
**Russ Conte**

**Predict 411- DL - 55**


# Report on Improving Wine Sales in our Stores:

November 20, 2016

# Table of contents



# Executive Summary

## Understanding the problem/opportunity

Dear Management:

We are in a very enviable position. Our company has been presented with a data set about wine sales in our stores. Careful analysis of the data will allow us to predict the number of sales per store, but also to make recommendations to maximize wine purchases and minimize unsold wine in our stores. These two steps will allow us to maximize our profit for our wine division. This analysis will shed light about how our company can to do that, based on the data we have received.

If this opportunity works for the Wine division, we will recommend similar techniques be applied to other divisions in our company.

## Summary of Business Recommendations–Full Recommendations at End of Report

The data will show that wine with 0 stars is much less likely to sell than even wine with one star. Our analysis will further show that wine with three or more stars are going to sell at least one bottle, hopefully many more. Thus our business recommendation will be to maximize the ordering and sale of brands of wine with the most stars, and minimize the ordering and sale of wines with zero stars. The analysis predicts this will maximize our profit in our wine division.

The summary will show our best business move today is to stock light wines that are at least three stars and have strong label appeal. These are the best sellers in our stores. We can use more analytics to reach these goals.

## Possible Next Steps

A second recommendation will be to do more gathering of data. Will it make a difference what time of day the purchase is made? What day of the week? The weather outside? Who is working at the counter? What is the price of the wine? We would do well to investigate all the factors that have the highest contribution to purchase of wine in our stores, allowing us to increase our sales beyond what we are able to establish with the current dat

# Overview of the current data

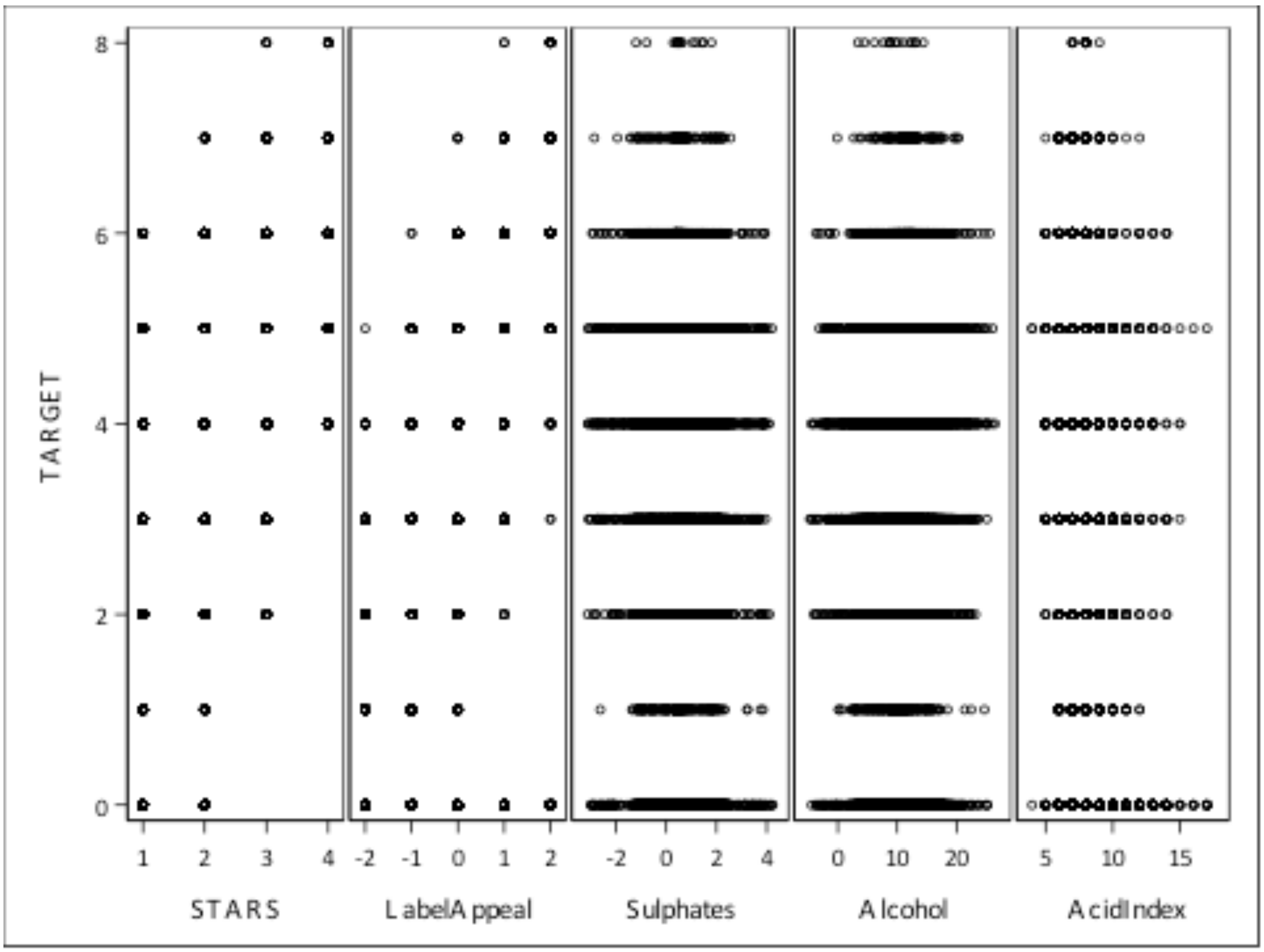


## Overview of the Wine data

The Wine data consists of 16 columns and 12,795 rows of data. Each column is a factor (for example, pH) and each row is a purchase. We do not have information on the individual doing the purchasing, nor dates, nor prices. Several factors are missing data (such as Stars) and we will address that in our analysis.

*Which individual factors are most predictive of sales?*

For this section, we investigate which of the individual factors in our data set are most predictive of sales. A simple scatter plot shows that Stars has the highest predictive value, followed by Label Appeal. All the remaining factors are not predictive of sales by themselves.



### Summary statistics about the data:

We can see a lot of valuable information from the table below. The most valuable piece of information here is that the average person purchased three bottles of wine from us, as can be seen in the highlighted box below:

| Variable | N | N Miss | Minimum | Maximum | Mean | Median |
|---|---|---|---|---|---|---|
| INDEX | 12795 | 0 | 1 | 16129 | 8069.98 | 8110 |
| TARGET | 12795 | 0 | 0 | 8 | 3.0290739 | 3 |
| FixedAcidity | 12795 | 0 | -18.1 | 34.4 | 7.0757171 | 6.9 |
| VolatileAcidity | 12795 | 0 | -2.79 | 3.68 | 0.3241039 | 0.28 |
| CitricAcid | 12795 | 0 | -3.24 | 3.86 | 0.3084127 | 0.31 |
| ResidualSugar | 12179 | 616 | -127.8 | 141.15 | 5.4187331 | 3.9 |
| Chlorides | 12157 | 638 | -1.171 | 1.351 | 0.0548225 | 0.046 |
| FreeSulfurDioxide | 12148 | 647 | -555 | 623 | 30.8455713 | 30 |
| TotalSulfurDioxide | 12113 | 682 | -823 | 1057 | 120.7142326 | 123 |
| Density | 12795 | 0 | 0.88809 | 1.09924 | 0.9942027 | 0.99449 |
| pH | 12400 | 395 | 0.48 | 6.13 | 3.2076282 | 3.2 |
| Sulphates | 11585 | 1210 | -3.13 | 4.24 | 0.5271118 | 0.5 |
| Alcohol | 12142 | 653 | -4.7 | 26.5 | 10.4892363 | 10.4 |
| LabelAppeal | 12795 | 0 | -2 | 2 | -0.009066 | 0 |
| AcidIndex | 12795 | 0 | 4 | 17 | 7.7727237 | 8 |
| STARS | 9436 | 3359 | 1 | 4 | 2.041755 | 2 |

It's also noteworthy that many of our variables have missing values, and we will deal with those next.

### Cleaning up the data

The table above shows thousands of missing data points. We have several options: Delete the rows or columns with missing data; replace the missing values with the mean for each variable; attempt to run the analysis without replacing the data. The best option is the second choice–replace all missing values with the mean for each variable. For example, the 616 missing values for ResidualSugar will be replaced with 5.4187331, and the same process will be done for all the other missing values.

Once the data are clean, we will proceed to make three models and evaluate them. That is the next section.

# First possible model: Regression

## We'll evaluate Stepwise, Forward and Backward regression.

Our first regression model was a forward regression model. A summary of the model is:

| Analysis of Variance for Model #1: Stepwise Regression | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 12 | 25547 | 2128.92261 | 1240.85 | <.0001 |
| **Error** | 12782 | 21930 | 1.7157 | | |
| **Corrected Total** | 12794 | 47477 | | | |

This table shows the model is significant with an F-Value of 1240.85, and a p-value <0.0001. The $R^2$ for this model is 0.5381 and Mallow's C(p) = 12.9350.

| Analysis of Variance for Model #2: Forward Regression | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 13 | 25550 | 1965.40804 | 1145.62 | <.0001 |
| **Error** | 12781 | 12927 | 1.71558 | | |
| **Corrected Total** | 12794 | 47477 | | | |

This table shows that model #2 is very similar to model #1. The $R^2$ for model #2 is 0.5382, and Mallow's C(p) is 13.0505.

| Analysis of Variance for Model #3: Backward Regression | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 12 | 25547 | 2128.92261 | 1240.85 | <.0001 |
| **Error** | 12782 | 21930 | 1.71558 | | |
| **Corrected Total** | 12794 | 47477 | | | |

This table shows that Model #3: Backward Regression is extremely similar to Model #1 and Model #2. The R2 for tis model is 0.5381 and Mallow's C(p) is 12.9350.

**Summary of regression models:**

All three models are basically equal, any of the three can be used with equal results.

# Second possible models: Poisson, Gamma, Normal distributions

This table summarizes the results from these three models:

| | Poisson | Gamma | Normal |
|---|---|---|---|
| AIC (smaller is better) | 45744.9334 | 25741.7988 | 43229.8159 |
| AICC (smaller is better) | 45744.9507 | 25741.8251 | 43229.8366 |
| BICC (smaller is better) | 45819.5015 | 25821.1795 | 43311.8408 |
| Full Log Likelihood | -22862.4667 | -12859.8994 | -21603.9080 |

The Gamma distribution is clearly the best with the smallest values for all three measures: AIC, AICC and BIC.

# Model Criteria and Selection

There is no single factor that can be used to compare all six models. The selection method will be the model with the lowest AIC and the Gamma model meets this criteria. The estimates for the model are:

| | |
|---|---|
| **Intercept** | 1.5005 |
| **VolatileAcidity** | -0.0105 |
| **AcidIndex** | -0.0158 |
| **IMP_STARS** | 0.0942 |
| **IMP_Density** | -0.3057 |
| **IMP_Alcohol** | 0.007 |
| **IMP_LabelAppeal** | 0.2316 |
| **IMP_Chlorides** | -0.0227 |
| **M_STARS** | -0.1594 |


# 7 Business Recommendations based on the data and analysis:

#1: Four of the factors are ***negatively*** related to purchase of wine in our stores. Thus we should look for wines that are low in Volatile Acidity, Acid Index, Density, and Chlorides. The biggest bang for our buck with this group of factors is to find wines that are lowest in density, that is more than 15 times bigger than each of the other negative factors. In short, it is in our best interest to stock Lite (or Light) Wines. These are much more likely to sell than any other factor out of this data set.

#2: The strongest ***positive*** factor is label appeal. Thus it is in our best interest to stock wines that have a strong label appeal. We can use analytics to determine which wines have strong label appeal. There are excellent survey methods that we can use both in and out of our store, I'm very interested in discussing these possibilities with management.

#3: Stock more wines that have at least two stars. These wines always sell at least one bottle out of out stock, and usually a lot more. No other single factor is predictive of wine sales, but this one is very good.

#4: Stock fewer wines with one star, and eliminate wines with zero stars. The analysis shows these wines do not sell very well, and I would predict we send many of them back. Let's prevent this problem from the beginning.

#4: More analytics and more data. We would do well to get our customer's feedback (maybe using a survey on the receipt?), adding data such as time of day and weather to the next data set, and looking for other factors that impact our sales–either positively or negatively.

#5: Look at deals for three bottles of wine. Our data shows the most common purchase is three bottles, so let's create incentives for everyone to purchase three bottles (or more) of wine.

#6: Look at our pricing. This is not in our data, but research has consistently shown that price is the biggest factor in wine purchase decisions. For example, here is a summary from a report on wine sales (emphasis mine):

> Food|beverage retailing practices have placed increasing emphasis on shelf management. While ***price is the major economic factor influencing sales of individual products***, other nonprice factors such as number of facings, shelf height, and season of the year, are also paramount to the sales of products.[1]

#7: Do research on educating our customers about wine. It is predicted that a customer educated about wine will make more purchases.

# Conclusion:

This report investigates data about wine sales in our stores. The data show that certain factors were favorable to wine sales, other factors were negatively related to wine sales. Business recommendations can be implemented starting today, including the use of analytics, to help our customers have the highest satisfaction with our store, our service, and the wine we offer. I happily offer my services as a data scientist to help achieve these goals. The future is bright, I look forward to even better days ahead.

1. https://ideas.repec.org/a/wly/agribz/v9y1993i6p595-603.html