IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lucía Pavón
July 16th, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions
- Appendixes

Executive Summary

I developed an end-to-end pipeline to analyze and predict SpaceX Falcon 9 first stage landing outcomes.

To do this, I first collected and processed data using web scraping and APIs. Then, I performed EDA using visualizations, SQL, Folium, and interactive dashboards.

Finally, I built multiple classification models and found the best-performing one to predict mission outcomes.

Introduction

Project background

SpaceX is a company that designs, manufactures, and launches rockets and spacecraft. Falcon 9 is a reusable, two-stage rocket created by SpaceX whose launches have a cost of USD62M, as opposed to other companies whose launches have costs over USD165M. This difference in cost is because SpaceX can reuse the first stage of the rocket launch.

Objective

To predict whether the first stage of the rocket launch will land successfully. This will allow me to determine the price of each launch.

Section 1

Methodology

Methodology

Executive Summary

- Collected launch data via Wikipedia scraping and SpaceX API.
- Cleaned and wrangled data to obtain a structured format.
- Performed visual EDA and SQL analysis.
- Built an interactive dashboard with Dash and maps with Folium.
- Trained classification models using GridSearchCV.

Data Collection

- Scraped Wikipedia for Falcon 9 raw launch data.

Wikipedia url → HTTP GET (requests.get) → HTML response → BeautifulSoup Parser → Extract launch records → Clean & normalize columns → Pandas Dataframe

- Queried SpaceX API for launch records and metadata.

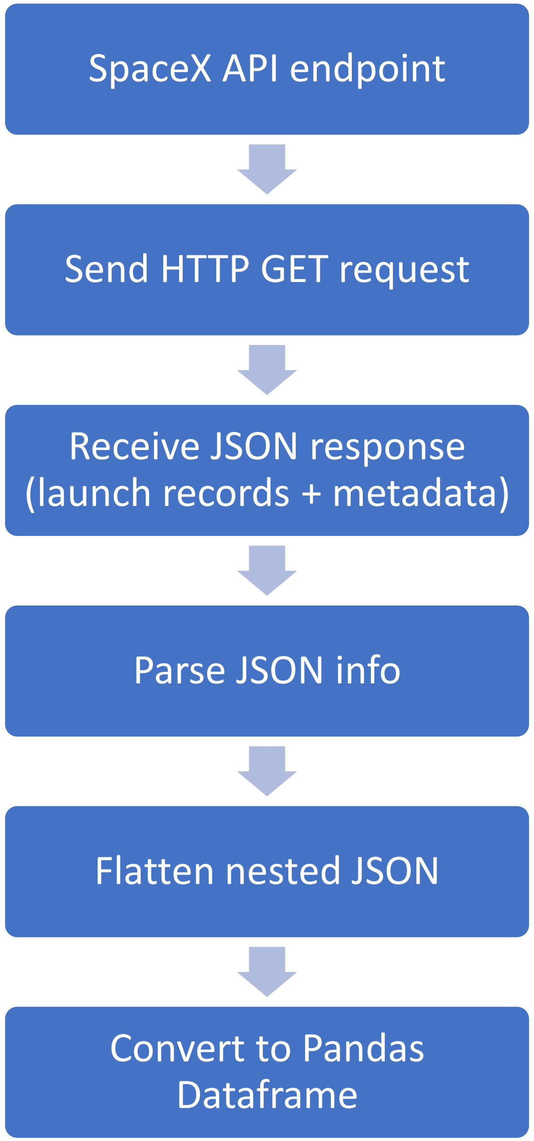
SpaceX REST API → HTTP GET (requests.get) → JSON response → Extract launch attributes → Flatten ested JSON (json_normalize) → Clean & transform → Pandas Dataframe

- Imported .csv datasets from SpaceX datasets for structured analysis.

.csv url → pd.read_csv() → Structured table (launch metrics) → Merge with scraped/API data

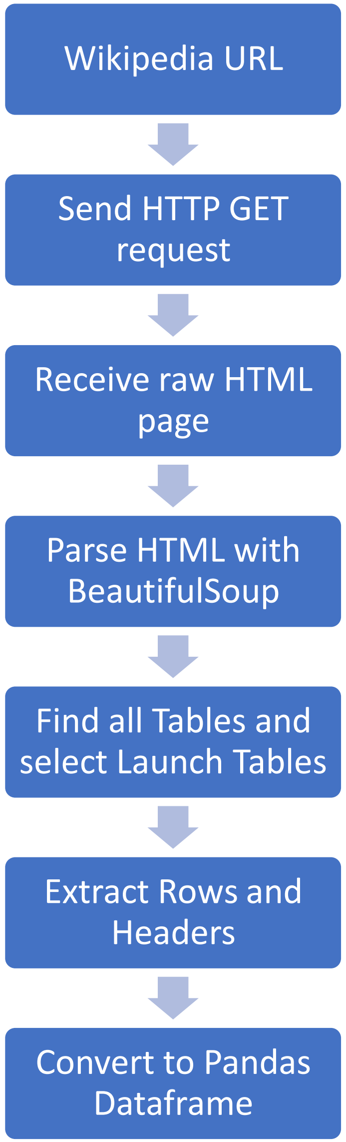
Data Collection – SpaceX API

- Requested SpaceX REST API for launch records.
- Extracted mission attributes (rocket type, payload mass, launchpad, etc.) from data in nested JSON format.
- Used pandas.json_normalize to flatten nested structures into a DataFrame.
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/LuciaPavon/capstone-project/blob/main/jupyter-labs-spacex-data-collection-api-SOLVED.ipynb



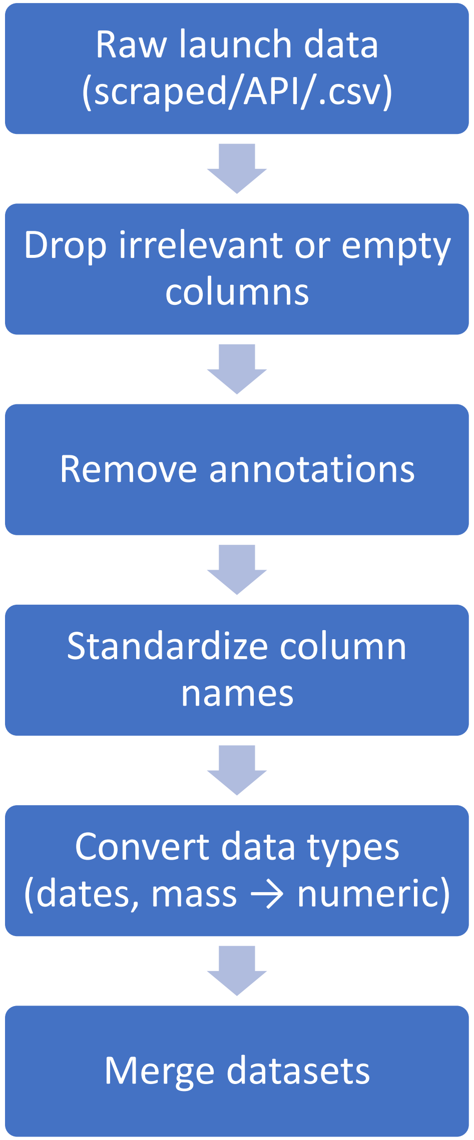
Data Collection - Scraping

- Performed HTTP GET request on Wikipedia Falcon 9 launch history page.
- Parsed HTML using BeautifulSoup to locate launch tables.
- Extracted column headers and table rows containing launch data.
- Cleaned irregular formatting and removed inline annotations.
- Loaded structured content into a Pandas Dataframe.
- GitHub URL of the completed web scraping notebook: https://github.com/LuciaPavon/capstone-project/blob/main/jupyter-labs-webscraping-SOLVED.ipynb



Data Wrangling

- Cleaned noisy HTML tables, removed annotations (e.g., "[1]"), and standardized column names.
- Converted payload mass to numeric, parsed date/time fields.
- Merged datasets from different sources.
- GitHub URL of the completed data wrangling related notebooks: https://github.com/LuciaPavon/capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling-SOLVED.ipynb



EDA with Data Visualization

Used the following:

- Scatter plots to study relationships between payload, launch site, flight number, orbit type, and outcome.
- Bar chart to analyze outcome by orbit type.
- Line chart to visualize success rate trends over time.
- GitHub URL of the completed EDA with data visualization notebook: https://github.com/LuciaPavon/capstone-project/blob/main/edadataviz-SOLVED.ipynb

EDA with SQL

- Queried unique launch sites.
- Calculated total/average payload mass by booster.
- Identified first successful landings by type.
- Ranked outcomes within a time window.
- GitHub URL of the completed EDA with SQL notebook: https://github.com/LuciaPavon/capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite-SOLVED.ipynb

Interactive Map with Folium (I)

- Created the following map objects:
  - Markers → Placed at each SpaceX launch site to visually identify geographic locations.
  - Colored Markers (Red/Green) → Represented individual launch outcomes (green = success, red = failure) for intuitive interpretation of performance per site.
  - Circles → Used around NASA Johnson Space Center for emphasis and labeling, to help locate and label each launch site precisely.
  - Lines (PolyLines) → Connected launch sites to key proximities like nearest city, railway, highway, and coastline. These landmarks are useful for evaluating safety, accessibility, and strategic placement of launch sites.

Interactive Map with Folium (II)

- DivIcons→ Displayed numeric distances (e.g., “12.5 KM”) between launch sites and nearby infrastructure for contextual reference.

- GitHub URL of the completed interactive map with Folium map: https://github.com/LuciaPavon/capstone-project/blob/main/lab_jupyter_launch_site_location-SOLVED-V2.ipynb

Dashboard with Plotly Dash (I)

- Added the following graphs and interactions:
  - Dropdown Menu → Allows users to select a specific launch site or view all sites, making the dashboard flexible and interactive.
  - Pie Chart → Dynamically displays the distribution of successful launches by site (when “All” is selected) or success vs. failure for an individual site.
  - Payload Range Slider → Enables users to filter launches by payload mass to explore correlations between payload weight and success, and to identify optimal payload thresholds for successful landings.

Dashboard with Plotly Dash (II)

- Scatter Plot → Plots payload mass vs. mission outcome with data points color-coded by booster version, showing correlations between payload and outcome. It also updates based on dropdown and slider selections.

- GitHub URL of the completed Plotly Dash lab: https://github.com/LuciaPavon/capstone-project/blob/main/spacex-dash-app.py

Predictive Analysis (Classification) (I)

- Built a model development and evaluation pipeline by following these steps:
  - Applied StandardScaler to normalize feature distributions.
  - Used train_test_split to separate training and testing data (80/20 split).
  - Trained and evaluated 4 models:

    → Logistic Regression
    → Support Vector Machine (SVM)
    → Decision Tree
    → K-Nearest Neighbors (KNN).
  - Performed hyperparameter tuning with GridSearchCV (cv=10).
  - Measured model performance using accuracy on test data with score method.

Predictive Analysis (Classification) (II)

- Selected the best model based on highest test accuracy.

- GitHub URL of the completed predictive analysis lab: https://github.com/LuciaPavon/capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5-SOLVED-V2.ipynb

Results Summary (I)

- Exploratory Data Analysis (EDA) Results:
  - Launch success rates have improved significantly since 2013.
  - KSC LC-39A has the highest number of successful missions.
  - Payloads in the 3000–4000 kg range achieved the best landing outcomes.

- Interactive Analytics (Dashboard & Folium):
  - Folium Map helped visualize:
    - Launch site locations.
    - Outcome patterns (success/failure markers).
    - Distances to coastlines, highways, cities, and railways.

Results Summary (II)

- Dash Dashboard enabled:
    - Filtering by site and payload range.
    - Visual discovery of success trends by booster version.

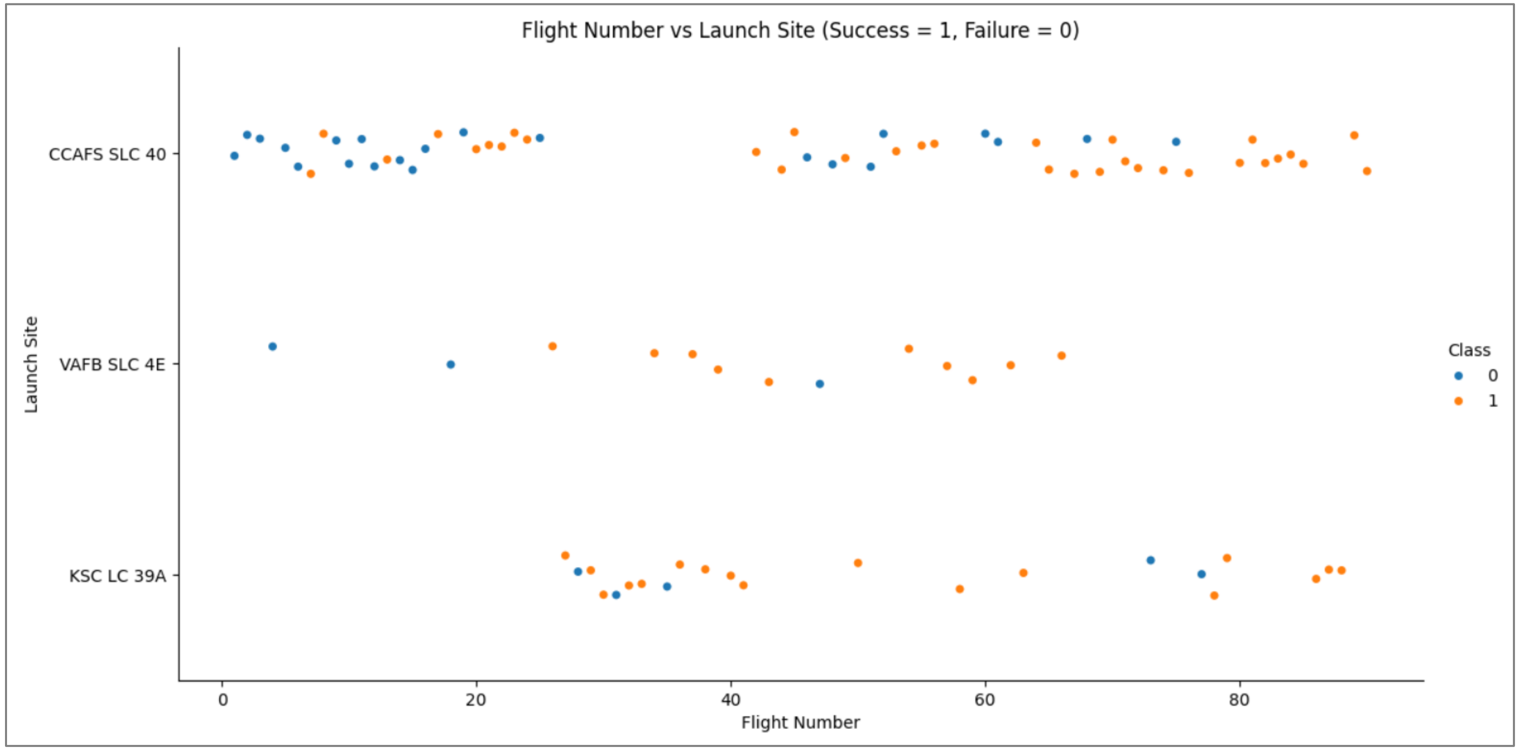
- Predictive Analysis Results:
    - Multiple classification models trained: Logistic Regression, SVM, Decision Tree, and KNN.
    - Logistic Regression achieved the best accuracy on test data.
    - Confusion matrix showed strong ability to identify successful missions.

Section 2

Insights drawn from EDA

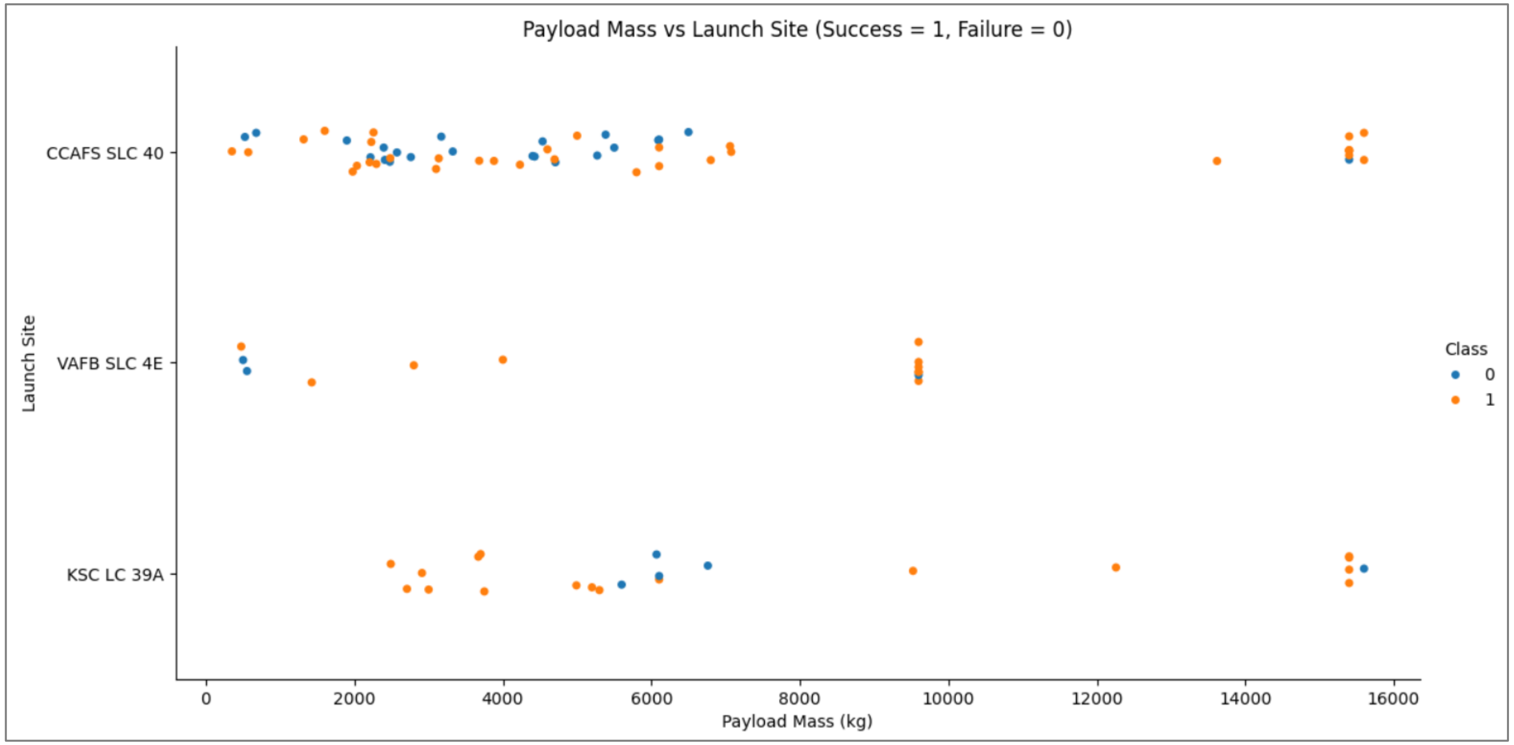
Flight Number vs. Launch Site

- Failures were more common for the earliest flight numbers. However, as the flight number increases over time, the number of successful launches increases at all launch sites.
- Some launch sites, like CCAFS SLC 40 and KSC LC 39A, have a higher concentration of successful launches.



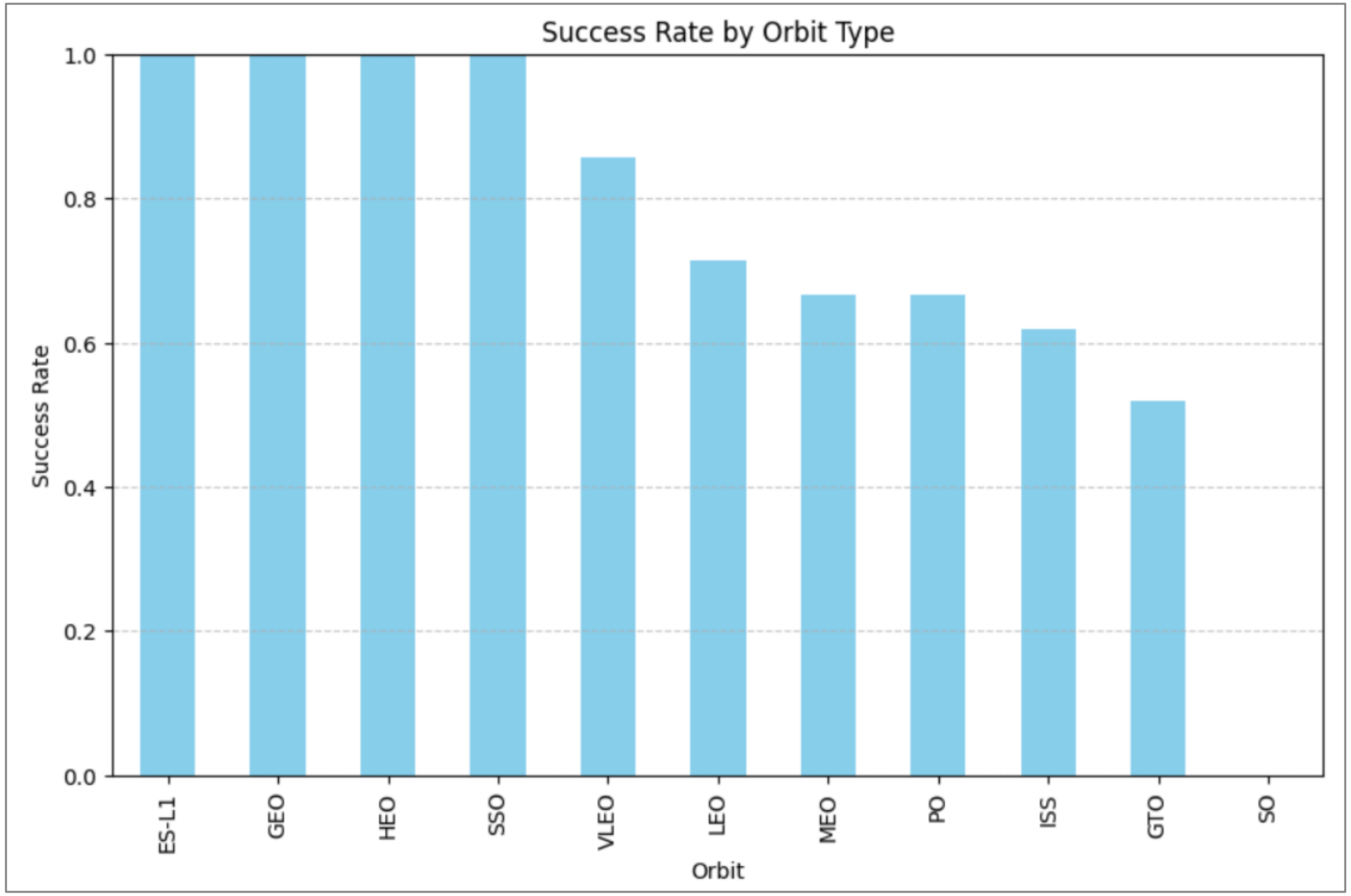
Payload Mass vs. Launch Site

- There are fewer launches with heavier payloads, but most of them still achieve successful landings.



Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have the highest success rates.

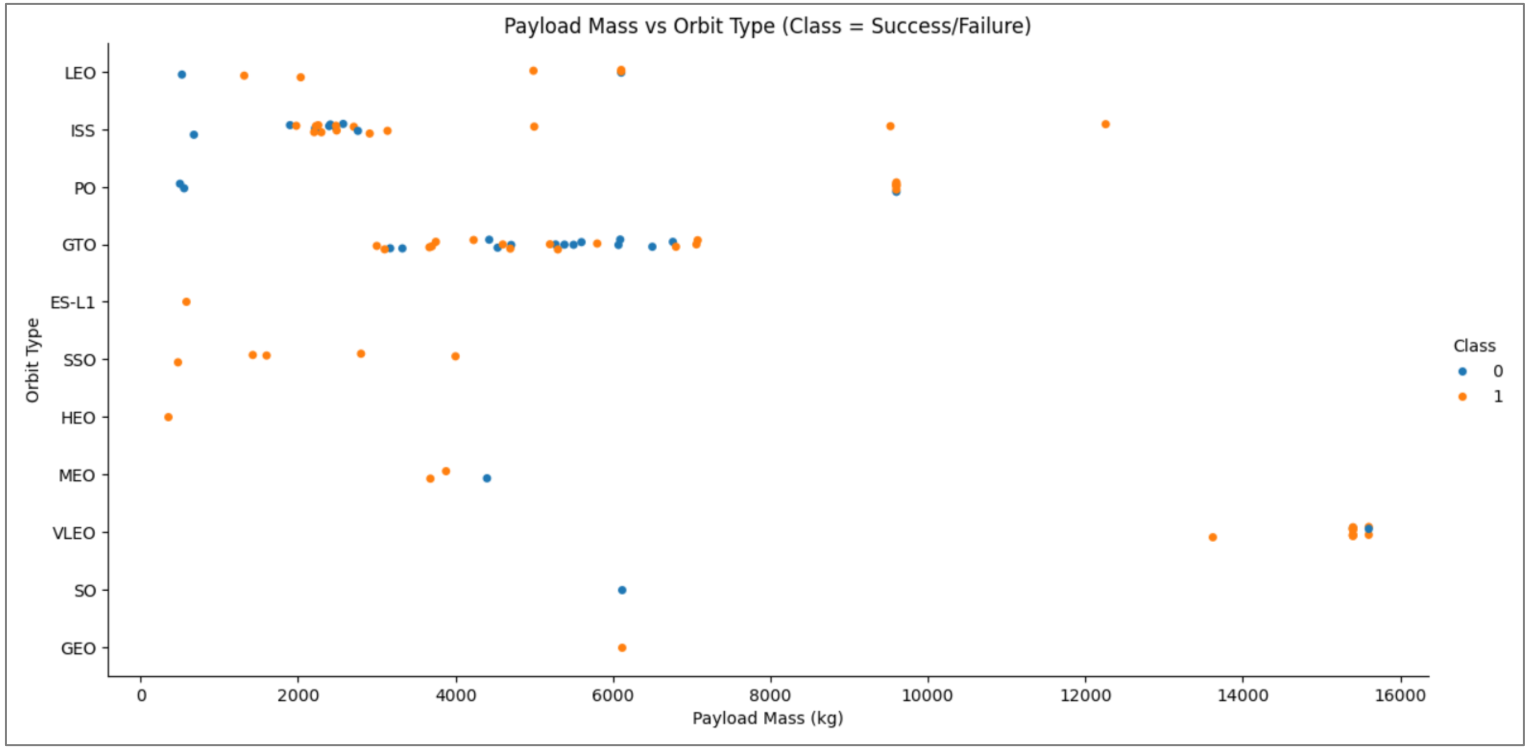


Flight Number vs. Orbit Type

- There is no strong correlation, but more recent flights show greater orbit diversity.

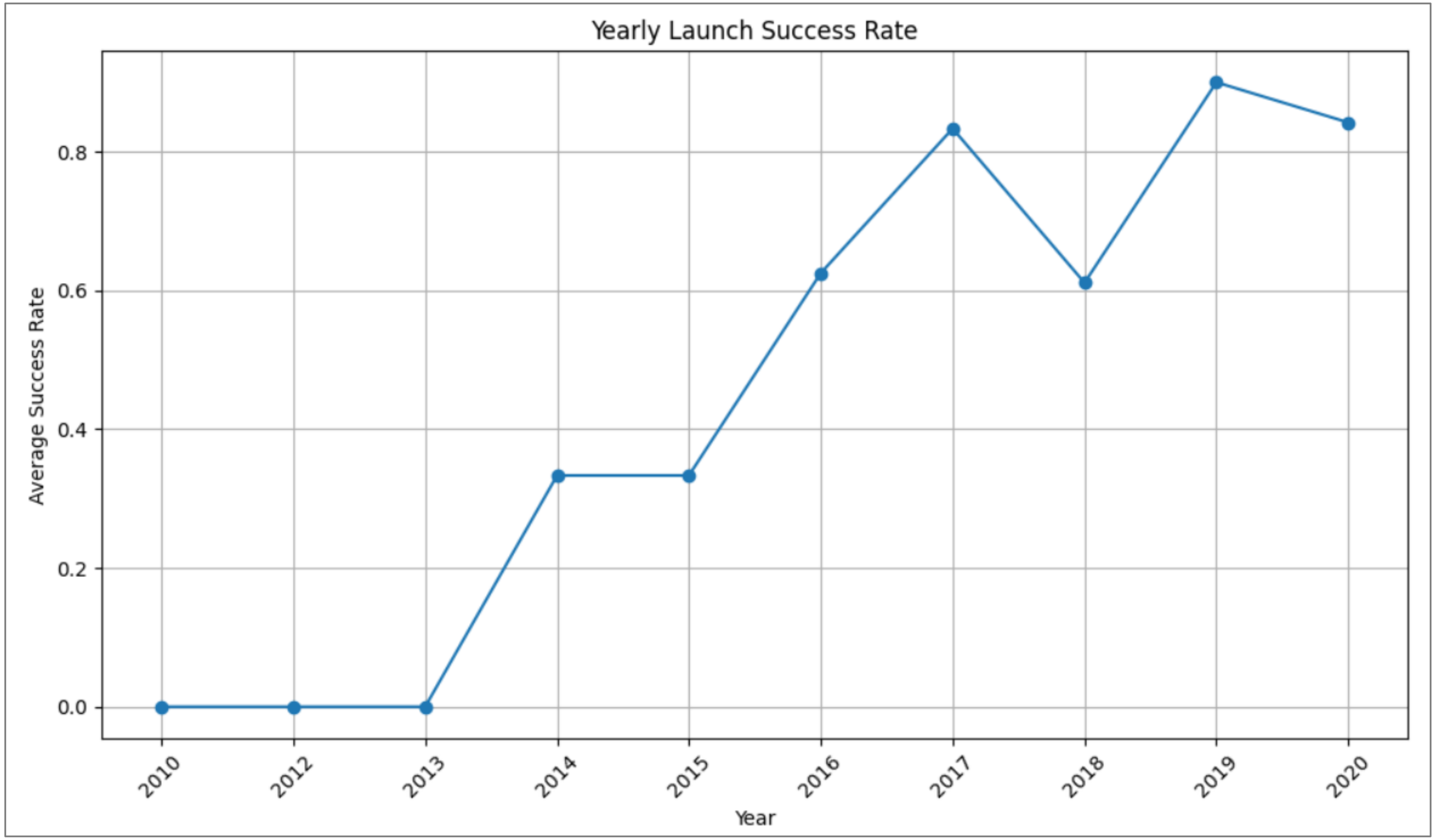
Payload Mass vs. Orbit Type

- Payload mass varies across orbit types.
- With heavy payloads the successful landing rate is higher for PO, ISS, and VLEO.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings, as both are present.



Launch Success Yearly Trend

- Success rates improved from 2013 to 2020, even though they dipped during 2017-2018.



Launch Site Names

There are 4 unique launch sites in the dataset:

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

Launch Site Names that Begin with 'CCA'

This is a sample of 5 records where launch sites begin with ‘CCA’:

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

The missions were successful; however, the records show that for booster versions B0003-4 the landings failed, and for B0005-6-7 the landings were not attempted.

Total Payload Mass

This is the total payload mass carried by boosters from NASA:

SUM("Payload_Mass__kg_")
45596

Calculating the total payload mass for each customer is useful for evaluating which customer launched the boosters that handled the most weight overall.

Average Payload Mass by F9 v1.1

This is the average payload mass carried by booster version F9 v1.1:

AVG("Payload_Mass__kg_")
2928.4

Calculating the average payload mass for this rocket version is useful for comparing performance across versions.

First Successful Ground Landing Date

This is the date of the first successful landing outcome on ground pad:

First_Ground_Pad_Success_Date
2015-12-22

The first successful ground landing occurred in 2015, two years after the first successful launch.

Successful Drone Ship Landing with Payload between 4000 and 6000

These are the boosters which have successfully landed on drone ships with payload masses between 4000 and 6000 kg.:

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

These medium-weight payload boosters are reliable for drone ship missions.

Total Number of Successful and Failure Mission Outcomes

This is the total number of successful and failed mission outcomes:

Mission_Outcome	Total_Outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

With only one failure, the missions were overwhelmingly successful.

Boosters which Carried Maximum Payload

Here are the names of the boosters which have carried the maximum payload mass, 15600kg.

It appears that the most capable versions are variations of B1048-49-51-56-60.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

These are the booster versions that had failed landing outcomes on drone ships during 2015, along with their launch sites:

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

These early landing failures occurred only months before the first successful ground landing.

Ranking of Landing Outcomes Between 2010-06-04 and 2017-03-20

Here is the ranking of landing outcomes between 2010-06-04 and 2017-03-20, in descending order.

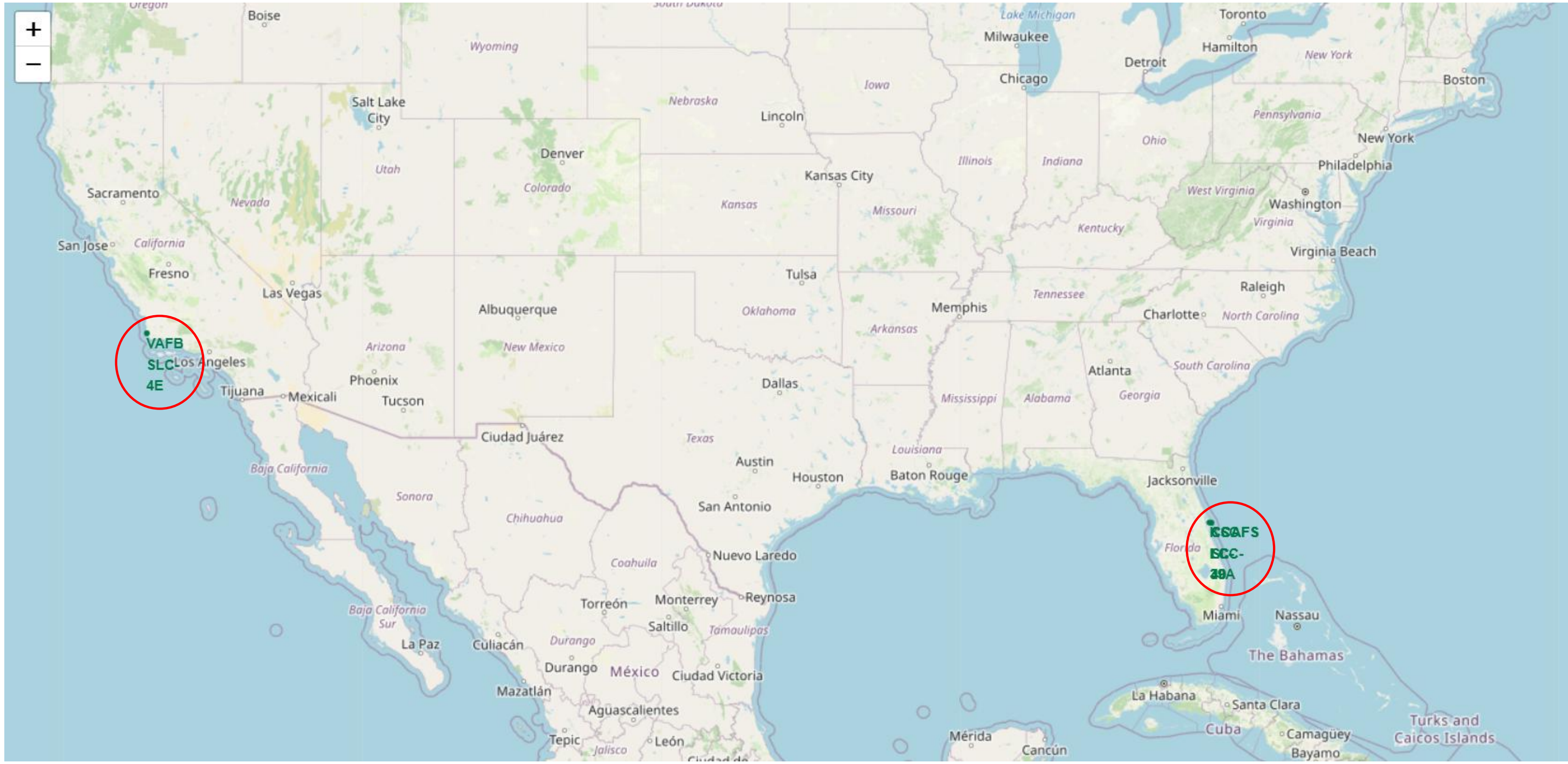
During this time, both success and failure on drone ship landing were the most common attempted outcomes, as the most common outcome was ‘No attempt’.

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

Section 3

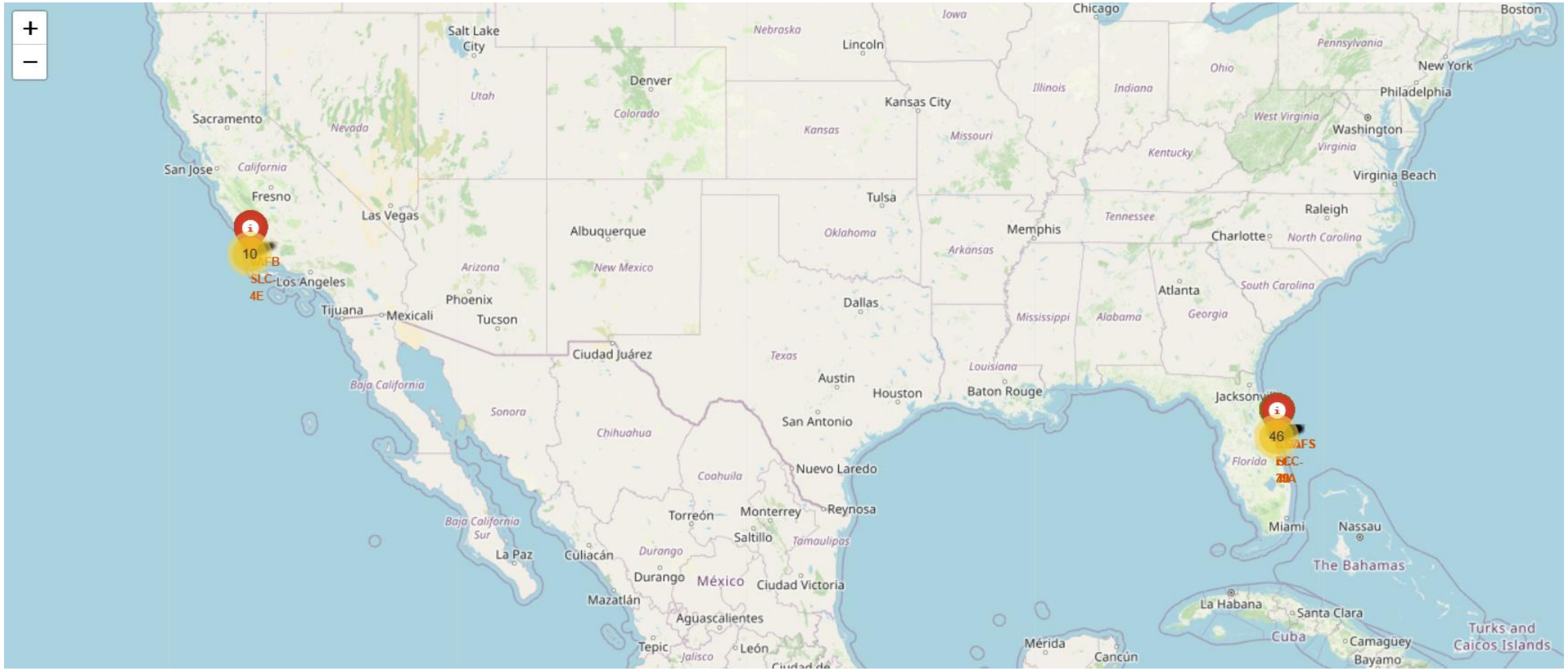
Launch Sites Proximities Analysis

Launch Site Locations Map

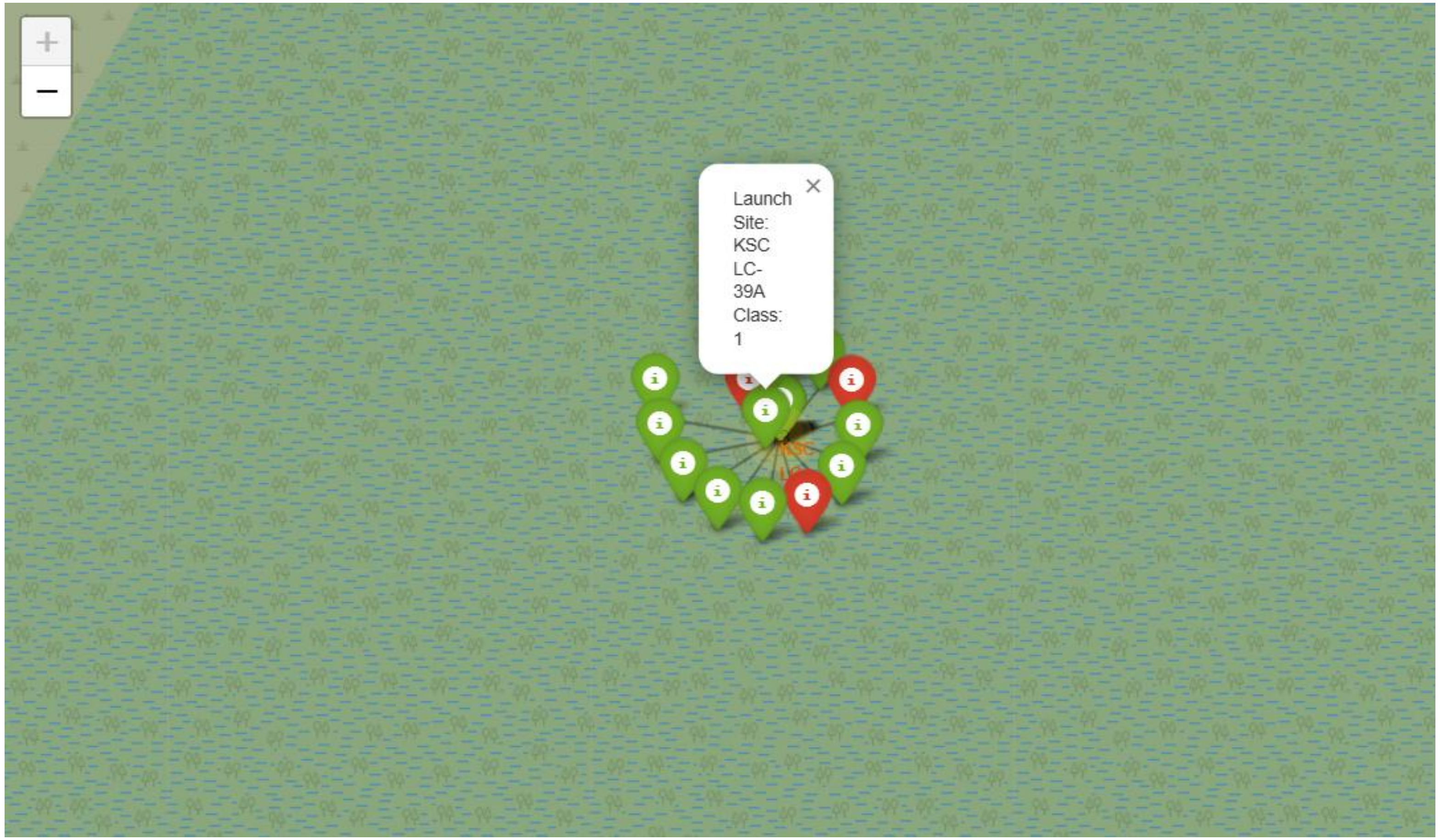


All launch sites are in the Northern Hemisphere and relatively close to the Equator (between ~26° and ~35° N), in the South of the US. All launch sites are located on coastlines, which is safer than other locations, because if anything goes wrong during ascent, debris can fall into the ocean instead of on populated areas.

Launch Outcomes Map (I)

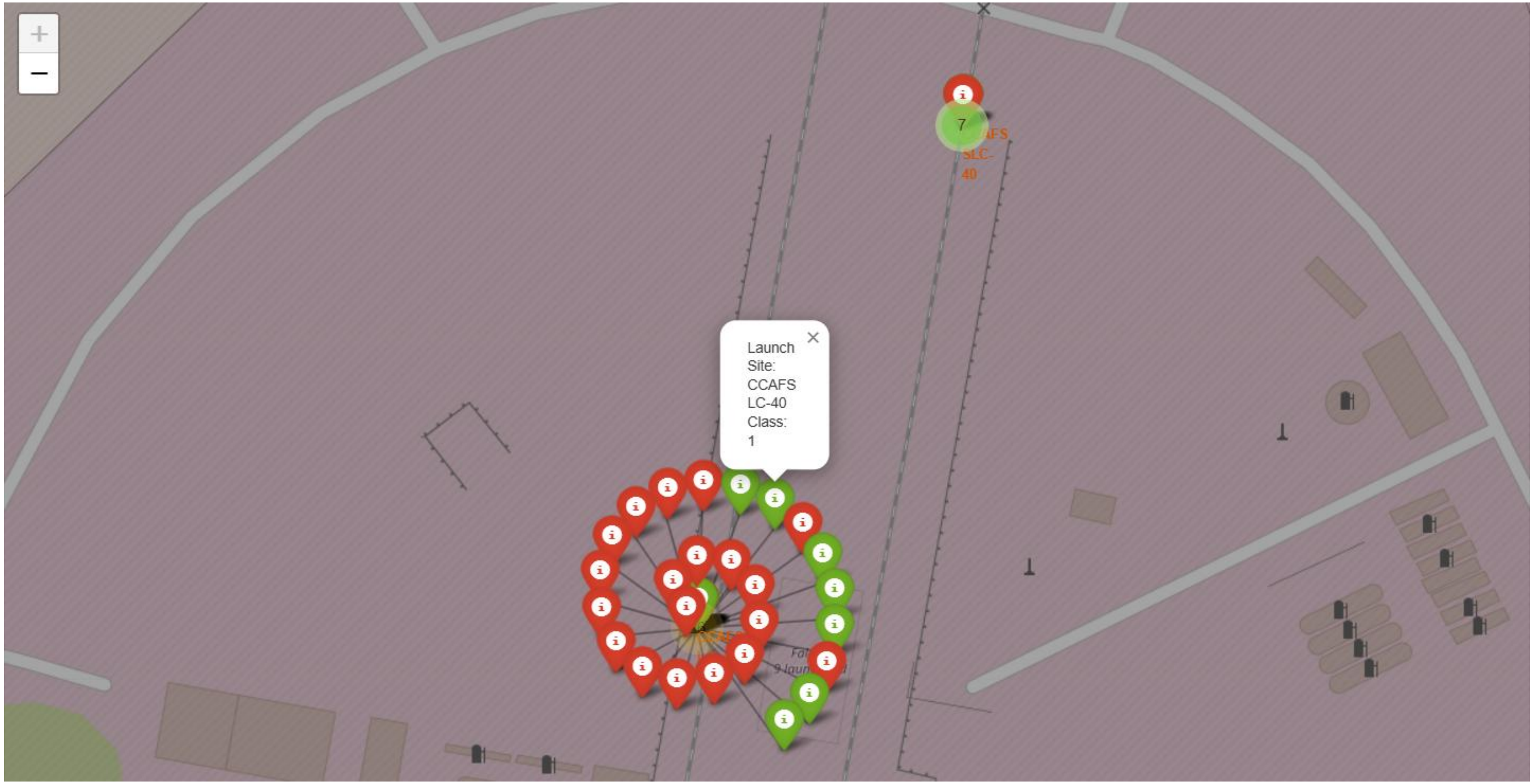


Launch Outcomes Map (II) (Zoom-in)



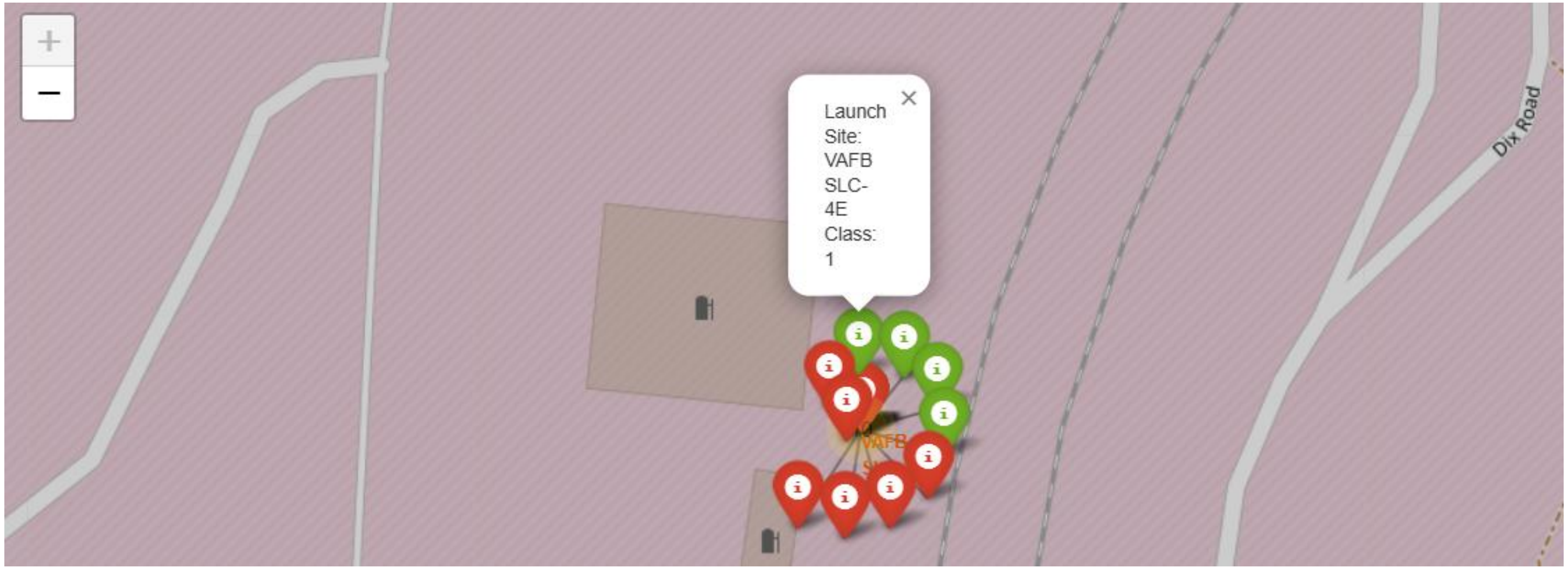
KSC LC-39A has the highest concentration of successful outcomes.

Launch Outcomes Map (III) (Zoom-in)



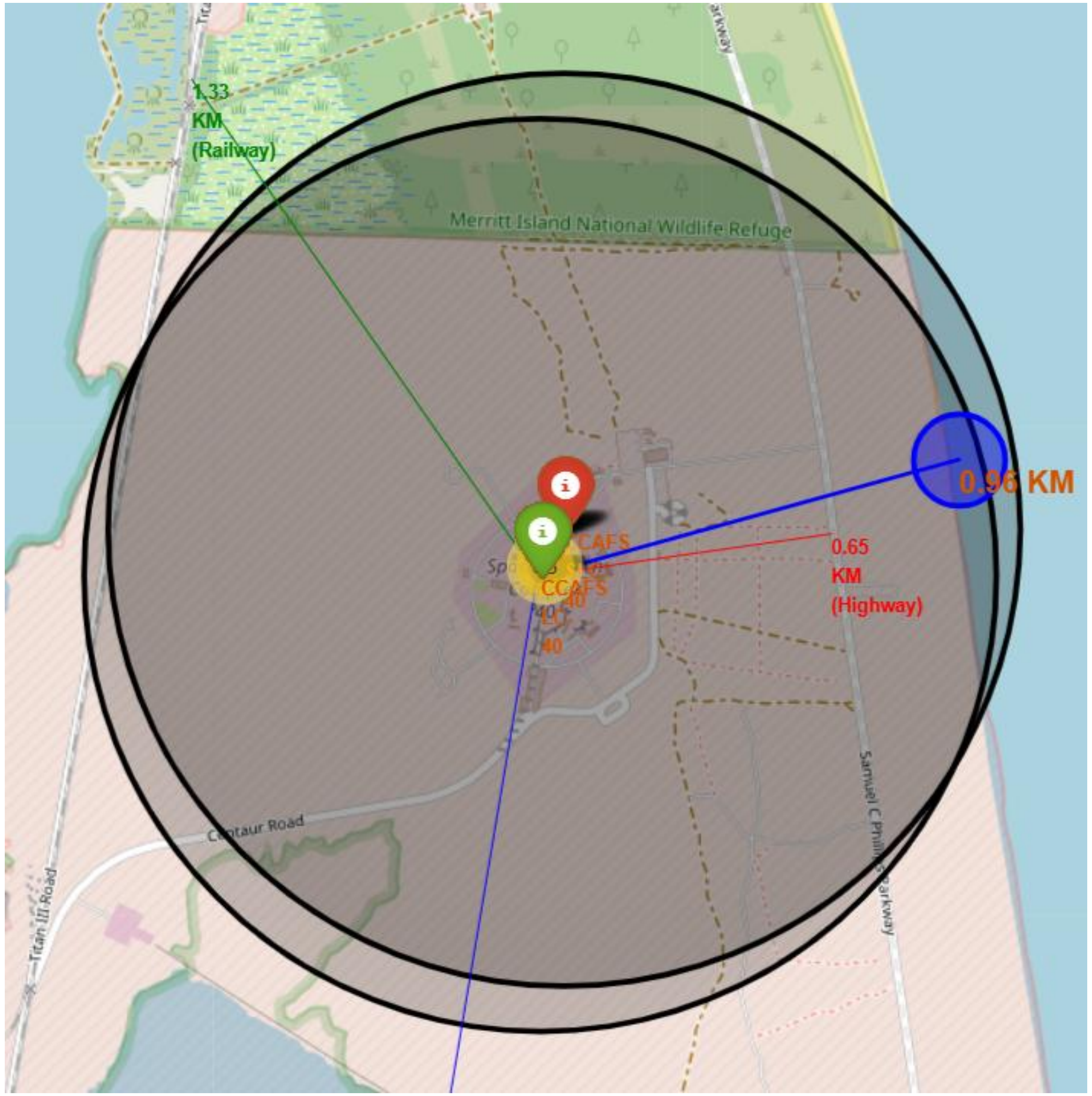
CCAFS LC-40 and CCAFS SLC-40 have higher concentrations of failed outcomes.

Launch Outcomes Map (IV) (Zoom-in)



VAFB SLC-4E has a higher concentration of failed outcomes.

Launch Site Proximity Insights



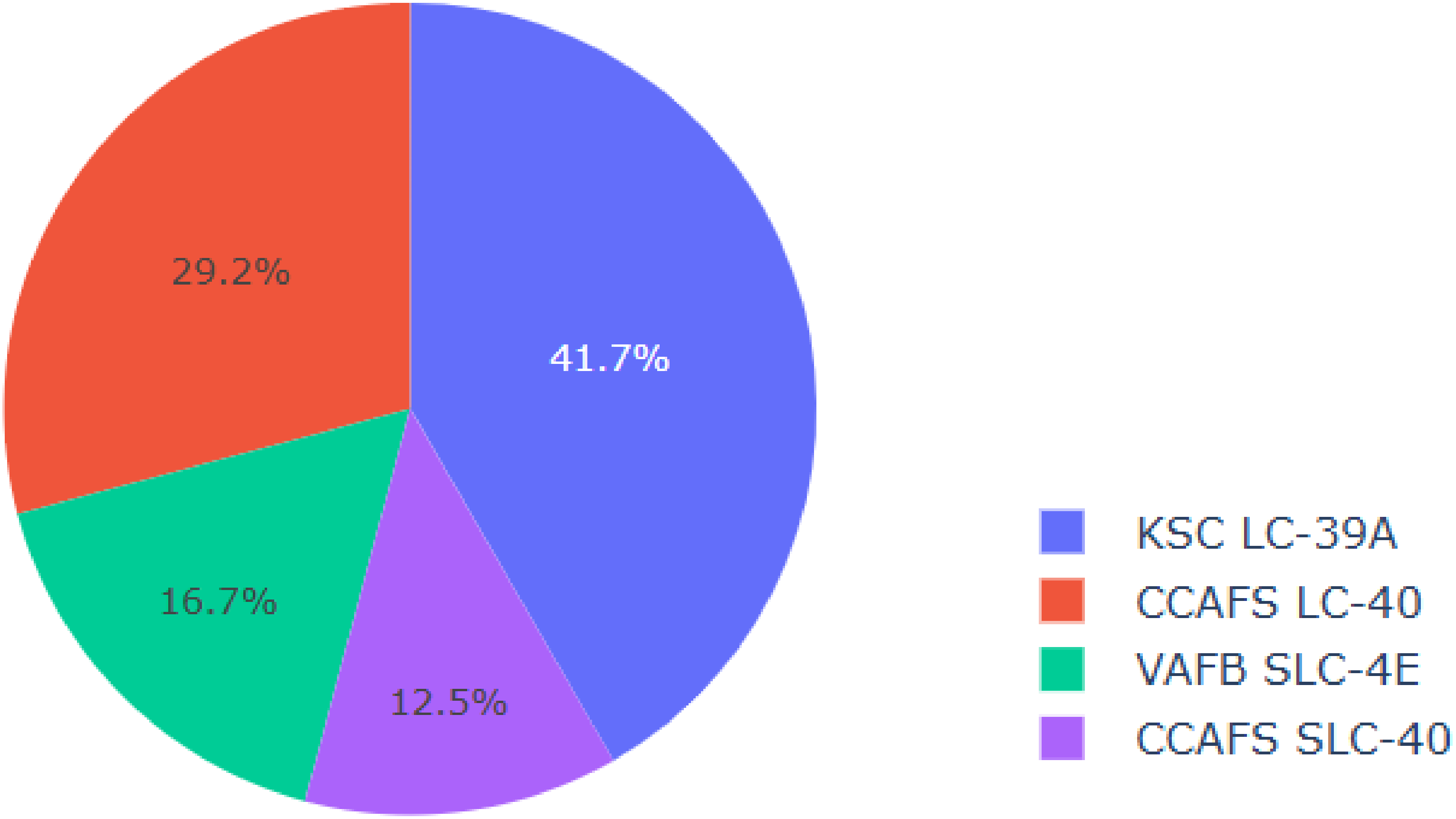
The launch site CCAFS LC-40 is near the coastline and major transport routes, while also far from urban areas.
The site is 0.96 km from the coastline, which supports the need for sea-based landing zones. It is also 0.65 km from a highway and 1.33 km from a railway, which facilitates logistics.
Beyond the scope of the graph, the nearest city is 17.12 km away.

Section 4

Build a Dashboard with Plotly Dash

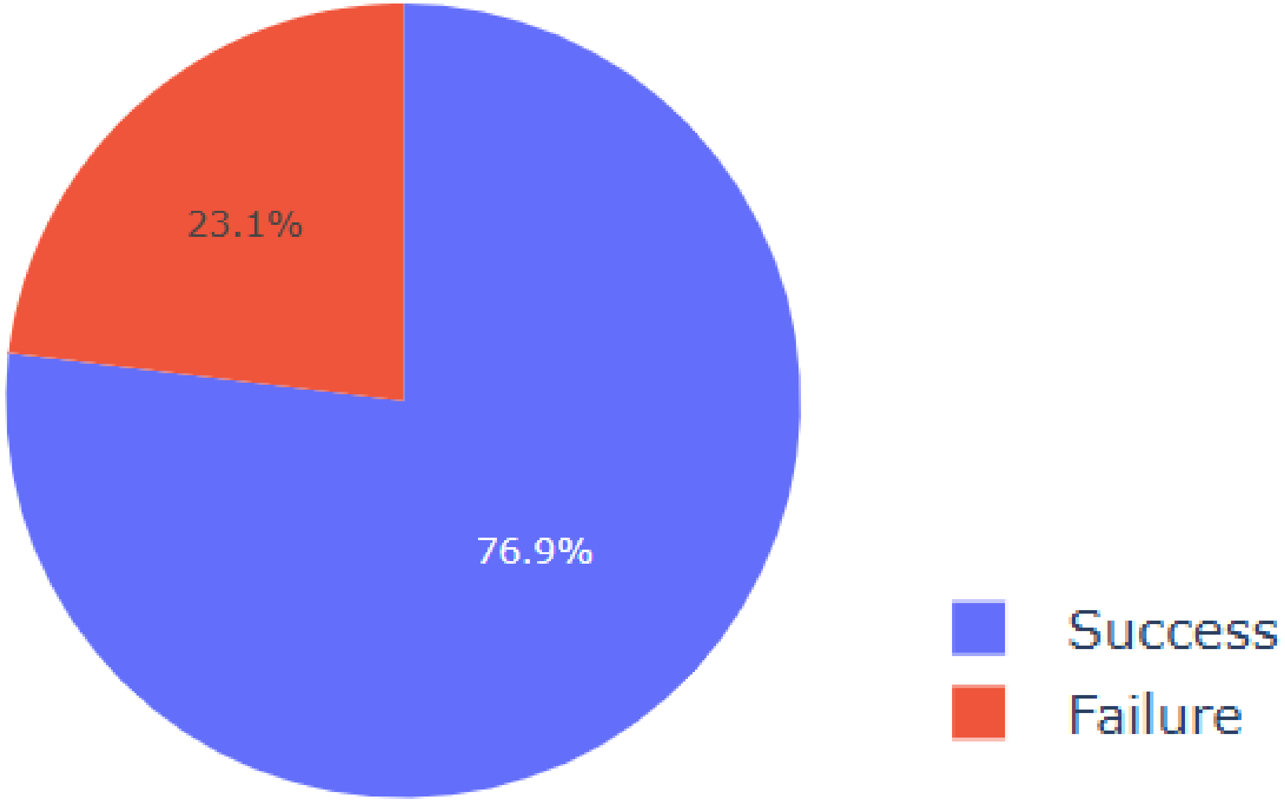
Total Successful Launches by Site

This pie chart shows the total successful launches by site. KSC LC-39A has the highest number of successes.



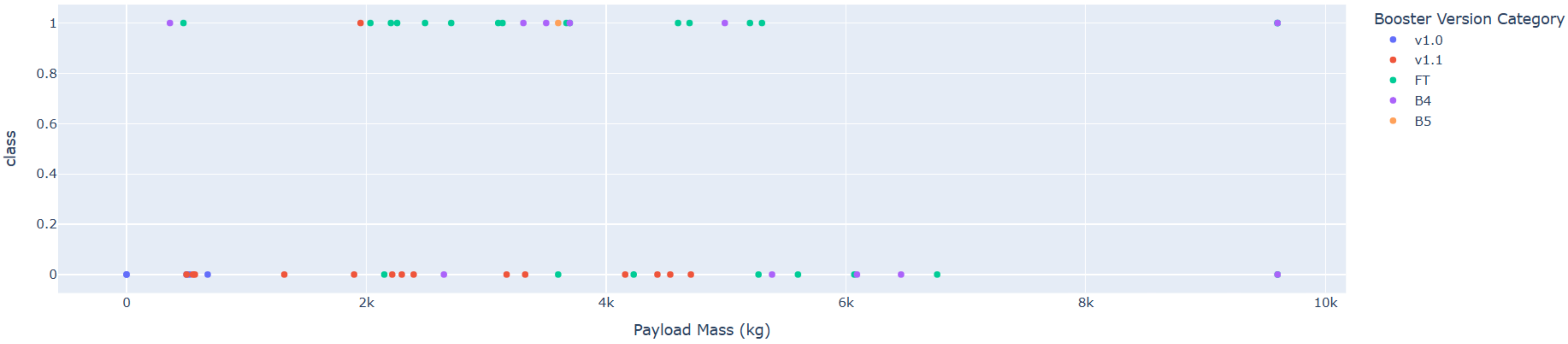
Launch Site with Highest Success Rate

This pie chart shows the proportion of total successful launches for site launch KSC LC-39A, the one with the highest success rate: 76.9%



Payload vs. Outcome (I)

This scatterplot shows the correlation between payload mass up to 10000 kg. and success for all sites:



Light payloads (<2000 kg) and heavy payloads (>6000 kg) show more frequent failures.

Payload vs. Outcome (II)

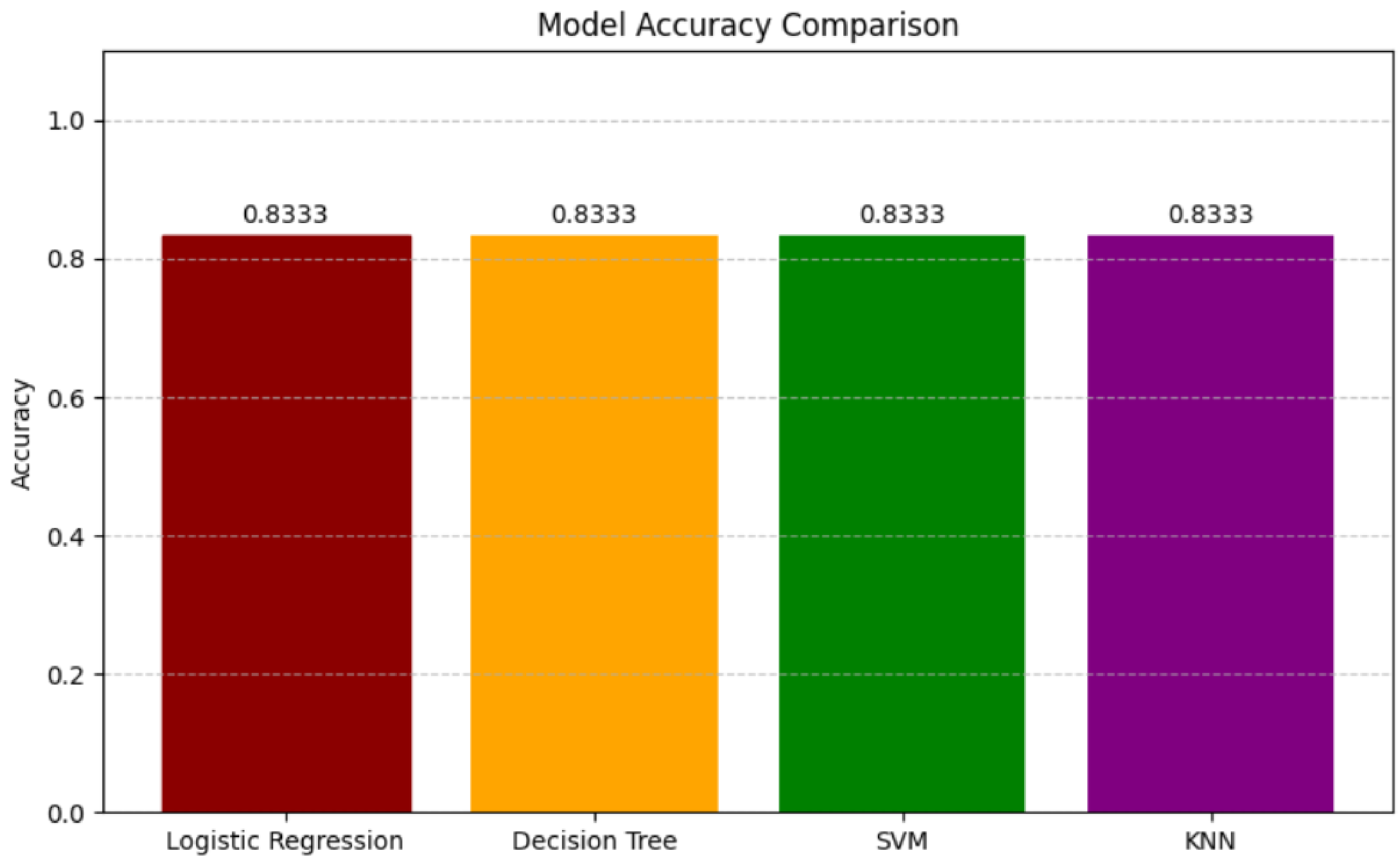
This scatterplot shows the correlation between payload mass up to 7500 kg and success for all sites.

Mid-range payloads (3000–4000kg) have the highest rate of success.
Booster version B5 had the highest rate of success.

Section 5

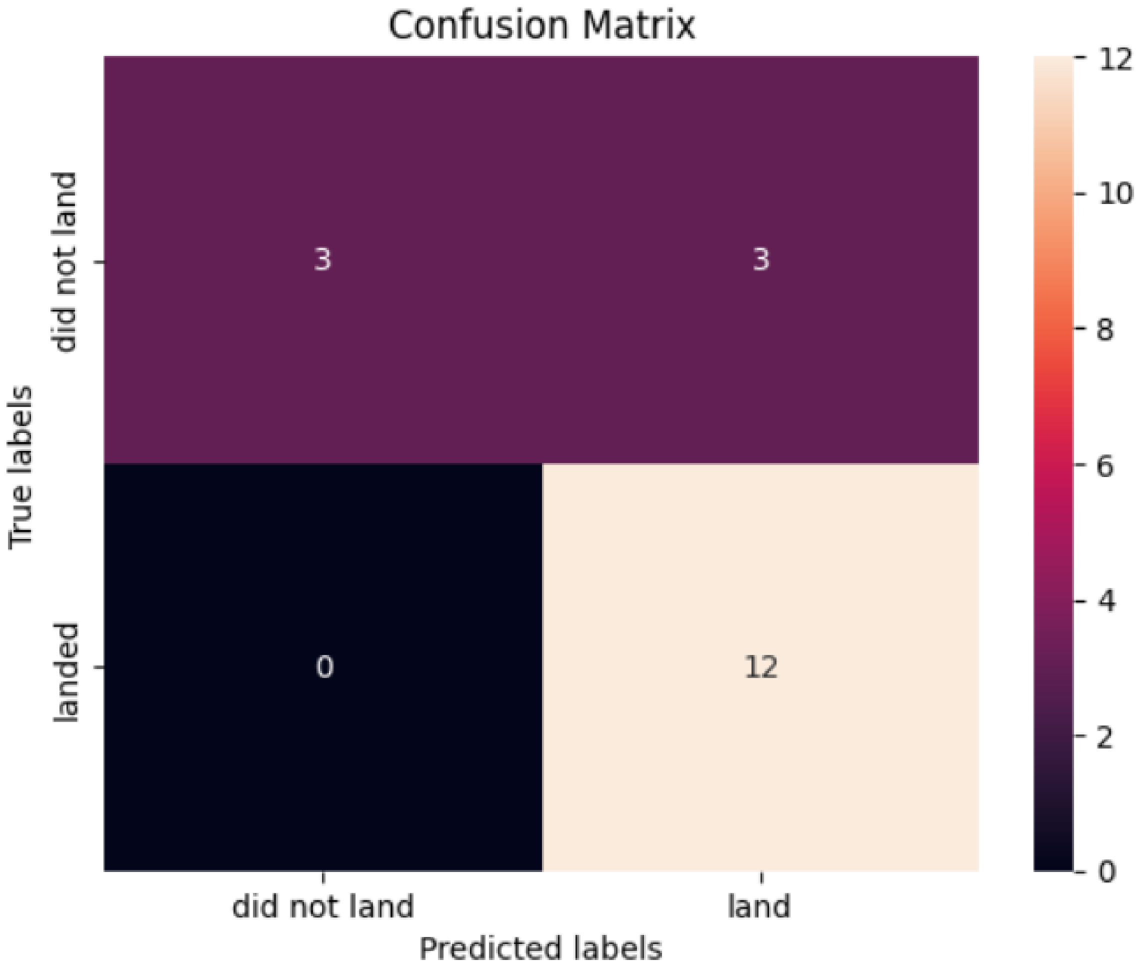
Predictive Analysis (Classification)

Classification Model Accuracy



The best performing model is the Logistic Regression, with an accuracy of 0.8333.

Logistic Regression Confusion Matrix



This is the confusion matrix for the Logistic Regression, the best performing classification model.

The matrix shows very few false positives/negatives, and the high count of true positives suggests that the model is especially good at identifying missions likely to succeed.

Conclusions

- KSC LC-39A recorded the highest number of successful launches among all sites. This highlights its strategic importance for Falcon 9 missions.
- Payloads in the 3000–4000 kg range demonstrated the highest success rate. This suggests that the optimal performance for Falcon 9 with first-stage recovery occurs under moderate payload conditions.
- Booster version F9 B5 outperformed all others in landing success.
- Logistic Regression was the best-performing model in predicting landing success. After training and tuning multiple classifiers, including SVM, Decision Tree, and KNN, Logistic Regression achieved the highest test accuracy. This model provided a balance between simplicity and predictive power, making it ideal for binary classification of mission outcomes.

Appendix A. Feature Engineering with Python

- After data collection and wrangling, I select the features that will be used to predict launch outcomes:

```
features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial']]
features.head()
```

- I then create dummy variables for categorical columns by applying OneHotEncoder:

```
features_one_hot = features_one_hot.astype('float64')
features_one_hot.dtypes.head()
```

- Finally, I convert the numeric columns to float64:

```
features_one_hot = pd.get_dummies(features,
                                  columns=['Orbit', 'LaunchSite', 'LandingPad', 'Serial'])
features_one_hot.head()
```

Appendix B. Sample SQL for EDA (I)

- Total payload mass carried by boosters launched by NASA (CRS):

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

- Names of the boosters that have success outcomes in drone ships and a payload mass between 4000 and 6000 kg.:

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
  AND "Payload_Mass__kg_" > 4000
  AND "Payload_Mass__kg_" < 6000;
```

Appendix B. Sample SQL for EDA (II)

- List of the booster versions that have carried the maximum payload mass:

```
%%sql
SELECT "Booster_Version", "Payload_Mass__kg_"
FROM SPACEXTABLE
WHERE "Payload_Mass__kg_" = (
    SELECT MAX("Payload_Mass__kg_")
    FROM SPACEXTABLE
);
```

- Ranking of the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order:

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

Appendix C. GridSearchCV Model Parameters

- Logistic Regression:

tuned hpyerparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs’}

- Support Vector Machine:

tuned hpyerparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid’}

- Decision Tree:

tuned hpyerparameters :(best parameters) {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random’}

- K-Nearest Neighbors:

tuned hpyerparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

Appendix D. Model Accuracy Score

- This is the accuracy of each model for the test data, calculated using the score method:

Model	Test Accuracy
Logistic Regression	0.8333
Support Vector Machine	0.8333
Decision Tree	0.8333
K-Nearest Neighbors	0.8333

Appendix E. GitHub and Sources

- GitHub repository containing Jupyter Notebooks and dashboard:

https://github.com/LuciaPavon/capstone-project

- Data sources: Wikipedia, SpaceX API, IBM Cloud-hosted CSVs.

Thank you!