International Journal of Business Marketing and Management (IJBMM)
Volume 3 Issue 1 January 2018, P.P.01-18
ISSN: 2456-4559
www.ijbmm.com

Brand Choice Prediction in Wine Consumption

Andy W. Chen[1]
[1] *University of British Columbia, Vancouver, BC, Canada*

ABSTRACT: *Increases in the world-wide consumption of wine over the past decade have placed pressure on wine-makers to more adequately meet consumer demand through targeted marketing strategies. It is of interest for wine producers to develop an understanding of the influential factors that drive demand within their target markets. Some of the factors influencing individual consumers are gender, age, income, location and product availability. This report provides an exploration of some of the factors influencing the preference of consumers for one brand of wine over another by addressing the following research question: Which of the variables among age, gender and income best predict the likelihood for wine consumers of choosing Brand B over Brand A? Using the brglm (bias reduction in binomial-response GLMs), the final selected model includes the variables* $Age^{-0.5}$ *and Gender. Both variables were found to be significant with a p-value less than 0.05. In order to assess the goodness of fit several evaluation methods were employed, including the Hosmer-Lomeshow Test, the Pearson Chi-Square Test, Pearson residuals, and Deviance Residuals. The results of these evaluations agree that the selected model provides a good fit to the data.*

KEYWORDS: *wine industry, brand choice, wine marketing, marketing, logistic regression*

I. INTRODUCTION

Trends in the wine industry have shown world-wide increases in consumption. It is estimated that approximately 240,915 hectoliters of wine were consumed in 2006, 244,294 hectoliters were consumed in 2007 and 245,012 hectoliters were consumed in 2008, according to a study by The Wine Institute. These trends show an increase in the total world-wide wine consumption (by volume) of 3.5% from 2004 to 2008.

With increases in wine consumption and increasing international competition, it is of interest to wine producers to develop a deeper understanding of the factors which influence the sale and consumption of their products. The wine industry offers some interesting challenges for market researchers and wine producers. In terms of marketing strategy, wine is a unique and complex product; in contrast to many other traditional products, the quality of a bottle of wine cannot be evaluated until after it is consumed. In addition to the marketing, pricing and presentation of wine, another key factor is present which must be carefully considered in the wine industry; the ability of potential consumers to select a bottle of wine is rather complex and depends largely upon the availability of appropriate information, prior knowledge and personal experience. There are a number of factors which stand to influence potential consumers in their selection process. Beverage preference, for example, is influenced by several factors, including gender, age, income, location, product availability and identity association.

This paper aims to explore the effect of gender, age and income on the preference of consumers between two brands of wine. This study examines the influence of these three factors on the preferences of 30 consumers of varying ages and income levels, and addresses the following research question: Which of the variables among age, gender and income best predict the likelihood for wine consumers of choosing Brand B over Brand A?

Past research includes different works on the brand choice using common products such as coffee and ketcup. Guadani and Little [1] studied consumer brand choice using purchase data of ground coffee recorded by 100 households and found brand loyalty, size loyalty, presence/absence of store promotion, regular shelf price

and promotional price cut to be highly significant. Winer [2] found that a brand choice models produced higher accuracy when including both reference and observed prices in a study of coffee purchases. Shimp [3] studied how advertising affects consumers' brand choices and found significant association between advertising and brand decisions. Nedungadi [4] conducted experiments to study how memory affects brand choices and found evidence for the influence of memory during the brand-choice process. Erdem and Swait [5] explored the role of brand credibility on brand choice for multiple products and found varying degrees of influence.

II. METHODOLOGY

The data contains two quantitative variables and two qualitative variables from a survey. These variables include two discrete quantitative variables and two qualitative variables which each have two levels. In order to analyze this dataset appropriately, I need to first gain an overall impression of the data through initial data exploration. The dataset in this study contains the following variables: Age (years), Annual income ($, thousands), Gender (M or F). Brand choice (A or B) is the response variable. I employ data exploration methods in order to get an initial impression of the dataset, which is shown below:

Subject ID	Age	Gender	Income	Choice
1	21	0	3	0
2	22	0	8	0
3	29	0	7	0
4	25	0	5	0
5	29	0	7	0
6	41	0	9	0
7	35	0	12	0
8	42	0	12	0
9	35	0	15	0
10	30	0	6	0
11	25	1	8	0
12	26	1	9	0
13	23	1	11	0
14	26	1	11	0
15	25	1	12	0
16	65	1	13	1
17	89	1	12	1
18	77	1	11	1
19	50	1	15	1
20	40	1	16	1
21	38	1	10	1
22	35	1	12	1
23	28	1	13	1
24	29	1	15	1
25	40	1	20	1
26	44	1	22	1
27	56	1	25	1
28	22	1	9	1
29	30	1	7	1
30	32	1	8	1

Figure 1. Survey data

Shown below are histograms of the gender and brand choice distributions. The distribution of choice is symmetric; half of the individuals in our sample chose Brand A and half chose Brand B. Shown below are histograms of income and age. The distribution of income is nearly symmetrical, but has a slight left skewness. On the right-hand side I can see that the distribution of age is asymmetrical with strong left skewness.

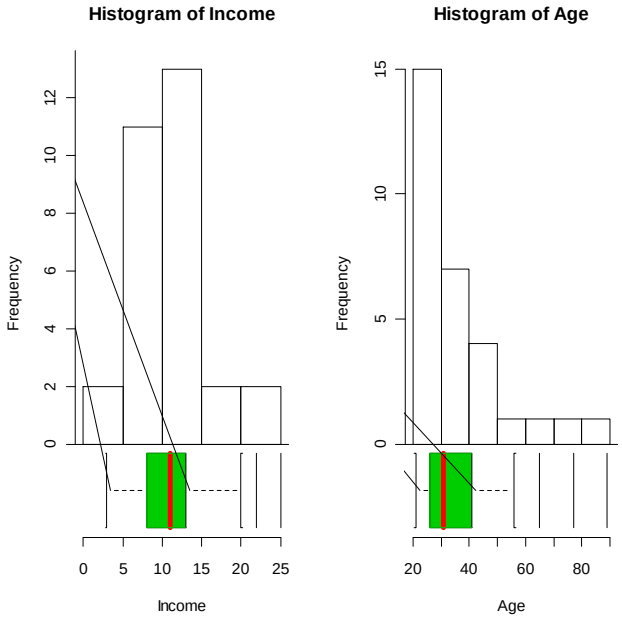

Figure 2. Histograms of income and age

The scatterplot below shows the relationship between age and income (by gender). It can be seen on this plot that a strong linear relationship exists between age and income among females, whereas a relatively weak linear relationship is found among males. There are some outliers in the apparently weak linear relationship between age and income among males.

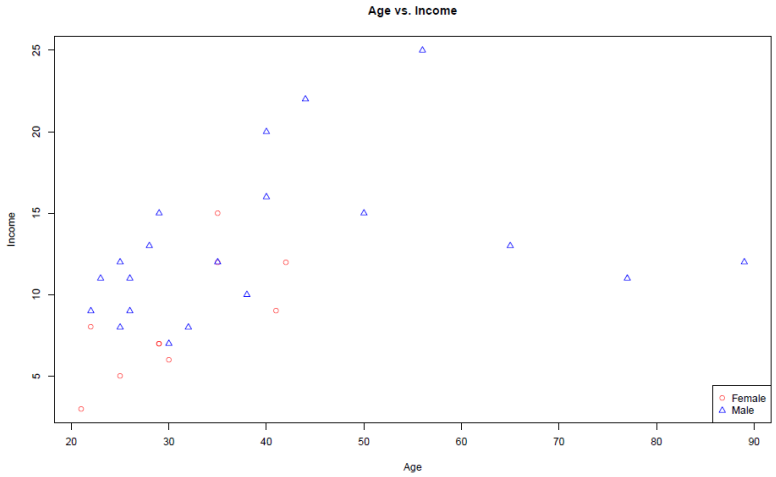

Figure 3. Scatterplot of age versus income by gender

Shown below is the distribution of our choice variable plotted against income. I can clearly see that consumers with income levels above the $15,000 mark prefer Brand B over Brand A. I also note the higher median income as indicated by the boxplot (to the right) for consumers who prefer Brand B.

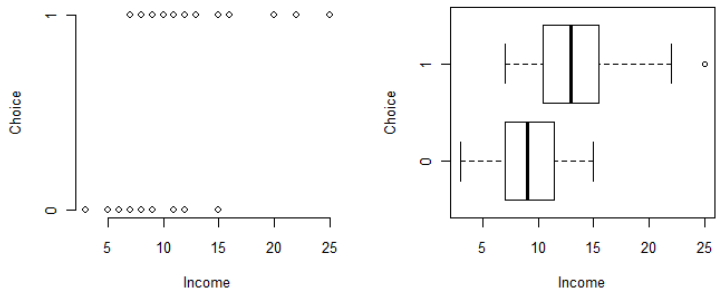

Figure 4. Distribution of brand choice versus income

The plot below shows the logit of the proportions of success (where success occurs when choice = 1) versus the midpoint values of the corresponding sub-ranges of income, as discussed above. I can see a strong linear relationship in this plot as would be required to satisfy the assumption of linearity needed to construct a binary logistic regression model of brand choice on income.

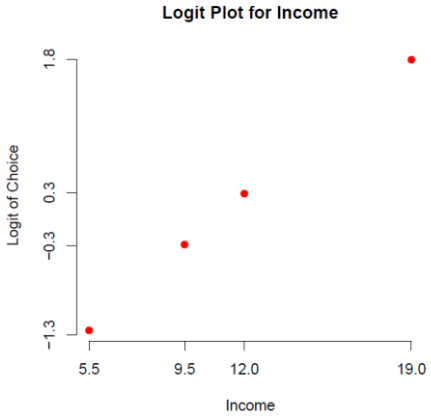

Figure 5. Logit of choice versus income

Shown below are the distributions of choice between Brand A (choice = 0) and Brand B (choice = 1) by gender. I observe that Brand A was preferred by 10 of the females and 5 of the males in the sample, whereas Brand B was selected by 15 of the males in the sample and none of the females. It must be noted for consideration in further analysis that none of the females in the sample chose Brand B over Brand A.

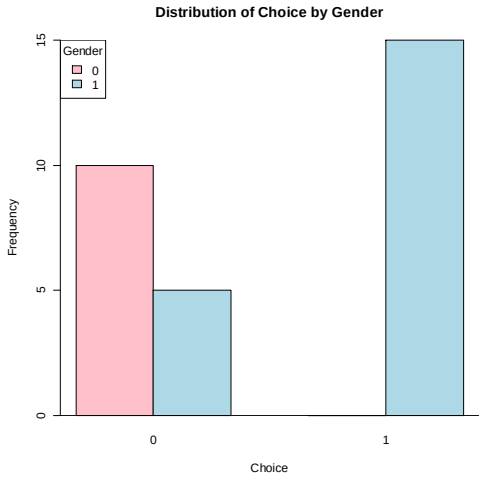

Figure 6. Distribution of age by gender

Shown below are the distributions for Brand A and Brand B (as represented by the brand choice variable) plotted against age. I can see that none of the individuals over the age of 43 included in our sample chose Brand A over Brand B. I do not, however, see a similar distinction in terms of Brand B. Also, from the boxplot on the right-hand side I can see the difference in the distribution of preference among consumers between these two brands.

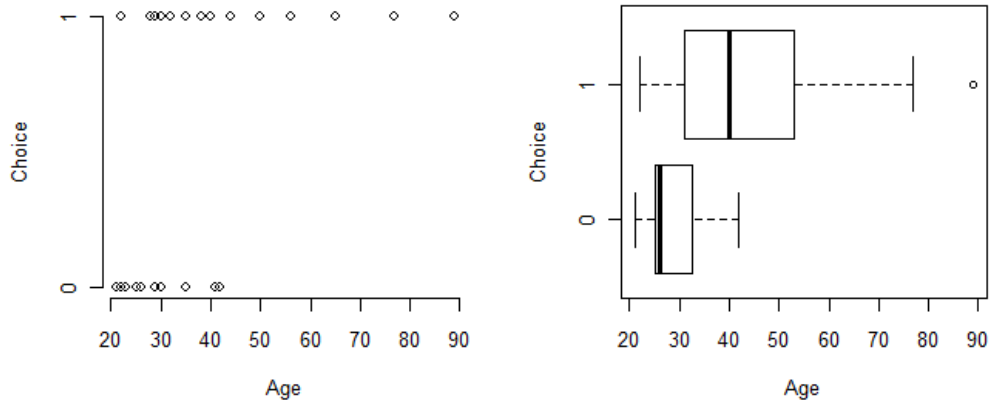

Figure 7. Distribution of brand choice versus age

One way of showing the relationship between Age and the predictor variable Choice is to divide Age into four contiguous sub-ranges and plot it against the variable "Choice". For each Age sub-range, the plot computes the proportion of successes (where a success occurs when Choice=1 i.e. when people choose Brand B) and plots the proportion of subjects with Age corresponding to the mid-point of that sub-range. When looking at the trend in these proportions, I hope to see that I can draw an imaginary S-shaped curve through these proportions. I do not observe an S-shape here but it is reasonably close to an S-shape given that our sample size is very small. When I plot Age versus Choice, I get the following graph where the blue dots represent the proportion of people who choose Brand B with Age in the first quartile, the second quartile, the third quartile and greater than the third quartile. The plot shows that as Age goes up, the proportion of successes (proportion of people choosing Brand B) goes up as well.

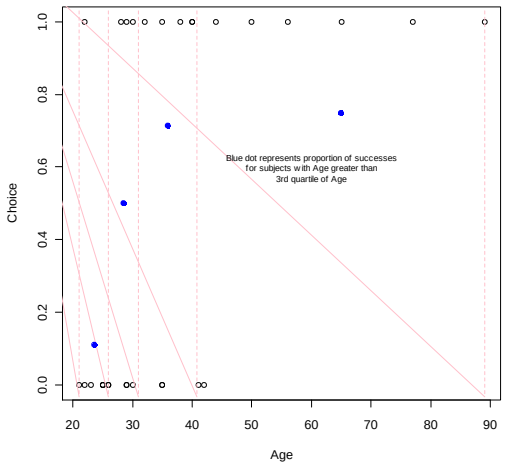

Figure 8. Distribution of brand choice versus age

The logit plot of Age versus Choice is not linear as shown in the graph below. Since our binary logistic regression assumes a linear relationship between the predictor variables and the response variable, I decided to test various transformations of the Age variable to make it linear.

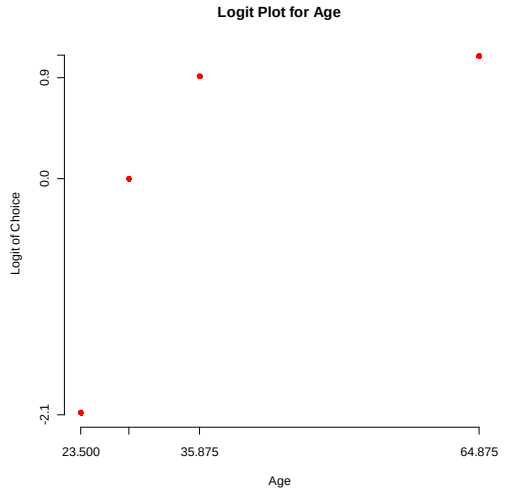

Figure 9. Logit of choice versus age

The various transformations used were $Age^{0.5}$, $\log_{10}(Age)$, $Age^{-0.5}$, Age^{-1}, Age^{-2}, Age^{-3}, Age^{-4} and Age^{-5}. The logit plots of all these transformations are given below. The variables $Age^{-0.5}$, Age^{-1}, Age^{-2}, Age^{-3}, Age^{-4} and Age^{-5} all have fairly linear logit plots. In order to decide which one of them to choose, I plot their histograms to see whether they have a symmetric distribution or not.

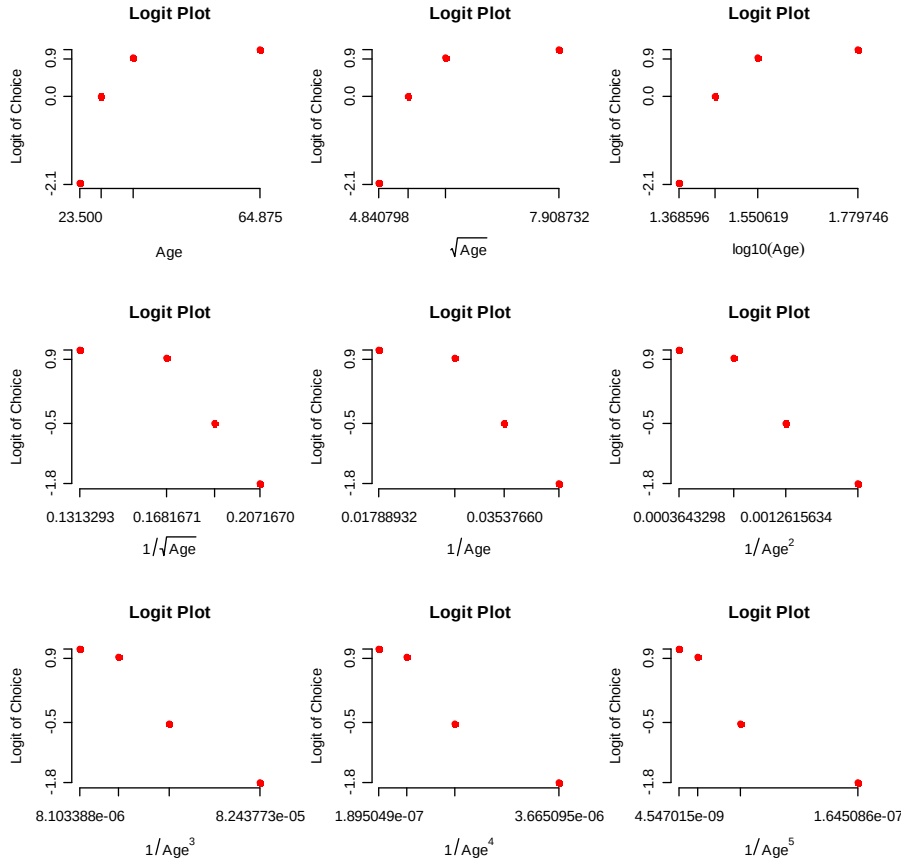

Figure 10. Logit plots of choice

III. RESULTS

To find a suitable binary logistic regression model, I would need to find the model with the lowest AIC. However, the data provided indicates that no females chose Brand B, so I cannot use the glm model because it generates an unreliable estimate and standard error for the effect of Gender on the probability of choosing Brand B. In fact, R generates an unusually high value for the binary logistic regression coefficient of Gender and for the ratio of odds of choosing Brand B for males compared to females because the number of females choosing Brand B is the denominator in the ratio and is 0. Using the bias-reduction method is not ideal for AIC-based model selection because AIC is intended for models fitted via the maximum likelihood method. This method suffices for the analysis. Nevertheless, I should be aware of the potential drawbacks when using this method in other scenarios.

I decided to keep Gender in the model, but use the bias-reduction method developed by Firth in 1993 with the use of the package brglm in R. I decided to test the model using Age, Age-0.5, Age-0.2, Age-2, Age-5, Age-1, Income, Gender, as well as the interactions between any two of Income, Gender, and one transformed Age. Using a heuristic approach I started by building the model including all the variables and found its AIC. If removing a variable from the model reduced the AIC, I would drop that variable. Conversely, if removing a variable resulted in a higher AIC, I would add that variable back to the model. The following table shows that models that I tested to find the one with the lowest AIC.

Variables Used	AIC
Age1 + Income + Gender + Age1:Income + Age1:Gender + Income:Gender	29.626
Age1 + Income + Gender + Age1:Income + Age1:Gender	27.078
Age1 + Income + Gender + Age1:Income	24.251
Age1 + Income + Gender	21.584
Age1 + Income	35.022
Age1 + Gender	18.918*
Income + Gender	26.649
Age2 + Income + Gender + Age2:Income + Age2:Gender + Income:Gender	29.486
Age2 + Income + Gender + Age2:Income + Age2:Gender	26.938
Age2 + Income + Gender + Age2:Income	24.107
Age2 + Income + Gender	21.485
Age2 + Income	34.910
Age2 + Gender	18.796
Age + Income + Gender + Age:Income + Age:Gender + Income:Gender	28.946
Age + Income + Gender + Age:Income + Age:Gender	26.372
Age + Income + Gender + Age:Income	23.585
Age + Income + Gender	21.198
Age + Income	34.723
Age + Gender	18.427
Age4 + Income + Gender + Age4:Income + Age4:Gender + Income:Gender	30.044
Age4 + Income + Gender + Age4:Income + Age4:Gender	27.489
Age4 + Income + Gender + Age4:Income	24.649
Age4 + Income + Gender	22.159
Age4 + Income	35.746
Age4 + Gender	19.669
Age5 + Income + Gender + Age5:Income + Age5:Gender + Income:Gender	31.536
Age5 + Income + Gender + Age5:Income + Age5:Gender	28.745

Age5 + Income + Gender + Age5:Income	25.064
Age5 + Income + Gender	24.334
Age5 + Income	37.004
Age5 + Gender	22.351
Age6 + Income + Gender + Age6:Income + Age6:Gender + Income:Gender	29.811
Age6 + Income + Gender + Age6:Income + Age6:Gender	27.258
Age6 + Income + Gender + Age6:Income	24.459
Age6 + Income + Gender	21.748
Age6 + Income	35.243
Age6 + Gender	19.136

Table 1. Variables used for different models and the respective AICs (*Final model)

The results indicate that the most suitable model is the one using $Age^{-0.5}$ and Gender as the variables. The final model is $\text{Choice} = \beta_0 + \beta_1 \cdot \text{Age1} + \beta_2 \cdot \text{Gender}$. The AIC is higher than the model with variables Age and Gender. However, Age1 (equivalent of $Age^{-0.5}$) is the transformation of Age that has most symmetrical and normally distributed distribution significant at the 5% level. Therefore, I chose the model with Age1 and Gender as the final model.

The intercept $\beta_0 = 10.791$ means that the for a female with age 0, the log odds of choosing Brand B is 10.791.The coefficient $\beta_1 = -81.149$ means that for each increase in $Age^{-0.5}$ for a particular gender, the log odds of choosing Brand B is estimated to decrease by 81.149. The coefficient $\beta_2 = 5.051$ means that for each increase in Gender, the log odds of choosing Brand B is estimated to increase by 5.051. In other words, the log odds of choosing Brand B for males (1) is estimated to be 5.051 greater than the log odds of choosing Brand B for females (0).

IV. DISCUSSION

In order to assess the goodness of fit, I will use several methods, such as the Hosmer-Lomeshow Test and the Pearson Chi-Square Test.

Examining the appropriateness of the fitted logistic regression model before using it is necessary. The Hosmer-Lemeshow test is a statistical test for goodness of fit of the logistic regression models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The statistical Hosmer–Lemeshow test specifically identifies subgroups of fitted values. The test statistic asymptotically follows a χ^2 distribution with n-2 degrees of freedom. The number of subgroups should be 6 for the dataset, each of which including 5 observations. The Hosmer-Lomeshow test examines the following competing hypotheses:

$$\begin{cases} H_0 = \text{Binary logistic model provides a satisfactory fit to the data} \\ H_1 = \text{Binary logistic model does not provide a satisfactory fit to the data} \end{cases}$$

The Hosmer–Lemeshow test (p-value= 0.485) indicates that the wine brand choices of individuals are not significantly different from those predicted by the model, and that the overall model fit is good. In other words, based on the Homer-Lemshow test, I fail to reject H_0 and therefore conclude that the model fits the data well. Further testing is done to assess the goodness of this fit in the Model Diagnostics section near the end of this report.

I can produce an analysis of deviance for the sequential addition of each variable by using the anova function, specifying the chi-squared test to examine for difference between models.

$$\text{Choice} = \beta_0 + \beta_1 \cdot \text{Age1} + \beta_2 \cdot \text{Gender} + \varepsilon \qquad (1)$$

The very small P-value (5.946e-07) strongly suggests that this null hypothesis is false. I conclude that Age1 and Gender is needed in the model and therefore, the obtained model is a well fitted model.

To estimate the probability that people would choose Brand A or B, a logistic regression model was used and the following logit model was estimated:

$$\log\left[\frac{P(Choice_i=1|Age_i,Gender)}{1-P(Choice_i=1|Age_i,Gender)}\right] = \beta_0 + \beta_1 Age^{-0.5}{}_i + \beta_2 Gender_i + \varepsilon_i \qquad (2)$$

The above model uses the logarithm of the odds of choosing Brand B. The model can be written on the odds scale as follows:

$$\frac{P(Choice_i=1|Age_i,Gender)}{1-P(Choice_i=1|Age_i,Gender)} = exp(\beta_0 + \beta_1 Age^{-0.5}{}_i + \beta_2 Gender_i + \varepsilon_i) \qquad (5)$$

The results are summarized in the following table.

Variable	Coefficient	Std. Error	z-value	P (>|z|)
Intercept	10.791	6.829	1.580	0.1141
Age	-81.150	40.774	-1.990	0.0466
Gender	5.051	2.065	2.447	0.0144

V. Table 2. Logistic regression results

β_0 is estimated to be 10.791 and it is insignificant which is alright because it is not an interpretable quantity. β_1 is estimated to be -81.150 which means that for every one unit change in $Age^{-0.5}$ for any given gender, the log odds of choosing Brand B decrease by -81.50. On the odds scale it implies that for every one unit change in $Age^{-0.5}$ for any given gender, the odds of choosing Brand B decrease by 100% {(5.714823e-36 – 1)× 100%}. β_2 is estimated to be 5.051 which means that the difference in log odds of choosing Brand B between males and females of a given age is 5.051. On the odds scale it means the odds of choosing Brand B for males is 156 (exp (5.051)) times the odds of choosing Brand B for females. β_1 and β_2 are significant with a p-value less than 0.05. This shows that the effect of $Age^{-0.5}$ and gender on the choice of Brand A and B is statistically significant. .

VI. FIGURES AND TABLES

Subject ID	Age	Gender	Income	Choice
1	21	0	3	0
2	22	0	8	0
3	29	0	7	0
4	25	0	5	0
5	29	0	7	0
6	41	0	9	0
7	35	0	12	0
8	42	0	12	0
9	35	0	15	0
10	30	0	6	0
11	25	1	8	0
12	26	1	9	0
13	23	1	11	0
14	26	1	11	0
15	25	1	12	0
16	65	1	13	1
17	89	1	12	1
18	77	1	11	1
19	50	1	15	1
20	40	1	16	1
21	38	1	10	1
22	35	1	12	1
23	28	1	13	1
24	29	1	15	1
25	40	1	20	1
26	44	1	22	1
27	56	1	25	1
28	22	1	9	1
29	30	1	7	1
30	32	1	8	1

Figure 1. Survey data

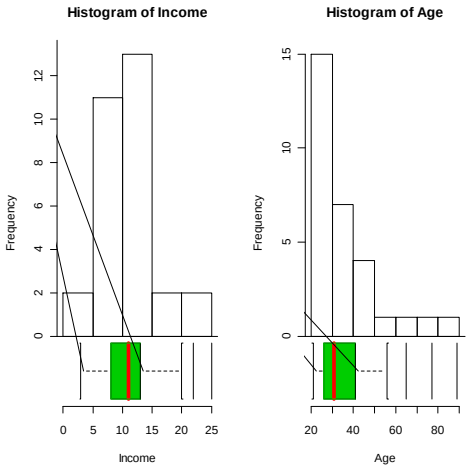

Figure 2. Histograms of income and age

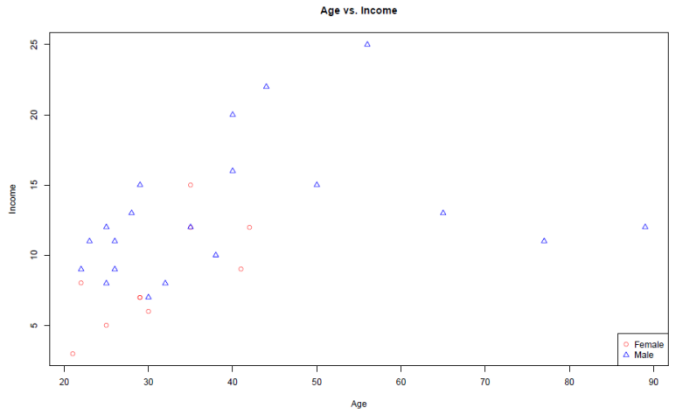

Figure 3. Scatterplot of age versus income by gender

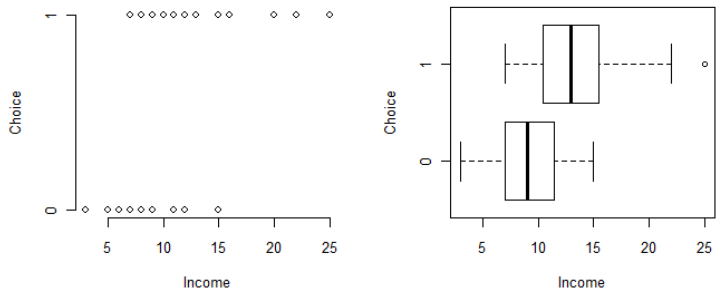

Figure 4. Distribution of brand choice versus income

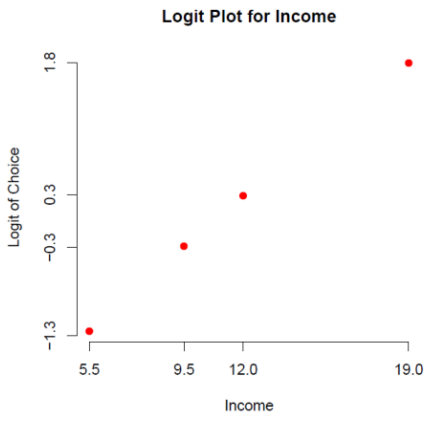

Figure 5. Logit of choice versus income

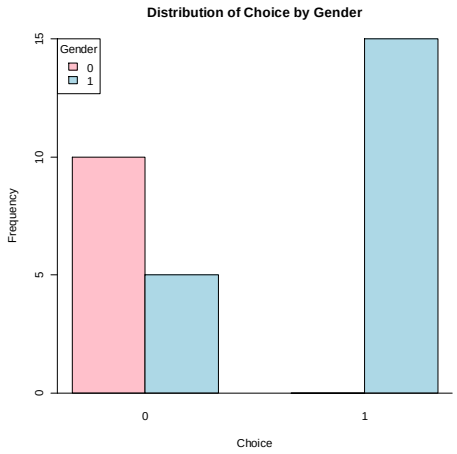

Figure 6. Distribution of age by gender

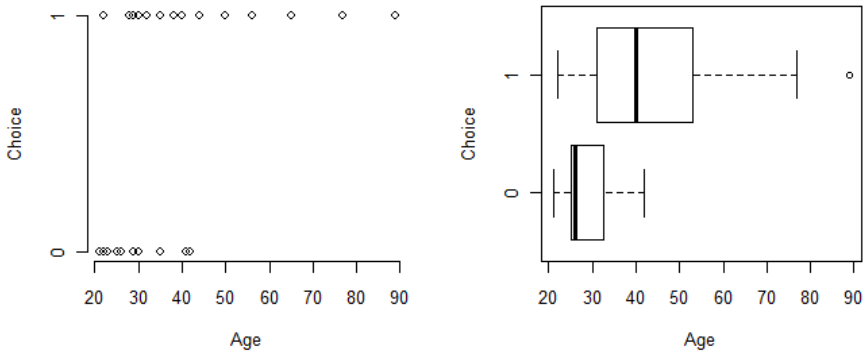

Figure 7. Distribution of brand choice versus age

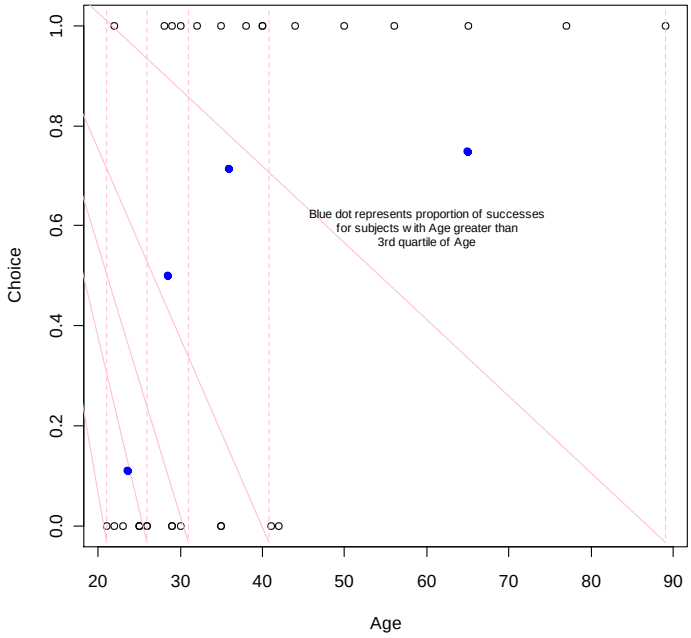

Figure 8. Distribution of brand choice versus age

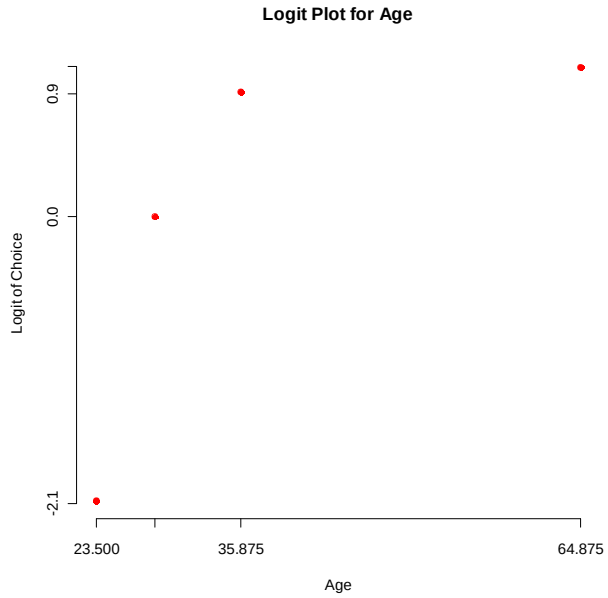

Figure 9. Logit of choice versus age

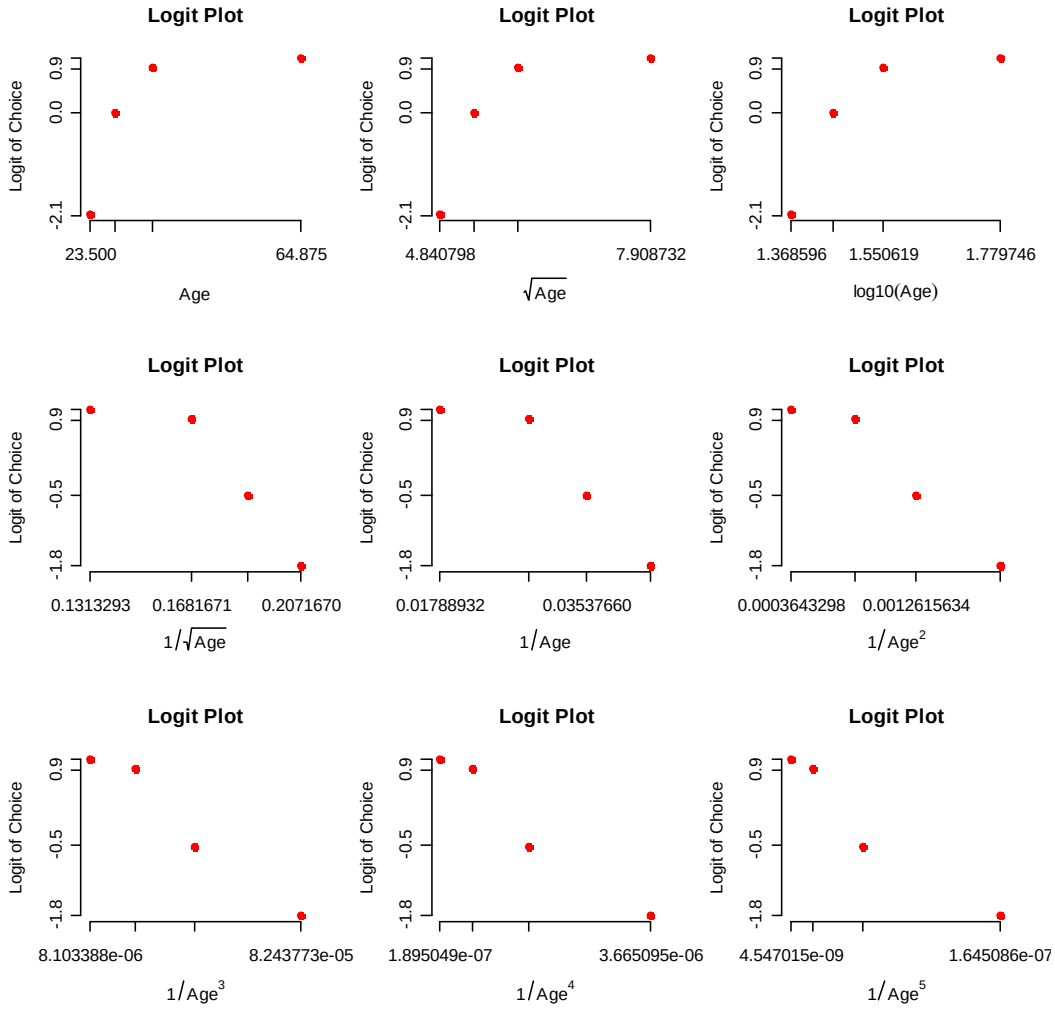

Figure 10. Logit plots of choice

Variables Used	AIC
Age1 + Income + Gender + Age1:Income + Age1:Gender + Income:Gender	29.626
Age1 + Income + Gender + Age1:Income + Age1:Gender	27.078
Age1 + Income + Gender + Age1:Income	24.251
Age1 + Income + Gender	21.584
Age1 + Income	35.022
Age1 + Gender	18.918*
Income + Gender	26.649
Age2 + Income + Gender + Age2:Income + Age2:Gender + Income:Gender	29.486
Age2 + Income + Gender + Age2:Income + Age2:Gender	26.938

Age2 + Income + Gender + Age2:Income	24.107
Age2 + Income + Gender	21.485
Age2 + Income	34.910
Age2 + Gender	18.796
Age + Income + Gender + Age:Income + Age:Gender + Income:Gender	28.946
Age + Income + Gender + Age:Income + Age:Gender	26.372
Age + Income + Gender + Age:Income	23.585
Age + Income + Gender	21.198
Age + Income	34.723
Age + Gender	18.427
Age4 + Income + Gender + Age4:Income + Age4:Gender + Income:Gender	30.044
Age4 + Income + Gender + Age4:Income + Age4:Gender	27.489
Age4 + Income + Gender + Age4:Income	24.649
Age4 + Income + Gender	22.159
Age4 + Income	35.746
Age4 + Gender	19.669
Age5 + Income + Gender + Age5:Income + Age5:Gender + Income:Gender	31.536
Age5 + Income + Gender + Age5:Income + Age5:Gender	28.745
Age5 + Income + Gender + Age5:Income	25.064
Age5 + Income + Gender	24.334
Age5 + Income	37.004
Age5 + Gender	22.351
Age6 + Income + Gender + Age6:Income + Age6:Gender + Income:Gender	29.811
Age6 + Income + Gender + Age6:Income + Age6:Gender	27.258
Age6 + Income + Gender + Age6:Income	24.459
Age6 + Income + Gender	21.748
Age6 + Income	35.243
Age6 + Gender	19.136

Table 1. Variables used for different models and the respective AICs (*Final model)

Variable	Coefficient	Std. Error	z-value	P (>|z|)
Intercept	10.791	6.829	1.580	0.1141
Age	-81.150	40.774	-1.990	0.0466
Gender	5.051	2.065	2.447	0.0144

Table 2. Logistic regression results

VII. CONCLUSION

The aim of this report was to explore the effect of gender, age and income on the preference of consumers between two brands of wine, by addressing the following research question: Which of the variables among age, gender and income best predict the likelihood for wine consumers of choosing Brand B over Brand A?

In addressing this question I examined a dataset comprising of 30 observations, each of which having data in the form of two quantitative variables and two qualitative variables, including the age, gender, income and brand choice of each consumer in the sample.

These data were examined and an appropriate corresponding statistical model was selected through the use of Akaike's information criterion. I made use of bias reduction in binomial-response GLMs (brglm) because I wished to include females in the model even though no females had chosen Brand B within the sample population.

A transformation of the variable Age was made use of in the selected model, which includes $\text{Age}^{-0.5}$ instead of Age. This transformation is able to provide a symmetrical and normally distributed distribution which is significant at the 5% level. The parameters β_0, β_1 and β_2 were each estimated and it was found that both β_1 and β_2 are significant at the 95% confidence level, while β_0 is not. The value of β_0, though, does not have an inherent practical meaning in this case, and it is estimated that $\beta_0 = 10.791$. The estimated coefficient $\beta_1 = -81.149$ means that for each increase in $\text{Age}^{-0.5}$ for a particular gender, the log odds of choosing Brand B is estimated to decrease by 81.149. The estimated coefficient $\beta_2 = 5.051$ means that for each increase in Gender, the log odds of choosing Brand B is estimated to increase by 5.051. In other words, the log odds of choosing Brand B for males (1) is estimated to be 5.051 greater than the log odds of choosing Brand B for females (0).

In order to evaluate this model further I employed the Hosmer–Lemeshow test, the Pearson Chi-Square Test, Pearson residuals, and Deviance Residuals. The results of this test concluded that the model is a good fit to the data and that brand choices of individuals in the sample are not significantly different from those predicted by the model. Brand choice predictions for the sample population based on the model were in line with the actual outcomes; the model was able to correctly predict 97% of the respondents' wine preferences.

I employed model diagnostic checks in order to assess the validity of this model further. These checks concluded that the major assumptions required for logistic regression modeling are very near completely satisfied in the model.

The results of the statistical analysis and model selection brought forth a model which stands to provide a good fit to the data while reasonably satisfying the major assumptions for logistic regression models. This model provides predictive capabilities based on the age and gender of individuals who are selecting wine between Brand A and Brand B.

REFERENCES

[1] P Guadagni, J Little, A Logit Model of Brand Choice Calibrated on Scanner Data, *Marketing Science, Vol. 2, No. 3,* 1983, pp. 203-238.

[2] R Winer, A Reference Price Model of Brand Choice for Frequently Purchased Products, *Journal of Consumer Research, Volume 13, Issue 2,* 1986, pp. 250–256.

[3] T Shimp, Attitude toward the AD as a Mediator of Consumer Brand Choice, *Journal of Advertising, Vol. 10, No. 2,* 1981, pp. 9-15.

[4] P Nedungadi, Recall and Consumer Consideration Sets: Influencing Choice without Altering Brand Evaluations, *Journal of Consumer Research, Volume 17, Issue 3,* 1990, pp. 263–276.

[5] T Erdem, J Swait, Brand Credibility, Brand Consideration, and Choice, *Journal of Consumer Research, Volume 31, Issue 1,* 2004, pp. 191–198.