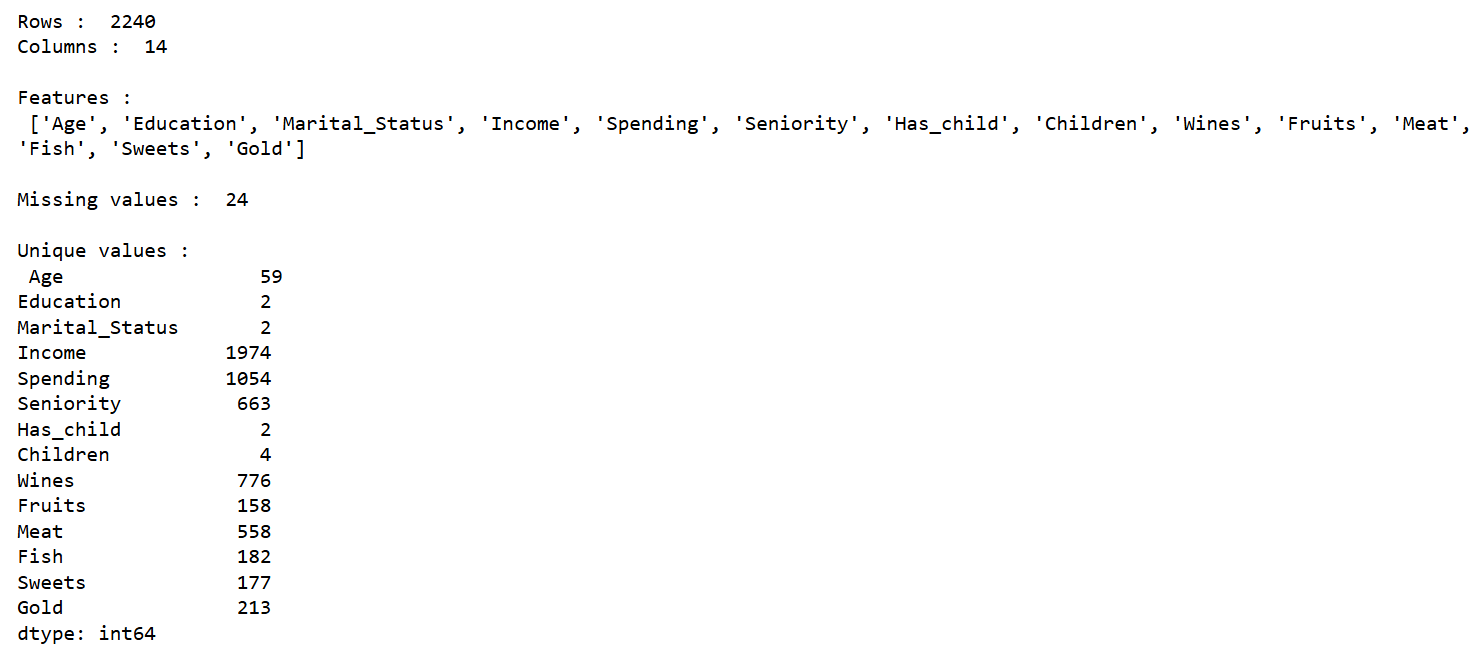
Following the conversion of raw data collected from our online retail system, we now have a refined dataset that highlights the essential products required for the entirety of our project. This dataset provides invaluable insights that will drive the strategic planning and execution of our project, ensuring that we are well-equipped to meet our objectives.

**Product Demand Trends**: By analyzing sales patterns and customer preferences, we can identify high-demand products, allowing us to optimize popular items. This ensures that we can meet customer demands efficiently.

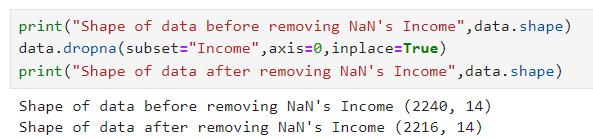
**Customer Segmentation**: The data allows us to categorize our customer base based on purchasing behavior and product preferences. This segmentation facilitates understanding which types of products will attract specific customer segments, enabling precise analysis.

In conclusion, the refined data equips us with a comprehensive understanding of our product landscape, empowering us to make data-driven decisions.

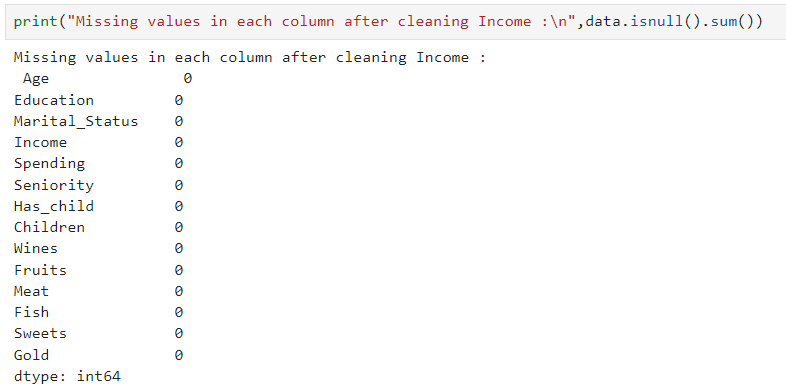
1. **EXPLORATORY DATA ANALYSIS**

**Missing Values**  
 

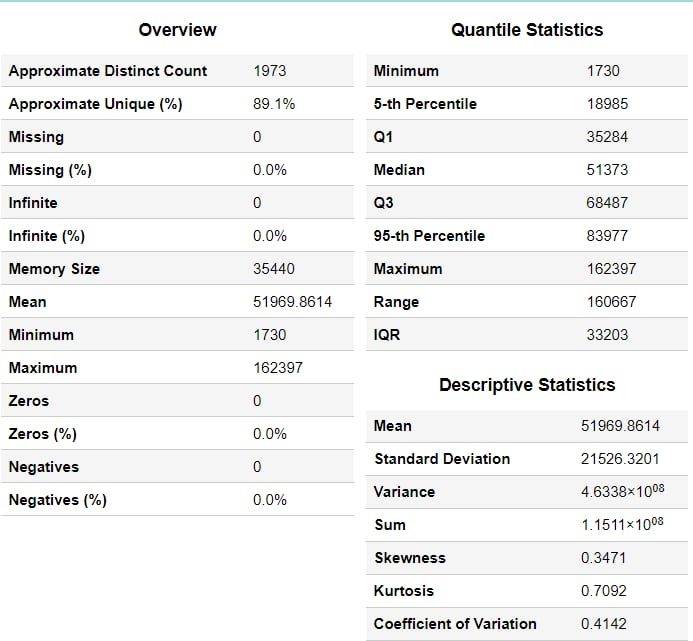
**Dealing With Missing Values**



**Missing Values After Cleaning**



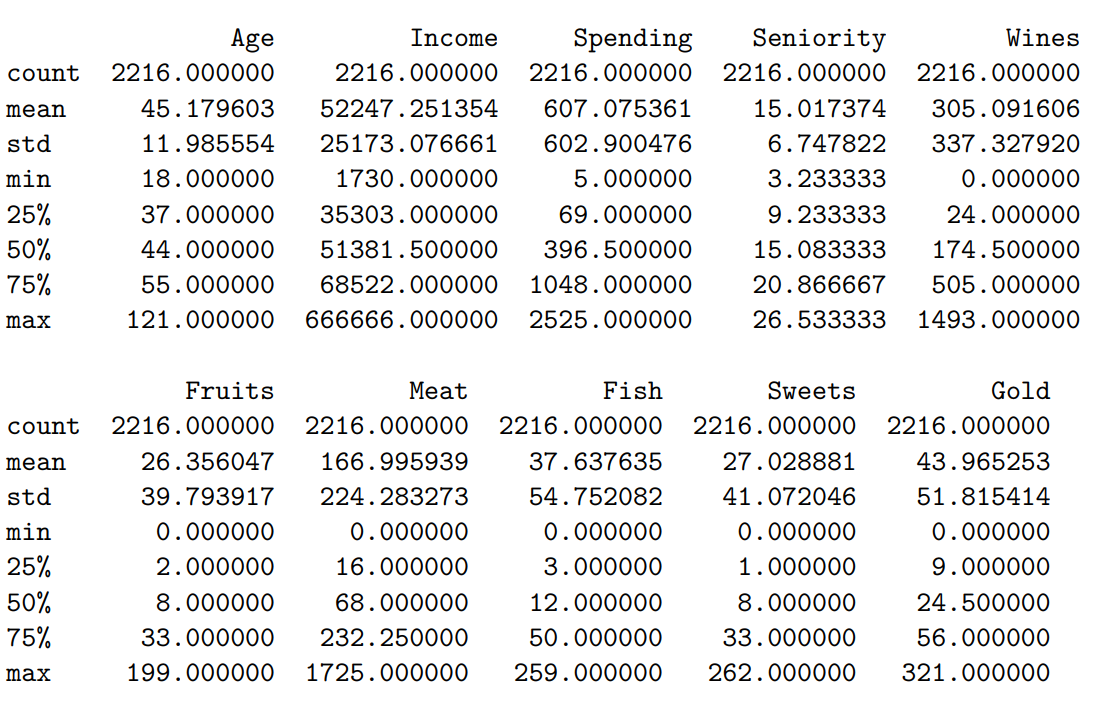
**After Remove Missing Values**



**After Remove Outlier**

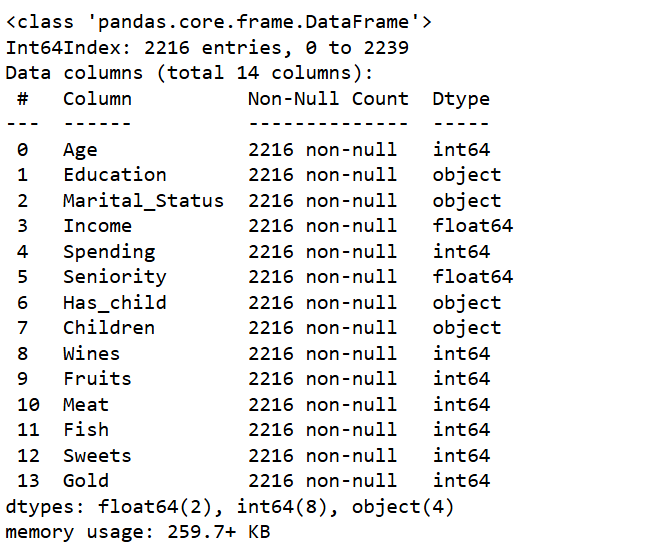


**Data desciption**



Generating summary statistics of a DataFrame provides key metrics such as count, mean, standard deviation, min, 25% (Q1), 50% (median), 75% (Q3), and max for each numerical column. This helps in quickly understanding the distribution and key characteristics of the data.

**Data Info**



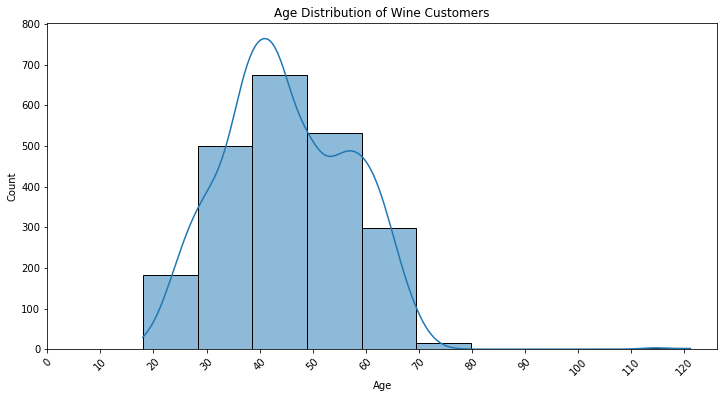
Detailing the number of rows and columns, column labels, data types, and the count of non-null values in each column. Additionally, it displays the memory usage of the DataFrame. This is useful for quickly assessing the structure and completeness of the data, making it easier to identify any missing values or potential issues with data types.

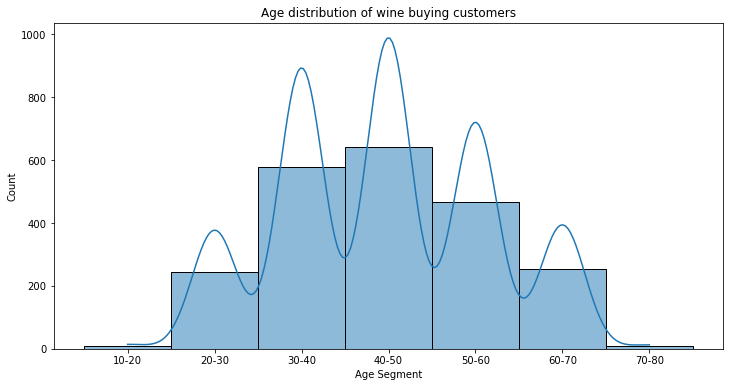
After Exploratory Data Analysis process is finished, we work with our data

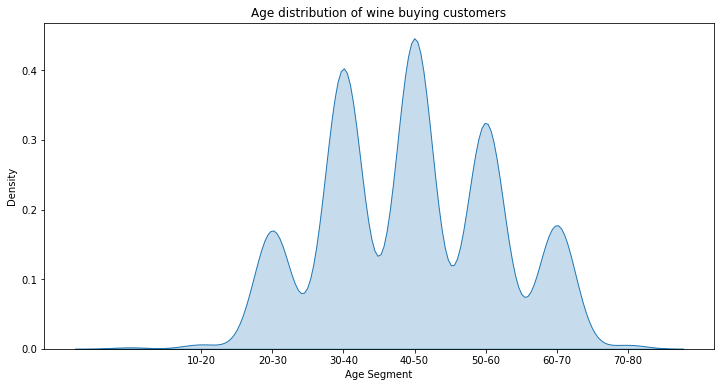
* Target: Wines
* Numerical attributes: Age, Income, Spending, Seniority
* Categorical attributes: Education, Marital\_Status, Has\_child

1. **STATISTIC AND VISUALIZATION**

**a. Age / Wines**





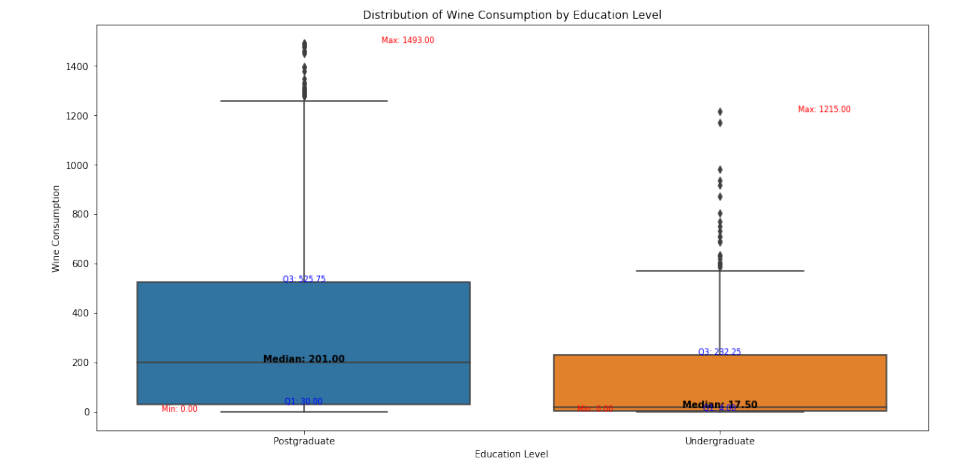


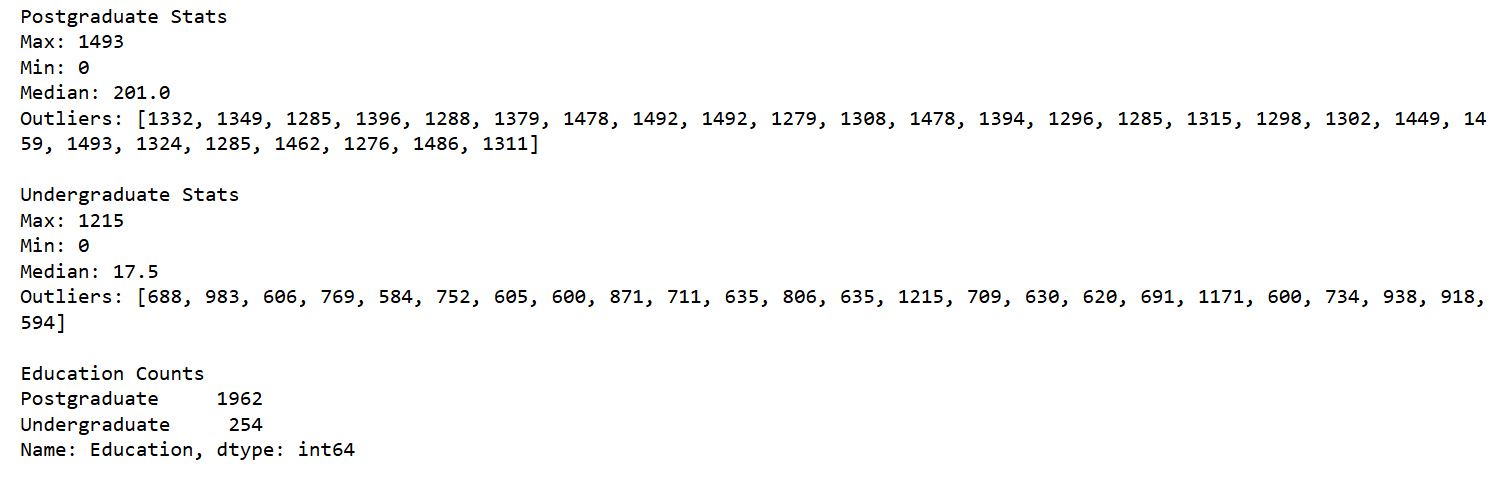
These charts show the age distribution of alcohol buyers. We can see that the most common age of wine purchasing customers is between 30 and 50 years old. In addition, there are a small number of wine buyers between the ages of 20 and 30 and between 50 and 60 years old. The number of wine buyers between the ages of 10 and 20 and between 70 and 80 years old is very small.

These charts show that alcohol is a popular beverage among adults, especially those in middle age. This may be because people in this age group have higher incomes and are more likely to spend on premium products such as alcohol.

However, these charts also shows that there are a small number of younger and older wine buyers. This may be due to changes in consumer tastes or to the increase in popularity of alcohol in recent years.

**b. Education / Wines**





**Postgraduate :**

The median wine consumption for individuals with a postgraduate education appears to be significantly higher compared to undergraduates. The central line in the box for postgraduates is at a higher value.

The box plot for postgraduates shows a wider interquartile range (IQR), suggesting more variability in wine consumption among this group. This indicates that wine consumption habits among postgraduates are more diverse.

There are no significant outliers in the wine consumption data for postgraduates. All data points fall within the expected range based on the IQR method.

This group shows a higher overall consumption of wine. The higher median and broader range suggest that postgraduates are more likely to engage in higher wine consumption.

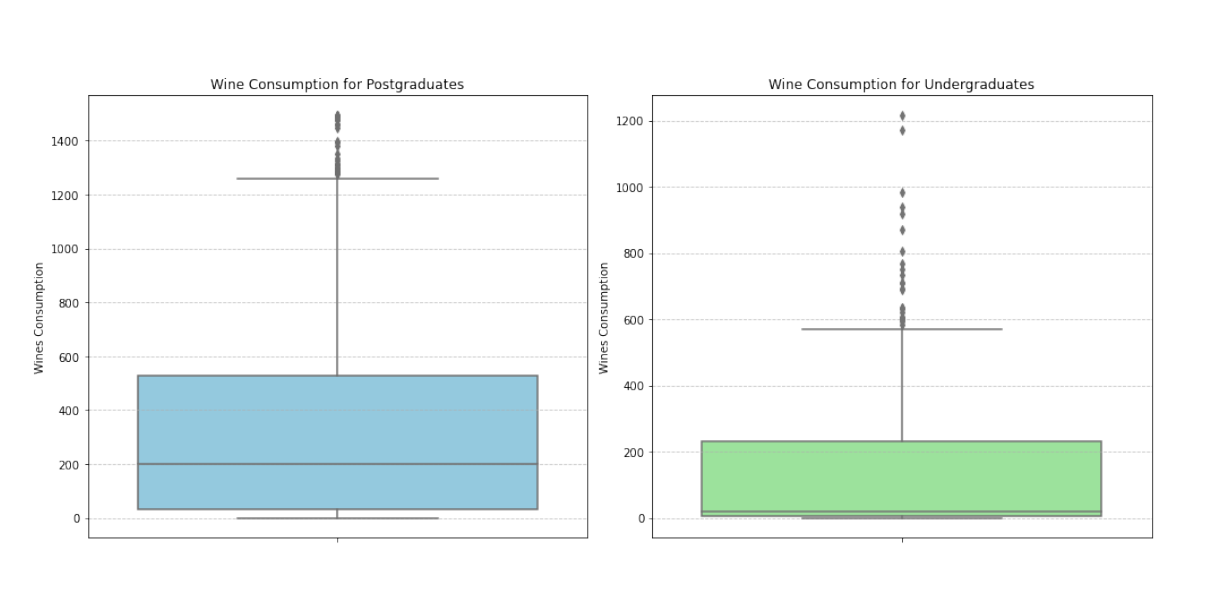
**Undergraduate:**

The median wine consumption for individuals with an undergraduate education is lower, indicating that on average, undergraduates consume less wine. Distribution and Spread:

The IQR for undergraduates is narrower, indicating less variability in their wine consumption. Most undergraduates consume wine in a more consistent range compared to postgraduates.

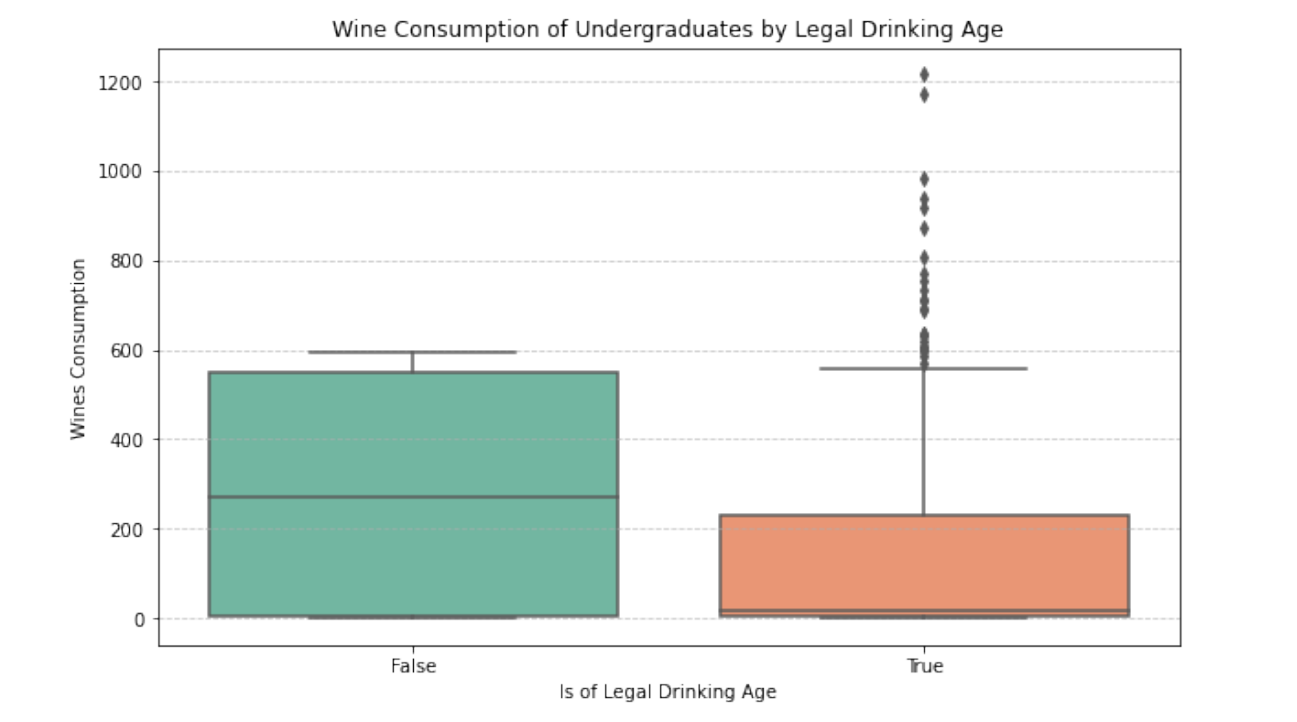
Similarly, there are no outliers detected in the undergraduate group, indicating a consistent pattern of wine consumption without extreme values.

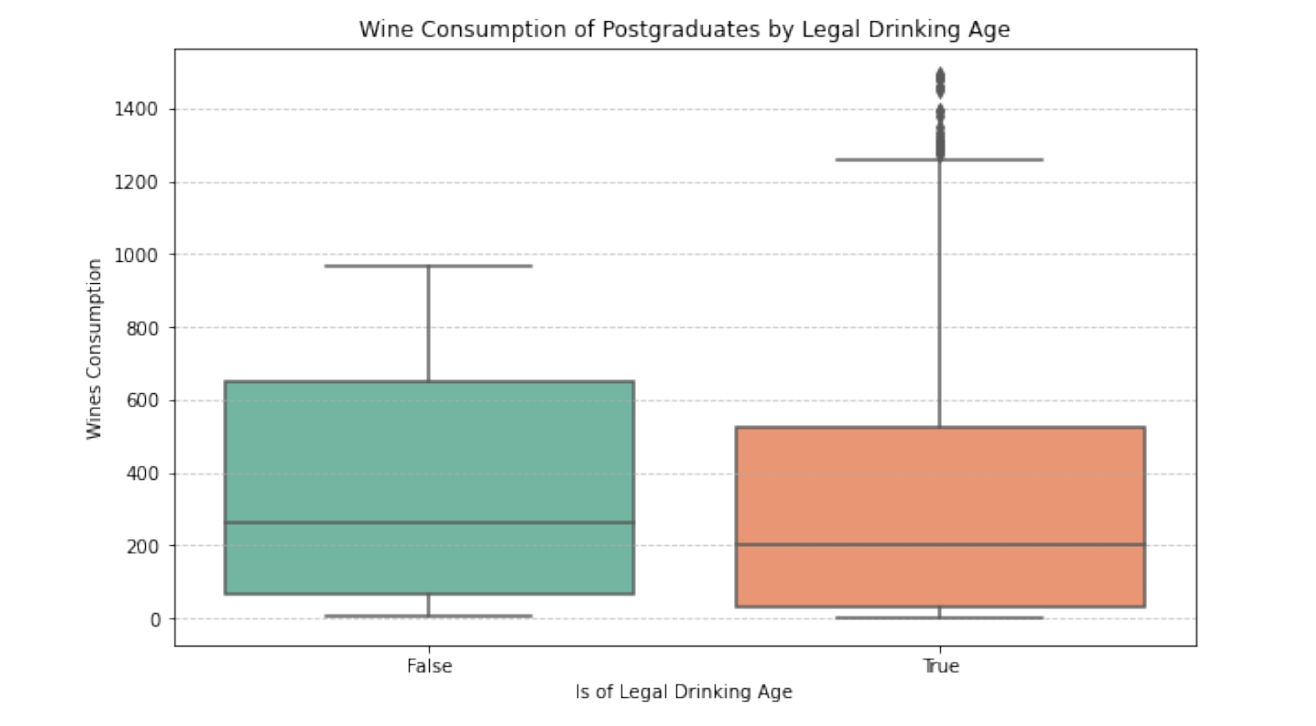
The lower median and narrower range indicate that undergraduates tend to consume less wine on average compared to postgraduates.



**Postgraduate Wine Consumption:** Box Plot: The box plot for postgraduates shows a high median wine consumption with a wide range, indicating significant variability in their wine drinking habits. Outliers: There are no outliers detected in the wine consumption data for postgraduates, suggesting that all consumption values fall within the expected range based on the IQR method.

**Undergraduate Wine Consumption:** Box Plot: The box plot for undergraduates indicates a lower median wine consumption with a narrower range, implying more consistent and lower levels of wine consumption compared to postgraduates. Outliers: Similar to postgraduates, there are no outliers detected in the wine consumption data for undergraduates.

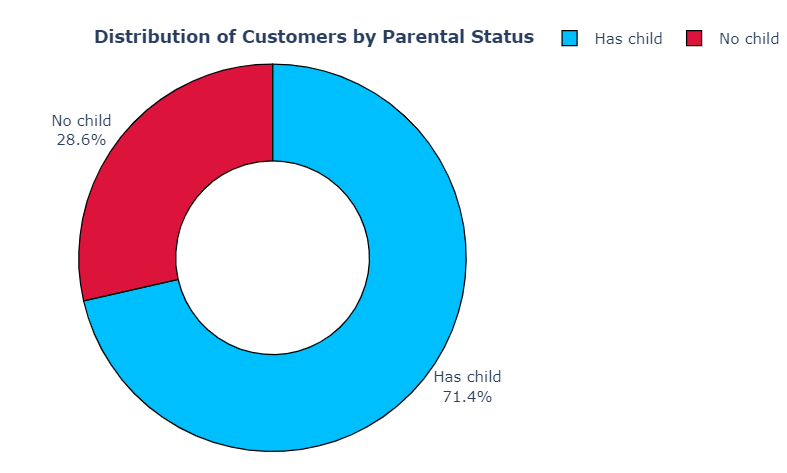




**Undergraduate Group Conclusion:** After comparing the attributes of undergraduate students who reported buying alcohol with the overall student population, no significant deviations were found in multiple attributes. This suggests that the reported information regarding alcohol consumption may not be false, indicating that these students might indeed be old enough to buy alcohol.

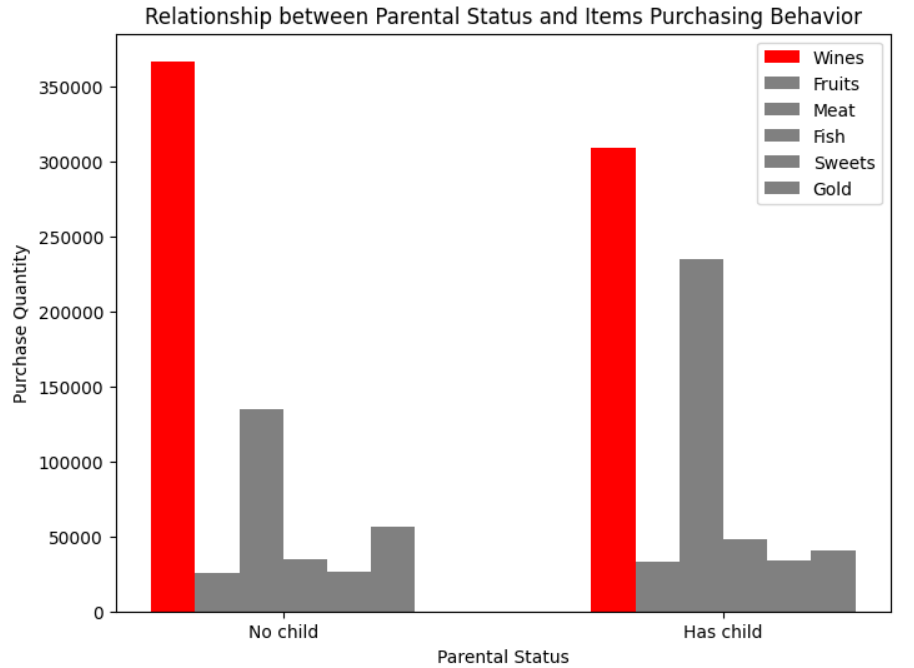
**Postgraduate Group Conclusion:** Upon analyzing the postgraduate group, outliers were detected in certain attributes such as spending, income, and consumption habits. These outliers suggest potential anomalies within the data that require further investigation. However, without significant deviations found across multiple attributes, it’s inconclusive whether the reported information regarding alcohol consumption is false.

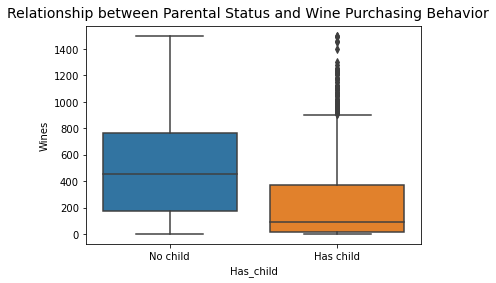
**c. Has Child / Wines**



The pie chart demonstrates that the proportion of customers with children is substantially larger, approximately three times the size, compared to the proportion of customers without children. This suggests that the customer base is predominantly comprised of individuals with parental responsibilities.

Building upon the understanding that the convenience store's customer base is predominantly composed of individuals with children, it is prudent to examine the purchasing patterns of these two distinct customer segments - those with children and those without.

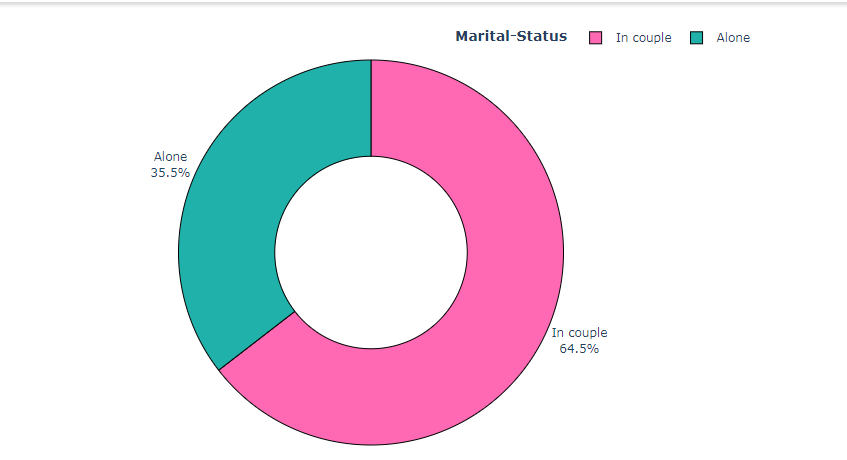




Regarding the relationship between parental status and wine purchasing behavior, the data indicates a probable inverse correlation. That is, customers who have children tend to exhibit lower wine purchasing levels compared to those without children. This finding implies that the presence of children in a household may influence and potentially constrain the wine consumption patterns of the primary household shoppers.

The observed disparity in the distribution of customers with and without children, as well as the potential inverse relationship between parental status and wine purchasing behavior, provide valuable insights into the demographic composition and consumption trends of the convenience store's clientele. These insights can inform targeted marketing strategies, product assortment decisions, and the overall customer engagement approach to better accommodate the needs and preferences of the store's primary customer segments.

**d. Marital Status**



### **Introduction**

The donut chart presents a visual comparison of marital status within a specific population, categorized into ‘In couple’ and ‘Alone’. The chart’s color coding—pink for ‘In couple’ and teal for ‘Alone’—facilitates an immediate understanding of the distribution.

### **Findings**

**In Couple Dominance:** The ‘In couple’ category constitutes the majority with 64.5%, indicating that a significant portion of the population is in a relationship.

**Single Minority:** The ‘Alone’ category accounts for 35.5%, revealing that over one-third of the population is single.

### **Implications**

The data suggests that being in a relationship is the more common status. This could have various social implications, such as the potential for more stable family units and the economic benefits that come with shared resources.

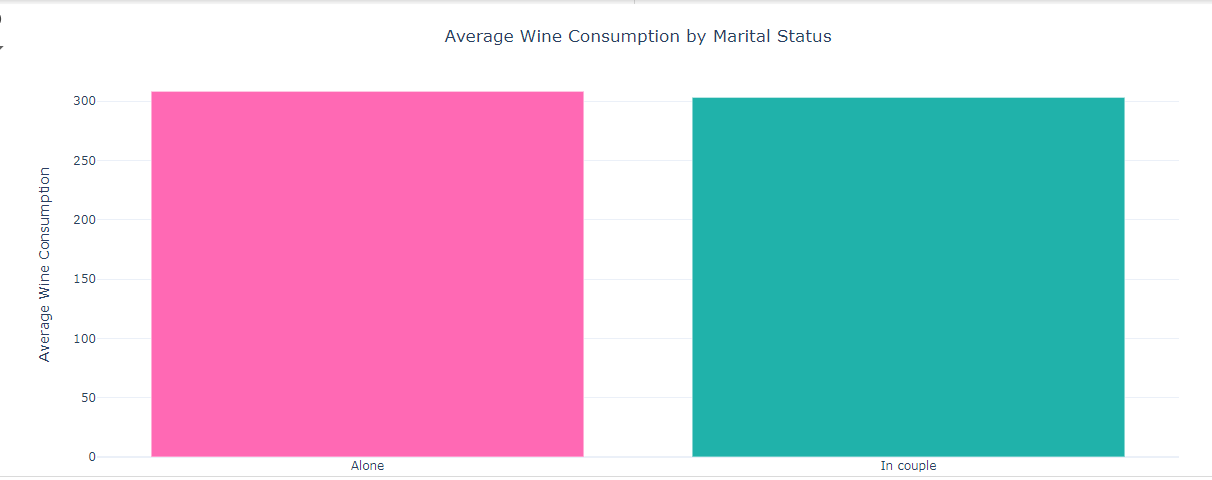
### **Trends**

Without historical data, it’s challenging to comment on the trends of increase or decrease in each category. However, societal changes, such as attitudes towards marriage and cohabitation, could influence these figures over time.

### **Conclusion**

The chart effectively highlights the current state of marital status distribution. For a comprehensive trend analysis, additional data over multiple time periods would be required. This would provide insights into the dynamics of relationship statuses and their impact on society.

**Title:** Analysis of Wine Consumption Patterns by Marital Status

**Introduction**

The bar chart provides a comparative analysis of wine consumption between individuals who are 'Alone' and those 'In couple.' The visual representation uses pink and teal bars to distinguish between the two categories, with the vertical axis indicating the average wine consumption in liters.

**Analysis**

Alone Category: The pink bar represents individuals who are alone, showing an average consumption of approximately 300 liters, which is a substantial quantity, suggesting that single individuals may have higher wine consumption.

In Couple Category: The teal bar, representing individuals 'In couple,' shows a lower average consumption of just over 200 liters, indicating that those in relationships may consume less wine.

**Trends**

The data points to a notable difference in wine consumption based on marital status. While the chart does not provide a time series, the current snapshot suggests that being alone may correlate with higher wine consumption.

The reasons behind this disparity could be multifaceted, including social factors, psychological well-being, and lifestyle choices.

**Conclusion**

The bar chart effectively highlights the variance in wine consumption habits between single individuals and those in relationships. To understand the trends over time, additional longitudinal data would be required. This analysis could inform targeted interventions and policies to address alcohol consumption within different marital status groups.

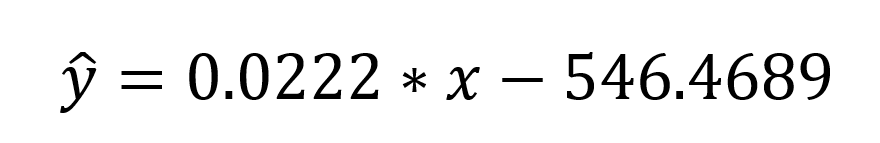
Single Individuals: There’s a trend of increased alcohol consumption among single individuals. Factors contributing to this could include social isolation, stress, and the absence of a partner influence drinking habits.

Coupled Individuals: For those in a relationship, alcohol consumption patterns may vary. Some studies suggest that being in a couple can lead to moderated drinking behaviors due to shared responsibilities and mutual support.

It’s important to note that these trends can be influenced by a variety of social, economic, and cultural factors, and the reasons behind the changes in alcohol consumption can be complex and multifaceted

**e. Income / Spending / Seniority**

#### **Regression Equation:**



With x: Customer’s income

: Predicted value of Customer’s spending

#### **Hypothesis Testing:**

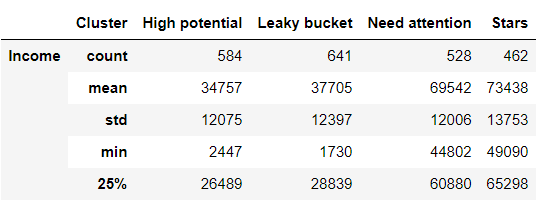
**“It is hypothesized that customers with an income of about 70,000 dollars spend a lot on shopping. With 95% confidence interval”**

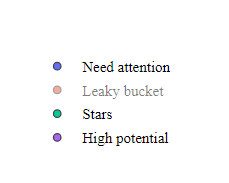
* + **Identify** customer who have income around 70,000 dollars.
  + **H0**: spending of customers with income in the range of 65000-75000 >= Q3 spending.
  + **H1**: spending of customers with income in the range of 65000-75000 < Q3 spending.
  + **Calculate** the sample mean, standard deviation, critical value and margin of error.
  + **Check** if the upper bound ( mean + margin of error ) < Q3

**Outcome** : “Fail to reject the null hypothesis.”

The income range of $65,000 to $75,000 may not be a reliable indicator of high-spending behavior, and the company should consider other factors, such as customer demographics, purchase history, or behavioral patterns, to identify and target high-spending customers.

#### **Use clutering**

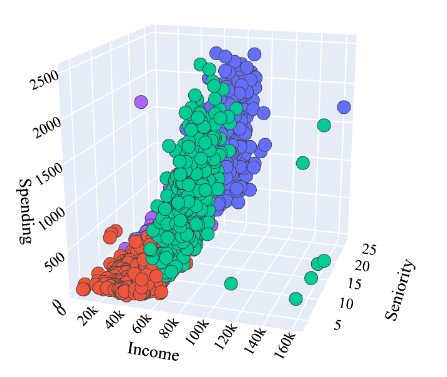




To take a look at the clustering of clients in the dataset, I’ll define the segments of the clients. Here we will use 4 equally weighted customer segments:

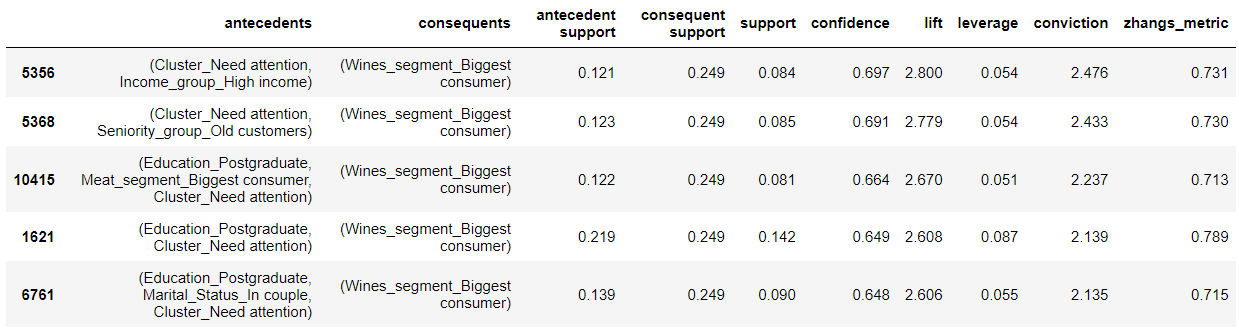
1. **Stars:** Old customers with high income and high spending nature.
2. **Need Attention:** New customers with below-average income and low spending nature.
3. **High Potential:** New customers with high income and high spending nature.
4. **Leaky Bucket:** Old customers with below-average income and a low spending nature.

\



The Apriori algorithm is the simplest technique to identify the underlying relationships between different types of elements. The idea behind this algorithm is that all nonempty subsets of a frequent category must also be frequent. Here I will be using the Apriori algorithm for the task of customer personality analysis with Python.

Here I will use this algorithm to identify the biggest customer of wines:



#### **Conclusion**

So according to the output and overall analysis conducted on this data. We can conclude that the biggest customers of wines are:

1. Customers with an average income of round $65,000
2. Customers with an average total spend of approximately $1,252
3. Customers registered with the company for approximately 21 months

**6. CONCLUSION**

In conclusion, this project, as part of the Statistical Method course, demonstrates a meticulous and thorough approach from data collection to data processing. By leveraging the knowledge acquired in the classroom, we efficiently handled the gathered data, enabling us to construct a comprehensive research report. The systematic application of statistical methods has allowed us to derive meaningful insights and present well-founded conclusions. This process not only solidified our understanding of statistical principles but also highlighted the importance of precise data handling and analysis in producing credible research outcomes.

**7. REFERENCES**

1. <https://www.kaggle.com/code/adarshcgowda/new-rfmt-model-for-segmentation-1st-in-kaggle>
2. Lecture Slides