Question 1

a)

$H0$: All means are equal
$H1$: At least mean of one sample is different.

Since the p-value is less than 0.05, the difference in means is statistically significant.

Hence, there is a difference in the tensile strength between the four varieties.

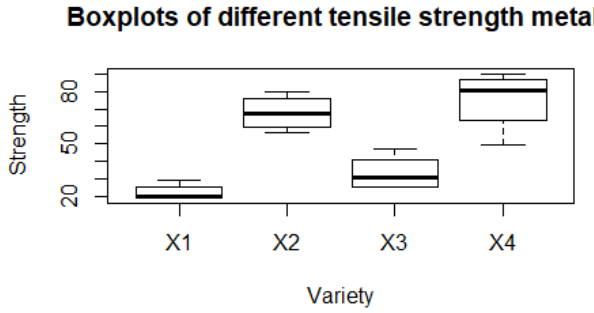

Figure 1: BoxPlot

Output 1: Q1(a) R Code Output

```
> tensileData <- read.table("AirCondition.txt", header=TRUE)
> data <- tensileData %>%gather(Sample,Strength,X1:X4)
> factorData <- factor(data$Sample)
>
> analysis <- aov(data$Strength ~ factorData)
> summary(analysis)
             Df Sum Sq Mean Sq F value   Pr(>F)
factorData    3   7978  2659.4   19.04 7.4e-05 ***
Residuals    12   1676   139.6
```

b)

VCA Component1: 629.94
VCA Component2: 139.65

Output 2: Q1(b) R Code Output

```
> data$Sample <- as.factor(data$Sample)
> anova_vca <- fitVCA(Strength ~ Sample, data,"anova")
> anova_vca
```

```
Result Variance Component Analysis:

  Name    DF         SS          MS           VC          %Total
total   3.994851                              769.583333  100
Sample  3          7978.1875   2659.395833   629.9375    50.768254
error   12         1675.75     139.645833    139.645833  18.145642

  SD          CV[%]
27.741365   56.114013
81.854358   25.098556
11.817184   23.90328
Mean: 49.4375 (N = 16)
```

Question 2

a)

$H0$: All means are equal

$H1$: At least mean of one sample is different.

Since the p-value is less than 0.05, the difference in means is statistically significant.

Hence, there is a difference in the diameter of the organism in different mediums.

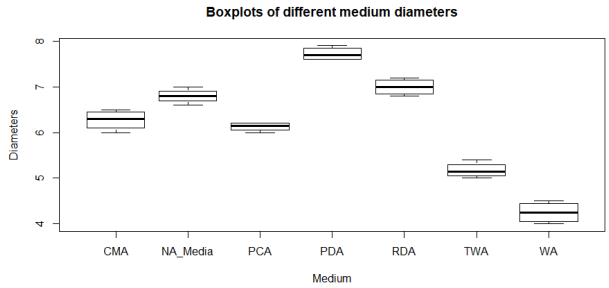

Figure 2: BoxPlot

Output 3: Q2(a) R Code Output

```
> organism_data <- read.csv("OrganismData.csv")
> factors <- factor(organism_data$Medium)
> anova_analysis <- aov(organism_data$Diameters ~ factors)
> summary(anova_analysis)
            Df Sum Sq Mean Sq F value Pr(>F)
factors      6  32.76   5.460   168.6 <2e-16 ***
Residuals   21   0.68   0.032
```

b)

This exercise is a Post hoc comparison because we want to analyze and compare difference of diameter means in different mediums after knowing from Anova test that at least one of the means are different from the other.

Output 4: Q2(b) R Code Output

```
> posthoc <- TukeyHSD(x=anova_analysis, "factors", conf.level=0.95)
> posthoc
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = organism_data$Diameters ~ factors)

$factors
                      diff        lwr        upr     p adj
NA_Media-CMA         0.525  0.1113645  0.9386355 0.0073956
PCA-CMA             -0.150 -0.5636355  0.2636355 0.8943879
PDA-CMA              1.450  1.0363645  1.8636355 0.0000000
RDA-CMA              0.725  0.3113645  1.1386355 0.0002062
TWA-CMA             -1.100 -1.5136355 -0.6863645 0.0000004
WA-CMA              -2.025 -2.4386355 -1.6113645 0.0000000
PCA-NA_Media        -0.675 -1.0886355 -0.2613645 0.0005017
PDA-NA_Media         0.925  0.5113645  1.3386355 0.0000068
RDA-NA_Media         0.200 -0.2136355  0.6136355 0.7004477
TWA-NA_Media        -1.625 -2.0386355 -1.2113645 0.0000000
WA-NA_Media         -2.550 -2.9636355 -2.1363645 0.0000000
PDA-PCA              1.600  1.1863645  2.0136355 0.0000000
RDA-PCA              0.875  0.4613645  1.2886355 0.0000156
TWA-PCA             -0.950 -1.3636355 -0.5363645 0.0000045
WA-PCA              -1.875 -2.2886355 -1.4613645 0.0000000
RDA-PDA             -0.725 -1.1386355 -0.3113645 0.0002062
TWA-PDA             -2.550 -2.9636355 -2.1363645 0.0000000
WA-PDA              -3.475 -3.8886355 -3.0613645 0.0000000
TWA-RDA             -1.825 -2.2386355 -1.4113645 0.0000000
WA-RDA              -2.750 -3.1636355 -2.3363645 0.0000000
WA-TWA              -0.925 -1.3386355 -0.5113645 0.0000068
```

Tuckey method declares that all means are different except for PCA-CMA and RDA-NA since their p value > 0.05

Question 3

a)

H0 : beta1 = 0
H1: beta1 $\neq$ 0

p value < 0.05, we reject the NULL hypothesis. Hence, there is a significant relationship between the two variables.

Output 5: Q3(a) R Code Output

```
> r_squared <- 0.18
> beta_1 <- -0.62
> n = 178
> f = ((n-2)*r_squared)/(1-r_squared)
> f
[1] 38.63415
> pf(f,1,n-2,lower.tail = F)
[1] 3.598011e-09
```

b)

From part a we know that the two variables are significant. We have beta1 = -0.62, which is a negative relation. Hence for 1 unit increase in neighborhood social disorder, satisfaction with police declines by 0.62.

Question 4

a)

H0: Beta1 = 0
H1: Beta1 $\neq$ 0

Regression Equation: y = 6.50 + 1.50x
For every unit change in x , y is increased by 1.50 units.

Result: Since p(0.02135) < 0.05, we reject the null hypothesis that beta1 = 0. Hence, there is a significant relationship between the variables.

Output 6: Q4 (a) R Code Output

```
> myData <- read.csv("q4.csv",header = TRUE)
> colnames(myData) <- c("Y","X")
> model <- lm(Y ~ X ,data = myData)
> summary(model)

Call:
lm(formula = Y ~ X, data = myData)
```

```
Residuals:
   Min      1Q  Median      3Q     Max
 -3.00   -1.75    0.00    1.00    4.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5000     0.5669  11.465 8.03e-08 ***
X             1.5000     0.5669   2.646   0.0213 *

Residual standard error: 2.121 on 12 degrees of freedom
Multiple R-squared:  0.3684,    Adjusted R-squared:  0.3158
F-statistic:      7 on 1 and 12 DF,  p-value: 0.02135
```

b)

The value of x = 1 or x = -1 makes it a categorical variable with 2 levels.

Although, we get the same p value in Part a and Part b, but the interpretation of the equation has totally changed. In Part b, now x is a categorical variable and both these categories can be compared. But in Part a since there was no distinction in the independent variable , we couldn't have compared the categories.

Output 7: Q4 (b) R Code Output

```
> myData$X <- as.factor(myData$X)
> model_1 <- lm(Y ~ X, data = myData)
> summary(model_1)

Call:
lm(formula = Y ~ X, data = myData)

Residuals:
   Min      1Q  Median      3Q     Max
 -3.00   -1.75    0.00    1.00    4.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0000     0.8018   6.236 4.34e-05 ***
X1            3.0000     1.1339   2.646   0.0213 *

Residual standard error: 2.121 on 12 degrees of freedom
Multiple R-squared:  0.3684,    Adjusted R-squared:  0.3158
```

```
F-statistic:      7 on 1 and 12 DF,  p-value: 0.02135
```

Question 5

a)

H0 : beta(i)=0
H1: beta(i) $\neq 0$

y = 530.26 + 1.29(value) + 0.46(Doct) - 0.88(Nurse) + 2.14(VN)

Result : Since overall p < 0.05, we reject the null hypothesis that all beta(i) = 0.
Hence, there is a significant relationship between the variables. Although, DOCT and NURSE columns are not significant because their p value > 0.05. But since, VIF values are very high, it suggests that the coefficeints are poorly calculated and p value be not very accurate.

VIF values of all the independent variables are greater than 5 , hence very high multicollinearity exists in this model. This model is unstable. .

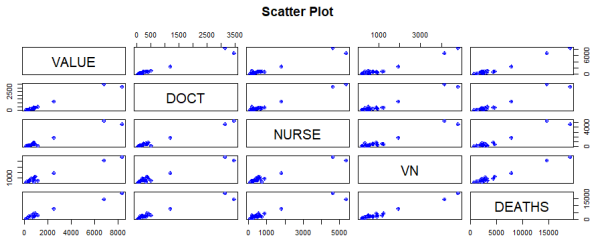

Figure 3: ScatterPlot

Output 8: Q5(a) R Code Output

```
> healthData <- read.table("Health.txt",header = TRUE)
> multiModel <- lm(DEATHS ~ VALUE + DOCT+NURSE+VN, data = healthData)
> summary(multiModel)

Call:
lm(formula = DEATHS ~ VALUE + DOCT + NURSE + VN, data = healthData)

Residuals:
     Min       1Q   Median       3Q      Max
-969.24  -338.57    48.59   269.80  1388.77

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 530.2651    263.5774    2.012   0.06139 .
VALUE         1.2958      0.5146    2.518   0.02282 *
DOCT          0.4639      1.7138    0.271   0.79010
NURSE        -0.8801      0.9071   -0.970   0.34636
VN            2.1422      0.6820    3.141   0.00631 **

Residual standard error: 651 on 16 degrees of freedom
Multiple R-squared:  0.9837,    Adjusted R-squared:  0.9797
F-statistic: 241.8 on 4 and 16 DF,  p-value: 4.365e-14

> x_cor = cor(healthData[,c(2,3,4,5)])
> x_cor = round(x_cor,2)
> x_cor
       VALUE DOCT NURSE   VN
VALUE   1.00 0.98  0.97 0.98
DOCT    0.98 1.00  0.99 0.97
NURSE   0.97 0.99  1.00 0.97
VN      0.98 0.97  0.97 1.00
> library(car)
> vif(multiModel)
    VALUE      DOCT     NURSE        VN
 56.71094 121.30731  77.55363  31.66861
```

b)
Yes, this model has a problem of multicollinearity. All the independent variables have high VIF values and highly correlated with each other.
Converting all the variables to Per Capita basis will solve the multicollinearity because it will scale down the variables which will lower the VIF values. Here, multicollinearity occurs because of high correlation between the variables.

c)
Although this model has low VIF values but it has a low R-squared value and p value $>$ 0.05. Hence, this model doesn't explain much of the variation and it is not significant.
In part-a, we had a high r-squared value and a p value $<$ 0.05. It was able to explain a lot of variation and also it was significant but with multicollinearity.

There are a lot of factors that needs to be taken into consideration before comparing models and only comparing r-square value is not enough.

Output 9: Q5(c) R Code Output

```
> perCapitaHealthData <- sqldf("select
cast (DEATHS as real)/POP
as DEATHS_PER_CAPITA,
```

```
cast(VALUE as real)/POP
as VALUE_PER_CAPITA,
cast(NURSE as real)/POP
as NURSE_PER_CAPITA,
cast(VN as real)/POP
as VN_PER_CAPITA,
cast(DOCT as real)/POP
as DOCT_PER_CAPITA from healthData")

> perCapitaMultiModel <- lm(DEATHS_PER_CAPITA ~ VALUE_PER_CAPITA
+ DOCT_PER_CAPITA
+ NURSE_PER_CAPITA
+ VN_PER_CAPITA, data = perCapitaHealthData)
> summary(perCapitaMultiModel)

Call:
lm(formula = DEATHS_PER_CAPITA ~
VALUE_PER_CAPITA + DOCT_PER_CAPITA +
    NURSE_PER_CAPITA + VN_PER_CAPITA,
    data = perCapitaHealthData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6276 -0.5940  0.2000  0.8019  2.7162

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.1162     2.5626   2.777   0.0135 *
VALUE_PER_CAPITA  -0.8424     1.5153  -0.556   0.5859
DOCT_PER_CAPITA    0.1742     2.9573   0.059   0.9538
NURSE_PER_CAPITA  -0.2844     1.0070  -0.282   0.7813
VN_PER_CAPITA      1.4995     0.6515   2.302   0.0351 *

Residual standard error: 1.771 on 16 degrees of freedom
Multiple R-squared:  0.3263,    Adjusted R-squared:  0.1578
F-statistic: 1.937 on 4 and 16 DF,  p-value: 0.1533

> x_cor_perC = cor(perCapitaHealthData[,c(2,3,4,5)])
> x_cor_perC = round(x_cor,2)
Error: object 'x_cor' not found
> x_cor_perC
                  VALUE_PER_CAPITA NURSE_PER_CAPITA VN_PER_CAPITA
```

```
VALUE_PER_CAPITA      1.0000000      0.4780082     -0.1492442
NURSE_PER_CAPITA      0.4780082      1.0000000      0.1904229
VN_PER_CAPITA        -0.1492442      0.1904229      1.0000000
DOCT_PER_CAPITA       0.8028668      0.5047947     -0.2134725
                    DOCT_PER_CAPITA
VALUE_PER_CAPITA         0.8028668
NURSE_PER_CAPITA         0.5047947
VN_PER_CAPITA           -0.2134725
DOCT_PER_CAPITA          1.0000000
>
> vif(perCapitaMultiModel)
VALUE_PER_CAPITA   DOCT_PER_CAPITA  NURSE_PER_CAPITA     VN_PER_CAPITA
        2.871505        3.164477        1.562935          1.197643
```