IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kenton Swanson
July 16th, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Methodologies Used:
  - Conducted data exploration using SQL queries to extract unique launch site names, success rates, and payload statistics.
  - Visualized key trends through scatter plots, bar charts, and line graphs.
  - Applied machine learning models like Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors to predict mission success. Evaluated using test accuracy.
- Key Results:
  - Launch success rates varied significantly across orbit types, due to the limited amount of data and innovations in rocket designs.
  - Higher flight numbers correlated with higher mission success likelihood.
  - Decision Tree model achieved the highest classification accuracy at 88.89%
  - Launch site activity and performance showed that KSC LC-39A and CCAFS LC-40 handled the most successful high-payload missions.

Introduction

Project Background and Context:

SpaceX revolutionized the science and economics by discovering how to recover and reuse the Falcon 9 first stage, allowing them to advertise missions at around $62M compared to $165M+ for other competitors. Since the first stage is the largest and most expensive element of the operation, the ability to predict landing success is directly tied to estimating mission cost.

In this capstone, I will be acting as a Data Scientist for Space Y, using publicly available information to aggregate launch records, wrangle data, and train classification models to predict whether the first stage will land and become reusable in order to have competitive pricing.

Key Questions:

- How has Falcon 9 landing outcomes evolved over time?
- Which mission factors contribute the most to landing success?
- Which graphs and plots will provide the most relevant data to achieving our goals?
- Which Classification Model will have the highest classification accuracy?

Section 1

Methodology

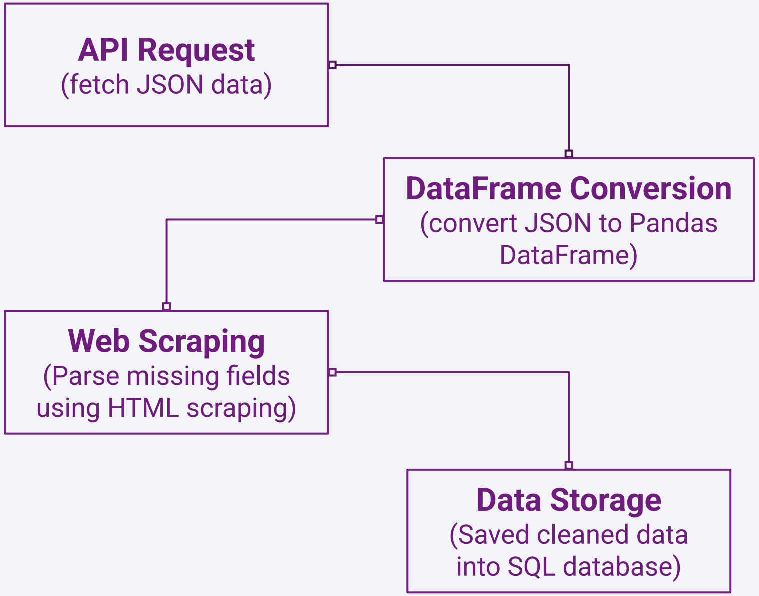
Methodology

Executive Summary:

- Data collection methodology:
  - Public Falcon 9 launch data was sourced from SpaceX APIs and datasets.
- Perform data wrangling
  - CSV files were cleaned, merged, and preprocessed using Pandas.
- Perform exploratory data analysis (EDA) using visualization and SQL.
  - Visual EDA used to find trends in Landing success. SQL queries used to segment and summarize mission outcomes.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Used to visualize launch sites, mission clustering, and model insights interactively.
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated classification models.

Data Collection

- Rest API Integration:
  - Used SpaceX’s public API to extract data on past Falcon 9 launches, including payload mass, launch site, orbit, rocket version, and landing outcome.
- Web Scraping:
  - Supplemented missing features like payload mass using BeautifulSoup to scrape SpaceX launch archive web pages.
- SQL Queries:
  - Processed and filtered structured launch data stored in a local SQLite database to support targeted analysis.



API Request
(fetch JSON data)

DataFrame Conversion
(convert JSON to Pandas DataFrame)

Web Scraping
(Parse missing fields using HTML scraping)

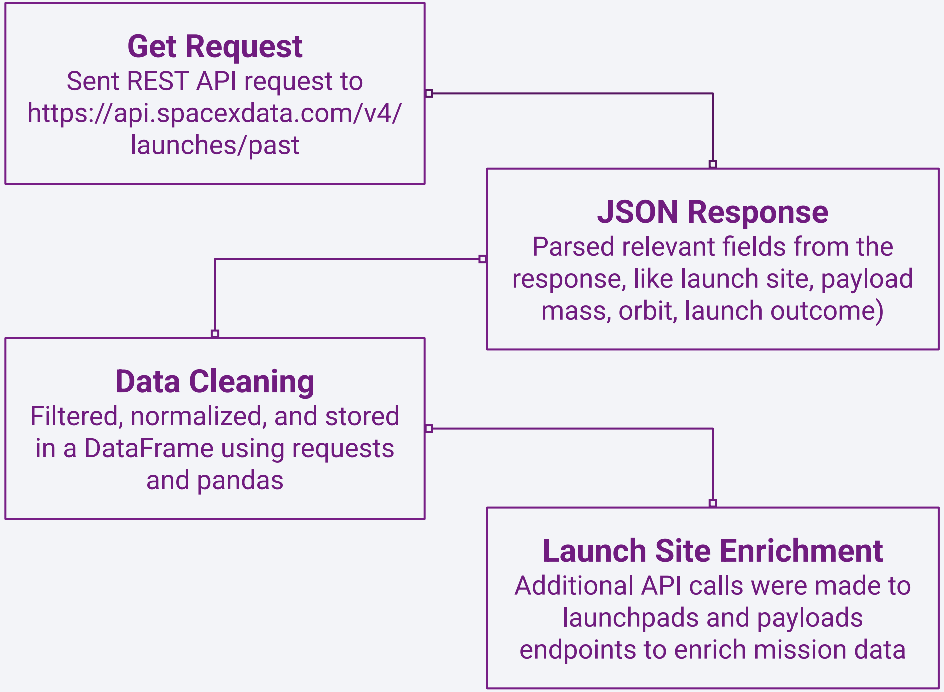
Data Storage
(Saved cleaned data into SQL database)

Data Collection – SpaceX API

Purpose:

- To extract important data required for EDA, visualization, and classification model training.

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Get Request
Sent REST API request to https://api.spacexdata.com/v4/launches/past

JSON Response
Parsed relevant fields from the response, like launch site, payload mass, orbit, launch outcome)

Data Cleaning
Filtered, normalized, and stored in a DataFrame using requests and pandas

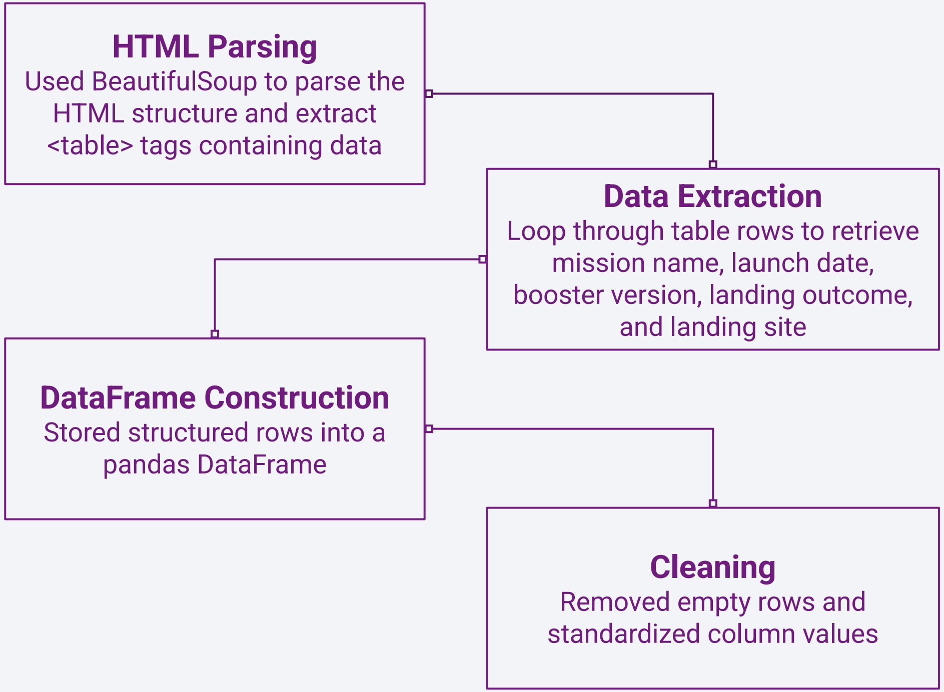
Launch Site Enrichment
Additional API calls were made to launchpads and payloads endpoints to enrich mission data

Data Collection - Scraping

Purpose:

- Supplement API and SQL datasets with historic booster landing data not available from SpaceX datasets. This will improve input quality.

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/jupyter-labs-webscraping.ipynb



HTML Parsing
Used BeautifulSoup to parse the HTML structure and extract <table> tags containing data

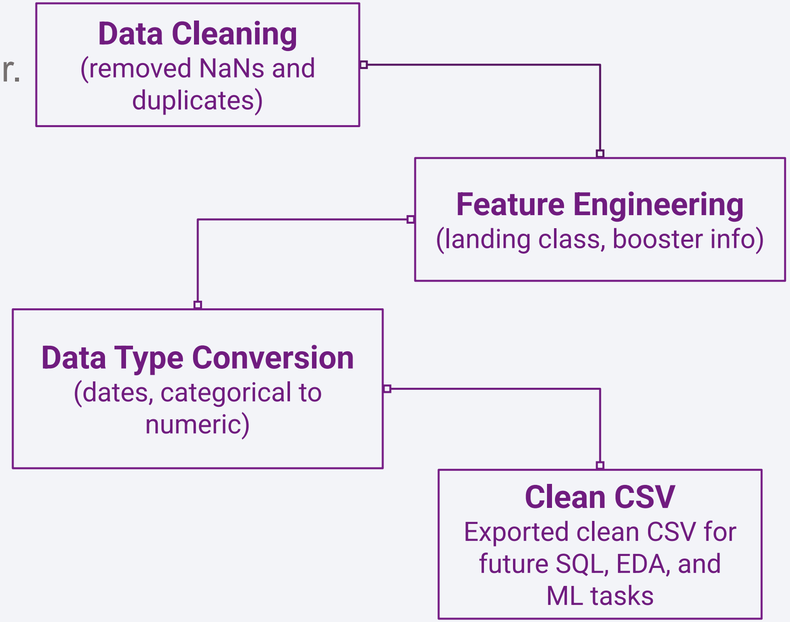
Data Extraction
Loop through table rows to retrieve mission name, launch date, booster version, landing outcome, and landing site

DataFrame Construction
Stored structured rows into a pandas DataFrame

Cleaning
Removed empty rows and standardized column values

Data Wrangling

- Data Cleaning:
  - Removed null values, duplicate records, and irrelevant columns like Unnamed and Flight Number. Also standardized feature naming for consistency.
- Feature Engineering:
  - Created binary landing outcome feature (Class) to represent success/failure and extracted features Such as BoosterVersion, Orbit, and LaunchSite.
- Data Transformation:
  - Converted data types (dates to datetime format, booleans to integers) and encoded categorical variables for model-readiness.



Data Cleaning
(removed NaNs and duplicates)

Feature Engineering
(landing class, booster info)

Data Type Conversion
(dates, categorical to numeric)

Clean CSV
Exported clean CSV for future SQL, EDA, and ML tasks

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Success Rate per Launch Site Chart:
  - Created bar charts to compare the success rate of the first stage landing across different SpaceX launch sites. This helped identify which sites have the highest landing success.
- Payload vs. Success Correlation Chart:
  - Used scatter plots with success/failure coloring to examine how payload mass influences landing outcome. This detected performance thresholds related to payload size.
- Booster Version and Outcome Chart:
  - Visualized landing outcomes per booster version using grouped bar charts. This analyzed whether newer or specific booster versions lead to higher success.
- Orbit Type and Outcome Chart:
  - Used categorical plots to assess how orbit types (e.g., LEO, GTO) relate to landing success. This determined if mission target or orbit complexity affects reusability.

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Performed aggregate counts grouped by launch site:

  ```
  SELECT Launch_Site, COUNT(*) AS Launch_Count FROM SPACEXTBL GROUP BY Launch_Site;
  ```

- Retrieved top 5 successful launches based on payload mass:

  ```
  SELECT * FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' ORDER BY Payload_Mass__kg_ DESC LIMIT 5;
  ```

- Filtered launch records for missions in specific orbits:

  ```
  SELECT * FROM SPACEXTBL WHERE Orbit = 'GTO';
  ```

- Identified launches with failed landing attempts:

  ```
  SELECT * FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%';
  ```

- Performed inner join to compare landing outcomes and booster versions:

  ```
  SELECT Landing_Outcome, Booster_Version, COUNT(*) AS Count FROM SPACEXTBL GROUP BY Landing_Outcome, Booster_Version;
  ```

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch-site Markers:
  - Created a Folium Marker for every Falcon 9 launch, color-coded by green (landed) vs red (failed). These are bundled with the function marker_cluster to avoid overplotting.
- Proximity Circles:
  - Created Folium Circles (1 km) around each pad to highlight the immediate safety zone.
- Distance Polylines
  - Created Folium Polylines drawn from the pad to the nearest coastline, highway, & railway. These demonstrate the distance from the launch site to various Points of Interest for cargo delivery and water access.

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

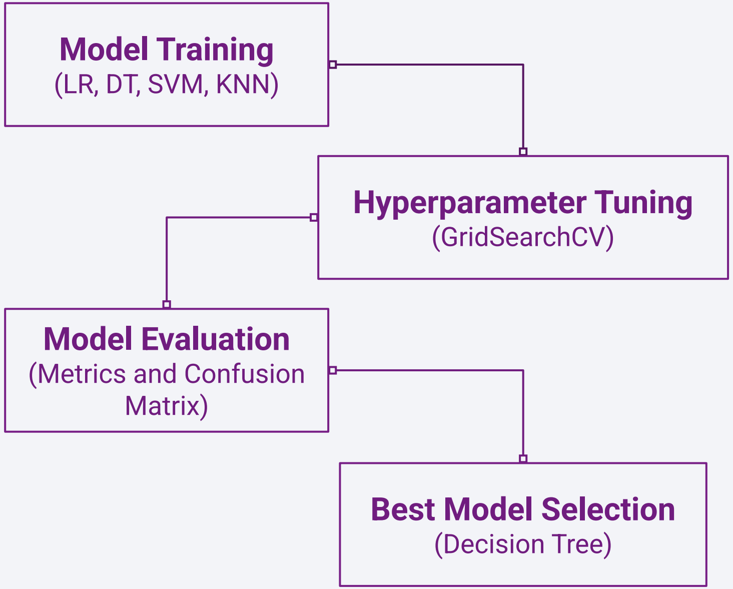
Build a Dashboard with Plotly Dash

- Launch Outcome Pie Chart:
  - Created a pie chart showing the total successful launches. When "All Sites" is selected, it displays overall success counts. When a specific site is selected, it shows the ratio of successful vs failed launches for that site.
- Launch Site Dropdown Filter:
  - Added a dropdown menu to filter visualizations by individual launch site or show all combined. This allows users to compare site performance interactively.
- Payload Range Slider:
  - Integrated a slider to adjust the payload mass range displayed in the scatter plot. This enables focused analysis on specific payload intervals.
- Payload vs. Outcome Scatter Plot:
  - Constructed a scatter plot showing payload mass vs launch success. They are color-coded by booster version and indicate success/failure, showing payload size vs. mission outcomes.

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/tree/main/Dash%20Lab

Predictive Analysis (Classification)

- Model Building & Evaluation:
  - Four models were built: Logistic Regression, Decision Tree, Support Vector Machine, and K-Nearest Neighbors. Each model was trained on SpaceX launch data.
- Hyperparameter Tuning & Optimization:
  - GridSearchCV was applied to optimize hyperparameters: C, penalty for LR, max_depth for DT, C, kernel for SVM, n_neighbors, weights, for KNN. Best selected on accuracy.
- Performance Comparison:
  - All models were evaluated using accuracy, precision, recall, and F1-score. Confusion matrices were generated for each model to visualize prediction outcomes.



Model Training
(LR, DT, SVM, KNN)

Hyperparameter Tuning
(GridSearchCV)

Model Evaluation
(Metrics and Confusion Matrix)

Best Model Selection
(Decision Tree)

GitHub Link:
https://github.com/kentonswanson/DS-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

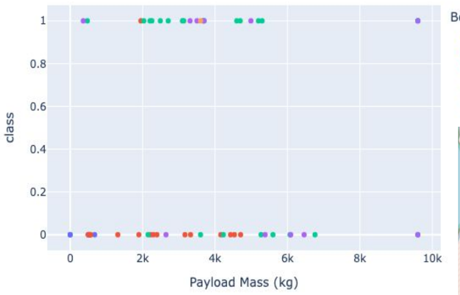
Results

Exploratory Data Analysis

- High first-stage success rate is around 78 % overall.
- LEO & ISS orbits have the highest landing success, while GTO is riskier than the others.
- Success probability increases with Flight Number (experience curve).

Interactive Analytics

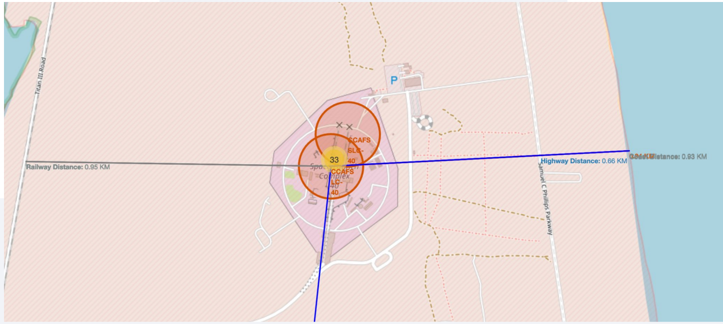
Payload vs. Outcome for All Sites



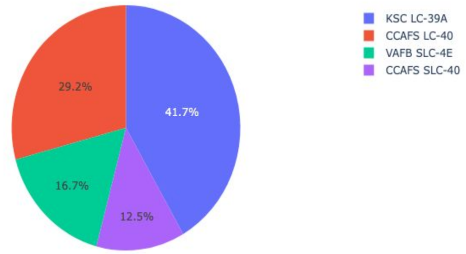
^^^^^^^^^^^^

Shows more failure beyond 6 tons

0.66km highway distance
0.95km railway distance
17.21km city distance
vvvvvvvvvvvv



Total Successful Launches by Site

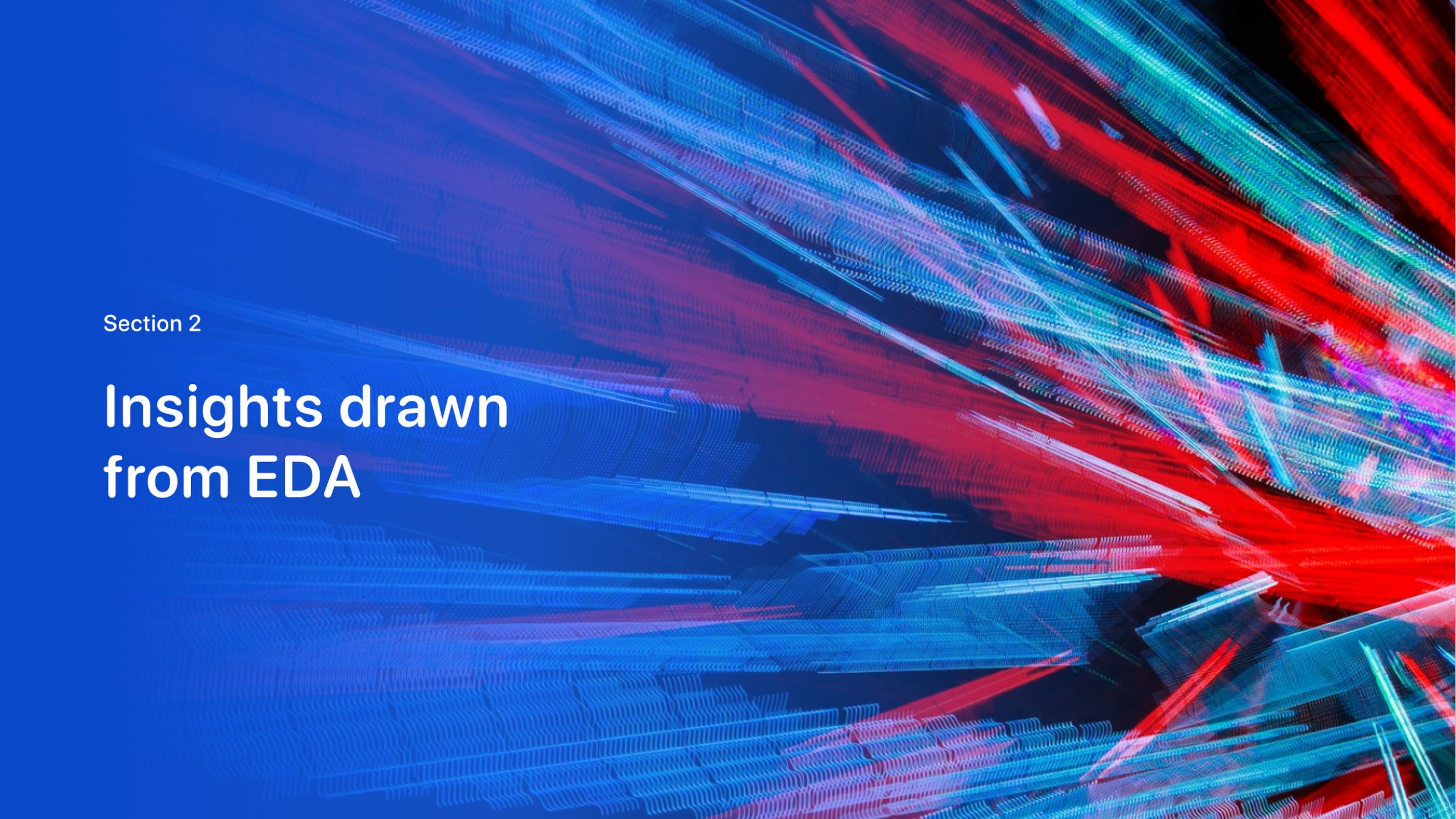


^^^^^^^^^^^^^

KSC LC-39A most success
CCAFS SLC-40 least

Predictive Analysis

- Decision Tree was selected as best model (88.9 % accuracy).
- Confusion matrix: 11 / 1 / 1 / 5

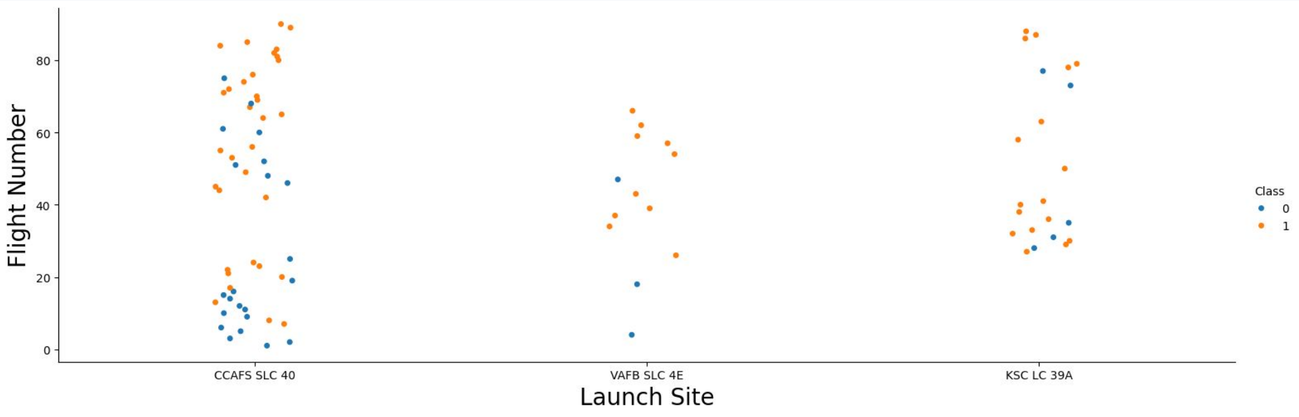


Section 2

Insights drawn from EDA

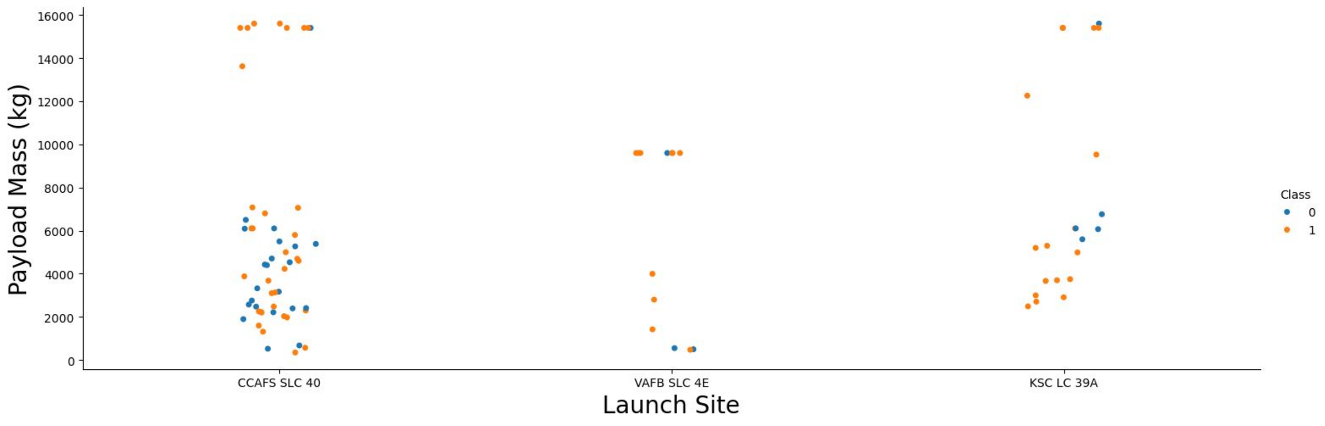
Flight Number vs. Launch Site

- This scatter plot displays the Flight Number (y-axis) against Launch Site (x-axis), with each point color-coded by the mission outcome (Class 1 = Success, Class 0 = Failure).
  - CCAFS SLC 40 has the most missions, however there is a high concentration of failures among the early missions and a high concentration of successes among the later missions. KSC LC 39A has the most missions at higher flight numbers, meaning it may be a site to test more mature booster designs.



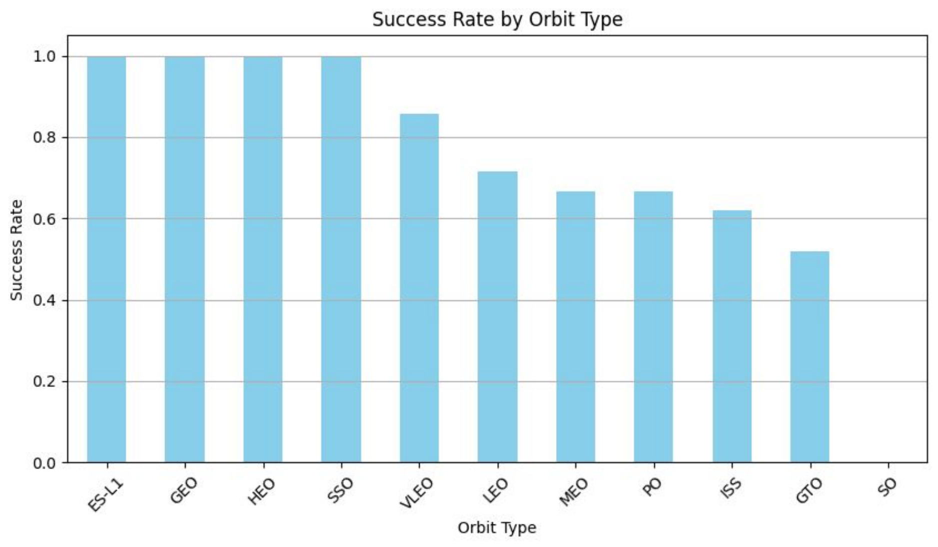
Payload vs. Launch Site

- This scatter plot displays the Payload Mass (y-axis) against Launch Site (x-axis), with each point color-coded by the mission outcome (Class 1 = Success, Class 0 = Failure).
  - CCAFS SLC 40 has the widest range of payload masses, higher payloads tend to succeed. VAFB SLC 4E has fewer launches than the others. KSC LC 39A has roughly the same successes as CCAFS SLC 40 at higher payload masses, so these two sites seem optimized for heavy-lift missions.



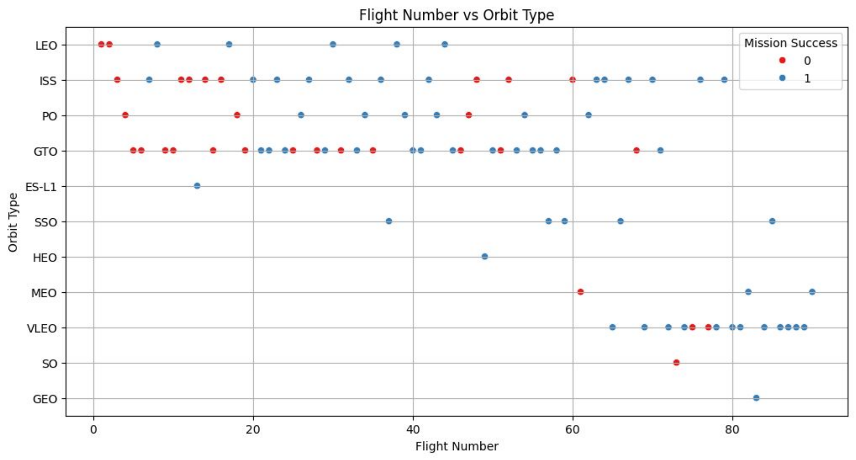
Success Rate vs. Orbit Type

- This bar chart displays the success rate of Falcon 9 launches grouped by orbit.
  - ES-L1, GEO, HEO, and SSO achieved a 100% success rate, however GEO, HEO, and VLEO had 1 success, while SSO had 5 successes.
  - VLEO had the highest non-perfect success rate
  - SO had the lowest success rate with 0%, however there was only flight in that orbit.



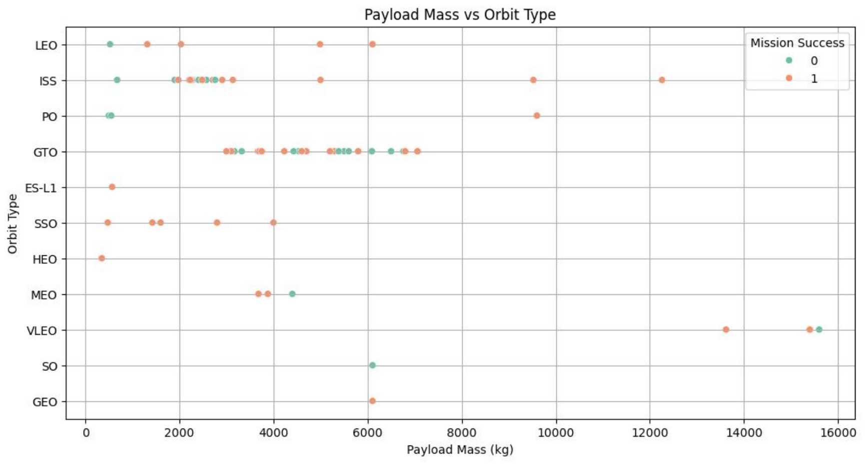
Flight Number vs. Orbit Type

- This scatter plot visualizes the relationship between Flight Number (chronological order of launches) and Orbit Type, with each point color-coded by the mission outcome (Class 1 = Success, Class 0 = Failure).
  - Early flights show a higher concentration of failures, particularly for GTO, PO, and ISS orbits. Later flights are predominantly successful, indicating improved reliability over time. Some orbits like SSO, HEO, VLEO, and GEO only appear in later flight numbers and show consistent success.



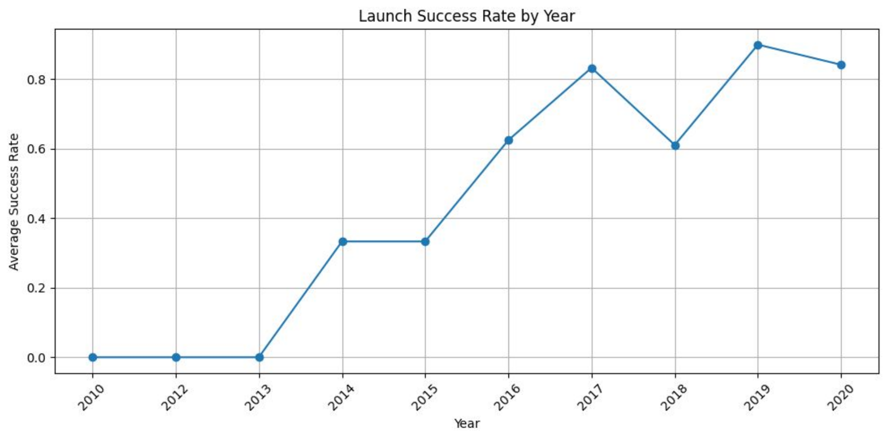
Payload vs. Orbit Type

- This scatter plot displays the Payload Mass (kg) against Orbit Type, with each point color-coded by the mission outcome (Class 1 = Success, Class 0 = Failure).
  - Mission success rates tend to increase at moderate payload levels, especially in orbits like SSO and HEO. Failures tend to be concentrated in mid-range payloads, particularly within GTO and PO missions. The VLEO orbit type stands out by supporting very high payloads (>15,000 kg), with a strong record of success.



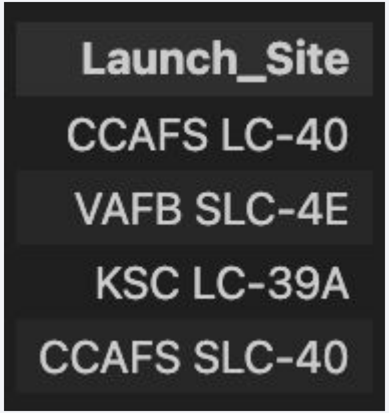
Launch Success Yearly Trend

- The line chart displays the average mission success rate per year, from 2010 to 2020.
  - 2010-2013: Success stayed consistent at 0%.
  - 2014-2017: Major growth, pushing the success rate up to 33% and to 83%.
  - 2018: Dip in success rate down to 60%
  - 2019-2020: Maintained a high success rate at above 80%, demonstrating consistency.



All Launch Site Names

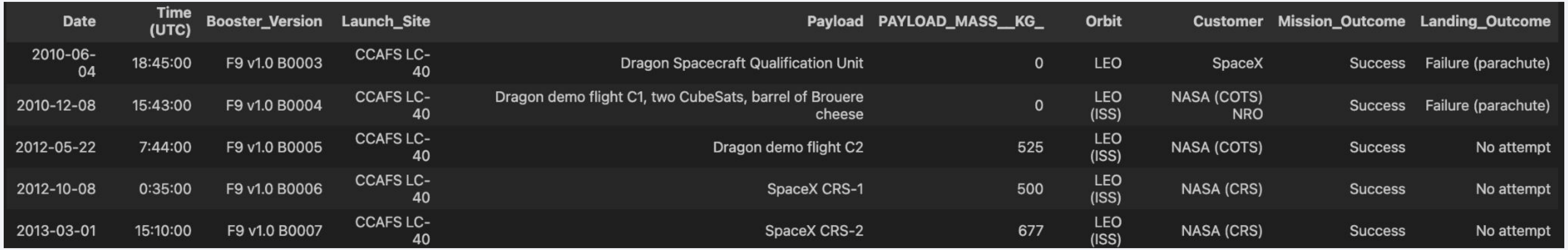
- Here are all the unique launch site names:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40



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

Launch Site Names Begin with 'CCA'

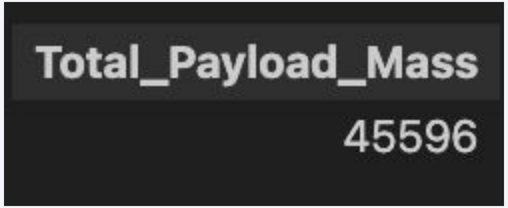
- Here are 5 records where launch sites begin with `CCA`:



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

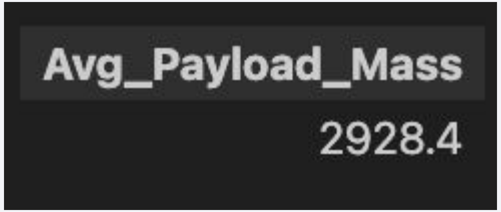
- The total payload mass of Boosters carried by NASA is 45,596 kg.



Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

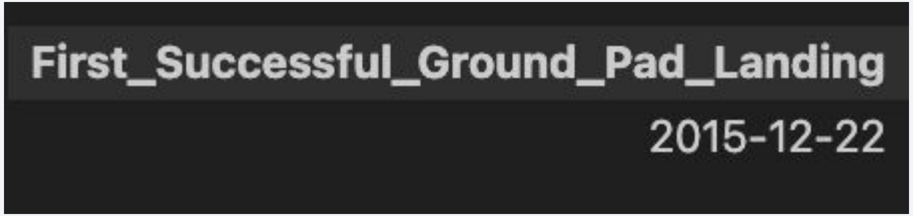
- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.



Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

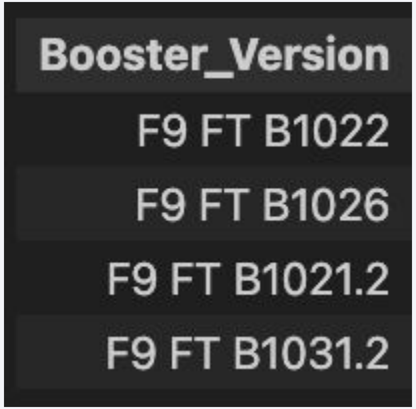
- The date of the first successful landing outcome on ground pad is December 22nd, 2015.



First_Successful_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

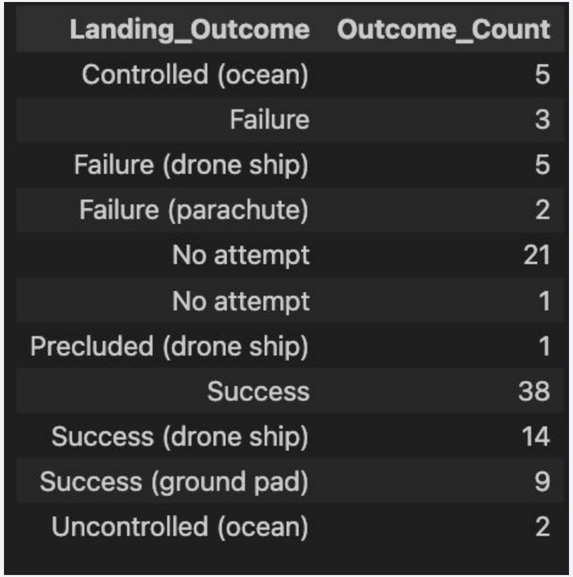
- Here are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2



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

Total Number of Successful and Failure Mission Outcomes

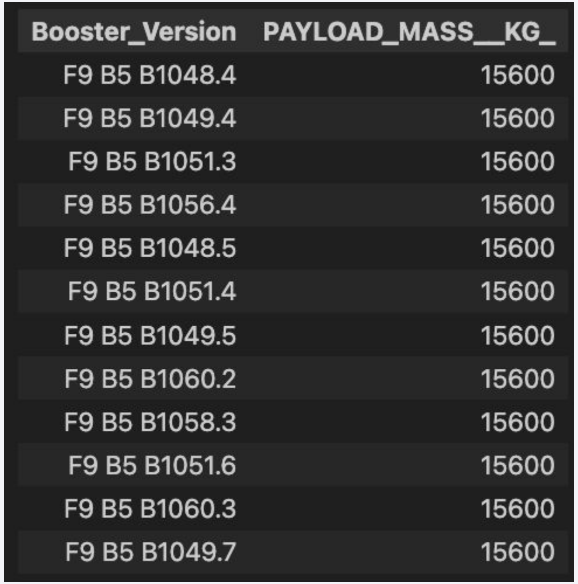
- Here is the total number of successful and failure mission outcomes, grouped by Landing_Outcome:



Landing_Outcome	Outcome_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

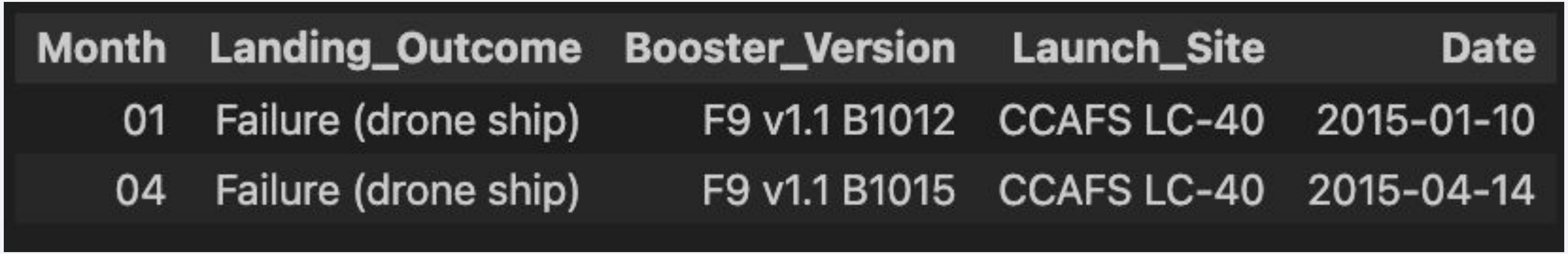
- Here is the list of booster names which have carried the maximum payload mass:



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

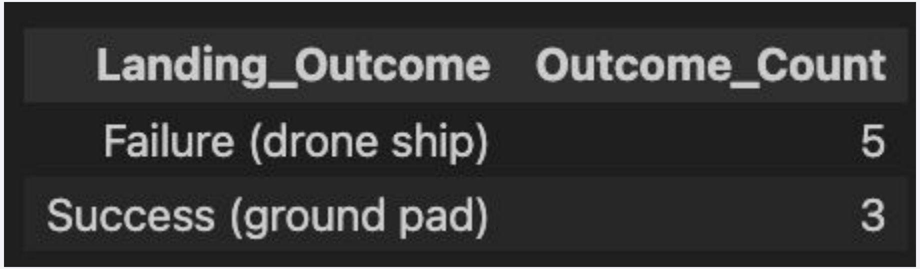
- Here are the failed landing_outcomes in drone ship, their booster versions, and launch site names in the year 2015:



Month	Landing_Outcome	Booster_Version	Launch_Site	Date
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

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



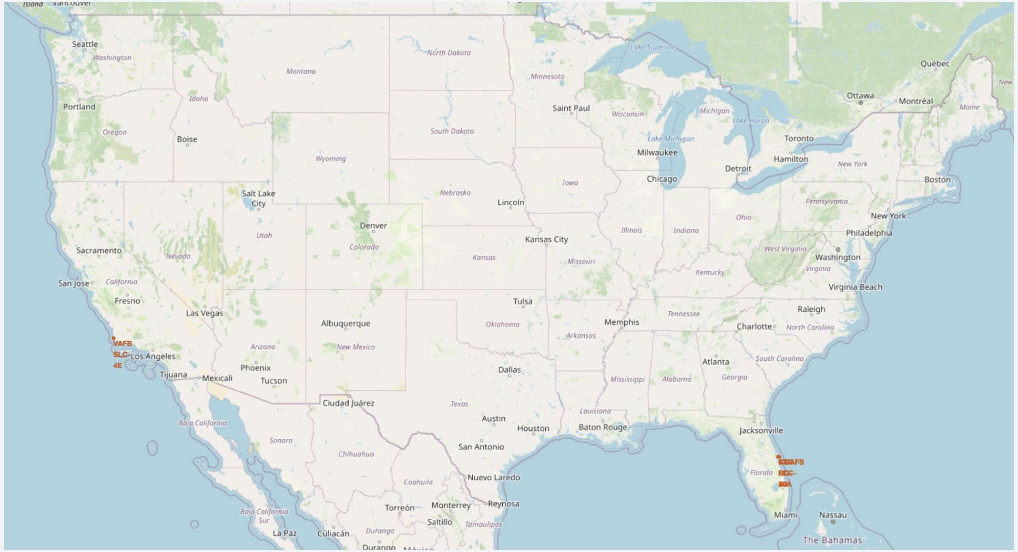
Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

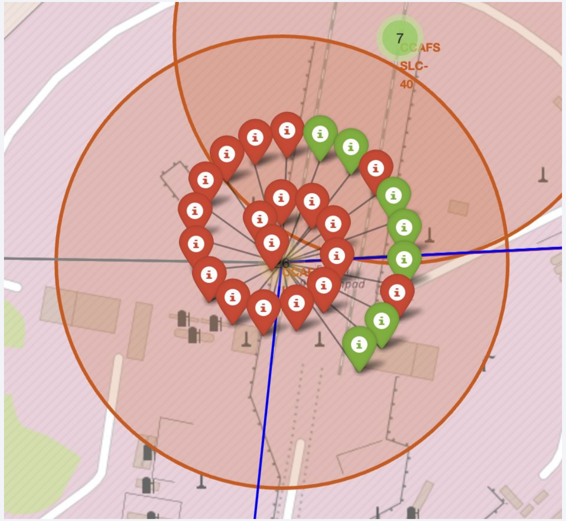
Map of the Launch Sites

- From the map we can see that 3 of the launch sites are in Florida (CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A) and 1 launch site is in California (VAFB SLC-4E).



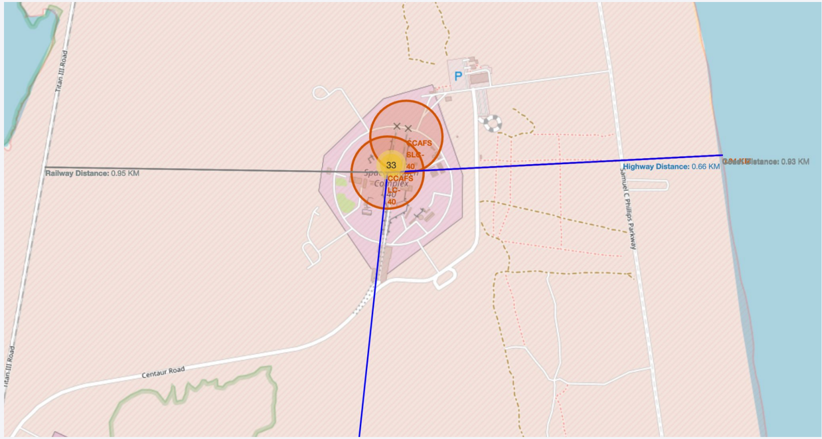
Colored Label Launch Outcomes

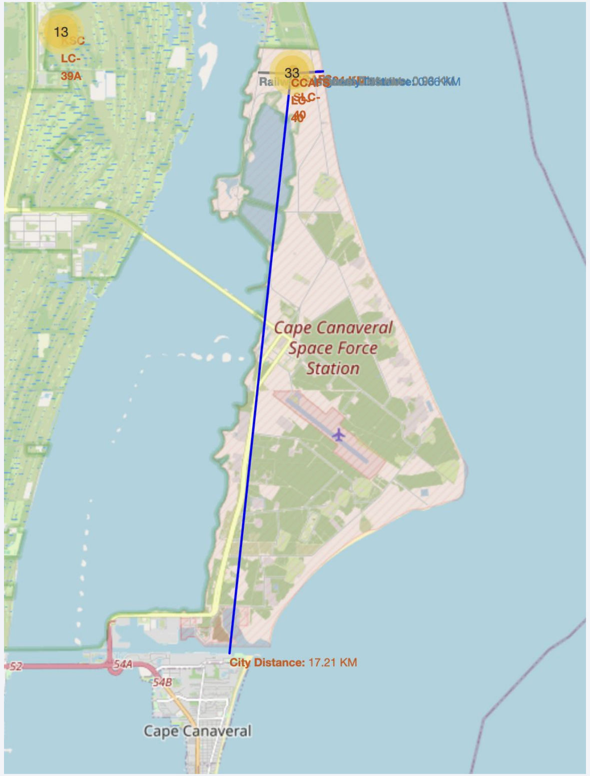
- At each launch site, every success is represented by a green circle while every failure is represented by a red circle. Site CCAFS SLC 40 has had 26 launches, 7 successes and 19 failures, demonstrated by the circles.

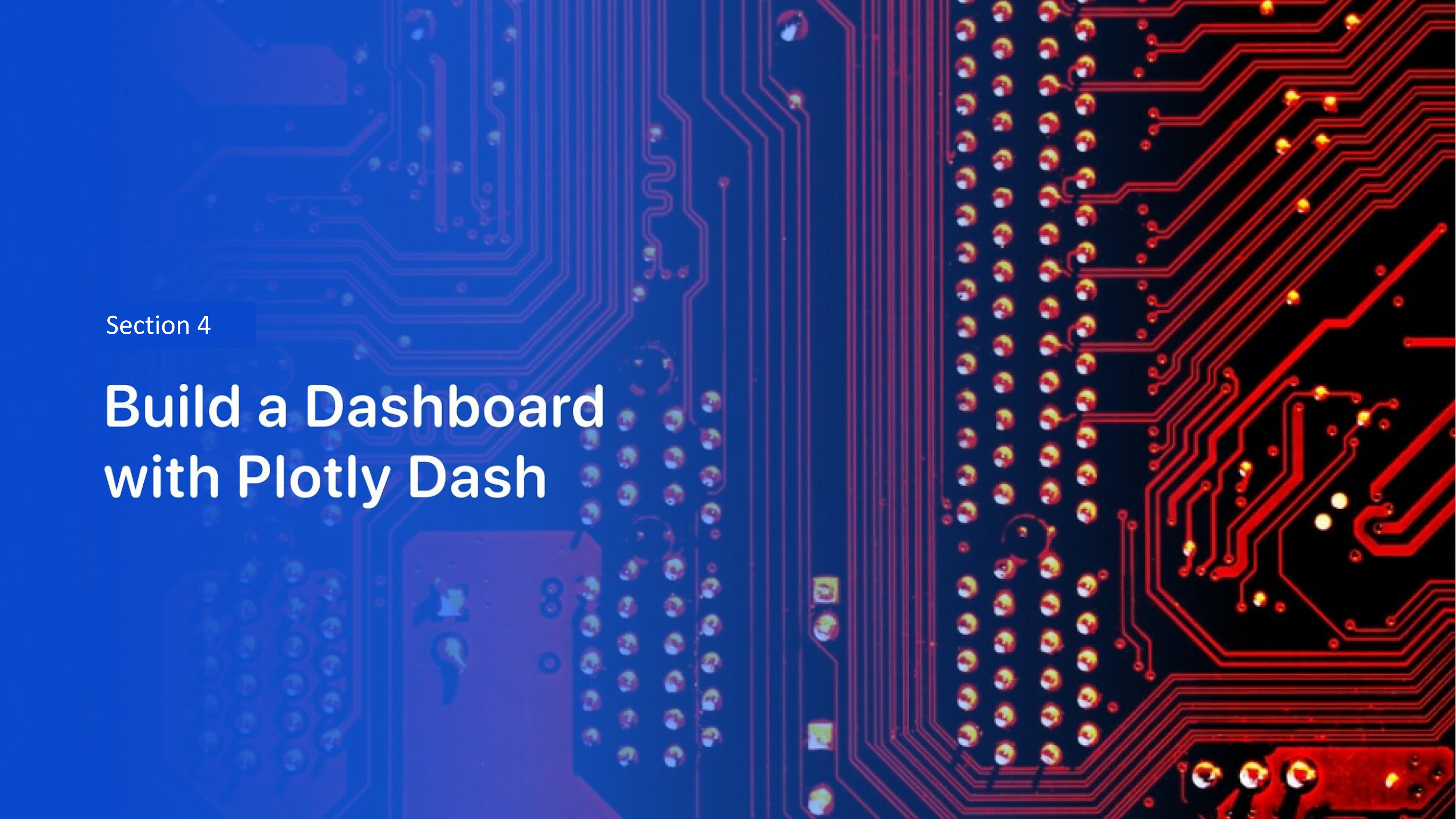


CCAFS SLC 40 and Its Proximities

- CCAFS SLC 40 is:
  - 0.66 kilometers away from a highway
  - 0.93 kilometers away from an ocean
  - 0.95 kilometers away from a railway
  - 17.21 kilometers away from a city


Titan III Road
P
CCAFS SLC-40
33
CCAFS LC-40
Space Complex
Railway Distance: 0.95 KM
Highway Distance: 0.66 KM
Coastline Distance: 0.93 KM
Samuel C Phillips Parkway
Centaur Road



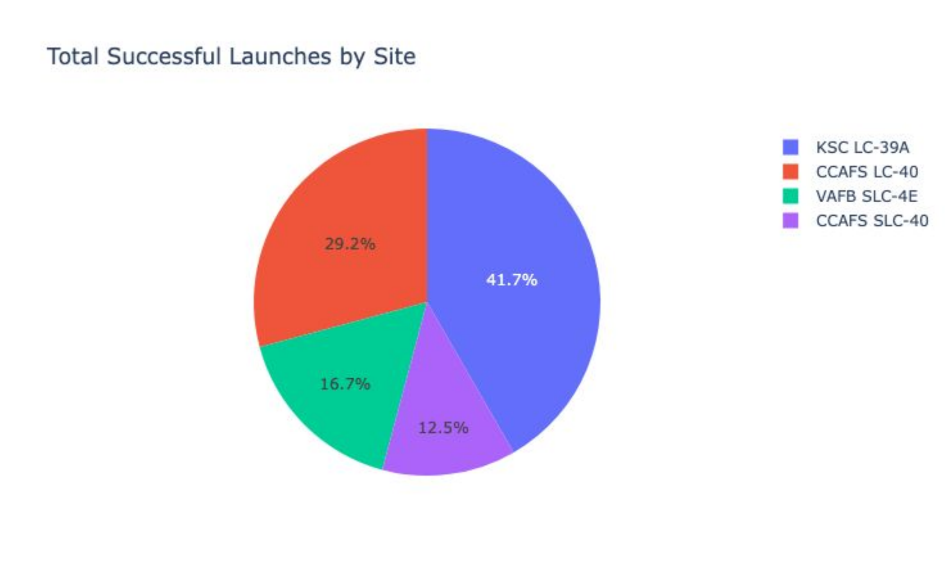


Section 4

Build a Dashboard with Plotly Dash

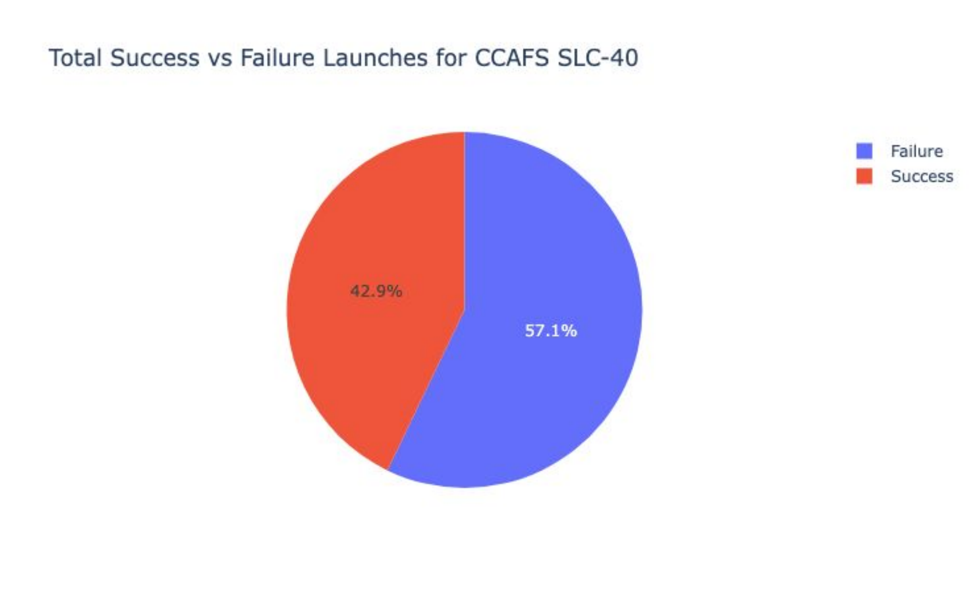
Total Successful Launches by Site

- According to the pie chart, KSC LC-39A had the highest amount of successful launches at 41.7%, while CCAFS SLC-40 had the smallest amount of successful launches at 12.5%.



Total Success vs Failure Launches for CCAFS SLC-40

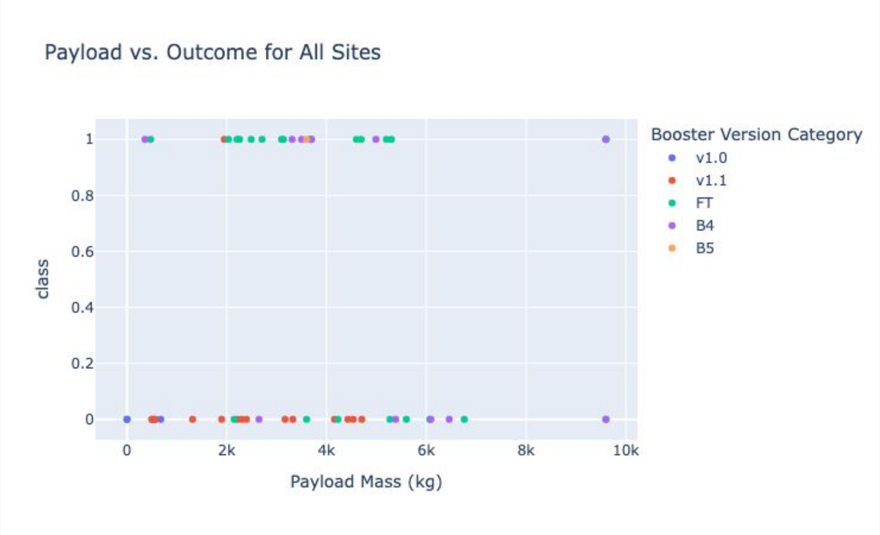
- Despite being the launch site with the least amount of successful landings, CCAFS SLC-40 has the highest percentage of successful landings, at 42.9%.



Payload vs. Outcome for All Sites

- According to the plot, the most effective Booster Version is category FT, while the least effective Booster Version is v1.0, closely followed by v1.1.

- A payload size below 6,000 kg has a higher rate of success, while a payload size above 6,000 kg has a much lower rate of success.



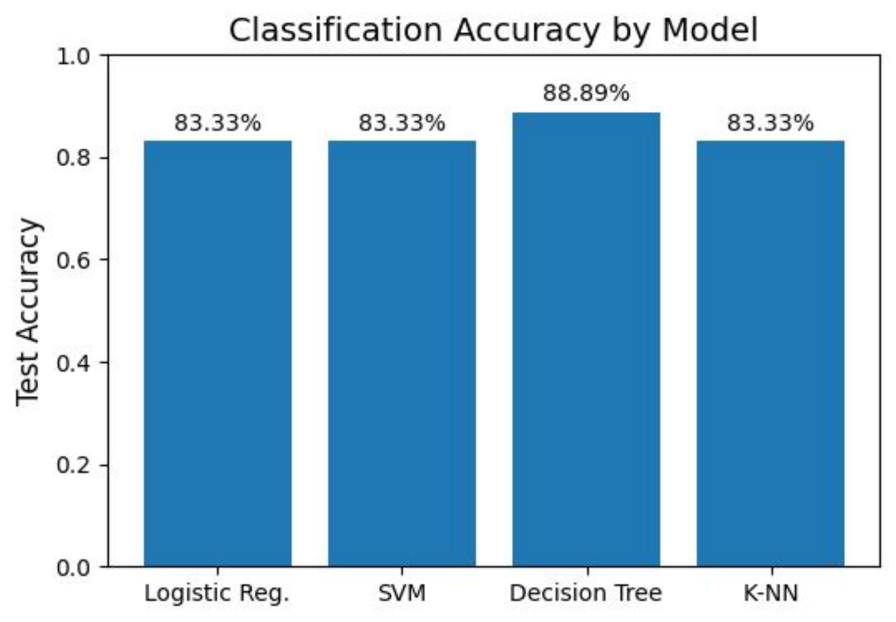
Section 5

Predictive Analysis (Classification)

Classification Accuracy

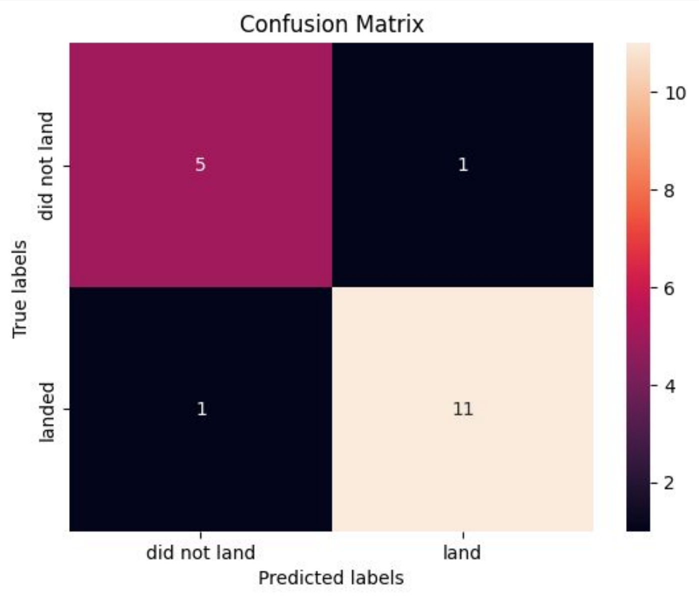
- According to the bar chart and the Machine Learning Notebook, Decision Tree has the highest classification accuracy at 88.89%

- Logistic Regression, SVM, and K-NN all are slightly below Decision Tree, at 83.33%.



Confusion Matrix

- The Confusion Matrix of Decision Tree shows that the Model correctly predictions 11 to land and 5 not to land, and incorrectly predicts 1 to land and 1 not to land. This yields a 16/18, or 88.9%, classification accuracy.



Conclusions

- Reusable-booster prediction is possible, based on public data:
  - Using the SpaceX API, web-scraped launch tables, and engineered features yields a clean SQL dataset that explains approximately 70 % of first-stage landing variability.
- Key drivers of landing success:
  - 1. Lower payload mass (< 4 t), 2. LEO / ISS orbits, 3. Booster versions ≥ “FT”, 4. Higher prior flight count (due to experience).
- Site-specific performance differs greatly:
  - CCAFS SLC-40 shows the widest payload range but the highest failure rate. However, KSC LC-39A achieves the most consistent landings.
- Machine Learning has revealed that Decision Tree is the top classifier:
  - Logistic Regression, SVM, and K-NN all scored 83% accuracy, while Decision Tree yielded 88.9% accuracy.

Thank you!