graphical-analysis

March 5, 2023

1 Graphical analysis of data

[24]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[27]:
```
df = pd.read_csv('/content/drive/MyDrive/Student.csv', na_values = ['??','???'])
df.dropna(inplace = True)
df.reset_index(drop = True, inplace = True)
df.head(5)
```

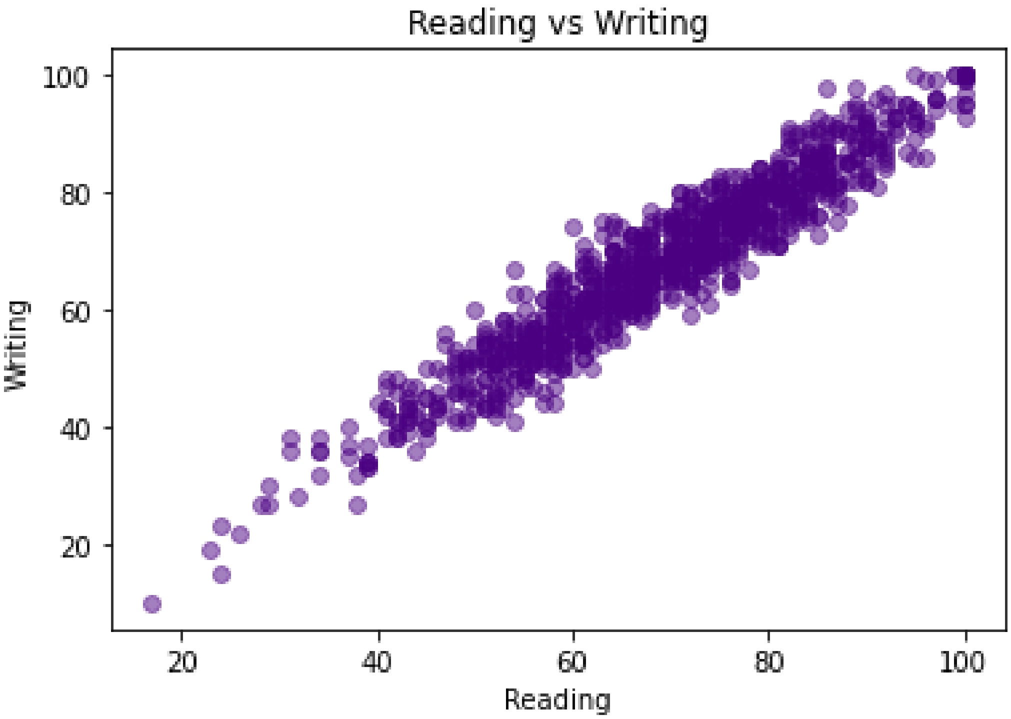
[27]:
```
   Gender Ethnicity                 PLE         Lunch        TPE  Math score  \
0  female   group B   bachelor's degree      standard       none          72
1  female   group C        some college      standard  completed          69
2  female   group B     master's degree      standard       none          90
3    male   group A  associate's degree  free/reduced       none          47
4    male   group C        some college      standard       none          76

   Reading score  Writing score
0             72             74
1             90             88
2             95             93
3             57             44
4             78             75
```

2 Using MATPLOTLIB

1. Scatter plot

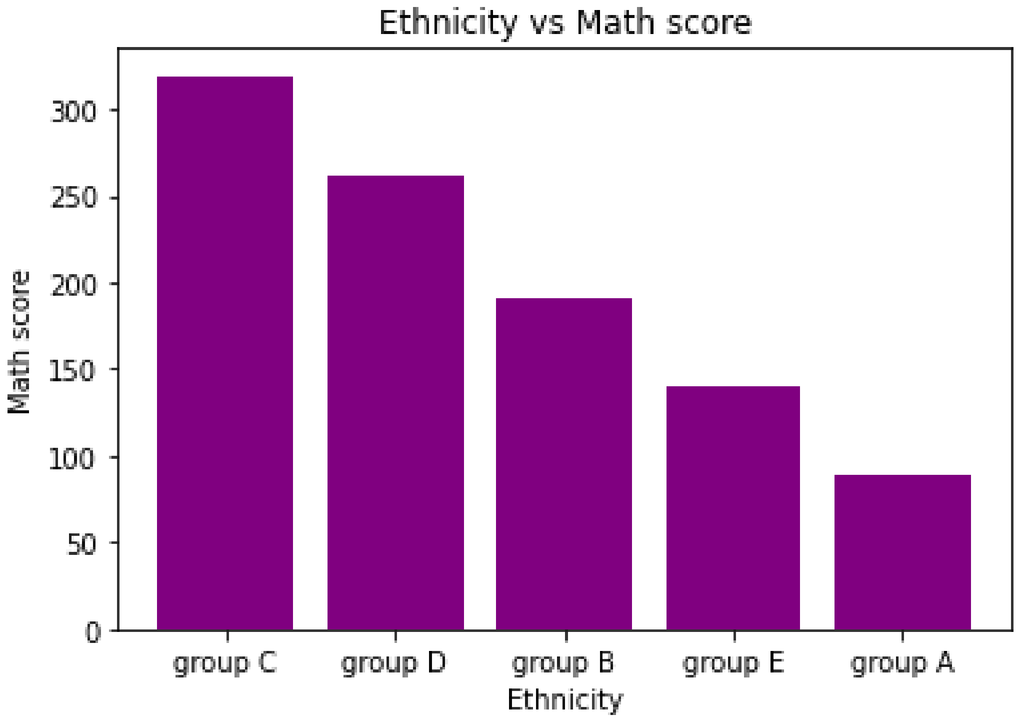
[8]:
```
plt.scatter(df['Reading score'], df['Writing score'], color = 'indigo', alpha =␣
  ↪0.5)
plt.title('Reading vs Writing')
plt.xlabel('Reading')
plt.ylabel('Writing')

plt.show()
```

```
# The scatter plot below shows the variation of reading score with respect to
# writing score. The variation is linear and depicts that as the reading score
# increases the writing score increases as well.
```



2. Bar plot

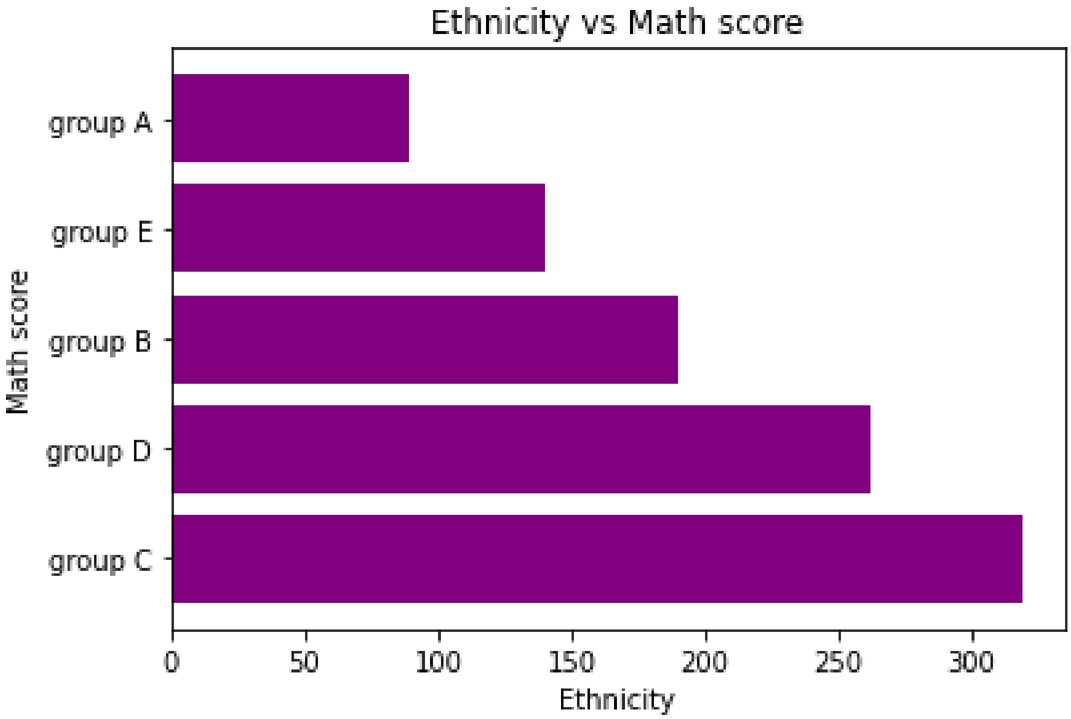
[9]:
```
data = df['Ethnicity'].value_counts()
x = data.index
y = data.values

plt.bar(x, y, color = 'purple')
plt.title("Ethnicity vs Math score")
plt.xlabel("Ethnicity")
plt.ylabel("Math score")

plt.show()

# The bar plot below shows unique value of each group from ethnicity and the
# height represents its peak value. It is clear that group C achieved the
# highest math score while group achieved the lowest.
```



2.1 Horizontal bar plot

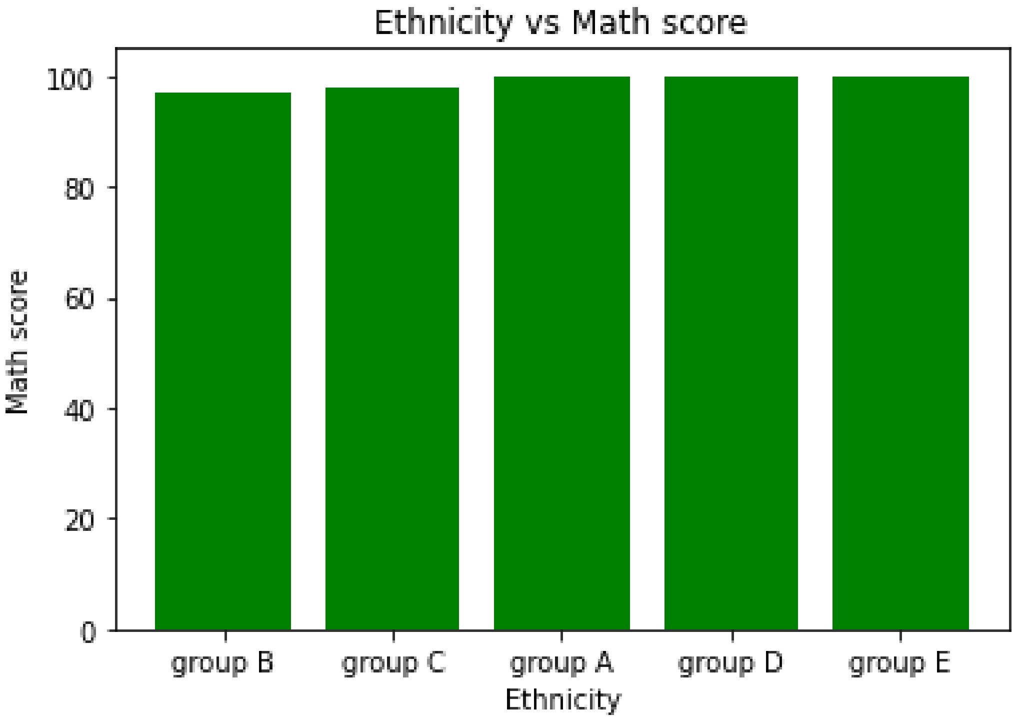
[10]:

```python
data = df['Ethnicity'].value_counts()
x = data.index
y = data.values

plt.barh(x, y, color = 'purple')
plt.title("Ethnicity vs Math score")
plt.xlabel("Ethnicity")
plt.ylabel("Math score")

plt.show()

# The inference is same as above. The only thing that changed is the␣
 ↪orientation.
# This type of plot is highly preferable in applications like dashboards.
```



2.2 Bar plot for every value

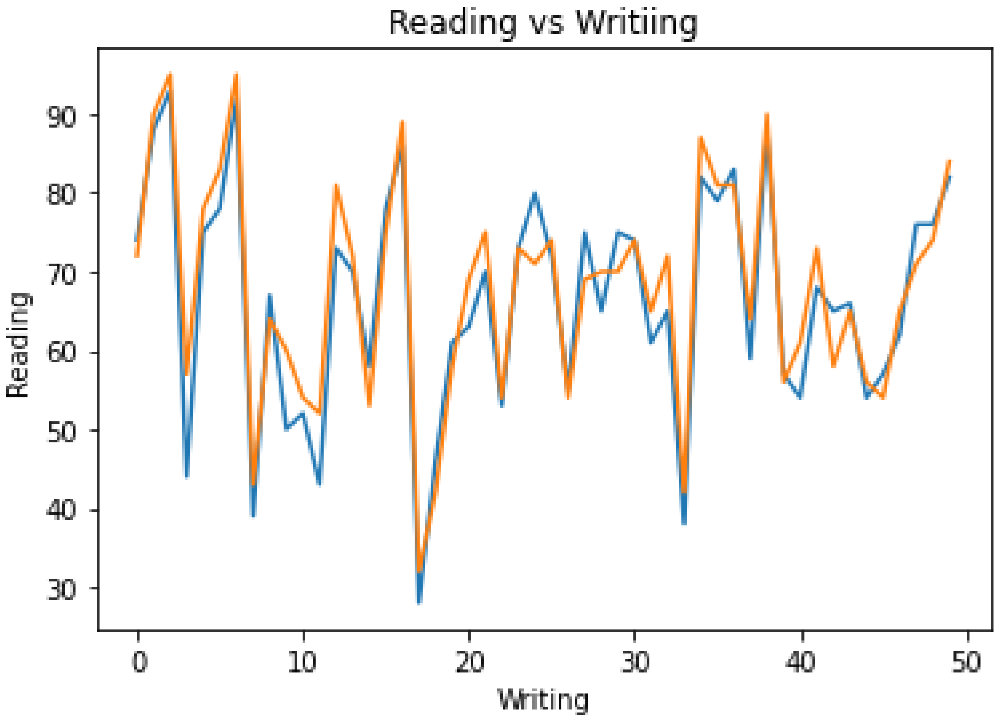
[87]:
```
X = list(df['Ethnicity'])
Y = list(df['Math score'])

plt.bar(X, Y, color ='g')
plt.title("Ethnicity vs Math score")
plt.xlabel("Ethnicity")
plt.ylabel("Math score")

plt.show()

# Below bar plot shows all the value of Ethnicity with respect to math score
# achieved by each group
```



3. Line chart

[57]:

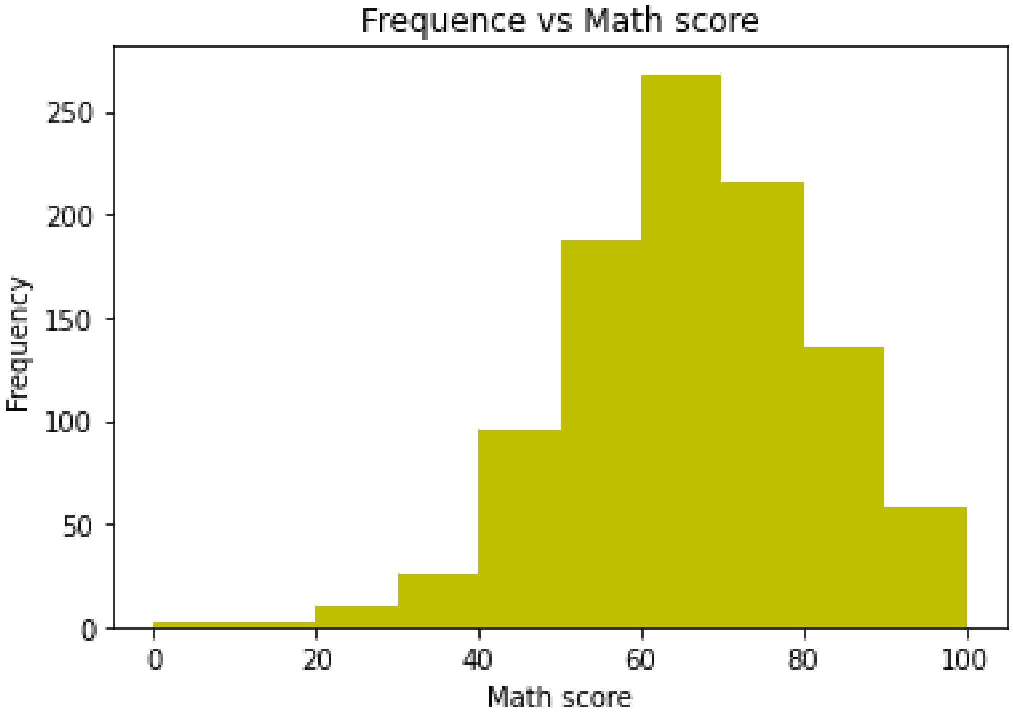
```
c = ['Writing score', 'Reading score']

x = range(50)
for i in c:
  plt.plot(x, df[i].head(50))
plt.title('Reading vs Writiing')
plt.xlabel('Writing')
plt.ylabel('Reading')

plt.show()

# From below line chart it can be infered that the spikes of writing score
# varies equivalent to reading score which means that they are dependent on␣
  ↪each
# other and if reading increases so is the writing.
```



4. Histogram

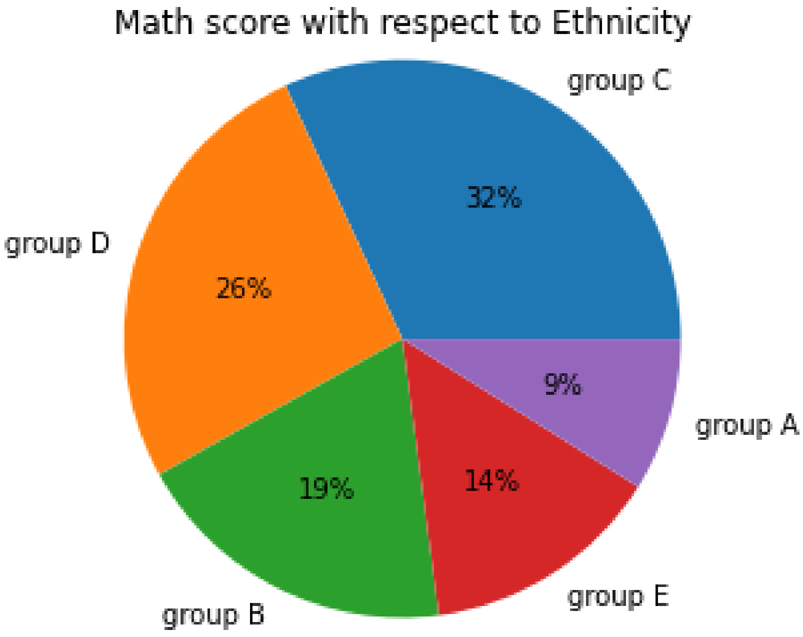
```
[90]: plt.hist(df['Math score'], color = 'y')
      plt.title('Frequence vs Math score')
      plt.xlabel('Math score')
      plt.ylabel('Frequency')

      # The histogram below shows the peak of math score using frequency by taking
      # score as continuous interval. It can be seen that the math score is achieved
      # maximum for range 60-70 and minimum for range 0-20 i.e., most of the students
      # achieved math score in the range 60-70.
```

```
[90]: Text(0, 0.5, 'Frequency')
```



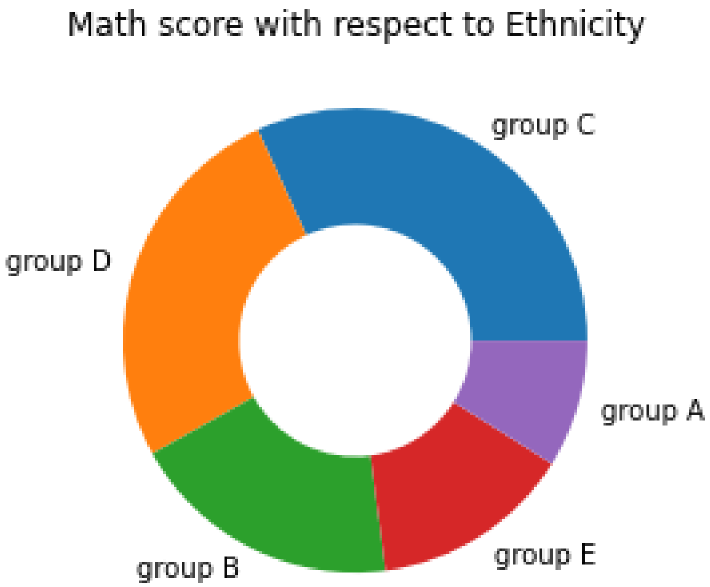
5. Pie chart

```
[80]: data = df['Ethnicity'].value_counts()
x = data.index
y = data.values
plt.pie(y, labels = x, radius = 1.2, autopct = '%1.0f%%')
plt.title('Math score with respect to Ethnicity')

plt.show()

# The pie chart below represents the composition of each group of ethnicity
# with respect to the maths score in terms of percentage. It shows that most of
# the people who scored well in maths belong to group C while those who scored
# minimum belongs to group A.
```



6. Donut chart

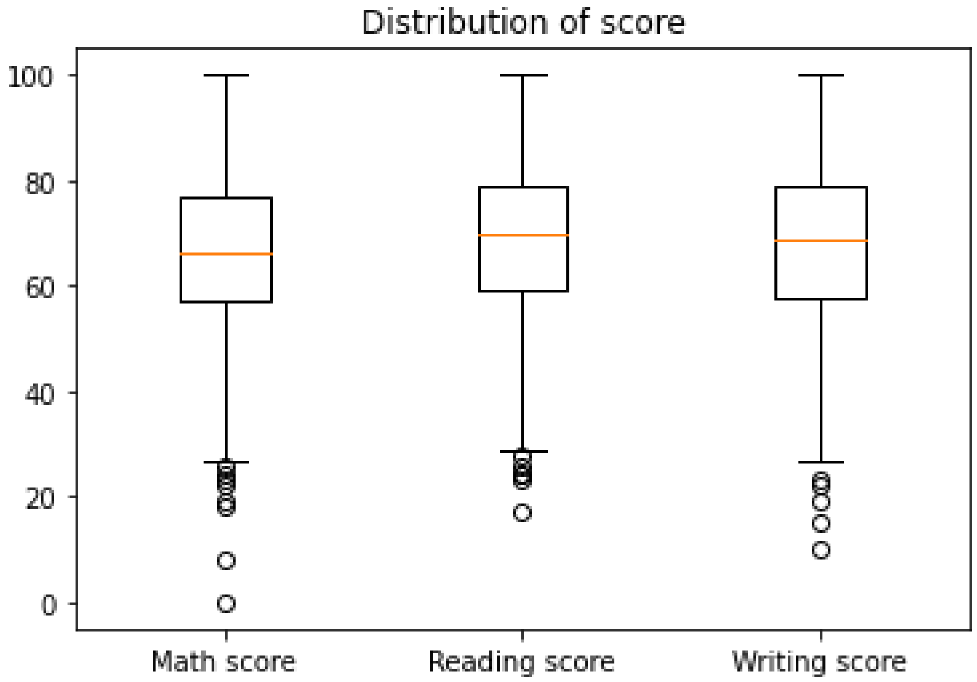
```
[88]: plt.pie(y, labels = x, wedgeprops = dict(width=0.5))
plt.title('Math score with respect to Ethnicity')
plt.show()

# The donut chart below represents the same thing as pie chart above. The only
# difference is that the donut chart is more appealing than pie chart and is
# widely used in dashboards for better visualization of data.
```



7. Box plot

[55]:

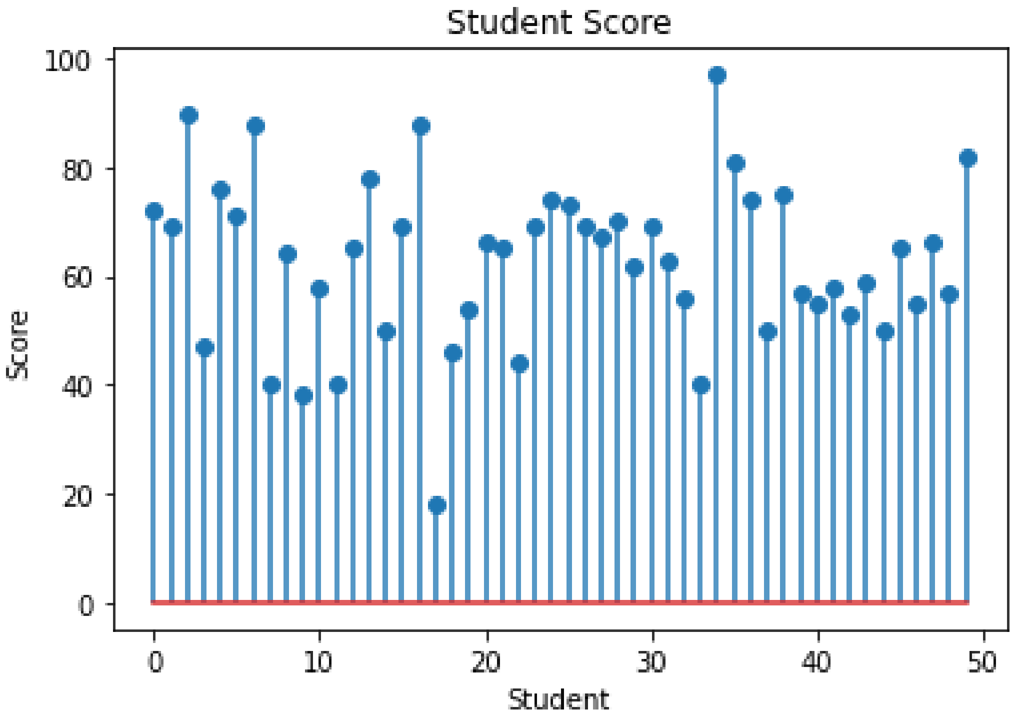
```
plt.boxplot([df['Math score'], df['Reading score'], df['Writing score']])
plt.xticks([1, 2, 3], ['Math score', 'Reading score', 'Writing score'])
plt.title('Distribution of score')
plt.show()

# The box plot below represnts the statistical aspects of numerical data. Here
# numerical data is different type of scores. It can be percieved that the
# median of reading and writing score is same but the whiskers of writing score
# extends to greater length compared to reading score. Also, most of the values
# of math score lies in upper quatile.
```



8. Stem plot

[109]:

```
marks = df['Math score'].head(50)
plt.stem(marks)
plt.xlabel('Student')
plt.ylabel('Score')
plt.title('Student Score')

plt.show()

# The stem plot below shows the first 50 students score. The student are plotted
# along x-axis which is the leaf while score achieved by each individual is
# plotted along y-axis shown using stem. It can be inferred that the 34th
# student achieved the highest score.
```

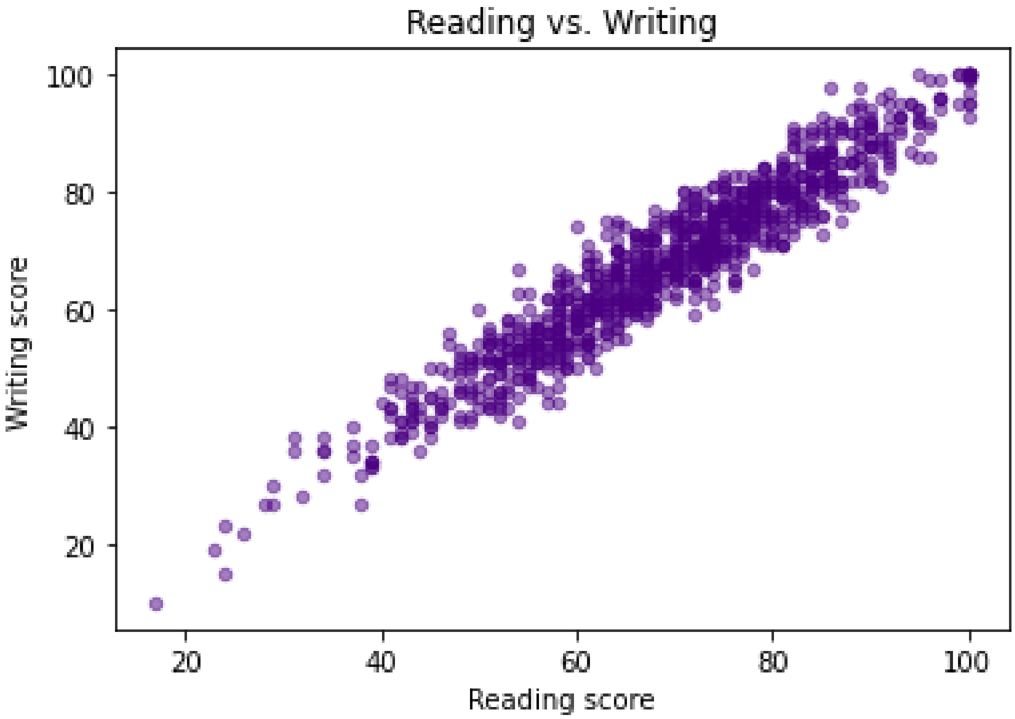


3 Using PANDAS

1. Scatter plot

```
[114]: df.plot.scatter(x = 'Reading score', y = 'Writing score', title = 'Reading vs.␣
        ↪Writing', color = 'indigo', alpha = 0.5)
```

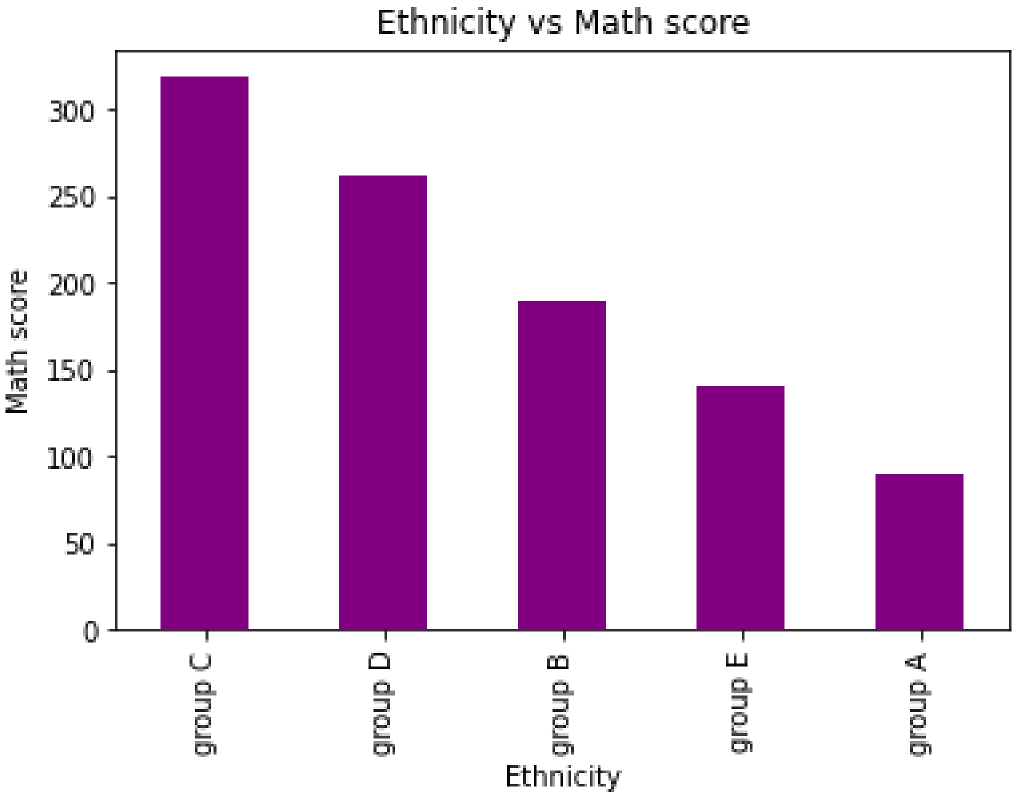
```
[114]: <AxesSubplot:title={'center':'Reading vs. Writing'}, xlabel='Reading score',
       ylabel='Writing score'>
```



2. Bar plot

```
[123]: df['Ethnicity'].value_counts().plot.bar(color = 'purple', title = 'Ethnicity vs␣
        ↪Math score', xlabel = 'Ethnicity', ylabel = 'Math score')
```

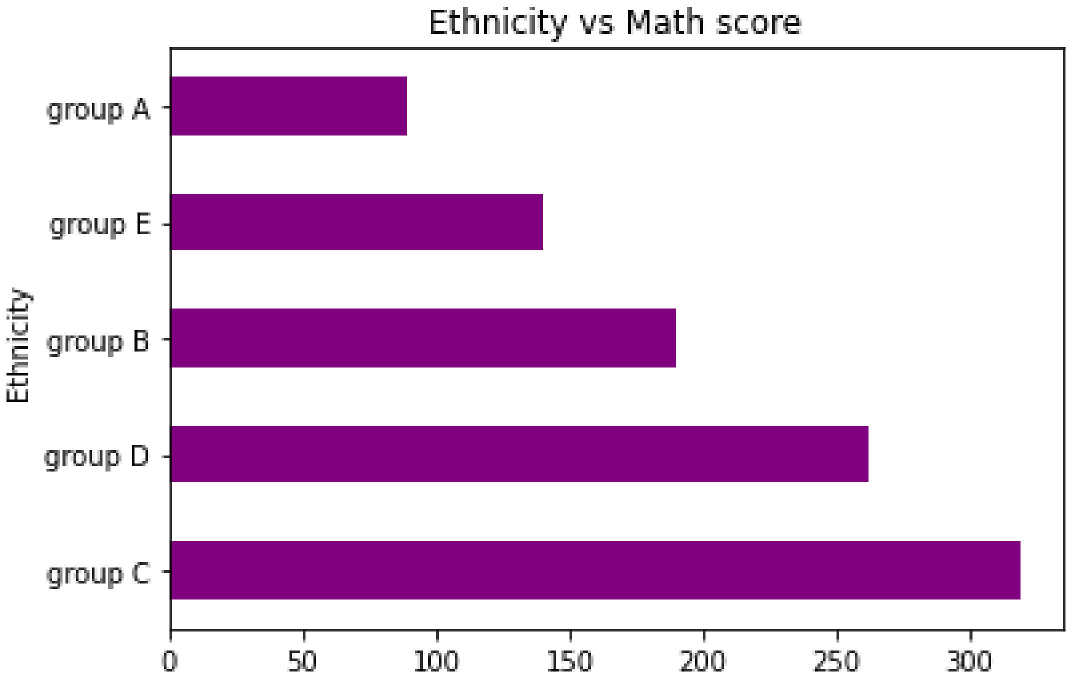
```
[123]: <AxesSubplot:title={'center':'Ethnicity vs Math score'}, xlabel='Ethnicity',
       ylabel='Math score'>
```



2.1 Horizontal bar plot

```
[127]: df['Ethnicity'].value_counts().plot.barh(color = 'purple', title = 'Ethnicity␣
       ↪vs Math score', xlabel = 'Ethnicity', ylabel = 'Math score')
```

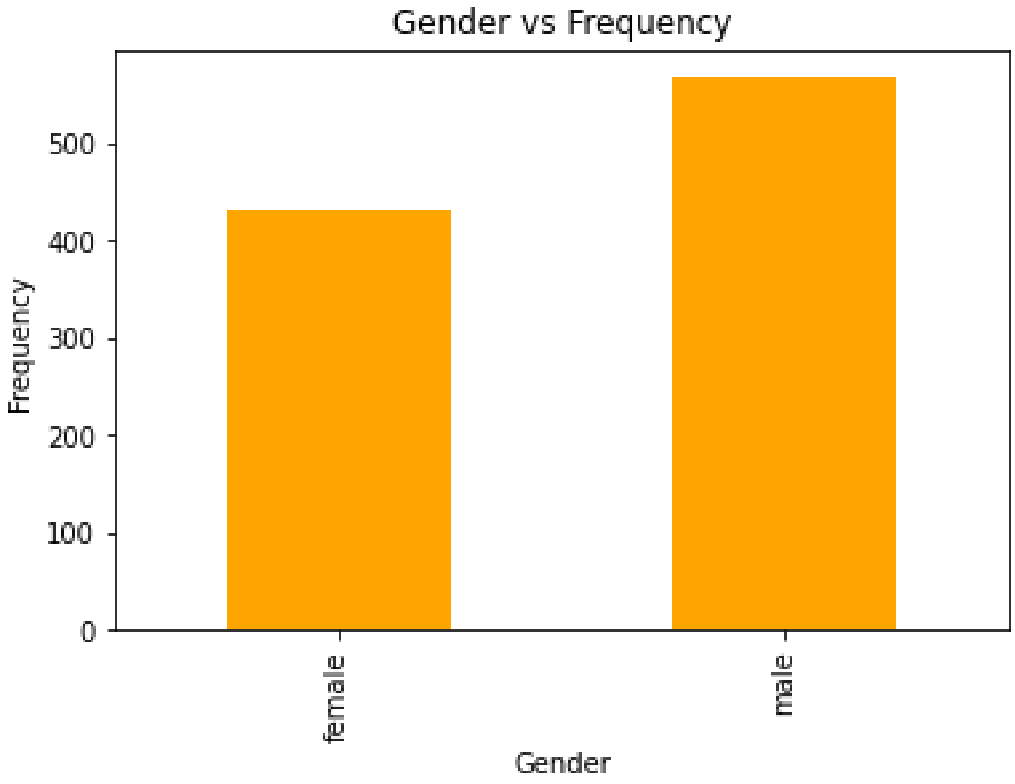
```
[127]: <AxesSubplot:title={'center':'Ethnicity vs Math score'}, ylabel='Ethnicity'>
```



2.2 Bar plot for sorted index

```
[20]: df['Gender'].value_counts().sort_index().plot.bar(title = 'Gender vs␣
        ↪Frequency', xlabel = 'Gender', ylabel= 'Frequency', color = 'orange')
```

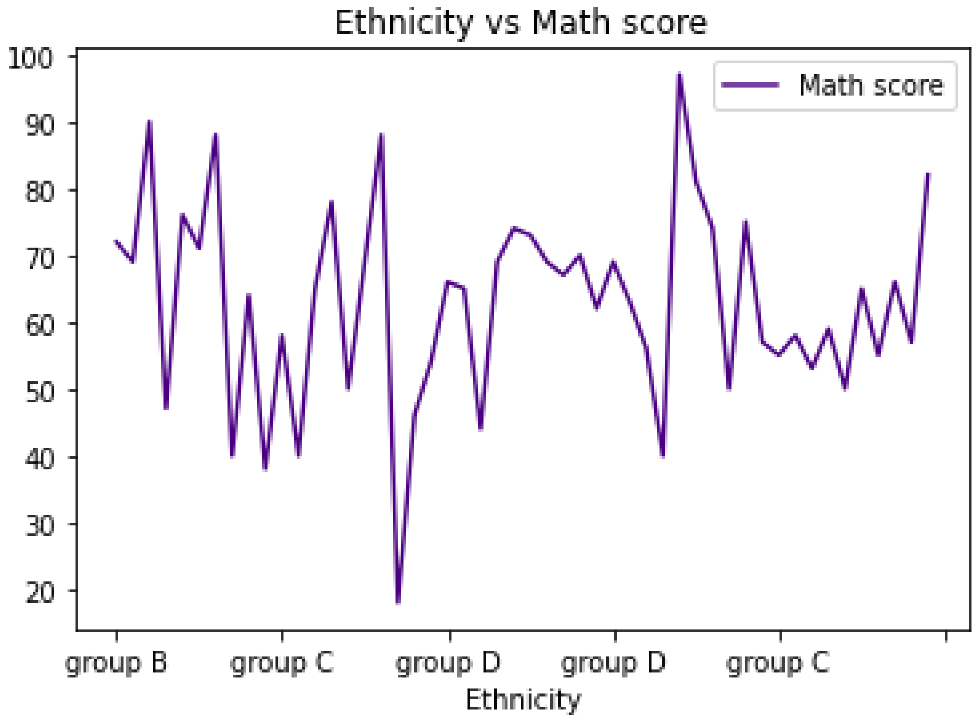
```
[20]: <AxesSubplot:title={'center':'Gender vs Frequency'}, xlabel='Gender',
      ylabel='Frequency'>
```



3. Line chart

```
[31]: df.head(50).plot.line(x = 'Ethnicity', y = 'Math score', title = 'Ethnicity vs␣
        ↪Math score', color = 'indigo')
```

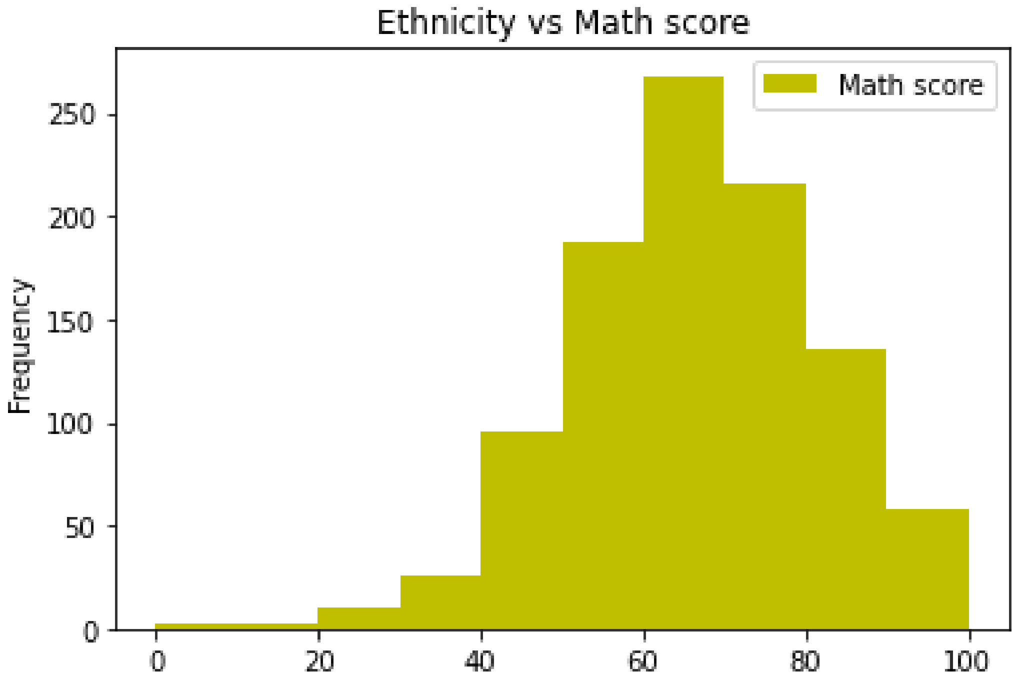
```
[31]: <AxesSubplot:title={'center':'Ethnicity vs Math score'}, xlabel='Ethnicity'>
```



4. **Histogram**

```
[33]: df.plot.hist(x = 'Ethnicity', y = 'Math score', title = 'Ethnicity vs Math␣
       ↪score', color = 'y')
```

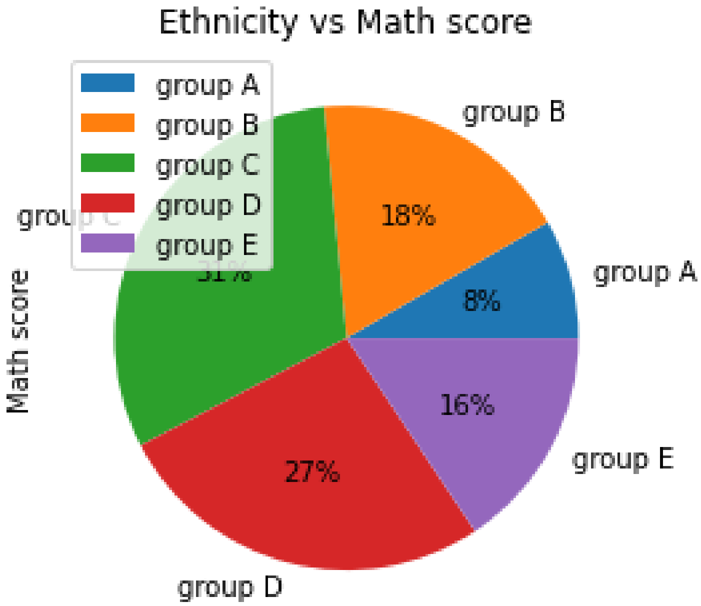
```
[33]: <AxesSubplot:title={'center':'Ethnicity vs Math score'}, ylabel='Frequency'>
```



5. **Pie chart**

```
[85]: df.groupby(['Ethnicity']).sum().plot.pie(y = 'Math score', autopct = '%1.0f%%',␣
      ↪title = 'Ethnicity vs Math score')
```

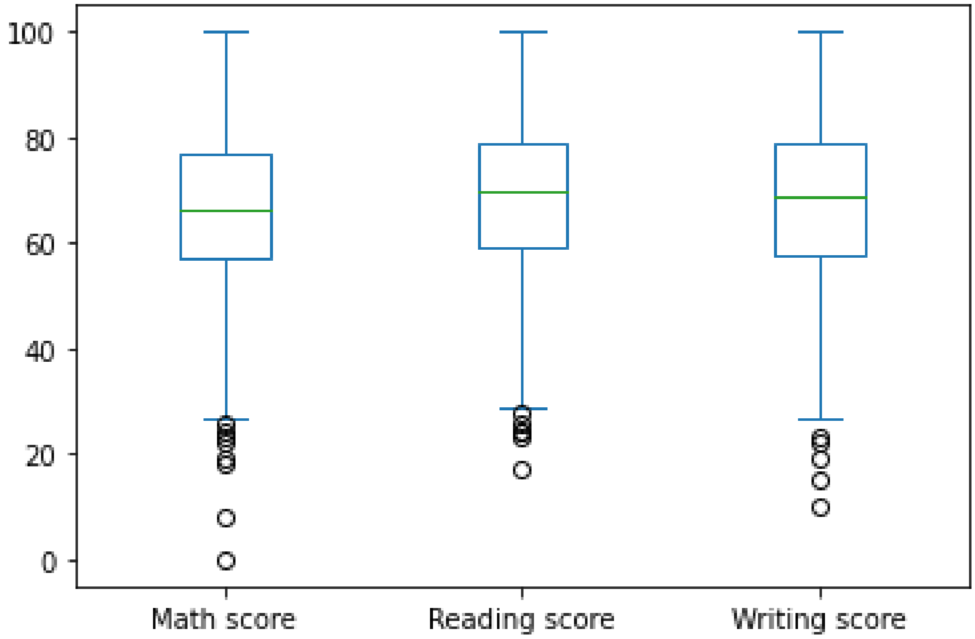
```
[85]: <AxesSubplot:title={'center':'Ethnicity vs Math score'}, ylabel='Math score'>
```



6. Box plot

```
[60]: df.plot.box(['Math score', 'Reading score', 'Writing score'])
```

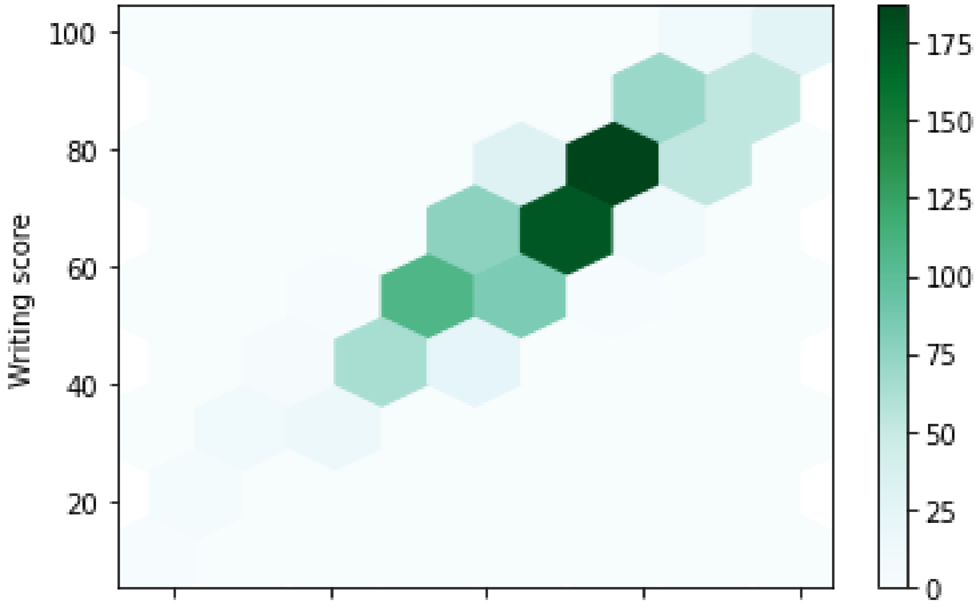
```
[60]: <AxesSubplot:>
```



7. Hexbin plot

```
[89]: df.plot.hexbin(x = 'Reading score', y = 'Writing score', gridsize = 7)

      # The hexbin plot below shows the trend between reading score and writing score
      # which is similar to scatter plot shown above. The only thing which makes it
      # better to visualize data is that it shows where all the data is scattered the
      # most. It can be infered that most of the scores lie in the range 75-85 which
      # is why the hexagonal bin is darker compared to other bins.
```

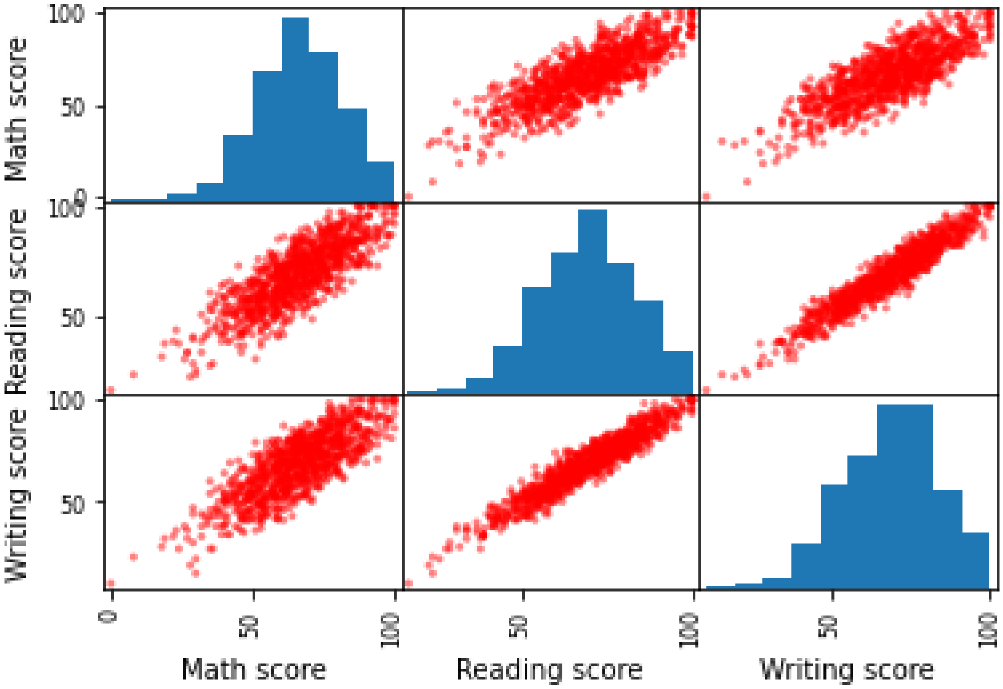
```
[89]: <AxesSubplot:xlabel='Reading score', ylabel='Writing score'>
```



8. Scatter matrix

```
[67]: pd.plotting.scatter_matrix(df, diagonal='hist', color = 'red')

      # The scatter matrix plot below is just the multivariate represntation of data.
      # Since there are three quantitative attributes therefore 9 squares were formed
      # (3*3). This shows the variation of all data with respect to each other.
```

```
[67]: array([[<AxesSubplot:xlabel='Math score', ylabel='Math score'>,
              <AxesSubplot:xlabel='Reading score', ylabel='Math score'>,
              <AxesSubplot:xlabel='Writing score', ylabel='Math score'>],
             [<AxesSubplot:xlabel='Math score', ylabel='Reading score'>,
              <AxesSubplot:xlabel='Reading score', ylabel='Reading score'>,
              <AxesSubplot:xlabel='Writing score', ylabel='Reading score'>],
             [<AxesSubplot:xlabel='Math score', ylabel='Writing score'>,
              <AxesSubplot:xlabel='Reading score', ylabel='Writing score'>,
              <AxesSubplot:xlabel='Writing score', ylabel='Writing score'>]],
            dtype=object)
```


Math score
Reading score
Writing score
100
50
0
100
50
100
50
0
50
100
0
50
100
50
100
50
100
Math score
Reading score
Writing score