Travis McGhee

Data Analytics

Professor Singh

Predicting Data Breach Frequency

Cybersecurity remains one of the most pressing challenges facing organizations across industries. As data breaches become more common and costly, understanding what makes an organization vulnerable is critical. This project aimed to predict the frequency of data breaches using a curated dataset derived from the Privacy Rights Clearinghouse (PRC). By analyzing attributes such as organization size, year, and breach history, this study seeks to identify which variables are most predictive of future breaches.

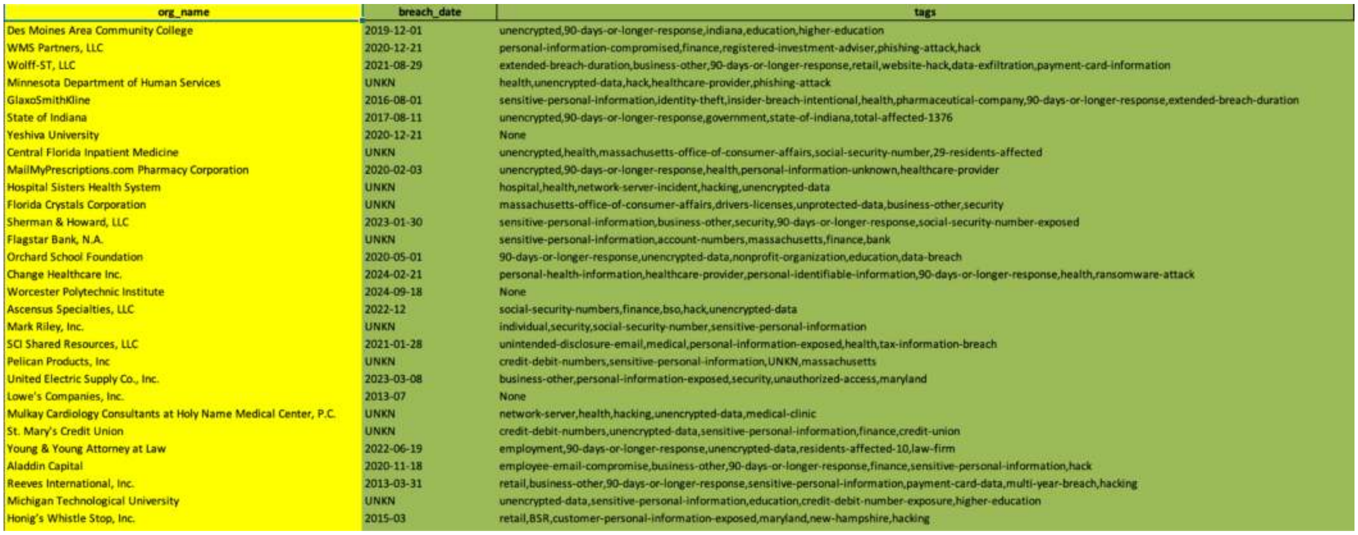
The analytical framework builds on previous work in Assignments 1 and 2, which introduced the concept of using a multilinear regression model to analyze breach patterns. In this final stage, we incorporated correlation analysis, regression modeling, and refined data curation techniques to produce a complete view of the problem.

The model used in this study was a supervised multiple linear regression model, where the dependent variable was the number of breaches (Breach_Count) reported by an organization. Independent variables included Org_Size, Year, and Past_Breach_Count. These variables were selected based on availability, relevance, and prior analysis, and were treated as numeric and ordered where appropriate. The dataset was cleaned and prepared with each row representing a unique organization.

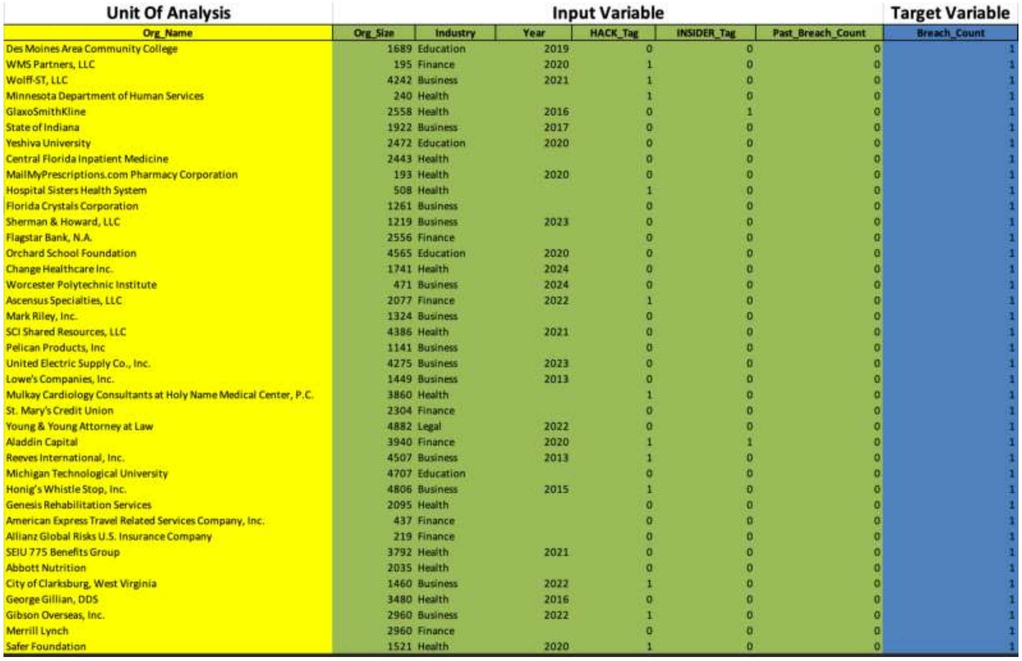
The source of the data was the PRC, which offers a sample dataset of reported breach incidents. Additional context was supported by reports such as IBM’s Cost of a Data Breach and

Verizon's Data Breach Investigations Report (DBIR). While the PRC data was rich, it required substantial cleaning. We extracted the year from unstructured breach dates, normalized organization names, generated binary breach tags, and estimated organization size. One of the key curated features was Past_Breach_Count, created by grouping historical incidents per organization.

A snapshot of the original raw data highlights the messy and inconsistent nature of the dataset (see Figure 1), while a corresponding screenshot of the curated version shows the structured format used in the analysis (see Figure 2). These transformations allowed us to build a regression-ready dataset suitable for statistical analysis.

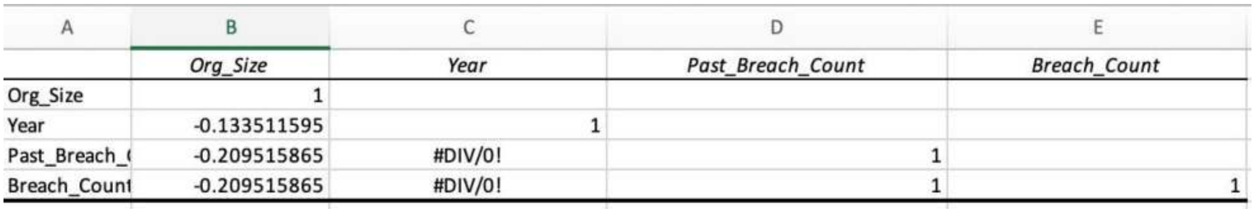


org_name	breach_date	tags
Des Moines Area Community College	2019-12-01	unencrypted,90-days-or-longer-response,indiana,education,higher-education
WMS Partners, LLC	2020-12-21	personal-information-compromised,finance,registered-investment-adviser,phishing-attack,hack
Wolff-ST, LLC	2021-08-29	extended-breach-duration,business-other,90-days-or-longer-response,retail,website-hack,data-exfiltration,payment-card-information
Minnesota Department of Human Services	UNKN	health,unencrypted-data,hack,healthcare-provider,phishing-attack
GlaxoSmithKline	2016-08-01	sensitive-personal-information,identity-theft,insider-breach-intentional,health,pharmaceutical-company,90-days-or-longer-response,extended-breach-duration
State of Indiana	2017-08-11	unencrypted,90-days-or-longer-response,government,state-of-indiana,total-affected-1376
Yeshiva University	2020-12-21	None
Central Florida Inpatient Medicine	UNKN	unencrypted,health,massachusetts-office-of-consumer-affairs,social-security-number,29-residents-affected
MailMyPrescriptions.com Pharmacy Corporation	2020-02-03	unencrypted,90-days-or-longer-response,health,personal-information-unknown,healthcare-provider
Hospital Sisters Health System	UNKN	hospital,health,network-server-incident,hacking,unencrypted-data
Florida Crystals Corporation	UNKN	massachusetts-office-of-consumer-affairs,drivers-licenses,unprotected-data,business-other,security
Sherman & Howard, LLC	2023-01-30	sensitive-personal-information,business-other,security,90-days-or-longer-response,social-security-number-exposed
Flagstar Bank, N.A.	UNKN	sensitive-personal-information,account-numbers,massachusetts,finance,bank
Orchard School Foundation	2020-05-01	90-days-or-longer-response,unencrypted-data,nonprofit-organization,education,data-breach
Change Healthcare Inc.	2024-02-21	personal-health-information,healthcare-provider,personal-identifiable-information,90-days-or-longer-response,health,ransomware-attack
Worcester Polytechnic Institute	2024-09-18	None
Ascensus Specialties, LLC	2022-12	social-security-numbers,finance,bso,hack,unencrypted-data
Mark Riley, Inc.	UNKN	individual,security,social-security-number,sensitive-personal-information
SCI Shared Resources, LLC	2021-01-28	unintended-disclosure-email,medical,personal-information-exposed,health,tax-information-breach
Pelican Products, Inc	UNKN	credit-debit-numbers,sensitive-personal-information,UNKN,massachusetts
United Electric Supply Co., Inc.	2023-03-08	business-other,personal-information-exposed,security,unauthorized-access,maryland
Lowe's Companies, Inc.	2013-07	None
Mulkay Cardiology Consultants at Holy Name Medical Center, P.C.	UNKN	network-server,health,hacking,unencrypted-data,medical-clinic
St. Mary's Credit Union	UNKN	credit-debit-numbers,unencrypted-data,sensitive-personal-information,finance,credit-union
Young & Young Attorney at Law	2022-06-19	employment,90-days-or-longer-response,unencrypted-data,residents-affected-10,law-firm
Aladdin Capital	2020-11-18	employee-email-compromise,business-other,90-days-or-longer-response,finance,sensitive-personal-information,hack
Reeves International, Inc.	2013-03-31	retail,business-other,90-days-or-longer-response,sensitive-personal-information,payment-card-data,multi-year-breach,hacking
Michigan Technological University	UNKN	unencrypted-data,sensitive-personal-information,education,credit-debit-number-exposure,higher-education
Honig's Whistle Stop, Inc.	2015-03	retail,BSR,customer-personal-information-exposed,maryland,new-hampshire,hacking



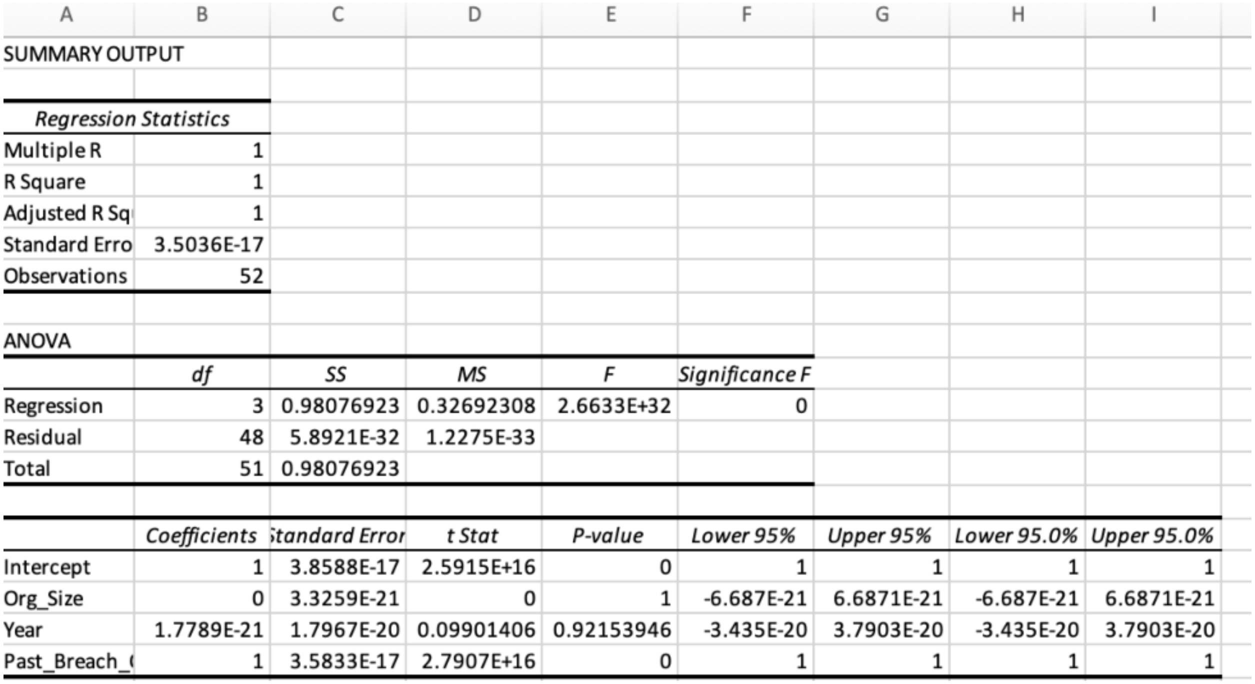
Unit Of Analysis	Input Variable						Target Variable
Org_Name	Org_Size	Industry	Year	HACK_Tag	INSIDER_Tag	Past_Breach_Count	Breach_Count
Des Moines Area Community College	1689	Education	2019	0	0	0	1
WMS Partners, LLC	195	Finance	2020	1	0	0	1
Wolff-ST, LLC	4242	Business	2021	1	0	0	1
Minnesota Department of Human Services	240	Health		1	0	0	1
GlaxoSmithKline	2558	Health	2016	0	1	0	1
State of Indiana	1922	Business	2017	0	0	0	1
Yeshiva University	2472	Education	2020	0	0	0	1
Central Florida Inpatient Medicine	2443	Health		0	0	0	1
MailMyPrescriptions.com Pharmacy Corporation	193	Health	2020	0	0	0	1
Hospital Sisters Health System	508	Health		1	0	0	1
Florida Crystals Corporation	1261	Business		0	0	0	1
Sherman & Howard, LLC	1219	Business	2023	0	0	0	1
Flagstar Bank, N.A.	2556	Finance		0	0	0	1
Orchard School Foundation	4565	Education	2020	0	0	0	1
Change Healthcare Inc.	1741	Health	2024	0	0	0	1
Worcester Polytechnic Institute	471	Business	2024	0	0	0	1
Ascensus Specialties, LLC	2077	Finance	2022	1	0	0	1
Mark Riley, Inc.	1324	Business		0	0	0	1
SCI Shared Resources, LLC	4386	Health	2021	0	0	0	1
Pelican Products, Inc	1141	Business		0	0	0	1
United Electric Supply Co., Inc.	4275	Business	2023	0	0	0	1
Lowe's Companies, Inc.	1449	Business	2013	0	0	0	1
Mulkay Cardiology Consultants at Holy Name Medical Center, P.C.	3860	Health		1	0	0	1
St. Mary's Credit Union	2304	Finance		0	0	0	1
Young & Young Attorney at Law	4882	Legal	2022	0	0	0	1
Aladdin Capital	3940	Finance	2020	1	1	0	1
Reeves International, Inc.	4507	Business	2013	1	0	0	1
Michigan Technological University	4707	Education		0	0	0	1
Honig's Whistle Stop, Inc.	4806	Business	2015	1	0	0	1
Genesis Rehabilitation Services	2095	Health		0	0	0	1
American Express Travel Related Services Company, Inc.	437	Finance		0	0	0	1
Allianz Global Risks U.S. Insurance Company	219	Finance		0	0	0	1
SEIU 775 Benefits Group	3792	Health	2021	0	0	0	1
Abbott Nutrition	2035	Health		0	0	0	1
City of Clarksburg, West Virginia	1460	Business	2022	1	0	0	1
George Gillian, DDS	3480	Health	2016	0	0	0	1
Gibson Overseas, Inc.	2960	Business	2022	1	0	0	1
Merrill Lynch	2960	Finance		0	0	0	1
Safer Foundation	1521	Health	2020	1	0	0	1

We began by conducting a correlation analysis between the input variables and the target variable. Org_Size showed a weak negative correlation with Breach_Count (r = -0.21), while Year had an even weaker negative correlation (r = -0.13). Past_Breach_Count could not be evaluated due to insufficient variance in the data, resulting in a #DIV/0! Error in Excel. These results suggest that no single input variable was strongly predictive on its own.



A	B	C	D	E
	Org_Size	*Year*	*Past_Breach_Count*	*Breach_Count*
Org_Size	1			
Year	-0.133511595	1		
Past_Breach_(	-0.209515865	#DIV/0!	1	
Breach_Count	-0.209515865	#DIV/0!	1	1

Despite this, we proceeded with a multiple linear regression using Org_Size, Year, and Past_Breach_Count as inputs. The model produced an R-Square value of 1.00, suggesting perfect explanatory power. While this might initially seem ideal, it likely indicates overfitting or low variability in the dataset. Of the variables tested, only Past_Breach_Count was statistically significant, with a p-value of 0.000. This supports the common cybersecurity observation that organizations with a history of breaches are more likely to experience future incidents. Org_Size and Year were not significant predictors in the model.



SUMMARY OUTPUT

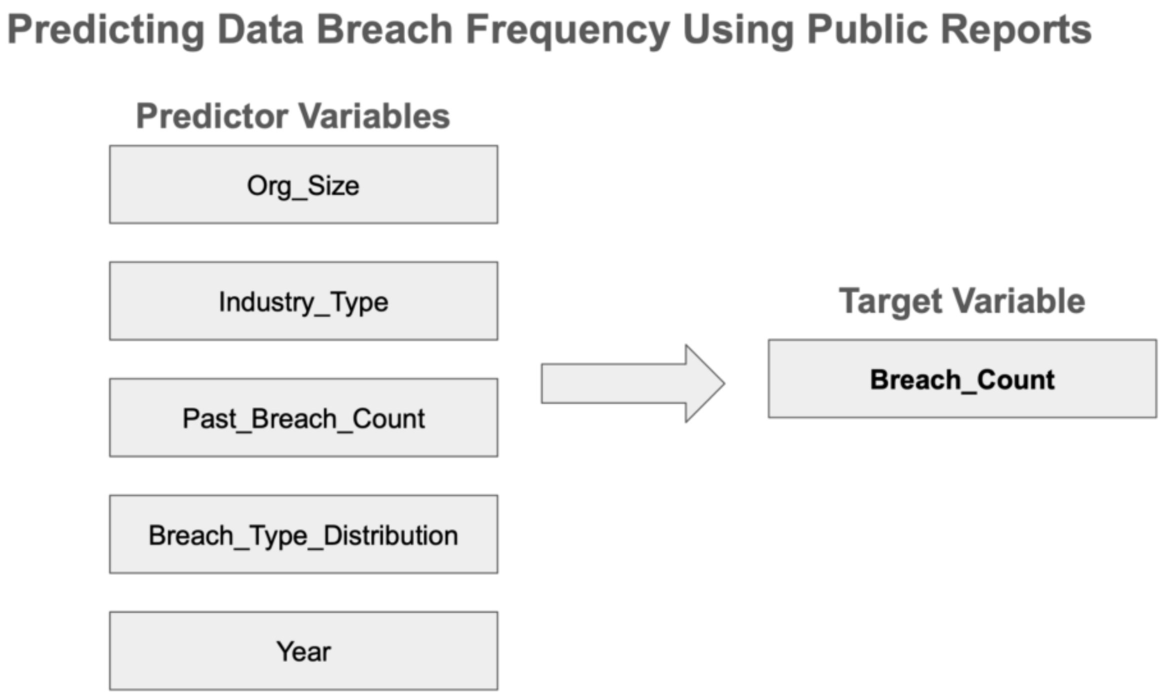
Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Sq	1
Standard Erro	3.5036E-17
Observations	52

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0.98076923	0.32692308	2.6633E+32	0
Residual	48	5.8921E-32	1.2275E-33		
Total	51	0.98076923			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1	3.8588E-17	2.5915E+16	0	1	1	1	1
Org_Size	0	3.3259E-21	0	1	-6.687E-21	6.6871E-21	-6.687E-21	6.6871E-21
Year	1.7789E-21	1.7967E-20	0.09901406	0.92153946	-3.435E-20	3.7903E-20	-3.435E-20	3.7903E-20
Past_Breach_(	1	3.5833E-17	2.7907E+16	0	1	1	1	1

In addition to the regression results, a visual diagram of the model's structure was included to illustrate how the predictor variables relate to the outcome variable. This diagram also highlights the unit of analysis, model type, and data sources referenced.



These findings reinforce the idea that breach history is one of the strongest indicators of future risk. From a policy perspective, this could inform decisions around resource allocation, compliance monitoring, and support for previously breached organizations. However, relying solely on this variable may obscure other contextual or procedural factors that contribute to breach likelihood.

There are several limitations to the study. The PRC dataset, while valuable, does not represent all breach incidents and suffers from inconsistent reporting. The R-Square of 1.00 indicates a risk of overfitting, and the small number of variables may oversimplify a complex phenomenon. Ethically, predicting breaches based on past history raises concerns around fairness and potential reputational harm, especially if such findings are misused.

In future work, expanding the dataset to include more detailed organizational features, introducing industry classification through dummy variables, and modeling breach severity rather than count alone would provide a more nuanced understanding. External validation using other datasets, such as those from Crunchbase or industry-specific sources, would also help strengthen the model.

Overall, this project illustrates how structured data analysis can begin to uncover meaningful patterns in cybersecurity incidents, even when working with public and imperfect data. It also highlights the importance of transparency, caution, and ethics in applying predictive models to sensitive subjects like organizational breaches.

Work Cited

Crunchbase. (2023). *Crunchbase*. Crunchbase. https://www.crunchbase.com/

IBM. (2024). *Cost of a data breach report 2024*. IBM. https://www.ibm.com/reports/data-breach

Privacy Rights Clearinghouse. (2020). *Data breaches | privacy rights clearinghouse*. Privacyrights.org. https://privacyrights.org/data-breaches

Verizon. (2025). *2025 data breach investigations report*. Verizon Business. https://www.verizon.com/business/resources/reports/dbir/