Data 621 - Homework 5

Group 4 Esteban Aramayo, Dmitriy Burtsev, Ian Costello, & Layla Quinones

12/6/2021

Overview

In this homework assignment, we explore, analyze and model a data set containing approximately 12,000 records representing commercially available wines. Each record has a target variable representing the number of cases purchased. Along with the target variable, there are fourteen predictor variables we will use to construct a count regression model. This model will seek to predict the number of cases that will be sold given certain properties of wine.

Libraries Used

We use the standard libraries such as `tidyverse`, `ggplot2`, and `caret`.

Data Exploration

As usual, our data are stored on GitHub at our team's main repository for easy access across team members **(Code Appendix 1.2)**. With our initial glimpse of the data, we know that all our data set is coded correctly as either doubles or integers. **(Code Appendix 1.3)**.

```
## Rows: 12,795
## Columns: 16
## $ ï..INDEX           <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19~
## $ TARGET             <int> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4, 0, 0, ~
## $ FixedAcidity       <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14.8, 5.5,~
## $ VolatileAcidity    <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0.290, -1~
## $ CitricAcid         <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0.40, 0.34~
## $ ResidualSugar      <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21.50, 1.40~
## $ Chlorides          <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0.060, 0.~
## $ FreeSulfurDioxide  <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213, 62, 551~
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA, 180, 65~
## $ Density            <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457, 0.9994~
## $ pH                 <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3.20, 4.9~
## $ Sulphates          <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, NA, 0.26~
## $ Alcohol            <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6, 15.0, 1~
## $ LabelAppeal        <int> 0, -1, -1, -1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 2, 0, 0, ~
## $ AcidIndex          <int> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8, 9, 8, 9~
## $ STARS              <int> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3, NA, NA~
```

Missing Values

Looking at the missing values for the data set, `STARS` has over 3,000 missing values. This makes sense since a team of experts can't realistically rate every bottle of wine. For the others it may make sense to impute the other missing values. For the `STARS` variable we will derive one more variable `STARSRating`, indicating whether a rating was conducted (no matter the score) or not.

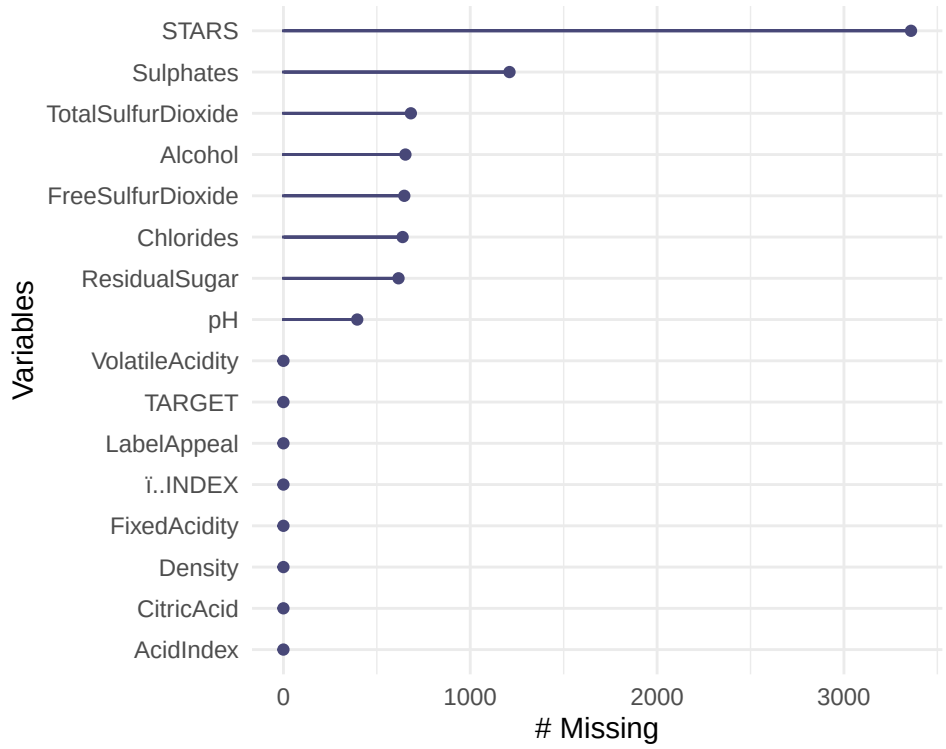


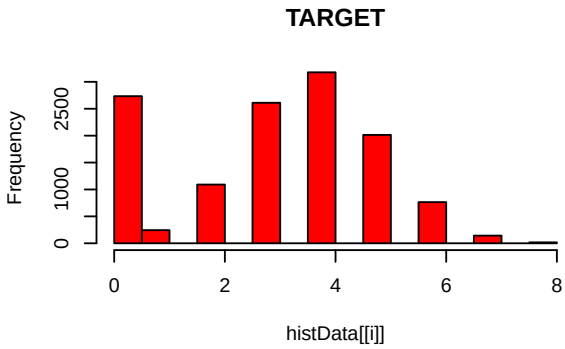
Figure 1. Plot of Missing Values

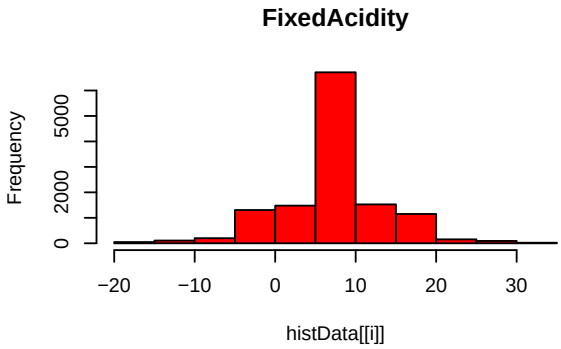
Summary Statistics

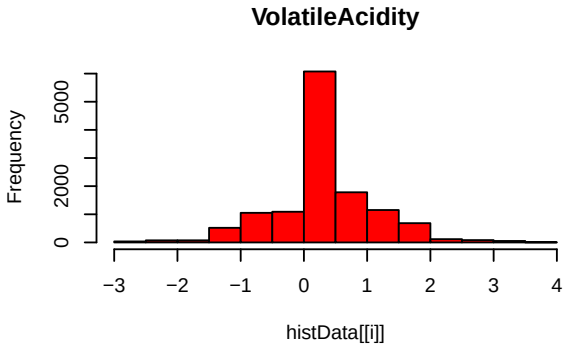
```
##                    vars     n    mean      sd  median trimmed     mad     min
## ï..INDEX              1 12795 8069.98 4656.91 8110.00 8071.03 5977.84    1.00
## TARGET                2 12795    3.03    1.93    3.00    3.05    1.48    0.00
## FixedAcidity          3 12795    7.08    6.32    6.90    7.07    3.26  -18.10
## VolatileAcidity       4 12795    0.32    0.78    0.28    0.32    0.43   -2.79
## CitricAcid            5 12795    0.31    0.86    0.31    0.31    0.42   -3.24
## ResidualSugar         6 12179    5.42   33.75    3.90    5.58   15.72 -127.80
## Chlorides             7 12157    0.05    0.32    0.05    0.05    0.13   -1.17
## FreeSulfurDioxide     8 12148   30.85  148.71   30.00   30.93   56.34 -555.00
## TotalSulfurDioxide    9 12113  120.71  231.91  123.00  120.89  134.92 -823.00
## Density              10 12795    0.99    0.03    0.99    0.99    0.01    0.89
## pH                   11 12400    3.21    0.68    3.20    3.21    0.39    0.48
## Sulphates            12 11585    0.53    0.93    0.50    0.53    0.44   -3.13
## Alcohol              13 12142   10.49    3.73   10.40   10.50    2.37   -4.70
## LabelAppeal          14 12795   -0.01    0.89    0.00   -0.01    1.48   -2.00
## AcidIndex            15 12795    7.77    1.32    8.00    7.64    1.48    4.00
## STARS                16  9436    2.04    0.90    2.00    1.97    1.48    1.00
## STARSRating          17 12795    0.74    0.44    1.00    0.80    0.00    0.00
```

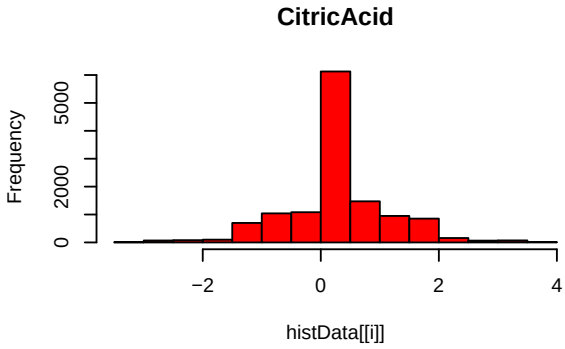
	max	range	skew	kurtosis	se
ï..INDEX	16129.00	16128.00	0.00	-1.20	41.17
TARGET	8.00	8.00	-0.33	-0.88	0.02
FixedAcidity	34.40	52.50	-0.02	1.67	0.06
VolatileAcidity	3.68	6.47	0.02	1.83	0.01
CitricAcid	3.86	7.10	-0.05	1.84	0.01
ResidualSugar	141.15	268.95	-0.05	1.88	0.31
Chlorides	1.35	2.52	0.03	1.79	0.00
FreeSulfurDioxide	623.00	1178.00	0.01	1.84	1.35
TotalSulfurDioxide	1057.00	1880.00	-0.01	1.67	2.11
Density	1.10	0.21	-0.02	1.90	0.00
pH	6.13	5.65	0.04	1.65	0.01
Sulphates	4.24	7.37	0.01	1.75	0.01
Alcohol	26.50	31.20	-0.03	1.54	0.03
LabelAppeal	2.00	4.00	0.01	-0.26	0.01
AcidIndex	17.00	13.00	1.65	5.19	0.01
STARS	4.00	3.00	0.45	-0.69	0.01
STARSRating	1.00	1.00	-1.08	-0.84	0.00

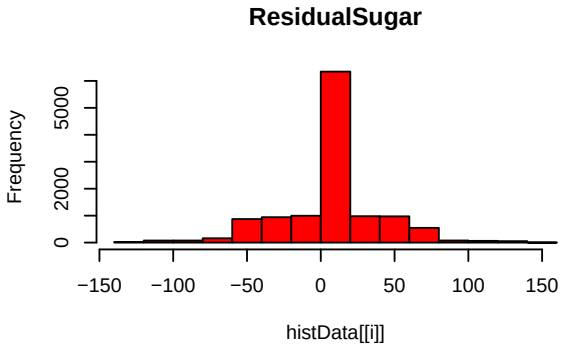
Table 1. Summary stats and description of data set

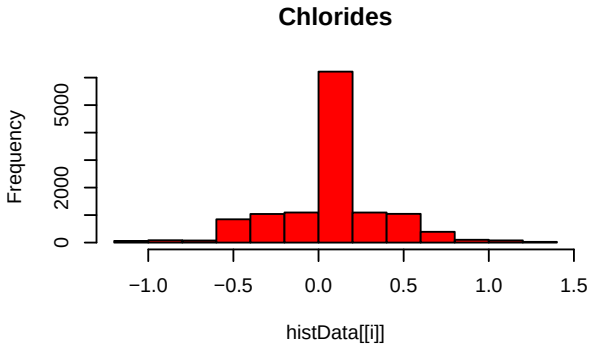

TARGET
Frequency
0
1000
2500
0
2
4
6
8
histData[[i]]

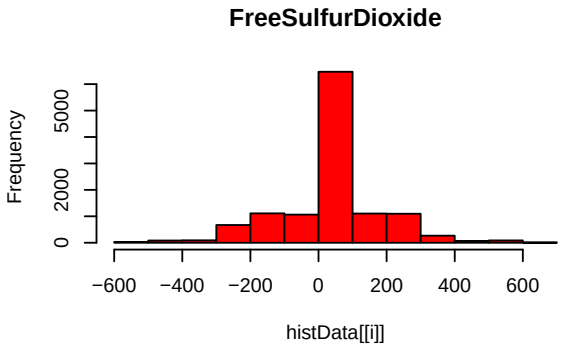

FixedAcidity
Frequency
0
2000
5000
−20
−10
0
10
20
30
histData[[i]]

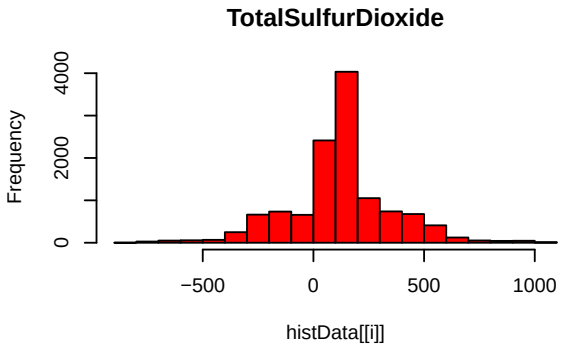

VolatileAcidity
Frequency
0
2000
5000
−3
−2
−1
0
1
2
3
4
histData[[i]]

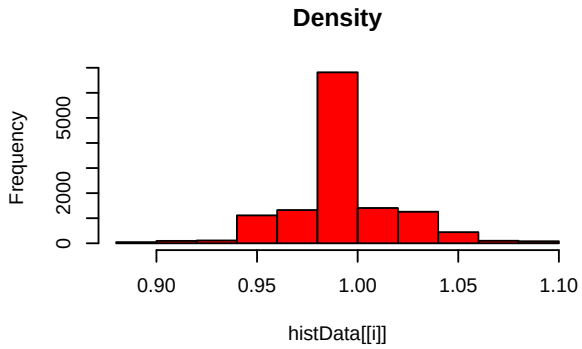

CitricAcid
Frequency
0
2000
5000
−2
0
2
4
histData[[i]]

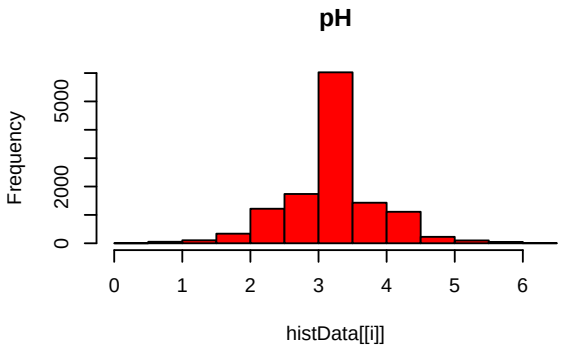

ResidualSugar
Frequency
0
2000
5000
−150
−100
−50
0
50
100
150
histData[[i]]

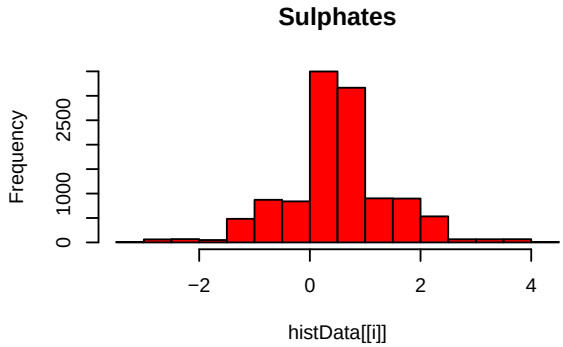

Chlorides
Frequency
0
2000
5000
−1.0
−0.5
0.0
0.5
1.0
1.5
histData[[i]]

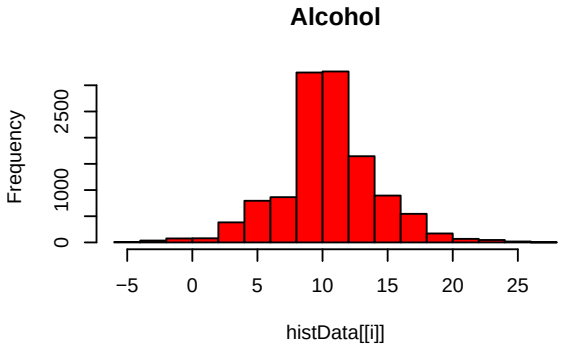

FreeSulfurDioxide
Frequency
0
2000
5000
−600
−400
−200
0
200
400
600
histData[[i]]

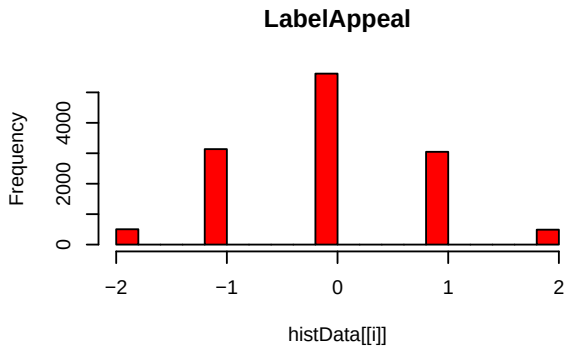

TotalSulfurDioxide
Frequency
0
2000
4000
−500
0
500
1000
histData[[i]]

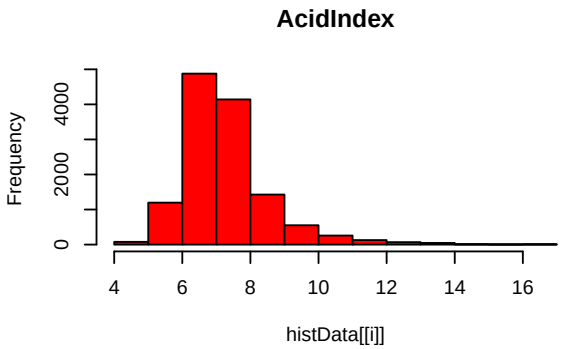

Density
Frequency
0
2000
5000
0.90
0.95
1.00
1.05
1.10
histData[[i]]

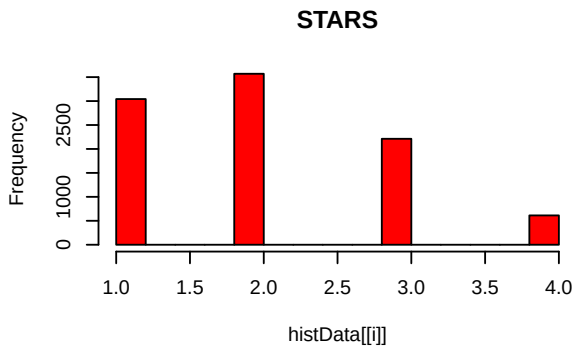

pH
Frequency
0
2000
5000
0
1
2
3
4
5
6
histData[[i]]

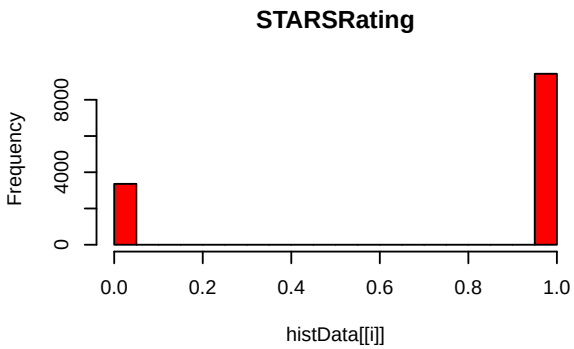

Sulphates
Frequency
0
1000
2500
−2
0
2
4
histData[[i]]


Alcohol
Frequency
0
1000
2500
−5
0
5
10
15
20
25
histData[[i]]


LabelAppeal
Frequency
0
2000
4000
−2
−1
0
1
2
histData[[i]]


AcidIndex
Frequency
0
2000
4000
4
6
8
10
12
14
16
histData[[i]]


STARS
Frequency
0
1000
2500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
histData[[i]]


STARSRating
Frequency
0
4000
8000
0.0
0.2
0.4
0.6
0.8
1.0
histData[[i]]