Cybersecurity

Analyzing Cybercrime Losses Across U.S. States

Team MIHH-GKAN

Introduction

Technology in the 21st century are more prevalent than ever due to constant innovation, with its use becoming integrated in the daily lives of people and is becoming a great importance to nations, economies, the government, and institutions worldwide. On the internet, there are no boundaries. Because of the immense step towards efficiency and connectivity, there is a downside that poses a threat that ranges to politics and the economy. Threat actors exists behind the scenes, driven primarily by financial gain. As a result, the team developed an interest in exploring damages this has caused on the lives of the people and business functionalities. We took an initiative to examine the area of cybersecurity regarding the losses and damage from the exploitation of vulnerabilities and blind spots that come through technology usage and how it affected society.

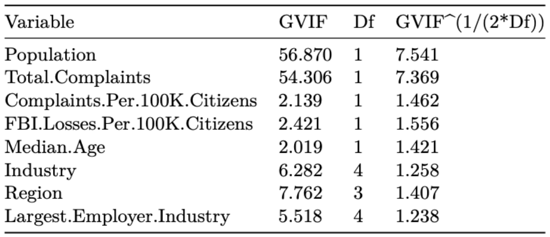
Background research source: https://www.ibm.com/think/topics/cybersecurity.
2024-Report-on-the-Cybersecurity-Posture-of-the-United-States.pdf
www.snhu.edu/about-us/newsroom/stem/what-is-cyber-security.

Research Questions

Our analysis addresses the following research questions:

1. Do states with higher populations have higher monetary losses?
2. Do states with a higher number of complaints have higher total monetary losses?
3. Do states with Finance & Insurance as their dominant industry have higher losses from cybersecurity attacks compared to states dominated by other industries?

Variable Selection + Multicollinearity

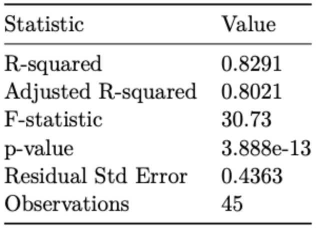


Variable	GVIF	Df	GVIF^(1/(2*Df))
Population	56.870	1	7.541
Total.Complaints	54.306	1	7.369
Complaints.Per.100K.Citizens	2.139	1	1.462
FBI.Losses.Per.100K.Citizens	2.421	1	1.556
Median.Age	2.019	1	1.421
Industry	6.282	4	1.258
Region	7.762	3	1.407
Largest.Employer.Industry	5.518	4	1.238

- Checked for multicollinearity by calculating Variance Inflation Factors (VIF) for all variables in our initial full model.
- Removed Total Complaints from the model
- Used stepwise in both directions with AIC criterion for optimal level
- The process converged on three predictors: Population, FBI Losses Per 100K Citizens, and Industry.

Final Model Assessment

Final model explains 80.21% of the variation in cybercrime losses (Adj.R^2 = 0.8021). Model is highly statistically significant (F=30.73, p<0.0001).



Statistic	Value
R-squared	0.8291
Adjusted R-squared	0.8021
F-statistic	30.73
p-value	3.888e-13
Residual Std Error	0.4363
Observations	45

Limitations:

- Variable Transformation - Transforming variables may make model more complex.
- Statistical Significance – Other reasons for a low p-value may be a 'bad' sample or by random chance.

Model Building

Step 1: Quantitative Predictors

- Began exploratory analysis identifying Population and FBI Losses per 100K as the strongest quantitative predictors of cybercrime losses.
- Started with a simple model using *Population* alone.
- Added *FBI Losses per 100K* to improve explanatory power.
- A nested F-test ($p < 0.05$) confirmed that adding *FBI Losses per 100K* significantly improved model fit.

Step 2: Qualitative Predictors

- Included Industry Type as a categorical variable to test for qualitative differences among states.
- Individual t-tests ($p < 0.001$) showed both quantitative predictors were highly significant.

Step 3: Model Significance

- The global F-test ($p < 0.001$) verified that the overall model is statistically significant.
- Adjusted R^2 = 0.93, indicating excellent explanatory power.

Diagnostics

- Residual diagnostics confirmed that the log transformation satisfied normality and variance assumptions.
- Minor multicollinearity remains between *Population* and exposure-related variables.

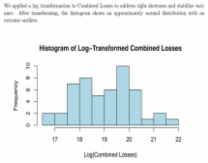
Final Model Summary

$$\log(\text{Combined Losses}) = 16.92 + (1.187 \times 10^{-7}) \times \text{Population} + (2.709 \times 10^{-7}) \times \text{FBI.Losses.Per.100K} + 0.711 \times \text{Manufacturing} + 0.534 \times \text{Real.Estate} + 0.498 \times \text{Other} - 0.563 \times \text{Mining.Oil.Gas}$$

Where industry coefficients represent deviations from the reference category (Finance & Insurance)

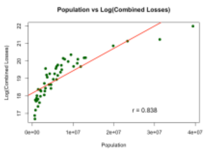
The final model is parsimonious, statistically valid, and captures the dominant factors, influencing cybercrime losses across the U.S. states.

Exploratory Data Analysis

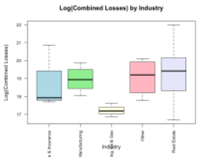


We applied a log transformation to Combined Losses to address right skewness and stabilize variance. After transforming, the histogram shows an approximately normal distribution with no extreme outliers.

- After transformation: approximately normal distribution (bell shaped)
- Original data was right skewed and unsuitable for regression
- Range: [16.68, 21.99] (no extreme outliers



- R = 0.838 -> strong positive relationship
- States with more people tend to lose more money overall
- Points are in an upward trend, so population is a strong predictor of how big losses are



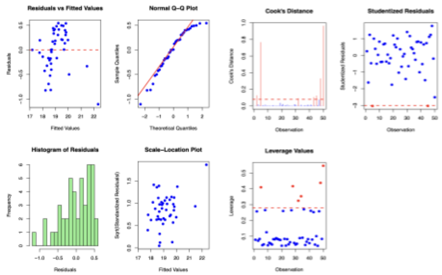
- Real Estate and Finance & Insurance states have higher median losses
- Mining and Oil & Gas are lower, those states lose less money to cybercrime
- Population based factors dominate, while industry and region have marginal effects

Residual Analysis

Checking for Regression Assumptions:

Linearity: Residuals scattered randomly around zero with no clear pattern.
Constant Variance: Spread of residuals remains consistent across all fitted values.
Normality: Points closely follow diagonal line on Q-Q plot.
Independence: Cross-sectional data of different states ensures independence.
No violations were made.

Influential Observations:
Observations that exceeded the threshold of Cook's' Distance, Studentized Residuals, and Leverage Values were removed.
Five states: California, New York, Vermont, West Virginia, and Wyoming were removed.



Additional Techniques

With PRESS ≈ RSS (7.23), Predicted R^2 = 0.80, and RMSE ≈ 0.5 log-units (~12–15%), the model shows strong fit, no overfitting, and reliable predictive accuracy for new data.

Data Summary

Population of Interest

- Cybercrime losses across 50 U.S. states

Data Sources

- **FBI Internet Crime Complaint Center (IC3, 2024):** Provided state-level data on complaints and total monetary losses
- **U.S. Census Bureau (2024):** Provided state population estimates for per-capita calculations
- **Economic Industry Reports:** State's dominant industry identified from *Visual Capitalist* and *World Population Review*

Variables in the Model

- **Response Variable:** Combined Losses (FBI + FTC losses, in dollars; log-transformed)
- **Quantitative Predictors:** Population, FBI Losses per 100K citizens
- **Qualitative Predictor:** Industry Type

Data Preparation Steps

- Removed commas from dollar values for consistency
- Applied natural log transformation to Combined Losses to correct right skewness
- Merged datasets by state name to form the final dataset of 50 observations

Conclusion

- Population and FBI losses per 100K residents are the strongest predictors of total cybercrime losses across U.S. states.
- The model explains 80% of variation in combined losses (Adjusted R^2 = 0.80), demonstrating high reliability.
- Both quantitative predictors are highly significant ($p < 0.001$), meaning states with larger populations and higher exposure experience greater losses.
- Industry effects are modest:
  - *Real Estate* and *Manufacturing* states show slightly higher losses.
  - *Mining/Oil & Gas* states report slightly lower losses than *Finance & Insurance*.
- The log transformation stabilized variance and satisfied model assumptions.

Key insight: Cybercrime losses scale mainly with population and exposure, while industry plays a secondary role.

Future Reseasrch

- Demographics: Incorporate age group data from the *2024 IC3 Report* to assess which populations are most vulnerable and how that varies by region.
- Attack Types: Analyze types of cybercrime (e.g., phishing, ransomware) to identify which cause the greatest financial damage and prioritize prevention efforts.
- Global Comparison: Expand analysis beyond the U.S. to explore how countries with lower cybercrime rates succeed in mitigation.

Works Cited

Research Background

Jonker, A., Lindemulder, G., & Kosinski, M. (2025, November 17). *What is Cybersecurity?* IBM. https://www.ibm.com/think/topics/cybersecurity
Office of the National Cyber Director. (2024, May). *2024 Report on the Cybersecurity Posture of the United States.* National Archives.
Patterson, N. (2025, September 19). *What Is Cybersecurity and Why Is It Important?* Southern New Hampshire University. https://www.snhu.edu/about-us/newsroom/stem/what-is-cyber-security

Data Sources

Federal Bureau of Investigation. (2024). *Internet Crime Report 2024.* Internet Crime Complaint Center (IC3). https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf
U.S. Census Bureau. (n.d.). *Census Regions and Divisions of the United States.* https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
Ross, J. (2025, July 28). *America's Economic Engines: The Biggest Industry in Every State.* Visual Capitalist. https://www.visualcapitalist.com/sp/ter01-the-biggest-industry-in-every-state/
Upwind. (2025, February 27). *Fraud, Phishing, and Internet Scams: Which States Lose the Most Money to Cybercrime?* https://www.upwind.io/industry-research/cybercrime-cost-by-state
World Population Review. (2025). *Largest Employer by State 2025.* https://worldpopulationreview.com/state-rankings/largest-employer-by-state
World Population Review. (2025). *Median Age by State 2025.* https://worldpopulationreview.com/state-rankings/median-age-by-state