Predicting Job success after graduating from College based on Student's Credentials

Using Supervised and Unsupervised Learning Algorithms

Naimish Sharma

Department of Electrical Engineering and Computer Science

Understanding the problem

The aim of this project is to recognize important factors that influences job success for engineers in India . The trends and patterns in the data are analysed and visualized to aid the design and development of machine learning models that would predict the job salary of a graduate . Aspiring Minds' Employability Outcomes (AMEO) 2015 dataset is analysed and pre-processed for the machine learning pipeline to build machine learning models which will be able to predict the salary of an engineering graduate given his credentials . Standard data analytics and visualizations are used to understand and prepare the data. The data is cleaned of unnecessary attributes . The categorical features are encoded using one-hot encoding scheme. For categorical classification , the target points are placed into class bins based on quantile distribution . These methods help feature engineering to enable a good enough prediction ability.

Data Understanding

Dataset Information –

Training Dataset – "train.xlsx" containing 3998 rows and 39 columns

Testing Dataset – "test.xlsx" containing 1500 rows and 39 columns

Dataset is available on –

http://research.aspiringminds.com/wp-content/uploads/2016/08/datachallenge-cods2016.zip

Attributes Information –

Attributes given in datasets are -

'ID','Salary', 'DOJ', 'DOL', 'Designation', 'JobCity' , 'Gender', 'DOB', '10percentage','10board','12graduation','12percentage','12board','CollegeTier', 'Degree','Specialization','CollegeGPA','CollegeID','CollegeCityTier','CollegeState','GraduationYear','English','Logical','Quant','Domain','ComputerProgramming','ElectronicsAndSemicon','ComputerScience','MechanicalEngg','ElectricalEngg','TelecomEngg','CivilEngg','conscientiousness','agreeablness','extraversion','nueroticism', 'openess_to_experience'

Detailed Information about attributes is given in "**data_description_doc.xlsx"** file which is given in dataset folder downloaded from above link .

Data Preprocessing

This project is implemented with Python Programming Language using various scikit learn machine learning libraries .

1) **Unnecessary features are dropped from data to increase training speed , model interpretability and performance of model on testing dataset .**

Dropped Features are -

['CollegeID','DOB','10board','12board','12graduation','DOJ','DOL','CollegeCityID','CollegeCityTier', 'CollegeState','Designation','Unnamed: 0','JobCity']

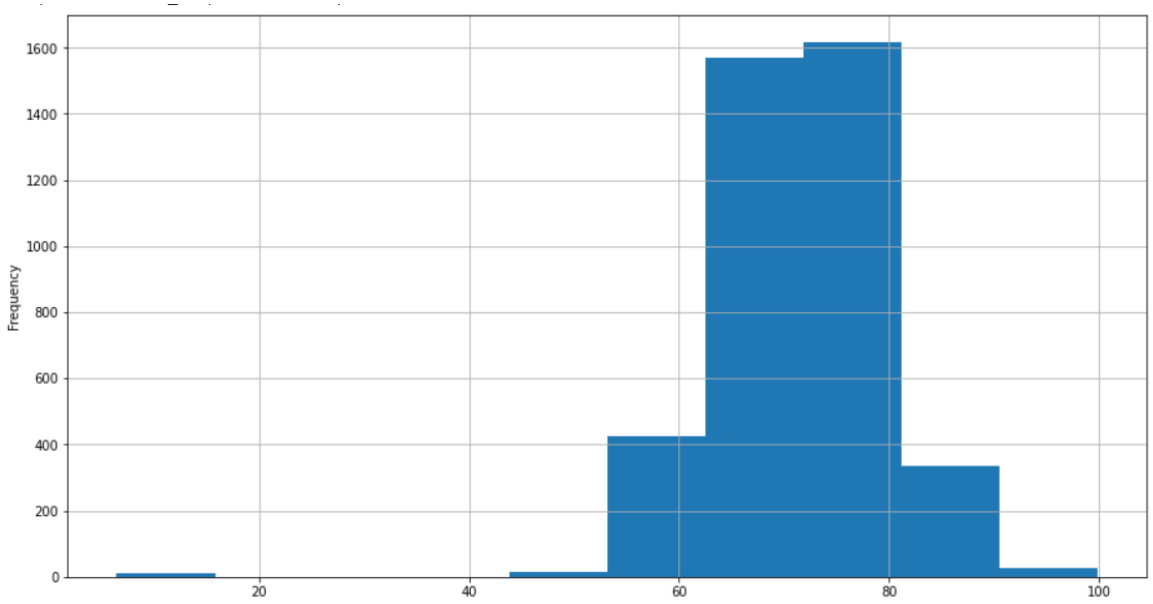
2) **Replacing Negative values(-1 for not attempted) in scores of AMCAT(Aspiring Minds Computer Adaptive Test) by 0**

There were many negative values so it was necessary to replace them by 0

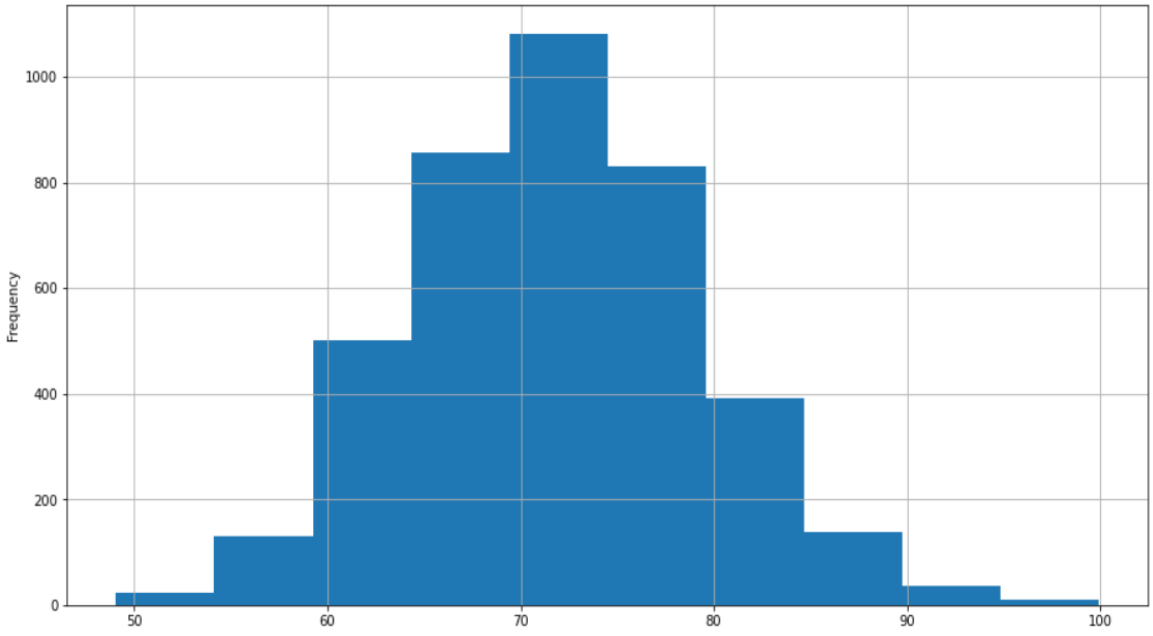
```
Domain                   246
ComputerProgramming      868
ElectronicsAndSemicon   2854
ComputerScience         3096
MechanicalEngg          3763
ElectricalEngg          3837
TelecomEngg             3624
CivilEngg               3956
```

3) **Data for attribute 'collegeGPA' is present on scale of 10 and 100 both . So all data for attribute 'collegeGPA' is converted on scale of 100**

Before Converting –

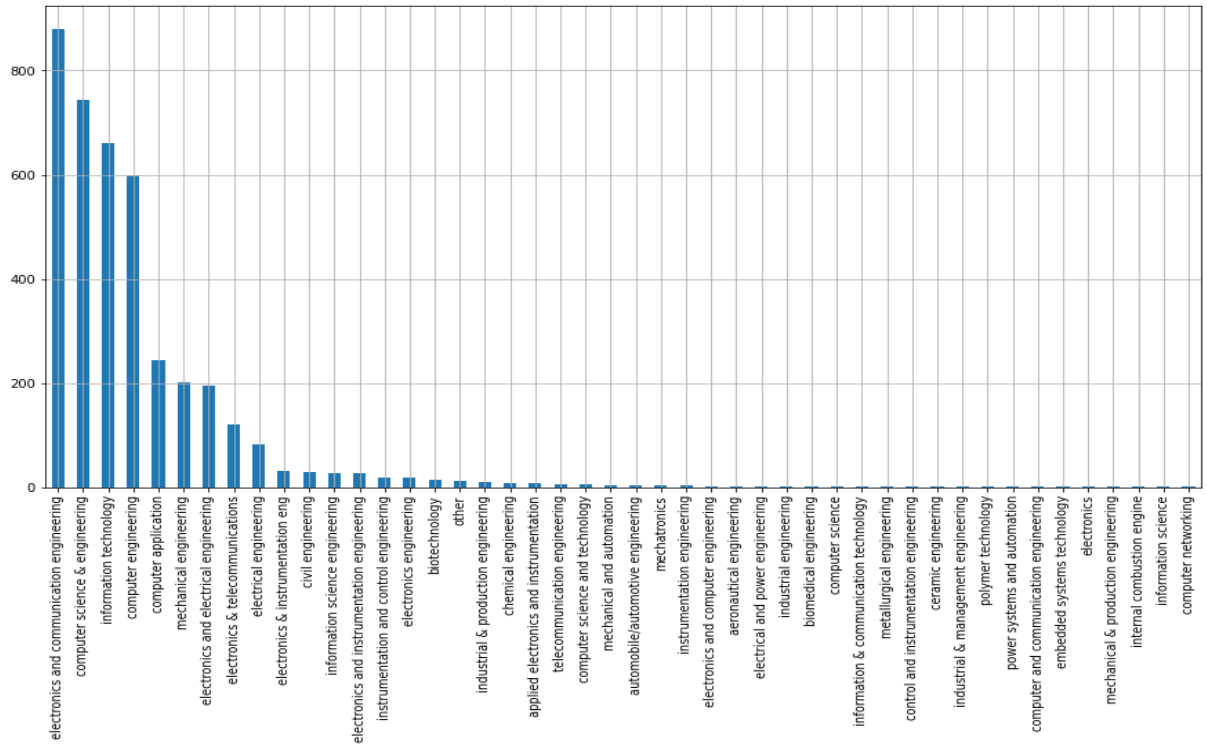


After Converting -

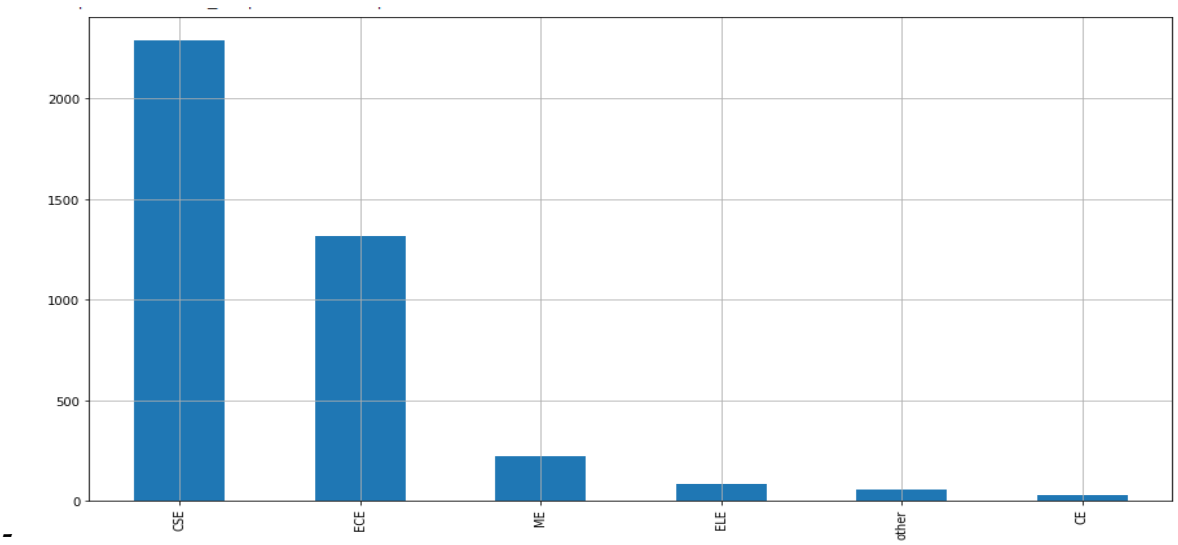


4) **Different specialization are mapped to core disciplines such as CSE(Computer Science&Enginnering) , ECE(Electronics & Communication Engineering) , ME(Mechanical Engineering) , ELE(Electrical Engineering) , CE(Civil Engineering) and others.**

Before mapping -



After mapping numerous Specializations into Core desciplines -



-

5) **One-hot Encode categorical features**

As many machine learning algorithms cannot work with categorical data directly so one-hot encoding is done .

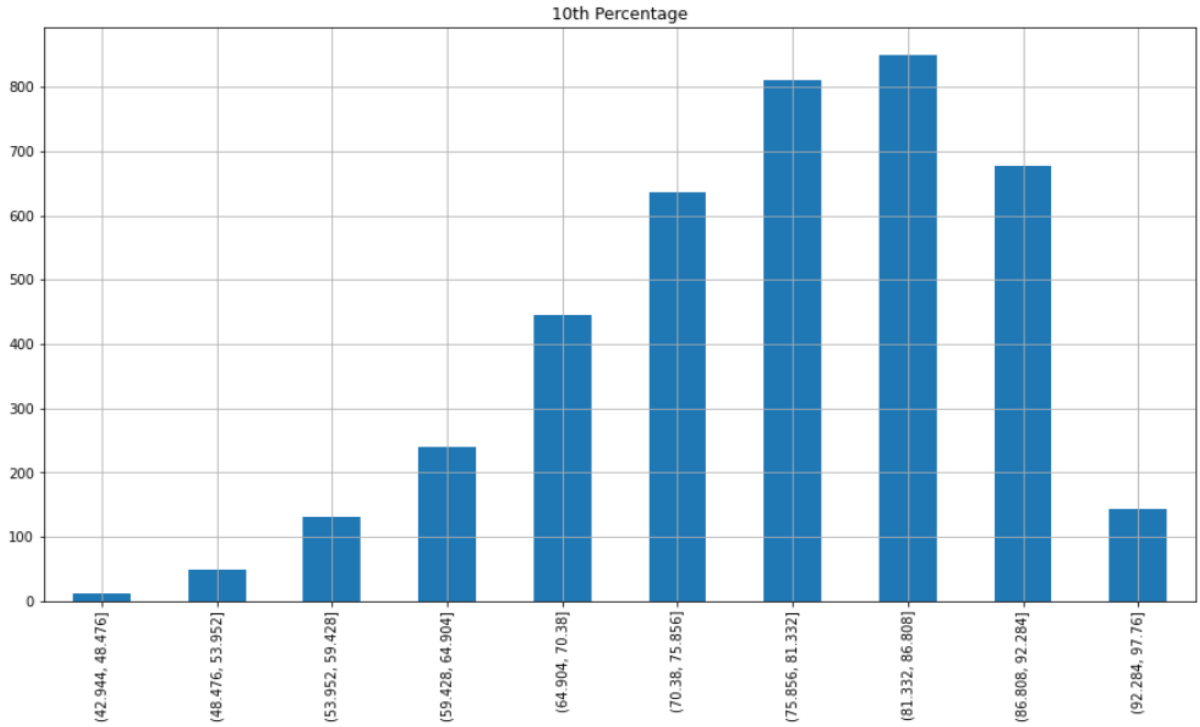
There were many features having categorical data such as -

```
Gender: 2
CollegeTier: 2
Degree: 4
Specialization: 6
GraduationYear: 11
```

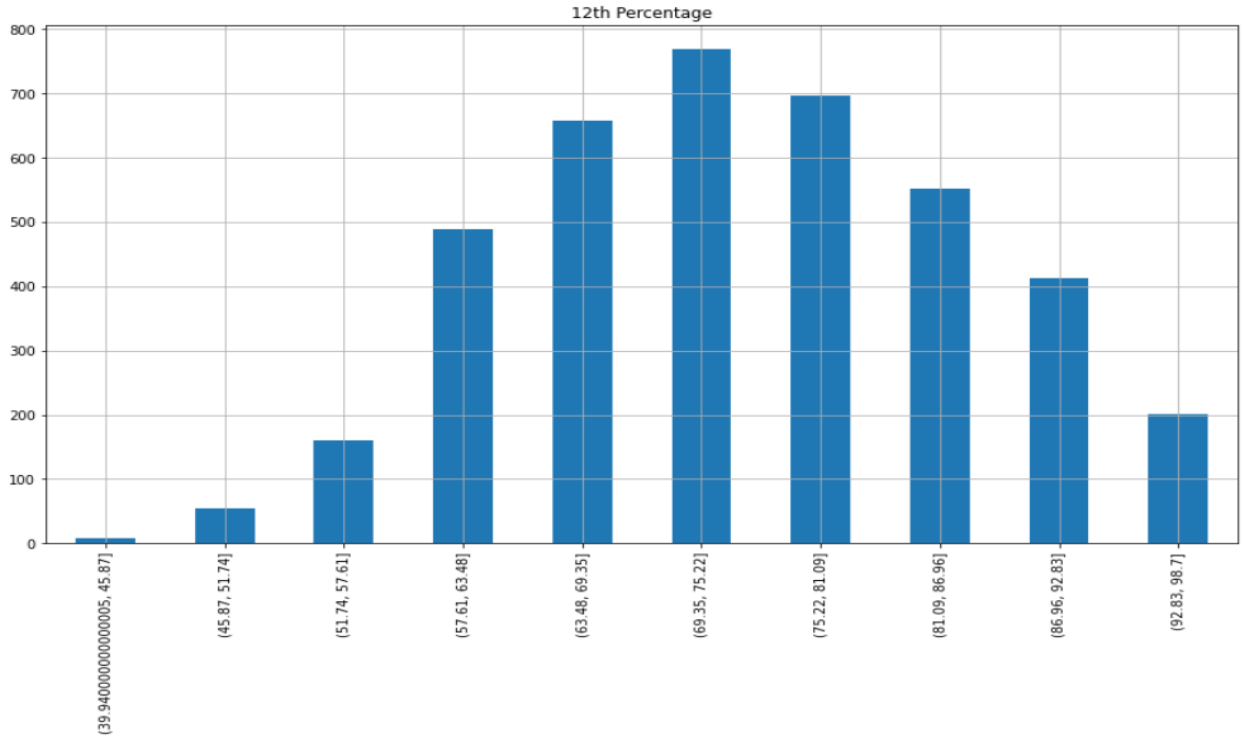
Data Analysis (Pre Analysis to get more information about data before data modelling) –

- **Distribution of 10th percentage –**

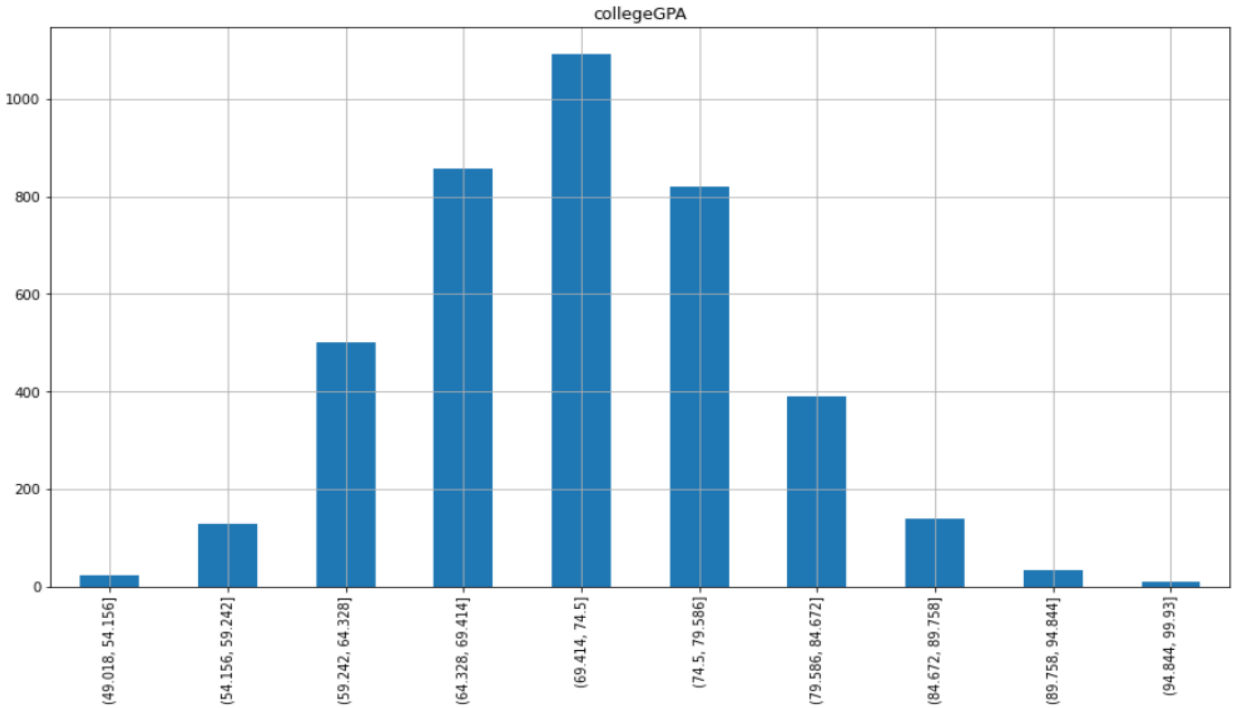
  More than 75% of graduates have 10th percentage in the range of 70.38% to 92.284 %



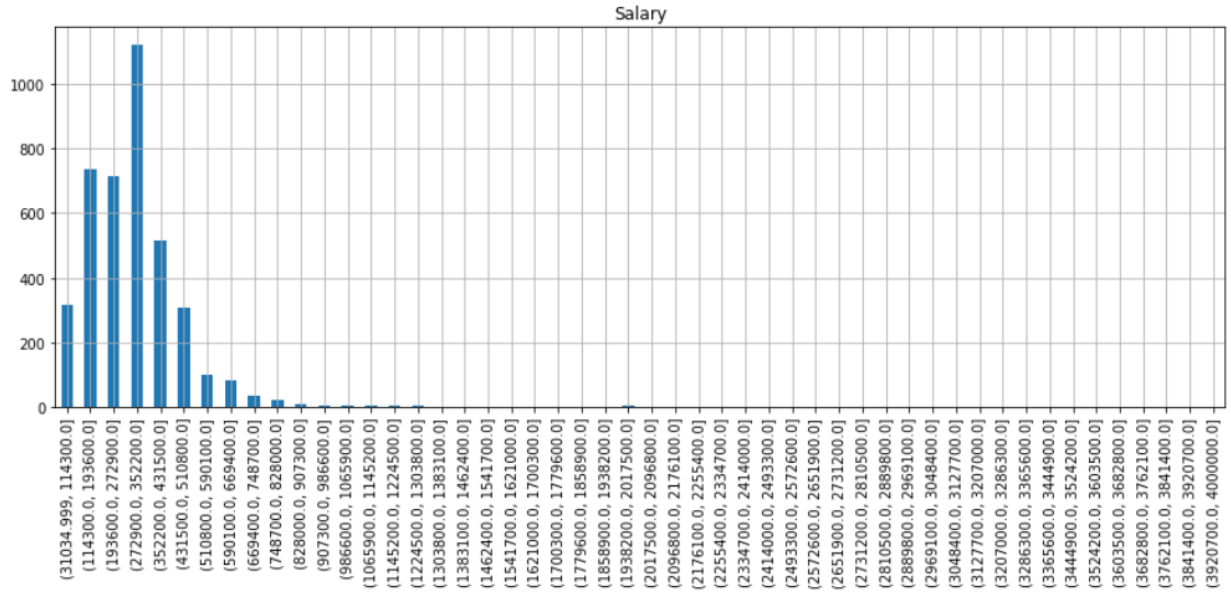
- **Distribution of 12th percentage –**
  More than 75% of graduates have 12th percentage in the range of 63.48% to 92.83 %



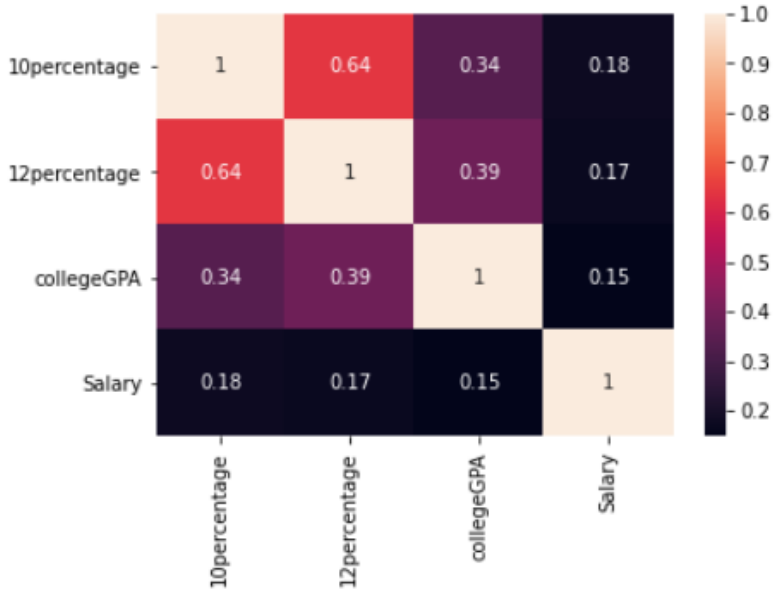
- **Distribution of collegeGPA** –
  More than 70% of graduates have collegeGPA in the range of 64.328% to 79.586 %



- **Distribution of Salary –**
  More than 75% of graduates have Salary in the range of 114300 to 431500 Indian Rupees per annum

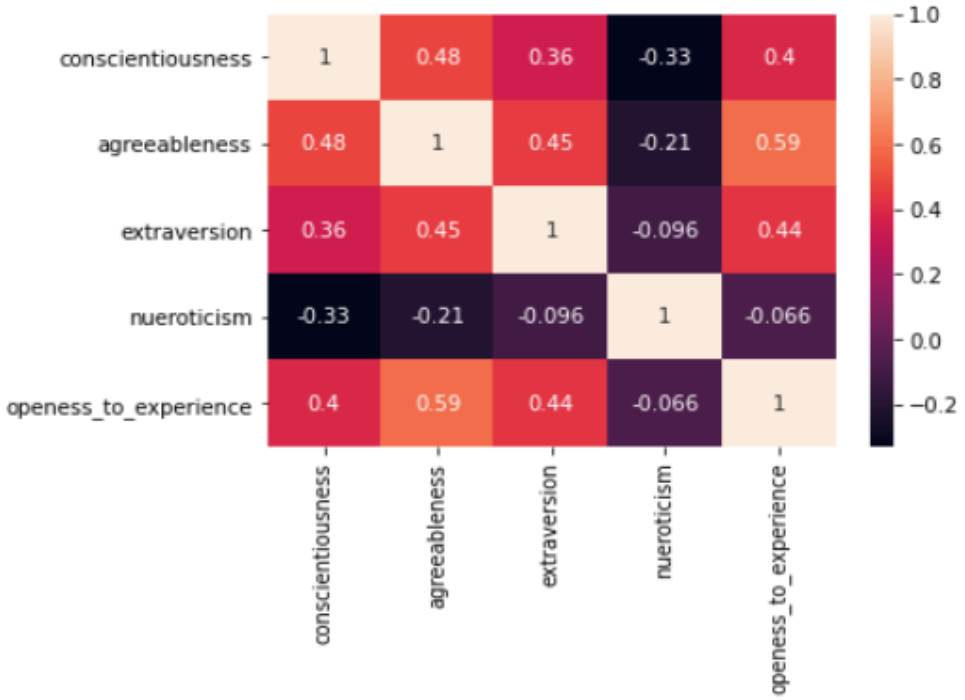


- **Correlation of 10th percentage , 12th percentage and collegeGPA**



10th percentage and 12th percentage have moderately positive relationship.

- **Correlation of different sections of AMCAT's personality test**



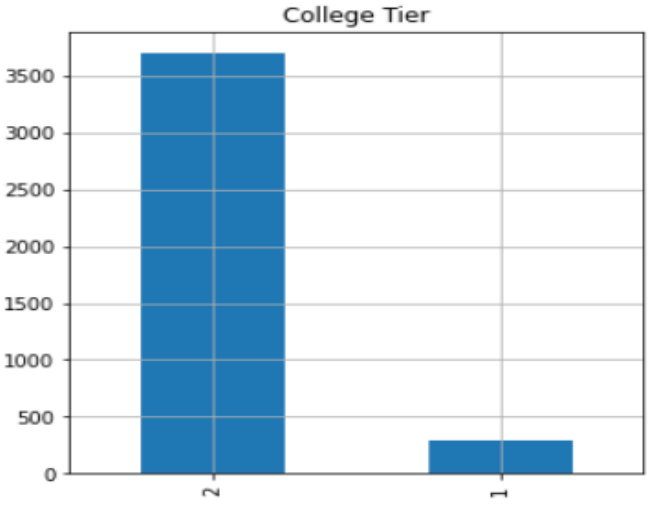
Openess_to_experience and agreeableness have moderately positive relationship . And neuroticism and conscientiousness are moderately negatively correated .

- **Correlation of Standard Test Scores**

No clear correlation is visible in AMCAT Standard tests .

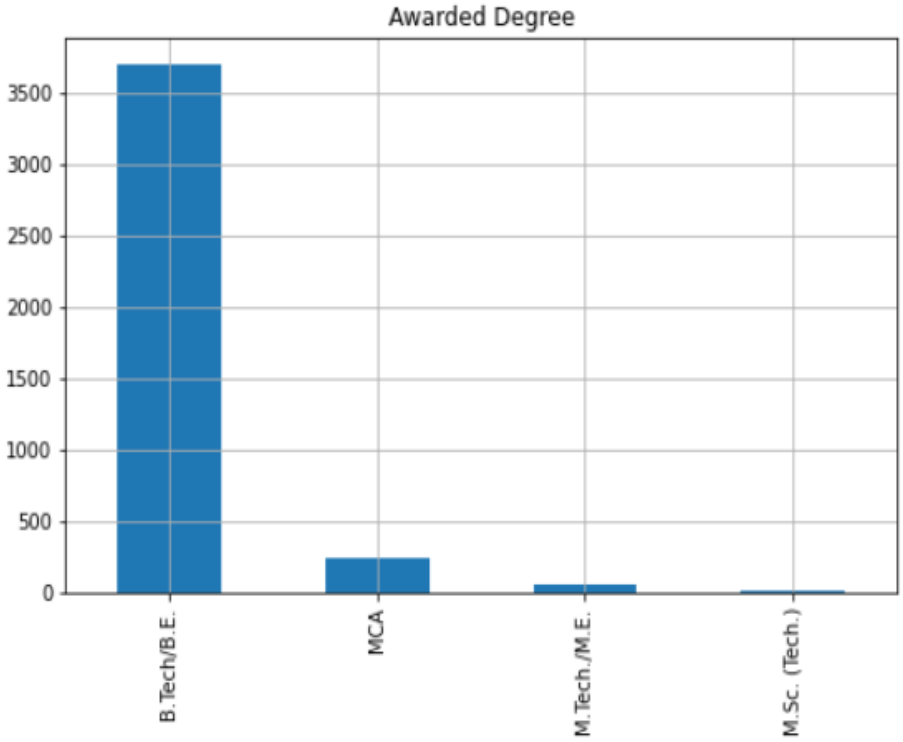
- **College Tier Distribution**

Large chunk of graduate (more than 90%) are from tier 2 College



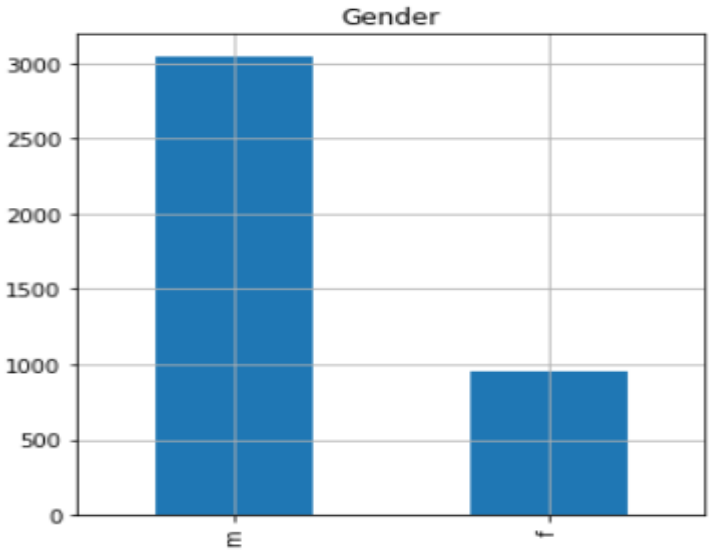
- **Distribution of Awarded Degree**

More than 90% graduates were awarded B.Tech/B.E. Degree



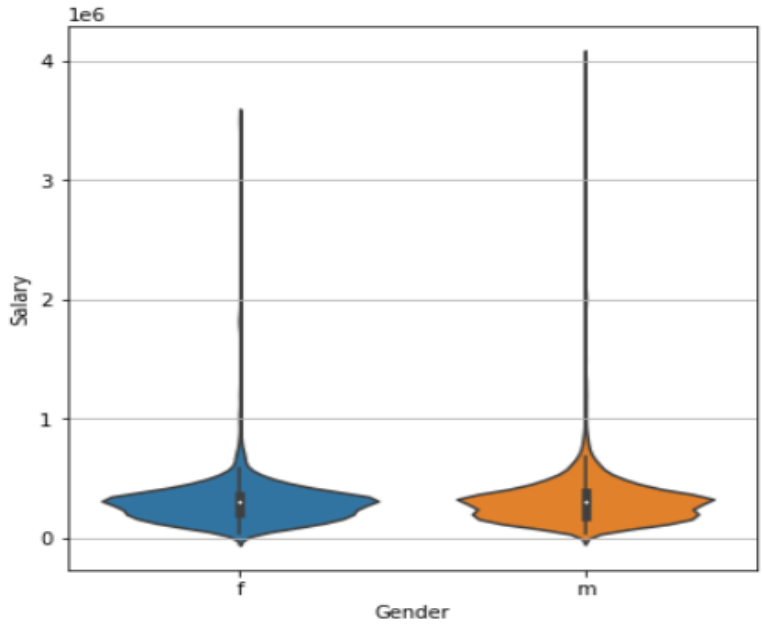
- **Distribution of Gender**

More than 75% of graduates were male



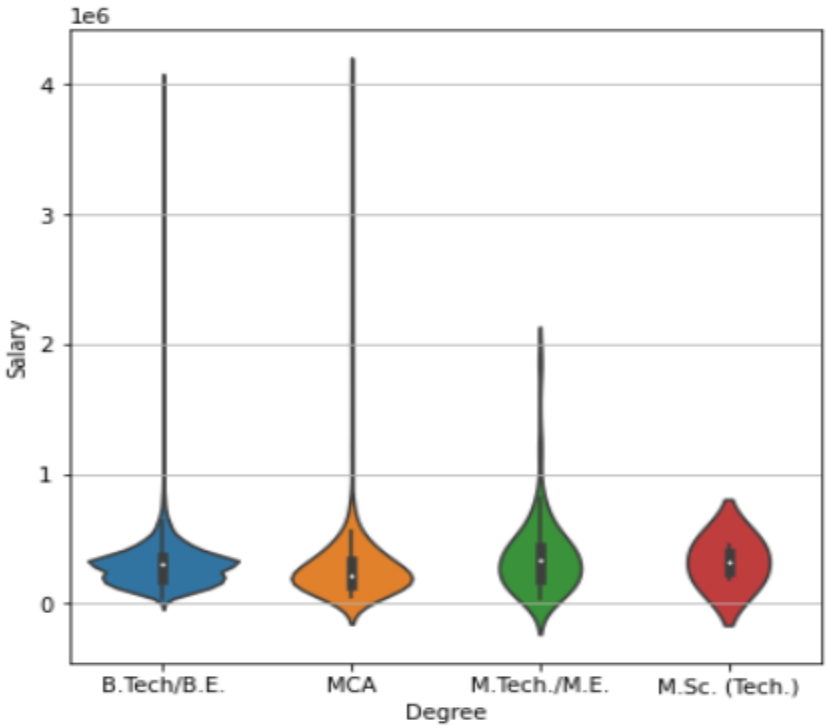
- **Does Gender Affects Salaries of Graduates**

It appears that there is no such gender biasness as observed from the plot though there are more outliers for male's salaries as compared to female's salaries



- **Does degree awarded affects Salaries of Graduates**

Distributions for all four awarded degrees are almost same though there are some outliers in B.Tech/B.E. and MCA as compared to M.Tech/M.E. and M.Sc.(Tech.)



Defining 'Salary' as target variable and rest all columns as features .

Data Modelling -

1) Linear Regression –

Regression estimates are used to explain the relationship between one dependent variable and one or more independent variables .

Linear Regression model is not able to predict the salaries of the graduates accurately as mean squared error and R2 score were quite poor for both training and testing . Linear model made error very high .

```
Mean Squared Error
Train: 38177058053.395   Test: 35218032641.643
R2 score
Train: 0.168     Test: 0.195
```

Tried to use regularized linear regression models –

Ridge Regression (i.e. Linear Regression with L2 regularization) -

With alpha = 0.01 , maximum iteration = 100000

The performance of model remains almost same on using ridge regression which is visible from the values of mean squared error and R2 score

```
Mean Squared Error
Train: 38177058340.606   Test: 35218119191.137
R2 score
Train: 0.168     Test: 0.195
```

Lasso Regression (i.e. Linear Regression with L1 regularization) -

With alpha = 0.01 , maximum iteration = 100000

The performance of model doesn't changes on using lasso regression also .

```
Mean Squared Error
Train: 38177058053.922   Test: 35209946628.721
R2 score
Train: 0.168     Test: 0.195
```

Analysis –

From the R2 scores measured by the different regression model on the dataset , linear regression models with or without regularization were not able to perform efficiently . Subset selection was not used on the processed dataset as we saw that most of the features in data are highly uncorrelated so subset selection would have biased our model towards selected features .

2) Logistic Regression –

It measures the relationship between the categorical dependent variable and one or more independent variables by estimating the probability of occurrence of an event using its logistics function

Column 'Salary' is changed to 5 classes based on percentiles -

Classes are divided on basis of 0 , 20 , 40 , 60 , 100 percentile values of Column 'Salary'

1. Class 1 contains (34999.999, 180000.0]
2. Class 2 contains (180000.0, 240000.0]
3. Class 3 contains (240000.0, 325000.0]
4. Class 4 contains (325000.0, 400000.0]
5. Class 5 contains (400000.0, 4000000.0]

Used all types of Solvers but accuracy remains around 0.43 in train and 0.42 in test dataset split .

Solver	Train	Test
lbfgs	43.46	42.00
liblinear	43.17	42.25
sag	43.46	42.00
newton-cg	43.46	42.00
saga	43.39	42.00

Analysis –

On trying different combinations of attributes for our logistic regression model accuracy was less than 35% . So , other than 'Salary_Class' and 'Salary' all other features were used to fit the model which give us accuracy around 43% . Therefore it signifies that simple logistic regression will be able to classify salaries of graduate with less accuracy because of presence of high dense data points . To increase the accuracy dimensionality reduction or more complex learning models has to be used .

3) Support Vector Machine –

Support vector machines (SVMs) are powerful yet flexible supervised machine learning methods used for classification, regression, and, outliers' detection. SVMs are very efficient in high dimensional spaces and generally are used in classification problems.The main goal of SVMs is to divide the datasets into number of classes in order to find a maximum marginal hyperplane (MMH)

SVC with linear and rbf kernels are used for multiclass classification -

Solver	Train	Test
rbf	53.57	40.5
linear	42.60	42.5

In certain situations some points lie across the margin and get misclassified . To get rid of this SVM have **slack variables** which allows some instances to fall of the margin but penalize them .

Analysis –
SVM model was trained with different kernels , but best results were obtained using linear kernel . Using this we got train accuracy of 0.426 and test accuracy as 0.425 respectively . On observing this we can say that there is large overlap of datapoints as we are still getting test accuracy of 0.425 despite using such a complex model(SVM)

4) Random Forest –

It computes the locally optimal feature/split combination. In Random forest, each decision tree in the ensemble is built from a sample drawn with replacement from the training set and then gets the prediction from each of them and finally selects the best solution by means of voting. It can be used for both classification as well as regression tasks. Random forest is used over decision tree to prevent overfitting of model on training dataset .

```
Training Accuracy :  100.0
Testing Accuracy :  40.42
```

Analysis –
The distribution of marks in 10th was skewed towards the right as most students scored better than their peers. The distribution of marks in 12th moved to a bell curve as most students performed average as compared to their peers. The distribution of college grades is slightly skewed towards the left as most students were performing worse than their peers . The distribution of salary is heavily skewed towards the left. One of the observations is the large overlap of data points .

5) Neural Network Analysis(ANNs) –

“categorical_crossentropy” is chosen as loss function , ‘Adagrad’ is chosen as the optimizer for the model . Relu and sigmoid are chosen as the activation function for the network architecture .

Train and Test accuracy for the model was -

	Train	Test
Accuracy	24.8	26.5

Analysis –

We observed that when relu function was chosen as an activation function of the model for input and hidden layers and sigmoid function for output layers , model performed better in classifying the data points to their correct salary levels compared to the predictions done by model using other activation functions. Also observed that accuracy of classification depends on the number of iterations , number of layers in the network architecture. Accuracy was just around 0.25 using all features while building model . So , accuracy may increase if we use subset of features during model building but it may account for biasness of model as the features remaining after data preprocessing are mostly uncorrelated and important as they affect the salaries of graduates .

6) Unsupervised learning –

Decision Tree on Principal Component Analysis(PCA) is used .

```
Train Accuracy: 24.803
Test Accuracy: 26.500
```

Analysis -

We concluded that accuracy was around 0.25 for train and 0.265 for test dataset with high dimensionality as well as high number of components. Due the presence of large number of features and samples, the reduced attributes will be better able to recognize the patterns present in the dataset. Moreover, with more hyperparameter tuning the performance of the model could be further improve .

Comparative Analysis of Models -

As there are lots of important features from the data which would be judging the salary of the graduates **linear regression** was unable to perform good taking into account all the useful features . But on applying linear regression considering individual features we got that **Logical Scores in the AMCAT tests** has the highest ranking showing that the aptitude levels play an important role.

From **Logistic Regression** we got accuracy of approximately 0.43 and 0.42 respectively on training and testing data by using any of the solver taking into account all the features for fitting of model. But on taking combinations of features we got to know that higher collegeGPA, quantitative score and 12th standard scores contribute to a higher salary .

Using SVM model but best results were obtained using linear kernel . Using this we got train accuracy of 0.426 and test accuracy as 0.425 respectively . On observing this we can say that there is large overlap of datapoints as we are still getting test accuracy of 0.425 despite using such a complex model(SVM)

Using Random Forest testing accuracy was 0.40 which is better when compared to other models . The distribution of marks in 10th was skewed towards the right as most students scored better than their peers. The distribution of marks in 12th moved to a bell curve as most students performed average as compared to their peers. The distribution of college grades is slightly skewed towards the left as most students were performing worse than their peers . The distribution of salary is heavily skewed towards the left .

Using Artificial Neural Network we got training and testing accuracy of 0.248 and 0.265 which was quite low compared to Logistic Regression , SVM and Random Forest . Accuracy could be increased if we use subset of features during model building .

Using Unsupervised Learning we got accuracy of 0.25 for train and 0.265 for test dataset with high dimensionality as well as high number of components which is quite low . Hyperparameter tuning could have further improved the performance .

Both the Classification and Regression have given useful insights into the factors affecting the salaries of the graduates in India. Some of the overlapping results from both the problems indicate that the English speaking skills, Quant skills, Specilization predominantly affect the salary.