Winning Space Race with Data Science

Kimberley Usher
24 February 2026

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This project applies data science methods to one of the most consequential cost reduction stories in modern aerospace: SpaceX's development of reusable rocket technology.
- First-stage booster recovery is the mechanism through which SpaceX sustains its cost advantage in commercial launch. This project uses publicly available mission data to test whether that outcome is predictable from observable launch parameters - payload mass, orbit type, launch site, and booster version.
- Data was collected programmatically from the SpaceX REST API and supplemented with historical records via web scraping.
- Four supervised classification models were trained, tuned using GridSearchCV with 10-fold cross-validation, and evaluated on a held-out test set.
- **The Decision Tree** classifier achieved the strongest performance at **94.4%** test accuracy, correctly classifying all 12 successful landings with a single misclassification.
- Beyond predictive accuracy, the exploratory analysis highlights consistent patterns across launch sites, orbit types, and booster generations that are consistent with SpaceX's broader technological trajectory.
- Taken together, the findings suggest that publicly available mission data contains sufficient signal to support meaningful prediction of landing outcomes, with clear implications for launch economics.

Introduction

- The commercial space industry is often framed as an engineering story, but it is equally an economic one.
- SpaceX's ability to offer Falcon 9 launches at approximately $62M, compared to competitors charging $165M or more, is rooted in a strategic commitment to first-stage reusability.
- Recovering and reusing boosters reduces marginal launch costs and has materially altered the competitive structure of a historically state-dominated industry.
- This project begins from that economic premise. If landing success underpins cost advantage, then predicting landing outcomes becomes operationally and strategically significant.
- Using publicly available mission data, this project models first-stage landing outcomes as a binary classification problem.
- Specifically, it tests whether observable launch parameters, including payload mass, orbit type, launch site, and booster version, contain sufficient predictive signal to reliably classify landing success.
- The results indicate that they do.

Section 1

Methodology

Methodology

The analysis follows a four-stage pipeline, with each stage informing the next.

Data Collection

- Data were sourced from two complementary inputs: the SpaceX REST API (v4), which provides structured mission-level variables such as payload mass, launch site, booster version, and landing outcome; and Wikipedia's historical Falcon 9 launch tables, accessed via web scraping.
- The API offers operational granularity, while Wikipedia adds historical completeness. These sources were merged into a consolidated dataset of 90 Falcon 9 launches.

Data Wrangling

- Missing PayloadMass values were imputed using the column mean. Falcon 1 launches were excluded to maintain platform consistency.
- A binary target variable, Class, was engineered to represent landing outcome (1 = success, 0 = failure).
- These preprocessing decisions defined the analytical scope and ensured the model was aligned with the specific prediction task.

Exploratory Analysis

- Exploratory analysis was conducted using Python visualizations and SQL queries to examine distributional patterns across launch sites, orbit types, payload ranges, and booster versions.
- This stage directly informed feature selection and clarified the relationships the classification models were designed to evaluate.

Predictive Modelling

- Four supervised classifiers were trained and compared: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors.
- All models were optimized using GridSearchCV with 10-fold cross-validation.
- **The Decision Tree** achieved the highest test accuracy of **94.4 %** and was selected as the final model based on both performance and interpretability.

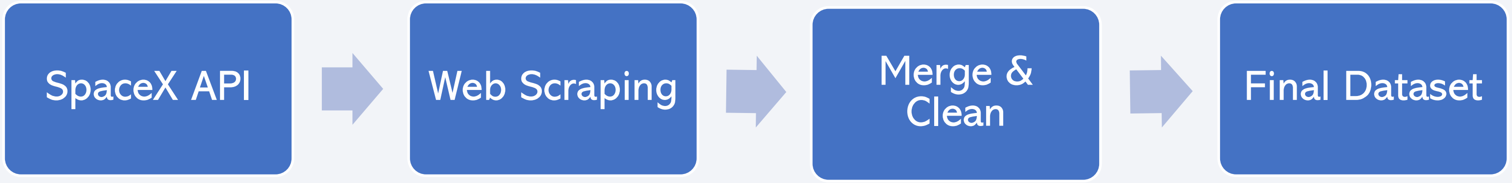
Data Collection

Data underpins everything that follows. Two sources were used to build the most complete picture of SpaceX launches possible:

1. The SpaceX REST API (v4) provided structured launch records including rocket type, payload, launch site, and landing outcome.
2. Wikipedia provided historical Falcon 9 launch tables collected via web scraping.

Both datasets were merged and cleaned to produce a unified dataset of 90 Falcon 9 launches used for analysis and modeling.

Data collection process



Key fields collected:

- Flight Number,
- Date,
- Booster Version,
- Payload Mass (kg),
- Orbit,
- Launch Site,
- Landing Outcome,
- Mission Outcome.

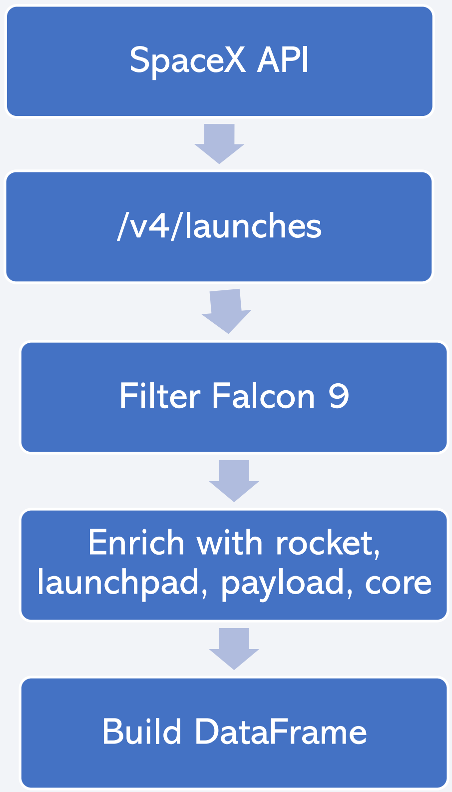
Data Collection - SpaceX API

Step 1: GET /v4/launches - retrieved all Falcon 9 launch records.

Step 2: Used launch IDs to call helper endpoints:

- /v4/rockets/{id} for booster name
- /v4/launchpads/{id} for site name and coordinates
- /v4/payloads/{id} for mass and orbit
- /v4/cores/{id} for landing type and outcome

Step 3: Compiled all fields into a structured Pandas DataFrame with 90 rows and 18 columns.



Data Collection - Web Scraping

Step 1

- Identified Wikipedia's Falcon 9 launch history table as the scraping target.

Step 2

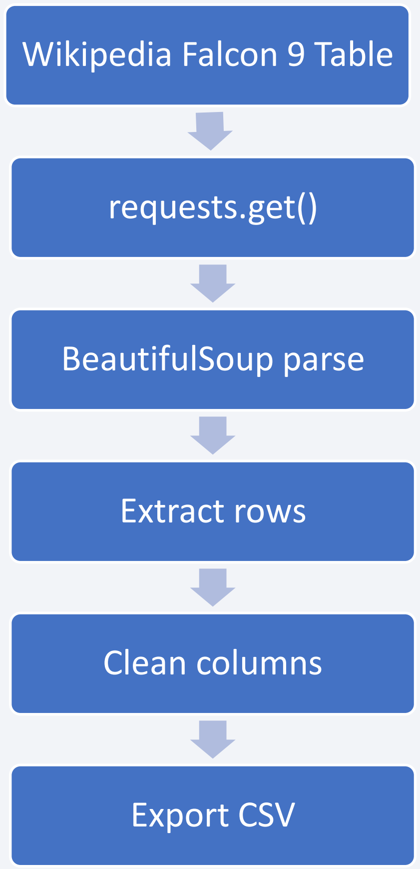
- Used Python requests and BeautifulSoup to parse the HTML table.

Step 3

- Extracted columns including Flight No., Launch Site, Payload, Orbit, Customer, Launch Outcome, and Booster Landing result.

Step 4

- Cleaned and exported data to CSV for merging with the API dataset.



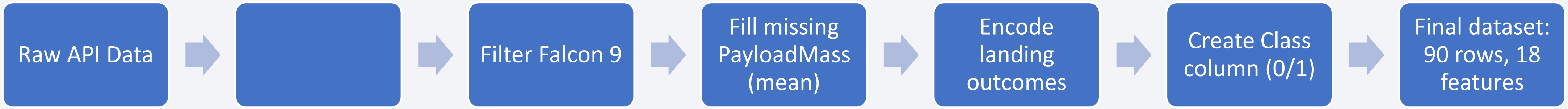
Data Wrangling

- Filtered dataset to include only Falcon 9 launches (excluding Falcon 1).
- Handled missing PayloadMass values by replacing with column mean.
- Created a binary target variable ‘Class’:

  1 = successful landing

  0 = unsuccessful landing

- Verified no duplicate records and confirmed final dataset of 90 launches.
- With a clean, consistent dataset of 90 Falcon 9 launches and a clearly defined target variable, exploratory analysis could begin.



EDA with Data Visualization

Before modelling, exploratory analysis was used to understand which features were likely to influence landing outcome using the following charts:

Flight Number vs. Launch Site scatter plot

- Revealed that CCAFS SLC-40 had the most launches.

Payload Mass vs. Orbit chart

- *S*howed heavier payloads tend to go to LEO.

Flight Number vs. Orbit

- Showed success rates improved over time for ISS missions.

Launch success rate by orbit

- Showed ES-L1, GEO, HEO, and SSO all achieved 100% success.

Bar chart of launch site success rates

- Showed KSC LC-39A had the highest success ratio.

EDA with SQL

SQL queries were used to validate and extend these findings numerically.

Summary of SQL queries performed:

- Queried total payload mass carried by NASA (CRS) missions.
- Identified average payload mass by booster version.
- Listed launch sites with at least 10 successful ground landings.
- Counted total successful and failed mission outcomes.
- Queried landing outcomes between 2010-06-04 and 2017-03-20 ranked by date.

Together, the visual and SQL-based analysis revealed clear patterns across launch sites, payload ranges, and booster versions that would later inform feature selection for the classification models.

Build an Interactive Map with Folium

Interactive geospatial mapping revealed that all SpaceX launch sites are coastal, a deliberate design choice for trajectory safety and booster recovery.

The following objects were created and added to a folium map:

Polylines: Connected sites to landmarks to visualise prCircles (1 km radius):

- Added to each SpaceX launch site (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E) to highlight their exact locations.
- Popups were included to display the site name when clicked.

Text Markers (DivIcon):

- Placed at each site to label the locations directly on the map for easy identification.

MarkerCluster with Colour-Coded Markers:

- Used to group launch outcomes.
- Green markers represent successful launches (class = 1) and red markers represent failed launches (class = 0), enabling visual comparison of success rates.

MousePosition Plugin:

- Added to capture coordinates while hovering over the map, helping identify nearby landmarks.

Markers for Nearby Features:

- Used to display the coastline, highway, railway, and nearest city, including calculated distances in kilometres.

Polylines:

- Drawn to connect launch sites to nearby landmarks, visually showing proximity and highlighting strategic site placement near coasts and transport routes while remaining distant from populated areas.

Build a Dashboard with Plotly Dash

The Dash dashboard was built to make the patterns from EDA interactively explorable allowing any stakeholder to filter by site and payload range without needing to run code.

Plots Added:

Pie Chart

- Shows total successful launches by site, or success vs failure rate for a selected site.
- This was chosen because it quickly and clearly shows proportions, making it easy to see which site performs best at a glance.

Scatter Plot

- Shows correlation between payload mass and launch outcome, color-coded by booster version.
- This was chosen because it is ideal for revealing patterns between two variables, helping identify whether heavier payloads or certain boosters affect mission success.

Interactions Added

Dropdown Menu

- Allows user to select All Sites or a specific launch site, updating both charts automatically.
- This was added so users can drill down into individual sites rather than being overwhelmed by all the data at once.

Range Slider

- Filters launches by payload mass (0 -10,000kg), updating the scatter plot in real time.
- This was added because payload mass varies widely, so filtering by range helps isolate patterns that might not be visible when viewing all data together.

Predictive Analysis (Classification)

With the data thoroughly explored, the question became: can these features reliably predict landing outcomes? Four classification models were trained to find out.

- **Build**

Trained 4 classifiers: Logistic Regression, SVM, Decision Tree, KNN

- **Evaluate**

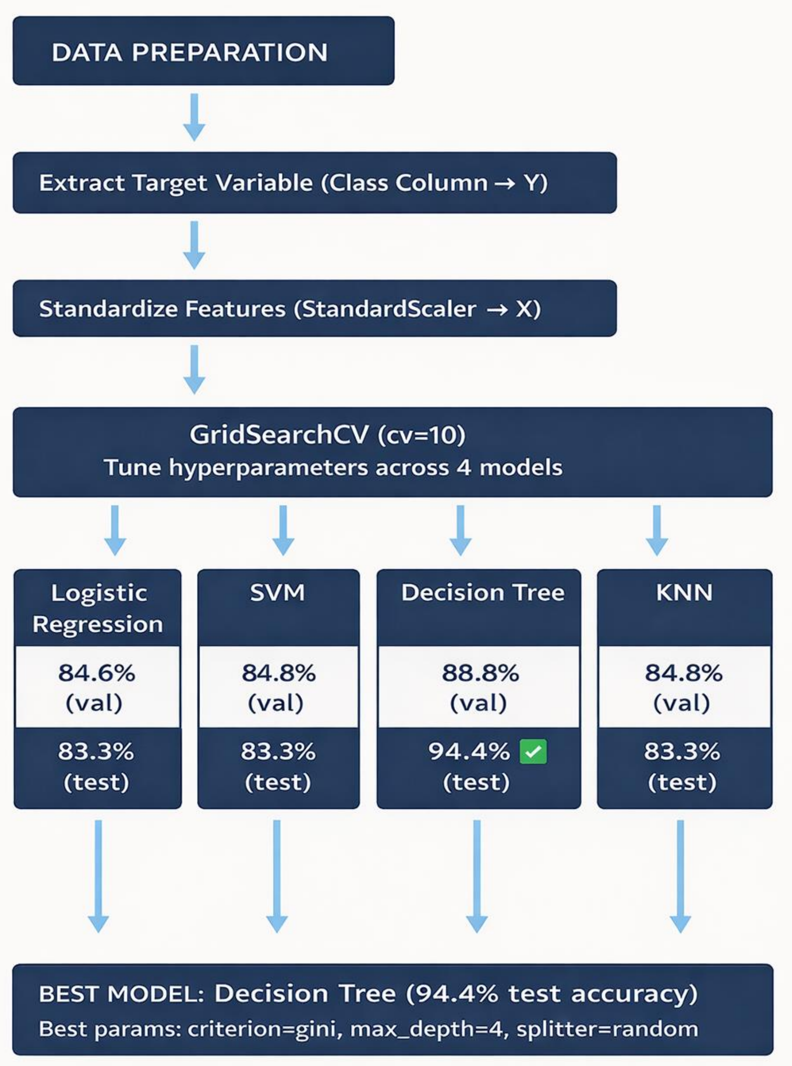
Used confusion matrices and accuracy scores on an 18-sample test set

- **Improve**

Applied GridSearchCV with 10-fold cross-validation to tune hyperparameters for each model

- **Best Model**

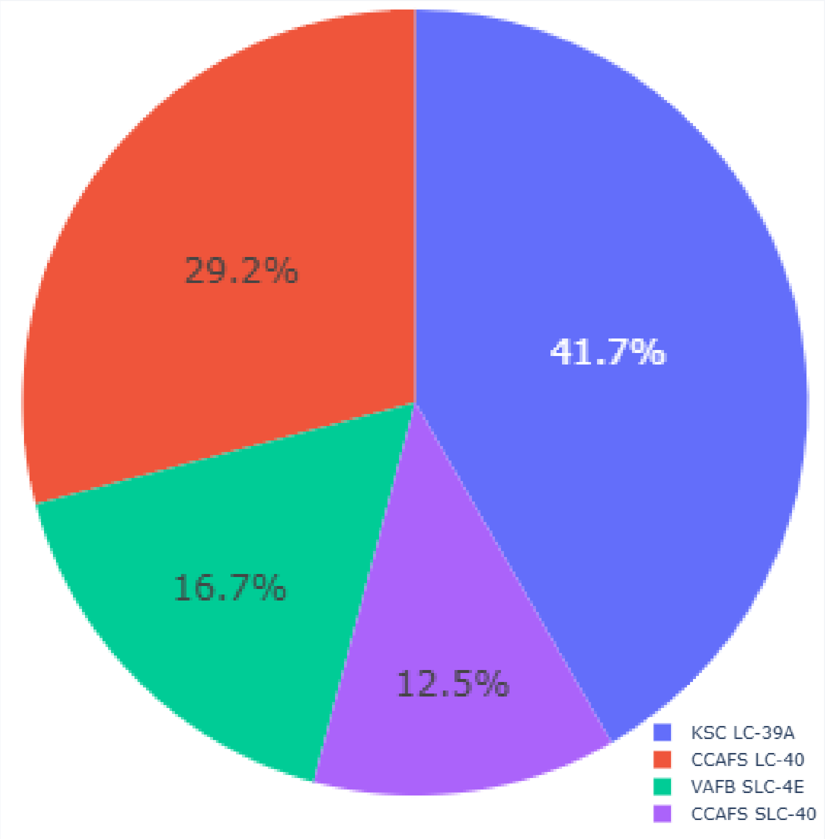
Decision Tree achieved the highest test accuracy at **94.4%**, outperforming all other models which plateaued at 83.3%



Results

Dashboard analysis reinforced the EDA findings that payload range and booster version were consistent predictors of success.

- **Pie chart showed KSC LC-39A contributes 41.7% of all successful launches across all sites**
- **Payload range 2,0r plot revealed that FT Booster version 00–5,500 kg showed the highest launch success rate**
- **Payload range 6,000+ kg showed the lowest launch success rate**
- **Scatter plot revealed that FT Booster version has the highest launch success rate across all payload ranges**
- **Confusion matrices and accuracy charts from the classification models confirmed the Decision Tree as the clear best performer**



Results

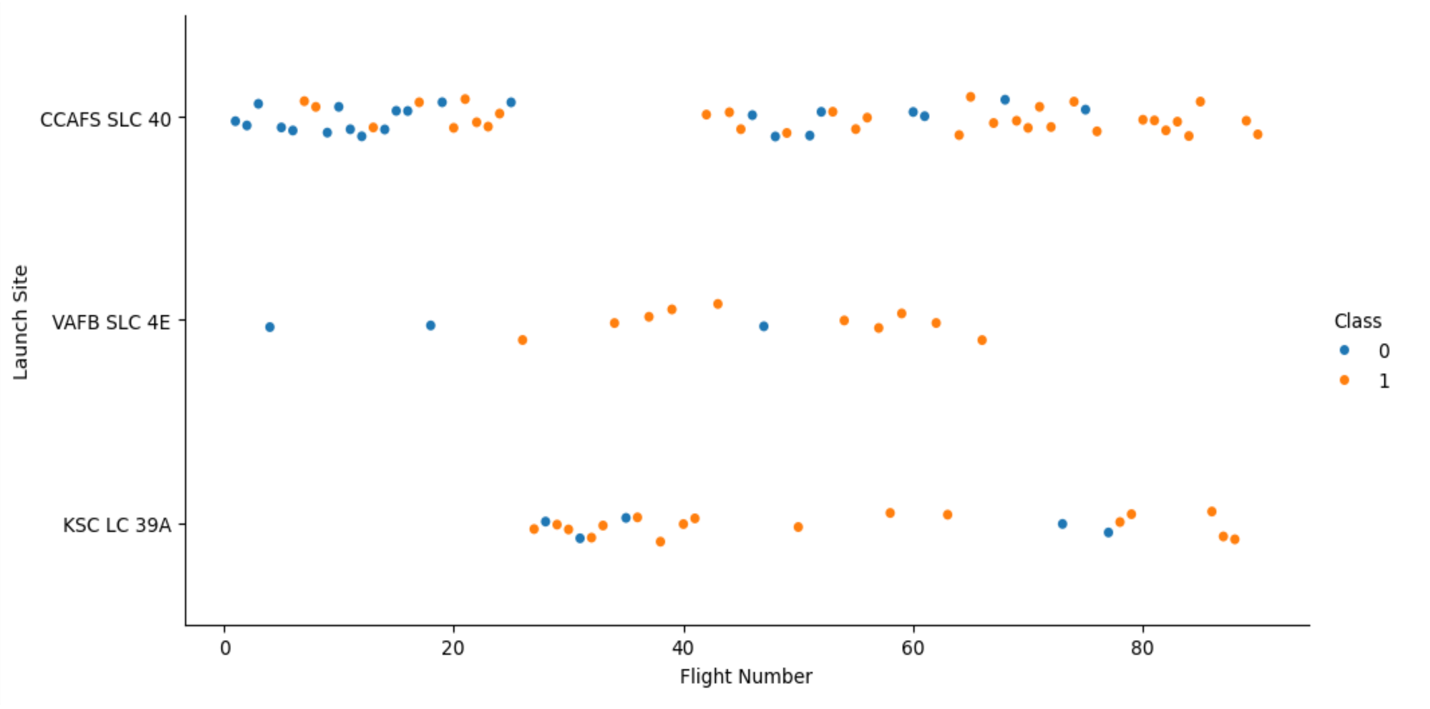

Payload range (Kg):
0
0
2500
5000
7500
10000
9600
Correlation between Payload and Success for all Sites
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

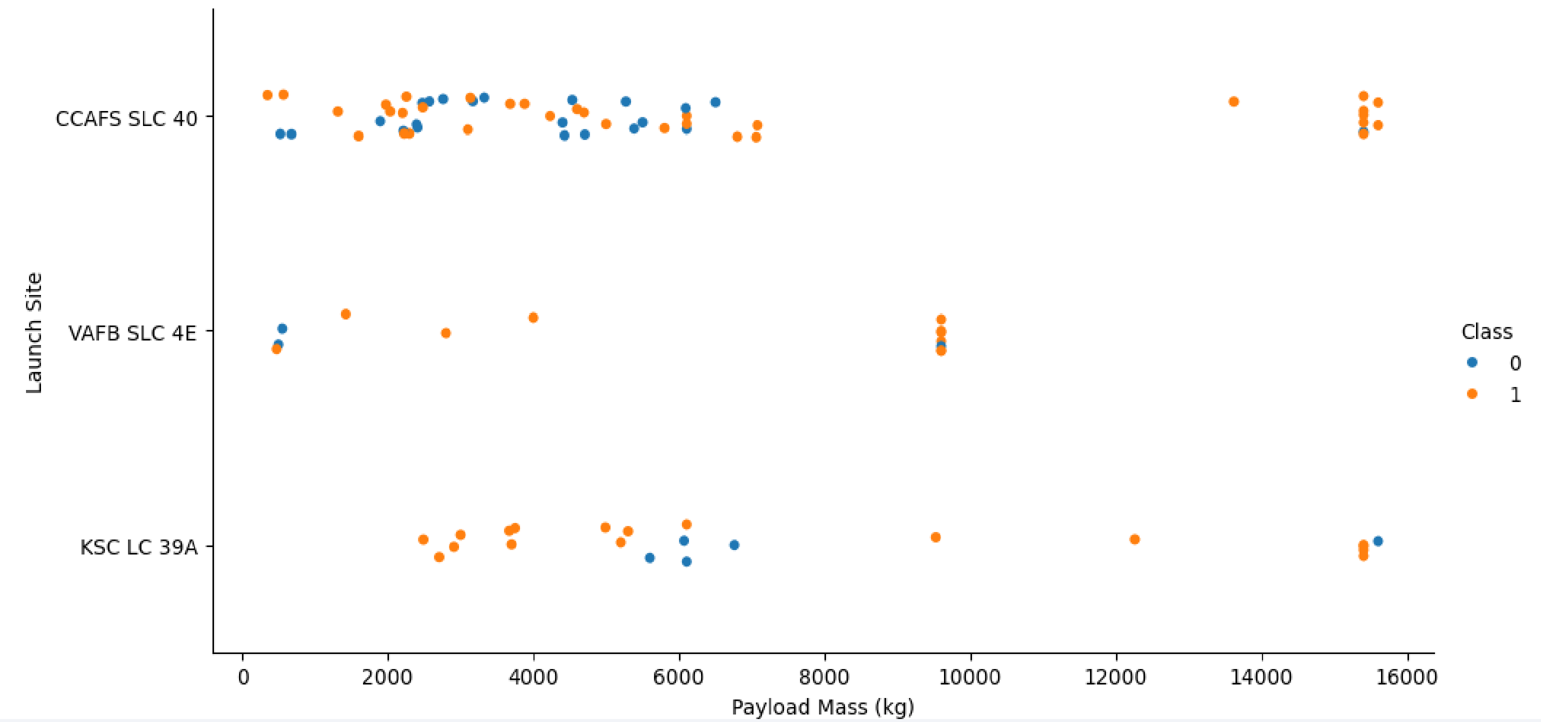
This scatter plot shows which launch site was used for each flight, colored by mission success



- The relationship between flight number and launch site tells a story of operational consolidation.
- CCAFS SLC-40 dominated early launches precisely because it was SpaceX's primary facility during the period when the Falcon 9 programme was still proving itself.
- The progressive improvement in success rates at higher flight numbers reflects the compounding effect of iterative learning.

- Each mission, whether successful or not, generated engineering data that fed directly into the next.
- This chart shows remarkable institutional learning, demonstrating how the company systematically converted experience into reliability.

Payload vs. Launch Site



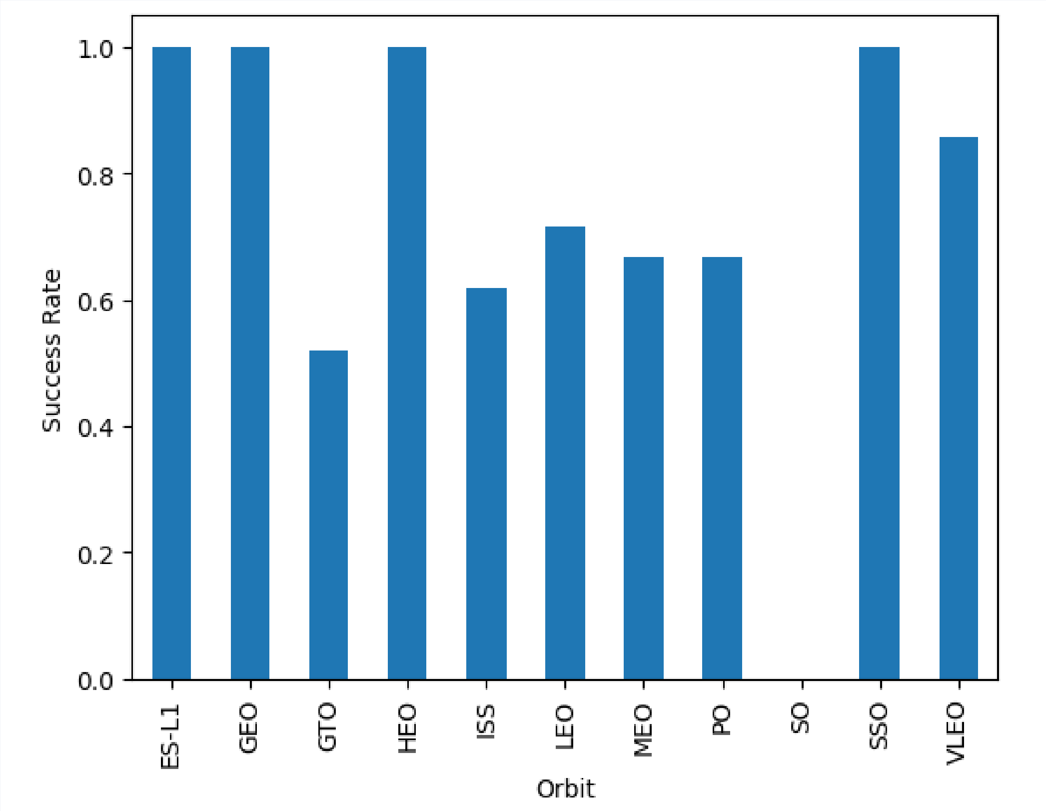
This scatter plot shows the payload mass carried per mission at each launch site, colored by mission success.

- The absence of a clear pattern between payload mass and mission failure across launch sites is itself analytically significant. It suggests that payload mass alone was not the primary determinant of landing success - a finding that pushes the analysis toward more complex interactions between variables.

- **CCAFS SLC-40** handles the widest range of payload sizes, from very light to extremely heavy (~15,500 kg), making it SpaceX's most versatile launch site. Its capacity to handle the widest payload range reflects its role as SpaceX's most operationally mature facility during this period, but the scattered failure distribution confirms that site and payload mass must be understood in combination with booster generation and orbit type to explain outcomes meaningfully.

Success Rate vs. Orbit Type

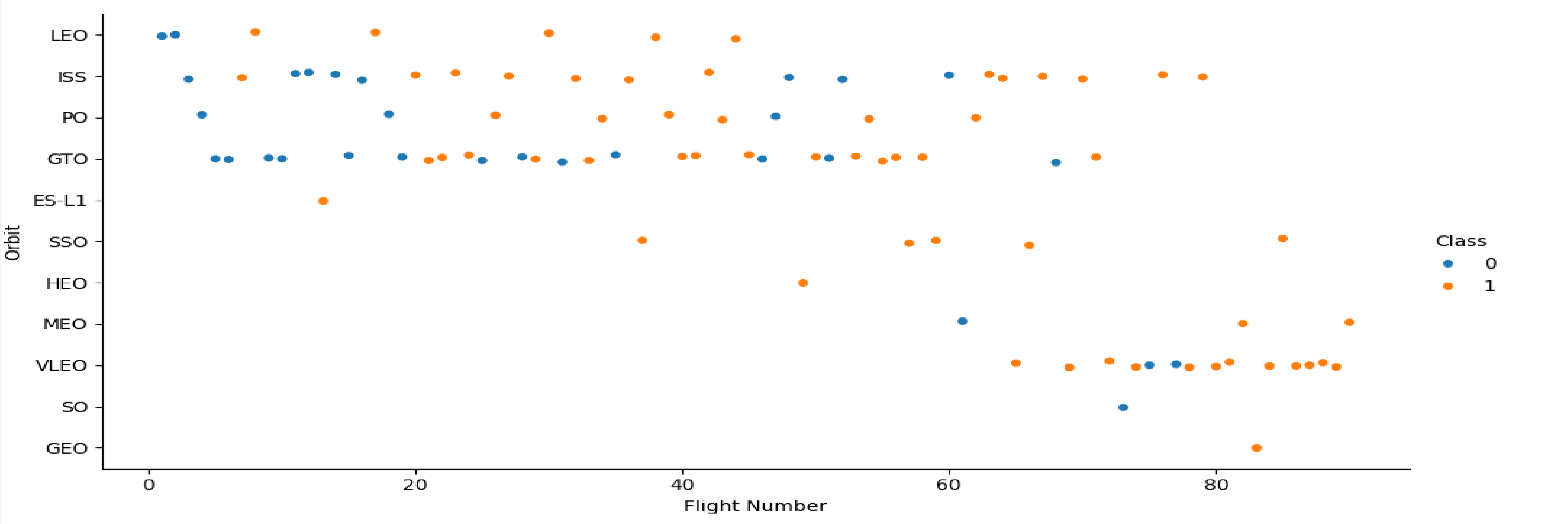
This bar chart shows the mission success rate for each orbit type SpaceX targeted.



- Orbit type emerges here as one of the strongest structural predictors in the dataset.
- The perfect success rates for ES-L1, GEO, HEO, and SSO reflect missions that were attempted only after SpaceX had achieved sufficient operational maturity.
- These orbit types appear later in the dataset, by which point landing technology was well established.
- GTO's low success rate (~52%) is not simply a technical limitation; it reflects a fundamental tension between mission demands and recovery feasibility.
- Geostationary transfer orbits require maximum propellant expenditure, leaving minimal margin for the controlled descent burn.
- This trade-off between payload delivery and booster recovery is central to understanding where reusability has limits.

Flight Number vs. Orbit Type

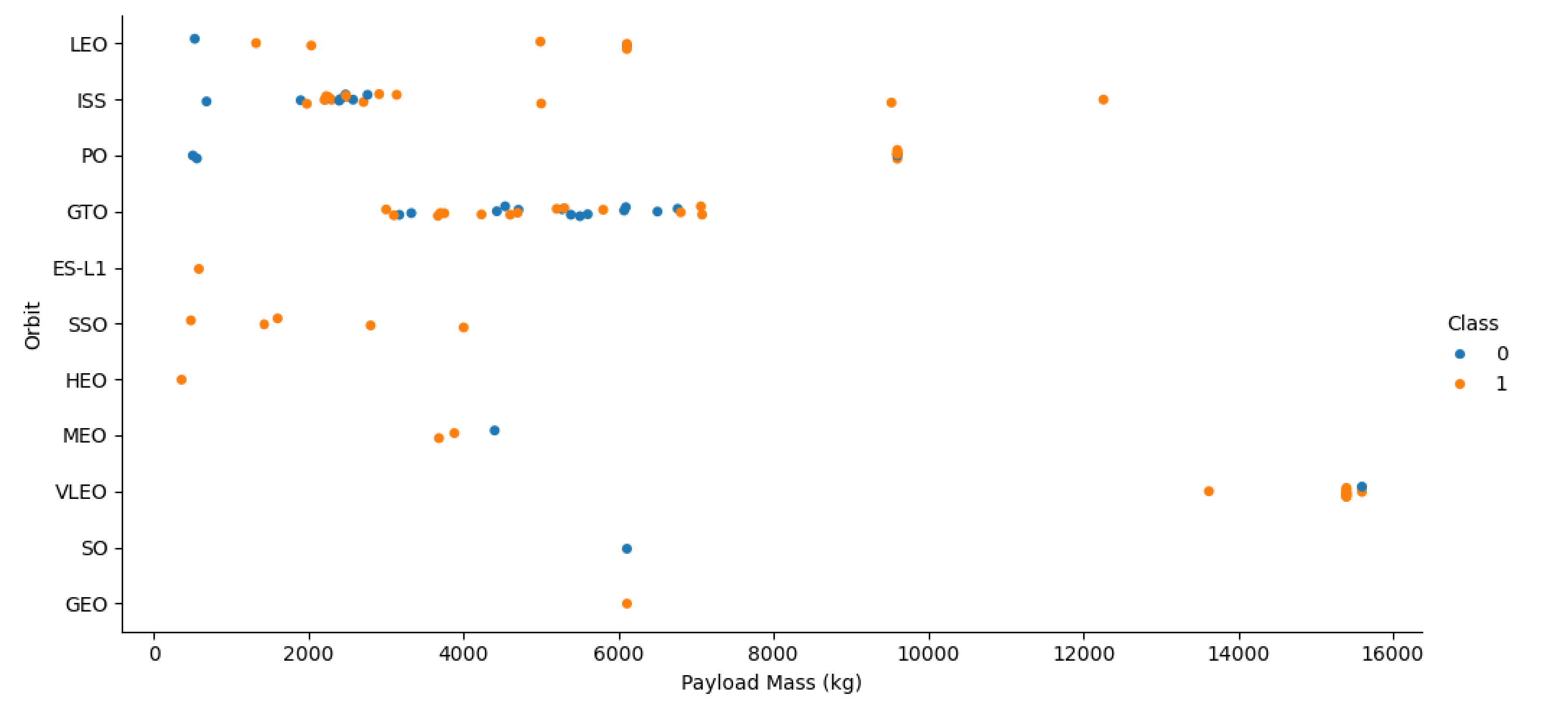
This chart captures SpaceX's expanding mission portfolio over time.



- The early concentration of GTO and ISS missions reflects the commercial and institutional contracts that sustained the company during its formative years.
- The emergence of VLEO and SSO missions at higher flight numbers signals the Starlink era - a strategic pivot toward high-frequency, high-volume deployment that fundamentally changed SpaceX's operational profile.
- The shift from predominantly failed to predominantly successful outcomes across all orbit types over time is consistent with the broader narrative: technical maturity translating into commercial dominance

Payload vs. Orbit Type

This scatter plot shows the payload mass carried for each orbit type.



- The clustering of GTO missions between 3,000 - 6,500 kg with the highest failure concentration reinforces the orbit type finding from the previous slide and adds granularity.

- GTO missions are more difficult at a specific payload range where the competing demands of delivery and recovery are most acute.

- The VLEO cluster at 13,000 - 15,500 kg with near-perfect success rates appears paradoxical at first - heavier payloads, better outcomes - but is explained by mission profile: Starlink deployments to very low Earth orbit require less propellant for the delivery burn, leaving more margin for recovery.

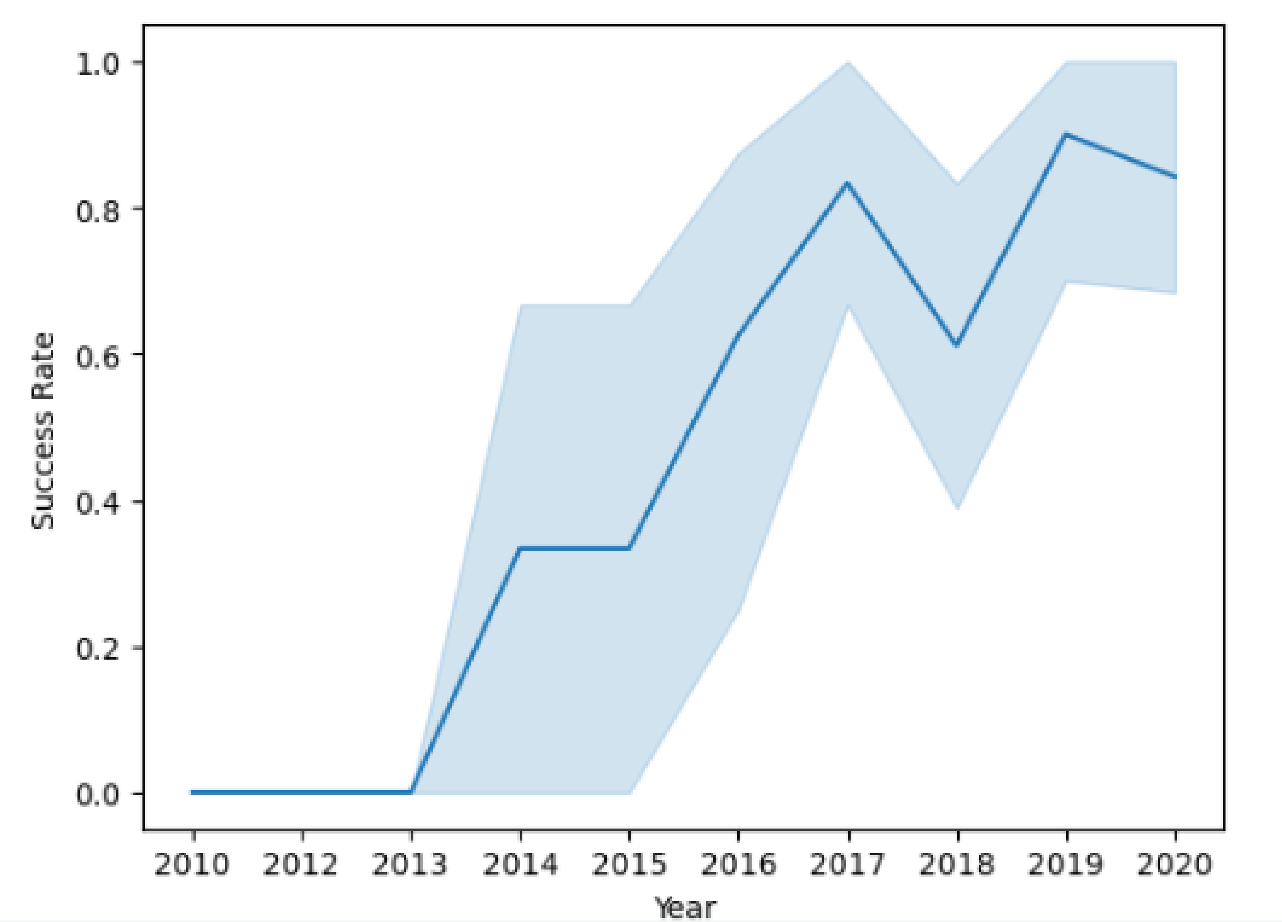
- Payload mass and orbit type interact in ways that a single-variable analysis would miss entirely.

Launch Success Yearly Trend

This line chart showing SpaceX's average mission success rate per year, with a shaded confidence interval band.

The yearly trend chart is perhaps the most analytically compelling visualization in the exploratory analysis:

- The trajectory from 0% success in 2010 - 2013 to sustained rates above 80% from 2017 onwards is punctuated by the wide confidence band of 2015 - 2018, which reflects a period of intense experimentation.
- This was the window in which SpaceX was simultaneously attempting to master drone ship landings, absorbing the lessons of early failures, and scaling launch frequency.
- The narrowing of the confidence interval toward 2020 signals the transition from experimental to operational: landing was no longer an aspiration but a routine expectation.
- That transition is what made the economics of reusability viable at scale.



All Launch Site Names

This query retrieves all unique launch sites used in SpaceX missions, returning 4 distinct locations:

- **CCAFS LC-40 -** Cape Canaveral
- **CCAFS SLC-40** - Cape Canaveral
- **VAFB SLC-4E** - Vandenberg Air Force Base.
- **KSC LC-39A** - Kennedy Space Centre

```
%%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

Launch Site Names 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 dataset shows the earliest SpaceX Falcon 9 launches (2010–2013), all from **CCAFS LC-40**, carrying Dragon spacecraft and NASA CRS missions with zero or low payloads, where landing was either failed via parachute or not attempted at all - reflecting SpaceX's early experimental phase before reusable landing was developed.

Total Payload Mass- (NASA =45596kg)

This query calculates the total payload mass (in kg) across all SpaceX launches where NASA was the customer under the Commercial Resupply Services (CRS) program, by summing the `PAYLOAD_MASS__KG_` column and filtering rows where the Customer is exactly `'NASA (CRS)'`.

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

- NASA's Commercial Resupply Services program was one of SpaceX's earliest and most consistent customers, and this figure reflects that relationship in concrete terms.

- The 45,596 kg total represents cargo carried across multiple CRS missions to the International Space Station - science experiments, crew supplies, and equipment critical to sustained human presence in orbit.

- What makes this figure significant is not just its scale, but what it represents economically:
  - SpaceX was trusted to carry irreplaceable cargo for the world's largest space agency at a fraction of what previous providers charged.
  - This was the commercial contract that helped validate SpaceX's model and fund the development of the very reusability technology this project analyses.

Average Payload Mass by F9 v1.1 = 2928,4kg

This query calculates the average payload mass carried by the F9 v1.1 booster version, returning 2928.4 kg representing the typical cargo weight this early Falcon 9 variant was carrying per mission.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

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

Avg_Payload_Mass
2928.4

- The F9 v1.1 was SpaceX's second major Falcon 9 configuration, introduced in 2013 as an upgrade to the original v1.0.
- At an average payload of 2,928.4 kg per mission, it was carrying relatively modest loads by modern standards - a reflection of both the rocket's capabilities at that stage and the types of missions being undertaken.
- Crucially, this was the era before booster recovery was a reliable operational reality.
- The v1.1 was being used to push boundaries, and its average payload tells the story of a company still in the process of proving itself.
- By comparison, the later F9 FT and B5 variants would regularly carry payloads two to five times heavier, demonstrating just how far the technology evolved in a short period.

First Successful Ground Landing - 22 December 2015

This query finds the earliest date of a successful Falcon 9 first-stage landing on a ground pad.

```
%%sql
SELECT MIN(Date) AS First_Successful_Ground_Landing
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Ground_Landing
2015-12-22

- It reveals that 22 December 2015 was the historic milestone when SpaceX first successfully landed a booster back on solid ground marking one of the most consequential dates in modern spaceflight.
- When the Falcon 9 first stage returned to Cape Canaveral and landed upright at Landing Zone 1, it demonstrated for the first time that an orbital-class rocket booster could be recovered intact after delivering its payload.
- This was not merely a technical achievement - it was a proof of concept for an entirely new economic model for space access.
- Every successful landing that followed, and every cost saving that SpaceX passed on to customers, traces back to this moment.
- The SQL query that surfaces this date is a small piece of code, but the date it returns carries enormous historical weight.

Successful Drone Ship Landing with Payload between 4000 and 6000

This query identifies the **booster versions** that successfully landed on a drone ship while carrying a payload between 4,000 -6,000 kg, returning 4 boosters.

- The four boosters identified here - B1022, B1026, B1021.2, and B1031.2 - all belong to the F9 FT (Full Thrust) generation and represent SpaceX's operational mastery of drone ship landings at medium payload ranges.
- Drone ship landings are technically more demanding than ground pad recoveries because the booster has less fuel remaining for the landing burn after delivering a heavier payload further downrange.

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

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

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

- That all four of these boosters succeeded at the 4,000 - 6,000 kg range is significant - it shows SpaceX had cracked the problem of recovering boosters even under more challenging mission profiles.
- The B1021.2 and B1031.2 designations are also notable: the ".2" suffix indicates these were reflown boosters, meaning SpaceX was not only landing rockets but reusing them - the true goal of the entire programme.

Total Number of Successful and Failure Mission Outcomes

This query groups and counts missions by their outcome, showing that SpaceX had an overwhelming **98 successes** against just **1 in-flight failure**, with 1 additional ambiguous success reflecting an exceptionally high mission success rate across the dataset.

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

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

- The near-total dominance of successful mission outcomes in this dataset is striking.
- What this distribution reflects is the result of an aggressive iteration and quality improvement process.
- SpaceX's approach - launching frequently, failing fast, and incorporating lessons rapidly - produced a reliability record that rivals or exceeds legacy launch providers who had decades of operational history.
- The single failure - the CRS-7 mission in June 2015 - which was lost due to a strut failure in the second stage - is important to acknowledge precisely because it contextualizes the 98 successes: they were hard won, not guaranteed.

Boosters Maximum Payload

This query uses a subquery to find the **maximum payload mass** in the dataset, then lists all boosters that carried that exact weight - showing that **12 F9 B5 boosters** all carried the maximum payload, highlighting that the latest Block 5 generation was SpaceX's most capable and frequently used heavy-lift variant.

- The dominance of the F9 B5 booster at maximum payload capacity is a clear signal of where SpaceX's technology had arrived by the later years of this dataset.
- The Block 5 variant, introduced in 2018, was designed from the outset for reusability - built to fly ten or more times with minimal refurbishment.
- That 12 B5 boosters all carried the dataset's maximum payload weight speaks to its reliability under the most demanding mission conditions.
- It also reflects the Starlink era: the bulk of these maximum-payload missions were deploying large batches of Starlink satellites into very low Earth orbit, where the combination of heavy payloads and short downrange distances made recovery both feasible and routine.
- The B5 did not just carry the most - it redefined what routine heavy-lift looked like.

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

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

2015 Launch Records

This query uses substr() to extract the month and year from the date (since SQLite lacks date functions), revealing that in 2015 SpaceX had 2 drone ship landing failures in January and April

```
%%sql
SELECT substr(Date, 6, 2) AS Month,
       Landing_Outcome,
       Booster_Version,
       Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
AND substr(Date, 0, 5) = '2015';
```

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

The two drone ship failures in January and April 2015 are easy to overlook in a dataset dominated by successes, but they deserve careful attention.

- Both used the F9 v1.1 booster from CCAFS LC-40, and both failed at the point of landing - not because the missions themselves failed, but because the recovery technology was not yet mature.
- SpaceX was attempting something that had never been done: landing a rocket on a ship in the middle of the ocean.
- The January failure saw the booster run out of hydraulic fluid; the April failure resulted from a leg lockout failure on touchdown.
- These were not the same mistake twice - they were two different engineering problems, both identified, both fixed.
- The methodical way SpaceX responded to these failures is precisely what made the eventual successes possible, and it is reflected clearly in the landing outcome data that follows.

Landing Outcomes Ranking between June 2010 and March 2017

This query counts and ranks all landing outcomes between 04/06/2010 and 20/03/2017, showing SpaceX's full learning curve.

- Between June 2010 and March 2017, SpaceX conducted 31 landing attempts across 8 distinct outcomes, painting a clear picture of a company rapidly evolving its reusable rocket technology.
- The dominance of "No Attempt" (10 missions) reflects SpaceX's earliest flights, where recovery wasn't yet a priority - the focus was simply getting payloads to orbit.
- As confidence grew, SpaceX began experimenting with ocean splashdowns, first uncontrolled (2), then controlled (3), using the ocean as a safe testing ground before committing to full landings.
- The earliest recovery method - parachutes (2 failures) - was quickly abandoned after both attempts failed in 2010, leading SpaceX to develop the propulsive, powered landing system that defines them today.
- By 2015, drone ship and ground pad landings were being attempted regularly. Drone ship results were evenly split - 5 successes vs. 5 failures - capturing the intense trial-and-error phase of mastering ocean landings, with the 2015 failures in January and April eventually giving way to consistent successes.
- Ground pad landings (3 successes) proved more reliable once achieved, with the historic first on December 22, 2015 marking a turning point for the entire space industry.
- A single precluded attempt serves as a reminder that even beyond technical challenges, logistics and weather played a role in SpaceX's journey.

```
%%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
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

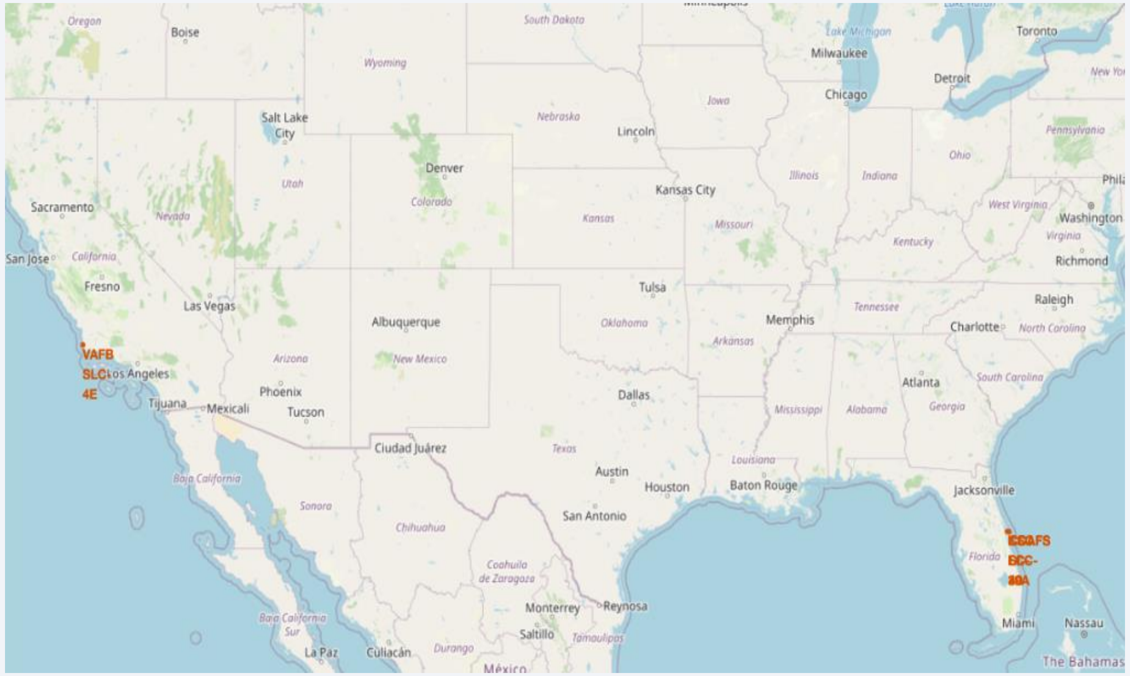
- Taken together, the data tells a story of deliberate, methodical progress - from parachutes to precision landings - that ultimately revolutionized the economics of space travel.

Section 3

Launch Sites Proximities Analysis

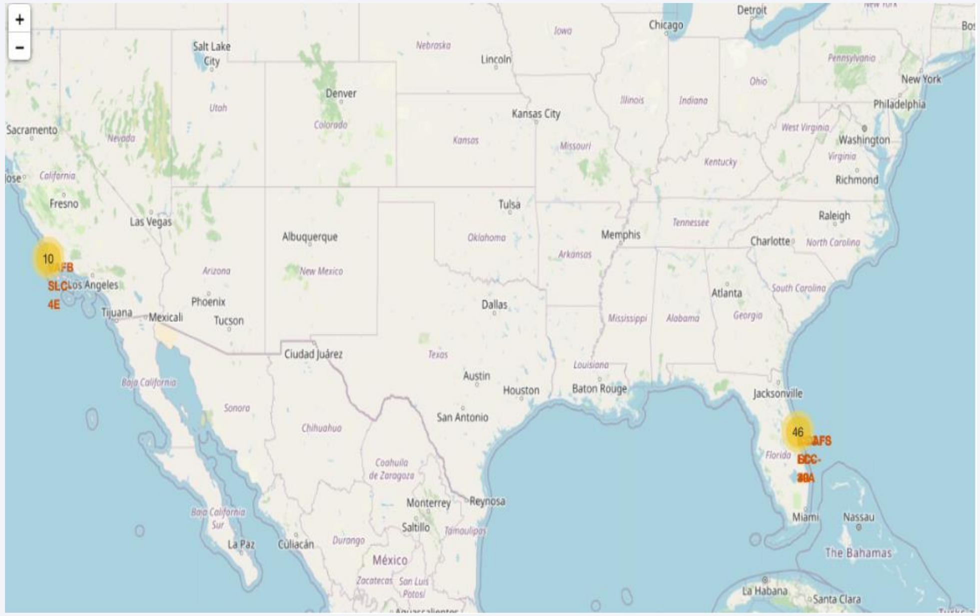
Global SpaceX Launch Site Locations

- The global map displays all 4 SpaceX launch sites marked with circle icons: CCAFS SLC-40 and KSC LC-39A in Florida, VAFB SLC-4E in California, and CCSFS SLC-40.
- All sites are coastal, confirming that launch sites are deliberately located near the ocean for safety and trajectory requirements.
- All launch sites are located along coastlines, enabling safe downrange trajectories over water.
- KSC and CCAFS in Florida support eastward equatorial launches; VAFB in California supports polar orbit launches.



Launch Outcomes by Site - Color-Coded Map

- Each launch marker is color-coded by outcome: green for successful landing, red for failed landing.
- KSC LC-39A shows a high concentration of green markers, indicating it has the best success rate.
- Earlier launches (lower flight numbers) show more red markers, reflecting the learning curve in SpaceX's landing technology.
- Color-coded outcomes confirm that landing success improved significantly after 2017, with KSC LC-39A achieving the highest success rate among all sites.
- Green markers cluster at KSC LC-39A; red markers are more frequent at CCAFS SLC-40, especially in earlier flight numbers



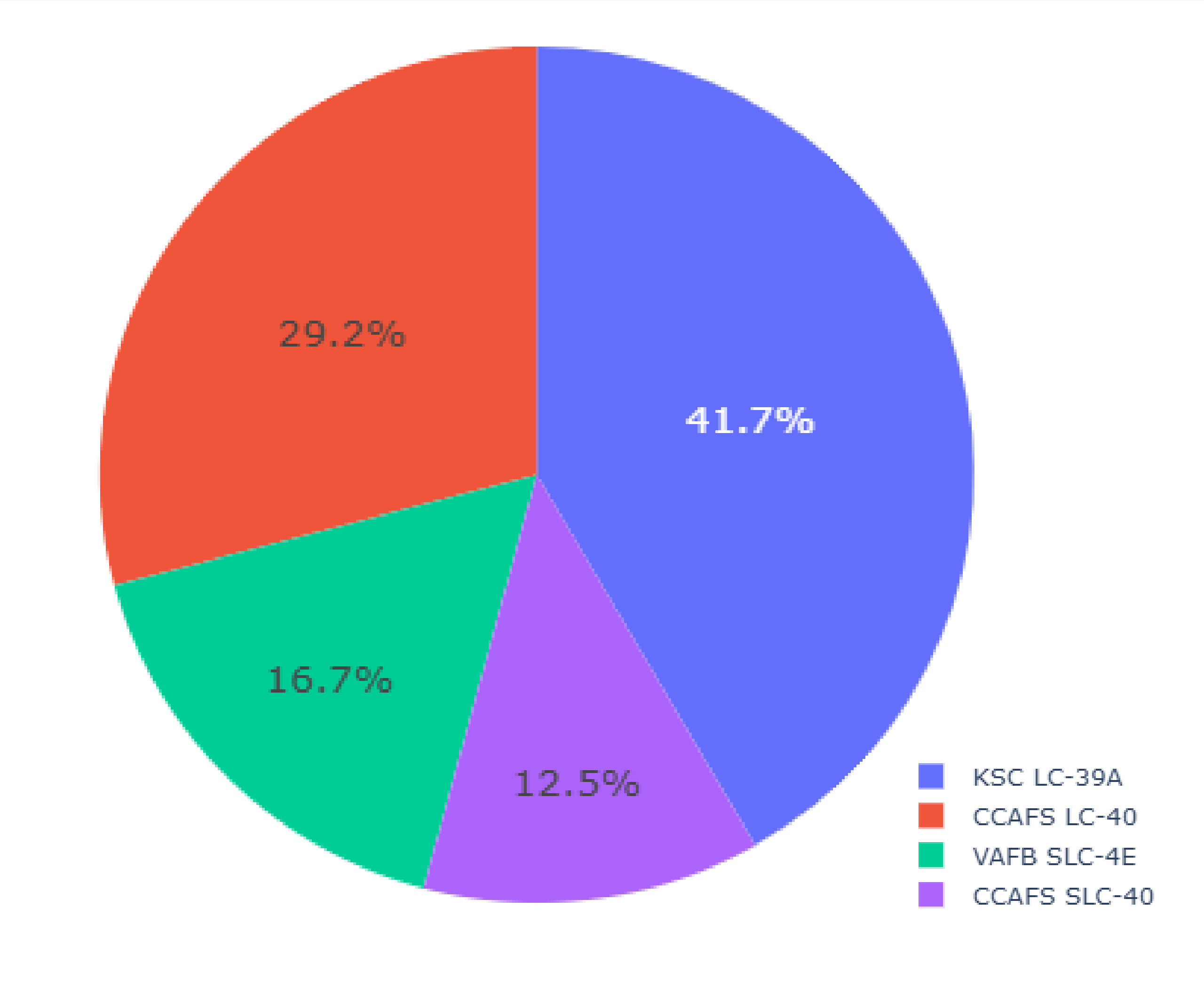
Launch Site Proximity Analysis - KSC LC-39A

- The proximity map for KSC LC-39A shows distances to nearby infrastructure.
- The site is approximately 1.6 km from the coastline, 6.5 km from the nearest highway, and 18 km from the nearest railway.
- KSC LC-39A's proximity to the coast minimizes safety exclusion zone requirements and provides direct ocean access for booster recovery operations.

Section 4

Build a Dashboard with Plotly Dash

Total Success Launches by Site

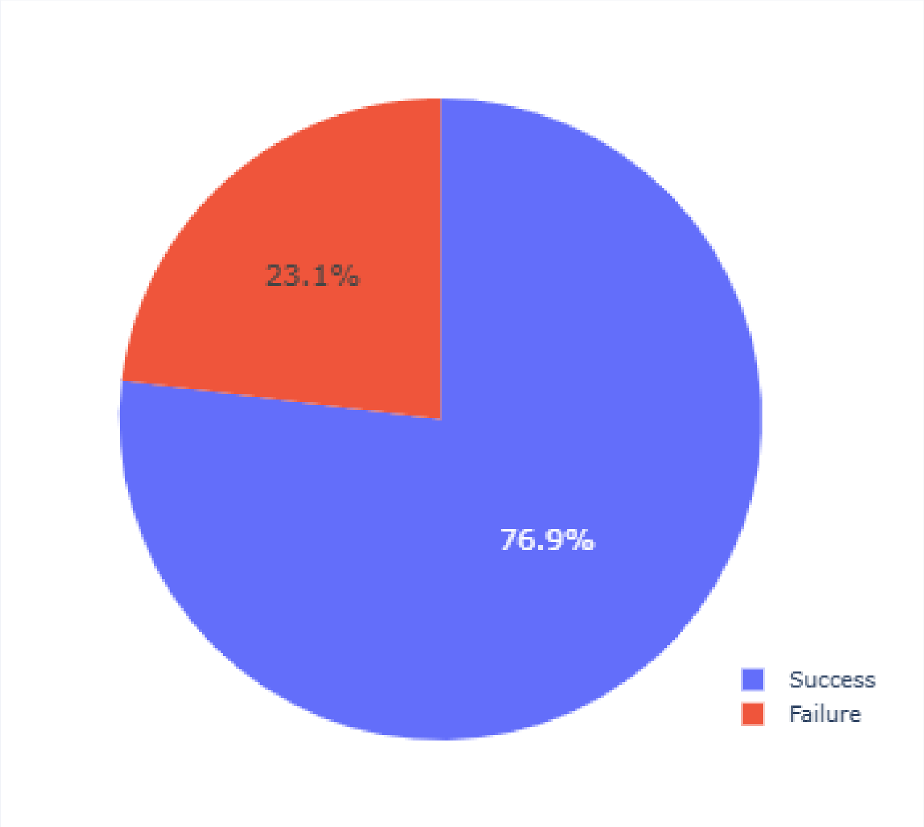


The pie chart shows the total launch success count across all sites.

- **KSC LC-39A-** the most reliable site - leads with **41.7%** of all successful launches.
- **CCAFS LC-40** - one of SpaceX's earliest active sites - is second at **29.2%.**
- **VAFB SLC-4E** - mainly used for polar orbit missions - contributes **16.7%**
- **CCAFS SLC-40** contributes the least at **12.5%**

Highest Success Rate Launch Site- KSC LC-39A

- KSC LC-39A achieved a success rate of approximately 76.9%, the highest among all Falcon 9 launch sites in the dataset.
- KSC LC-39A historically hosted NASA's Apollo and Space Shuttle programs, giving it a legacy infrastructure advantage that contributes to its high launch success reliability.
- KSC LC-39A's combination of infrastructure maturity and favorable launch geometry makes it SpaceX's most reliable site



Interactive Payload Analysis

•The following 4 charts explore the relationship between payload mass and launch outcomes for all of the sites.

•Each view filters a different payload range using an interactive slider

•The scatter plot shows payload mass (kg) on the x-axis and launch outcome on the y-axis

•Color coding: 🔵 Failure = 0, 🟠 Success = 1

0kg -10000kg (Full Payload)

The Full payload scatterplot shows a baseline view showing all launches across every site.

- CCAFS SLC-40 dominates in launch frequency across all payload sizes
- Failures are scattered with no clear payload-specific pattern at this level

0kg - 2000kg (Low Payload)

Zooming into lighter payloads reveals early mission behavior and site preferences.

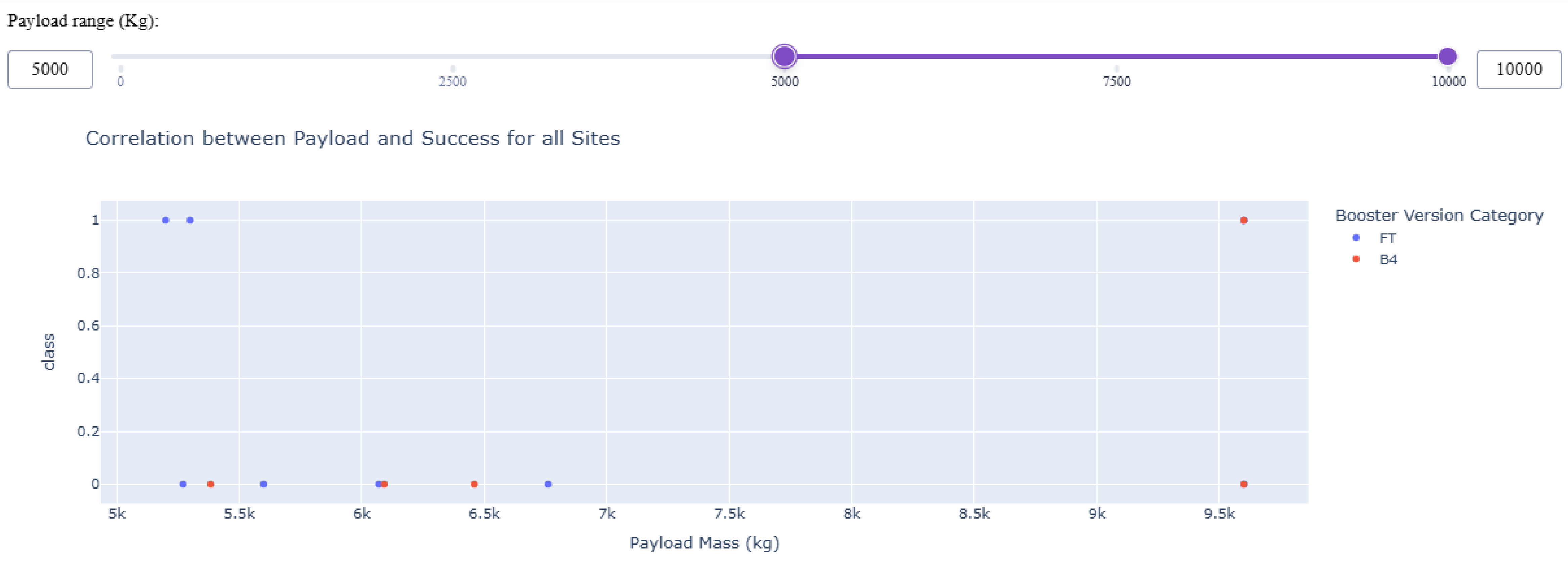
- Lighter missions were mostly early flights with lower success rates
- VAFB SLC-4E features more prominently here with smaller, polar orbit payloads

2000kg - 5000kg (Medium Payload)

The mid payload range is the most competitive and mixed range where most GTO missions fall.

- Highest concentration of failures, particularly for GTO-bound missions
- B5 booster begins appearing here with noticeably better success rates

5000kg -10000kg (High Payload)



The heaviest payloads tell a story of maturity and mastery.

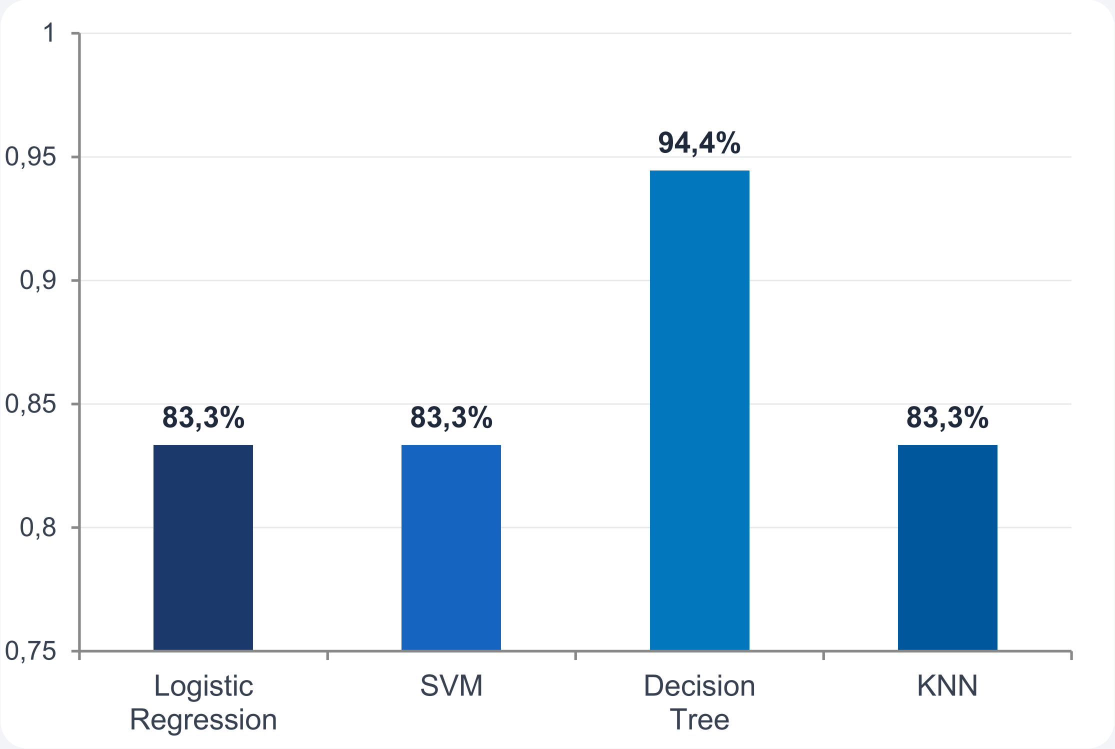
- Almost exclusively successful, dominated by Starlink/VLEO missions on the B5 booster
- KSC LC-39A handles the bulk of heavy payload launches with near-perfect reliability

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Four classification models were trained and evaluated using GridSearchCV with 10-fold cross-validation.



- The Decision Tree classifier achieved the highest test accuracy at 94.44%, outperforming Logistic Regression, SVM, and KNN which all scored 83.33%.

- The Decision Tree model was selected as the final model for its superior accuracy and interpretability, making it well-suited for binary classification of landing outcomes.

Confusion Matrix

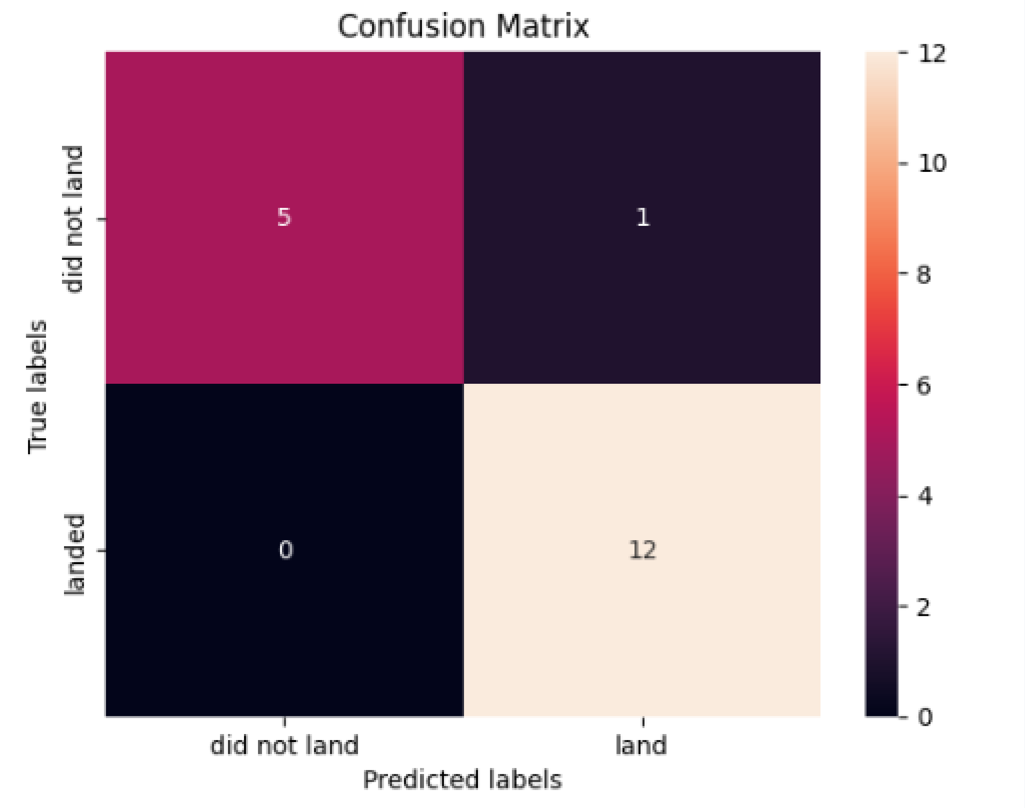
The Decision Tree model's confusion matrix on the test set shows:

- True Negatives (did not land, correctly predicted): 5
- False Positives (did not land, predicted land): 1
- False Negatives (landed, predicted did not land): 0
- True Positives (landed, correctly predicted): 12.

The model correctly identified all 12 successful landings with zero misses (0 false negatives).

Only 1 non-landing was incorrectly classified as a landing.

This gives a precision of 92.3%, recall of 100%, and F1-score of 96.0% for the positive class.



Conclusions

- This project tested a clear proposition: that observable launch parameters contain sufficient predictive signal to classify Falcon 9 first-stage landing outcomes. The results confirm that they do.
- The Decision Tree classifier achieved 94.4% test accuracy on a held-out test set, demonstrating strong separability within the feature space.
- However, the more substantive insight lies beyond model performance. Exploratory analysis shows that landing success is not an isolated technical event. It reflects the interaction of orbit type, payload mass, booster generation, and launch site within SpaceX's evolving operational model.
- The transition from experimental to operational reusability between 2015 and 2017 is visible in the data. The constraints associated with geostationary transfer orbits are visible in the data. The impact of the Block 5 booster and the Starlink programme is visible in the data. The model predicts outcomes, but the analysis explains why those outcomes occur.
- There are important limitations. The dataset includes 90 launches, which is sufficient for identifying patterns but limited for broader generalization. The binary outcome variable simplifies variation in landing conditions. Publicly available data also excludes proprietary operational variables that would influence internal mission planning.
- Ultimately, this project demonstrates that publicly available mission data contains meaningful structural signal. Applied data science methods can surface that structure in analytically useful ways.
- As commercial spaceflight continues to scale, the economics of reusability will shape competitive dynamics across the industry. Understanding the drivers of landing success is therefore not only a technical exercise, but an economic one.

Full code and analysis available at:github.com/kimannu/spacex-falcon9-landing-prediction

Thank you!