Part 2: Classification on Wine Review Data Set

Code

Rafael Melendez

Hide

```
wdata <- read.csv("C:\\Users\\bayon\\OneDrive\\Documents\\winemag-data_first150k.csv", header = TRUE, sep = ",")
wdata
```

X <int>	country <chr>
0	US
1	Spain
2	US
3	US
4	France
5	Spain
6	Spain
7	Spain
8	US
9	US

1-10 of 150,930 rows | 1-2 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
w_rem <- wdata[1:15000, ]
# remove rows with missing values in the price column
w_rem <- w_rem[!is.na(w_rem$price),]
w_rem <- w_rem[, -1]
w_rem <- w_rem[!is.na(w_rem$region_1),]
w_rem <- w_rem[!is.na(w_rem$region_2),]
```

Hide

```
#A
str(wdata)
```

```
'data.frame':   150930 obs. of  11 variables:
 $ X          : int  0 1 2 3 4 5 6 7 8 9 ...
 $ country    : chr  "US" "Spain" "US" "US" ...
 $ description: chr  "This tremendous 100% varietal wine hails from Oakville and was aged over t
hree years in oak. Juicy red-cherry f"| __truncated__ "Ripe aromas of fig, blackberry and cassis
are softened and sweetened by a slathering of oaky chocolate and vani"| __truncated__ "Mac Watso
n honors the memory of a wine once made by his mother in this tremendously delicious, balanced a
nd com"| __truncated__ "This spent 20 months in 30% new French oak, and incorporates fruit from
Ponzi's Aurora, Abetina and Madrona vin"| __truncated__ ...
 $ designation: chr  "Martha's Vineyard" "Carodorum Selección Especial Reserva" "Special Selecte
d Late Harvest" "Reserve" ...
 $ points     : int  96 96 96 96 95 95 95 95 95 95 ...
 $ price      : num  235 110 90 65 66 73 65 110 65 60 ...
 $ province   : chr  "California" "Northern Spain" "California" "Oregon" ...
 $ region_1   : chr  "Napa Valley" "Toro" "Knights Valley" "Willamette Valley" ...
 $ region_2   : chr  "Napa" "" "Sonoma" "Willamette Valley" ...
 $ variety    : chr  "Cabernet Sauvignon" "Tinta de Toro" "Sauvignon Blanc" "Pinot Noir" ...
 $ winery     : chr  "Heitz" "Bodega Carmen Rodríguez" "Macauley" "Ponzi" ...
```

Hide

```
set.seed(1234)
i <- sample(1:nrow(wdata), nrow(wdata)*0.8, replace=FALSE)
train <- wdata[i,]
test <- wdata[-i,]
cl <- wdata[i,]

names(wdata)
```

```
 [1] "X"           "country"     "description"
 [4] "designation" "points"      "price"
 [7] "province"    "region_1"    "region_2"
[10] "variety"     "winery"
```

Hide

```
ncol(wdata)
```

```
[1] 11
```

Hide

```
tail(wdata, n = 10)
```

	X	country
	<int>	<chr>
150921	150920	Italy

	X <int>	country <chr>
150922	150921	France
150923	150922	Italy
150924	150923	France
150925	150924	France
150926	150925	Italy
150927	150926	France
150928	150927	Italy
150929	150928	France
150930	150929	Italy

1-10 of 10 rows | 1-3 of 11 columns

Hide

```
colSums(is.na(wdata))
```

```
         X     country description designation 
         0           0           0           0 
    points       price    province    region_1 
         0       13695           0           0 
  region_2     variety      winery 
         0           0           0 
```

Hide

```
summary(wine_train)
```

```
      X            country          description
 Min.   :     0   Length:120744      Length:120744
 1st Qu.: 37745   Class :character   Class :character
 Median : 75500   Mode  :character   Mode  :character
 Mean   : 75483
 3rd Qu.:113249
 Max.   :150929

 designation            points           price
 Length:120744      Min.   : 80.00   Min.   :   4.00
 Class :character   1st Qu.: 86.00   1st Qu.:  16.00
 Mode  :character   Median : 88.00   Median :  24.00
                    Mean   : 87.89   Mean   :  33.14
                    3rd Qu.: 90.00   3rd Qu.:  40.00
                    Max.   :100.00   Max.   :2300.00
                                     NA's   :10916
   province           region_1           region_2
 Length:120744      Length:120744      Length:120744
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character




   variety             winery
 Length:120744      Length:120744
 Class :character   Class :character
 Mode  :character   Mode  :character
```

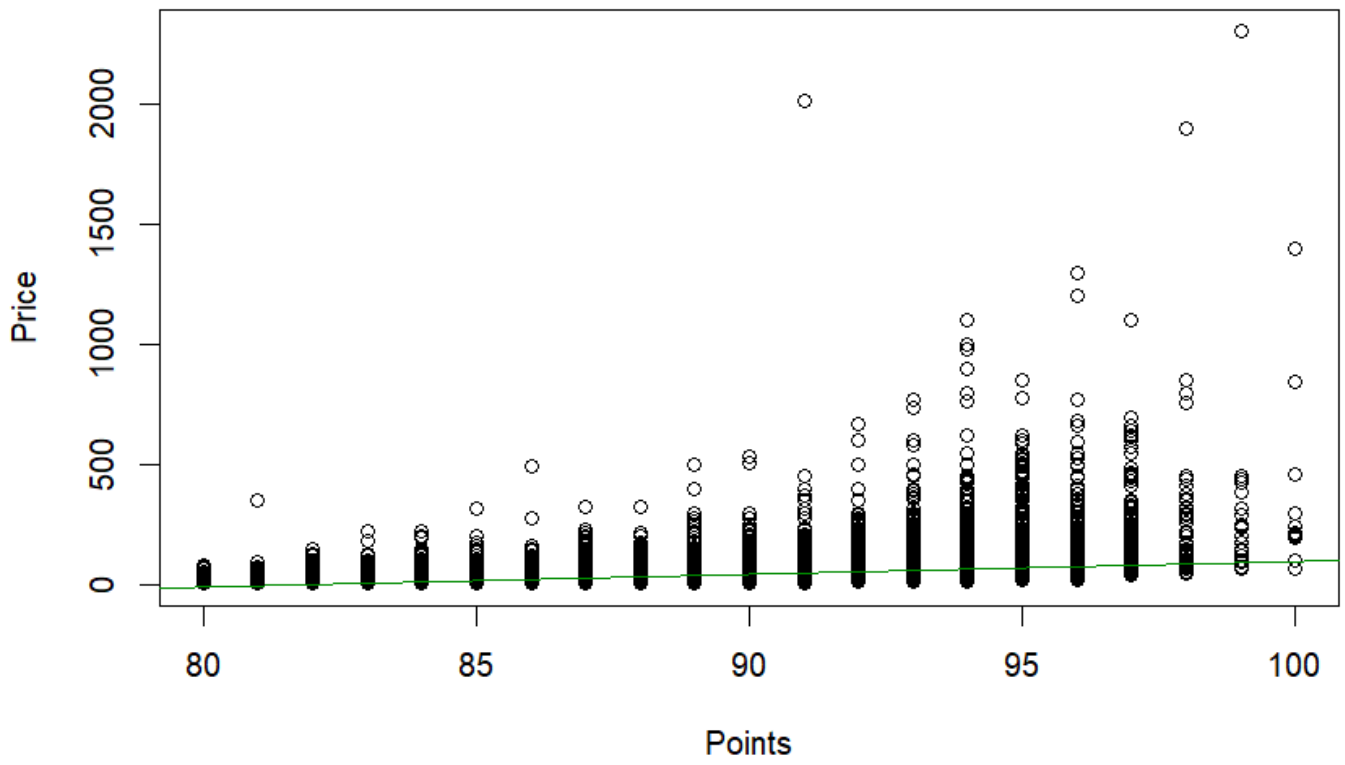
Hide

```
summary(wine_test)
```

```
      X            country          description
 Min.   :     3   Length:30186       Length:30186
 1st Qu.: 37685   Class :character   Class :character
 Median : 75319   Mode  :character   Mode  :character
 Mean   : 75389
 3rd Qu.:112982
 Max.   :150924

 designation            points           price
 Length:30186       Min.   : 80.00   Min.   :   4.00
 Class :character   1st Qu.: 86.00   1st Qu.:  16.00
 Mode  :character   Median : 88.00   Median :  24.00
                    Mean   : 87.87   Mean   :  33.11
                    3rd Qu.: 90.00   3rd Qu.:  40.00
                    Max.   :100.00   Max.   :1000.00
                                     NA's   :2779
   province           region_1           region_2
 Length:30186       Length:30186       Length:30186
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character




   variety             winery
 Length:30186       Length:30186
 Class :character   Class :character
 Mode  :character   Mode  :character
```

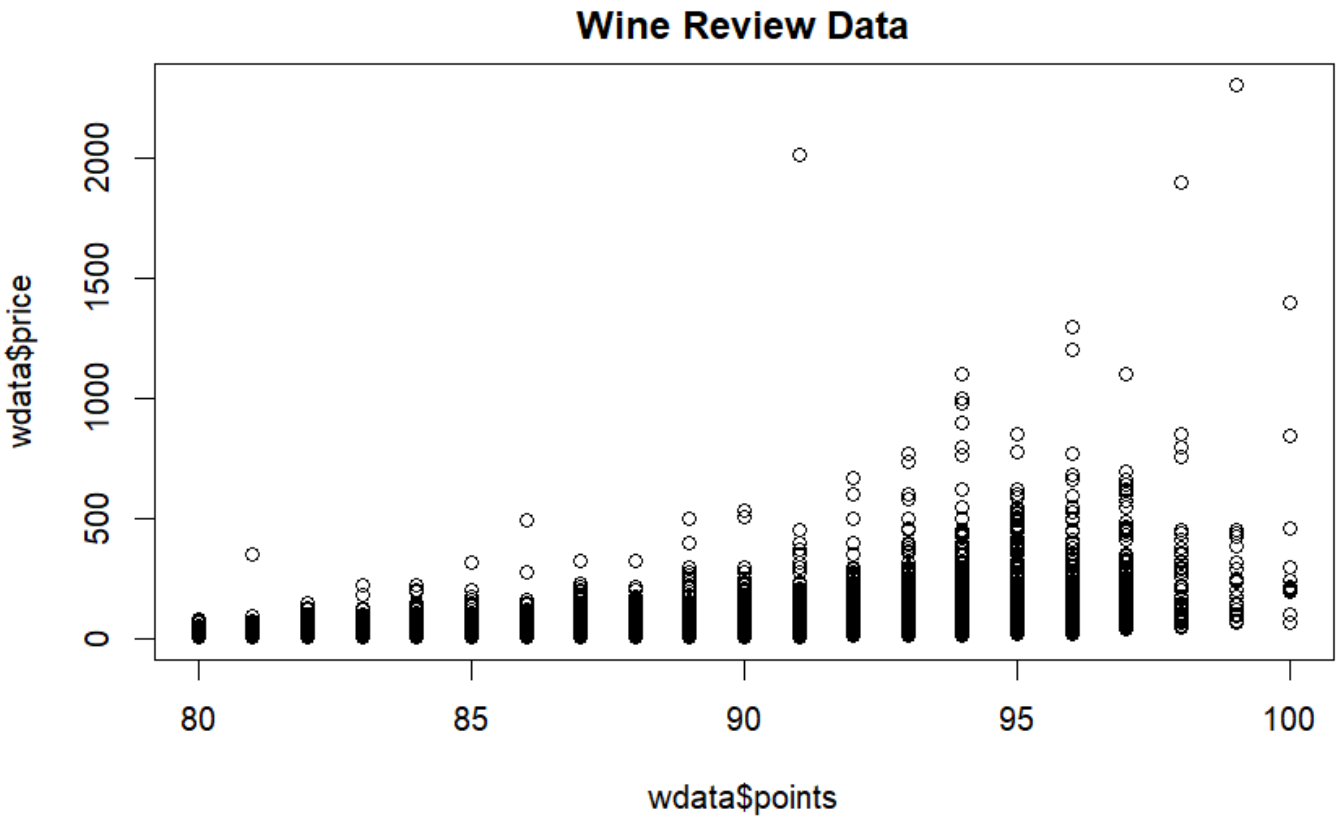
Hide

```
#B
plot(wdata$price~wdata$points, xlab = "Points", ylab = "Price")
abline(lm(wdata$price~wdata$points), col="green4")
```



Hide

```
#plot(wdata$points~wdata$variety, xlab = "Variety", ylab = "Points")
#abline(lm(wdata$points~wdata$variety), col="green4")

plot(wdata$points, wdata$price, pch=21, bg=c("green4","blue4", "orange2", "red3", "yellow2", "pu
rple4")
     [unclass(wdata$country)], main="Wine Review Data")
```



Hide

```
#C: Logistic Model
glm1 <- glm(as.factor(wdata$price)~wdata$points, data = train, family = "binomial")
summary(glm1)
```

```
Call:
glm(formula = as.factor(wdata$price) ~ wdata$points, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2190   0.0082   0.0116   0.0165   0.0474

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -21.35175    7.32423  -2.915  0.00355 **
wdata$points   0.35176    0.08652   4.066 4.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 357.80  on 137234  degrees of freedom
Residual deviance: 338.81  on 137233  degrees of freedom
  (13695 observations deleted due to missingness)
AIC: 342.81

Number of Fisher Scoring iterations: 12
```

Hide

```
pred <- predict(glm1, newdata = test)
```

```
Warning: 'newdata' had 30186 rows but variables found have 150930 rows
```

Hide

```
pred <- exp(pred1)
cor <- cor(pred, test$points)
```

```
Error in cor(pred, test$points) : incompatible dimensions
```

Hide

```
#C: kNN Classification

set.seed(1234)
i <- sample(2, nrow(wdata), replace=TRUE, prob=c(0.8, 0.2))
wtrain <- wdata[i==1, 1:4]
wtest <- wdata[i==2, 1:4]
wTrainL <- wdata[i==1, 5]
wTestL <- wdata[i==2, 5]

library(class)
wine_pred <- knn(train=iris.train, test=iris.test, cl=iris.trainLabels, k=3)
```

```
Warning: NAs introduced by coercionWarning: NAs introduced by coercionError in knn(train = iris.
train, test = iris.test, cl = iris.trainLabels,  :
  NA/NaN/Inf in foreign function call (arg 6)
```

Hide

```
#C: Decision Trees
library(rpart)
#install.packages("tree")
library(tree)
wine_tree <- rpart(variety~., data = wdata, method = "class")
```