Wine analysis: Using Multiple Linear Regression to Determine What Personal Factors Make Consumers Buy Wine

Diya Kamath

Introduction:

Consumerism is growing at an exponential rate and companies have to adapt and cater to the people they are marketing to. One of the ways in which companies can achieve this is by using statistical methods to understand the demographic of the people they are marketing to in order to predict their purchasing patterns so that they can employ an appropriate marketing method. Studies have previously been done to study the impact of demographic factors on consumer behaviour (Kumar, 2014), and more specifically understand if there are lifestyle differences in wine drinkers, by understanding their consumers through segmentation (Bruwer and Li, 2007) as well as studies done to gain insight into the product style preferences of consumers and how that affects wine product marketing (Bruwer et al, 2011). Deriving inspiration from previous studies. my research question explores if there are certain personal factors that consumers have that make them more likely to buy wine.

Methods:

The data set that was used had a research question in mind and had a selected response variable as well as several variables that were relevant for their research question. However, this study will use ‘money spent on wine’ as the response variable and will remove the variable that is not necessary for this study. In general, the variables in the original data set were split up into groups, the groups were: personal information of consumers, the money spent on different products in the store, and promotional factors the store used. The research question for this study did not require the variables that were under the category of money spent on other products or promotional factors that the store used. Due to this there were 8 variables used , 1 being the response and the other 7 being the variables that fell under the category of personal information of the consumers. The 8 variables were: the amount of money spent on wine (response), marital status, education level, number of kids at home, number of teenagers at home, date the customer joined the store, birth year and income.

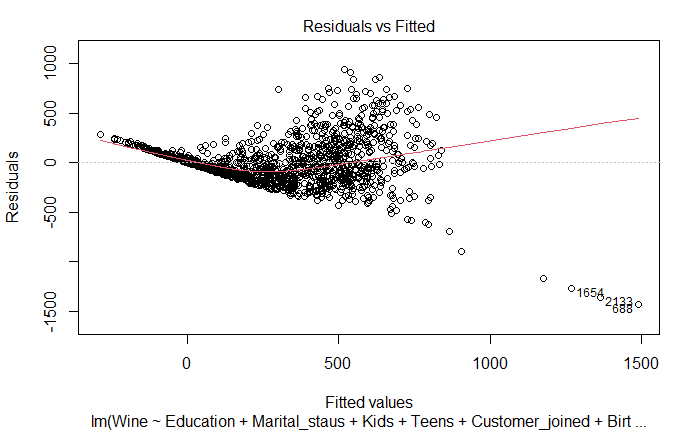
Exploratory data analysis done to take a look at the data before working with it to note any irregularities in the data, or to note factors that might cause problems when trying to fit a linear model to the data. First, a general scatterplot of all the variables plotted against each other were made to identify collinearity and to see whether or not the variables showed a linear relationship at all (figure 1). Then ,each variable was looked at via histograms or bar plots (if they were qualitative variables) to check the distribution of the variables and to make sure each variable followed a normal distribution – as it is an important factor when fitting a linear regression model (which will be discussed later). All the variables did follow a normal distribution however it was noted that the response variable was extremely right skewed.

Figure1.  plot of all the predictors 


Figure : Plot of predictors

Before fitting a linear model, to make the model usable to make predictions, model assumptions need to be satisfied . The first assumption is model linearity, this was checked in the residual plots (figure 2), there was no curve or systematic pattern in the residuals and therefore was satisfied. The second assumption is uncorrelated errors, the same plot was used to check for large clusters of residuals that are separated from the rest of the values; this assumption was satisfied as well. The third assumption is constant variance, once again the residual plot can be used to search for any pattern, there sems to be a fanning pattern in the residual plot which show s a violation of constant variance.

Since there was a violation in constant variance, conditions were checked, and transformations were applied. Before the application of the transformation, it was important to check the residual vs fitted plot of the variables to check the relationship between the response variable and each individual predictor. While analysing these plots, the fanning pattern was noticed in the residual vs fitted plot between the response variable and the predictor variable “income”. Since it was an issue of constant variance using a Poisson response variable removed the dependence of the error variance on the predictor values as it was a variance stabilizing transformation. There was no longer a fanning pattern in the residual plot (figure 3 )

Chart, scatter chart

Description automatically generated

Figure 2: Residual vs fitted plot pre transformation Figure 3: Residual vs fitted plot after transformation

After the transformations were applied, the model could then be tested to see which predictors would be the most beneficial when applying a linear regression model on the data. To do this, an ANOVA test was done to check if there was an overall linear relationship in the model, there were statistically significant results for the following predictors: education, kids, when the customer joined and income. Teens, birth year and marital status did not. To confirm this, individual T tests on the predictors were performed, which yielded the same result (Table 1).The two results were run through a partial F test which showed that under the null hypothesis the predictors marital status, number of teens, and birth year are not in the true relationship and can be removed from the predictor and confirmed the previous findings.

Graphical user interface, table

Description automatically generated

Table 1: Original model

Before the model removed the predictors, a test to check collinearity and problematic observations was performed. There was no collinearity between the predictors and there were some problematic observations observed in the following predictors: marital status, number of teens at home, Birth year and income. However, before transformations and further analysis this did not provide relevant information. Once the predictors were removed and the new model was formed, the model was re-tested for collinearity and problematic observations, the results showed no collinearity between predictors and most of the predictors that had problematic observations in the original model were removed, except for income (the implications of this will be discussed later).

To make sure the model is equally good within the population. In the beginning of the analysis, the data was separated into testing and training data to make sure the results found in the sample can be transferred to the population. Once the training data was used to make the model, this same model was applied to the testing data. Numerical summaries showed that the model fit the testing data and there was no statistically significant differences.

Results:

After the analysis, the linear model that was the best for the data involved the predictors presented in table below. This model shows that there seems to be a linear relationship between the independent variables income, education, the amount of kids and the data the customer joined are all personal factors that contribute and can predict to the amount of money spent on wine.

Table

Description automatically generated

Table 2: New model

Discussion:

The multiple linear regression model produced shows that out of all the personal factors collected. When all other predictors are fixed, the decrease in the number of children the customer has, along with if they did not join in 2013 or 2014, if they held a graduate, masters or PHD degree, combined with an increase in income showed an increase in wine sales (figure). This model is true only when we interpret each predictor by holding the other predictors fixed, without any other predictors. The R adjusted value being 0.651 shows a decent linear relationship.

This information can be applied to the general population and can provide insights to the wine industry in regard to targeting a specific demographic of people (e.g. people who have done graduate studies) to reap the most monetary benefits.

Although this model was carefully analyzed, it is still important to discuss the problematic observations and biases that could have caused error in the model. First, the data collected was specific to a store – this could cause several demographic biases (selection bias). Due to this there could be issues in using this model for the larger population. Furthermore, the problematic observations noted in the income predictor could account for more variability in the model which could have caused a problem with inference. All in all, the multiple linear regression approach to this data yielded a beneficial model that could be used by wine stores, grocery stores and marketing companies. Future research could collect and test more predictors to see if the model changes or remains the same.

References:

Bruwer, J., & Li, E. (2007). Wine-related lifestyle (WRL) market segmentation: Demographic and behavioural factors. *Journal of Wine Research*, *18*(1), 19–34. https://doi.org/10.1080/09571260701526865

Bruwer, J., Saliba, A., & Miller, B. (2011). Consumer behaviour and sensory preference differences: Implications for wine product marketing. *Journal of Consumer Marketing*, *28*(1), 5–18. https://doi.org/10.1108/07363761111101903

Kumar, R. (2014). Impact of demographic factors on consumer behaviour - A consumer behaviour survey in Himachal pradesh. *Global Journal of Enterprise Information System*, *6*(2), 35. https://doi.org/10.15595/gjeis/2014/v6i2/51844