Question 1

a.)

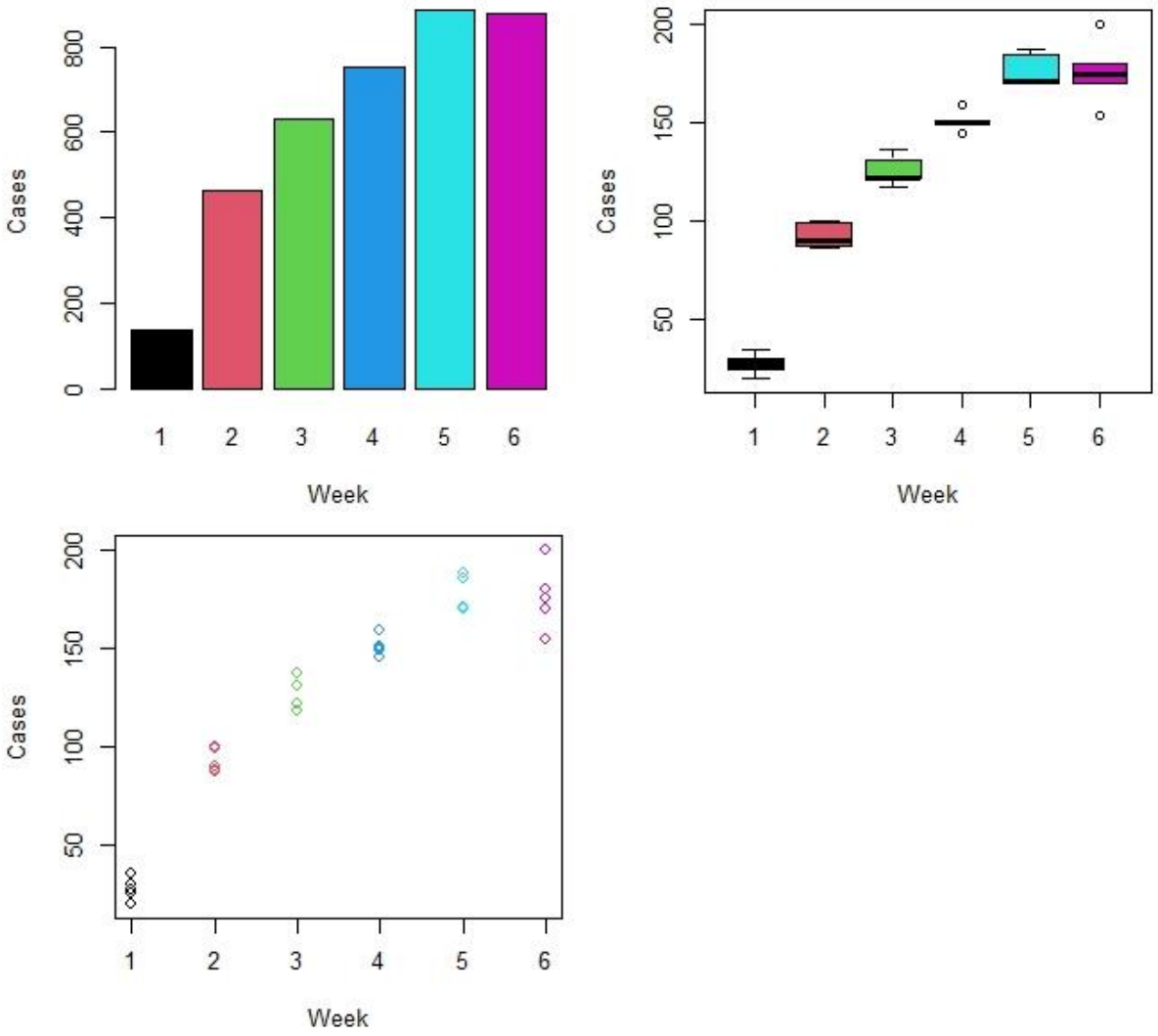

Figure 1: Various plot between week and cases

The bar plot shows the representation of the above table and it can be seen that week 5 had the most cases recorded in the 5 cities combined and week 1 had the least number of cases recorded. The boxplot shows the statistical estimates of the data and it can be seen that some cities had some outlying cases different from the mean for week 4 and 6. And lastly, the scatter plot shows that the number of cases increases as the weeks increases which is a positive relation. Also, it can be noticed that a linear model could be a possible fit for this data, but would not fit the data properly therefore, a curvilinear model might be best suitable for this data.

b.)

Using the first Order

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.053	8.018	2.875	0.00763 **
week	29.109	2.059	14.139	2.83e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.26 on 28 degrees of freedom

Multiple R-squared: 0.8771, Adjusted R-squared: 0.8728

F-statistic: 199.9 on 1 and 28 DF, p-value: 2.835e-14

Output 1: Regression Summary of 1st order polynomial

Analysis of Variance Table

Response: cases

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Week	1	74140	74140	199.91	2.835e-14 ***
Residuals	28	10384	371		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Output 2: ANOVA Summary of 1st order polynomial

The first order shows that the predictor is still significant with R-squared and adjusted R-squared values of 0.8771 and 0.8728 respectively, and the ANOVA output also indicated that the model is significant.

Using second order

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-37.0800	7.9042	-4.691	6.98e-05 ***
week	74.2086	5.1711	14.351	3.74e-14 ***
I(week^2)	-6.4429	0.7232	-8.909	1.59e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.88 on 27 degrees of freedom

Multiple R-squared: 0.9688, Adjusted R-squared: 0.9665

F-statistic: 419.4 on 2 and 27 DF, p-value: < 2.2e-16

Output 3: Regression Summary of 2nd order polynomial

Analysis of Variance Table

Response: cases

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
week	1	74140	74140	759.475	< 2.2e-16 ***
I(week^2)	1	7749	7749	79.376	1.592e-09 ***
Residuals	27	2636	98		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Output 4: ANOVA Summary of 2nd order polynomial

The second order polynomial shows that the addition of the higher power predictor improves the model and increases the R-squared and adjusted R-squared value to 0.9688 and 9.9665. Also, the ANOVA table indicates that the overall model is still significant with the addition of the new term.

Using third order

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-49.8667	15.9759	-3.121	0.00437 **
week	90.2934	18.2031	4.960	3.73e-05 ***
I(week^2)	-11.7706	5.8249	-2.021	0.05372
I(week^3)	0.5074	0.5504	0.922	0.36509

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.908 on 26 degrees of freedom

Multiple R-squared: 0.9698, Adjusted R-squared: 0.9663

F-statistic: 278.3 on 3 and 26 DF, p-value: < 2.2e-16

Output 5: Regression Summary of 3rd order polynomial

Analysis of Variance Table

Response: cases

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
week	1	74140	74140	755.2489	< 2.2e-16 ***
I(week^2)	1	7749	7749	78.9340	2.344e-09 ***
I(week^3)	1	83	83	0.8498	0.3651
Residuals	26	2552	98		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Output 6: ANOVA Summary of 3rd order polynomial

It can be observed from the ANOVA output table of the addition of a third order polynomial will make the term insignificant to the model with a p-value of 0.3651. The summary value also indicated that the third order term makes is insignificant as a predictor and it also reduces the adjusted R-square from 0.9665 to 0.9663.

Therefore, the best order of polynomial to create a suitable model is the second order polynomial as it has all predictors significant with the highest adjusted R-squared value

c.)

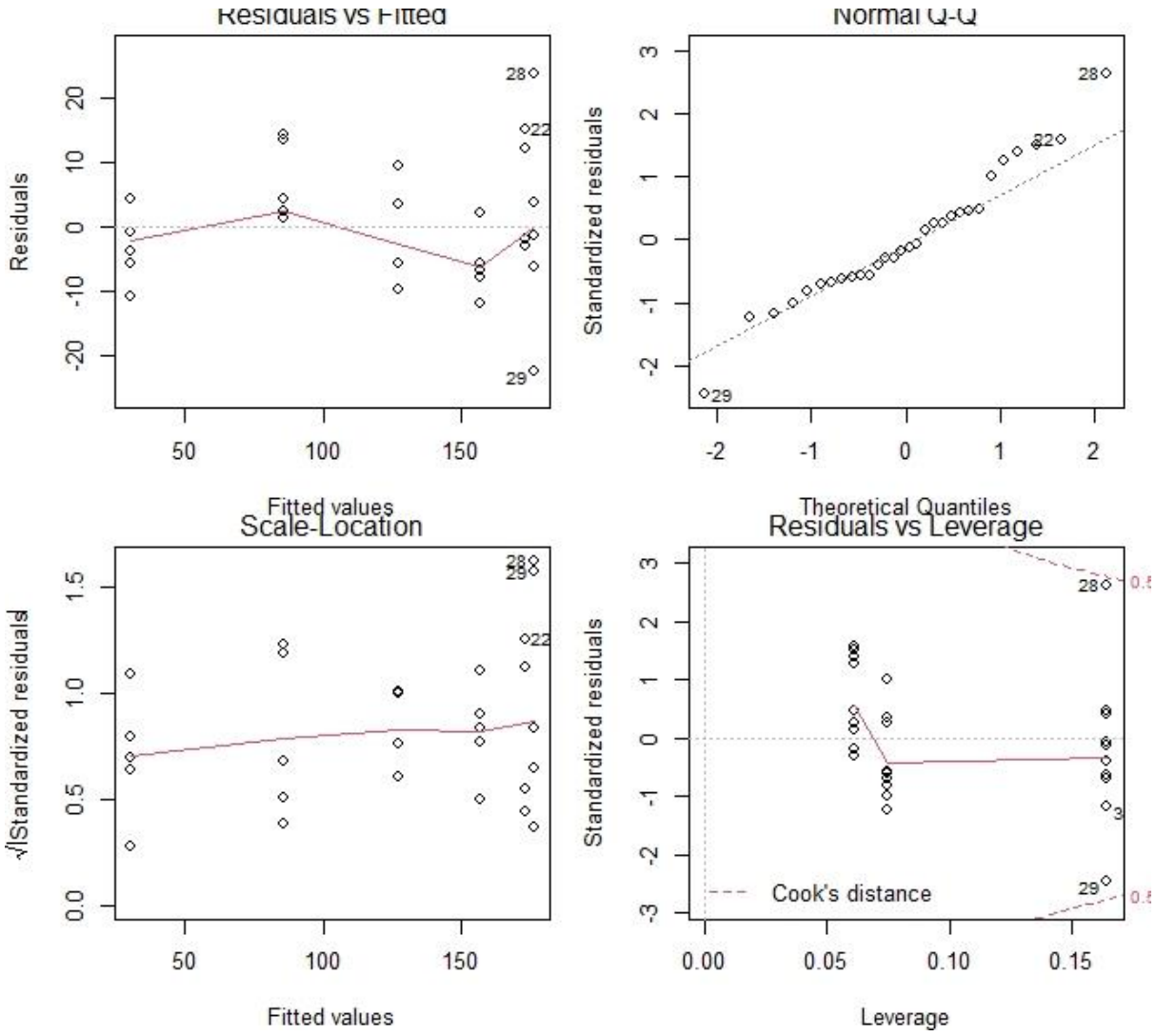

Figure 2: Assumption Plots

- The model assumptions for the second order polynomial suggests that the residuals are fairly randomly distributed along the line which indicates that the model provides a decent fit to the data. This can be observed in the Residuals vs Fitted plot.
- The Normal Q-Q plot provides a fairly safe assumption that the data is normally distributed as majority of the observed data points are in close proximity to the line.
- The scale location plot satisfies our assumption for homoscedasticity as the red line is roughly horizontal and the spread of residuals around the line is randomly scattered showing patterned distribution. Therefore, there is a roughly equal variability with the fitted values.
- Residuals vs Leverage plot does not show any high influence points but there are some points that should be further investigated as they are potentially influential (28 and 29) and also, these points are potential outliers as well.

d.)

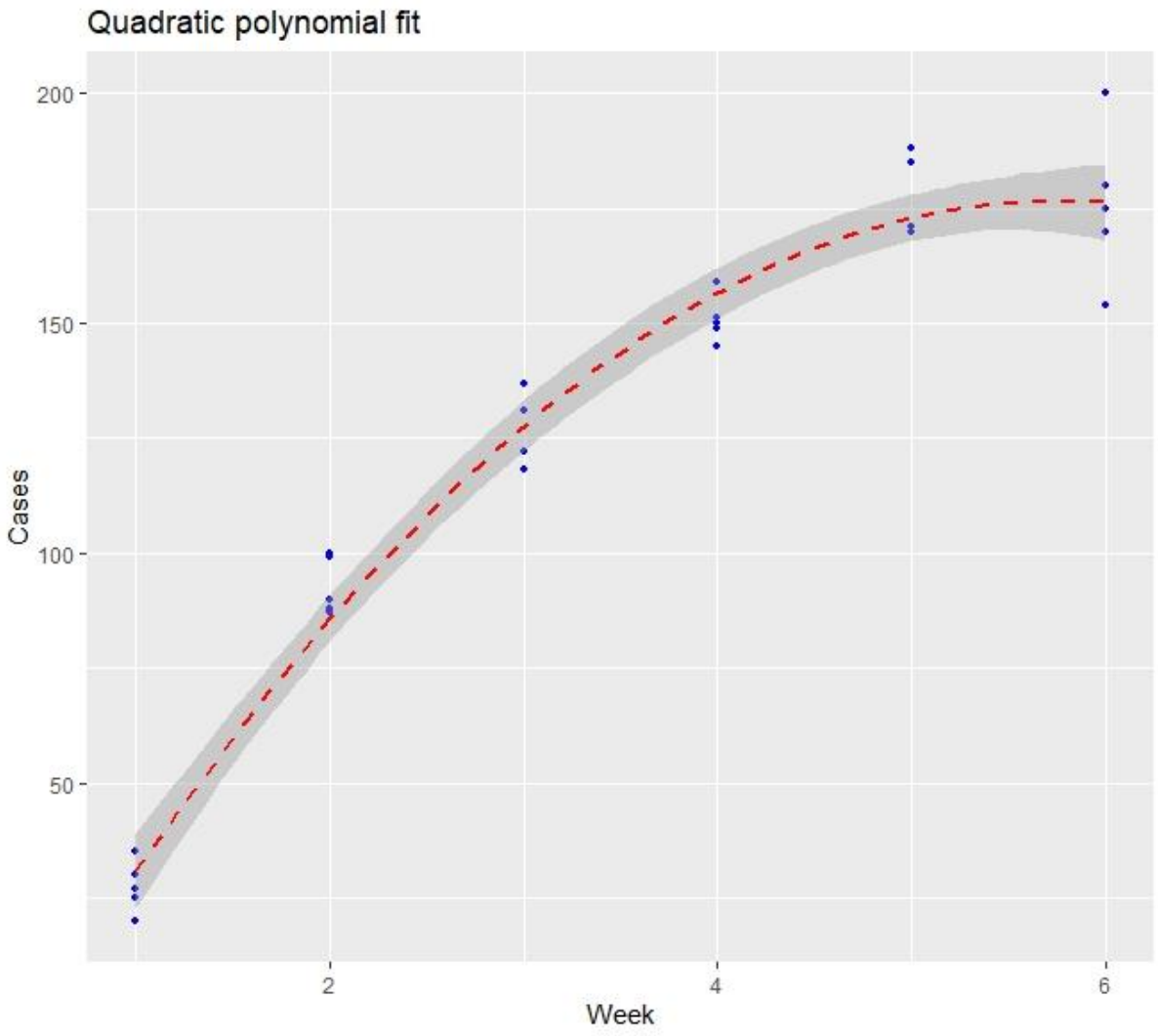

Figure 3: Quadratic Model for data

It can be seen that after plotting the scattered plot to the data, the second order polynomial is fit nicely with the data, and the 95% confidence band contains most of the data points. Therefore, the regression model can be considered appropriate for the data.

Question 2

a.)

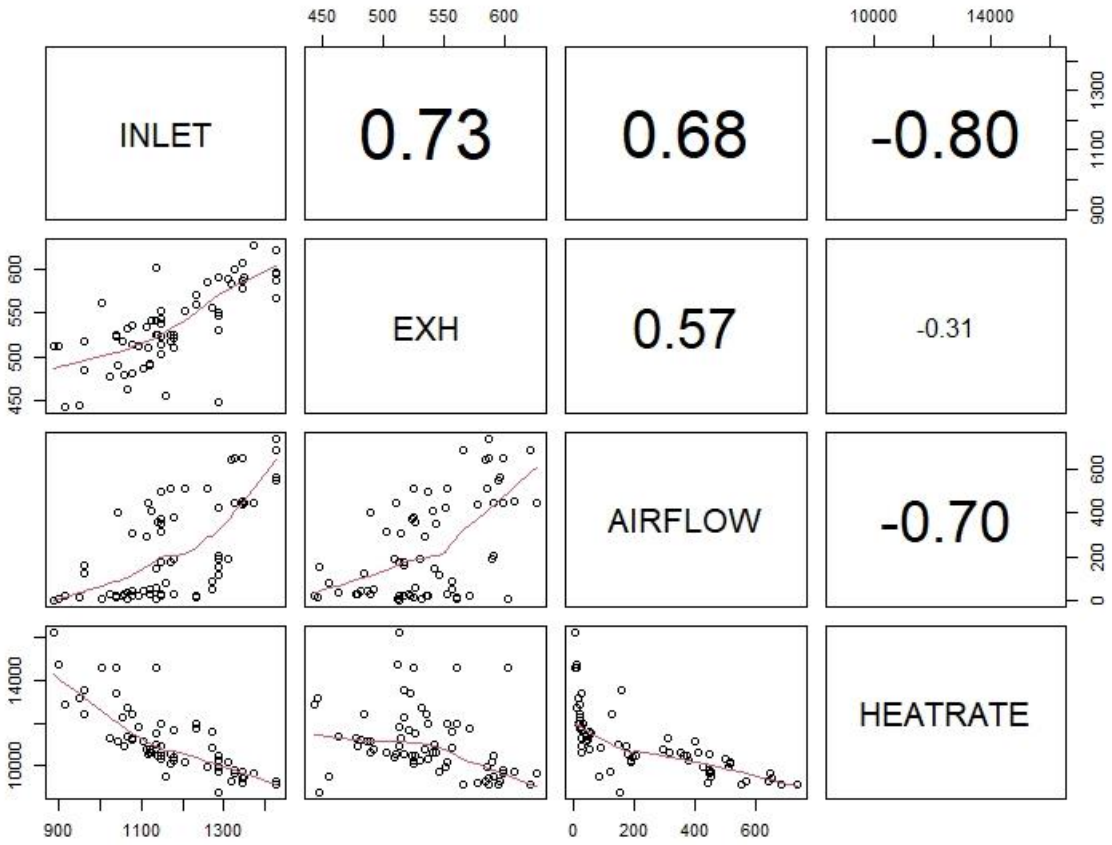

Figure 4: Pairs plot showing correlation between numeric variables

It will be observed from the plot above that the HEATRATE of the gas turbines has a strong negative correlation with INLET and AIRFLOW with a correlation value of -0.8 and -0.7 respectively. However, it has a weak negative correlation with EXH of -0.31.

Also, it can be seen that INLET has a high positive correlation with both EXH and AIRFLOW with values 0.73 and 0.68 respectively. There is a moderate positive correlation between EXH and AIRFLOW of value 0.57.

As we are trying to determine the best combination of measurements to analyse the gas turbine measurement, a regression model would be the preferred choice for the model; the suggested regression model deduced from the pairs plot is the linear regression model as this helps predict the value of the response variable (HEATRATE) based on the values of the other predictors combine The suggested model would be the linear regression model as this helps explain the weight each predictor has on the response variable therefore explaining the high correlation.

b.)

	INLET	**EXH**	**AIRFLOW**	**HEATRATE**
INLET	1.000	0.728	0.681	-0.801
EXH	0.728	1.000	0.567	-0.314
AIRFLOW	0.681	0.567	1.000	-0.703
HEATRATE	-0.801	-0.314	-0.703	1.000

Output 7: Correlation Output between INLET, EXH, AIRFLOW, AND HEATRATE

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13170.6662	1049.6177	12.548	< 2e-16 ***
INLET	-11.6268	0.8755	-13.280	< 2e-16 ***
EXH	22.7362	2.4228	9.384	1.41e-13 ***
AIRFLOW	-2.6555	0.4413	-6.017	9.92e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 589.4 on 63 degrees of freedom

Multiple R-squared: 0.8697, Adjusted R-squared: 0.8634

F-statistic: 140.1 on 3 and 63 DF, p-value: < 2.2e-16

Output 8: Regression Output of the Main effects Model

Analysis of Variance Table

Response: HEATRATE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
INLET	1	107601712	107601712	309.747	< 2.2e-16 ***
EXH	1	25831694	25831694	74.360	2.917e-12 ***
AIRFLOW	1	12578470	12578470	36.209	9.916e-08 ***
Residuals	63	21885332	347386		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Output 9: ANOVA output of Main effects Model

INLET	EXH	AIRFLOW
2.751	2.173	1.902

Output 10: VIF output of predictors

- The correlation shows that there is a strong positive correlation between INLET and both (EXH and AIRFLOW) with values (0.728, 0.681) respectively and there is a moderate correlation between EXH and AIRFLOW. From this output, it COULD be deduced that multicollinearity exists between the predictors of turbine's HEATRATE.
- Comparing the correlation output to the regression output, it can be seen that the response variable (HEATRATE) has a negative correlation with all the predictors on the correlation output, but on the regression output, it is notices that the relationship between HEATRATE and EXH is positive which is contrary to their correlation. This output also indicates the presence of multicollinearity.
- The regression output and the ANOVA output indicate that all predictors are significant to the model and the overall model is significant as indicated by the p-value of the F-statistics. With no contradictions between the F and the T statistics, this indicator does not show that the model has multicollinearity.
- The above table is the variance inflation factor (VIF) output of the model; this does not detect any sign of multicollinearity as none of the values are above 10.

Therefore, after considering all indicators of multicollinearity, I found out that out of the 4 indicators, only two ("VIF output" and "F-test and t-test comparison) do not provide indications that multicollinearity exists in the model.

c.)

Start: AIC=989.19

HEATRATE ~ 1

	Df	Sum of Sq	RSS	AIC
+ INLET	1	107601712	60295496	922.58
+ AIRFLOW	1	83000387	84896822	945.50
+ EXH	1	16586588	151310620	984.22
<none>			167897208	989.19

Step: AIC=922.58

HEATRATE ~ INLET

	Df	Sum of Sq	RSS	AIC
+ EXH	1	25831694	34463802	887.10
+ AIRFLOW	1	7818472	52477024	915.27
<none>			60295496	922.58

Step: AIC=887.1

HEATRATE ~ INLET + EXH

	Df	Sum of Sq	RSS	AIC
+ AIRFLOW	1	12578470	21885332	858.67
<none>			34463802	887.10
+ INLET: EXH	1	47855	34415948	889.01

Step: AIC=858.67

HEATRATE ~ INLET + EXH + AIRFLOW

	Df	Sum of Sq	RSS	AIC
+ INLET: AIRFLOW	1	6056999	15828333	838.97
<none>			21885332	858.67
+ INLET: EXH	1	522081	21363251	859.06
+ EXH: AIRFLOW	1	448267	21437065	859.29

Step: AIC=838.97

HEATRATE ~ INLET + EXH + AIRFLOW + INLET: AIRFLOW

	Df	Sum of Sq	RSS	AIC
+ EXH: AIRFLOW	1	4806495	11021838	816.72
+ INLET: EXH	1	1213545	14614788	835.62
<none>			15828333	838.97

Step: AIC=816.72

HEATRATE ~ INLET + EXH + AIRFLOW + INLET: AIRFLOW + EXH: AIRFLOW

	Df	Sum of Sq	RSS	AIC
<none>			11021838	816.72
+ INLET: EXH	1	20129	11001709	818.59

Output 11: Forward Stepwise output including interaction

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.394e+04	1.044e+03	13.353	< 2e-16 ***
INLET	-1.514e+01	7.775e-01	-19.470	< 2e-16 ***
EXH	2.884e+01	2.304e+00	12.519	< 2e-16 ***
AIRFLOW	-6.895e-01	3.628e+00	-0.190	0.85
INLET: AIRFLOW	2.277e-02	2.999e-03	7.592	2.22e-10 ***
EXH: AIRFLOW	-5.430e-02	1.053e-02	-5.158	2.87e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 425.1 on 61 degrees of freedom

Multiple R-squared: 0.9344, Adjusted R-squared: 0.929

F-statistic: 173.6 on 5 and 61 DF, p-value: < 2.2e-16

Output 12: Summary output of Stepwise Regression

Analysis of Variance Table

Response: HEATRATE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
INLET	1	107601712	107601712	595.518	< 2.2e-16 ***
EXH	1	25831694	25831694	142.965	< 2.2e-16 ***
AIRFLOW	1	12578470	12578470	69.615	1.130e-11 ***
INLET: AIRFLOW	1	6056999	6056999	33.522	2.628e-07 ***
EXH: AIRFLOW	1	4806495	4806495	26.601	2.869e-06 ***
Residuals	61	11021838	180686		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Output 13: ANOVA output of Stepwise Regression

The Stepwise regression output above uses the forward direction to find the predictors that are significant to the model in predicting the HEATRATE provided all possible interaction terms are included. The best model of AIC value of 816.72 is achieved when the INLET and EXH interaction is not considered as a predictor.

The summary output of the stepwise regression model shows that the new model explains 92.9% of the variability of the performance. Also, the p-value of the F-statistic shows that the overall model is significant (p-value < 2.2e-16)

From the ANOVA table above, it will be seen that the addition of the interaction term and AIRFLOW predictor variables are significant to the model, likewise the overall model is significant as all the predictors are significant.

d.)

From the initial regression model with no interaction, it can be seen that the model explains about 86.34% of the response term variability and all the main effect terms are significant to the model. Also, the correlation check indicated the presence of multicollinearity as indicated by the change of expected relationship between EXH and HEATRATE, and the output of the correlation showing high relationship among the predictor variables. An opposite result was found from the VIF output and comparison between F and T test output. Since correlation does not imply causation, the effect of a predictor variable over another is not necessarily caused by their correlation. Since the main effect regression summary shows that all predictor terms are significant in the regression, therefore, it can be said that the presence of high correlation in the data among the predictors have no effect in predicting the response variable.

The result of the forward stepwise model when considering the interaction terms between the predictors provide a better output as it explains a much larger variability of the response variable of 92.9%. This model gives the lowest AIC value of 816.72 is chosen as the preferred model. The least significant interaction term is the interaction between EXH and AIRFLOW. It can be concluded that for the data collected, a better model can be obtained when the interaction terms are considered despite the presence of high correlation amongst the predictor variables. The final model equation for the analysis is:

E(HEATRATE) = 13940 -15.14(INLET) + 28.84(EXH) – 0.6895(AIRFLOW) + 0.02277(INLET*AIRFLOW) - 0.0543(EXH*AIRFLOW)

Appendix: R-Code

Question 1

```
# Import Data
data_virus <- read.table(file.choose(),header = TRUE)
head(data_virus,5)
# Explore Data
str(data_virus) # get structure of the dataset
summary(data_virus)
# Find Missing Values
sum(is.na(data_virus))
# Plot cases against week
total_cases <- data.frame(tapply(data_virus$cases,data_virus$week,sum)) # get the total cases each week
colnames(total_cases)<-"Total"
total_cases
par(mfrow = c(2,2), mar = c(4,4,1,1))
# Barplot
barplot(tapply(data_virus$cases,data_virus$week,sum),col=c(1,2,3,4,5,6),
      xlab = "Week",
      ylab = "Cases")
# Boxplot
boxplot(data_virus$cases~data_virus$week,col=c(1,2,3,4,5,6),
   xlab = "Week",
   ylab = "Cases")
plot(cases~week,data=data_virus,col=week,
   xlab = "Week",
   ylab = "Cases")
par(mfrow = c(1,1), mar = c(4,4,1,1))
# Polynomial Regression
# Order 1
mod <- lm(cases~week,data_virus)
```

```
summary(mod)
anova(mod)
# Order 2
mod2 <- lm(cases~week+I(week^2),data_virus)
summary(mod2)
anova(mod2)
# Order 3
mod3 = lm(cases ~week+I(week^2)+I(week^3),data_virus)
summary(mod3)
anova(mod3)
par(mfrow = c(2,2), mar = c(4,4,1,1))
plot(mod2)
par(mfrow = c(1,1), mar = c(4,4,1,1))
library(ggplot2)
ggplot(data = data_virus, aes(x=week,y=cases))+
      geom_point(pch=20,color = "blue",size = 2)+
      geom_smooth(method = "lm", formula = y~poly(x,2), color="red",linetype= 2, se = TRUE)+
      labs(title = "Quadratic polynomial fit",x="Week",y="Cases")
```

Question 2

```
# Import Data
data_gas <- read.table(file.choose(),header = TRUE)
head(data_gas,5)
# Explore Data
str(data_gas) # get structure of the dataset
summary(data_gas)
# Find Missing Values
sum(is.na(data_gas))
# pairs plot to visualize correlation of the numeric values
pairs(data_gas[2:ncol(data_gas)],
    lower.panel=panel.smooth,
```

```
    upper.panel = panel.cor) # visual plot to show bivariate relationship between all variable pairs (visualizes correlation)

set.seed(600)

# fit model using linear regression

mod <- lm(HEATRATE~INLET+EXH+AIRFLOW,data_gas)

summary(mod) # model summary

round(cor(data_gas[2:(ncol(data_gas)-1)]),3) #correlation between numeric predictors

anova(mod)

# Get VIF of model predictors

# install.packages("car")

library(car)

round(vif(mod),3)

# Using stepwise Regression

# minimal model (lower model)

formL <- formula(~ 1)

# maximum model (upper model)

formU <- formula(~INLET*EXH*AIRFLOW)


# forward selection

start.model <- lm(HEATRATE ~ 1, data=data_gas)

fstep.model <- step(start.model,

                    direction="forward",

                    scope=list(lower= formL,upper=formU))

summary(fstep.model)

anova(fstep.model)
```