# DATA 621: BUSINESS ANALYTICS AND DATA MINING HOMEWORK#5 Assignment Requirements

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited November 22, 2023

## Contents

## 1 Overview

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

| VARIABLE NAME | DEFINITION | THEORETICAL EFFECT |
| --- | --- | --- |
| INDEX | Identification Variable (do not use) | None |
| TARGET | Number of Cases Purchased | None |
| | | |
| | | |
| AcidIndex | Proprietary method of testing total acidity of wine by using a weighted average | |
| Alcohol | Alcohol Content | |
| Chlorides | Chloride content of wine | |
| CitricAcid | Citric Acid Content | |
| Density | Density of Wine | |
| FixedAcidity | Fixed Acidity of Wine | |
| FreeSulfurDioxide | Sulfur Dioxide content of wine | |
| LabelAppeal | Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design. | Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. |
| ResidualSugar | Residual Sugar of wine | |
| STARS | Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor | A high number of stars suggests high sales |
| Sulphates | Sulfate conten of wine | |
| TotalSulfurDioxide | Total Sulfur Dioxide of Wine | |
| VolatileAcidity | Volatile Acid content of wine | |
| pH | pH of wine | |

## 1.1 Deliverables

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned predictions (number of cases of wine sold) for the evaluation data set.
- Include your R statistical programming code in an Appendix.

## 1.2 Write Up:

### 1.2.1 1. DATA EXPLORATION (25 Points)

Describe the size and the variables in the wine training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

### 1.2.2 2. DATA PREPARATION (25 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

  a. Fix missing values (maybe with a Mean or Median value)
  b. Create flags to suggest if a variable was missing
  c. Transform data by putting it into buckets
  d. Mathematical transforms such as log or square root (or use Box-Cox)
  e. Combine variables (such as ratios or adding or multiplying) to create new variables

### 1.2.3 3. BUILD MODELS (25 Points)

Using the training data set, build at least two different poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different variables (or the same variables with different transformations). Sometimes poisson and negative binomial regression models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models. Although not covered in class, you may also want to consider building zero-inflated poisson and negative binomial regression models. You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say "pH seems to have a major positive impact in my poisson regression model, but a negative effect in my multiple linear regression model". Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

### 1.2.4 4. SELECT MODELS (25 Points)

Decide on the criteria for selecting the best count regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure to explain how you can make inferences from the model, and discuss other relevant model output. If you like the multiple linear regression model the best, please say why. However, you must select a count regression model for model deployment. Using the training data set, evaluate the performance of the count regression model. Make predictions using the evaluation data set.

# 2 Import Data

```
df_wine_eval <-
  read.csv(paste0(url_git,"wine-evaluation-data.csv"))

head(df_wine_eval)
```

```
##   IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
```

```
## 1  3      NA          5.4          -0.860        0.27       -10.7      0.092
## 2  9      NA         12.4           0.385       -0.76       -19.7      1.169
## 3 10      NA          7.2           1.750        0.17       -33.0      0.065
## 4 18      NA          6.2           0.100        1.80         1.0     -0.179
## 5 21      NA         11.4           0.210        0.28         1.2      0.038
## 6 30      NA         17.6           0.040       -1.15         1.4      0.535
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                23                398 0.98527 5.02      0.64   12.30
## 2               -37                 68 0.99048 3.37      1.09   16.00
## 3                 9                 76 1.04641 4.61      0.68    8.55
## 4               104                 89 0.98877 3.20      2.11   12.30
## 5                70                 53 1.02899 2.54     -0.07    4.80
## 6              -250                140 0.95028 3.06     -0.02   11.40
##   LabelAppeal AcidIndex STARS
## 1          -1         6    NA
## 2           0         6     2
## 3           0         8     1
## 4          -1         8     1
## 5           0        10    NA
## 6           1         8     4
```

```
df_wine_train <-
  read.csv(paste0(url_git,"wine-training-data.csv"))
head(df_wine_train)
```

```
##   INDEX TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1     1      3          3.2           1.160      -0.98          54.2    -0.567
## 2     2      3          4.5           0.160      -0.81          26.1    -0.425
## 3     4      5          7.1           2.640      -0.88          14.8     0.037
## 4     5      3          5.7           0.385       0.04          18.8    -0.425
## 5     6      4          8.0           0.330      -1.26           9.4        NA
## 6     7      0         11.3           0.320       0.59           2.2     0.556
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                NA                268 0.99280 3.33     -0.59     9.9
## 2                15               -327 1.02792 3.38      0.70      NA
## 3               214                142 0.99518 3.12      0.48    22.0
## 4                22                115 0.99640 2.24      1.83     6.2
## 5              -167                108 0.99457 3.12      1.77    13.7
## 6               -37                 15 0.99940 3.20      1.29    15.4
##   LabelAppeal AcidIndex STARS
## 1           0         8     2
## 2          -1         7     3
## 3          -1         8     3
## 4          -1         6     1
## 5           0         9     2
## 6           0        11    NA
```