# Appendix

## Appendix A (Histograms)





## Appendix B (Correlation Coefficients)

| **Pearson Correlation Coefficients, N = 6497 Prob > |r| under H0: Rho=0** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **fixed\_acidity** | **volatile\_acidity** | **citric\_acid** | **residual\_sugar** | **chlorides** | **free\_sulfur\_dioxide** |
| **fixed\_acidity** | 1.00000 | 0.21901 <.0001 | 0.32444 <.0001 | -0.11198 <.0001 | 0.29819 <.0001 | -0.28274 <.0001 |
| **volatile\_acidity** | 0.21901 <.0001 | 1.00000 | -0.37798 <.0001 | -0.19601 <.0001 | 0.37712 <.0001 | -0.35256 <.0001 |
| **citric\_acid** | 0.32444 <.0001 | -0.37798 <.0001 | 1.00000 | 0.14245 <.0001 | 0.03900 0.0017 | 0.13313 <.0001 |
| **residual\_sugar** | -0.11198 <.0001 | -0.19601 <.0001 | 0.14245 <.0001 | 1.00000 | -0.12894 <.0001 | 0.40287 <.0001 |
| **chlorides** | 0.29819 <.0001 | 0.37712 <.0001 | 0.03900 0.0017 | -0.12894 <.0001 | 1.00000 | -0.19504 <.0001 |
| **free\_sulfur\_dioxide** | -0.28274 <.0001 | -0.35256 <.0001 | 0.13313 <.0001 | 0.40287 <.0001 | -0.19504 <.0001 | 1.00000 |
| **total\_sulfur\_dioxide** | -0.32905 <.0001 | -0.41448 <.0001 | 0.19524 <.0001 | 0.49548 <.0001 | -0.27963 <.0001 | 0.72093 <.0001 |
| **density** | 0.45891 <.0001 | 0.27130 <.0001 | 0.09615 <.0001 | 0.55252 <.0001 | 0.36261 <.0001 | 0.02572 0.0382 |
| **pH** | -0.25270 <.0001 | 0.26145 <.0001 | -0.32981 <.0001 | -0.26732 <.0001 | 0.04471 0.0003 | -0.14585 <.0001 |
| **sulphates** | 0.29957 <.0001 | 0.22598 <.0001 | 0.05620 <.0001 | -0.18593 <.0001 | 0.39559 <.0001 | -0.18846 <.0001 |
| **alcohol** | -0.09545 <.0001 | -0.03764 0.0024 | -0.01049 0.3977 | -0.35941 <.0001 | -0.25692 <.0001 | -0.17984 <.0001 |
| **quality** | -0.07674 <.0001 | -0.26570 <.0001 | 0.08553 <.0001 | -0.03698 0.0029 | -0.20067 <.0001 | 0.05546 <.0001 |
| **grape** | -0.48674 <.0001 | -0.65304 <.0001 | 0.18740 <.0001 | 0.34882 <.0001 | -0.51268 <.0001 | 0.47164 <.0001 |

| **Pearson Correlation Coefficients, N = 6497 Prob > |r| under H0: Rho=0** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **total\_sulfur\_dioxide** | **density** | **pH** | **sulphates** | **alcohol** | **quality** | **grape** |
| **fixed\_acidity** | -0.32905 <.0001 | 0.45891 <.0001 | -0.25270 <.0001 | 0.29957 <.0001 | -0.09545 <.0001 | -0.07674 <.0001 | -0.48674 <.0001 |
| **volatile\_acidity** | -0.41448 <.0001 | 0.27130 <.0001 | 0.26145 <.00010.47164 | 0.22598 <.0001 | -0.03764 0.0024 | -0.26570 <.0001 | -0.65304 <.0001 |
| **citric\_acid** | 0.19524 <.0001 | 0.09615 <.0001 | -0.32981 <.0001 | 0.05620 <.0001 | -0.01049 0.3977 | 0.08553 <.0001 | 0.18740 <.0001 |
| **residual\_sugar** | 0.49548 <.0001 | 0.55252 <.0001 | -0.26732 <.0001 | -0.18593 <.0001 | -0.35941 <.0001 | -0.03698 0.0029 | 0.34882 <.0001 |
| **chlorides** | -0.27963 <.0001 | 0.36261 <.0001 | 0.04471 0.0003 | 0.39559 <.0001 | -0.25692 <.0001 | -0.20067 <.0001 | -0.51268 <.0001 |
| **free\_sulfur\_dioxide** | 0.72093 <.0001 | 0.02572 0.0382 | -0.14585 <.0001 | -0.18846 <.0001 | -0.17984 <.0001 | 0.05546 <.0001 | <.0001 |
| **total\_sulfur\_dioxide** | 1.00000 | 0.03239 0.0090 | -0.23841 <.0001 | -0.27573 <.0001 | -0.26574 <.0001 | -0.04139 0.0008 | 0.70036 <.0001 |
| **density** | 0.03239 0.0090 | 1.00000 | 0.01169 0.3463 | 0.25948 <.0001 | -0.68675 <.0001 | -0.30586 <.0001 | -0.39065 <.0001 |
| **pH** | -0.23841 <.0001 | 0.01169 0.3463 | 1.00000 | 0.19212 <.0001 | 0.12125 <.0001 | 0.01951 0.1159 | -0.32913 <.0001 |
| **sulphates** | -0.27573 <.0001 | 0.25948 <.0001 | 0.19212 <.0001 | 1.00000 | -0.00303 0.8071 | 0.03849 0.0019 | -0.48722 <.0001 |
| **alcohol** | -0.26574 <.0001 | -0.68675 <.0001 | 0.12125 <.0001 | -0.00303 0.8071 | 1.00000 | 0.44432 <.0001 | 0.03297 0.0079 |
| **quality** | -0.04139 0.0008 | -0.30586 <.0001 | 0.01951 0.1159 | 0.03849 0.0019 | 0.44432 <.0001 | 1.00000 | 0.11932 <.0001 |
| **grape** | 0.70036 <.0001 | -0.39065 <.0001 | -0.32913 <.0001 | -0.48722 <.0001 | 0.03297 0.0079 | 0.11932 <.0001 | 1.00000 |

## SAS Code – EDA

data wine;

infile '/folders/myfolders/Experimental Statistics I/winequality.csv' firstobs = 2 delimiter = ';';

input fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides free\_sulfur\_dioxide total\_sulfur\_dioxide density pH sulphates alcohol quality grape;

proc univariate data = wine;

histogram;

run;

/\*close to normal distribution, slight right skew. Enough levels to consider like a cont var./\*

/\*Included in write up\*/

proc sgscatter data = wine;

matrix total\_sulfur\_dioxide free\_sulfur\_dioxide grape alcohol density volatile\_acidity;

run;

proc corr data = wine plots(maxpoints=none)=scatter;

var fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides free\_sulfur\_dioxide total\_sulfur\_dioxide density pH sulphates alcohol quality grape;

run;

## Objective 1 – SAS Output

## Assumptions

## Model Selection

## SAS Code for Regressions

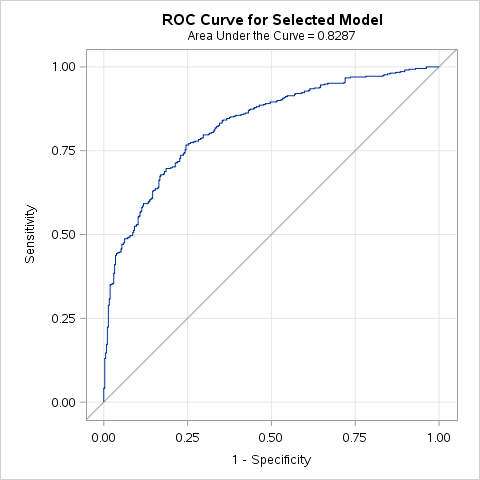
## Objective 2

## Logistic Regression 2

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 4.8489 | 8 | 0.7736 |
| **Model Fit Statistics** | | |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 1106.827 | 825.142 |
| **SC** | 1111.511 | 862.619 |
| **-2 Log L** | 1104.827 | 809.142 |

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -11.6910 | 3.0092 | 15.0941 | 0.0001 |
| **log.chlorides** | 1 | -1.0685 | 0.3215 | 11.0442 | 0.0009 |
| **log.total.sulfur.dio** | 1 | -0.3460 | 0.1332 | 6.7514 | 0.0094 |
| **log.sulphates** | 1 | 3.5737 | 0.4936 | 52.4218 | <.0001 |
| **log.alcohol** | 1 | 9.0324 | 1.1476 | 61.9431 | <.0001 |
| **volatile.acidity** | 1 | -3.2189 | 0.6516 | 24.4033 | <.0001 |
| **citric.acid** | 1 | -2.1847 | 0.6547 | 11.1336 | 0.0008 |
| **pH** | 1 | -2.0224 | 0.7030 | 8.2758 | 0.0040 |

| **Classification Table** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Prob Level** | **Correct** | | **Incorrect** | | **Percentages** | | | | |
| **Event** | **Non- Event** | **Event** | **Non- Event** | **Correct** | **Sensi- tivity** | **Speci- ficity** | **False POS** | **False NEG** |
| **0.020** | 429 | 0 | 371 | 0 | 53.6 | 100.0 | 0.0 | 46.4 | . |
| **0.040** | 429 | 3 | 368 | 0 | 54.0 | 100.0 | 0.8 | 46.2 | 0.0 |
| **0.060** | 429 | 9 | 362 | 0 | 54.8 | 100.0 | 2.4 | 45.8 | 0.0 |
| **0.080** | 427 | 14 | 357 | 2 | 55.1 | 99.5 | 3.8 | 45.5 | 12.5 |
| **0.100** | 426 | 21 | 350 | 3 | 55.9 | 99.3 | 5.7 | 45.1 | 12.5 |
| **0.120** | 422 | 39 | 332 | 7 | 57.6 | 98.4 | 10.5 | 44.0 | 15.2 |
| **0.140** | 420 | 50 | 321 | 9 | 58.8 | 97.9 | 13.5 | 43.3 | 15.3 |
| **0.160** | 417 | 63 | 308 | 12 | 60.0 | 97.2 | 17.0 | 42.5 | 16.0 |
| **0.180** | 415 | 85 | 286 | 14 | 62.5 | 96.7 | 22.9 | 40.8 | 14.1 |
| **0.200** | 408 | 104 | 267 | 21 | 64.0 | 95.1 | 28.0 | 39.6 | 16.8 |
| **0.220** | 406 | 125 | 246 | 23 | 66.4 | 94.6 | 33.7 | 37.7 | 15.5 |
| **0.240** | 399 | 132 | 239 | 30 | 66.4 | 93.0 | 35.6 | 37.5 | 18.5 |
| **0.260** | 392 | 154 | 217 | 37 | 68.3 | 91.4 | 41.5 | 35.6 | 19.4 |
| **0.280** | 386 | 170 | 201 | 43 | 69.5 | 90.0 | 45.8 | 34.2 | 20.2 |
| **0.300** | 381 | 187 | 184 | 48 | 71.0 | 88.8 | 50.4 | 32.6 | 20.4 |
| **0.320** | 378 | 197 | 174 | 51 | 71.9 | 88.1 | 53.1 | 31.5 | 20.6 |
| **0.340** | 370 | 208 | 163 | 59 | 72.3 | 86.2 | 56.1 | 30.6 | 22.1 |
| **0.360** | 367 | 219 | 152 | 62 | 73.3 | 85.5 | 59.0 | 29.3 | 22.1 |
| **0.380** | 363 | 230 | 141 | 66 | 74.1 | 84.6 | 62.0 | 28.0 | 22.3 |
| **0.400** | 358 | 239 | 132 | 71 | 74.6 | 83.4 | 64.4 | 26.9 | 22.9 |
| **0.420** | 352 | 241 | 130 | 77 | 74.1 | 82.1 | 65.0 | 27.0 | 24.2 |
| **0.440** | 342 | 247 | 124 | 87 | 73.6 | 79.7 | 66.6 | 26.6 | 26.0 |
| **0.460** | 338 | 259 | 112 | 91 | 74.6 | 78.8 | 69.8 | 24.9 | 26.0 |
| **0.480** | 333 | 266 | 105 | 96 | 74.9 | 77.6 | 71.7 | 24.0 | 26.5 |
| **0.500** | 329 | 275 | 96 | 100 | 75.5 | 76.7 | 74.1 | 22.6 | 26.7 |
| **0.520** | 317 | 280 | 91 | 112 | 74.6 | 73.9 | 75.5 | 22.3 | 28.6 |
| **0.540** | 308 | 285 | 86 | 121 | 74.1 | 71.8 | 76.8 | 21.8 | 29.8 |
| **0.560** | 301 | 291 | 80 | 128 | 74.0 | 70.2 | 78.4 | 21.0 | 30.5 |
| **0.580** | 294 | 297 | 74 | 135 | 73.9 | 68.5 | 80.1 | 20.1 | 31.3 |
| **0.600** | 287 | 305 | 66 | 142 | 74.0 | 66.9 | 82.2 | 18.7 | 31.8 |
| **0.620** | 276 | 309 | 62 | 153 | 73.1 | 64.3 | 83.3 | 18.3 | 33.1 |
| **0.640** | 270 | 312 | 59 | 159 | 72.8 | 62.9 | 84.1 | 17.9 | 33.8 |
| **0.660** | 260 | 317 | 54 | 169 | 72.1 | 60.6 | 85.4 | 17.2 | 34.8 |
| **0.680** | 254 | 320 | 51 | 175 | 71.8 | 59.2 | 86.3 | 16.7 | 35.4 |
| **0.700** | 243 | 326 | 45 | 186 | 71.1 | 56.6 | 87.9 | 15.6 | 36.3 |
| **0.720** | 230 | 329 | 42 | 199 | 69.9 | 53.6 | 88.7 | 15.4 | 37.7 |
| **0.740** | 216 | 334 | 37 | 213 | 68.8 | 50.3 | 90.0 | 14.6 | 38.9 |
| **0.760** | 209 | 343 | 28 | 220 | 69.0 | 48.7 | 92.5 | 11.8 | 39.1 |
| **0.780** | 200 | 349 | 22 | 229 | 68.6 | 46.6 | 94.1 | 9.9 | 39.6 |
| **0.800** | 187 | 351 | 20 | 242 | 67.3 | 43.6 | 94.6 | 9.7 | 40.8 |
| **0.820** | 176 | 358 | 13 | 253 | 66.8 | 41.0 | 96.5 | 6.9 | 41.4 |
| **0.840** | 158 | 359 | 12 | 271 | 64.6 | 36.8 | 96.8 | 7.1 | 43.0 |
| **0.860** | 144 | 362 | 9 | 285 | 63.3 | 33.6 | 97.6 | 5.9 | 44.0 |
| **0.880** | 126 | 365 | 6 | 303 | 61.4 | 29.4 | 98.4 | 4.5 | 45.4 |
| **0.900** | 108 | 366 | 5 | 321 | 59.3 | 25.2 | 98.7 | 4.4 | 46.7 |
| **0.920** | 74 | 367 | 4 | 355 | 55.1 | 17.2 | 98.9 | 5.1 | 49.2 |
| **0.940** | 55 | 369 | 2 | 374 | 53.0 | 12.8 | 99.5 | 3.5 | 50.3 |
| **0.960** | 33 | 370 | 1 | 396 | 50.4 | 7.7 | 99.7 | 2.9 | 51.7 |
| **0.980** | 11 | 371 | 0 | 418 | 47.8 | 2.6 | 100.0 | 0.0 | 53.0 |
| **1.000** | 0 | 371 | 0 | 429 | 46.4 | 0.0 | 100.0 | . | 53.6 |



## SAS Code for Regressions

proc import datafile="/home/anhainguyen820/sasuser.v94/wine\_train.csv"

dbms=dlm out=wine replace;

delimiter=',';

getnames=yes;

run;

data wine;

set wine;

"log.residual.sugar"N = log("residual.sugar"N);

"log.chlorides"N = log(chlorides);

"log.free.sulfur.dioxide"N = log("free.sulfur.dioxide"N);

"log.total.sulfur.dioxide"N = log("total.sulfur.dioxide"N);

"log.sulphates"N = log(sulphates);

"log.alcohol"N = log(alcohol);

run;

/\*Logistic regression\*/

proc logistic data=wine;

class Outcome / param=ref;

model Outcome(event='fine') = "log.residual.sugar"N

"log.chlorides"N

"log.free.sulfur.dioxide"N

"log.total.sulfur.dioxide"N

"log.sulphates"N

"log.alcohol"N

"fixed.acidity"N

"volatile.acidity"N

"citric.acid"N

density

pH/ selection=forward scale=none lackfit ctable;

run;

## QDA

**Test of Homogeneity of Within Covariance Matrices**

| **Chi-Square** | **DF** | **Pr > ChiSq** |
| --- | --- | --- |
| 264.243869 | 66 | <.0001 |

| **Number of Observations and Percent Classified into Outcome** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **From Outcome** | | | **fine** | | **poor** | **Total** |
| **fine** | | | 312  72.73 | | 117  27.27 | 429  100.00 |
| **poor** | | | 76  20.49 | | 295  79.51 | 371  100.00 |
| **Total** | | | 388  48.50 | | 412  51.50 | 800  100.00 |
| **Priors** | | | 0.5 | | 0.5 |  |
| **Error Count Estimates for Outcome** | | | | |
|  | **fine** | **poor** | | **Total** |
| **Rate** | 0.2727 | 0.2049 | | 0.2388 |
| **Priors** | 0.5000 | 0.5000 | |  |

| **Number of Observations and Percent Classified into Outcome** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **From Outcome** | | | **fine** | | **poor** | **Total** |
| **fine** | | | 313  73.47 | | 113  26.53 | 426  100.00 |
| **poor** | | | 106  28.42 | | 267  71.58 | 373  100.00 |
| **Total** | | | 419  52.44 | | 380  47.56 | 799  100.00 |
| **Priors** | | | 0.5 | | 0.5 |  |
| **Error Count Estimates for Outcome** | | | | |
|  | **fine** | **poor** | | **Total** |
| **Rate** | 0.2653 | 0.2842 | | 0.2747 |
| **Priors** | 0.5000 | 0.5000 | |  |

## SAS Code for QDA

proc import datafile="/home/anhainguyen820/sasuser.v94/wine\_train.csv"

dbms=dlm out=wine replace;

delimiter=',';

getnames=yes;

run;

proc import datafile="/home/anhainguyen820/sasuser.v94/wine\_test.csv"

dbms=dlm out=test replace;

delimiter=',';

getnames=yes;

run;

data wine; set wine;

"log.residual.sugar"N = log("residual.sugar"N);

"log.chlorides"N = log(chlorides);

"log.free.sulfur.dioxide"N = log("free.sulfur.dioxide"N);

"log.total.sulfur.dioxide"N = log("total.sulfur.dioxide"N);

"log.sulphates"N = log(sulphates);

"log.alcohol"N = log(alcohol);

run;

data test; set test;

"log.residual.sugar"N = log("residual.sugar"N);

"log.chlorides"N = log(chlorides);

"log.free.sulfur.dioxide"N = log("free.sulfur.dioxide"N);

"log.total.sulfur.dioxide"N = log("total.sulfur.dioxide"N);

"log.sulphates"N = log(sulphates);

"log.alcohol"N = log(alcohol);

run;

proc discrim data=wine pool=test testdata=test;

class Outcome;

var "log.residual.sugar"N

"log.chlorides"N

"log.free.sulfur.dioxide"N

"log.total.sulfur.dioxide"N

"log.sulphates"N

"log.alcohol"N

"fixed.acidity"N

"volatile.acidity"N

"citric.acid"N

density

pH;

run;

# Objective 2 – Nonparametric approach

library(ggplot2)

library(tree)

library(ISLR)

library(randomForest)

library(FNN)

library(reprtree)

setwd("D:/My\_Docs/univer/Stat2/Project2/MSDS6372\_Project\_2-master")

wine<-read.csv("winequality-red.csv",header=T)

attach(wine)

names(wine) <- gsub(x = names(wine),

pattern = "\\.",

replacement = "\_")

QualityCat=ifelse(quality<=5,"Poor","Fine")

wine=data.frame(wine,QualityCat)

set.seed(2)

train=sample(1:nrow(wine), nrow(wine)\*2/3) #split 2/3 train 1/3 tes

wine.test=wine[-train,]

wine.train=wine[train,] #test data set

Quality.test=QualityCat[-train] #test

set.seed(5)

rf.wine<-randomForest(factor(QualityCat)~.-quality,wine,mtry=5,subset=train,importance=F,ntree=100)

rf.wine

rf.wine$err.rate[,1]

plot(rf.wine$err.rate[,1],xlab="Number of Trees",ylab="Error Rate")

fit.pred<-predict(rf.wine,newdata=wine.test,type="response")

table(fit.pred,wine.test$QualityCat)

confusionMatrix(fit.pred, wine.test$QualityCat)

# Get importance

importance <- importance(rf.wine)

varImportance <- data.frame(Variables = row.names(importance),

Importance = round(importance[ ,'MeanDecreaseGini'],2))

# Create a rank variable based on importance

rankImportance <- varImportance %>%

mutate(Rank = paste0('#',dense\_rank(desc(Importance))))

# Use ggplot2 to visualize the relative importance of variables

ggplot(rankImportance, aes(x = reorder(Variables, Importance),

y = Importance, fill = Importance)) +

geom\_bar(stat='identity') +

geom\_text(aes(x = Variables, y = 0.5, label = Rank),

hjust=0, vjust=0.55, size = 4, colour = 'red') +

labs(x = 'Variables') +

coord\_flip() +

theme\_classic()