College Readiness:Admissions

May 12, 2021

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

1 Part A: College Readiness

1.1 Data Processing

[2]:
```
# Load each year of district/district data
sd_2010_2011 = pd.read_csv('district/district_2010-2011.csv')
sd_2011_2012 = pd.read_csv('district/district_2011-2012.csv')
sd_2012_2013 = pd.read_csv('district/district_2012-2013.csv')
sd_2013_2014 = pd.read_csv('district/district_2013-2014.csv')
sd_2014_2015 = pd.read_csv('district/district_2014-2015.csv')
sd_2015_2016 = pd.read_csv('district/district_2015-2016.csv')
sd_2016_2017 = pd.read_csv('district/district_2016-2017.csv')
sd_2017_2018 = pd.read_csv('district/district_2017-2018.csv')
sd_2018_2019 = pd.read_csv('district/district_2018-2019.csv')
```

[3]:
```
# Add a new column for the year each entry is from
sd_2010_2011 = sd_2010_2011.assign(Year = '2010-2011')
sd_2011_2012 = sd_2011_2012.assign(Year = '2011-2012')
sd_2012_2013 = sd_2012_2013.assign(Year = '2012-2013')
sd_2013_2014 = sd_2013_2014.assign(Year = '2013-2014')
sd_2014_2015 = sd_2014_2015.assign(Year = '2014-2015')
sd_2015_2016 = sd_2015_2016.assign(Year = '2015-2016')
sd_2016_2017 = sd_2016_2017.assign(Year = '2016-2017')
sd_2017_2018 = sd_2017_2018.assign(Year = '2017-2018')
sd_2018_2019 = sd_2018_2019.assign(Year = '2018-2019')
```

[4]:
```
# Concatenate the dataframes along axis 0 (vertically stacked by year)
sd_dfs = [sd_2010_2011, sd_2011_2012, sd_2012_2013, sd_2013_2014, sd_2014_2015,␣
↪sd_2015_2016, sd_2016_2017, sd_2017_2018, sd_2018_2019]
sd_all = pd.concat(sd_dfs, axis = 0)
```

[5]:
```
# Dataset containing all districts from each year
sd_all
```

[5]:
```
                                    District Name County Name (District)  \
0                     Calistoga Joint Unified (Napa)                  Napa
1                        Death Valley Unified (Inyo)                  Inyo
2                     Golden Valley Unified (Madera)                Madera
3                         Warner Unified (San Diego)             San Diego
4              Acton-Agua Dulce Unified (Los Angeles)           Los Angeles
..                                                ...                   ...
340                   Woodland Joint Unified (Yolo)                  Yolo
341                       Yosemite Unified (Madera)                Madera
342                      Yuba City Unified (Sutter)                Sutter
343  Yucaipa-Calimesa Joint Unified (San Bernardino)         San Bernardino
344                                             NaN                   NaN

    District Type (District) Census Day Enrollment (District)  \
0    Unified School District                              858
1    Unified School District                               51
2    Unified School District                             1925
3    Unified School District                              286
4    Unified School District                             1696
..                       ...                              ...
340  Unified School District                            10031
341  Unified School District                             2060
342  Unified School District                            13111
343  Unified School District                             9982
344                      NaN                              NaN

    English Learners % (District) Free/Reduced Meals % (District)  \
0                   Not Certified                             78.4
1                   Not Certified                             84.3
2                             6.2                             46.1
3                             8.4                             75.2
4                   Not Certified                             30.3
..                            ...                              ...
340                          23.3                             60.3
341                             2                             48.2
342                          22.1                             74.2
343                           7.6                             50.8
344                           NaN                              NaN

    Ethnic Diversity Index (District) Cohort Graduates % (District)  \
0                                   5                         82.19
1                                  28                           100
2                                  36                         91.03
3                                  56                            90
```

```
4                                   34                                  90
..                                  …                                   …
340                                 32                                92.3
341                                 36                                83.8
342                                 53                                85.8
343                                 39                                90.5
344                                NaN                                 NaN

    Grads Mtg UC/CSU % (District) Grads Mtg UC/CSU # (District)  …  \
0                             100                            63  …
1                        redacted                       redacted  …
2                             100                           143  …
3                             100                            31  …
4                            99.3                           138  …
..                            …                            …  …
340                          47.7                           327  …
341                          49.3                            74  …
342                          37.7                           314  …
343                          46.2                           306  …
344                           NaN                           NaN  …

     CAASPP-Math Standard Exceeded or Met (Levels 3 and 4) (District)  \
0                                                   NaN
1                                                   NaN
2                                                   NaN
3                                                   NaN
4                                                   NaN
..                                                  …
340                                               27.89
341                                               34.89
342                                               29.59
343                                               36.33
344                                                 NaN

     Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                12806.0
1                                                25438.0
2                                                 7633.0
3                                                13610.0
4                                                 7795.0
..                                                   …
340                                              12175.0
341                                              12771.0
342                                              12987.0
343                                              11629.0
344                                                  NaN
```

	Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0	604.0
1	1733.0
2	256.0
3	1019.0
4	431.0
..	…
340	422.0
341	587.0
342	793.0
343	346.0
344	NaN

	Total Gen Fund Revenues Per Student (District)
0	13952.0
1	27294.0
2	8320.0
3	15644.0
4	8361.0
..	…
340	13128.0
341	14195.0
342	13557.0
343	12174.0
344	NaN

	Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0	239.0
1	312.0
2	107.0
3	56.0
4	77.0
..	…
340	92.0
341	83.0
342	133.0
343	149.0
344	NaN

	Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
0	8543.0
1	12543.0
2	4329.0
3	7675.0
4	4778.0

```
..                                                    …
340                                                7653.0
341                                                6795.0
342                                                8195.0
343                                                7204.0
344                                                   NaN

    Gen Fund Exp by Activity - 5000-5999 Community Services Per Student #
(District)  \
0                                                    31
1                                                     0
2                                                     0
3                                                     0
4                                                     0
..                                                    …
340                                                  19
341                                                  38
342                                                   0
343                                                   0
344                                                 NaN

     Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                1250.0
1                                                4458.0
2                                                 781.0
3                                                1998.0
4                                                 687.0
..                                                    …
340                                               619.0
341                                              1157.0
342                                               679.0
343                                               661.0
344                                                 NaN

     Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                 781.0
1                                                4449.0
2                                                 533.0
3                                                2465.0
4                                                 909.0
..                                                    …
340                                              1296.0
341                                              1391.0
342                                              1114.0
343                                              1460.0
```

```
344                                                 NaN

          Year
0    2010-2011
1    2010-2011
2    2010-2011
3    2010-2011
4    2010-2011
..         …
340  2018-2019
341  2018-2019
342  2018-2019
343  2018-2019
344  2018-2019

[3076 rows x 26 columns]
```

[6]:
```
# Remove all districts with nan and specified string values
sd_all_cleaned = sd_all.copy()
for col in sd_all_cleaned.columns:
  sd_all_cleaned = sd_all_cleaned[~(sd_all_cleaned[col].isin(["redacted",␣
 ↪"(1)", "Not Certified"]) | pd.isna(sd_all_cleaned[col]))]

# Average CAASPP scores in one column as target variable
english = sd_all_cleaned["CAASPP-ELA Standard Exceeded or Met (Levels 3 and 4)␣
 ↪(District)"].astype(float)
math = sd_all_cleaned["CAASPP-Math Standard Exceeded or Met (Levels 3 and 4)␣
 ↪(District)"].astype(float)
sd_all_cleaned["CAASPP % Passing"] = (english + math) / 2
sd_all_cleaned
```

[6]:
```
                                    District Name County Name (District)  \
0                          ABC Unified (Los Angeles)            Los Angeles
2                          Alameda Unified (Alameda)                Alameda
3                       Albany City Unified (Alameda)                Alameda
4                      Alhambra Unified (Los Angeles)            Los Angeles
5                            Alpaugh Unified (Tulare)                 Tulare
..                                                 …                    …
339                        Woodlake Unified (Tulare)                Tulare
340                     Woodland Joint Unified (Yolo)                  Yolo
341                        Yosemite Unified (Madera)                Madera
342                       Yuba City Unified (Sutter)                Sutter
343  Yucaipa-Calimesa Joint Unified (San Bernardino)         San Bernardino

    District Type (District) Census Day Enrollment (District)  \
0    Unified School District                            20768
2    Unified School District                            11201
```

```
3    Unified School District                          3702
4    Unified School District                         17071
5    Unified School District                           793
..                       …                            …
339  Unified School District                          2118
340  Unified School District                         10031
341  Unified School District                          2060
342  Unified School District                         13111
343  Unified School District                          9982

    English Learners % (District) Free/Reduced Meals % (District)  \
0                            20.8                            49.3
2                            16.4                            28.6
3                            18.2                            17.3
4                            22.1                            64.5
5                            23.8                            72.5
..                            …                               …
339                          28.3                            88.9
340                          23.3                            60.3
341                             2                            48.2
342                          22.1                            74.2
343                           7.6                            50.8

    Ethnic Diversity Index (District) Cohort Graduates % (District)  \
0                                  56                           91.1
2                                  69                             87
3                                  60                           91.9
4                                  41                           94.2
5                                  37                           58.2
..                                 …                             …
339                                10                           85.9
340                                32                           92.3
341                                36                           83.8
342                                53                           85.8
343                                39                           90.5

    Grads Mtg UC/CSU % (District) Grads Mtg UC/CSU # (District)  …  \
0                            58.7                            910  …
2                              59                            461  …
3                              62                            176  …
4                            56.7                            995  …
5                             4.3                              2  …
..                             …                          …  …
339                          33.6                             45  …
340                          47.7                            327  …
341                          49.3                             74  …
342                          37.7                            314  …
```

```
343                          46.2                          306  …

     Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                   10410.0
2                                                   11570.0
3                                                   12165.0
4                                                   11511.0
5                                                   14629.0
..                                                      …
339                                                 13190.0
340                                                 12175.0
341                                                 12771.0
342                                                 12987.0
343                                                 11629.0

     Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                    439.0
2                                                    540.0
3                                                    405.0
4                                                    485.0
5                                                   1005.0
..                                                     …
339                                                  906.0
340                                                  422.0
341                                                  587.0
342                                                  793.0
343                                                  346.0

     Total Gen Fund Revenues Per Student (District)  \
0                                                 11160.0
2                                                 11919.0
3                                                 12649.0
4                                                 12495.0
5                                                 14325.0
..                                                    …
339                                               15803.0
340                                               13128.0
341                                               14195.0
342                                               13557.0
343                                               12174.0

     Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                    0.0
2                                                  200.0
3                                                  124.0
```

```
4                                                    64.0
5                                                   151.0
..                                                    …
339                                                 256.0
340                                                  92.0
341                                                  83.0
342                                                 133.0
343                                                 149.0

     Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                                  6761.0
2                                                  7634.0
3                                                  7952.0
4                                                  7204.0
5                                                  8688.0
..                                                    …
339                                                7624.0
340                                                7653.0
341                                                6795.0
342                                                8195.0
343                                                7204.0

    Gen Fund Exp by Activity - 5000-5999 Community Services Per Student #
(District)  \
0                                                     0
2                                                    14
3                                                   116
4                                                     1
5                                                     0
..                                                    …
339                                                  10
340                                                  19
341                                                  38
342                                                   0
343                                                   0

    Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                  580.0
2                                                  727.0
3                                                 1031.0
4                                                  722.0
5                                                 1750.0
..                                                    …
339                                                733.0
340                                                619.0
```

```
341                                                 1157.0
342                                                  679.0
343                                                  661.0

     Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                    849.0
2                                                    875.0
3                                                    559.0
4                                                   1381.0
5                                                    812.0
..                                                     …
339                                                 1664.0
340                                                 1296.0
341                                                 1391.0
342                                                 1114.0
343                                                 1460.0

          Year  CAASPP % Passing
0    2016-2017            59.270
2    2016-2017            62.050
3    2016-2017            76.625
4    2016-2017            59.240
5    2016-2017            12.990
..         …                …
339  2018-2019            28.945
340  2018-2019            35.360
341  2018-2019            43.655
342  2018-2019            38.630
343  2018-2019            42.790

[919 rows x 27 columns]
```

[7]:
```
# Split into training, validation and test sets
years = sd_all_cleaned["Year"].unique()

# Use 2016 and 2017 for training
sd_train = sd_all_cleaned[sd_all_cleaned["Year"].isin(years[:2])]

# Use 2018 for validation and testing
sd_2018 = sd_all_cleaned[sd_all_cleaned["Year"] == years[2]].sample(frac = 1) #␣
 ↪Shuffle remaining rows
num_half = len(sd_2018) // 2

# Assign half of the randomized data to validation and the other half to testing
sd_val = sd_2018.iloc[:num_half + 1]
sd_test = sd_2018.iloc[num_half + 1:]
```

```
# Choose the features to be used
cols = list(sd_all_cleaned.columns)[3:len(sd_all_cleaned.columns)-2]
cols.remove("CAASPP-ELA Standard Exceeded or Met (Levels 3 and 4) (District)")
cols.remove("CAASPP-Math Standard Exceeded or Met (Levels 3 and 4) (District)")

# Split into X and y sets
X_train = sd_train[cols].astype(float)
y_train = sd_train["CAASPP % Passing"]

X_val = sd_val[cols].astype(float)
y_val = sd_val["CAASPP % Passing"]

X_test = sd_test[cols].astype(float)
y_test = sd_test["CAASPP % Passing"]
X_train
```

[7]:

	Census Day Enrollment (District)	English Learners % (District)	\
0	20768.0	20.8	
2	11201.0	16.4	
3	3702.0	18.2	
4	17071.0	22.1	
5	793.0	23.8	
..	…	…	
338	2135.0	29.5	
339	10041.0	24.5	
340	2176.0	2.1	
341	13236.0	22.9	
342	10063.0	7.8	

	Free/Reduced Meals % (District)	Ethnic Diversity Index (District)	\
0	49.3	56.0	
2	28.6	69.0	
3	17.3	60.0	
4	64.5	41.0	
5	72.5	37.0	
..	…	…	
338	89.3	9.0	
339	62.4	33.0	
340	49.5	34.0	
341	72.1	53.0	
342	51.6	38.0	

	Cohort Graduates % (District)	Grads Mtg UC/CSU % (District)	\
0	91.1	58.7	
2	87.0	59.0	
3	91.9	62.0	

```
4                              94.2                              56.7
5                              58.2                               4.3
..                              …                                 …
338                            88.7                               9.0
339                            89.6                              30.5
340                            87.2                              36.9
341                            86.7                              35.7
342                            92.1                              42.0

     Grads Mtg UC/CSU # (District)  ACT Test Takers # (District)  \
0                             910.0                          361.0
2                             461.0                          275.0
3                             176.0                          128.0
4                             995.0                          511.0
5                               2.0                            2.0
..                              …                              …
338                            12.0                           34.0
339                           205.0                          126.0
340                            73.0                           42.0
341                           291.0                          148.0
342                           305.0                           74.0

     Per Pupil Ratio: Teacher (District)  Avg Years Teaching (District)  \
0                                   24.0                            14.0
2                                   19.1                             9.0
3                                   18.6                             8.0
4                                   22.9                            14.0
5                                   23.8                             3.0
..                                   …                               …
338                                 21.0                            11.0
339                                 19.0                            11.0
340                                 21.0                            12.0
341                                 18.0                            12.0
342                                 21.0                            10.0

     Teacher Salary-Avg (District)  Chronic Absenteeism % (District)  \
0                          82191.0                              7.0
2                          73140.0                              9.0
3                          72154.0                              4.9
4                          82543.0                              6.9
5                          57398.0                             24.1
..                             …                                …
338                        68000.0                              9.9
339                        69150.0                             15.2
340                        65298.0                             14.4
341                        73168.0                             12.3
342                        84535.0                             11.6
```

```
     Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                  10410.0
2                                                  11570.0
3                                                  12165.0
4                                                  11511.0
5                                                  14629.0
..                                                     …
338                                                12470.0
339                                                11751.0
340                                                11312.0
341                                                11826.0
342                                                10933.0

     Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                   439.0
2                                                   540.0
3                                                   405.0
4                                                   485.0
5                                                  1005.0
..                                                     …
338                                                 693.0
339                                                 805.0
340                                                 514.0
341                                                 811.0
342                                                 461.0

     Total Gen Fund Revenues Per Student (District)  \
0                                               11160.0
2                                               11919.0
3                                               12649.0
4                                               12495.0
5                                               14325.0
..                                                  …
338                                             14110.0
339                                             11842.0
340                                             11634.0
341                                             11729.0
342                                             11117.0

     Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                   0.0
2                                                 200.0
3                                                 124.0
4                                                  64.0
```

```
5                                                151.0
..                                                 …
338                                              261.0
339                                               90.0
340                                                0.0
341                                              119.0
342                                              113.0

     Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                               6761.0
2                                               7634.0
3                                               7952.0
4                                               7204.0
5                                               8688.0
..                                                 …
338                                             7177.0
339                                             7430.0
340                                             6074.0
341                                             7270.0
342                                             6797.0

     Gen Fund Exp by Activity - 5000-5999 Community Services Per Student #
(District)  \
0                                                 0.0
2                                                14.0
3                                               116.0
4                                                 1.0
5                                                 0.0
..                                                 …
338                                               4.0
339                                              23.0
340                                              68.0
341                                               0.0
342                                               0.0

     Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                580.0
2                                                727.0
3                                               1031.0
4                                                722.0
5                                               1750.0
..                                                 …
338                                              733.0
339                                              779.0
340                                             1071.0
```

```
341                                                 671.0
342                                                 663.0

     Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)
0                                                   849.0
2                                                   875.0
3                                                   559.0
4                                                  1381.0
5                                                   812.0
..                                                    …
338                                                1702.0
339                                                1160.0
340                                                1207.0
341                                                1009.0
342                                                1271.0

[611 rows x 20 columns]
```

1.2 Linear Regression

[8]:
```
import statsmodels.api as sm

X1 = sm.add_constant(X_train)

linreg1 = sm.OLS(y_train, X1).fit()
print(linreg1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.862
Model:                            OLS   Adj. R-squared:                  0.857
Method:                 Least Squares   F-statistic:                     184.2
Date:                Wed, 12 May 2021   Prob (F-statistic):          2.77e-238
Time:                        14:58:23   Log-Likelihood:                -1947.9
No. Observations:                 611   AIC:                             3938.
Df Residuals:                     590   BIC:                             4030.
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
47.3247      4.753      9.957      0.000      37.990      56.660
```

```
Census Day Enrollment (District)
-0.0002   5.97e-05     -3.081      0.002      -0.000   -6.67e-05
English Learners % (District)
-0.0950      0.029     -3.261      0.001      -0.152      -0.038
Free/Reduced Meals % (District)
-0.3595      0.021    -17.314      0.000      -0.400      -0.319
Ethnic Diversity Index (District)
0.0512      0.022      2.343      0.019       0.008       0.094
Cohort Graduates % (District)
0.0050      0.032      0.158      0.875      -0.057       0.067
Grads Mtg UC/CSU % (District)
0.1829      0.021      8.630      0.000       0.141       0.225
Grads Mtg UC/CSU # (District)
0.0053      0.002      3.109      0.002       0.002       0.009
ACT Test Takers # (District)
-0.0007      0.002     -0.461      0.645      -0.004       0.002
Per Pupil Ratio: Teacher (District)
-0.1380      0.077     -1.795      0.073      -0.289       0.013
Avg Years Teaching (District)
-0.0275      0.127     -0.216      0.829      -0.277       0.223
Teacher Salary-Avg (District)
0.0002    3.7e-05      5.590      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6130      0.068     -9.055      0.000      -0.746      -0.480
Current Exp of Educ per ADA (Ed Code 41372) (District)
-3.449e-05      0.001     -0.050      0.960      -0.001       0.001
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0.0009      0.001      0.657      0.512      -0.002       0.004
Total Gen Fund Revenues Per Student (District)
-0.0004      0.000     -0.886      0.376      -0.001       0.000
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0069      0.003      2.413      0.016       0.001       0.013
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
0.0006      0.001      0.804      0.422      -0.001       0.002
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
0.0034      0.005      0.708      0.479      -0.006       0.013
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    -0.0013      0.001     -0.970      0.333      -0.004       0.001
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
0.0015      0.001      1.350      0.177      -0.001       0.004
==============================================================================
Omnibus:                        4.094   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.129   Jarque-Bera (JB):                4.322
Skew:                          -0.113   Prob(JB):                        0.115
Kurtosis:                       3.345   Cond. No.                     1.57e+06
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.57e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

We want to remove the features with high multicollinearity by examining their VIF values. We also keep an eye out for drastic changes in the R2 score.

[9]:
```
# Import VIF package
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools import add_constant

# Examine the Variance Inflation Factor for each feature
def VIF(df, features):
    values = add_constant(df[features]).values
    vif = [variance_inflation_factor(values, i) for i in range(len(features)+1)]
    return pd.Series(vif[1:], index = features)

VIF(X1, X_train.columns)
```

[9]:
```
Census Day Enrollment (District)
88.660833
English Learners % (District)
2.247280
Free/Reduced Meals % (District)
3.912751
Ethnic Diversity Index (District)
1.997281
Cohort Graduates % (District)
1.383734
Grads Mtg UC/CSU % (District)
2.261284
Grads Mtg UC/CSU # (District)
99.807969
ACT Test Takers # (District)
21.980686
Per Pupil Ratio: Teacher (District)
1.385171
Avg Years Teaching (District)
1.562008
Teacher Salary-Avg (District)
2.499295
Chronic Absenteeism % (District)
2.209894
Current Exp of Educ per ADA (Ed Code 41372) (District)
44.560756
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
2.046311
```

```
Total Gen Fund Revenues Per Student (District)
18.421001
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
1.380040
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
17.981814
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
1.127043
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     4.283697
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
2.726350
dtype: float64
```

[10]:
```
X_train2 = X_train.drop("Grads Mtg UC/CSU # (District)", axis = 1)
X2 = sm.add_constant(X_train2)

linreg2 = sm.OLS(y_train, X2).fit()
print(linreg2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.860
Model:                           OLS   Adj. R-squared:                  0.855
Method:                Least Squares   F-statistic:                     190.6
Date:               Wed, 12 May 2021   Prob (F-statistic):          2.33e-237
Time:                       14:58:23   Log-Likelihood:                -1952.8
No. Observations:                611   AIC:                             3946.
Df Residuals:                    591   BIC:                             4034.
Df Model:                         19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
46.4820      4.780      9.724      0.000      37.094      55.870
Census Day Enrollment (District)
-2.013e-05   2.83e-05     -0.712      0.477   -7.57e-05    3.54e-05
English Learners % (District)
-0.0972      0.029     -3.311      0.001      -0.155      -0.040
Free/Reduced Meals % (District)
-0.3672      0.021    -17.688      0.000      -0.408      -0.326
Ethnic Diversity Index (District)
0.0522      0.022      2.372      0.018       0.009       0.095
Cohort Graduates % (District)
```

```
0.0157      0.032      0.495      0.620      -0.047       0.078
Grads Mtg UC/CSU % (District)
0.2009      0.021      9.785      0.000       0.161       0.241
ACT Test Takers # (District)
0.0010      0.001      0.676      0.499      -0.002       0.004
Per Pupil Ratio: Teacher (District)
-0.1331      0.077     -1.719      0.086      -0.285       0.019
Avg Years Teaching (District)
-0.0296      0.128     -0.231      0.817      -0.281       0.222
Teacher Salary-Avg (District)
0.0002   3.73e-05      5.412      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6264      0.068     -9.205      0.000      -0.760      -0.493
Current Exp of Educ per ADA (Ed Code 41372) (District)
-0.0002      0.001     -0.245      0.806      -0.002       0.001
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0.0011      0.001      0.787      0.432      -0.002       0.004
Total Gen Fund Revenues Per Student (District)
-0.0004      0.000     -0.897      0.370      -0.001       0.000
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0076      0.003      2.632      0.009       0.002       0.013
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
0.0007      0.001      0.940      0.348      -0.001       0.002
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
0.0038      0.005      0.787      0.431      -0.006       0.013
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    -0.0007      0.001     -0.522      0.602      -0.003       0.002
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
0.0016      0.001      1.397      0.163      -0.001       0.004
==============================================================================
Omnibus:                        3.361   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.186   Jarque-Bera (JB):                3.407
Skew:                          -0.104   Prob(JB):                        0.182
Kurtosis:                       3.301   Cond. No.                     1.57e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.57e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[11]:
```
VIF(X2, X_train2.columns)
```

[11]:
```
Census Day Enrollment (District)
19.617688
English Learners % (District)
```

```
2.246039
Free/Reduced Meals % (District)
3.856113
Ethnic Diversity Index (District)
1.996854
Cohort Graduates % (District)
1.367340
Grads Mtg UC/CSU % (District)
2.092020
ACT Test Takers # (District)
19.242446
Per Pupil Ratio: Teacher (District)
1.384598
Avg Years Teaching (District)
1.561961
Teacher Salary-Avg (District)
2.493882
Chronic Absenteeism % (District)
2.200857
Current Exp of Educ per ADA (Ed Code 41372) (District)
44.383059
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
2.042468
Total Gen Fund Revenues Per Student (District)
18.420386
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
1.372408
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
17.944489
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
1.126211
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     4.194304
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
2.725455
dtype: float64
```

[12]:
```
X_train3 = X_train2.drop("Current Exp of Educ per ADA (Ed Code 41372)␣
 ↪(District)", axis = 1)
X3 = sm.add_constant(X_train3)

linreg3 = sm.OLS(y_train, X3).fit()
print(linreg3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.860
```

```
Model:                            OLS   Adj. R-squared:                  0.855
Method:                 Least Squares   F-statistic:                     201.5
Date:                Wed, 12 May 2021   Prob (F-statistic):          1.67e-238
Time:                        14:58:23   Log-Likelihood:                -1952.9
No. Observations:                 611   AIC:                             3944.
Df Residuals:                     592   BIC:                             4028.
Df Model:                          18
Covariance Type:            nonrobust
================================================================================
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
const
46.4450      4.774      9.729      0.000      37.069      55.821
Census Day Enrollment (District)
-2.026e-05   2.82e-05     -0.717      0.473   -7.57e-05    3.52e-05
English Learners % (District)
-0.0976      0.029     -3.335      0.001      -0.155      -0.040
Free/Reduced Meals % (District)
-0.3673      0.021    -17.703      0.000      -0.408      -0.327
Ethnic Diversity Index (District)
0.0518      0.022      2.362      0.019       0.009       0.095
Cohort Graduates % (District)
0.0160      0.032      0.506      0.613      -0.046       0.078
Grads Mtg UC/CSU % (District)
0.2005      0.020      9.805      0.000       0.160       0.241
ACT Test Takers # (District)
0.0010      0.001      0.678      0.498      -0.002       0.004
Per Pupil Ratio: Teacher (District)
-0.1312      0.077     -1.705      0.089      -0.282       0.020
Avg Years Teaching (District)
-0.0285      0.128     -0.223      0.824      -0.280       0.223
Teacher Salary-Avg (District)
0.0002   3.72e-05      5.412      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6274      0.068     -9.243      0.000      -0.761      -0.494
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0.0011      0.001      0.788      0.431      -0.002       0.004
Total Gen Fund Revenues Per Student (District)
-0.0004      0.000     -1.193      0.233      -0.001       0.000
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0074      0.003      2.634      0.009       0.002       0.013
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
0.0006      0.001      1.124      0.261      -0.000       0.002
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
0.0039      0.005      0.828      0.408      -0.005       0.013
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
```

```
(District)    -0.0009      0.001     -0.733      0.464      -0.003       0.001
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
0.0014      0.001      1.485      0.138      -0.000       0.003
==============================================================================
Omnibus:                        3.404   Durbin-Watson:                   1.910
Prob(Omnibus):                  0.182   Jarque-Bera (JB):                3.466
Skew:                          -0.104   Prob(JB):                        0.177
Kurtosis:                       3.305   Cond. No.                     1.55e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.55e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[13]:
```
VIF(X3, X_train3.columns)
```

[13]:
```
Census Day Enrollment (District)
19.610662
English Learners % (District)
2.237411
Free/Reduced Meals % (District)
3.856102
Ethnic Diversity Index (District)
1.983559
Cohort Graduates % (District)
1.365018
Grads Mtg UC/CSU % (District)
2.078324
ACT Test Takers # (District)
19.241066
Per Pupil Ratio: Teacher (District)
1.370235
Avg Years Teaching (District)
1.560150
Teacher Salary-Avg (District)
2.492504
Chronic Absenteeism % (District)
2.193175
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
2.042444
Total Gen Fund Revenues Per Student (District)
13.557217
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
1.326501
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
```

```
8.196267
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
1.106483
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     3.212579
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
1.996710
dtype: float64
```

[14]:
```
X_train4 = X_train3.drop("Census Day Enrollment (District)", axis = 1)
X4 = sm.add_constant(X_train4)

linreg4 = sm.OLS(y_train, X4).fit()
print(linreg4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.860
Model:                           OLS   Adj. R-squared:                  0.856
Method:                Least Squares   F-statistic:                     213.5
Date:               Wed, 12 May 2021   Prob (F-statistic):          1.45e-239
Time:                       14:58:23   Log-Likelihood:                -1953.1
No. Observations:                611   AIC:                             3942.
Df Residuals:                    593   BIC:                             4022.
Df Model:                         17
Covariance Type:            nonrobust
================================================================================
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
const
46.5388      4.770      9.756      0.000      37.171      55.907
English Learners % (District)
-0.0957      0.029     -3.286      0.001      -0.153      -0.039
Free/Reduced Meals % (District)
-0.3710      0.020    -18.471      0.000      -0.410      -0.332
Ethnic Diversity Index (District)
0.0500      0.022      2.297      0.022       0.007       0.093
Cohort Graduates % (District)
0.0180      0.031      0.571      0.569      -0.044       0.080
Grads Mtg UC/CSU % (District)
0.2013      0.020      9.859      0.000       0.161       0.241
ACT Test Takers # (District)
-2.377e-05      0.000     -0.068      0.946      -0.001       0.001
Per Pupil Ratio: Teacher (District)
-0.1299      0.077     -1.689      0.092      -0.281       0.021
```

```
Avg Years Teaching (District)
-0.0346      0.128     -0.271      0.786      -0.285       0.216
Teacher Salary-Avg (District)
0.0002   3.72e-05      5.402      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6287      0.068     -9.268      0.000      -0.762      -0.495
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0.0012      0.001      0.809      0.419      -0.002       0.004
Total Gen Fund Revenues Per Student (District)
-0.0004      0.000     -1.179      0.239      -0.001       0.000
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0073      0.003      2.601      0.010       0.002       0.013
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
0.0006      0.001      1.092      0.275      -0.000       0.002
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
0.0039      0.005      0.820      0.413      -0.005       0.013
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    -0.0008      0.001     -0.668      0.505      -0.003       0.002
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
0.0014      0.001      1.476      0.140      -0.000       0.003
==============================================================================
Omnibus:                        3.171   Durbin-Watson:                   1.907
Prob(Omnibus):                  0.205   Jarque-Bera (JB):                3.202
Skew:                          -0.098   Prob(JB):                        0.202
Kurtosis:                       3.296   Cond. No.                     1.52e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.52e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[15]:
```
VIF(X4, X_train4.columns)
```

[15]:
```
English Learners % (District)
2.219225
Free/Reduced Meals % (District)
3.617011
Ethnic Diversity Index (District)
1.959312
Cohort Graduates % (District)
1.354885
Grads Mtg UC/CSU % (District)
2.072834
ACT Test Takers # (District)
1.172989
```

```
Per Pupil Ratio: Teacher (District)
1.369523
Avg Years Teaching (District)
1.553293
Teacher Salary-Avg (District)
2.491684
Chronic Absenteeism % (District)
2.191726
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
2.040861
Total Gen Fund Revenues Per Student (District)
13.551558
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
1.322838
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
8.178234
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
1.106337
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     3.183279
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
1.996354
dtype: float64
```

[16]:
```
X_train5 = X_train4.drop("Total Gen Fund Revenues Per Student (District)", axis␣
 ↪= 1)
X5 = sm.add_constant(X_train5)

linreg5 = sm.OLS(y_train, X5).fit()
print(linreg5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.859
Model:                            OLS   Adj. R-squared:                  0.855
Method:                 Least Squares   F-statistic:                     226.6
Date:                Wed, 12 May 2021   Prob (F-statistic):          1.90e-240
Time:                        14:58:23   Log-Likelihood:                -1953.9
No. Observations:                 611   AIC:                             3942.
Df Residuals:                     594   BIC:                             4017.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
==========================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------
--------------------------------------------------------------------------
```

```
const
46.2185      4.764      9.702      0.000      36.862      55.575
English Learners % (District)
-0.0970      0.029     -3.329      0.001      -0.154      -0.040
Free/Reduced Meals % (District)
-0.3747      0.020    -18.888      0.000      -0.414      -0.336
Ethnic Diversity Index (District)
0.0505      0.022      2.319      0.021       0.008       0.093
Cohort Graduates % (District)
0.0194      0.031      0.615      0.539      -0.042       0.081
Grads Mtg UC/CSU % (District)
0.2028      0.020      9.949      0.000       0.163       0.243
ACT Test Takers # (District)
-6.313e-05      0.000     -0.181      0.857      -0.001       0.001
Per Pupil Ratio: Teacher (District)
-0.1234      0.077     -1.608      0.108      -0.274       0.027
Avg Years Teaching (District)
-0.0325      0.128     -0.254      0.799      -0.283       0.218
Teacher Salary-Avg (District)
0.0002   3.72e-05      5.347      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6175      0.067     -9.191      0.000      -0.749      -0.486
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
0.0008      0.001      0.599      0.550      -0.002       0.004
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0066      0.003      2.411      0.016       0.001       0.012
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
9.333e-05      0.000      0.282      0.778      -0.001       0.001
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
0.0026      0.005      0.561      0.575      -0.006       0.012
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    -0.0015      0.001     -1.507      0.132      -0.003       0.000
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
0.0010      0.001      1.087      0.278      -0.001       0.003
==============================================================================
Omnibus:                        4.231   Durbin-Watson:                   1.904
Prob(Omnibus):                  0.121   Jarque-Bera (JB):                4.618
Skew:                          -0.102   Prob(JB):                       0.0994
Kurtosis:                       3.374   Cond. No.                     1.50e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.5e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[17]:
```
VIF(X5, X_train5.columns)
```

[17]:
```
English Learners % (District)
2.216392
Free/Reduced Meals % (District)
3.526692
Ethnic Diversity Index (District)
1.958652
Cohort Graduates % (District)
1.352983
Grads Mtg UC/CSU % (District)
2.064775
ACT Test Takers # (District)
1.162352
Per Pupil Ratio: Teacher (District)
1.362337
Avg Years Teaching (District)
1.552970
Teacher Salary-Avg (District)
2.485253
Chronic Absenteeism % (District)
2.148817
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
1.969585
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
1.268037
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
3.260921
Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District)
1.046808
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    2.311511
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
1.658012
dtype: float64
```

After removing the variables with high VIF values, we now want to remove variables that are not significant based on their p-values. We will remove all variables with p-value > 0.05. For simplicity, we removed the variables all at once, but we did check that each variable is still insignificant after each removal.

[18]:
```
# Remove variables with high p-values (>0.05)
p_cols = ["Cohort Graduates % (District)", "ACT Test Takers # (District)", "Per␣
↪Pupil Ratio: Teacher (District)", "Avg Years Teaching (District)",
          "Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student␣
↪(District)",
```

```
        "Gen Fund Exp by Activity - 1000-1999 Instruction Per Student #␣
↪(District)",
        "Gen Fund Exp by Activity - 5000-5999 Community Services Per Student␣
↪# (District)",
        "Gen Fund Exp by Activity - 7000-7999 General Administration Per␣
↪Student # (District)",
        "Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #␣
↪(District)"]

X_train6 = X_train5.drop(p_cols, axis = 1)
X6 = sm.add_constant(X_train6)

linreg6 = sm.OLS(y_train, X6).fit()
print(linreg6.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.857
Model:                            OLS   Adj. R-squared:                  0.856
Method:                 Least Squares   F-statistic:                     518.4
Date:                Wed, 12 May 2021   Prob (F-statistic):          2.48e-250
Time:                        14:58:23   Log-Likelihood:                -1957.6
No. Observations:                 611   AIC:                             3931.
Df Residuals:                     603   BIC:                             3967.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                                                                              
==============================================================================
                                                                              
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
45.7131      2.755     16.591      0.000      40.302      51.124
English Learners % (District)
-0.0917      0.028     -3.243      0.001      -0.147      -0.036
Free/Reduced Meals % (District)
-0.3786      0.019    -20.135      0.000      -0.416      -0.342
Ethnic Diversity Index (District)
0.0446       0.020      2.209      0.028       0.005       0.084
Grads Mtg UC/CSU % (District)
0.2034       0.020     10.350      0.000       0.165       0.242
Teacher Salary-Avg (District)
0.0002    2.79e-05      7.249      0.000       0.000       0.000
Chronic Absenteeism % (District)
-0.6014      0.060     -9.963      0.000      -0.720      -0.483
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
0.0080       0.003      3.143      0.002       0.003       0.013
```

```
==============================================================================
Omnibus:                        3.293   Durbin-Watson:                   1.887
Prob(Omnibus):                  0.193   Jarque-Bera (JB):                3.252
Skew:                          -0.115   Prob(JB):                        0.197
Kurtosis:                       3.274   Cond. No.                     8.62e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 8.62e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Our final linear regression model has a high R2 score of 0.857, so we will consider using it in our blended model.

1.3 Decision Tree

[19]:
```
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold

grid_values = {'ccp_alpha': np.linspace(0, 1, 500), # 500 cp values in [0,1]
               'min_samples_leaf': [5],
               'min_samples_split': [20],
               'max_depth': [30],
               'random_state': [88]}

dtr = DecisionTreeRegressor()

cv = KFold(n_splits = 5, random_state = 88, shuffle = True) # 5 fold cross␣
↪validation for each cp value

dtr_cv = GridSearchCV(dtr, param_grid = grid_values, scoring = 'r2', cv = cv,␣
↪verbose = 1)

dtr_cv.fit(X_train, y_train)
```

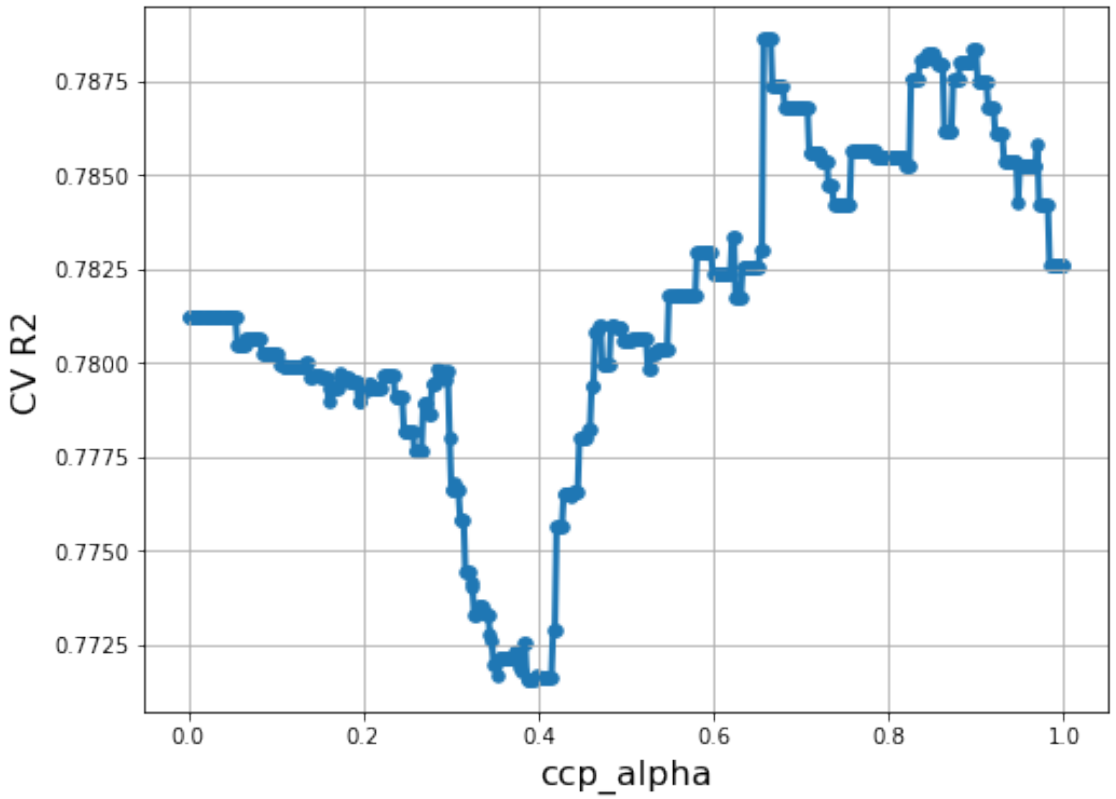
```
Fitting 5 folds for each of 500 candidates, totalling 2500 fits

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 2500 out of 2500 | elapsed:   11.9s finished
```

[19]:
```
GridSearchCV(cv=KFold(n_splits=5, random_state=88, shuffle=True),
             estimator=DecisionTreeRegressor(),
             param_grid={'ccp_alpha': array([0.        , 0.00200401, 0.00400802,
0.00601202, 0.00801603,
```

```
0.01002004, 0.01202405, 0.01402806, 0.01603206, 0.01803607,
0.02004008, 0.02204409, 0.0240481 , 0.0260521 , 0.02805611,
0.03006012, 0.03206413, 0.03406814, 0.03607214, 0.03807615,
0.04008016, 0…
0.95190381, 0.95390782, 0.95591182, 0.95791583, 0.95991984,
0.96192385, 0.96392786, 0.96593186, 0.96793587, 0.96993988,
0.97194389, 0.9739479 , 0.9759519 , 0.97795591, 0.97995992,
0.98196393, 0.98396794, 0.98597194, 0.98797595, 0.98997996,
0.99198397, 0.99398798, 0.99599198, 0.99799599, 1.        ]),
                  'max_depth': [30], 'min_samples_leaf': [5],
                  'min_samples_split': [20], 'random_state': [88]},
      scoring='r2', verbose=1)
```

Below is the plot examining the R2 for each ccp_alpha between 0 and 1. We will choose the optimal complexity parameter based on the highest CV R2.

[20]:
```
ccp_alpha_dtr = dtr_cv.cv_results_['param_ccp_alpha'].data
r2_scores = dtr_cv.cv_results_['mean_test_score']

plt.figure(figsize=(8, 6))
plt.xlabel('ccp_alpha', fontsize=16)
plt.ylabel('CV R2', fontsize=16)
plt.scatter(ccp_alpha_dtr, r2_scores, s=30)
plt.plot(ccp_alpha_dtr, r2_scores, linewidth=3)
plt.grid(True, which='both')

plt.show()
```



[21]:
```
print('Best ccp_alpha', dtr_cv.best_params_)
```

```
Best ccp_alpha {'ccp_alpha': 0.657314629258517, 'max_depth': 30,
'min_samples_leaf': 5, 'min_samples_split': 20, 'random_state': 88}
```

[22]:
```
dtr_cv.best_score_
```

[22]:
```
0.7886094987036202
```

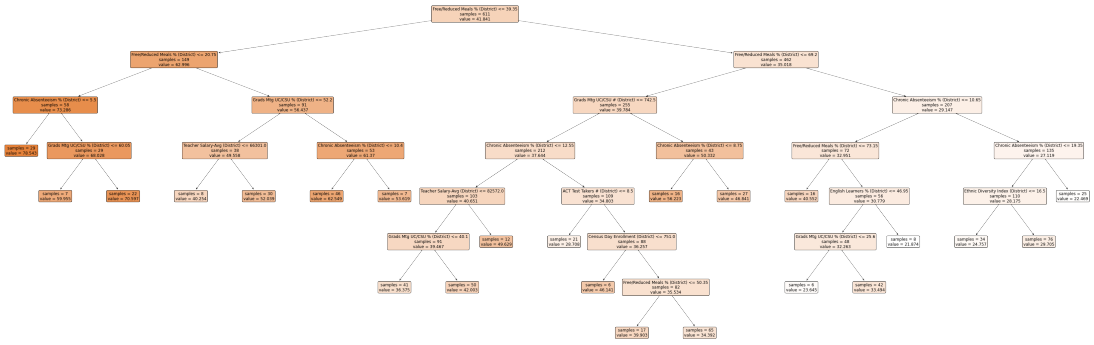
We observe a moderately high R2 for our optimal ccp_alpha of 0.657. So, we will consider using our decision tree regressor in our blended model.

Below is a visualization of our decision tree, seeing which features are considered first when determining the CAASPP % Passing for each district.

[23]:
```
from sklearn.tree import plot_tree

ccp_dtr = dtr_cv.best_estimator_

print('Node count =', ccp_dtr.tree_.node_count)

plt.figure(figsize=(60,20))
```

```
plot_tree(ccp_dtr,
          feature_names=X_train.columns,
          class_names=['0','1'],
          filled=True,
          impurity=False,
          rounded=True,
          fontsize=12)
plt.show()
```

```
Node count = 45
```



1.4 Random Forest

[24]:
```
from sklearn.ensemble import RandomForestRegressor

grid_values2 = {'max_features': np.linspace(1, len(X_train.columns),␣
 ↪len(X_train.columns), dtype='int32'), # Consider 1 to the total number of␣
 ↪features
                'min_samples_leaf': [5],
                'n_estimators': [500],
                'random_state': [88]}

rf_cv = RandomForestRegressor()

cv = KFold(n_splits = 5, random_state = 88, shuffle = True) # 5 fold cross␣
 ↪validation for each max_features value

rf_cv = GridSearchCV(rf_cv, param_grid = grid_values2, scoring = 'r2', cv = cv,␣
 ↪verbose = 1)

rf_cv.fit(X_train, y_train)
```

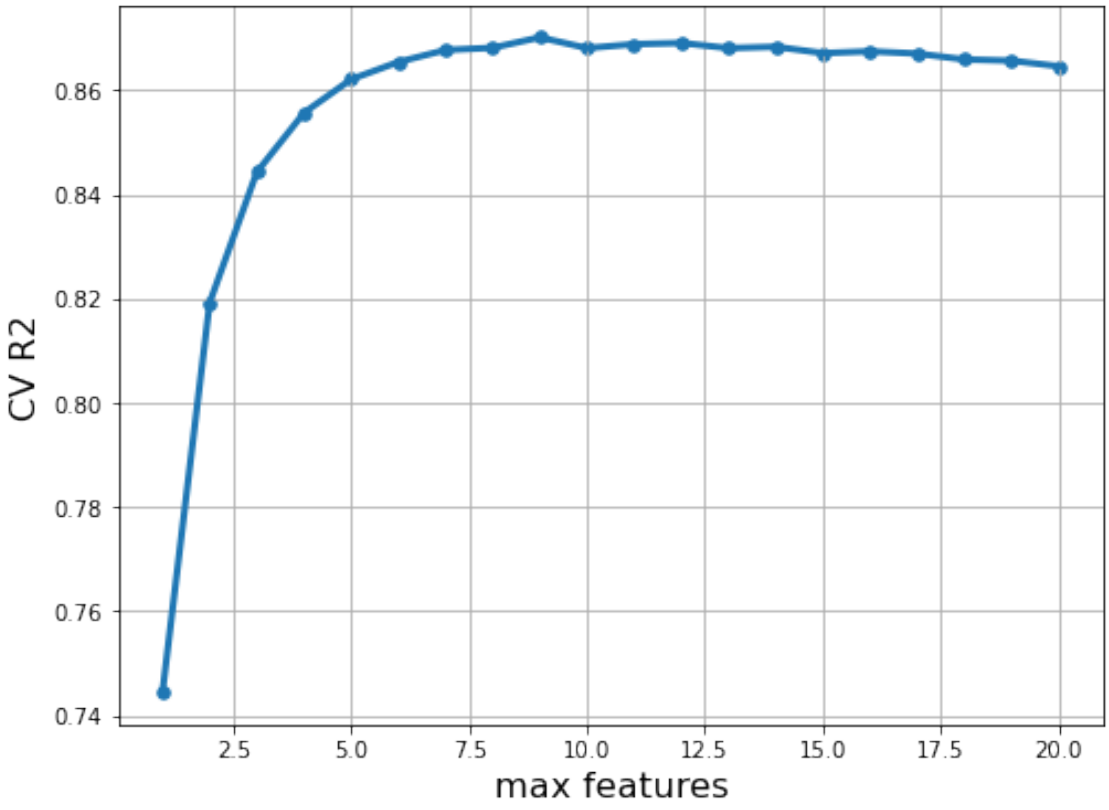
```
Fitting 5 folds for each of 20 candidates, totalling 100 fits
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:  1.3min finished
```

[24]:
```
GridSearchCV(cv=KFold(n_splits=5, random_state=88, shuffle=True),
             estimator=RandomForestRegressor(),
             param_grid={'max_features': array([ 1,  2,  3,  4,  5,  6,  7,  8,
9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20], dtype=int32),
                         'min_samples_leaf': [5], 'n_estimators': [500],
                         'random_state': [88]},
             scoring='r2', verbose=1)
```

Below is the plot examining the R2 for each max_features values between 1 and 20. We will choose the max features parameter based on the highest CV R2.

[25]:
```
max_features = rf_cv.cv_results_['param_max_features'].data
r2_scores_rf = rf_cv.cv_results_['mean_test_score']

plt.figure(figsize=(8, 6))
plt.xlabel('max features', fontsize=16)
plt.ylabel('CV R2', fontsize=16)
plt.scatter(max_features, r2_scores_rf, s=30)
plt.plot(max_features, r2_scores_rf, linewidth=3)
plt.grid(True, which='both')

plt.show()
```



[26]:
```
print('Best max_features', rf_cv.best_params_)

rf = rf_cv.best_estimator_
```

```
Best max_features {'max_features': 9, 'min_samples_leaf': 5, 'n_estimators':
500, 'random_state': 88}
```

[27]:
```
rf_cv.best_score_
```

[27]:
```
0.8701230466287683
```

We observe a relatively high R2 for our optimal max_features value of 0.657. So, we will consider using our random forest model in our blended model.

Below is the importance scores of all features sorted in descending order. We see which features are most important in determining the CAASPP % Passing for each district in our random forest model.

[28]:
```
imp = pd.DataFrame({'Feature' : X_train.columns,
              'Importance score': 100*rf.feature_importances_}).round(1)
imp = imp.sort_values("Importance score", ascending = False)
imp
```

[28]:

	Feature	Importance score
2	Free/Reduced Meals % (District)	46.6
5	Grads Mtg UC/CSU % (District)	22.5
11	Chronic Absenteeism % (District)	10.1
14	Total Gen Fund Revenues Per Student (District)	3.2
1	English Learners % (District)	3.1
6	Grads Mtg UC/CSU # (District)	2.1
7	ACT Test Takers # (District)	1.9
10	Teacher Salary-Avg (District)	1.8
4	Cohort Graduates % (District)	1.7
3	Ethnic Diversity Index (District)	1.5
13	Gen Fund Exp by Object Code - 4000-4999 Books …	0.9
0	Census Day Enrollment (District)	0.9
18	Gen Fund Exp by Activity - 7000-7999 General A…	0.7
12	Current Exp of Educ per ADA (Ed Code 41372) (D…	0.6
9	Avg Years Teaching (District)	0.4
8	Per Pupil Ratio: Teacher (District)	0.4
15	Gen Fund Exp by Activity - 4000-4999 Ancillary…	0.4
16	Gen Fund Exp by Activity - 1000-1999 Instructi…	0.4
19	Gen Fund Exp by Activity - 3000-3999 Pupil Ser…	0.4
17	Gen Fund Exp by Activity - 5000-5999 Community…	0.3

1.5 Blending

[29]:
```
# Gather predictions of the three models on the validation set
X_blend_val = pd.DataFrame(data = {'val_pred_linreg': linreg6.predict(sm.
↪add_constant(X_val[X_train6.columns])),
                    'val_pred_dtr': ccp_dtr.predict(X_val),
                    'val_pred_rf': rf.predict(X_val)})
X_blend_val
```

[29]:

	val_pred_linreg	val_pred_dtr	val_pred_rf
292	39.584727	39.902647	41.797308
254	36.184094	29.704868	35.660884
226	41.698115	34.391769	39.949619
27	24.791700	28.707619	28.803548
100	39.985414	34.391769	38.019899
..	…	…	…
21	33.479326	29.704868	32.367204
184	59.721657	62.549457	59.071124
35	57.619577	62.549457	57.751663
293	19.857898	34.391769	33.443627
305	62.299306	62.549457	62.780062

[155 rows x 3 columns]

[30]:
```
# Blend three models together through ols linear regression
blending_ols = sm.add_constant(X_blend_val)

blending_res = sm.OLS(y_val, blending_ols).fit()

print(blending_res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       CAASPP % Passing   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     560.7
Date:                Wed, 12 May 2021   Prob (F-statistic):           1.31e-81
Time:                        14:59:53   Log-Likelihood:                -454.50
No. Observations:                 155   AIC:                             917.0
Df Residuals:                     151   BIC:                             929.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---
const             -0.5325      1.162     -0.458      0.648      -2.829
1.764
val_pred_linreg    0.0297      0.098      0.302      0.763      -0.165
0.224
val_pred_dtr      -0.1233      0.086     -1.439      0.152      -0.293
0.046
val_pred_rf        1.1452      0.139      8.224      0.000       0.870
1.420
==============================================================================
Omnibus:                        3.563   Durbin-Watson:                   2.073
Prob(Omnibus):                  0.168   Jarque-Bera (JB):                3.339
Skew:                           0.359   Prob(JB):                        0.188
Kurtosis:                       3.030   Cond. No.                         246.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

[31]:
```
# Add the predictions of the blended model to the dataframe
val_pred_blended = blending_res.predict(sm.add_constant(X_blend_val))
X_blend_val['val_pred_blended']= val_pred_blended
```

```
X_blend_val
```

[31]:

	val_pred_linreg	val_pred_dtr	val_pred_rf	val_pred_blended
292	39.584727	39.902647	41.797308	43.587532
254	36.184094	29.704868	35.660884	37.716735
226	41.698115	34.391769	39.949619	42.213810
27	24.791700	28.707619	28.803548	29.648754
100	39.985414	34.391769	38.019899	39.953098
..	…	…	…	…
21	33.479326	29.704868	32.367204	33.864593
184	59.721657	62.549457	59.071124	61.174230
35	57.619577	62.549457	57.751663	59.600831
293	19.857898	34.391769	33.443627	34.115261
305	62.299306	62.549457	62.780062	65.498150

[155 rows x 4 columns]

1.6 Test Performance

[32]:
```
# Get predictions for linear regression, decision tree regression and random␣
↪forest on the test set
X_blend_test = pd.DataFrame(data = {'val_pred_linreg': linreg6.predict(sm.
↪add_constant(X_test[X_train6.columns])),
                    'val_pred_dtr': ccp_dtr.predict(X_test),
                    'val_pred_rf': rf.predict(X_test)})
X_blend_test
```

[32]:

	val_pred_linreg	val_pred_dtr	val_pred_rf
269	77.185455	78.543276	79.206574
134	21.885844	22.468800	27.080265
105	28.244188	33.494405	30.069988
133	51.571911	49.629167	44.190577
8	42.521771	39.902647	39.691409
..	…	…	…
51	34.178287	34.391769	34.469879
129	34.830547	34.391769	32.327586
290	36.849253	34.391769	35.509749
26	68.778618	70.597045	69.878636
268	53.157226	52.038667	54.073523

[153 rows x 3 columns]

[33]:
```
def OSR2(model, X_test, y_test, y_train):
    y_pred = model.predict(X_test)
    SSE = np.sum((y_test - y_pred)**2)
    SST = np.sum((y_test - np.mean(y_train))**2)
```

```
    return (1 - SSE/SST)

def MAE(model, X_test, y_test):
  diff = y_test - model.predict(X_test)
  return np.mean(abs(diff))

def RMSE(model, X_test, y_test):
  diff = y_test - model.predict(X_test)
  return np.sqrt(np.mean(diff**2))
```

[34]:
```
print('Baseline MAE:', np.mean(np.abs(y_test - np.mean(y_train))))
print('Linear Regression MAE:', MAE(linreg6, sm.add_constant(X_test[X_train6.
 ↪columns]), y_test))
print('Regression Tree MAE:', MAE(ccp_dtr, X_test, y_test))
print('Random Forest MAE:', MAE(rf, X_test, y_test))
print('Blended Model MAE:', MAE(blending_res, sm.add_constant(X_blend_test),␣
 ↪y_test), '\n')

print('Baseline RMSE:', np.sqrt(np.mean((y_test - np.mean(y_train))**2)))
print('Linear Regression RMSE:', RMSE(linreg6, sm.add_constant(X_test[X_train6.
 ↪columns]), y_test))
print('Regression Tree RMSE:', RMSE(ccp_dtr, X_test, y_test))
print('Random Forest RMSE:', RMSE(rf, X_test, y_test))
print('Blended Model RMSE:', RMSE(blending_res, sm.add_constant(X_blend_test),␣
 ↪y_test), '\n')

print('Linear Regression OSR2:', OSR2(linreg6, sm.add_constant(X_test[X_train6.
 ↪columns]), y_test, y_train))
print('Regression Tree OSR2:', OSR2(ccp_dtr, X_test, y_test, y_train))
print('Random Forest OSR2:', OSR2(rf, X_test, y_test, y_train))
print('Blended Model OSR2:', OSR2(blending_res, sm.add_constant(X_blend_test),␣
 ↪y_test, y_train), '\n')
```

```
Baseline MAE: 12.027233133296967
Linear Regression MAE: 5.184440220438947
Regression Tree MAE: 5.203101993363607
Random Forest MAE: 4.156863814631019
Blended Model MAE: 3.8434294968747262

Baseline RMSE: 15.405715510462189
Linear Regression RMSE: 6.379729951191242
Regression Tree RMSE: 6.306335897483519
Random Forest RMSE: 5.198373783738898
Blended Model RMSE: 4.796145360379238

Linear Regression OSR2: 0.8285091929626793
Regression Tree OSR2: 0.8324322451844091
```

```
Random Forest OSR2: 0.8861399788446034
Blended Model OSR2: 0.9030783214698206
```

Our blended model produces the lowest error and highest OSR2 of 0.89

1.7 Part B: UC Admissions

1.8 Data Processing

We will data process three datasets: school level, admissions and district level. We will eventually merge them on District Name and School Name. Therefore, we have to make sure that the values in those columns are in the same format and that the merging columns have the same name.

1.8.1 School level data

[35]:
```
# Load the school datasets
s1 = pd.read_csv("school/schools_1.csv")
s2 = pd.read_csv('school/schools_2.csv')
s3 = pd.read_csv('school/schools_3.csv')
s4 = pd.read_csv("school/schools_4.csv")
s5 = pd.read_csv('school/schools_5.csv')
s6 = pd.read_csv('school/schools_6.csv')
s7 = pd.read_csv("school/schools_7.csv")
s8 = pd.read_csv('school/schools_8.csv')
s9 = pd.read_csv('school/schools_9.csv')
s10 = pd.read_csv('school/schools_10.csv')
```

[36]:
```
# Concatenate the dataframes along axis 0 to combine school data from all years
all_schools = pd.concat([s1, s2, s3, s4, s5, s6, s7, s8, s9, s10], axis = 0)
```

[37]:
```
# Only keep high school data
all_hs = all_schools[all_schools['School Type (School)'] == 'High School']
all_hs
```

[37]:
```
                                          School Name  \
0              Abraham Lincoln High (San Jose Unified)
1     Abraham Lincoln Senior High (Los Angeles Unified)
2                     Abraxis Charter (Santa Rosa High)
3     Academies of Education and Empowerment at Cars…
4     Academies of the Antelope Valley (Antelope Val…
…                                                   …
1346  YouthBuild Charter School of California (Inyo …
1347                     Yreka High (Yreka Union High)
1348                 Yuba City High (Yuba City Unified)
1349     Yucaipa High (Yucaipa-Calimesa Joint Unified)
```

```
1350                Yucca Valley High (Morongo Unified)

               District Name (School) School Type (School)  \
0                    San Jose Unified          High School
1                 Los Angeles Unified          High School
2                     Santa Rosa High          High School
3                 Los Angeles Unified          High School
4           Antelope Valley Union High          High School
…                                   …                   …
1346  Inyo County Office of Education          High School
1347                  Yreka Union High          High School
1348                 Yuba City Unified          High School
1349   Yucaipa-Calimesa Joint Unified          High School
1350                   Morongo Unified          High School

     County Name (School)
0              Santa Clara
1              Los Angeles
2                   Sonoma
3              Los Angeles
4              Los Angeles
…                      …
1346                 Inyo
1347              Siskiyou
1348                Sutter
1349       San Bernardino
1350       San Bernardino

[4032 rows x 4 columns]
```

[38]:
```
# Remove parentheses and district name from the school name, important for␣
 ↪future merging on column
import re

all_hs['School Name'] = all_hs['School Name'].apply(lambda x: re.sub(' \(.*\)',␣
 ↪'', str(x)))
all_hs
```

```
<ipython-input-38-b306db095972>:4: 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
  all_hs['School Name'] = all_hs['School Name'].apply(lambda x: re.sub('
\(.*\)', '', str(x)))
```

[38]:
```
                                          School Name  \
0                                Abraham Lincoln High
1                            Abraham Lincoln Senior High
2                                      Abraxis Charter
3     Academies of Education and Empowerment at Cars…
4                        Academies of the Antelope Valley
…                                                    …
1346            YouthBuild Charter School of California
1347                                        Yreka High
1348                                     Yuba City High
1349                                       Yucaipa High
1350                                  Yucca Valley High

               District Name (School) School Type (School)  \
0                    San Jose Unified          High School
1                  Los Angeles Unified          High School
2                      Santa Rosa High          High School
3                  Los Angeles Unified          High School
4           Antelope Valley Union High          High School
…                                  …                    …
1346  Inyo County Office of Education          High School
1347                 Yreka Union High          High School
1348                Yuba City Unified          High School
1349   Yucaipa-Calimesa Joint Unified          High School
1350                  Morongo Unified          High School

     County Name (School)
0              Santa Clara
1              Los Angeles
2                   Sonoma
3              Los Angeles
4              Los Angeles
…                        …
1346                 Inyo
1347              Siskiyou
1348                Sutter
1349       San Bernardino
1350       San Bernardino

[4032 rows x 4 columns]
```

[39]:
```
# Fix Columns
all_hs.rename({'District Name (School)': 'District Name'}, inplace=True, axis=1)
all_hs.rename({'County Name (School)': 'County'}, inplace=True, axis=1)

all_hs = all_hs.drop(['School Type (School)'], axis=1)
all_hs
```

```
/Users/tarsus/opt/anaconda3/lib/python3.8/site-
packages/pandas/core/frame.py:4296: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return super().rename(
```

[39]:

	School Name	District Name	County
0	Abraham Lincoln High	San Jose Unified	Santa Clara
1	Abraham Lincoln Senior High	Los Angeles Unified	Los Angeles
2	Abraxis Charter	Santa Rosa High	Sonoma
3	Academies of Education and Empowerment at Cars…	Los Angeles Unified	Los Angeles
4	Academies of the Antelope Valley	Antelope Valley Union High	Los Angeles
…	…	…	…
1346	YouthBuild Charter School of California	Inyo County Office of Education	Inyo
1347	Yreka High	Yreka Union High	Siskiyou
1348	Yuba City High	Yuba City Unified	Sutter
1349	Yucaipa High	Yucaipa-Calimesa Joint Unified	San Bernardino
1350	Yucca Valley High	Morongo Unified	San Bernardino

[4032 rows x 3 columns]

1.8.2 Admissions data

[40]:
```
# Import college admissions data for each year
adm_1 = pd.read_csv('admission/adm_1.csv')
adm_2 = pd.read_csv('admission/adm_2.csv')
adm_3 = pd.read_csv('admission/adm_3.csv')
adm_4 = pd.read_csv('admission/adm_4.csv')
adm_5 = pd.read_csv('admission/adm_5.csv')
adm_6 = pd.read_csv('admission/adm_6.csv')
adm_7 = pd.read_csv('admission/adm_7.csv')
```

```
adm_8 = pd.read_csv('admission/adm_8.csv')
adm_9 = pd.read_csv('admission/adm_9.csv')
```

[41]:
```
# Add a new column for the year
adm_1 = adm_1.assign(Year = '2010-2011')
adm_2 = adm_2.assign(Year = '2011-2012')
adm_3 = adm_3.assign(Year = '2012-2013')
adm_4 = adm_4.assign(Year = '2013-2014')
adm_5 = adm_5.assign(Year = '2014-2015')
adm_6 = adm_6.assign(Year = '2015-2016')
adm_7 = adm_7.assign(Year = '2016-2017')
adm_8 = adm_8.assign(Year = '2017-2018')
adm_9 = adm_9.assign(Year = '2018-2019')
```

[42]:
```
# Combine all the admission data into one dataframe
adm_data = pd.concat([adm_1, adm_2, adm_3, adm_4, adm_5, adm_6, adm_7, adm_8,␣
↪adm_9], axis=0)
adm_data.head()
```

[42]:

	Calculation1	City	County/State/ Territory	Count	Gender	School	Pivot Field Values	Year
0	A B MILLER HIGH SCHOOL50944	Fontana	San Bernardino	Enr	Male	A B MILLER HIGH SCHOOL	10	2010-2011
1	A B MILLER HIGH SCHOOL50944	Fontana	San Bernardino	Adm	Male	A B MILLER HIGH SCHOOL	14	2010-2011
2	A B MILLER HIGH SCHOOL50944	Fontana	San Bernardino	App	Male	A B MILLER HIGH SCHOOL	21	2010-2011
3	A B MILLER HIGH SCHOOL50944	Fontana	San Bernardino	Enr	Female	A B MILLER HIGH SCHOOL	21	2010-2011
4	A B MILLER HIGH SCHOOL50944	Fontana	San Bernardino	Adm	Female	A B MILLER HIGH SCHOOL	30	2010-2011

[43]:
```
# Drop unwanted columns
adm_data = adm_data.drop('Calculation1', axis=1)
```

[44]:
```
# Rename the school column to match column name of the school level dataset
adm_data.rename({'School':'School Name'}, inplace=True, axis=1)

# Rename columns for readability
adm_data = adm_data.rename({'Pivot Field Values': 'Student Count'}, axis = 1)
adm_data = adm_data.rename({'County/State/ Territory': 'County'}, axis = 1)
```

[45]:
```
# Fixing up school names to resemble those of school level dataset
adm_data['School Name'] = adm_data['School Name'].apply(lambda x: ' '.
↪join([word.capitalize() for word in str(x).split()]))
```

```
adm_data['School Name'] = adm_data['School Name'].apply(lambda x: x.rsplit(' ',␣
↪1)[0])
adm_data
```

[45]:

	City	County	Count	Gender	School Name	Student Count	Year
0	Fontana	San Bernardino	Enr	Male	A B Miller High	10	2010-2011
1	Fontana	San Bernardino	Adm	Male	A B Miller High	14	2010-2011
2	Fontana	San Bernardino	App	Male	A B Miller High	21	2010-2011
3	Fontana	San Bernardino	Enr	Female	A B Miller High	21	2010-2011
4	Fontana	San Bernardino	Adm	Female	A B Miller High	30	2010-2011
…	…	…	…	…	…	…	…
9084	Yucca Valley	San Bernardino	Adm	Female	Yucca Valley High	13	2018-2019
9085	Yucca Valley	San Bernardino	App	Female	Yucca Valley High	13	2018-2019
9086	Yucca Valley	San Bernardino	Enr	All	Yucca Valley High	12	2018-2019
9087	Yucca Valley	San Bernardino	Adm	All	Yucca Valley High	21	2018-2019
9088	Yucca Valley	San Bernardino	App	All	Yucca Valley High	24	2018-2019

[82469 rows x 7 columns]

[46]:
```
# Remove male and female rows since we are only concerned about the total␣
↪admissions
adm_data = adm_data[adm_data["Gender"] == "All"]
adm_data = adm_data.drop('Gender', axis=1)
adm_data
```

[46]:

	City	County	Count	School Name	Student Count	Year
6	Fontana	San Bernardino	Enr	A B Miller High	31	2010-2011
7	Fontana	San Bernardino	Adm	A B Miller High	44	2010-2011
8	Fontana	San Bernardino	App	A B Miller High	61	2010-2011
15	Los Angeles	Los Angeles	Enr	Abraham Lincoln High	55	2010-2011
16	Los Angeles	Los Angeles	Adm	Abraham Lincoln High	74	2010-2011
…	…	…	…	…	…	…
9078	Yucaipa	San Bernardino	Adm	Yucaipa Senior High	51	2018-2019
9079	Yucaipa	San Bernardino	App	Yucaipa Senior High	86	2018-2019

```
9086  Yucca Valley  San Bernardino   Enr     Yucca Valley High            12
9087  Yucca Valley  San Bernardino   Adm     Yucca Valley High            21
9088  Yucca Valley  San Bernardino   App     Yucca Valley High            24

          Year
6     2010-2011
7     2010-2011
8     2010-2011
15    2010-2011
16    2010-2011
…         …
9078  2018-2019
9079  2018-2019
9086  2018-2019
9087  2018-2019
9088  2018-2019

[28829 rows x 6 columns]
```

1.8.3 District level data

[47]:
```
# Using the district level data imported in Part A
sd_all
```

[47]:
```
                                    District Name County Name (District)  \
0                    Calistoga Joint Unified (Napa)                  Napa
1                       Death Valley Unified (Inyo)                  Inyo
2                    Golden Valley Unified (Madera)                Madera
3                        Warner Unified (San Diego)             San Diego
4            Acton-Agua Dulce Unified (Los Angeles)           Los Angeles
..                                              …                     …
340                  Woodland Joint Unified (Yolo)                  Yolo
341                      Yosemite Unified (Madera)                Madera
342                     Yuba City Unified (Sutter)                Sutter
343  Yucaipa-Calimesa Joint Unified (San Bernardino)        San Bernardino
344                                             NaN                   NaN

    District Type (District) Census Day Enrollment (District)  \
0    Unified School District                             858
1    Unified School District                              51
2    Unified School District                            1925
3    Unified School District                             286
4    Unified School District                            1696
..                       …                               …
340  Unified School District                           10031
341  Unified School District                            2060
```

```
342  Unified School District                            13111
343  Unified School District                             9982
344                      NaN                              NaN

    English Learners % (District) Free/Reduced Meals % (District)  \
0                   Not Certified                             78.4
1                   Not Certified                             84.3
2                             6.2                             46.1
3                             8.4                             75.2
4                   Not Certified                             30.3
..                              …                            …
340                          23.3                             60.3
341                             2                             48.2
342                          22.1                             74.2
343                           7.6                             50.8
344                           NaN                              NaN

    Ethnic Diversity Index (District) Cohort Graduates % (District)  \
0                                   5                          82.19
1                                  28                            100
2                                  36                          91.03
3                                  56                             90
4                                  34                             90
..                                …                            …
340                                32                           92.3
341                                36                           83.8
342                                53                           85.8
343                                39                           90.5
344                               NaN                            NaN

    Grads Mtg UC/CSU % (District) Grads Mtg UC/CSU # (District)  …  \
0                              100                            63  …
1                         redacted                      redacted  …
2                              100                           143  …
3                              100                            31  …
4                             99.3                           138  …
..                               …                          …  …
340                           47.7                           327  …
341                           49.3                            74  …
342                           37.7                           314  …
343                           46.2                           306  …
344                           NaN                            NaN  …

     CAASPP-Math Standard Exceeded or Met (Levels 3 and 4) (District)  \
0                                                   NaN
1                                                   NaN
2                                                   NaN
```

```
3                                                    NaN
4                                                    NaN
..                                                   …
340                                                27.89
341                                                34.89
342                                                29.59
343                                                36.33
344                                                  NaN

     Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                  12806.0
1                                                  25438.0
2                                                   7633.0
3                                                  13610.0
4                                                   7795.0
..                                                     …
340                                                12175.0
341                                                12771.0
342                                                12987.0
343                                                11629.0
344                                                    NaN

     Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                   604.0
1                                                  1733.0
2                                                   256.0
3                                                  1019.0
4                                                   431.0
..                                                     …
340                                                 422.0
341                                                 587.0
342                                                 793.0
343                                                 346.0
344                                                   NaN

     Total Gen Fund Revenues Per Student (District)  \
0                                              13952.0
1                                              27294.0
2                                               8320.0
3                                              15644.0
4                                               8361.0
..                                                 …
340                                            13128.0
341                                            14195.0
342                                            13557.0
343                                            12174.0
```

```
344                                                NaN

     Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                 239.0
1                                                 312.0
2                                                 107.0
3                                                  56.0
4                                                  77.0
..                                                  …
340                                                92.0
341                                                83.0
342                                               133.0
343                                               149.0
344                                                 NaN

    Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)  \
0                                                8543.0
1                                               12543.0
2                                                4329.0
3                                                7675.0
4                                                4778.0
..                                                  …
340                                              7653.0
341                                              6795.0
342                                              8195.0
343                                              7204.0
344                                                 NaN

    Gen Fund Exp by Activity - 5000-5999 Community Services Per Student #
(District)  \
0                                                   31
1                                                    0
2                                                    0
3                                                    0
4                                                    0
..                                                   …
340                                                 19
341                                                 38
342                                                  0
343                                                  0
344                                                NaN

     Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                1250.0
1                                                4458.0
```

```
2                                                    781.0
3                                                   1998.0
4                                                    687.0
..                                                     …
340                                                  619.0
341                                                 1157.0
342                                                  679.0
343                                                  661.0
344                                                    NaN

     Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                    781.0
1                                                   4449.0
2                                                    533.0
3                                                   2465.0
4                                                    909.0
..                                                     …
340                                                 1296.0
341                                                 1391.0
342                                                 1114.0
343                                                 1460.0
344                                                    NaN

          Year
0    2010-2011
1    2010-2011
2    2010-2011
3    2010-2011
4    2010-2011
..         …
340  2018-2019
341  2018-2019
342  2018-2019
343  2018-2019
344  2018-2019

[3076 rows x 26 columns]
```

[48]:
```
# Remove unwanted characters in District Name
sd_all['District Name'] = sd_all['District Name'].apply(lambda x: re.sub(' \(.
 ↪*\)', '', str(x)))
sd_all.head()
```

[48]:
```
              District Name County Name (District) District Type (District)  \
0   Calistoga Joint Unified                   Napa  Unified School District
1      Death Valley Unified                   Inyo  Unified School District
```

```
2     Golden Valley Unified                Madera  Unified School District
3            Warner Unified             San Diego  Unified School District
4  Acton-Agua Dulce Unified           Los Angeles  Unified School District

  Census Day Enrollment (District) English Learners % (District)  \
0                              858                 Not Certified
1                               51                 Not Certified
2                             1925                           6.2
3                              286                           8.4
4                             1696                 Not Certified

  Free/Reduced Meals % (District) Ethnic Diversity Index (District)  \
0                            78.4                                 5
1                            84.3                                28
2                            46.1                                36
3                            75.2                                56
4                            30.3                                34

  Cohort Graduates % (District) Grads Mtg UC/CSU % (District)  \
0                         82.19                          100
1                           100                     redacted
2                         91.03                          100
3                            90                          100
4                            90                         99.3

  Grads Mtg UC/CSU # (District)  …  \
0                            63  …
1                      redacted  …
2                           143  …
3                            31  …
4                           138  …

   CAASPP-Math Standard Exceeded or Met (Levels 3 and 4) (District)  \
0                                                 NaN
1                                                 NaN
2                                                 NaN
3                                                 NaN
4                                                 NaN

   Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                            12806.0
1                                            25438.0
2                                             7633.0
3                                            13610.0
4                                             7795.0

   Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
```

```
(District)  \
0                                                604.0
1                                               1733.0
2                                                256.0
3                                               1019.0
4                                                431.0

   Total Gen Fund Revenues Per Student (District)  \
0                                              13952.0
1                                              27294.0
2                                               8320.0
3                                              15644.0
4                                               8361.0

   Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                239.0
1                                                312.0
2                                                107.0
3                                                 56.0
4                                                 77.0

  Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)  \
0                                               8543.0
1                                              12543.0
2                                               4329.0
3                                               7675.0
4                                               4778.0

  Gen Fund Exp by Activity - 5000-5999 Community Services Per Student #
(District)  \
0                                                  31
1                                                   0
2                                                   0
3                                                   0
4                                                   0

   Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                               1250.0
1                                               4458.0
2                                                781.0
3                                               1998.0
4                                                687.0

   Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
\
```

```
0                                                  781.0
1                                                 4449.0
2                                                  533.0
3                                                 2465.0
4                                                  909.0

        Year
0  2010-2011
1  2010-2011
2  2010-2011
3  2010-2011
4  2010-2011

[5 rows x 26 columns]
```

1.8.4 Merging

Having modified each dataset, we will now merge them all together. We want to end up with a final dataset containing UC admission rates at the district level along with the same features used in Part A.

However, our UC admissions dataset does not have District Name to merge with the district level data. Therefore, we first merge with the school level data to get District Name and merge again with the district level data.

[49]:
```
# Merge school level and admissions data to get District Name and college␣
 ↪admissions statistics in one table
hs_adm = adm_data.merge(all_hs, on=['School Name', 'County'])
hs_adm = hs_adm.drop_duplicates()
hs_adm.sort_values(by='School Name', inplace=True)

# Check that each school has enrolled, applied and admitted statistics for each␣
 ↪school year
hs_adm.head(28) # 3 x 9 = 27, 28th should be next school
```

[49]:
```
           City        County Count           School Name  Student Count  \
0     San Jose   Santa Clara   Enr  Abraham Lincoln High             16
78    San Jose   Santa Clara   App  Abraham Lincoln High            105
75    San Jose   Santa Clara   Adm  Abraham Lincoln High             56
72    San Jose   Santa Clara   Enr  Abraham Lincoln High             23
69    San Jose   Santa Clara   App  Abraham Lincoln High            100
66    San Jose   Santa Clara   Adm  Abraham Lincoln High             58
63    San Jose   Santa Clara   Enr  Abraham Lincoln High             26
60    San Jose   Santa Clara   App  Abraham Lincoln High            123
54    San Jose   Santa Clara   Enr  Abraham Lincoln High             33
51    San Jose   Santa Clara   App  Abraham Lincoln High             78
48    San Jose   Santa Clara   Adm  Abraham Lincoln High             45
```

45	San Jose	Santa Clara	Enr	Abraham Lincoln High	19
42	San Jose	Santa Clara	App	Abraham Lincoln High	95
57	San Jose	Santa Clara	Adm	Abraham Lincoln High	72
36	San Jose	Santa Clara	Enr	Abraham Lincoln High	37
39	San Jose	Santa Clara	Adm	Abraham Lincoln High	64
6	San Jose	Santa Clara	App	Abraham Lincoln High	60
9	San Jose	Santa Clara	Enr	Abraham Lincoln High	30
12	San Jose	Santa Clara	Adm	Abraham Lincoln High	47
15	San Jose	Santa Clara	App	Abraham Lincoln High	63
18	San Jose	Santa Clara	Enr	Abraham Lincoln High	23
3	San Jose	Santa Clara	Adm	Abraham Lincoln High	46
24	San Jose	Santa Clara	App	Abraham Lincoln High	70
27	San Jose	Santa Clara	Enr	Abraham Lincoln High	25
30	San Jose	Santa Clara	Adm	Abraham Lincoln High	48
33	San Jose	Santa Clara	App	Abraham Lincoln High	84
21	San Jose	Santa Clara	Adm	Abraham Lincoln High	47
123	Lafayette	Contra Costa	App	Acalanes High	158

	Year	District Name
0	2010-2011	San Jose Unified
78	2018-2019	San Jose Unified
75	2018-2019	San Jose Unified
72	2018-2019	San Jose Unified
69	2017-2018	San Jose Unified
66	2017-2018	San Jose Unified
63	2017-2018	San Jose Unified
60	2016-2017	San Jose Unified
54	2016-2017	San Jose Unified
51	2015-2016	San Jose Unified
48	2015-2016	San Jose Unified
45	2015-2016	San Jose Unified
42	2014-2015	San Jose Unified
57	2016-2017	San Jose Unified
36	2014-2015	San Jose Unified
39	2014-2015	San Jose Unified
6	2010-2011	San Jose Unified
9	2011-2012	San Jose Unified
12	2011-2012	San Jose Unified
15	2011-2012	San Jose Unified
18	2012-2013	San Jose Unified
3	2010-2011	San Jose Unified
24	2012-2013	San Jose Unified
27	2013-2014	San Jose Unified
30	2013-2014	San Jose Unified
33	2013-2014	San Jose Unified
21	2012-2013	San Jose Unified
123	2014-2015	Acalanes Union High

[50]:
```
# Now we can merge on district name to get college admissions and district level data in the same table
sd_acceptance = hs_adm.merge(sd_all, on=['District Name', 'Year'], how='inner')
# Sort by school name
sd_acceptance.sort_values(by=['School Name', 'Year'], inplace=True)
sd_acceptance.head(28)
```

[50]:

	City	County	Count	School Name	Student Count	Year	District Name	County Name (District)
0	San Jose	Santa Clara	Enr	Abraham Lincoln High	16	2010-2011	San Jose Unified	Santa Clara
1	San Jose	Santa Clara	App	Abraham Lincoln High	60	2010-2011	San Jose Unified	Santa Clara
2	San Jose	Santa Clara	Adm	Abraham Lincoln High	46	2010-2011	San Jose Unified	Santa Clara
90	San Jose	Santa Clara	Enr	Abraham Lincoln High	30	2011-2012	San Jose Unified	Santa Clara
91	San Jose	Santa Clara	Adm	Abraham Lincoln High	47	2011-2012	San Jose Unified	Santa Clara
92	San Jose	Santa Clara	App	Abraham Lincoln High	63	2011-2012	San Jose Unified	Santa Clara
105	San Jose	Santa Clara	Enr	Abraham Lincoln High	23	2012-2013	San Jose Unified	Santa Clara
106	San Jose	Santa Clara	App	Abraham Lincoln High	70	2012-2013	San Jose Unified	Santa Clara
107	San Jose	Santa Clara	Adm	Abraham Lincoln High	47	2012-2013	San Jose Unified	Santa Clara
120	San Jose	Santa Clara	Enr	Abraham Lincoln High	25	2013-2014	San Jose Unified	Santa Clara
121	San Jose	Santa Clara	Adm	Abraham Lincoln High	48	2013-2014	San Jose Unified	Santa Clara
122	San Jose	Santa Clara	App	Abraham Lincoln High	84	2013-2014	San Jose Unified	Santa Clara
75	San Jose	Santa Clara	App	Abraham Lincoln High	95	2014-2015	San Jose Unified	Santa Clara
76	San Jose	Santa Clara	Enr	Abraham Lincoln High	37	2014-2015	San Jose Unified	Santa Clara
77	San Jose	Santa Clara	Adm	Abraham Lincoln High	64	2014-2015	San Jose Unified	Santa Clara
60	San Jose	Santa Clara	App	Abraham Lincoln High	78	2015-2016	San Jose Unified	Santa Clara
61	San Jose	Santa Clara	Adm	Abraham Lincoln High	45	2015-2016	San Jose Unified	Santa Clara
62	San Jose	Santa Clara	Enr	Abraham Lincoln High	19	2015-2016	San Jose Unified	Santa Clara
45	San Jose	Santa Clara	App	Abraham Lincoln High	123	2016-2017	San Jose Unified	Santa Clara
46	San Jose	Santa Clara	Enr	Abraham Lincoln High	33	2016-2017	San Jose Unified	Santa Clara
47	San Jose	Santa Clara	Adm	Abraham Lincoln High	72	2016-2017	San Jose Unified	Santa Clara
30	San Jose	Santa Clara	App	Abraham Lincoln High	100	2017-2018	San Jose Unified	Santa Clara
31	San Jose	Santa Clara	Adm	Abraham Lincoln High	58	2017-2018	San Jose Unified	Santa Clara
32	San Jose	Santa Clara	Enr	Abraham Lincoln High	26	2017-2018	San Jose Unified	Santa Clara
15	San Jose	Santa Clara	App	Abraham Lincoln High	105	2018-2019	San Jose Unified	Santa Clara
16	San Jose	Santa Clara	Adm	Abraham Lincoln High	56	2018-2019	San Jose Unified	Santa Clara
17	San Jose	Santa Clara	Enr	Abraham Lincoln High	23	2018-2019	San Jose Unified	Santa Clara
171	Agoura Hills	Los Angeles	Adm	Agoura High	98			

```
120  2013-2014      San Jose Unified            Santa Clara
121  2013-2014      San Jose Unified            Santa Clara
122  2013-2014      San Jose Unified            Santa Clara
75   2014-2015      San Jose Unified            Santa Clara
76   2014-2015      San Jose Unified            Santa Clara
77   2014-2015      San Jose Unified            Santa Clara
60   2015-2016      San Jose Unified            Santa Clara
61   2015-2016      San Jose Unified            Santa Clara
62   2015-2016      San Jose Unified            Santa Clara
45   2016-2017      San Jose Unified            Santa Clara
46   2016-2017      San Jose Unified            Santa Clara
47   2016-2017      San Jose Unified            Santa Clara
30   2017-2018      San Jose Unified            Santa Clara
31   2017-2018      San Jose Unified            Santa Clara
32   2017-2018      San Jose Unified            Santa Clara
15   2018-2019      San Jose Unified            Santa Clara
16   2018-2019      San Jose Unified            Santa Clara
17   2018-2019      San Jose Unified            Santa Clara
171  2010-2011  Las Virgenes Unified            Los Angeles

    District Type (District) Census Day Enrollment (District)  …  \
0    Unified School District                            33018  …
1    Unified School District                            33018  …
2    Unified School District                            33018  …
90   Unified School District                            33306  …
91   Unified School District                            33306  …
92   Unified School District                            33306  …
105  Unified School District                            33184  …
106  Unified School District                            33184  …
107  Unified School District                            33184  …
120  Unified School District                            33152  …
121  Unified School District                            33152  …
122  Unified School District                            33152  …
75   Unified School District                            32938  …
76   Unified School District                            32938  …
77   Unified School District                            32938  …
60   Unified School District                            32454  …
61   Unified School District                            32454  …
62   Unified School District                            32454  …
45   Unified School District                            32004  …
46   Unified School District                            32004  …
47   Unified School District                            32004  …
30   Unified School District                            31713  …
31   Unified School District                            31713  …
32   Unified School District                            31713  …
15   Unified School District                            31114  …
16   Unified School District                            31114  …
```

```
17   Unified School District                          31114  …
171  Unified School District                          11393  …

    CAASPP-ELA Standard Exceeded or Met (Levels 3 and 4) (District)  \
0                                                          NaN
1                                                          NaN
2                                                          NaN
90                                                         NaN
91                                                         NaN
92                                                         NaN
105                                                        NaN
106                                                        NaN
107                                                        NaN
120                                                        NaN
121                                                        NaN
122                                                        NaN
75                                                          51
76                                                          51
77                                                          51
60                                                          54
61                                                          54
62                                                          54
45                                                       54.19
46                                                       54.19
47                                                       54.19
30                                                       55.54
31                                                       55.54
32                                                       55.54
15                                                       54.91
16                                                       54.91
17                                                       54.91
171                                                        NaN

    CAASPP-Math Standard Exceeded or Met (Levels 3 and 4) (District)  \
0                                                          NaN
1                                                          NaN
2                                                          NaN
90                                                         NaN
91                                                         NaN
92                                                         NaN
105                                                        NaN
106                                                        NaN
107                                                        NaN
120                                                        NaN
121                                                        NaN
122                                                        NaN
75                                                          40
```

76	40
77	40
60	43
61	43
62	43
45	43.62
46	43.62
47	43.62
30	44.23
31	44.23
32	44.23
15	43.56
16	43.56
17	43.56
171	NaN

	Current Exp of Educ per ADA (Ed Code 41372) (District) \
0	8963.0
1	8963.0
2	8963.0
90	8985.0
91	8985.0
92	8985.0
105	9040.0
106	9040.0
107	9040.0
120	9400.0
121	9400.0
122	9400.0
75	10076.0
76	10076.0
77	10076.0
60	11075.0
61	11075.0
62	11075.0
45	12097.0
46	12097.0
47	12097.0
30	11766.0
31	11766.0
32	11766.0
15	12363.0
16	12363.0
17	12363.0
171	7966.0

	Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student

(District) \	
0	365.0
1	365.0
2	365.0
90	302.0
91	302.0
92	302.0
105	325.0
106	325.0
107	325.0
120	401.0
121	401.0
122	401.0
75	389.0
76	389.0
77	389.0
60	396.0
61	396.0
62	396.0
45	413.0
46	413.0
47	413.0
30	406.0
31	406.0
32	406.0
15	410.0
16	410.0
17	410.0
171	169.0

	Total Gen Fund Revenues Per Student (District) \
0	9112.0
1	9112.0
2	9112.0
90	9047.0
91	9047.0
92	9047.0
105	9038.0
106	9038.0
107	9038.0
120	9500.0
121	9500.0
122	9500.0
75	10416.0
76	10416.0
77	10416.0
60	12117.0

61	12117.0
62	12117.0
45	11674.0
46	11674.0
47	11674.0
30	12108.0
31	12108.0
32	12108.0
15	13282.0
16	13282.0
17	13282.0
171	8865.0

	Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District) \
0	135.0
1	135.0
2	135.0
90	128.0
91	128.0
92	128.0
105	130.0
106	130.0
107	130.0
120	129.0
121	129.0
122	129.0
75	135.0
76	135.0
77	135.0
60	137.0
61	137.0
62	137.0
45	152.0
46	152.0
47	152.0
30	154.0
31	154.0
32	154.0
15	166.0
16	166.0
17	166.0
171	1.0

	Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District) \
0	5235.0

1	5235.0
2	5235.0
90	5366.0
91	5366.0
92	5366.0
105	5373.0
106	5373.0
107	5373.0
120	5516.0
121	5516.0
122	5516.0
75	5912.0
76	5912.0
77	5912.0
60	6492.0
61	6492.0
62	6492.0
45	7140.0
46	7140.0
47	7140.0
30	6898.0
31	6898.0
32	6898.0
15	7198.0
16	7198.0
17	7198.0
171	5563.0

Gen Fund Exp by Activity - 5000-5999 Community Services Per Student # (District) \

0	0
1	0
2	0
90	0
91	0
92	0
105	0
106	0
107	0
120	0
121	0
122	0
75	0
76	0
77	0
60	0
61	0

62	0
45	0
46	0
47	0
30	0
31	0
32	0
15	0
16	0
17	0
171	0

Gen Fund Exp by Activity - 7000-7999 General Administration Per Student # (District) \

0	403.0
1	403.0
2	403.0
90	427.0
91	427.0
92	427.0
105	425.0
106	425.0
107	425.0
120	518.0
121	518.0
122	518.0
75	517.0
76	517.0
77	517.0
60	570.0
61	570.0
62	570.0
45	629.0
46	629.0
47	629.0
30	594.0
31	594.0
32	594.0
15	671.0
16	671.0
17	671.0
171	334.0

Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)

0	727.0
1	727.0

```
2                                                    727.0
90                                                   773.0
91                                                   773.0
92                                                   773.0
105                                                  732.0
106                                                  732.0
107                                                  732.0
120                                                  836.0
121                                                  836.0
122                                                  836.0
75                                                   861.0
76                                                   861.0
77                                                   861.0
60                                                  1197.0
61                                                  1197.0
62                                                  1197.0
45                                                  1294.0
46                                                  1294.0
47                                                  1294.0
30                                                  1156.0
31                                                  1156.0
32                                                  1156.0
15                                                  1217.0
16                                                  1217.0
17                                                  1217.0
171                                                  610.0

[28 rows x 31 columns]
```

Now we want to calculate our target variable, which is UC admissions rate for each district, by dividing the Admitted count by the Applied count. We accomplished this by pivotting the table.

[51]:
```
# Aggregate on school district, get all admitted, enrolled and applied data per␣
↪year per school district
sd_acceptance = sd_acceptance.groupby(['District Name', 'Count', 'Year']).sum().
↪reset_index()
sd_acceptance
```

[51]:
```
           District Name Count       Year  Student Count  \
0           ABC Unified   Adm  2010-2011            205
1           ABC Unified   Adm  2011-2012            207
2           ABC Unified   Adm  2012-2013            183
3           ABC Unified   Adm  2013-2014            200
4           ABC Unified   Adm  2014-2015            182
…                   …   …        …              …
6601  Yuba City Unified   Enr  2014-2015             43
6602  Yuba City Unified   Enr  2015-2016             31
```

```
6603  Yuba City Unified   Enr  2016-2017             40
6604  Yuba City Unified   Enr  2017-2018             27
6605  Yuba City Unified   Enr  2018-2019             41

      ACT Test Takers # (District)  Per Pupil Ratio: Teacher (District)  \
0                              0.0                                  49.0
1                              0.0                                  49.6
2                              0.0                                  48.2
3                              0.0                                  49.2
4                              0.0                                  48.0
…                              …                                    …
6601                           0.0                                  40.0
6602                         402.0                                  39.6
6603                         332.0                                  39.4
6604                         296.0                                  36.0
6605                         232.0                                  38.6

      Avg Years Teaching (District)  Teacher Salary-Avg (District)  \
0                              30.0                         143614.0
1                              30.0                         143774.0
2                              28.0                         143186.0
3                              28.0                         152394.0
4                              28.0                         156382.0
…                              …                            …
6601                           24.0                         128330.0
6602                           24.0                         132862.0
6603                           24.0                         137350.0
6604                           24.0                         146336.0
6605                           24.0                         151670.0

      Chronic Absenteeism % (District)  \
0                                  0.0
1                                  0.0
2                                  0.0
3                                  0.0
4                                  0.0
…                                  …
6601                               0.0
6602                               0.0
6603                              36.4
6604                              24.6
6605                              24.4

      Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                    16108.0
1                                                    16216.0
2                                                    16010.0
```

```
3                                                    16818.0
4                                                    17850.0
…                                                      …
6601                                                 18236.0
6602                                                 19472.0
6603                                                 21636.0
6604                                                 23652.0
6605                                                 25974.0

      Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                     688.0
1                                                     756.0
2                                                     598.0
3                                                     768.0
4                                                     990.0
…                                                       …
6601                                                 1224.0
6602                                                 1322.0
6603                                                 1288.0
6604                                                 1622.0
6605                                                 1586.0

      Total Gen Fund Revenues Per Student (District)  \
0                                                 16586.0
1                                                 16356.0
2                                                 16480.0
3                                                 17876.0
4                                                 18804.0
…                                                   …
6601                                              18916.0
6602                                              22078.0
6603                                              23672.0
6604                                              23458.0
6605                                              27114.0

      Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                     0.0
1                                                     0.0
2                                                     0.0
3                                                     0.0
4                                                     0.0
…                                                       …
6601                                                152.0
6602                                                186.0
6603                                                274.0
```

```
6604                                                238.0
6605                                                266.0

      Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                                 10458.0
1                                                 10592.0
2                                                 10534.0
3                                                 11010.0
4                                                 11618.0
…                                                   …
6601                                              11812.0
6602                                              12314.0
6603                                              13470.0
6604                                              14540.0
6605                                              16390.0

      Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                   952.0
1                                                   954.0
2                                                   912.0
3                                                   938.0
4                                                   968.0
…                                                   …
6601                                                888.0
6602                                               1086.0
6603                                               1402.0
6604                                               1342.0
6605                                               1358.0

      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)
0                                                 1116.0
1                                                 1142.0
2                                                 1102.0
3                                                 1194.0
4                                                 1320.0
…                                                   …
6601                                              1520.0
6602                                              1586.0
6603                                              1764.0
6604                                              2018.0
6605                                              2228.0

[6606 rows x 16 columns]
```

[52]:
```
# Pivot to make Count as columns
pivoted = sd_acceptance.pivot(index=['District Name', 'Year'], columns='Count',␣
↪values='Student Count').reset_index()
pivoted
```

[52]:
```
Count      District Name       Year    Adm    App    Enr
0            ABC Unified  2010-2011  205.0  270.0  120.0
1            ABC Unified  2011-2012  207.0  281.0  131.0
2            ABC Unified  2012-2013  183.0  258.0  108.0
3            ABC Unified  2013-2014  200.0  275.0  105.0
4            ABC Unified  2014-2015  182.0  271.0   98.0
…                    …        …    …    …    …
2232   Yuba City Unified  2014-2015   64.0   94.0   43.0
2233   Yuba City Unified  2015-2016   62.0   98.0   31.0
2234   Yuba City Unified  2016-2017   74.0  111.0   40.0
2235   Yuba City Unified  2017-2018   56.0   87.0   27.0
2236   Yuba City Unified  2018-2019   75.0  114.0   41.0

[2237 rows x 5 columns]
```

[53]:
```
# Retain only one entry for each district per year
sd_adm = sd_acceptance[sd_acceptance['Count'] == 'Adm']

dist_acceptance = pd.merge(sd_adm, pivoted, on=['District Name', 'Year'],␣
↪how='inner')
dist_acceptance
```

[53]:
```
          District Name Count       Year  Student Count  \
0           ABC Unified   Adm  2010-2011            205
1           ABC Unified   Adm  2011-2012            207
2           ABC Unified   Adm  2012-2013            183
3           ABC Unified   Adm  2013-2014            200
4           ABC Unified   Adm  2014-2015            182
…                   …   …        …              …
2229  Yuba City Unified   Adm  2014-2015             64
2230  Yuba City Unified   Adm  2015-2016             62
2231  Yuba City Unified   Adm  2016-2017             74
2232  Yuba City Unified   Adm  2017-2018             56
2233  Yuba City Unified   Adm  2018-2019             75

      ACT Test Takers # (District)  Per Pupil Ratio: Teacher (District)  \
0                              0.0                                 49.0
1                              0.0                                 49.6
2                              0.0                                 48.2
3                              0.0                                 49.2
4                              0.0                                 48.0
…                              …                                   …
```

```
2229                          0.0                                  40.0
2230                        402.0                                  39.6
2231                        332.0                                  39.4
2232                        296.0                                  36.0
2233                        232.0                                  38.6

      Avg Years Teaching (District)  Teacher Salary-Avg (District)  \
0                              30.0                       143614.0
1                              30.0                       143774.0
2                              28.0                       143186.0
3                              28.0                       152394.0
4                              28.0                       156382.0
…                               …                            …
2229                           24.0                       128330.0
2230                           24.0                       132862.0
2231                           24.0                       137350.0
2232                           24.0                       146336.0
2233                           24.0                       151670.0

      Chronic Absenteeism % (District)  \
0                                  0.0
1                                  0.0
2                                  0.0
3                                  0.0
4                                  0.0
…                                   …
2229                               0.0
2230                               0.0
2231                              36.4
2232                              24.6
2233                              24.4

      Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                    16108.0
1                                                    16216.0
2                                                    16010.0
3                                                    16818.0
4                                                    17850.0
…                                                      …
2229                                                 18236.0
2230                                                 19472.0
2231                                                 21636.0
2232                                                 23652.0
2233                                                 25974.0

      Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
```

```
0                                                  688.0
1                                                  756.0
2                                                  598.0
3                                                  768.0
4                                                  990.0
…                                                    …
2229                                              1224.0
2230                                              1322.0
2231                                              1288.0
2232                                              1622.0
2233                                              1586.0

      Total Gen Fund Revenues Per Student (District)  \
0                                              16586.0
1                                              16356.0
2                                              16480.0
3                                              17876.0
4                                              18804.0
…                                                ...
2229                                           18916.0
2230                                           22078.0
2231                                           23672.0
2232                                           23458.0
2233                                           27114.0

      Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                0.0
1                                                0.0
2                                                0.0
3                                                0.0
4                                                0.0
…                                                  …
2229                                           152.0
2230                                           186.0
2231                                           274.0
2232                                           238.0
2233                                           266.0

      Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                              10458.0
1                                              10592.0
2                                              10534.0
3                                              11010.0
4                                              11618.0
…                                                ...
```

```
2229                                                 11812.0
2230                                                 12314.0
2231                                                 13470.0
2232                                                 14540.0
2233                                                 16390.0

      Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                     952.0
1                                                     954.0
2                                                     912.0
3                                                     938.0
4                                                     968.0
…                                                     …
2229                                                  888.0
2230                                                 1086.0
2231                                                 1402.0
2232                                                 1342.0
2233                                                 1358.0

      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                    1116.0
1                                                    1142.0
2                                                    1102.0
3                                                    1194.0
4                                                    1320.0
…                                                     …
2229                                                 1520.0
2230                                                 1586.0
2231                                                 1764.0
2232                                                 2018.0
2233                                                 2228.0

        Adm    App    Enr
0     205.0  270.0  120.0
1     207.0  281.0  131.0
2     183.0  258.0  108.0
3     200.0  275.0  105.0
4     182.0  271.0   98.0
…     …    …    …
2229   64.0   94.0   43.0
2230   62.0   98.0   31.0
2231   74.0  111.0   40.0
2232   56.0   87.0   27.0
2233   75.0  114.0   41.0
```

```
[2234 rows x 19 columns]
```

[54]:
```
# Add new column of admission rates by dividing admission by applied
dist_acceptance['Admission Rate to UC System'] = dist_acceptance['Adm'] /␣
↪dist_acceptance['App']
dist_acceptance
```

[54]:
```
           District Name Count       Year  Student Count  \
0            ABC Unified   Adm  2010-2011            205
1            ABC Unified   Adm  2011-2012            207
2            ABC Unified   Adm  2012-2013            183
3            ABC Unified   Adm  2013-2014            200
4            ABC Unified   Adm  2014-2015            182
…                    …   …        …              …
2229  Yuba City Unified   Adm  2014-2015             64
2230  Yuba City Unified   Adm  2015-2016             62
2231  Yuba City Unified   Adm  2016-2017             74
2232  Yuba City Unified   Adm  2017-2018             56
2233  Yuba City Unified   Adm  2018-2019             75

      ACT Test Takers # (District)  Per Pupil Ratio: Teacher (District)  \
0                              0.0                                 49.0
1                              0.0                                 49.6
2                              0.0                                 48.2
3                              0.0                                 49.2
4                              0.0                                 48.0
…                              …                                   …
2229                           0.0                                 40.0
2230                         402.0                                 39.6
2231                         332.0                                 39.4
2232                         296.0                                 36.0
2233                         232.0                                 38.6

      Avg Years Teaching (District)  Teacher Salary-Avg (District)  \
0                              30.0                         143614.0
1                              30.0                         143774.0
2                              28.0                         143186.0
3                              28.0                         152394.0
4                              28.0                         156382.0
…                              …                             …
2229                           24.0                         128330.0
2230                           24.0                         132862.0
2231                           24.0                         137350.0
2232                           24.0                         146336.0
2233                           24.0                         151670.0

      Chronic Absenteeism % (District)  \
```

```
0                                    0.0
1                                    0.0
2                                    0.0
3                                    0.0
4                                    0.0
…                                    …
2229                                 0.0
2230                                 0.0
2231                                36.4
2232                                24.6
2233                                24.4

      Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                    16108.0
1                                                    16216.0
2                                                    16010.0
3                                                    16818.0
4                                                    17850.0
…                                                      …
2229                                                 18236.0
2230                                                 19472.0
2231                                                 21636.0
2232                                                 23652.0
2233                                                 25974.0

      Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                     688.0
1                                                     756.0
2                                                     598.0
3                                                     768.0
4                                                     990.0
…                                                       …
2229                                                 1224.0
2230                                                 1322.0
2231                                                 1288.0
2232                                                 1622.0
2233                                                 1586.0

      Total Gen Fund Revenues Per Student (District)  \
0                                               16586.0
1                                               16356.0
2                                               16480.0
3                                               17876.0
4                                               18804.0
…                                                 …
2229                                            18916.0
```

2230	22078.0
2231	23672.0
2232	23458.0
2233	27114.0

Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District) \

0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
…	…
2229	152.0
2230	186.0
2231	274.0
2232	238.0
2233	266.0

Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District) \

0	10458.0
1	10592.0
2	10534.0
3	11010.0
4	11618.0
…	…
2229	11812.0
2230	12314.0
2231	13470.0
2232	14540.0
2233	16390.0

Gen Fund Exp by Activity - 7000-7999 General Administration Per Student # (District) \

0	952.0
1	954.0
2	912.0
3	938.0
4	968.0
…	…
2229	888.0
2230	1086.0
2231	1402.0
2232	1342.0
2233	1358.0

```
      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                    1116.0
1                                                    1142.0
2                                                    1102.0
3                                                    1194.0
4                                                    1320.0
…                                                     …
2229                                                 1520.0
2230                                                 1586.0
2231                                                 1764.0
2232                                                 2018.0
2233                                                 2228.0

        Adm    App    Enr  Admission Rate to UC System
0     205.0  270.0  120.0                     0.759259
1     207.0  281.0  131.0                     0.736655
2     183.0  258.0  108.0                     0.709302
3     200.0  275.0  105.0                     0.727273
4     182.0  271.0   98.0                     0.671587
…     …    …    …                          …
2229   64.0   94.0   43.0                     0.680851
2230   62.0   98.0   31.0                     0.632653
2231   74.0  111.0   40.0                     0.666667
2232   56.0   87.0   27.0                     0.643678
2233   75.0  114.0   41.0                     0.657895

[2234 rows x 20 columns]
```

1.9 Data Cleaning

[55]:
```
# Remove nan and specified string values
sd_all_cleaned = dist_acceptance.copy()
for col in sd_all_cleaned.columns:
  sd_all_cleaned = sd_all_cleaned.replace('', np.nan)
  sd_all_cleaned = sd_all_cleaned[~(sd_all_cleaned[col].isin(["redacted",␣
 ↪"(1)", "Not Certified"]) | pd.isna(sd_all_cleaned[col]))]
```

[56]:
```
# Split into training, validation and test sets
years = sd_all_cleaned["Year"].unique()

# Use 2010-2016 for training
sd_train = sd_all_cleaned[sd_all_cleaned["Year"].isin(years[:len(years)-2])]

sd_val = sd_all_cleaned[sd_all_cleaned['Year'] == "2017-2018"] # Use 2017 for␣
 ↪validation
```

```
sd_test = sd_all_cleaned[sd_all_cleaned['Year'] == "2018-2019"] # Use 2018 for␣
↪validation

# Choose the features to be used
cols = list(sd_all_cleaned.columns)[4:len(sd_all_cleaned.columns)-4]


# Split into X and y sets
X_train = sd_train[cols].astype(float)
y_train = sd_train["Admission Rate to UC System"]

X_val = sd_val[cols].astype(float)
y_val = sd_val["Admission Rate to UC System"]

X_test = sd_test[cols].astype(float)
y_test = sd_test["Admission Rate to UC System"]

sd_train
```

[56]:

	District Name	Count	Year	Student Count	\
0	ABC Unified	Adm	2010-2011	205	
1	ABC Unified	Adm	2011-2012	207	
2	ABC Unified	Adm	2012-2013	183	
3	ABC Unified	Adm	2013-2014	200	
4	ABC Unified	Adm	2014-2015	182	
…	…	…	…	…	
2227	Yuba City Unified	Adm	2012-2013	72	
2228	Yuba City Unified	Adm	2013-2014	52	
2229	Yuba City Unified	Adm	2014-2015	64	
2230	Yuba City Unified	Adm	2015-2016	62	
2231	Yuba City Unified	Adm	2016-2017	74	

	ACT Test Takers # (District)	Per Pupil Ratio: Teacher (District)	\
0	0.0	49.0	
1	0.0	49.6	
2	0.0	48.2	
3	0.0	49.2	
4	0.0	48.0	
…	…	…	
2227	0.0	41.6	
2228	0.0	40.6	
2229	0.0	40.0	
2230	402.0	39.6	
2231	332.0	39.4	

	Avg Years Teaching (District)	Teacher Salary-Avg (District)	\
0	30.0	143614.0	

```
1                                30.0                        143774.0
2                                28.0                        143186.0
3                                28.0                        152394.0
4                                28.0                        156382.0
…                                …                            …
2227                             22.0                        116650.0
2228                             22.0                        121336.0
2229                             24.0                        128330.0
2230                             24.0                        132862.0
2231                             24.0                        137350.0

      Chronic Absenteeism % (District)  \
0                                   0.0
1                                   0.0
2                                   0.0
3                                   0.0
4                                   0.0
…                                   …
2227                                0.0
2228                                0.0
2229                                0.0
2230                                0.0
2231                               36.4

      Current Exp of Educ per ADA (Ed Code 41372) (District)  \
0                                                    16108.0
1                                                    16216.0
2                                                    16010.0
3                                                    16818.0
4                                                    17850.0
…                                                      …
2227                                                 15306.0
2228                                                 16684.0
2229                                                 18236.0
2230                                                 19472.0
2231                                                 21636.0

      Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
0                                                      688.0
1                                                      756.0
2                                                      598.0
3                                                      768.0
4                                                      990.0
…                                                      …
2227                                                   644.0
2228                                                   806.0
```

```
2229                                                1224.0
2230                                                1322.0
2231                                                1288.0

      Total Gen Fund Revenues Per Student (District)  \
0                                                16586.0
1                                                16356.0
2                                                16480.0
3                                                17876.0
4                                                18804.0
…                                                  …
2227                                             15410.0
2228                                             17072.0
2229                                             18916.0
2230                                             22078.0
2231                                             23672.0

      Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                    0.0
1                                                    0.0
2                                                    0.0
3                                                    0.0
4                                                    0.0
…                                                   …
2227                                               126.0
2228                                               144.0
2229                                               152.0
2230                                               186.0
2231                                               274.0

      Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                                10458.0
1                                                10592.0
2                                                10534.0
3                                                11010.0
4                                                11618.0
…                                                  …
2227                                              9974.0
2228                                             10762.0
2229                                             11812.0
2230                                             12314.0
2231                                             13470.0

      Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
```

```
0                                                    952.0
1                                                    954.0
2                                                    912.0
3                                                    938.0
4                                                    968.0
…                                                      …
2227                                                 852.0
2228                                                 864.0
2229                                                 888.0
2230                                                1086.0
2231                                                1402.0

      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                   1116.0
1                                                   1142.0
2                                                   1102.0
3                                                   1194.0
4                                                   1320.0
…                                                     …
2227                                                1232.0
2228                                                1404.0
2229                                                1520.0
2230                                                1586.0
2231                                                1764.0

        Adm    App    Enr  Admission Rate to UC System
0     205.0  270.0  120.0                      0.759259
1     207.0  281.0  131.0                      0.736655
2     183.0  258.0  108.0                      0.709302
3     200.0  275.0  105.0                      0.727273
4     182.0  271.0   98.0                      0.671587
…       …      …      …                           …
2227   72.0  108.0   40.0                      0.666667
2228   52.0   77.0   26.0                      0.675325
2229   64.0   94.0   43.0                      0.680851
2230   62.0   98.0   31.0                      0.632653
2231   74.0  111.0   40.0                      0.666667

[1671 rows x 20 columns]
```

1.10 Linear Regression

[57]:
```
X1 = sm.add_constant(X_train)

linreg1 = sm.OLS(y_train, X1).fit()
print(linreg1.summary())
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.037
Model:                                    OLS   Adj. R-squared:
0.030
Method:                         Least Squares   F-statistic:
5.305
Date:                        Wed, 12 May 2021   Prob (F-statistic):
7.39e-09
Time:                                14:59:54   Log-Likelihood:
1403.4
No. Observations:                        1671   AIC:
-2781.
Df Residuals:                            1658   BIC:
-2710.
Df Model:                                  12
Covariance Type:                    nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
0.6685      0.005    132.221      0.000       0.659       0.678
ACT Test Takers # (District)
1.56e-07   5.12e-07      0.305      0.761   -8.48e-07    1.16e-06
Per Pupil Ratio: Teacher (District)
0.0005      0.000      1.484      0.138      -0.000       0.001
Avg Years Teaching (District)
0.0013      0.001      2.176      0.030       0.000       0.003
Teacher Salary-Avg (District)
-1.038e-07   1.27e-07     -0.816      0.415   -3.53e-07    1.46e-07
Chronic Absenteeism % (District)
-0.0005      0.000     -1.685      0.092      -0.001    8.22e-05
Current Exp of Educ per ADA (Ed Code 41372) (District)
-2.084e-06   3.89e-06     -0.536      0.592   -9.71e-06    5.54e-06
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
4.927e-06   8.72e-06      0.565      0.572   -1.22e-05     2.2e-05
```

```
Total Gen Fund Revenues Per Student (District)
7.279e-07   1.96e-06      0.371      0.710   -3.12e-06    4.57e-06
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.664e-05   1.49e-05      3.802      0.000    2.74e-05    8.59e-05
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
-3.356e-06   4.37e-06     -0.768      0.443   -1.19e-05    5.22e-06
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -1.071e-05   1.05e-05     -1.021      0.307   -3.13e-05    9.86e-06
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
5.989e-06   7.08e-06      0.846      0.398    -7.9e-06    1.99e-05
==============================================================================
Omnibus:                       24.329   Durbin-Watson:                   1.509
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.430
Skew:                           0.088   Prob(JB):                     1.66e-09
Kurtosis:                       3.741   Cond. No.                     3.63e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.63e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

We want to remove the features with high multicollinearity by examining their VIF values. We also keep an eye out for drastic changes in the R2 score when removing.

[58]:
```
VIF(X1, X_train.columns)
```

[58]:
```
ACT Test Takers # (District)
1.589048
Per Pupil Ratio: Teacher (District)
25.840407
Avg Years Teaching (District)
27.805104
Teacher Salary-Avg (District)
36.693027
Chronic Absenteeism % (District)
1.462258
Current Exp of Educ per ADA (Ed Code 41372) (District)
641.959520
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
6.862443
Total Gen Fund Revenues Per Student (District)
187.257928
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
3.099618
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
336.746454
```

```
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     12.327774
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
17.815472
dtype: float64
```

[59]:
```
X_train2 = X_train.drop(['Current Exp of Educ per ADA (Ed Code 41372)␣
↪(District)'],axis=1)
X2 = sm.add_constant(X_train2)

linreg2 = sm.OLS(y_train, X2).fit()
print(linreg2.summary())
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.037
Model:                                    OLS   Adj. R-squared:
0.030
Method:                         Least Squares   F-statistic:
5.763
Date:                        Wed, 12 May 2021   Prob (F-statistic):
3.34e-09
Time:                                14:59:54   Log-Likelihood:
1403.2
No. Observations:                        1671   AIC:
-2782.
Df Residuals:                            1659   BIC:
-2717.
Df Model:                                  11
Covariance Type:                    nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
0.6686      0.005    132.264      0.000       0.659       0.678
ACT Test Takers # (District)
1.436e-07   5.11e-07      0.281      0.779   -8.59e-07    1.15e-06
Per Pupil Ratio: Teacher (District)
0.0005      0.000       1.494      0.135      -0.000       0.001
Avg Years Teaching (District)
0.0014      0.001       2.254      0.024       0.000       0.003
Teacher Salary-Avg (District)
-1.072e-07   1.27e-07     -0.843      0.399   -3.56e-07    1.42e-07
```

```
Chronic Absenteeism % (District)
-0.0005      0.000     -1.716      0.086      -0.001     7.3e-05
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
5.112e-06   8.71e-06      0.587      0.557    -1.2e-05    2.22e-05
Total Gen Fund Revenues Per Student (District)
1.399e-07   1.63e-06      0.086      0.931   -3.05e-06    3.33e-06
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.498e-05   1.46e-05      3.774      0.000    2.64e-05    8.36e-05
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
-5.23e-06   2.62e-06     -1.993      0.046   -1.04e-05   -8.31e-08
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -1.297e-05    9.6e-06     -1.351      0.177   -3.18e-05    5.85e-06
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
4.253e-06    6.3e-06      0.676      0.499    -8.1e-06    1.66e-05
==============================================================================
Omnibus:                       24.568   Durbin-Watson:                   1.508
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.840
Skew:                           0.090   Prob(JB):                     1.35e-09
Kurtosis:                       3.744   Cond. No.                     3.60e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.6e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[60]:
```
VIF(X2, X_train2.columns)
```

[60]:
```
ACT Test Takers # (District)
1.585811
Per Pupil Ratio: Teacher (District)
25.832795
Avg Years Teaching (District)
27.425935
Teacher Salary-Avg (District)
36.603462
Chronic Absenteeism % (District)
1.458321
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
6.851727
Total Gen Fund Revenues Per Student (District)
128.708855
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.966398
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
121.431333
```

```
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     10.334302
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
14.093279
dtype: float64
```

[61]:
```
X_train3 = X_train2.drop(['Total Gen Fund Revenues Per Student␣
↪(District)'],axis=1)
X3 = sm.add_constant(X_train3)
linreg3 = sm.OLS(y_train, X3).fit()
print(linreg3.summary(),'\n')
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.037
Model:                                     OLS   Adj. R-squared:
0.031
Method:                          Least Squares   F-statistic:
6.342
Date:                         Wed, 12 May 2021   Prob (F-statistic):
1.27e-09
Time:                                 14:59:54   Log-Likelihood:
1403.2
No. Observations:                         1671   AIC:
-2784.
Df Residuals:                             1660   BIC:
-2725.
Df Model:                                   10
Covariance Type:                     nonrobust
==============================================================================
==========================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------
--------------------------------------------------------------------------
const
0.6686      0.005    132.369      0.000       0.659       0.678
ACT Test Takers # (District)
1.537e-07   4.98e-07      0.309      0.757   -8.22e-07    1.13e-06
Per Pupil Ratio: Teacher (District)
0.0005      0.000      1.499      0.134      -0.000       0.001
Avg Years Teaching (District)
0.0014      0.001      2.254      0.024       0.000       0.003
Teacher Salary-Avg (District)
-1.063e-07   1.27e-07     -0.839      0.401   -3.55e-07    1.42e-07
Chronic Absenteeism % (District)
```

```
-0.0005      0.000     -1.718      0.086      -0.001    7.25e-05
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
5.321e-06   8.36e-06      0.637      0.524   -1.11e-05    2.17e-05
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.54e-05   1.37e-05      4.032      0.000    2.84e-05    8.23e-05
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
-5.075e-06   1.91e-06     -2.660      0.008   -8.82e-06   -1.33e-06
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -1.267e-05   8.95e-06     -1.417      0.157   -3.02e-05    4.87e-06
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
4.407e-06   6.04e-06      0.730      0.465   -7.43e-06    1.62e-05
==============================================================================
Omnibus:                       24.533   Durbin-Watson:                   1.508
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               40.772
Skew:                           0.090   Prob(JB):                     1.40e-09
Kurtosis:                       3.744   Cond. No.                     3.57e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.57e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[62]:
```
VIF(X3, X_train3.columns)
```

[62]:
```
ACT Test Takers # (District)
1.502436
Per Pupil Ratio: Teacher (District)
25.786986
Avg Years Teaching (District)
27.348602
Teacher Salary-Avg (District)
36.363642
Chronic Absenteeism % (District)
1.458112
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
6.317612
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.639742
Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
64.205917
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     8.984514
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
12.960447
```

```
dtype: float64
```

[63]:
```
X_train4 = X_train3.drop(['Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)'],axis=1)
X4 = sm.add_constant(X_train4)
linreg4 = sm.OLS(y_train, X4).fit()
print(linreg4.summary(),'\n')
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.033
Model:                                     OLS   Adj. R-squared:
0.027
Method:                          Least Squares   F-statistic:
6.238
Date:                         Wed, 12 May 2021   Prob (F-statistic):
1.05e-08
Time:                                 14:59:54   Log-Likelihood:
1399.7
No. Observations:                         1671   AIC:
-2779.
Df Residuals:                             1661   BIC:
-2725.
Df Model:                                    9
Covariance Type:                     nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
0.6714      0.005    135.632      0.000       0.662       0.681
ACT Test Takers # (District)
1.164e-07   4.98e-07      0.234      0.815   -8.61e-07    1.09e-06
Per Pupil Ratio: Teacher (District)
0.0005      0.000       1.525      0.128      -0.000       0.001
Avg Years Teaching (District)
0.0008      0.001       1.464      0.143      -0.000       0.002
Teacher Salary-Avg (District)
-2.538e-07   1.14e-07     -2.225      0.026   -4.78e-07   -3.01e-08
Chronic Absenteeism % (District)
-0.0007      0.000     -2.247      0.025      -0.001   -8.36e-05
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
-2.808e-06   7.79e-06     -0.360      0.719   -1.81e-05    1.25e-05
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
```

```
4.941e-05   1.36e-05      3.639      0.000    2.28e-05     7.6e-05
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -2.438e-05    7.8e-06     -3.124      0.002   -3.97e-05   -9.07e-06
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
-1.571e-06   5.61e-06     -0.280      0.780   -1.26e-05    9.44e-06
==============================================================================
Omnibus:                       26.262   Durbin-Watson:                   1.507
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.858
Skew:                           0.104   Prob(JB):                     2.99e-10
Kurtosis:                       3.766   Cond. No.                     3.48e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[64]:
```
VIF(X4, X_train4.columns)
```

[64]:
```
ACT Test Takers # (District)
1.501246
Per Pupil Ratio: Teacher (District)
25.784075
Avg Years Teaching (District)
24.446694
Teacher Salary-Avg (District)
29.391663
Chronic Absenteeism % (District)
1.407756
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
5.473385
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.568927
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     6.810806
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
11.164455
dtype: float64
```

[65]:
```
X_train5 = X_train4.drop(['Teacher Salary-Avg (District)'],axis=1)
X5 = sm.add_constant(X_train5)
linreg5 = sm.OLS(y_train, X5).fit()
print(linreg4.summary(),'\n')
```

```
                                OLS Regression Results
==============================================================================
```

```
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.033
Model:                                    OLS   Adj. R-squared:
0.027
Method:                         Least Squares   F-statistic:
6.238
Date:                        Wed, 12 May 2021   Prob (F-statistic):
1.05e-08
Time:                                14:59:54   Log-Likelihood:
1399.7
No. Observations:                        1671   AIC:
-2779.
Df Residuals:                            1661   BIC:
-2725.
Df Model:                                   9
Covariance Type:                    nonrobust
================================================================================
================================================================================
           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
const
0.6714      0.005    135.632      0.000       0.662       0.681
ACT Test Takers # (District)
1.164e-07   4.98e-07      0.234      0.815   -8.61e-07    1.09e-06
Per Pupil Ratio: Teacher (District)
0.0005      0.000      1.525      0.128      -0.000       0.001
Avg Years Teaching (District)
0.0008      0.001      1.464      0.143      -0.000       0.002
Teacher Salary-Avg (District)
-2.538e-07   1.14e-07     -2.225      0.026   -4.78e-07   -3.01e-08
Chronic Absenteeism % (District)
-0.0007      0.000     -2.247      0.025      -0.001   -8.36e-05
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
-2.808e-06   7.79e-06     -0.360      0.719   -1.81e-05    1.25e-05
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
4.941e-05   1.36e-05      3.639      0.000    2.28e-05     7.6e-05
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -2.438e-05    7.8e-06     -3.124      0.002   -3.97e-05   -9.07e-06
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
-1.571e-06   5.61e-06     -0.280      0.780   -1.26e-05    9.44e-06
==============================================================================
Omnibus:                       26.262   Durbin-Watson:                   1.507
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.858
Skew:                           0.104   Prob(JB):                     2.99e-10
Kurtosis:                       3.766   Cond. No.                     3.48e+05
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[66]:
```
VIF(X5, X_train5.columns)
```

[66]:
```
ACT Test Takers # (District)
1.500355
Per Pupil Ratio: Teacher (District)
16.889439
Avg Years Teaching (District)
21.745809
Chronic Absenteeism % (District)
1.346675
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
5.291020
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.503291
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     6.734335
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
11.091477
dtype: float64
```

[67]:
```
X_train6 = X_train5.drop(['Avg Years Teaching (District)'],axis=1)
X6 = sm.add_constant(X_train6)
linreg6 = sm.OLS(y_train, X6).fit()
print(linreg6.summary(),'\n')
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.029
Model:                                     OLS   Adj. R-squared:
0.025
Method:                          Least Squares   F-statistic:
7.213
Date:                         Wed, 12 May 2021   Prob (F-statistic):
1.57e-08
Time:                                 14:59:54   Log-Likelihood:
1396.9
No. Observations:                         1671   AIC:
-2778.
```

```
Df Residuals:                                1663   BIC:
-2734.
Df Model:                                       7
Covariance Type:                        nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
0.6722      0.005    140.476      0.000       0.663       0.682
ACT Test Takers # (District)
1.509e-07   4.92e-07      0.307      0.759   -8.14e-07    1.12e-06
Per Pupil Ratio: Teacher (District)
0.0003      0.000      1.719      0.086   -3.76e-05       0.001
Chronic Absenteeism % (District)
-0.0008      0.000     -2.876      0.004      -0.001      -0.000
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
-5.834e-06   7.67e-06     -0.761      0.447   -2.09e-05    9.21e-06
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.435e-05   1.34e-05      4.050      0.000     2.8e-05    8.07e-05
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -2.591e-05   7.76e-06     -3.340      0.001   -4.11e-05   -1.07e-05
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
-1.33e-06   5.36e-06     -0.248      0.804   -1.18e-05    9.18e-06
==============================================================================
Omnibus:                       27.765   Durbin-Watson:                   1.504
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.611
Skew:                           0.125   Prob(JB):                     1.25e-10
Kurtosis:                       3.770   Cond. No.                     1.20e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.2e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[68]:
```
VIF(X6, X_train6.columns)
```

[68]:
```
ACT Test Takers # (District)
1.461569
Per Pupil Ratio: Teacher (District)
4.810530
Chronic Absenteeism % (District)
1.328715
```

```
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
5.288022
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.503026
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)     6.716536
Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student # (District)
10.151273
dtype: float64
```

[69]:
```
X_train7 = X_train6.drop(['Gen Fund Exp by Activity - 3000-3999 Pupil Services␣
↪Per Student # (District)'],axis=1)
X7 = sm.add_constant(X_train7)
linreg7 = sm.OLS(y_train, X7).fit()
print(linreg7.summary(),'\n')
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.029
Model:                                     OLS   Adj. R-squared:
0.026
Method:                          Least Squares   F-statistic:
8.410
Date:                         Wed, 12 May 2021   Prob (F-statistic):
5.23e-09
Time:                                 14:59:54   Log-Likelihood:
1396.9
No. Observations:                         1671   AIC:
-2780.
Df Residuals:                             1664   BIC:
-2742.
Df Model:                                    6
Covariance Type:                     nonrobust
==============================================================================
==========================================================================
           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------
--------------------------------------------------------------------------
const
0.6726      0.005    147.922      0.000       0.664       0.681
ACT Test Takers # (District)
1.332e-07   4.87e-07      0.274      0.784   -8.22e-07    1.09e-06
Per Pupil Ratio: Teacher (District)
0.0002      0.000       1.855      0.064   -1.41e-05       0.001
Chronic Absenteeism % (District)
```

```
-0.0008      0.000     -2.976      0.003      -0.001      -0.000
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
-6.471e-06   7.23e-06     -0.895      0.371   -2.06e-05     7.7e-06
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.431e-05   1.34e-05      4.049      0.000     2.8e-05    8.06e-05
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -2.677e-05   6.95e-06     -3.850      0.000   -4.04e-05   -1.31e-05
==============================================================================
Omnibus:                       27.616   Durbin-Watson:                   1.504
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.232
Skew:                           0.125   Prob(JB):                     1.51e-10
Kurtosis:                       3.766   Cond. No.                     1.13e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.13e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[70]:
```
VIF(X7, X_train7.columns)
```

[70]:
```
ACT Test Takers # (District)
1.430676
Per Pupil Ratio: Teacher (District)
3.542321
Chronic Absenteeism % (District)
1.282451
Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student (District)
4.696959
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.502720
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    5.396638
dtype: float64
```

After removing the variables with high VIF values, we now want to remove variables that are not significant based on their p-values. We will remove all variables with p-value > 0.05. For simplicity, we removed the variables all at once, but we did check that each variable is still insignificant after each removal.

We initially also removed "Gen Fund Exp by Activity - 7000-7999 General Administration Per Student # (District)" but that drastically reduced our R2. Therefore, we removed "Per Pupil Ratio: Teacher (District)" instead.

[71]:
```
# Remove variables with high p-values (>0.05)
p_cols = ['ACT Test Takers # (District)',
```

```
        'Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student␣
↪(District)',
        'Per Pupil Ratio: Teacher (District)']

X_train_p = X_train7.drop(p_cols, axis = 1)
X_p = sm.add_constant(X_train_p)

linreg_p = sm.OLS(y_train, X_p).fit()
print(linreg_p.summary())
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.027
Model:                                     OLS   Adj. R-squared:
0.026
Method:                          Least Squares   F-statistic:
15.65
Date:                         Wed, 12 May 2021   Prob (F-statistic):
4.88e-10
Time:                                 14:59:54   Log-Likelihood:
1395.1
No. Observations:                         1671   AIC:
-2782.
Df Residuals:                             1667   BIC:
-2761.
Df Model:                                    3
Covariance Type:                     nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------------------------------------------
const
0.6735      0.004    156.470      0.000       0.665       0.682
Chronic Absenteeism % (District)
-0.0009      0.000     -3.510      0.000      -0.001      -0.000
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
5.54e-05   1.29e-05      4.308      0.000    3.02e-05    8.06e-05
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District) -2.273e-05   4.68e-06     -4.861      0.000   -3.19e-05   -1.36e-05
==============================================================================
Omnibus:                       24.019   Durbin-Watson:                   1.508
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               39.975
Skew:                           0.084   Prob(JB):                     2.09e-09
Kurtosis:                       3.739   Cond. No.                     2.31e+03
```

```
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.31e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[72]:
```
VIF(X_p, X_train_p.columns)
```

[72]:
```
Chronic Absenteeism % (District)
1.139194
Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student # (District)
2.299605
Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)    2.441869
dtype: float64
```

Our final linear regression model has a really low R2 score of 0.027, so we will most likely not use it in our blended model.

1.11 Decision Tree

[73]:
```
grid_values = {'ccp_alpha': np.linspace(0, 0.001, 500), # 500 cp values in [0,0.
↪001]
               'min_samples_leaf': [5],
               'min_samples_split': [20],
               'max_depth': [30],
               'random_state': [88]}

dtr = DecisionTreeRegressor()

cv = KFold(n_splits = 5, random_state = 88, shuffle = True) # 5 fold cross␣
↪validation for each cp value

dtr_cv = GridSearchCV(dtr, param_grid = grid_values, scoring = 'r2', cv = cv,␣
↪verbose = 1)

dtr_cv.fit(X_train, y_train)
```

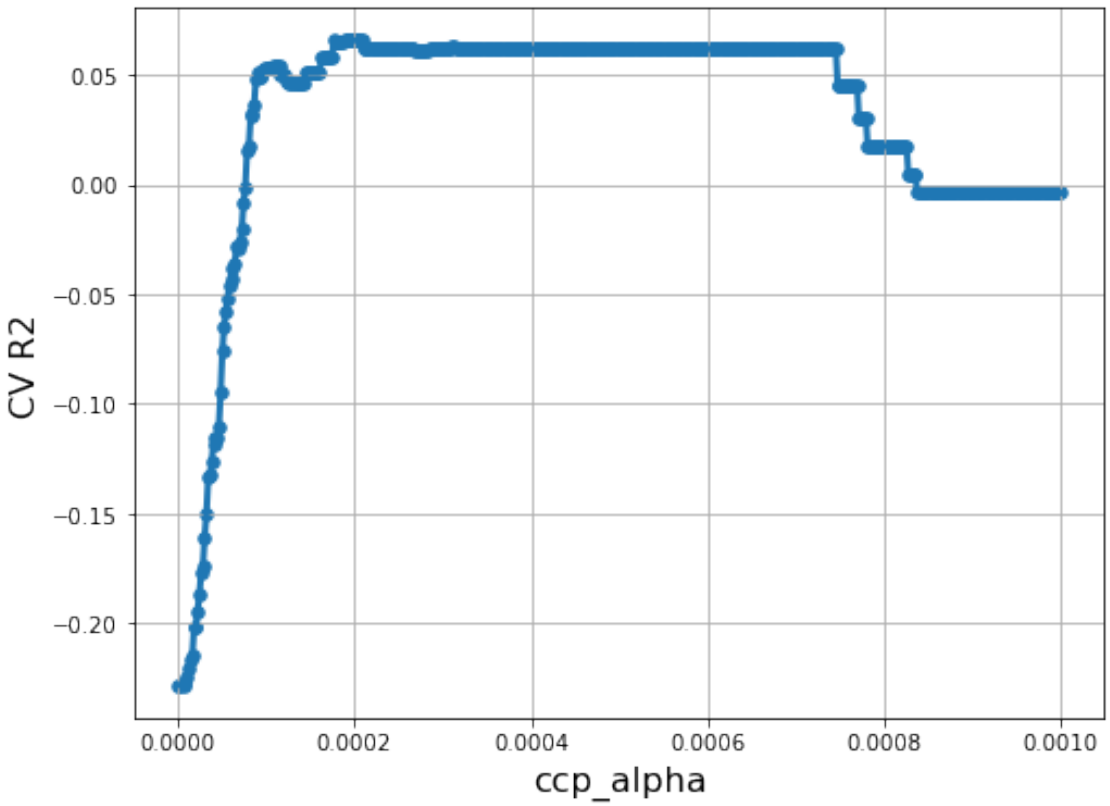
```
Fitting 5 folds for each of 500 candidates, totalling 2500 fits

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 2500 out of 2500 | elapsed:   21.1s finished
```

[73]:
```
GridSearchCV(cv=KFold(n_splits=5, random_state=88, shuffle=True),
             estimator=DecisionTreeRegressor(),
```

```
             param_grid={'ccp_alpha': array([0.00000000e+00, 2.00400802e-06,
4.00801603e-06, 6.01202405e-06,
       8.01603206e-06, 1.00200401e-05, 1.20240481e-05, 1.40280561e-05,
       1.60320641e-05, 1.80360721e-05, 2.00400802e-05, 2.20440882e-05,
       2.40480962e-05, 2.60521042e-05, 2.80561122e-…
       9.69939880e-04, 9.71943888e-04, 9.73947896e-04, 9.75951904e-04,
       9.77955912e-04, 9.79959920e-04, 9.81963928e-04, 9.83967936e-04,
       9.85971944e-04, 9.87975952e-04, 9.89979960e-04, 9.91983968e-04,
       9.93987976e-04, 9.95991984e-04, 9.97995992e-04, 1.00000000e-03]),
                         'max_depth': [30], 'min_samples_leaf': [5],
                         'min_samples_split': [20], 'random_state': [88]},
             scoring='r2', verbose=1)
```

Below is the plot examining the R2 for each ccp_alpha between 0 and 0.001. We will choose the optimal complexity parameter based on the highest CV R2.

[74]:
```
ccp_alpha_dtr = dtr_cv.cv_results_['param_ccp_alpha'].data
r2_scores = dtr_cv.cv_results_['mean_test_score']

plt.figure(figsize=(8, 6))
plt.xlabel('ccp_alpha', fontsize=16)
plt.ylabel('CV R2', fontsize=16)
plt.scatter(ccp_alpha_dtr, r2_scores, s=30)
plt.plot(ccp_alpha_dtr, r2_scores, linewidth=3)
plt.grid(True, which='both')

plt.show()
```



[75]:
```
print('Best ccp_alpha', dtr_cv.best_params_)
```

```
Best ccp_alpha {'ccp_alpha': 0.00018837675350701403, 'max_depth': 30,
'min_samples_leaf': 5, 'min_samples_split': 20, 'random_state': 88}
```

[76]:
```
dtr_cv.best_score_
```

[76]:
```
0.06574800019913989
```

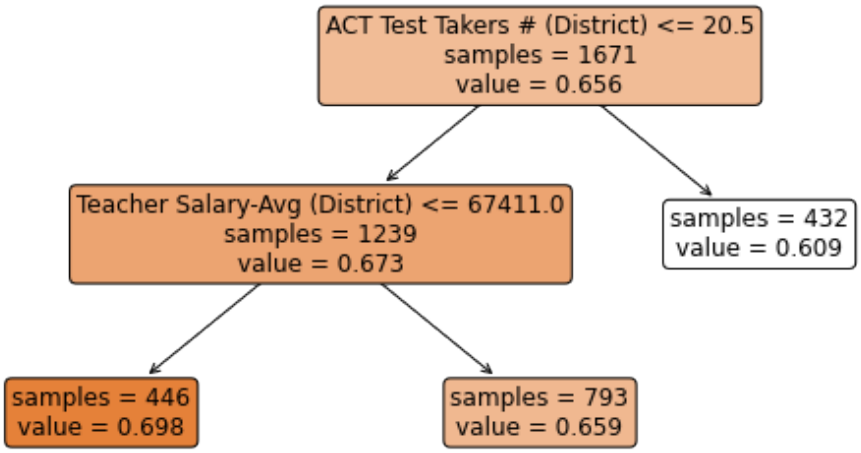
We observe a really low R2 for our optimal ccp_alpha of 0.000188. So, we will most likely not use our decision tree regressor in our blended model.

Below is a visualization of our decision tree, seeing which features are considered first when determining the UC Admissions rate for each district.

[77]:
```
ccp_dtr = dtr_cv.best_estimator_

print('Node count =', ccp_dtr.tree_.node_count)

plt.figure(figsize=(10,5))
plot_tree(ccp_dtr,
          feature_names=X_train.columns,
```

```
          class_names=['0','1'],
          filled=True,
          impurity=False,
          rounded=True,
          fontsize=12)
plt.show()
```

```
Node count = 5
```



1.12 Random Forest

[78]:
```
# Consider 1 to the total number of features (for max_features) while keeping␣
 ↪the rest of the parameters constant
grid_values2 = {'max_features': np.linspace(1, len(X_train.columns),␣
 ↪len(X_train.columns), dtype='int32'),
                'min_samples_leaf': [5],
                'n_estimators': [500],
                'random_state': [88]}

rf_cv = RandomForestRegressor()

cv = KFold(n_splits = 5, random_state = 88, shuffle = True) # 5 fold cross␣
 ↪validation for each max_features value

rf_cv = GridSearchCV(rf_cv, param_grid = grid_values2, scoring = 'r2', cv = cv,␣
 ↪verbose = 1)
```

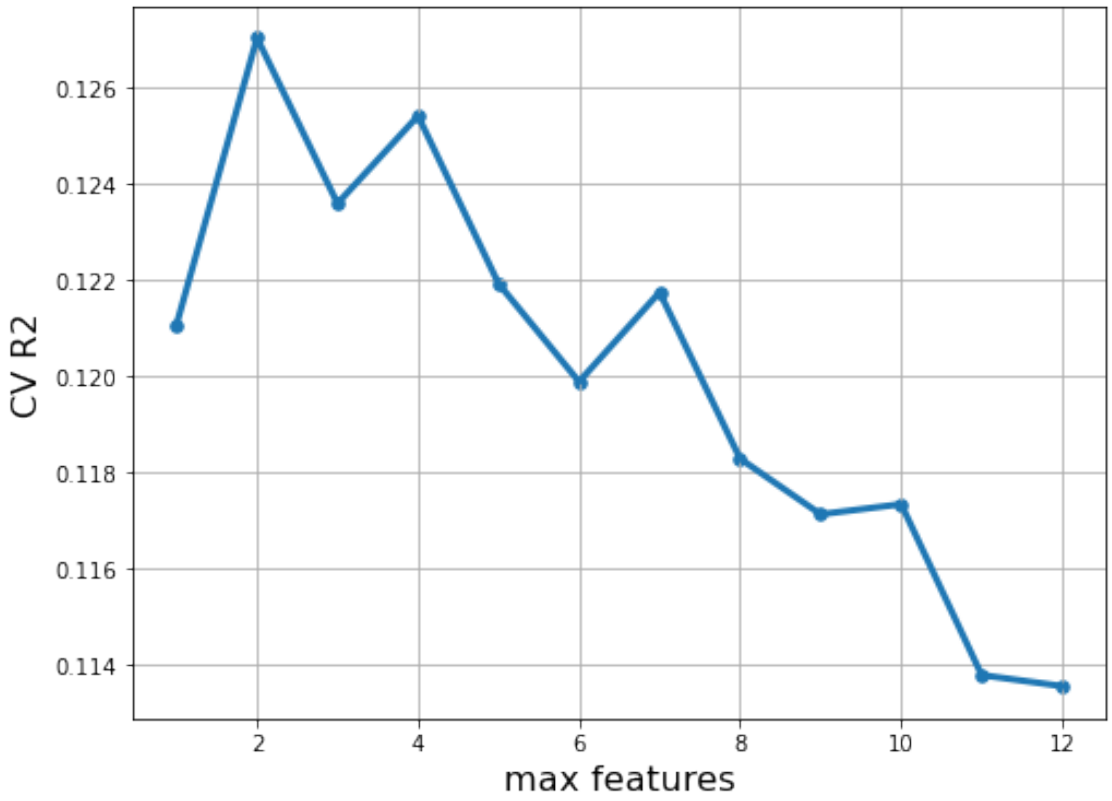
```
rf_cv.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  60 out of  60 | elapsed:  1.4min finished
```

[78]:
```
GridSearchCV(cv=KFold(n_splits=5, random_state=88, shuffle=True),
             estimator=RandomForestRegressor(),
             param_grid={'max_features': array([ 1,  2,  3,  4,  5,  6,  7,  8,
 9, 10, 11, 12], dtype=int32),
                         'min_samples_leaf': [5], 'n_estimators': [500],
                         'random_state': [88]},
             scoring='r2', verbose=1)
```

Below is the plot examining the R2 for each max_features values between 1 and 12. We will choose the max features parameter based on the highest CV R2.

[79]:
```
max_features = rf_cv.cv_results_['param_max_features'].data
r2_scores_rf = rf_cv.cv_results_['mean_test_score']

plt.figure(figsize=(8, 6))
plt.xlabel('max features', fontsize=16)
plt.ylabel('CV R2', fontsize=16)
plt.scatter(max_features, r2_scores_rf, s=30)
plt.plot(max_features, r2_scores_rf, linewidth=3)
plt.grid(True, which='both')

plt.show()
```



[80]:
```
print('Best max_features', rf_cv.best_params_)

rf = rf_cv.best_estimator_
```

```
Best max_features {'max_features': 2, 'min_samples_leaf': 5, 'n_estimators':
500, 'random_state': 88}
```

[81]:
```
rf_cv.best_score_
```

[81]:
```
0.1270296826029237
```

For our random forest model, we observe the highest R2 among our three models with our optimal max_features of 2. So, we will most likely use our random forest model alone as our final model.

Below is the importance scores of all features sorted in descending order. We see which features are most important in determining the UC Admissions Rate for each district in our random forest model.

[82]:
```
imp = pd.DataFrame({'Feature' : X_train.columns,
                    'Importance score': 100*rf.feature_importances_}).round(1)
imp = imp.sort_values("Importance score", ascending = False)
imp
```

[82]:

	Feature	Importance score
3	Teacher Salary-Avg (District)	11.5
1	Per Pupil Ratio: Teacher (District)	9.9
5	Current Exp of Educ per ADA (Ed Code 41372) (D…	9.8
0	ACT Test Takers # (District)	9.4
8	Gen Fund Exp by Activity - 4000-4999 Ancillary…	9.4
7	Total Gen Fund Revenues Per Student (District)	9.1
9	Gen Fund Exp by Activity - 1000-1999 Instructi…	9.0
11	Gen Fund Exp by Activity - 3000-3999 Pupil Ser…	8.6
10	Gen Fund Exp by Activity - 7000-7999 General A…	7.9
6	Gen Fund Exp by Object Code - 4000-4999 Books …	7.3
2	Avg Years Teaching (District)	5.2
4	Chronic Absenteeism % (District)	2.9

1.13 Time Series

We want to determine if we can incorporate a time series model to improve our final model, especially since our previous three models yielded very low R2 scores.

[83]:
```
# Make new column containing the starting year for each year range
sd_all_cleaned['Year_new'] = sd_all_cleaned['Year'].str[0:4]
sd_all_cleaned['Year_new'] = pd.to_datetime(sd_all_cleaned['Year_new'],␣
 ↪infer_datetime_format=True)
sd_all_cleaned
```

[83]:

	District Name	Count	Year	Student Count	\
0	ABC Unified	Adm	2010-2011	205	
1	ABC Unified	Adm	2011-2012	207	
2	ABC Unified	Adm	2012-2013	183	
3	ABC Unified	Adm	2013-2014	200	
4	ABC Unified	Adm	2014-2015	182	
…	…	…	…	…	
2229	Yuba City Unified	Adm	2014-2015	64	
2230	Yuba City Unified	Adm	2015-2016	62	
2231	Yuba City Unified	Adm	2016-2017	74	
2232	Yuba City Unified	Adm	2017-2018	56	
2233	Yuba City Unified	Adm	2018-2019	75	

	ACT Test Takers # (District)	Per Pupil Ratio: Teacher (District)	\
0	0.0	49.0	
1	0.0	49.6	
2	0.0	48.2	
3	0.0	49.2	
4	0.0	48.0	
…	…	…	
2229	0.0	40.0	

```
2230                          402.0                                    39.6
2231                          332.0                                    39.4
2232                          296.0                                    36.0
2233                          232.0                                    38.6

      Avg Years Teaching (District)  Teacher Salary-Avg (District)  \
0                              30.0                         143614.0
1                              30.0                         143774.0
2                              28.0                         143186.0
3                              28.0                         152394.0
4                              28.0                         156382.0
…                              …                             …
2229                           24.0                         128330.0
2230                           24.0                         132862.0
2231                           24.0                         137350.0
2232                           24.0                         146336.0
2233                           24.0                         151670.0

      Chronic Absenteeism % (District)  \
0                                  0.0
1                                  0.0
2                                  0.0
3                                  0.0
4                                  0.0
…                                  …
2229                               0.0
2230                               0.0
2231                              36.4
2232                              24.6
2233                              24.4

      Current Exp of Educ per ADA (Ed Code 41372) (District)  …  \
0                                                 16108.0       …
1                                                 16216.0       …
2                                                 16010.0       …
3                                                 16818.0       …
4                                                 17850.0       …
…                                                   …       …
2229                                              18236.0       …
2230                                              19472.0       …
2231                                              21636.0       …
2232                                              23652.0       …
2233                                              25974.0       …

      Total Gen Fund Revenues Per Student (District)  \
0                                           16586.0
1                                           16356.0
```

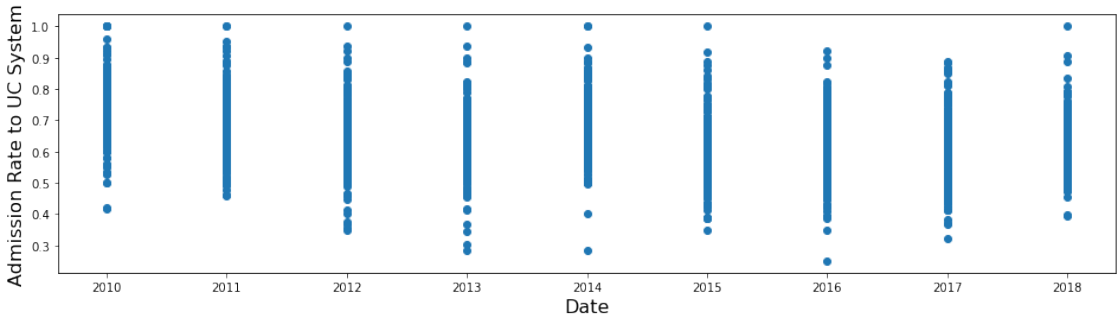
```
2                                                16480.0
3                                                17876.0
4                                                18804.0
…                                                  …
2229                                             18916.0
2230                                             22078.0
2231                                             23672.0
2232                                             23458.0
2233                                             27114.0

      Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
0                                                    0.0
1                                                    0.0
2                                                    0.0
3                                                    0.0
4                                                    0.0
…                                                    …
2229                                               152.0
2230                                               186.0
2231                                               274.0
2232                                               238.0
2233                                               266.0

      Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
0                                                10458.0
1                                                10592.0
2                                                10534.0
3                                                11010.0
4                                                11618.0
…                                                  …
2229                                             11812.0
2230                                             12314.0
2231                                             13470.0
2232                                             14540.0
2233                                             16390.0

      Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
0                                                  952.0
1                                                  954.0
2                                                  912.0
3                                                  938.0
4                                                  968.0
…                                                    …
2229                                               888.0
```

```
2230                                                  1086.0
2231                                                  1402.0
2232                                                  1342.0
2233                                                  1358.0

      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
0                                                     1116.0
1                                                     1142.0
2                                                     1102.0
3                                                     1194.0
4                                                     1320.0
…                                                      …
2229                                                  1520.0
2230                                                  1586.0
2231                                                  1764.0
2232                                                  2018.0
2233                                                  2228.0

        Adm    App    Enr  Admission Rate to UC System   Year_new
0     205.0  270.0  120.0                     0.759259 2010-01-01
1     207.0  281.0  131.0                     0.736655 2011-01-01
2     183.0  258.0  108.0                     0.709302 2012-01-01
3     200.0  275.0  105.0                     0.727273 2013-01-01
4     182.0  271.0   98.0                     0.671587 2014-01-01
…      …      …      …                          …          …
2229   64.0   94.0   43.0                     0.680851 2014-01-01
2230   62.0   98.0   31.0                     0.632653 2015-01-01
2231   74.0  111.0   40.0                     0.666667 2016-01-01
2232   56.0   87.0   27.0                     0.643678 2017-01-01
2233   75.0  114.0   41.0                     0.657895 2018-01-01

[2147 rows x 21 columns]
```

[84]:
```python
def plot_entire_series(x, y, **kwargs):

    plt.figure(figsize=(16, 4))
    plt.scatter(x,y, linewidth=1)
    #plt.plot(x, y, linewidth=1, color='black')

    for key, value in kwargs.items():
        plt.plot(x, value, linewidth=1, color="red")

    plt.xlabel('Date', fontsize=16)
    plt.ylabel('Admission Rate to UC System', fontsize=16)
    plt.show()
```

```
    return None
```

[85]:
```
# Make a time series plot to observe potential trend
x = sd_all_cleaned['Year_new'].to_numpy()
y = sd_all_cleaned['Admission Rate to UC System'].to_numpy()

plot_entire_series(x, y)
```



[86]:
```
time_period = list(range(len(sd_all_cleaned)))

series_lm = sd_all_cleaned.copy().drop(columns=['Year','Year_new'])
series_lm = series_lm.rename(columns = {'Admission Rate to UC System':␣
 ↪'AdmissionRateToUCSystem'}, inplace = False)
series_lm['TimePeriod'] = time_period
series_lm.tail()
```

[86]:

	District Name	Count	Student Count	ACT Test Takers # (District)	\
2229	Yuba City Unified	Adm	64	0.0	
2230	Yuba City Unified	Adm	62	402.0	
2231	Yuba City Unified	Adm	74	332.0	
2232	Yuba City Unified	Adm	56	296.0	
2233	Yuba City Unified	Adm	75	232.0	

	Per Pupil Ratio: Teacher (District)	Avg Years Teaching (District)	\
2229	40.0	24.0	
2230	39.6	24.0	
2231	39.4	24.0	
2232	36.0	24.0	
2233	38.6	24.0	

	Teacher Salary-Avg (District)	Chronic Absenteeism % (District)	\
2229	128330.0	0.0	
2230	132862.0	0.0	
2231	137350.0	36.4	
2232	146336.0	24.6	

```
2233                      151670.0                              24.4

     Current Exp of Educ per ADA (Ed Code 41372) (District)  \
2229                                              18236.0
2230                                              19472.0
2231                                              21636.0
2232                                              23652.0
2233                                              25974.0

     Gen Fund Exp by Object Code - 4000-4999 Books & Supplies Per Student
(District)  \
2229                                               1224.0
2230                                               1322.0
2231                                               1288.0
2232                                               1622.0
2233                                               1586.0

     Total Gen Fund Revenues Per Student (District)  \
2229                                            18916.0
2230                                            22078.0
2231                                            23672.0
2232                                            23458.0
2233                                            27114.0

     Gen Fund Exp by Activity - 4000-4999 Ancillary Services Per Student #
(District)  \
2229                                             152.0
2230                                             186.0
2231                                             274.0
2232                                             238.0
2233                                             266.0

     Gen Fund Exp by Activity - 1000-1999 Instruction Per Student # (District)
\
2229                                            11812.0
2230                                            12314.0
2231                                            13470.0
2232                                            14540.0
2233                                            16390.0

     Gen Fund Exp by Activity - 7000-7999 General Administration Per Student #
(District)  \
2229                                             888.0
2230                                            1086.0
2231                                            1402.0
2232                                            1342.0
2233                                            1358.0
```

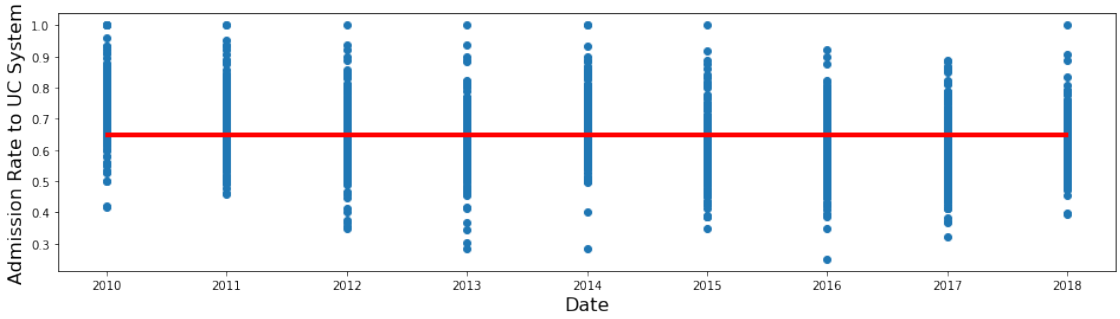
```
      Gen Fund Exp by Activity - 3000-3999 Pupil Services Per Student #
(District)  \
2229                                                 1520.0
2230                                                 1586.0
2231                                                 1764.0
2232                                                 2018.0
2233                                                 2228.0

       Adm    App   Enr  AdmissionRateToUCSystem  TimePeriod
2229  64.0   94.0  43.0                 0.680851        2142
2230  62.0   98.0  31.0                 0.632653        2143
2231  74.0  111.0  40.0                 0.666667        2144
2232  56.0   87.0  27.0                 0.643678        2145
2233  75.0  114.0  41.0                 0.657895        2146
```

[87]:
```
# Build a simple linear regressive time series model
import statsmodels.formula.api as smf

lm = smf.ols(formula='AdmissionRateToUCSystem ~ TimePeriod', data=series_lm).
 ↪fit()

print(lm.summary())
```

```
                                OLS Regression Results
==============================================================================
===
Dep. Variable:     AdmissionRateToUCSystem   R-squared:
0.001
Model:                                 OLS   Adj. R-squared:
0.000
Method:                      Least Squares   F-statistic:
1.693
Date:                     Wed, 12 May 2021   Prob (F-statistic):
0.193
Time:                             15:01:39   Log-Likelihood:
1797.2
No. Observations:                     2147   AIC:
-3590.
Df Residuals:                         2145   BIC:
-3579.
Df Model:                                1
Covariance Type:                 nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6527      0.005    144.323      0.000       0.644       0.662
```

```
TimePeriod -4.749e-06   3.65e-06     -1.301      0.193   -1.19e-05    2.41e-06
==============================================================================
Omnibus:                       40.061   Durbin-Watson:                   1.397
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               60.302
Skew:                           0.189   Prob(JB):                     8.05e-14
Kurtosis:                       3.729   Cond. No.                     2.48e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.48e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[88]:
```
# Draw linear regression line onto time series plot
x = sd_all_cleaned['Year_new'].to_numpy()
y = sd_all_cleaned['Admission Rate to UC System'].to_numpy()
y_pred = lm.predict(series_lm).to_numpy()

plot_entire_series(x, y, red=y_pred)
```



Despite the long time span, our data does not fit a time series trend, bearing a R-squared of 0.001. Therefore, we will not incorporate this into our blended or final model (only using random forest).

1.14 Blending

Although linear regression and the decision tree regressor yielded much lower R2 scores, there is still a possibility of an improved blended model with a higher score than random forest alone. This is because the two models might include insight not well encompassed in the random forest model.

[89]:
```
# Gather predictions of the three models on the validation set
X_blend_val = pd.DataFrame(data = {'val_pred_linreg': linreg_p.predict(sm.
 ↪add_constant(X_val[X_train_p.columns])),
                    'val_pred_dtr': ccp_dtr.predict(X_val),
                    'val_pred_rf': rf.predict(X_val)})
```

```
X_blend_val
```

[89]:

	val_pred_linreg	val_pred_dtr	val_pred_rf
7	0.633624	0.608834	0.589683
16	0.623523	0.698060	0.626044
25	0.657130	0.608834	0.635691
34	0.650184	0.608834	0.662579
43	0.615364	0.608834	0.635348
…	…	…	…
2198	0.644490	0.658796	0.663241
2205	0.662171	0.608834	0.617423
2214	0.646749	0.608834	0.600927
2223	0.635865	0.608834	0.594032
2232	0.633453	0.608834	0.593689

[240 rows x 3 columns]

[90]:
```
# Train linear regression of blended model using linear regression, decision␣
 ↪tree and random forest
blending_ols = sm.add_constant(X_blend_val)

blending_res = sm.OLS(y_val, blending_ols).fit()

print(blending_res.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.091
Model:                                     OLS   Adj. R-squared:
0.080
Method:                          Least Squares   F-statistic:
7.919
Date:                         Wed, 12 May 2021   Prob (F-statistic):
4.70e-05
Time:                                 15:01:39   Log-Likelihood:
227.74
No. Observations:                          240   AIC:
-447.5
Df Residuals:                              236   BIC:
-433.6
Df Model:                                    3
Covariance Type:                     nonrobust
================================================================================
===
                     coef    std err          t      P>|t|      [0.025
```

```
0.975]
------------------------------------------------------------------------------
---
const              -0.3251      0.238     -1.368      0.173      -0.793
0.143
val_pred_linreg     0.0691      0.225      0.307      0.759      -0.374
0.512
val_pred_dtr        0.4604      0.323      1.424      0.156      -0.177
1.097
val_pred_rf         0.9813      0.269      3.649      0.000       0.451
1.511
==============================================================================
Omnibus:                        0.849   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.654   Jarque-Bera (JB):                0.558
Skew:                           0.055   Prob(JB):                        0.757
Kurtosis:                       3.209   Cond. No.                         88.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Unfortunately, the blended model with all three models yields an R2 score *lower* than random forest alone. We want to remove the linear regression and decision tree predictions from our blended model because they show up as insignificant ($p > 0.05$).

[91]:
```
# Train linear regression of blended model using linear regression, decision
↪tree and random forest
X_blend_val2 = X_blend_val.drop(['val_pred_linreg', 'val_pred_dtr'], axis = 1)
blending_ols2 = sm.add_constant(X_blend_val2)

blending_res2 = sm.OLS(y_val, blending_ols2).fit()

print(blending_res2.summary())
```

```
                                OLS Regression Results
==============================================================================
=======
Dep. Variable:     Admission Rate to UC System   R-squared:
0.083
Model:                                     OLS   Adj. R-squared:
0.079
Method:                          Least Squares   F-statistic:
21.55
Date:                         Wed, 12 May 2021   Prob (F-statistic):
5.69e-06
Time:                                 15:01:39   Log-Likelihood:
226.63
```

```
No. Observations:                        240   AIC:
-449.3
Df Residuals:                            238   BIC:
-442.3
Df Model:                                  1
Covariance Type:                   nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -0.0954      0.151     -0.633      0.528      -0.392       0.202
val_pred_rf     1.1408      0.246      4.643      0.000       0.657       1.625
==============================================================================
Omnibus:                        2.333   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.311   Jarque-Bera (JB):                2.000
Skew:                           0.161   Prob(JB):                        0.368
Kurtosis:                       3.310   Cond. No.                         55.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Although the R2 dropped, this blended model is more reliable because the previously high R2 was due to insignificant models that were specific to the training set. This model should perform better on the test set.

[92]:
```
# Add the predictions of the blended model to the dataframe
val_pred_blended = blending_res.predict(sm.add_constant(X_blend_val))
X_blend_val['val_pred_blended']= val_pred_blended

X_blend_val
```

[92]:

	val_pred_linreg	val_pred_dtr	val_pred_rf	val_pred_blended
7	0.633624	0.608834	0.589683	0.577685
16	0.623523	0.698060	0.626044	0.653750
25	0.657130	0.608834	0.635691	0.624457
34	0.650184	0.608834	0.662579	0.650363
43	0.615364	0.608834	0.635348	0.621235
…	…	…	…	…
2198	0.644490	0.658796	0.663241	0.673623
2205	0.662171	0.608834	0.617423	0.606879
2214	0.646749	0.608834	0.600927	0.589625
2223	0.635865	0.608834	0.594032	0.582107
2232	0.633453	0.608834	0.593689	0.581604

[240 rows x 4 columns]

1.15 Test Performance

The blended model might perform differently on the test set.

[93]:
```
# Get predictions for linear regress, decision tree, random forest on the test␣
↪set
X_blend_test = pd.DataFrame(data = {'val_pred_linreg': linreg_p.predict(sm.
↪add_constant(X_test[X_train_p.columns])),
                    'val_pred_dtr': ccp_dtr.predict(X_test),
                    'val_pred_rf': rf.predict(X_test)})
X_blend_test
```

[93]:

	val_pred_linreg	val_pred_dtr	val_pred_rf
8	0.630074	0.608834	0.582759
26	0.654605	0.608834	0.616496
35	0.653596	0.608834	0.625848
44	0.612098	0.608834	0.625592
53	0.654471	0.608834	0.587694
…	…	…	…
2199	0.642684	0.658796	0.653733
2206	0.662912	0.608834	0.630168
2215	0.650959	0.608834	0.587809
2224	0.636935	0.608834	0.590910
2233	0.634825	0.608834	0.599742

[236 rows x 3 columns]

[94]:
```
print('Blended Model MAE:', MAE(blending_res, sm.add_constant(X_blend_test),␣
↪y_test), '\n')
print('Blended Model RMSE:', RMSE(blending_res, sm.add_constant(X_blend_test),␣
↪y_test), '\n')
print('Blended Model OSR2:', OSR2(blending_res, sm.add_constant(X_blend_test),␣
↪y_test, y_train), '\n')
```

```
Blended Model MAE: 0.06686308054211872

Blended Model RMSE: 0.08688349902376523

Blended Model OSR2: 0.02505428589669001
```

As expected, the first blended model performed worse than the random forest model alone on the test set. We will now use our final blended model with only random forest.

[95]:
```
# Get predictions for random forest on the test set
X_blend_test2 = pd.DataFrame(data = {'val_pred_rf': rf.predict(X_test)})
X_blend_test2
```

[95]:
```
     val_pred_rf
0       0.582759
1       0.616496
2       0.625848
3       0.625592
4       0.587694
..           …
231     0.653733
232     0.630168
233     0.587809
234     0.590910
235     0.599742

[236 rows x 1 columns]
```

[96]:
```
print('Baseline MAE:', np.mean(np.abs(y_test - np.mean(y_train))))
print('Linear Regression MAE:', MAE(linreg_p, sm.add_constant(X_test[X_train_p.
 ↪columns]), y_test))
print('Regression Tree MAE:', MAE(ccp_dtr, X_test, y_test))
print('Random Forest MAE:', MAE(rf, X_test, y_test))
print('Blended Model MAE:', MAE(blending_res2, sm.add_constant(X_blend_test2),␣
 ↪y_test), '\n')

print('Baseline RMSE:', np.sqrt(np.mean((y_test - np.mean(y_train))**2)))
print('Linear Regression RMSE:', RMSE(linreg_p, sm.
 ↪add_constant(X_test[X_train_p.columns]), y_test))
print('Regression Tree RMSE:', RMSE(ccp_dtr, X_test, y_test))
print('Random Forest RMSE:', RMSE(rf, X_test, y_test))
print('Blended Model RMSE:', RMSE(blending_res2, sm.
 ↪add_constant(X_blend_test2), y_test), '\n')

print('Linear Regression OSR2:', OSR2(linreg_p, sm.
 ↪add_constant(X_test[X_train_p.columns]), y_test, y_train))
print('Regression Tree OSR2:', OSR2(ccp_dtr, X_test, y_test, y_train))
print('Random Forest OSR2:', OSR2(rf, X_test, y_test, y_train))
print('Blended Model OSR2:', OSR2(blending_res2, sm.
 ↪add_constant(X_blend_test2), y_test, y_train), '\n')
```

```
Baseline MAE: 0.06867056140947875
Linear Regression MAE: 0.06798468632447917
Regression Tree MAE: 0.06400386900143737
Random Forest MAE: 0.06406630206852267
Blended Model MAE: 0.09098084993931609

Baseline RMSE: 0.08799278946978878
Linear Regression RMSE: 0.08782124484865726
Regression Tree RMSE: 0.08423867666145476
```

```
Random Forest RMSE: 0.08426492808290466
Blended Model RMSE: 0.10753554490504622

Linear Regression OSR2: 0.0038952602010740023
Regression Tree OSR2: 0.08350753153017043
Random Forest OSR2: 0.08293622676125567
Blended Model OSR2: 0.8417885485436253
```

Our final blended model has a high OSR2 of 0.84.