WESTERN SYDNEY UNIVERSITY

School of Computer, Data and Mathematical Sciences

COMP 7006 Data Science

Computer Based Assignment – PART B

Spring, 2025 – Due date October 10th (not Oct 3rd) Midnight

STUDENT ID:	22220716
STUDENT FIRST NAME:	Mayur
STUDENT SURNAME:	Kalpe

QUESTIONS FORMAT:	Word processed document in PDF format; logically presenting answers to each question incorporating R outputs including graphs and charts.
TOTAL MARKS:	**60 Marks**
UNIT CO-ORDINATOR:	Dr. Liwan Liyanage
TUTOR:	Mr Timofei Latskov
TOTAL PAGES:	16

INSTRUCTIONS

Please note that you are expected to answer the questions clearly in this document. Use the template included where relevant to answer. Give the R outputs, comments, and discussion clearly and logically in the rectangular box provided under each question. Attach all the R commands in the Appendix. Write the resulting **model equation** to the relevant questions. Once completed submit the answer scripts as a **PDF** via TurnItin link within vUWS site.

Please note that **10 Marks** are allocated for organization, reasoning, logical flow, and the inclusion of all correct R codes and outputs in the Appendix for both Part A and Part B.

SCENARIO
Recent public health data indicate a troubling increase in asthma disease rates within specific suburban areas, attracting significant attention from public health practitioners. Determined to uncover the root causes and identify actionable risk factors to address this issue, the public health team has embarked on a comprehensive study. They have collected patient records and relevant information on medical factors and environmental factors, as provided in the **EnvData.csv** dataset.

Data Description:

Variables	Description
patient_id	Patient ID
triage	Triage
postcode	Postcode of the admitted patients
age	Age at the time of admission
gender	Gender
suburb	Suburb of the admitted patients
co	Hourly records of atmospheric concentration of Carbon monoxide in parts per million (ppm)
o3	Hourly records of atmospheric concentration of ground level Ozone in parts per billion (ppb)
no2	Hourly records of atmospheric concentration of Nitrogen dioxide in parts per billion (ppb)
so2	Hourly records of atmospheric concentration of Sulphur dioxide in parts per billion (ppb)
ppm10	Hourly records of atmospheric concentration of particulate matter less than 10µm in diameter ($\mu g/m^3$)
visibility_reduction	Hourly records of visibility reduction i.e. minimum visible distance – 20km
aqi	Air Quality Index
precipitation	Half-hourly records of rainfall in millimetre (mm)
relativehumidity	Half-hourly records of Relative Humidity (%)
vapourpressure	Half-hourly records of Vapour Pressure
windspeed	Half-hourly records of Wind Speed (km/hr.)

winddirection	Half-hourly records of Wind Direction
maxwindspeed	Half-hourly records of Maximum Wind Speed (km/hr.)
asthma	Binary indicator of asthma attacks

* Please note that this is a simulated data generated to resemble the real-world data for the purpose of this assignment.

Consider the scenario described, the data set provided and your answers in Part A to answer the following questions.

1. Build a logistic regression model incorporating polynomial terms. Clearly outline and explain each step of the process involved. [*This question is designed to assess your critical thinking and analytical skills. Please note that guidance on how to complete the task will not be provided.*] **(8 Marks)**

The logistic regression model was built with the polynomial term by the following manner:

1. Data wrangling and exploration
2. Building a base logistic regression model using ag, gender, o3,no2 and RH variables
3. Adding the polynomial terms to check whether it improves the model
4. Model comparison: used AIC, loocv, k-fold method

The code is attached below in the Appendix.

2. Give the resultant accepted model (i.e. write the model equation) based on your findings above. Justify your answer clearly. **(3 Marks)**

The Accepted model in the logistic regression model with the polynomial term is the one with *quadratic term with o3*. It also includes age, gender, relative humidity and no2 as other base factors. The reason to select that model was- As per the k10 method, fit_o3 model has the lowest value 0.219 amongst others which was more than 2.22

The resultant model is written as below:

fit_final = glm(asthma ~ age +gender+relativehumidity+no2 + poly(o3, 2, raw = TRUE), data = envdata, family = binomial)

or it can be also written as:-

logit P (asthma = yes) = ß0 + ∑ ß(age_group) + ß Male + B relativehumidity + ß no2 + ßo3^2

(Where all the ß values can be found in summary (fit_final) coefficients)

3. Use decision tree model to answer the research question. Clearly outline and explain each step of the process involved [*Hint: model building, improvement and evaluation*]. **(12 Marks)**

The decision tree model is created as the below bullet points, the codes, charts and outputs are attached in the appendix:

- Model building: create a training dataset (50-50 split) for the decision tree using tree function
- Improvement: Model was improved using prune.misclass in the cross validation
- Accuracy- predicted accuracy for the test and pruned dataset

The codes and outputs attached below in appendix

4. Give the resultant model and interpret it. Clearly describe the terminal nodes [i.e. list the profiles]. *[Include the relevant R output]* **(5 Marks)**

Decision tree model is:

pruned_dt1 <- prune.misclass(dt2, best = 3)

Terminal nodes:

1) root 54 70.05 No (0.6481 0.3519) - Root node is 54, 64% is No and 35% is yes
  2) o3 < 13.5 19 25.01 Yes (0.3684 0.6316)
    4) vapourpressure < 16.85 14 11.48 Yes (0.1429 0.8571) *
    5) vapourpressure > 16.85 5 0.00 No (1.0000 0.0000) *
  3) o3 > 13.5 35 35.03 No (0.8000 0.2000) *

Interpretation:

- The main node of the resultant model is o3 (13.5), that means for high ozone the model mostly predicts no asthma.
- After o3, vapour pressure is the second most relevant factor for asthma. For lower vapour pressure, the possibility of asthma increases.

5. Compare the different resultant models (Part A Question 5, Part B Question 2 and Question 4) you obtained above. **(12 Marks)**

Following are the models were created from part A and B combines.

1. Logistic regression (Part A) : $x=\beta0+\beta1 \cdot age\_mid+\beta2 \cdot O3$
2. Polynomial Logistic regression (Part B) : ***fit_final = glm(asthma ~ age +gender+relativehumidity+no2 + poly(o3, 2, raw = TRUE), data = envdata, family = binomial)***
3. Decision tree : ***pruned_dt1 <- prune.misclass(dt2, best = 3)***

For the logistic regressions modelling, I have used a common splitting the dataset technique either 50-50 or 70-30 split for training and test consecutively.

For the glm's AIC tests along with loocv and k10 method was performed successfully.

All the models above were close enough for their test accuracy, however the Part A logistic model has given a slightly better AIC value and test accuracy than that of other two.

6. Give the final accepted model based on your findings above and Part A. Justify your answer. **(5 Marks)**

The selected final model amongst all the three is:

1. Logistic regression (Part A) : $x=\beta0+\beta1 \cdot age\_mid+\beta2 \cdot O3$

The reason for selecting this model as the final accepted model is the model complexity (Least complex amongst others) and model accuracy. This model uses just 2 predictors age_mid and o3.

By the model we can interpret that for every unit increase of the age_mid variable, it multiplies the odds by $\beta1$, and for the o3, it multiplies by $\beta2$.

7. Apply an unsupervised learning technique of your choice to identify any interesting or hidden patterns in the dataset. Provide a clear explanation of the technique used and thoroughly describe your findings. **(10 Marks)**

I have applied Principal Component Analysis technique as an unsupervised learning technique for above research question.

The pve (proportion of variance explained) explains that PC1 shows the maximum variance.

The cumulative sum of proportion of variance explained aka "cumsum(pve)" increases in the first few principal components. pca_1$rotation shows the loadings in our PC analysis. Larger the reading in their absolute version, stronger it is.

For the PC1- the largest absolute reading is of visibility_reduction that is 0.47. the data can be interpreted from this is lower the visibility, higher the chances of asthma.

Ppm10 and aqi are the other two top loaders in PC1.

Biplot interpretation: As PC1 and PC2 is having most variance, biplot shows the same.

Acute angle between two arrows suggests positive co relation, obtuse angle interprets negative correlation.

Values showing up towards the direction of the variable arrow suggests higher values of that variable.

The codes and outputs are attached below in the appendix.

8. What are your conclusion and recommendations for this problem? [Hint: Use your results and findings from previous questions to answer this question] **(5 Marks)**

Conclusion:

We have concluded that the final selected model will be from the part A that is:

$x=\beta 0+\beta 1 \cdot age\_mid+\beta 2 \cdot O3$

- This particular model is simple enough and having slightly higher accuracy rate than the others. In the polynomial regression model, adding o3^2 did not deliver better results.
- For the **decision tree**, the pruned tree interpretation is as follows:

Node 1: $O_3 < 13.5$ & VapourPressure < 16.85 = "Yes" profile

Node 2: O3 >13.5 = "No" profile

This model was also simple enough with less variables, with slightly lower accuracy than the Part A model.

- In Principal component analysis the dominant variables (or factors) in the PC1 and PC2 are Air quality and visibility reduction.

Recommendations:

- To use the final accepted model as a primary operation in the health space.
- They can use the Node 1 of the decision tree model generically. For example, which localities have low o3 and low vapour pressure assign them as high risk of asthma.

- If any other predictors seem relevant in future, do add it in the dataset and re-assess the models.

--- End of questions ---

APPENDIX

[Attach all your R codes and outputs here.]

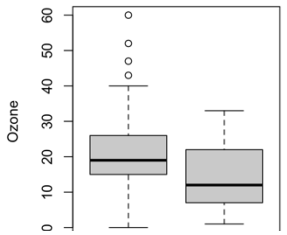
Q1. Model building:

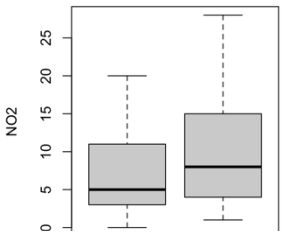
- envdata <- read.csv("EnvData.csv")
- attach(envdata)
- names(envdata)
- head(envdata)
- summary(envdata$asthma)

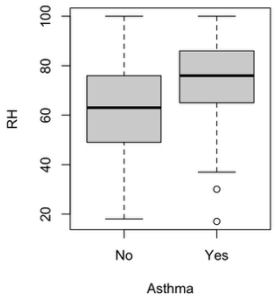
The Boxplot for the analysis are:

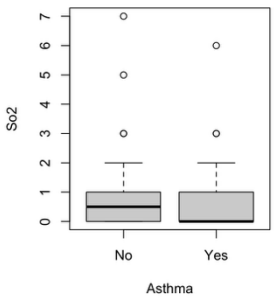
par(mfrow = c(1,2))

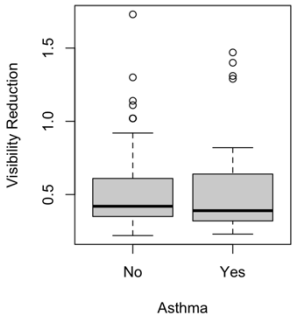
- boxplot(o3 ~ asthma, data = envdata, ylab="Ozone", xlab="Asthma")
- boxplot(no2 ~ asthma, data = envdata, ylab="NO2", xlab="Asthma")
- boxplot(relativehumidity ~ asthma, data = envdata, ylab="RH", xlab="Asthma")
- boxplot(so2 ~ asthma, data = envdata, ylab="So2", xlab="Asthma")
- boxplot(visibility_reduction ~ asthma, data = envdata, ylab="Visibility Reduction", xlab="Asthma")











#lets create a baseline logistic model (without poly)

- fit_base <- glm(asthma ~ age + gender + o3 + no2 + +relativehumidity, data = envdata, family = binomial)
- summary (fit_base)
- AIC(fit_base) # *will compare later with the poly models. Lower the better*

Output:

```
Call:
glm(formula = asthma ~ age + gender + o3 + no2 + +relativehumidity,
    family = binomial, data = envdata)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.57661    1.78126   0.885  0.37610
age05 to 09       -2.09689    1.34912  -1.554  0.12012
age10 to 14       -3.74805    1.42298  -2.634  0.00844 **
age15 to 19       -3.13916    1.41825  -2.213  0.02687 *
age20 to 24       -1.58871    1.42056  -1.118  0.26341
age25 to 29       -2.17551    1.49962  -1.451  0.14686
age30 to 34       -1.64802    1.43054  -1.152  0.24931
age35 to 39       -4.27575    1.67588  -2.551  0.01073 *
age40 to 44       16.37556 3956.18052   0.004  0.99670
age45 to 49       -3.59931    1.67655  -2.147  0.03181 *
age50 to 54       -2.27525    2.15709  -1.055  0.29153
age55 to 59       -3.33207    1.53001  -2.178  0.02942 *
age60 to 64       -2.22373    1.57062  -1.416  0.15682
age65 to 69      -19.09708 2243.59801  -0.009  0.99321
age70 to 74      -19.87813 2771.17141  -0.007  0.99428
age80 to 84      -19.24470 2637.27175  -0.007  0.99418
age90 to 94       14.38322 3956.18057   0.004  0.99710
genderMale        -0.64959    0.55755  -1.165  0.24399
o3                -0.05802    0.03574  -1.623  0.10452
no2                0.01942    0.05303   0.366  0.71428
relativehumidity   0.02588    0.01539   1.681  0.09270 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 144.342  on 107  degrees of freedom
Residual deviance:  98.683  on  87  degrees of freedom
AIC: 140.68

Number of Fisher Scoring iterations: 16

[1] 140.6832
```

 # Now let's try adding the polynomial terms first in Ozone (o3)

- fit_o3 <- glm(asthma ~ age +gender+relativehumidity+no2 + poly(o3, 2, raw = TRUE), data = envdata, family = binomial)

#Now the poly model with NO2 only

- fit_no2 <- glm(asthma ~ age +gender+relativehumidity + o3 + poly(no2, 2, raw = TRUE), data = envdata, family = binomial)

#The next model will consist of both o3 and no2

- fit_both <- glm(asthma ~ age +gender+relativehumidity + poly(o3,2,raw = TRUE) + poly(no2, 2, raw = TRUE), data = envdata, family = binomial)

#Model comparison

- AIC(fit_base, fit_o3, fit_no2, fit_both)
- library(boot)
- set.seed(4)
- fit_all <- list (fit_base=fit_base, fit_o3=fit_o3, fit_no2=fit_no2, fit_both=fit_both)
- loocv <- sapply(fit_all, function(f) cv.glm(envdata,f)$delta[1])
- k10 <- sapply (fit_all, function(f) cv.glm (envdata, f, K=10)$delta[1])
- loocv
- k10

Outputs:

	df <dbl>	AIC <dbl>
fit_base	21	140.6832
fit_o3	7	138.3020
fit_no2	7	137.7805
fit_both	8	139.7342

```
fit_base    fit_o3   fit_no2  fit_both
0.2184372 0.2201327 0.2225193 0.2244563
 fit_base    fit_o3   fit_no2  fit_both
0.2142399 0.2214688 0.2302847 0.2298773
```

As per the k10 method, **fit_o3** model has the lowest value 0.219. Hence:-

```{r}

- fit_final <- fit_o3
- summary(fit_final)
```

Question 2:

- summary(fit_final)

```
Call:
glm(formula = asthma ~ age + gender + relativehumidity + no2 +
    poly(o3, 2, raw = TRUE), family = binomial, data = envdata)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.5363371  1.6487365   0.325    0.745
age                       -0.0184370  0.0112322  -1.641    0.101
genderMale                -0.4067644  0.4504937  -0.903    0.367
relativehumidity           0.0173139  0.0127117   1.362    0.173
no2                        0.0106062  0.0481020   0.220    0.825
poly(o3, 2, raw = TRUE)1 -0.1025330  0.0826514  -1.241    0.215
poly(o3, 2, raw = TRUE)2  0.0007104  0.0016209   0.438    0.661

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 144.34  on 107  degrees of freedom
Residual deviance: 124.30  on 101  degrees of freedom
AIC: 138.3

Number of Fisher Scoring iterations: 4
```

Q3: and Q4- Decsion trees:

Code:

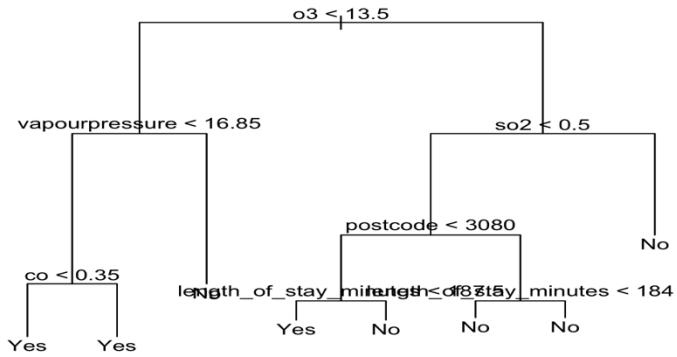
- dt1 <- tree(asthma~., data = envdata)
- plot(dt1)
- text(dt1, pretty = 0)
- summary (dt1)

```
> summary (dt1)

Classification tree:
tree(formula = asthma ~ ., data = envdata)
Variables actually used in tree construction:
[1] "o3"                  "vapourpressure"      "aqi"                 "visibility_reduction" "so2"                 "relativehumidity"
[7] "windspeed"
Number of terminal nodes:  12
Residual mean deviance:  0.4194 = 40.27 / 96
Misclassification error rate: 0.09259 = 10 / 108
>
```

#Divide the data into training and testing 50-50 split.

- set.seed(4)
- trainId = sample (1:nrow(envdata),nrow(envdata)*0.5)
- #Build DT for the training dataset
- dt2 <- tree(asthma~., data = envdata, subset = trainId)

- plot (dt2)
- text (dt2, pretty = 0)



Build a test frame and keep only complete cases (so predict() and observed match)

- test <- envdata[-trainId,]
- ok <- complete.cases(test) # rows with no NA across used columns
- test <- test[ok,]

##

#model acccuracy for test dataset

- observed1 <- test$asthma
- pred1 <- predict(dt2, test, type = "class")
- tab1 <- table(observed1, pred1)
- rate1 <- (tab1[1,2] + tab1[2,1])/sum(tab1)
- rate1

o/p: -

rate1

[1] 0.3148148

Misclassification rate for the test data set is 31.48%

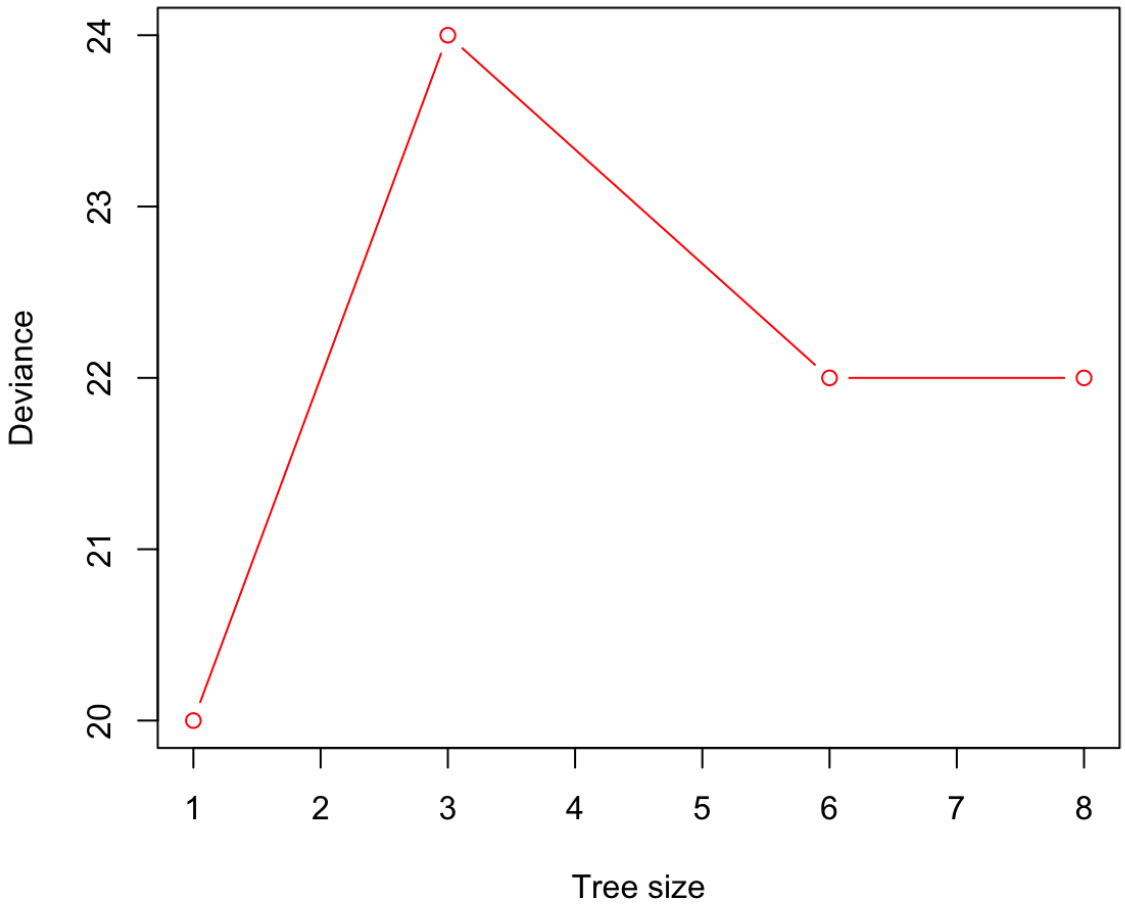
#Pruning the model below. To check if the chopping the tree is required or not.

- dt_cv1 <- cv.tree(dt2, FUN = prune.misclass)
- dt_cv1

o/p:-

```
> dt_cv1
$size
[1] 8 6 3 1

$dev
[1] 22 22 24 20

$k
[1] -Inf    0    1    5

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
>
```

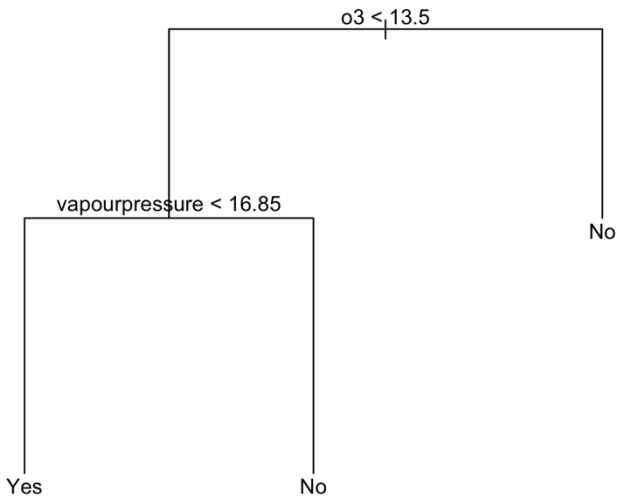
- set.seed(4)
- plot(dt_cv1$size, dt_cv1$dev, type = "b", col = "red", xlab = "Tree size", ylab = "Deviance")



- #The best tree size is 3
- pruned_dt1 <- prune.misclass(dt2, best = 3)
- plot(pruned_dt1)
- text (pruned_dt1, pretty = 0)

#Interpretation: When the o3 is less than 13.5 and vapourpressure is less than 16.85 then we get a yes profile.

If o3 is greater than 13.5, we get the no profile



- pruned_dt1

```
> pruned_dt1
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 54 70.05 No ( 0.6481 0.3519 )
  2) o3 < 13.5 19 25.01 Yes ( 0.3684 0.6316 )
    4) vapourpressure < 16.85 14 11.48 Yes ( 0.1429 0.8571 ) *
    5) vapourpressure > 16.85 5  0.00 No ( 1.0000 0.0000 ) *
  3) o3 > 13.5 35 35.03 No ( 0.8000 0.2000 ) *
>
```

Question 7 - PCA

- dim(envdata)
- #mean of the variables:
- sapply(envdata[,c(3:4,8:20)], mean)
- #Variance
- sapply(envdata[,c(3:4,8:20)], var)
- #pca
- pca1 <- prcomp(envdata[,c(3:4,8:20)], scale. = TRUE)
- pca1$rotation

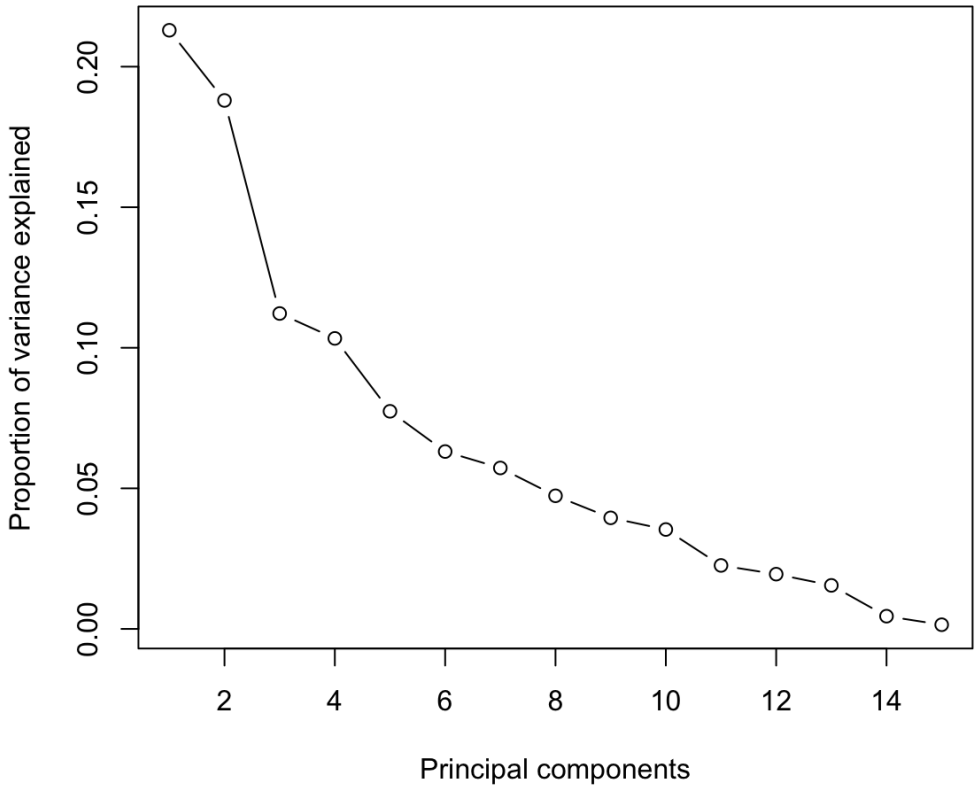
O/p : -

```
> pca1$rotation
                              PC1         PC2         PC3         PC4          PC5         PC6         PC7          PC8         PC9
length_of_stay_minutes  0.12455479 -0.15454188  0.15744316 -0.21910353  0.002774439  0.83562816 -0.18840608  0.092326989 -0.10874756
postcode               -0.00462400 -0.01355168 -0.36847678  0.21084514  0.511633884  0.07299161  0.48444857  0.315881440 -0.39790897
co                     -0.30721987 -0.25352791  0.17232200 -0.03065251  0.310376789 -0.23665296  0.01003716  0.108735528  0.35678784
o3                     -0.08264053  0.42368904 -0.34581558 -0.05895368 -0.095658400  0.12954485  0.02276153  0.112242065  0.38200223
no2                    -0.31099981 -0.30342460  0.33611379 -0.12581036  0.178235395  0.03203578 -0.05395334  0.048552888 -0.07709620
so2                    -0.27258444  0.03431855  0.23540767  0.02311625  0.056854155  0.28322223  0.63777137 -0.502022068  0.14823884
ppm10                  -0.38967835  0.30461989  0.05997612 -0.11701953  0.058961852  0.07323713 -0.15743517  0.139319445 -0.32575555
visibility_reduction   -0.47675973  0.09094985  0.03354996  0.09093539 -0.085419098 -0.03546878 -0.10151337 -0.004131747  0.16099014
aqi                    -0.43054614  0.30780367  0.02106819 -0.03759596  0.067328170  0.03562888 -0.19314437  0.158753529 -0.21362398
precipitation           0.07248265 -0.01527268 -0.01354556  0.47861884  0.490239790  0.24379834 -0.31805606  0.078560028  0.42907443
relativehumidity       -0.03206218 -0.25537184  0.22193887  0.55584356 -0.186190489 -0.09960902 -0.13211765 -0.024763175 -0.33265392
vapourpressure         -0.17625353  0.10204602 -0.10621726  0.55444512 -0.402097480  0.23415555  0.07076157 -0.035975861  0.01288679
windspeed               0.21872346  0.35295218  0.48961128  0.10142059  0.025774050 -0.04512661  0.13697826  0.239221166  0.01479621
winddirection           0.11066309  0.31734775 -0.01550846  0.05424292  0.377193425 -0.06576449 -0.29669218 -0.683888047 -0.23205887
maxwindspeed            0.22422250  0.37675667  0.46466676  0.06861561  0.053000193 -0.06268390  0.12932071  0.187752397  0.04506721
                             PC10        PC11        PC12        PC13         PC14         PC15
length_of_stay_minutes  0.31946813 -0.09739680 -0.07104825  0.12503994 -0.031556635 -0.0213131155
postcode                0.19624931  0.12494348 -0.05100639  0.02620978 -0.046181954 -0.0108654840
co                      0.48151685 -0.48106974  0.22612240  0.05192067 -0.042081418 -0.0161396100
o3                      0.22680575 -0.11896298 -0.38647138 -0.53801588  0.005611715 -0.0342633387
no2                    -0.01743378  0.47803329  0.07122219 -0.62745396  0.103993133  0.0004158906
so2                    -0.23594150 -0.15710273 -0.14145803  0.03789368  0.050655255  0.0025386372
ppm10                  -0.32153966 -0.29986142  0.15860178 -0.11228135 -0.585121137 -0.0740351666
visibility_reduction    0.24388417  0.50729760 -0.32377307  0.44864351 -0.295305917  0.0096535845
aqi                    -0.08605730 -0.12470797 -0.02998517  0.18548241  0.736101448  0.0630741771
precipitation          -0.41339964  0.03759074 -0.01181259  0.03754368 -0.011426850  0.0072583364
relativehumidity        0.08469172 -0.30151143 -0.52366878 -0.15420840 -0.010955446 -0.0775336760
vapourpressure          0.20354449  0.08229588  0.59571416 -0.11297389  0.030969835  0.0071067907
windspeed               0.10060654  0.02347807 -0.01672000 -0.02644946 -0.077723119  0.6935251257
winddirection           0.33303456  0.05249712  0.06762883 -0.08811441 -0.023841726  0.0332686480
maxwindspeed            0.09667245  0.10442133  0.04238292  0.04096834  0.049515996 -0.7072731425
>
```

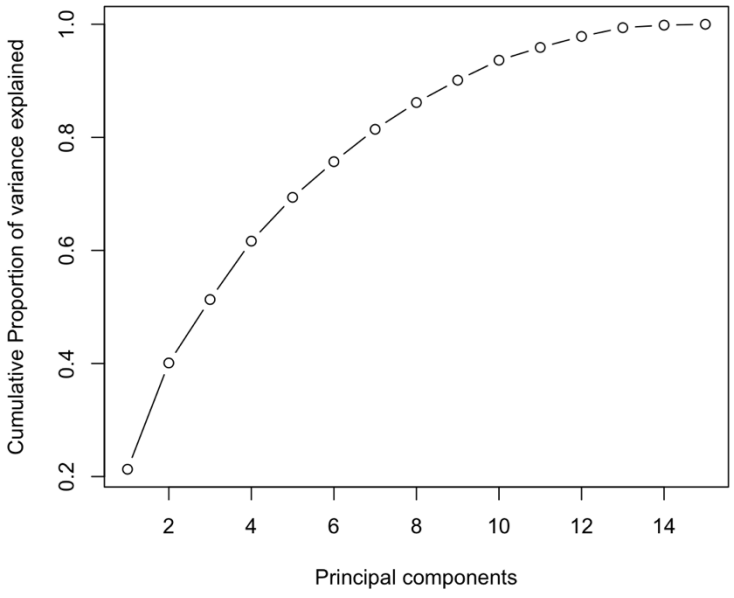
- # To check the most affecting and relevant loading vector from the output of above command, check the values with the highest number in the PC
- #column
- # Note only check the absolute values while checking the number. For example: in PC1 - visibility_reduction has the highest value i.e. 0.47
- pr.var <- pca1$sdev^2
- pr.var
- pve <- pr.var/sum(pr.var)
- pve
- cumsum(pve)
- par(mfrow = c(1,1))
- plot(x = c(1:15), y = pve,

  xlab = "Principal components",

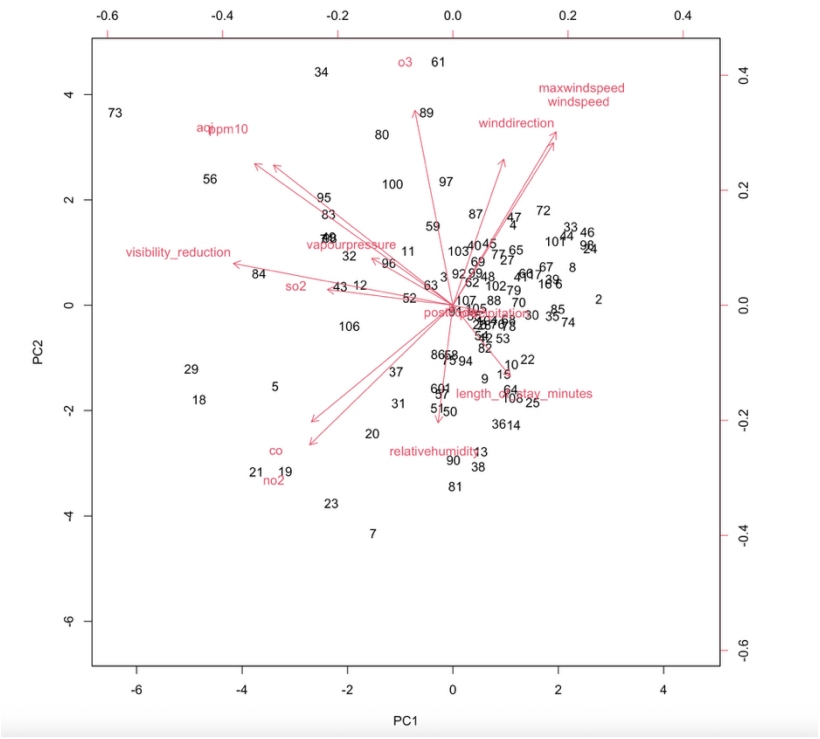
  ylab = "Proportion of variance explained", type = "b")



- plot(x = c(1:15), y = cumsum(pve),
  xlab = "Principal components",
  ylab = "Cumulative Proportion of variance explained", type = "b")



- biplot(pca1, scale = 0)



----------######### End of the document #########----------