## /***WRITTEN PROJECT REPORT***

### **Purpose of Team Project**

In the team project you will work jointly to collect, describe, analyze, and interpret results from a dataset of your choice using the tools learned in BUDT758T. You will have to demonstrate your ability to phrase the practical questions as scientific hypotheses; to select the models that are suitable for presenting and analyzing your data; and to interpret the analyses and use them for decision-making. The project is an opportunity to display mastery of technique, as well as creativity and originality. An executive summary should be used to present your key findings using non-technical language.

### **Structure of the Proposal Written Document**

The proposal should be printed on single-sided 8 1/2'' x 11'' paper using at least an 11-point type font. It should be **stapled** with pages numbered. An electronic version of the document has to be uploaded on Canvas. The written document should effectively communicate your ideas, using appropriate visualizations.

1. **Length of Report**

There is no pre-set length of the report. Approximately 7-10 pages (1.5 spacing) is the norm for the narrative. Teams may additionally attach a results appendix with output from the analysis. Please be selective if attaching output, and present it so that it is readable.

The report is meant to be minimalistic – it should get the point of your study, its significance, and the findings as concisely as possible. A short description is all I am looking for (and definitely *not* an account of everything you thought and did). Since students often ask be how long the study must be, I am providing the above suggested guidelines.

1. **Deliverables and deadlines**

Reports are due in class on the presentation date (TBA). The report must be uploaded prior to the presentation. Please also upload your PowerPoint deck, data (unless protected by confidentiality agreements) and scripts to replicate your key results.

1. **Parts of the Narrative**

See the following pages (lengths are meant to be suggestive; feel free to depart from these suggestions).

1. **Suggested Cover Page:**

Data Mining for Business (BUDT758T)

Project Title: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Be explicit about the explanatory or predictive flavor of your, using terms such as “explaining”, “predicting,” etc.).

Team Members: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(SIGN THE FOLLOWING STATEMENT AND INCLUDE IT ON THE COVER PAGE OF YOUR PROPOSAL)

## ***ORIGINAL WORK STATEMENT***

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

|  | **Typed Name** | **Signature** |
| --- | --- | --- |
| Contact Author |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. **Executive Summary**

In today's fiercely competitive business landscape, understanding customers' individual preferences and buying behaviors is paramount. Personalized customer experiences are no longer a luxury but a necessity, and a key differentiator for businesses. Our Customer Personality Analysis project, therefore, is a strategic initiative aimed at harnessing customer data to personalize services, enhance customer satisfaction, and ultimately, drive growth.

We procured a rich dataset from Kaggle, comprising over 2000 customer records, with detailed attributes such as age, income, number of children, purchasing habits, response to promotions, and preferred order placement channels. Our initial analysis involved an exploratory data analysis (EDA) and data cleaning, ensuring our study's foundational integrity. The EDA allowed us to uncover significant trends, most notably, the dominance of wine purchases in the last two years, leading us to focus primarily on this category.

The next phase of our project involved Recency-Frequency-Monetary (RFM) segmentation based on the customer data. This widely recognized marketing technique allowed us to categorize customers based on their purchase history, identifying the recency, frequency, and monetary value of purchases. This understanding of customer behavior, particularly regarding wine purchases, was instrumental in our subsequent steps.

Following RFM segmentation, we built several predictive models, including regression trees, naive bayes, logistic regression, and random forest algorithms. Each model offered unique insights and was evaluated based on performance and suitability for various purposes. This multi-model approach provided us a comprehensive view of customer behaviors and allowed us to predict future buying patterns accurately.

Our next step was to construct association rules, which helped us delve deeper into customer personality analysis. These rules served to identify patterns and connections between different customer attributes, enabling us to determine the personalized service to offer new customers.

In conclusion, the Customer Personality Analysis project underscores the importance of a deep, nuanced understanding of our customers in an intensely competitive market. By leveraging cutting-edge data analysis techniques, we can tailor our services to meet individual customer needs better, thereby increasing loyalty and driving sales. We remain committed to enhancing our analytical capabilities and incorporating more sophisticated technologies and information to continually refine our understanding of customer personalities. This continuous improvement is key to remaining competitive and delivering superior customer experiences.

1. **Data Description (1 page)**

Data Source: [BUDT758T\_TeamS\_Data\_Source](https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?datasetId=1546318&sortBy=commentCount&searchQuery=rfm)

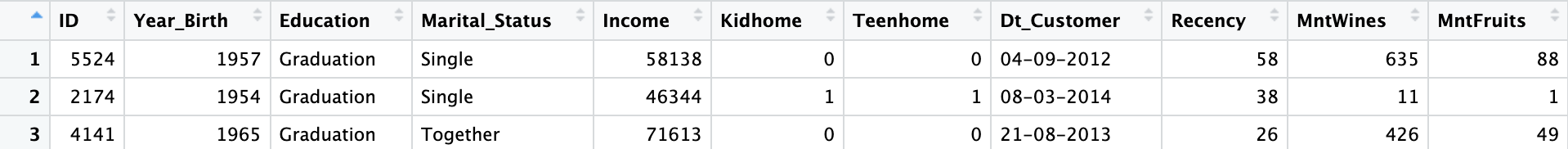
Number of Variables (k): 29

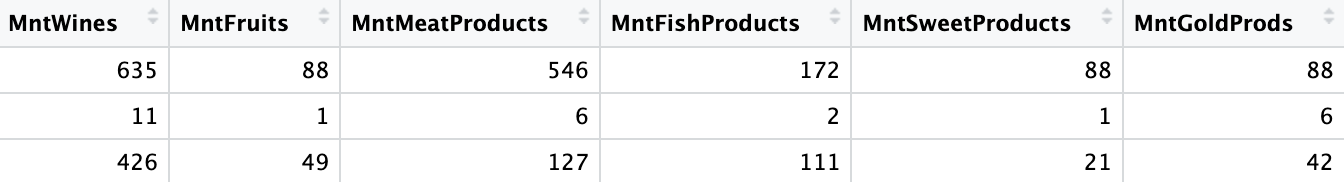
Sample Size (n): 2240

Data Dictionary:

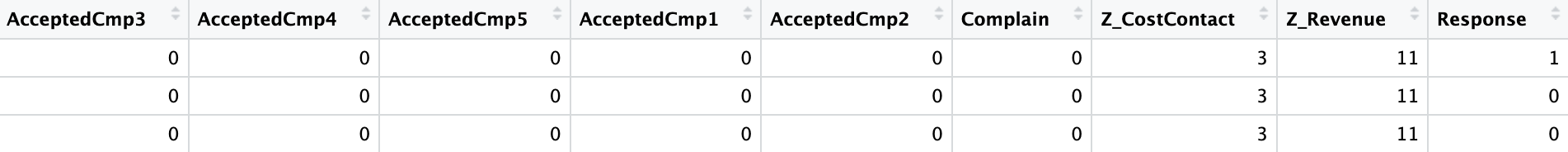
| Column Name | Type | Unit | Column Description |
| --- | --- | --- | --- |
| ID | int | NaN | Customer's unique identifier |
| Year\_Birth | int | NaN | Customer's birth year |
| Education | chr | NaN | Customer's education level |
| Marital\_Status | chr | NaN | Customer's marital status |
| Income | int | USD | Customer's yearly household income |
| Kidhome | int | NaN | Number of children in customer's household |
| Teenhome | int | NaN | Number of teenagers in customer's household |
| Dt\_Customer | chr | NaN | Date of customer's enrollment with the company |
| Recency | int | NaN | Number of days since customer's last purchase |
| Complain | int | NaN | 1 if the customer complained in the last 2 years, 0 otherwise |
| MntWines | int | USD | Amount spent on wine in last 2 years |
| MntFruits | int | USD | Amount spent on fruits in last 2 years |
| MntMeatProducts | int | USD | Amount spent on meat in last 2 years |
| MntFishProducts | int | USD | Amount spent on fish in last 2 years |
| MntSweetProducts | int | USD | Amount spent on sweets in last 2 years |
| MntGoldProds | int | USD | Amount spent on gold in last 2 years |
| NumDealsPurchases | int | NaN | Number of purchases made with a discount |
| AcceptedCmp1 | int | NaN | 1 if customer accepted the offer in the 1st campaign, 0 otherwise |
| AcceptedCmp2 | int | NaN | 1 if customer accepted the offer in the 2nd campaign, 0 otherwise |
| AcceptedCmp3 | int | NaN | 1 if customer accepted the offer in the 3rd campaign, 0 otherwise |
| AcceptedCmp4 | int | NaN | 1 if customer accepted the offer in the 4th campaign, 0 otherwise |
| AcceptedCmp5 | int | NaN | 1 if customer accepted the offer in the 5th campaign, 0 otherwise |
| Response | int | NaN | 1 if customer accepted the offer in the last campaign, 0 otherwise |
| NumWebPurchases | int | NaN | Number of purchases made through the company’s website |
| NumCatalogPurchases | int | NaN | Number of purchases made using a catalogue |
| NumStorePurchases | int | NaN | Number of purchases made directly in stores |
| NumWebVisitsMonth | int | NaN | Number of visits to company’s website in the last month |
| Z\_CostContact | int | NaN | Null |
| Z\_Revenue | int | NaN | Null |

Samples:





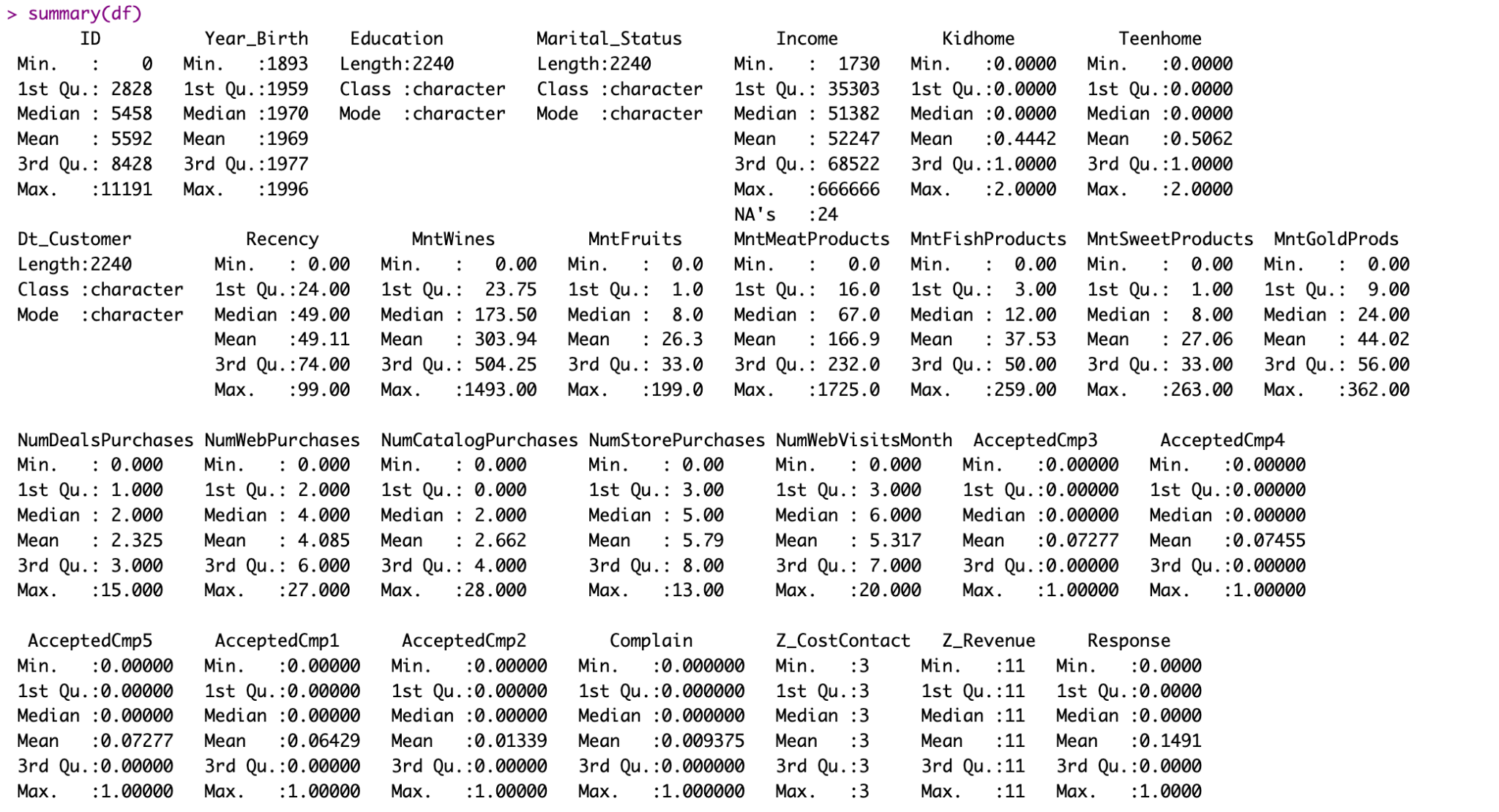


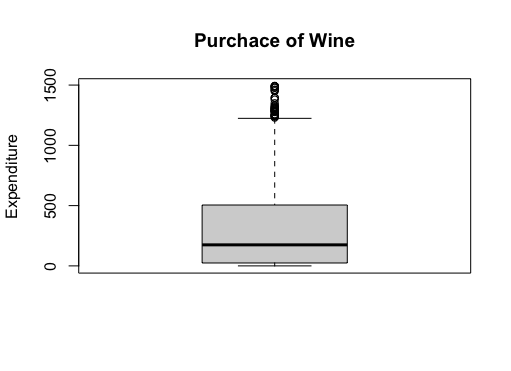
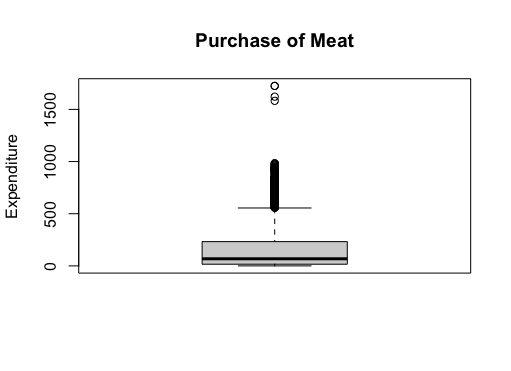
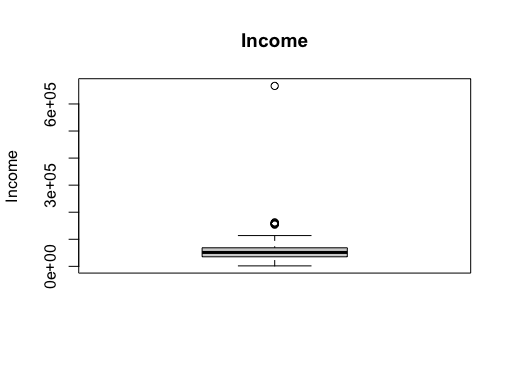


Interest:

The dataset encompasses various facets for detailed scrutiny. It incorporates characteristics about the customer, such as age, income, and education. Additionally, it includes information on the customer's consumption patterns, such as expenditure across different categories, preferred locations for placing orders, and their inclination towards promotional offers. Leveraging this data can enable us to deliver a more insightful and accurate interpretation of customer behavior analysis.

EDA:





Data cleaning and preprocessing:

Dropping Na’s (24 rows)

Column transformation: Year\_born -> Age

Dropping outliers:

Upper limit: 85% quartile + 1.5\*IQR; Lower limit: 15% quartile - 1.5\*IQR

Dropping duplicated columns: Z\_CostContact, Z\_Revenue

Provide information on the data you used for your study. This should include

* + Data source (where you got the data from)
  + What the data are (what is measured by each variable, in what units, whether you will treat it numerical or categorical, etc.)
  + Sample size (*n*) and number of variables (*k*)
  + A small sample of observations from the data (a few observations to demonstrate the above points).
  + Why the data are of interest.

If the dataset was gathered on the Internet, please provide a complete link to the website. If it was entered manually, please provide a complete reference.

Unless you are bound by non-disclosure agreements, please submit the data and scripts to replicate results.

# **III. Research Questions (1 page)**

# From a small neighborhood convenience store to a famous top multinational retail corporation that operates a chain of hypermarkets, they all share one thing in common which is a group of satisfied, happy customers. Without a steady customer flow, no business can survive. And this is why what we do is important – Customer Analysis.

# We provide our clients(from groceries stores to big name operations) with detailed customer analysis results. Not just presenting them with a bunch of raw numbers or random model outputs, instead we help them understand how these numbers would help them to improve their business performance and how to apply them in the real world in solid steps. Other than that, we also provide our clients with detailed promotional strategies and future marketing policies that they would act right now to increase their KPI and ROI.

# Speaking of improving KPI and ROI, which is a big blueprint for every business, but how can we really help? And what is our starting point? We will take three main steps into place, they are past customer analysis, future customer prediction, and potential marketing strategies.

# The first question is among all the customers that we have had in the past, who are our most important customers? Who would be our returning customers? Which group of customers are at risk of losing them? After we have sorted them out, what solutions can we develop to save those at risk and maintain our loyal customers?

In order to help our clients to improve their business performance, analyzing past customers is not enough, identifying future customers is also important because they would likely bring us future profit.

Therefore, we need to know what our current customers could tell us about them. After a little bit processing of the customer data, we have found out that a large portion of our customers are more likely to spend the max amount on the wine products among all of their purchases. Since wine would bring us the max revenue in this dataset, thus, we decide to predict for each customer whether they will spend the max on wine?

However, we would like to do more than just predicting our customer’s decision on spending max on wine or not, we also want to know what factors of the customers behaviors, background, characteris, any information that we could gather would affect their decision of spending max wine? Among those, what are our most useful customer factors in this prediction?

In reality, do customers always stay the same? The answer is no. For example, some customers might suddenly decide to be a vegan, therefore the amount that they spend on meat will decrease. How would this change affect their decision on spending the max on wine? There are customers who like to use deals, and coupons, and there customers who do not like discounted products. Does this kind of change change their choices on spending the max on wine?

What are our next steps after predicting potential customers who are more likely to purchase wine? What are the actionable advice that we could provide right now that can help our clients directly on next month marketing? How do we identify customer segments?

After accomplishing and finding the answers to the above questions, we have successfully developed a business model on the customer analysis that could actually help clients to act right now to improve their business performance in every aspect.

1. **Methodology (1 page)**

To gain a succinct understanding of our customers, we've chosen to employ the Recency-Frequency-Monetary (RFM) model as a quantification technique to better comprehend customer behavior. Our initial step involves constructing an RFM matrix, factoring in the customer's recency, total orders in correlation to their membership duration (frequency), and the overall expenditure (monetary value). Subsequently, we transform Recency, Frequency, and Monetary metrics into scores ranging from 1 to 5, utilizing *rank* and *cut*. These scores are then recorded as recency\_score, frequency\_score, and monetary\_score. We further streamline this by combining recency\_score and frequency\_score into a singular variable, which we denote as RF\_SCORE. Finally, to enhance our understanding of customer distribution, we leverage the RF analysis to generate pie charts for a more visually intuitive representation.

In order to better understand our customers, an appropriate and accurate model is required to determine strategies for target customers. To predict whether a customer will spend the most on wine, we used four supervised methods in machine learning, including regression tree, random forest, and logistic regression, so we can use the labeled datasets to train models to predict outcomes. Before generating the model, we split the dataset into 60% of training data and 40% of test data. After using the training data to train the model, we used a confusion matrix to calculate the accuracy score and the ROC curve to compare the performance of these models.

The regression tree can quickly reveal the relationship in the data and provide a human-readable tree, with information about what variables are used and how variables are used to predict the outcome. It is easy to understand and useful to present to our client. We also pruned our original tree to remove parts of the tree that do not provide powerful predictions. Random forest is similar to a regression tree, but it contains several decision trees to produce a predictive model. It is a good option to predict the outcome more accurately. Logistic regression estimates the probability that a customer will spend the most on wine. We used a cut-off of 0.5 to determine whether a customer spends the most on wine. If the probability of a customer spending the most on wine is greater than 0.5, we classify the outcome as 1, otherwise, the outcome is 0. We begin with a logistic regression of all variables and use backward feature selection to choose the most relevant variables. With the help of step() function, the model is more interpretable and the chance of overfitting is reduced. The naive Bayes model provides the conditional probability of a customer will spend the most on wine, based on a given class or category. The output is easy to interpret and easy for our client to understand.

We also discovered some interesting relationships between variables by using association rules, which is one of the unsupervised machine learning techniques that would also provide helpful insights. However, we have to convert the dataset into item matrix format. We first convert all numerical columns to categorical columns. For some numerical columns such as Age and Income, we use ifelse() function to determine the age level and income level, so they can be categorical columns. The transaction() function automatically creates the item matrix to be used in the association rule.

1. **Results and Finding (varies considerably in length depending on study)**

**Past Customer Analysis**

Upon constructing the RFM matrix and proceeding with RFM segmentation, we have segregated our customers into 10 distinct clusters. Presently, approximately 16% of these are steadfast and loyal customers, for whom we emphatically suggest our clients to formulate personalized communication and deliver outstanding customer service. Roughly 24% of the current members are potential loyalists; for them, we recommend our clients to devise a strategy that encompasses follow-up and nurturing tactics to sustain engagement. As for the customers who are less likely to contribute substantial value to the market (around 16% of the customer base), we advise our client to consider disengaging with them strategically, allowing for better resource allocation.

**Future Customer Action Prediction**

In order to analyze different aspects of our customers, for different approaches we had built different models. In order to predict whether our customers would spend the max amount of money on wine products we have set our dependent variable as the max spend on wine. The test accuracy provides us a percentage of chance that our prediction is correct or not. For example, for our random forest classification, it provides us a test accuracy of 0.8223 which means if we need to predict 1000 customers, and find out whether they are more likely to spend the max on wine, we will be able to 822 customers correctly. This also applies to other models too. Since our random forest model has the highest test accuracy and AUC score, we would like to use this method in prediction of our future return customers whether they would purchase wine at their next visit.

**Examining Relationship between Independent Variable and Dependent Variable**

By analyzing the output of these models, we have identified important customer behaviors, background, and characteristics that would play an important role in predicting our dependent variable. By looking at the split of the pruned tree, we have seen that decisions could be changed based on the income level, education background and number of teens at home. For the regression tree, each of these variables have a low p-value. For example, the naive bayes model’s summary clearly shows there are differences between the mean income of customers who are likely to spend max on wine or not. Other than that, we could conclude from the Random Forest Report that the Mean Decrease Gini of Income, Education, Teenhome is also high. Therefore, after interpreting the model’s output, we had found out that Income, Education, Teenhome, NumWebVisitsMonth would affect whether the customer would decide to spend the max on wine the most.

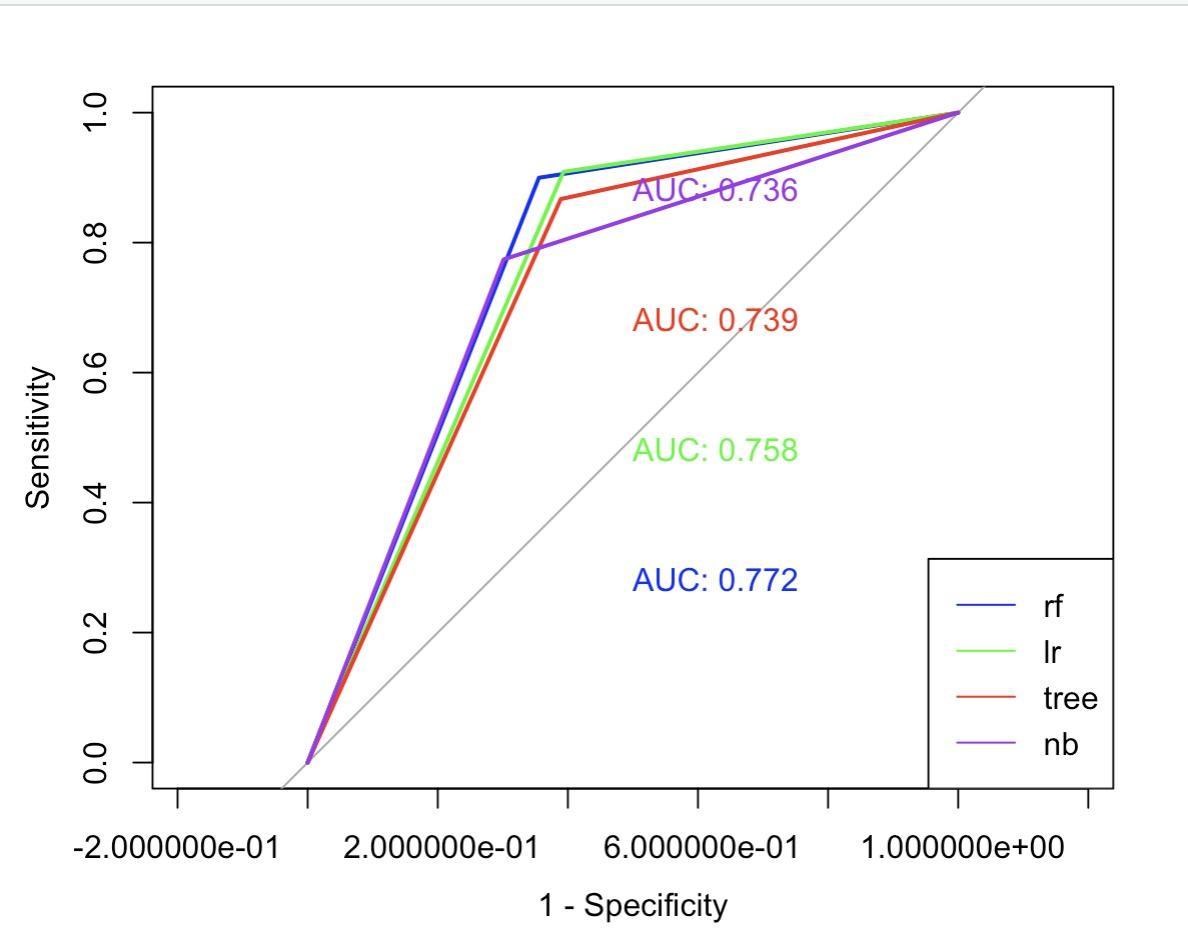
**Random Forest:**

* Customers with annual Income less than $30359.5 are less likely to spent more on wine.
* For Customers with annual income greater than $30359.5:
* Those have teenagers at home are more likely to buy wine
  + Those don’t have teenagers at home, with education level of basic, master or PhD are more likely to buy wine For Customers with annual income greater than $30359.5, no teenagers at home, education level of 2n cycle and college: Those annual income less than $47869.5 are less likely to buy wine Those annual income greater than $47869.5 are more likely to buy wine

We have interpreting the coefficient of the logistic regressions

We evaluated the performance of our models based on the accuracy score and ROC curve for these four predictive models. They are summarized as follows:

|  | Regression tree | Random forest | Naive Bayes | Logistic Regression |
| --- | --- | --- | --- | --- |
| Accuracy | 0.7893 | 0.8223 | 0.7513 | 0.8071 |
| AUC | 0.739 | 0.772 | 0.736 | 0.748 |



The regression tree is pruned to remove parts of trees that are not useful for prediction. The tree finally ends with 7 final nodes, which have the lowest deviance. For logistic regression, we used a backward stepwise feature selection method to remove unnecessary independent variables, and the model finally ends with variables including Education, Income, Teenhome, Recency, NumDealsPurchases, NumWebVisitsMonth, AcceptedCmp, and Response. Based on the p-value, most of the variables are statistically significant, except recency, NumDealsPurchases, and AcceptedCmp2. The random forest not only provides the prediction outcomes but also the feature importance for each variable. It shows that Income is the most important variable for the prediction.

According to the ROC curve, the ROC curve for these models is very close, but the curve for the random forest is further out the most compared to others. Overall, since the random forest has high accuracy, it is a powerful model to predict whether or not a customer will spend the most on wine.

For association rules, we discovered what will associate with the max spending on wine and what will associate with preferring deals. We set the minimum support threshold to 0.1 and the minimum confidence to 0.7. {Income=median, FishProducts=no, Complain=no}

=> {max\_spent=Wines}

**FUTURE CUSTOMER**  
  
In this section you would include tables from your data analysis. *But please be selective!* Too many tables detract from the study – use the appendix for this.

1. **Future Improvement and Business Model Application**
2. **Conclusion**
3. **Appendix (Any additional information to be submitted):**

**Grading Notes:**

The principal criterion is technical quality of the work. In addition, I will grade

* The interestingness and originality of the project. [There is an element of subjectivity here – however this is a criterion on which recruiters and others will judge you and, moreover, the onus is on you to convince me that the work is interesting and original.]
* The ambitiousness of the project. [All other things being equal, choosing a more challenging project will result in higher scores.]
* The effectiveness of presentation. This includes (1) the quality of your PowerPoint deck, and (2) the quality of oral presentation. [Was there a logical flow to the presentation? Did you engage the audience?]
* The quality of writing, and written and visual presentation of statistical analyses. [Clear, crisp and concise writing is rewarded. The report should be professional looking.]
* A portion of the grade will reflect peer evaluations (I will preserve confidentiality of feedback). [Project scores tend to be correlated with team cohesiveness and an ethos in which everyone is committed to making significant contributions.]

**Peer evaluations are to be submitted on Canvas no more than two days after the presentation in class.**