IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zachery Lahti
1st July 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

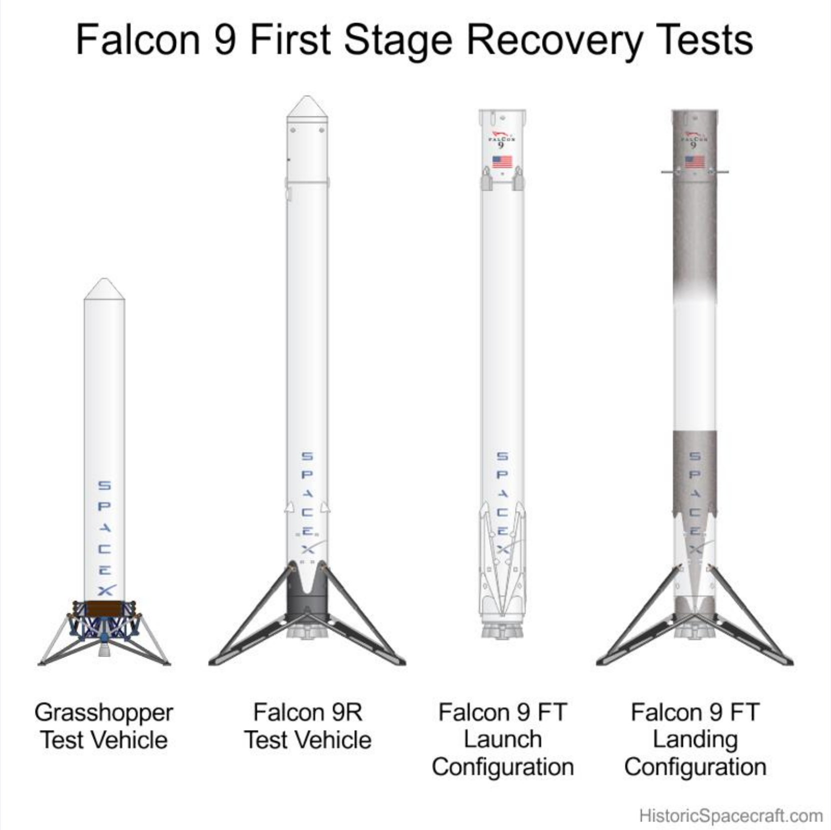
With the departure of NASA from the space exploration industry, the opportunity has been passed onto the private sector. Virgin Galactic, Rocket Lab, Blue Origin and SpaceX have all entered the field, intent on launching technologies into orbit, transporting payloads and astronauts to the International Space Station, and potentially developing a space tourism industry.

However, development of newer propulsion technologies comes at a high cost. Currently, it costs around $165 million USD to launch a rocket into orbit. SpaceX is currently leading the technology race, having developed a reusable first-stage rocket, thereby decreasing launch costs of the Falcon 9 rocket to around $62 million. The data presented here evaluates whether SpaceY, a new entrant into the field, can submit a competitive bid against SpaceX for a rocket launch.

Data was extracted from SpaceX API and SpaceX's Wikipedia page. SQL, data visualization, folium maps and plotly dashboards, were utilized to explore the data, while GridSearchCV and logistic regression was used to determine potential parameters for predictive machine learning models. Successful landings were overpredicted and we propose that further data is warranted.

Introduction

- Closure of NASA's space program has shifted space flight programs to the private sector
- Single-use rocket launches cost ≈ $165 million per launch
- 1st stage reusable rockets cost ≈ $62 million *per launch*
- SpaceY aims to predict:
  - Most optimal launch locations for successful landing
  - Most optimal payload mass for successful landing


Falcon 9 First Stage Recovery Tests
Grasshopper Test Vehicle
Falcon 9R Test Vehicle
Falcon 9 FT Launch Configuration
Falcon 9 FT Landing Configuration
HistoricSpacecraft.com

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Public data was collected through SpaceX API and scraped from SpaceX's Wikipedia
- Perform data wrangling
    - Data was processed by using Pandas and NumPy, focusing on Launch Sites, Orbit and Outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Models were tuned using standardized data and best parameters using GridSearchCV

Data Collection

SpaceX API Variables Collected:

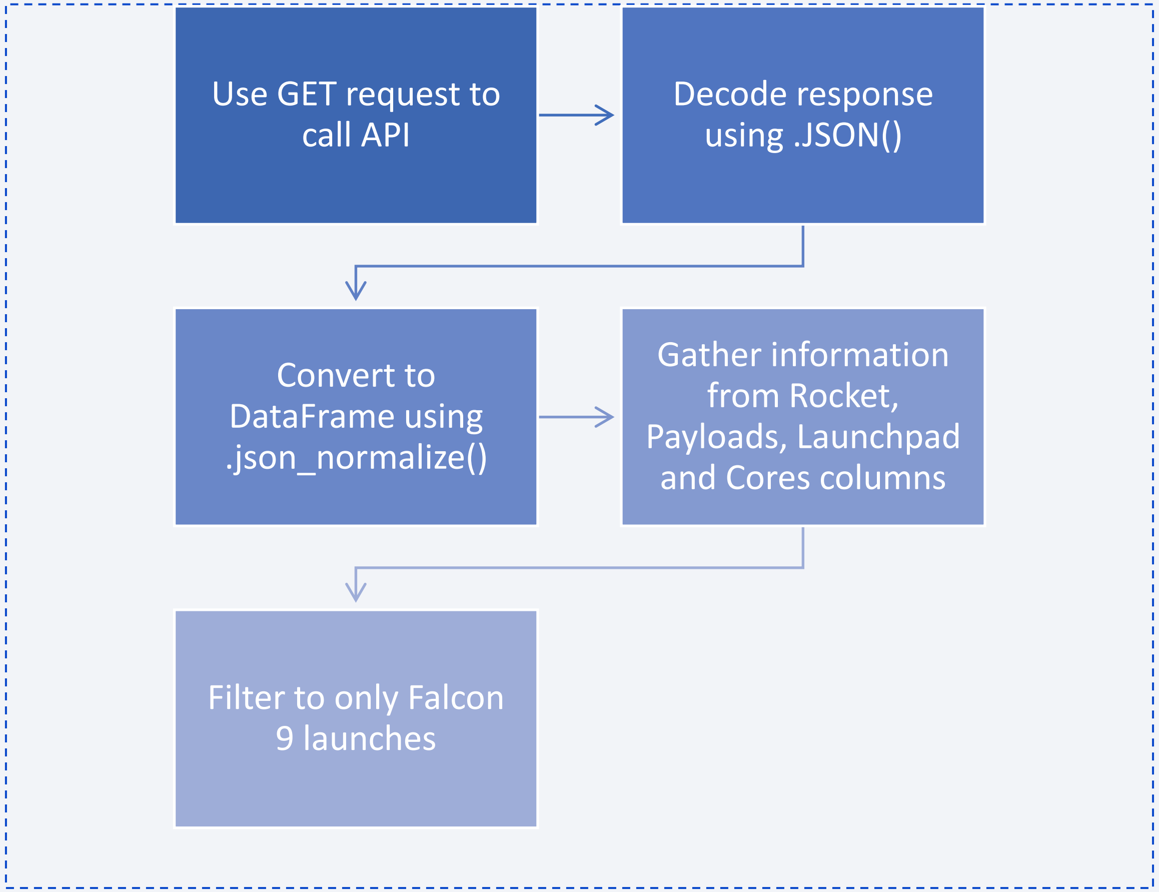
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
    - SpaceX URL (here)

Web Scraping Variables Collected:

- Flight No, Launch Site, Payload Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
    - Static URL (here)

Data Collection – SpaceX API

- GitHub URL



Use GET request to call API

Decode response using .JSON()

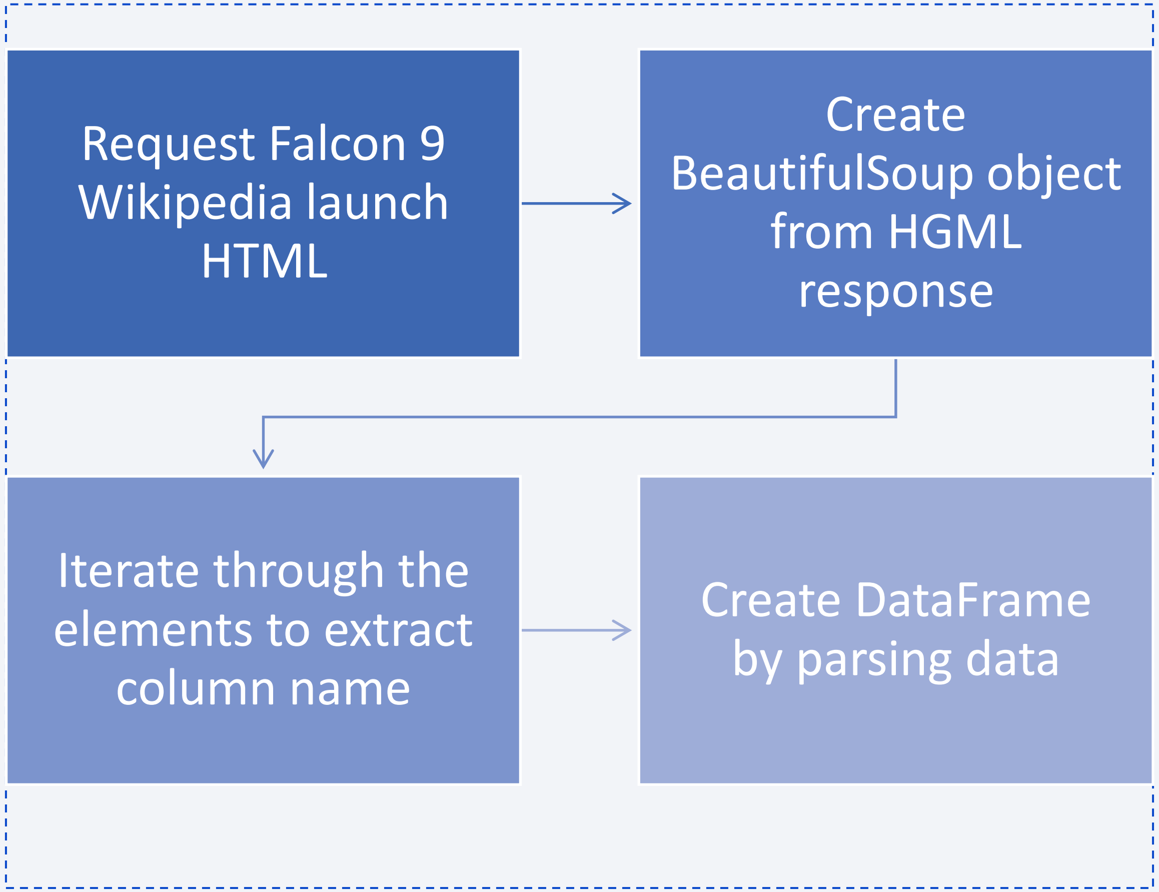
Convert to DataFrame using .json_normalize()

Gather information from Rocket, Payloads, Launchpad and Cores columns

Filter to only Falcon 9 launches

Data Collection - Scraping

- GitHub URL



Request Falcon 9 Wikipedia launch HTML

Create BeautifulSoup object from HGML response

Iterate through the elements to extract column name

Create DataFrame by parsing data

Data Wrangling

Launch sites were defined as:

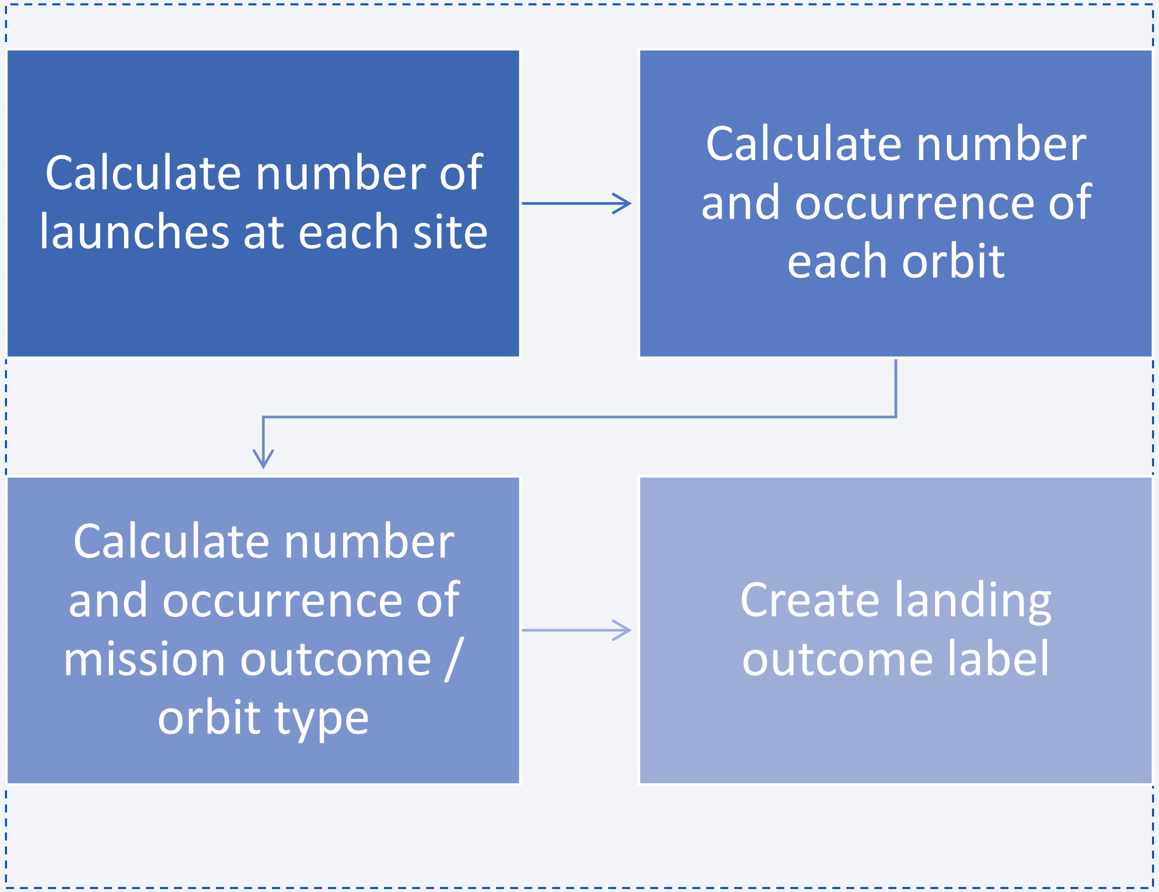
- Cape Canaveral Space Launch Complex 40 = CCAFS (S)LC 40
- Vandenberg Air Force Base Space Launch Complex 4E = VAFB SLC 4E
- Kennedy Space Center Launch Complex 39A = KSC LC 39A

Landing outcomes defined as:

- Landing_class = 0 (bad outcome)
- Landing_class = 1 (successful outcome)

Data Wrangling

- GitHub URL



Calculate number of launches at each site

Calculate number and occurrence of each orbit

Calculate number and occurrence of mission outcome / orbit type

Create landing outcome label

EDA with Data Visualization

Plots Analyzed:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Yearly Trend vs. Success Rate

Libraries Used:

- Pandas, NumPy, Matplotlib.pyplot, Seaborn

Plot Type Used:

- Scatter
- Bar
- Line

- GitHub URL

EDA with SQL

SQL (using SQLAlchemy)

- Identified average payload mass carried by booster version F9 v1.1
- Identified Boosters that had a successful or failed landing on drone ship, with payload mass between 4000 and 6000 (kg)
- Identified Boosters that had successful or failed landing on land
- Rank landing outcomes on both drone ship and land

- GitHub URL

Build an Interactive Map with Folium

Folium used to:

- Mark all launch sites
- Mark success / failed launches for each site

Additions:

- Markers, circles and lines used to highlight launch sites, successful / failed landings, and proximity to various locations / structures

- GitHub URL

Build a Dashboard with Plotly Dash

- Plotly Dash used to:
  - Visualize successful landings across launch sites, payload mass, booster version

- Inputs used:
  - All sites
  - Individual launch site

- Plots used:
  - Pie chart
  - Scatter plot

- GitHub URL

Predictive Analysis (Classification)

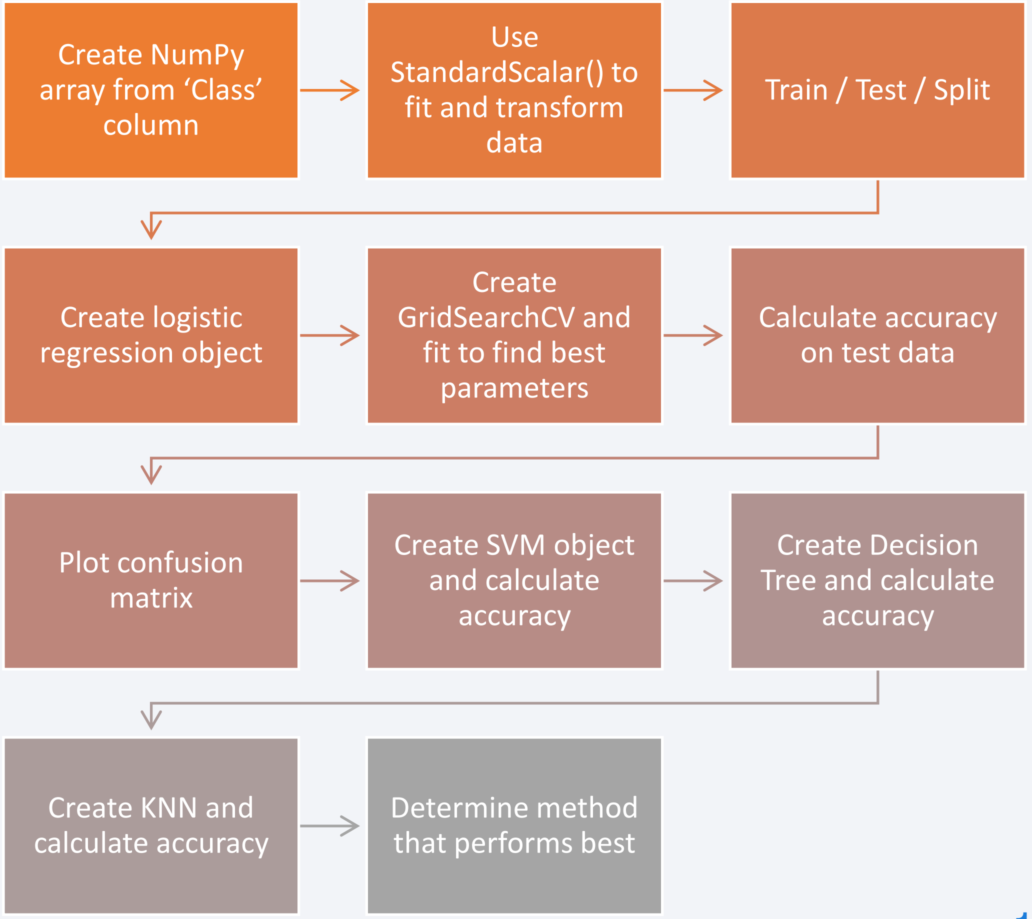
Train / Test / Split

- Transform data using preprocessing.StandardScaler()
- test_size=0.2
- random_state=2

GridSearchCV to find optimal parameters

- cv=10
- Logistic regression
- SVM
- Decision Tree
- K Nearest Neighbors

- GitHub URL


Create NumPy array from 'Class' column
Use StandardScalar() to fit and transform data
Train / Test / Split
Create logistic regression object
Create GridSearchCV and fit to find best parameters
Calculate accuracy on test data
Plot confusion matrix
Create SVM object and calculate accuracy
Create Decision Tree and calculate accuracy
Create KNN and calculate accuracy
Determine method that performs best

Results

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

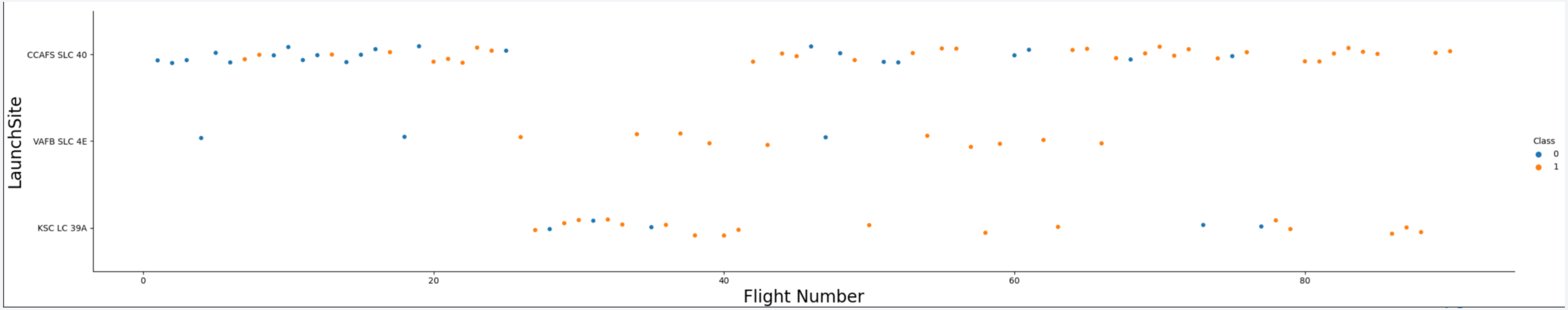
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

**Blue = Failed landing / Orange = Successful landing

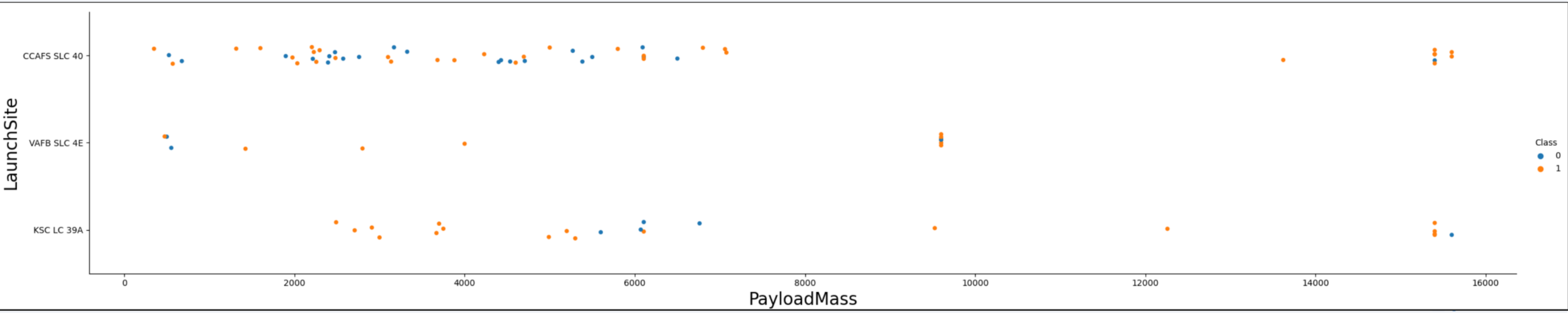
- CCAFS SLC 40 has conducted the most launches
- Landing success has increased over time



Payload vs. Launch Site

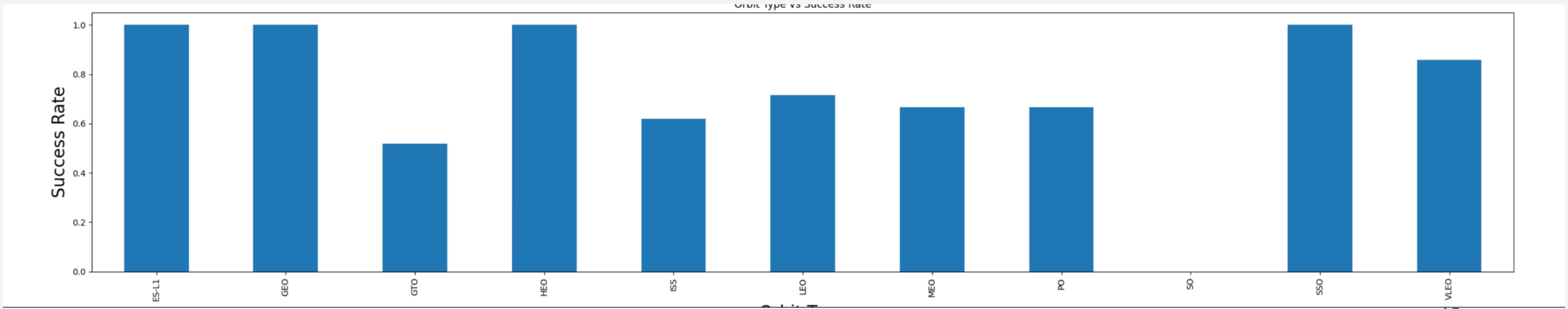
**Blue = Failed landing / Orange = Successful landing

- Majority of launches are with payload mass < 10,000 (kg)
- No heavy payload launches > 10,000 (kg) for VAFB-SLC



Success Rate vs. Orbit Type

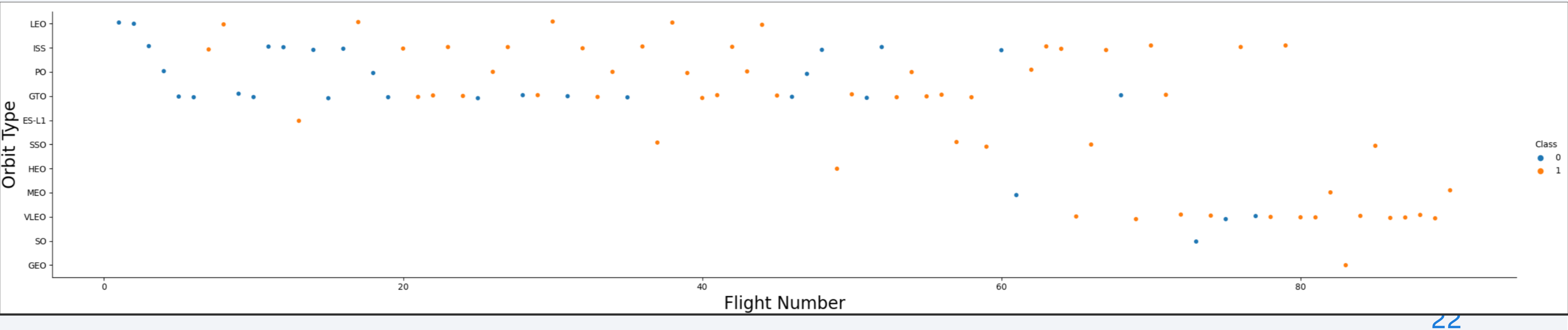
- ES-L1, GEO, HEO and SSO all have the highest success rate
- GTO has the lowest success rate



Flight Number vs. Orbit Type

**Blue = Failed landing / Orange = Successful landing

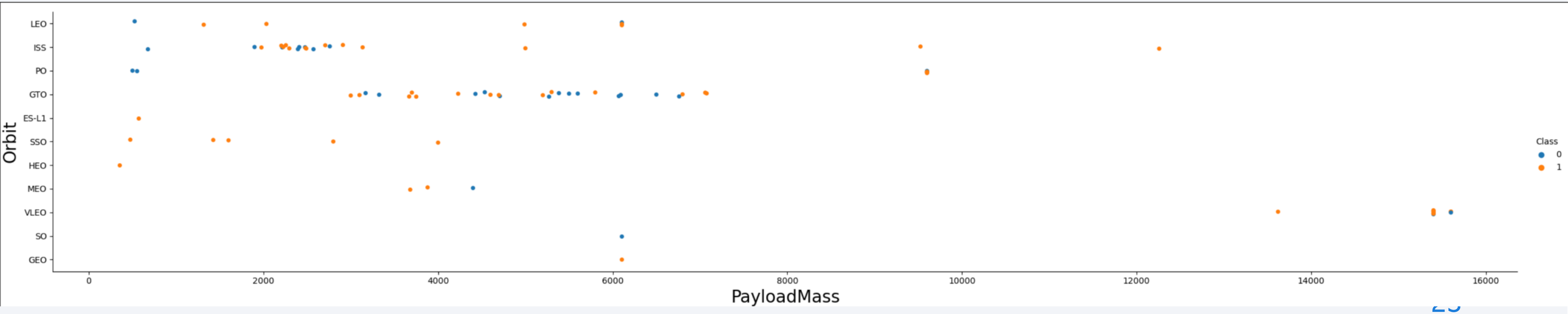
- Lower earth orbits and sun-synchronous orbits appear to have higher success rates



Payload vs. Orbit Type

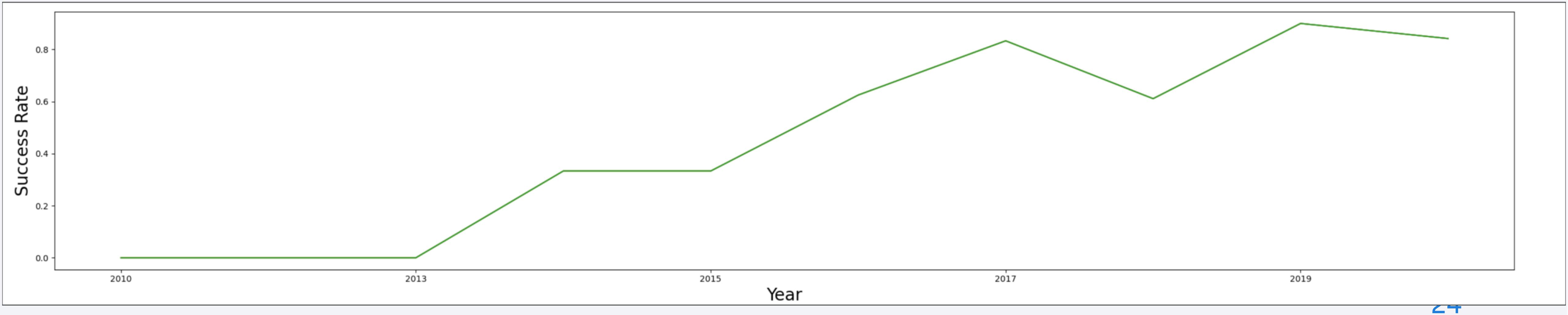
**Blue = Failed landing / Orange = Successful landing

- LEO, ISS and PO orbits appear to produce more successful landings
- GTO and payload mass does not appear to produce a consistent success rate
- Minimal, inconclusive data for higher orbits



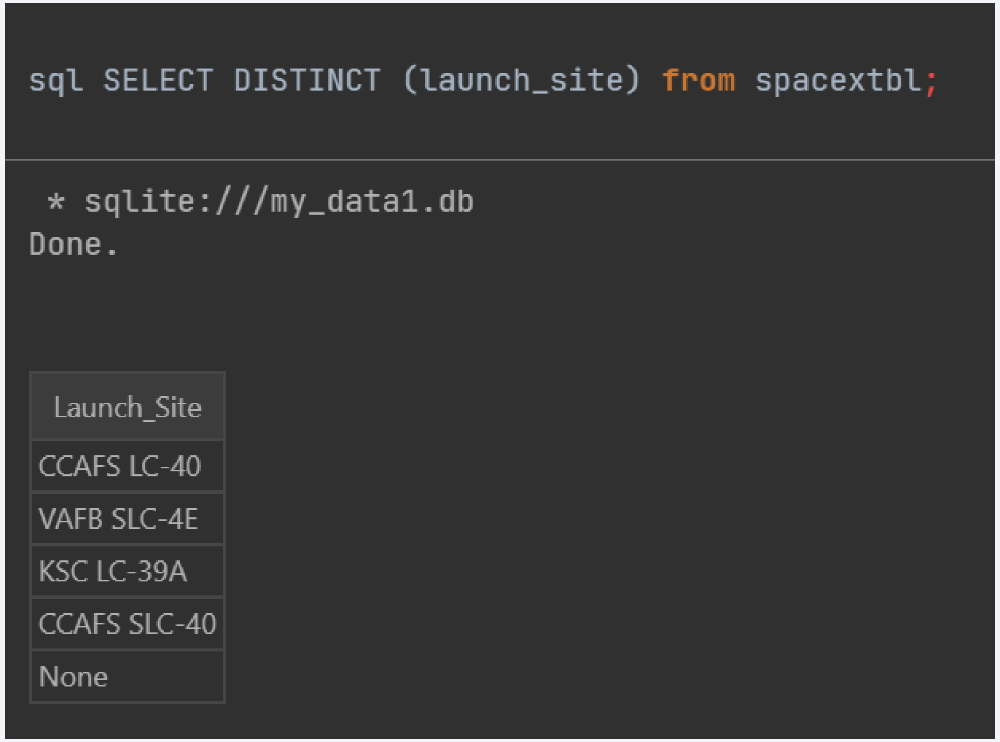
Launch Success Yearly Trend

- Rate of successful landings increases over time
- 2019 success rate exceeded 80%



All Launch Site Names

- Data query of launch_site produced four distinct launch sites
- Note that CCAFS LC-40 and CCAFS SLC-40 are two distinct sites on the same premises and are considered the same site for our analysis



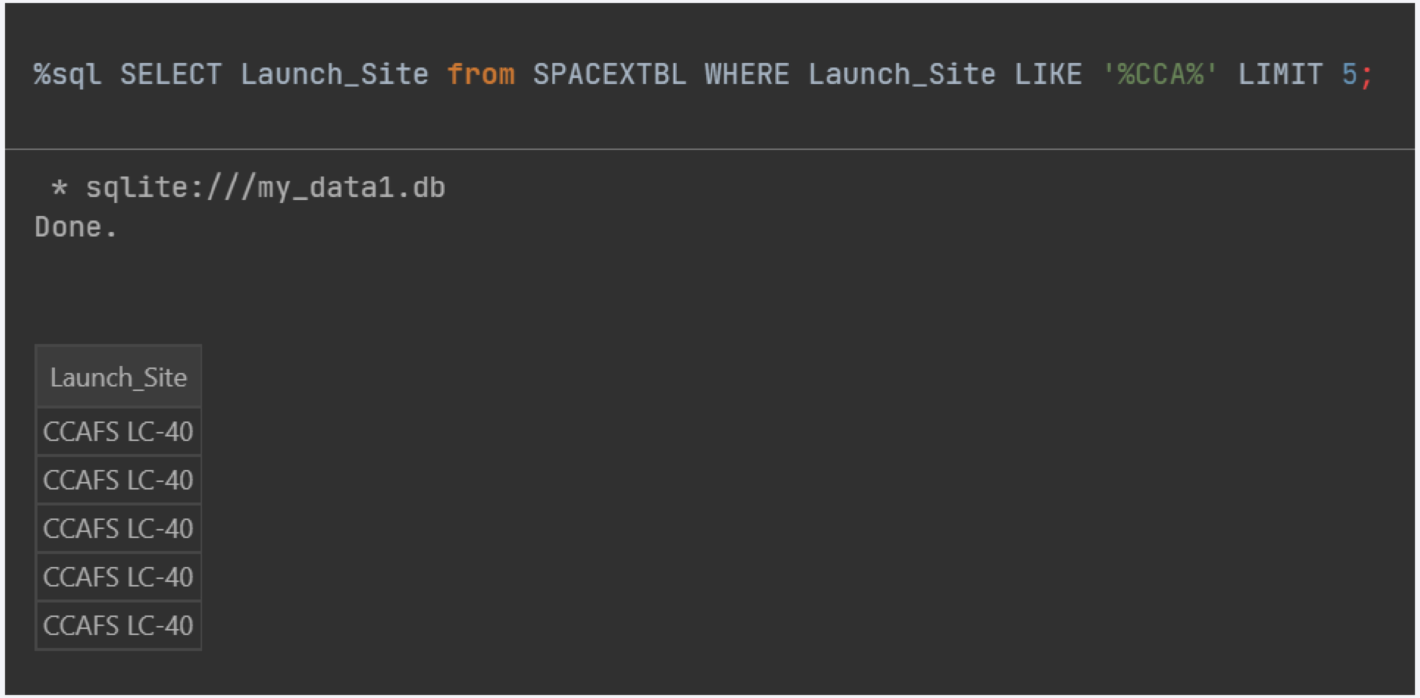
```
sql SELECT DISTINCT (launch_site) from spacextbl;
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- Using the query to determine launch sites begin with `CCA`, the following output was produced



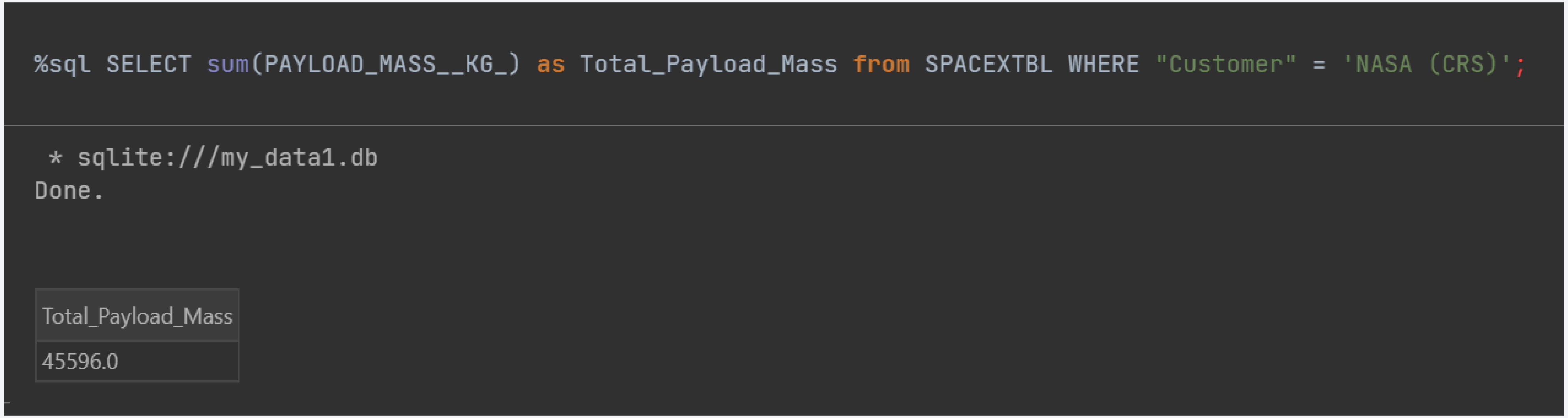
```
%sql SELECT Launch_Site from SPACEXTBL WHERE Launch_Site LIKE '%CCA%' LIMIT 5;
```

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

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

Total Payload Mass

- Where NASA was the customer, a total mass of 45,596 (kg) was launched



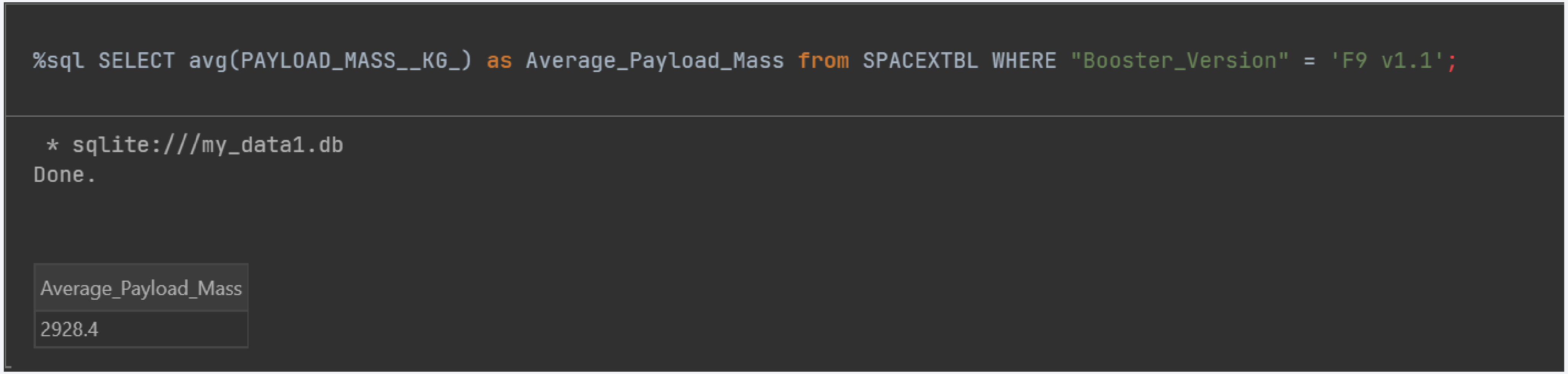
```
%sql SELECT sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTBL WHERE "Customer" = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596.0

Average Payload Mass by F9 v1.1

- Average payload mass / launch by the F9 v1.1 booster was 2,928.4 (kg)



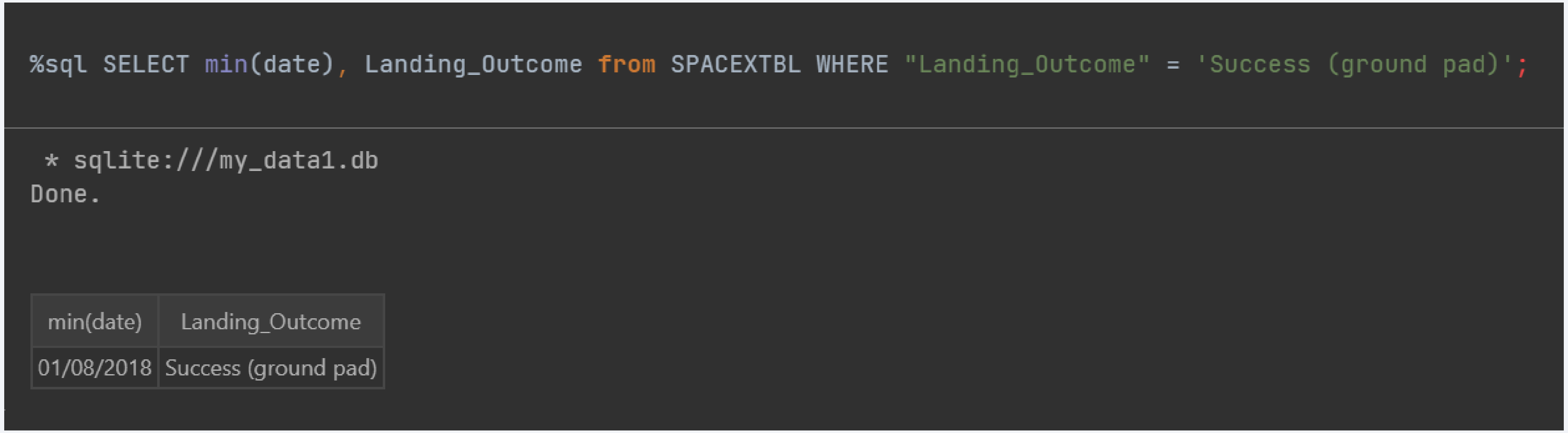
```
%sql SELECT avg(PAYLOAD_MASS__KG_) as Average_Payload_Mass from SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- The query used to return the first successful group pad landing date produced a date in August. However, due to the internal settings of the local computer set to European standards, we believe this is incorrect. A deeper look into the dataset revealed that the correct data is 12.22.2015



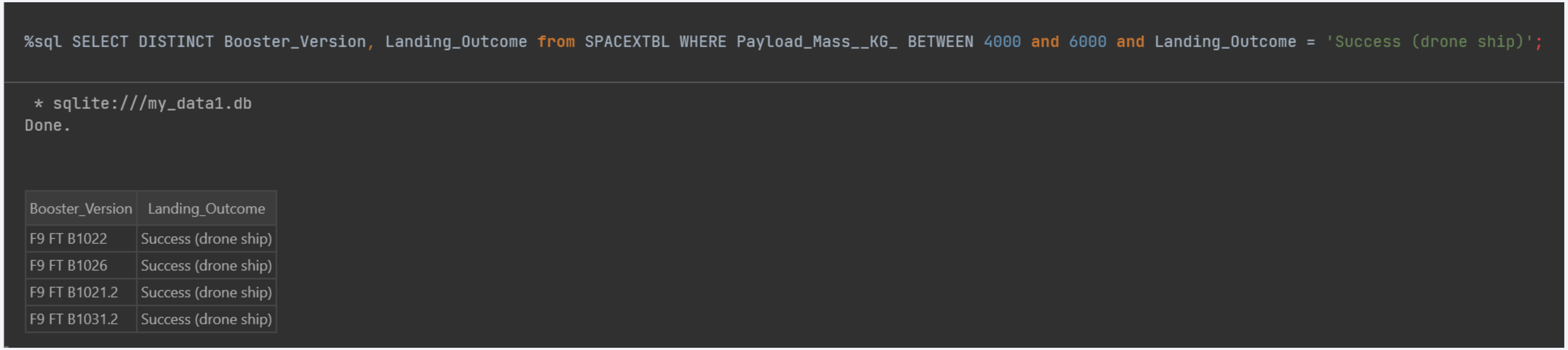
```
%sql SELECT min(date), Landing_Outcome from SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

min(date)	Landing_Outcome
01/08/2018	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- When payload mass (kg) was between 4000 and 6000, four booster versions produced a successful drone ship landing



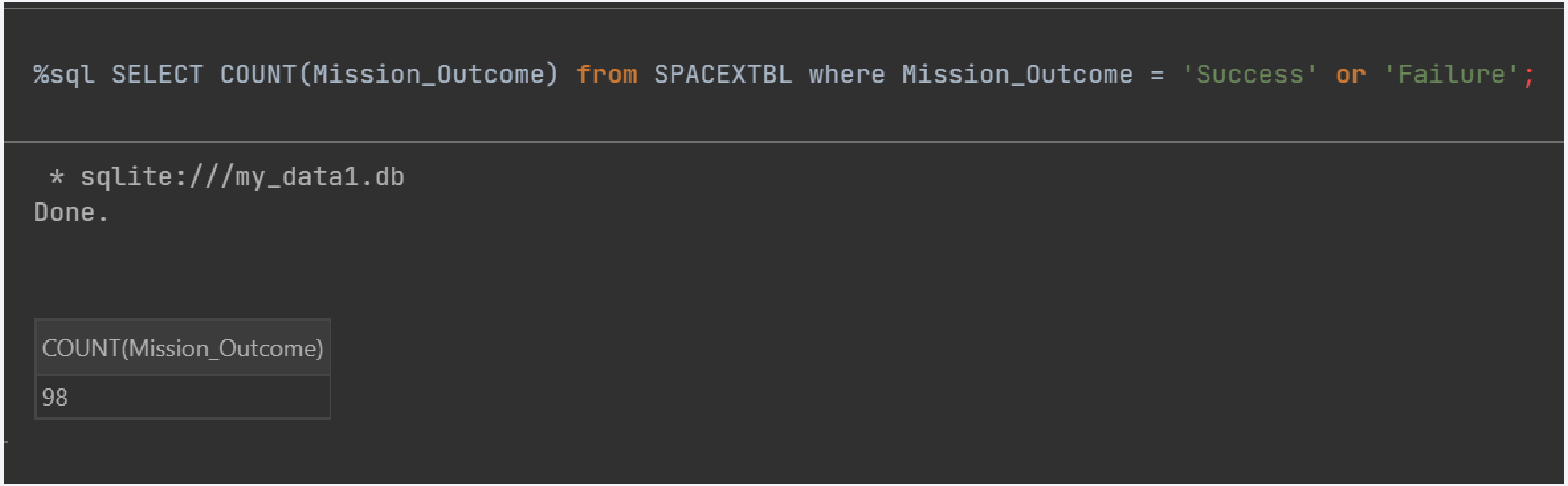
```
%sql SELECT DISTINCT Booster_Version, Landing_Outcome from SPACEXTBL WHERE Payload_Mass__KG_ BETWEEN 4000 and 6000 and Landing_Outcome = 'Success (drone ship)';
```

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

Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- This query returned a total of 98 successful mission outcomes



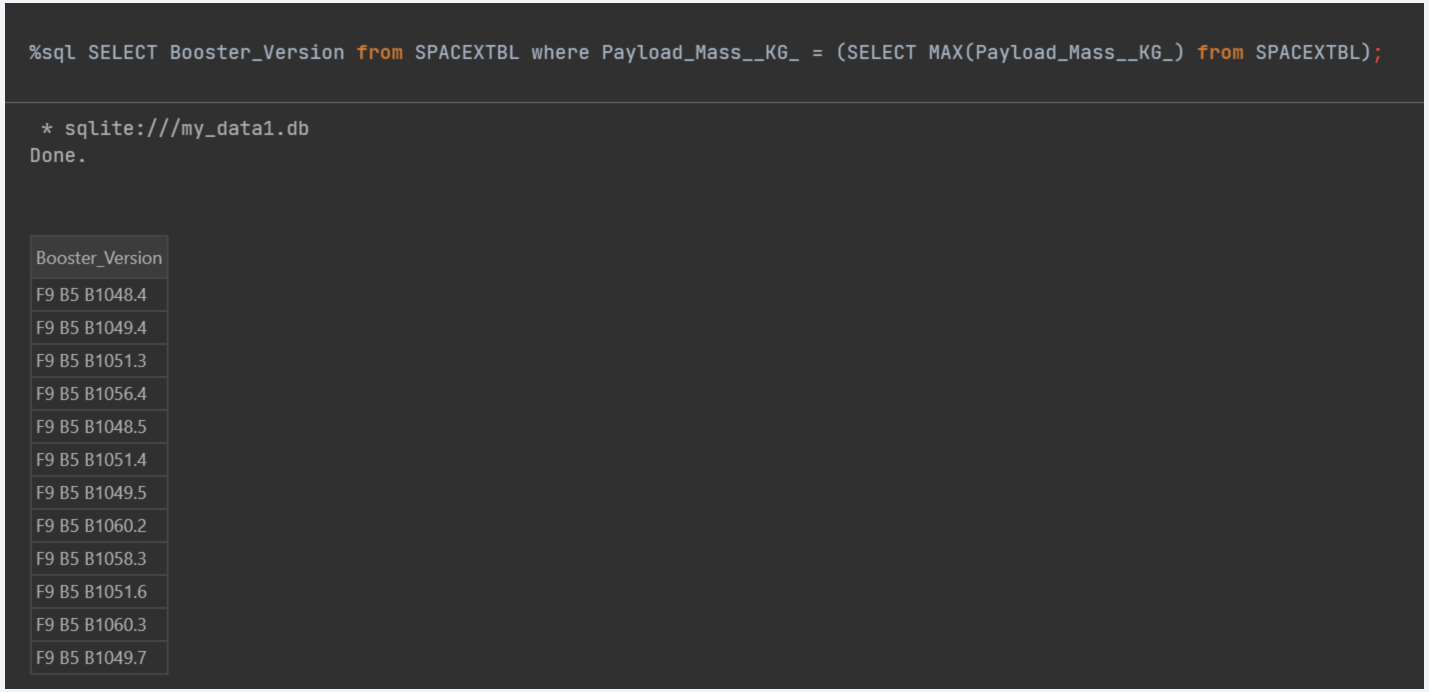
```
%sql SELECT COUNT(Mission_Outcome) from SPACEXTBL where Mission_Outcome = 'Success' or 'Failure';
```

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

COUNT(Mission_Outcome)
98

Boosters Carried Maximum Payload

- Maximum payloads were all carried by the F9 B5 B10xx.x booster



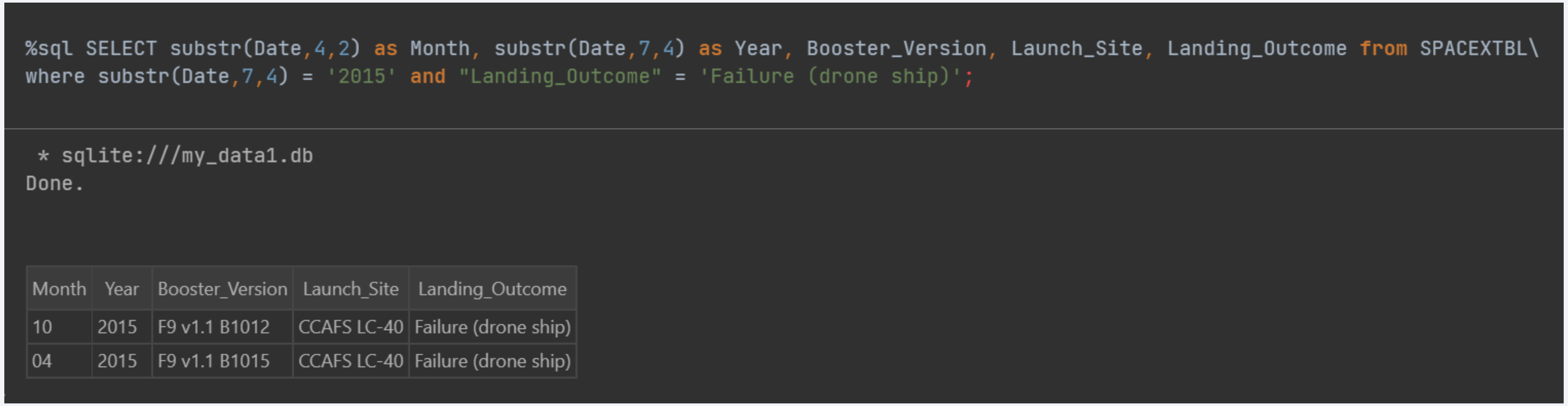
```
%sql SELECT Booster_Version from SPACEXTBL where Payload_Mass__KG_ = (SELECT MAX(Payload_Mass__KG_) from SPACEXTBL);
```

```
 * 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

- This query produced the launches for April and October 2015, showing that both experienced failed landings on a drone ship when launched from the CCAFS LC-40 site



```
%sql SELECT substr(Date,4,2) as Month, substr(Date,7,4) as Year, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTBL\
where substr(Date,7,4) = '2015' and "Landing_Outcome" = 'Failure (drone ship)';
```

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

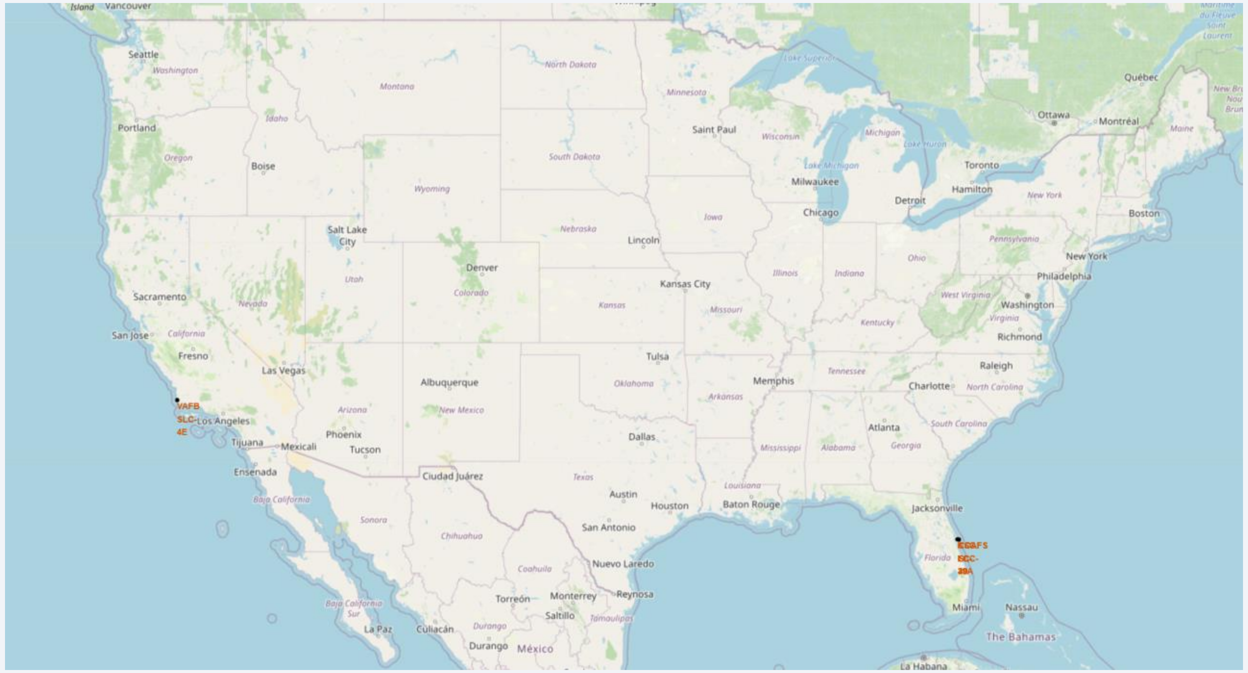
Month	Year	Booster_Version	Launch_Site	Landing_Outcome
10	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

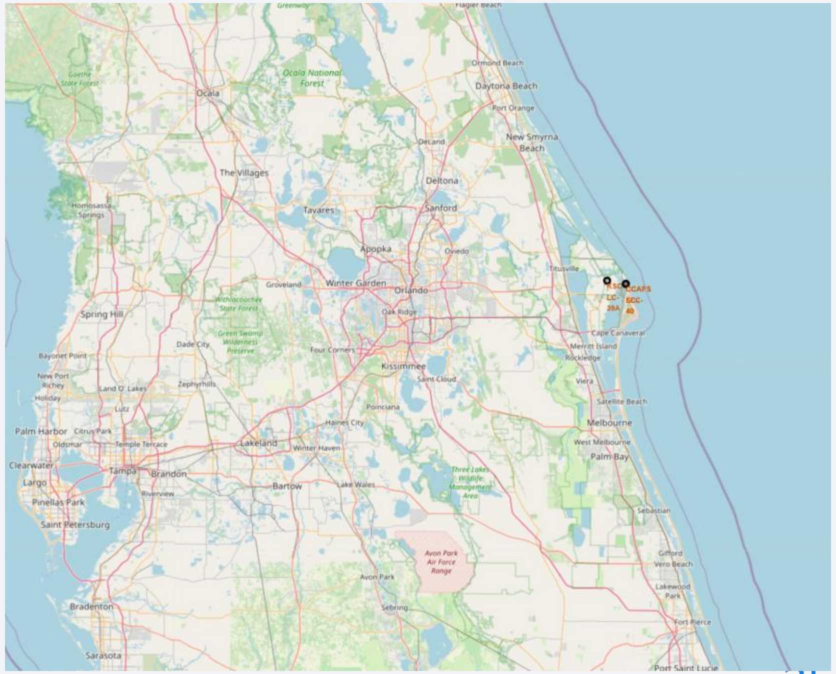
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

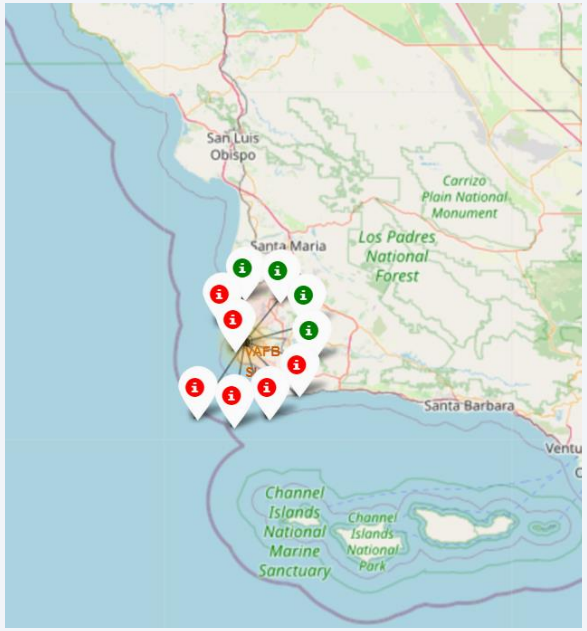
- One launch site location VAFB SLC-4E, located on the west coast
- Two launch sites; KSC LC-39A and CCAFS-40, located on the east coast

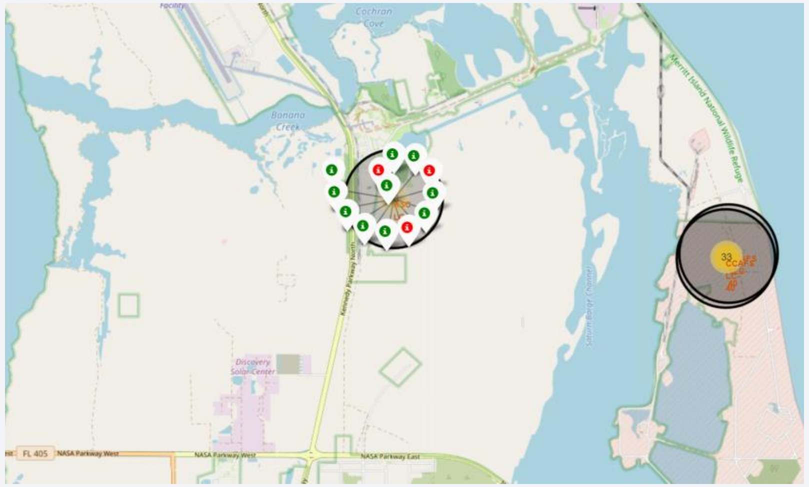


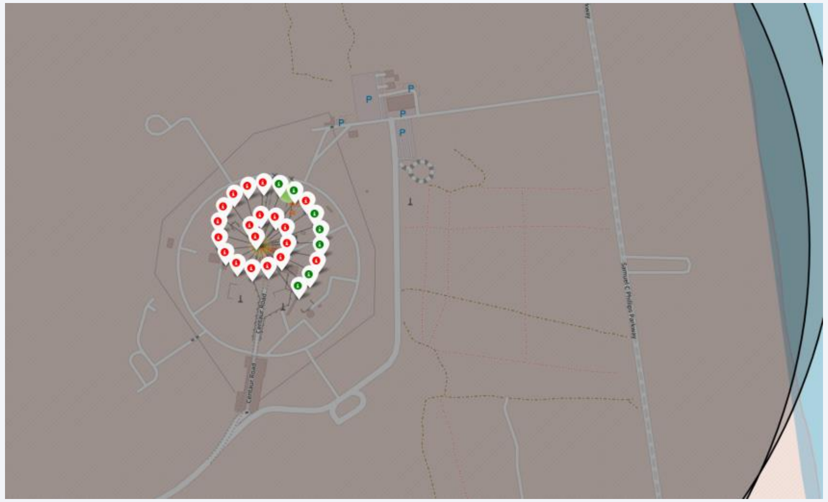


Launch Markers

- Folium map detailing successful landings (in green) and failed landings (in red) for VAFB SLC-4E (4 successful), KSC LC-39A (10 successful), and CCAFS LC-40 (10 successful; *not shown = 3 CCAFS SLC-40 site*) launch sites, respectively

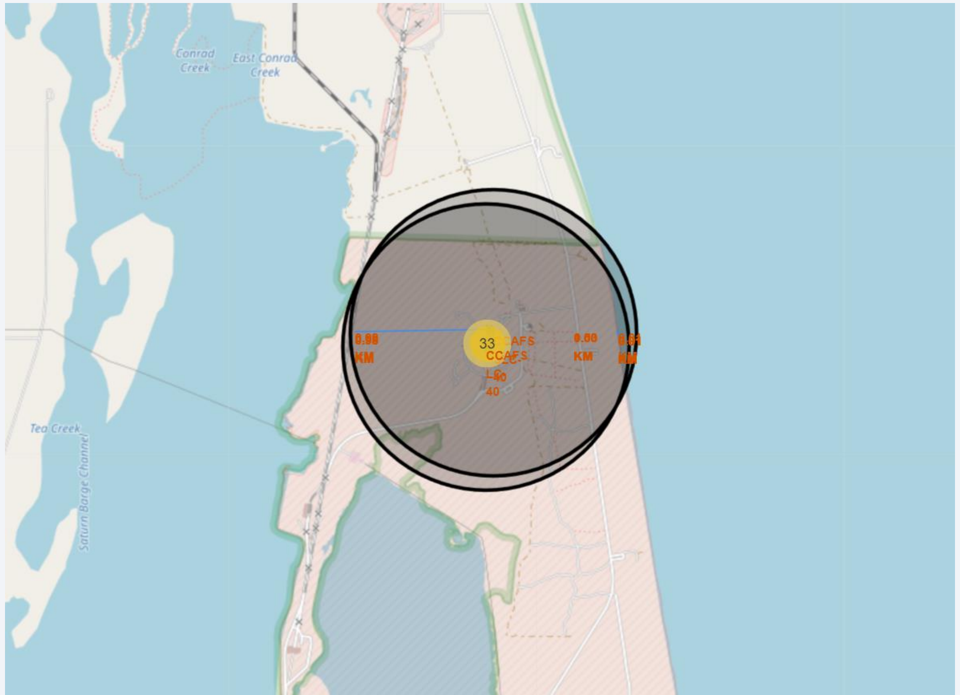

San Luis Obispo
Carrizo Plain National Monument
Santa Maria
Los Padres National Forest
VAFB
Santa Barbara
Ventu
Channel Islands National Marine Sanctuary
Channel Islands National Park

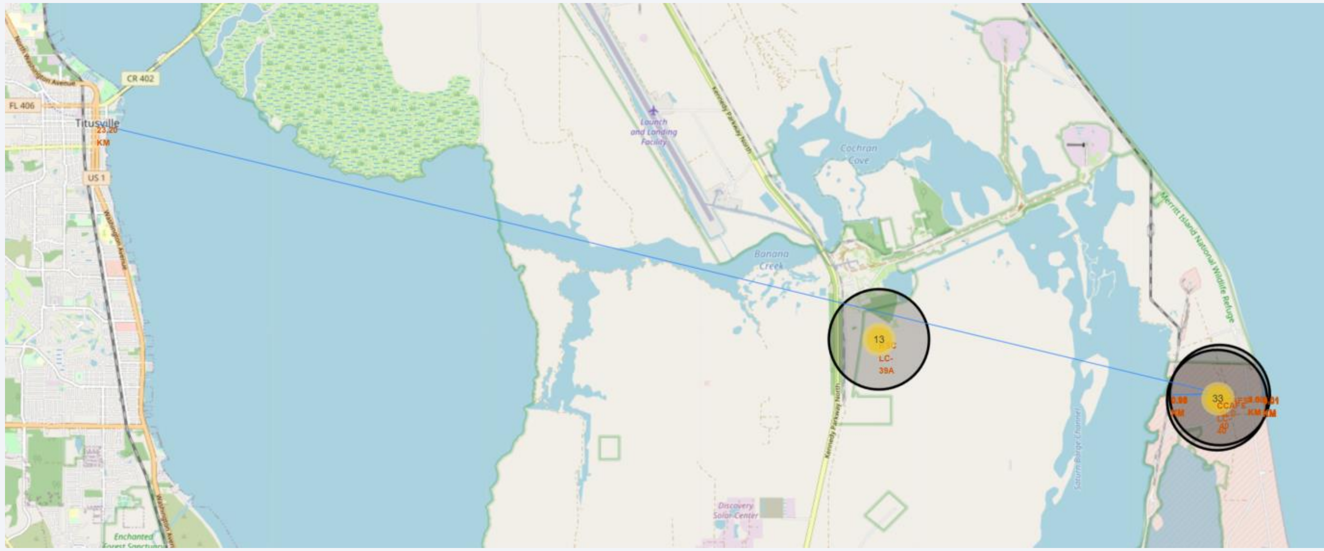


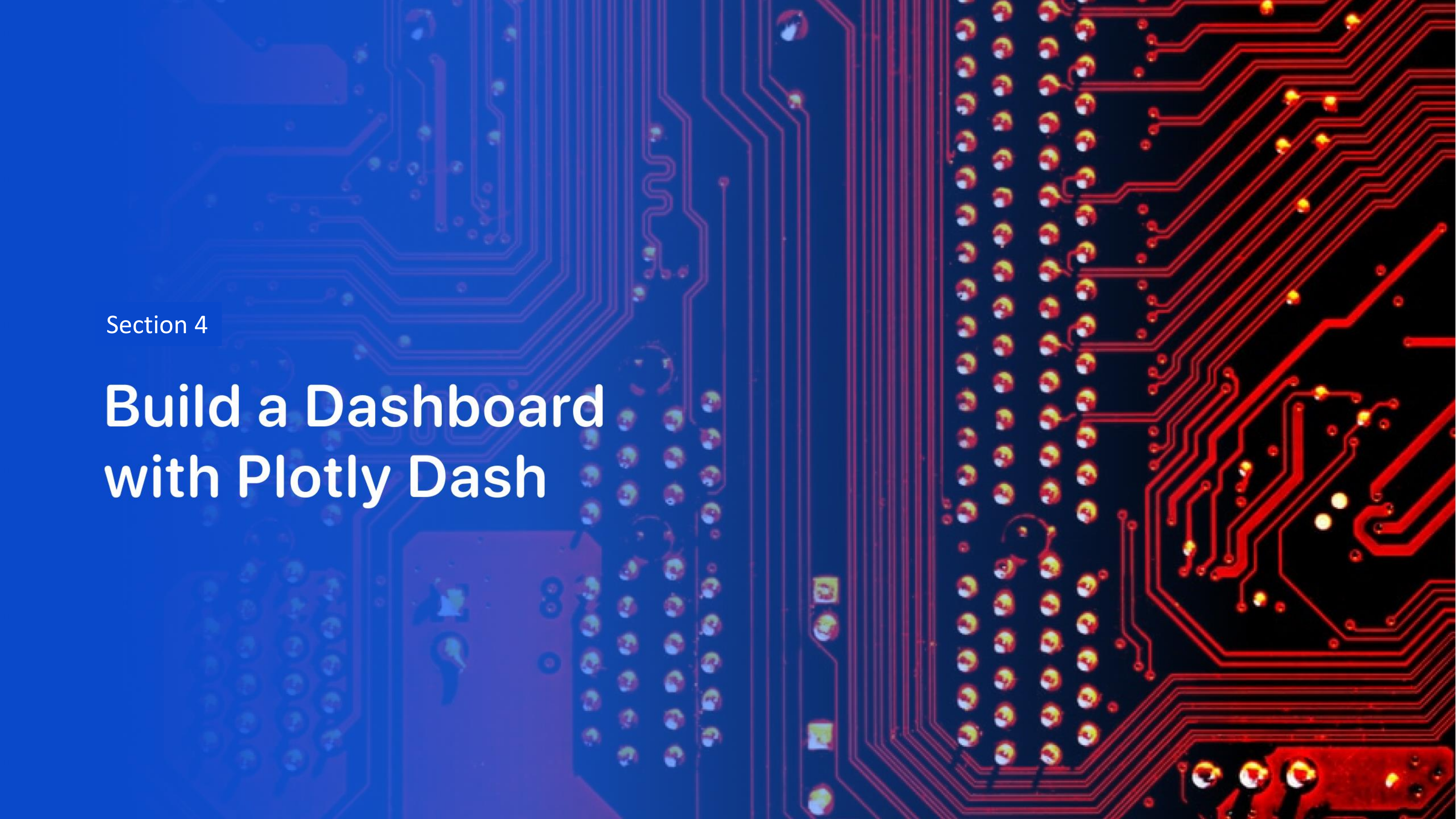


Location Proximities

- Image on left shows the distance from CCFAS SLC-40 to the railway at 0.98 km
- Image on the right shows the distance of CCFAS SLC-40 to Titusville, FL at 23.20 km
- Launch sites appear to be at moderate distances from areas with potential high population in order to decrease the risks associated with failed launches / landings







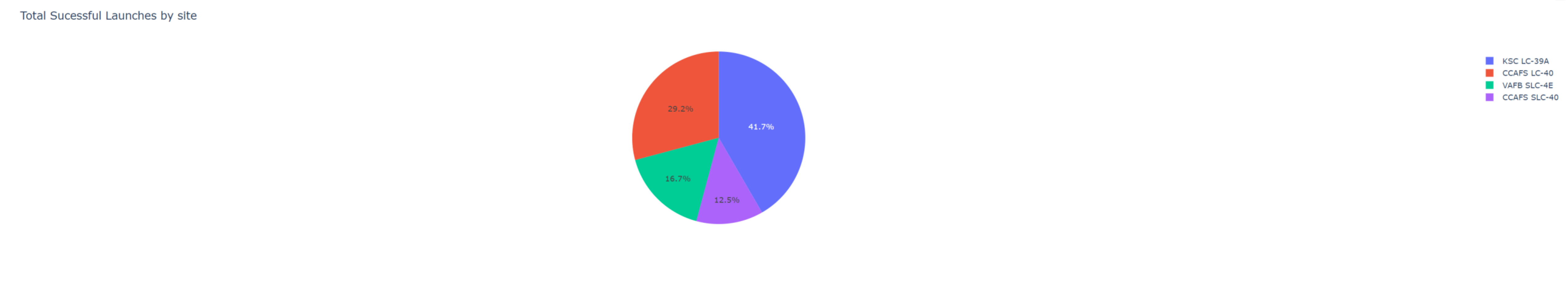
Section 4

Build a Dashboard with Plotly Dash

Successful Landings

Across All Sites:

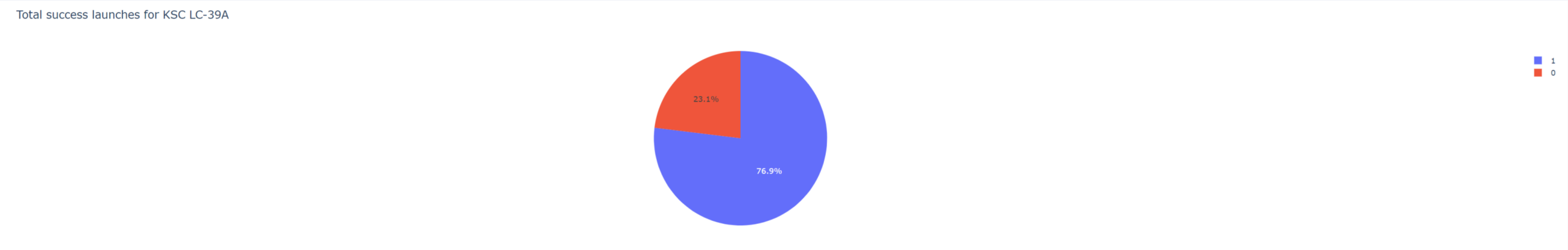
- KSC LC-39A demonstrated the highest rate of successful landings with 41.7%
- CCAFS SLC-40 produced the lowest success rate at 12.5%



Launch Site with Highest Launch Success Ratio

KSC LC-39A:

- At the KSC LC-39A launch site, 76.9% of the landings were successful; 23.1% were classified as a failure

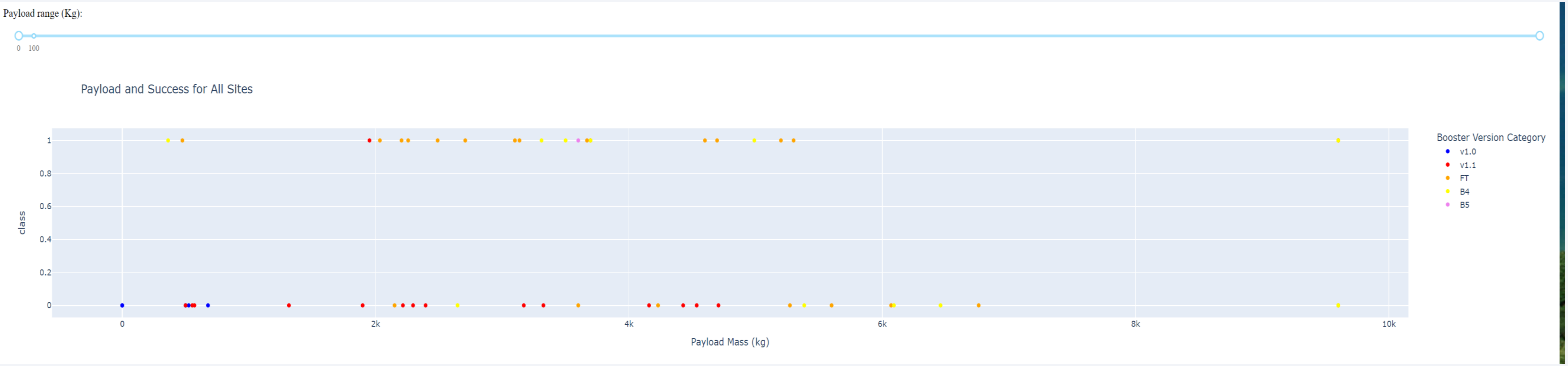


Payload vs. Launch Outcome

**class 1 = Successful landing / class 0 = Failed landing

Payload Range Selector:

- The FT booster represents the majority of the successful landings when carrying a payload mass between 0 – 7,500 (kg)
- The v1.1 booster appears to be associated with the highest rate of failed landings



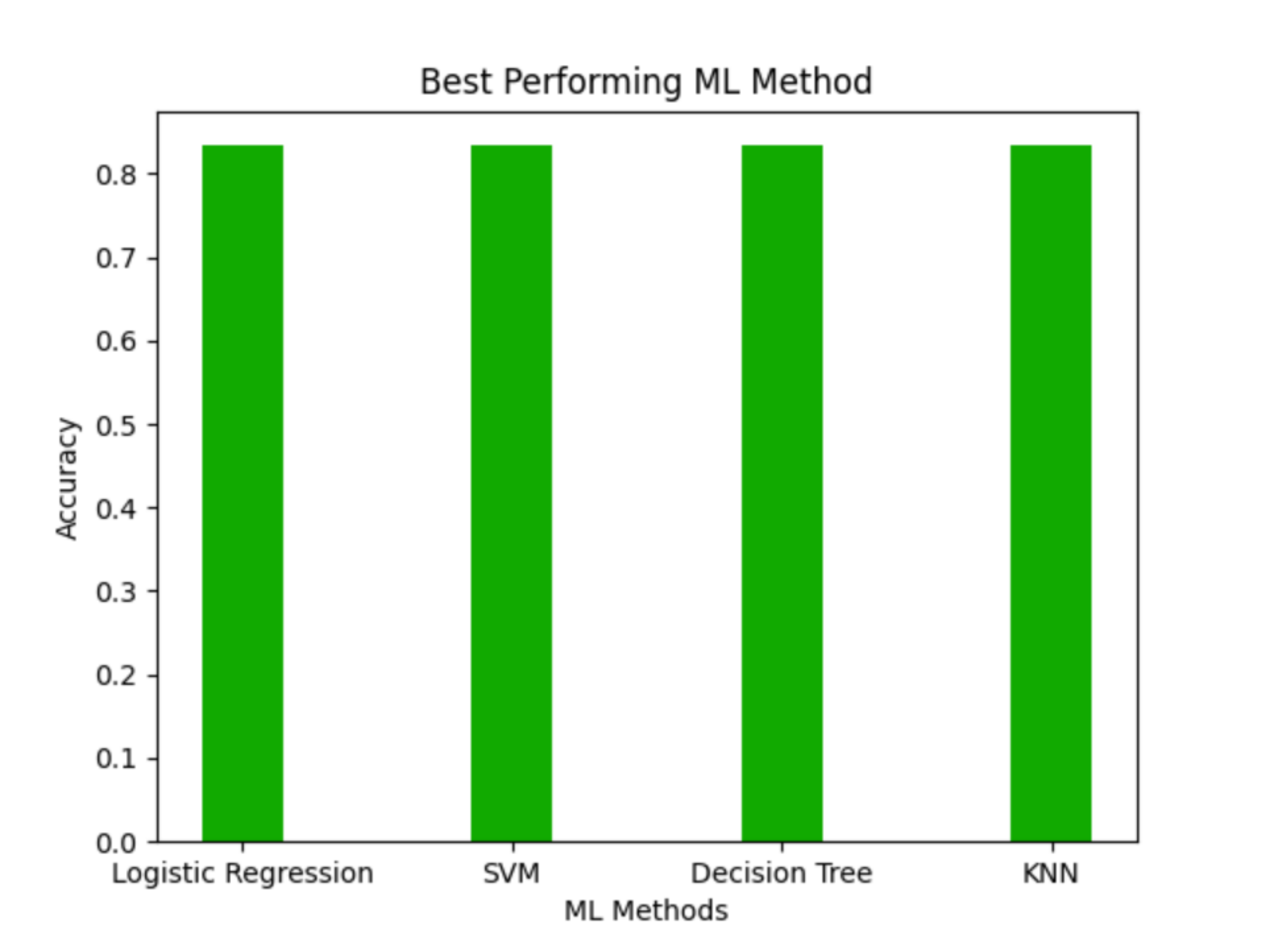
Payload range (Kg):

Section 5

Predictive Analysis (Classification)

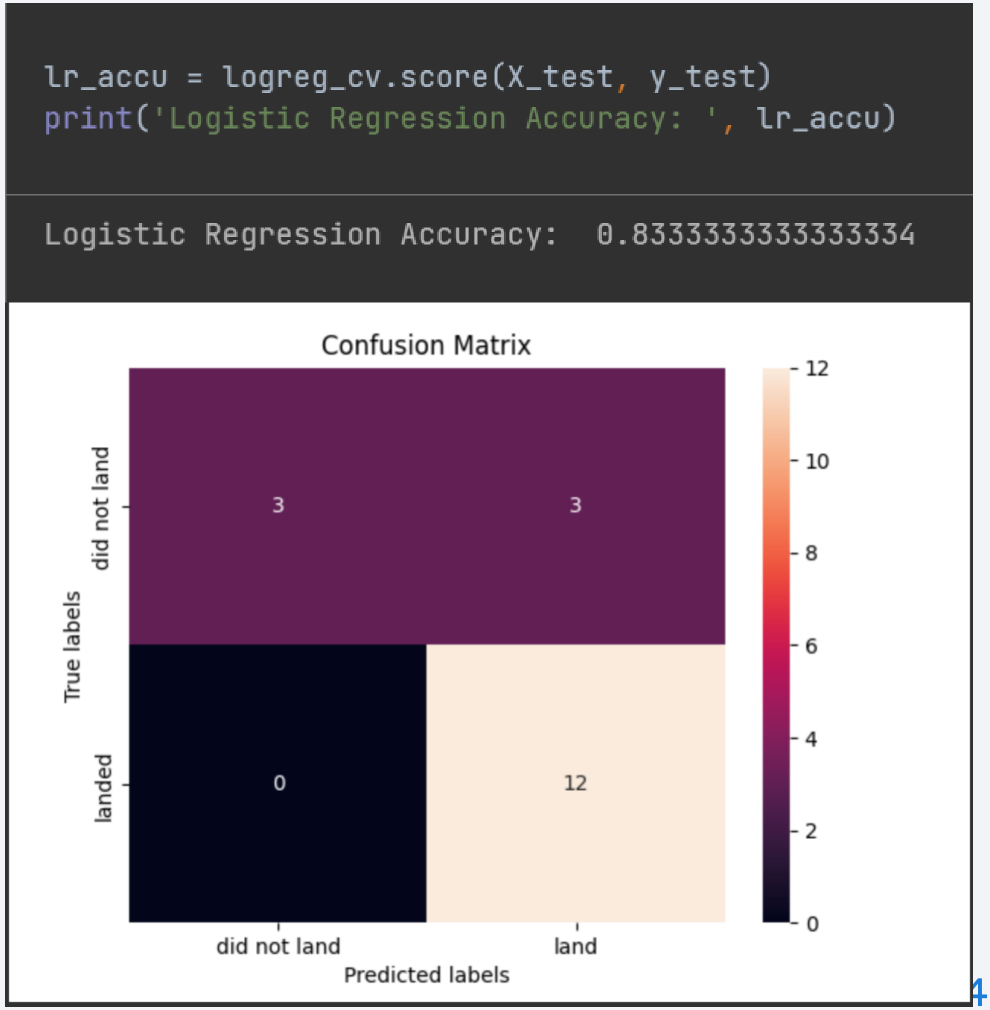
Classification Accuracy

- LR, SVM and KNN appeared to have the relatively same accuracy of 83.33% on the test set
- Decision Tree produced an accuracy rate of 84.82%



Confusion Matrix

- LR accuracy of 0.833, therefore distinguishing between different classes of success vs. failure
- The models predicted 3 successful and unsuccessful landings when the 'true label' = 'did not land'
- Successful landings were overpredicted



```
lr_accu = logreg_cv.score(X_test, y_test)
print('Logistic Regression Accuracy: ', lr_accu)
```

```
Logistic Regression Accuracy:  0.8333333333333334
```

Conclusions

- Optimal launch locations appear to be the KSC LC-39A and CCAFS LC-40 sites in Florida
- Though the CCAFS SLC-40 site has a low success rate, the number of launches is only 7 and may not be indicative of long-term success
- Payload mass appears to have an impact on landing success rate and should therefore be strongly considered when determining costs and potential success of landing
- Orbit type had an impact on success rate, which was not anticipated, and may warrant further investigation

Thank you!