

Data Collection Report

Github repo link: <https://github.com/mubasharjaved3567/Project-Data-Mosaic>

Group Information:

- **Group Number:**16
- **Student 1:** Zafir Muhammad, ID: 24280031
- **Student 2:** Muhammad Mubashar Ali, ID: 24280037

Contributions:

- **Zafir Muhammad:** Implemented data collection scripts, integrated APIs.
- **Muhammad Mubashar Ali:** Conducted data analysis, performed dataset cleaning, and wrote the report.

1. Overview of the Topic Chosen Topic: Electric Vehicles (EVs)

Reason for Selection: Electric Vehicles (EVs) are a rapidly growing industry due to their environmental benefits and technological advancements. The transition to sustainable transportation is a critical global movement. Understanding the adoption, market trends, and public perception of EVs can provide valuable insights into the future of this industry.

Expected Data:

- **Reddit Discussions:** Public sentiment, common concerns, and emerging trends.
- **Stock Market Data:** Trends in EV company stocks such as Tesla, General Motors, and Ford.
- **EV Sales Data:** Growth trends, country-wise adoption, and model-wise sales.

2. Data Collection Process

Reddit Data:

- **Sources:** r/ElectricVehicles, r/TeslaMotors
- **Keywords Used:** "electric vehicle," "Tesla," "EV charging"
- **Fields Collected:** title, text, author, date, upvotes, subreddit
- **Challenges Faced:** Reddit API rate limits, missing or deleted posts, and ensuring compliance with Reddit's Terms of Service.

Yahoo Finance Data:

- **Tickers Monitored:** "TSLA" (Tesla), "GM" (General Motors), "F" (Ford), "NIO" (Nio)

- **Fields Collected:** date, open, high, low, close, volume
- **Challenges Faced:** Missing data points in stock history, handling API failures, and ensuring data consistency.

EV Sales Dataset:

- **Source:** datahub.io (historic-sales-of-electric-vehicles)
- **Fields Collected:** Sales by country, model, and year
- **Challenges Faced:** Some missing values, inconsistent formatting, and ensuring proper dataset versioning.

3. Initial Observations Using pandas, we generated summary statistics of the datasets. Below is a brief overview:

- **Reddit Data:** High engagement in discussions around charging infrastructure and range anxiety.
- **Yahoo Finance Data:** Tesla stock experiences significant fluctuations around major EV announcements.
- **Public EV Sales Data:** A strong upward trend in EV adoption, with China and the US leading in sales.

Summary statistics for datasets/raw/TSLA_stock.csv:

	Date	Open	...	Dividends	Stock Splits
count	251	251.000000	...	251.0	251.0
unique	251	NaN	...	NaN	NaN
top	2024-02-15 00:00:00-05:00	NaN	...	NaN	NaN
freq	1	NaN	...	NaN	NaN
mean	NaN	252.709841	...	0.0	0.0
std	NaN	87.456641	...	0.0	0.0
min	NaN	140.559998	...	0.0	0.0
25%	NaN	182.050003	...	0.0	0.0
50%	NaN	221.190002	...	0.0	0.0
75%	NaN	328.815002	...	0.0	0.0
max	NaN	475.899994	...	0.0	0.0

Summary statistics for datasets/raw/reddit_posts.csv:

	title	...	subreddit
count	25	...	25
unique	25	...	2
top	Executive order signed to reverse US electric	...	ElectricVehicles
freq	1	...	15
mean	NaN	...	NaN
std	NaN	...	NaN
min	NaN	...	NaN
25%	NaN	...	NaN
50%	NaN	...	NaN
75%	NaN	...	NaN
max	NaN	...	NaN

Summary statistics for datasets/raw/NIO_stock.csv:

	Date	Open	...	Dividends	Stock Splits
count	251	251.000000	...	251.0	251.0
unique	251	NaN	...	NaN	NaN
top	2024-02-15 00:00:00-05:00	NaN	...	NaN	NaN
freq	1	NaN	...	NaN	NaN
mean	NaN	4.842869	...	0.0	0.0
std	NaN	0.709933	...	0.0	0.0
min	NaN	3.680000	...	0.0	0.0
25%	NaN	4.355000	...	0.0	0.0
50%	NaN	4.630000	...	0.0	0.0
75%	NaN	5.290000	...	0.0	0.0
max	NaN	7.700000	...	0.0	0.0

Summary statistics for datasets/raw/GM_stock.csv:						
	Date	Open	...	Dividends	Stock	Splits
count	251	251.000000	...	251.000000		251.0
unique	251	NaN	...	NaN		NaN
top	2024-02-15 00:00:00-05:00	NaN	...	NaN		NaN
freq	1	NaN	...	NaN		NaN
mean	NaN	47.181985	...	0.001912		0.0
std	NaN	4.640577	...	0.015057		0.0
min	NaN	37.955583	...	0.000000		0.0
25%	NaN	44.311004	...	0.000000		0.0
50%	NaN	46.399288	...	0.000000		0.0
75%	NaN	50.440001	...	0.000000		0.0
max	NaN	59.027030	...	0.120000		0.0

Summary statistics for datasets/raw/TSLA_stock.csv:						
	Date	Open	...	Dividends	Stock	Splits
count	251	251.000000	...	251.0		251.0
unique	251	NaN	...	NaN		NaN
top	2024-02-15 00:00:00-05:00	NaN	...	NaN		NaN
freq	1	NaN	...	NaN		NaN
mean	NaN	252.709841	...	0.0		0.0
std	NaN	87.456641	...	0.0		0.0
min	NaN	140.559998	...	0.0		0.0
25%	NaN	182.050003	...	0.0		0.0
50%	NaN	221.190002	...	0.0		0.0
75%	NaN	328.815002	...	0.0		0.0
max	NaN	475.899994	...	0.0		0.0

4. Terms of Service s Privacy Issues

- **Reddit:** Reddit’s API Terms of Service restrict storing and redistributing user-generated content. Data should be anonymized before analysis.

5. Multi-Source Data Integration Challenges

- **Data Inconsistencies:** Differences in sources can lead to conflicting insights.
- **Varying Data Formats:** Reddit provides unstructured text, whereas stock market and sales data are structured.
- **Bias in Data:** Online discussions may not reflect real-world adoption rates due to demographic differences.

6. Data Storage s Integration Strategy

- **PostgreSQL:** To store structured financial and sales data.
- **MongoDB:** For unstructured Reddit discussions.
- **Data Warehouse:** Combining all sources for efficient querying and analysis.

Part 3: Pipeline Architecture

