IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kin Fook, Lok
2024-12-19

Outline

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

Executive Summary

Summary of methodologies

- Data collection via API
- Data collection with Web Scraping
- Data wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Machine Learning Prediction

Summary of all results

- Success Rate over Time: An improvement in success rate was observed over time.
- Success Rate by Launch Site: KSC LC-39A shows the highest success rate of all launch sites.
- Success Rate by Orbit: ES-L1, SSO, HEO, and GEO were observed to have the highest success rates.
- Payload: Heavier payloads have a high failure rate early on with a significant improvement over time.
- Predictive Analysis: The DecisionTreeClassifier algorithm has proven highly accurate in predicting landing outcomes.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars.
- Other providers cost upward of 165 million dollars each.
- Much of the savings is due to SpaceX's ability to reuse the first stage.
- If we can determine if the first stage will land, we can determine the cost of a launch.
- Our company, IBM Developer Skills Network, can use this information to successfully bid against SpaceX for a rocket launch.

Introduction

The project endeavors to…

- Collect as much publicly available data as possible.
- Wrangle the data to improve its quality and prepare it for analysis.
- Explore the processed data with SQL and various visualizations.
- Drill down into finer levels of detail by splitting the data into groups defined by categorical variables.
- Build, evaluate, and refine predictive models for discovering additional insights.

Introduction

Problem Statement

- Predict the successful landing of SpaceX's Falcon 9 first stage, a critical factor in determining the cost-efficiency of reusable rocket launches.

Why we should solve this problem?

- Solving this problem allows IBM Developer Skills Network to make competitive bids against SpaceX by accurately estimating launch costs.

Questions

- What is the historical success rate of Falcon 9 first stage landings?
- What factors most significantly influence the success or failure of a landing?
- Can we develop a predictive model that accurately forecasts the outcome of a landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Retrieval and consolidation from multiple SpaceX API endpoints
  - Web scaping tabular data from Wikipedia
- Perform data wrangling
  - Extracted relevant records
  - Flattened fields and resolved missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualize variable relationships
  - Look at the data in aggregate
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mark all launch sites on a map
  - Mark successful and failed launches
  - Calculate distances to proximate locations
  - Provide for interactive exploration of the data
- Perform predictive analysis using classification models
  - Build, evaluate, and compare several predictive classification models

Data Collection

Data was collected using a combination of retrieval techniques:

1. HTTP requests against various SpaceX API endpoints:
   - Initial launch data was obtained from: /v4/launches/past
   - Additional data was backfilled from:
     - /v4/rockets
     - /v4/launchpads
     - /v4/payloads
     - /v4/cores
2. Tabular data from the List of Falcon 9 and Falcon Heavy launches Wikipedia page.

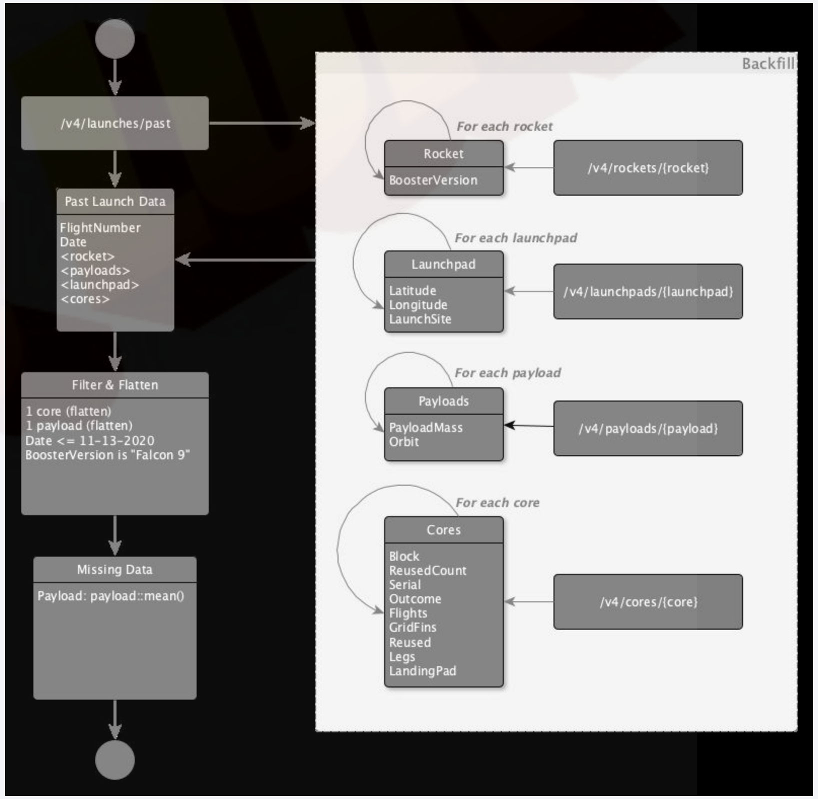
Data Collection – SpaceX API

The initial data was obtained from the `/v4/launches/past` API endpoint.

Additional data was backfilled from the `rocket`, `launchpad`, `payloads`, and `cores` API endpoints, for records with extant corresponding IDs.

GitHub URL:

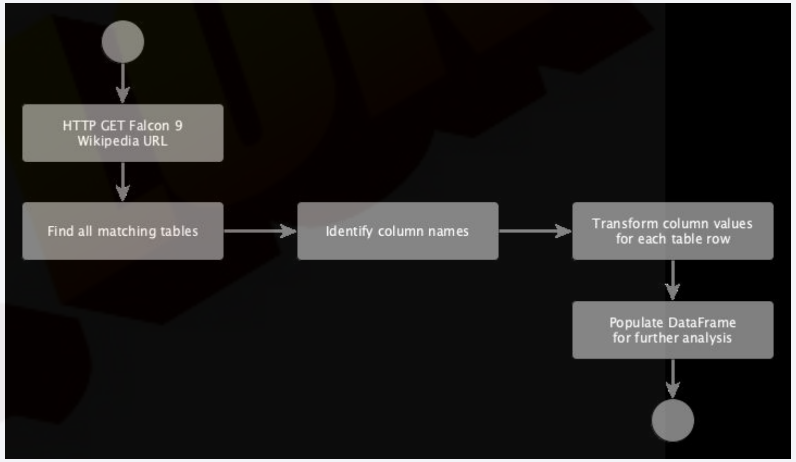
[IBM_Data_Science_Capstone/01 Hands-on Lab Complete the Data Collection API Lab.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]



Data Collection - Scraping

Web scraping workflow:

1. Send HTTP Get to Falcon 9 Launch page at Wikipedia, to retrieve the HTML source
2. Create a BeautifulSoup object to extract the tables containing launch data
3. Populate a Pandas DataFrame using the extracted columns



GitHub URL:

[IBM_Data_Science_Capstone/02 Hands-on Lab Complete the Data Collection with Web Scraping lab.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]

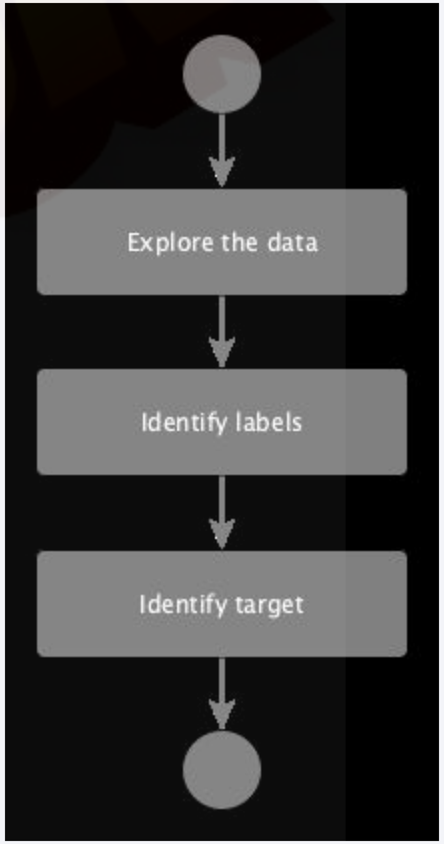
Data Wrangling

Goals:

- Use Exploratory Data Analysis (EDA) to find patterns in the data.
- Determine the labels for training supervised models.

Steps:

- Find missing values as a percentage of each attribute
- Identify column types, numerical or categorical
- Show launches per Launch Site
- Show distribution of Orbit type in the data set
- Explore outcomes, and group them by binary outcome (success or failure)
- Store outcome as "Class" label, to be used as target value in training



GitHub URL:

[IBM_Data_Science_Capstone/03 Hands-on Lab Data Wrangling.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]

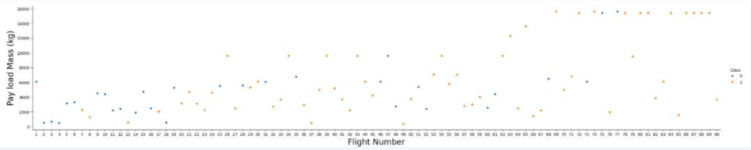
EDA with Data Visualization

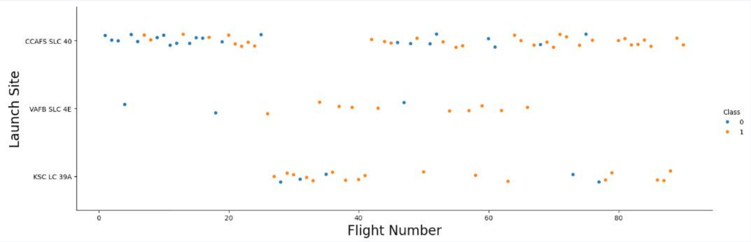
- Visualize relationships to gain insight into the importance of each variable:
  - Flight Number and Outcome
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Orbit and Outcome
  - Flight Number and Orbit
  - Payload and Orbit
  - Yearly success rate
- Expand categorical variables into "dummy" columns
- Convert numerical columns into float64

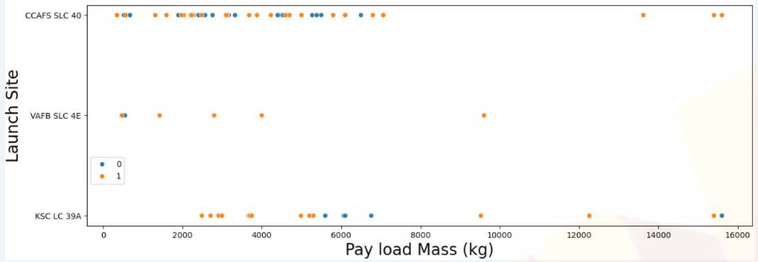
GitHub URL:

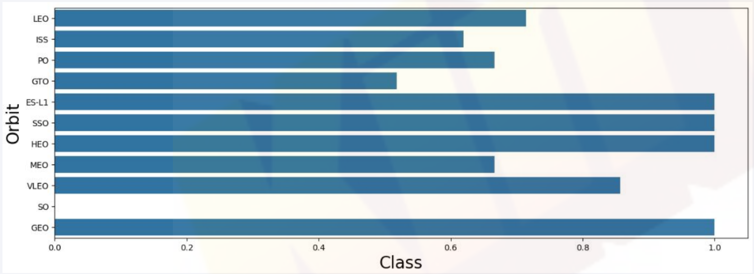
[IBM_Data_Science_Capstone/05 EDA with Visualization Lab.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]

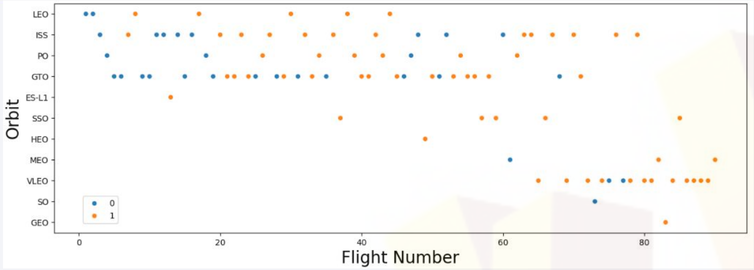
EDA with Data Visualization

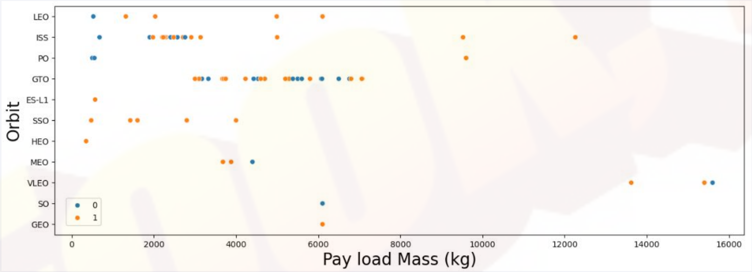


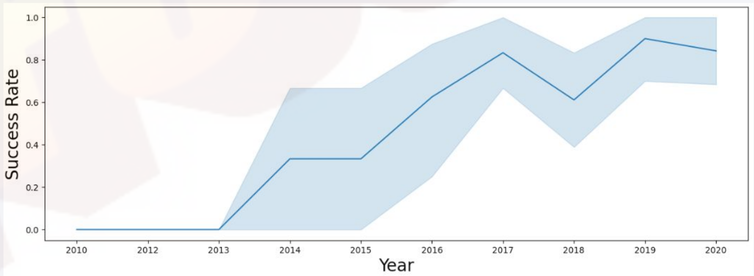












	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_I
0	1	6104.959412	1	False	False	False	1.0	0	False	False	...	False	
1	2	525.000000	1	False	False	False	1.0	0	False	False	...	False	
2	3	677.000000	1	False	False	False	1.0	0	False	False	...	False	
3	4	500.000000	1	False	False	False	1.0	0	False	False	...	False	
4	5	3170.000000	1	False	False	False	1.0	0	False	False	...	False	

EDA with SQL

SQL Queries Performed:

- Show each unique launch site
- Show 5 records where launch site names begin with 'CCA'
- Display the total payload mass carried by boosters launched by 'NASA (CRS)'
- Display the average payload mass carried by the v1.1 Falcon 9 booster
- List the date of the first successful ground landing outcome
- List the booster versions with successful outcomes landing on the drone ship with payloads between 4000kg and 6000kg.
- List the number of successful and failed mission outcomes
- List all of the booster versions that carried the max payload mass
- List the month name, outcome, booster version, and launch site for missions with failure outcomes landing on a drone ship in 2015.
- Show the distribution of outcomes between June 4th, 2010 and March 20th, 2017

GitHub URL:

IBM_Data_Science_Capstone/04 Hands-on Lab Complete the EDA with SQL.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub

Build an Interactive Map with Folium

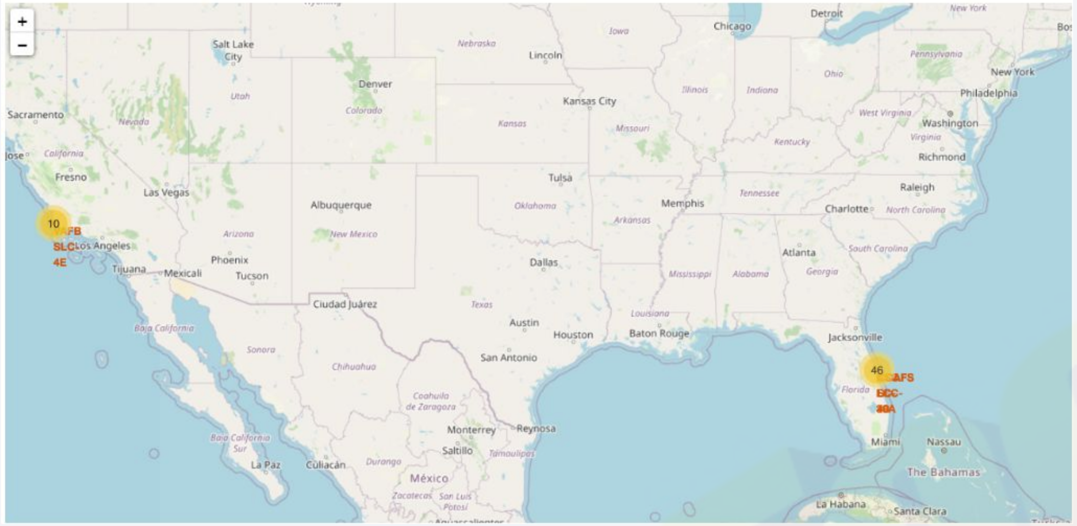
To find geographical patterns in the data the following items were marked on a map of launch sites:

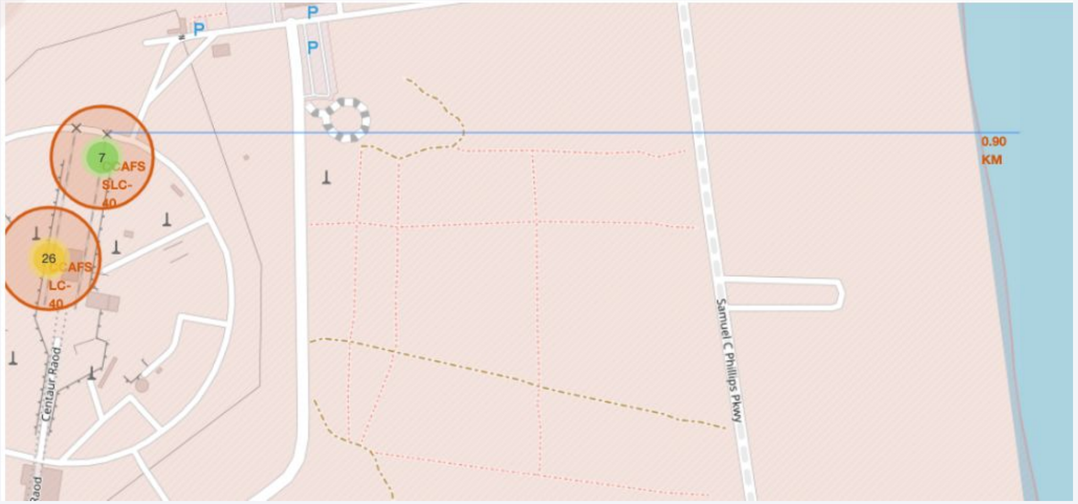
- All Launch Sites
- Successful and Failed Launches
- Distances between a launch site and proximate landmarks

GitHub URL:

IBM_Data_Science_Capstone/06 Hands-on Lab Interactive Visual Analytics with Folium lab.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub

Build an Interactive Map with Folium





Build a Dashboard with Plotly Dash

To enable interactive exploration of the data, a Plotly Dash dashboard was developed to include:

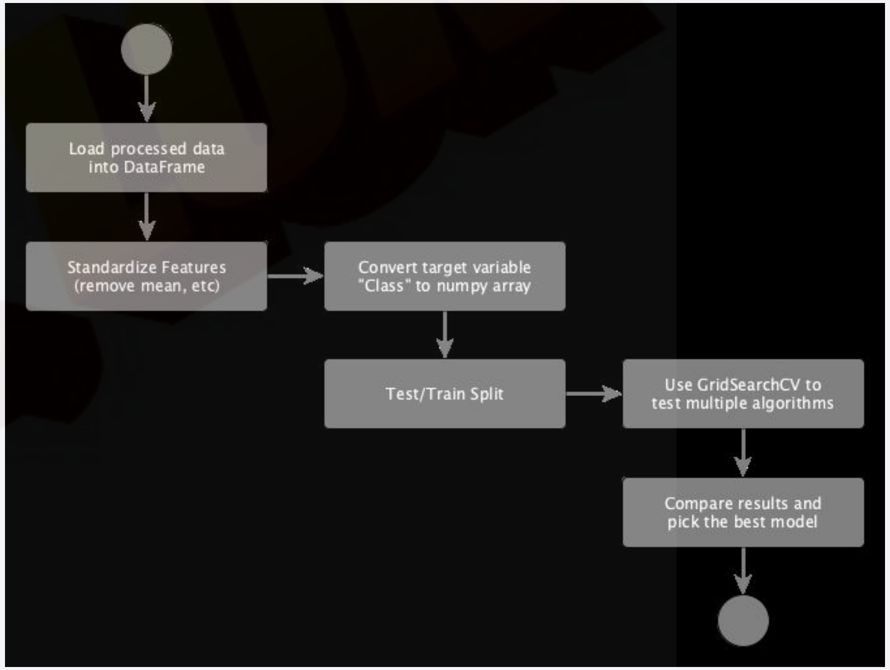
- A dropdown selector to choose a Launch Site, affecting:
  - A pie chart
    - All sites selected: shows the breakdown of successful outcomes across all sites
    - A launch site selected: shows the breakdown of successful vs failed launches for the given site
  - A scatter plot
    - All sites selected: shows outcome by payload mass and booster version for all sites
    - A launch site selected: shows outcome by payload mass and booster version for the given site
- A Payload Mass range selector that filters data points on the scatter plot

GitHub URL:

[IBM_Data_Science_Capstone/08 DASH.py at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]

Predictive Analysis (Classification)

1. Load data
2. Apply StandardizedScaler on X
3. Convert Y to numpy array
4. Split training and testing data
5. Use GridSearchCV to test hyperparameters for multiple algorithms:
   - Logistic Regression
   - SVC
   - Decision Tree Classifier
   - K Neighbors Classifier



GitHub URL:

[IBM_Data_Science_Capstone/07 Hands-on Lab Complete the Machine Learning Prediction lab.ipynb at main · kinfooklok/IBM_Data_Science_Capstone · GitHub]

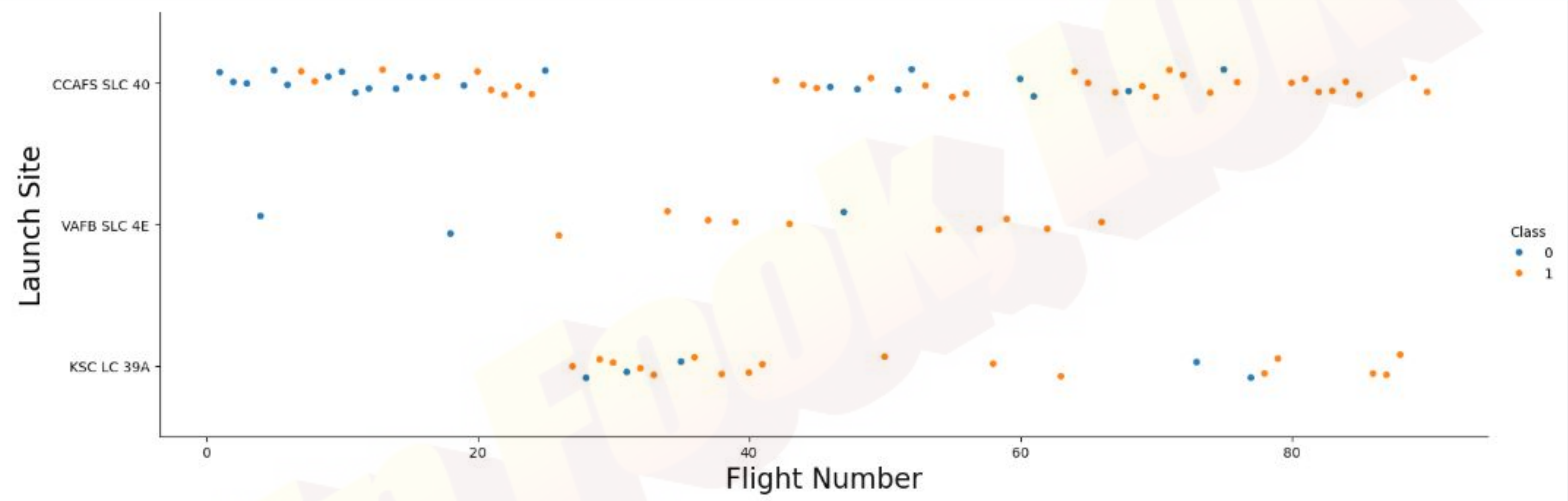
Results

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

Section 2

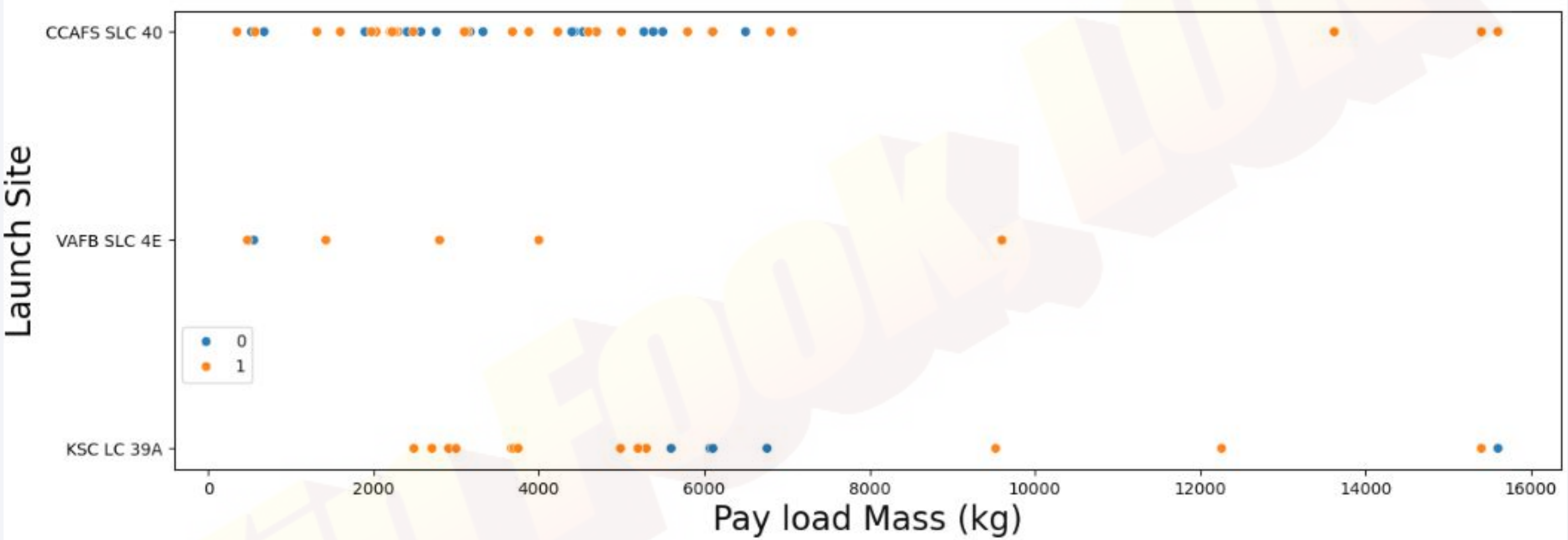
Insights drawn from EDA

Flight Number vs. Launch Site



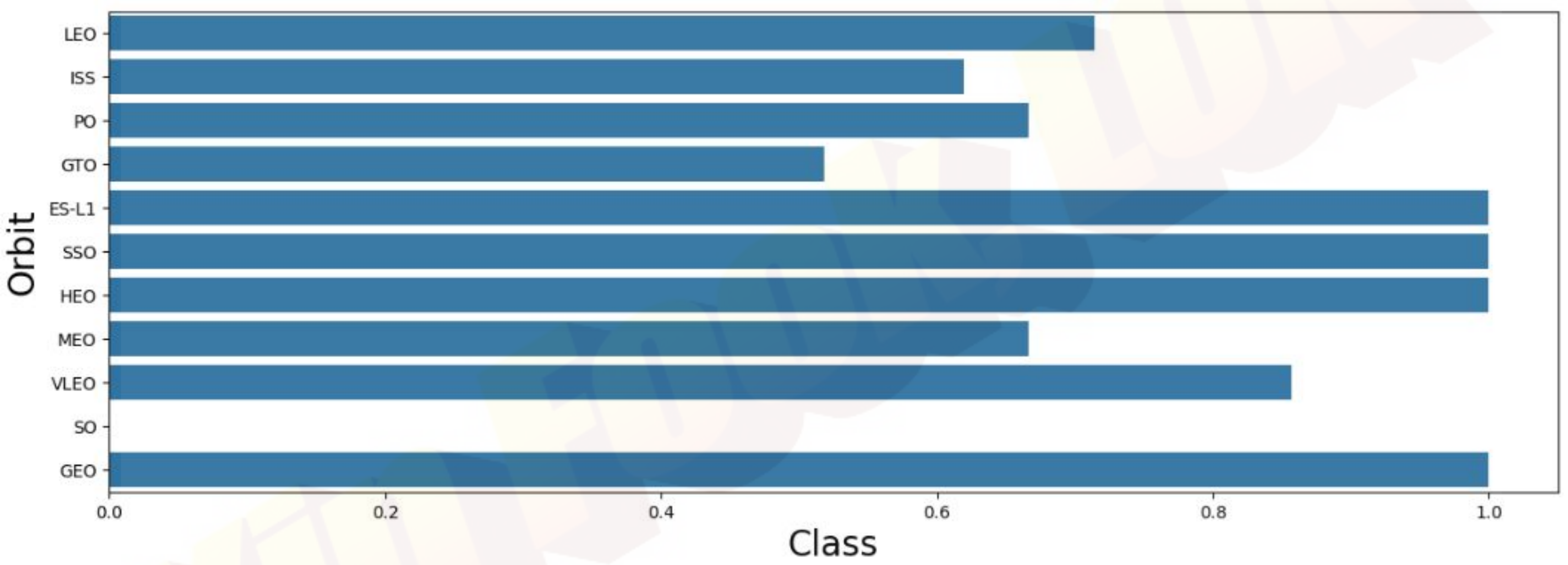
- All sites show a mix of first stage landing successes and failures, with successes increasing over time.
- Early flights predominantly resulted in failures, indicating improvements to technology or process.
- While CCAFS SLC 40 has the most total flights, VAFB SLC 4E appears to have a relatively higher proportion of successful landing outcomes.

Payload vs. Launch Site



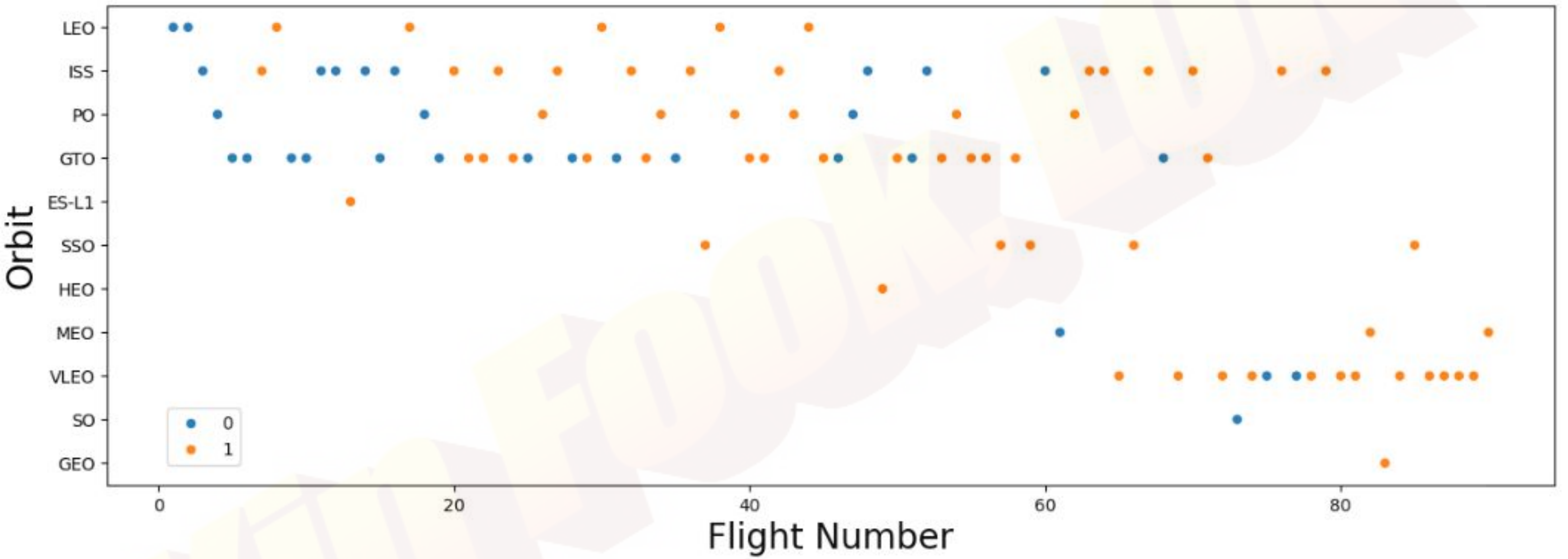
- All sites show a wide range of payload weights, from light to heavy payloads.
- Early flights trend toward lighter payloads, representing the bulk of the landing failures.
- This suggests technological or operational improvements lead to a greater rate of success with heavier payloads.

Success Rate vs. Orbit Type



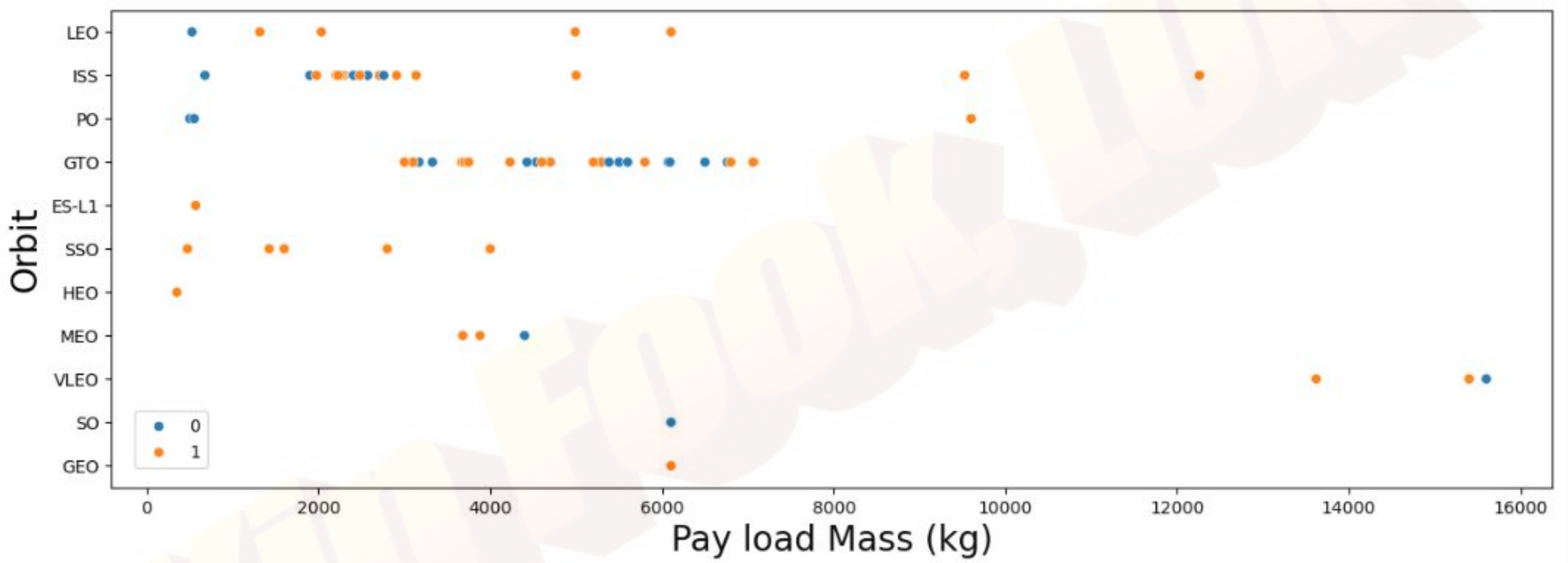
- Some orbits, such as ES-L1, SSO, HEO, and GEO consistently show high success rates.
- Others such as GTO show more mixed outcomes, suggesting some orbit types may introduce operational or technological challenges.
- With only one launch, there is not enough data for the SO orbit type to provide an accurate analysis.

Flight Number vs. Orbit Type



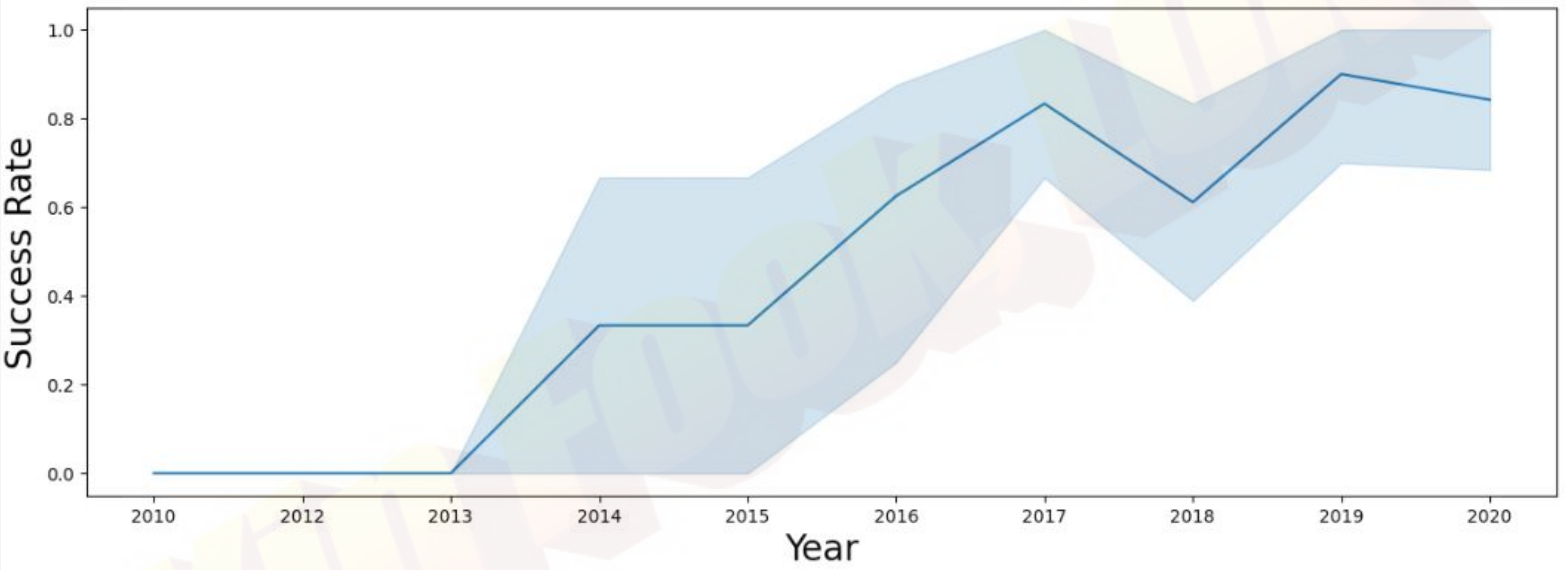
- Many orbits are represented throughout the flight number range, some orbits are not attempted until later flights.
- There is a noticeable improvement in landing success as flight numbers increase, indicating an accumulation of experience and ongoing improvements.

Payload vs. Orbit Type



- Many orbits are represented across a wide range of payload masses, but others like SSO, MEO, HEO and GEO show a generally lower range.
- Orbits with a constrained payload range, tend to show a higher rate of landing success.
- While payload mass does not appear to directly determine mission success, its interplay with orbit suggests a significant correlation.

Launch Success Yearly Trend



- Yearly trend shows a consistent progression from early challenges to high reliability in first stage landings over time.
- From 2016 onward, SpaceX experienced year over year improvement in success rate with a minor setback in 2018.

All Launch Site Names

There are four unique Launch Sites

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

SELECT DISTINCT Launch_Site from SPACEXTABLE;

Launch Site Names Begin with 'CCA'

	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

*SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;*

Total Payload Mass

The total payload carried by boosters from NASA (CRS) is
45, 596 kg.

SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is
2, 534.67 kg.

SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';

First Successful Ground Landing Date

The first successful landing outcome on ground pad occurred on **December 22nd, 2015.**

```
SELECT MIN(Date) as LaunchDate
    FROM SPACEXTABLE
        WHERE Landing_Outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

BOOSTER	PAYLOAD MASS
F9 FT B1022	4, 696 KG
F9 FT B1026	4, 600 KG
F9 FT B1021.2	5,300 KG
F9 FT V1031.2	5, 200 KG

```
SELECT Booster_Version, PAYLOAD_MASS__KG_
    FROM SPACEXTABLE
        WHERE Landing_Outcome = 'Success (drone ship)'
            AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

MISSION STATUS	COUNT
FAILURE	1
SUCCESS	100

```
SELECT CASE
	WHEN Mission_Outcome LIKE 'Success%' THEN 'Success'
	WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
END as Mission_Status, COUNT(*)
	FROM SPACEXTABLE
		GROUP BY Mission_Status;
```

Boosters Carried Maximum Payload

The maximum payload sent was **15, 600 kg.**
The boosters that carried the maximum payload are:

- 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

```
SELECT
    DISTINCT Booster_Version,
    PAYLOAD_MASS__KG_
FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (
        SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
ORDER BY Booster_Version;
```

2015 Launch Records

List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

MONTH	OUTCOME	BOOSTER	LAUNCH SITE
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
SELECT
    CASE strftime('%m', Date)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as Month,
    Landing_Outcome, Booster_Version, Launch_Site, Date
FROM SPACEXTABLE
WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

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

```
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;
```

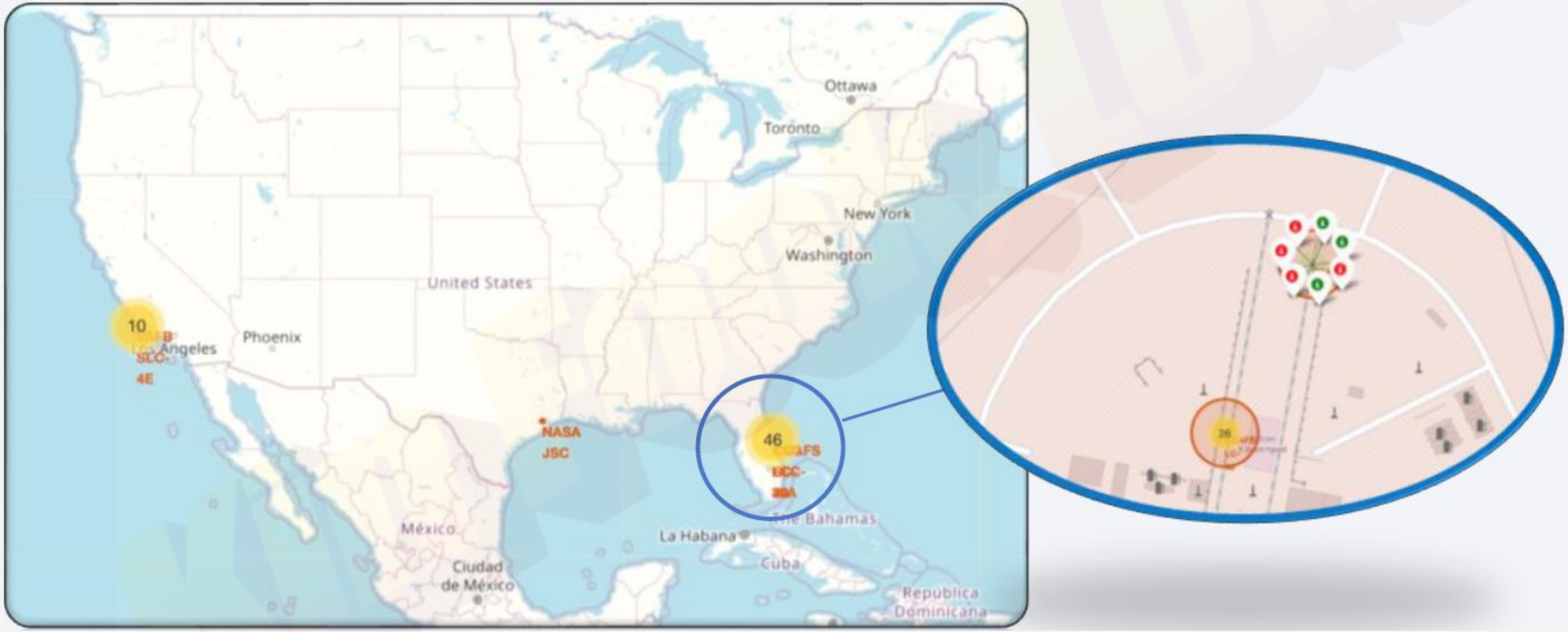
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

Launch sites are located near coastal regions in Florida and California to reduce risk of catastrophic failures affecting human activities.

Launch Outcomes



Notable Proximate Locations

Notable locations shown

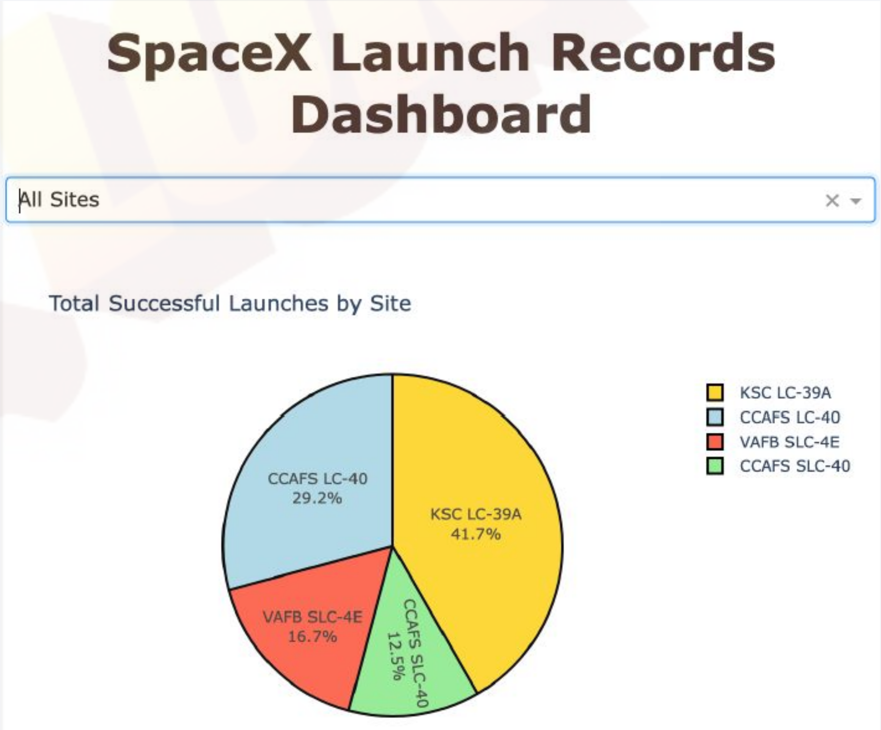
1. Railway (1.02KM)
2. Roadway (0.59KM)
3. Coast (0.86KM)

Section 4

Build a Dashboard with Plotly Dash

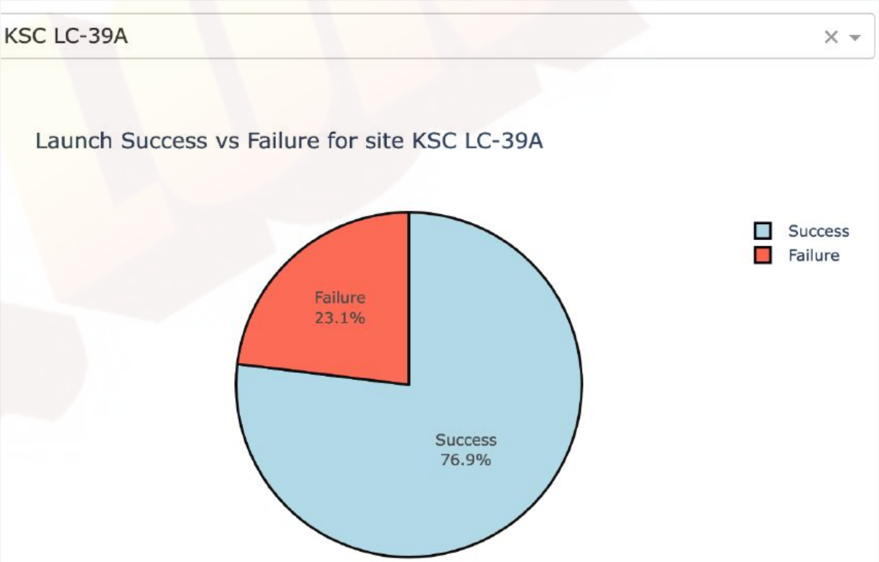
All Launch Sites: Successful Landings

- KSC LC-39A experienced the highest proportion of successful landings, followed by CCAFS LC-40.
- VAFB SLC-4E and CCAFS SLC-40 the lowest.



Per-site Launch Success Ratio: High

- KSC LC-39A had the highest ratio of successful landings



Payload Range

- With a payload mass between 3, 000 kg and 5, 000 kg, v1.1 boosters performed the worst.
- In the same payload range, B4 and B5 boosters had the best success rate, followed by FT.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Of the algorithms tested, the DecisionTreeClassifier was the most accurate.

Confusion Matrix

- 14 observations were correctly predicted as successful landings (true positive)
- 3 observations were correctly predicted as failed landings (true negative)
- 1 observation was incorrectly predicted as a successful landing (false positive)
- No observations were incorrectly predicted as a failed landing (false negative)

Conclusions

- Success rates increase over time, across all factors, which indicates continuous and incremental operational improvements and technological advancements.

- Different orbits have varying success rates, with ES-L1, SSO, HEO, and GEO showing consistently successful outcomes.

- Launch site was a highly predictive factor, with KSC LC-39A being a top performer, closely followed by CCAFS LC-40.

- Many of the predictive models evaluated were able to predict landing outcome with an acceptable level of accuracy. In the testing performed, DecisionTreeClassifier produced best results with high accuracy, precision, and recall.

Appendix

Data Sources

SpaceX API	r/SpaceX API Docs
Collected Data	IBM_Data_Science_Capstone/01 Hands-on Lab Complete the Data Collection API Lab.csv at main · kinfooklok/IBM_Data_Science_Capstone · GitHub
Wikipedia: List of Falcon 9 and Falcon Heavy Launches (June 2021)	List of Falcon 9 and Falcon Heavy launches - Wikipedia
Data after wrangling	IBM_Data_Science_Capstone/03 Hands-on Lab Data Wrangling.csv at main · kinfooklok/IBM_Data_Science_Capstone · GitHub

Thank you!