

CSA1668-DATA WAREHOUSING AND DATA MINING FOR PATTERN ANALYSIS

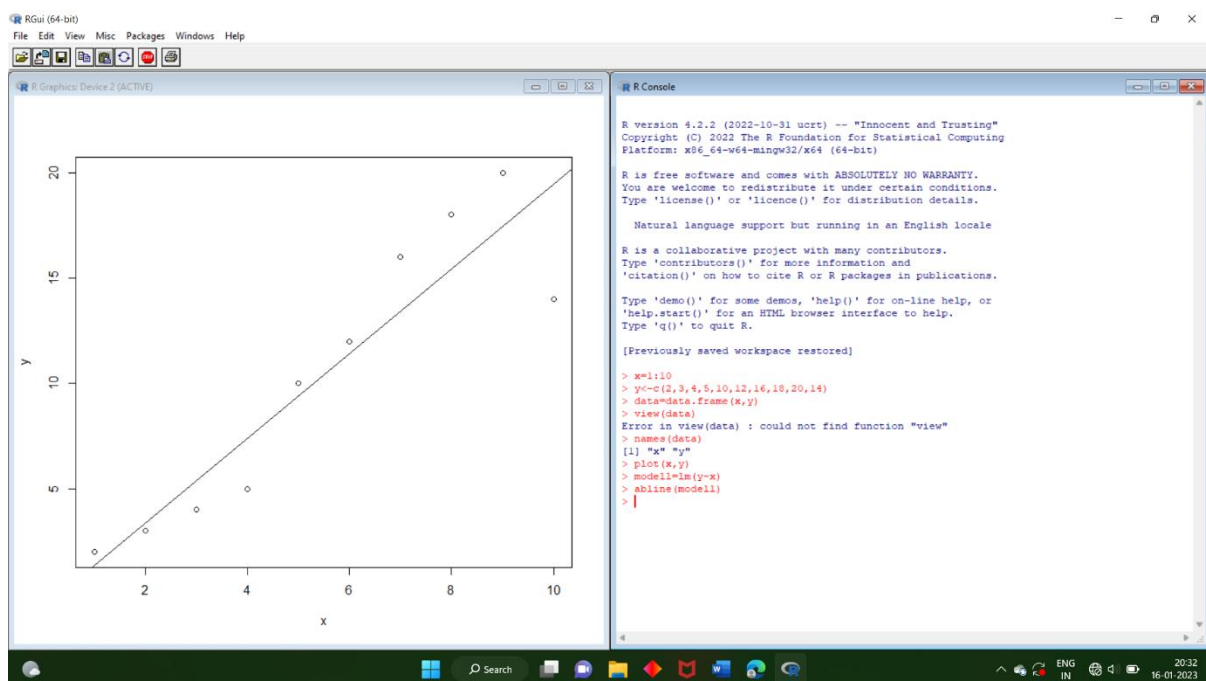
NAME:B.VINEETHA

REG.NO:192110487

ETL AND OLAP OPERATION USING KNIME DATA ANALYTICS PLATFORM.

PREDICTION ANALYSIS USING LINEAR REGRESSION THROUGH R TOOL.

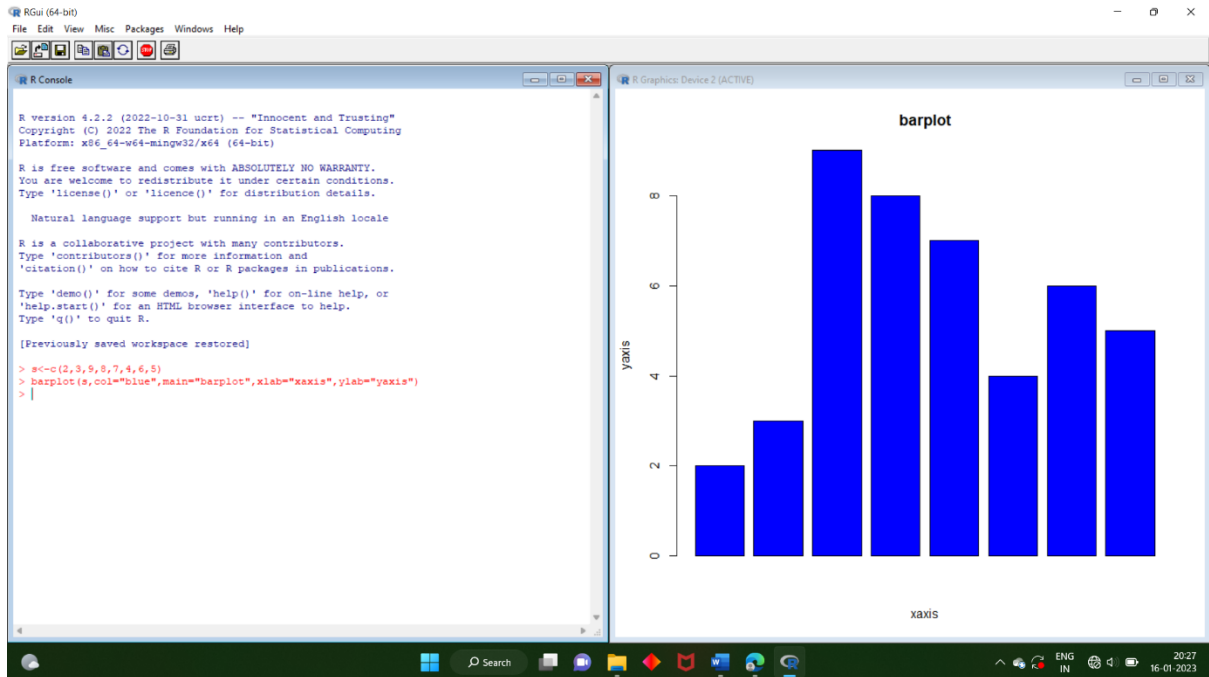
Output:



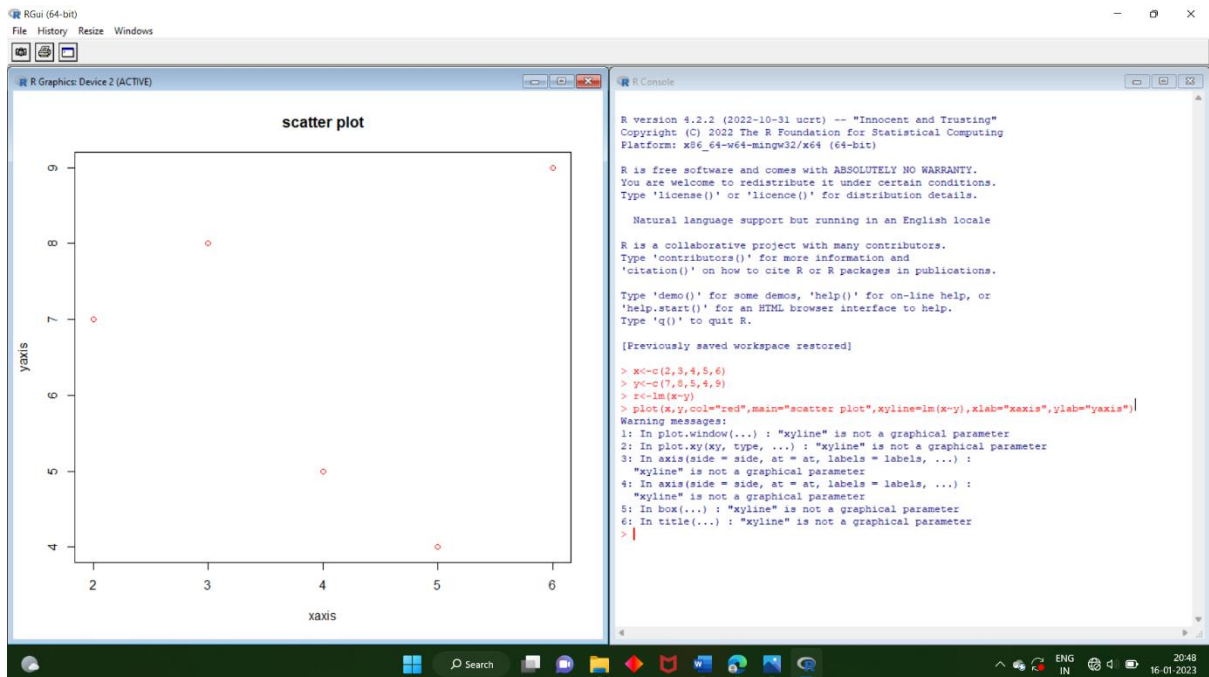
PLOTTING GRAPHS USING R TOOL

Output:

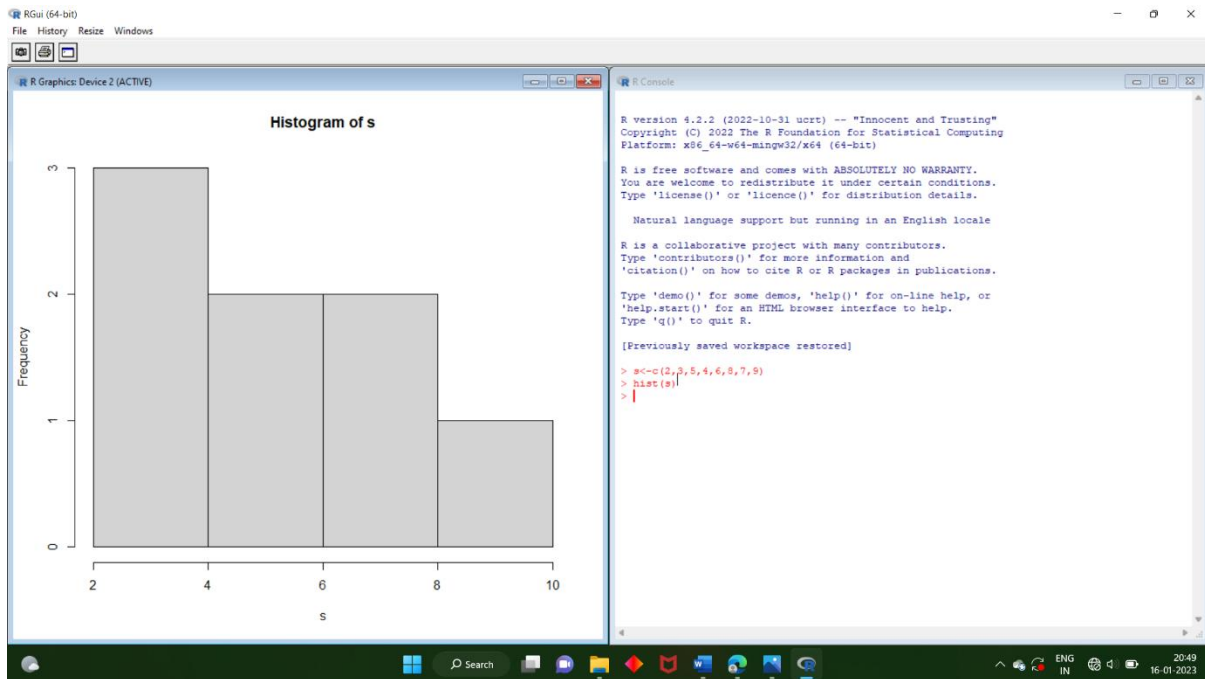
Barplot:



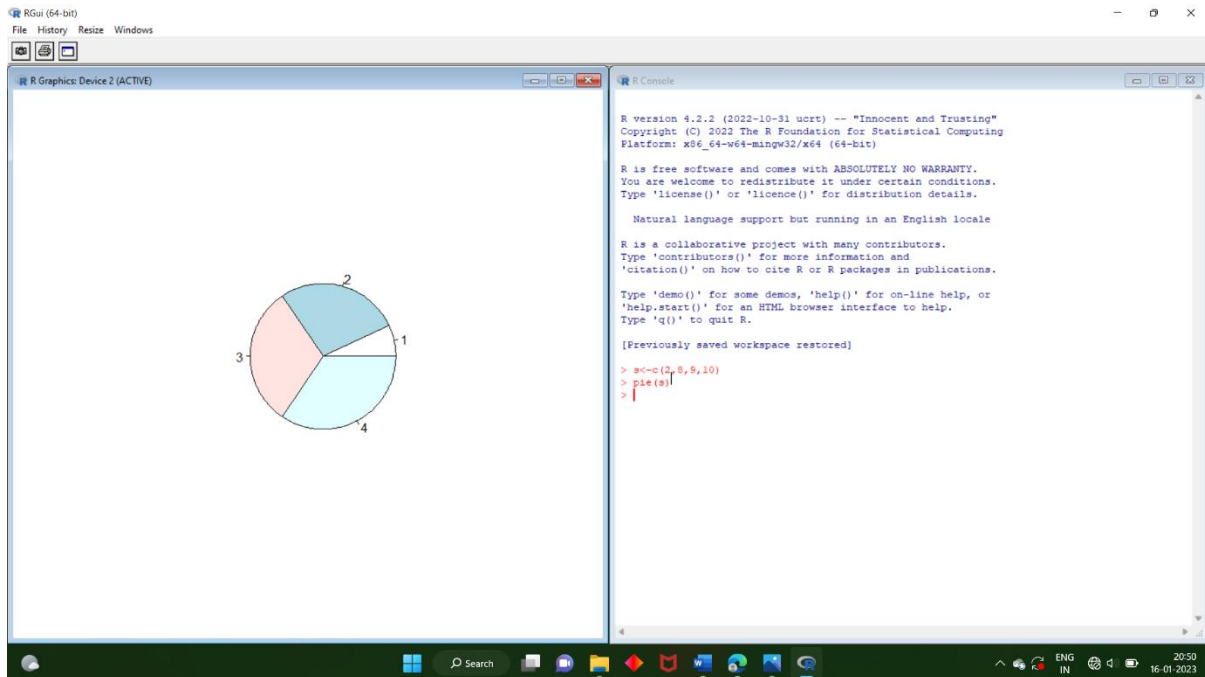
Scatterplot:



Histogram:



Piechart:



CENTRAL TENDENCY AND DATA DISPERSION MEASURES USING R-TOOL.

Output:

Mean and Median :

```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> s<-c(7,6,8,4,5,5,7,3,6,7)
> result.m<-mean(s)
> print(result.m)
[1] 6.2
>
> s<-c(6,47,49,15,43,41,7,39,43,41,36)
> result.m<-mean(s)
> print(result.m)
[1] 33.36364
> s<-c(7,6,8,4,5,5,7,3,6,7)
> result.median(s)
> print(result)
[1] 6.5
> s<-c(6,47,49,15,43,41,7,39,43,41,36)
> result.median(s)
> print(result)
[1] 41
>
```

Mode:

```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> my_mode<-function(x){
+   unique_x<-unique(x)
+   tabulate_x<-tabulate(match(x,unique_x))
+   unique_x[tabulate_x==max(tabulate_x)]
+ }
> s<-c(14,21,18,21,14,35)
> my_mode(s)
[1] 14 21
> my_mode<-function(x){
+   unique_x<-unique(x)
+   tabulate_x<-tabulate(match(x,unique_x))
+   unique_x[tabulate_x==max(tabulate_x)]
+ }
> s<-c(13,15,16,16,19,20,20,20,21,22,22,22,25,25,25,30,33,40,45,46,52,52,70)
> my_mode(s)
[1] 20 22 25
>
```

IQR,Range,Fivenumber summary,Boxplot :

```
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

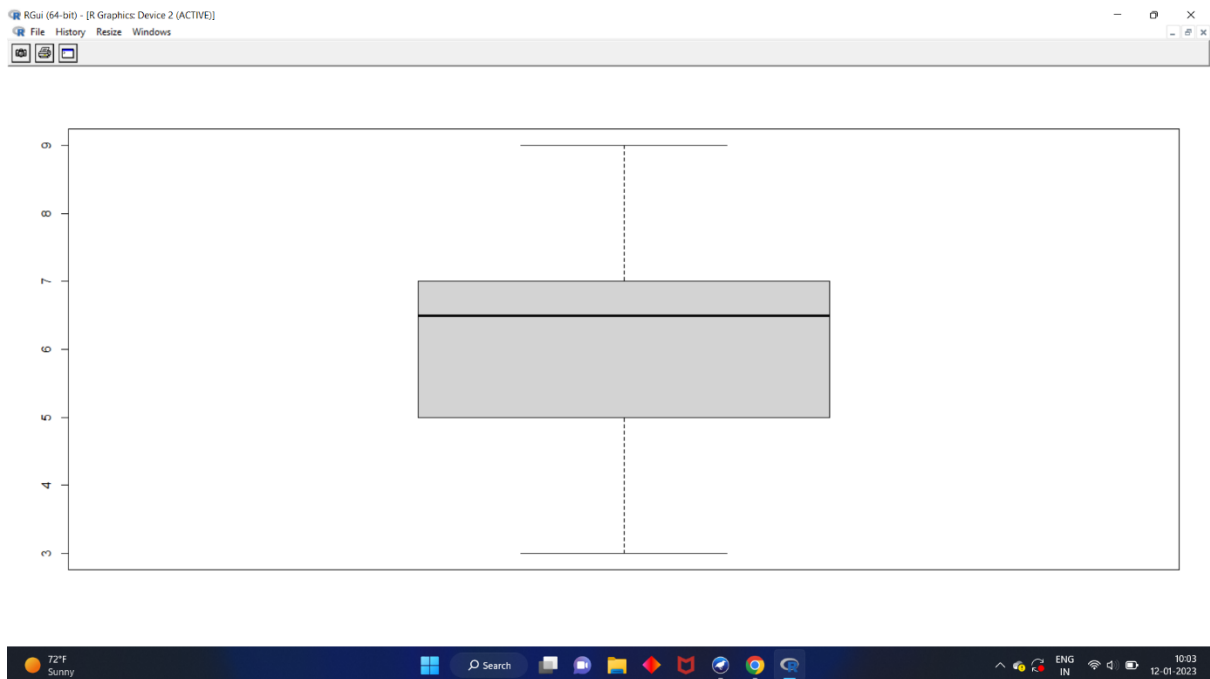
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

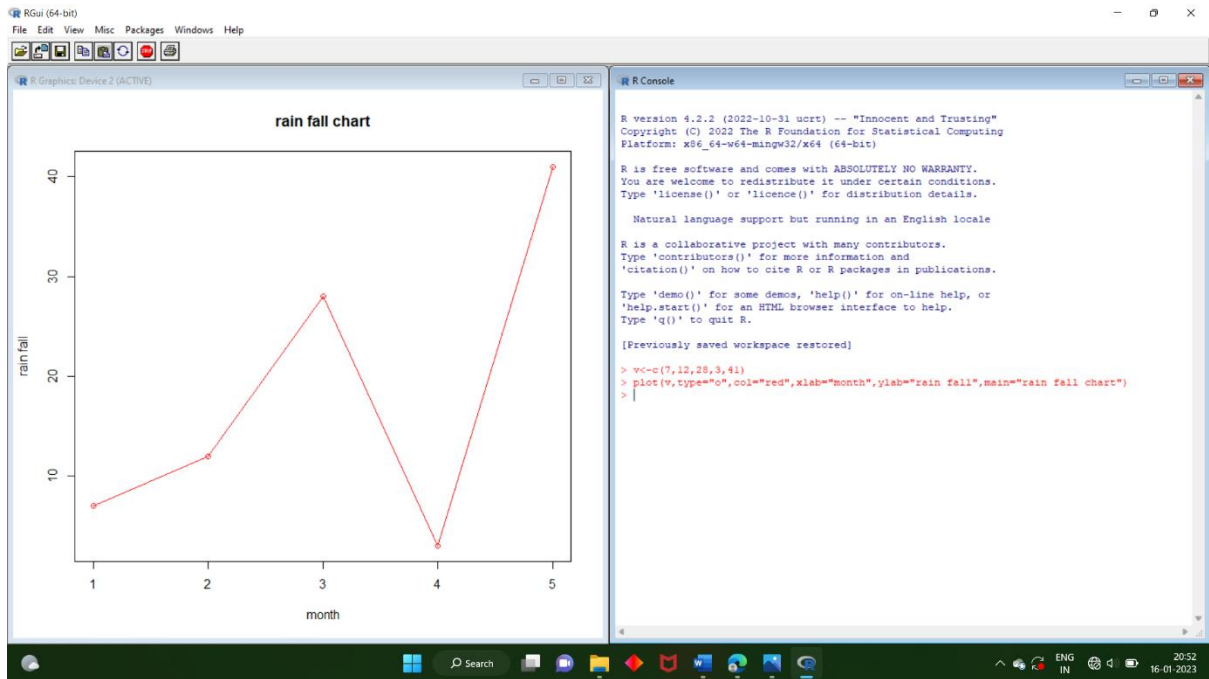
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

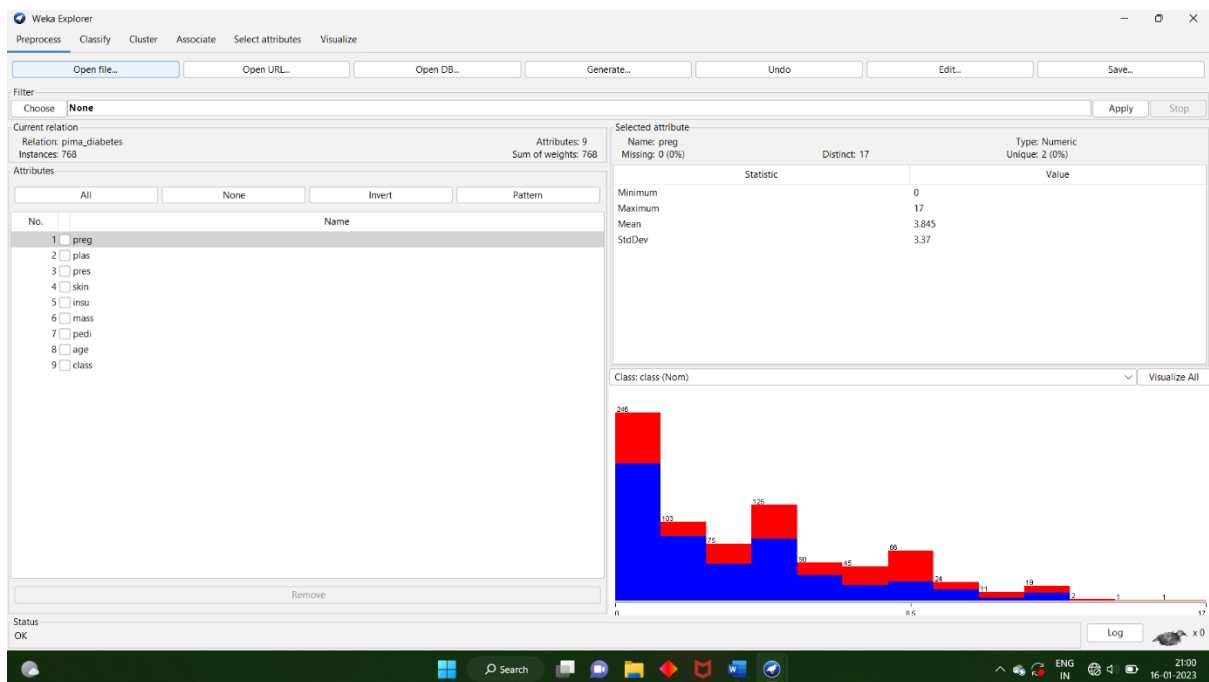
> s<-c(7,6,8,4,5,9,7,3,6,7)
> result=IQR(s)
> print(result)
[1] 1.75
> s<-c(7,6,8,4,5,9,7,3,6,7)
> result=range(s)
> print(result)
[1] 3 9
> s<-c(7,6,8,4,5,9,7,3,6,7)
> result=fivenum(s)
> print(result)
[1] 3.0 5.0 6.5 7.0 9.0
> s<-c(7,6,8,4,5,9,7,3,6,7)
> result=boxplot(s)
> |
```

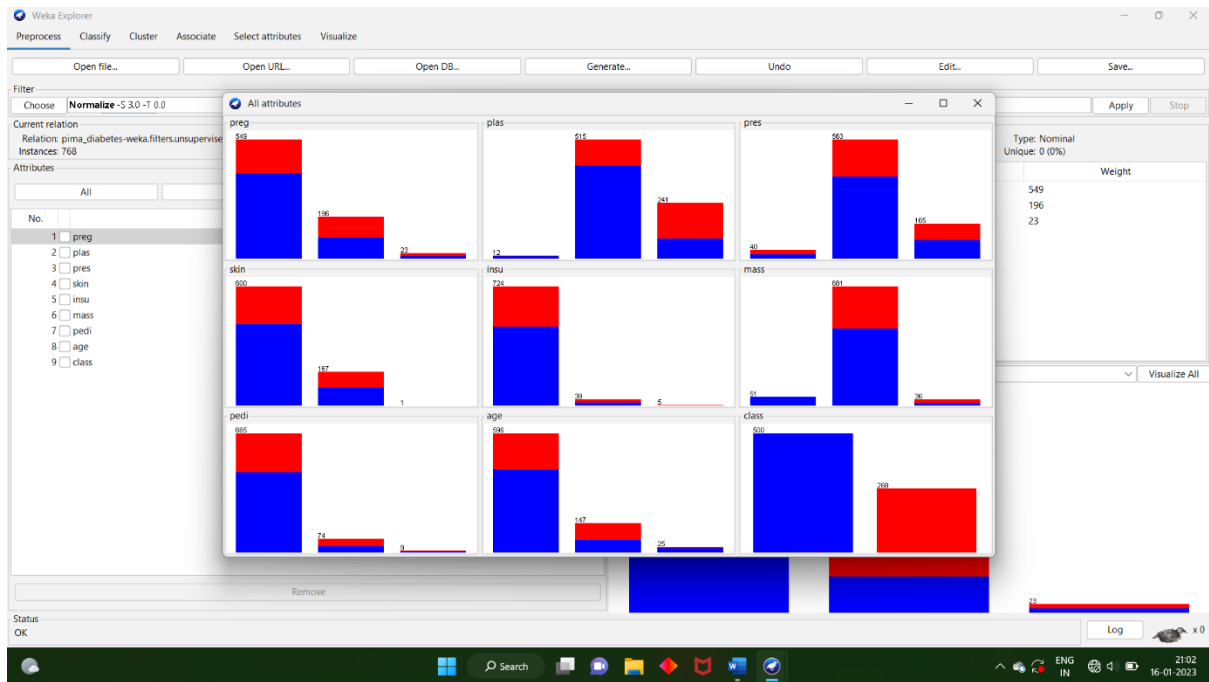


REGRESSION ANALYSIS USING R TOOL.

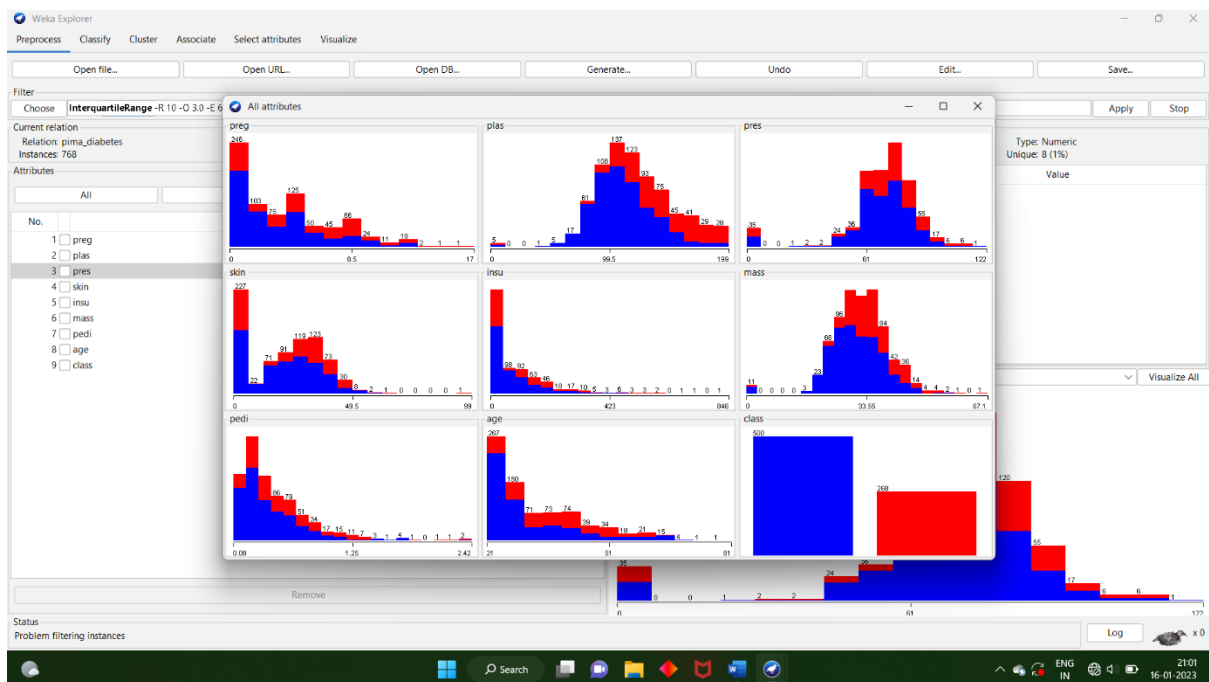


PERFORM CORRECTION ANALYSIS AND NORMALIZATION.





DATA PREPROCESSING AND PREPARATION FOR KNOWLEDGE ANALYSIS USING WEKA.



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **J48 -C 0.25 -M 2**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) class

Start Stop

Result list (right-click for options)

21:03:17 - trees.J48

21:03:51 - trees.J48

Classifier output

```

I mass = '(-inf-22.366667)': tested_negative (5.0)
I mass = '(22.366667-44.733333)': tested_positive (218.0/77.0)
I mass = '(44.733333-inf)': tested_positive (18.0/5.0)

Number of Leaves : 17
Size of the tree : 25

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      561      73.0469 %
Incorrectly Classified Instances    207      26.9531 %
Kappa statistic                    0.3926
Mean absolute error                 0.3667
Root mean squared error             0.4418
Relative absolute error             80.6754 %
Root relative squared error         92.6901 %
Total Number of Instances          768

=== Detailed Accuracy By Class ===

      TP Rate  PP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
-----
0.820  0.437  0.770  0.820  0.798  0.394  0.708  0.781  tested_negative
0.563  0.180  0.627  0.563  0.593  0.394  0.708  0.542  tested_positive
Weighted Avg.  0.730  0.347  0.725  0.730  0.727  0.394  0.708  0.698

=== Confusion Matrix ===

  a  b  <-- classified as
410 90 | a = tested_negative
117 151 | b = tested_positive

```

Status OK

Log

K-MEANS CLUSTER ANALYSIS USING WEKA.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Cluster

Choose **SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10**

Cluster mode

☒ Use training set

☐ Supplied test set

☐ Percentage split % **66**

☐ Classes to clusters evaluation

(Nom) class

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

21:11:46 - SimpleKMeans

Cluster output

```

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance"
Relation: pima_diabetes-weka.filters.unsupervised.attribute.Discretize-83-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.Normalize-83.0-T0
Instances: 768
Attributes: 9
preg
plas
pres
skin
insu
mass
pedi
age
class

Test mode: evaluate on training data

=== Clustering model (full training set) ===

KMeans
=====

Number of iterations: 4
Within cluster sum of squared errors: 1204.0

Initial starting points (random):

Cluster 0: '(-inf-5.666667)\',\','(66.333333-132.666667)\',\','(40.666667-81.333333)\',\','(-inf-33)\',\','(-inf-282)\',\','(22.366667-44.733333)\',\','(-inf-
Cluster 1: '\'(5.666667-11.333333)\',\','(66.333333-132.666667)\',\','(40.666667-81.333333)\',\','(-inf-33)\',\','(-inf-282)\',\','(22.366667-44.733333)\',\','(-

Missing values globally replaced with mean/mode

Final cluster centroids:

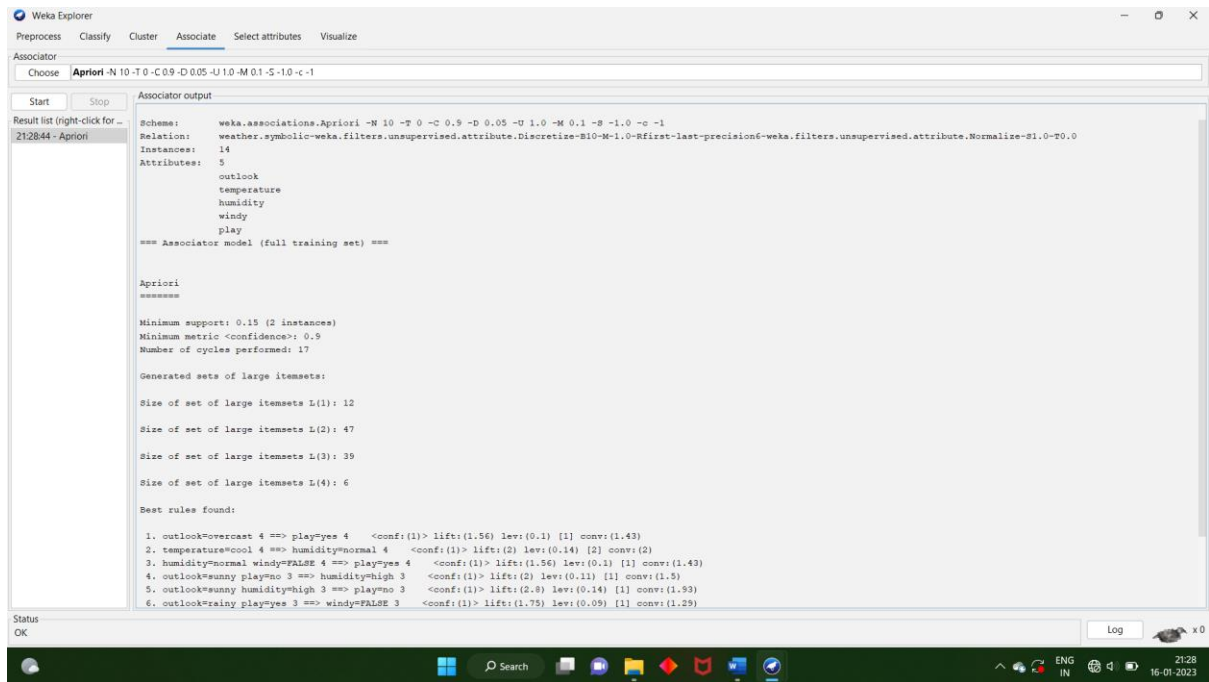
Attribute      Full Data      Cluster#
0              0              1

```

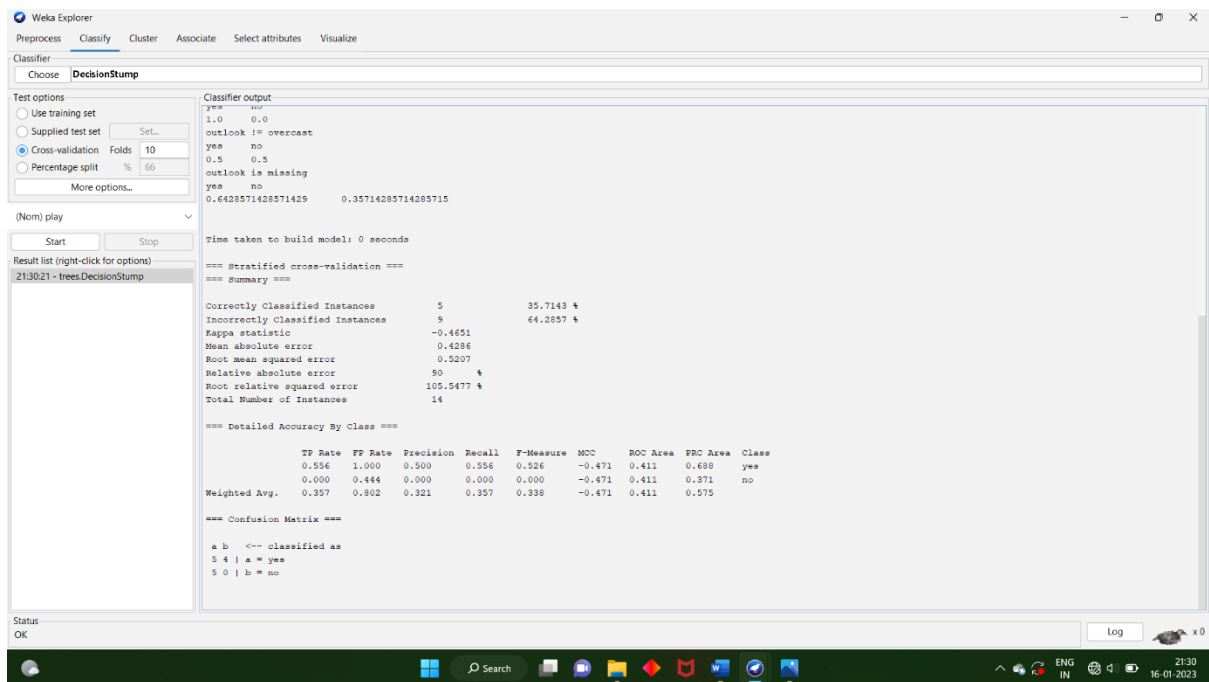
Status OK

Log

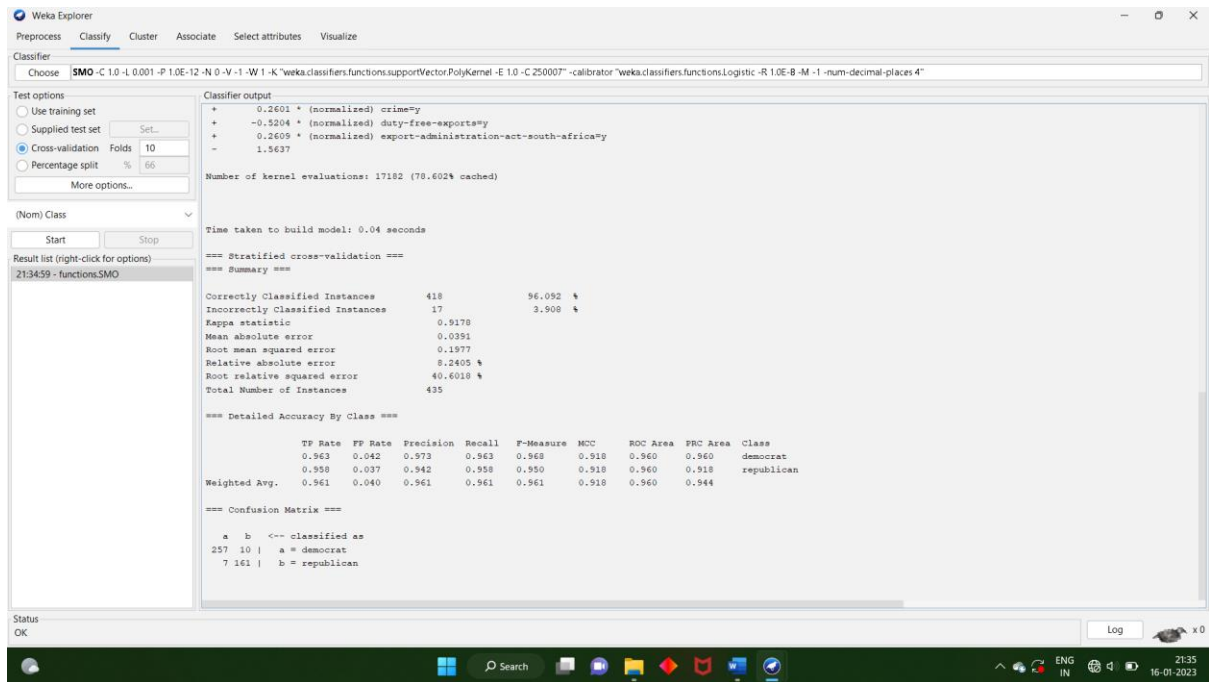




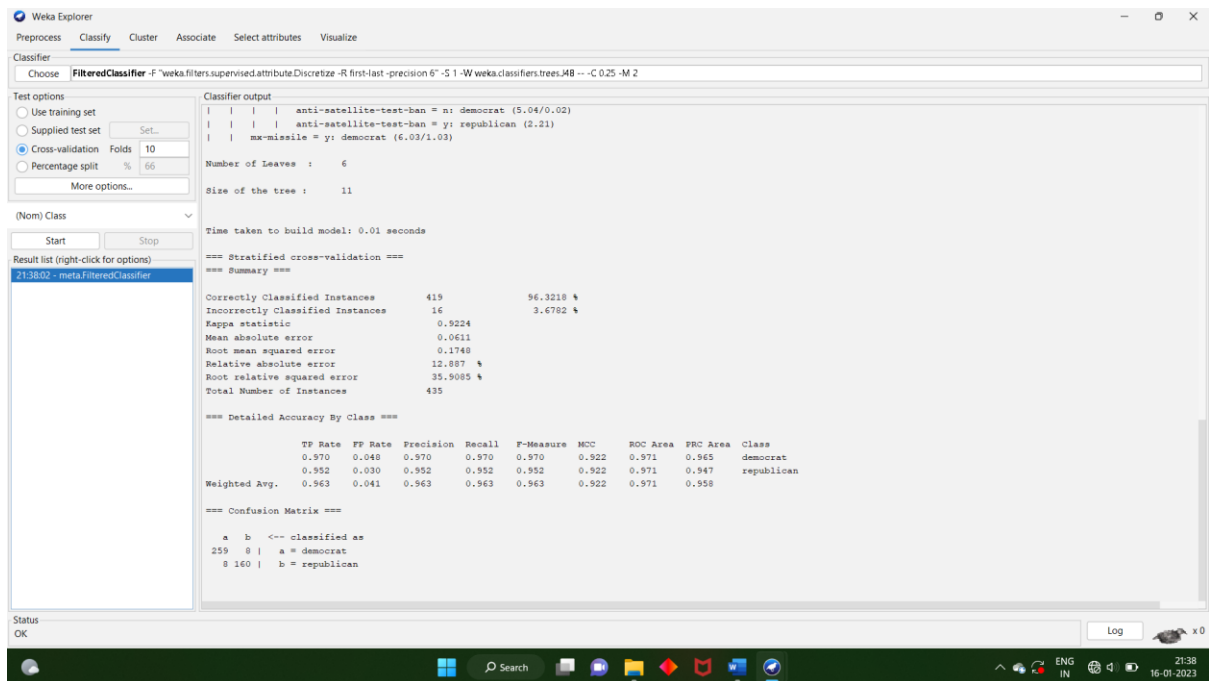
PREDICTION OF CATEGORICAL DATA USING DECISION TREE ALGORITHM USING WEKA.



PREDICTION OF CATEGORICAL DATA USING SMO ALGORITHM USING WEKA.



EVALUATING THE ACCURACY OF THE CLASSIFIERS USING WEKA



PREDICTION OF CATEGORICAL DATA USING BAYESIAN ALGORITHM USING WEKA.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options:
☐ Use training set
☐ Supplied test set
☒ Cross-validation Folds: **10**
☐ Percentage split % **66**
 More options...

(Nom) Class: **democrat**
 Start Stop

Result list (right-click for options):
 21:38:02 - meta.FilteredClassifier
21:40:50 - bayes.NaiveBayes

Classifier output:

```

y
[total]                253.0    159.0

export-administration-act-south-africa
n                      13.0     51.0
y                      174.0    57.0
[total]                187.0    148.0
  
```

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	392	90.1145 %
Incorrectly Classified Instances	43	9.8851 %
Kappa statistic	0.7949	
Mean absolute error	0.0995	
Root mean squared error	0.2977	
Relative absolute error	20.9815 %	
Root relative squared error	61.1406 %	
Total Number of Instances	435	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MDC	ROC Area	PRC Area	Class
Weighted Avg.	0.901	0.109	0.842	0.891	0.877	0.797	0.973	0.557	democrat
	0.901	0.053	0.905	0.501	0.502	0.797	0.973	0.973	republican

=== Confusion Matrix ===

	a	b	<-- classified as
238	29		a = democrat
14	154		b = republican

Status: OK

DATA ANALYSIS BY DENSITY BASED CLUSTERING ALGORITHM USING WEKA.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer: Choose **MakeDensityBasedClusterer** -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last -I 500 -num-slots 1 -S 10

Cluster mode:
☒ Use training set
☐ Supplied test set
☐ Percentage split % **66**
☐ Classes to clusters evaluation
 (Nom) Class: **democrat**
☒ Store clusters for visualization
 Ignore attributes
 Start Stop

Result list (right-click for options):
 21:42:20 - MakeDensityBasedClusterer

Clusterer output:

```

=== Run information ===

Scheme:      weka.clusterers.MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last -I 500 -num-slots 1 -S 10
Relation:    votes
Instances:   435
Attributes:  17
handicapped-infants
water-project-coast-sharing
adoption-of-the-budget-resolution
physician-fee-freeze
el-salvador-aid
religious-groups-in-schools
anti-satellite-test-ban
aid-to-nicaraguan-contras
mx-missile
immigration
synfuels-corporation-cutback
education-spending
superfund-right-to-sue
crime
duty-free-exports
export-administration-act-south-africa
Class
Test mode:   evaluate on training data

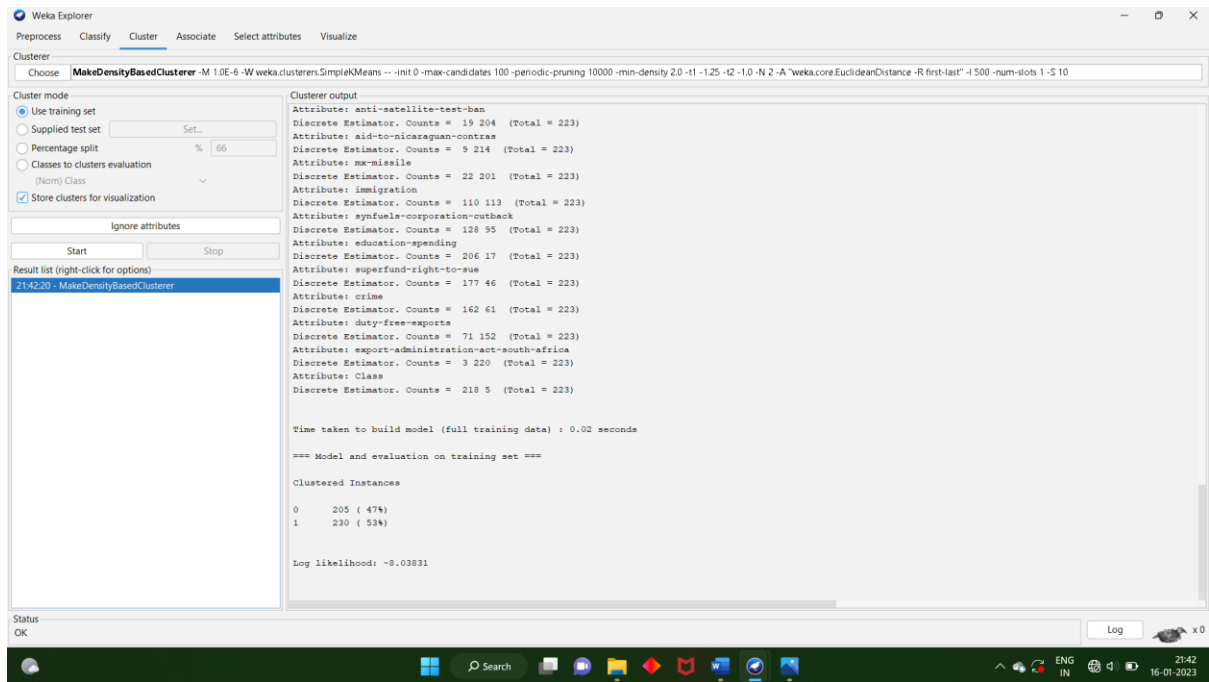
=== Clustering model (full training set) ===

MakeDensityBasedClusterer:

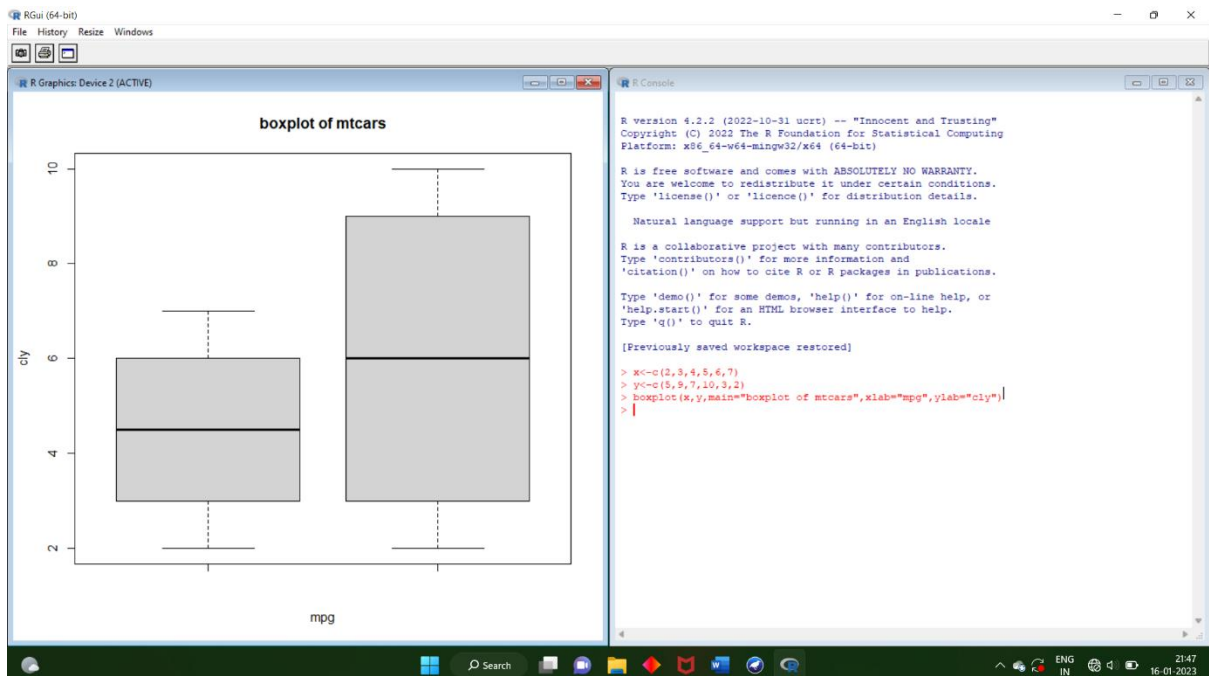
Wrapped clusterer:
kMeans
=====

Number of iterations: 3
Within cluster sum of squared errors: 1510.0
  
```

Status: OK



CREATE A BOXPLOT GRAPH FOR THE RELATION BETWEEN "MPG"(MILES PER GALLOON) AND "CYL"(NUMBER OF CYLINDERS) FOR THE DATASET "MTCARS" AVAILABLE IN R ENVIRONMENT



**USING R PROGRAM MAKE A HISTOGRAM FOR THE “AIRPASSENGERS
“DATASET, START AT 100 ON THE X-AXIS, AND FROM VALUES 200 TO 700,
MAKE THE BINS 150 WIDE.**

