IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Edward Baldwin IV
2024-04-22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL, Pandas, and Pyplot
  - Interactive Analytics with Folium
  - Predictive Analytics
- Summary of all results
  - Predictive models with tuned hyperparameters (Logistic Regression, SVM, Decision Tree, and KNN)
  - Best predictive model

Introduction

- Space flight has become less expensive with the advent of reusable first stages on rockets such as SpaceX's Falcon 9. On average, the Falcon 9's reusable first stage has reduced the cost of a launch from $162 million to just $62 million. Space Y would like to compete with SpaceX.
- To compete with SpaceX, we will use historical data to determine the cost of a launch based on whether SpaceX will land its first, reusable stage on a Falcon 9.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX's REST API
  - Web Scraping SpaceX's tabular data publicly available on Wiki
- Perform data wrangling
  - Initial data was collected as JSON and HTML objects and then processed as Pandas Data Frames for analysis and further visualization.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple classification models with optimal hyperparameters compared for accuracy on test data

Data Collection

Method 1

SpaceX REST API → GET requests → JSON objects of historical SpaceX launch data → JSON -> Pandas Data Frame

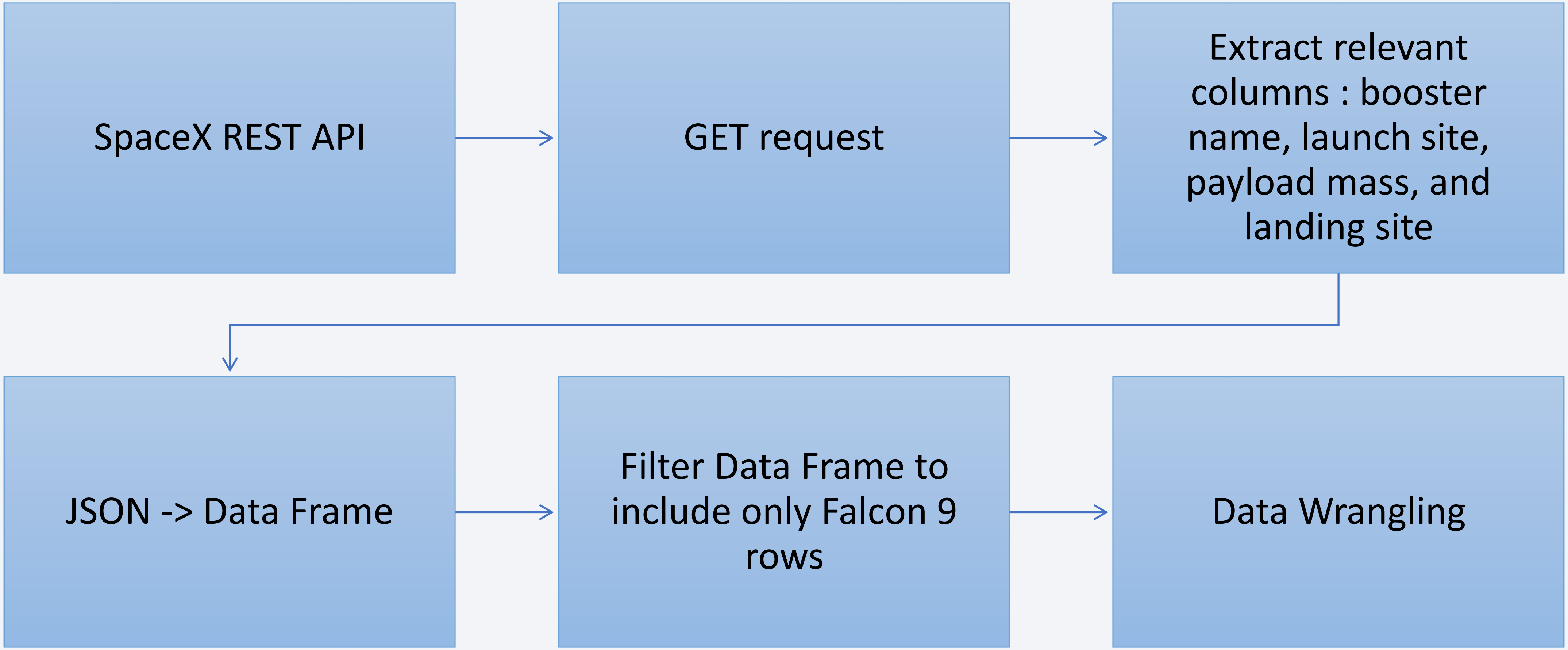
Method 2

Web scraping SpaceX launch records → BeautifulSoup objects to parse and process HTML → Parsed HTML -> Pandas Data Frame

Data Collection – SpaceX API

Data Collection Notebook

Collecting data from SpaceX REST API for further analysis



Data Collection - Scraping

[Web Scraping Notebook]

Collecting SpaceX launch data from publicly available web-hosted URL in the form of raw HTML

URL to web page hosting Falcon 9 Launch table → BeautifulSoup object to work with raw HTML from URL → HTML Parsing for relevant features and attributes → Parsed HTML - > Data Frame → Data Wrangling

Data Wrangling

Data Wrangling Notebook

Exploratory Data Analysis to familiarize ourselves with the data, identifying patterns and determining relevant features

Identification and statistical processing of NULL values

Encoding of Categorical Features

Normalization of Data

Number of launches per launch site

Number of launches and mission outcome by orbit

Boolean classification of landing outcomes as Failure or Success

EDA with Data Visualization

Data Visualization Notebook

Summary of plotted charts:

- Catplot – Flight Number v. Payload
- Catplot – Flight Number v Launch Site
- Catplot – Payload v Launch Site
- Bar chart – Success rate by Orbit Category
- Catplot – Flight Number v Orbit Category
- Catplot – Payload v Orbit Category
- Line Chart – Yearly trends for successful landing outcomes

EDA with SQL

[SQL Notebook]

- **Display the names of unique launch sites**

```
SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

- **Display the total payload mass carried by boosters launched by NASA (CRS)**

```
SELECT SUM(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

- **Display average payload mass carried by booster version F9 v1.1**

```
SELECT AVG(PAYLOAD_MASS__KG_) as AVG_PAYLOAD_MASS_F9 FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

- **List the date when the first successful landing outcome in ground pad was achieved**

```
SELECT MIN(Date) as FIRST_SUCCESSFUL_GROUND_PAD FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

EDA with SQL

SQL Notebook

- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**

```
SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'S AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

- **List the total number of successful and failure mission outcomes**

```
SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE GROUP BY Landing_Outcome
```

- **List the names of the booster_versions which have carried the maximum payload mass.**

```
SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

- **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.**

```
SELECT 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 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.**

```
SELECT Landing_Outcome, Date FROM SPACEXTABLE WHERE substr(Date,0,11) BETWEEN '2010-06-04' and '2017-03-20'
```

Build an Interactive Map with Folium

Geospatial Analysis Notebook

- folium.Circle and folium.Marker
  - These objects are used to mark and encircle a given coordinate with a text label on the sitemap
- MarkerCluster
  - Since our markers are labelling multiple launches in a small area, a MarkerCluster helps us to organize these markers into a group rather than multiple, overlapping markers. The individual markers can be viewed by zooming in and left-clicking on the cluster.
- MousePosition
  - MousePosition is used to display the latitude and longitude for the current position of the mouse on the site map
- folium.PolyLine
  - The Polyline is used to draw labeled lines between relevant sites and markers on the site map. In this instance, we use polylines to show the distance between a launch, the nearest coastline, nearby cities, railways, and highways.

Build a Dashboard with Plotly Dash

Interactive Dashboard

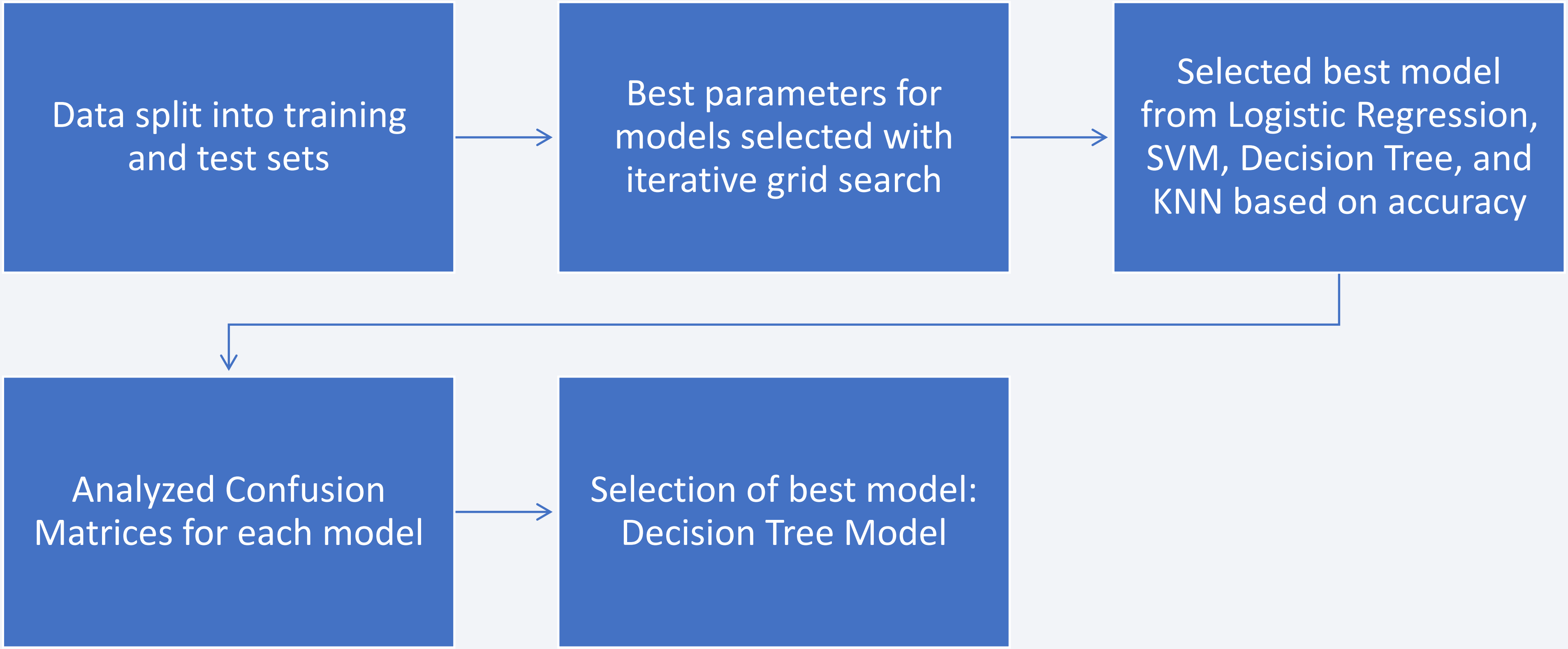
Summary:

- A dropdown menu permits the user to select a launch site, or all launch sites
- A pie chart
    - If all launch sites are selected from the drop down, the pie chart displays the ratio of all successful landing outcomes by launch site
    - If a specific launch site is selected, the pie chart displays the ratio of successful versus failed landings
- A sliding bar is used to select the range for the Payload Mass (kg)
- A scatter plot below the sliding bar displays landing outcome (0 for failed, 1 for success) with the x-axis showing the Payload Mass (kg) within the range specified on the sliding bar.
    - A legend on the side associates the booster version used in the launch via color. (You can also hover over the point on the plot to reveal more detailed information)
- If a specific launch site is chosen from the dropdown menu, the scatter plot will display the details for only that launch site

Predictive Analysis (Classification)

[Predictive Analysis Notebook]

Data was split into training and testing sets. Using the training set, we performed a grid search to determine the best hyperparameters for each predictive model used. After determining the best parameters, models were scored on their ability to accurately classify the test data.



Results

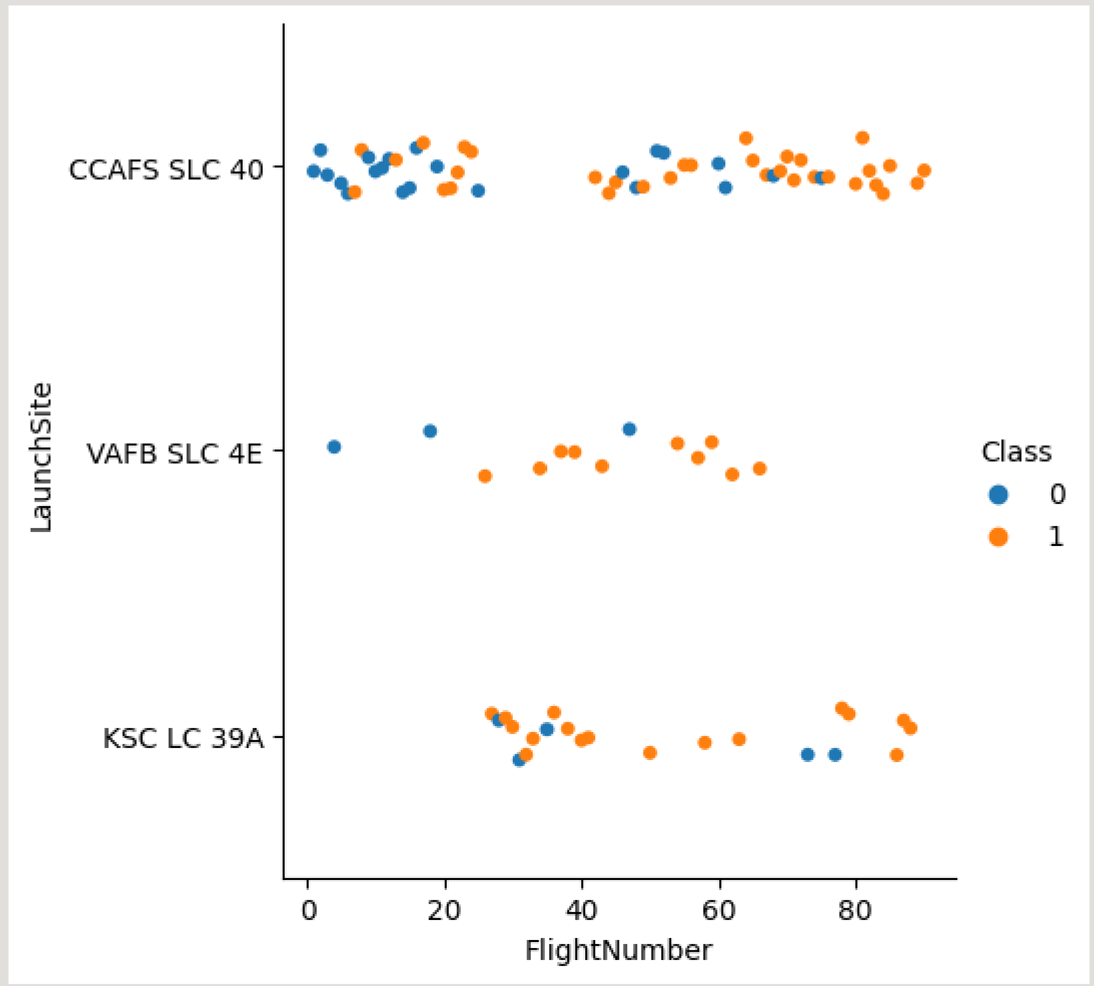
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

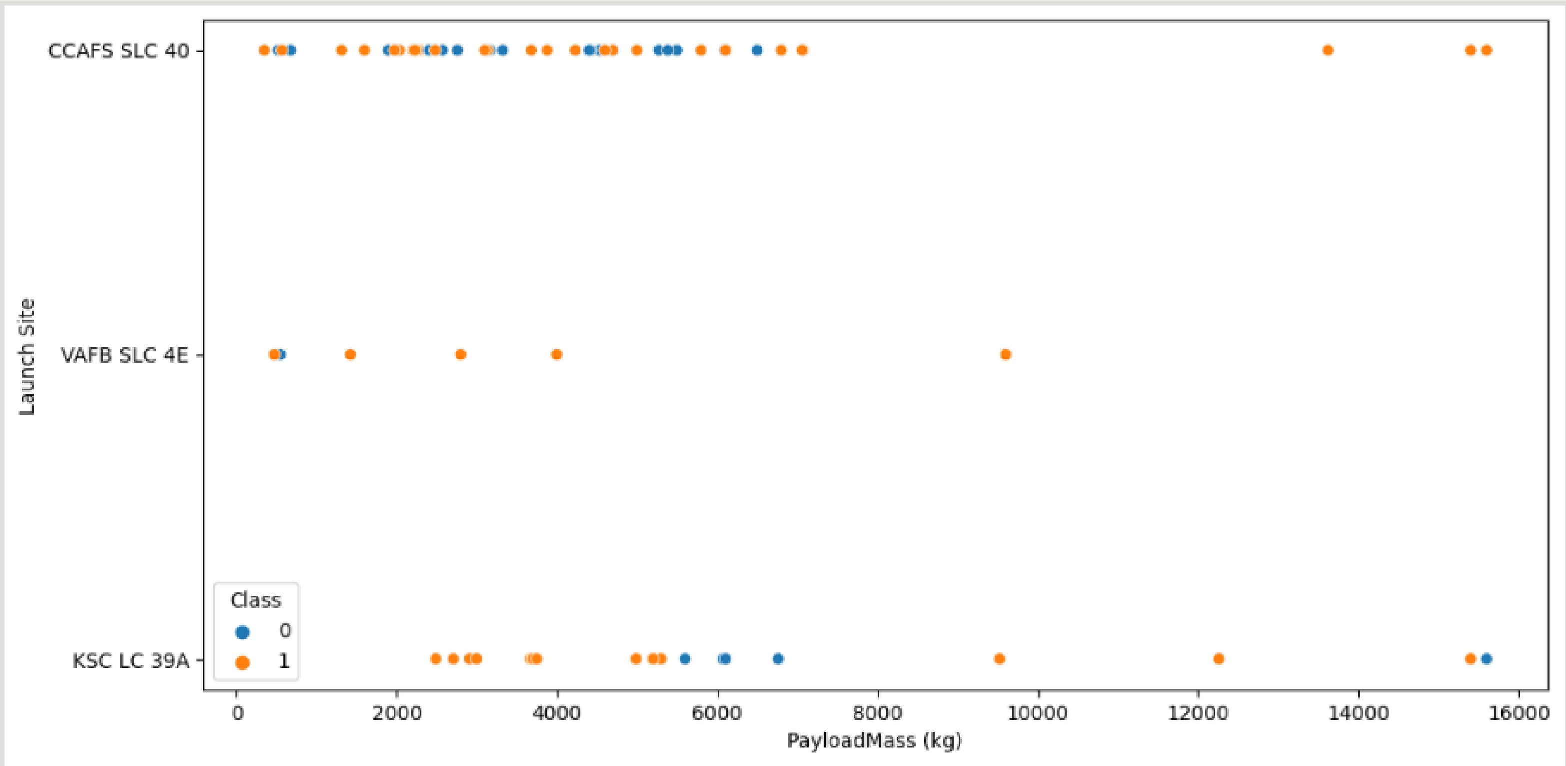
Flight Number vs. Launch Site

- Over time, the frequency of failed landings have decreased.



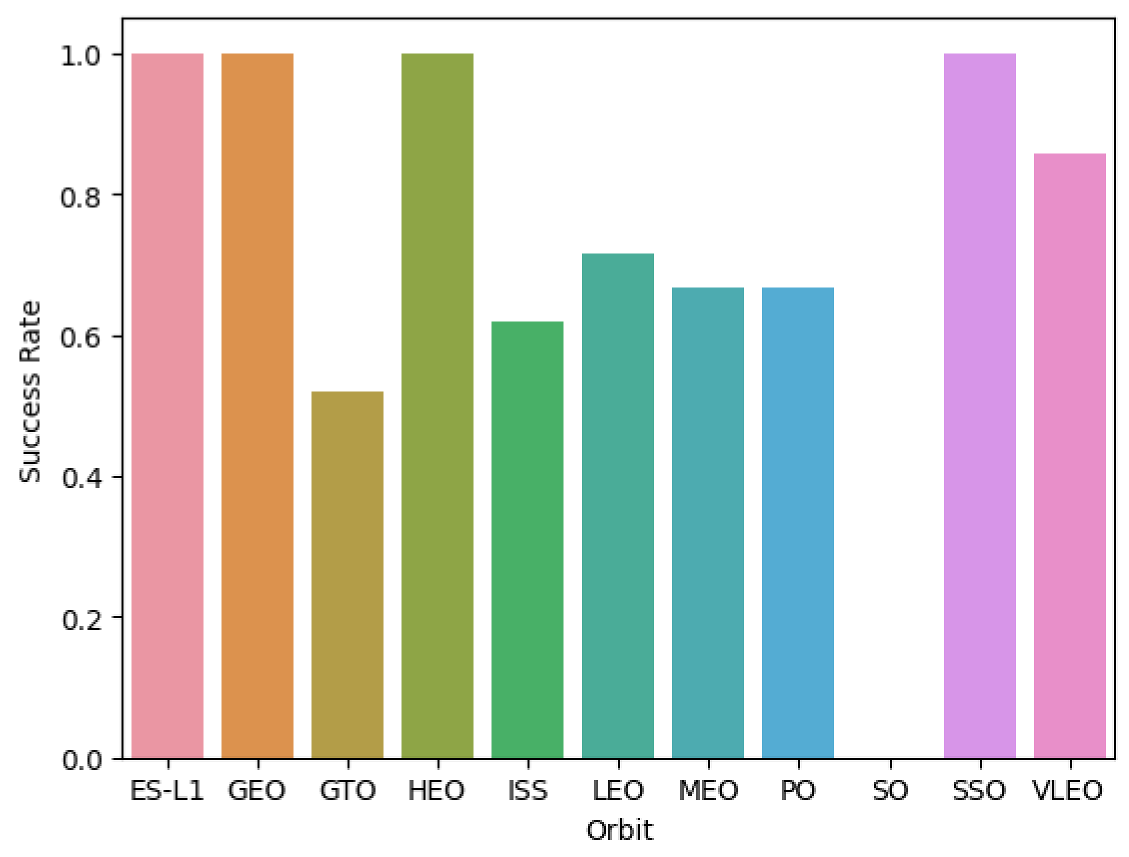
Payload vs. Launch Site

- For VAFB SLC4E, there are no payload masses greater than 10,000 kg
- For KSC LC 39A, though there are fewer launches, this site has performed well for payload masses less than ~5800 kg
- For CCAFS SLC40, though it is responsible for most launches, there is a noticeable gap between ~7500kg to ~13,500 kg.



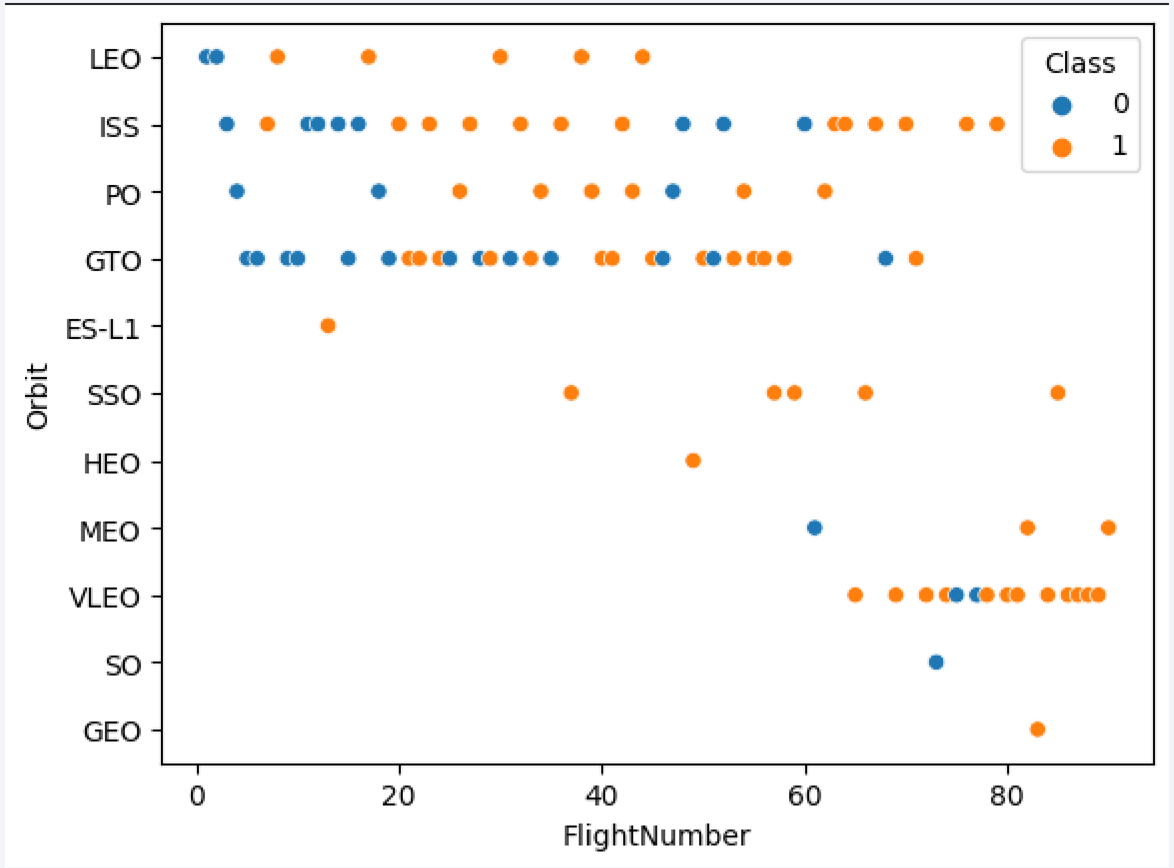
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO show a success rate of 100%
- To gain a complete picture of the success, or lack thereof, for certain sites, we need to account for the number of flights



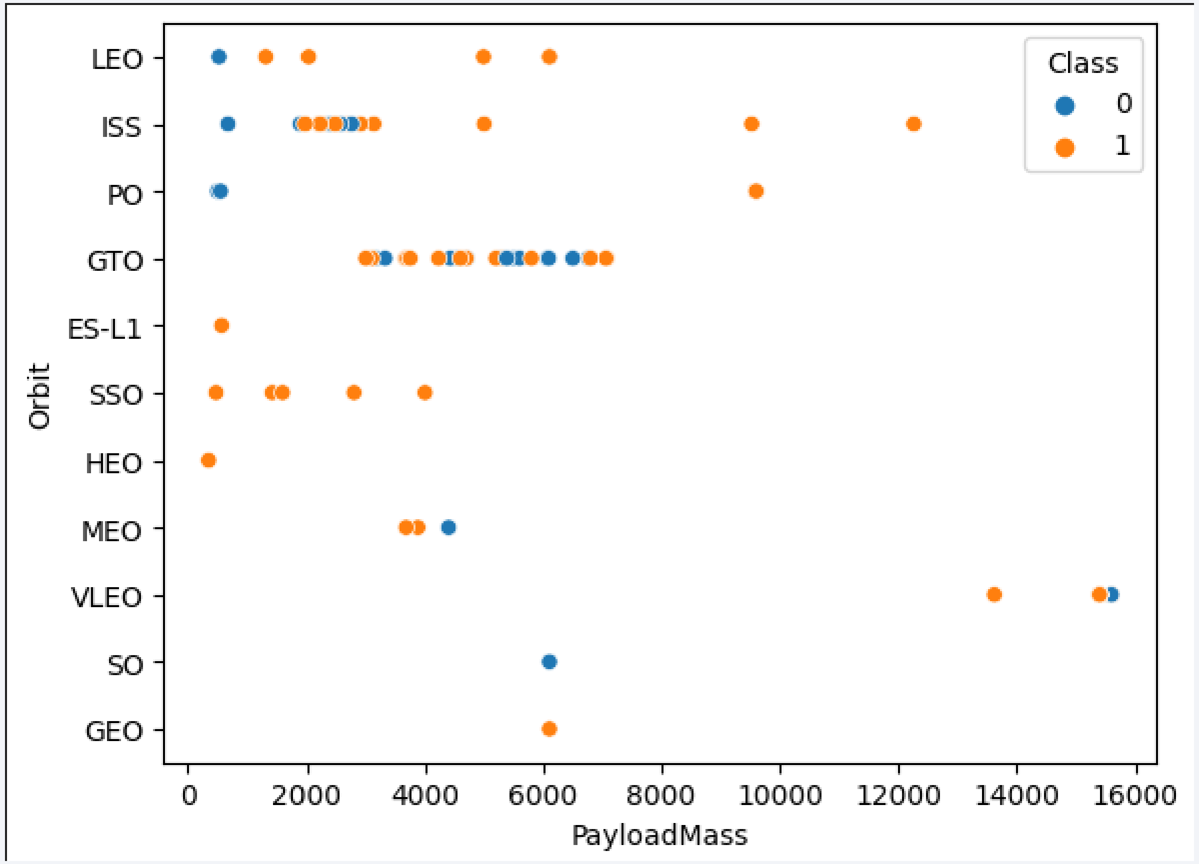
Flight Number vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO show a success rate of 100%
- ES-L1: 1 Flight
- SSO: 5 Flights
- HEO: 1 Flight
- GEO: 1 Flight
- LEO, ISS, PO, GTO, and VLEO are the target orbit for a majority of flights



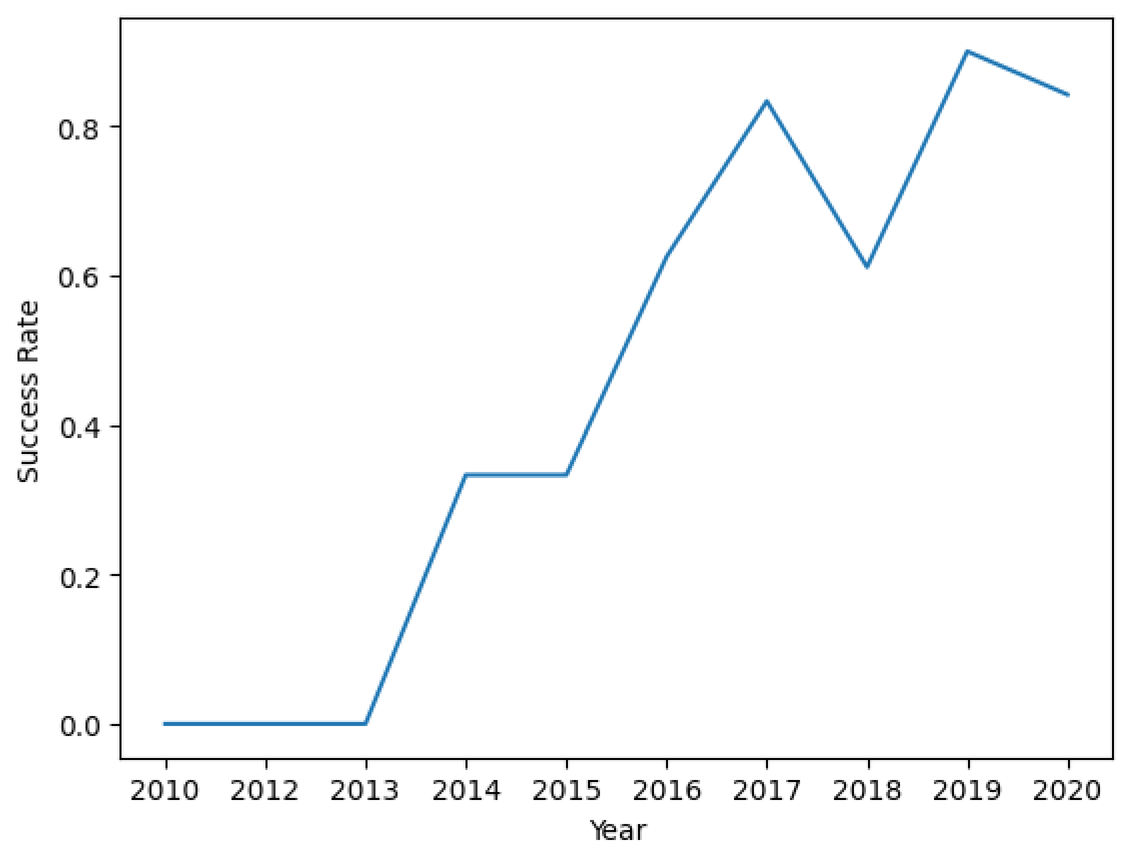
Payload vs. Orbit Type

- Most flights to *any* orbit hold a payload mass of less than 8000kg.



Launch Success Yearly Trend

- 2010-2013: 0% success rate
- 2014-2015: ~37% success rate
- 2015-2017: + ~20% per year
- 2017-2018: - ~20%
- 2018-2019: + ~20% per year



All Launch Site Names

Unique launch sites from SpaceX data

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

Launch Site Names Begin with 'CCA'

5 records of Launch Sites beginning with ‘CCA’

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Total payload mass (kg) carried in NASA (CRS) launches

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

Average payload mass (kg) carried F9 v1.1 launches

AVG_PAYLOAD_MASS_F9
2928.4

First Successful Ground Landing Date

Date of first successful ground landing

FIRST_SUCCESSFUL_GROUND_PAD
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

Landing_Outcome	COUNT(Landing_Outcome)
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

Boosters which have carried the maximum payload mass of 15,600kg

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 in 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

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

Ranked 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

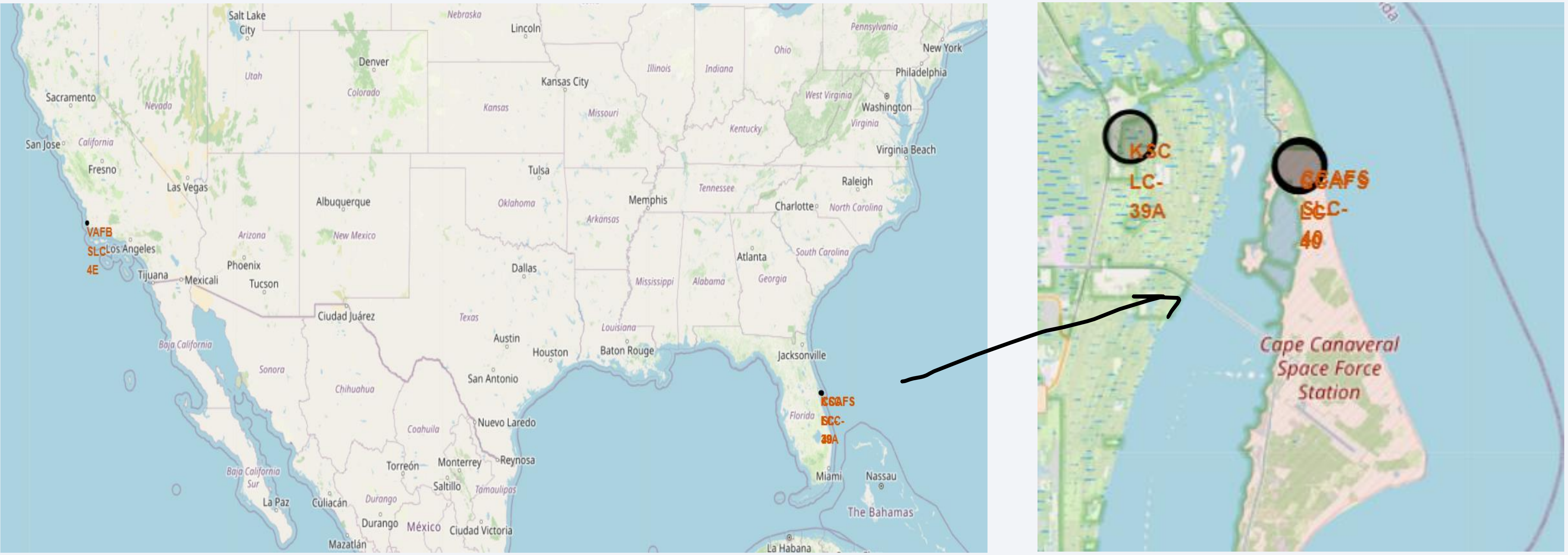
Landing_Outcome	count	Most_Recent_Date
No attempt	10	2017-03-16
Success (drone ship)	5	2017-01-14
Failure (drone ship)	5	2016-06-15
Success (ground pad)	3	2017-02-19
Controlled (ocean)	3	2015-02-11
Uncontrolled (ocean)	2	2014-09-21
Failure (parachute)	2	2010-12-08
Precluded (drone ship)	1	2015-06-28

Section 3

Launch Sites Proximities Analysis

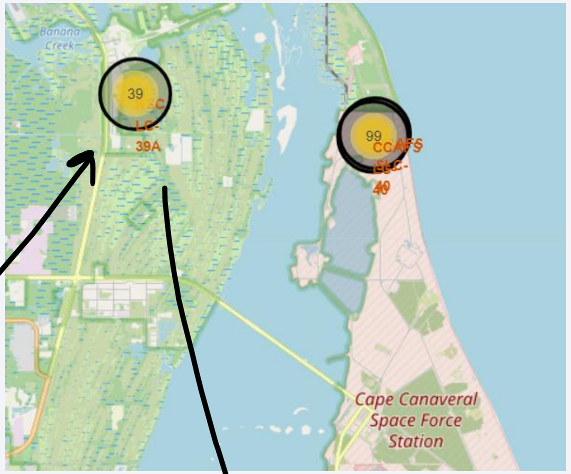
All Launch Sites

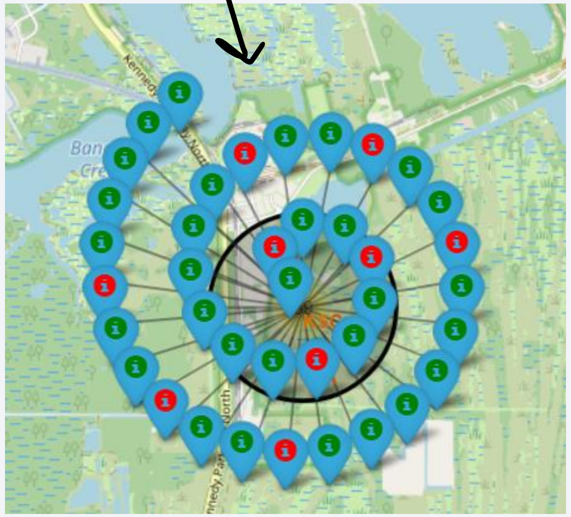
All launch sites are in coastal regions and in restricted areas



Success/Failed Launches for Each Site

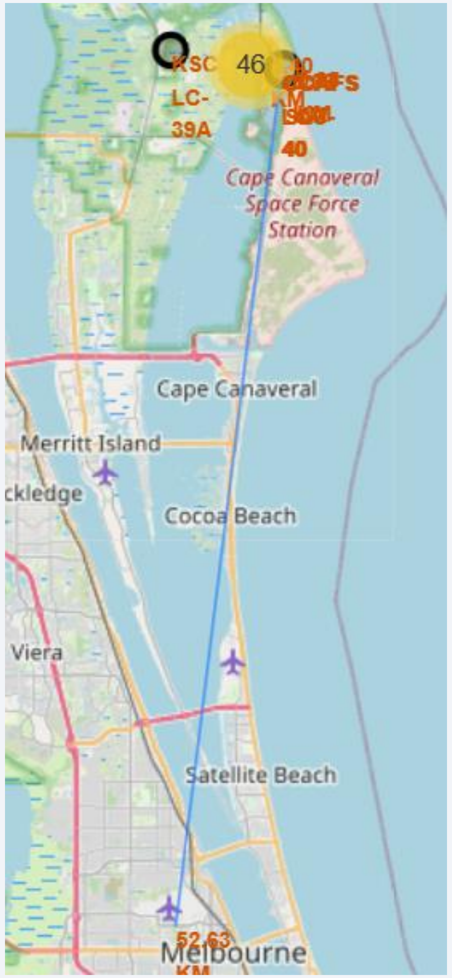
The largest of the maps shows cluster of launch information.
As you zoom into a single region, you'll find landings marked as failures (red) or success (green).

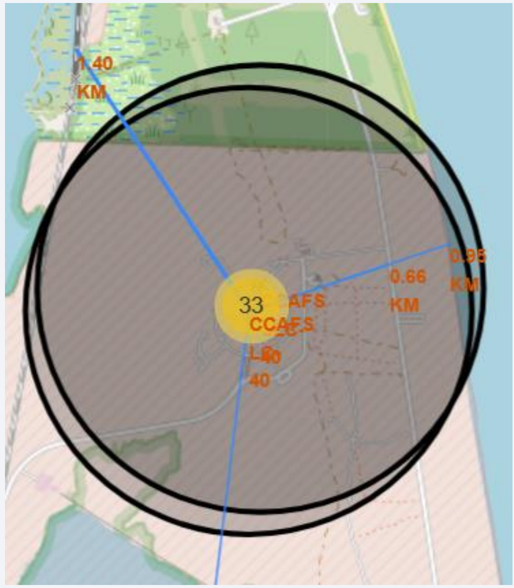




Launch Sites and Nearby Landmarks

- Typical of most launch sites is their near proximity to cities, railways, highways, and coastlines. Indicating population and movement.
- The images to the right show Melbourne is ~52km away, while railways, highways, and the coastline are all within ~1.5 km from the launch site.



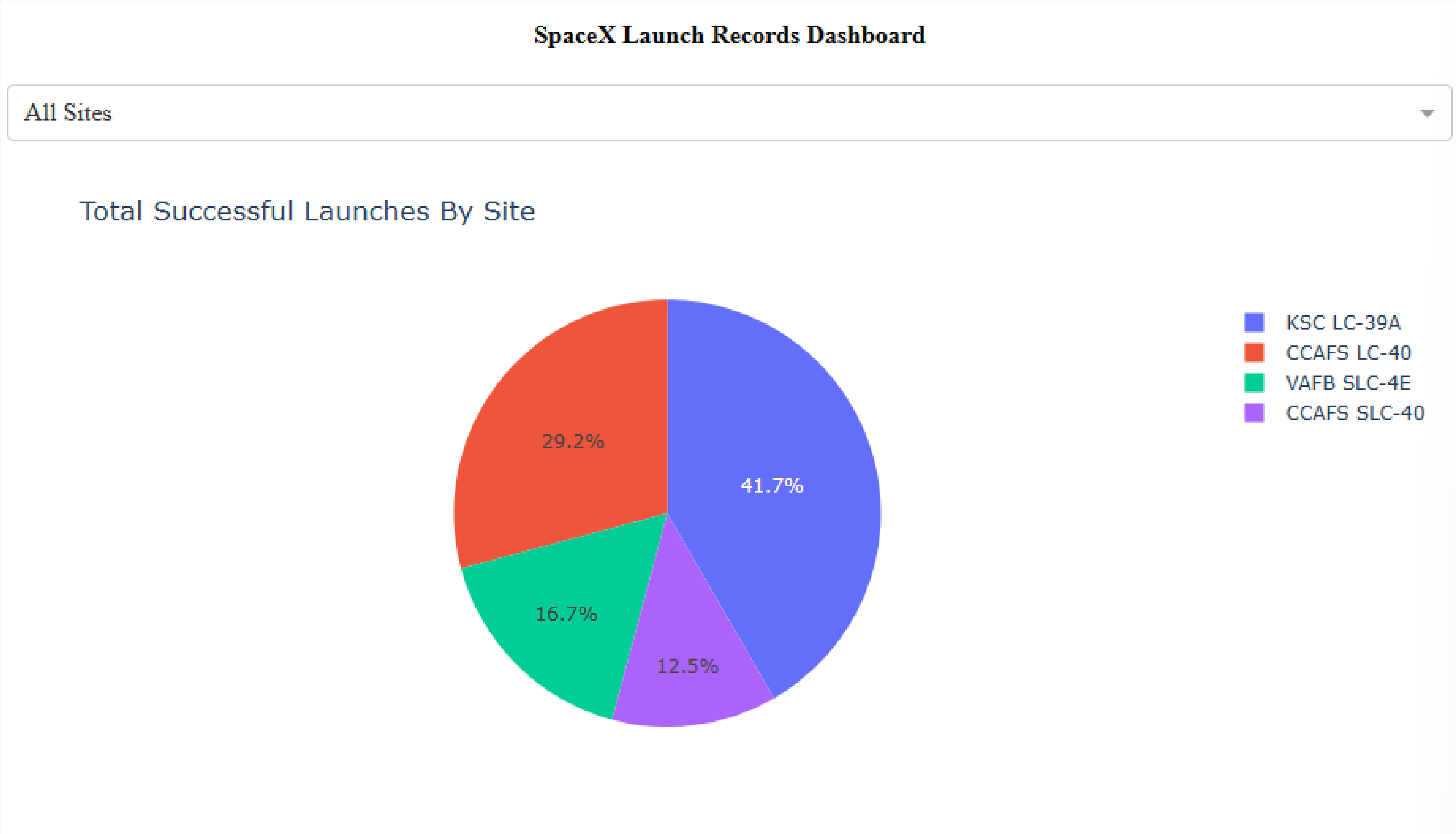


Section 4

Build a Dashboard with Plotly Dash

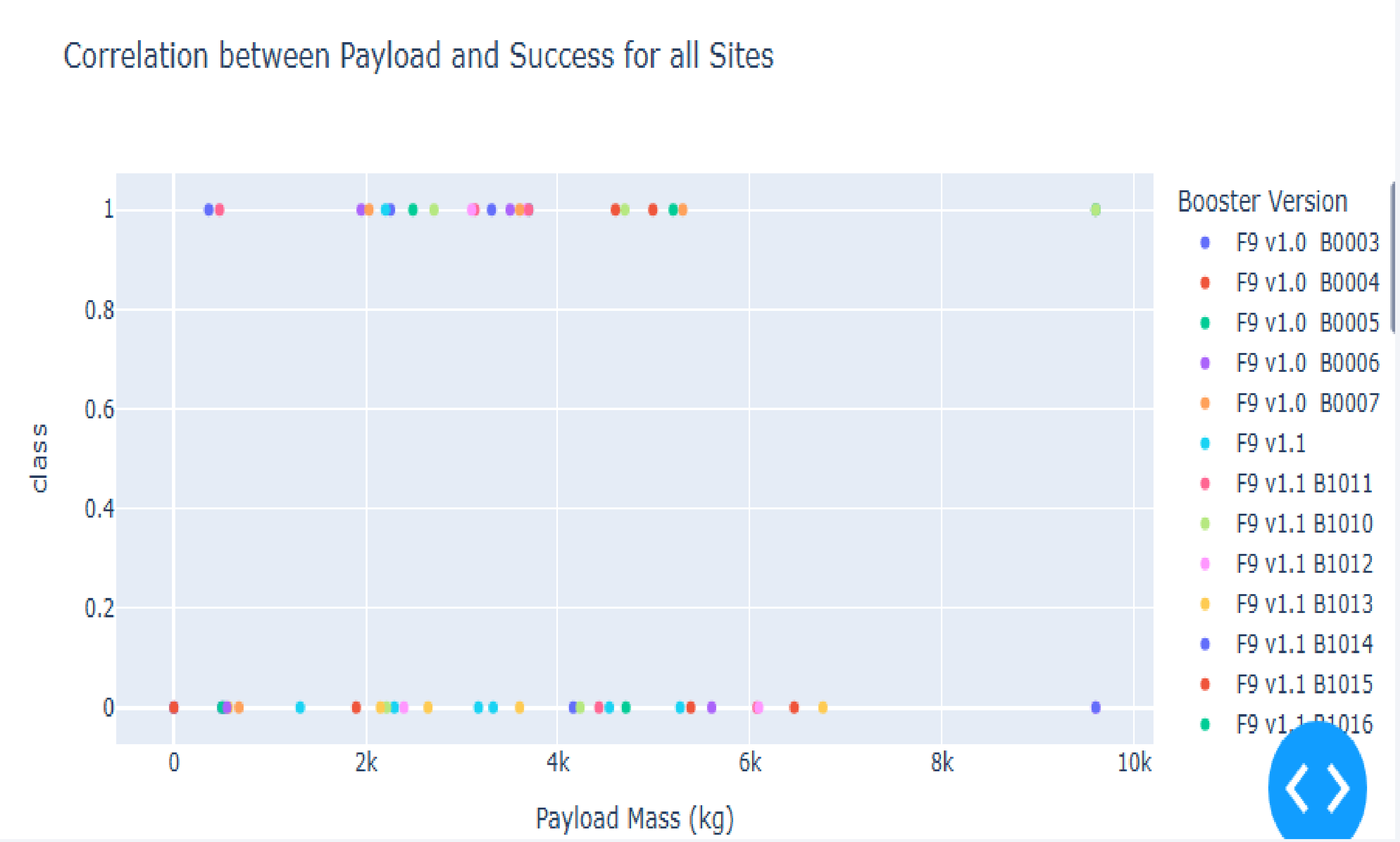
Total Success Launches by Site

- The interactive dashboard provides a dropdown menu, where one can view the individual success rate of a launch site, or the ratio of successful landings for all sites.
- A clear highlight is that KSC LC-39A has had the most successful launches, while CCAFS SLC-40 has performed poorly.



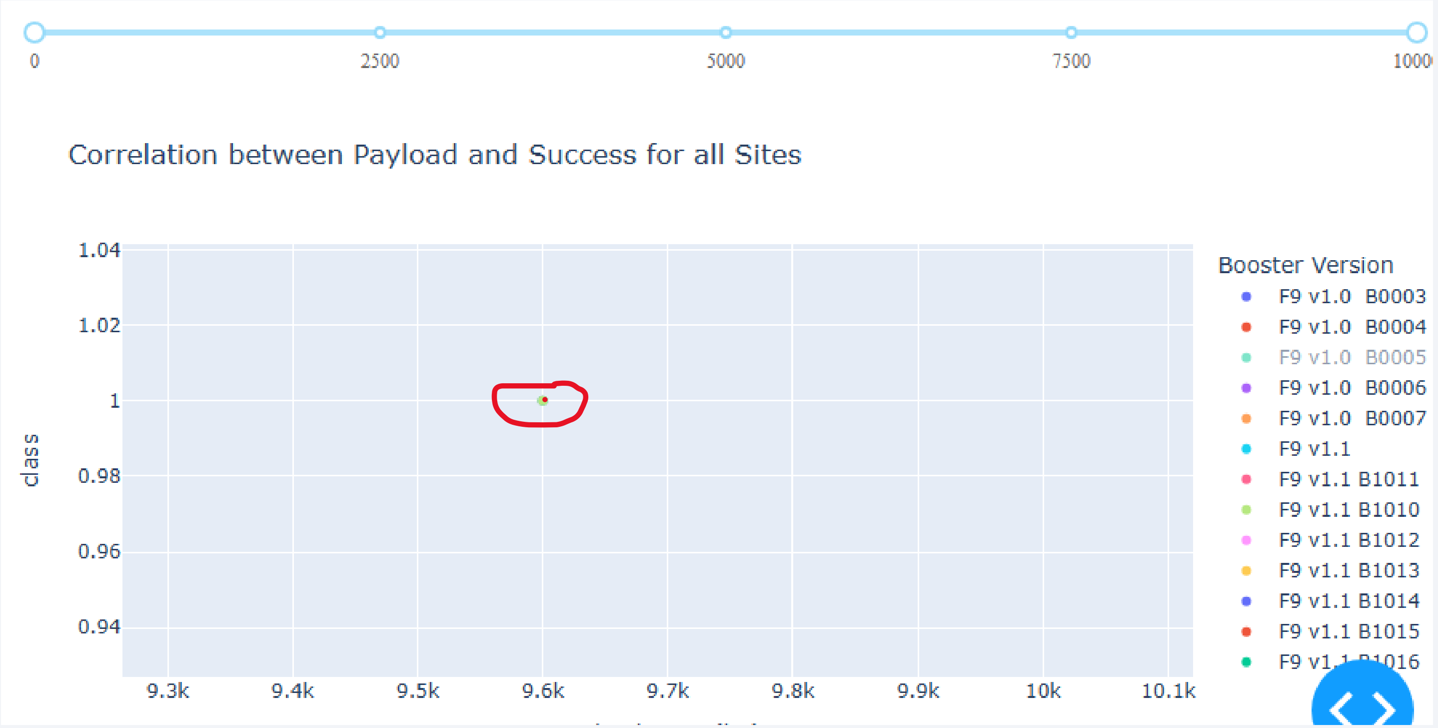
Payload vs. Launch Outcome

- The interactive dashboard allows one to choose a range of payload masses and further narrow down investigation by booster version.
- A critical finding for all boosters is the scarcity of success after 6,000 kg.



Top Payload Mass Success

- The only successful launch after 9,000 kg is the F9 B4.

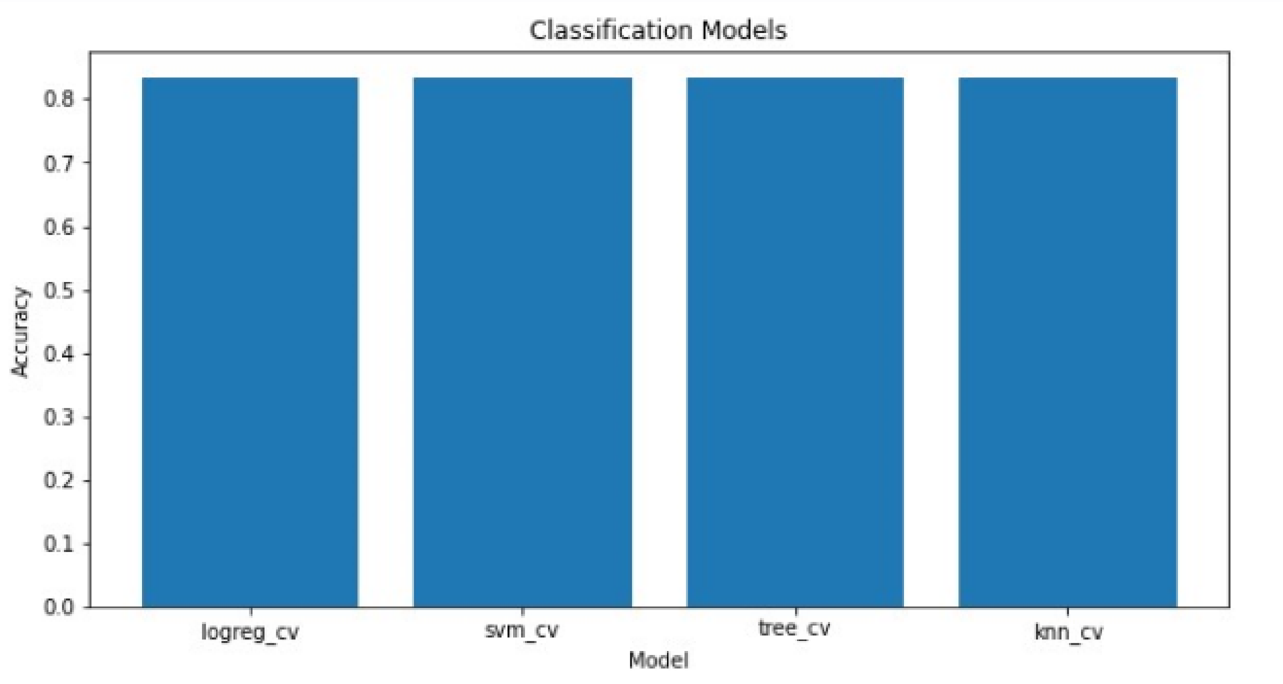


Section 5

Predictive Analysis (Classification)

Classification Accuracy

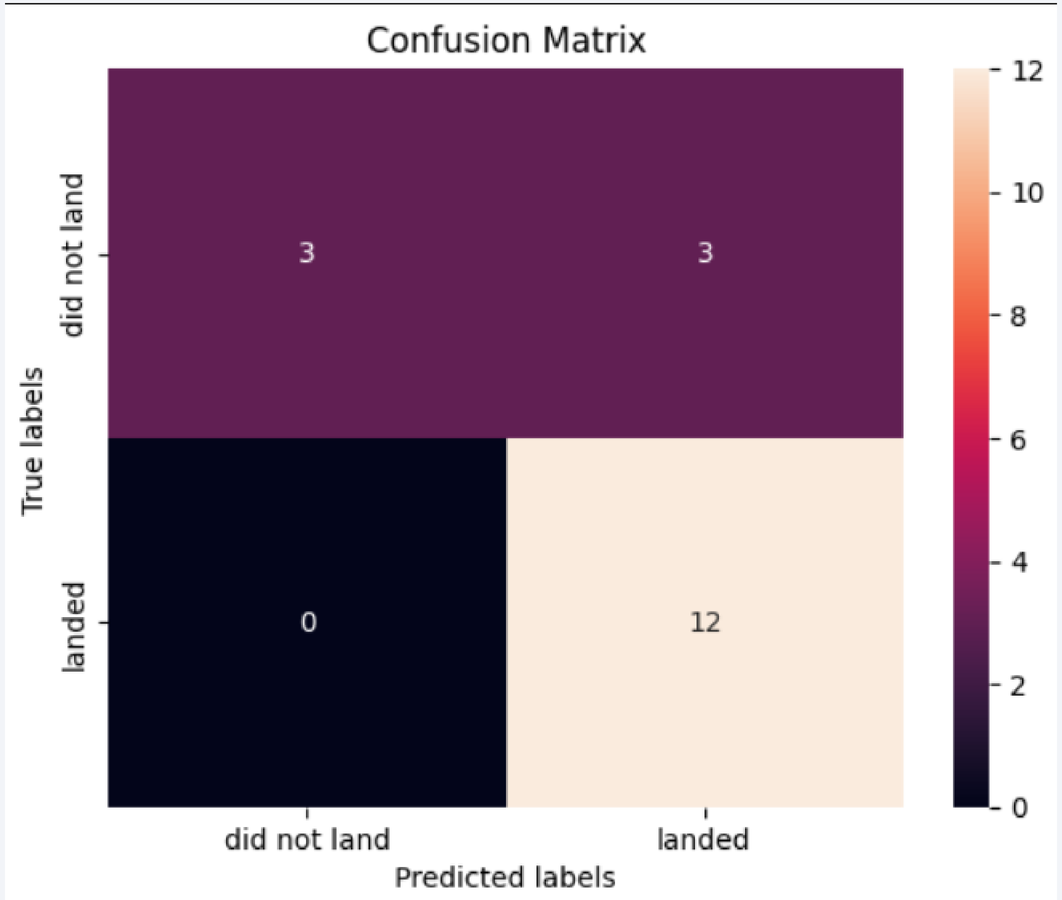
Accuracy was approximately equivalent for all models.



Confusion Matrix

The confusion matrix was the same for all models.

- True Positives = 12
- True Negatives = 3
- False Positives = 0
- False Negatives = 3



Conclusions

- Given that our models perform nearly the same, it is not unreasonable to select a model based on computational simplicity.
- By using any of the models, we can guarantee an approximate accuracy of 84%

Appendix

For notebooks, scripts, and dataset interactions, navigate to my Github:

[Cost of a SpaceX Falcon9 Launch]

Thank you!