# Summary

The purpose of this research is to explore the relationship between Phishing and Phishing Education & Awareness Training. Phishing is an Initial Access technique as defined within the MITRE ATT&CK framework and is the intentional sending of misleading or malicious emails for fraudulent purposes (*Phishing, n.d.*). It is estimated that roughly 90% of all successful cyberattacks start with phishing (*CISA*, *n.d.*). Phishing Education & Awareness Training is a common security control or mitigation utilized by many organizations to bolster their security posture by repeatedly testing employees with mock phishing attempts (*User Training, n.d.*). This research will explore its effectiveness.

# Objective

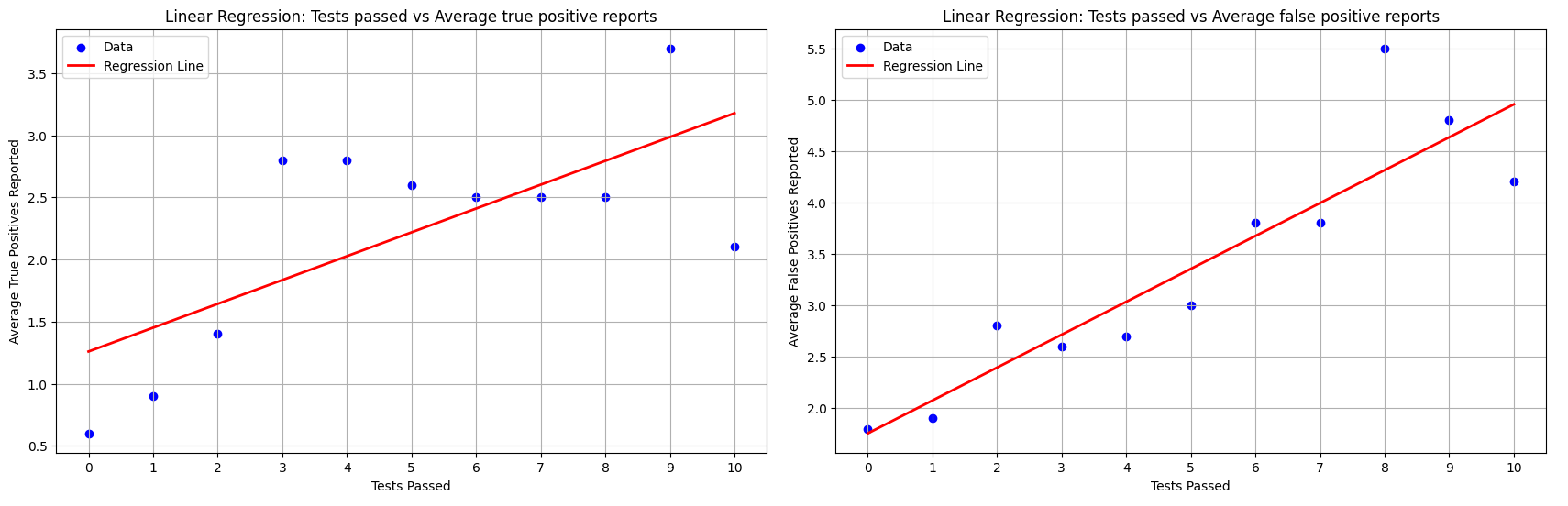
The objective of this research is to explore the validity of Phishing Education & Awareness Training in improving an organization’s security posture against phishing. Numerous metrics and approaches could be used to analyze metrics to help answer this question. Specifically, within this research, an analysis will be performed to help answer the following questions: Do employees with higher training pass rates report more emails for review, and are those reported emails more likely to be accurately identified as phishing?

# Process

The process for this analysis will involve a synthesized report from an organization. This synthesized report is a combination of metrics from a Phish Testing & Awareness Training Platform, and metrics from the organizations' Phish Reporting Queue, lovingly referred to as the *Phish Tank*. The dataset includes roughly one year worth of data. This dataset will be run through a python program to ensure accurate data ingestion, cleaning and filtering of data, statistical analysis of data, and the production of visual elements to help showcase the data. Lastly, the resulting output will be analyzed to help answer the questions outlined within the objective.

# Results

The results from the data analysis indicate that only **66%** of the employees within the sample had made attempts to report emails during the timeframe. Focusing on the employees who did engage, analysis of the data initially indicates that engaging employees do tend to report more real-world emails. This was correlated with the number of successful phishing tests passed. However, the analysis also indicated that those employees also reported more false positive emails as well. For example, employees who had passed nine tests reported an average of **3.7** true positives. However, that same cohort reported an average of **4.8** false positives. In fact, the correlation between the number of tests passed and the number of false positive reported was slightly stronger as seen below.



The above figure was generated with the following code:

#Calculate min and max for column 'test\_sum\_Reported' for x axis ticks

tsr\_min=selected\_df['test\_sum\_Reported'].min()

tsr\_max=selected\_df['test\_sum\_Reported'].max()

# Reshape data

X = merged\_df[['test\_sum\_Reported']]

y = merged\_df['avg\_real\_truepositive']

# Fit linear regression model

model = LinearRegression()

model.fit(X, y)

# Predict y values

y\_pred = model.predict(X)

# Plot the data and the regression line

plt.figure(figsize=(10, 6))

plt.scatter(X, y, color='blue', label='Tests Passed')

plt.plot(X, y\_pred, color='red', linewidth=2, label='Regression Line')

plt.title('Linear Regression: Tests passed vs Average true positive reports')

plt.xlabel('Tests Passed')

plt.ylabel('Average True Positives Reported')

plt.xticks(range(tsr\_min,tsr\_max+1,1))

plt.legend(loc='upper left')

plt.grid(True)

plt.show()

#Calculate min and max for column 'test\_sum\_Reported' for x axis ticks

tsr\_min=selected\_df['test\_sum\_Reported'].min()

tsr\_max=selected\_df['test\_sum\_Reported'].max()

# Reshape data

X = merged\_df[['test\_sum\_Reported']]

y = merged\_df['avg\_real\_falsepositive']

# Fit linear regression model

model = LinearRegression()

model.fit(X, y)

# Predict y values

y\_pred = model.predict(X)

# Plot the data and the regression line

plt.figure(figsize=(10, 6))

plt.scatter(X, y, color='blue', label='Data')

plt.plot(X, y\_pred, color='red', linewidth=2, label='Regression Line')

plt.title('Linear Regression: Tests passed vs Average false positive reports')

plt.xlabel('Tests Passed')

plt.ylabel('Average False Positives Reported')

plt.xticks(range(tsr\_min,tsr\_max+1,1))

plt.legend(loc='upper left')

plt.grid(True)

plt.show()

# Conclusion

The analysis performed above does initially indicate that people who have passed more tests do report more emails as potential phishing. However, the accuracy of those reports does not seem to improve over time, but rather only the volume. This could be attributed to examining a large period of time homogeneously.

Additionally, there was a significant portion of the population that did not meaningfully engage with reporting emails. This population could be addressed in a different manner to improve outcomes as potential issues could include a lack of receiving suspicious emails, technical issues, or knowledge gaps.

# Further Steps

To further improve this research, it is recommended that the organization attempts to determine why some employees do not engage, and correct that if possible. This would allow the organization to gather more data for analysis. Additionally, this analysis could be improved by comparing month by months results to see how the employee responded following a pass or fail during the testing period.

# References:

1. *Phishing*. Phishing, Technique T1566 - Enterprise | MITRE ATT&CK®. (n.d.). https://attack.mitre.org/techniques/T1566/
2. Shields up: Guidance for families: CISA, Cybersecurity and Infrastructure Security Agency CISA. (n.d.). https://www.cisa.gov/shields-guidance-families3
3. *User Training*. User Training, Mitigation M1017 - Enterprise | MITRE ATT&CK®. (n.d.). https://attack.mitre.org/mitigations/M1017/