IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lukah Connolly Sams
24/04/2025

Outline

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

Executive Summary

Summary of methodologies:

- Import and clean dataset using REST API's, SQL, and Pandas
- Use Seaborn to conduct EDA
- Visualize data and trends interactively with Folium, Plotly, and Dash
- Determine best ML model for data

Summary of all results:

- Showed that over the years, the success rate of landings has increased
- The Tree Classification model was the best model for the dataset

Introduction

Project background and context:

- Predicting factors that affect landing outcomes, of Falcon 9 rockets, can be used to improve success rates and save money.

What we want to know (examples):

- Which launch site is the most successful?
- Is the yearly success rate increasing?
- How well can we predict if a launch will succeed?

Section 1

Methodology

Methodology

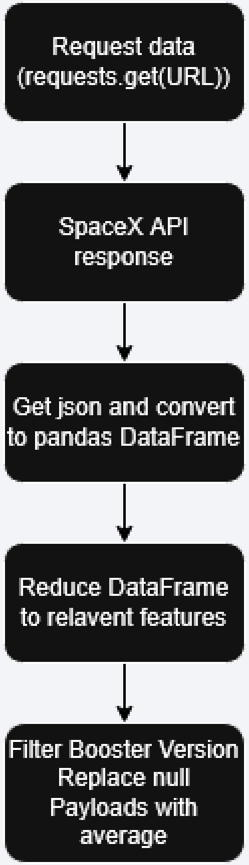
Executive Summary

- Data collection methodology:
  - Data was collected through web scraping and the SpaceX API
- Perform data wrangling
  - Null values were cleaned, only falcon 9 rocket record were kept in data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

- Requested data from SpaceX REST API and get Json object, then convert data into pandas DataFrame whilst filtering and cleaning columns.

- GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb



Request data (requests.get(URL))

SpaceX API response

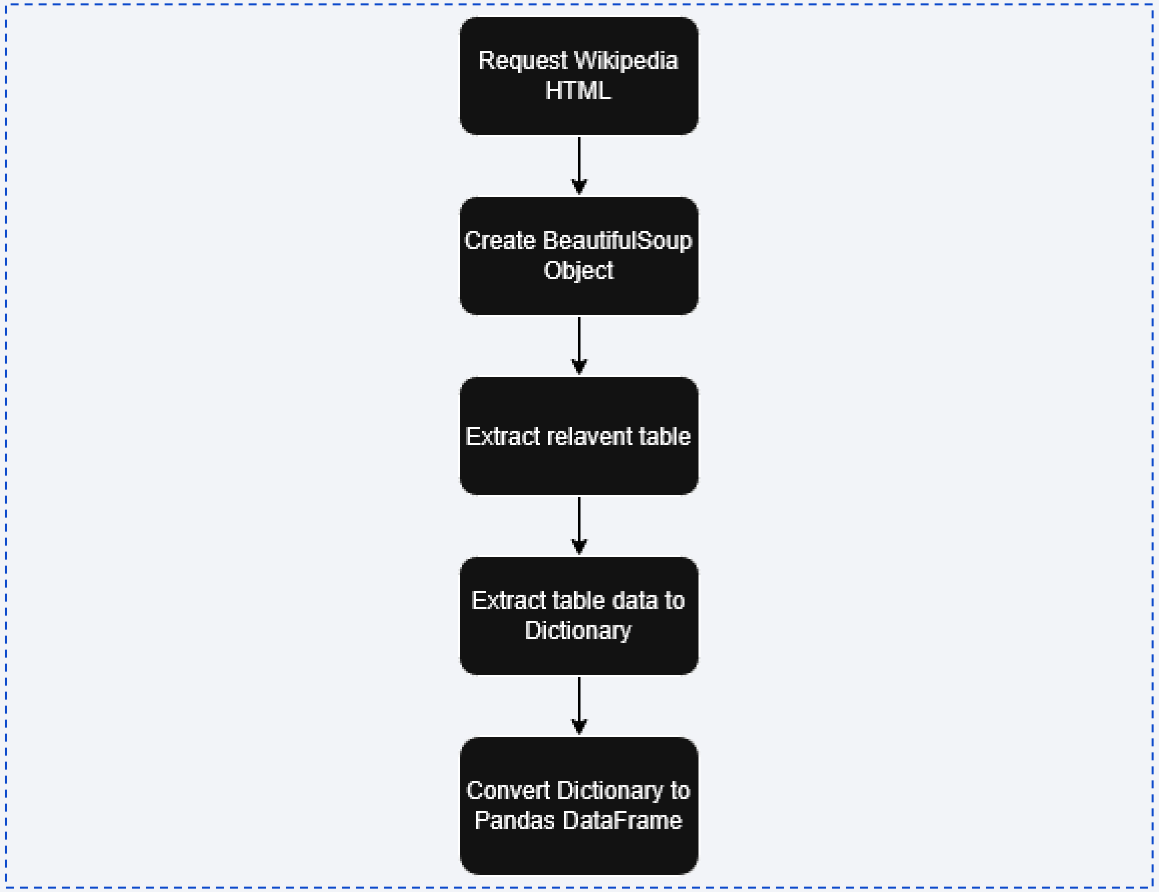
Get json and convert to pandas DataFrame

Reduce DataFrame to relavent features

Filter Booster Version Replace null Payloads with average

Data Collection - Scraping

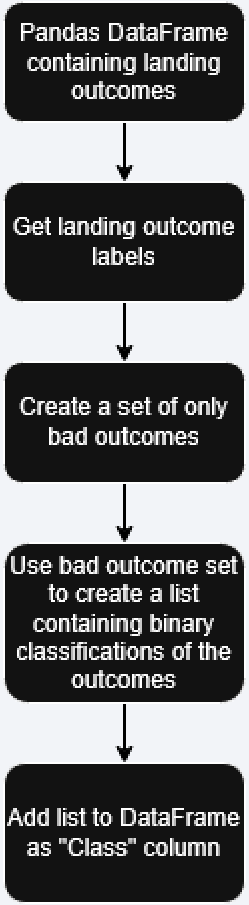
- Request Wikipedia HTML, use BeautifulSoup4 to parse HTML, and extract relevant table data to dictionary, then convert to DataFrame.
- GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/jupyter-labs-webscraping.ipynb


Request Wikipedia HTML
Create BeautifulSoup Object
Extract relavent table
Extract table data to Dictionary
Convert Dictionary to Pandas DataFrame

Data Wrangling

- Data was processed by creating a new column (binary classification column) for landing outcome success.

- GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



Pandas DataFrame containing landing outcomes

Get landing outcome labels

Create a set of only bad outcomes

Use bad outcome set to create a list containing binary classifications of the outcomes

Add list to DataFrame as "Class" column

EDA with Data Visualization

Charts:

- Payload Mass vs Flight Number (with Class hue) – Scatter, payload trends in newer missions (flight number).
- Launch Site vs Flight Number (with Class hue) – Scatter, favoring/Succes of different sites.
- Launch Site vs Payload Mass (with Class hue) – Scatter, payload bias/success in different sites.
- Class vs Orbit Type – Bar, success rate of different mission orbits.
- Orbit Type vs Flight Number (with Class hue) - Scatter, orbit type and success trend over with newer missions (flight number).
- Orbit vs Payload Mass (with Class hue) – Scatter, tendencies of payload mass with respect to orbits.
- Overall Success Rate vs Year – Line, success trend over time (tracks improvement of mission success).

GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

SQL queries:

- Get list of launch sites (unique values)
- Display first 5 records from sites with “CCA”
- Total Payload Mass launched by “NASA (CRS)”
- Average Payload Mass from F9 v1.1 rockets
- First Successful Ground Pad Landing
- List of Boosters that successfully landed with a “drone ship” (unique values)
- Count of each type of mission outcome
- List of Boosters that carried the maximum payload amount
- Display Month, Outcome, Booster Version, Launch Site for Failed Drone Ship landings in 2015
- Rank by count the types of Landing Outcomes recorded

GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added Circle markers for NASA and JSC
- Added Circle markers for each launch site
- Added marker clusters for failed and successful landings, per launch site
- Added lines to nearest points of interest from one launch site

Doing this provides helpful visual geographical context to our data and could help discover important features that might affect the outcome.

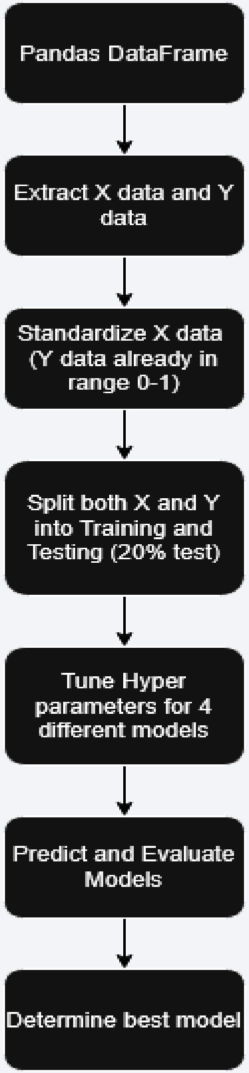
GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- Pie charts for Percentages of total launches between sites, and percentage of successes/failures in a specific site.
- Scatter plots for an adjustable payload mass range (can be filtered per launch site).
- Useful for quickly visualizing subsets of data for different scenarios.
- GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/spacex_dash_app.py

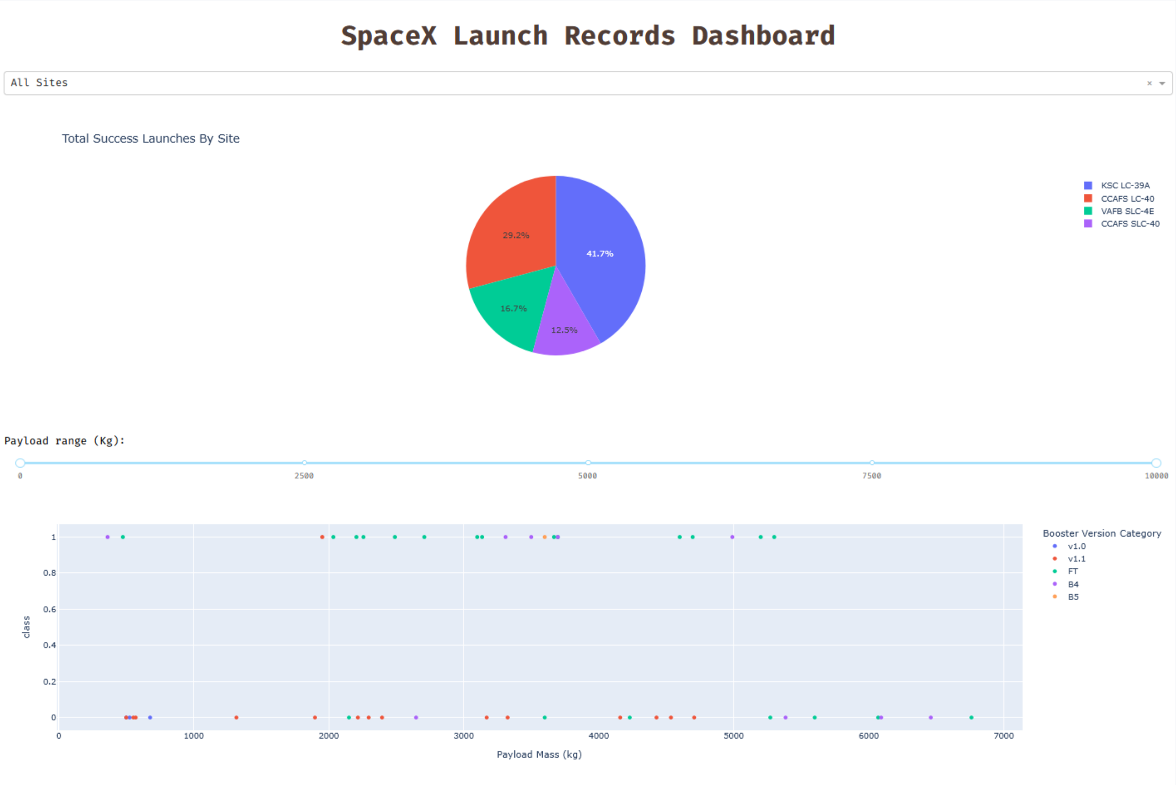
Predictive Analysis (Classification)

- Standardized the different columns of the X data.
- Split the data into training and testing (testing at 20%)
- Built 4 models, and tuned their hyperparameters using GridSearchCV():
  - Logistic Regression
  - Support Vector Machine
  - Tree classification
  - K Nearest Neighbors
- Evaluated models using the best hyperparameter scores, confusion matrices, and scores from the testing data
- Found best model using the model with the highest accuracy
- GitHub URL: https://github.com/LukahConnollySams/Capstone/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb


Pandas DataFrame
Extract X data and Y data
Standardize X data (Y data already in range 0-1)
Split both X and Y into Training and Testing (20% test)
Tune Hyper parameters for 4 different models
Predict and Evaluate Models
Determine best model

Results

- Exploratory data analysis results:
  - _____
- Predictive analysis results:
  - Best model was the Tree Classification model

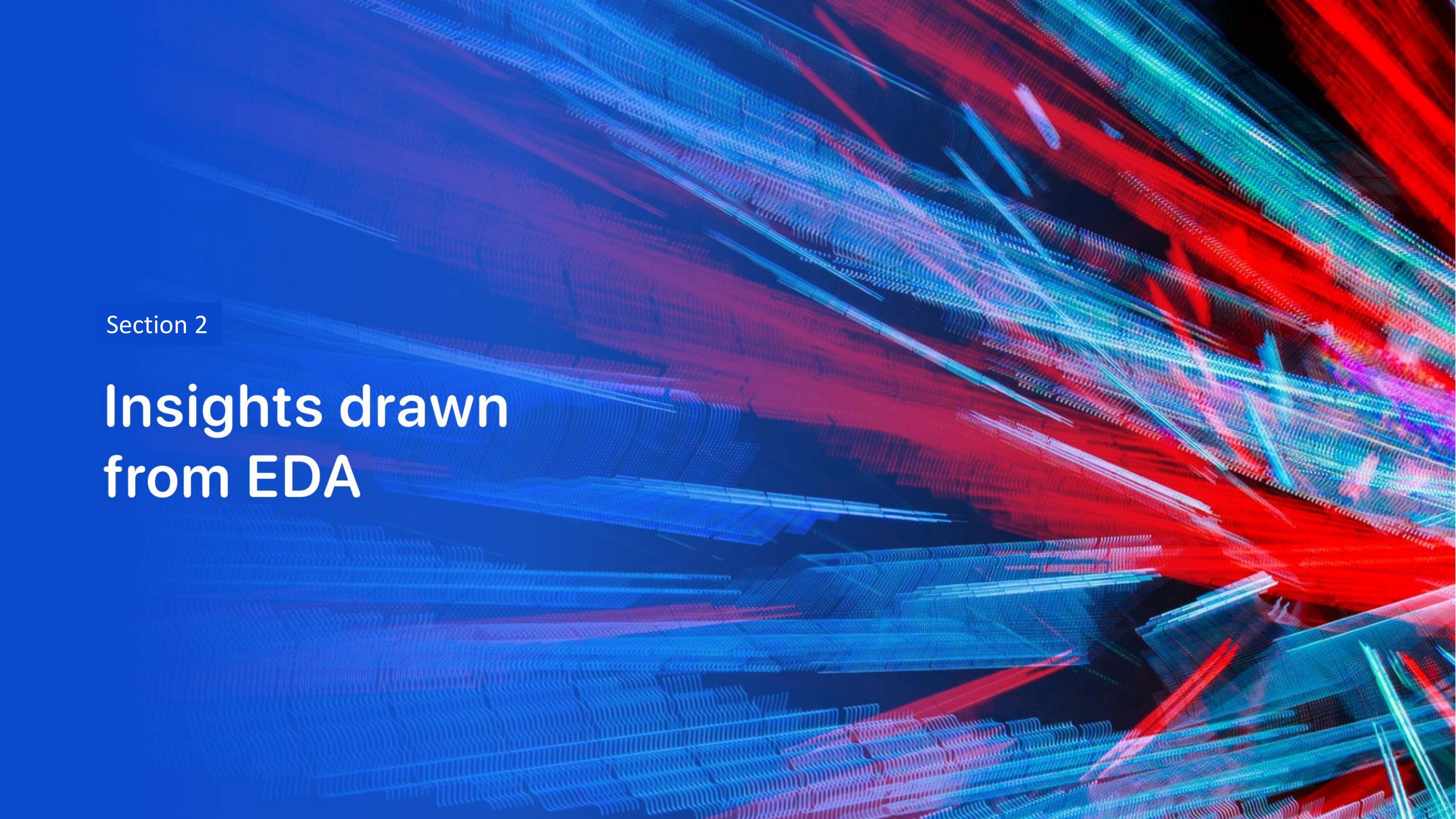


SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

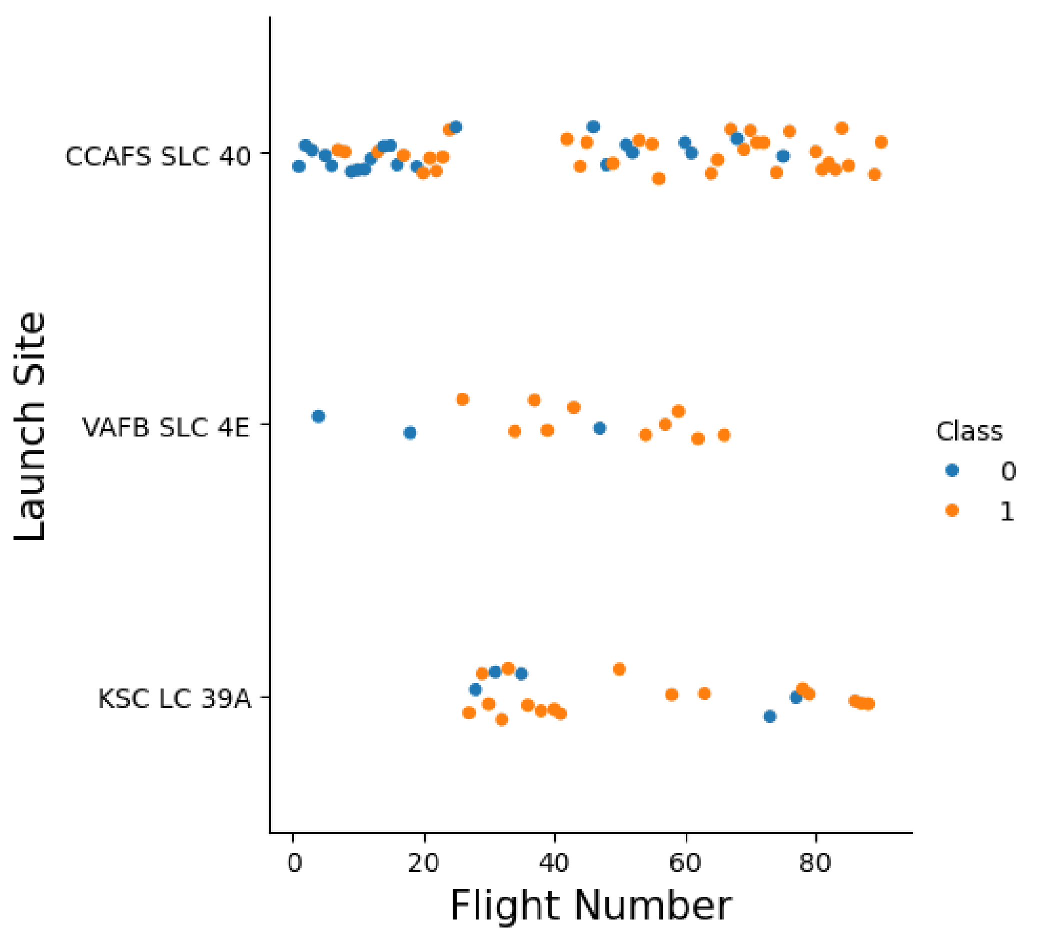
Plotly Interactive Dashboard



Section 2

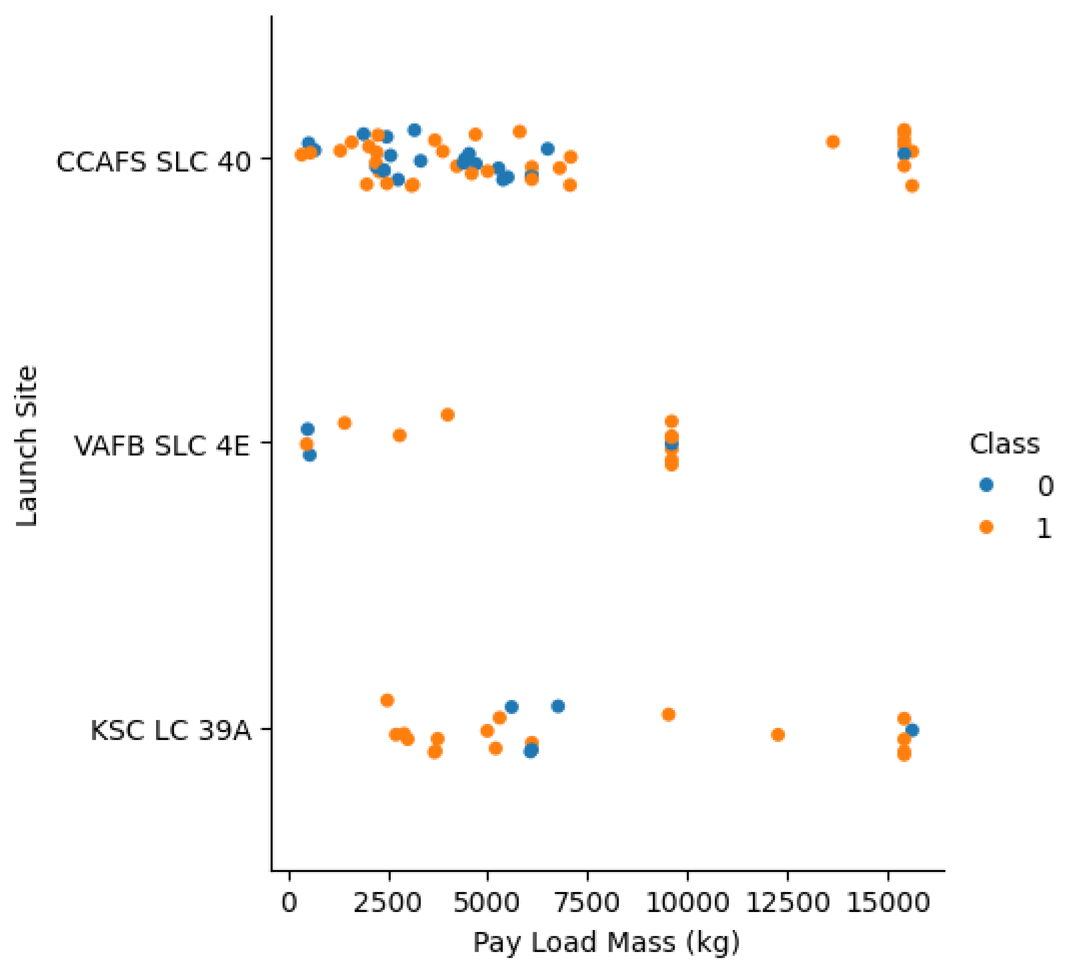
Insights drawn from EDA

Flight Number vs. Launch Site



- KSC LC 39-A and VAFB SLC 4E have fewer failures.
- VAFB SLC 4E has significantly less flights than the other two sites
- CCAFS SLC 40 has a region void of flights

Payload vs. Launch Site

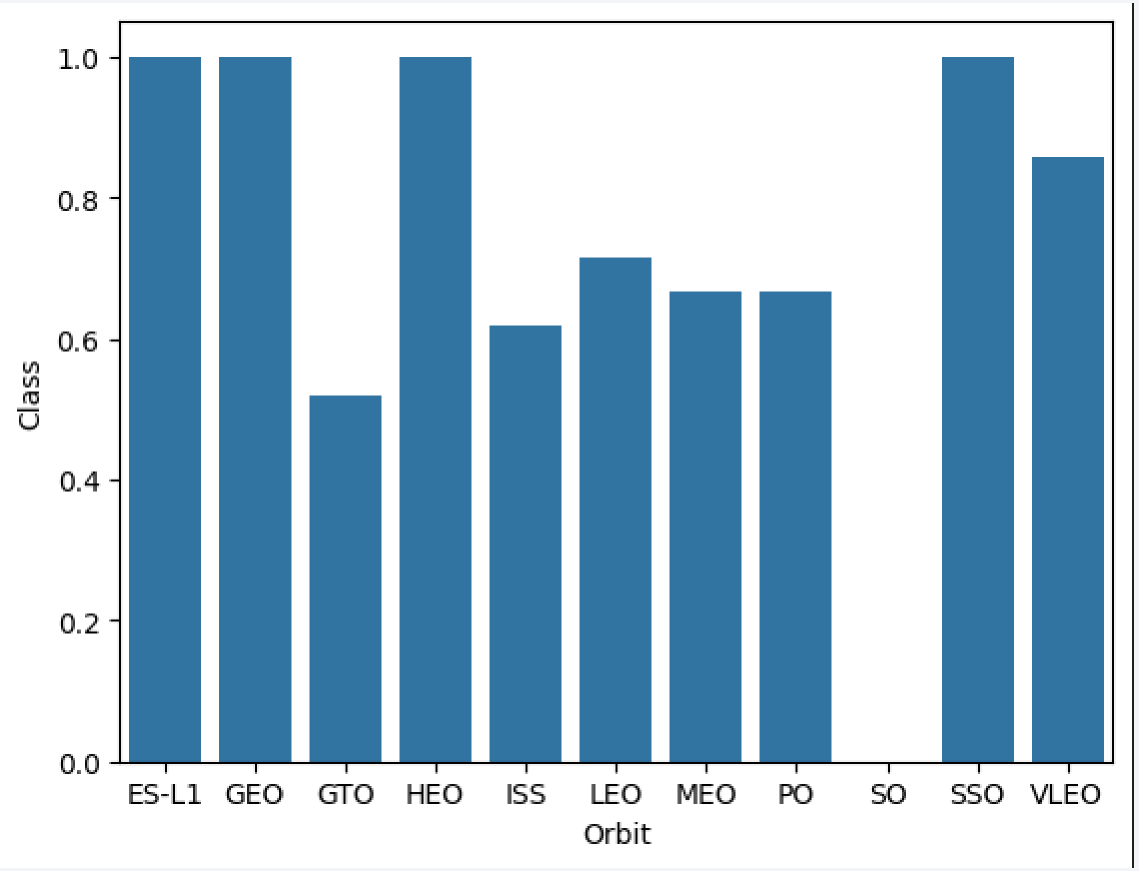


- VAFB SLC 4E doesn’t deal with high payloads

- Most of CCAFS SLC 40’s payloads are on the lower end of the payload range

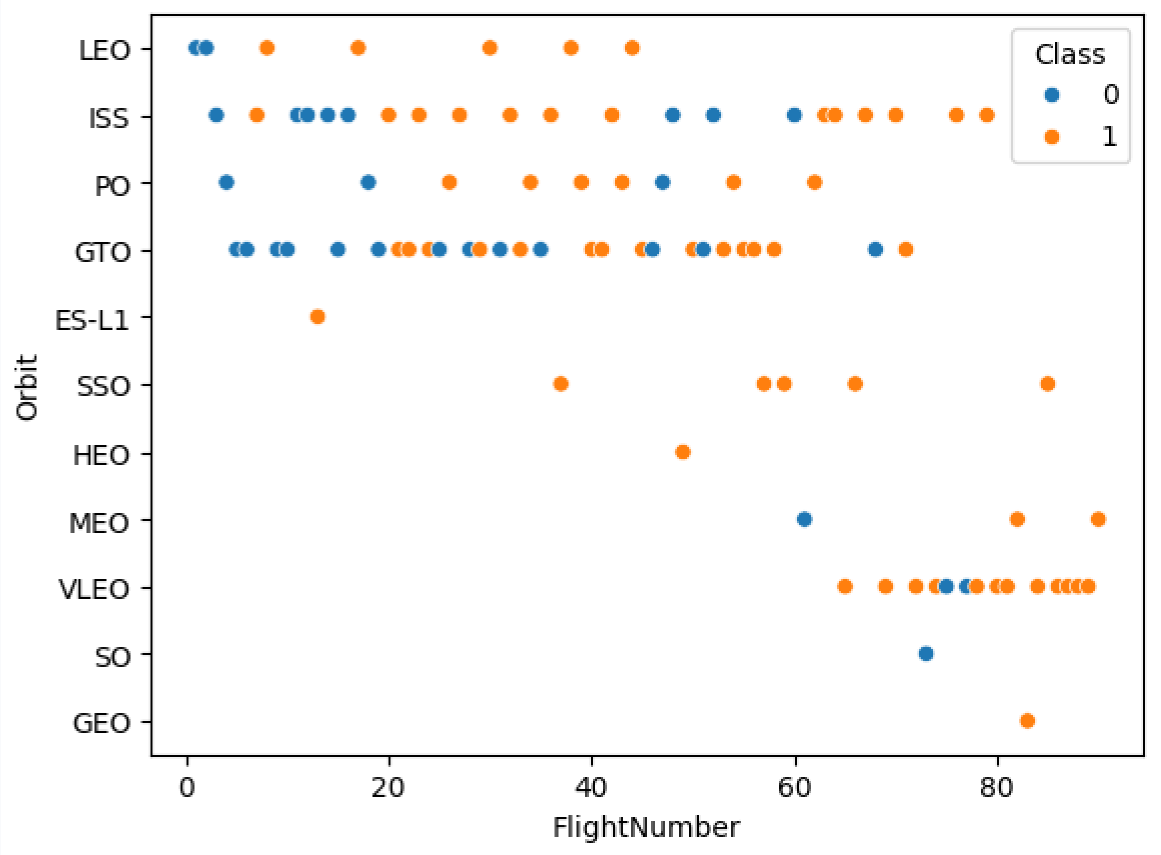
- Higher payloads appear to have a better success rate

Success Rate vs. Orbit Type



- The SO orbit has no successes
- There doesn't appear to be a common factor among the orbits that were relatively more successful or relatively less successful.

Flight Number vs. Orbit Type

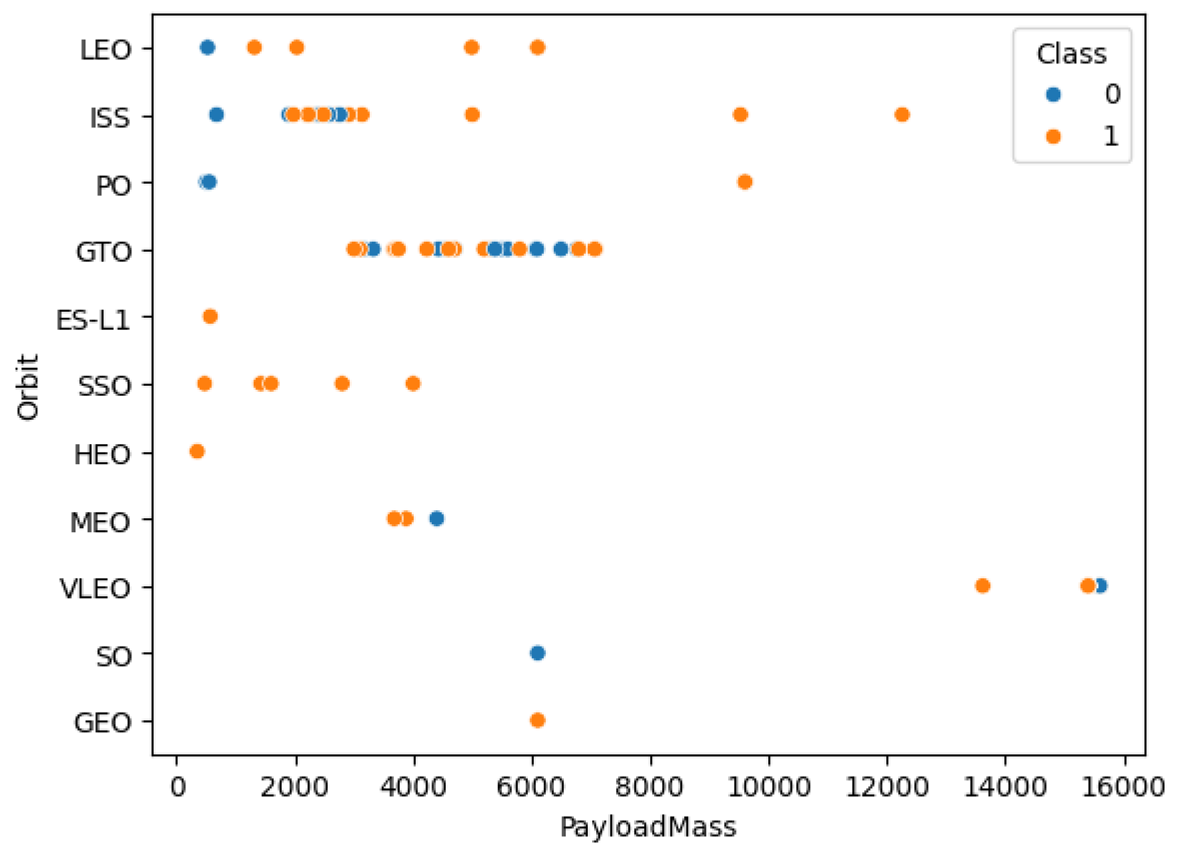


- Success of LEO flights increases with flight number

- A few of the orbits only appear in later flight numbers

- ISS and GTO seem to be the most popular orbit types, and have lower success rates

Payload vs. Orbit Type

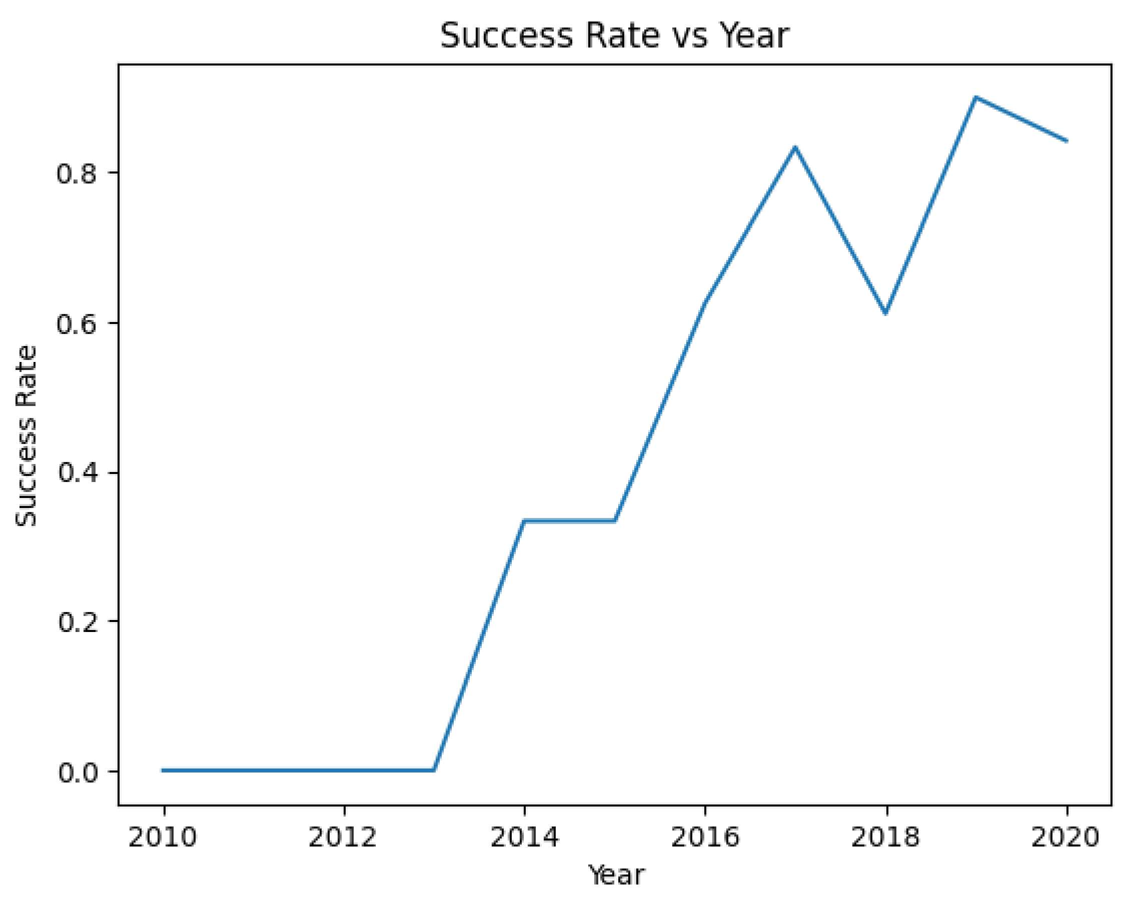


- LEO, ES-L1, SSO, MEO, HEO all seem to have exclusively low payload masses

- VLEO appears to have only high payload masses

- ISS has the widest variety of payload masses

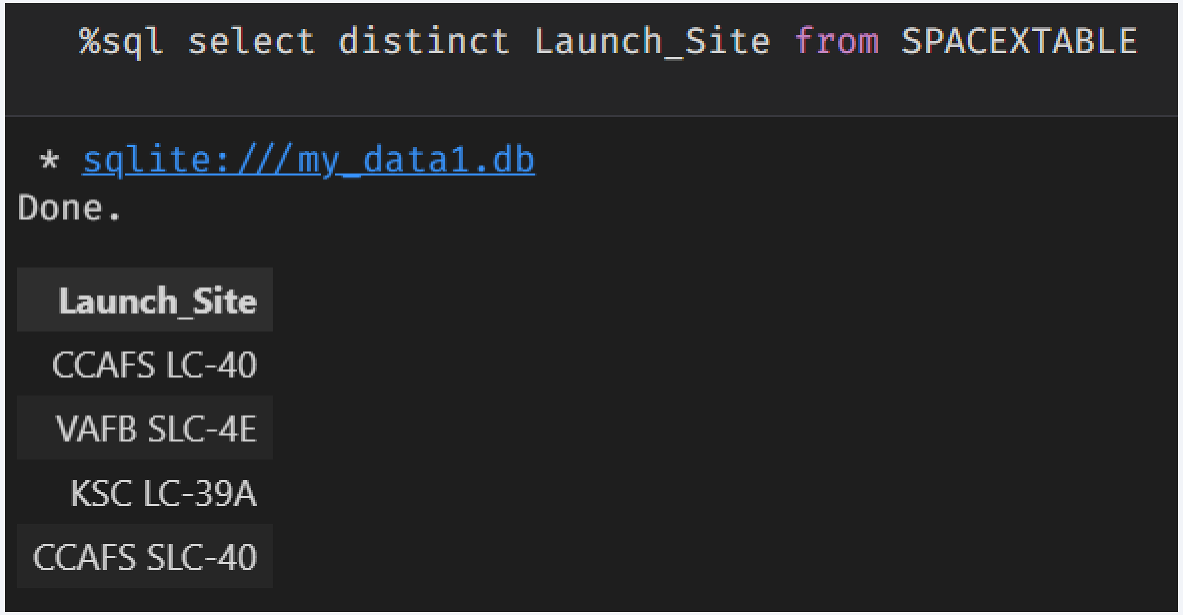
Launch Success Yearly Trend



- The overall trend is an increase in success rate over time

- For the first couple years, the success rate did not improve at all above 0%.

All Launch Site Names



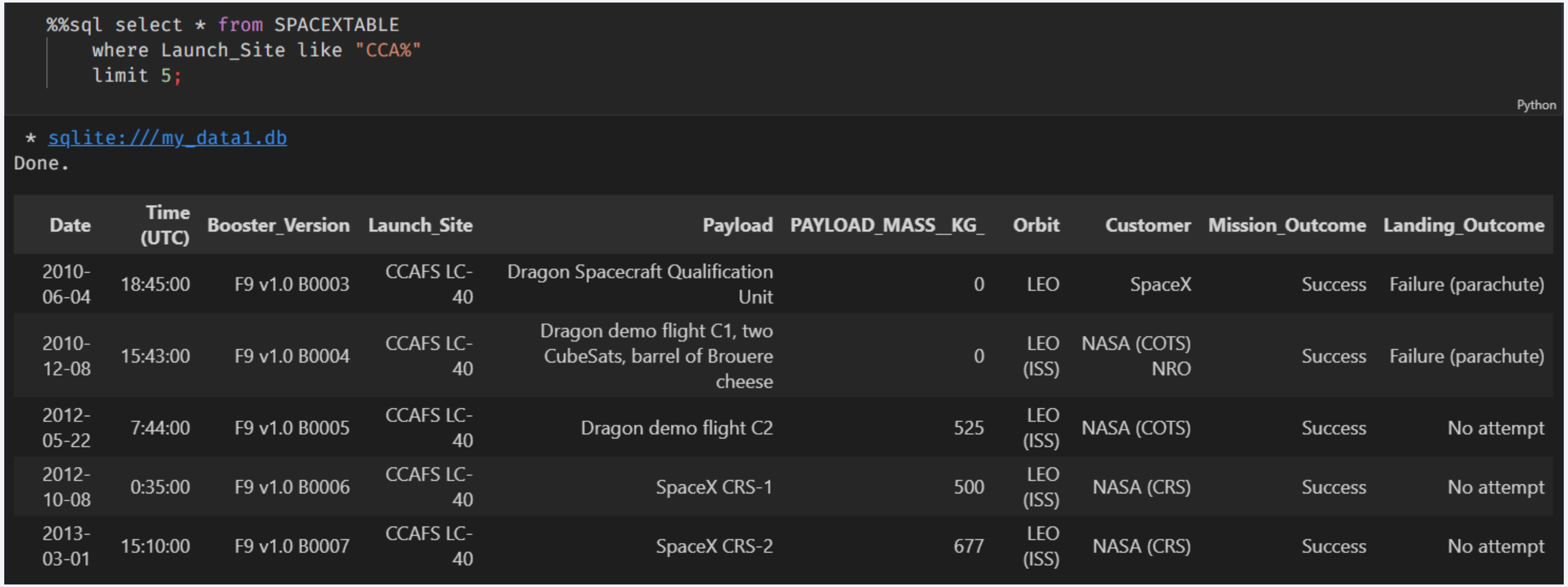
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

- All unique launch site names recorded in table

Launch Site Names Begin with 'CCA'



```
%%sql select * from SPACEXTABLE
    where Launch_Site like "CCA%"
    limit 5;
```

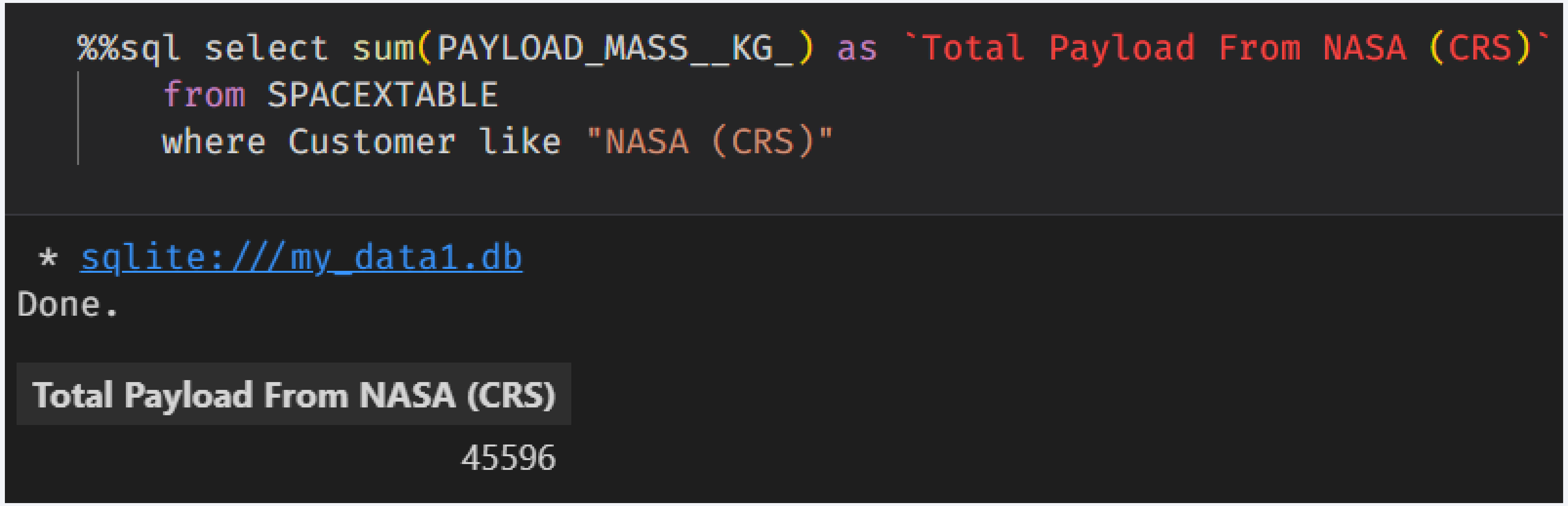
Python

```
* 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

- The first 5 records with a launch site containing “CAA”

Total Payload Mass



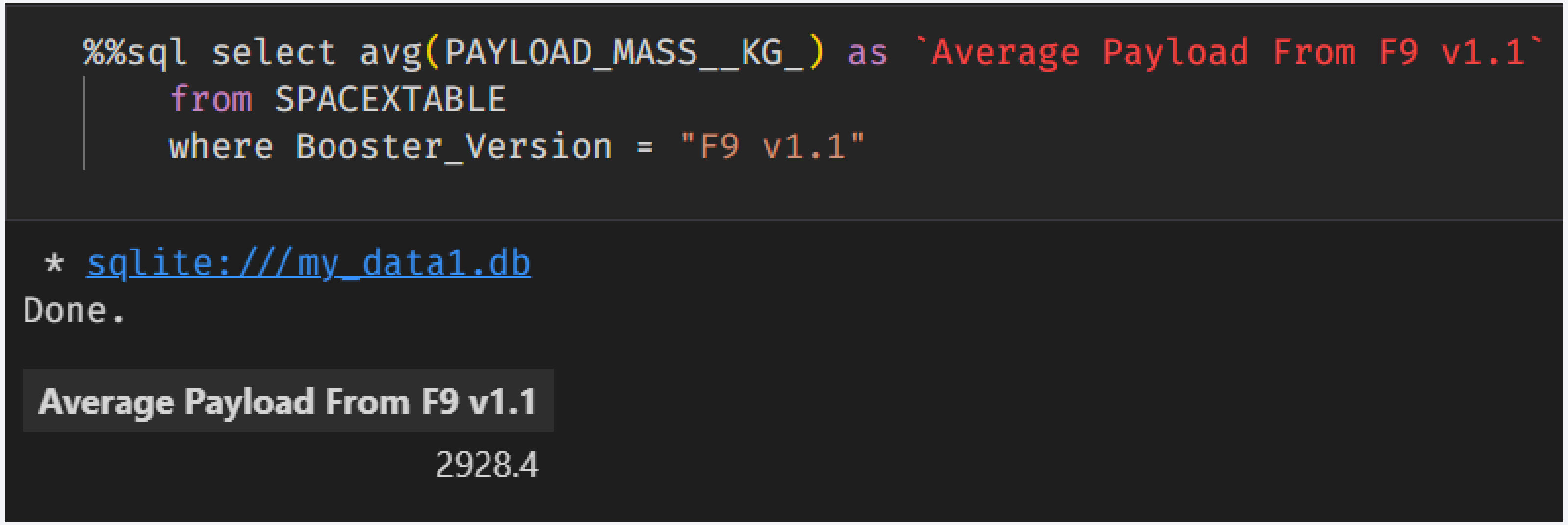
```
%%sql select sum(PAYLOAD_MASS__KG_) as `Total Payload From NASA (CRS)`
    from SPACEXTABLE
    where Customer like "NASA (CRS)"
```

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

Total Payload From NASA (CRS)
                        45596
```

- Total Payload Mass (sum) from NASA (CRS)

Average Payload Mass by F9 v1.1



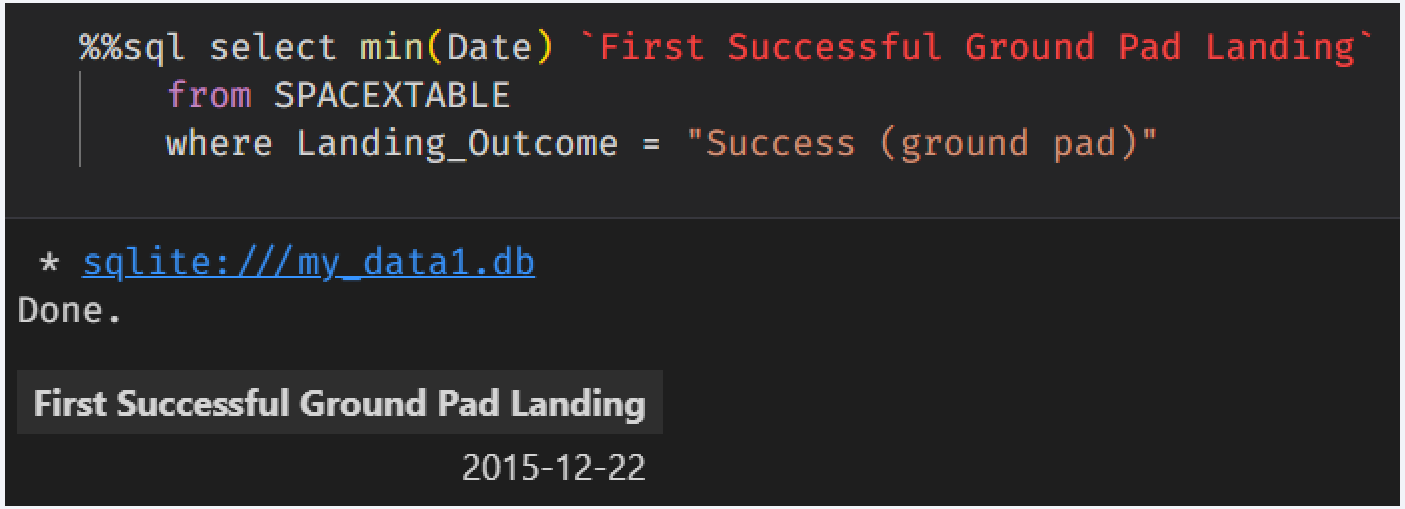
```
%%sql select avg(PAYLOAD_MASS__KG_) as `Average Payload From F9 v1.1`
    from SPACEXTABLE
    where Booster_Version = "F9 v1.1"
```

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

Average Payload From F9 v1.1
2928.4

- F9 v1.1 average Payload Mass

First Successful Ground Landing Date



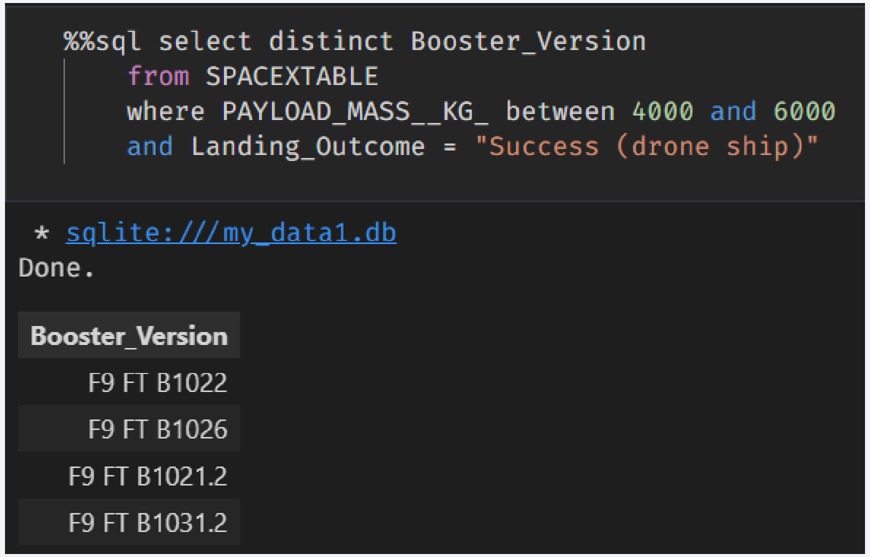
```
%%sql select min(Date) `First Successful Ground Pad Landing`
    from SPACEXTABLE
    where Landing_Outcome = "Success (ground pad)"
```

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

First Successful Ground Pad Landing
2015-12-22

- Date of the first successful ground pad landing

Successful Drone Ship Landing with Payload between 4000 and 6000



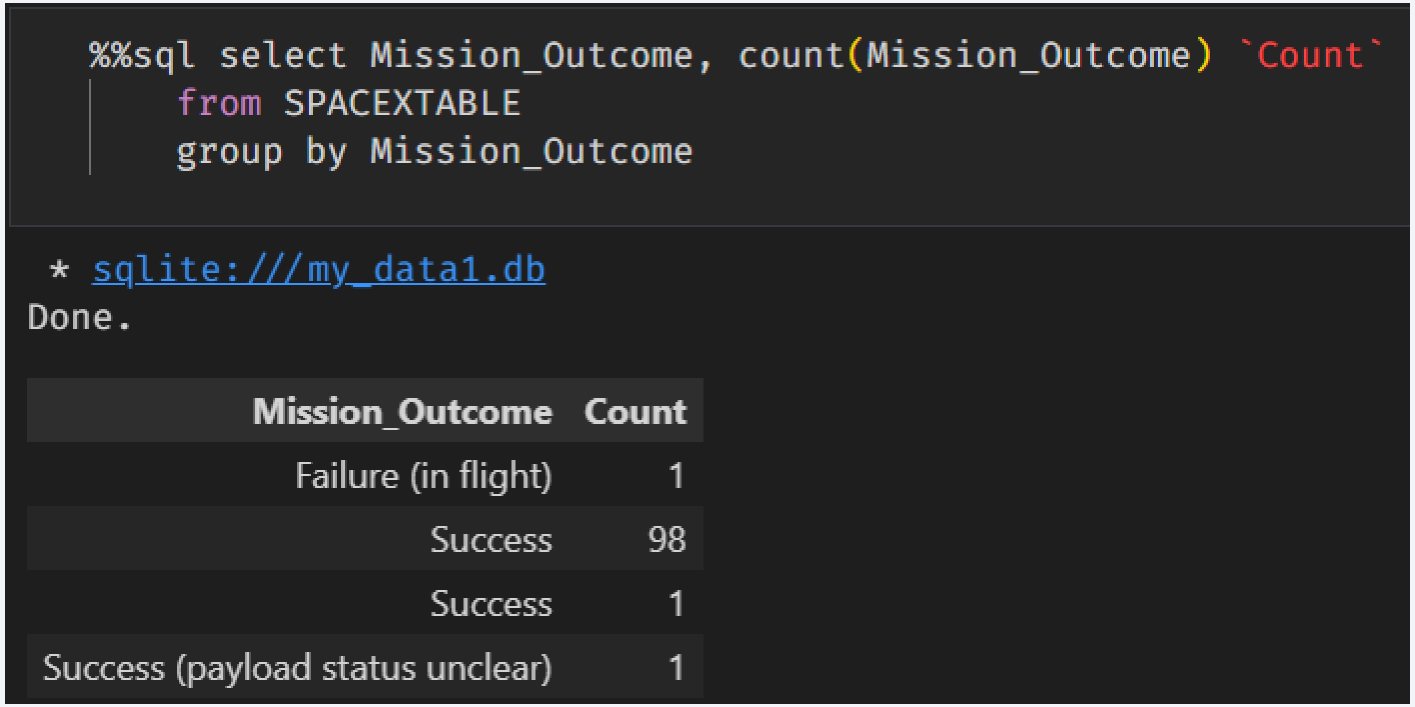
```
%%sql select distinct Booster_Version
    from SPACEXTABLE
    where PAYLOAD_MASS__KG_ between 4000 and 6000
    and Landing_Outcome = "Success (drone ship)"
```

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

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

- Names of the boosters which have successful drone ship landings and a payload mass between 4000kg and 6000kg

Total Number of Successful and Failure Mission Outcomes



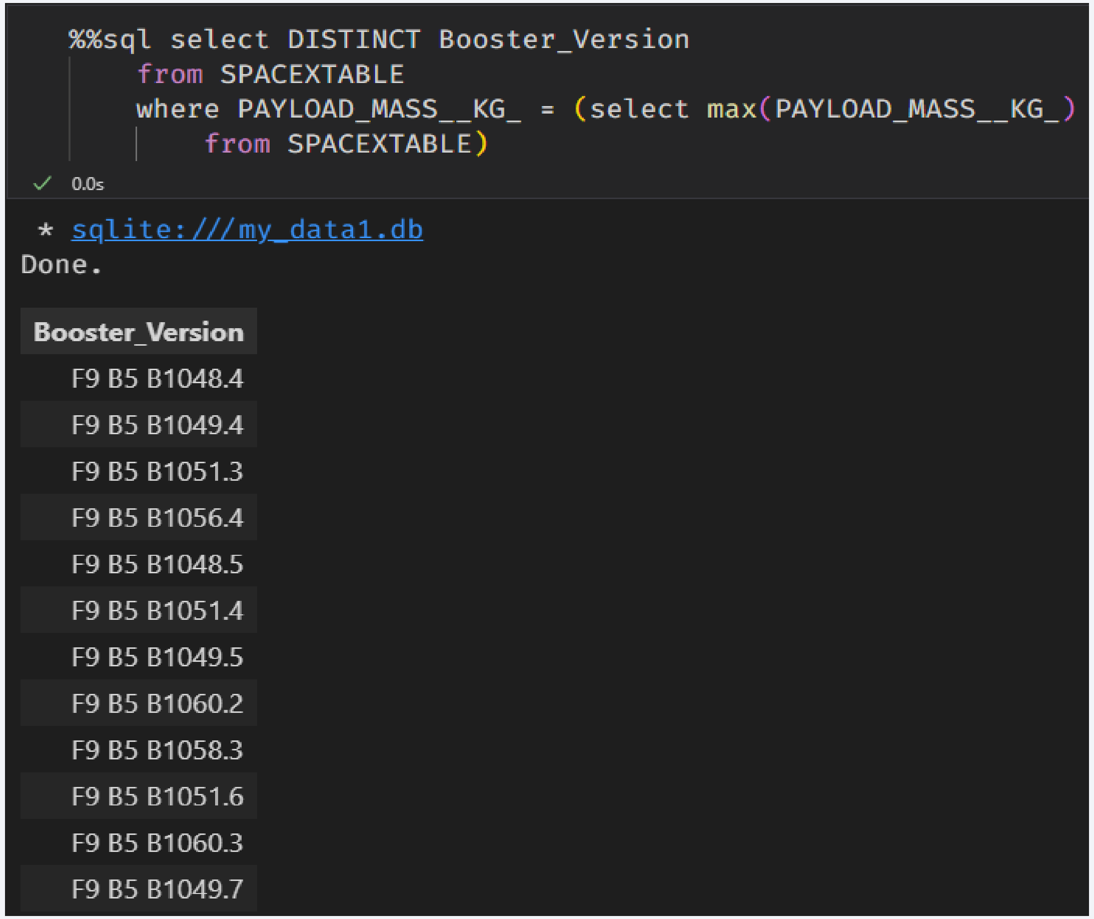
```
%%sql select Mission_Outcome, count(Mission_Outcome) `Count`
    from SPACEXTABLE
    group by Mission_Outcome
```

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

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

- Count of each type of mission outcome

Boosters Carried Maximum Payload



```
%%sql select DISTINCT Booster_Version
    from SPACEXTABLE
    where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)
        from SPACEXTABLE)
```

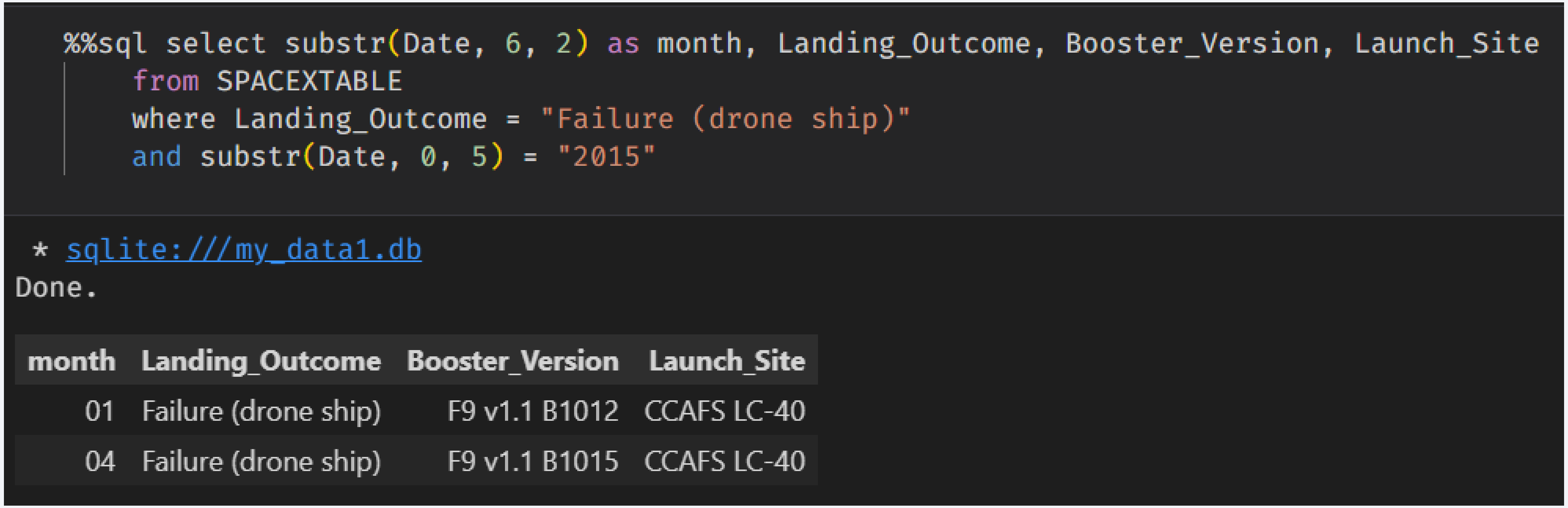
✓ 0.0s

```
 * 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

- Names of boosters which have carried the highest payload amount

2015 Launch Records



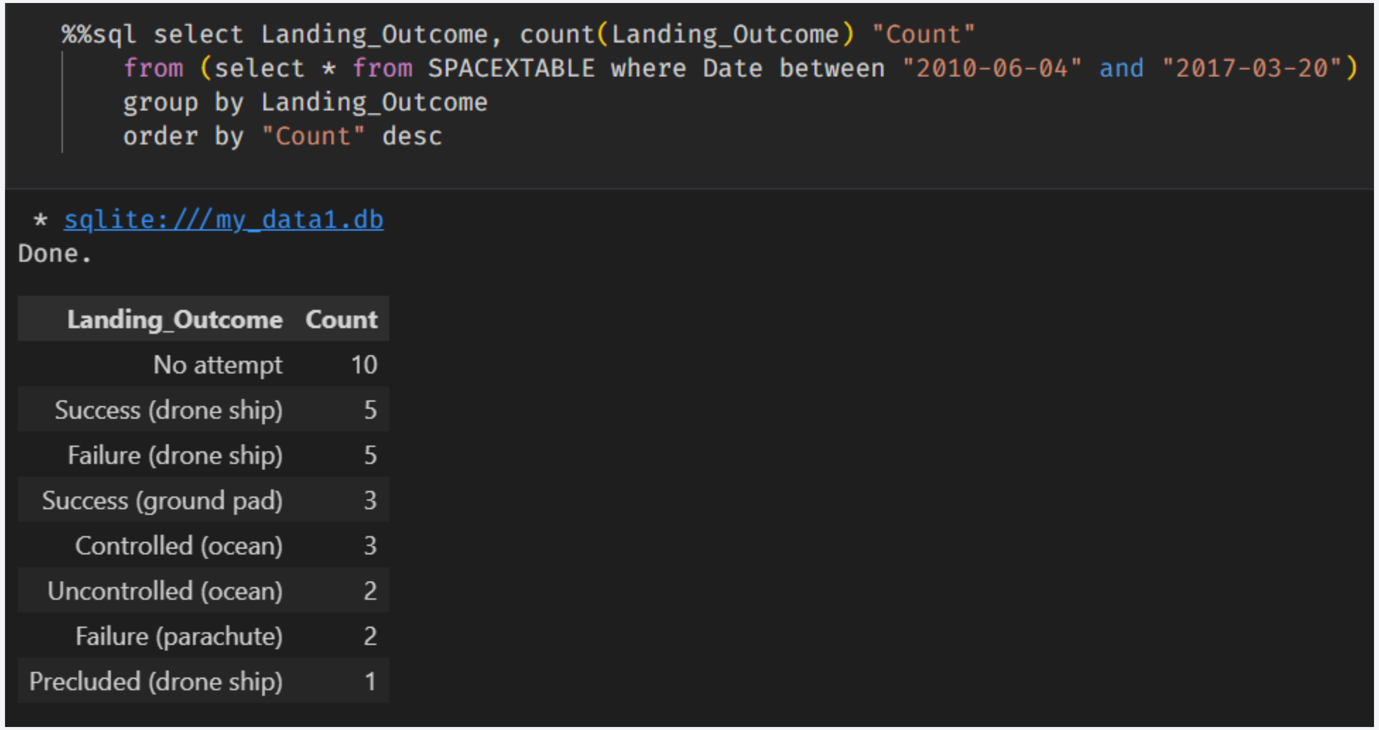
```
%%sql select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site
    from SPACEXTABLE
    where Landing_Outcome = "Failure (drone ship)"
    and substr(Date, 0, 5) = "2015"
```

```
 * 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

- Failed drone ship landing records in the year 2015, showing: month, outcome, booster version, and launch site

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



```
%%sql select Landing_Outcome, count(Landing_Outcome) "Count"
    from (select * from SPACEXTABLE where Date between "2010-06-04" and "2017-03-20")
    group by Landing_Outcome
    order by "Count" desc
```

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

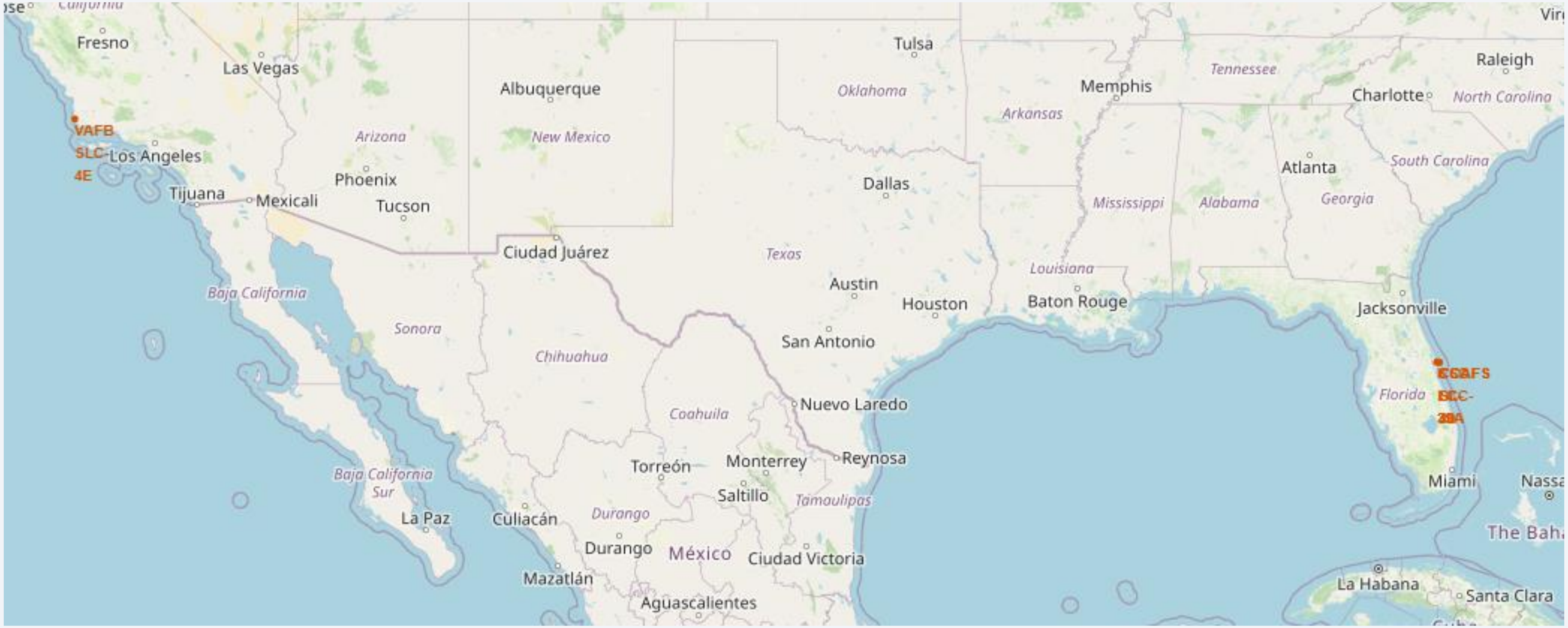
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Ranking the counts of different types of landing outcome between 4/6/2010 and 20/3/2017

Section 3

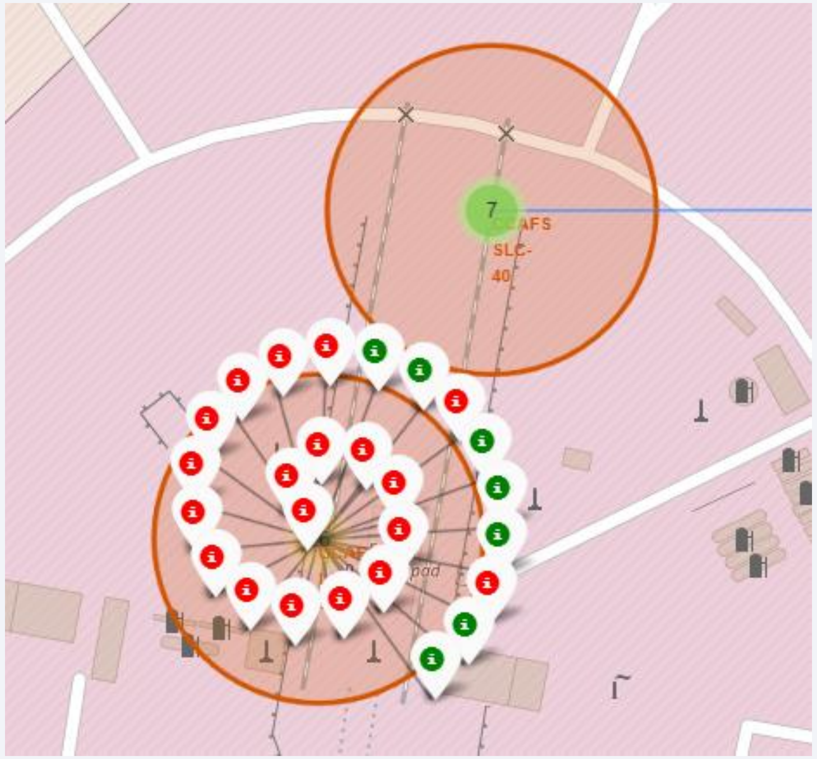
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



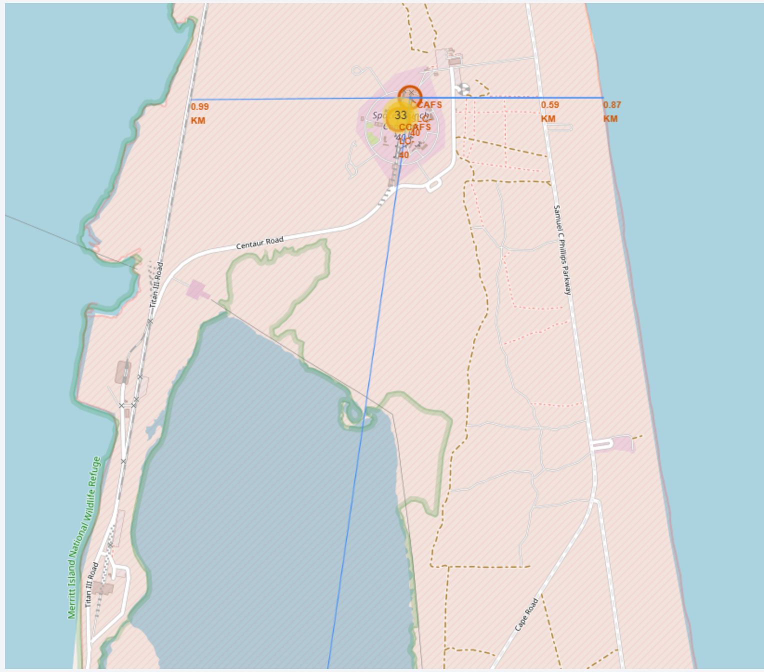
- All launch sites are relatively near the coast

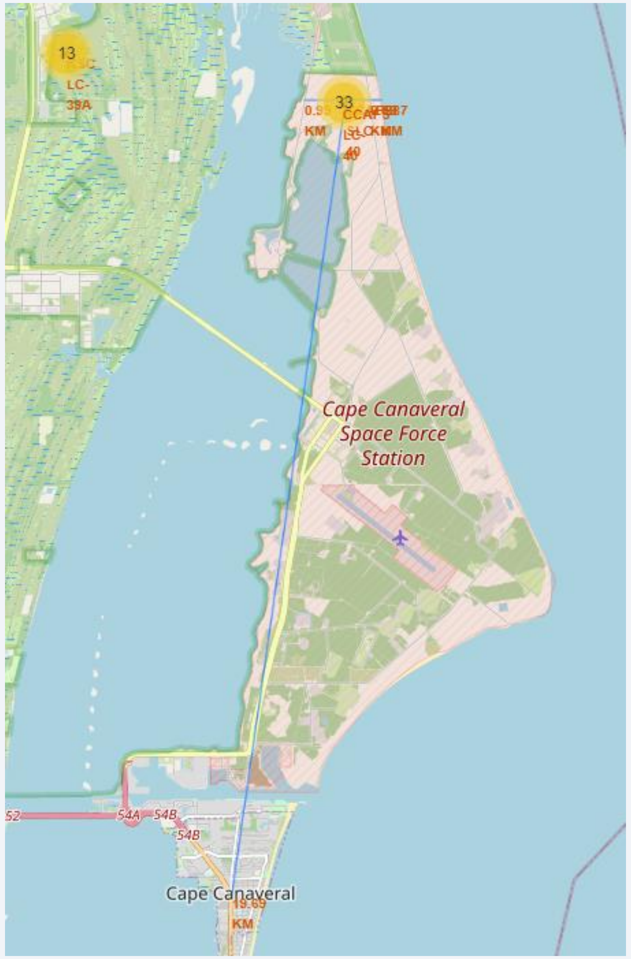
Launch Site Landing Outcome Clusters



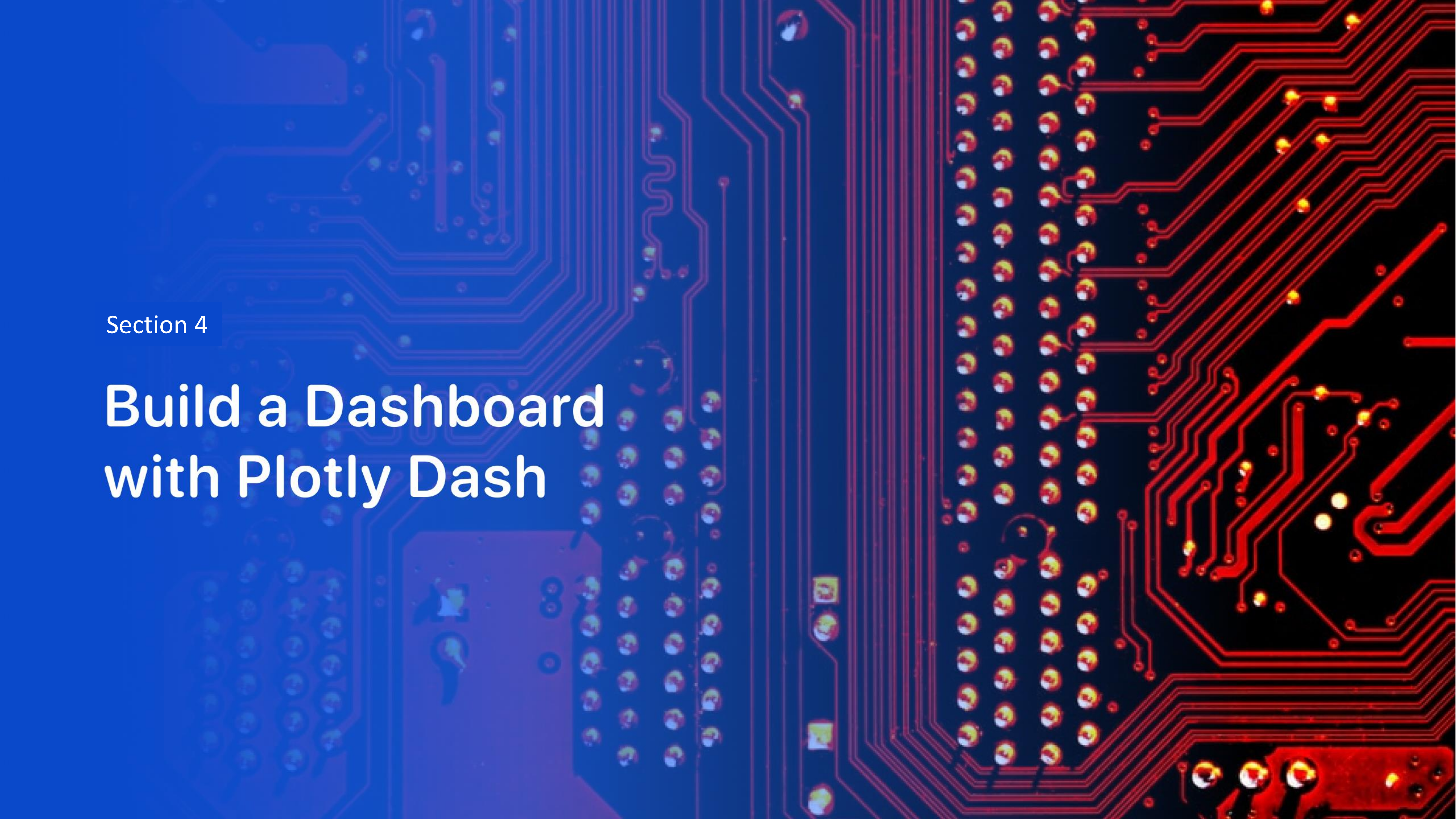
- This launch site has a high amount of failures

CCAFS SLC-40 Infrastructure Proximities





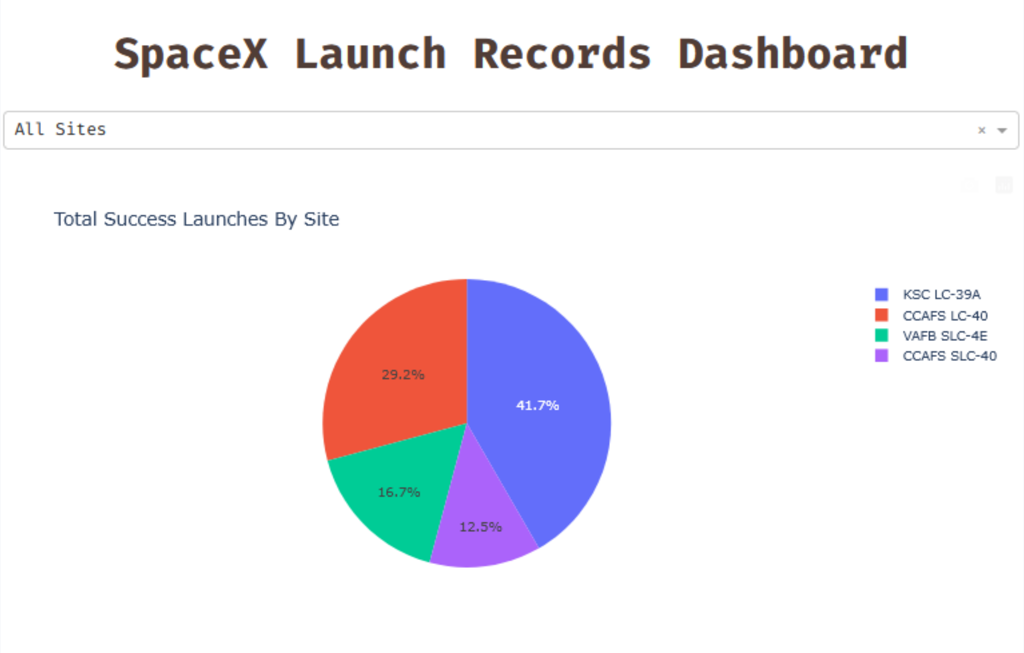
- Launch site is close to coast and railway, but far from nearest city (in comparison).



Section 4

Build a Dashboard with Plotly Dash

All Sites Total Successful Launches

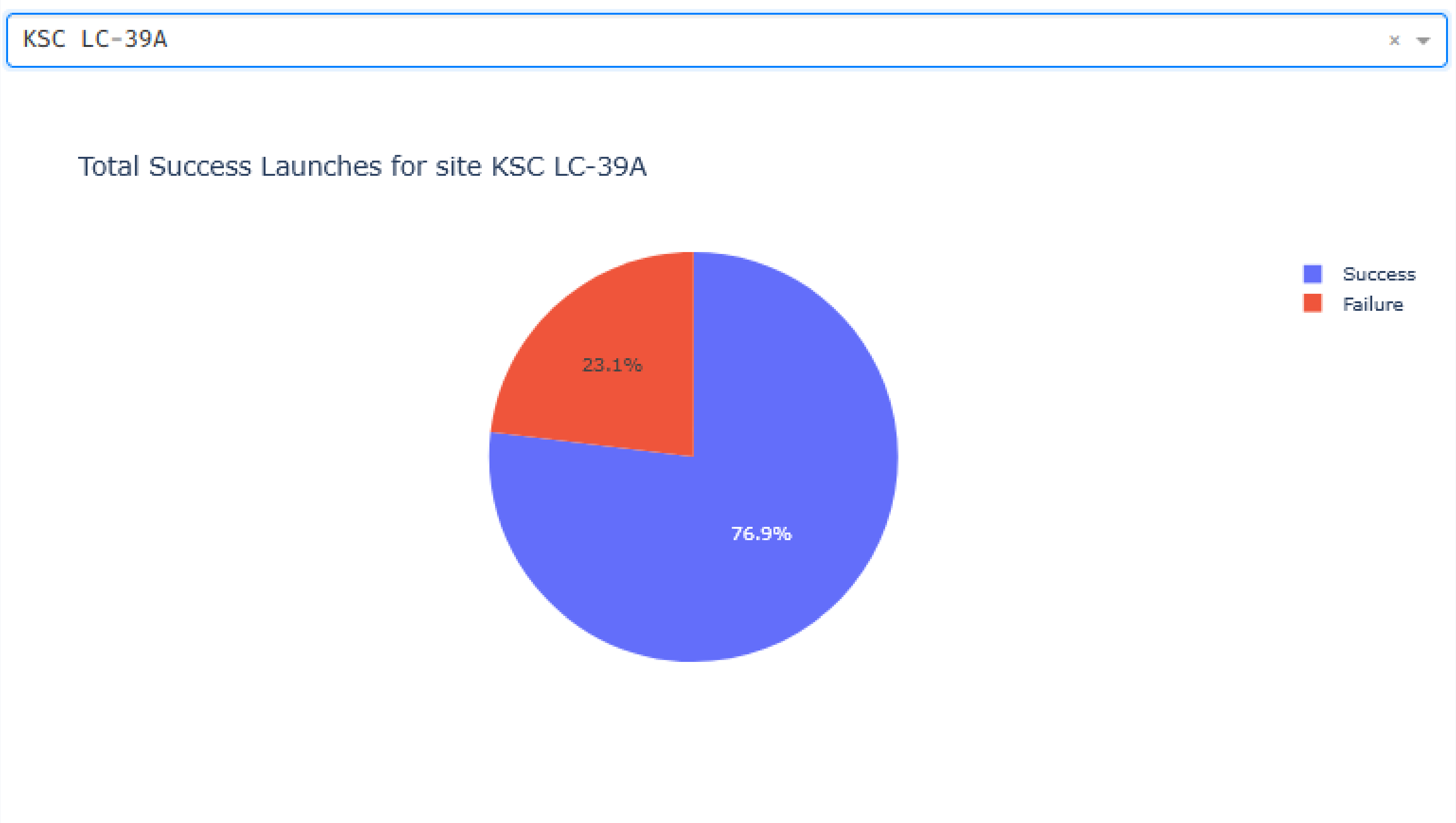


SpaceX Launch Records Dashboard

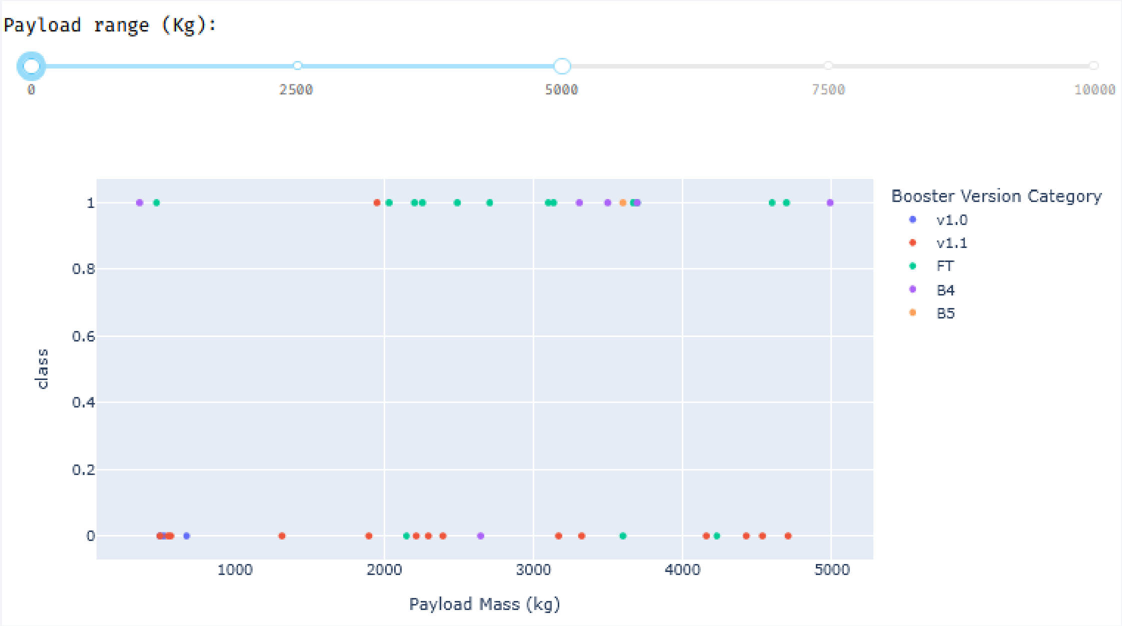
All Sites

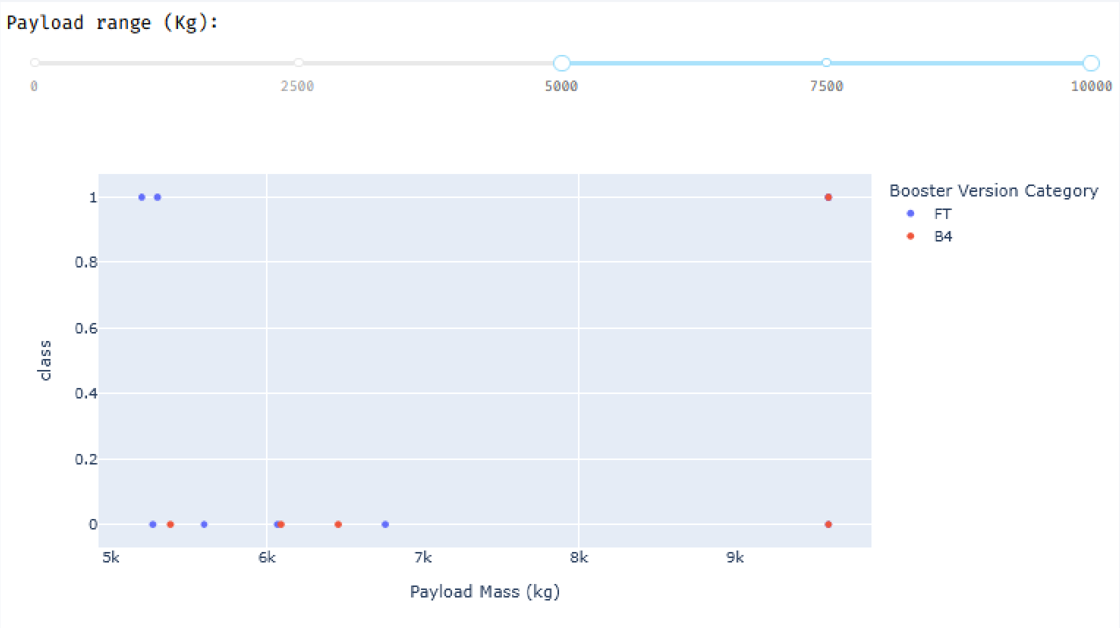
- Pie chart of Total successful launches, per site

Highest Launch Success Rate Site



Payload Range, Booster Category, and Success Rate



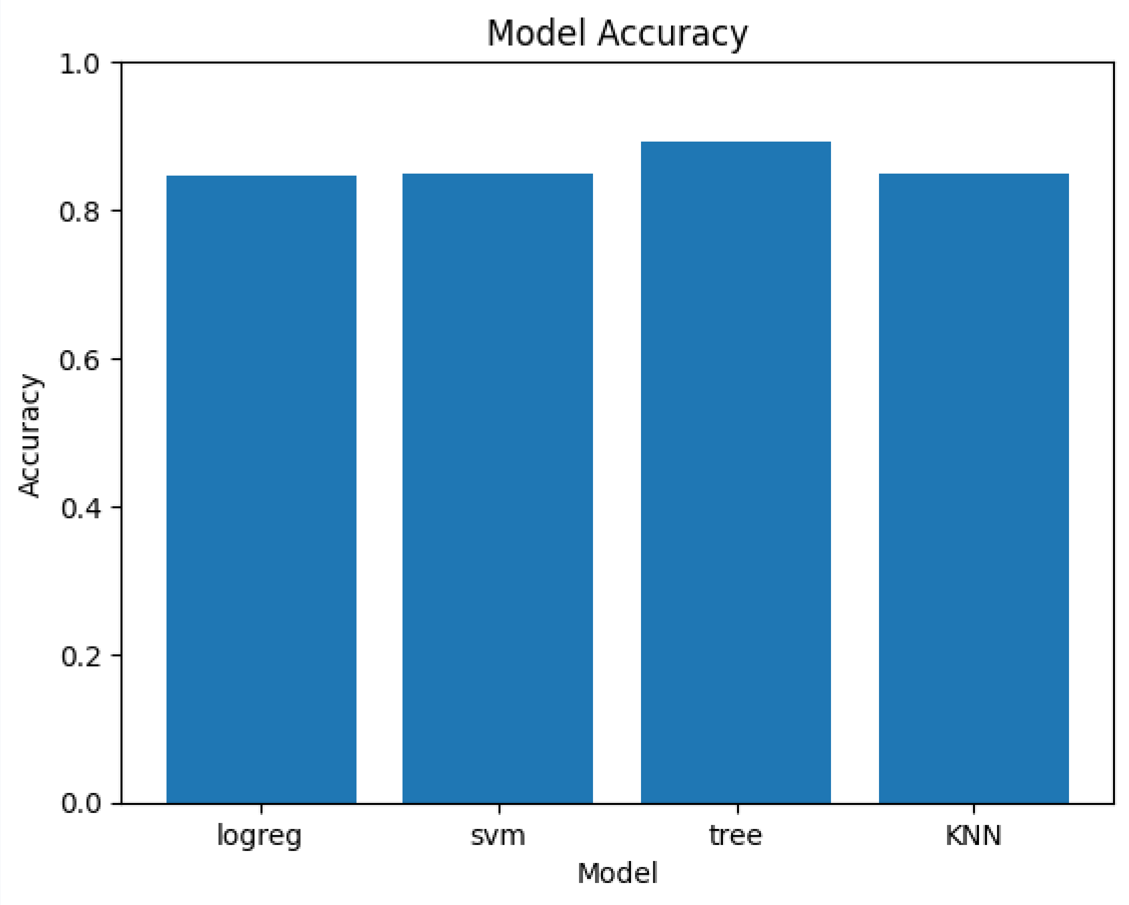


- FT booster has a high success rate
- FT and B4 are the only booster versions to carry high payloads
- All booster categories have carried low payloads

Section 5

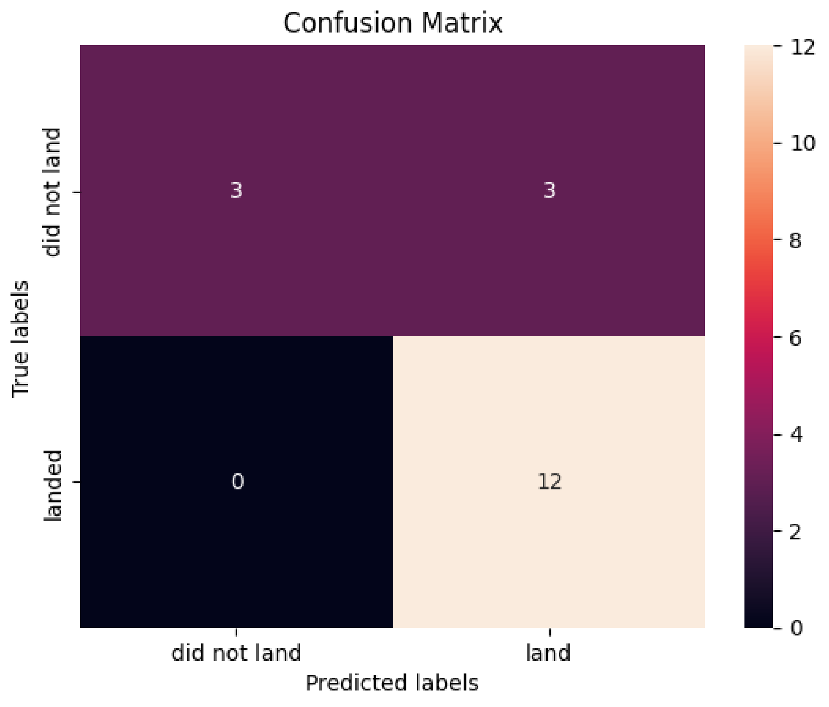
Predictive Analysis (Classification)

Classification Accuracy



- The Tree Classification model has the highest accuracy

Tree Classification Confusion Matrix



- The model can accurately predict with a high accuracy when a rocket did land, but incorrectly predicts (50% of the time) when it did not land

Conclusions

- The yearly success rate’s general trend is an increasing one
- All launch sites appear near coastlines and relatively far from cities/towns
- KSC LC 39-A had the highest success rate
- The Tree Classification model performed the best

Appendix

- All code can be found on GitHub at: https://github.com/LukahConnollySams/Capstone

Thank you!