cs412_assignment1

1.

```
col_names = c("id","midterm","final")
df = read.csv("data.online.scores.txt",header = FALSE,sep = "\t",
              col.names = col_names)
```

a.

max() function displays the maximum number in data

```
max(df['midterm'])
```

```
## [1] 100
```

min() function displays the minimum number in data

```
min(df['midterm'])
```

```
## [1] 37
```

b

```
x = df[['midterm']]
quantile(x)
```

```
##   0%  25%  50%  75% 100%
##   37   68   77   87  100
```

First quantile is 68, median is 77, third quantile is 87.

c

mean() is calculating the average

```
m = mean(x)
m
```

```
## [1] 76.715
```

d used table() to make a table of the number of times that each data point appear in the whole dataset, and subset the largest data point

```
t = table(x)
t[t==max(t)]
```

```
## x
## 77 83
## 37 37
```

The mode are 77 and 83.

e var() is calling sample variance

```
variance = var(x)
round(variance,3)
```

```
## [1] 173.279
```

The variance is 173.279 based on the formula

$$s = 1/(n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2.

a

```
var_before = var(x)
x_norm = (x-m)/sqrt(var_before)
var_after = var(x_norm)
round(var_before,3)
```

```
## [1] 173.279
```

```
round(var_after,3)
```

```
## [1] 1
```

After normalization, the variance is one compared to the variance before normalization is 173.2791.

The equation of normalization is

$$z = \frac{x - \bar{x}}{\sqrt{\sigma^2}}$$

b

```
round((90-m)/sqrt(var_before),3)
```

```
## [1] 1.009
```

c

```
x2 = df[["final"]]
correlation = cor(x,x2)
round(correlation,3)
```

```
## [1] 0.544
```

The formula of Pearson's correlation coefficient is

$$r_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2)}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

d

```
covariance = cov(x,x2)
round(covariance,3)
```

```
## [1] 78.254
```

The formula of covariance is

$$cov_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

3

a

```
round(58/(120+2+58),3)
```

```
## [1] 0.322
```

The formula of Jaccard coefficient is

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

b

```
df1 = read.table("data.libraries.inventories.txt",sep = "\t",header = TRUE)
df1$library = NULL
rownames(df1) = c('Citadel','Castle')
```

1. h = 1

dist() function using the specified distance measure to calculates and returns the distance matrix

```
dist(df1,method = "manhattan")
```

```
##        Citadel
## Castle    6152
```

2. h = 2

```
round(dist(df1,method = "euclidean"),3)
```

```
##        Citadel
## Castle 715.328
```

3. h = $-\infty$

```
dist(df1,method = "maximum")
```

```
##        Citadel
## Castle     170
```

c

```
cml = df1[1,]
cbl = df1[2,]
cosine = sum(cml*cbl)/(sqrt(sum(cml^2))*sqrt(sum(cbl^2)))
round(cosine,3)
```

```
## [1] 0.841
```

The formula of cosine similarity is

$$cos(d_1, d_2) = \frac{d_1 * d_2}{||d_1|| * ||d_2||}$$

d

```
epsi = 0.0001
require(flexmix)
```

```
## Loading required package: flexmix
```

```
## Loading required package: lattice
```

```
cml1 = as.numeric(as.vector(df1[1,]))
cbl1 = as.numeric(as.vector(df1[2,]))
y = cbind(cml1,cbl1)
kl = KLdiv(y)
round(kl[1,2],3)
```

```
## [1] 0.207
```

Or by formula

$$D_{KL}(p(x)||q(x)) = \sum p(x) ln \frac{p(x)}{q(x)}$$

```
c1 = sum(cml1)
c2 = sum(cbl1)
round(sum(cml1/c1*log((cml1/c1/(cbl1/c2)))),3)
```

```
## [1] 0.207
```

4.

By the formula of chi-square correlation

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

```r
beer = matrix(c(150, 40, 15, 3300),ncol = 2,byrow = TRUE)
beer
```

```
##      [,1] [,2]
## [1,]  150   40
## [2,]   15 3300
```

```r
row.names(beer) = c("buy beer","do not buy beer")
colnames(beer) = c("buy diaper","do not buy diaper")
sums = 150+40+15+3300
E_b_d = (150+40)*(150+15)/sums
E_b_nd = (150+40)*(40+3300)/sums
E_nb_d = (15 + 3300)*(150+15)/sums
E_nb_nd = (15 + 3300)*(40+3300)/sums
chi = (150-E_b_d)^2/E_b_d + (40-E_b_nd)^2/E_b_nd +
  (15 - E_nb_d)^2/E_nb_d + (3300-E_nb_nd)^2/E_nb_nd
chi
```

```
## [1] 2468.183
```

Since value of chi-square test is large, we reject the hypothesis that the two variables are independent to each other.