IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lyn Pratt

October 03, 2024

Outline

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

Executive Summary

Summary of methodologies:
This study utilized public datasets to analyze the factors contributing to successful space launches. Data preprocessing included cleaning, standardizing the data, and feature engineering. Multiple machine learning models were trained and evaluated, including logistic regression, support vector machines, decision trees, and k nearest neighbour. Hyperparameter tuning was performed using grid search cross-validation to optimize each model.

Summary of all results:
The analysis revealed that factors such as launch site location, payload mass, and orbit type influence launch success. Moreover, the launch success rate was seen to increase over time. The best-performing model, classification tree, achieved an accuracy of 94.4% on the test data. These findings suggest that data science can play a critical role in predicting launch success. Future work will explore the use of more advanced models and incorporate real-time data for improved prediction success.

Introduction

Project Background and Context:

- SpaceY is looking to compete with SpaceX
- SpaceX generates and shares detailed information about the success of their missions
- This project attempts to use data science to optimize launch success by analyzing this historical data

Problems/Questions to Answers:

1. What are the most significant factors that contribute to launch success?
   - essential for improving launch reliability and mitigating risks.

2. Can we build a predictive model to forecast the probability of successful launches based on historical data?
   - helps make informed decisions about launch planning and risk assessment.

Section 1

Methodology

Methodology - Executive Summary

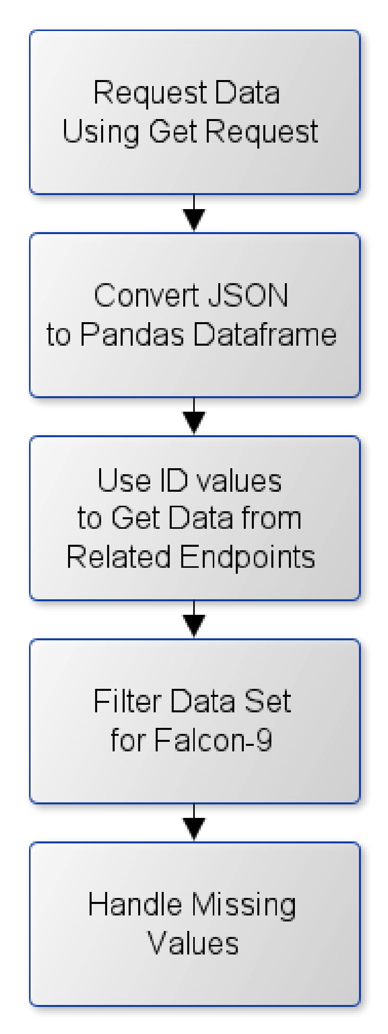
- Data collection methodology:
  - Data sources: Publicly available information, REST APIs and a Web Page
- Perform data wrangling
  - Replaced missing values,
  - Ensured data types were consistent, and standardized units of measurement.
  - Encoded Categorical Variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected Models to Evaluate: Logistic Regression, Support Vector Machines (SVM), Decision Trees, K Nearest Neighbours
  - Data was standardized and divided into training and test data sets.
  - Performed Hyperparameter Optimization using Gridsearch. Training data was used to train the models.
  - Evaluated each model's performance.
  - The best performing model was selected

Data Collection

- **Data Sources :**
  - Publicly available SpaceX REST API
  - Publicly available Web Page
- **Data Collection Techniques**:
  - REST API requests: Using Python's requests and JSON
  - Web Scraping: Using Beautiful Soup.

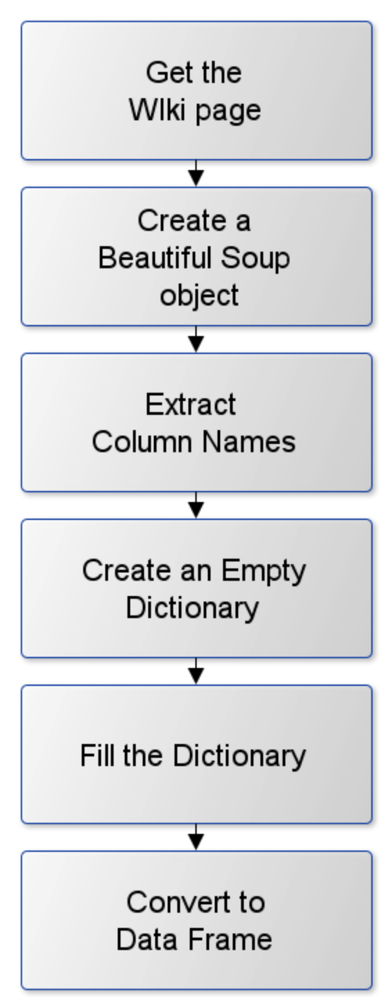
Data Collection – SpaceX API

- Request the SpaceX launch data using the GET request
- Decode the response content as a Json using .json()
- Turn it into a Pandas data frame using .json_normalize()
- Use ID values to request related information
    - From the rocket get booster name
    - From the payload get mass of the payload and the orbit
    - From the launchpad get the launch site being used, the longitude, and the latitude.
    - From cores get the outcome of the landing
- Construct a data frame and filter it to include only Falcon 9 launches
- Handle for missing values
    - For Payload Mass, convert missing values to the mean
- Save the data frame as a csv file
- https://github.com/lyn797/ds-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



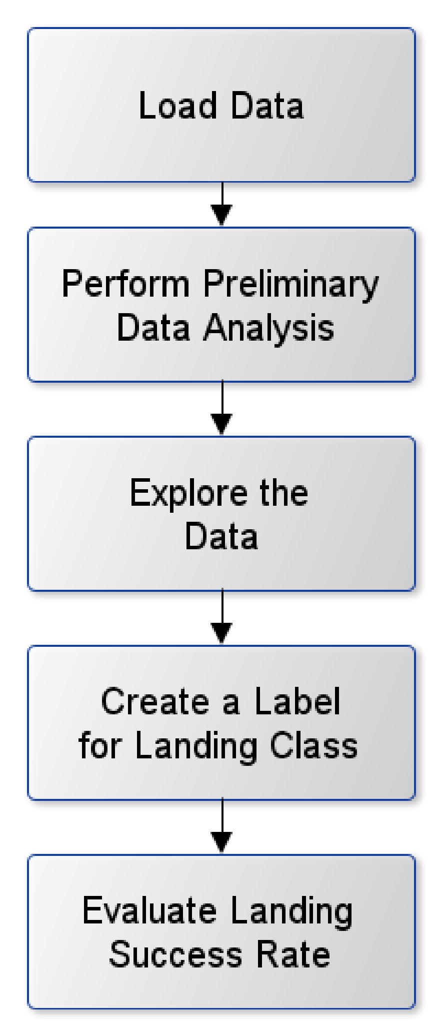
Data Collection - Scraping

- Use a get request to retrieve the Wiki page
- Create a Beautiful Soup object from the response
- Extract column names from the first data table
- Create an empty dictionary with keys from the extracted column names
- fill up the dictionary with launch records from all data tables
    - Handle unexpected annotations and other types of noises
- Convert the filled dictionary to a data frame
- Save the data frame as a csv file
- https://github.com/lyn797/ds-capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Read the SpaceX launch data from a CSV file.
- Check for missing values and identify data types of columns.
- Explore the Data
  - Analyze launch sites to understand launch frequency.
  - Analyze orbits to understand mission types.
  - Analyze landing outcomes to understand success and failure rates.
  - Create a set, bad_outcomes, containing outcomes where the landing was unsuccessful
- Create a new column ("Class") to label successful (1) and unsuccessful (0) landing outcomes using bad_comes as a reference.
- Determine the overall success rate of landings.
- Save the processed data.
- https://github.com/lyn797/ds-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



Load Data

Perform Preliminary Data Analysis

Explore the Data

Create a Label for Landing Class

Evaluate Landing Success Rate

EDA with Data Visualization

- Flight Number vs. Payload Mass Scatter Plot:
  - Shows trends in landing success based on launch attempts and payload weight.
- Flight Number vs. Launch Site Bar Chart:
  - Compares launch site activity and success rates across different sites.
- Payload Mass vs. Launch Site Scatter Plot:
  - launch site choice is related to payload weight and landing success.
- Success Rate by Orbit Bar Chart:
  - Compares success rates across different orbit types (e.g., LEO, GTO).
- Flight Number vs. Orbit Scatter Plot:
  - Checks if success rate varies with flight number for different orbit types.
- Payload Mass vs. Orbit Scatter Plot:
  - Analyzes the relationship between payload weight and orbit type to see if it impacts landing success.
- Launch Success Yearly Trend Line Chart:
  - Tracks the overall success rate of landings over time.
- https://github.com/lyn797/ds-capstone/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Find all the unique launch sites used in the dataset.
  - Show the first 5 launch records from sites starting with "CCA".
  - Calculate the total payload mass launched by NASA (CRS).
  - Find the average payload mass of rockets using the "F9 v1.1" booster.
  - Show the date of the first successful landing for each type of landing outcome.
  - List booster versions that successfully landed on drone ships with payloads between 4000 and 6000 kg.
  - Count the number of successful and failed missions.
  - Find the booster versions that have carried the heaviest payload.
  - Identify the month and landing outcome for all drone ship landing failures in 2015.
  - Rank the number of different landing outcomes between June 4th, 2010, and March 20th, 2017, from highest to lowest.
- https://github.com/lyn797/ds-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Map Objects:

1. **Markers**: Pinpoint locations (e.g., launch sites, cities) on the map.
2. **Circles**: Highlight areas or zones around points (e.g., a buffer around a launch site).
3. **PolyLines**: Connect points to show distances (e.g., a line between a launch site and the coastline).
4. **MousePosition**: Displays the coordinates of the cursor's position on the map, helping users find the location of proximities
5. **MarkerCluster**: Groups closely spaced markers together to reduce clutter and improve map readability.

Why They Are Used:

- **Visualizing Launch Sites:** Markers help show where launches have taken place.
- **Highlighting Success:** Color-coded markers help identify launch sites with high success rates.
- **Finding Proximity:** Mouse position helps locate nearby features (coastlines, cities, etc.).
- **Showing Relationships:** Lines connect launch sites to features to visually show proximity and possible connections.

- https://github.com/lyn797/ds-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dashboard Components

1. **Launch Site Dropdown:** Allows users to choose a specific launch site or view data for all sites.
2. **Success Pie Chart:** Shows the proportion of successful vs. failed launches for the chosen site (or all sites).
3. **Payload Range Slider:** Lets users filter the data by selecting a range of payload masses.
4. **Success vs. Payload Scatter Chart:** Displays the relationship between launch success, payload mass and booster version, allowing users to filter by site and payload range.

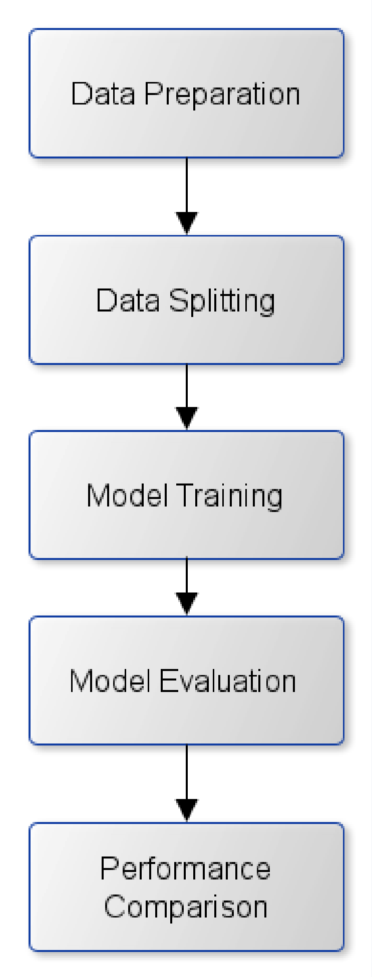
This dashboard helps users analyze SpaceX launch data interactively. They can:

- Focus on specific sites and payload ranges.
- See how launch success relates to launch site and payload mass.
- Identify potential patterns and correlations that might influence launch outcomes.

https://github.com/lyn797/ds-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The 'Class' column was converted to a NumPy array, features were standardized.
- The data is split into training and testing sets.
- Four models were trained using GridSearchCV for hyperparameter optimization:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- The accuracy of each model was assessed on both the training data (cross-validation) and the test data.
- Confusion matrices were used to analyze the types of errors each model makes.
- A bar chart visualizes the performance of the models was created
- https://github.com/lyn797/ds-capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb



Data Preparation

Data Splitting

Model Training

Model Evaluation

Performance Comparison

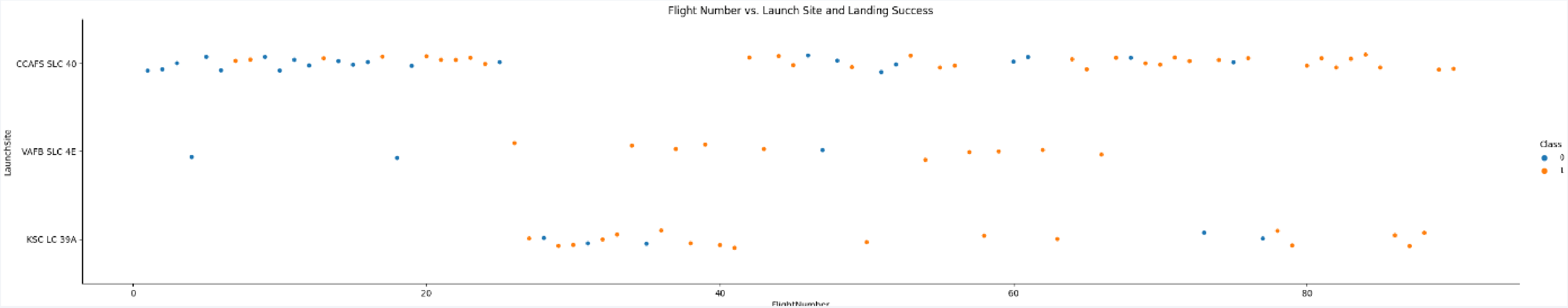
Results

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

Section 2

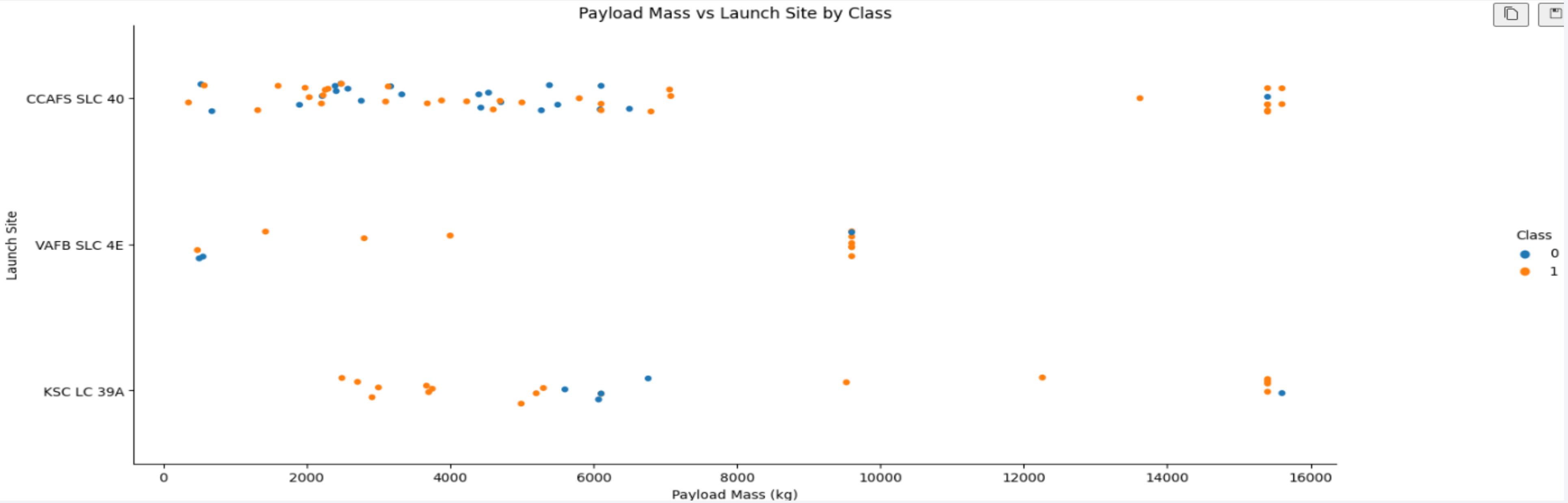
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40: The most frequent launch site.
- VAFB SLC 4E: A smaller number of launches compared to CCAFS SLC 40.
- CCAFS SLC 40: Early launches show a higher frequency of unsuccessful landings. Later launches appear to have a better landing success rate.
- The plot shows that successful landings appear more frequently in later flight numbers.

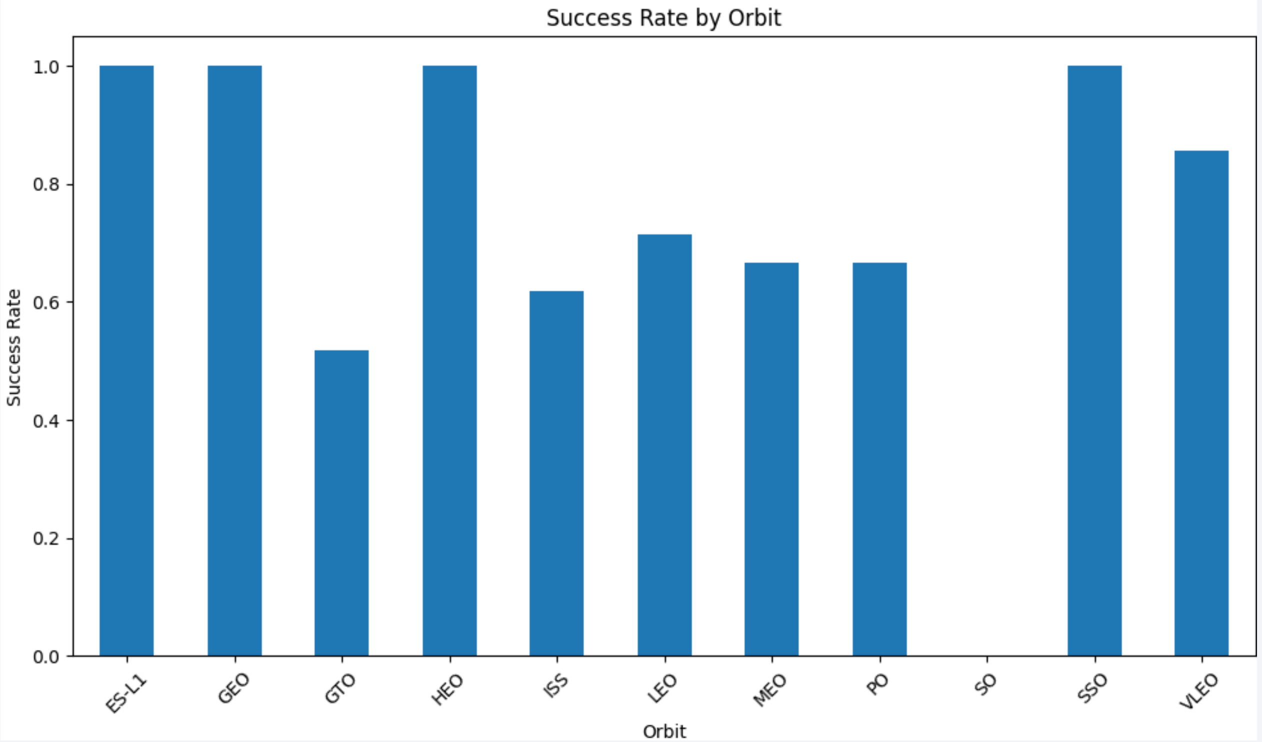
Payload vs. Launch Site



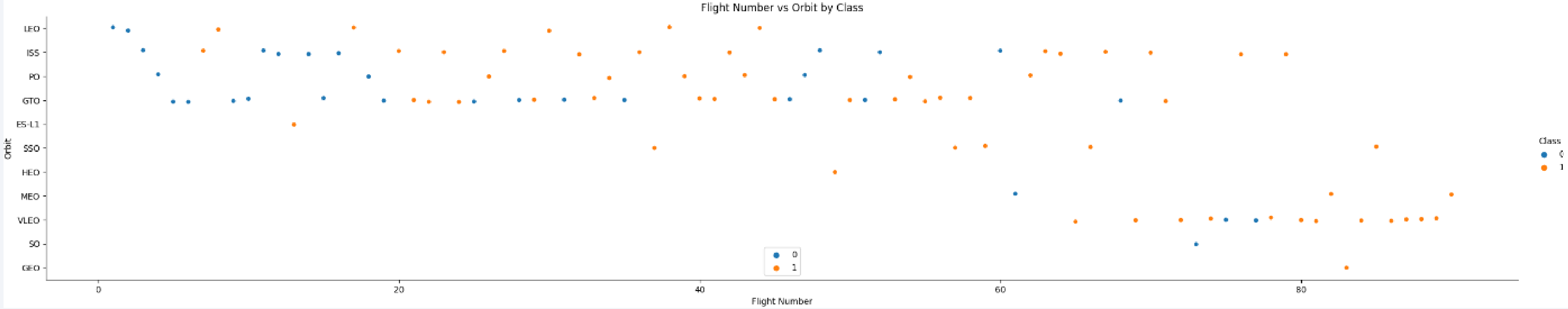
- The plot shows that payload mass varies, with most launches falling in the range of 0 to 6,000 kg.
- Payload Mass:
  - CCAFS SLC 40: has the widest range of payload masses, both lighter and heavier payloads.
  - VAFB SLC 4E: Primarily used for launches with smaller payloads,
- Landing Success:
  - Fewer Heavy loads were launched; however, the success rate appears higher than for lighter loads.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO, and VLEO have a consistently high success rate, nearing or exceeding 90%. This suggests that SpaceX has achieved a high level of reliability in launching to these specific orbits.
- GTO has a significantly lower success rate than the other orbits, suggesting a higher risk associated with launching to this orbit.
- Orbits ISS, LEO, MEO, and PO show moderate success rates, ranging from 60% to 70%. This indicates a reasonable success rate for these orbital destinations



Flight Number vs. Orbit Type



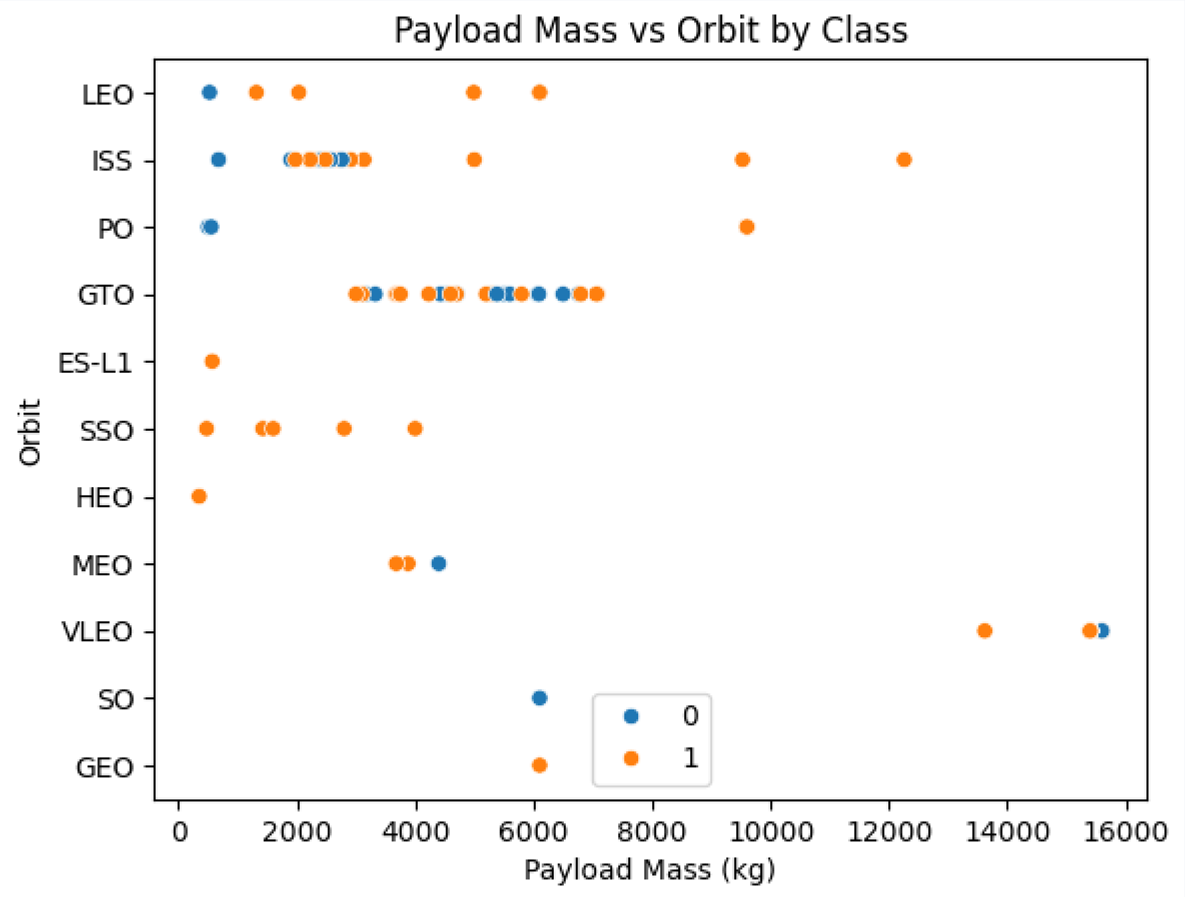
- LEO: noticeable trend towards more successful landings in later flight numbers.
- SSA: high success rate. All missions successful
- GTO: there appears to be no relationship between flight number and success
- VLEO: A significant number of successful landings are observed in later flight numbers

Payload vs. Orbit Type

- GTO : both successful and unsuccessful landings
- Polar, LEO, and ISS: Success improves with heavier payloads

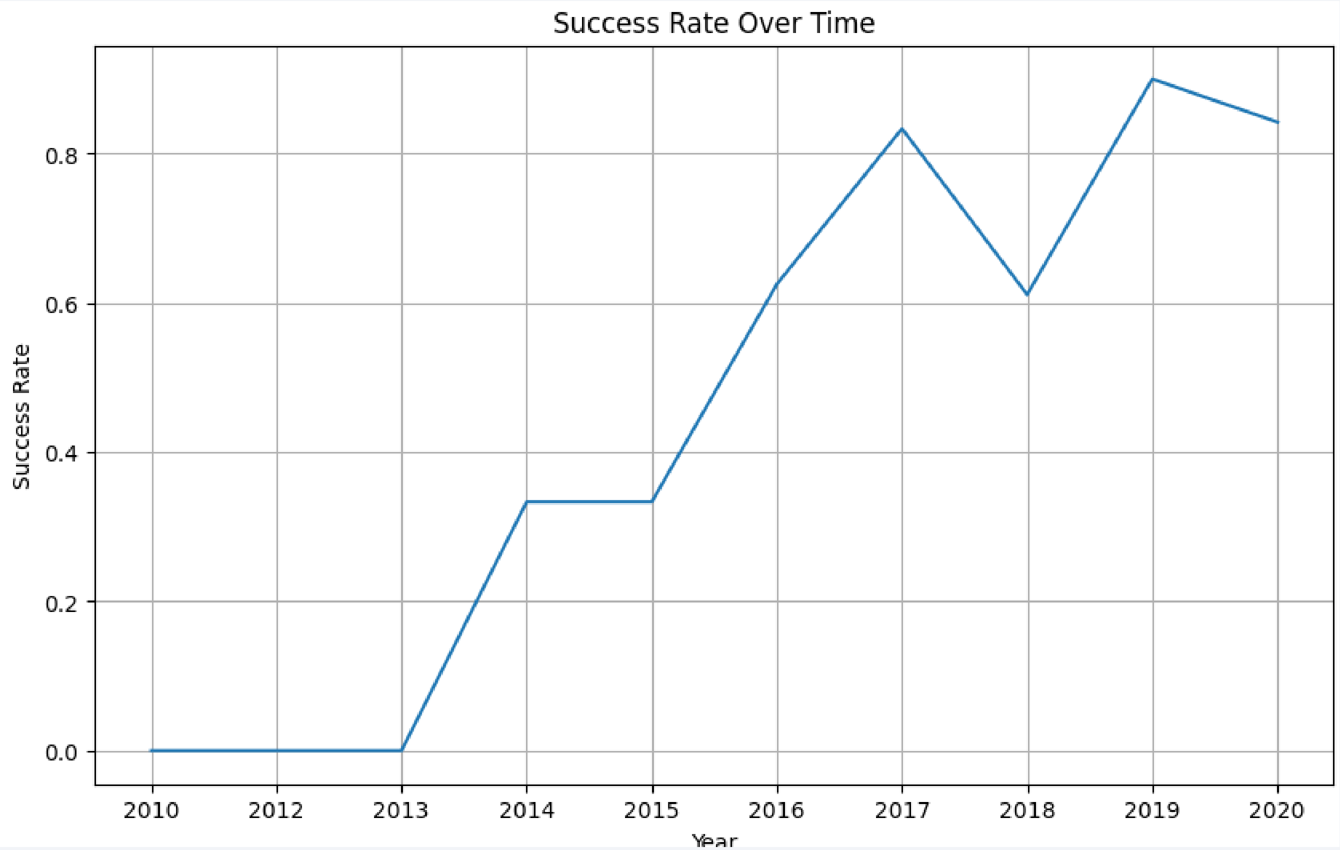
Insights:

- The plot suggests a connection between payload mass and the choice of orbit. For example, GTO missions tend to have lighter payloads compared to VLEO.
- The data highlights that landing success can vary across different orbits. It seems that landing success is not solely determined by payload mass



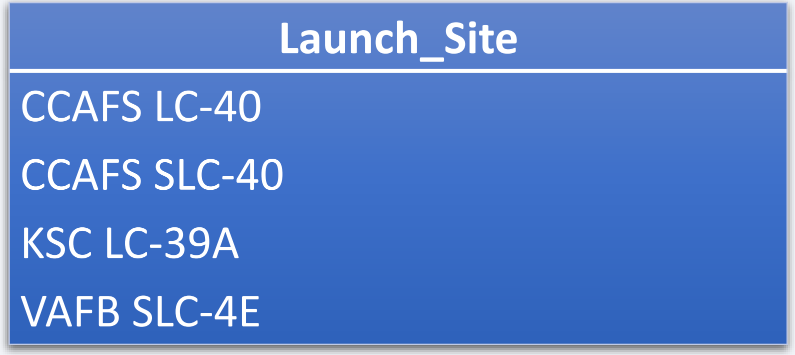
Launch Success Yearly Trend

- The success rate has steadily increased over time, demonstrating a commitment to improving launch reliability.
- Failures in the initial years likely led to improvements in the design, manufacturing, and operational procedures.
- The slight dip in 2018 suggests that SpaceX is continuously innovating and enhancing the Falcon 9's performance.



All Launch Site Names

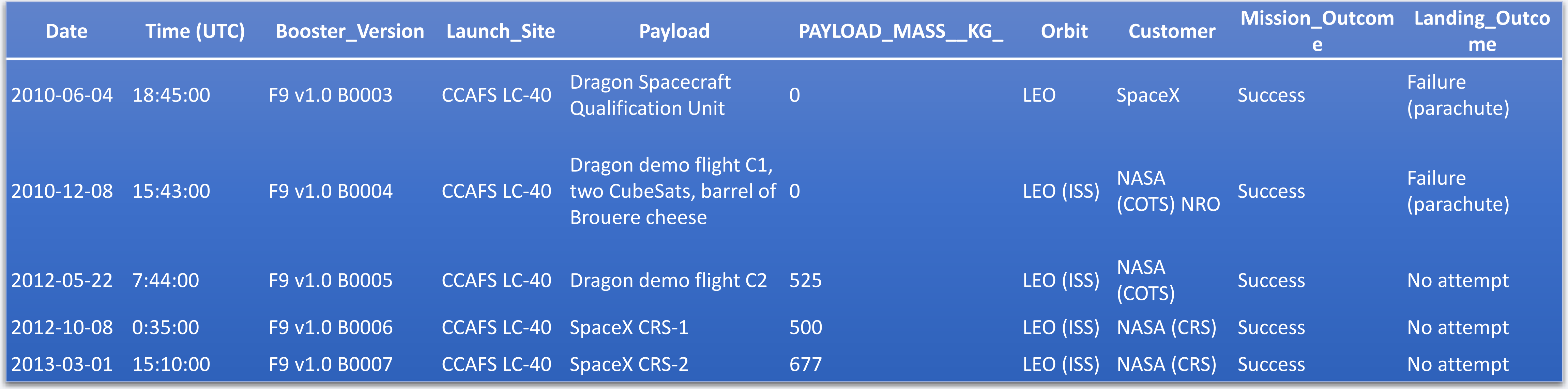
- Selecting the unique occurrences of "launch_site" (group by) from the dataset returns 4 unique launch sites



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

Launch Site Names Begin with 'CCA'

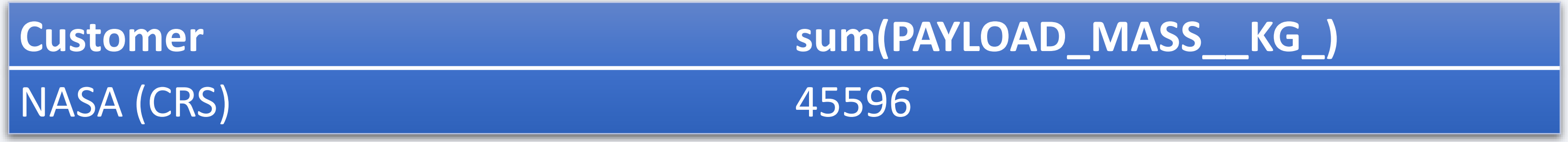
- The following table displays 5 records where launch sites begin with `CCA`
- It is restricted to 5 records by the limit clause



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

- Calculate the total payload carried by boosters from NASA
- This query groups the data by customer (group by), sums (sum) the payload mass for each customer and filters the results (having) to include only the desired customer, NASA (CRS).



Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The query calculates the average (avg) payload mass (in kilograms) for all space launches (where) that used the 'F9 v1.1'

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Landing_Outcome	min(date)
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- This query filters the results to include only launches where:
    1. The Landing_Outcome is 'Success (drone ship)'
    2. The PAYLOAD_MASS__KG_ is between 4000kg and 6000kg.

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

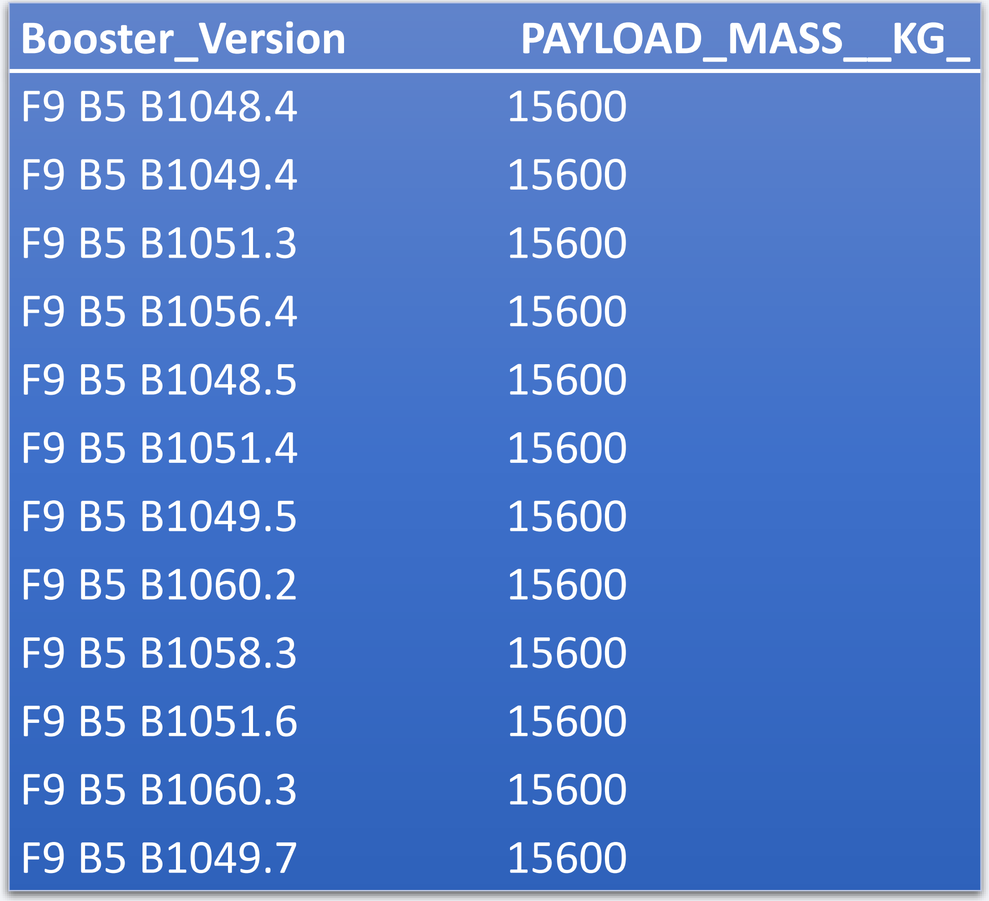
Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- The data set was grouped by mission outcome. The first attempt showed that there were 4 mission outcomes: Failure (in flight), Success, Success [with a non-printing character], and Success (payload status unclear)
- To correct only the first 7 digits were evaluated. This group all 3 “Success” records together

substr(mission_Outcome,0,8)	count(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

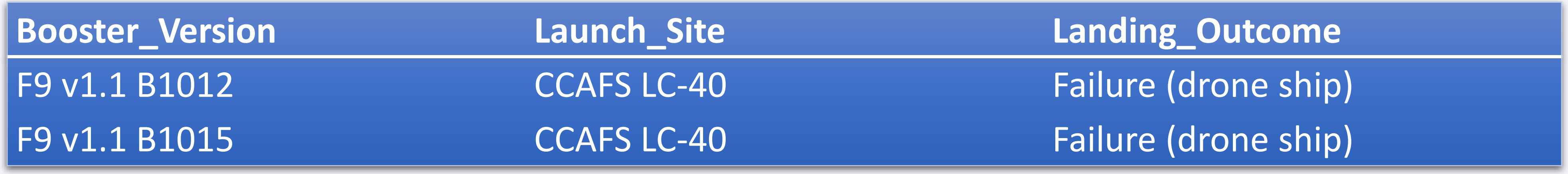
- List the names of the booster which have carried the maximum payload mass
- The query first finds the maximum payload mass value in the table using a subquery. Then, it retrieves the Booster_Version and PAYLOAD_MASS__KG_ for the launch(es) that have this maximum payload mass.



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

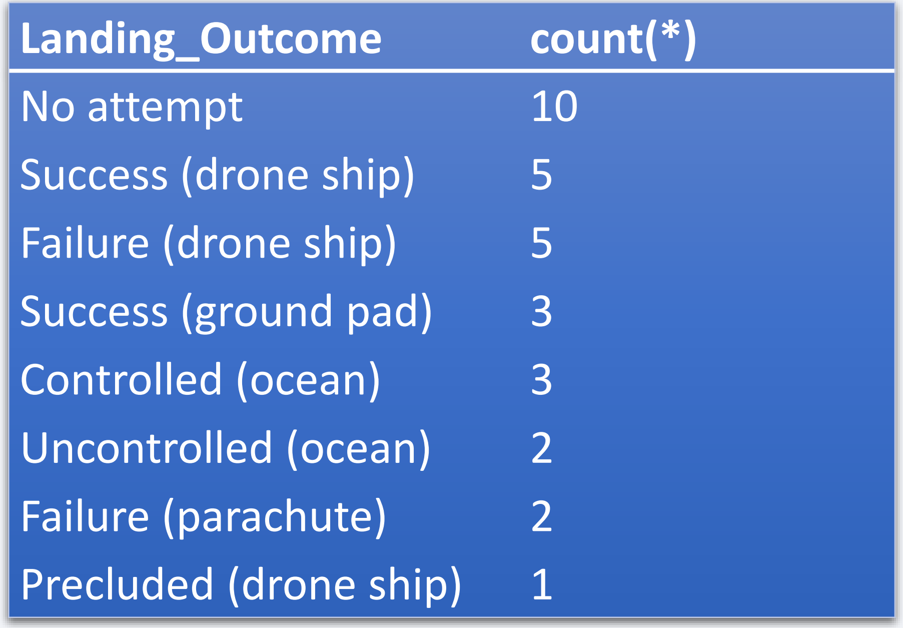
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There were only 2 launches in 2015 that failed in drone ship.



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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- This query highlights the significance of “No attempt”



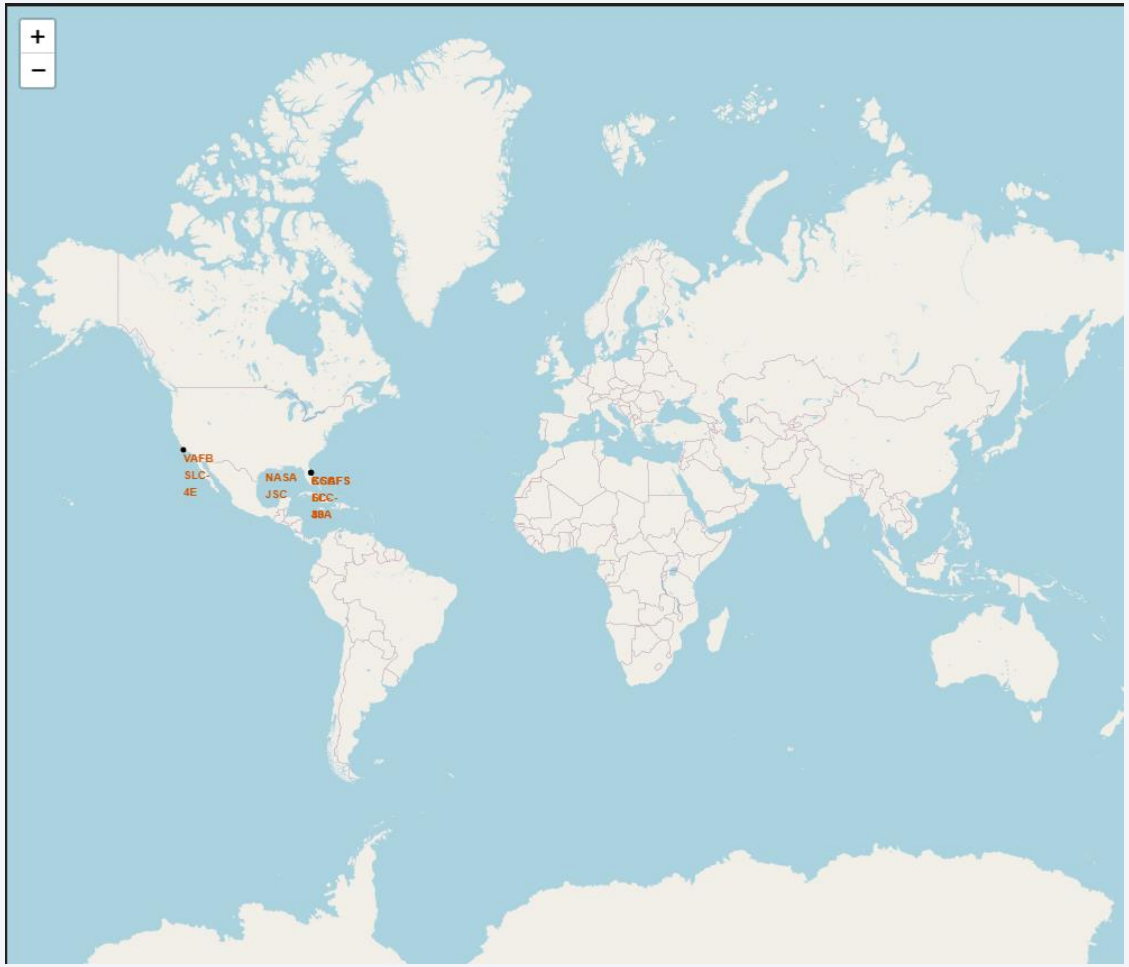
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

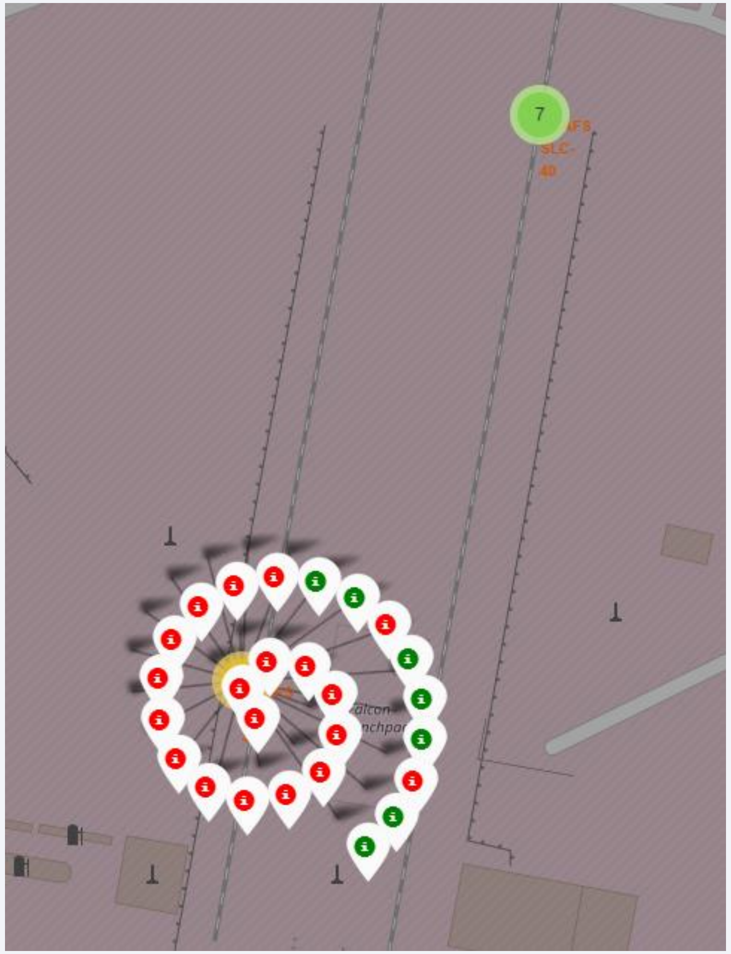
Launch Sites from a Global Perspective

- All launch sites are:
  - In the southern U.S.
  - Near a Coast



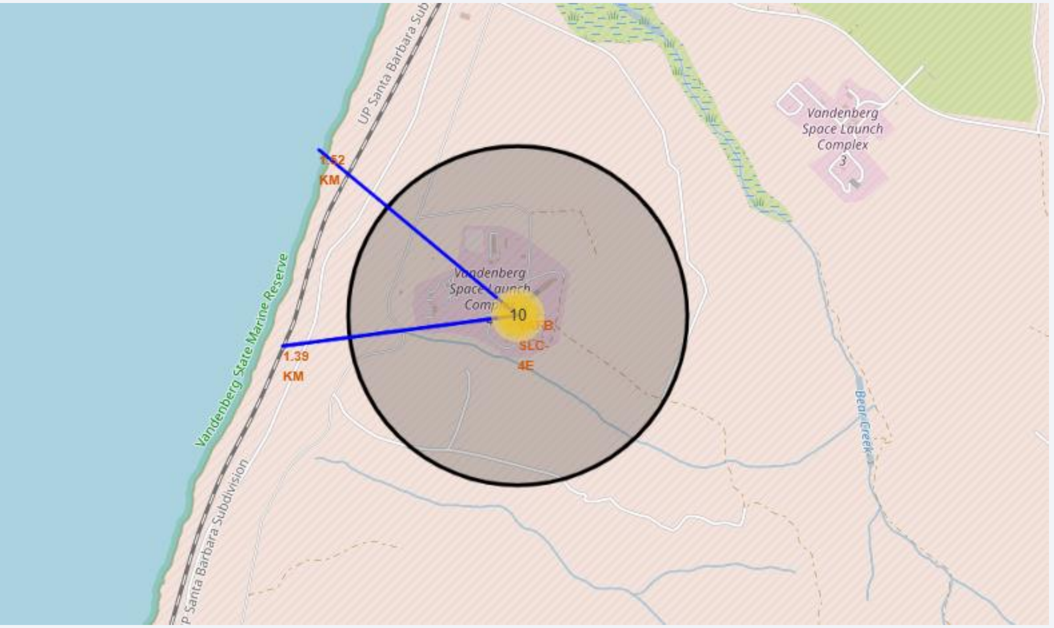
Launch Outcomes for CCAFS LC-40

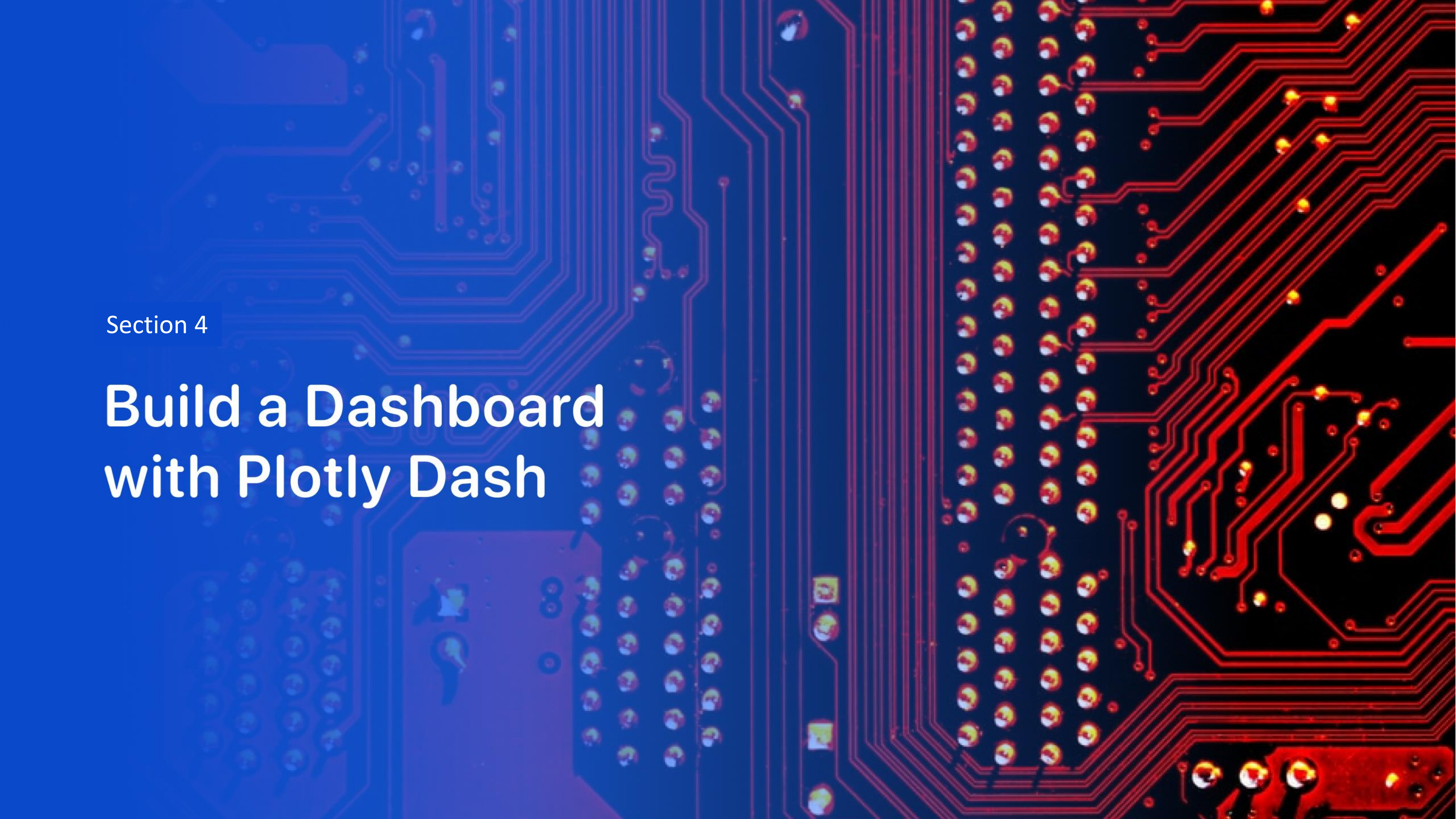
- This map shows the launch outcomes for CCAFS LC 40
- There are more red indicators than green. This visually depicts that more launches from this site have been unsuccessful than successful
- Notice the green circle at CCASF SLC 40 with a 7 in the centre. This indicates that there are 7 outcomes for this site



VAFB SLC-4E Proximities

- VAFB SLC-4E is located within 2 kilometers from the coast and a rail line
- This screen shot also shows that is a similar distance to a road


UP Santa Barbara Sub
1.52
KM
Vandenberg Space Launch Complex 3
Vandenberg State Marine Reserve
Vandenberg
Space Launch
Comp
10
B
SLC-
4E
1.39
KM
Bear Creek
UP Santa Barbara Subdivision

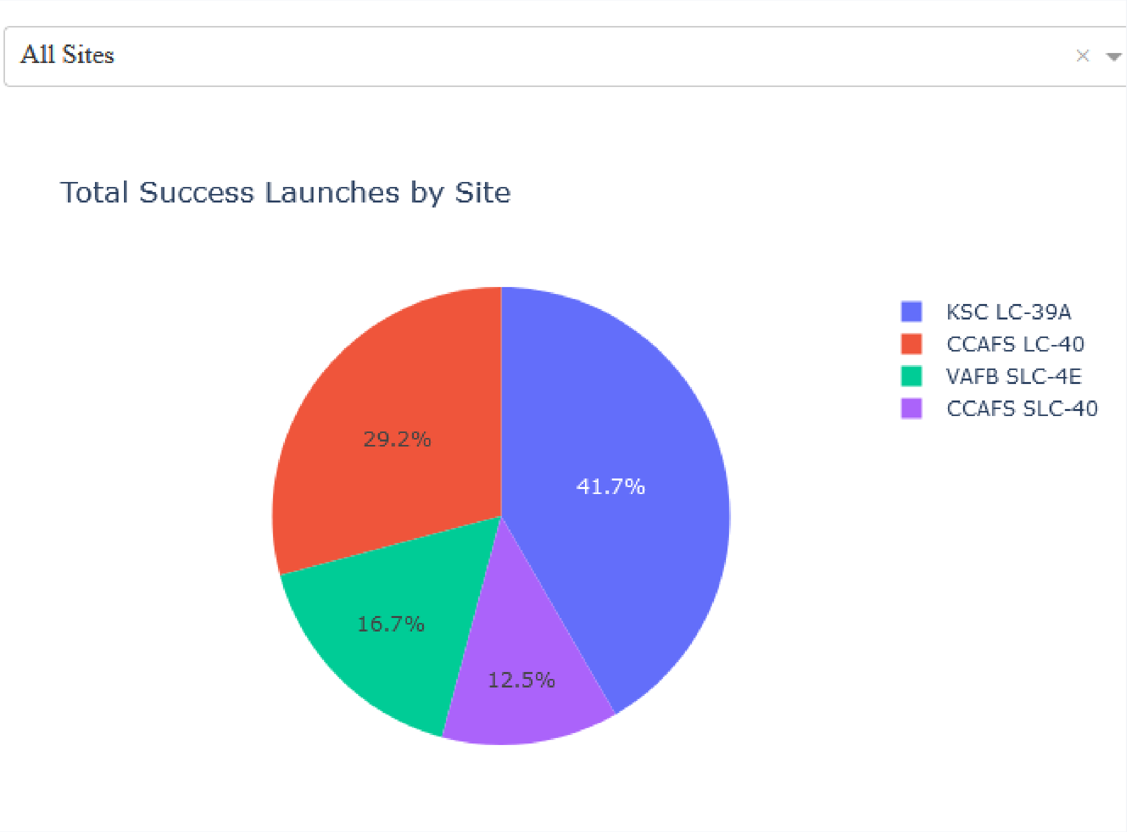


Section 4

Build a Dashboard with Plotly Dash

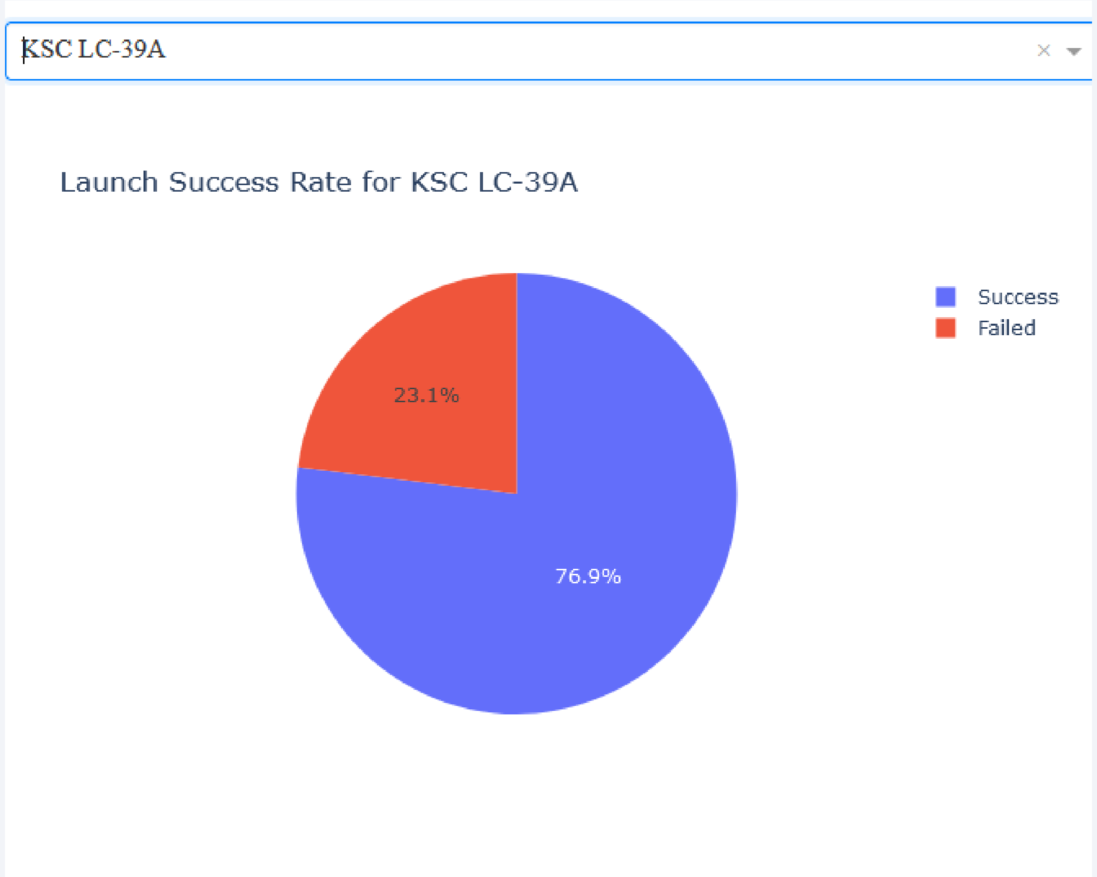
Launch Success – All Sites

- KSC-LC-39A had the highest percent of successful mission
- Launch site appears to be an important factor of success



Launch Success Ratio for KSC LC-39A

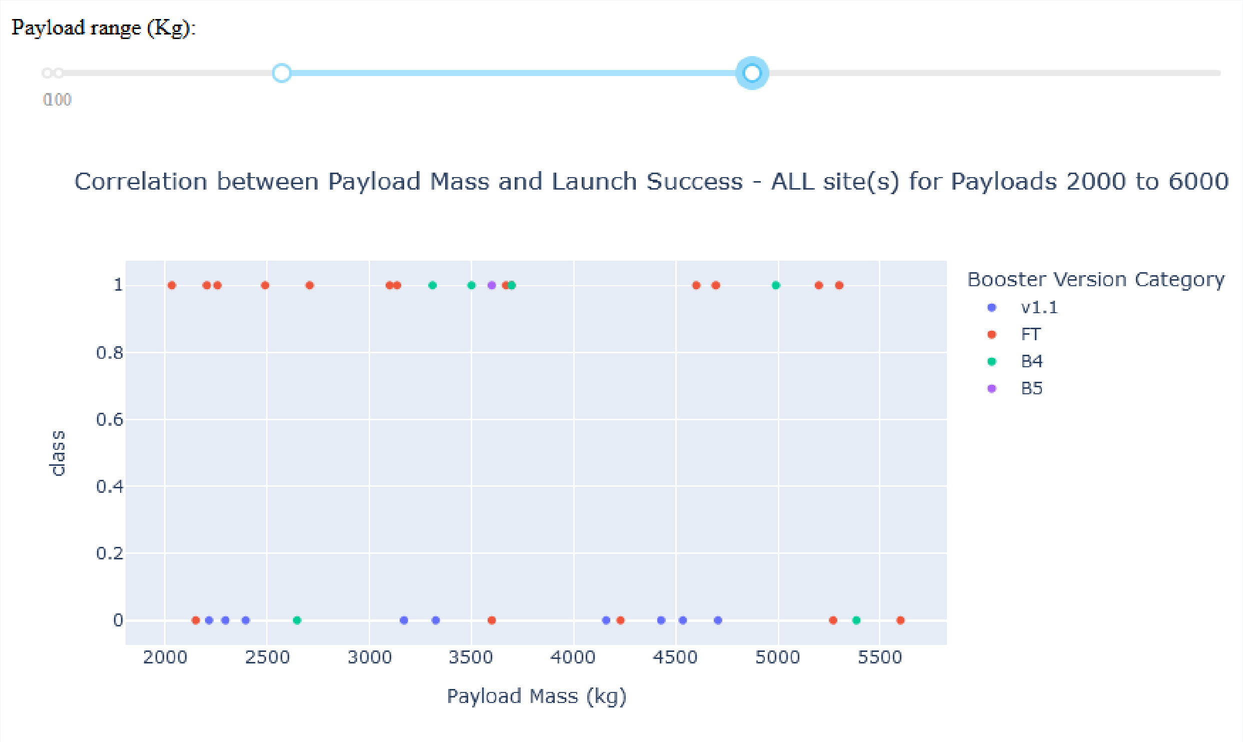
- 76.9% of launches from this site were successful



KSC LC-39A

Payload vs. Launch Outcome

- For FT boosters there are more successful landings than unsuccessful when the payload is between 2000 and 6000
- For V1.1 boosters there all launches were unsuccessful



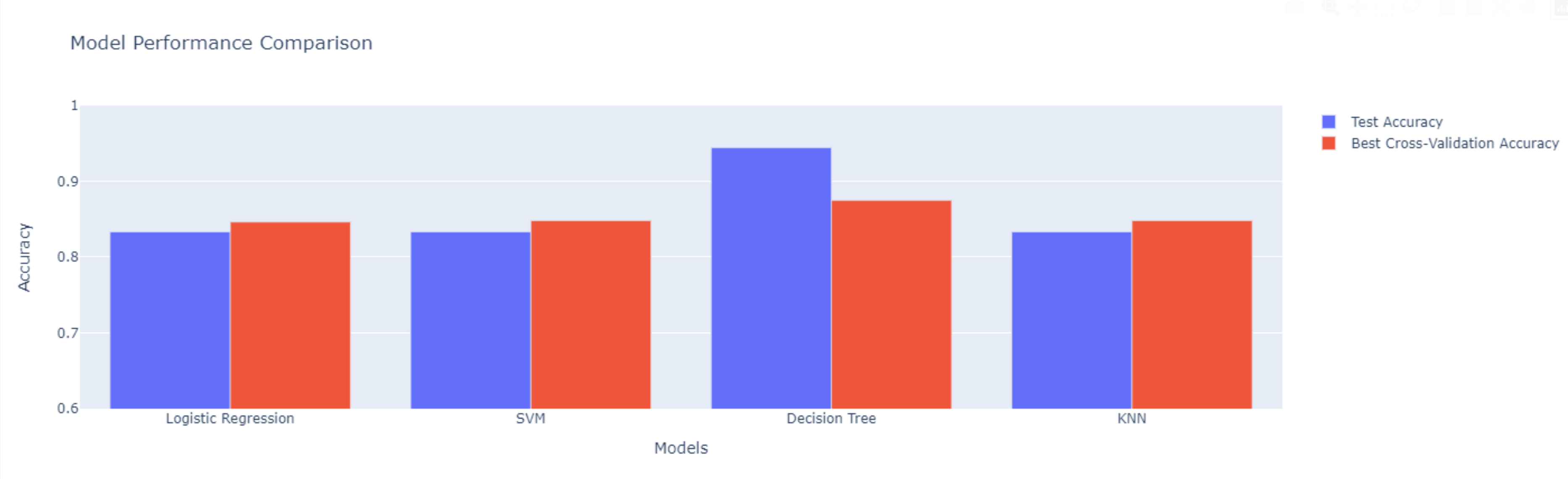
Payload range (Kg):

Correlation between Payload Mass and Launch Success - ALL site(s) for Payloads 2000 to 6000

Section 5

Predictive Analysis (Classification)

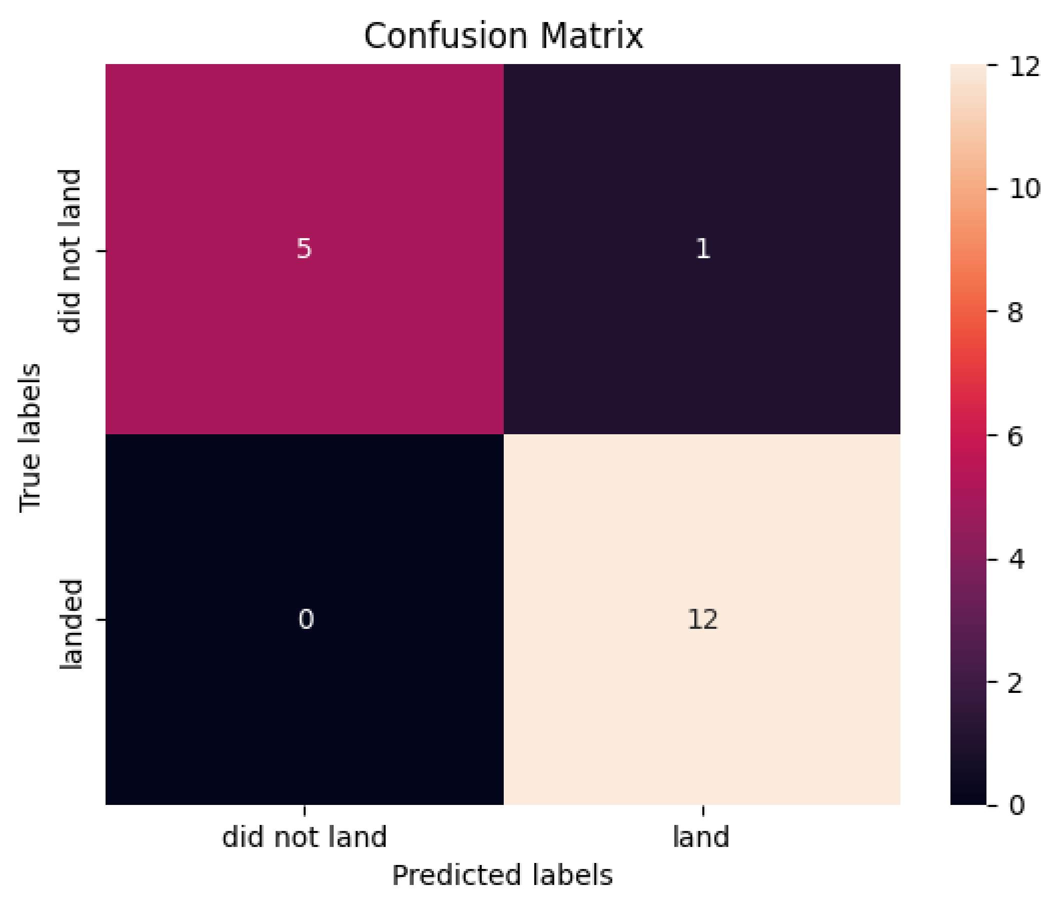
Classification Accuracy



- The above chart visualizes the built model accuracy for all built classification models
- The model with the highest classification accuracy is Decision Tree Classifier

Confusion Matrix

- The following shows the confusion matrix of the best performing model – Confusion Matrix
- It shows a high degree of accuracy with only one false positive



Conclusions

- Multiple data sources were analyzed to refine conclusions throughout the research process.
- KSC LC-39A was identified as the optimal launch site.
- Launches exceeding 7,000kg were found to have a lower risk profile.
- While mission outcomes are predominantly successful, the success rate of landing outcomes has shown improvement over time, reflecting the evolution of launch processes and rocket technology.
- A Decision Tree Classifier can be employed to predict successful landings, potentially enhancing profitability.

Thank you!