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 adding `col.names` in the code `read.table`. Such that, one can produce a `data.frame` object with appropriate columns names.

```r
wines_data <- read.table("wine.data",
                         sep = ",",
                         col.names=c('class_number',
                                     'Alcohol',
                                     'Malic_acid',
                                     'Ash',
                                     'Alcalinity_of_ash',
                                     'Magnesium',
                                     'Total_phenols',
                                     'Flavanoids',
                                     'Nonflavanoid_phenols',
                                     'Proanthocyanins',
                                     'Color_intensity',
                                     'Hue',
                                     'OD280_OD315_of_diluted_wines',
                                     'Proline'))
```

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.

```r
num_class_one <- sum(wines_data['class_number'] == 1)
num_class_two <- sum(wines_data['class_number'] == 2)
num_class_three <- sum(wines_data['class_number'] == 3)
```

The results are:

```r
num_class_one
```

```
[1] 59
```

```r
num_class_two
```

```
[1] 71
```

```r
num_class_three
```

```
[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:

```r
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_one + num_class_two + num_class_three)
percentage
```

```
[1] 8.426966
```

d

```
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(class_two$Alcohol),
                                             mean(class_three$Alcohol)),

                            Mean_Malic_acid = c(mean(wines_data$Malic_acid),
                                                mean(class_one$Malic_acid),
                                                mean(class_two$Malic_acid),
                                                mean(class_three$Malic_acid)),

                            Mean_Ash = c(mean(wines_data$Ash),
                                         mean(class_one$Ash),
                                         mean(class_two$Ash),
                                         mean(class_three$Ash)),

                            Mean_Alcalinity_of_ash = c(
                              mean(wines_data$Alcalinity_of_ash),
                              mean(class_one$Alcalinity_of_ash),
                              mean(class_two$Alcalinity_of_ash),
                              mean(class_three$Alcalinity_of_ash)),

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

```
                              mean(class_two$Magnesium),
                              mean(class_three$Magnesium)),

Mean_Total_phenols = c(mean(wines_data$Total_phenols),
                        mean(class_one$Total_phenols),
                        mean(class_two$Total_phenols),
                        mean(class_three$Total_phenols)),

Mean_Flavanoids = c(mean(wines_data$Flavanoids),
                     mean(class_one$Flavanoids),
                     mean(class_two$Flavanoids),
                     mean(class_three$Flavanoids)),

Mean_Nonflavanoid_phenols = c(
  mean(wines_data$Nonflavanoid_phenols),
  mean(class_one$Nonflavanoid_phenols),
  mean(class_two$Nonflavanoid_phenols),
  mean(class_three$Nonflavanoid_phenols)),

Mean_Proanthocyanins = c(mean(wines_data$Proanthocyanins),
                          mean(class_one$Proanthocyanins),
                          mean(class_two$Proanthocyanins),
                          mean(class_three$Proanthocyanins)),

Mean_Color_intensity = c(mean(wines_data$Color_intensity),
                          mean(class_one$Color_intensity),
                          mean(class_two$Color_intensity),
                          mean(class_three$Color_intensity)),

Mean_Hue = c(mean(wines_data$Hue), mean(class_one$Hue),
              mean(class_two$Hue), mean(class_three$Hue)),

Mean_OD280_OD315_of_diluted_wines = c(
    mean(wines_data$OD280_OD315_of_diluted_wines),
    mean(class_one$OD280_OD315_of_diluted_wines),
    mean(class_two$OD280_OD315_of_diluted_wines),
    mean(class_three$OD280_OD315_of_diluted_wines)),

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

```r
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

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:

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

For existing R function.

```r
t_test_1_2 <- t.test(class_one_phenols, class_two_phenols)
t_test_1_2
```

```
    Welch Two Sample t-test
```

```
data:  class_one_phenols and class_two_phenols
t = 7.4206, df = 119.14, p-value = 1.889e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4261870 0.7364055
sample estimates:
mean of x mean of y 
 2.840169  2.258873 
```

```r
t_test_1_3 <- t.test(class_one_phenols, class_three_phenols)
t_test_1_3
```

```

    Welch Two Sample t-test

data:  class_one_phenols and class_three_phenols
t = 17.12, df = 98.356, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.026801 1.296038
sample estimates:
mean of x mean of y 
 2.840169  1.678750 
```

```r
t_test_2_3 <- t.test(class_two_phenols, class_three_phenols)
t_test_2_3
```

```

    Welch Two Sample t-test

data:  class_two_phenols and class_three_phenols
t = 7.0125, df = 116.91, p-value = 1.622e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4162855 0.7439610
sample estimates:
mean of x mean of y 
 2.258873  1.678750 
```

For manually conducting the t-test.

Then, 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, assuming that the variances are different, 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, μ_1 and μ_2 are the means, S_1^2 and S_2^2 are the sample variances, n_1 and n_2 are the sizes.

Since the null hypothesis is that there is no difference between each class, $\mu_1 - \mu_2 = 0$, thus the t-statistics are:

```
t_1_2 <- (mean_one - mean_two) /
  (sqrt((variance_one / num_class_one) + variance_two / num_class_two))
t_1_2
```

```
[1] 7.420649
```

```
t_1_3 <- (mean_one - mean_three) /
  (sqrt((variance_one / num_class_one) + variance_three / num_class_three))
t_1_3
```

```
[1] 17.12025
```

```
t_2_3 <- (mean_two - mean_three) /
  (sqrt((variance_two / num_class_two) + variance_three / num_class_three))
t_2_3
```

```
[1] 7.012505
```

Next the degrees of freedom are defined as $\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)}{n_2 - 1}}$, then rounding it down to find the degree of freedom. The results are:

```
nu_1_2 <- floor(
  (variance_one / num_class_one + variance_two / num_class_two)^2 /
    ((variance_one / num_class_one)^2 / (num_class_one - 1) +
       (variance_two / num_class_two)^2 / (num_class_two - 1))
  )
nu_1_2
```

```
[1] 119
```

```
nu_1_3 <- floor(
  (variance_one / num_class_one + variance_three / num_class_three)^2 /
    ((variance_one / num_class_one)^2 / (num_class_one - 1) +
       (variance_three / num_class_three)^2 / (num_class_three - 1))
  )
nu_1_3
```

```
[1] 98
```

```r
nu_2_3 <- floor(
  (variance_two / num_class_two + variance_three / num_class_three)^2 /
    ((variance_two / num_class_two)^2 / (num_class_two - 1) +
       (variance_three / num_class_three)^2 / (num_class_three - 1))
  )
nu_2_3
```

```
[1] 116
```

Define a function to manually compute the p-value of give t-statistics and degree of freedom:

```r
compute_two_tail_p_value <- function(t_statistics, df){
  #inputs: t_statistics : the t-statistics, df: the degree of freedom
  #outputs: the p-value
  t_pdf <- function(x, df){
    # inputs x: variable, df: the degree of freedom
    # output the value of the probability density function
    return(gamma((df+1)/2) / (sqrt(df*pi) * gamma(df/2)) * (1 + (x^2)/df)^(-(df+1)/2))
  }

  p_value_two_tail <- 2 * integrate(t_pdf, t_statistics, Inf, df = df)$value

  return(p_value_two_tail)
}
p_1_2 <- compute_two_tail_p_value(t_1_2, nu_1_2)
p_1_2
```

```
[1] 1.897952e-11
```

```r
p_1_3 <- compute_two_tail_p_value(t_1_3, nu_1_3)
p_1_3
```

```
[1] 3.267661e-31
```

```r
p_2_3 <- compute_two_tail_p_value(t_2_3, nu_2_3)
p_2_3
```

```
[1] 1.664716e-10
```

Through calculation, one can observe that the p-values of all three comparisons are extremely small. Thus one can argue that there is extremely strong evidence against the null hypothesis for each pairwise comparison. The differences in phenol levels between all the classes are statistically significant.

Problem 2

a

Import the data as `raw_table`.

```r
raw_table <- read.table("AskAManager.csv", sep = ",", header = TRUE)
head(raw_table)
```

```
  X          Timestamp How.old.are.you. What.industry.do.you.work.in.
1 1 4/27/2021 11:02:10            25-34  Education (Higher Education)
2 2 4/27/2021 11:02:22            25-34             Computing or Tech
3 3 4/27/2021 11:02:38            25-34 Accounting, Banking & Finance
4 4 4/27/2021 11:02:41            25-34                    Nonprofits
5 5 4/27/2021 11:02:42            25-34 Accounting, Banking & Finance
6 6 4/27/2021 11:02:46            25-34  Education (Higher Education)
                                 Job.title
1       Research and Instruction Librarian
2 Change & Internal Communications Manager
3                     Marketing Specialist
4                          Program Manager
5                       Accounting Manager
6           Scholarly Publishing Librarian
  If.your.job.title.needs.additional.context..please.clarify.here.
1
2
3
4
5
6
  What.is.your.annual.salary...You.ll.indicate.the.currency.in.a.later.question..If.you.are.p
1
2
3
4
5
```

```
6
  How.much.additional.monetary.compensation.do.you.get..if.any..for.example..bonuses.or.overt
1
2
3
4
5
6
  Please.indicate.the.currency If..Other...please.indicate.the.currency.here..
1                          USD
2                          GBP
3                          USD
4                          USD
5                          USD
6                          USD
  If.your.income.needs.additional.context..please.provide.it.here.
1
2
3
4
5
6
  What.country.do.you.work.in.
1               United States
2              United Kingdom
3                          US
4                         USA
5                          US
6                         USA
  If.you.re.in.the.U.S...what.state.do.you.work.in. What.city.do.you.work.in.
1                                     Massachusetts                    Boston
2                                                                   Cambridge
3                                         Tennessee               Chattanooga
4                                         Wisconsin                 Milwaukee
5                                    South Carolina                Greenville
6                                     New Hampshire                   Hanover
  How.many.years.of.professional.work.experience.do.you.have.overall.
1                                                           5-7 years
2                                                        8 - 10 years
3                                                         2 - 4 years
4                                                        8 - 10 years
5                                                        8 - 10 years
6                                                        8 - 10 years
```

```
  How.many.years.of.professional.work.experience.do.you.have.in.your.field.
1                                                                 5-7 years
2                                                                 5-7 years
3                                                               2 - 4 years
4                                                                 5-7 years
5                                                                 5-7 years
6                                                               2 - 4 years
  What.is.your.highest.level.of.education.completed. What.is.your.gender.
1                                    Master's degree                  Woman
2                                     College degree             Non-binary
3                                     College degree                  Woman
4                                     College degree                  Woman
5                                     College degree                  Woman
6                                    Master's degree                    Man
  What.is.your.race...Choose.all.that.apply..
1                                       White
2                                       White
3                                       White
4                                       White
5                                       White
6                                       White
```

b

In order to clean up the variable names, a rename will be conducted. The new variable names will be `id`, `timestamp`, `age`, `work_industry`, `job`, `job_context`, `annual_salary`, `compensation`, `currency`, `other_currency`, `income_context`, `country`, `state`, `city`, `overall_work_years`, `specific_work_years`, `education`, `gender`, `race`.

```r
colnames(raw_table) <- c('id',
                         'timestamp',
                         'age',
                         'work_industry',
                         'job',
                         'job_context',
                         'annual_salary',
                         'compensation',
                         'currency',
                         'other_currency',
                         'income_context',
                         'country', 'state',
                         'city', 'overall_work_years',
```

```r
                         'specific_work_years',
                         'education',
                         'gender',
                         'race')
head(raw_table)
```

```
  id           timestamp   age                 work_industry
1  1 4/27/2021 11:02:10 25-34  Education (Higher Education)
2  2 4/27/2021 11:02:22 25-34             Computing or Tech
3  3 4/27/2021 11:02:38 25-34 Accounting, Banking & Finance
4  4 4/27/2021 11:02:41 25-34                   Nonprofits
5  5 4/27/2021 11:02:42 25-34 Accounting, Banking & Finance
6  6 4/27/2021 11:02:46 25-34  Education (Higher Education)
                                        job job_context annual_salary
1       Research and Instruction Librarian                      55000
2 Change & Internal Communications Manager                      54600
3                     Marketing Specialist                      34000
4                          Program Manager                      62000
5                       Accounting Manager                      60000
6           Scholarly Publishing Librarian                      62000
  compensation currency other_currency income_context        country
1            0      USD                                United States
2         4000      GBP                               United Kingdom
3           NA      USD                                           US
4         3000      USD                                          USA
5         7000      USD                                           US
6           NA      USD                                          USA
          state        city overall_work_years specific_work_years
1  Massachusetts      Boston          5-7 years           5-7 years
2                  Cambridge       8 - 10 years           5-7 years
3     Tennessee Chattanooga        2 - 4 years         2 - 4 years
4     Wisconsin   Milwaukee       8 - 10 years           5-7 years
5 South Carolina  Greenville       8 - 10 years           5-7 years
6  New Hampshire     Hanover       8 - 10 years         2 - 4 years
        education     gender  race
1 Master's degree      Woman White
2  College degree Non-binary White
3  College degree      Woman White
4  College degree      Woman White
5  College degree      Woman White
6 Master's degree        Man White
```

c

In order to restrict the data to those being paid in USD, a logistical judgment has been down, which will yield the index of entries whose currency is USD or they have USD as their `other_currency`. After that, using mask to get the restricted table which is `usd_table`.

```
usd_table <- raw_table[raw_table['currency'] == 'USD'
                       | raw_table['other_currency'] == 'USD', ]
head(usd_table)
```

```
  id          timestamp   age                  work_industry
1  1 4/27/2021 11:02:10 25-34  Education (Higher Education)
3  3 4/27/2021 11:02:38 25-34 Accounting, Banking & Finance
4  4 4/27/2021 11:02:41 25-34                    Nonprofits
5  5 4/27/2021 11:02:42 25-34 Accounting, Banking & Finance
6  6 4/27/2021 11:02:46 25-34  Education (Higher Education)
7  7 4/27/2021 11:02:51 25-34                    Publishing
                                  job job_context annual_salary compensation
1 Research and Instruction Librarian                       55000            0
3               Marketing Specialist                       34000           NA
4                    Program Manager                       62000         3000
5                 Accounting Manager                       60000         7000
6    Scholarly Publishing Librarian                       62000           NA
7               Publishing Assistant                       33000         2000
  currency other_currency income_context       country          state
1      USD                              United States  Massachusetts
3      USD                                         US      Tennessee
4      USD                                        USA      Wisconsin
5      USD                                         US South Carolina
6      USD                                        USA  New Hampshire
7      USD                                        USA South Carolina
         city overall_work_years specific_work_years       education gender
1      Boston          5-7 years           5-7 years Master's degree  Woman
3 Chattanooga        2 - 4 years         2 - 4 years  College degree  Woman
4   Milwaukee       8 - 10 years           5-7 years  College degree  Woman
5  Greenville       8 - 10 years           5-7 years  College degree  Woman
6     Hanover       8 - 10 years         2 - 4 years Master's degree    Man
7    Columbia        2 - 4 years         2 - 4 years  College degree  Woman
   race
1 White
3 White
4 White
```

```
5 White
6 White
7 White
```

For the number of observation:

```r
total_num <- nrow(raw_table)
total_num
```

```
[1] 28062
```

```r
usd_num <- nrow(usd_table)
usd_num
```

```
[1] 23382
```

```r
diff_num <- total_num - usd_num
diff_num
```

```
[1] 4680
```

By restricting the data to those being paid in USD, the number of observations decreases by `4680`.

d

Assume everyone starts working at least they are 18. The impossible entry is that the maximum possible value of its age minus the lowest value in its years of experience in their field, and years of experience total respectively. If the result smaller than 18, this entry will be seen as impossible.

```r
larger_age <- unlist(lapply(usd_table$age,
                     function(x) max(
                       as.numeric(
                         unlist(
                           regmatches(
                             x, gregexpr("\\d+", x)))))))

smaller_overall_work <- unlist(lapply(usd_table$overall_work_years,
```

```r
                     function(x) min(
                       as.numeric(
                         unlist(
                           regmatches(
                             x, gregexpr("\\d+", x)))))))

smaller_specific_work <- unlist(lapply(usd_table$specific_work_years,
                     function(x) min(
                       as.numeric(
                         unlist(
                           regmatches(
                             x, gregexpr("\\d+", x)))))))
```

Thus the impossible index are as following, where `TRUE` means impossible.

```r
overall_diff <- larger_age - smaller_overall_work
specific_diff <- larger_age - smaller_specific_work
overall_impossible <- overall_diff < 18
specific_impossible <- specific_diff < 18
impossible_index <- overall_impossible | specific_impossible
head(impossible_index)
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE
```

Then the cleaned table is:

```r
possible_usd_table <- usd_table[!impossible_index, ]
head(possible_usd_table)
```

```
  id           timestamp   age                 work_industry
1  1 4/27/2021 11:02:10 25-34  Education (Higher Education)
3  3 4/27/2021 11:02:38 25-34 Accounting, Banking & Finance
4  4 4/27/2021 11:02:41 25-34                    Nonprofits
5  5 4/27/2021 11:02:42 25-34 Accounting, Banking & Finance
6  6 4/27/2021 11:02:46 25-34  Education (Higher Education)
7  7 4/27/2021 11:02:51 25-34                    Publishing
                              job job_context annual_salary compensation
1 Research and Instruction Librarian                   55000            0
3              Marketing Specialist                    34000           NA
4                   Program Manager                    62000         3000
```

```
5                Accounting Manager                         60000        7000
6     Scholarly Publishing Librarian                        62000          NA
7             Publishing Assistant                          33000        2000
  currency other_currency income_context       country          state
1      USD                              United States  Massachusetts
3      USD                                         US      Tennessee
4      USD                                        USA      Wisconsin
5      USD                                         US South Carolina
6      USD                                        USA  New Hampshire
7      USD                                        USA South Carolina
         city overall_work_years specific_work_years       education gender
1      Boston          5-7 years           5-7 years Master's degree  Woman
3 Chattanooga        2 - 4 years         2 - 4 years  College degree  Woman
4   Milwaukee       8 - 10 years           5-7 years  College degree  Woman
5  Greenville       8 - 10 years           5-7 years  College degree  Woman
6     Hanover       8 - 10 years         2 - 4 years Master's degree    Man
7    Columbia        2 - 4 years         2 - 4 years  College degree  Woman
   race
1 White
3 White
4 White
5 White
6 White
7 White
```

For the number of observations:

```r
possible_num <- nrow(possible_usd_table)
possible_num
```

```
[1] 23321
```

```r
diff_possible_num <- usd_num - possible_num
diff_possible_num
```

```
[1] 61
```

By restricting the data to those being paid in USD, the number of observations decreases by 61.

e

In this section, the IQR(interquartile range) will be used to identify the outliers, which means that the data fall below $Q1 - 1.5$ IQR or above $Q3 + 1.5$ IQR will be considered as outliers, then removed.

First, sorting the salary in ascending order, then calculating the Q1 and Q3. Finally, using $Q3 - Q1$ to get IQR.

```r
sorted_salary <- sort(possible_usd_table$annual_salary)
Q_1 <- (sorted_salary[floor(1 + (possible_num - 1) / 4)] +
          sorted_salary[ceiling(1 + (possible_num - 1) / 4)]) / 2
Q_3 <- (sorted_salary[floor(1 + (possible_num - 1) * 3 / 4)] +
          sorted_salary[ceiling(1 + (possible_num - 1) * 3 / 4)]) / 2
IQR <- Q_3 - Q_1
IQR
```

```
[1] 55840
```

Then one can use this IQR to find the outliers:

```r
min_salary <- Q_1 - 1.5 * IQR
max_salary <- Q_3 + 1.5 * IQR
final_table <- possible_usd_table[
  possible_usd_table['annual_salary'] >= min_salary &
    possible_usd_table['annual_salary'] <= max_salary, ]
head(final_table)
```

```
  id          timestamp   age                 work_industry
1  1 4/27/2021 11:02:10 25-34  Education (Higher Education)
3  3 4/27/2021 11:02:38 25-34 Accounting, Banking & Finance
4  4 4/27/2021 11:02:41 25-34                    Nonprofits
5  5 4/27/2021 11:02:42 25-34 Accounting, Banking & Finance
6  6 4/27/2021 11:02:46 25-34  Education (Higher Education)
7  7 4/27/2021 11:02:51 25-34                    Publishing
                                job job_context annual_salary compensation
1 Research and Instruction Librarian                     55000            0
3              Marketing Specialist                     34000           NA
4                   Program Manager                     62000         3000
5                Accounting Manager                     60000         7000
6     Scholarly Publishing Librarian                     62000           NA
7              Publishing Assistant                     33000         2000
```

```
  currency other_currency income_context       country          state
1      USD                                United States  Massachusetts
3      USD                                           US      Tennessee
4      USD                                          USA      Wisconsin
5      USD                                           US South Carolina
6      USD                                          USA  New Hampshire
7      USD                                          USA South Carolina
         city overall_work_years specific_work_years       education gender
1      Boston          5-7 years           5-7 years Master's degree  Woman
3 Chattanooga        2 - 4 years         2 - 4 years  College degree  Woman
4   Milwaukee       8 - 10 years           5-7 years  College degree  Woman
5  Greenville       8 - 10 years           5-7 years  College degree  Woman
6     Hanover       8 - 10 years         2 - 4 years Master's degree    Man
7    Columbia        2 - 4 years         2 - 4 years  College degree  Woman
   race
1 White
3 White
4 White
5 White
6 White
7 White
```

For the final sample size:

```r
final_num <- nrow(final_table)
final_num
```

```
[1] 22407
```