Untitled2

December 29, 2024

[7]:
```
#import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#step-1;load the data set
url="https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.
 ↪zip"
dataset_path="student-mat.csv"

#load the dataset
import urllib.request
import zipfile

#download the dataset
urllib.request.urlretrieve(url,"student.zip")

#extract the data
with zipfile.ZipFile("student.zip","r") as zip_ref:
    zip_ref.extractall(".")

#load the data into data frame
data = pd.read_csv("student-mat.csv", sep=";")
print("data loaded successfully")

#step:data exploration
print(data.head()) #display the first few rows
print("\ndataset info:")
print(data.info()) #checks data types and missing values

#step-3:data cleaning
#check for missing values
print("\nMissing values:")
print(data.isnull().sum())

#remove duplicates
data = data.drop_duplicates()
```

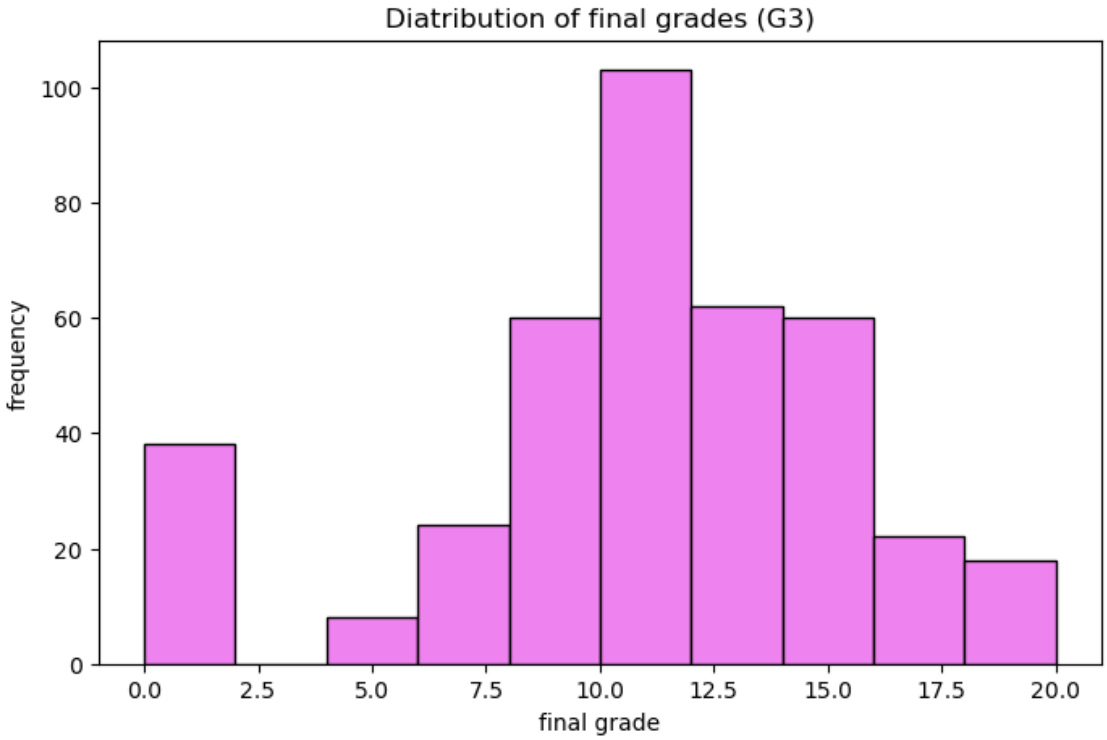
```
#step-4: data analysis
#question 1: what is the average score in math(G3)?
average_score = data['G3'].mean()
print(f"\nAverage math score (G3) : {average_score:.2f}")

#question 2: how many students scored above 15 in their final grade (G3)?
students_above_15 = len(data[data['G3'] > 15])
print(f"Number of students scoring above 15: {students_above_15}")

#question 3: is there any correlation between study time and final grade?
correlation = data['studytime'].corr(data['G3'])
print(f"correlation between study time and final grade: {correlation:.2f}")

#question 4: which gender has a higher average final grade?
average_grade_by_gender = data.groupby('sex')['G3'].mean()
print("\nAverage final grade by gender:")
print(average_grade_by_gender)

#step 5 : data visualization
#histogram of final grades
plt.figure(figsize=(8,5))
plt.hist(data['G3'],bins=10, color='violet', edgecolor='black')
plt.title("Diatribution of final grades (G3)")
plt.xlabel("final grade")
plt.ylabel("frequency")
plt.show()

#scatter plot of studytime vs final grade
plt.figure(figsize=(8,5))
sns.scatterplot(data=data,x='studytime',y='G3',hue='sex')
plt.title("study time vs final grade")
plt.xlabel("study time (hours)")
plt.ylabel("final grade")
plt.legend(title="Gender")
plt.show()

#bar chart of average scores by gender
plt.figure(figsize=(8,5))
average_grade_by_gender.plot(kind='bar',color=['red','green'])
plt.title("Average final grade by gender")
plt.ylabel("Average final grade")
plt.xlabel("Average final grade")
plt.xticks(rotation=0)
plt.show()
```

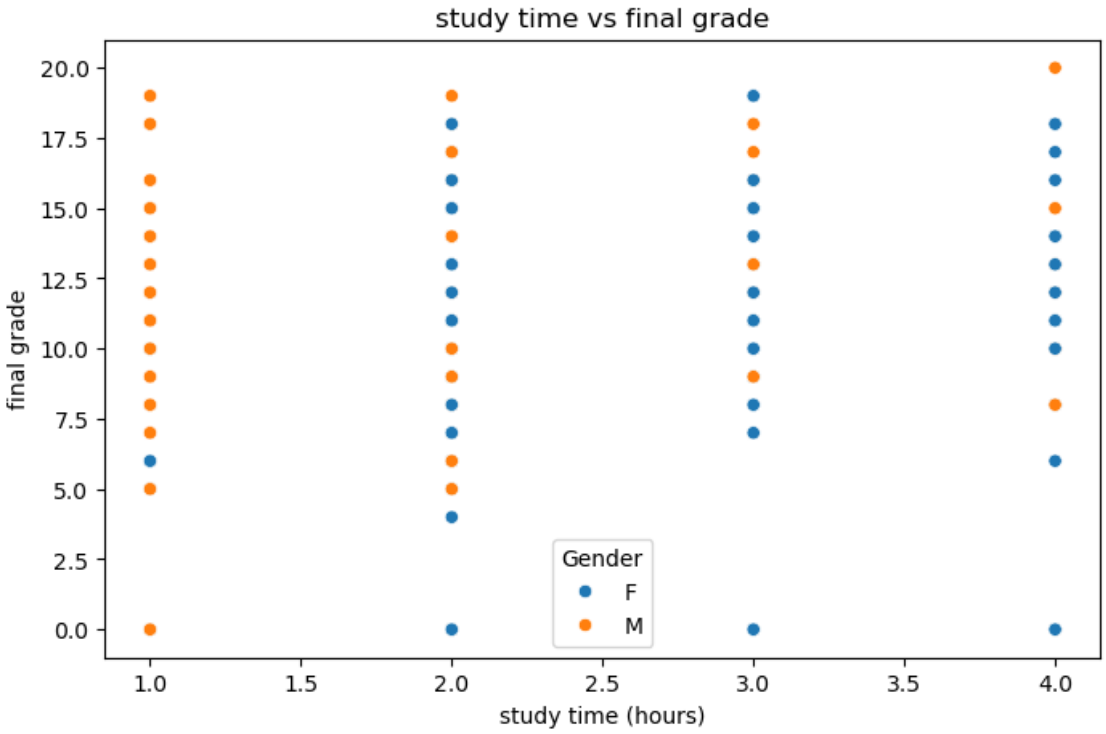
```
data loaded successfully
```

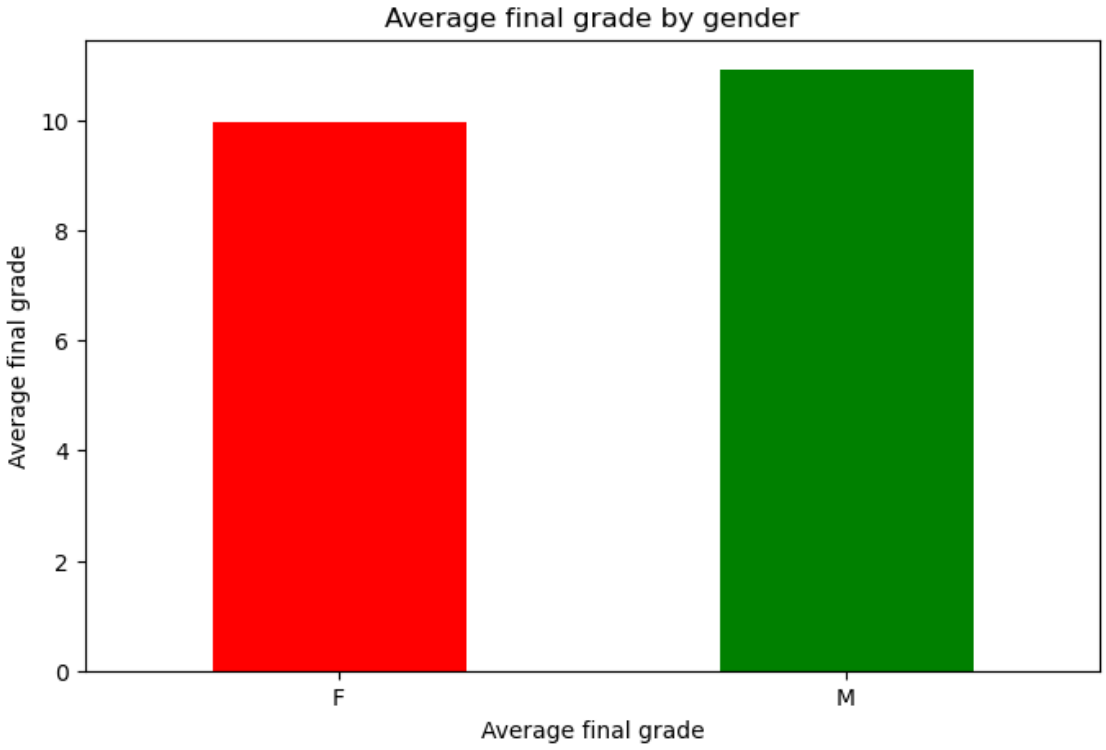
```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  …  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher  …
1     GP   F   17       U     GT3       T     1     1  at_home     other  …
2     GP   F   15       U     LE3       T     1     1  at_home     other  …
3     GP   F   15       U     GT3       T     4     2   health  services  …
4     GP   F   16       U     GT3       T     3     3    other     other  …

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        6   5   6   6
1      5        3      3     1     1      3        4   5   5   6
2      4        3      2     2     3      3       10   7   8  10
3      3        2      2     1     1      5        2  15  14  15
4      4        3      2     1     2      5        4   6  10  10

[5 rows x 33 columns]

dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
```

```
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
None

Missing values:
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64

Average math score (G3) : 10.42
```

```
Number of students scoring above 15: 40
correlation between study time and final grade: 0.10

Average final grade by gender:
sex
F     9.966346
M    10.914439
Name: G3, dtype: float64
```




study time vs final grade
final grade
study time (hours)
Gender
F
M
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Average final grade by gender
Average final grade
Average final grade
F
M
10
8
6
4
2
0

[ ]: