# The Variables to Success

Che Guan, Judge Hiciano, Nicholas Lee, Tatianna Martinez
*School of Information, University of California Berkeley*
*W200, Introduction to Data Science Programming*
*Thursday 630pm, Spring 2022*

# Research Questions

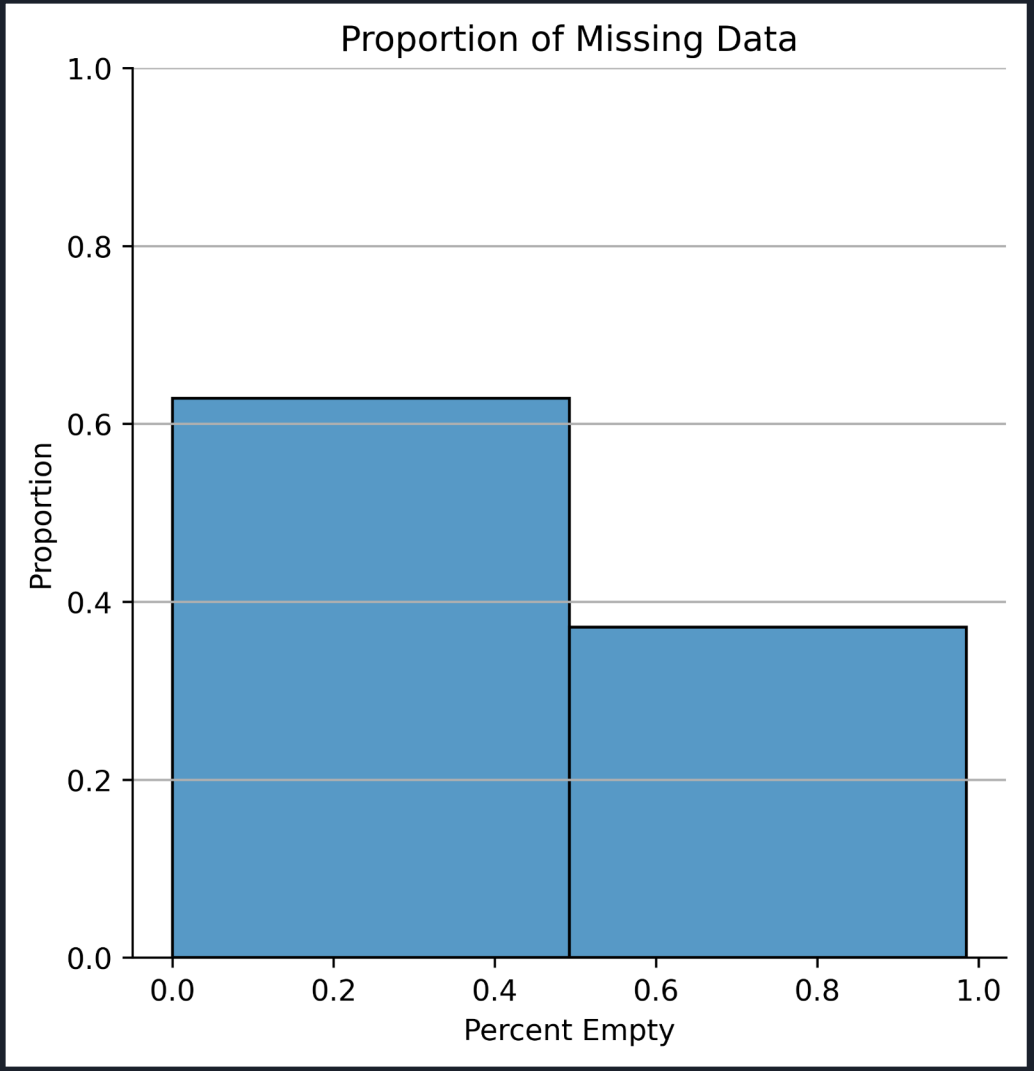
- Do students have higher earnings if their degree program was one of the top five degree programs for 2019?
- What top ten features are important in predicting higher earnings for students post graduation for 2019?

# Data Cleansing

- 24 CSV files from school years 1997 - 2020
- Transformation:
  - (170026, 2990) -> (48477, 175)
- From certificate to Graduate degrees

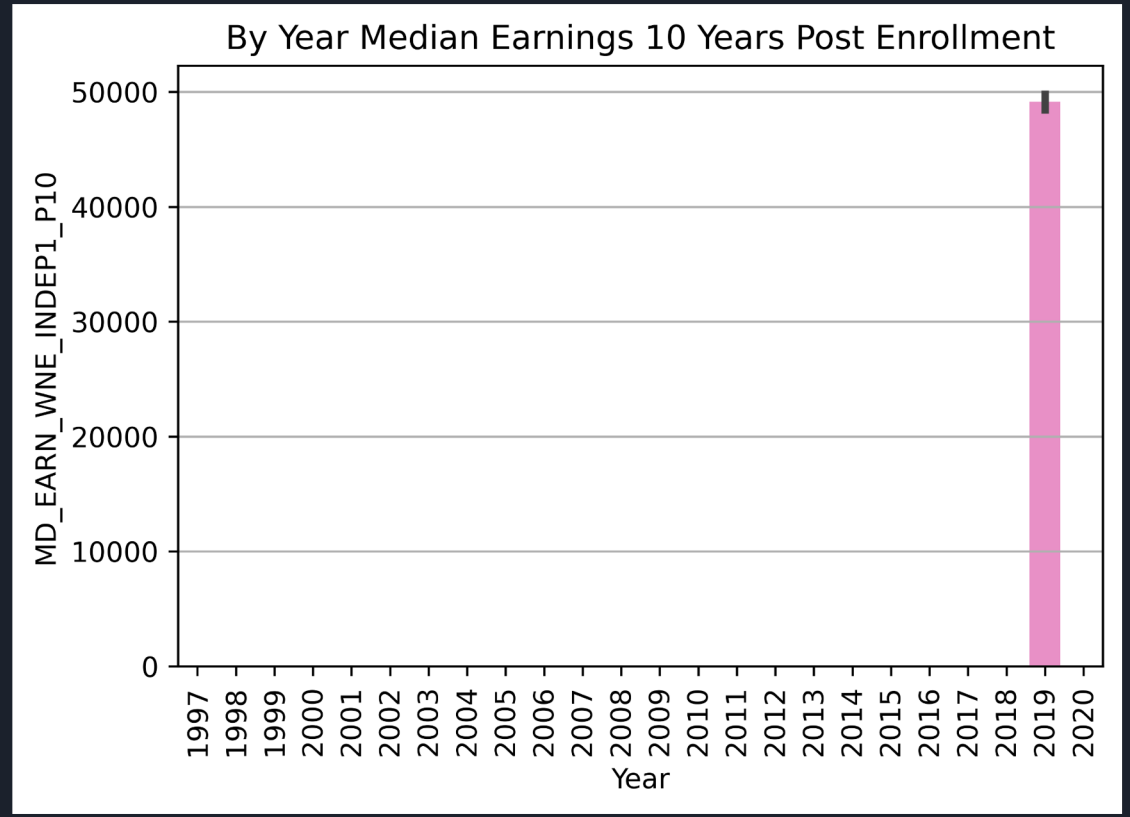
# Limitations

- ~40% of the columns have more than half of the data missing.
- Privacy Suppressed



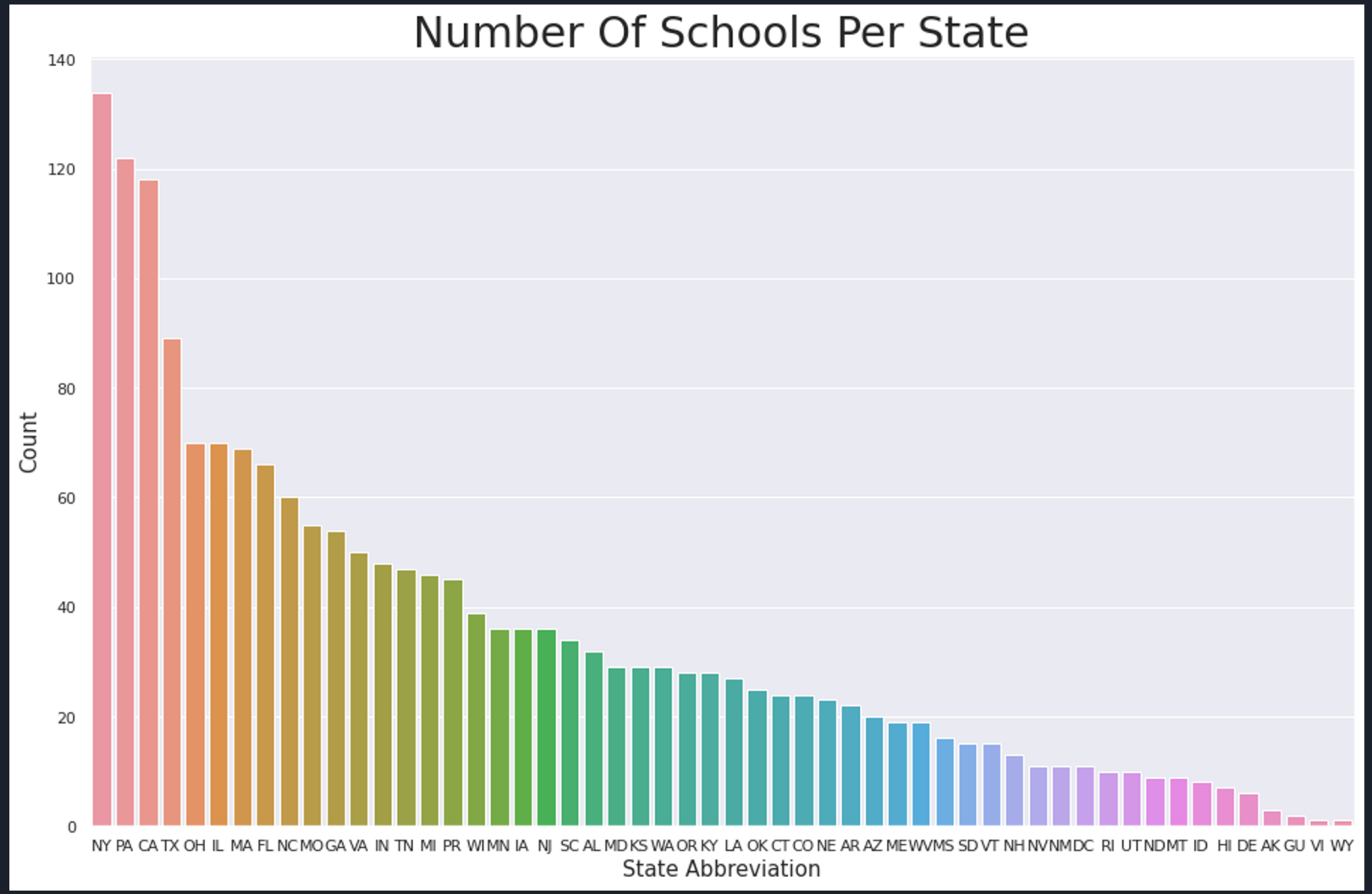
# Limitations

- Variables added as years passed
  - Median earnings of independent students working and not enrolled 10 years after entry (MD_EARN_WNE_INDEP1_P10)

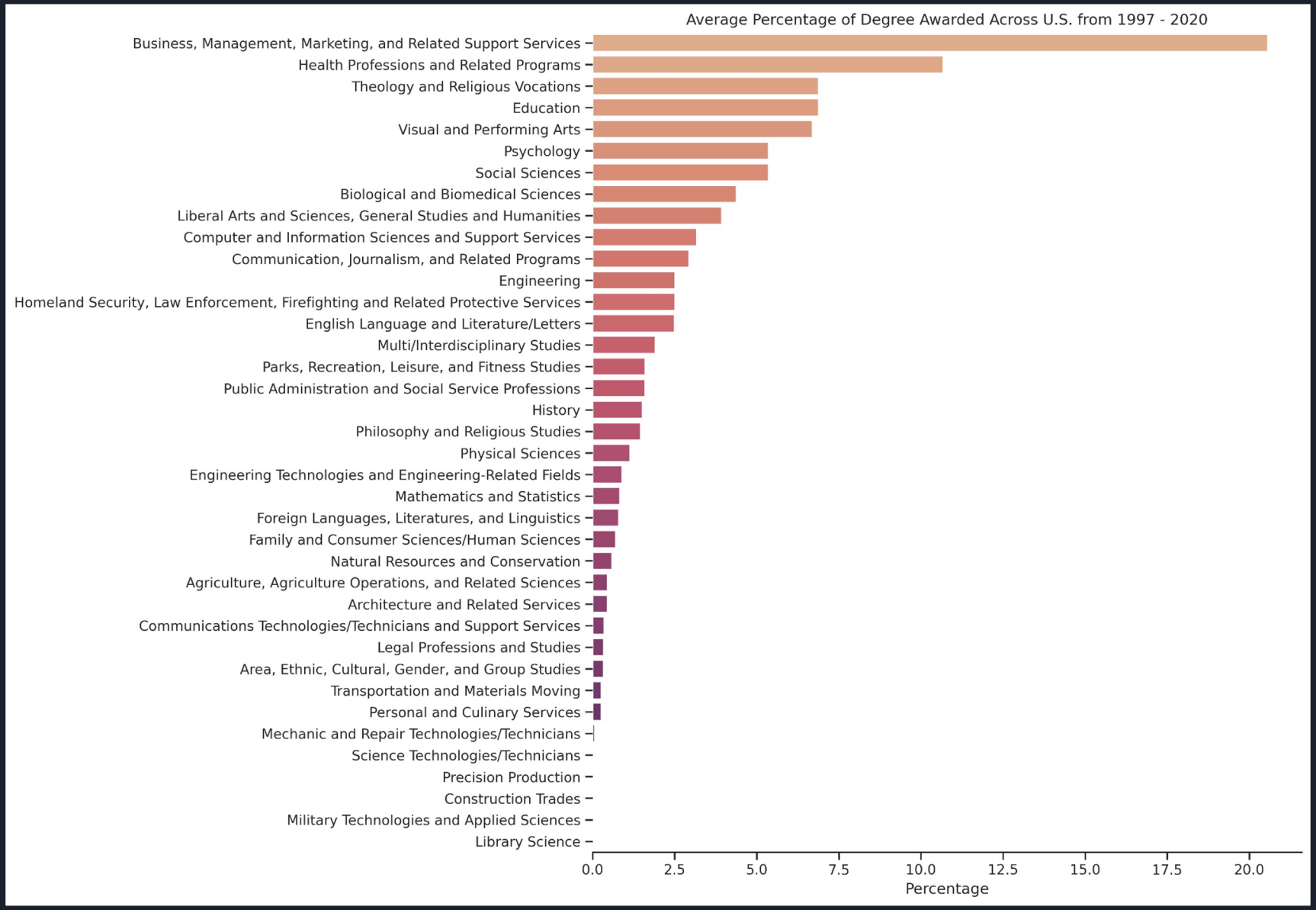


# Exploratory Data Analysis

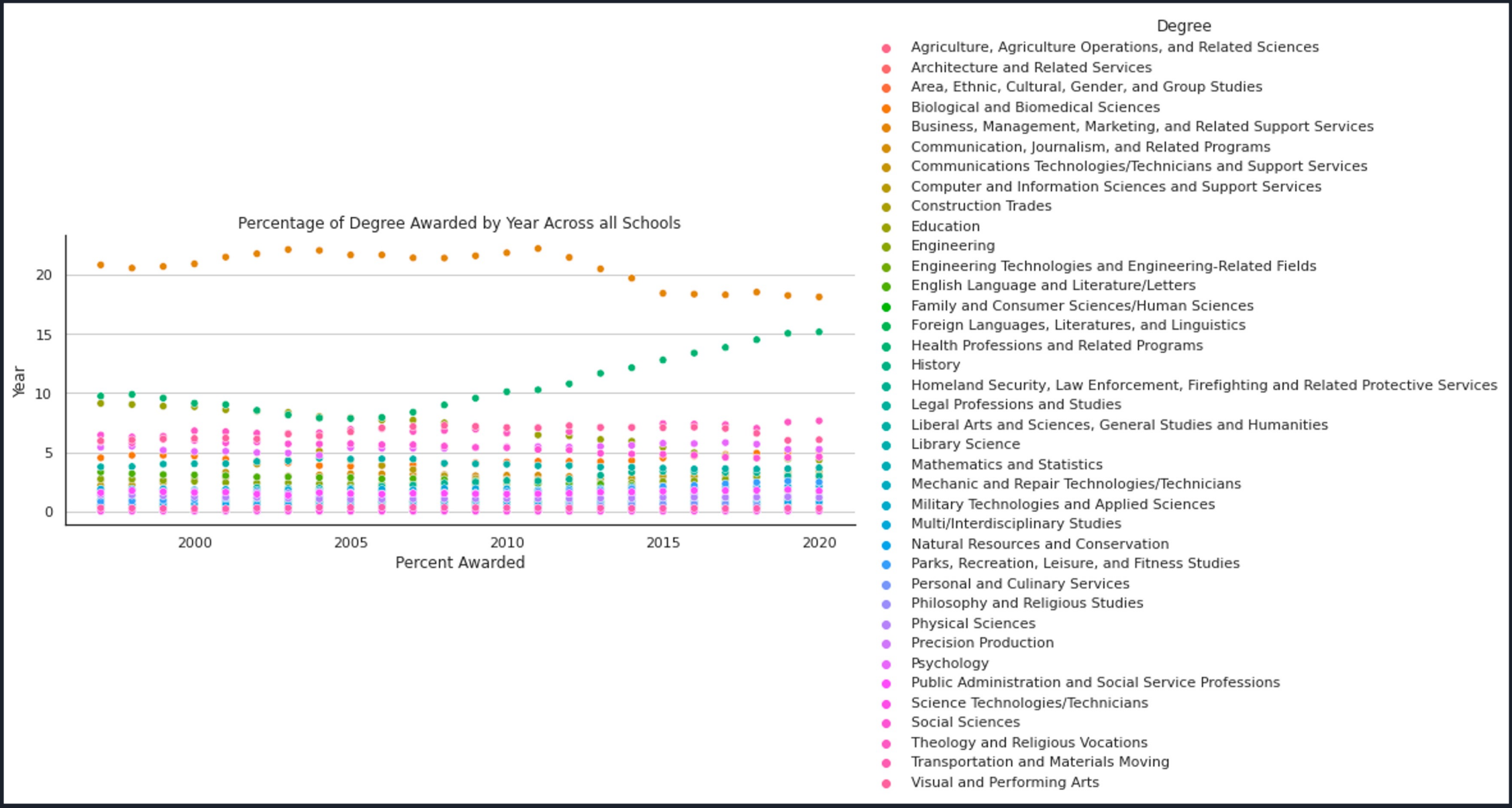
- Dataset consist of self reported information from 2,780 universities and colleges around the United States from 1997-2020
- The dataset does include US territories: Puerto Rico, Guam, and the Virgin islands which we left in the data set.
- The dataset itself has evolved over time but has never dropped any columns, so there were many columns that had Null values.
- Based on the question we are exploring, there were only 71 entries that had null values for the columns that we identified we needed to investigate our question.

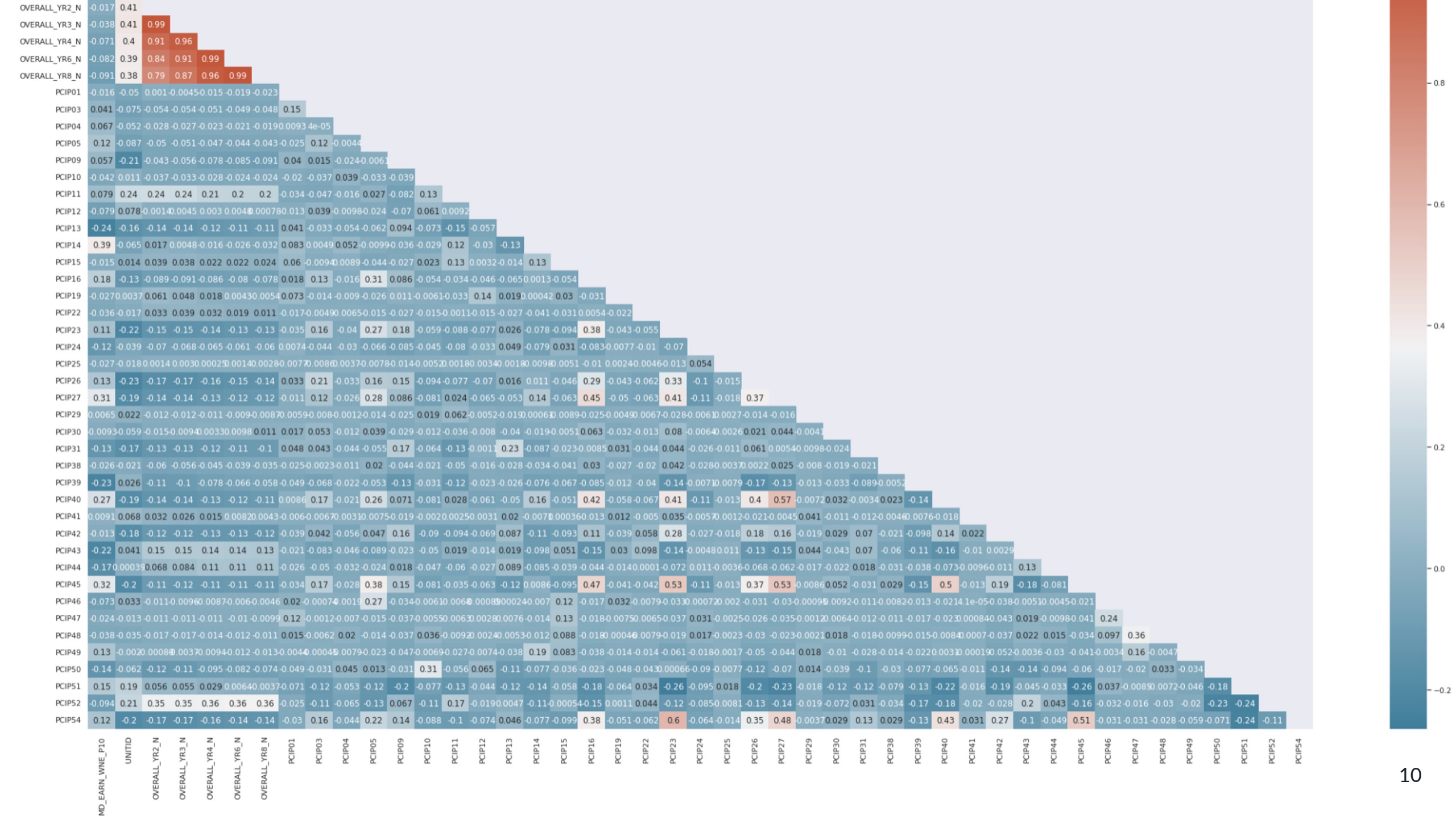


# Exploratory Data Analysis



# Exploratory Data Analysis

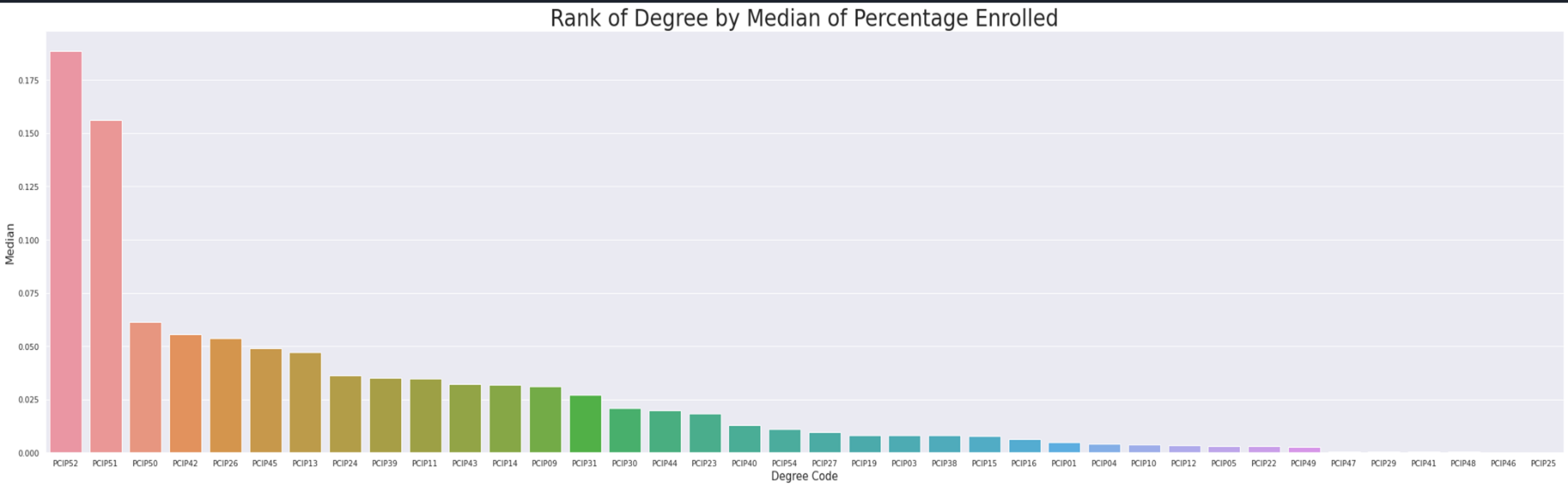



MD_EARN_WNE_P10
UNITID
OVERALL_YR2_N
OVERALL_YR3_N
OVERALL_YR4_N
OVERALL_YR6_N
OVERALL_YR8_N
PCIP01
PCIP03
PCIP04
PCIP05
PCIP09
PCIP10
PCIP11
PCIP12
PCIP13
PCIP14
PCIP15
PCIP16
PCIP19
PCIP22
PCIP23
PCIP24
PCIP25
PCIP26
PCIP27
PCIP29
PCIP30
PCIP31
PCIP38
PCIP39
PCIP40
PCIP41
PCIP42
PCIP43
PCIP44
PCIP45
PCIP46
PCIP47
PCIP48
PCIP49
PCIP50
PCIP51
PCIP52
PCIP54
0.8
0.6
0.4
0.2
0.0
−0.2

# Exploratory Data Analysis

Top 5 degree programs by the median of students enrolled compared to total student body are:

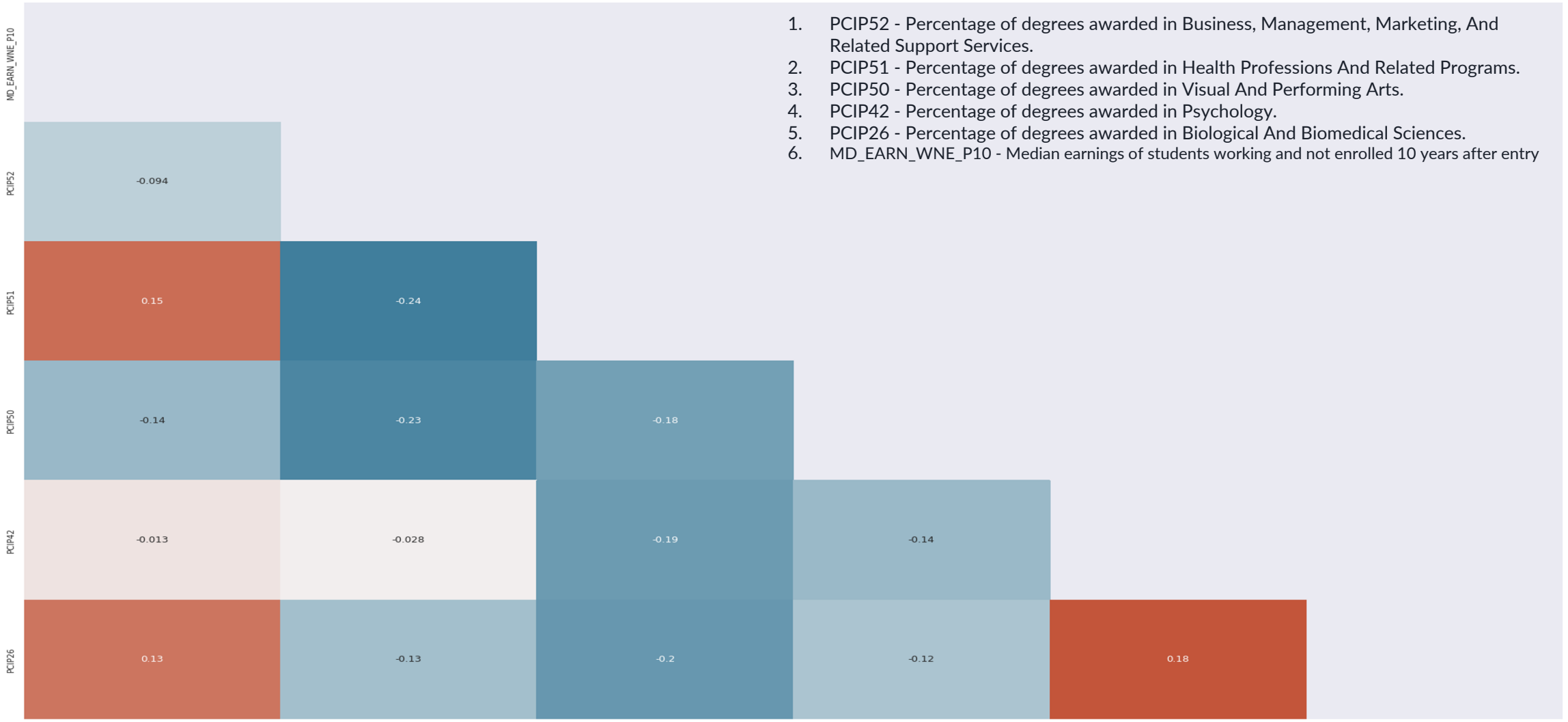
1. PCIP52 - Percentage of degrees awarded in Business, Management, Marketing, And Related Support Services.
2. PCIP51 - Percentage of degrees awarded in Health Professions And Related Programs.
3. PCIP50 - Percentage of degrees awarded in Visual And Performing Arts.
4. PCIP42 - Percentage of degrees awarded in Psychology.
5. PCIP26 - Percentage of degrees awarded in Biological And Biomedical Sciences.



# Exploratory Data Analysis

## Correlation of top 5 Programs and Median Earning

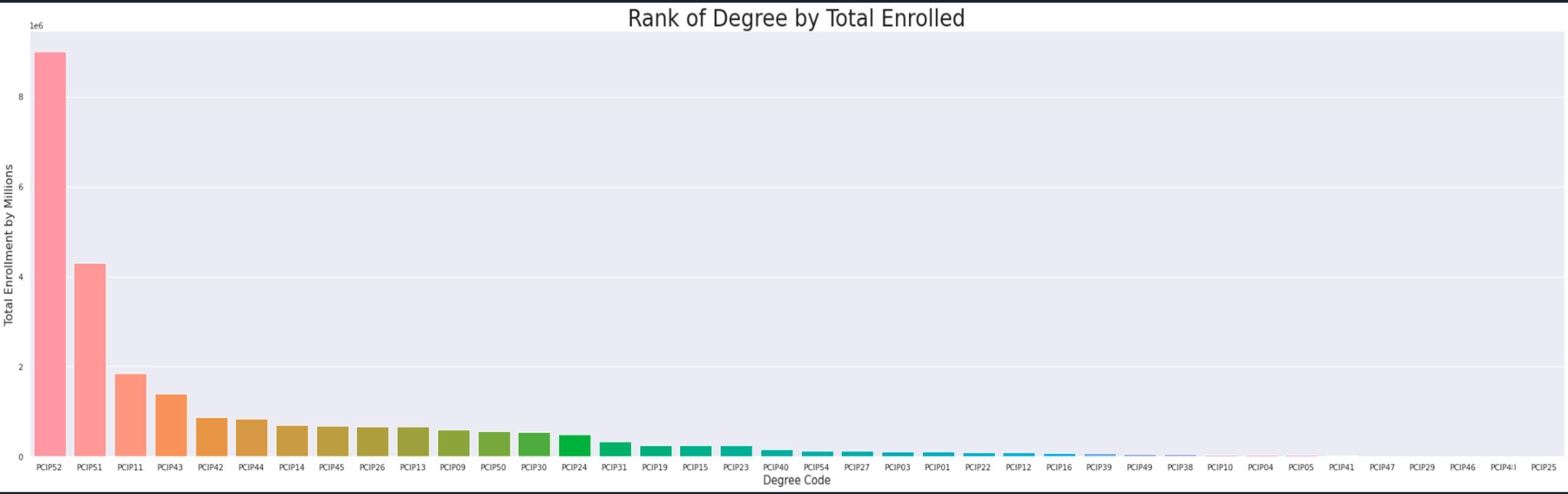
1. PCIP52 - Percentage of degrees awarded in Business, Management, Marketing, And Related Support Services.
2. PCIP51 - Percentage of degrees awarded in Health Professions And Related Programs.
3. PCIP50 - Percentage of degrees awarded in Visual And Performing Arts.
4. PCIP42 - Percentage of degrees awarded in Psychology.
5. PCIP26 - Percentage of degrees awarded in Biological And Biomedical Sciences.
6. MD_EARN_WNE_P10 - Median earnings of students working and not enrolled 10 years after entry



# Exploratory Data Analysis

Top 5 degree programs by enrollment  are:

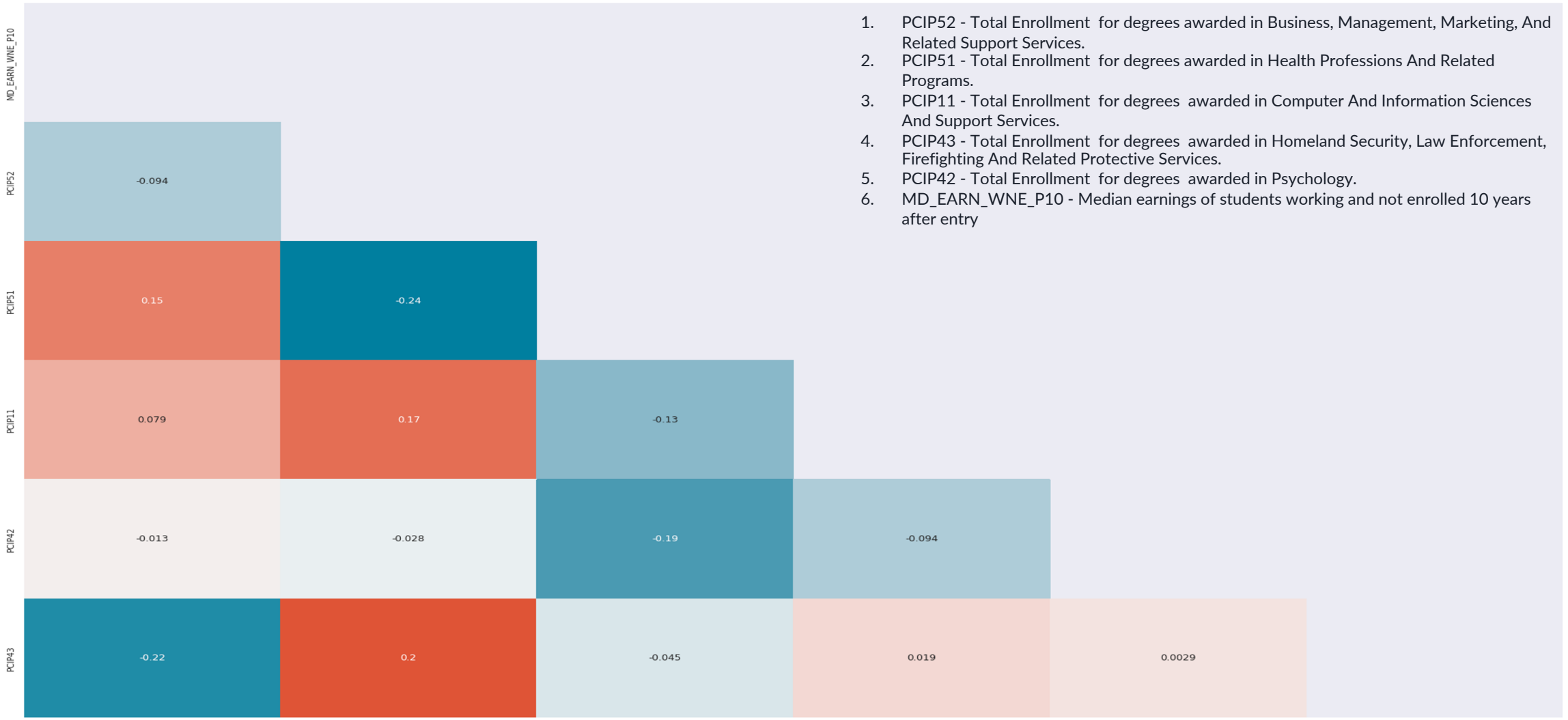
1. PCIP52 - Total Enrollment  for degrees awarded in Business, Management, Marketing, And Related Support Services.
2. PCIP51 - Total Enrollment  for degrees awarded in Health Professions And Related Programs.
3. PCIP11 - Total Enrollment  for degrees  awarded in Computer And Information Sciences And Support Services.
4. PCIP43 - Total Enrollment  for degrees  awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.
5. PCIP42 - Total Enrollment  for degrees  awarded in Psychology.



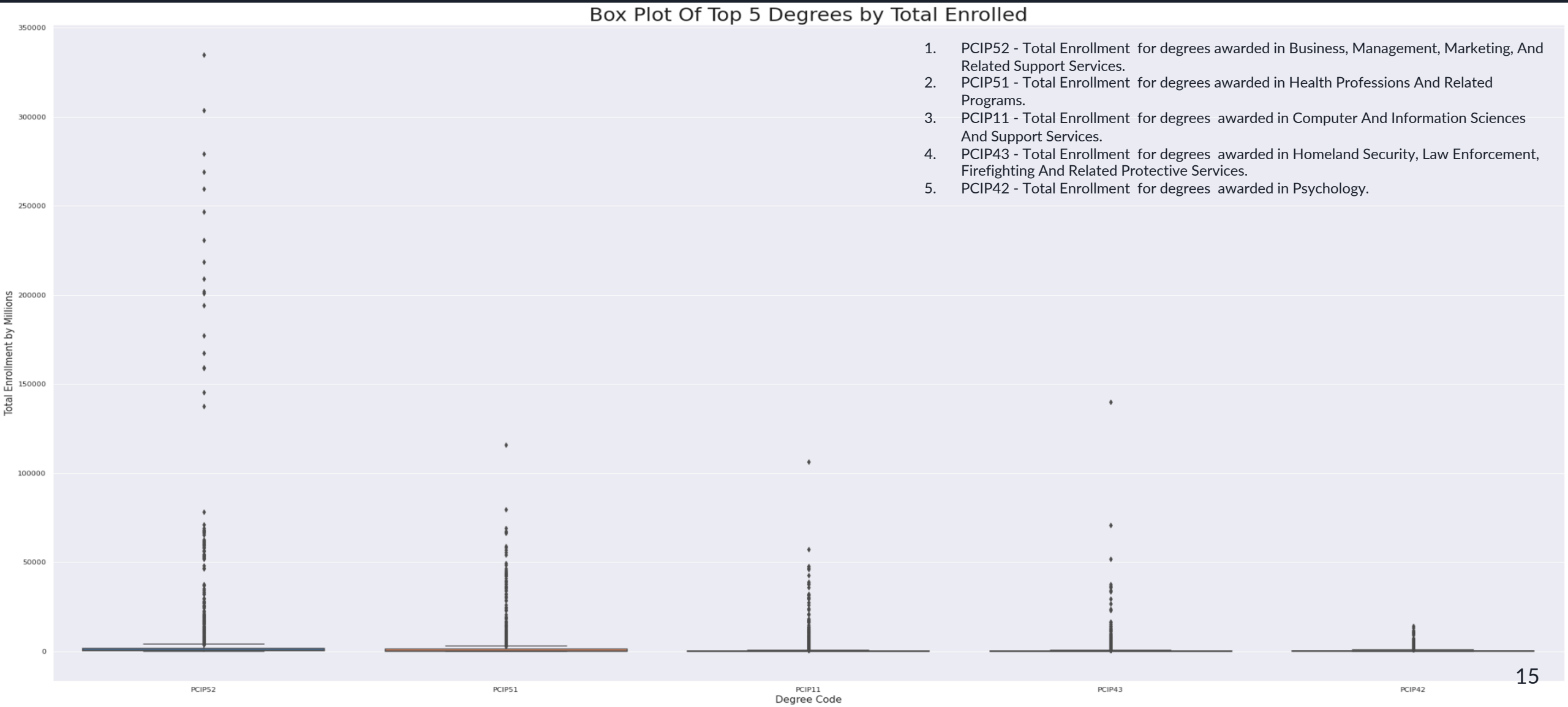
# Exploratory Data Analysis

## Correlation of top 5 programs and median earning

1. PCIP52 - Total Enrollment for degrees awarded in Business, Management, Marketing, And Related Support Services.
2. PCIP51 - Total Enrollment for degrees awarded in Health Professions And Related Programs.
3. PCIP11 - Total Enrollment for degrees awarded in Computer And Information Sciences And Support Services.
4. PCIP43 - Total Enrollment for degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.
5. PCIP42 - Total Enrollment for degrees awarded in Psychology.
6. MD_EARN_WNE_P10 - Median earnings of students working and not enrolled 10 years after entry



# Exploratory Data Analysis

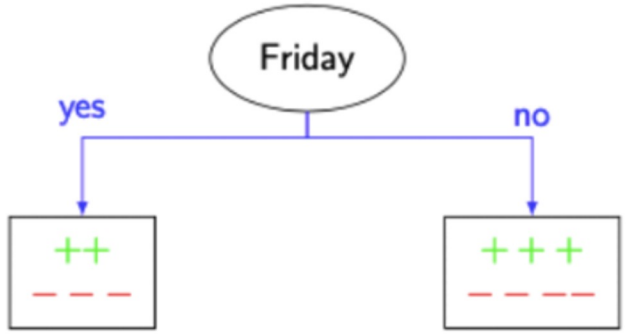



1. PCIP52 - Total Enrollment for degrees awarded in Business, Management, Marketing, And Related Support Services.
2. PCIP51 - Total Enrollment for degrees awarded in Health Professions And Related Programs.
3. PCIP11 - Total Enrollment for degrees awarded in Computer And Information Sciences And Support Services.
4. PCIP43 - Total Enrollment for degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.
5. PCIP42 - Total Enrollment for degrees awarded in Psychology.

## Summary

- Create a tree with maximum depth
  - Either until every leaf is a single data point
  - Or use all features
- Pick a subtree (a node and all leaves)
- Aggregate that leaves all the way to the node
- Compute new error
  - Misclassification
  - Entropy
- Pruning is expensive
- Pruning is counter-intuitive
  - Fixing a bad model
- How about directly building a better model

## Decision Tree




Loss: 0.97928

- $H(S_1) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = .97$
- $H(S_2) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = .99$
- Entropy of Split: .

$$.97 * \frac{5}{12} + .99 * \frac{7}{12} = .979$$

- $IG = H(Y) - .979 = 1 - .929 = .021$
- *Not a great split.*

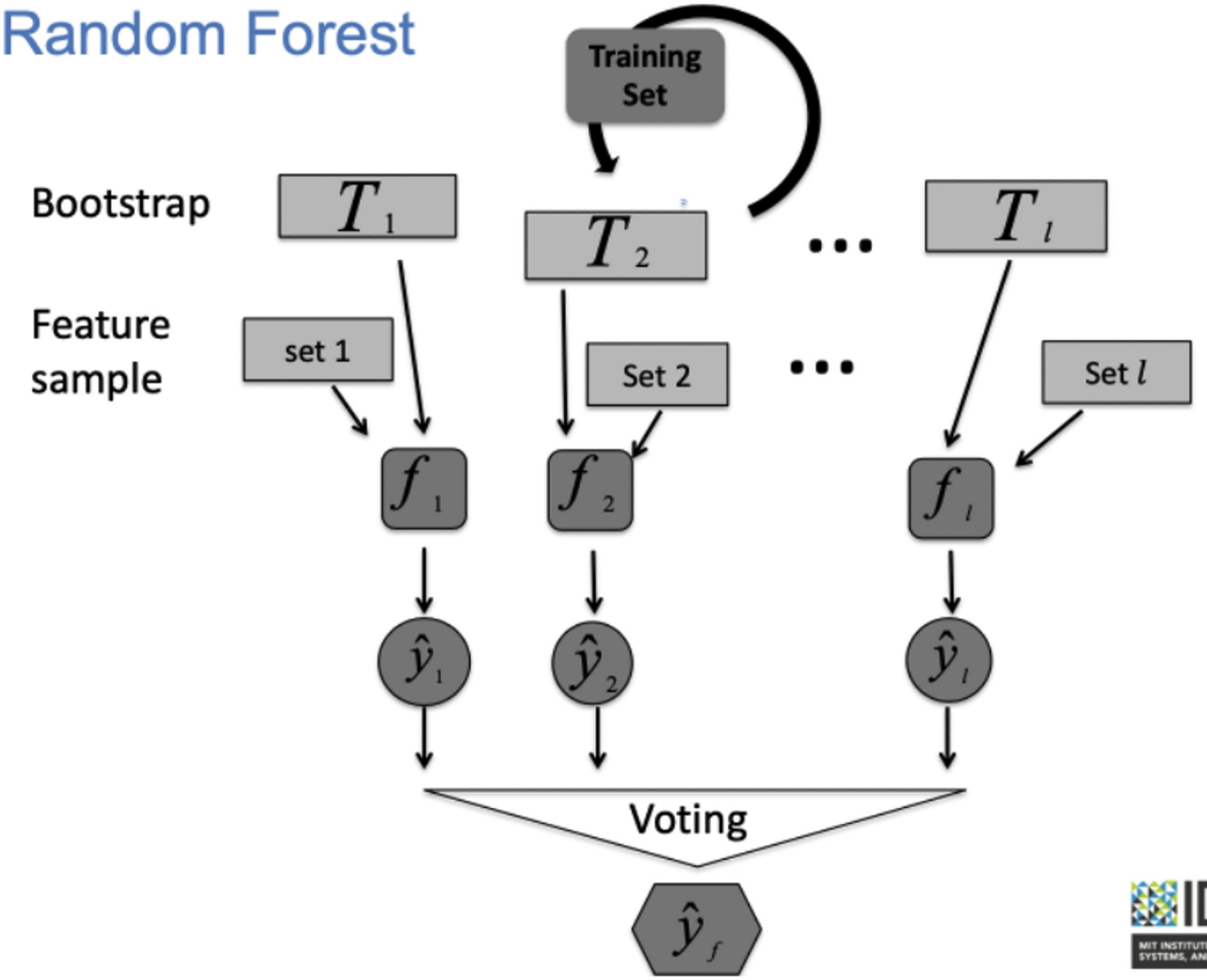
## Ensemble Learning/Random Forest

- Ensemble Methods are the key idea behind Random Forests
- Motivated by averaging techniques
- You can reduce the variance if you average a number of independent RVs

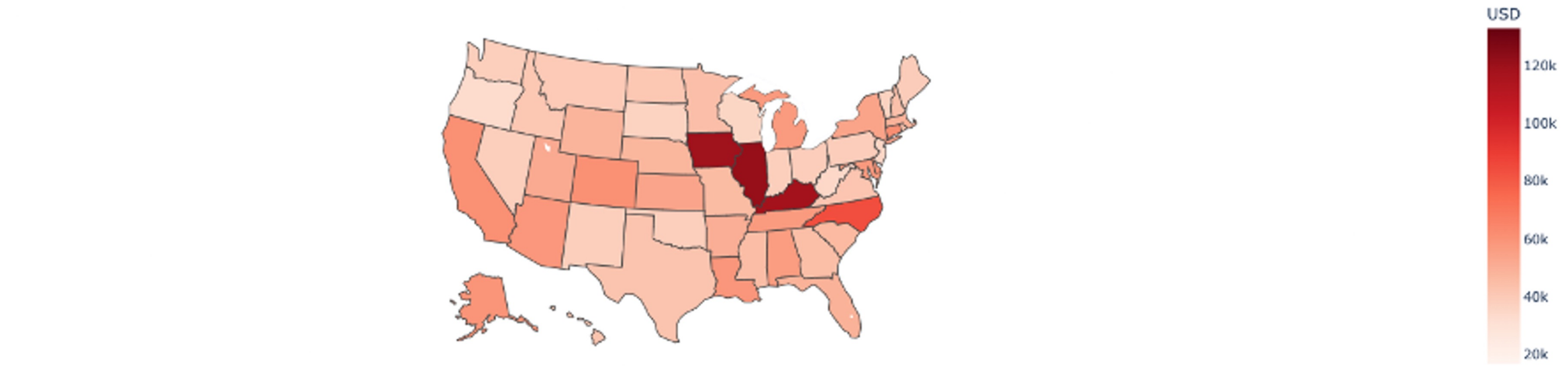
## Summary: Bagging

- Bagging= Bootstrap + Aggregation
- Bootstrap: sample with replacement
- Build a tree classifier with each sample
- Aggregate through majority voting

# Random Forest



# Students’ Median Earning by State



# Models and Performance

- Feature Engineering:
  - Categorical variables:
    - Group school level data by regions and impute missing values with the mode
    - Convert each categorical variable into dummy
  - Continuous variables:
    - Group data by school and year, and impute missing with the mean
    - Convert continuous variables into categorical ones by creating bins
- Target (Binary) - MD_EARN_WNE_P10 (Median earnings of students working and not enrolled 10 years after entry)
  - 1, if MD_EARN_WNE_P10> median
  - 0, else
- Random forest is more accurate than single decision trees, but it is not as comprehensible as decision trees.
- Important features from both models include:
  - Average cost of attendance
  - Enrollment of degree-seeking students and diversity of students
  - Average net price of family income
  - Degree programs
  - Type of school

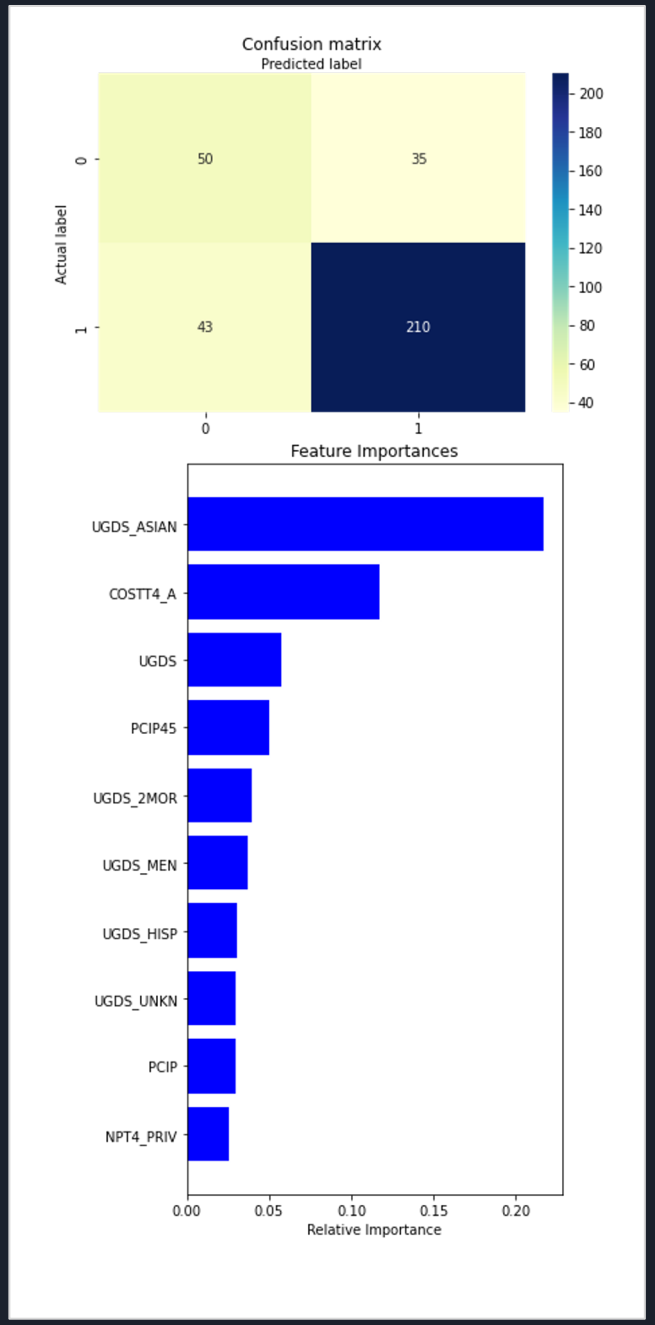
Classification Report for

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.50 | 0.64 | 113 |
| 1 | 0.85 | 0.97 | 0.91 | 338 |
| accuracy | | | 0.86 | 451 |
| macro avg | 0.86 | 0.74 | 0.77 | 451 |
| weighted avg | 0.86 | 0.86 | 0.84 | 451 |

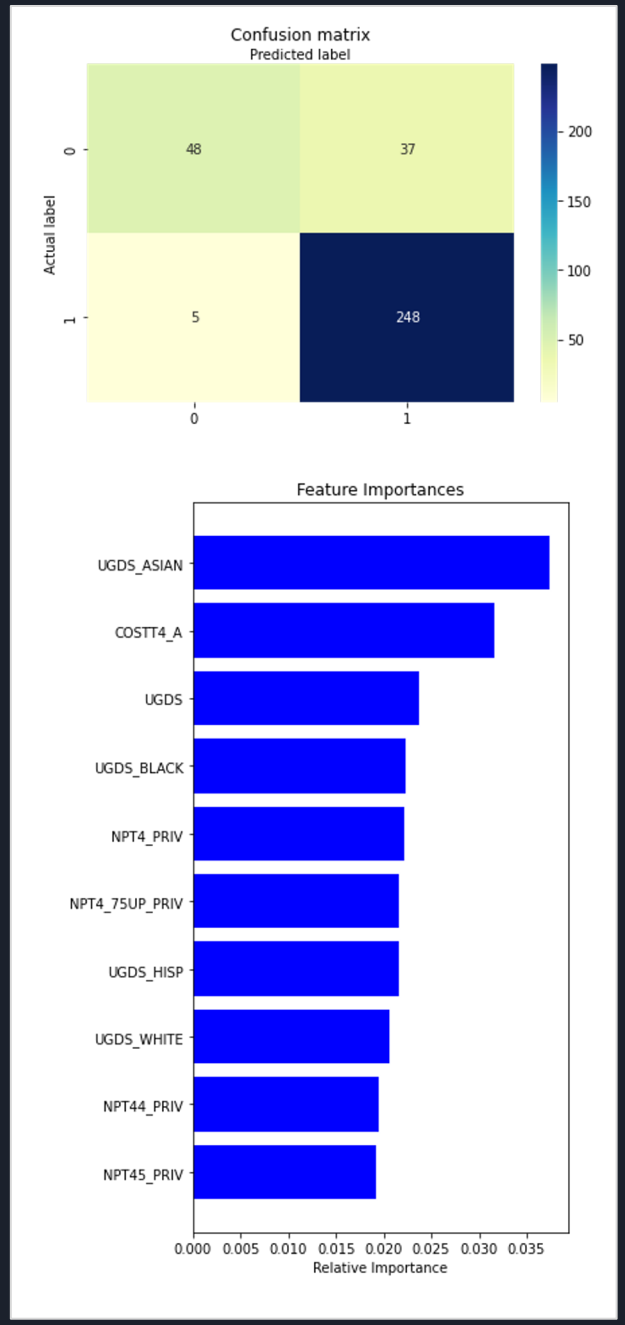
Classification Report for RF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.50 | 0.64 | 113 |
| 1 | 0.85 | 0.97 | 0.91 | 338 |
| accuracy | | | 0.86 | 451 |
| macro avg | 0.86 | 0.74 | 0.77 | 451 |
| weighted avg | 0.86 | 0.86 | 0.84 | 451 |

Confusion Matrix and Top 10 Features from DT




Confusion Matrix and Top 10 Features from



# References

U.S Department of Education (2020). *College Scorecard Institution-Level Data* [Data Files]. Available from College Scorecard https://collegescorecard.ed.gov/data/