Part-2 Answers

AI-620: Fundamentals of Data Engineering

Name: M Danish Zaheer
Roll no: 25280092

Note: For this task I have transformed and cleaned only 3 datasets out of 4 which I have extracted.

(a) -Answer:
For Google Pytrends: Dropped isPartial column because it is basically a metadata flag from Google Trends and not an actual feature needed for analysis, so keeping it does not add much value in our trend comparison.
For Ransomware dataset(Allcyberattacks): Dropped the high missing infostealer related columns (27 columns) because these fields have extremely high missingness. Keeping them makes the dataset very sparse and reduces usability for analysis. Also removed the rows having >= 1 missing value because the missing rows count% is around 0.008% only so removing them gives a clean dataset with no missing values. Parsed and normalized the date column to ensure consistent time based aggregation (monthly/yearly) and to avoid any formatting issues. Replaced False values in claim_gang/claim_url with "unknown" because False here is not a real category, it only represents missing/unknown information, and converting it to "unknown" makes it meaningful for grouping and plotting.
For Kaggle dataset: Although the data is already in a cleaned state, still for the numerical features I have applied standardization on Incident Resolution Time, Financial Loss, and Number of Affected Users because these variables are on very different scales, and standardization makes them comparable for analysis and reduces scale dominance when comparing patterns.

(b) -Answer:
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. Seeing the feature totals visualization clearly shows phishing has the highest overall trend count, followed by cybersecurity, then ransomware, and lastly SQL injection. 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. 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. 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.

(c) -Answer:
Current visualizations do lack some in depth filtering and proper segmentation before visualizing, so a few plots are still very high level. Also, the cleaned dataset distributions are not explored in detail, since I am not doing any predictive analysis and my focus is more on descriptive analysis. My current visualizations are mostly exploratory, so they show trends but don't clearly explain why changes happen. Correlation heatmaps show relationships but they dont show causation. From the dataset useful metrics can also be generated to inform business stakeholders, like industry wise risk ranking top attack type per industry and month by month incident change, so stakeholders can stay on the safe side and take better decisions.