Assignment-1 Report

AI-620: Fundamentals of Data Engineering

Name: M Danish Zaheer
Roll no: 25280092

Thematic Focus:
The thematic domain I have choosen is cybersecurity with primary focus on analyzing threat landscape and incident trends in cyberattacks like phishing, ransomware, and SQL injection. In this thematic focus, I am mainly looking at how these threats change over time, how their overall trend patterns behave and how they relate with each other using correlation. Along with this, I also focus on the incident side analysis such as monthly increase in incidents, countrywise distribution of ransomware attack, common claim gangs, and victim domain patterns, I also analyze which industries are more affected and how different attack types impact different industries.

Key Findings:
1. Clearly, the Visualizations notebook time-series graph shows a sharp late-2025 rise in cybersecurity, with phishing and ransomware also bumping upward, while SQL injection remains low and largely flat throughout.
2. Seeing the feature totals visualization clearly shows phishing has the highest overall trend count, followed by cybersecurity, then ransomware, and lastly SQL injection.
3. Cybersecurity and phishing show a moderate positive correlation of about 0.56. From the correlation heatmap, cybersecurity is similarly moderately correlated with both ransomware (0.56) and SQL injection (0.56), while SQL injection has stronger correlations with phishing (0.97) and ransomware (0.82) overall.
4. In the monthly incidents trend visualization, incident levels stay relatively steady through 2023, then drop sharply at the start of 2024 before gradually rising again and spiking toward the end of 2025.
5. From the incidents distribution perspective for ransomware, the US shows the highest number of incidents (1400+), and among the known claim gangs, Lockbit3 appears as the most frequent (around 90 incidents). Also, the .com TLD is the most common in victim domains, followed by .org.
6. “Data Breach Notification” is most commonly used word while ransomware attacks.
7. From 2015 to 2024, the lowest number of incidents is observed in 2019 overall industry wise.
8. From the target industry distribution, the IT industry is the most affected industry from a cyber threats perspective.
9. From the attacktype vs industry heatmap phishing impacts the banking industry the most, highest attack by phishing column.

Note: These are the main findings from the visualizations more insights can be derived if we go in more depth.

Challenges:
During the pipeline and analysis one challenge I have faced is the schema difference across data sources where each dataset has different column names different feature availability and different detail making it difficult to directly mapping to one single source. Secondly, issue was rate limiting from google pytrends which limited how much data could be extracted in a single run, rate limiting is not faced for sites with APIs. I also faced data inconsistencies such as mixed formats in categorical fields and uneven reporting across countries and industries. Missing values were common in some sources, which required dropping high missing columns and removing a small number of incomplete rows to keep the cleaned dataset consistent. At last, different datasets followed different time formats some were recorded by year, while others were in week or month by year or full date format, so date parsing and normalization steps were made at needed place.

Conclusion:
I have tried my best to map the different data sources and map one to another, but due to a clear gap in their schemas and the differences in attack data extracted from each site, there are still many attack types that can be further extracted from there respected sources and can be included in our pipeline. In future, improving the schema alignment and adding more extraction coverage can make the pipeline more complete and consistent for analysis.