IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashritha Anandal
18th Jan 2026

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodologies used to analyse data:

- Data collection
- Data wrangling
- EDA with SQL
- EDA with Visualisation
- Building an interactive map with folium
- Building a dashboard with plotly dash
- Predictive analysis

Summary of all results:

- It was possible to aggregate and prepare high-quality datasets by mining through open-sources resources
- EDA showed the specific variables that served as the strongest indicators for a successful launch
- Machine Learning prediction identified an optimal model to forecast mission outcomes.

Introduction

Project background and context

- SpaceX has transformed the commercial space industry by reducing launch costs.
- Falcon 9 missions cost about $62 million, significantly lower than traditional providers.
- Cost reduction is mainly achieved through reusable first-stage rockets.
- Using machine learning and public data, we can predict first-stage landing success, helping estimate mission cost.

Problems you want to find answers

- How do specific factors like payload weight, chosen launch site, flight count, target orbit etc influence landing success?
- Best place for launches?
- Has the frequency of successful landings shown a positive trend over the years?

Section 1

Methodology

Methodology

- Data collection methodology:
    - Official SpaceX API data
    - Web scraping wikipedia
- Perform data wrangling
    - Filter data and handle missing values
    - Prepare data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

•**SpaceX REST API**: Used an official API to pull technical flight details like flight number, payload mass, and landing pad information.

•**Web Scraping**: Extracted historical launch tables from Wikipedia to gather extra details about customers and specific landing outcomes.

•**Data Integration**: Combined both sources to create a complete dataset for more accurate analysis.

•**Key Columns**: Gathered critical features such as booster version, orbit type, and launch site to understand success factors.

Data Collection – SpaceX API

- **API Retrieval**: Pulled technical launch records directly from the official SpaceX REST API to ensure data accuracy.
- **Data Selection**: Custom functions to extract critical features like Flight Number, Booster Version, Payload Mass, Orbit Type, and Launch Site.
- **Data Storage**: Raw data was initially converted from a JSON format into a Python dictionary and then into a structured Pandas dataframe.

Request API: Call the SpaceX endpoint to retrieve all historical launch data.

JSON Parsing: Decode the response using .json() and flatten the data into a table format

Filtering: Remove all rockets except for **Falcon 9** to keep the analysis focused.

Handling Missing Values: Identify any null entries in the **Payload Mass** column and fill them using the **mean** (average) value.

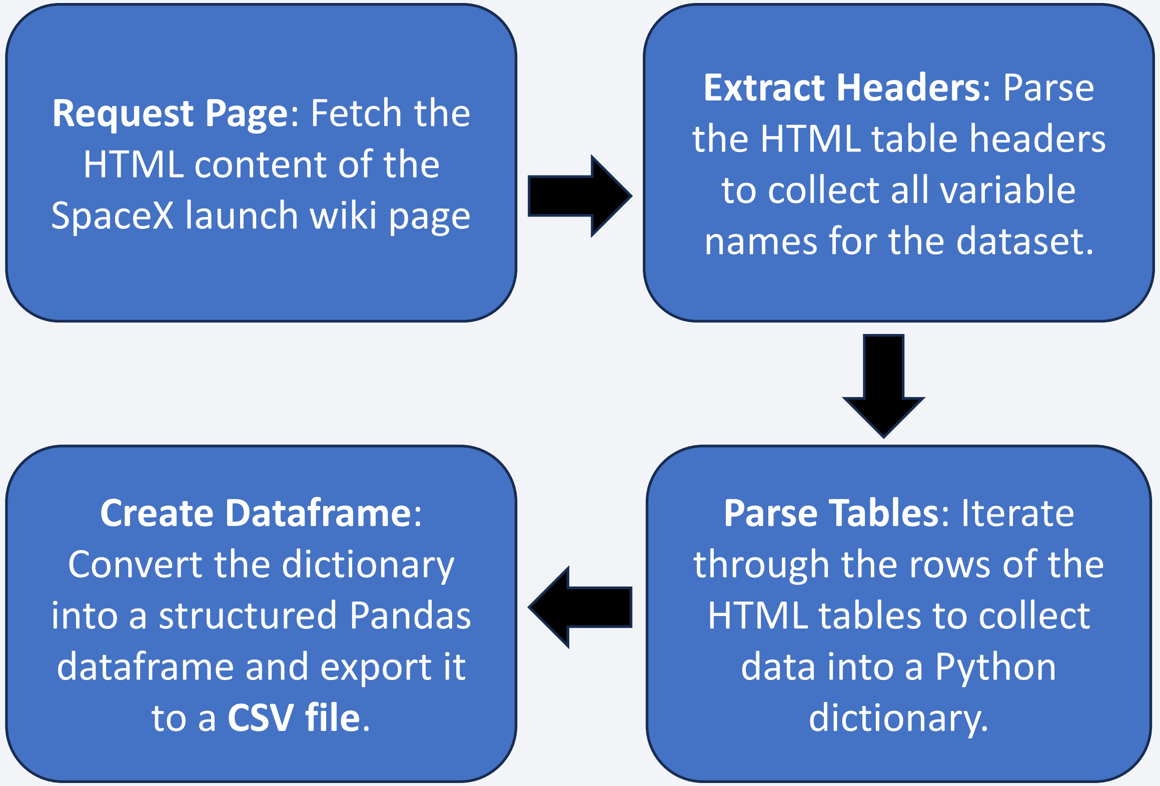
https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/data%20collection%20api.ipynb

Data Collection - Scraping

•**Wikipedia Data Retrieval**: Downloaded historical Falcon 9 launch records from Wikipedia.

•**HTML Parsing**: Used the BeautifulSoup library to convert the raw HTML response into a searchable object to find relevant data tables.

•**Feature Discovery**: Extracted column headers to identify key variables such as Flight No., Launch Site, Payload, Orbit, Customer, and Launch Outcome.



https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/webscraping.ipynb

Data Wrangling

- **Data Enrichment**: Converted raw landing descriptions into clear Training Labels, where "1" = successful booster landing and "0" =unsuccessful attempt.
- **Statistical Summaries**: Calculated the total number of launches per site and the frequency of mission outcomes for each orbit type to identify performance trends.
- **Feature Engineering**: Used One Hot Encoding to transform categorical columns into a numerical format, making the dataset compatible with binary classification models.

Initial Exploration: Conducted **Exploratory Data Analysis (EDA)** to determine the necessary training labels.

Outcome Classification: Categorized complex outcomes (like "True RTLS" or "False ASDS") into a simplified landing success column.

Data Cleaning: Handled the dataset by filtering irrelevant data and managing missing values in the payload mass column.

Final Export: Converted the processed data into a clean **CSV file** ready for the machine learning phase.

https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/data%20wrangling.ipynb

EDA with Data Visualization

- **Visualization Strategy**:
  - Used Scatter Plots to show the relationship between continuous variables (like Flight Number and Payload Mass)
  - Bar Charts to compare discrete categories (like Orbit types).

- **Trend Analysis**: Utilized Line Charts to track the yearly average success rate, allowing us to see if SpaceX's landing technology improved over time.

- **Feature Relationship Discovery**: By plotting variables like Payload Mass vs. Launch Site, we identified patterns that help determine which features should be included in our machine learning models.

https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/eda%20data%20visualization.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/eda%20sql.ipynb

Build an Interactive Map with Folium

Map Objects Used:

- Markers: Show launch sites and important locations
- Circles: Highlight key areas on the map
- Marker Clusters: Group multiple launches at the same site
- Colored Markers: Green = Success, Red = Failure
- Lines: Show distances to nearby city, coast, highway, and railway

Purpose:

- Visualize launch site locations
- Analyze launch success and failure patterns
- Understand proximity to important infrastructure

https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/interactive%20map%20with%20folium.ipynb

Build a Dashboard with Plotly Dash

Plots & Interactions Used:

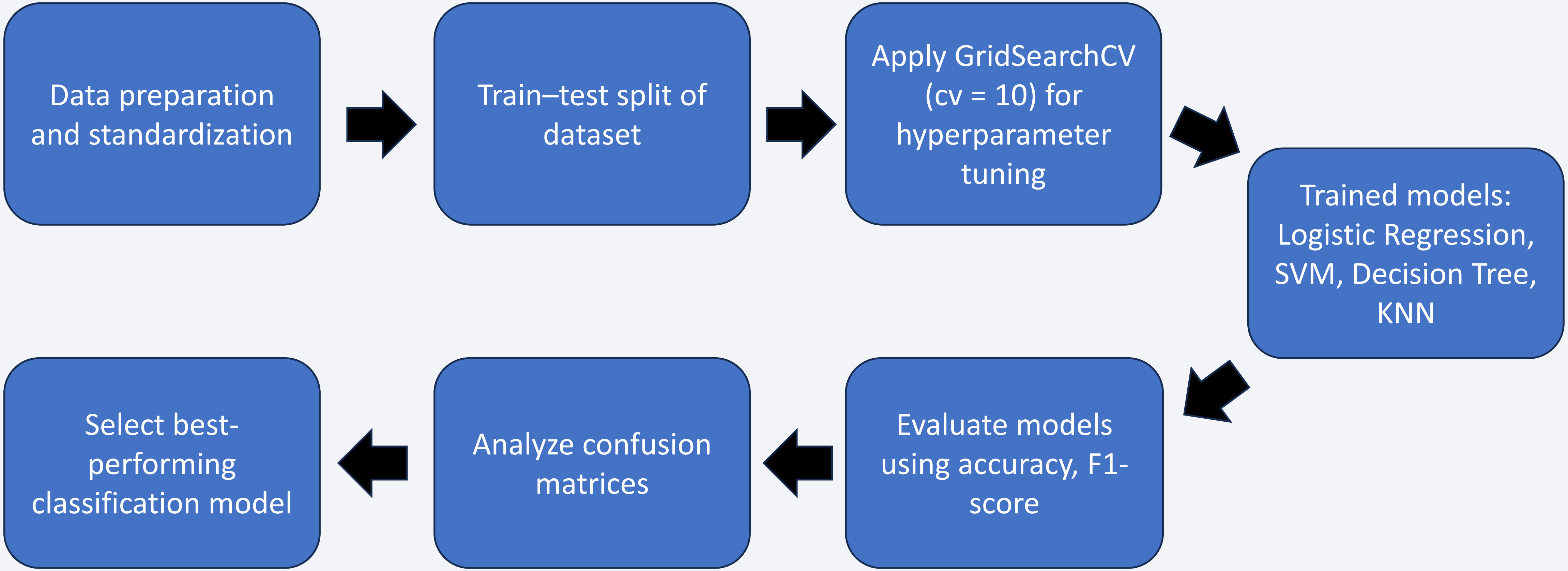
- <u>Dropdown</u>: Select launch site
- <u>Pie Chart</u>: Shows successful launches (all sites / selected site)
- <u>Slider</u>: Select payload mass range
- <u>Scatter Plot</u>: Payload mass vs launch success by booster version

Purpose:

- Compare launch success across sites
- Analyze impact of payload on launch outcome
- Enable interactive and dynamic data exploration

https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/ashrithaxx/Applied-data-science-capstone/blob/main/machine%20learning%20prediction.ipynb

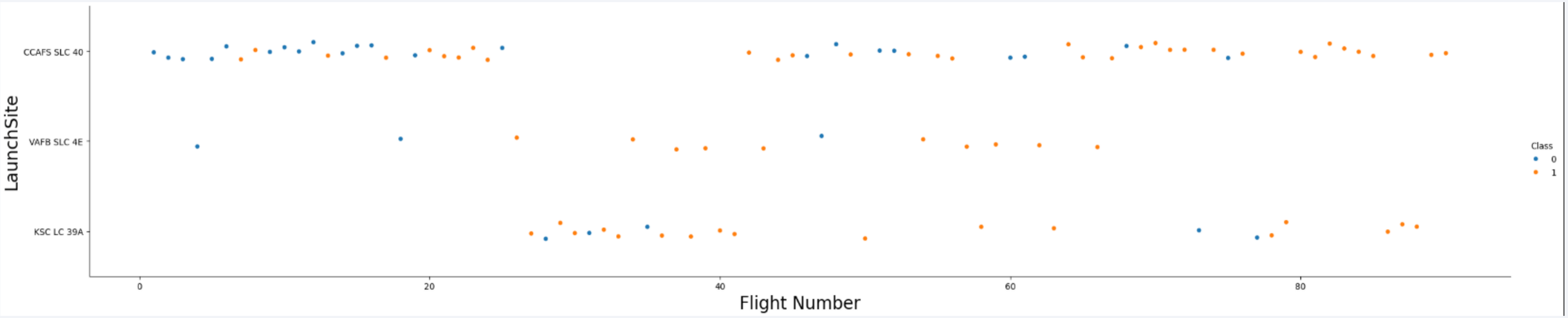
Results

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

Section 2

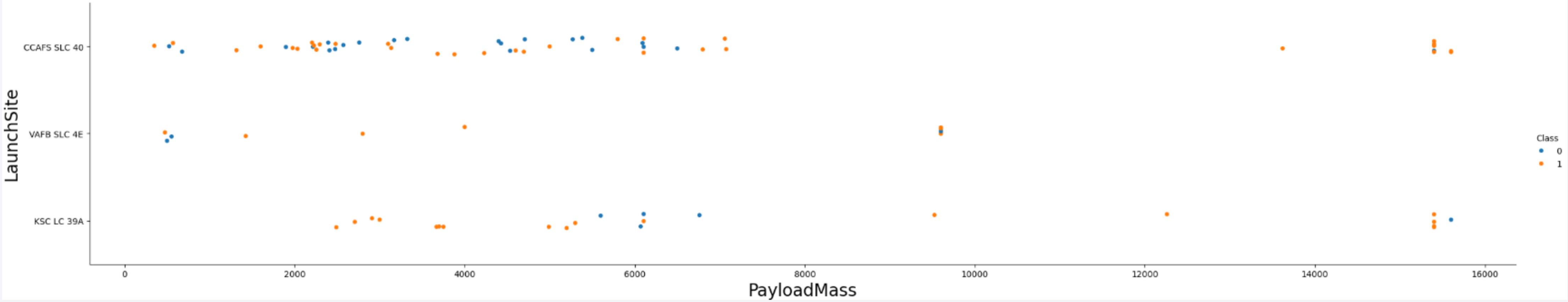
Insights drawn from EDA

Flight Number vs. Launch Site



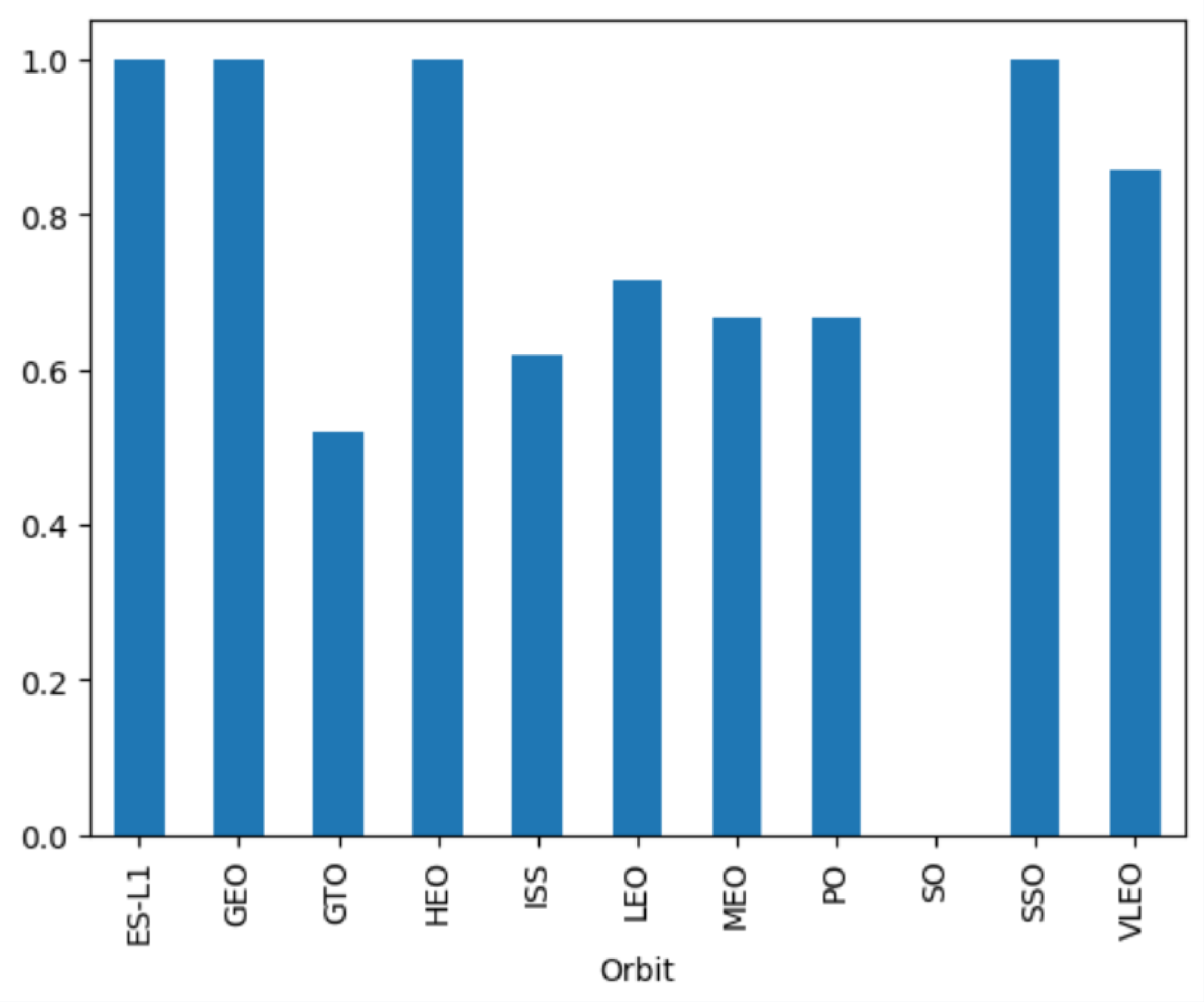
- CCAFS SLC-40 has the highest number of launches across all flight numbers
- KSC LC-39A and VAFB SLC-4E show higher success rates in later missions
- Overall, launch success improves as flight number increases

Payload vs. Launch Site



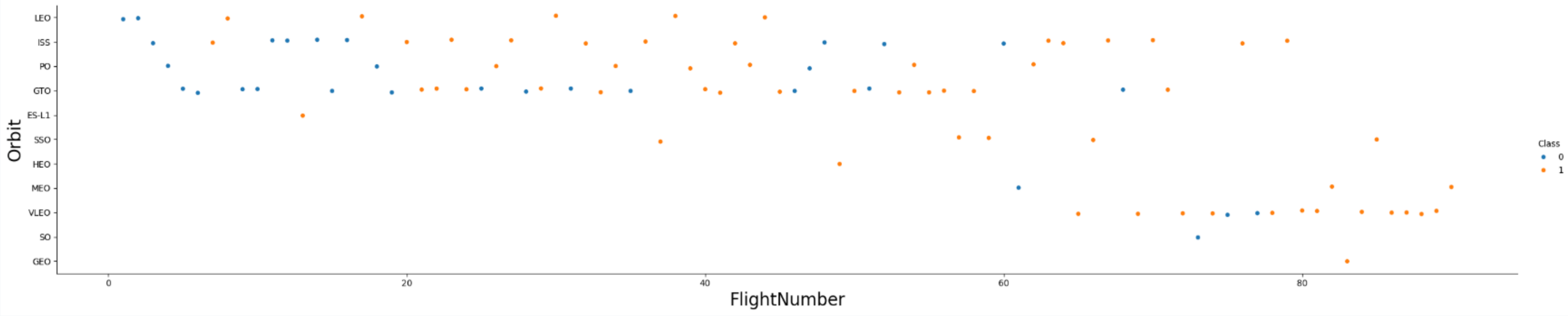
- Higher payload mass generally leads to higher success rates across all launch sites
- Most launches with payload above 7000 kg were successful
- KSC LC-39A achieved 100% success for payloads below 5500 kg

Success Rate vs. Orbit Type



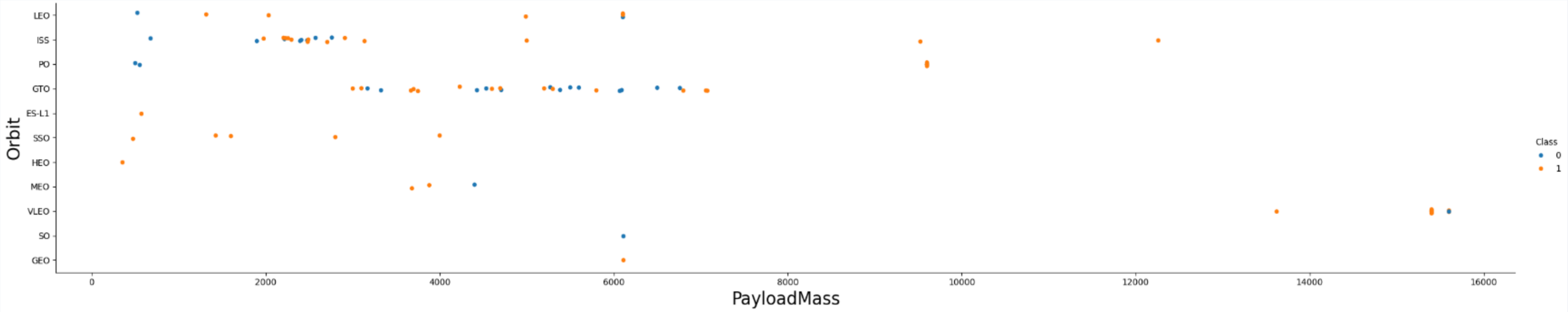
0% success rate: SO
50%–85% success rate: GTO, ISS, LEO, MEO, PO
100% success rate: ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type



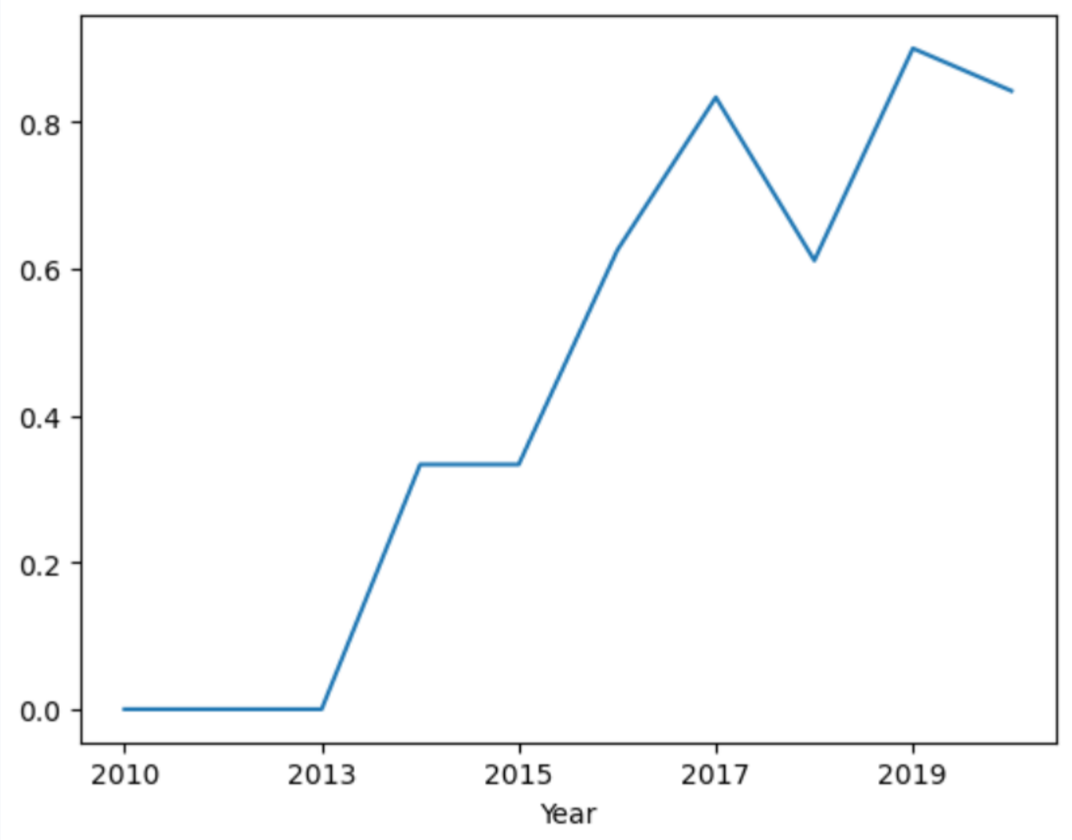
- In LEO, launch success increases with flight number
- In GTO, no clear relationship with flight number
- Overall success rate improves over time across all orbits
- VLEO shows rising frequency, indicating a new business opportunity

Payload vs. Orbit Type



•GTO missions are mostly concentrated in the mid-payload range with mixed outcomes
•VLEO launches appear only at higher payload masses, indicating emerging usage
•Failures are more common at lower payload masses, while heavier payloads tend to succeed

Launch Success Yearly Trend



- Launch success rate was very low before 2013
- From 2013 onwards, success rate steadily increased
- Peak success observed during 2019–2020
- Overall trend shows continuous improvement over time

All Launch Site Names

- Displaying names of unique launch sites

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records of launch sites starting with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (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 Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Displaying total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Displaying the avg payload mass carried by booster version F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission

Mission_Outcome	Total_Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- Booster versions that carried max payload

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015

```
Done.
```

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

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

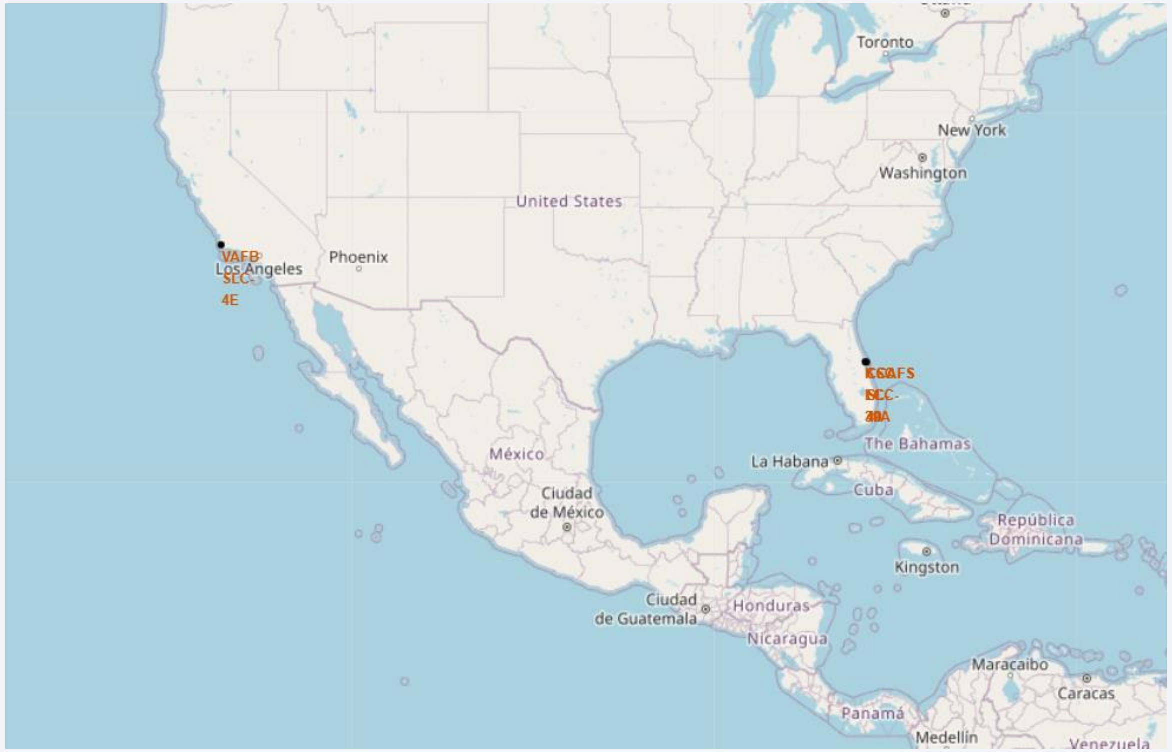
Done.

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- “No attempt” must be taken into account

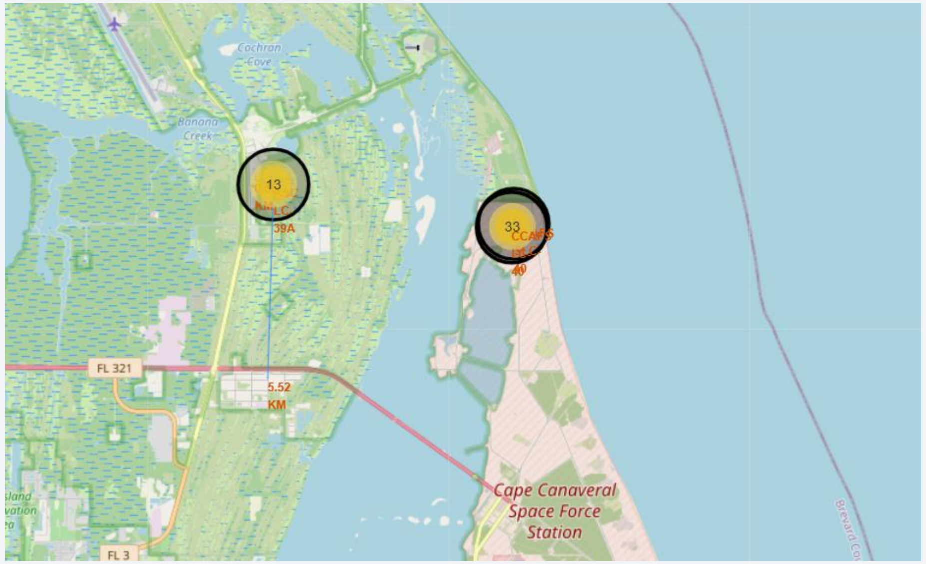

Section 3
Launch Sites
Proximities Analysis

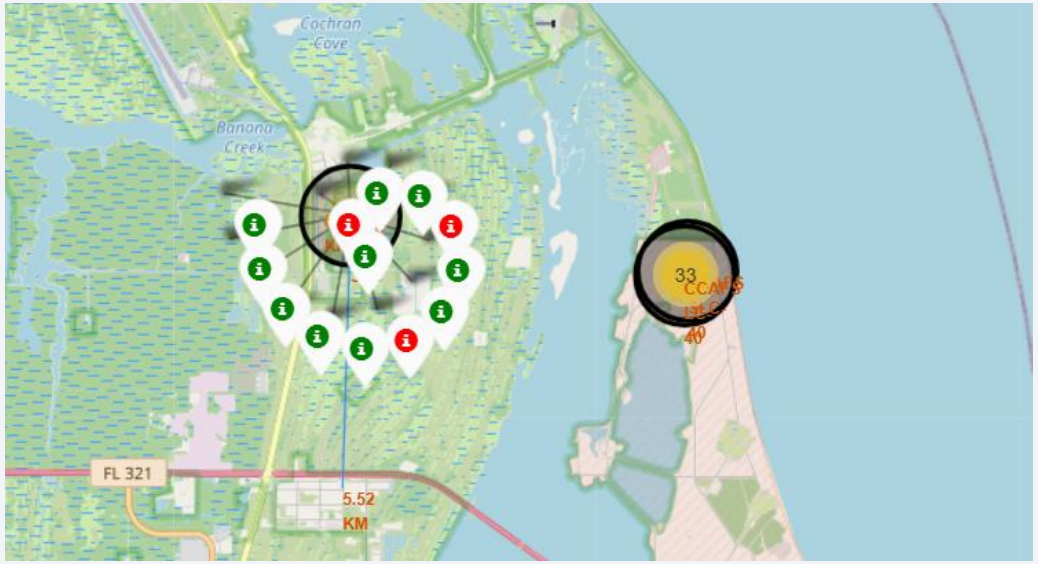
All Launch Sites



- •All SpaceX launch sites are located within the United States
- •Launch sites are distributed on both east and west coasts
- •Major sites include CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E
- •Sites are located near the equator compared to higher latitudes, supporting orbital launches

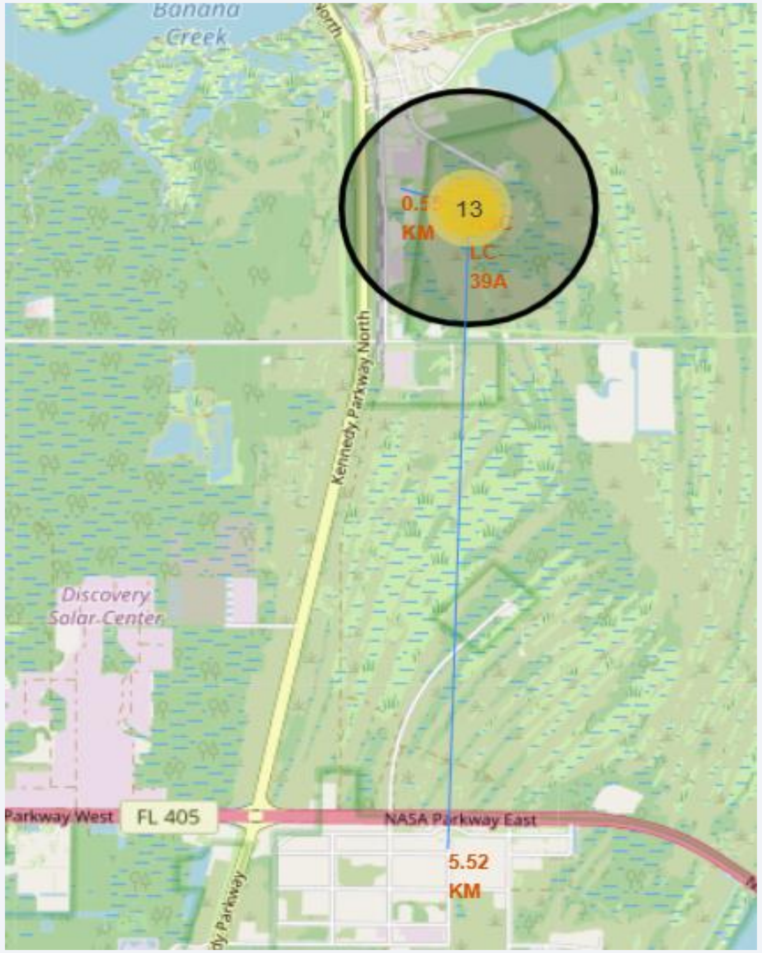
Launch Outcomes by Site





- Colour-coded markers help identify launch success at each site
- **Green marker:** Successful launch
- **Red marker:** Failed launch
- Enables quick comparison of success rates across launch sites

Logistics

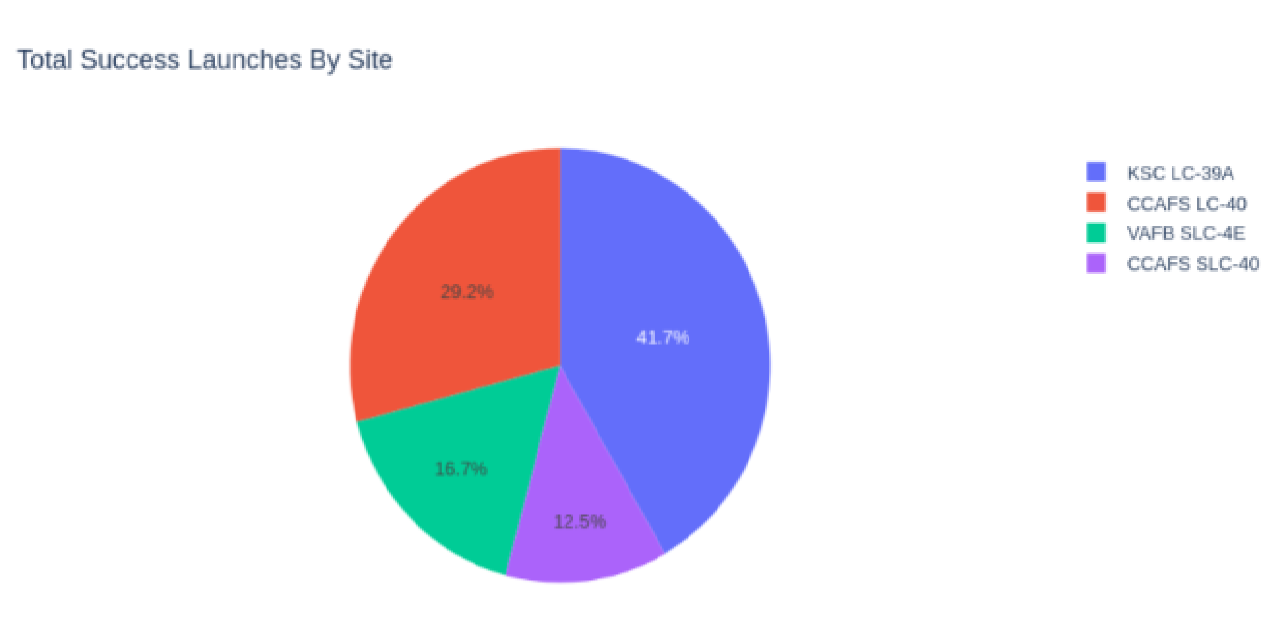


- KSC LC-39A has good logistics support, located near roads and railways and away from densely populated areas

Section 4

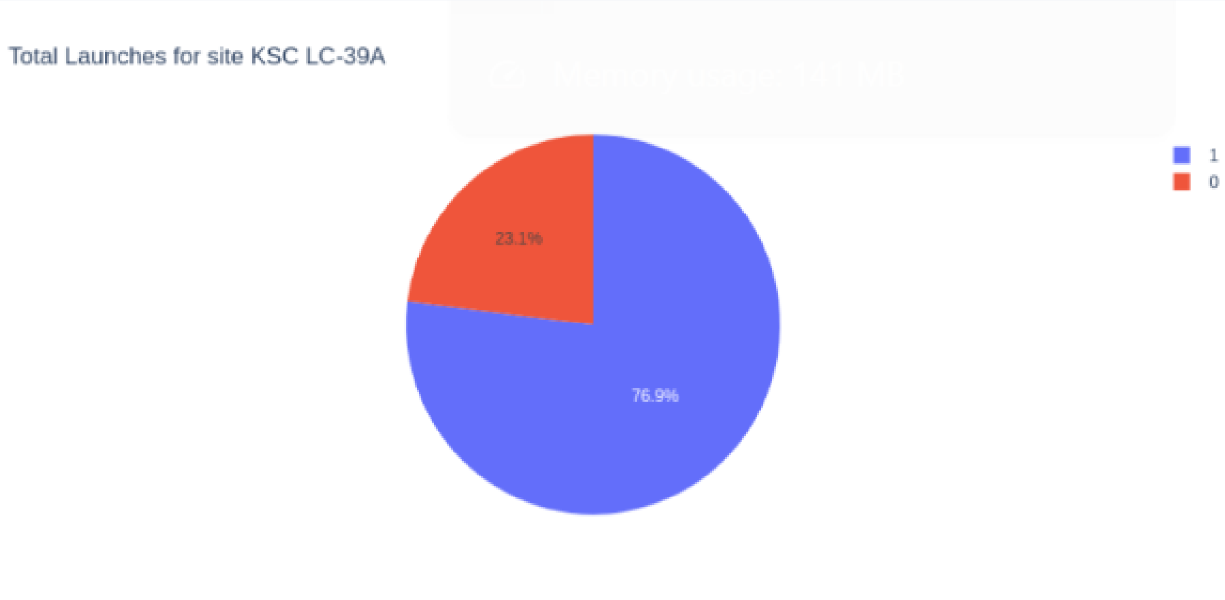
Build a Dashboard with Plotly Dash

Successful Launches



- KSC LC-39A has the highest number of successful launches
- CCAFS SLC-40 is the second most successful launch site
- VAFB SLC-4E and CCAFS LC-40 contribute comparatively fewer successes
- Launch site location plays an important role in mission success

Launch Site with Highest Success Ratio



•KSC LC-39A has the highest launch success rate (76.9%)

•Out of total launches, 10 were successful and only 3 failed

•This indicates strong reliability and consistency at this launch site

Payload Mass vs Launch Outcome

•Payloads between 2000–5500 kg show the highest success rate

•Payloads under 6000 kg perform better overall

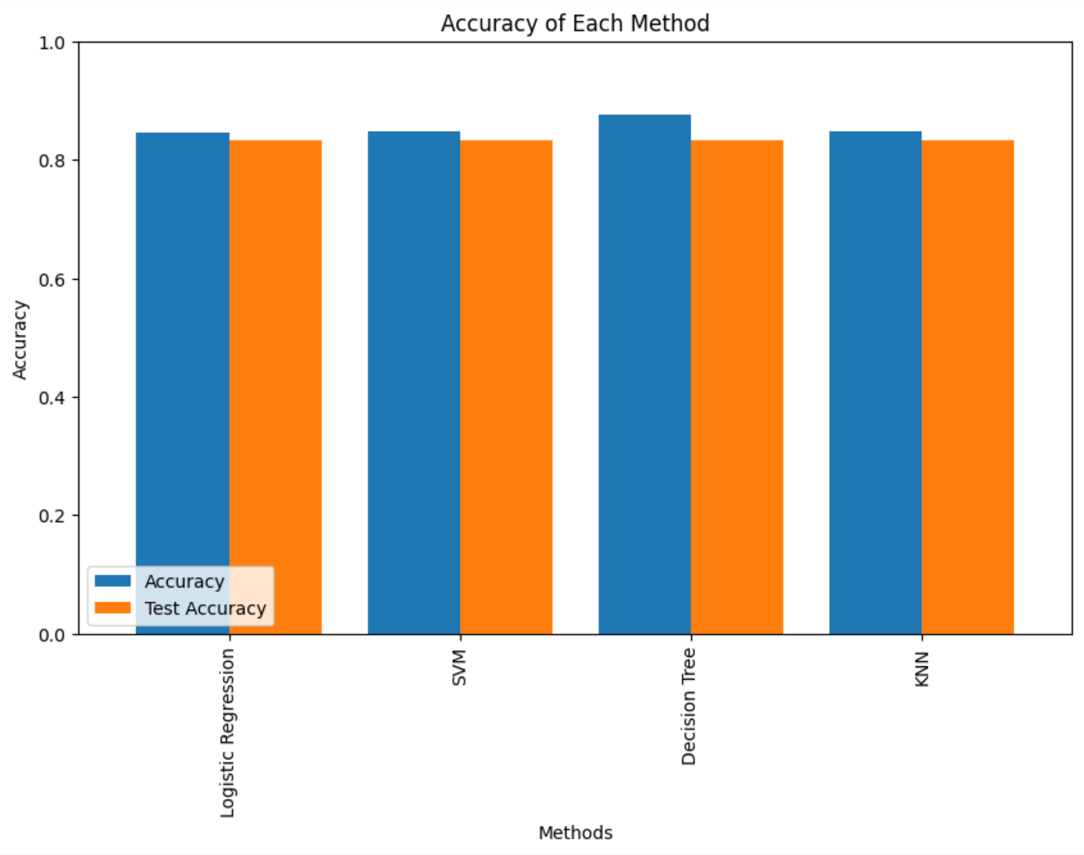
•FT booster versions have the most successful launches

•Higher payload ranges show mixed success outcomes

Section 5

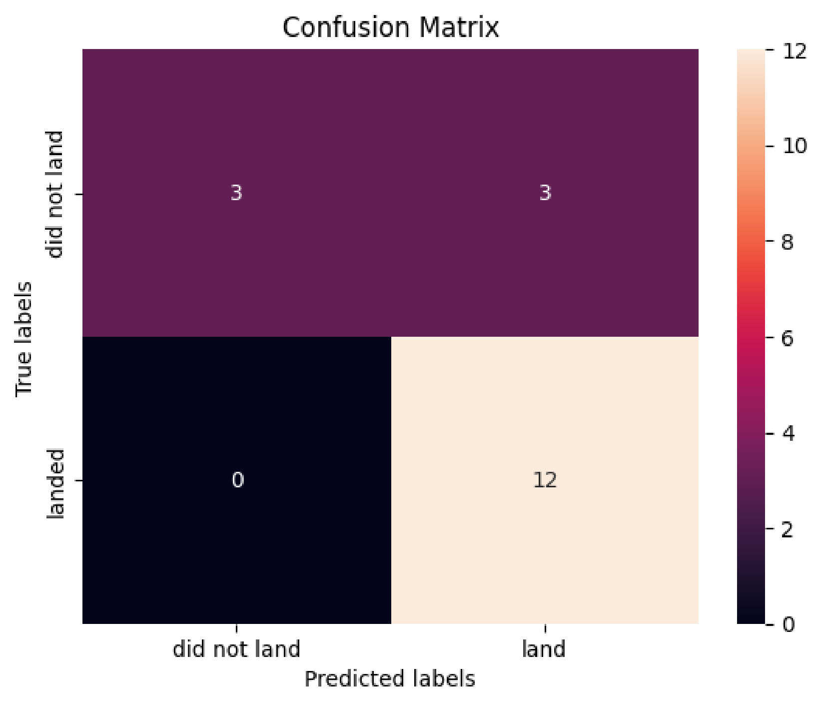
Predictive Analysis (Classification)

Classification Accuracy



- **Four classification models were evaluated**: Logistic Regression, SVM, Decision Tree, and KNN
- Accuracy results were compared side by side
- **Decision Tree Classifier** achieved the highest accuracy
- Best performance observed at above 87% accuracy

Confusion Matrix



•**True Positives (12):** Correctly predicted successful landings
•**True Negatives (3):** Correctly predicted failed landings
•**False Positives (3):** Predicted landing, but actually failed
•**False Negatives (0):** No successful landings were missed

INSIGHT:

• The model perfectly identifies successful landings (no false negatives)
• Most errors occur when failures are predicted as successes
• Overall, the confusion matrix confirms high accuracy and strong reliability of the Decision Tree model

Conclusions

- Multiple data sources were analyzed to derive reliable insights.
- **KSC LC-39A** is the most reliable launch site with the highest success rate.
- Launch success has improved steadily over the years.
- Some orbits (ES-L1, GEO, HEO, SSO) achieved 100% success.
- The **Decision Tree model** performed best for predicting successful landings.

Key Takeaways and Business Impact

- Predicting landing success supports cost-efficient mission planning
- Reliable launch sites reduce operational risk
- Machine learning enables data-driven decision making for future launches
- Insights from this analysis can guide launch strategy optimization

Thank you!