academic-data-eda

January 17, 2024

[1]:
```
# load the data

import pandas as pd

acad_data = pd.read_csv('DBS.csv')
acad_data.head()
```

[1]:
```
  access;tests;tests_grade;exam;project;project_grade;assignments;result_points;
result_grade;graduate;year;acad_year
0  1256;57;A;19;91.54;A;40;189.92;A;1;2019;2019/2020
1  985;42.87;B;19;75.96;A;13.7;189.43;A;1;2017;20…
2  1455;54.5;A;16;96.79;A;40;188.91;A;1;2019;2019…
3  998;54.5;A;16;93.36;A;40;186.85;A;1;2019;2019/…
4  1347;55;A;16;92.86;A;39;186.38;A;1;2019;2019/2020
```

[2]:
```
acad_data[['access','tests','tests_grade','exam','project','project_grade','assignments','resu
    'result_grade','graduate','year','acad_year']] = acad_data['access;tests;
 ↪tests_grade;exam;project;project_grade;assignments;result_points;
 ↪result_grade;graduate;year;acad_year'].str.split(';', expand=True)
```

[3]:
```
acad_data = acad_data.drop(['access;tests;tests_grade;exam;project;
 ↪project_grade;assignments;result_points;result_grade;graduate;year;
 ↪acad_year'],axis=1)
acad_data
```

[3]:

	access	tests	tests_grade	exam	project	project_grade	assignments	\
0	1256	57	A	19	91.54	A	40	
1	985	42.87	B	19	75.96	A	13.7	
2	1455	54.5	A	16	96.79	A	40	
3	998	54.5	A	16	93.36	A	40	
4	1347	55	A	16	92.86	A	39	
..	…	…	…	…	…	…	…	
256	340	0	FX	0	0	FX	0	
257	429	0	FX	0	0	FX	0	
258	26	0	FX	0	0	FX	0	
259	126	0	FX	0	0	FX	0	
260	28	0	0	0	0	0	0	

```
     result_points result_grade graduate  year  acad_year
0           189.92            A        1  2019  2019/2020
1           189.43            A        1  2017  2017/2018
2           188.91            A        1  2019  2019/2020
3           186.85            A        1  2019  2019/2020
4           186.38            A        1  2019  2019/2020
..             …          …      …   …       …
256              0           FX        0  2016  2016/2017
257              0           FX        0  2016  2016/2017
258              0           FX        0  2018  2018/2019
259              0           FX        0  2018  2018/2019
260              0           FX        0  2019  2019/2020

[261 rows x 12 columns]
```

The educational data used in this project represents 261 unique students enrolled in the e-learning course over four academic years. The course used as the primary source of data contained 13 sections with more than 30 interactive activities, which required continual students'activity. These activities could be divided into assignments, tests, project, and exam.

Interpretation of some features:

Access: represent the total number of course views by a student in the observed period. Assignments: represent a total score from different types of evaluated activities within the observed period. Tests: represent a total score from the midterm and final tests during the semester. Project: a total score from the final project. result_points: represents the total sum of partial points, which the student could get during the course.

[4]:
```
acad_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 261 entries, 0 to 260
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   access         261 non-null    object
 1   tests          261 non-null    object
 2   tests_grade    261 non-null    object
 3   exam           261 non-null    object
 4   project        261 non-null    object
 5   project_grade  261 non-null    object
 6   assignments    261 non-null    object
 7   result_points  261 non-null    object
 8   result_grade   261 non-null    object
 9   graduate       261 non-null    object
 10  year           261 non-null    object
 11  acad_year      261 non-null    object
dtypes: object(12)
```

```
memory usage: 24.6+ KB
```

[5]:
```
acad_data.describe()
```

[5]:

	access	tests	tests_grade	exam	project	project_grade	assignments
count	261	261	261	261	261	261	261
unique	227	162	7	17	202	7	116
top	323	0	E	0	0	A	40
freq	2	17	58	45	42	106	22

	result_points	result_grade	graduate	year	acad_year
count	261	261	261	261	261
unique	241	6	2	4	4
top	190	C	1	2018	2018/2019
freq	8	70	210	74	74

[6]:
```
for col in acad_data.columns:
    print('\nUnique values of',col, 'are : ',acad_data[col].unique(), "\nand␣
 ↪Number of Unique values:",
              len(acad_data[col].unique()))
```

```
Unique values of access are :  ['1256' '985' '1455' '998' '1347' '1000' '1216'
'737' '782' '799' '1506'
 '776' '699' '615' '1065' '482' '459' '592' '779' '619' '738' '2392' '595'
 '424' '1047' '930' '1275' '349' '697' '359' '1085' '457' '620' '701'
 '816' '969' '841' '945' '2089' '912' '680' '1537' '1757' '616' '630'
 '1081' '1054' '674' '1354' '978' '849' '712' '1056' '2135' '936' '967'
 '426' '442' '696' '485' '973' '513' '561' '1084' '1053' '925' '515' '466'
 '498' '275' '397' '172' '1470' '768' '649' '625' '598' '623' '843' '433'
 '723' '583' '577' '411' '504' '614' '383' '469' '1197' '1153' '708'
 '1664' '521' '682' '736' '472' '1162' '893' '1118' '777' '1021' '810'
 '464' '556' '815' '520' '1039' '1111' '496' '941' '352' '710' '659'
 '1265' '676' '772' '379' '570' '475' '487' '558' '501' '534' '306' '863'
 '439' '582' '319' '305' '303' '384' '666' '375' '668' '905' '405' '500'
 '1941' '429' '518' '997' '1583' '1046' '1311' '837' '403' '512' '1292'
 '575' '792' '519' '1043' '1167' '942' '957' '528' '898' '910' '605' '508'
 '574' '727' '790' '1158' '456' '378' '828' '419' '331' '281' '323' '356'
 '208' '440' '267' '505' '569' '286' '686' '1259' '1026' '259' '314' '373'
 '524' '540' '289' '618' '360' '645' '298' '548' '608' '369' '434' '545'
 '313' '334' '691' '250' '204' '624' '177' '1198' '470' '463' '13' '448'
 '522' '479' '79' '179' '537' '1079' '173' '1007' '918' '635' '151' '18'
 '527' '78' '603' '340' '26' '126' '28']
and Number of Unique values: 227

Unique values of tests are :  ['57' '42.87' '54.5' '55' '51' '56' '52' '49'
'55.5' '52.5' '53.5' '56.5'
 '54' '51.5' '38.84' '43.33' '38.44' '49.07' '37.4' '48.84' '47.33'
```

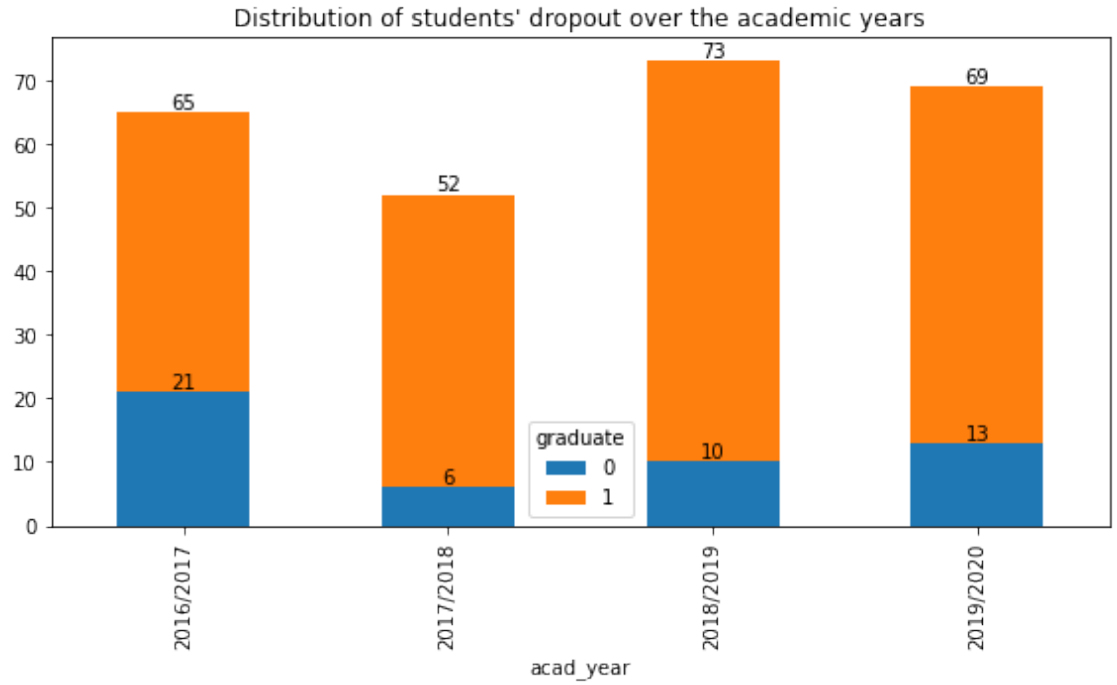
```
 '45.33' '0' '41.07' '40.7' '42.53' '36.48' '37' '40.43' '34.57' '49.5'
 '48' '46.5' '40.23' '50' '45' '46' '45.5' '42.5' '53' '53.82' '46.56'
 '47.5' '50.5' '34.82' '32.96' '33.5' '31.8' '39.9' '34.4' '39.63' '35.78'
 '36.3' '28.67' '35.8' '30.53' '37.13' '35.07' '43' '38.4' '31.75' '35.7'
 '41.5' '31.02' '32.26' '43.67' '34.65' '31.48' '31.69' '36.87' '39'
 '43.64' '40' '42' '40.82' '40.13' '41' '44' '50.91' '42.76' '44.87'
 '44.53' '47.07' '47' '43.2' '43.5' '39.16' '36.22' '43.82' '44.5' '45.64'
 '41.56' '38.5' '43.42' '29.73' '29.62' '31.47' '27.5' '32.93' '32.52'
 '29.58' '30.3' '39.5' '36.5' '36.91' '38.31' '38.53' '36' '40.51' '40.6'
 '39.07' '35.31' '29' '35' '42.69' '37.84' '37.81' '37.87' '38.79' '40.5'
 '43.78' '21.5' '34.09' '34.69' '44.67' '31.93' '30.37' '31.5' '31' '34'
 '33.7' '42.4' '26' '34.5' '37.5' '32.5' '28' '24' '31.55' '16' '25.5'
 '48.33' '50.93' '47.2' '39.24' '41.87' '42.58' '39.6' '18.5' '38.49'
 '36.93' '34.73' '41.22' '31.52' '32.36' '33' '20.5' '7' '11' '42.27' '27'
 '32']
and Number of Unique values: 162

Unique values of tests_grade are :  ['A' 'B' 'C' 'E' 'D' 'FX' '0']
and Number of Unique values: 7

Unique values of exam are :  ['19' '16' '13' '15' '20' '14' '17' '18' '0' '11'
'12' '10' '9' '5' '6'
 '7' '8']
and Number of Unique values: 17

Unique values of project are :  ['91.54' '75.96' '96.79' '93.36' '92.86' '94.55'
'98.77' '96.7' '93.17'
 '90.93' '97.54' '89.18' '87.82' '86.64' '90' '85.23' '84.41' '87.34'
 '88.41' '70.15' '81.09' '77.7' '82.55' '66.54' '77.44' '60.4' '0' '75.84'
 '76.2' '73.94' '79.52' '77.13' '75.88' '79.19' '88.59' '95.74' '93.94'
 '94.56' '96.59' '93.44' '97.91' '77.9' '88.06' '99.48' '96.08' '85.89'
 '90.79' '92.98' '90.03' '98.13' '79.79' '88.02' '90.99' '89.1' '90.85'
 '93.85' '87' '84.14' '75.13' '86.8' '88.98' '80.66' '85.77' '85.16'
 '71.8' '66.27' '85.98' '75.14' '84.36' '73.3' '84.27' '69.34' '73.78'
 '80' '76.32' '76.53' '76.65' '77.01' '77.98' '73.58' '67.88' '67.43'
 '68.85' '69.7' '71.84' '70.12' '94.18' '68.48' '73.33' '71.27' '88.4'
 '70.87' '69.49' '89.88' '68' '80.63' '70.75' '73.62' '72.2' '70.59'
 '89.41' '78.47' '89.75' '82.54' '82.17' '81.48' '85.36' '78.95' '84.61'
 '91.21' '79.29' '84.34' '80.32' '85.45' '73.37' '77.66' '70.06' '81.68'
 '76.76' '75.89' '74.35' '76.59' '80.25' '70.32' '87.55' '66.41' '75.38'
 '84.07' '70.23' '79.57' '69.46' '72.04' '68.81' '68.04' '74.63' '74.06'
 '74.98' '61.49' '68.44' '95.6' '88.55' '78.24' '80.1' '91.14' '74.22'
 '72.89' '67.4' '86.72' '71.93' '76.89' '70.62' '87.44' '81.85' '69.37'
 '68.86' '63.62' '71.39' '72.55' '62.86' '64.91' '59.8' '77.39' '78.89'
 '57.68' '52.5' '83.59' '60.68' '63.07' '73.77' '73.11' '66.88' '64.09'
 '73.45' '65.25' '75.47' '84.5' '89.21' '82.06' '73.88' '65.85' '66.83'
 '92.24' '74.68' '79.23' '73.09' '75.95' '66.61' '74.09' '83.06' '59.52'
 '82.33' '43.88' '71.15' '58.32' '22.5' '24.48' '52.29' '56.25' '83.16'
```

```
 '63.19' '65.97' '92.33']
and Number of Unique values: 202

Unique values of project_grade are :  ['A' 'C' 'D' 'FX' 'B' 'E' '0']
and Number of Unique values: 7

Unique values of assignments are :  ['40' '13.7' '39' '37' '32' '24' '23' '19'
'21' '14.86' '12.36' '14.59'
 '15.31' '11.8' '14.4' '14.5' '11.48' '14' '11.5' '12.7' '11.6' '13.4'
 '11.7' '31' '35' '33' '36' '20' '15.42' '14.13' '16' '15' '2' '9.7' '12'
 '8.7' '7.7' '9.4' '11.9' '11.4' '13.5' '9.5' '13.2' '9.1' '10.27' '8.6'
 '7.9' '10.4' '9.6' '5.9' '6.1' '7.4' '13.17' '34' '11.74' '15.41' '12.3'
 '18' '12.81' '11.77' '17' '13.23' '9.63' '9.94' '13.31' '13.45' '15.27'
 '12.5' '8.5' '10.75' '13.9' '8.4' '6.3' '25' '30' '11.79' '12.14' '14.68'
 '12.58' '13' '11.99' '38' '9.37' '9.3' '12.17' '11.72' '9' '11' '4'
 '9.41' '9.92' '6.5' '10' '28' '27' '7.63' '8' '26' '7' '0' '18.33' '16.6'
 '15.78' '9.23' '16.67' '15.56' '9.2' '11.67' '12.4' '9.83' '10.28'
 '11.22' '4.19' '8.1' '8.58' '5']
and Number of Unique values: 116

Unique values of result_points are :  ['189.92' '189.43' '188.91' '186.85'
'186.38' '184.56' '184.26' '183.02'
 '181.23' '180.22' '180.19' '175.16' '173.57' '171.96' '170.2' '169.55'
 '169.01' '168.29' '167.27' '190' '22' '186.83' '186.35' '184.35' '182.57'
 '180.89' '180.65' '179.99' '179.82' '179.11' '178.86' '178.4' '177.96'
 '177.73' '176.25' '174.62' '174.5' '174.02' '171.65' '171.53' '171.14'
 '170.79' '170.35' '169.88' '169.54' '169.48' '168.93' '167.46' '167.34'
 '167.14' '165.14' '161.98' '161.4' '160.45' '160.3' '159.71' '158.5'
 '158.44' '158.15' '154.69' '154.55' '154.16' '153.81' '153.32' '150.35'
 '144.66' '52.05' '172.3' '168.5' '168.1' '168.07' '167.01' '166.81'
 '165.86' '165.2' '165.08' '164.11' '164.04' '164.02' '163.99' '163.17'
 '162.88' '162.43' '162.23' '162.21' '162.03' '161.93' '161.3' '160.77'
 '160.05' '159.61' '159.04' '158.22' '158.08' '155.65' '154.65' '154.52'
 '154.13' '154.02' '153.55' '152.95' '152.54' '152.06' '151.48' '150.77'
 '150.36' '150.12' '150.03' '149.48' '148.95' '148.76' '148.4' '148.36'
 '148.2' '147.72' '147.41' '147.15' '146.87' '146.72' '145.94' '145.86'
 '145.15' '144.97' '143.41' '143.32' '142.74' '141.96' '141.53' '141.35'
 '140.63' '140.56' '139.87' '201.92' '157.31' '154.98' '154.54' '152.45'
 '150.98' '150.05' '149.99' '149.97' '148.13' '147.11' '144.9' '144.68'
 '142.42' '142.37' '142.11' '142.03' '141.1' '140.73' '139.88' '138.86'
 '138.8' '138.45' '138.29' '137.79' '137.36' '137.14' '136.92' '136.33'
 '136.31' '134.45' '134.02' '133.68' '131.87' '130.36' '129.57' '128.91'
 '128.36' '127.69' '126.91' '126.11' '120.92' '54.04' '27.72' '186.69'
 '181.44' '180.71' '142.75' '137.82' '135.6' '135.2' '130.86' '128.9'
 '127.99' '127.03' '126.12' '124.68' '121.98' '120.62' '118.72' '116.59'
 '115.75' '110.67' '134' '97.3' '96.4' '95.38' '89.35' '88.7' '88.33'
 '76.32' '76.19' '72.33' '71.7' '70.17' '68.95' '68.68' '64.85' '61.99'
 '60.14' '59.36' '58.02' '56.96' '55.8' '54.6' '54.21' '52.9' '52.07'
```

```
 '47.18' '46.11' '36.59' '36.5' '33.09' '22.44' '21.67' '20.52' '20.48'
 '166.9' '119.58' '118.23' '116.92' '115.92' '114.06' '17.1' '8.55' '26'
 '21' '16' '0']
and Number of Unique values: 241

Unique values of result_grade are :  ['A' 'B' 'C' 'D' 'E' 'FX']
and Number of Unique values: 6

Unique values of graduate are :  ['1' '0']
and Number of Unique values: 2

Unique values of year are :  ['2019' '2017' '2018' '2016']
and Number of Unique values: 4

Unique values of acad_year are :  ['2019/2020' '2017/2018' '2018/2019'
'2016/2017']
and Number of Unique values: 4

Number of duplicated rows:  2
```

1 EDA

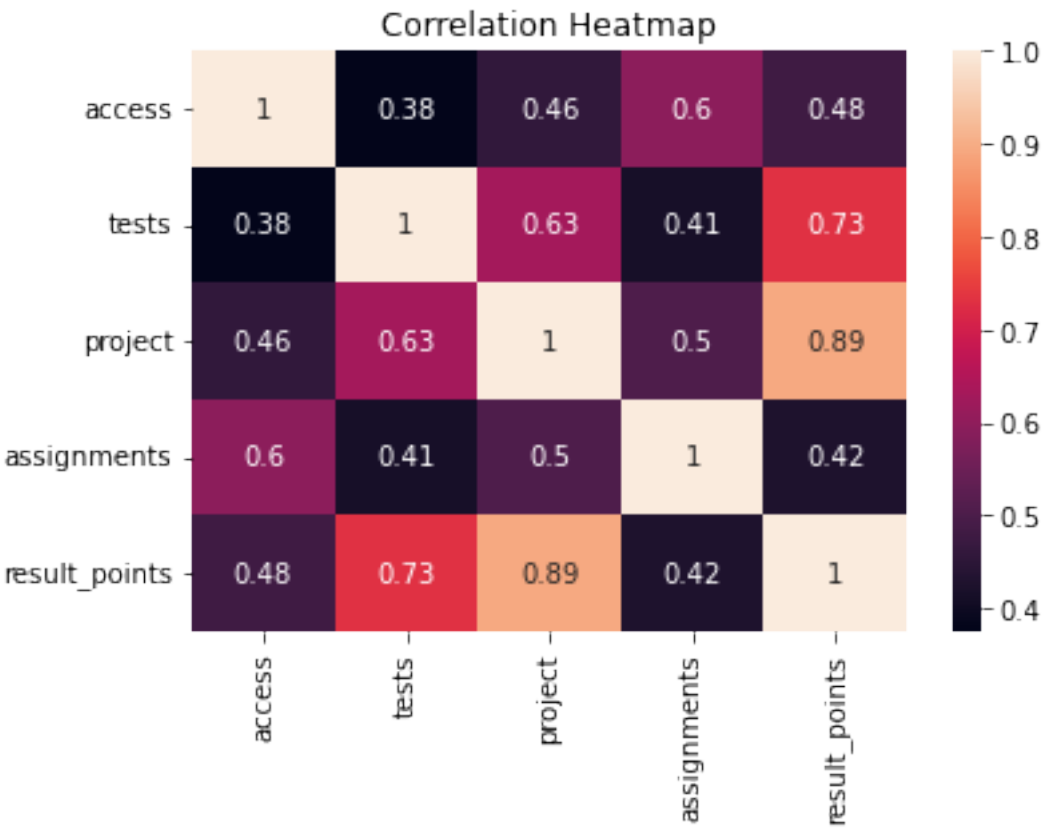
[38]:
```
import matplotlib.pyplot as plt

ax = acad_data.groupby(['acad_year','graduate'])['access'].count().unstack().
 ↪plot(kind='bar',stacked=True,

 ↪  figsize=(8, 5))
ax.bar_label(ax.containers[0], label_type='edge')
ax.bar_label(ax.containers[1], label_type='edge')
#plt.subplots()
plt.title("Distribution of students' dropout over the academic years")
plt.tight_layout()
```



[46]:
```
# change the dtype of numerical columns

acad_data[['access','tests','project','assignments','result_points']]=acad_data[['access',
 ↪'tests','project','assignments','result_points']].astype(float)
```

```
C:\Users\B590\AppData\Local\Temp\ipykernel_10124\2715135247.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  acad_data[['access','tests','project','assignments','result_points']]=acad_dat
a[['access','tests','project','assignments','result_points']].astype(float)
```

[47]:
```
import seaborn as sns

sns.
 ↪heatmap(acad_data[['access','tests','project','assignments','result_points']].
 ↪corr(), annot=True)
plt.title('Correlation Heatmap')
plt.show()
```



The variables project and tests have a slightly stronger correlation with result_points than the other variables.

[ ]: