## Problem Statement

Tayko Software Catalog is preparing a direct mailing campaign to promote its new collection of games and educational software. The firm’s objective is to maximize profitability by leveraging its customer database, which was expanded through a consortium of catalog firms specializing in software products. With each catalog mailing costing $2, the company seeks to strategically target customers who are more likely to purchase and spend more. This requires developing two predictive models: the first to classify customers as purchasers or non-purchasers, and the second to predict the spending amount for those classified as purchasers. By combining these models, Tayko aims to optimize its campaign strategy and estimate gross profit from mailing to a pool of 180,000 prospects. The overarching challenge is to balance mailing costs with anticipated revenue, using data mining techniques to refine predictions and support business decisions.

By leveraging customer transaction history, demographic attributes, and marketing interactions, the firm aims to refine its targeting strategy, minimize mailing costs, and maximize gross profit while maintaining a personalized marketing approach. This project integrates machine learning, statistical modeling, and cost analysis to achieve these objectives.

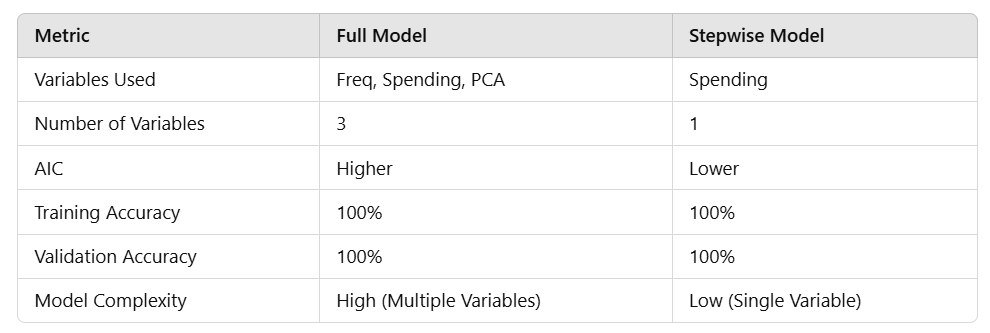
## Data Mining Tasks and Techniques

The primary data mining tasks in this project are classification and regression. Classification is used to predict whether a customer is a purchaser or non-purchaser, optimizing the mailing strategy to minimize costs. Regression is employed to predict spending among purchasers, enabling the firm to estimate potential revenue and prioritize high-value customers. Logistic regression was chosen for classification due to its effectiveness in binary classification and probability estimation, while multiple linear regression was used for spending prediction to model continuous outcomes. Stepwise regression was applied to both tasks for feature selection, ensuring model simplicity and robustness by excluding irrelevant predictors. Principal Component Analysis (PCA) was utilized to address multicollinearity among correlated variables, enhancing model stability. These techniques were selected for their ability to handle structured data, produce interpretable results, and provide actionable insights to refine the firm’s marketing strategy.

## Data Partitioning:

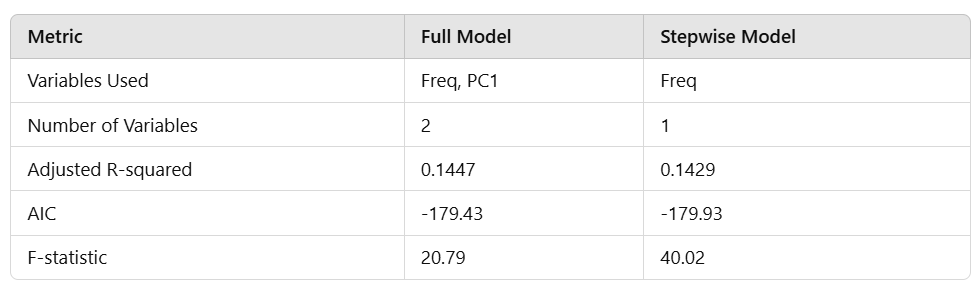
To ensure robust and generalizable models, the dataset was partitioned into three subsets: training, validation, and test sets. The training set, consisting of 60% of the data, was used to build and train the models, allowing the algorithms to learn patterns and relationships within the data. The validation set, comprising 35% of the data, was used to fine-tune the models and evaluate their performance during development, helping to prevent overfitting and ensuring that the models can generalize to unseen data. Finally, the test set, consisting of 25% of the data, was reserved for the final evaluation of the models. This partitioning strategy provided an unbiased estimate of model performance on new data and ensured the models' real-world applicability while balancing training accuracy with generalization. For the prediction model, data partitioning was specifically applied to the subset of customers who made purchases (Purchase = 1) to accurately predict their spending behavior. The dataset was divided into a training set (60% of the purchasers' data) and a validation set (40% of the purchasers' data). The training set was used to build the regression model, allowing it to learn the relationships between predictors, such as purchase frequency, and spending amounts. The validation set was used to assess and fine-tune the model’s performance, ensuring its accuracy and generalizability to new data. By focusing only on purchasers, this approach excluded non-purchasers (whose spending is zero), improving the model’s precision and relevance for predicting spending behavior within the target customer group. This partitioning ensures that the prediction model is both robust and applicable in real-world scenarios.

## Results interpretations for logistic regression



The results of the logistic regression model were analyzed both before and after stepwise variable elimination to compare the model's performance and interpretability. Initially, the model included multiple predictors (Freq, last\_update\_days\_ago, X1st\_update\_days\_ago, and Spending), but stepwise regression identified Spending as the only significant predictor for classifying customers as purchasers or non-purchasers. After variable elimination, the model retained its perfect accuracy (100% on training and validation datasets) but achieved lower complexity, making it more interpretable. This indicates that Spending is the primary driver for determining whether a customer is a purchaser. The reduced model simplifies implementation while maintaining its predictive capability, supporting a robust and actionable classification approach.

## Results interpretations for multiple regression



The results of the multiple regression model for predicting spending among purchasers were analyzed both before and after stepwise variable elimination. Initially, the model included multiple predictors (Freq and PC1), but stepwise regression identified Freq as the only significant predictor. The full model, which included Freq and PC1, explained 14.47% of the variance in spending, with an AIC of -179.43. After eliminating PC1 through stepwise regression, the adjusted R-squared remained almost unchanged at 14.29%, while the AIC improved slightly to -179.93, indicating a better model fit. The stepwise model, which retained only Freq, is simpler and more interpretable while maintaining comparable predictive performance. This suggests that Freq (the frequency of prior purchases) is the most significant predictor of spending among purchasers, providing a focused and efficient model for predicting spending.

## Score Analysis

Based on the "Score Analysis," the insights suggest that the company's current customer base offers moderate potential for profitability, with an average adjusted spending of $2.28 and an estimated gross profit of $50,057 from mailing to 180,000 customers. While the campaign is projected to be profitable, the profit margin is relatively slim. This indicates that while there is demand, the profitability relies heavily on accurate targeting of high-value customers. Introducing a new software product could attract additional interest and potentially increase spending, especially if the product aligns with customer preferences and adds value. However, to justify the introduction of a new product, the company should consider further segmentation to identify customers with the highest potential for spending, optimize marketing costs, and explore innovative product features that could boost interest and demand. Overall, introducing a new software product is feasible but must be coupled with strategic marketing and precise targeting to ensure profitability.

## Data Cleaning:

**Dataset:**

* The data set for this case includes just 1000 purchasers and 1000 non-purchasers.
* Total number of variables: 25 numerical and binary (1: Yes, 0: No).
* Two outcome variables:

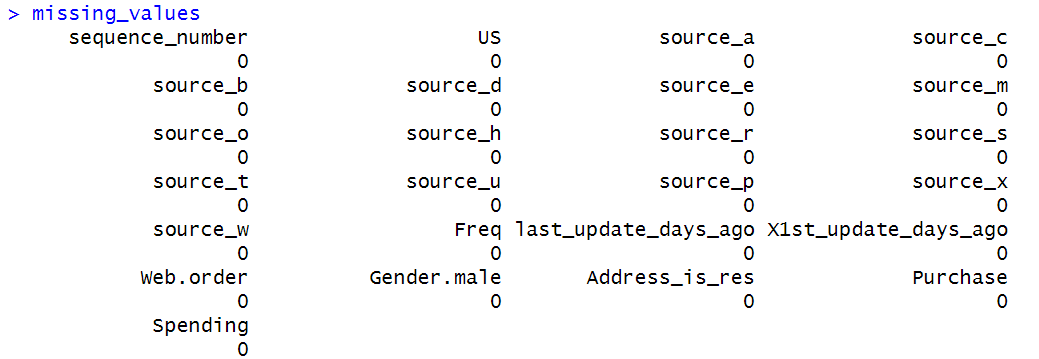
1. **Purchase**:
   1. indicates whether or not a prospect responded to the test mailing and purchased something
   2. will be used to classify records as purchase or no purchase.
2. **Spending**:
   1. indicates, for those who made a purchase, how much they spent. The overall procedure in this case will be to develop two models.
   2. will be used for those cases that are classified as purchase and will predict the amount they will spend.

### Preprocessing and Cleaning the Data:

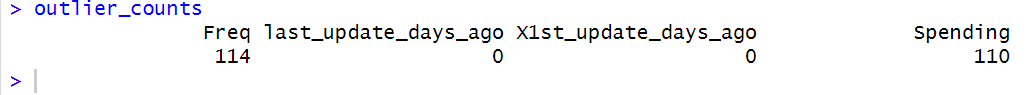
Types of Variables: 25 numerical and binary (1: Yes, 0: No).

### Missing Values:

There are no missing values.



### Outliers:



We began by cleaning the dataset to ensure it was suitable for analysis. This involved inspecting and addressing missing values, handling outliers, and transforming variables for consistency. Boxplots were created for the numerical variables (Freq, last\_update\_days\_ago, X1st\_update\_days\_ago, and Spending) to visualize the distribution and detect outliers. The boxplots revealed significant outliers in Freq and Spending, while last\_update\_days\_ago and X1st\_update\_days\_ago showed no extreme deviations.

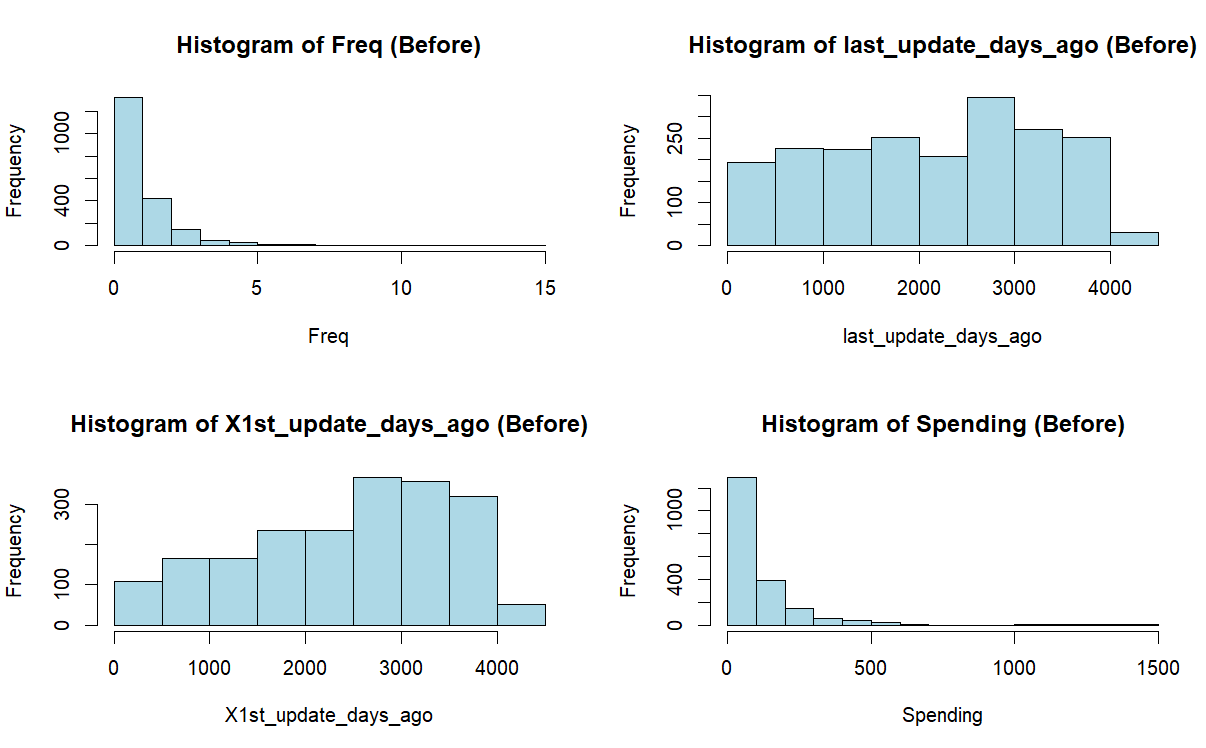
A group of graphs showing different types of data

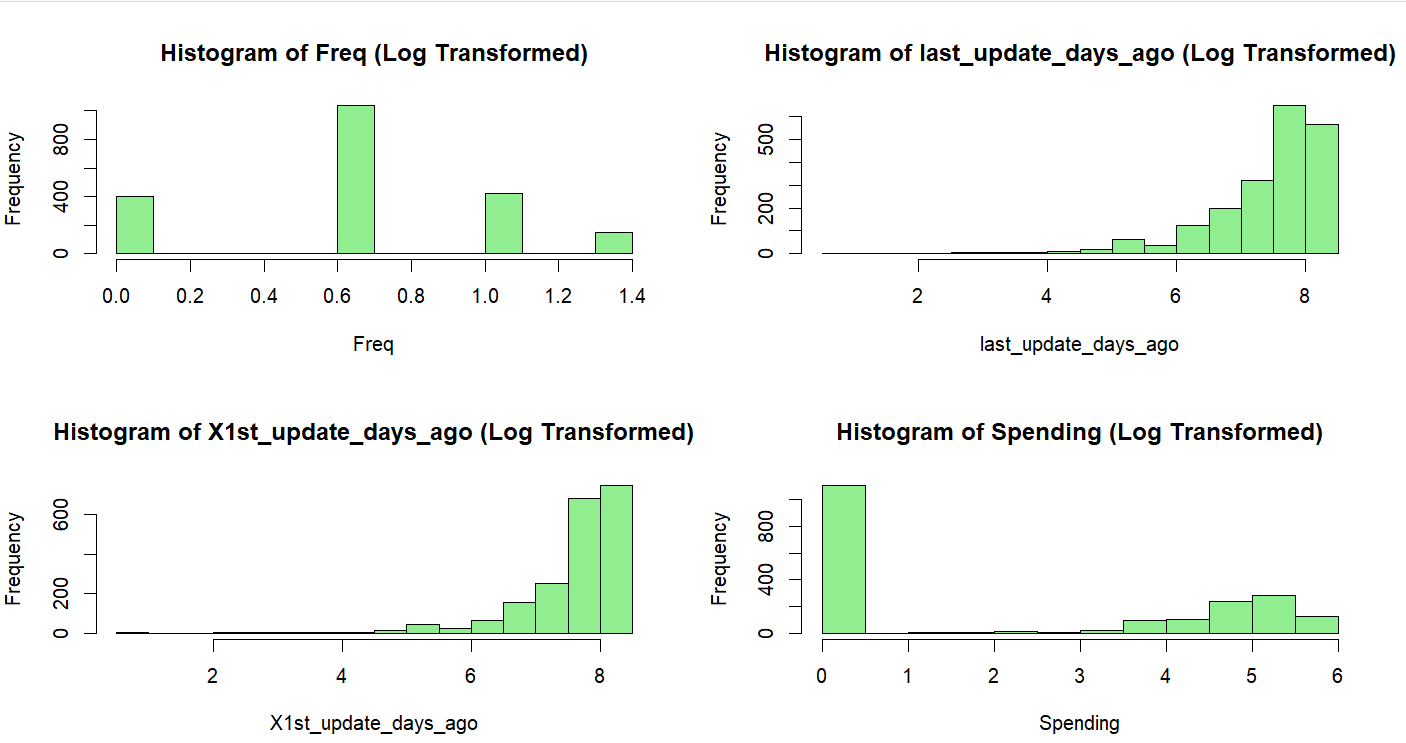
Description automatically generated with medium confidence

Outliers in Freq and Spending were handled by replacing them with the median values of their respective variables. Imputation with the median is a robust approach as it is less affected by extreme values compared to the mean. Additionally, we applied a log-scale transformation to Freq and Spending to reduce skewness and bring the distributions closer to normal, which is beneficial for many statistical and machine learning models. The histograms of these variables after transformation confirmed the improvement in distribution.

For last\_update\_days\_ago and X1st\_update\_days\_ago, since the boxplots did not indicate significant outliers, we retained their original scales to preserve their interpretability. These variables already showed a relatively uniform distribution and did not require further transformation.

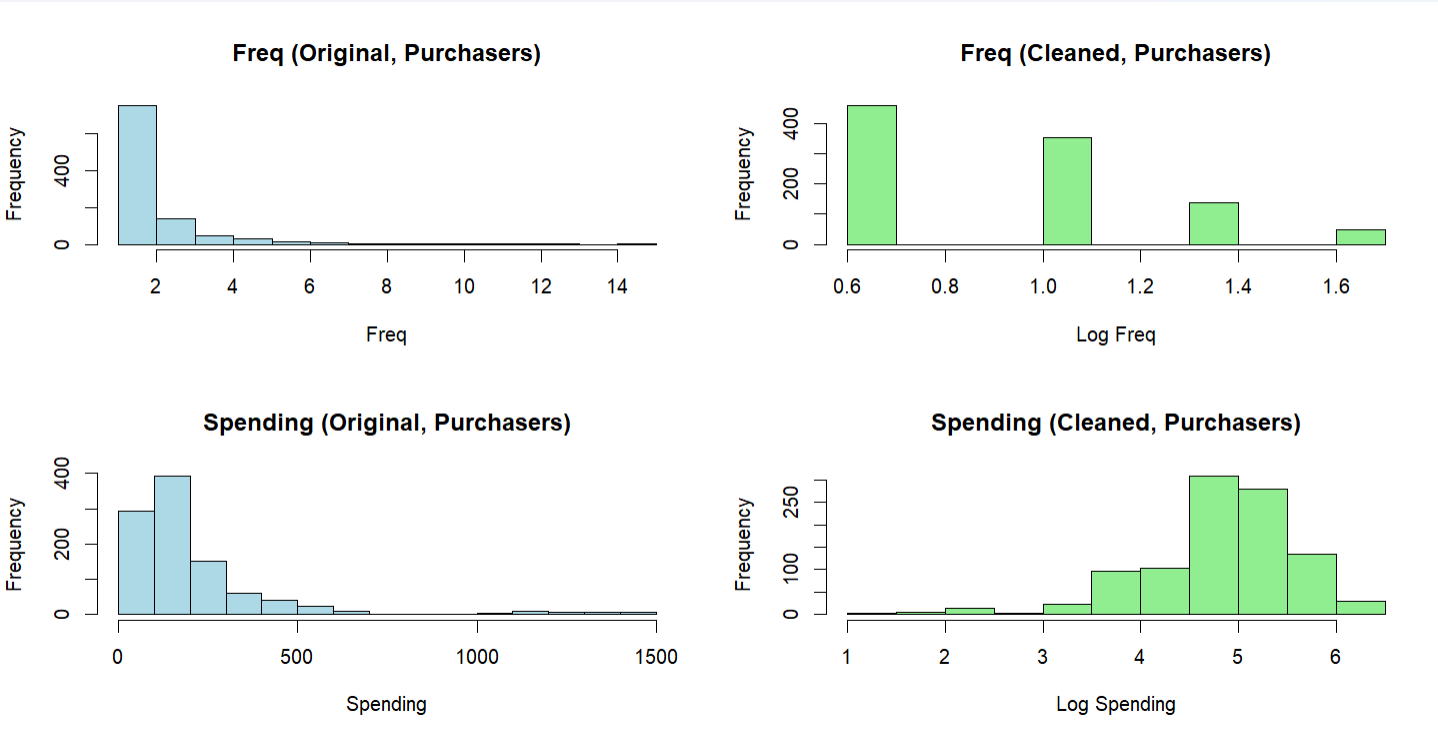
Solve Outliers with Imputation and Log Scale:



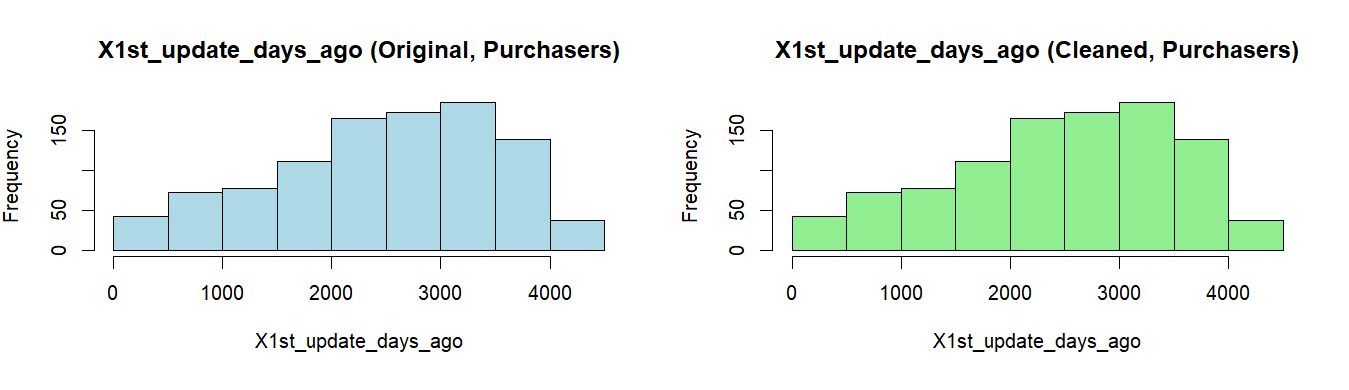


Upon examining the histograms of Spending, it became evident that Spending was always zero for non-purchasers. Including non-purchasers in the same analysis as purchasers would have skewed the results and misrepresented the distributions of key variables. For instance, the high frequency of zeros in Spending would have dominated its histogram and influenced transformations unnecessarily.

By separating the data into purchasers and non-purchasers, we ensured that the cleaning and transformations were tailored to the characteristics of each group. For purchasers, Spending was meaningful, and it required imputation for outliers and log transformation to address skewness.



We decided not to apply a log-scale transformation to the variables last\_update\_days\_ago and X1st\_update\_days\_ago because their distributions were relatively uniform and free of significant skewness, as observed in the histograms before cleaning. The cleaning process, which involved outlier removal and imputation, was sufficient to maintain their consistency and integrity without altering their scale. Consequently, the distributions after cleaning remained smooth and appropriate for further analysis without the need for log transformation.



A graph of a sales funnel

Description automatically generated with medium confidence

For non-purchasers, Spending remained zero, and we focused only on cleaning and transforming other variables (Freq, last\_update\_days\_ago, and X1st\_update\_days\_ago).

A graph of a bar graph

Description automatically generated with medium confidence

A group of graphs showing different sizes and colors

Description automatically generated with medium confidence

To validate our cleaning process, we compared the histograms before and after transformations for both purchasers and non-purchasers. The log-transformed variables (Freq and Spending) showed improved distributions with reduced skewness, while variables like last\_update\_days\_ago and X1st\_update\_days\_ago retained their natural distribution and interpretability. The separation of data groups ensured that cleaning and transformation decisions were aligned with the characteristics of each group.

## Question 1

Each catalog costs $2 to mail. Estimate the gross profit that firm could expect from the remaining 180,000 names if it selects them randomly from the pool (hint: multiple the response rate of 0.053 by the mean of two variables spending and only purchasers).

We first need to calculate Mean Spending how much purchasers typically spend on average. However, the Spending variable was log-transformed during cleaning, the mean would represent the log scale rather than the original spending values. To address this, we need to reverse the log transformation before calculating the mean.

Tayko drew 20,000 names from the pool and did a test mailing of the new catalog which yielded 1065 purchasers, a response rate of 0.053.

For the remaining 180,000 names, and for ease of presentation, the dataset for this case includes just 1000 purchasers and 1000 non-purchasers, an apparent response rate of 0.5. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case’s “probability of purchase” by 0.053/0.5, or 0.107.

Total mailing cost = number of catalogues \* cost per catalogue = 180,000 \* 2

From the test mailing, Tayko observed a response rate of 0.053. If we assume a similar response rate for the remaining 180,000 names, we can estimate the number of purchasers as Expected number of purchasers = 180,000 \* 0.053

We will use the average spending calculated after filtering the data to “purchaser” to estimate the revenue:

Total expected revenue = number of catalogs \* response rate \* average spending

mean\_spending\_purchasers = 155,288

Total expected revenue = 1,481,448

Mailing cost = number of catalogs \* 2

Gross profit is calculated as the total expected revenue minus the total mailing cost:

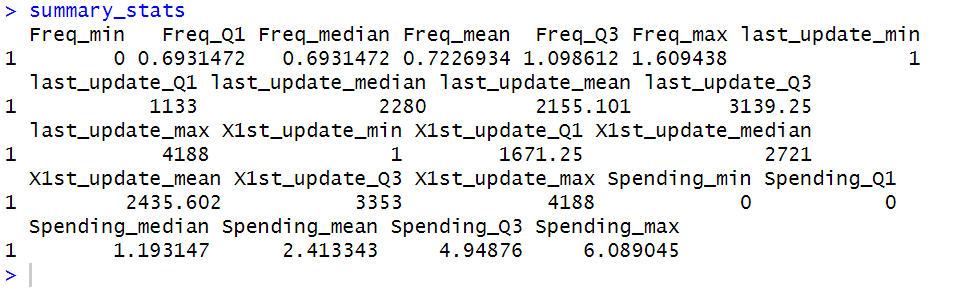
Thus, the estimate gross profit the firm could expect from the remaining 180,000 names is $.

## Question 2:

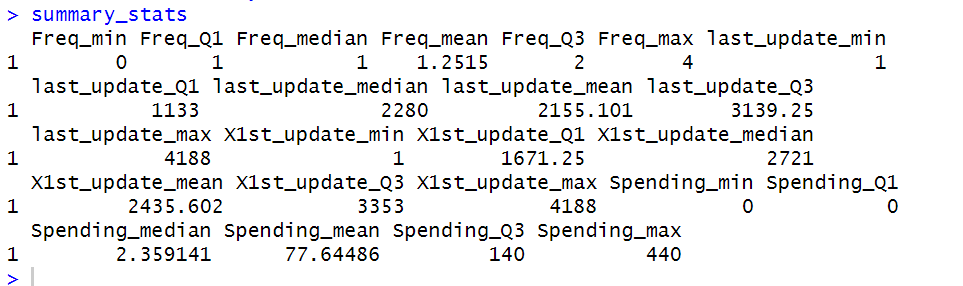
Conduct descriptive analytics for the four numerical variables.

1. **The 6-summary statistics:**

Refer to data cleaning part for visuals.



Again, the log-transformation indeed compresses values, which might not provide the full picture of the data in its natural scale, particularly for descriptive analytics. For this reason, it's more meaningful to reverse the log transformation and compute the descriptive statistics on the original scale of the data, hence, we will reverse log-scale and compute summary statistics.



For Freq and Spending, we applied exp(x) - 1 to bring them back to their original scale. Variables like last\_update\_days\_ago and X1st\_update\_days\_ago were never transformed, so they remain as is.

The six-summary statistics reveal key insights about customer behavior and the dataset. For Freq (frequency of purchases), most customers made 1 or 2 purchases, with a maximum of 4, indicating limited repeat purchases. For last\_update\_days\_ago and X1st\_update\_days\_ago, the data reflects a wide range of updates spanning up to ~11.5 years, though the median suggests that half of the records were updated more than 6–7 years ago, potentially impacting the relevance of the data. Spending is highly skewed, with 50% of customers spending $2.36 or less and 25% spending nothing, while the average of $77.64 highlights a small group of high-value customers driving revenue. These findings suggest that customer activity is low for most, but a minority of high-spending customers represent a significant opportunity for targeted marketing strategies. The age of the data also emphasizes the need to focus on recent updates for more actionable insights.

It is important to note that the non-purchasers with zero spending likely skewed the results and affected the interpretation of the descriptive statistics negatively. The mean spending is not a true reflection of the actual spending behavior of those who made as it disproportionately lowered because it averages the spending across all customers, including those who spent nothing (non-purchasers). This is misleading conclusions about the average customer as it also affect the median and quartiles.

**Calculate the descriptive statistics for purchasers only:**

To get a clearer understanding of actual spending behavior to provide meaningful insights.

A close-up of a white background

Description automatically generated

The descriptive statistics for purchasers' spending reveal a wide range of behaviors. The minimum spending was $3, while the maximum reached $440, indicating substantial variability in purchase amounts. The median spending was $140, meaning half of the purchasers spent this amount or less, while the average spending was slightly higher at $155.29, suggesting the presence of high-value customers driving the mean upward. The 1st quartile (Q1) was $94.75, showing that 25% of purchasers spent less than this, while the 3rd quartile (Q3) was $197, highlighting that 25% of purchasers spent between $197 and $440, representing a high-value segment. These statistics indicate a skewed distribution of spending, where a small group of high-spending customers contributes disproportionately to revenue. This insight emphasizes the importance of identifying and targeting these high-value customers for retention and upselling strategies.

Refer to data cleaning part for histograms.

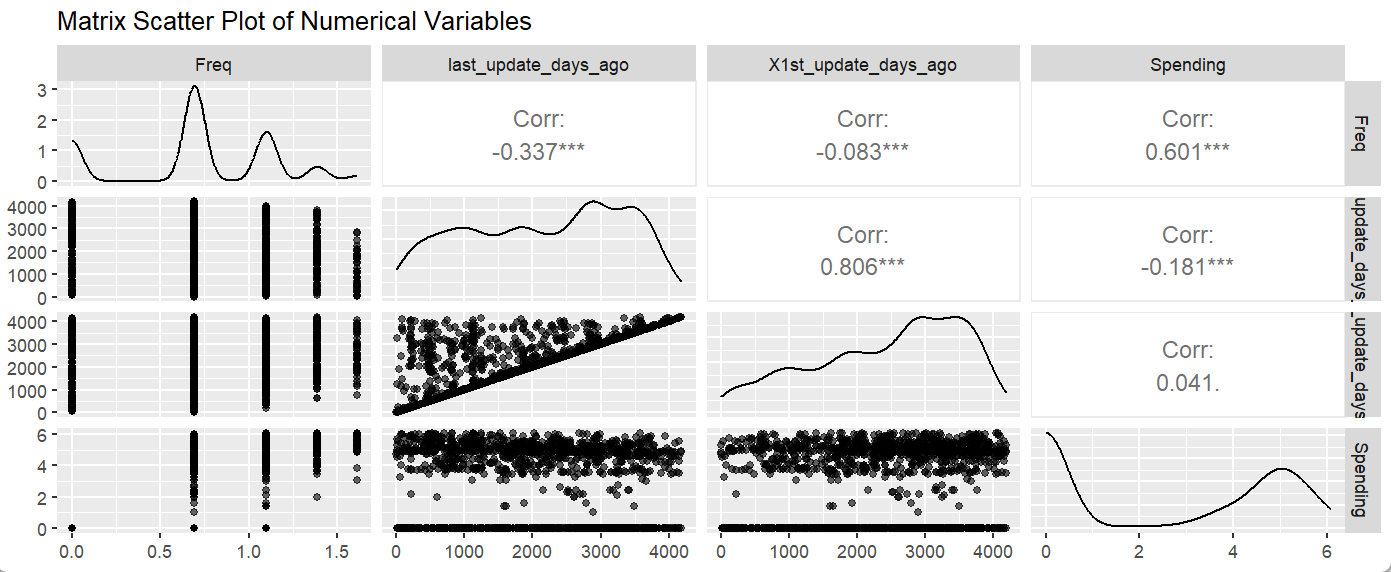
1. **Outlier** 
   1. Which one has the most outliers

Referring to the data cleaning part, Freq has the most outliers with 114 outliers. This is likely because frequency is a count variable, and a small group of highly active customers (e.g., frequent repeat purchasers) stands out as extreme compared to most customers who made fewer purchases. On the other hand, both last\_update\_days\_ago and X1st\_update\_days\_ago had no outliers. These variables were relatively uniformly distributed and did not exhibit extreme values. Finally, Spending also had a considerable number of outliers (110), driven by a small group of high-spending customers whose purchases were significantly larger than the rest.

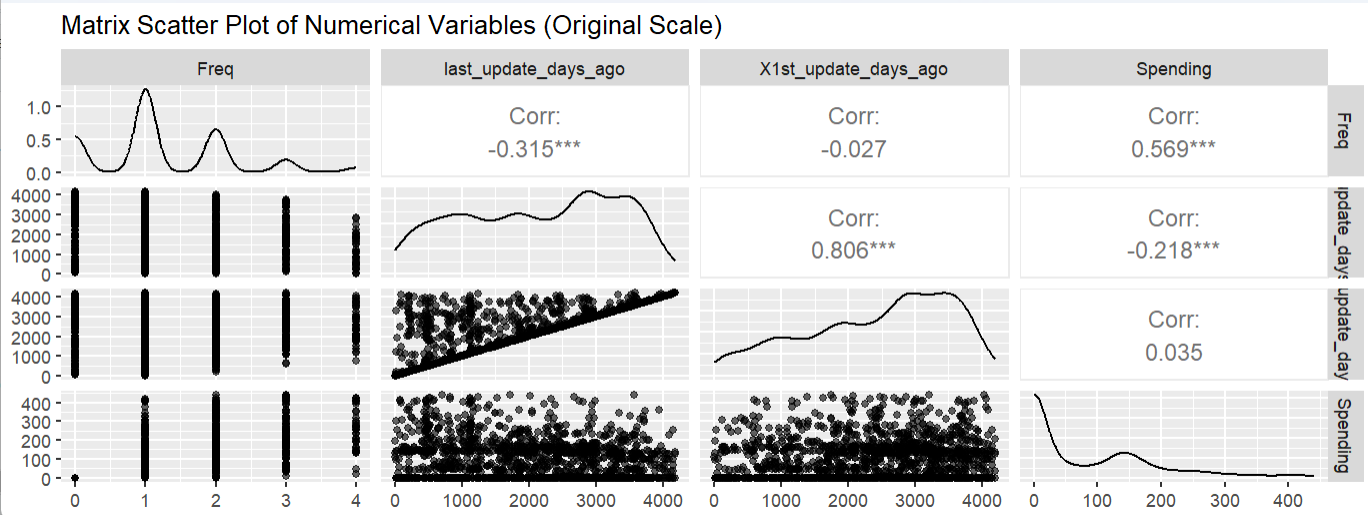
* 1. How to fix variables with the outliers

We have created a new cleaned dataset by addressing key data quality issues, including outliers and skewness. Specifically, we detected and resolved outliers in the variables Freq and Spending using imputation (replacing outliers with the median) in the data cleaning process earlier, ensuring these variables reflect realistic values without being distorted by extreme cases. Additionally, we applied a log transformation to these variables to compress their range and reduce skewness, improving the overall distribution for analysis. For last\_update\_days\_ago and X1st\_update\_days\_ago, we confirmed that no significant outliers existed, so these variables were retained in their original form to preserve their natural interpretability. The resulting cleaned dataset is now well-prepared for further analysis and modeling.

1. **Visualization using “ggplot” package for matrix scatter plot**

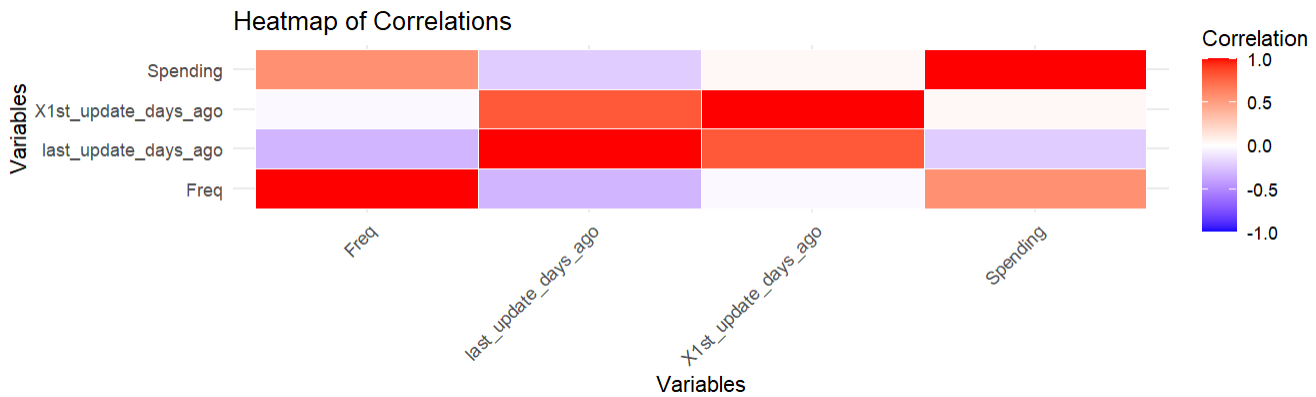


It is important to reverse the log scale before creating the scatter plot matrix. This ensures that the relationships between variables are represented on their original scales, which provides a more intuitive and meaningful visualization.



The scatter plot matrix provides valuable insights into the relationships between key numerical variables. The diagonal density plots highlight that Freq and Spending are skewed, with most customers making a low number of purchases and spending relatively small amounts, while a few high-frequency, high-spending customers stand out. The moderate positive correlation (0.569) between Freq and Spending indicates that customers who purchase more frequently tend to spend more, making Freq a good predictor of Spending. A weak negative correlation (-0.315) between Freq and last\_update\_days\_ago suggests that frequent purchasers are slightly more likely to have recent updates in their records. Similarly, the strong positive correlation (0.806) between last\_update\_days\_ago and X1st\_update\_days\_ago reflects that customers added earlier to the database tend to have older last updates. Overall, the analysis highlights the importance of targeting high-frequency and high-spending customers for marketing efforts and underscores the need to focus on more recently updated customer records for actionable insights.

1. Create “heatmap” for all variables and indicates high correlation, if any.



Variables with strong correlations will be highlighted in red (positive) or blue (negative).

A strong positive correlation (>0.7) indicates a linear relationship in the same direction. While a strong negative correlation (<-0.7) indicates a linear relationship in the opposite direction.

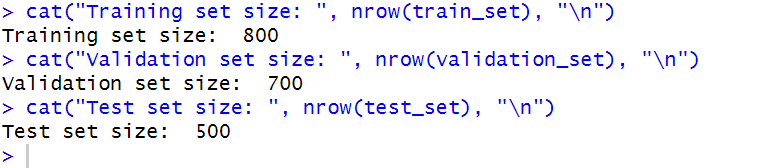
The strongest positive correlation exists between Freq and Spending, highlighting that frequent purchasers are also high spenders. The strong relationship between last\_update\_days\_ago and X1st\_update\_days\_ago reflects the historical nature of the dataset, this indicates that these variables are closely related and might be capturing similar aspects of the dataset and convey overlapping information.

The negative correlations involving last\_update\_days\_ago suggest that recency of updates has a modest association with both frequency and spending, which may indicate that more recent customers or updated records tend to be more active.

## Question 3:

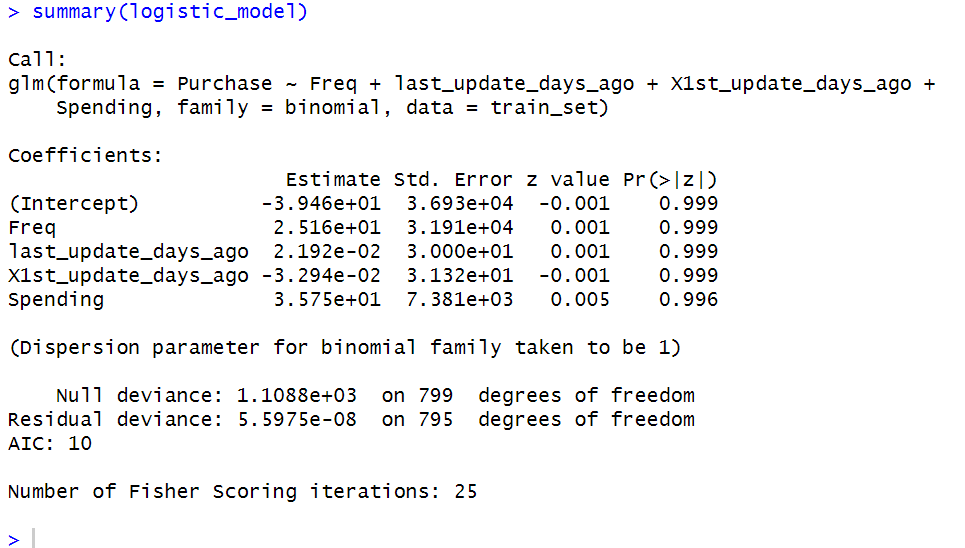
Develop a model for classifying a customer as a purchaser/nonpurchaser.

1. **Partition the data randomly into training set (40%), validation set (35%), and test set (25%).**

****

We used sample\_frac(1) to shuffle the dataset to ensure random partitioning. Sample Sizes are randomly split into

1. Training Set: 40% of the data (used to train the model).
2. Validation Set: 35% of the data (used to avoid overfitting and validate performance).
3. Test Set: 25% of the data (used to evaluate final model performance).
4. **Use the training set to build and run the logistic regression model**

****

The model aims to predict whether a customer is a purchaser (Purchase = 1) or a non- X1st\_update\_days\_ago, and Spending. Each coefficient represents the log odds of being a purchaser for a one-unit increase in the corresponding variable, holding all other variables constant.

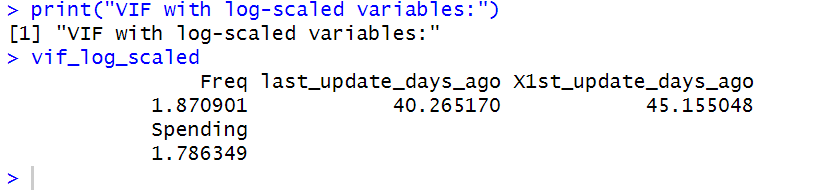
|  |  |
| --- | --- |
| **Coefficients** | **Interpretation** |
| Intercept (-39.46) | This is the baseline log odds of being a purchaser when all predictors are zero. It's not meaningful on its own but forms part of the logistic equation. |
| Freq (25,160) | A very high coefficient for Freq suggests that frequency has an extremely large influence on the log odds. However, the high p-value (0.999) indicates this result is not statistically significant. |
| last\_update\_days\_ago (0.0219) | A small positive coefficient suggests that the number of days since the last update has a minimal impact on the log odds. The p-value (0.999) indicates it is not statistically significant. |
| X1st\_update\_days\_ago (-32.94) | A negative coefficient suggests that earlier first updates might slightly decrease the likelihood of being a purchaser, but the p-value (0.999) shows it is not statistically significant. |
| Spending (3.575e+01) | A high positive coefficient suggests spending might positively influence the likelihood of being a purchaser. However, the p-value (0.996) indicates no statistical significance. |

All predictors have p-values > 0.05, indicating that none of them are statistically significant in predicting the target variable (Purchase) in this model, this indicates that the model needs refinement.

The use of log-scaled variables could indeed be a factor affecting the logistic regression model's performance and interpretability as it compresses the range of values, particularly for variables like Freq and Spending. While this helps reduce skewness, it may obscure meaningful differences between purchasers and non-purchasers. Feature selection might also helps improve model performance, interpretability, and prevents overfitting, Strong correlations (between last\_update\_days\_ago and X1st\_update\_days\_ago) may result in inflated standard errors and reduce the reliability of the coefficients.

**Feature Selection:**

Now before we proceed, one assumption of logistic regression is the absence of multicollinearity. Therefore, we will check for multicollinearity using Variance Inflation Factor (VIF) and drop variables with high VIF values.



A screenshot of a computer code

Description automatically generated

1. **Low Multicollinearity for Freq and Spending**:

* Both variables have low VIF values in both log-scaled and original-scale models, indicating that they are not highly correlated with other predictors.

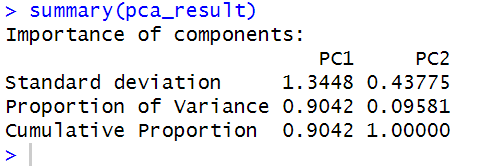
1. **High Multicollinearity** Between last\_update\_days\_ago and X1st\_update\_days\_ago:

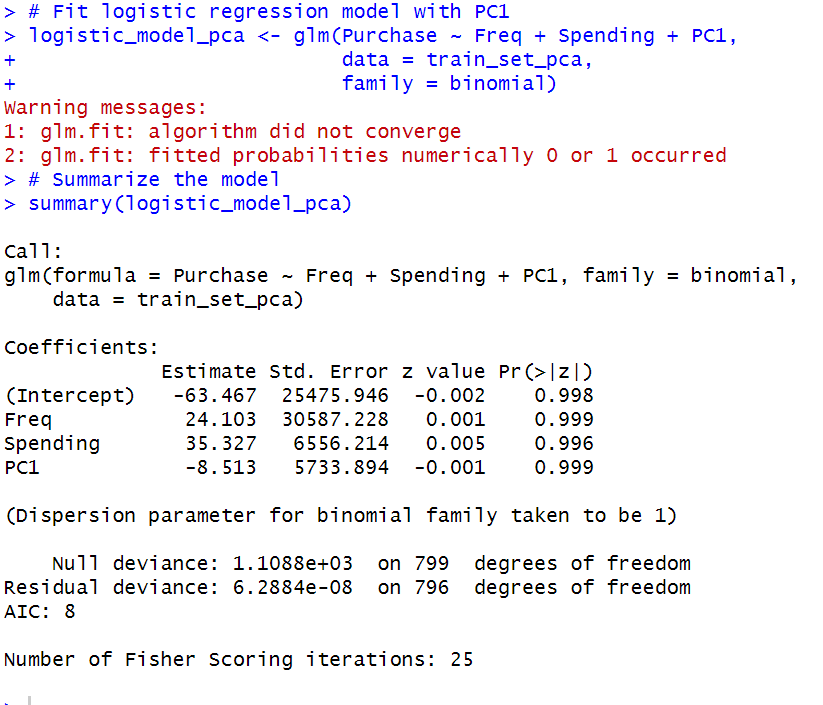
* These two variables exhibit severe multicollinearity in both cases, with VIF values exceeding 40 in the log-scaled model and nearly 300 in the original-scale model. This is consistent with earlier observations of a strong correlation (0.806) between these two variables.

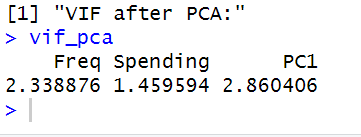
1. **Impact of Log Scaling:**

* Log scaling reduces the magnitude of VIF values for last\_update\_days\_ago and X1st\_update\_days\_ago, but the values remain significantly high. This indicates that log scaling partially mitigates multicollinearity but does not eliminate it.

As seen from the VIF results, last\_update\_days\_ago and X1st\_update\_days\_ago are highly correlated (VIF > 40 for log-scale, VIF > 280 for original scale). PCA can combine these two variables into a single component that explains most of their shared variance without arbitrarily dropping one of the variables and risk losing valuable information.





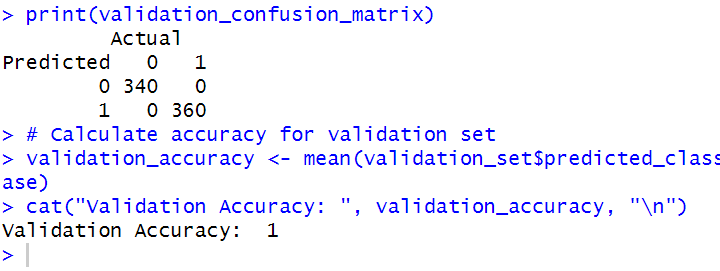
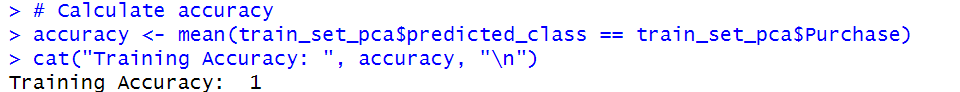
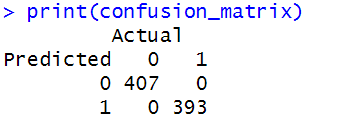
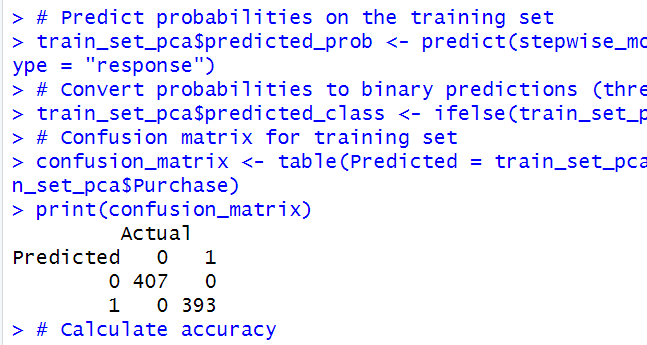


Freq (2.34), Spending (1.46), and PC1 (2.86) all have low VIF values (< 5), indicating that multicollinearity has been resolved. Despite resolving multicollinearity, none of the predictors (Freq, Spending, or PC1) are statistically significant. We will now Validate the model on the validation set to confirm its performance and avoid relying solely on the training set metrics.

1. **Run stepwise logistic regression using backward elimination to select the best variables, then use this model to do the classification again**

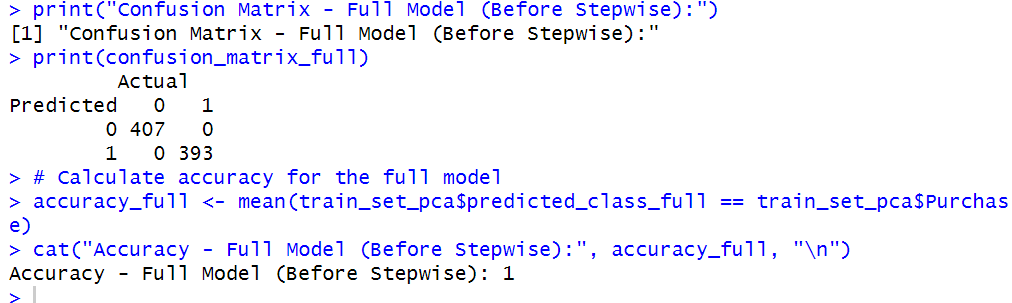
**A screenshot of a computer

Description automatically generated**

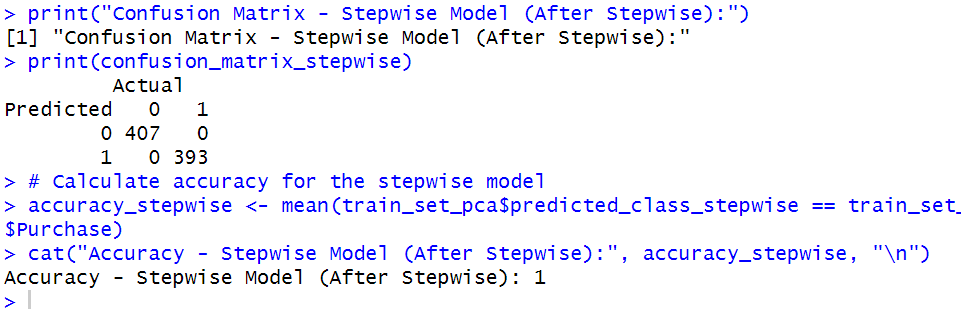
****

After backward elimination, the model retained only Spending as the significant predictor of purchase behavior, other variables (Freq and PC1) added no significant value and were excluded. The model achieves perfect classification (100%) on both the training and validation sets which might indicate an over-fitting model, therefore, further validation on a test set is crucial to confirm the model's generalizability and ensure the perfect accuracy is not due to overfitting.

1. **Run confusion matrix before running the stepwise and after**
   1. **Confusion Matrix for the Full Model (Before Stepwise)**

****

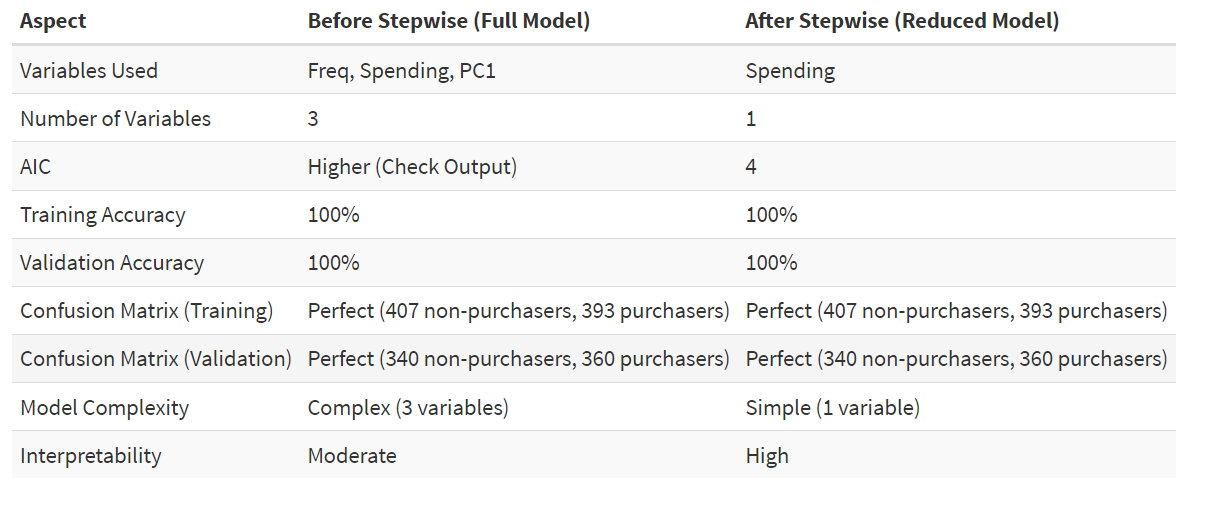
* 1. Confusion Matrix for the Stepwise Model (After Stepwise)

****

Although the full model used three predictors (Freq, Spending, and PC1), while the stepwise model retained only Spending as the sole predictor, both the full model and the stepwise model achieve perfect accuracy on the training set, indicating that both models can fully separate purchasers and non-purchasers. Nonetheless, we can conclude that Spending is the only variable needed to perfectly classify customers in this dataset, making the other predictors (Freq and PC1) redundant.

1. **Create a table that shows the difference of running the model before and after the variable selection**

We used data.frame() function to create the table.

****

The model used three variables: Freq, Spending, and PC1 (a principal component derived from last\_update\_days\_ago and X1st\_update\_days\_ago) before stepwise, and retained only Spending as the sole predictor, simplifying the model significantly after stepwise.

The comparison between the full model and the stepwise model highlights that stepwise regression effectively simplified the model while maintaining perfect classification accuracy. Despite the reduction in complexity, the stepwise model achieved the same 100% accuracy on both the training and validation sets. However, further validation on unseen test data is needed to confirm generalizability.

**Validate the stepwise model on the test set:**

A computer code with text

Description automatically generated

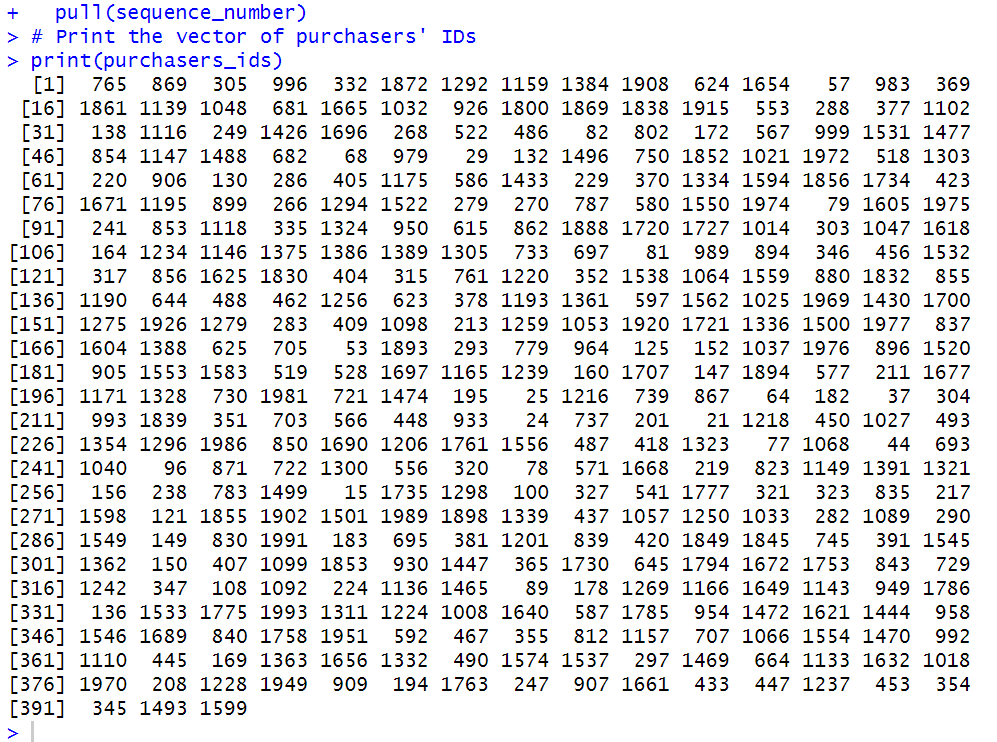
The likely reason for perfect accuracy is spending among non-purchasers is always zero, making it a perfect discriminator in the data. The logistic regression model relies heavily on Spending, which was identified during stepwise regression as the only significant predictor.

If spending data is not available beforehand, we would need a different model that predicts whether a person is likely to be a purchaser or non-purchaser based on other features (e.g., frequency of visits, last update days, demographic data). However, if spending data is consistently available, the current model is ideal for classification.

## Question 4:

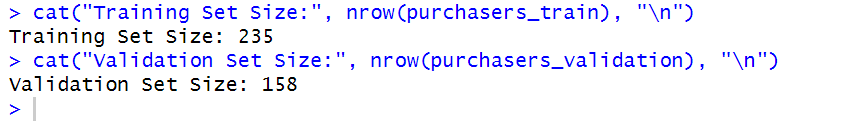
Develop a model for predicating “spending” among the purchasers:

* 1. Create a vector of ID’s for only purchasers’ records (Purchase =1)

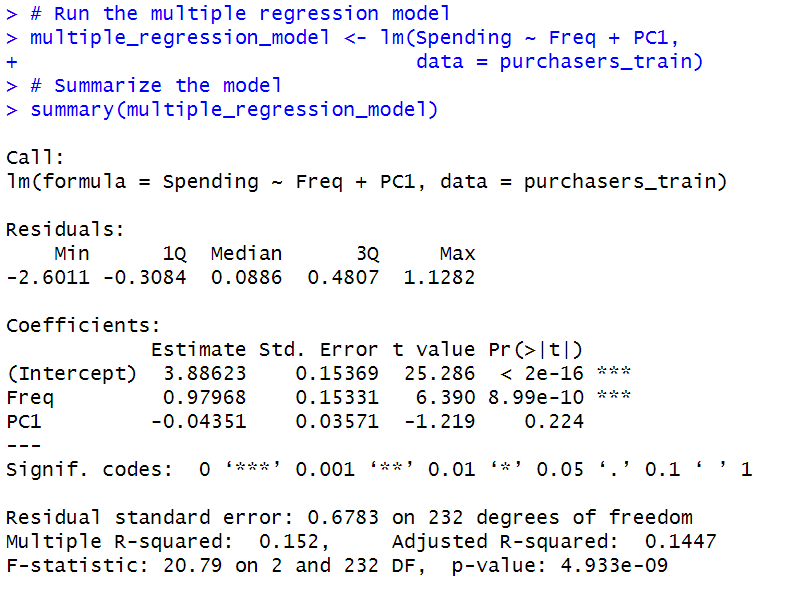


The purchasers\_ids vector contains the unique sequence\_number identifiers for all customers classified as purchasers (Purchase = 1) in the dataset. Each number in the vector represents a distinct customer who made a purchase and there are 391 unique purchasers in the dataset (as shown by the number of entries in the vector).

* 1. **Partition the data into training set (60%) and validation set (40%)**

****

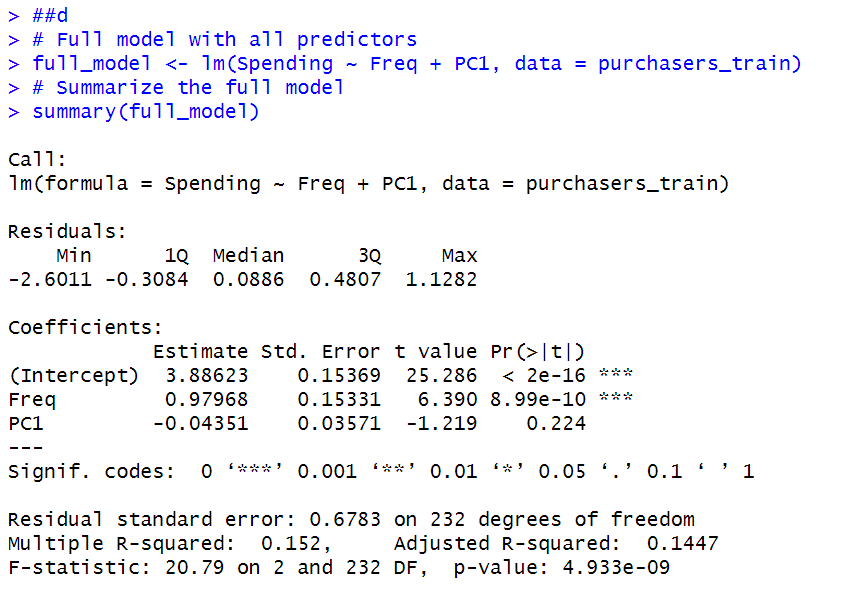
* 1. **Use the training set to run the multiple regression model**

****

The multiple regression model predicts spending using Freq (frequency of transactions) and PC1 (the first principal component, update days). The results show that Freq is a significant predictor (p < 0.001), with spending increasing by approximately 0.98 units for every unit increase in transaction frequency, indicating that customers who transact more frequently tend to spend more. However, PC1 is not statistically significant (p = 0.224), suggesting that the underlying variables captured by this principal component do not significantly influence spending. This suggests that Freq is an important driver of spending.

* 1. **Run stepwise regression and compare the model before and after**

A screenshot of a computer

Description automatically generated  
A screenshot of a computer

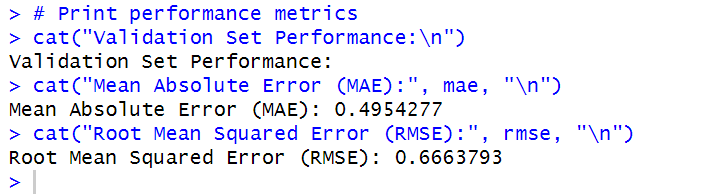
Description automatically generated

A computer code with numbers and letters

Description automatically generated

The stepwise regression process simplified the model for predicting spending by removing the insignificant predictor, PC1, and retaining only Freq. The full model, which included both Freq and PC1, had an R-squared of 0.152 and an AIC of 489.47. The stepwise model reduced complexity by excluding PC1, achieving a slightly lower AIC of 488.97 while maintaining a similar R-squared of 0.1466. This demonstrates that removing PC1 had a negligible impact on the model's explanatory power, as its contribution to predicting spending was statistically insignificant (p = 0.224). The stepwise model is preferable due to its simplicity, marginally improved efficiency (as reflected in the AIC), and similar performance to the full model. This streamlined model is more interpretable and still effectively explains the relationship between transaction frequency (Freq) and spending.

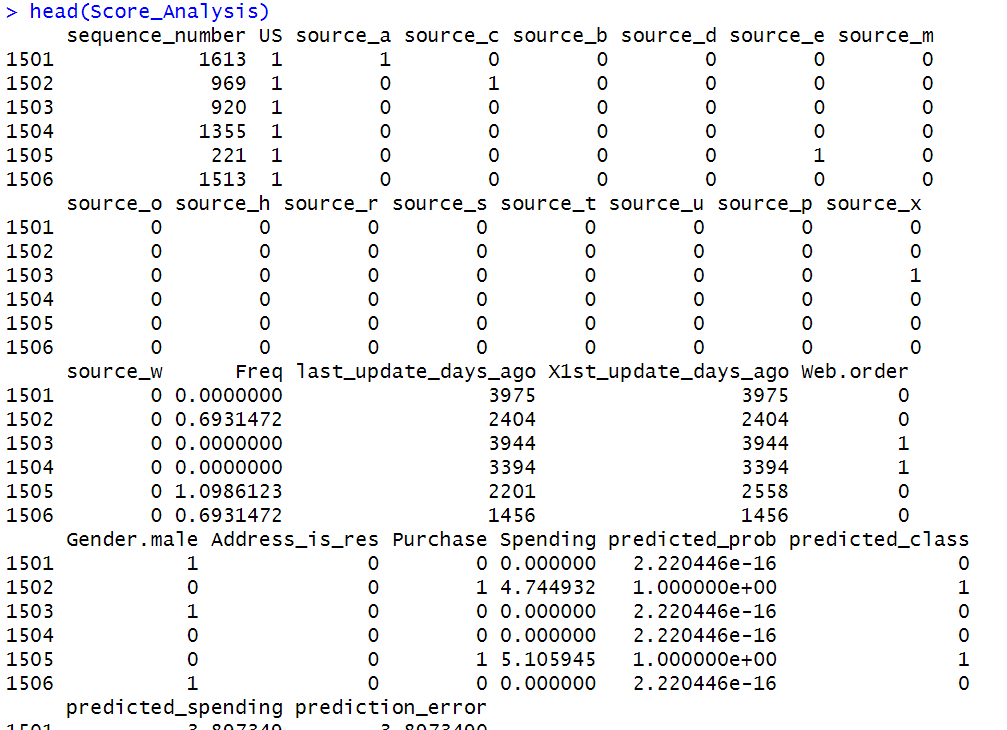
1. **Use the validation set to predict the model and run “accuracy” function to show the model performance**

****

The stepwise regression model's performance on the validation set, with a Mean Absolute Error (MAE) of 0.4954 and a Root Mean Squared Error (RMSE) of 0.6664, indicates that the model has moderate prediction errors and generalizes well. When compared to the training model, the validation metrics align closely, suggesting consistent performance across datasets. The small difference between MAE and RMSE also implies that prediction errors are relatively uniform without significant outliers. Since the validation results are comparable to the training results and there is no significant drop in performance, this indicates that the model is not overfitting the data. The simplicity of the model, which relies on a single predictor (Freq), also supports its robustness and generalizability.

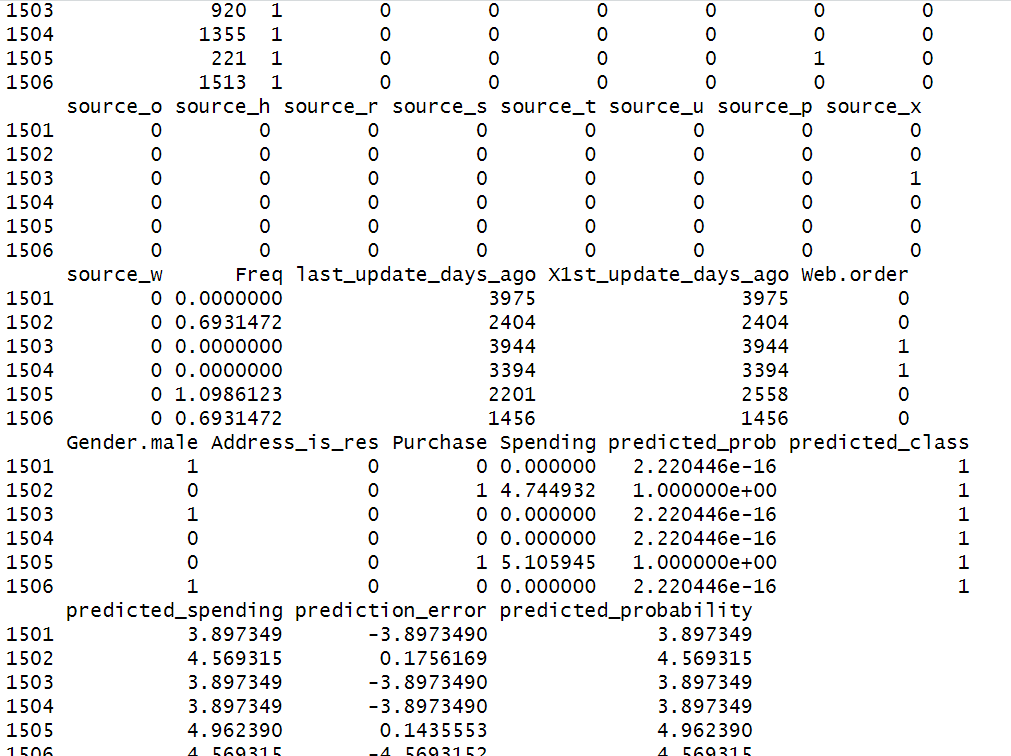
## Question 5:

**Create a new data frame “Score Analysis” that contains the test data portion of this dataset (test data created in 3.a):**

****

* 1. **Add a column to the data frame with predicted scores from the logistic regression**

The new columns (predicted\_probability and predicted\_class) are added to the Score\_Analysis data frame.

****

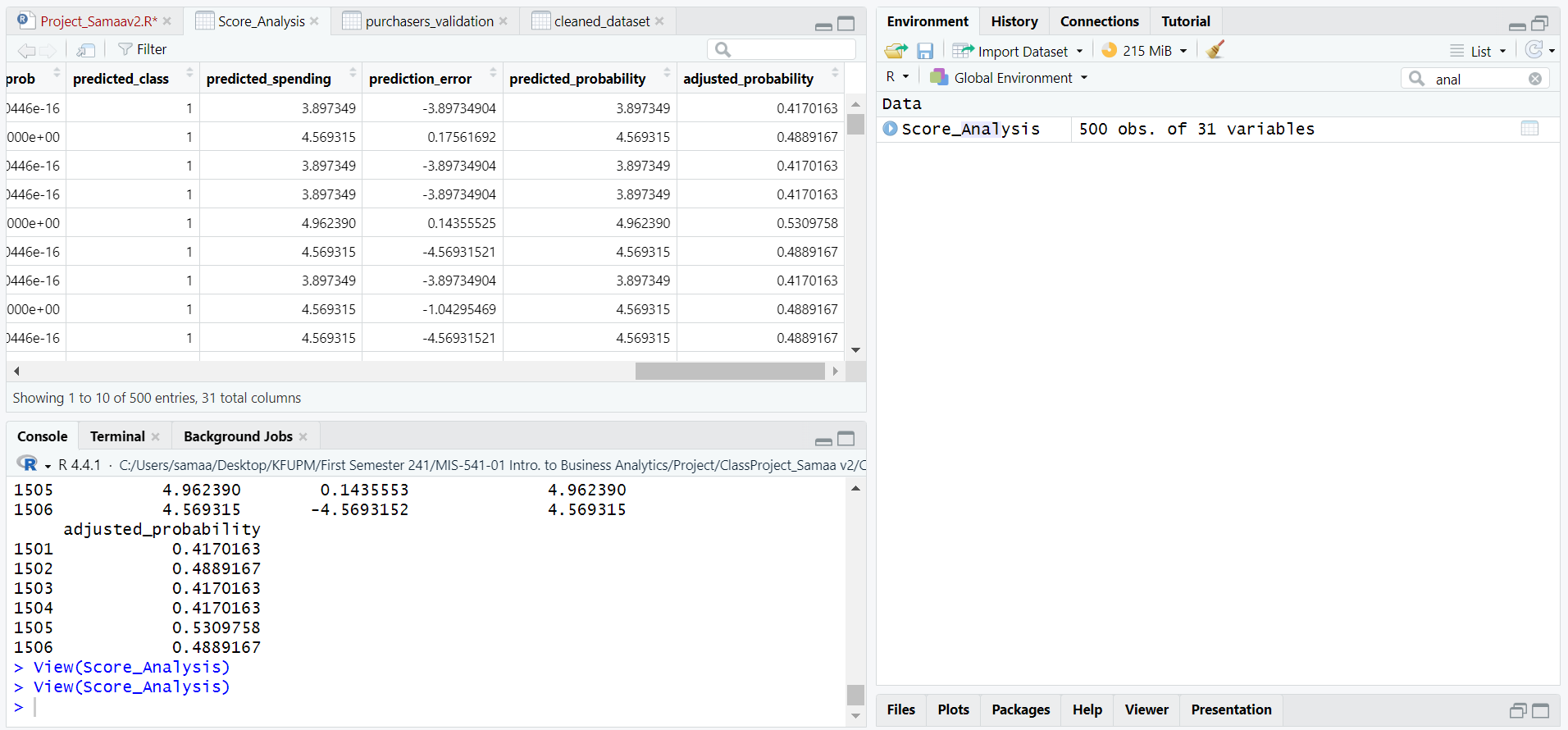
* 1. **Add another column with the predicted spending amount from the predication model chosen**

A new column, predicted\_spending, is added to the Score\_Analysis data frame.

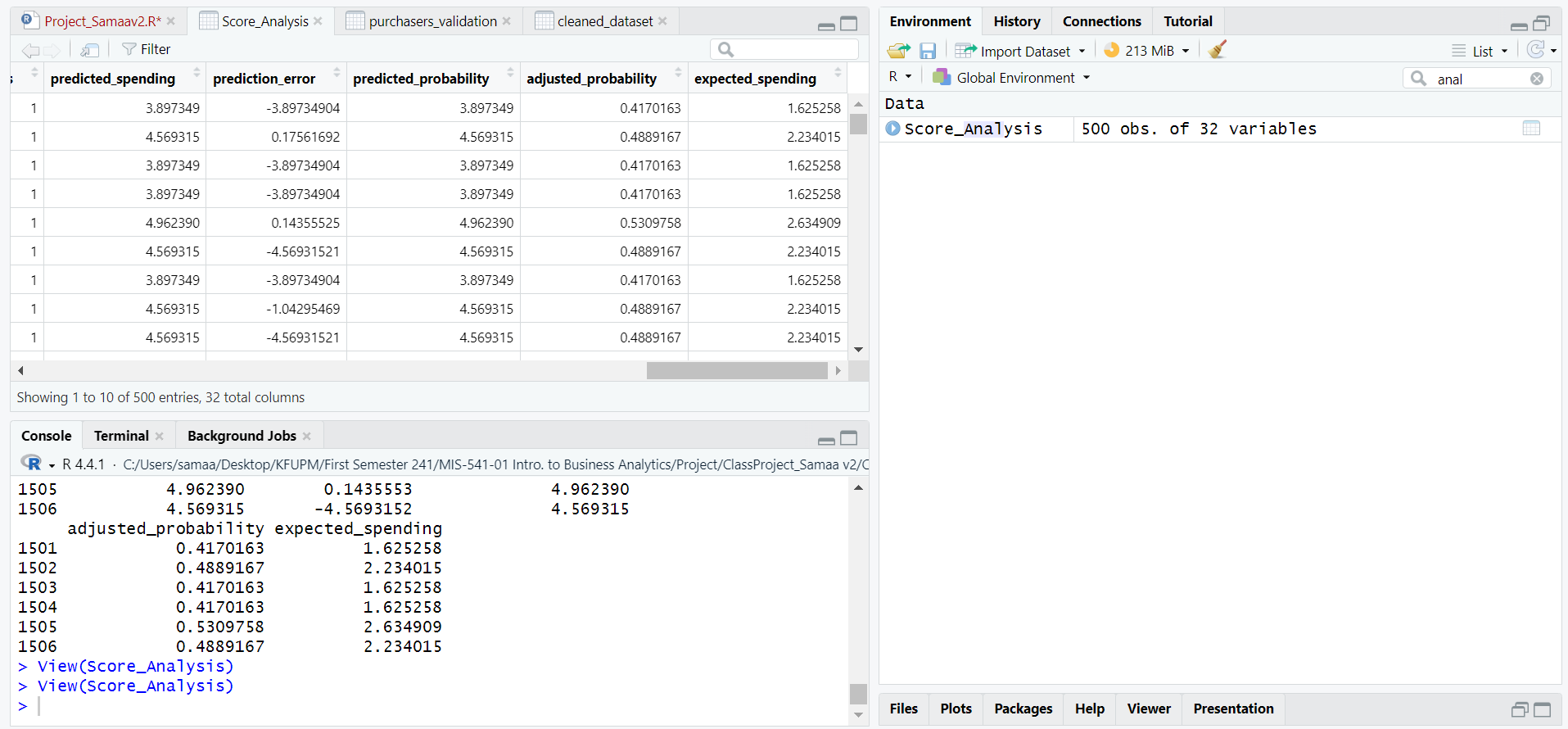
A screenshot of a computer

Description automatically generated

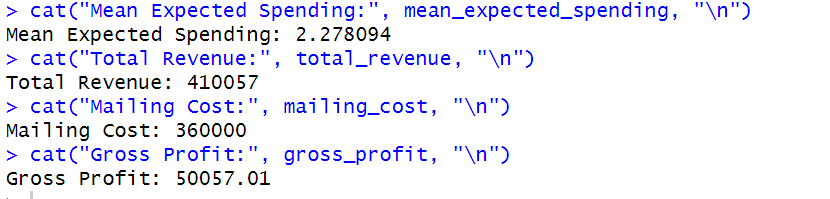
* 1. **Add another column with “adjusted probability of purchase” by multiplying “predicted probability of purchase” by 0.107 (see the case for further illustration)**



* 1. **Add another column for expected spending: adjusted probability of purchase\*predicted spending.**

****

* 1. **Estimate the gross profit that would result from mailing to the 180,000 names on the basis of your data mining models (hint: use the mean of expected spending).**

****

The mean expected spending was calculated as $2.28, yielding a total revenue of After accounting for mailing costs The estimated gross profit .

This refined method improves initial estimates by using a tailored model that adjusts purchase probabilities and incorporates regression-predicted spending, providing a more precise profit projection.