IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franz Brincat
21st Sep 2025

Outline

Executive Summary

- **Summary of methodologies**

Over this module we assembled a complete Falcon-9 analytics workflow: we collected launch data via the public API and complemented it with targeted web scraping of Wikipedia tables; cleaned and standardized the dataset (including type casting and imputation); explored patterns using both SQL queries and pandas; visualized relationships with seaborn and produced an interactive folium map to examine geography and outcomes; and finally engineered features (including one-hot encodings) to support downstream machine-learning prediction

- **Summary of all results**

The accompanying report organizes the takeaways from each stage, as exploratory insights from the SQL/pandas analysis, screenshots illustrating the interactive visual analytics, and a concise statement of predictive-model performance. Full tables, charts, and numerical results appear in the subsequent sections of this report.

Introduction

- **Project background and context**

Falcon 9 launches are priced well below many competitors largely because the first stage can be recovered and reused. The economic value of any mission therefore hinges on whether that stage lands successfully. This project builds an end-to-end analytics pipeline - API data + curated web tables → cleaning and feature engineering → exploratory analysis and mapping - to understand the drivers of landing outcomes and to prepare a prediction model. The same insights can inform pricing, risk management, and bid strategies for organizations planning missions or evaluating launch providers.

- **Problems you want to find answers**
    - Which technical and operational factors are most associated with a successful landing?
    - How do interactions among these factors influence success?
    - How has landing reliability evolved over time, and is there evidence of a learning curve effect?
    - What conditions appear necessary to sustain a repeatable, high-success landing program?
    - Can we engineer a compact feature set that yields an accurate, interpretable classifier for landing success prediction?

Section 1

Methodology

Methodology

Executive Summary

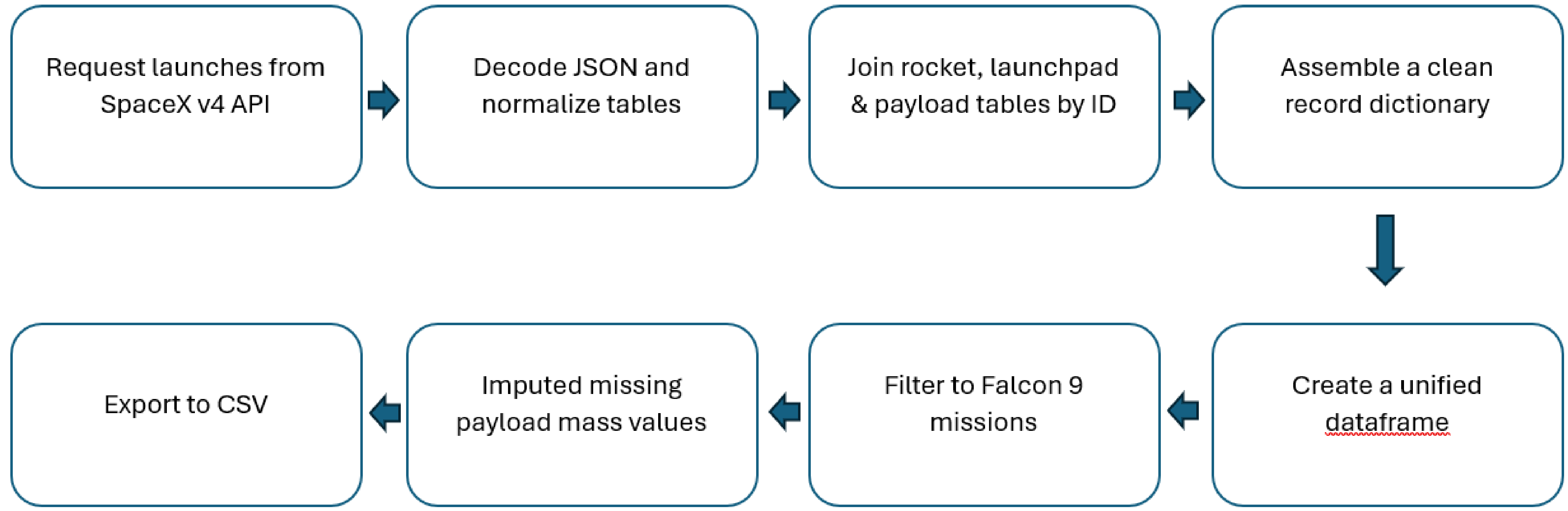
- **Data collection methodology:**
    - Using SpaceX Rest API
    - Using Web Scraping from Wikipedia
- **Perform data wrangling**
    - Filtering the data
    - Dealing with missing values
    - Used One Hot Encoding to prepare the data to a binary classification
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
    - Trained on a stratified split, set a logistic baseline, cross-validated and tuned candidate classifiers optimizing ROC-AUC, selected a decision threshold from ROC/PR, confirmed performance on a hold-out set with full metrics and a confusion matrix, reviewed feature importances, then retrained and saved the best model.

Data Collection

The dataset has been assembled from two public sources:

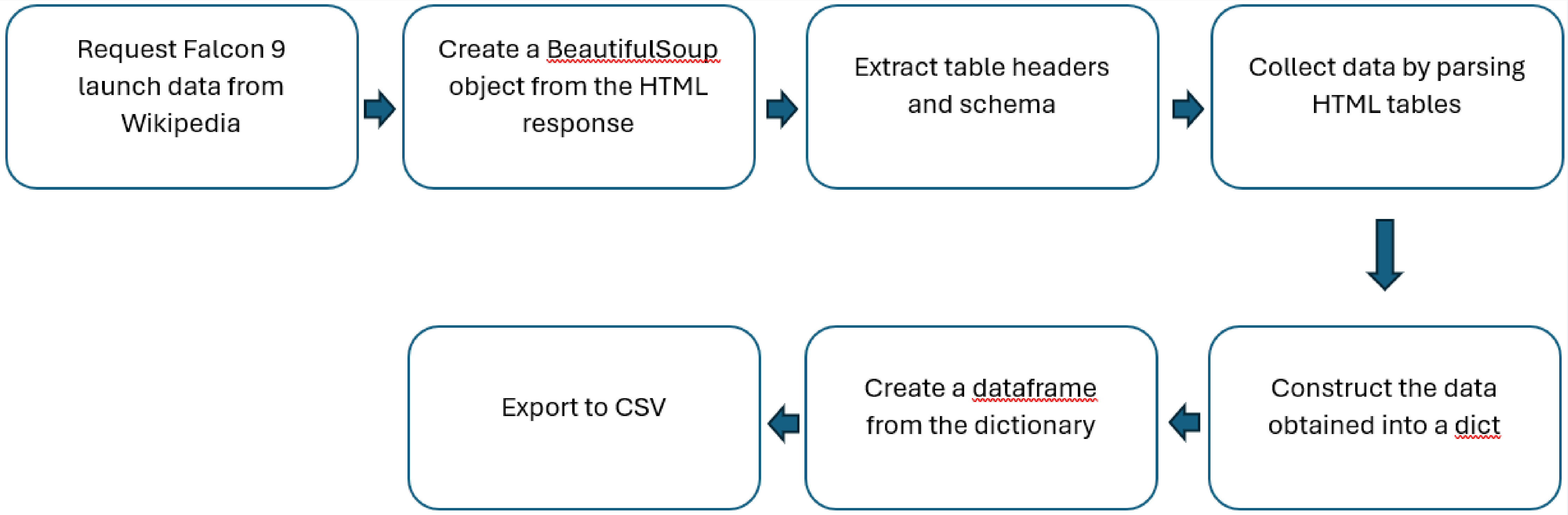
- **SpaceX REST API (v4)** - queried via HTTP GET. Primary endpoints used:
  - Launches: https://api.spacexdata.com/v4/launches
  - Rockets (for metadata like “Falcon 9”): https://api.spacexdata.com/v4/rockets
  - Launchpads (site names/coords): https://api.spacexdata.com/v4/launchpads
  - Payloads (mass/orbit where available): https://api.spacexdata.com/v4/payloads
    We filtered to Falcon-9 launches and saved pulls to CSV/SQLite for reproducibility.
- **Wikipedia wikitables (for gaps/cross-checks)** — scraped with requests + BeautifulSoup
- List of Falcon 9 and Falcon Heavy launches:
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API



GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

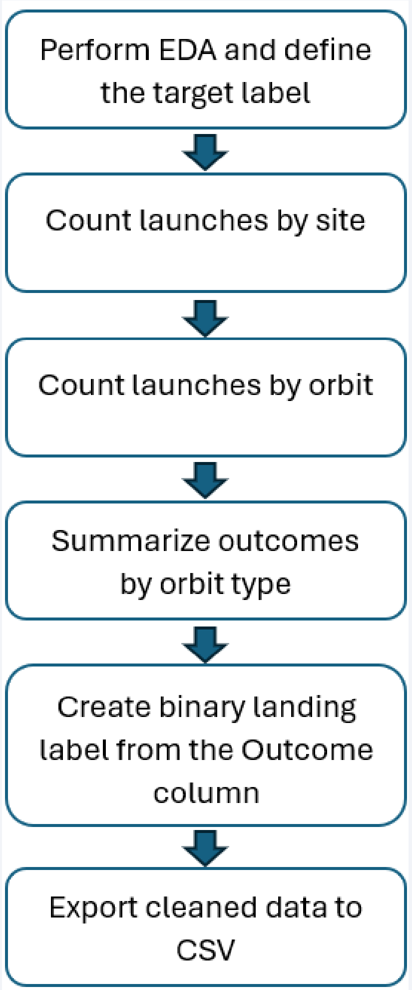


GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- In the dataset, mission outcomes record whether the booster landed successfully and where.
- **Ocean (True/False):** 'True Ocean' = landed in the designated ocean area; "False Ocean" = attempted but did not land successfully in that area.
- **RTLS (True/False):** 'True RTLS' = successful Return To Launch Site landing on a ground pad; "False RTLS" = unsuccessful ground-pad attempt.
- **ASDS (True/False):** 'True ASDS' = successful landing on a drone ship; 'False ASDS' = unsuccessful drone-ship attempt.
- For modeling, we convert these outcomes into a binary training label: 1 = successful landing, 0 = unsuccessful.

GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- **Summarize what charts were plotted and why you used those charts**
  - FlightNumber vs PayloadMass (Scatter), FlightNumber vs LaunchSite (Scatter)
  - Payload Mass vs LaunchSite (Scatter), Success Rate vs Orbit Type (Bar)
  - FlightNumber vs Orbit Type (Scatter), PayloadMass vs Orbit Type (Scatter)
  - SuccessRate vs Year (Line)

- **Scatter** highlights relationships and non-linearities between numeric/discrete features and the target that guide feature engineering and model choice.
- **Bar** charts summarize categorical effects (site/orbit) on probability of success - the most actionable comparisons.
- **Line** charts reveal temporal trend/learning curve useful for context and potential time features.

GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- **SQL queries performed**
  - Listed distinct launch site names present in the dataset.
  - Retrieved five sample rows where the site label starts with 'CCA'.
  - Computed the total payload mass for missions whose Customer = 'NASA (CRS)'.
  - Calculated the mean payload mass for booster version F9 v1.1.
  - Identified the earliest date on which a landing succeeded on a ground pad.
  - Found booster names that succeeded on a drone ship with payload mass > 4000 and < 6000 kg.
  - Produced counts of mission outcomes grouped into Success vs Failure.
  - Determined which booster version(s) carried the maximum payload mass.
  - For calendar year 2015, listed failed drone-ship landings with their booster versions and launch sites.
  - Ranked the frequency of landing outcomes (e.g., *Failure (drone ship)* vs *Success (ground pad)*) within the date window 2010-06-04 to 2017-03-20, sorted descending

GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Site circles + labels: For each launch site we drew a small folium.Circle at its latitude/longitude and attached a popup/tooltip with the site name.

Why: Gives immediate geographic context (where sites are, proximity to coast/equator) and stable anchors when zooming.

Clustered launch markers (outcomes): Every launch became a folium.Marker added to a MarkerCluster; the icon uses a white pin with a red/green "info" glyph (green = success, red = failure) and a popup summarizing site and outcome. Clustering was configured with disableClusteringAtZoom=12 so individual markers appear when zoomed in.

Why: Lets us scan density at low zoom and inspect individual outcomes near a pad at high zoom without overplotting.

Layer control and map fit: Added LayerControl and fit_bounds to frame all sites on load.

Why: Improves usability—users can toggle layers and the map opens centered on relevant data.

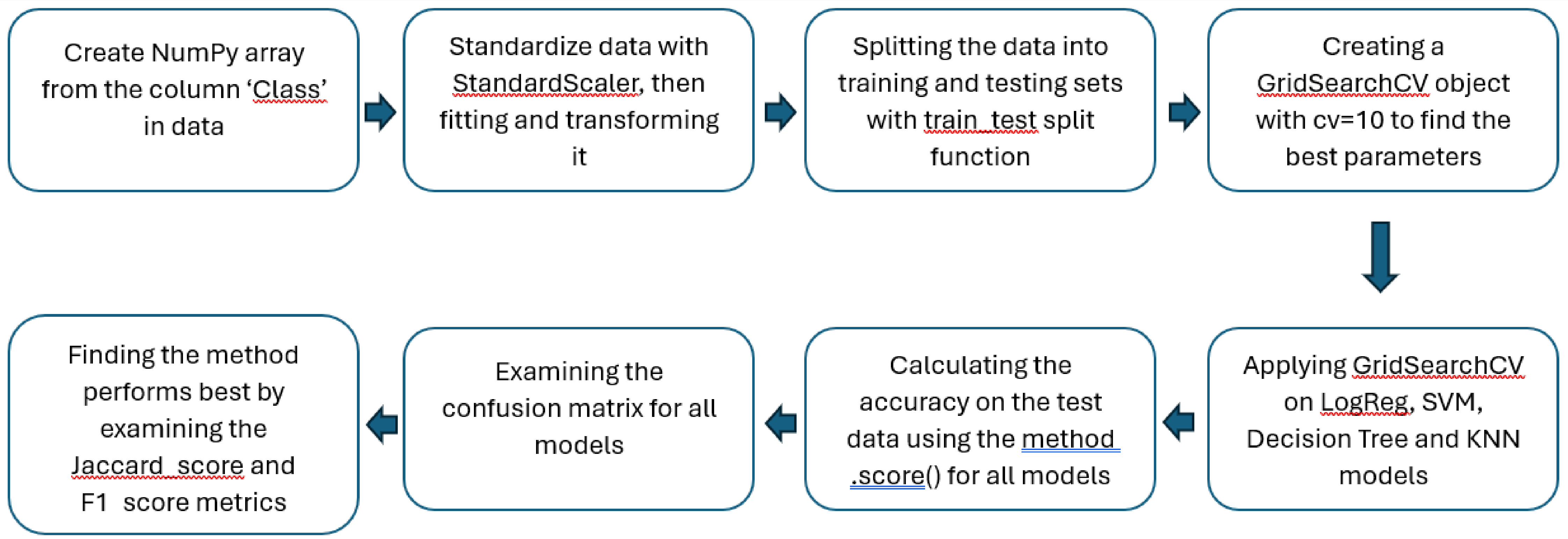
GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- **Components (plots & interactions)**
  - Launch Site dropdown (All Sites + each site): drives all charts via callbacks.
  - Success vs Failure pie (global or selected site): shows outcome mix at a glance.
  - Payload Mass range slider: filters all plots by payload window.
  - Scatter: PayloadMass vs FlightNumber, color=Class, facet by Booster Version: reveals learning-by-flight and mass effects across hardware.
  - Bar/Point: Success Rate by Orbit (filtered by site/range): categorical comparison on the target.
  - Line: Yearly Success Rate (filtered): trend over time.
  - KPI cards: total launches, success rate, selected-site stats.
- **Why these choices**
  - The dropdown + slider provide intuitive global filters without clutter.
  - Pie communicates overall balance of outcomes instantly.
  - Scatter exposes interactions (mass × experience × hardware) relevant for modeling.
  - Bar/Point isolates categorical influence (orbit/site) on success probability.
  - Line contextualizes improvements and stability over time.
  - KPIs give quick status for stakeholders before diving into the charts.

GitHub URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



Github URL: https://github.com/Franz-B14/Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

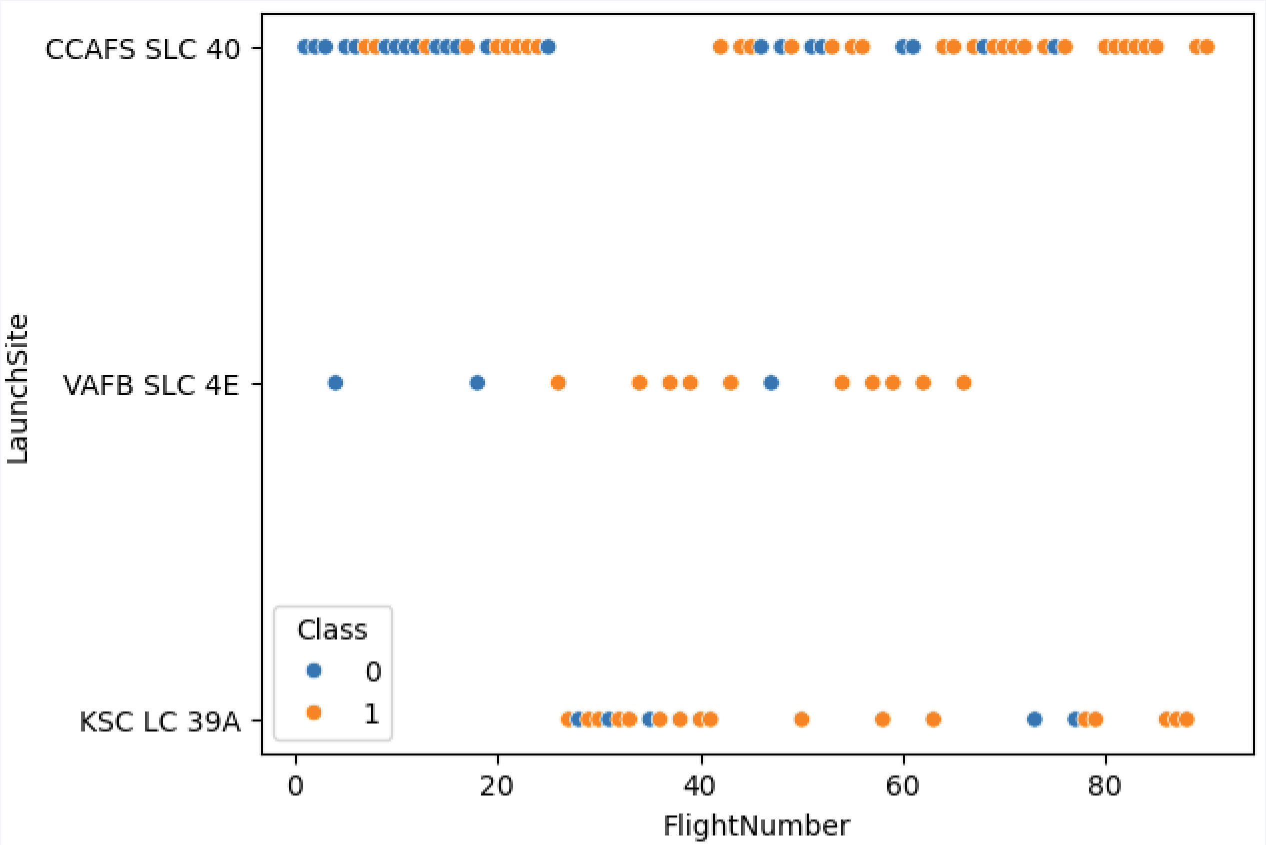
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

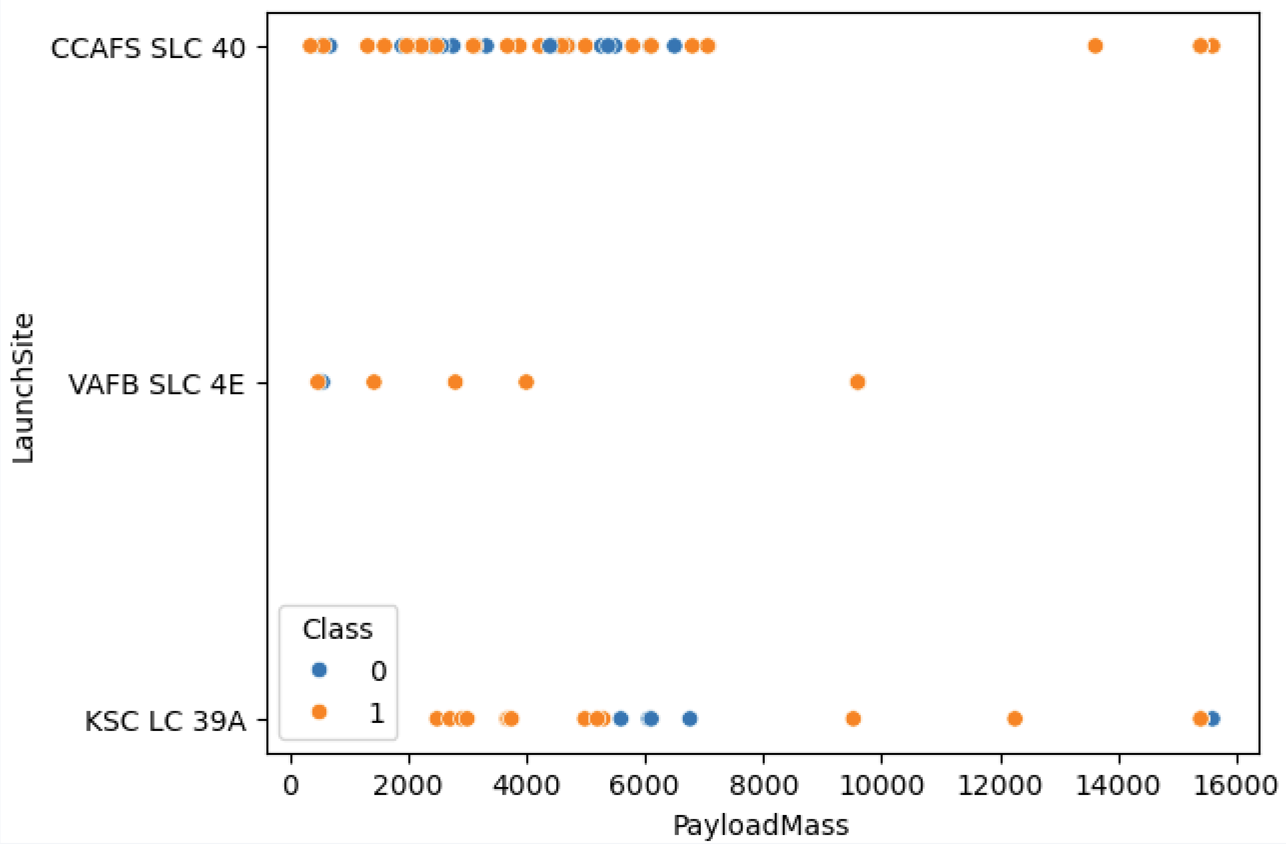
Flight Number vs. Launch Site

- The earliest missions show mostly failures, while later flights are predominantly successful.
- **CCAFS SLC-40** accounts for roughly half of all launches in the dataset.
- **VAFB SLC-4E** and **KSC LC-39A** exhibit comparatively higher success proportions.
- Overall, success probability appears to increase with flight sequence, suggesting a learning effect.



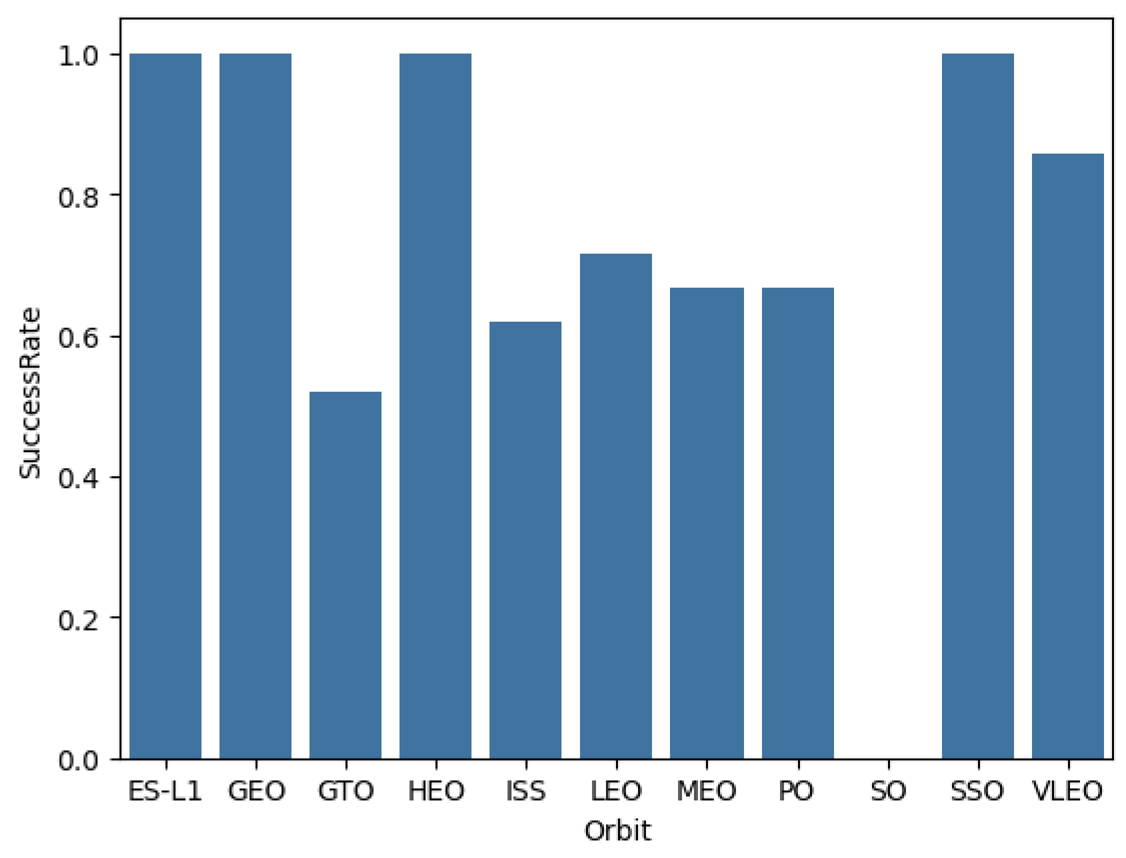
Payload vs. Launch Site

- Across all sites, greater payload masses are generally associated with higher landing success.
- The majority of missions carrying more than 7,000 kg payloads were successful.
- At **KSC LC-39A**, missions with payloads below 5,500 kg also achieved a 100% success rate.



Success Rate vs. Orbit Type

- **Orbits with perfect success:** ES-L1, GEO, HEO, and SSO show 100% landing success in our sample.
- **Orbits with no success:** SO records a 0% success rate.
- **Mid-tier performance (≈50–85%):** GTO, ISS, LEO, MEO, and PO fall in the moderate success range.

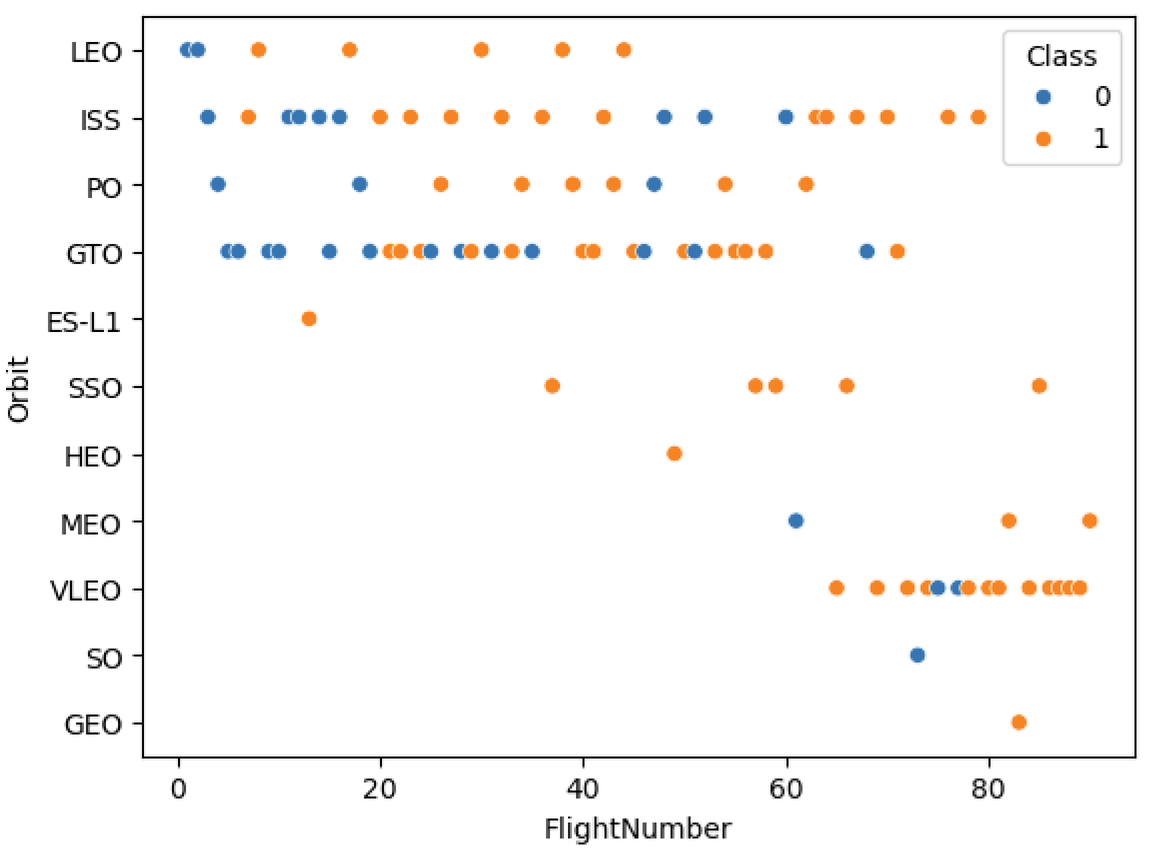


Flight Number vs. Orbit Type

- In **LEO**, success (Class = 1) increases with FlightNumber, suggesting a clear learning-curve effect.

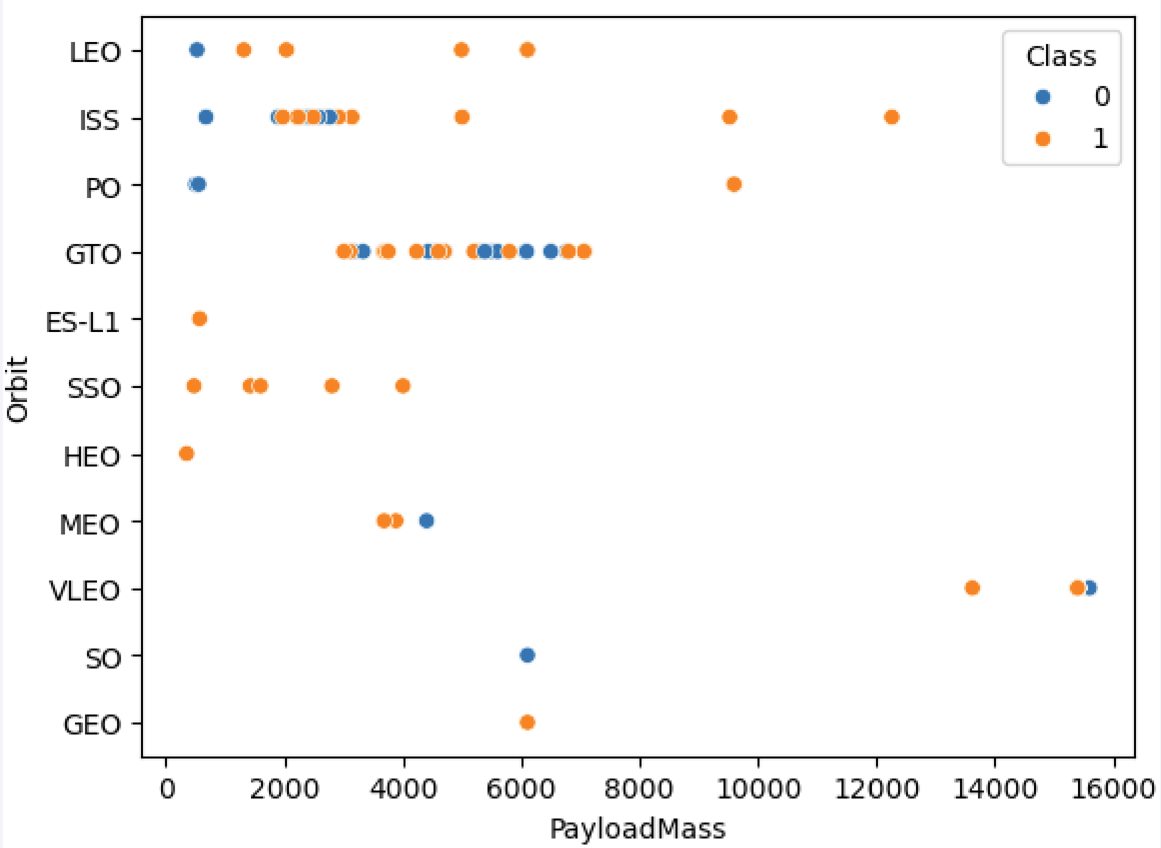
- In **GTO**, outcomes look flat with respect to FlightNumber—no obvious relationship between experience and success.

- Other orbits show mixed or sparse patterns, so any trend with flight sequence is weak or inconclusive.



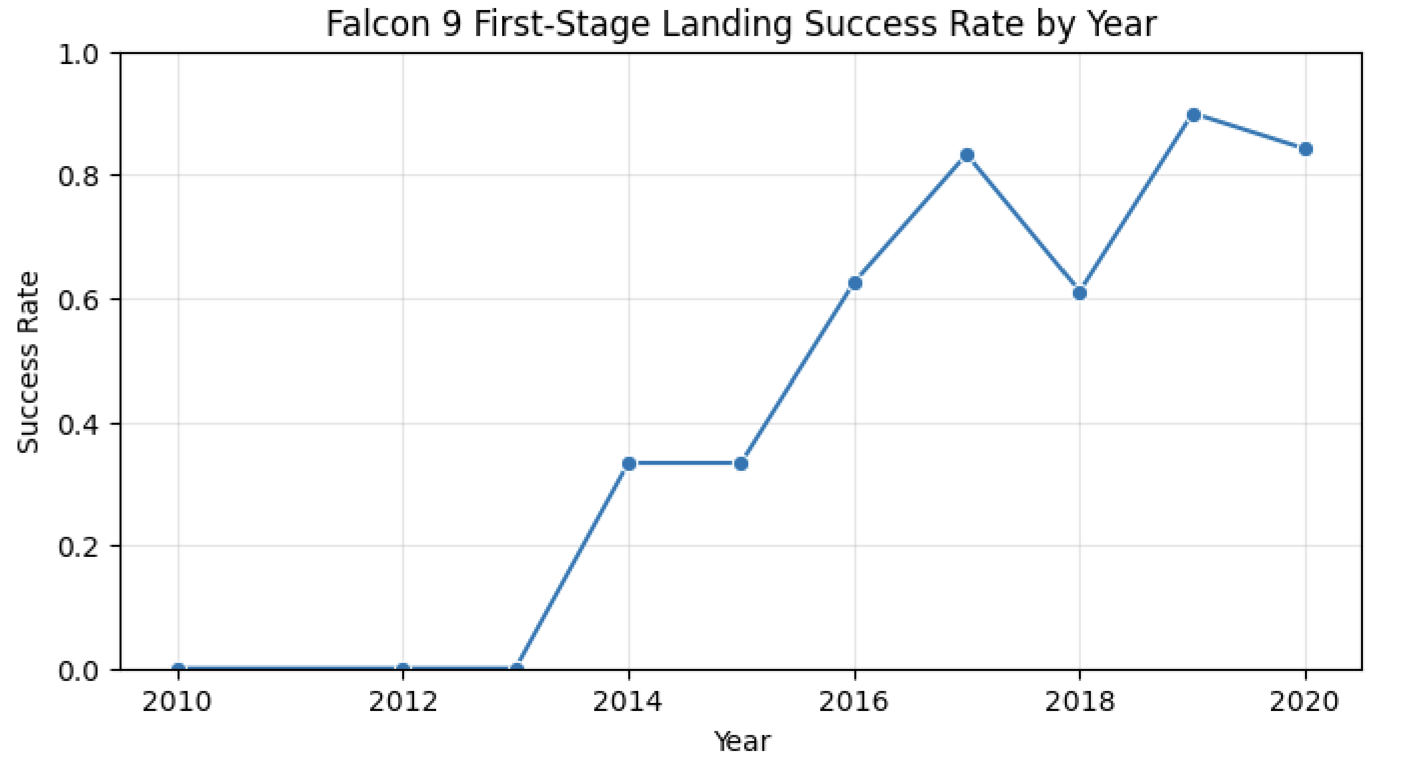
Payload vs. Orbit Type

- The relationship between payload mass and landing success depends on the orbit - mass alone doesn’t dictate the outcome.
- GTO shows many successful missions at moderate - high masses with a few mid-mass failures.
- LEO/ISS flights are mostly successful across a wide mass range; lower-mass launches show more variability.
- Very heavy payloads (≥13 t) are rare but tend to be successful in our sample, so orbit-specific effects should be interpreted with care



Launch Success Yearly Trend

- From the line chart we can notice that success rate for first-stage landing kept increasing from 2013 until 2019



All Launch Site Names

[12]:
```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE;
```

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

[12]:

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

- Explanation:
  - List of unique launch sites in space mission

Launch Site Names Begin with 'CCA'

[34]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

[34]:

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

- Explanation:
  - Displaying 5 records where launch site names begin with ‘CCA’

Total Payload Mass

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

[17]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

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

[17]:

total_payload_mass
45596

- Explanation:
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

[19]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

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

[19]:

avg_payload_mass
2928.4

- Explanation:
  - Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[22]: %%sql
      SELECT MIN(Date) AS first_success_ground_pad
      FROM SPACEXTABLE
      WHERE Landing_Outcome = 'Success (ground pad)';
```

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

[22]:

first_success_ground_pad
2015-12-22

- Explanation:
  - Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[24]:
```
%%sql
SELECT DISTINCT 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.
```

[24]:

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

- Explanation:
  - Listing the name of the boosters that had success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

[27]:
```
%%sql
SELECT
  CASE
    WHEN LOWER(TRIM("Mission_Outcome")) LIKE 'success%' THEN 'Success'
    ELSE 'Failure'
  END AS Result,
  COUNT(*) AS Count
FROM SPACEXTABLE
GROUP BY Result;
```

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

[27]:

Result	Count
Failure	1
Success	100

- Explanation:
  - Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

- Explanation:
  - Listing all the booster versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

[31]:
```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

[31]:

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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[35]:
```
%%sql
SELECT
  CAST(substr(Date,6,2) AS INTEGER) AS Month,
  CASE
    WHEN LOWER("Landing_Outcome") LIKE '%drone ship%' THEN 'drone ship'
    WHEN LOWER("Landing_Outcome") LIKE '%ground pad%'  THEN 'ground pad'
    WHEN LOWER("Landing_Outcome") LIKE '%ocean%'       THEN 'ocean'
    WHEN LOWER("Landing_Outcome") LIKE '%parachute%'   THEN 'parachute'
    ELSE COALESCE(TRIM("Landing_Outcome"), 'unknown')
  END AS Ship,
  "Booster_Version",
  "Launch_Site",
  COUNT(*) AS FailureCount
FROM SPACEXTABLE
WHERE substr(Date,1,4) = '2015'
  AND LOWER(TRIM("Mission_Outcome")) NOT LIKE 'success%'
GROUP BY Month, Ship, "Booster_Version", "Launch_Site"
ORDER BY Month, FailureCount DESC;
```

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

[35]:

Month	Ship	Booster_Version	Launch_Site	FailureCount
6	drone ship	F9 v1.1 B1018	CCAFS LC-40	1

- Explanation:
  - Listing the records displaying the month in which there was a failed landing outcome, the ship type, booster version and launch site in the year 2015

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

[36]:

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

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

[36]:

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

- Explanation
  - Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20

Section 3

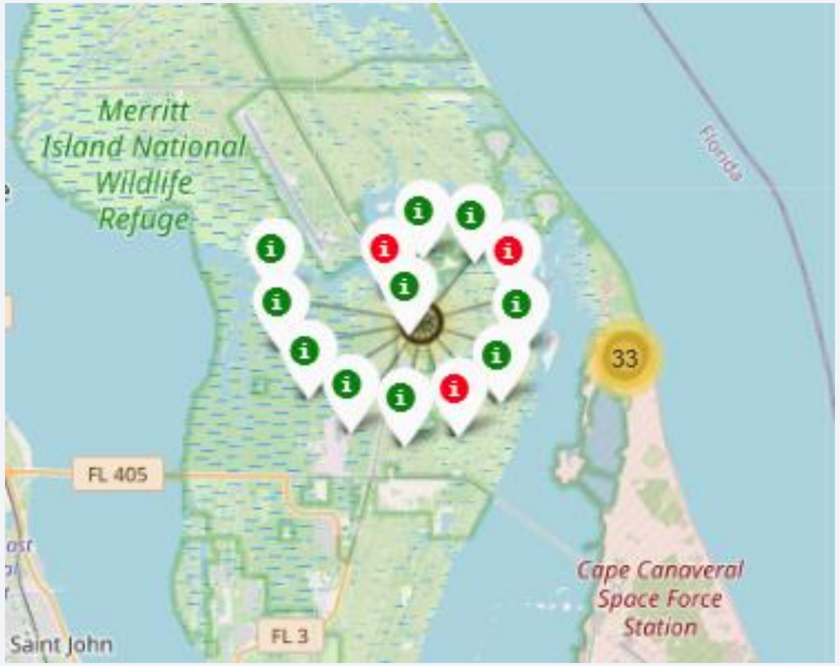
Launch Sites Proximities Analysis

Launch sites' location markers on global map

- The four pads cluster on U.S. coasts: KSC LC-39A and CCAFS SLC-40 on Florida's Atlantic coast; VAFB SLC-4E on California's Pacific coast. Coasts allow down-range flight over ocean, reducing public-safety risk from debris or aborts.
- Florida sites sit at low latitude, leveraging the Earth's rotational speed to boost eastward, equatorial-to-inclined orbits (energy savings and higher payload).
- Vandenberg (SLC-4E) is ideal for polar/sun-synchronous orbits because launches can head south over the Pacific without overflying populated areas.
- All markers are near sea level and accessible infrastructure, which simplifies logistics and recovery operations.

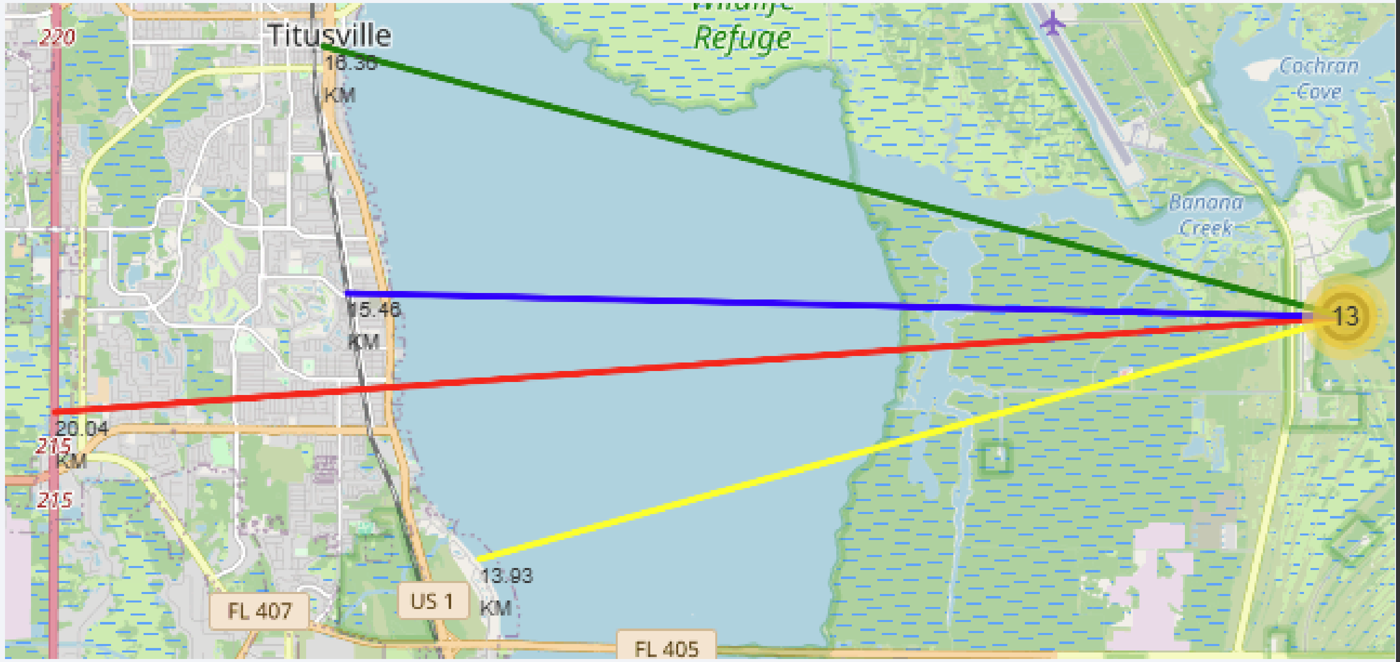
Color-labeled launch outcomes on the map

- Each pin is a launch record at the pad location; green = success, red = failure
- Pins are placed inside a MarkerCluster so overlapping launches stack neatly; zooming in fans out the markers to reveal individual events.
- The brown circle overlay highlights the pad vicinity and helps orient the cluster spatially.
- At the shown site (e.g., KSC LC-39A), the green-to-red ratio is high, indicating a strong success rate after the early flights.
- Popups on each marker summarize site + outcome, allowing quick auditing of nearby successes/failures and visual comparison across pads.



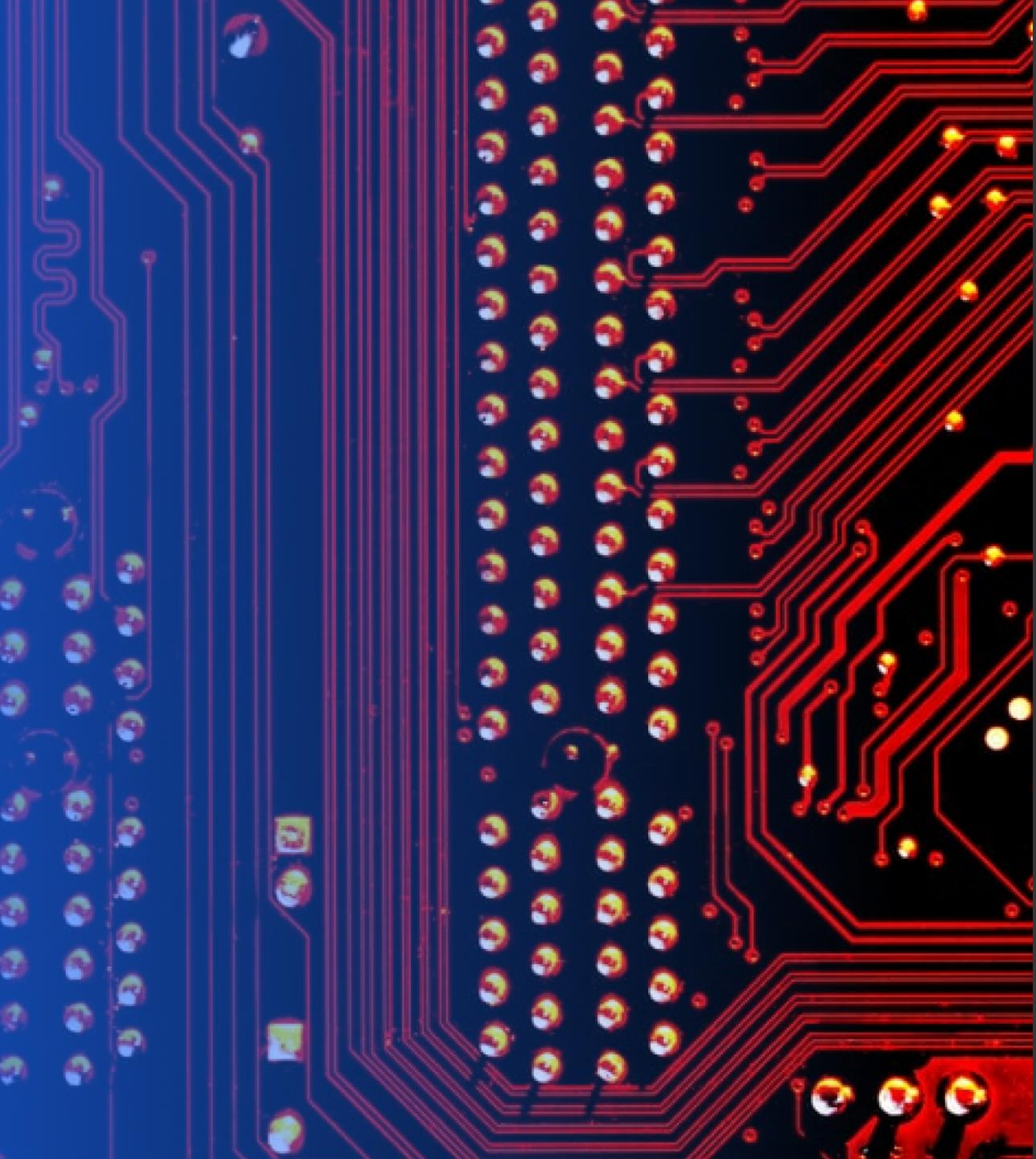
Distance of KSC LC-39A to its proximities

- LC-39A sits ~15–20 km from major transport corridors (railway, highway) and ~14 km from the coastline
- Nearest city: Titusville (~16 km) - within a distance a debris cloud could reach in minutes in a worst-case failure.
- The site is surrounded by water/wetlands, which helps buffer risk downrange but concentrates evacuation routes along a few bridges/roads.

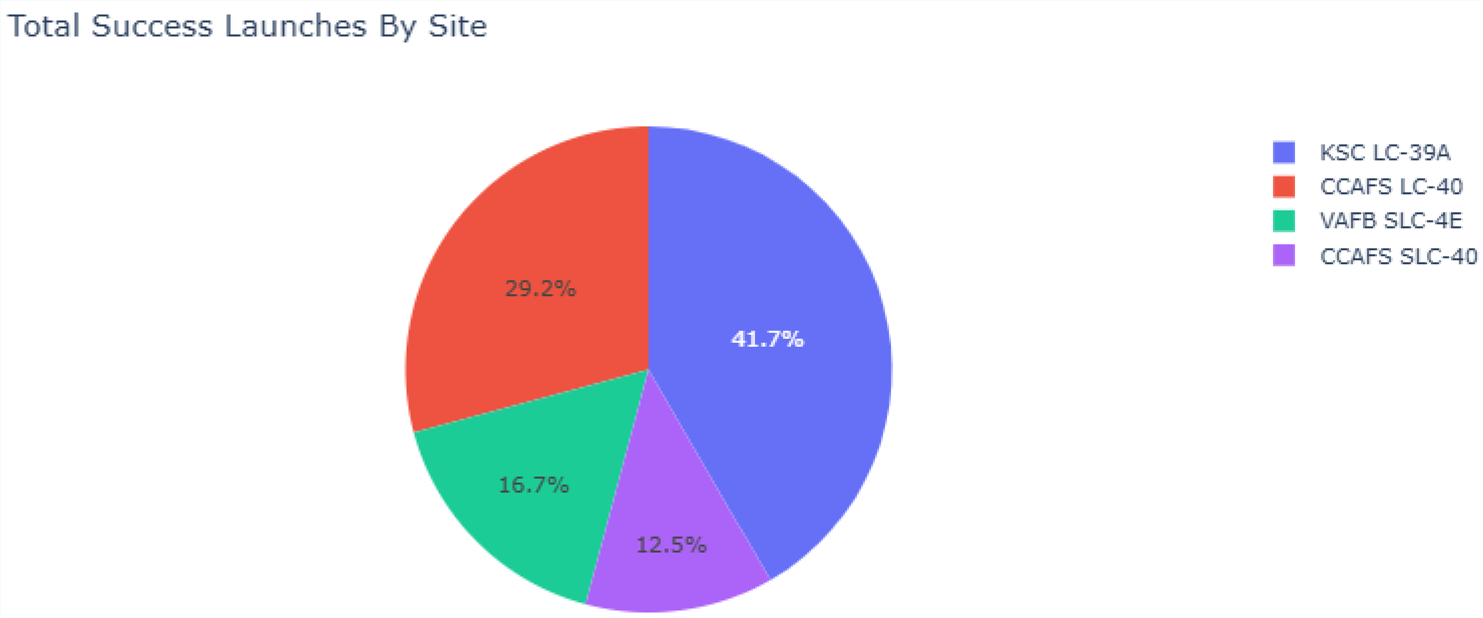


Section 4

Build a Dashboard with Plotly Dash

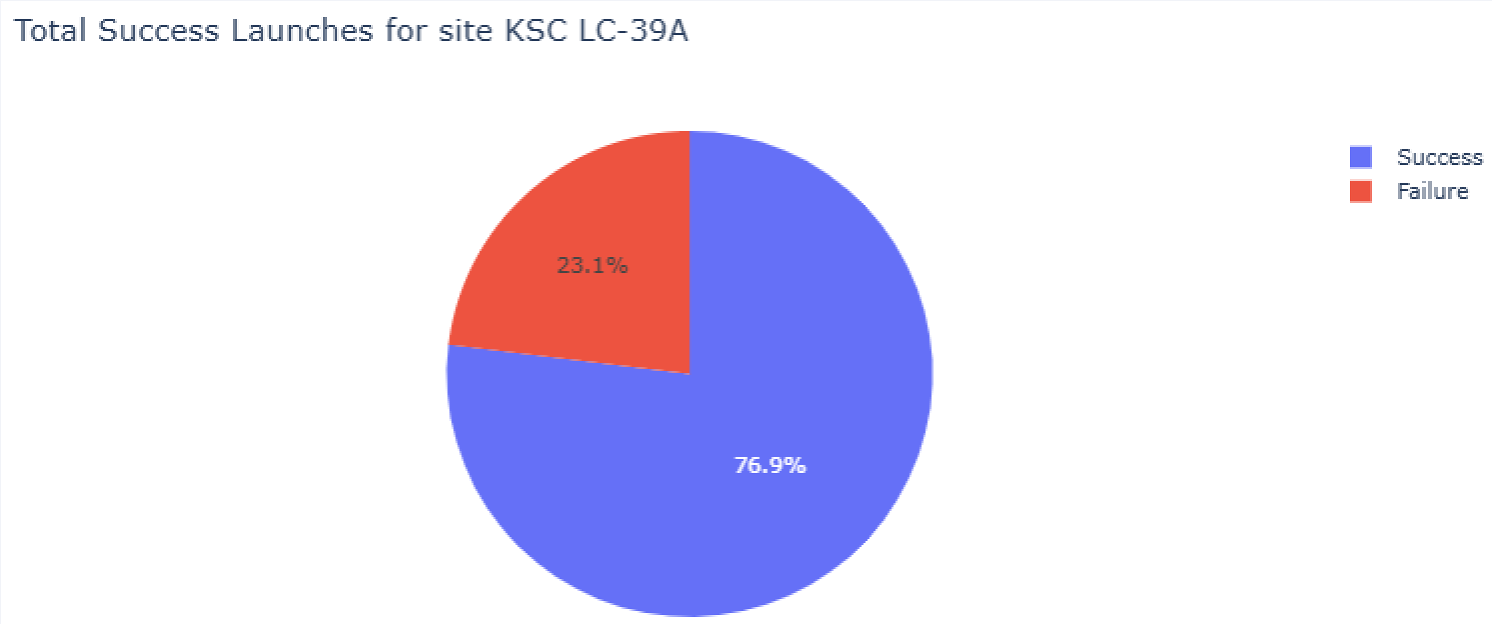


Total Success Launches by Site



- The chart is showing the total successful launches by site, with KCS LC-39A having the most successful ones

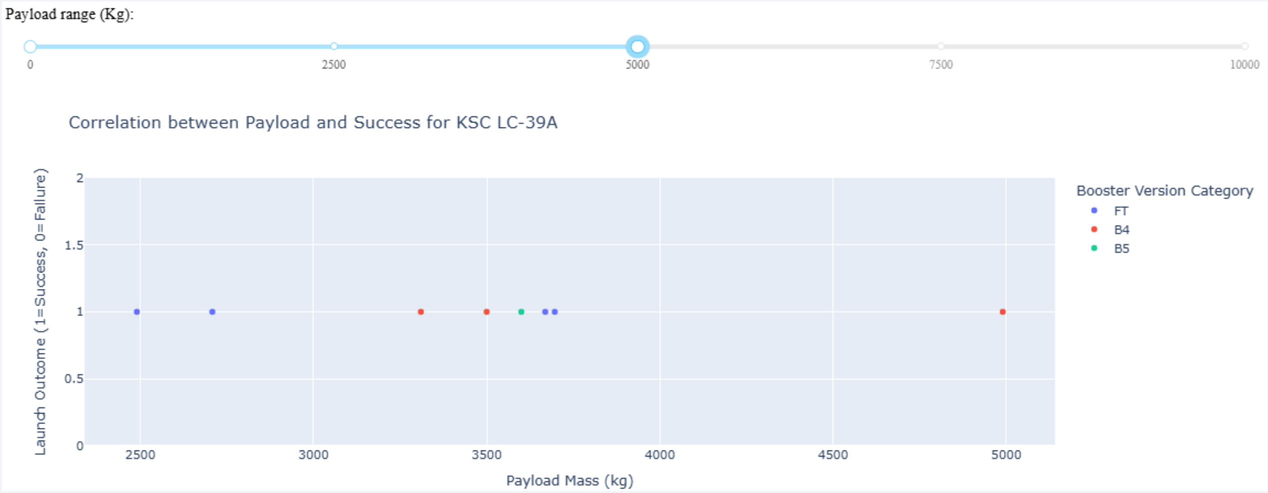
Launch site with highest launch success

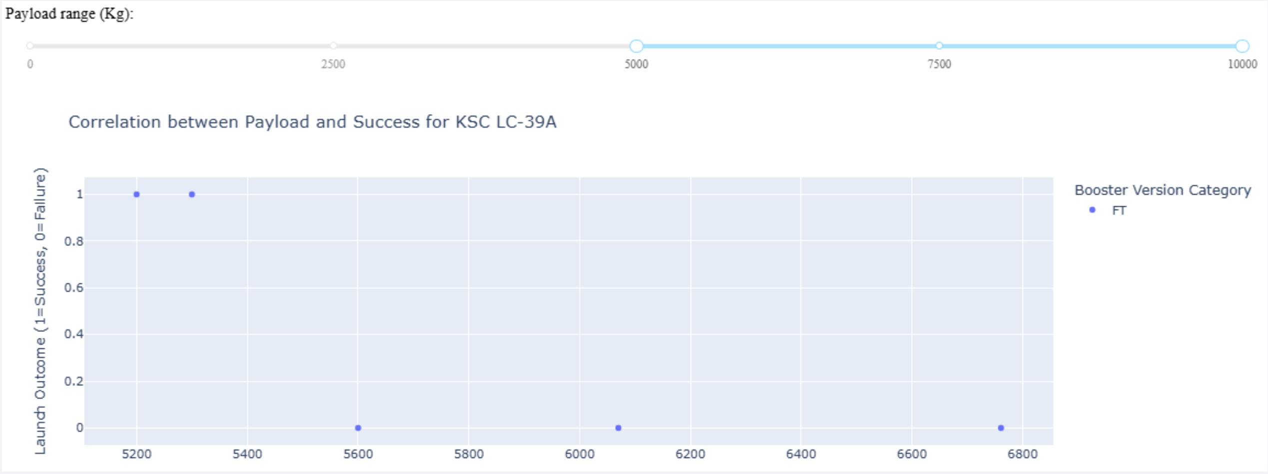


- KSC LC-39A has the highest launch success rate at 76.9%
- 10 launches were successful while 3 have failed

Payload Mass vs Launch Outcome (all sites)

- After filtering across all sites, most successful launches cluster between ~2,000 and 5,500 kg; this band shows the densest concentration of green points.
- Very light payloads (<1,000 kg) have mixed outcomes, and mid-heavy masses (≈6–8 t) include several failures; however, counts there are lower.
- Where visible, newer booster versions (e.g., FT / Block 5) dominate the successful points at the higher end of the range, suggesting hardware maturity helps heavier missions.



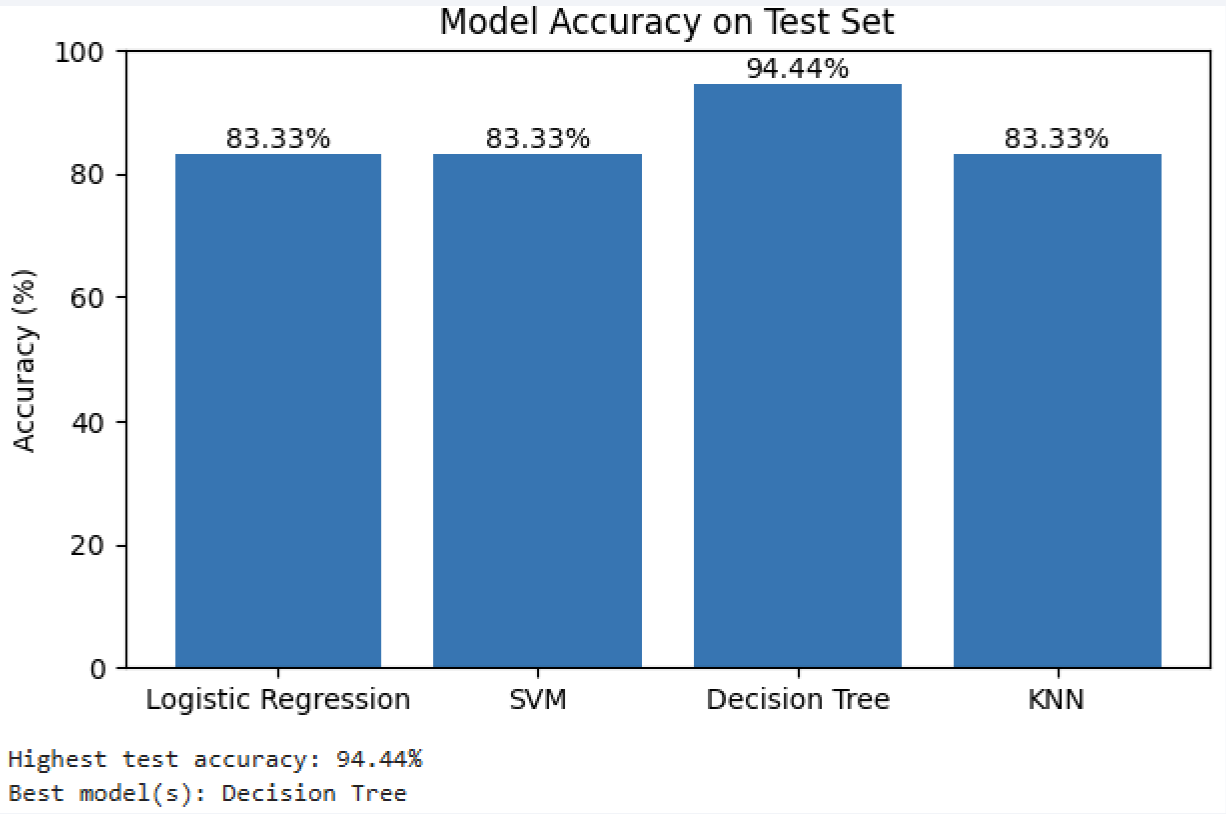


Section 5

Predictive Analysis (Classification)

Classification Accuracy

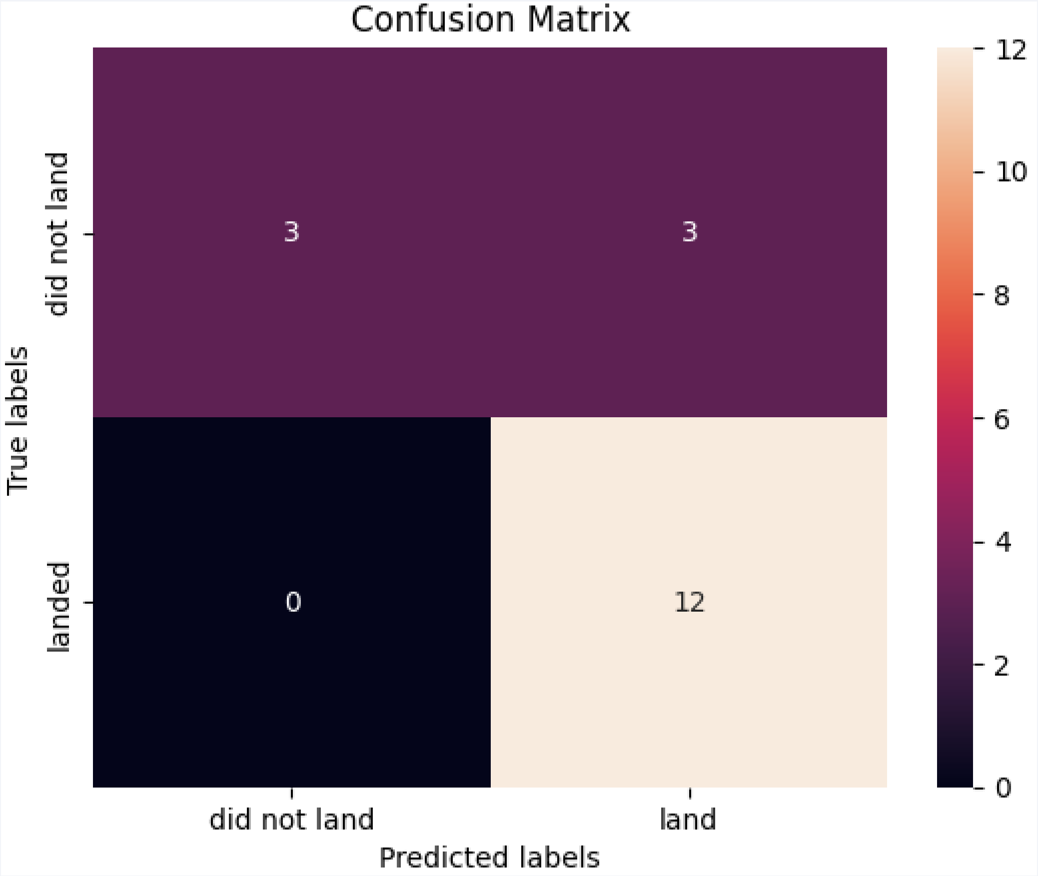
- Highest test accuracy: 94.44%
- Best model: Decision Tree at 94.44% on test
- Others: SVM, KNN, Logistic Regression tie at 83.33%



Confusion Matrix

- **True negatives (3):** correctly predicted “did not land.”
- **False positives (3):** predicted land but it actually did not (20% of positive predictions were wrong → precision 0.80).
- **False negatives (0):** none missed; every true landing was predicted as landing
- **True positives (12):** correctly predicted landings.

The model captures all real landings (no misses), but occasionally over-predicts success (3 false alarms). It’s ideal if missing a successful landing is worse than a false alarm; if false positives are costly, one should consider a higher decision threshold or a model tuned for higher precision.



Conclusions

1) Decision Tree emerged as the top-performing model for this dataset.
2) Missions with lower payload mass tend to achieve higher landing success than heavier payloads.
3) Launch sites are mostly near the equator and close to the coastline.
4) The success rate has improved over time, showing an upward trend across years.
5) KSC LC-39A records the highest success rate among all launch sites.
6) For orbits ES-L1, GEO, HEO, and SSO, observed missions show a 100% success rate in this data.

Thank you!