Part 1

Q1. Data Heterogeneity:

(a) Structured Data:

The UPI transactions dataset is a structured data, which is downloaded from Kaggle in the form of a CSV file. It has fixed columns like UPI Banks, Value (Cr), Volume, Month, Year etc. The rows indicate the volume and value of transaction of a given bank within a certain month and year. Each row is of the same format and which made it easy to group, filter and calculate totals. For example, I can categorize the information by the bank name and add up the total volume of transactions without having to change the structure. It is also simple to operate with this form of data since all the data is already organized into rows and columns.

(b) Semi- Structured Data:

The data on the cryptocurrencies was retrieved via an API (Coingecko) and was returned as JSON. Such data is semi-structured since not every cryptocurrency had the same information. It has columns like name, market cap, current price, max supply, ROI etc. However, some coins had missing values for roi and max supply. For example some coin did not have value in ROI column whereas other coins did. This shows that there was inconsistency between records in the structure. Also, the JSON data required to be converted into table format using Pandas before analysis. This shows that API data is frequently more difficult to clean and format than a simple CSV file.

(c) Time Series Data:

The stock data (yf_df) is a structured data set that includes time-dependent information in it, which is not similar to the static tabular data because it has a time dependent nature. The stock price data comprises of the closing price of Paypal, MasterCard, and Visa on a daily basis in 2021 to 2026. It is time based information, whereby each row is linked to a particular date. UPI data set had aggregated monthly prices whereas this data set includes daily closing prices from 2021 to 2026. The data is non-uniform and sequential because of non-trading days (weekends and holidays) and brings temporal irregularity. This type of data requires chronological sorting, date and time conversion etc.

The structure, semi-structured, and time based data in a single pipeline shows the heterogeneity characteristic of financial technology ecosystems.

Q2. Extraction Challenges:

Some technical and practical issues were found during the data extraction process when accessing the API, time-series data, and UPI transaction dataset. These issues outline real-life challenges in the integration of multiple financial data.

(a) Inconsistent Data Fields in API Response:

The cryptocurrency data was accessed using the API in the format of JSON. Main problem was that not every cryptocurrency had the same fields. For example, there were missing values in the columns roi and max supply for some coins. The resulting inconsistency caused problems in the process of creating a structured table in Pandas with the JSON response. Such missing values were required to be identified and processed accordingly, before to carrying out ranking or aggregation processes. Without this, there would have been errors in calculations like sorting by market value or making numeric comparisons. This proves the fact that API data is not complete or consistent and should be thoroughly checked after extraction.

(b) Time-Based Data Gaps in Stock Prices:

The data on the Yahoo Finance included the daily stock price of PYPL, MA, and V. Financial markets are, however, not open on the weekends or on the public holidays. This means the data is not continuous and the dataset has natural data gaps by date. It requires transforming the date column to proper datetime format, maintaining proper chronological sorting and formatting the x-axis carefully during visualization to avoid overcrowding. Time-series data may need more preprocessing than static tabular data because it relies on chronological continuity.

(c) Handling Missing and Null Values:

The API displayed the null values of some of the numerical fields. These null values were to be transformed by putting 0 or mean value, so that they could be further analyzed. This was necessary since it is not possible to do numerical operations in Python with null values. The analysis pipeline would have been disrupted without dealing with missing data.

(d) Combining Temporal Fields in UPI Dataset:

Month and Year were put in different columns of transaction dataset. This arrangement made visualization of time-series difficult. These two fields were required to be combined into one single column, the month-year column in order to produce the trend per month. This shows that structured CSV data can still be subject to transformation before being analysis-ready.

Q3. Storage Justification:

The choice to store data in CSV format and to store data in JSON were conscious decisions in this assignment that shows good data engineering. This data of the cryptocurrency was retrieved as an API in JSON format, which is the most appropriate to save the raw response as it is. JSON is also flexible enough to have records in cases where some of the fields are not present in the record, or the values are different across different record values, like the roi and max supply values that were not present in all the coins. Storing the raw data as JSON will make the data transparent and enable reprocessing of the data in the future without placing another API call. The data was cleaned and then sorted into a coherent tabular format and was saved as CSV since CSV is more effective and convenient with structured analysis. CSV files have the lightweight

nature, are easy to read and can be analyzed with analytical packages such as Pandas, Excel, and visualization libraries. Certainly, in data engineering, JSON is normally used when processing the raw responses of APIs or other flexible data formats, and CSV is used when the data is organized, cleansed, and prepared to be analyzed or reported. The reproducibility, flexibility, and the ease of downstream processing are facilitated by the use of both formats.

Part 2

Q1. Cleaning Rationale:

(a) **Handling Missing Values in Cryptocurrency Data**

In the crypto-data, there were blank values in the block like roi and max supply. Such missing values were dealt with by replacing the values with 0. These were all numerical columns that would be left as null, which would lead to sort errors or aggregate errors. Major ranking measure employed in the analysis was the market cap rather than roi and max supply. Assigning averages or approximated values would make untestable artificial assumptions regarding financial measures. The replacement with 0 is also a clear indication of no data being reported but still is numerically consistent.

(b) Data Type Standardization

There were no gaps in the UPI data set. The Date columns however were changed into date time. Month and Year were merged together into one time reference in order to be able to chart them properly in a chronological manner. It was a structural adjustment and not correction of mistakes. Date time formatting makes time based visualization accurate and standardization makes it reliable in grouping and aggregation.

(c) Outlier Treatment

Outlier elimination was not done deliberately. Extreme values in financial data are commonly reflective of true market phenomena. The elimination of these values would decrease the authenticity of financial behavior. Because this was a descriptive analysis and not predictive modeling, it would be more appropriate to retain real market variation than smooth or cut off extremes.

Q2. Visualization Insights:

(a) **Cryptocurrency Market Cap Distribution**
The pie chart of the top 10 cryptocurrencies indicates that Bitcoin is the largest share (63) of total market capitalization, which is followed by Ethereum (11), Tether (8), and XRP (4) with the remaining part of other cryptocurrencies having less than 4. This shows how the cryptocurrency market is excessively controlled by a limited number of large actors and the role played by big digital assets in developing blockchain-based financial infrastructure, investments, and applications of decentralized finance.

(b) **Top Banks by UPI Transaction Volume**
The horizontal bar graph indicates that the category of Google Pay is second in about

14,000 million transactions, whereas PhonePe is first in about 17,500 million transactions, and Paytm Payment Bank was third. This indicates that in India, there are a few leading platforms that control the digital payments. It also demonstrates the changing nature of traditional banking and payment services by technology-based platforms, which make mass adoption of mobile and app-based transactions possible and gives an overview of the competition within the digital finance industry.

(c) **Monthly UPI Transaction Trend**

The trend line indicated monthly demonstrates the evident upward trend of transaction volumes in UPI. The dips are minor around April and May then there is a steep rise surpassing 4,500 million transactions compared to the start of 2,500 million. This trend indicates the use and the confidence in the digital payment systems. These trends reflect in practice how technological innovation influenced the consumer payment behavior and how FinTech solutions are replacing traditional cash transactions, and how they are driving financial inclusion.

(d) **Stock Price Trends of FinTech Companies**

The line plots of Visa, MasterCard and PayPal indicate that Visa and MasterCard have comparable growth trends, which have been generally upward whereas PayPal has had slightly a downward or a less steep growth trend in the same duration. Correlation analysis has shown that Visa and MasterCard have very high positive correlation (0.99), but the correlation between Visa and PayPal (-0.4) and MasterCard and PayPal (-0.44) exhibit moderate negative correlations. These analyses reveal the contrasting market forces and future growth trends of the traditional payment processors and the new digital payment systems.

Q3. Visualization Critique:

Although the existing visualizations do offer valuable information, they have certain limitations that can be improved to help understand them better and be more meaningful to the audience. In the case of a cryptocurrency market cap pie chart, 10 coins is sufficient to give a good picture of what is happening in the market, however, the performance of smaller coins, which may be of interest in a detailed market analysis, is obscured. An interactive visualization or stacked bar chart would give the opportunity to view the entire distribution.

The UPI transaction volume bar charts do not prioritize the regional differences or the demographics of the users; however, it emphasizes on the top banks and general trends. The visualizations might be more informative by including extra dimensions e.g. city-wise or age-wise adoption to business stakeholders intending to plan the business market strategies. Equally, the monthly trend line now displays aggregate volumes with no annotations of crucial events or policy adjustments that could be the cause of spikes or dips.

Stock price trends are effective in illustrating the differences between Visa, MasterCard and PayPal but line plots can be difficult to interpret especially when it comes to correlations to the non technical audiences. The trends may be made more accessible to the business decision-

makers through the use of a combination of line plots and shaded correlation matrices or summary tables.