DATA 621 Homework 5

Critical Thinking Group 1

November 22, 2021

Contents

Prepared for:

Prof. Dr. Nasrin Khansari

City University of New York, School of Professional Studies - Data 621

DATA 621 – Business Analytics and Data Mining

Home Work 5

Prepared by:

Critical Thinking Group 1

Vic Chan

Gehad Gad

Evan McLaughlin

Bruno de Melo

Anjal Hussan

Zhouxin Shi

Sie Siong Wong

Introduction

Problem

Our goal is to explore, analyze and model a dataset 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.

The 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.

1. DATA EXPLORATION

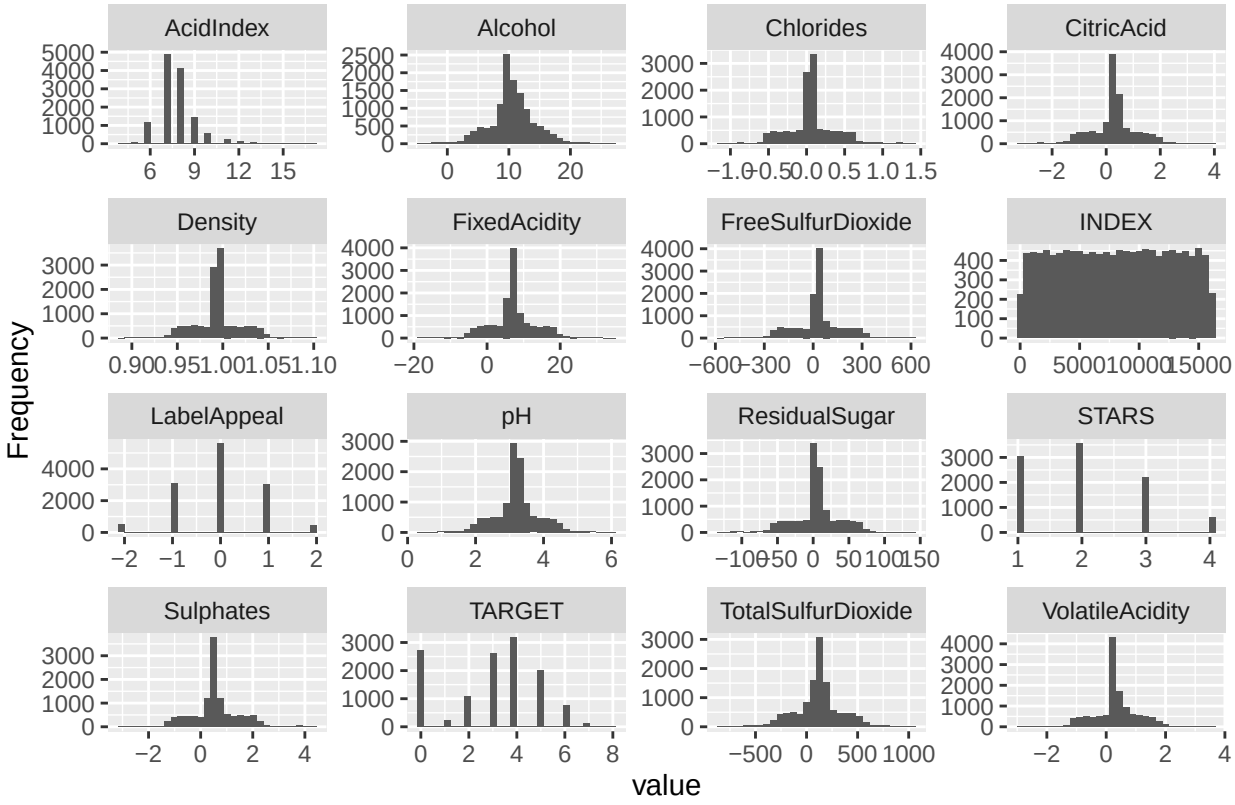
Below we'll display a few basic EDA techniques to gain insight into our wine dataset.

Basic Statistics

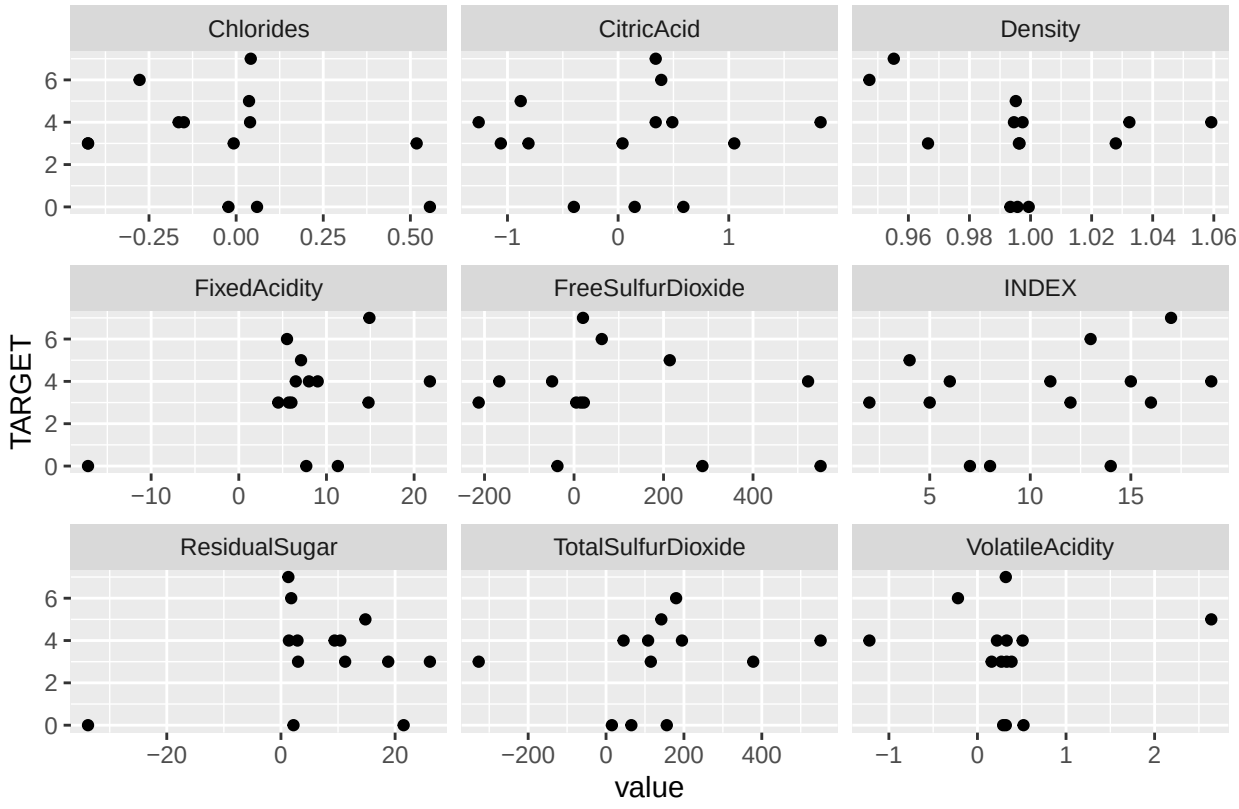
The data is 1.3 Mb in size. There are 12,795 rows and 15 columns (features). Of all 15 columns, 0 are discrete, 15 are continuous, and 0 are all missing. There are 8,200 missing values out of 191,925 data points.

```
##                         n    mean      sd  median     min      max  skew
## INDEX               12795 8069.98 4656.91 8110.00    1.00 16129.00  0.00
## TARGET              12795    3.03    1.93    3.00    0.00     8.00 -0.33
## FixedAcidity        12795    7.08    6.32    6.90  -18.10    34.40 -0.02
## VolatileAcidity     12795    0.32    0.78    0.28   -2.79     3.68  0.02
## CitricAcid          12795    0.31    0.86    0.31   -3.24     3.86 -0.05
## ResidualSugar       12179    5.42   33.75    3.90 -127.80   141.15 -0.05
## Chlorides           12157    0.05    0.32    0.05   -1.17     1.35  0.03
## FreeSulfurDioxide   12148   30.85  148.71   30.00 -555.00   623.00  0.01
## TotalSulfurDioxide 12113  120.71  231.91  123.00 -823.00  1057.00 -0.01
## Density             12795    0.99    0.03    0.99    0.89     1.10 -0.02
## pH                  12400    3.21    0.68    3.20    0.48     6.13  0.04
## Sulphates           11585    0.53    0.93    0.50   -3.13     4.24  0.01
## Alcohol             12142   10.49    3.73   10.40   -4.70    26.50 -0.03
## LabelAppeal         12795   -0.01    0.89    0.00   -2.00     2.00  0.01
## AcidIndex           12795    7.77    1.32    8.00    4.00    17.00  1.65
##                    kurtosis
## INDEX                 -1.20
## TARGET                -0.88
## FixedAcidity           1.67
## VolatileAcidity        1.83
## CitricAcid             1.84
## ResidualSugar          1.88
## Chlorides              1.79
## FreeSulfurDioxide      1.84
## TotalSulfurDioxide     1.67
## Density                1.90
## pH                     1.65
## Sulphates              1.75
## Alcohol                1.54
## LabelAppeal           -0.26
## AcidIndex              5.19
```

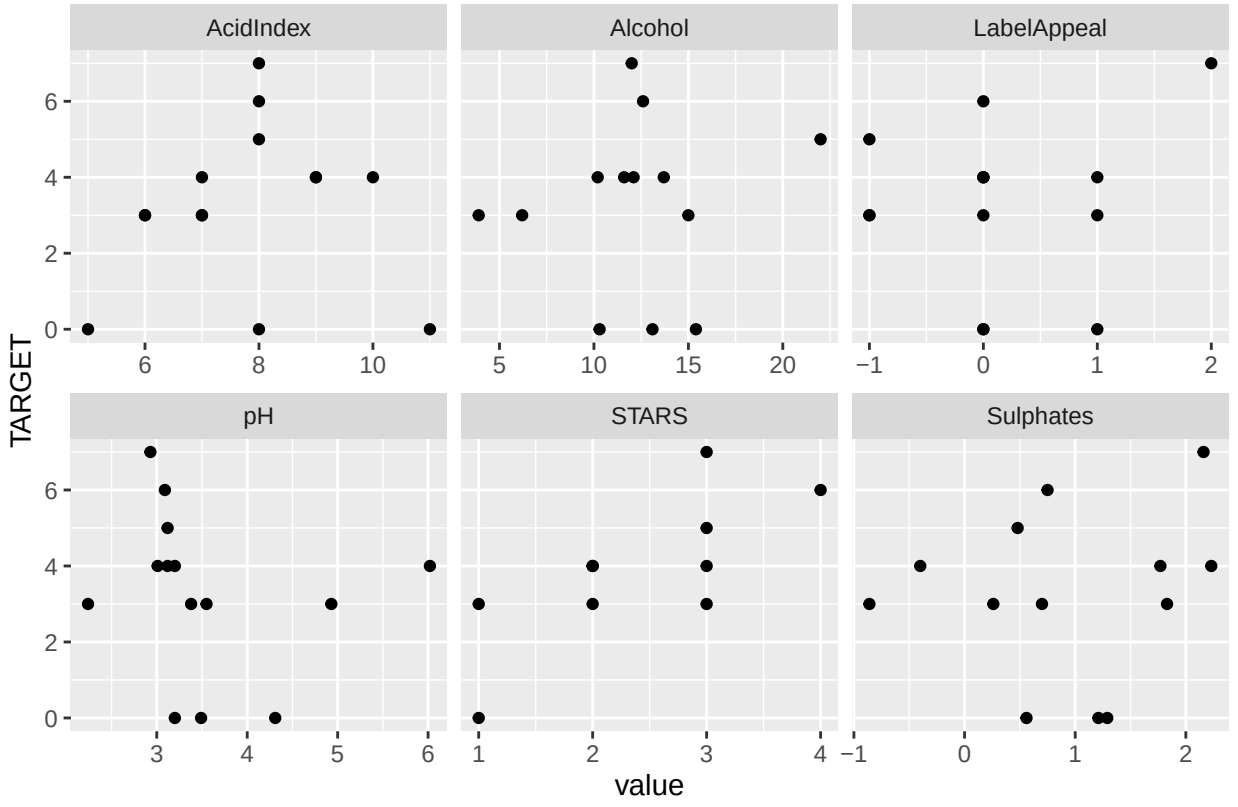
Histogram of Variables



Relationship of Predictors to Target



Page 1



Page 2