Intro to Data Science - EDA - Correlation

March 27, 2020

[13]:
```
# This Python 3 environment comes with many helpful analytics libraries␣
 ↪installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/
 ↪docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list␣
 ↪all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Any results you write to the current directory are saved as output.
```

```
/kaggle/input/default-of-credit-card-clients-dataset/UCI_Credit_Card.csv
```

0.1 ### Correlation and Heatmaps

Correlation is a mathematical technique that shows how strongly two variables are linked. It quantifies the strength of the relationship. For instance, we know that the weight and height of a person are correlated. Taller people tend to have more weight. Hence, we say that height and weight are correlated.

Correlation is measured in terms of a number called correlation coefficient, which ranges from -1 to 1. The value of 1 or -1 denotes complete correlation, while 0 indicates that no correlation is present between the two variables. Negative values mean there is an inverse relationship between the two variables, while a positive value denotes a direct relationship.

[14]:
```
df = pd.read_csv('/kaggle/input/default-of-credit-card-clients-dataset/
 ↪UCI_Credit_Card.csv')
correlation = df.corr()
print(correlation)
```

```
                                  ID  LIMIT_BAL       SEX  EDUCATION  \
ID                          1.000000   0.026179  0.018497   0.039177
LIMIT_BAL                   0.026179   1.000000  0.024755  -0.219161
SEX                         0.018497   0.024755  1.000000   0.014232
EDUCATION                   0.039177  -0.219161  0.014232   1.000000
MARRIAGE                   -0.029079  -0.108139 -0.031389  -0.143464
AGE                         0.018678   0.144713 -0.090874   0.175061
PAY_0                      -0.030575  -0.271214 -0.057643   0.105364
PAY_2                      -0.011215  -0.296382 -0.070771   0.121566
PAY_3                      -0.018494  -0.286123 -0.066096   0.114025
PAY_4                      -0.002735  -0.267460 -0.060173   0.108793
PAY_5                      -0.022199  -0.249411 -0.055064   0.097520
PAY_6                      -0.020270  -0.235195 -0.044008   0.082316
BILL_AMT1                   0.019389   0.285430 -0.033642   0.023581
BILL_AMT2                   0.017982   0.278314 -0.031183   0.018749
BILL_AMT3                   0.024354   0.283236 -0.024563   0.013002
BILL_AMT4                   0.040351   0.293988 -0.021880  -0.000451
BILL_AMT5                   0.016705   0.295562 -0.017005  -0.007567
BILL_AMT6                   0.016730   0.290389 -0.016733  -0.009099
PAY_AMT1                    0.009742   0.195236 -0.000242  -0.037456
PAY_AMT2                    0.008406   0.178408 -0.001391  -0.030038
PAY_AMT3                    0.039151   0.210167 -0.008597  -0.039943
PAY_AMT4                    0.007793   0.203242 -0.002229  -0.038218
PAY_AMT5                    0.000652   0.217202 -0.001667  -0.040358
PAY_AMT6                    0.003000   0.219595 -0.002766  -0.037200
default.payment.next.month -0.013952  -0.153520 -0.039961   0.028006

                            MARRIAGE       AGE     PAY_0     PAY_2     PAY_3  \
ID                         -0.029079  0.018678 -0.030575 -0.011215 -0.018494
LIMIT_BAL                  -0.108139  0.144713 -0.271214 -0.296382 -0.286123
SEX                        -0.031389 -0.090874 -0.057643 -0.070771 -0.066096
EDUCATION                  -0.143464  0.175061  0.105364  0.121566  0.114025
MARRIAGE                    1.000000 -0.414170  0.019917  0.024199  0.032688
AGE                        -0.414170  1.000000 -0.039447 -0.050148 -0.053048
PAY_0                       0.019917 -0.039447  1.000000  0.672164  0.574245
PAY_2                       0.024199 -0.050148  0.672164  1.000000  0.766552
PAY_3                       0.032688 -0.053048  0.574245  0.766552  1.000000
PAY_4                       0.033122 -0.049722  0.538841  0.662067  0.777359
PAY_5                       0.035629 -0.053826  0.509426  0.622780  0.686775
PAY_6                       0.034345 -0.048773  0.474553  0.575501  0.632684
BILL_AMT1                  -0.023472  0.056239  0.187068  0.234887  0.208473
BILL_AMT2                  -0.021602  0.054283  0.189859  0.235257  0.237295
BILL_AMT3                  -0.024909  0.053710  0.179785  0.224146  0.227494
BILL_AMT4                  -0.023344  0.051353  0.179125  0.222237  0.227202
BILL_AMT5                  -0.025393  0.049345  0.180635  0.221348  0.225145
BILL_AMT6                  -0.021207  0.047613  0.176980  0.219403  0.222327
PAY_AMT1                   -0.005979  0.026147 -0.079269 -0.080701  0.001295
PAY_AMT2                   -0.008093  0.021785 -0.070101 -0.058990 -0.066793
```

```
PAY_AMT3                   -0.003541  0.029247 -0.070561 -0.055901 -0.053311
PAY_AMT4                   -0.012659  0.021379 -0.064005 -0.046858 -0.046067
PAY_AMT5                   -0.001205  0.022850 -0.058190 -0.037093 -0.035863
PAY_AMT6                   -0.006641  0.019478 -0.058673 -0.036500 -0.035861
default.payment.next.month -0.024339  0.013890  0.324794  0.263551  0.235253

                               PAY_4  …  BILL_AMT4  BILL_AMT5  BILL_AMT6  \
ID                         -0.002735  …   0.040351   0.016705   0.016730
LIMIT_BAL                  -0.267460  …   0.293988   0.295562   0.290389
SEX                        -0.060173  …  -0.021880  -0.017005  -0.016733
EDUCATION                   0.108793  …  -0.000451  -0.007567  -0.009099
MARRIAGE                    0.033122  …  -0.023344  -0.025393  -0.021207
AGE                        -0.049722  …   0.051353   0.049345   0.047613
PAY_0                       0.538841  …   0.179125   0.180635   0.176980
PAY_2                       0.662067  …   0.222237   0.221348   0.219403
PAY_3                       0.777359  …   0.227202   0.225145   0.222327
PAY_4                       1.000000  …   0.245917   0.242902   0.239154
PAY_5                       0.819835  …   0.271915   0.269783   0.262509
PAY_6                       0.716449  …   0.266356   0.290894   0.285091
BILL_AMT1                   0.202812  …   0.860272   0.829779   0.802650
BILL_AMT2                   0.225816  …   0.892482   0.859778   0.831594
BILL_AMT3                   0.244983  …   0.923969   0.883910   0.853320
BILL_AMT4                   0.245917  …   1.000000   0.940134   0.900941
BILL_AMT5                   0.242902  …   0.940134   1.000000   0.946197
BILL_AMT6                   0.239154  …   0.900941   0.946197   1.000000
PAY_AMT1                   -0.009362  …   0.233012   0.217031   0.199965
PAY_AMT2                   -0.001944  …   0.207564   0.181246   0.172663
PAY_AMT3                   -0.069235  …   0.300023   0.252305   0.233770
PAY_AMT4                   -0.043461  …   0.130191   0.293118   0.250237
PAY_AMT5                   -0.033590  …   0.160433   0.141574   0.307729
PAY_AMT6                   -0.026565  …   0.177637   0.164184   0.115494
default.payment.next.month  0.216614  …  -0.010156  -0.006760  -0.005372

                            PAY_AMT1  PAY_AMT2  PAY_AMT3  PAY_AMT4  PAY_AMT5  \
ID                          0.009742  0.008406  0.039151  0.007793  0.000652
LIMIT_BAL                   0.195236  0.178408  0.210167  0.203242  0.217202
SEX                        -0.000242 -0.001391 -0.008597 -0.002229 -0.001667
EDUCATION                  -0.037456 -0.030038 -0.039943 -0.038218 -0.040358
MARRIAGE                   -0.005979 -0.008093 -0.003541 -0.012659 -0.001205
AGE                         0.026147  0.021785  0.029247  0.021379  0.022850
PAY_0                      -0.079269 -0.070101 -0.070561 -0.064005 -0.058190
PAY_2                      -0.080701 -0.058990 -0.055901 -0.046858 -0.037093
PAY_3                       0.001295 -0.066793 -0.053311 -0.046067 -0.035863
PAY_4                      -0.009362 -0.001944 -0.069235 -0.043461 -0.033590
PAY_5                      -0.006089 -0.003191  0.009062 -0.058299 -0.033337
PAY_6                      -0.001496 -0.005223  0.005834  0.019018 -0.046434
BILL_AMT1                   0.140277  0.099355  0.156887  0.158303  0.167026
BILL_AMT2                   0.280365  0.100851  0.150718  0.147398  0.157957
```

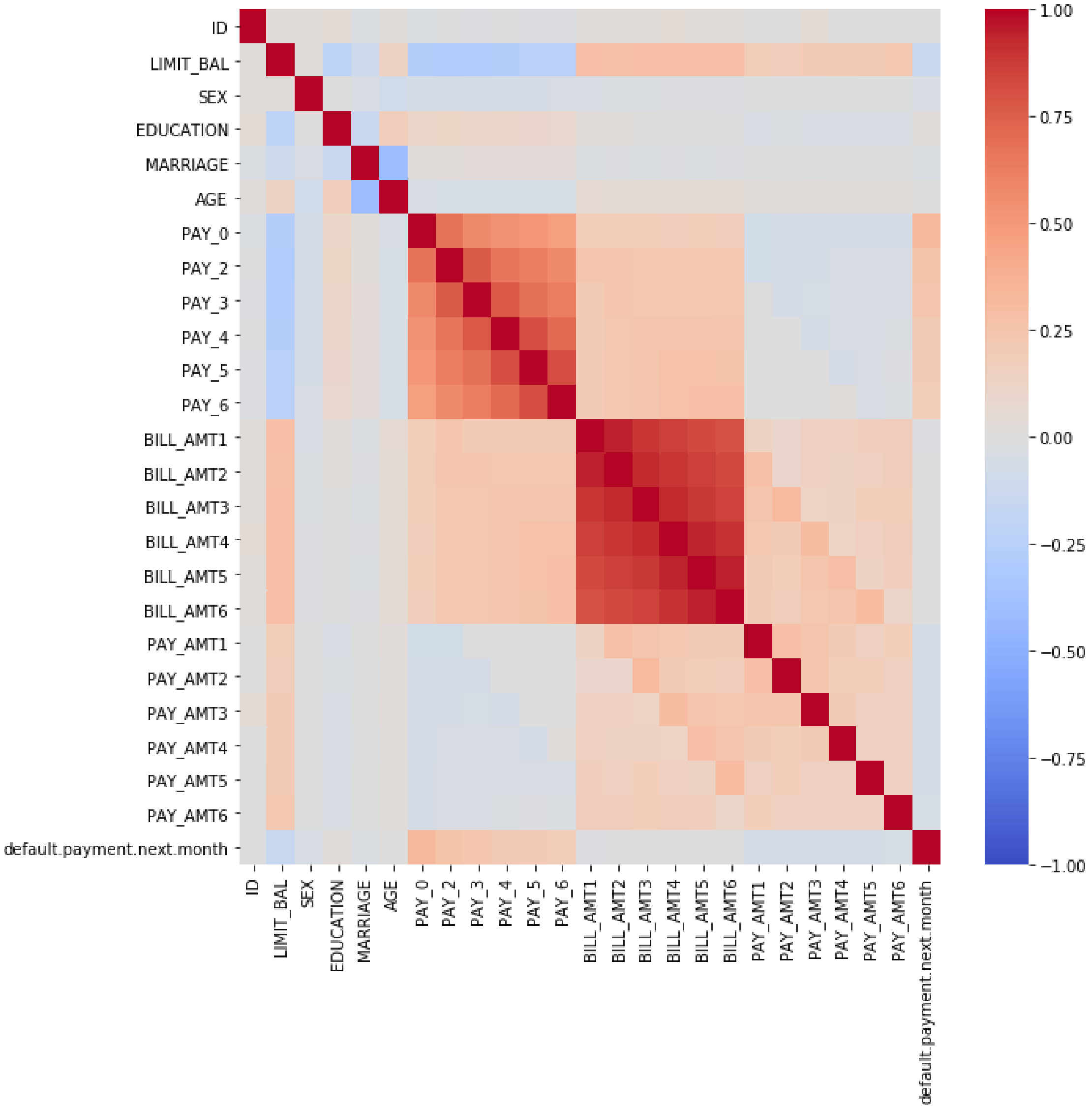
```
BILL_AMT3                   0.244335  0.316936  0.130011  0.143405  0.179712
BILL_AMT4                   0.233012  0.207564  0.300023  0.130191  0.160433
BILL_AMT5                   0.217031  0.181246  0.252305  0.293118  0.141574
BILL_AMT6                   0.199965  0.172663  0.233770  0.250237  0.307729
PAY_AMT1                    1.000000  0.285576  0.252191  0.199558  0.148459
PAY_AMT2                    0.285576  1.000000  0.244770  0.180107  0.180908
PAY_AMT3                    0.252191  0.244770  1.000000  0.216325  0.159214
PAY_AMT4                    0.199558  0.180107  0.216325  1.000000  0.151830
PAY_AMT5                    0.148459  0.180908  0.159214  0.151830  1.000000
PAY_AMT6                    0.185735  0.157634  0.162740  0.157834  0.154896
default.payment.next.month -0.072929 -0.058579 -0.056250 -0.056827 -0.055124

                            PAY_AMT6  default.payment.next.month
ID                          0.003000                   -0.013952
LIMIT_BAL                   0.219595                   -0.153520
SEX                        -0.002766                   -0.039961
EDUCATION                  -0.037200                    0.028006
MARRIAGE                   -0.006641                   -0.024339
AGE                         0.019478                    0.013890
PAY_0                      -0.058673                    0.324794
PAY_2                      -0.036500                    0.263551
PAY_3                      -0.035861                    0.235253
PAY_4                      -0.026565                    0.216614
PAY_5                      -0.023027                    0.204149
PAY_6                      -0.025299                    0.186866
BILL_AMT1                   0.179341                   -0.019644
BILL_AMT2                   0.174256                   -0.014193
BILL_AMT3                   0.182326                   -0.014076
BILL_AMT4                   0.177637                   -0.010156
BILL_AMT5                   0.164184                   -0.006760
BILL_AMT6                   0.115494                   -0.005372
PAY_AMT1                    0.185735                   -0.072929
PAY_AMT2                    0.157634                   -0.058579
PAY_AMT3                    0.162740                   -0.056250
PAY_AMT4                    0.157834                   -0.056827
PAY_AMT5                    0.154896                   -0.055124
PAY_AMT6                    1.000000                   -0.053183
default.payment.next.month -0.053183                    1.000000

[25 rows x 25 columns]
```

we can see here a list of correlation value, it is very difficult to study let's create a **Heatmap** of correlation that is much easier to study. A **heatmap** is a graphical representation of data where individual values are represented as colors. The intensity of the colors indicates the values.

[18]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# correlation table
corr = df.corr()
#plot heatmap
#vmin = minimum number of color
#vmax = maximum number of color
plt.figure(figsize=(10,10))
sns.heatmap(data=corr, vmin=-1, vmax=1, cmap='coolwarm')
```

```
[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6320db22e8>
```



let’s study on this heatmap,

1. we can see **BILL_AMT1, BILL_AMT2 .... BILL_AMT6** are positively correlated with each other.
2. **LIMIT_BAL** is has positive correlation with **BILL_AMT1, BILL_AMT2 .... BILL_AMT6**, which implies that people who were given more credit (higher values of LIMIT_BAL) tend to have larger bills.
3. **LIMIT_BAL** has negative correlation with payment delay variables **PAY_0, ... PAY_6** which implies higher limite balance tends to fewer payment delay.
4. There are not correlation in **SEX**, **MARRIAGE**.