# Income Segregation and Intergenerational Mobility Across Colleges in the United States

Jacob Sharadin

# Paper Background

Main idea: Investigating how higher education shapes mobility in the US.

How could changes in the distributions of students from different backgrounds in college affect segregation by parental income and intergenerational mobility?

# Paper Background

Paper Outline - Three Main Sections:

1. Constructing statistics on parent income at different colleges
2. Earning outcomes of students at each college
3. Changes in income segregation if students were allocated to colleges evenly

# Paper Findings

Four Key Findings:

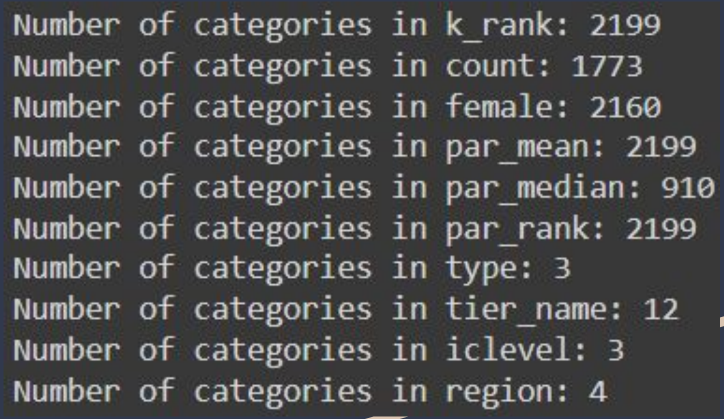
1. Low and middle income students attend selective schools at much lower rates than students from higher income
2. Very low middle class representation at most selective schools (Ivy-Plus)
3. Would need to raise attendance rates for low income students from ~7% to ~26%
    a. Low-income students would need to attend all schools at rates similar to those with 160 point higher SAT scores
4. By equalizing attendance rates for students with the same test scores, outcome gap would be decreased by 15%

Main takeaway: By changing how students are allocated to colleges, segregation could be decreased and intergenerational mobility increased

# The Dataset

- College attendance: federal tax records, Department of Education records 1999-2013
- Incomes: federal income tax 1996-2014, information returns (like W-2)
- Parent Income: total pre-tax income at the household level
  - Averaged over the five years when child aged 15-19
  - Parents then assigned income percentiles through ranking with other parents w/ children in same birth cohort
- Child Income: pre-tax individual earnings 2014
  - Ranked relative to other children in same birth cohort
- ~2200 observations

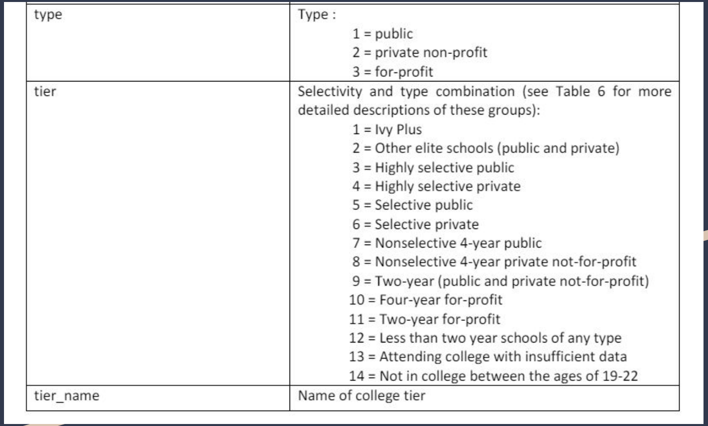
# Predicting k_rank



```
Number of categories in k_rank: 2199
Number of categories in count: 1773
Number of categories in female: 2160
Number of categories in par_mean: 2199
Number of categories in par_median: 910
Number of categories in par_rank: 2199
Number of categories in type: 3
Number of categories in tier_name: 12
Number of categories in iclevel: 3
Number of categories in region: 4
```

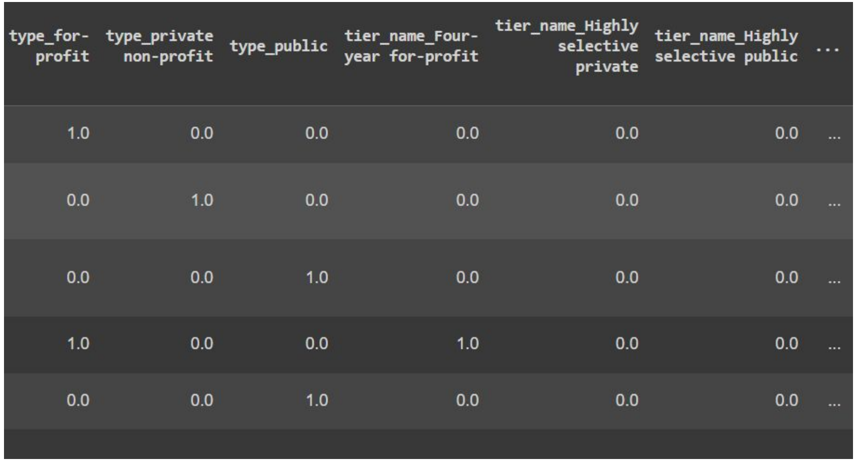
- Numerical variables
  - k_rank (Dependent Variable)
  - count
  - female
  - par_mean
  - par_median
  - par_rank
- Categorical variables
  - type
  - tier_name
  - iclevel
  - region

# Handling Categorical Variables



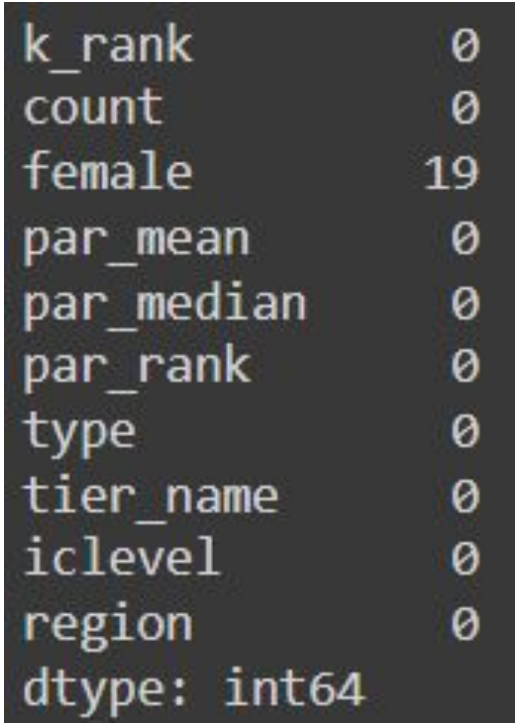
| | |
|---|---|
| type | Type :<br>1 = public<br>2 = private non-profit<br>3 = for-profit |
| tier | Selectivity and type combination (see Table 6 for more detailed descriptions of these groups):<br>1 = Ivy Plus<br>2 = Other elite schools (public and private)<br>3 = Highly selective public<br>4 = Highly selective private<br>5 = Selective public<br>6 = Selective private<br>7 = Nonselective 4-year public<br>8 = Nonselective 4-year private not-for-profit<br>9 = Two-year (public and private not-for-profit)<br>10 = Four-year for-profit<br>11 = Two-year for-profit<br>12 = Less than two year schools of any type<br>13 = Attending college with insufficient data<br>14 = Not in college between the ages of 19-22 |
| tier_name | Name of college tier |

One-hot encoding



| type_for-profit | type_private non-profit | type_public | tier_name_Four-year for-profit | tier_name_Highly selective private | tier_name_Highly selective public | ... |
|---|---|---|---|---|---|---|
| 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... |
| 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... |
| 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | ... |
| 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | ... |
| 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | ... |

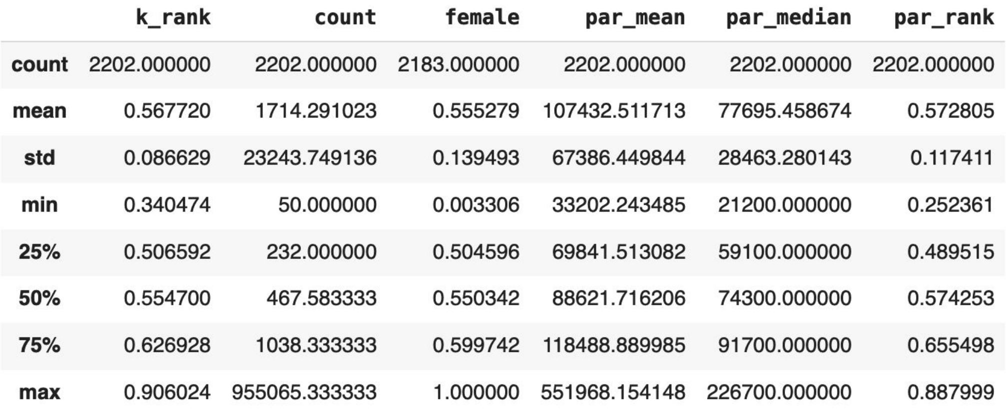
# Data Cleaning



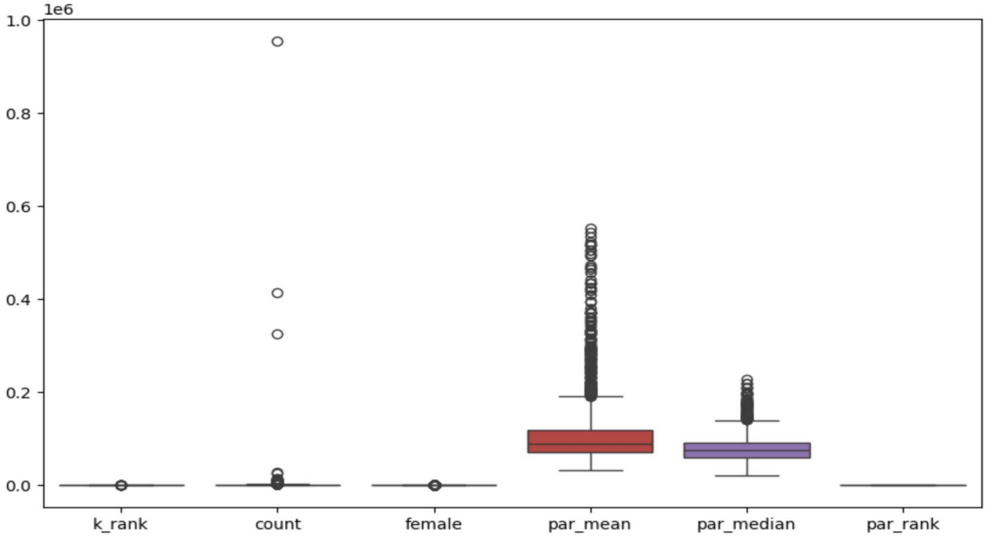
```
k_rank         0
count          0
female        19
par_mean       0
par_median     0
par_rank       0
type           0
tier_name      0
iclevel        0
region         0
dtype: int64
```

- Only 19 missing values

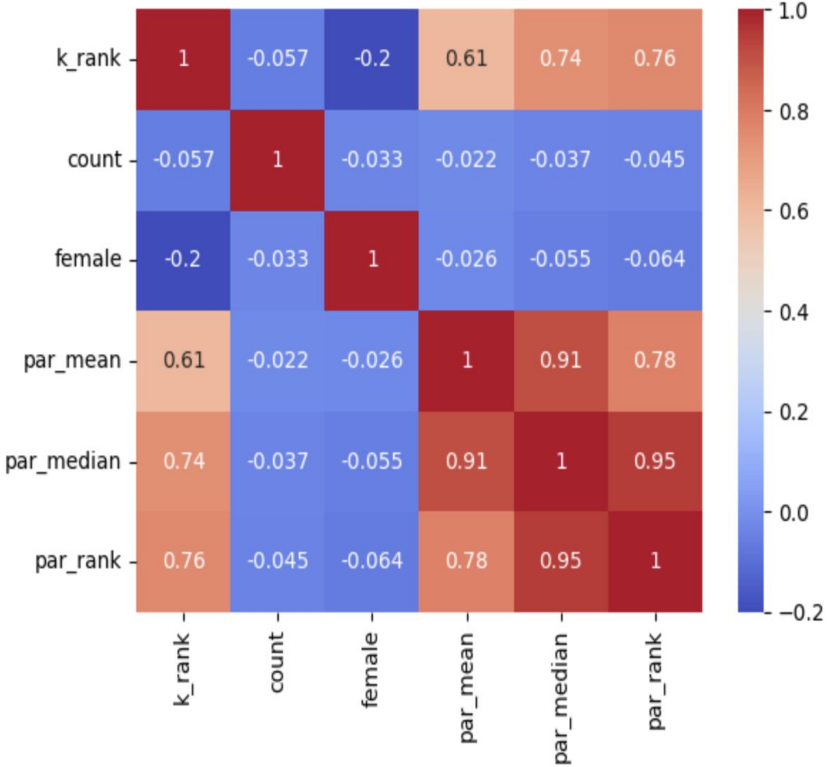
# Summary Statistics and Boxplots



| | k_rank | count | female | par_mean | par_median | par_rank |
|---|---|---|---|---|---|---|
| **count** | 2202.000000 | 2202.000000 | 2183.000000 | 2202.000000 | 2202.000000 | 2202.000000 |
| **mean** | 0.567720 | 1714.291023 | 0.555279 | 107432.511713 | 77695.458674 | 0.572805 |
| **std** | 0.086629 | 23243.749136 | 0.139493 | 67386.449844 | 28463.280143 | 0.117411 |
| **min** | 0.340474 | 50.000000 | 0.003306 | 33202.243485 | 21200.000000 | 0.252361 |
| **25%** | 0.506592 | 232.000000 | 0.504596 | 69841.513082 | 59100.000000 | 0.489515 |
| **50%** | 0.554700 | 467.583333 | 0.550342 | 88621.716206 | 74300.000000 | 0.574253 |
| **75%** | 0.626928 | 1038.333333 | 0.599742 | 118488.889985 | 91700.000000 | 0.655498 |
| **max** | 0.906024 | 955065.333333 | 1.000000 | 551968.154148 | 226700.000000 | 0.887999 |



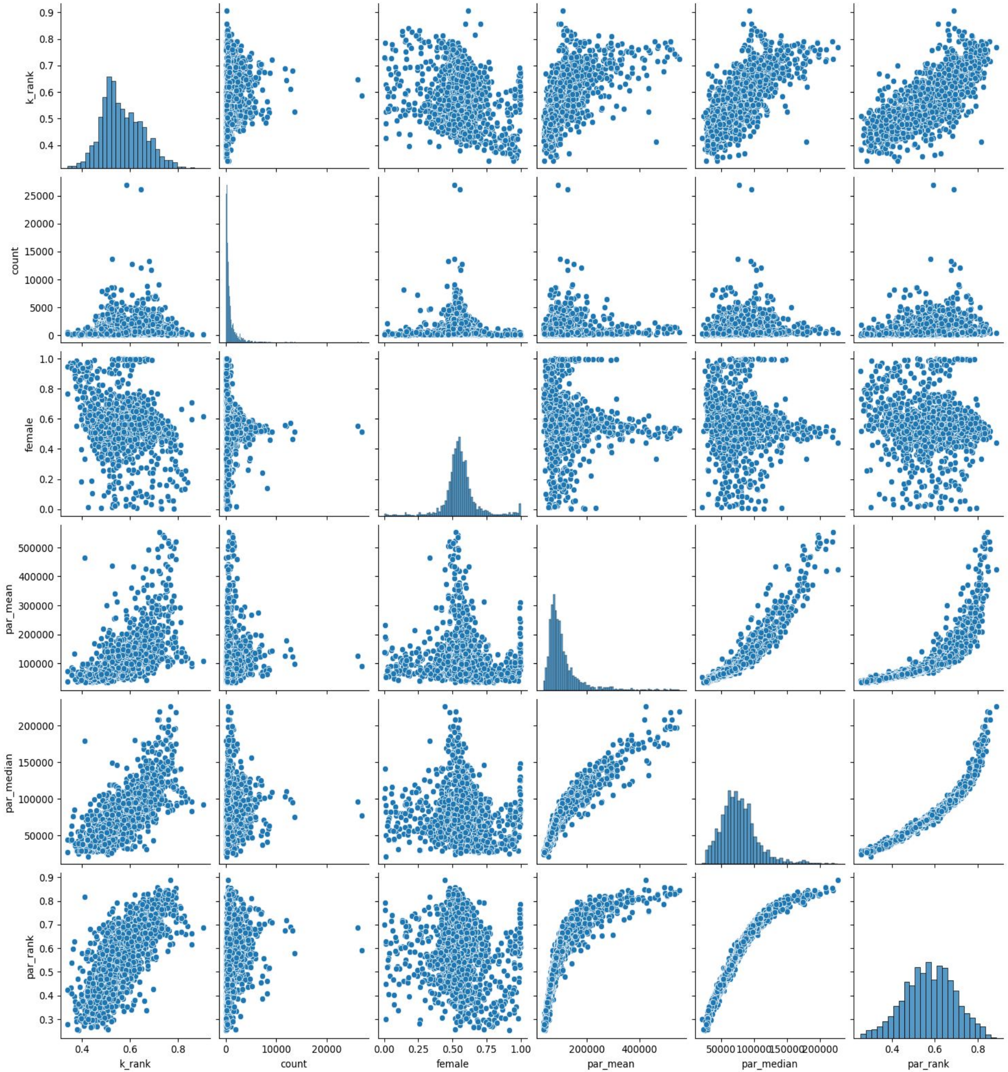
# Identifying Highly Correlated Variables



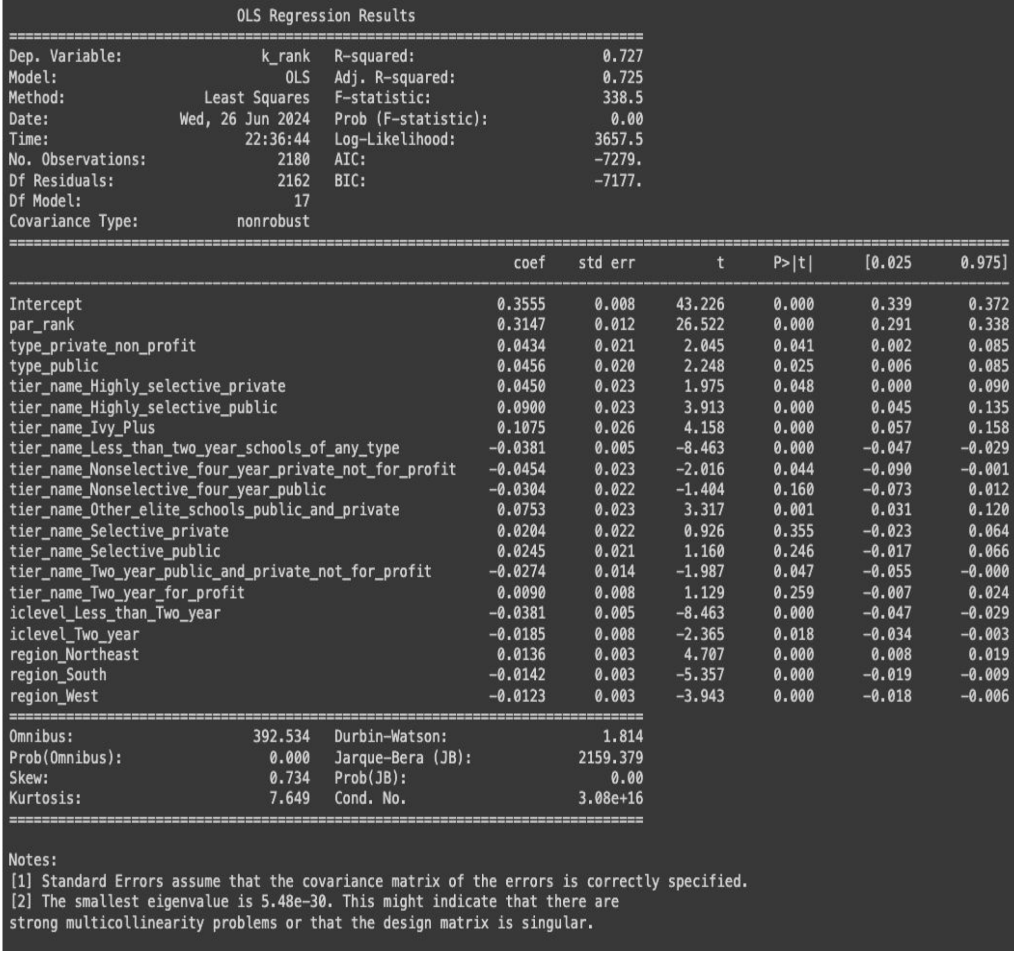
|  | k_rank | count | female | par_mean | par_median | par_rank |
|---|---|---|---|---|---|---|
| k_rank | 1 | -0.057 | -0.2 | 0.61 | 0.74 | 0.76 |
| count | -0.057 | 1 | -0.033 | -0.022 | -0.037 | -0.045 |
| female | -0.2 | -0.033 | 1 | -0.026 | -0.055 | -0.064 |
| par_mean | 0.61 | -0.022 | -0.026 | 1 | 0.91 | 0.78 |
| par_median | 0.74 | -0.037 | -0.055 | 0.91 | 1 | 0.95 |
| par_rank | 0.76 | -0.045 | -0.064 | 0.78 | 0.95 | 1 |

- Multicollinearity between variables

# Pairplot Visualization



# Regression Analysis

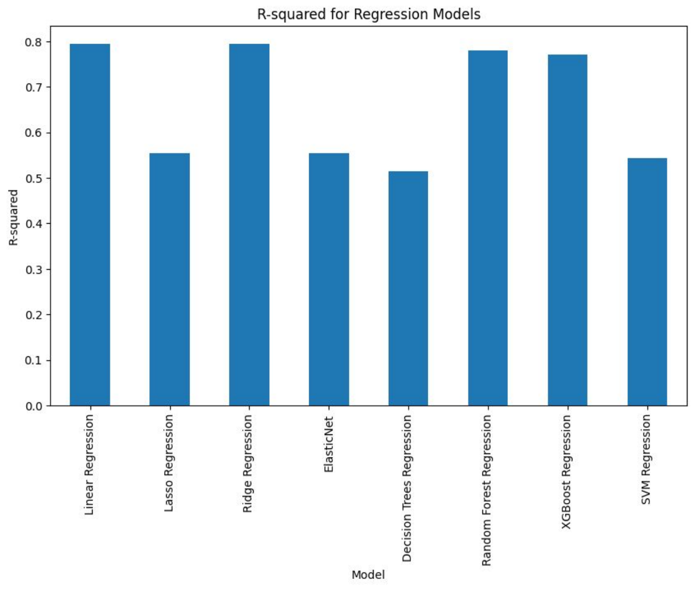


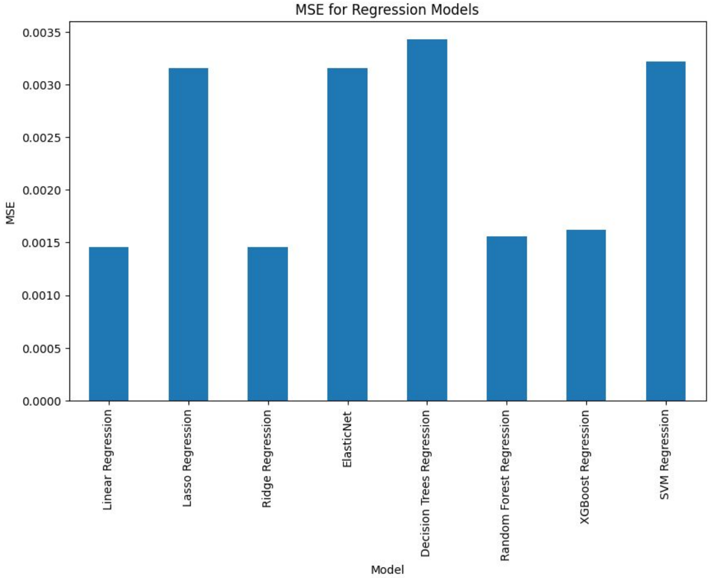
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 k_rank   R-squared:                       0.727
Model:                            OLS   Adj. R-squared:                  0.725
Method:                 Least Squares   F-statistic:                     338.5
Date:                Wed, 26 Jun 2024   Prob (F-statistic):               0.00
Time:                        22:36:44   Log-Likelihood:                 3657.5
No. Observations:                2180   AIC:                            -7279.
Df Residuals:                    2162   BIC:                            -7177.
Df Model:                          17
Covariance Type:            nonrobust
==================================================================================================================
                                                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
Intercept                                              0.3555      0.008     43.226      0.000       0.339       0.372
par_rank                                               0.3147      0.012     26.522      0.000       0.291       0.338
type_private_non_profit                                0.0434      0.021      2.045      0.041       0.002       0.085
type_public                                            0.0456      0.020      2.248      0.025       0.006       0.085
tier_name_Highly_selective_private                     0.0450      0.023      1.975      0.048       0.000       0.090
tier_name_Highly_selective_public                      0.0900      0.023      3.913      0.000       0.045       0.135
tier_name_Ivy_Plus                                     0.1075      0.026      4.158      0.000       0.057       0.158
tier_name_Less_than_two_year_schools_of_any_type      -0.0381      0.005     -8.463      0.000      -0.047      -0.029
tier_name_Nonselective_four_year_private_not_for_profit -0.0454   0.023     -2.016      0.044      -0.090      -0.001
tier_name_Nonselective_four_year_public               -0.0304      0.022     -1.404      0.160      -0.073       0.012
tier_name_Other_elite_schools_public_and_private       0.0753      0.023      3.317      0.001       0.031       0.120
tier_name_Selective_private                            0.0204      0.022      0.926      0.355      -0.023       0.064
tier_name_Selective_public                             0.0245      0.021      1.160      0.246      -0.017       0.066
tier_name_Two_year_public_and_private_not_for_profit  -0.0274      0.014     -1.987      0.047      -0.055      -0.000
tier_name_Two_year_for_profit                          0.0090      0.008      1.129      0.259      -0.007       0.024
iclevel_Less_than_Two_year                            -0.0381      0.005     -8.463      0.000      -0.047      -0.029
iclevel_Two_year                                      -0.0185      0.008     -2.365      0.018      -0.034      -0.003
region_Northeast                                       0.0136      0.003      4.707      0.000       0.008       0.019
region_South                                          -0.0142      0.003     -5.357      0.000      -0.019      -0.009
region_West                                           -0.0123      0.003     -3.943      0.000      -0.018      -0.006
==============================================================================
Omnibus:                      392.534   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2159.379
Skew:                           0.734   Prob(JB):                         0.00
Kurtosis:                       7.649   Cond. No.                     3.08e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 5.48e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

- This result suggests that this model is not valid

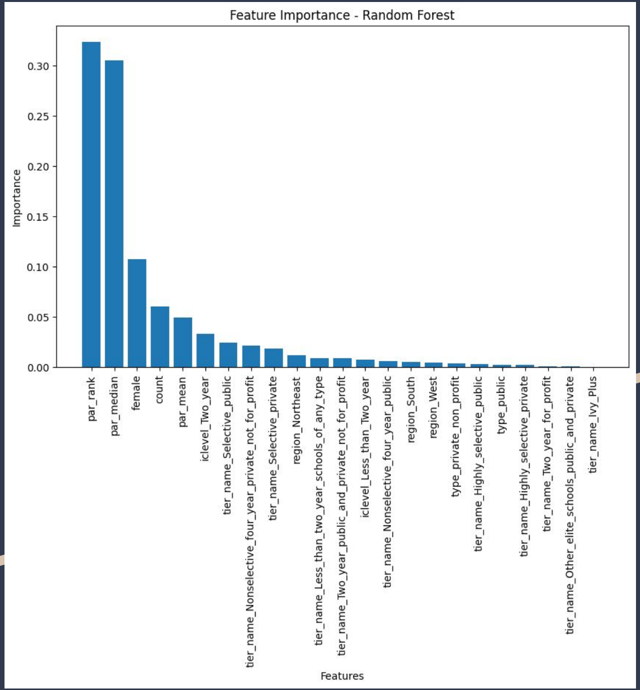
# Prediction with the use of ML algorithms

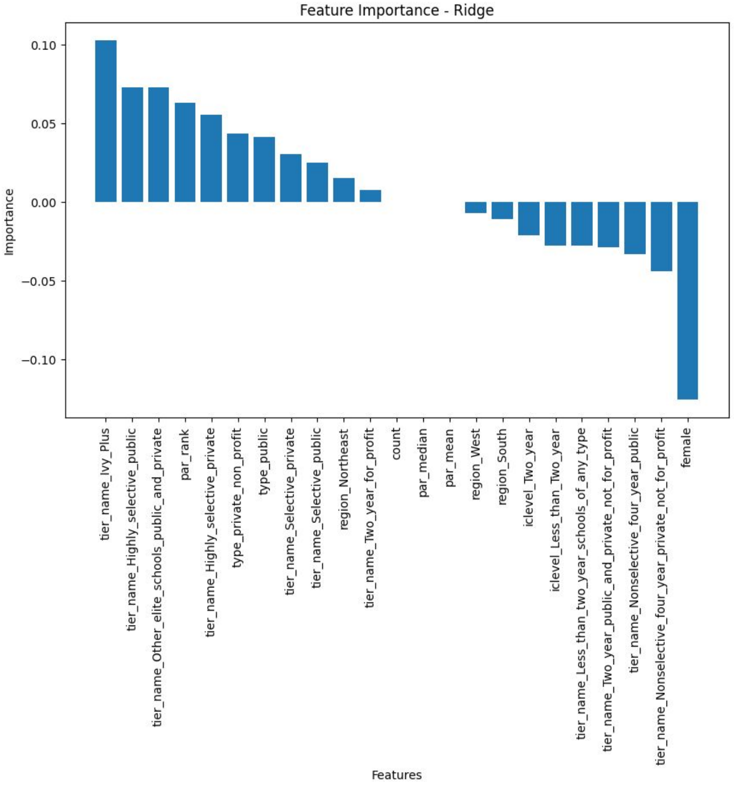
Model #1: Include all features to get initial results





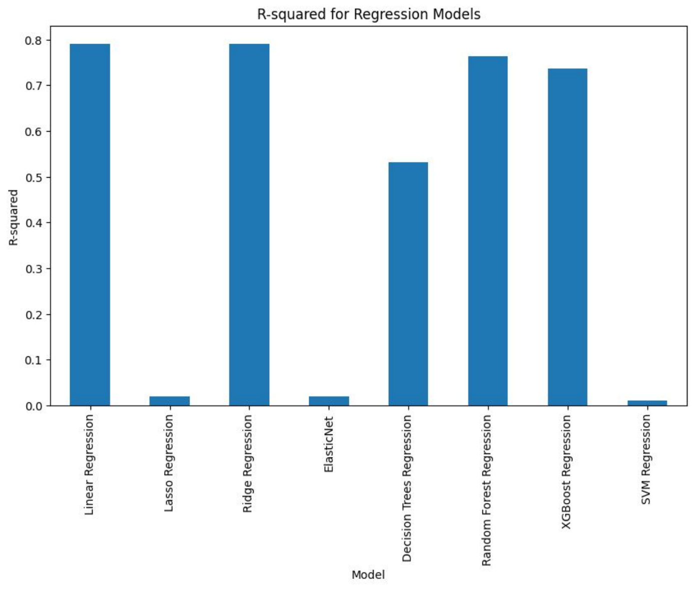
# Model #1 Feature Importances

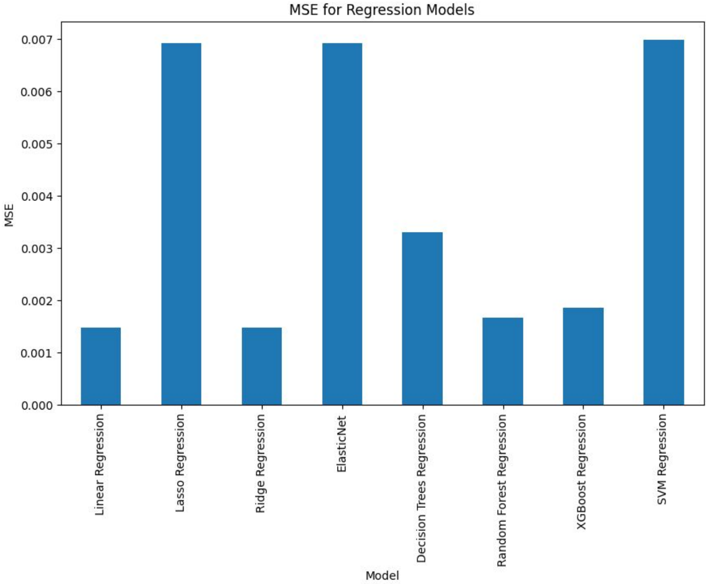




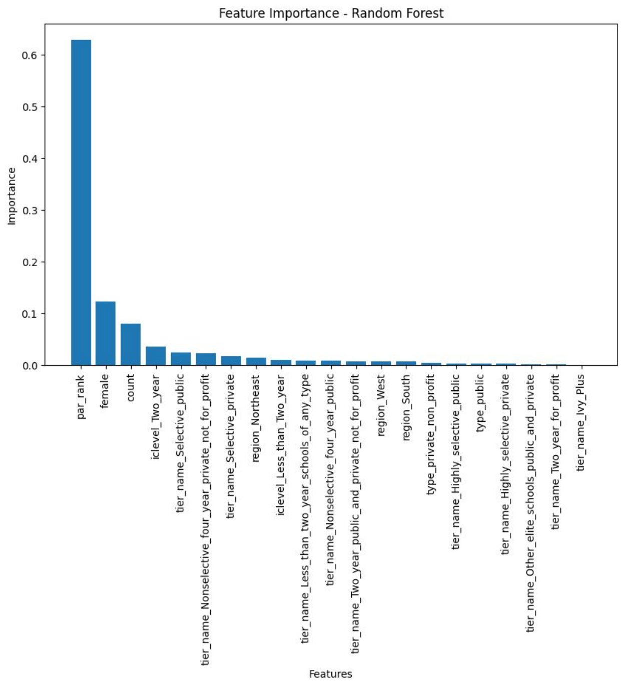
# Prediction Model #2

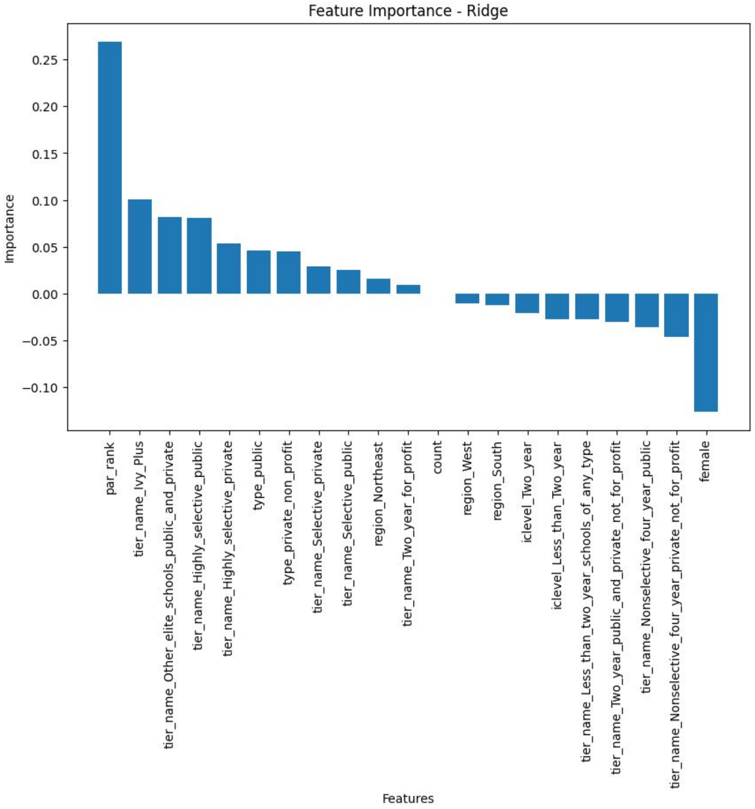
Model #2: Drop par_mean and par_median



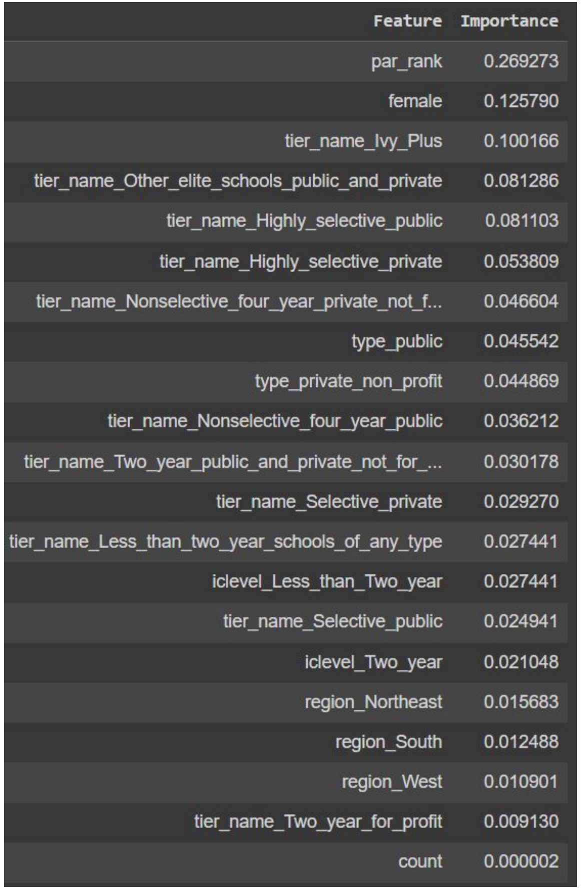


# Model #2 Feature Importances





# Ridge Regression Feature Importances

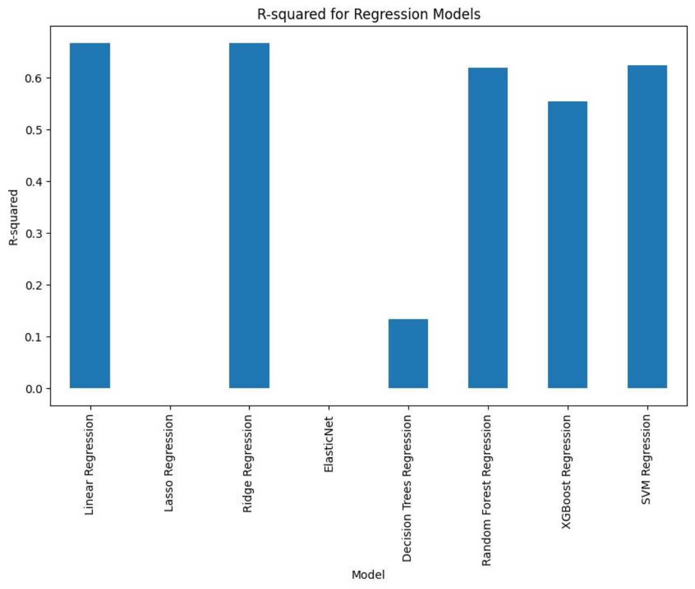


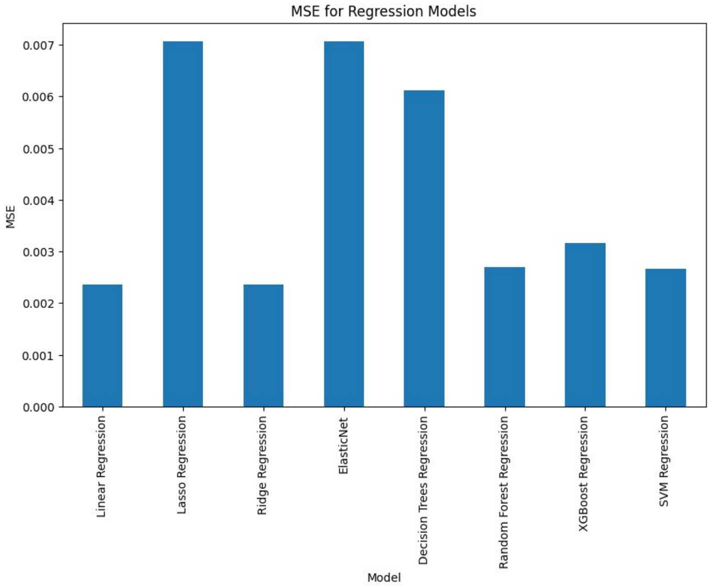
| Feature | Importance |
|---|---|
| par_rank | 0.269273 |
| female | 0.125790 |
| tier_name_Ivy_Plus | 0.100166 |
| tier_name_Other_elite_schools_public_and_private | 0.081286 |
| tier_name_Highly_selective_public | 0.081103 |
| tier_name_Highly_selective_private | 0.053809 |
| tier_name_Nonselective_four_year_private_not_f... | 0.046604 |
| type_public | 0.045542 |
| type_private_non_profit | 0.044869 |
| tier_name_Nonselective_four_year_public | 0.036212 |
| tier_name_Two_year_public_and_private_not_for_... | 0.030178 |
| tier_name_Selective_private | 0.029270 |
| tier_name_Less_than_two_year_schools_of_any_type | 0.027441 |
| iclevel_Less_than_Two_year | 0.027441 |
| tier_name_Selective_public | 0.024941 |
| iclevel_Two_year | 0.021048 |
| region_Northeast | 0.015683 |
| region_South | 0.012488 |
| region_West | 0.010901 |
| tier_name_Two_year_for_profit | 0.009130 |
| count | 0.000002 |

Absolute value taken for each feature importance value

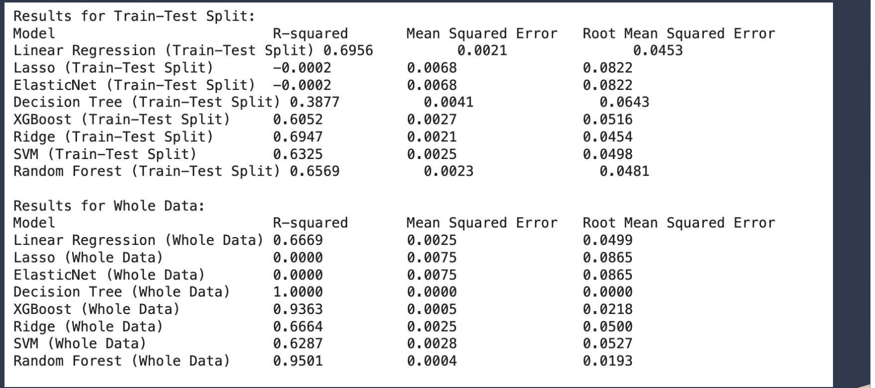
# Prediction Model #3

Model #3: Keeping the 10 most important features

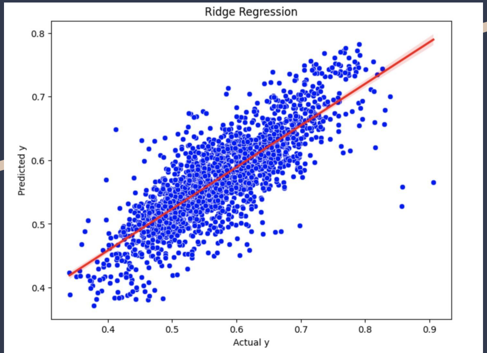




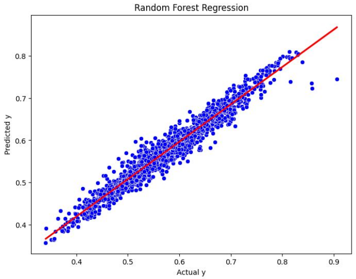
# Interpretations



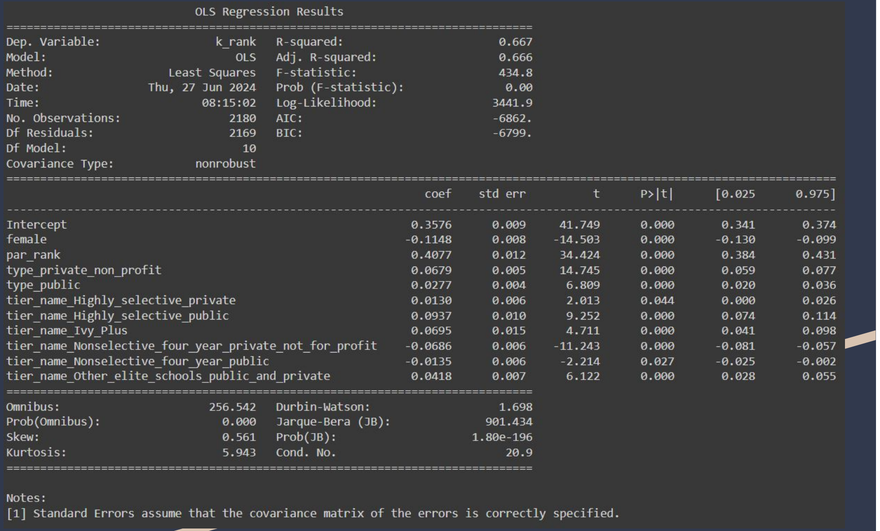
```
Results for Train-Test Split:
Model                         R-squared       Mean Squared Error   Root Mean Squared Error
Linear Regression (Train-Test Split) 0.6956        0.0021              0.0453
Lasso (Train-Test Split)      -0.0002         0.0068               0.0822
ElasticNet (Train-Test Split)  -0.0002        0.0068               0.0822
Decision Tree (Train-Test Split) 0.3877          0.0041               0.0643
XGBoost (Train-Test Split)    0.6052          0.0027               0.0516
Ridge (Train-Test Split)      0.6947          0.0021               0.0454
SVM (Train-Test Split)        0.6325          0.0025               0.0498
Random Forest (Train-Test Split) 0.6569          0.0023               0.0481

Results for Whole Data:
Model                         R-squared       Mean Squared Error   Root Mean Squared Error
Linear Regression (Whole Data) 0.6669         0.0025               0.0499
Lasso (Whole Data)            0.0000          0.0075               0.0865
ElasticNet (Whole Data)       0.0000          0.0075               0.0865
Decision Tree (Whole Data)    1.0000          0.0000               0.0000
XGBoost (Whole Data)          0.9363          0.0005               0.0218
Ridge (Whole Data)            0.6664          0.0025               0.0500
SVM (Whole Data)              0.6287          0.0028               0.0527
Random Forest (Whole Data)    0.9501          0.0004               0.0193
```



- While the Random Forest model has an r squared value of 0.9501 and decision tree of 1.0000. Ridge model performs best on new data.
- Ridge Regression is the model we selected to characterize our dataset.
- Feature importance (top 10)
- Results from OLS regression with top 10 features
- Scatterplot of ridge regression model with line of best fit

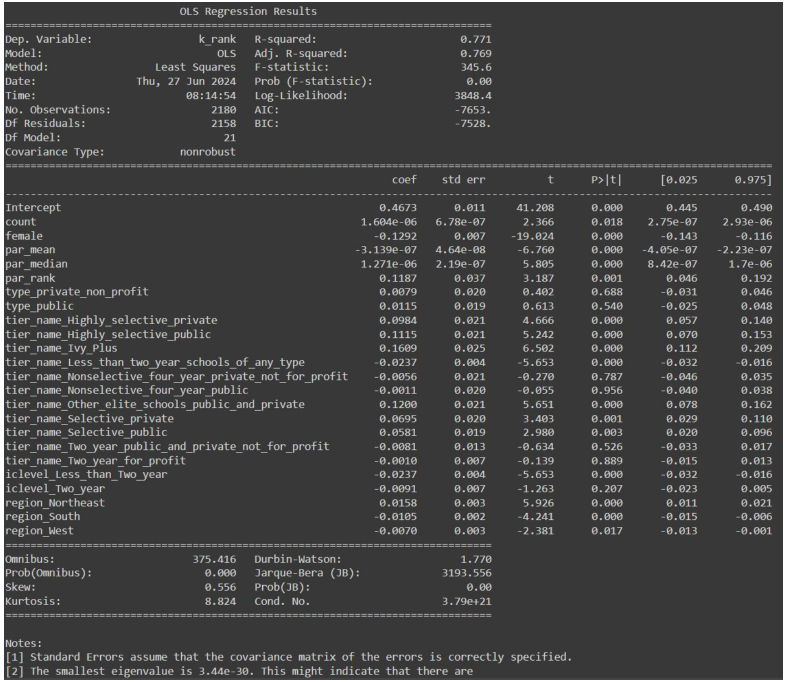


# Final Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 k_rank   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     434.8
Date:                Thu, 27 Jun 2024   Prob (F-statistic):               0.00
Time:                        08:15:02   Log-Likelihood:                 3441.9
No. Observations:                2180   AIC:                            -6862.
Df Residuals:                    2169   BIC:                            -6799.
Df Model:                          10
Covariance Type:            nonrobust
==========================================================================================================
                                                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
Intercept                                       0.3576      0.009     41.749      0.000       0.341       0.374
female                                         -0.1148      0.008    -14.503      0.000      -0.130      -0.099
par_rank                                        0.4077      0.012     34.424      0.000       0.384       0.431
type_private_non_profit                         0.0679      0.005     14.745      0.000       0.059       0.077
type_public                                     0.0277      0.004      6.809      0.000       0.020       0.036
tier_name_Highly_selective_private              0.0130      0.006      2.013      0.044       0.000       0.026
tier_name_Highly_selective_public               0.0937      0.010      9.252      0.000       0.074       0.114
tier_name_Ivy_Plus                              0.0695      0.015      4.711      0.000       0.041       0.098
tier_name_Nonselective_four_year_private_not_for_profit    -0.0686      0.006    -11.243      0.000      -0.081      -0.057
tier_name_Nonselective_four_year_public        -0.0135      0.006     -2.214      0.027      -0.025      -0.002
tier_name_Other_elite_schools_public_and_private     0.0418      0.007      6.122      0.000       0.028       0.055
==============================================================================
Omnibus:                      256.542   Durbin-Watson:                   1.698
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              901.434
Skew:                           0.561   Prob(JB):                    1.80e-196
Kurtosis:                       5.943   Cond. No.                         20.9
==============================================================================

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

- Regression results
  - Using top 10 features from Ridge



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 k_rank   R-squared:                       0.771
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     345.6
Date:                Thu, 27 Jun 2024   Prob (F-statistic):               0.00
Time:                        08:14:54   Log-Likelihood:                 3848.4
No. Observations:                2180   AIC:                            -7653.
Df Residuals:                    2158   BIC:                            -7528.
Df Model:                          21
Covariance Type:            nonrobust
=========================================================================================================================
                                                             coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
Intercept                                                  0.4673      0.011     41.208      0.000       0.445       0.490
count                                                   1.604e-06   6.78e-07      2.366      0.018    2.75e-07    2.93e-06
female                                                    -0.1292      0.007    -19.024      0.000      -0.143      -0.116
par_mean                                               -3.139e-07   4.64e-08     -6.760      0.000   -4.05e-07   -2.23e-07
par_median                                              1.271e-06   2.19e-07      5.805      0.000    8.42e-07     1.7e-06
par_rank                                                   0.1187      0.037      3.187      0.001       0.046       0.192
type_private_non_profit                                    0.0079      0.020      0.402      0.688      -0.031       0.046
type_public                                                0.0115      0.019      0.613      0.540      -0.025       0.048
tier_name_Highly_selective_private                         0.0984      0.021      4.666      0.000       0.057       0.140
tier_name_Highly_selective_public                          0.1115      0.021      5.242      0.000       0.070       0.153
tier_name_Ivy_Plus                                         0.1609      0.025      6.502      0.000       0.112       0.209
tier_name_Less_than_two_year_schools_of_any_type          -0.0237      0.004     -5.653      0.000      -0.032      -0.016
tier_name_Nonselective_four_year_private_not_for_profit   -0.0056      0.021     -0.270      0.787      -0.046       0.035
tier_name_Nonselective_four_year_public                   -0.0011      0.020     -0.055      0.956      -0.040       0.038
tier_name_Other_elite_schools_public_and_private           0.1200      0.021      5.651      0.000       0.078       0.162
tier_name_Selective_private                                0.0695      0.020      3.403      0.001       0.029       0.110
tier_name_Selective_public                                 0.0581      0.019      2.980      0.003       0.020       0.096
tier_name_Two_year_public_and_private_not_for_profit      -0.0081      0.013     -0.634      0.526      -0.033       0.017
tier_name_Two_year_for_profit                             -0.0010      0.007     -0.139      0.889      -0.015       0.013
iclevel_Less_than_Two_year                                -0.0237      0.004     -5.653      0.000      -0.032      -0.016
iclevel_Two_year                                          -0.0091      0.007     -1.263      0.207      -0.023       0.005
region_Northeast                                           0.0158      0.003      5.926      0.000       0.011       0.021
region_South                                              -0.0105      0.002     -4.241      0.000      -0.015      -0.006
region_West                                               -0.0070      0.003     -2.381      0.017      -0.013      -0.001
==============================================================================
Omnibus:                      375.416   Durbin-Watson:                   1.770
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3193.556
Skew:                           0.556   Prob(JB):                         0.00
Kurtosis:                       8.824   Cond. No.                     3.79e+21
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.44e-30. This might indicate that there are
```

# Some Drawbacks and Improvements

- More extensive hyperparameter tuning could’ve been done
  - Used alpha value = 0.1
- Selection of top 10 features
  - Ex. In model 3, choose top 10 features
  - Somewhat arbitrary, could’ve chosen 9, 11, etc.

# Conclusion

# Works Cited

1. Raj Chetty, John N Friedman, Emmanuel Saez, Nicholas Turner, Danny Yagan, Income Segregation and Intergenerational Mobility Across Colleges in the United States, *The Quarterly Journal of Economics*, Volume 135, Issue 3, August 2020, Pages 1567–1633, https://doi.org/10.1093/qje/qjaa005