Assignment 1

Greg Strohl

10/20/2019

Explain your code where appropriate.

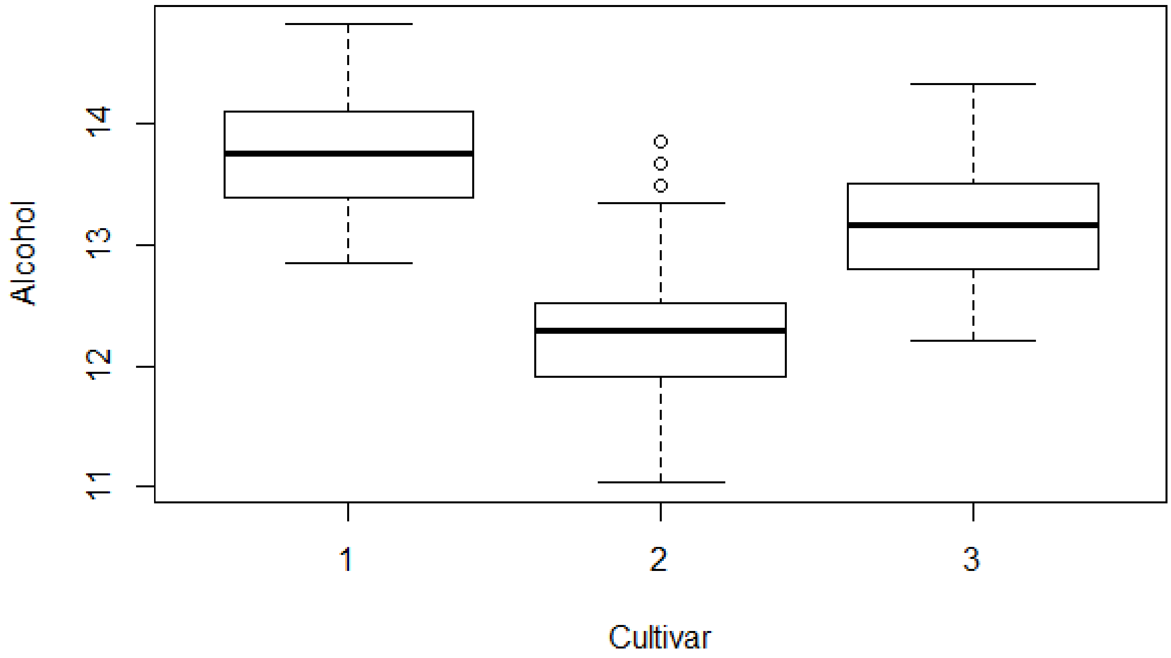
Reference: UCI Machine Learning Repository: Wine Dataset

The wine data set consists of chemical measurements on 13 constituents found in each of the three types of wines (cultivars) grown in the same region of Italy.

1. Compute the Alcohol means by `Cultivar`. Briefly discuss the mean differences among the cultivars.

```
Cultivar<-as.factor(Wine[,"Cultivar"])

## Error in is.factor(x): object 'Wine' not found

Alcohol<-Wine[,"Alcohol"]

## Error in eval(expr, envir, enclos): object 'Wine' not found

# Split `alcohol` by `cultivar` to form a list with each element containing
the data for a group.
# Look at the function `sapply`.
# Put your R code here.
AlcbyCult<-cbind(Cultivar,Alcohol)

## Error in cbind(Cultivar, Alcohol): object 'Cultivar' not found

sapply(split(Alcohol,Cultivar),mean)

## Error in split(Alcohol, Cultivar): object 'Alcohol' not found

alcoholList<-split(Alcohol,Cultivar)

## Error in split(Alcohol, Cultivar): object 'Alcohol' not found

sapply(alcoholList,mean)

## Error in lapply(X = X, FUN = FUN, ...): object 'alcoholList' not found

summary(Wine)

## Error in summary(Wine): object 'Wine' not found

plot(Alcohol~Cultivar)

## Error in eval(predvars, data, env): object 'Alcohol' not found
```

```
       1        2        3
13.74475 12.27873 13.15375
       1        2        3
13.74475 12.27873 13.15375
    Cultivar        Alcohol        Malic_acid         Ash
 Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
 1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210
 Median :2.000   Median :13.05   Median :1.865   Median :2.360
 Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367
 3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558
 Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230
 Alcalinity_ash    Magnesium      Total_phenols     Flavanoids
 Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340
 1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
 Median :19.50   Median : 98.00   Median :2.355   Median :2.135
 Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029
 3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875
 Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080
 Nonflavanoid_phenols Proanthocyanins Color_intensity       Hue
 Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800
 1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825
 Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650
 Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574
 3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200
 Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100
 OD280toOD315_diluted    Proline
 Min.   :1.270        Min.   : 278.0
 1st Qu.:1.938        1st Qu.: 500.5
 Median :2.780       Median : 673.5
 Mean   :2.612      Mean   : 746.9
 3rd Qu.:3.170      3rd Qu.: 985.0
 Max.   :4.000      Max.   :1680.0
```



I combined the alcohol and cultivar into a matrix called *AlcByCult* using the ‘cbind’ function. From there, I used the ‘sapply’ function to calculate the mean alcohol per cultivar.

The mean alcohol for cultivar 1 was 13.74475.

The mean alcohol for cultivar 2 was 12.27873.

The mean alcohol for cultivar 3 was 13.15375.

2. Compute the number of observations in each `cultivar`.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

dim(Wine)

## Error in eval(expr, envir, enclos): object 'Wine' not found

arrange(Wine)

## Error in arrange(Wine): object 'Wine' not found

Wine %>%
  group_by(Cultivar) %>%
  summarize(n())

## Error in eval(lhs, parent, parent): object 'Wine' not found

sapply(alcoholList,length)

## Error in lapply(X = X, FUN = FUN, ...): object 'alcoholList' not found
```

```
[1] 178  14

 1  2  3
59 71 48
```

R Console

Cultivar <int>	Alcohol <dbl>	Malic_acid <dbl>	Ash <dbl>	Alcalinity_ash <dbl>	Magnesium <int>
1	14.23	1.71	2.43	15.6	127
1	13.20	1.78	2.14	11.2	100
1	13.16	2.36	2.67	18.6	101
1	14.37	1.95	2.50	16.8	113
1	13.24	2.59	2.87	21.0	118
1	14.20	1.76	2.45	15.2	112
1	14.39	1.87	2.45	14.6	96
1	14.06	2.15	2.61	17.6	121
1	14.83	1.64	2.17	14.0	97
1	13.86	1.35	2.27	16.0	98

Next
123456
…
18

Previous
1-10 of 178 rows | 1-6 of 14 columns

data.frame
178 x 14

Cultivar <int>	n() <int>
1	59
2	71
3	48

3 rows

tbl_df
3 x 2

Cultivar <int>	Alcohol <dbl>	Malic_acid <dbl>	Ash <dbl>	Alcalinity_ash <dbl>	Magnesium <int>	Total_phenols <dbl>
1	14.23	1.71	2.43	15.6	127	2.80
1	13.20	1.78	2.14	11.2	100	2.65
1	13.16	2.36	2.67	18.6	101	2.80
1	14.37	1.95	2.50	16.8	113	3.85
1	13.24	2.59	2.87	21.0	118	2.80
1	14.20	1.76	2.45	15.2	112	3.27
1	14.39	1.87	2.45	14.6	96	2.50
1	14.06	2.15	2.61	17.6	121	2.60
1	14.83	1.64	2.17	14.0	97	2.80
1	13.86	1.35	2.27	16.0	98	2.98
1	14.10	2.16	2.30	18.0	105	2.95
1	14.12	1.48	2.32	16.8	95	2.20
1	13.75	1.73	2.41	16.0	89	2.60
1	14.75	1.73	2.39	11.4	91	3.10
1	14.38	1.87	2.38	12.0	102	3.30
1	13.63	1.81	2.70	17.2	112	2.85
1	14.30	1.92	2.72	20.0	120	2.80
1	13.83	1.57	2.62	20.0	115	2.95
1	14.19	1.59	2.48	16.5	108	3.30
1	13.64	3.10	2.56	15.2	116	2.70
1	14.06	1.63	2.28	16.0	126	3.00
1	12.93	3.80	2.65	18.6	102	2.41

Next
123456
...
9
Previous
1-22 of 178 rows | 1-8 of 14 columns

Using the wine data set, I used the dplyr function to group and summarize the data by Cultivar.

There were 59 observations for cultivar 1.

There were 71 observations for cultivar 2.

There were 48 observations for cultivar 3.

3. Create a function to perform a one-way analysis of variance. The input argument z should be a list consisting of (possibly) named components, one for each group. The output should be a named list containing components for the between SS (SS_B), the within SS (SS_W), the between degrees of freedom, and the within degrees of freedom. Note: $SS_B = \sum_i n_i \, (\overline{y}_i - \overline{y})^2$ and $SS_W = \sum_i \, (n_i - 1)s_i^2$ where n_i is the sample size of group i, $\overline{y}_i$ is the mean of group i, and s_i^2 is the variance of group i. These group statistics can easily be computed using `sapply`. For the grand mean, $\overline{y}$, think about using `unlist` on `z`. Let g be the number of groups and $n = \sum_i n_i$ be the total sample size, which can also be computed by `unlist`ing `z`. n and g are needed to compute the between and within degrees of freedom.
   Note: The code should be general for any g and n_i.

```r
# Look at sapply for summarizing over the elements of a list.
oneway <- function(z){
  # Put your R code here.
  summary(Wine)
  n <- length(unlist(z,recursive = TRUE))
  n_i <-sapply(z,length)
  s_i <- sapply(z,var)
  g <- length(z)
  y_bar <-mean(unlist(z,recursive = TRUE))
  y<- sapply(z,mean)
  ssb<- n_i*(sapply(z,mean)-y_bar)^2
  ssw<- sum(n_i-1)*s_i^2
  return(list(ssb=sum(ssb),ssw=sum(ssw),n=n,g=g))
}
x<-oneway(alcoholList)

## Error in summary(Wine): object 'Wine' not found

x

## Error in eval(expr, envir, enclos): object 'x' not found

df1 = x[[4]]-1

## Error in eval(expr, envir, enclos): object 'x' not found
```

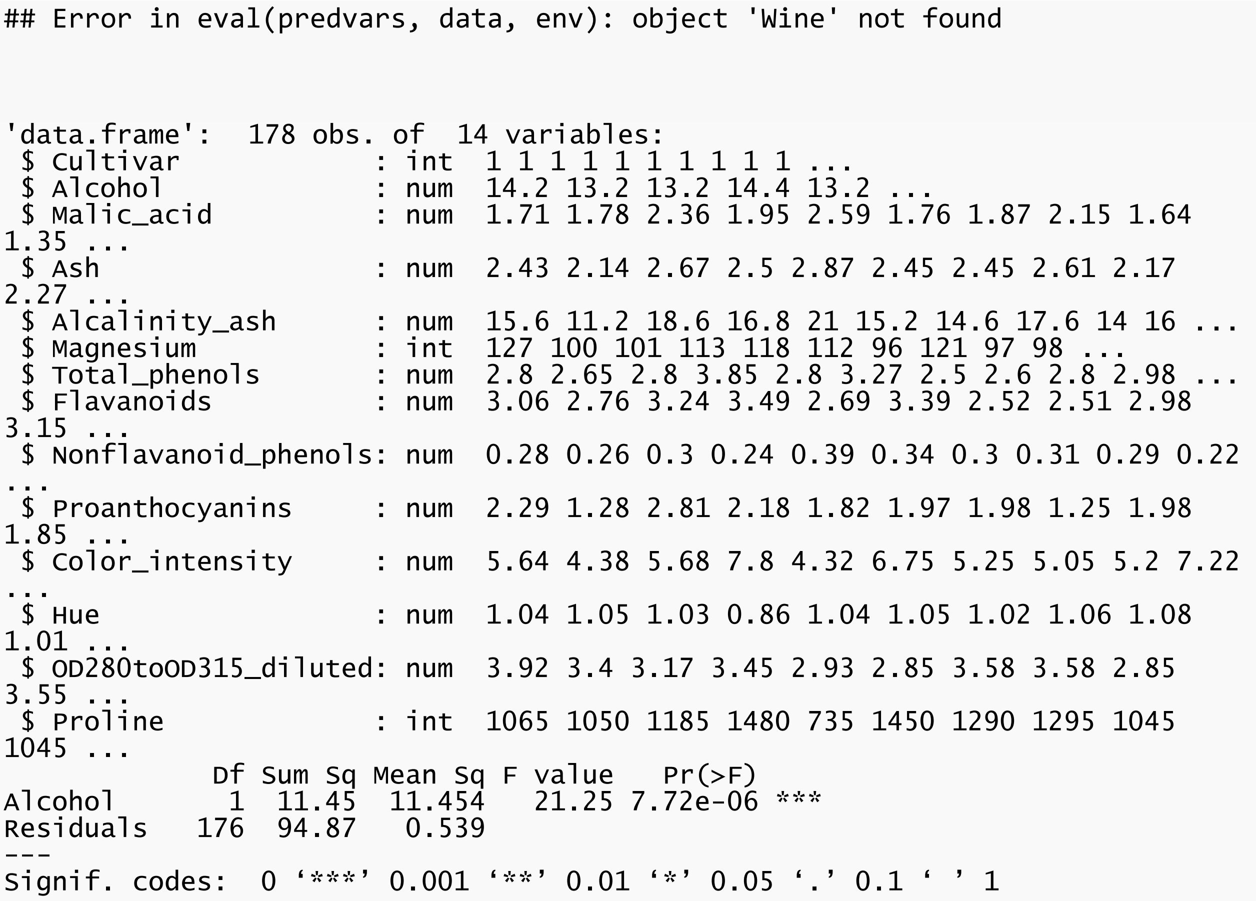
```
df2 = x[[3]]-x[[4]]

## Error in eval(expr, envir, enclos): object 'x' not found

p = pf(x[[2]],df1,df2)

## Error in pf(x[[2]], df1, df2): object 'x' not found

df1

## Error in eval(expr, envir, enclos): object 'df1' not found

df2

## Error in eval(expr, envir, enclos): object 'df2' not found

p

## Error in eval(expr, envir, enclos): object 'p' not found

$ssb

[1] 70.79485


$ssw

[1] 36.4721


$n

[1] 178


$g

[1] 3


[1] 2

[1] 175

[1] 1
```

4. Create a function to summarize the output in a one-way ANOVA table, including the F test and *p*-value. The input argument is the output named list in the previous question. The output should be one-way ANOVA table.
   Note: For computing the *p*-value look at the R function `pf`.

```
# For your output, mimic the tabular output of the builtin `summary` function
applied to the output of the builtin `aov` function.
# Look at the function `printCoefmat` to form a table.
oneway.table <- function(x){
  # Put your R code here.
  df1 = x[[4]]-1
  df2 = x[[3]]-x[[4]]
  p = pf(x[[1]],df1,df2)
  ss = sum(x[[1]]^2)
}
```

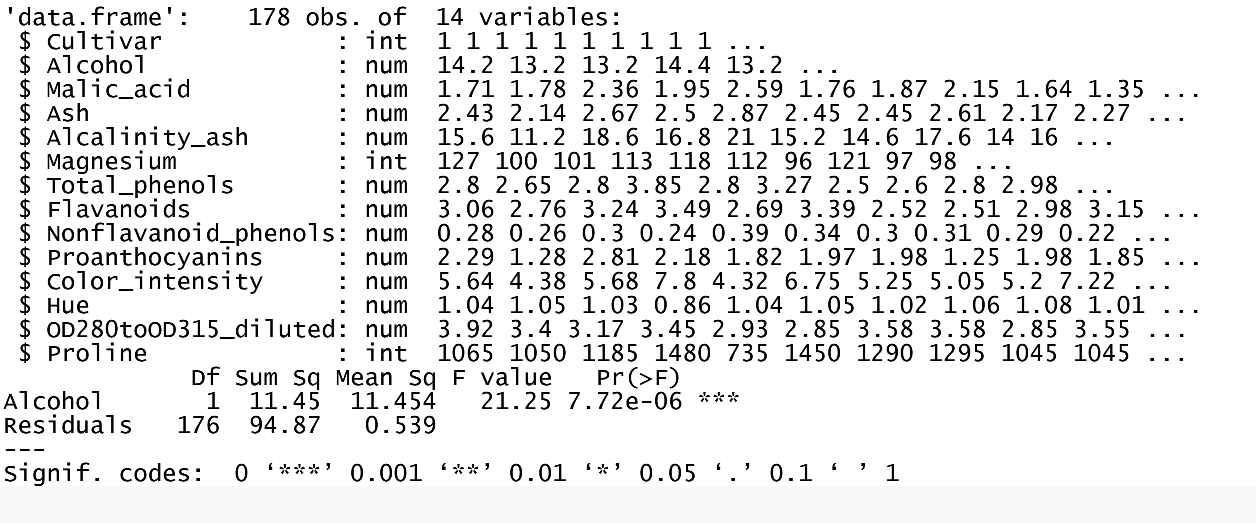
5. Your functions should be illustrated with the `wine` data set. The data consists of 178 samples measuring alcohol (the outcome variable) divided among three (3) cultivars (the input variable).

```
# Split `alcohol` by `cultivar` to call `oneway`.
# Put your R code here.
attach(Wine)

## Error in attach(Wine): object 'Wine' not found

data(Wine)
str(Wine)

## Error in str(Wine): object 'Wine' not found

# Summary of the analysis
wine.aov <- aov(Cultivar~Alcohol, data = Wine)

## Error in terms.formula(formula, "Error", data = data): object 'Wine' not
found

summary(wine.aov)

## Error in summary(wine.aov): object 'wine.aov' not found

model1<- aov(Cultivar ~ Alcohol)

## Error in eval(predvars, data, env): object 'Cultivar' not found

par(mfrow=c(2,2))
plot(wine.aov, 2)

## Error in plot(wine.aov, 2): object 'wine.aov' not found

boxplot(Wine$Alcohol ~ Cultivar,
        vertical = TRUE,
        main="AlcByCult",
        col = "blue")
```



```
## Error in eval(predvars, data, env): object 'Wine' not found



'data.frame':  178 obs. of  14 variables:
 $ Cultivar           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol            : num  14.2 13.2 13.2 14.4 13.2 ...
 $ Malic_acid         : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64
1.35 ...
 $ Ash                : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17
2.27 ...
 $ Alcalinity_ash     : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ Magnesium          : int  127 100 101 113 118 112 96 121 97 98 ...
 $ Total_phenols      : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ Flavanoids         : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98
3.15 ...
 $ Nonflavanoid_phenols: num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22
...
 $ Proanthocyanins    : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98
1.85 ...
 $ Color_intensity    : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22
...
 $ Hue                : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08
1.01 ...
 $ OD280toOD315_diluted: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85
3.55 ...
 $ Proline            : int  1065 1050 1185 1480 735 1450 1290 1295 1045
1045 ...
             Df Sum Sq Mean Sq F value   Pr(>F)    
Alcohol       1  11.45  11.454   21.25 7.72e-06 ***
Residuals   176  94.87   0.539                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R Console



```
'data.frame':   178 obs. of  14 variables:
 $ Cultivar           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol            : num  14.2 13.2 13.2 14.4 13.2 ...
 $ Malic_acid         : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
 $ Ash                : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
 $ Alcalinity_ash     : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ Magnesium          : int  127 100 101 113 118 112 96 121 97 98 ...
 $ Total_phenols      : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ Flavanoids         : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
 $ Nonflavanoid_phenols: num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
 $ Proanthocyanins    : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
 $ Color_intensity    : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
 $ Hue                : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
 $ OD280toOD315_diluted: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
 $ Proline            : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
             Df Sum Sq Mean Sq F value   Pr(>F)    
Alcohol       1  11.45  11.454   21.25 7.72e-06 ***
Residuals   176  94.87   0.539                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```