WESTERN SYDNEY
UNIVERSITY

School of Computing, Engineering and Mathematics

Complete your details in this section when instructed by the Exam Supervisor at the start of the exam. You should also complete your details on any extra answer papers provided.	
STUDENT ID:	MARKING SCHEME
STUDENT FIRST NAME:	
STUDENT SURNAME:	

UNIT NUMBER:	301044		
UNIT NAME:	Data Science		
QUESTIONS FORMAT:	Word processed document in PDF format; logically presenting answers to each question incorporating R outputs including graphs and charts.		
WEIGHT:	Total exam marks: **60** - **50%** of total assessment.		
UNIT CO-ORDINATOR:	Dr. Liwan Liyanage		
LECTURER:			
TIME ALLOWED:	2 Hours	TOTAL PAGES:	

Final Exam INSTRUCTIONS

Please note that you are expected to answer the questions clearly. Give the R commands, analysis, comments and discussion clearly and logically. Once completed submit the answer scripts via TurnItin link within vUWS site. You also need to show the file you uploaded before leaving the examination room.

Resources:

Open book. Students are allowed to use any material related to the subject Lecture notes and practical notes available on vUWS. Summaries and handy hints given in class or done by yourself; useful links and readings listed and uploaded in vUWS.

Question 1 (3 + 2 + 3 + 2 + 2 + 3 = 15)

This Question uses the data set “Envdata”. The data represents the pollution conditions and maximum wind speed together with the prevalence of Asthma (present or absent of Asthma) associate with several patients in Victoria State.

The Envdata are given for each patient under investigation as follows:

X1=co
X2=so2
X3=o3
X4=ppm10
X5=no2
X6= maxwindspeed
Y = asthma

a. Use K Means Clustering method and identify two clusters with K=2. [MARK = 3]

R code:

```
library(readxl)
env=read.csv("RELEVANT PATH")
head(env)
Y = env_PCA[ , c(-1,-8)]
# Check whether the Response variable is removed before clustering MARK (1)

km = kmeans(Y, 2, nstart = 20) #Writing the correct coding with k value MARK (1)
km
```

R output: ***MARK (1)***

```
K-means clustering with 2 clusters of sizes 22, 86

Cluster means:
         co       o3      no2       so2    ppm10 maxwindspeed
1 0.2090909 31.45455 8.136364 1.2727273 38.66364     6.072727
2 0.2279070 15.65116 8.395349 0.6046512 14.30000     5.805814

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2
2 1 2 1 2 2 2 2 2 2 2 2
 [69] 1 2 1 2 1 2 2 2 2 2 2 2 1 2 2 1 1 2 2 1 2 1 2 2 2 1 2 1 1 1 2 2 1 2 2 2 2 2 1 2 2

Within cluster sum of squares by cluster:
[1]  9187.442 13622.711
 (between_SS / total_SS =  39.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
```

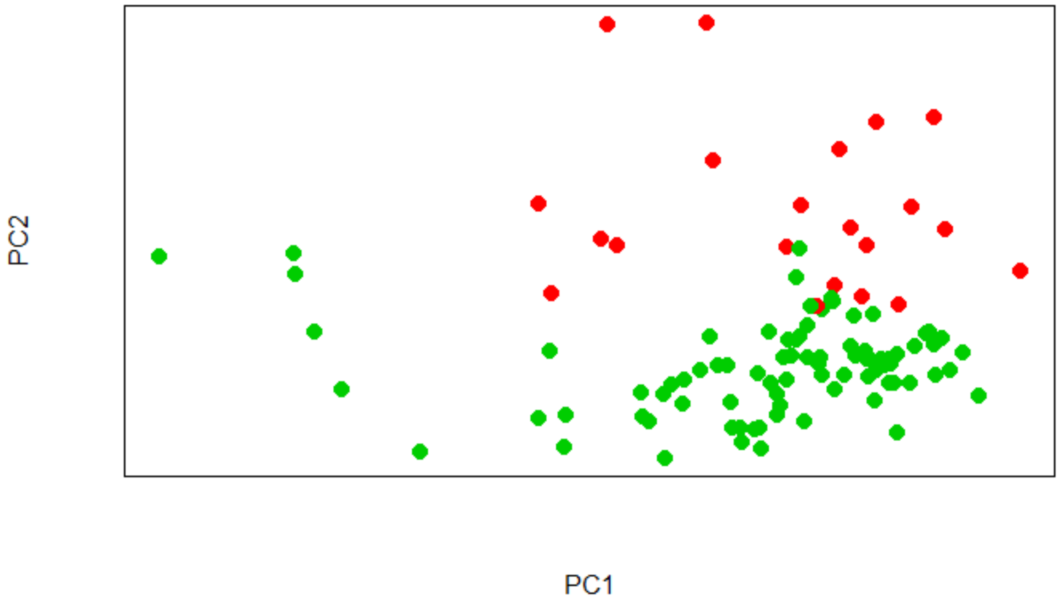
b. In order to visually display the two clusters obtained in part a, plot the first two principal components and colour according to the k-means classes. [MARK = 2]

R code:

```
pp = prcomp(Y, scale = TRUE) #do PCA MARK (1)
```

plot(pp$x[,1:2], col = fitted(km, "classes")+1, xaxt = "n", yaxt = "n", pch = 20, cex = 2) # use appropriate colour coding ***MARK (1)***

R output:



c. Construct the misclassification table and misclassification rate and discuss the accuracy of predicting presence of Asthma. [MARK = 3]

R code:

mis = table(truth = env_PCA[,8], cluster = fitted(km, "classes"))
#identifying response for true value ***MARK (1)***
mis

R output:

```
     cluster
truth   1  2
  FALSE 19 47
  TRUE   3 39
```

\# Output **MARK (1)**

Interpretation:

From above table it can be seen clearly that Cluster 1 (or 2) includes most of the patients with Asthma while Cluster 2 (or 1) shows the once without Asthma. Even though the cluster shows good results in classifying Asthma patients it did not perform well in classifying the non-Asthma patients. The misclassification rate is 46.29%. Even though, the misclassification rate is high, this clustering technique performs well in identifying Asthma patients.

\# Interpretation **MARK (1)**

d. Alternatively use K Means Clustering method to identify three clusters using K=3. [MARK = 2]

R code:

km2 = kmeans(Y, 3, nstart = 20) #Writing the correct coding with k value ***MARK (1)***
km2

R output: ***MARK (1)***

```
K-means clustering with 3 clusters of sizes 21, 64, 23

Cluster means:
         co        o3       no2       so2    ppm10 maxwindspeed
1 0.2000000 32.333333  7.571429 1.285714 38.84762     6.142857
2 0.1703125 18.859375  5.718750 0.453125 12.76406     6.265625
3 0.3956522  6.608696 16.347826 1.043478 19.46522     4.473913

Clustering vector:
  [1] 2 2 3 2 3 2 3 2 3 2 1 2 2 2 2 2 2 3 3 3 3 2 3 2 2 3 2 2 3 2 3 1 2 1 2 3 2 3 2 2 2 2 1 2 2 2 2 2 1 3 2 3 2 3 2 1 2
2 1 2 1 2 2 2 2 2 2 2
 [69] 1 2 1 2 1 2 2 2 2 2 2 1 3 2 1 1 2 2 1 2 1 2 3 2 1 2 1 1 1 2 2 1 2 2 3 3 2 3 2 3

Within cluster sum of squares by cluster:
[1] 8665.179 5456.977 3451.198
 (between_SS / total_SS =  53.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
```

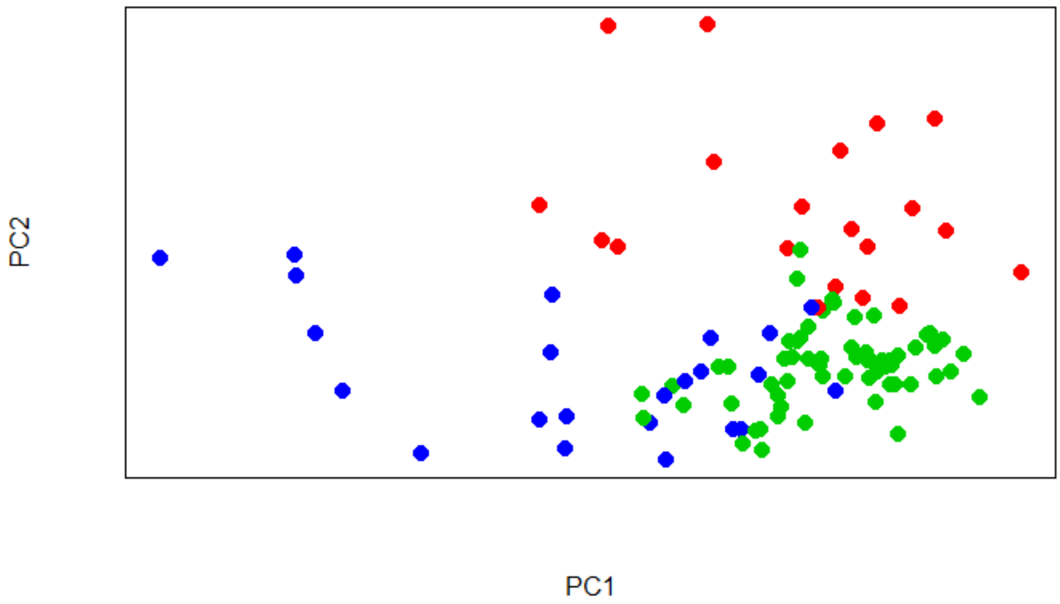
e. In order to visually display the three clusters obtained in part d, plot the first two principal components and colour according to the k-means classes. [MARK = 2]

R code:

plot(pp$x[,1:2], col = fitted(km2, "classes")+1, xaxt = "n", yaxt = "n", pch =20, cex = 2)

 # use appropriate colour coding ***MARK (1)***

R output: ***MARK (1)***



f. Compare results obtained in parts “a and b” with parts “d and e” and justify most suitable number of clusters for this data set using total within cluster variation and total between cluster variation. [MARK = 3]

Interpretation:
Total within cluster variation when 2 clusters considered = 22810.15
Total within cluster variation when 3 clusters considered = 17573.35

There is a huge reduction in the total within SS variation when the cluster number is 3. Also the between SS variation is 39.3% for two clusters whereas it was 53.3% for three clusters. Thus, it is better to cluster the data into 3 sets than 2 sets.

Mentioning the values of Total within cluster variation **MARK (1)**
Mentioning the value of total between cluster variation **MARK (1)**
final summary **MARK (1)**

#Mention R^2 value **MARK (1)**
#Give RSE. **MARK (1)**
OR
#Only ANOVA table . **MARK (0.5)**

Question 2 (2 + 4 + 3 + 2 + 4 + 2 + 3= 20)

This Question uses the data set “Envdata” used in Question 1

a. Calculate the mean and the variance for each variable and discuss if scaling is necessary and justify your findings. [MARK = 2]

R code:
library(readxl)
env = read.csv("RELEVANT PATH")

```
head(env)
Y_num = env[ , c(-1,-8)]
apply(Y_num, 2, mean)
apply(Y_num, 2, var)
# Using proper R code removing inappropriate variables MARK (1)
```

R output:

```
   co         o3        no2        so2      ppm10 maxwindspeed
 0.2240741  18.8703704   8.3425926   0.7407407  19.2629630   5.8601852
        co          o3         no2          so2        ppm10 maxwindspeed
0.05119072 117.59051575  38.82545864   1.50224991 179.09637245  14.28335324
```

Interpretation:
Since the value of mean and variance values of selected variables differ largely, it is advisable to scale the variables (standardise the variables).
Provide justification stating need for scaling **MARK (1)**

b. Apply scaling and derive the principal components. (R code and output) [MARK = 4]

R code:

```
pp = prcomp(Y_num, scale = TRUE) #Make sure in the code "scale = TRUE" MARK (1)
pp
# Use the appropriate R code MARK (1)
#Using appropriate Variables in PCA MARK (1)
```

R output: MARK (1)

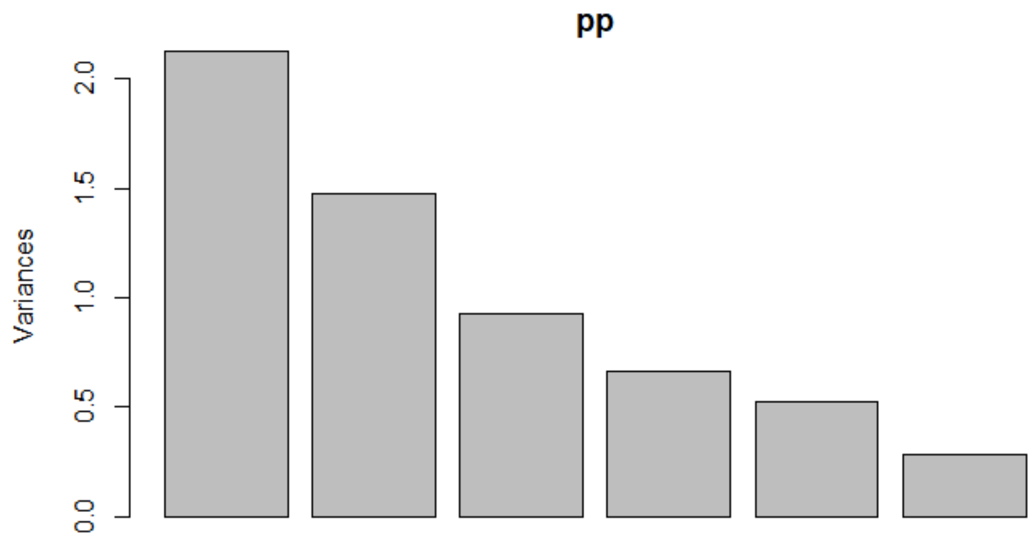
```
Standard deviations (1, .., p=6):
[1] 1.4587570 1.2144469 0.9637641 0.8132041 0.7254836 0.5297909

Rotation (n x k) = (6 x 6):
                  PC1        PC2        PC3         PC4        PC5         PC6
co          -0.5450045 0.02732184 -0.1096117 -0.31339377  0.7120811  0.29143389
o3           0.3612149 0.52018122 -0.4198813  0.01863647  0.4073065 -0.50634875
no2         -0.6047801 0.02365796  0.1104536 -0.22452450 -0.2543653 -0.71159567
so2         -0.2997724 0.45848213  0.3808967  0.73427757  0.1111391  0.05773126
ppm10       -0.1117122 0.69792161 -0.1372112 -0.37152683 -0.4504230  0.37508112
maxwindspeed  0.3230972 0.17551320  0.7972300 -0.41692930  0.2170425 -0.09104995
```

c. Give the Scree Plot and give the percentage variation captured by each principal component. [MARK = 3]

R code:

```
screeplot(pp) MARK (1)
summary(pp) #Any graphs showing variation captured or the given output MARK (1)
```

R output:



Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.4588	1.2144	0.9638	0.8132	0.72548	0.52979
Proportion of Variance	0.3547	0.2458	0.1548	0.1102	0.08772	0.04678
Cumulative Proportion	0.3547	0.6005	0.7553	0.8655	0.95322	1.00000

Interpretation: **MARK (1)**

By looking at the scree plot, we can see clearly that most of the variation is explained by first PC. After 3rd PC there is no much variation explained by PCs. About 36% of the variation is explained by first PC. First two PCs together explain about 60% of the variation in the data. First three PCs capture about 76% of the variation in the data.

d. Select the number of principal components most suitable to represent the dataset and justify your answer. [MARK = 2]

Interpretation:

From part c, the optimal number of PCs that can be used to explain variation is first three PCs. About 76% of the variation in the data is explained by the first three PCs.

#Number of PCs mentioned **MARK (1)**

#Justification **MARK (1)**

Since the answer is subjective marks is given if justified properly

e. Derive and give the principal component loading vectors for the given dataset and explain the results/output. [MARK = 4]

R code:

pp$rotation ***MARK (2)***

R output:

PC1 PC2 PC3 PC4 PC5 PC6

```
co          -0.5450045 0.02732184 -0.1096117 -0.31339377  0.7120811  0.29143389
o3           0.3612149 0.52018122 -0.4198813  0.01863647  0.4073065 -0.50634875
no2         -0.6047801 0.02365796  0.1104536 -0.22452450 -0.2543653 -0.71159567
so2         -0.2997724 0.45848213  0.3808967  0.73427757  0.1111391  0.05773126
ppm10       -0.1117122 0.69792161 -0.1372112 -0.37152683 -0.4504230  0.37508112
maxwindspeed  0.3230972 0.17551320  0.7972300 -0.41692930  0.2170425 -0.09104995
```

Interpretation: **MARK (2)**

no2 and co contribute most to PC1 with small contribution from o3, maximum wind speed and so2. ppm10 and o3 contribute mostly to PC2 with small contribution from so2 and maximum wind speed. PC3 mostly represents maximum wind speed.

f. Give the first two principal components using loading parameters obtained in part e. [MARK = 2]

Interpretation:

#Each **MARK (1)**

$$PC1 = -0.5450045 \times co + 0.3612149 \times o3 - 0.6047801 \times no2 - 0.2997724 \times so2 - 0.1117122 \times ppm10 + 0.3230972 \times maxwindspeed$$

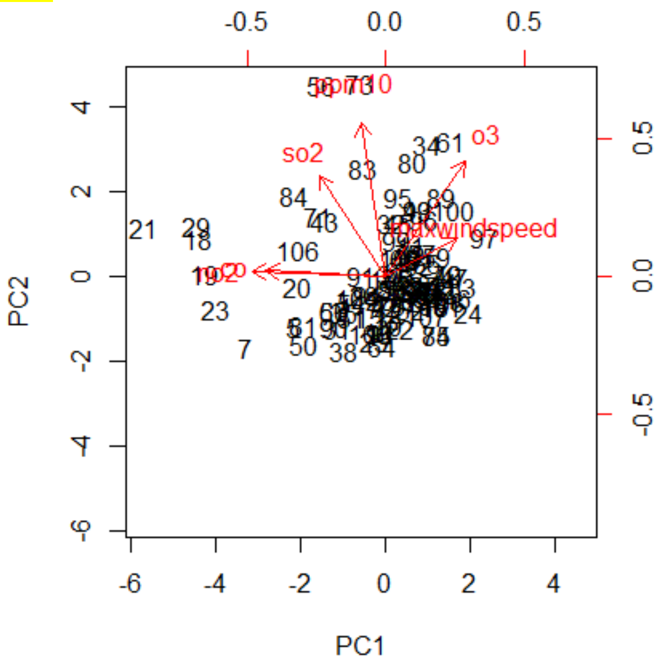
$$PC2 = 0.02732184 \times co + 0.52018122 \times o3 + 0.02365796 \times no2 + 0.45848213 \times so2 + 0.69792161 \times ppm10 + 0.17551320 \times maxwindspeed$$

g. Construct the Biplot and interpret it in terms of original variable contributions. [MARK = 3]

R code:

biplot(pp, scale = 0) ***MARK (1)***

R output: **MARK (1)**



Interpretation: **MARK (1)**

Above graph depicts that co and no2 provide high contribution to first PC while ppm10 highly contribute to second PC. co and no2 seem have a strong positive association while showing a negative association with maximum wind speed. so2 and ppm 10 shows a significant positive association. Similarly, ppm10 and o3 also shows a significant positive association. o3 provides almost same contribution to both the PCs.

Question 3 (3 + 3 + 2 + 3 + 2 + 3 + 4 + 2 + 3 = 25)

This Question uses the data set “Admission”.

This dataset contains the following variables.

Serial. No: observation number
GRE : Graduate Record Examinations score (out of 340)
TOEFL : Test of English as a Foreign Language Scores (out of 120)
Uni_R : University rating (out of 5)
SOP : Statement of Purpose score (out of 5)
CGPA : Undergraduate GPA (out of 10)
Chance : Chance of admission

a. Construct the matrix plot and correlation matrix and comment. [MARK = 3]

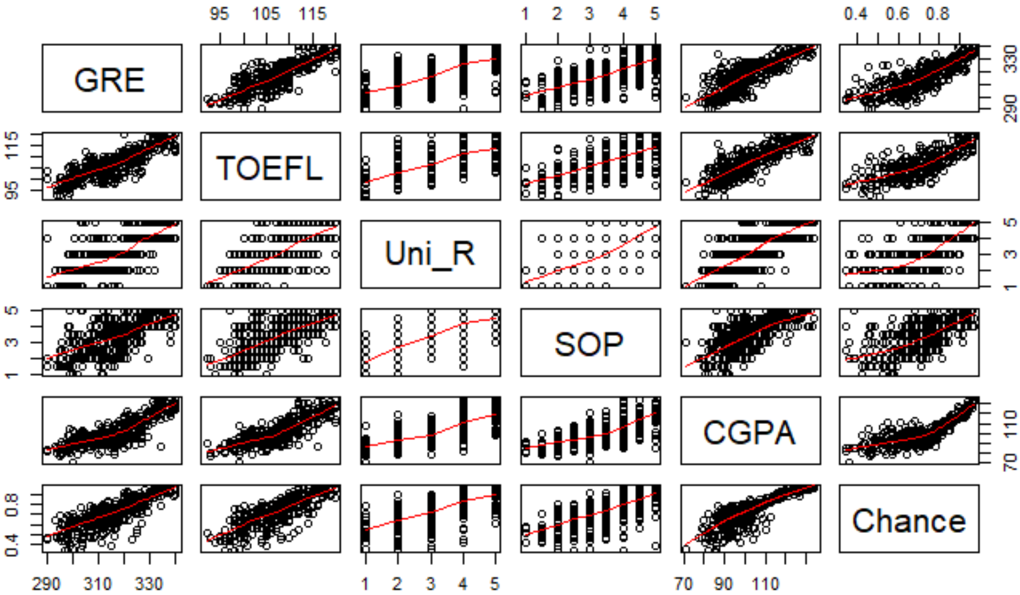
R code:

```
admin=read.csv("RELEVANT PATH")
head(admin)
X = admin[,c(2:7)]
str(X)
cor(X)
pairs(X, panel = panel.smooth)
#Considering only appropriate variables finding correlations MARK (1)
```

R output:

```
'data.frame':	400 obs. of  6 variables:
 $ GRE   : int  298 297 300 303 299 304 315 311 304 299 ...
 $ TOEFL : int  98 96 99 99 97 101 105 104 100 94 ...
 $ Uni_R : num  2 2 1 3 3 2 2 2 4 1 ...
 $ SOP   : num  4 2.5 3 2 5 2 2 2 1.5 1 ...
 $ CGPA  : num  85.3 81.7 71 85.8 85.8 ...
 $ Chance: num  0.34 0.34 0.36 0.36 0.38 0.38 0.39 0.42 0.42 0.42 ...
             GRE     TOEFL     Uni_R       SOP      CGPA    Chance
GRE    1.0000000 0.8359768 0.6689759 0.6128307 0.8162660 0.8026105
TOEFL  0.8359768 1.0000000 0.6955898 0.6579805 0.8201615 0.7915940
Uni_R  0.6689759 0.6955898 1.0000000 0.7345228 0.7419442 0.7112503
SOP    0.6128307 0.6579805 0.7345228 1.0000000 0.6949374 0.6757319
CGPA   0.8162660 0.8201615 0.7419442 0.6949374 1.0000000 0.8285208
Chance 0.8026105 0.7915940 0.7112503 0.6757319 0.8285208 1.0000000
```

Displaying Matrix plot **MARK (1)**



Interpretation:

Chance is highly positively correlated with GRE and CGPA. Chance is also positively correlated with TOEFL University Ranking and Statement of Purpose. It can be seen clearly that all the predictor variables are also positively correlated among each other.

From the matrix plot it can be seen clearly that there is a strong positive correlation between GRE and TOEFL, GRE and CGPA, GRE and Chance. TOEFL and CGPA, TOEFL and Chance.The plot also suggests a non-linear relationship between Chance and CGPA. Uni_R and SOP doesnot show any significant pattern.

Proper interpretation **MARK (1)**

b. Derive a multiple linear regression model to describe the "Chance of admission" in terms of other numeric variables and give the estimated model. (No need to prove the significance of the model) [MARK = 3]

R code:

model1 = lm(Chance ~ GRE+TOEFL+Uni_R+SOP+CGPA, data = admin) #Proper R Code ***MARK (1)***

summary(model1)

R output: **MARK (1)**

```
Call:
lm(formula = Chance ~ GRE + TOEFL + Uni_R + SOP + CGPA)

Residuals:
      Min        1Q    Median        3Q       Max
-0.305813 -0.021621  0.009739  0.045337  0.148117

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1965199  0.1387597  -8.623  < 2e-16 ***
GRE          0.0033627  0.0006195   5.428 9.99e-08 ***
TOEFL        0.0035547  0.0012068   2.946  0.00341 **
```

```
Uni_R        0.0121451  0.0052507   2.313  0.02123 *
SOP          0.0145156  0.0055349   2.623  0.00907 **
CGPA         0.0038074  0.0005840   6.519 2.17e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07071 on 394 degrees of freedom
Multiple R-squared:  0.7572,     Adjusted R-squared:  0.7541
F-statistic: 245.8 on 5 and 394 DF,  p-value: < 2.2e-16
```

Interpretation:

The estimated model is

$$Chance = -1.1965199 + 0.0033627 \times GRE + 0.0035547 \times TOEFL + 0.0145156 \times SOP + 0.0038074 \times CGPA$$

#Writing the correct equation **MARK (1)**

c. Describe the model accuracy (Not the model assumptions) [MARK = 2]

R code:

anova(model1)

R output:

```
Analysis of Variance Table

Response: Chance
          Df Sum Sq Mean Sq  F value    Pr(>F)
GRE        1 5.2273  5.2273 1045.445 < 2.2e-16 ***
TOEFL      1 0.3921  0.3921   78.421 < 2.2e-16 ***
Uni_R      1 0.2370  0.2370   47.390 2.304e-11 ***
SOP        1 0.0757  0.0757   15.145 0.0001169 ***
CGPA       1 0.2125  0.2125   42.498 2.169e-10 ***
Residuals 394 1.9700  0.0050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

$R^2 = 75.72\%.$

Thus, about 76% of the variation in the data is explained by the model.

Residual Standard Error = 0.07071 ($= \sqrt{0.0050}$ is very small

#Mention R^2 value **MARK (1)**

#Give RSE. **MARK (1)**

OR

#Only ANOVA table . **MARK (0.5)**

d. Select one suitable explanatory variable and derive the best polynomial regression model and give the estimated model to describe "Chance of admission". (No need to prove the significance of the model) [MARK = 3]

Explanation:

From Question 3.a, it can be seen clearly using matrix plot that there is non-linear relationship between CGPA and Chance. Hence it is appropriate to fit a polynomial regression with chance as Response variable using CGPA as predictor.

#Selecting the variable with non-linear relationship **MARK (1)**

#Output **MARK (1)**

R code:

```
model2 = lm(Chance ~CGPA + I(CGPA*CGPA), data = admin)
summary(model2)
```

R output:

```
Call:
lm(formula = Chance ~ CGPA + I(CGPA * CGPA), data = admin)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30145 -0.02735  0.01181  0.04439  0.20229 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -6.751e-01  2.333e-01  -2.894  0.00401 ** 
CGPA            1.812e-02  4.481e-03   4.043 6.34e-05 ***
I(CGPA * CGPA) -4.209e-05  2.128e-05  -1.978  0.04864 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07966 on 397 degrees of freedom
Multiple R-squared:  0.6895,	Adjusted R-squared:  0.6879 
F-statistic: 440.8 on 2 and 397 DF,  p-value: < 2.2e-16
```

Interpretation:

The estimated model is:

$$Chance = -0.6751 + 0.01812 \times CGPA - 0.00004209 \times CGPA^2$$

#Writing the equation **MARK (1)**

e. Describe the model accuracy of the polynomial model. (Not the model assumptions) [MARK = 2]

R code:

```
anova(model2)
```

R output:

```
Analysis of Variance Table

Response: Chance
                Df Sum Sq Mean Sq  F value  Pr(>F)    
CGPA             1 5.5703  5.5703 877.6962 < 2e-16 ***
I(CGPA * CGPA)   1 0.0248  0.0248   3.9119 0.04864 *  
Residuals      397 2.5195  0.0063                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

$R^2 = 68.95\%$.

Thus, about 69% of the variation in the data is explained by the model.

Residual standard error is 0.07966

The variance of error is 0.0063 which is very small.

#Mention R^2 value **MARK (1)**

#Give RSE. **MARK (1)**

OR

#Only ANOVA table . **MARK (0.5)**

f. Improve the model by combining the two models derived from parts b and d and obtain the best final model. (No need to prove the significance of the model) [MARK = 3]

R code:

```
model3 = lm(Chance ~GRE + TOEFL +Uni_R + SOP +  CGPA + I(CGPA*CGPA), data = admin)
summary(model3)
# Proper R code with correct predictors MARK (1)
```

R output:

```
Call:
lm(formula = Chance ~ GRE + TOEFL + Uni_R + SOP + CGPA3 + I(CGPA3 *
    CGPA3), data = admin)

Residuals:
      Min        1Q    Median        3Q       Max
-0.310213 -0.022095  0.009725  0.042875  0.153072

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.468e+00  2.475e-01  -5.932 6.57e-09 ***
GRE              3.370e-03  6.190e-04   5.445 9.14e-08 ***
TOEFL            3.478e-03  1.207e-03   2.881  0.00418 **
Uni_R            1.212e-02  5.246e-03   2.311  0.02135 *
SOP              1.388e-02  5.550e-03   2.501  0.01277 *
CGPA3            9.156e-03  4.081e-03   2.244  0.02541 *
I(CGPA3 * CGPA3) -2.517e-05  1.901e-05  -1.324  0.18616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07064 on 393 degrees of freedom
Multiple R-squared:  0.7583,	Adjusted R-squared:  0.7546
F-statistic: 205.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

Interpretation:

When the two models are combined, the polynomial with two degrees of freedom becomes insignificant. Hence, the best model is model 1 with larger R^2 value.

Selecting the best model **MARK (1)**

Giving proper justification for the choice **MARK (1)**

g. Test the significance of the best model obtained in part f (Show all steps in the significance test when testing for one parameter and describe logically the significance of other parameters briefly). [MARK = 4]

Interpretation:
To test the significance of the parameters, we use the following hypothesis test.

$$H_0 : \beta = 0 \;\; (The\, slope\, parameter\, is\, not\, significantly\, different\, from\, zero)$$
$$H_1 : \beta \neq 0 \;\; (The\, slope\, parameter\, is\, significantly\, different\, from\, zero)$$

Mentioning the relevant hypothesis statement **MARK (1)**

We considering model1 as the best model, the slope parameter of GRE is significant since the p-value is less than 0.05 suggesting a significant linear relationship between GRE and Chance.

#Clearly mention the decision criteria using p-value **MARK (1)**

Similarly,
There is a significant linear relationship between TOEFL and Chance.
There is a significant linear relationship between Uni_R and Chance.
There is a significant linear relationship between SOP and Chance.
There is a significant linear relationship between CGPA and Chance.

#Properly mentioning the significance of all the variables in the best model **MARK (2)**

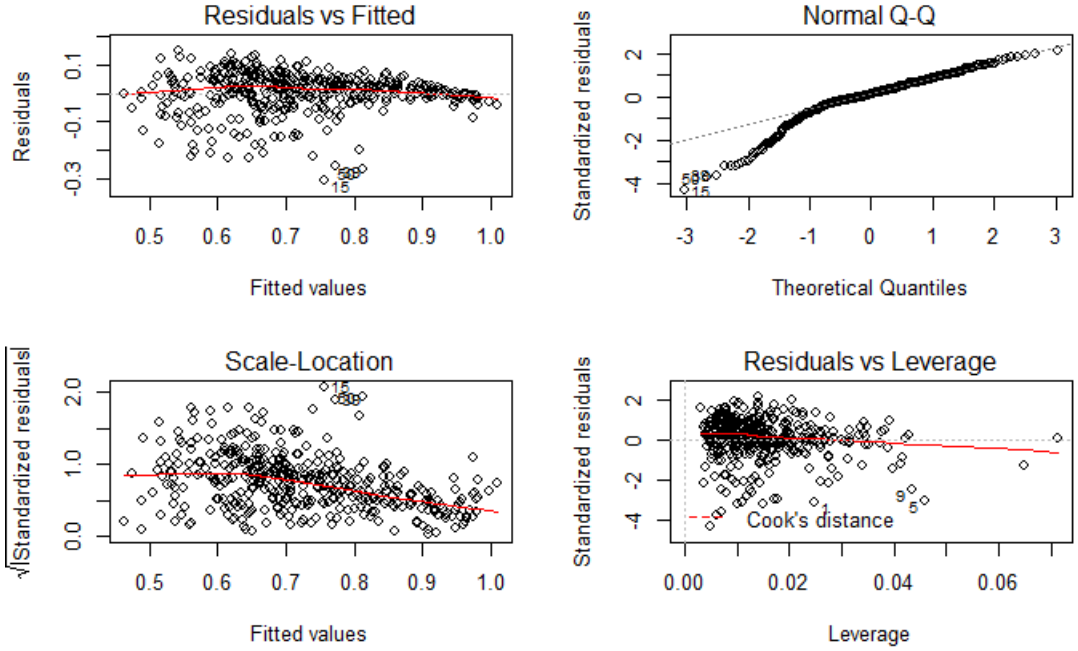
Thus, the model is adequate.

h. Test the model assumptions of the final model obtained in part f. (Using 4 default diagnostic plots) [MARK = 2]

R code:

```
par(mfrow = c(2,2))
plot(model1)
```

R output:



Interpretation:

Graph 1:

It appears to have a V-shape suggesting heteroscedasticity. Thus, the constant variance assumption seems to be violated.

Graph 2: (Normal Q-Q)

It does not appear to be a straight line although it is linear to some extent. Therefore normality assumption is not met.

Graph 3:

Same as graph 1.

Graph 4:

There are few influential observations such as 93, 375.

 # Each Graph with interpretation ***MARK (0.5)***

i. Describe the model accuracy of the Final model using all the above findings. [MARK = 3]

Interpretation:

The p-value (2.2e-16) is very small supporting strong linear relationship. Residual Standard error = 0.07966 (RSE) is a very small values. Also, 75.7% of the variation is explained by the regression.

Even though, most of the variation is captured by the model, model assumptions were ***not met*** (using Q3.h). This suggests that there is some non-linearity in the data which is not captured by the model.

#Mention RSE **MARK (1)**

Mention R2 with interpretation **MARK (1)**

State the violation of model assumptions **MARK (1)**

- END -