Data_Wrangling

Vishal

2025-02-14

R Markdown

```
df <- read.csv("D:/Prog for DS/Vit_reviews_data_final.csv")
head(df)
```

```
##                        Name Rating                                          Department
## 1 Mukesh Yashvanth     3.4                      B.Sc, Animation & Multimedia
## 2    Saamya Kunhibi     3.6                      B.Sc, Animation & Multimedia
## 3        Punya Oswal    4.0 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal     5.0          B.Tech, Computer Science and Engineering
## 5             GDivya    3.3             B.Tech + M.Tech, Software Engineering
## 6        Review guru    3.9          B.Tech, Computer Science and Engineering
##     Review_Date
## 1 Feb 13, 2025
## 2 Feb 13, 2025
## 3 Feb 12, 2025
## 4 Feb 12, 2025
## 5 Feb 12, 2025
## 6 Feb 11, 2025
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

#Dicover

```
dim(df)  # Returns the number of rows and columns
```

```
## [1] 3268    5
```

```
str(df)  # Displays structure, including column types
```

```
## 'data.frame':    3268 obs. of  5 variables:
##  $ Name       : chr  "Mukesh Yashvanth" "Saamya Kunhibi" "Punya Oswal" "Soumyajit Ghosal"
...
##  $ Rating     : num  3.4 3.6 4 5 3.3 3.9 3.5 3.9 3.6 4 ...
##  $ Department : chr  "B.Sc, Animation & Multimedia" "B.Sc, Animation & Multimedia" "B.Tec
h, Electrical And Electronics Engineering" "B.Tech, Computer Science and Engineering" ...
##  $ Review_Date: chr  "Feb 13, 2025" "Feb 13, 2025" "Feb 12, 2025" "Feb 12, 2025" ...
##  $ Review_Text: chr  "One of my senior got placed in Couch Base for 34LPA CTC another got
placed in Unilever for 42LPA but this was f"| __truncated__ "I don't like how the student and
faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams yearly 3 exams in a "| __truncated_
_ "The university is vast to have best experience  Best faculty crew providing best source of
education  Security "| __truncated__ "Transportation is very bad making it difficult for stud
ents  The infrastructure of the university gives you a p"| __truncated__ ...
```

```
summary(df)  # Provides a summary of numerical and categorical columns
```

```
##      Name              Rating        Department        Review_Date
##  Length:3268       Min.   :0.500   Length:3268       Length:3268
##  Class :character  1st Qu.:3.700   Class :character  Class :character
##  Mode  :character  Median :4.100   Mode  :character  Mode  :character
##                    Mean   :4.031
##                    3rd Qu.:4.400
##                    Max.   :5.000
##                    NA's   :253
##  Review_Text
##  Length:3268
##  Class :character
##  Mode  :character
##
##
##
##
```

```
colSums(is.na(df))  # Counts missing values in each column
```

```
##        Name       Rating   Department  Review_Date  Review_Text
##         253          253          504          253            0
```

#Histogram Distibution of Rating

```
hist(df$Rating, main="Histogram of column_name", xlab="column_name")
```

Histogram of column_name

column_name
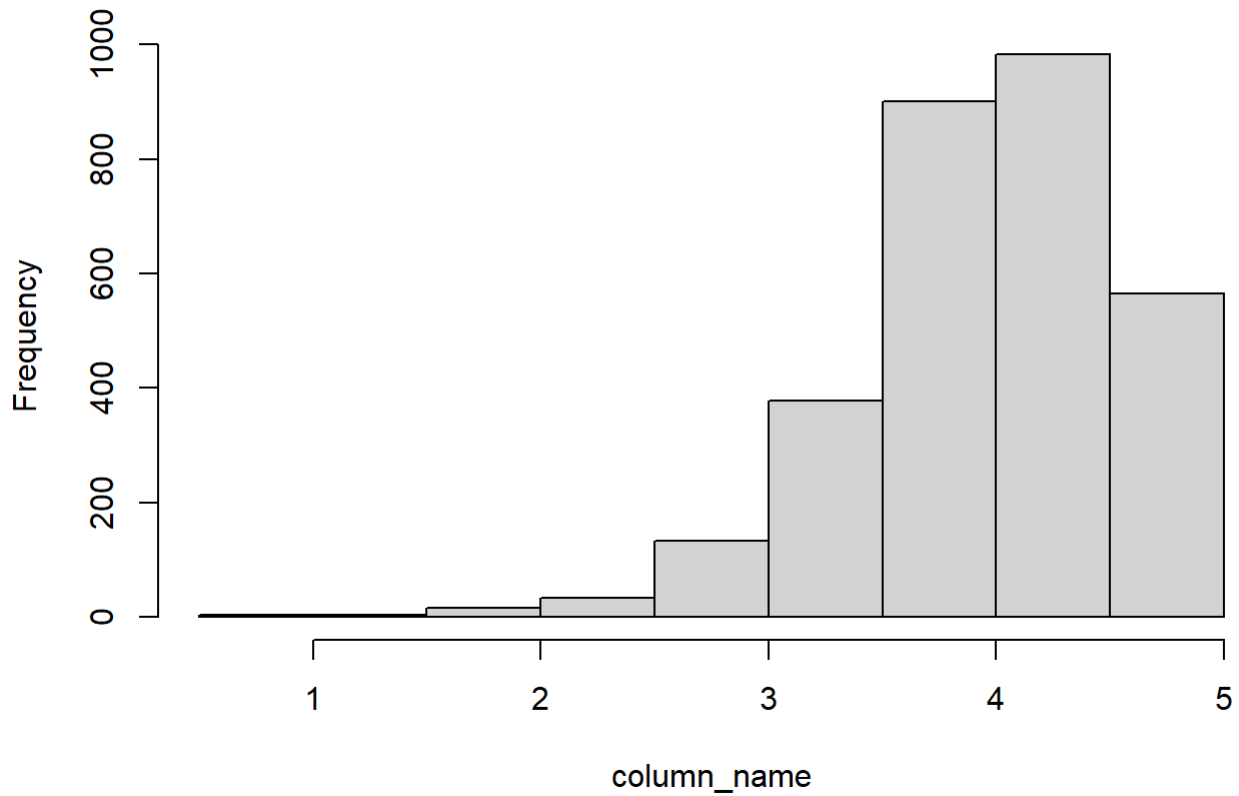
#Frequesncy Distribution of Departmentwise

```
sort(table(df$Department), decreasing = TRUE)
```

```
##
##                                     B.Tech, Computer Science and Engineering
##                                                                         1314
##                               B.Tech, Electronics & Communication Engineering
##                                                                          245
##                                                  B.Tech, Mechanical Engineering
##                                                                          171
##                         B.Tech, Artificial Intelligence & Machine Learning
##                                                                          114
##                                  B.Tech, Electrical And Electronics Engineering
##                                                                           93
##                                           B.Tech, Information Technology
##                                                                           65
##                                     B.Tech + M.Tech, Software Engineering
##                                                                           62
##                                                                          MCA
##                                                                           54
##                               B.Tech, Computer Science and Business Systems
##                                                                           45
##                                  B.Tech, Electronics & Computer Engineering
##                                                                           40
##                               B.Tech, Cyber Security and Digital Forensics
##                                                                           35
##                                                       B.Tech, Biotechnology
##                                                                           33
##                                                   B.Tech, Civil Engineering
##                                                                           31
##                          B.Tech + M.Tech, Computer Science & Engineering
##                                                                           27
##                                                                 BCA, General
##                                                                           24
##                                                B.Tech, Chemical Engineering
##                                                                           20
##                                             B.Tech, Cyber Physical Systems
##                                                                           18
##                                                      B.Sc, Computer Science
##                                                                           17
##          B.Tech + M.Tech, Computer Science and Engineering + Data Science
##                                                                           15
##                                              B.Tech, Biomedical Engineering
##                                                                           14
##                                                        B.Tech, Data Sciences
##                                                                           14
##                                               B.Tech, Information Security
##                                                                           14
##                                                                 BBA, General
##                                                                           14
##                                                              B.Com, General
##                                                                           13
##                               M.Tech, Computer Science And Engineering
##                                                                           13
##                                                         M.Sc, Data Science
##                                                                           12
##                                B.Tech + M.Tech, Artificial Intelligence
##                                                                           11
```

```
##                                            B.Tech + M.Tech, Business Analytics
##                                                                             10
##                     B.Tech, Electronics And Instrumentation Engineering
##                                                                             10
##                                                     B.Tech, Internet of Things
##                                                                             10
##                                                  B.Sc + M.Sc, Biotechnology
##                                                                              9
##                                                     B.Tech, Cloud Computing
##                                                                              9
##                     B.E / B.Tech (Computer Science and Engineering)
##                                                                              8
##                                                     B.Sc {Hons.}, Agriculture
##                                                                              8
##                                                  B.Sc + M.Sc, Data Science
##                                                                              7
##                     B.Tech + M.Tech, Computational and Data Science
##                                                                              7
##                                            B.Tech + M.Tech, Cyber security
##                                                                              7
##                                                  M.Sc, Applied Microbiology
##                                                                              7
##                                                  B.Tech, Aerospace Engineering
##                                                                              6
##                                                     B.Tech, Gaming Technology
##                                                                              6
##                              B.Tech, Mechatronics & Automation Engineering
##                                                                              6
##                                                              BALLB {Hons.}
##                                                                              6
##                  B.Sc + M.Sc, Computational Statistics and Data Analytics
##                                                                              5
##                                            B.Sc, Animation & Multimedia
##                                                                              5
##                                                     B.Tech, Data Analytics
##                                                                              5
##                              B.Tech, Electrical And Computer Engineering
##                                                                              5
##                                                     B.Tech, Bioengineering
##                                                                              4
##                                                     B.Tech, Bioinformatics
##                                                                              4
##                                                           B.Tech, Blockchain
##                                                                              4
##                                                  B.Tech, Fashion Technology
##                                                                              4
##                                            B.Tech, Healthcare Informatics
##                                                                              4
##                                                           B.Tech, VLSI Design
##                                                                              4
##                                                              M.Sc, Chemistry
##                                                                              4
##                                                        M.Tech, Mechatronics
##                                                                              4
##                                                                 MBA, Finance
##                                                                              4
```

```
##                                              B.Des, Industrial Design
##                                                                      3
##                                                    B.Tech, E-Commerce
##                                                                      3
##                                                      M.Sc, Biotechnology
##                                                                      3
##                                                          MBA, Marketing
##                                                                      3
##                                                                 B.Arch
##                                                                      2
##                                                         B.Com, Finance
##                                                                      2
##            B.E / B.Tech (Electronics and Communication Engineering)
##                                                                      2
##                                                  B.Sc + M.Sc, Chemistry
##                                                                      2
##                                                        B.Tech, Robotics
##                                                                      2
##                                                        BBALLB {Hons.}
##                                                                      2
##                                                         M.Sc, Physics
##                                                                      2
##                            M.Tech, Cyber Security & Digital Forensics
##                                                                      2
##                                                M.Tech, Embedded Systems
##                                                                      2
##                                    M.Tech, Power Electronics And Drives
##                                                                      2
##                                      MBA, Human Resource Management
##                                                                      2
## B.E / B.Tech (CSE - Specialization in Artificial Intelligence and Machine Learning)
##                                                                      1
##                        B.E / B.Tech (Electrical and Electronics Engineering)
##                                                                      1
##                      B.E / B.Tech (Electronics and Computer Engineering)
##                                                                      1
##                                                B.Sc {Hons.}, Economics
##                                                                      1
##                                            B.Sc + M.Sc, Food Technology
##                                                                      1
##                                              B.Sc + M.Sc, Mathematics
##                                                                      1
##                                                  B.Sc + M.Sc, Physics
##                                                                      1
##                        B.Sc(HHA), Hospitality And Hotel Administration
##                                                                      1
##                                              B.Sc, Visual Communication
##                                                                      1
##                                          B.Tech, Artificial Intelligence
##                                                                      1
##                                                  B.Tech, Cyber Security
##                                                                      1
##                                        B.Tech, Manufacturing Engineering
##                                                                      1
##                                B.Tech, Robotics and Artificial Intelligence
##                                                                      1
```

```
##                                                 BBA {Hons.}, General
##                                                                    1
##                                                BBA, Business Analytics
##                                                                    1
##                                              BBALLB {Hons.}, General
##                                                                    1
##                                           Integrated (CSE - Data Science)
##                                                                    1
##                                             M.Sc, Biomedical Genetics
##                                                                    1
##                                                   M.Sc, Statistics
##                                                                    1
##                            M.Tech, Artificial Intelligence & Data Science
##                                                                    1
##                          M.Tech, Artificial Intelligence and Machine Learning
##                                                                    1
##                                           M.Tech, Automotive Engineering
##                                                                    1
##                                            M.Tech, Big Data Analytics
##                                                                    1
##                           M.Tech, Constructions Technology & Management
##                                                                    1
##                                           M.Tech, Control & Automation
##                                                                    1
##                                          M.Tech, Structural Engineering
##                                                                    1
##                                                M.Tech, Vlsi Design
##                                                                    1
##                                            M.Tech, Vlsi System Design
##                                                                    1
##                                              MBA, Business Analytics
##                                                                    1
##                                                    Ph.D, Physics
##                                                                    1
```

```r
# Calculate frequency of each department and sort in descending order
dept_freq <- sort(table(df$Department), decreasing = TRUE)

# Select the top 10 departments
top10_depts <- dept_freq[1:10]

# Increase plot size and decrease text size
par(mai = c(1, 1, 1, 1))  # Adjust margins
par(cex = 0.7)  # Reduce text size

# Create a larger pie chart
pie(top10_depts, main = "Top 10 Departments Distribution", col = rainbow(10), cex.main = 1)
```

Top 10 Departments Distribution
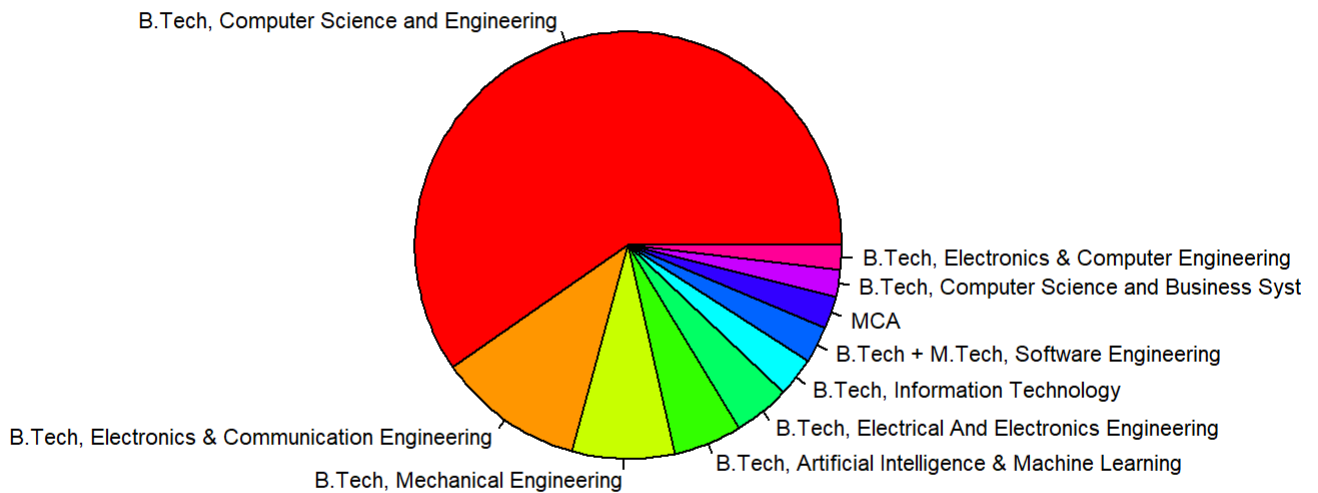

```r
# Basic Statistics for the 'Rating' column
mean_rating <- mean(df$Rating, na.rm = TRUE)  # Mean
median_rating <- median(df$Rating, na.rm = TRUE)  # Median
sd_rating <- sd(df$Rating, na.rm = TRUE)  # Standard Deviation
var_rating <- var(df$Rating, na.rm = TRUE)  # Variance
min_rating <- min(df$Rating, na.rm = TRUE)  # Minimum
max_rating <- max(df$Rating, na.rm = TRUE)  # Maximum
range_rating <- range(df$Rating, na.rm = TRUE)  # Range

# Mode function (since R does not have a built-in mode function)
mode_rating <- as.numeric(names(sort(table(df$Rating), decreasing = TRUE)[1]))

# Print results
cat("Mean:", mean_rating, "\n")
```

```
## Mean: 4.030912
```

```r
cat("Median:", median_rating, "\n")
```

```
## Median: 4.1
```

```r
cat("Mode:", mode_rating, "\n")
```

```
## Mode: 4.3
```

```
cat("Standard Deviation:", sd_rating, "\n")
```

```
## Standard Deviation: 0.6009446
```

```
cat("Variance:", var_rating, "\n")
```

```
## Variance: 0.3611344
```

```
cat("Min:", min_rating, "\n")
```

```
## Min: 0.5
```

```
cat("Max:", max_rating, "\n")
```

```
## Max: 5
```

```
cat("Range:", range_rating, "\n")
```

```
## Range: 0.5 5
```

Structure

```
colnames(df)  # Displays all column names
```

```
## [1] "Name"        "Rating"      "Department"  "Review_Date" "Review_Text"
```

```
head(df)
```

```
##                  Name Rating                              Department
## 1 Mukesh Yashvanth    3.4               B.Sc, Animation & Multimedia
## 2    Saamya Kunhibi    3.6               B.Sc, Animation & Multimedia
## 3       Punya Oswal    4.0 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal    5.0        B.Tech, Computer Science and Engineering
## 5            GDivya    3.3          B.Tech + M.Tech, Software Engineering
## 6       Review guru    3.9        B.Tech, Computer Science and Engineering
##    Review_Date
## 1 Feb 13, 2025
## 2 Feb 13, 2025
## 3 Feb 12, 2025
## 4 Feb 12, 2025
## 5 Feb 12, 2025
## 6 Feb 11, 2025
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

```r
# Convert 'Review_Date' to Date format
df$Review_Date <- as.Date(df$Review_Date, format = "%b %d, %Y")

# Change format to "DD-MM-YYYY"
df$Review_Date <- format(df$Review_Date, "%d-%m-%Y")

# Verify the change
head(df$Review_Date)
```

```
## [1] "13-02-2025" "13-02-2025" "12-02-2025" "12-02-2025" "12-02-2025"
## [6] "11-02-2025"
```

```
head(df)
```

```
##                Name Rating                              Department
## 1 Mukesh Yashvanth    3.4             B.Sc, Animation & Multimedia
## 2   Saamya Kunhibi    3.6             B.Sc, Animation & Multimedia
## 3      Punya Oswal    4.0 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal    5.0     B.Tech, Computer Science and Engineering
## 5           GDivya    3.3       B.Tech + M.Tech, Software Engineering
## 6      Review guru    3.9     B.Tech, Computer Science and Engineering
##   Review_Date
## 1  13-02-2025
## 2  13-02-2025
## 3  12-02-2025
## 4  12-02-2025
## 5  12-02-2025
## 6  11-02-2025
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

Cleaning

```
# Count missing values in each column
colSums(is.na(df))
```

```
##         Name      Rating   Department Review_Date Review_Text
##          253         253          504         268            0
```

```r
df <- na.omit(df)  # Removes all rows with NA values
colSums(is.na(df))
```

```
##         Name      Rating   Department Review_Date Review_Text
##            0           0            0           0            0
```

```r
dim(df)
```

```
## [1] 2750     5
```

```r
df$Rating <- as.numeric(df$Rating)  # Ensure Rating is numeric
df$Department <- as.factor(df$Department)  # Ensure Department is categorical
df$Review_Date <- as.Date(df$Review_Date, format="%d-%m-%Y")  # Ensure correct date format
```

```r
head(df)
```

```
##                Name Rating                               Department
## 1 Mukesh Yashvanth    3.4                  B.Sc, Animation & Multimedia
## 2   Saamya Kunhibi    3.6                  B.Sc, Animation & Multimedia
## 3      Punya Oswal    4.0 B.Tech, Electrical And Electronics Engineering
## 4 Soumyajit Ghosal    5.0         B.Tech, Computer Science and Engineering
## 5           GDivya    3.3          B.Tech + M.Tech, Software Engineering
## 6      Review guru    3.9         B.Tech, Computer Science and Engineering
##   Review_Date
## 1  2025-02-13
## 2  2025-02-13
## 3  2025-02-12
## 4  2025-02-12
## 5  2025-02-12
## 6  2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

```r
# Check for duplicate rows
sum(duplicated(df))
```

```
## [1] 0
```

```r
# Remove duplicate rows
df <- df[!duplicated(df), ]
```

```r
dim(df)
```

```
## [1] 2750    5
```

```
# Check summary statistics
summary(df$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50    3.70    4.10    4.02    4.40    5.00
```

```
# Identify outliers using boxplot
boxplot(df$Rating, main = "Rating Outliers", col = "red")
```
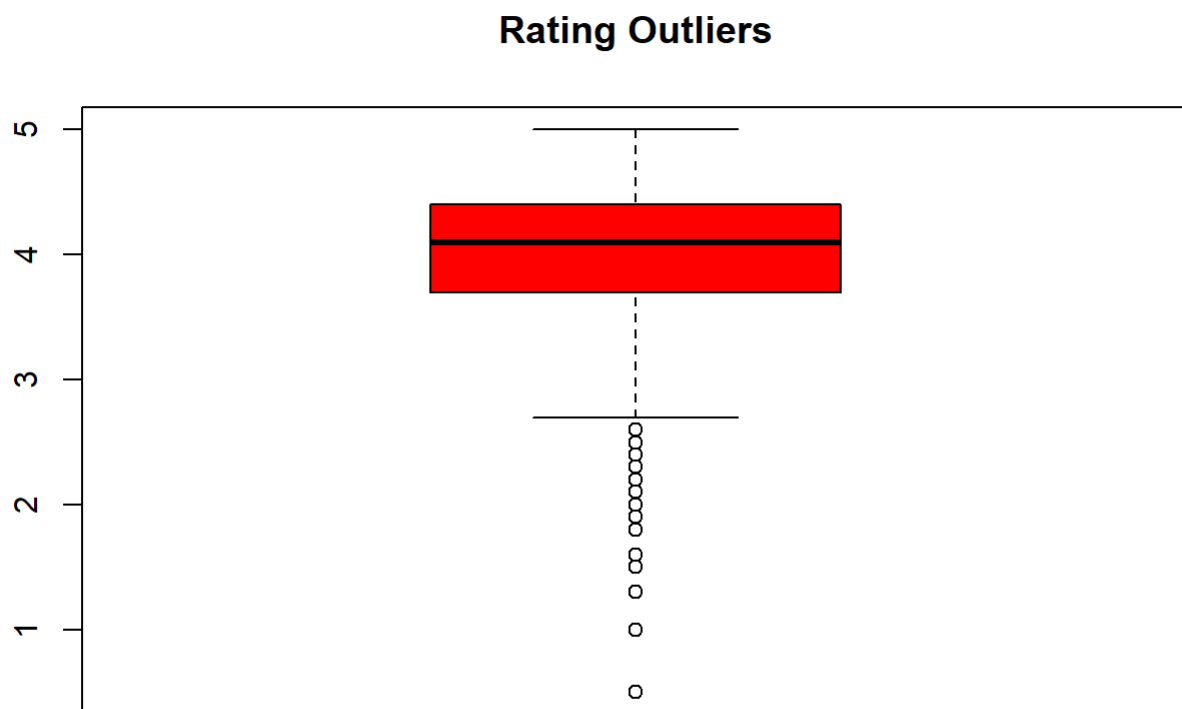
Rating Outliers


```
# Handle outliers (Example: Remove ratings outside 1-5 range)
df <- df[df$Rating >= 1 & df$Rating <= 5, ]
```

```
str(df)  # Check data structure
```

```
## 'data.frame':    2748 obs. of  5 variables:
## $ Name       : chr  "Mukesh Yashvanth" "Saamya Kunhibi" "Punya Oswal" "Soumyajit Ghosal"
...
## $ Rating     : num  3.4 3.6 4 5 3.3 3.9 3.5 3.9 3.6 4 ...
## $ Department : Factor w/ 94 levels "B.Arch","B.Com, Finance",..: 15 15 45 37 24 37 48 22
56 21 ...
## $ Review_Date: Date, format: "2025-02-13" "2025-02-13" ...
## $ Review_Text: chr  "One of my senior got placed in Couch Base for 34LPA CTC another got
placed in Unilever for 42LPA but this was f"| __truncated__ "I don't like how the student and
faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams yearly 3 exams in a "| __truncated_
_ "The university is vast to have best experience  Best faculty crew providing best source of
education  Security "| __truncated__ "Transportation is very bad making it difficult for stud
ents  The infrastructure of the university gives you a p"| __truncated__ ...
## - attr(*, "na.action")= 'omit' Named int [1:518] 2751 2752 2753 2754 2755 2756 2757 2758
2759 2760 ...
##   ..- attr(*, "names")= chr [1:518] "2751" "2752" "2753" "2754" ...
```

#Enrich

1️⃣ Degree & Specialization (Extracted from Department)

- Splitting 'Department' into 'Degree' and 'Specialization' allows for

more granular analysis of academic performance, student feedback, and trends.

- Helps in comparing different degrees (B.Tech, M.Tech) separately.

- Enables analysis of which specializations receive better reviews or ratings.

```
library(tidyr)

# Example: Splitting 'Course' into 'Degree' and 'Specialization'
df <- df %>%
  separate(Department, into = c("Degree", "Specialization"), sep = ", ", extra = "merge", fil
l = "right")

# Check the results
head(df)
```

```
##                   Name Rating              Degree
## 1 Mukesh Yashvanth    3.4                  B.Sc
## 2   Saamya Kunhibi    3.6                  B.Sc
## 3     Punya Oswal     4.0              B.Tech
## 4 Soumyajit Ghosal    5.0              B.Tech
## 5           GDivya    3.3 B.Tech + M.Tech
## 6       Review guru    3.9              B.Tech
##                             Specialization Review_Date
## 1               Animation & Multimedia   2025-02-13
## 2               Animation & Multimedia   2025-02-13
## 3 Electrical And Electronics Engineering  2025-02-12
## 4       Computer Science and Engineering  2025-02-12
## 5                   Software Engineering  2025-02-12
## 6       Computer Science and Engineering  2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
```

```
head(df,10)
```

```
##                     Name Rating             Degree
## 1      Mukesh Yashvanth    3.4               B.Sc
## 2       Saamya Kunhibi    3.6               B.Sc
## 3         Punya Oswal    4.0             B.Tech
## 4     Soumyajit Ghosal    5.0             B.Tech
## 5              GDivya    3.3 B.Tech + M.Tech
## 6         Review guru    3.9             B.Tech
## 7  Srikar Tandulwadikar    3.5             B.Tech
## 8     Mohammed Shahid    3.9 B.Tech + M.Tech
## 9            Avinash    3.6             B.Tech
## 10           jyotinadh    4.0 B.Tech + M.Tech
##                                         Specialization Review_Date
## 1                          Animation & Multimedia  2025-02-13
## 2                          Animation & Multimedia  2025-02-13
## 3          Electrical And Electronics Engineering  2025-02-12
## 4              Computer Science and Engineering  2025-02-12
## 5                        Software Engineering  2025-02-12
## 6              Computer Science and Engineering  2025-02-11
## 7      Electronics And Instrumentation Engineering  2025-02-10
## 8  Computer Science and Engineering + Data Science  2025-02-09
## 9                      Mechanical Engineering  2025-02-09
## 10               Computer Science & Engineering  2025-02-09
##
Review_Text
## 1  One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever
for 42LPA but this was for CSE and related branches while few of my seniors got decent packag
e from mechanical branch too 8-12L in companies like Mercedes a Euler   ~ Anonymous, B.Tech FF
CS. A study who is facing difficulty to study more, can take less number of credits and it wi
ll reduce the burden of the students. And the student can also Complete the entire course in
4 years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc C
ampus Life in Vit is simply magnificent.All types of facilities are provided inside the campu
s including sport complex and restaurants.The girls to boys ratio in Mechanical engineering i
s surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
## 7
The Academics is very good and opportunities for internship is also good  The Mess In the foo
d is ok not good as your home but fine Rules are totally strict you cannot go after 9 outside
the campus
## 8
```

```
I didn't like the rule of freshers cannot go out of campus for first 3 months Faculty was exc
ellent.They taught us very well Teaching methods were excellent.All modern methods were used
## 9
Curriculum was equivalent to IIT standard. There is nothing to dislike.Every thing was good
Yes I can say ragging was prevalent.That is the negetive
## 10
All I can say you need to adjust with the new environment  Flexibility to choose timetable, f
aculty,courses .Due to this flexibility we can opt out desired timings for classes Exam patte
rn that we have 2 cats like mids and one fat like finals
```

```
sum(is.na(df$Review))# Check for empty or unexpected values
```

```
## [1] 0
```

```
unique(df$Review)  # Count NA values
```

```
## NULL
```

- Quantifies review length, helping in sentiment correlation studies.

- Longer reviews might indicate more detailed feedback (either positive or negative).

- Useful for detecting fake reviews (e.g., extremely short reviews may be suspicious).

```
library(stringr)
df$Word_Count <- str_count(df$Review_Text, "\\S+")
```

```
colnames(df)
```

```
## [1] "Name"           "Rating"        "Degree"         "Specialization"
## [5] "Review_Date"    "Review_Text"   "Word_Count"
```

```
head(df$Word_Count)
```

```
## [1] 139  50  22  29  28  25
```

```
head(df)
```

```
##                 Name Rating          Degree
## 1 Mukesh Yashvanth    3.4             B.Sc
## 2   Saamya Kunhibi    3.6             B.Sc
## 3     Punya Oswal    4.0           B.Tech
## 4 Soumyajit Ghosal    5.0           B.Tech
## 5           GDivya    3.3 B.Tech + M.Tech
## 6     Review guru    3.9           B.Tech
##                           Specialization Review_Date
## 1             Animation & Multimedia    2025-02-13
## 2             Animation & Multimedia    2025-02-13
## 3 Electrical And Electronics Engineering  2025-02-12
## 4      Computer Science and Engineering   2025-02-12
## 5                Software Engineering    2025-02-12
## 6      Computer Science and Engineering   2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
##    Word_Count
## 1         139
## 2          50
## 3          22
## 4          29
## 5          28
## 6          25
```

Sentiment Analysis & Scaled Sentiment Score (0-5) # - Sentiment_Score provides a **numerical** way to gauge positive or negative feedback. # - Scaling it between 0-5 aligns it with the existing Rating scale for better comparability. # - While 'Rating' is manually given by users, Sentiment_Score is derived from text, # helping identify discrepancies (e.g., positive text but low rating = inconsistency). # - Enables deeper analysis, such as sentiment trends over time.

```r
library(syuzhet)

# Get sentiment scores for each review
df$Sentiment_Score <- get_sentiment(df$Review_Text, method = "bing")

# Categorize sentiment as Positive, Negative, or Neutral
df$Sentiment_Label <- ifelse(df$Sentiment_Score > 0, "Positive",
                      ifelse(df$Sentiment_Score < 0, "Negative", "Neutral"))

# Check results
table(df$Sentiment_Label)
```

```
##
## Negative  Neutral Positive
##      452      576     1720
```

```r
291+394+1333
```

```
## [1] 2018
```

```r
dim(df)
```

```
## [1] 2748    9
```

```r
head(df)
```

```
##                   Name Rating            Degree
## 1 Mukesh Yashvanth    3.4               B.Sc
## 2   Saamya Kunhibi    3.6               B.Sc
## 3     Punya Oswal     4.0              B.Tech
## 4 Soumyajit Ghosal    5.0              B.Tech
## 5          GDivya     3.3 B.Tech + M.Tech
## 6     Review guru      3.9             B.Tech
##                              Specialization Review_Date
## 1               Animation & Multimedia   2025-02-13
## 2               Animation & Multimedia   2025-02-13
## 3 Electrical And Electronics Engineering   2025-02-12
## 4        Computer Science and Engineering   2025-02-12
## 5                 Software Engineering   2025-02-12
## 6        Computer Science and Engineering   2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
##    Word_Count Sentiment_Score Sentiment_Label
## 1       139               0         Neutral
## 2        50               0         Neutral
## 3        22               0         Neutral
## 4        29               0         Neutral
## 5        28               1        Positive
## 6        25               2        Positive
```

```
range(df$Sentiment_Score)
```

```
## [1] -5 13
```

```r
# Convert columns to numeric (if not already)
df$Sentiment_Score <- as.numeric(df$Sentiment_Score)
df$Rating <- as.numeric(df$Rating)

# Check min and max values, ignoring NA values
sentiment_min <- min(df$Sentiment_Score, na.rm = TRUE)
sentiment_max <- max(df$Sentiment_Score, na.rm = TRUE)

rating_min <- min(df$Rating, na.rm = TRUE)
rating_max <- max(df$Rating, na.rm = TRUE)

# Print results
cat("Sentiment_Score -> Min:", sentiment_min, "| Max:", sentiment_max, "\n")
```

```
## Sentiment_Score -> Min: -5 | Max: 13
```

```r
cat("Rating -> Min:", rating_min, "| Max:", rating_max, "\n")
```

```
## Rating -> Min: 1 | Max: 5
```

```r
# Convert Sentiment_Score to numeric
df$Sentiment_Score <- as.numeric(df$Sentiment_Score)

# Define old and new range
old_min <- -4
old_max <- 11
new_min <- 0
new_max <- 5

# Apply Min-Max Scaling and round to 1 decimal place
df$Scaled_Sentiment_Score <- round(((df$Sentiment_Score - old_min) /
                                    (old_max - old_min)) *
                                    (new_max - new_min) + new_min, 1)

# Check new range
range(df$Scaled_Sentiment_Score, na.rm = TRUE)
```

```
## [1] -0.3  5.7
```

```r
# View results
head(df[c("Sentiment_Score", "Scaled_Sentiment_Score")])
```

```
##    Sentiment_Score Scaled_Sentiment_Score
## 1                0                    1.3
## 2                0                    1.3
## 3                0                    1.3
## 4                0                    1.3
## 5                1                    1.7
## 6                2                    2.0
```

```r
df$Sentiment_Score <- NULL

# Rename Scaled_Sentiment_Score to Sentiment_Score
names(df)[names(df) == "Scaled_Sentiment_Score"] <- "Sentiment_Score"

# View updated dataframe
head(df)
```

```
##                   Name Rating              Degree
## 1 Mukesh Yashvanth    3.4                   B.Sc
## 2   Saamya Kunhibi    3.6                   B.Sc
## 3      Punya Oswal    4.0                 B.Tech
## 4 Soumyajit Ghosal    5.0                 B.Tech
## 5           GDivya    3.3 B.Tech + M.Tech
## 6      Review guru    3.9                 B.Tech
##                              Specialization Review_Date
## 1                    Animation & Multimedia  2025-02-13
## 2                    Animation & Multimedia  2025-02-13
## 3 Electrical And Electronics Engineering    2025-02-12
## 4        Computer Science and Engineering    2025-02-12
## 5                    Software Engineering    2025-02-12
## 6        Computer Science and Engineering    2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
##   Word_Count Sentiment_Label Sentiment_Score
## 1        139         Neutral             1.3
## 2         50         Neutral             1.3
## 3         22         Neutral             1.3
## 4         29         Neutral             1.3
## 5         28         Positive            1.7
## 6         25         Positive            2.0
```

```
# Define intervals (bins)
df$Score_Interval <- cut(df$Sentiment_Score,
                         breaks = seq(0, 5, by = 1),
                         include.lowest = TRUE,
                         right = FALSE)

# Count occurrences in each interval
table(df$Score_Interval)
```

```
##
## [0,1) [1,2) [2,3) [3,4) [4,5]
##   141  1490   984   119     9
```

#Validation

```
#  🛠 **1. Check Data Structure**
str(df)  # Verify column types and structure
```

```
## 'data.frame':    2748 obs. of  10 variables:
##  $ Name          : chr  "Mukesh Yashvanth" "Saamya Kunhibi" "Punya Oswal" "Soumyajit Ghos
al" ...
##  $ Rating        : num  3.4 3.6 4 5 3.3 3.9 3.5 3.9 3.6 4 ...
##  $ Degree        : chr  "B.Sc" "B.Sc" "B.Tech" "B.Tech" ...
##  $ Specialization : chr  "Animation & Multimedia" "Animation & Multimedia" "Electrical And
Electronics Engineering" "Computer Science and Engineering" ...
##  $ Review_Date   : Date, format: "2025-02-13" "2025-02-13" ...
##  $ Review_Text   : chr  "One of my senior got placed in Couch Base for 34LPA CTC another
got placed in Unilever for 42LPA but this was f"| __truncated__ "I don't like how the student
and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams yearly 3 exams in a "| __trunca
ted__ "The university is vast to have best experience  Best faculty crew providing best sourc
e of education  Security "| __truncated__ "Transportation is very bad making it difficult for
students  The infrastructure of the university gives you a p"| __truncated__ ...
##  $ Word_Count    : int  139 50 22 29 28 25 38 31 24 43 ...
##  $ Sentiment_Label: chr  "Neutral" "Neutral" "Neutral" "Neutral" ...
##  $ Sentiment_Score: num  1.3 1.3 1.3 1.3 1.7 2 1.3 2.7 1.3 1.7 ...
##  $ Score_Interval : Factor w/ 5 levels "[0,1)","[1,2)",..: 2 2 2 2 2 3 2 3 2 2 ...
```

```
#  🛠 **2. Summary Statistics**
summary(df)  # Provides mean, median, min, max, and quartiles for numerical columns
```

```
##      Name              Rating          Degree          Specialization
## Length:2748        Min.   :1.000   Length:2748     Length:2748
## Class :character   1st Qu.:3.700   Class :character  Class :character
## Mode  :character   Median :4.100   Mode  :character  Mode  :character
##                    Mean   :4.022
##                    3rd Qu.:4.400
##                    Max.   :5.000
##   Review_Date          Review_Text        Word_Count      Sentiment_Label
## Min.   :2020-02-27   Length:2748        Min.   : 13.00   Length:2748
## 1st Qu.:2021-08-13   Class :character   1st Qu.: 30.00   Class :character
## Median :2023-08-20   Mode  :character   Median : 38.00   Mode  :character
## Mean   :2023-01-04                      Mean   : 46.35
## 3rd Qu.:2024-03-10                      3rd Qu.: 52.00
## Max.   :2025-02-13                      Max.   :149.00
## Sentiment_Score   Score_Interval
## Min.   :-0.300    [0,1): 141
## 1st Qu.: 1.300    [1,2):1490
## Median : 1.700    [2,3): 984
## Mean   : 1.732    [3,4): 119
## 3rd Qu.: 2.000    [4,5):   9
## Max.   : 5.700    NA's :   5
```

```r
# 🛠️ **3. Check for Missing Values**
colSums(is.na(df))  # Count of missing values in each column
```

```
##         Name          Rating          Degree    Specialization     Review_Date
##            0               0               0              64               0
##  Review_Text      Word_Count Sentiment_Label Sentiment_Score  Score_Interval
##            0               0               0               0               5
```

```r
head(df)
```

```
##             Name Rating        Degree
## 1 Mukesh Yashvanth   3.4           B.Sc
## 2   Saamya Kunhibi   3.6           B.Sc
## 3     Punya Oswal    4.0         B.Tech
## 4 Soumyajit Ghosal   5.0         B.Tech
## 5          GDivya    3.3 B.Tech + M.Tech
## 6     Review guru    3.9         B.Tech
##                            Specialization Review_Date
## 1                  Animation & Multimedia  2025-02-13
## 2                  Animation & Multimedia  2025-02-13
## 3 Electrical And Electronics Engineering  2025-02-12
## 4      Computer Science and Engineering  2025-02-12
## 5                 Software Engineering  2025-02-12
## 6      Computer Science and Engineering  2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
##   Word_Count Sentiment_Label Sentiment_Score Score_Interval
## 1        139         Neutral             1.3          [1,2)
## 2         50         Neutral             1.3          [1,2)
## 3         22         Neutral             1.3          [1,2)
## 4         29         Neutral             1.3          [1,2)
## 5         28        Positive             1.7          [1,2)
## 6         25        Positive             2.0          [2,3)
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Safe function to compute mode (returns NA if no valid values exist)
get_mode <- function(x) {
  x <- na.omit(x)  # Remove NAs
  if (length(x) == 0) return(NA)  # Return NA if no non-NA values exist
  tab <- table(x)  # Count occurrences
  names(tab)[which.max(tab)]  # Return most frequent value
}
```

```r
# Find mode specialization for each degree
mode_specialization <- df %>%
  group_by(Degree) %>%
  summarise(Mode_Specialization = get_mode(Specialization)) %>%
  ungroup()  # Ungroup after summarizing

print(mode_specialization)  # Check the mode values
```

```
## # A tibble: 19 × 2
##    Degree          Mode_Specialization
##    <chr>           <chr>
##  1 B.Arch          <NA>
##  2 B.Com           General
##  3 B.Des           Industrial Design
##  4 B.Sc            Computer Science
##  5 B.Sc + M.Sc     Biotechnology
##  6 B.Sc {Hons.}    Agriculture
##  7 B.Sc(HHA)       Hospitality And Hotel Administration
##  8 B.Tech          Computer Science and Engineering
##  9 B.Tech + M.Tech Software Engineering
## 10 BALLB {Hons.}   <NA>
## 11 BBA             General
## 12 BBA {Hons.}     General
## 13 BBALLB {Hons.}  General
## 14 BCA             General
## 15 M.Sc            Data Science
## 16 M.Tech          Computer Science And Engineering
## 17 MBA             Finance
## 18 MCA             <NA>
## 19 Ph.D            Physics
```

```r
# Merge mode specializations into the main dataset
df <- df %>%
  left_join(mode_specialization, by = "Degree") %>%
  mutate(Specialization = ifelse(is.na(Specialization), Mode_Specialization, Specialization))
%>%
  select(-Mode_Specialization)  # Remove helper column

# Verify if any NA values remain
sum(is.na(df$Specialization))  # Should be 0 now
```

```
## [1] 62
```

```r
# View all rows where Specialization is still NA
df %>% filter(is.na(Specialization)) %>% select(Degree)
```

```
##            Degree
## 1              MCA
## 2    BALLB {Hons.}
## 3    BALLB {Hons.}
## 4              MCA
## 5    BALLB {Hons.}
## 6              MCA
## 7              MCA
## 8              MCA
## 9              MCA
## 10             MCA
## 11             MCA
## 12             MCA
## 13             MCA
## 14             MCA
## 15             MCA
## 16             MCA
## 17             MCA
## 18             MCA
## 19             MCA
## 20             MCA
## 21             MCA
## 22             MCA
## 23             MCA
## 24   BALLB {Hons.}
## 25             MCA
## 26             MCA
## 27             MCA
## 28             MCA
## 29             MCA
## 30             MCA
## 31             MCA
## 32             MCA
## 33             MCA
## 34             MCA
## 35             MCA
## 36             MCA
## 37             MCA
## 38             MCA
## 39             MCA
## 40             MCA
## 41          B.Arch
## 42   BALLB {Hons.}
## 43             MCA
## 44          B.Arch
## 45             MCA
## 46             MCA
## 47             MCA
## 48             MCA
## 49             MCA
## 50   BALLB {Hons.}
## 51             MCA
## 52             MCA
## 53             MCA
## 54             MCA
```

```
## 55              MCA
## 56              MCA
## 57              MCA
## 58              MCA
## 59              MCA
## 60              MCA
## 61              MCA
## 62              MCA
```

```r
df <- df %>%
  mutate(Specialization = ifelse(is.na(Specialization), "Core", Specialization))

# Verify if any NA values remain
sum(is.na(df$Specialization))  # Should be 0 now
```

```
## [1] 0
```

```r
colSums(is.na(df))
```

```
##             Name           Rating           Degree   Specialization      Review_Date
##                0                0                0                0                0
##      Review_Text       Word_Count  Sentiment_Label  Sentiment_Score   Score_Interval
##                0                0                0                0                5
```

```r
nrow(df) - nrow(unique(df)) # Number of duplicate rows
```

```
## [1] 0
```

```r
# Validate Data Ranges**
range(df$Rating, na.rm = TRUE)  # Should be within expected limits (e.g., 1-5)
```

```
## [1] 1 5
```

```r
range(df$Sentiment_Score, na.rm = TRUE)  # Should be between 0 and 5
```

```
## [1] -0.3  5.7
```

```r
unique(df$Sentiment_Label)
```

```
## [1] "Neutral"  "Positive" "Negative"
```

```r
colnames(df)
```

```
##  [1] "Name"            "Rating"          "Degree"          "Specialization"
##  [5] "Review_Date"     "Review_Text"     "Word_Count"      "Sentiment_Label"
##  [9] "Sentiment_Score" "Score_Interval"
```

```
head(df)
```

```
##                   Name Rating         Degree
## 1 Mukesh Yashvanth    3.4            B.Sc
## 2   Saamya Kunhibi    3.6            B.Sc
## 3      Punya Oswal    4.0          B.Tech
## 4 Soumyajit Ghosal    5.0          B.Tech
## 5           GDivya    3.3 B.Tech + M.Tech
## 6      Review guru    3.9          B.Tech
##                              Specialization Review_Date
## 1                Animation & Multimedia  2025-02-13
## 2                Animation & Multimedia  2025-02-13
## 3 Electrical And Electronics Engineering  2025-02-12
## 4        Computer Science and Engineering  2025-02-12
## 5                   Software Engineering  2025-02-12
## 6        Computer Science and Engineering  2025-02-11
##
Review_Text
## 1 One of my senior got placed in Couch Base for 34LPA CTC another got placed in Unilever f
or 42LPA but this was for CSE and related branches while few of my seniors got decent package
from mechanical branch too 8-12L in companies like Mercedes a Euler  ~ Anonymous, B.Tech FFC
S. A study who is facing difficulty to study more, can take less number of credits and it wil
l reduce the burden of the students. And the student can also Complete the entire course in 4
years instead of 5years  by taking maximum no of credits.  ~ Supriya R Patil, B.Sc + M.Sc Cam
pus Life in Vit is simply magnificent.All types of facilities are provided inside the campus
including sport complex and restaurants.The girls to boys ratio in Mechanical engineering is
surprisinly high of about 1:7.  ~ Mithilesh Angal, B.Tech
## 2
I don't like how the student and faculty ratio is 60-40  ~ Praveen Kumar, BCA Lot of exams ye
arly 3 exams in a semester 2-3 quiz and 1-2 DA  ~ Kartik Navnath Narare, B.Tech Water is real
ly bad. 750+ TDS. It makes you go bald. Really.  ~ Piyush Prajapati, B.Des
## 3
The university is vast to have best experience  Best faculty crew providing best source of ed
ucation  Security guards of universities are bad.
## 4
Transportation is very bad making it difficult for students  The infrastructure of the univer
sity gives you a perfect learning environment  The faculties are highly qualified to train th
e students
## 5
The library is the best part out of the university  Too many rules and regulations to be foll
owed  Arrear fees is 6000 per subject which is too much
## 6
Attendance minimum criteria is seventify percentage  The College has many hostel buildings an
d enough space  The nature of college and greenery is the best part
##   Word_Count Sentiment_Label Sentiment_Score Score_Interval
## 1        139         Neutral            1.3        [1,2)
## 2         50         Neutral            1.3        [1,2)
## 3         22         Neutral            1.3        [1,2)
## 4         29         Neutral            1.3        [1,2)
## 5         28        Positive            1.7        [1,2)
## 6         25        Positive            2.0        [2,3)
```

Stats measure

```r
# Select only numerical columns
num_cols <- c("Rating", "Word_Count", "Sentiment_Score")

# Compute statistics
stats <- data.frame(
  Metric = c("Min", "Max", "Range", "Mean", "Median", "Std_Dev"),
  Rating = c(min(df$Rating, na.rm = TRUE),
             max(df$Rating, na.rm = TRUE),
             diff(range(df$Rating, na.rm = TRUE)),
             mean(df$Rating, na.rm = TRUE),
             median(df$Rating, na.rm = TRUE),
             sd(df$Rating, na.rm = TRUE)),

  Word_Count = c(min(df$Word_Count, na.rm = TRUE),
                 max(df$Word_Count, na.rm = TRUE),
                 diff(range(df$Word_Count, na.rm = TRUE)),
                 mean(df$Word_Count, na.rm = TRUE),
                 median(df$Word_Count, na.rm = TRUE),
                 sd(df$Word_Count, na.rm = TRUE)),

  Sentiment_Score = c(min(df$Sentiment_Score, na.rm = TRUE),
                      max(df$Sentiment_Score, na.rm = TRUE),
                      diff(range(df$Sentiment_Score, na.rm = TRUE)),
                      mean(df$Sentiment_Score, na.rm = TRUE),
                      median(df$Sentiment_Score, na.rm = TRUE),
                      sd(df$Sentiment_Score, na.rm = TRUE))
)

# Print statistics table
print(stats)
```

```
##     Metric    Rating Word_Count Sentiment_Score
## 1     Min  1.0000000   13.00000      -0.3000000
## 2     Max  5.0000000  149.00000       5.7000000
## 3   Range  4.0000000  136.00000       6.0000000
## 4    Mean  4.0222707   46.35153       1.7318777
## 5  Median  4.1000000   38.00000       1.7000000
## 6 Std_Dev  0.5959568   25.34854       0.6353463
```

```r
# Remove the Score_Interval column
df <- df %>% select(-Score_Interval)

# Verify the column is removed
colnames(df)
```

```
## [1] "Name"            "Rating"        "Degree"       "Specialization"
## [5] "Review_Date"     "Review_Text"   "Word_Count"   "Sentiment_Label"
## [9] "Sentiment_Score"
```

```r
colSums(is.na(df))  # Should return 0 for all critical columns
```

```
##          Name        Rating        Degree Specialization      Review_Date
##             0             0             0             0                0
##   Review_Text    Word_Count Sentiment_Label Sentiment_Score
##             0             0             0             0
```

```
nrow(df)
```

```
## [1] 2748
```

```
write.csv(df, "D:/Prog for DS/FINALVIT_Reviews_Students_Data2.csv", row.names = FALSE)
```