Assignment Cover Sheet

School of Computer, Data and Mathematical Sciences

WESTERN SYDNEY UNIVERSITY

Student Name	Grishma Mainali
Student Number	[redacted]
Unit Name and Number	MATH7061 The Nature of Data
Lecturer	[redacted]
Due Date	13th October 2023
Date Submitted	13th October 2023

DECLARATION

I hold a copy of this assignment that I can produce if the original is lost or damaged.

I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.

No part of this assignment/product has been written/produced for me by another person except where such collaboration has been authorized by the subject lecturer/tutor concerned.

I am aware that this work may be reproduced and submitted to plagiarism detection software programs to detect possible plagiarism (which may retain a copy on its database for future plagiarism checking).

I hereby certify that I have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

***Signature**:* 

Note: An examiner or lecturer/tutor has the right not to mark this assignment if the above declaration has not been signed)

Dataset

Loading the provided data set into R using 'read.csv()' function.

```
data = read.csv("F:/WSU/Year 1/Semester 1/Nature of
data/Assignment/Air_Quality.csv")
```

Analyzing the data set

```
head(data)
```

```
##   region_ID state AQI PM2.5 PM10 CO NO2 temperature humidity population
## 1   Region1   NSW  48     6   11  6  13          17       65      91160
## 2   Region2   QLD  17     8   22  9  12          27       89     105532
## 3   Region3    WA  28     2    6  4   6          14       66     106314
## 4   Region4   NSW  19    14   35 13  21          27       88     156018
## 5   Region5   VIC   7    19   29 15  28          25       94     169420
## 6   Region6   QLD  24     5   13  7   8          22       85      77657
##   industrial traffic proximity_nr vegetation distance_to_coast
## 1          2       2         high          9           coastal
## 2          3       4          low          7          moderate
## 3          2       1         high          7           coastal
## 4          4       5     moderate          4      near coastal
## 5          5       5     moderate          4      near coastal
## 6          2       2         high          9           coastal
```

```
dim(data)
```

```
## [1] 236  15
```

The provided data set has 15 variables with 236 observations. Variables "region_ID", "state", "proximity_nr", "distance_to_coast" stores values in character format and variables "AQI", "PM2.5", "PM10", "CO", "NO2", "temperature", "humidity", "population", "industrial", "traffic", "vegetation" stores values in integer format

```r
str(data)
```

```
## 'data.frame':    236 obs. of  15 variables:
##  $ region_ID        : chr  "Region1" "Region2" "Region3" "Region4" ...
##  $ state            : chr  "NSW" "QLD" "WA" "NSW" ...
##  $ AQI              : int  48 17 28 19 7 24 38 13 14 37 ...
##  $ PM2.5            : int  6 8 2 14 19 5 2 5 10 4 ...
##  $ PM10             : int  11 22 6 35 29 13 4 13 27 7 ...
##  $ CO               : int  6 9 4 13 15 7 5 7 11 5 ...
##  $ NO2              : int  13 12 6 21 28 8 10 10 13 11 ...
##  $ temperature      : int  17 27 14 27 25 22 11 20 29 11 ...
##  $ humidity         : int  65 89 66 88 94 85 76 75 91 77 ...
##  $ population       : int  91160 105532 106314 156018 169420 77657 82119
147601 116797 86233 ...
##  $ industrial       : int  2 3 2 4 5 2 2 3 4 2 ...
##  $ traffic          : int  2 4 1 5 5 2 2 3 5 2 ...
##  $ proximity_nr     : chr  "high" "low" "high" "moderate" ...
##  $ vegetation       : int  9 7 7 4 4 9 9 4 6 8 ...
##  $ distance_to_coast: chr  "coastal" "moderate" "coastal" "near coastal"
...
```

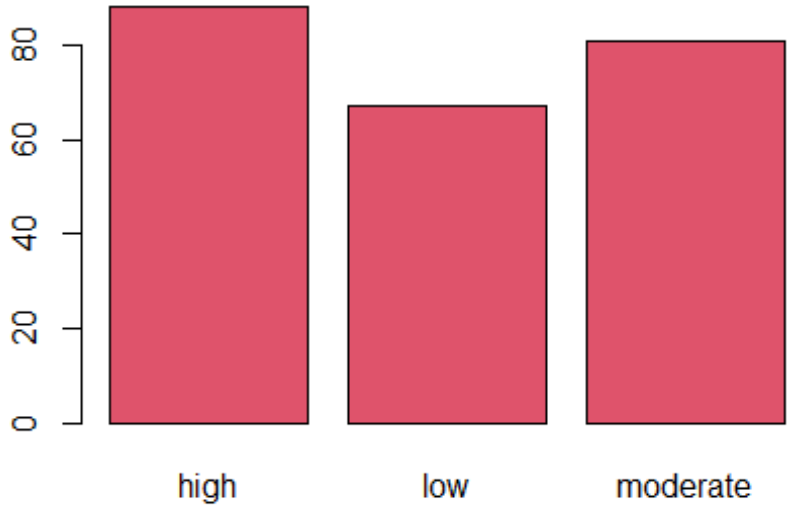
Question 1

Test whether there is a relationship between the state in which regions are located and their proximity to natural reserves in Australia.

To test this relationship, I have used chi squared test.

Firstly, the observed proximity is extracted from data set and stored in a table and a built in chi squared test is performed.

```
table1 = table(data$proximity_nr)
table1

##
##     high      low moderate
##       88       67       81

barplot(table1, col = 2)
```



Hypothesis for the test:

Null Hypothesis (Ho): There is a significant relationship between state and proximity to natural reserves

Alternative Hypothesis (Ha): There is no significant relationship between state and proximity to natural reserves

```
chi_test = chisq.test(table1, simulate.p.value = TRUE)
chi_test$expected

##     high      low moderate
## 78.66667 78.66667 78.66667
```

Expected counts are greater than 5, so we can use the p-value obtained from inbuilt chi squared test.

```
chi_test

##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  table1
## X-squared = 2.9068, df = NA, p-value = 0.2419

chi_test$p.value

## [1] 0.2418791
```

The obtained p-value is more than the chosen significance level (0.05). Hence, there is not enough evidence to reject the null hypothesis.

It concludes that there is evidence of a significant relationship between state and proximity to natural reserves.

Question 2

Test whether the mean Air Quality Index (AQI) for the regions in NSW are lower than the regions in VIC.

To test the difference in means for the two regions, t-test is performed.

Firstly, subsetting the data set to extract the data for regions NSW and VIC.

```
data_nsw = subset(data, state=="NSW")
data_vic = subset(data, state == "VIC")
```

Now, extracting only the AQI values of those two regions

```
nsw_aqi = data_nsw$AQI
vic_aqi = data_vic$AQI
```

Checking the variance and mean of the two regions.

```
sd(nsw_aqi)

## [1] 12.14409

sd(vic_aqi)

## [1] 9.515073
```

There is not significant difference in variance of AQI between the two regions.

```
mean(nsw_aqi)

## [1] 30.71642

mean(vic_aqi)

## [1] 27.4127
```

There is not significant difference in mean of AQI between the two regions

Checking the sample size to check if inbuilt function can be used for t-test instead of simulation.

```
length(nsw_aqi)

## [1] 67
```

```r
length(vic_aqi)
```

```
## [1] 63
```

Both region sizes are larger than 30, so the normality assumption is satisfied and inbuilt function can be used for t-test.

Hypothesis:

Null hypothesis (Ho): The mean AQI for regions in NSW is not lower than the mean AQI for regions in VIC.

Alternative hypothesis (Ha): The mean AQI for regions in NSW is lower than the mean AQI for regions in VIC.

Here in the t-test function, alternative = “less” is set because we want to test if the mean AQI in NSW is less than that in VIC, as asked by the question.

```r
t = t.test(nsw_aqi, vic_aqi, alternative = "less" )
t$p.value
```

```
## [1] 0.957123
```

The obtained p-value is greater than the chosen significance level (0.05). Hence, there is not enough evidence to reject the null hypothesis.

It concludes that there is not enough statistical evidence to conclude that the mean AQI for regions in NSW is lower than the mean AQI for regions in VIC.

Question 3

What is the 94% confidence interval for the difference in mean Temperature between inland regions and coastal regions in Queensland (QLD)?

Firstly, we subset the data to only obtain values for the state QLD

```
data_qld = subset(data, state=="QLD")
```

Then from the obtained data of QLD, we separate the data into inland and coastal regions and extract only the temperature values for that regions

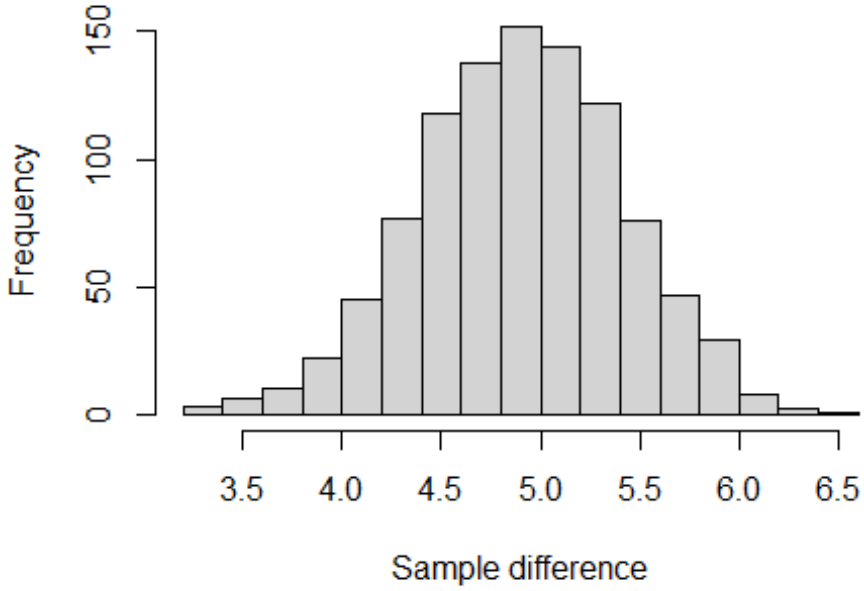
```
inland_qld = data_qld[data_qld$distance_to_coast=="inland", "temperature"]
coastal_qld = data_qld[data_qld$distance_to_coast=="coastal", "temperature"]
```

Checking the sample size to check if inbuilt function can be used for t-test instead of simulation.

```
length(inland_qld)

## [1] 14

length(coastal_qld)

## [1] 22
```

Both region sizes are smaller than 30 hence the inbuilt function cannot be used and simulation is required.

```
d = replicate(1000,{
in_sample = sample(inland_qld, replace=TRUE)
co_sample = sample(coastal_qld, replace=TRUE)
mean(in_sample) - mean(co_sample)
})

hist(d, xlab="Sample difference", main="", breaks=20)
```



We want a 94% confidence interval so we can use the d to estimate the points that have 3% below and 3% above using quantile

```
quantile(d, c(0.03, 0.97))
```

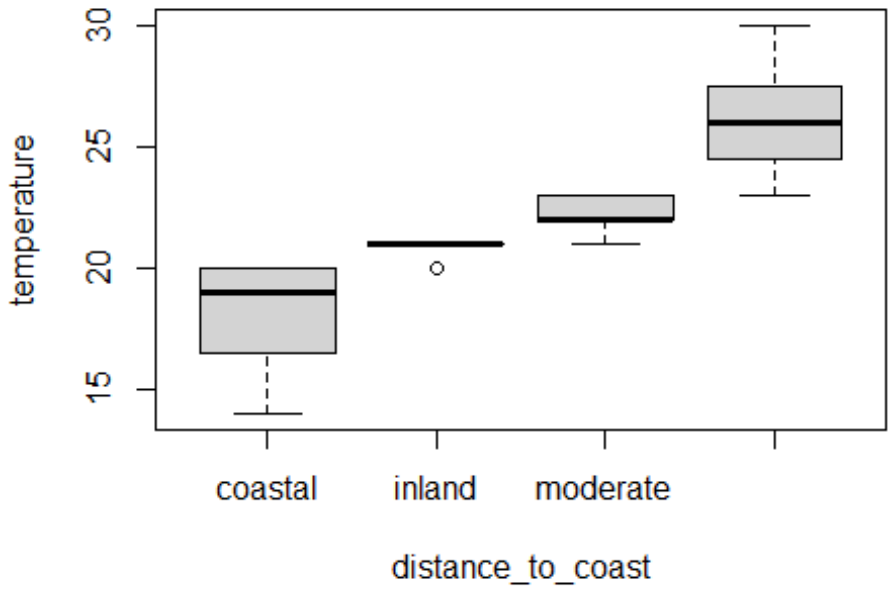
```
##       3%      97% 
## 3.909091 5.870130
```

From the obtained statistics, we are 94% confident that the difference in mean temperature between inland regions and coastal regions in Queensland are between 3.902403 and 5.915584

Question 4

Test whether the mean temperature varies among regions located at varying distances from the coast in NSW. If so, determine which categories have statistically different means.

```
boxplot(temperature~distance_to_coast, data = data_nsw)
```




```
ns = table(data_nsw$distance_to_coast)
ns

##
##      coastal       inland     moderate near coastal
##           20           13           10           24
```

The data size is small so we have to calculate using permutation

Firstly, testing equality of Means

```
m = oneway.test(temperature~distance_to_coast, data = data_nsw, var.equal =
TRUE)
F0 = m$statistic
F0

##        F
## 85.92763
```

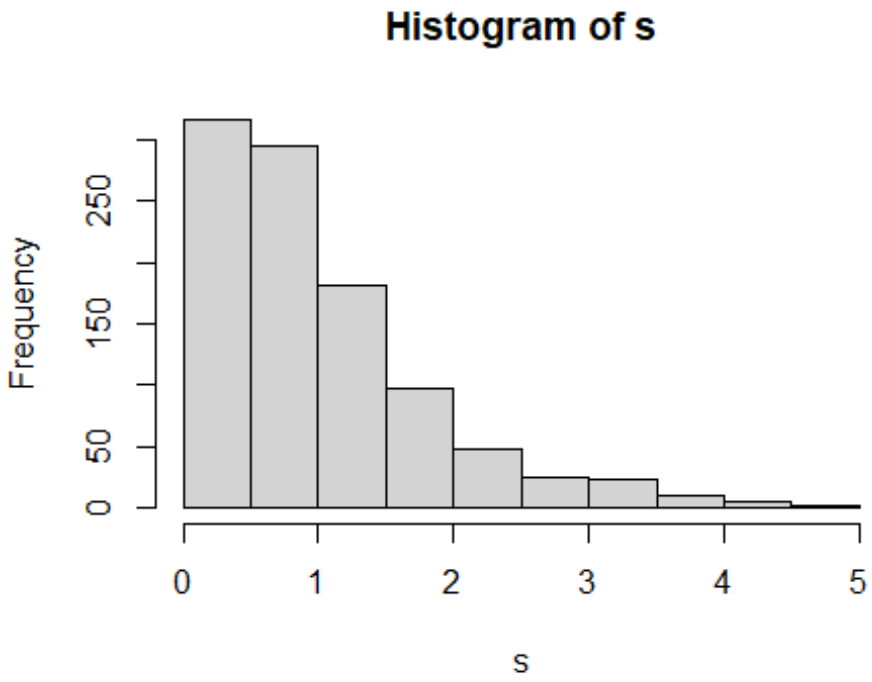
Hypothesis:

Null Hypothesis (Ho): There is no difference in mean temperature among the regions located at varying distances from the coast in NSW

Alternative Hypothesis (Ha): At least one region has a difference in mean temperature among the regions located at varying distances from the coast in NSW

Then we do simulation when the null hypothesis is true

```
s = replicate(1000,{
  sdistance = sample(data_nsw$distance_to_coast)
  test = oneway.test(temperature~sdistance, data = data_nsw, var.equal =
TRUE)
  test$statistic
})
hist(s)
```



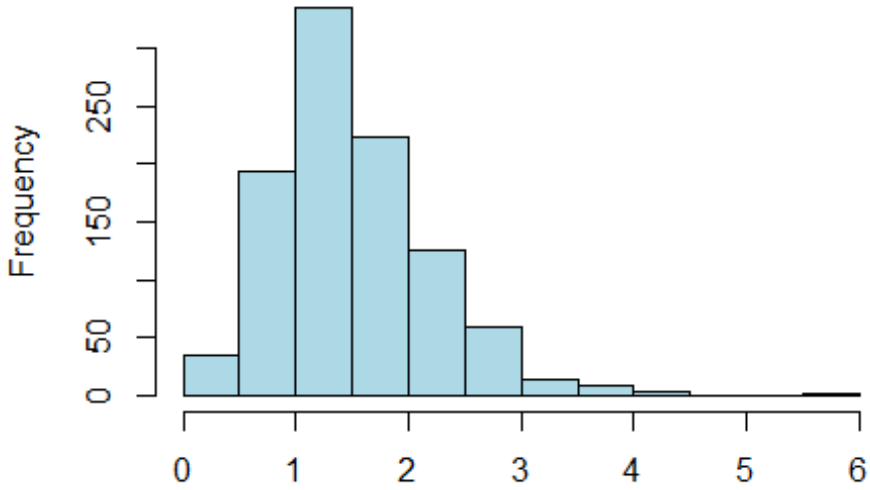

```
pval = mean(s>F0)
pval

## [1] 0
```

P-value is 0, so we can reject the null hypothesis. There is evidence of a difference and at least one region has a different means.

We detected a difference and now we need to find which one is different.

Computing t-statistics for simulated data

```
x = replicate(1000,{
distance.perm = sample(data_nsw$distance_to_coast) # shuffle the distances
fit0 = aov(temperature ~ distance.perm, data = data_nsw) # compute ANOVA to
obtain MSE
MSE = summary(fit0)[[1]][2,3] # Extract the MSE
means = aggregate(temperature ~ distance.perm, data = data_nsw, mean)[,2] #
compute means of categories
Ts = outer(means, means, "-")/sqrt(outer(1/ns,1/ns, "+")) # t-statistics
Ts = Ts/sqrt(MSE) # Scale by pooled standard deviation
max(abs(Ts)) # keep largest t statistic
})
hist(x, col = "lightblue", main = "", xlab = "")
```



Now, computing original t-statistics from all pairs

```
fit = aov(temperature ~ distance_to_coast, data = data_nsw)
MSE = summary(fit)[[1]][2,3] ## Extract the MSE
means = aggregate(temperature ~ distance_to_coast, data = data_nsw, mean)[,2]
Ts = outer(means, means, "-")/sqrt(outer(1/ns,1/ns, "+"))
Ts = Ts/sqrt(MSE)
Ts

##
##                  coastal    inland  moderate near coastal
##   coastal       0.000000 -4.456319 -6.256089   -15.732444
##   inland        4.456319  0.000000 -1.985993    -9.221555
##   moderate      6.256089  1.985993  0.000000    -6.217710
##   near coastal 15.732444  9.221555  6.217710     0.000000
```

Now, we compare the t-statistics of simualated and original data, and caluclate p-value.

```
p=matrix(rep(0,16),nrow=4)
rownames(p)=rownames(Ts)
colnames(p)=colnames(Ts)
for(i in 1:4){
  for(j in 1:4){
    p[i,j]=mean(x>Ts[i,j])
  }
}
p
```

```
##              coastal inland moderate near coastal
## coastal        1.000  1.000        1            1
## inland         0.001  1.000        1            1
## moderate       0.000  0.221        1            1
## near coastal   0.000  0.000        0            1
```

From obtained p-values we can conclude that:

There is a difference in mean temperature between inland and coastal as the obtained p-value (0) is less than 0.05

There is a difference in mean temperature between moderate and coastal as the obtained p-value (0) is less than 0.05

There is a difference in mean temperature between near-coastal and coastal as the obtained p-value (0) is less than 0.05

There is a difference in mean temperature between near-coastal and inland as the obtained p-value (0) is less than 0.05

There is a difference in mean temperature between near-coastal and moderate as the obtained p-value (0) is less than 0.05

Question 5

Test whether there is a linear relationship between the Population size and the Particulate Matter (PM2.5) levels in regions located in New South Wales (NSW) with Traffic Density ratings of 4 or 5. Provide the equation of the linear line.

Firstly, subset the data set to only extract the data for state NSW.

```
nsw = subset(data, state=="NSW")
```

Then, extracting the values of Population and PM2.5 of regions in NSW with traffic density ratings of 4 or 5.

```
nsw_pop = nsw[nsw$traffic=="4" | nsw$traffic=="5", "population"]

nsw_pm = nsw[nsw$traffic=="4" | nsw$traffic=="5", "PM2.5"]
```

If there is a linear relation between population size and PM2.5, the slope cannot be zero as non-zero slopes indicate a significant impact of the predictor variable from the response variable.

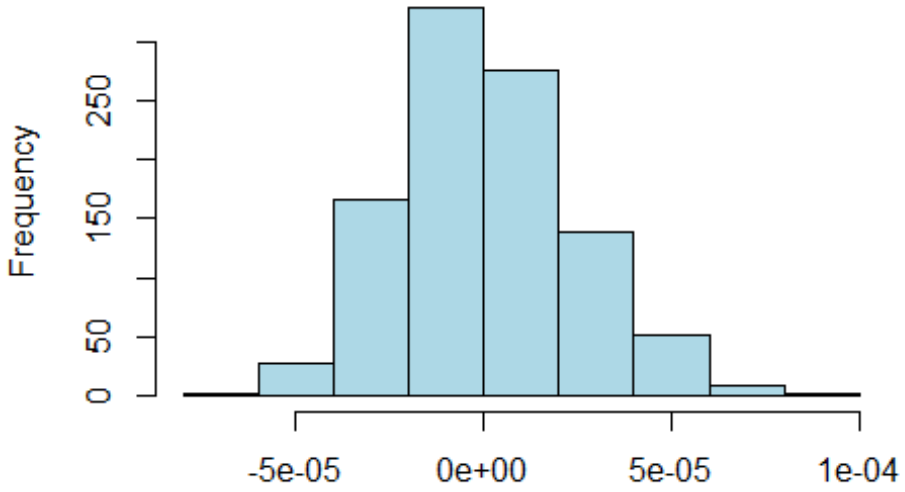
Hypothesis test (b = 0)

Null hypothesis (Ho): Slope(b) = 0

Alternative hypothesis (Ho): Slope(b) != 0

Creating a bootstrap distribution of b where the null hypothesis is true (b=0)

```
x= replicate(1000, {
pop.perm = sample(nsw_pop) # shuffle one variable to force population b = 0
fit = lm(nsw_pm~pop.perm) # fit the straight line model
coef(fit)[2] # return the fitted b
})
## examine the distribution of b, when the population b = 0
hist(x, col="lightblue", main="", xlab="")
```



Performing linear regression on original data with PM2.5 as the dependent variable and population as the independent variable.

```
lm1 = lm(nsw_pm~nsw_pop)
lm1

##
## Call:
## lm(formula = nsw_pm ~ nsw_pop)
##
## Coefficients:
## (Intercept)      nsw_pop
##  -5.8127093    0.0001265
```

Extracting the slope of original data

```
slope = coef(lm1)[2]
slope

##      nsw_pop
## 0.0001264759
```

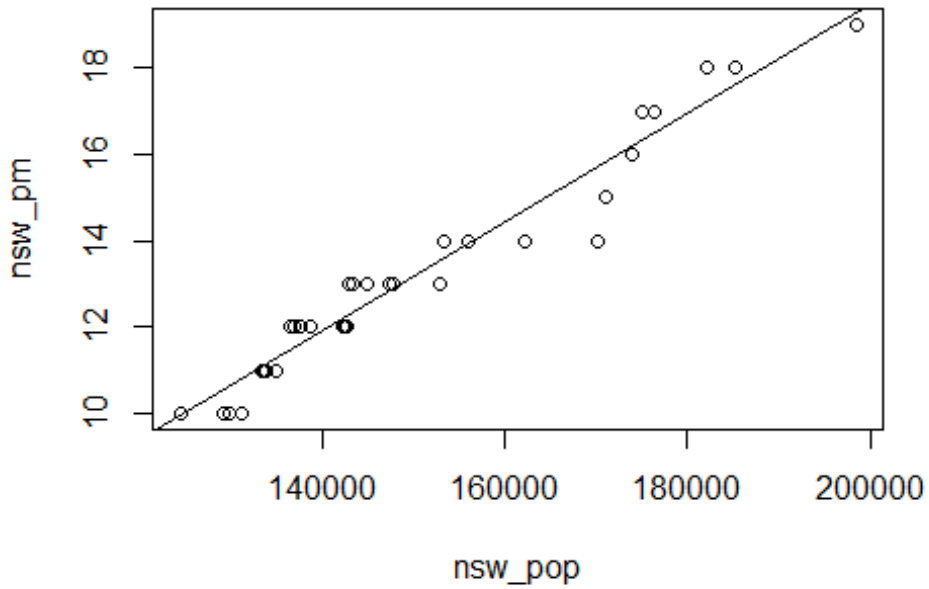
Computing the p-value

```
pValue = mean(x > slope) + mean(x < (-slope))
pValue
```

```
## [1] 0
```

The p value is small, so we reject H0, meaning that b != 0. So there is an association between the population size and PM2.5.

```
plot(nsw_pm~nsw_pop)
abline(lm1)
```




```
summary(lm1)
```

```
##
## Call:
## lm(formula = nsw_pm ~ nsw_pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71982 -0.25572  0.06456  0.46839  0.76260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.813e+00  8.100e-01  -7.176 4.56e-08 ***
```

```
## nsw_pop      1.265e-04  5.353e-06  23.629  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5721 on 31 degrees of freedom
## Multiple R-squared:  0.9474, Adjusted R-squared:  0.9457
## F-statistic: 558.3 on 1 and 31 DF,  p-value: < 2.2e-16
```

From the summary above, we can see that coefficient is significant as it has low p-value and the R-squared value is high indicating that a larger proportion of the variance in PM2.5 can be explained by Population.

Hence, we can conclude that there is a linear relationship between Population and PM2.5 levels in regions of NSW with Traffic Density ratings of 4 or 5.

```
coefficients(lm1)

##   (Intercept)       nsw_pop
## -5.8127093448  0.0001264759
```

The equation of the linear line is as follows:

nsw_pm = -5.8127093448 + nsw_pop * 0.0001264759

- where nsw_pm: Particulate Matter (PM2.5) levels in regions located in New South Wales (NSW) with Traffic Density ratings of 4 or 5.

nsw_pop : Population size in regions located in New South Wales (NSW) with Traffic Density ratings of 4 or 5.

Question 6

Investigate how good your model in Question 5.

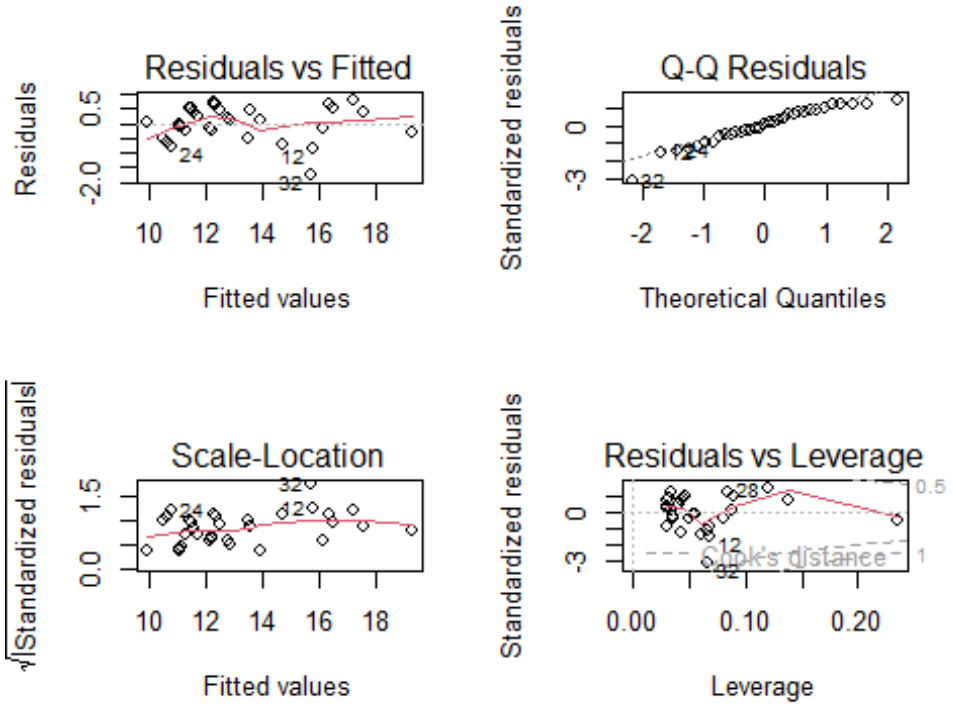
```
summary(lm1)
```

```
##
## Call:
## lm(formula = nsw_pm ~ nsw_pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71982 -0.25572  0.06456  0.46839  0.76260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.813e+00  8.100e-01  -7.176 4.56e-08 ***
## nsw_pop      1.265e-04  5.353e-06  23.629  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5721 on 31 degrees of freedom
## Multiple R-squared:  0.9474, Adjusted R-squared:  0.9457
## F-statistic: 558.3 on 1 and 31 DF,  p-value: < 2.2e-16
```

For the linear model developed, from summary we can see that the coefficient of the Population variables is significant which means that population does have an effect on PM2.5 levels.

Also, the R-squared value obtained is 0.94 which is quite high, indicating that a larger proportion of variance in PM2.5 levels is explained by population variable.

```
par(mfrow=c(2,2))
plot(lm1)
```

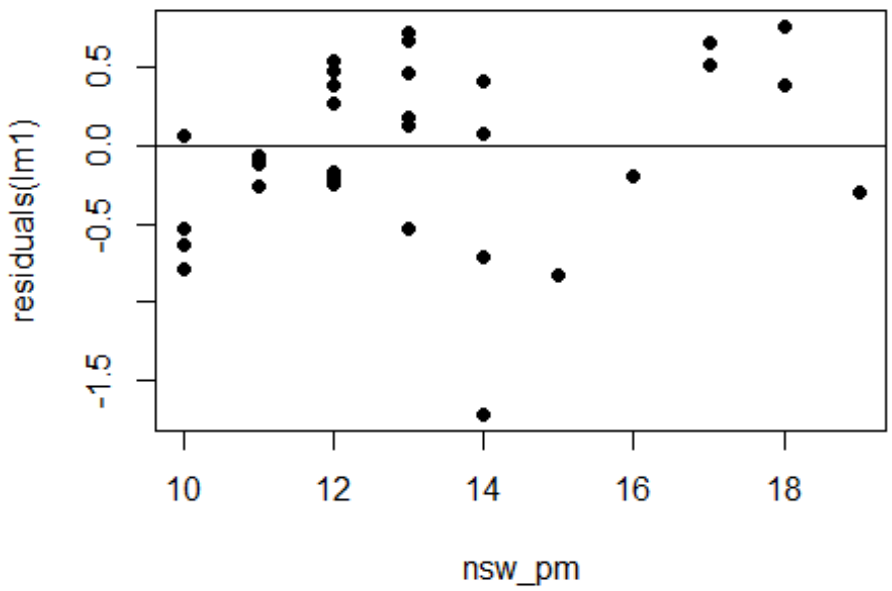


In the first box, the plots are scattered and do not follow a pattern, which validates linearity assumption

In second box, most of the values are aligned in straight line

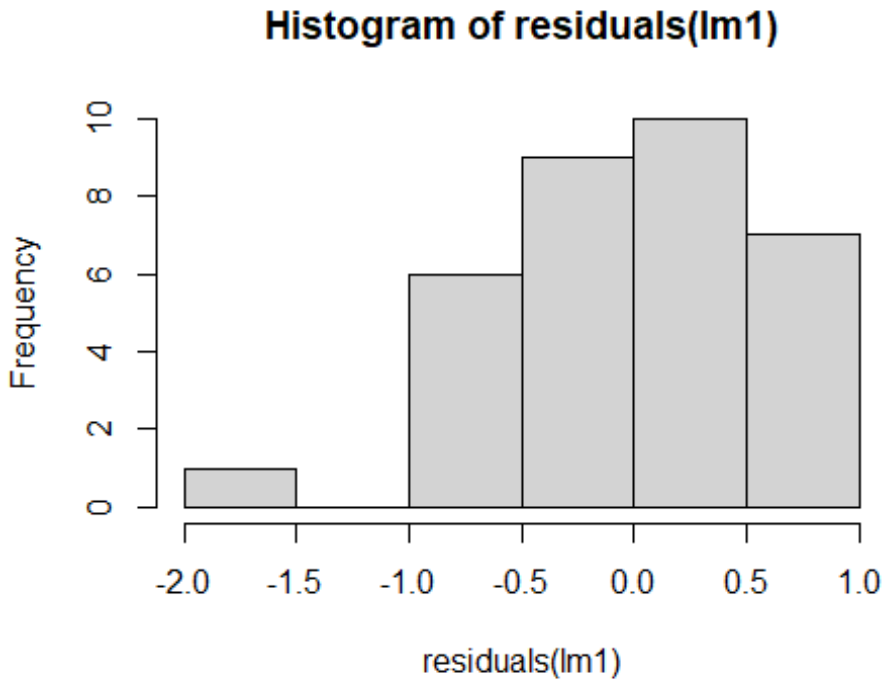
In third box, the variance is not consistant throughout, but its not that bad as well.

```
plot(residuals(lm1) ~ nsw_pm, pch = 16) # plot the residuals
abline(h = 0)
```



The residuals are scattered and do not seem to follow a pattern.

```
hist(residuals(lm1))
```



The distribution of residuals look normal.

Through all these information, we can conclude that the model looks like good as the variables in the model is significant and most of the variance of the data is captured by the model.

Question 7

Provide an interval estimate for Air Quality Index when Humidity is 84

Firstly, building a linear model between AQI and humidity variables.

```
lm2 = lm(AQI~humidity, data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = AQI ~ humidity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.617  -1.117   2.113   4.140  44.192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.23177    4.85590   23.94   <2e-16 ***
## humidity     -1.12171    0.06001  -18.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.969 on 234 degrees of freedom
## Multiple R-squared:  0.5989, Adjusted R-squared:  0.5971
## F-statistic: 349.3 on 1 and 234 DF,  p-value: < 2.2e-16
```

The built model has all significant values and it proves that humdity does have an effect on AQI.

Now, we make predictions from the built model by passing the value of humidity as “84”.

```
new_data <- data.frame(humidity = 84)

predict(lm2, newdata = new_data, interval="conf")
```

```
##        fit      lwr      upr
## 1 22.00852 20.90369 23.11336
```

From the obtained results, we can conclude that when the value of humidity is 84, AQI is 22.00852. Also, we are 95% confident that for that value of humidity AQI value ranges from 20.90369 to 23.11336.

Question 8

Critically discuss the sample used in this study.

Discuss its representativeness and the quality of data collection. Are there any potential biases or limitations associated with the sample that might impact the validity of the results?

```
dim(data)
```

```
## [1] 236  15
```

The data set includes information on air quality measurements, weather conditions, and demographic factors for 236 regions. The data set contains observations on the following 15 variables of randomly selected 236 regions across Australia. Data is randomly generated for the assignment.

```
str(data)
```

```
## 'data.frame':    236 obs. of  15 variables:
##  $ region_ID        : chr  "Region1" "Region2" "Region3" "Region4" ...
##  $ state            : chr  "NSW" "QLD" "WA" "NSW" ...
##  $ AQI              : int  48 17 28 19 7 24 38 13 14 37 ...
##  $ PM2.5            : int  6 8 2 14 19 5 2 5 10 4 ...
##  $ PM10             : int  11 22 6 35 29 13 4 13 27 7 ...
##  $ CO               : int  6 9 4 13 15 7 5 7 11 5 ...
##  $ NO2              : int  13 12 6 21 28 8 10 10 13 11 ...
##  $ temperature      : int  17 27 14 27 25 22 11 20 29 11 ...
##  $ humidity         : int  65 89 66 88 94 85 76 75 91 77 ...
##  $ population       : int  91160 105532 106314 156018 169420 77657 82119
147601 116797 86233 ...
##  $ industrial       : int  2 3 2 4 5 2 2 3 4 2 ...
##  $ traffic          : int  2 4 1 5 5 2 2 3 5 2 ...
##  $ proximity_nr     : chr  "high" "low" "high" "moderate" ...
##  $ vegetation       : int  9 7 7 4 4 9 9 4 6 8 ...
##  $ distance_to_coast: chr  "coastal" "moderate" "coastal" "near coastal"
...
```

The data set has 236 observations and 15 variables. All the variables are stored in proper suitable datatypes and there are no missing values in the data set which is a good thing for analysis.

For some questions that I solved, such as question 2 (related to t-test), the sample size was greater than 30, which meant that the data was equally covered for all the regions in NSW.

However, for question 4 (related to anova test), the sample size was smaller than 30, which meant that the data was not equally covered for all regions located at varying distances from the coast in NSW.

This indicates that the data might not be covering all the aspects of variations in variables within the data set.