IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tatiana Bouza Fortunato
October 16, 2024

Outline

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

Executive Summary

SpaceX stands as a leading force in the commercial space industry, primarily due to its cost-effective Falcon 9 rocket launches, priced at $62M compared to competitors' $165M. A key factor in this cost efficiency is SpaceX's ability to reuse the rocket's first stage. This project aims to estimate launch costs for a new company, SpaceY, by predicting the likelihood of first-stage reuse using machine learning.

Methodology Overview:

- Data collection via SpaceX API and web scraping from Wikipedia.
- Data wrangling, exploratory data analysis (EDA), and feature engineering.
- Launch site analysis using Folium and interactive dashboards with Plotly Dash.
- Machine learning pipeline to predict first-stage landing success, evaluating SVM, decision trees, and logistic regression models.

Key Findings:

- 61% of launches occurred at Cape Canaveral SLC-40.
- GTO (30%) and ISS (23%) were the most frequent orbits.
- 66% of landings were successful; overall, 99% of missions were successful.
- Most successful landings occurred at KSC LC-39 and CCAFS LC-40.
- The first successful ground-pad landing was achieved in 2015.
- Accuracy in predicting landing success on test data was nearly identical across Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors.

Introduction

The commercial space industry is rapidly evolving, with SpaceX emerging as a leader due to its relatively inexpensive rocket launches. A key factor contributing to SpaceX's cost-efficiency is the reuse of the Falcon 9 rocket's first stage, which allows them to offer launch services for $62 million—far below the $165 million typical of other providers. The ability to predict whether the first stage will land successfully is critical in determining overall launch costs.

This project aims to address that challenge by developing a machine learning model to **estimate launch prices for a new competitor**, Space Y, which seeks to enter the market and compete with SpaceX. Using public data on SpaceX's launches, **the model will predict the likelihood of first-stage reuse**, providing key insights into pricing strategies for Space Y.

Section 1

Methodology

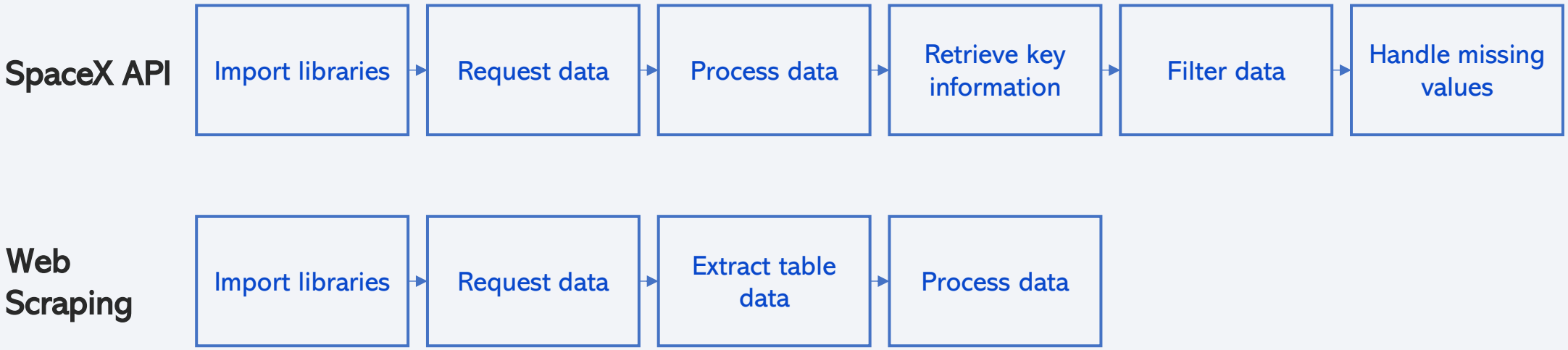
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Open Source REST API
  - Web scrape data from the Wikipedia page on Falcon 9 launches
- Perform data wrangling
  - Identify and handle null values, analyze column data types, and create a new column called 'class' derived from the outcome column for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform predictive analysis using classification models by standardizing the data, splitting it into training and test sets, and using GridSearchCV to find the best hyperparameters for SVM, Classification Trees, and Logistic Regression.

Data Collection

- Two datasets on SpaceX Falcon 9 launches were obtained: one from the SpaceX API and the other through web scraping a Wikipedia page. The data collection process is illustrated as follows:



Data Collection – SpaceX API

Import libraries and define auxiliary functions

Libraries:	Auxiliary functions:
Requests	def get BoosterVersion
Pandas	def get LaunchSite
NumPy	def get PayloadData
Datetime	def get CoreData

Request rocket launch data from SpaceX API

Use GET request method

Decode the response

Convert to JSON and Pandas dataframe.
Functions: *.json()* and *.json_normalize()*

Retrieve launch information using provided IDs

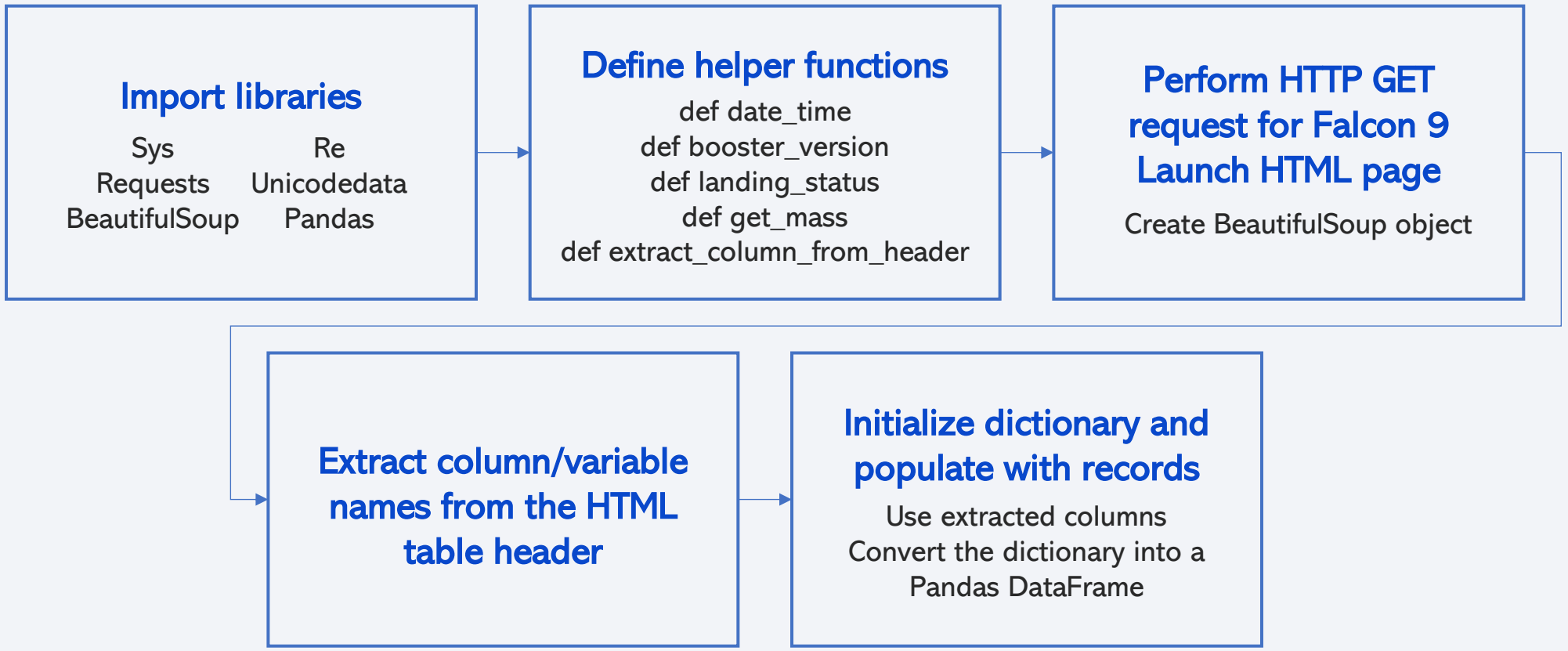
Focus on columns: rocket, payloads, launchpad, cores
Use auxiliary functions for additional data

Filter to include only Falcon 9 launches

Handle missing values

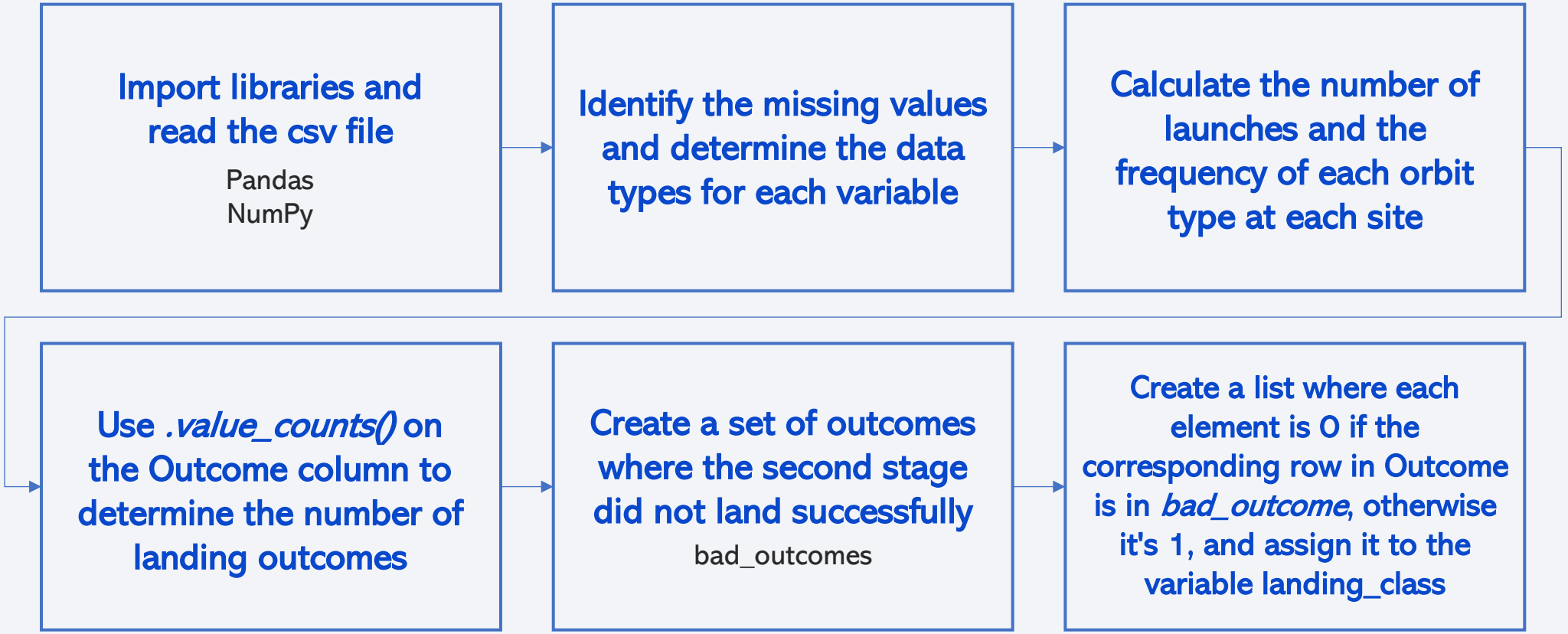
Click here to access the **GitHub Notebook**

Data Collection - Scraping



Click here to access the **GitHub Notebook**

Data Wrangling



Click here to access the **GitHub Notebook**

EDA with Data Visualization

Scatterplots were used to analyze the relationships between payload mass and flight number, launch site and flight number, launch site and payload mass, flight number and orbit type, and payload mass and orbit, with the class as the hue to indicate landing success. A bar chart was created to visualize the average success rate for each orbit type, and a line chart was used to show the yearly trend in launch success.

Click here to access the **GitHub Notebook**

EDA with SQL

SQL queries performed:

- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites began with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed the average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome on a ground pad was achieved.
- Listed the names of the boosters that successfully landed on a drone ship and carried a payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failed mission outcomes.
- Listed the names of the booster versions that carried the maximum payload mass using a subquery.
- Listed the records that displayed the month names, failed landing outcomes on a drone ship, booster versions, and launch sites for the months in the year 2015.
- Ranked the count of landing outcomes (e.g., Failure on a drone ship or Success on a ground pad) between the dates 2010-06-04 and 2017-03-20, in descending order.

Click here to access the **GitHub Notebook**

Build an Interactive Map with Folium

- Placed a circle at the coordinates of NASA Johnson Space Center with a popup label displaying its name.
- Added a marker at NASA Johnson Space Center's coordinates, using an icon with a text label to show its name.
- Created and added both a circle and a marker for each launch site on the map.
- Plotted markers for all launch records, using a green marker for successful launches and a red marker for failed ones.
- Implemented a MousePosition feature to display the coordinates of the mouse pointer as it hovers over any point on the map.
- Drew a polyline from each launch site to its nearest coastline, placing a marker at the closest point on the coastline and displaying the distance between the two points.

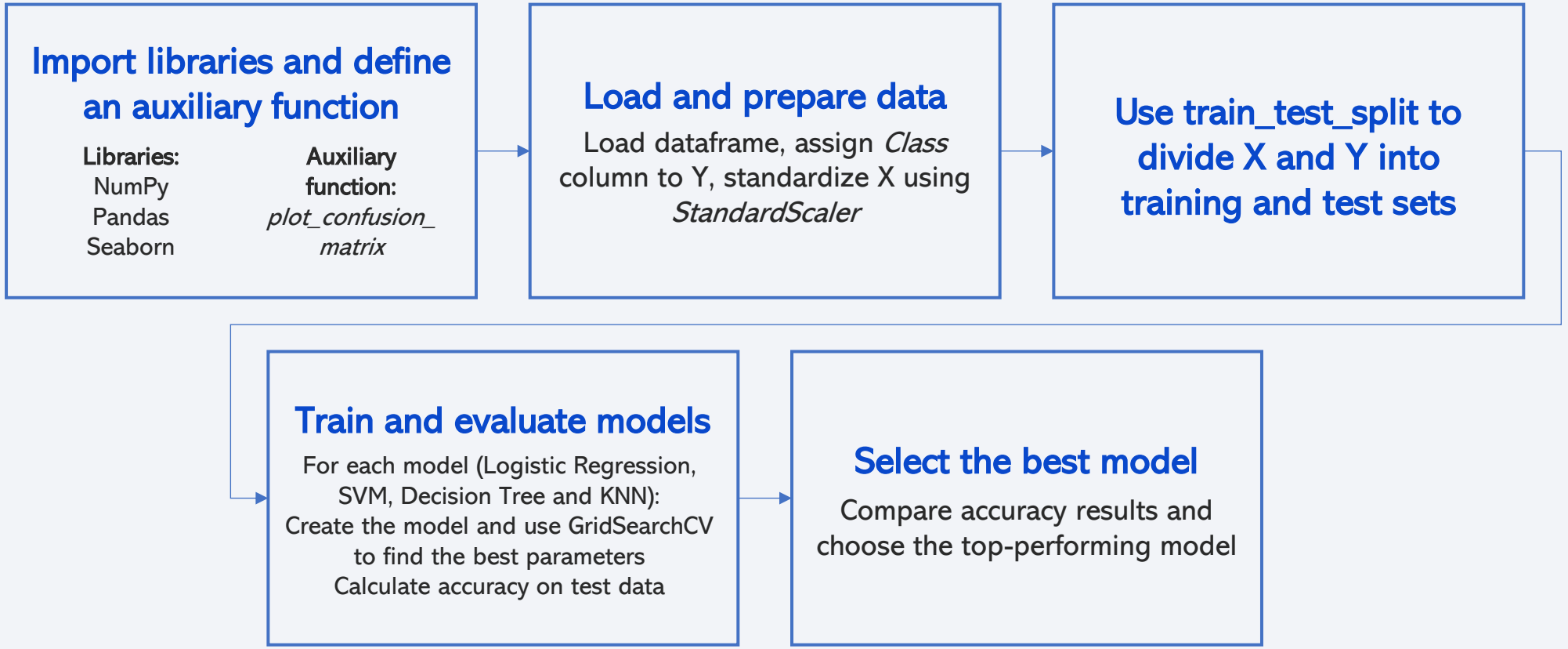
Click here to access the **GitHub Notebook**

Build a Dashboard with Plotly Dash

- Added a launch site dropdown input component to allow users to select different launch sites.
- Implemented a callback function that renders a pie chart displaying the total number of successful launches for the selected launch site.
- Integrated a range slider to filter payload ranges and explore how payload size may correlate with mission outcomes based on the selected launch site.
- Added a callback function to generate a scatter plot, where the x-axis represents payload and the y-axis represents launch outcome. Each scatter point is color-coded by booster version, allowing us to observe how different boosters impact mission outcomes.

Click here to access the **Plotly Dash App**

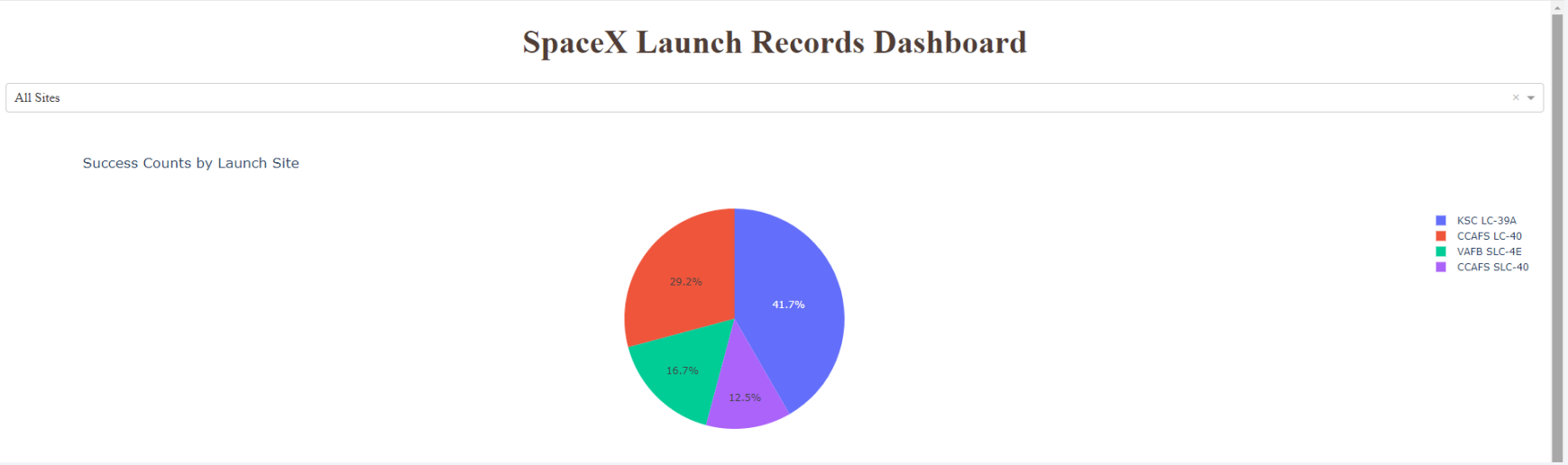
Predictive Analysis (Classification)

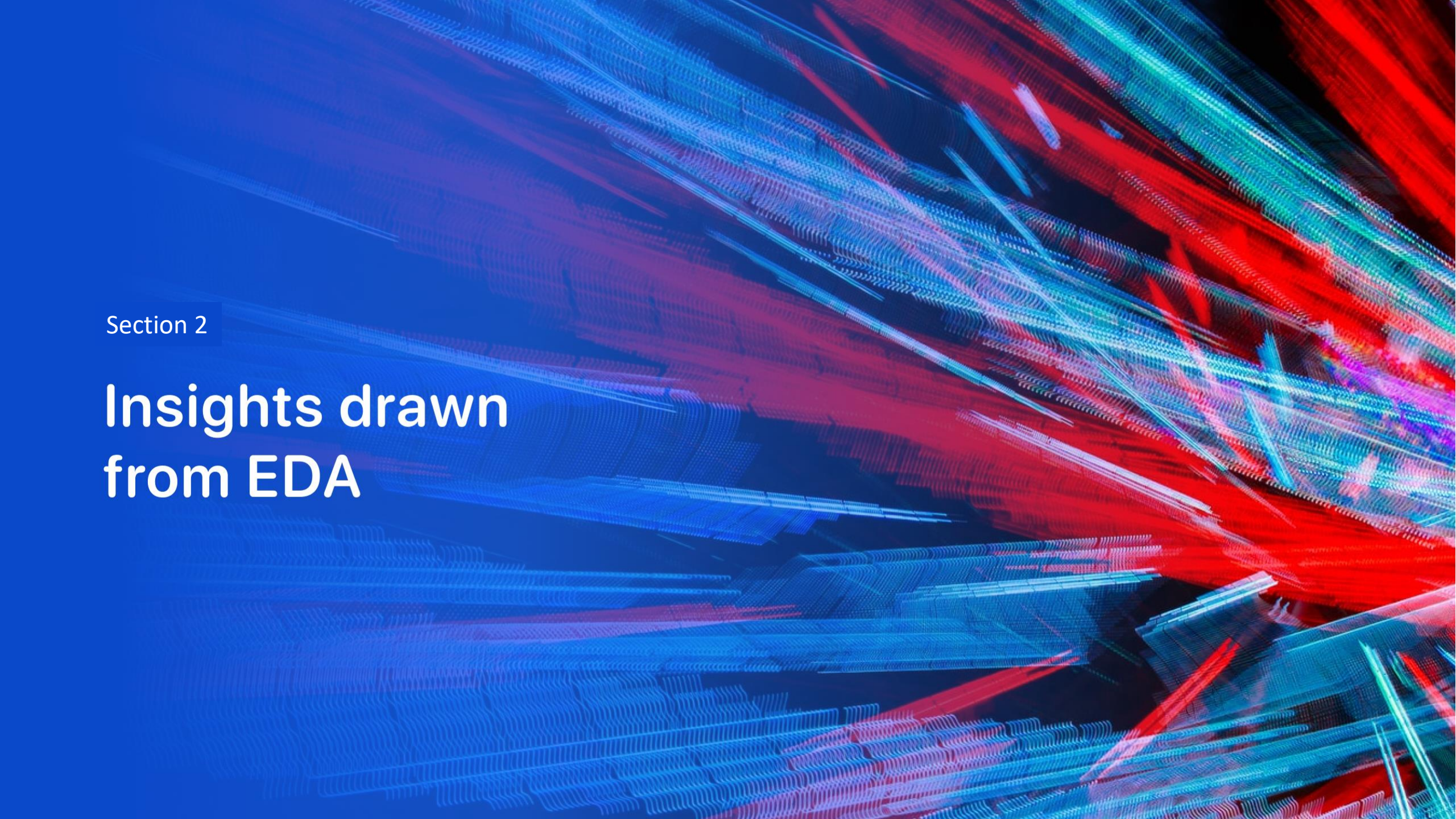


Click here to access the **GitHub Notebook**

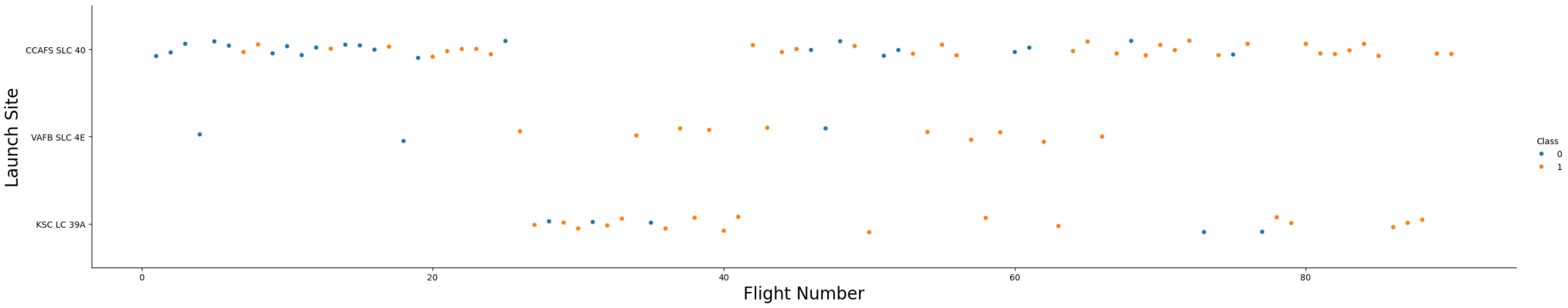
Results

- 61% of launches occurred at Cape Canaveral SLC-40.
- GTO (30%) and ISS (23%) were the most frequent orbits.
- 66% of landings were successful; overall, 99% of missions were successful.
- Most successful landings occurred at KSC LC-39 and CCAFS LC-40.
- The first successful ground-pad landing was achieved in 2015.
- Accuracy in predicting landing success on test data was nearly identical across Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors.



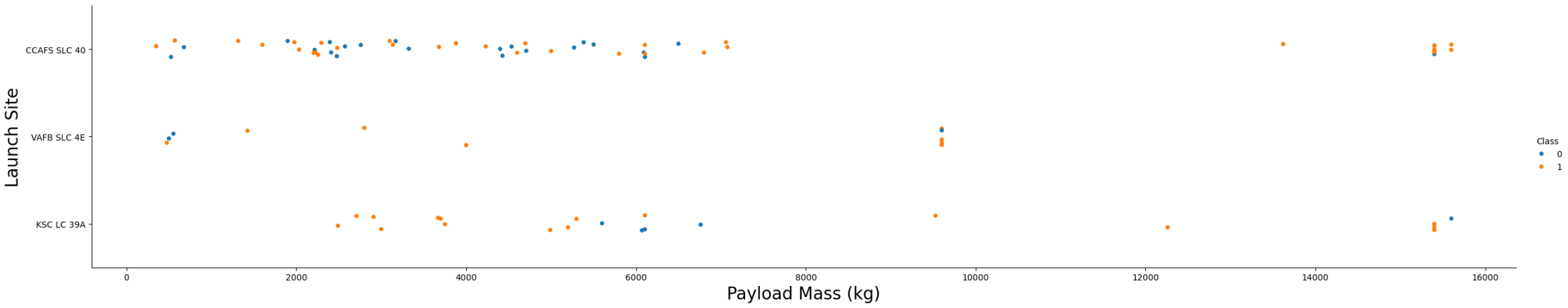

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



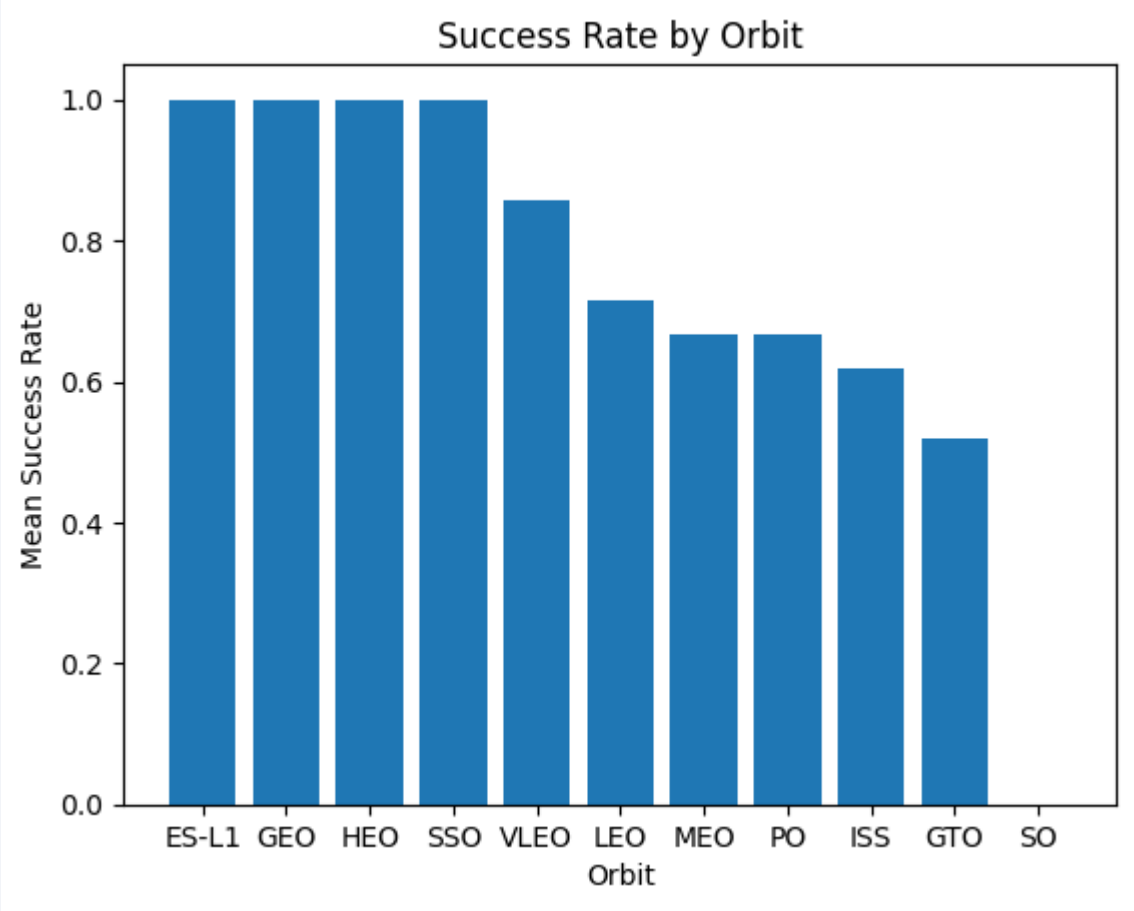
Launches from CCAFS SLC 40 were significantly more frequent than from other sites. In its early missions, SpaceX primarily used CCAFS SLC 40, with approximately 39% of flights failing to land successfully. After about 25 flights, SpaceX shifted to KSC LC 39A, where most landings were successful. They also launched from VAFB SLC 4E, achieving a 100% success rate. Around the 40th flight, SpaceX resumed launches at CCAFS SLC 40 and began seeing improved landing success compared to earlier missions.

Payload vs. Launch Site



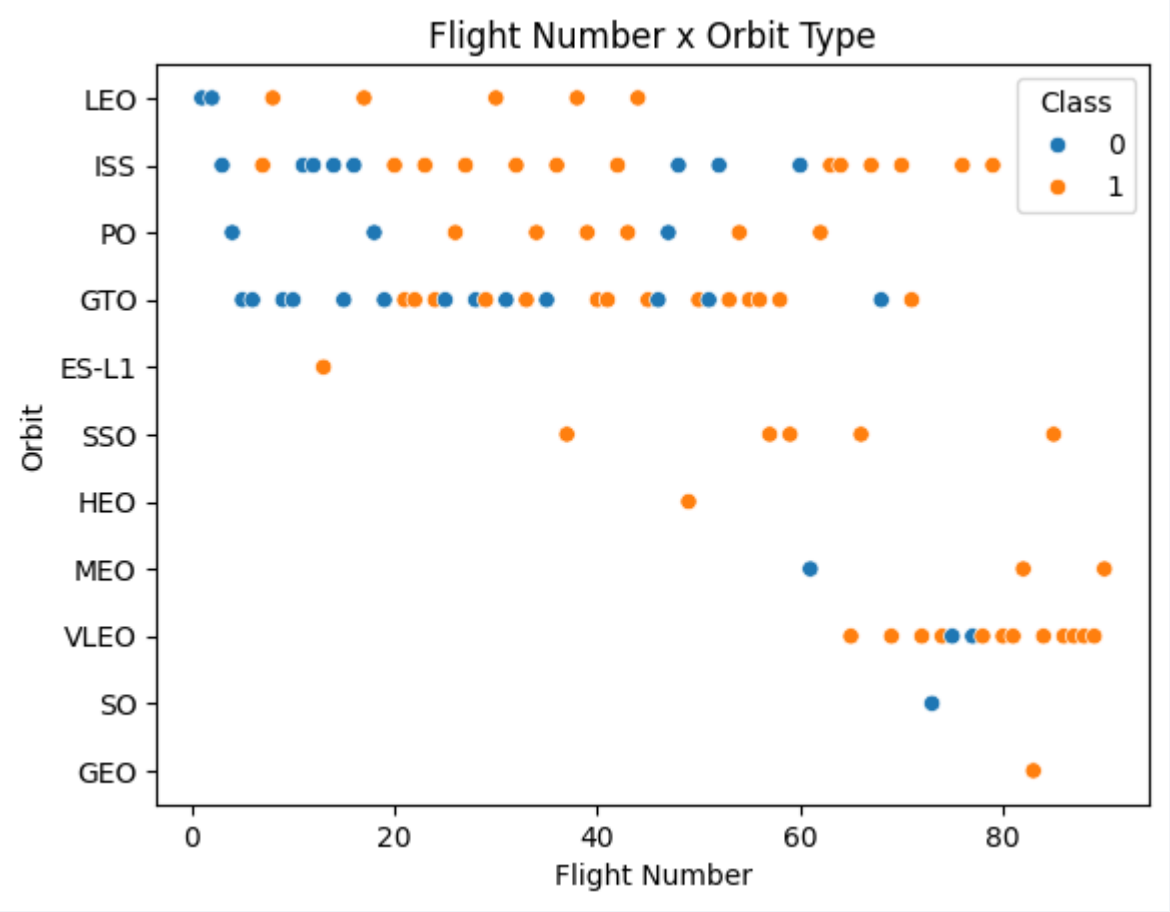
Most launches carry a payload mass of less than 8,000 kg. Additionally, no rockets with heavy payloads (over 10,000 kg) have been launched from the VAFB SLC site.

Success Rate vs. Orbit Type



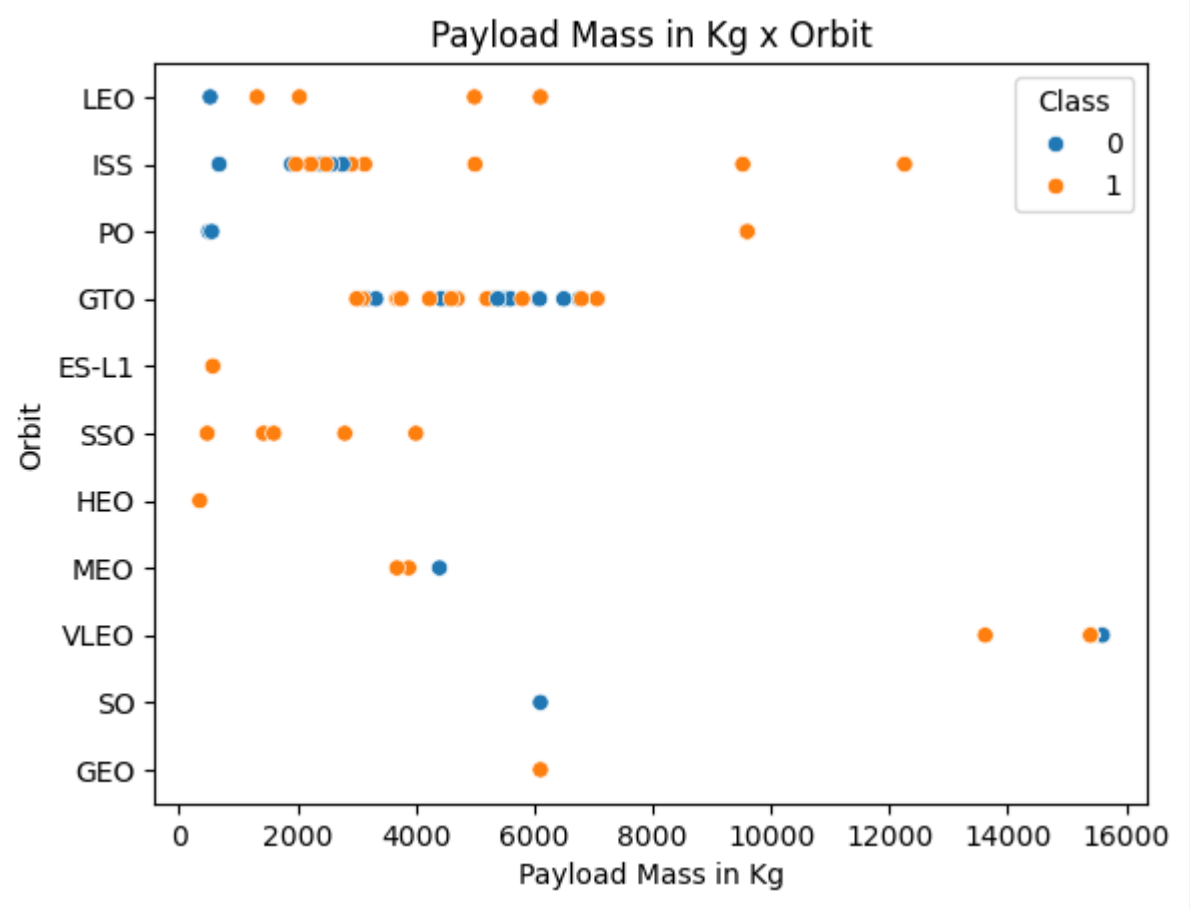
ES-L1, GEO, HEO and SSO orbits have the highest success rates.

Flight Number vs. Orbit Type



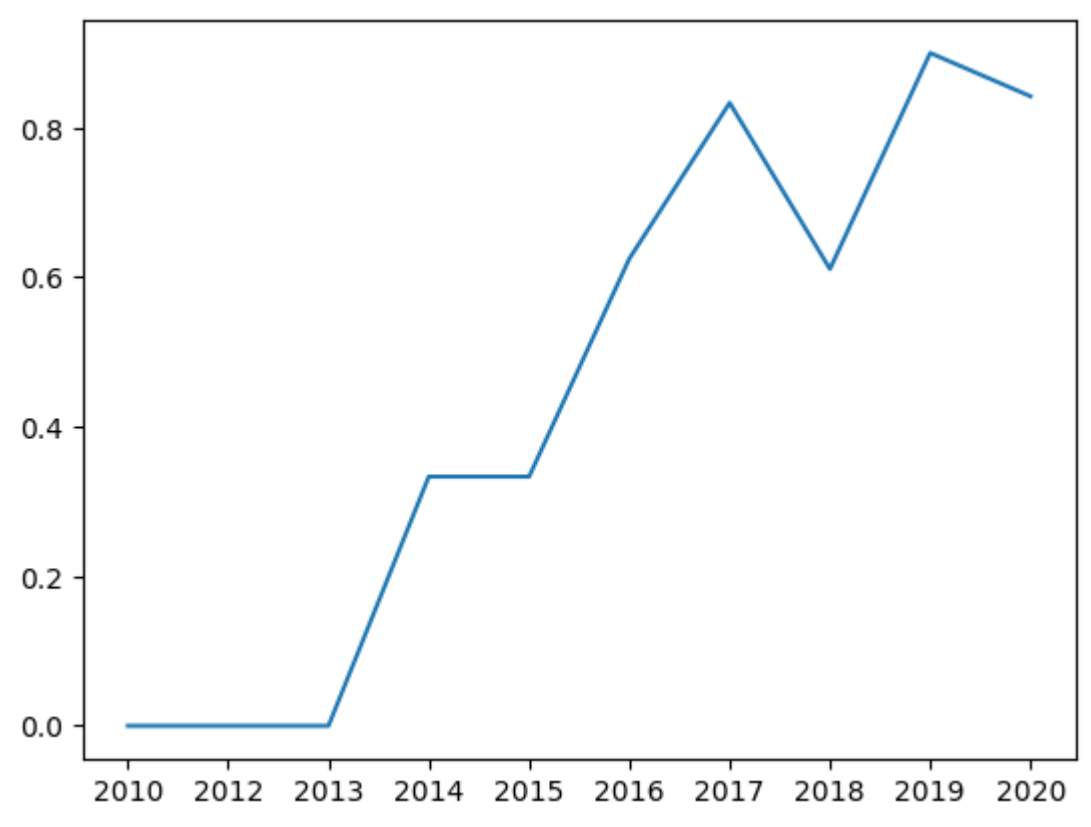
In LEO orbit, success appears to be linked to the number of flights. However, in GTO orbit, there seems to be no correlation between flight number and success.

Payload vs. Orbit Type



For heavy payloads, the landing success rate is higher in LEO, ISS, and PO orbits. However, in GTO orbit, it is difficult to distinguish between successful and unsuccessful landings, as both outcomes are common.

Launch Success Yearly Trend



The success rate steadily increased from 2013 to 2020. However, around 2018, there was a dip from approximately 0.8 to 0.6, followed by a quick recovery to above 0.8.

All Launch Site Names

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

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

The names of the unique launch sites are:

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

Launch Site Names Begin with 'CCA'

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

```
 * sqlite:///my_data1.db
Done.
```

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

Here are 5 records where the launch site names start with “CCA”.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer=="NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

The total payload carried by NASA boosters is 45,596 kg.

Average Payload Mass by F9 v1.1

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

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by the F9 v1.1 booster version is 2,928.4 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" == "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

The first successful ground pad landing occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
 * sqlite:///my_data1.db
Done.
```

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

Here are the names of boosters that successfully landed on a drone ship with a payload mass between 4,000 and 6,000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

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

Here are the total numbers of successful and failed mission outcomes.

Boosters Carried Maximum Payload

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

```
 * sqlite:///my_data1.db
Done.
```

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

Here are the names of the booster versions that have carried the maximum payload mass.

2015 Launch Records

```
%sql SELECT SUBSTR(Date, 6, 2) AS "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = "2015" AND "Landing_Outcome" = "Failure (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

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

Here are the failed landing outcomes in drone ship, along with their booster versions and launch site names in 2015.

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

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

```
 * sqlite:///my_data1.db
Done.
```

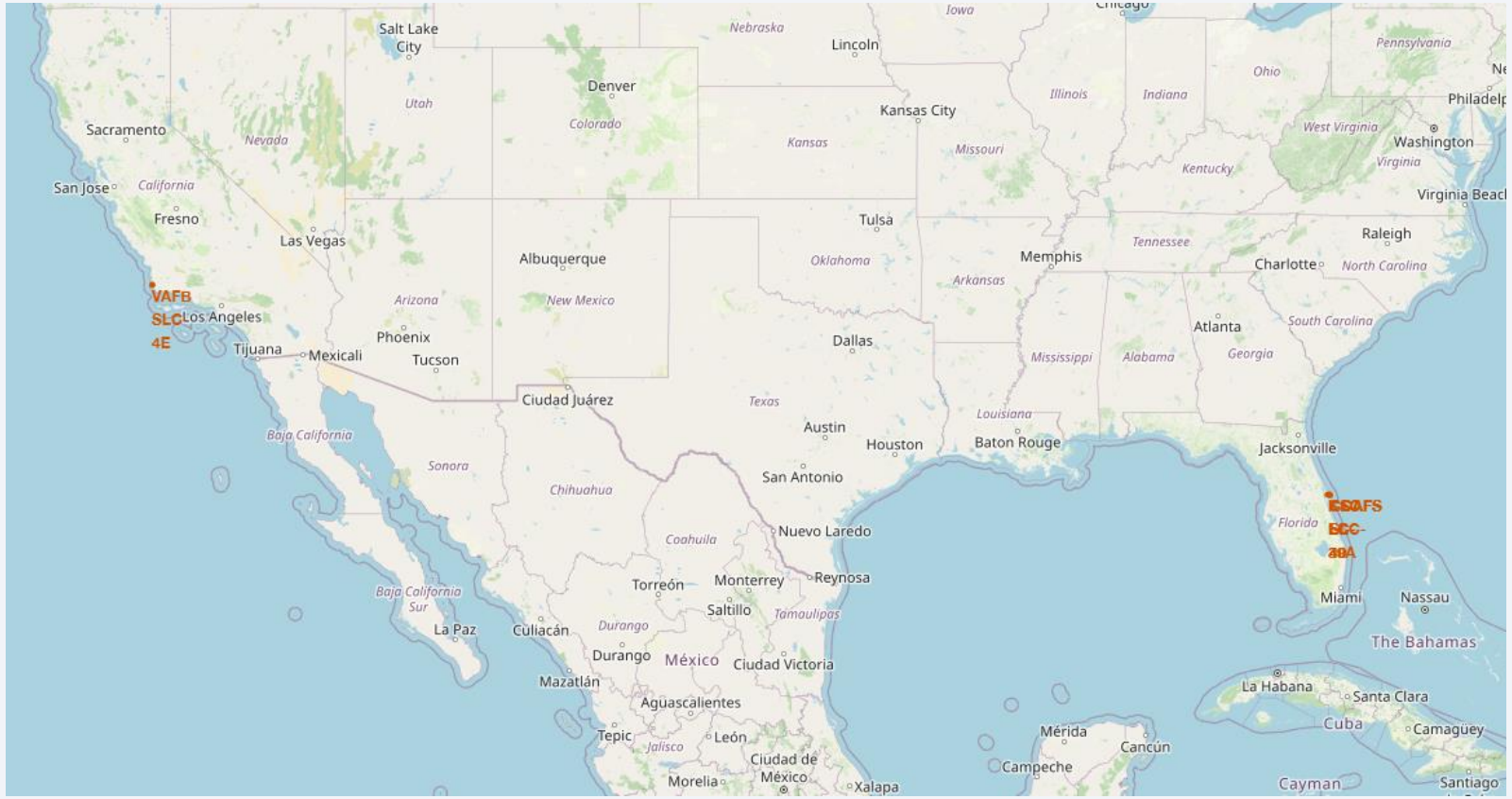
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Here are the ranked counts of landing outcomes between June 4, 2010, and March 20, 2017.

Section 3

Launch Sites Proximities Analysis

Map of All Launch Sites

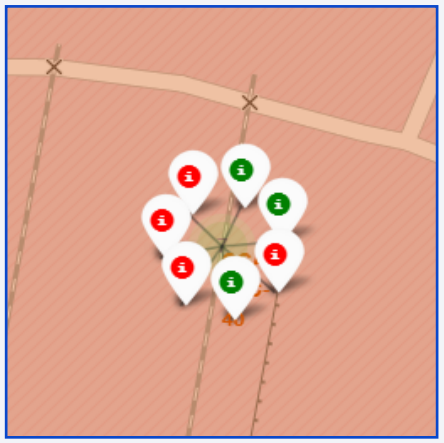


All launch sites are located in close proximity to the coast.

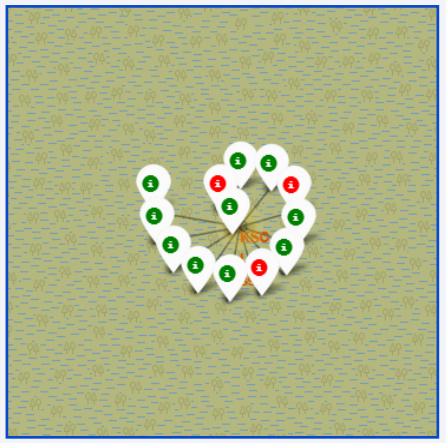
Launch outcomes color-coded on the map

CCAFS LC-40

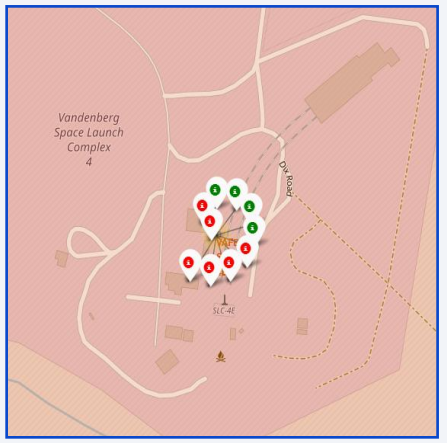
CCAFS SLC-40



KSC LC-39A



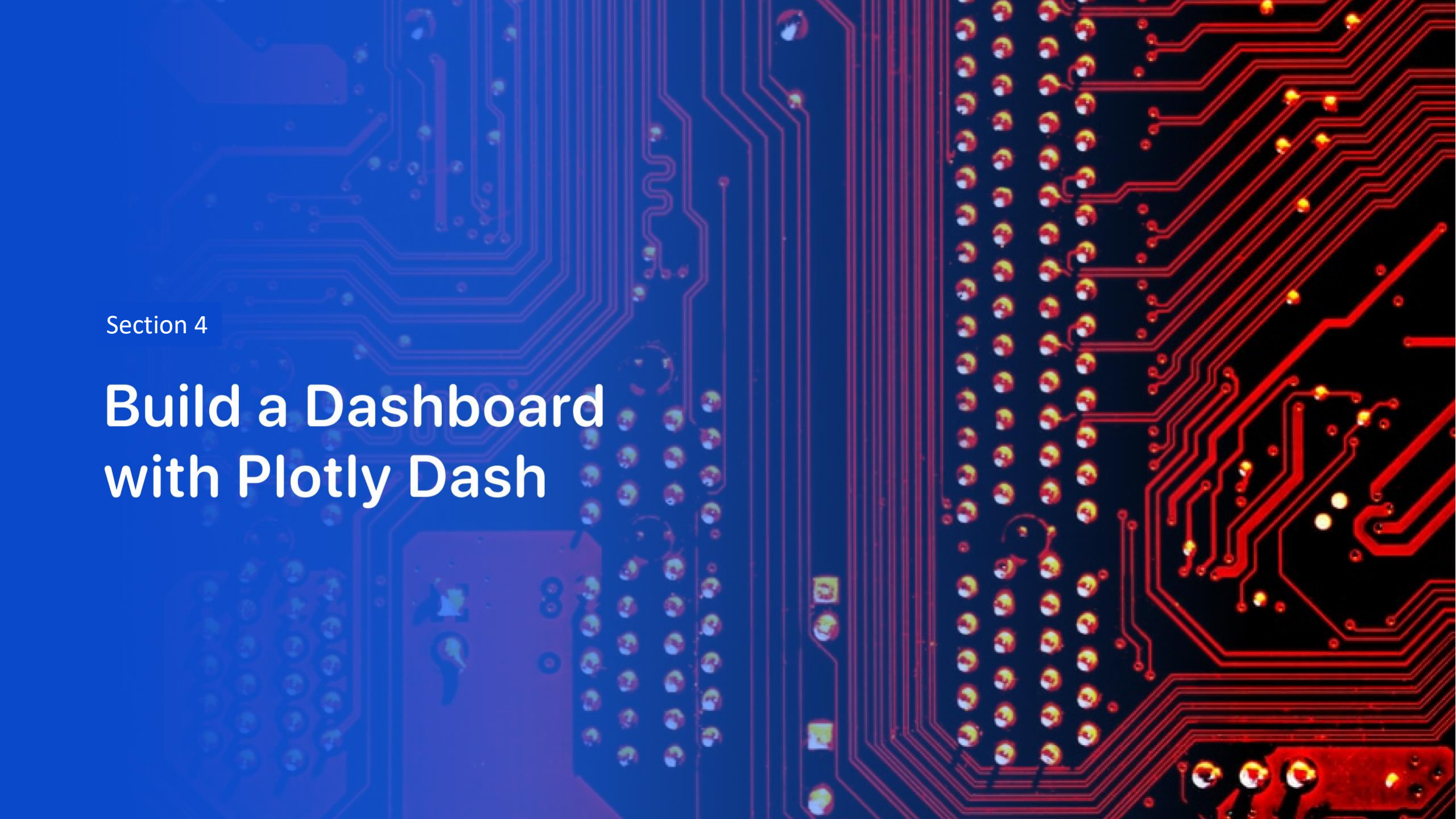
VAFB SLC-4E



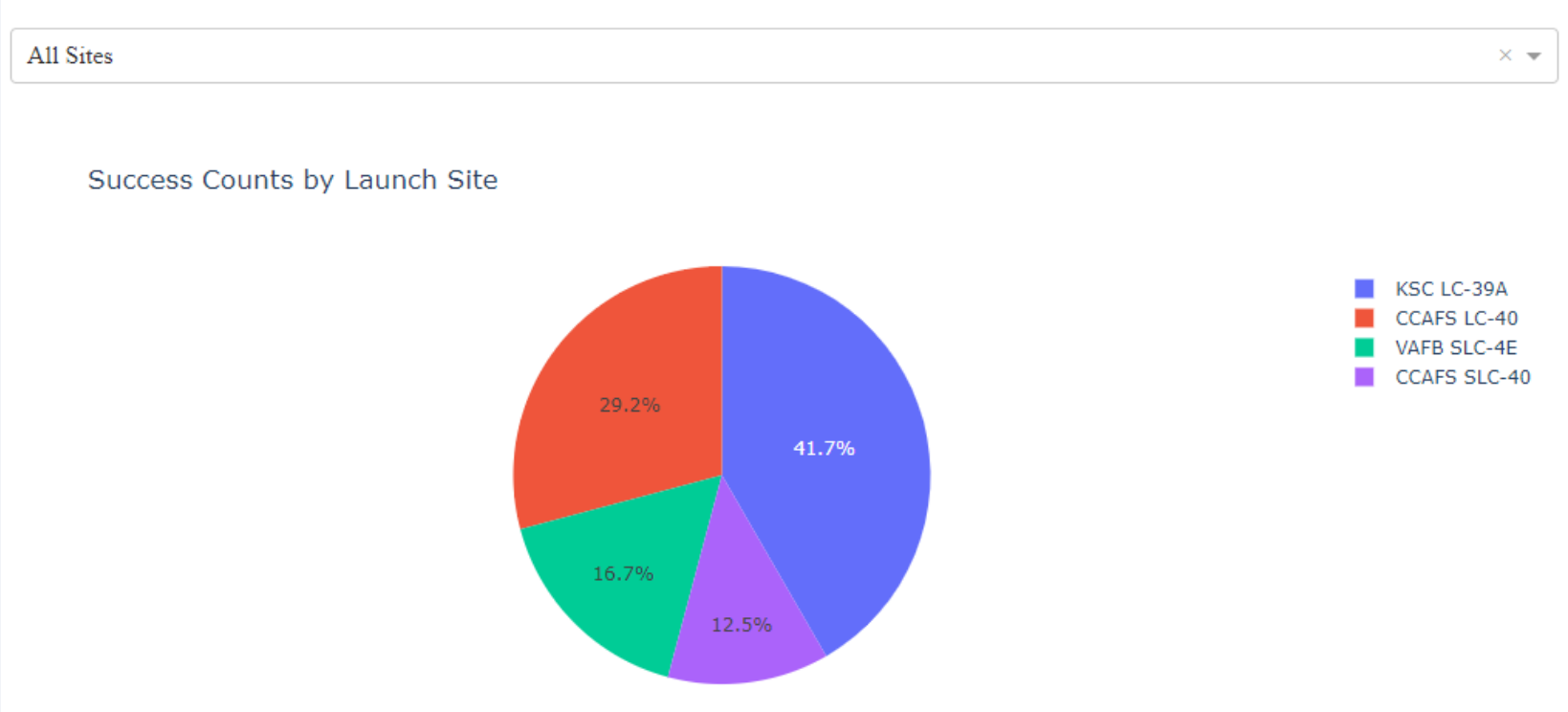
While most launches took place at Cape Canaveral Space Launch Complex 40, the majority of successful landings occurred at both KSC LC-39 and CCAFS LC-40.

Map of the CCAFS SLC-40 site and its surrounding areas

Lines were drawn to measure the distance from the CCAFS SLC-40 site to the nearest highway (0.58 km) and coastline (0.85 km).

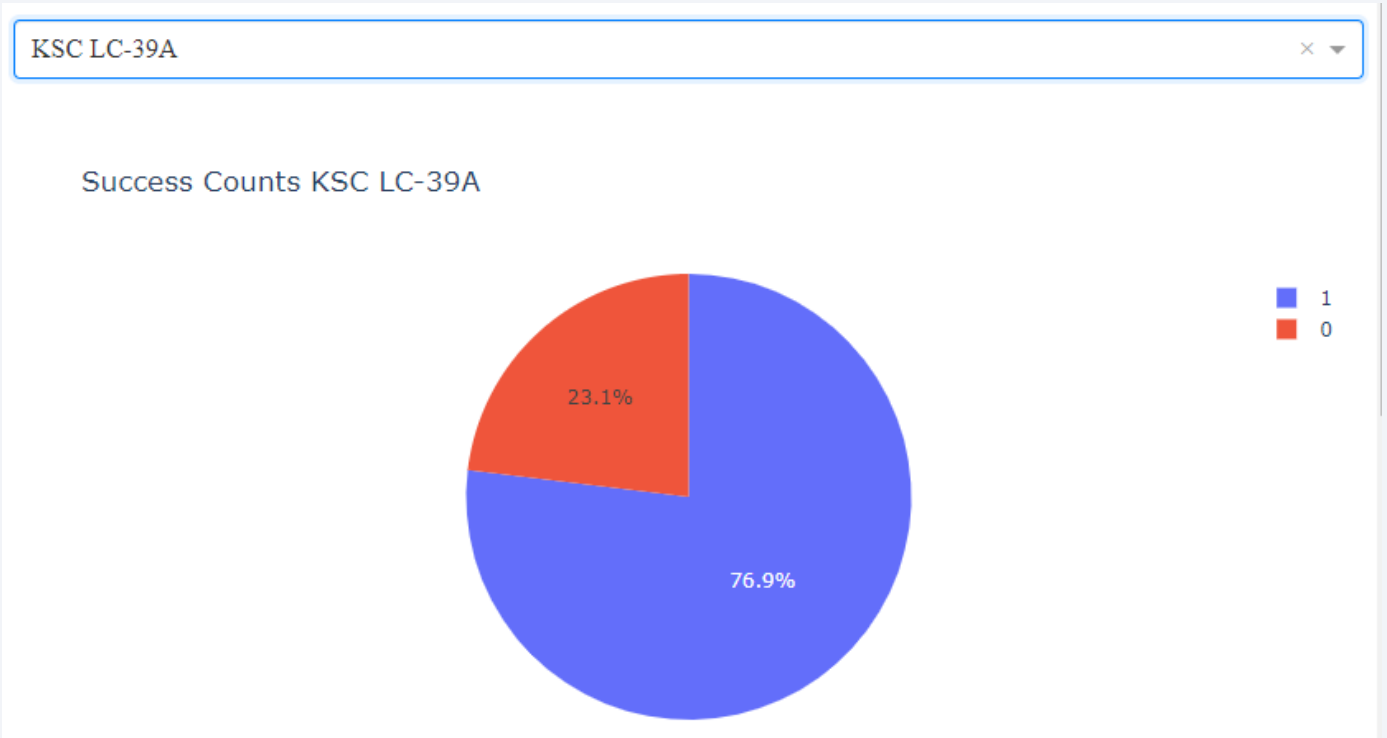

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count by Site



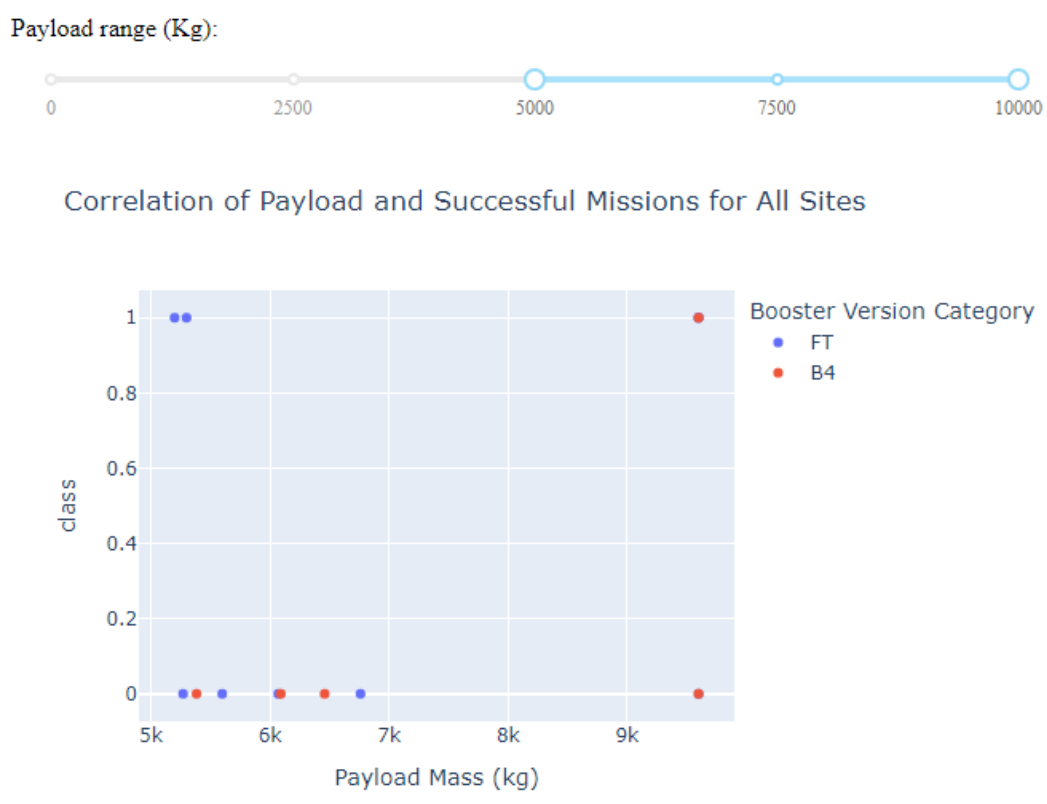
KSC LC-39A had the highest number of successful launches, while CCAFS SLC-40 had the fewest.

Launch Success Breakdown at KSC LC-39A



"At KSC LC-39A, the site with the highest number of successful launches, 76.9% of all launches were successful."

Payload vs. Launch Outcome Across All Sites

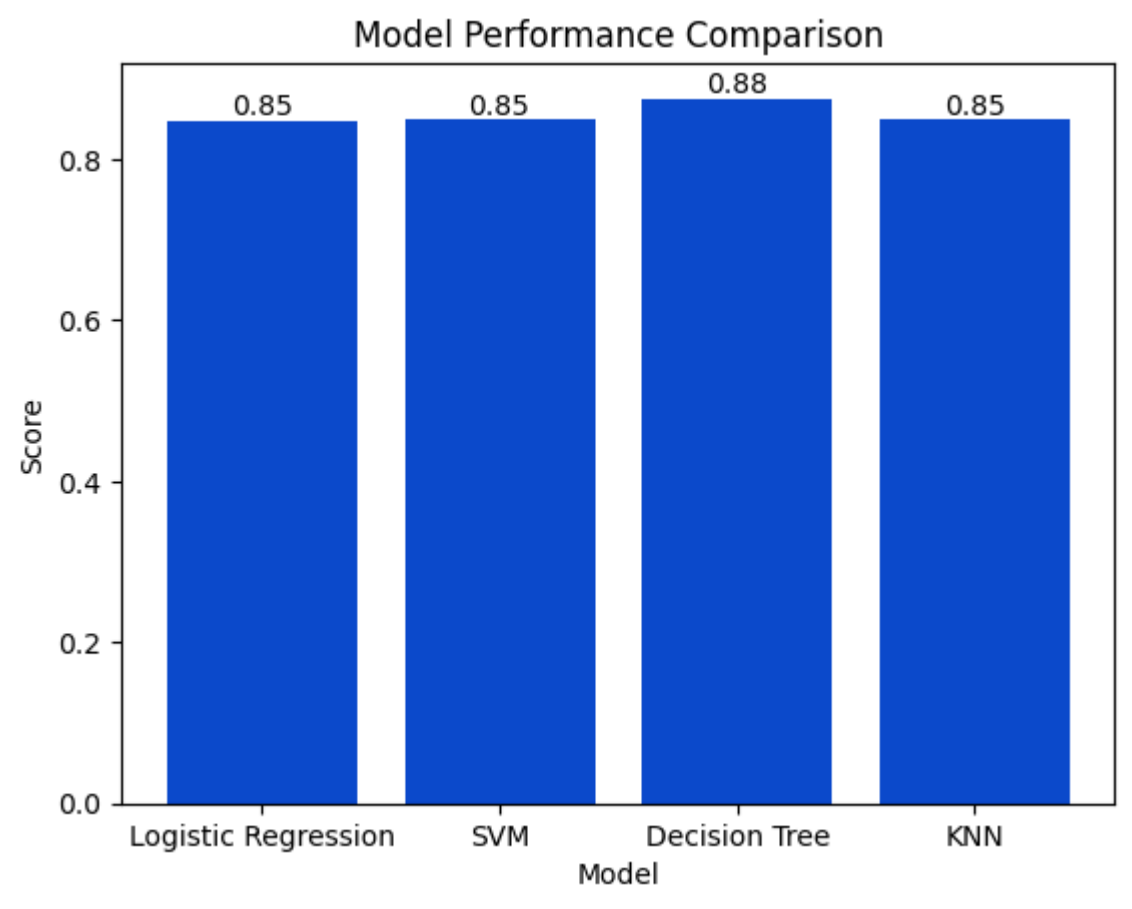


The payload range with the highest success rate is under 5000 kg, while the booster version with the best success rate is FT. Conversely, most missions with a payload mass exceeding 5000 kg resulted in failure.

Section 5

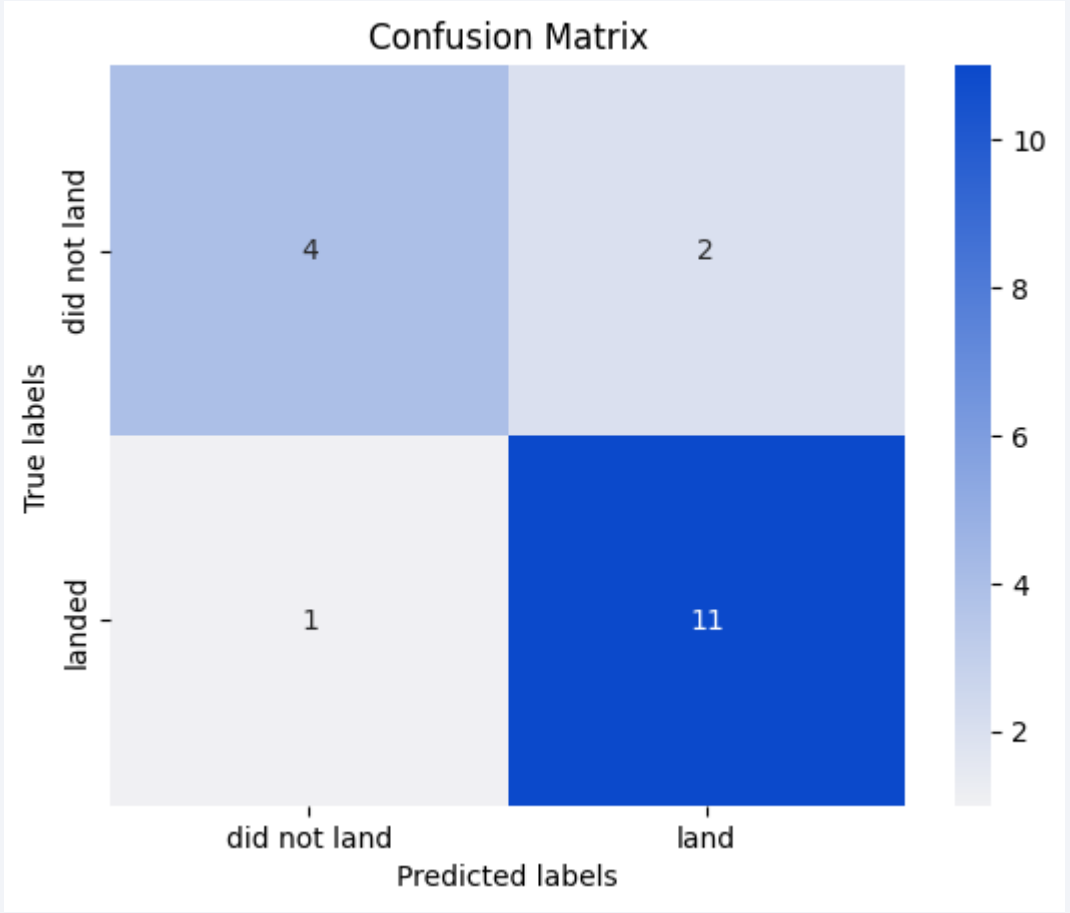
Predictive Analysis (Classification)

Classification Accuracy



The model that achieved the highest classification accuracy is the Decision Tree.

Confusion Matrix



This is the confusion matrix for the Decision Tree model, our best-performing model. It reveals 1 False Negative and 2 False Positives.

Conclusions

- 61% of launches occurred at Cape Canaveral SLC-40.
- GTO (30%) and ISS (23%) were the most frequent orbits.
- 66% of landings were successful; overall, 99% of missions were successful.
- Most successful landings occurred at KSC LC-39 and CCAFS LC-40.
- The first successful ground-pad landing was achieved in 2015.
- Accuracy in predicting landing success on test data was nearly identical across Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors.

Thank you!