Data Analytics with Python

```
import pandas as pd
import numpy as np

#to read the file
df = pd.read_csv('student-mat.csv')

#df = pd.read_csv("C:/Users/Viswanathan/Desktop/new/New
folder/student-mat.csv")

#to display the rows
df.head()
```

```
  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]
```

```
#to check the missing values
df.isnull().sum()
```

```
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
```

```
#to display the column data
df.dtypes
```

```
school        object
sex           object
age            int64
address       object
famsize       object
Pstatus       object
Medu           int64
Fedu           int64
Mjob          object
Fjob          object
reason        object
guardian      object
traveltime     int64
studytime      int64
failures       int64
schoolsup     object
famsup        object
paid          object
activities    object
nursery       object
higher        object
internet      object
romantic      object
famrel         int64
```

```
freetime       int64
goout          int64
Dalc           int64
Walc           int64
health         int64
absences       int64
G1             int64
G2             int64
G3             int64
dtype: object
```

```
df.dtypes
```

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

```
#to understand the dataset's size
df.shape
```

```
(395, 33)
```

```
#to find the duplicates
df.duplicated(keep=False)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
390    False
391    False
392    False
393    False
394    False
Length: 395, dtype: bool
```

```
#to remove the duplicates
df = df.drop_duplicates()
```

```
#to find the average score in math
average_g3 = df['G3'].mean()
print ("Average score in Math:",average_g3)
```

```
Average score in Math: 10.415189873417722
```

```
#to find the students scored above 15 in their final grade
count = df['G3'][df['G3'] > 15].value_counts().sum()
print ('Total number of students scored more than 15 is:',count)
```

```
Total number of students scored more than 15 is: 40
```

```
#to read the columns
df.columns
```

```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus',
'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime',
'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities',
'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime',
'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

```
#to find the correlation between study time (study time) and the final
grade (G3)
correlation = df['studytime'].corr(df['G3'])
print(correlation)
```

```
0.09781968965319633
```

```
# Grouping by gender and calculating average G3
average_g3_by_gender = df.groupby('sex')['G3'].mean()
print(average_g3_by_gender)
```

```
sex
F     9.966346
M    10.914439
Name: G3, dtype: float64
```

```
#to find higher average final grade (G3) in gender
highest_avg_gender = average_g3_by_gender.idxmax()
print(f"The gender with the highest average G3 is:
{highest_avg_gender}")
```

```
The gender with the highest average G3 is: M
```