JUST THE FACT(OR)S, MA'AM

AN ANALYSIS OF SOME FACTORS WHICH DETERMINE POST-GRADUATION SALARY

BACKGROUND

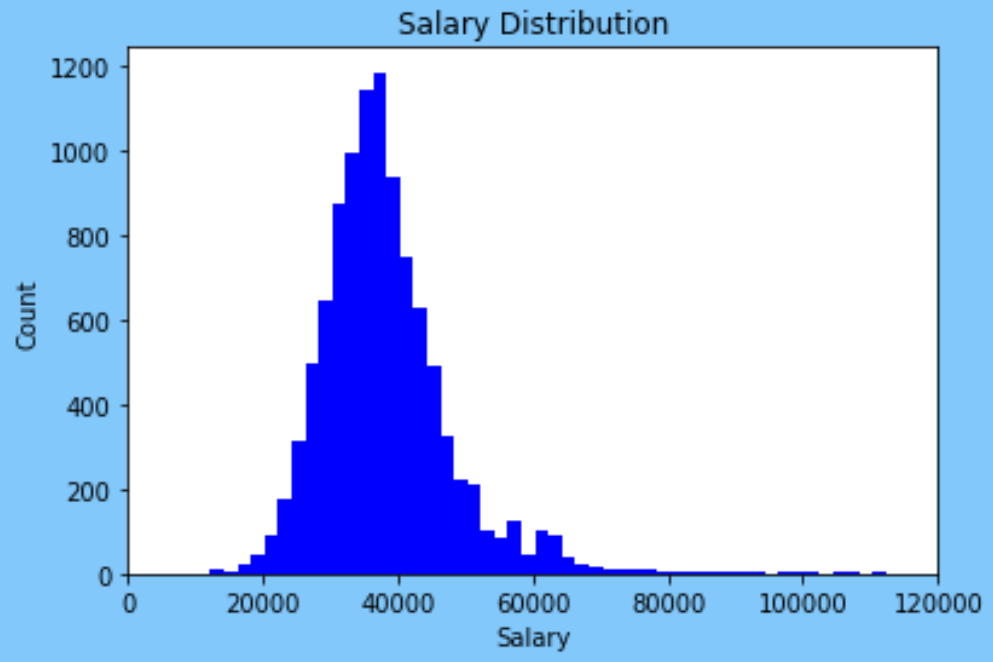
- The cost of a college education increased roughly 67% between 2002 and 2012.
- During that same period, wages decreased approximately 5%.
- Given this, we can ask: How do we make the most of our investment?
- In this presentation, I will provide a high-level overview of the relationship between salary and the following predictors: tuition, pre-admission placement score, geographic region, programs offered, and whether the institution is public or private.

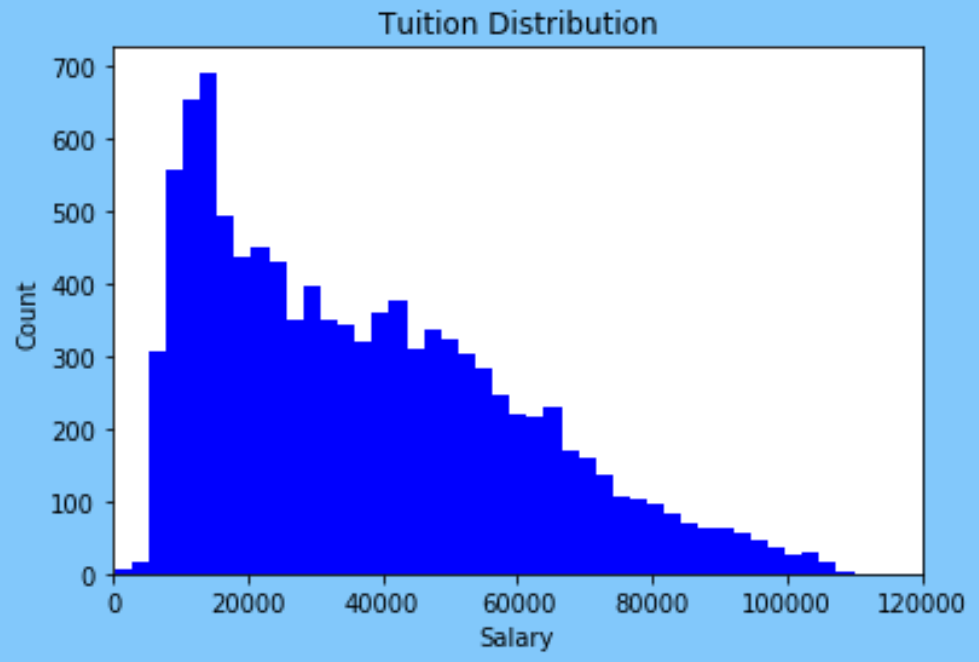
EXPLORATORY DATA ANALYSIS

- Let's begin with some cursory data exploration:
  - After loading the data into a Pandas DataFrame, I used the shape attribute and the describe() and info() methods to determine that:
    - The original dataset consisted of approximately 140,000 records with roughly 8,000 fields per record, of which a significant amount of data consisted of null values.
    - The salary field contained string values used to indicate that the data was suppressed to obfuscate personally identifiable information.
    - The ACT and SAT fields were missing a significant amount of data.
  - Once the data was cleansed, roughly 10,000 records remained with 9 fields per record.

EXPLORATORY DATA ANALYSIS

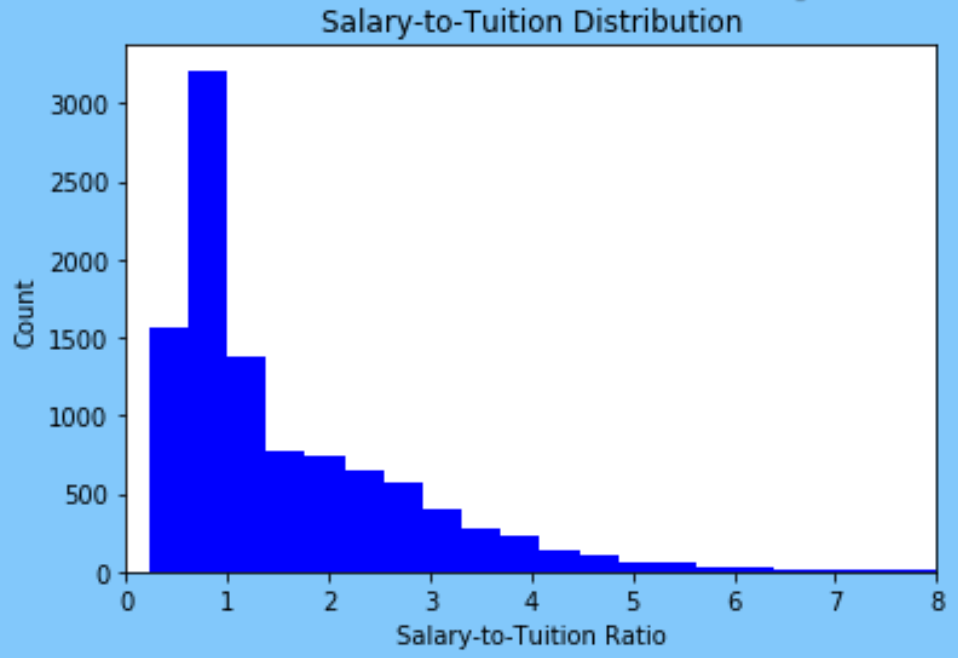
- Let's examine the salary and tuition distributions:





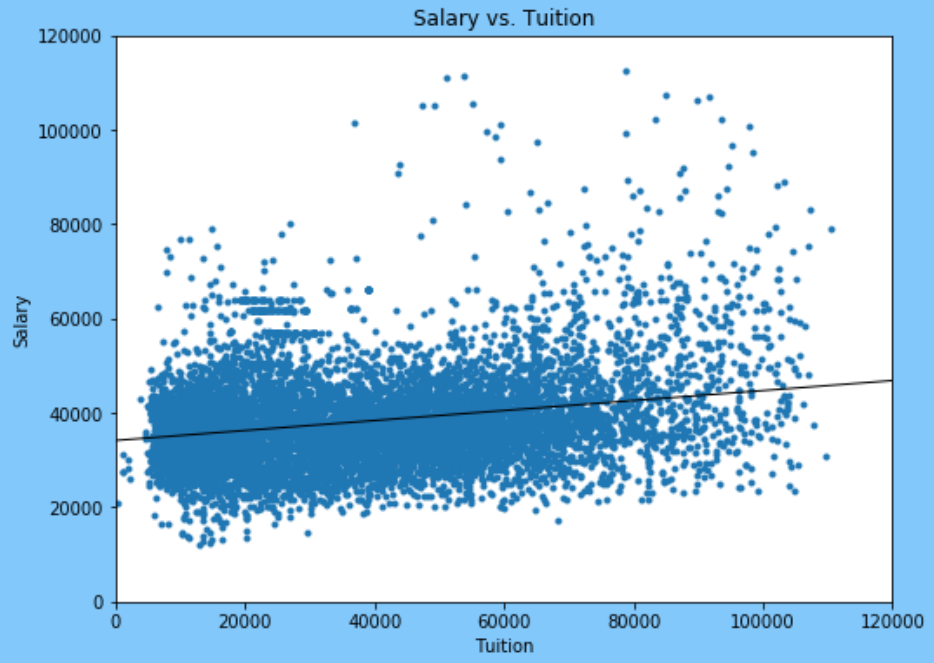
EXPLORATORY DATA ANALYSIS

- Then let's examine the salary-to-tuition ratio distribution.
- Note here that 47% of institutions are below the salary-to-tuition ratio value of 1.



EXPLORATORY DATA ANALYSIS

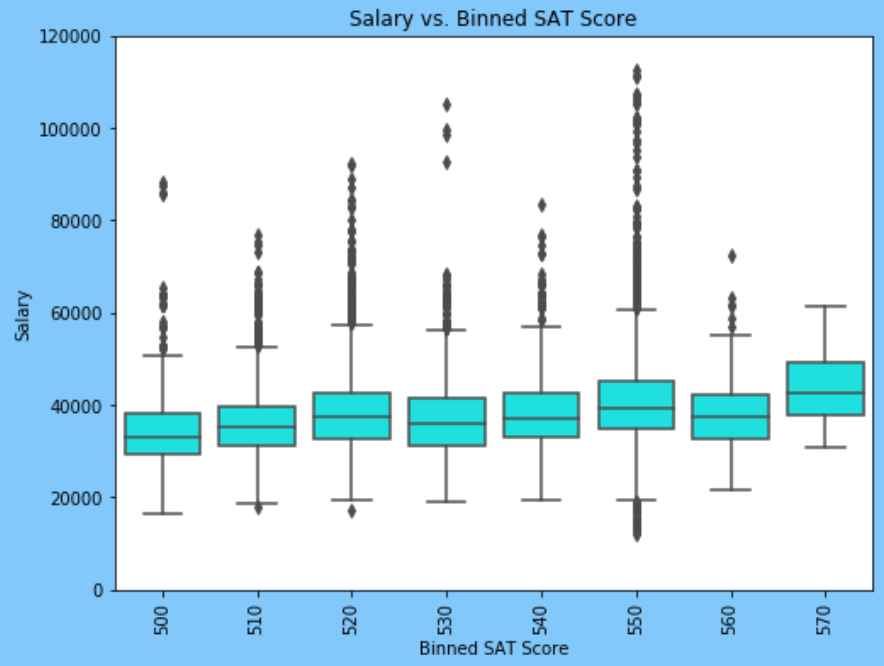
- We can then ask, *How does salary vary with tuition?*

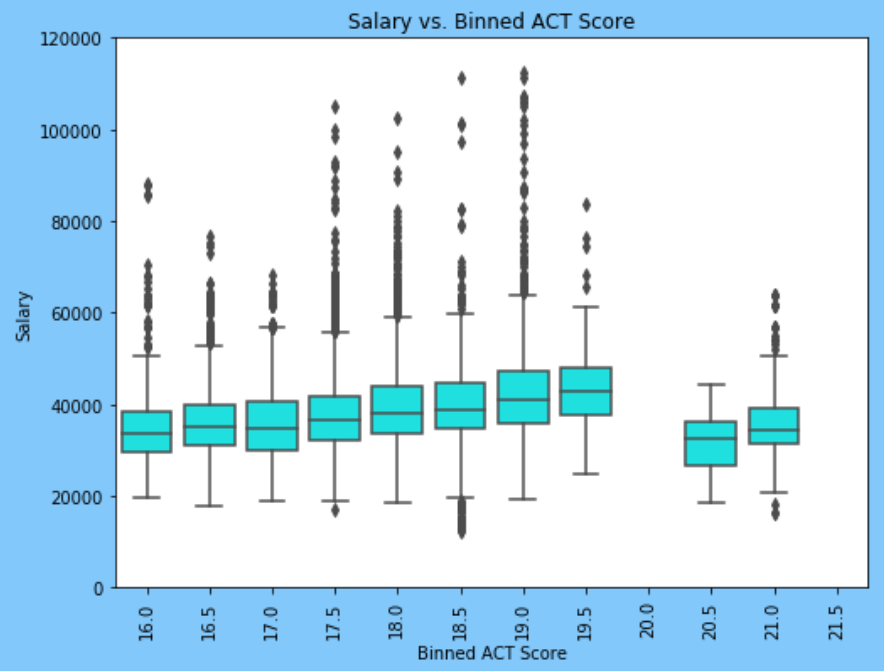


Slope	0.106
Intercept	34,157
R-squared	0.062
Pearson correlation coefficient	0.250
p-value	< 0.01

EXPLORATORY DATA ANALYSIS

- And, *How does salary varies with placement score?*





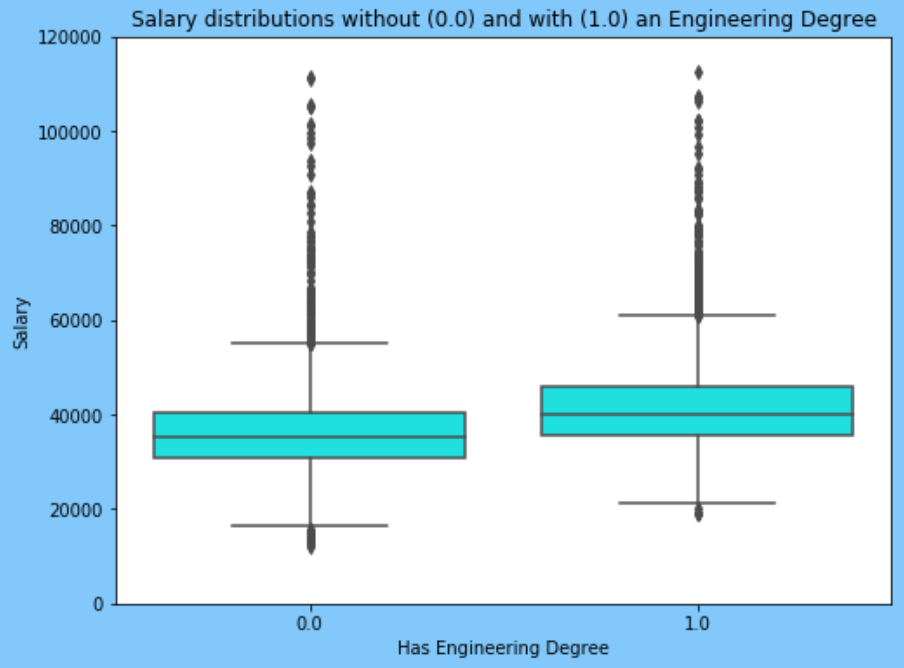
EXPLORATORY DATA ANALYSIS

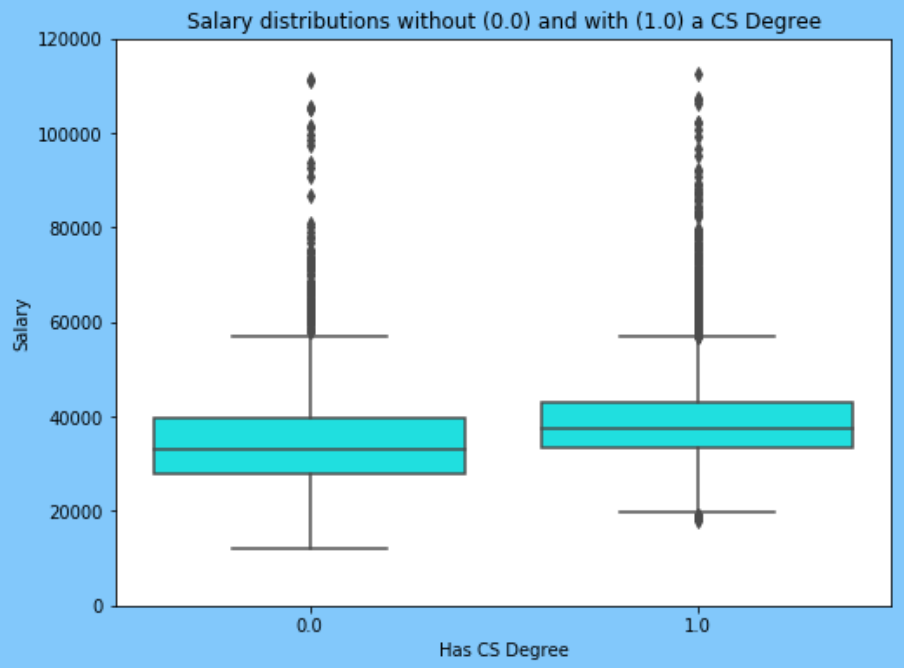
- And continue with, *How does salary varies with placement score?*

Sample set	Pearson correlation coefficient	p-value
Salary vs. Mean SAT score	0.202	< 0.01
Salary vs. Mean ACT score	0.104	< 0.01

EXPLORATORY DATA ANALYSIS

- We can then ask, *How does salary depend on the programs offered.*





EXPLORATORY DATA ANALYSIS

- And we continue by reviewing how salary depends on programs offered (cont.).

Sample set	Difference in means	Margin of error	p-value
Engineering	5436	419	< 0.01
Computer Science	3777	487	< 0.01

MACHINE LEARNING

- We chose a linear regression model because it allows us to show a relation between a dependent variable (label) and one or more independent variables (features).
- Used OLS to fit the data. The fit produced an R-squared = 0.179.

Feature	Coefficient	Coefficient type	p-value
Tuition	0.079	Multiplier	< 0.01
Mean ACT Score	216	Multiplier	0.011
Mean SAT Score	39	Multiplier	< 0.01
Private Institution	9633	Addend	< 0.01
Public Institution	8591	Addend	< 0.01
Region – New England	-1184	Addend	< 0.01
Region – Mid Eastern	830	Addend	0.015
Region – Southwest	-1333	Addend	< 0.01
Engineering program offered	5018	Addend	< 0.01
Computer Science program offered	2801	Addend	< 0.01

CONCLUSION

- Selected a dataset and explored it.
- Used ordinary least squares to model the relationship between the salary (label) and the tuition, etc. (features).
- Based on the model, we recommended that a student attend a private school in the Mid-Atlantic region which offers both an engineering degree and a computer science degree. This would offer the best opportunity, above and beyond solid placement test scores, to attain a robust salary.
- This model is limited. What about other factors (i.e. family of origin, gender, ethnicity, or even height)?
- This work requires more data and more analysis. For example, a complete set of placement scores, more detailed salary and tuition data, and perhaps even information at the program level would be quite beneficial. Also, examining how the data behaves for public vs. private institutions or how the data changes with respect to time could also provide more insights.