IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eslam Elmasry

January 7, 2026

Outline

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

Executive Summary

Summary of methodologies:

- Data Collection - spaceX API
- Data Collection - Web Scraping
- Data Wrangling
- EDA using SQL
- EDA with Data Viz
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Machine learning prediction

Summary of methodologies:

- Data Analysis Results
- Data Visuals
- Interactive Dashboard
- Predictive Model Analysis Results

Introduction

- **Objective**: In this capstone project, we will develop a model to predict whether the first stage of SpaceX's Falcon 9 rocket will successfully land. This prediction capability is valuable because SpaceX's ability to reuse the first stage enables them to offer launches at $62 million - significantly less than competitors who charge over $165 million per launch. If we can accurately forecast landing success, we can effectively determine the true cost of a launch, which provides crucial competitive intelligence for other companies considering bidding against SpaceX for rocket launch contracts.
- **Context:** It's important to note that not all unsuccessful landings represent failures. SpaceX sometimes intentionally performs controlled ocean landings as part of their operational planning. These planned ocean landings should be distinguished from actual landing failures in our analysis.
- **Core Predictive Question:** The central problem we're addressing is: Given a specific set of features characterizing a Falcon 9 rocket launch - including payload mass, orbit type, launch site, and other relevant parameters - can we accurately predict whether the rocket's first stage will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - SpaceX REST API
  - Web scrapping Falcon 9 and Falcon Heavy Launch records from Wikipedia
- **Perform data wrangling**
  - Filtering the Data
  - Dealing with the missing values
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - We created training labels for supervised machine learning by converting the mission outcomes into binary classification labels, where unsuccessful outcomes were encoded as 0 and successful outcomes were encoded as 1.

Data Collection

- **Data Collection Methodology**:
- We employed a dual approach to gather comprehensive launch data, utilizing both API requests to the SpaceX REST API and web scraping techniques to extract information from SpaceX's Wikipedia entry table. This combination was necessary to obtain a complete dataset for thorough analysis.
- **Data from SpaceX REST API**:The following data columns were acquired through the SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- **Data from Wikipedia Web Scraping**: The following data columns were extracted from the Wikipedia table through web scraping:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API

Data for this project was sourced from the SpaceX API at api.spacexdata.com/v4/rockets/. The dataset was filtered to include only Falcon 9 rocket launches. Missing values were imputed using the mean of their respective columns. The final processed dataset comprises 90 instances with 17 features each.

GitHub URL

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

90 rows × 17 columns

Data Collection - Scraping

To supplement the API data, a web scraper was used to extract Falcon 9 launch records from a specified Wikipedia revision. The resulting dataset comprises 121 observations (launches) with 11 associated variables per observation, as seen in the initial rows below.

GitHub URL

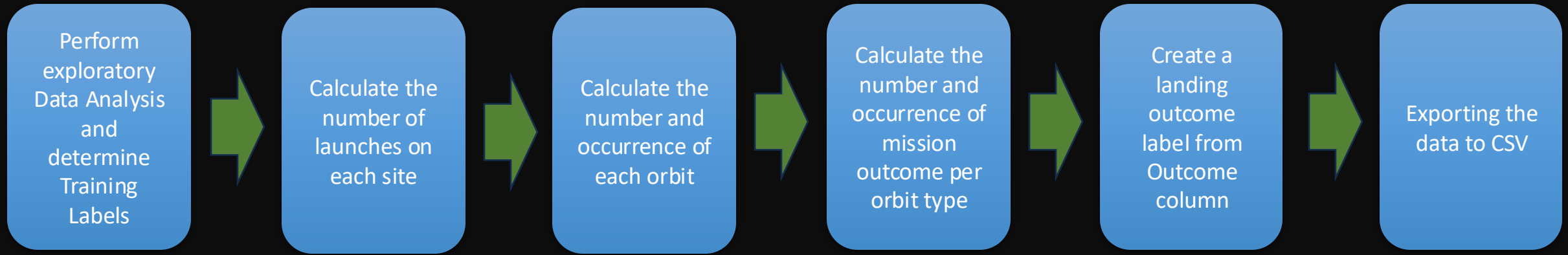
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10

Data Wrangling

The dataset contains multiple scenarios where booster landings were unsuccessful. These include situations where landing attempts failed due to accidents. Specifically:

- **True Ocean**: The mission outcome involved a successful landing in a designated ocean region
- **False Ocean**: The mission outcome involved an unsuccessful landing attempt in a designated ocean region
- **True RTLS**: The mission outcome involved a successful landing on a ground pad
- **False RTLS**: The mission outcome involved an unsuccessful landing attempt on a ground pad
- **True ASDS**: The mission outcome involved a successful landing on an autonomous drone ship
- **False ASDS**: The mission outcome involved an unsuccessful landing attempt on an autonomous drone ship

For supervised learning purposes, we primarily converted these mission outcomes into binary training labels, where "1" represents successful booster landings and "0" represents unsuccessful booster landings.



GitHub URL

EDA with Data Visualization

A series of visualizations were created to explore the dataset and identify relationships for predictive modeling:

- **Scatter Plots** (e.g., Flight Number vs. Payload Mass) were used to detect potential correlations between continuous or ordinal variables that could serve as features for a machine learning model.
- **Bar Charts** (e.g., Orbit Type vs. Success Rate) were employed to compare success metrics across discrete categories like launch sites and orbit types.
- **A Line Chart** illustrated the yearly trend in launch success over time, highlighting temporal patterns in the data.

GitHub URL

EDA with SQL

The following specific SQL queries were performed on the mission database:

- Identified all distinct launch site names.
- Retrieved 5 records where the launch site name starts with 'CCA'.
- Calculated the total payload mass delivered by boosters launched for NASA's Commercial Resupply Services (CRS).
- Determined the average payload mass carried by the booster version `F9 v1.1`.
- Found the date of the first successful booster landing on a ground pad.
- Listed booster names that achieved a successful drone ship landing with a payload mass between 4,000 and 6,000 units.
- Counted the total number of mission outcomes, categorized as 'Success' and 'Failure'.
- Identified the booster version(s) that carried the maximum recorded payload mass.
- Retrieved details (outcome, booster version, launch site) for failed drone ship landing attempts that occurred in the year 2015.
- Ranked the frequency of each distinct landing outcome (e.g., "Failure (drone ship)") between June 4, 2010, and March 20, 2017, in descending order.

GitHub URL

Build an Interactive Map with Folium

The geographical data was visualized using an interactive mapping library. The implementation involved three main stages:

- **Geospatial Plotting:** All launch site latitudes and longitudes were used to place map markers with interactive circles and labels, initiating the view at NASA Johnson Space Center.
- **Outcome Clustering:** A MarkerCluster layer was implemented to group individual launch outcome markers by location. Successes and failures were distinguished using green and red markers, respectively, creating a density-based view of success rates.
- **Proximity Mapping:** As a spatial analysis example, direct-line distances were rendered on the map from launch site KSC LC-39A to its nearest significant features (railway, highway, coastline, and city), with each connection line colored and labeled for clarity.

GitHub URL

Build a Dashboard with Plotly Dash

An interactive dashboard was implemented to facilitate dynamic data exploration. Key components include:

- **Launch Site Selector:** A dropdown menu allowing users to filter data by individual launch sites or view an aggregated "All Sites" summary.
- **Success Rate Visualization:** A pie chart that dynamically updates to display either the overall count of successful launches across all sites or a detailed breakdown of success versus failure counts for a selected site.
- **Payload Filter:** An adjustable slider enabling interactive selection of a payload mass range.
- **Success-Payload Correlation:** A scatter chart illustrating the relationship between payload mass and launch success, with data points segmented by booster version to analyze potential correlations across rocket models.

Predictive Analysis (Classification)

The machine learning pipeline was implemented using the Scikit-learn library. The workflow consisted of the following key phases:

- **Data Preprocessing:** Features were standardized to ensure consistent scaling.
- **Data Partitioning:** The dataset was split into training and testing subsets.
- **Model Development:** Four distinct classification algorithms were instantiated: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- **Model Training & Optimization:** Each model was fitted on the training data, followed by hyperparameter tuning to identify the optimal configuration for each algorithm.
- **Model Evaluation:** Final performance was assessed using accuracy scores and confusion matrices on the held-out test set.

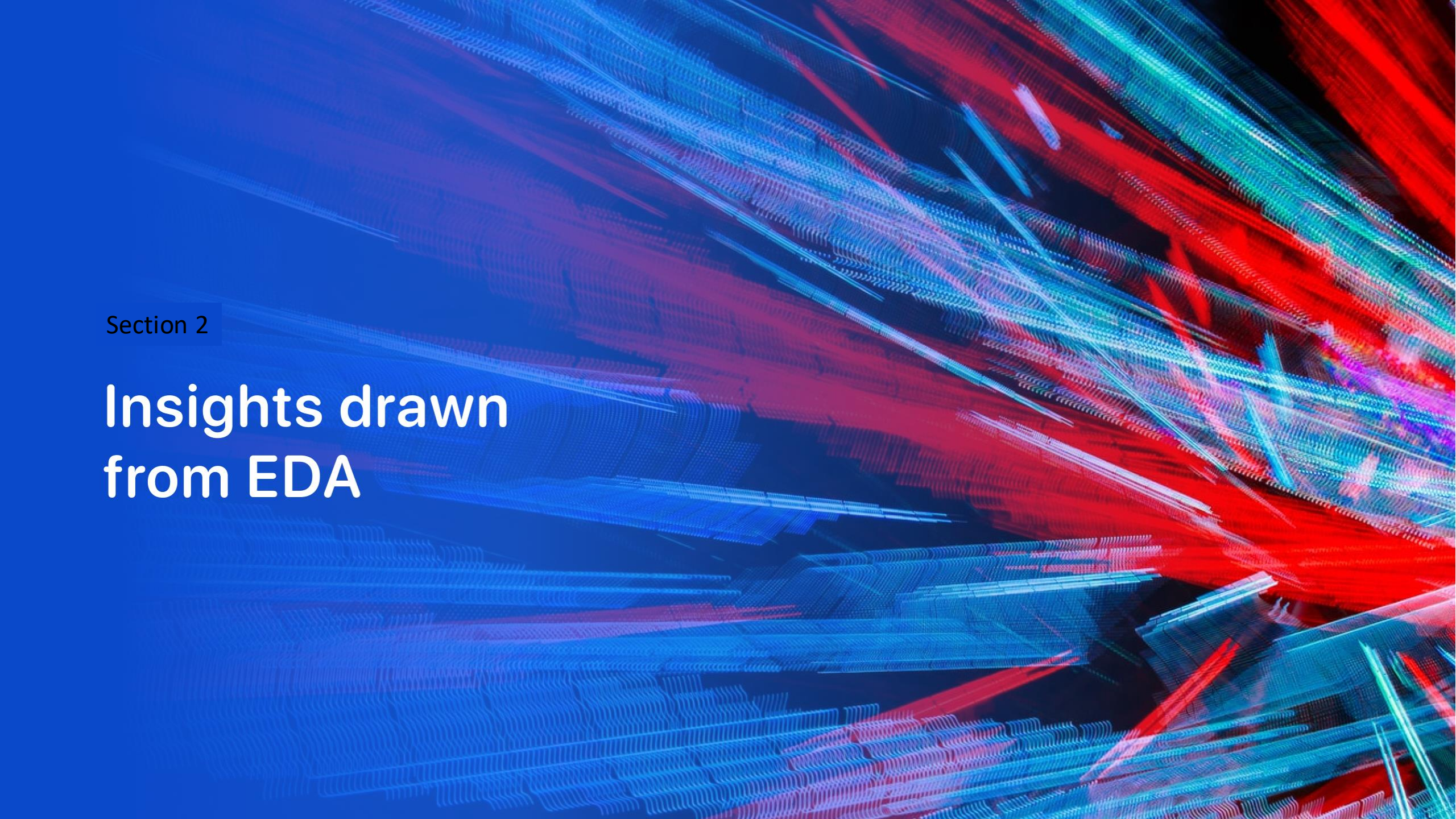
GitHub URL

Results

- The results are presented in the following five sections, outlining the sequential workflow:
- **SQL:** Initial exploratory data analysis (EDA) performed via database queries.
- **Matplotlib & Seaborn:** Advanced EDA and static visualizations.
- **Folium:** Creation of interactive geographical maps.
- **Dash:** Development of a comprehensive, interactive web application dashboard.
- **Predictive Analysis:** Application and evaluation of machine learning models.

Section 2

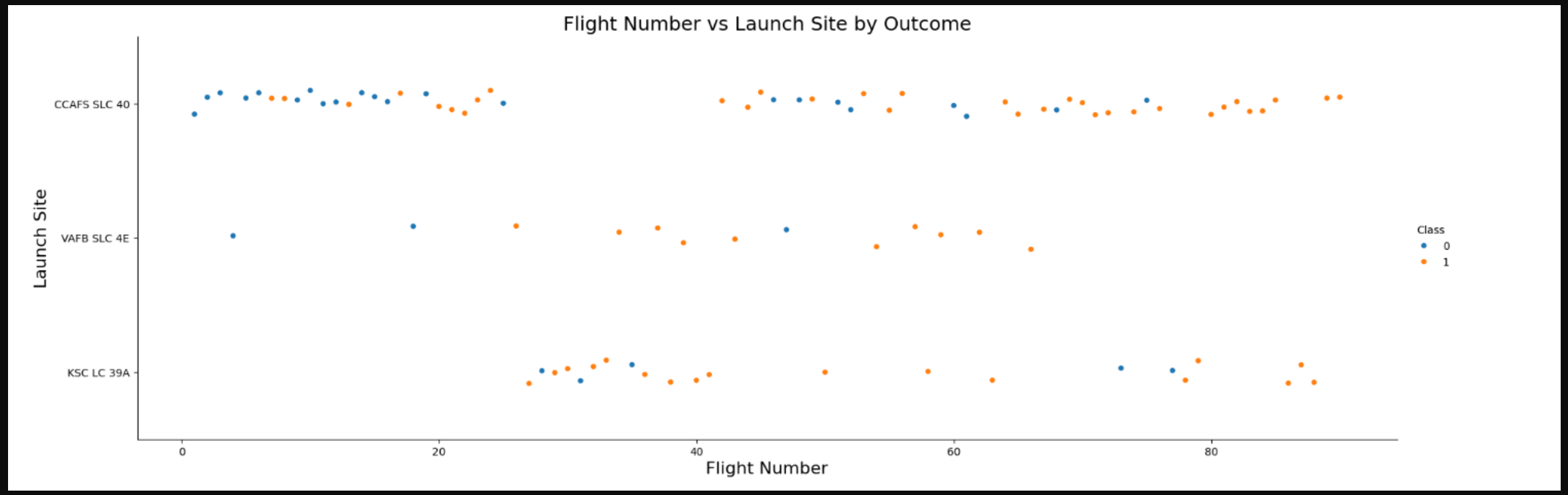
Insights drawn from EDA



Flight Number vs. Launch Site

Site Performance: Launch pads such as CCAFS SLC-40 and KSC LC-39A show no inherent advantage, with both successful and unsuccessful landings, pointing to other critical factors.

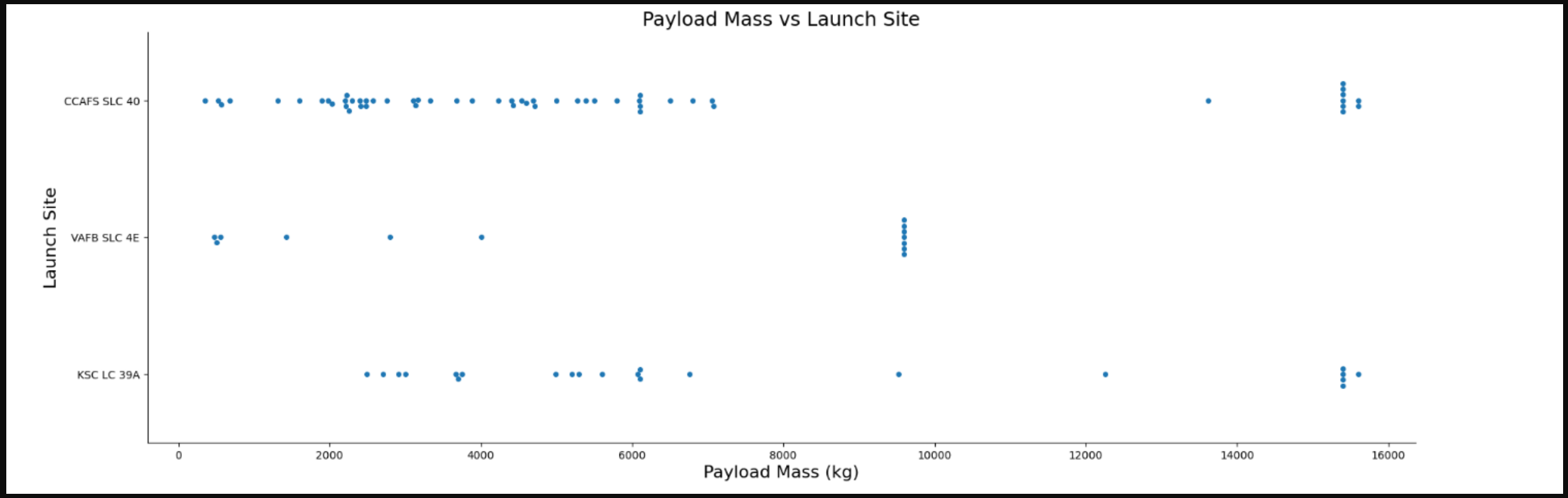
Temporal Pattern: Activity is distributed evenly across sequential missions, indicating landing success rates have remained consistent over time without a marked upward or downward trend.



Payload vs. Launch Site

Payload Specialization by Site: CCAFS SLC-40 is predominantly used for lighter payloads (under 10,000 kg), whereas VAFB SLC-4E and KSC LC-39A accommodate a more diverse mass range, reflecting varied mission types.

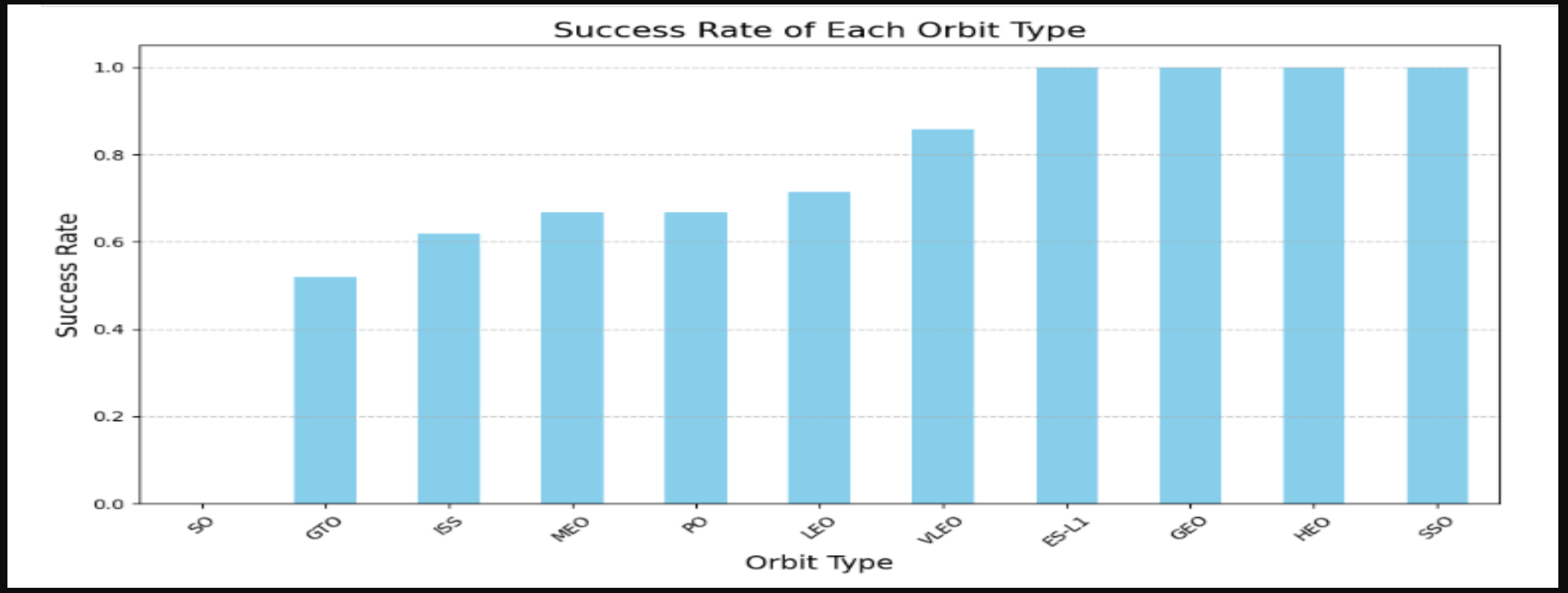
Heavy-Lift Role: KSC LC-39A stands out as the leading site for heavy-lift missions, with a high frequency of launches carrying payloads over 15,000 kg.



Success Rate vs. Orbit Type

High-Reliability Orbits: Missions to VLEO, ES-L1, GEO, HEO, and SSO have demonstrated perfect landing success, indicating these trajectories are highly reliable for first-stage recovery.

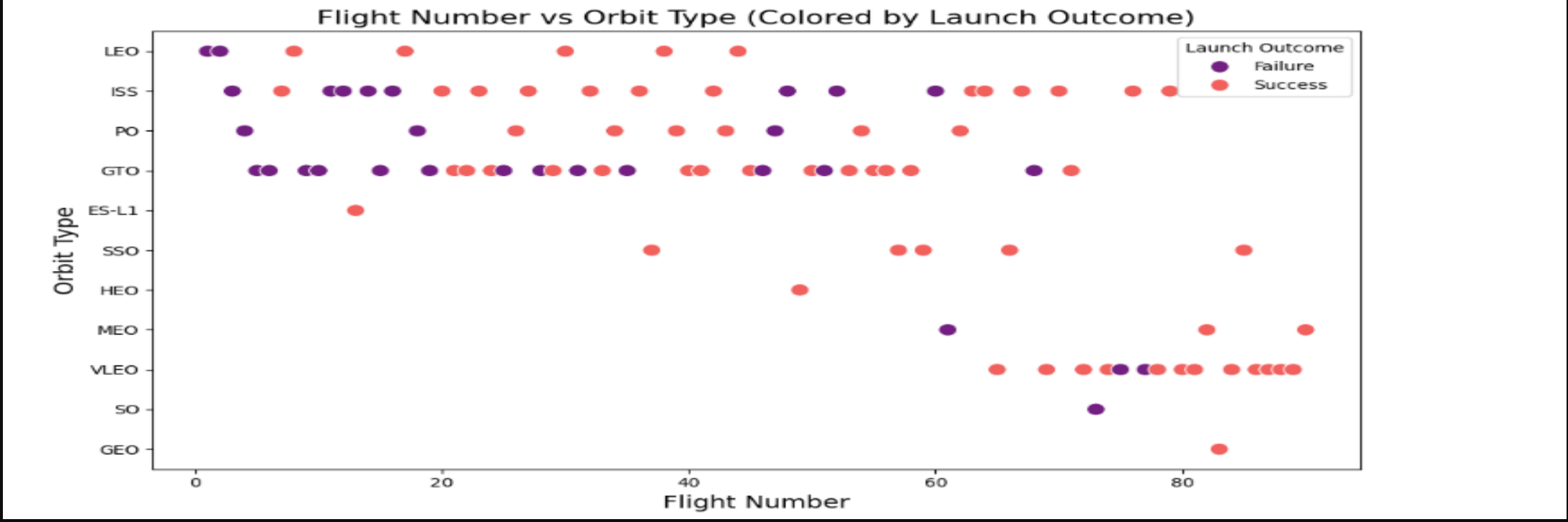
Notable Challenge: In contrast, GTO missions exhibit a substantially lower success rate, suggesting the unique demands of this orbit profile present greater complexity or difficulty for successful landings.



Flight Number vs. Orbit Type

Trend of Improvement: Falcon 9's launch success rate improves significantly with higher flight numbers, highlighting the impact of operational experience and iterative design enhancements.

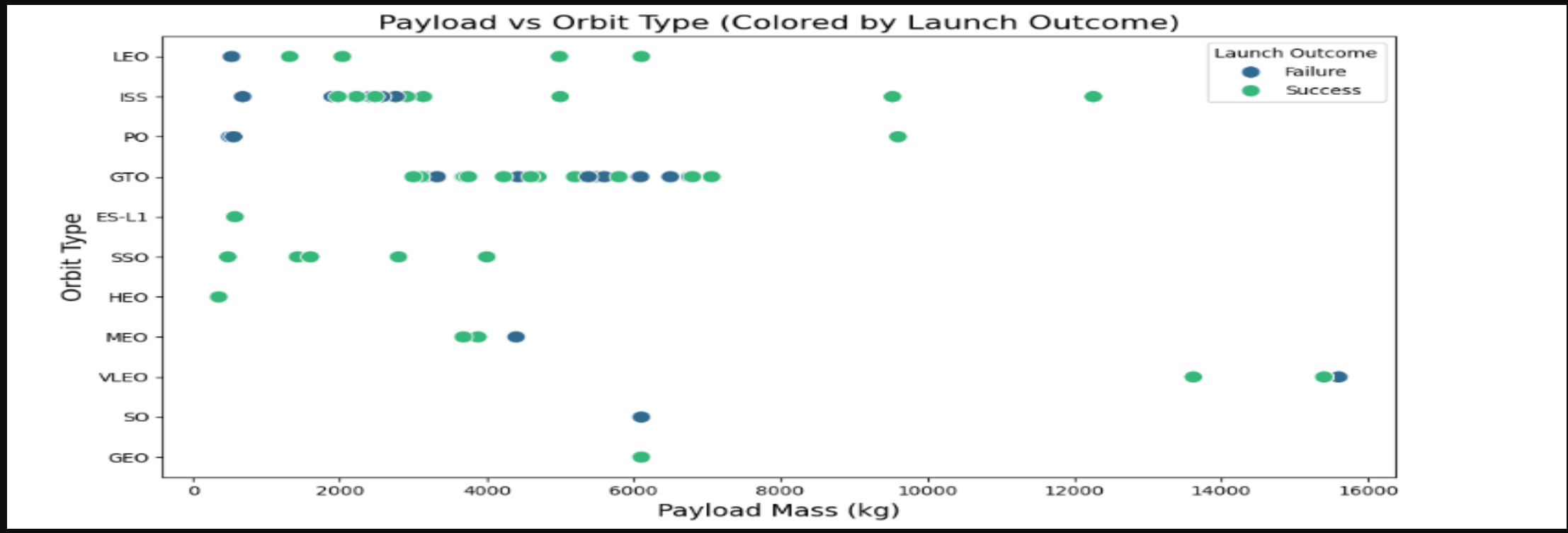
Overcoming Specific Challenges: Orbits such as GTO and ISS, which saw mixed results early on, now achieve a higher success rate, reflecting refined mission planning and execution for these complex profiles.



Payload vs. Orbit Type

Light Payloads (High Success): Missions carrying under 6,000 kg achieve reliable landing success across all orbit types.

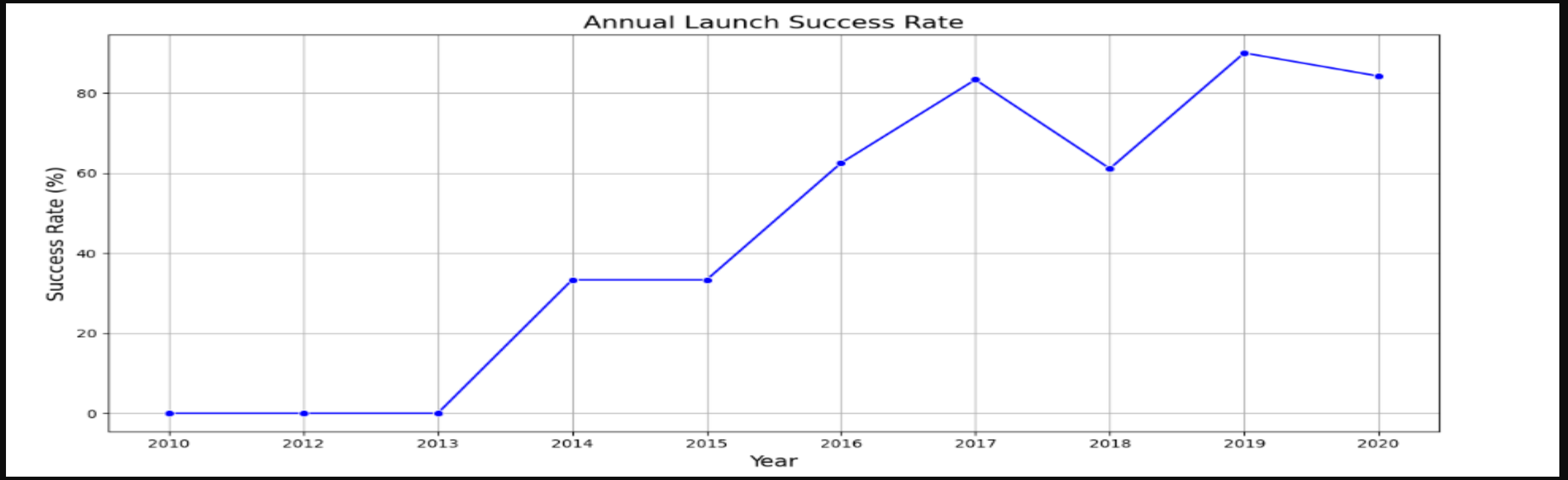
Heavy Payloads (Variable Outcomes): Payloads above 10,000 kg result in a mix of successes and failures, demonstrating increased difficulty for high-mass recovery.



Launch Success Yearly Trend

Clear Upward Trajectory: Since 2013, the annual launch success rate has shown significant improvement, reaching over 80% by 2020.

Resilient Trend: Although 2018 saw a temporary decline, the overarching pattern reflects a consistent increase in launch reliability over time.



All Launch Site Names

Generate a list of unique launch site identifiers from the space mission records.

```
%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'

Retrieve and display a sample of 5 records for which the launch site identifier begins with the characters 'CCA'.

```
%sql SELECT * FROM SPACEXTABLE 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	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

Total Payload Mass

Retrieve the total payload mass transported on missions where the customer is NASA CRS.

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

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

Present the average payload capacity delivered by the **Falcon 9 Version 1.1** launch vehicle.

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

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

Retrieve the historical date marking the first successful landing outcome on a terrestrial landing pad.

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

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Show all booster names where:

- Landing outcome = successful (drone ship)
- Payload mass > 4000 kg and < 6000 kg

```
%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.
```

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

Total Number of Successful and Failure Mission Outcomes

Calculate and present the total quantity of successful versus failed mission outcomes.

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTABLE WHERE "Mission_Outcome" IN ('Success', 'Failure') GROUP BY "Mission_Outcome";
```

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

Mission_Outcome	Total
Success	98

Boosters Carried Maximum Payload

Retrieve the names of the booster variants associated with the maximum payload mass in the mission history.

```
%sql SELECT DISTINCT "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

Show records where:

- Year = 2015
- Landing outcome = failed (drone ship)
  Include: booster version, launch site.

```
%%sql

SELECT
    CASE
        WHEN substr("Date", 6, 2) = '01' THEN 'January'
        WHEN substr("Date", 6, 2) = '02' THEN 'February'
        WHEN substr("Date", 6, 2) = '03' THEN 'March'
        WHEN substr("Date", 6, 2) = '04' THEN 'April'
        WHEN substr("Date", 6, 2) = '05' THEN 'May'
        WHEN substr("Date", 6, 2) = '06' THEN 'June'
        WHEN substr("Date", 6, 2) = '07' THEN 'July'
        WHEN substr("Date", 6, 2) = '08' THEN 'August'
        WHEN substr("Date", 6, 2) = '09' THEN 'September'
        WHEN substr("Date", 6, 2) = '10' THEN 'October'
        WHEN substr("Date", 6, 2) = '11' THEN 'November'
        WHEN substr("Date", 6, 2) = '12' THEN 'December'
        ELSE 'Unknown'
    END AS "Month_Name",
    "Mission_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTABLE
WHERE
    substr("Date", 0, 5) = '2015';
```

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

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

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

Show landing outcome types and their occurrence count between 2010-06-04 and 2017-03-20, ranked highest to lowest.

```
%%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

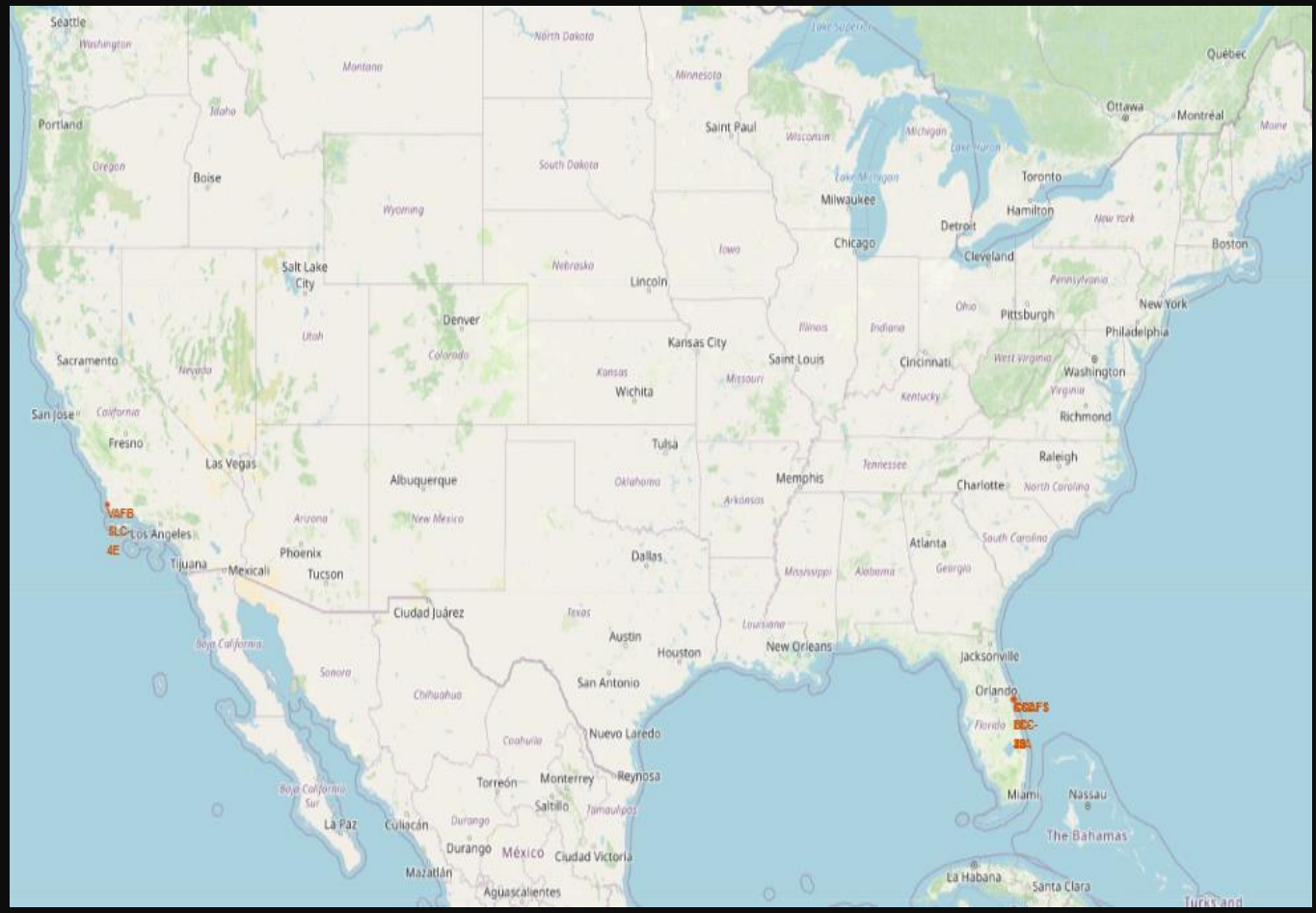
Section 3

Launch Sites Proximities Analysis

Global Map of Launch Site Locations

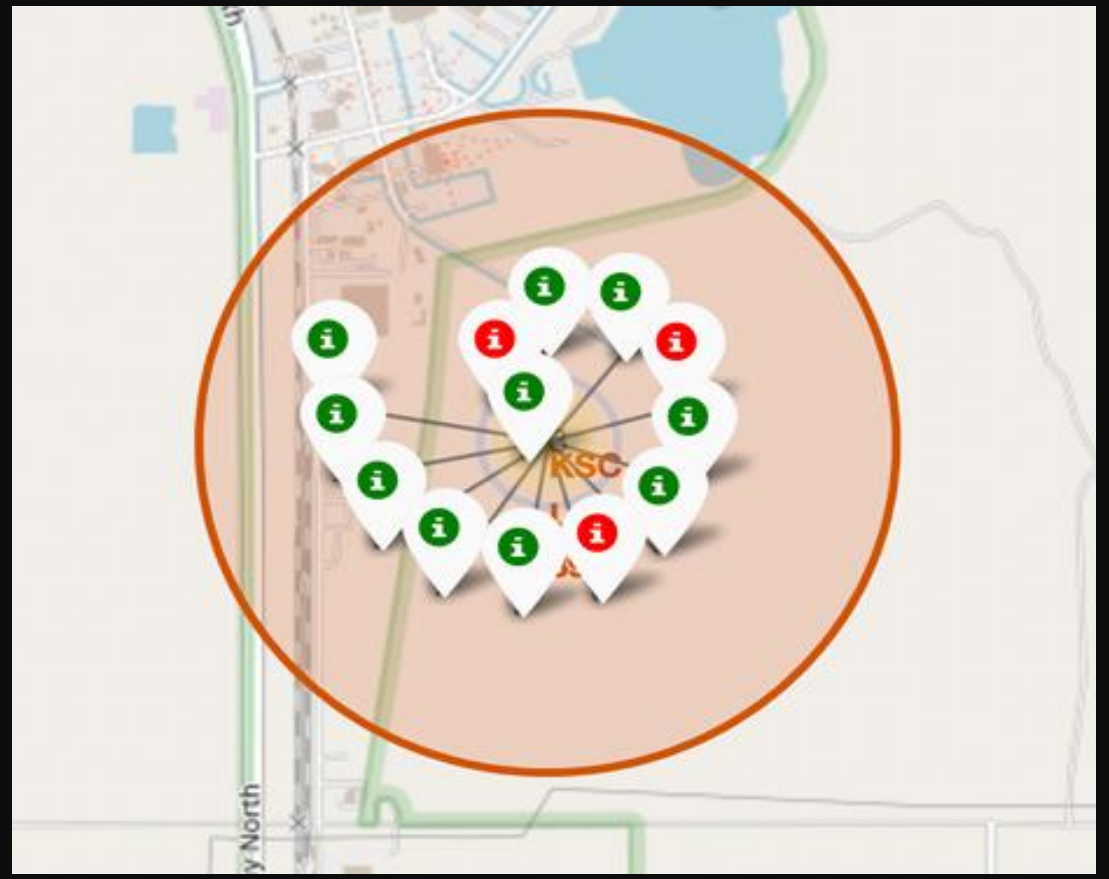
The geographic selection of launch sites is governed by two primary principles:

- **Orbital Mechanics:** Locations close to the equator harness the Earth's rotational momentum, giving rockets a "head start" of roughly 1,670 km/h, which is conserved in space due to inertia and aids in achieving stable orbit.
- **Risk Mitigation:** Coastal positions allow for launch trajectories over the ocean, creating a safe downrange corridor that isolates hazards such as booster separation debris or in-flight incidents from populated landmasses.



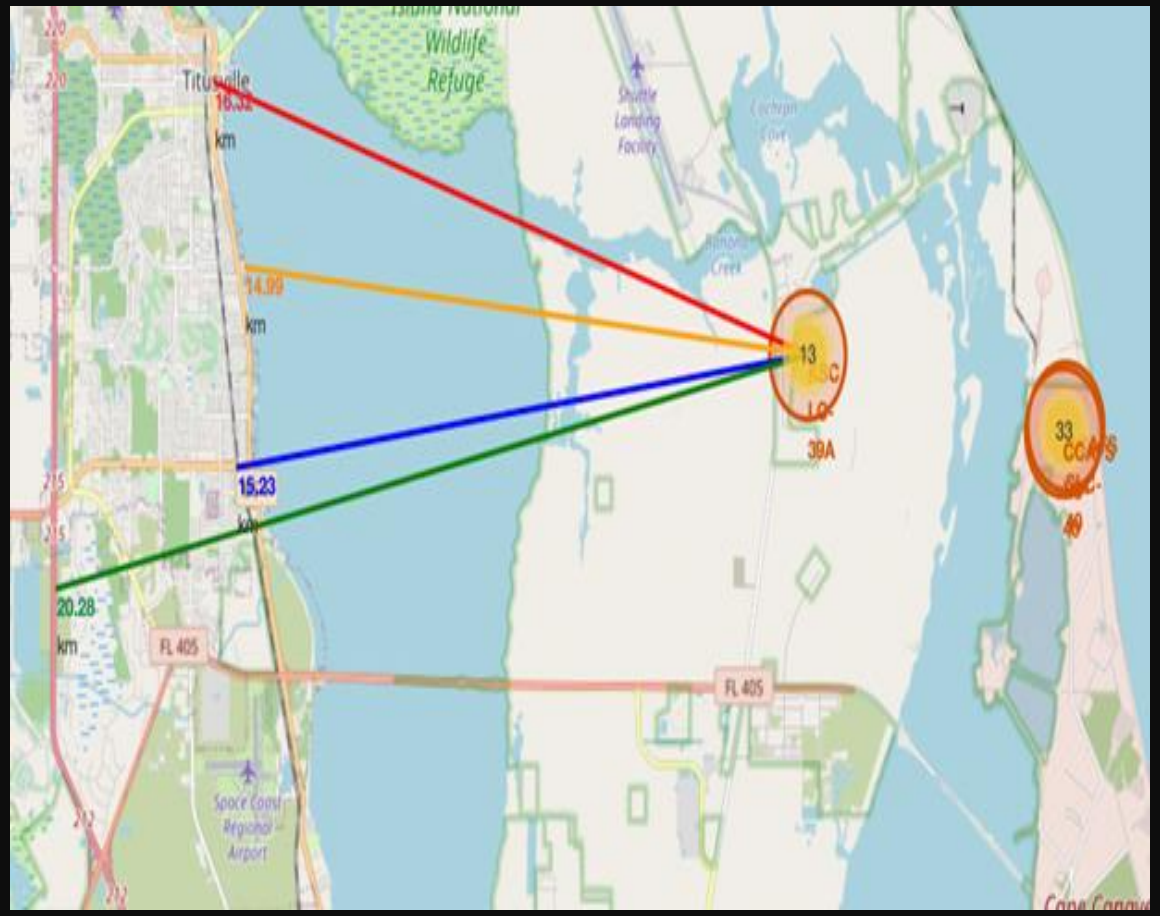
Visualization of Launch Outcomes by Geographic Location

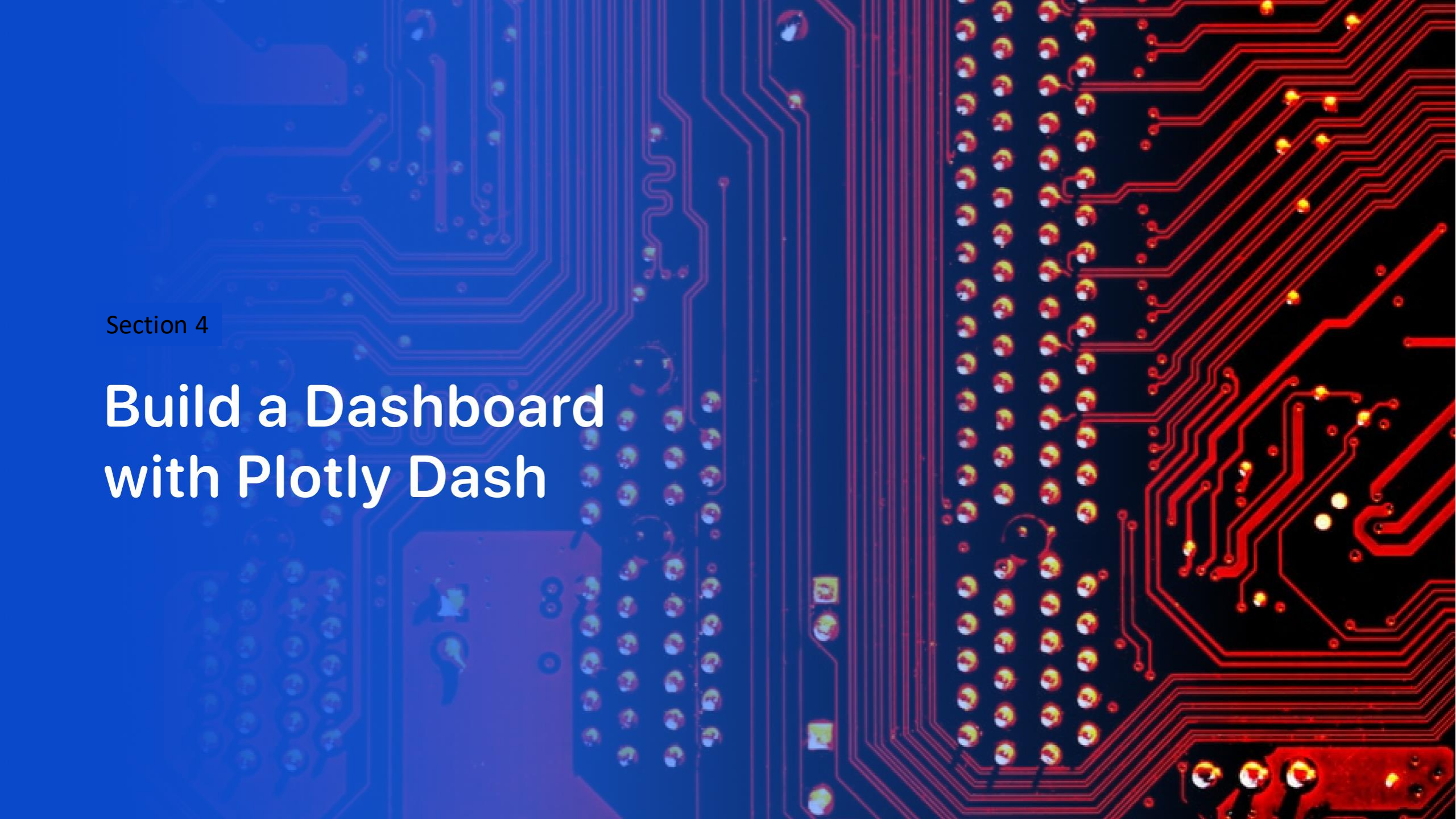
A visual analysis of the color-coded launch sites—where green represents success and red represents failure—reveals significant variance in reliability. KSC LC-39A is a notable standout, with its marker color indicating a very high success rate compared to other locations.



KSC LC-39A Proximity Distances

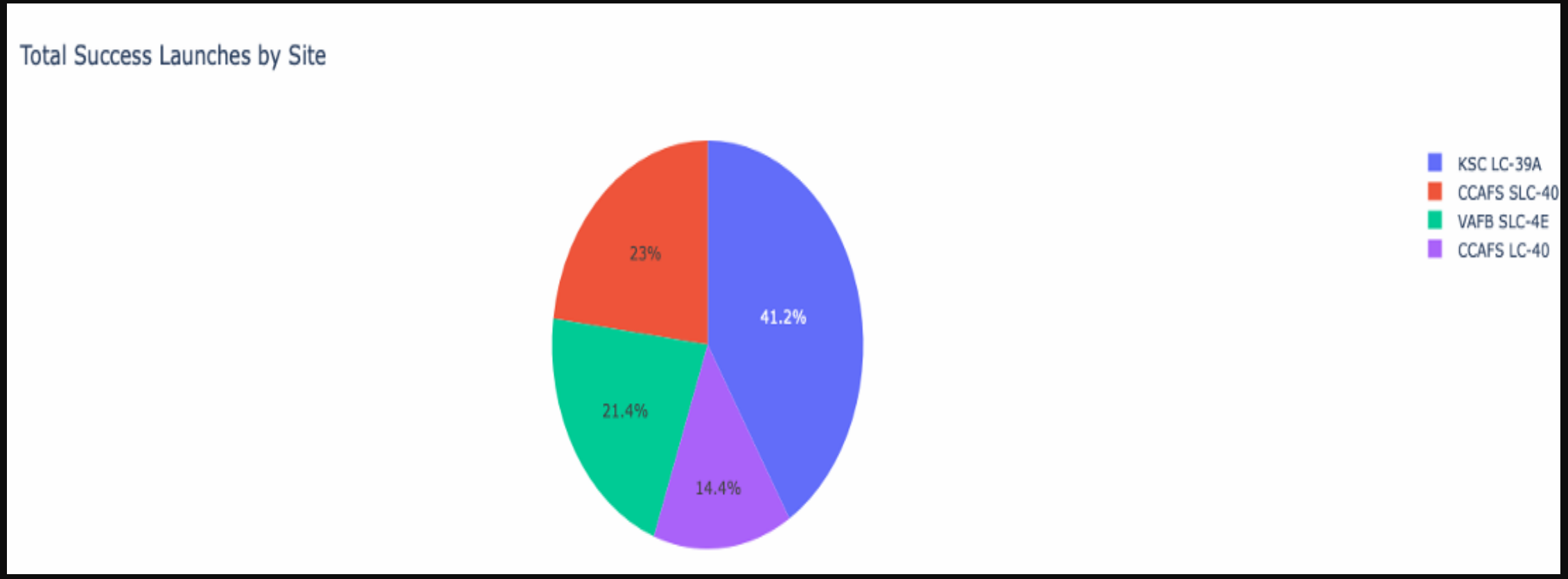
- **Proximity Analysis:** Launch site KSC LC-39A is located relatively close to:
  - Railway: 15.23 km
  - Highway: 20.28 km
  - Coastline: 14.99 km
  - Nearest City (Titusville): 16.32 km
- **Risk Implication:** Given the speed of a failing rocket, these distances of 15-20 km could be traversed in mere seconds, indicating a potential hazard to nearby populated areas and infrastructure.




Section 4
Build a Dashboard with Plotly Dash

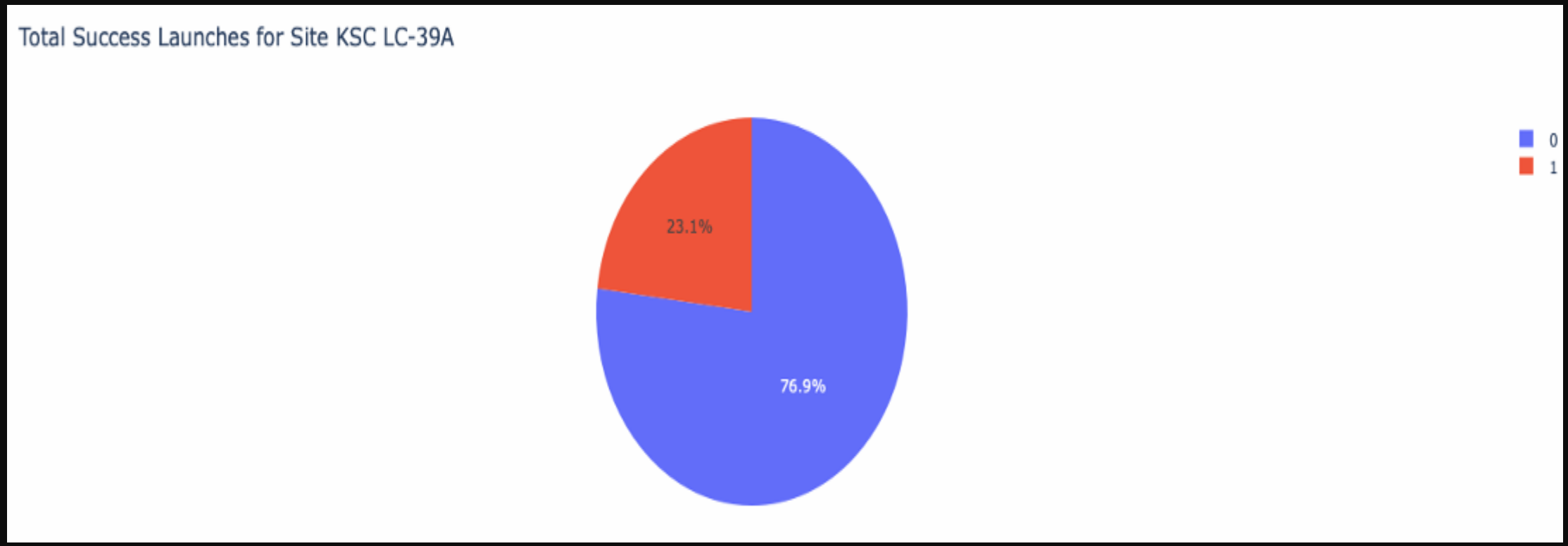
Successful Launch Count by Site

Based on the chart data, KSC LC-39A leads all launch sites in total successful launches.



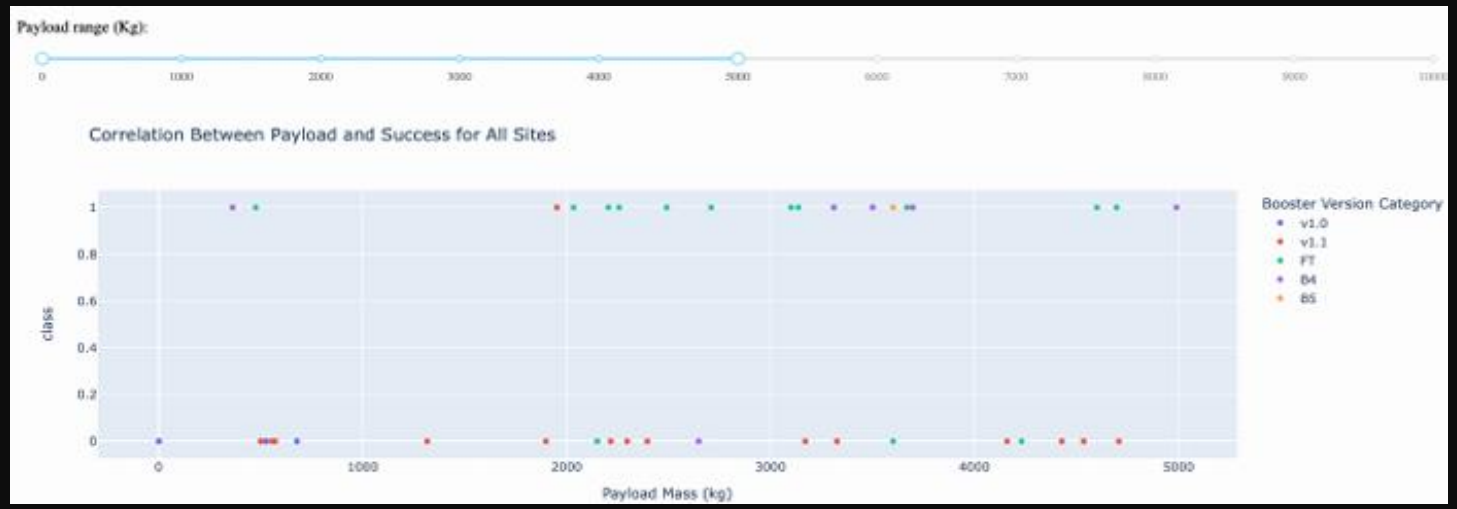
Analysis of Peak Launch Success by Site

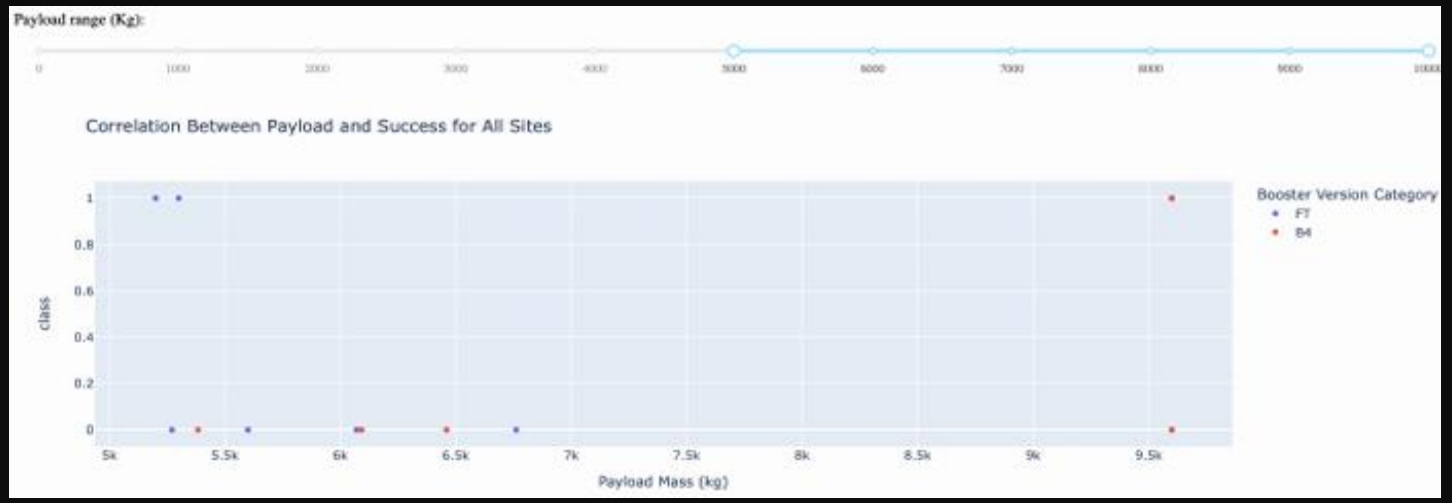
With a success rate of 76.9% (10 successful vs. 3 failed landings), KSC LC-39A demonstrates the highest reliability among all launch sites.



Payload Mass vs. Launch Outcome for all sites

The data visualization reveals that the highest mission success rate corresponds to payload masses ranging from 2,000 to 5,500 kg.



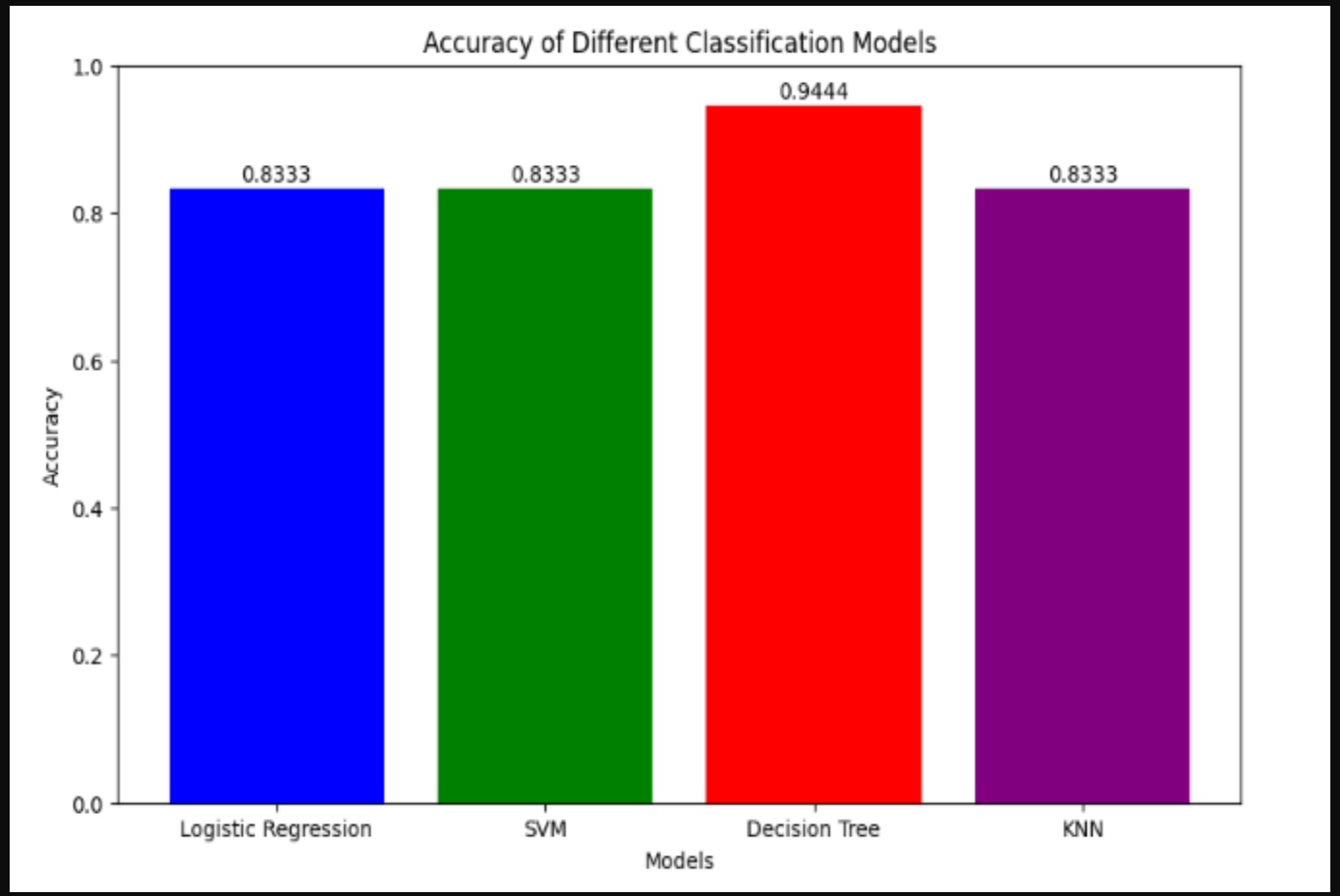


Section 5

Predictive Analysis (Classification)

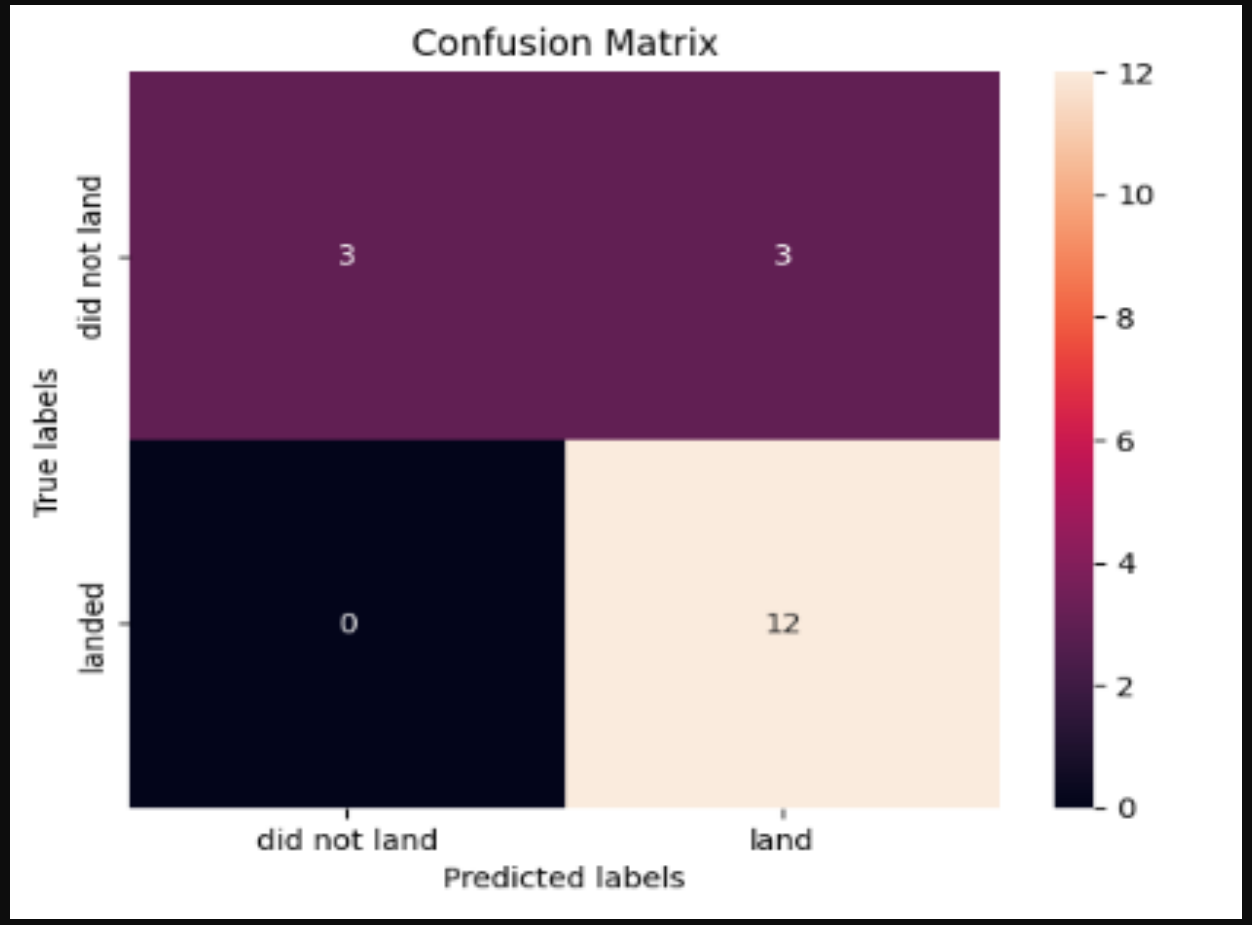
Classification Accuracy

Comparative analysis shows that the Decision Tree classifier achieves the highest classification accuracy on the test set (0.9444). In contrast, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors each yield a lower, identical accuracy of 0.8333, indicating the Decision Tree model is best suited for this specific dataset.



Confusion Matrix

Analysis of the confusion matrix shows that while logistic regression successfully separates the classes, its key weakness is a significant rate of false positives.



Conclusions

- **Optimal Model:** The Decision Tree algorithm delivers the best predictive performance for this dataset.
- **Payload Impact:** Success rates are inversely related to payload mass, with lighter payloads achieving better outcomes.
- **Site Geography:** Launch sites are strategically located near the equator to harness rotational velocity and along coastlines for safety.
- **Temporal Trend:** Launch success rates have shown consistent improvement over time.
- **Top-Performing Site:** KSC LC-39A has the highest success rate among all launch sites.
- **High-Reliability Orbits:** Orbits ES-L1, GEO, HEO, and SSO have maintained a 100% success rate.

Thank you!