Nicholas Angeramo and Zac Nelson-Marois
Professor Ethan Levien
Introduction to Linear Models - MATH 50
20 November 2022

How One Model Changed the American Higher Education System 'Forever'

The book *Weapons of Math Destruction* by Cathy O'Neil highlights the frequency in which algorithms make important decisions in close to every domain of our lives, including education, criminal justice, finances, healthcare, advertising, employment, politics, insurance, and social media.[1] O'Neil argues throughout the novel that the lack of transparency, accessibility, and disputability of these mathematical models, which usually do not accurately model society, threatens to perpetuate discriminatory practices. These algorithms begin as a relatively harmless idea, and aim to effectively manage large amounts of data in a more productive manner, but lead to success of companies and therefore profits. This ends up being disadvantageous to outliers within marginalized communities who do not fit within these "mainstream" algorithms. The book points out various examples of algorithms overlooking these outliers and contributing to inequalities, suggesting that there is bias present within the data sets used to train the algorithms and the people who create them.

There are multiple instances throughout the novel which suggest that models are simplifications and are incapable of suggesting the world's true variability. One specific example of this is contained in Chapter 3, which discusses the *U.S. News and World Report's* rise to popularity, for better or for worse. What began as a struggling magazine looking for profits turned into an ambitious evaluation of more than 1,800 colleges and universities. While at face value, one may see a rigorous evaluation of our colleges and universities as a great benefit to society, it resulted in adverse effects that would change the American education system forever.

The *US News and World Report's* College and University Rankings took into account various factors in order to determine the excellence of an academic institution. Twenty-five percent of the ranking was determined by qualitative measures through surveying other professionals in higher education regarding their perception of their peer institutions. Seventy-five percent of the ranking consisted of quantitative factors. As stated in the book, it would be near impossible to quantitatively measure learning, happiness, confidence, friendships, and other aspects of students' experiences while in college. Therefore, the US News has to rely on proxies to attempt to measure some of these attributes. In data science, a proxy is when an indirect measurement is used due to its correlation with the desired outcome. In this case, the model the US News uses is based upon metrics surrounding a school's operational and admissions procedures. These metrics included: SAT scores, student-teacher ratios, acceptance rates, retention rates, graduation rates, and living alumni donation rates. While this may seem like a sound way to measure a college's performance, these proxies are not necessarily indicative

[1] O'Neil, *Weapons of Math Destruction*.

of the strength of their practices. While it is easy to assume a low student-teacher ratio results in better results in the classroom based on prior educational studies (CITE), the metric is in no way indicative of the type of instruction students are receiving once in the classroom. Despite serving as a good predictor, it is impossible to truly determine whether or not a low student-teacher ratio causes a better academic result at that institution.

The biggest downside of using a proxy is how vulnerable it makes the model to the subjects it attempts to analyze. Its fatal flaw is that proxies can directly impact the behavior of colleges and universities in order to conform to what the model deems as an important predictor of educational excellence; this is exactly what happened after the US News launched their ranking system. Instead of being a reflection of the state of education in America, the rankings have created a system described as "an endless spiral of destructive feedback loops." Schools have shifted their entire institutional goals towards performing well on these rankings in order to attract the best talent in students and professors. As discussed in the book, Baylor University paid admitted students to retake the SAT in order to increase their school's average score. Bucknell University and Claremont McKenna reported inflated SAT scores to the US News. Iona College inflated not only its SAT scores when reporting, but also every other metric the US News considers in its college rankings. Just this year, Columbia University, who was previously ranked second on the rankings, were found to be falsifying their institutions statistics, and they are now ranked eighteenth.[2] This dishonesty across American institutions shows that they are valuing prestige in the eyes of the public and revenue from incoming students they seek to attract over everything else. This highlights that the student experience and academic quality while attending is not their top priority.

This behavior shift goes beyond misreporting practices as well. Colleges now have shifted to be much more profit-oriented, raising tuition at rates significantly higher than inflation over the past 30 years. This increased revenue becomes necessary in order to expand their budgets in order to afford heightened capital expenditures. More money is being poured into athletic programs and campus improvement projects in order to increase the image of the school. The goal is then for the investments to result in a lower acceptance rate for the school, another metric accounted for by the US News & World Report's rankings. We have seen this occur throughout history with the success of Division 1 sports teams. O'Neil cites the effect that Doug Flutie and Patrick Ewing had on their university's acceptance rates, Boston College and Georgetown respectively, due to the increase in applications the schools saw surrounding their exceptional athletic performances in the 1980s. This is yet another example where schools have changed their behavior to center around profit and lowering their acceptance rate to achieve a higher ranking. This results in a neglect of the students actually at the school and the education the university is actually providing.

While conforming operations to a model affects current university students, it also affects aspiring college students and the education system as a whole. As colleges are actively making it harder to attend their universities, high school students then need to work even harder to conform

[2] Selcho, "Columbia Fell from No. 2 to No. 18 in Rankings of Universities."

to university's consistently rising standards. This causes many aspiring college students to seek help outside of school in order to get in, opening up an entire industry of college consultants that charge outrageous prices for their services. The result is the wealthy paying the price set by college counselors and gaining an edge in the admissions process. However, this comes at the expense of middle and lower class students who can not afford these services and are worse off in the college admissions process. In the long term, this will only exacerbate wealth inequality in America.

Weapons of Math Destruction clearly showed the impactful role that the US News & World Report college rankings played from 1983 - 2016 (when the book was published). However, this left a large gap in terms of whether things have changed since then. To answer how the model has changed in the last six years, we wanted to look into the *US News & World Report* in its current condition, and ask ourselves the role that various proxies, or in this case, independent variables, play in terms of predicting *US News & World Report* ranking. According to *Weapons of Math Destruction*, tuition, fees, and student debt were typically left out of the ranking, whereas success metrics (SAT scores, student teacher ratios, etc.) accounted for a large portion of the score. We want to answer the question as to what extent this is still the case, if at all, and identify other variables that could play a role in indicating rank, since it is not exactly deducible upon intuition, nor is there a formula directly provided anywhere. This will help us answer the overarching question as to whether or not models need to be reformed, and if we would be better off in certain instances if they were not applied.

This question is intriguing to us because of our interests beyond this class. Zac is majoring in computer engineering as well as minoring in statistics, and has always wanted to better understand the college admissions process and the legitimacy of various college ranking sites. As a computer engineer, he wants to understand how one could implement such a model through coding and mathematics effectively so that it is representative of a large sample of colleges. Nick is a QSS major writing a thesis surrounding the energy transition. Algorithms are an integral piece of the puzzle when it comes to managing a smarter and more complex electrical grid. They are also utilized heavily when anticipating demand throughout the energy transition and modeling the energy needs as we plan for a more sustainable society.

The author's approach to determining what factors most influence ranking was reinforced through qualitative examples, finding various instances in which colleges and universities attempted to boost their rank, as discussed previously. However, a more concrete example of this is in 2008, TCU, who was falling in the US News and World Report rankings, gamed the system by launching a fundraising drive that brought in $434 million (fundraising is one of the metrics) and used a lot of the money to improve campus, including a state-of-the-art training facility and the football program. This led to more school spirit, more applications, higher test scores, more rejections, and thus, higher acceptance rate, making TCU the second most selective university in Texas by 2013.

In contrast, our approach is to look at statistical data and analyze it to determine patterns among variables. To do this, we identified our dependent variable that we are trying to predict, Y,

and a host of explanatory variables (X1, X2, X3, X4, etc.), including acceptance rate, tuition, percent of students on financial aid, average SAT score, whether a university is public or private, average ACT score, average high school GPA, and average cost after financial aid to analyze the relationship between rank and either a single or multiple independent variables whose values are known. We performed single regression models as well as multiple regression models because sometimes, a multiple regression model is not a good estimate of Y with many variables since some of the variables may be correlated with one another, so looking at more specific trends can sometimes result in a higher R^2 value, and is therefore a possible better estimation of the data set. We narrowed the data set to national universities, since this is likely representative of a lot of universities and it would be difficult to accurately form conclusions with too much data, some of which may be outliers or not show clear trends in the data.

We performed simple linear regressions for Y on each independent variable to identify necessary test statistics, specifically looking for the R^2 and p-values. To observe the relationship as a whole, we then plotted the data versus the regression line as predicted from the model. The regression line (best fit) is represented by the model Y = aX + b (+ e), where X is the independent variable being used to predict the rank. First, a is represented by the average difference in rank for which X grows by 1 unit. Second, b is the average increase in rank proceeding a period with zero growth in X. Lastly, sigma is the standard deviation in rank for those with the same growth in X.

We also performed multiple linear regressions for Y on multiple independent variables and identified necessary test statistics again. The regression model is represented by Y = a1X1 + a2X2 + a3X3 + … + anXn + b (+ e). For all of these regression coefficients, it must be noted that a given regression coefficient a represents the slope of Y vs. X1 given that all of the other variables (let's denote as X2, X3, etc. if applicable) are conditioned on X2, X3, etc. if applicable (these variables remain fixed/constant). a represents the average difference in rank for which the dependent variable differs by one unit (value, %, $) and the other variables remain constant. Note that a_public represents the average difference in rank between a school that is public as compared to a school that is not public (categorical variable that is not quantifiable). All of the physical coding analyses can be found at the bottom of this document.

After performing various regression analyses and comparing different models, we were able to come to some conclusions regarding which variables have the most influence in predicting US News & World Ranking for national universities. To perform the regression analyses, we found data from 2018 regarding various metrics from the US News & World Report on Github. Though not current, it should give an approximate estimation as to how the model has changed over the course of many years, and hopefully should answer the question as to what variables play the largest role in influencing rank.

Let us discuss the results regarding the simple linear regressions and their respective models. The first statistic that we must pay attention to is the R^2 value. We notice that for the independent variables of tuition (.355), percent of students on financial aid (.046), whether a school is public (.005), high school GPA (.273), and average cost after financial aid (.00l), the

R^2 value is relatively small and is less than 0.4. The R^2 of the model suggests that approximately (R^2 value * 100)% of the variability observed in US News and World Report rank can be explained by the regression model, or the variance in the independent variable. This seems to indicate that the regression model does not explain the observed data well due to this low R^2 value, thereby suggesting that the regression model does not accurately estimate the data set. However, even though the data appears to indicate a low correlation and thus a bad predictive model, there are sometimes great models with low R^2 values, and we should not only rely on this value since other factors such as the causation relationship between rank and the independent variable cannot determine the correctness of the regression model alone. Thus, due to there being a possibility of a possible trend in the data which is not explained by the value, perhaps due to the influence of other confounding variables or biases, we cannot make a definitive conclusion since the conclusion is solely based on the value of R^2 at the moment.

On the contrary, we notice that for the independent variables of acceptance rate (.676), SAT score (.711), and ACT score (.698), the R^2 value is relatively large and is greater than 0.65. This seems to indicate that the regression model does explain the observed data well due to this high R^2 value, thereby suggesting that the regression model does seem to accurately estimate the data set. Yet again, we cannot only rely on this statistic, but it is a good indicator that there may be a strong relationship between these independent variables and rank.

Next, let us interpret the p values for each regression coefficient. The p-value expresses the likelihood that the data occurs by random chance (or that the null hypothesis is true), such that the smaller the p-value, the stronger the evidence that the null hypothesis should be rejected. Specifically, if the p-value is less than or equal to 0.05, then it is statistically significant such that we reject the null hypothesis and accept the alternative hypothesis. We note that the p-values for acceptance rate (0.000), tuition (0.000), percent of students on financial aid (0.003), SAT score (0.000), ACT score (0.000), and high school GPA (0.000) are all <= 0.05, which indicates that they are statistically significant and that the data likely does not occur by random chance. However, the p-values for public (0.384) and cost after aid (0.777) are significantly larger than 0.05, and are thus not statistically significant (most likely occurred by random chance). This makes relative sense in both instances why the p-values are high; for public, the data is separated into two bins similar to a histogram, so we would expect the regression results to be skewed, and for cost after aid, the data has almost no correlation and appears to be random (very difficult to estimate and likely occurs by chance).

These pieces of information seem to align with our generated models. We plotted the actual data against the regression line of best fit to measure the variability from our model to the actual data. These visualizations show some intriguing trends. As acceptance rate decreases with ranking decreasing, the variability between the regression line and actual data seems to decrease as well. As cost of tuition increases and rank decreases, a similar pattern is observed in that the variability decreases. Average high school GPA does seem to matter to a variety of schools, but less than one may think, as the graph is relatively dispersed, although there is a clear increasing GPA trend as rank decreases. SAT and ACT scores are mostly evenly distributed on both sides of

the regression line and appear very linear, hence the highly correlation; also, both SAT and ACT scores show a very clear increasing trend as rank decreases. Variables such as cost after aid and percent of students given financial aid seem to be very randomly distributed with no clear pattern. Lastly, private schools tend to be lower in the ranks than public schools, with the lowest public school being ranked around 20.

The single regression analyses seem to indicate that generally, success-related predictor variables have the most significant influence on US News & World Report rank (i.e. acceptance rate, test scores). We find this particularly intriguing because with many top institutions going test-optional in recent years, we see an increased application volume, and therefore decreased acceptance rate. At Dartmouth alone, we see that the regular decision acceptance rate for the class of 2024 was cut by a third, with the class of 2024 at 6.9% and the class of 2026 at 4.7%, and the effects are even more dramatic at public universities.[3] Colleges have almost no incentive to revert back to normal because students know to submit scores only if it is high compared to most others, and this drives up their test score averages while decreasing their acceptance rate averages. We do not see this same effect for high school GPA, which has a spread out distribution albeit still increasing trend in GPA as rank decreases. Likewise, financial-related predictor variables (i.e. cost, aid) seem to have limited role in rank. This supports the book in the sense that middle and poor class family students do not benefit from a system that prioritizes test success, for which only wealthy individuals have access. There is no equal access provision to those who need it; rather, they must do the best they can with the minimal free opportunities available.

These trends are observed as well when comparing this hypothesis with the multiple regression models generated. First, let us consider four models: rank as a function of acceptance rate, tuition, percent aid, SAT, public or not, ACT, HS GPA, and cost after aid (which is not the best model because many of the variables are correlated with one another, but still a good indicator of general trend), rank as a function of test scores (SAT, ACT), rank as a function of monetary factors (tuition, percent aid), and rank as a function of highly correlated variables (acceptance rate, ACT, SAT).

Next, let us interpret R^2 for each of the models. The R^2 for each of the models as listed above, respectively, are 0.784, 0.440, 0.717, and 0.742, suggesting the proportion of variability observed in a given child's test score that can be explained by the independent variables. The first model with all 8 variables seems to model the data well, but this is most likely due to the fact that the model is not a good estimate with this many variables, and multicollinearity effects are probably in place. Therefore, even though the R^2 value is high, the regression model may not predict the observed data well. The second model with monetary factors does not seem to explain the observed data very well since there is a relatively low R^2 value, suggesting the regression model does not accurately estimate the data set. However, this could be due to the fact that the model is not a good estimate with this many variables, and looking at more specific trends may very well produce a much better R^2 value (indicating a possible better estimation of the data

[3] "Dartmouth Admissions Acceptance Rates & Statistics."

set). The other two models have high R^2 values and seem to explain the observed data relatively well based on the R^2 and there not being too many variables. There may be some multicollinearity effects, but overall, the model seems like a good predictor for rank based on the observed information to this point.

Furthermore, let us compare the p-values for each regression coefficient and for each model. We note that the p-values of x1, x2, and x3 (acceptance rate, tuition, percent aid) are all <= 0.05, which indicates that they are statistically significant and that the data likely does not occur by random chance. Unfortunately, the p-values of x4-x8 are larger than 0.05, and are not statistically significant and occurred by random chance. This is largely due to the fact that many of the variables are correlated with one another, thereby influencing the results. Likewise, the p-values of the monetary model are all <= 0.05, which indicates that they are statistically significant. In the test score model, x1 (ACT) is not statistically significant whereas x2 (SAT) is statistically significant, but the value for x1 is close to 0.05 (=0.078), so it is not a major cause for concern. Finally, in the correlation model, x1 (0.080) and x2 (0.058) aren't statistically significant, whereas x3 is statistically significant. Yet again, the values are close to 0.05, so it is not a major cause for concern. Even though there may not be as strong evidence to reject the null hypothesis, for the sake of the samples given and the somewhat odd data present, we will assume the data for the last three models does not occur by random chance, or at least, close to it.

The multiple regression analyses indicate similar trends to the single regression analyses. The first large sum model, while somewhat useful, does not provide much information except that the independent variables are largely correlated, which we already knew. However, the other three models do provide some useful information. The monetary factors model confirms that with a low R^2 and data that likely does not occur by chance, we can presume it is not a great estimator for rank. Likewise, the test score factors model confirms that with a high R^2 and unlikely by chance data that it is a good predictor of rank. When acceptance rate is also factored in with test scores, this seems to be an even better predictor with a higher R^2 value, although it is important to note that R^2 is not necessarily a quantifiable comparative measure and rather an estimate for that particular data set. This follows the scope of the success-related indicators mentioned throughout the course of the novel.

The analysis revealed that acceptance rate, SAT score average, and ACT score average are all metrics that schools likely target the most in terms of boosting their US News & World Report ranking. This is indicated by their strong individual relationships with rank, as well as their collective relationships with rank, both of which display high correlation coefficients squared as well as, for the most part, statistical significance. This suggests that the data is highly correlated and likely does not occur by chance, meaning that the observable trends are present and likely. These are all generally variables that are a quantifiable measure of "success", both as a university and the types of students that attend the school in a standardized, distributive manner. In a similar vein, other variables such as tuition, percent of students with aid, and average cost after aid show a much lesser existence of trends that support ranking both individually and collectively, which suggests that financial factors have a limited effect on rank.

This means that schools likely target the other sources to increase their ranking, and focus much less on the financial background of the student they are accepting (possibly as a function of the fact that national universities have large endowments and are not at all concerned with money). High school GPA seems to have a limited influence on rank due to the poor individual relationship with rank, which is somewhat surprising; however, it makes sense because high school GPA is not an easily standardizable measure by means of a normal distribution or other form of distribution, so it would not make sense for the US News & World Report or colleges to prioritize these values. Lastly, it appears that the first 20 or so schools are likely not public universities, which also makes sense as there is a larger proportion of wealthy alumni who donate back to private institutions as compared to at public universities with many more students.

The purpose of this analysis was to answer the question as to what variables were primarily impactful in ranking and therefore confirm or deny the reports in the book that financial implications are not considered as much whereas success-related indicators are, which was confirmed. Much of the book was based upon inference rather than statistical analysis, so we attempted to confirm this though a representative sample. Before performing my regression analysis, I was skeptical; many other sources seem to suggest that algorithms are transformative and revolutionary, and biased for intention rather than necessity. For example, when Hurricane Ian destroyed much of Florida in late September 2022 due to high winds and flooding, around 3,500 residents of Collier, Charlotte, and Lee Counties received a push notification offering $700 in assistance.[4] This was due to a nonprofit called GiveDirectly who partnered with Google and was able to determine specific neighborhoods who needed the most assistance through satellite imaging. However, after the analysis, it is clear that schools are attempting to game the system, as financial factors, arguably one of the most important factors, are not even considered; thus, as evidenced by rapidly declining acceptance rates and near-perfect test scores, it is clear schools are targeting students which maximize these values and do not negatively impact their data, as the analysis proves it. Not only does the analysis we performed show that colleges are gaming the system, so do recent trends in data. For example, Northeastern, a school with an 18% acceptance rate in the class of 2025 now boasts an acceptance rate of 6.7% for the class of 2026, a decrease of more than 10% in a year alone.[5] The fact that the acceptance rate dropped so quickly is a reflection of schools' successful attempts to get more students to apply and boost their statistics.

In the book, O'Neil mentions that it may be possible to fix the rankings issue, but at a cost. She mentions that President Obama suggested a new model which aligns more with national priorities and middle-class families. This model would essentially redistribute the power imbalance that is in place for private institutions by tying a new set of metrics to affordability, percentage of poor and minority students, and post graduation job placement, as well as graduation rate. In this system, colleges would be cut off from the federal student loan market, a

[4] WIRED, "Hurricane Ian Destroyed Their Homes. Algorithms Sent Them Money."
[5] Armanini, "Northeastern Acceptance Rate Drops to 6.7%."

$180 billion dollar industry, if they did not meet certain benchmarks. However, there are clearly drawbacks to even supposedly good ideas such as this one. For example, raising graduation rates could simply mean reducing difficult course requirements, but this would severely limit the breadth of the educational system; lowering costs is as simple as removing highly paid tenured faculty, but this limits the quality of education as well.

Perhaps the best method to fight the system is to go against it completely and let students, the true consumers, rank colleges themselves. The US Department of Education has a bunch of unbiased data released, such as class sizes, graduation rates, and average debt. Making models that are "transparent, controlled by the user, and personal" is the best way to combat the issue of the unethical and impractical models regarding college rankings that exist. Regardless, it is evident that models, while sometimes positive, can have very negative consequences, and ones that are capable of shaping entire societies, whether intentional or not.

We attempted to determine the role that certain variables play in predicting college rank, and it is very clear from our analysis that there are some factors missing that would contribute to a better model for the average student. The knowledge from the book Weapons of Math Destruction supports the analysis we performed, and implies that there is much corruptness behind the scenes. College admissions seemed like a game when we were applying, but little did we realize how much our test scores mattered, all so colleges can be boosted on a should-be-meaningless algorithm. The unfortunate truth, however, is that rankings such as the *US News & World Report* have significant weight in regards to how societies view colleges and higher education, and the only way to change this system is to reform it altogether. Algorithms aren't going anywhere, and many of the systems which they control are too unprincipled to enact as well as expect change. And after all, even the best models are susceptible to loopholes and exploitation.

Works Cited

Armanini, Kate. "Northeastern Acceptance Rate Drops to 6.7%." *The Huntington News* (blog), April 21, 2022. https://huntnewsnu.com/68615/campus/northeastern-acceptance-rate-drops-to-6-7/.

Top Tier Admissions. "Dartmouth Admissions Acceptance Rates & Statistics." Accessed November 22, 2022. https://toptieradmissions.com/resources/college-admissions-statistics/dartmouth-college-acceptance-rates/.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown, 2016.

Selcho, Madison. "Columbia Fell from No. 2 to No. 18 in Rankings of Universities." Deseret News, September 14, 2022. https://www.deseret.com/2022/9/13/23351917/us-news-world-report-columbia-falls.

WIRED. "Hurricane Ian Destroyed Their Homes. Algorithms Sent Them Money." Ars Technica, October 11, 2022. https://arstechnica.com/information-technology/2022/10/hurricane-ian-destroyed-their-homes-algorithms-sent-them-money/.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
%config InlineBackend.figure_format = "svg"
#from pymc3 import *


#Zac Nelson-Marois & Nick Angeramo
#Final Project Colab Notebook


#I found this US News and World Report Data (from 2018) on GitHub.
#The link should work as provided. If it does not work, here is the
#google URL:
#https://github.com/kajchang/USNews-College-Scraper
#Note that there may have been some changes here and there since the data is
#not current. However, I made the reasonable assumption that the data is
#closely representative as to what we would see today since trends have
#not drastically changed and it is safe to assume that the data I observed
#closely represents the data observed today.
#Read file from GitHub
data = pd.read_csv("https://raw.githubusercontent.com/kajchang/USNews-College-Scrap
#Set the school type as the variable to be compared and make it the column all
#the way to the left
data_with_index = data.set_index("institution.schoolType")
data_with_index.head()
#Drop all rows which are not national universities, as this will allow me to
#work with a smaller data set. There are too many schools to observe trends
#effectively, so the best way to do so is to eliminate all other schools
#which comes out to a population of about 400.
#Note: There is likely a faster way to do this, but the data I am working with
#is somewhat confusing with its many entries and it not operating the same
#as previous assignments. Thus, some of my actions may not be completely
#representative of what we have done throughout the class, but it still gets
#the job done relatively efficiently and produces the desired outcomes.
data_with_index.drop(["regional-universities-west","regional-colleges-south","unrar
data_with_index
```

institution.schoolType	institution.displayName	institution.aliasNames	insti
national-universities	Adelphi University	NaN	
national-universities	Alliant International University	NaN	
national-universities	American University	NaN	
national-universities	Andrews University	NaN	
national-universities	Arizona State University--Tempe	ASU	
...	...	...	
national-universities	Wingate University	NaN	
national-universities	Worcester Polytechnic Institute	WPI	
national-universities	Wright State University	NaN	
national-universities	Yale University	NaN	
national-universities	Yeshiva University	NaN	

399 rows × 32 columns

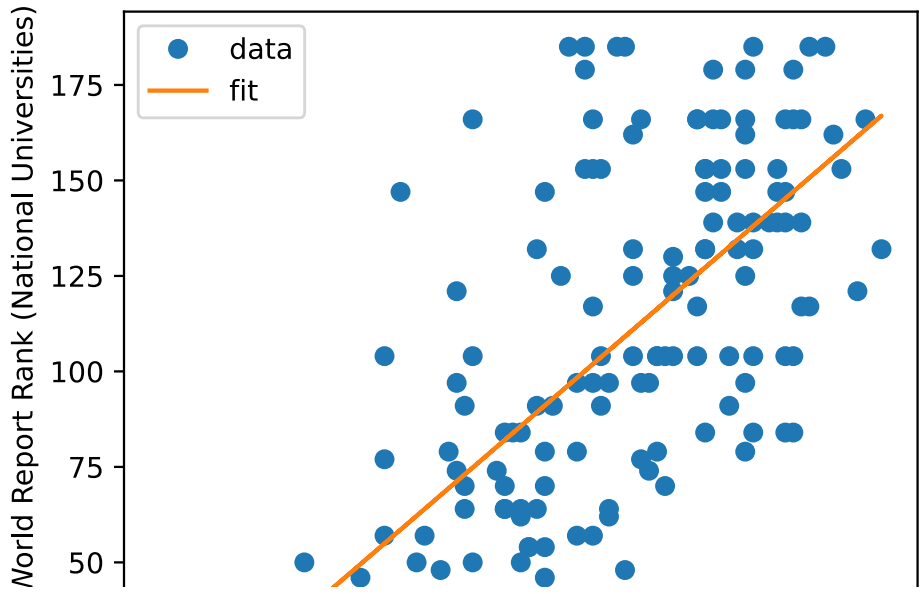
```
#Here, I am defining the variables for which I wish to compare
data_x = data_with_index[["ranking.sortRank","searchData.acceptanceRate.rawValue","
#I drop all rows which have NaN so that observable trends are relevant. I
#had around 400 rows before, so cutting that in half does little damage in terms
#of the trends that we will or will not see.
data_x = data_x.dropna()
```

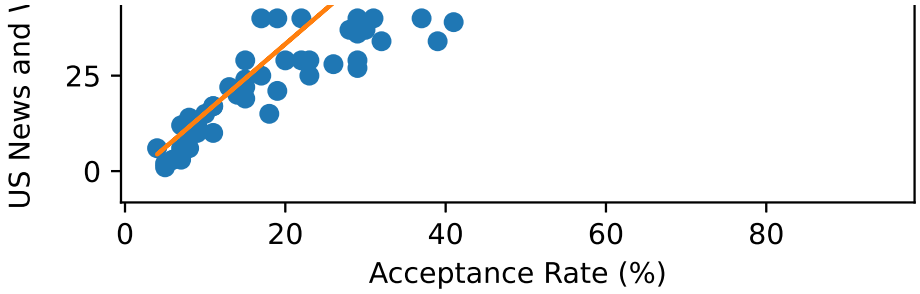
```
#Below, I attempted to collect arrays in a similar manner to the notes,
#but unfortunately, the variable names throw an error and do not allow me to
#collect directly from data_x. As a result, I iterated through all of the
#possible values of the table that I just created, thereby gathering the values
#of a given column by keeping the column constant and iterating through all of
#the rows. Most of my variables were quantitative.
#US News and World Report Ranking (#)
#Define an array the size of all of the values
rank = np.zeros(len(data_x))
#Iterate through all of the values
for i in range(len(data_x)):
  #Find the rank at each school that has not been omitted from before
  rank[i] = data_x.values[i,0]
#Acceptance Rate (%)
acceptance_rate = np.zeros(len(data_x))
for i in range(len(data_x)):
  acceptance_rate[i] = data_x.values[i,1]
#Average Tuition ($)
tuition = np.zeros(len(data_x))
for i in range(len(data_x)):
  tuition[i] = data_x.values[i,2]
#Percent of Individuals Recieving Aid (%)
percent_aid = np.zeros(len(data_x))
for i in range(len(data_x)):
  percent_aid[i] = data_x.values[i,3]
#Average SAT Score
sat = np.zeros(len(data_x))
for i in range(len(data_x)):
  sat[i] = data_x.values[i,4]
#Public University or Not
#Define an empty array of strings
public = ["" for x in range(len(data_x))]
#Iterate through array and replace with "public", "private", etc. strings
for i in range(len(data_x)):
  public[i] = str(data_x.values[i,5])
#Iterate through newly formed array
for i in range(len(public)):
  #Replace "public" with 1 (occurrence) and other mostly "private", etc. with 0
  if public[i] == 'public':
    public[i] = 1
  else:
    public[i] = 0
#Average ACT Score
act = np.zeros(len(data_x))
for i in range(len(data_x)):
```

```
  act[i] = data_x.values[i,6]
#Average High School GPA
hs_gpa = np.zeros(len(data_x))
for i in range(len(data_x)):
  hs_gpa[i] = data_x.values[i,7]
#Average Cost After Aid ($)
cost_after_aid = np.zeros(len(data_x))
for i in range(len(data_x)):
  cost_after_aid[i] = data_x.values[i,8]


#Now that my data has been cut so that only the "good" data is left, the
#variables have been narrowed down to those of interest, and the data is in
#array form, I can perform regression analysis. Note that all of my regression
#analysis is to predict US N&WR Rank from independent variable(s). First, I
#wanted to explore the relationship between rank and an arbitrary IV. I graphed
#and performed a regression analysis for each. My conclusions regarding relevant
#data will be included in the final paper.


#Predicting rank from acceptance rate
#Generate regression model
#Rank array (dependent variable)
Y = rank
#Independent variable array, with a constant added to the array for proper
#regression analysis (to find b value as well)
X = sm.add_constant(acceptance_rate)
#Calculate ordinary least squares regression and store it
model = sm.OLS(Y,X)
#Train model so predictions can be made
results = model.fit()
#Print results
print(results.summary())
#Store values of intercept and slope
b_fit,a_fit = results.params
#Store standard deviation
sigma_fit = np.sqrt(results.mse_resid)

#Generate plot
#Store independent variable in x
x = acceptance_rate
#Generate a plot of length and width 5
fig,ax = plt.subplots(figsize=(5,5))
#Plot the actual data as found previously
ax.plot(x, Y, 'o', label="data")
```

```
#Plot the predicted data as found from the regression model
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
#Set titles and legend
ax.set_xlabel("Acceptance Rate (%)")
ax.set_ylabel("US News and World Report Rank (National Universities)")
ax.legend()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.676
Model:                            OLS   Adj. R-squared:                  0.675
Method:                 Least Squares   F-statistic:                     376.2
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           5.81e-46
Time:                        08:35:26   Log-Likelihood:                -880.72
No. Observations:                 182   AIC:                             1765.
Df Residuals:                     180   BIC:                             1772.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.8541      5.368     -0.532      0.596     -13.447       7.739
x1             1.8064      0.093     19.395      0.000       1.623       1.990
==============================================================================
Omnibus:                       16.937   Durbin-Watson:                   1.730
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.974
Skew:                           0.707   Prob(JB):                     7.58e-05
Kurtosis:                       3.709   Cond. No.                         136.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
<matplotlib.legend.Legend at 0x7f923c9d3fd0>
```

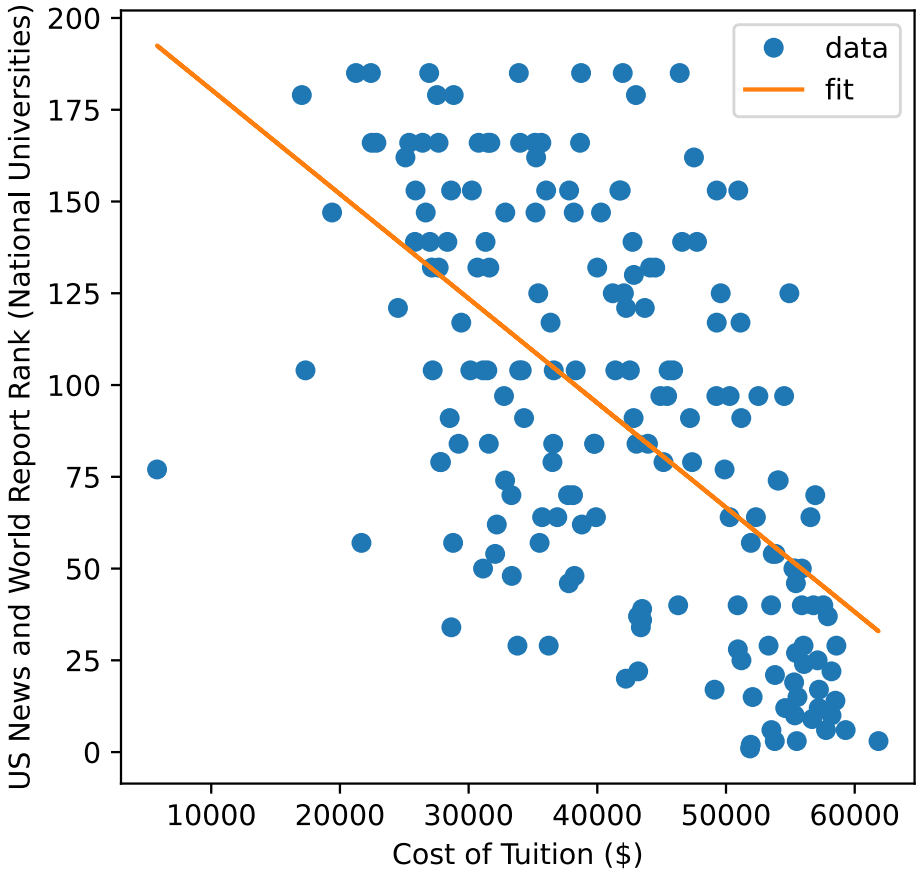





```
#Predicting rank from tuition (note that the condition number is of no concern
#due to the unorthodox range of the data)
Y = rank
X = sm.add_constant(tuition)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = tuition
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
ax.set_xlabel("Cost of Tuition ($)")
ax.set_ylabel("US News and World Report Rank (National Universities)")
ax.legend()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.355
Model:                            OLS   Adj. R-squared:                  0.352
Method:                 Least Squares   F-statistic:                     99.15
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           6.98e-19
Time:                        08:35:30   Log-Likelihood:                -943.45
No. Observations:                 182   AIC:                             1891.
Df Residuals:                     180   BIC:                             1897.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        208.9447     12.234     17.079      0.000     184.805     233.085
x1            -0.0028      0.000     -9.957      0.000      -0.003      -0.002
==============================================================================
Omnibus:                        3.539   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.170   Jarque-Bera (JB):                2.414
Skew:                           0.089   Prob(JB):                        0.299
```

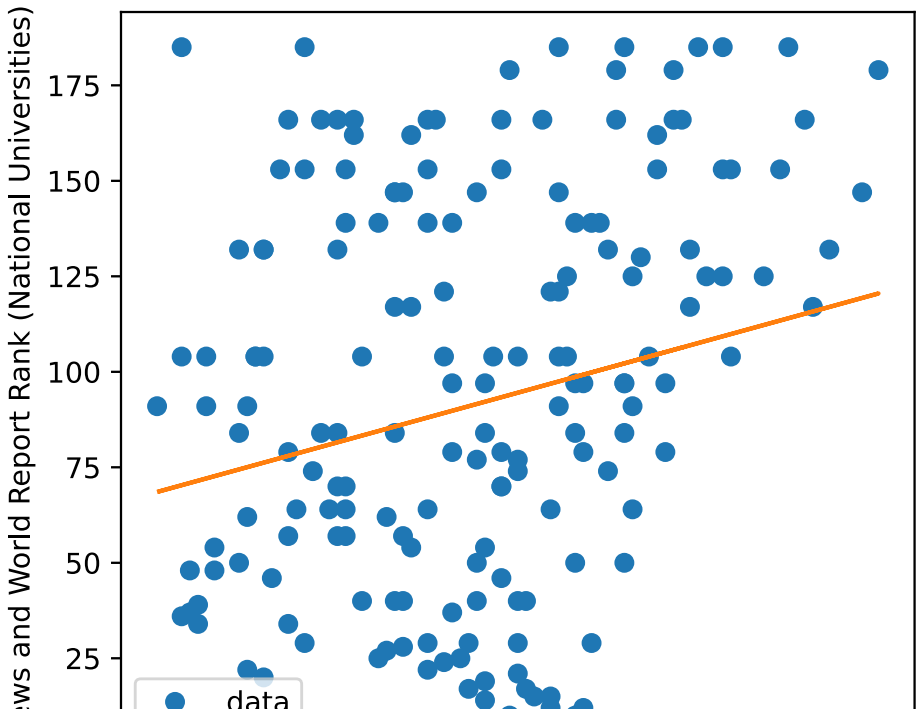
```
Kurtosis:                       2.464   Cond. No.                     1.63e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.63e+05. This might indicate that there ar
strong multicollinearity or other numerical problems.
<matplotlib.legend.Legend at 0x7f924006a750>
```

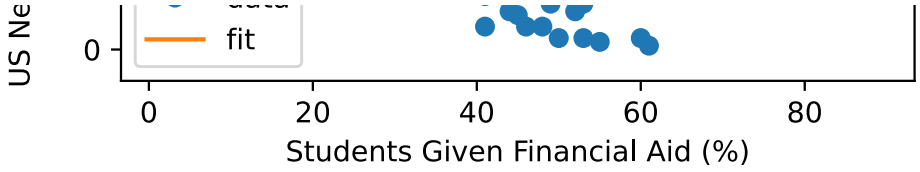



```
#Predicting rank from Percent of Students on Financial Aid
Y = rank
X = sm.add_constant(percent_aid)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = percent_aid
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
```

```
ax.set_xlabel("Students Given Financial Aid (%)")
ax.set_ylabel("US News and World Report Rank (National Universities)")
ax.legend()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     8.760
Date:                Tue, 22 Nov 2022   Prob (F-statistic):            0.00349
Time:                        08:35:33   Log-Likelihood:                -979.06
No. Observations:                 182   AIC:                             1962.
Df Residuals:                     180   BIC:                             1969.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         67.9519      8.841      7.686      0.000      50.507      85.397
x1             0.5906      0.200      2.960      0.003       0.197       0.984
==============================================================================
Omnibus:                       24.034   Durbin-Watson:                   1.897
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                6.839
Skew:                          -0.020   Prob(JB):                       0.0327
Kurtosis:                       2.051   Cond. No.                         100.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
<matplotlib.legend.Legend at 0x7f923c8cead0>
```

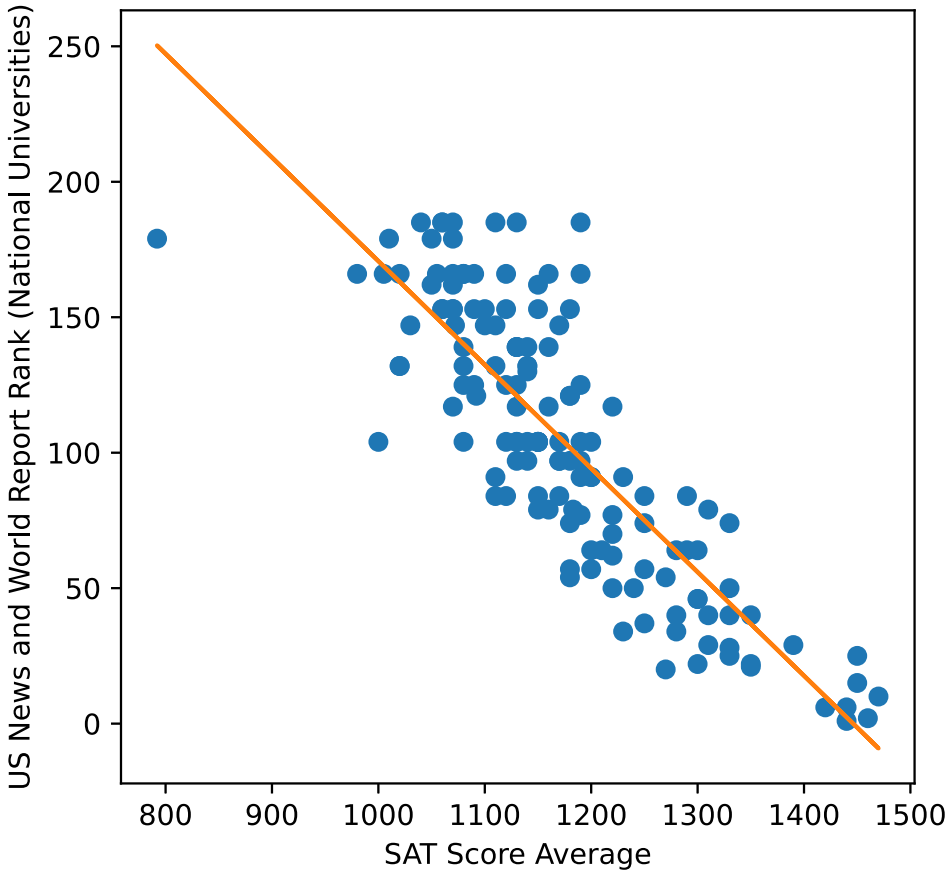





```
#Predicting rank from SAT score (note the condition number is yet again of no
#concern since the range is unorthodox and the values start at around 1000)
Y = rank
X = sm.add_constant(sat)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = sat
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
ax.set_xlabel("SAT Score Average")
ax.set_ylabel("US News and World Report Rank (National Universities)")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.711
Model:                            OLS   Adj. R-squared:                  0.709
Method:                 Least Squares   F-statistic:                     344.6
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           1.43e-39
Time:                        10:07:58   Log-Likelihood:                -670.01
No. Observations:                 142   AIC:                             1344.
Df Residuals:                     140   BIC:                             1350.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        553.3447     24.329     22.745      0.000     505.246     601.444
x1            -0.3826      0.021    -18.563      0.000      -0.423      -0.342
==============================================================================
Omnibus:                        2.199   Durbin-Watson:                   1.844
Prob(Omnibus):                  0.333   Jarque-Bera (JB):                1.725
Skew:                           0.236   Prob(JB):                        0.422
Kurtosis:                       3.263   Cond. No.                     1.25e+04
==============================================================================

Notes:
```

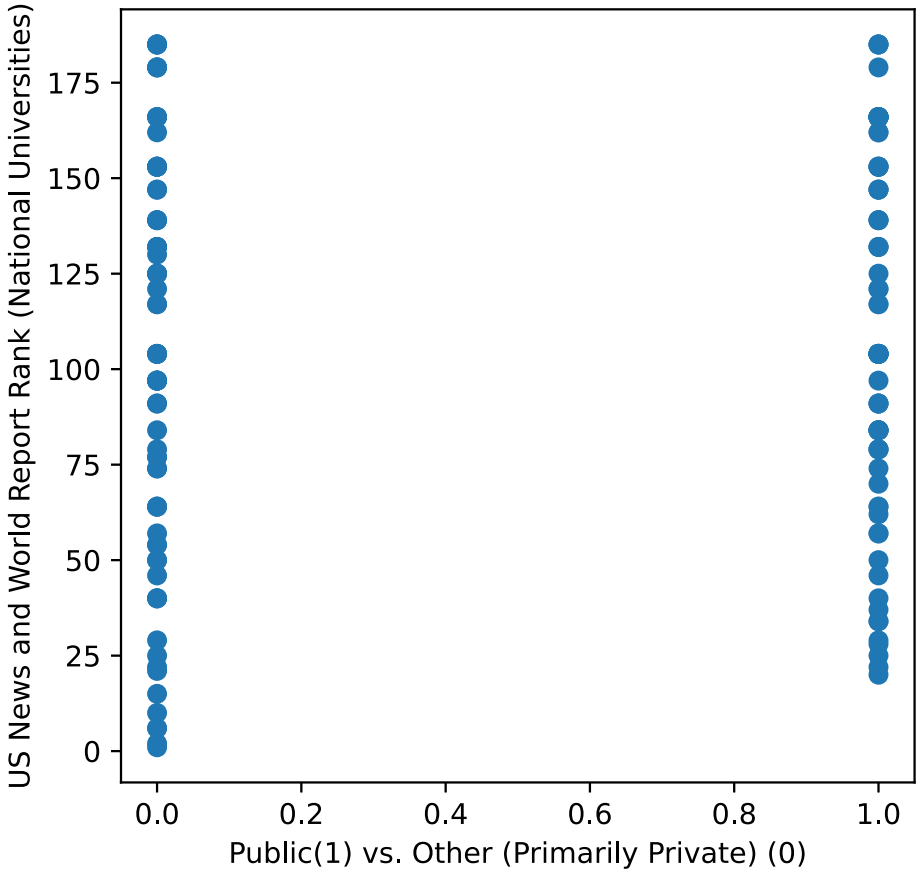
```
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.25e+04. This might indicate that there ar
strong multicollinearity or other numerical problems.
Text(0, 0.5, 'US News and World Report Rank (National Universities)')
```




```
#Predicting rank from whether a person attends public or another form of school
Y = rank
X = sm.add_constant(public)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = public
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.set_xlabel("Public(1) vs. Other (Primarily Private) (0)")
ax.set_ylabel("US News and World Report Rank (National Universities)")
```

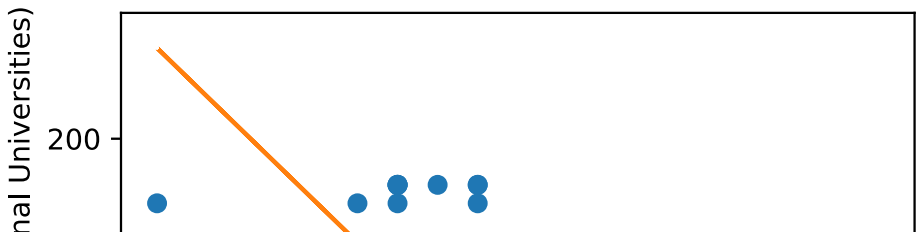
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.005
```

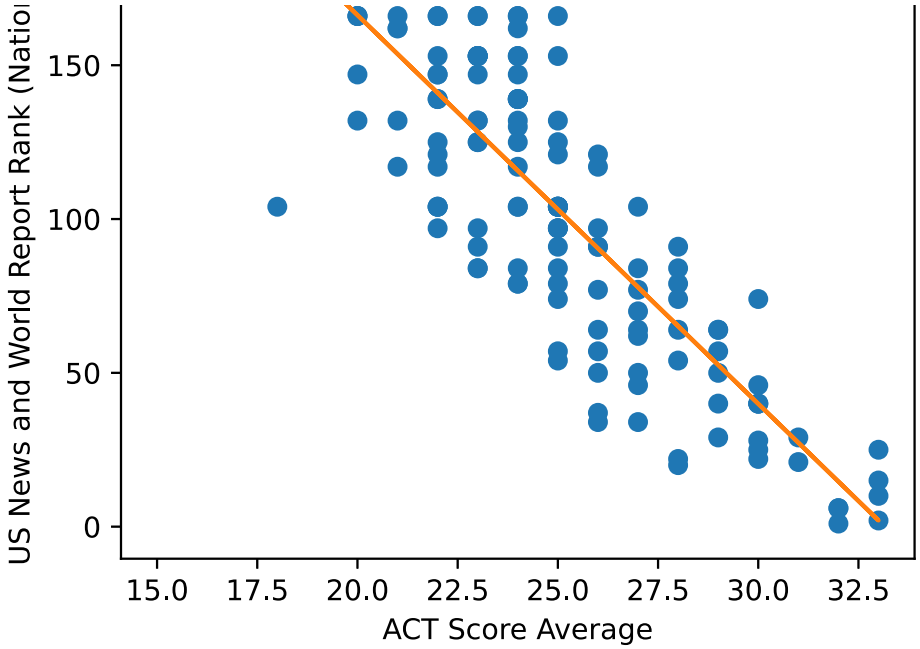
```
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                    0.7628
Date:                Tue, 22 Nov 2022   Prob (F-statistic):              0.384
Time:                        10:12:14   Log-Likelihood:                -757.78
No. Observations:                 142   AIC:                             1520.
Df Residuals:                     140   BIC:                             1525.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        100.0282      6.009     16.646      0.000      88.148     111.909
x1             7.4225      8.498      0.873      0.384      -9.379      24.224
==============================================================================
Omnibus:                       23.838   Durbin-Watson:                   1.913
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                6.657
Skew:                          -0.161   Prob(JB):                       0.0358
Kurtosis:                       1.989   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
Text(0, 0.5, 'US News and World Report Rank (National Universities)')
```



```
#Predicting rank from Average ACT Score
Y = rank
X = sm.add_constant(act)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = act
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
ax.set_xlabel("ACT Score Average")
ax.set_ylabel("US News and World Report Rank (National Universities)")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.698
Model:                            OLS   Adj. R-squared:                  0.696
Method:                 Least Squares   F-statistic:                     324.3
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           2.85e-38
Time:                        08:38:50   Log-Likelihood:                -673.04
No. Observations:                 142   AIC:                             1350.
Df Residuals:                     140   BIC:                             1356.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        419.0314     17.663     23.724      0.000     384.111     453.952
x1           -12.6366      0.702    -18.009      0.000     -14.024     -11.249
==============================================================================
Omnibus:                        1.025   Durbin-Watson:                   1.915
Prob(Omnibus):                  0.599   Jarque-Bera (JB):                1.100
Skew:                          -0.194   Prob(JB):                        0.577
Kurtosis:                       2.813   Cond. No.                         190.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
Text(0, 0.5, 'US News and World Report Rank (National Universities)')
```

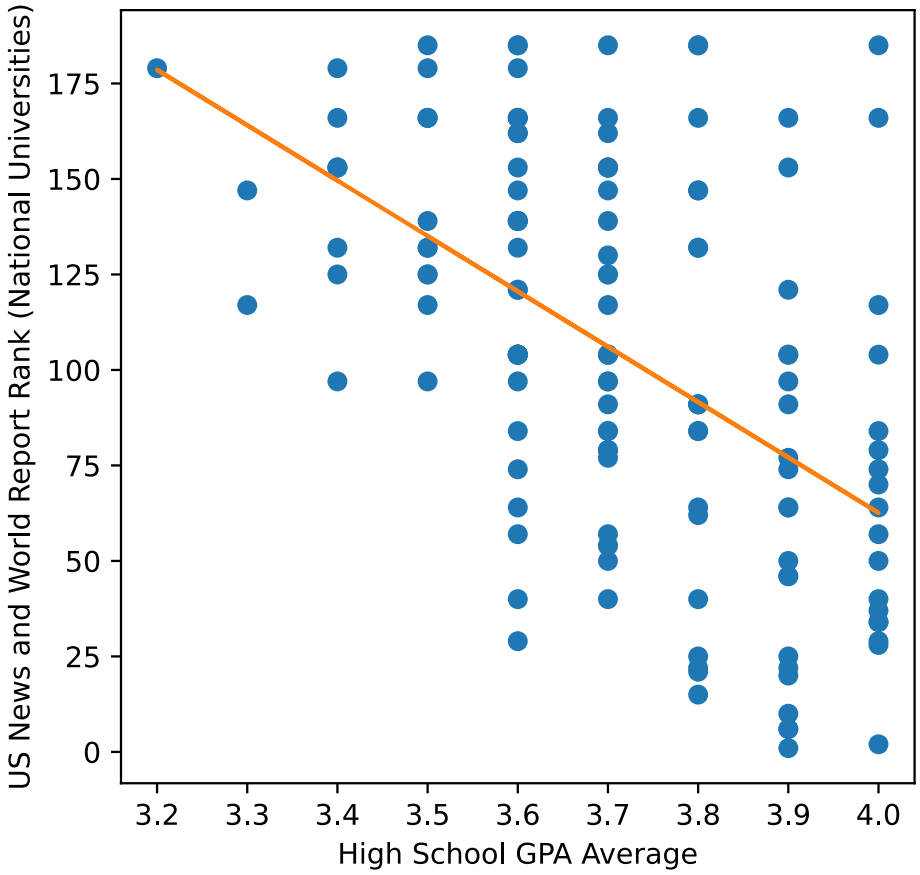





```
#Predicting rank from Average High School GPA
Y = rank
X = sm.add_constant(hs_gpa)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)

x = hs_gpa
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
ax.set_xlabel("High School GPA Average")
ax.set_ylabel("US News and World Report Rank (National Universities)")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.273
Model:                            OLS   Adj. R-squared:                  0.268
Method:                 Least Squares   F-statistic:                     52.64
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           2.50e-11
Time:                        08:45:04   Log-Likelihood:                -735.50
No. Observations:                 142   AIC:                             1475.
Df Residuals:                     140   BIC:                             1481.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```

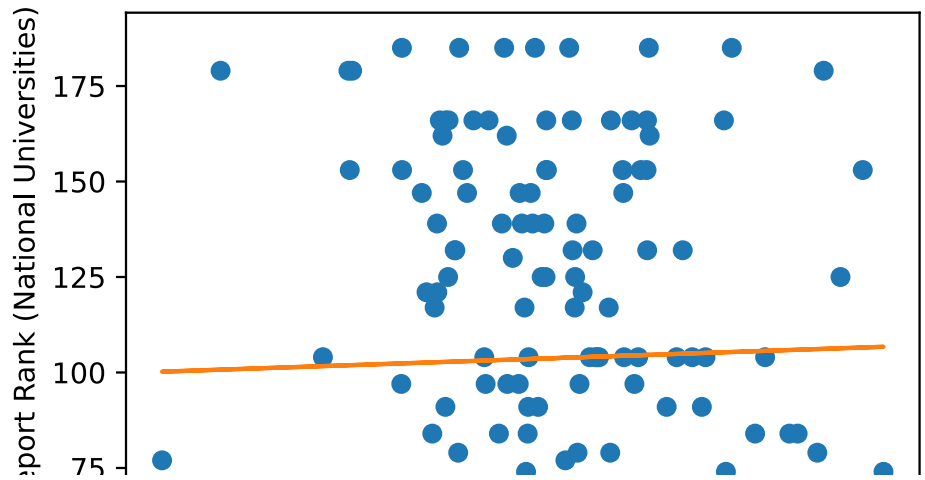
```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        642.4712     74.339      8.642      0.000     495.498     789.444
x1          -144.9686     19.980     -7.256      0.000    -184.471    -105.467
==============================================================================
Omnibus:                        1.874   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.392   Jarque-Bera (JB):                1.924
Skew:                           0.239   Prob(JB):                        0.382
Kurtosis:                       2.688   Cond. No.                         81.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
Text(0, 0.5, 'US News and World Report Rank (National Universities)')
```

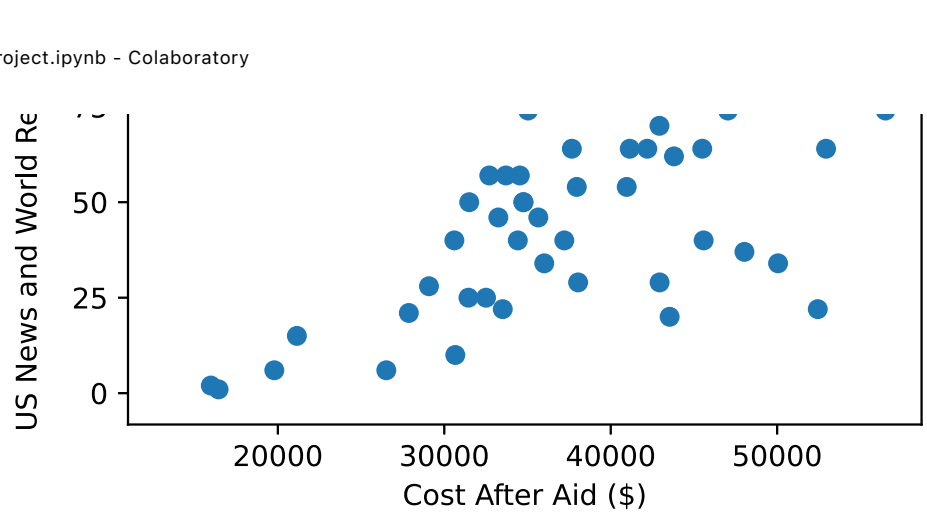



```
#Predicting rank from Average Cost After Financial Aid
Y = rank
X = sm.add_constant(cost_after_aid)
model = sm.OLS(Y,X)
results = model.fit()
print(results.summary())
b_fit,a_fit = results.params
sigma_fit = np.sqrt(results.mse_resid)
```

```
x = cost_after_aid
fig,ax = plt.subplots(figsize=(5,5))
ax.plot(x, Y, 'o', label="data")
ax.plot(x,x*a_fit + b_fit,"-", label="fit")
ax.set_xlabel("Cost After Aid ($)")
ax.set_ylabel("US News and World Report Rank (National Universities)")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.007
Method:                 Least Squares   F-statistic:                   0.08071
Date:                Tue, 22 Nov 2022   Prob (F-statistic):              0.777
Time:                        08:57:41   Log-Likelihood:                -758.13
No. Observations:                 142   AIC:                             1520.
Df Residuals:                     140   BIC:                             1526.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         98.2325     19.846      4.950      0.000      58.995     137.470
x1             0.0002      0.001      0.284      0.777      -0.001       0.001
==============================================================================
Omnibus:                       24.696   Durbin-Watson:                   1.909
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                6.795
Skew:                          -0.167   Prob(JB):                       0.0335
Kurtosis:                       1.982   Cond. No.                     1.75e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.75e+05. This might indicate that there ar
strong multicollinearity or other numerical problems.
Text(0, 0.5, 'US News and World Report Rank (National Universities)')
```






```
#After these calculations, I calculated multiple regression models. Note that
#many have a high condition number because there is multicollinearity effects
#of some of the variables being correlated with one another. Still, there are
#many conclusions to be drawn, which are discussed in the final paper. Also note
#that this is why, for example, there may appear to be strong trends in the data
#that may not be completely representative, especially since the correlation
#seems to increase with additional variables added to the model.
```

```
#Predicting rank from all 8 IV's as outlined above
Y = rank
#Form an array of values, adding a constant as well
X = sm.add_constant(np.transpose(np.array([acceptance_rate,tuition,percent_aid,sat,
model = sm.OLS(Y,X)
results = model.fit()
b,a1,a2,a3,a4,a5,a6,a7,a8 = results.params
sigma = np.sqrt(results.mse_resid)
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.784
Model:                            OLS   Adj. R-squared:                  0.771
Method:                 Least Squares   F-statistic:                     60.49
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           1.27e-40
Time:                        10:15:02   Log-Likelihood:                -649.22
No. Observations:                 142   AIC:                             1316.
Df Residuals:                     133   BIC:                             1343.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        274.6331     67.707      4.056      0.000     140.712     408.555
x1             0.6489      0.157      4.144      0.000       0.339       0.959
x2            -0.0012      0.000     -3.417      0.001      -0.002      -0.001
x3             0.4602      0.163      2.830      0.005       0.139       0.782
x4            -0.0714      0.079     -0.899      0.370      -0.229       0.086
x5           -13.4740      7.374     -1.827      0.070     -28.059       1.111
x6            -4.7390      2.514     -1.885      0.062      -9.711       0.233
x7             2.9207     15.539      0.188      0.851     -27.815      33.656
x8             0.0005      0.000      1.685      0.094   -9.53e-05       0.001
==============================================================================
Omnibus:                        2.551   Durbin-Watson:                   2.104
Prob(Omnibus):                  0.279   Jarque-Bera (JB):                2.078
Skew:                           0.210   Prob(JB):                        0.354
Kurtosis:                       3.419   Cond. No.                     1.85e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.85e+06. This might indicate that there ar
strong multicollinearity or other numerical problems.
```

```
#Predicting rank from monetary factors
Y = rank
X = sm.add_constant(np.transpose(np.array([tuition,percent_aid])))
model = sm.OLS(Y,X)
results = model.fit()
sigma = np.sqrt(results.mse_resid)
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.440
Model:                            OLS   Adj. R-squared:                  0.432
Method:                 Least Squares   F-statistic:                     54.51
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           3.33e-18
Time:                        09:08:09   Log-Likelihood:                -717.05
No. Observations:                 142   AIC:                             1440.
Df Residuals:                     139   BIC:                             1449.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        174.9275     12.483     14.014      0.000     150.247     199.608
x1            -0.0029      0.000     -9.584      0.000      -0.004      -0.002
x2             1.1033      0.164      6.735      0.000       0.779       1.427
==============================================================================
Omnibus:                        3.642   Durbin-Watson:                   2.108
Prob(Omnibus):                  0.162   Jarque-Bera (JB):                3.257
Skew:                          -0.365   Prob(JB):                        0.196
Kurtosis:                       3.136   Cond. No.                     1.60e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.6e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
#Predicting rank from test scores
Y = rank
X = sm.add_constant(np.transpose(np.array([act,sat])))
model = sm.OLS(Y,X)
results = model.fit()
sigma = np.sqrt(results.mse_resid)
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.717
Model:                            OLS   Adj. R-squared:                  0.713
Method:                 Least Squares   F-statistic:                     176.5
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           7.04e-39
Time:                        09:10:06   Log-Likelihood:                -668.42
No. Observations:                 142   AIC:                             1343.
Df Residuals:                     139   BIC:                             1352.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        510.0072     34.358     14.844      0.000     442.075     577.940
x1            -4.7383      2.673     -1.773      0.078     -10.023       0.546
x2            -0.2451      0.080     -3.056      0.003      -0.404      -0.087
==============================================================================
Omnibus:                        0.185   Durbin-Watson:                   1.871
Prob(Omnibus):                  0.912   Jarque-Bera (JB):                0.051
Skew:                           0.042   Prob(JB):                        0.975
Kurtosis:                       3.037   Cond. No.                     1.79e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 1.79e+04. This might indicate that there ar
strong multicollinearity or other numerical problems.
```

```
#Predicting rank from 3 highest individual correlations
Y = rank
X = sm.add_constant(np.transpose(np.array([sat,act,acceptance_rate])))
model = sm.OLS(Y,X)
results = model.fit()
sigma = np.sqrt(results.mse_resid)
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.737
Method:                 Least Squares   F-statistic:                     132.5
Date:                Tue, 22 Nov 2022   Prob (F-statistic):           1.92e-40
Time:                        09:00:35   Log-Likelihood:                -661.91
No. Observations:                 142   AIC:                             1332.
Df Residuals:                     138   BIC:                             1344.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        360.7060     52.596      6.858      0.000     256.707     464.705
x1            -0.1444      0.082     -1.767      0.080      -0.306       0.017
x2            -4.8918      2.562     -1.909      0.058      -9.958       0.175
x3             0.6047      0.166      3.641      0.000       0.276       0.933
==============================================================================
Omnibus:                        1.731   Durbin-Watson:                   1.894
Prob(Omnibus):                  0.421   Jarque-Bera (JB):                1.453
Skew:                           0.245   Prob(JB):                        0.484
Kurtosis:                       3.072   Cond. No.                     2.85e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
[2] The condition number is large, 2.85e+04. This might indicate that there ar
strong multicollinearity or other numerical problems.
```

Colab paid products - Cancel contracts here