Exploratory Data Analysis on Student Performance Using Python and Pandas.

Preet Dhabuwala

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Loading and Cleaning the Data

```
df = pd.read_csv("student_performance_prediction.csv")
```

```
df
```

	Student_ID	Study_Hour_per_Week	Attendance_Rate	Previous_Grades	Participation_in_Extracurricular_Activities	Parent_Education_Leve
0	S00001	12.5	NaN	75.0	Yes	Mast
1	S00002	9.3	95.3	60.6	No	High Scho
2	S00003	13.2	NaN	64.0	No	Associa
3	S00004	17.6	76.8	62.4	Yes	Bachel
4	S00005	8.8	89.3	72.7	No	Mast
...	...	...	...	...	...	
39995	S39996	15.6	93.8	51.4	Yes	Mast
39996	S39997	11.3	66.4	64.2	No	Doctora
39997	S39998	13.1	65.6	38.1	No	Bachel
39998	S39999	14.1	74.9	NaN	Yes	Mast
39999	S40000	11.8	55.1	68.5	No	Bachel

```
df.head(10)
```

	Student_ID	Study_Hour_per_Week	Attendance_Rate	Previous_Grades	Participation_in_Extracurricular_Activities	Parent_Education_Level
0	S00001	12.5	NaN	75.0	Yes	Master
1	S00002	9.3	95.3	60.6	No	High School
2	S00003	13.2	NaN	64.0	No	Associate
3	S00004	17.6	76.8	62.4	Yes	Bachelor
4	S00005	8.8	89.3	72.7	No	Master
5	S00006	8.8	73.8	69.3	Yes	High School
6	S00007	17.9	38.6	93.6	No	Doctorate
7	S00008	13.8	95.8	59.2	Yes	Doctorate
8	S00009	7.7	100.1	91.9	No	Bachelor
9	S00010	[illegible]	[illegible]	[illegible]	Yes	High School

```
df.tail(10)
```

	Student_ID	Study_Hour_ per_Week	Attendance_Rate	Previous_Grades	Participation_in_Extracurricular_Activities	Parent_Education_Leve
39990	S39991	19.1	NaN	94.9	No	Na
39991	S39992	15.0	52.1	70.4	Yes	Mast
39992	S39993	10.8	46.3	73.5	Yes	Doctora
39993	S39994	7.0	86.3	65.5	No	Bachel
39994	S39995	5.1	92.1	46.1	Yes	Doctora
39995	S39996	15.6	93.8	51.4	Yes	Mast
39996	S39997	11.3	66.4	64.2	No	Doctora
39997	S39998	13.1	65.6	38.1	No	Bachel
39998	S39999	14.1	74.9	NaN	Yes	Mast

```
df.sample(10)
```

	Student_ID	Study_Hour_ per_Week	Attendance_Rate	Previous_Grades	Participation_in_Extracurricular_Activities	Parent_Education_Leve
11102	S11103	5.2	54.9	72.1	Yes	Doctora
10712	S10713	2.8	104.3	NaN	No	Doctora
22984	S22985	11.9	62.4	40.6	No	Bachel
1871	S01872	8.9	92.6	88.4	No	Associa
33948	S33949	4.0	NaN	54.8	Yes	Bachel
35997	S35998	11.8	85.4	60.7	No	Na
13570	S13571	8.5	99.5	63.8	No	Bachel
39778	S39779	9.0	55.5	50.7	Yes	Associa
12698	S12699	3.1	82.4	56.0	Yes	Associa

```
df.isnull().sum()
```

```
Student_ID                                        0
Study_Hour_ per_Week                           1995
Attendance_Rate                                1992
Previous_Grades                                1994
Participation_in_Extracurricular_Activities    2000
Parent_Education_Level                         2000
Passed                                         2000
dtype: int64
```

```
df['Study_Hour_ per_Week']=df['Study_Hour_ per_Week'].fillna(df['Study_Hour_ per_Week'].mean())
```

```
df.shape
```

```
(40000, 7)
```

```
df.columns
```

```
Index(['Student_ID', 'Study_Hour_ per_Week', 'Attendance_Rate',
       'Previous_Grades', 'Participation_in_Extracurricular_Activities',
       'Parent_Education_Level', 'Passed'],
      dtype='object')
```

```
df.info()
df.dtypes
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 7 columns):
 #   Column                                       Non-Null Count  Dtype
---  ------                                       --------------  -----
 0   Student_ID                                   40000 non-null  object
 1   Study_Hour_ per_Week                         38005 non-null  float64
 2   Attendance_Rate                              38008 non-null  float64
 3   Previous_Grades                              38006 non-null  float64
 4   Participation_in_Extracurricular_Activities  38000 non-null  object
 5   Parent_Education_Level                       38000 non-null  object
 6   Passed                                       38000 non-null  object
```

```
dtypes: float64(3), object(4)
memory usage: 2.1+ MB
Student_ID                                     object
Study_Hour_ per_Week                          float64
Attendance_Rate                               float64
Previous_Grades                               float64
Participation_in_Extracurricular_Activities    object
Parent_Education_Level                         object
Passed                                         object
dtype: object
```

```
df.describe()
```

	Study_Hour_ per_Week	Attendance_Rate	Previous_Grades
count	38005.000000	38008.000000	38006.000000
mean	9.962744	75.276323	65.440107
std	5.031154	20.393418	16.503119
min	-12.300000	-14.300000	8.300000
25%	6.600000	61.600000	55.100000
50%	10.000000	75.300000	65.200000
75%	13.400000	88.800000	75.200000
max	32.400000	150.200000	200.000000

```
df.describe(include='object')
```

	Student_ID	Participation_in_Extracurricular_Activities	Parent_Education_Level	Passed
count	40000	38000	38000	38000
unique	40000	2	5	2
top	S00001	No	Bachelor	Yes
freq	1	19028	7685	19011

```
df.Passed.unique()
```

```
array(['Yes', 'No', nan], dtype=object)
```

```
df.Passed.value_counts(dropna=False)
```

```
Passed
Yes    19011
No     18989
NaN     2000
Name: count, dtype: int64
```

Start coding or generate with AI.

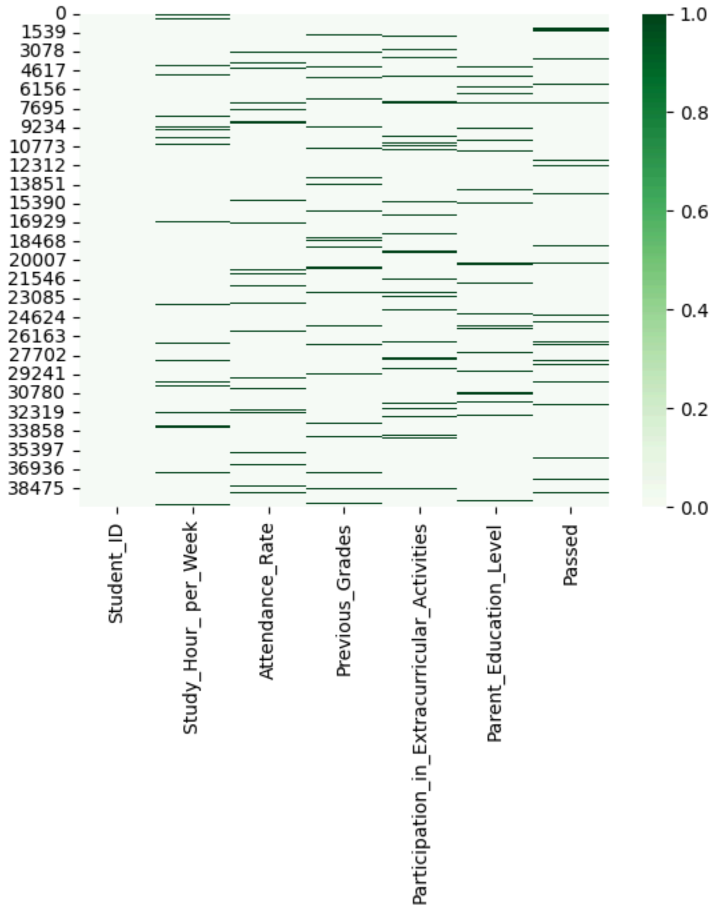
```
df['Student_ID']=df['Student_ID'].str.replace('S','').fillna(0).astype(int)
```

```
df.head(10)
```

	Student_ID	Study_Hour_ per_Week	Attendance_Rate	Previous_Grades	Participation_in_Extracurricular_Activities	Parent_Education_Level
0	1	12.5	NaN	75.0	Yes	Master
1	2	9.3	95.3	60.6	No	High School
2	3	13.2	NaN	64.0	No	Associate
3	4	17.6	76.8	62.4	Yes	Bachelor
4	5	8.8	89.3	72.7	No	Master
5	6	8.8	73.8	69.3	Yes	High School
6	7	17.9	38.6	93.6	No	Doctorate
7	8	13.8	95.8	59.2	Yes	Doctorate
8	9	7.7	100.1	91.9	No	Bachelor
9	10	[illegible]	[illegible]	[illegible]	Yes	High School

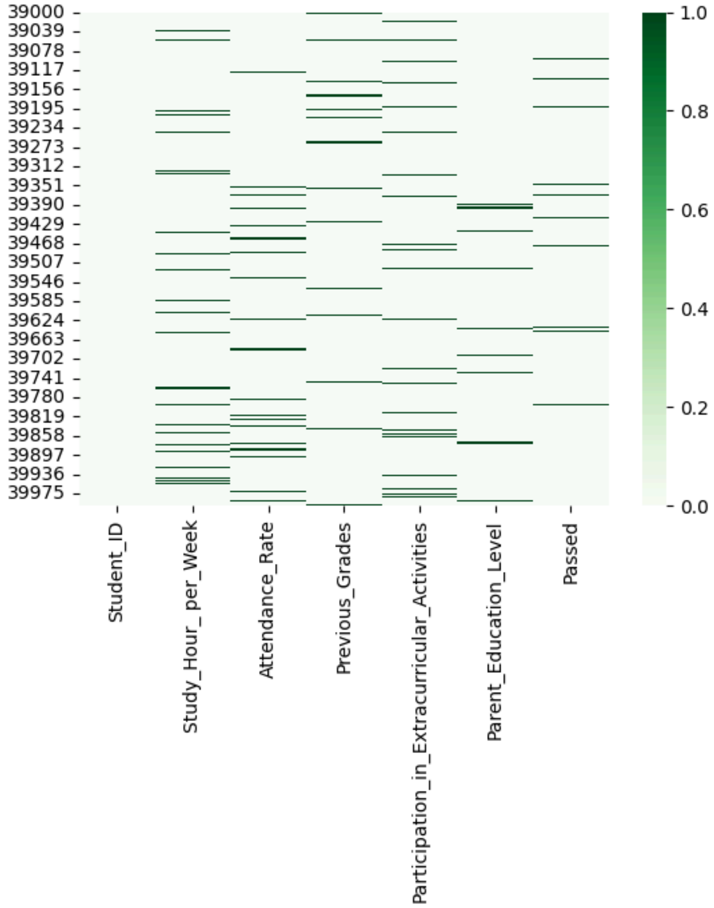
```
sns.heatmap(df.isna(),cmap='Greens')
```

```
<Axes: >
```



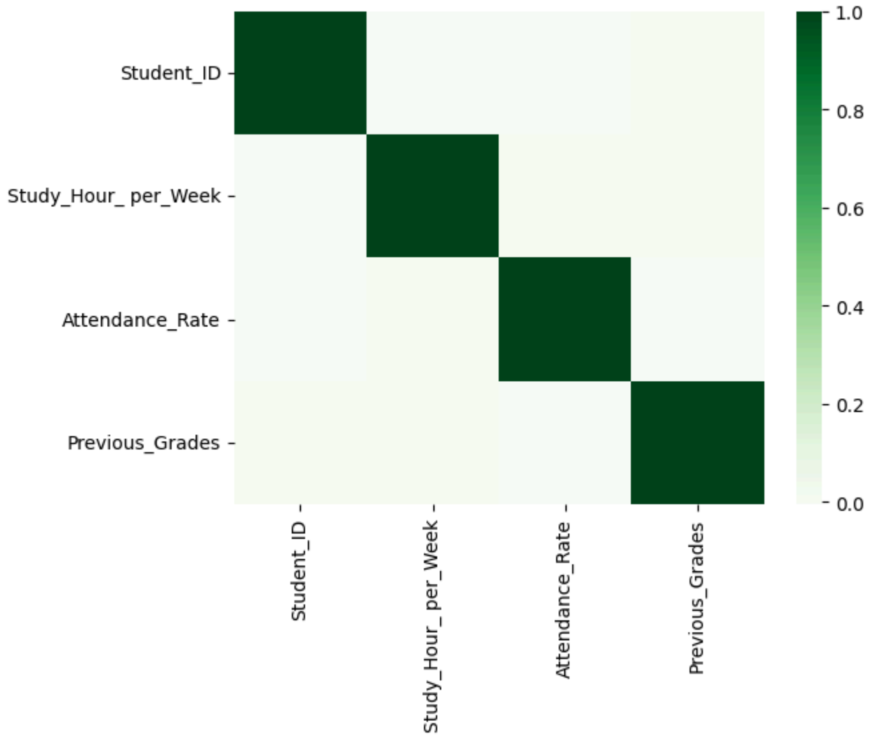
```
sns.heatmap(df.tail(1000).isna(),cmap='Greens')
```

```
<Axes: >
```



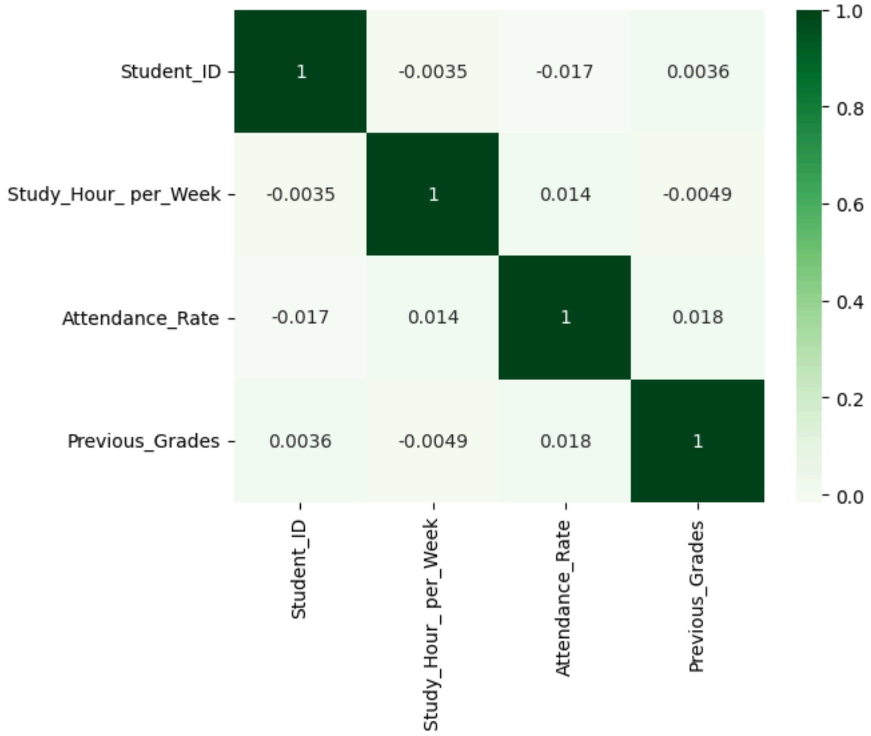
```
sns.heatmap(df.corr(numeric_only=True),cmap='Greens',annot=False)
```

```
<Axes: >
```



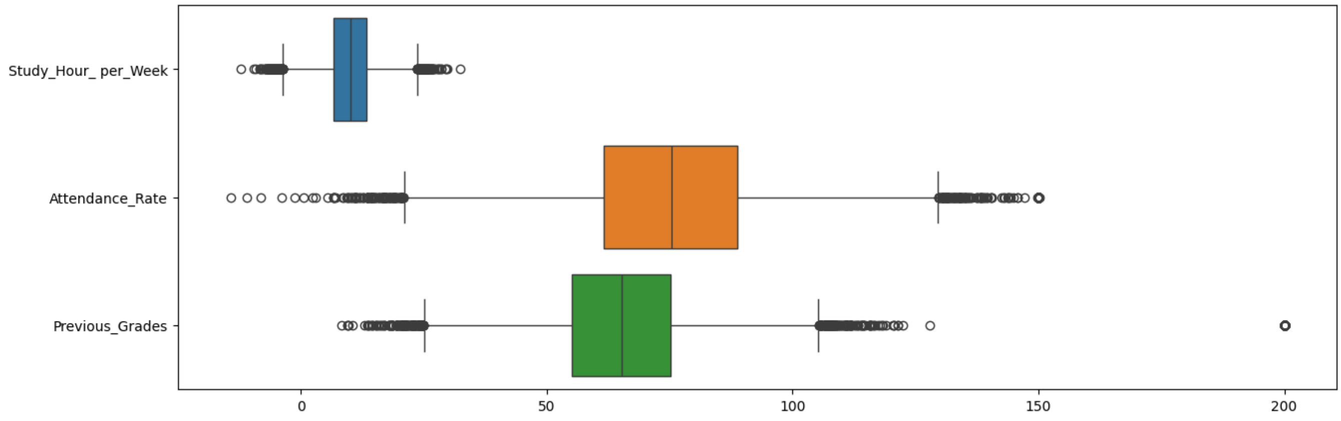
```
sns.heatmap(df[df.Parent_Education_Level=='Master'].corr(numeric_only=True),cmap='Greens',annot=True)
```

```
<Axes: >
```



```
cols = ['Study_Hour_ per_Week','Attendance_Rate','Previous_Grades',]
plt.figure(figsize=(15,5))
sns.boxplot(data=df[cols], orient='h')
```

```
<Axes: >
```



```
df[['Study_Hour_ per_Week','Previous_Grades']].head(18).describe()
```

	Study_Hour_ per_Week	Previous_Grades
count	18.000000	18.000000
mean	9.688889	64.161111
std	4.727005	17.444067
min	0.400000	37.800000
25%	7.700000	50.750000
50%	9.050000	61.500000
75%	12.650000	72.600000