Milestone 2 – Data Project

Cybersecurity Analytics

November 8, 2024

Deliverable Requirements:

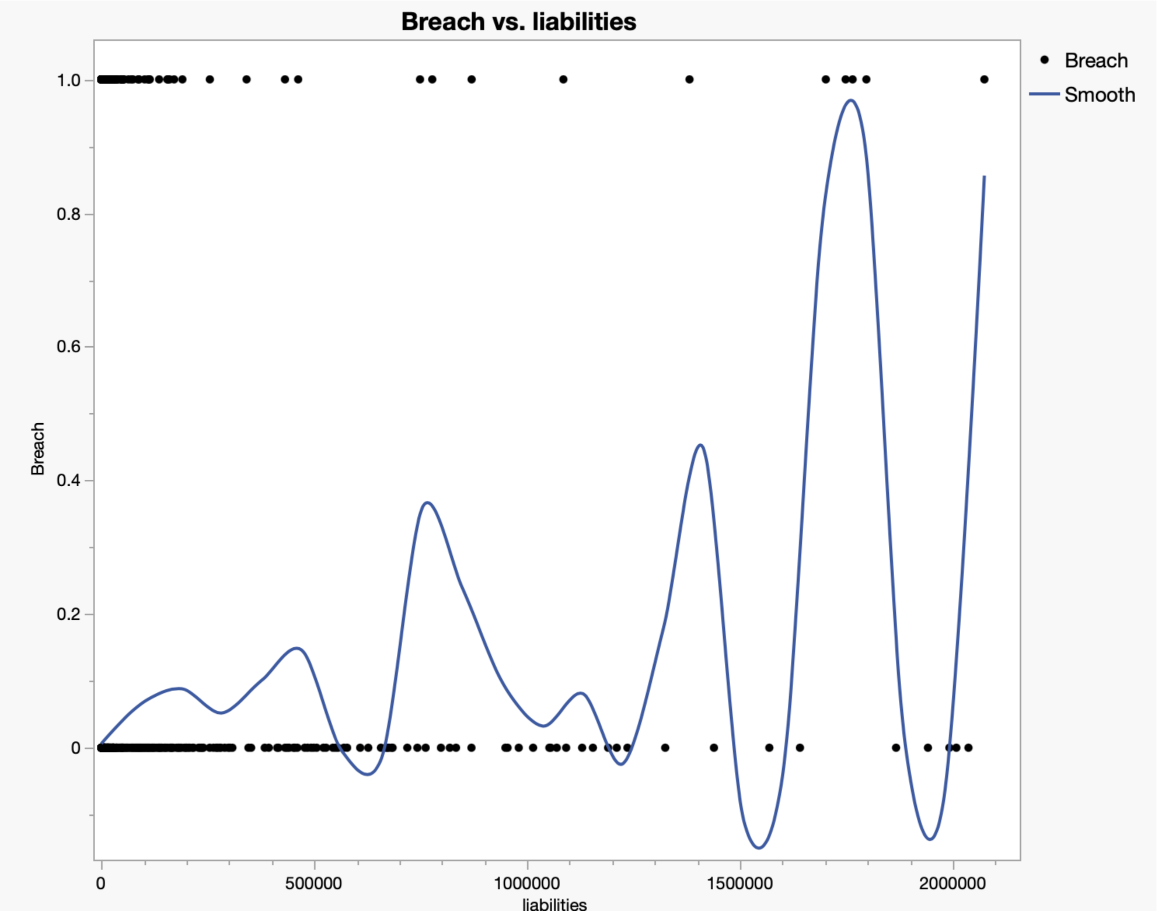
A set of descriptive analysis that should include at a minimum: correlation matrices between the feature set and the key outcome and graphical exploration of key relationships (e.g. histograms, scatter plots, change in variables over time). An appropriate submission will include 15-20 analyses. However, in order to receive the full points here, the group must identify useful relationships of interest and provide credible rationales for them. In other words, I'm not looking for quantity of analysis but for quality of analysis.

Variables of importance:

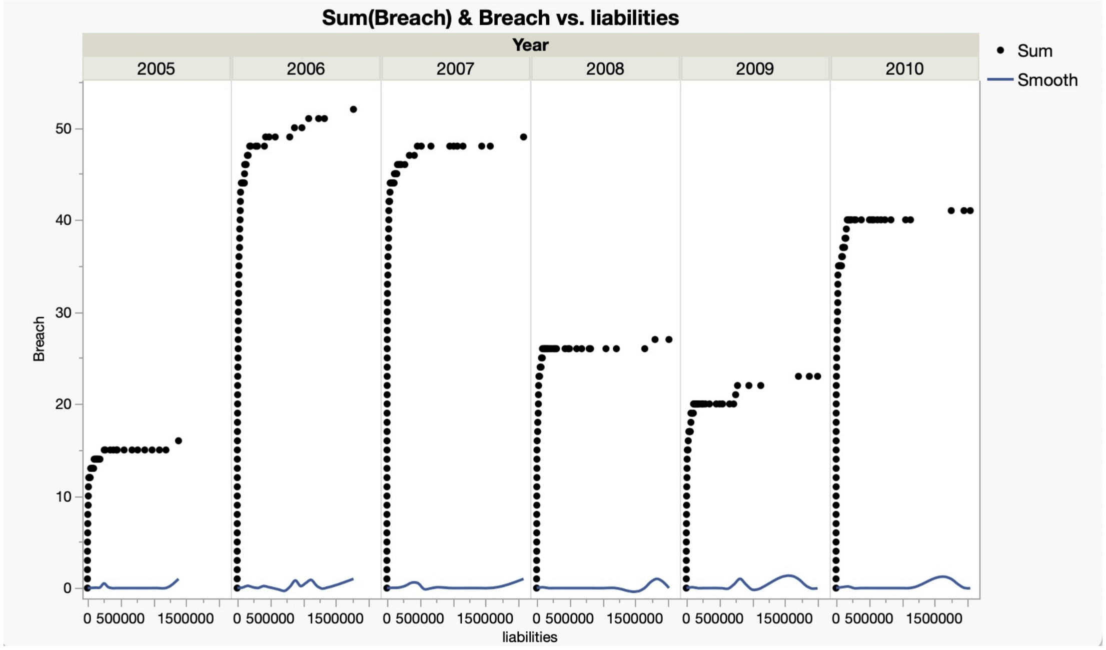
1. Employees
2. Capexp
3. Breach
4. Stockholderequity
5. Retainedearnings
6. Netincome
7. Cash
8. Liabilities
9. CashToLiabilities
10. IsCommunicationServices
11. Year
12. Isfinancialservices
13. Totalassests
14. Netincometoliabilities
15. ROE
16. Newshits

Analysis:

Liabilities vs Breach

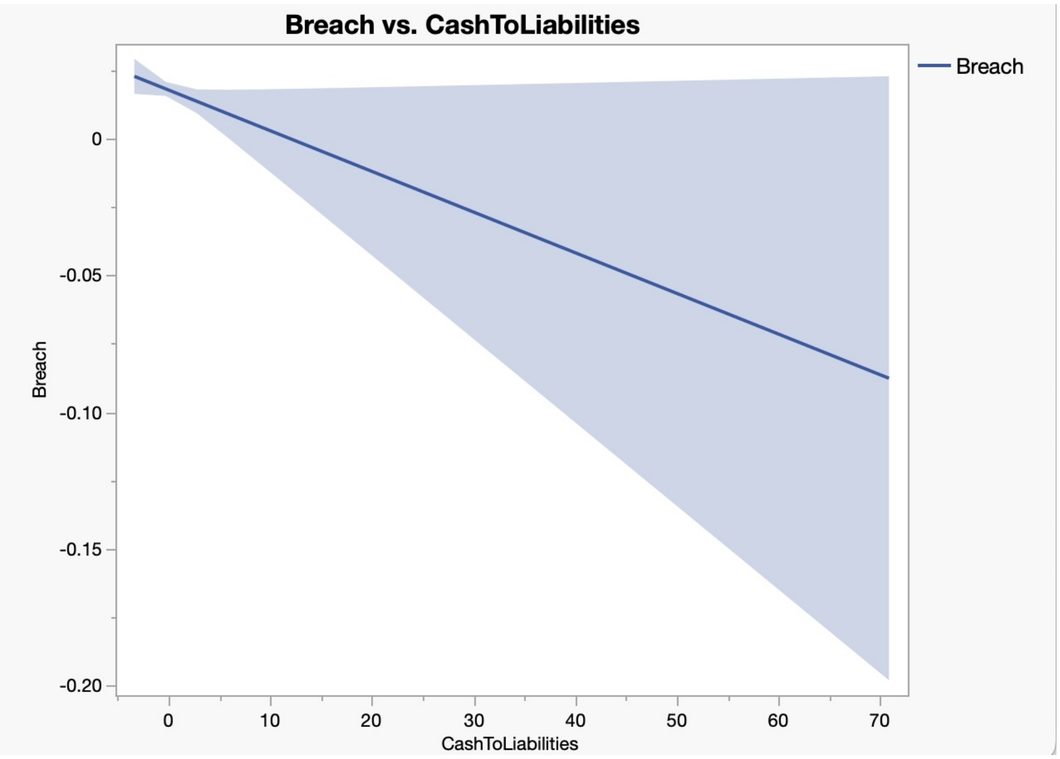


The two variables we are focused on in this graph are breach and liabilities. The variable "liabilities" is the accounting measure of the total obligations that a company owes. The variable breach is a binary variable that describes whether a company has been breached or not, 1 = breached and 0 = not breached. Our correlation matrix shows that there is a positive correlation between the two variables with a value of 0.1272. From the graph above you cannot really tell too much so we decided to separate the number of breaches by year for the liability amounts and got a graph that is much more descriptive.



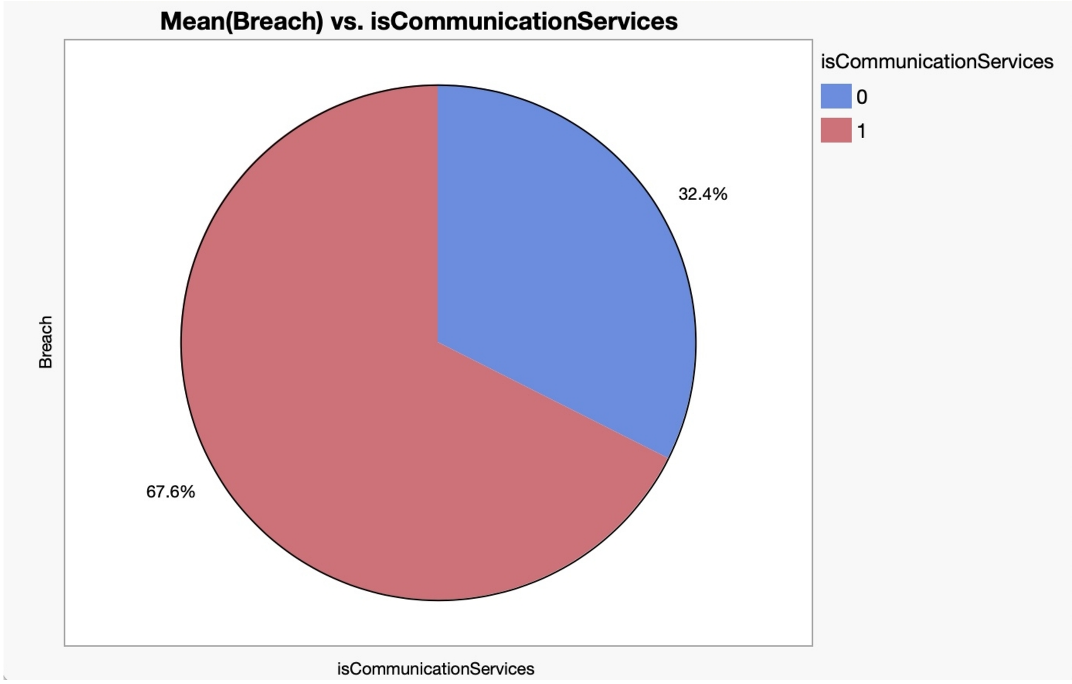
This graph shows us a lot more about the correlation between the two variables. You can see that the number of breaches that occur within a year increases as the amount of liabilities increases. In 2006 the number of breaches increases to a maximum of 53 for the company with the highest liability. In 2005, 2008, 2009, and 2010 you can see that the logarithmic curve for the number of breaches levels our much more than those in 2006 and 2007 and this could be because of the invention of cloud computing in 2006.

CashToLiabilities v Breach



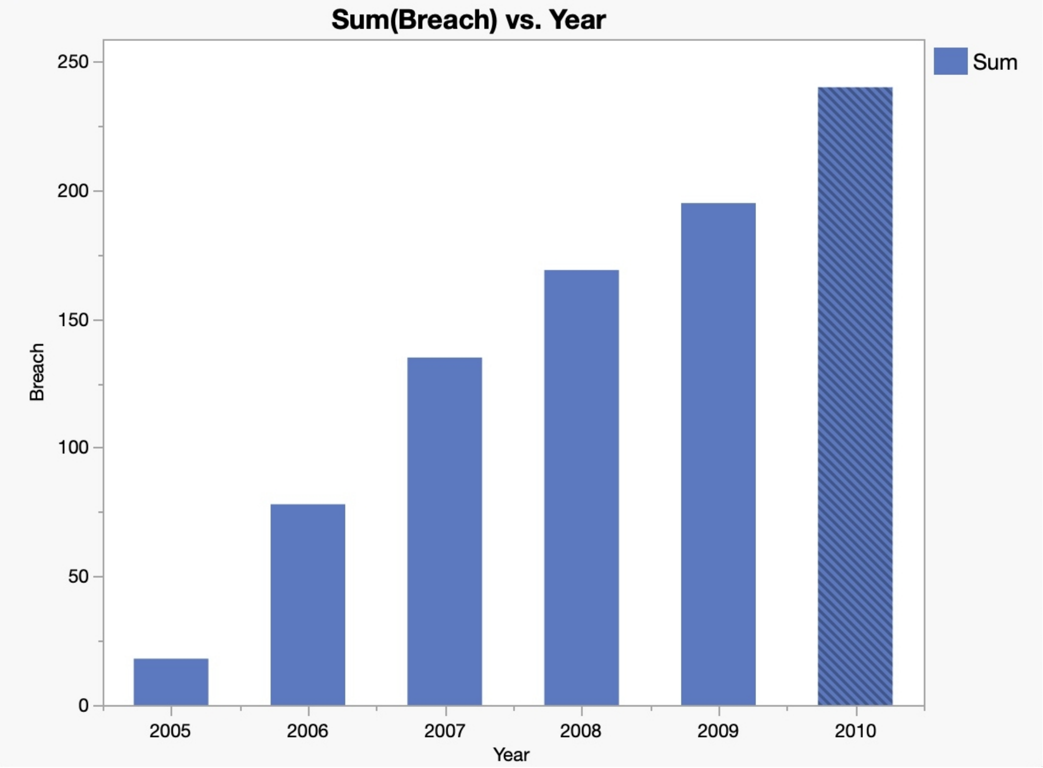
The two variables we are focused on in this graph are breach and CashToLiabilities. The variable "CashToLiabilities" is a cash to liabilities ratio that calculates the amount of liabilities for a company. From our correlation matrix, we can see that breach and cashtoliabilities have a negative correlation of –0.0170. This is significant because that means the lower a company's liabilities ratio is, the less of a chance that the company gets breached. This could be because a threat actor could assume that the company has good security or maybe isn't as big as another company and not work attacking.

IsCommunicationServices v Breach



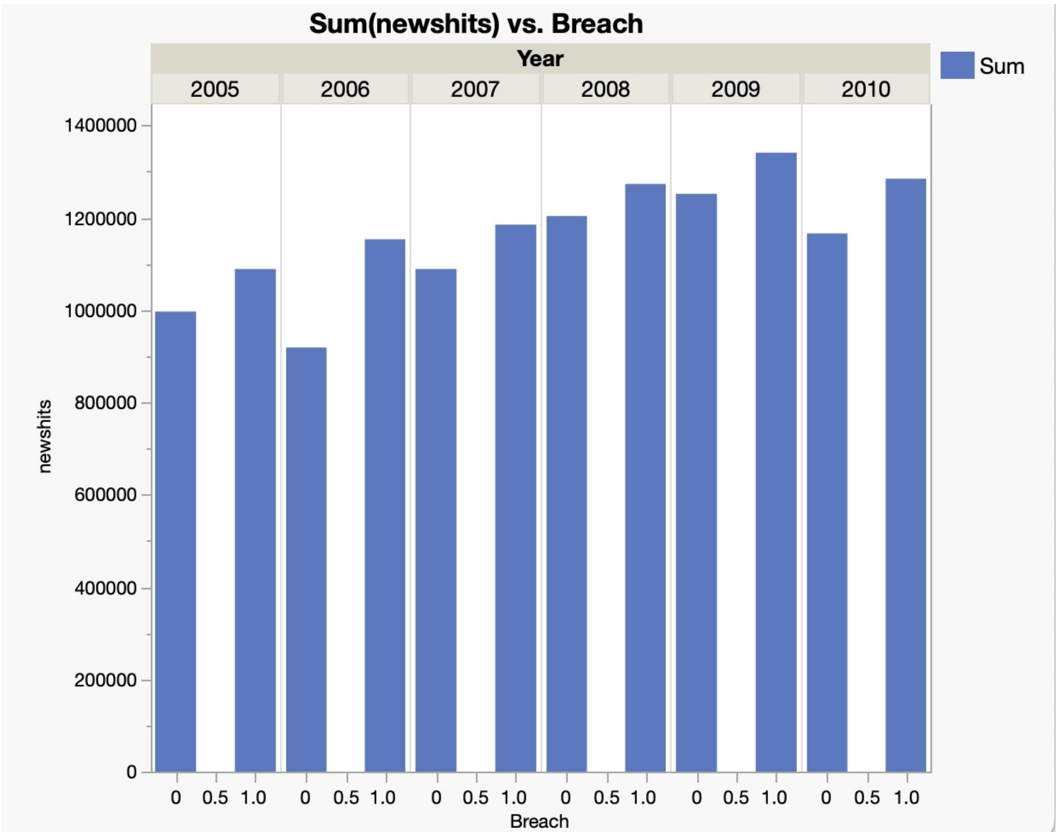
The two variables we are focused on in this graph are breach and isCommunicationServices. The variable "isCommunicationServices" is a binary variable in which it tells us whether the company is a communication service or not, 1 = is a communication service and 0 = is not a communication service. From the correlation matrix, we got a positive correlation value of 0.0230. So, to further analyze this correlation, we decided to make a pie chart of the percentage of companies that have been breached and we can see that 67.6% of the companies that have been breached are communication service companies and the remaining 32.4% are other companies in different sectors. This is a high percentage of companies that are breached that are communication services and this could be because communication services like Verizon and T-Mobile are very susceptible to data breaches because they contain so much data on their costumers and provide to millions of people worldwide.

Year v Breach



The two variables we are focused on in this graph are breach and year. The variable “year” is the variable quantifying what year the company in question has been possibly breached. From the correlation matrix we got a positive correlation value of 0.0010. When initially seeing this value, it does not look like there is a high correlation between the year and whether a company has been breached or not. So, we decided to make a histogram for the years in question and the number of breaches that occurred in that year. You can see from the graph above as the years go on the number of breaches that occur are much higher and the data in the histogram is right skewed. This is possible because technology is becoming much more popular as the years go on and much more people’s and companies' data are moving online giving threat actors more opportunity to attack a company online and breach their data.

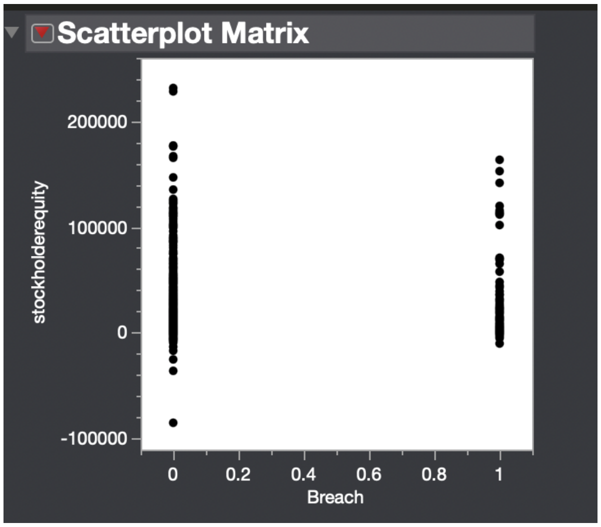
NewsHits v Breach



The two variables we are focused on in this graph are breach and NewsHits. The variable "NewsHits" is the number of times a company is mentioned in the news. From the correlation matrix we have a positive correlation value of 0.1334. Again, this does not tell us too much of anything, so we decided to make a histogram to further analyze the data. After making the histogram we decided that splitting up the data by year would help us further see what the correlation looked like. From the graph above you can see that on average companies with a larger amount of news hits are more likely to have a data breach and this is true for each year between 2005 and 2010. This is probably because companies that are in the news more often are flashed right in the face of threat actors and are sought to be more profitable than others. News hits could also mean a problem within a company that would make it much easier for a threat actor to attack and breach their data.

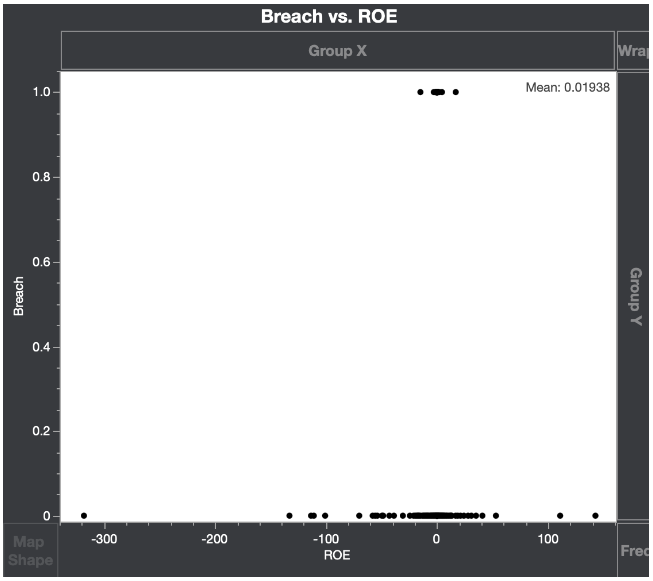
Stockholder Equity vs Breach

Based on the Correlation Matrix, we found a moderate, positive relationship of 0.1435 between the stockholder equity and breach variables. Using a Scatterplot Matrix, we also see their linear relationship between the variables. This implies that there's a minor tendency for breach occurrences to rise in parallel with stockholder equity.



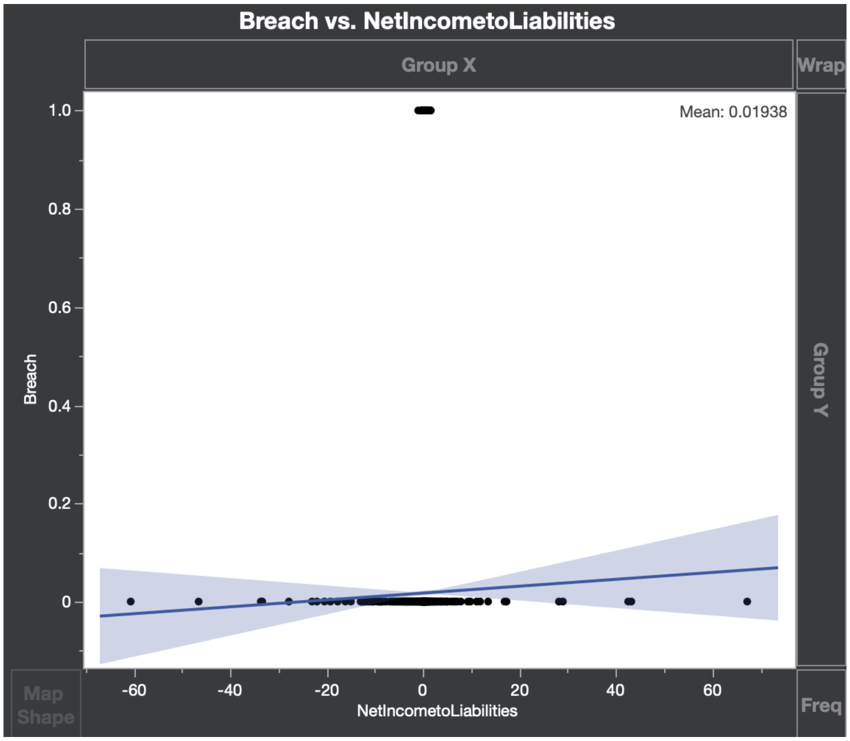
ROE vs Data Breach

Based on the Correlation Matrix, the Total Return on Assets and Data Breach variables have an extremely weak, positive relationship of 0.0050. Using a Scatterplot Matrix, we also see the variables possess a weak linear relationship. This suggests that the ROE has little to no effect on the data breach predictability.



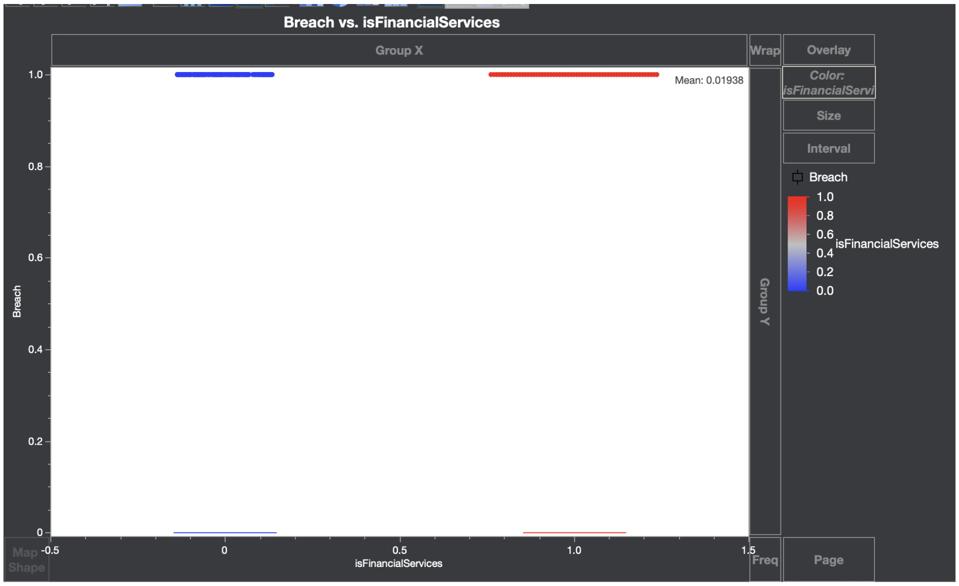
NetIncomeToLiabilities vs Breach

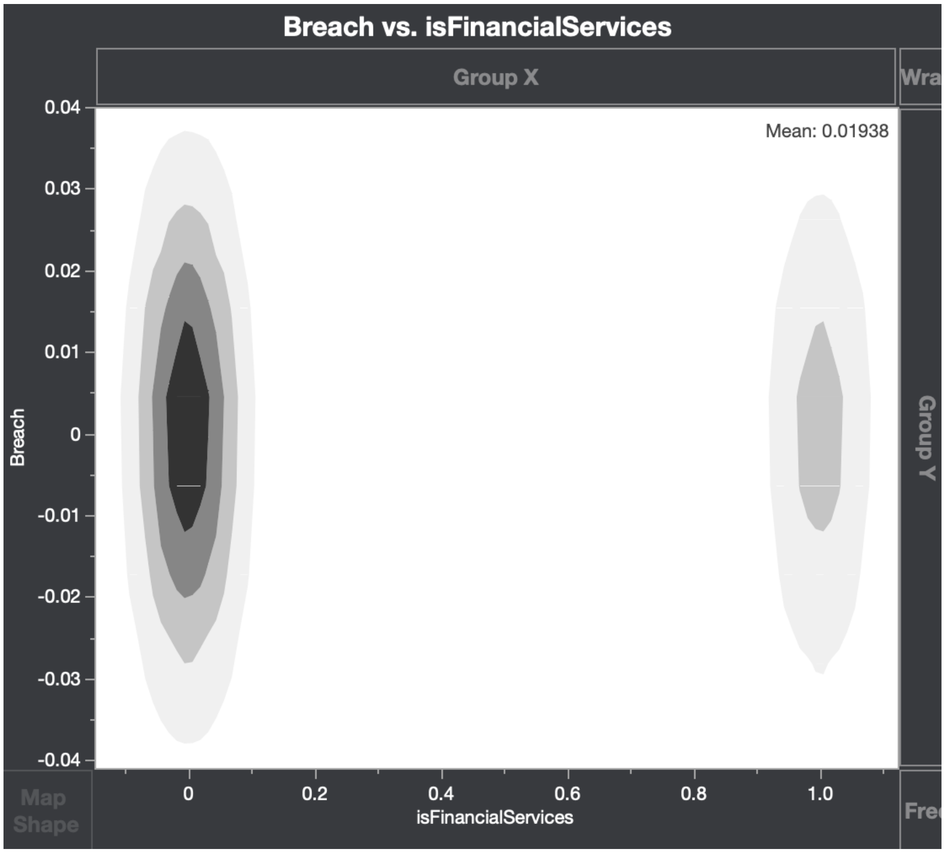
Based on the Correlation Matrix, the NetIncometoLiabilities and breach variables have a -relationship of 0.0086. This very low correlation indicates that Net Income to Liabilities and breach frequency are not correlated, which we can also see in the scatter points.



IsFinancialService vs Breach

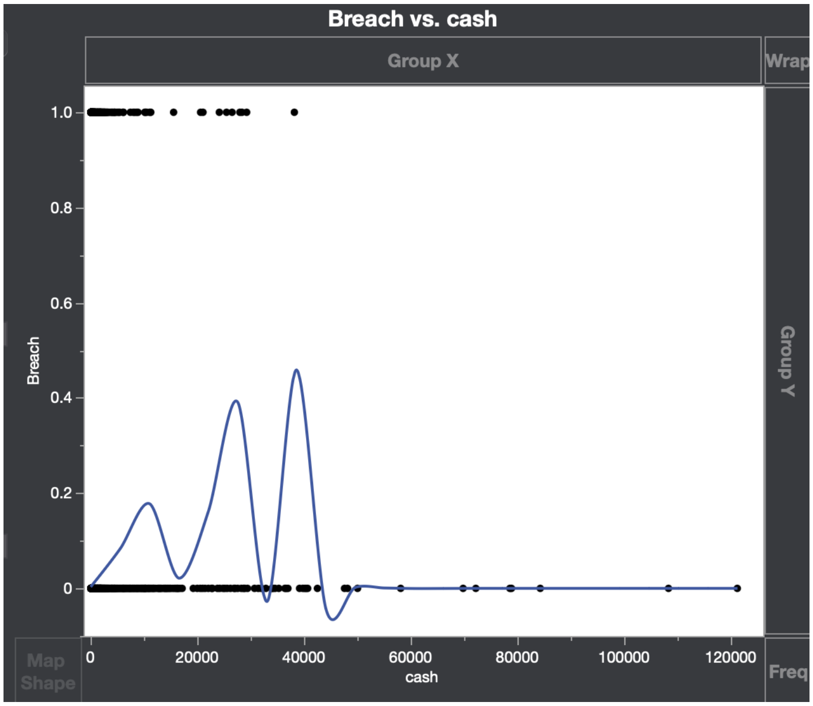
Based on the Correlation Matrix, the breach probability's relationship to whether a company is in the financial service sector has a weak, positive relationship of 0.0061, indicating a little to no effect.

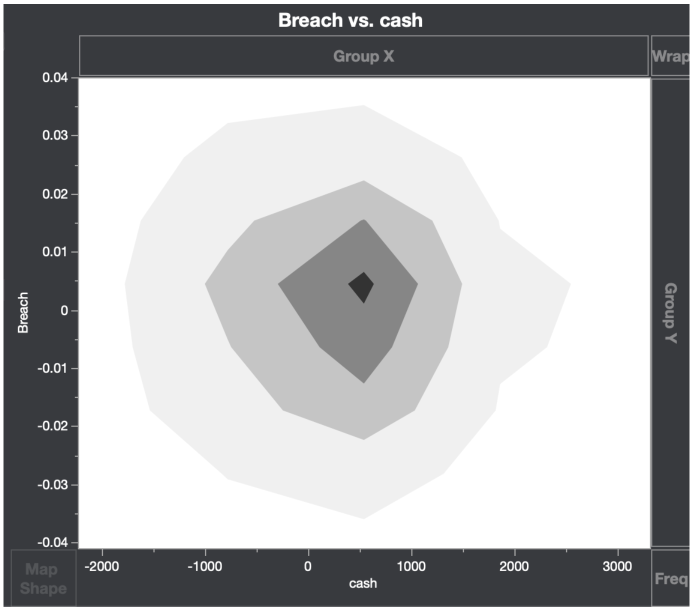




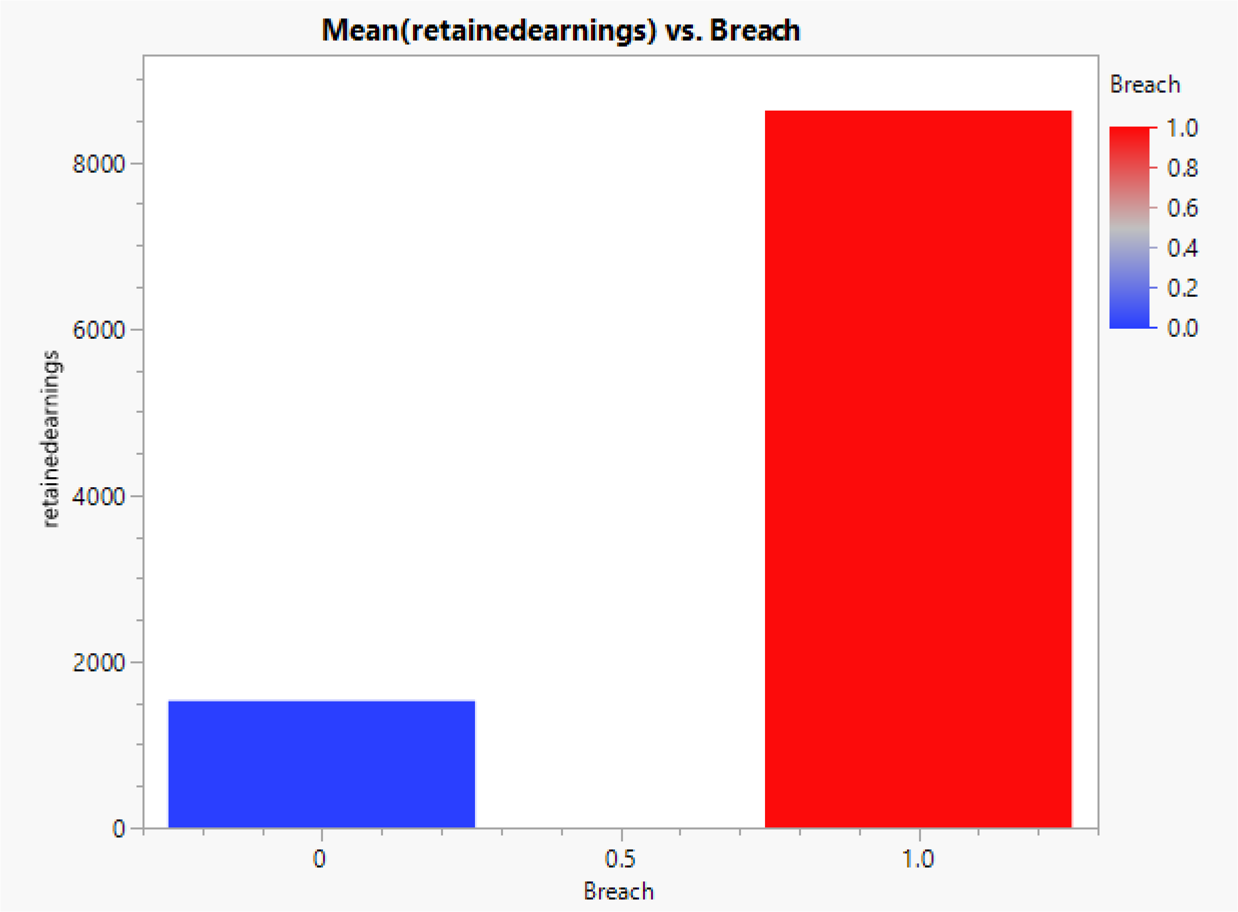
Cash

Based on the Correlation Matrix, the Cash and Data Breach variables have an low positive relationship of 0.0945, suggesting that there is little correlation between an firm's monetary values and data breach incidents. Any correlation between cash and breaches is too weak to be significant or predictive in this situation, as also seen in the density and scatter plots.



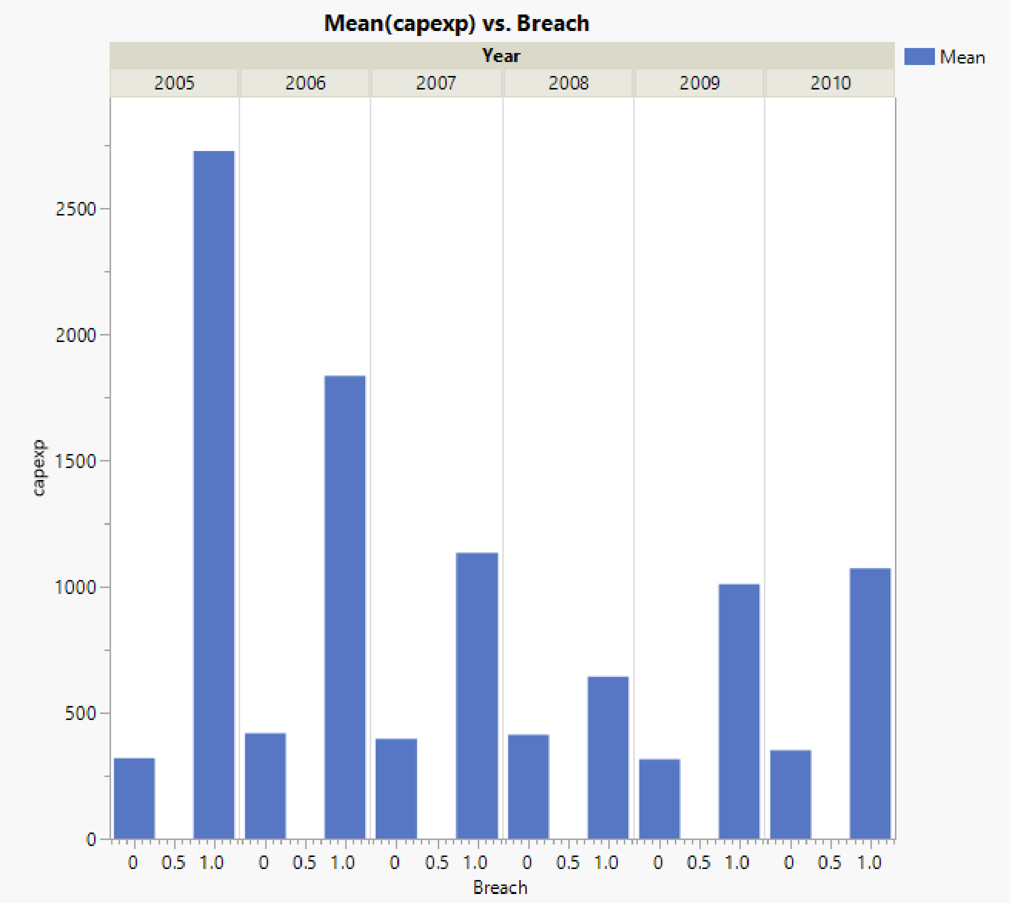


Retained Earnings vs. Breach:



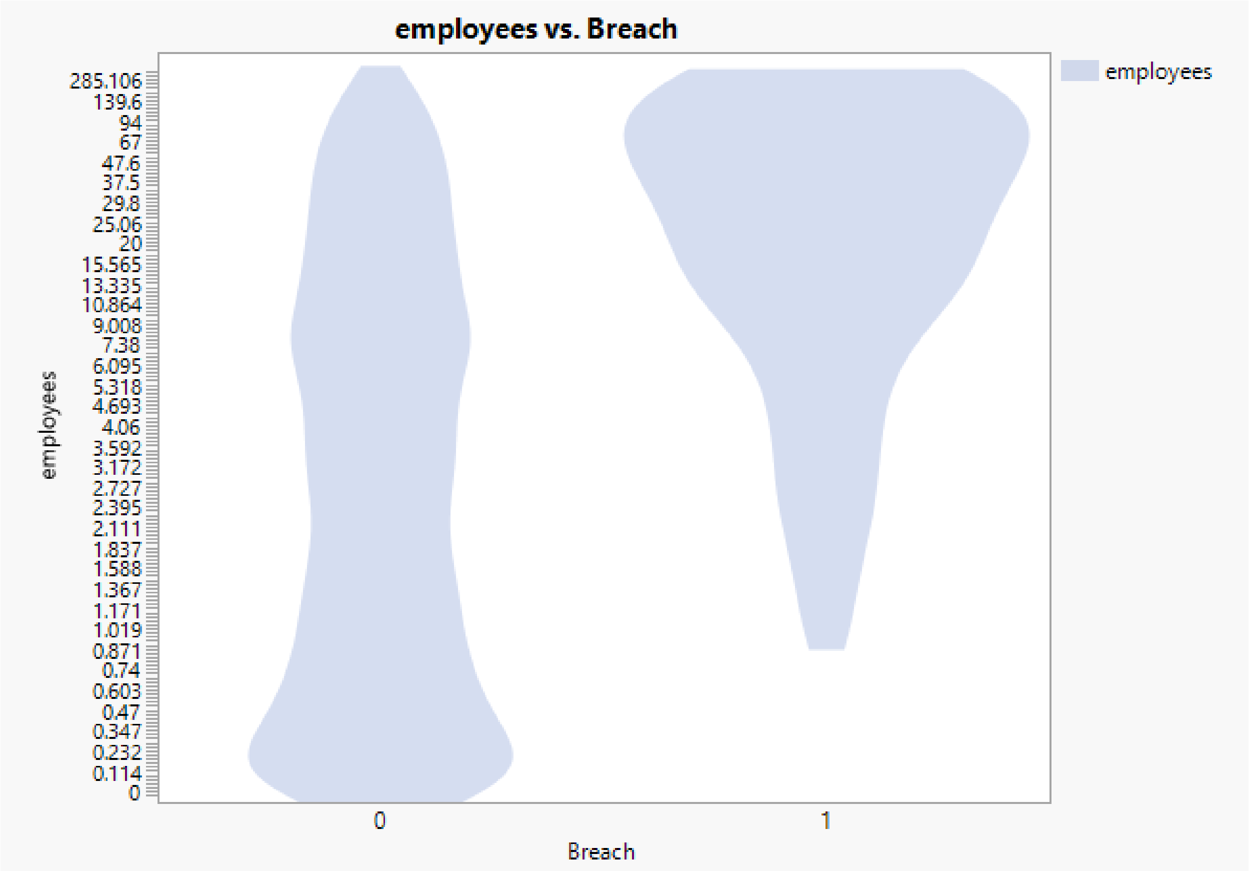
The correlation matrix gives breach vs. retained earnings a positive correlation of + 0.0982, a rating that cannot be passed as directly correlated. Even though there is a slight relationship there is not a significant relation between them to make an informed decision. In the bar chart you can see that the mean retained earnings is higher for breached companies than it is for non-breached companies. This means that there are many more breached companies with higher retained earnings than those with lower.

CapeXP vs. Breach:



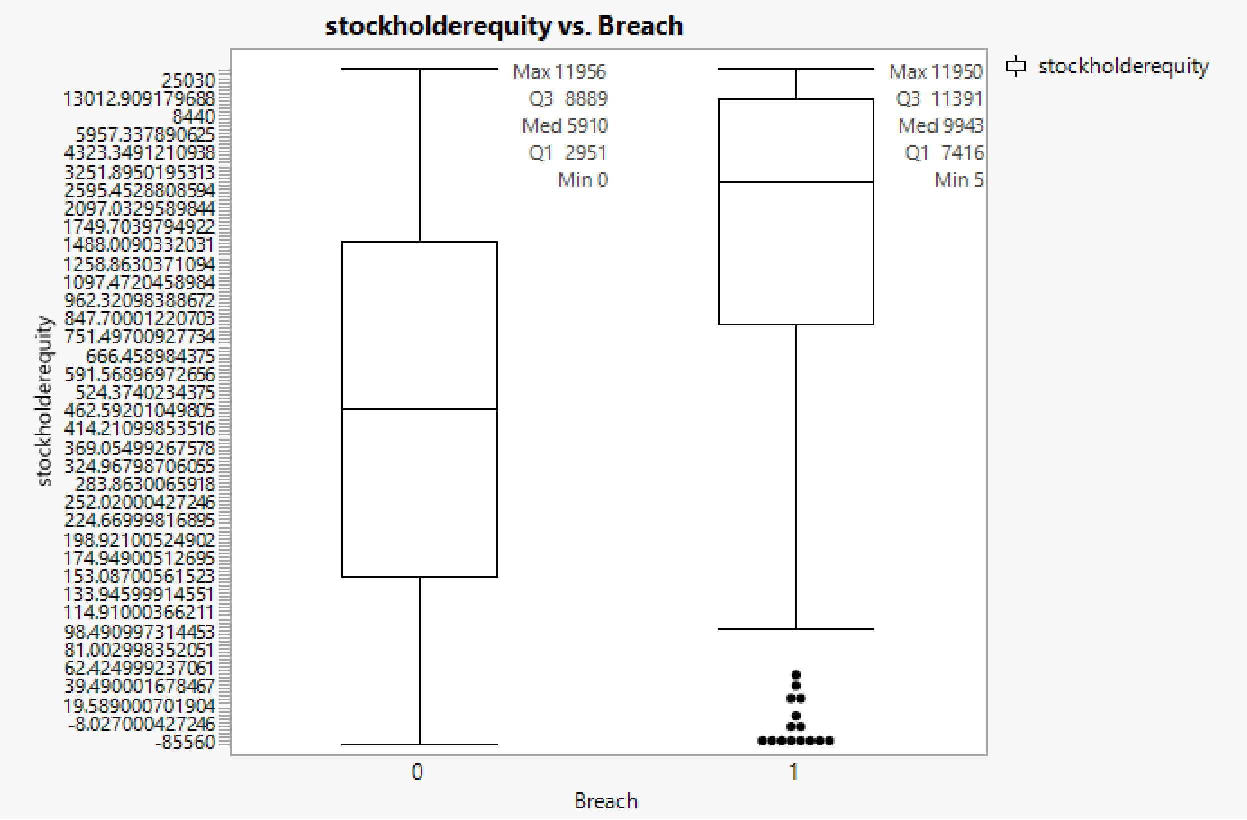
The variables compared here are Breach and Captial Expenditure. Breach is displayed as a binary variable 0 or 1; not breached and breached. We compare this data against the mean capital expenditure over the years 2005-2010. When analyzing the data on the matrix we see a positive correlation between capital expenditure and breaches of + 0.0730. Analyzing the graph, we see the trend of capital expenditure and fate of being breached. Although the graph can display that lower mean capital expenditure is somewhat associated with less security breaches. There is ultimately not enough evidence to say these two variables are directly correlated.

Employees vs. Breach:



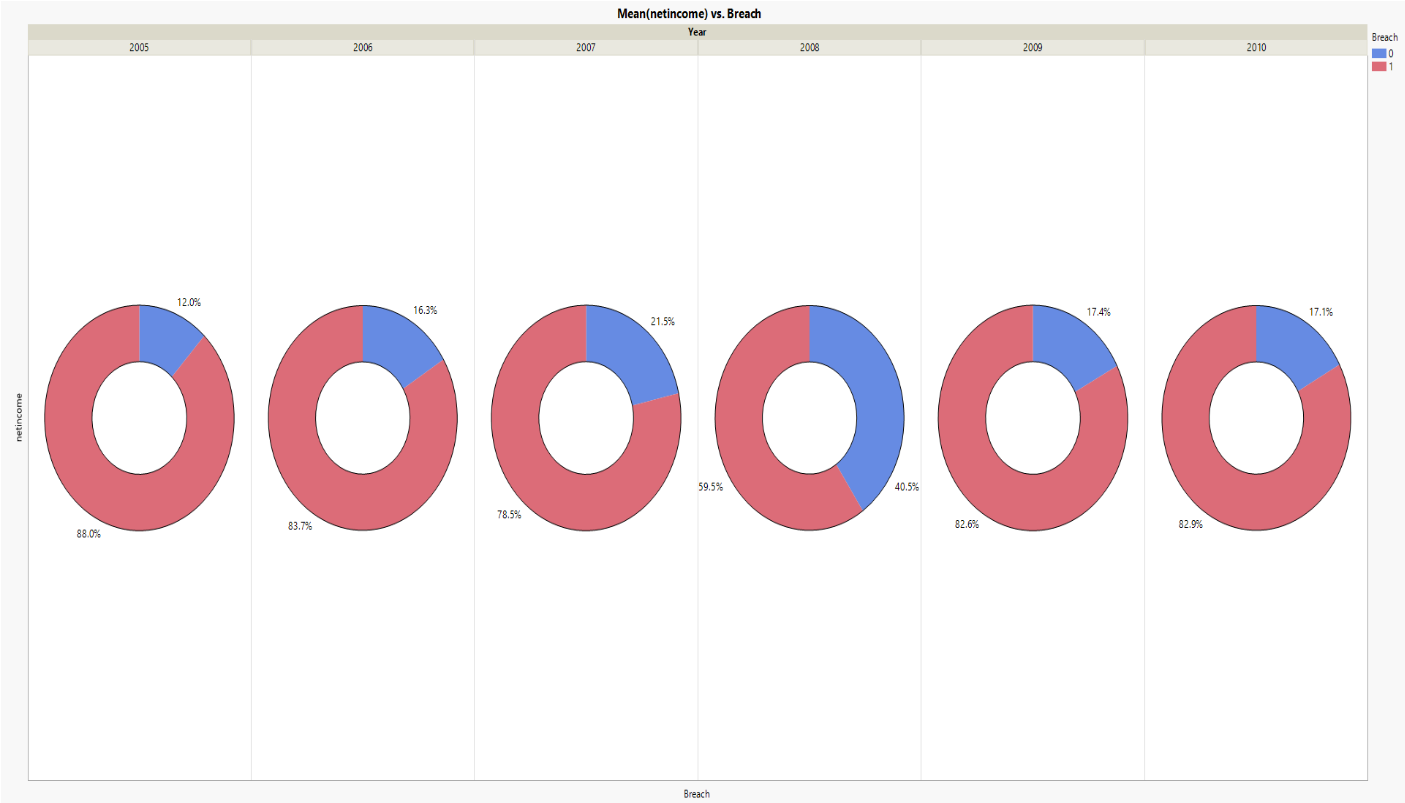
Here we're comparing the number of employees and breaches. Using a heat map, we can see when breaches do and do not occur, and the number of employees present at that time. The thicker the more congested that area is with data points and vice versa. According to our matrix there is a positive correlation of + 0.1056 which means there is a medium relationship between these two variables. No breaches can happen to any company, but looking at the cases where there are breaches can show us the pattern we may be looking for. In this case you can see there are more breaches when there are more employees than when there are fewer.

Stockholder Equity vs. Breach:



Here we are comparing breach and stockholder equity which we made into nominal values. The correlation matrix gave these two a + 0.1435 positive correlations, making it stronger than most other variables in our group; meaning it is technically more correlated than most. The box plot also visually shows us the quartiles in the values of stockholder equity. The box plot also gives us the Min and Max, first and third quartiles, and median for breached and not breached companies. The median stockholder equity for breached companies is much higher than that of non-breached.

Net Income vs. Breach:



In this graph we're looking at multiple pie charts over the years 2005-2010. We compare the mean total net income vs. Breaches. Each pie chart is the total mean net income for that year. The correlation matrix says there is a positive correlation of + 0.0760 which doesn't make it the strongest correlation in our data set nor the weakest. What we see here is that there have been a great number of breaches on companies that make up the total mean net income for that year.