Problem Set 1

Yifan Li

Problem 1

a

Using `read.table` to read a file written in txt format. For the separation, using ',' Then according to the description file, 'wine.names', there are 14 attributes in the data file with a class number listed in the first column. So i add `col.names` in the code `read.table`. Such that, I can produce a `data.frame` object with appropriate columns names.

```
wines_data <- read.table("wine.data", sep = ",", col.names=c('class_number', 'Alcohol', 'Mali
```

b

First, using `wines_data['class_number]==i` for `i in [1, 2, 3]` to create a new `data.frame` that has `True` only if the class numbers match with `i`. After that, using a `sum` function to compute the number of `True`, which is the number of the wine class.

```
num_class_1 <- sum(wines_data['class_number'] == 1)
num_class_2 <- sum(wines_data['class_number'] == 2)
num_class_3 <- sum(wines_data['class_number'] == 3)
```

The results are:

```
num_class_1
```

```
[1] 59
```

```
num_class_2
```

```
[1] 71
```

```
num_class_3
```

```
[1] 48
```

So, the number of wines within each class is correct as reported in the file "wine.names".

c

1. The correlation between alcohol content and color intensity can be derived from a function `cor`. The alcohol content has variable name `Alcohol`, the color intensity has variable name `Color_intensity`. So the input of the function will be:

```
cor(wines_data['Alcohol'], wines_data['Color_intensity'])
```

```
        Color_intensity
Alcohol       0.5463642
```

2. For each class, first the whole data from that class will be extracted, then the correlation between alcohol content and color intensity will be calculated.

For class one:

```
class_one <- wines_data[wines_data['class_number'] == 1, ]
class_one_cor <- cor(class_one['Alcohol'], class_one['Color_intensity'])
class_one_cor
```

```
        Color_intensity
Alcohol       0.4082913
```

For class two:

```
class_two <- wines_data[wines_data['class_number'] == 2, ]
class_two_cor <- cor(class_two['Alcohol'], class_two['Color_intensity'])
class_two_cor
```

```
        Color_intensity
Alcohol       0.2697891
```

For class three:

```
class_three <- wines_data[wines_data['class_number'] == 3, ]
class_three_cor <- cor(class_three['Alcohol'], class_three['Color_intensity'])
class_three_cor
```

```
        Color_intensity
Alcohol       0.3503777
```

Through comparison, one will find that class one has the highest correlation which is `0.4082913`, while class two has the lowest correlation which is `0.2697891`.

3. To find the wine with highest color intensity, using **`which.max`** function, with attributes **`wines_data$Color_intensity`**. This will yield the index of the wine with highest color intensity. Then using this index to find the wine, after that extract its alcohol content.

```
index <- which.max(wines_data$Color_intensity)
target_wine <- wines_data[index, ]
target_wine$Alcohol
```

```
[1] 14.34
```

Finally extract the alcohol content from the target wine, which is `14.34`.

4. First, find the number of wines that have a higher content of proanthocyanins than ash. Then divide it by the sum of three classes of wines, which will give us the percentage of wines had a higher content of proanthocyanins compare to ash, which is `8.426966%`.

```
num <- sum(wines_data$'Proanthocyanins' > wines_data$'Ash')
percentage <- num * 100 / (num_class_1 + num_class_2 + num_class_3)
percentage
```

```
[1] 8.426966
```

d

```r
average_table <- data.frame(id = 1: 4,
                            class_number = c('overall', '1', '2', '3'),

                            Mean_Alcohol = c(mean(wines_data$Alcohol), mean(class_one$Alcohol

                            Mean_Malic_acid = c(mean(wines_data$Malic_acid), mean(class_one$M

                            Mean_Ash = c(mean(wines_data$Ash), mean(class_one$Ash), mean(clas

                            Mean_Alcalinity_of_ash = c(mean(wines_data$Alcalinity_of_ash), me

                            Mean_Magnesium = c(mean(wines_data$Magnesium), mean(class_one$Mag

                            Mean_Total_phenols = c(mean(wines_data$Total_phenols), mean(class

                            Mean_Flavanoids = c(mean(wines_data$Flavanoids), mean(class_one$M

                            Mean_Nonflavanoid_phenols = c(mean(wines_data$Nonflavanoid_phenol

                            Mean_Proanthocyanins = c(mean(wines_data$Proanthocyanins), mean(c

                            Mean_Color_intensity = c(mean(wines_data$Color_intensity), mean(c

                            Mean_Hue = c(mean(wines_data$Hue), mean(class_one$Hue), mean(clas
                            Mean_OD280_OD315_of_diluted_wines = c(mean(wines_data$OD280_OD315

                            Mean_Proline = c(mean(wines_data$Proline), mean(class_one$Proline
average_table
```

```
  id class_number Mean_Alcohol Mean_Malic_acid Mean_Ash Mean_Alcalinity_of_ash
1  1      overall     13.00062        2.336348 2.366517               19.49494
2  2            1     13.74475        2.010678 2.455593               17.03729
3  3            2     12.27873        1.932676 2.244789               20.23803
4  4            3     13.15375        3.333750 2.437083               21.41667
  Mean_Magnesium Mean_Total_phenols Mean_Flavanoids Mean_Nonflavanoid_phenols
1       99.74157           2.295112       2.0292697                 0.3618539
2      106.33898           2.840169       2.9823729                 0.2900000
3       94.54930           2.258873       2.0808451                 0.3636620
4       99.31250           1.678750       0.7814583                 0.4475000
  Mean_Proanthocyanins Mean_Color_intensity  Mean_Hue
1             1.590899             5.058090 0.9574494
```

```
2             1.899322             5.528305 1.0620339
3             1.630282             3.086620 1.0562817
4             1.153542             7.396250 0.6827083
  Mean_OD280_OD315_of_diluted_wines Mean_Proline
1                         2.611685     746.8933
2                         3.157797    1115.7119
3                         2.785352     519.5070
4                         1.683542     629.8958
```

e

For manually conducting the t-test. Since there are three different classes, one will need to do 3 comparisons, class 1 vs. class 2, class 1 vs class 3 and class 2 vs class 3.

Firstly, extracting the data of level of phenols of each classes:

```
class_one_phenols <- class_one['Total_phenols']
class_two_phenols <- class_two['Total_phenols']
class_three_phenols <- class_three['Total_phenols']
```

Secondly, calculating the mean, variance for each groups:

```
mean_one <- mean(class_one_phenols[,])
mean_one
```

```
[1] 2.840169
```

```
variance_one <- var(class_one_phenols[,])
variance_one
```

```
[1] 0.1148948
```

```
mean_two <- mean(class_two_phenols[,])
mean_two
```

```
[1] 2.258873
```

```
variance_two <- var(class_two_phenols[,])
variance_two
```

```
[1] 0.2974187
```

```
mean_three <- mean(class_three_phenols[,])
mean_three
```

```
[1] 1.67875
```

```
variance_three <- var(class_three_phenols[,])
variance_three
```

```
[1] 0.1274282
```

For different comparisons, first compute the t-statistics with formula: $t = \frac{(\hat{X}_1 - \hat{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, where $\hat{X}_1$ and $\hat{X}_2$ are the sample means, S_1^2 and S_2^2 are the sample variances, n_1 and n_2 are the sizes.

```
t_1_2 <- 1
t_1_3 <- 2
t_2_3 <- 3
```