IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Siphesihle Masuku
07 December 2024

Outline

- ❖Executive Summary
- ❖Introduction
- ❖Methodology
- ❖Results
- ❖Conclusion
- ❖Appendix

Executive Summary

Summary of methodologies

- Data collected via SpaceX API and web scraping.
- Data wrangling included cleaning, feature engineering
- Conducted EDA using visualizations to explore payload, launch sites, and outcomes.
- SQL queries analyzed payload metrics and success rates.
- Interactive Folium maps and Plotly Dash dashboard enabled dynamic analysis.
- Built machine learning models (Logistic Regression, SVM, Decision Trees) to predict landing success.

Summary of all results

- Exploratory Data Analysis Results
- Interactive Maps and Dashboard
- Predictive Analytics Results

Introduction

SpaceX aims to revolutionize space travel by making it cost-effective through reusable rocket boosters. Analyzing historical launch data can provide insights into factors influencing mission outcomes, such as payload, launch site, and booster performance.

The problems included:

- ❖What factors contribute most to a successful launch?
- ❖How does payload size/type affect outcomes?
- ❖Are certain launch sites more successful than others?
- ❖What trends exist in success rates over time?
- ❖What conditions improve booster reusability?

Section 1

Methodology

Methodology

Executive Summary

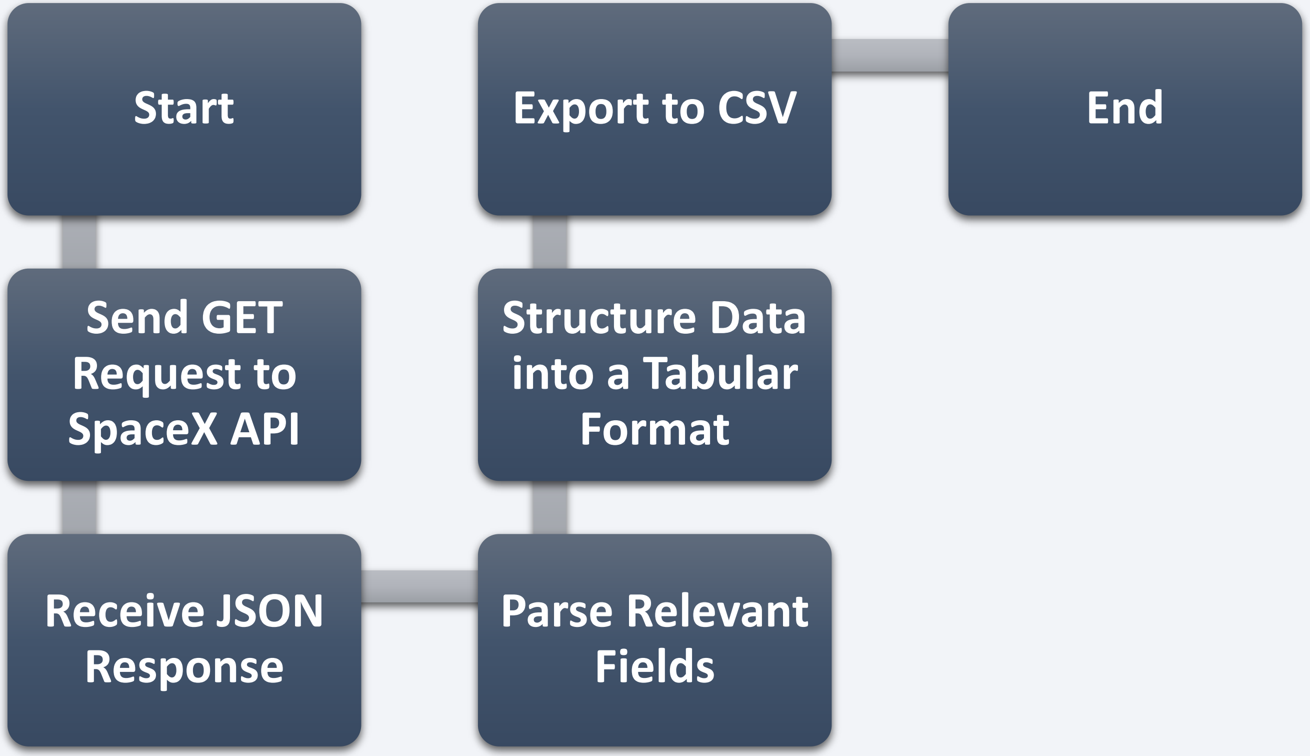
- ❖ Data collection methodology:
  - ❖ SpaceX REST API & Web Scrapping from Wikipedia (List of Falcon 9 and Falcon Heavy Launches)
- ❖ Perform data wrangling
  - ❖ The data was processed by cleaning missing values, removing duplicates, engineering features like landing_class, standardizing numerical columns, and merging data from multiple sources into a unified dataset for analysis.
- ❖ Perform exploratory data analysis (EDA) using visualization and SQL
- ❖ Perform interactive visual analytics using Folium and Plotly Dash
- ❖ Perform predictive analysis using classification models
  - ❖ Classification models were built using Logistic Regression, SVM, Decision Trees, and KNN, tuned with GridSearchCV for optimal hyperparameters, and evaluated using metrics like accuracy, F1-score, and confusion matrices to identify the best-performing model.

Data Collection

1. SpaceX REST API: Data on rocket launches, payloads, and outcomes was collected programmatically using SpaceX's API.
2. Web Scraping: Additional data on booster specifications and mission details was extracted from SpaceX's website using Python libraries like BeautifulSoup.
3. Data Integration: Combined API and web-scraped data into a single dataset for analysis, ensuring completeness and consistency.

Data Collection – SpaceX API

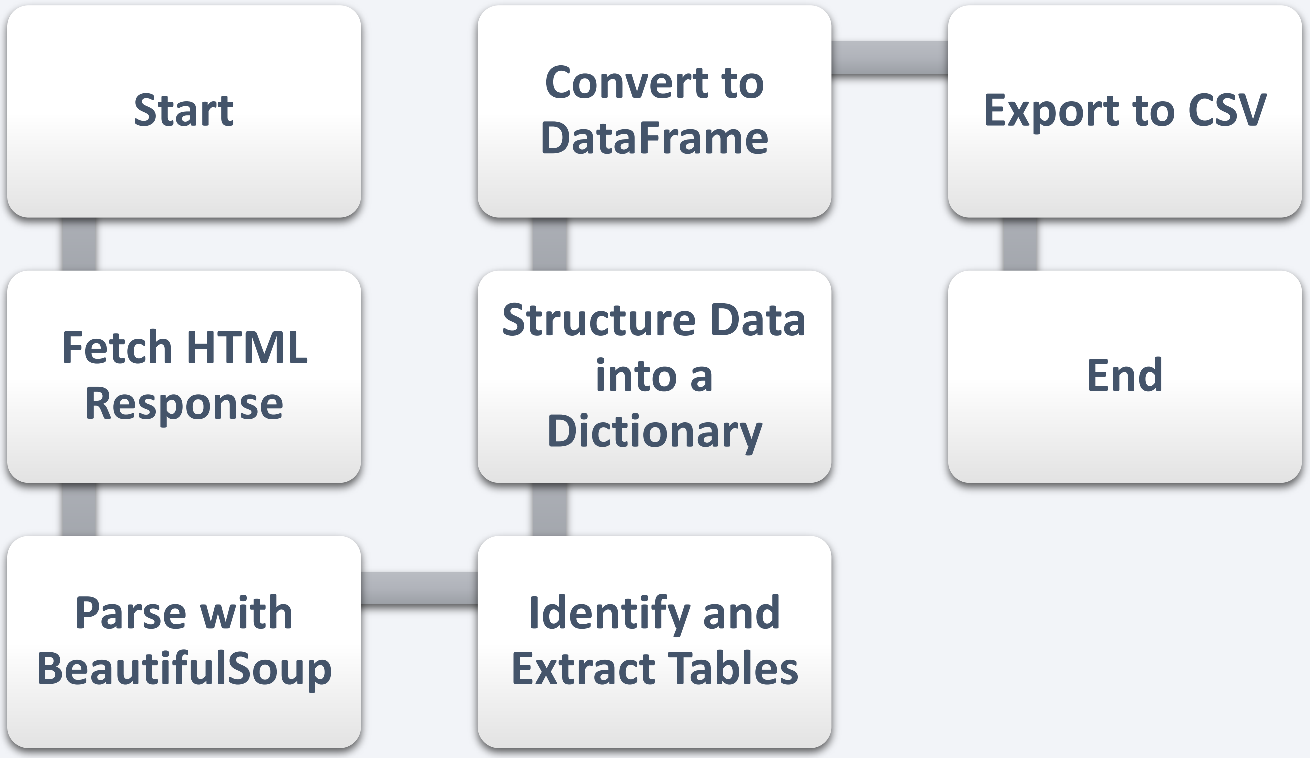
1. API Endpoint:
   - ❖ Used SpaceX's public REST API endpoint to retrieve launch data, including payloads, launch sites, and mission outcomes.
2. Request Handling:
   - ❖ Sent HTTP GET requests to the API endpoint using Python's requests library.
3. Data Parsing:
   - ❖ Extracted JSON responses and parsed relevant fields like flight_number, payload_mass, launch_site, and landing_success.
4. Data Cleaning:
   - ❖ Structured the extracted data into a tabular format for analysis.
5. Exporting:
   - ❖ Converted the cleaned data into a Pandas DataFrame and exported it to a CSV file for downstream tasks.



SpaceX API Code Link

Data Collection - Scraping

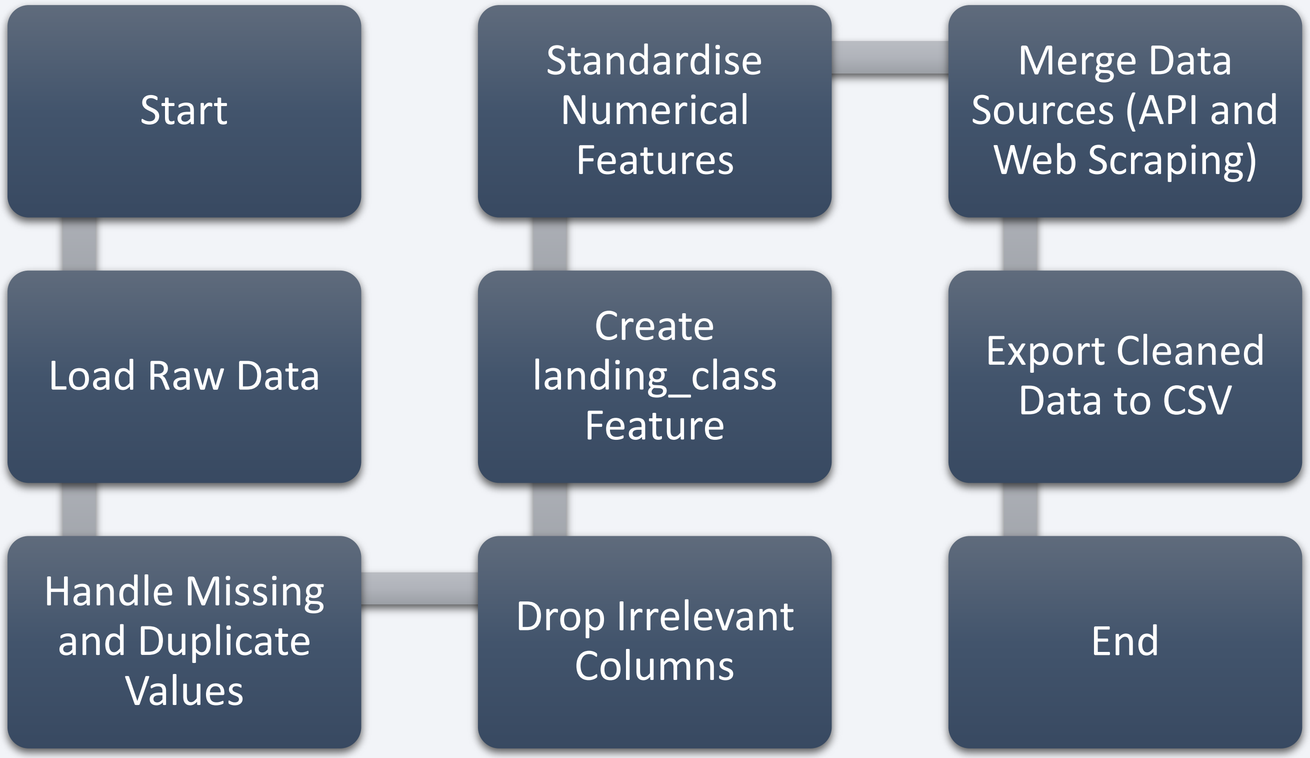
1. HTML Response:
   - ❖ Used requests.get() to fetch the webpage content containing launch data.
2. Parse HTML:
   - ❖ Created a BeautifulSoup object to parse the HTML structure for extracting tables.
3. Extract Data:
   - ❖ Located tables and column names, iterated over rows to retrieve relevant launch data.
4. Store Data:
   - ❖ Created a dictionary to structure the extracted data into columns like Launch Site, Payload Mass, and Mission Outcome.
5. Export:
   - ❖ Converted the dictionary into a Pandas DataFrame and exported it to a CSV file for further analysis.



[Scrapping Code Link]

Data Wrangling

1. Data Cleaning:
   - ❖ Removed duplicates and missing values.
   - ❖ Dropped irrelevant columns and formatted data types.
2. Feature Engineering:
   - ❖ Created a new "landing_class" feature.
   - ❖ Extracted and transformed key fields.
3. Standardization:
   - ❖ Normalized numerical fields (e.g., Payload Mass).
4. Integration:
   - ❖ Merged API and web-scraped data into a unified dataset.
5. Export:
   - ❖ Saved processed data as CSV files for analysis and modeling.



[Data Wrangling Code Link]

EDA with Data Visualization

1. **Flight Number vs. Payload Mass Scatter Plot:**
   - ❖ Demonstrates that success rates improve with higher flight numbers due to operational experience, while heavier payloads negatively impact landing success.
2. **Flight Number vs. Launch Site Scatter Plot:**
   - ❖ Highlights the distribution of launches across sites, showing that CCAFS SLC 40 and KSC LC 39A host more frequent launches, while VAFB SLC 4E maintains balanced success rates.
3. **Categorical Plot (Flight Number vs Launch Site with Outcome):**
   - ❖ Examines success rates at different launch sites, showcasing variations in site performance and trends.

EDA - Dataviz Code Link

EDA with SQL

❖Created a new table: Created a table (SPACEXTABLE) by filtering non-null records from the dataset.

❖Query 1: Retrieved distinct launch site names to identify unique locations of SpaceX launches.

❖Query 2: Filtered launch site names beginning with CCA to explore launches from specific facilities.

❖Query 3: Calculated the total payload mass carried by NASA missions to understand their contribution.

❖Query 4: Computed the average payload mass for missions using the F9 v1.1 booster version to analyze booster-specific performance.

EDA - SQL Code Link

Build an Interactive Map with Folium

1. Circles:
   - ❖ Placed around each launch site to indicate their geographical location with a visual radius of influence.
   - ❖ Purpose: To highlight the region surrounding the launch site, aiding in proximity analysis for potential impacts of geography on launch outcomes.
2. Markers:
   - ❖ Added at the exact coordinates of each launch site with labels displaying the site's name.
   - ❖ Purpose: To provide a clear and interactive reference for identifying specific launch sites.
3. Cluster Markers:
   - ❖ Used for grouping nearby launch events and displaying success or failure outcomes.
   - ❖ Purpose: To efficiently visualize clustered launch events, especially in areas with frequent launches.
4. Popup Labels:
   - ❖ Integrated with the circle objects to display additional information about the launch site upon clicking.
   - ❖ Purpose: To offer interactive details about launch sites, such as the site's name and significance.
5. Lines (for Proximity Analysis):
   - ❖ Lines connecting a launch site to nearby infrastructures such as railways, highways, or coastlines.
   - ❖ Purpose: To calculate and visualise the distances between the launch site and critical infrastructure, analysing logistical advantages or constraints.

Launch Site Location Code Link

Build a Dashboard with Plotly Dash

1. Pie Chart: Visualizes total successful launches for all sites or selected sites (success vs failure).
   - Purpose: Summarizes overall and site-specific performance.
2. Scatter Plot: Shows correlation between payload mass and launch success, colour-coded by booster version.
   - Purpose: Analyses payload impact and booster trends.
3. Interactive Features:
   - Dropdown: Filters data by site or shows all sites.
   - Payload Slider: Filters payload range for detailed analysis.

These features enable dynamic exploration of launch data and trends.

SpaceX Dash App Code Link

Predictive Analysis (Classification)

1. Data Preparation:
   - ❖ Standardized data using StandardScaler.
   - ❖ Created training and test datasets with train_test_split.
2. Model Selection:
   - ❖ Used three classification algorithms: Logistic Regression, SVM, and Decision Trees.
   - ❖ Applied GridSearchCV to tune hyperparameters for each model.
3. Evaluation:
   - ❖ Models were evaluated using metrics like accuracy and confusion matrices.
   - ❖ Comparison focused on identifying the best performing model using test data.
4. Findings:
   - ❖ All models performed similarly, highlighting balanced predictive power across methods.

Data Cleaning & Standardisation:

Ensured consistency and normalized features.

Data Splitting:

Training (80%) vs Testing (20%).

Model Training:

Logistic Regression, SVM, and Decision Trees with hyperparameter tuning via GridSearchCV.

Model Evaluation:

Performance metrics compared using test datasets and confusion matrix visualization.

[SpaceX Machine Learning Prediction Code Link]

Results

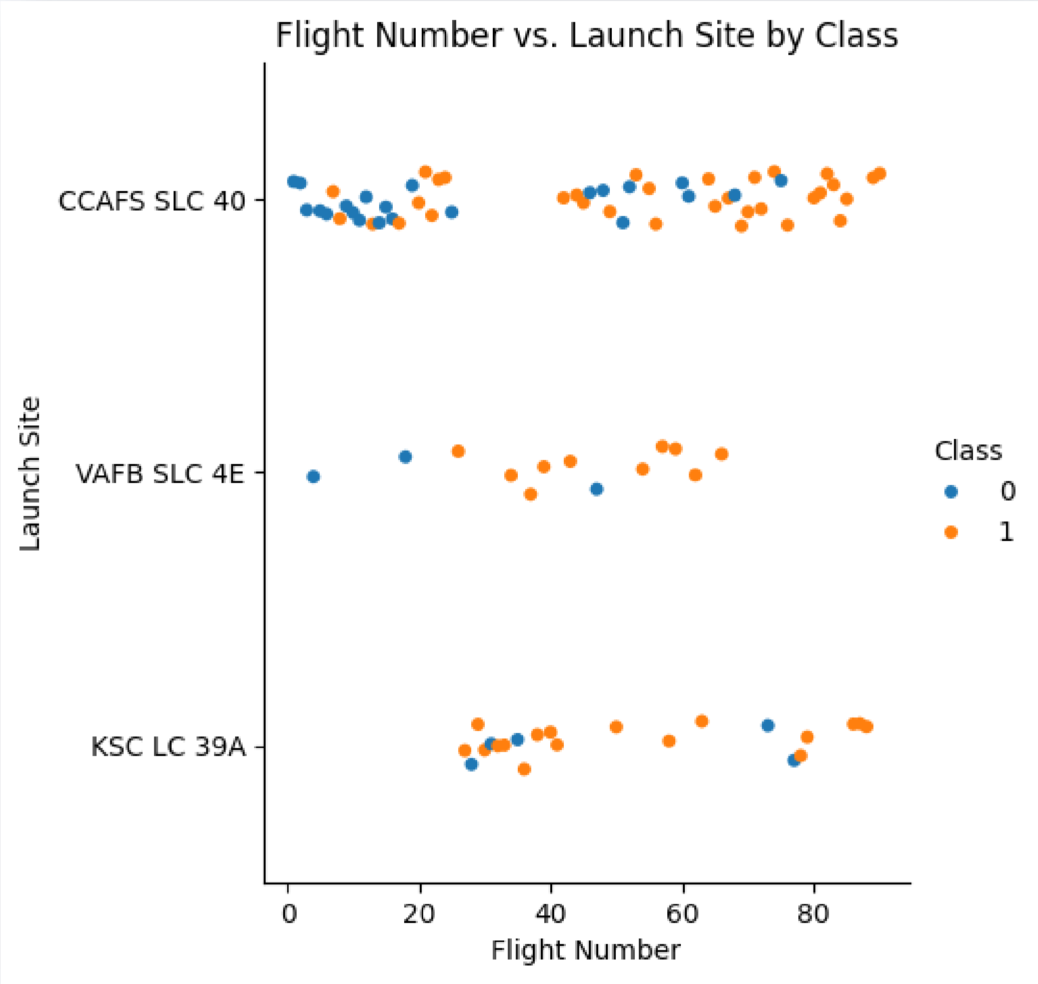
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

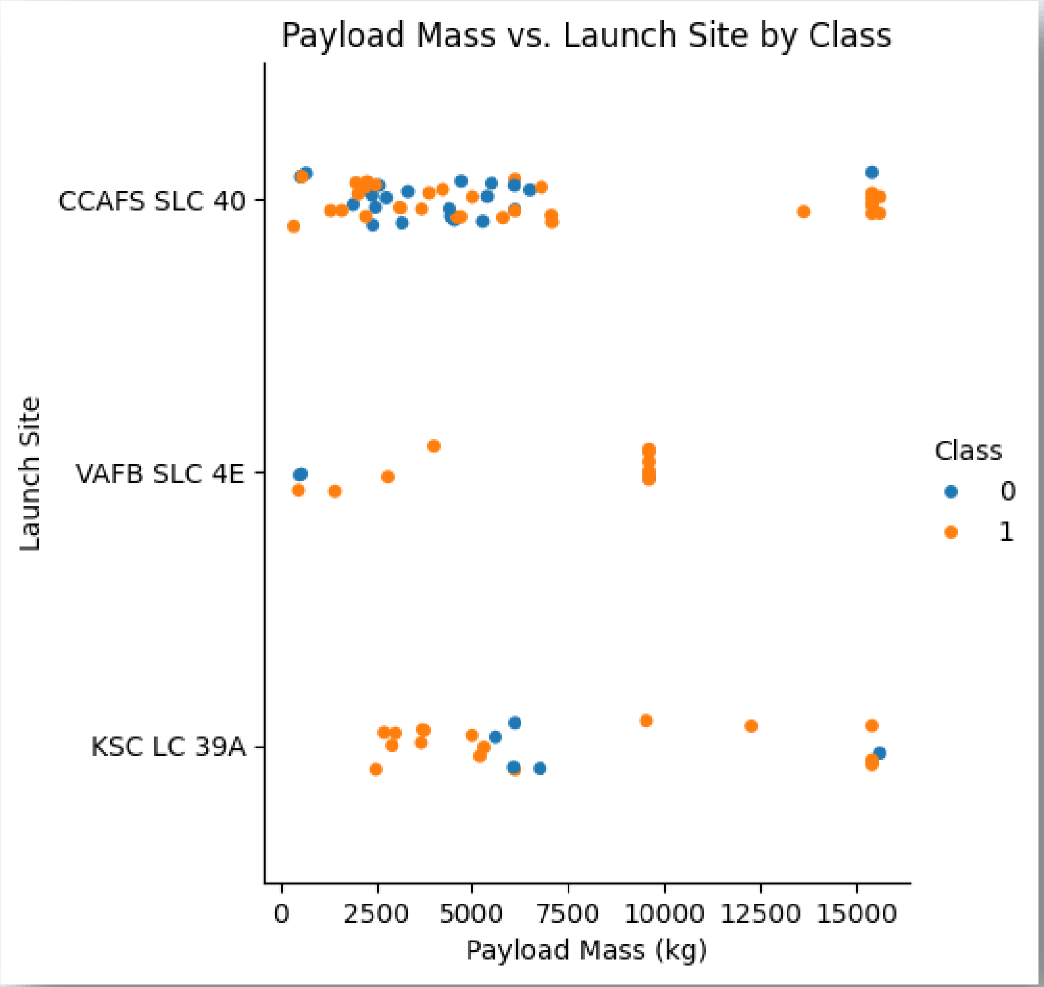
Flight Number vs. Launch Site

- Flight numbers increase over time.
- CCAFS SLC 40 and KSC LC 39A have more launches, especially towards higher flight numbers.
- VAFB SLC 4E has fewer launches but a more balanced success rate.



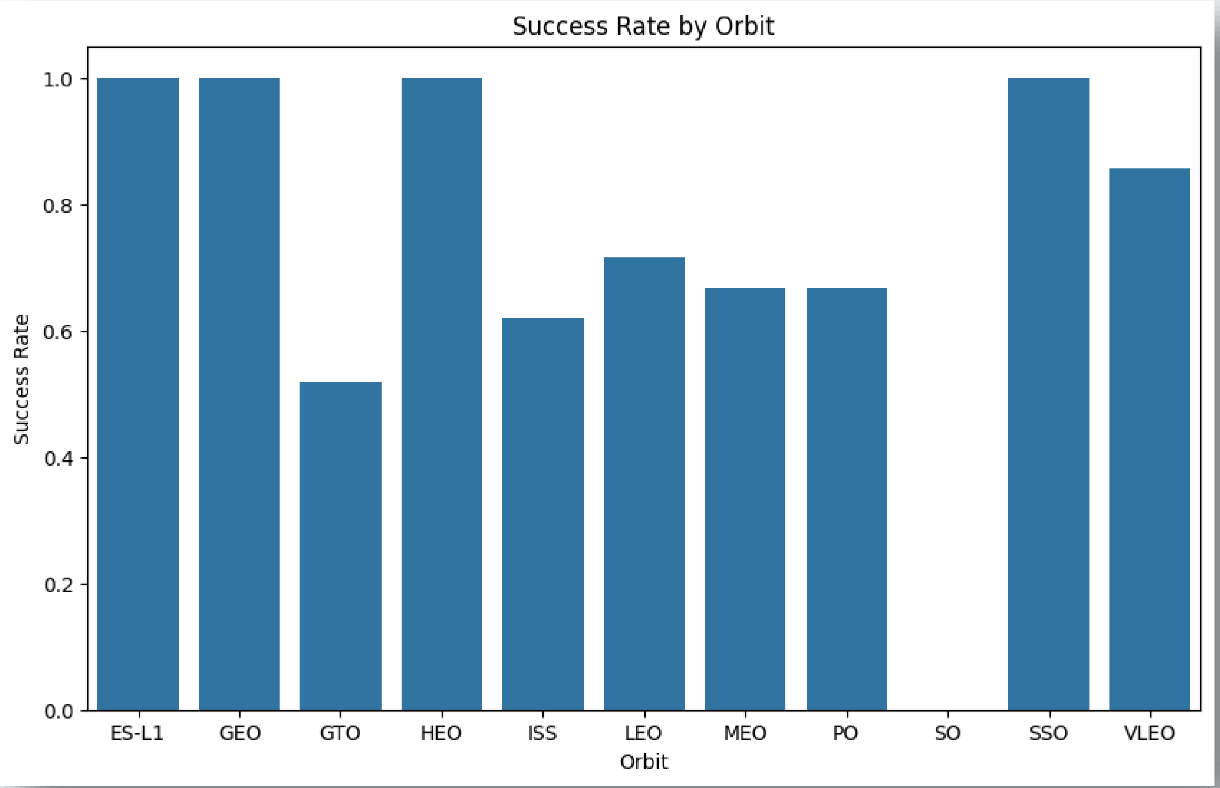
Payload vs. Launch Site

- Payload capacity varies by site, possibly influenced by site-specific infrastructure or mission objectives.
- Successful launches are observed across all payload ranges, but the outcomes differ by site and payload size, hinting at optimized matching of payload types to launch locations.



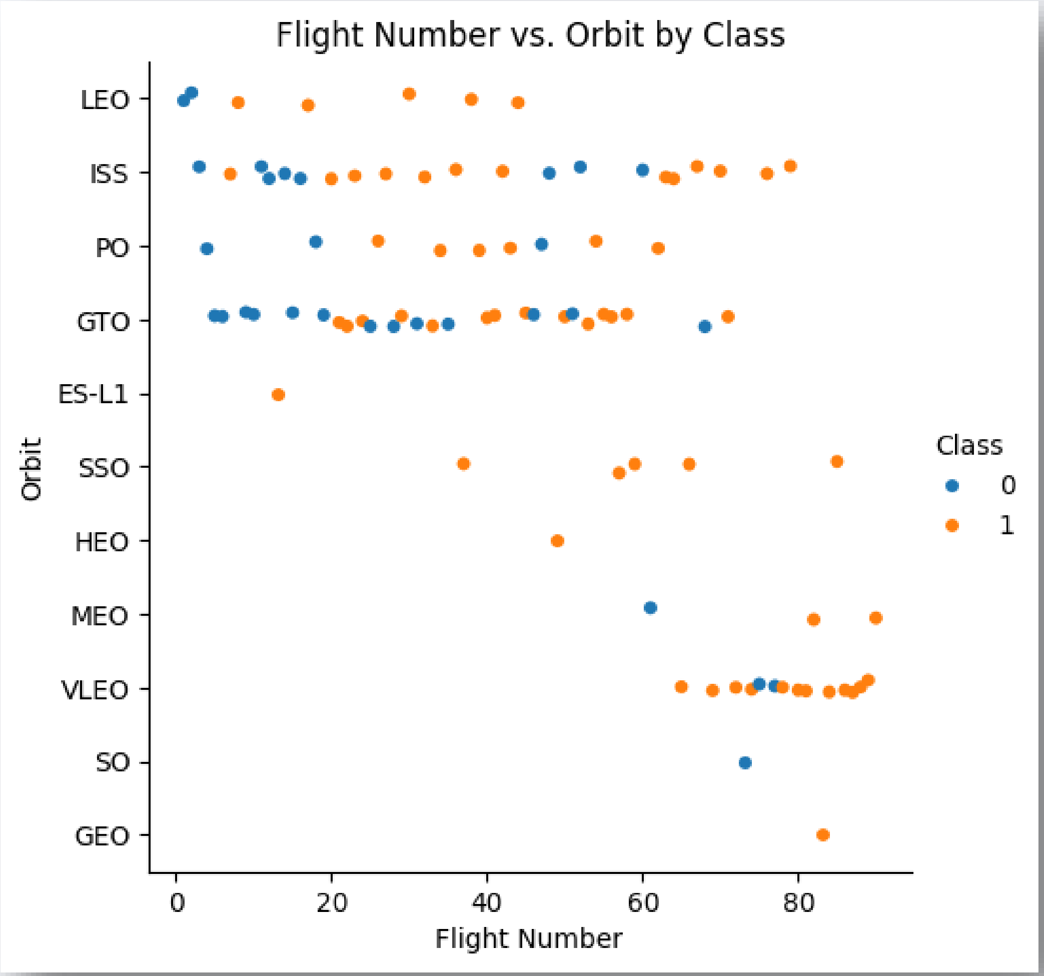
Success Rate vs. Orbit Type

- Orbits like ES-L1, GEO, and HEO have a 100% success rate, indicating reliable missions likely due to specialized planning or mission-critical payloads.
- GTO has the lowest success rate, reflecting challenges associated with reaching higher altitudes or transitioning to specific geostationary positions.
- LEO (Low Earth Orbit) and VLEO (Very Low Earth Orbit) show decent success rates, often used for satellites and space station resupply missions.



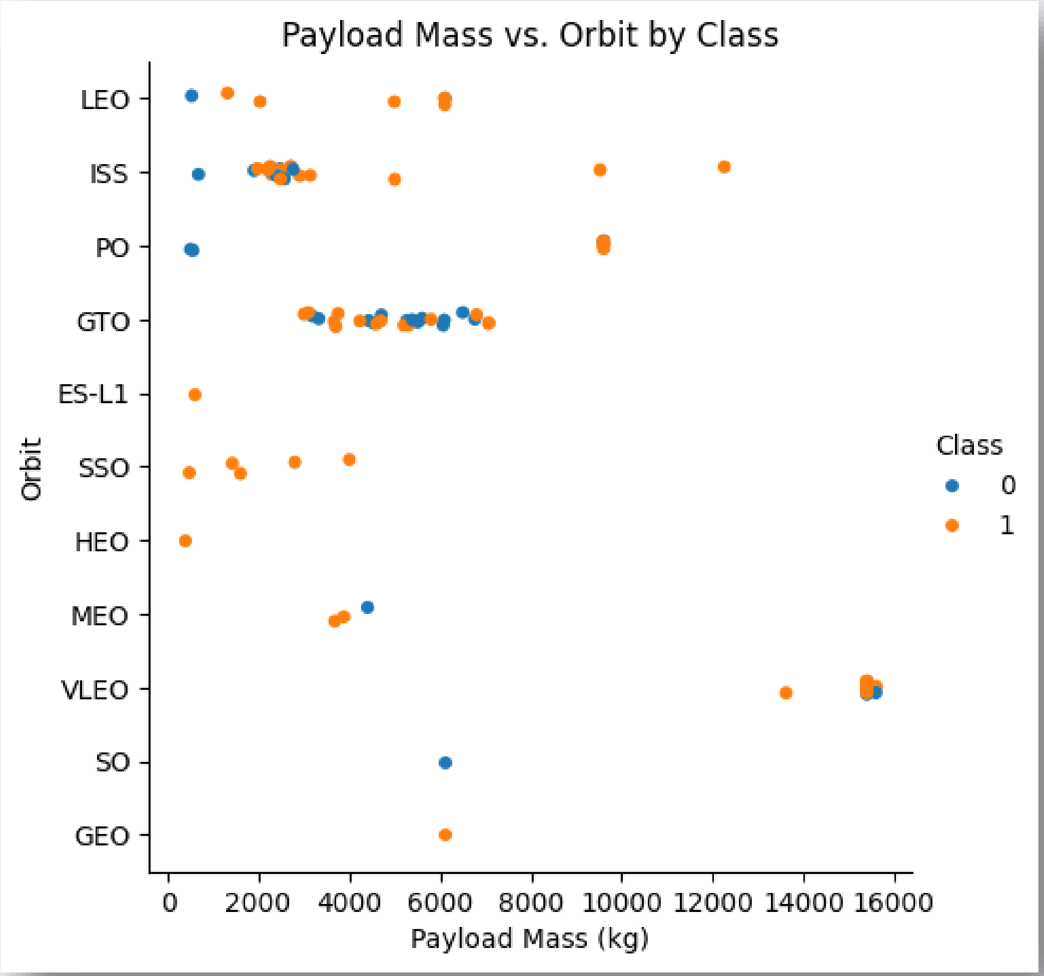
Flight Number vs. Orbit Type

- Higher flight numbers (indicating more recent launches) show an increased proportion of successful launches (Class 1), reflecting learning and operational improvements.
- Low Earth Orbits (LEO and VLEO) see more frequent launches, potentially due to their suitability for satellite deployment and reusability testing. Success rates improve with increased operational experience.
- Higher complexity orbits such as GTO and GEO exhibit more challenges, with fewer flights and a mixed success rate.



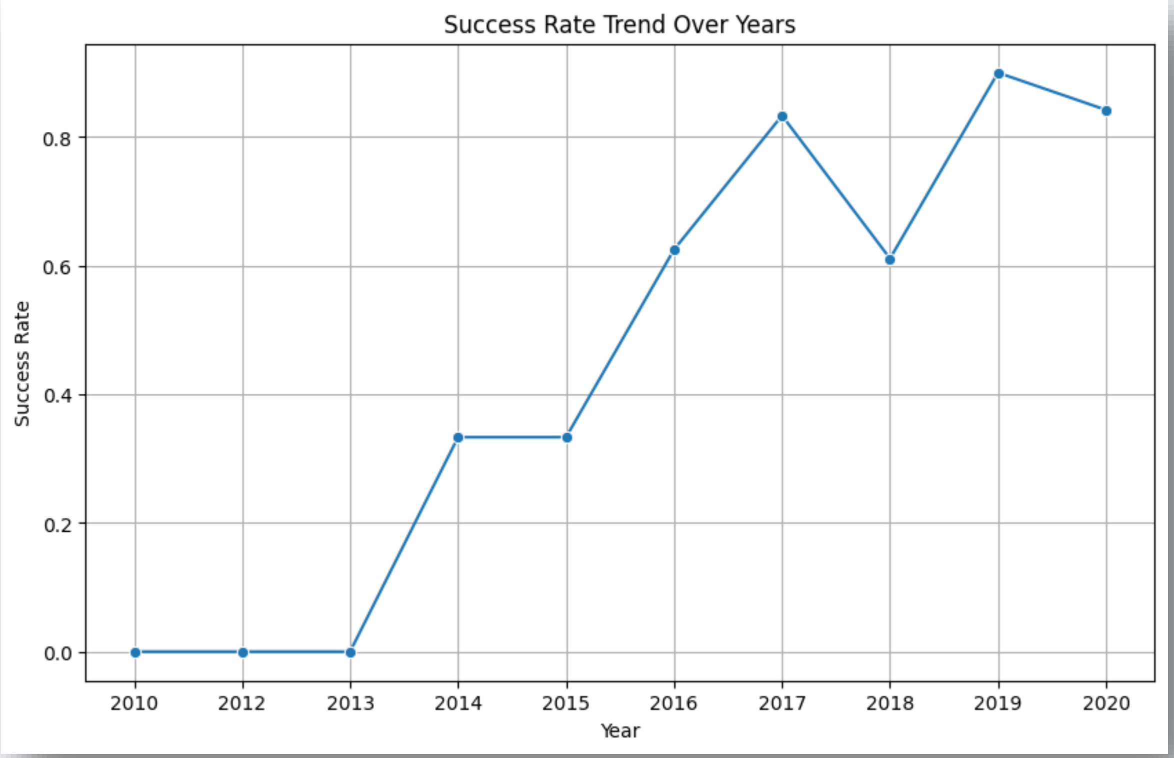
Payload vs. Orbit Type

- ❖ As payload mass increases, particularly beyond 10,000 kg, the likelihood of successful launches (Class 1) decreases, reflecting operational and technical challenges.
- ❖ Higher orbits like GTO and GEO typically involve mid-range payloads, suggesting limitations due to energy requirements for reaching these altitudes.
- ❖ Successful launches (Class 1) are observed across a wide payload range, especially for LEO and VLEO, demonstrating flexibility and reliability.



Launch Success Yearly Trend

- A sharp increase after 2014 indicates pivotal breakthroughs, likely influenced by iterative testing and learning.
- Post-2016, the success rate stabilizes above 60%, showing consistent operational reliability with occasional dips due to challenges in complex missions or payloads.
- The dip in 2018 may indicate issues related to high-profile or experimental launches, but rapid recovery by 2019 shows resilience and corrective measures.



All Launch Site Names

The query identified the following unique launch sites:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

Explanation: This result highlights the four distinct launch sites used by SpaceX.

Launch Site Names Begin with 'CCA'

The query retrieved the following records.

Explanation: This query filters records for launches conducted at CCAFS and limits the result to the first five entries.

DateTime (UTC)	Booster Version	Launch Site	Payload	Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04 18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08 15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22 07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08 08:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01 15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The query returned the following result:

Total Payload Mass: 45,596 kg

Explanation: This query calculates the total payload mass carried by SpaceX on behalf of NASA (CRS) (Commercial Resupply Services). It highlights the significant contribution of SpaceX in transporting cargo to the International Space Station (ISS) under the CRS programme, showcasing their reliability for government missions.

Average Payload Mass by F9 v1.1

The query returned the following result:

Average Payload Mass: 2,928.4 kg

Explanation: This query calculates the average payload mass for launches using the F9 v1.1 booster version. The result reflects the typical payload capacity handled by this booster, showcasing its operational range and efficiency for medium-weight payloads. This information can be used to evaluate the performance and utility of the F9 v1.1 for various mission types.

First Successful Ground Landing Date

The query returned the following result:

Date of First Successful Ground Pad Landing: 2015-12-22

Explanation: This query identifies the earliest date when SpaceX successfully landed a rocket's first stage on a ground pad. The achievement on 22nd December 2015 marks a significant milestone in SpaceX's efforts toward reusable rocket technology, setting the foundation for cost-effective space exploration.

Successful Drone Ship Landing with Payload between 4000 and 6000

The query returned the following booster versions:

1. F9 FT B1032.1
2. F9 B4 B1040.1
3. F9 B4 B1043.1

Explanation: These boosters achieved successful landings on a ground pad while carrying payloads between 4,000 kg and 6,000 kg. This result showcases the capability of these specific boosters to handle medium-heavy payloads with high precision, demonstrating their operational reliability in achieving mission objectives.

Total Number of Successful and Failure Mission Outcomes

The query returned the following results for SpaceX mission outcomes:

Total Missions: 101

Successful Missions: 23

Failed Missions: 78

Explanation: This query categorizes SpaceX missions based on landing outcomes. The high number of failed missions in the dataset reflects the early challenges faced in achieving reusable rocket landings. However, the 23 successful landings highlight significant progress and technological advancements over time.

Boosters Carried Maximum Payload

The query returned the following booster versions that carried the maximum payload mass:

1. F9 B5 B1048.4
2. F9 B5 B1049.4
3. F9 B5 B1051.3
4. F9 B5 B1056.4
5. F9 B5 B1048.5
6. F9 B5 B1051.4
7. F9 B5 B1049.5
8. F9 B5 B1060.2
9. F9 B5 B1058.3
10. F9 B5 B1051.6
11. F9 B5 B1060.3
12. F9 B5 B1049.7

Explanation: The repeated appearance of specific boosters (e.g., F9 B5 series) highlights their advanced capacity and reliability for heavy payload missions. This demonstrates SpaceX's focus on optimizing booster performance for high-stakes launches.

2015 Launch Records

The query returned the following records for failed drone ship landings in 2015:

Month	Landing Outcome	Booster Version	Launch Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation: This query filters records for 2015, specifically focusing on failed landings on drone ships. It also extracts the month, booster versions, and launch site details. The results indicate two failed attempts in January and April 2015, both using the F9 v1.1 booster at the CCAFS LC-40 site. These early failures reflect the challenges faced during the developmental phase of reusable rocket technology.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

The query ranked the count of landing outcomes between 2010-06-04 and 2017-03-20 as follows:

Landing Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3

Explanation:

- ❖ "No attempt" is the most frequent outcome during this period, reflecting missions where landing was not attempted, possibly due to technological or mission constraints.
- ❖ "Success (drone ship)" and "Failure (drone ship)" share the second-highest counts, demonstrating early experimentation with drone ship landings.
- ❖ "Success (ground pad)" occurred less often, showing the gradual adoption of ground pad landings.
- ❖ The lower counts of outcomes like "Controlled (ocean)" and "Uncontrolled (ocean)" indicate they were fallback options during early recovery attempts.

Section 3

Launch Sites Proximities Analysis

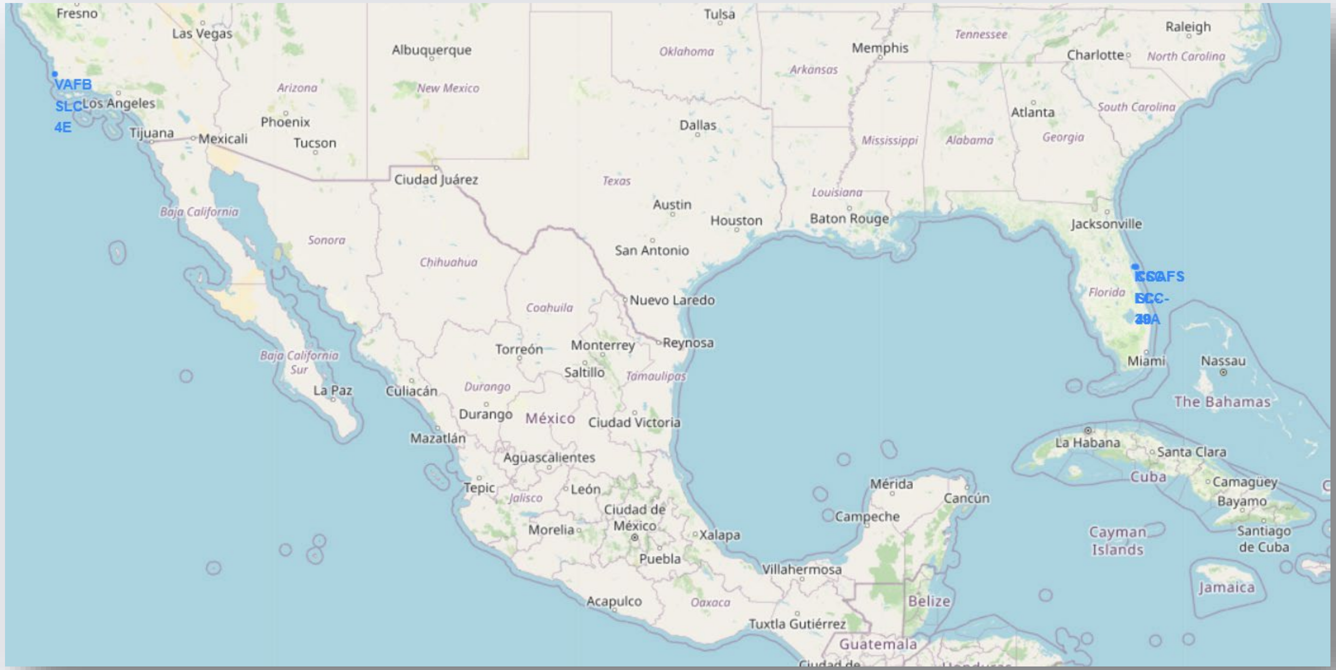
Falcon 9 Launch Site Locations

California, USA

- ❖ VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E

Florida, USA

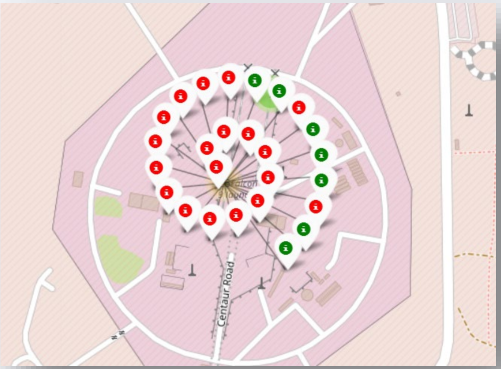
- ❖ KSC LC-39A | Kennedy Space Center Launch Complex 39A
- ❖ CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40

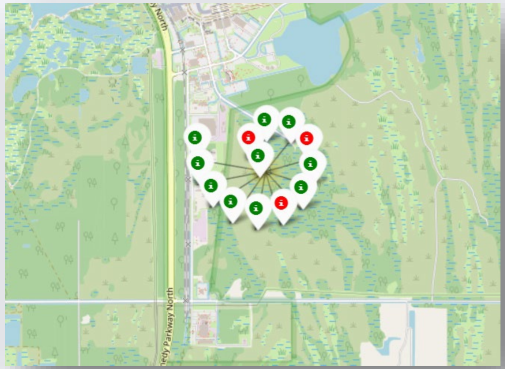


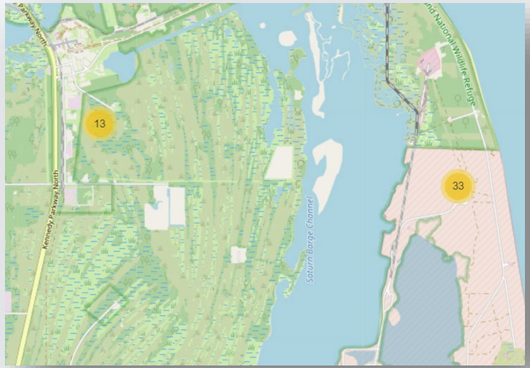
Map Markers of Success/Failed Landings

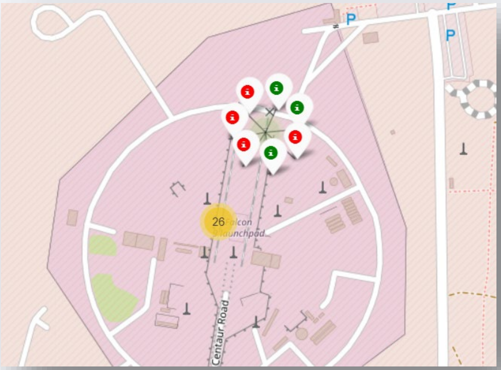
The markers represent the mission outcomes (Success or Failure) for Falcon 9 first-stage landings, positioned on the map according to the geographical coordinates of each launch site.

- ❖ The success rate of a launch site for Falcon 9 landings can be inferred by comparing the proportion of green success markers to red failure markers.



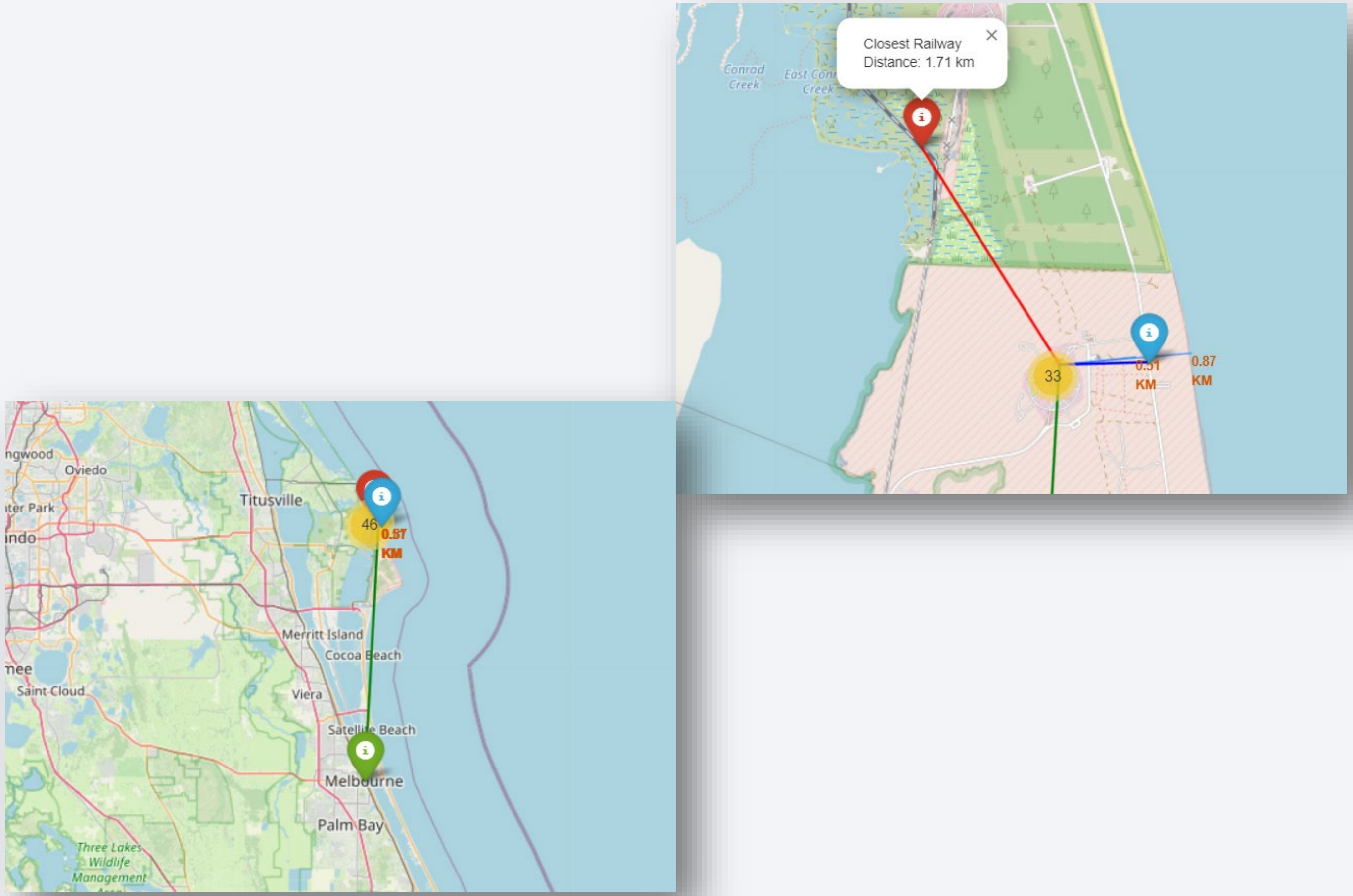






Distance from Launch Site to Proximities

- The perimeter road around CCAFS LC-40 is 0.51 km away from the launch site coordinates.
- The coastline is 0.87 km away from CCAFS LC-40.
- The rail line is 1.71 km away from CCAFS LC-40.
- The closest city Melbourne is 54.03 km away from CCAFS LC-40.



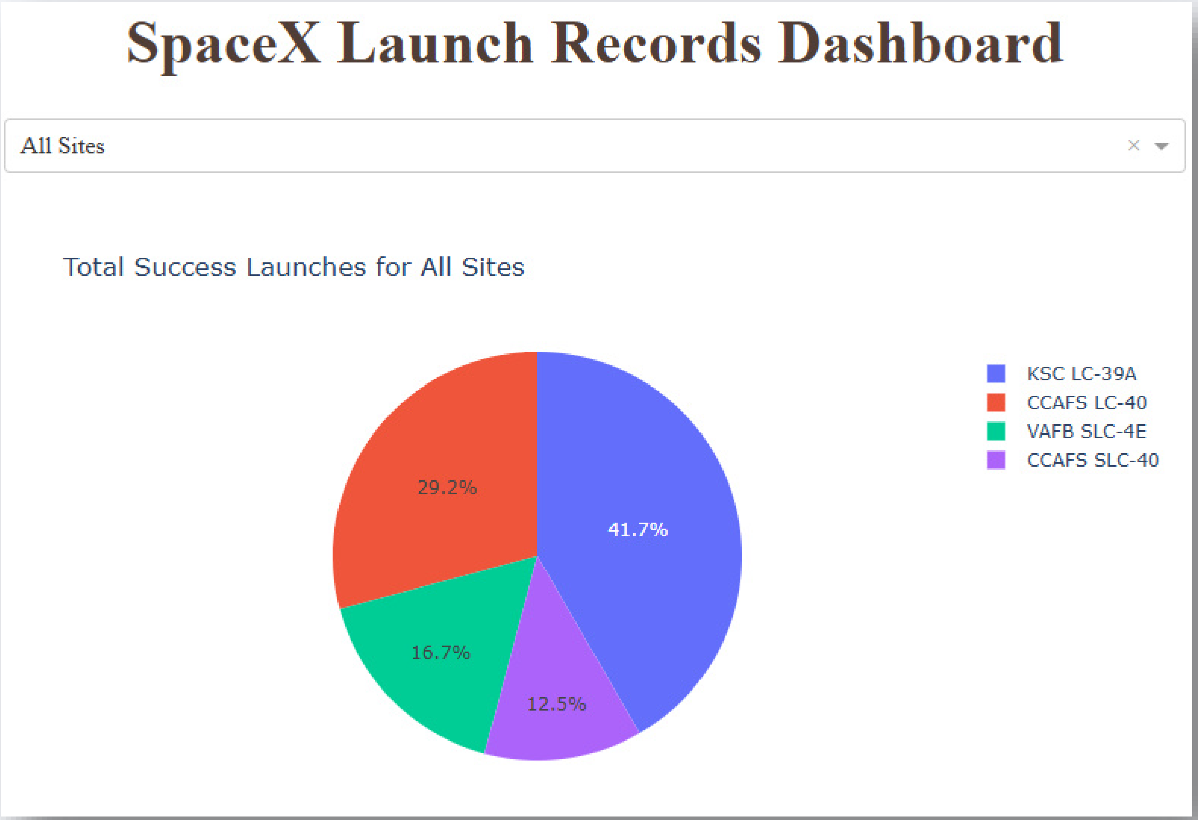
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

When all launch sites are selected, the pie chart illustrates the distribution of successful Falcon 9 first-stage landings across various launch sites.

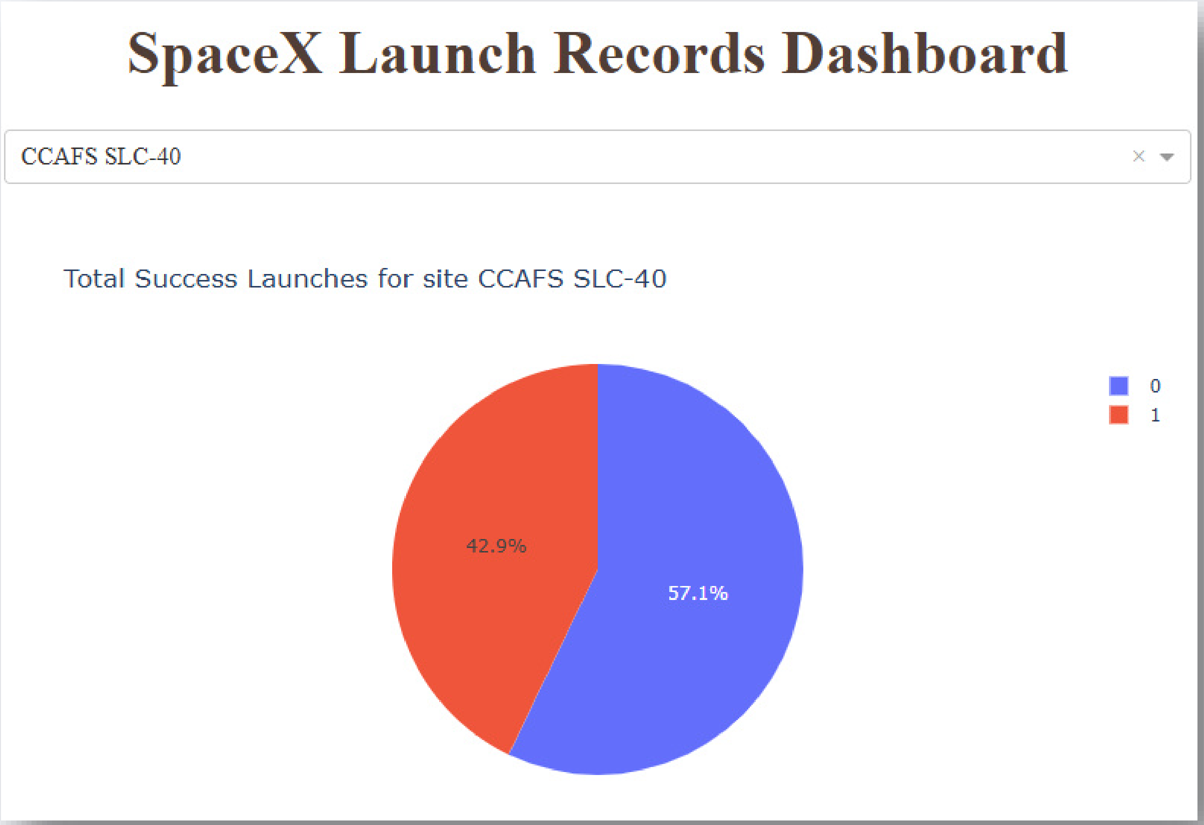
❖ The majority of successful landings, accounting for 41.7% of the total, took place at KSC LC-39A.



Launch Site with the Highest Launch Success Ratio

CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).

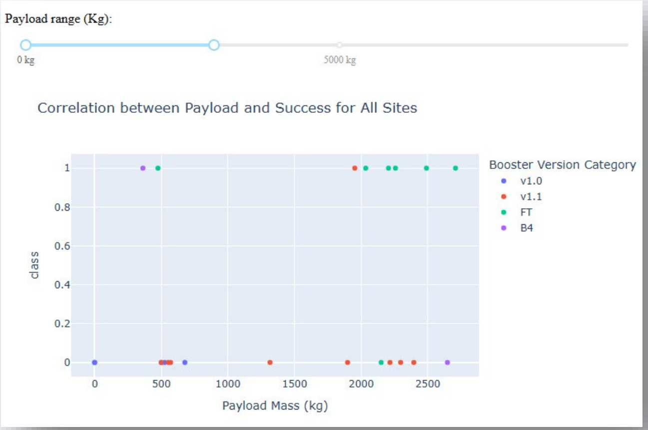
(failed landings are indicated by the ‘0’ Class and successful landings by the ‘1’)

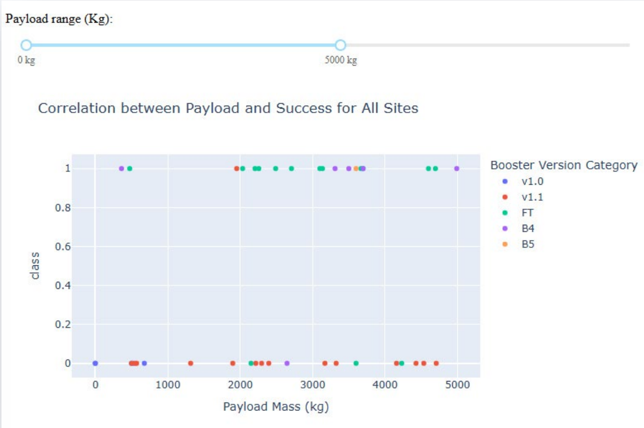


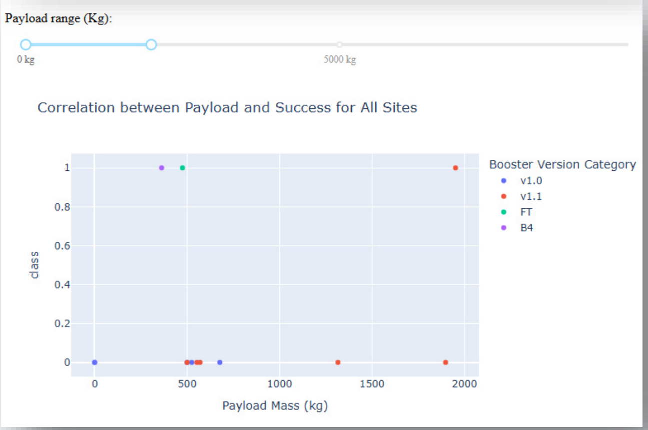
Payload vs. Launch Outcome Scatter Plot

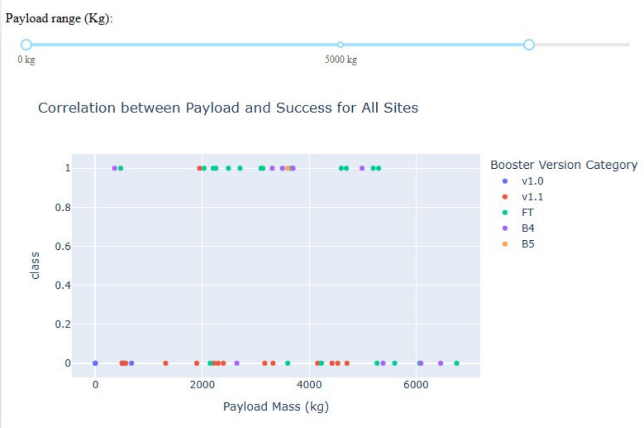
These scatter plots display Payload vs. Launch Outcome for all sites, with payload ranges adjustable using the range slider.

- ❖ The highest success rate is observed for payloads between approximately 2,000 kg and 5,000 kg.







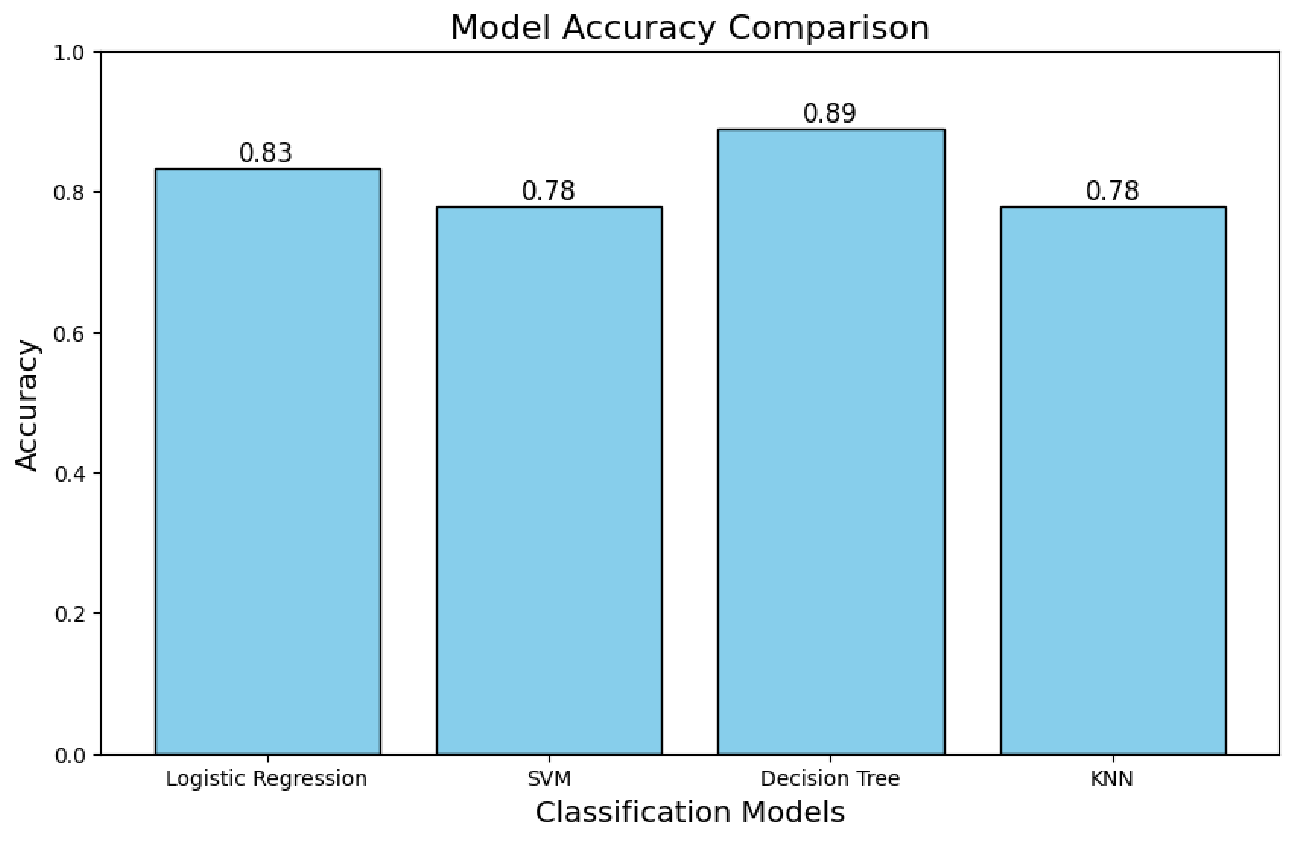


Section 5

Predictive Analysis (Classification)

Classification Accuracy

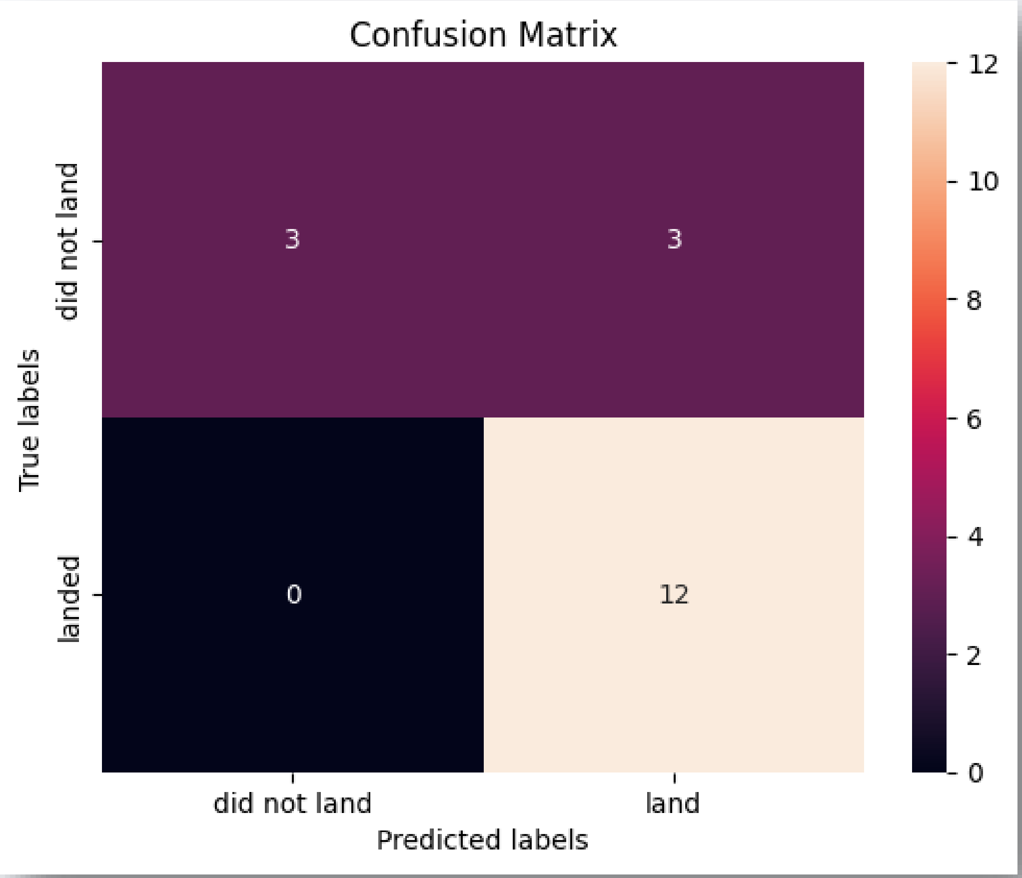
When using from sklearn.metrics import accuracy_score, we obtain different accuracy scores, which are used to plot the bar graph. Among all the models, the Decision Tree performs the best.



Confusion Matrix

The Logistic Regression performs slightly better as it avoids false negatives entirely, meaning it correctly identifies all successful landings.

❖ Depending on the importance of avoiding false negatives or false positives, this model may be preferred, especially if false negatives (missing a successful landing) are more critical.



Conclusions

- ❖ SpaceX has been increasingly successful in landing the first stage of its Falcon 9 rockets.
- ❖ This upward trend suggests that future launches are likely to have even higher success rates.
- ❖ Additionally, machine learning models can now be used to forecast the outcome of future Falcon 9 first stage landings.
- ❖ The insights support SpaceX in refining mission planning and resource allocation, while future work should focus on incorporating environmental factors and optimizing complex orbital missions. This analysis underscores the value of data-driven strategies in advancing space exploration.

Appendix

- ❖ SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- ❖ Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- ❖ SpaceX (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- ❖ Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- ❖ • Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Thank you!