IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kunle
6 June,2025

Outline

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

Executive Summary

- In the rapidly evolving aerospace sector, SpaceX has emerged as a transformative force, offering cost-effective launch services that have redefined industry benchmarks. By advancing reusable rocket technology, the company has significantly reduced mission costs, charging approximately $62 million per launch compared to the traditional $165 million, securing contracts from major agencies such as NASA. Within this competitive landscape, the project simulates the role of a data scientist within a new entrant, Space Y, tasked with leveraging data science methodologies to assess and enhance competitiveness against established players.
- To achieve this objective, a comprehensive data pipeline was implemented, beginning with targeted data acquisition via APIs and web scraping. Preprocessing steps using Python ensured data integrity, followed by SQL-based querying to uncover critical mission attributes. Exploratory and visual analytics highlighted operational variables associated with landing outcomes. The analysis culminated in the deployment of supervised machine learning models, with the top-performing classifier achieving an accuracy of 83% in predicting successful landings. These findings provide actionable intelligence for strategic planning in space mission logistics and cost optimization

Introduction

- Since the inception of orbital exploration in the mid-20th century, space missions have remained capital-intensive, with per-launch costs historically exceeding hundreds of millions of dollars. The advent of reusable rocket stages has disrupted this paradigm. Among leading disruptors, SpaceX has demonstrated consistent success in recovering the first stage of its launch vehicles, thereby drastically reducing hardware attrition and overall mission expenditure. The capacity to reliably land and reuse rocket stages is not merely a technical milestone but a central economic advantage in modern aerospace logistics.
- This project investigates the operational variables that influence first-stage landing outcomes, including payload characteristics, launch geography, and mission trajectory. Through comprehensive data-driven modeling and visualization, the objective is to evaluate which factors most strongly affect recovery success and to construct predictive tools capable of enhancing launch planning and cost-efficiency for emerging entrants in the space launch ecosystem.

Section 1

Methodology

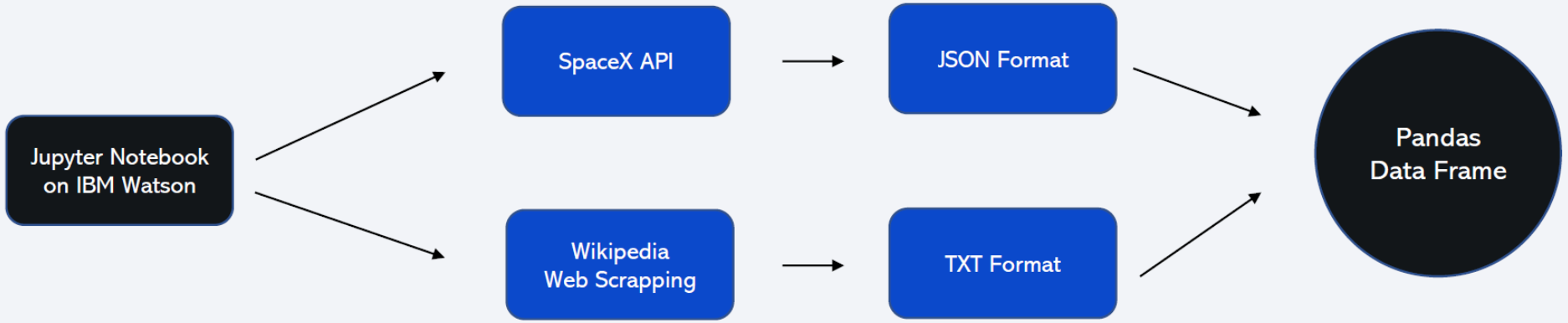
Methodology

Executive Summary

- Data Acquisition
  - Data was sourced through the SpaceX REST API and supplemented with web-scraped content from Wikipedia to ensure comprehensive coverage.
- Data Wrangling
  - Data preprocessing was conducted using Pandas and NumPy, involving techniques such as one-hot encoding, removal of irrelevant columns, normalization, and standardization.
- Exploratory Data Analysis (EDA)
  - Exploratory insights were derived using SQL queries and visualized through Seaborn and Matplotlib to uncover patterns and relationships among key variables.
- Interactive Visual Analytics
  - Dynamic and geographical visualizations were created using Folium and Plotly Dash to enhance data interpretation and engagement.
- Predictive Modeling
  - Classification models were developed following a structured machine learning pipeline, splitting the dataset, performing hyperparameter tuning via Grid Search, and deploying the optimal model configuration for accurate prediction of landing outcomes.

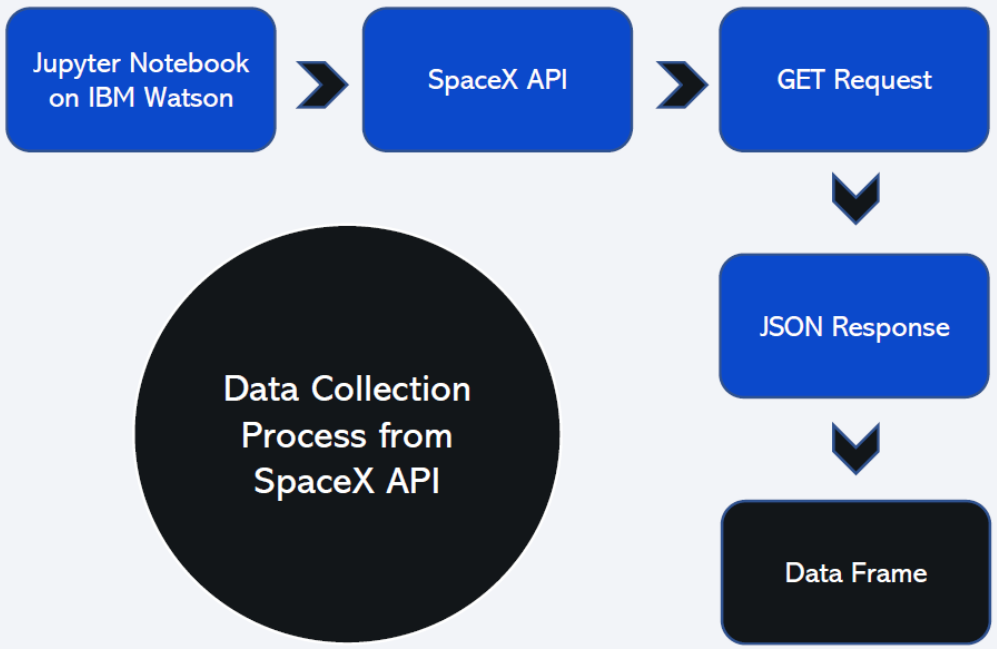
Data Collection

- To support predictive modeling and visual analytics, two complementary data acquisition methods were employed: RESTful API integration and web scraping. These techniques ensured the collection of up-to-date and historical records of Falcon 9 launch missions. The workflow began with structured API calls to SpaceX's open-source endpoints, followed by HTML parsing of publicly available mission logs hosted on Wikipedia. This dual approach enabled the compilation of a comprehensive dataset, combining both technical parameters and manually curated launch outcomes.
- Flowchart – Data Collection Workflow



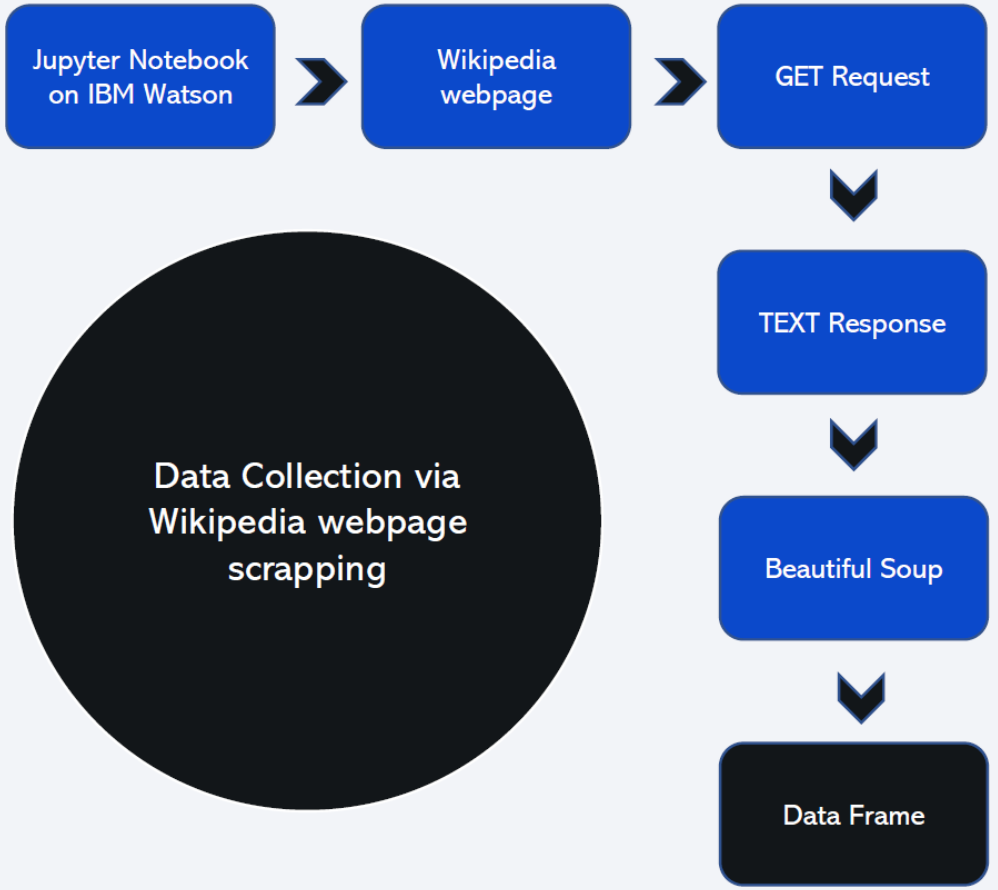
Data Collection – SpaceX API

- The first dataset was sourced from the SpaceX REST API, which offers programmatic access to launch records, rocket specifications, core components, payload data, and more. The data acquisition process began with the import of necessary libraries including requests, pandas, and JSON. Custom helper functions were implemented to automate GET requests across multiple endpoints, each identified by unique launch IDs.
- The responses were returned in JSON format, decoded and normalized to ensure structural consistency, and then converted into a Pandas DataFrame. Relevant attributes collected included launch site, orbit, payload mass, booster version, and landing outcome. This pipeline ensured that dynamic and structured launch event records were captured reliably.
- 🔗 GitHub – Completed SpaceX API Calls Notebook



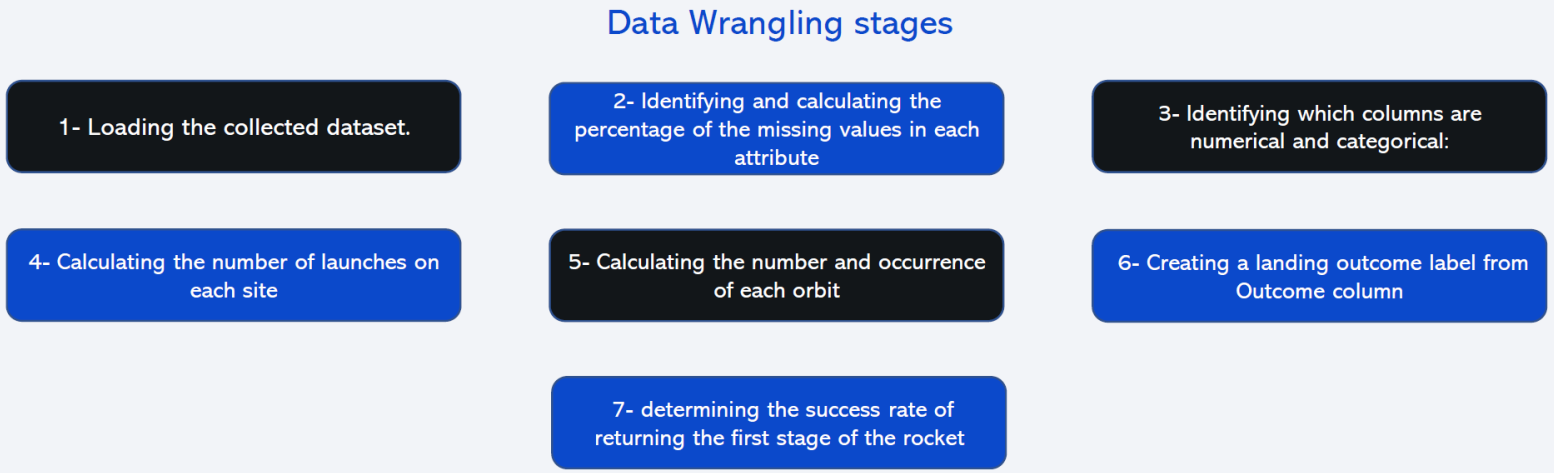
Data Collection - Scraping

- To complement the API-sourced dataset, web scraping was conducted on a Wikipedia page containing historical launch records of the Falcon 9 program. The scraping process involved the use of requests to issue an HTTP GET request to the target URL, followed by BeautifulSoup for parsing the HTML content.
- Specific HTML table elements were located and filtered using CSS selectors, and the extracted data was then transformed into a structured DataFrame using Pandas. Key fields retrieved included launch date, mission name, landing outcome, and customer organization. This enriched the dataset with descriptive labels and historical commentary that were not included in the API data.
- 🔗 GitHub – Completed Web Scraping Notebook



Data Wrangling

- **Processing Overview:**
  - Imported raw datasets collected via SpaceX API and Wikipedia web scraping
  - Removed redundant and irrelevant columns to minimize noise
  - Handled missing or inconsistent data entries for accuracy
  - Applied normalization and standardization to numerical features
  - Created binary classification target (Landing Outcome) from mission status data
  - Performed one-hot encoding for categorical variables such as launch site and rocket type
  - Saved the cleaned dataset to a structured .csv format for downstream modeling
- **GitHub Reference:**
- 🔗 Completed Data Wrangling Notebook



EDA with Data Visualization

- **Visualizations and Purpose:**
  - Bar Charts: Used to compare launch success rates across sites
  - Scatter Plots: Explored relationships between payload mass and success rates
  - Heatmaps: Visualized feature correlations to identify predictive attributes
  - Boxplots: Assessed distribution and variability of payloads across mission types
  - Line Charts: Illustrated trends in launch frequency over time
  - Pie Charts: Displayed orbit type distributions to identify dominance in launch patterns
  - These plots provided both descriptive and inferential insights, supporting feature selection and hypothesis formulation.
- **GitHub Reference:**
- 🔗 EDA Visualization Notebook

EDA with SQL

- **Summary of SQL Queries:**
  - Retrieved unique launch site names
  - Filtered launch records starting with ‘CCA’
  - Aggregated total payload mass for NASA (CRS) missions
  - Identified the date of the first successful landing on a ground pad
  - Listed boosters with successful drone ship landings and payloads between 4000–6000 kg
  - Counted total successful and failed mission outcomes
  - Found booster versions that carried the maximum payload mass
  - Extracted failed drone ship landings in 2015 with associated booster and site data
  - Ranked landing outcomes by frequency between June 2010 and March 2017
- **GitHub Reference:**
- 🔗 SQL EDA Notebook

Build an Interactive Map with Folium

- To visually represent launch site data and mission outcomes, an interactive map was constructed using the Folium library. Key launch locations, including CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E, were marked using Circle and Marker objects to display both geographic coordinates and landing outcomes.
- Circles provided spatial emphasis on launch sites, while markers highlighted successful or failed first-stage recoveries.
- Additionally, PolyLine objects were added to illustrate calculated distances from the CCAFS LC-40 site to nearby landmarks such as the closest city, coastline, and highway, offering contextual insights into geographic proximity and logistical considerations.
- These elements were integrated to enhance the interpretability of spatial patterns in Falcon 9 operations.
- GitHub URL of the completed interactive Folium map: Click Here

Build a Dashboard with Plotly Dash

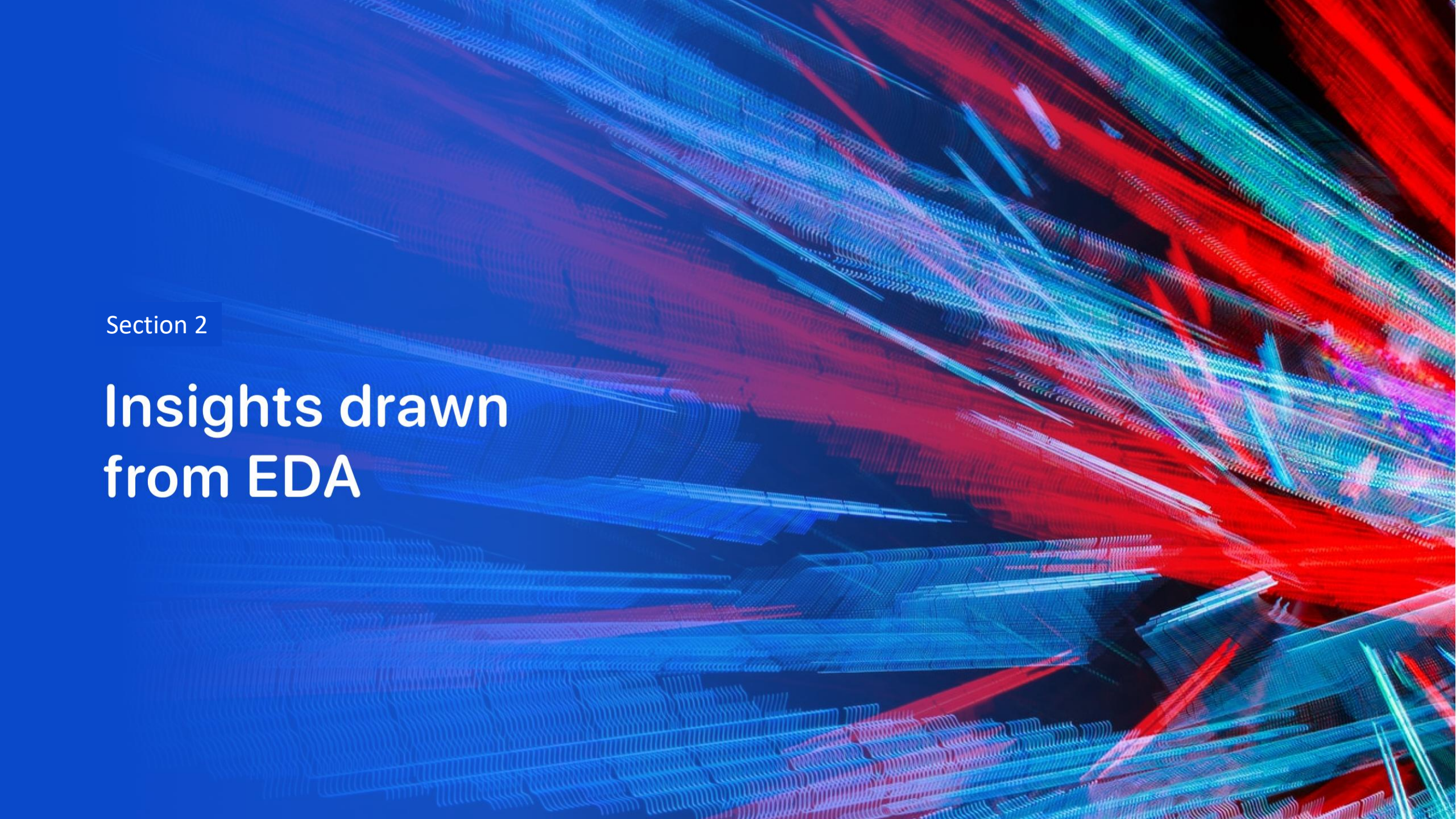
- An interactive dashboard was developed using Plotly Dash to enable dynamic exploration of Falcon 9 launch data.
- A dropdown component was implemented to allow users to filter by specific launch sites, namely All Sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, and KSC LC-39A, enhancing the ability to perform site-specific analysis.
- A pie chart was included to display the distribution of successful launches across selected sites, providing a quick overview of launch success rates. To further examine launch outcomes in relation to payload mass, a range slider was added, allowing users to filter payload values between 0 and 10,000 kg.
- Complementing this, a scatter plot was integrated to reveal the correlation between payload mass and mission success, offering visual insights into operational performance under varying payload conditions. These interactive components were chosen to support granular, data-driven decision-making.
- GitHub URL of the completed Plotly Dash lab: Click Here

Predictive Analysis (Classification)

- To forecast the success of Falcon 9 first-stage landings, a structured machine learning workflow was implemented. The process began with importing essential libraries and loading the preprocessed dataset.
- Feature standardization was applied to ensure balanced model performance across varying scales. The dataset was partitioned into training (80%) and testing (20%) subsets to facilitate unbiased evaluation.
- Four classification algorithms: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors were initialized.
- A Grid Search technique was used to optimize hyperparameters for each model.
- Model performance was rigorously assessed using evaluation metrics such as confusion matrix, F1-score, and Jaccard index.
- This systematic approach led to the selection of the most effective model, which achieved a predictive accuracy of 83%, demonstrating strong capability in identifying factors associated with successful landings.
- 📎 GitHub URL of the completed predictive analysis lab: Click Here

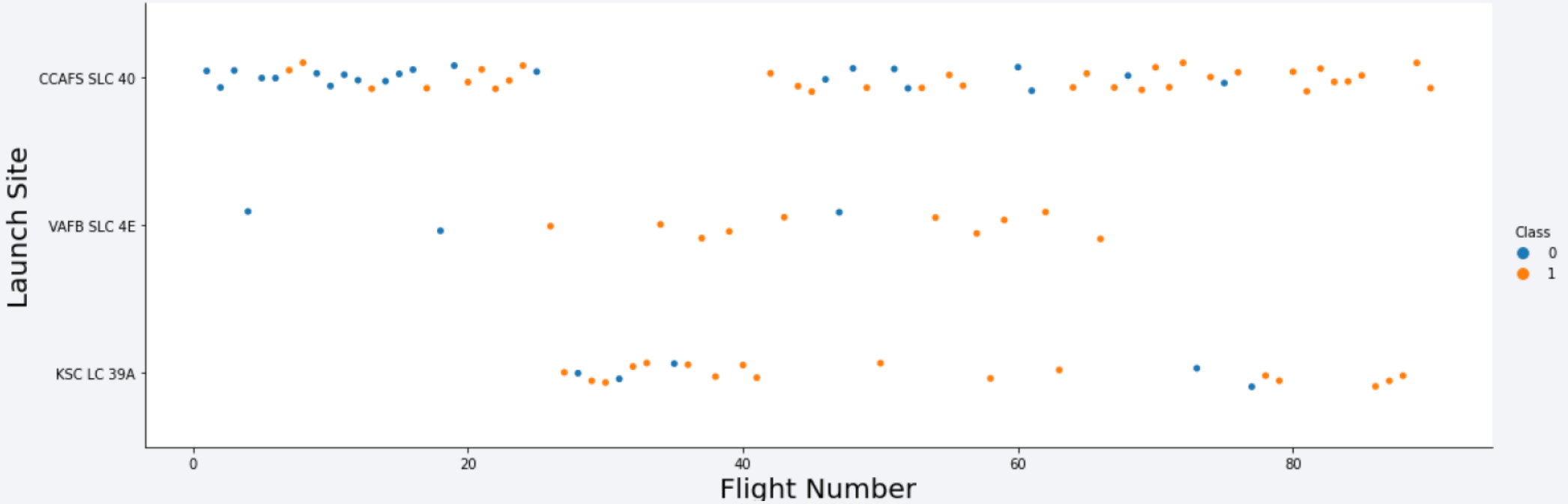
Results

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


Section 2
Insights drawn from EDA

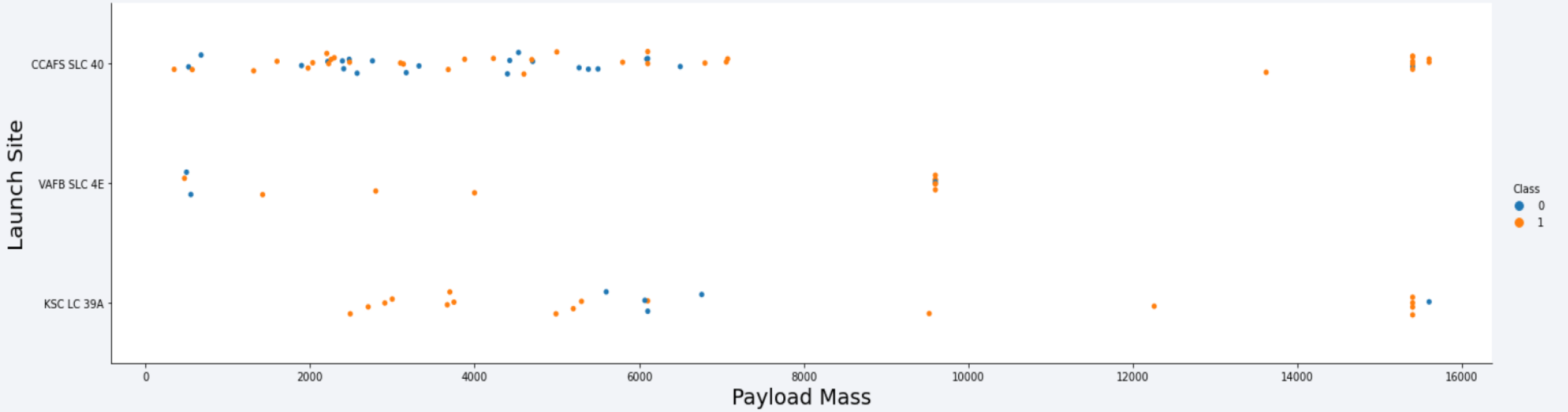
Flight Number vs. Launch Site

- CCAFS SLC 40 emerged as the most frequently used launch site with 55 launches, achieving a 60% success rate (33 successful, 22 failed).
- VAFB SLC 4E recorded the fewest launches (13 total) but maintained a high success rate of 77% (10 successful).
- KSC LC 39A showed moderate usage with 22 launches and similarly achieved a 77% success rate (17 successful).



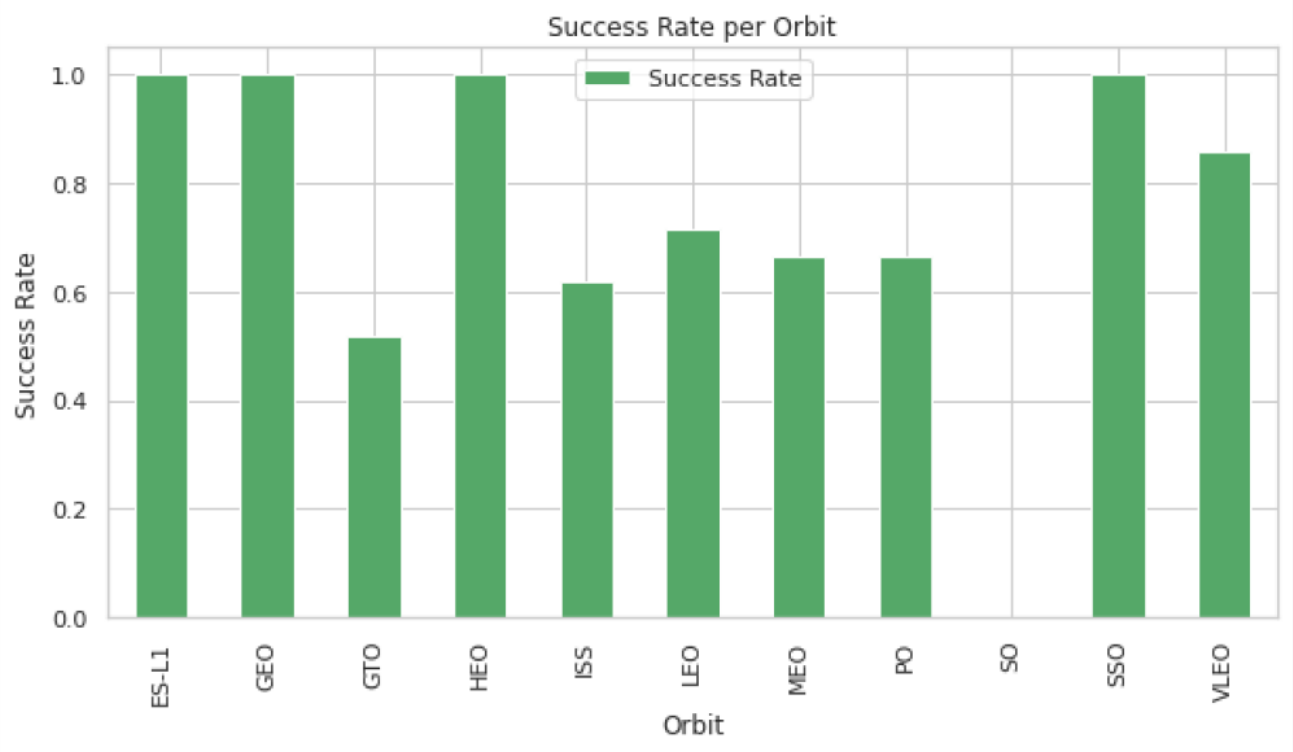
Payload vs. Launch Site

- Analysis indicates no strong correlation between payload mass and first-stage landing success.
- Both successful and failed landings occurred across a wide range of payload weights, suggesting other factors are more influential.



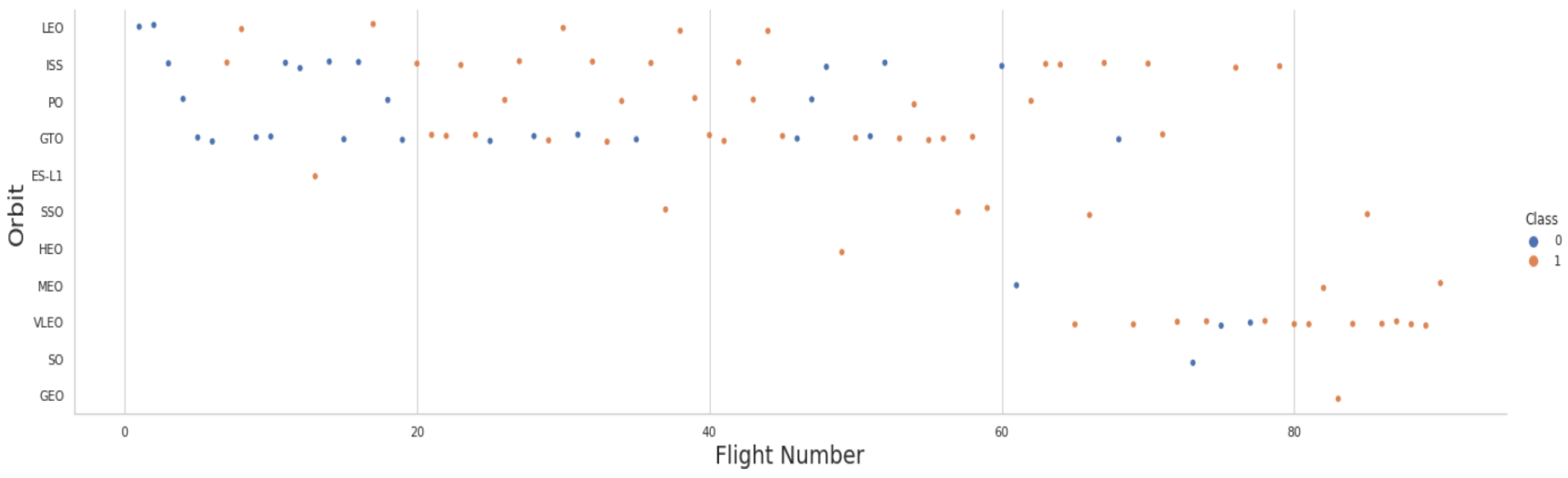
Success Rate vs. Orbit Type

- The highest landing success rates were observed for the following orbits: ES-L1, GEO, HEO, and SSO.
- GTO (Geostationary Transfer Orbit) exhibited the lowest success rate, highlighting a need for further investigation to mitigate risks associated with this orbit.



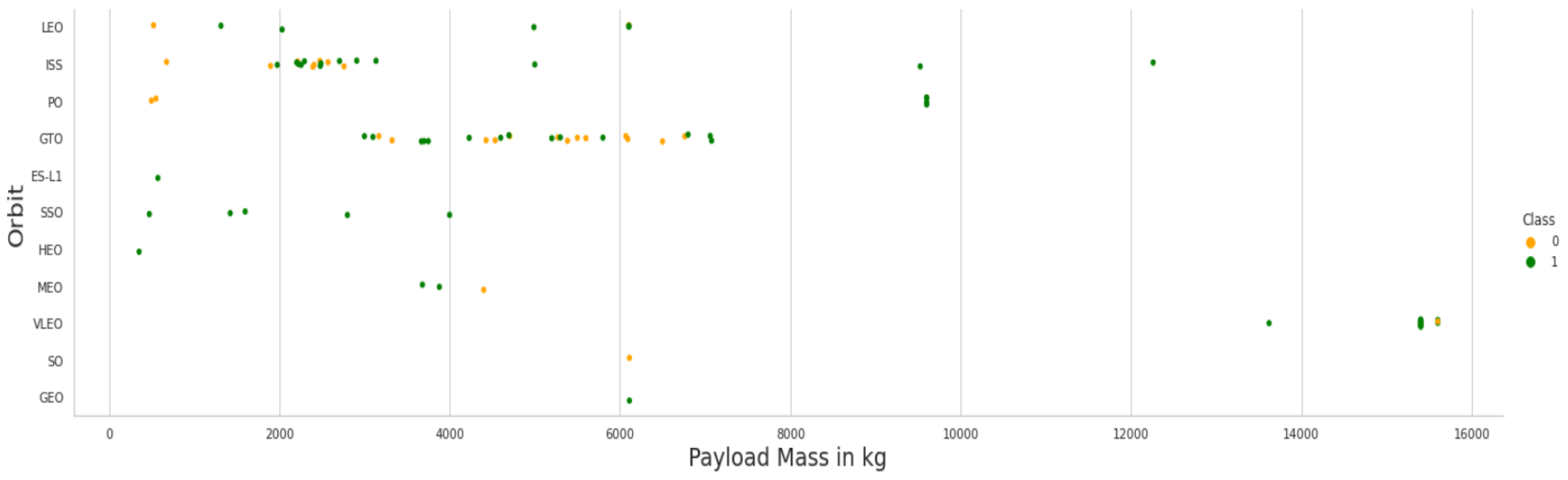
Flight Number vs. Orbit Type

- A positive correlation is observed between flight number and landing success in LEO orbit.
- No clear relationship is found between flight number and success for GTO orbit missions.



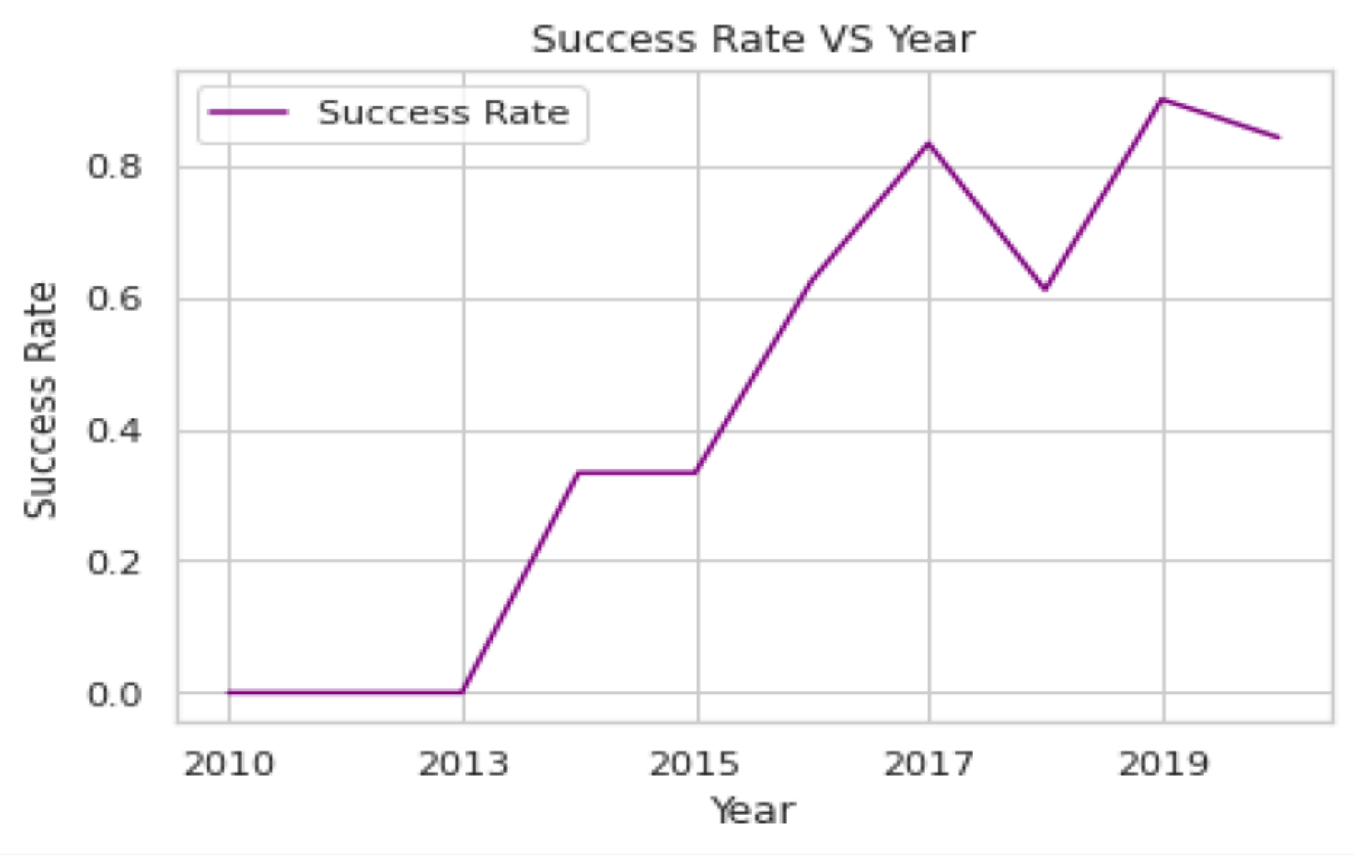
Payload vs. Orbit Type

- Heavy payloads tend to negatively affect success in GTO orbits.
- Positive performance is noted for heavy payloads in Polar LEO and ISS-targeted orbits.



Launch Success Yearly Trend

- From 2013 to 2020, there is a steady upward trend in successful mission outcomes, showing operational improvement over time.



All Launch Site Names

- A total of 4 unique launch sites were identified and analyzed in the dataset:
  - CCAFS LC 40, CCAFS SLC 40, KSC LC 39A, and VAFB SLC 4E.

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

```
%sql select distinct launch_site from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- The query showed five records where launch sites begin with ‘CCA’.
- The booster version are F9 v1 series with success mission outcome.

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

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcom
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- NASA’s total payload mass launched via SpaceX is 45,596 kg, equivalent to approximately 50.26 US tons.

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

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

- The F9 v1.1 booster carried an average payload of 2,928 kg, highlighting its moderate lifting capability.

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

```
%sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

avg_mass_f9
2928

First Successful Ground Landing Date

- The first successful landing on a ground pad occurred on December 22, 2015, marking a major milestone in SpaceX recovery technology.

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters that successfully landed on drone ships with medium payloads (4,000–6,000 kg) include:
- F9 FT B1029.1
- F9 FT B1036.1
- F9 B4 B1041.1

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

```
%sql select booster_version from SPACEXTBL\
where (landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and  payload_mass__kg_ > 6000));
```

booster_version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

Total Number of Successful and Failure Mission Outcomes

- An overwhelmingly high success rate: 99 successful missions versus only 1 failure, reinforcing SpaceX's operational reliability.

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

mission_outcome	counts
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The F9 B5 series, specifically B1048 to B1060, handled the heaviest payloads, showcasing advancements in booster capacity.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select distinct booster_version from SPACEXTBL\
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- In 2015, there were two failed drone ship landings, both from CCAFS LC 40 and using the F9 v1.1 booster version.

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select landing__outcome, booster_version, launch_site from SPACEXTBL\
where (landing__outcome = 'Failure (drone ship)' and date like '2015%')
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- No Attempt: Most common (10 times), showing early missions lacked recovery efforts.
- Success vs. Failure (Drone Ship): Equal frequency (5 each), indicating mixed offshore landing results.
- Controlled Ocean & Ground Pad Success: Occurred 3 times each.
- Other Outcomes: Rare parachute failures, uncontrolled ocean landings (2 each), and one precluded attempt.

```
%sql select landing__outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL\
where DATE between '2010-06-04' and  '2017-03-20' group by landing__outcome\
order by count(landing__outcome) desc
```

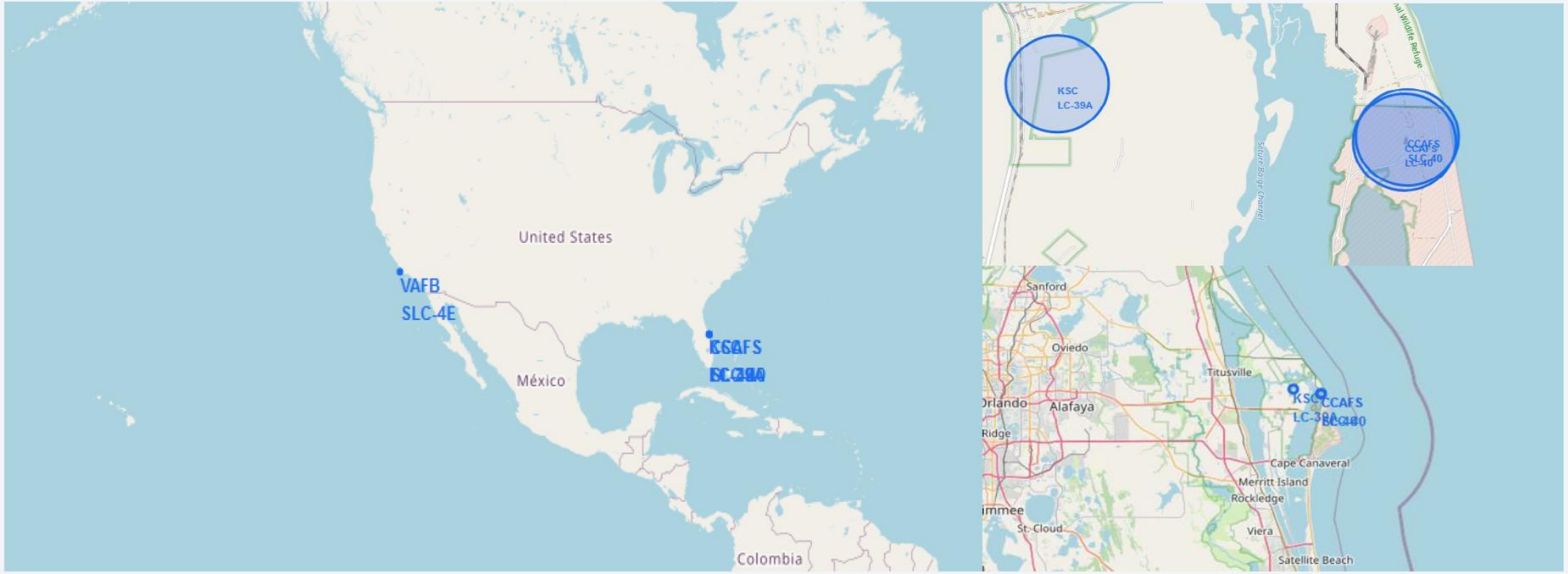
landing__outcome	counts_of_landing_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Folium Map: Launch Sites

- All launch sites are located near coastlines and close to the equator.
- Strategically chosen to minimize risk and optimize launch trajectory.
- Geographically distributed between Florida and California.



Folium Map: Launch Success Rates

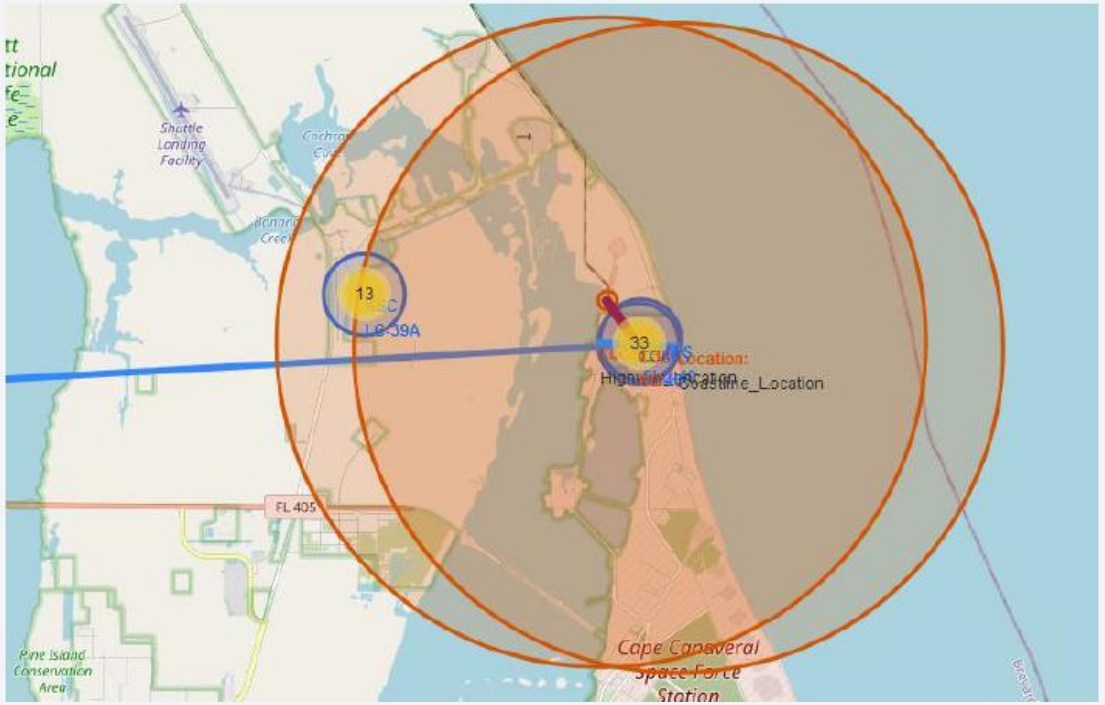
- Green markers indicate successful returns; Red markers indicate failures.
- Visual clustering reveals that some sites consistently outperform others in landing success.

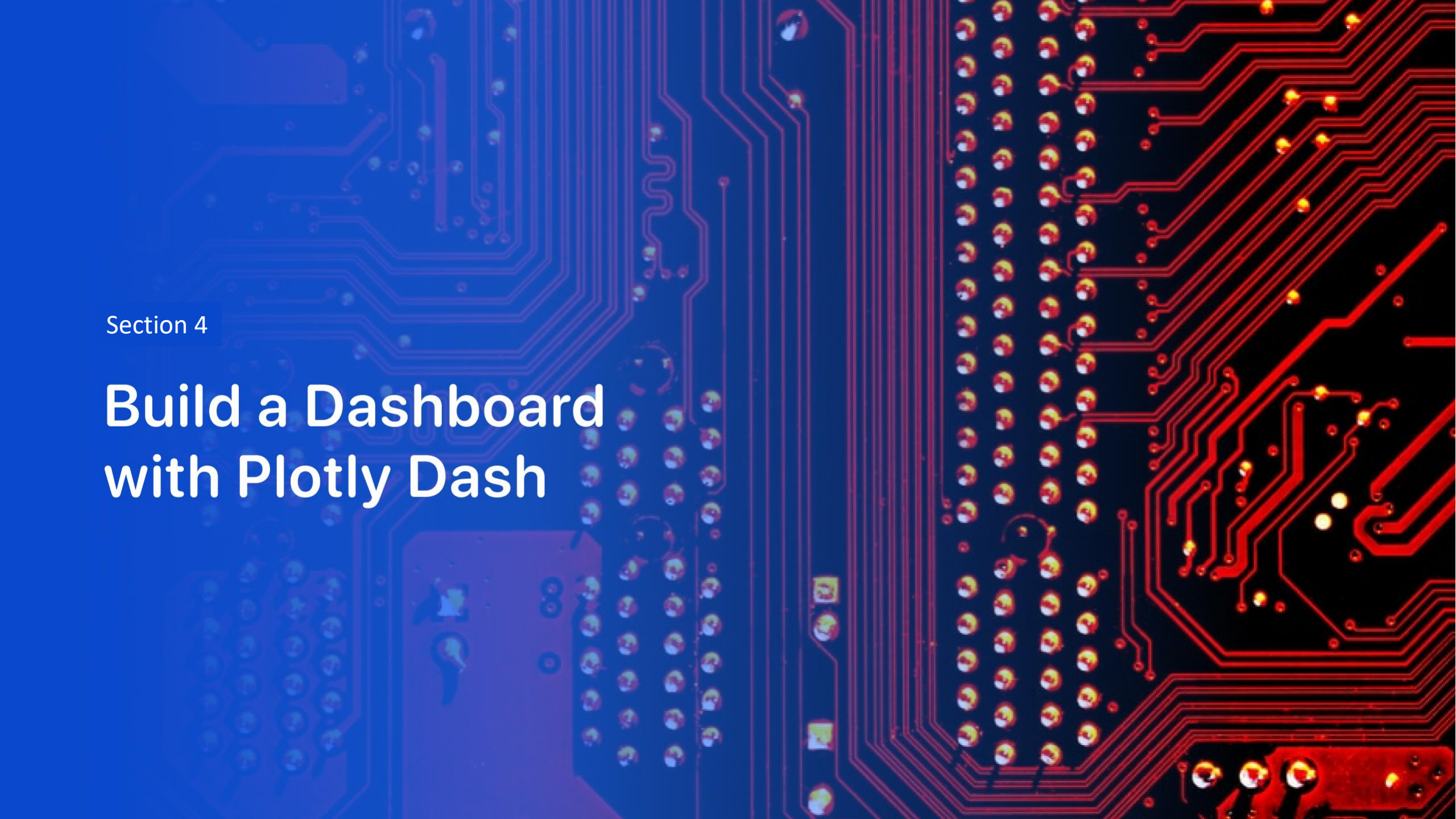
FoI Map: CCAFS LC 40 Proximities

- Calculated distances from launch site:
  - Orlando City: ~78.8 km
  - Coastline: ~0.97 km
  - Nearest Highway: ~0.95 km

- Proximities

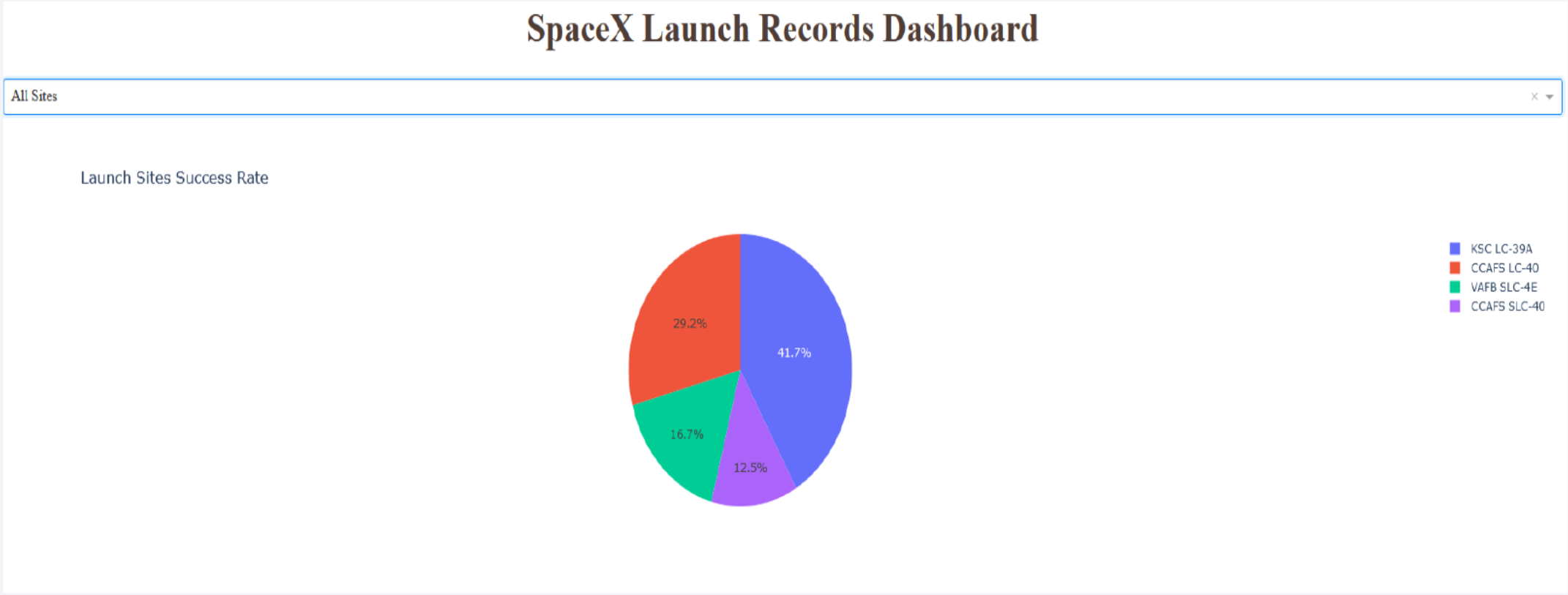
N	Location	Lat	Long
0	Orlando_Location	28.523	-81.3826
1	Coastline_Location	28.56146	-56.746
2	Highway_Location	28.5627	-80.587




Section 4
Build a Dashboard with Plotly Dash

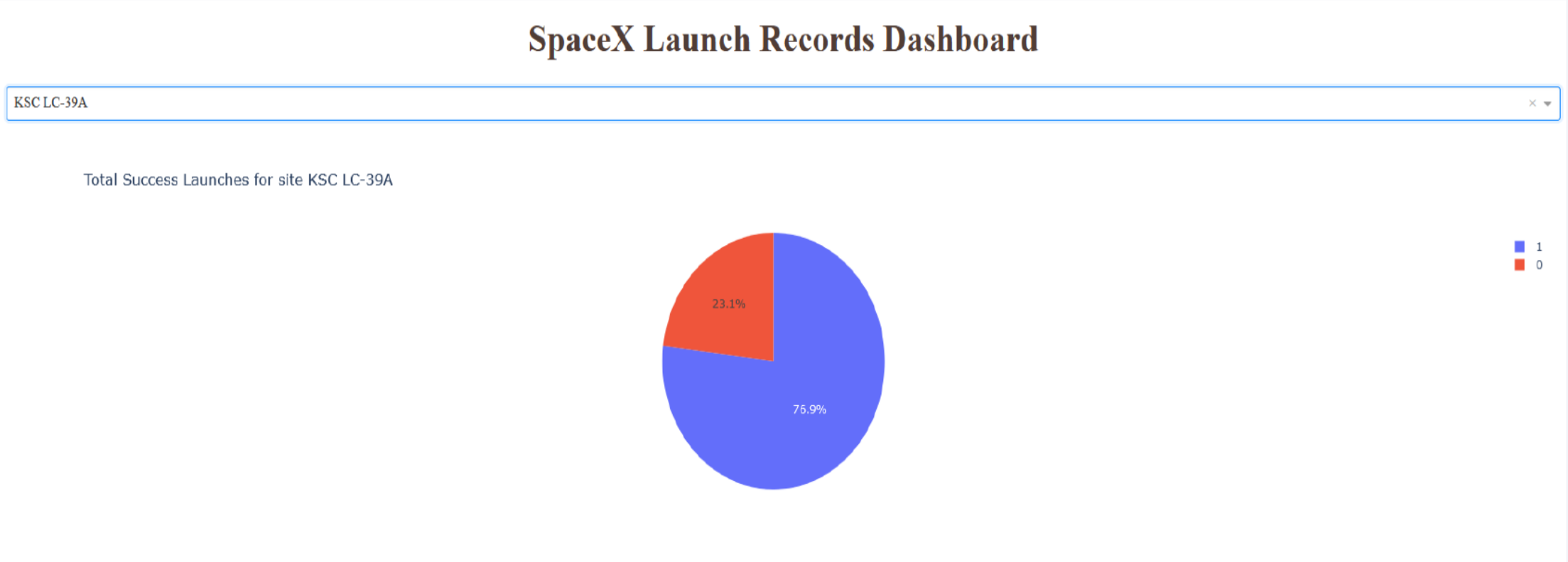
Dashboard: Launch Success by Site

- KSC LC 39A: Highest share of successful returns (41.7%)
- CCAFS SLC 40: Lowest contribution to success (12.5%)



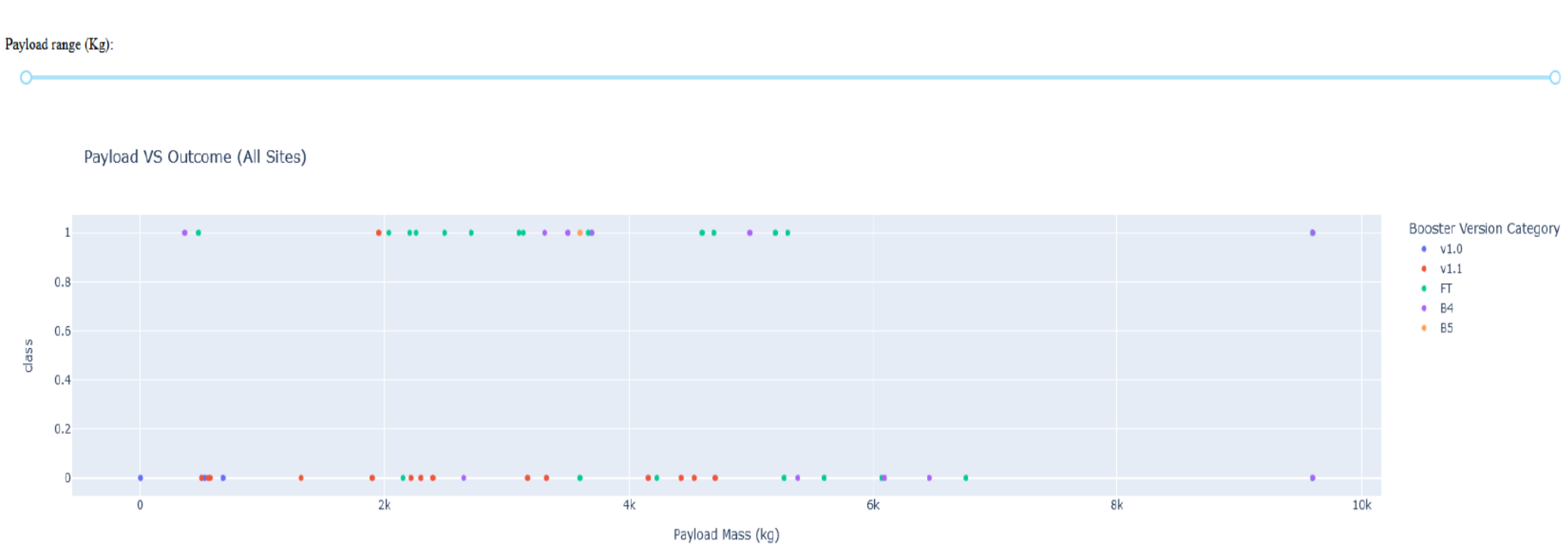
Dashboard: KSC LC 39A Launch Breakdown

- 76.9% of launches were successful
- 23.1% ended in failure



Dashboard: Payload vs. Launch Outcome

- Interactive scatter plot shows:
  - Boosters with < 4000 kg payloads had higher success rates.
  - Clear trend showing payload mass impacts outcome likelihood.

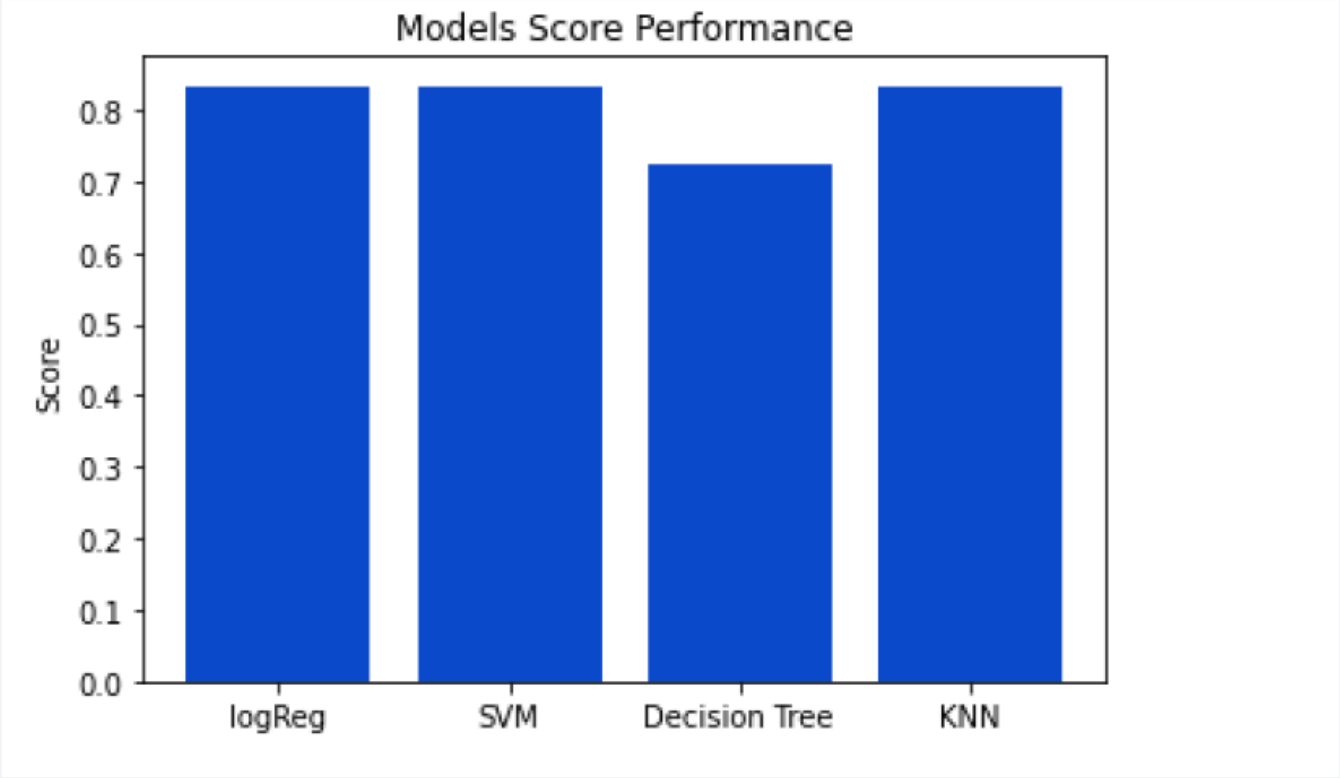


Section 5

Predictive Analysis (Classification)

Classification Accuracy

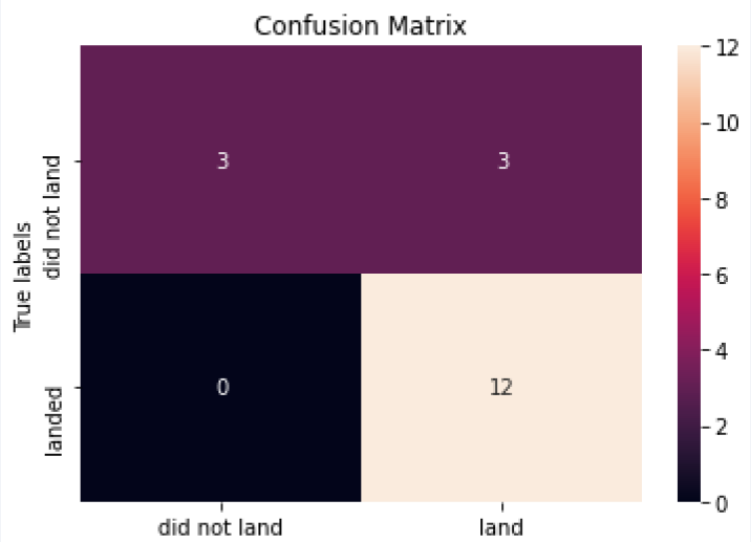
- Logistic Regression, SVM, and KNN each achieved a Jaccard Score of 0.8
- Decision Tree performed the worst among all tested models.




```
Logistic Regression:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
SVM:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
Decision Tree:
Jaccard Score of = 0.6666666666666666
F1 Score = 0.6727272727272727
====================================
KNN:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
```

Confusion Matrix

- **Classification Model Evaluation (LR, SVM, KNN)**
- All three models produced identical results:
  - True Positives (TP): 12
  - False Positives (FP): 0
  - True Negatives (TN): 3
  - False Negatives (FN): 3
- **Performance Highlights:**
  - High precision: No false positives
  - Moderate recall: Some missed successful landings (false negatives)
  - Indicates the models are conservative in predicting success but highly reliable when they do



Conclusions

- A comprehensive data-driven approach was applied to analyze and model SpaceX Falcon 9 launches.
- Data was collected from SpaceX API and Wikipedia, then wrangled, visualized, and analyzed using Python, SQL, and interactive tools.
- Key insights were revealed on launch success trends, site performance, payload impact, and orbit outcomes.
- An interactive dashboard and geospatial map enhanced interpretability of mission outcomes.
- Classification models were developed to predict first-stage landing success with up to 83% accuracy, supporting future mission planning and strategy.
- These findings demonstrate the value of data science in guiding competitive decision-making in the aerospace sector.

Appendix

- **Python Code Snippets**
    - API requests and parsing JSON data into DataFrames
    - Web scraping using requests and BeautifulSoup
    - Data preprocessing: OneHotEncoding, normalization, label creation
    - Classification models (Logistic Regression, SVM, KNN, Decision Tree)
    - GridSearchCV for hyperparameter tuning
- **SQL Queries**
- **Charts & Visuals**
    - Correlation heatmaps, bar plots, scatter plots
    - Folium map visualizations for geospatial data
    - Plotly Dash dashboard components: dropdowns, pie charts, scatter plots
- **Notebook Outputs**
    - Confusion matrices, Jaccard & F1 scores
    - Interactive components showcasing filtered launch success
- 🔗 **GitHub Repository:** Click Here

Thank you!