IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amr Ezzat
08 June 2025

Outline

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

Executive Summary

- This project comprehensively utilized data from the SpaceX API and web scraping, followed by extensive EDA (visual & SQL) and interactive analytics (Folium, Dash), culminating in the development and optimization of machine learning classification models to predict Falcon 9 landing outcomes.
- The Decision Tree Classifier performed the best on the test data with an accuracy of 87.5%

Introduction

- Project Goal:
  - To predict if a rocket will achieve a successful landing, given various flight predictors.
- Key Questions:
  - Identify the most influential predictors for landing outcomes and determine the accuracy of our predictions.

Section 1

Methodology

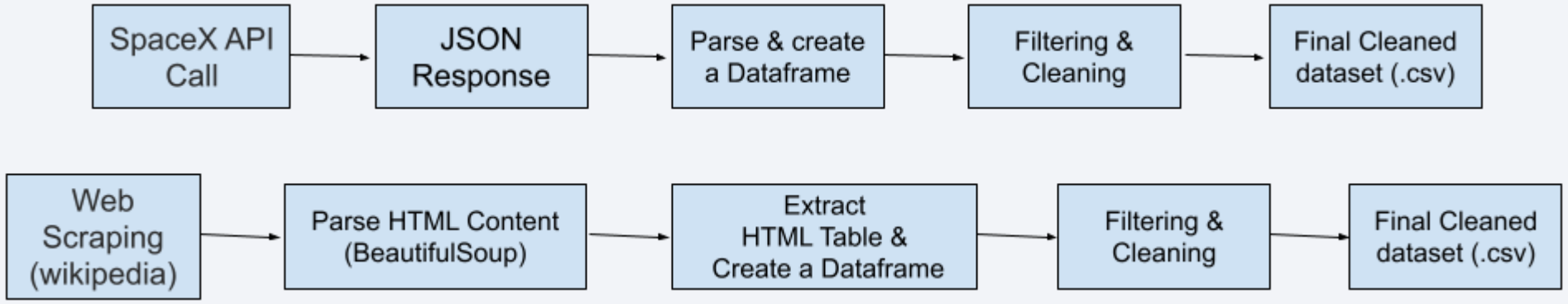
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
  - Replacing null values from payload mass with the mean value.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building models with scikit-learn, tune using GridSearchCV, and evaluate using test prediction accuracy.

Data Collection

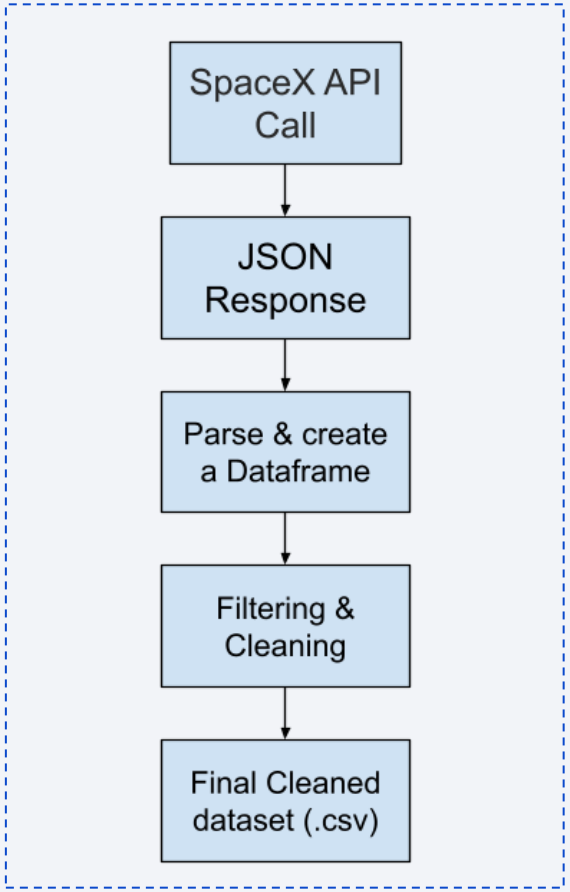
- Dual-Source Acquisition Strategy:
  - SpaceX API Calls (Endpoint): Programmatically collected Falcon 9 launch data as JSON objects, converted to Pandas DataFrames.
  - Web Scraping (Wikipedia): Extracted historical landing outcome specifics from structured HTML tables using BeautifulSoup.



Data Collection – SpaceX API

- Using SpaceX API (Endpoint): Programmatically collected Falcon 9 launch data as JSON objects, converted to Pandas DataFrame and convert it into .csv file with different predictors columns.

- GitHub link for "SpaceX data collection API"

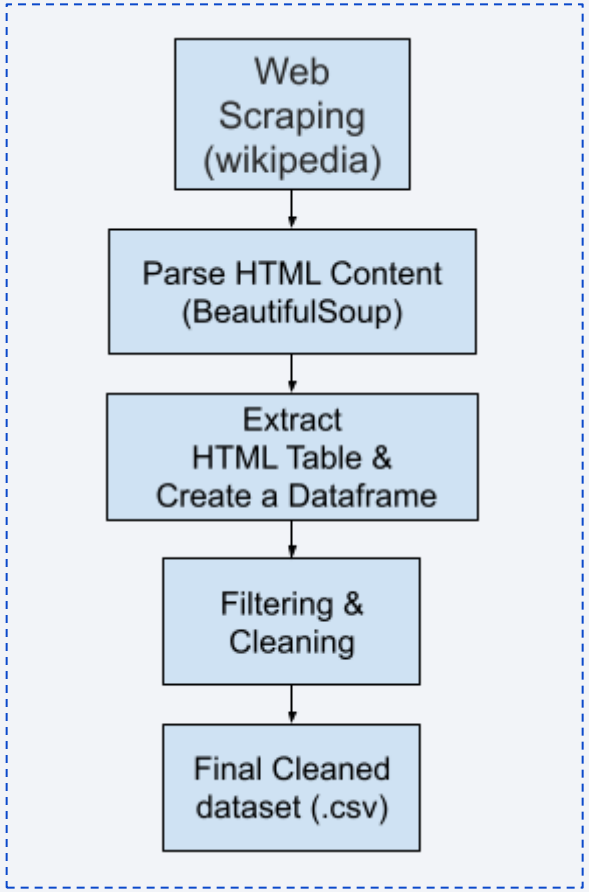


performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches

Data Collection - Scraping

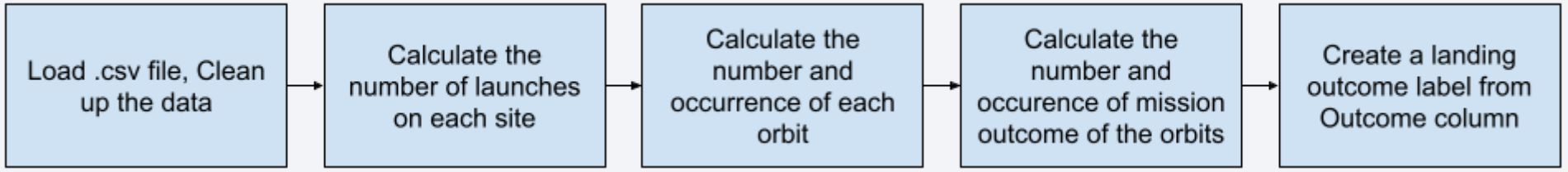
- Used Wikipedia for scraping the HTML Table using BeautifulSoup, then converted it into a DataFrame and to a .csv file with different predictors columns.

- GitHub link for “Web Scraping”



Data Wrangling

- Calculate the number of launches for each site.
- Calculate the number of occurrence of each orbit.
- Calculate the number and occurence of mission outcome of the orbits.
- Convert each labels into a class of good, or bad (1 or 0).
- GitHub link for “Data Wrangling”



EDA with Data Visualization

- **Scatter Plot: Flight Number vs Payload Mass (with Class Hue)**
  - To examine the relationship between flight experience (FlightNumber), payload mass, and launch success.
- **Scatter Plot: Flight Number vs Launch Site (with Class Hue)**
  - To analyze how different launch sites perform across flight attempts and identify site-specific success patterns.
- **Scatter Plot: Payload Mass vs Launch Site (with Class Hue)**
  - To understand payload capacity limitations and success rates at different launch sites
- **Bar Chart: Success Rate by Orbit Type**
  - To examine if mission experience affects success in different orbital missions
- **Scatter Plot: Payload Mass vs Orbit Type (with Class Hue)**
  - To analyze how payload mass affects success across different orbital destinations.
- **Line Chart: Launch Success Rate by Year**
  - To visualize temporal trends in SpaceX's landing success rate
- GitHub Link for “EDA with Visualization”

EDA with SQL

- 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 succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List all the booster_versions that have carried the maximum payload mass. Use a subquery.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.
- GitHub URL of your completed EDA with SQL notebook

Build an Interactive Map with Folium

- **Circles** - To clearly identify and highlight the geographic locations of all SpaceX launch sites for spatial analysis.
- **Text Labels** - To provide immediate identification of launch site names without requiring user interaction.
- **MarkerCluster** - To prevent map clutter when multiple launches occur at the same location while maintaining data visibility.
- **Color-Coded Markers** - To instantly visualize launch success patterns and identify which sites have better performance rates.
- **Distance Markers** - To quantify proximity relationships and analyze how infrastructure affects launch site selection.
- **PolyLines** - To visually connect and measure distances between launch sites and critical infrastructure for geographic pattern analysis.

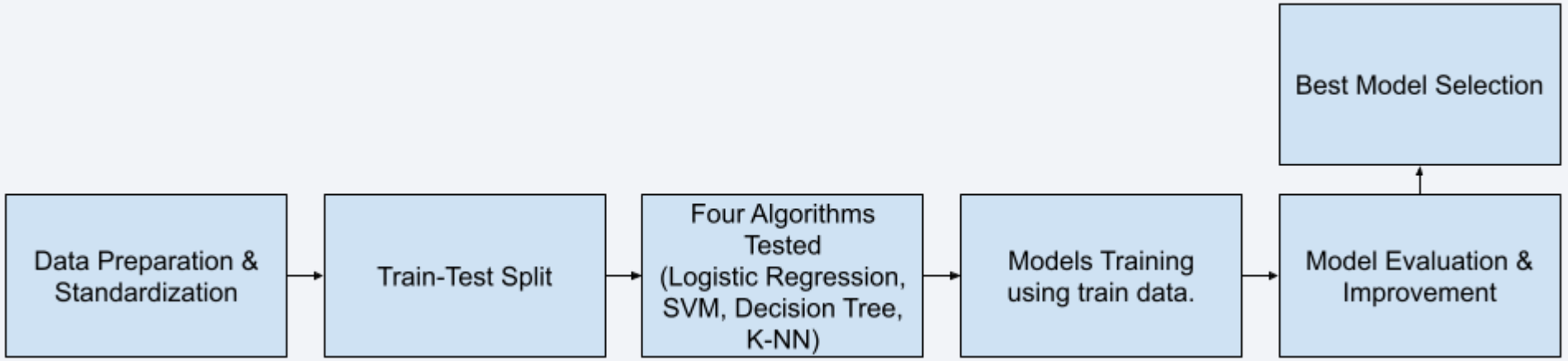
- GitHub URL of your completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- **Pie Chart** - Quick visual comparison of success rates across sites or success/failure ratio for individual sites.
- **Scatter Plot** - Identifies relationships between payload weight and mission success, helps spot patterns by booster type.
- **Dropdown** - Enables site-specific analysis vs. overall comparison.
- **Range Slider** - Allows filtering out outliers or focusing on specific payload categories for clearer insights.
- GitHub URL of your completed Plotly Dash.

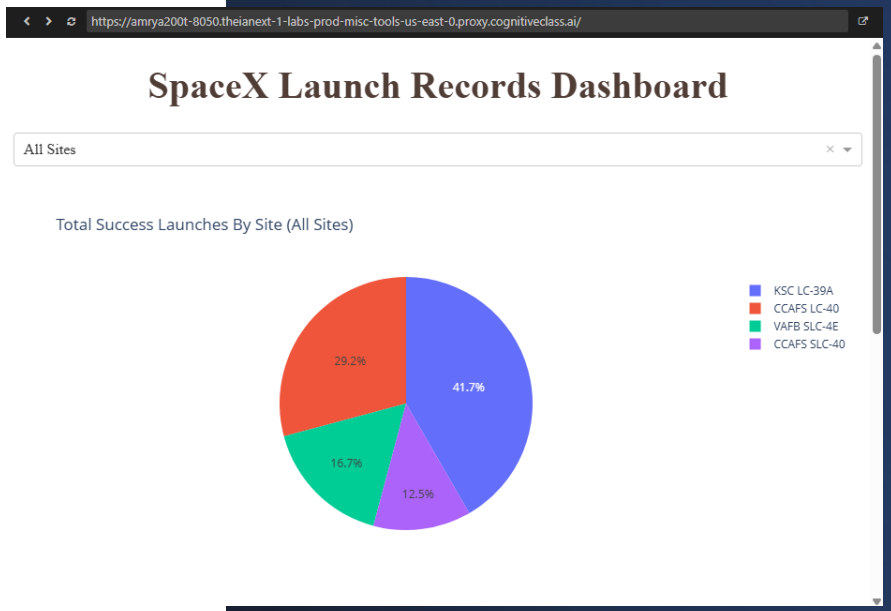
Predictive Analysis (Classification)

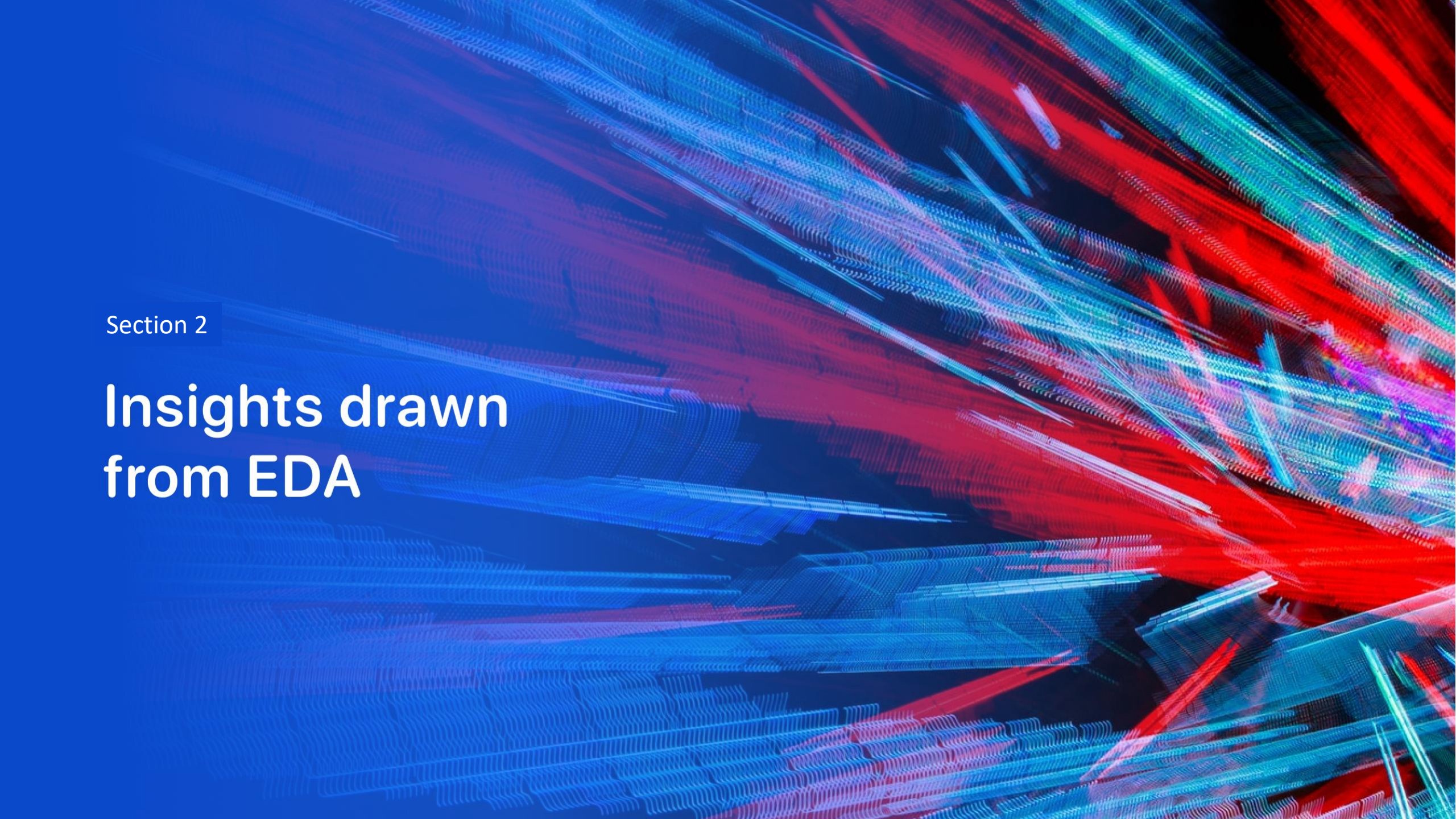
- Used 4 different ML models from Scikit-Learn to choose the best model for the data using for binary classification.
- Used GridSearchCV to find the best parameters for the model.
- GitHub URL of your completed predictive analysis lab



Results

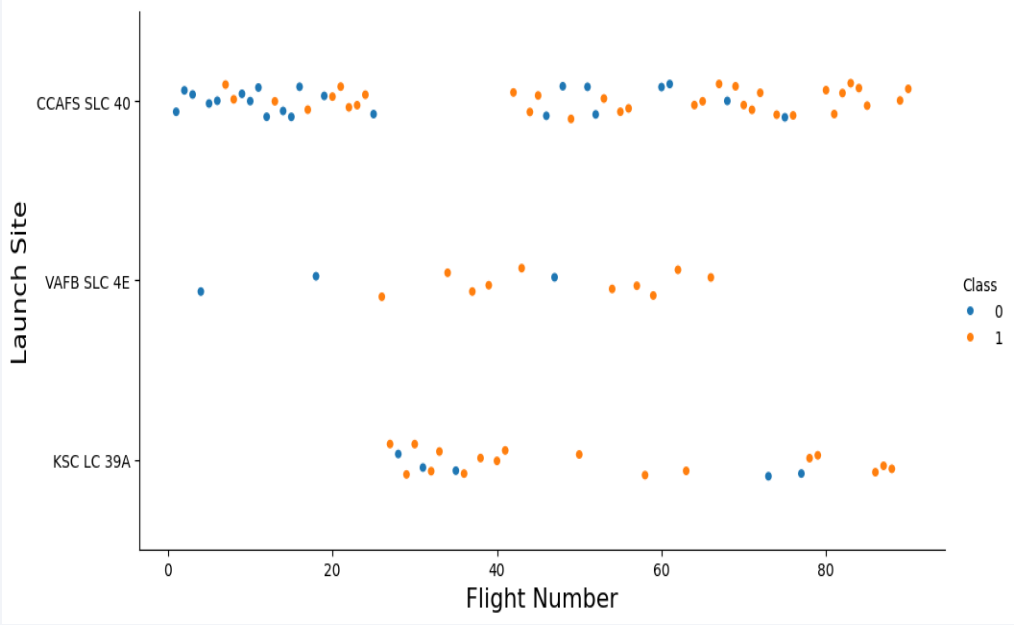
- The success rate of launches and landing improve overtime.
- The Decision Tree Model had the best classification rate of about 87.5% compared to other models 83%.




Section 2
Insights drawn from EDA

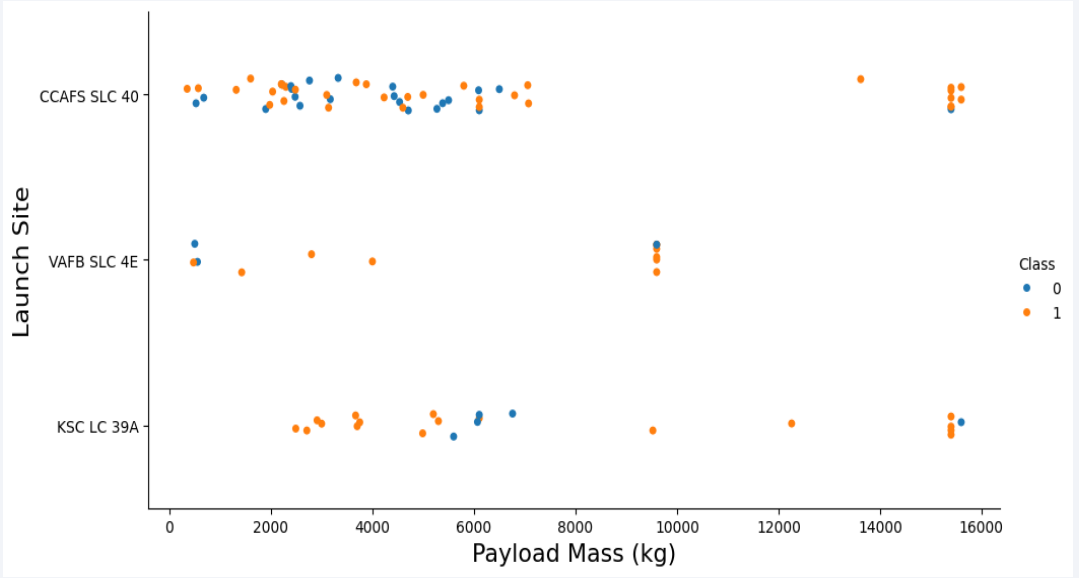
Flight Number vs. Launch Site

- **CCAFS SLC-40** dominates early flight numbers while **KSC LC-39A** is used more in later flights, showing SpaceX's facility evolution.
- Success rate increases with higher flight numbers across all sites, with more green dots (successful landings) appearing in later launches.



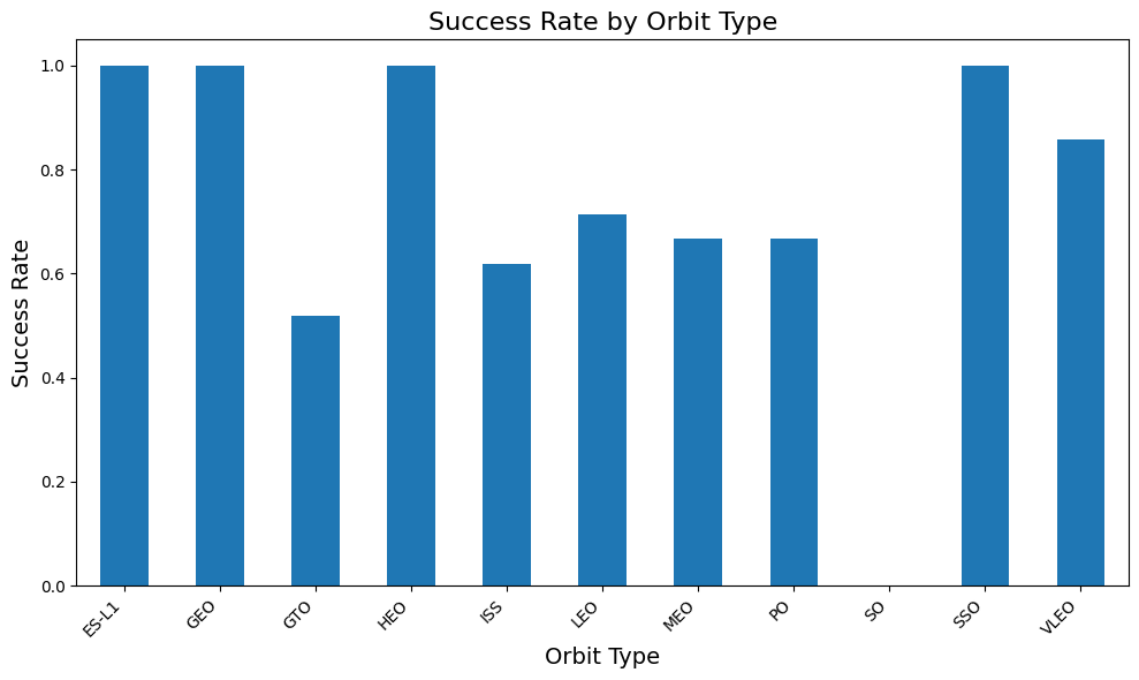
Payload vs. Launch Site

- **VAFB SLC-4E** has no launches for heavy payloads (>10,000 kg), indicating site-specific payload limitations.
- **CCAFS SLC-40** and **KSC LC-39A** handle the full range of payload masses with mixed success rates across different weights.



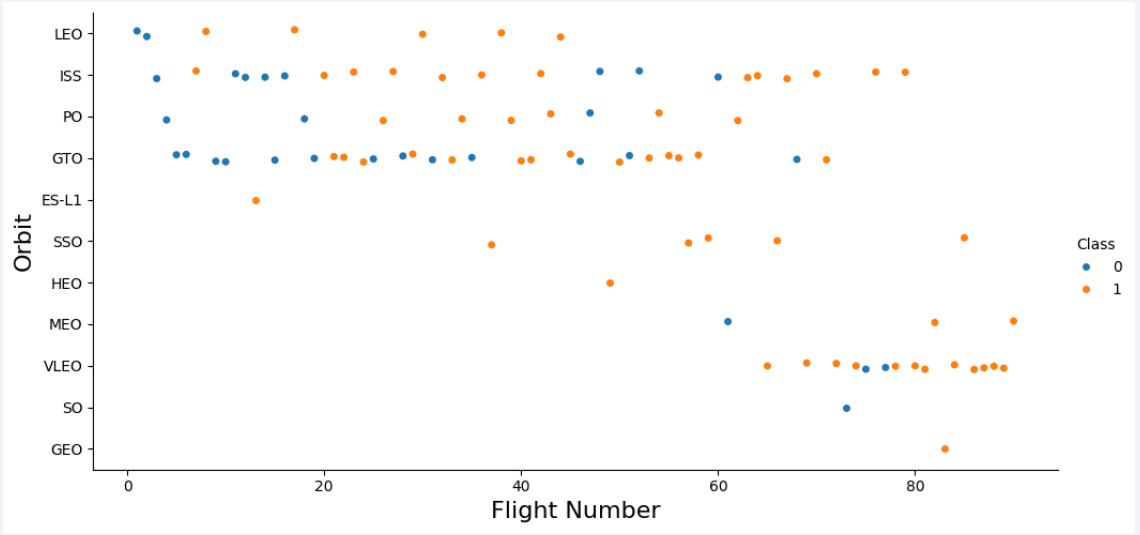
Success Rate vs. Orbit Type

- **ES-L1**, **GEO**, **HEO**, and **SSO** orbits show **100%** success rates, indicating optimal conditions for first stage recovery.
- **GTO** orbit has the lowest success rate, suggesting heavier payloads or higher energy requirements make landing more challenging.



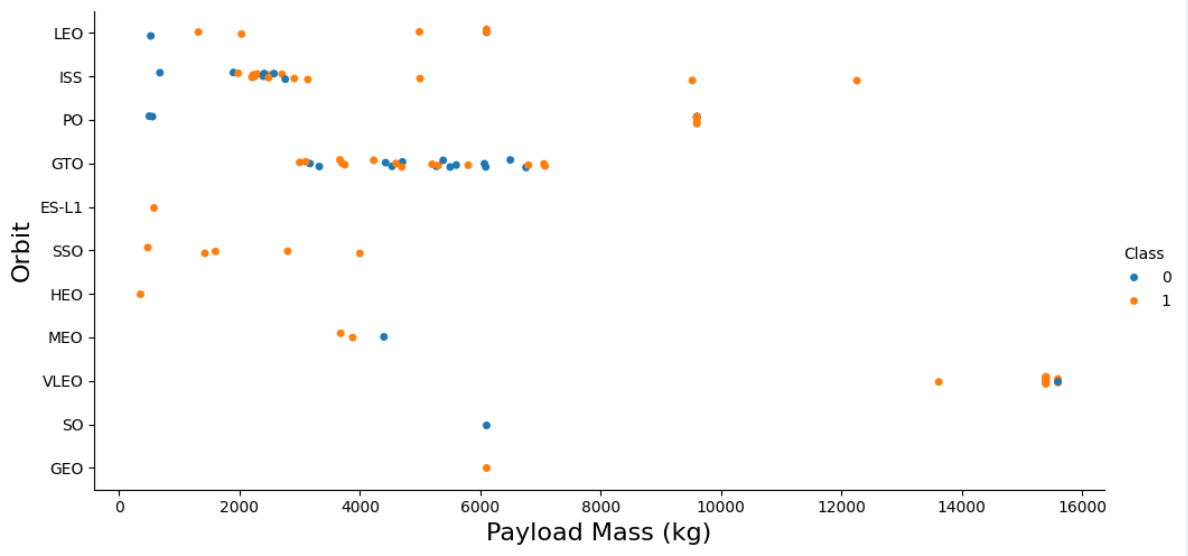
Flight Number vs. Orbit Type

- **LEO** orbit shows success rate improvement with higher flight numbers, indicating SpaceX's learning curve for low Earth orbit missions.
- **GTO** orbit displays no clear relationship between flight number and success, suggesting consistent difficulty regardless of experience level.



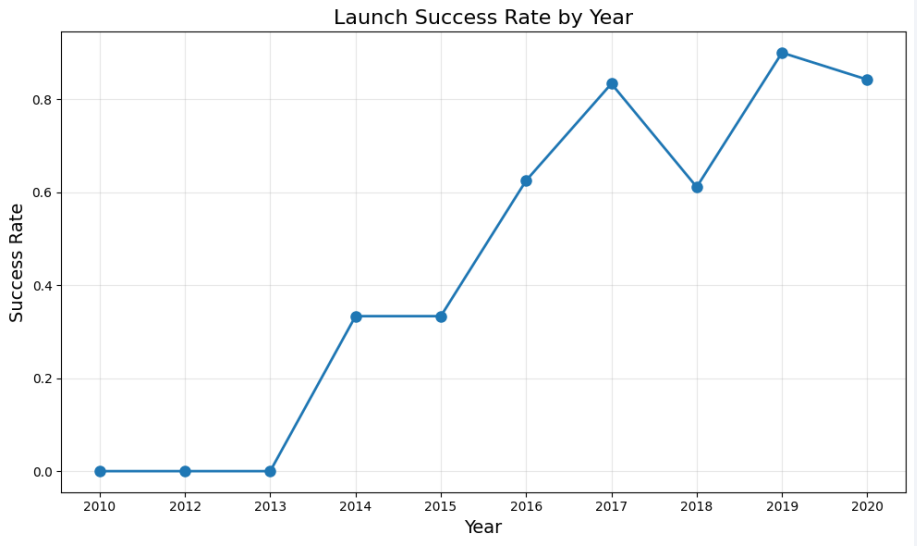
Payload vs. Orbit Type

- **Heavy** payloads show higher success rates for **PO**, **LEO**, and **ISS** orbits, indicating better landing conditions for these mission types.
- **GTO** orbit displays mixed results with both successful and failed landings across different payload masses, making success prediction difficult.



Launch Success Yearly Trend

- Success rate consistently increased from 2013-2017 (with stability in 2014), showing SpaceX's continuous improvement over time.
- Post-2015 acceleration in success rate demonstrates SpaceX mastering first stage landing technology after initial learning period.



All Launch Site Names

- Find the names of the unique launch sites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
%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

Launch Site Names Begin with 'CCA'

- Display 5 records where launch sites begin with the string 'CCA'

```
%sql Select * from SPACEXTABLE Where Launch_Site Like 'CCA%' Limit 5
```

Python

```
 * 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

Total Payload Mass

- 45,596 kg - Total payload mass carried by boosters launched by NASA (CRS)

```
%sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE Where Customer like 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- 2,534.67 kg - Average payload mass carried by F9 v1.1 booster versions.

```
%sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE Where Booster_Version like 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- 2015-12-22 - Date of first successful landing outcome.

```
%sql Select min(Date) from SPACEXTABLE where Landing_outcome like 'Success%'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four boosters that successfully landed on drone ships.
  - F9 FT B1022 | F9 FT B1026
  - F9 FT B1021.2 | F9 FT B1031.2

```
%sql Select Distinct Booster_Version from SPACEXTABLE Where Landing_Outcome like 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND
PAYLOAD_MASS__KG_ < 6000
```

Python

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

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

Total Number of Successful and Failure Mission Outcomes

- 100 Successful mission outcomes.
- 1 Failure mission outcomes.

```
%sql Select Count(Mission_Outcome) from SPACEXTABLE where Mission_Outcome like 'Success%'
```

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

Count(Mission_Outcome)
100

```
%sql Select Count(Mission_Outcome) from SPACEXTABLE where Mission_Outcome like 'Failure%'
```

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

Count(Mission_Outcome)
1

Boosters Carried Maximum Payload

- 12 Falcon 9 Block 5 boosters carried the maximum payload mass:

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

```
%sql Select Distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ == (Select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

Python

```
 * 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

2015 Launch Records

- 2 Failed drone ship landings in 2015:
  - **January (01)**: F9 v1.1 B1012 at CCAFS LC-40
  - **April (04)**: F9 v1.1 B1015 at CCAFS LC-40

```
%sql Select substr(Date, 6, 2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5) = '2015'
```

Python

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

substr(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- **No attempt**: **9** - Most common outcome (early missions)
- **Failure (drone ship)**: **5** - Failed sea landings during learning phase
- **Success (drone ship)**: **4** - Successful autonomous ship landings
- **Controlled (ocean)**: **3** - Intentional ocean landings
- **Success (ground pad)**: **2** - Successful land-based recoveries
- **Uncontrolled (ocean)**: **2** - Unplanned ocean impacts
- **Precluded (drone ship)**: **1** - Landing attempt prevented

Landing_Outcome	Count(Landing_Outcome)
Controlled (ocean)	3
Failure (drone ship)	5
No attempt	9
Precluded (drone ship)	1
Success (drone ship)	4
Success (ground pad)	2
Uncontrolled (ocean)	2

```
%sql Select Landing_Outcome, Count(Landing_Outcome) from SPACEXTABLE where substr(Date,0,5) > '2010' and substr(Date, 0, 5) < '2017' group by
Landing_Outcome
```

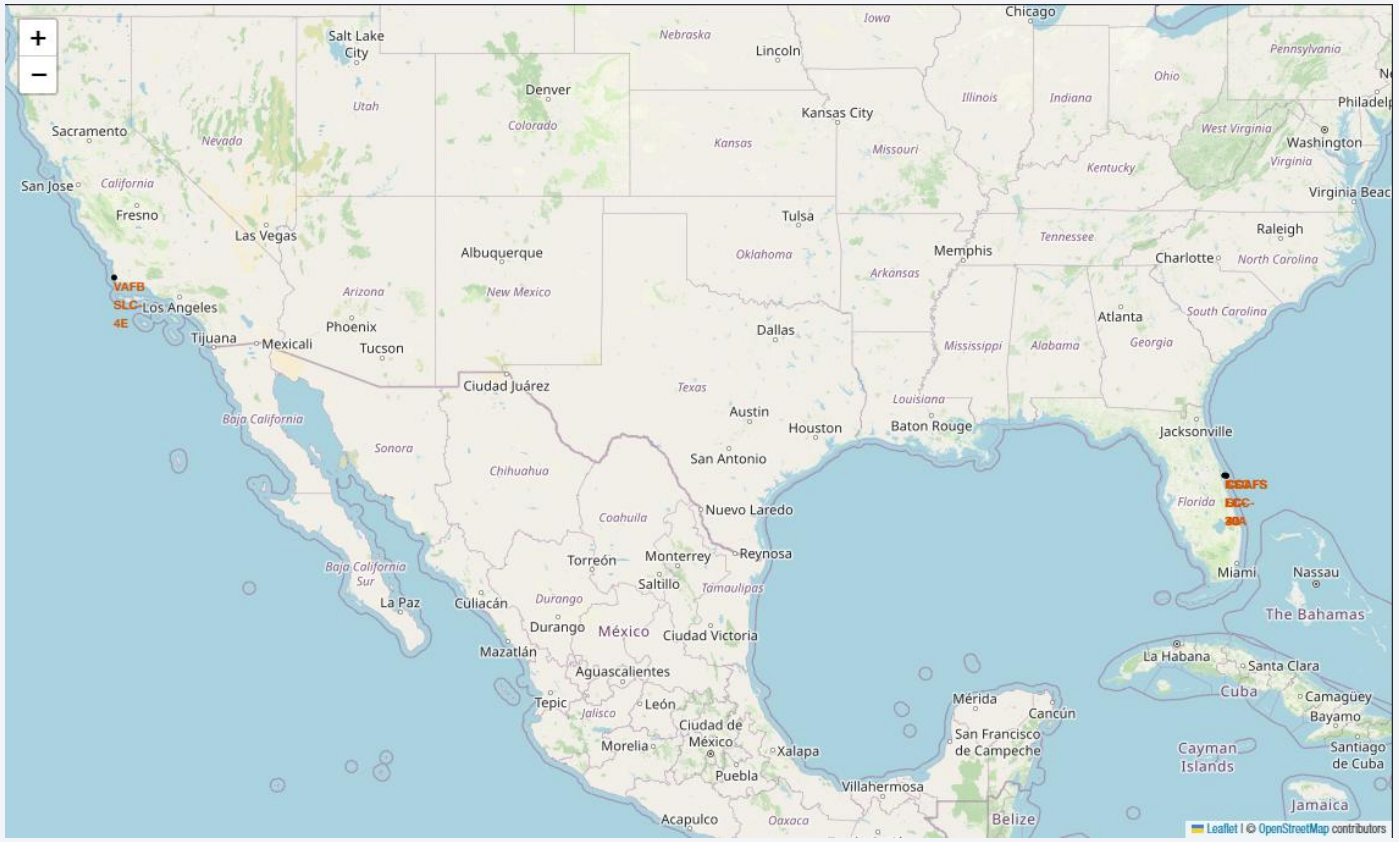
Python

Section 3

Launch Sites Proximities Analysis

Launch site locations

- Launch site locations are focused between Florida and California.

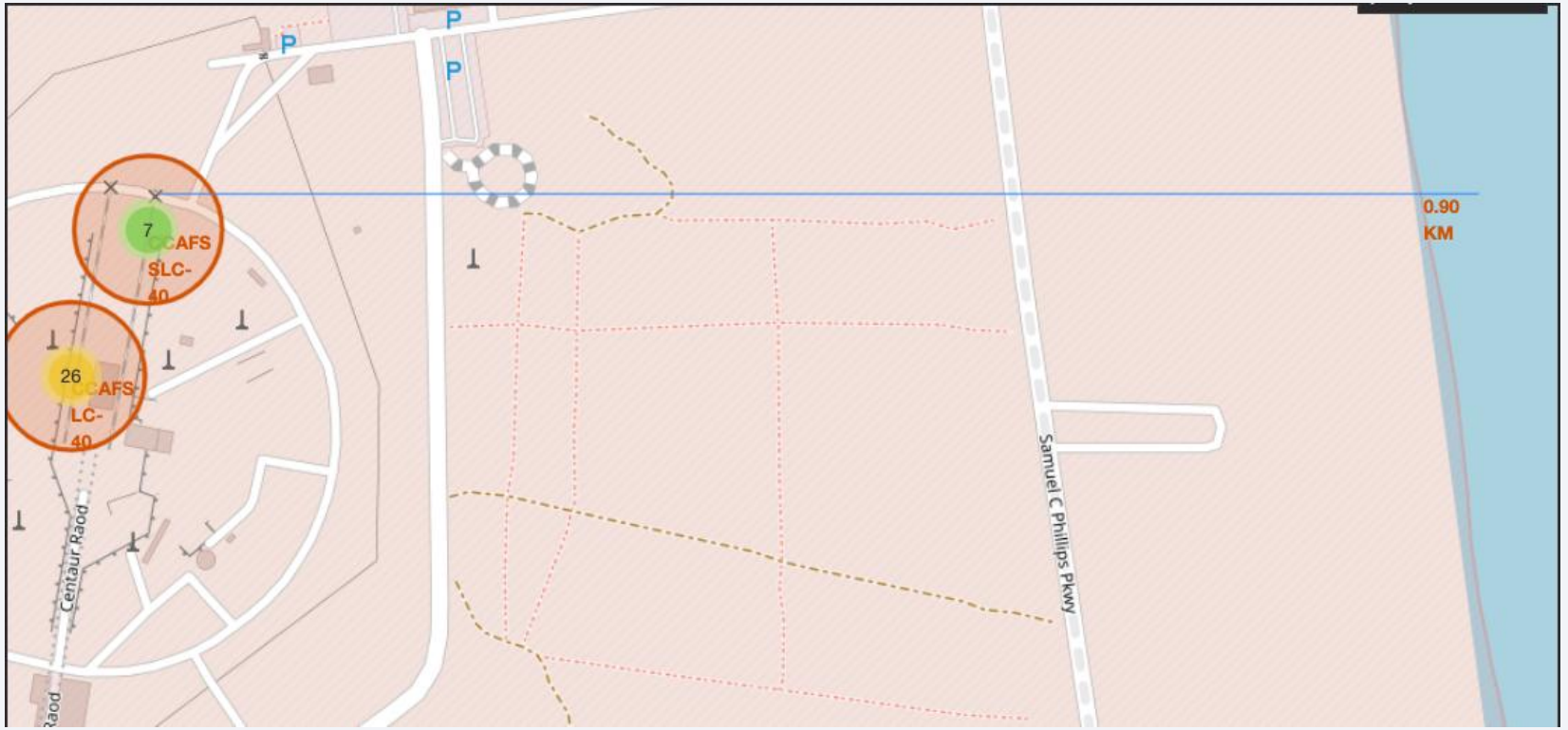


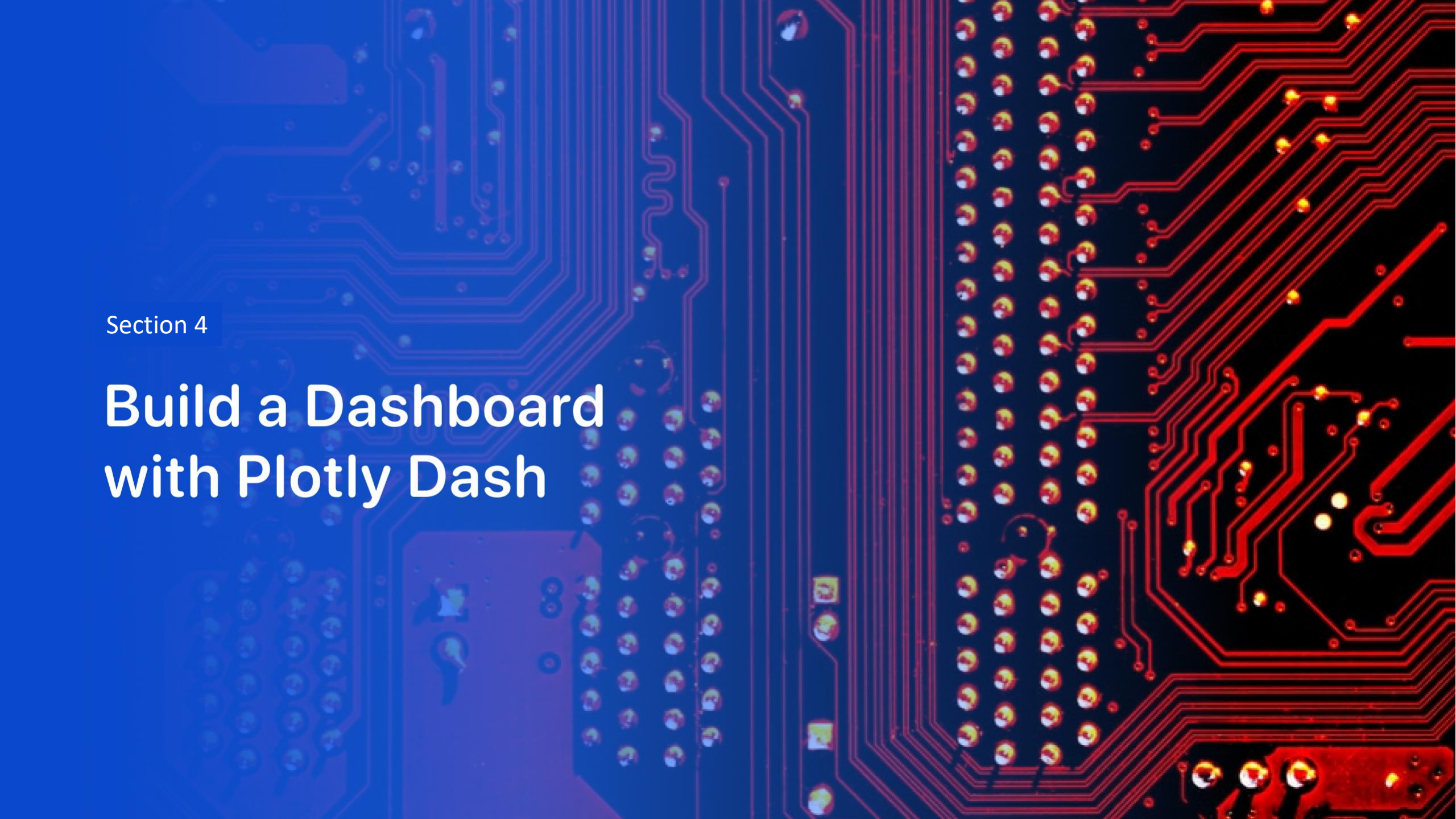
Launches for Each Location.

- The majority of launches originated from Florida (46 launches), significantly outnumbering those from California (10 launches).

Distance between CCAFS SLC-40 and coastline

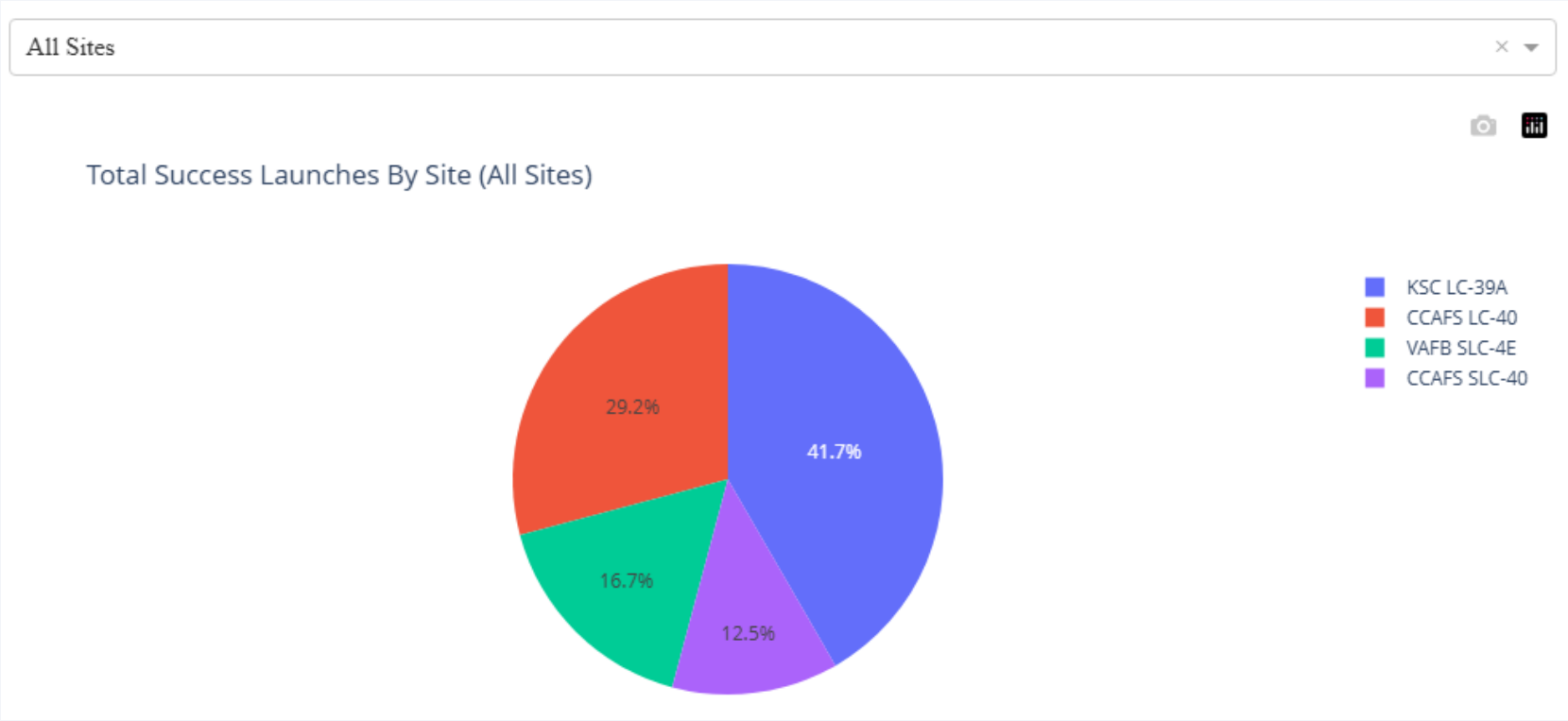
- The CCAFS SLC-40 launch site is located approximately 0.9 km from the coastline.




Section 4
Build a Dashboard with Plotly Dash

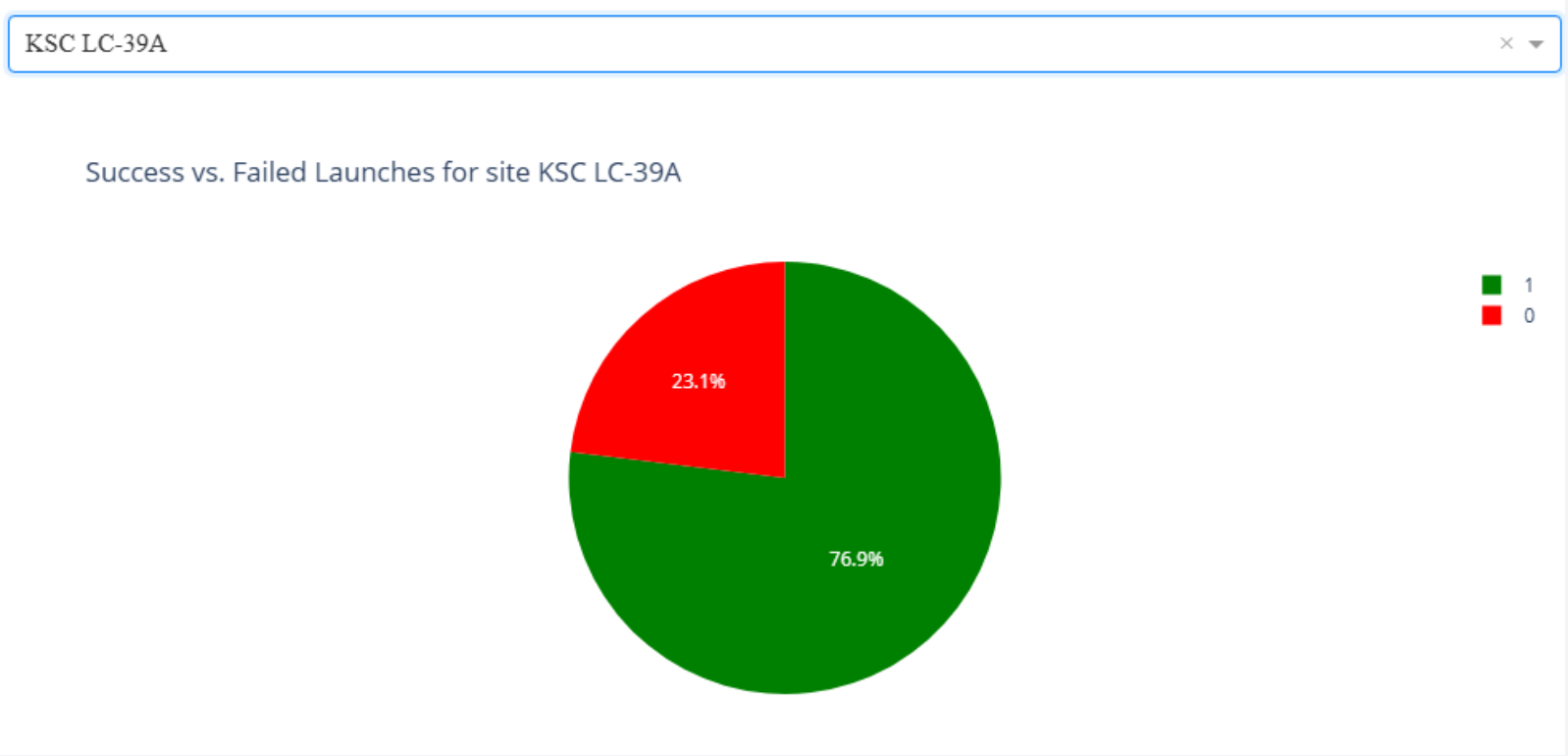
Total Success Launches By Site (All Sites)

- KSC LC39A has the most successful launches rate 41.7% while CCAFS SLC-40 has the least successful rate 12.5%.



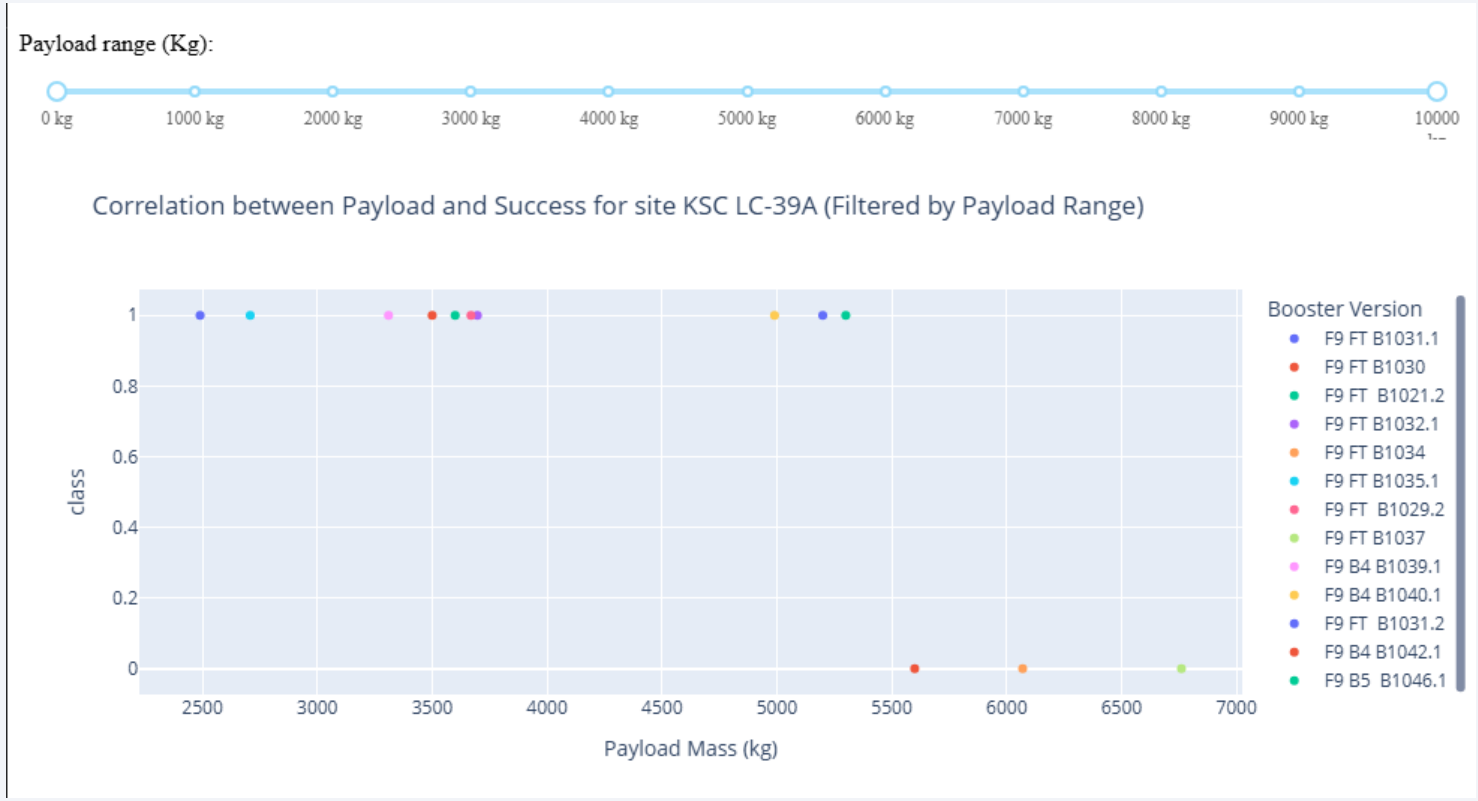
Success vs. Failed Launches for site KSC LC-39A

- The analyzed data indicates a 76.9% success rate for launches, with 23.1% resulting in failure.



Payload Mass and booster success rate

- Launches with payload masses under 5500 kg consistently resulted in 100% success.

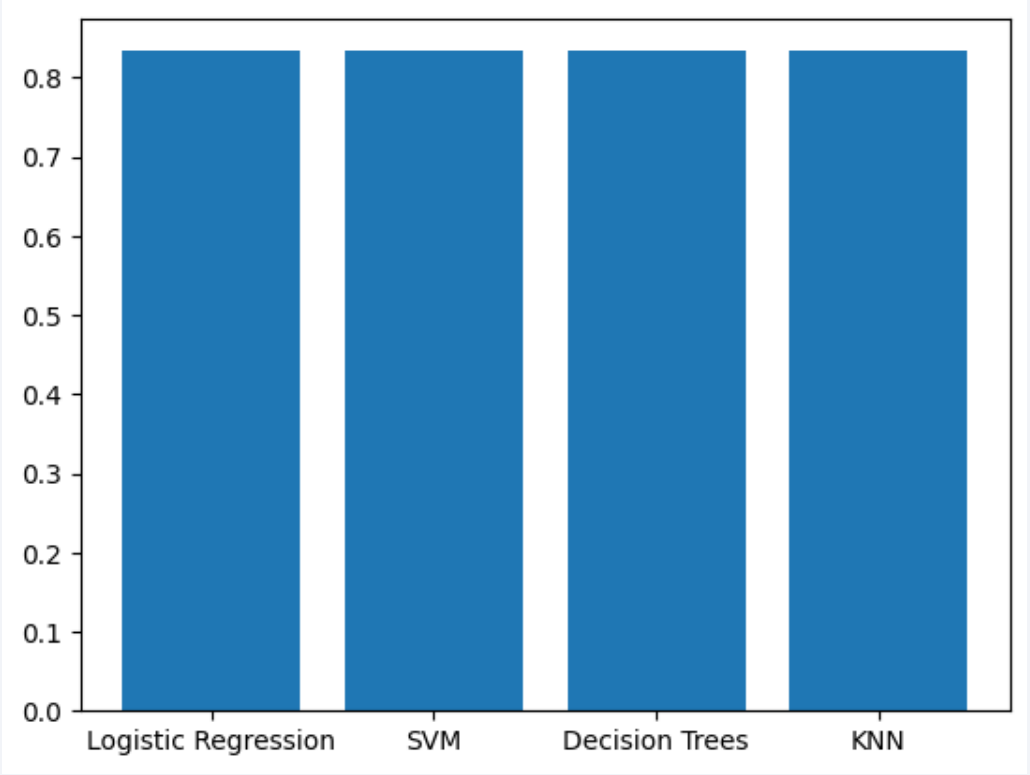


Section 5

Predictive Analysis (Classification)

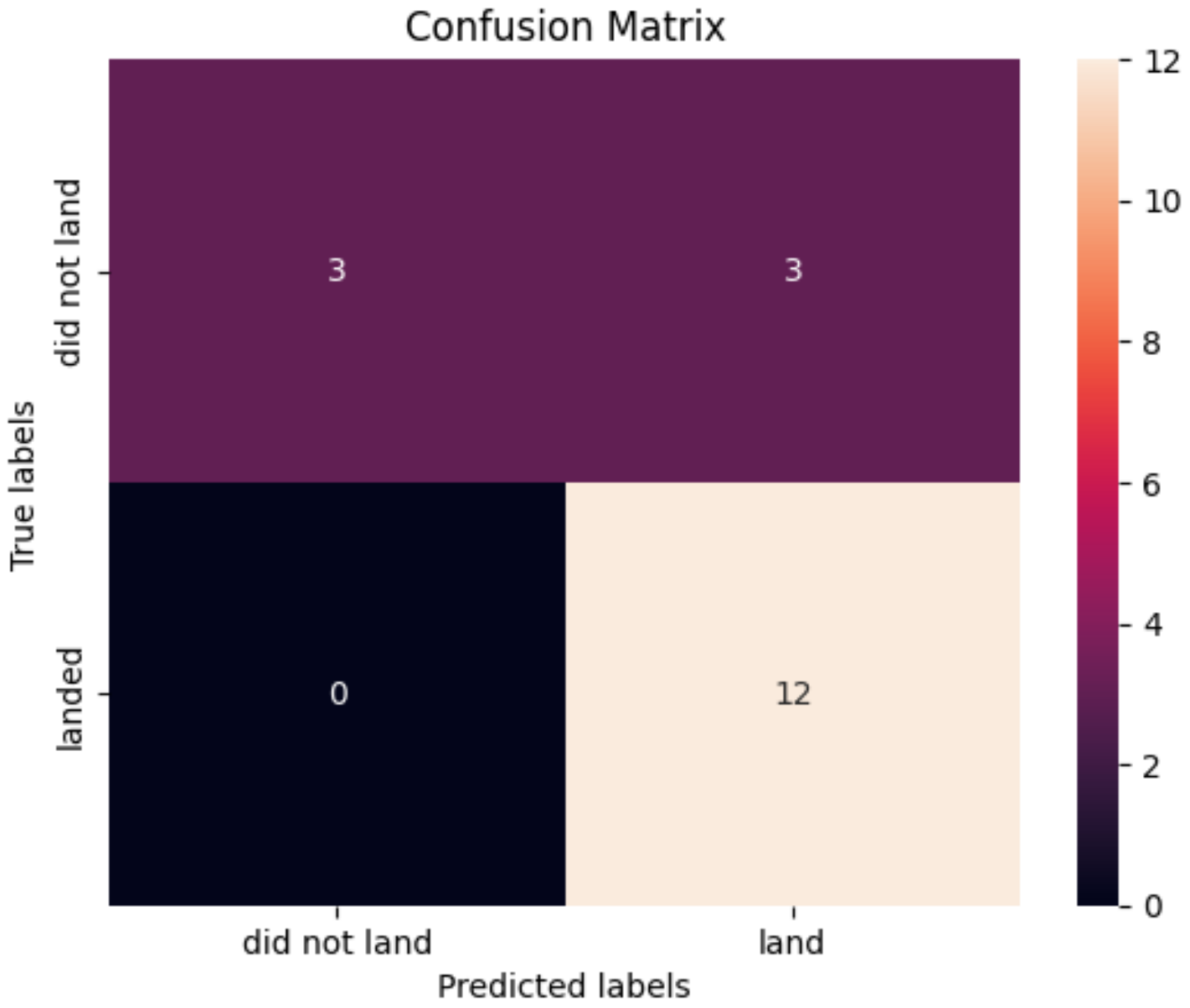
Classification Accuracy

- The Decision Tree model demonstrated the highest classification accuracy, reaching about 87.5%. While the rest got about 83%.
- Notably, the performance of all considered models was very similar, with only slight variations in their classification accuracies.



We can see that it is correctly identifies all successful landings, while incorrectly predicted 3 failed landings as successful.

Confusion Matrix



Conclusions

- **Key Predictors Identified**: Launch number, launch site, payload mass, and orbit are confirmed as significant predictors for landing success.
- **Optimal Model Performance**: A Decision Tree Classifier demonstrates superior performance, achieving approximately 87.5% accuracy in predicting landing outcomes.
- **Production Readiness**: Given its high accuracy, the Decision Tree Classifier is a suitable candidate for implementation in production settings, pending stakeholder validation.

Appendix

- https://github.com/amrya200t/Applied-Data-Science-Capstone/blob/main/Module%203/spacex-dash-app.py

Thank you!