Winning Space Race with Data Science

Jhermien Paul Alejandria
Data Scientist

github.com/jhermienpaul

Outline

- Executive Summary
- Introduction
- Methodology
- Findings/Insights
- Conclusion
- Appendix

Executive Summary

This project focuses on predicting the success of SpaceX Falcon 9 first-stage landings, critical for reducing launch costs through reusability. By accurately forecasting landing outcomes, it enables better cost estimation, potentially benefiting competitive bidding processes against SpaceX. Utilizing data collected via REST APIs and web scraping, the analysis applies data wrangling, exploratory data analysis (EDA), interactive visualizations, and machine learning classification models to identify key success factors.

The analysis revealed significant predictors such as payload mass, orbit type, and launch sites influencing landing outcomes. Interactive dashboards and maps provided intuitive insights, while machine learning models achieved high accuracy in predicting landing successes, underscoring the potential for cost-effective and strategic planning for rocket launches.

Introduction

- Project Background
  - This project was conducted in the context of understanding and modeling Falcon 9 landing outcomes essential for maximizing reusability, minimizing launch costs, and driving SpaceX's competitive advantage in commercial spaceflight..
- Statement of the Problem
  - Determine the key drivers and build a predictive model for Falcon 9 first-stage landing success to enable accurate, data-driven cost and operational decisions for SpaceX launches.
- Objectives
  - Collect and integrate historical launch data for Falcon 9 missions
  - Perform data wrangling to clean and combine datasets into a structured form
  - Conduct exploratory data analysis (EDA) to uncover patterns and relationships
  - Build and evaluate a machine learning pipeline to predict landing success

Section 1

Methodology

Methodology

Summary of Methodology

- Data collection method
  - REST API and Web Scraping
- Data wrangling method
  - Explore, clean, transform, integrate, and validate data
- Exploratory data analysis (EDA) using Python and SQL
- Interactive visualizations using Folium and Plotly Dash
- Predictive analysis using classification models
  - Logistic Regression, Decision Trees, KNN, and SVM

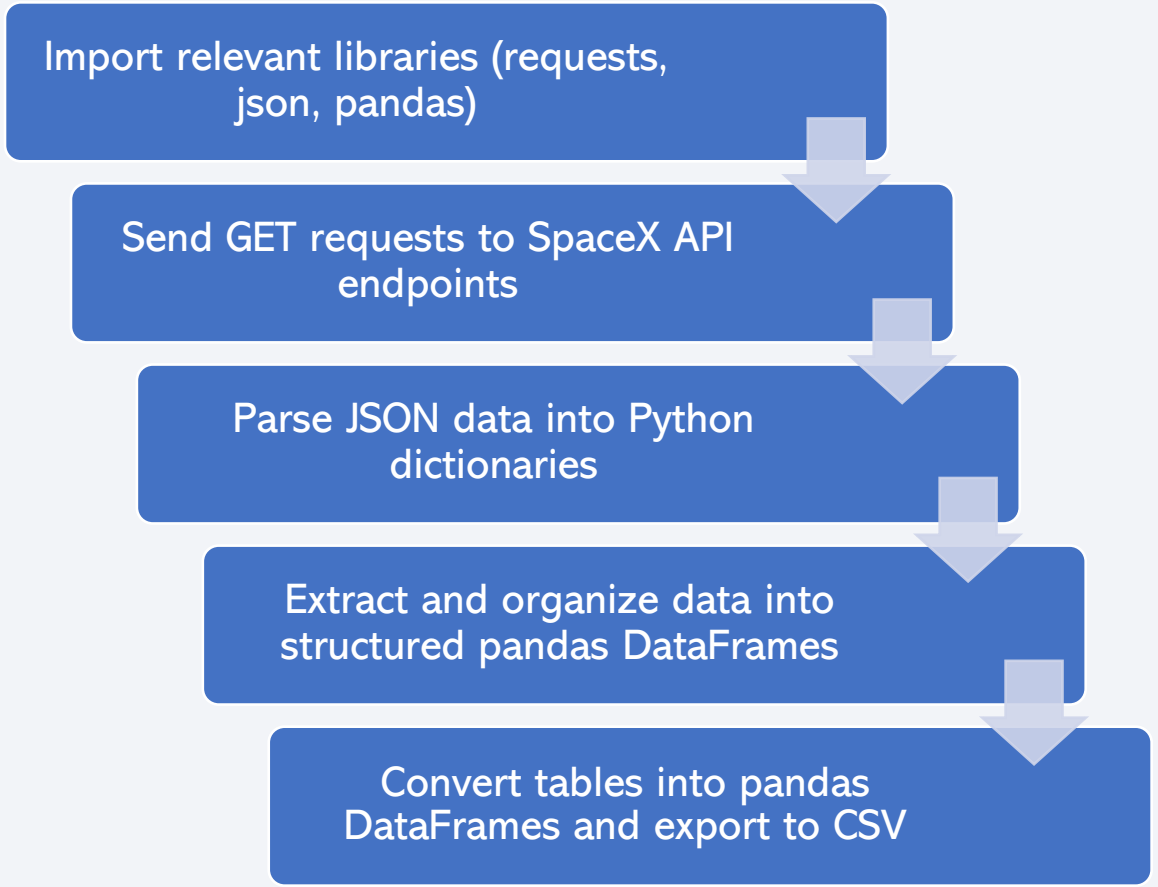
Data Collection

- REST API (SpaceX API: *Rockets, Launchpads, Payloads, Cores*)
  - Provides accurate, real-time data on launches, vehicles, and mission parameters essential for building reliable landing prediction models.
  - Ensures data integrity and reproducibility through standardized, structured endpoints, minimizing manual errors and version inconsistencies.
  - Direct access to operational features (e.g., payload mass, rocket specs, site details) enables granular analysis and precise cost modeling aligned with business goals.
- Web Scraping (Wikipedia: *List of Falcon 9 and Falcon Heavy launches*)
  - Captures historical launch outcomes and additional mission metadata not available in the official API, expanding the feature set for more robust predictive modeling.
  - Fills critical data gaps and cross-validates API results, strengthening overall data completeness and reliability for decision-making.
  - Aggregates community-curated records, including anomalies and edge cases, to support comprehensive cost estimation and competitive analysis scenarios.

Data Collection with REST API

- Accessing the SpaceX REST API provides authoritative, real-time metrics (e.g., payload mass, orbit type, core reuse details) that form the core features for predicting first-stage landing success and downstream cost estimates.
- By parsing JSON into structured pandas DataFrames and exporting to CSV, we ensure a reproducible, high-fidelity data pipeline that directly feeds our machine learning models for reliable launch cost analysis.
- Jupyter notebook:

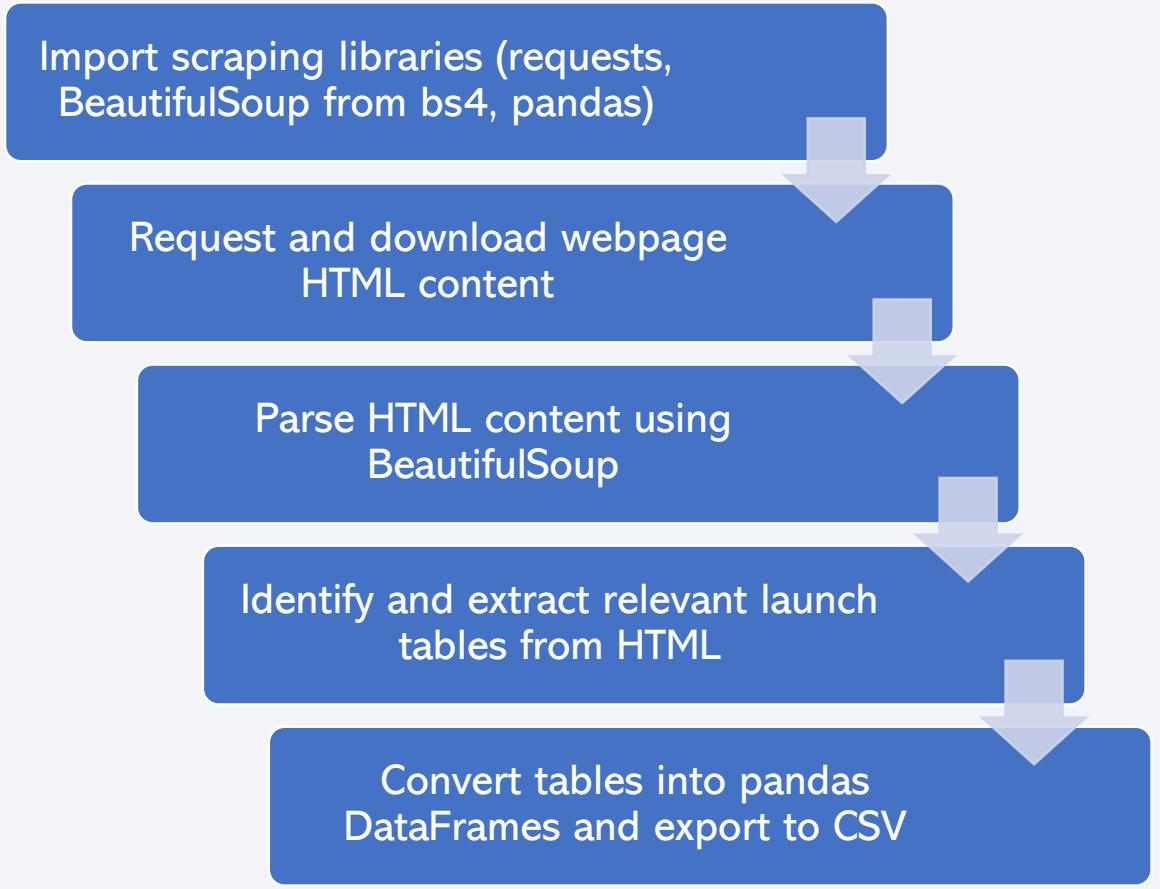
  Data Collection with API.ipynb



Data Collection with Web Scraping

- Scraping Wikipedia's Falcon 9 and Heavy launch tables captures historical anomalies, landing outcomes, and mission metadata not exposed via API, enhancing feature diversity and model robustness for landing prediction.
- Converting HTML tables into pandas DataFrames aligns this supplemental dataset with API outputs, creating a unified, gap-free foundation for cost-driven analytical workflows.
- Jupyter notebook:

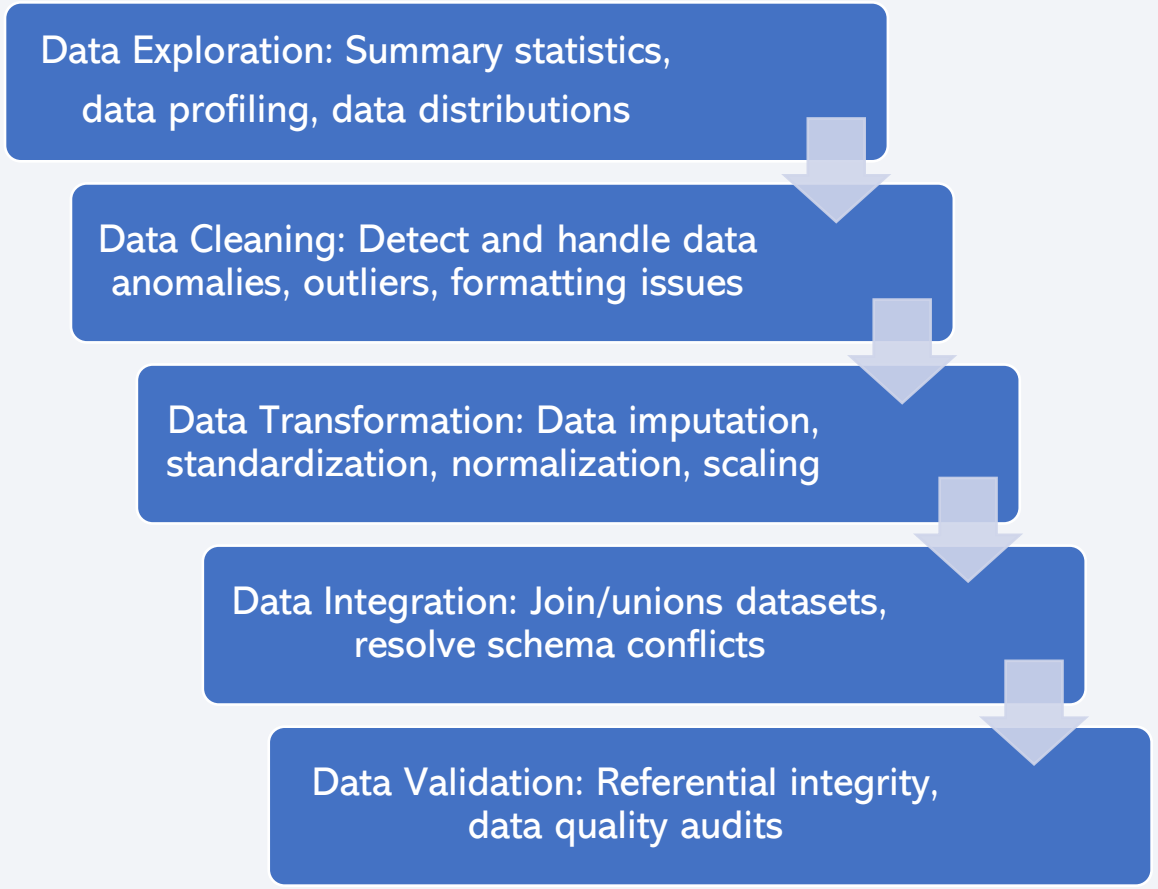
  Data Collection with Web Scraping.ipynb



Data Wrangling

- Implementing a full-spectrum wrangling workflow—from exploration and cleaning through transformation, integration, and validation—ensures data accuracy, consistency, and completeness, which are non-negotiable for high-stakes landing forecasts.
- Leveraging pandas, numpy, and sklearn at each phase fortifies our dataset against outliers, schema mismatches, and integrity violations, delivering trusted inputs for predictive modeling and cost optimization.
- Jupyter notebook:

  Data Wrangling.ipynb



EDA with Python Visualization

- Summary of Charts Plotted:
  - Scatter Plots: Explored relationships between launch sites, orbit types, flight numbers, payload mass, and mission outcomes. This helped identify patterns and anomalies in launch success and failure rates.
  - Bar Chart: Compared launch success rates across different orbit types to quickly pinpoint which orbits have the highest/lowest reliability.
  - Line Chart: Visualized the average annual success rate, highlighting SpaceX's dramatic improvements in launch reliability over time.

- Jupyter notebook: Exploratory Data Analysis with Visualization.ipynb

EDA with SQL Query

- Summary of SQL Queries Performed:
  - Queried and summarized key mission stats (unique launch sites, mission outcomes, payload totals/averages) to understand overall launch patterns.
  - Filtered records to answer specific business questions (e.g., booster performance, landing outcomes, milestone dates) using flexible SQL conditions and aggregations.
  - Used subqueries and grouping to benchmark technical achievements (e.g., max payloads, monthly/yearly trends, recovery methods).

- Jupyter notebook: Exploratory Data Analysis with SQL.ipynb

Interactive Mapping with Folium

- Summary of Map Objects and Rationale
  - Markers: Plotted each launch site as an interactive marker to visualize SpaceX's geographic distribution and enable site-specific insights.
  - Marker Clusters: Used marker clusters to avoid overlapping points and make dense regions (like Florida) easier to explore.
  - Distance Lines: Drew lines from launch pads to the nearest coastline to analyze site safety and proximity to the ocean.

- Jupyter notebook: Interactive Mapping with Folium.ipynb

Interactive Dashboard with Plotly Dash

- Plots, Graphs, and Interactions Added:
  - Pie charts to visualize launch success counts by site and by outcome (success vs. failure).
  - Dropdown menu for selecting specific launch sites, instantly updating all charts.
  - Scatter plot showing the relationship between payload mass, launch outcome, and booster version.
  - Payload range slider that lets users filter data by payload mass, updating the scatter plot in real time.

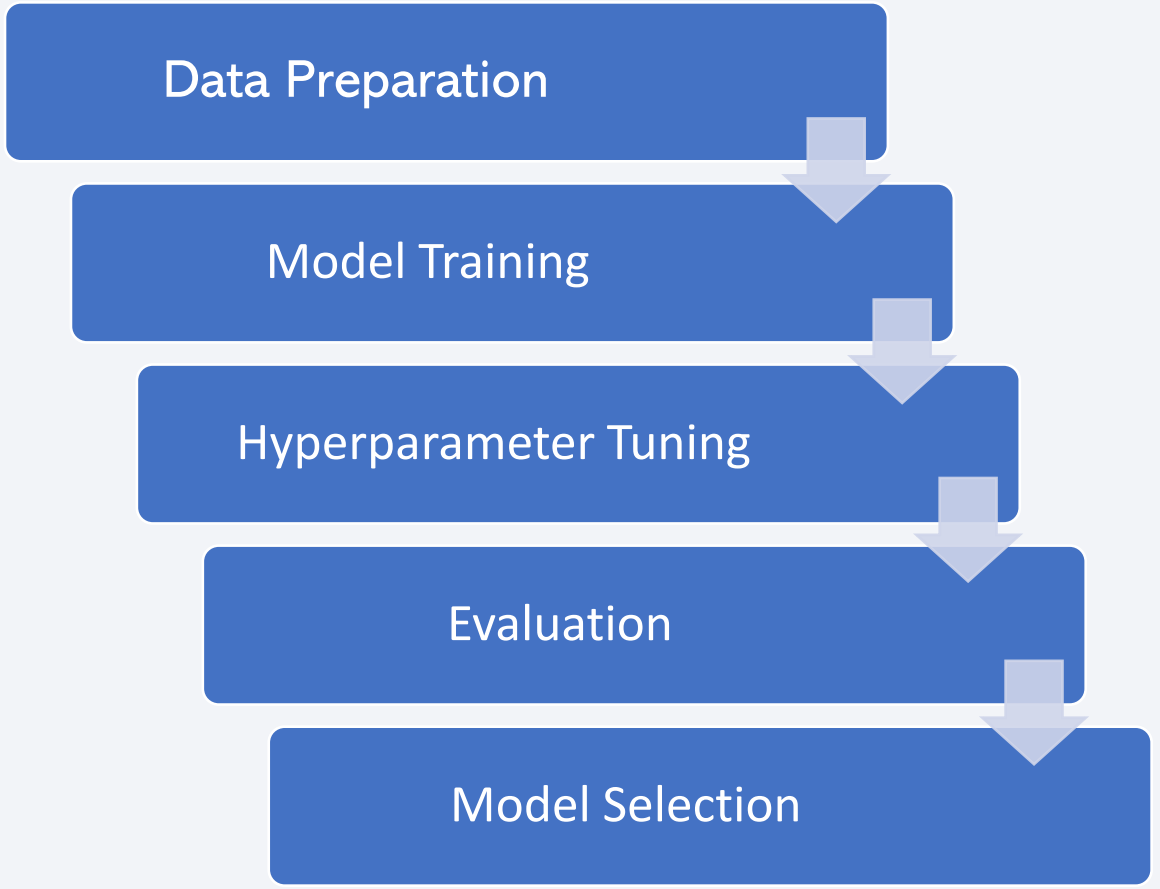
- Plotly Dash: Interactive Dashboard with Plotly Dash.py

Predictive Analysis with Classification Models

- ML Pipeline Stages
  - Built, tuned, and evaluated multiple classification models (Logistic Regression, SVM, Decision Tree, KNN) using a consistent pipeline with cross-validation, and selected the best performer based on test accuracy.

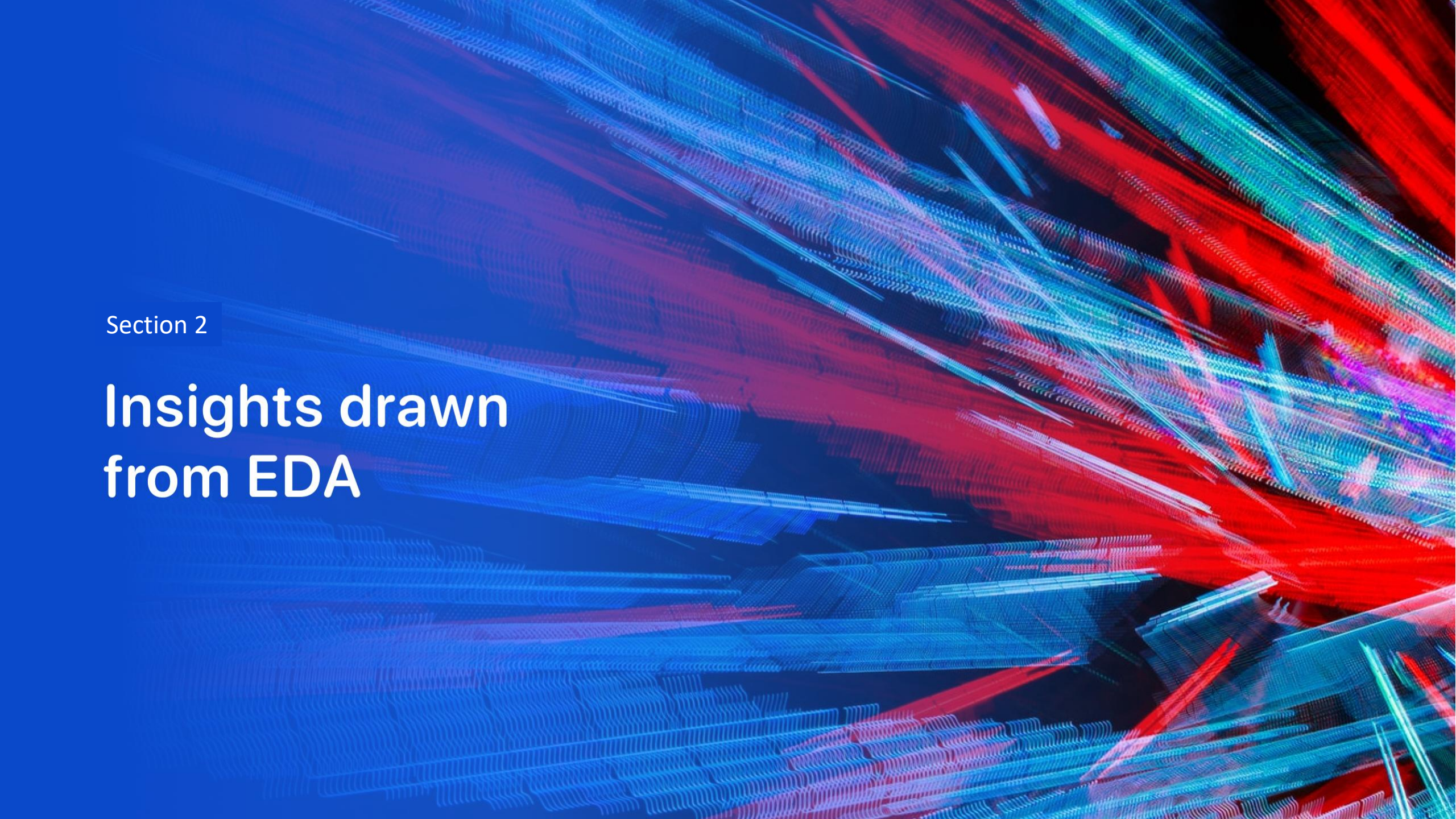
- Jupyter notebook:

  Machine Learning Model Pipeline.ipynb



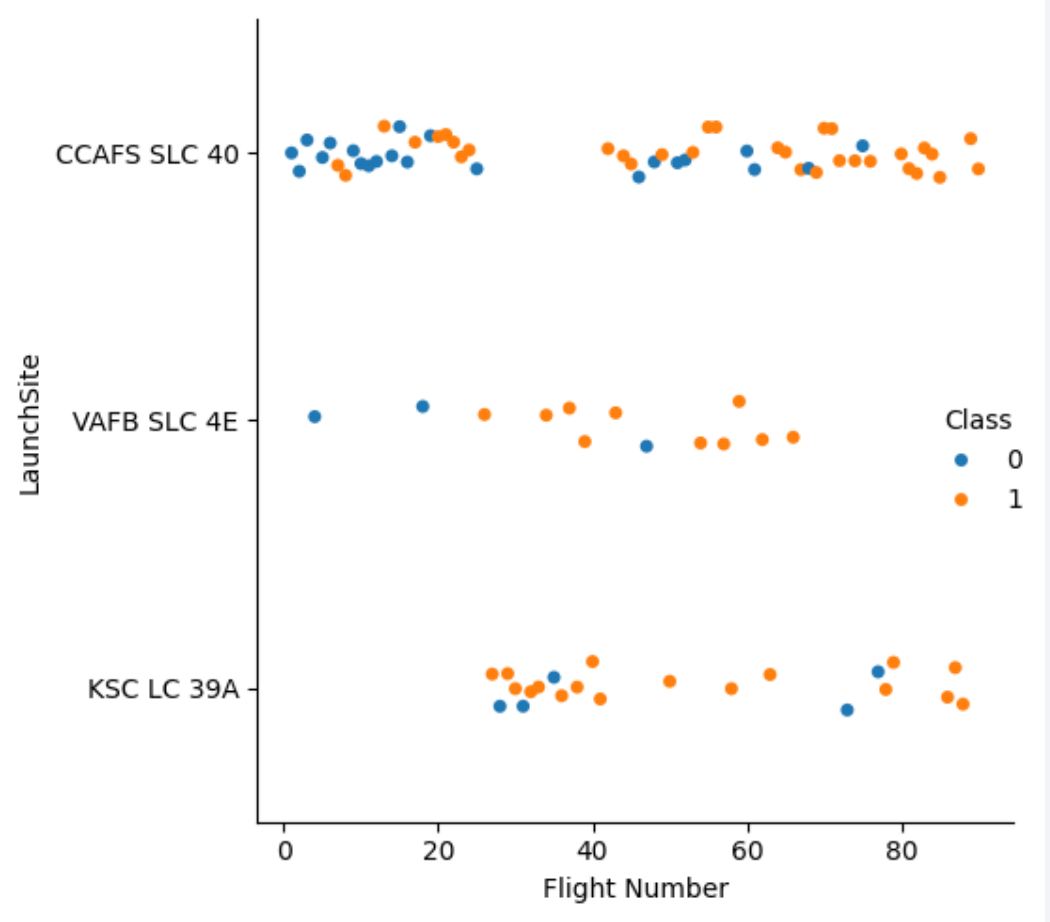
Findings/Insights

- Exploratory Data Analysis (EDA):
  - Found that most launches happened at two main sites in Florida and California, with launch frequencies and payload masses visualized through charts and interactive maps.
- Descriptive Analytics:
  - Showcased dynamic dashboards and maps where users can explore launch success rates by site, booster versions, and payload mass. Some launch sites consistently performed better, but larger payloads didn't always mean lower success. The dashboards made it easy to spot trends and outliers at a glance.
- Predictive Analysis:
  - Multiple machine learning models (Logistic Regression, SVM, Decision Tree, KNN) were built to predict landing success, all achieving 83% accuracy—solid performance, no clear "winner."


Section 2
Insights drawn from EDA

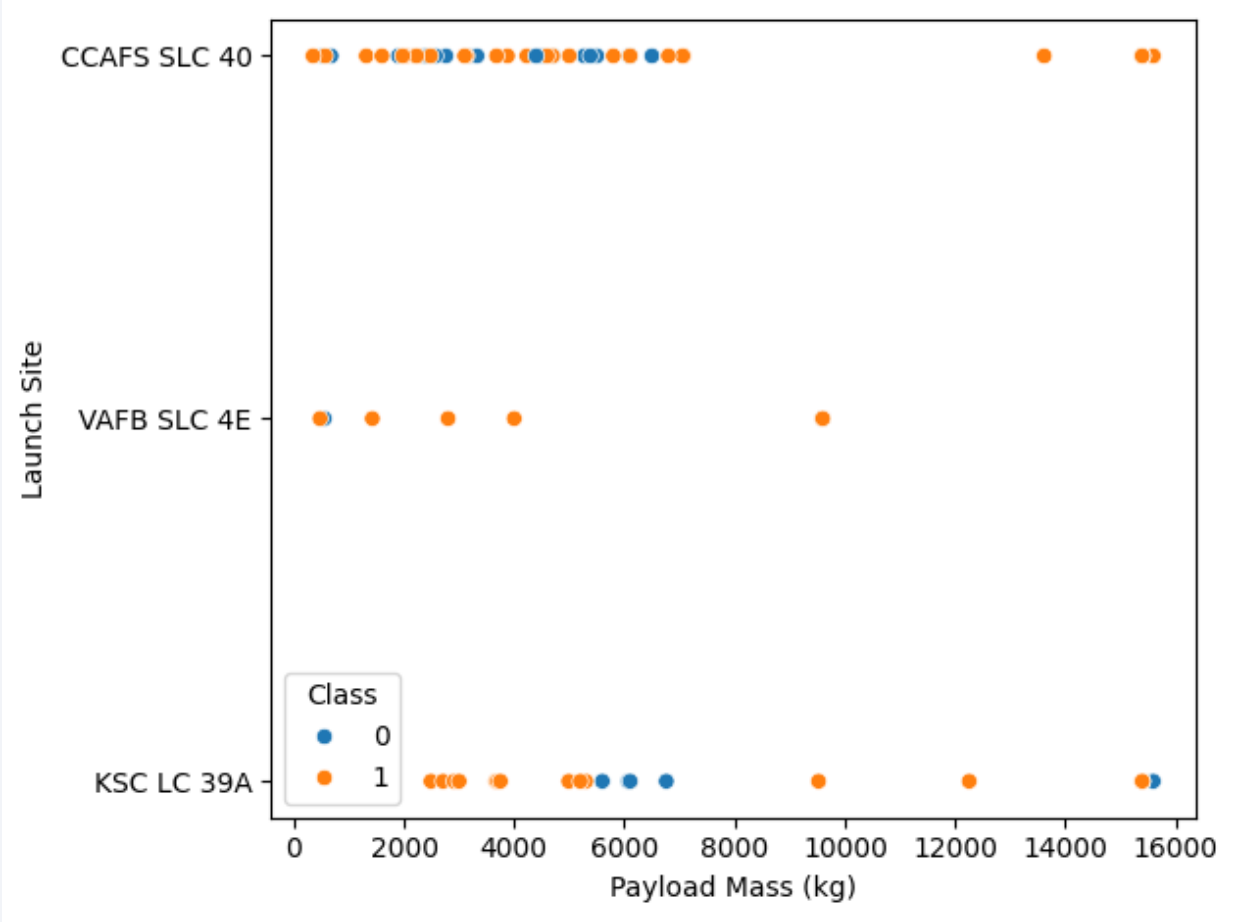
Flight Number vs. Launch Site

- CCAFS SLC 40 has the highest concentration of launches across the full flight number range, showing consistent launch activity.
- KSC LC 39A launches are grouped at higher flight numbers, indicating more recent usage for this site.
- VAFB SLC 4E has fewer launches overall, with more spread-out flight numbers.
- Across all sites, the proportion of successful landings (Class 1, orange) increases as flight numbers go up, suggesting learning and process improvement over time.



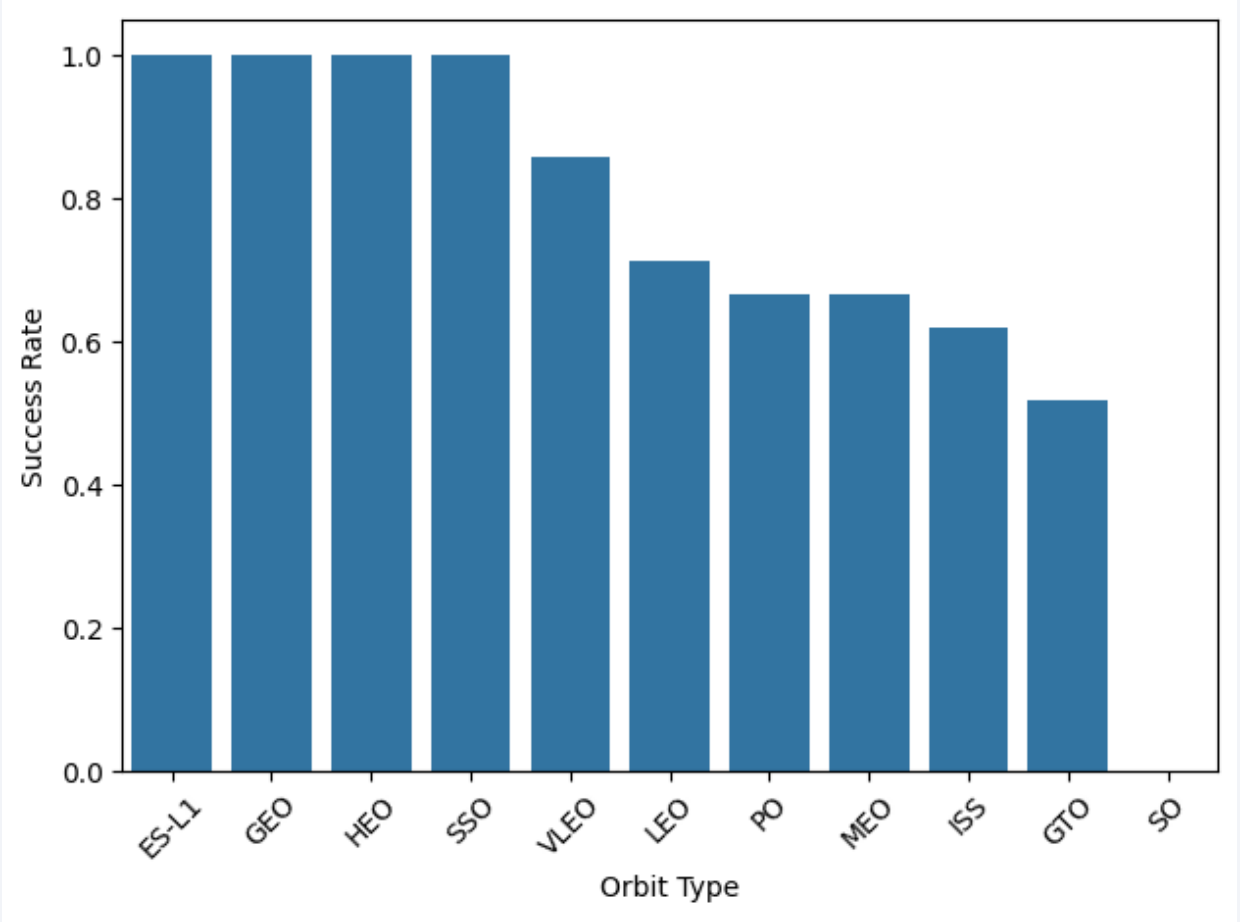
Payload vs. Launch Site

- Most launches from CCAFS SLC 40 occur with payloads less than 6,000 kg, with several successful landings (Class 1) at these lower payloads.
- KSC LC 39A features both low and high payload launches, but successful landings are present across all payload mass ranges.
- VAFB SLC 4E launches are fewer but skew toward lower payloads, with all showing successful landings.
- Very heavy payload launches (10,000–16,000 kg) are rare but mostly successful, especially from CCAFS SLC 40 and KSC LC 39A.



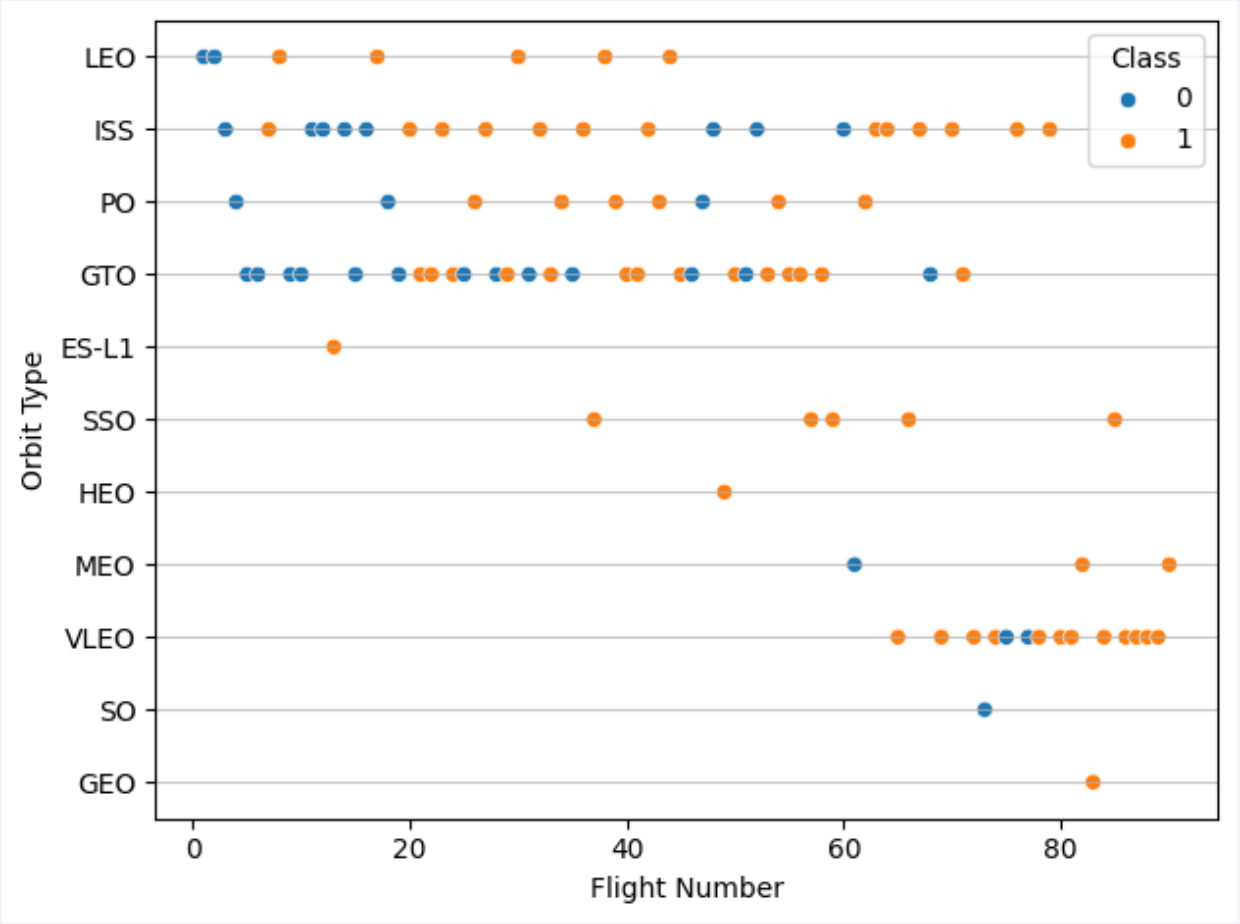
Success Rate vs. Orbit Type

- Certain orbit types like ES-L1, GEO, HEO, and SSO have a perfect (100%) landing success rate.
- Common orbit types like LEO and PO have lower success rates, possibly due to more challenging mission profiles or higher frequency.
- GTO (Geostationary Transfer Orbit) has the lowest observed landing success rate among the listed orbits.
- The drop-in success rate for some orbits suggests specific technical or operational challenges associated with those missions.



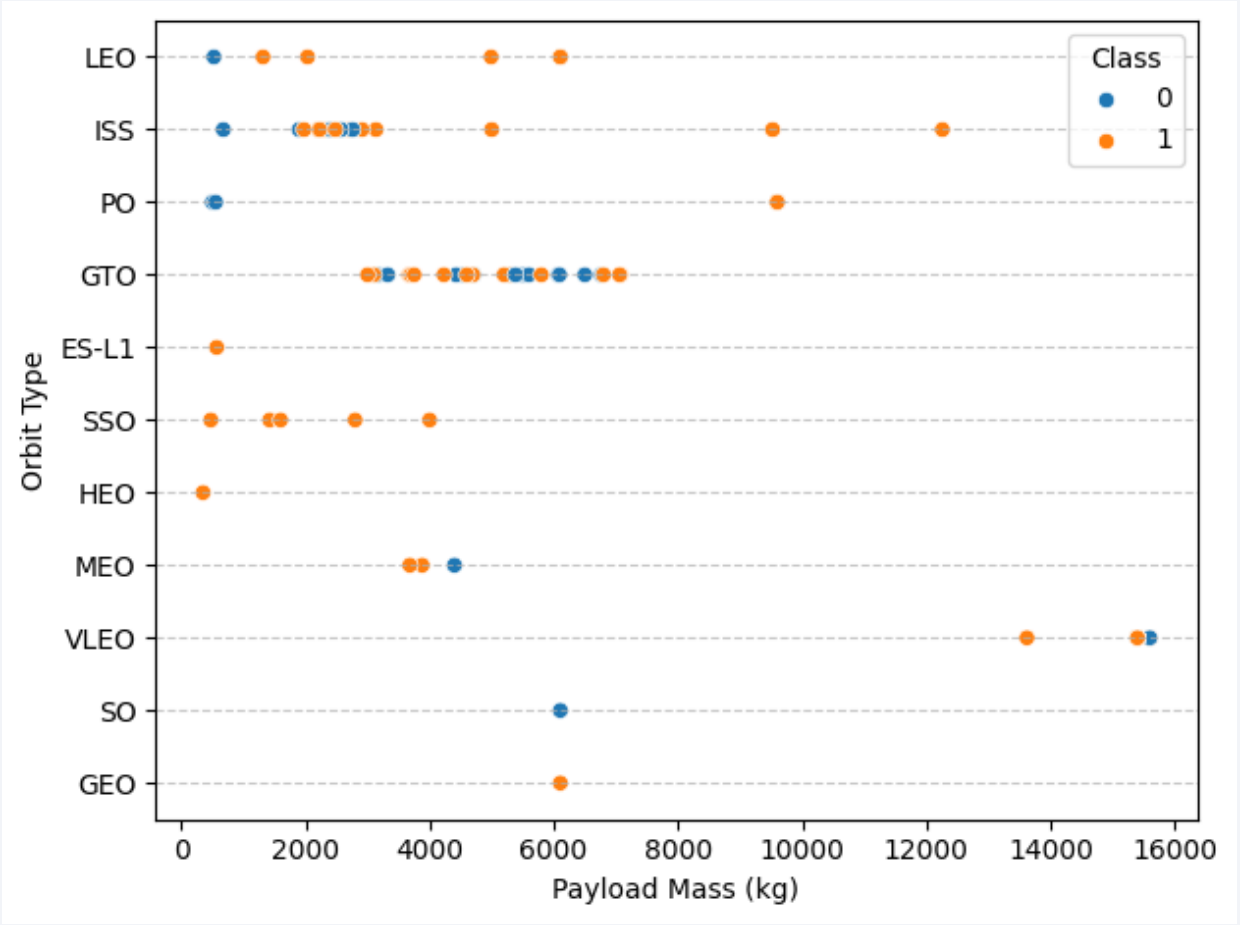
Flight Number vs. Orbit Type

- Most VLEO and MEO missions occur at higher flight numbers, aligning with more recent mission types.
- GTO and LEO orbits are distributed throughout the flight history, but success rates (orange) increase in later flights.
- ISS and PO launches are consistent across flight numbers, but with varying landing success.
- Missions to rare orbits (ES-L1, GEO, HEO, SSO) appear at select flight numbers and all end in successful landings.



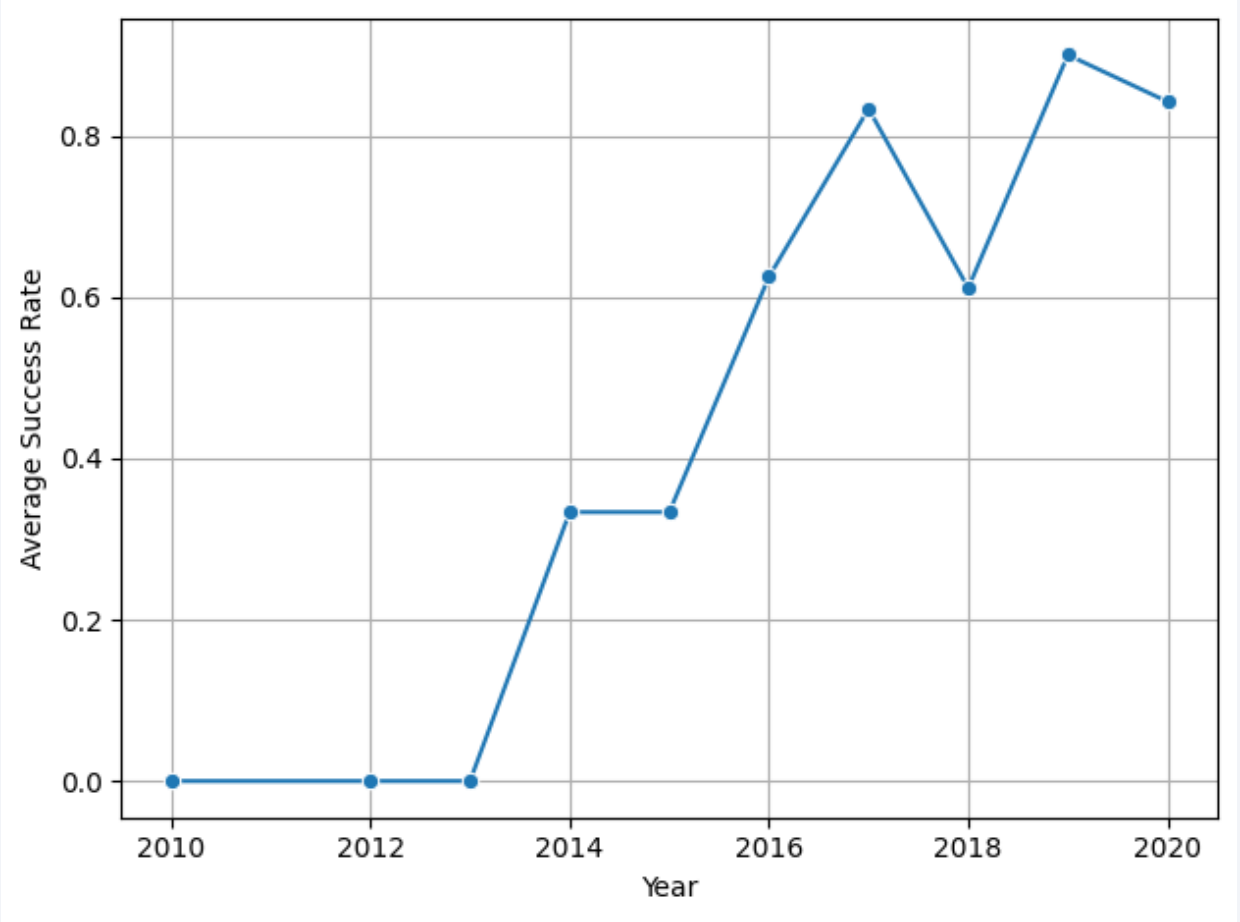
Payload vs. Orbit Type

- ISS missions typically cluster around 2,000–4,000 kg payloads, with most resulting in successful landings.
- GTO missions center on 5,000–6,000 kg payloads but show a mix of success and failure, reflecting the challenging nature of these missions.
- High-payload missions (10,000+ kg) to various orbits are rare but mostly successful, possibly indicating refined technology for such launches.
- Some orbits, like VLEO and SSO, display high landing success even at varying payload masses.



Launch Success Yearly Trend

- There is a clear upward trend in average success rate over the years, especially after 2014.
- Early years (2010–2013) saw little to no successful landings, reflecting the experimental phase.
- Rapid improvement is evident from 2015 onward, peaking above 80% average success in the latest years.
- The trendline demonstrates SpaceX's increasing mastery in first-stage landings, likely due to technological advancements and operational experience.



All Launch Site Names

- The dataset includes three main launch sites: CCAFS LC-40, VAFB SLC-4E, and KSC LC-39A.
- CCAFS LC-40 appears twice, indicating possible data entry or label inconsistency.
- The diversity of launch sites allows SpaceX to serve a wide range of orbits and customer needs.
- Site usage frequency can be analyzed further to assess launch patterns or site reliability.

Display the names of the unique launch sites in the space mission

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

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

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

Launch Site Names Begin with 'CCA'

- All five displayed records originate from CCAFS LC-40, showing its heavy utilization, especially in early Falcon 9 launches.
- Early missions (2010–2013) had payload masses of zero or were demo flights, highlighting the test phase.
- NASA and SpaceX were the primary customers for these launches, focusing on LEO and ISS missions.
- All missions shown resulted in "Success," indicating a solid track record from this site in the displayed period.

Display 5 records where launch sites begin with the string 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- NASA (CRS) missions have launched a combined payload mass of 45,596 kg via SpaceX boosters.
- This showcases SpaceX's significant partnership with NASA for cargo resupply services.
- High aggregate payload demonstrates the operational reliability and frequency of these launches.
- The result could be used for reporting to stakeholders on SpaceX's delivery volume for major clients.

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

In [12]:

```
%sql SELECT SUM("Payload_Mass__kg_") AS TOTAL_PAYLOAD
```

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

Out[12]:

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by the F9 v1.1 booster version is 2,928.4 kg.
- This statistic gives a benchmark for mission planning with this booster type.
- It helps compare payload capability trends between different Falcon 9 versions.
- Useful for tracking booster upgrades and performance improvements over time.

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

In [13]:

```
%sql SELECT AVG("Payload_Mass__kg_") AS Avg_Payload_M
```

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

Out[13]:

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

- The first successful ground landing was achieved on 2015-12-22.
- This marks a historic milestone in SpaceX's reusability journey.
- The date serves as a reference for improvement in landing technology.
- Landing recoveries after this point became more frequent, boosting SpaceX's cost efficiency.

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

Hint:Use min function

In [15]:

```
%sql SELECT MIN(Date) AS First_Succesful_landing FROM
```

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

Out[15]:

First_Succesful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four unique booster versions (F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2) achieved successful drone ship landings within the 4,000–6,000 kg payload range.
- These boosters demonstrate operational success for mid-heavy payload missions.
- Success in this mass range highlights technical capability in challenging recovery scenarios.
- Insights help SpaceX market these boosters' reliability for similar future missions.

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

In [16]:

```
%sql SELECT "Booster_version" FROM SPACEXTBL WHERE "Landing_(
```

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

Out[16]:

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

Total Number of Successful and Failure Mission Outcomes

- The vast majority of recorded missions resulted in "Success" (98), with only a few failures or ambiguous outcomes.
- There is only one "Failure (in flight)," indicating a strong overall reliability.
- "Success (payload status unclear)" indicates occasional data or mission reporting ambiguities.
- High success count is a key selling point for SpaceX's commercial viability.

List the total number of successful and failure mission outcomes

```
In [17]:
%sql SELECT "Mission_Outcome", COUNT(*) AS TOTAL FROM SPACEXTE
 * sqlite:///my_data1.db
Done.
Out[17]:
```

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

Boosters Carried Maximum Payload

- Multiple booster versions have carried the maximum payload of 15,600 kg, all using the Falcon 9 Block 5 variant.
- This consistency across different Block 5 boosters shows design robustness and repeatability at maximum payload capacity.
- Indicates Block 5's role as the workhorse for heavy-lift missions.
- Demonstrates that heavy-lift capability is not limited to a single "lucky" booster, but is standard across the fleet.

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

```
Payload_Mass__kg_" FROM SPACEXTBL WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTBL);
```

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

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

- In 2015, two failures on drone ship landings were recorded: one in January (B1012) and one in April (B1015), both from CCAFS LC-40.
- This pinpoints a specific period of operational challenge in drone ship landings.
- Both failures involved the F9 v1.1 booster version, indicating possible technical or weather-related hurdles at the time.
- Temporal analysis like this can inform improvements for future missions.

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.

In [19]:

```
%sql SELECT SUBSTR(Date, 6, 2) AS Month, "Landing_Outcome", "Booster_V
```

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

Out[19]:

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

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

- "No attempt" was the most frequent outcome (10), reflecting either test flights or missions without recovery goals in the early period.
- Successes and failures on drone ships are evenly split (5 each), with ground pad successes (3) and ocean recoveries (controlled/uncontrolled) making up the rest.
- Parachute and precluded landings were rare.
- The data reflects SpaceX's trial-and-error evolution from expendable boosters to regular, successful landings.

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.

```
In [20]:
%sql SELECT "Landing_Outcome", Count(*) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '
 * sqlite:///my_data1.db
Done.
Out[20]:
```

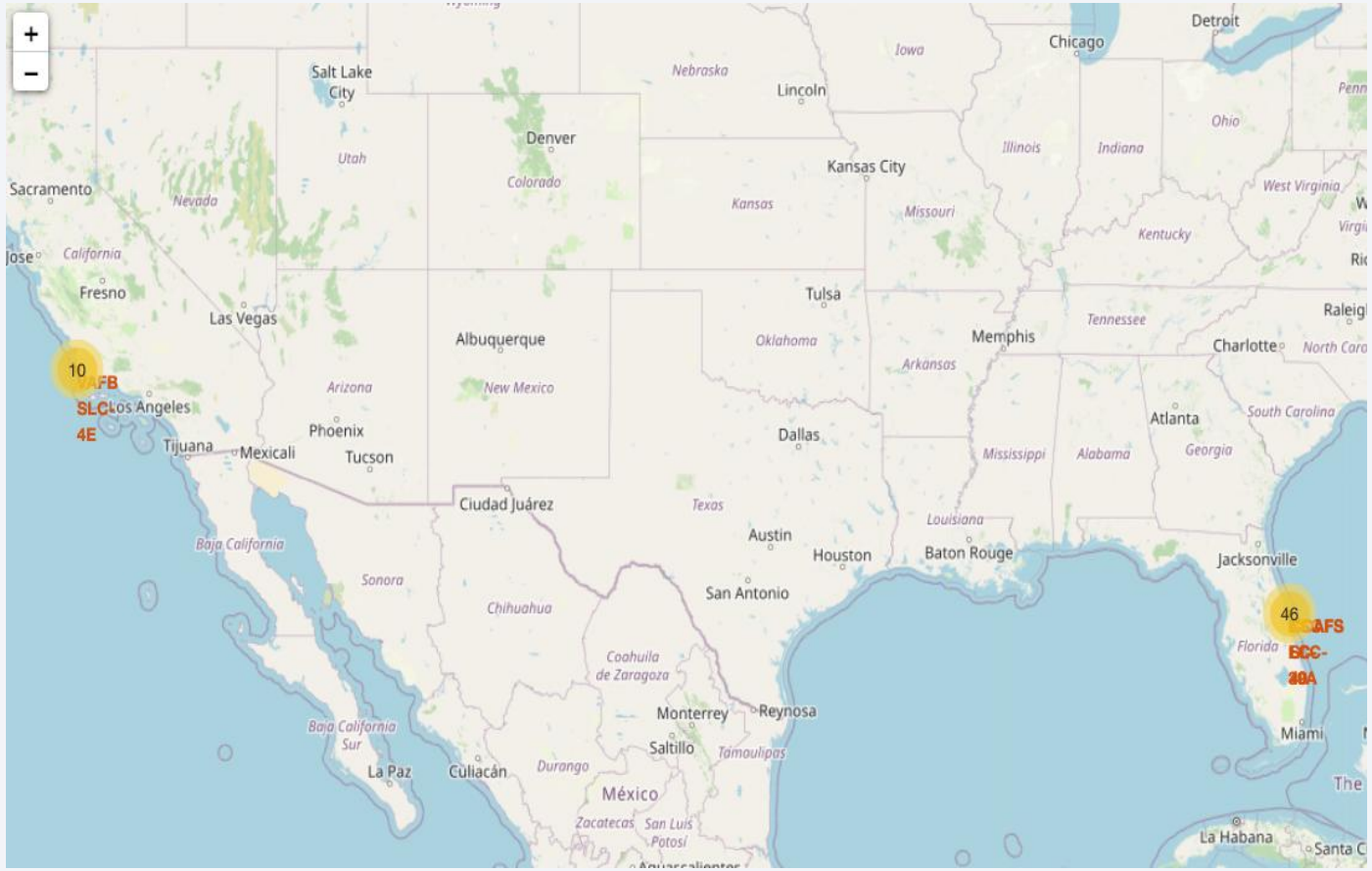
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

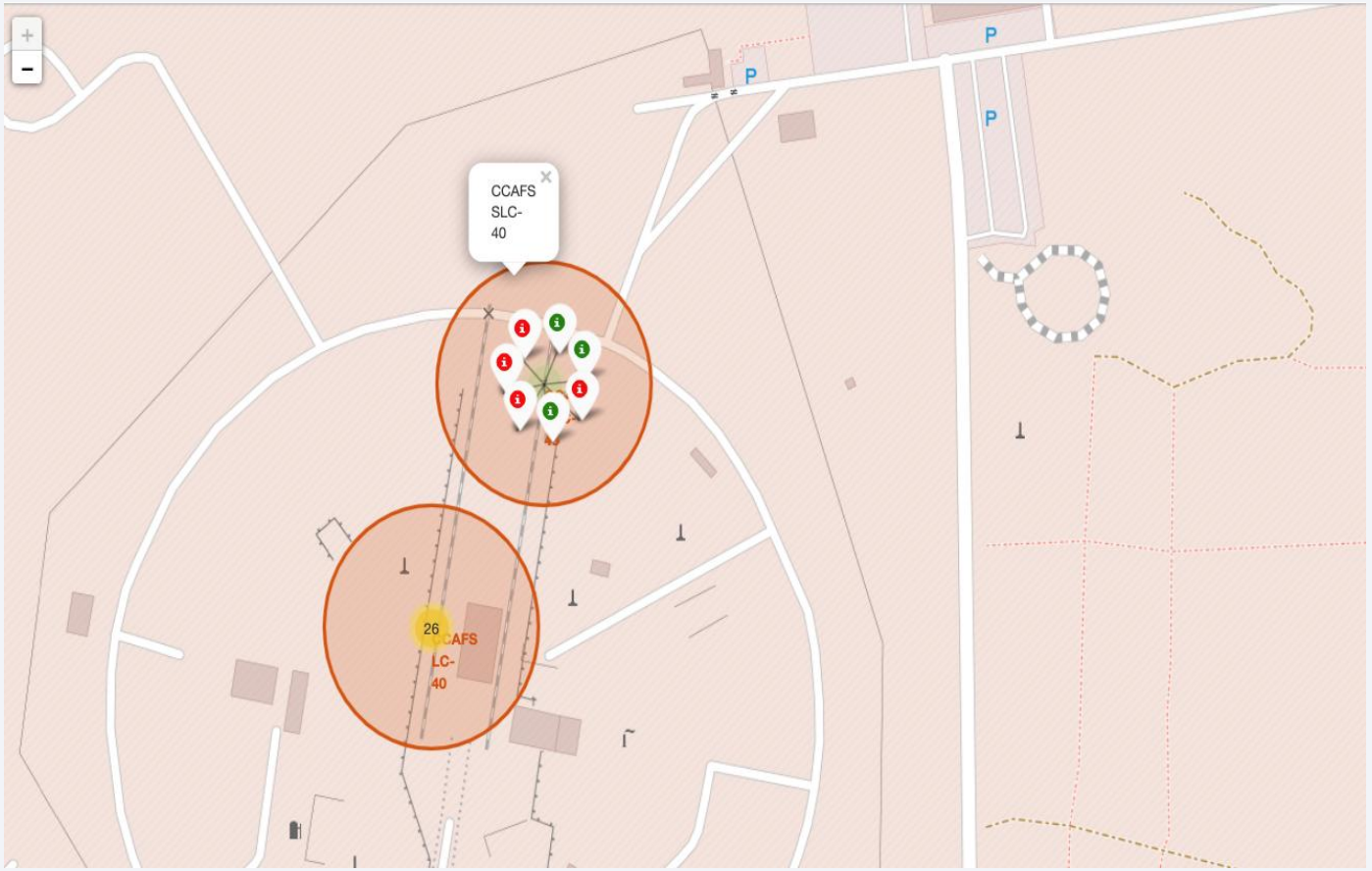
SpaceX Launch Sites Mapped Across the United States

- CCAFS LC-40 (Florida) is the most frequently used site, followed by VAFB SLC-4E (California) and KSC LC-39A (Florida).
- Launch activity is heavily concentrated on the East Coast (Florida), making it SpaceX's operational hotspot.
- VAFB SLC-4E on the West Coast handles fewer launches, likely for specific polar or sun-synchronous missions.
- Strategic site distribution allows SpaceX to serve orbital inclinations and customer needs.



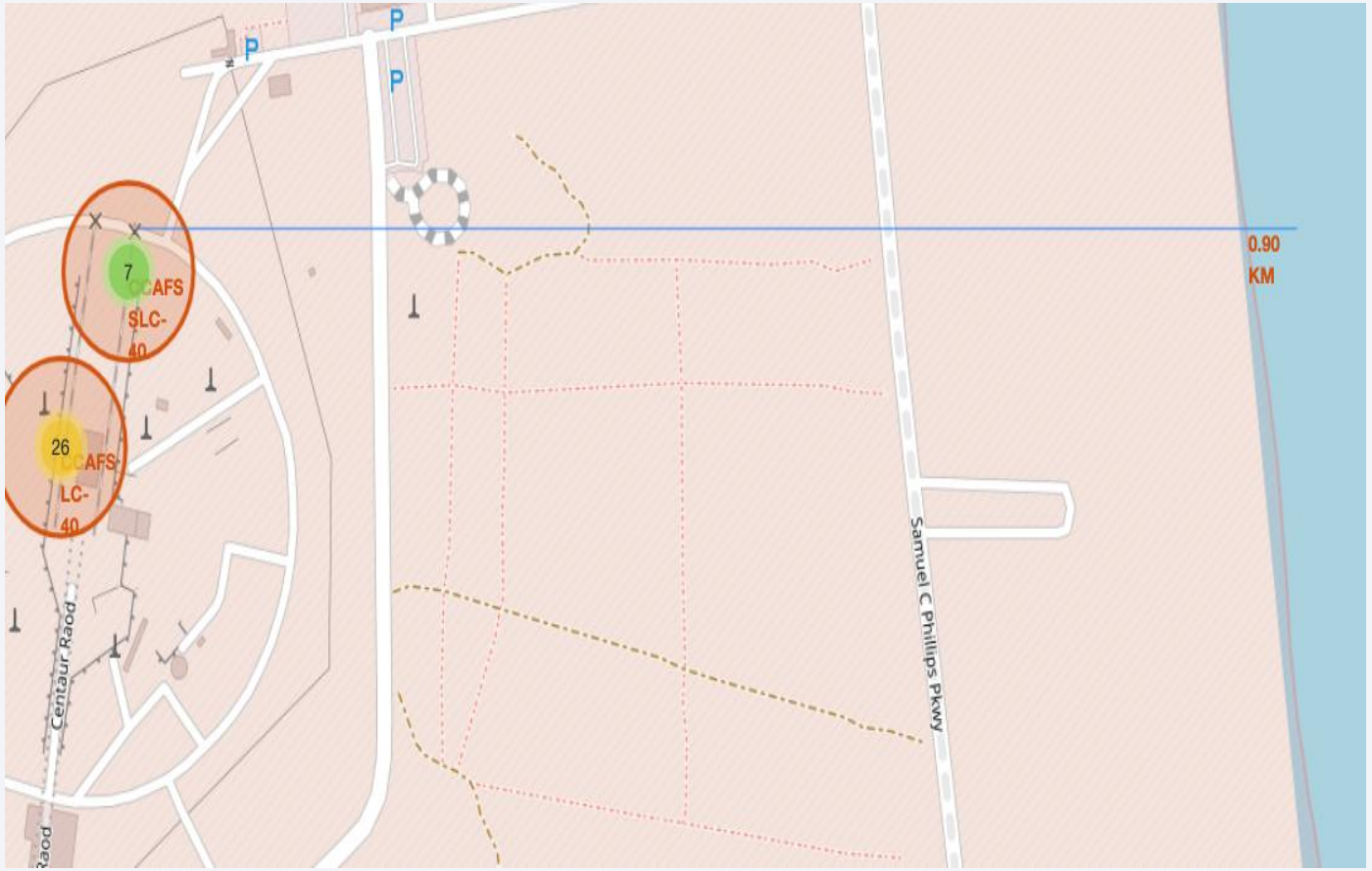
Micro-Cluster Analysis of Launch Outcomes at CCAFS SLC-40

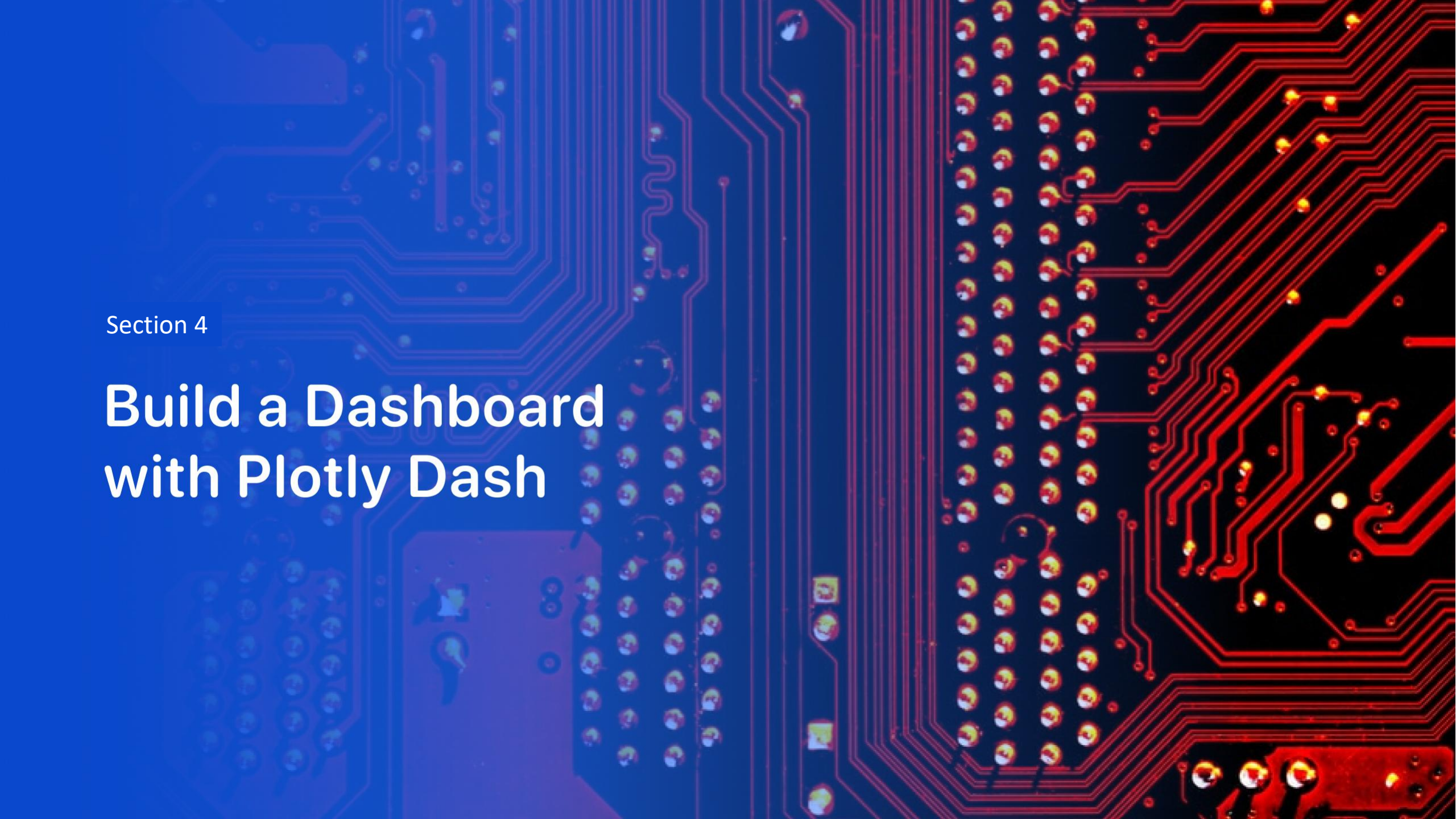
- Multiple landing attempts (success and failure) are clustered at the same pad, highlighting iterative testing at CCAFS LC-40.
- Green markers indicate successful landings, while red show failed attempts, illustrating gradual improvement.
- The pad serves as both a launch and landing site, maximizing facility utility.
- Spatial overlap shows landing risk is managed within a compact zone.



Proximity of CCAFS SLC-40 Launch Site to Coastline

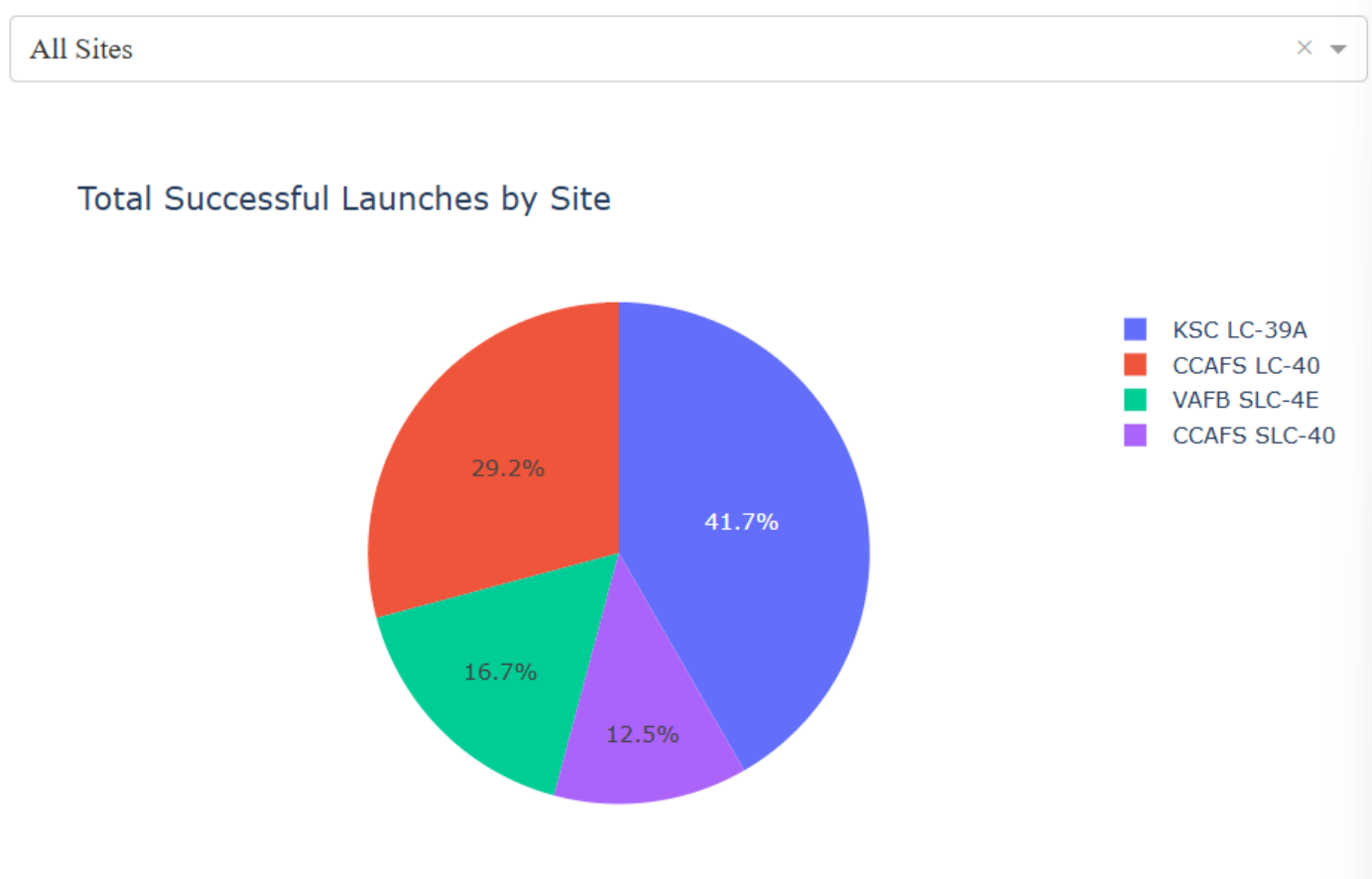
- The launch pad is only 0.9 km from the coast, which minimizes overland risk and allows direct launch trajectories over the ocean.
- Proximity to water is optimal for both safety and regulatory compliance.
- The close distance also supports drone ship operations for offshore booster landings.
- The map underlines how physical geography is critical in launch site selection.




Section 4
Build a Dashboard with Plotly Dash

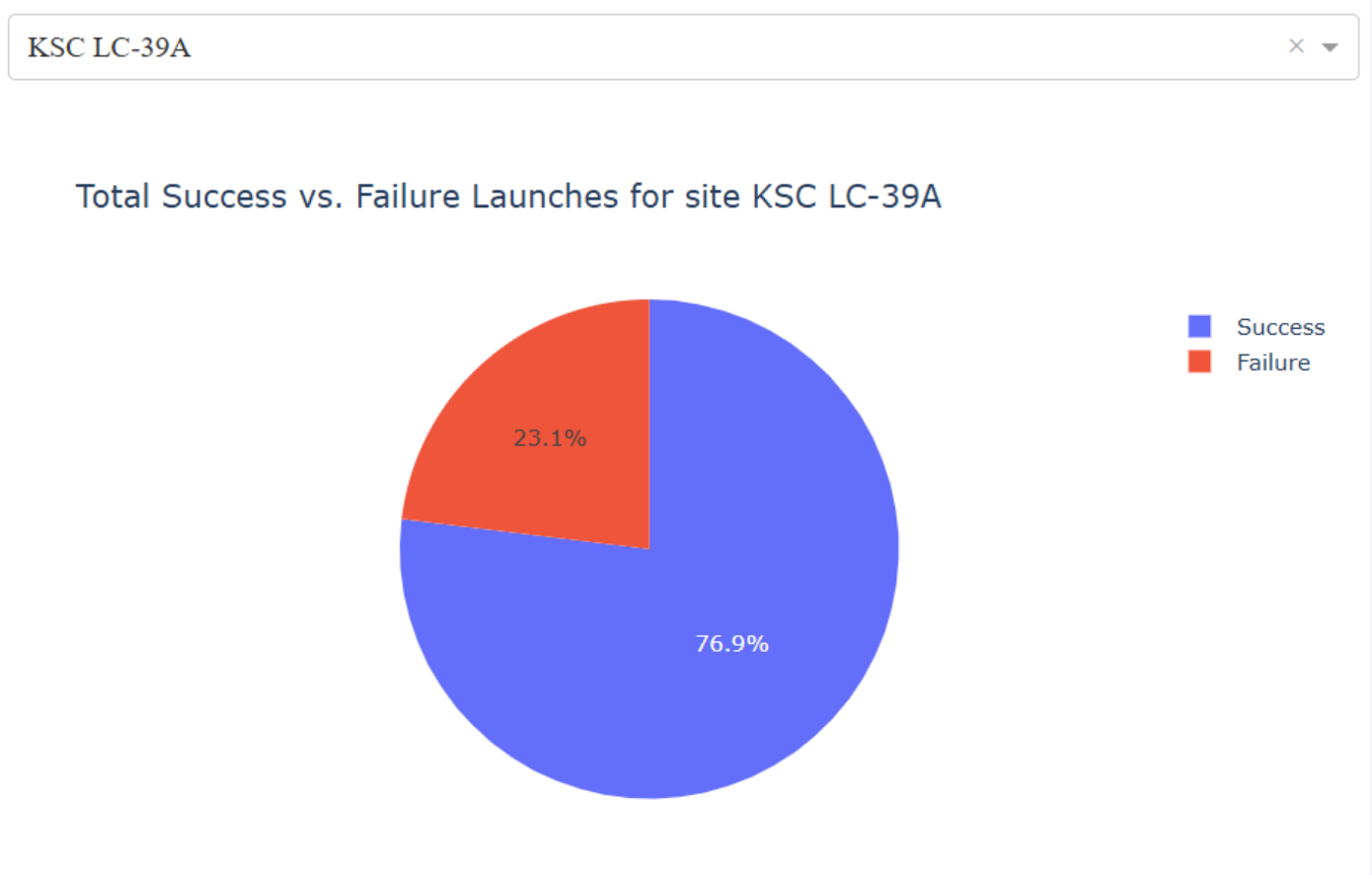
Distribution of Successful Launches by Site

- KSC LC-39A accounts for the highest share of successful launches (41.7%), dominating other sites.
- CCAFS LC-40 is next (29.2%), followed by VAFB SLC-4E (16.7%) and a minor share for CCAFS SLC-40 (12.5%).
- The chart exposes a possible data labeling or duplication issue with “CCAFS LC-40” vs. “CCAFS SLC-40.”
- Success distribution highlights the reliability and volume of Florida sites.



Success vs. Failure at KSC LC-39A

- KSC LC-39A has a high launch success rate: 76.9% successful vs. 23.1% failures.
- The majority of launches from this site are reliable, confirming its role as a prime pad for high-profile missions.
- Failure rate is notable but does not undermine overall reliability.
- This breakdown can guide risk assessment for future missions launched from LC-39A.



Correlation between Payload and Landing Success

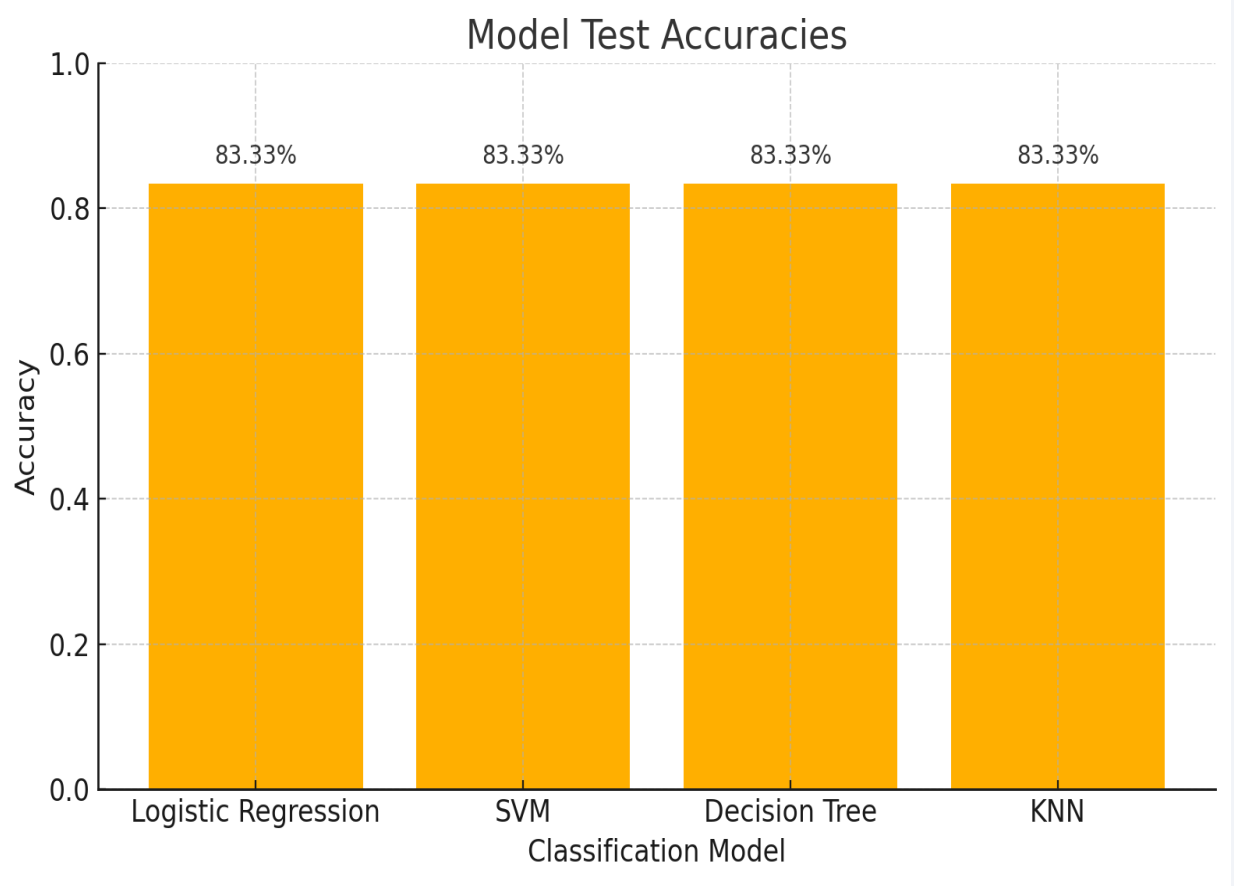
- The scatter plot shows the relationship between payload mass (kg) and launch outcome (1 = Success, 0 = Failure) for different booster versions.
- Most successful launches (class = 1) occur in the lower to mid payload range (up to ~6000 kg), with FT boosters showing the highest concentration of successes.
- Failures (class = 0) are more common with earlier booster versions (v1.0, v1.1) and tend to cluster at lower payloads.

Section 5

Predictive Analysis (Classification)

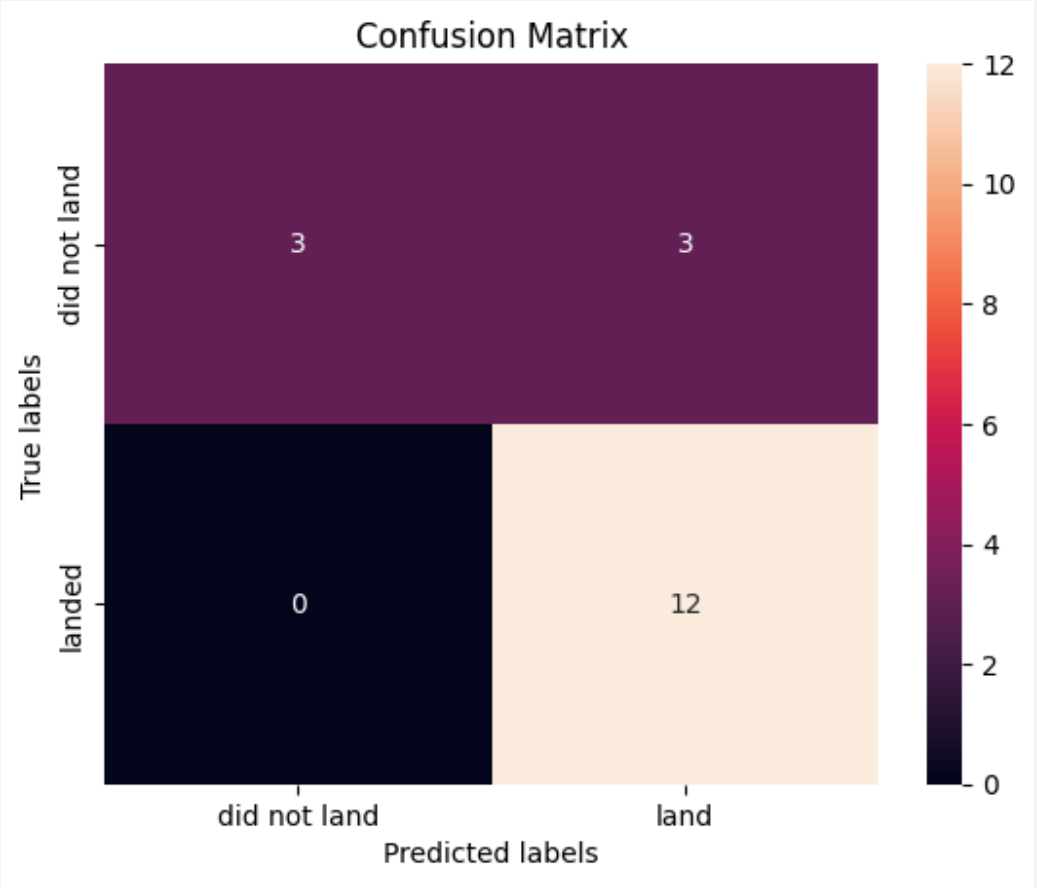
Classification Accuracy

- Consistent 83.33% across all algorithms
  - All models—Logistic Regression, SVM, Decision Tree, KNN—achieve identical test accuracy.
- Algorithm choice has limited marginal gain
  - Uniform performance implies the need for feature enrichment or hyperparameter tuning.
- Baseline reliability for cost forecasting
  - An 83% accuracy rate gives a strong base for predicting landing outcomes and supporting cost analysis.



Confusion Matrix

- Zero false negatives
  - The model correctly identifies all actual landings (12/12), ensuring no successful landing is overlooked—vital for maximizing stage reuse and cost savings.
- Balanced specificity gap
  - With 3 true negatives vs. 3 false positives, specificity sits at 50%, indicating some over-optimism in predicting landings and a target area for reducing risk of underestimating failure.
- Recall prioritized over precision
  - Achieving 100% recall for successful landings aligns with business priorities—capturing every potential cost saving—even at the expense of a few false-success forecasts.



Conclusions

- Consistent Baseline Performance
  - All four models yielded 83.33% accuracy, indicating that core features (payload mass, orbit type, launch site) robustly capture landing drivers.
- Risk-Averse Error Profile
  - The confusion matrix shows 0 false negatives (100% recall for landed boosters) and 3 false positives, prioritizing recovery opportunities over minimizing redundant re-launch scenarios.
- Model Selection for Interpretability
  - Logistic Regression is recommended for its coefficient transparency—confirming positive payload mass and VLEO orbit effects—while matching the accuracy of more complex algorithms.
- Next-Steps for Optimization
  - Implement probability threshold tuning, feature interaction terms, and ensemble stacking to boost precision and further mitigate cost uncertainties in competitive bidding.

Appendix

All code, notebooks, and supporting materials for this project are available in this public GitHub repository: [github/jhermienpaul/ibm-data-science-program/Applied Data Science Capstone]

- Data Collection with API.ipynb
- Data Collection with Web Scraping.ipynb
- Data Wrangling.ipynb
- Exploratory Data Analysis with SQL.ipynb
- Exploratory Data Analysis with Visualization.ipynb
- Interactive Mapping with Folium.ipynb
- Interactive Dashboard with Plotly Dash.py
- Machine Learning Model Pipeline.ipynb

Thank you!