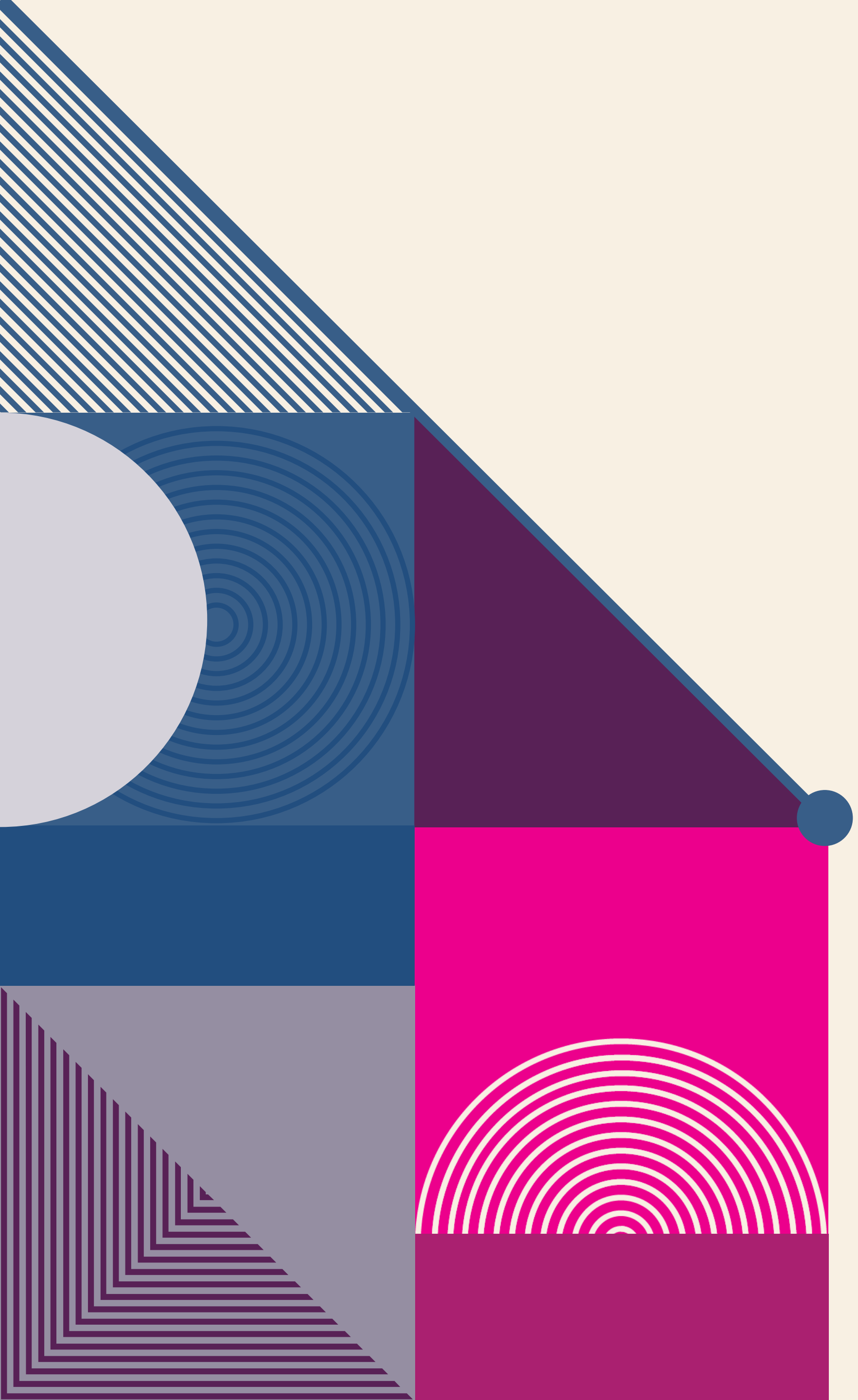
DATA SCIENCE CAPSTONE PROJECT

IJAZ AHMED
01.09.2025



OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

EXECUTIVE SUMMARY

Studied what makes a launch successful by fetching and preparing data for different Predictive Models which and then making a Dashboard for the live analytics

INTRODUCTION

SpaceX is a successful company for making the space travel affordable (62 Million Dollars instead of 165 Million Dollars) because they can reuse the first stage. Using ML Models we want to see what is the best Model to make predictions when they can reuse the first stage.

But

- How often, and what factors are important?

METHODOLOGY

Data Collection using APIs and Scrapping Wikipedia

Data Prepearation

EDA with Python and SQL

Predictive Analysis using ML Models

Finding the **Best Model**

Interactive Visuals using Folium and Plotly Dash

DATA ACCUMULATION

API-calls: - Data Collection API

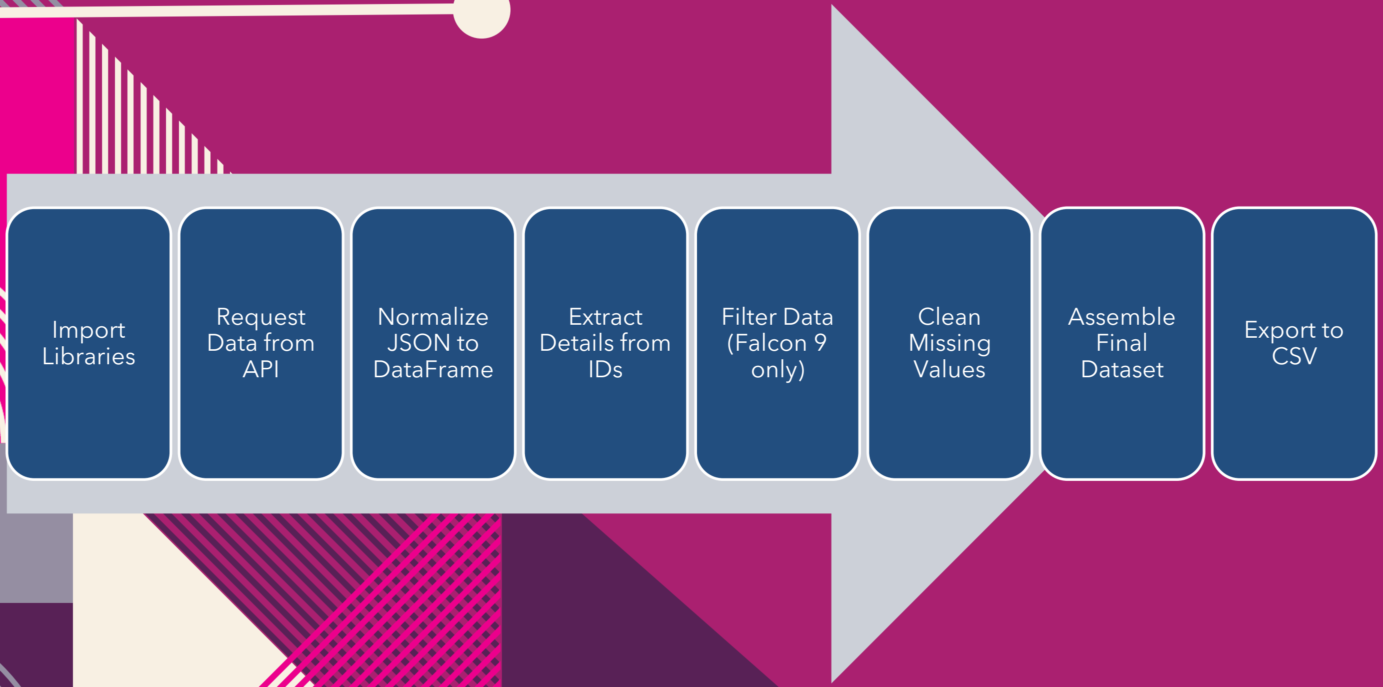
SpaceX Api

Scrapping:

Using requests module

Using Beautiful Soap and html parser

DATA COLLECTION:



Import Libraries

Request Data from API

Normalize JSON to DataFrame

Extract Details from IDs

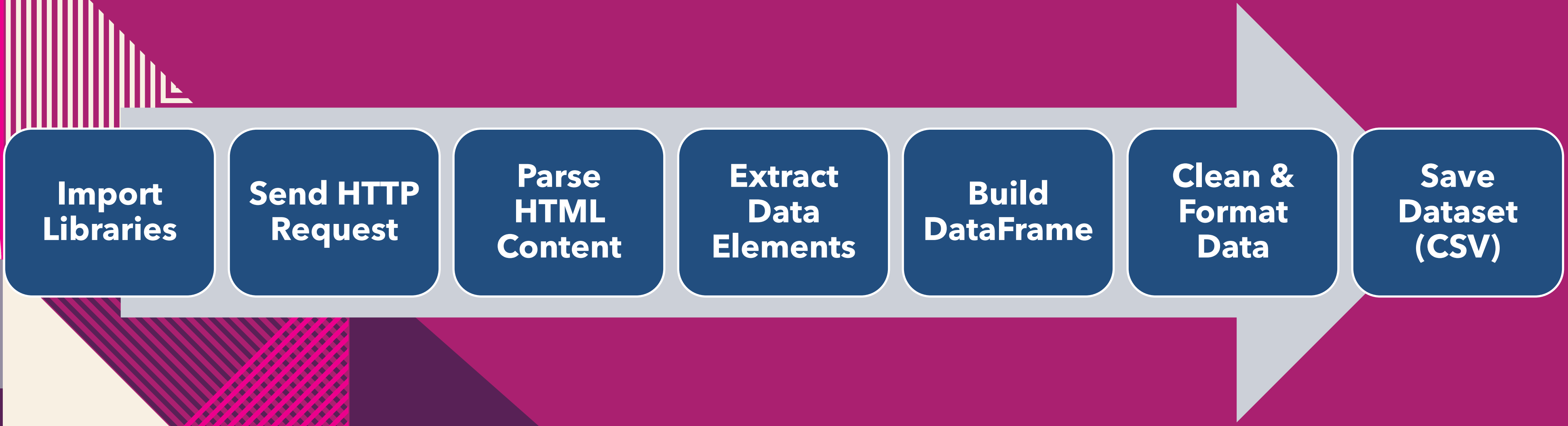
Filter Data (Falcon 9 only)

Clean Missing Values

Assemble Final Dataset

Export to CSV

WEB SCRAPING



Import Libraries

Send HTTP Request

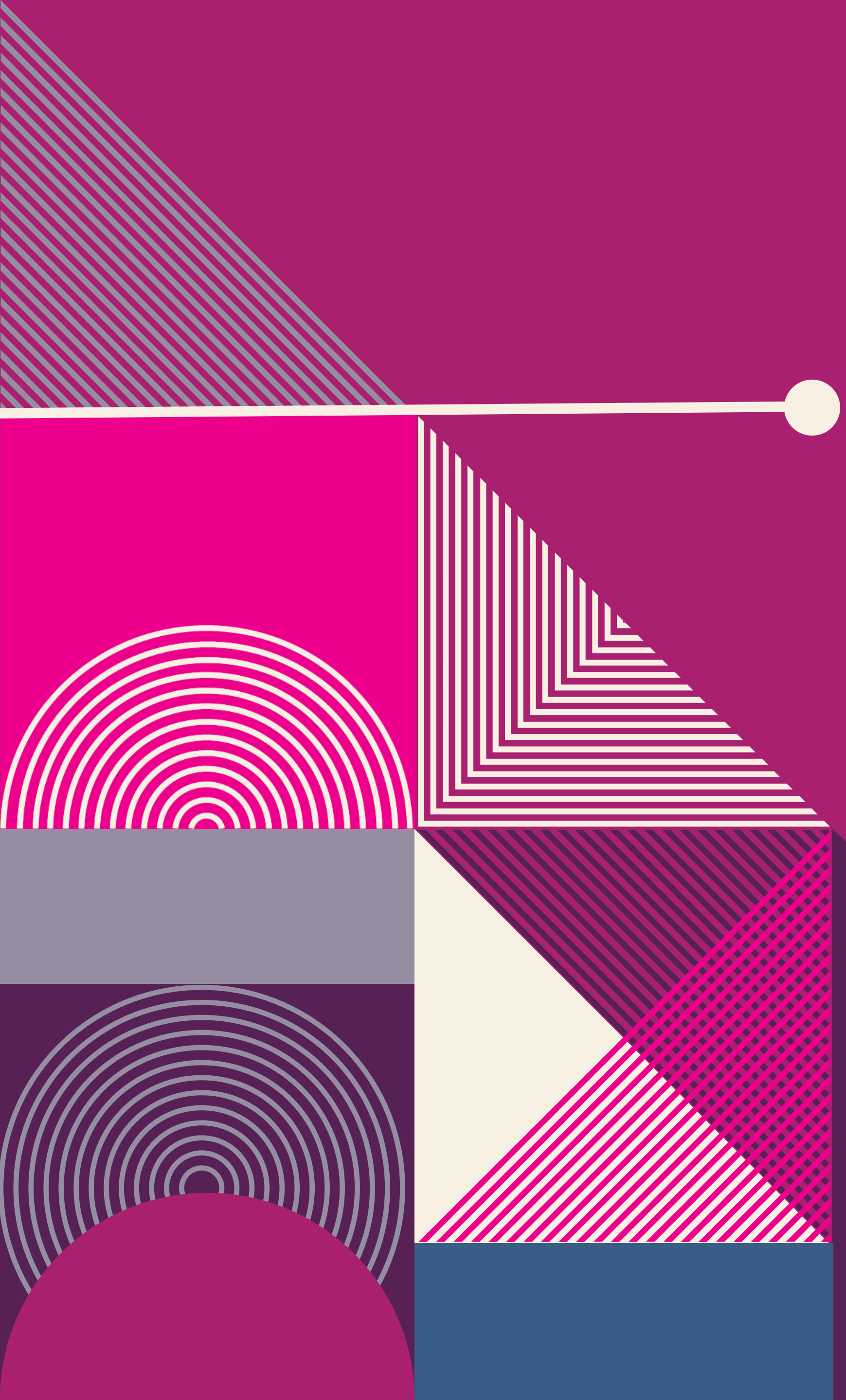
Parse HTML Content

Extract Data Elements

Build DataFrame

Clean & Format Data

Save Dataset (CSV)



Import Libraries

Tools: requests, BeautifulSoup, pandas.

Purpose: enable HTTP requests, HTML parsing, and structured data storage.

GETTING HTTP RESPONSE

Send HTTP Request

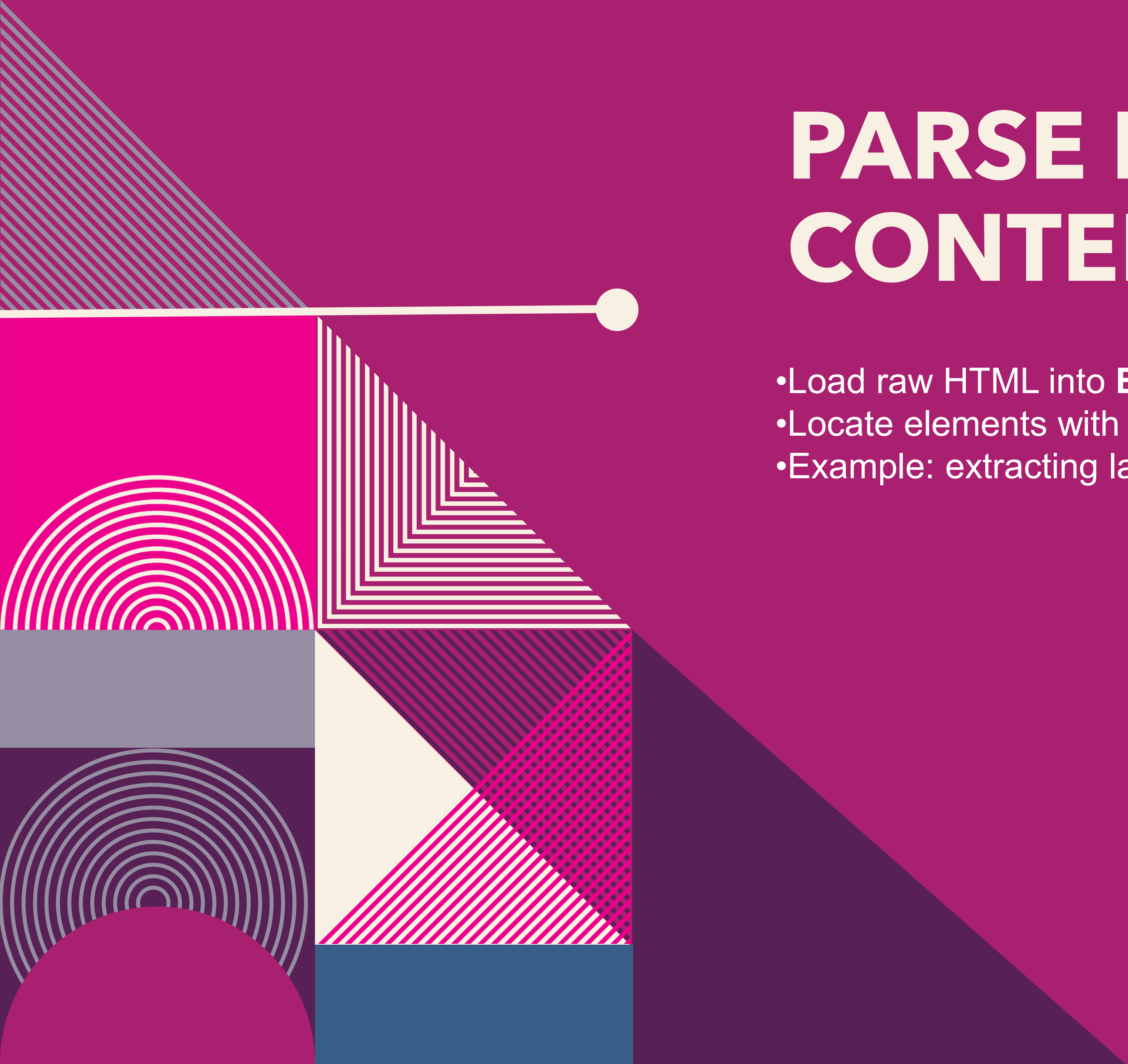
Use requests.get(url) to retrieve webpage content.

Example: scraping Wikipedia page on Falcon 9 launches.

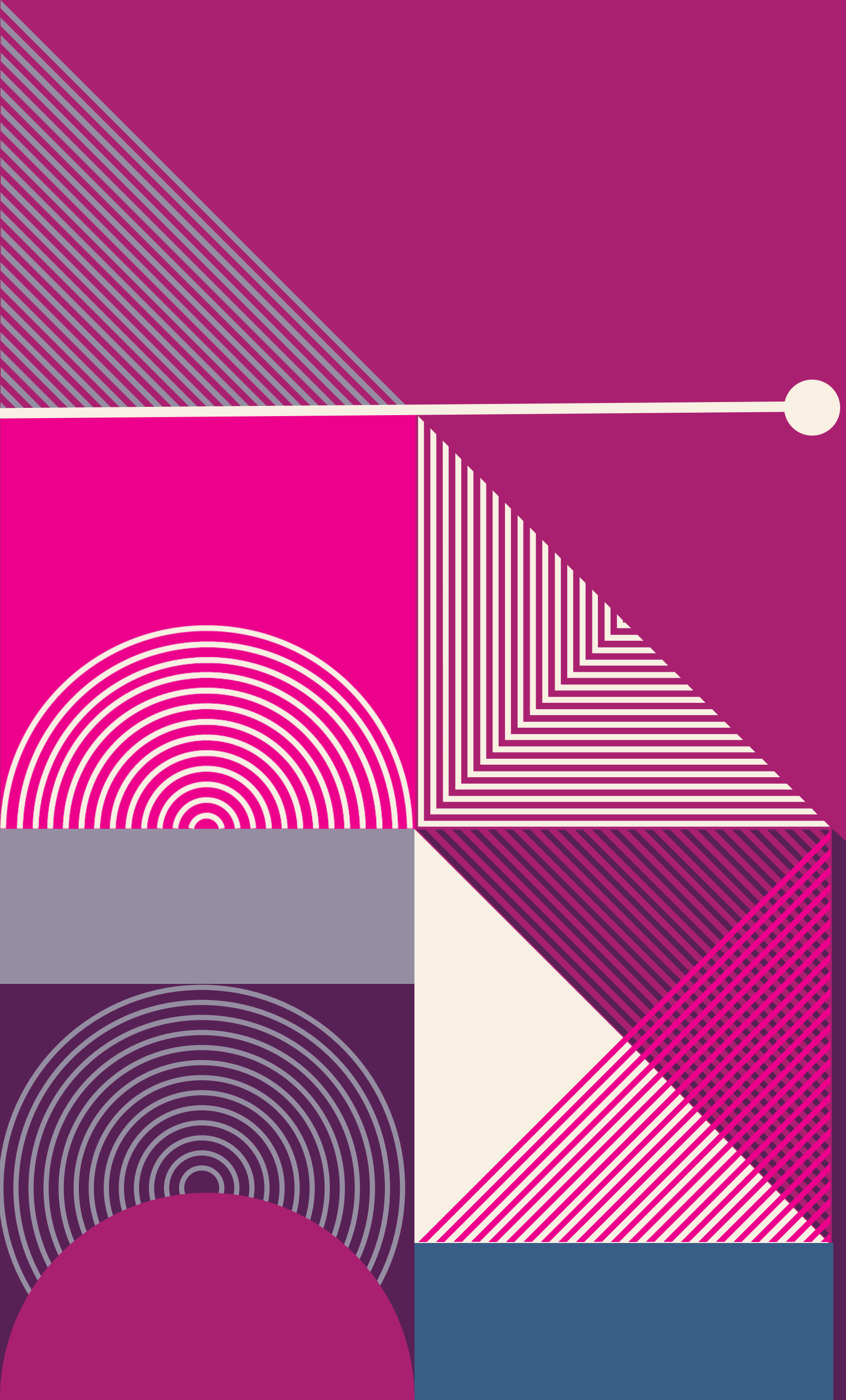
Response codes (200 = success).

We initially get 403 error for scraping Wikipedia which can be easily resolved using a User-Agent.

PARSE HTML CONTENT



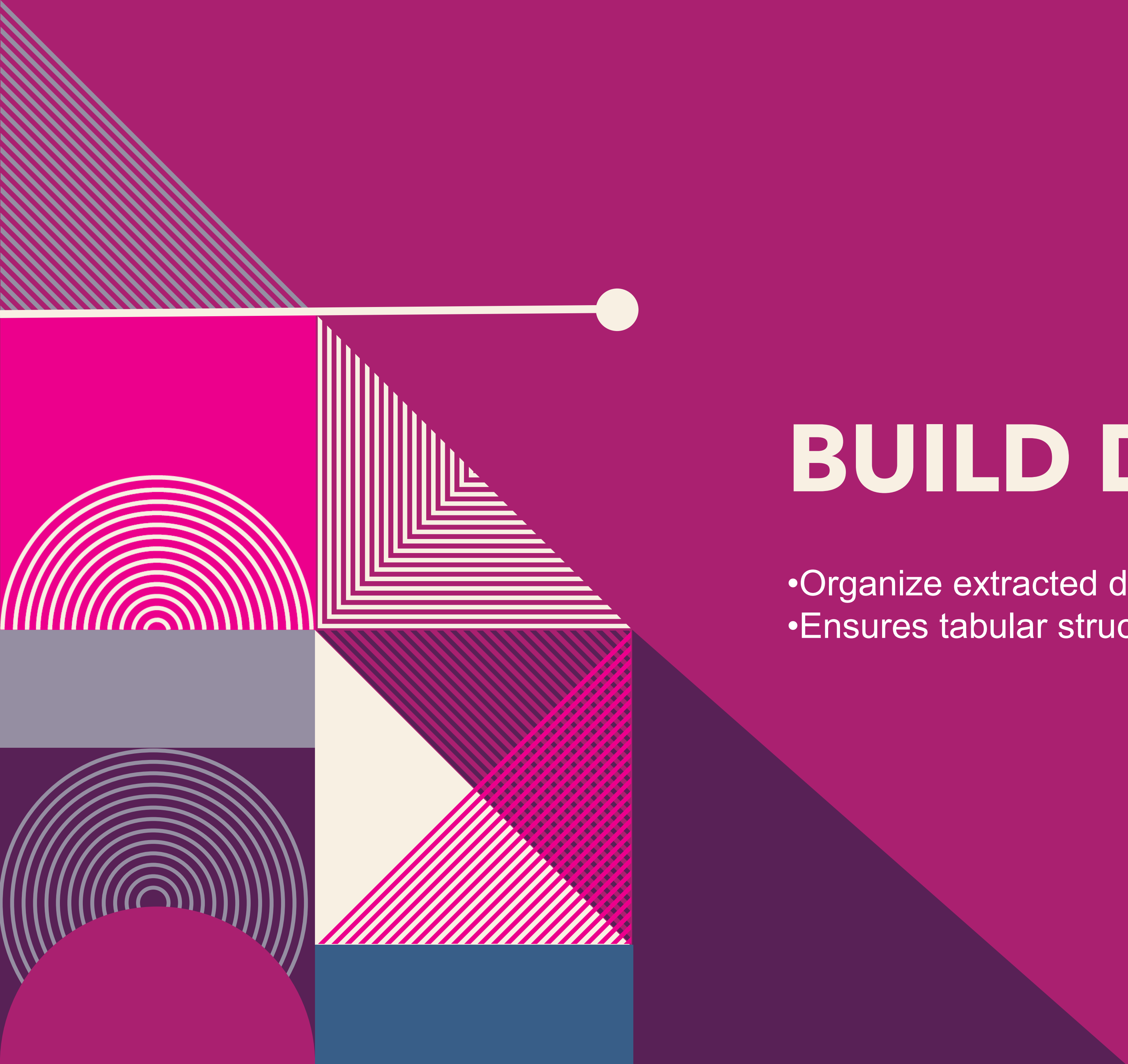
- Load raw HTML into **BeautifulSoup**.
- Locate elements with tags (<table>, <tr>, <td>) or CSS selectors.
- Example: extracting launch tables from Wikipedia.



EXTRACT DATA ELEMENTS

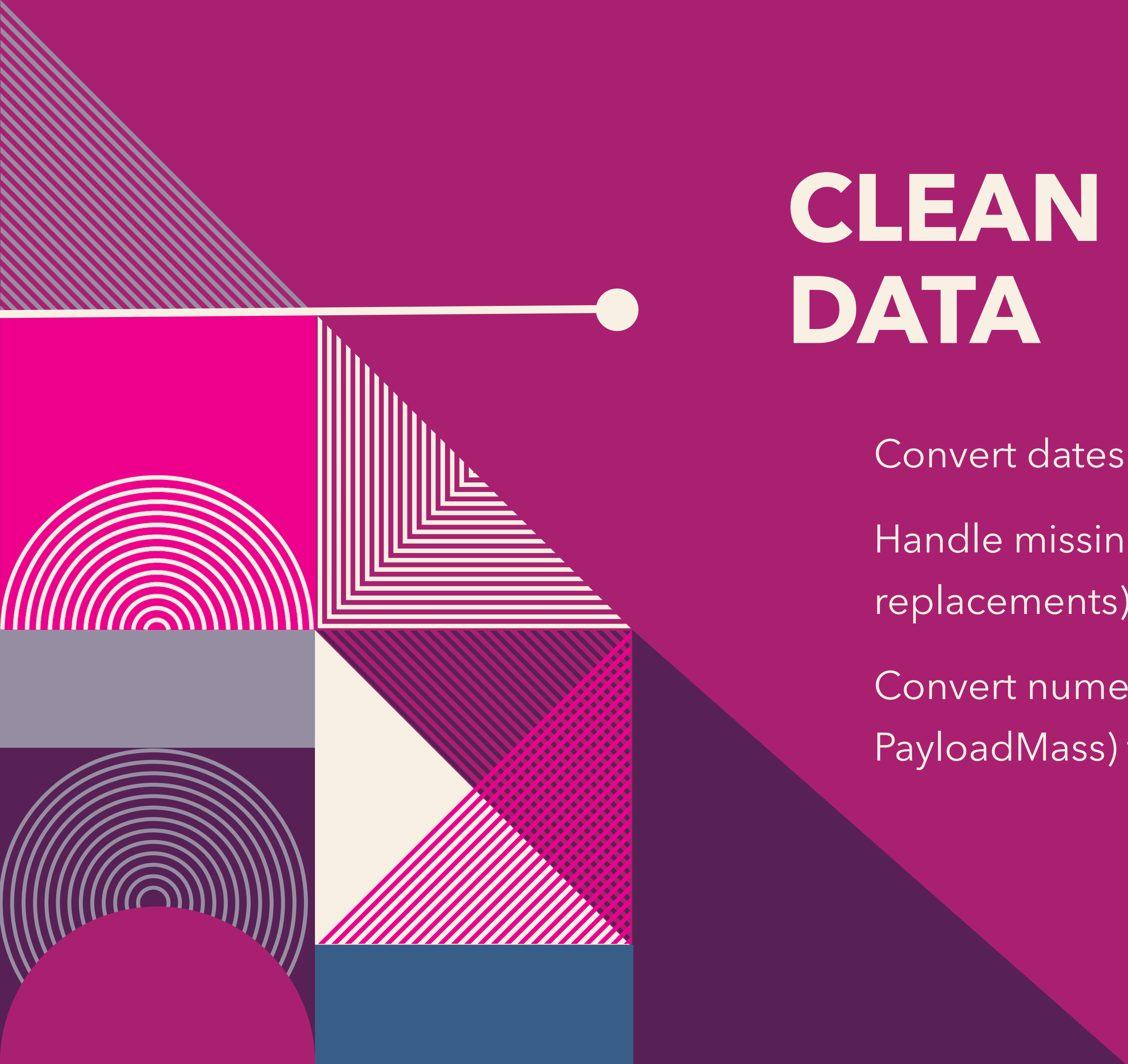
Identify relevant features: Date, Booster version, Payload mass, Orbit, Landing outcome.

Loop through table rows to collect structured values.



BUILD DATAFRAME

- Organize extracted data into a **pandas DataFrame**.
- Ensures tabular structure for analysis.



CLEAN & FORMAT DATA

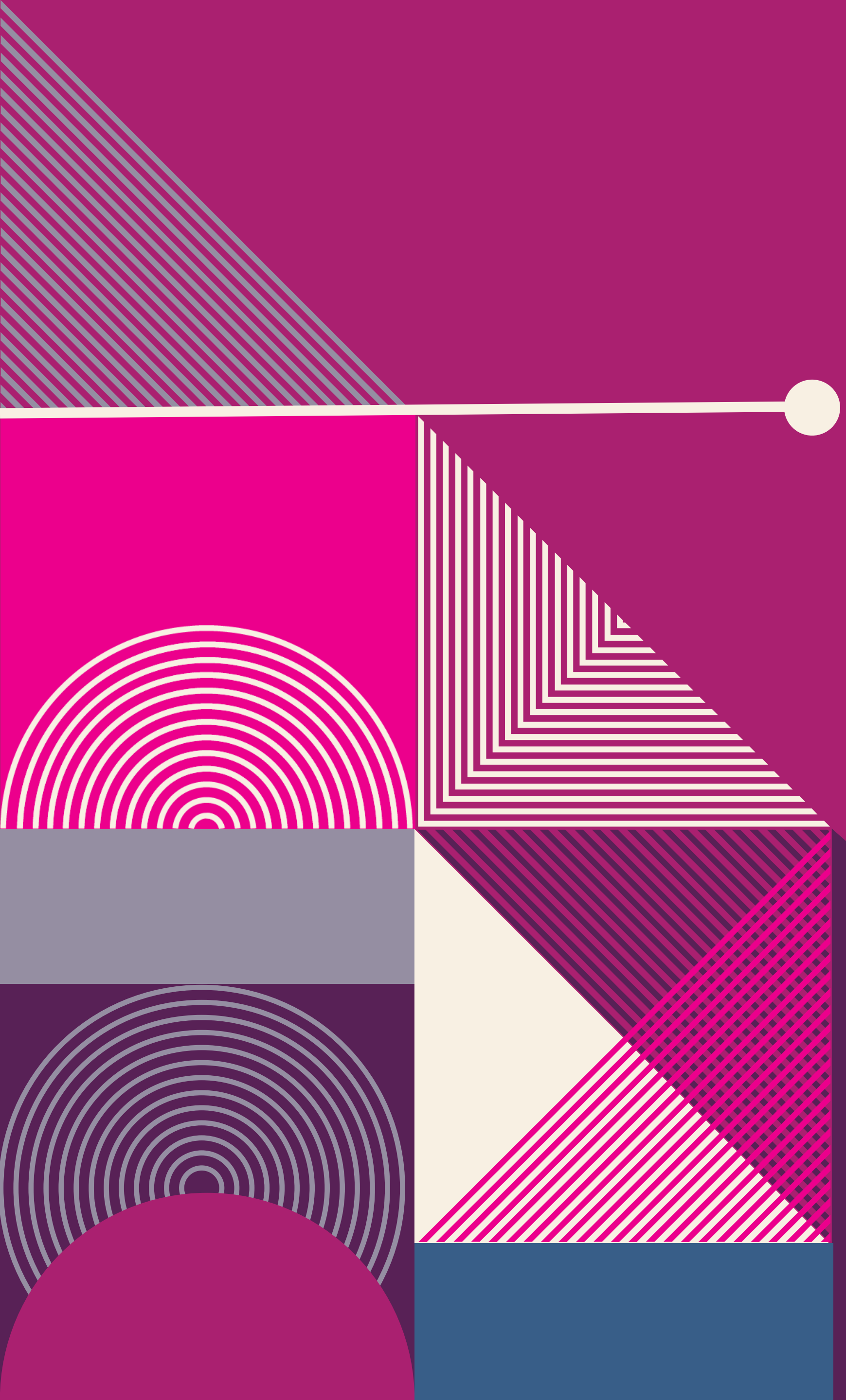
Convert dates → datetime format.

Handle missing values (NaN replacements).

Convert numeric columns (e.g., PayloadMass) to float.

SAVE DATASET (CSV)

We save csv files for every notebook wherever possible



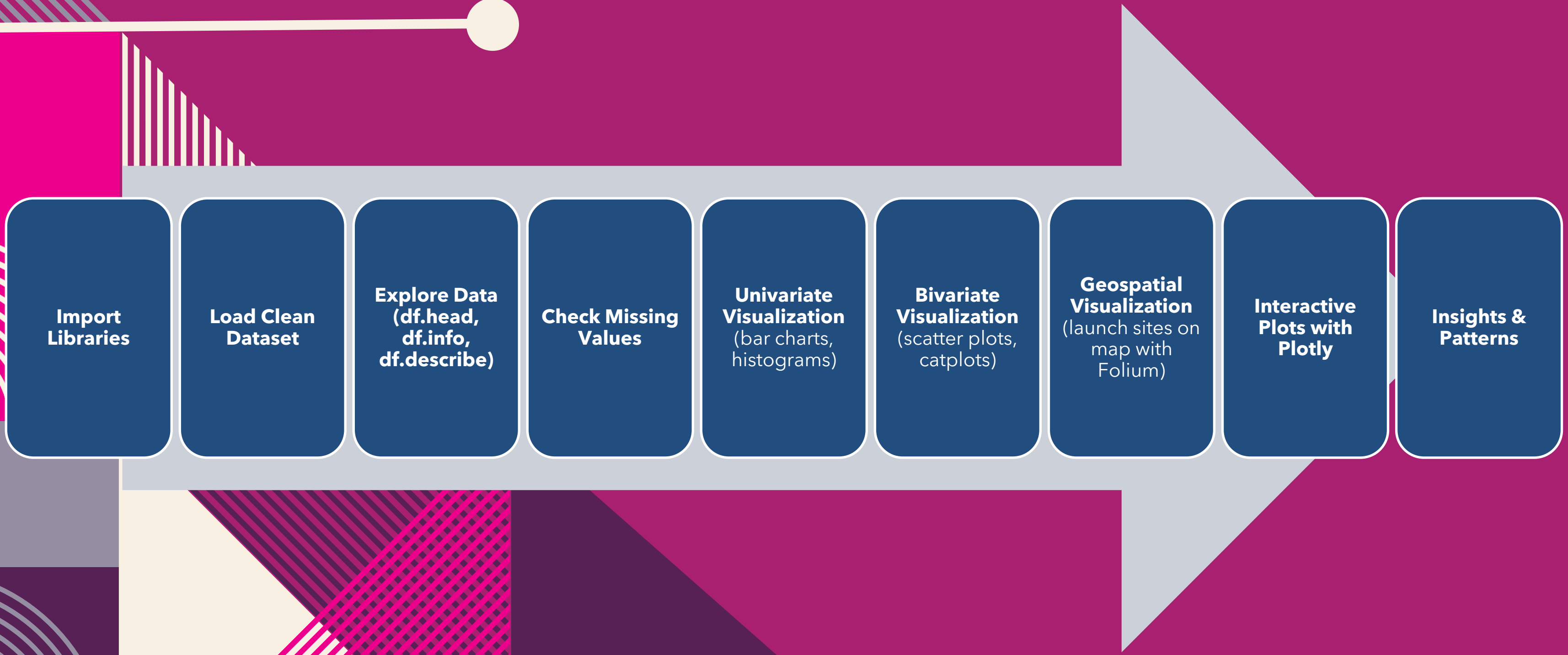
EXPLORATORY DATA ANALYSIS

Focused on Falcon9

Fixed the null values

Extracted info for each Orbit, Launch Site Payload

PROCESS - EDA



Import Libraries

Load Clean Dataset

Explore Data (df.head, df.info, df.describe)

Check Missing Values

Univariate Visualization (bar charts, histograms)

Bivariate Visualization (scatter plots, catplots)

Geospatial Visualization (launch sites on map with Folium)

Interactive Plots with Plotly

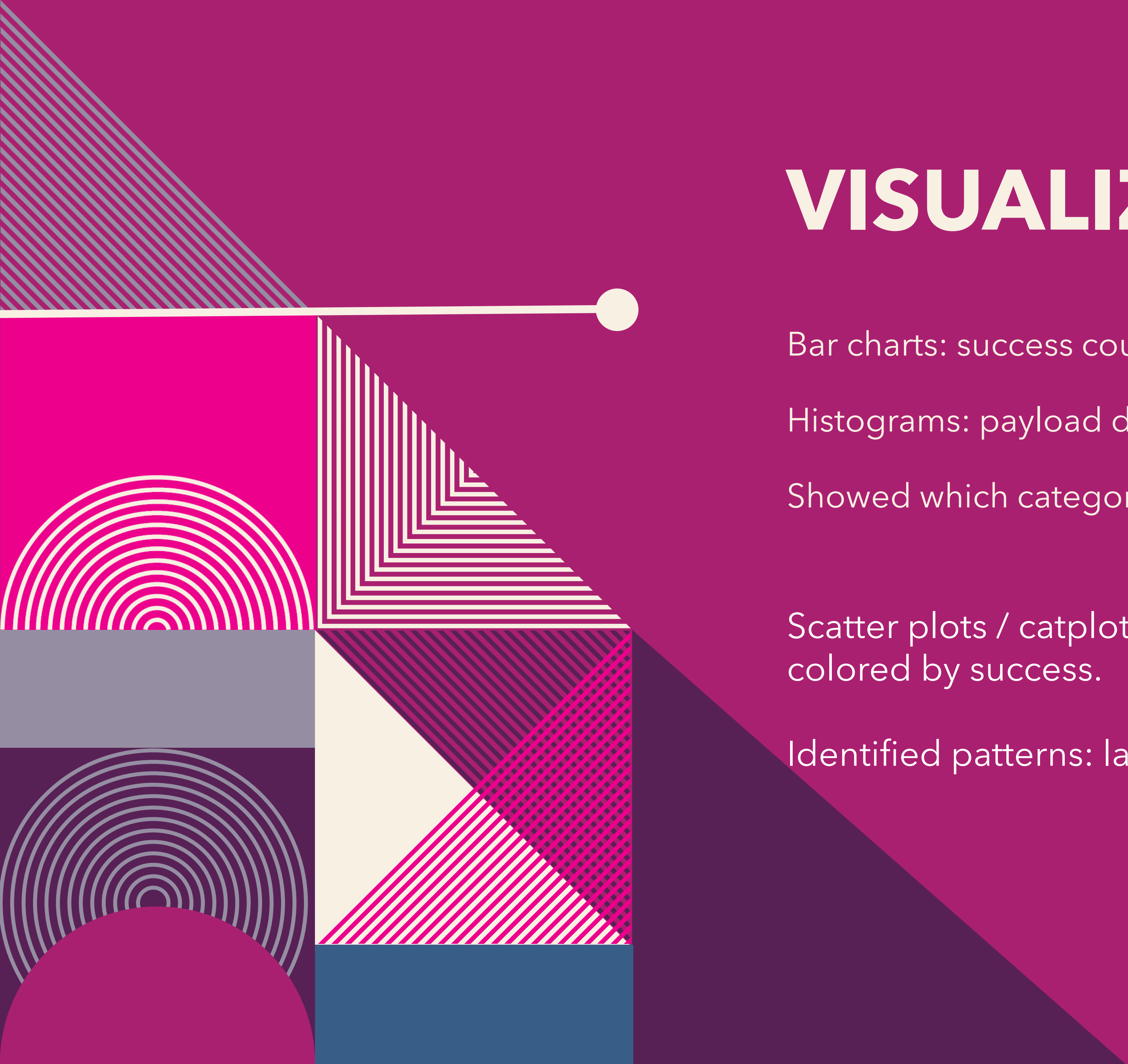
Insights & Patterns

EXPLORE DATA

Dataset from previous lab (dataset_part_2.csv).
Contains Falcon 9 launches with engineered features.

FIXING MISSING VALUES

Ensured completeness.
Only certain categorical columns



VISUALIZATIONS - EDA

Bar charts: success counts by orbit, launch site.

Histograms: payload distribution.

Showed which categories dominate.

Scatter plots / catplots: FlightNumber vs. PayloadMass colored by success.

Identified patterns: later flights → higher success

FOLIUM AND PLOTLY

Folium maps to display launch sites with markers.

Circle markers and clusters to explore success patterns geographically.

Dropdown menus to filter by orbit or launch site.

Scatter and line plots showing success by payload/flight.

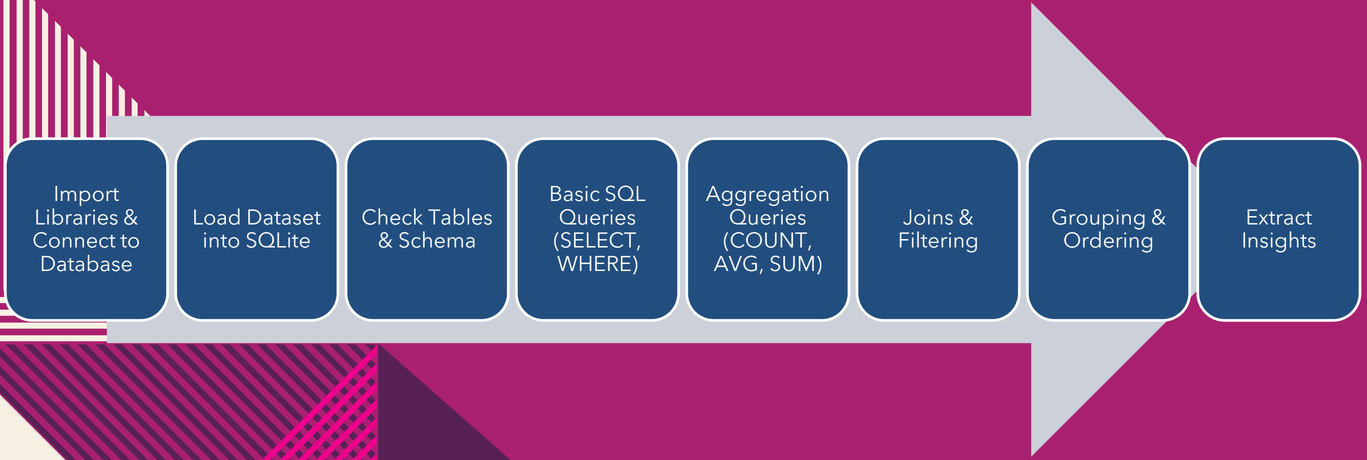
EDA - INSIGHTS

Success rate improved as FlightNumber increased.

Higher payloads correlated with different orbits (e.g., GTO riskier than LEO).

Certain launch sites performed more reliably.

EDA - SQL



Import Libraries & Connect to Database

Load Dataset into SQLite

Check Tables & Schema

Basic SQL Queries (SELECT, WHERE)

Aggregation Queries (COUNT, AVG, SUM)

Joins & Filtering

Grouping & Ordering

Extract Insights

Import Libraries & Connect to Database

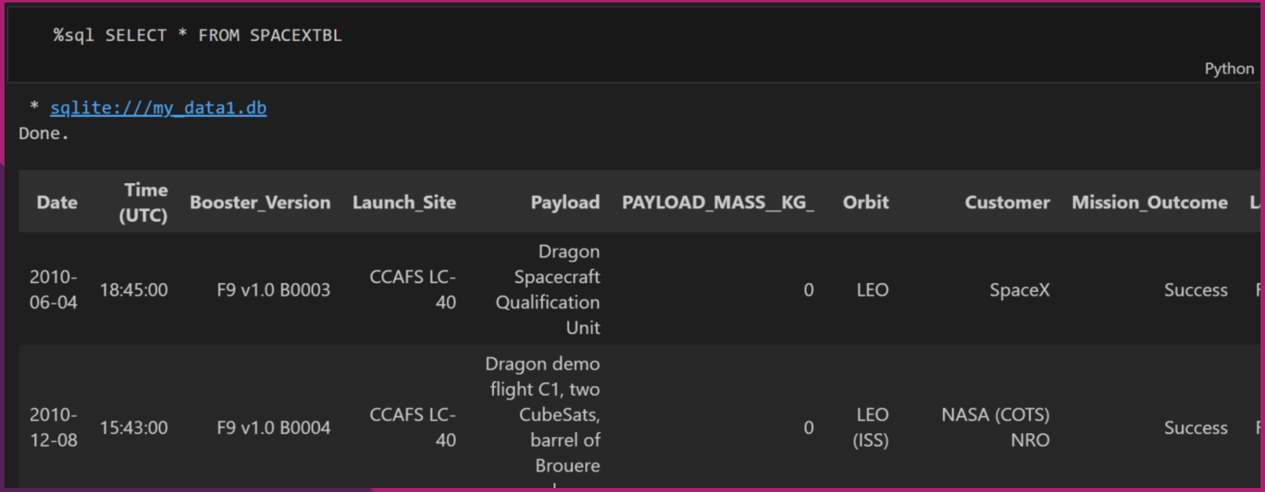
Used sqlite3 and pandas for database connection and query execution.

Created a cursor for running SQL commands.

Load Dataset into SQLite

Loaded cleaned Falcon 9 dataset (dataset_part_2.csv).

Stored in an SQLite database for query-based exploration.



```
%sql SELECT * FROM SPACEXTBL
```

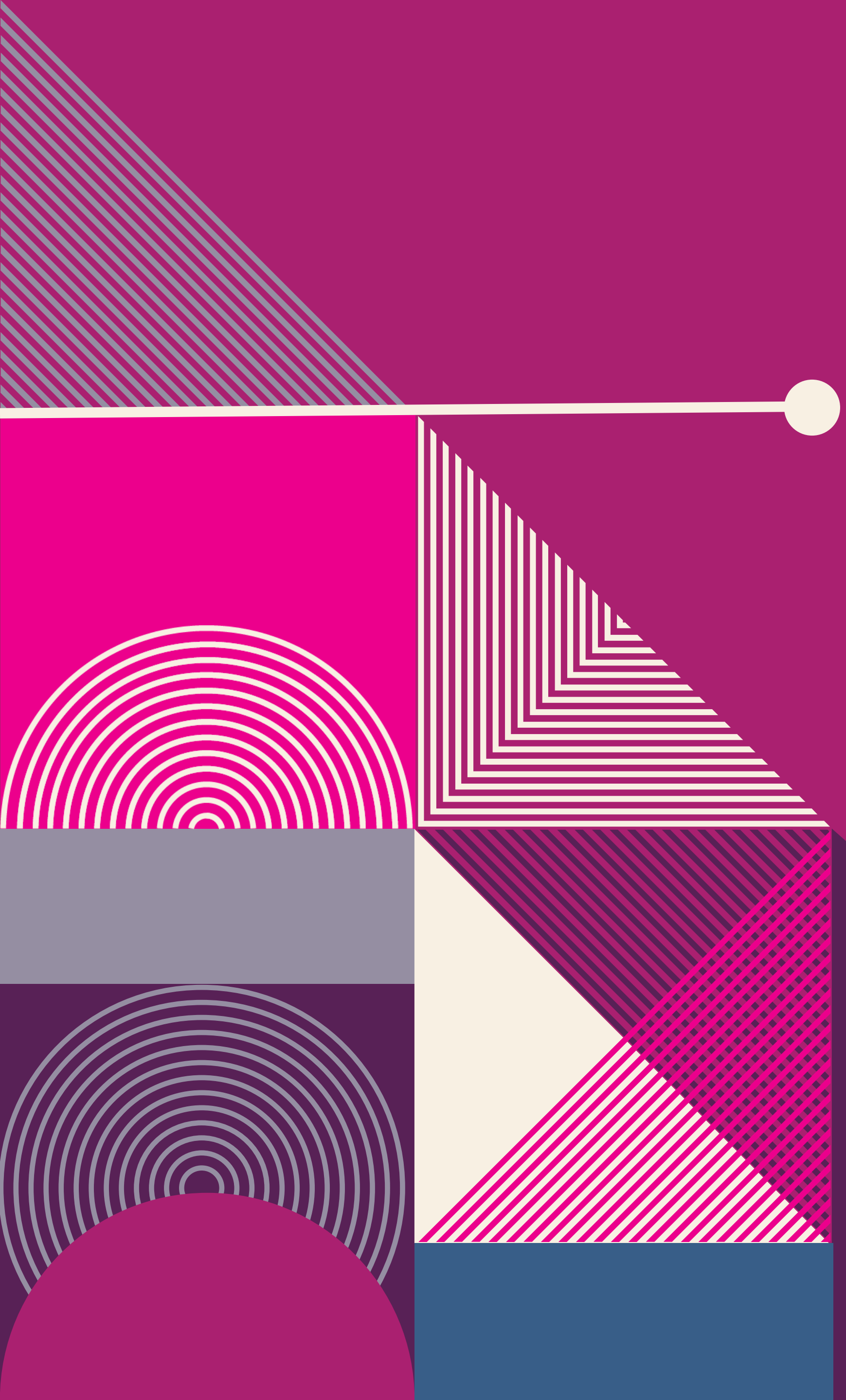
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	L
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere	0	LEO (ISS)	NASA (COTS) NRO	Success	F

Check Tables & Schema

Verified database structure with SELECT name FROM sqlite_master.

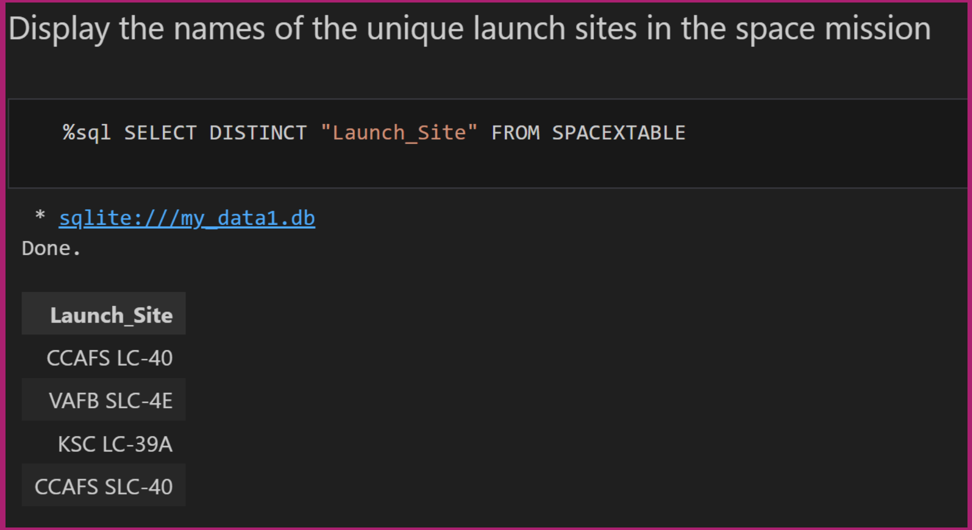
Checked table column names and data types.



Basic SQL Queries

Queried launch data (flight number, payload, orbit).

Applied WHERE conditions to filter by orbit or launch site.



Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

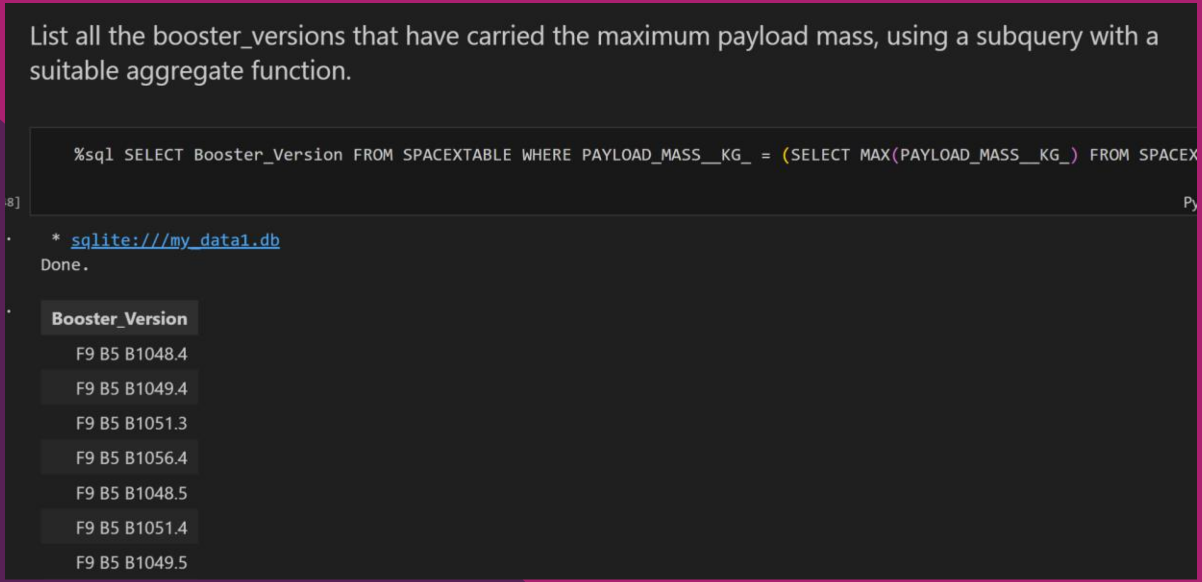
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Aggregation Queries

Used COUNT() for number of launches.

Used AVG() for mean payload mass.

Summarized launch outcomes.



List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5

Joins & Filtering

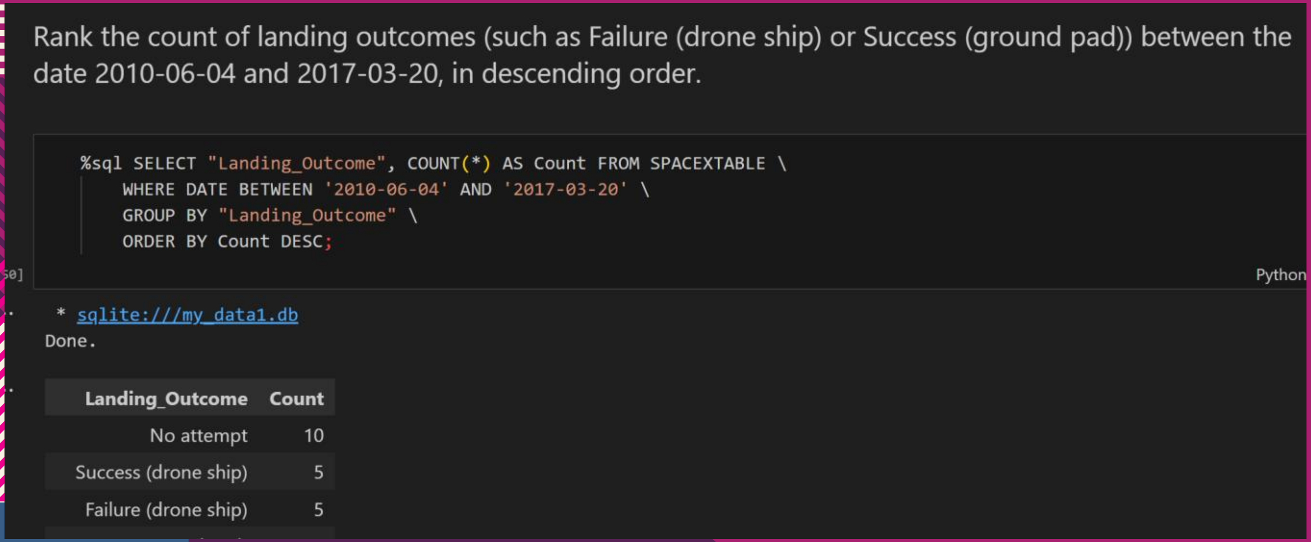
Joined tables (if split into payloads/launches).

Filtered data to answer specific business questions.

Grouping & Ordering

Grouped results by orbit, launch site, and booster version.

Ordered by payload mass, success rate, or flight number.



Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS Count FROM SPACEXTABLE \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY "Landing_Outcome" \
    ORDER BY Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5

SQL EDA - INSIGHTS

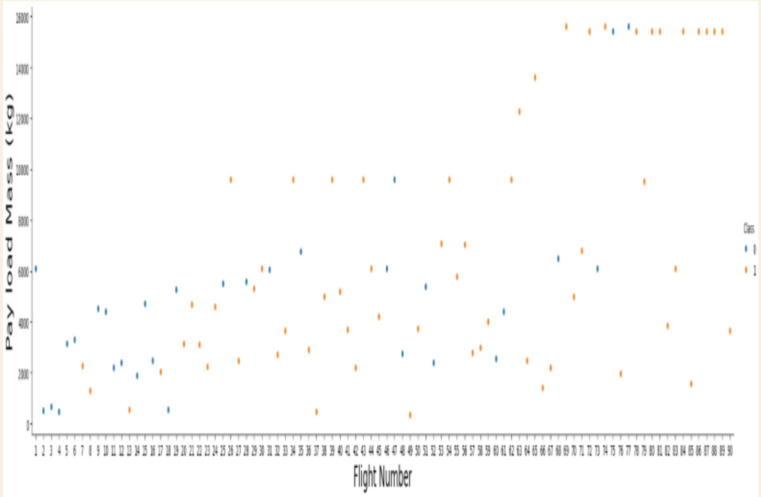
Found relationships between payload mass, orbit, and success rate.

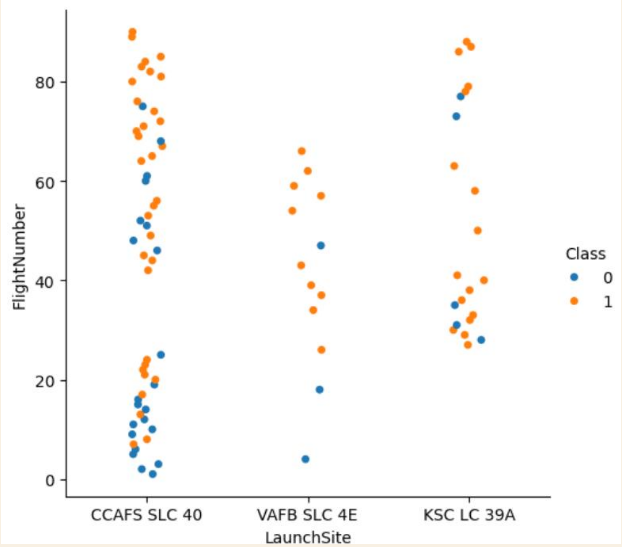
Identified which launch sites and booster versions were most reliable.

EDA: CHARTS

Scatter Plots were plotted to check the correlation of variables with each other:

- Flight Number
- Payload Mass
- Launch Site
- Orbit
- Success Rate, and Yearly trend for the success





EDA: WITH SQL

Explored further patterns within the data for payload (4000<payload<6000)

Successful missions and their relevant parameters

Ranking the Landing outcomes for drones and groundpad

INTERACTIVE FOLIUM MAP

- The Project also included a map in Folium, to interactively see the sites and their outcomes using Markers

INTERACTIVE DASHBOARD

- Interactive Dashboard for visualization for the factors that can be manually selected

PREDICTIVE ANALYSIS

- Machine Learning Models were trained with GridSearch CV, cv=10 and the models were:
- Logistic Regression
- Decision Tree
- SVM
- KNN

MODELS

- SVM and KNN performed better than others, generally.
- Decision Tree worked best for Validation Data only.

Metric	**Test Accuracy**	**Validation Accuracy**
Logistic Regression	0.833	0.8464
SVM	**0.833**	**0.848**
DecisionTree	0.722	0.9017
KNN	**0.833**	**0.848**

CONCLUSIONS

- Best Models were KNN and SVM
- Reused Boosters perform better
- Orbits like LEO, ISS have higher success rates than GTO
- Heavier payloads reduce success probability

SPEAKING ENGAGEMENT METRICS

Impact factor	Measurement	Target	Achieved
Audience interaction	Percentage (%)	85	88
Knowledge retention	Percentage (%)	75	80
Post-presentation surveys	Average rating	4.2	4.5
Referral rate	Percentage (%)	10	12

APPENDIX

Special Thanks to Coursera, Instructors, and Fellow-Peers.