IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Magdalena Ivanova
27/08/2025

Outline

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

Executive Summary

Summary of methodologies

Data sources: Combined SpaceX REST API pulls with targeted web scraping to enrich mission details.
Data prep: Used pandas/NumPy to tidy fields, handle nulls, tally missions by site/orbit/outcome, and create a binary “landing success” label for modeling.
Exploration: Verified aggregates with SQL and built visuals with Seaborn/Matplotlib; mapped launch sites in Folium and built an interactive view in Plotly Dash.
Modeling: Trained several classifiers (e.g., Decision Tree, Logistic Regression, KNN, SVM), tuned hyperparameters, and evaluated on a hold-out set with accuracy and confusion matrices.

Summary of all results

Best site performance: Kennedy Space Center posted the highest landing success rate (~76.9%).
Heaviest successful payload: Achieved at Cape Canaveral SLC.
Payload sweet spot: Most successful landings fell in the ~2,000–4,000 kg range.
Top booster: The Falcon 9 FT variant recorded the most successful landings.
Predictive models: A tuned Decision Tree led training accuracy (~88.8%), but on the test split all four models tied at ~83.3% accuracy.
Error pattern: Confusion matrices were similar across models, with errors skewing toward **false positives** more than **false negatives**

Introduction

Project background & context

SpaceX’s strategy for driving down launch costs centers on first-stage reusability. Falcon 9 missions are marketed at roughly $62M per launch—about a third of typical competitors—an economics story that only works if boosters routinely land and fly again. With many successful recoveries, especially from newer Falcon 9 variants, the approach looks promising. This project examines the data behind those outcomes.

Questions I set out to answer

- Which launch sites deliver the highest booster-landing success rates?
- Which orbits are most commonly associated with successful recoveries?
- Is there a payload “sweet spot” where success is more likely?
- Do model features (site, orbit, payload, booster version) reliably predict landing success for future missions?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Spacex API.
  - Webscraping.
- Perform data wrangling
  - Convert collected data to generate binary target variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into training and test sets.
  - Determine optimal parameters and test models for accuracy.

Data Collection

- The main chunk of data was fetched from the official SpaceX API.
- A script to automatically collect some additional facts from a detailed Wikipedia list of all their launches was also used

Data Collection - SpaceX API

- **How I pulled the data:** Issued REST requests to the SpaceX API using Python (requests).

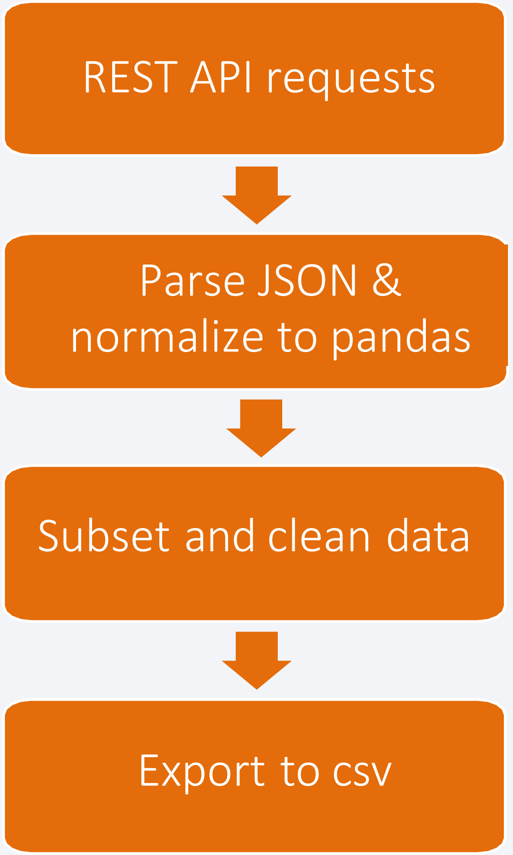
- **How I structured it:** Parsed the JSON and flattened nested fields into a tidy pandas dataframe.

- **What I kept & cleaned:** Selected analysis-ready columns (date, site, orbit, booster, payload, landing outcome), applied small helper functions to extract embedded attributes, fixed types, checked for missing values/duplicates, and standardized labels (including a binary class target).

- **Output:** Saved the curated dataset to CSV for the later EDA, SQL, mapping, dashboard, and modeling steps.

- **References:**

1. API notebook with code
2. Cleaned dataset (CSV)


REST API requests

Parse JSON & normalize to pandas

Subset and clean data

Export to csv

Data Collection - Scraping

•**Where I scraped:** Retrieved the Falcon 9 / Falcon Heavy launch tables from Wikipedia with a simple HTTP GET.

•**How I parsed:** Converted the HTML to a BeautifulSoup object and identified the table headers to define column names.

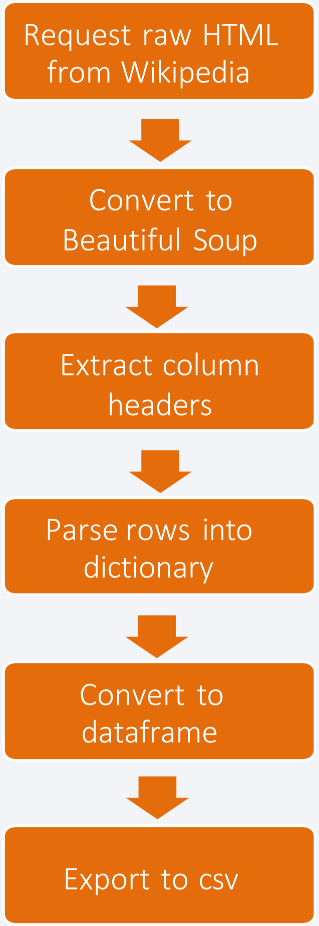
•**How I extracted rows:** Walked each

•**How I structured & saved:** Built a tidy pandas DataFrame and exported the cleaned result to CSV for downstream EDA, SQL, Folium, Dash, and modeling.

External references

Web scraping notebook

Scraped dataset (CSV)



Request raw HTML from Wikipedia

Convert to Beautiful Soup

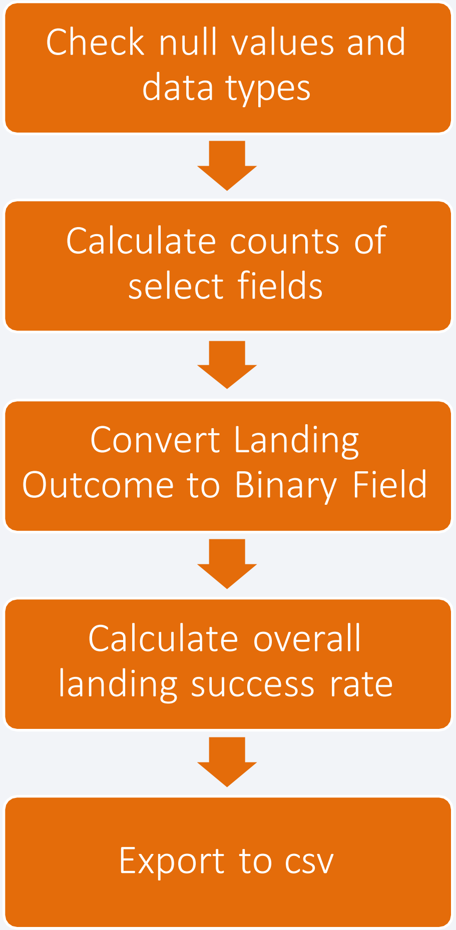
Extract column headers

Parse rows into dictionary

Convert to dataframe

Export to csv

Data Wrangling


Check null values and data types
Calculate counts of select fields
Convert Landing Outcome to Binary Field
Calculate overall landing success rate
Export to csv

- **Schema & quality checks:** Verified data types, scanned for missing values, and fixed minor inconsistencies.
- **Quick profiling:** Counted distributions for Launch Site, Orbit, and Landing Outcome to spot issues and understand balance.
- **Target engineering:** Converted landing outcome into a binary label
- **Baseline metric:** Calculated the overall booster-landing success rate.
- **Output:** Saved the cleaned, analysis-ready dataset to CSV.

References

[Data wrangling notebook]

[Cleaned dataset (CSV)]

EDA with Data Visualization

- EDA notebook (code + visuals + notes)
- Model-ready dataset

Plots used	Why I used them
Scatter plot: Launch number vs. launch site (colored by landing success)	To explore correlations between launch sites, orbits, and outcomes
Scatter plot: Payload mass vs. launch site (colored by landing success)	To check how payload mass influences the likelihood of recovery
Bar chart: Success rate grouped by orbit type	To identify which orbits yield the best success rates
Scatter plot: Launch number vs. orbit (colored by landing success)	To observe improvements in landing reliability over time
Scatter plot: Payload mass vs. orbit (colored by landing success)	To highlight anomalies or unusual data points before modeling
Line plot: Change in overall success rate across time	To guide feature selection for predictive models

EDA with SQL

SQL queries were performed to find:

- Listed the distinct launch sites in the dataset.
- Pulled five sample launches where the site name starts with “CCA”.
- Computed the total payload mass for missions that involved NASA.
- Calculated the average payload for Falcon 9 v1.1 flights.
- Found the earliest date of a successful ground-pad landing.
- Identified boosters with successful drone-ship landings carrying 4,000 - 6,000 kg payloads.
- Tallied the number of successes vs. failures.
- Returned the booster(s) that carried the maximum payload.
- Listed 2015 launches with failed drone-ship landings (outcome, booster, site, date).
- Ranked landing outcomes between 2010-06-04 and 2017-03-03.

Reference

- EDA with SQL notebook

Interactive Map with Folium

Map features	Why I added them
Circle markers placed at each launch site using latitude/longitude	Pinpoint launch pads and explore the surrounding geography
Rocket markers clustered by site (green = success, red = failure)	See successes vs. failures at a glance for each site
Popups/tooltips with site name and quick stats	Compare sites quickly through on-hover details
Lines with labels showing measured distances (km) to nearby points of interest	Quantify proximity to features like coastline, highways, or rail

Interactive Folium notebook

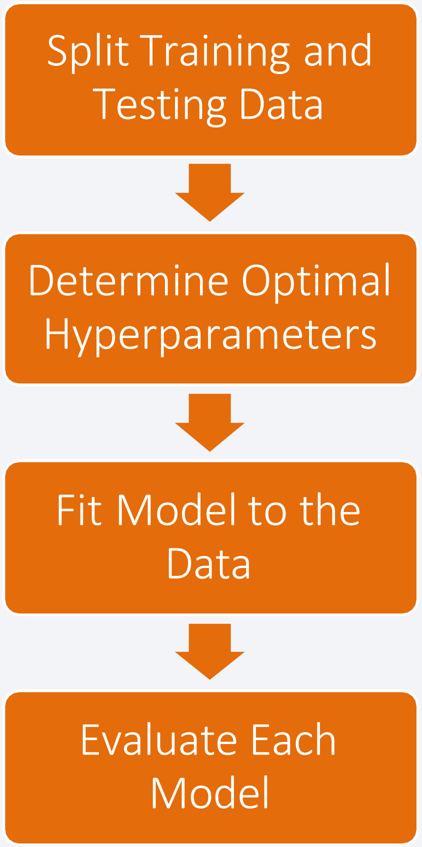
Dashboard with Plotly Dash

Dashboard features	Why these features
Pie view: overall split of successes vs. failures (and per-site when filtered).	Quickly see the overall success share and how it changes by site.
Scatter view: Payload mass vs. landing outcome, colored by booster/site.	Explore how payload ranges relate to landing outcomes.
Filters: site dropdown, optional orbit selector, and a payload range slider.	Let reviewers slice the data by site/orbit without touching code.
Live interactions: hover tooltips, click-to-filter behavior, and instant updates via Dash callbacks.	Support model building by revealing patterns/outliers interactively.

Dash app (source + callbacks)

Dataset used by the app

Predictive Analysis (Classification)



Split Training and Testing Data

Determine Optimal Hyperparameters

Fit Model to the Data

Evaluate Each Model

- **Feature prep:** Used payload mass as numeric and one-hot encoded categorical fields (site, orbit, booster). Scaled numeric features where appropriate.
- **Hold-out evaluation:** Performed a stratified train/test split to preserve class balance.
- **Model candidates:** Trained Logistic Regression, Decision Tree, KNN, and SVM.
- **Hyperparameter tuning:** Ran GridSearchCV (k-fold CV) for each model to find strong settings.
- **Metrics & comparison:** Evaluated on the test set using accuracy, precision/recall, and the confusion matrix (and ROC-AUC where available).
- **Selection & artifacts:** Chose the top performer based on test metrics and saved the model pipeline for reuse.

Reference

- Predictive analysis notebook

Results

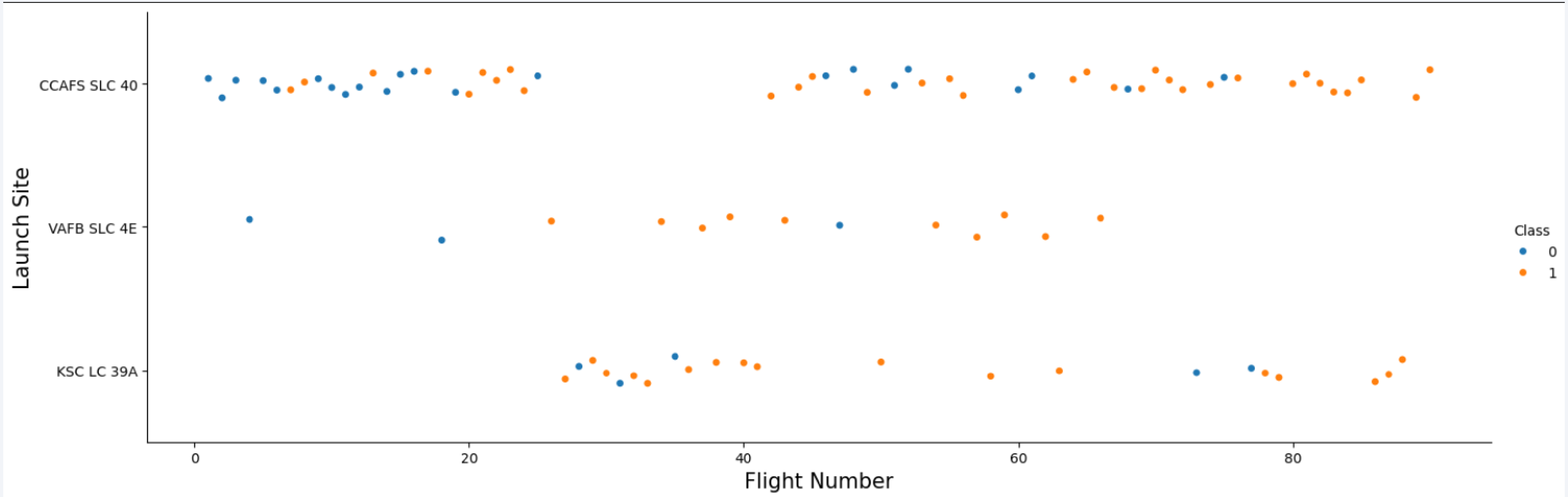
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

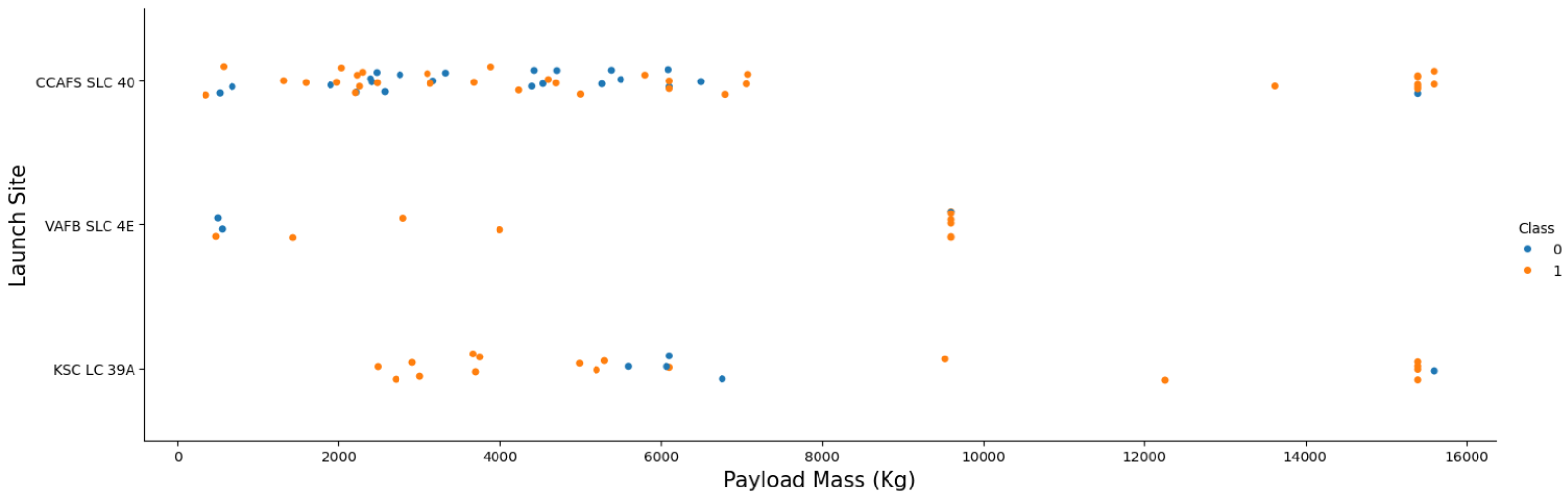
Flight Number vs. Launch Site

- Early missions with low flight numbers - mostly departing from Cape Canaveral SLC - belong to the program's first years and are dominated by unsuccessful landings (class = 0). In contrast, Kennedy Space Center started launching later and posts the highest recovery rate, likely benefiting from lessons learned after those early setbacks at the other pads.



Payload vs. Launch Site

- Payloads above 10,000 kg were flown mainly from Kennedy Space Center and Cape Canaveral SLC, and most of those missions ended in successful recoveries. Cape Canaveral SLC also holds the record for the heaviest successful payload - about 15,600 kg (see *Boosters Carrying Maximum Payload*). By contrast, many more launches used payloads below 8,000 kg, and their landing outcomes were much more mixed.

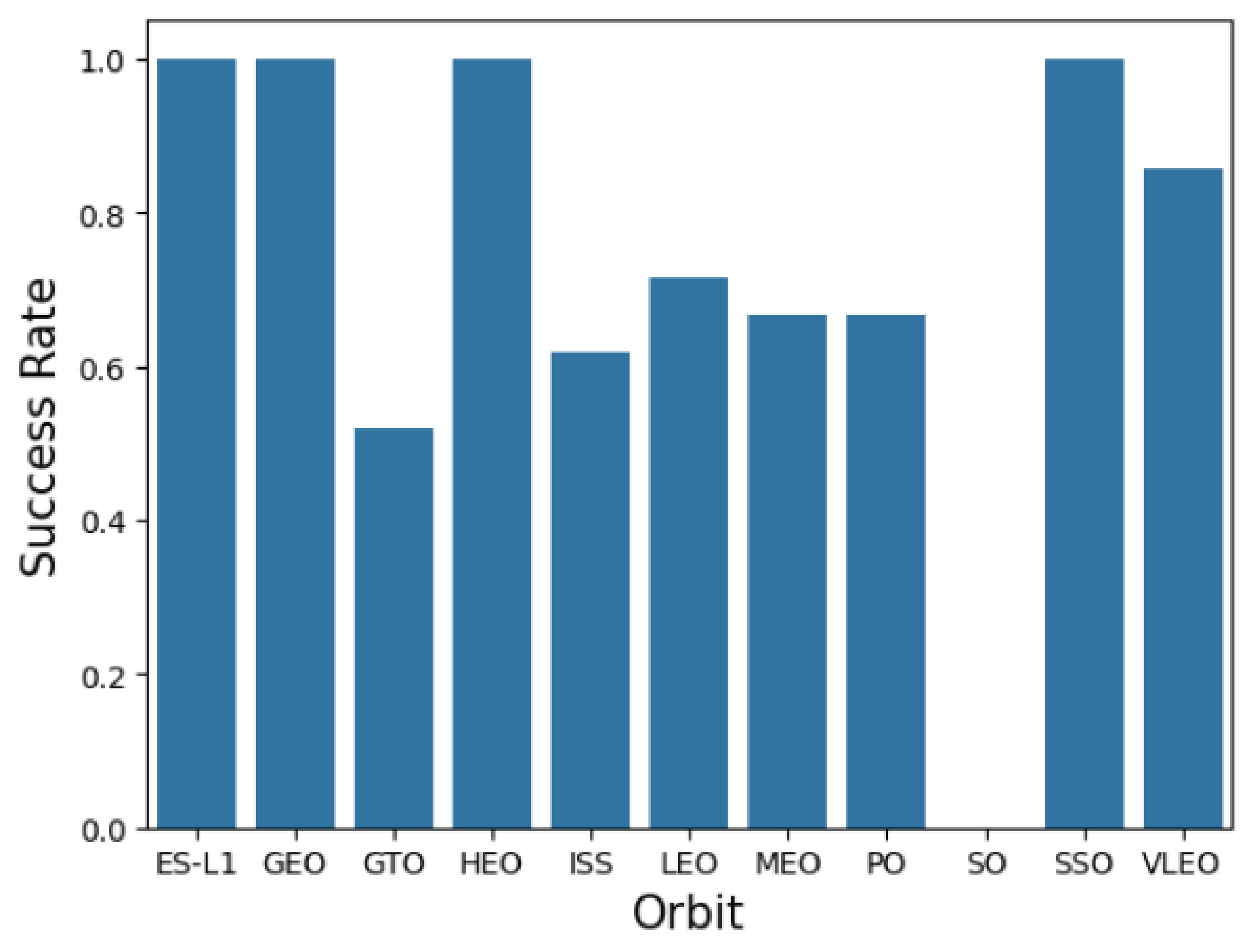


Success Rate vs. Orbit Type

ES-L1, GEO, HEO, and SSO all show a spotless record so far - 100% success - but three of those orbits are based on just a single attempt. The more meaningful standout is VLEO, which posts 86% success across 14 missions.

Counts of each orbit from initial data wrangling:

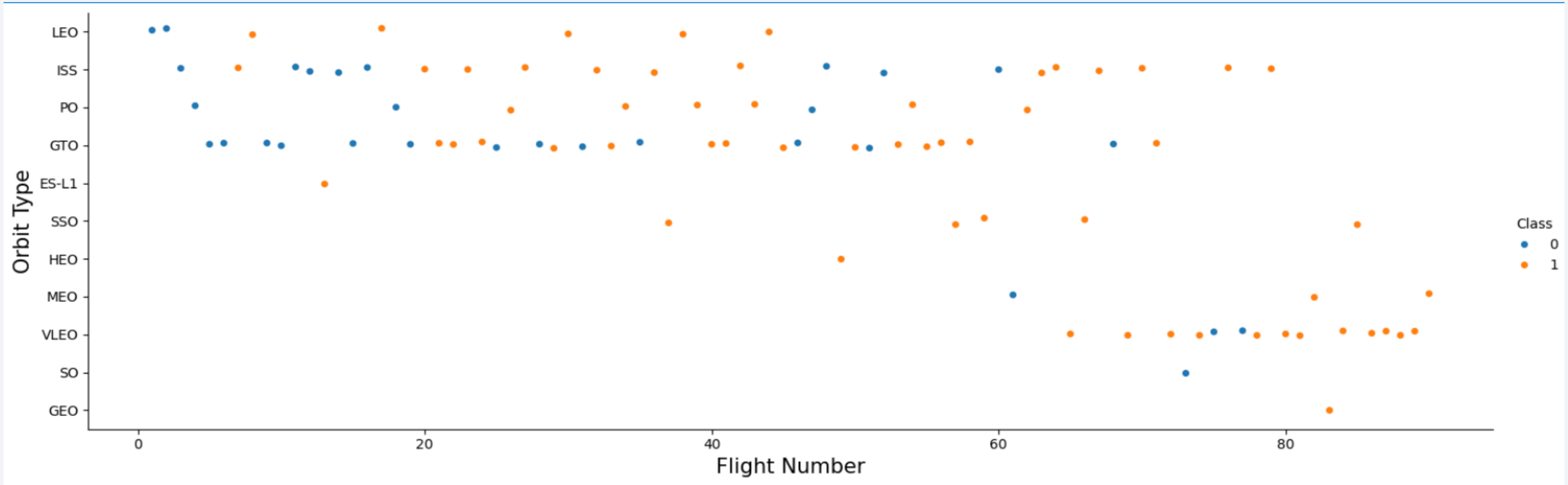
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3



Flight Number vs. Orbit Type

- A cluster of low flight numbers - roughly 2010–2013 - lines up with many failed landings (see Launch Success Yearly Trend).

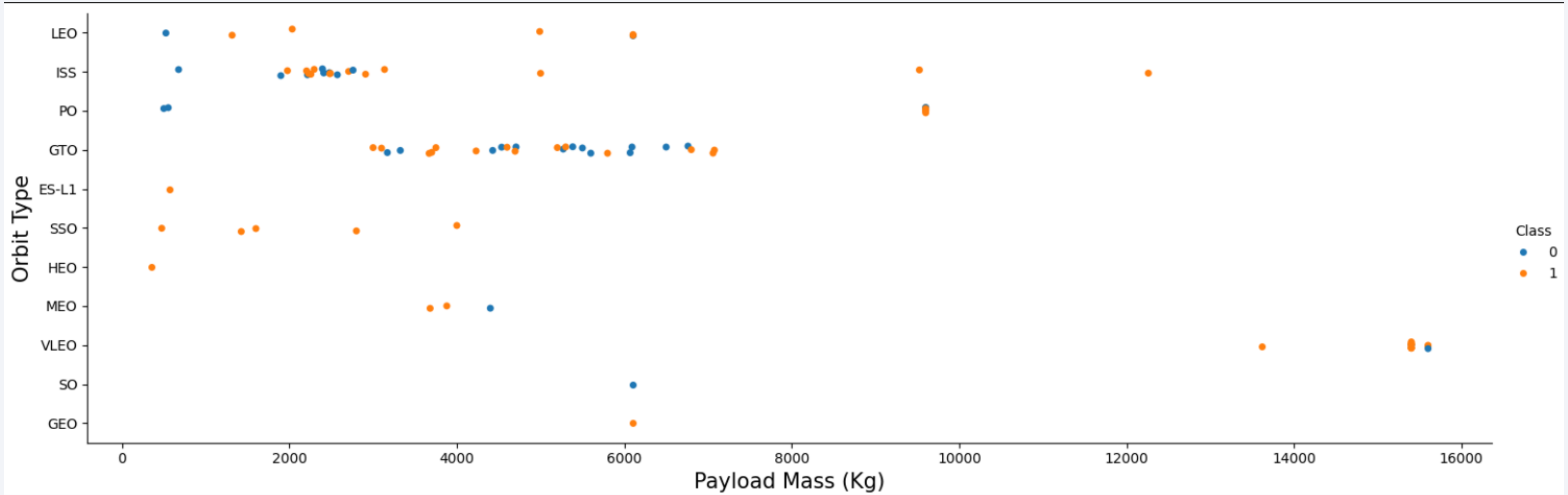
- As the program matured, LEO missions became steadily more reliable, with success rates rising alongside flight number.



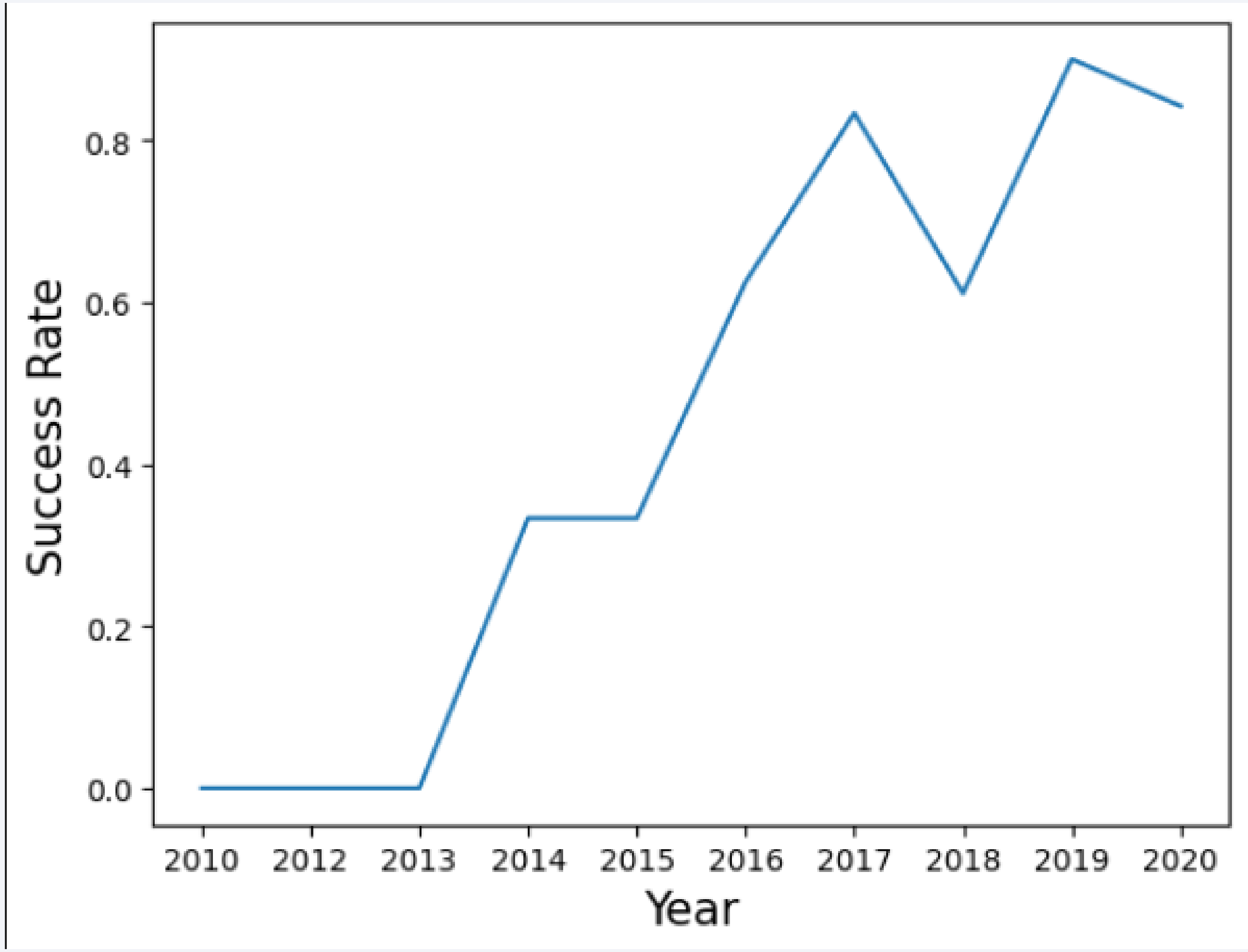
Payload vs. Orbit Type

- Most of the successful heavy-lift recoveries occurred on ISS, Polar (PO), and VLEO missions; no other orbit category in this dataset carried payloads above 8,000 kg.

- By contrast, every successful SSO landing involved a much lighter payload≈4,000 kg or less.



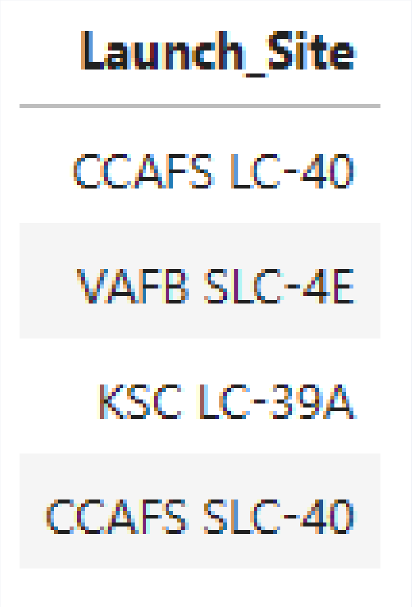
Launch Success Yearly Trend



- Landing reliability climbed steadily after 2013, peaking at about 90% in 2019.

All Launch Site Names

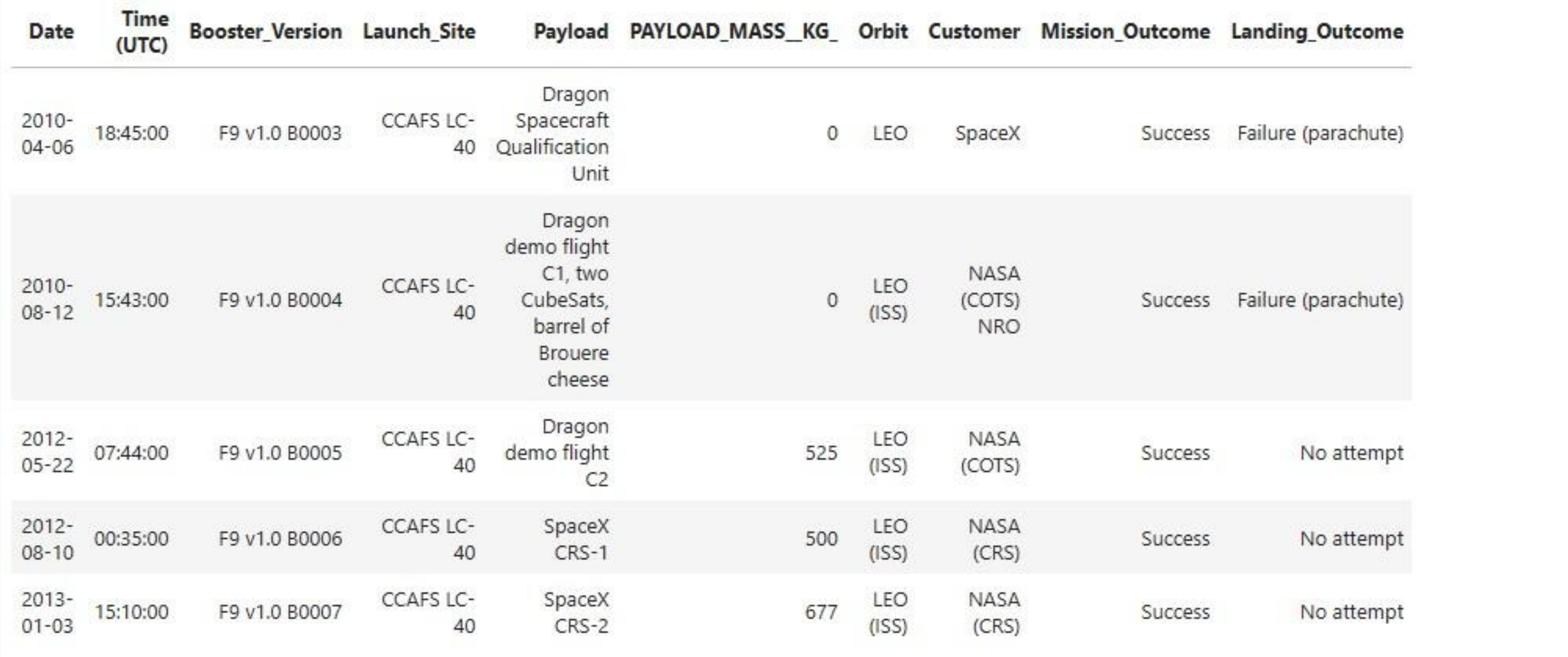
Cape Canaveral Launch Complex
Vandenberg Space Force Base
Kennedy Space Center
Cape Canaveral Space Launch Complex



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

Launch Site Names Begin with 'CCA'

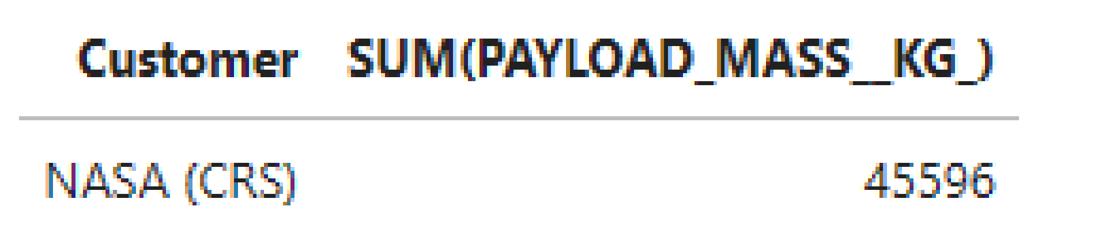
- First five launches from Cape Canaveral. All landings failed or were not attempted.



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

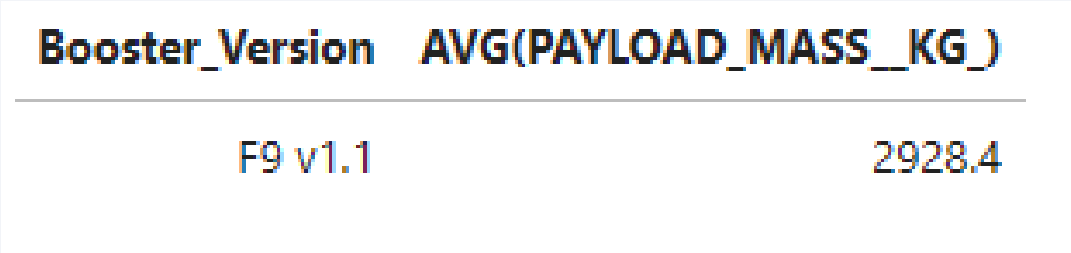
- The total mass of payloads carried by boosters from NASA was 45,596 Kg.



Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

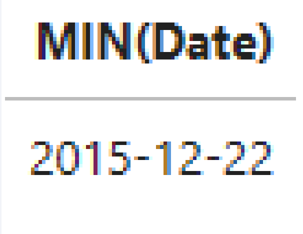
- The average payload mass carried by booster version F9 v1.1 was 2928.4 Kg.



Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

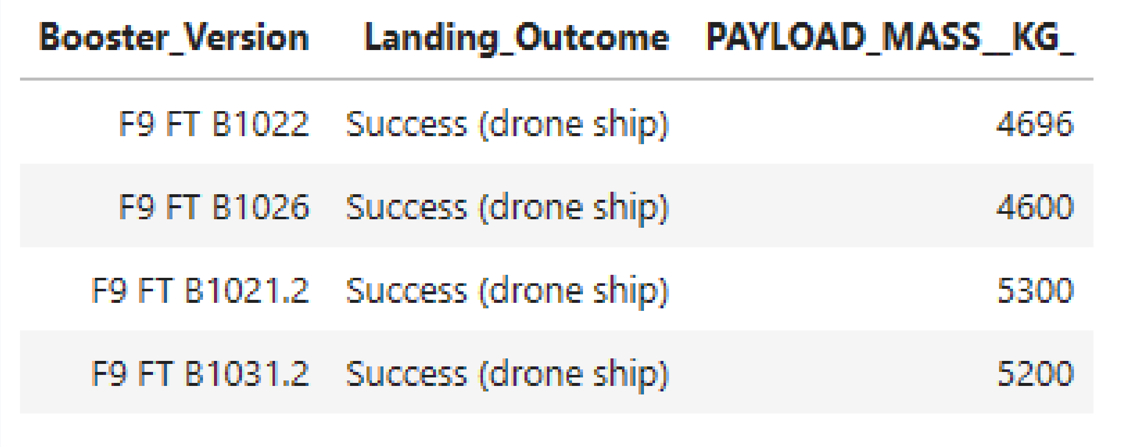
- The date of the first successful landing outcome on ground pad was December 22, 2015.



MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

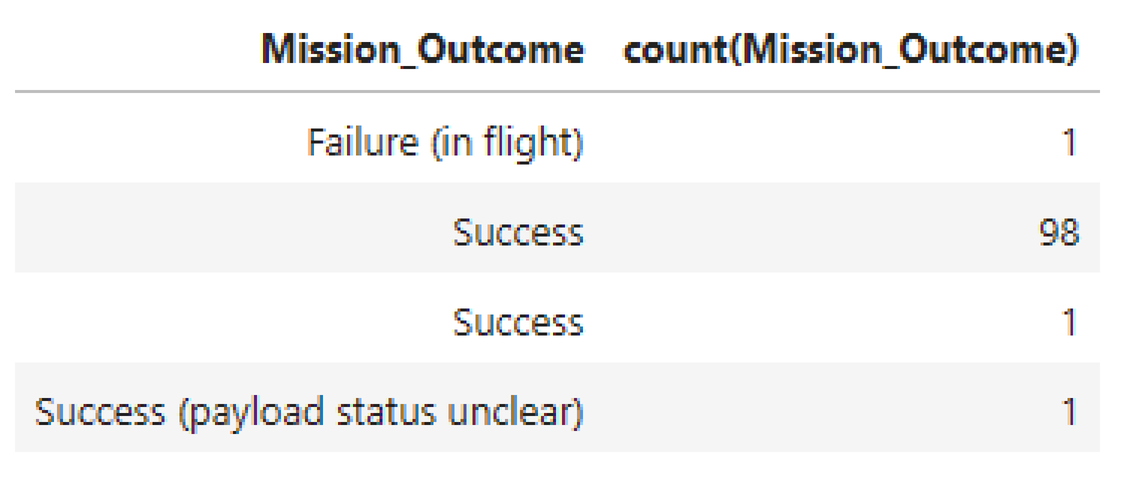
- Table with the names of boosters which have successfully landed on drone ship with payload mass between 4,000 and 6,000 Kg.



Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

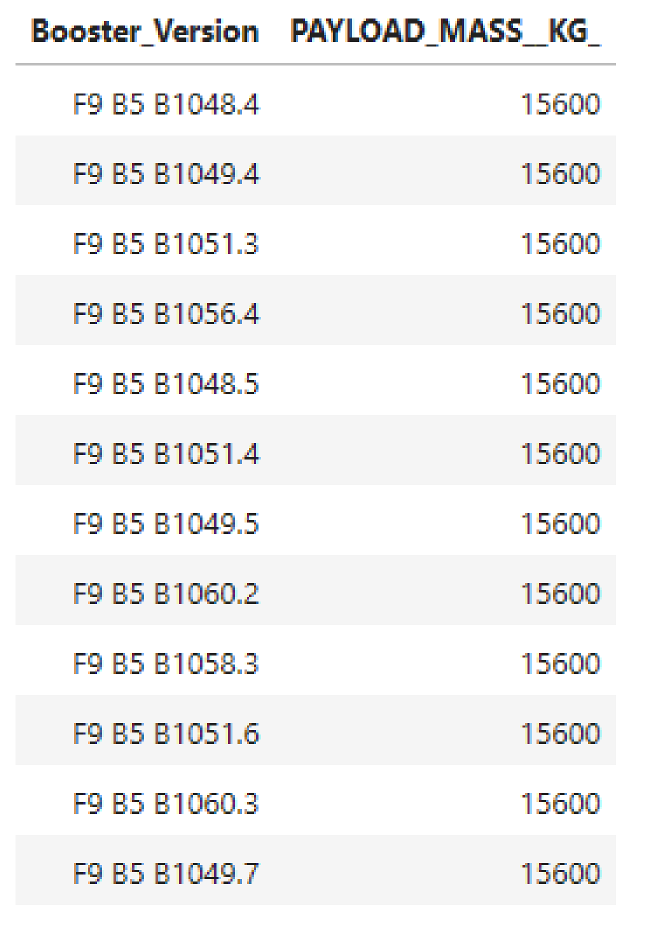
- Table showing the number of successful and failure mission outcomes.



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

Boosters Carrying Maximum Payload

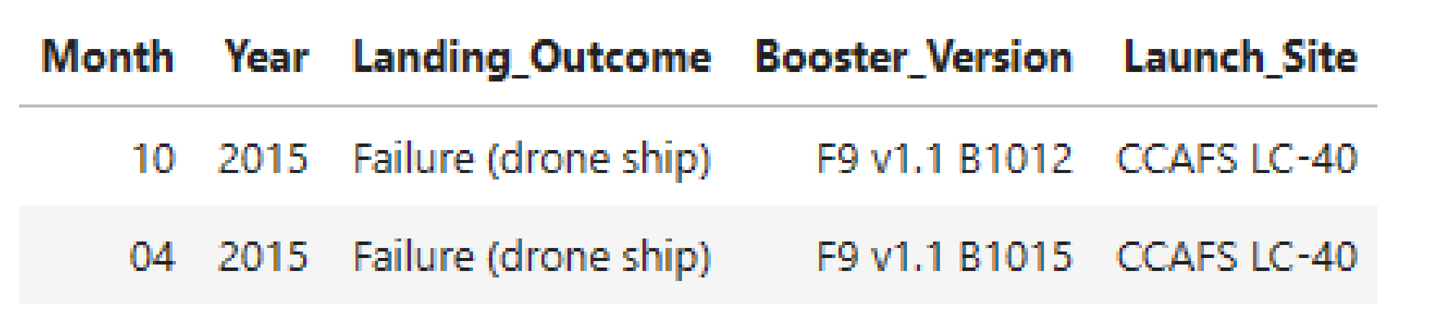
- Table with the names of boosters which have carried the maximum payload mass (15,600 Kg).



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

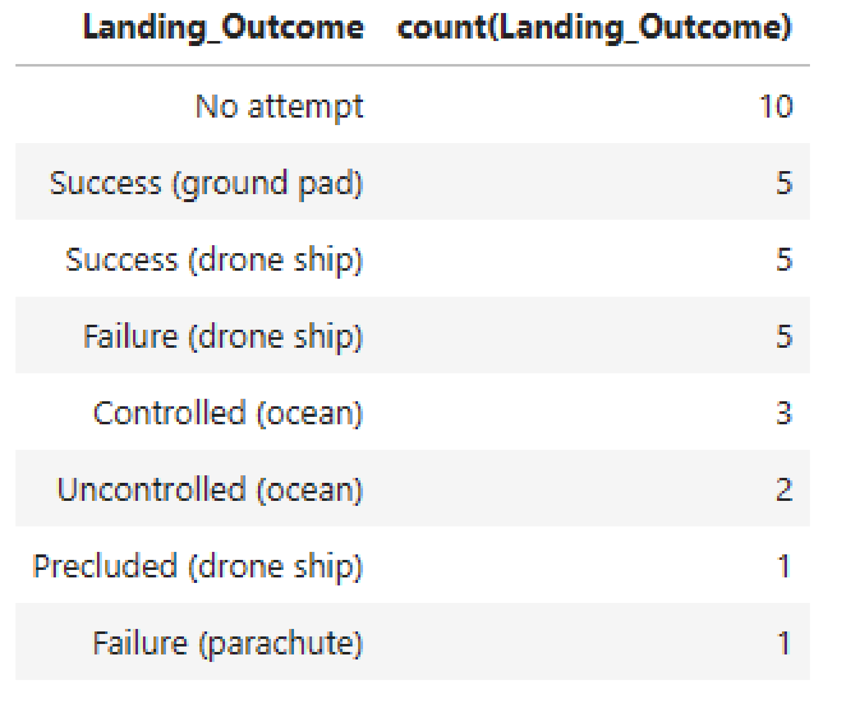
2015 Launch Records

- Table of launches with failed landings in drone ship during 2015 with their booster versions and launch site names.



Month	Year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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



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

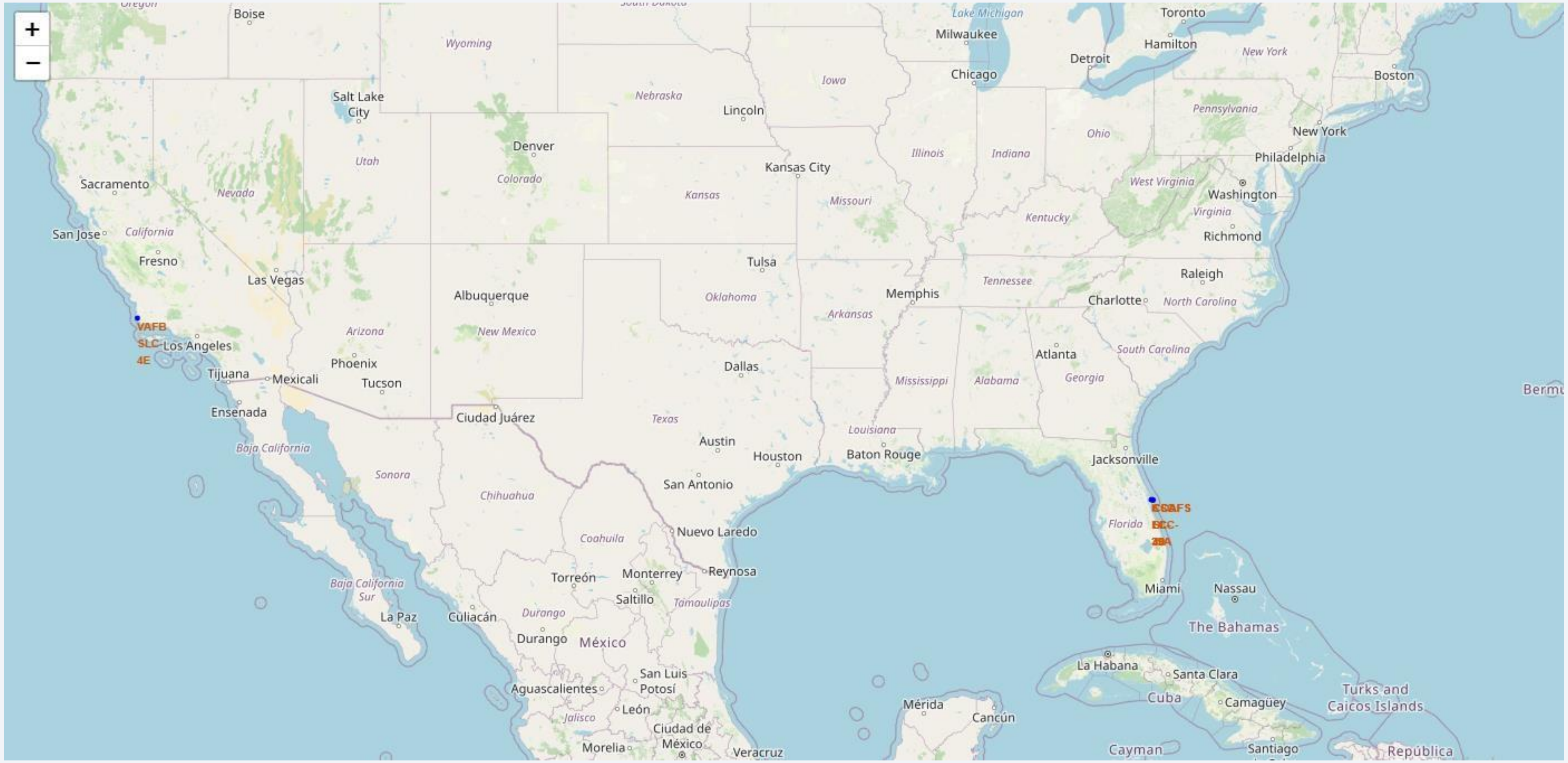
- Ranking of the count of landing outcomes between June 4, 2010 and March 3, 2017, in descending order

Section 3

Launch Sites Proximities Analysis

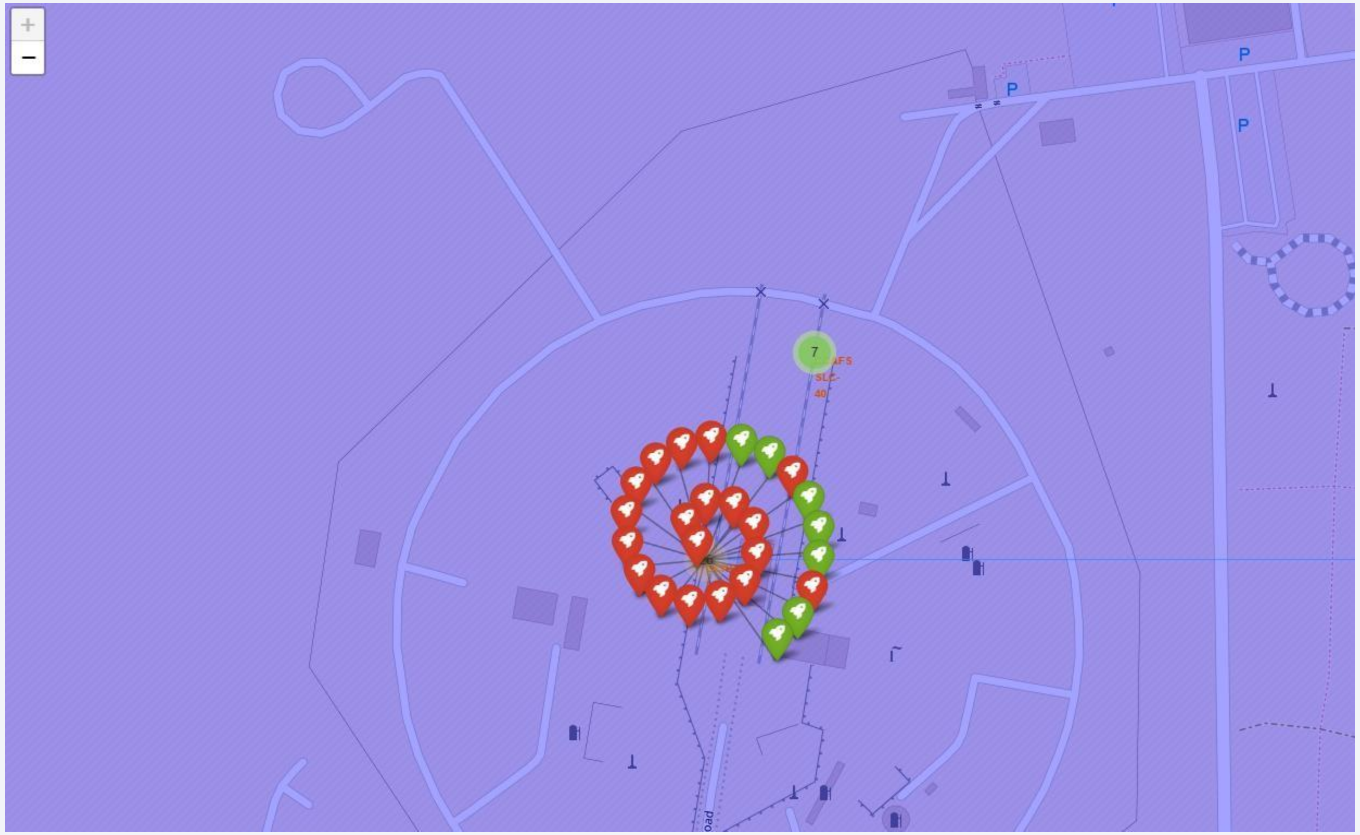
Launch Site Locations

All pads sit on the coastline. Vandenberg SFB is the only West Coast site (California, on the Pacific), while the remaining complexes - KSC and Cape Canaveral - are clustered on Florida’s Atlantic coast, just a short distance from one another.



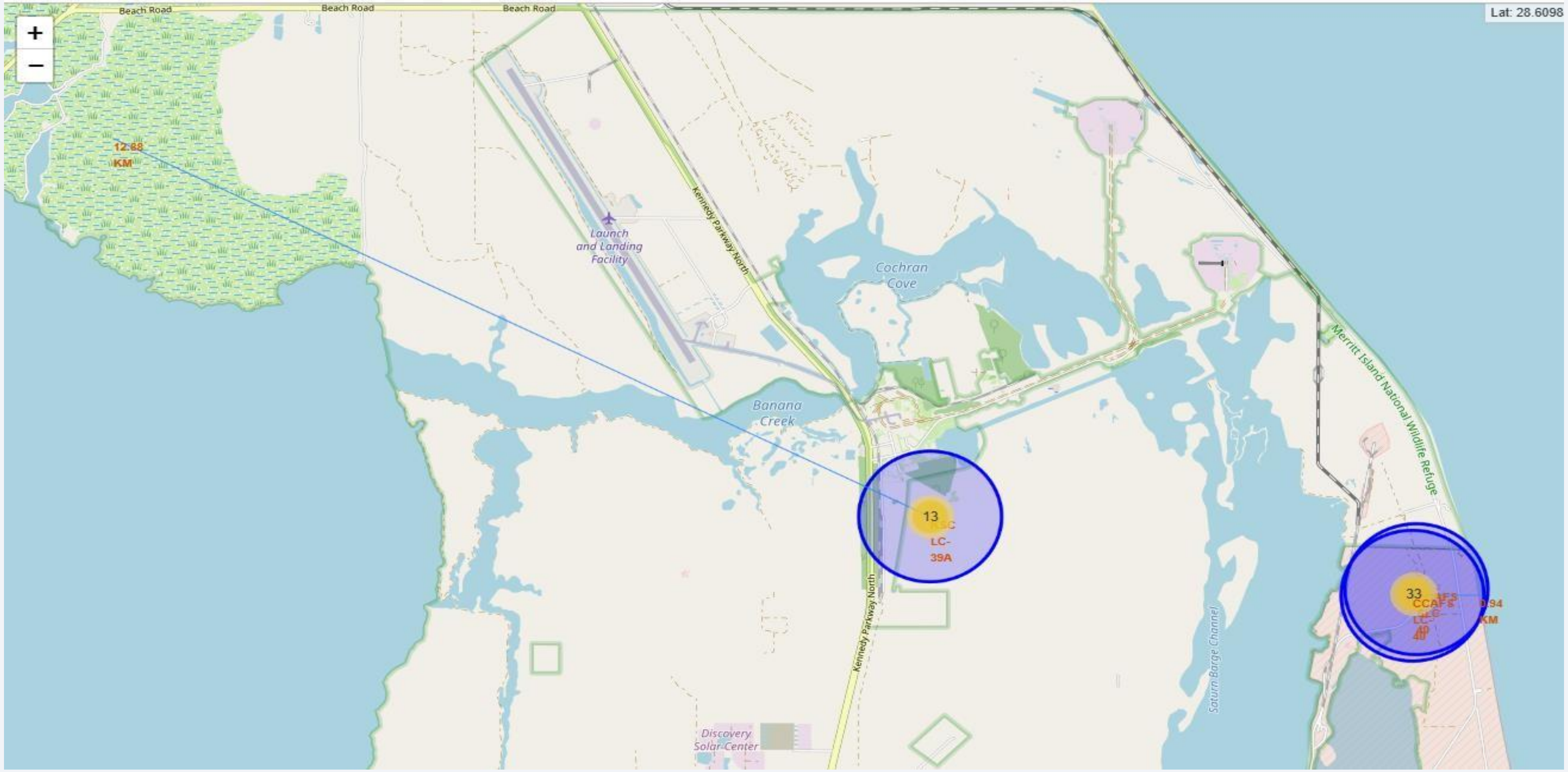
Folium Map Screenshot 1

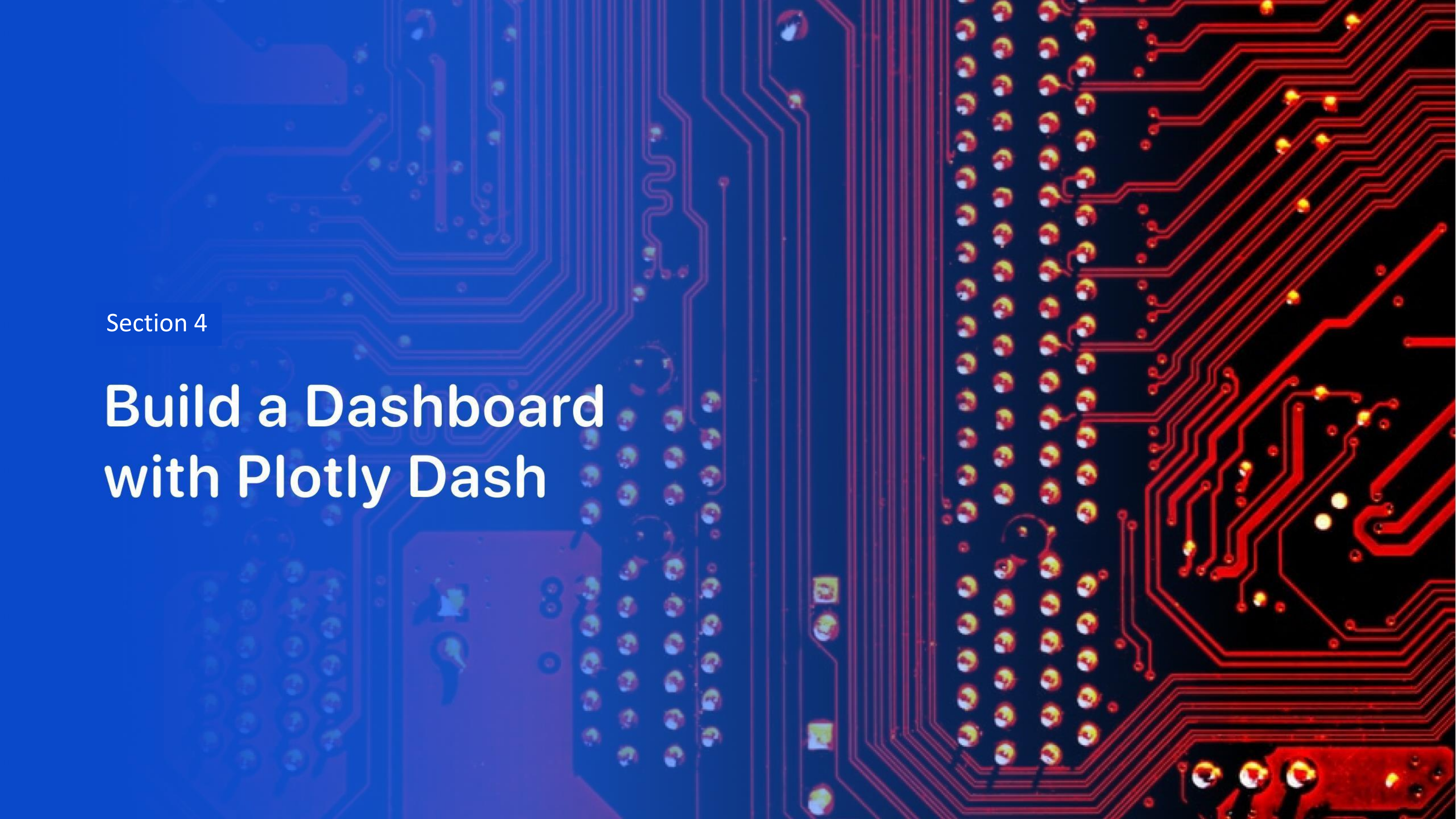
- Map of Cape Canaveral Launch Complex with color-coded launch icons indicating success or failure of landing.



Folium Map Screenshot 2

- Map showing distance of Cape Canaveral Launch Complex to the coast (0.94 km) and proximity of Kennedy Space Center to the Merritt Island National Wildlife Refuge (12.88 km).



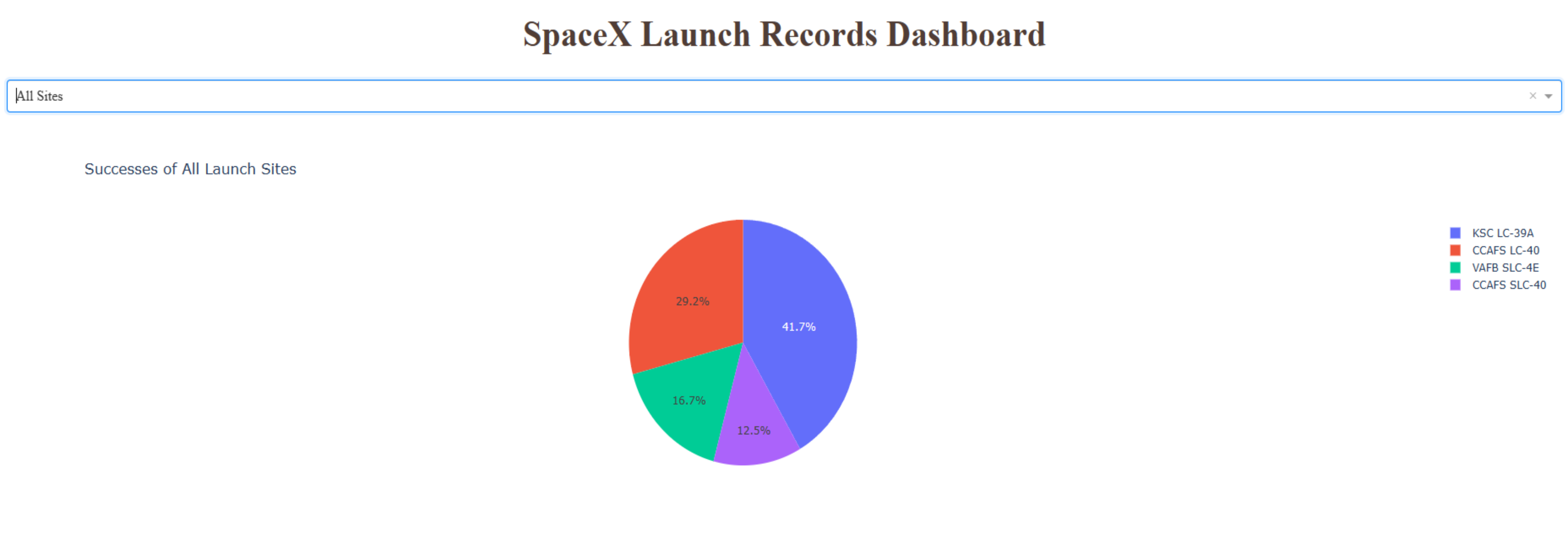


Section 4

Build a Dashboard with Plotly Dash

Success Rates of All Launch Sites

Kennedy Space Center accounts for the largest portion of successful recoveries - 41.7% of all successes in the dataset. Notably, if you combine Cape Canaveral’s two pads - the earlier CCAFS LC-40 and the newer CCAFS SLC-40 - their combined share of successes is essentially on par with KSC.

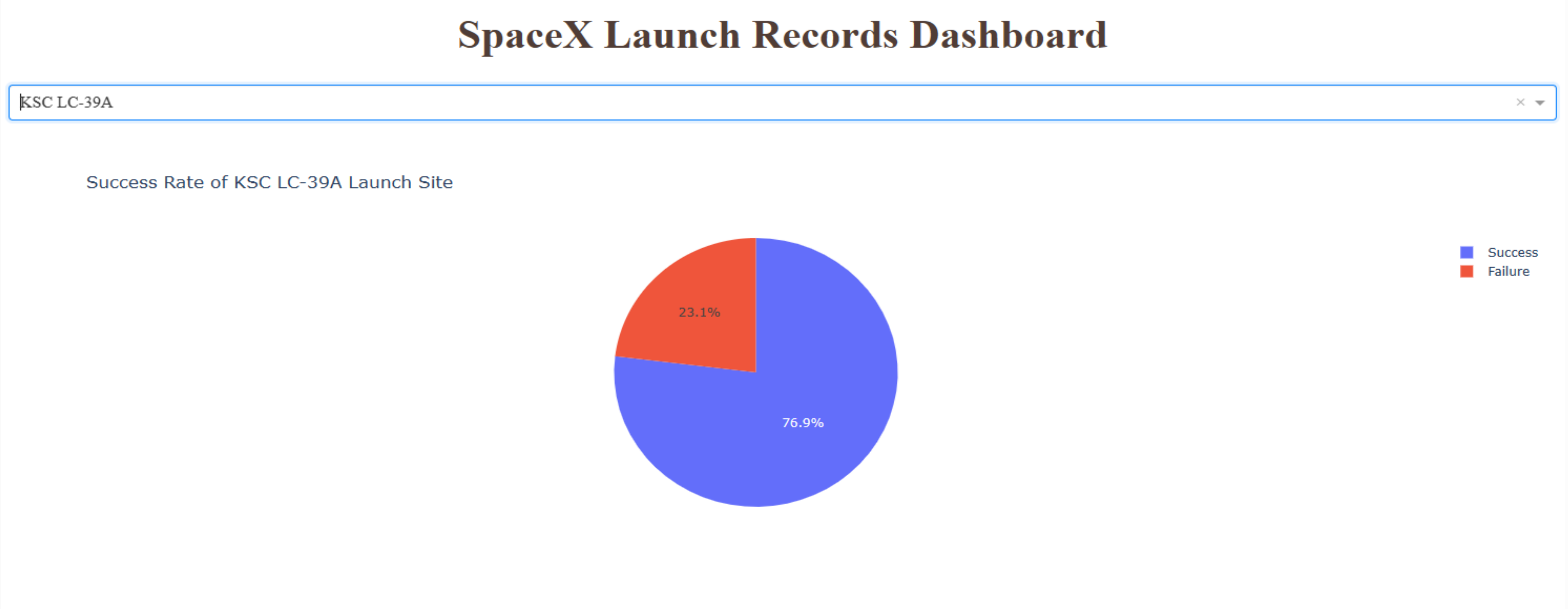


SpaceX Launch Records Dashboard

All Sites

Launch Site with Highest Rate of Successful Landings

Kennedy Space Center leads on both fronts: it contributes the largest share of all successful recoveries and records the top site-level success rate - about 76.9% of its launches end in a successful landing.



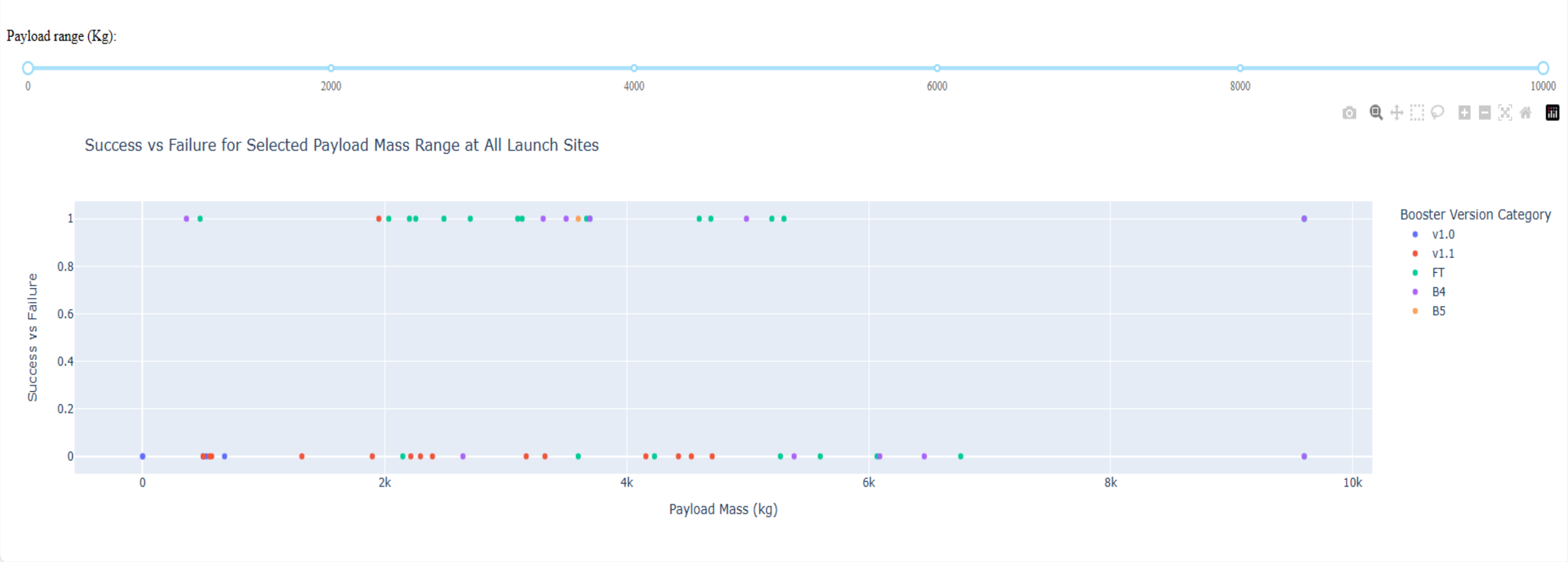
SpaceX Launch Records Dashboard

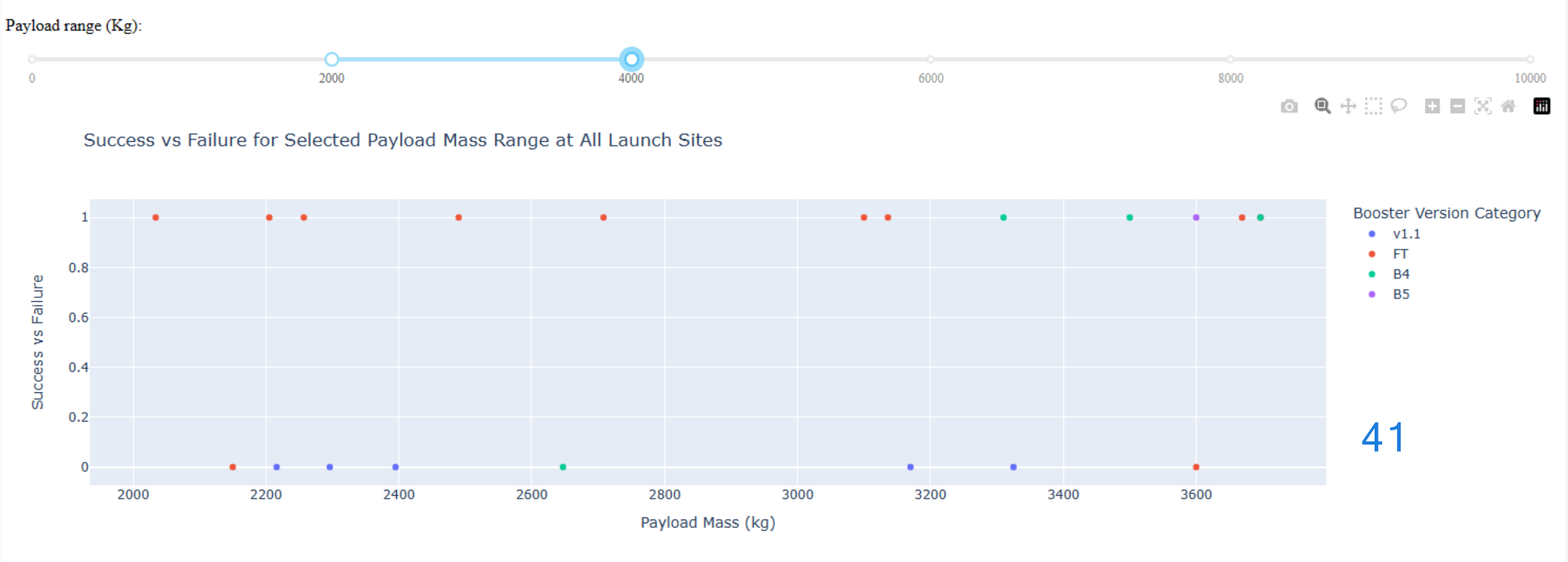
KSC LC-39A

Success Rate of KSC LC-39A Launch Site

Payload Mass vs Launch Outcome

- The scatter plots suggest the FT booster lands more successfully than other variants. The effect is most pronounced for payloads in the 2,000 - 4,000 kg range, which also shows the highest concentration of successful recoveries in the payload distribution.



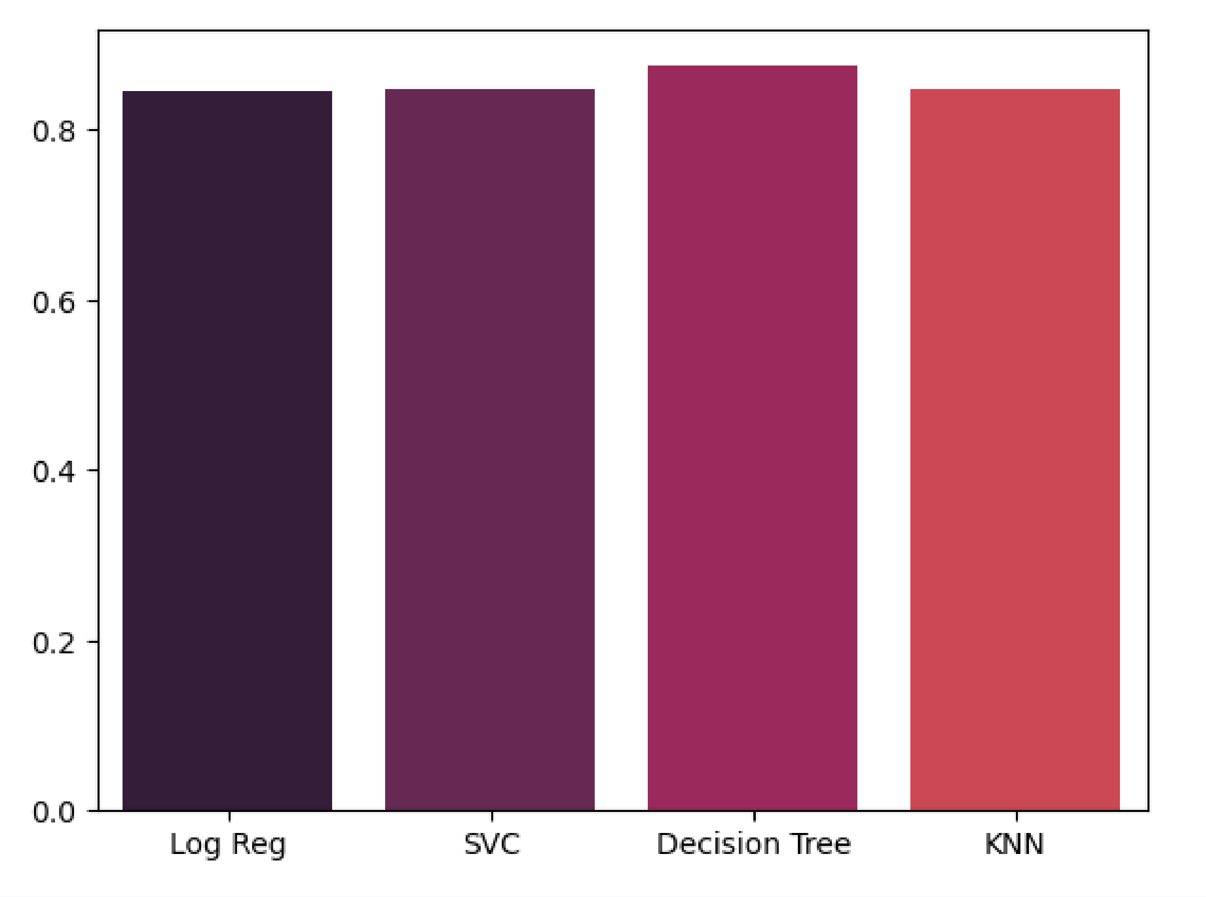


Section 5

Predictive Analysis (Classification)

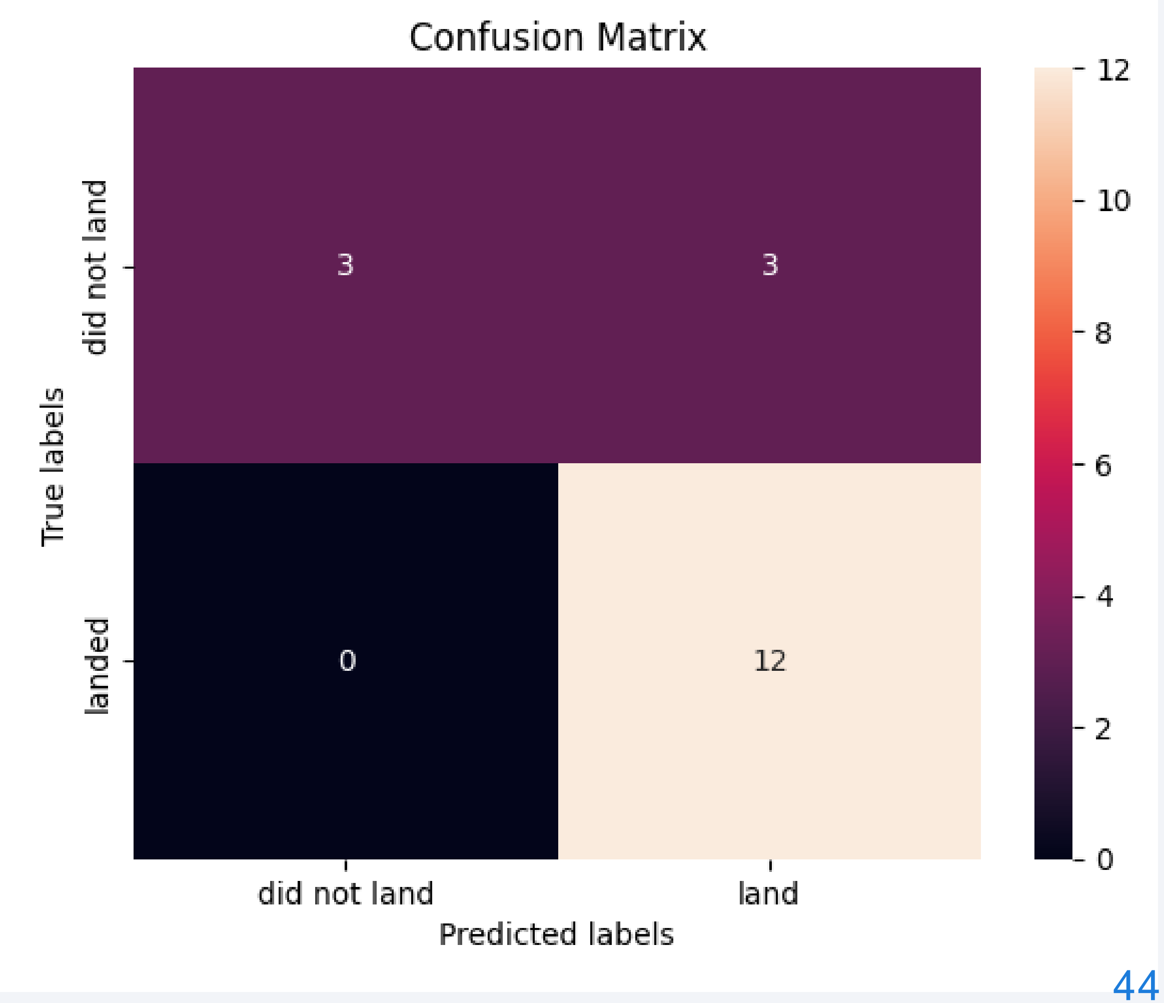
Classification Accuracy

The bar chart compares accuracy across the four classifiers I tested, with the Decision Tree coming out on top at 87.5%.



Confusion Matrix

- On the test split, all four models behaved the same: 3 false positives, 0 false negatives, and 83.33% accuracy. A larger test sample - or stronger validation (e.g., k-fold CV or repeated splits) - would help tease out real differences between them.



Conclusions

- SpaceX’s push for lower-cost access to orbit is clearly paying off: first-stage landing reliability has climbed steadily from the early years and has held above 90% since 2019.
- While heavier missions are now landing more often, the 2,000 - 4,000 kg payload band remains both the most common and the most consistently successful.
- Successful recoveries span 10 of 11 orbit types in the data, though several are based on single attempts; among the frequently used orbits, VLEO stands out with a particularly high success rate across 14+ missions.
- In model building, a Decision Tree delivered the top training accuracy (~87.5%). However, on the held-out test set, all four models tied at ~83.3%, suggesting the current sample size limits separation between approaches.
- As new launches are added, we expect sharper insights and better model discrimination—both from more data and from iterating on features and validation strategy.

Appendix

- API Data Collection notebook with code
- Cleaned dataset with the collected data (CSV)
- Web scraping notebook
- Scraped dataset (CSV)
- Data wrangling notebook
- Cleaned dataset (CSV)
- EDA notebook (code + visuals + notes)
- Model-ready dataset
- EDA with SQL notebook
- Interactive Folium notebook
- Dash app (source + callbacks)
- Dataset used by the app
- Predictive analysis notebook

Thank you!