IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mehdi Hosseinpour
04/19/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

This project focuses on the landing outcome of the Falcon 9 first stage. Data was collected from the SpaceX API, capturing key information such as rocket type, payload mass (kg), launch and landing sites, and landing outcome (successful vs. failed). The data was then cleaned, structured, and visualized to explore the relationship between different factors and the target outcome—landing success. Next, four machine learning algorithms, including logistic regression, Support Vector Machine (SVM), decision tree classifier, and K-Nearest Neighbors (KNN), were developed to predict the likelihood of successful first-stage landings. Each model was tuned using GridSearchCV with 10-fold cross-validation. The comparison results showed that the decision tree classifier slightly outperformed the others in terms of cross-validation accuracy, though all models achieved similar results on the test set.

Introduction

Several aerospace companies—such as Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX—are working to make space travel more affordable and accessible. Among them, SpaceX stands out as a pioneer, primarily due to its innovative use of first-stage reusability, which significantly reduces launch costs to around $62 million per launch. However, not all first-stage landings have been successful. To address this challenge, it is crucial to develop a machine learning model capable of accurately predicting the likelihood of a successful first-stage landing. The prediction outcomes from such a model may provide valuable insights to competing aerospace companies or contractors aiming to bid competitively against SpaceX for launch services. Therefore, the main objective of this project is to determine the probability that the first stage will successfully land.

Section 1
Methodology

Methodology

- Data collection from relevant sources (e.g., SpaceX API)
- Perform data wrangling (e.g., one-hot encoding, handling missing values)
- Conduct exploratory data analysis (EDA) using visualization and SQL tools
- Create interactive visual analytics using Folium and Plotly Dash
- Build predictive analysis using classification models

Data Collection

Data was collected from two primary sources:

(1) SpaceX API:

Retrieved structured data including rocket type, payload mass (kg), launch and landing sites, and landing outcome (successful vs. failed).

(2) Web Scraping from Wikipedia:

Collected historical launch data for Falcon 9 rockets from a Wikipedia page listing Falcon 9 missions.

Data Collection – SpaceX API

- Send the GET request to SpaceX API and decode the response content as a Json
- Turn Json file into a Pandas data frame
- Filter the data to only include Falcon 9 launches

GitHub URL

Data Collection - Web Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract the HTML table containing Falcon 9 launch records
- Parse the table and convert it into a Pandas data frame

GitHub URL

Data Wrangling

- Handle missing values

- Calculate the number and occurrence of each orbit
- Compute the number of launches on each orbit

- Calculate the number and occurrence of mission outcome of the orbits

- Create a new binary variable, named Class, representing the outcome of each launch (1: successful landing; 0: otherwise)

GitHub URL

EDA with Data Visualization & Feature Engineering

Seaborn *catplot* to Visualize:

- Launch Site vs. Flight Number, grouped by Landing Outcome (Class 0 = Failure, Class 1 = Success)
- Flight Number vs. Orbit type, grouped by Landing Outcome

Seaborn *scatterplot* to Visualize:

- Payload Mass vs. Launch Site, grouped by Landing Outcome
- Payload Mass vs. Orbit type, grouped by Landing Outcome

Seaborn *barplot* to Visualize:

- The relationship between success rate of each orbit type

Seaborn *lineplot* to Visualize:

- The launch success yearly trend

Pandas *get_dummies* Function:

- Create dummy variables for categorical columns

[GitHub URL]

EDA with SQL

SQL queries were executed to:

- 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 ground pad was achieved
- List the total number of successful and failure mission outcomes
- List all the booster versions that have carried the maximum payload mass
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

[GitHub URL]

Build an Interactive Map with Folium

Map Object	Why It Was Used
Marker	To highlight key launch sites, landing sites, or observation points on the map
Circle	To display launch site vicinities
CircleMarker	For plotting dense points like landing sites grouped by landing outcome
Polyline	To represent flight paths or the trajectory of a rocket launch
Popup	To provide more detailed information such as site name, or landing outcome (0, 1)
MarkerCluster	To visualizing multiple markers on the map
MousePosition	Marks down a point on the closest coastline to a specific launch site.

GitHub URL

Build a Dashboard with Plotly Dash

Map Object	Why It Was Used
Plotly Express (px)	Create interactive plots (e.g., pie charts, scatter plots)
px.pie()	Create interactive pie chart displaying the percentage of success launches by site
px.scatter()	Create interactive scatter plots for payload mass vs. landing outcome
Dash HTML Component	Used to structure and format the style of web applications
Dash Core Component	Allow to create interactive elements like dropdowns, sliders, buttons, and graphs
dcc.RangeSlider()	Add a Range Slider to select different ranges of payloads
Dash @app.callback()	Render interactive pie-chart and scatter plots by updating the content

GitHub URL

Predictive Analysis (Classification)

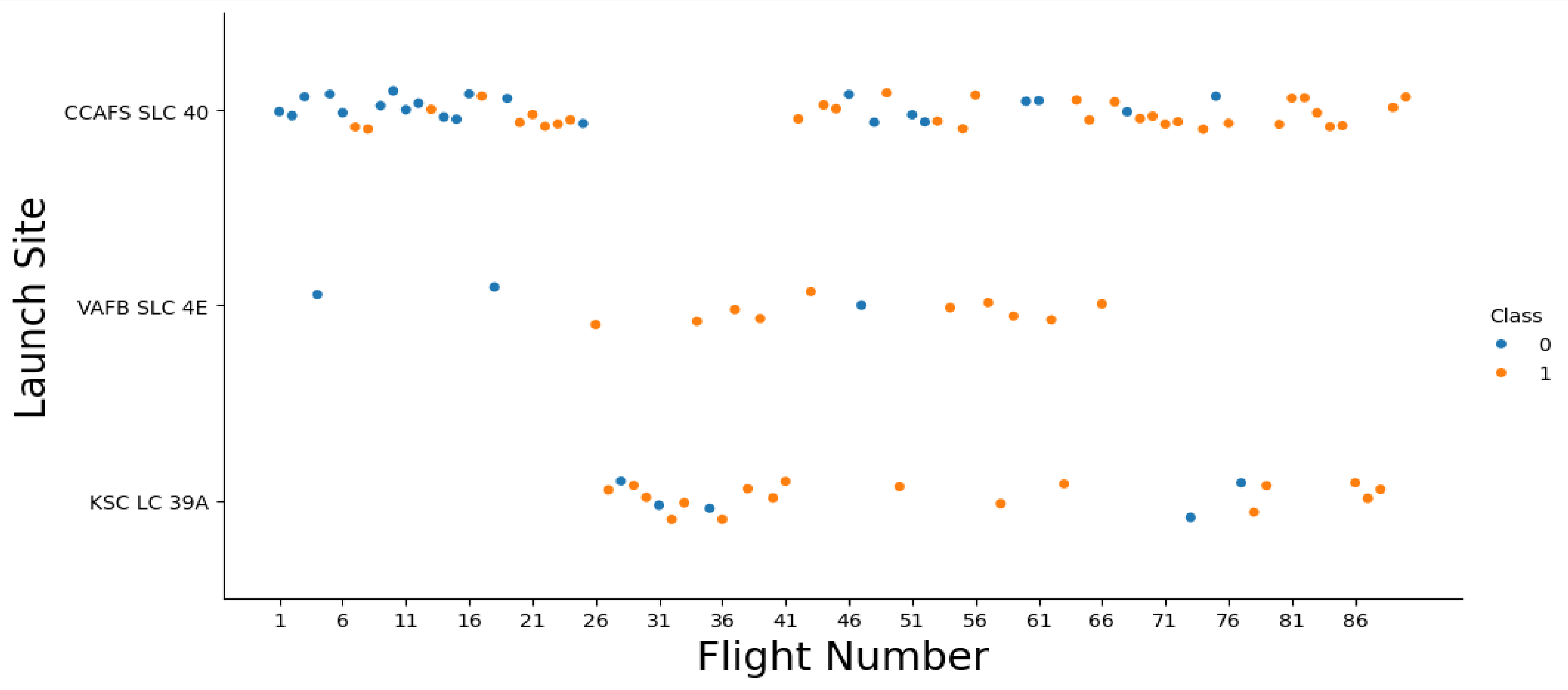
1. **Load and prepare the data** (e.g., perform feature scaling)
2. **Standardize the features and assign the them to the variable scaled_X**
3. **Split the data** (80% for training, 20% for testing)
4. **Train classification models** (Binary Logistic Regression, SVC, Decision Tree, and KNN)
5. **Tune and fit models** (Use Scikit-learn GridSearchCV with 10-fold cross-validation)
6. **Evaluate model performance** (Based on cross-validation score, confusion matrix, and test accuracy)

[GitHub URL]

Section 2

Insights drawn from EDA

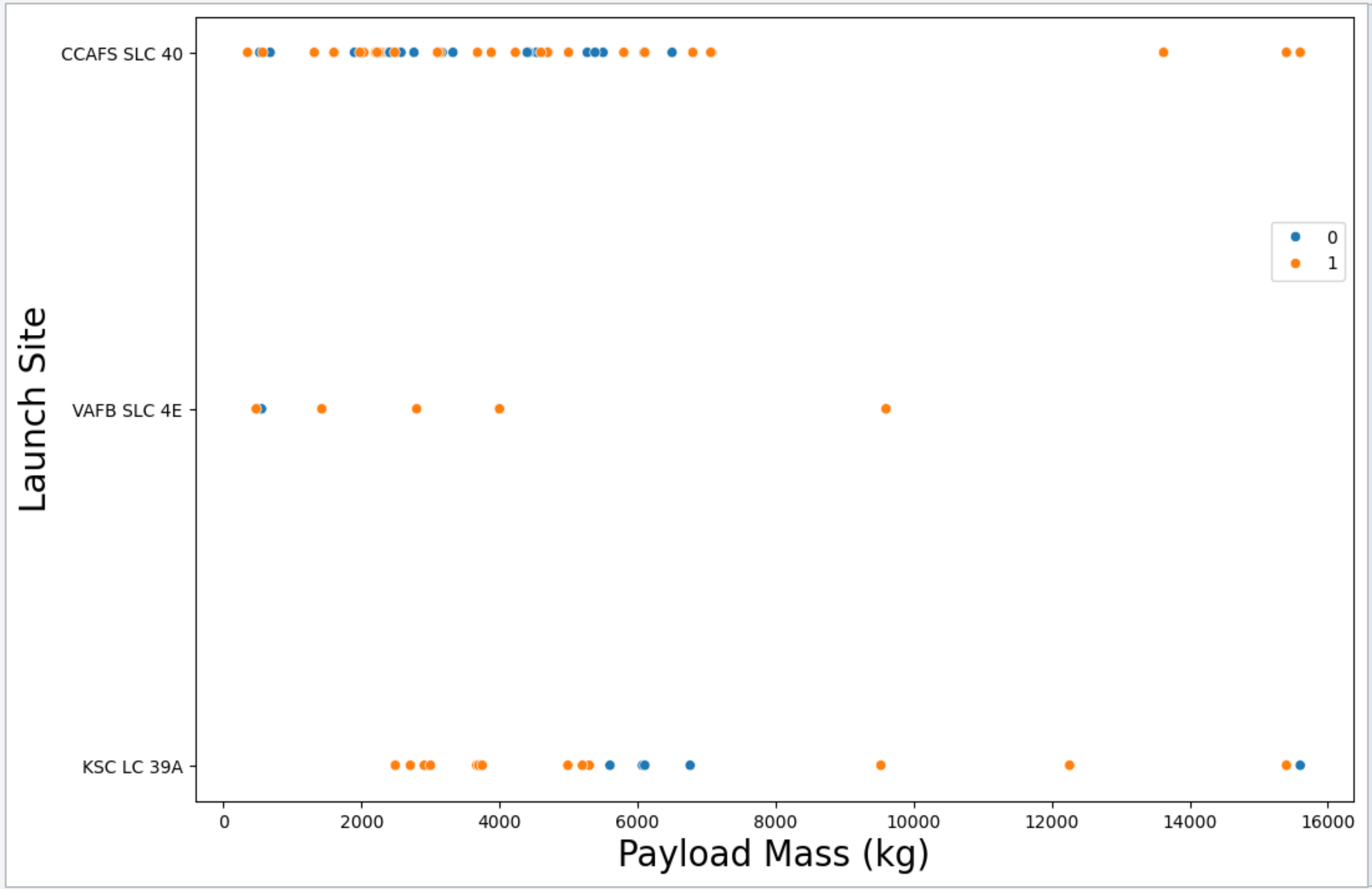
Flight Number vs. Launch Site



[GitHub URL]

- The likelihood of successful launches tends to increase as the flight number increases.
- More launches were conducted and successfully landed at the CCAFS SLC 40 launch site.

Payload vs. Launch Site

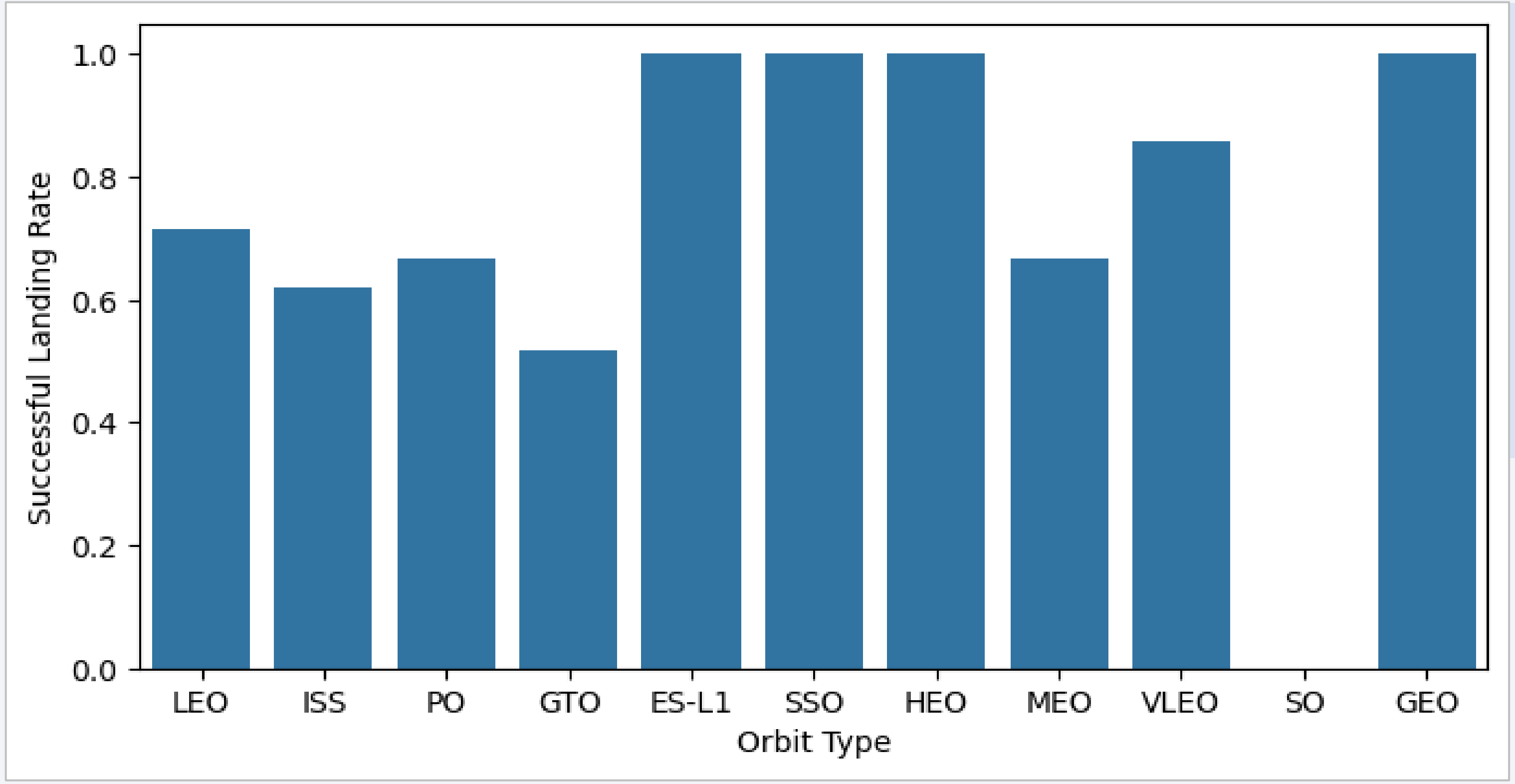


- For the CCAFS SLC 40 and VAFB SLC 4E launch sites, the likelihood of successful launches tend to increase with heavier payload mass.

- For the KSC LC 39A launch site, launches with payload mass around 6000 kg had unsuccessful landing, while other payload ranges tend to have successful landing.

[GitHub URL]

Success Rate vs. Orbit Type



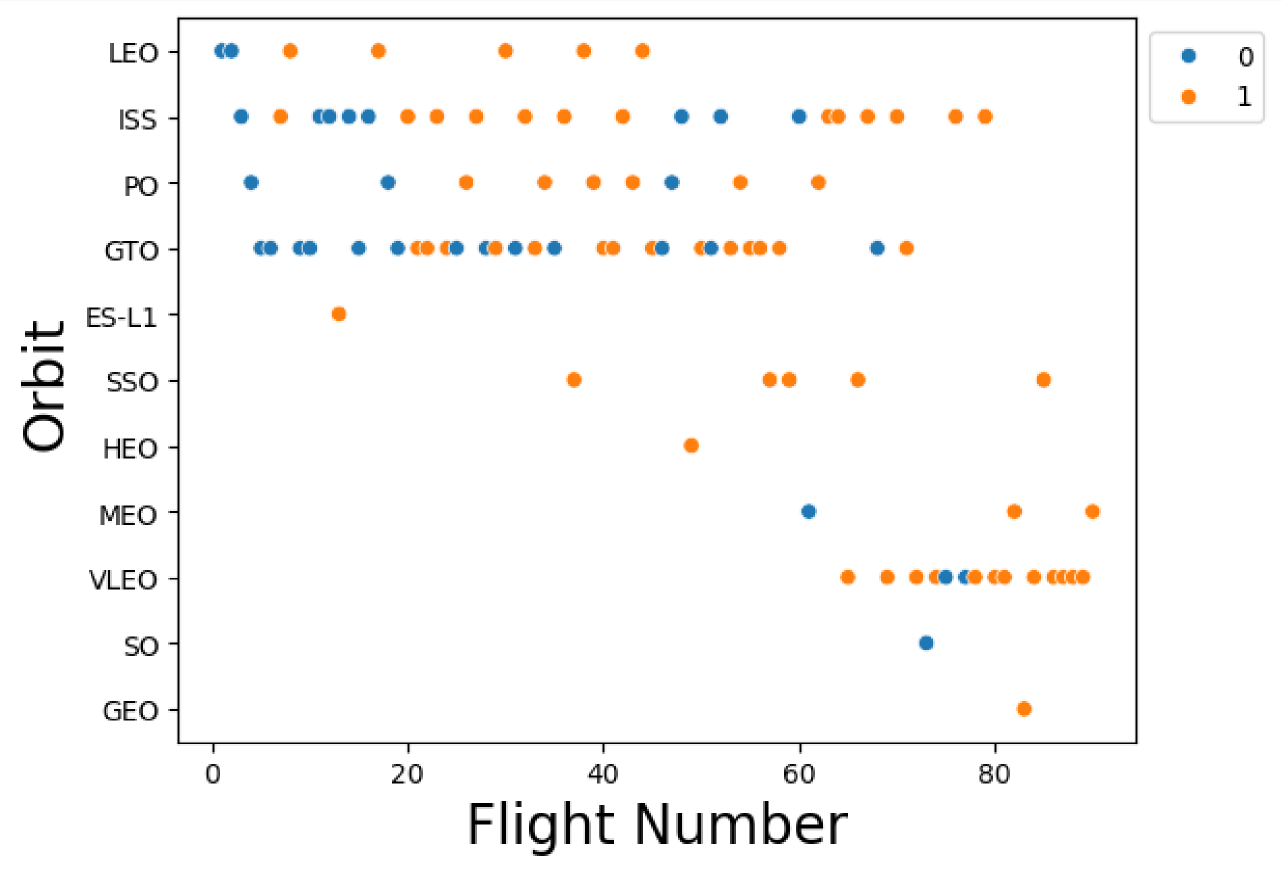
- The VLEO, LEO, PO, and ISS orbits had moderate successful landing rates, ranging from 63% to 87%.

GitHub URL

- The ES-L1, SSO, HEO, and GEO orbits achieved a 100% successful landing rate, while the SO orbit had the lowest success rate.
- Although not shown in the barplot above, the ES-L1, HEO, GEO, and SO orbits each had a single launch with the SO orbit resulting in a failed landing.

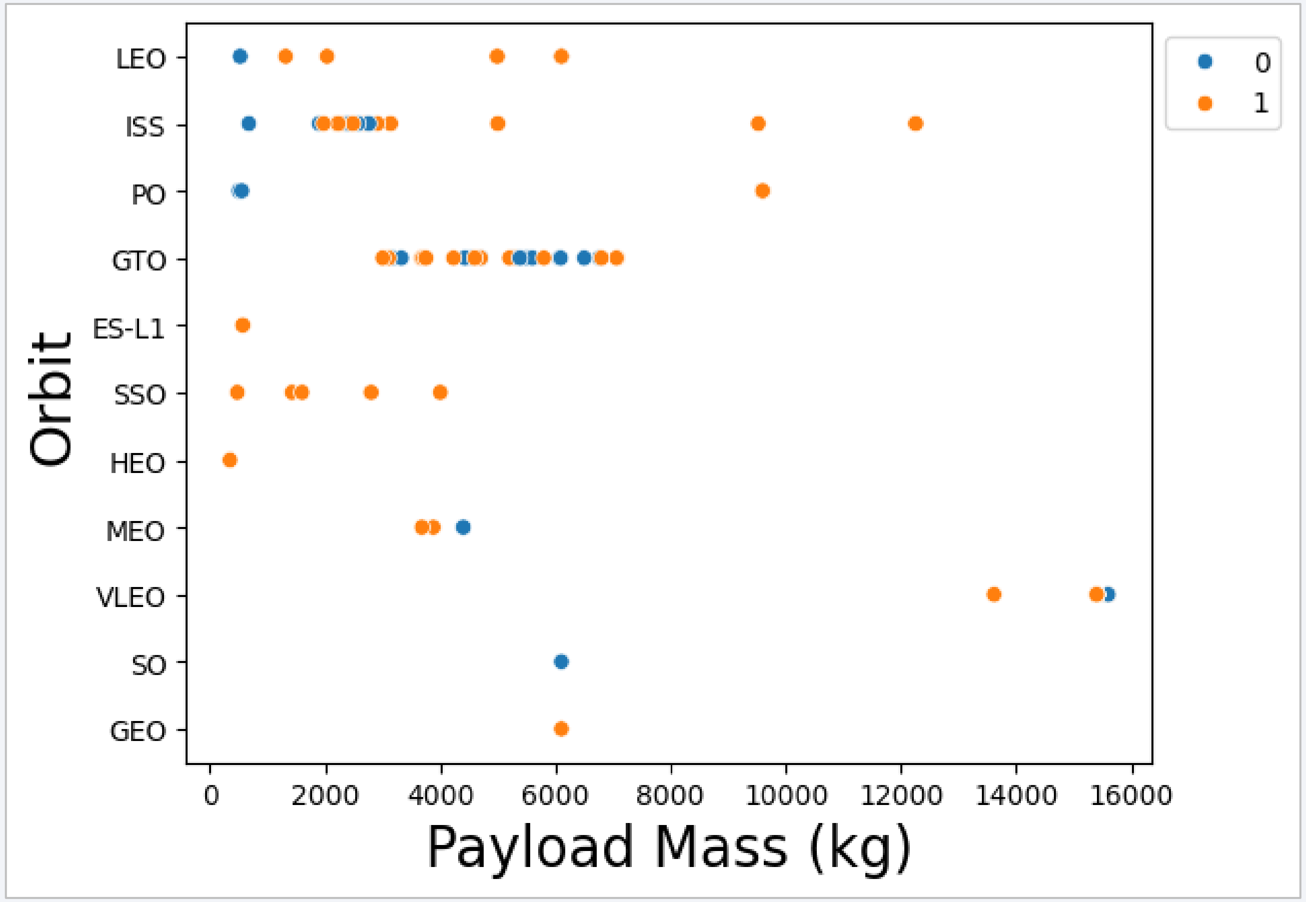
Flight Number vs. Orbit Type

- The GTO, ISS, PO, and LEO orbits had launches spanning a wider range of flight numbers.
- The VLEO orbit type was associated with launches in the higher flight number, ranging from 65 to 89.
- The ES-L1, HEO, GEO, and SO orbits each had a single launch.



Payload vs. Orbit Type

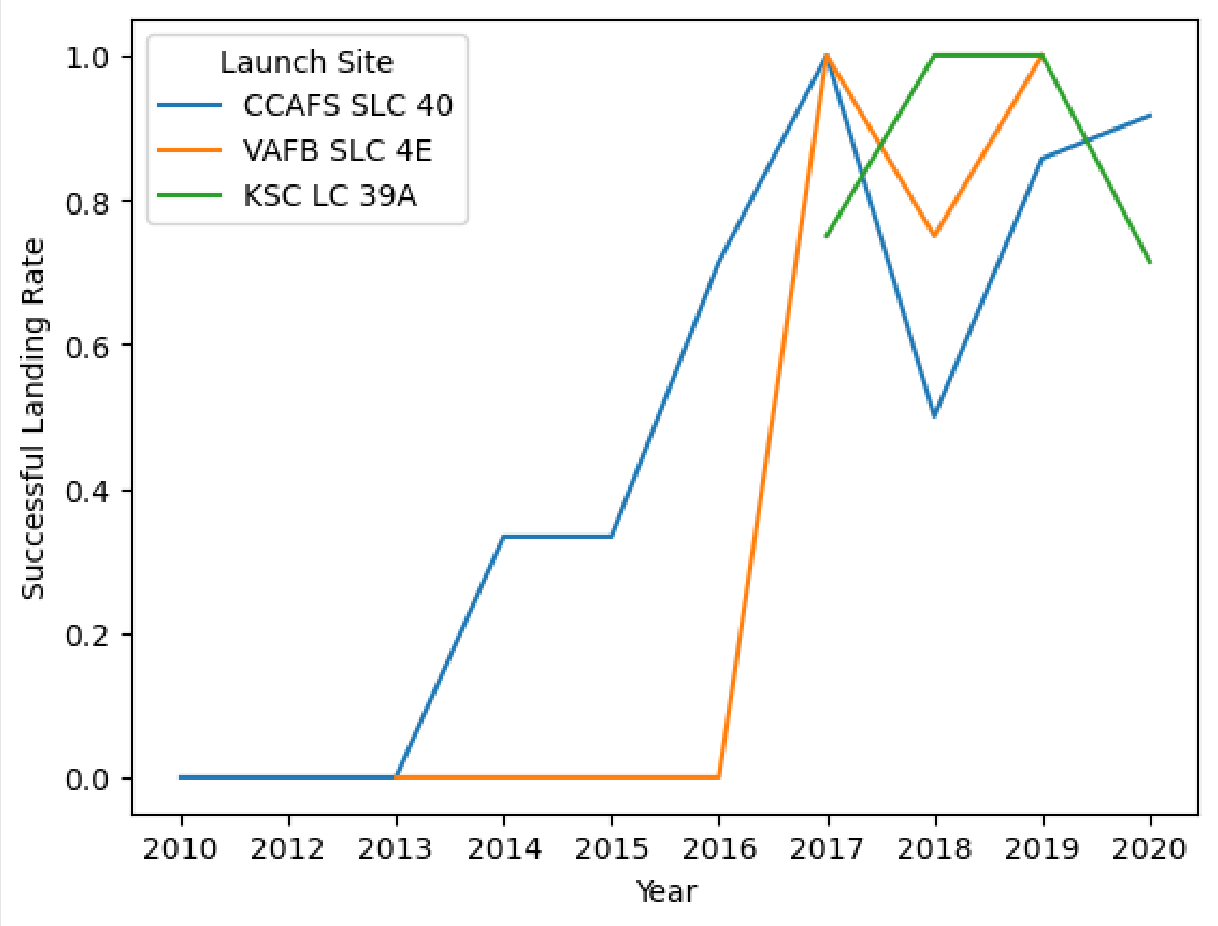
The scatter plot indicates that certain orbits, such as ISS, LEO, and GTO, had a wider range of payload masses. However, due to the limited number of launches associated with other orbit types (e.g., PO, VLEO, SO), there is no clear pattern linking orbit type, payload mass, and landing outcome.



GitHub URL

Launch Success Yearly Trend

- The line chart depicts the average success landing by launch site over years.
- Overall, the rate of successful landings increased over the years, especially after 2016, with the CCAFS SLC 40 and VAFB SLC 4E launch sites having notable improvements in success rates.



GitHub URL

All Launch Site Names

- **SQL Query to find the names of the unique launch sites:**

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

- **Explanation:** The SQL query above returns a list of unique launch sites from the SPACEXTABLE

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

GitHub URL

Launch Site Names Begin with 'CCA'

- **SQL Query to find 5 records where launch sites begin with "CCA“:**

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

Explanation: This SQL query displays the first 5 launches from the SPACEXTABLE 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

GitHub URL

Total Payload Mass

- **SQL Query to calculate the total payload carried by boosters from NASA:**

%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';

- **Explanation:** This SQL query returns **total payload mass (in kg)** launched for missions where NASA (CRS) was the customer.

SUM(PAYLOAD_MASS__KG_)
45596

GitHub URL

Average Payload Mass by F9 v1.1

- **SQL Query to** calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE
Booster_Version = 'F9 v1.1’;
```

- **Explanation:** the query above returns the average payload mass for all flights using the F9 v1.1 booster version.

AVG(PAYLOAD_MASS__KG_)
2928.4

GitHub URL

First Successful Ground Landing Date

- **SQL Query** to find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN(Date) AS First_Successful_GroundPad_Landing FROM
SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)’;
```

- **Explanation:** To find and display the dates of the first successful landing occurred on a ground pad.

First_Successful_GroundPad_Landing
2015-12-22

GitHub URL

Successful Drone Ship Landing with Payload between 4000 and 6000

- **SQL Query** to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Explanation: The query above was executed to list the names of boosters that successfully landed on a drone ship and carried a payload between 4000 kg and 6000 kg.

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

GitHub URL

Total Number of Successful and Failure Mission Outcomes

- **SQL Query** to calculate the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

Explanation:

The query above was run to calculate and list the total number of successful and failure mission outcomes.

Mission_Outcome	Outcome_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

GitHub URL

Boosters Carried Maximum Payload

- **SQL Query** to list the names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

- **Explanation:** The query above was used to find those boosters in the SPACEXTABLE dataset that carried the heaviest payload.

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

GitHub URL

2015 Launch Records

- **SQL Query** to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 1, 4) = '2015';
```

- **Explanation**: This SQL query was executed to find and list failed landings on a drone ship that took place in the year 2015.

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

GitHub URL

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

- **SQL Query** to 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

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

Landing_Outcome	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

- **Explanation:** This SQL query was used to rank in descending order the count of landing outcomes between June 4, 2010 and March 20, 2017.

[GitHub URL]()

Section 3

Launch Sites Proximities Analysis

Folium Map Locations of Launch Sites

- There are four launch sites, including CCAFS LC-40 (FL), CCAFS SLC-40 (FL), KSC LC-39A (FL), and VAFB SLC-4E (CA).
- The launch sites are located on the map using the folium.Circle and folium.Marker methods.

GitHub URL

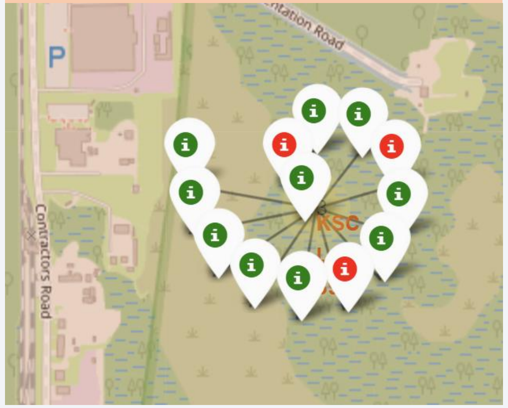
Folium Map of Launch Outcomes for Each Site

- For each launch site on the map, the landing outcomes are depicted using the folium.Circle and folium.Marker methods.
- Green and red markers indicate successful (class=1) and failed (class=0) landings, respectively.
- Among the four launch sites, KSC LC-39A achieved the highest success rate (~77%), whereas CCAFS LC-40 recorded the lowest (~27%).

[GitHub URL]

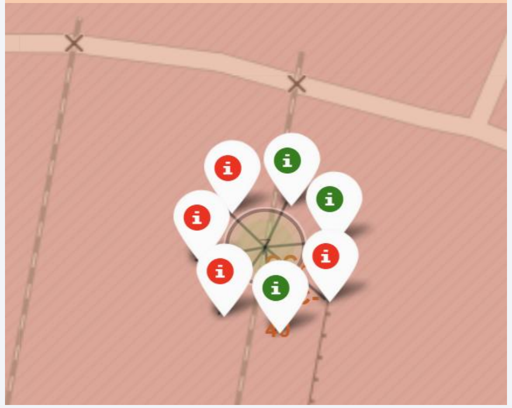
VAFB SLC-4E (CA)

KSC LC-39A (FL)

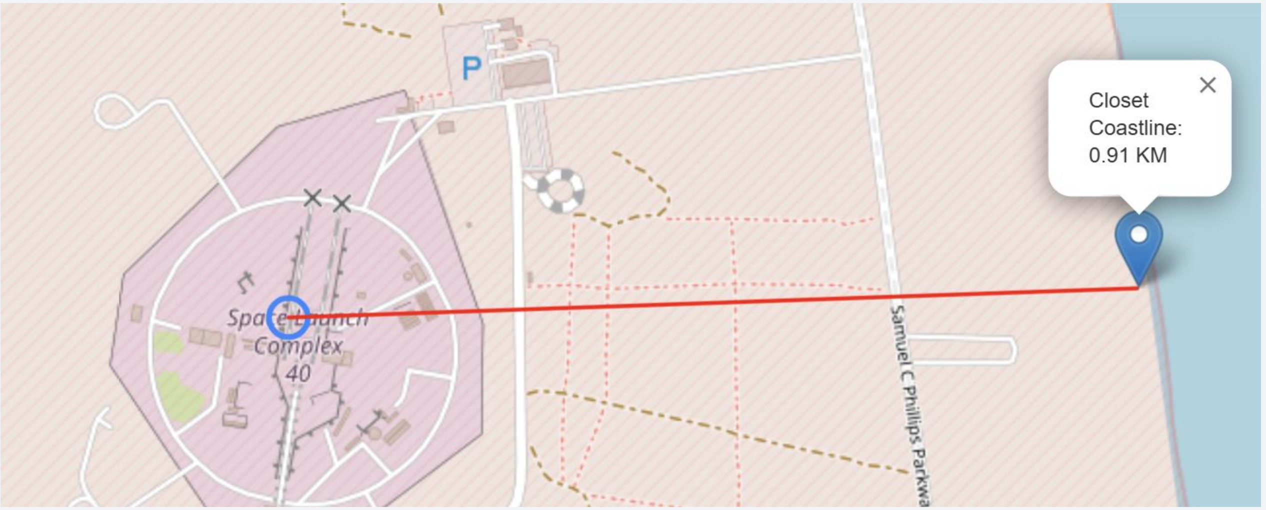


CCAFS LC-40 (FL)

CCAFS SLC-40 (FL)

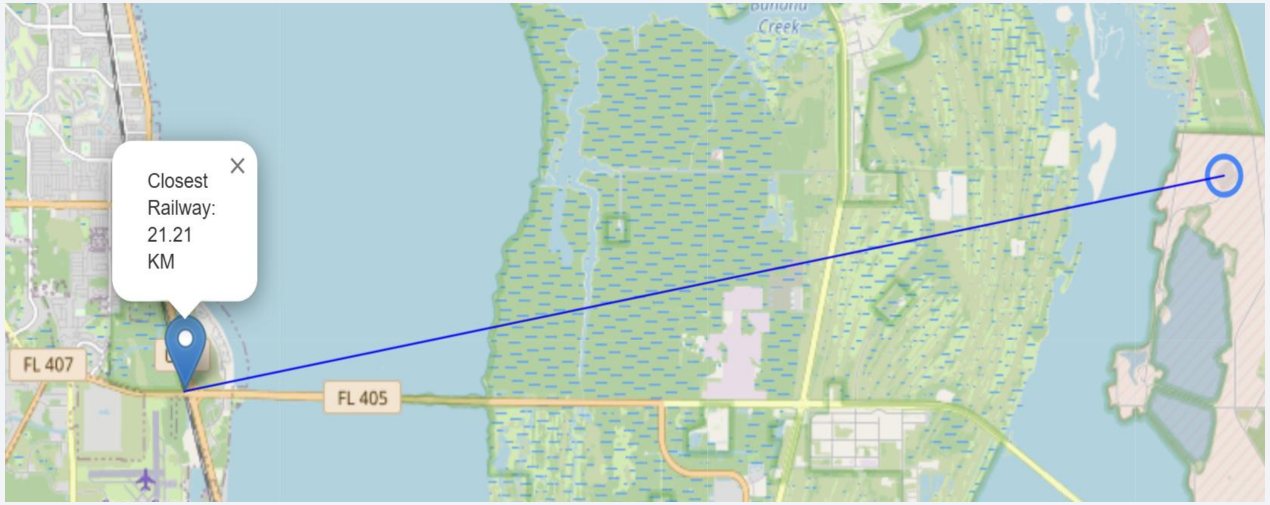


Folium Map of CCAFS SLC-40 Site to Its Proximities



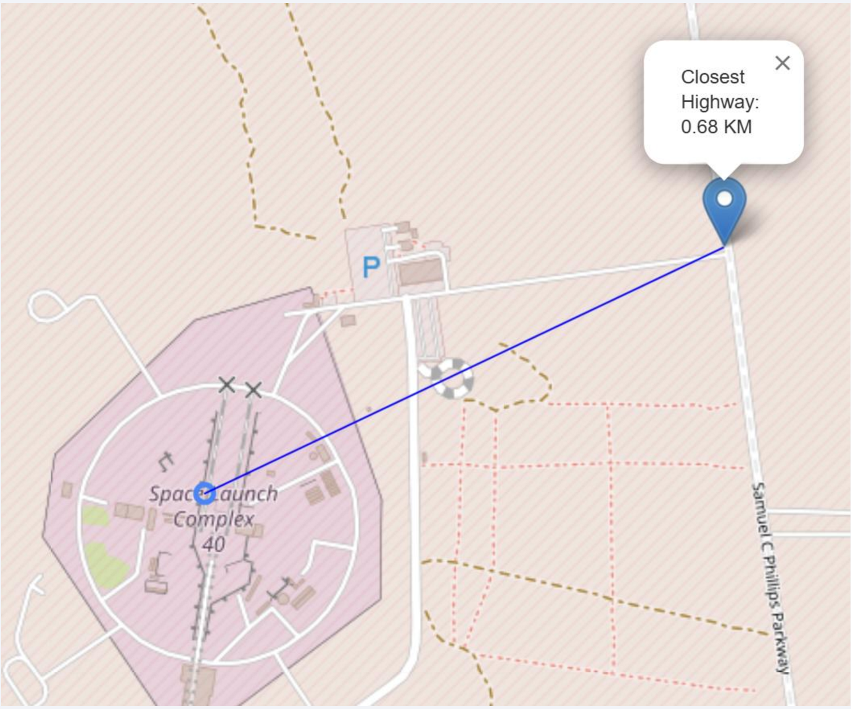
Proximity map showing the CCAFS SLC-40 Launch Site and the nearest coastline

[GitHub URL]

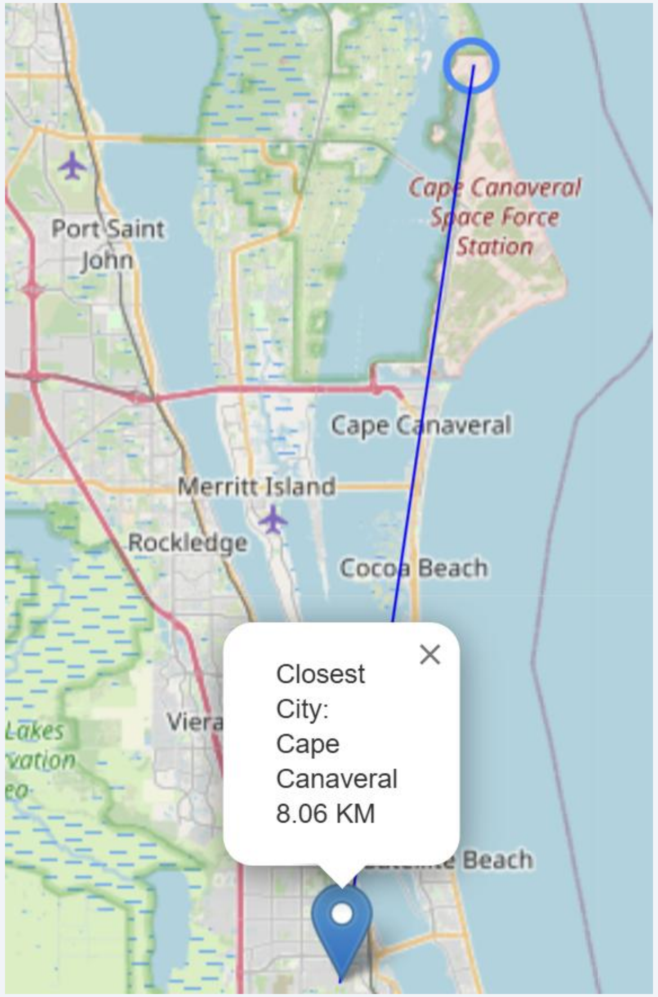


Proximity map displaying the CCAFS SLC-40 Launch Site and the nearest railway line

Folium Map of CCAFS SLC-40 Site to Its Proximities (*cont.*)



Proximity map showing the CCAFS SLC-40 Launch Site and the nearest highway



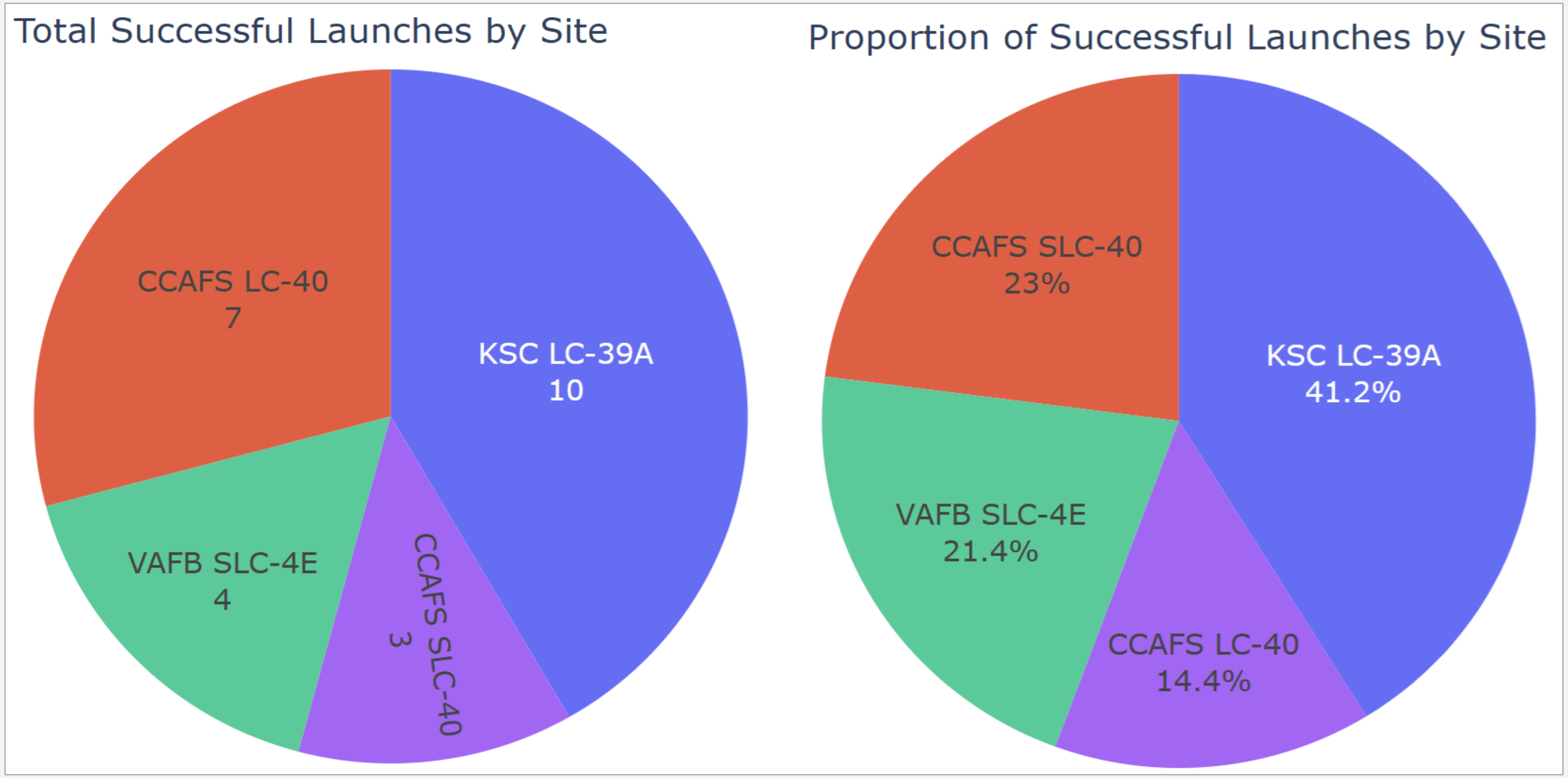
Proximity map displaying the CCAFS SLC-40 Launch Site and the nearest city

GitHub URL

Section 4

Build a Dashboard with Plotly Dash

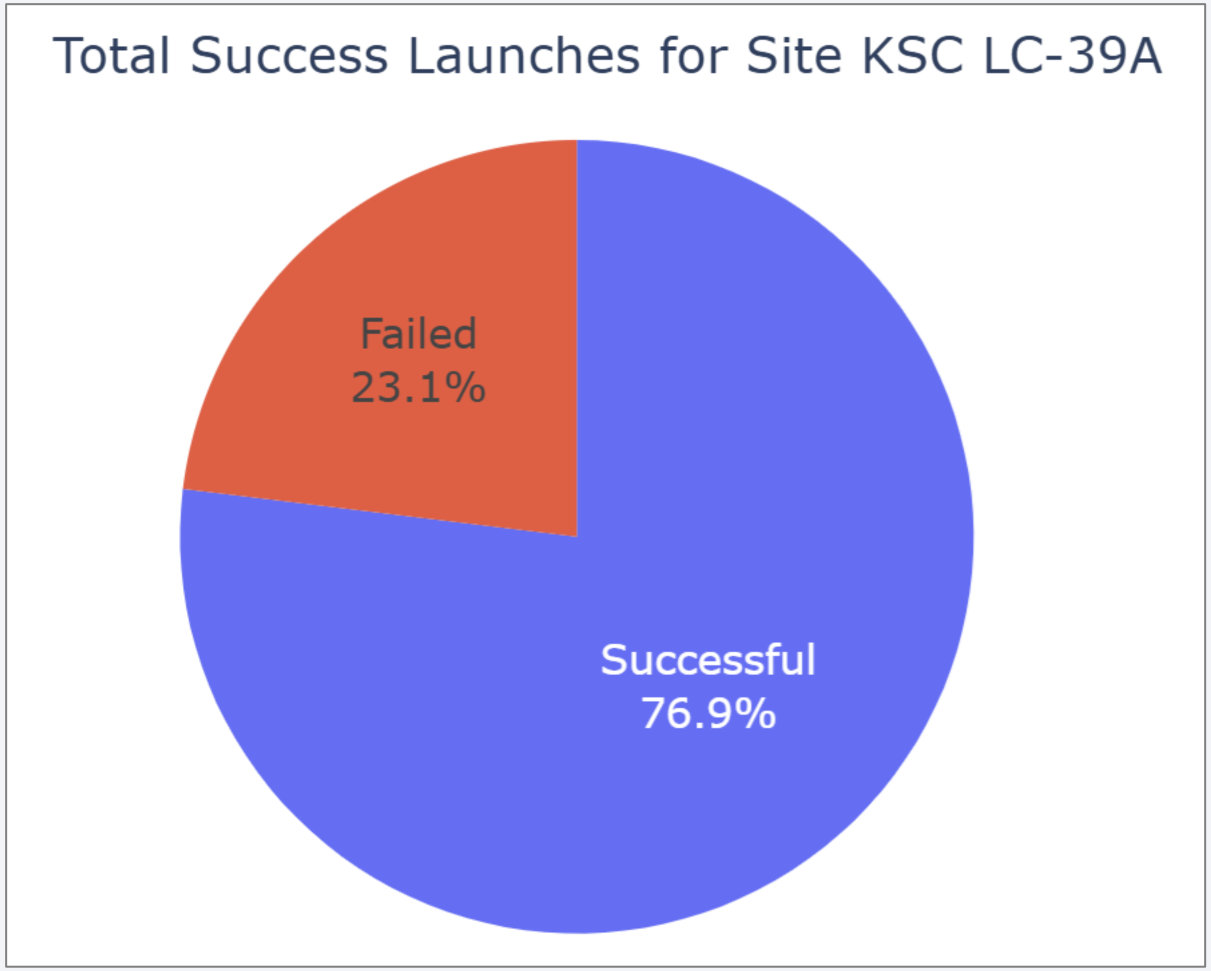
Pie Chart of Success Launches Across Sites



[GitHub URL]

The pie charts presented above show that the KSC LC-39A site accounted for 41.2% of all successful landings, whereas the CCAFS LC-40 site contributed only 14.4%.

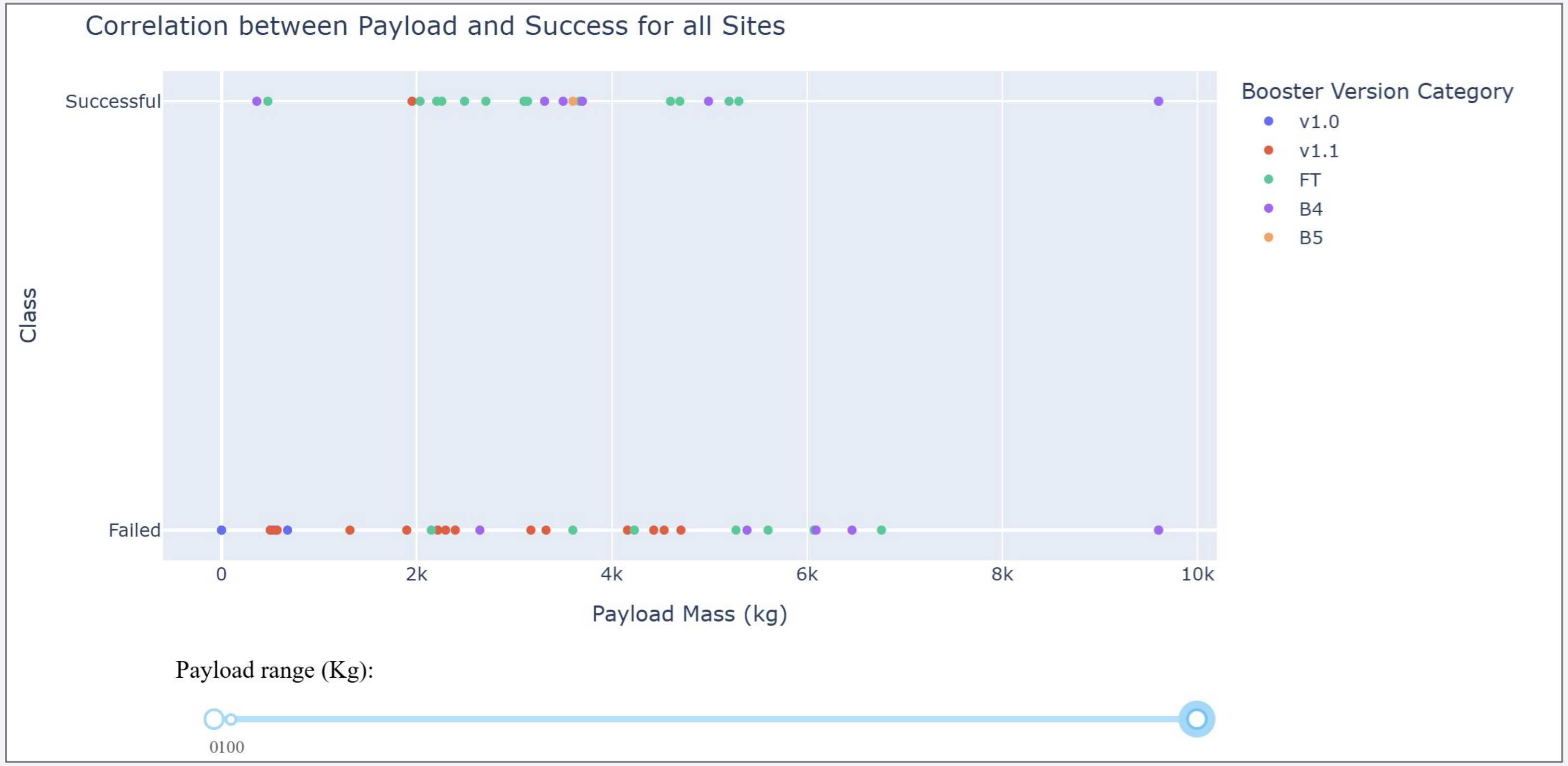
Pie Chart of Success Launches for KSC LC-39A Site



The pie chart illustrates that nearly 77 percent of launches at the KSC LC-39A site were successful, making it the most reliable launch site compared to others.

GitHub URL

Scatter Plot of Payload Mass vs. Landing Outcome



[GitHub URL]

As shown in the scatter plot above, the FT booster version had higher success rate, particularly for the payload range between 2,000 and 5,600 kg.

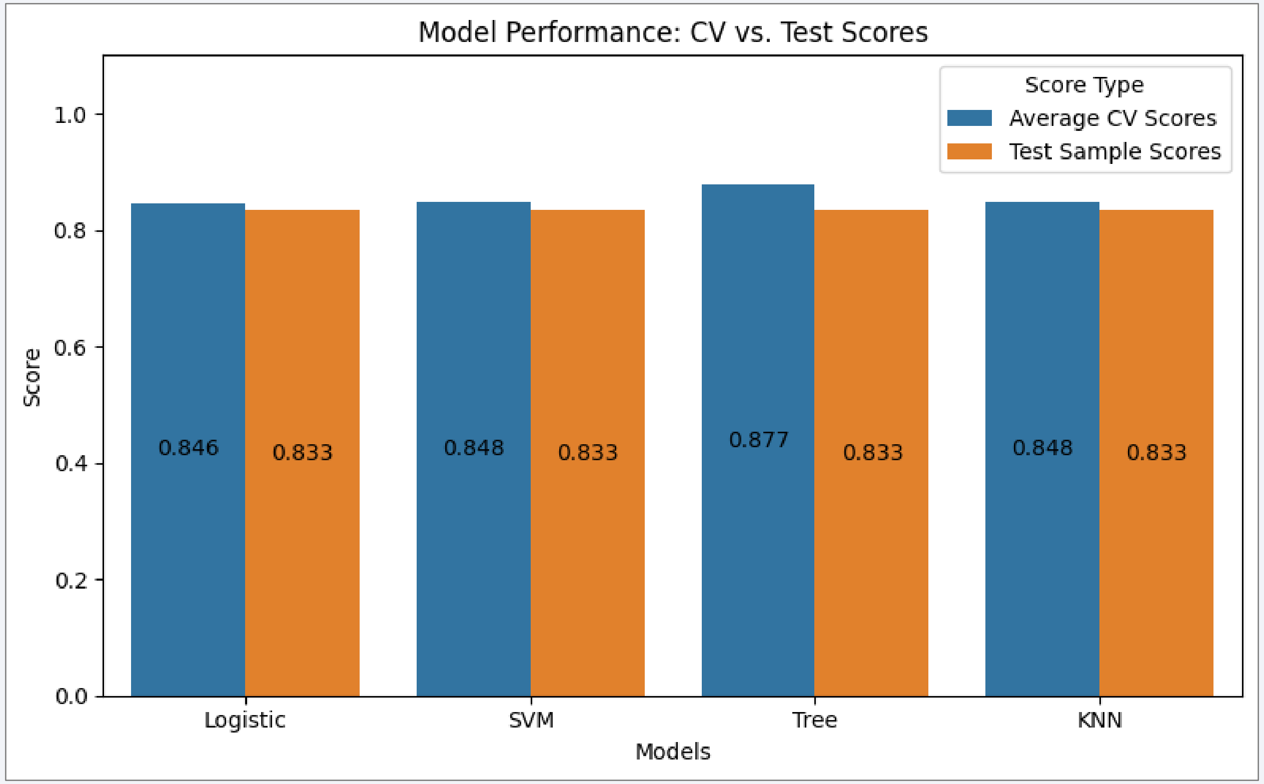
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Models	Average CV Scores	Test Sample Scores
Logistic	0.846	0.833
SVM	0.848	0.833
Tree	0.877	0.833
KNN	0.848	0.833

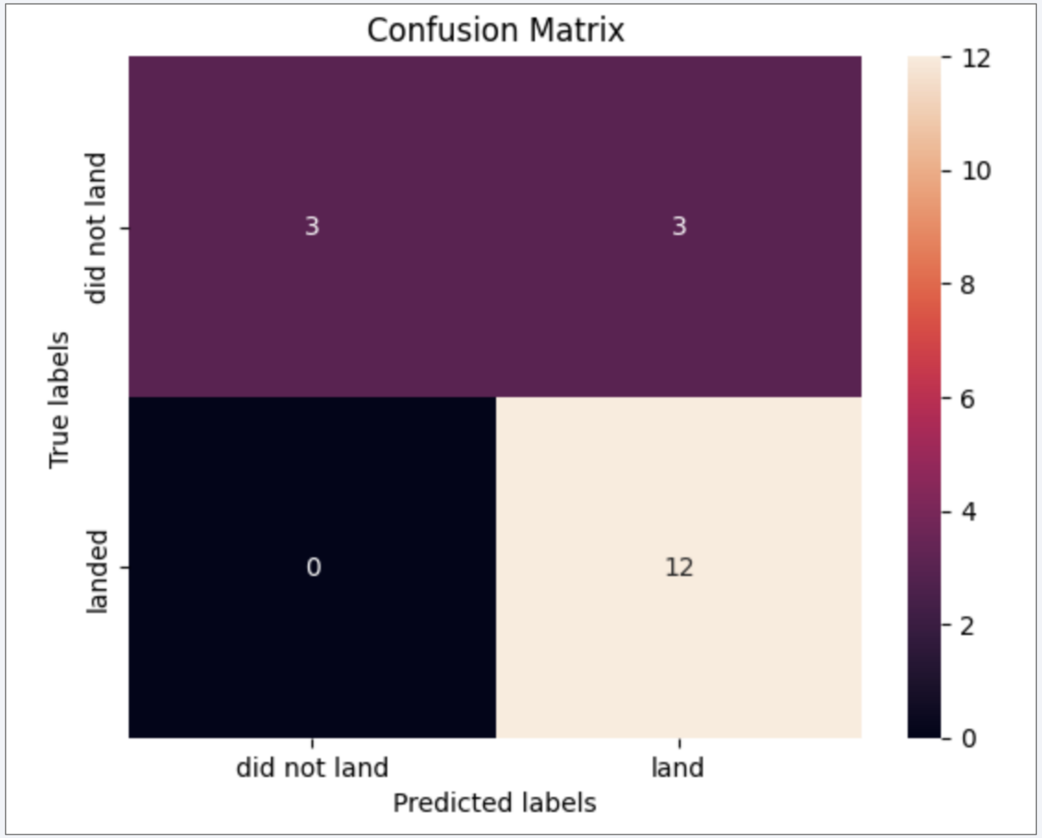
The results show that the decision tree classifier was preferred over the other candidate models based on its cross-validation score. However, all models demonstrated similar prediction performance on the test set.



[GitHub URL]

Confusion Matrix

The confusion matrix of the decision tree classifier, which is found as the best performing model according to the cross-validation accuracy measure. Note that all the four candidate models had similar performance on the test set.



[GitHub URL](#)

Conclusions

- This project focused on predicting the likelihood of successful first-stage landings of the Falcon 9 rocket.
- The rate of successful landings has increased over the years, especially after 2016, indicating improved performance over time.
- Over 41.2% of all successful landings occurred at the KSC LC-39A site, whereas 14.4% took place at the CCAFS LC-40 site.
- Seventy-seven percent of launches at the KSC LC-39A site landed successfully, making it making it the most reliable site for landing compared to others. In contrast, CCAFS LC-40 recorded the lowest success rate (~27%).

Conclusions (*cont.*)

- Compared to other orbit types, the VLEO, SSO, and LEO orbits had higher successful landings.
- The FT booster version exhibited higher success rate, particularly for the payload range between 2,000 and 5,600 kg.
- According to the cross-validation score, the decision tree model marginally outperformed the other candidate models. However, all four models demonstrated similar prediction performance on the test set.

Thank you!