Cyber Safe- A look into global cybersecurity

Data Understanding & Cleaning

My analysis of the cybersecurity data sheet began by examining the structure of the data set. I did this by viewing the first few rows of the column names, looking at the data types and confirming the absence of null values.The dataset consisted of 3000 Cyber attack incidents across multiple different countries and industries. The key variables I analyzed were Attack Type, Target Industry, Financial Loss, Number of effective usages, and Incident resolution time. After I was able to understand the data I started to standardize all of the categorical texts by getting rid of the white space from some of the columns. I checked for duplicates but was not able to find any, but I still applied a duplicate removal step. Since the year was categorical, I converted it from an integer to a categorical data type so the grouping was able to be clean. After that was completed I checked the numeric columns to make sure that the values were valid and I specifically used box plots to see if there were any outliers in the financial area in affected users or in resolution time. Next I created a loss per user column so I could figure out around an area of how much per capita these attacks were affecting people financially.

Exploratory Data Analysis

In this section I started by generating summary statistics for affected users. Using key variables like Attack Type and Target industry I was able to generate the amounts of attacks and the percent of each kind of attack that occurred in a country. I was able to find that out of all the attacks, the split of type was about 17% per each attack. In addition, the percentage of attack per country was around 10%.

Data Visualizations

My initial visualization, the heat map, demonstrated that there was not much correlation between the Financial Loss in Millions, Number of affected Users, Incident Resolution Time, and Loss per user. This finding was very interesting because without a correlation it could show that we need more data to get more accurate results. When looking at the data from the “Top 10 Countries by financial Loss”, I was pleased to see a slight trend. The country with the highest financial loss is the UK followed by Germany and Brazil. It would make sense that these countries have more cybersecurity attacks than countries like India where access to technology is not as widespread.

It is to also note that in my third visualization “Year-Wise Attack Type Frequency, it was an almost even amount of attacks per year, even when compared to other countries. My trend of not getting much variation continued into my next visualization of the “Distribution of financial loss”. Although there was a variety in the amount of financial loss, it intrigued me to see that the frequency of the attacks was very similar throughout the different amounts of financial loss.

Through looking at my Resolution Time by attack type barplot, I was able to find out that all attack types take roughly the same 30 to 45 hours of time to resolve on average. There was no correlation found between the three variables – Financial Loss in Millions, Number of Affected Users and Incident resolution time. In addition the map I created in GeoMap, emphasised my point of seeing more cybersecurity issues in wealthier countries.

Insights and Generalizations

After examining the data using my visualizations, I was able to definitively say a few things. First being that the countries that suffered the most from these attacks are the UK, Germany, Brazil and the US. Additionally, these cyberattackers specifically targeted IT, banking, healthcare and retail organizations. It was interesting to see how much of the data was evenly distributed. The types of attacks that affected the users and the defense mechanisms against these attacks both had very even distribution. Potentially this could mean that we need to increase cybersecurity technology since it is not currently significantly decreasing our attacks. Although the data may have been limited due to lack of true global coverage and reporting bias, I think it serves as a start for a larger conversation around cybersecurity and how we can remain safe online.

Optional Questions

I chose to answer the following questions and will refer to them in my paragraph below. “Which countries experience the most diverse types of cyberattacks? Is resolution time decreasing over the years? Do certain industries recover faster than others?”

When looking at the types of attacks and the frequencies, it remained puzzling to me as to why there were no trends in the data. After creating a bar chart to understand the diversity of attacks, I found that it was completely equal amongst the countries. In continuation, my graph for the resolution time over the years showed no specific trends and my boxplots for industry recovery time all had a mean around 35. It was very interesting to see how similar each attack frequency and recovery time was especially in different countries.