CAPSTONE PROJECT N.Aakash 21BTRCD045(DataScience)

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

In [11]:
```
df = pd.read_csv("D:/the_Boys_Mrksheet.csv")
df
```

Out[11]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NaN
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	NaN
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	NaN
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	NaN
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	NaN
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	NaN
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	NaN
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	NaN
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	NaN A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	NaN

In [12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Student_ID                 10 non-null     int64
 1   Gender                     10 non-null     object
 2   Age                        10 non-null     int64
 3   Contact Information        10 non-null     int64
 4   GPA                        10 non-null     float64
 5   Grade                      10 non-null     object
 6   Course Code                10 non-null     object
 7   Extracurricular Activities  10 non-null     object
 8   Unnamed: 8                 0 non-null      float64
 9   Nationality                10 non-null     object
dtypes: float64(2), int64(3), object(5)
memory usage: 928.0+ bytes
```

In [14]:
```
#DATA CLEANING
df.isnull().sum()
```

Out[14]:
```
Student_ID                    0
Gender                        0
Age                           0
Contact Information           0
GPA                           0
Grade                         0
Course Code                   0
Extracurricular Activities    0
Unnamed: 8                   10
Nationality                   0
dtype: int64
```

In [15]:
```
df.nunique()
```

Out[15]:
```
Student_ID                   10
Gender                        2
Age                           3
Contact Information          10
GPA                          10
Grade                         3
Course Code                   9
Extracurricular Activities    6
Unnamed: 8                    0
Nationality                   5
dtype: int64
```

In [16]:
```
df.fillna(10)
```

Out[16]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	10.0
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	10.0
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	10.0
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	10.0
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	10.0
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	10.0
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	10.0
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	10.0
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	10.0 A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	10.0

In [17]:
```
df.head()
```

Out[17]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8	N
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NaN	
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	NaN	
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	NaN	
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	NaN	
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	NaN	

In [18]:
```
df.dropna()
```

Out[18]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8	Nationa

In [19]:
```
df.dropna(axis=1)
```

Out[19]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Nationality
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NEPAL
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	INDIAN
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	UAE
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	INDIAN
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	INDIAN
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	NEPAL
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	NIGERIA
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	INDIAN
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	AFGANISTAN
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	INDIAN

In [29]:
```
df.dropna(thresh = 2)
```

Out[29]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8	
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NaN	
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	NaN	
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	NaN	
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	NaN	
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	NaN	
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	NaN	
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	NaN	
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	NaN	
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	NaN	A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	NaN	

In [30]:
```
df.fillna(1000)
```

Out[30]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8	
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	1000.0	
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	1000.0	
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	1000.0	
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	1000.0	
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	1000.0	
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	1000.0	
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	1000.0	
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	1000.0	
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	1000.0	A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	1000.0	

In [31]:
```
df.replace()
```

Out[31]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NaN
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	NaN
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	NaN
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	NaN
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	NaN
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	NaN
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	NaN
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	NaN
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	NaN A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	NaN

In [42]:
```
#DATA ENCODING
df['Gender '] = df['Gender '].astype('category')
print(df.dtypes)
```

```
Student_ID                    int64
Gender                     category
Age                           int64
Contact Information           int64
GPA                         float64
Grade                        object
Course Code                  object
Extracurricular Activities   object
Unnamed: 8                  float64
Nationality                  object
dtype: object
```

In [43]:
```
df
```

Out[43]:

	Student_ID	Gender	Age	Contact Information	GPA	Grade	Course Code	Extracurricular Activities	Unnamed: 8	
0	1	Male	20	7555420856	8.65	A	21CIC33	Anova Club	NaN	
1	2	Female	19	9055509046	8.44	B+	21CIC44	Cultral Club	NaN	
2	3	Male	20	9155509816	8.07	B	21CSEDS403	Football	NaN	
3	4	Female	19	9779858834	7.56	B	21CSEDS40L	Vollyball	NaN	
4	5	Male	19	8723475863	8.23	B+	21CSEDS404	Cricket	NaN	
5	6	Female	20	7555065432	8.01	B+	21CSSP41	Football	NaN	
6	7	Male	20	8555737464	8.53	A	21CSSP42	Cultral Club	NaN	
7	8	Female	18	9055560746	8.26	B+	21PC4ED46	Vollyball	NaN	
8	9	Male	19	9779853734	7.46	B	21CIC331	Anova Club	NaN	A
9	10	Female	20	9779855505	8.93	A	21CSSP42	GDSC Club	NaN	

In [47]:
```
Gender_dummies  = pd.get_dummies(df['Gender '], prefix = 'Gender ' )
df = pd.concat([df,Gender_dummies], axis = 1)
print(df)
```

```
   Student_ID Gender   Age  Contact Information   GPA Grade Course Code  \
0           1    Male   20           7555420856  8.65     A     21CIC33
1           2  Female   19           9055509046  8.44    B+     21CIC44
2           3    Male   20           9155509816  8.07     B  21CSEDS403
3           4  Female   19           9779858834  7.56     B  21CSEDS40L
4           5    Male   19           8723475863  8.23    B+  21CSEDS404
5           6  Female   20           7555065432  8.01    B+    21CSSP41
6           7    Male   20           8555737464  8.53     A    21CSSP42
7           8  Female   18           9055560746  8.26    B+   21PC4ED46
8           9    Male   19           9779853734  7.46     B    21CIC331
9          10  Female   20           9779855505  8.93     A    21CSSP42

  Extracurricular Activities  Unnamed: 8 Nationality  Gender _Female  \
0                Anova Club         NaN       NEPAL               0
1              Cultral Club         NaN      INDIAN               1
2                   Football        NaN         UAE               0
3                  Vollyball        NaN      INDIAN               1
4                    Cricket        NaN      INDIAN               0
5                   Football        NaN       NEPAL               1
6              Cultral Club         NaN     NIGERIA               0
7                  Vollyball        NaN      INDIAN               1
8                 Anova Club        NaN  AFGANISTAN               0
9                  GDSC Club        NaN      INDIAN               1

   Gender _Male  Gender _Female  Gender _Male  Gender _Female  Gender _Male
0             1               0             1               0             1
1             0               1             0               1             0
2             1               0             1               0             1
3             0               1             0               1             0
4             1               0             1               0             1
5             0               1             0               1             0
6             1               0             1               0             1
7             0               1             0               1             0
8             1               0             1               0             1
9             0               1             0               1             0
```

In [53]:

```
df['Grade'] = df['Grade'].astype('category')
df['Extracurricular Activities'] = df['Extracurricular Activities'].astype('categor
df['Nationality'] = df['Nationality'].astype('category')
print(df.dtypes)
```

```
Student_ID                      int64
Gender                       category
Age                             int64
Contact Information             int64
GPA                           float64
Grade                        category
Course Code                    object
Extracurricular Activities   category
Unnamed: 8                    float64
Nationality                  category
Gender _Female                  uint8
Gender _Male                    uint8
Gender _Female                  uint8
Gender _Male                    uint8
Gender _Female                  uint8
Gender _Male                    uint8
Grade_A                         uint8
Grade_B                         uint8
Grade_B+                        uint8
Extracurricular Activitie    category
dtype: object
```

In [48]:

```
Grade_dummies  = pd.get_dummies(df['Grade'], prefix = 'Grade' )
df = pd.concat([df,Grade_dummies], axis = 1)
print(df)
```

```
   Student_ID Gender   Age  Contact Information   GPA Grade Course Code  \
0           1    Male   20           7555420856  8.65     A     21CIC33
1           2  Female   19           9055509046  8.44    B+     21CIC44
2           3    Male   20           9155509816  8.07     B  21CSEDS403
3           4  Female   19           9779858834  7.56     B  21CSEDS40L
4           5    Male   19           8723475863  8.23    B+  21CSEDS404
5           6  Female   20           7555065432  8.01    B+    21CSSP41
6           7    Male   20           8555737464  8.53     A    21CSSP42
7           8  Female   18           9055560746  8.26    B+   21PC4ED46
8           9    Male   19           9779853734  7.46     B    21CIC331
9          10  Female   20           9779855505  8.93     A    21CSSP42

  Extracurricular Activities  Unnamed: 8 Nationality  Gender _Female  \
0                Anova Club         NaN       NEPAL               0
1              Cultral Club         NaN      INDIAN               1
2                   Football        NaN         UAE               0
3                  Vollyball        NaN      INDIAN               1
4                    Cricket        NaN      INDIAN               0
5                   Football        NaN       NEPAL               1
6              Cultral Club         NaN     NIGERIA               0
7                  Vollyball        NaN      INDIAN               1
8                 Anova Club        NaN  AFGANISTAN               0
9                  GDSC Club        NaN      INDIAN               1

   Gender _Male  Gender _Female  Gender _Male  Gender _Female  Gender _Male  \
0             1               0             1               0             1
1             0               1             0               1             0
2             1               0             1               0             1
3             0               1             0               1             0
4             1               0             1               0             1
5             0               1             0               1             0
6             1               0             1               0             1
7             0               1             0               1             0
8             1               0             1               0             1
9             0               1             0               1             0

   Grade_A  Grade_B  Grade_B+
0        1        0         0
1        0        0         1
2        0        1         0
3        0        1         0
4        0        0         1
5        0        0         1
6        1        0         0
7        0        0         1
8        0        1         0
9        1        0         0
```

In [54]:

```
ExtraCur_dummies  = pd.get_dummies(df['Extracurricular Activities'], prefix = 'Extr
df = pd.concat([df,ExtraCur_dummies], axis = 1)
print(df)
```

```
   Student_ID Gender   Age  Contact Information   GPA Grade Course Code  \
0           1    Male   20           7555420856  8.65     A     21CIC33
1           2  Female   19           9055509046  8.44    B+     21CIC44
2           3    Male   20           9155509816  8.07     B  21CSEDS403
3           4  Female   19           9779858834  7.56     B  21CSEDS40L
4           5    Male   19           8723475863  8.23    B+  21CSEDS404
5           6  Female   20           7555065432  8.01    B+    21CSSP41
6           7    Male   20           8555737464  8.53     A    21CSSP42
7           8  Female   18           9055560746  8.26    B+   21PC4ED46
8           9    Male   19           9779853734  7.46     B    21CIC331
9          10  Female   20           9779855505  8.93     A    21CSSP42

  Extracurricular Activities  Unnamed: 8 Nationality  ...  Grade_A  Grade_B  \
0                Anova Club         NaN       NEPAL  ...        1        0
1              Cultral Club         NaN      INDIAN  ...        0        0
2                  Football         NaN         UAE  ...        0        1
3                 Vollyball         NaN      INDIAN  ...        0        1
4                   Cricket         NaN      INDIAN  ...        0        0
5                  Football         NaN       NEPAL  ...        0        0
6              Cultral Club         NaN     NIGERIA  ...        1        0
7                 Vollyball         NaN      INDIAN  ...        0        0
8                Anova Club         NaN  AFGANISTAN  ...        0        1
9                 GDSC Club         NaN      INDIAN  ...        1        0

   Grade_B+  Extracurricular Activitie  Extracurricular Activities_Anova Club  \
0         0                Anova Club                                      1
1         1              Cultral Club                                      0
2         0                  Football                                      0
3         0                 Vollyball                                      0
4         1                   Cricket                                      0
5         1                  Football                                      0
6         0              Cultral Club                                      0
7         1                 Vollyball                                      0
8         0                Anova Club                                      1
9         0                 GDSC Club                                      0

   Extracurricular Activities_Cricket  \
0                                   0
1                                   0
2                                   0
3                                   0
4                                   1
5                                   0
6                                   0
7                                   0
8                                   0
9                                   0

   Extracurricular Activities_Cultral Club   \
0                                         0
1                                         1
2                                         0
3                                         0
4                                         0
5                                         0
6                                         1
```

```
7                                          0
8                                          0
9                                          0

   Extracurricular Activities_Football  Extracurricular Activities_GDSC Club  \
0                                    0                                     0
1                                    0                                     0
2                                    1                                     0
3                                    0                                     0
4                                    0                                     0
5                                    1                                     0
6                                    0                                     0
7                                    0                                     0
8                                    0                                     0
9                                    0                                     1

  Extracurricular Activities_Vollyball
0                                    0
1                                    0
2                                    0
3                                    1
4                                    0
5                                    0
6                                    0
7                                    1
8                                    0
9                                    0

[10 rows x 26 columns]
```

In [55]:
```
Nationality_dummies  = pd.get_dummies(df['Nationality'], prefix = 'Nationality' )
df = pd.concat([df,Nationality_dummies], axis = 1)
print(df)
```

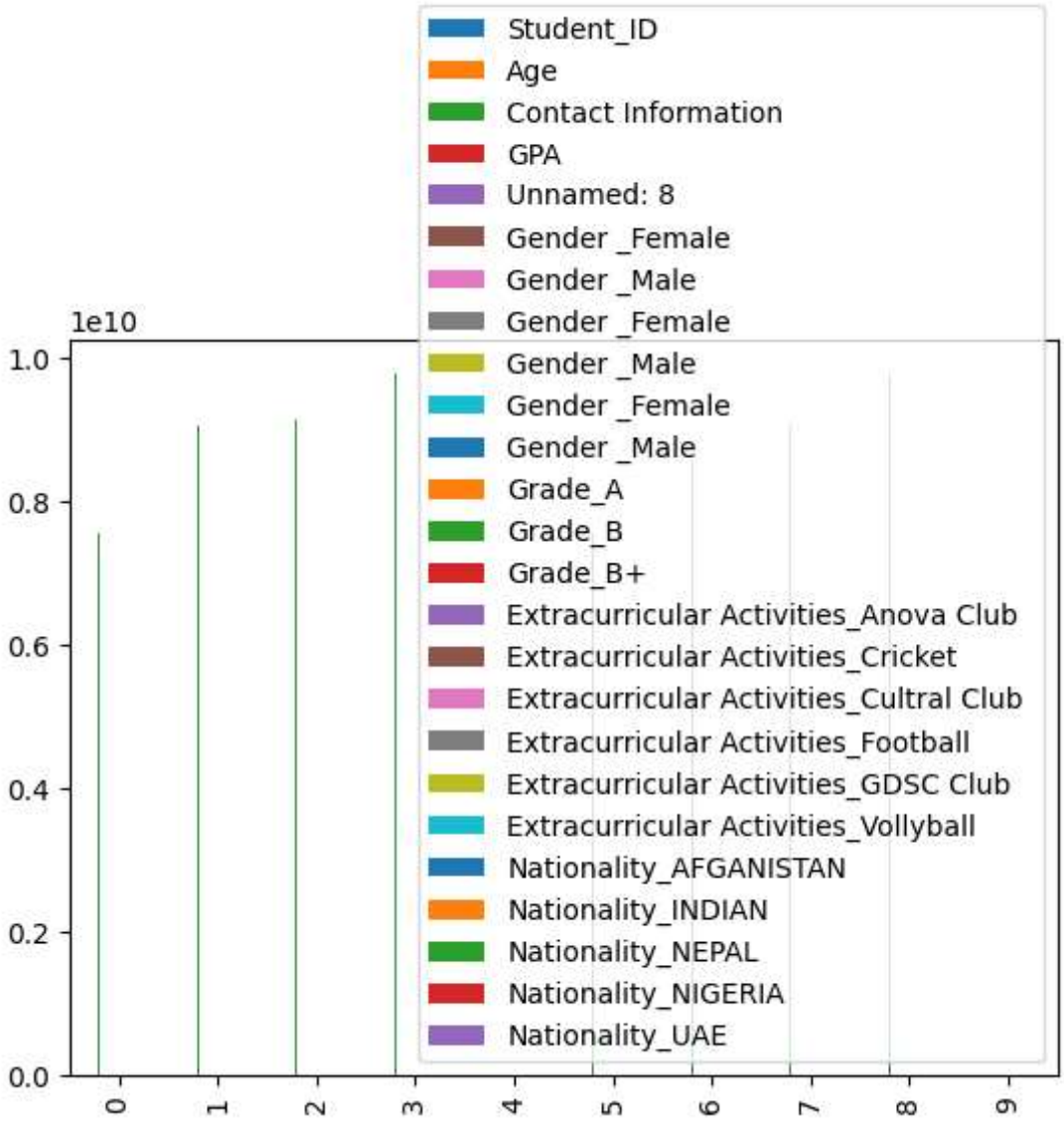
```
   Student_ID Gender   Age  Contact Information   GPA Grade Course Code  \
0           1    Male   20           7555420856  8.65     A     21CIC33
1           2  Female   19           9055509046  8.44    B+     21CIC44
2           3    Male   20           9155509816  8.07     B  21CSEDS403
3           4  Female   19           9779858834  7.56     B  21CSEDS40L
4           5    Male   19           8723475863  8.23    B+  21CSEDS404
5           6  Female   20           7555065432  8.01    B+    21CSSP41
6           7    Male   20           8555737464  8.53     A    21CSSP42
7           8  Female   18           9055560746  8.26    B+   21PC4ED46
8           9    Male   19           9779853734  7.46     B    21CIC331
9          10  Female   20           9779855505  8.93     A    21CSSP42

  Extracurricular Activities  Unnamed: 8 Nationality  ...  \
0                Anova Club         NaN       NEPAL  ...
1             Cultral Club          NaN      INDIAN  ...
2                  Football         NaN         UAE  ...
3                 Vollyball         NaN      INDIAN  ...
4                   Cricket         NaN      INDIAN  ...
5                  Football         NaN       NEPAL  ...
6             Cultral Club          NaN     NIGERIA  ...
7                 Vollyball         NaN      INDIAN  ...
8                Anova Club         NaN  AFGANISTAN  ...
9                 GDSC Club         NaN      INDIAN  ...

   Extracurricular Activities_Cricket  \
0                                   0
1                                   0
2                                   0
3                                   0
4                                   1
5                                   0
6                                   0
7                                   0
8                                   0
9                                   0

   Extracurricular Activities_Cultral Club   \
0                                         0
1                                         1
2                                         0
3                                         0
4                                         0
5                                         0
6                                         1
7                                         0
8                                         0
9                                         0

   Extracurricular Activities_Football  Extracurricular Activities_GDSC Club  \
0                                   0                                      0
1                                   0                                      0
2                                   1                                      0
3                                   0                                      0
4                                   0                                      0
5                                   1                                      0
6                                   0                                      0
```

```
7                                    0                                    0
8                                    0                                    0
9                                    0                                    1

   Extracurricular Activities_Vollyball  Nationality_AFGANISTAN  \
0                                     0                       0
1                                     0                       0
2                                     0                       0
3                                     1                       0
4                                     0                       0
5                                     0                       0
6                                     0                       0
7                                     1                       0
8                                     0                       1
9                                     0                       0

   Nationality_INDIAN  Nationality_NEPAL  Nationality_NIGERIA Nationality_UAE
0                   0                  1                    0               0
1                   1                  0                    0               0
2                   0                  0                    0               1
3                   1                  0                    0               0
4                   1                  0                    0               0
5                   0                  1                    0               0
6                   0                  0                    1               0
7                   1                  0                    0               0
8                   0                  0                    0               0
9                   1                  0                    0               0

[10 rows x 31 columns]
```

In [ ]:

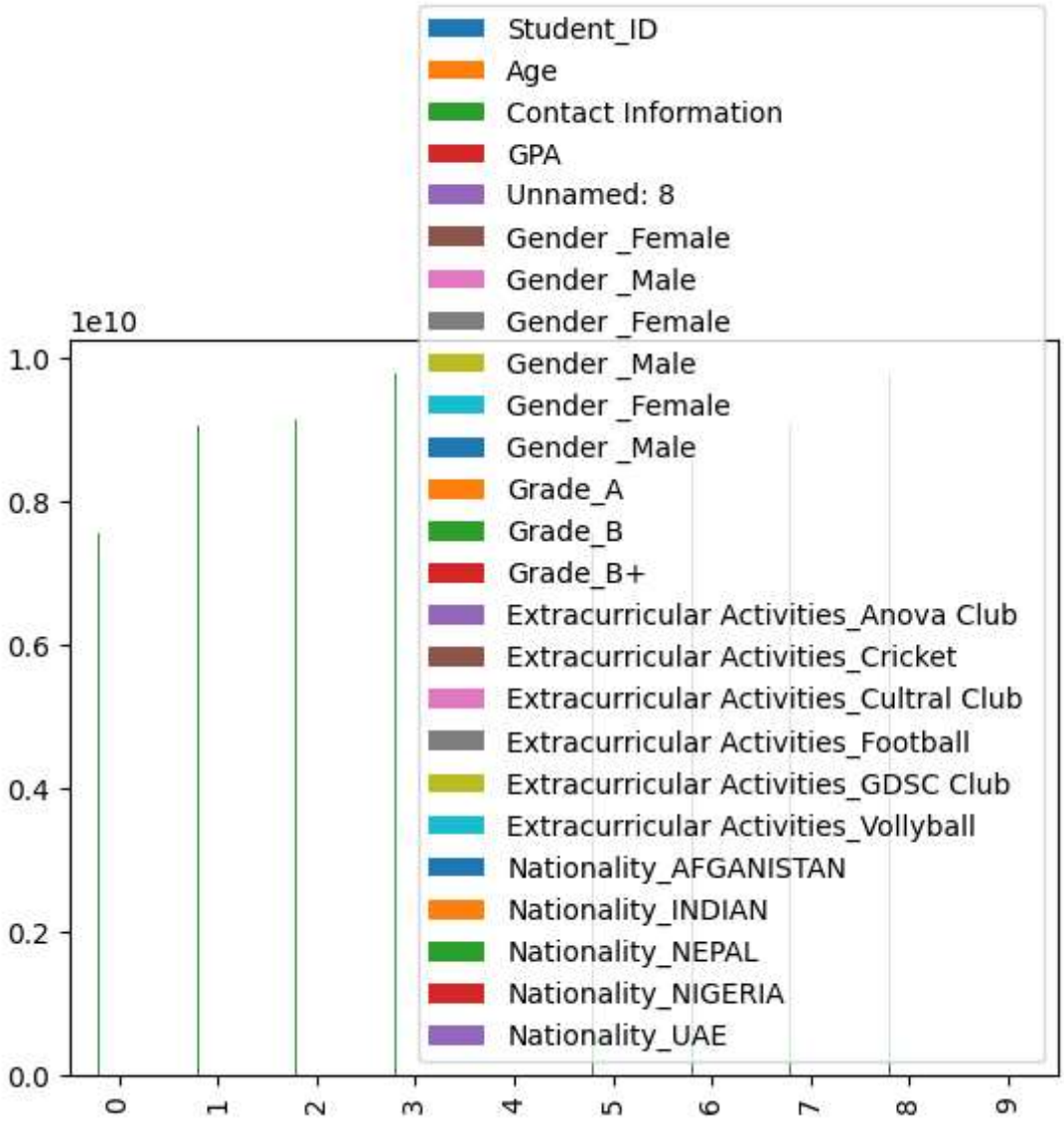
In [56]:
```
#DATA VISUALIZATION
df.plot(kind = 'bar')
```

Out[56]: <AxesSubplot: >



In [58]:
```
df.drop('Extracurricular Activities', axis = 1)
df.plot(kind = 'bar')
```

Out[58]: <AxesSubplot: >



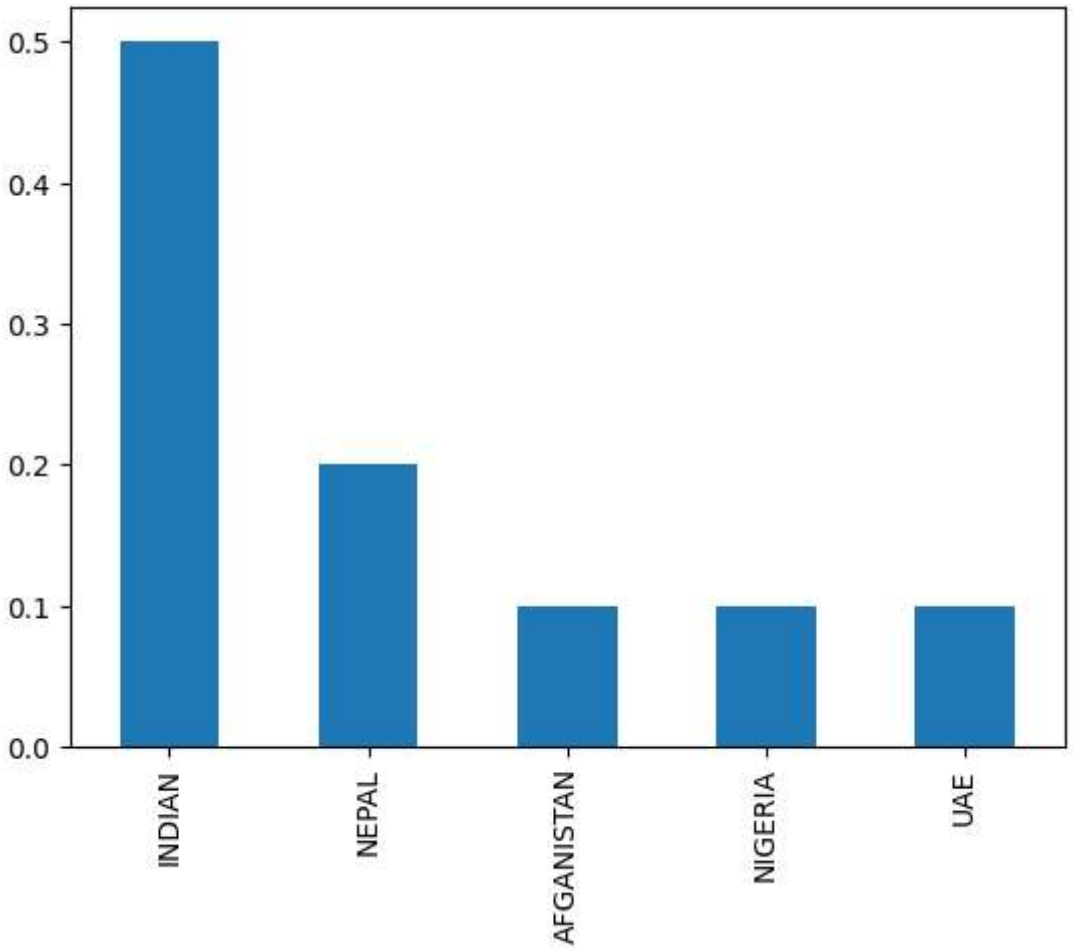
In [59]:
```
df['Nationality'].value_counts()
```

Out[59]:
```
INDIAN        5
NEPAL         2
AFGANISTAN    1
NIGERIA       1
UAE           1
Name: Nationality, dtype: int64
```

In [62]:
```
print('Percentage', df.Nationality.value_counts(normalize=True))
df.Nationality.value_counts(normalize=True).plot(kind='bar')
```

```
Percentage INDIAN        0.5
NEPAL         0.2
AFGANISTAN    0.1
NIGERIA       0.1
UAE           0.1
Name: Nationality, dtype: float64
```

Out[62]:
```
<AxesSubplot: >
```



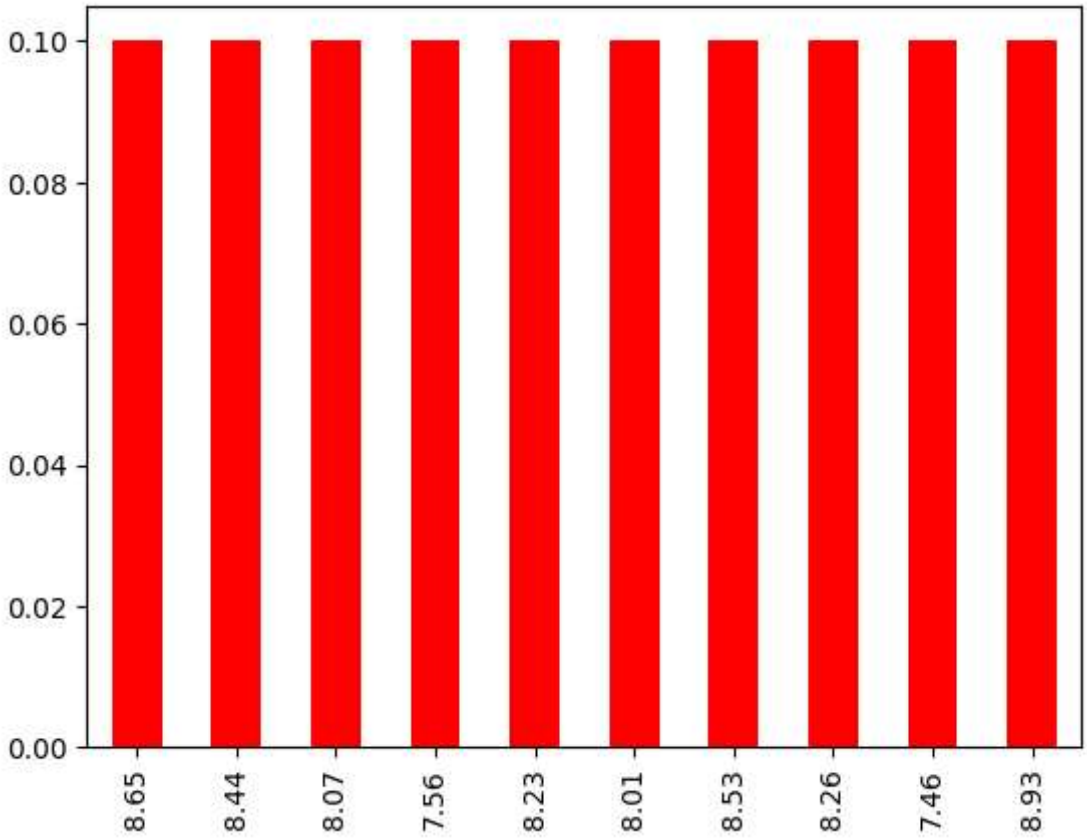
In [63]:
```
df['GPA'].value_counts()
```

Out[63]:
```
8.65    1
8.44    1
8.07    1
7.56    1
8.23    1
8.01    1
8.53    1
8.26    1
7.46    1
8.93    1
Name: GPA, dtype: int64
```

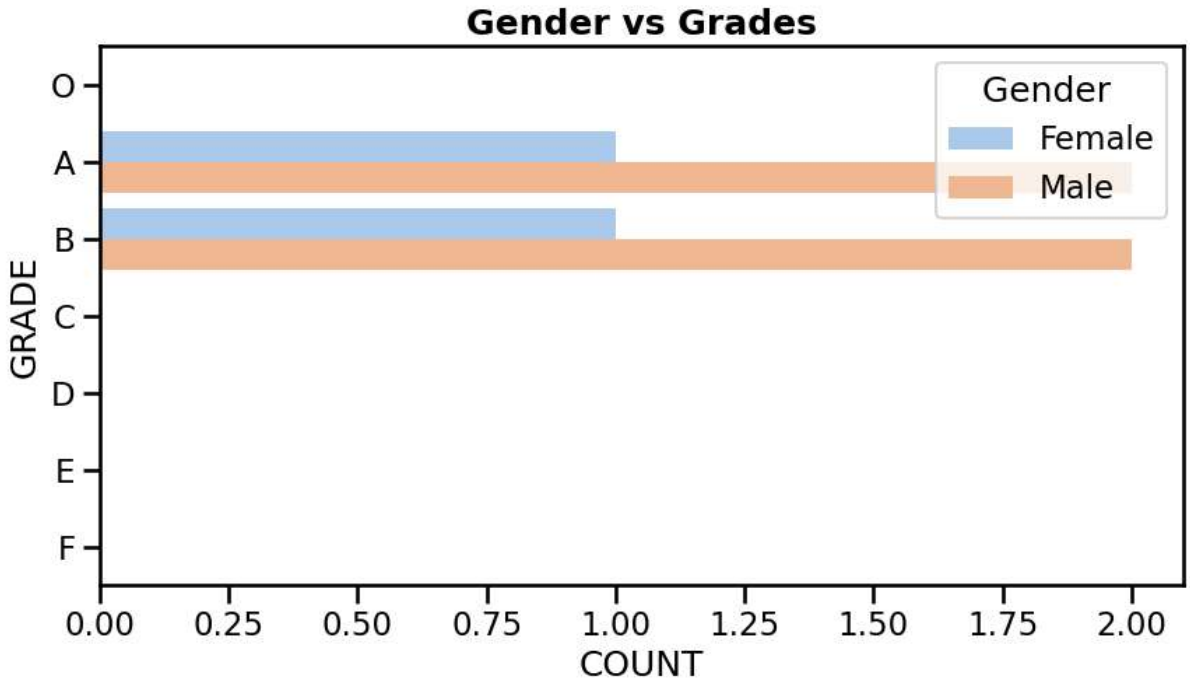
In [65]:
```
print('GPA', df.GPA.value_counts(normalize=True))
df.GPA.value_counts(normalize=True).plot(kind='bar', color = 'red')
```

```
GPA 8.65    0.1
8.44    0.1
8.07    0.1
7.56    0.1
8.23    0.1
8.01    0.1
8.53    0.1
8.26    0.1
7.46    0.1
8.93    0.1
Name: GPA, dtype: float64
```

Out[65]: <AxesSubplot: >



In [67]:
```
plt.figure(figsize=(10,5))
sns.set_context('talk', font_scale = 1)
sns.set_palette('pastel')
ax =sns.countplot(y= 'Grade', hue = 'Gender ', data = df, order = ['O', 'A', 'B', '
plt.title('Gender vs Grades', fontsize = 18, fontweight = 'bold')
ax.set(xlabel = 'COUNT', ylabel = 'GRADE')
plt.show()
```


Gender vs Grades
Gender
Female
Male
GRADE
O
A
B
C
D
E
F
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
COUNT