STUDENTS' EXAM SCORES EXPLORATORY DATA ANALYSIS PROJECT

Objective of the project:

1) To get statistical insight of the dataset. 2) Gender distribution of the students. 3) Effect of parents' education on the score sheet of the students. 4) Effect of the type of lunch provided to the students on the score sheet of the students. 5) Effect of the marital status of the parents on the score card of the students. 6) Effect of weekly study hours on the marksheet of the students. 7) Distribution of the students according to various ethnic groups.

Tools used in this project

1) NumPy 2) Pandas 3) matplotlib 4) Seaborn

Importing the Dependencies

In [1]:

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

The Dataset

In [2]:

```
student_score=pd.read_csv(r'C:\Users\HP\Downloads\archive (4)\Expanded_data_with_more_features.csv')
```

In [3]:

```
student_score.head()
```

Out[3]:

	Unnamed: 0	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	PracticeSport	IsFirstChild	NrSiblings	Transport
0	0	female	NaN	bachelor's degree	standard	none	married	regularly	yes	3.0	schc
1	1	female	group C	some college	standard	NaN	married	sometimes	yes	0.0	
2	2	female	group B	master's degree	standard	none	single	sometimes	yes	4.0	schc
3	3	male	group A	associate's degree	free/reduced	none	married	never	no	1.0	
4	4	male	group C	some college	standard	none	married	sometimes	yes	0.0	schc

In [4]:

```
student_score.shape
```

Out[4]:

```
(30641, 15)
```

Statistical insight of the dataset

In [5]:

```
student_score.describe()
```

Out[5]:

	Unnamed: 0	NrSiblings	MathScore	ReadingScore	WritingScore
count	30641.000000	29069.000000	30641.000000	30641.000000	30641.000000
mean	499.556607	2.145894	66.558402	69.377533	68.418622
std	288.747894	1.458242	15.361616	14.758952	15.443525
min	0.000000	0.000000	0.000000	10.000000	4.000000
25%	249.000000	1.000000	56.000000	59.000000	58.000000
50%	500.000000	2.000000	67.000000	70.000000	69.000000
75%	750.000000	3.000000	78.000000	80.000000	79.000000
max	999.000000	7.000000	100.000000	100.000000	100.000000

It can be observed that the minimum marks obtained by the students is 0 but that for reading and writing are 10 and 4 respectively. The highest score obtained in mathematics,reading and writing are 100.From this insight we can say that some students are very poor in mathematics and this subject has to be taught more carefully.

Checking for null values

In [6]:

```
student_score.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30641 entries, 0 to 30640
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Unnamed: 0           30641 non-null  int64  
 1   Gender               30641 non-null  object 
 2   EthnicGroup          28801 non-null  object 
 3   ParentEduc           28796 non-null  object 
 4   LunchType            30641 non-null  object 
 5   TestPrep             28811 non-null  object 
 6   ParentMaritalStatus  29451 non-null  object 
 7   PracticeSport        30010 non-null  object 
 8   IsFirstChild         29737 non-null  object 
 9   NrSiblings           29069 non-null  float64
 10  TransportMeans       27507 non-null  object 
 11  WklyStudyHours       29686 non-null  object 
 12  MathScore            30641 non-null  int64  
 13  ReadingScore         30641 non-null  int64  
 14  WritingScore         30641 non-null  int64  
dtypes: float64(1), int64(4), object(10)
memory usage: 3.5+ MB
```

In [7]:

```
student_score.isnull().sum()
```

Out[7]:

```
Unnamed: 0                0
Gender                    0
EthnicGroup            1840
ParentEduc             1845
LunchType                 0
TestPrep               1830
ParentMaritalStatus    1190
PracticeSport           631
IsFirstChild            904
NrSiblings             1572
TransportMeans         3134
WklyStudyHours          955
MathScore                 0
ReadingScore              0
WritingScore              0
dtype: int64
```

Removing 'Unnamed: 0' column

In [9]:

```
student_score.drop(columns='Unnamed: 0',axis=1,inplace=True)
```

In [10]:

```
student_score.shape
```

Out[10]:

```
(30641, 14)
```

In [11]:

```
student_score.head()
```

Out[11]:

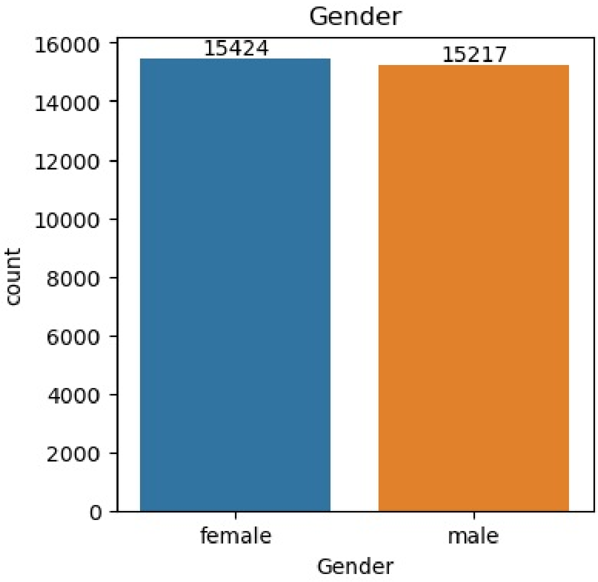
	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	PracticeSport	IsFirstChild	NrSiblings	TransportMeans	Wkl
0	female	NaN	bachelor's degree	standard	none	married	regularly	yes	3.0	school_bus	
1	female	group C	some college	standard	NaN	married	sometimes	yes	0.0	NaN	
2	female	group B	master's degree	standard	none	single	sometimes	yes	4.0	school_bus	
3	male	group A	associate's degree	free/reduced	none	married	never	no	1.0	NaN	
4	male	group C	some college	standard	none	married	sometimes	yes	0.0	school_bus	

Gender Distribution

In [56]:

```
plt.figure(figsize=(4,4))
ax=sns.countplot(data=student_score,x='Gender')
ax.bar_label(ax.containers[0])
```

```
plt.title('Gender')
plt.show()
```



There are 15424 female students and 15217 male students.

Effect of parents' education on the score sheet

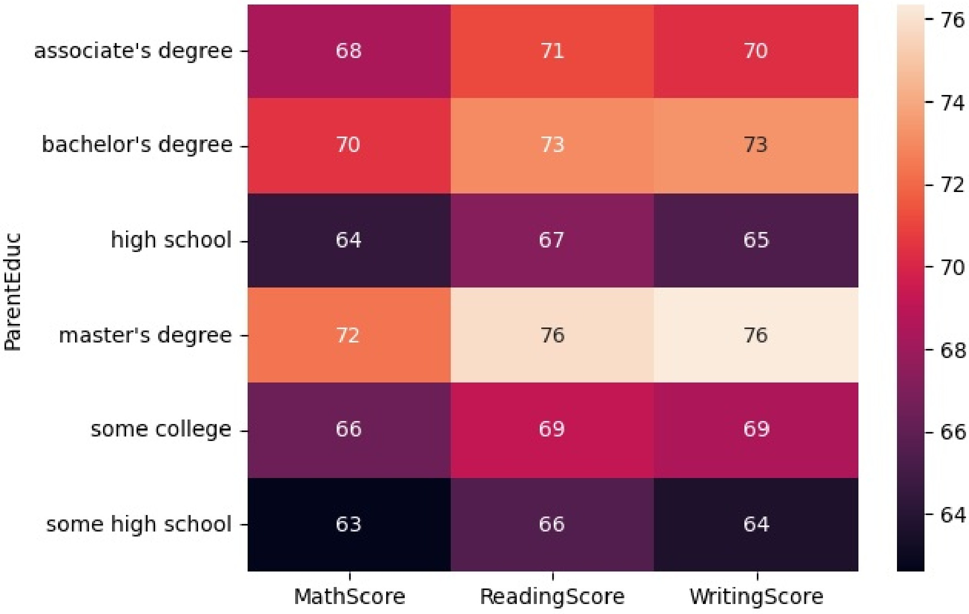
In [17]:
```
groupby_value=student_score.groupby("ParentEduc").agg({'MathScore':'mean','ReadingScore':'mean','WritingScore':
```

In [18]:
```
print(groupby_value)
```

```
                   MathScore  ReadingScore  WritingScore
ParentEduc
associate's degree  68.365586     71.124324     70.299099
bachelor's degree   70.466627     73.062020     73.331069
high school         64.435731     67.213997     65.421136
master's degree     72.336134     75.832921     76.356896
some college        66.390472     69.179708     68.501432
some high school    62.584013     65.510785     63.632409
```

In [20]:
```
sns.heatmap(groupby_value,annot=True)
plt.show
```

Out[20]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



From the above heatmap it can be concluded that those students whose parents have higher education degrees such as bachelor's degree, master's degree etc. score better than those students whose parents have lower educational background.

Effect of the type of lunch on the score sheet

In [21]:
```
groupby_value=student_score.groupby("LunchType").agg({'MathScore':'mean','ReadingScore':'mean','WritingScore':'
```

In [22]:

```
print(groupby_value)
```

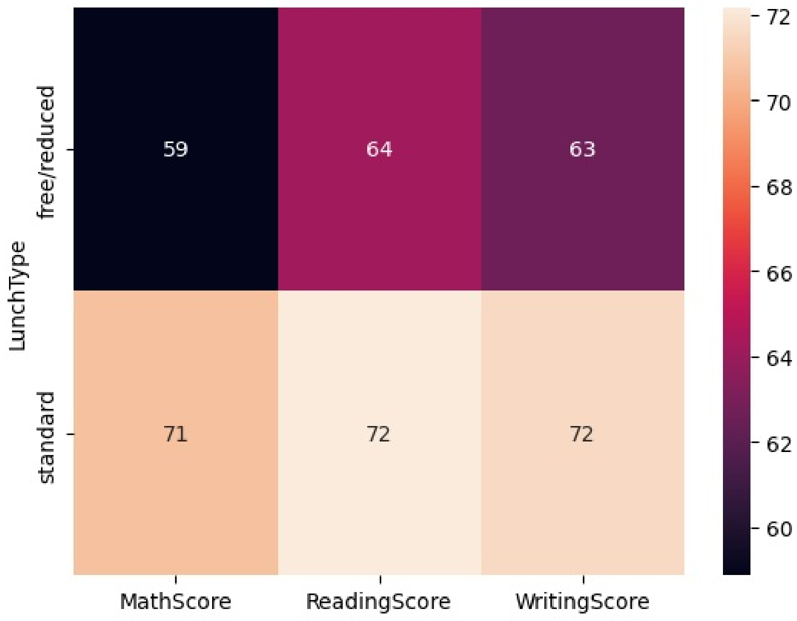
```
               MathScore  ReadingScore  WritingScore
LunchType
free/reduced   58.862332     64.189735     62.650522
standard       70.709370     72.175634     71.529716
```

In [23]:

```
sns.heatmap(groupby_value,annot=True)
plt.show
```

Out[23]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



It can be concluded that those students who are getting standard lunch score better than thos estudents who get free/reduced lunch.

Effect of parents' marital status on the score sheet of the students

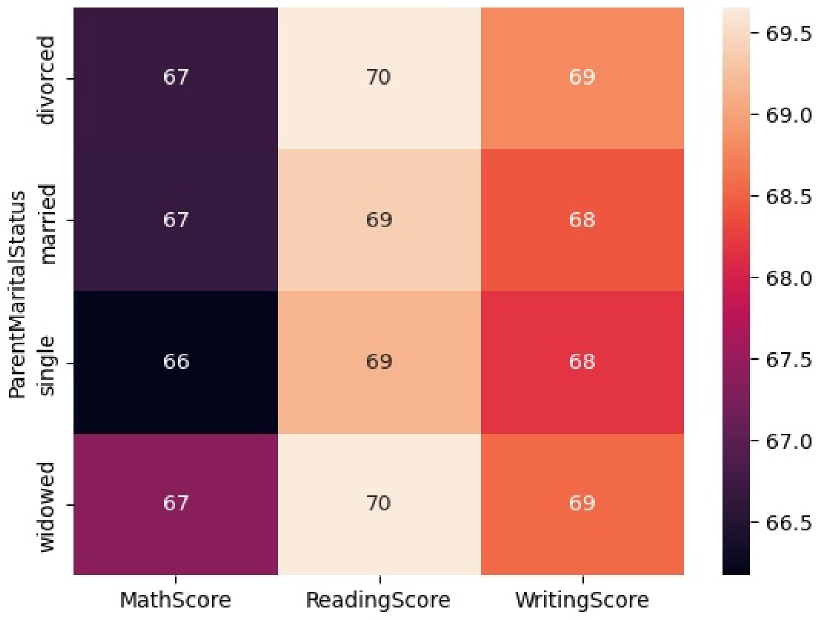
In [24]:

```
groupby_value=student_score.groupby("ParentMaritalStatus").agg({'MathScore':'mean','ReadingScore':'mean','Writi
print(groupby_value)
```

```
                     MathScore  ReadingScore  WritingScore
ParentMaritalStatus
divorced             66.691197     69.655011     68.799146
married              66.657326     69.389575     68.420981
single               66.165704     69.157250     68.174440
widowed              67.368866     69.651438     68.563452
```

In [25]:

```
sns.heatmap(groupby_value,annot=True)
plt.show()
```



The children of single parents are very poor in mathematics while they have score a better score in reading test and writing test. The children of widowed,married and divorced prents are also poor in mathematics but they are a little bit better in mathematics than the children of single parents. The children of widowed and divorced parents are very brilliant in reading ,and writing skill.

Effect of weekly study hours on the mark sheet

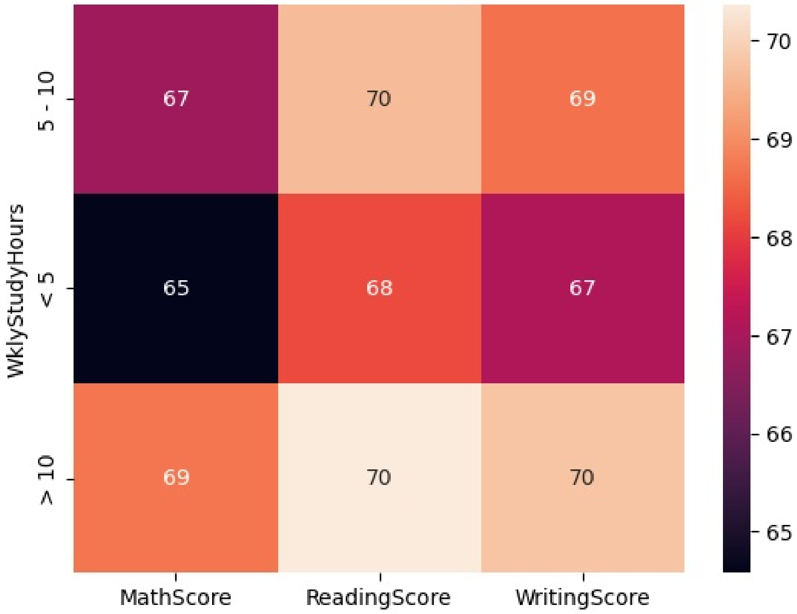
In [67]:

```
groupby_value=student_score.groupby("WklyStudyHours").agg({'MathScore':'mean','ReadingScore':'mean','WritingSco
print(groupby_value)
```

```
                MathScore  ReadingScore  WritingScore
WklyStudyHours
5 - 10          66.870491     69.660532     68.636280
< 5             64.580359     68.176135     67.090192
> 10            68.696655     70.365436     69.777778
```

In [68]:

```
sns.heatmap(groupby_value,annot=True)
plt.show()
```



From the above heatmap it can be concluded that those student who have weekly study hours of less than 5 hours are very poor in mathematics whereas those who have weekly study hour of (5-10) hours or more than 10 hours are good enogh in reading and writing.

Unique values in 'EthnicGroup'

In [30]:

```
print(student_score['EthnicGroup'].unique())
```

```
[nan 'group C' 'group B' 'group A' 'group D' 'group E']
```

Distribution of ethnic groups provided in the dataset

For ethnicity: group A

In [35]:

```
group_A=student_score.loc[(student_score['EthnicGroup']=='group A')].count()
```

In [33]:

```
print(group_A)
```

```
Gender                 2219
EthnicGroup            2219
ParentEduc             2078
LunchType              2219
TestPrep               2081
ParentMaritalStatus    2121
PracticeSport          2167
IsFirstChild           2168
NrSiblings             2096
TransportMeans         1999
WklyStudyHours         2146
MathScore              2219
ReadingScore           2219
WritingScore           2219
dtype: int64
```

For ethnicity: group B

```
In [36]: group_B=student_score.loc[(student_score['EthnicGroup']=='group B')].count()
```

```
In [37]: print(group_B)
```

```
Gender                 5826
EthnicGroup            5826
ParentEduc             5470
LunchType              5826
TestPrep               5488
ParentMaritalStatus    5605
PracticeSport          5704
IsFirstChild           5649
NrSiblings             5546
TransportMeans         5238
WklyStudyHours         5642
MathScore              5826
ReadingScore           5826
WritingScore           5826
dtype: int64
```

For ethnicity : group C

```
In [38]: group_C=student_score.loc[(student_score['EthnicGroup']=='group C')].count()
```

```
In [39]: print(group_C)
```

```
Gender                 9212
EthnicGroup            9212
ParentEduc             8652
LunchType              9212
TestPrep               8652
ParentMaritalStatus    8858
PracticeSport          9050
IsFirstChild           8929
NrSiblings             8763
TransportMeans         8280
WklyStudyHours         8933
MathScore              9212
ReadingScore           9212
WritingScore           9212
dtype: int64
```

For ethnicity : group D

```
In [40]: group_D=student_score.loc[(student_score['EthnicGroup']=='group D')].count()
```

```
In [41]: print(group_D)
```

```
Gender                 7503
EthnicGroup            7503
ParentEduc             7056
LunchType              7503
TestPrep               7070
ParentMaritalStatus    7218
PracticeSport          7343
IsFirstChild           7285
NrSiblings             7106
TransportMeans         6713
WklyStudyHours         7270
MathScore              7503
ReadingScore           7503
WritingScore           7503
dtype: int64
```

For ethnicity : group E

```
In [42]: group_E=student_score.loc[(student_score['EthnicGroup']=='group E')].count()
```

```
In [43]: print(group_E)
```

```
Gender                 4041
EthnicGroup            4041
ParentEduc             3814
LunchType              4041
TestPrep               3804
ParentMaritalStatus    3892
PracticeSport          3954
IsFirstChild           3918
NrSiblings             3820
TransportMeans         3624
WklyStudyHours         3924
MathScore              4041
ReadingScore           4041
WritingScore           4041
dtype: int64
```

Pie Chart

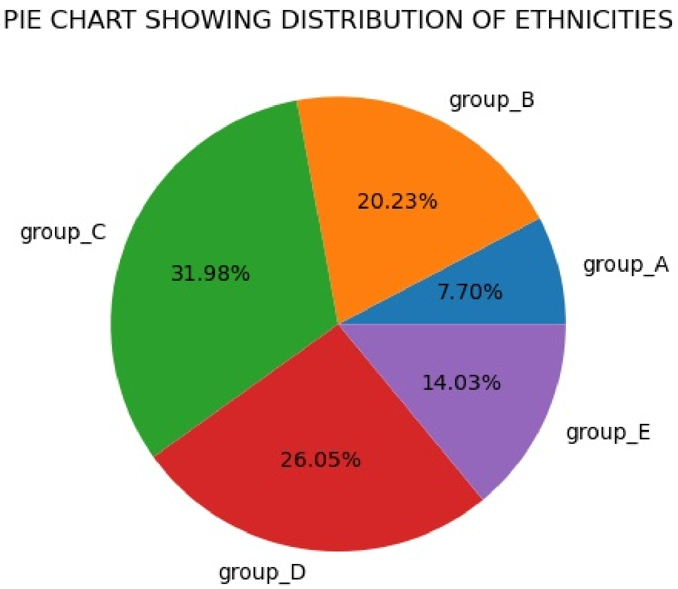
In [66]:

```
group_A=student_score.loc[(student_score['EthnicGroup']=='group A')].count()
group_B=student_score.loc[(student_score['EthnicGroup']=='group B')].count()
group_C=student_score.loc[(student_score['EthnicGroup']=='group C')].count()
group_D=student_score.loc[(student_score['EthnicGroup']=='group D')].count()
group_E=student_score.loc[(student_score['EthnicGroup']=='group E')].count()

l=['group_A','group_B','group_C','group_D','group_E']
mlist=[group_A['EthnicGroup'],group_B['EthnicGroup'],group_C['EthnicGroup'],group_D['EthnicGroup'],group_E['Eth
print(mlist)

plt.pie(mlist,labels=l,autopct="%1.2f%%")
plt.title('PIE CHART SHOWING DISTRIBUTION OF ETHNICITIES')
plt.show()
```

```
[2219, 5826, 9212, 7503, 4041]
```

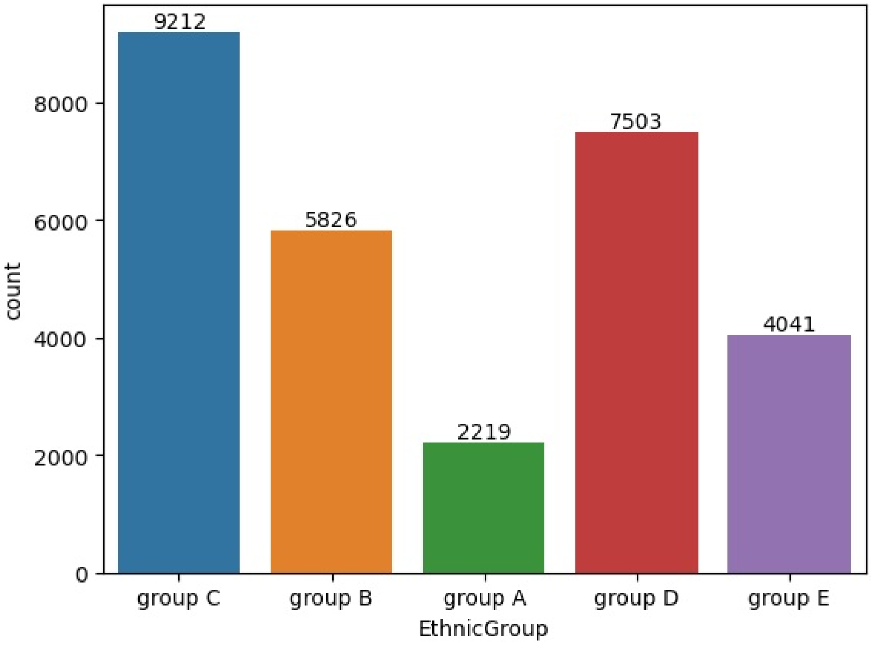


In [65]:

```
ax=sns.countplot(data=student_score,x='EthnicGroup')
ax.bar_label(ax.containers[0])
```

Out[65]:

```
[Text(0, 0, '9212'),
 Text(0, 0, '5826'),
 Text(0, 0, '2219'),
 Text(0, 0, '7503'),
 Text(0, 0, '4041')]
```



From the above graphs it is clearly observed that students belonging to ethnicity of group C are the highest in number.

Conclusion

1) It can be observed that the minimum marks obtained by the students is 0 but that for reading and writing are 10 and 4 respectively. The highest score obtained in mathematics,reading and writing are 100.From this insight we can say that some students are very poor in mathematics and this subject has to be taught more carefully. 2) There are 15424 female students and 15217 male students. 3) From the above heatmap it can be concluded that those students whose parents have higher education degrees such as bachelor's degree, master's degree etc. score better than those students whose parents have lower educational background. 4) It can be concluded that those students who are getting standard lunch score better than thos estudents who get free/reduced lunch. 5) The children of single parents are very poor in mathematics while they have score a better score in reading test and writing test. The children of widowed,married and divorced prents are also poor in mathematics but they are a little bit better in mathematics than the children of single parents. The children of widowed and divorced parents are very brilliant in reading ,and writing skill. 6) From the above heatmap it can be concluded that those student who have weekly study hours of less than 5 hours are very poor in mathematics whereas those who have weekly study hour of (5-10) hours or more than 10 hours are good enogh in reading and writing. 7) From the above graphs it is clearly observed that students belonging to ethnicity of group C are the highest in number.

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js