Investigating Differences Between Schools With Different Male to Female Ratios

Author: Qi Wang

In this notebook we will compare the different polarity trends between schools of different male to female ratios.

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
# Read in College Data CSV
college_data = pd.read_csv('college-records.csv')
```

In [3]:
```
college_data.head()
```

Out[3]:

	Name	Public / Private	Annual Tuition (2023)	Undergraduate Population (2023)	Acceptance rate (2021)	Ranking (top 100 or not)	Geographic location	Male : Female (ratio)	Type	...	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
0	us-west	Stanford	Private	$57,692	7,761	4.34%	6	Suburban	49 : 51	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	UC Berkeley	Public	$43,980	32,143	17.50%	49	Urban	46 : 54	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	Santa Clara	Private	$58,017	5,895	52.00%	NaN	Suburban	53 : 47	Jesuit	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	Caltech	Private	$58,479	901	6.70%	2	Suburban	65 : 35	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	USC	Private	$63,468	20,790	21.10%	43	Urban	46 : 54	Research	...	-1.443932	-2.522296	-4.243743	-5.392547	-7.123969	-3.919732	-2.692837	-5.143249	-2.996033	-0.472082

5 rows × 29 columns

Peforming Linear Regression

Here, we will use the Linear Regression model from sci-kit learn to give us the slope of the polarity across years.

In [9]:
```
# Initialize Linear Regression Model
from sklearn.linear_model import LinearRegression

model = LinearRegression()
```

In [35]:
```
slopes = []

for _, row in college_data.iterrows():
    y = []
    for year in range(2013, 2023):
        if (not pd.isna(row[str(year)])):
            y.append(row[str(year)])

    if len(y) == 0:
        slopes.append(np.nan)
        continue

    X = np.arange(1, len(y) + 1, 1).reshape(-1,1)
    Y = np.array(y)

    model.fit(X, Y)
#     print(model.coef_)
    slopes.append(model.coef_[0])

college_data['slopes'] = slopes
```

In [36]:
```
college_data.head(10)
```

Out[36]:

	Name	Public / Private	Annual Tuition (2023)	Undergraduate Population (2023)	Acceptance rate (2021)	Ranking (top 100 or not)	Geographic location	Male : Female (ratio)	Type	...	2014	2015	2016	2017	2018	2019	2020	2021	2022	slopes
0	us-west	Stanford	Private	$57,692	7,761	4.34%	6	Suburban	49 : 51	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	UC Berkeley	Public	$43,980	32,143	17.50%	49	Urban	46 : 54	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	Santa Clara	Private	$58,017	5,895	52.00%	NaN	Suburban	53 : 47	Jesuit	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	Caltech	Private	$58,479	901	6.70%	2	Suburban	65 : 35	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	USC	Private	$63,468	20,790	21.10%	43	Urban	46 : 54	Research	...	-2.522296	-4.243743	-5.392547	-7.123969	-3.919732	-2.692837	-5.143249	-2.996033	-0.472082	0.074160
5	NaN	University of Washington	Public	$40,740	30,856	53.50%	NaN	Urban	45 : 55	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	Harvey Mudd College	Private	$62,516	905	10%	28	Suburban	50 : 50	Science & Engineering	...	NaN	NaN	NaN	-0.075550	NaN	NaN	0.761180	0.151384	0.425327	0.089283
7	NaN	Pomona College	Private	$59,238	1,764	6.60%	16	Suburban	45 : 55	Liberal Arts	...	-2.709579	-4.401501	-5.927530	-8.535499	-8.483771	-5.880843	-7.086419	-6.771482	NaN	-0.524290
8	NaN	UCLA	Public	$13,804	31,600	10.80%	35	Urban	44 : 56	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	NaN	BYU	Private	$6,304	31,633	59.20%	NaN	Suburban	50 : 50	Research	...	2.906494	1.335666	0.280031	2.320210	1.889712	2.347615	1.088530	0.601041	2.321979	-0.171053

10 rows × 30 columns

Now we will create a new column with fractional values for the Male to Female ratios

In [40]:
```
ratios = []

for _, row in college_data.iterrows():
    rat = row['Male : Female (ratio)'].split(' : ')

    ratios.append(int(rat[0]) / int(rat[1]))

college_data['Gender Ratio'] = ratios
```

In [41]:
```
college_data.head(10)
```

Out[41]:

	Name	Public / Private	Annual Tuition (2023)	Undergraduate Population (2023)	Acceptance rate (2021)	Ranking (top 100 or not)	Geographic location	Male : Female (ratio)	Type	...	2015	2016	2017	2018	2019	2020	2021	2022	slopes	Gender Ratio
0	us-west	Stanford	Private	$57,692	7,761	4.34%	6	Suburban	49 : 51	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.960784
1	NaN	UC Berkeley	Public	$43,980	32,143	17.50%	49	Urban	46 : 54	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.851852
2	NaN	Santa Clara	Private	$58,017	5,895	52.00%	NaN	Suburban	53 : 47	Jesuit	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.127660
3	NaN	Caltech	Private	$58,479	901	6.70%	2	Suburban	65 : 35	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.857143
4	NaN	USC	Private	$63,468	20,790	21.10%	43	Urban	46 : 54	Research	...	-4.243743	-5.392547	-7.123969	-3.919732	-2.692837	-5.143249	-2.996033	-0.472082	0.074160	0.851852
5	NaN	University of Washington	Public	$40,740	30,856	53.50%	NaN	Urban	45 : 55	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.818182
6	NaN	Harvey Mudd College	Private	$62,516	905	10%	28	Suburban	50 : 50	Science & Engineering	...	NaN	NaN	-0.075550	NaN	NaN	0.761180	0.151384	0.425327	0.089283	1.000000
7	NaN	Pomona College	Private	$59,238	1,764	6.60%	16	Suburban	45 : 55	Liberal Arts	...	-4.401501	-5.927530	-8.535499	-8.483771	-5.880843	-7.086419	-6.771482	NaN	-0.524290	0.818182
8	NaN	UCLA	Public	$13,804	31,600	10.80%	35	Urban	44 : 56	Research	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.785714
9	NaN	BYU	Private	$6,304	31,633	59.20%	NaN	Suburban	50 : 50	Research	...	1.335666	0.280031	2.320210	1.889712	2.347615	1.088530	0.601041	2.321979	-0.171053	1.000000

10 rows × 31 columns

Plotting

We will plot the avaliable polarities with the different gender ratios

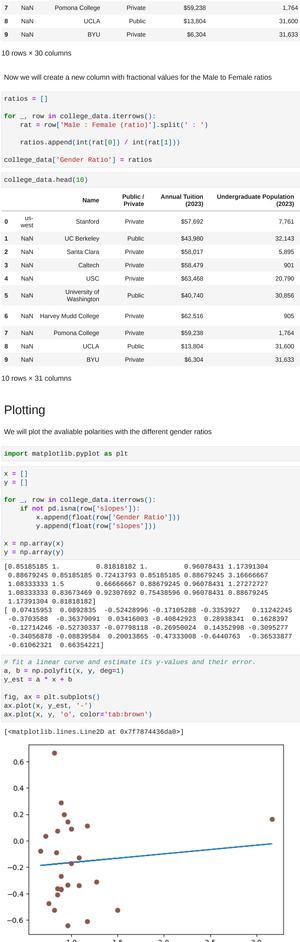
In [42]:
```
import matplotlib.pyplot as plt
```

In [61]:
```
x = []
y = []

for _, row in college_data.iterrows():
    if not pd.isna(row['slopes']):
        x.append(float(row['Gender Ratio']))
        y.append(float(row['slopes']))

x = np.array(x)
y = np.array(y)
```

```
[0.85185185 1.         0.81818182 1.         0.96078431 1.17391304
 0.88679245 0.85185185 0.72413793 0.85185185 0.88679245 3.16666667
 1.08333333 1.5        0.66666667 0.88679245 0.96078431 1.27272727
 1.08333333 0.83673469 0.92307692 0.75438596 0.96078431 0.88679245
 1.17391304 0.81818182]
[ 0.07415953  0.0892835  -0.52428996 -0.17105288 -0.3353927   0.11242245
 -0.3703588  -0.36379091  0.03416003 -0.40842923  0.28938341  0.1628397
 -0.12714246 -0.52730337 -0.07798118 -0.26950024  0.14352998 -0.3095277
 -0.34656878 -0.08839584  0.20013865 -0.47333008 -0.6440763  -0.36533877
 -0.61062321  0.66354221]
```

In [56]:
```
# fit a linear curve and estimate its y-values and their error.
a, b = np.polyfit(x, y, deg=1)
y_est = a * x + b

fig, ax = plt.subplots()
ax.plot(x, y_est, '-')
ax.plot(x, y, 'o', color='tab:brown')
```

Out[56]:
```
[<matplotlib.lines.Line2D at 0x7f7874436da0>]
```

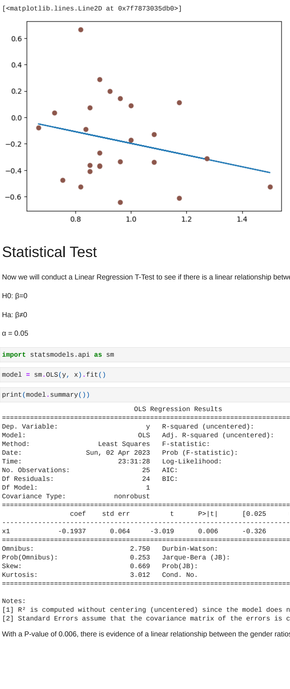


Let's remove the outlier

In [62]:
```
x = []
y = []

for _, row in college_data.iterrows():
    if not pd.isna(row['slopes']) and row['Gender Ratio'] < 2:
        x.append(float(row['Gender Ratio']))
        y.append(float(row['slopes']))

x = np.array(x)
y = np.array(y)
```

In [63]:
```
# fit a linear curve and estimate its y-values and their error.
a, b = np.polyfit(x, y, deg=1)
y_est = a * x + b

fig, ax = plt.subplots()
ax.plot(x, y_est, '-')
ax.plot(x, y, 'o', color='tab:brown')
```

Out[63]:
```
[<matplotlib.lines.Line2D at 0x7f7873035db0>]
```



Statistical Test

Now we will conduct a Linear Regression T-Test to see if there is a linear relationship between the gender ratio and the polarity trend across years.

H0: $\beta=0$

Ha: $\beta\neq0$

$\alpha = 0.05$

In [64]:
```
import statsmodels.api as sm
```

In [65]:
```
model = sm.OLS(y, x).fit()
```

In [66]:
```
print(model.summary())
```

```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.275
Model:                            OLS   Adj. R-squared (uncentered):              0.245
Method:                 Least Squares   F-statistic:                              9.112
Date:                Sun, 02 Apr 2023   Prob (F-statistic):                     0.00594
Time:                        23:31:28   Log-Likelihood:                         -5.7642
No. Observations:                  25   AIC:                                      13.53
Df Residuals:                      24   BIC:                                      14.75
Df Model:                           1                                                  
Covariance Type:            nonrobust                                                  
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.1937      0.064     -3.019      0.006      -0.326      -0.061
==============================================================================
Omnibus:                        2.750   Durbin-Watson:                   2.087
Prob(Omnibus):                  0.253   Jarque-Bera (JB):                1.867
Skew:                           0.669   Prob(JB):                        0.393
Kurtosis:                       3.012   Cond. No.                         1.00
==============================================================================


Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

With a P-value of 0.006, there is evidence of a linear relationship between the gender ratios and polarity slopes