KARNAVATI UNIVERSITY

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted BY

Panchal Rushin

(Enrl. No.: 20220701044)

Course Code and Title: 21BSCS23C02 – R Programming

B.Sc. (Hons.) Computer Science / Data Science / AIML
III Semester – July – Nov 2023

USCI

Nov/Dec 2023

Index :

- Introduction
- Aim of the Project
- Intended outcomes of the Project
- Description of the Dataset
- Dataset
- Functions
  - Description of data function
  - Head function
  - Summary function
  - Tail function
  - Name function
  - Structure function
  - Missing values function
  - Unique function
  - Co relation function
  - Dimension function
  - Unique value function

- Statistical Analysis
  - Shapiro.test function
  - Mean function
  - Standard Deviation function
  - Quantile function
  - Variance function
  - Minimum & Maximum function
  - Median function
  - Co relation function between two variables

- Data Visualization
  - BarPlot
  - Histogram
  - Barchart
  - Scatter plot
  - Pie chart

- Linear regression

- ANOVA Function

The Joyner–Boore Attenuation Data

Introduction:

In the realm of seismic hazard assessment, the Joyner–Boore Attenuation Data holds paramount significance. This dataset serves as a cornerstone for understanding the attenuation of ground motion during earthquakes, providing valuable insights into the characteristics of seismic waves as they traverse through various geological structures. In this project, we delve into the exploration and analysis of the Joyner–Boore Attenuation Data using the R programming language, aiming to unravel hidden patterns, trends, and essential information embedded within the dataset.

Aim of the Project:

The primary objective of this project is to employ advanced data analysis techniques using R programming to discern the intricacies of the Joyner–Boore Attenuation Data. By leveraging statistical methods, visualizations, and machine learning algorithms, our goal is to gain a comprehensive understanding of the factors influencing ground motion attenuation during earthquakes. This analysis can contribute to the broader understanding of seismic hazards, aiding in the development of more resilient structures in earthquake-prone regions.

Intended Outcomes of the Project:

Utilizing the capabilities of R programming to conduct thorough exploratory data analysis (EDA) on the Joyner–Boore Attenuation Data. This involves summarizing key statistics, visualizing data distributions, and identifying potential outliers or patterns within the dataset.

Statistical and machine learning techniques in R to develop predictive models that capture the relationship between earthquake magnitude, distance from the epicenter, and ground motion amplitude.

Description of the Dataset:

The Joyner–Boore Attenuation dataset, in R programming, consists of 182 observations organized into five key columns. Each row in the dataset represents a distinct seismic event, providing valuable information for studying the attenuation of seismic waves. The dataset's columns are defined as follows:

Event Number*:* The event number serves as a unique identifier for each seismic event in the dataset. This column facilitates the tracking and referencing of individual occurrences.

Magnitude: The magnitude column contains numerical values representing the moment magnitude of the seismic events. Earthquake magnitude is a crucial parameter, indicating the energy released during an earthquake. This column helps assess the influence of earthquake strength on the attenuation of seismic waves.

Distance: The distance column contains numerical values representing the distance from the seismic event's hypocenter to the station. Distance is a key factor influencing the attenuation of seismic waves; as waves travel through the Earth's crust, their amplitudes often decrease with increasing distance from the source.

Station: The station column contains categorical data identifying the recording station associated with each observation. Stations play a critical role in capturing seismic data, and variations in station characteristics may impact recorded ground motion amplitudes.

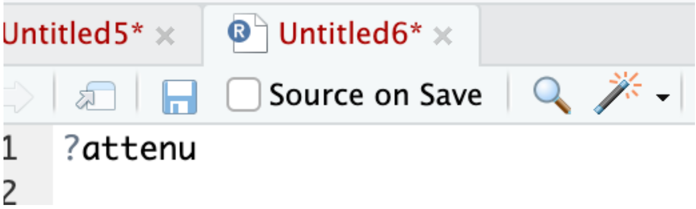
Acceleration: The acceleration column contains numerical values representing the recorded ground motion amplitude associated with each seismic event-station pair. Ground motion amplitude is a fundamental measure of the intensity of seismic waves experienced at a given station. This column serves as the response variable in the analysis, reflecting the impact of seismic attenuation under varying earthquake magnitudes and distances.

Dataset: https://1drv.ms/x/s!AhN-beO5cKv7ggfj4aKc3j8X-1Jp?e=nvutU5

Functions, Statistical Analysis and Data Visualization of Attenu Dataset -

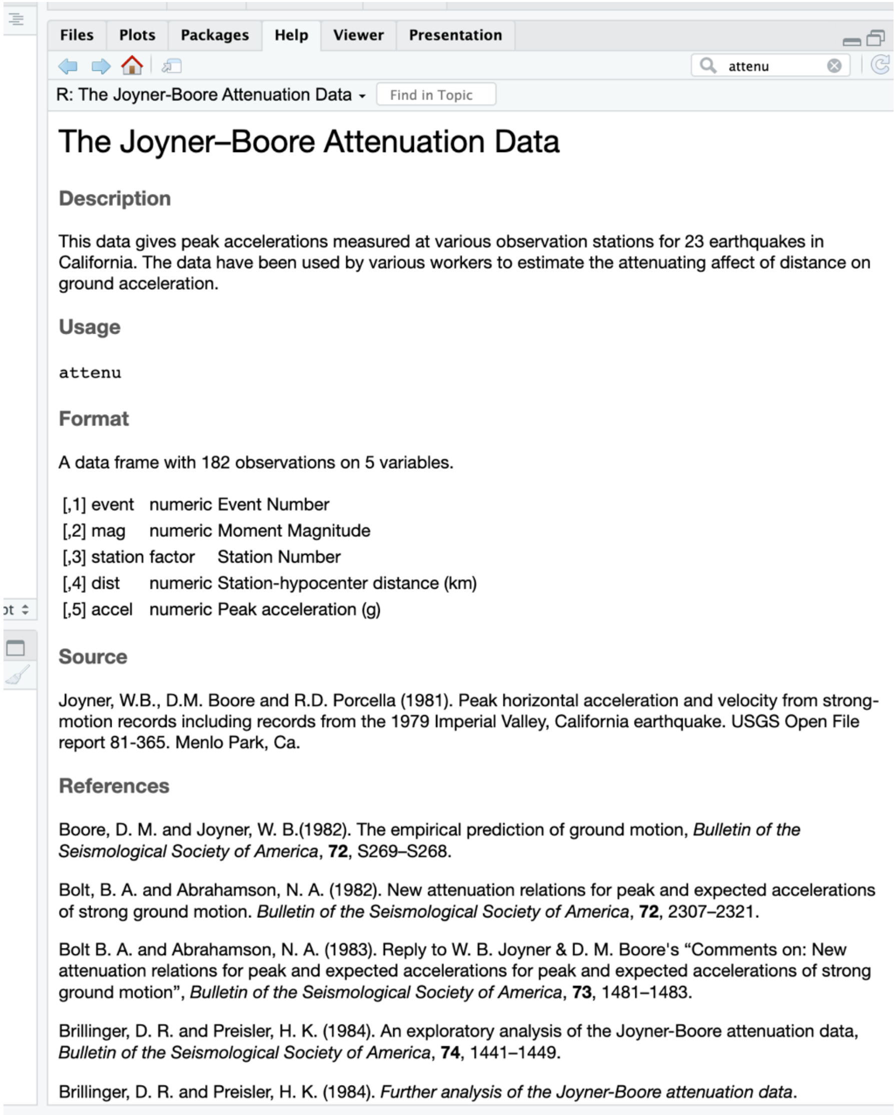
1. **?attenu**: It is used to access the documentation or help page of the dataset.

Input -



```
?attenu
```

Output –



The Joyner–Boore Attenuation Data

Description

This data gives peak accelerations measured at various observation stations for 23 earthquakes in California. The data have been used by various workers to estimate the attenuating affect of distance on ground acceleration.

Usage

```
attenu
```

Format

A data frame with 182 observations on 5 variables.

[,1]	event	numeric Event Number
[,2]	mag	numeric Moment Magnitude
[,3]	station	factor Station Number
[,4]	dist	numeric Station-hypocenter distance (km)
[,5]	accel	numeric Peak acceleration (g)

Source

Joyner, W.B., D.M. Boore and R.D. Porcella (1981). Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California earthquake. USGS Open File report 81-365. Menlo Park, Ca.

References

Boore, D. M. and Joyner, W. B.(1982). The empirical prediction of ground motion, *Bulletin of the Seismological Society of America*, **72**, S269–S268.

Bolt, B. A. and Abrahamson, N. A. (1982). New attenuation relations for peak and expected accelerations of strong ground motion. *Bulletin of the Seismological Society of America*, **72**, 2307–2321.

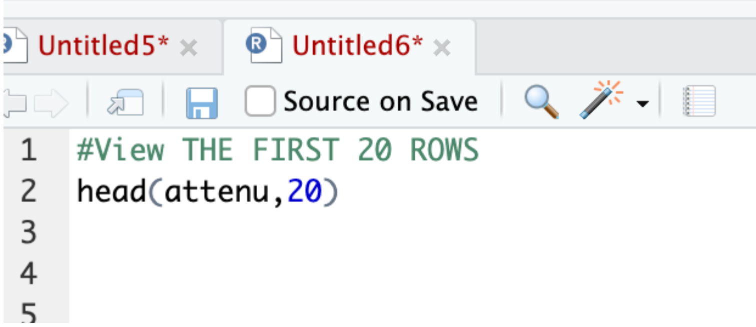
Bolt B. A. and Abrahamson, N. A. (1983). Reply to W. B. Joyner & D. M. Boore's “Comments on: New attenuation relations for peak and expected accelerations for peak and expected accelerations of strong ground motion”, *Bulletin of the Seismological Society of America*, **73**, 1481–1483.

Brillinger, D. R. and Preisler, H. K. (1984). An exploratory analysis of the Joyner-Boore attenuation data, *Bulletin of the Seismological Society of America*, **74**, 1441–1449.

Brillinger, D. R. and Preisler, H. K. (1984). *Further analysis of the Joyner-Boore attenuation data*.

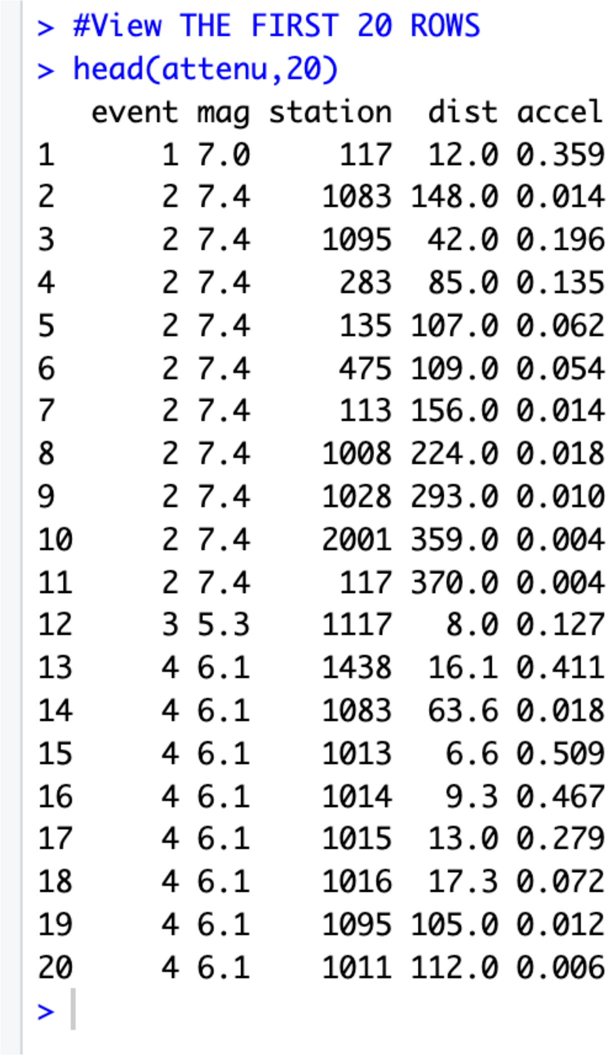
2. **Head**: It is used for viewing the first few rows.

Input –



```
#View THE FIRST 20 ROWS
head(attenu,20)
```

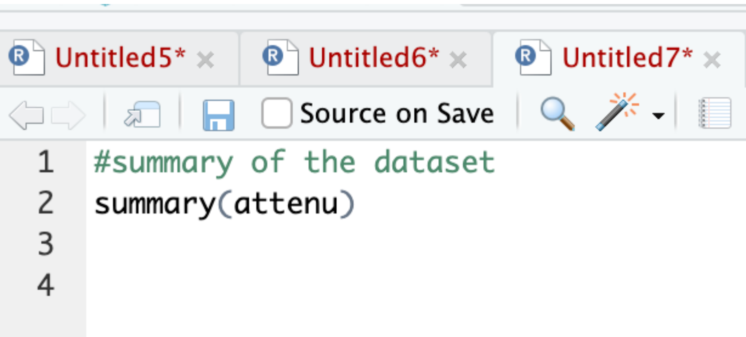
Output –



```
> #View THE FIRST 20 ROWS
> head(attenu,20)
   event mag station  dist accel
1      1 7.0     117  12.0 0.359
2      2 7.4    1083 148.0 0.014
3      2 7.4    1095  42.0 0.196
4      2 7.4     283  85.0 0.135
5      2 7.4     135 107.0 0.062
6      2 7.4     475 109.0 0.054
7      2 7.4     113 156.0 0.014
8      2 7.4    1008 224.0 0.018
9      2 7.4    1028 293.0 0.010
10     2 7.4    2001 359.0 0.004
11     2 7.4     117 370.0 0.004
12     3 5.3    1117   8.0 0.127
13     4 6.1    1438  16.1 0.411
14     4 6.1    1083  63.6 0.018
15     4 6.1    1013   6.6 0.509
16     4 6.1    1014   9.3 0.467
17     4 6.1    1015  13.0 0.279
18     4 6.1    1016  17.3 0.072
19     4 6.1    1095 105.0 0.012
20     4 6.1    1011 112.0 0.006
> 
```

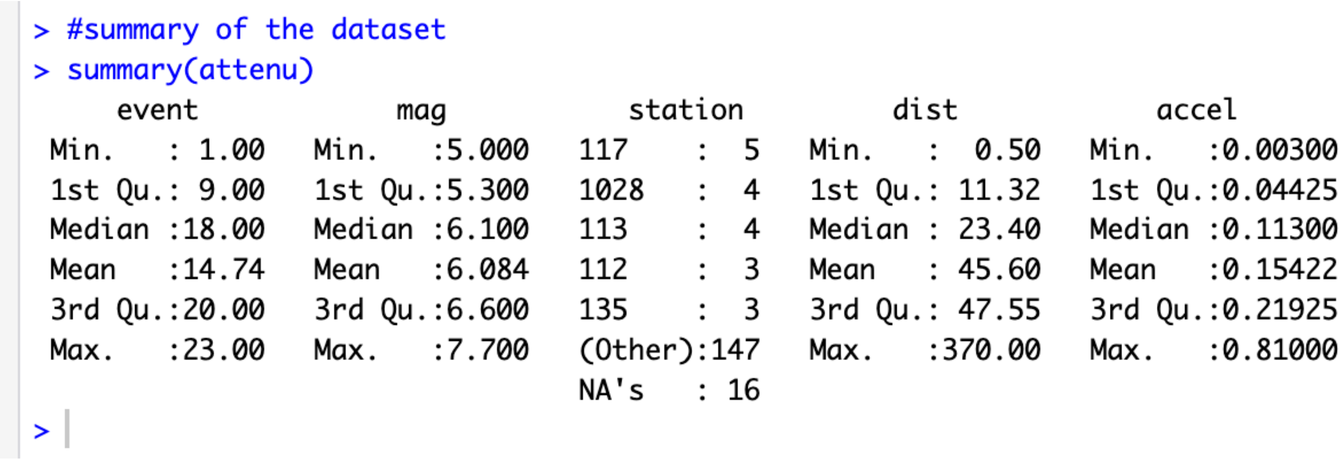
3. **Summary**: It is used to access the summary of the dataset.

Input –



```
#summary of the dataset
summary(attenu)
```

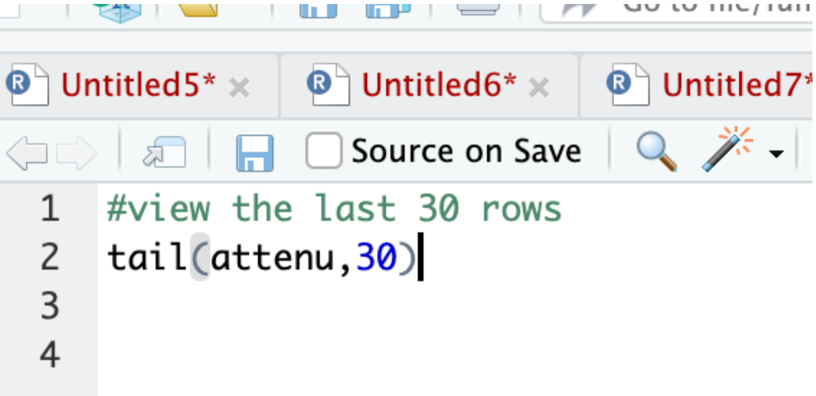
Output –



```
> #summary of the dataset
> summary(attenu)
     event            mag           station          dist            accel        
 Min.   : 1.00   Min.   :5.000   117    :  5   Min.   :  0.50   Min.   :0.00300  
 1st Qu.: 9.00   1st Qu.:5.300   1028   :  4   1st Qu.: 11.32   1st Qu.:0.04425  
 Median :18.00   Median :6.100   113    :  4   Median : 23.40   Median :0.11300  
 Mean   :14.74   Mean   :6.084   112    :  3   Mean   : 45.60   Mean   :0.15422  
 3rd Qu.:20.00   3rd Qu.:6.600   135    :  3   3rd Qu.: 47.55   3rd Qu.:0.21925  
 Max.   :23.00   Max.   :7.700   (Other):147   Max.   :370.00   Max.   :0.81000  
                                 NA's   : 16                                    
> 
```

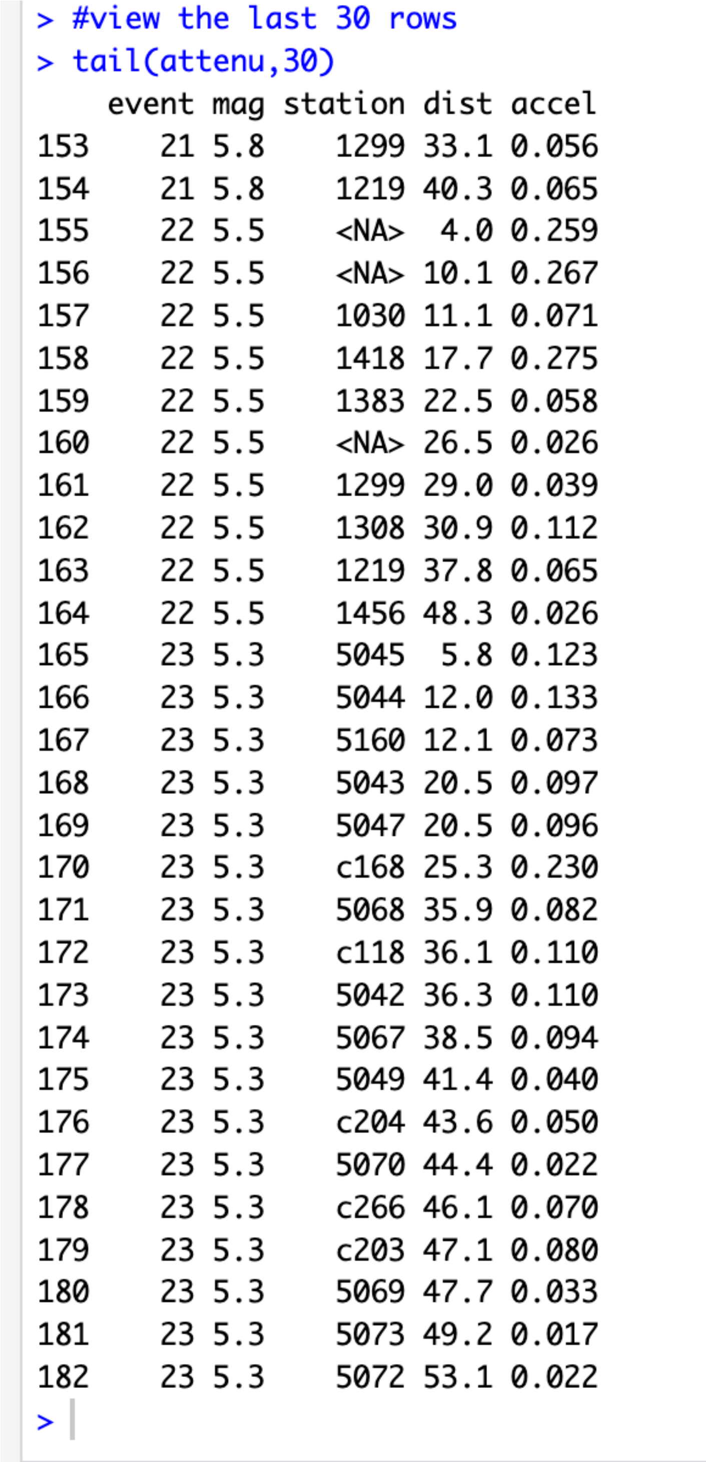
4. **Tail** : It is used for viewing the last few rows.

Input –



```
#view the last 30 rows
tail(attenu,30)
```

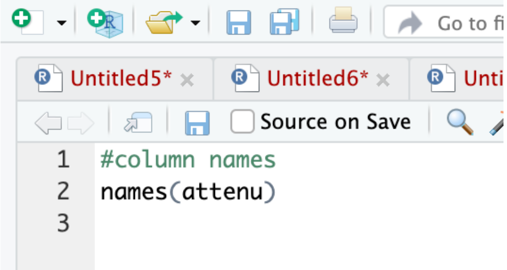
Output –



```
> #view the last 30 rows
> tail(attenu,30)
    event mag station dist accel
153    21 5.8    1299 33.1 0.056
154    21 5.8    1219 40.3 0.065
155    22 5.5    <NA>  4.0 0.259
156    22 5.5    <NA> 10.1 0.267
157    22 5.5    1030 11.1 0.071
158    22 5.5    1418 17.7 0.275
159    22 5.5    1383 22.5 0.058
160    22 5.5    <NA> 26.5 0.026
161    22 5.5    1299 29.0 0.039
162    22 5.5    1308 30.9 0.112
163    22 5.5    1219 37.8 0.065
164    22 5.5    1456 48.3 0.026
165    23 5.3    5045  5.8 0.123
166    23 5.3    5044 12.0 0.133
167    23 5.3    5160 12.1 0.073
168    23 5.3    5043 20.5 0.097
169    23 5.3    5047 20.5 0.096
170    23 5.3    c168 25.3 0.230
171    23 5.3    5068 35.9 0.082
172    23 5.3    c118 36.1 0.110
173    23 5.3    5042 36.3 0.110
174    23 5.3    5067 38.5 0.094
175    23 5.3    5049 41.4 0.040
176    23 5.3    c204 43.6 0.050
177    23 5.3    5070 44.4 0.022
178    23 5.3    c266 46.1 0.070
179    23 5.3    c203 47.1 0.080
180    23 5.3    5069 47.7 0.033
181    23 5.3    5073 49.2 0.017
182    23 5.3    5072 53.1 0.022
> 
```

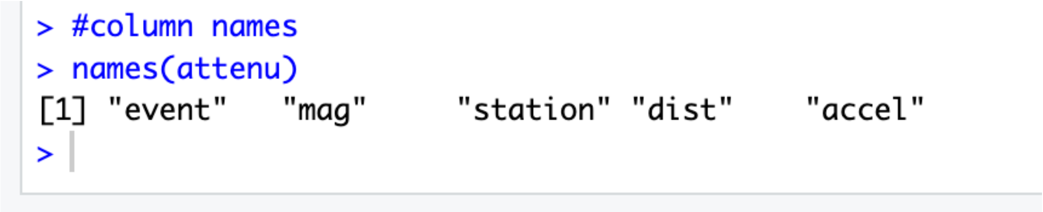
5. **Name(attenu)** : It is used to showcase the column names of the dataset.

Input –



```
#column names
names(attenu)
```

Output –

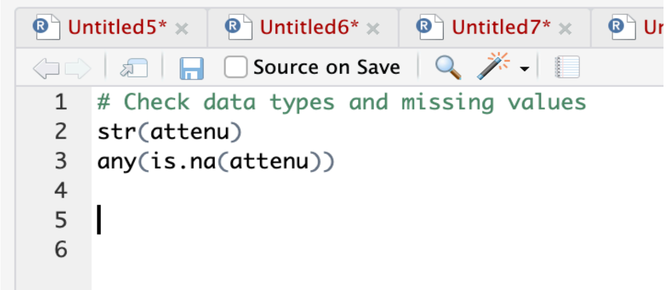


```
> #column names
> names(attenu)
[1] "event"   "mag"     "station" "dist"    "accel"
> 
```

6. **Str(attenu)** : It is used to view the structure of the dataset.

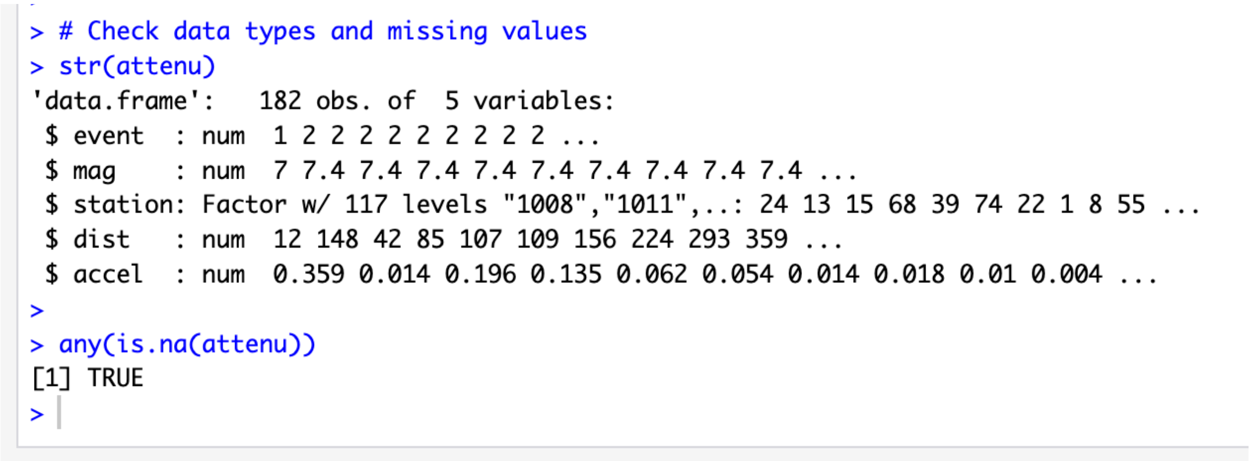
any(is.na(attenu)) : It is used to check missing values.

Input –



```
# Check data types and missing values
str(attenu)
any(is.na(attenu))
```

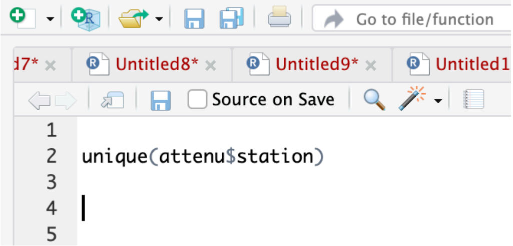
Output –



```
> # Check data types and missing values
> str(attenu)
'data.frame':	182 obs. of  5 variables:
 $ event  : num  1 2 2 2 2 2 2 2 2 2 ...
 $ mag    : num  7 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 7.4 ...
 $ station: Factor w/ 117 levels "1008","1011",..: 24 13 15 68 39 74 22 1 8 55 ...
 $ dist   : num  12 148 42 85 107 109 156 224 293 359 ...
 $ accel  : num  0.359 0.014 0.196 0.135 0.062 0.054 0.014 0.018 0.01 0.004 ...
> 
> any(is.na(attenu))
[1] TRUE
> 
```

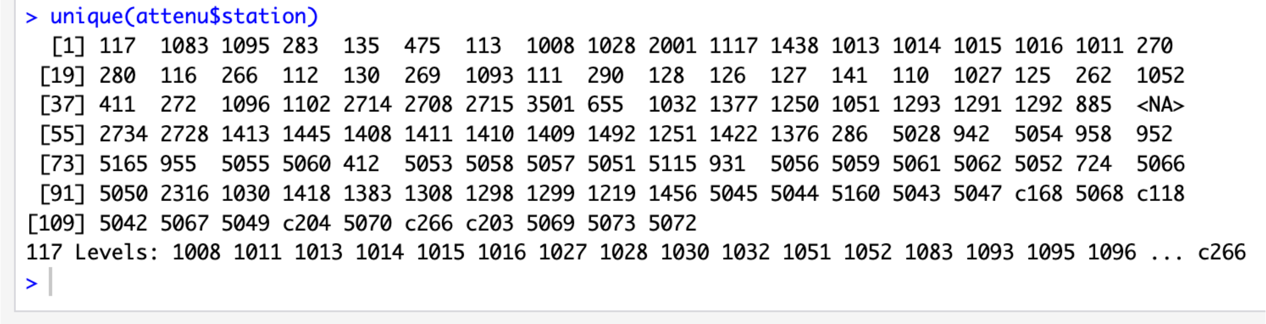
7. **unique(attenu)** : It is used to extract unique values present in the 'station' column of the dataset.

Input –



```
unique(attenu$station)

|
```

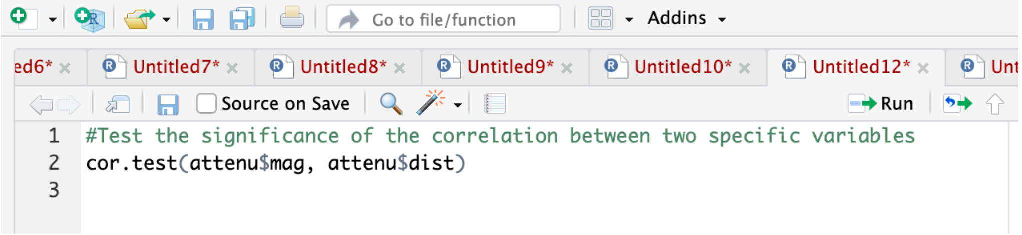
Output –



```
> unique(attenu$station)
  [1] 117  1083 1095 283  135  475  113  1008 1028 2001 1117 1438 1013 1014 1015 1016 1011 270 
 [19] 280  116  266  112  130  269  1093 111  290  128  126  127  141  110  1027 125  262  1052
 [37] 411  272  1096 1102 2714 2708 2715 3501 655  1032 1377 1250 1051 1293 1291 1292 885  <NA>
 [55] 2734 2728 1413 1445 1408 1411 1410 1409 1492 1251 1422 1376 286  5028 942  5054 958  952 
 [73] 5165 955  5055 5060 412  5053 5058 5057 5051 5115 931  5056 5059 5061 5062 5052 724  5066
 [91] 5050 2316 1030 1418 1383 1308 1298 1299 1219 1456 5045 5044 5160 5043 5047 c168 5068 c118
[109] 5042 5067 5049 c204 5070 c266 c203 5069 5073 5072
117 Levels: 1008 1011 1013 1014 1015 1016 1027 1028 1030 1032 1051 1052 1083 1093 1095 1096 ... c266
> |
```

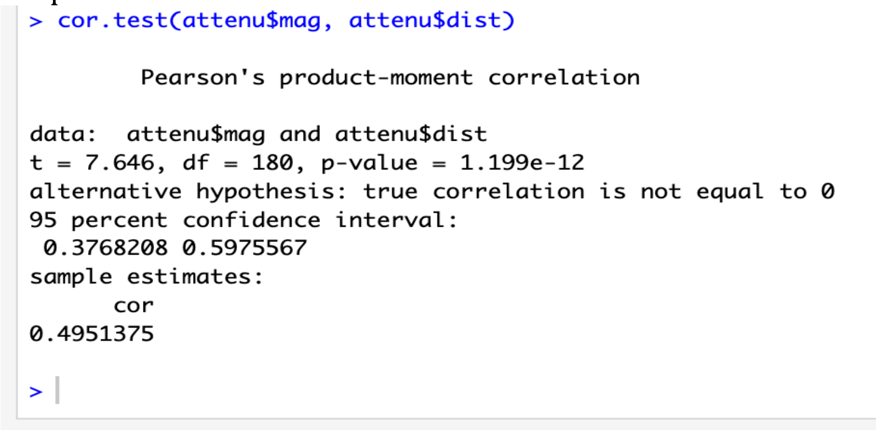
8. **cor.test(attenu$mag, attenu$dist)** : This function tests the significance of the correlation between two specific variables

Input –



```
#Test the significance of the correlation between two specific variables
cor.test(attenu$mag, attenu$dist)

```

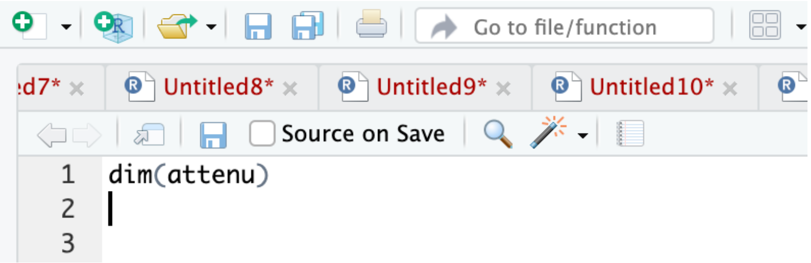
Output –



```
> cor.test(attenu$mag, attenu$dist)

	Pearson's product-moment correlation

data:  attenu$mag and attenu$dist
t = 7.646, df = 180, p-value = 1.199e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3768208 0.5975567
sample estimates:
      cor 
0.4951375 

> |
```

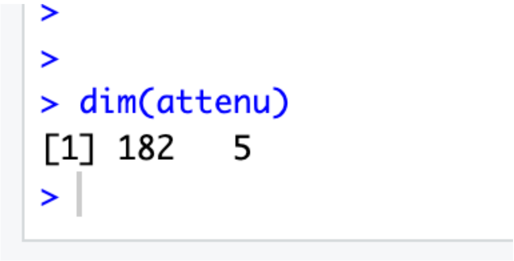
9. **dim(attenu)**: It is used to get the number of rows and columns in the dataset.

Input –



```
dim(attenu)
```

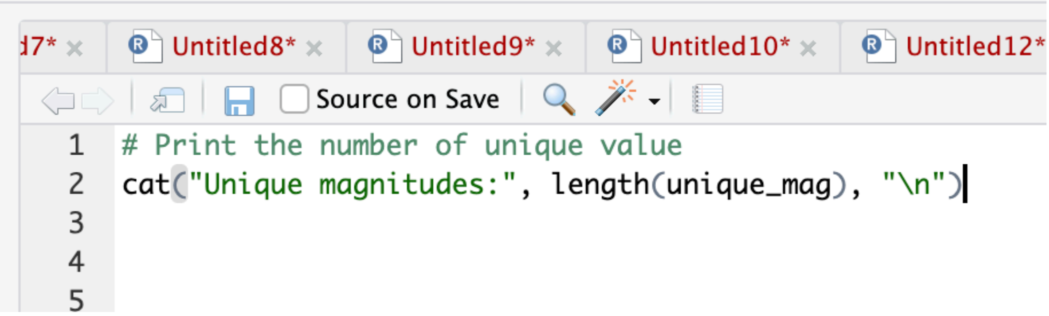
Output –



```
>
>
> dim(attenu)
[1] 182   5
> 
```

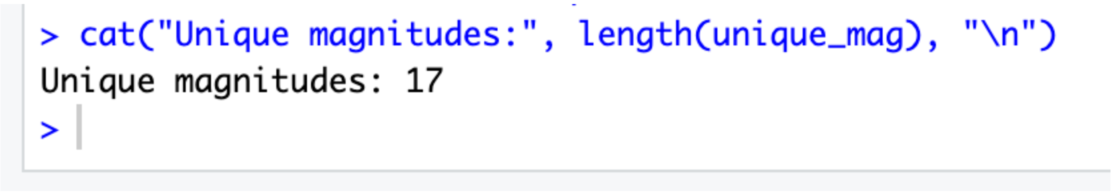
10. **cat("unique magnitudes:", length(unique_mag), "\n")** = It is used to print the number of unique value.

Input –



```
# Print the number of unique value
cat("Unique magnitudes:", length(unique_mag), "\n")
```

Output –

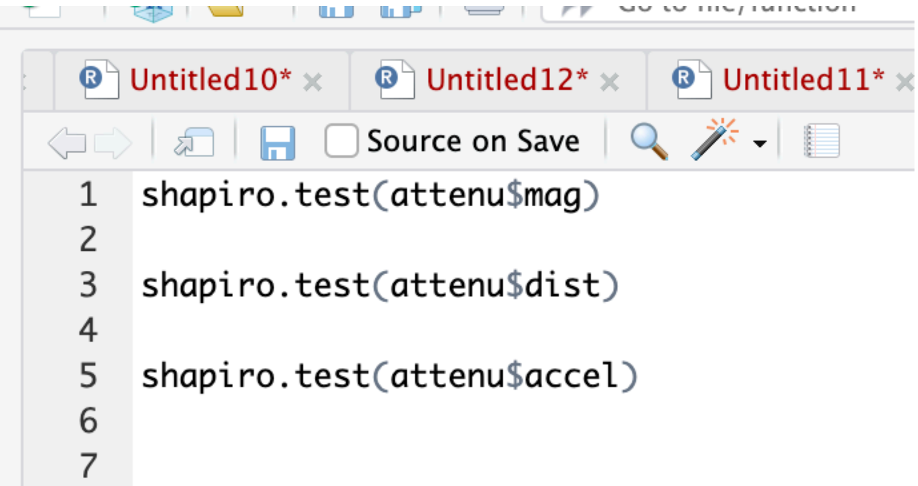


```
> cat("Unique magnitudes:", length(unique_mag), "\n")
Unique magnitudes: 17
> 
```

Stastical Analysis of attenu Dataset:

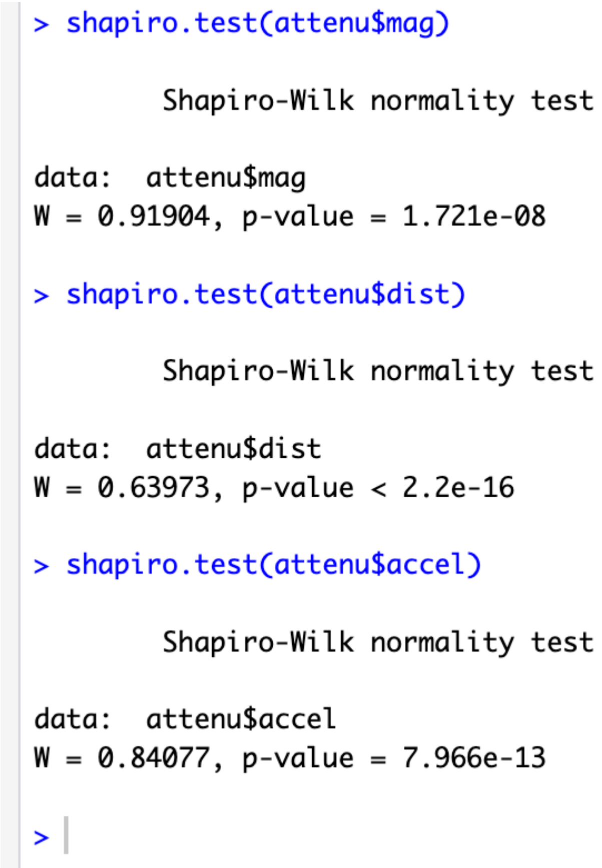
11. shapiro.test(attenu$variables): It is used to check the normality of a variable using the Shapiro-Wilk test.

Input –



```
shapiro.test(attenu$mag)

shapiro.test(attenu$dist)

shapiro.test(attenu$accel)
```

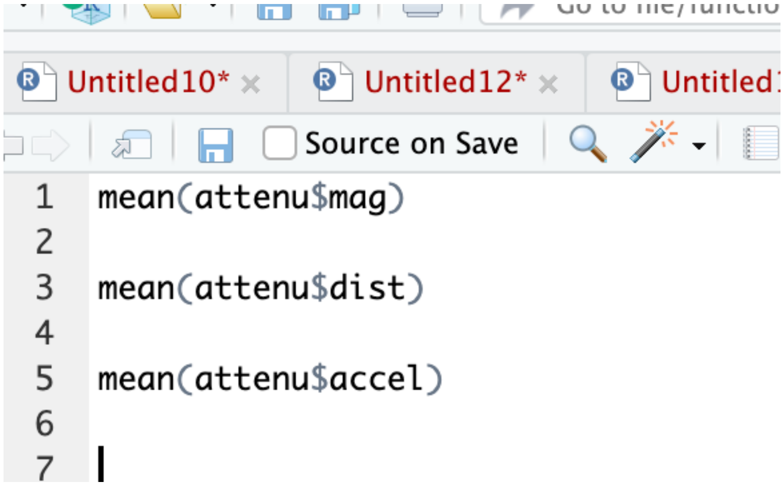
Output –



```
> shapiro.test(attenu$mag)

        Shapiro-Wilk normality test

data:  attenu$mag
W = 0.91904, p-value = 1.721e-08

> shapiro.test(attenu$dist)

        Shapiro-Wilk normality test

data:  attenu$dist
W = 0.63973, p-value < 2.2e-16

> shapiro.test(attenu$accel)

        Shapiro-Wilk normality test

data:  attenu$accel
W = 0.84077, p-value = 7.966e-13

> 
```

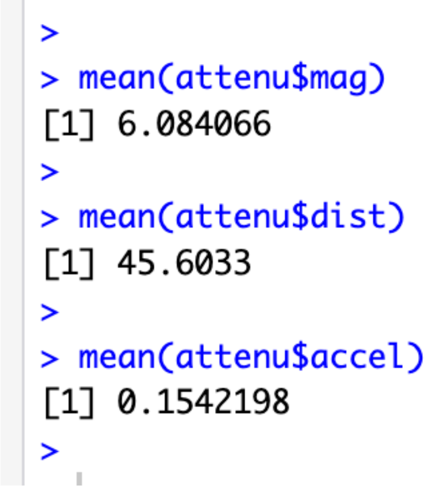
12. mean(variable): Sum of all observation divided by the total number of observation in the dataset.

Input –



```
mean(attenu$mag)

mean(attenu$dist)

mean(attenu$accel)

```

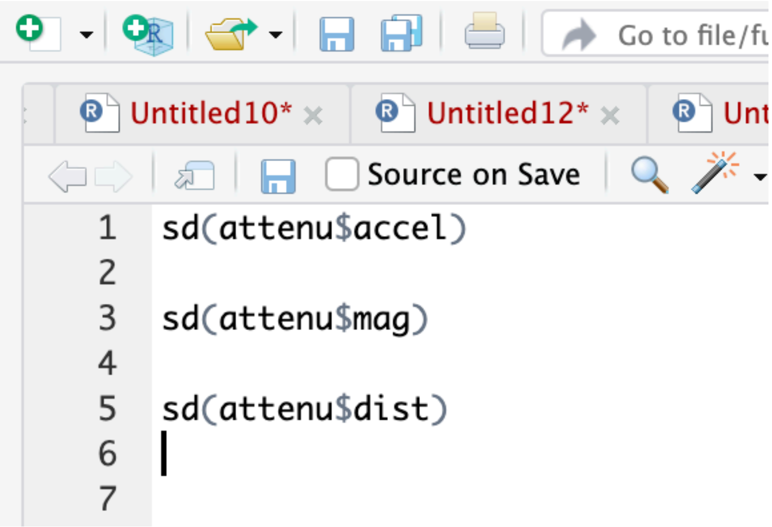
Output –



```
>
> mean(attenu$mag)
[1] 6.084066
>
> mean(attenu$dist)
[1] 45.6033
>
> mean(attenu$accel)
[1] 0.1542198
>
```

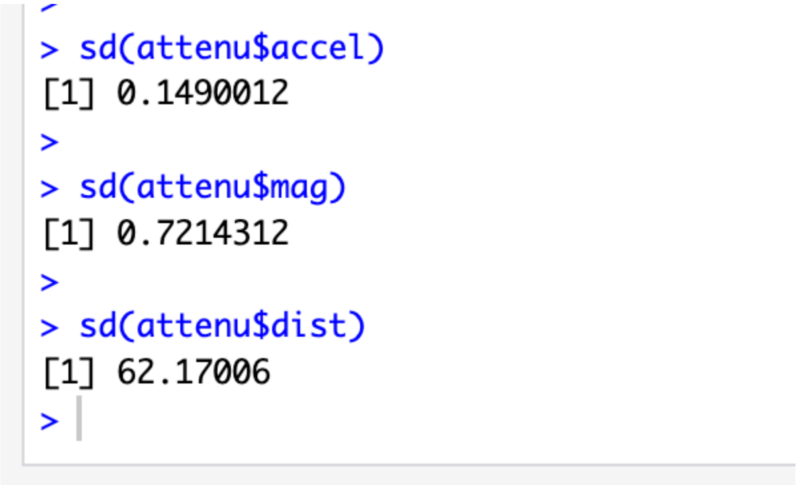
13. **sd(variable):** A statistical tool used to quantify the degree of variation or dispersion in a set of data values is the standard deviation.

Input –



```
sd(attenu$accel)

sd(attenu$mag)

sd(attenu$dist)

```

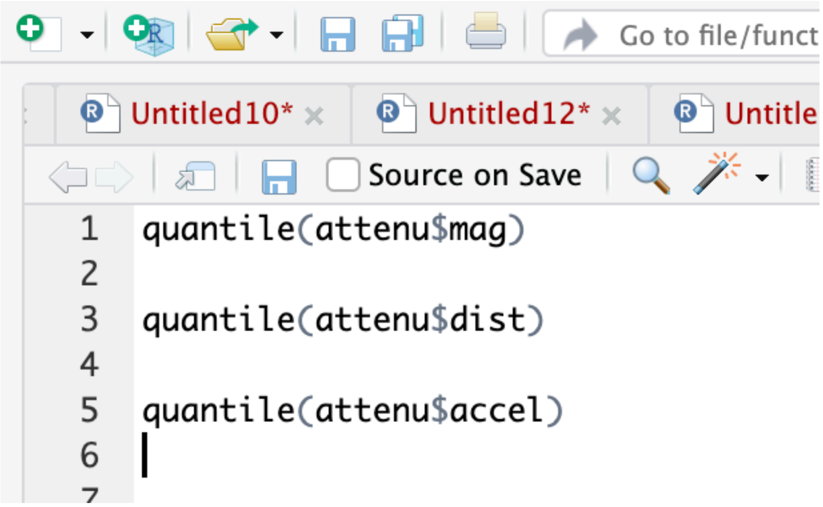
Output –



```
> sd(attenu$accel)
[1] 0.1490012
>
> sd(attenu$mag)
[1] 0.7214312
>
> sd(attenu$dist)
[1] 62.17006
>
```

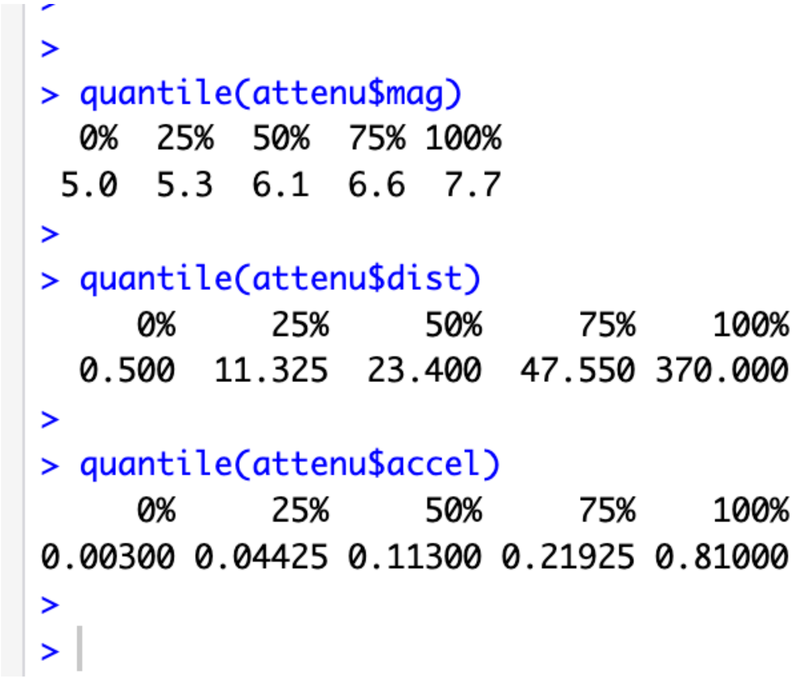
14. quantile(iris): Values known as quantiles divide a dataset into intervals with equal probability.

Input –



```
quantile(attenu$mag)

quantile(attenu$dist)

quantile(attenu$accel)
```

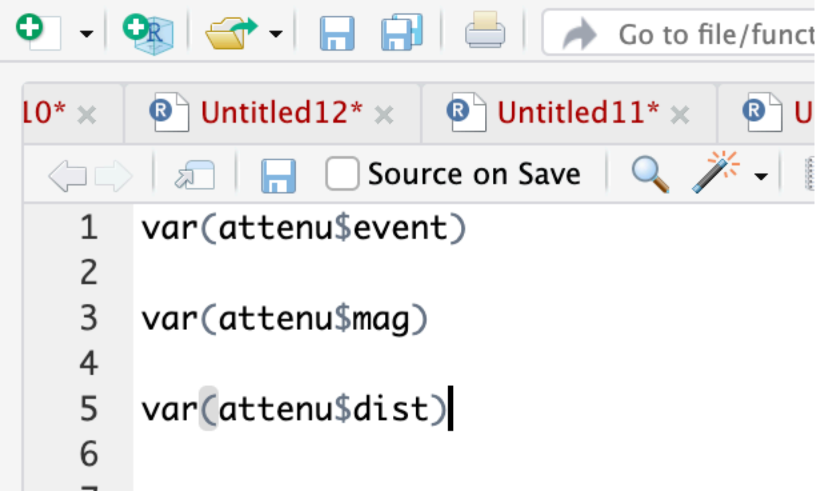
Output –



```
>
> quantile(attenu$mag)
 0%  25%  50%  75% 100% 
5.0  5.3  6.1  6.6  7.7 
>
> quantile(attenu$dist)
     0%     25%     50%     75%    100% 
  0.500  11.325  23.400  47.550 370.000 
>
> quantile(attenu$accel)
     0%     25%     50%     75%    100% 
0.00300 0.04425 0.11300 0.21925 0.81000 
>
> 
```

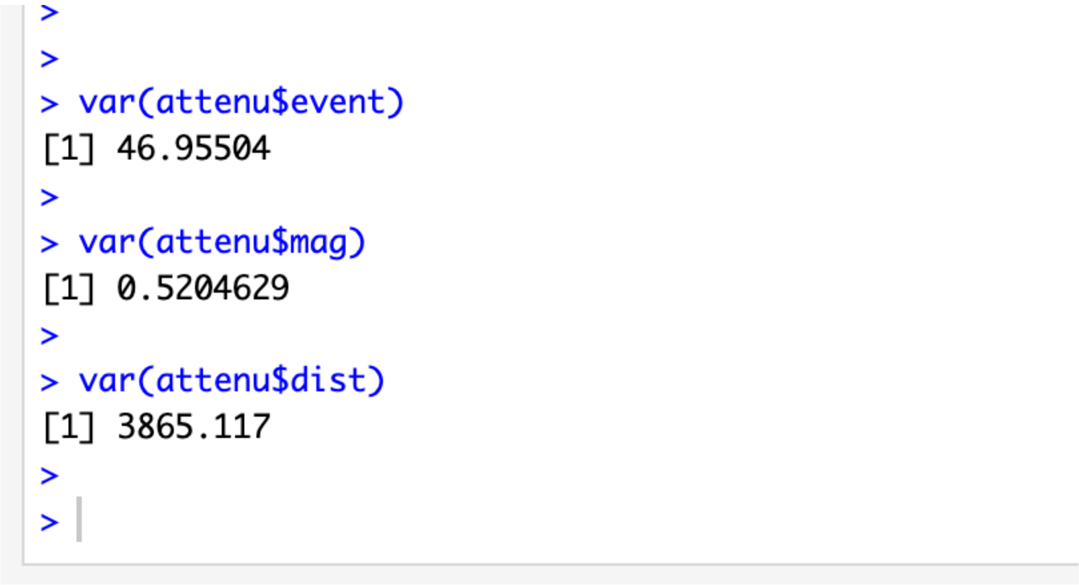
15. Variance : It is used to find variance of the variable.

Input –



```
var(attenu$event)

var(attenu$mag)

var(attenu$dist)

```

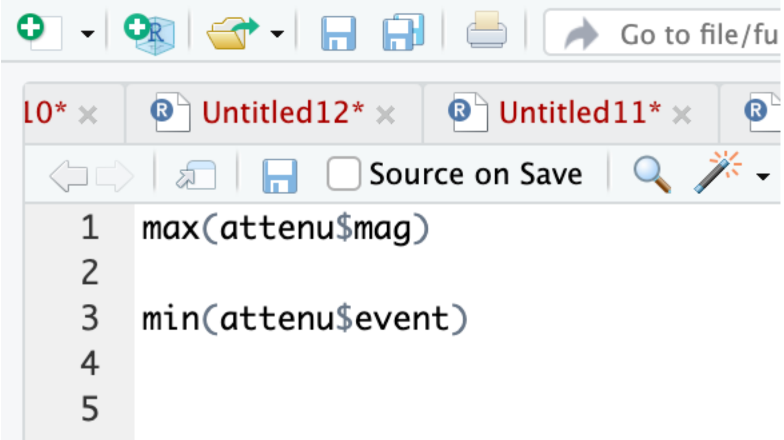
Output –



```
>
>
> var(attenu$event)
[1] 46.95504
>
> var(attenu$mag)
[1] 0.5204629
>
> var(attenu$dist)
[1] 3865.117
>
> 
```

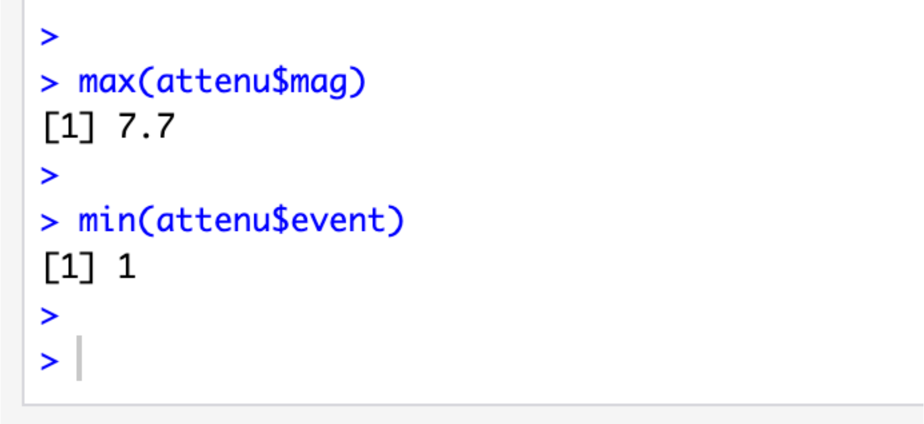
16. Finding the maximum and minimum values :

Input –



```
max(attenu$mag)

min(attenu$event)


```

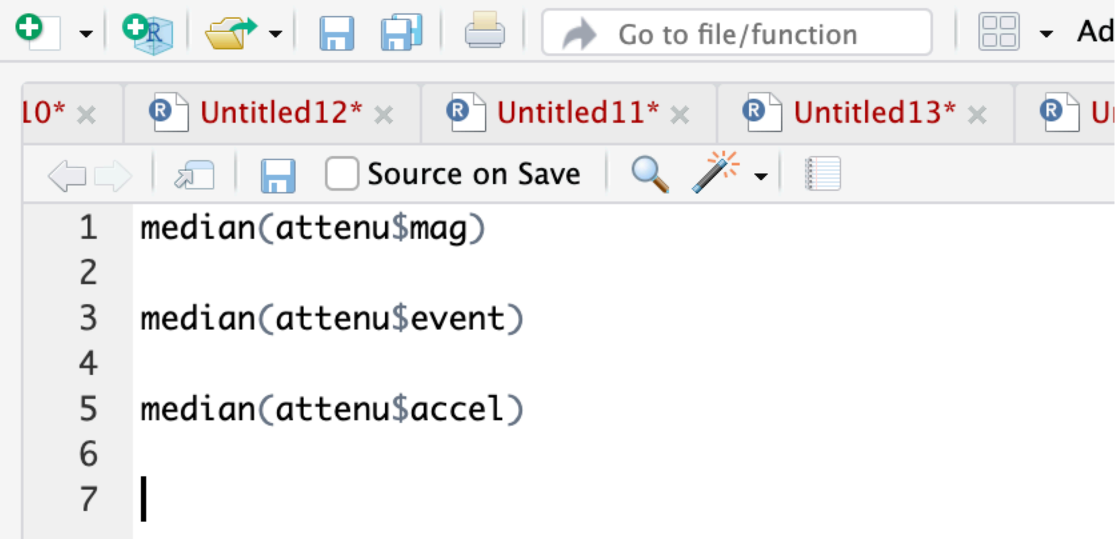
Output –



```
>
> max(attenu$mag)
[1] 7.7
>
> min(attenu$event)
[1] 1
>
> 
```

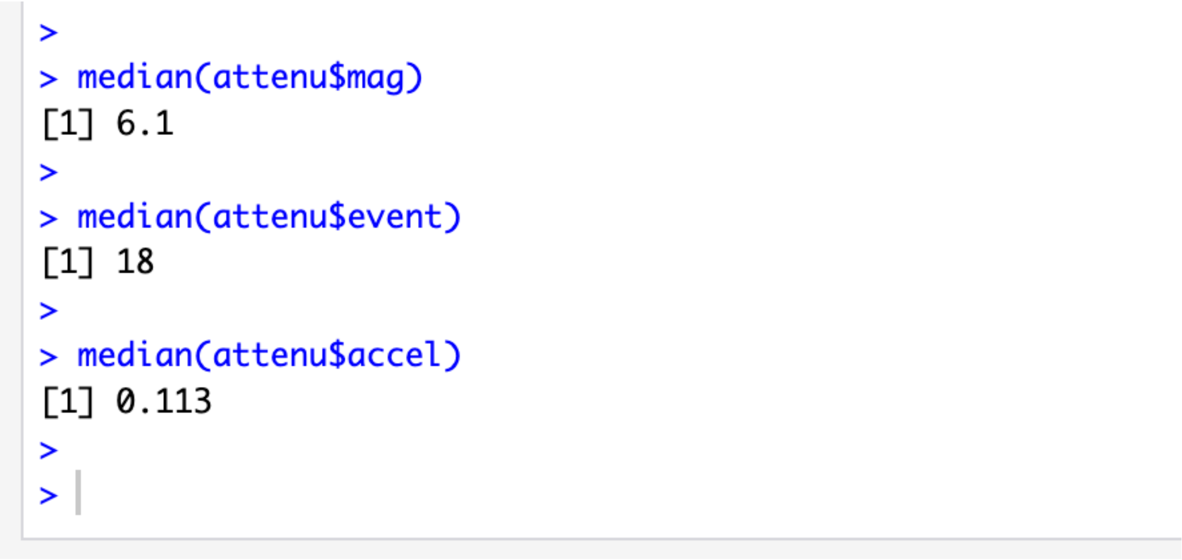
17.Median : It is used to find median of variables .

Input –



```
median(attenu$mag)

median(attenu$event)

median(attenu$accel)

```

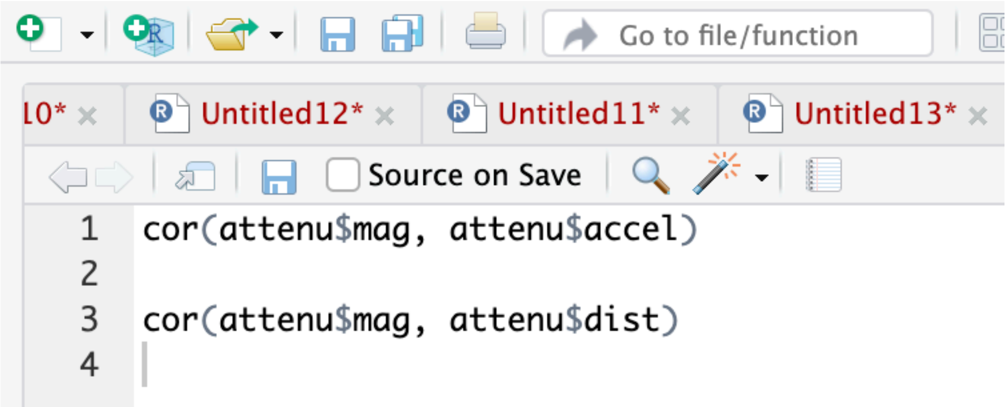
Output –



```
>
> median(attenu$mag)
[1] 6.1
>
> median(attenu$event)
[1] 18
>
> median(attenu$accel)
[1] 0.113
>
> 
```

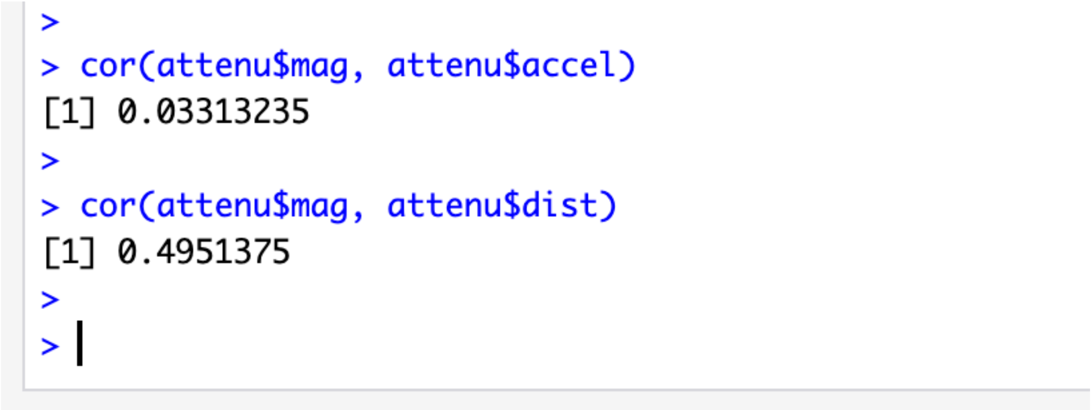
18. Co relation : It is used to find co relation between two variables .

Input –



```
cor(attenu$mag, attenu$accel)

cor(attenu$mag, attenu$dist)

```

Output –

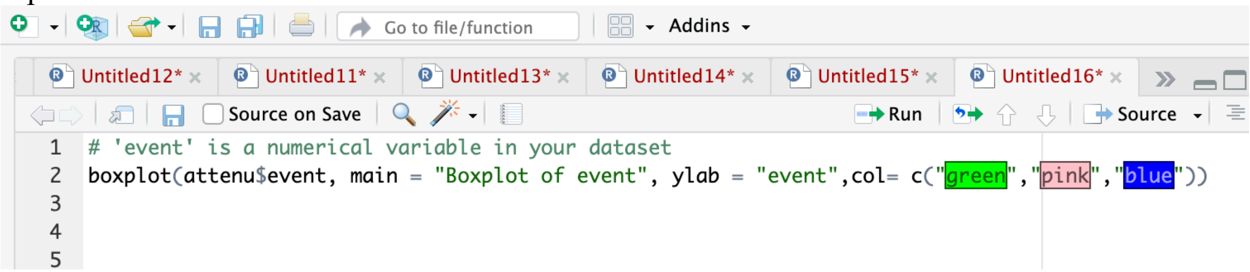


```
>
> cor(attenu$mag, attenu$accel)
[1] 0.03313235
>
> cor(attenu$mag, attenu$dist)
[1] 0.4951375
>
> 
```

Data visulaization :

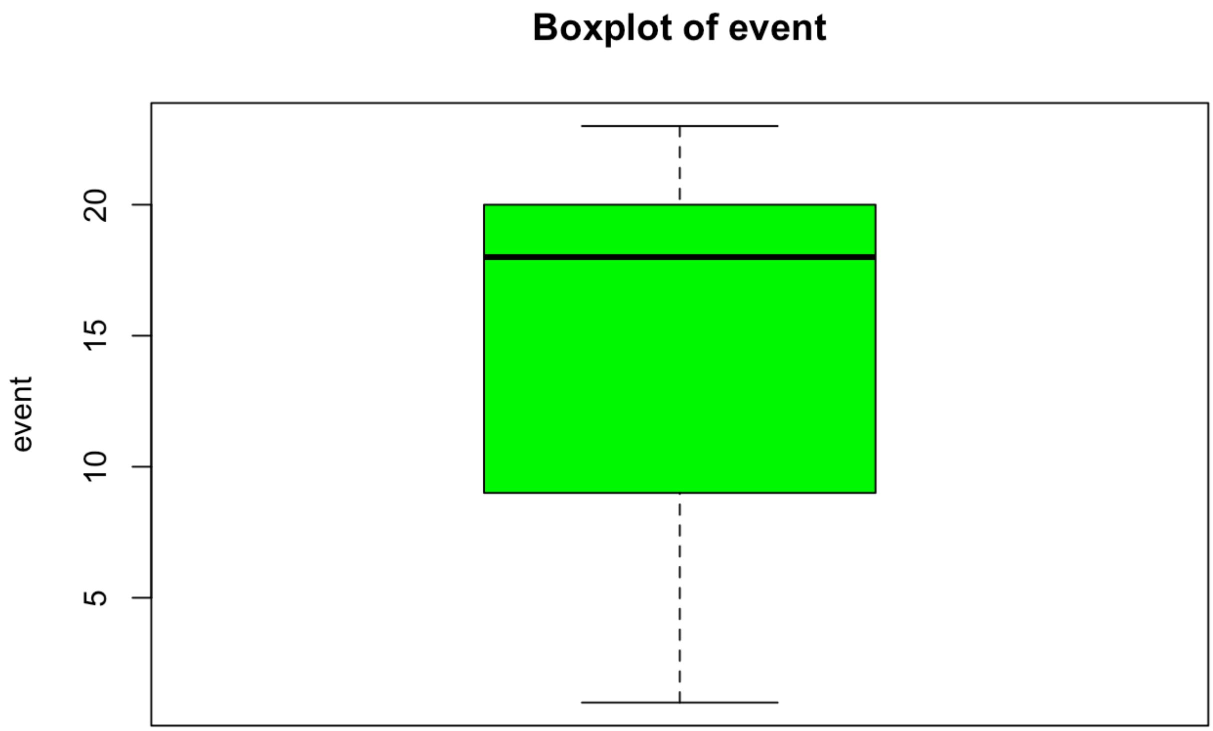
19. Boxplot :

Input –



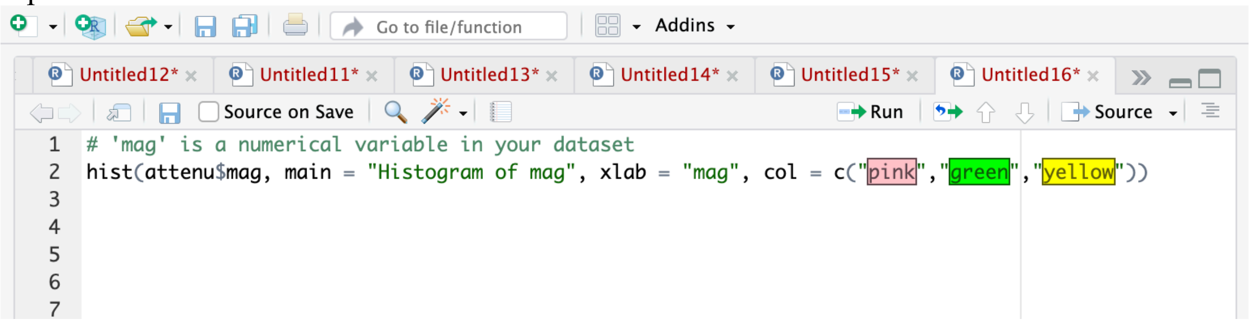
```
# 'event' is a numerical variable in your dataset
boxplot(attenu$event, main = "Boxplot of event", ylab = "event",col= c("green","pink","blue"))
```

Output –



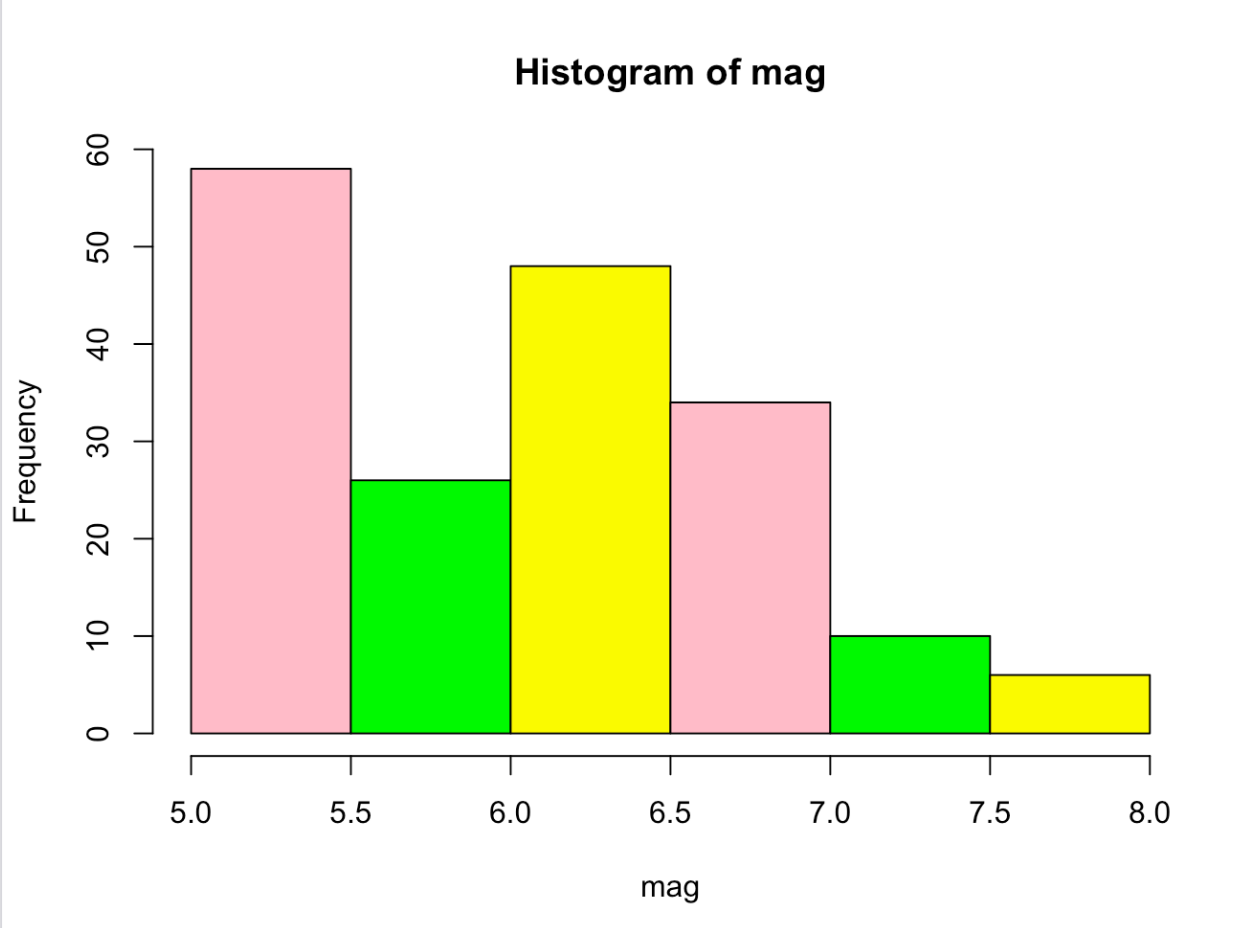
20. Histogram :

Input –



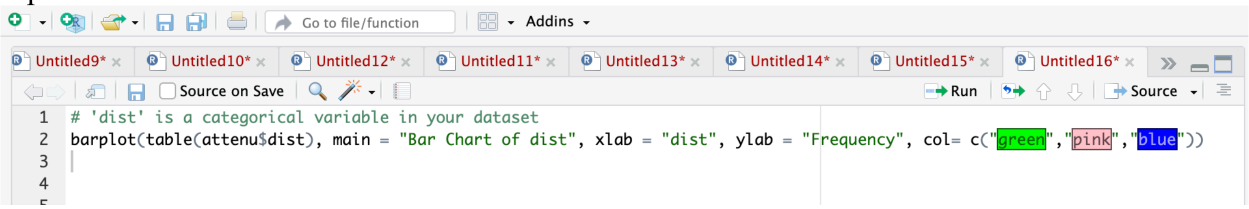
```
# 'mag' is a numerical variable in your dataset
hist(attenu$mag, main = "Histogram of mag", xlab = "mag", col = c("pink","green","yellow"))
```

Output –



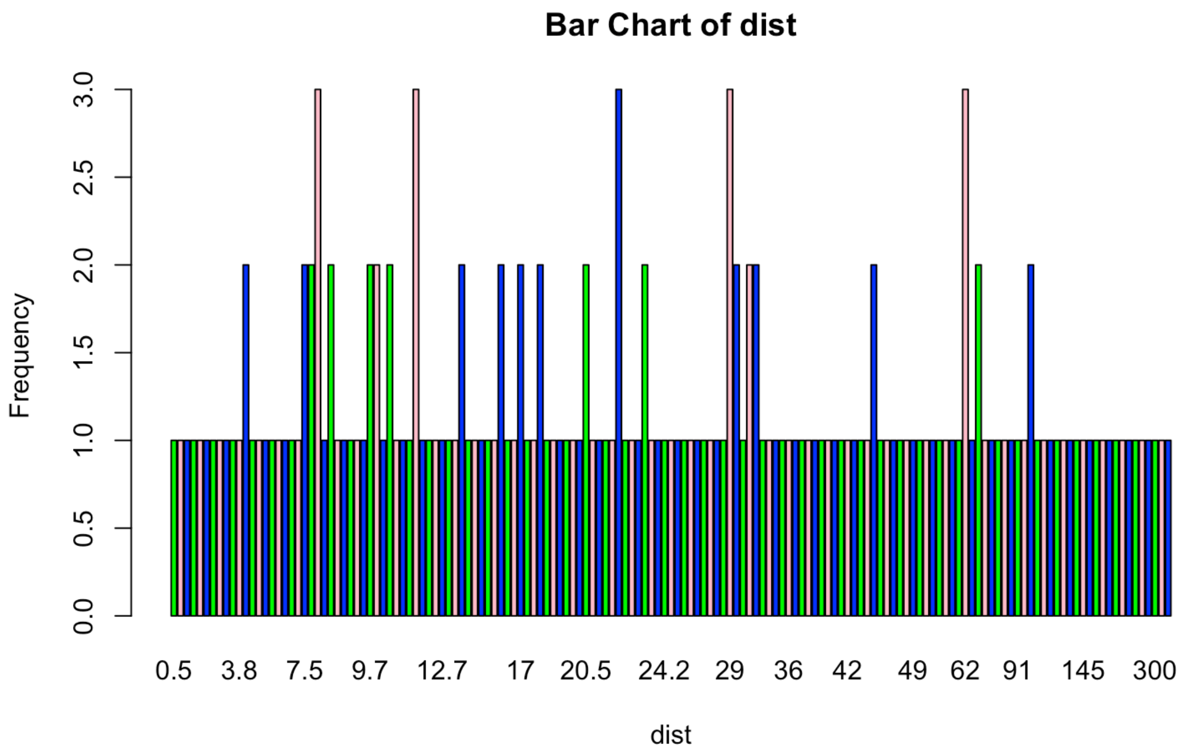
21. Barchart :

Input –



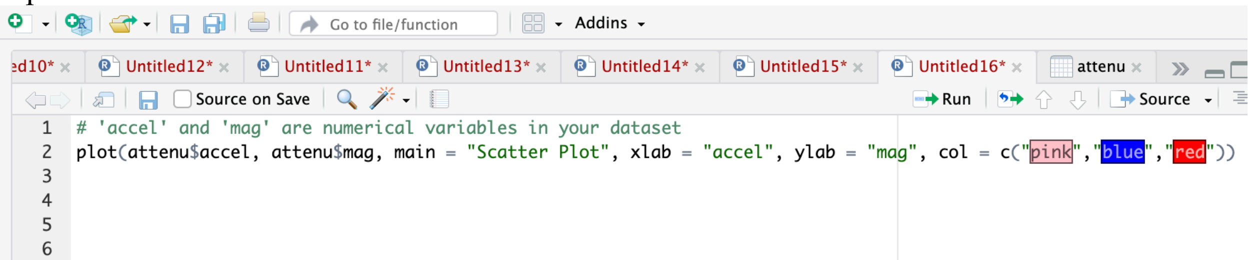
```r
# 'dist' is a categorical variable in your dataset
barplot(table(attenu$dist), main = "Bar Chart of dist", xlab = "dist", ylab = "Frequency", col= c("green","pink","blue"))
```

Output –



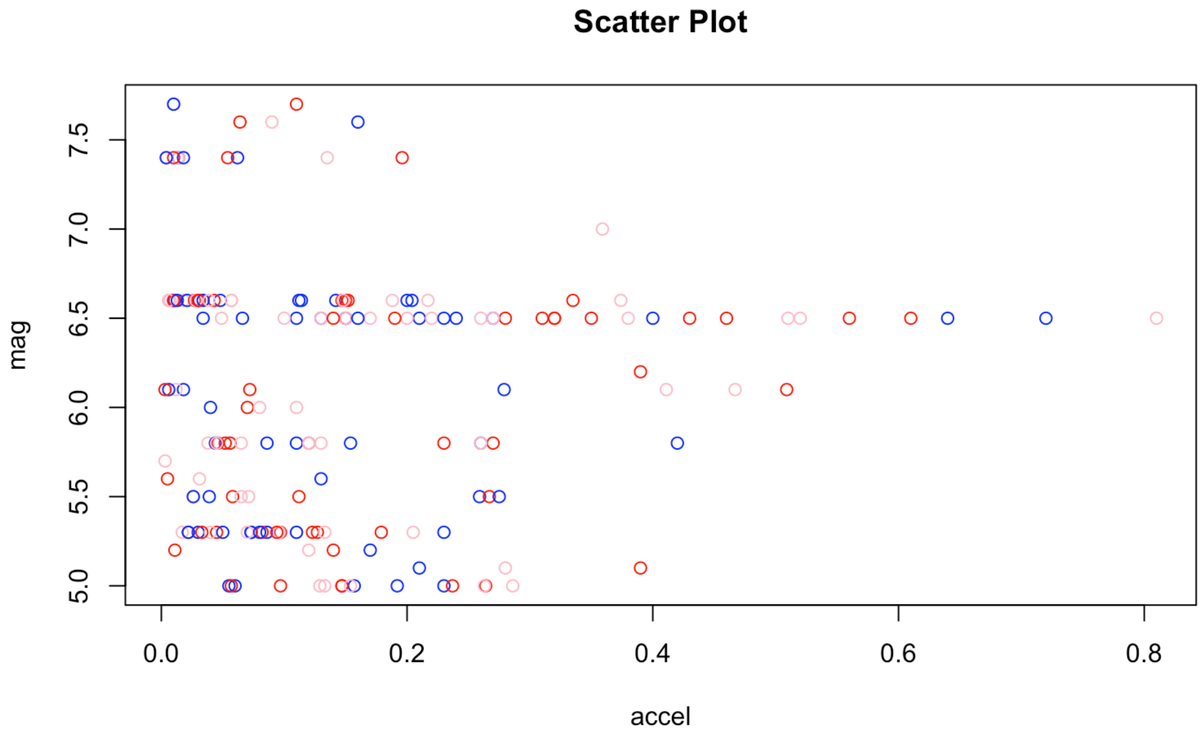
22. Scatter plot :

Input –



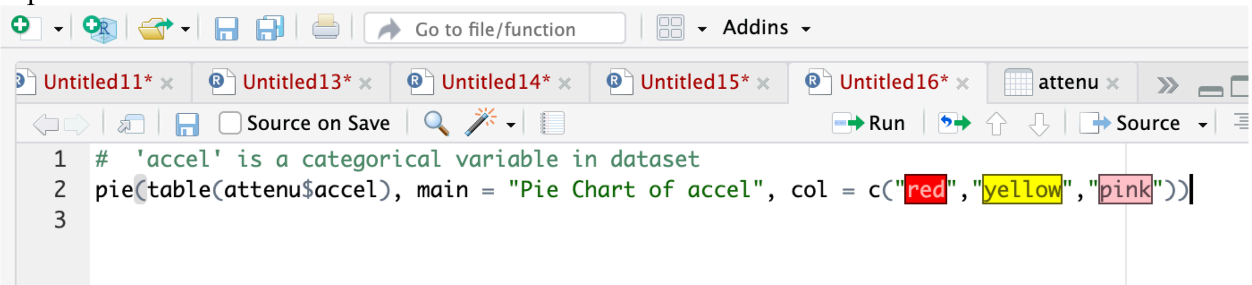
```
# 'accel' and 'mag' are numerical variables in your dataset
plot(attenu$accel, attenu$mag, main = "Scatter Plot", xlab = "accel", ylab = "mag", col = c("pink","blue","red"))
```

Output –



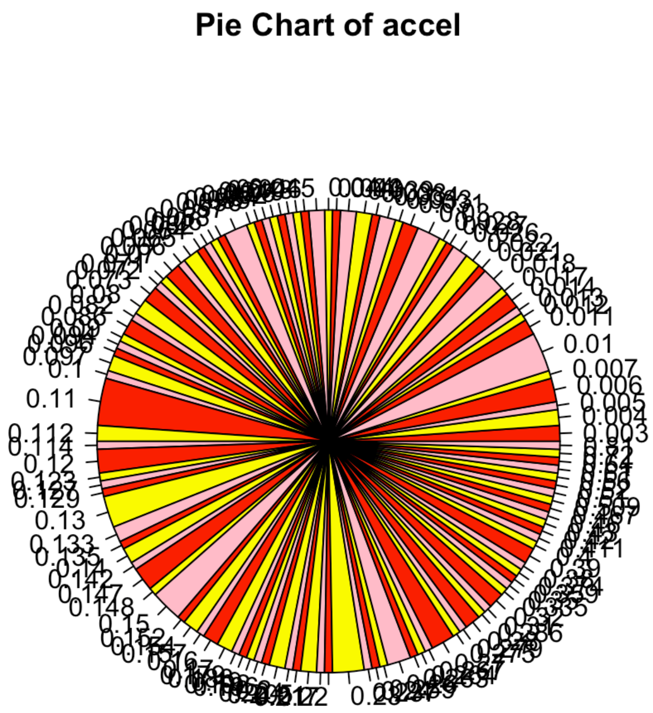
23. Pie chart :

Input –



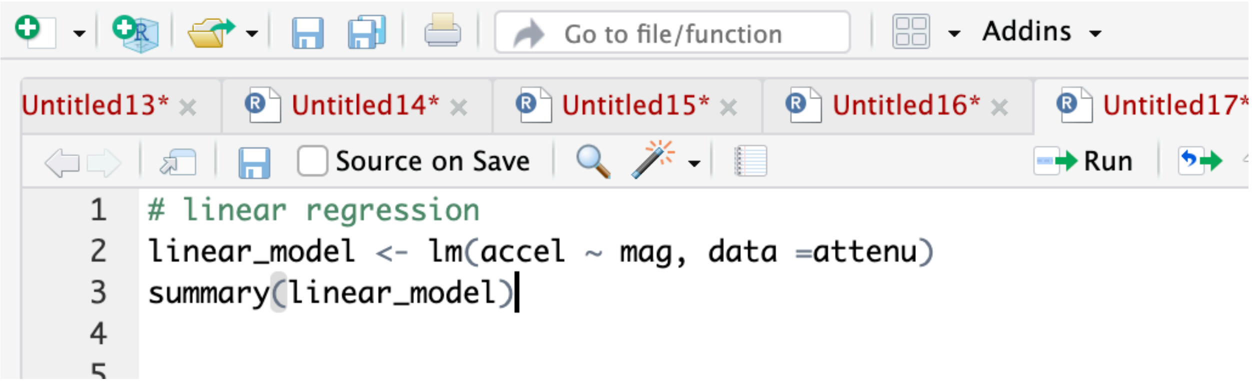
```
#  'accel' is a categorical variable in dataset
pie(table(attenu$accel), main = "Pie Chart of accel", col = c("red","yellow","pink"))
```

Output –



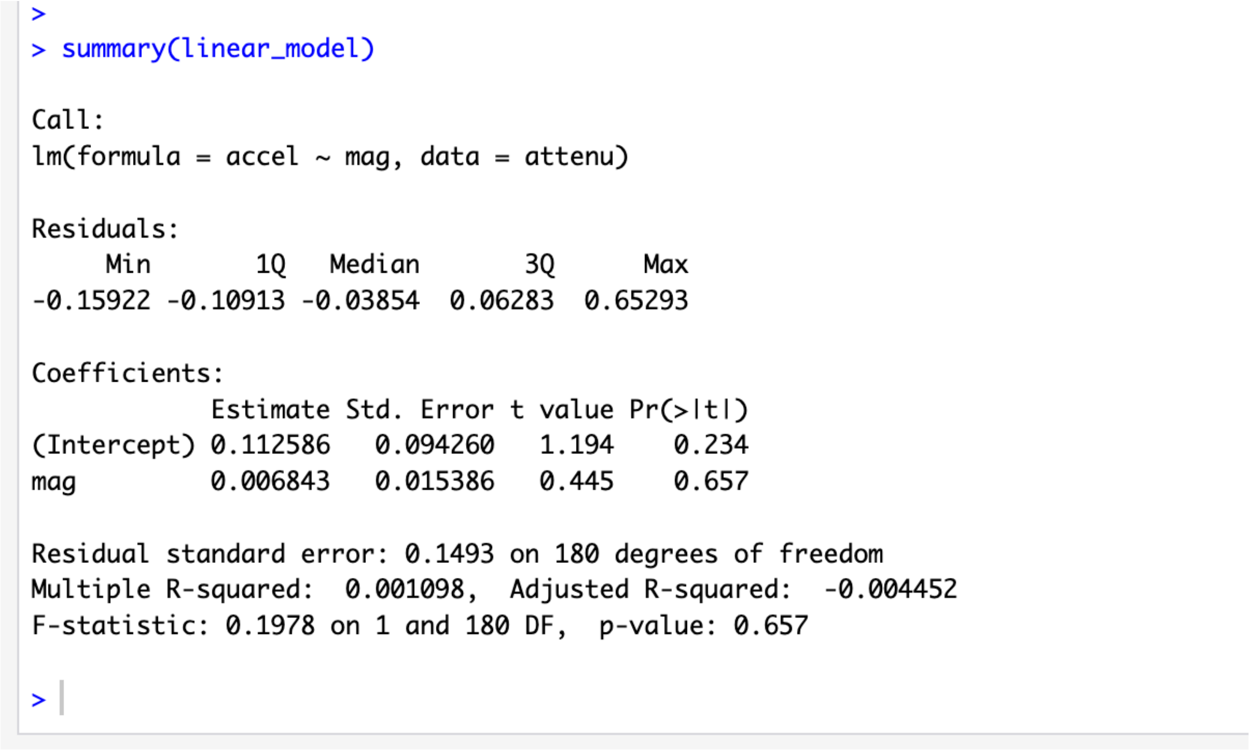
24. Linear regression :

Input –



```
# linear regression
linear_model <- lm(accel ~ mag, data =attenu)
summary(linear_model)
```

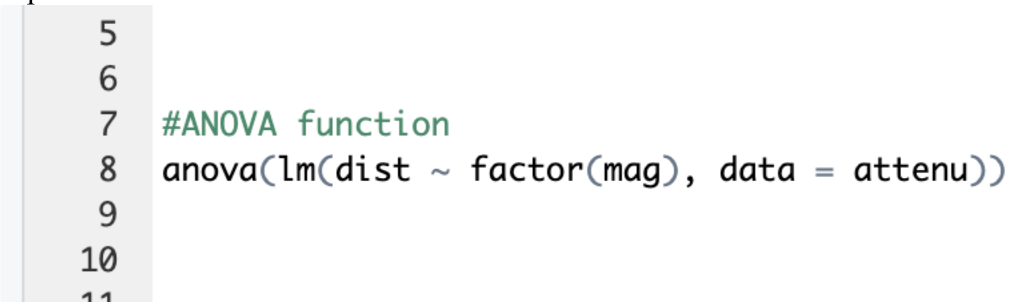
Output –



```
>
> summary(linear_model)

Call:
lm(formula = accel ~ mag, data = attenu)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.15922 -0.10913 -0.03854  0.06283  0.65293 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.112586   0.094260   1.194    0.234
mag         0.006843   0.015386   0.445    0.657

Residual standard error: 0.1493 on 180 degrees of freedom
Multiple R-squared:  0.001098,	Adjusted R-squared:  -0.004452 
F-statistic: 0.1978 on 1 and 180 DF,  p-value: 0.657

> 
```

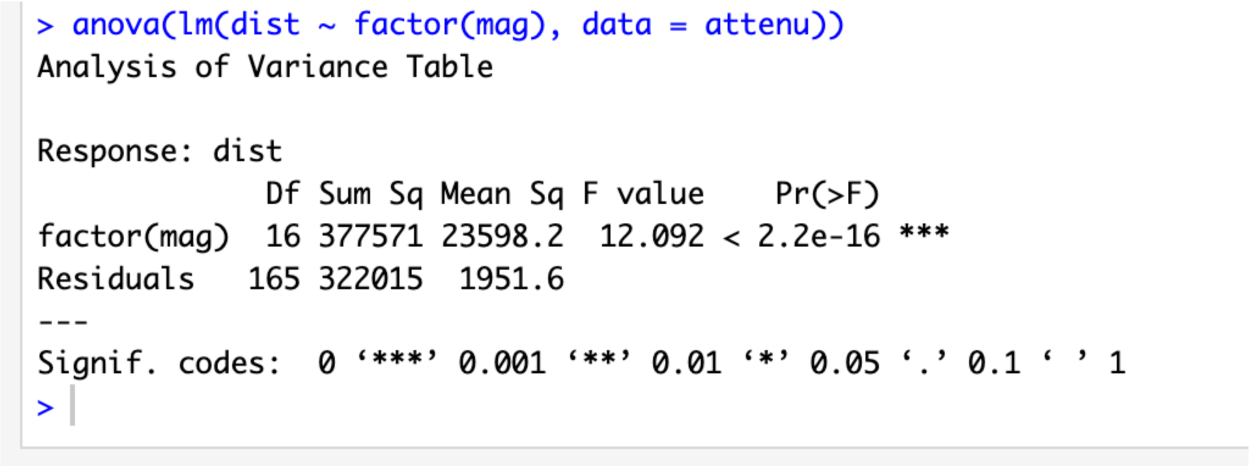
25. Anova function :

Input –



```
#ANOVA function
anova(lm(dist ~ factor(mag), data = attenu))
```

Output –



```
> anova(lm(dist ~ factor(mag), data = attenu))
Analysis of Variance Table

Response: dist
             Df Sum Sq Mean Sq F value    Pr(>F)    
factor(mag)  16 377571 23598.2  12.092 < 2.2e-16 ***
Residuals   165 322015  1951.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```