EDA_CollegeInsight_Polina

September 28, 2021

LOAD THE DATA AND PACKAGES

[1]:
```
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
```

```
/opt/conda/lib/python3.8/site-packages/IPython/core/interactiveshell.py:3165:
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,
```

EDA AND DATA CLEANING

[2]:
```
df.head()
```

[2]:

	data_yr_string	entity_id	entity_type	name	\
0	2000-01	111100654	1	ALABAMA A & M UNIVERSITY	
1	2000-01	111100663	1	UNIVERSITY OF ALABAMA AT BIRMINGHAM	
2	2000-01	111100706	1	UNIVERSITY OF ALABAMA IN HUNTSVILLE	
3	2000-01	111100724	1	ALABAMA STATE UNIVERSITY	
4	2000-01	111100751	1	UNIVERSITY OF ALABAMA	

	city	state	state_fips	cong_dist	webaddr	sector	…	\
0	NORMAL	AL	1.0	NaN	WWW.AAMU.EDU	1.0	…	
1	BIRMINGHAM	AL	1.0	NaN	www.uab.edu	1.0	…	
2	HUNTSVILLE	AL	1.0	NaN	www.uah.edu	1.0	…	
3	MONTGOMERY	AL	1.0	NaN	www.alasu.edu	1.0	…	
4	TUSCALOOSA	AL	1.0	NaN	www.ua.edu	1.0	…	

	compl_rpy_7yr_n	noncom_rpy_7yr_n	dep_rpy_7yr_n	ind_rpy_7yr_n	\
0	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	NaN	

```
4              NaN               NaN            NaN            NaN

   pell_rpy_7yr_n  nopell_rpy_7yr_n female_rpy_7yr_n  male_rpy_7yr_n  \
0             NaN               NaN              NaN             NaN
1             NaN               NaN              NaN             NaN
2             NaN               NaN              NaN             NaN
3             NaN               NaN              NaN             NaN
4             NaN               NaN              NaN             NaN

   firstgen_rpy_7yr_n  notfirstgen_rpy_7yr_n
0                 NaN                    NaN
1                 NaN                    NaN
2                 NaN                    NaN
3                 NaN                    NaN
4                 NaN                    NaN

[5 rows x 408 columns]
```

[3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 139149 entries, 0 to 139148
Columns: 408 entries, data_yr_string to notfirstgen_rpy_7yr_n
dtypes: float64(396), int64(3), object(9)
memory usage: 433.1+ MB
```

[4]:
```
# all columns
my_list = list(df)

print (my_list)
```

```
['data_yr_string', 'entity_id', 'entity_type', 'name', 'city', 'state',
'state_fips', 'cong_dist', 'webaddr', 'sector', 'control', 'level', 'title_iv',
'hbcu', 'unitid', 'opeid6', 'opeid8', 'reporting_type', 'deggrant', 'ugoffer',
'carnegie2000', 'carnegie2010', 'carnegie2015', 'carnegie2018',
'fa_loans_debt_p', 'fa_loans_debt_avg_d', 'fa_loans_debt_pc_d',
'fa_loans_fed_p', 'fa_loans_fed_avg_d', 'fa_loans_fed_pc_d',
'fa_loans_nfed_pc_d', 'fa_loans_fed_vol_p', 'fa_loans_nfed_vol_p',
'fa_loans_debt_cohort_n', 'fa_loans_debt_n', 'fa_ftft_loans_any_n',
'fa_ftft_loans_any_p', 'fa_ftft_loans_any_avg_d', 'fa_ug_nd_n', 'fa_ug_nd_p',
'fa_ug_nd_met_p', 'fa_ug_award_n', 'fa_ug_nd_full_n', 'fa_ug_nd_full_p',
'fa_ftft_aid_n', 'fa_ftft_aid_p', 'fa_grants_tot_d', 'fa_grants_fed_tot_d',
'fa_grants_fed_nb_d', 'fa_grants_fed_nnb_d', 'fa_grants_state_tot_d',
'fa_grants_state_nb_d', 'fa_grants_state_nnb_d', 'fa_grants_inst_tot_d',
'fa_grants_inst_nb_d', 'fa_grants_inst_nnb_d', 'fa_grants_inst_nb_p',
'fa_grants_ext_tot_d', 'fa_grants_ext_nb_d', 'fa_grants_ext_nnb_d',
'fa_ug_avg_grant_d', 'grad_debt_mdn', 'wdraw_debt_mdn', 'fa_fws_n',
'fa_fws_tot_d', 'fa_fws_avg_d', 'fa_ows_n', 'fa_ows_tot_d', 'fa_ows_avg_d',
```

'fa_ftft_grants_any_n', 'fa_ftft_grants_any_p', 'fa_ftft_grants_any_avg_d',
'fa_ftft_grants_any_pc_d', 'fa_ftft_grants_fed_n', 'fa_ftft_grants_fed_p',
'fa_ftft_grants_fed_avg_d', 'fa_ftft_grants_fed_pc_d', 'fa_ftft_grants_sta_n',
'fa_ftft_grants_sta_p', 'fa_ftft_grants_sta_avg_d', 'fa_ftft_grants_sta_pc_d',
'fa_ftft_grants_inst_n', 'fa_ftft_grants_inst_p', 'fa_ftft_grants_inst_avg_d',
'fa_ftft_grants_inst_pc_d', 'fa_ftft_loans_any_pc_d', 'fa_ftft_loans_fed_n',
'fa_ftft_loans_fed_p', 'fa_ftft_loans_fed_avg_d', 'fa_ftft_loans_fed_pc_d',
'fa_ftft_loans_nfed_n', 'fa_ftft_loans_nfed_p', 'fa_ftft_loans_nfed_avg_d',
'fa_ftft_loans_nfed_pc_d', 'coa_tuit_fees_d', 'coa_books_supp_d',
'coa_on_room_board_d', 'coa_on_other_d', 'coa_on_tcoa_d', 'coa_off_tcoa_d',
'coa_cipcode_prog1', 'coa_length_prog1', 'coa_off_room_board_d',
'coa_off_other_d', 'coa_room_board_d', 'coa_other_d', 'coa_tcoa_d',
'fa_sfa_pell_p', 'fa_sfa_pell_avg_d', 'fa_sfa_pell_n', 'fa_sfa_pell_tot_d',
'en_amerind_n', 'en_amerind_p', 'en_asn_pac_n', 'en_asn_pac_p', 'en_black_n',
'en_black_p', 'en_hispanic_n', 'en_hispanic_p', 'en_white_n', 'en_white_p',
'en_intl_n', 'en_intl_p', 'en_multiracial_n', 'en_multiracial_p',
'en_race_unknown_n', 'en_race_unknown_p', 'en_12mo_amerind_n',
'en_12mo_amerind_p', 'en_12mo_asn_pac_n', 'en_12mo_asn_pac_p',
'en_12mo_black_n', 'en_12mo_black_p', 'en_12mo_hispanic_n',
'en_12mo_hispanic_p', 'en_12mo_white_n', 'en_12mo_white_p', 'en_12mo_intl_n',
'en_12mo_intl_p', 'en_12mo_multiracial_n', 'en_12mo_multiracial_p',
'en_12mo_race_unknown_n', 'en_12mo_race_unknown_p', 'en_12mo_ipeds_tot_n',
'en_12mo_ipeds_ug_n', 'en_12mo_ipeds_grad_n', 'en_fall_tot_n',
'en_fall_ug_tot_n', 'en_fall_grad_n', 'en_fall_ug_ft_n', 'en_fall_ug_pt_n',
'en_fall_ug_ft_p', 'en_fresh_n', 'en_fresh_ft_n', 'en_fall_fte_n',
'en_under25_n', 'en_under25_p', 'en_25plus_n', 'en_25plus_p',
'comp_fresh_ret_p', 'fa_ftft_ug_n', 'fa_ftft_ug_p', 'fa_dep_n',
'ef_undg_allstudentsenrolled', 'ef_undg_students_distanceonly',
'ef_undg_students_distanceonly_p', 'ef_undg_distance_samestate',
'ef_undg_instate_dist_p', 'ef_undg_students_distancesome',
'ef_undg_students_distancesome_p', 'ef_undg_students_distancenone',
'ef_undg_students_distancenone_p', 'ef_undg_distance_US_notsamestate',
'ef_undg_distance_US_stateunknown', 'ef_undg_distance_outside_US',
'ef_undg_distance_unknown', 'ef_all_allstudentsenrolled',
'ef_all_students_distanceonly', 'ef_all_students_distanceonly_p',
'ef_all_distance_samestate', 'ef_all_instate_dist_p',
'ef_all_students_distancesome', 'ef_all_students_distancesome_p',
'ef_all_students_distancenone', 'ef_all_students_distancenone_p',
'ef_all_distance_US_notsamestate', 'ef_all_distance_US_stateunknown',
'ef_all_distance_outside_US', 'ef_all_distance_unknown',
'fa_sfa_living_oncamp_p', 'fa_sfa_living_offfam_p', 'fa_sfa_living_offnotfam_p',
'fa_sfa_living_unkn_p', 'comp_grad_p', 'pell_comp_gr_p', 'nonpell_comp_gr_p',
'ba_gr_p', 'pell_ba_gr_p', 'nonpell_ba_gr_p', 'comp_certif_n', 'comp_assoc_n',
'comp_bach_n', 'comp_grad_certif_n', 'comp_ma_n', 'comp_phd_n', 'om4_total_p',
'om4_pell_p', 'om4_nonpell_p', 'ftft_om4_total_p', 'ftpt_om4_total_p',
'nftft_om4_total_p', 'nftpt_om4_total_p', 'om6_total_p', 'om6_pell_p',
'om6_nonpell_p', 'ftft_om6_total_p', 'ftpt_om6_total_p', 'nftft_om6_total_p',
'nftpt_om6_total_p', 'om8_total_p', 'om8_pell_p', 'om8_nonpell_p',

'ftft_om8_total_p', 'ftpt_om8_total_p', 'nftft_om8_total_p',
'nftpt_om8_total_p', 'fa_sfa_np_avg_d', 'fa_sfa_np_0_30_avg_d',
'fa_sfa_np_30_48_avg_d', 'fa_sfa_np_48_75_avg_d', 'fa_sfa_np_75_110_avg_d',
'fa_sfa_np_110plus_avg_d', 'fsa_dl_sub_recip_n', 'fsa_dl_sub_orig_n',
'fsa_dl_sub_orig_amt', 'fsa_dl_sub_dis_n', 'fsa_dl_sub_dis_amt',
'fsa_dl_sub_ug_recip_n', 'fsa_dl_sub_ug_orig_n', 'fsa_dl_sub_ug_orig_amt',
'fsa_dl_sub_ug_dis_n', 'fsa_dl_sub_ug_dis_amt', 'fsa_dl_sub_grad_recip_n',
'fsa_dl_sub_grad_orig_n', 'fsa_dl_sub_grad_orig_amt', 'fsa_dl_sub_grad_dis_n',
'fsa_dl_sub_grad_dis_amt', 'fsa_dl_unsub_recip_n', 'fsa_dl_unsub_orig_n',
'fsa_dl_unsub_orig_amt', 'fsa_dl_unsub_dis_n', 'fsa_dl_unsub_dis_amt',
'fsa_dl_unsub_ug_recip_n', 'fsa_dl_unsub_ug_orig_n', 'fsa_dl_unsub_ug_orig_amt',
'fsa_dl_unsub_ug_dis_n', 'fsa_dl_unsub_ug_dis_amt', 'fsa_dl_unsub_grad_recip_n',
'fsa_dl_unsub_grad_orig_n', 'fsa_dl_unsub_grad_orig_amt',
'fsa_dl_unsub_grad_dis_n', 'fsa_dl_unsub_grad_dis_amt',
'fsa_dl_parent_plus_recip_n', 'fsa_dl_parent_plus_orig_n',
'fsa_dl_parent_plus_orig_amt', 'fsa_dl_parent_plus_dis_n',
'fsa_dl_parent_plus_dis_amt', 'fsa_dl_grad_plus_recip_n',
'fsa_dl_grad_plus_orig_n', 'fsa_dl_grad_plus_orig_amt',
'fsa_dl_grad_plus_dis_n', 'fsa_dl_grad_plus_dis_amt', 'fsa_gr_pell_recip_n',
'fsa_gr_pell_dis_amt', 'fsa_gr_acg_recip_n', 'fsa_gr_acg_dis_amt',
'fsa_gr_smart_recip_n', 'fsa_gr_smart_dis_amt', 'fsa_gr_teach_recip_n',
'fsa_gr_teach_dis_amt', 'fsa_gr_iasg_recip_n', 'fsa_gr_iasg_dis_amt',
'fsa_cba_fws_recip_n', 'fsa_cba_fws_fedaw_amt', 'fsa_cba_fws_dis_amt',
'fsa_cba_pl_recip_n', 'fsa_cba_pl_fedaw_amt', 'fsa_cba_pl_dis_amt',
'fsa_cba_fseog_recip_n', 'fsa_cba_fseog_fedaw_amt', 'fsa_cba_fseog_dis_amt',
'fsa_ffel_sub_recip_n', 'fsa_ffel_sub_orig_n', 'fsa_ffel_sub_orig_amt',
'fsa_ffel_sub_dis_n', 'fsa_ffel_sub_dis_amt', 'fsa_ffel_unsub_recip_n',
'fsa_ffel_unsub_orig_n', 'fsa_ffel_unsub_orig_amt', 'fsa_ffel_unsub_dis_n',
'fsa_ffel_unsub_dis_amt', 'fsa_ffel_parent_plus_recip_n',
'fsa_ffel_parent_plus_orig_n', 'fsa_ffel_parent_plus_orig_amt',
'fsa_ffel_parent_plus_dis_n', 'fsa_ffel_parent_plus_dis_amt',
'fsa_ffel_grad_plus_recip_n', 'fsa_ffel_grad_plus_orig_n',
'fsa_ffel_grad_plus_orig_amt', 'fsa_ffel_grad_plus_dis_n',
'fsa_ffel_grad_plus_dis_amt', 'md_earn_wne_p10', 'gt_25k_p10', 'gt_28k_p10',
'md_earn_wne_p8', 'gt_25k_p8', 'gt_28k_p8', 'md_earn_wne_p6', 'gt_25k_p6',
'gt_28k_p6', 'cdr2', 'cdr3', 'rpy_3yr_rt_supp', 'lo_inc_rpy_3yr_rt_supp',
'md_inc_rpy_3yr_rt_supp', 'hi_inc_rpy_3yr_rt_supp', 'compl_rpy_3yr_rt_supp',
'noncom_rpy_3yr_rt_supp', 'dep_rpy_3yr_rt_supp', 'ind_rpy_3yr_rt_supp',
'pell_rpy_3yr_rt_supp', 'nopell_rpy_3yr_rt_supp', 'female_rpy_3yr_rt_supp',
'male_rpy_3yr_rt_supp', 'firstgen_rpy_3yr_rt_supp',
'notfirstgen_rpy_3yr_rt_supp', 'rpy_5yr_rt', 'lo_inc_rpy_5yr_rt',
'md_inc_rpy_5yr_rt', 'hi_inc_rpy_5yr_rt', 'compl_rpy_5yr_rt',
'noncom_rpy_5yr_rt', 'dep_rpy_5yr_rt', 'ind_rpy_5yr_rt', 'pell_rpy_5yr_rt',
'nopell_rpy_5yr_rt', 'female_rpy_5yr_rt', 'male_rpy_5yr_rt',
'firstgen_rpy_5yr_rt', 'notfirstgen_rpy_5yr_rt', 'rpy_7yr_rt',
'lo_inc_rpy_7yr_rt', 'md_inc_rpy_7yr_rt', 'hi_inc_rpy_7yr_rt',
'compl_rpy_7yr_rt', 'noncom_rpy_7yr_rt', 'dep_rpy_7yr_rt', 'ind_rpy_7yr_rt',
'pell_rpy_7yr_rt', 'nopell_rpy_7yr_rt', 'female_rpy_7yr_rt', 'male_rpy_7yr_rt',

```
'firstgen_rpy_7yr_rt', 'notfirstgen_rpy_7yr_rt', 'cdr2_denom', 'cdr3_denom',
'rpy_3yr_n', 'lo_inc_rpy_3yr_n', 'md_inc_rpy_3yr_n', 'hi_inc_rpy_3yr_n',
'compl_rpy_3yr_n', 'noncom_rpy_3yr_n', 'dep_rpy_3yr_n', 'ind_rpy_3yr_n',
'pell_rpy_3yr_n', 'nopell_rpy_3yr_n', 'female_rpy_3yr_n', 'male_rpy_3yr_n',
'firstgen_rpy_3yr_n', 'notfirstgen_rpy_3yr_n', 'rpy_5yr_n', 'lo_inc_rpy_5yr_n',
'md_inc_rpy_5yr_n', 'hi_inc_rpy_5yr_n', 'compl_rpy_5yr_n', 'noncom_rpy_5yr_n',
'dep_rpy_5yr_n', 'ind_rpy_5yr_n', 'pell_rpy_5yr_n', 'nopell_rpy_5yr_n',
'female_rpy_5yr_n', 'male_rpy_5yr_n', 'firstgen_rpy_5yr_n',
'notfirstgen_rpy_5yr_n', 'rpy_7yr_n', 'lo_inc_rpy_7yr_n', 'md_inc_rpy_7yr_n',
'hi_inc_rpy_7yr_n', 'compl_rpy_7yr_n', 'noncom_rpy_7yr_n', 'dep_rpy_7yr_n',
'ind_rpy_7yr_n', 'pell_rpy_7yr_n', 'nopell_rpy_7yr_n', 'female_rpy_7yr_n',
'male_rpy_7yr_n', 'firstgen_rpy_7yr_n', 'notfirstgen_rpy_7yr_n']
```

[5]:
```
print(df['data_yr_string'])
```

```
0         2000-01
1         2000-01
2         2000-01
3         2000-01
4         2000-01
           ...
139144    2018-19
139145    2018-19
139146    2018-19
139147    2018-19
139148    2018-19
Name: data_yr_string, Length: 139149, dtype: object
```

[6]:
```
df[['data_yr_string1','data_yr_string2']] = df['data_yr_string'].str.
 ↪split("-",expand=True,)
print(df['data_yr_string2'])
# I use df['data_yr_string2'] cus it is a School year coded as for example␣
 ↪2018-19,
# so I am using the last year
```

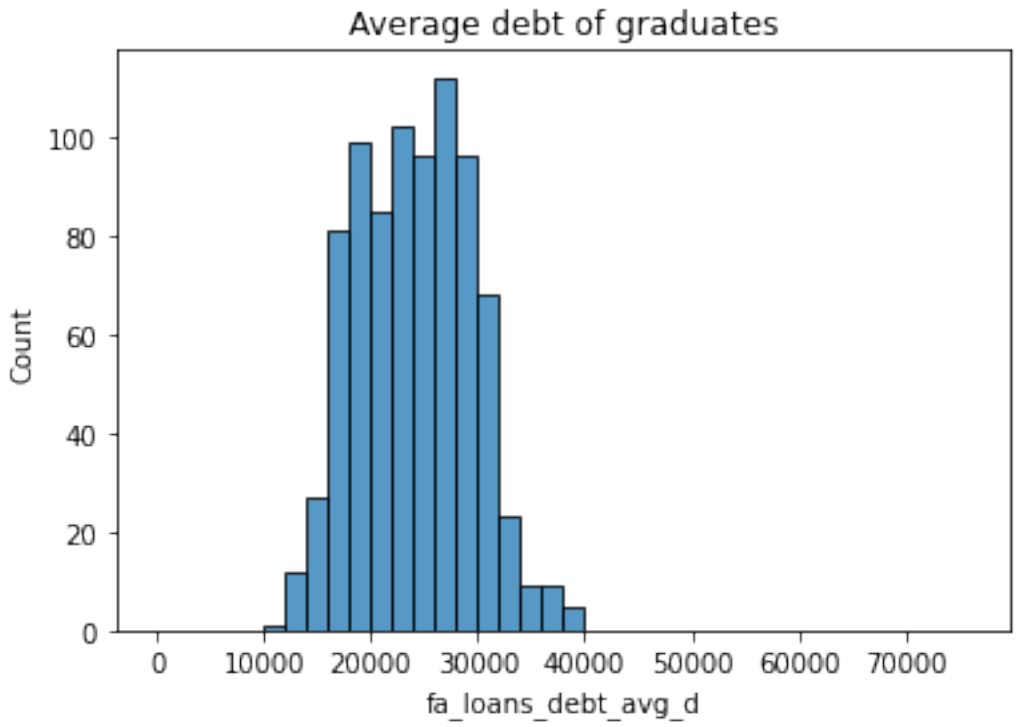
```
0         01
1         01
2         01
3         01
4         01
          ..
139144    19
139145    19
139146    19
139147    19
139148    19
Name: data_yr_string2, Length: 139149, dtype: object
```

[7]:
```
# entity type 6 is state total and entity type 10 is nation total
x = df[df['entity_type'] == 6]
y = df[df['entity_type'] == 6]
new_df = x + y
new_df.head()
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 867 entries, 145 to 139132
Columns: 410 entries, data_yr_string to data_yr_string2
dtypes: float64(396), int64(3), object(11)
memory usage: 2.7+ MB
```

[8]:
```
sns.histplot(data=x, x='fa_loans_debt_avg_d', binrange=(0,75000),␣
↪binwidth=2000).set_title('Average debt of graduates')
```

[8]:
```
Text(0.5, 1.0, 'Average debt of graduates')
```



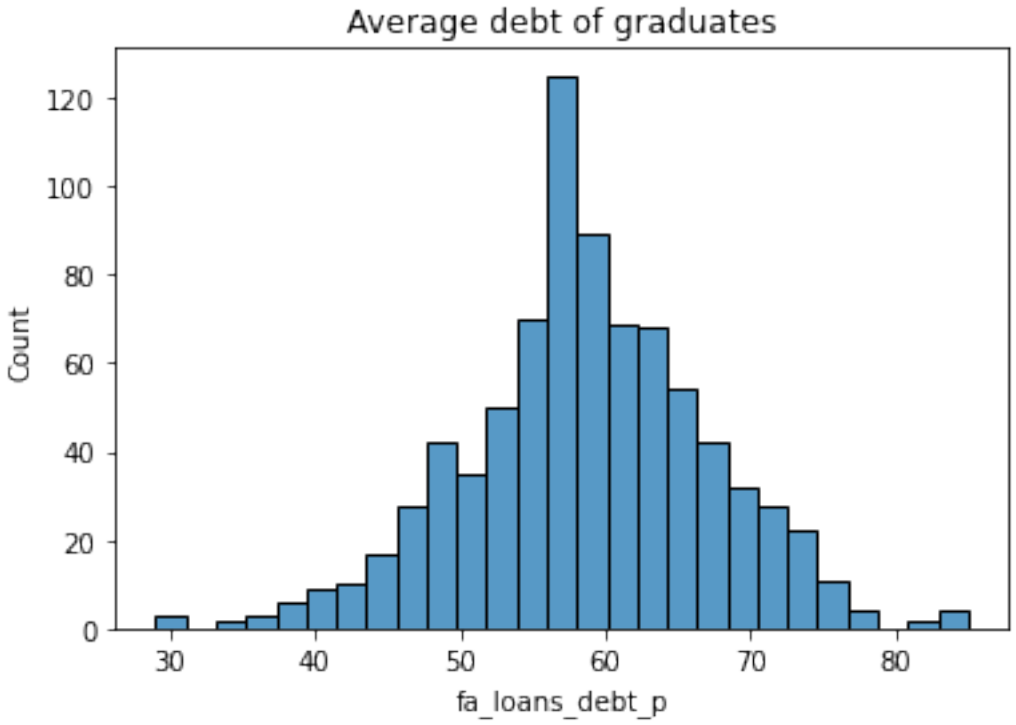
[9]:
```
x['fa_loans_debt_p'] = x['fa_loans_debt_p'] * 100
```

```
<ipython-input-9-891b78c0a8cd>: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
  x['fa_loans_debt_p'] = x['fa_loans_debt_p'] * 100
```

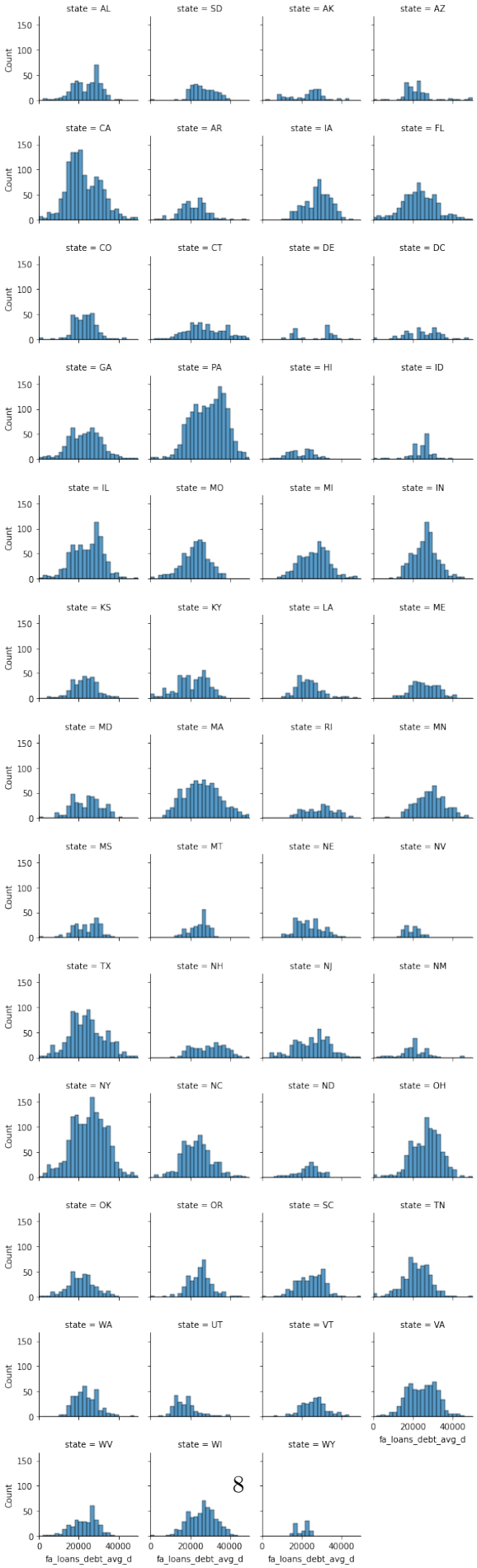
[10]:
```
#Percent of graduates from 4-year public and private nonprofit colleges who are␣
 ↪carrying student debt
sns.histplot(data=x, x='fa_loans_debt_p').set_title('Average debt of graduates')
```

[10]:
```
Text(0.5, 1.0, 'Average debt of graduates')
```



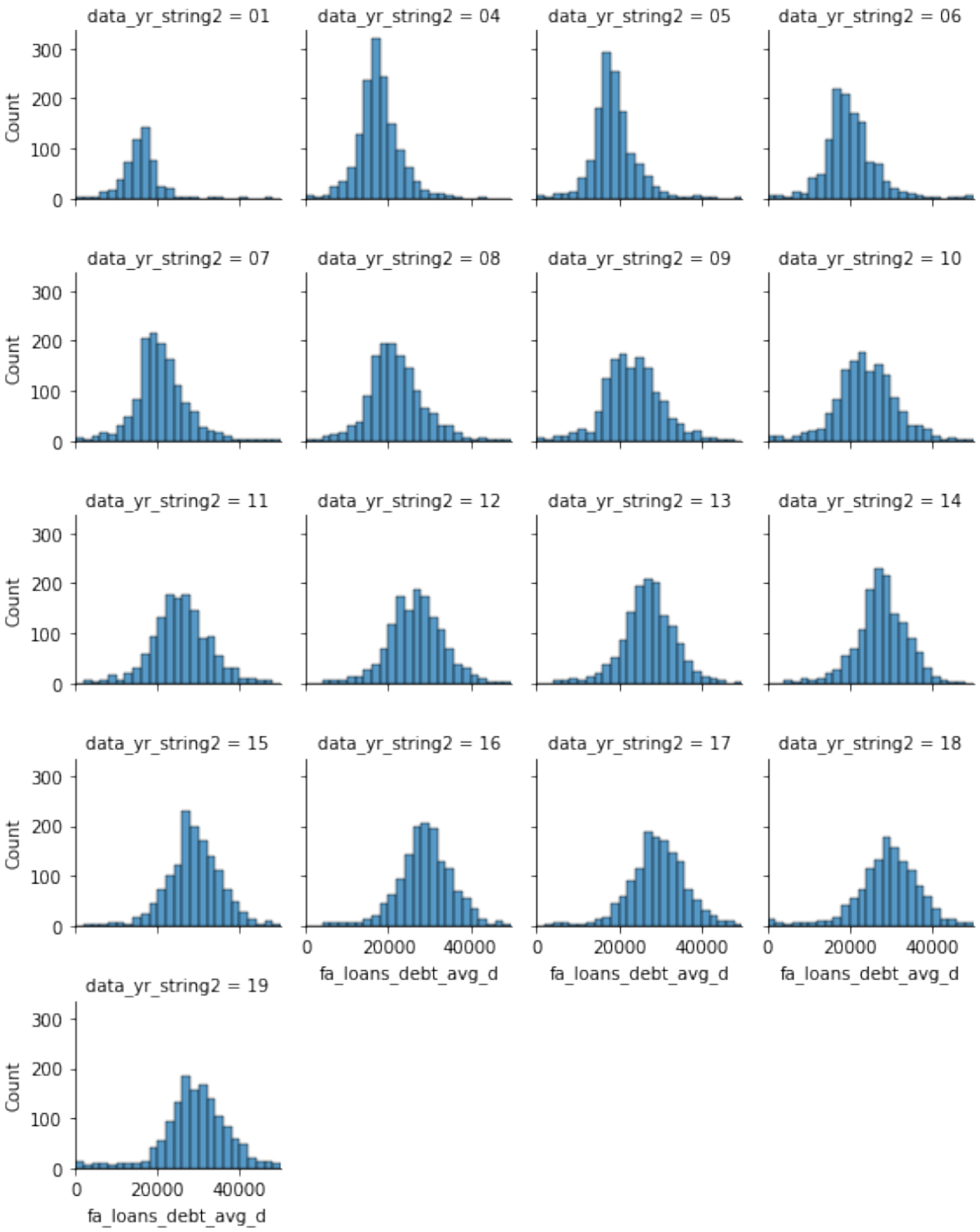
[18]:
```
## Average debt of graduates by state
g = sns.FacetGrid(df, col="state", col_wrap=4, height=2, xlim=(0,50000))
g.map(sns.histplot, "fa_loans_debt_avg_d",binrange=(0,50000), binwidth=2000)
```

[18]:
```
<seaborn.axisgrid.FacetGrid at 0x7f4096afb700>
```


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
Count
0
50
100
150
0
20000
40000
fa_loans_debt_avg_d

[19]:
```
## Average debt of graduates by state
g = sns.FacetGrid(df, col="data_yr_string2", col_wrap=4, height=2,␣
 ↪xlim=(0,50000))
g.map(sns.histplot, "fa_loans_debt_avg_d",binrange=(0,50000), binwidth=2000)
```

[19]:
```
<seaborn.axisgrid.FacetGrid at 0x7f408d3032e0>
```


data_yr_string2 = 01
data_yr_string2 = 04
data_yr_string2 = 05
data_yr_string2 = 06
data_yr_string2 = 07
data_yr_string2 = 08
data_yr_string2 = 09
data_yr_string2 = 10
data_yr_string2 = 11
data_yr_string2 = 12
data_yr_string2 = 13
data_yr_string2 = 14
data_yr_string2 = 15
data_yr_string2 = 16
data_yr_string2 = 17
data_yr_string2 = 18
data_yr_string2 = 19
Count
0
100
200
300
0
20000
40000
fa_loans_debt_avg_d