**Question 0: Salary Prediction**

**Preliminary Visualization**

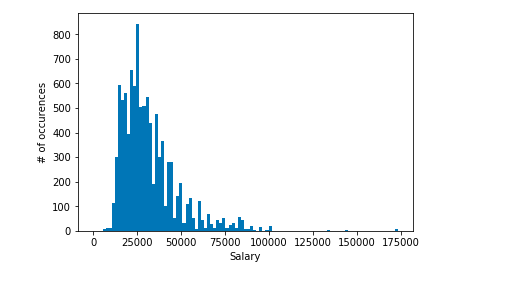
salary = originalsalary

plt.hist(originalsalary['SalaryNormalized'], bins=[n\*1750 **for** n **in** range(100)], histtype='stepfilled')

plt.xlabel("Salary")

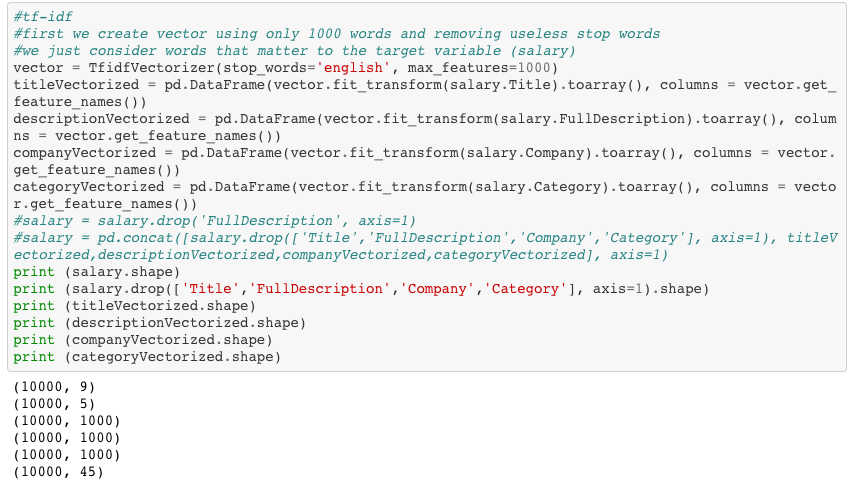
plt.ylabel("# of occurences")

plt.show()



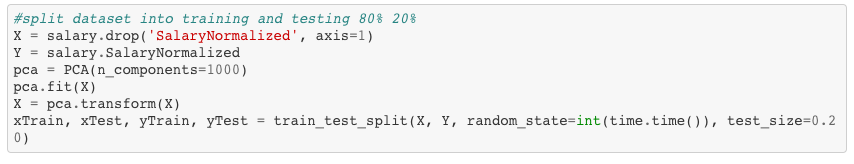
**Data Preprocessing**

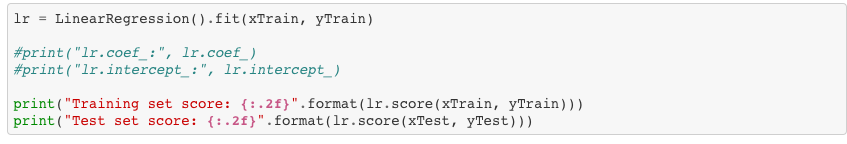
****

****

****

**Training and Testing Models**

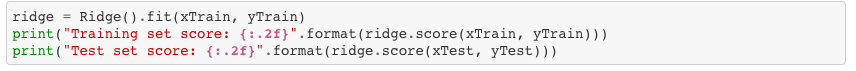
**  
Linear Regression**

****

Training set score: 0.73

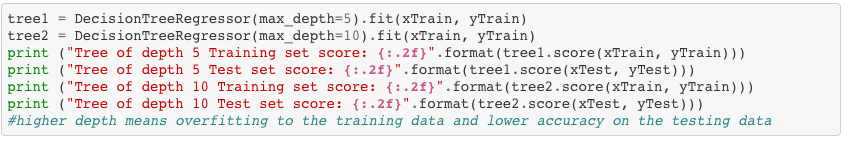
Test set score: 0.60

**Ridge Regression**

****

Training set score: 0.79

Test set score: 0.60 **Decision Trees**

****

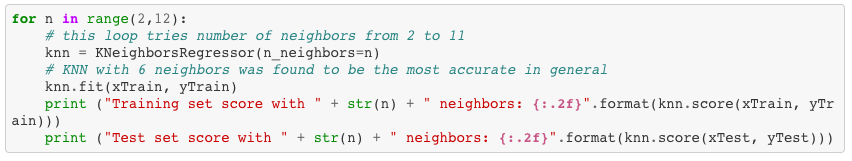
Tree of depth 5 Training set score: 0.50

Tree of depth 5 Test set score:0.30

Tree of depth 10 Training set score:0.77

Tree of depth 10 Test set score:0.55

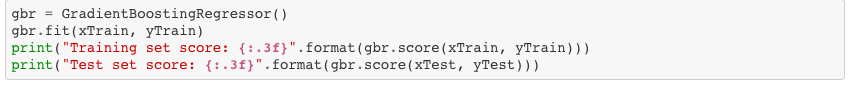
**K-Nearest Neighbors**

****

Training set score with 6 neighbors: 0.39

Test set score with 6 neighbors: 0.18

**Gradient Boosted Regression**[**¶**](https://render.githubusercontent.com/view/ipynb?commit=ca63444e3db91b833abd909b24255e7d089c2700&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f726f68697472616a6373652f535441543531372f636136333434346533646239316238333361626439303962323432353565376430383963323730302f5175657374696f6e305f50726f322e6970796e62&nwo=rohitrajcse%2FSTAT517&path=Question0_Pro2.ipynb&repository_id=149576428&repository_type=Repository#Gradient-Boosted-Regression)

****

Training set score: 0.73

Test set score: 0.50

**Results and Conclusion:**

Datastet was imported and data preprocessing has been done by getting rid of all the useless columns, and replaced the null values with mode which is most frequent occurrence in the dataset, and then implemented tf-idf(term frequency-inverse document frequency) where I first created vector using only 1000 words and removing useless stop words because we just consider words that matter to the target variable (salary) then preprocess the categorical features and then splitted my data in 80% of training and 20% test set and applied different models in which It seems the linear and ridge regression the most accurate and best among all in training and test score than other models.