**Outcome of EDA:**

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

**Libraries Used:** I used important libraries/tools such as pandas, numpy, scipy, matplotlib, Seaborn, statsmodels, [Scikit-learn](https://www.educative.io/blog/scikit-learn-tutorial-linear-regression) to name a few. I read four datasets and then also combined two of the four datasets and performed EDA on the fifth combined dataset as well.

**Commands Used to understand data and variables within the Datasets:** I used various data analytics commands such as on each of the five datasets and its selected variables that were of interest to my analysis

* **dtypes:** check the data type of all columns
* **info:** Get information on DataFrame
* **head:** returns top n rows of a DataFrame or Series
* **tail:** shows you data from the end of a file.
* **min:** Find min values of a DataFrame
* **max:** Find max values of a DataFrame
* **groupby:** grouping the data according to the categories and apply a function to the categories.
* **sort\_values:** sorts the DataFrame by the specified label
* **isnull:** Detect missing values
* **describe:** Describe a summary of data statistics
* **concat:** Combine DataFrames across columns or rows: concatenation
* and more

Since most of my variables were non-numeric and most commands required them to be numeric, I learned and used the following application e.g. on my Incident Type variable and mapped it to its numeric counterpart. I used the scikit-learn library and in particular the LabelEncoder tool. I created the numeric representation of Type e.g. Espionage is now equal to 5.0, Data destruction Type is mapped to 1.0 etc.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

The commands above helped me understand the datasets, the number of records, number and type of variables, nulls or blank data, which variable values had the most occurrences, etc.

**Histograms and scatterplots:** I created histograms and scatterplots of the datasets I was analyzing to see visually the general pattern of the data and the variables being analyze. The histograms of each of the datasets showed that security vulnerabilities and Security Incidents reported increase as years increase. I also noticed that of all the security incidents, Espionage incident type is on the increase in recent years. So far, both these observations support my hypothesis that security incidents and vulnerabilities increase as years go by and that the espionage (state sponsored security incidents) are on the rise. I created scatterplots on a couple of variables of my interest (Year and incidents for the Combined dataset of security incidents (Dataset 2 and Dataset 4) and count of vulnerabilities by year for the vulnerability (dataset 1) and both show a linear positive correlation between the variables. When the points in the graph are rising (as in my case in both the datasets I used the scatterplots), moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another.

Combined Dataset’s Scatterplot of variable Year and Incident Type: Chart

Description automatically generated

Dataset#1 of Vulnerabilities by Year Scatterplot of variable Year and frequency of vulnerabilities:

Chart, scatter chart

Description automatically generated

**Correlation (Pearson) and Covariance:**

I calculated the correlation and covariance on two distinct datasets:

1. The Combined dataset of two datasets (Dataset2 – Cyber Security Incidents and Dataset 4 – Security Incidents) – aka “Combined Dataset’.
2. Dataset 1 that has vulnerabilities by year.

*Correlation and Covariance of Combined Dataset:* The Pearson correlation coefficient, often referred to as Pearson’s r, is a measure of linear correlation between two variables. This means that the Pearson correlation coefficient measures a normalized measurement of covariance (i.e., a value between -1 and 1 that shows how much variables vary together). The correlation between the two variables in the combined dataset (Year, and count of Incidents) is 0.91 showing a strong positive correlation.

Covariance is a measure of the relationship between two random variables. The covariance of the two variable in the combined dataset is positive which means these variables are moving in the same direction.

*Correlation and Covariance of Dataset1 (Vulnerabilities by year):* The correlation between the two variables in the vulnerability dataset (Year, and count of vulnerabilities) is 0.77 showing a positive correlation.The covariance of the two variable in the combined dataset is positive which means these variables are moving in the same direction.

We also performed the regression analysis and Hypothesis testing (please see the PowerPoint presentation for details) that supports our hypothesis that incidents and vulnerabilities increase with increase in year.

We also identified a Pareto distribution chart that shows that espionage for both datasets is the most prominent incident category that nations need to pay special attention to. See the Pareto chart for the combined dataset provided below:

*Chart, line chart

Description automatically generated*

**What do you feel was missed during the analysis?**

I feel I performed various exploratory data analysis (EDA)steps to understand the data. I am sure as I get more experience using EDA, I can expand my testing further and improve my data analysis skills.

**Were there any variables you felt could have helped in the analysis?**

Yes, I believe if we had more variables, I could have explored other aspects on evaluating the impact and relationship of vulnerabilities on Security incidents. I did not find variables or datasets that lists what vulnerabilities caused a certain security incident. I could have identified which vulnerability is causing the security incidents, how much damage/expense it is costing organizations, and isolate them so that they can be focused on remediation.

**Were there any assumptions made you felt were incorrect?**

Since we were asked in the assignment to combine two datasets, I was able to accomplish that but with an assumption that there were no duplicates when I combined the two datasets. However, I analyzed the single dataset and the combined dataset and the trending appears to be similar so I do not think that effected my analysis in a negative way.

**What challenges did you face, what did you not fully understand?**

I had challenge with understanding the concepts of PMF, CDF, etc. Those were hard for me to fully grasp including the calculation of them. The textbook referred to the ThinkStat package which referred in turn to various other functions and calculations so it was hard to figure out using the textbook version how to calculate them. I had to research the web to understand and calculate them.