Winning Space Race with Data Science

Gregory Siegfried
December 21st, 2021

Outline

Executive Summary

- Data collection, wrangling, and analysis have provided some insights on SpaceX. Features such as Launch Site, Payload Mass, and Orbit are associated with launch landing success. SpaceX has been increasing their landing success and are currently above 80%.

- Testing model analysis puts accuracy at 83.3% on our SpaceX data using Support Vector Machine with a C of 1.0, a gamma around 0.0316 and a sigmoid kernel.

Introduction

SpaceX, our biggest competitor, have been able to reuse their Falcon 9 first stage boosters, to allow an entry price of $62 million for orbital launches.

Our company, Space Y, need to figure out if a booster will be reused to influence contract bidding.

Our question:

Can we determine the price of a SpaceX launch, based on being able to reuse the first stage?

Section 1

Methodology

Methodology

Data collection methodology:

Using the SpaceX REST API, and Wikipedia HTML Web Scraping.

Perform data wrangling:

One Hot Encoding converts categorical to numerical data, missing numerical values were averaged where appropriate, while unnecessary data was removed.

Perform exploratory data analysis (EDA) using visualization and SQL:

Bar charts, scatter plots, box plots and queries foe analysis.

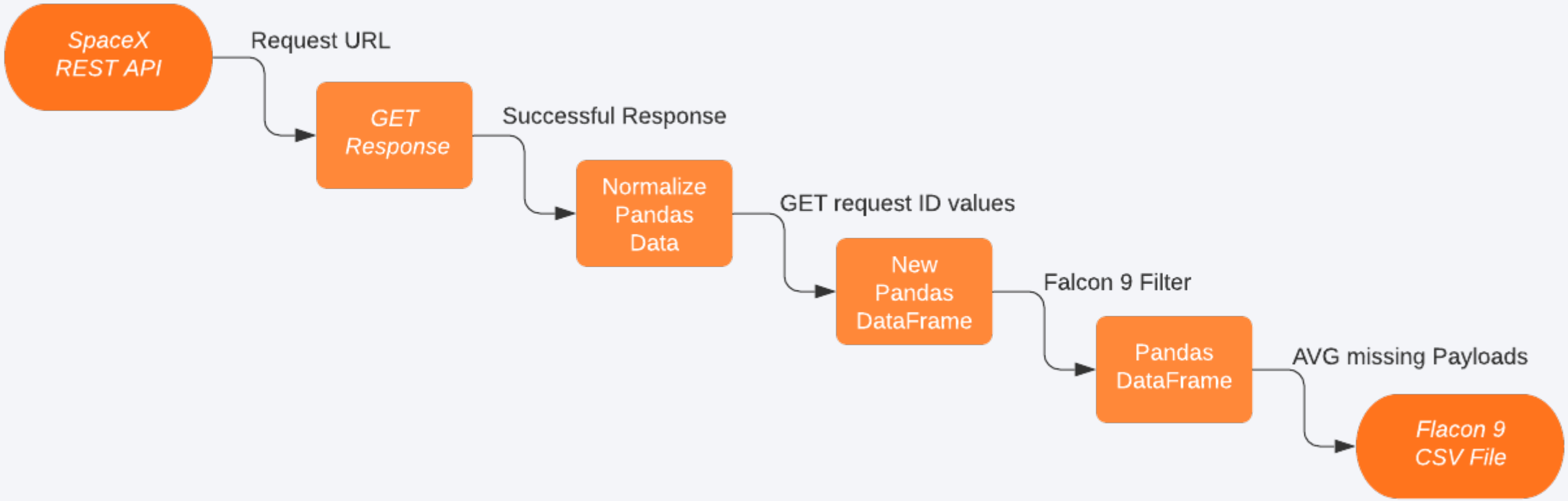
Perform interactive visual analytics using Folium and Plotly Dash:

Folium maps for geography insights and Plotly dashboards for interactive insights.

Perform predictive analysis using classification models:

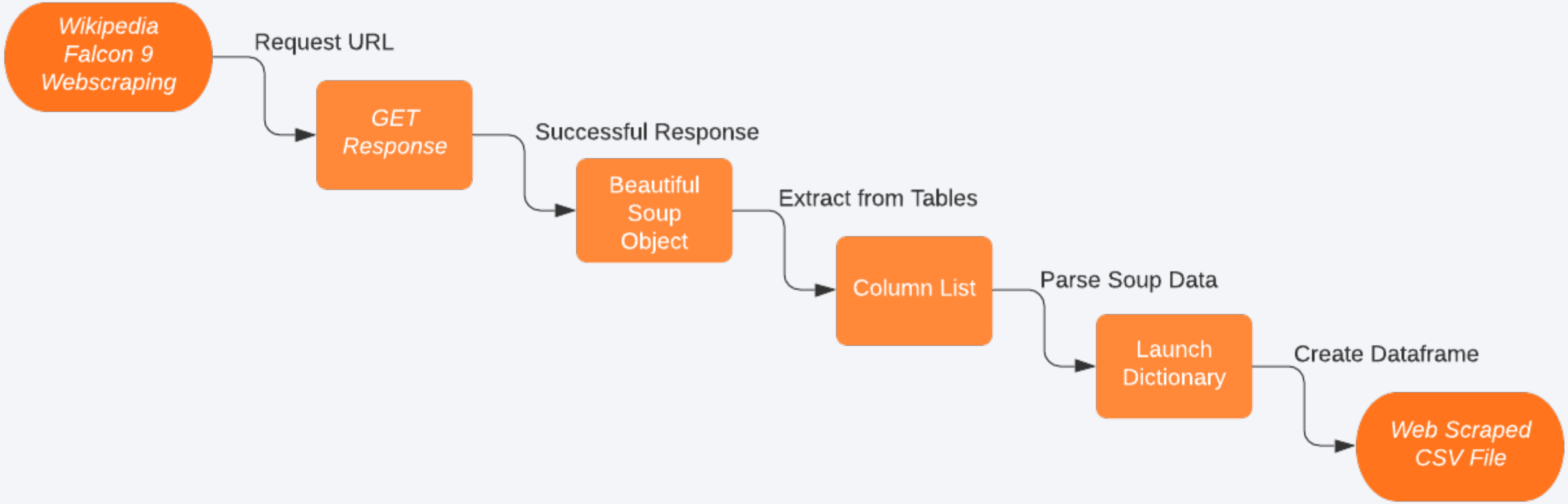
GridSearch for testing a variety of models and parameters for the best accuracy.

Data Collection – SpaceX API



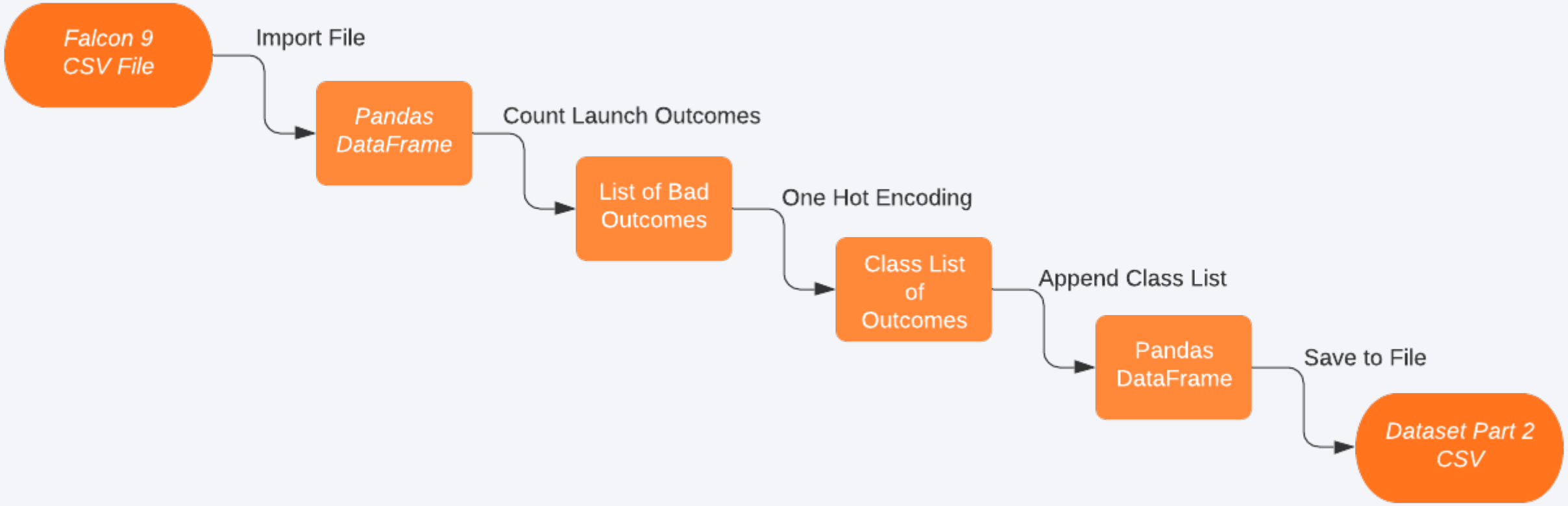
https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%201%20-%20SpaceX-API%20Calls.ipynb

Data Collection - Scraping



https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%202%20-%20SpaceX-Web%20Scraping.ipynb

Data Wrangling



https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%203%20-%20SpaceX-Data%20Wrangling.ipynb

EDA with Data Visualization

Using cat-plot, scatter plot, bar chart, and line plot, to visualize the data across Launch Site, Flight Number, Payload Mass, and Orbit.

All points being plotted are colored by a successful or unsuccessful launch

Examples: Slides 16-22

https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%204%20-%20EDA%20Data%20Visualization.ipynb

EDA with SQL

Selected Items to List:

- Distinct Launch Sites
- Total or Average Payload Across Features
- Booster Versions on Payload Ranges
- Landing Outcomes for Dates

Examples: Slides 23-32

https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%205%20-%20EDA%20SQL%20Queries.ipynb

Build an Interactive Map with Folium

Markers and Circles were added to aid in locating launch sites and zooming levels.

Marker Clusters were added to aid in visualizing launch outcomes for each site.

Feature Lines and Distances were added to show proximity to Cities, Rail Roads, Highways, and Coastlines.

Examples: Slides 33-36

https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%206%20-%20Folium%20Map.ipynb

Build a Dashboard with Plotly Dash

Selecting one or all launch sites you can:

1) see the launch outcomes as a percentage in a pie chart.

2) see the launch outcomes per booster across all payload ranges, using a slider to narrow down values in a scatter plot.

Examples: Slides 37-40

https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%207%20-%20SpaceX%20Dash%20App.py

Predictive Analysis (Classification)

X features uses a data set where all values are numeric, or one hot encoded. Y features is the launch outcome column ‘Class’.

The X and Y features were used to train-test-split the data sets.

GridSearch was used to find the best parameters for K Nearest Neighbors, Logistic Regression, Decision Tree, and Support Vector Machine.

Examples: Slides 41-43

https://github.com/siegfriedgreg/IBM-Data-Science-Coursera/blob/main/Capstone%20Data_Science/Lab%20-%208%20-%20SpaceX%20Predictive%20Analysis.ipynb

Results

Exploratory Data Analysis:

Flight success has increased over the years, with the most recent (2020) being a slight decrease. Payloads over 8000 kg are successful, and those in certain orbits are also. Data discrepancies exists for API data and SQL data, the former has 3 launch sites and the latter has 4.

Interactive Data Analysis:

Certain launch sites show to be more successful (KSC LC-39A). Booster versions don’t show to be more successful than others.

Predictive Data Analysis:

Both Logistic Regression and Support Vector Machines obtained a score of 83.3%. Their respective training accuracies were, 82.1% and 84.8%, and they both performed the same with respect to their confusion matrices.

Section 2

Insights drawn from EDA

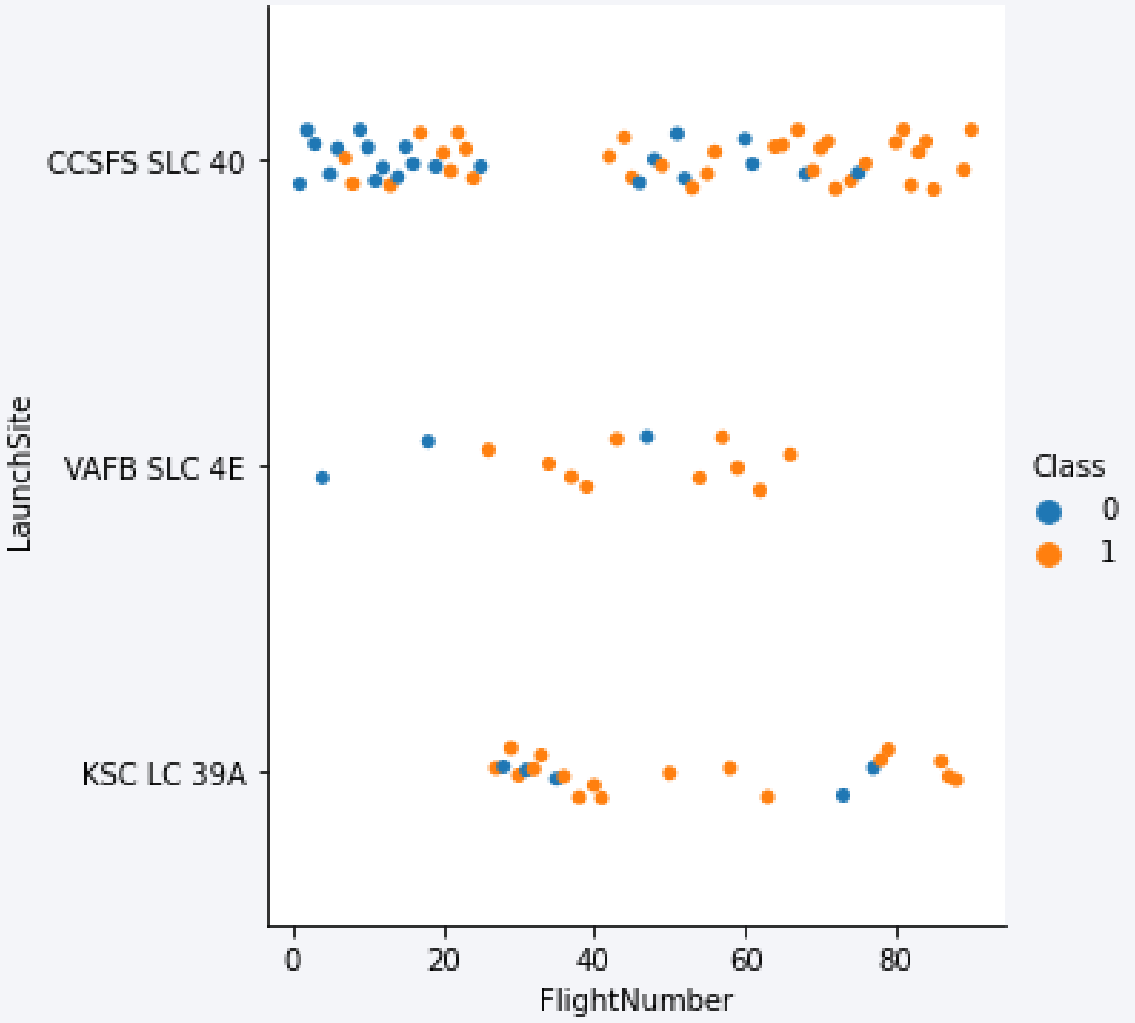
Flight Number vs. Launch Site

Class coloring indicates launch outcome.

0 is unsuccessful

1 is successful

Notice which sites are most active.



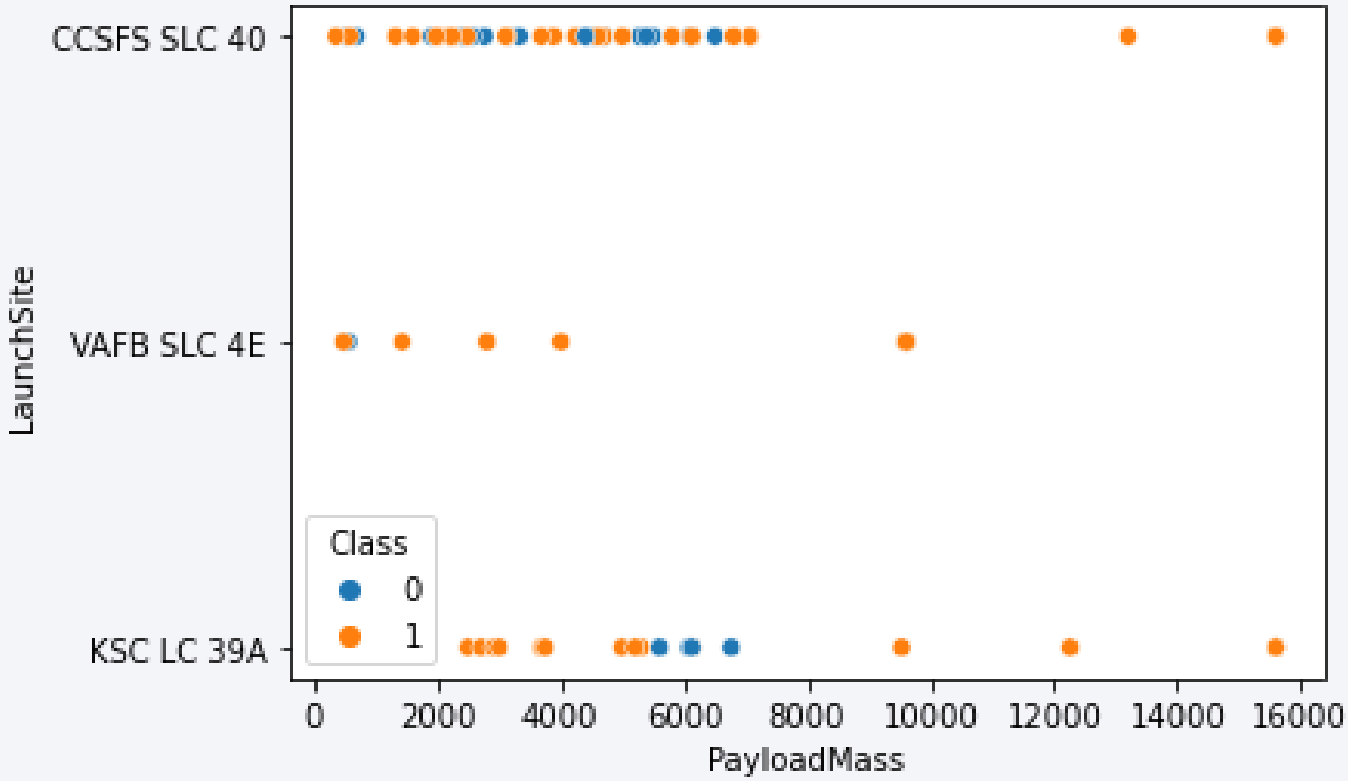
Payload Mass vs. Launch Site

Class coloring indicates launch outcome.

0 is unsuccessful

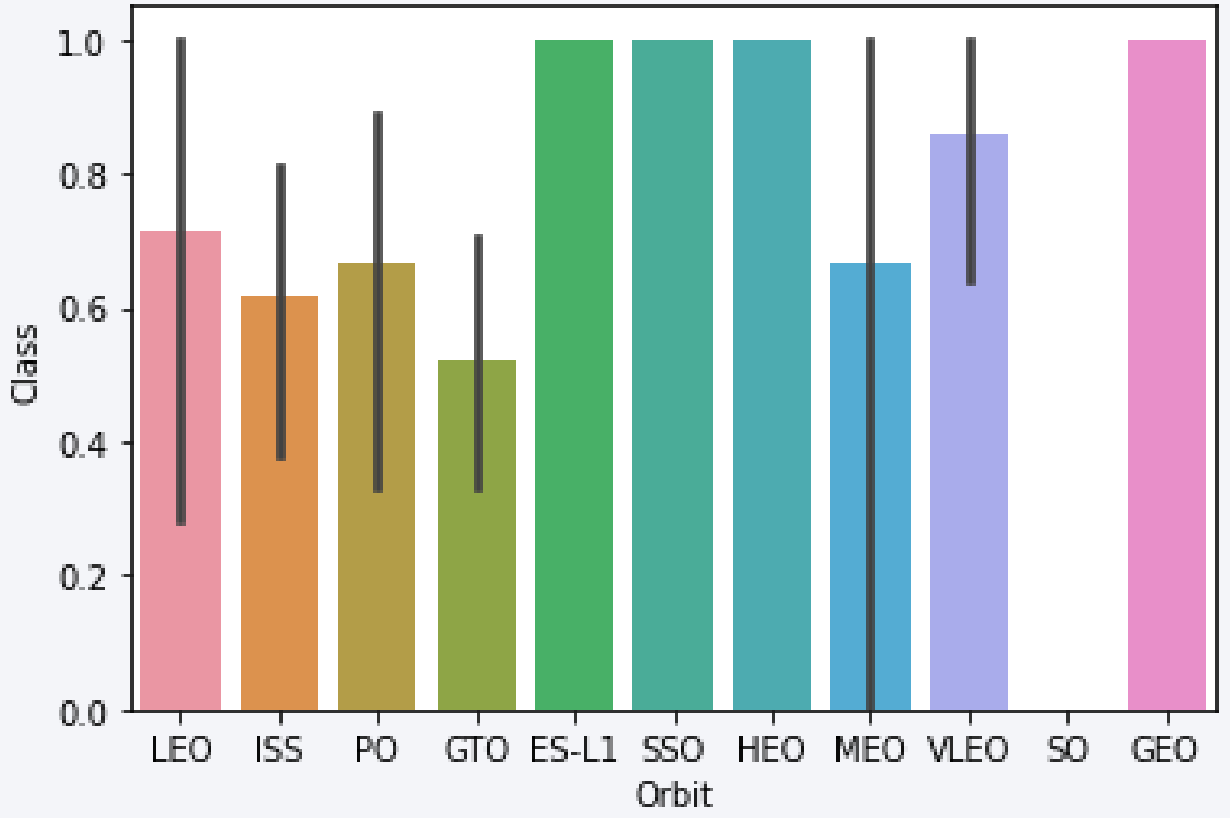
1 is successful

Notice which payload masses are most used at a site.



Orbit Type vs. Success Rate

- Notice which orbit destinations are the more successful, and ones that have the most variability.



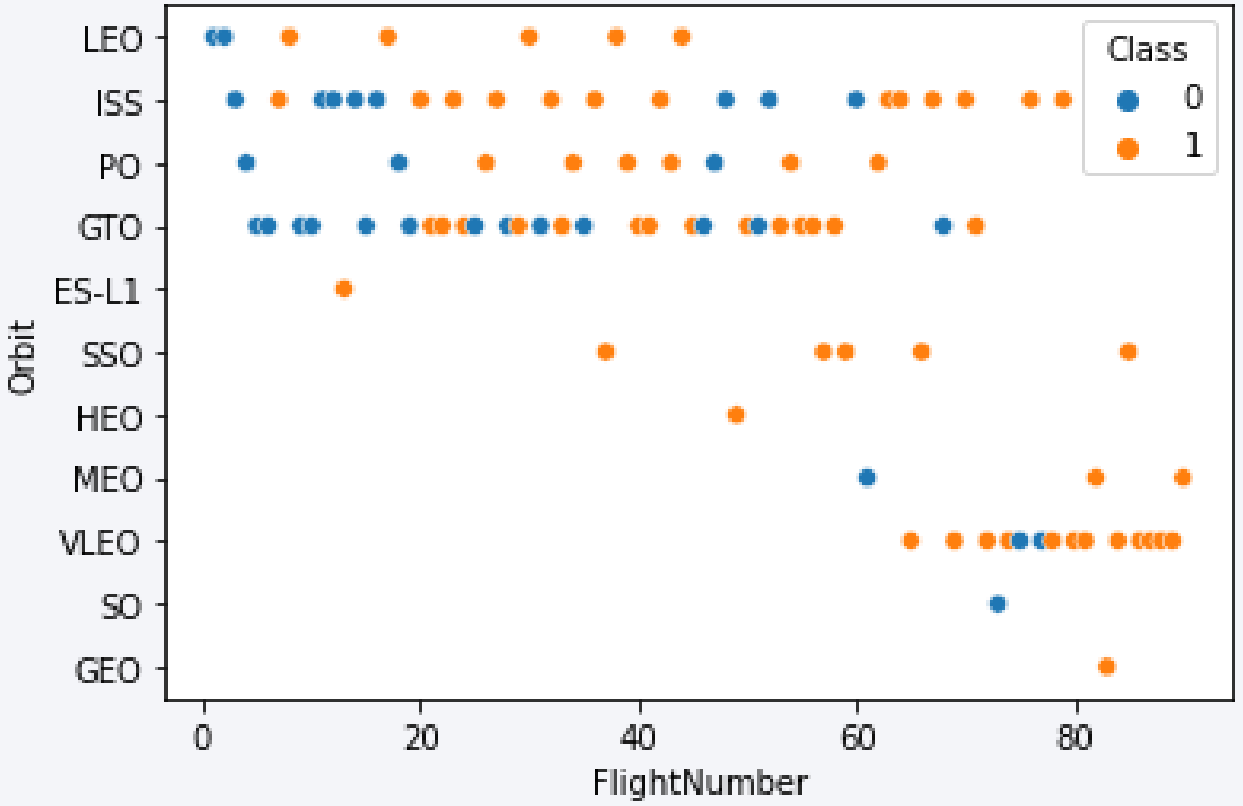
Flight Number vs. Orbit Type

Coloring indicates the launch outcome.

0 is unsuccessful

1 is successful

Notice, that few orbit destinations are without a successful landing.



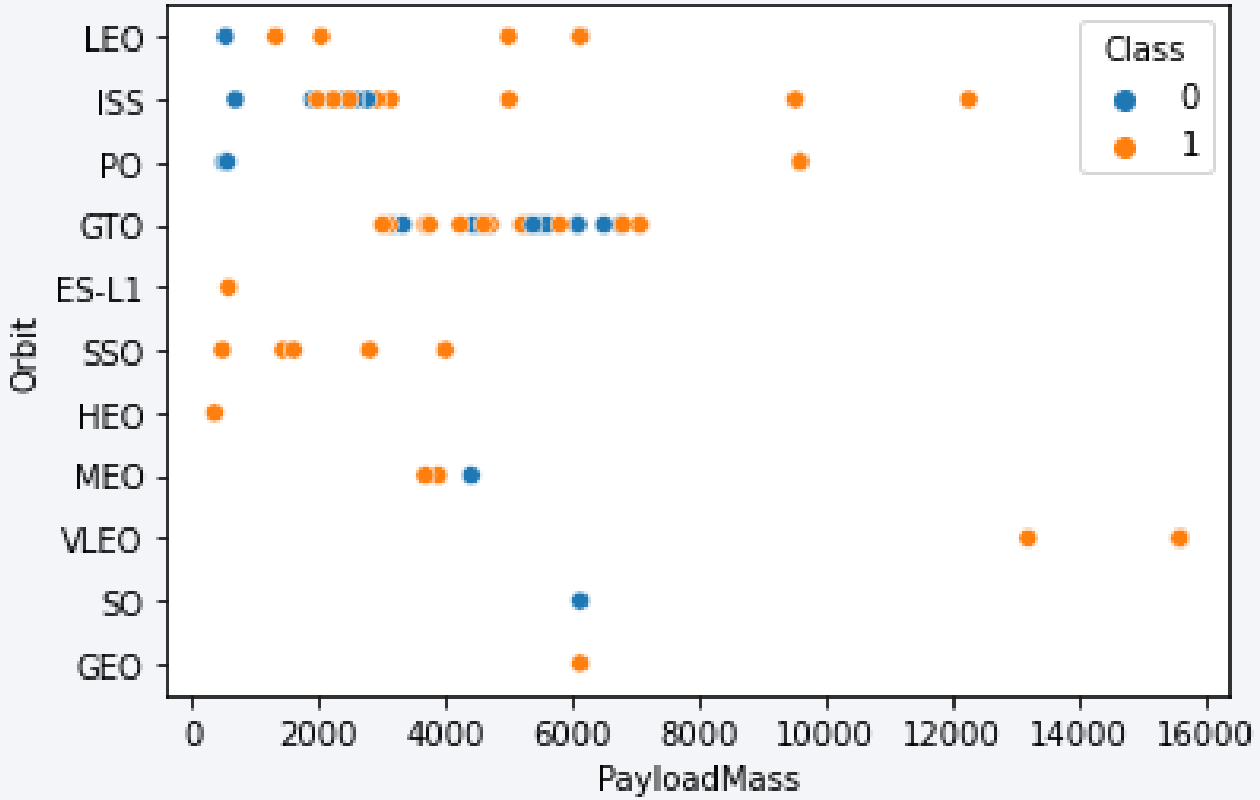
Payload Mass vs. Orbit Type

Coloring indicates landing outcomes.

0 is unsuccessful

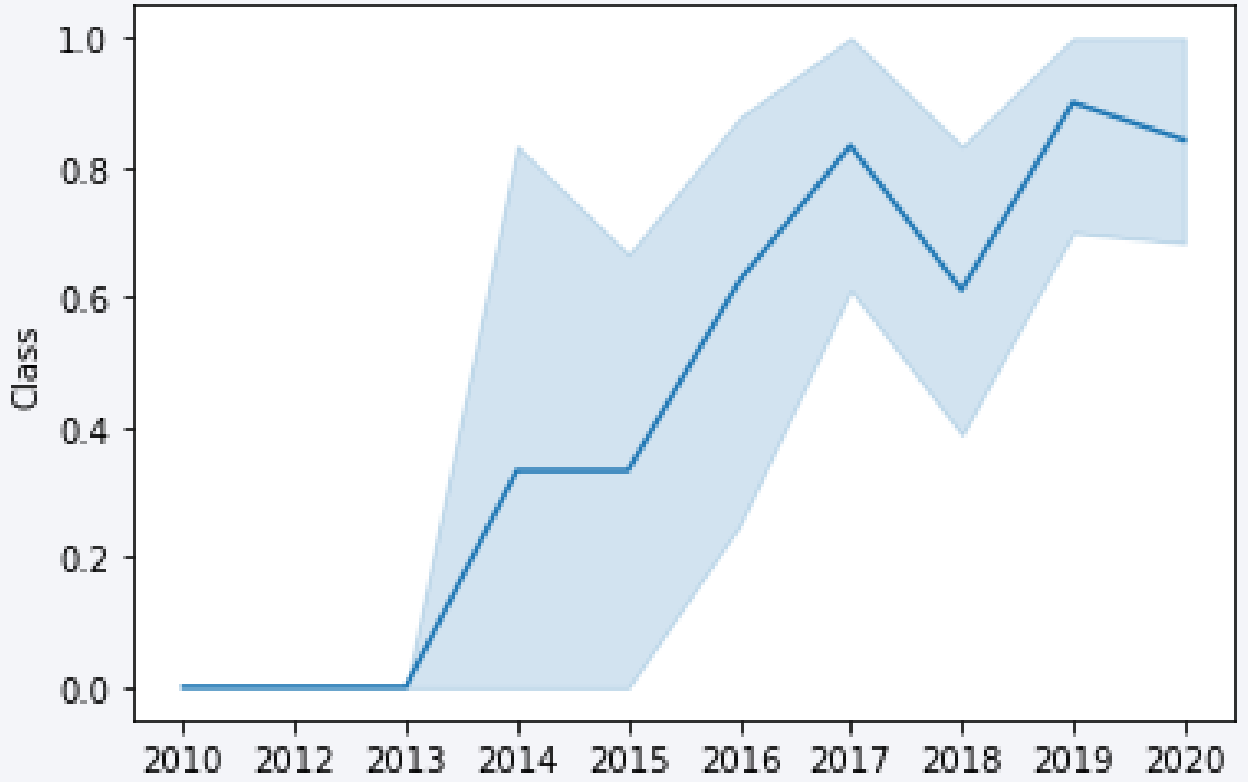
1 is successful

Notice the payload ranges for particular orbit destinations.



Launch Success Yearly Trend

Notice the trend from 2013 to 2019 increased in success. 2020 looks to be on a down trend.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%sql select distinct(launch_site) from spacextbl;
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

Task 2

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

```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5;
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Task 3

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

```
%sql SELECT SUM(payload_mass__kg_) FROM spacextbl WHERE customer LIKE 'NASA (CRS)';
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE booster_version LIKE 'F9 v1.1%';
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
2534

First Successful Ground Landing Date

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT DATE FROM spacextbl WHERE landing__outcome LIKE 'Success (ground pad)' ORDER BY DATE ASC LIMIT 1;
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%sql SELECT booster_version FROM spacextbl WHERE landing__outcome LIKE 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
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

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome AS Outcome, COUNT(mission_outcome) AS Number FROM spacextbl GROUP BY mission_outcome;
```

Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

outcome	number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl);
```

Python

```
* ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
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

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql SELECT landing__outcome AS Outcome, booster_version AS Version, launch_site AS Site FROM spacextbl
WHERE landing__outcome LIKE 'Failure (drone ship)' AND YEAR(DATE)=2015;
```

[12] ✓ 0.4s Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

outcome	VERSION	site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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(*) outcome_freq FROM spacextbl WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome LIKE 'Success%'
GROUP BY landing__outcome ORDER BY outcome_freq DESC;
```

[13] ✓ 0.8s Python

```
 * ibm_db_sa://cfj09142:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

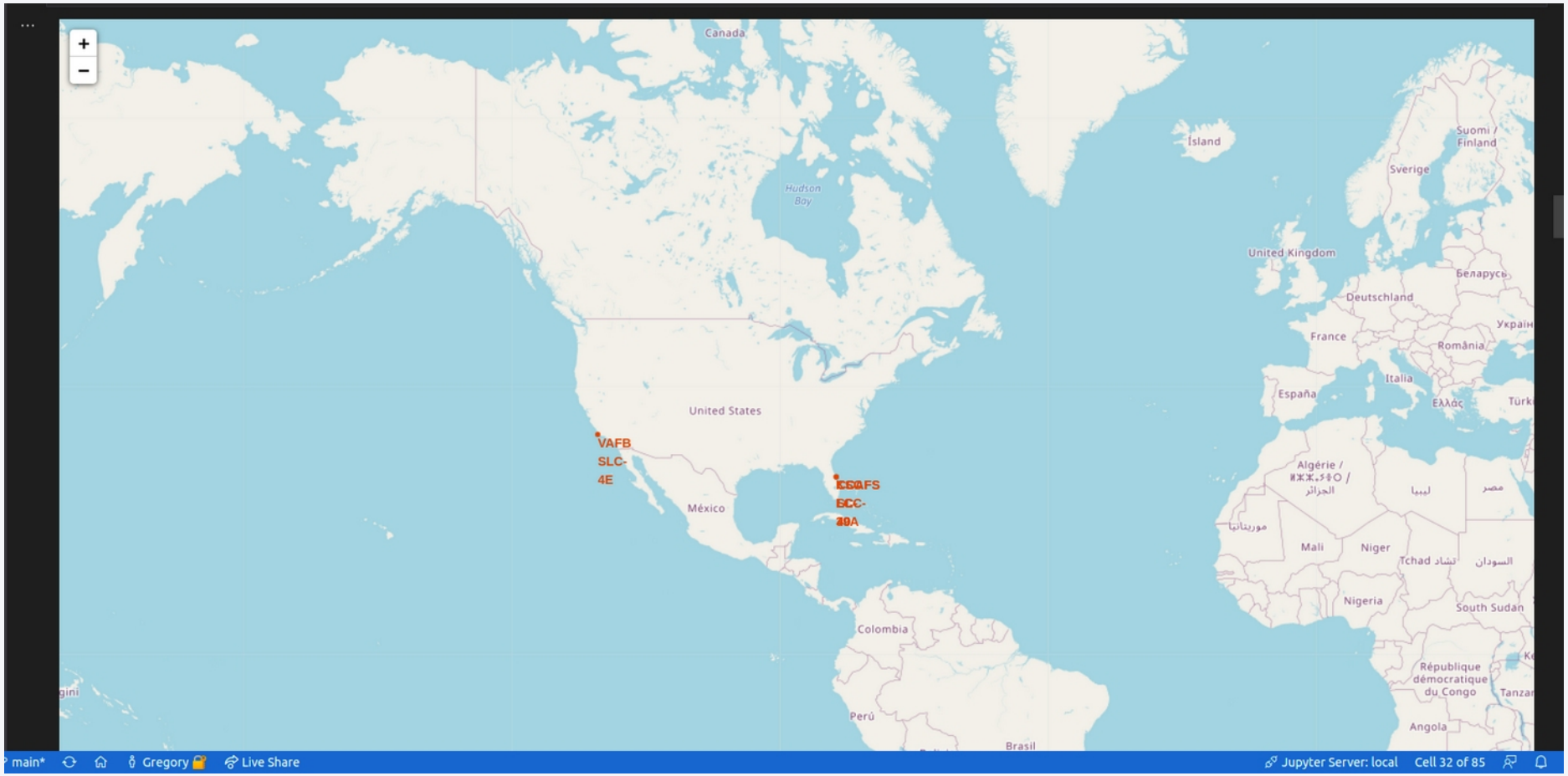
landing__outcome	outcome_freq
Success (drone ship)	5
Success (ground pad)	3

Section 4

Launch Sites Proximities Analysis

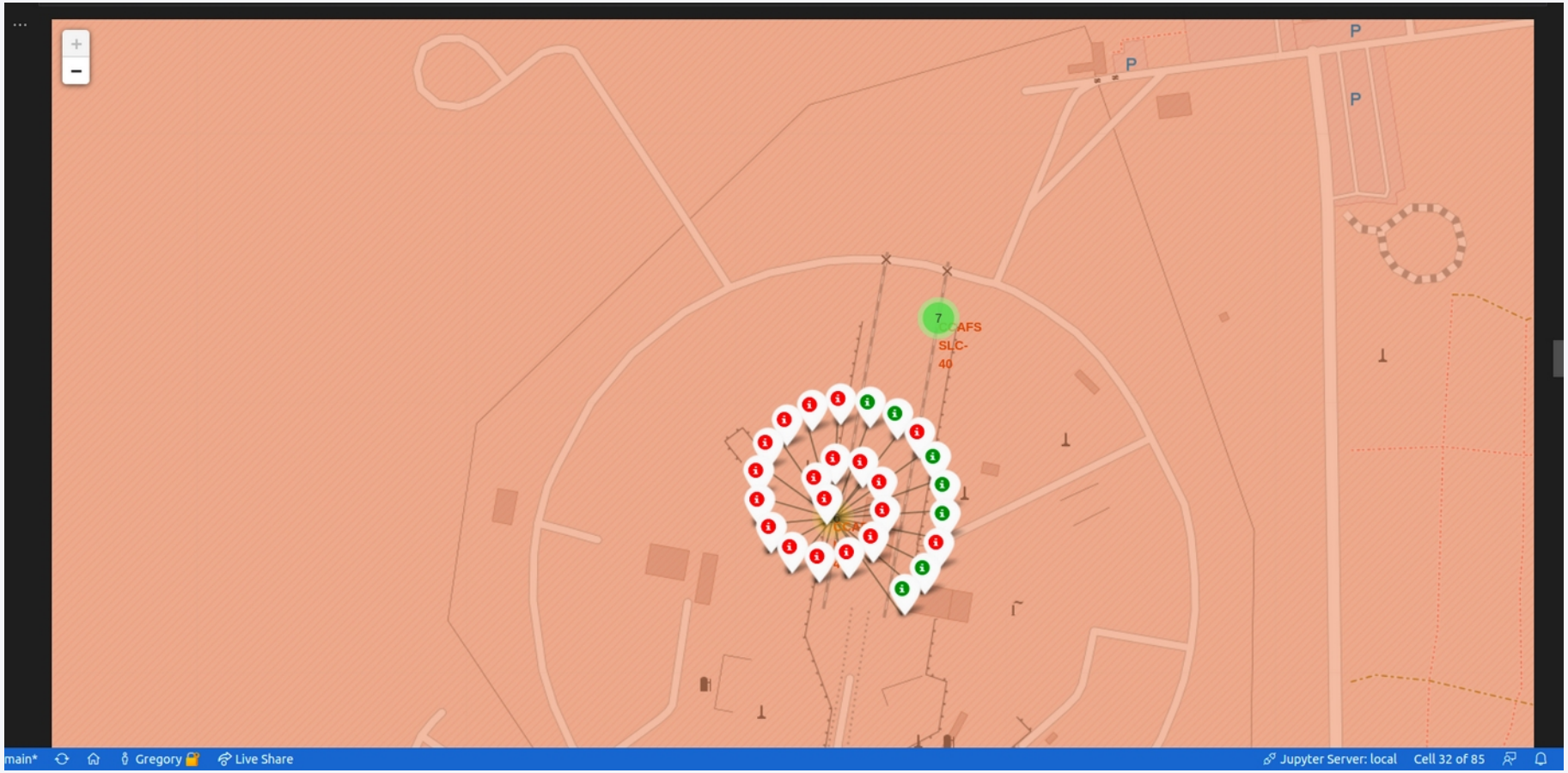
Folium Map - Launch Sites

Global screen shot of folium map with launch site locations.



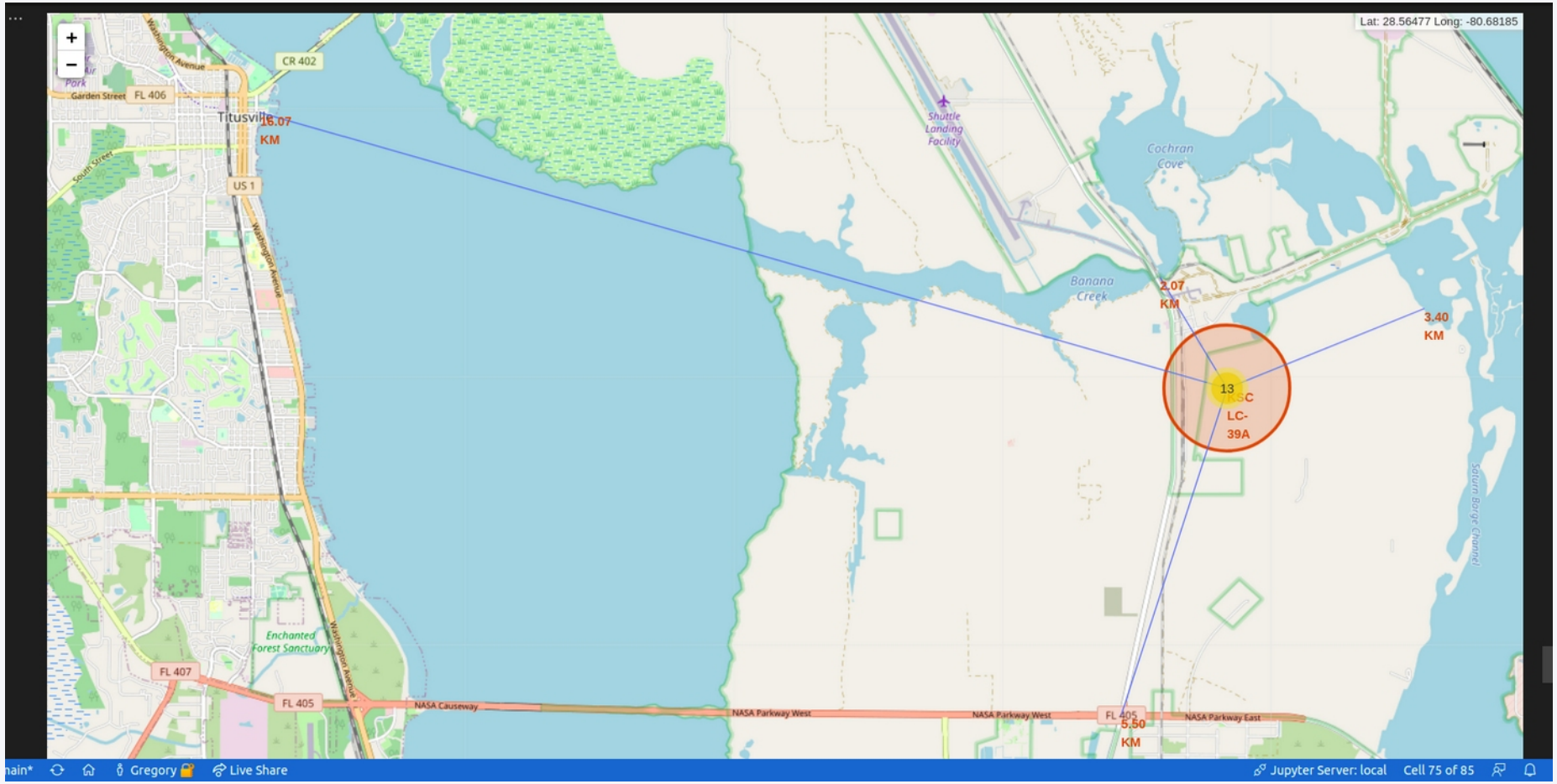
Folium Map – Launch Outcome Markers

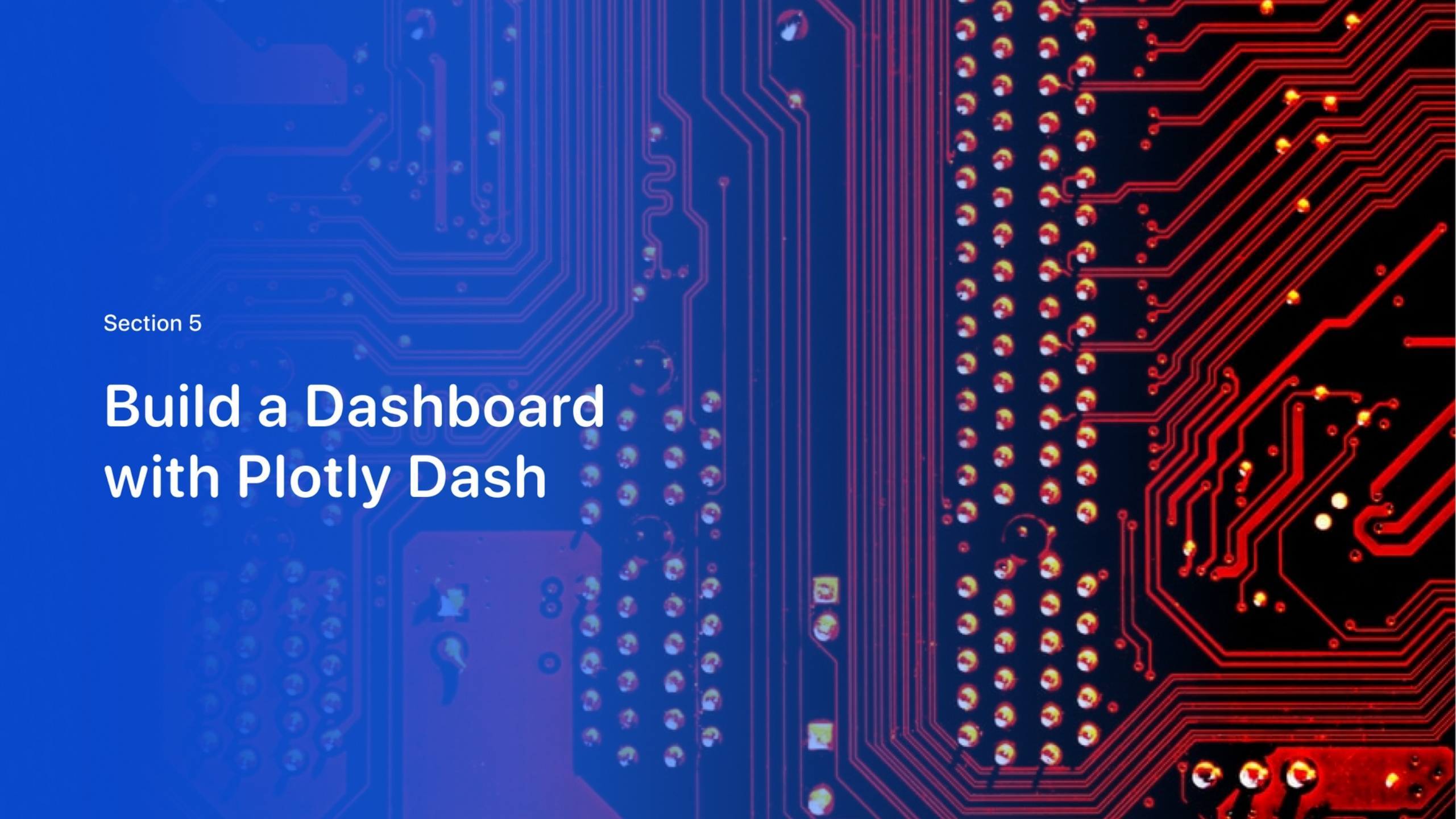
Launch outcome markers for site: CCAFS LC-40



Folium Map – Local Feature Markers

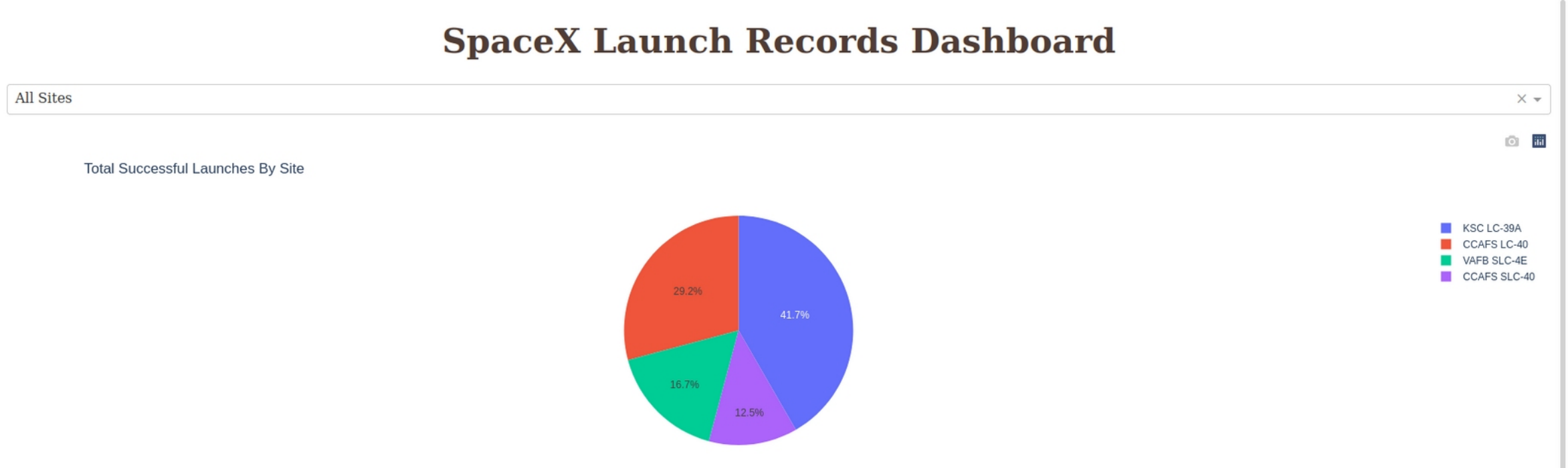
Locations of features around site: KSC LC-39A




Section 5
Build a Dashboard with Plotly Dash

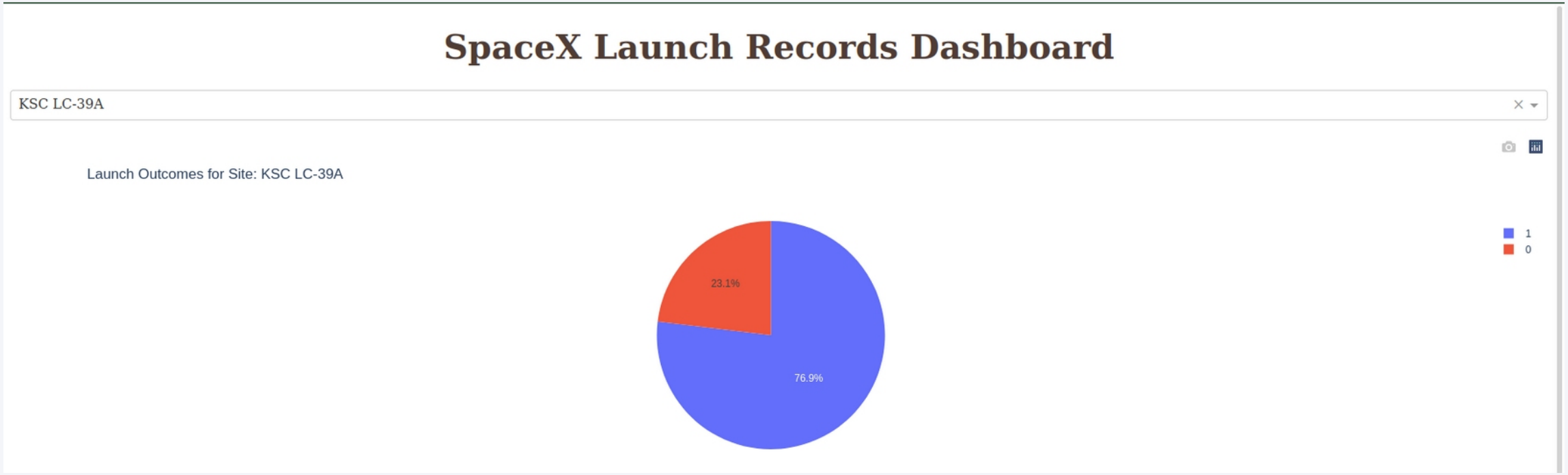
Dash – Successful Outcome by Site

Successful landings are shown across all launch sites



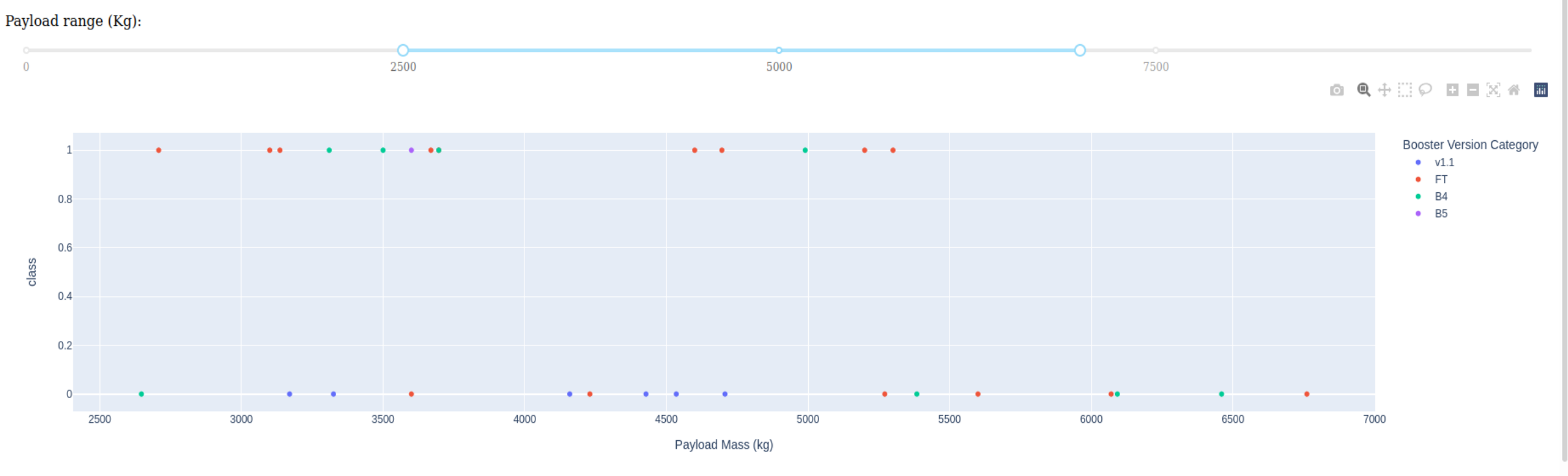
Dash – Most Successful Site

Successful landings are coded 1, and Unsuccessful as 0



Dash – Payload Mass vs. Launch Outcome

All sites with payload range between 2500 to 7000 kg, booster version by color.



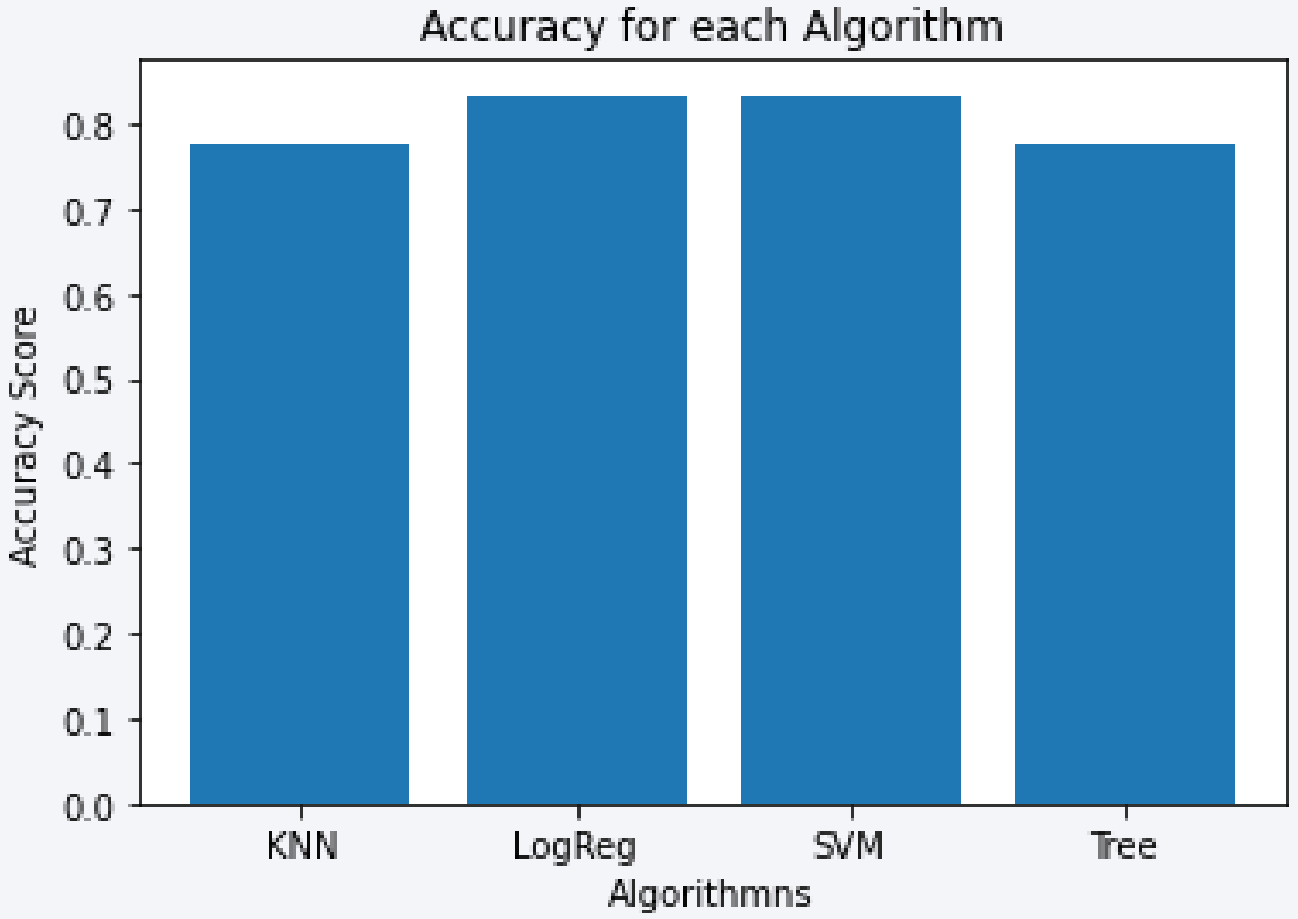
Section 6

Predictive Analysis (Classification)

Classification Accuracy

The two algorithms that both performed the same on testing data were Logistic Regression (LogReg) and Support Vector Machine (SVM).

When taking training accuracy into account, SVM performed better than LogReg at 84.8% to 82.1% respectively.



Confusion Matrix

Both Support Vector Machine and Logistic Regression had the same outcome with respect to their Confusion Matrix.

The results show that each algorithm had three false positives, where it predicted a successful landing and that was not the case.

Conclusions

- Support Vector Machine model with the best parameters: { 'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid' }, had a training accuracy of 84.8% and testing accuracy of 83.3%
- Landing success increases with a payload mass greater than 8000 kg.
- Landing success has increased from 2013 to 2019, and 2020 is the most recent decrease which currently sits above 80%.
- Landing success increases when launched from certain sites like KSC LC-39A.
- Landing success increases with certain orbit destinations.

Thank you!