Untitled

2025-04-17

Introduzione

Librerie

```
library(ggplot2)
library(scales)  # per formattare le percentuali
```

Import del dataset e analisi preliminare

```
ds <- read.csv("StudentPerformanceFactors.csv")


# Lista di variabili categoriali
categorical_vars <- c(
  "Parental_Involvement", "Access_to_Resources", "Extracurricular_Activities",
  "Motivation_Level", "Internet_Access", "Family_Income", "Teacher_Quality",
  "School_Type", "Peer_Influence", "Learning_Disabilities",
  "Parental_Education_Level", "Distance_from_Home", "Gender"
)

ds[categorical_vars] <- lapply(ds[categorical_vars], factor)

head(ds)
```

```
##   Hours_Studied Attendance Parental_Involvement Access_to_Resources
## 1            23         84                  Low                High
## 2            19         64                  Low              Medium
## 3            24         98               Medium              Medium
## 4            29         89                  Low              Medium
## 5            19         92               Medium              Medium
## 6            19         88               Medium              Medium
##   Extracurricular_Activities Sleep_Hours Previous_Scores Motivation_Level
## 1                         No           7              73              Low
## 2                         No           8              59              Low
## 3                        Yes           7              91           Medium
## 4                        Yes           8              98           Medium
## 5                        Yes           6              65           Medium
## 6                        Yes           8              89           Medium
##   Internet_Access Tutoring_Sessions Family_Income Teacher_Quality School_Type
## 1             Yes                 0           Low          Medium      Public
## 2             Yes                 2        Medium          Medium      Public
## 3             Yes                 2        Medium          Medium      Public
## 4             Yes                 1        Medium          Medium      Public
## 5             Yes                 3        Medium            High      Public
```

```
## 6             Yes                 3        Medium          Medium      Public
##   Peer_Influence Physical_Activity Learning_Disabilities
## 1       Positive                 3                    No
## 2       Negative                 4                    No
## 3        Neutral                 4                    No
## 4       Negative                 4                    No
## 5        Neutral                 4                    No
## 6       Positive                 3                    No
##   Parental_Education_Level Distance_from_Home Gender Exam_Score
## 1              High School               Near   Male         67
## 2                  College           Moderate Female         61
## 3             Postgraduate               Near   Male         74
## 4              High School           Moderate   Male         71
## 5                  College               Near Female         70
## 6             Postgraduate               Near   Male         71
```

```
colnames(ds)
```

Descrizione delle variabili

```
##  [1] "Hours_Studied"              "Attendance"
##  [3] "Parental_Involvement"       "Access_to_Resources"
##  [5] "Extracurricular_Activities" "Sleep_Hours"
##  [7] "Previous_Scores"            "Motivation_Level"
##  [9] "Internet_Access"            "Tutoring_Sessions"
## [11] "Family_Income"              "Teacher_Quality"
## [13] "School_Type"                "Peer_Influence"
## [15] "Physical_Activity"          "Learning_Disabilities"
## [17] "Parental_Education_Level"   "Distance_from_Home"
## [19] "Gender"                     "Exam_Score"
```

- **Hours_Studied** Numero di ore spese studiando a settimana.
- **Attendance** Percentuale di lezioni frequentate.
- **Parental_Involvement** Livello di coinvolgimento genitoriale nella formazione dello studente (Low, Medium, High).
- **Access_to_Resources** Disponibilità di risorse educative(Low, Medium, High).
- **Extracurricular_Activities** Partecipazione ad attività extracurriculari (Yes, No).
- **Sleep_Hours** Numero medio di ore di sonno a notte.
- **Previous_Scores** Punteggio degli esami precedenti.
- **Motivation_Level** Livello di motivazione dello studente (Low, Medium, High).
- **Internet_Access** Disponibilità di accesso ad Internet (Yes, No).
- **Tutoring_Sessions** Numero di sessioni di tutoraggio frequentata al mese.
- **Family_Income** Livello di reddito familiare (Low, Medium, High).
- **Teacher_Quality** Qualità dell'insegnamento (Low, Medium, High).
- **School_Type** Tipo di scuola frequentata (Public, Private).
- **Peer_Influence** Influenza dei pari sulla performance accademica (Positive, Neutral, Negative).
- **Physical_Activity** Numero medio di ore di attività fisica a settimana.
- **Learning_Disabilities** Presenza di difficoltà di apprendimento (Yes, No).
- **Parental_Education_Level** Livello più alto di educazione dei genitori (High School, College, Postgraduate).
- **Distance_from_Home** Distanza da casa a scuola (Near, Moderate, Far).
- **Gender** Genere dello studente (Male, Female).
- **Exam_Score** Punteggio dell' esame finale.

```r
summary(ds)
```

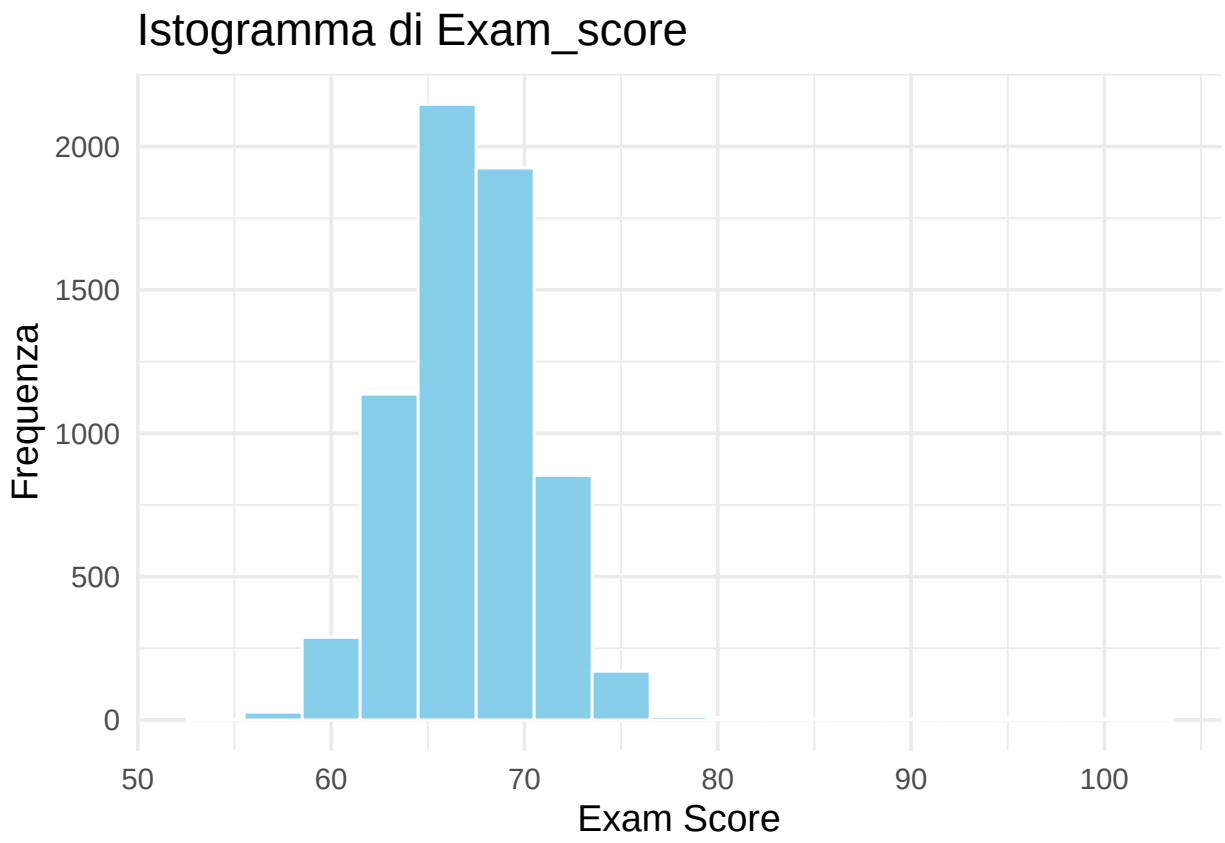
```
##  Hours_Studied     Attendance     Parental_Involvement Access_to_Resources
##  Min.   : 1.00   Min.   : 60.00   High  :1908          High  :1975
##  1st Qu.:16.00   1st Qu.: 70.00   Low   :1337          Low   :1313
##  Median :20.00   Median : 80.00   Medium:3362          Medium:3319
##  Mean   :19.98   Mean   : 79.98
##  3rd Qu.:24.00   3rd Qu.: 90.00
##  Max.   :44.00   Max.   :100.00
##  Extracurricular_Activities  Sleep_Hours     Previous_Scores  Motivation_Level
##  No :2669                   Min.   : 4.000   Min.   : 50.00   High  :1319
##  Yes:3938                   1st Qu.: 6.000   1st Qu.: 63.00   Low   :1937
##                             Median : 7.000   Median : 75.00   Medium:3351
##                             Mean   : 7.029   Mean   : 75.07
##                             3rd Qu.: 8.000   3rd Qu.: 88.00
##                             Max.   :10.000   Max.   :100.00
##  Internet_Access Tutoring_Sessions Family_Income Teacher_Quality  School_Type
##  No : 499        Min.   :0.000     High  :1269         :  78     Private:2009
##  Yes:6108        1st Qu.:1.000     Low   :2672   High  :1947     Public :4598
##                  Median :1.000     Medium:2666   Low   : 657
##                  Mean   :1.494                   Medium:3925
##                  3rd Qu.:2.000
##                  Max.   :8.000
##   Peer_Influence Physical_Activity Learning_Disabilities
##  Negative:1377   Min.   :0.000     No :5912
##  Neutral :2592   1st Qu.:2.000     Yes: 695
##  Positive:2638   Median :3.000
##                  Mean   :2.968
##                  3rd Qu.:4.000
##                  Max.   :6.000
##  Parental_Education_Level Distance_from_Home    Gender       Exam_Score
##              :  90                :  67      Female:2793   Min.   : 55.00
##  College     :1989        Far     : 658      Male  :3814   1st Qu.: 65.00
##  High School :3223        Moderate:1998                    Median : 67.00
##  Postgraduate:1305        Near    :3884                    Mean   : 67.24
##                                                            3rd Qu.: 69.00
##                                                            Max.   :101.00
```

```r
str(ds)
```

```
## 'data.frame':    6607 obs. of  20 variables:
##  $ Hours_Studied             : int  23 19 24 29 19 19 29 25 17 23 ...
##  $ Attendance                : int  84 64 98 89 92 88 84 78 94 98 ...
##  $ Parental_Involvement      : Factor w/ 3 levels "High","Low","Medium": 2 2 3 2 3 3 3 2 3 3 ...
##  $ Access_to_Resources       : Factor w/ 3 levels "High","Low","Medium": 1 3 3 3 3 3 2 1 1 3 ...
##  $ Extracurricular_Activities: Factor w/ 2 levels "No","Yes": 1 1 2 2 2 2 2 2 1 2 ...
##  $ Sleep_Hours               : int  7 8 7 8 6 8 7 6 6 8 ...
##  $ Previous_Scores           : int  73 59 91 98 65 89 68 50 80 71 ...
##  $ Motivation_Level          : Factor w/ 3 levels "High","Low","Medium": 2 2 3 3 3 3 2 3 1 3 ...
##  $ Internet_Access           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Tutoring_Sessions         : int  0 2 2 1 3 3 1 1 0 0 ...
##  $ Family_Income             : Factor w/ 3 levels "High","Low","Medium": 2 3 3 3 3 3 2 1 3 1 ...
##  $ Teacher_Quality           : Factor w/ 4 levels "","High","Low",..: 4 4 4 4 2 4 4 2 3 2 ...
##  $ School_Type               : Factor w/ 2 levels "Private","Public": 2 2 2 2 2 2 1 2 1 2 ...
```

```
##  $ Peer_Influence            : Factor w/ 3 levels "Negative","Neutral",..: 3 1 2 1 2 3 2 1 2 3 ...
##  $ Physical_Activity         : int  3 4 4 4 4 3 2 2 1 5 ...
##  $ Learning_Disabilities     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Parental_Education_Level  : Factor w/ 4 levels "","College","High School",..: 3 2 4 3 2 4 3 3 2 3
##  $ Distance_from_Home        : Factor w/ 4 levels "","Far","Moderate",..: 4 3 4 3 4 4 3 2 4 3 ...
##  $ Gender                    : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 2 2 2 2 2 ...
##  $ Exam_Score                : int  67 61 74 71 70 71 67 66 69 72 ...
```

```r
ggplot(ds, aes(x = Exam_Score)) +
  geom_histogram(
    binwidth = 3,               # larghezza del bin; modificala a seconda della granularità desiderata
    fill     = "skyblue",       # colore interno delle barre
    color    = "white"          # colore del bordo delle barre
  ) +
  labs(
    x     = "Exam Score",
    y     = "Frequenza",
    title = "Istogramma di Exam_score"
  ) +
  theme_minimal(base_size = 14)
```



Trasformiamo la variabile Exam_Score in un variabile categorica.

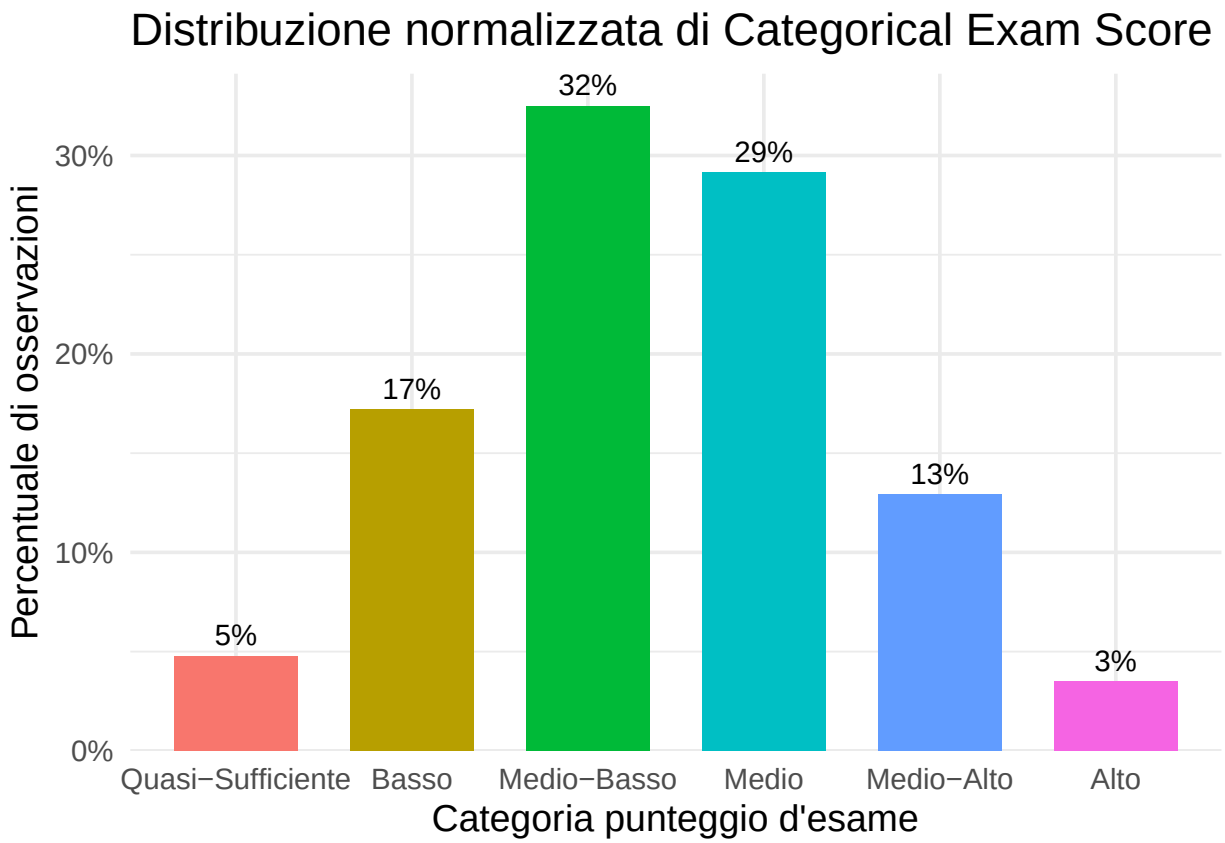
```r
ds$Categorical_Exam_Score <- cut(
  ds$Exam_Score,
  breaks = c(54, 61, 64, 67, 70, 73, 102),
  labels = c("Quasi-Sufficiente", "Basso", "Medio-Basso", "Medio", "Medio-Alto", "Alto"),
```

```
  include.lowest = FALSE,
  right = TRUE
)

table(ds$Categorical_Exam_Score)/length(ds$Categorical_Exam_Score)*100
```

```
##
## Quasi-Sufficiente             Basso       Medio-Basso             Medio
##          4.782806         17.193885         32.495838         29.135765
##        Medio-Alto              Alto
##         12.910549          3.481156
```

```
ggplot(ds, aes(x = Categorical_Exam_Score,
               fill = Categorical_Exam_Score)) +
  # barre con proporzione
  geom_bar(
    aes(y = after_stat(count) / sum(after_stat(count))),
    stat = "count",
    width = 0.7,
    show.legend = FALSE
  ) +
  # percentuali sopra le barre
  geom_text(
    aes(
      label = percent(after_stat(count) / sum(after_stat(count)), accuracy = 1),
      y     = after_stat(count) / sum(after_stat(count))
    ),
    stat = "count",
    vjust = -0.5
  ) +
  # scala y in percentuale e un po' di spazio in alto
  scale_y_continuous(
    labels = percent_format(accuracy = 1),
    expand = expansion(mult = c(0, 0.05))
  ) +
  labs(
    x     = "Categoria punteggio d'esame",
    y     = "Percentuale di osservazioni",
    title = "Distribuzione normalizzata di Categorical Exam Score"
  ) +
  theme_minimal(base_size = 14)
```



Analisi

Conclusioni