IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohamad Alhaidar
January 14, 2026

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected from SpaceX's REST API, integrated with data from web scraping.
  - The dataset was preprocessed and encoded to prepare for analysis and model training.
  - Exploratory data analysis using SQL and visualizations was performed to gain insights and select relevant features.
  - Four classification models (Linear Regression, SVM, Decision Tree, and KNN) were trained to predict the landing outcome of the first stage of SpaceX launches.
- Summary of all results
  - Several features were observed to have high correlation with the target variable and were chosen for model training.
  - The models showed similar performance on test data with an accuracy score of 83.33%.

Introduction

- SpaceX launches Falcon 9 rockets with a lower cost compared to other providers, by reusing the first stage.

- This project aims to collect data related to SpaceX Falcon 9 launches, analyze the data, and develop machine learning models to predict whether the first stage launches will land successfully or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launches data was collected and integrated from two sources:
    - SpaceX REST API endpoints: https://api.spacexdata.com/v4/rockets/
    - Web scraping from: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Data wrangling:
  - The data was cleaned, a column for the outcome (Class) was created as the target variable, and categorical features were encoded using One-Hot encoding.

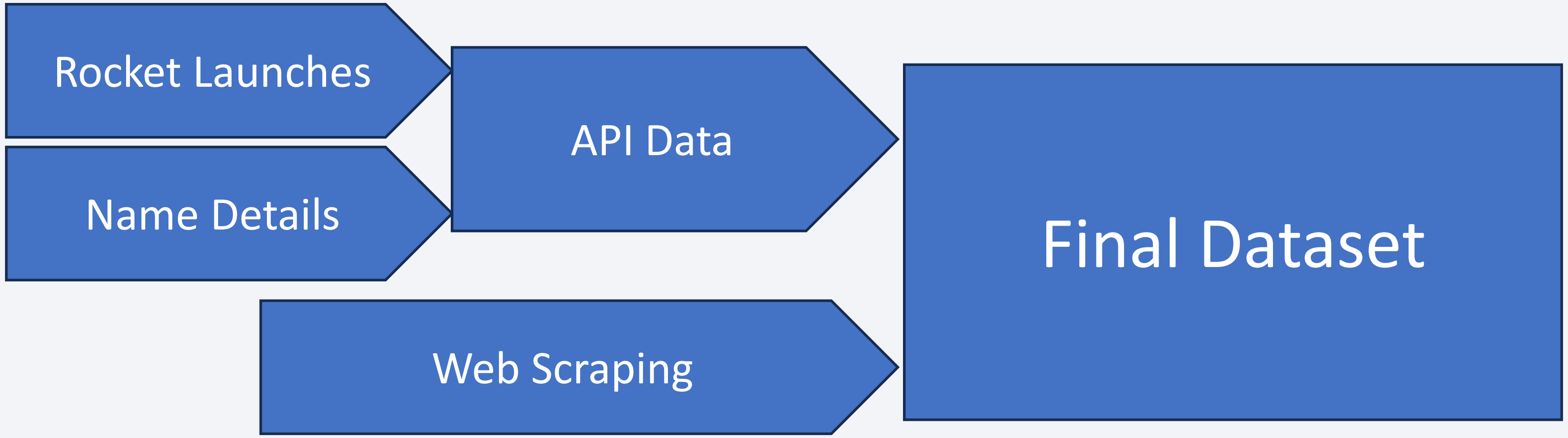
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
    - SQL was used to query the data and find important launch sites and models.
- Perform interactive visual analytics using Folium and Plotly Dash
    - Launch sites and relevant information were displayed on the map using Folium.
    - An interactive dashboard was created to display key information using Plotly Dash.
- Perform predictive analysis using classification models
    - The dataset was split into training and test sets.
    - The models were trained using cross-validation and were evaluated on the accuracy of predictions on the test set.

Data Collection

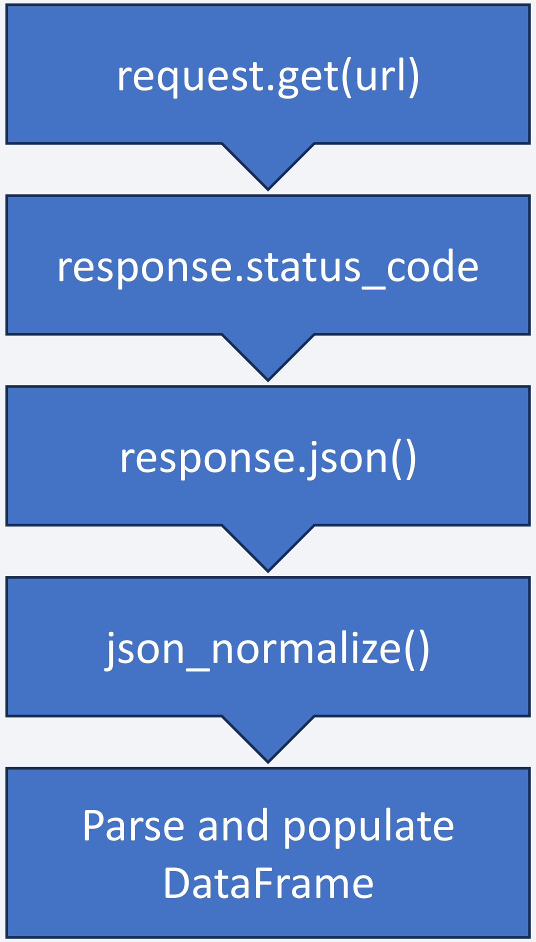
- The dataset was constructed by combining data collected from SpaceX REST API through its provided endpoints, joining data from several endpoints based on rocket IDs and location IDs, in addition to data collected from Wikipedia using web scraping tools, like Beautiful Soup.



Data Collection – SpaceX API

- SpaceX provides data through endpoints.
- Request the data through the URL.
- Confirm request success
- Extract JSON content
- Parse and convert into a DataFrame
- Clean and select relevant columns

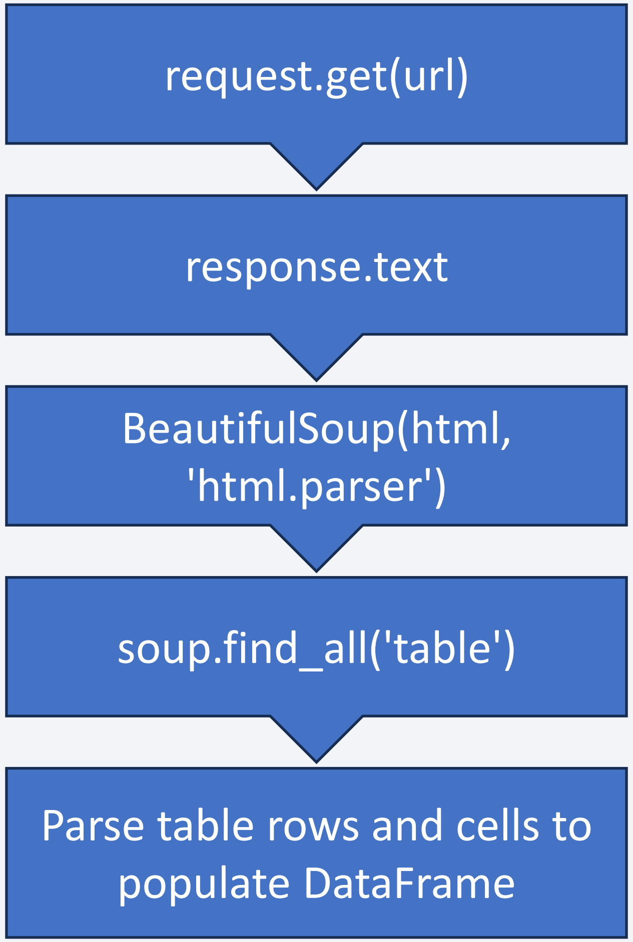
https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Initiate a request to the URL
- Extract text content
- Create a BeautifulSoup object
- Find the required table
- Find table instances
- Parse rows and cells to fill a DataFrame

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/jupyter-labs-webscraping.ipynb



request.get(url)

response.text

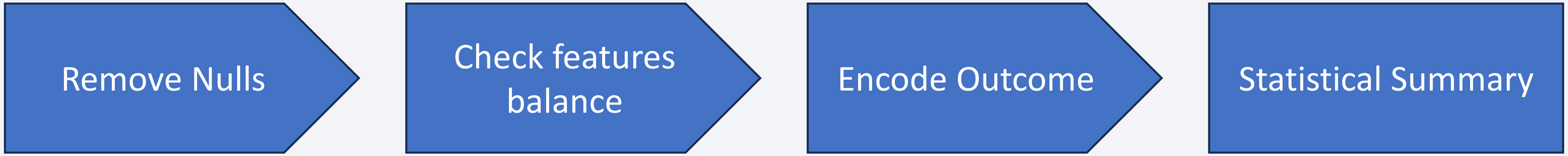
BeautifulSoup(html, 'html.parser')

soup.find_all('table')

Parse table rows and cells to populate DataFrame

Data Wrangling

- Rows with missing values were removed
- Value counts for launch sites and orbits were viewed to check for data imbalances
- Landing outcome unique values were encoded as 0 and 1 for future analysis and model training.
- Statistical summary was produced



Remove Nulls

Check features balance

Encode Outcome

Statistical Summary

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

The following charts were produced:

- Flight number vs Payload mass with outcome overlay:
  - Effect of payload mass on outcome over time
- Flight number vs Launch site with outcome overlay:
  - Effect of launch site on outcome over time
- Payload mass vs Launch site with outcome overlay:
  - The interaction of payload and launch site and its effect on the outcome
- Success rate by Orbit type bar chart:
  - Identify which orbits have the highest success rates
- Flight number vs Orbit type with outcome overlay:
  - Notice how some orbit types had a relation to flight number, while others did not
- Payload mass vs Orbit type with outcome overlay:
  - The interaction of payload mass and different orbit types
- Yearly Success rate plot:
  - Observe the trend in yearly success rate

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/edadataviz.ipynb

EDA with SQL

- EDA SQL Queries:
    - Display the names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA’
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1
    - List the date when the first successful landing outcome in the ground pad was achieved
    - List the names of the boosters that have been successful in drone ship and have a payload mass greater than 4000 and less than 6000
    - List the total number of successful and failed mission outcomes
    - List all the booster versions that have carried the maximum payload mass
    - Display the month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in the year 2015
    - Rank landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Color-coded location marker clusters, circles, name markers, lines, and distance markers were created and displayed on a Folium map.
- Color-coded marker clusters make it easy to visualize successful and failed launches on the map for overlapping launch sites.
- Circles and names help spot the main launch sites on the map.
- Lines and distance markers show proximity of launch sites to main features, like coastlines, highways, railroads, and main cities.

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/lab_jupyter_launch_site_location.ipynb

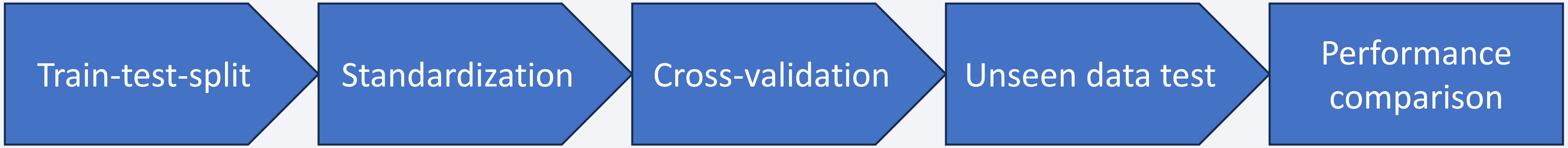
Build a Dashboard with Plotly Dash

- Graphs added to the interactive dashboard:
  - Pie chart of the success rate by launch site, or for each site individually.
  - Scatter plot of the payload mass against outcome with booster version overlay for the selected payload range.
- The pie chart helps visualize the success rate for launch sites and check the success rate for each site individually, while the scatter plot helps visualize the relation between the payload mass and success rate and compare different ranges of payload mass in an interactive way.

https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Data was split into training and test sets (20%). Each part was standardized separately to avoid data leakage.
- Four models were trained on training data using cross-validation with a set of hyperparameters.
- The models were tested on the test set, and their accuracy scores were compared.



Train-test-split

Standardization

Cross-validation

Unseen data test

Performance comparison

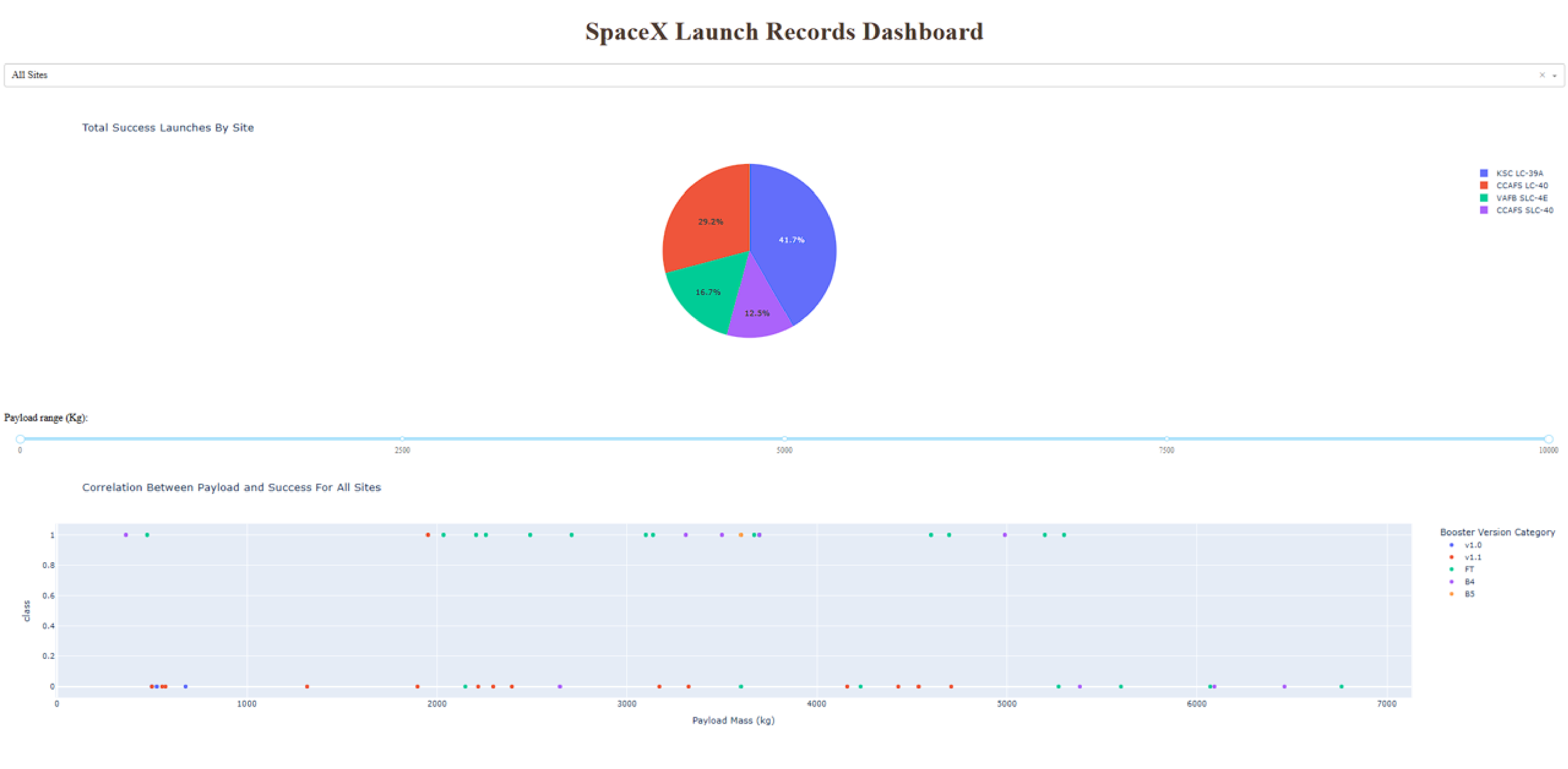
https://github.com/MohamadAlhaidar/coursera_ibm_data_science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - Flight number, Payload mass, Launch site, and Orbit type may be good predictors for the success of landings, as it appears in the graphs and plots.
  - The success rate of landings is increasing by year.
  - The success rate for drone ship landings is the highest, followed by ground pad landings.
  - The three main launch sites are located on the east and west coasts of the US, with proximity to the coastline, highways, and railways.
  - Launch site KSC LC-39A has the highest success rate.
  - Payload mass range 2500-5000 has the highest success rate.
  - Booster version B5 has the highest success rate, although it had only one launch attempt.

Results

- Interactive analytics demo screenshots:



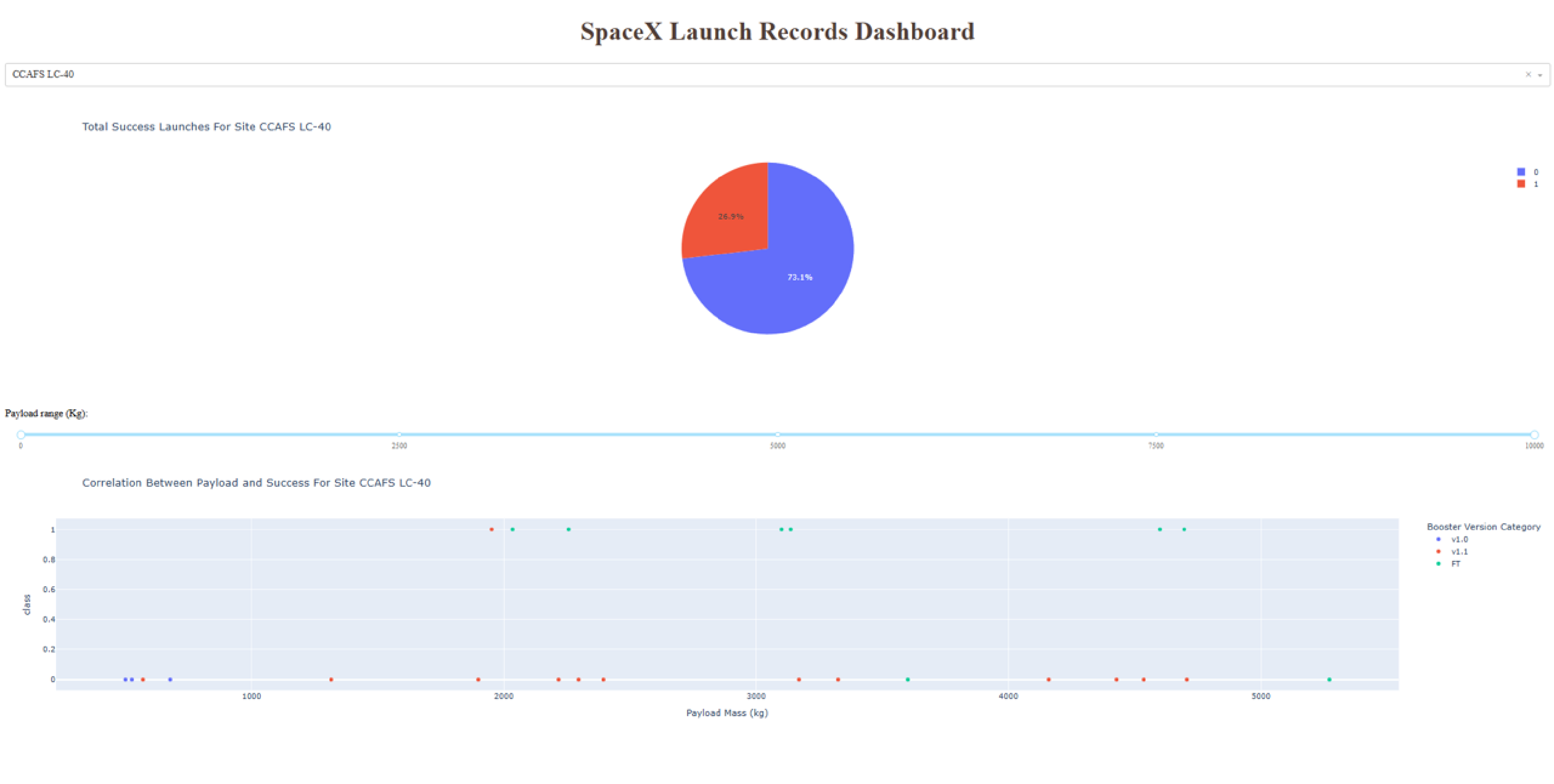
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Results

- Interactive analytics demo screenshots:



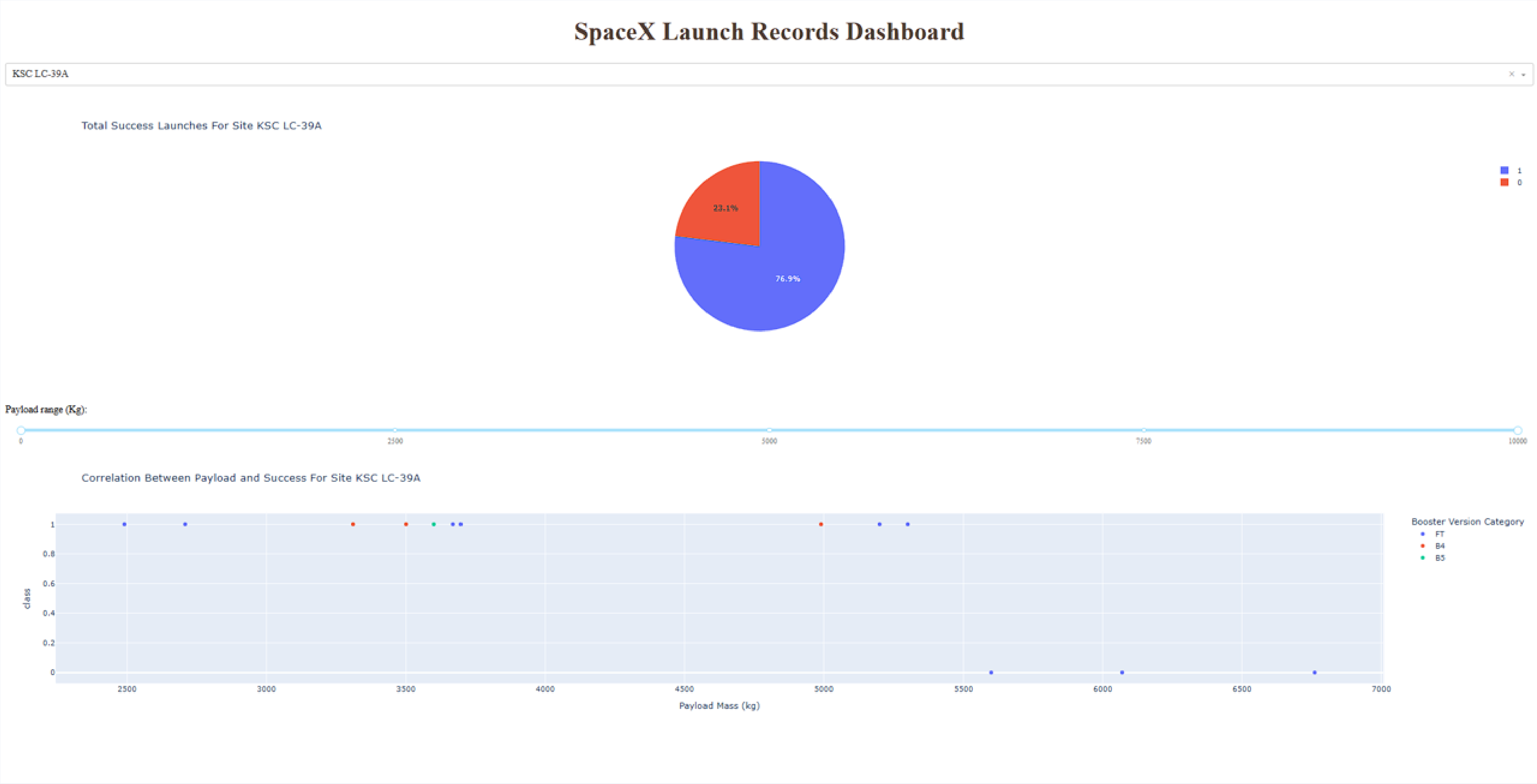
SpaceX Launch Records Dashboard

CCAFS LC-40

Payload range (Kg):

Results

- Interactive analytics demo screenshots:



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches For Site KSC LC-39A

Payload range (Kg):

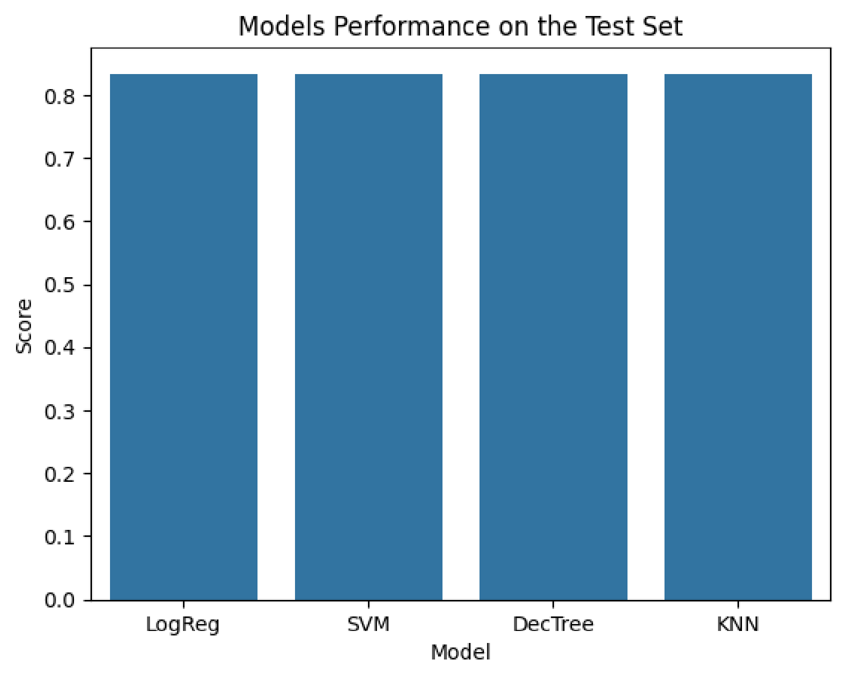
Correlation Between Payload and Success For Site KSC LC-39A

Results

- Predictive analysis results:

  - Four machine learning models were trained on the final dataset, namely, Linear Regression Classifier, Support Vector Machine Classifier, Decision Tree Classifier, and K-Nearest Neighbors Classifier.

  - The Decision Tree Classifier showed higher accuracy on the training data. However, all four models showed similar performance on the test data with an accuracy of 83.33%.

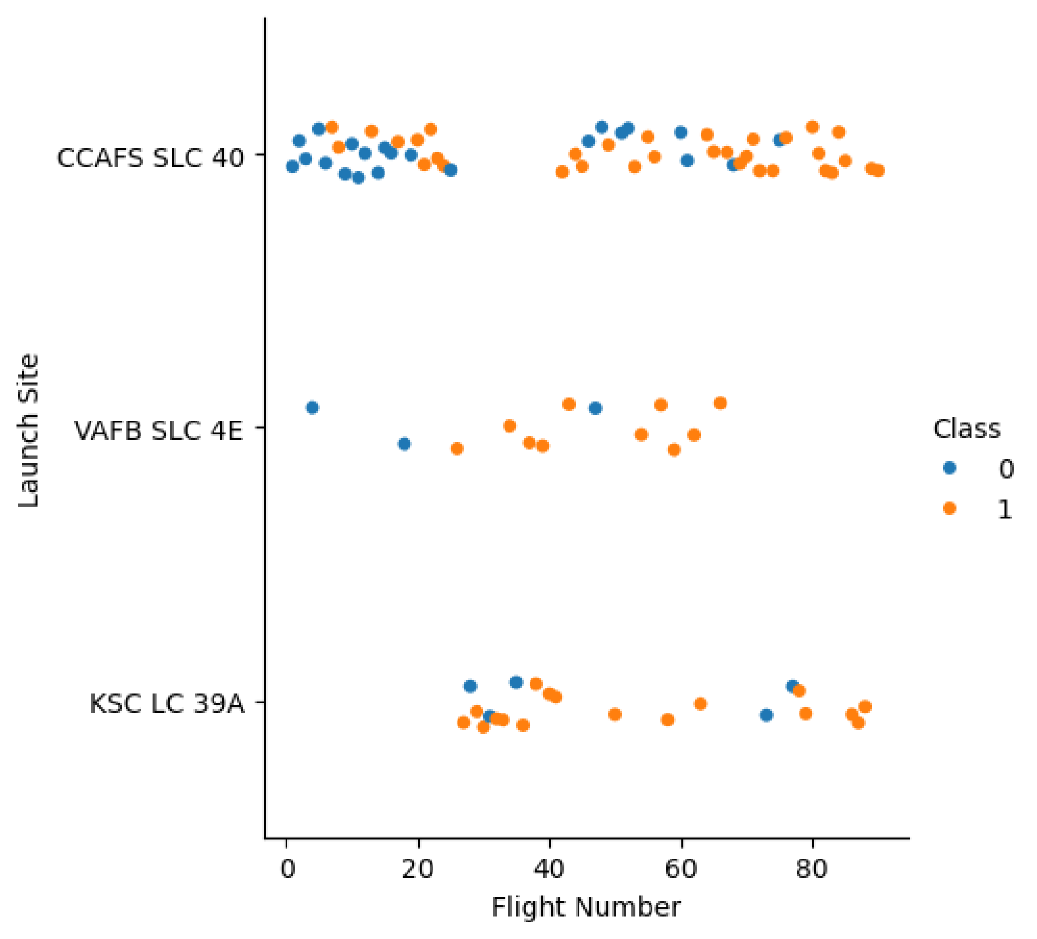


Section 2

Insights drawn from EDA

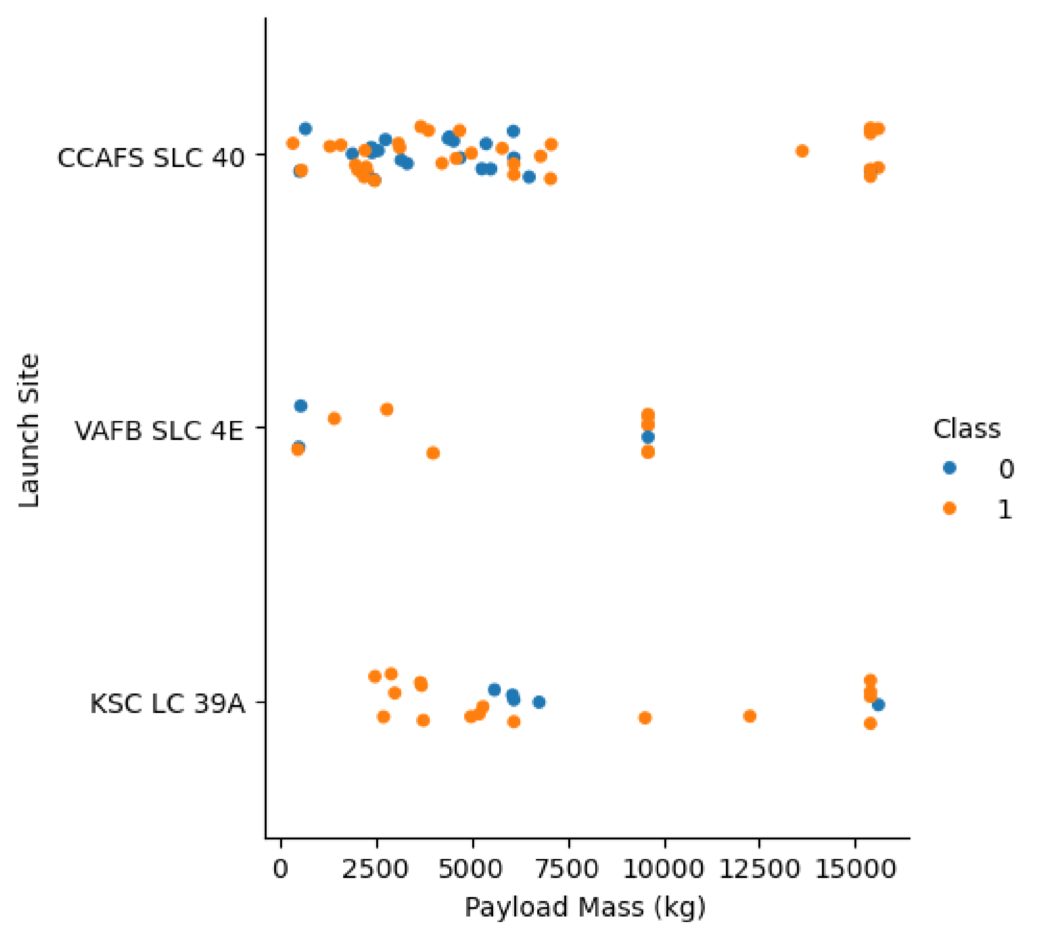
Flight Number vs. Launch Site

- The plot shows that KSC LC-39A has the highest success rate among launch sites.
- Launches at VAFB SLC-4E seem to have stopped after about 70 launches.
- CCAFS SLC-40 has the most launches, with many failed landings earlier and an increase in success rate later.



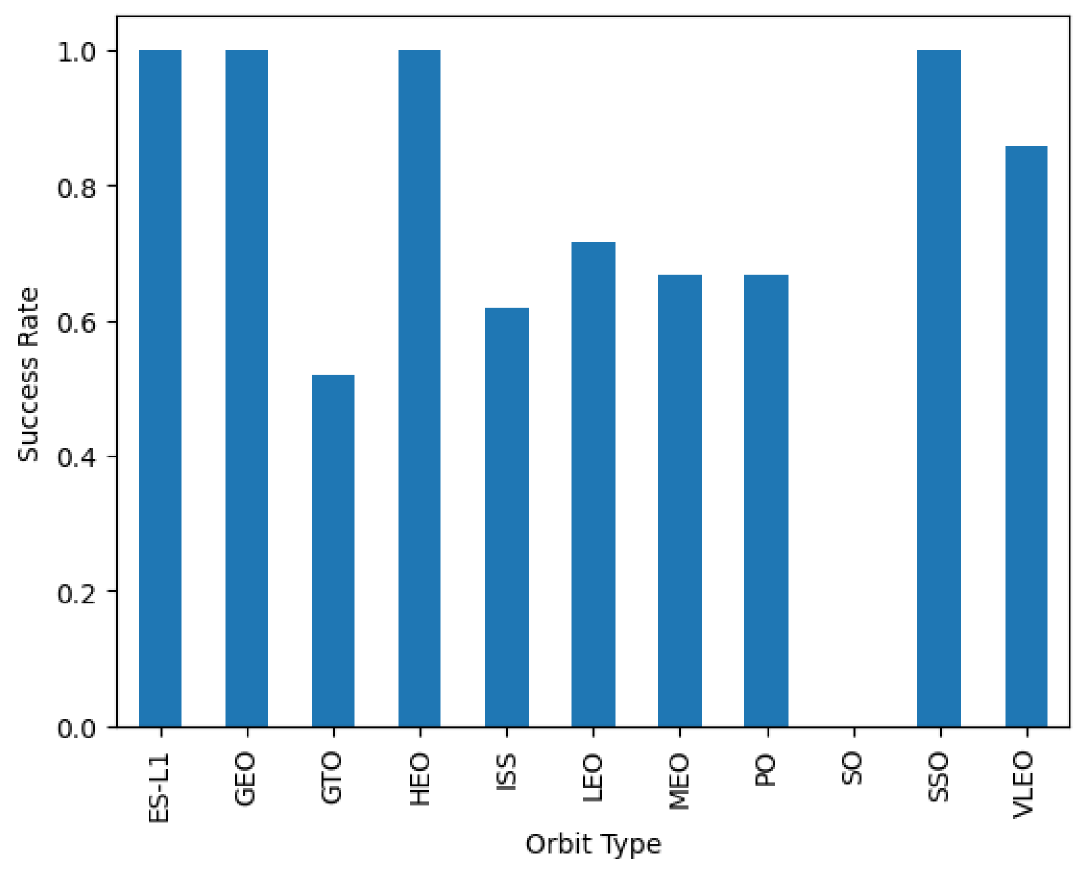
Payload vs. Launch Site

- All landings at CCAFS SLC-40 with payload masses higher than 12500 kg were successful.
- VAFB SLC-4E has no launches with payload masses higher than 10000 kg.
- In general, launches with high payload masses had a higher success rate.



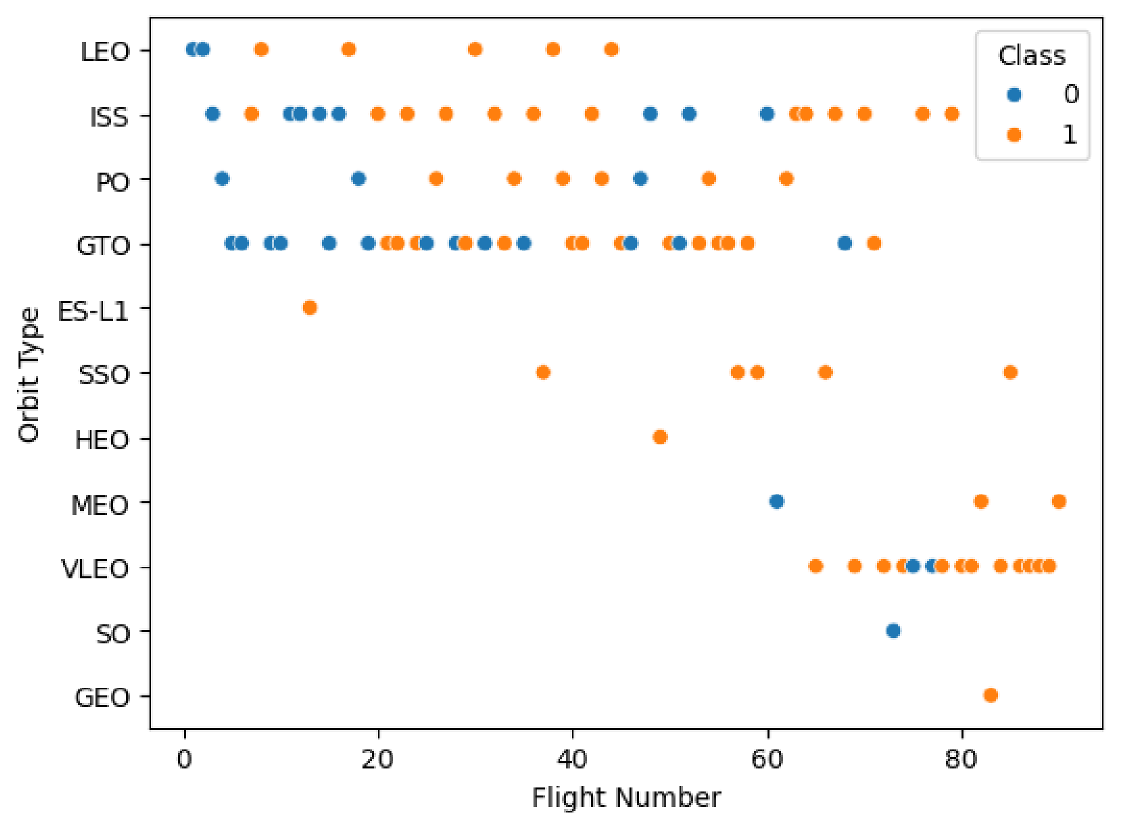
Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO, and SSO have the highest success rates.
- GEO orbits have a relatively lower success rate.
- SO orbits have zero success rate.



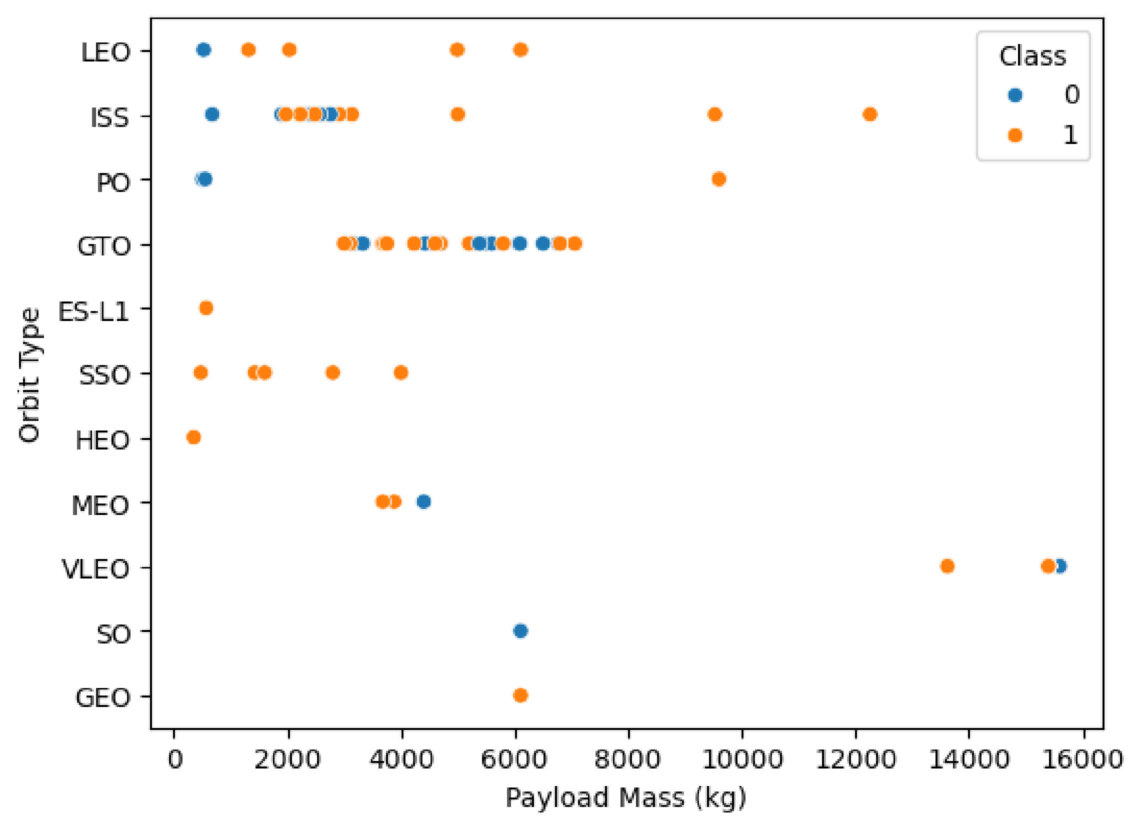
Flight Number vs. Orbit Type

- The orbit types seem to have shifted from LEO, ISS, PO, GTO, and ES-L1, to include the orbits SSO, MEO, and VLEO in later flights.
- The number of flights for some orbits has decreased or stopped, and increased significantly for VLEO.



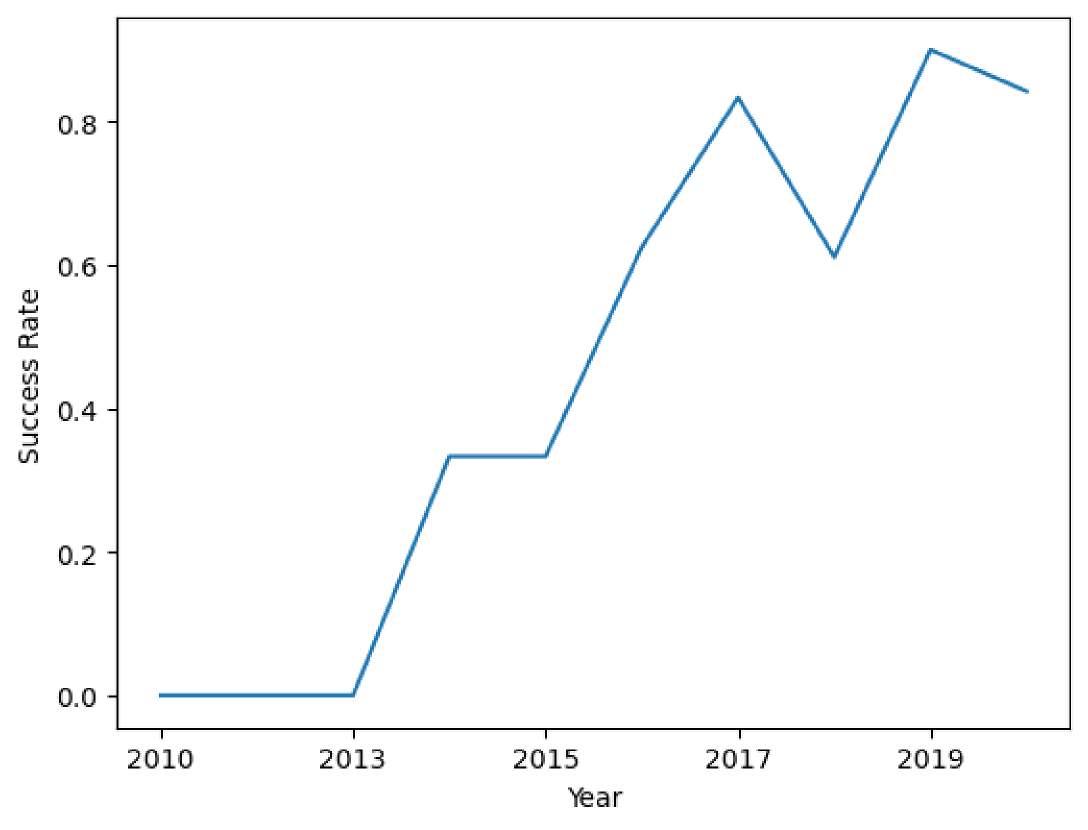
Payload vs. Orbit Type

- PO, LEO, and ISS orbits have a higher success rate with heavier payloads.
- GTO orbits have successes and failures.
- SSO orbits included only lighter payloads but had a high success rate.



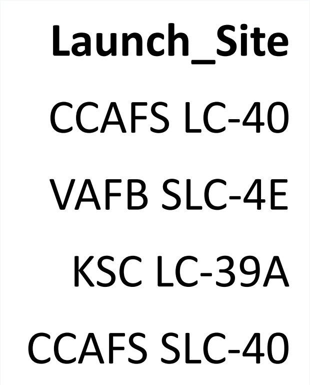
Launch Success Yearly Trend

- The yearly success rate for Falcon 9 landings shows a rising trend over time.
- The success rate started to improve after 2013.



All Launch Site Names

- Unique launch site names:

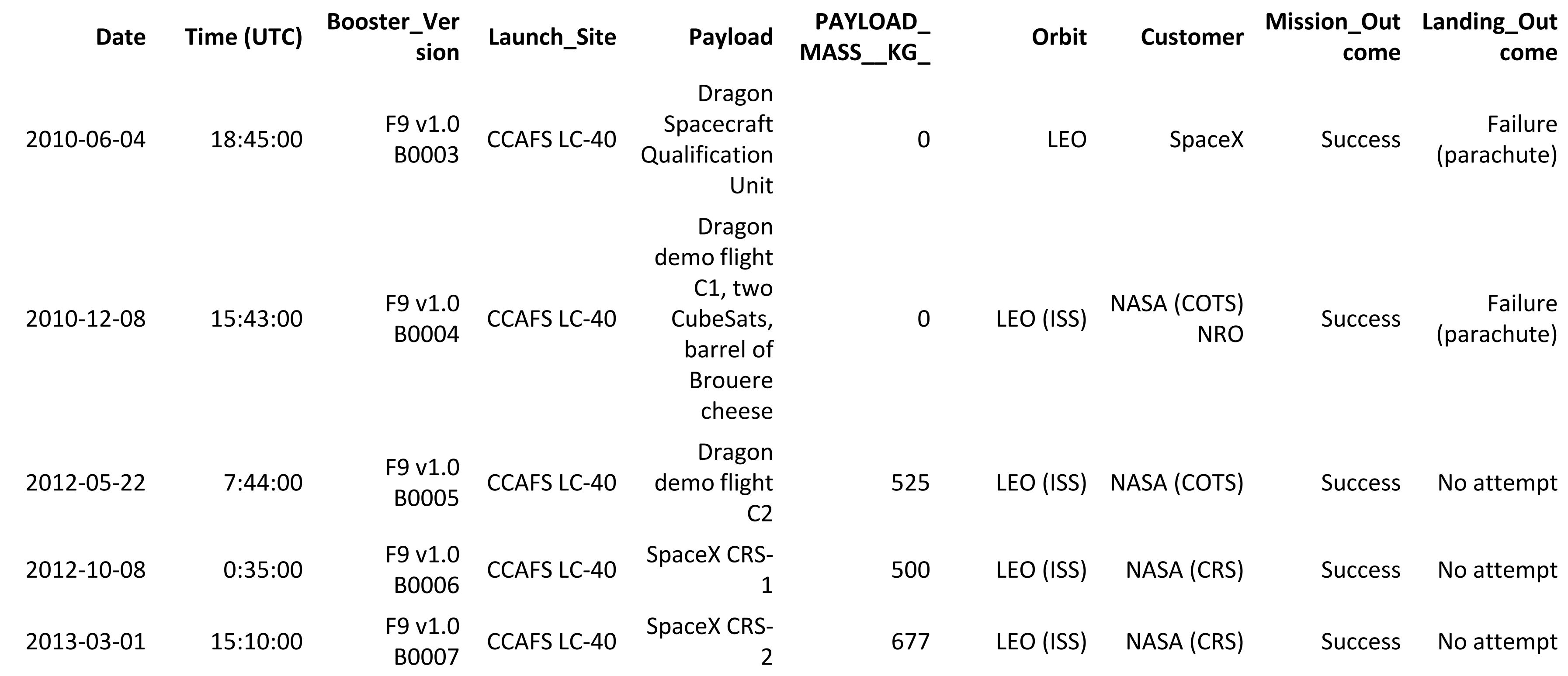


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- A SQL query was used to retrieve the unique launch site names from the database.

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`:

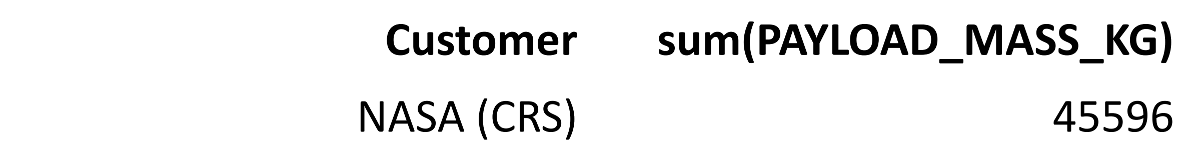


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Above are the first 5 records of the result set that includes launch sites beginning with ‘CCA’.

Total Payload Mass

- Total payload carried by boosters from NASA

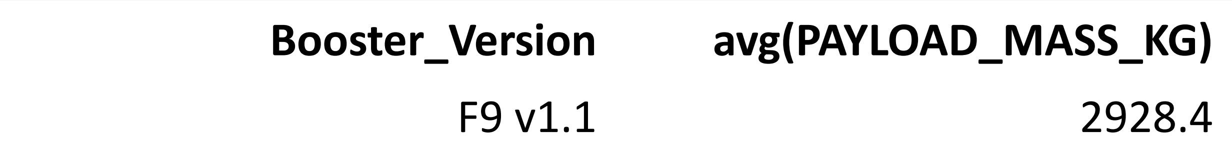


Customer	sum(PAYLOAD_MASS_KG)
NASA (CRS)	45596

- The sum was obtained by grouping the set by Customer and summing the payload mass values.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

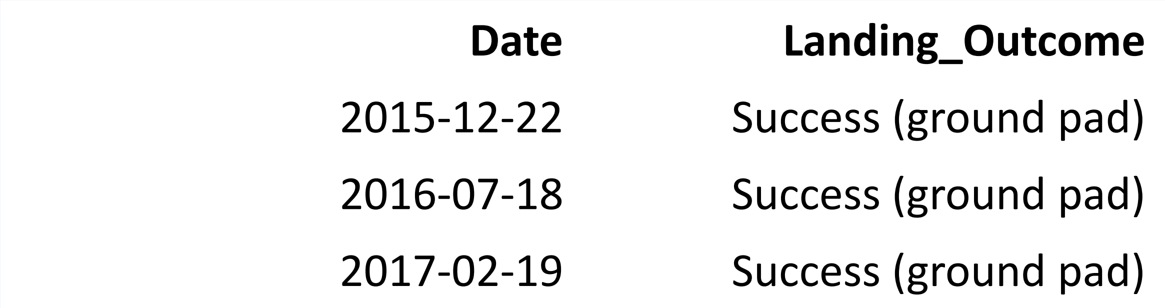


Booster_Version	avg(PAYLOAD_MASS_KG)
F9 v1.1	2928.4

- The value was obtained by grouping by Booster Version and displaying records having Booster Version F9 v1.1.

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pads.

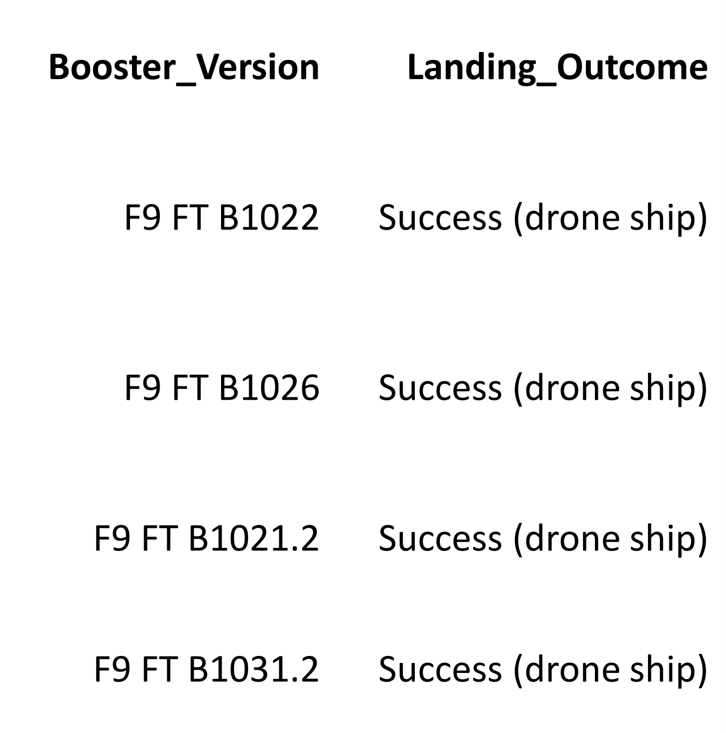


Date	Landing_Outcome
2015-12-22	Success (ground pad)
2016-07-18	Success (ground pad)
2017-02-19	Success (ground pad)

- The date and landing outcome were extracted for records that had a success on ground pads.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters that have successfully landed on a drone ship and had a payload mass greater than 4000 and less than 6000:

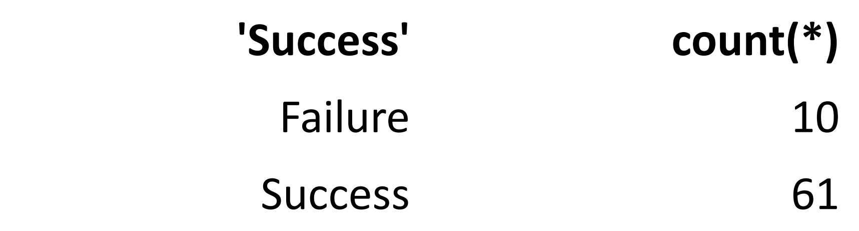


Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

- The Booster version and landing outcome were selected with a condition to limit the outcome to ‘Success (drone ship).

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failed mission outcomes:



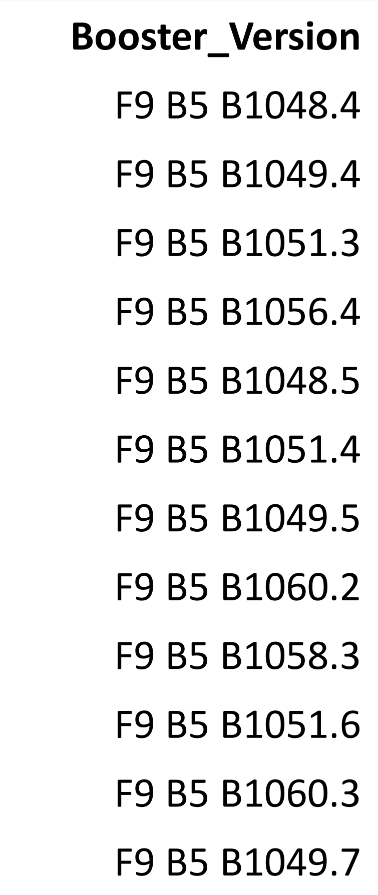
'Success'	count(*)
Failure	10
Success	61

- The counts were obtained using a union of the count of success (all types) and failure (all types) outcomes.

Boosters Carried Maximum Payload

- Names of the boosters that have carried the maximum payload mass:

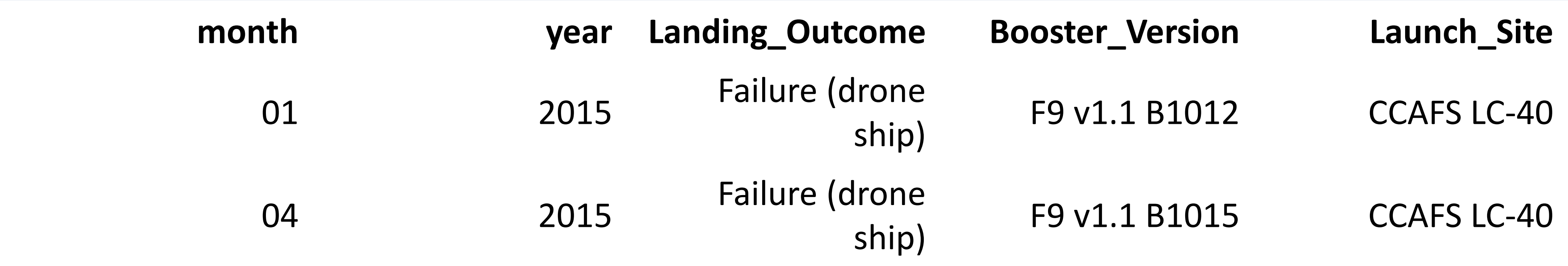
- The result was obtained using a subquery to specify the records with the maximum payload mass.



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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015:

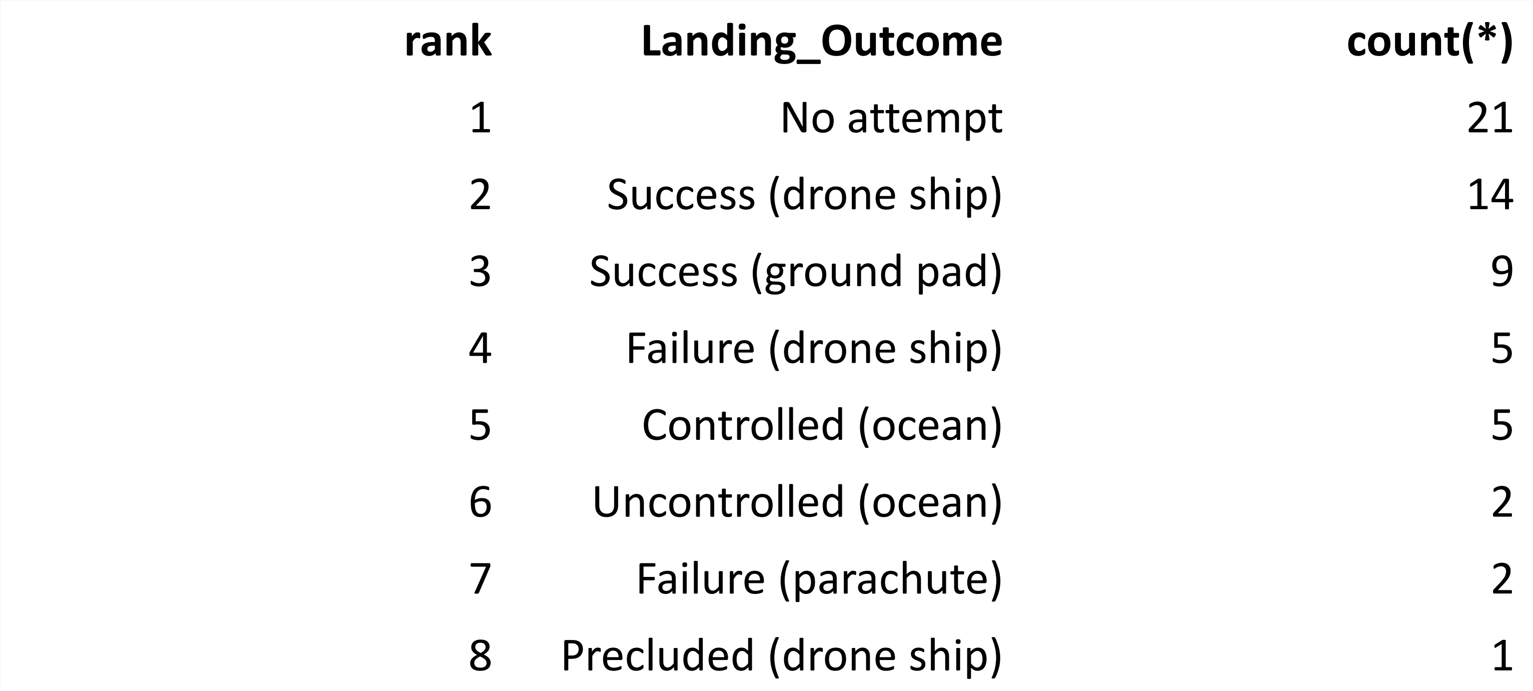


month	year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The function ‘substr()’ was employed in the query to extract the month and year out of the date column, in addition to grouping by landing outcome.

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

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



rank	Landing_Outcome	count(*)
1	No attempt	21
2	Success (drone ship)	14
3	Success (ground pad)	9
4	Failure (drone ship)	5
5	Controlled (ocean)	5
6	Uncontrolled (ocean)	2
7	Failure (parachute)	2
8	Precluded (drone ship)	1

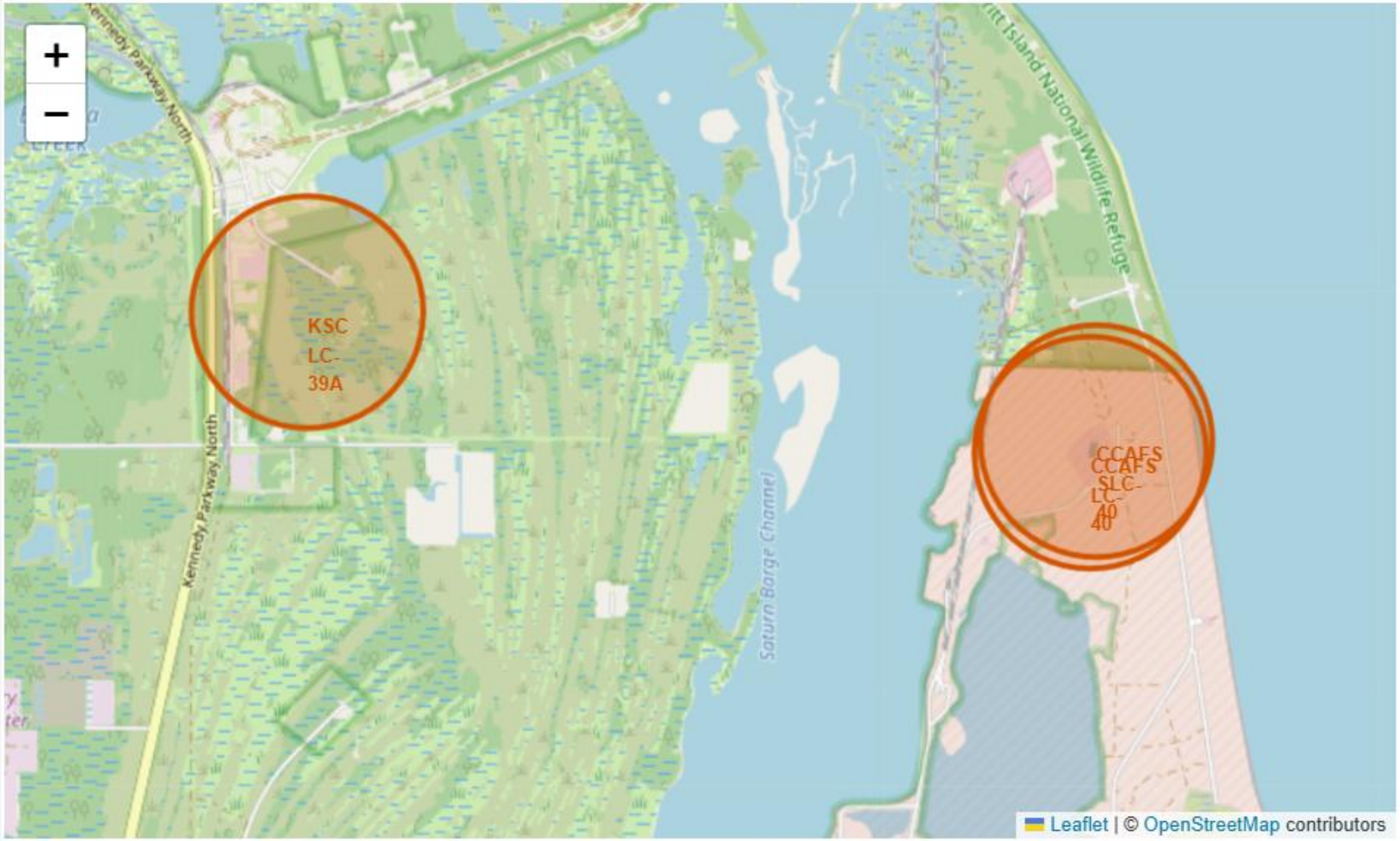
- The phrase ‘ROW_NUMBER() OVER (ORDER BY count(*) desc)’ was used to produce a ranking of the landing outcome.

Section 3

Launch Sites Proximities Analysis

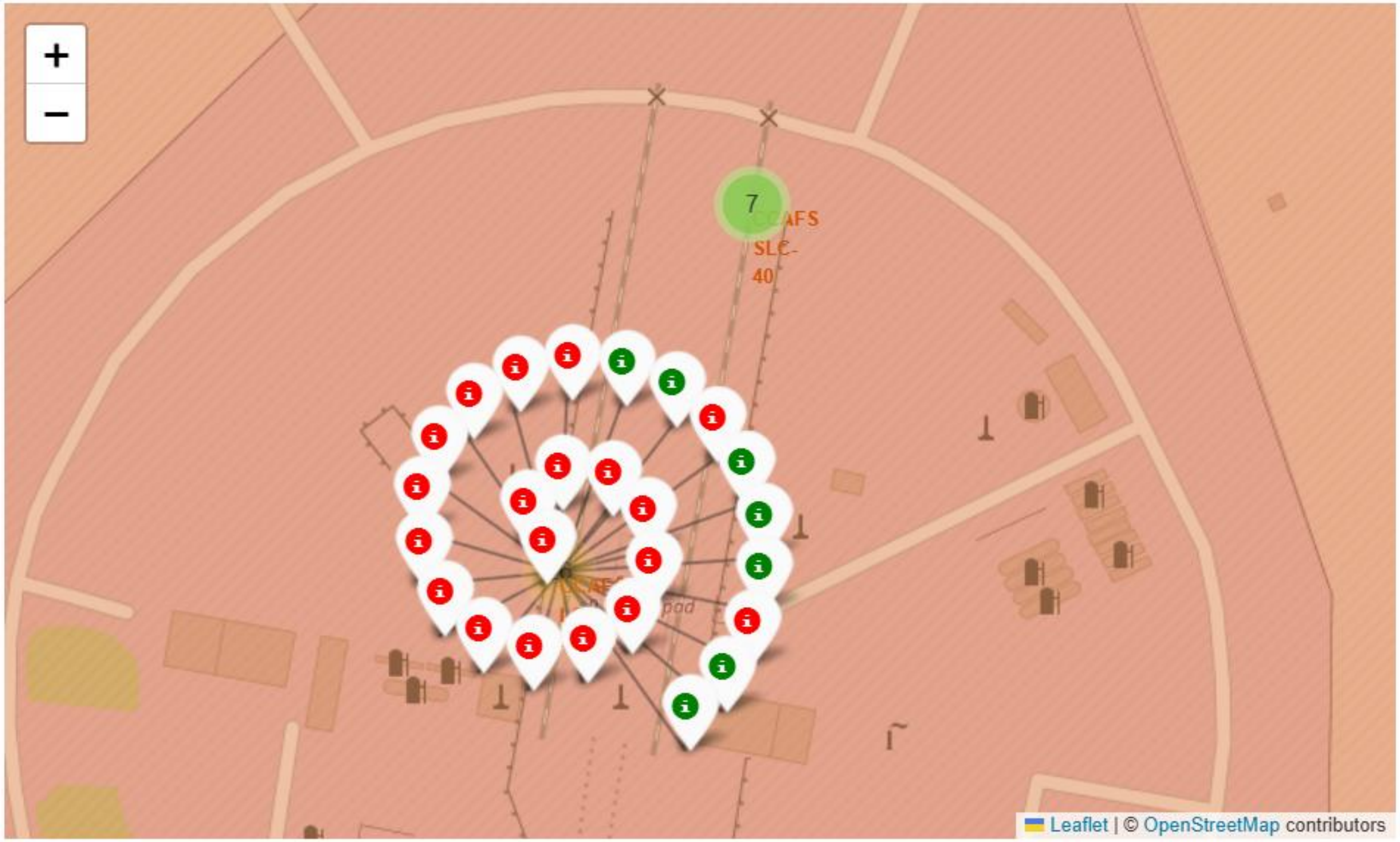
Launch Sites Location Markers

- Name markers and circles make it easier to spot the launch sites on the map.



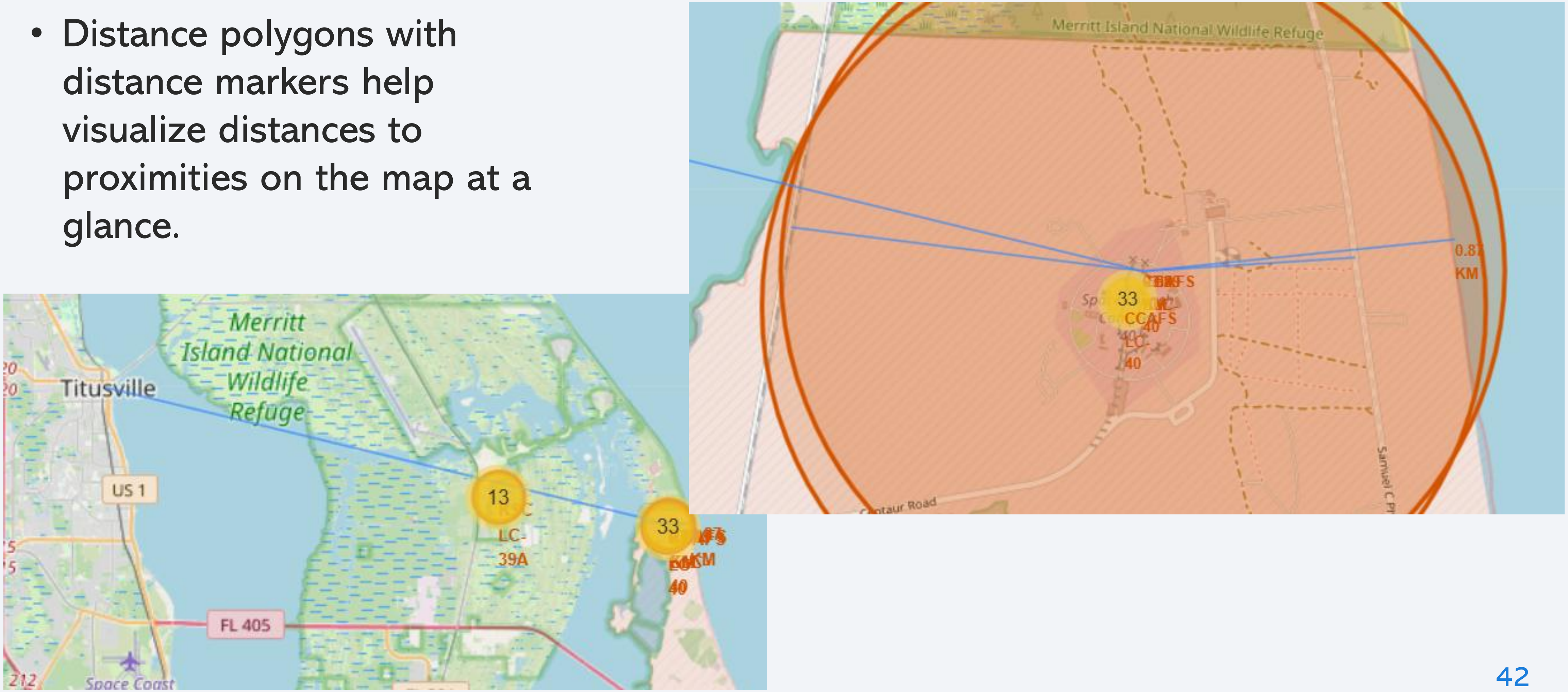
Color-Labeled Clustered Launch Outcomes

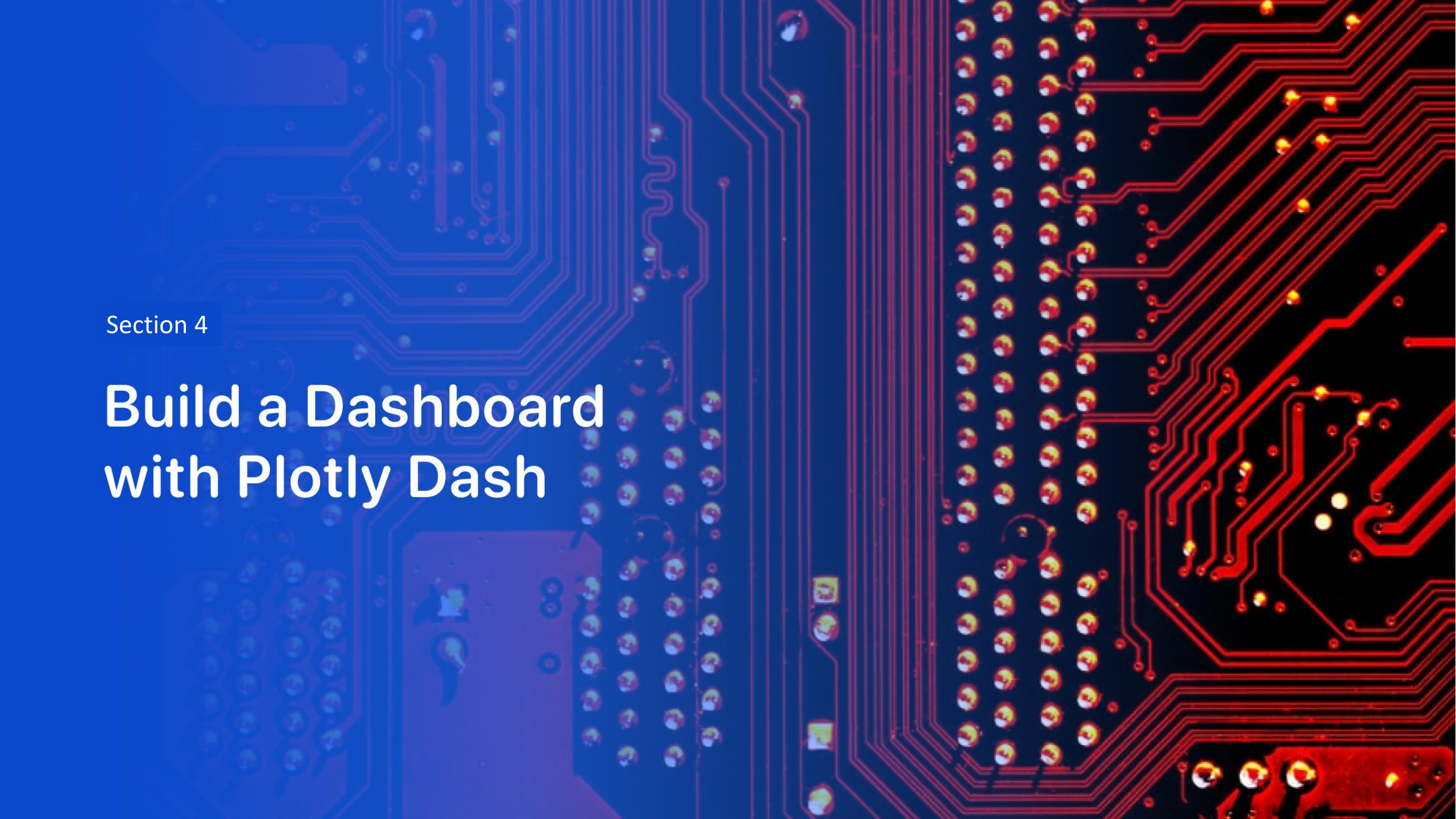
- Color-labeled outcome marker clusters help distinguish successful and failed landings with overlapping locations.



Proximities and Distance Markers

- Distance polygons with distance markers help visualize distances to proximities on the map at a glance.

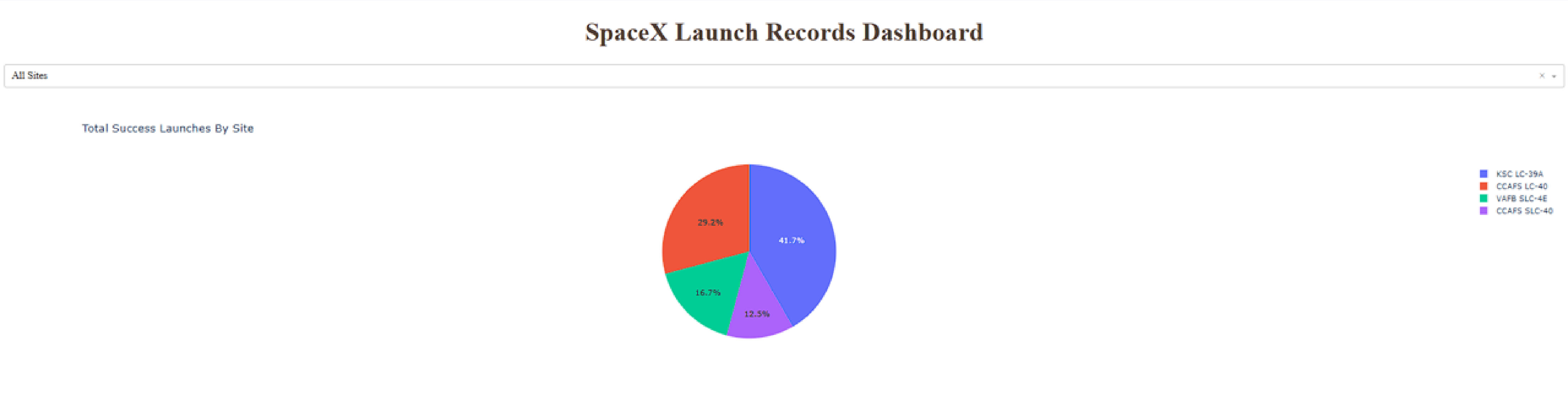




Section 4

Build a Dashboard with Plotly Dash

Success Rate for All Sites

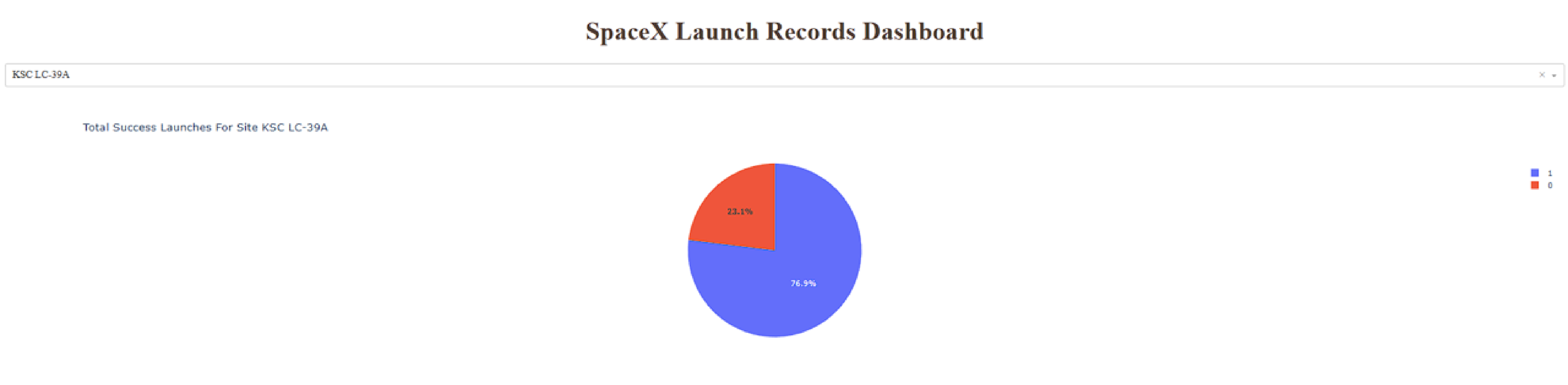


SpaceX Launch Records Dashboard

All Sites

- The pie chart displays the success rate for all launch sites.
- KSC LC-39A has the highest success rate (41.7%).
- CCAFS SLC-40 has the lowest success rate (12.5%).

Success vs Failure for the Most Successful Site



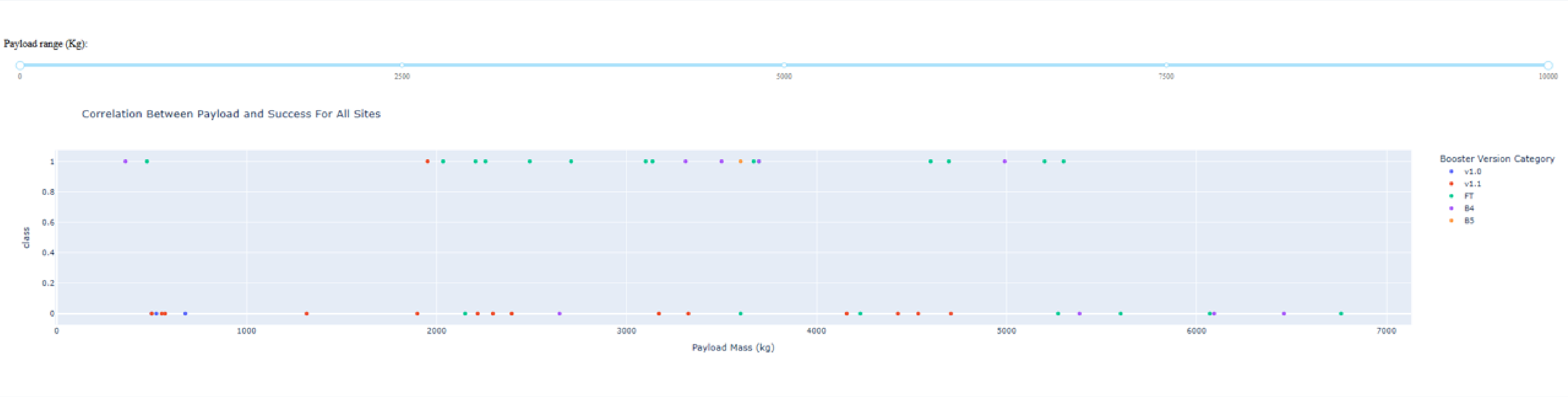
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches For Site KSC LC-39A

- Launch site KSC LC-39A has the highest success rate.
- 76.9% of its landings were successful.
- 23.1% of its landings were a failure.

Payload vs Launch Outcome for All Sites



- The scatter plot shows success and failure for different payload masses, with the booster version displayed as the color.

- B5 has the highest success rate (100%), but had only one launch.

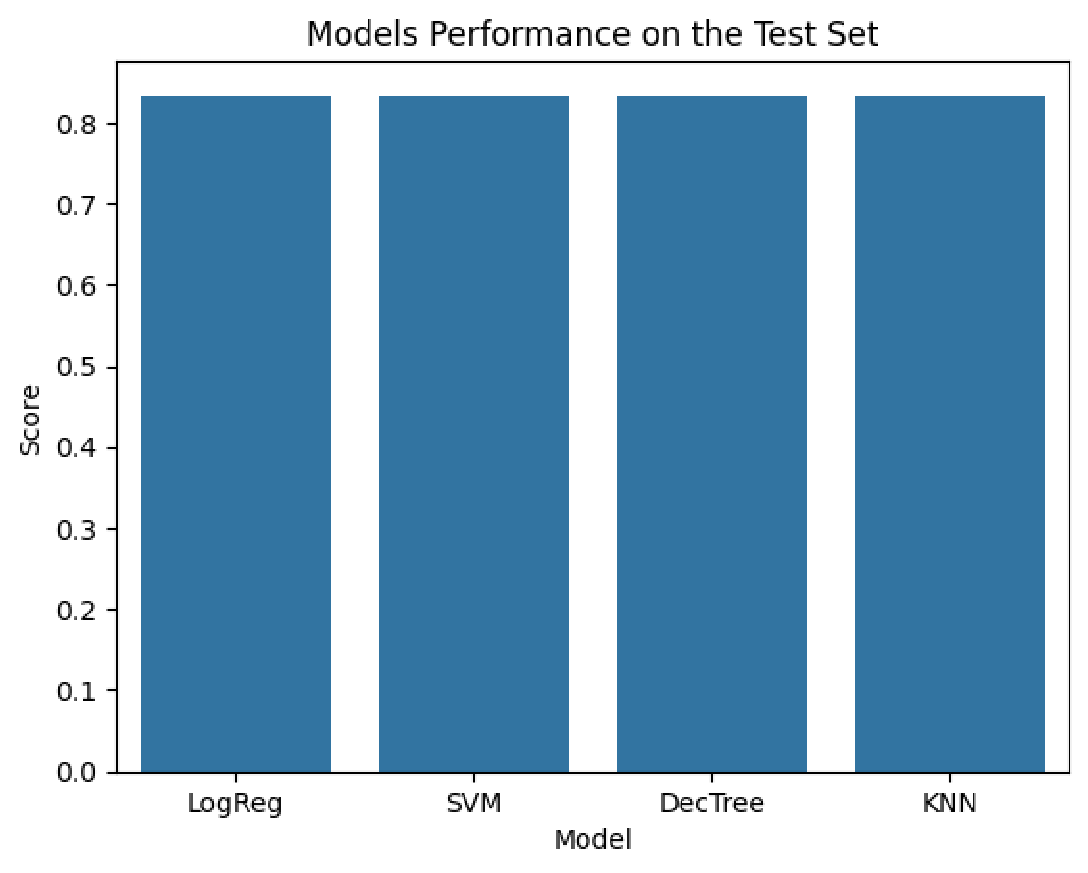
- FT has a (65%) success rate with 13 successful landings out of a total of 20.

Section 5

Predictive Analysis (Classification)

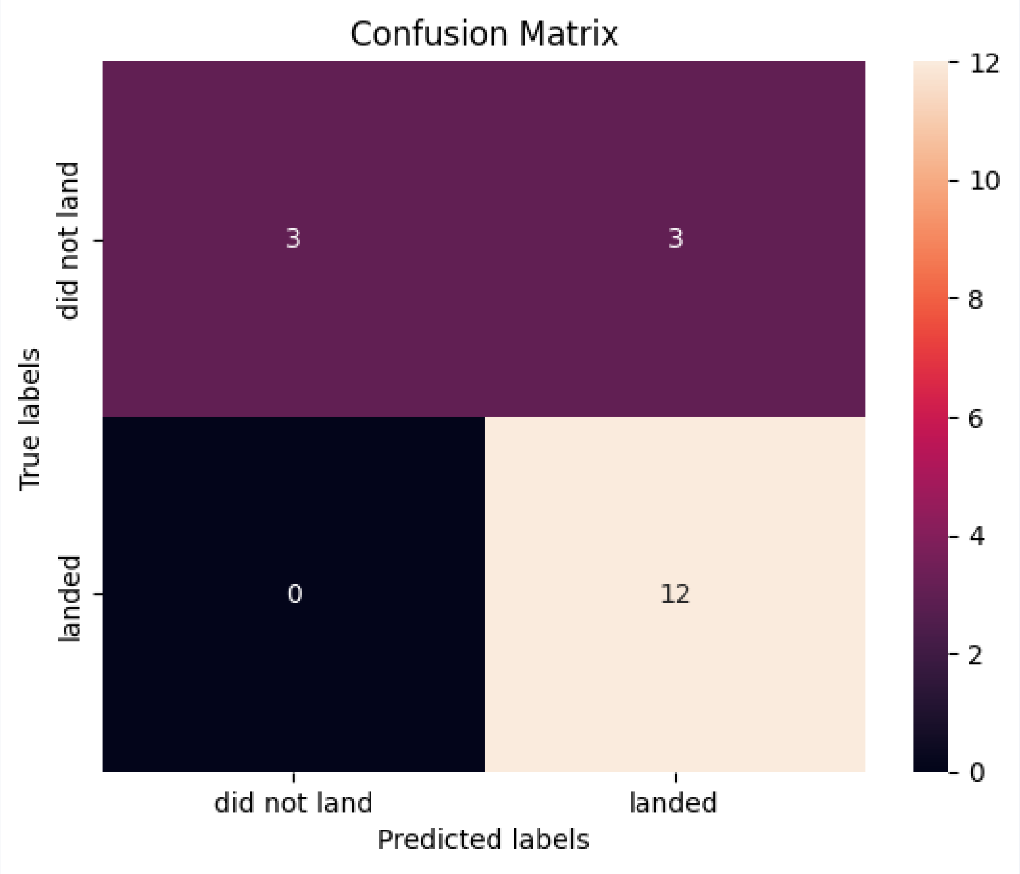
Classification Accuracy

- All four classification models (Logistic Regression, SVM, Decision Tree, and KNN) had similar accuracy scores when tested on test data.
- The accuracy score was 83.33%.
- The decision tree model had a slightly higher score when tested on training data.



Confusion Matrix

- The confusion matrix for the best model shows a perfect true positive score with 12 positive outcomes labeled as true.
- The model falsely predicted three negatives as true while correctly labeling three negatives.

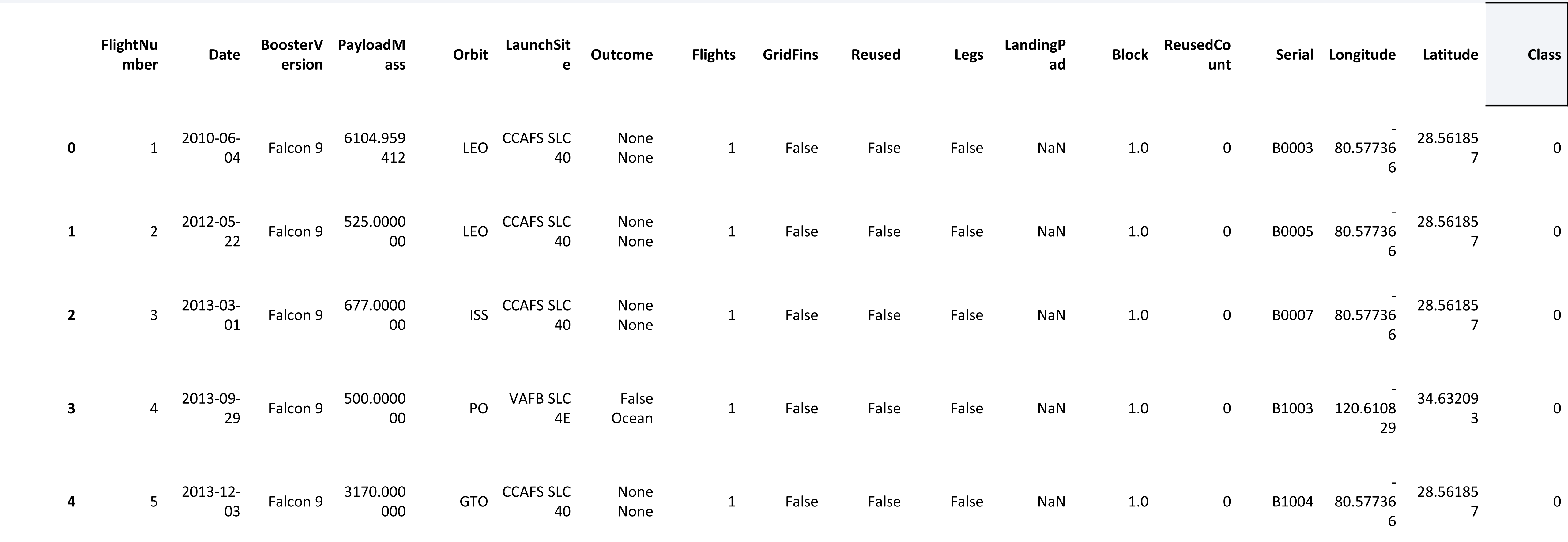


Conclusions

- The data was collected from several sources and integrated into a useful dataset.
- Several variables were identified as candidates for model development.
- Analysis shows a trend of increasing success rate for Falcon 9 first stage landing and reuse over time.
- A machine learning model was developed to predict future landing outcomes with reasonable accuracy.

Appendix

- First few records of the cleaned dataset:



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Thank you!