# capstone

October 18, 2020

## 1 TASK DETAILS

**The task aims at better describing the onboard task given with newly gained knowledge from python programming knowledge. Task further includes;**

1. For each variable, create a summary statistics that tells you about the data type.
2. For each variable, use a summary metric that helps you describe the data.
3. For each variable. provide a graphical representation of the data distribution.
4. Create a new variable "average_score" represented from "math", "reading" and "writing" scores.
5. Create another variable "average_score_cat" that categorises the "average_score" using WAEC grading system.
6. Find a relationship (if any) between each variable and the new variable "average_score_cat".
7. Create graphical representation of the relationship(s) discovered in 6.
8. Develop an hypothesis about which variables that can help predict the "average_score_cat" of a new student.

## 2 ABSTRACT

This project gives a descriptive analysis of a dataset consisting of test scores of student in three subjects and determining the relationship between the scores of the students and their gender, economic, personal and social attributes. The given data set was sourced from a Kaggle challenge and consisted of a 1000 rows and 8 columns of qualitative and quantitative data. Univariate and Multivariate analysis of the dataset was carried out using Python and its libraries.

## 3 DATA SET PREPARATION

The data Set contains 1000 rows and 8 columns. Data was checked for special character or white spaces to avoid difficulties in analyses.Data was also checked for missing values.There are 3 sets of quantitative variables and 5 sets of qualitative variables. Required libraries needed for the analyses were also installed appropriately.

[1]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
```

```
sns.set_style('whitegrid')
```

# 4 SUMMARY STATISTICS

[19]:
```
sd = pd.read_csv('student_data.csv')
sd.head()
```

[19]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 |

The above table shows our dataset consisting of qualitative and quantitative data.

[21]:
```
sd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

From the result obtained, we observe that we have a thousand entries in eight columns. Data tyes are given as three for integer based result columns and five for variable or object based columns. We also observe that we have no missing or null values.

[22]:
```
sd.describe(include='all')
```

[22]:

| | gender | race/ethnicity | parental level of education | lunch | \ |
|---|---|---|---|---|---|
| count | 1000 | 1000 | 1000 | 1000 | |
| unique | 2 | 5 | 6 | 2 | |
| top | female | group C | some college | standard | |
| freq | 518 | 319 | 226 | 645 | |
| mean | NaN | NaN | NaN | NaN | |
| std | NaN | NaN | NaN | NaN | |
| min | NaN | NaN | NaN | NaN | |
| 25% | NaN | NaN | NaN | NaN | |
| 50% | NaN | NaN | NaN | NaN | |
| 75% | NaN | NaN | NaN | NaN | |
| max | NaN | NaN | NaN | NaN | |

| | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|
| count | 1000 | 1000.00000 | 1000.000000 | 1000.000000 |
| unique | 2 | NaN | NaN | NaN |
| top | none | NaN | NaN | NaN |
| freq | 642 | NaN | NaN | NaN |
| mean | NaN | 66.08900 | 69.169000 | 68.054000 |
| std | NaN | 15.16308 | 14.600192 | 15.195657 |
| min | NaN | 0.00000 | 17.000000 | 10.000000 |
| 25% | NaN | 57.00000 | 59.000000 | 57.750000 |
| 50% | NaN | 66.00000 | 70.000000 | 69.000000 |
| 75% | NaN | 77.00000 | 79.000000 | 79.000000 |
| max | NaN | 100.00000 | 100.000000 | 100.000000 |

From the table, we observe a count of a thousand entries as stated earlier. For the gender group we observe females have the highest frequency, the group C have the highest frequency for race/ethnicity. Some college, standard and none also have the highest frequencies in their respective groups. We also observe several results for calculations of some central measures of dispersion. For the mean we have reading score having the highest mean, with wiing score having the highest standard deviation value. All three scores have a maximum value of 100, with maths having the highest range and lowse minimum, found by subtracting the its minimum value from its maximum. We also observed several unique values in our object data types.

[17]:
```
math_mode = sd['math score'].mode()

math_median = sd['math score'].median()

print('The mode of math score is',math_mode[0],' and the median is given as',␣
 ↪math_median,'.' '\n')


reading_mode = sd['reading score'].mode()

reading_median = sd['reading score'].median()
```

```
print('The mode of reading score is',reading_mode[0],' and the median is given␣
 ↪as', reading_median,'.' '\n')


writing_mode = sd['writing score'].mode()

writing_median = sd['writing score'].median()

print('The mode of writing score is',writing_mode[0],' and the median is given␣
 ↪as', writing_median,'.')
```

```
The mode of math score is 65  and the median is given as 66.0 .

The mode of reading score is 72  and the median is given as 70.0 .

The mode of writing score is 74  and the median is given as 69.0 .
```

The mode and median values were also computed.

[20]:
```
print(sd['gender'].value_counts(),'\n')
print(sd['race/ethnicity'].value_counts(), '\n')
print(sd['parental level of education'].value_counts(), '\n')
print(sd['lunch'].value_counts(), '\n')
print(sd['test preparation course'].value_counts(), '\n')
```

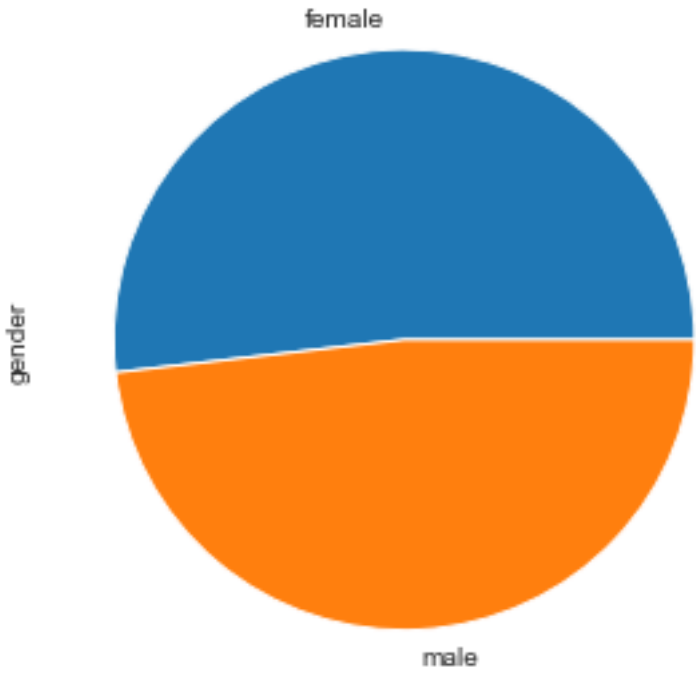
```
female    518
male      482
Name: gender, dtype: int64

group C    319
group D    262
group B    190
group E    140
group A     89
Name: race/ethnicity, dtype: int64

some college          226
associate's degree    222
high school           196
some high school      179
bachelor's degree     118
master's degree        59
Name: parental level of education, dtype: int64

standard        645
free/reduced    355
Name: lunch, dtype: int64
```

```
none         642
completed    358
Name: test preparation course, dtype: int64
```

The above is a breakdown of the groups showing frequencies of subgroups.

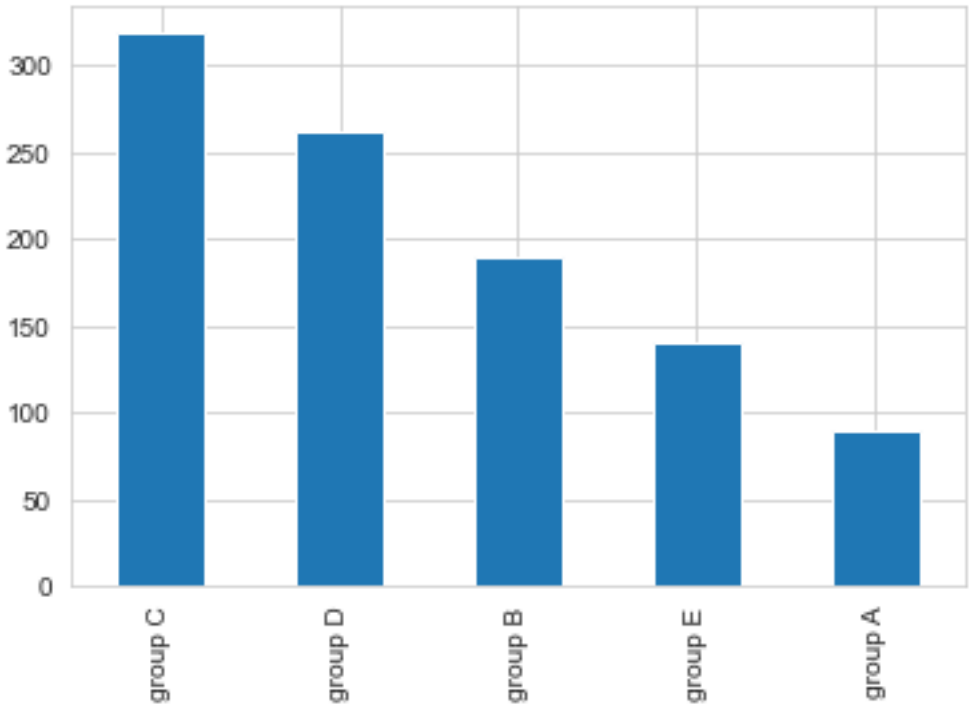
[31]:
```
sd['gender'].value_counts().plot.pie(figsize = (6, 5))
```

[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x10f0184fd30>
```

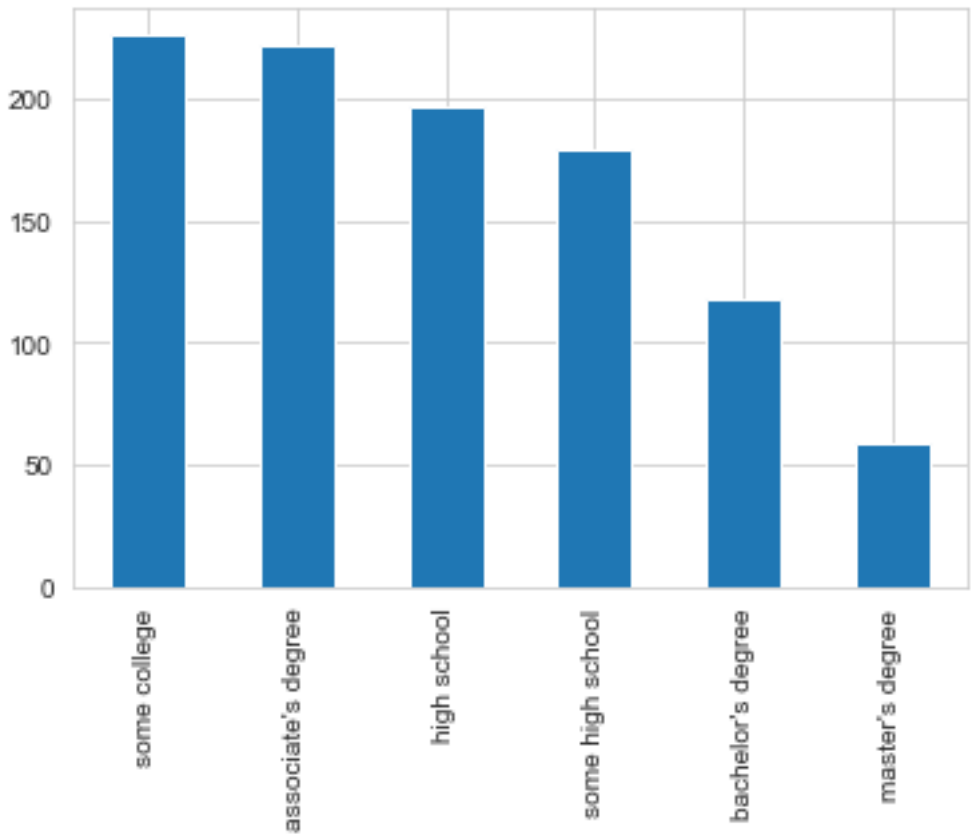



Piechart showing the distribution of males and females in gender.

[27]:
```
sd['race/ethnicity'].value_counts().plot.bar()
plt.show()
```

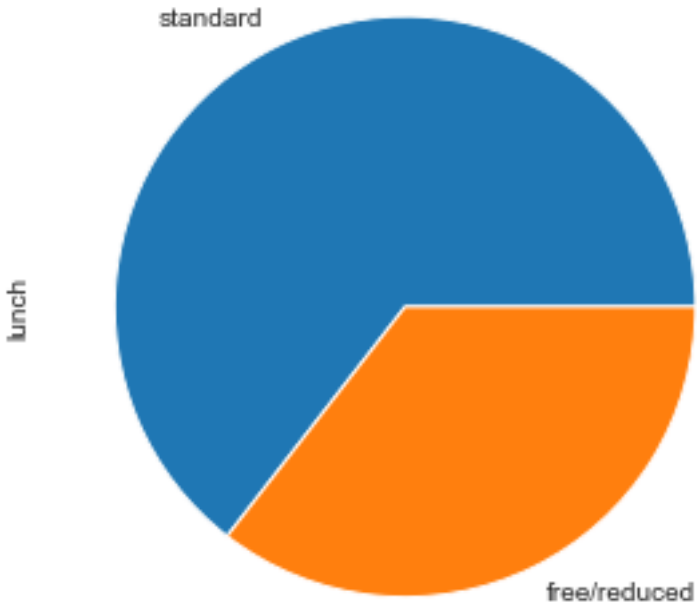



[28]:
```
sd['parental level of education'].value_counts().plot.bar()
plt.show()
```



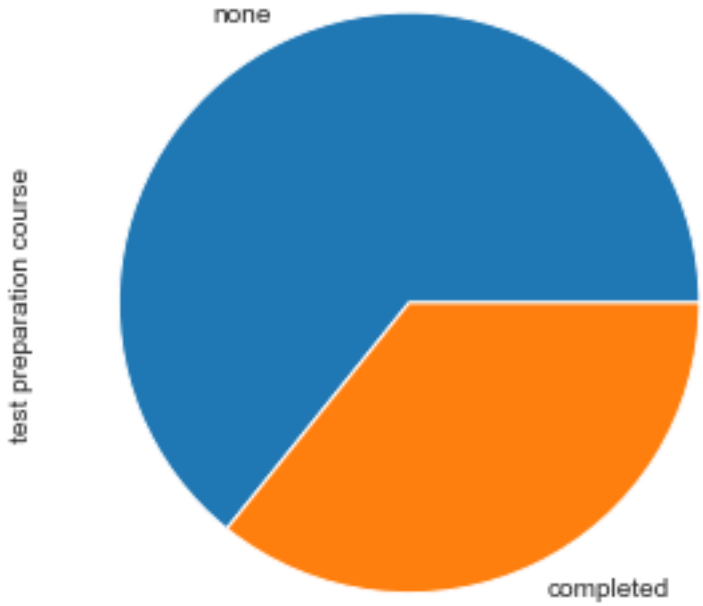

[29]:
```
sd['lunch'].value_counts().plot.pie(figsize = (6, 5))
```

[29]: <matplotlib.axes._subplots.AxesSubplot at 0x10f01765820>



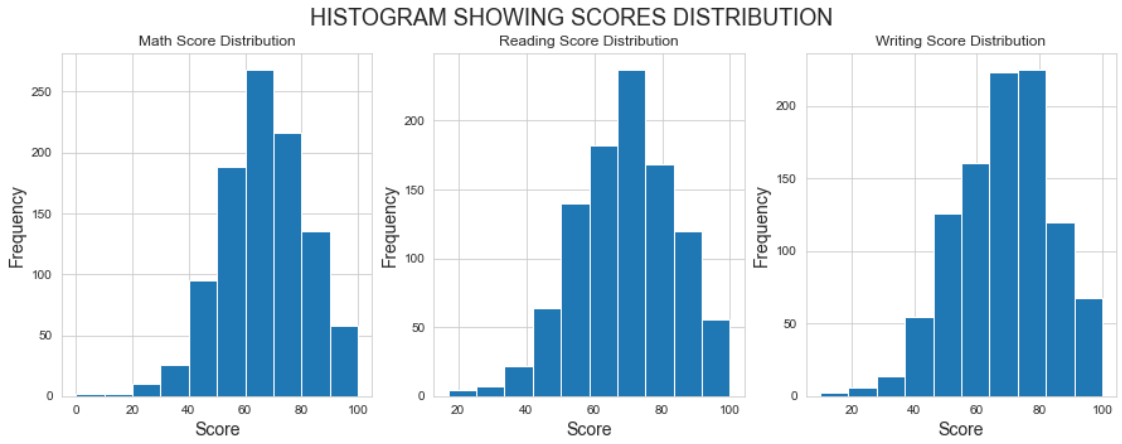

[30]:
```
sd['test preparation course'].value_counts().plot.pie(figsize = (6, 5))
```

[30]: <matplotlib.axes._subplots.AxesSubplot at 0x10f01816070>




[38]:
```
plt.figure(figsize=(15,5))
plt.subplot(1,3,1)
plt.hist(x = sd["math score"])
plt.title("Math Score Distribution")
plt.ylabel('Frequency', fontsize=14)
plt.xlabel('Score', fontsize=14)

plt.subplot(1,3,2)
plt.hist(x = sd["reading score"])
plt.title("Reading Score Distribution")
plt.ylabel('Frequency', fontsize=14)
plt.xlabel('Score', fontsize=14)

plt.subplot(1,3,3)
plt.hist(x = sd["writing score"])
plt.title("Writing Score Distribution")
plt.ylabel('Frequency', fontsize=14)
plt.xlabel('Score', fontsize=14)

plt.suptitle('HISTOGRAM SHOWING SCORES DISTRIBUTION', fontsize = 18)
```

[38]:
```
Text(0.5, 0.98, 'HISTOGRAM SHOWING SCORES DISTRIBUTION')
```




From the histogram, we observe that most students scored between 60-80%. We also observed a left-skewed curve meaning that most of the data values are found right of the curve as the median is found to have a slightly larger value than the mean.

# 5 CREATING AVERAGE_SCORE AND AVERAGE_SCORE_CAT

[39]:
```
sd['average_score'] = round(sd.mean(axis=1))
```

[40]:
```
sd.head()
```

[40]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score | average_score |
|---|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 | 73.0 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 | 82.0 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 | 93.0 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 | 49.0 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 | 76.0 |

[47]:
```
def waec_grade(row):
    if row >= 85:
        return 'A1'
    if (row >= 70) and (row < 85):
        return 'B2'
    if (row >= 65) and (row < 70):
        return 'B3'
    if (row >= 60) and (row < 65):
        return 'C4'
    if (row >= 55) and (row < 60):
        return 'C5'
    if (row >= 50) and (row < 55):
        return 'C6'
    if (row >= 45) and (row < 50):
        return 'D7'
    if (row >= 40) and (row < 45):
        return 'E8'
    else:
        return 'F9'


sd['average_score_cat'] = sd['average_score'].apply(waec_grade)
```

[44]:
```
sd.head()
```

[44]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score | average_score | average_score_cat |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 | 73.0 | B2 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 | 82.0 | B2 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 | 93.0 | A1 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 | 49.0 | D7 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 | 76.0 | B2 |

# 6 RELATIONSHIP BETWEEN SCORES AND AVERAGE_SCORE

[50]:
```
x = sd[['math score', 'reading score', 'writing score', 'average_score']]
x.corr()
```

[50]:

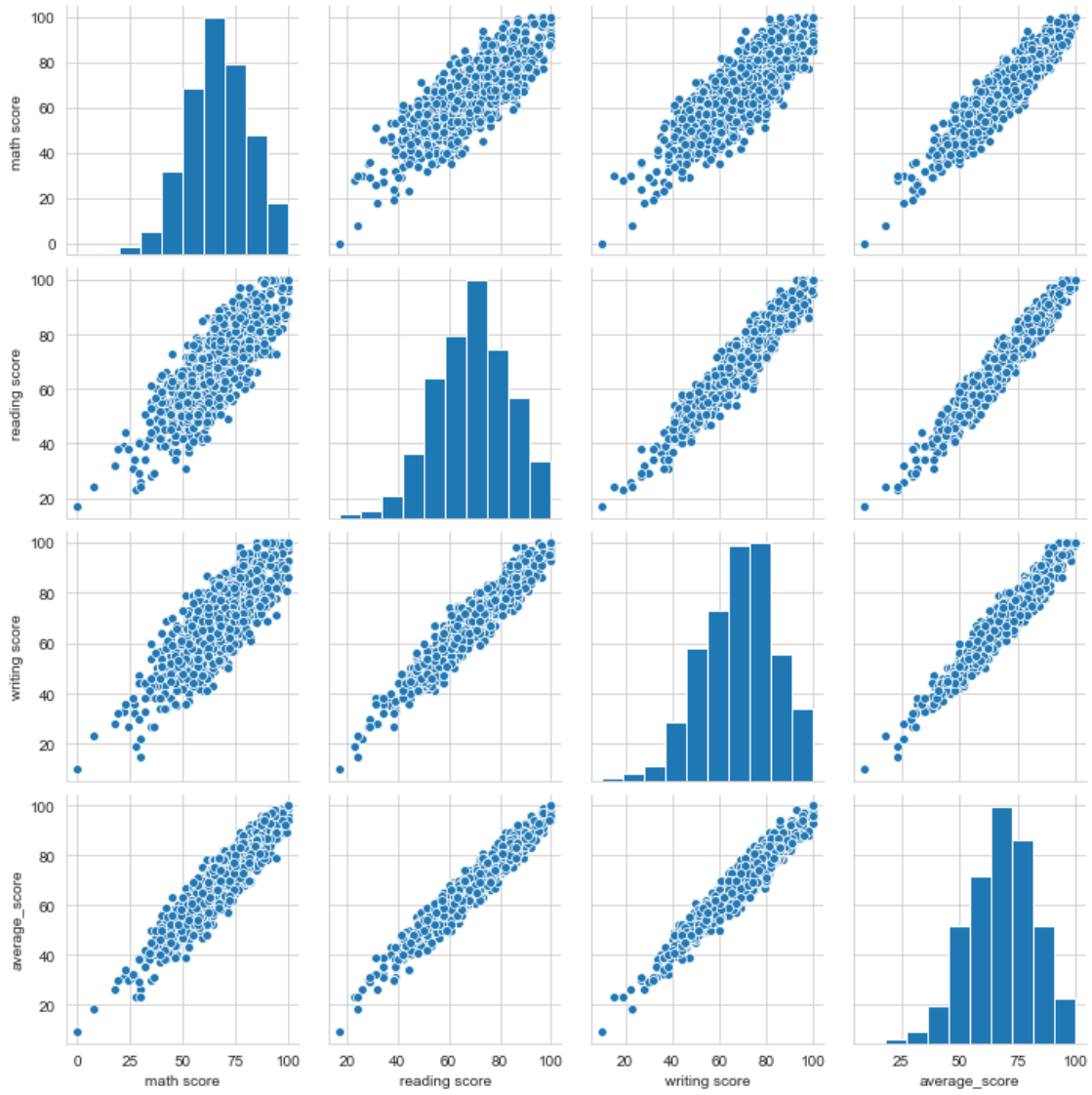
| | math score | reading score | writing score | average_score |
|---|---|---|---|---|
| math score | 1.000000 | 0.817580 | 0.802642 | 0.918442 |
| reading score | 0.817580 | 1.000000 | 0.954598 | 0.970143 |
| writing score | 0.802642 | 0.954598 | 1.000000 | 0.965643 |
| average_score | 0.918442 | 0.970143 | 0.965643 | 1.000000 |

The pearson correlation table above attempts to show a relationship between test scores and the average score. From the positive R values found in the table, it can be inferred that there is a strong relationship between test scores and their average.

## 6.1 Graphical Representation of The Relationship (Correlation)

[51]:
```
sns.pairplot(sd)
```

[51]:
```
<seaborn.axisgrid.PairGrid at 0x10f0217acd0>
```




From the graphs above, we observe considerable correlatons between them.

# 7 HYPOTHESIS AND CONCLUSION

From calculations carried out and trends observed, we observed the females generally perform better than their male counterparts, with males only out doing them in maths. We also observed that good knowledge and scores in all three subjects give better average scores. It was also generally observed that students from Race/Ethnicity E with Parents having a masters degree having completed the Test preparation course and taking Standard Lunch, do better in the test especially the females. Thus, student scores are thus affected by their economical, social and personal attributes.

[ ]: