Data Analytics Assignment 7: 2 pm Session
Shiwei Huang

**1. QUESTION ONE**
**Exploratory Data Analysis (3%) Explore the statistical aspects of both datasets. Analyze the distributions and provide summaries of the relevant statistics. Perform any cleaning, transformations, interpolations, smoothing, outlier detection/ removal, etc. required on the data. Include figures and descriptions of this exploration and a short description of what you concluded (e.g. nature of distribution, indication of suitable model approaches you would try, etc.). Min.1 page text + graphics (required).**

Dataset1: Red wine Quality Data
Dataset2: White wine Quality Data

I choose to perform the wine quality data set for the data analysis project. For the wine quality data set, the characteristics is multivariate; and the attribute characteristics is real. Moreover, it is better to perform regression and classification models on this set of data.

Regarding the characteristics of the data sets, they are related to red and white variants of the Portuguese "Vinho Verde" wine. According to the description, these datasets can be viewed as classification or regression tasks. There are twelve attributes related to the dataset which is:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. Quality

pca.R × | hw3.R × | titanic.R × | lab6.R × | a7.R* × | winered × | lab7.R ×

Filter

| | fixed.acidity.volatile.acidity.citric.acid.residual.sugar.chlorides.free.sulfur.dioxide.total.sulfur.dioxide.density.pH.sulphates.alcohol.quality |
|---|---|
| 1 | 7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5 |
| 2 | 7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5 |
| 3 | 7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5 |
| 4 | 11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6 |
| 5 | 7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5 |
| 6 | 7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5 |
| 7 | 7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5 |
| 8 | 7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7 |
| 9 | 7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7 |
| 10 | 7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5 |
| 11 | 6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5 |
| 12 | 7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5 |
| 13 | 5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5 |

Showing 1 to 14 of 1,599 entries, 1 total columns

Similarly with the red wine data set, I also did an summary analysis on the white whine data set. These are the summary based on the dataset attributes:

a7.R × | winered × | a7b.R* × | winewhite ×

Filter

| | fixed.acidity.volatile.acidity.citric.acid.residual.sugar.chlorides.free.sulfur.dioxide.total.sulfur.dioxide.density.pH.sulphates.alcohol.quality |
|---|---|
| 1 | 7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6 |
| 2 | 6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6 |
| 3 | 8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6 |
| 4 | 7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6 |
| 5 | 7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6 |
| 6 | 8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6 |
| 7 | 6.2;0.32;0.16;7;0.045;30;136;0.9949;3.18;0.47;9.6;6 |
| 8 | 7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6 |
| 9 | 6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6 |
| 10 | 8.1;0.22;0.43;1.5;0.044;28;129;0.9938;3.22;0.45;11;6 |
| 11 | 8.1;0.27;0.41;1.45;0.033;11;63;0.9908;2.99;0.56;12;5 |
| 12 | 8.6;0.23;0.4;4.2;0.035;17;109;0.9947;3.14;0.53;9.7;5 |
| 13 | 7.9;0.18;0.37;1.2;0.04;16;75;0.992;3.18;0.63;10.8;5 |

Showing 1 to 14 of 4,898 entries, 1 total columns

Console | Terminal × | Jobs ×

~/Desktop/DataAnalyticsSpring2020/lab4/

```
[1] 1
> View(winered)
> ncol(winered)
[1] 1
>
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
> read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-white.csv")->winewhite
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
>
> View(winewhite)
> ncol(winewhite)
[1] 1
>
```

I separated the data sets from one column to multiple columns:
The following is the wine red analysis:

a7.R* | winered | a7b.R* | winewhite

Filter

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 34 | 0.9978 |
| 2 | 7.8 | 0.880 | 0.00 | 2.60 | 0.098 | 25 | 67 | 0.9968 |
| 3 | 7.8 | 0.760 | 0.04 | 2.30 | 0.092 | 15 | 54 | 0.9970 |
| 4 | 11.2 | 0.280 | 0.56 | 1.90 | 0.075 | 17 | 60 | 0.9980 |
| 5 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 34 | 0.9978 |
| 6 | 7.4 | 0.660 | 0.00 | 1.80 | 0.075 | 13 | 40 | 0.9978 |
| 7 | 7.9 | 0.600 | 0.06 | 1.60 | 0.069 | 15 | 59 | 0.9964 |
| 8 | 7.3 | 0.650 | 0.00 | 1.20 | 0.065 | 15 | 21 | 0.9946 |
| 9 | 7.8 | 0.580 | 0.02 | 2.00 | 0.073 | 9 | 18 | 0.9968 |
| 10 | 7.5 | 0.500 | 0.36 | 6.10 | 0.071 | 17 | 102 | 0.9978 |
| 11 | 6.7 | 0.580 | 0.08 | 1.80 | 0.097 | 15 | 65 | 0.9959 |
| 12 | 7.5 | 0.500 | 0.36 | 6.10 | 0.071 | 17 | 102 | 0.9978 |
| 13 | 5.6 | 0.615 | 0.00 | 1.60 | 0.089 | 16 | 59 | 0.9943 |
| 14 | 7.8 | 0.610 | 0.29 | 1.60 | 0.114 | 9 | 29 | 0.9974 |

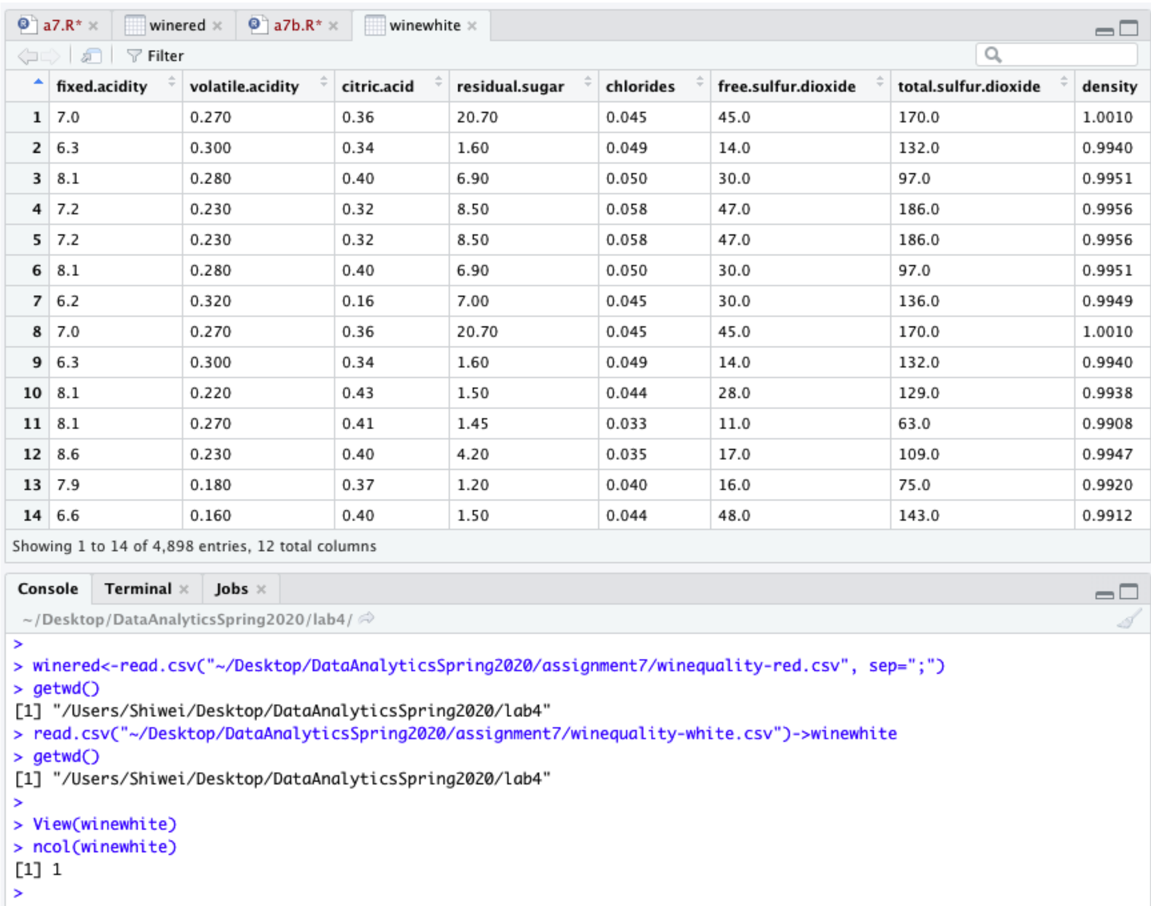
Showing 1 to 14 of 1,599 entries, 12 total columns

Console | Terminal | Jobs

~/Desktop/DataAnalyticsSpring2020/lab4/

```
> ncol(winered)
[1] 1
> winered<-read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-red.csv", sep=";")
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
> read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-red.csv")->winered
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
>
> View(winered)
> ncol(winered)
[1] 1
>
```

The following the wine white dataset:



a7.R* | winered | a7b.R* | winewhite

Filter

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.0 | 0.270 | 0.36 | 20.70 | 0.045 | 45.0 | 170.0 | 1.0010 |
| 2 | 6.3 | 0.300 | 0.34 | 1.60 | 0.049 | 14.0 | 132.0 | 0.9940 |
| 3 | 8.1 | 0.280 | 0.40 | 6.90 | 0.050 | 30.0 | 97.0 | 0.9951 |
| 4 | 7.2 | 0.230 | 0.32 | 8.50 | 0.058 | 47.0 | 186.0 | 0.9956 |
| 5 | 7.2 | 0.230 | 0.32 | 8.50 | 0.058 | 47.0 | 186.0 | 0.9956 |
| 6 | 8.1 | 0.280 | 0.40 | 6.90 | 0.050 | 30.0 | 97.0 | 0.9951 |
| 7 | 6.2 | 0.320 | 0.16 | 7.00 | 0.045 | 30.0 | 136.0 | 0.9949 |
| 8 | 7.0 | 0.270 | 0.36 | 20.70 | 0.045 | 45.0 | 170.0 | 1.0010 |
| 9 | 6.3 | 0.300 | 0.34 | 1.60 | 0.049 | 14.0 | 132.0 | 0.9940 |
| 10 | 8.1 | 0.220 | 0.43 | 1.50 | 0.044 | 28.0 | 129.0 | 0.9938 |
| 11 | 8.1 | 0.270 | 0.41 | 1.45 | 0.033 | 11.0 | 63.0 | 0.9908 |
| 12 | 8.6 | 0.230 | 0.40 | 4.20 | 0.035 | 17.0 | 109.0 | 0.9947 |
| 13 | 7.9 | 0.180 | 0.37 | 1.20 | 0.040 | 16.0 | 75.0 | 0.9920 |
| 14 | 6.6 | 0.160 | 0.40 | 1.50 | 0.044 | 48.0 | 143.0 | 0.9912 |

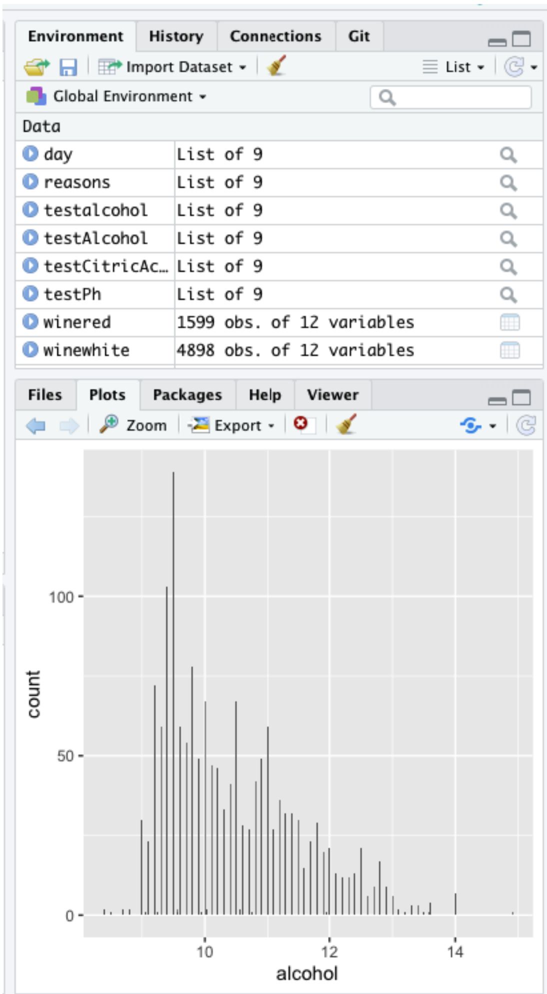
Showing 1 to 14 of 4,898 entries, 12 total columns

Console | Terminal | Jobs

~/Desktop/DataAnalyticsSpring2020/lab4/

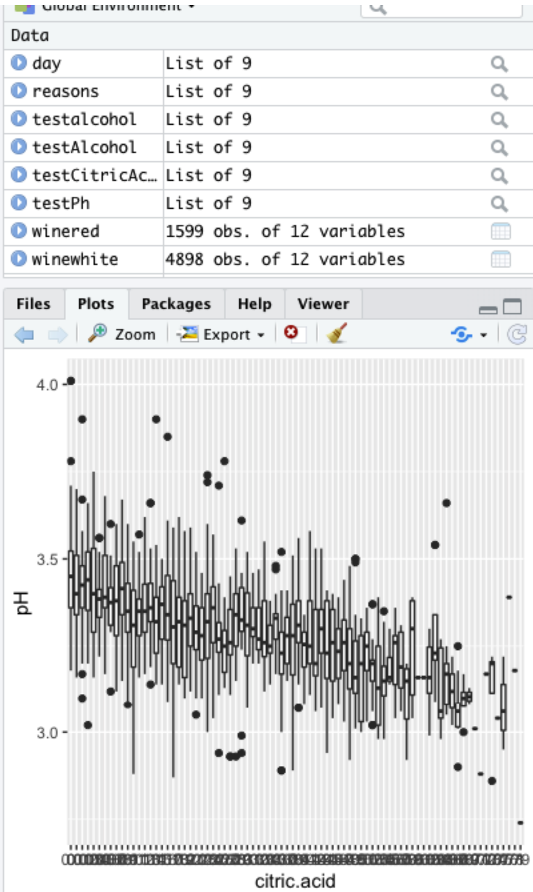
```
>
> winered<-read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-red.csv", sep=";")
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
> read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-white.csv")->winewhite
> getwd()
[1] "/Users/Shiwei/Desktop/DataAnalyticsSpring2020/lab4"
>
> View(winewhite)
> ncol(winewhite)
[1] 1
>
```

I made a ggplot on red wine with alcohol count, this is my finding:

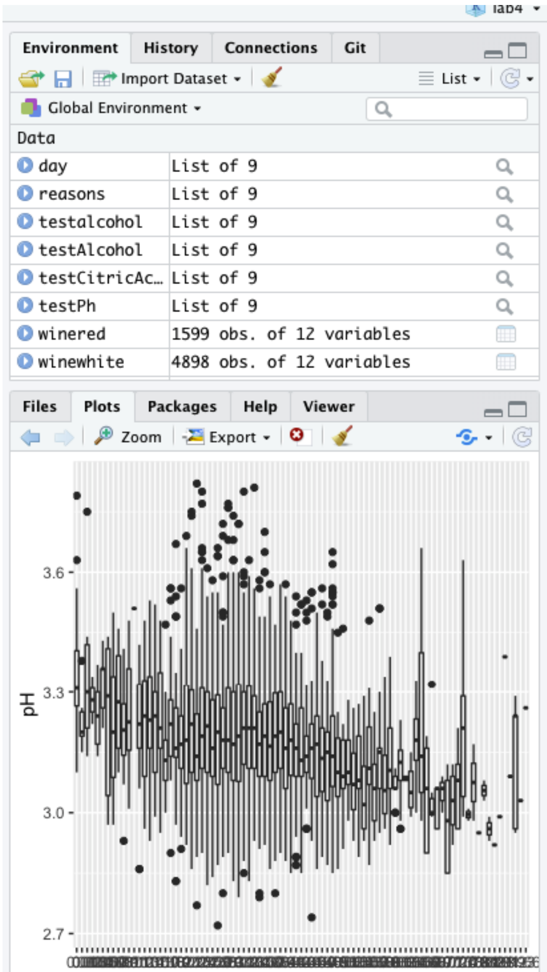



I made a ggplot on citric.acid with Ph level for the red wine:

```
ggplot(winered, aes_string(y=winered$pH, x=as.factor(winered$citric.acid)),)+
  geom_boxplot() + xlab('citric.acid') + ylab('pH')
```




This is the ggplot on pH data with citric.acid:




In addition, I wanted to understand what is the range for pH level for red wine, citric acid range for red wine, volatile acidity on red wine and chlorides range for red wine.

I performed analyses on the range function:
This is my finding:
> range(winered$pH)
[1] 2.74 4.01
> range(winered$pH)
[1] 2.74 4.01
> range(winered$citric.acid)
[1] 0 1
> range(winered$volatile.acidity)
[1] 0.12 1.58
> range(winered$chlorides)
[1] 0.012 0.611

This is the range for wine white pH level, citric acid, volatile acidity and chlorides:

> range(winewhite$pH)
[1] 2.72 3.82
> range(winewhite$citric.acid)
[1] 0.00 1.66
> range(winewhite$volatile.acidity)
[1] 0.08 1.10
> range(winewhite$chlorides)
[1] 0.009 0.346

**2. Question Two: Model Development, Validation, Optimization and Tuning (14%) Choose two (4000-level*) or three (6000-level) or more different models (e.g. a model with a different set/ number of variables/ features in a regression, or classification, etc. does NOT count as a different model). Explain why you chose them. Construct the models, test/ validate them. Explain the validation approach. You can use any method(s) covered in the course. Include your code in your submission. Compare model results if applicable. Report the results of the model (fits, coefficients, graphs, trees, other measures of fit/ importance, etc.), predictors, and summary statistics. Min. 4 pages of text + graphics (required). * 4000-level will receive extra credit for 6000-level responses.**

Dataset 1: red wine
Model 1: Multivariate linear regression
Multivariable linear regression on redwine:
Reasons I chose this model is I want to understand the dependency of the variables:

Result:

```
Call:
lm(formula = pH ~ ., data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35632 -0.05285 -0.00043  0.05227  0.46412 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -6.059e+01  2.771e+00 -21.868  < 2e-16 ***
fixed.acidity        -9.543e-02  2.756e-03 -34.621  < 2e-16 ***
volatile.acidity      1.304e-02  1.903e-02   0.685 0.493288    
citric.acid          -3.780e-02  2.238e-02  -1.689 0.091448 .  
residual.sugar       -2.644e-02  2.268e-03 -11.660  < 2e-16 ***
chlorides            -4.880e-01  6.334e-02  -7.704 2.78e-14 ***
free.sulfur.dioxide   1.224e-03  3.325e-04   3.681 0.000243 ***
total.sulfur.dioxide -7.105e-04  1.110e-04  -6.398 2.26e-10 ***
density               6.435e+01  2.776e+00  23.181  < 2e-16 ***
sulphates            -4.537e-02  1.853e-02  -2.448 0.014496 *  
alcohol               7.337e-02  3.637e-03  20.175  < 2e-16 ***
quality              -9.584e-03  3.864e-03  -2.480 0.013266 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08618 on 1187 degrees of freedom
Multiple R-squared:  0.6866,	Adjusted R-squared:  0.6837 
F-statistic: 236.4 on 11 and 1187 DF,  p-value: < 2.2e-16
```

```
#testing first model:
library(caTools)
set.seed(500)
```

54:1 (Top Level) ÷

Console Terminal × Jobs ×

~/Desktop/DataAnalyticsSpring2020/lab4/

```
> preddata=predict(regr,newdata = testdata)
>
> preddata
       3        8       12       15       20       24       27       32       36       39       44 
3.402389 3.427479 3.272577 3.347229 3.215343 3.387249 3.332919 3.425302 3.427790 3.356566 3.326493 
      48       51       56       60       63       68       72       75       80       84       87 
3.212647 3.308245 3.411097 3.302229 3.362573 3.392124 3.321309 3.201657 3.337794 3.307623 3.309695 
      92       96       99      104      108      111      116      120      123      128      132 
3.309695 3.353035 3.384558 3.331780 3.287301 3.346399 3.301607 3.388495 3.424680 3.385184 3.395751 
     135      140      144      147      152      156      159      164      168      171      176 
3.380412 3.346399 3.407155 3.416489 2.979365 3.249559 3.425613 3.311977 3.420013 3.399177 3.418561 
     180      183      188      192      195      200      204      207      212      216      219 
3.366099 3.413379 3.378438 3.330845 3.337897 3.372740 3.282115 3.111661 3.313326 3.213893 3.350753 
     224      228      231      236      240      243      248      252      255      260      264 
3.313740 3.353037 3.419805 3.428723 3.364652 3.386213 3.353035 3.430589 3.430589 3.234213 3.339969 
     267      272      276      279      284      288      291      296      300      303      308 
3.413796 3.224050 3.406638 3.242715 3.297045 3.388285 3.394507 3.226957 3.398863 3.369314 3.250803 
     312      315      320      324      327      332      336      339      344      348      351 
3.325455 3.313219 3.365271 3.239607 3.133851 3.153549 3.182894 3.172835 3.229237 3.131777 3.325871 
     356      360      363      368      372      375      380      384      387      392      396 
3.416697 3.143181 3.130741 3.153862 3.270501 3.155001 3.268429 3.260133 3.315709 3.131880 3.098183 
     399      404      408      411      416      420      423      428      432      435      440 
3.162053 3.207051 3.142559 3.286055 3.313326 3.420015 3.316958 3.319029 3.274029 3.191497 3.347229 
     444      447      452      456      459      464      468      471      476      480      483 
3.217003 3.170553 3.203109 3.123691 3.159977 3.107517 3.234007 3.151475 3.316125 3.375120 3.176359 
     488      492      495      500      504      507      512      516      519      524      528 
3.263558 3.214307 3.338725 3.283569 3.237323 3.243129 3.289789 3.203630 3.231308 3.219491 3.220735 
     531      536      540      543      548      552      555      560      564      567      572 
3.344737 3.344737 3.099221 3.313948 3.225089 3.331675 3.204252 3.215137 3.226333 3.314881 3.336029 
     576      579      584      588      591      596      600      603      608      612      615 
```

White wine data:

~/Desktop/DataAnalyticsSpring2020/lab4/

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.98464 -0.06068  0.00197  0.06211  0.34412 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -1.064e+02  2.274e+00 -46.771  < 2e-16 ***
fixed.acidity        -1.308e-01  2.477e-03 -52.782  < 2e-16 ***
volatile.acidity     -9.510e-02  1.827e-02  -5.204 2.05e-07 ***
citric.acid          -6.865e-02  1.474e-02  -4.657 3.32e-06 ***
residual.sugar       -4.472e-02  9.041e-04 -49.458  < 2e-16 ***
chlorides            -8.038e-01  8.383e-02  -9.588  < 2e-16 ***
free.sulfur.dioxide   3.577e-04  1.331e-04   2.687  0.00724 ** 
total.sulfur.dioxide -2.699e-05  5.926e-05  -0.455  0.64879    
density               1.101e+02  2.278e+00  48.351  < 2e-16 ***
sulphates            -6.747e-02  1.564e-02  -4.314 1.65e-05 ***
alcohol               1.213e-01  3.120e-03  38.885  < 2e-16 ***
quality               1.286e-02  2.234e-03   5.759 9.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.102 on 3662 degrees of freedom
Multiple R-squared:  0.5382,	Adjusted R-squared:  0.5369 
F-statistic: 388.1 on 11 and 3662 DF,  p-value: < 2.2e-16

>
> regr=lm(formula=pH~volatile.acidity+citric.acid,data=traindata)
> summary(regr)
```

```
>
> preddata=predict(regr,newdata = testdata)
>
> preddata
       3        8       12       15       20       24       27       32       36       39       44
3.174807 3.183616 3.178360 3.120315 3.225316 3.199734 3.187772 3.196903 3.168733 3.179674 3.203970
      48       51       56       60       63       68       72       75       80       84       87
3.192386 3.199424 3.187062 3.188590 3.228119 3.231605 3.207805 3.196582 3.180627 3.199317 3.120422
      92       96       99      104      108      111      116      120      123      128      132
3.169443 3.198499 3.156974 3.173030 3.188483 3.154278 3.241226 3.155767 3.201556 3.193846 3.193846
     135      140      144      147      152      156      159      164      168      171      176
3.211961 3.176831 3.177756 3.209655 3.169765 3.173600 3.236325 3.173600 3.206558 3.161237 3.178360
     180      183      188      192      195      200      204      207      212      216      219
3.193136 3.174807 3.207805 3.174807 3.165997 3.208623 3.188376 3.199531 3.185962 3.208623 3.207201
     224      228      231      236      240      243      248      252      255      260      264
3.201449 3.193136 3.214014 3.182409 3.215407 3.185855 3.206812 3.214697 3.179674 3.173171 3.174096
     267      272      276      279      284      288      291      296      300      303      308
3.200310 3.218881 3.179178 3.184112 3.190790 3.180384 3.180384 3.197574 3.162377 3.246555 3.198674
     312      315      320      324      327      332      336      339      344      348      351
3.208261 3.209012 3.216191 3.166708 3.174203 3.189797 3.198606 3.176228 3.158502 3.179178 3.194275
     356      360      363      368      372      375      380      384      387      392      396
3.185144 3.189797 3.204291 3.209722 3.157899 3.207698 3.196971 3.201838 3.150229 3.185319 3.232208
     399      404      408      411      416      420      423      428      432      435      440
3.202549 3.193739 3.192278 3.178963 3.182905 3.196971 3.204077 3.153849 3.231283 3.186458 3.242720
     444      447      452      456      459      464      468      471      476      480      483
3.149653 3.201127 3.210862 3.197963 3.184608 3.210433 3.210433 3.107671 3.155378 3.193243 3.209333
     488      492      495      500      504      507      512      516      519      524      528
```

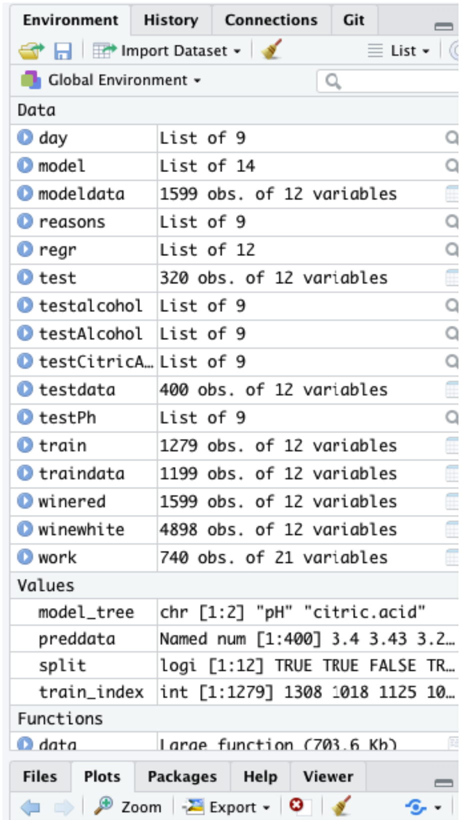
Model 2: decision tree model

Because the decision tree model can be used to solve regression and classification problems.

Redwine:

```
Global Environment
Data
day          List of 9
model        List of 12
modeldata    1599 obs. of 12 variables
reasons      List of 9
regr         List of 12
test         320 obs. of 12 variables
testalcohol  List of 9
testAlcohol  List of 9
testCitricA… List of 9
testdata     1224 obs. of 12 variables
testPh       List of 9
train        1279 obs. of 12 variables
traindata    3674 obs. of 12 variables
winered      1599 obs. of 12 variables
winewhite    4898 obs. of 12 variables
work         740 obs. of 21 variables
Values
model_tree   chr [1:2] "pH" "citric.acid"
preddata     Named num [1:1224] 3.17 3.18 3…
split        logi [1:12] TRUE TRUE FALSE TR…
train_index  int [1:1279] 1308 1018 1125 10…
Functions
data         Large function (703.6 Kb)
Files  Plots  Packages  Help  Viewer
```

Whitewine:




Model 3: random forest model

Random forest can handle the missing values and prune the data accuracy.

```r

#second model:
library(randomForest)
require(caTools)
modeldata<-winered
set.seed(1234)
train_index = sample(1:nrow(modeldata),0.8*nrow(modeldata))
train=modeldata[train_index,]
test=modeldata[-train_index,]


set.seed(1234)
model_tree(data)<-c("pH", "citric.acid")
head(data)
summary(model_tree)


#third model: decision tree
library(rpart)
model<-rpart(fixed.acidity~pH+citric.acid, data=traindata, method="class")
```

67:1 (Top Level) R Script

Console Terminal Jobs

~/Desktop/DataAnalyticsSpring2020/lab4/

```
function (..., list = character(), package = NULL, lib.loc = NULL,
    verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
{
    fileExt <- function(x) {
        db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
        ans <- sub(".*\\\\.", "", x)
> summary(model_tree)
   Length     Class      Mode 
        2 character character 
```

```r

#second model:
library(randomForest)
require(caTools)
modeldata<-winewhite
set.seed(1234)
train_index = sample(1:nrow(modeldata),0.8*nrow(modeldata))
train=modeldata[train_index,]
test=modeldata[-train_index,]


set.seed(1234)
model_tree(data)<-c("pH", "citric.acid")
head(data)
summary(model_tree)
```

89:1 (Top Level)

Console Terminal Jobs

~/Desktop/DataAnalyticsSpring2020/lab4/

```
> head(data)

1 function (..., list = character(), package = NULL, lib.loc = NULL,
2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
3 {
4     fileExt <- function(x) {
5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
6         ans <- sub(".*\\\\.", "", x)
> summary(model_tree)
   Length     Class      Mode
        2 character character
```

**3. Decisions (3%) Describe your conclusions in regard to the model fit, predictions and how well (or not) it could be used for decisions and why. Min. 1 page of text + graphics.**

My conclusions is that the dataset 1 and 3 provides a good summary on each variable. Dataset number 2 does not work well with the wine data.

winered | test.R* | work | a7.R* | a7b.R | winewhite

Filter

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dio |
|---|---|---|---|---|---|---|---|
| 1 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 34 |
| 2 | 7.8 | 0.880 | 0.00 | 2.60 | 0.098 | 25 | 67 |
| 3 | 7.8 | 0.760 | 0.04 | 2.30 | 0.092 | 15 | 54 |
| 4 | 11.2 | 0.280 | 0.56 | 1.90 | 0.075 | 17 | 60 |
| 5 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 34 |
| 6 | 7.4 | 0.660 | 0.00 | 1.80 | 0.075 | 13 | 40 |
| 7 | 7.9 | 0.600 | 0.06 | 1.60 | 0.069 | 15 | 59 |
| 8 | 7.3 | 0.650 | 0.00 | 1.20 | 0.065 | 15 | 21 |
| 9 | 7.8 | 0.580 | 0.02 | 2.00 | 0.073 | 9 | 18 |
| 10 | 7.5 | 0.500 | 0.36 | 6.10 | 0.071 | 17 | 102 |
| 11 | 6.7 | 0.580 | 0.08 | 1.80 | 0.097 | 15 | 65 |
| 12 | 7.5 | 0.500 | 0.36 | 6.10 | 0.071 | 17 | 102 |
| 13 | 5.6 | 0.615 | 0.00 | 1.60 | 0.089 | 16 | 59 |
| 14 | 7.8 | 0.610 | 0.29 | 1.60 | 0.114 | 9 | 29 |
| 15 | 8.9 | 0.620 | 0.18 | 3.80 | 0.176 | 52 | 145 |
| 16 | 8.9 | 0.620 | 0.19 | 3.90 | 0.170 | 51 | 148 |
| 17 | 8.5 | 0.280 | 0.56 | 1.80 | 0.092 | 35 | 103 |

Showing 1 to 18 of 1,599 entries, 12 total columns

I performed summary on winered and wine white:

This the summary on winered:

```
~/Desktop/DataAnalyticsSpring2020/lab4/
> summary(winered)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
       pH            sulphates          alcohol          quality
```

This is the summary on winewhite:

```
Console   Terminal   Jobs
~/Desktop/DataAnalyticsSpring2020/lab4/
> winewhite<-read.csv("~/Desktop/DataAnalyticsSpring2020/assignment7/winequality-
> summary(winewhite)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
 Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
 3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
 Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
   chlorides        free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
 1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
 Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
 Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
 3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
 Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
```

Doing the comparison, we could find the median summary on winered is 7.9 while the median on whitewhite is 6.8; in terms of the volatile acidity, the median for winered is 0.52 while the median for winewhite is 0.26; in terms of citric acid, the median for wine red is 0.26 while the median for wine white is 0.32; in terms of the residual sugar, wine red median is 2.2; wine white median is 5.2; for chlorides, the medium for winered is 0.079, the median for wine white is 0.043; in terms of free sulfur dioxide, the medium for wine red is 14 while the medium for wine white is 34; in terms of total sulfur dioxide, the medium for wine red is 38 while the median for wine white is 134; in terms of density, the median for wine red is 0.996 while wine white density is slightly less which is 0.9937.