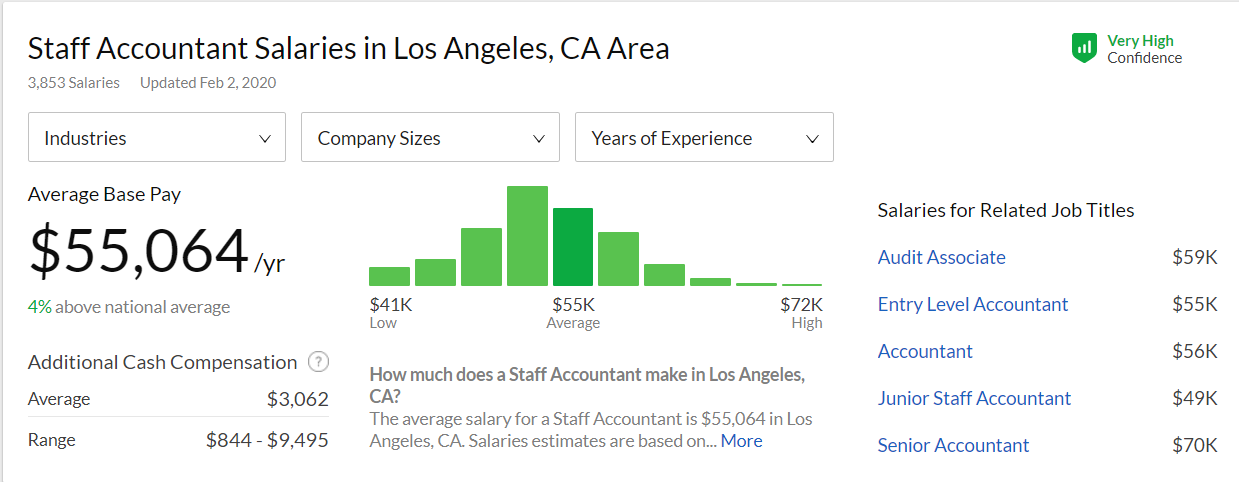
**Introduction**

For my first data science project, I chose to examine employment data to build a salary predictor.

Websites such as [Salary.com](http://salary.com/) [payscale.com](http://payscale.com/), or [Glassdoor.com](http://glassdoor.com/), utilize user imputed data to predict salaries among all industries. The partial screenshot below displays the wages for a staff accountant in the Los Angeles area. Given the industry, company size, and years of experience, the average salary for a Staff Accountant is $55,064.



Both employees and employers benefit from these websites by knowing the market value of a job. As you can imagine, websites that provide salary data are interested in providing accurate results, so that there is a likelihood that future employees and employers can be confident during salary negotiation.

Years of experience, college major, industry, and distance from metropolis are some of the factors that affect the salary predictor's formula accuracy in computing wages. However, most salaries recorded on [glassdoor.com](http://glassdoor.com/) are posted by users who are willing to divulge their actual pay. Since anyone can post their salary, the data from these websites might be skewed. For example, there is a 30-thousand-dollar difference between the average and the highest-paid Staff Accountants. Additionally, the section labeled "Salaries for Related Job Titles," reveals a similar trend to the Staff Accountant salary information and should be juxtaposed upon the Staff accountant bell curve.

Therefore, the intended value of this project is to provide insight into more tangible discriminating and modifiable predictors of salary data.

**Data**

The features of the dataset include job id, company id, job type, degree, major, industry, years of experience, and miles the employee lives from the metropolitan city. The actual size of the dataset is 1,000,000 rows. The data from this project was provided by Data Science Dream Job LLC.

**Preprocessing**

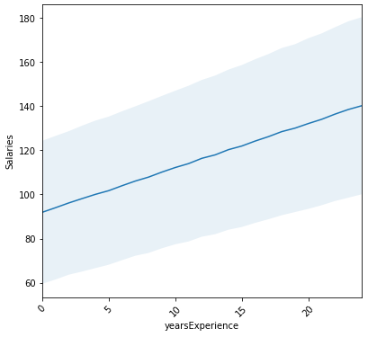
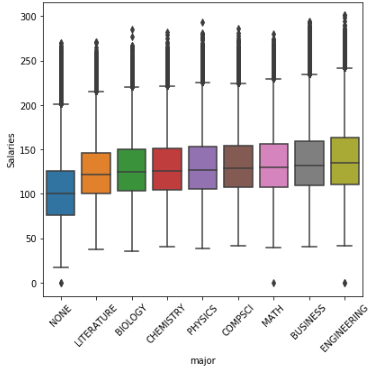
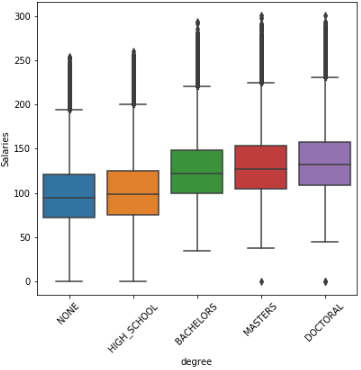
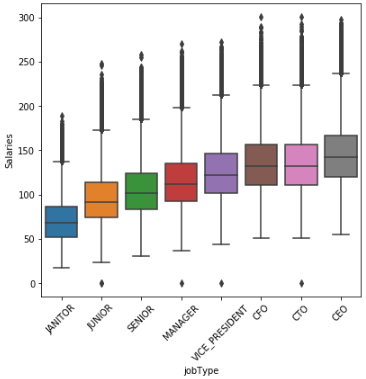
As part of data integrity, I implemented the following steps:

* Merged all features into a single data frame.
* Visualized the target variable.
* Identified potential outliers utilizing IQR.
* Removed salaries with a 0 value.

One notable outlier discovered are the 16 individuals with the title of "Junior" who are receiving a salary over $250. After inspecting the data, I found that these high paid "Junior' employees are primarily in the Oil and Finance fields. Those industries are extremely lucrative; therefore, the outliers discovered will be kept in the dataset.

**Exploratory Data Analysis**

Before developing predictive models, I did a fundamental exploratory data analysis to see the relationship between salary and the other attributes. I hypothesized that years of experience, industry background would have the most significant correlation with salary, and all other features would be insignificant. However, the box plots display a different story. Salary is mostly predicated upon a person's position, degree, major, and the years of experience they have. Thus, the data support the basic notion that higher education and expertise yields a higher salary.

****

**Modeling**

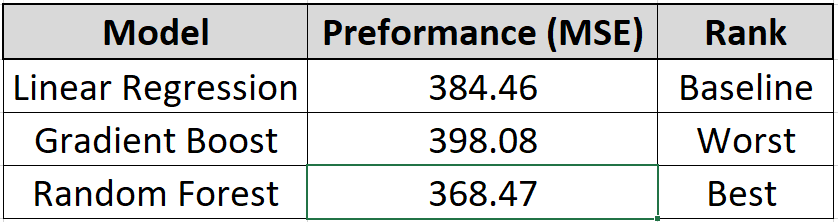
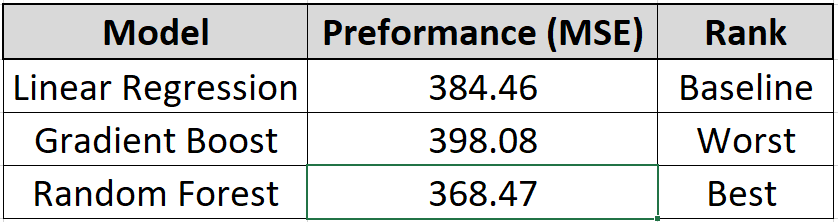
After EDA completed the following feature engineering.

* Defined categorical variables
* Consolidated the training data
* Re-indexed the training data.
* Encoded categorical data.

Next, created and tuned each model, the models chosen were:

* Linear Regression
* Pipeline (PCA)
* Random Forest (RF)
* Gradient Boost (GB)

The results are displayed in the image below, I excluded Pipeline because it yielded the same MSE as the baseline model.



The best preforming model, random forest, out preforms all models.  In attempting to make a more accurate model I added more trees in both gradient boost and random forest algorithms. The result did not develop a better MSE for either and continuing down that path would lead to overfilling. Sixty trees yielded the best result for random forest, and forty trees for gradient boot. Oddly, I assumed that GB would out preform RB due to its ability to learn from previous iterations; however, multiple randomized trees proved to be more accurate.

**Conclusion**

This project's greatest insight is that it supports the notion that getting a higher education, working hard and getting an executive position will yield a higher salary. And greatest factor to attaining a higher salary in any field of work is to many years of experience. In a future iteration of this project I would develop a more object oriented approach that has more feature engineering to reduce the MSE.