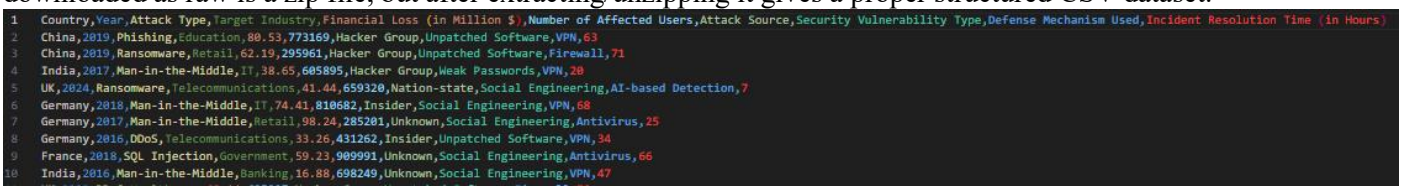
Part-1 Answers

AI-620: Fundamentals of Data Engineering

Name: M Danish Zaheer
Roll no: 25280092

(a) -Answer:

Kaggle dataset is a structured dataset, it is a CSV where every row is an observation and column names are consistent. This is structured because it has a fixed schema and is easy to load into Pandas as a dataframe. Although the file I downloaded as raw is a zip file, but after extracting/unzipping it gives a proper structured CSV dataset.



```
Country,Year,Attack Type,Target Industry,Financial Loss (in Million $),Number of Affected Users,Attack Source,Security Vulnerability Type,Defense Mechanism Used,Incident Resolution Time (in Hours)
China,2019,Phishing,Education,80.53,773169,Hacker Group,Unpatched Software,VPN,63
China,2019,Ransomware,Retail,62.19,295961,Hacker Group,Unpatched Software,Firewall,71
India,2017,Man-in-the-Middle,IT,38.65,605895,Hacker Group,Weak Passwords,VPN,20
UK,2024,Ransomware,Telecommunications,41.44,659320,Nation-state,Social Engineering,AI-based Detection,7
Germany,2018,Man-in-the-Middle,IT,74.41,810682,Insider,Social Engineering,VPN,68
Germany,2017,Man-in-the-Middle,Retail,98.24,285201,Unknown,Social Engineering,Antivirus,25
Germany,2016,DDoS,Telecommunications,33.26,431262,Insider,Unpatched Software,VPN,34
France,2018,SQL Injection,Government,59.23,909991,Unknown,Social Engineering,Antivirus,66
India,2016,Man-in-the-Middle,Banking,16.88,698249,Unknown,Social Engineering,VPN,47
```

Similarly, Google Trends provides time indexed values for the keywords given like ransomware, ddos etc. This is structured because each timestamp maps to numeric values for each keyword, forming a clean time series dataset. As shown in my output, it follows a clear schema and structure.

```
date,cybersecurity,SQL injection,ransomware,phishing,isPartial
2020-01-05,43,12,27,62,False
2020-01-12,45,14,27,71,False
2020-01-19,44,12,28,76,False
2020-01-26,41,13,32,66,False
2020-02-02,42,13,32,69,False
```

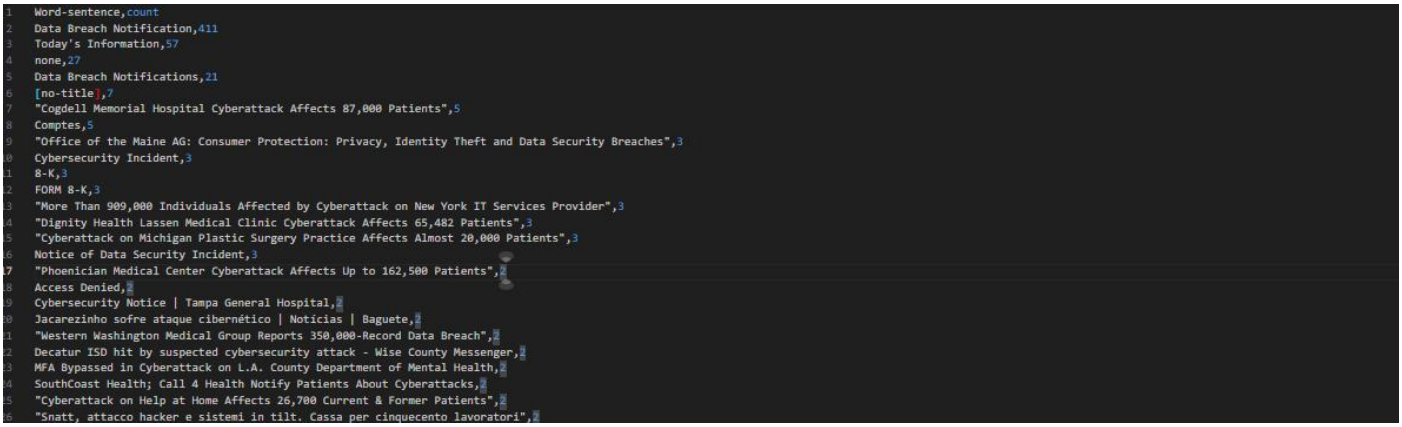
NVD raw file is semi-structured. It's not fully structured because it is a nested JSON and many fields are optional some entries contain more/less keys and nested lists. At the same time, its not unstructured because it has clear identification parameters and standard keys like ids, published date and references etc so it follows a schema but not in a fixed table format.


```
{"resultsPerPage":5,"startIndex":0,"totalResults":19551,"format":"NVD_CVE","version":"2.0","timestamp":"2026-02-11T14:32:23.539","vulnerabilities":[{"cve":{"id":"CVE-2000-1233","sourceIdentifier":"cve@mitre.org","published":"2000-12-31T05:00:00.000","lastModified":"2025-04-03T01:03:51.193","vulnStatus":"Deferred","cveTags":[],"descriptions":[{"lang":"en","value":"SQL injection vulnerability in read.php3 and other scripts in Phorum 3.0.7 allows remote attackers to execute arbitrary SQL queries via the sSQL parameter."}],"metrics":{"cvssMetricV2":[{"source":"nvd@nist.gov","type":"Primary","cvssData":
{"version":"2.0","vectorString":"AV:N/AC:L/Au:N/C:P/I:P/A:P","baseScore":7.5,"accessVector":"NETWORK","accessComplexity":"LOW","authentication":"NONE","confidentialityImpact":"PARTIAL","integrityImpact":"PARTIAL","availabilityImpact":"PARTIAL"},"baseSeverity":"HIGH","exploitabilityScore":10.0,"impactScore":6.4,"acInsufInfo":false,"obtainAllPrivilege":false,"obtainUserPrivilege":false,"obtainOtherPrivilege":true,"userInteractionRequired":false}]},"weaknesses":[{"source":"nvd@nist.gov","type":"Primary","description":
[{"lang":"en","value":"NVD-CWE-Other"}]}],"configurations":[{"nodes":
[{"operator":"OR","negate":false,"cpeMatch":[{"vulnerable":true,"criteria":"cpe:2.3:a:phorum:phorum:3.0.7:*:*:*:*:*:*:*","matchCriteriaId":"740F974F-D679-472C-966A-3E4C334E3C0A"}]}]}],"references":
[{"url":"http:\/\/cert.uni-stuttgart.de\/archive\/bugtraq\/2000\/01\/msg00215.html","source":"cve@mitre.org","tags":["Exploit","Patch","Vendor Advisory"]},{"url":"http:\/\/www.digitalsec.net\/stuff\/z-mirrors\/hispahack\/mi020.htm","source":"cve@mitre.org"},{"url":"http:\/\/cert.uni-stuttgart.de\/archive\/bugtraq\/2000\/01\/msg00215.html","source":"af854a3a-2127-422b-91ae-364da2661108","tags":["Exploit","Patch","Vendor Advisory"]},{"url":"http:\/\/hispahack.ccc.de\/mi020.html","source":"af854a3a-2127-422b-91ae-364da2661108"},{"url":"http:\/\/www.digitalsec.net\/stuff\/z-mirrors\/hispahack\/mi020.htm","source":"af854a3a-2127-422b-91ae-364da2661108"}]},{"cve":{"id":"CVE-2000-1236","sourceIdentifier":"cve@mitre.org","published":"2000-12-31T05:00:00.000","lastModified":"2025-04-03T01:03:51.193","vulnStatus":"Deferred","cveTags":[],"descriptions":[{"lang":"en","value":"SQL injection vulnerability in mod_sql in Oracle Internet Application Server (IAS) 3.0.7 and earlier allows remote attackers to execute arbitrary SQL commands via the query string of the URL."}],"metrics":{"cvssMetricV2":[{"source":"nvd@nist.gov","type":"Primary","cvssData":
{"version":"2.0","vectorString":"AV:N/AC:L/Au:N/C:P/I:P/A:P","baseScore":7.5,"accessVector":"NETWORK","accessComplexity":"LOW","authentication":"NONE","confidentialityImpact":"PARTIAL","integrityImpact":"PARTIAL","availabilityImpact":"PARTIAL"},"baseSeverity":"HIGH","exploitabilityScore":10.0,"impactScore":6.4,"acInsufInfo":false,"obtainAllPrivilege":false,"obtainUserPrivilege":false,"obtainOtherPrivilege":true,"userInteractionRequired":false}]},"weaknesses":[{"source":"nvd@nist.gov","type":"Primary","description":
[{"lang":"en","value":"NVD-CWE-Other"}]}],"configurations":[{"nodes":
[{"operator":"OR","negate":false,"cpeMatch":[{"vulnerable":true,"criteria":"cpe:2.3:a:oracle:application_server:*:*:*:*:*:*:*:*","versionEndIncluding":"3.0.7","matchCriteriaId":"7487A9CD-A2A1-4AC7-B67F-131923DE2A0F"}]}]}],"references":[{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0339.html","source":"cve@mitre.org"},{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0372.html","source":"cve@mitre.org"},
{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0463.html","source":"cve@mitre.org"},{"url":"http:\/\/online.securityfocus.com\/archive\/1\/155881","source":"cve@mitre.org"},
{"url":"http:\/\/www.iss.net\/security_center\/static\/5817.php","source":"cve@mitre.org","tags":["Patch"]},{"url":"http:\/\/www.securityfocus.com\/bid\/2150","source":"cve@mitre.org","tags":["Patch"]},{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0339.html","source":"af854a3a-2127-422b-91ae-364da2661108"},{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0372.html","source":"af854a3a-2127-422b-91ae-364da2661108"},{"url":"http:\/\/archives.neohapsis.com\/archives\/bugtraq\/2000-12\/0463.html","source":"af854a3a-2127-422b-91ae-364da2661108"},{"url":"http:\/\/online.securityfocus.com\/archive\/1\/155881","source":"af854a3a-2127-422b-91ae-364da2661108"},{"url":"http:\/\/www.iss.net\/security_center\/static\/5817.php","source":"af854a3a-2127-422b-91ae-364da2661108","tags":["Patch"]},{"url":"http:\/\/www.securityfocus.com\/bid\/2150","source":"af854a3a-2127-422b-91ae-364da2661108","tags":["Patch"]}]},{"cve":{"id":"CVE-2001-1348","sourceIdentifier":"cve@mitre.org","published":"2001-05-28T04:00:00.000","lastModified":"2025-04-03T01:03:51.193","vulnStatus":"Deferred","cveTags":[],"descriptions":[{"lang":"en","value":"TWIG 2.6.2 and earlier allows remote attackers to perform unauthorized database operations via a SQL injection attack on the id parameter."}],"metrics":{"cvssMetricV2":[{"source":"nvd@nist.gov","type":"Primary","cvssData":
```


For the ransomware/cyberattacks JSON, the dataset is also semi-structured overall, but it contains unstructured data inside it, especially the summary and sometimes title which is free text. Also, there is inconsistency because some entries have more keys than others example: some have extra nested info while others don't, and sometimes the title appears in different languages


```
[
  {
    "claim_gang": false,
    "claim_url": false,
    "country": "US",
    "date": "2022-02-28",
    "domain": "kannapolisnc.gov",
    "has_infostealer_info": false,
    "link": "https://www.ransomware.live/id/a2FubmFwb2xpc25jLmdvdkAyMDIyLTAyLTI4",
    "summary": "La ville de Kannapolis, en Caroline du Nord, a été victime d'une cyberattaque en février 2022 qui a perturbé son système de dispatch pour la police et les pompiers, mais n'a pas informé le public de l'incident. Les responsables de la ville ont déclaré
    "title": "9 Investigates: Kannapolis didn't alert public when cyberattack knocked out police dispatch - WSOC TV",
    "url": "https://www.wsoctv.com/news/local/9-investigates-kannapolis-didnt-alert-public-when-cyberattack-knocked-out-police-dispatch/UEMD6NRNOJGE3N5UHCRIUBTECE/",
    "victim": "Kannapolis"
  },
```


From an **unstructured data** perspective I have not found direct source for my choosen thematic but the tittle field is the unstructured part in ransomware_live(allcyberattacks.json) I have extracted as a separate .csv for visulization at a processing step. (it's a derived feature)



```
Word-sentence,count
Data Breach Notification,411
Today's Information,57
none,27
Data Breach Notifications,21
[no-title],7
"Cogdell Memorial Hospital Cyberattack Affects 87,000 Patients",5
Comptes,5
"Office of the Maine AG: Consumer Protection: Privacy, Identity Theft and Data Security Breaches",3
Cybersecurity Incident,3
8-K,3
FORM 8-K,3
"More Than 909,000 Individuals Affected by Cyberattack on New York IT Services Provider",3
"Dignity Health Lassen Medical Clinic Cyberattack Affects 65,482 Patients",3
"Cyberattack on Michigan Plastic Surgery Practice Affects Almost 20,000 Patients",3
Notice of Data Security Incident,3
"Phoenician Medical Center Cyberattack Affects Up to 162,500 Patients",2
Access Denied,2
Cybersecurity Notice | Tampa General Hospital,2
Jacarezinho sofre ataque cibernético | Notícias | Baguete,2
"Western Washington Medical Group Reports 350,000-Record Data Breach",2
Decatur ISD hit by suspected cybersecurity attack - Wise County Messenger,2
MFA Bypassed in Cyberattack on L.A. County Department of Mental Health,2
SouthCoast Health; Call 4 Health Notify Patients About Cyberattacks,2
"Cyberattack on Help at Home Affects 26,700 Current & Former Patients",2
"Snatt, attacco hacker e sistemi in tilt. Cassa per cinquecento lavoratori",2
```

(b) -Answer:

Extraction challenges varied by source. For dataset extraction, the main issue was rate limiting: I repeatedly received "Too Many Requests" errors, so I reduced the number of keyword requests and added delays and retry logic in the code. API key maintenance was also a challenge for sources that require keys, since they can expire over time. Another issue was data-format inconsistency. Different sources are hosted on different websites and follow different structures. The JSON data is nested and semi-structured, so some records have missing fields or extra keys, and the structure is not consistent across entries. Because of this, mapping or comparing data between two threat sources becomes more difficult. Finally, one source provided the raw dataset as a .zip file, which required an extra step to download and extract it before I could access the CSV.

(c) -Answer:

Storing data in multiple formats (CSV + JSON) is valuable in data engineering because different formats are useful for different purposes. In practical, CSV for flat tables and BI/analytics tools while JSON for API ingestion, logs, and nested records. CSV is best when the data is structured with a fixed schema (rows and columns). JSON is best when the data is semi-structured and contains nested fields or variable keys. It preserves the original structure of API responses (lists inside objects, optional fields) without losing information. Storing both formats is important because we can keep raw JSON for reproducibility and also create CSV for same data for easy analysis and reporting.