**Marketing Strategy for Greyson Corporation**

**MSBA 635: Data Analytics 2**

**Hudson Gifford, Nicholas Michal, Tyler Nunez, Dylan Scharff, and Ben Stoll**

**February 7, 2021**

**Executive Summary**

Greyson Corporation wishes to have an updated model for predicting customers who will respond to magazine renewal offers. Our team analyzed their historical customer data and created a logistic regression model for predicting customer response. Specifically, a LASSO logistic regression model was created due to model interpretability and the exclusion of unimportant factors. Variables such as income, amount paid, dollars per issue, and time between orders play a pivotal role in predicting whether customers will renew. Our recommendation is for Greyson Corporation to place their focus on these important factors and understand their impact on the customer response.

**Introduction**

Data analytics and modeling play a significant role in understanding markets and predicting customer behavior. Keeping track of detailed customer data can help a company draw insights and make better business decisions. Greyson Corporation *—* a media and marketing company that publishes magazines *—* uses customer data for advertising and marketing purposes. A current problem facing Greyson is how to identify magazine subscribers who will respond to mailed renewal offers. Therefore, our client is seeking a marketing strategy to boost customer response to magazine renewal offers.

To predict people who would renew their magazine subscription, we created a LASSO logistic regression model, where the response is renewal or no renewal. In this report, we explore historical Greyson customer data and evaluate different classification models. In addition, we provide our client with a few recommendations for how to boost customer response, such as using our logistic regression model and clustering the data into market segments. Our client should consider and understand how the important variables drawn from the LASSO logistic regression model affects their customer response.

**Exploratory Data Analysis**

Exploratory data analysis is a crucial part of analytics and is the first step in properly understanding the data. Upon reviewing the data, it was found that there were some missing observations as shown in the table below. Imputing the data mitigated this issue.

|  |  |  |
| --- | --- | --- |
| Variables | Total Observation | Expected Observations |
| Home Owner | 35431 | 36000 |
| Household Size | 35992 | 36000 |
| Home Value | 35995 | 36000 |
| Total Paid Orders | 35974 | 36000 |
| Months Since expire | 35955 | 36000 |
| Months Since Last Payment | 35681 | 36000 |
| Months Since 1st Order | 35866 | 36000 |
| Dollars Per Issue | 34666 | 36000 |
| Total Paid Orders | 35974 | 36000 |

*Figure 1: A list of predictor variables containing missing values in the original data set*

There are several variables that are greatly skewed because of their binary nature. The most notable of them being Paid Complaints and Number Gift Donation. However, several of the classification variables are also quite heavily skewed.

* Gender (Female 83% of observations)
* Dwelling type (Single Family 87% of observations)
* Marital status (Married 72% of observations)
* Child present (True for 61% of observations)
* Last payment type (Cash orders for 68% of observations)

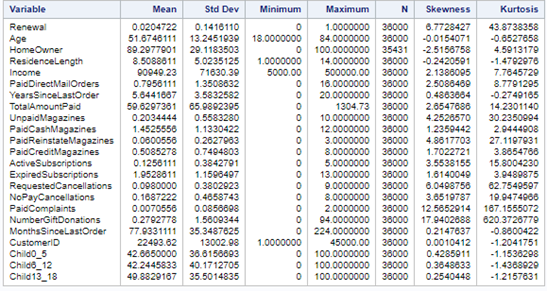


Figure 2: A basic summary of variables included in the data

Much of the data is in a binary format; however, there are still a few important variables for which we created histograms along with their density curve and box plots. Histograms help evaluate the normality of the variable as well as show us how impactful the outliers are. The income distribution seems skewed upon visual inspection and includes outliers. The considerable number of outliers is clearly visible in the box plot hence the skewness of 2.13 (no classification grouping).

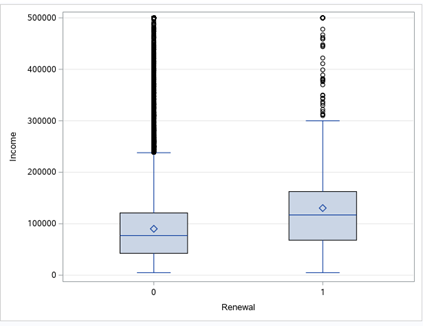
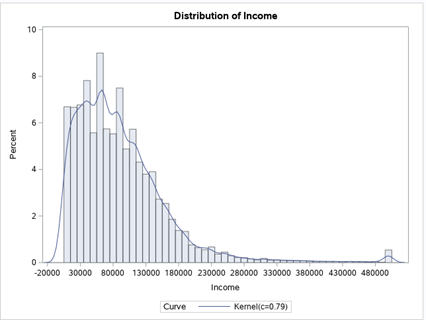


Figure 3: A histogram showing the distribution of the Figure 4: A box plot displaying the "Income" variable

"Income" variable against the response variable "Renewal"

Months Since Last Order is an important variable when it comes to renewals. The density curve is normal, and the overall skewness is only 0.21 (no classification grouping). Upon reviewing the box plot there are a few outlies on the No Renewal side, however they do not impose a significant impact on the data as you can see by the skewness.

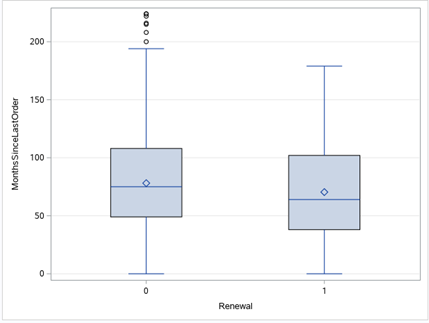
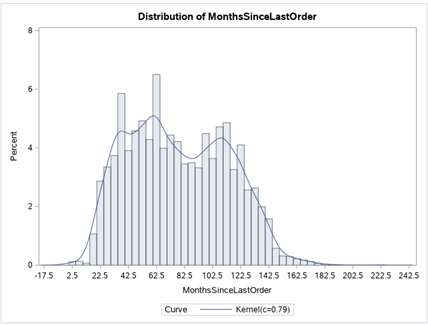
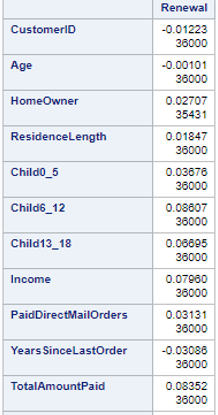


Figure 5: A histogram showing the distribution of the Figure 6: A box plot testing "MonthsSinceLastOrder against "MonthsSinceLastOrder" variable “Renewal”

There were correlations in the data, but it was clear from the correlation test that none of the variables had a strong positive or negative correlation. This is indicated by the values’ nearness to 0 in Figure 7, which displays the results of the correlation test run in SAS.

However, this does not necessarily mean our data is useless, as it is possible for weak correlations to translate to strong predictors depending on the model run on the data. Also, while all of the predictor variables show weak correlations, some are still meaningfully stronger than others.

For example, “TotalAmountPaid,” “Child6\_12,” and “Income,” all have correlation values near or above 0.08. While this value indicates a weak correlation, these correlations are much stronger than that of “NoPayCancellation,” which returned a correlation value below 0.004, which indicates an infinitesimal correlation.



*Figure 7: A list of correlation coefficients for each observation variable compared to “Renewal”*

**Variable Transformation**

After using an exploratory analysis to understand the data, we preprocessed our data to ensure model accuracy. This primarily consisted of removing unnecessary variables and imputing the remaining data to correct for null values.

The CustomerID variable was removed from the data because it is a number sequentially assigned to each customer and has no statistical value. While variable selection in later steps will likely account for CustomerID, it was clear that the variable had no usefulness to this analysis. Next, the remaining data was imputed, meaning null values within the data were converted to the mean of their respective variables. Many models are incapable of running when null or “NA” values are present, and in the case of this data, there were 2,338 observations that were corrected.

Rather than completely omitting those observations – which make up 6.5% of the data as a whole – the null values within those observations were transformed to the means of their respective variables. This mitigated their impact on the data while also allowing the chosen model to utilize the observations that would have otherwise been lost had they just been deleted from the data frame.

Once these processes were complete, it came time to determine which model would be best for identifying customers who would renew, thereby improving the response rate.

**Model Selection**

Ultimately, it was decided that a LASSO regression model would best serve the group in determining how to best improve the renewal rate of magazine subscriptions for Greyson. This is because it hit the best balance between interpretability, ease of implementation, and the time-cost associated with its use.

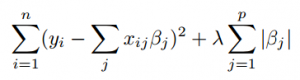
Other models were considered, but it became clear that some sort of logistic regression would be the best course of action. Due to the classification nature of the business problem, any linear regression model would be useless, as those are used to solve regression problems. Gradient boost and random forest models seemed intriguing at first due to their higher prediction accuracy, but ultimately took hours to run, making troubleshooting and fine-tuning the model difficult.

Another intriguing option considered was to run a principal components analysis and then run the principal components through another classification model. This, however, greatly sacrificed interpretability while not meaningfully improving prediction accuracy. In the end, this method was discarded in favor of a more understandable model.

The group then decided that a logistic regression model would serve best, and ultimately landed on LASSO regression due to its ease of interpretability and its accounting for all variables at its start. LASSO regression also excludes unimportant variables from the model, which will allow our client to focus on factors that affect magazine renewal.

**Model Implementation**

A LASSO regression is a shrinkage method that aims to significantly reduce variance (at the cost of a slight increase in bias) by eliminating predictor variables using the sum of squared error via the following equation for each variable:



While the above equation may look intimidating at first glance, what it essentially conveys is much simpler: the sum of squared error is added to the product of a chosen lambda (representing a penalty coefficient) value and the absolute value of the slope. This process effectively allows the model to reduce unimpactful predictors to 0 using lambda, leading to much higher interpretability than other shrinkage methods such as ridge regression, while also limiting the risk of overfitting that is associated with stepwise selection regressions.

The process is begun by splitting the given data, of which there are 36,000 observations into two sets: a training set containing 80 percent of the data (28,800 observations) and a testing set containing 20 percent of the data. This 80-20 split is generally considered a standard for such models. Splitting the data into training and testing sets ensures that we do not overfit the model. After the data was imputed, a dummy matrix is created to convert any categorical variables (in this case “DwellingType,” “Gender,” “Marital,” “ChildPresent,” “Occupation,” “MagazineStatus,” “LastPaymentType,” and “GiftDonor”) into numeric dummy variables in the form of a matrix. The data is then prepared for LASSO by creating matrices for the training and testing data on the X axis, and data frames for the training and testing data on the Y axis.

Then, a cross-validation is performed on the training data to determine the shrinkage parameter, which are visually represented by the vertical dotted lines seen in **Figure 8**.

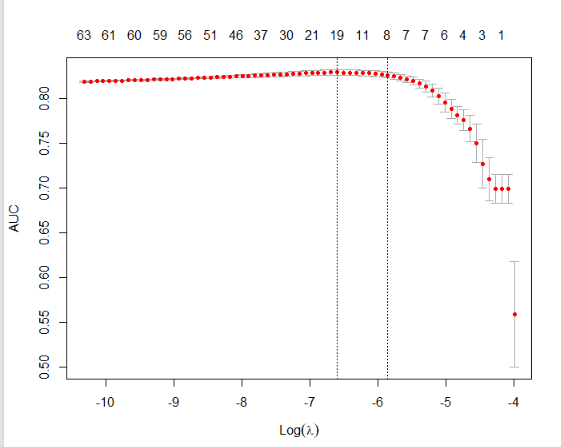


Figure 8: A plot visualizing the cross validation done by the LASSO model

In the graph, each predictor variable is represented as a red dot, and those that fall within the shrinkage parameter are those that will be used by the model to make predictions. Those being used can be seen in **Figure 9.**

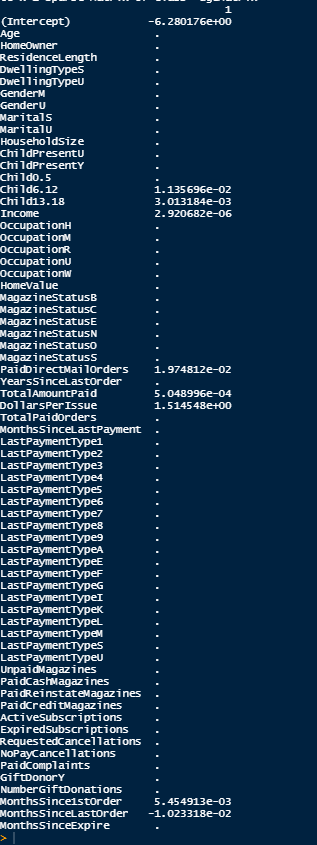


Figure 9: The resulting coefficients for each variable tested by the LASSO model. Those variables without coefficients were removed by the model.

As evident by these results, there are only eight variables with coefficients remaining in the model, including “Child6.12,” “Child13.18,” “Income,” “PaidDirectMailOrders,” “TotalAmountPaid,” “DollarsPerIssue,” “MonthsSince1stOrder,” and “MonthsSinceLastOrder.” The rest have been removed from the model as a result of a lack of usefulness to the model. This is one of the key advantages to using LASSO logistic regression: it gives those running the model a clear view of which predictor variables have a significant impact in predicting the response variable.

**Model Performance**

Now with these results in hand, it came time to find out just how good this model was at predicting renewing customers. In other words, we evaluated our model on the testing set.

To do this, the 7,200 observations in the testing data were run through this updated model, and a receiver operating characteristic (ROC) curve was created to break down how successful the model was. The ROC curve graphs the True Positive Rate (as the y-axis) against the False Positive Rate (as the x-axis) to give a representation of the performance of a classification model. In short, the sharper the curve, the more successful the model. This can be more specifically measured and represented numerically by calculating the area underneath the ROC curve. The higher that number, the better the performance.

The ROC curve produced by the LASSO regression model can be seen in **Figure 10.**

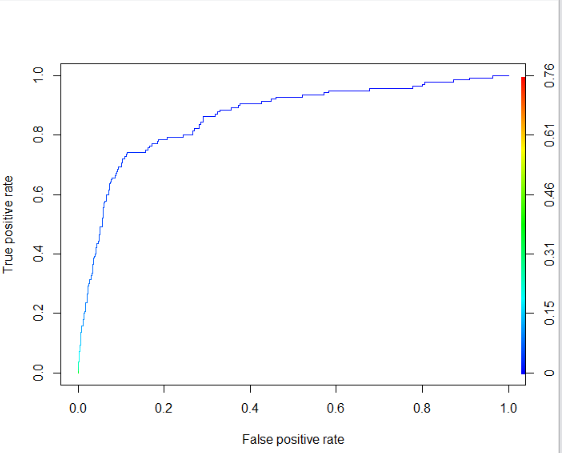


Figure 10: The resulting ROC Curve showing the accuracy of the LASSO model

The area under this ROC curve was calculated to 0.8637, indicating that the model has become an effective way to predict whether a customer will renew their magazine subscription. Further, this is an indication that the predictor variables ultimately used by the LASSO regression model are a reliable and useful source of information that can aid us in hopefully increasing the rate of renewal for Greyson Corporation.

**Recommendation for Greyson Corporation**

We recommend to Greyson Corporation that they can use this Logistic Regression model on future data to predict customer renewal behavior. In addition, we recommend that Greyson Corporation’s marketing department focus on the primary factors that drive renewal: “Child6.12,” “Child13.18,” “Income,” “PaidDirectMailOrders,” “TotalAmountPaid,” “DollarsPerIssue,” “MonthsSince1stOrder,” and “MonthsSinceLastOrder.”

One way that their marketing department could go about this task is preforming a Clustering Analysis on these variables to look for meaningful market segments. **Figure 11** shows a four-group clustering of all eight primary variables identified in the logistic regression model.

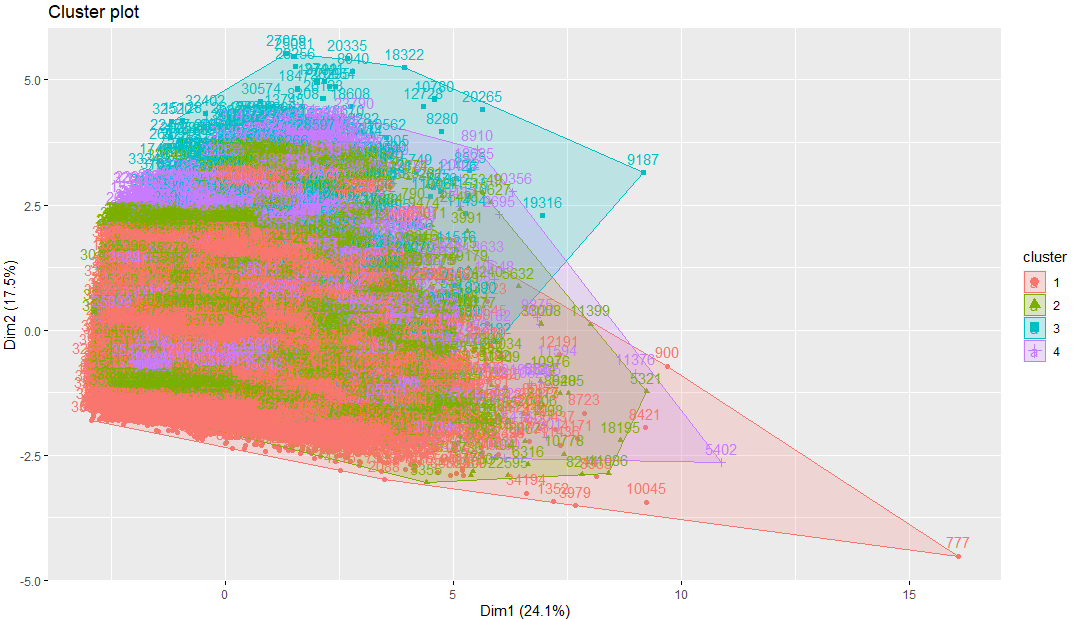


Figure 11: A cluster plot visualizing all eight primary variable models identified by the LASSO model

This is not very useful because too many variables are diluting the results and making an unorganized mess. However, if Greyson Corporation's marketing department were to cluster on fewer variables, say “Income” and “MonthsSinceLastOrder,” then a more meaningful and actionable cluster plot emerges (**Figure 12**).

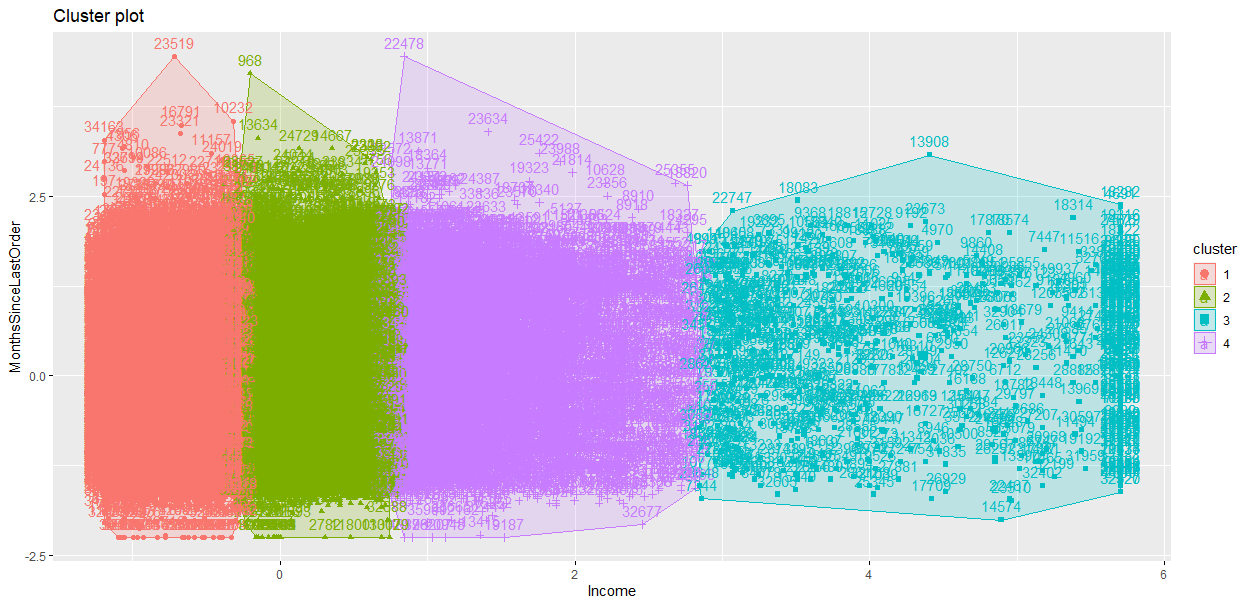


Figure 12: A cluster plot visualizing groupings derived by the “Income” and “MonthsSinceLastOrder” variables

With this example, there are four clear market segments that can be targeted with different campaigns to promote renewal rates. By experimenting with these identified variables the marketing department can create their own cluster plots for their focused campaigns.

**Conclusion and Next Steps**

As mentioned above, our client should consider the primary factors from the LASSO model when identifying potential customers that will renew. For example, “TotalAmountPaid” and “DollarsPerIssue” play a role in magazine subscription renewal. The pricing department at Greyson should focus on determining the best rates in order to retain their customer’s business. If prices are too high, a customer may be more inclined to let the magazine subscription expire. Furthermore, the client should perform research on how the time since first and last magazine order affects customer response. Greyson could first identify customers who have not ordered in a while and then focus its advertising efforts or offer special incentives to them.

The generated LASSO logistic regression model predicts customers who will renew their subscription with a high accuracy. We are confident that implementing this model will help Greyson Corporation boost their response rate.

**Code for Logistic Regression LASSO Analysis**

library(caret)

greyson\_data <- read.csv('Greyson.csv')

greyson\_data <- greyson\_data[-c(1)]

summary(greyson\_data)

str(greyson\_data)

#turn classification variables into factors so that future code can properly read data

greyson\_data$DwellingType <- as.factor(greyson\_data$DwellingType)

greyson\_data$Gender <- as.factor(greyson\_data$Gender)

greyson\_data$Marital <- as.factor(greyson\_data$Marital)

greyson\_data$ChildPresent <- as.factor(greyson\_data$ChildPresent)

greyson\_data$Occupation <- as.factor(greyson\_data$Occupation)

greyson\_data$MagazineStatus <- as.factor(greyson\_data$MagazineStatus)

greyson\_data$LastPaymentType <- as.factor(greyson\_data$LastPaymentType)

greyson\_data$GiftDonor <- as.factor(greyson\_data$GiftDonor)

#Split data between training and test data

set.seed(23)

index <- sample(nrow(greyson\_data),nrow(greyson\_data)\*0.80)

greyson\_train <- greyson\_data[index,]

greyson\_test <- greyson\_data[-index,]

#Impute data to account for null

library(imputeMissings)

trainMedians<- compute(greyson\_train, method = "median/mode")

greyson\_train <- impute(greyson\_train,object = trainMedians)

greyson\_test <- impute(greyson\_test, object = trainMedians)

summary(greyson\_train)

sum(is.na(greyson\_train))

#Variable selection with LASSO

library(dplyr)

dummy <- model.matrix(~., data = greyson\_train)

greyson\_train <- data.frame(dummy[,-1])

dummytest <- model.matrix(~., data = greyson\_test)

greyson\_test <- data.frame(dummytest[,-1])

greyson\_test$MagazineStatusN = 0 #add lost variable back into testing data

greyson\_test <- subset(greyson\_test, select=c(1:26,68,27:67))

names(greyson\_test)

greyson\_train\_X = as.matrix(select(greyson\_train, -Renewal))

greyson\_test\_X = as.matrix(select(greyson\_test, -Renewal))

greyson\_train\_Y = greyson\_train[,'Renewal']

greyson\_test\_Y = greyson\_test[, 'Renewal']

#Cross valildation of training data

library(glmnet)

greyson\_lasso\_cv <- cv.glmnet(x=greyson\_train\_X, y = greyson\_train\_Y, family = 'binomial', type.measure = 'auc', alpha = 1,nfolds=3)

plot(greyson\_lasso\_cv)

coef(greyson\_lasso\_cv, s = greyson\_lasso\_cv$lambda.1se)

predprob\_lasso <- predict(greyson\_lasso\_cv, s = greyson\_lasso\_cv$lambda.1se, newx = greyson\_test\_X, type = 'response')

#Get AUC and ROC curve for LASSO model

library(ROCR)

pred\_lasso <- prediction(predprob\_lasso, greyson\_test$Renewal)

perf\_lasso <- performance(pred\_lasso, 'tpr', 'fpr')

plot(perf\_lasso, colorize = TRUE)

auc\_lasso <- unlist(slot(performance(pred\_lasso, 'auc'), 'y.values'))

auc\_lasso

**Code for Clustering Analysis**

library(caret)

library(cluster)

library(factoextra)

#load DataSet

seg.df <- read.csv('Greyson.csv')

#Select variables from Logistic Regression Analysis

seg.df <- seg.df[,c(2,12,13,14,18,20,21,36,37)]

seg.df$Renewal<-as.factor(seg.df$Renewal)

#set as dataframe and omit na's

seg.df <- as.data.frame(seg.df)

seg.df <- na.omit(seg.df)

#convert all to numeric

seg.df$Renewal<-as.numeric(seg.df$Renewal)

seg.df$Child6.12<-as.numeric(seg.df$Child6.12)

seg.df$Child13.18<-as.numeric(seg.df$Child13.18)

seg.df$Income<-as.numeric(seg.df$Income)

seg.df$PaidDirectMailOrders<-as.numeric(seg.df$PaidDirectMailOrders)

seg.df$TotalAmountPaid<-as.numeric(seg.df$TotalAmountPaid)

seg.df$DollarsPerIssue<-as.numeric(seg.df$DollarsPerIssue)

seg.df$MonthsSince1stOrder<-as.numeric(seg.df$MonthsSince1stOrder)

seg.df$MonthsSinceLastOrder<-as.numeric(seg.df$MonthsSinceLastOrder)

seg.df.num <- seg.df

set.seed(23)

seg.k <- kmeans(seg.df.num, centers=4,nstart=25) #Left nstart at 25 random sets

seg.k$centers

#filter down variables from Logistic Regression Analysis to just two

seg.df.num <- seg.df.num[,c(4,9)]

fviz\_cluster(seg.k, data = seg.df.num)