

# **Battling Bots**

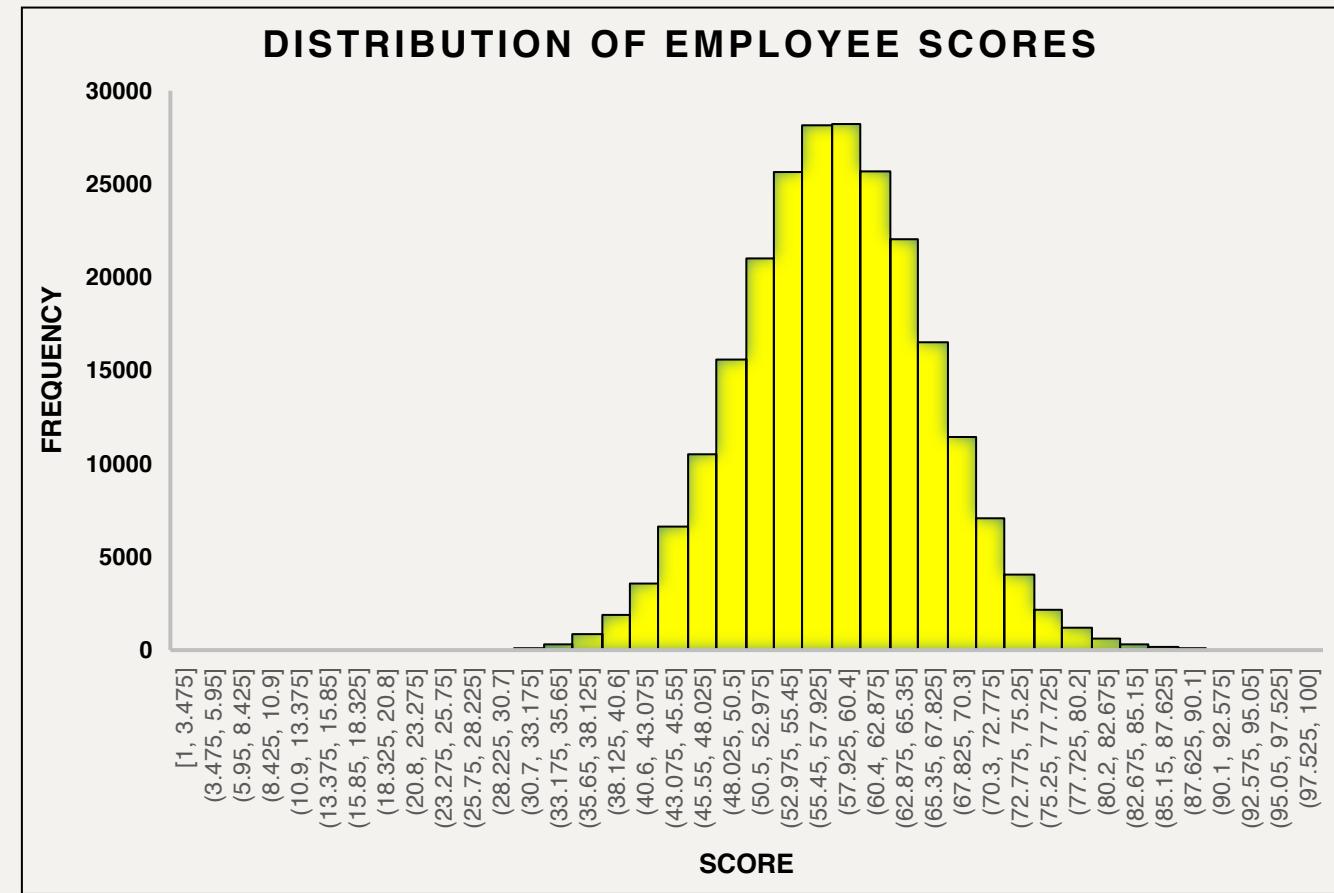
Kyndall Butler



# ***Understanding the Data:***

## Methodology:

- Is the data normally distributed?
- How many values are missing?
- Is it appropriate to impute the data before training/creating the model?



The **Shapiro-Wilk test** for normality failed to reject the null hypothesis that the data is normal with a **p-value = 2.5959**; the data is normally distributed.



# **Data Manipulation**

- Dropping observations with missing data:

233,936 observations -> 65,616 observations

- Evaluating the distribution of the new dataset (subset):

The Shapiro-Wilk test yielded identical results to the original dataset

- Converting variables to float and integer variables.



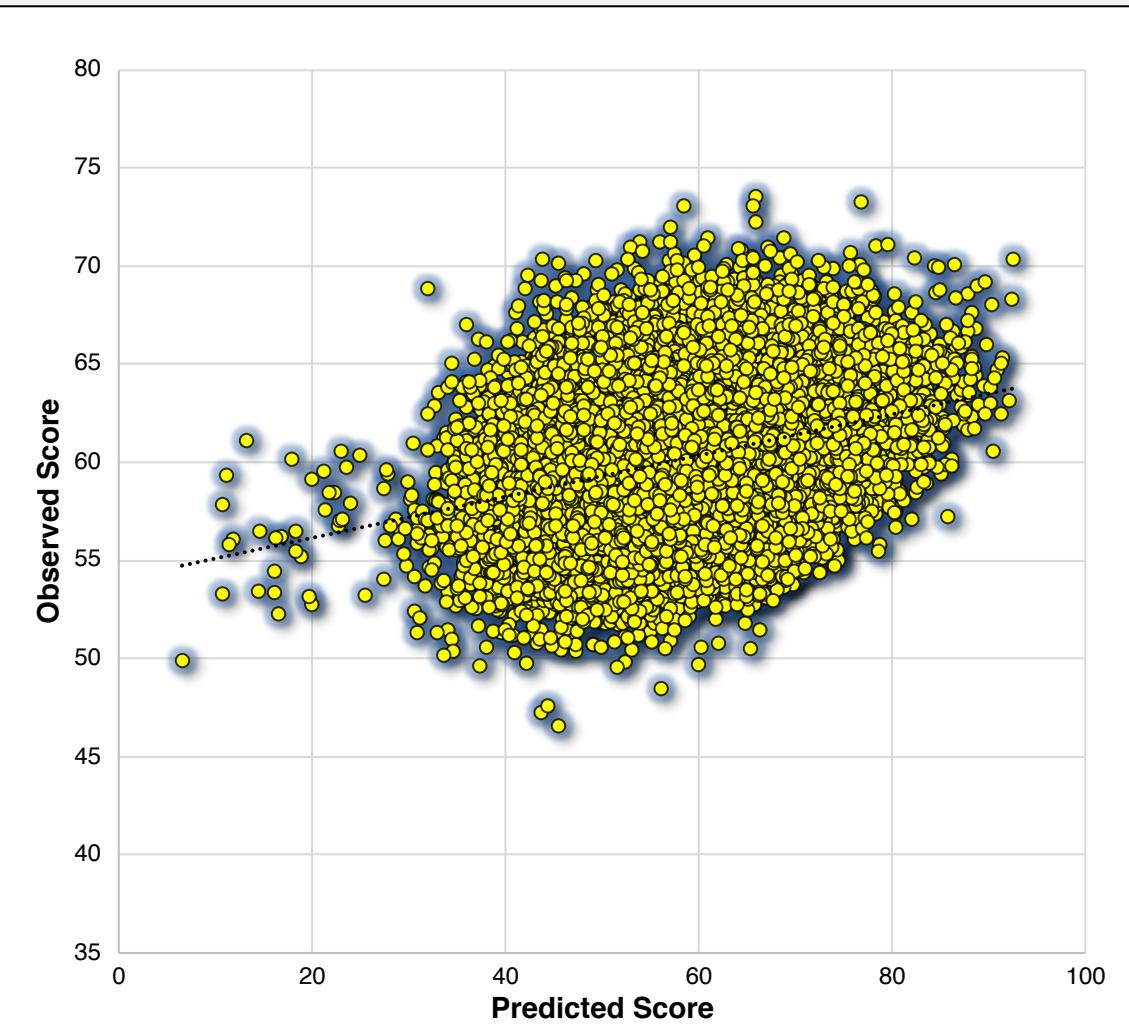
# ***Building the Model:***

- The regression model was produced from cross-validation.

- The employee scores were compared to scores predicted from the regression equation.

Score = 210.88 + 0.08(age) -4.2825(problem solving skill) - 0.2768(technology skill) - 0.0014(recent income) + 0.1231(english skill) + 0.0014(total jobs)

- The mean squared error between the employee scores and employee predicted scores was 3.897.



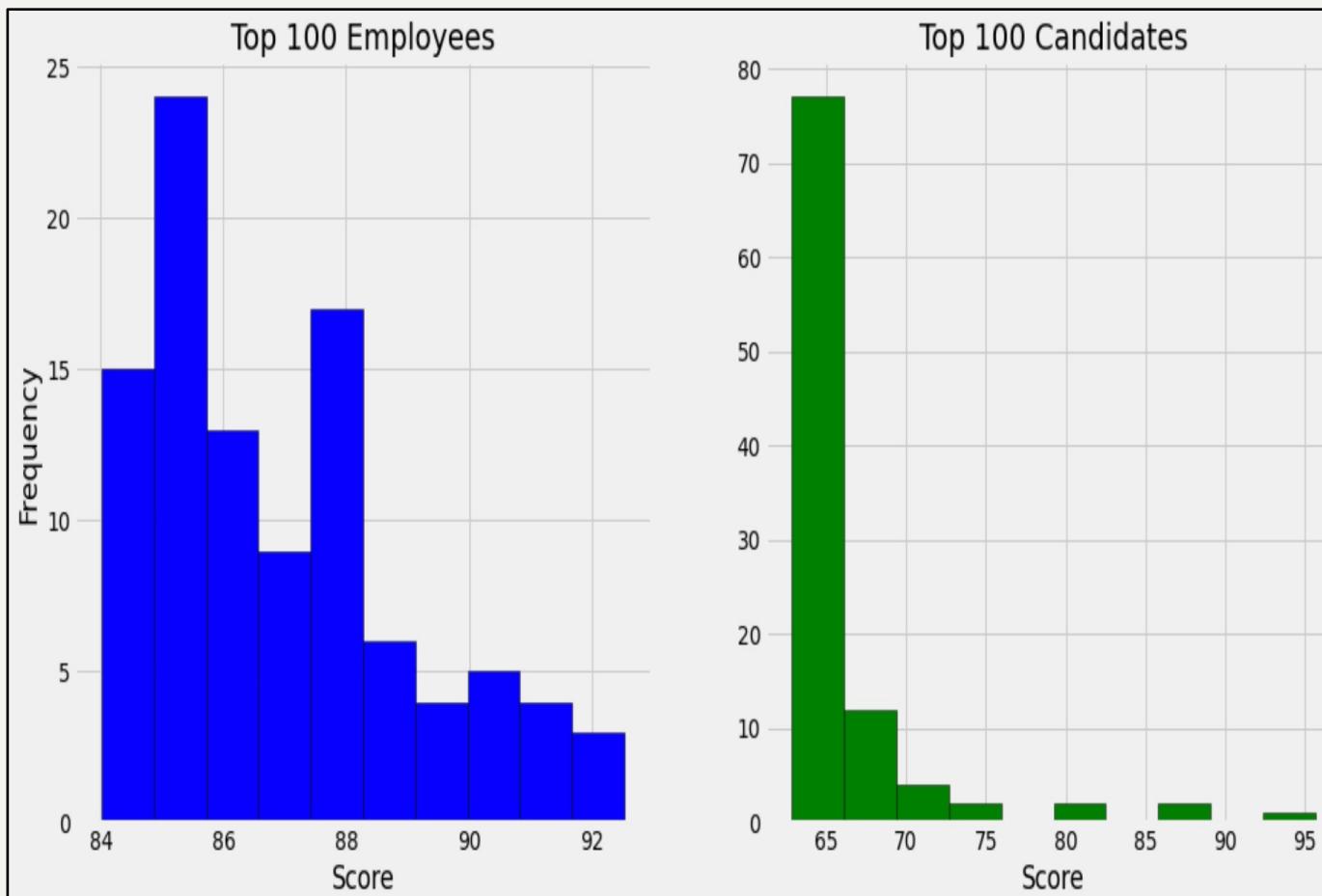
# **Model Deployment**

- The candidate scores were calculated using the regression equation.
  - The distribution of candidate scores seemed to mirror the distribution of employee scores:
    - + seemingly similar averages
    - + seemingly similar q1, q2, and q3
    - + Seemingly similar maximums

	<b>EMPLOYEE SCORES</b>	<b>CANDIDATE SCORES</b>
<b>mean</b>	58.2124	60.6036
<b>std</b>	8.1611	4.9316
<b>min</b>	25.5786	6.5300
<b>25%</b>	58.3162	52.7500
<b>50%</b>	60.5602	58.1400
<b>75%</b>	62.4661	63.6600
<b>max</b>	95.6380	92.5400



# **Screening the Candidates:**



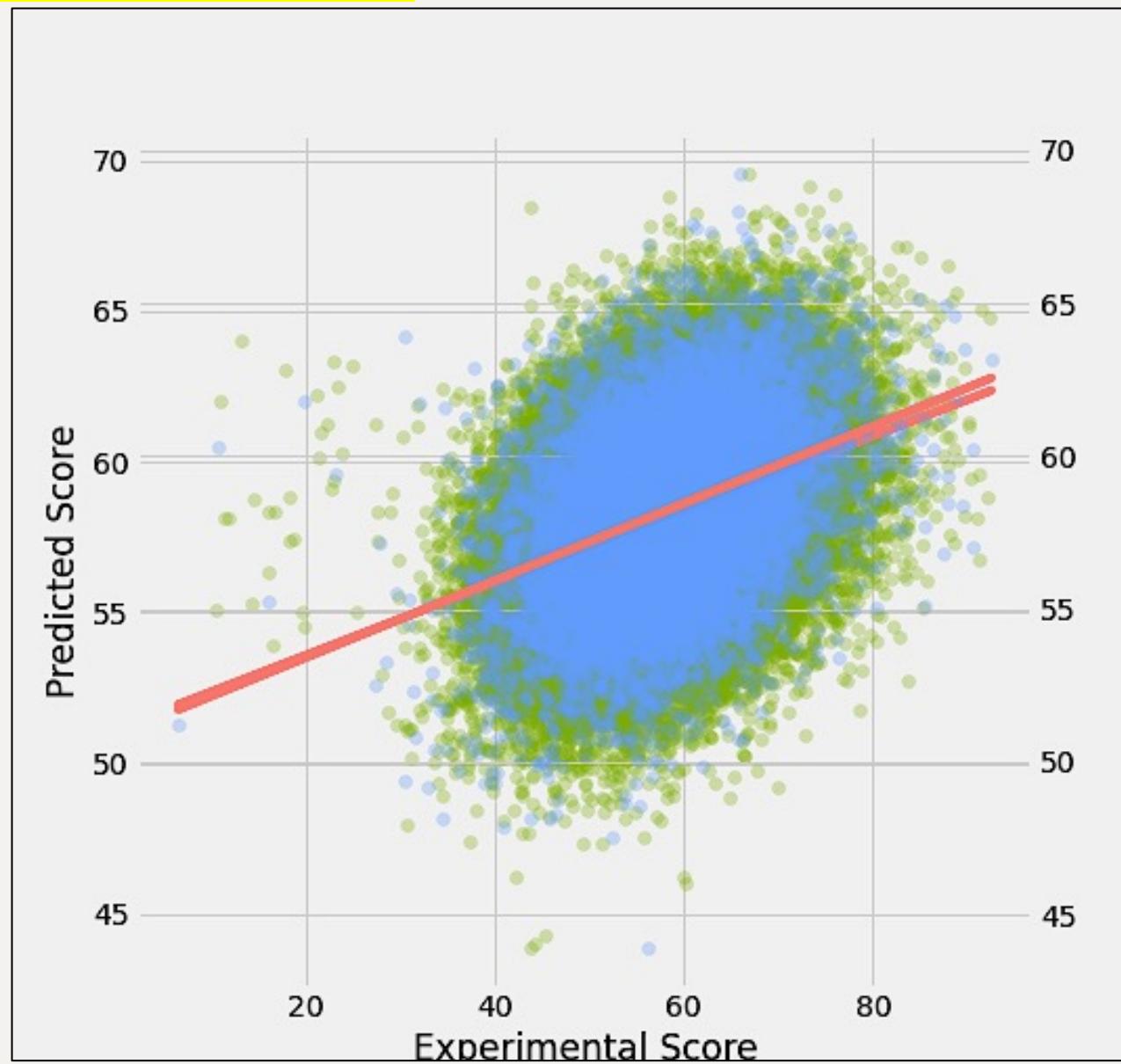
	EMPLOYEE SCORES	CANDIDATE SCORES
mean	86.9423	66.2126
std	2.136	5.3377
min	84.0100	62.8283
25%	85.1675	63.6040
50%	86.4200	64.7732
75%	88.1925	65.8644
max	92.5400	95.6380

- The top 100 candidates were chosen according to calculated scores and compared to the top 100 employees.
  - T-statistic=36.0561, pvalue=5.9784e-89
  - Predicted data follows trend of observed data (positive skewness).



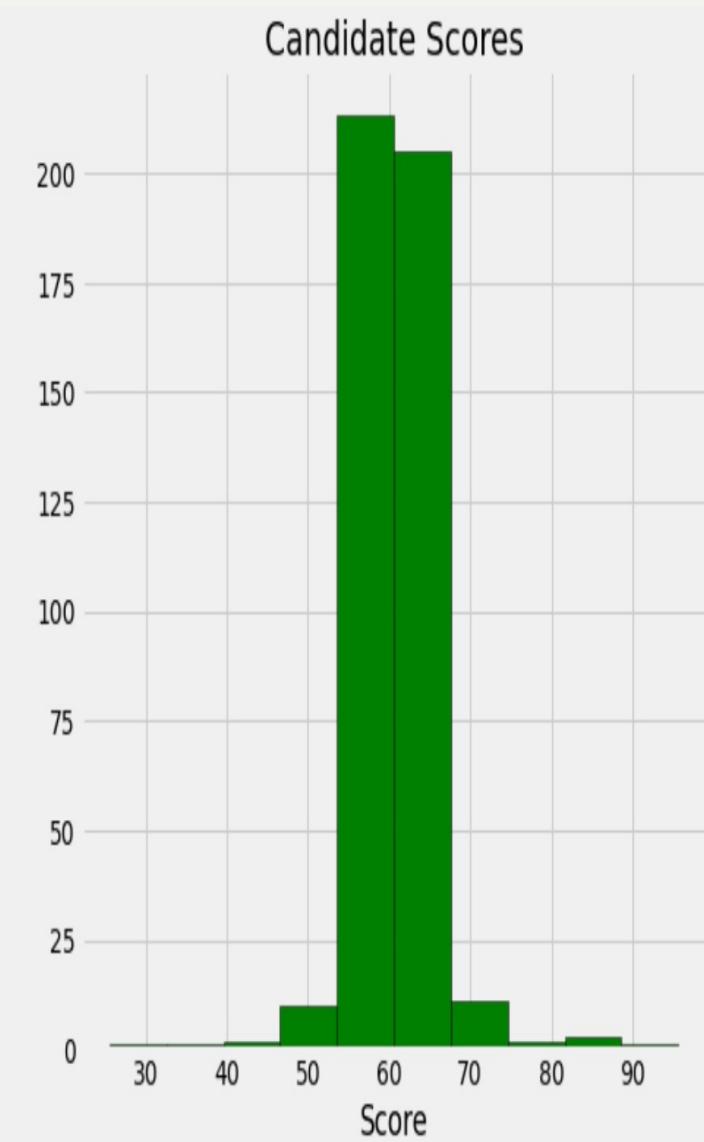
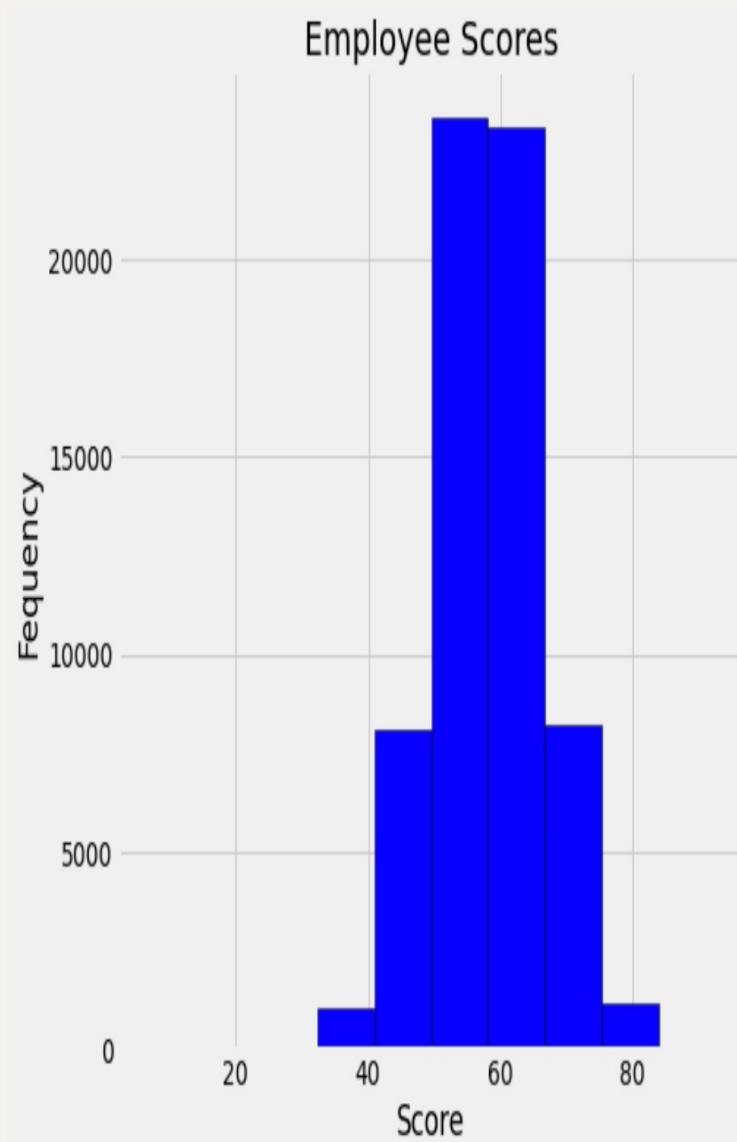
# **Model Evaluation**

- The original employee scores were compared to scores predicted from the regression model.
  - The test set is closer to the trendline and has less outliers.
  - **test set**
  - **train set**



# **Validating the Data:**

- The data was cross-validated (via sklearn in Python).
- Candidate scores almost mirror the distribution of employee scores.



## ***Further Analysis***

- Training the model by purposefully feeding it raw and randomized data.
- Imputing data for the outliers in the original “employees” dataset to train the model to account for more outliers.
- Evaluating the linear regression equation against a quadratic regression equation.
- Averaging more regression models to improve accuracy.

