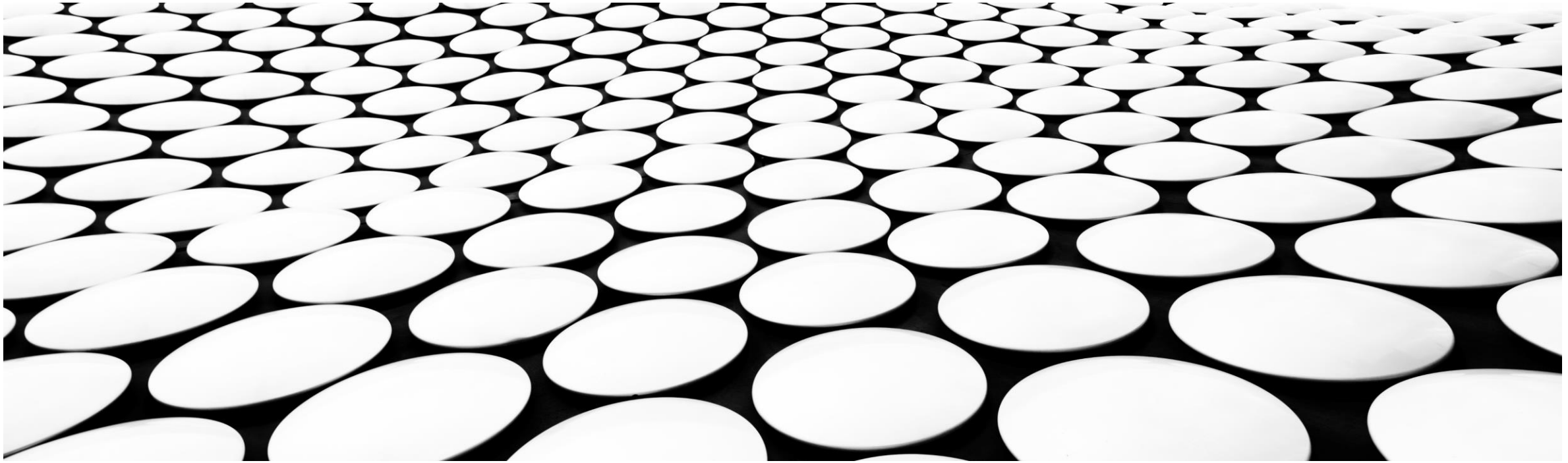
COURSERA CAPSTONE PROJECT

EMPLOYABILITY OUTCOMES OF ENGINEERING GRADUATES IN INDIA

KANDASAMY RAMANUJAM



PROBLEM STATEMENT

Some Facts

- 1.8 Million students enrolled in engineering courses in India
- 10,000+ engineering institutions across the country
- Only 42% of students are placed in jobs

Objective of this project

- understand the factors that influence the employability of engineers
- indicated by the initial salary

- Source:

All India Council for Technical Education: https://facilities.aicte-india.org/dashboard/pages/dashboardaicte.php

DESCRIPTION OF DATA

ASPIRING MINDS EMPLOYABILITY OUTCOMES 2015 (AMEO 2015)

- Scores from school final exams – 10th and 12th standard
- Scores from Engineering course
- Engineering branch
- Tier of college and the city in which the college is located
- Demographic data such as gender, state
- Scores on English, logical ability, quantitative aptitude, and Computer Programming from standardized assessment test conducted by Aspiring Minds
- Salary offered to candidate indicates Employability Outcome
- Data can be downloaded here

http://research.aspiringminds.com/resources/#ameo

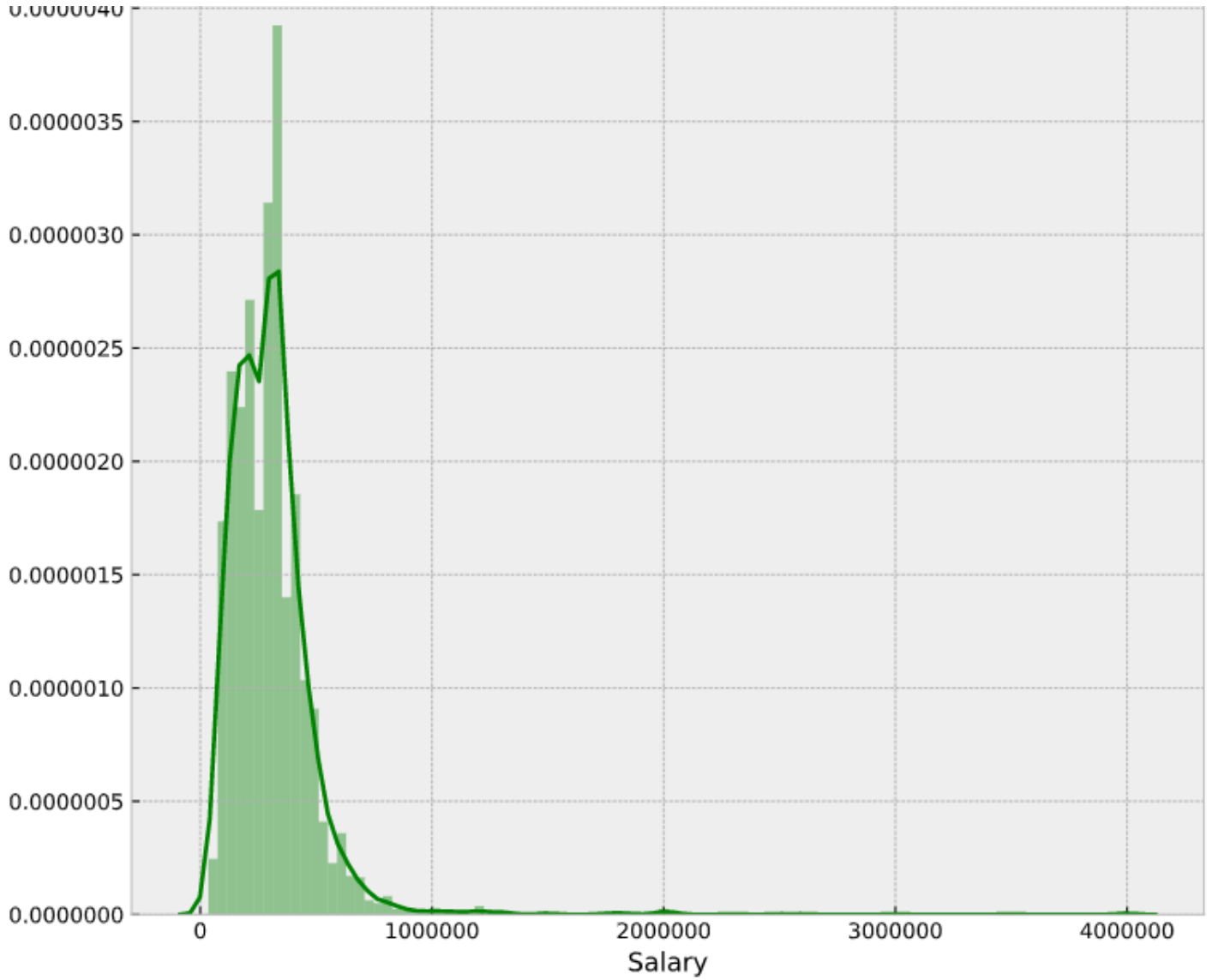
METHODOLOGY FOR ANALYSIS

- Exploratory Data Analysis to understand the data
- Regression Analysis using Salary as the outcome
- Classification by bucketing Salary into a set of ranges

EXPLORATORY DATA ANALYSIS

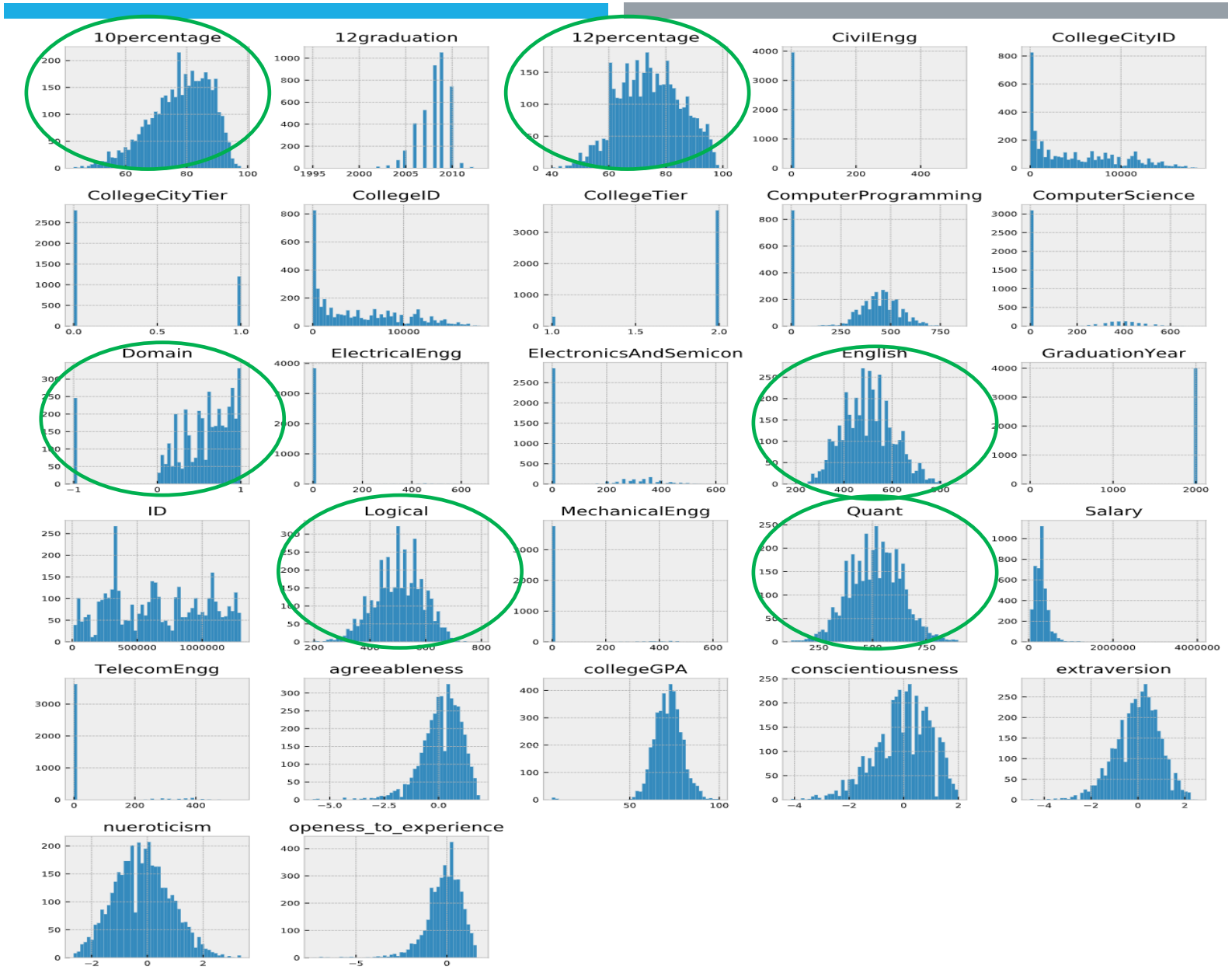
HIGHLIGHTS

- 3998 rows and 39 columns
  All non-null values
- Salary ranges from 0 to 4 million (Indian Rupees)



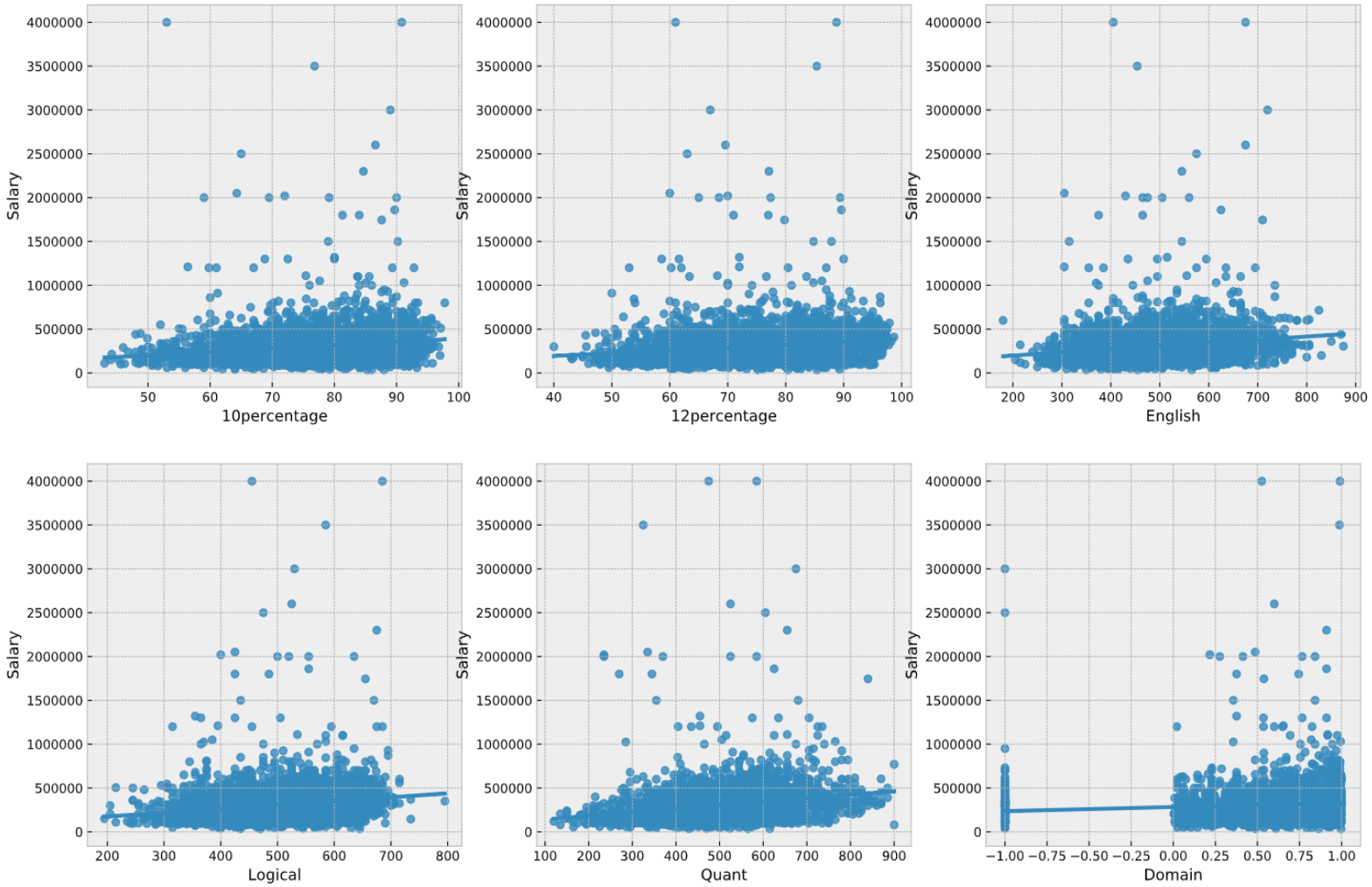
DATA DISTRIBUTION

- Features with good distribution of data chosen
- Features without a good definition of meaning of data are ignored



REGRESSION

- Regression charts do not indicate strong correlation though there is a general trend



RESULTS FROM MODELING

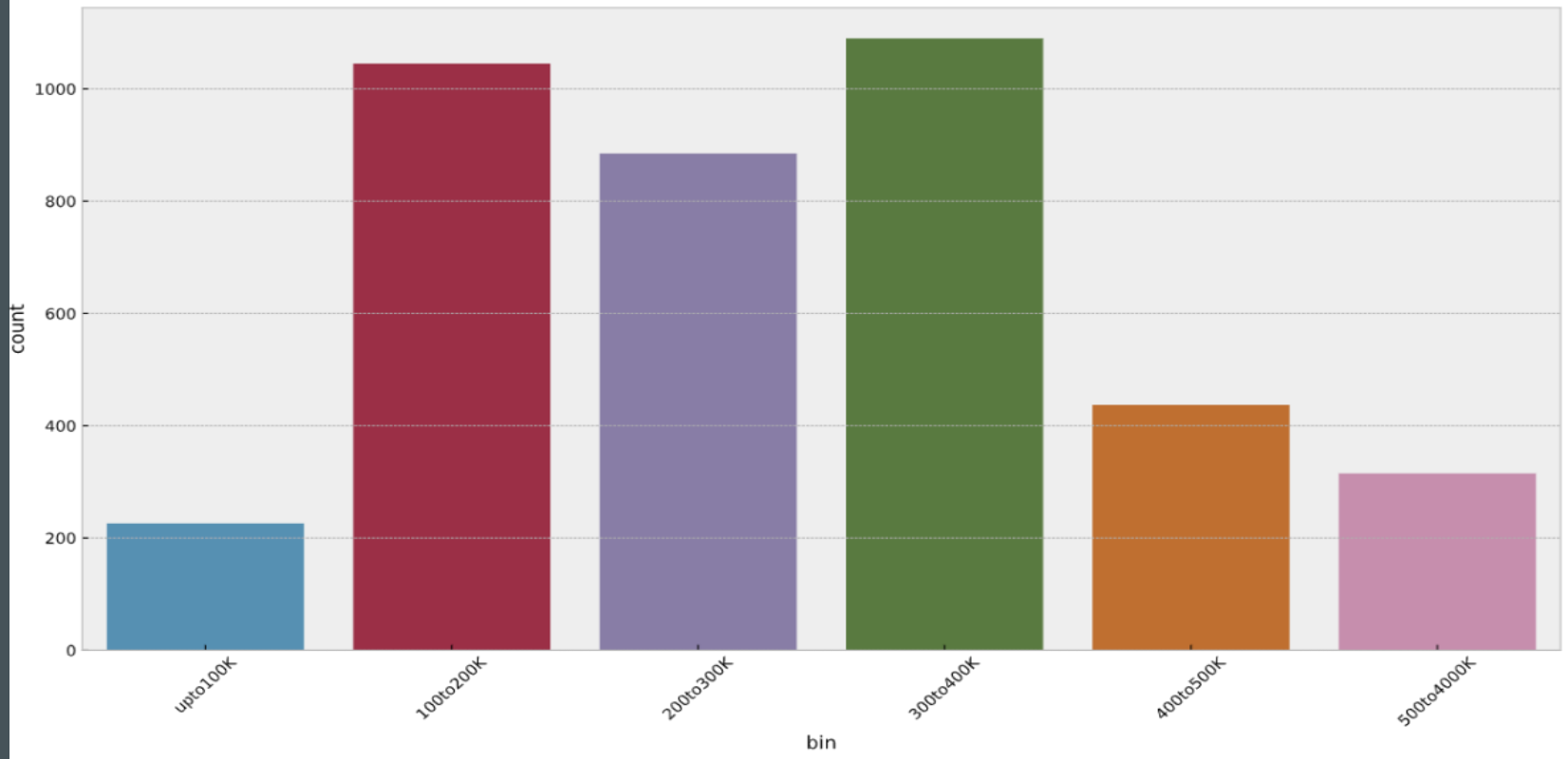
RESULTS FROM LINEAR REGRESSION

- R^2: 0.11
- 10th Percentage and 12th Percentage have a significant influence

```
'10percentage',     852.39702613,
'12percentage',     1416.99541726,
'English’,          176.97876278,
'Logical’,          86.31073346,
'Quant’,            212.99132784,
'Domain'            19859.35136246
```

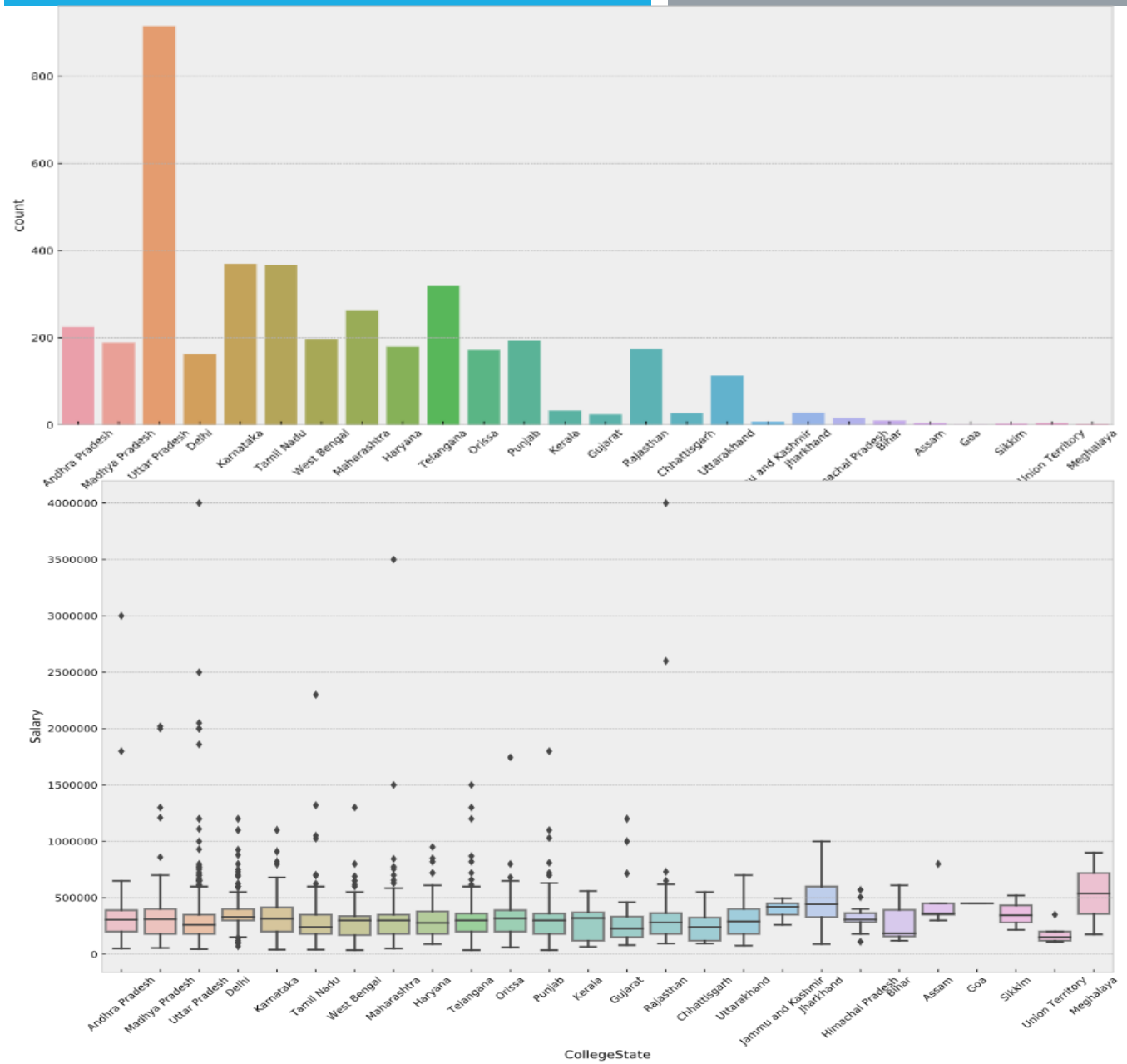
CLASSIFICATION

- Data divided into six bins based on Salary



CATEGORICAL FEATURES

- Additional features identified
- College State is an example



RESULTS FROM SVM

- R^2: 0.417

	precision	recall	f1-score	support
100to200K	0.42	0.67	0.51	230
200to300K	0.19	0.10	0.13	178
300to400K	0.43	0.64	0.51	228
400to500K	0.35	0.12	0.18	89
500to4000K	0.67	0.07	0.12	59
upto100K	0.00	0.00	0.00	57
accuracy			0.39	841
macro avg	0.34	0.27	0.24	841
weighted avg	0.35	0.39	0.33	841

```
[[153  28  48   1   0   0]
 [ 97  18  62   1   0   0]
 [ 49  24 146   8   1   0]
 [ 19   9  49  11   1   0]
 [ 11   6  28  10   4   0]
 [ 39   9   9   0   0   0]]
```

RESULTS FROM GRADIENT BOOST

- R^2: 0.436
- Gradient Boost identified by comparing KNN, SVM, Decision Tree, Random Forest, Ada Boost
- Marginally better score

	precision	recall	f1-score	support
100to200K	0.46	0.72	0.56	210
200to300K	0.26	0.15	0.19	160
300to400K	0.46	0.66	0.55	237
400to500K	0.42	0.17	0.24	87
500to4000K	0.60	0.12	0.19	78
upto100K	0.00	0.00	0.00	44
accuracy			0.44	816
macro avg	0.37	0.30	0.29	816
weighted avg	0.41	0.44	0.38	816

```
[[151  20  33   2   3   1]
 [ 70  24  65   1   0   0]
 [ 39  24 157  11   3   3]
 [ 17   9  45  15   0   1]
 [ 19  11  32   7   9   0]
 [ 34   4   6   0   0   0]]
```

CONCLUSION AND NEXT STEPS

- Low correlation in the set of features available in the data
- 10^{th} Percentage and 12^{th} Percentage seem to be the best predictors of employability
- Accuracy of models could be marginally improved using techniques such as hyper parameter tuning and cross validation
- Additional features with better correlation on outcome may improve the quality of models and predictions