IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Samir Paul
October 16, 2025

Outline

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

Executive Summary- SpaceX Falcon 9 First-Stage Landing Prediction Modeling

- The machine learning pipeline developed to predict SpaceX Falcon 9 first-stage landing success analyzed four models—Logistic Regression, Support Vector Classifier (SVC), Decision Tree, and K-Nearest Neighbors (KNN)—using a dataset with 83 features, including payload mass, orbit type, and launch site. The SVC model outperformed others with the highest test accuracy of 0.833 and only three incorrect classifications, compared to 0.778 test accuracy and four misclassifications for Logistic Regression, Decision Tree, and KNN. SVC also demonstrated balanced performance with a training accuracy of 0.8482, indicating minimal overfitting compared to Decision Tree's larger gap (training: 0.9031, test: 0.778). The robustness of SVC in handling high-dimensional data and capturing non-linear relationships via kernel functions (e.g., RBF) makes it the most reliable model for predicting landing outcomes, as validated through GridSearchCV hyperparameter tuning and 10-fold cross-validation.
- These findings can be leveraged in aerospace and space launch industries as we targetted for SPACE Y to optimize mission planning and cost estimation. By accurately predicting first-stage landing success, companies like SpaceY or competing providers can estimate launch costs more precisely, given SpaceX's reusable first stage reduces costs to $62 million compared to competitors' $165 million. This predictive model can guide resource allocation, inform bidding strategies for launch contracts, and enhance risk assessment for mission-critical decisions. Beyond SpaceY, similar machine learning approaches can be applied to other high-stakes industries, such as autonomous vehicles or satellite operations, where predicting operational outcomes based on complex, multi-dimensional data (e.g., sensor readings, environmental factors) is crucial for cost efficiency, safety, and strategic decision-making.

Introduction

- Background: The first stage of the SpaceX Falcon 9 rocket, powered by nine Merlin 1D engines generating up to 1.7 million pounds of thrust using liquid oxygen and rocket-grade kerosene, ignites at liftoff to propel the vehicle through Earth's dense lower atmosphere, reaching an altitude of about 70-80 km in roughly 2.5 minutes before separating from the upper stage. Its reusability is what makes it revolutionary, as it performs a controlled descent with reentry burns, grid fins for steering, and precise landings on drone ships or ground pads, enabling over 300 successful recoveries and drastically reducing launch costs. Nonetheless, despite putting best efforts, there are unsuccessful launches too.
- Objective: To develop a supervised machine learning model to predict the binary success of the Nova Streak*,* the edgier space craft of *SPACE Y*, the dream venture of Allon Mask first stage launch (1 for successful ascent and recovery/landing, 0 for failure or expendable without recovery attempt) using a dataset of historical launch records of Falcon 9 to explore the possibility of the explosive rebirth of a reusable rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: We used two sources for the data
  - SpaceX API and
  - Wikipedia web scrapping
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

As mentioned above, we extracted the following features for Falcon 9 launches using Space

- Rocket,
- Payloads
- Launch Pad
- Cores
- Flight Number and
- Launch Date

Data Collection – SpaceX API

A five-stage framework was followed to extract the data using REST APIs for launchpads, payloads, cores and landing types as follows:

Steps of Data Extraction

-Initial Data Fetch (Raw Launch Records)
-Enrichment via Secondary API Calls (ID- Based Lookups)
-Dataset Assembly – combined to a data dictionary and filtered for Falcon 9
-Data Cleaning and Handling Missing Values and
-Output and Export

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Falcon 9 launch data was also scrapped from the Wikipedia page for Falcon 9 launches, dated June 9, 2021. The steps have been outlines in the flowchart.

Web Scrapping of Wikipedia Page

- **Setup and Page Fetch**: largely used requests and BeautifulSoup packages for extracting the data from the url.
- **Column Name Extraction**: used a helper method to extract the column names
- **Row Data Extraction and Parsing**: iterated over the html table rows
- **DataFrame Creation and Export**: build a pandas dataframe to save to csv data

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

LOAD & INSPECT DATA USING PANDAS TO VIEW STRUCTURE

- EDA ON OUTCOME COLUMN TO CHECK UNIQUE VALUES OF KEY FEATURES, VISUALIZE INCLUDING A BAR CHART OF OUTCOME DISTRIBUTION AND FILTER & PLOT SUCCESSFUL LANDINGS (ASDS/RTLS/OCEAN) FOR PATTERNS.

- TARGET LABEL ('CLASS') - DEFINED FAILURES AND SUCCESS

- EXPORT CLEANED DATA WITH LABELED DF READY FOR MODELING (CLASS: 1=SUCCESS, 0=FAILURE).

https://github.com/samizard2016/coursera-ds/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization of Falcon 9 Launch data

- Cumulative Success Rate Over Time

- Landing Outcome Distribution: Pie chart visualizing the proportion of successful

- Success Rate by Launch Site: Grouped bar plot comparing landing success rates across sites (- Success Rate by Orbit Type: Bar plot of success rates for different orbits (e.g., LEO, ISS, GTO, VLEO), showing higher success in low-Earth orbits like LEO/ISS.

- Payload Mass vs. Success: Scatter plot of `PayloadMass` (kg) against `Class` (0/1), with a line for mean payload, revealing no strong correlation but denser failures at higher masses.

- Success Rate by Block Version: Bar plot comparing success rates across Falcon 9 block versions

- Success Rate by Grid Fins: Bar plot showing success rates with/without `GridFins` (True/False), emphasizing higher success when grid fins are deployed.

- Success Rate by Reused Booster: Bar plot of success rates for reused (`Reused=True`) vs. new (`False`) boosters

- Success Rate by Landing Legs: Bar plot comparing success with/without `Legs` (True/False), highlighting the role of legs in ground landings.

https://github.com/samizard2016/coursera-ds/blob/main/edadataviz_sp.ipynb

EDA with SQL

- Table Setup Query: `DROP TABLE IF EXISTS SPACEXTABLE;`

- Table Creation Query: `create table SPACEXTABLE as select * from SPACEXTBL where Date is not null`

Task 1 Query: `select DISTINCT Launch_Site from SPACEXTABLE;`

- Task 2 Query: `select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;`

- Task 3 Query: `select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)';`

- Task 4 Query : `select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1';`

- Task 5 Query: `select Date from SPACEXTABLE where Landing_Outcome='Success (ground pad)' ORDER BY Date ASC LIMIT 1;`

- Task 6 Query : `select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;`

- Task 7 Query : `select count(*) from SPACEXTABLE where Landing_Outcome LIKE '%Success%' OR Landing_Outcome LIKE '%Failure%';`

- Task 8 Query: `select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_=(Select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE);`

Task 9 Query: `select Date,substr(Date,6,2) as "Month", Landing_Outcome, Booster_Version,Launch_Site from SPACEXTABLE where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015';`

- Task 10 Query: `select Landing_Outcome, count(Landing_Outcome) as Count from SPACEXTABLE where Date BETWEEN date('2010-06-04') AND date('2017-03-20') Group by Landing_Outcome ORDER BY Count DESC;`

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Task	Map Objects	Objective
Task 1: Mark launch sites	Folium Map, Circles, Markers, Tile Layer	Show SpaceX launch site locations
Task 2: Mark success/failed launches	Marker Cluster, Markers	Visualize launch outcomes by site
Task 3: Calculate proximity distances	PolyLine, Marker, MousePosition	Measure distances to coastline, city, etc.

https://github.com/samizard2016/coursera-ds/blob/main/lab_jupyter_launch_site_location.ipynb

Dashboard with Plotly Dash

Plots and Interactions	**Objective**
Pie Chart (success-pie-chart)	Visualize success vs. failure rates for all or selected sites.
Range Slider (payload-slider)	Filter scatter plot by payload mass range.
Scatter Plot (success-payload-scatter-chart)	Explore payload mass vs. launch outcome relationship.

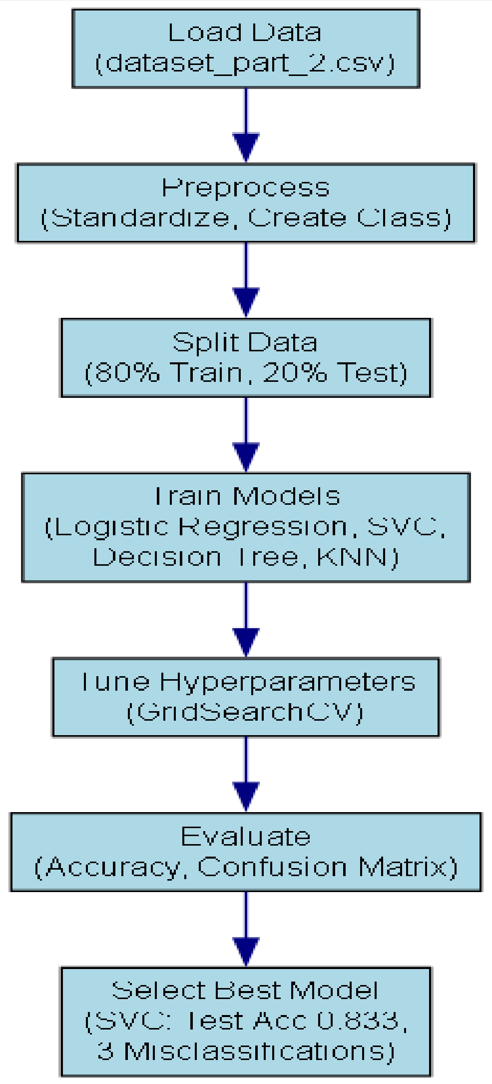
https://github.com/samizard2016/coursera-ds/blob/main/spacex_dashboard1.py

Classification Model Development Summary for SpaceX Falcon 9 First Stage Landing Prediction

Key Steps
- Data Loading: Loaded dataset (`dataset_part_2.csv`) with features like PayloadMass, Orbit, LaunchSite, and Class (landing outcome).
- Preprocessing: Standardized data using `sklearn.preprocessing`. Created Class column (1=landed, 0=did not land).
- Data Splitting: Split into training (80%) and test (20%) sets using `train_test_split` with a stratification on the target variable.
- Model Training: Trained Logistic Regression, SVM (SVC), Decision Tree, and KNN Classifier models.
- Hyperparameter Tuning: Used `GridSearchCV` to find optimal parameters for each model.
- Evaluation: Assessed models using accuracy (training/test) and confusion matrix (via `plot_confusion_matrix`).
- Best Model Selection: Compared test accuracy and incorrect classifications; SVC selected for highest test accuracy (0.833), lowest gap between training and test accuracies and misclassifications (3).

https://github.com/samizard2016/coursera-ds/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Classification Model Development Summary for SpaceX Falcon 9First Stage Landing Prediction



Machine Learning Modeling pipeline for SVC classifier

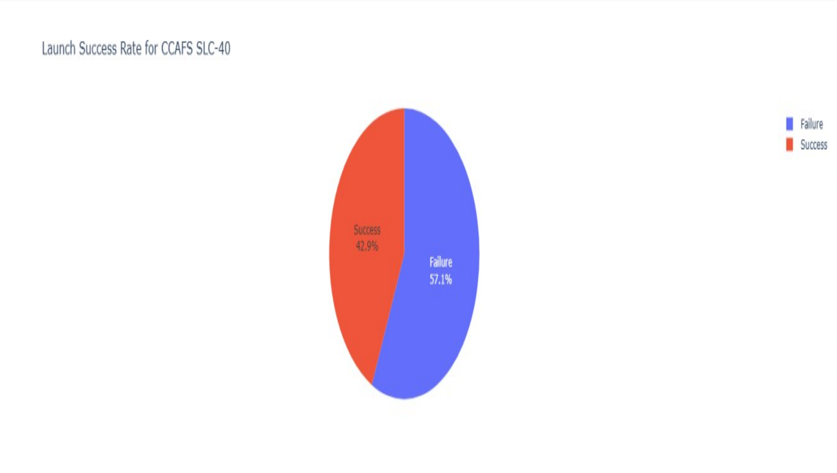
Results – key findings from queries

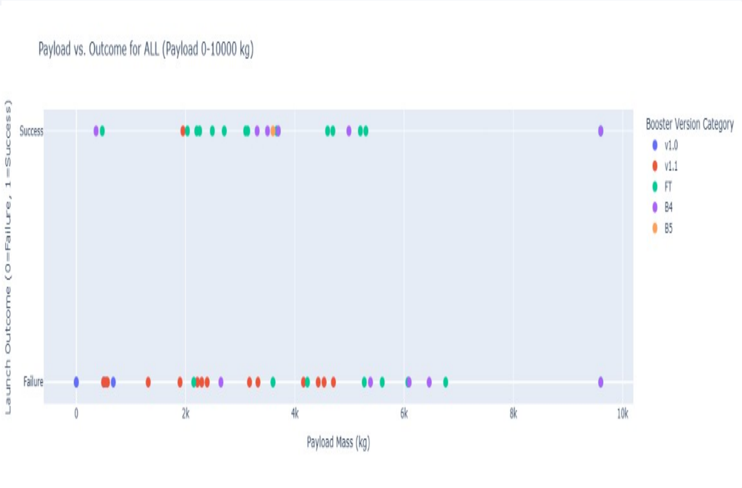
Key Findings	Comments
First successful ground pad landing: 2015-12-22.	Marks a milestone in reusable rocket technology, key to SpaceX's cost advantage.
Booster versions with maximum payload mass (15,600 kg): F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1051.3, F9 B5 B1056.4, F9 B5 B1048.5, F9 B5 B1051.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3, F9 B5 B1051.6, F9 B5 B1060.3, F9 B5 B1049.7.	Showcases Falcon 9 Block 5's capability for heavy payloads, key for competitive contracts.
Total payload mass for NASA (CRS) missions: 45,596 kg.	Demonstrates Falcon 9's reliability for high-profile NASA contracts, boosting credibility.

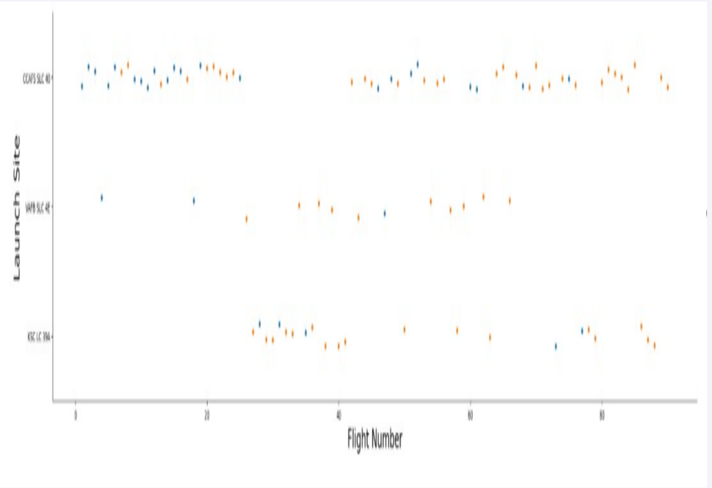
Results – key findings from queries

Key Findings	Comments
First successful ground pad landing: 2015-12-22.	Marks a milestone in reusable rocket technology, key to SpaceX's cost advantage.
Booster versions with maximum payload mass (15,600 kg): 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.	Showcases Falcon 9 Block 5's capability for heavy payloads, key for competitive contracts.
Total payload mass for NASA (CRS) missions: 45,596 kg.	Demonstrates Falcon 9's reliability for high-profile NASA contracts, boosting credibility.

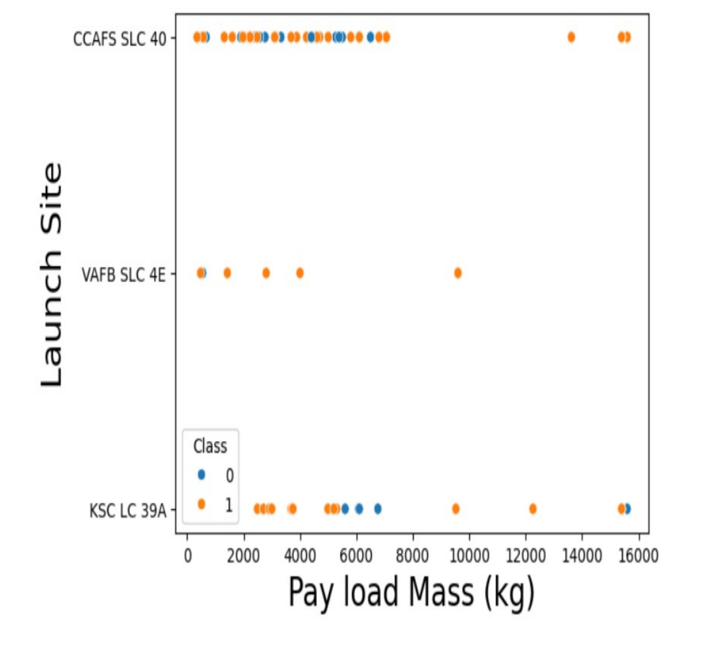
Results – key findings from plots

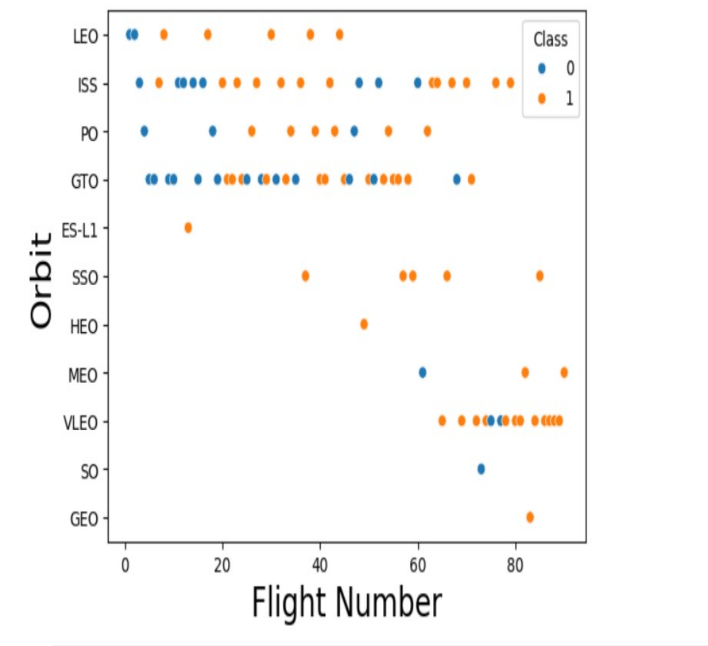






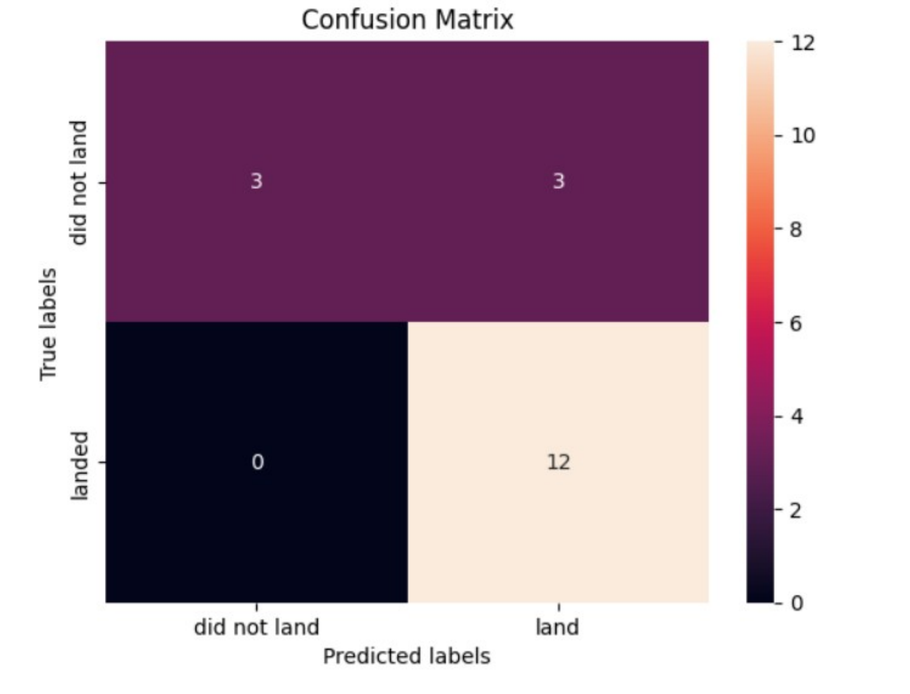
Analyze the plotted bar chart to identify which orbits have the highest success rates.





Results – key findings from model development

Confusion Matrix



Classification Report

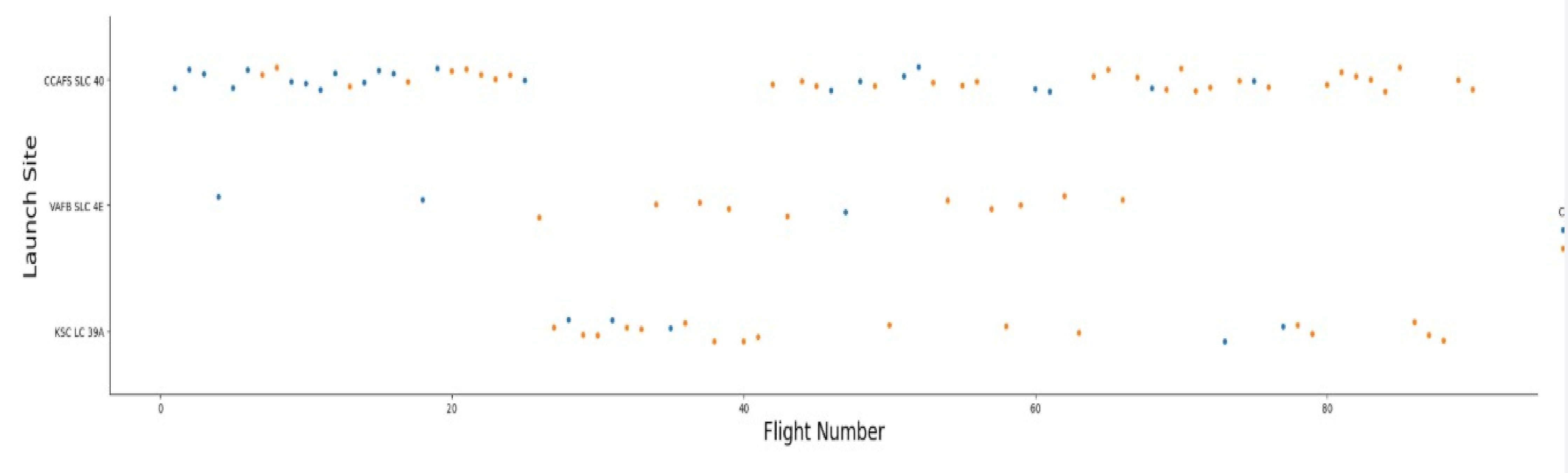
Class	Precision	Recall	F1-Score	Support
Failure	1.0	0.5	0.67	6
Success	0.8	1.0	0.89	12
Macro Avg	0.9	0.75	0.78	18
Weighted Avg	0.87	0.83	0.81	18

Confusion matrix and the Classification Report for the best Classifier - SVC

Section 2

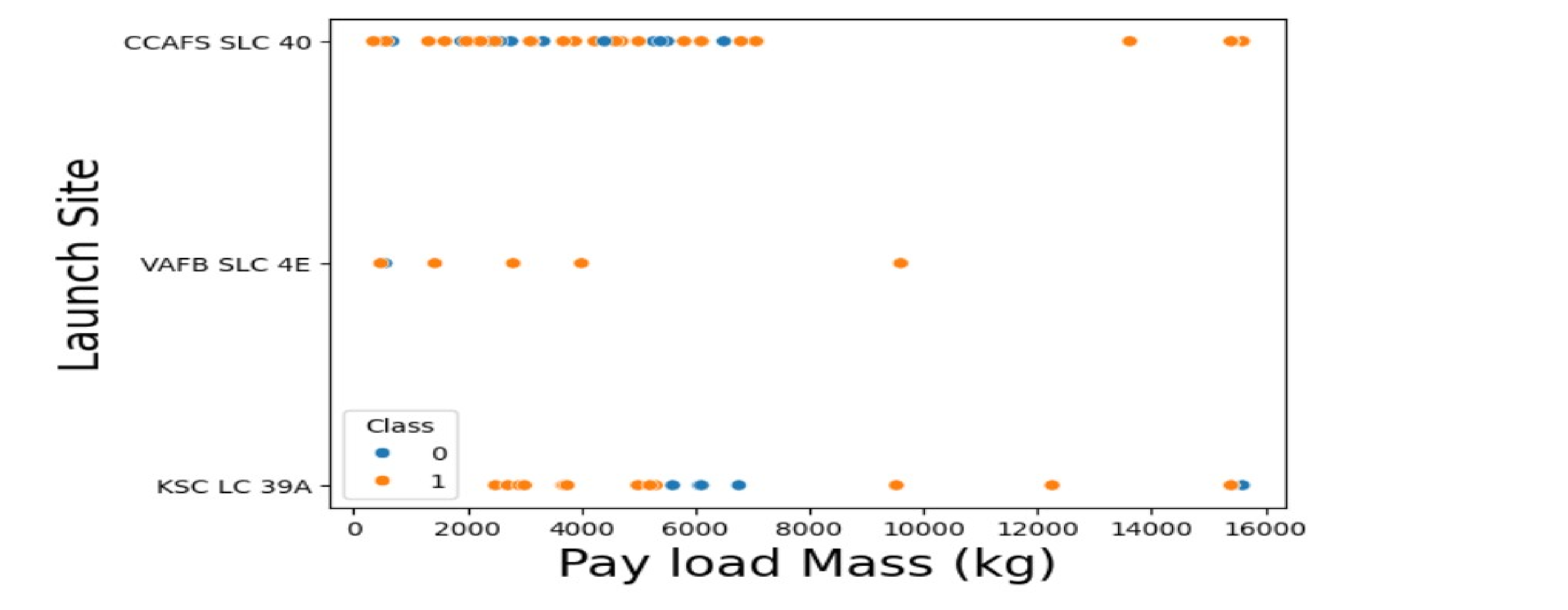
Insights drawn from EDA

Flight Number vs. Launch Site



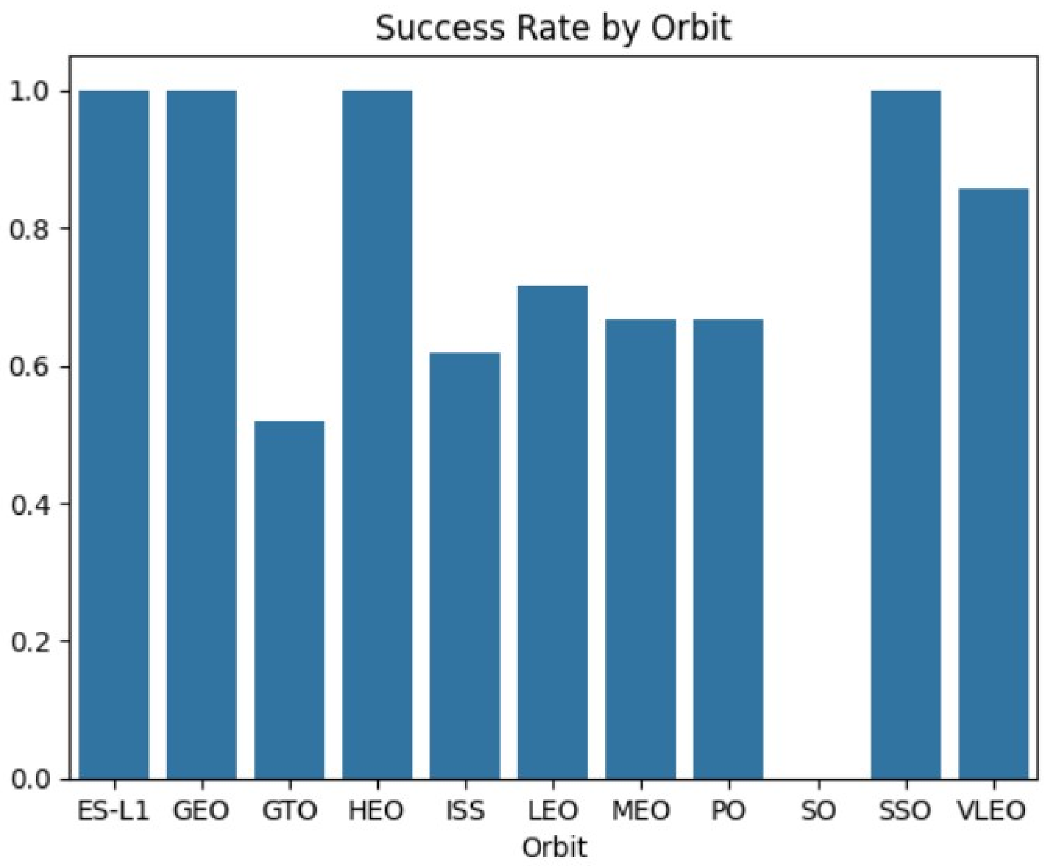
Cumulative Success Rate Over Time: Line plot showing the running success rate of landings as a function of `FlightNumber` , highlighting improving reliability post-early failures.

Payload vs. Launch Site



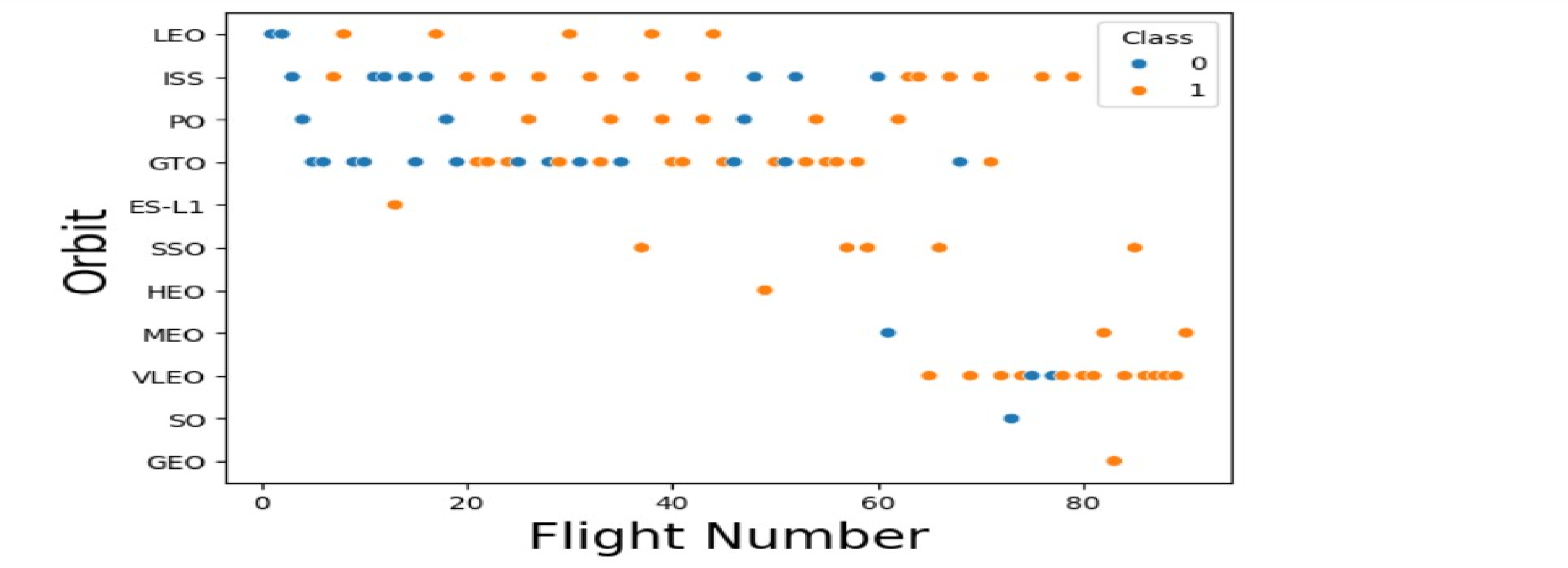
- Launch Site CCAFS SLC-40 seemed to have settled success with higher payloads

Success Rate vs. Orbit Type



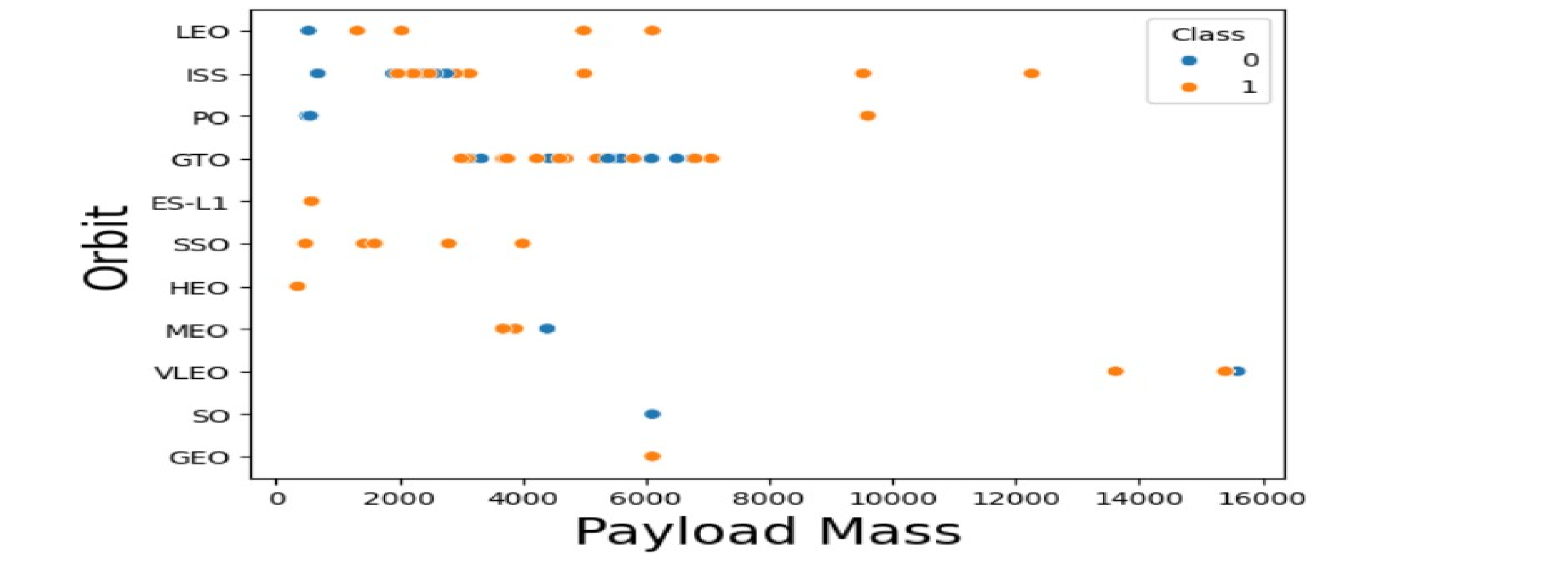
- Orbit types ES-L1, GEO, HEO have had absolute success rates whereas GTO seemed to have the lowest success

Flight Number vs. Orbit Type



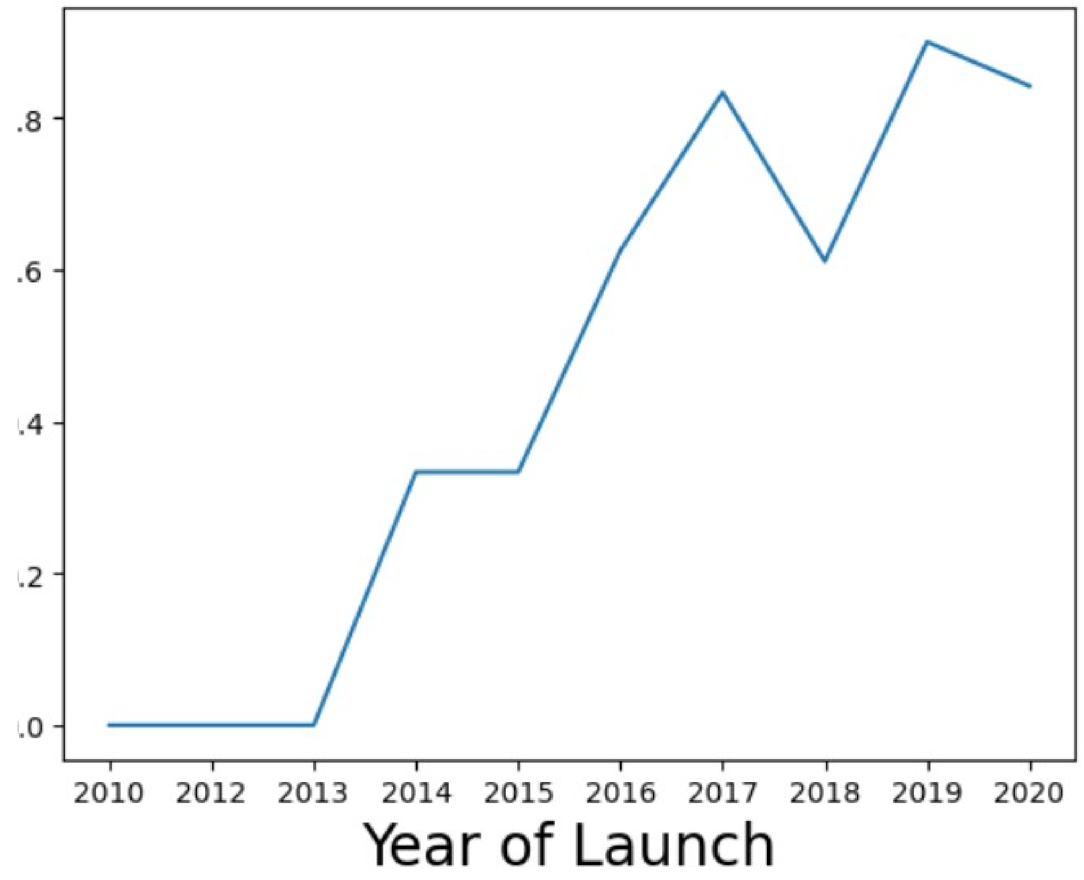
- Orbit types ISS and VLEO seemed to have highest flight number on the ground of settling success

Payload vs. Orbit Type



- Orbit type ISS seemed to have successes with moderate payload as opposed to VELO that experienced higher payloads with some failure though

Launch Success Yearly Trend



- Success seems to have been stabilized over the years and the highest success was experienced between 2019 and 2020

All Launch Site Names

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

From the query on Launch Site using DISTINCT clause I received 4 unqiue site names as shown in the table

%sql select DISTINCT Launch_Site from SPACEXTABLE;

Launch Site Names 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

Top 5 records with Launch Site names that start with 'CCA' was shown in the above table using the following query: *%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;*

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

Total Paylaod Mass was 45596 Kg using the following query
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)';

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1';

First Successful Ground Landing Date

first_successful_ground_landing
2015-12-22

First Successful Ground Landing was on 22 December of 2015 and was found using the following query:

%sql *SELECT MIN(Date) as first_successful_ground_landing FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%Ground Pad%' AND Landing_Outcome LIKE '%Success%';*

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

The list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 has been shown in the above table. The result was obtained using the following query:

```
%sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome='Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Total Number of Successful and Failure Mission Outcomes

count(*)
71

The total number of successful and failure mission outcomes was 71 and was found using the following query:

%sql select count(*) from SPACEXTABLE where Landing_Outcome LIKE '%Success%' OR Landing_Outcome LIKE '%Failure%';

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass are shown in the table and this was found using the following query:

%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_=(Select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE);

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

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

The two failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 has been shown in the table above and the result was the query below:

%sql select Date,substr(Date,6,2) as "Month", Landing_Outcome, Booster_Version,Launch_Site from SPACEXTABLE where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015';

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

Landing_Outcome	count	rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

The rank of 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 has been shown in the table and the result was found using the following query:

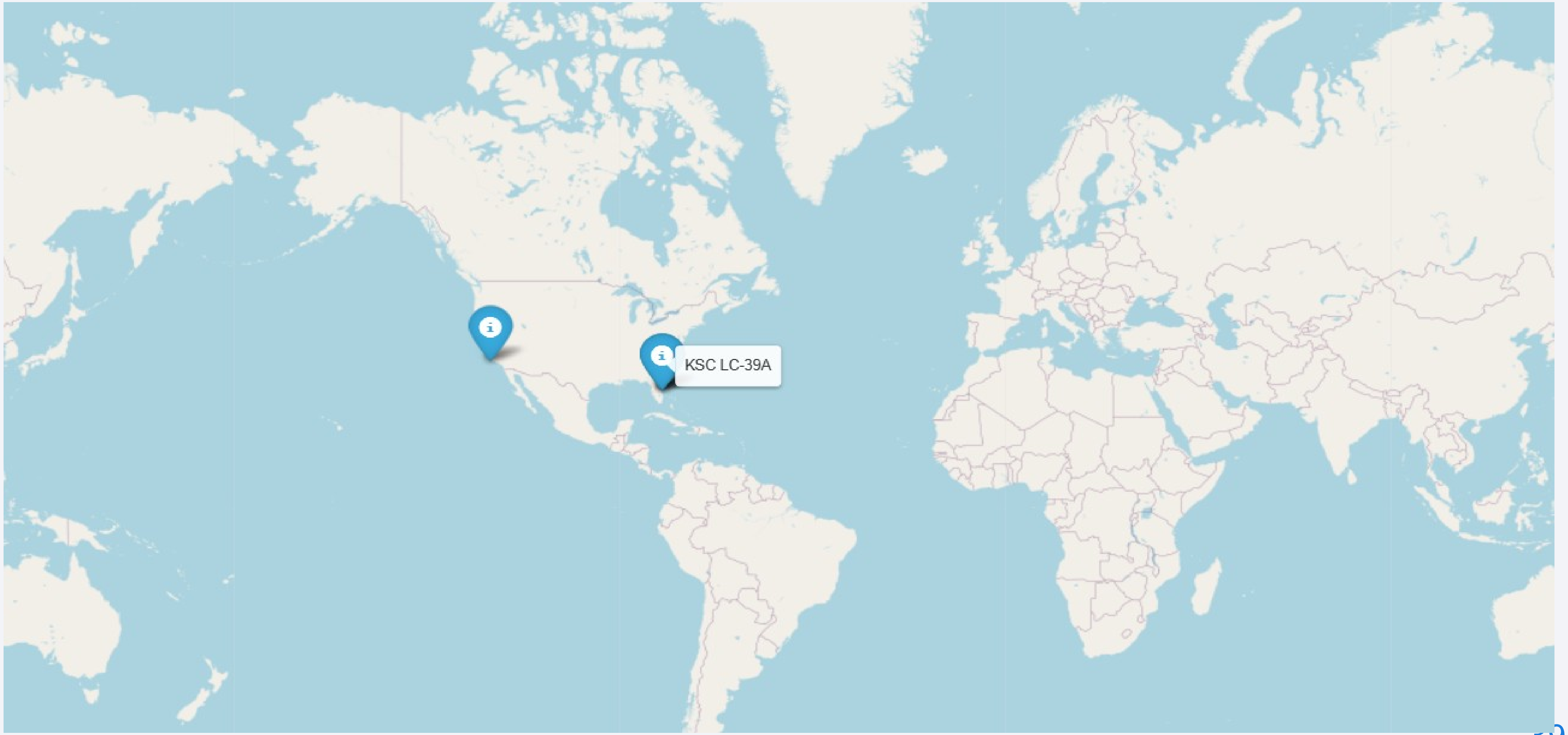
```
%%sql WITH outcome_counts AS (
  SELECT
    Landing_Outcome,
    COUNT(*) as count
  FROM SPACEXTABLE
  WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    AND Landing_Outcome IS NOT NULL
  GROUP BY Landing_Outcome
)
SELECT
  Landing_Outcome,
  count,
  RANK() OVER (ORDER BY count DESC) as rank
FROM outcome_counts

ORDER BY count DESC;
```

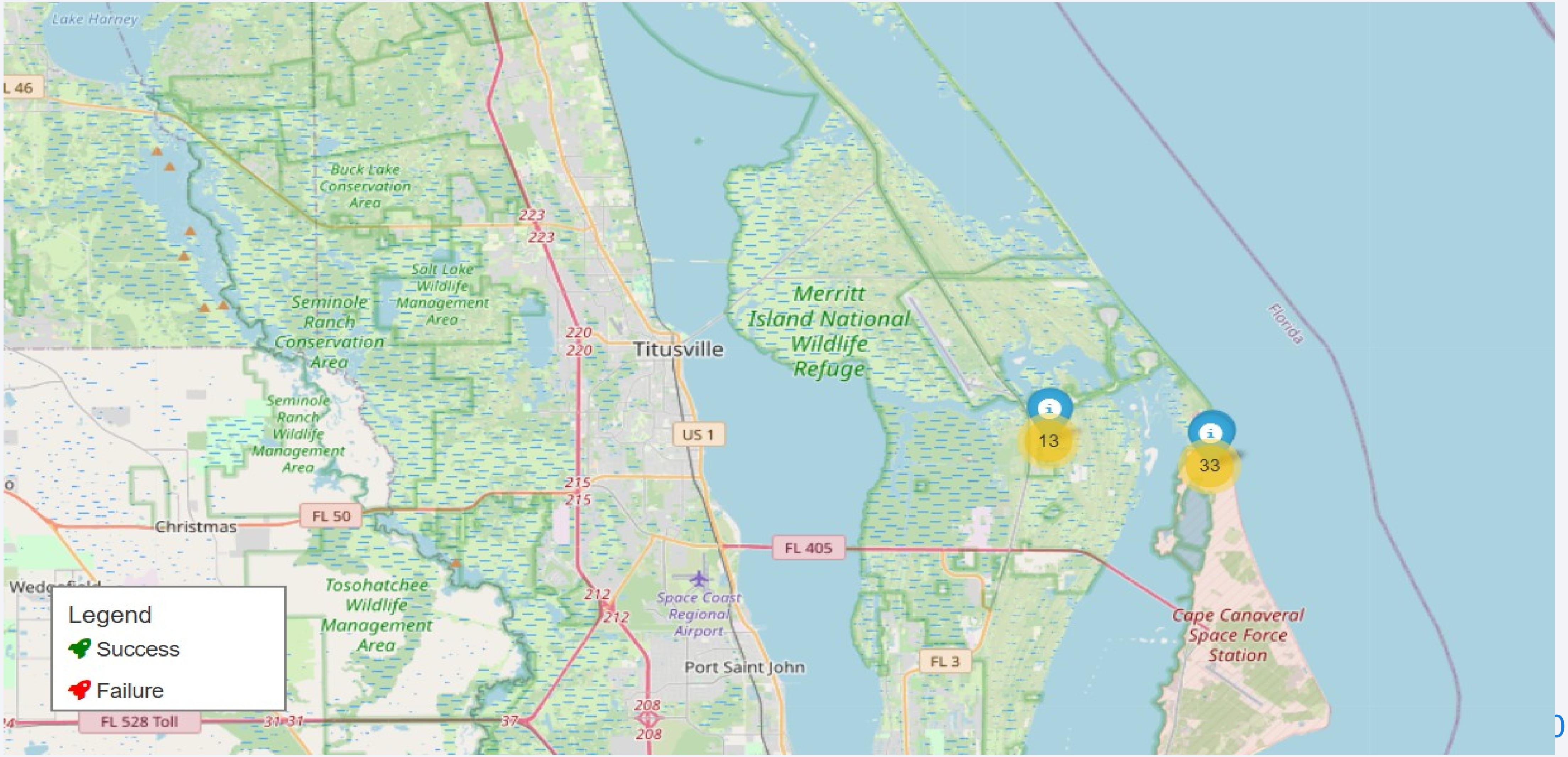
Section 3

Launch Sites Proximities Analysis

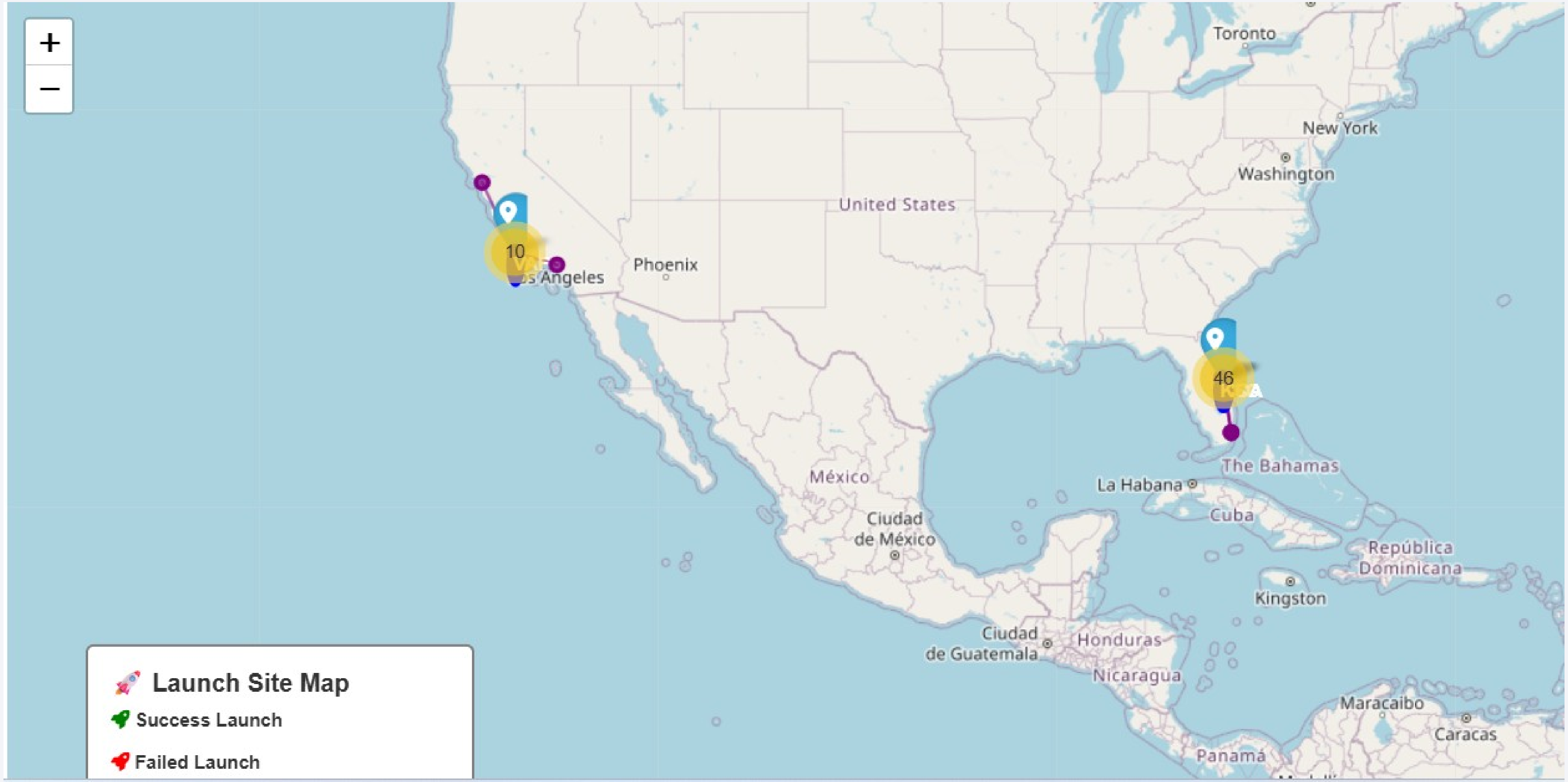
Folium Map with Launch Site Name on hover



Folium Map marked the Launch Sites as Success and Failure



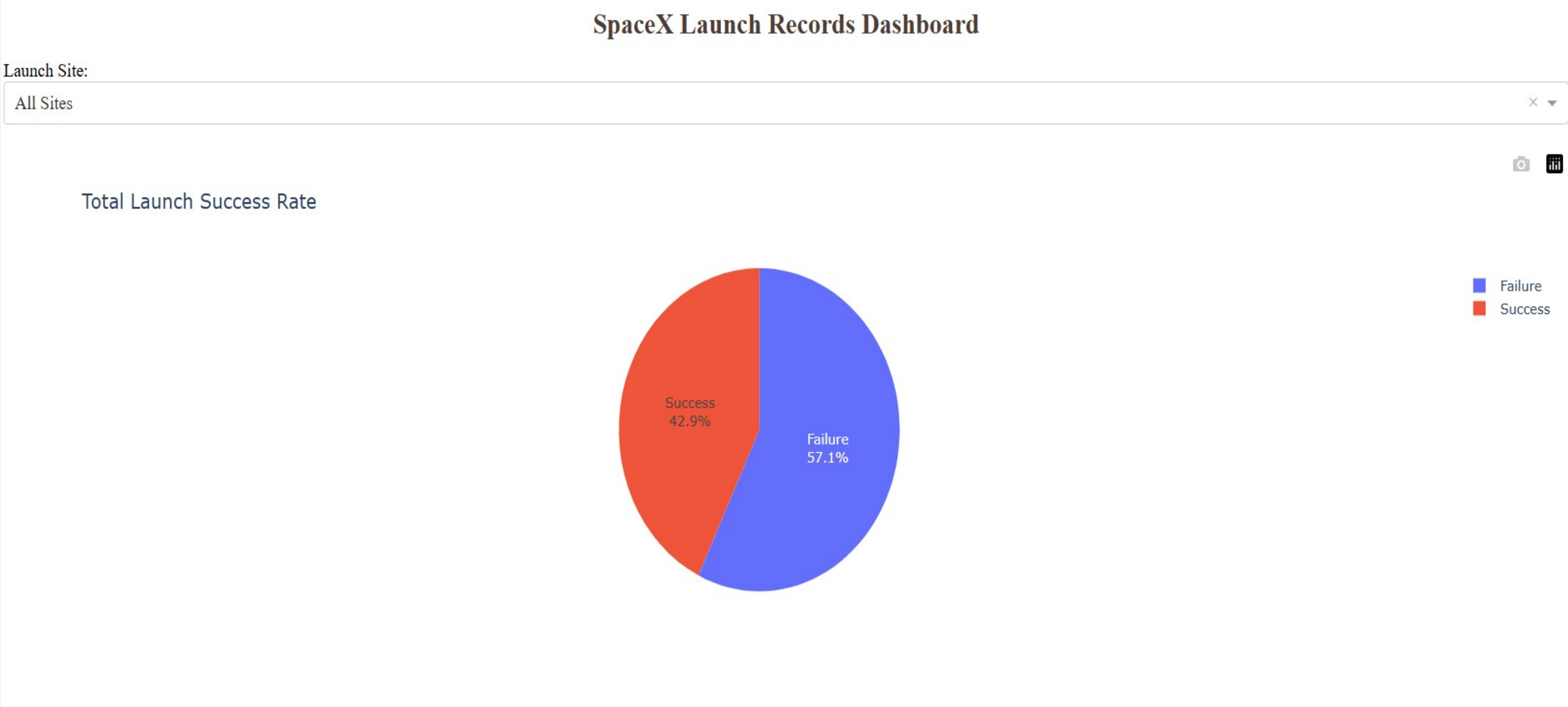
Folium Map with Launch Site Proximities



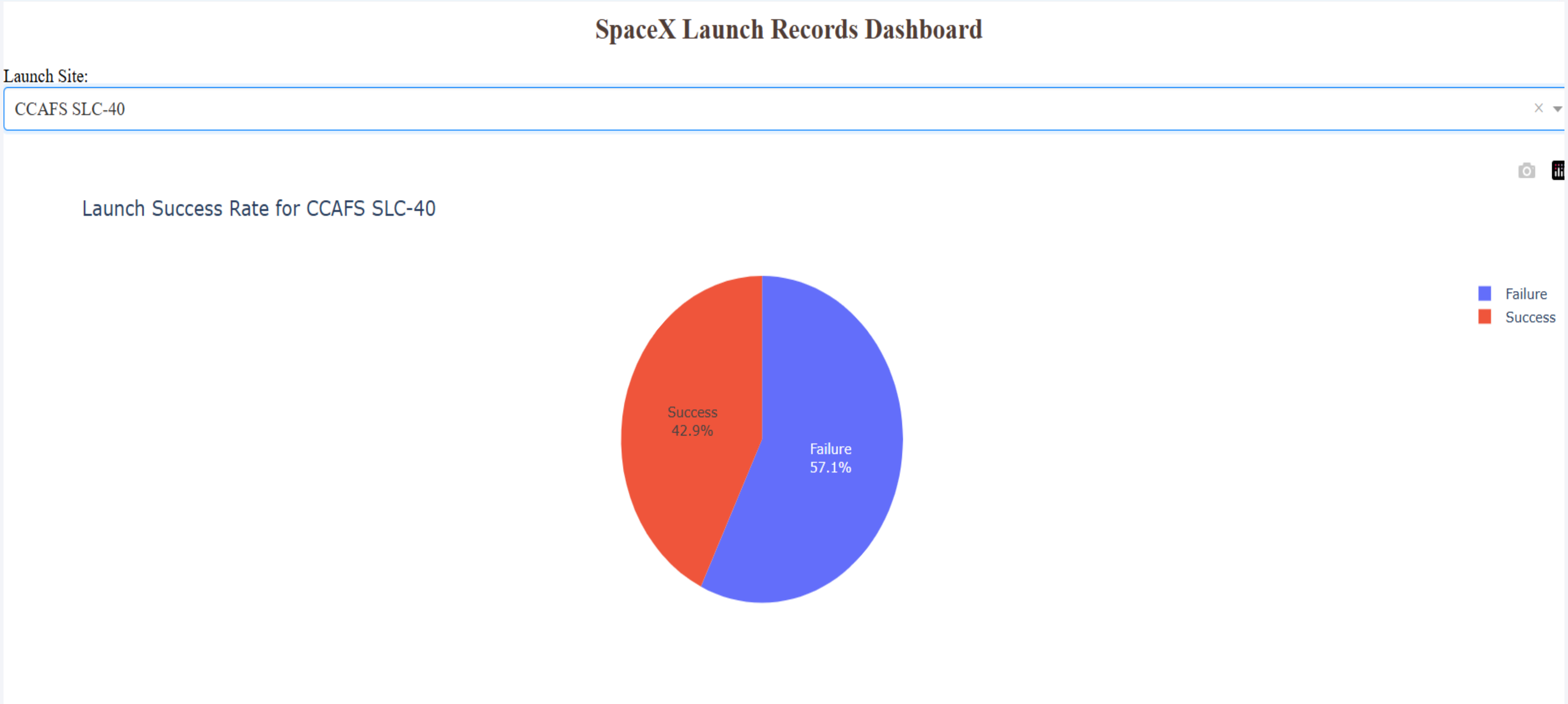
Section 4

Build a Dashboard with Plotly Dash

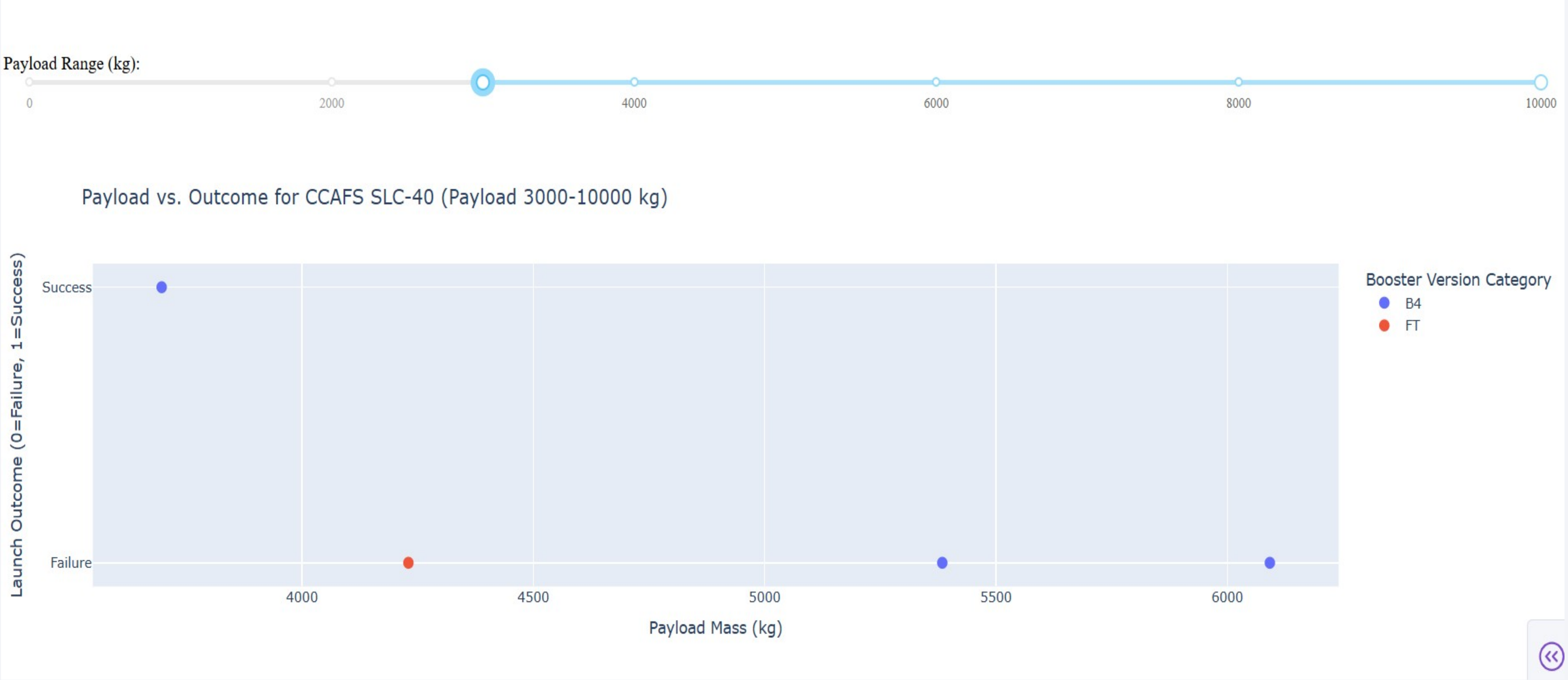
Dashboard showing overall Success and Failure rates



Dashboard showing Highest Success Rate for Site CCAFS SLC-40



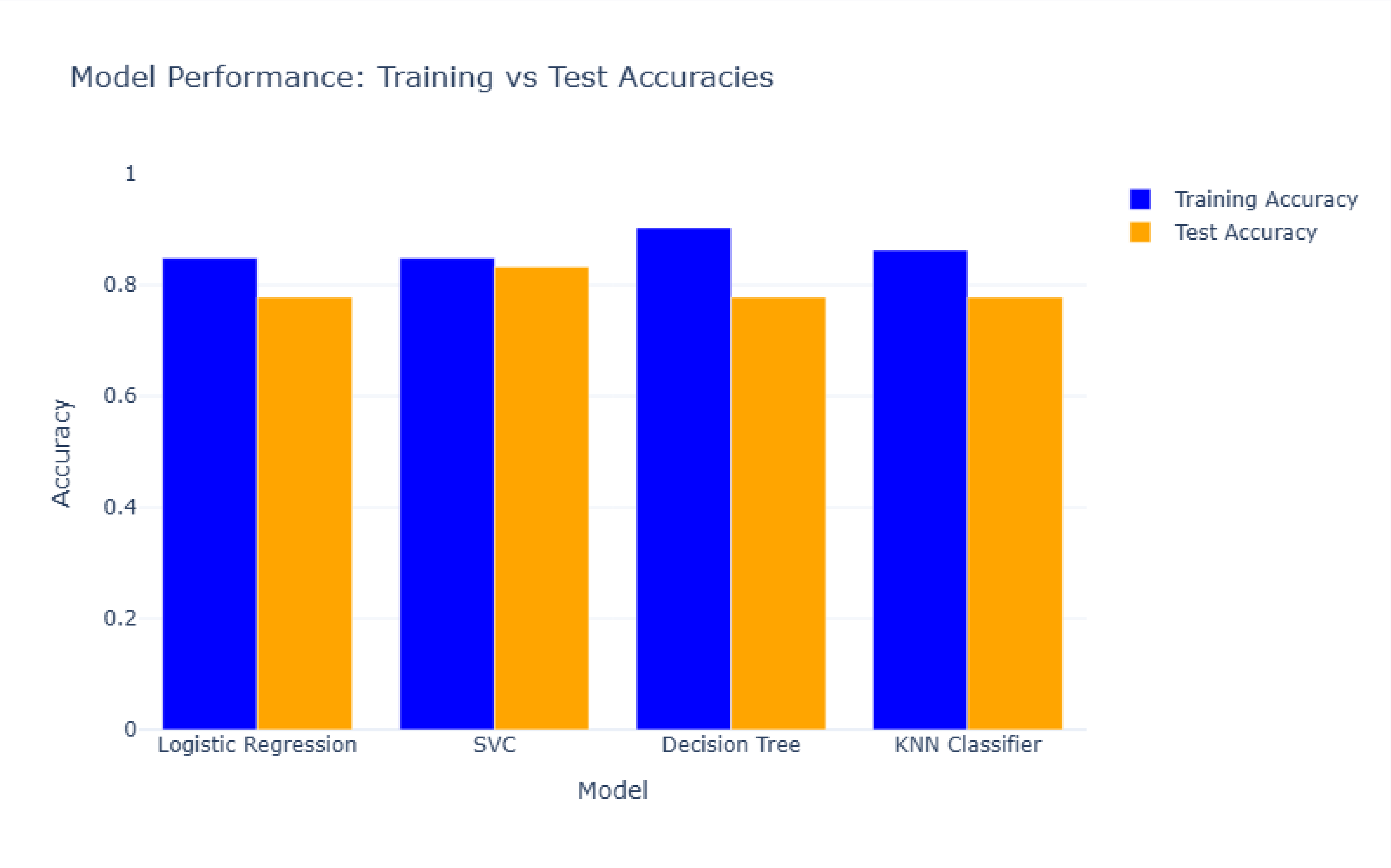
Dashboard showing outcomes for varying Payloads



Section 5

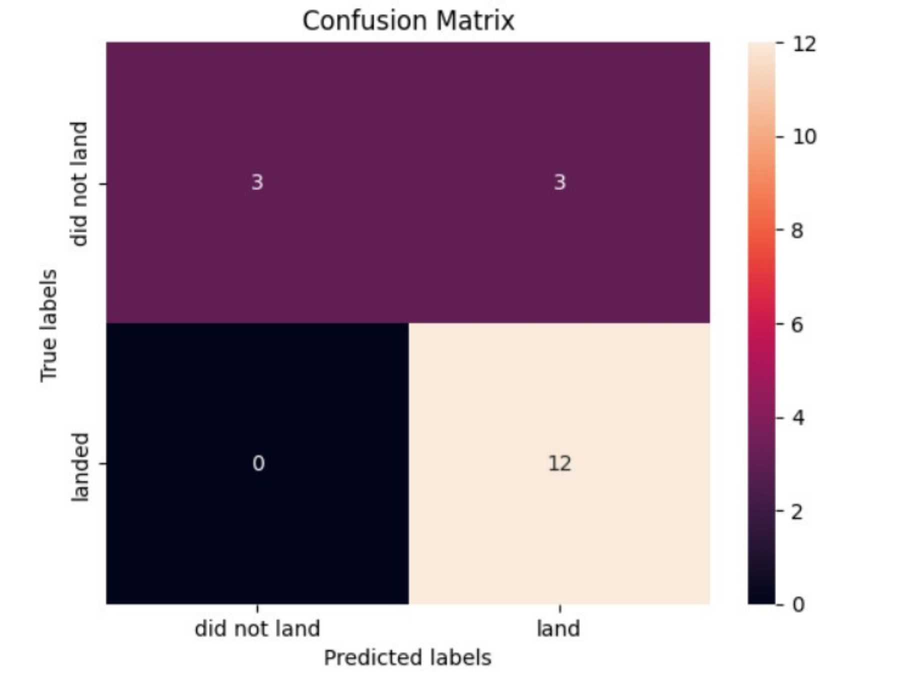
Predictive Analysis (Classification)

Classification Accuracy



Test Accuracy is the highest for Support Vector Machine Classifier (SVC) with 83.3%

Confusion Matrix for the best model - SVC



SVC ended up with the lowest misclassification of 3 – didn't classify a single launch that landed in reality as did not land

SVC should be the best model as

Reason	**Description**
Highest Test Accuracy	SVC achieves the best test accuracy (0.833) compared to others (0.778).
Fewest Misclassifications	Only 3 incorrect predictions on test data, outperforming others (4 errors).
Balanced Performance	Small gap between training (0.8482) and test (0.833) accuracy, reducing overfitting risk.
Robust to High-Dimensional Data	Effectively handles 83-column dataset with optimal hyperplane separation.
Captures Non-Linear Patterns	Kernel functions (e.g., RBF) model complex relationships in features like payload and orbit.

Appendix

https://github.com/samizard2016/coursera-ds/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/edadataviz_sp.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/spacex_dashboard1.py

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-webscraping.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/jupyter-labs-webscraping.py

https://github.com/samizard2016/coursera-ds/blob/main/lab_jupyter_launch_site_location.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

https://github.com/samizard2016/coursera-ds/blob/main/model%20accuracies.png

Thank you!