Rubina Noor
25280028

Part 1 Questions
(a) Data Heterogeneity: Explain how your chosen data sources represent different data types (structured, semi-structured, unstructured).
Provide concrete examples from your extracted data.

HealthData API Dataset:
It is structured with tabular rows and columns with a fixed schema.

	round	indicator	group	subgroup	sample_size	response	percent	standard_error	suppression	significant_1	significant_2
0	1	Provider offers telemedicine	Total	Total	6786.0	Total	100.0	NaN	NaN	NaN	NaN
1	1	Provider offers telemedicine	Total	Total	NaN	Yes	36.6	0.8	NaN	NaN	NaN
2	1	Provider offers telemedicine	Total	Total	NaN	No	46.9	0.8	NaN	NaN	NaN
3	1	Provider offers telemedicine	Total	Total	NaN	Do not know	5.2	0.4	NaN	NaN	NaN
4	1	Provider offers telemedicine	Total	Total	NaN	No usual place of care	11.3	0.7	NaN	NaN	NaN

```
round              int64
indicator            str
group                str
subgroup             str
sample_size        float64
response             str
percent            float64
standard_error     float64
suppression          str
significant_1      float64
significant_2      float64
dtype: object
```

Kaggle Dataset:
Format of dataset: CSV file
It is structured with tabular rows and columns with a fixed schema. There is some aspect of semi-structuredness in the dataset like the column Blood_Pressure_mmHg which combines two values and needs to be broken down into two columns.

	Patient_ID	Age	Gender	Symptom_1	Symptom_2	Symptom_3	Heart_Rate_bpm	Body_Temperature_C	Blood_Pressure_mmHg	Oxygen_Saturation_%	Diagnosis	Severity	Treatment_Plan
0	1	74	Male	Fatigue	Sore throat	Fever	69	39.4	132/91	94	Flu	Moderate	Medication and rest
1	2	66	Female	Sore throat	Fatigue	Cough	95	39.0	174/98	98	Healthy	Mild	Rest and fluids
2	3	32	Male	Body ache	Sore throat	Fatigue	77	36.8	136/60	96	Healthy	Mild	Rest and fluids
3	4	21	Female	Shortness of breath	Headache	Cough	72	38.9	147/82	99	Healthy	Mild	Rest and fluids
4	5	53	Male	Runny nose	Sore throat	Fatigue	100	36.6	109/106	92	Healthy	Mild	Rest and fluids

```
Patient_ID             int64
Age                    int64
Gender                   str
Symptom_1                str
Symptom_2                str
Symptom_3                str
Heart_Rate_bpm         int64
...
Diagnosis                str
Severity                 str
Treatment_Plan           str
dtype: object
```

Google Trends:
Format: Dataframe from Pytrends

Structure: It is a semi-structured data with time-indexed series representing searches for keywords. Each column is numeric and represents search interest for the keywords mentioned.

	date	telemedicine	digital health	mental health app	wearable health device	isPartial
0	2004-01-01	10	3	0	0	False
1	2004-02-01	10	2	0	0	False
2	2004-03-01	10	2	0	0	False
3	2004-04-01	9	1	0	0	False
4	2004-05-01	9	2	0	0	False

```
date                    str  ← needs formatting to_date
telemedicine          int64
digital health        int64
mental health app     int64
wearable health device int64
isPartial              bool
dtype: object
```

(b) Extraction Challenges: Discuss specific technical or practical challenges encountered while accessing different data sources (rate limits, authentication, data format inconsistencies, etc.).
Answer:
Pytrends often times returns HTTP 429 “Too Many Requests” which causes delays or retries.
The data extracted using API from HealthData.gov had missing values and nested dictionaries which required preprocessing.

(c) Storage Justification: Explain why storing data in multiple formats (CSV, JSON) is valuable in a data engineering context. When would you choose one format over another?
Answer:
CSV Format:
This format is used for simple, tabular and structured data.
It is ideal for local datasets, time series or data that needs to be analyzed in Excel and Pandas or SQL.

JSON Format:
JSON format supports data that is in forms of nested dictionaries or data that is semi-structured. It helps to preserve flexibility and original hierarchy for future transformations.
For example the data extracted from HealthData.gov API was nested and storing it in JSON preserved the original API response

Part 2 Questions
(a) Cleaning Rationale: Justify your data cleaning decisions. Why were specific approaches chosen for handling missing data or outliers?

For API Dataset

- To handle missing values for ‘percent’ and ‘standard_error’ columns, i used median of each column because median is less affected by outliers. This maintains the overall distribution of the data without being heavily influenced by the outliers.
- To handle missing values for ‘supression’ and ‘significance’ columns, i created indicator variables (1 and 0) to show whether the info is present or not because missing values in these columns may carry meaning.
- Used strip() to remove whitespaces from start and end and title() to capitalize first letter and make the rest lowercase to ensure consistent formatting for all categorical columns like ‘indicator’, ‘group’, ‘subgroup’, ‘response’.

For Loaded Dataset from Kaggle:

- The ‘Blood_Pressure_mmHg’column was broken down into two separate columns called ‘systolic’ and ‘diastolic’so it could be used for analysis purposes and be used numerically.
- Used strip() to remove whitespaces from start and end and title() to capitalize first letter and make the rest lowercase to ensure consistent formatting for all categorical columns ‘Gender’, ‘Diagnosis’,’Treatment’, ‘Severity’, ‘Symptom_1’, ‘Symptom_2’, ‘Symptom_3’.

For Google Dataset:

- Converted date to datetime format so it could be used for analysis purposes.
- Dropped the isPartial column since it was not needed.
- Set date as index as the data represents time-series so it could be used for timeseries analysis and could be accessed in terms of intervals too if needed.

(b) Visualization Insights: What key insights or patterns emerge from your visualizations? How do they relate to your chosen thematic domain?

Answer:
API Bar Chart
The majority of the responses fall under “Yes” and “No”, with few “Do not know” and “No usual place of care”.
This shows moderate adoption of telemedicine but it is still not universally available.
Thematic Relevance: this directly connects to digital health trends and accessibility.

Loaded DataSet Box Plot
Moderate severity cases show consistently high median body temperature.
Mild cases show generally lower median body temperature.
Severe cases show wider spread, suggesting variability.
Thematic relevance: this visualization connects patient health outcomes with measurable biological data.

Google Trends Time Series Graph
The time-series graph shows that interest in telemedicine and digital health has increased over time. The noticeable spikes in digital health searches show that major global events pushed people to look online for healthcare solutions.

Later spikes in wearable health devices and mental health apps shows that people are becoming more aware of and interested in digital health tools over time.
Thematic relevance: This helps connect public behavior with real-world healthcare practices.

(c) Visualization Critique: What limitations exist in your current visualizations? How could they be improved for different audiences (technical vs. business stakeholders)?

Bar Chart

- It shows frequency but not intervals.
- It does not reflect percentage proportions directly unless labeled.

Boxplot

- It does not show exact data points.
- The sample size per severity group is unclear.
- It may oversimplify variability.

Time-Series Plot

- The multiple lines make interpretation difficult.
- The search interest values are relative not absolute.
- It does not explain causality behind spikes.

Improvements for Different Audiences

For Technical Stakeholders

- Add regression lines and correlation coefficients.
- Include intervals and statistical testing.
- Use interactive dashboards (Plotly, Tableau).

For Business or Policy Stakeholders

- Add percentage labels directly on bars.
- Highlight key insights with annotations.
- Simplify time-series to focus on 1–2 main variables.
- Use summary style visuals with clear takeaway messages.

Attributed sources:
Referenced for pytrends and time-series diagram:
https://thepythoncode.com/article/extract-google-trends-data-in-python

HealthData API used from
:https://healthdata.gov/CDC/Telemedicine-Use-in-the-Last-4-Weeks/jnr3-qn3j/about_data