EDA_CollegeInsight

September 28, 2021

[3]:

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

# Load CSV
df = pd.read_csv('collegeinsight_data_nolabel_ICs_by_year.csv',␣
 ↪encoding="ISO-8859-1") # default encoding couldn't ready all characters
```

```
/Users/jhuang/opt/anaconda3/envs/MLBDenv/lib/python3.8/site-
packages/IPython/core/interactiveshell.py:3146: DtypeWarning: Columns (7,16)
have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

[4]:

```
## Correlation matrix
corr = df.corr()
```

[6]:

```
corr[['fa_loans_debt_avg_d']].sort_values(by=['fa_loans_debt_avg_d'])
```

[6]:

```
                                  fa_loans_debt_avg_d
fa_loans_fed_vol_p                          -0.352467
fa_sfa_living_offnotfam_p                   -0.205308
ef_undg_students_distancesome_p             -0.181024
ef_all_students_distancesome_p              -0.171980
fa_sfa_living_offfam_p                      -0.152660
…                                                 …
fa_loans_fed_pc_d                            0.685298
fa_loans_fed_avg_d                           0.740785
fa_loans_debt_avg_d                          1.000000
deggrant                                          NaN
fsa_cba_pl_fedaw_amt                              NaN

[399 rows x 1 columns]
```

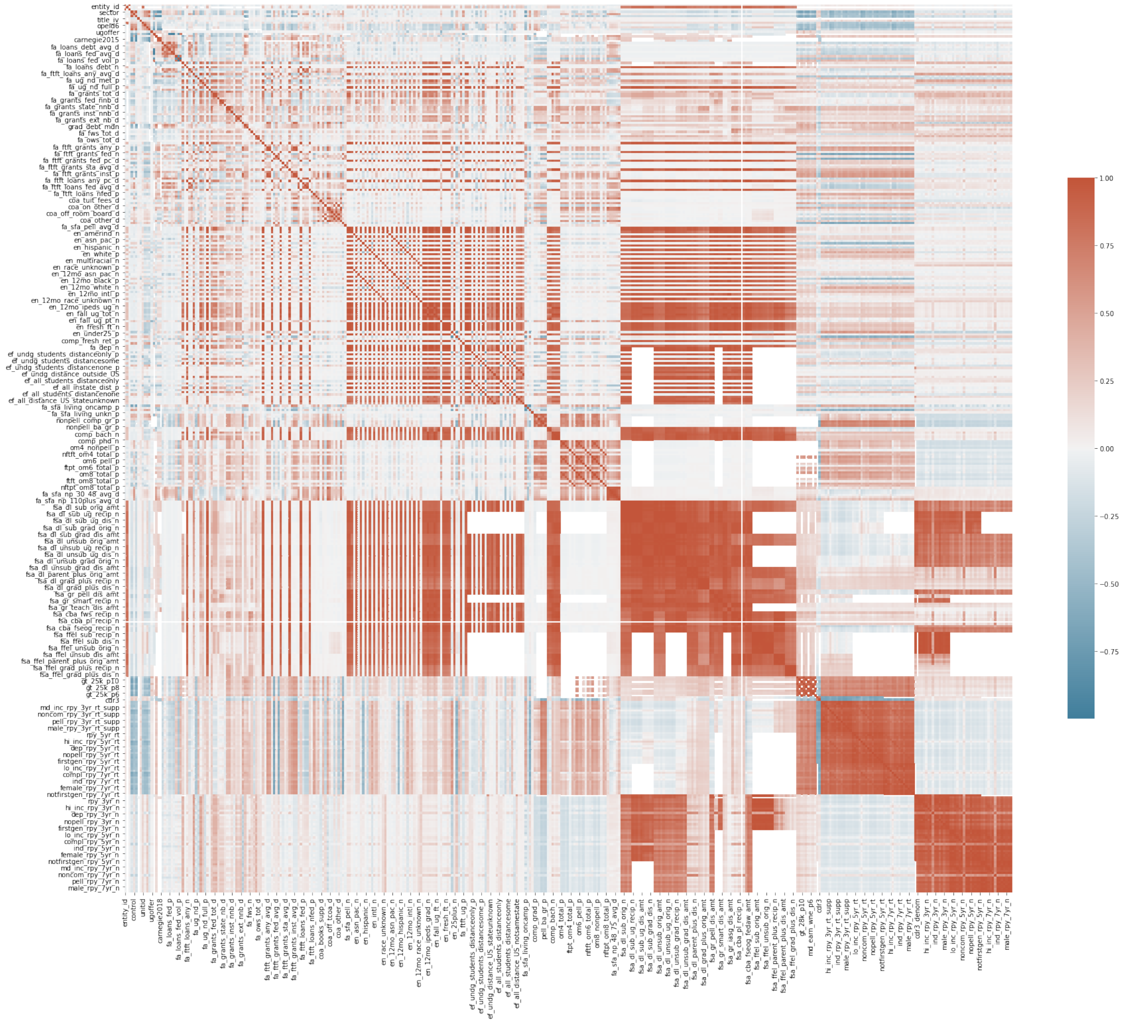
[3]:

```
## Plot correlation matrix

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(30, 30))
```

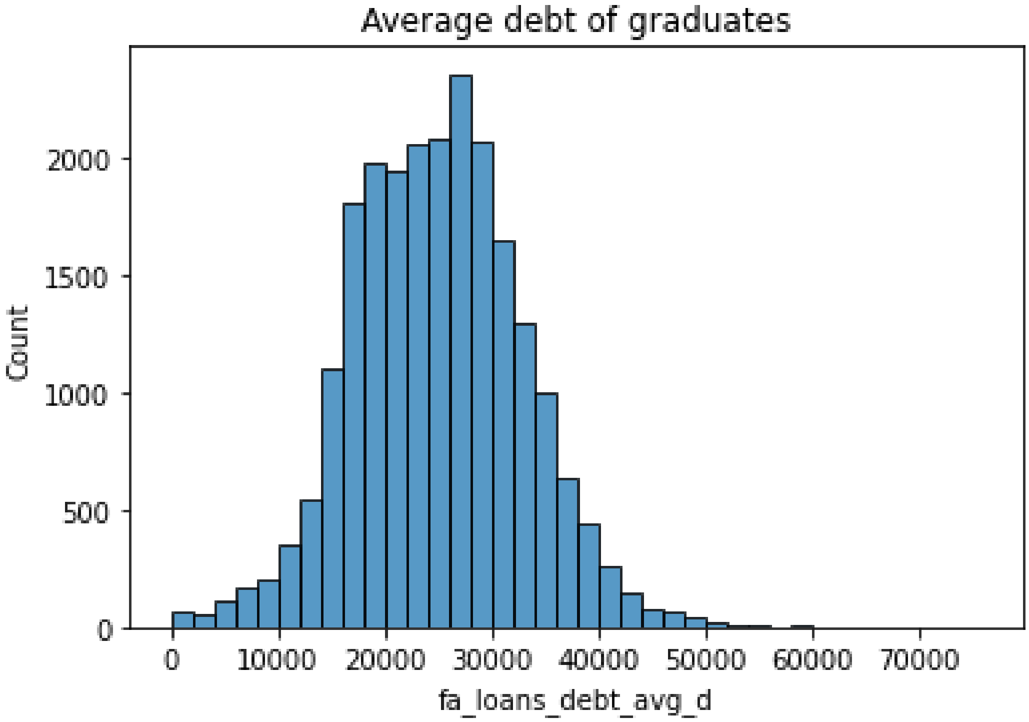
```python
# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)
sns.heatmap(corr, cmap=cmap, square=True, cbar_kws={"shrink": .5})
```

[3]:
```
<AxesSubplot:>
```



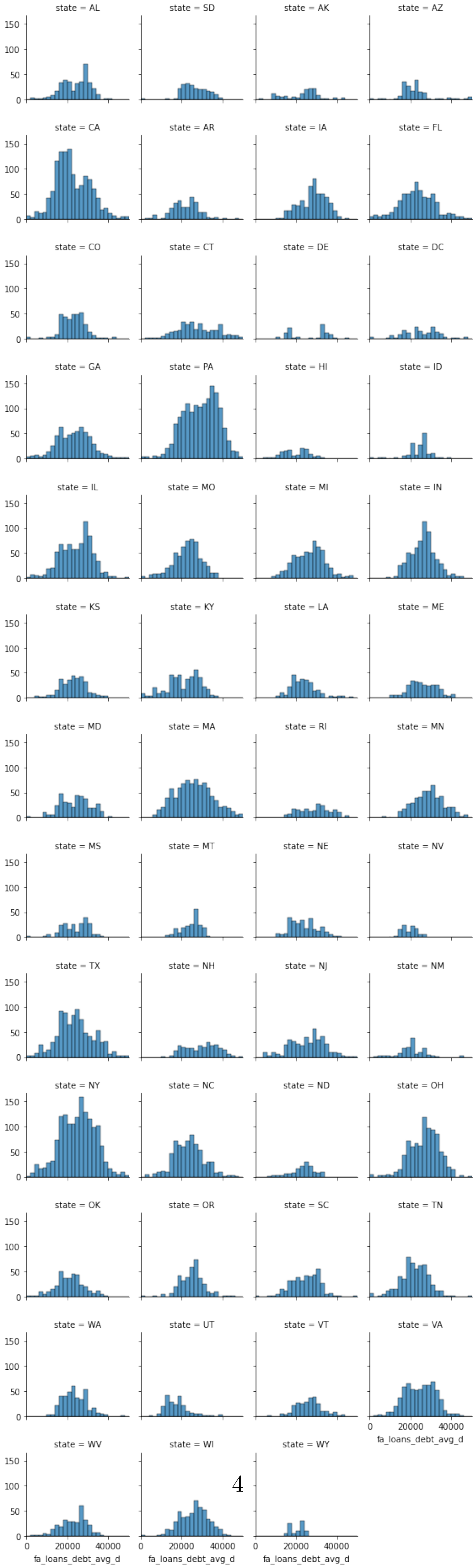
[4]:
```python
sns.histplot(data=df, x='fa_loans_debt_avg_d', binrange=(0,75000),␣
↪binwidth=2000).set_title('Average debt of graduates')
```

[4]:
```
Text(0.5, 1.0, 'Average debt of graduates')
```



```
[5]: ## Average debt of graduates by state
     g = sns.FacetGrid(df, col="state", col_wrap=4, height=2, xlim=(0,50000))
     g.map(sns.histplot, "fa_loans_debt_avg_d",binrange=(0,50000), binwidth=2000)
```

```
[5]: <seaborn.axisgrid.FacetGrid at 0x7fa90e433640>
```

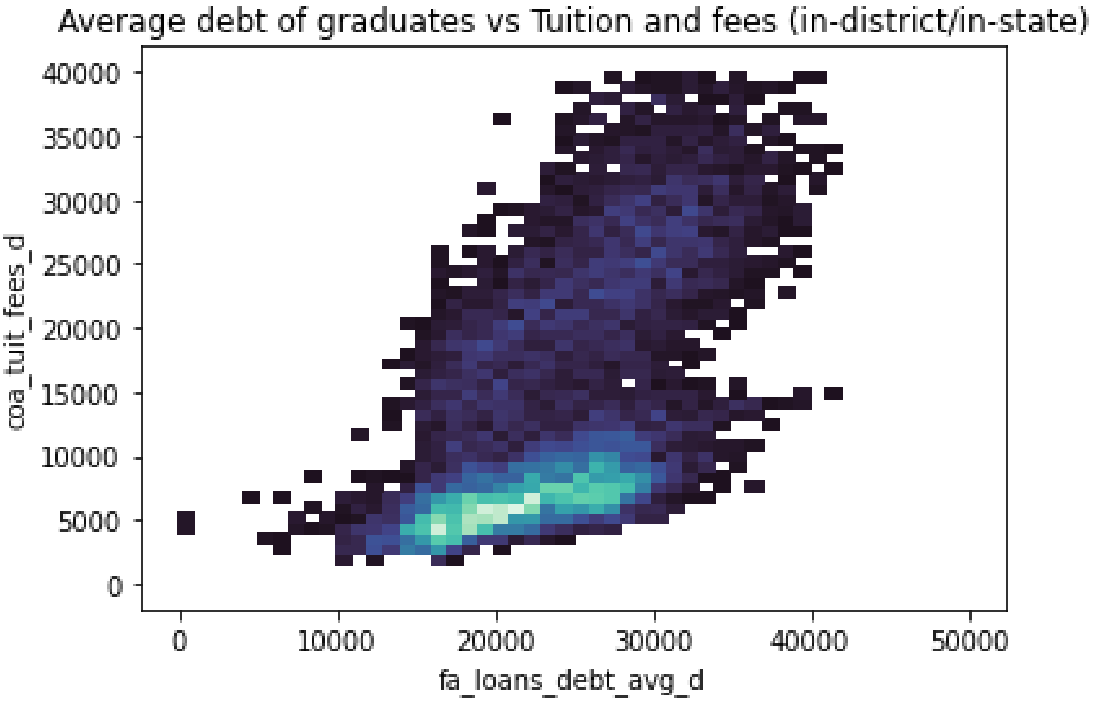

state = AL
state = SD
state = AK
state = AZ
state = CA
state = AR
state = IA
state = FL
state = CO
state = CT
state = DE
state = DC
state = GA
state = PA
state = HI
state = ID
state = IL
state = MO
state = MI
state = IN
state = KS
state = KY
state = LA
state = ME
state = MD
state = MA
state = RI
state = MN
state = MS
state = MT
state = NE
state = NV
state = TX
state = NH
state = NJ
state = NM
state = NY
state = NC
state = ND
state = OH
state = OK
state = OR
state = SC
state = TN
state = WA
state = UT
state = VT
state = VA
state = WV
state = WI
state = WY
150
100
50
0
0
20000
40000
fa_loans_debt_avg_d

[7]:

```
## Average debt of graduates vs Tuition and fees
## Bi-modal structure. Avg debt rises very quickly as tuition fees rise
df_cut = df[(df['fa_loans_debt_avg_d'].between(0, 50000)) &␣
 ↪(df["COA_TUIT_FEES_D".lower()].between(0, 40000))]
sns.histplot(df_cut, y="COA_TUIT_FEES_D".lower(), x="fa_loans_debt_avg_d",␣
 ↪bins=50, pthresh=.1, cmap="mako")
plt.title("Average debt of graduates vs Tuition and fees (in-district/
 ↪in-state)")
```

[7]:

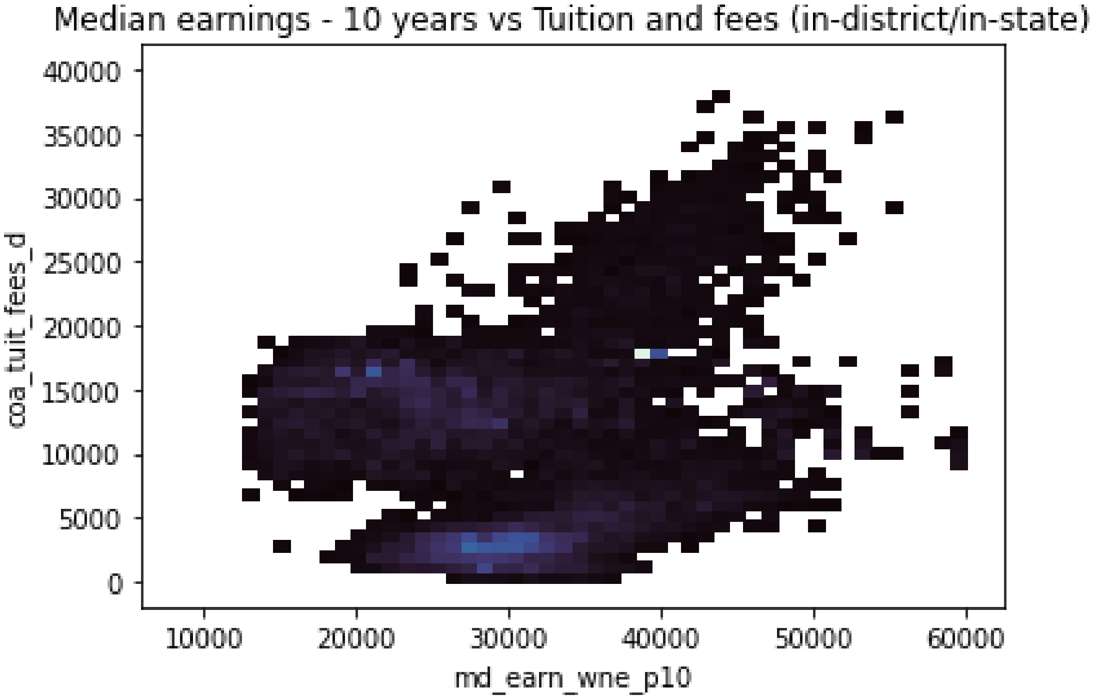
```
Text(0.5, 1.0, 'Average debt of graduates vs Tuition and fees (in-district/in-
state)')
```



[10]:

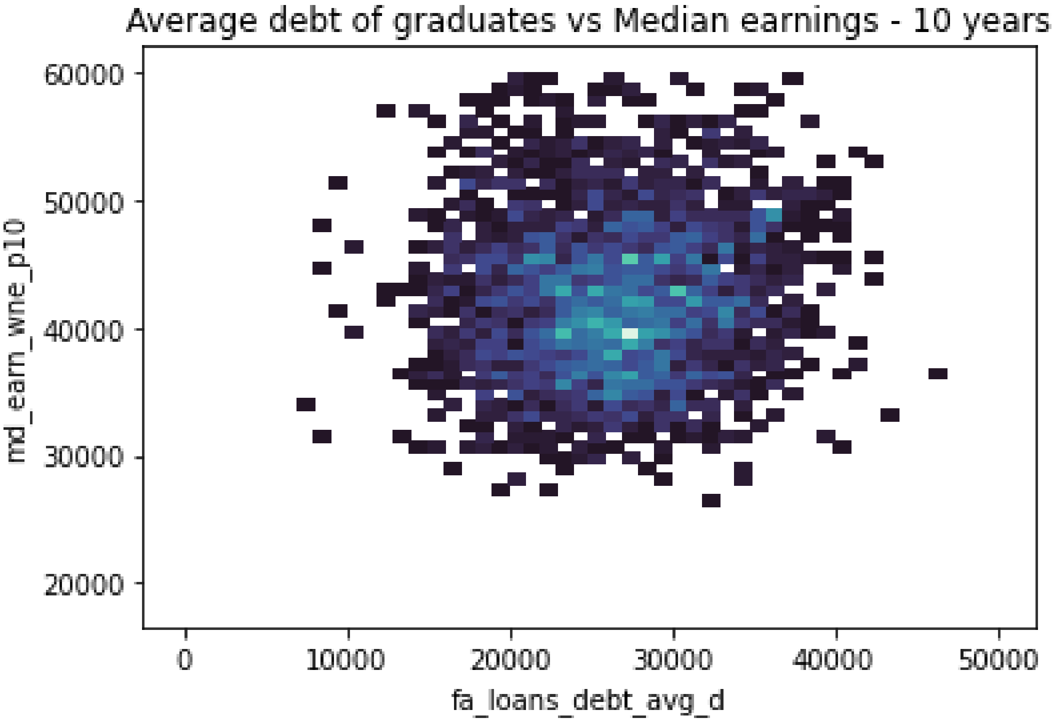
```
## Median earnings - 10 years vs Tuition and fees
df_cut2 = df[(df["MD_EARN_WNE_P10".lower()].between(0, 60000)) &␣
 ↪(df["COA_TUIT_FEES_D".lower()].between(0, 40000))]
sns.histplot(df_cut2, y="COA_TUIT_FEES_D".lower(), x="MD_EARN_WNE_P10".lower(),␣
 ↪bins=50, pthresh=.1, cmap="mako")
plt.title("Median earnings - 10 years vs Tuition and fees (in-district/
 ↪in-state)")
```

```
[10]: Text(0.5, 1.0, 'Median earnings - 10 years vs Tuition and fees (in-district/in-
      state)')
```



```
[14]: ## Average debt of graduates vs Tuition and fees
      ## Bi-modal structure. Avg debt rises very quickly as tuition fees rise
      df_cut3 = df[(df['fa_loans_debt_avg_d'].between(0, 50000)) &␣
       ↪(df["MD_EARN_WNE_P10".lower()].between(0, 60000))]
      sns.histplot(df_cut3, y="MD_EARN_WNE_P10".lower(), x="fa_loans_debt_avg_d",␣
       ↪bins=50, pthresh=.1, cmap="mako")
      plt.title("Average debt of graduates vs Median earnings - 10 years")
```

```
[14]: Text(0.5, 1.0, 'Average debt of graduates vs Median earnings - 10 years')
```


Average debt of graduates vs Median earnings - 10 years
md_earn_wne_p10
60000
50000
40000
30000
20000
0
10000
20000
30000
40000
50000
fa_loans_debt_avg_d