WAF Data Challenge

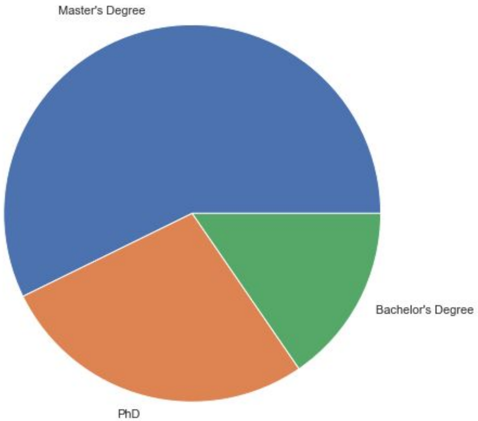
Andrew Zhang

EDA, Data Processing/Cleansing

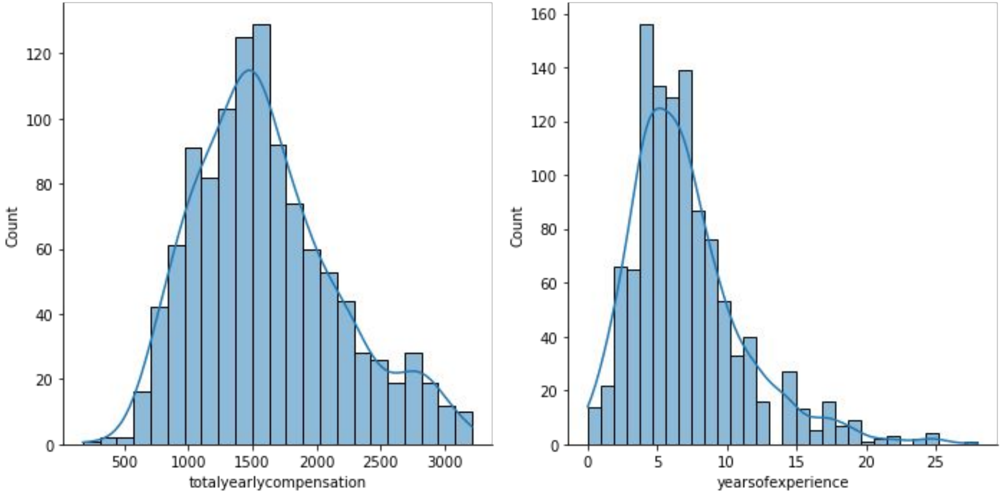
Data Preprocessing:

- Standardized total compensation based on cost of living
- Removed rows with invalid or Null entries
- Removed redundant/irrelevant columns
- Adjusted years of experience for Master's/PhD
- Log Transformed total compensation for hypothesis test

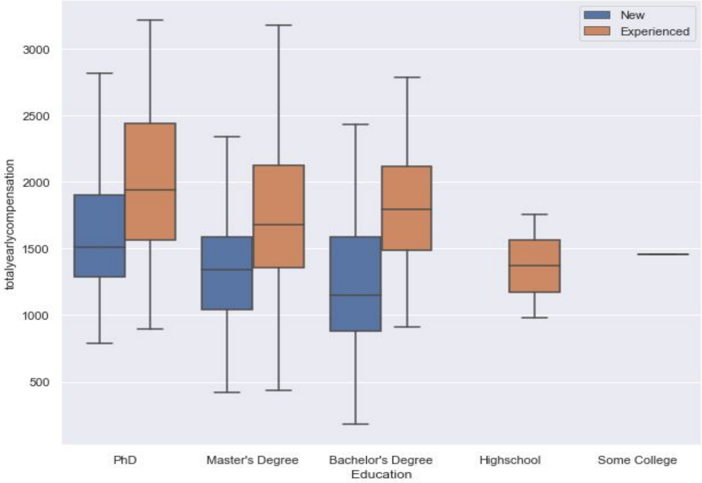
Composition of Education Levels



Distribution of Key Columns



Compensation at Different Levels

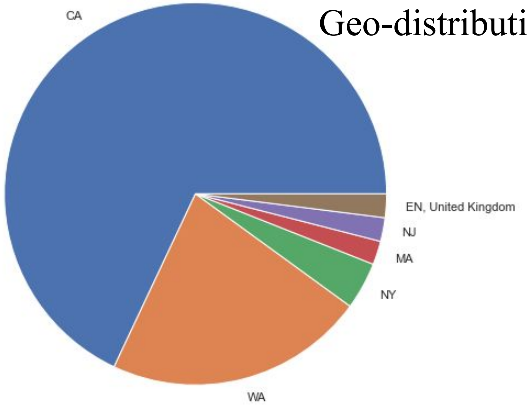


Evaluating Top 50 Earners

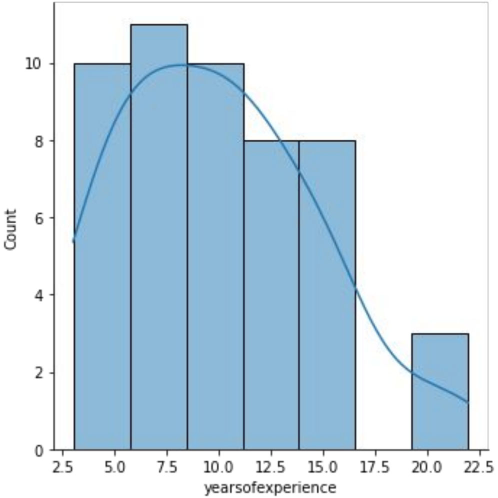
Insights:

- Top earners work for big tech or unicorns and are concentrated on the West Coast or North East.
- Only 14% have worked over 5 years at current company.
- Wide range of YOE from 2 to 20+ years.

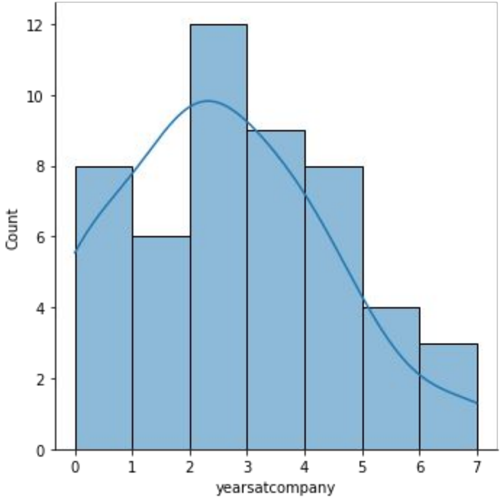
Geo-distribution



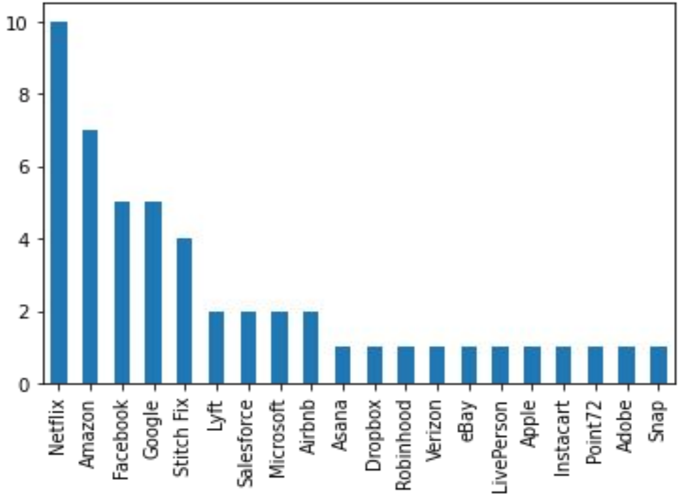
Distribution of YOE



Distribution of YAC

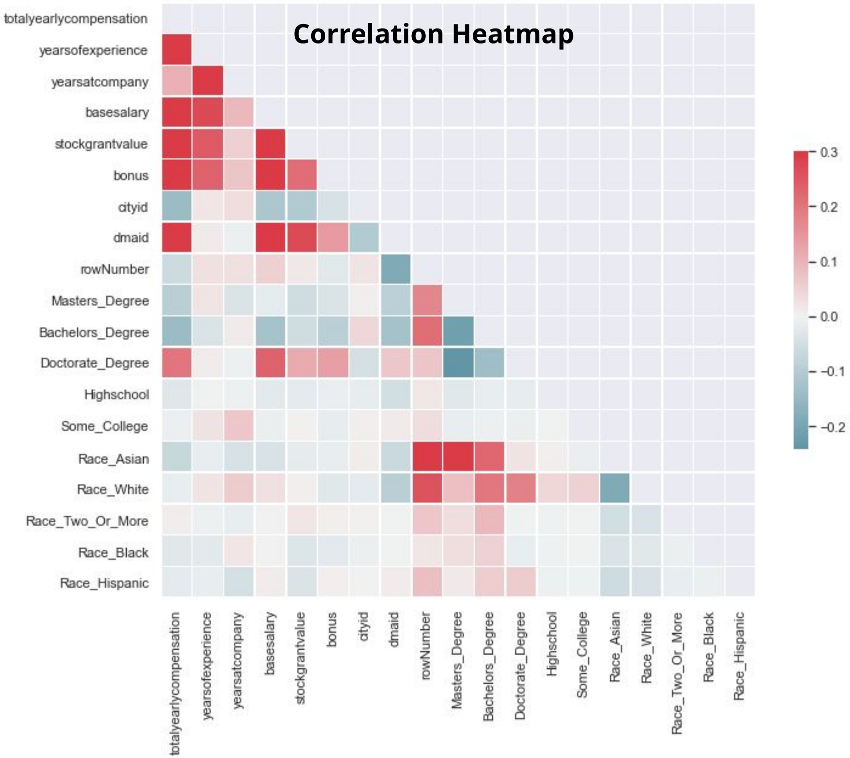


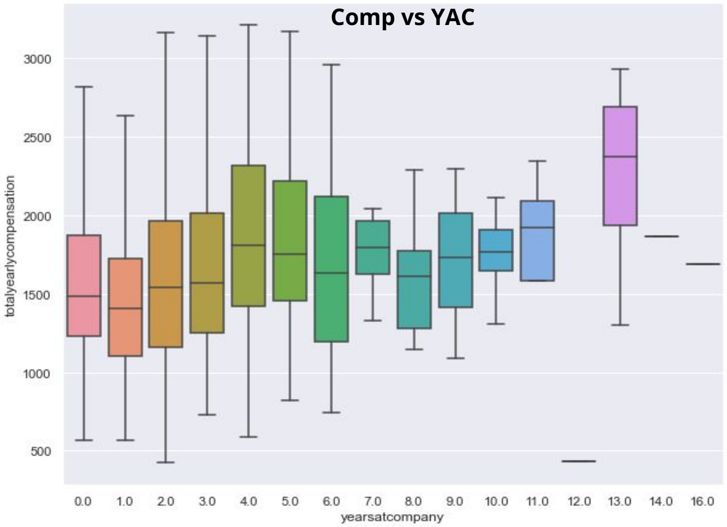
Distribution of Companies

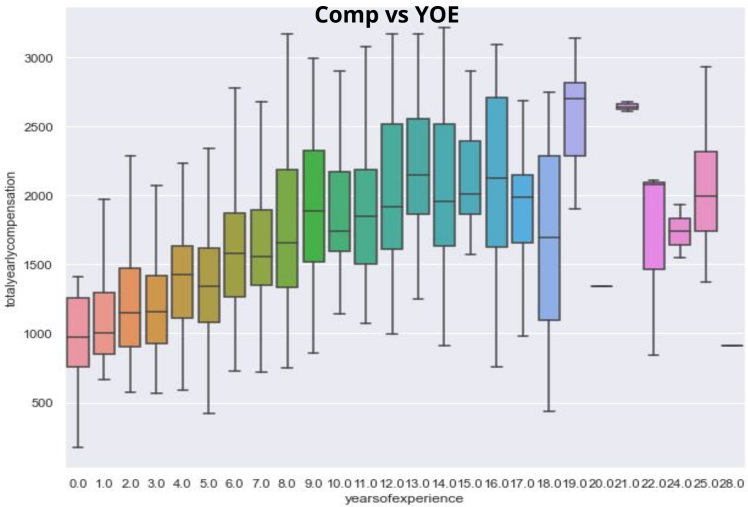


Correlation Between Features

- Both YAC and YOE are promising features for modeling comp.
- No features are strongly correlated (|all corr| < 0.3)


Correlation Heatmap


Comp vs YAC


Comp vs YOE

Is a Graduate Degree Worth It?

Problem: Does a graduate degree make a statistically significant difference in total compensation?

- Calculate P-value
- Calculate Confidence Interval

Model: Welch 2 Sample T-test

- **Null Hypothesis**: mean compensation between education levels are equal
- **Alternate Hypothesis**: mean compensation of graduate level is greater
- **P-value**= 2.28e-09

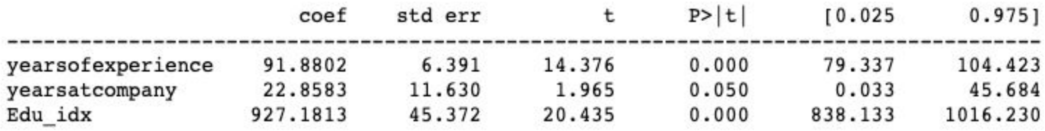
Takeaways:

- Reject Null Hypothesis – individuals with graduate degrees earn more
- Confidence Interval **(0.133 0.261) –** Graduate Degrees boost earnings by 14% to 29%

Predict Individualized Graduate Degree Premium

Objective: Isolate the contribution of graduate degrees in modelling total compensation.

- Multiple Linear Regression model
- Selected Features: YOE, YAC, Education Level



```
                       coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
yearsofexperience   91.8802      6.391     14.376      0.000      79.337     104.423
yearsatcompany      22.8583     11.630      1.965      0.050       0.033      45.684
Edu_idx            927.1813     45.372     20.435      0.000     838.133    1016.230
```

Takeaways:

- All else held equal, a graduate degree provides a premium of 927 * (cost of living index)

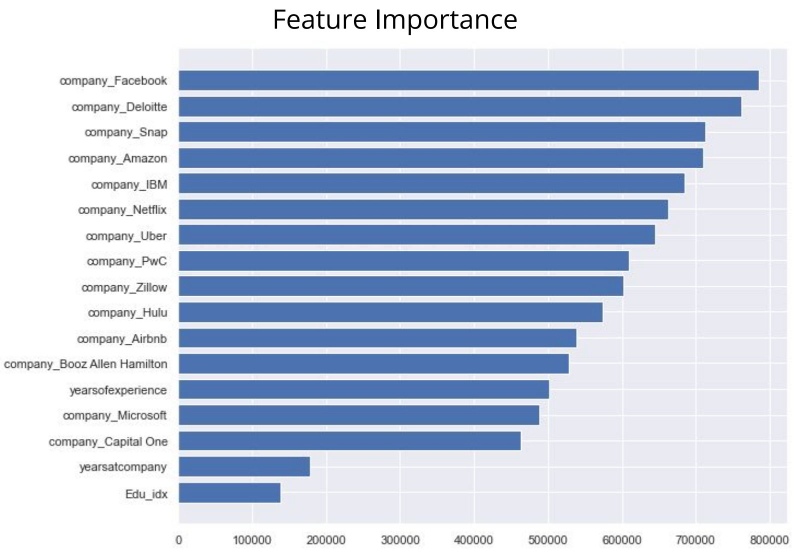
Quantifying Feature Importance with XGBoost

Objective: Determining importance of education relative to other features.

- XGBoost Regression Model
- Used one-hot encoding for company information

Takeaways:

- Education level is less influential than YOE, YAC, and top companies



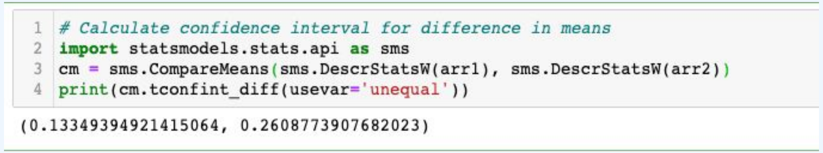
Further Work

- Model Lifetime Utility/Earnings to determine graduate degree value
- Normalize job-grade levels across companies to better model career trajectory
- Cost of living for global cities instead of US states
- Evaluate impact of degrees on access to top companies: Big tech and unicorns pay top-dollar for talent

Models: Welch 2-sample T-test, Linear Regression, XGBoost

This is the confidence interval for the difference in means for our 2-sample t-test. Since it is strictly positive, and the p-value is near 0, we reject the null hypothesis. Thus we establish that graduate degrees do significantly impact compensation positively.

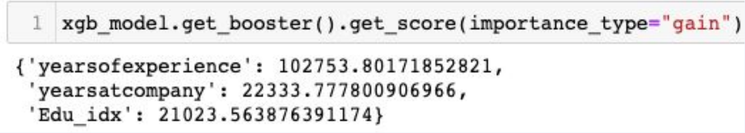
T-test



```
# Calculate confidence interval for difference in means
import statsmodels.stats.api as sms
cm = sms.CompareMeans(sms.DescrStatsW(arr1), sms.DescrStatsW(arr2))
print(cm.tconfint_diff(usevar='unequal'))
```

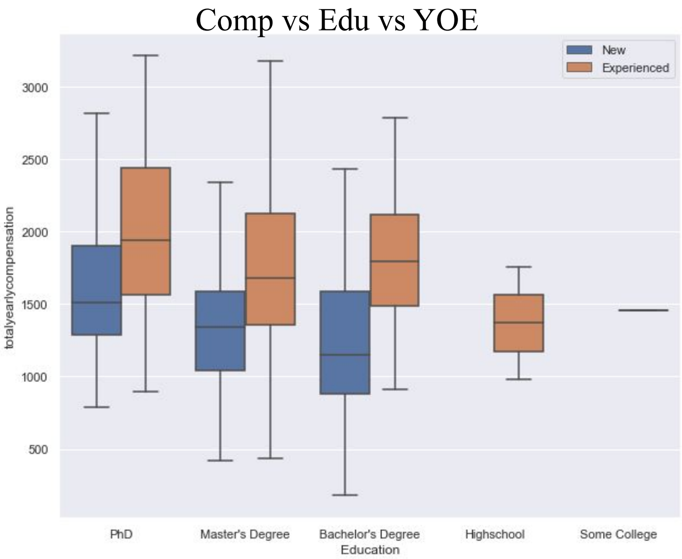
```
(0.13349394921415064, 0.2608773907682023)
```

XGBoost Regression Feature Importance
When determining total compensation, YOE is the most important feature. YAC and Education play similar roles of importance.



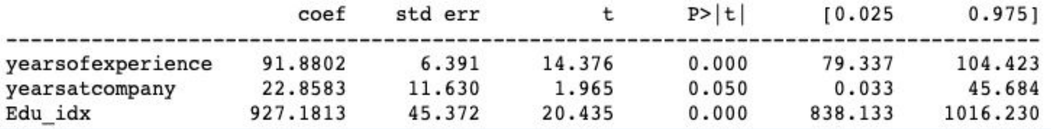
```
xgb_model.get_booster().get_score(importance_type="gain")
```

```
{'yearsofexperience': 102753.80171852821,
 'yearsatcompany': 22333.777800906966,
 'Edu_idx': 21023.563876391174}
```



Comp vs Edu vs YOE

Multiple Linear Regression Model All the features are positively correlated with total compensation, with all coefficient confidence intervals being strictly positive.



```
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
yearsofexperience    91.8802      6.391     14.376      0.000      79.337     104.423
yearsatcompany       22.8583     11.630      1.965      0.050       0.033      45.684
Edu_idx             927.1813     45.372     20.435      0.000     838.133    1016.230
```