IMD Cleaning and Merging

November 16, 2020

1 Imports

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

2 Preprocessing

[2]:
```
fileName = 'student-mat.csv'

data = pd.read_csv(fileName)

data.head()
```

[2]:
```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ...  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
3     GP   F   15       U     GT3       T     4     2   health  services  ...
4     GP   F   16       U     GT3       T     3     3    other     other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        6   5   6   6
1      5        3      3     1     1      3        4   5   5   6
2      4        3      2     2     3      3       10   7   8  10
3      3        2      2     1     1      5        2  15  14  15
4      4        3      2     1     2      5        4   6  10  10

[5 rows x 33 columns]
```

[3]:
```
fileName = 'student-por.csv'

data2 = pd.read_csv(fileName)

data2.head()
```

[3]:

```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ...  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
3     GP   F   15       U     GT3       T     4     2   health  services  ...
4     GP   F   16       U     GT3       T     3     3    other     other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        4   0  11  11
1      5        3      3     1     1      3        2   9  11  11
2      4        3      2     2     3      3        6  12  13  12
3      3        2      2     1     1      5        0  14  14  14
4      4        3      2     1     2      5        0  11  13  13

[5 rows x 33 columns]
```

[4]:

```
print("Maths Students Missing Data: \n", data.isnull().sum())
```

```
Maths Students Missing Data:
 school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
```

```
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

[5]:
```
print("Portugese Students Missing Data: \n", data2.isnull().sum())
```

```
Portugese Students Missing Data:
 school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

[6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```

[7]:

```
#No nulls recorded. We're going to merge the two datasets based on their␣
 ↪classes.

data.insert(0, 'class', 'Maths')
data.head()

data2.insert(0, 'class', 'Portugese')
data2.head()
```

[7]:

```
       class school sex  age address famsize Pstatus  Medu  Fedu     Mjob  \
0  Portugese     GP   F   18       U     GT3       A     4     4  at_home   
1  Portugese     GP   F   17       U     GT3       T     1     1  at_home   
2  Portugese     GP   F   15       U     LE3       T     1     1  at_home   
3  Portugese     GP   F   15       U     GT3       T     4     2   health   
4  Portugese     GP   F   16       U     GT3       T     3     3    other   

   … famrel freetime goout  Dalc  Walc  health absences  G1  G2  G3
0  …      4        3     4     1     1       3        4   0  11  11
1  …      5        3     3     1     1       3        2   9  11  11
2  …      4        3     2     2     3       3        6  12  13  12
3  …      3        2     2     1     1       5        0  14  14  14
4  …      4        3     2     1     2       5        0  11  13  13

[5 rows x 34 columns]
```

Cool. Let's do a merge.

[8]:

```
student_merge = pd.concat([data, data2], axis=0)
```

[9]:

```
student_merge
```

[9]:

```
         class school sex  age address famsize Pstatus  Medu  Fedu       Mjob  \
0        Maths     GP   F   18       U     GT3       A     4     4    at_home   
1        Maths     GP   F   17       U     GT3       T     1     1    at_home   
2        Maths     GP   F   15       U     LE3       T     1     1    at_home   
3        Maths     GP   F   15       U     GT3       T     4     2     health   
4        Maths     GP   F   16       U     GT3       T     3     3      other   
..         …    …  ..  …     …     …     …   …   …       …
644  Portugese     MS   F   19       R     GT3       T     2     3   services   
645  Portugese     MS   F   18       U     LE3       T     3     1    teacher   
646  Portugese     MS   F   18       U     GT3       T     1     1      other   
647  Portugese     MS   M   17       U     LE3       T     3     1   services   
648  Portugese     MS   M   18       R     LE3       T     3     2   services   

     … famrel freetime goout  Dalc  Walc  health absences  G1  G2  G3
0    …      4        3     4     1     1       3        6   5   6   6
1    …      5        3     3     1     1       3        4   5   5   6
2    …      4        3     2     2     3       3       10   7   8  10
3    …      3        2     2     1     1       5        2  15  14  15
4    …      4        3     2     1     2       5        4   6  10  10
..   …    …      …   …   …   …     …      …  ..   ..  ..
644  …      5        4     2     1     2       5        4  10  11  10
645  …      4        3     4     1     1       1        4  15  15  16
646  …      1        1     1     1     1       5        6  11  12   9
647  …      2        4     5     3     4       2        6  10  10  10
648  …      4        4     1     3     4       5        4  10  11  11
```

```
[1044 rows x 34 columns]
```

[10]:

```
student_merge.reset_index(drop=True, inplace=True)
print(student_merge)
duplicate = student_merge[student_merge.duplicated()]
print(duplicate)
#No duplicate rows.
```

```
          class school sex  age address famsize Pstatus  Medu  Fedu      Mjob  \
0         Maths     GP    F   18        U     GT3         A     4     4  at_home
1         Maths     GP    F   17        U     GT3         T     1     1  at_home
2         Maths     GP    F   15        U     LE3         T     1     1  at_home
3         Maths     GP    F   15        U     GT3         T     4     2   health
4         Maths     GP    F   16        U     GT3         T     3     3    other
…         …    …  ..  …     …     …     …   …   …       …
1039  Portugese     MS    F   19        R     GT3         T     2     3 services
1040  Portugese     MS    F   18        U     LE3         T     3     1  teacher
1041  Portugese     MS    F   18        U     GT3         T     1     1    other
1042  Portugese     MS    M   17        U     LE3         T     3     1 services
1043  Portugese     MS    M   18        R     LE3         T     3     2 services

      … famrel freetime goout  Dalc  Walc  health absences  G1  G2  G3
0     …      4        3     4     1     1       3        6   5   6   6
1     …      5        3     3     1     1       3        4   5   5   6
2     …      4        3     2     2     3       3       10   7   8  10
3     …      3        2     2     1     1       5        2  15  14  15
4     …      4        3     2     1     2       5        4   6  10  10
…   …    …      …   …   …   …     …      …  ..  ..  ..
1039  …      5        4     2     1     2       5        4  10  11  10
1040  …      4        3     4     1     1       1        4  15  15  16
1041  …      1        1     1     1     1       5        6  11  12   9
1042  …      2        4     5     3     4       2        6  10  10  10
1043  …      4        4     1     3     4       5        4  10  11  11

[1044 rows x 34 columns]
Empty DataFrame
Columns: [class, school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob,
Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup,
paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout,
Dalc, Walc, health, absences, G1, G2, G3]
Index: []

[0 rows x 34 columns]
```

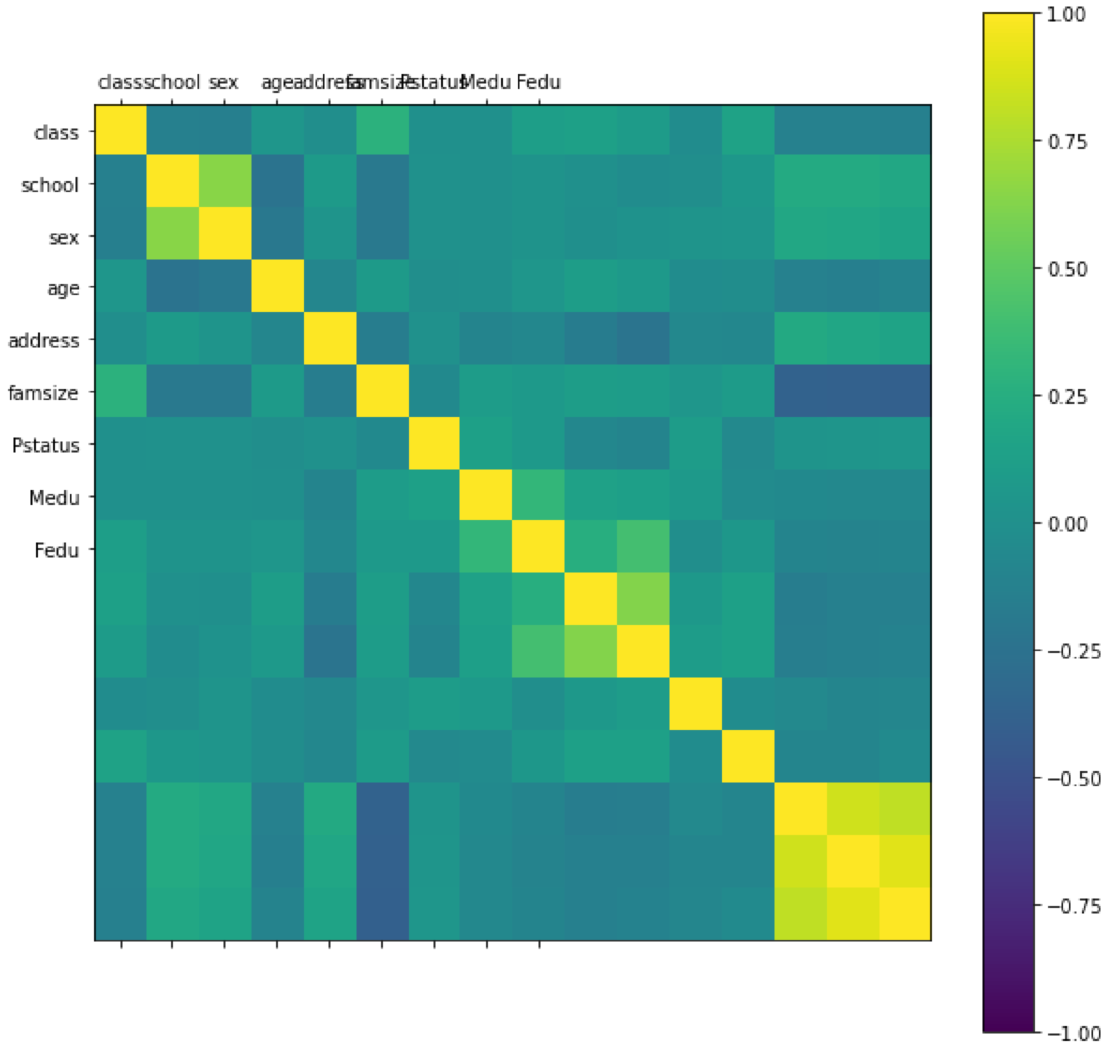
[ ]:

```

```

3 Minor EDA

[11]:
```
namesCol = list(student_merge.columns)


correlations = student_merge.corr()
print("Correlation Matrix")
# plot correlation matrix
fig = plt.figure(figsize=(10,10))
ax = fig.add_subplot(111)
cax = ax.matshow(correlations, vmin=-1, vmax=1)
fig.colorbar(cax)
ticks = np.arange(0,9,1)
ax.set_xticks(ticks)
ax.set_yticks(ticks)
ax.set_xticklabels(namesCol)
ax.set_yticklabels(namesCol)
plt.show()
print("Correlation Table")
# Correlation Table, note this does not export easily
corr = student_merge.corr()
corr.style.background_gradient().set_precision(2)
```

```
Correlation Matrix
```



```
Correlation Table
```

```
[11]: <pandas.io.formats.style.Styler at 0x22a0e90d970>
```

```
[ ]:
```