IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Espinoza
February 14, 2026

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection – Used APIs to access and collect launch data for historical SpaceX launches
- Data Processing – Formatted data and adjusted for errors. Gathered particular data for exploratory analysis using query languages
- Dashboards and Visualizations – Display data insights and trends using custom graphs and an interactive dashboard for real time hypothesis testing.
- Model Building and Machine Learning - Compared model types and hyperparameters to find most accurate model for analysis.

Summary of all Results

- Data Trends – Analyzed for factors that might aid in predicting success of first stage landings by creating custom graphs and comparative displays
- Model Testing – Decision Tree returned a 94.44% accuracy to lead tested models which include Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors
- Overall Result – Location of launch and payload mass were shown to be ideal predictors of first stage landing success

Introduction

Project Background and Context

- This project aims to win the space race by applying data science principles to historical SpaceX launch data. By determining predictive factors in a launch's first stage successful re-landing, we can gain a unique market advantage. This knowledge can aid in predictive financial reporting and creating accurate cost expectations without having to fully build and test costly units.

Problems to Solve

- Which accessible factors are shown to aid in predicting a successful landing?
- Which machine learning model will have the most success in predicting landing success?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection Methodology:
  - Data was collected through the use of an API. Said API sourced and extracted historical SpaceX launch data for additional analysis.
- Perform Data Wrangling
  - Data was inspected to remove/replace missing values, remove duplicate values, standardize data formats, and introduce new features. Preliminary analysis was done to find some descriptive analytics to aid in creating a framework to continue efficient hypothesis testing.
- Perform Exploratory Data Analysis (EDA) using Visualization and SQL
  - Creating preliminary charts and graphs to find general trends such as landing success rates vs payload masses or launch location. Used SQL commands to see additional descriptive table and trends.

Methodology Continued

Executive Summary

- Perform Interactive Visual Analytics using Folium and Plotly Dash
  - Took advantage of Folium to create custom maps to display relevant launch data such as launch location, landing success and additional location based data.
  - Used Plotly to create an interactive dashboard. Location dropdown and mass sliders enabled real time hypothesis testing. This enabled the ability to successfully analyze predictive trends.
- Perform predictive analysis using classification models
  - Created Logistic Regression, SVM, K-Nearest Neighbor, and Decision Tree classification models to test effectiveness in correct analysis of SpaceX launch data.
  - Performed hyperparameter tuning using GridSearchCV to make sure that each model performed optimally.
  - Compared models on overall accuracy in predicting correct result when compared to real launch data

Methodology Continued

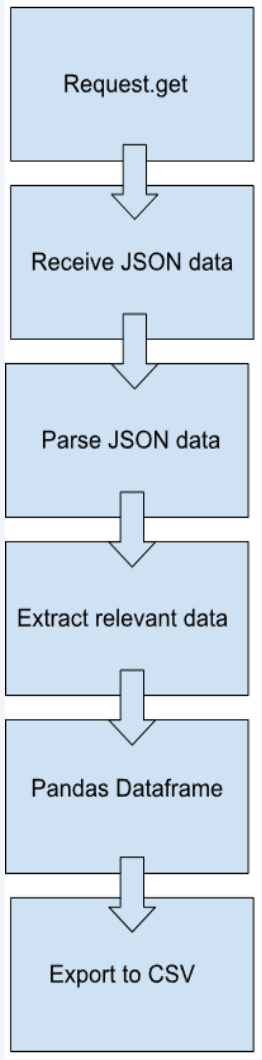
Github:

All methodology steps can be found at the following URL

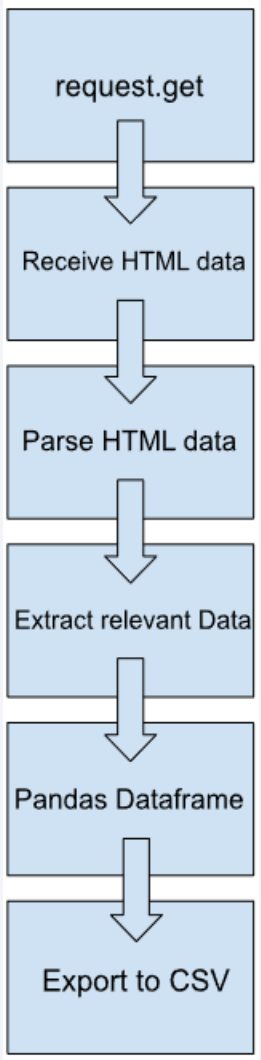
https://github.com/miguelespinoza00/Applied-Data-Science-Capstone-Project.git

Data Collection

SpaceX API

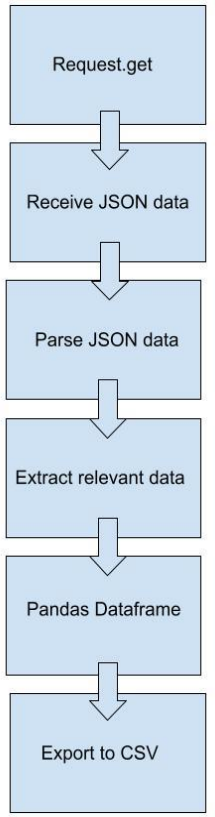


Web Scrape



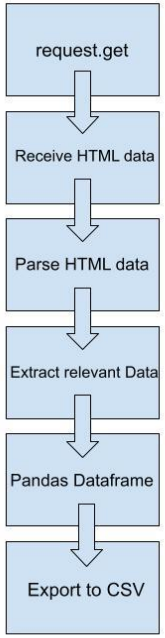
Data Collection – SpaceX API

- Initialize API using requests command to connect and collect from a desired location
- Process received JSON data to normalize and extract pertinent data
- Use pandas Python library to properly store dataframe
- Output formatted data as a CSV



Data Collection - Scraping

- Initialize API using requests command to connect and collect HTML data from a desired location
- Process received HTML data to normalize and extract pertinent data
- Use pandas Python library to properly store dataframe
- Output formatted data as a CSV



Data Wrangling

- Cleaning – Find all points with missing data values. Remove or replace values.
- Transforming - Standardizing data formats such as date formatting and string values. Creating new features to aid in future EDA and visualization demands.
- Integration - Combining data from SpaceX website and web scrape into a single dataset while maintaining data formatting.
- Verification - Recheck newly combined dataset for duplicated entries and ensure data has been properly merged

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Line Chart – Track data over a period of time.
- Bar – Compare data after grouping
- Histogram - Depict data distributions by a set number of group ranges
- Scatter Plots - Compare 2 data points to see if there are any computational relationships
- Heat map - Show the relationship between variables
- Box Plots - Displays the nature of data distribution when compared to itself in order to display outliers and trends of core data

EDA with SQL

- Count number of total launches, successful launches, failed launches
- Find success rate by location and rocket model
- Join tables together
- Filter to show results based on specific criteria
- Sorting data
- Order data for simplicity
- Calculate more descriptive analytics
- Calculate some more in depth analytics

Build an Interactive Map with Folium

- Markers – Used to display the locations of various launch sites on generated map. This aids viewer to visualize data on a global level without having prior geographical knowledge.
- Circles – Show proximity of launches. This aids viewer to visualize data on a local level without having prior geographical knowledge
- Lines – Show proximity of nearby locations that might have relevance to analysis. This aids viewer to visualize data on a local level without having prior geographical knowledge and also enables the viewer to take local locations outside of direct launch zone into account.

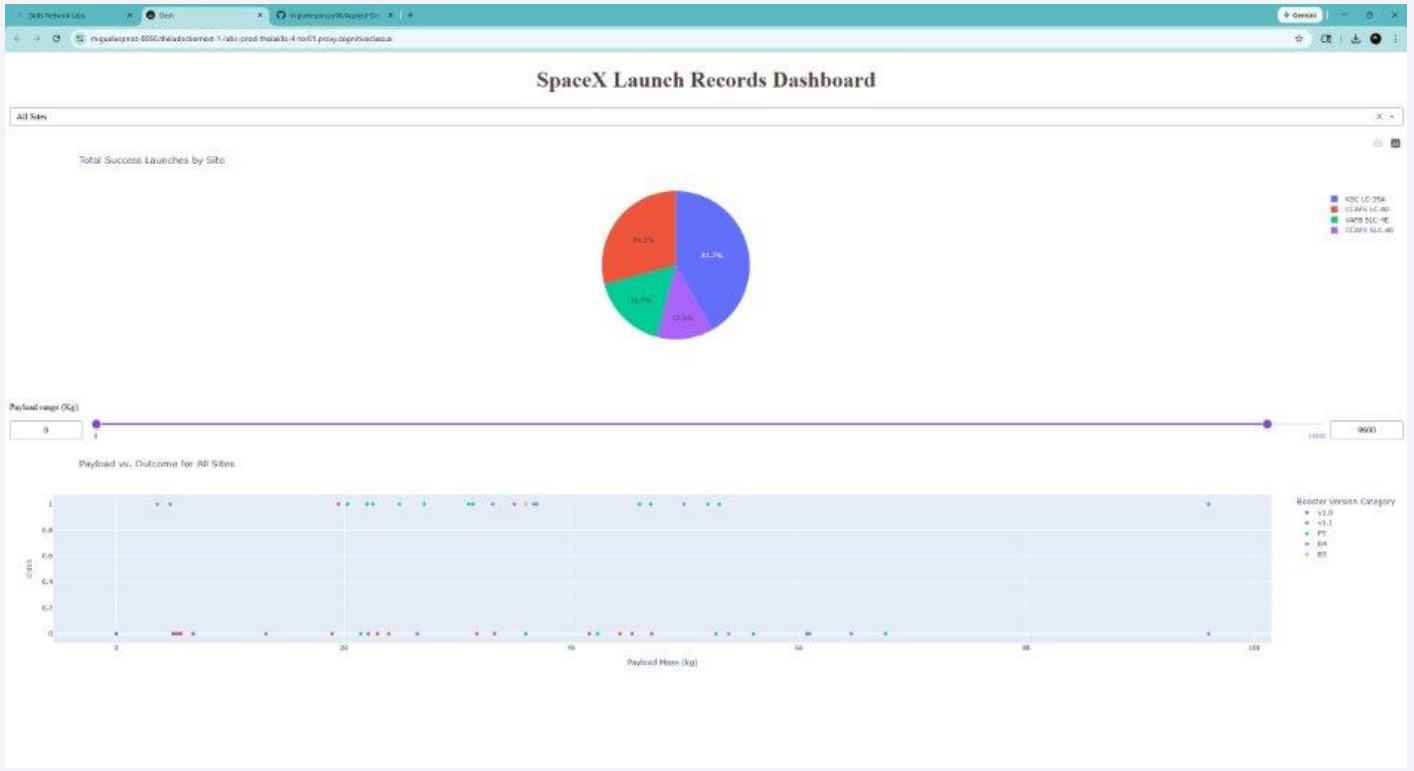
Build a Dashboard with Plotly Dash

- Pie Chart – Can be used to show the ratio at which each location is able to output successful or failed landings. Can also be used to compare the success rates of each location.
- Scatterplot - Shows the relationship between payload mass and launch success variables. This allows for analysis of any potential link that might aid in generating more successful landings.
- Launch Site Dropdown – This enables real time hypothesis testing by allowing for user to filter dashboard to their thoughts on location based data.
- Payload Mass Slider – This enables real time hypothesis testing by allowing for user to filter dashboard to their thoughts on mass based data.

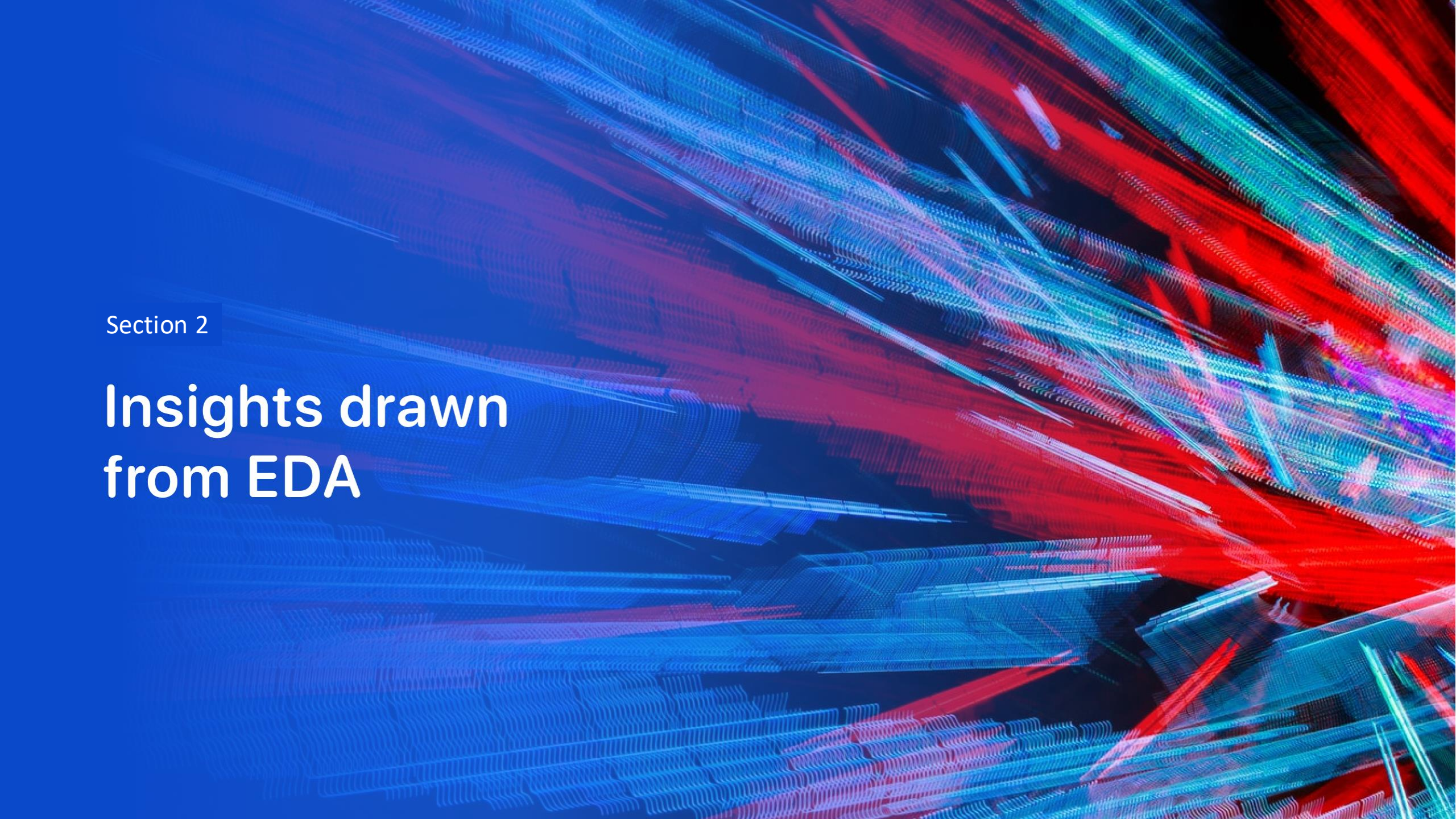
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- Data Processing – Ensure features have been standardized and are accounted for equally in analysis
- Model Inclusion – Included multiple algorithm variations for comparison
- Hyperparamater Tuning – Optimized the parameters to ensure analysis returned best case analytics
- Model Results – Cross-validated models to ensure model's abilities
- Model Iterations – Adjust models based on initial results to further optimize training
- Model Selection – Compare models based on final results and choose optimal algorithm

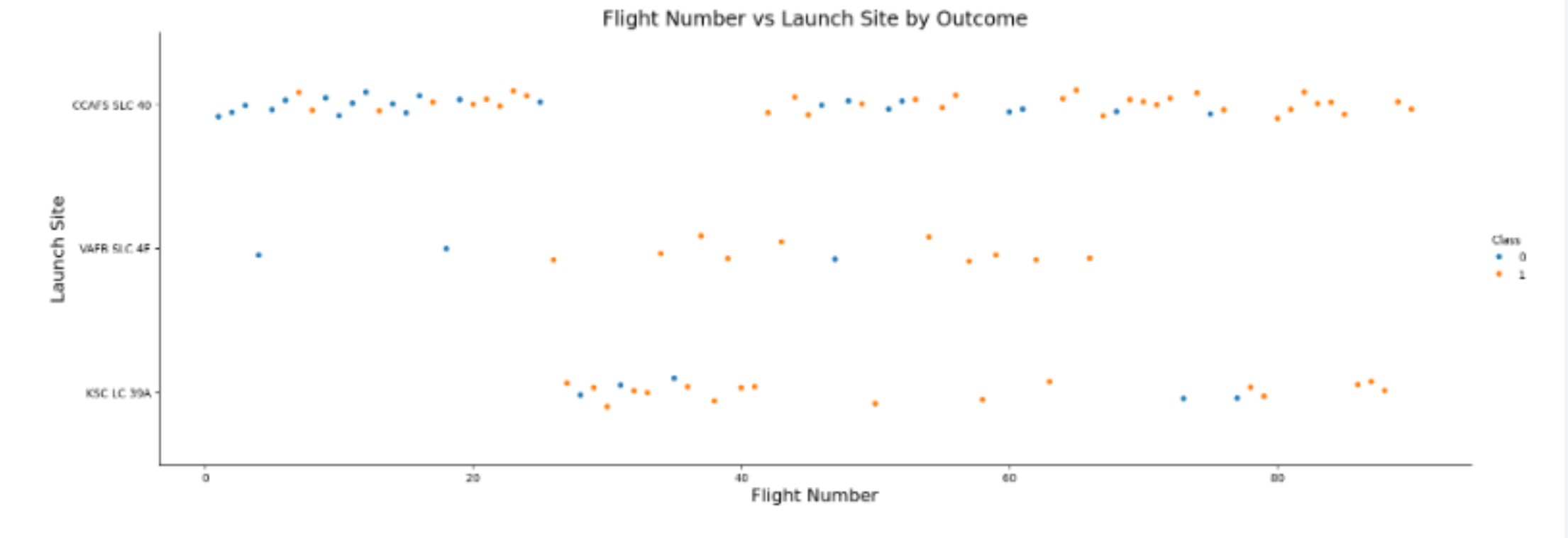
Results



- Exploratory data analysis results
- Interactive analytics
- Predictive analysis results

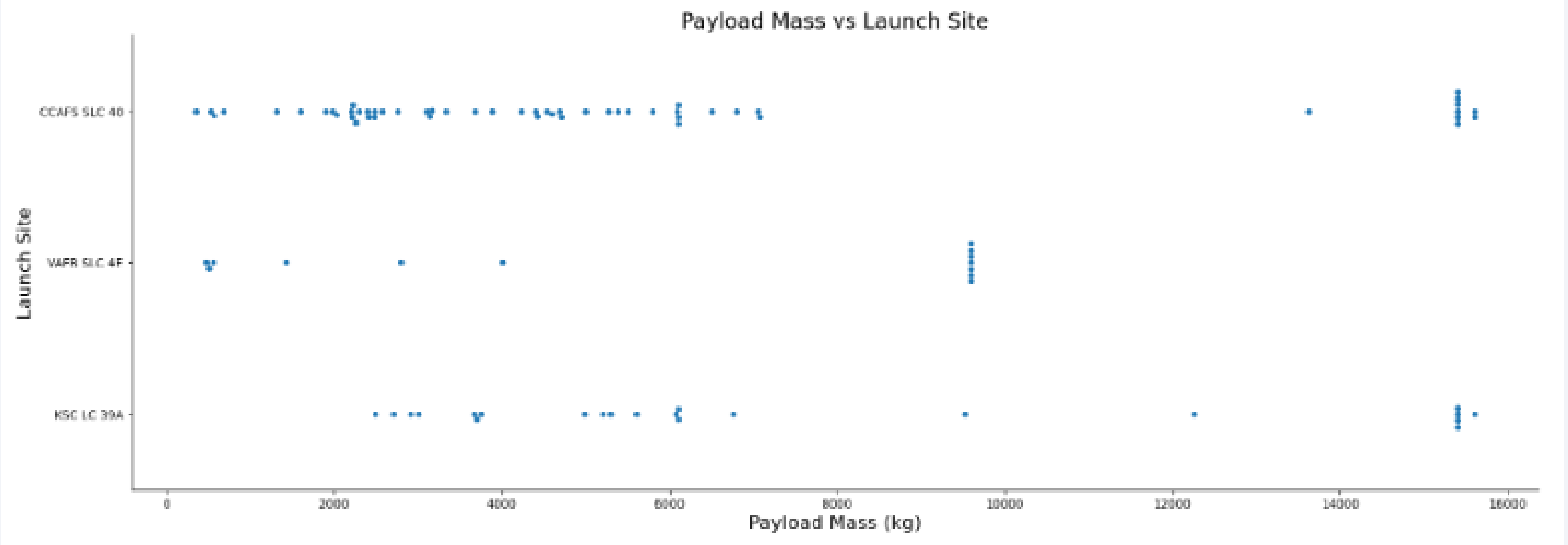

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



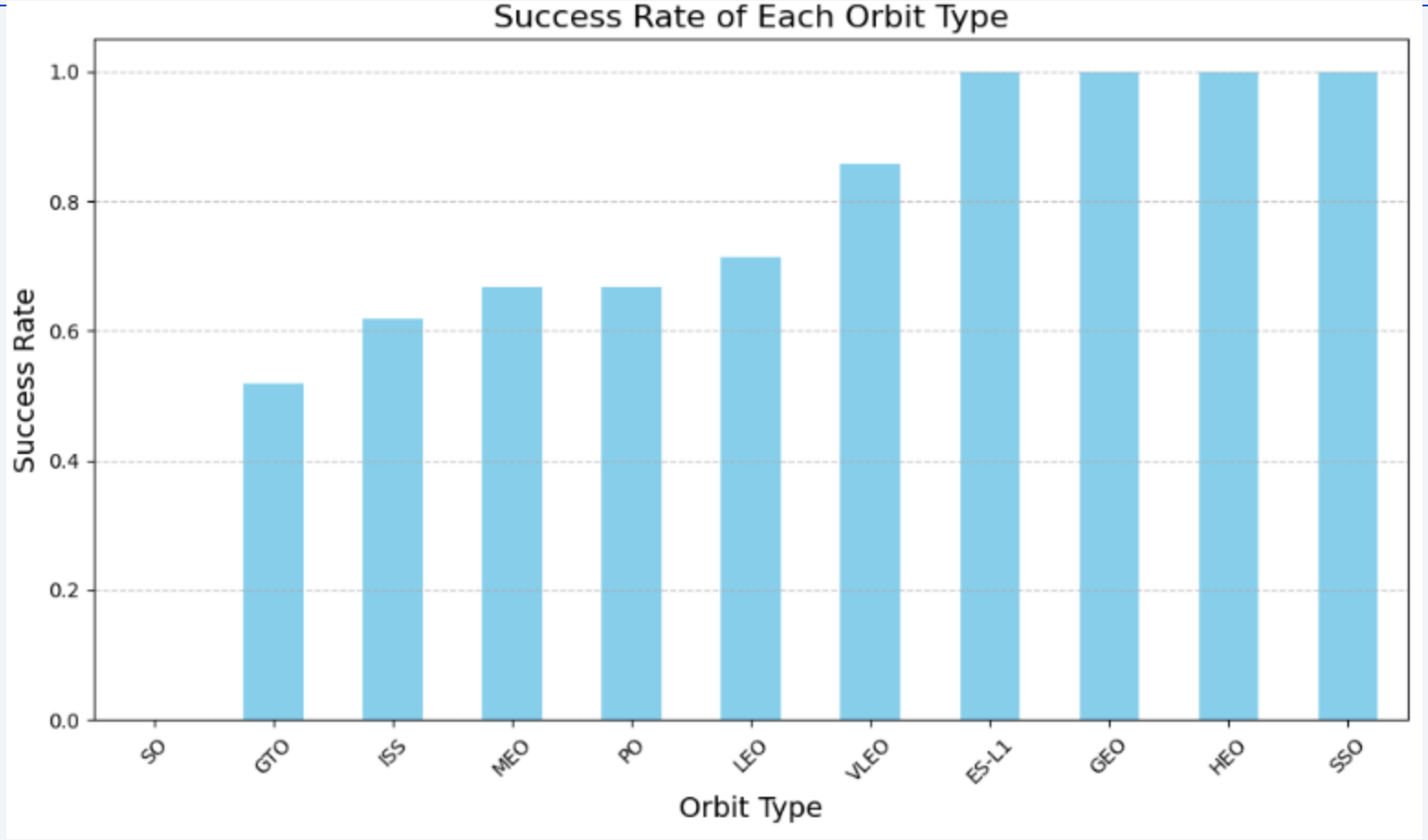
- The scatterplot shows that all 3 locations were concurrently active when concerning launches. This is important to consider since no one location is favorably getting data from the other 2 locations without having to test a rocket themselves which carries the risk of failure. All locations have a unique combination of failures and successes.

Payload vs. Launch Site



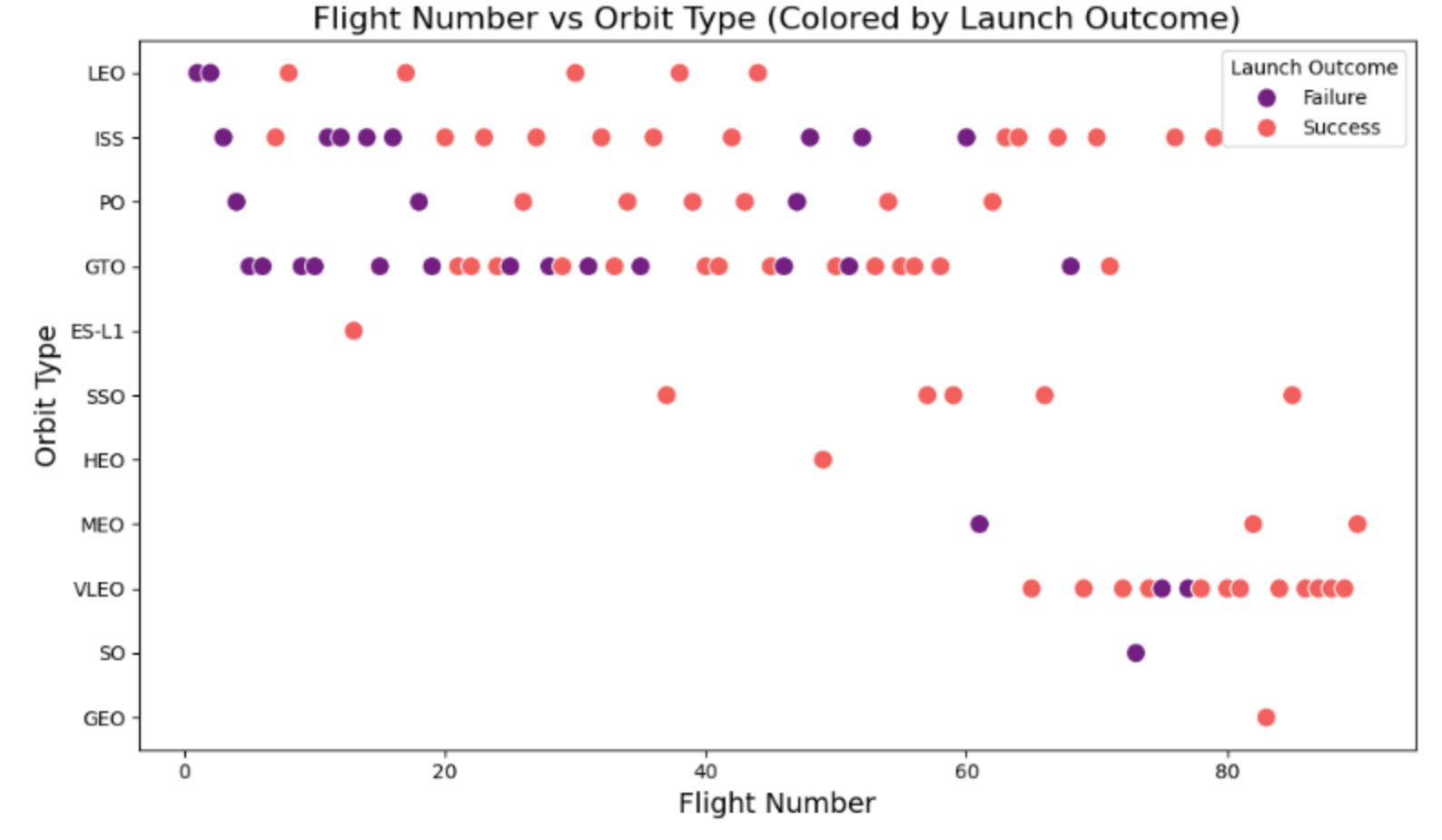
- The scatterplot shows the CCAFS SLC 40 location has a wider range of payload masses when compared to other 2 locations
- VAFB SLC 4E location has a larger number of payload masses near 10,000 kilograms
- CCAFS SLC 40 and KSC LC39A locations have more payload masses near 16,000 kilograms

Success Rate vs. Orbit Type



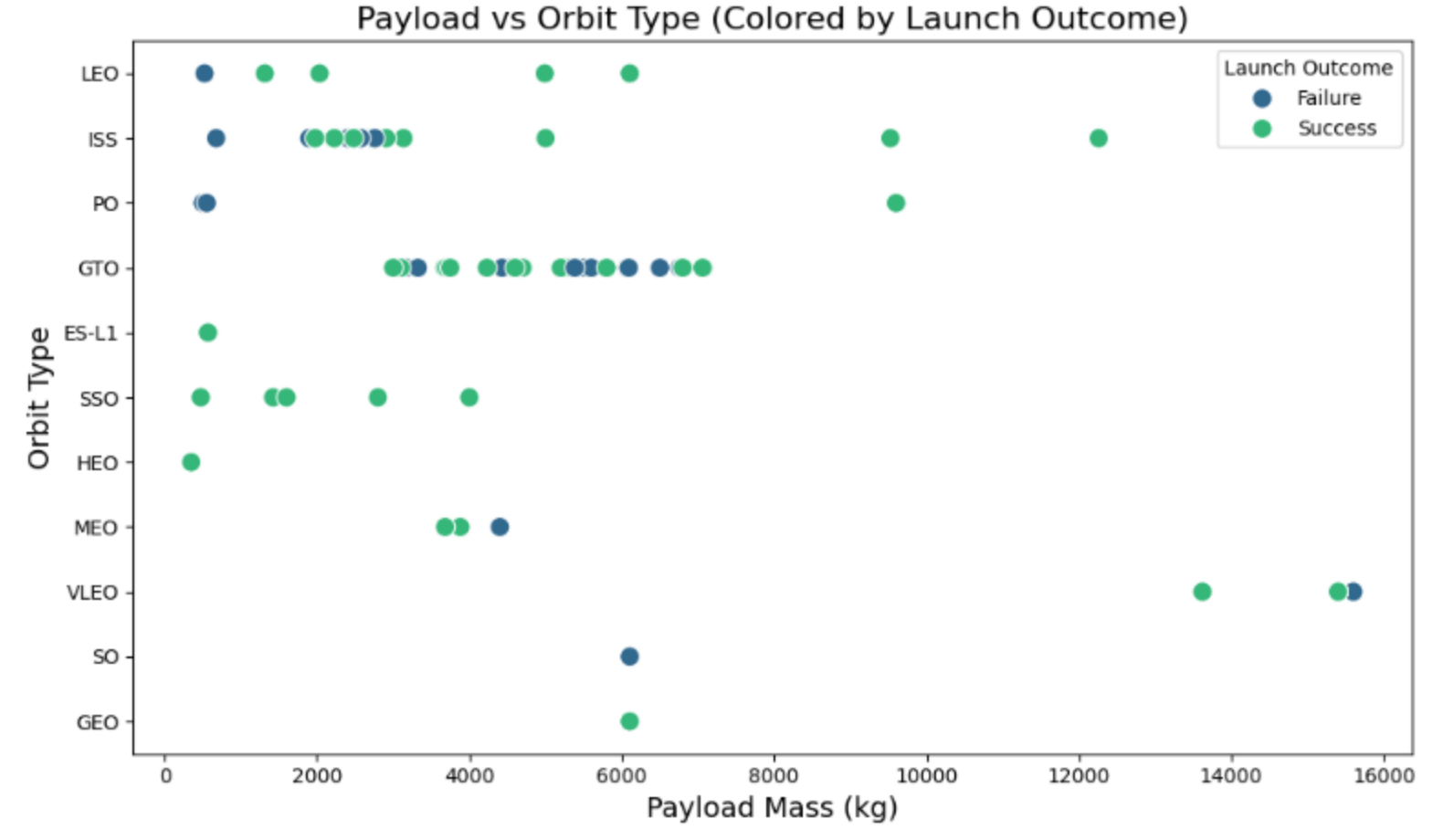
- ES-L1, GEO, HEO, and SSO orbit types have a 100% success rate which shows that the orbit type can be a key indicator of successful landing result.

Flight Number vs. Orbit Type



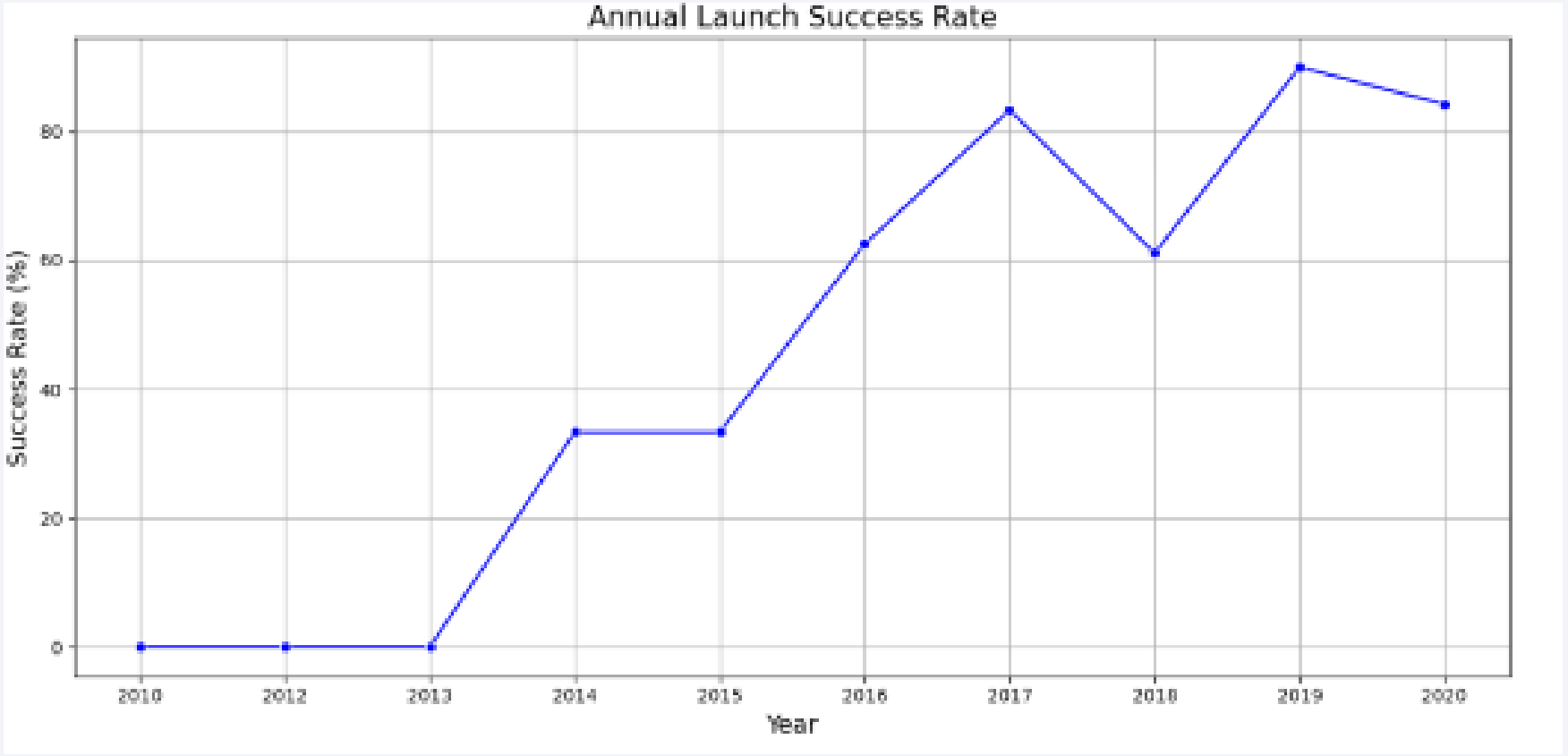
- Initial (lower) launch numbers were more likely to fail which indicates an overall improvement as more tests were done
- GTO has some issues that persist even as more launches are done which indicates there might be an additional factor affecting results.

Payload vs. Orbit Type



- GTO remains a troubled orbit type even though it's payload has been releatively consistent. So, that orbit type is more of a facotr than the payload mass.
- Other orbit types had success with varying payload masses.
- Most launches were done with payloads less than 8,000 kilograms

Launch Success Yearly Trend



- The overall trend shows that the success rate has improved across the years.
- There is some slight volatility as shown from the 2018 results but still strong growth that should remain in the 80-100 range going forward.

All Launch Site Names

Task 1

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

```
%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'

Task 2

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

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

Task 3

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

SUM("PAYLOAD_MASS__KG_")
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 SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

MIN("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 DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AN
```

```
* 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

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTABLE WHERE "Mission_Outcome" IN ('Success',
```

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

Mission_Outcome	Total
Success	98

Boosters Carried Maximum Payload

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MA
```

```
 * 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

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql

SELECT
    CASE
        WHEN substr("Date", 6, 2) = '01' THEN 'January'
        WHEN substr("Date", 6, 2) = '02' THEN 'February'
        WHEN substr("Date", 6, 2) = '03' THEN 'March'
        WHEN substr("Date", 6, 2) = '04' THEN 'April'
        WHEN substr("Date", 6, 2) = '05' THEN 'May'
        WHEN substr("Date", 6, 2) = '06' THEN 'June'
        WHEN substr("Date", 6, 2) = '07' THEN 'July'
        WHEN substr("Date", 6, 2) = '08' THEN 'August'
        WHEN substr("Date", 6, 2) = '09' THEN 'September'
        WHEN substr("Date", 6, 2) = '10' THEN 'October'
        WHEN substr("Date", 6, 2) = '11' THEN 'November'
        WHEN substr("Date", 6, 2) = '12' THEN 'December'
        ELSE 'Unknown'
    END AS "Month_Name",
    "Mission_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTABLE
WHERE
    substr("Date", 0, 5) = '2015';
```

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

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	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(*) AS "Count"
FROM
    SPACEXTABLE
WHERE
    "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    COUNT(*) DESC;
```

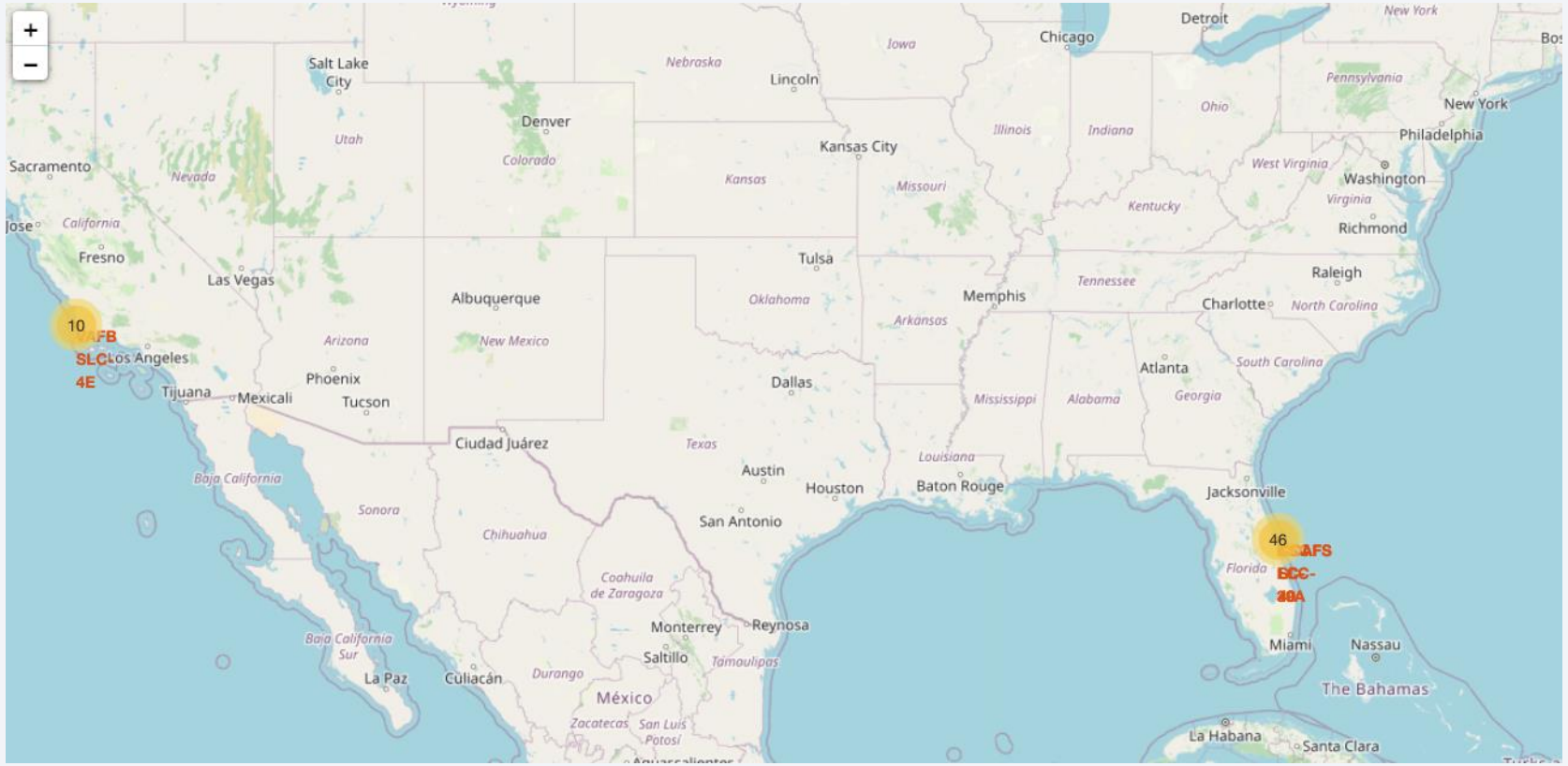
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

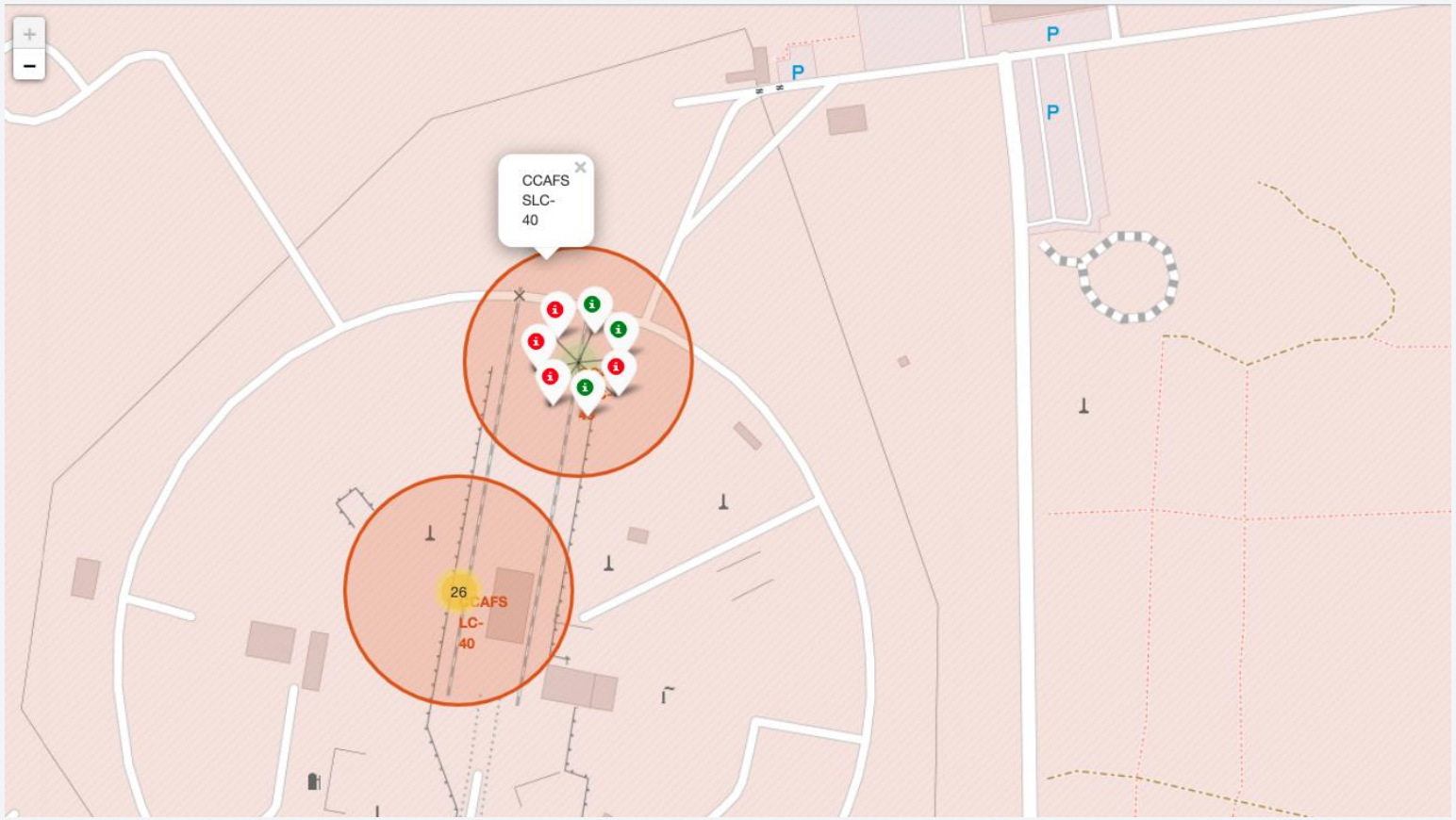
Launch Sites Proximities Analysis

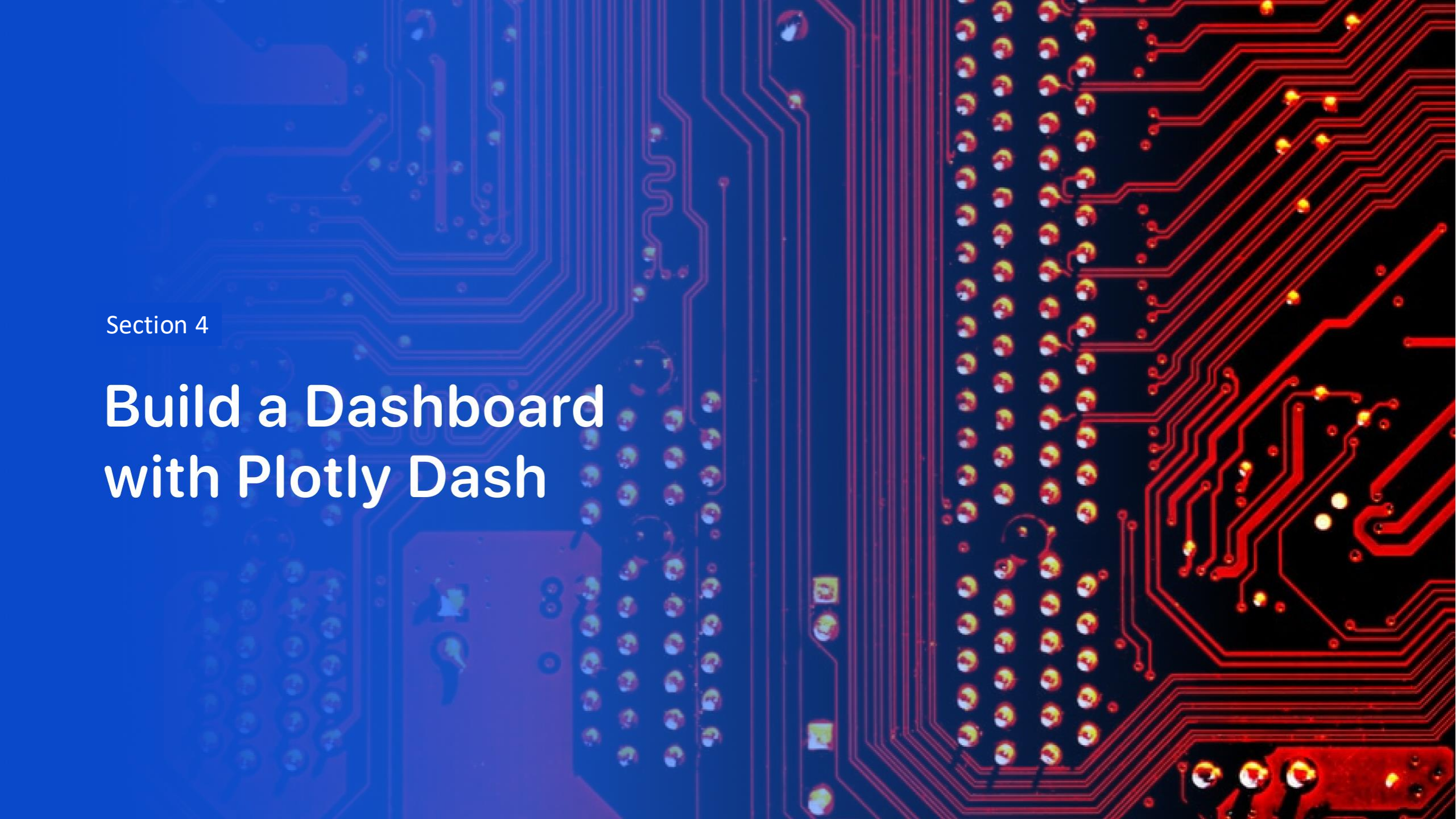
Folium Launch Locations Map



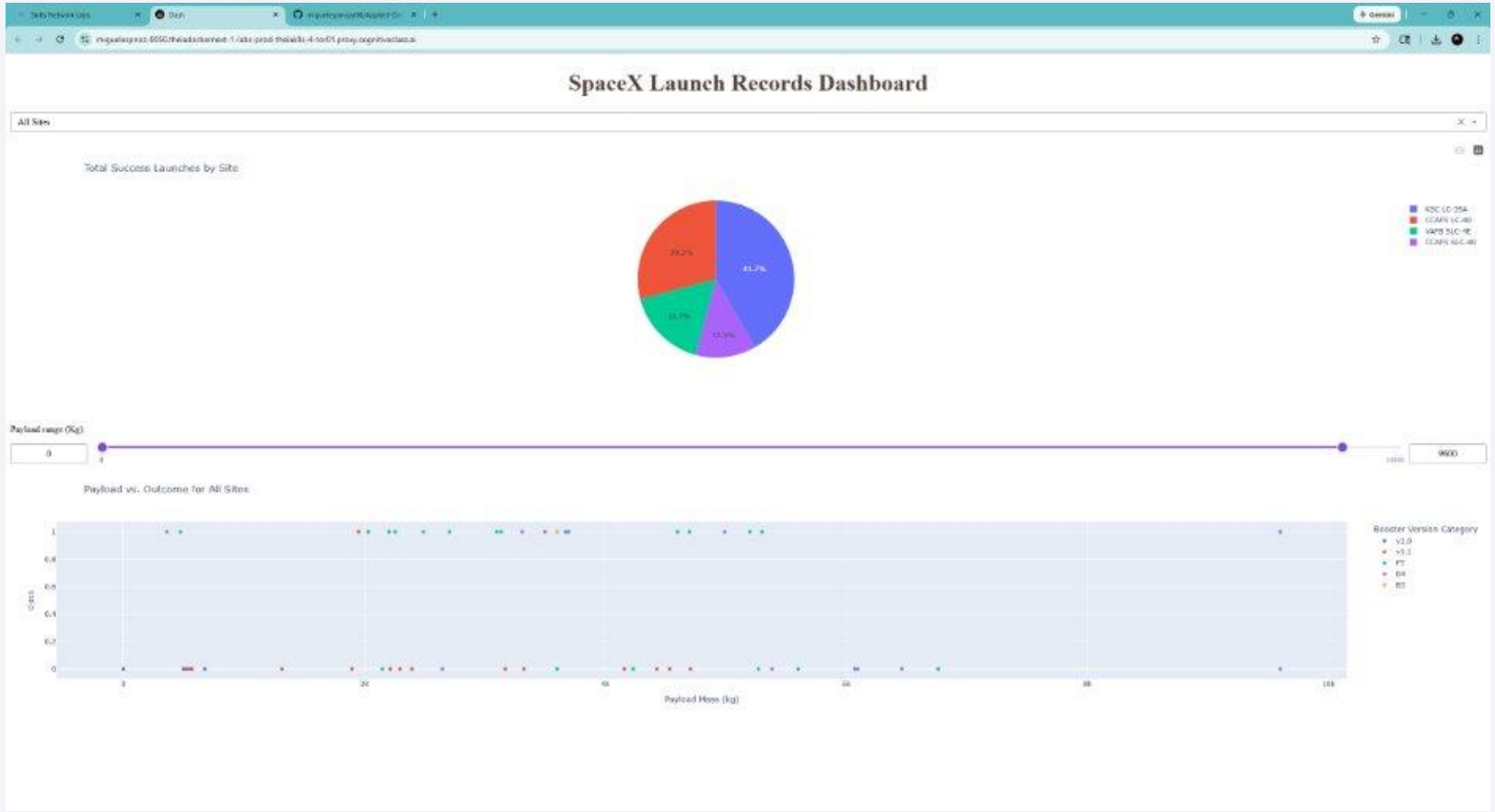
- Launch sites range from coast to coast of United States. This shows that proximity to a coastline/large body of water is ideal.

Folium Launch Success Map



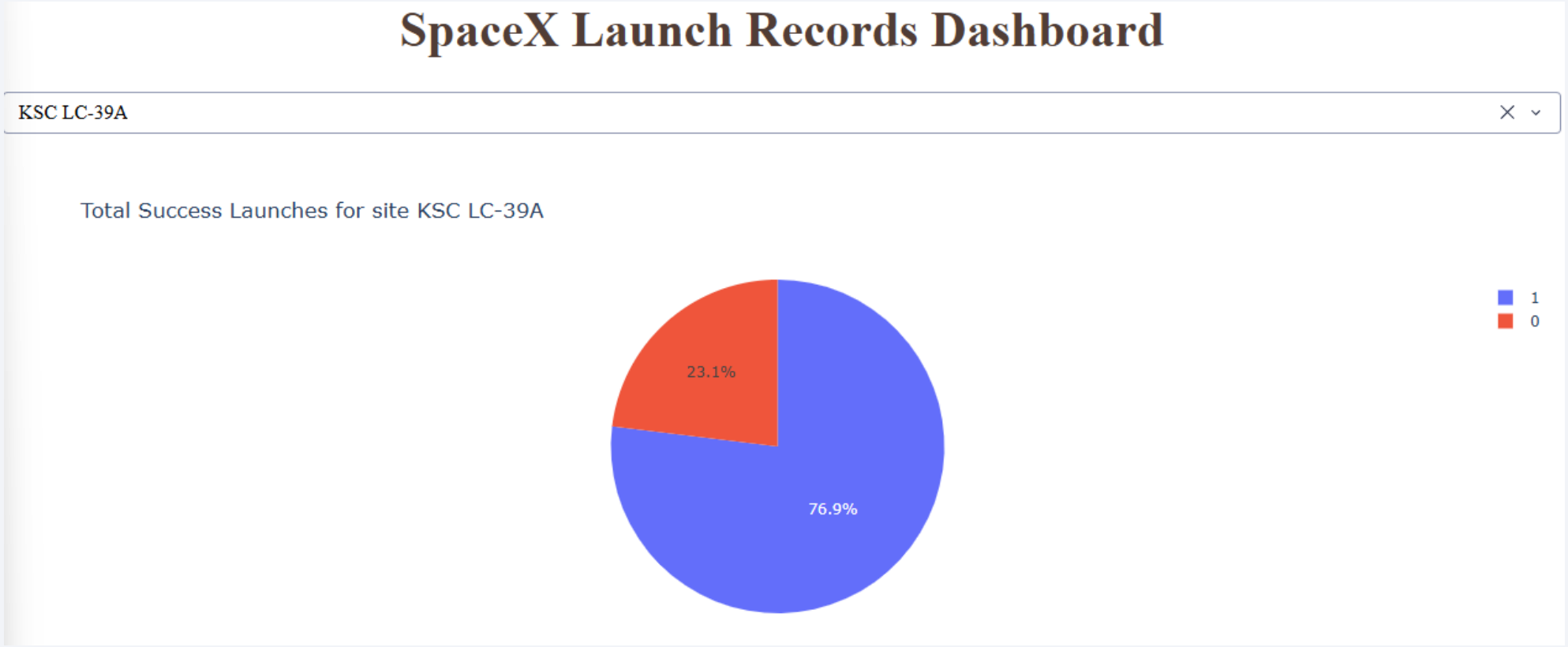

Section 4
Build a Dashboard with Plotly Dash

Plotly Dash Interactive Dashboard



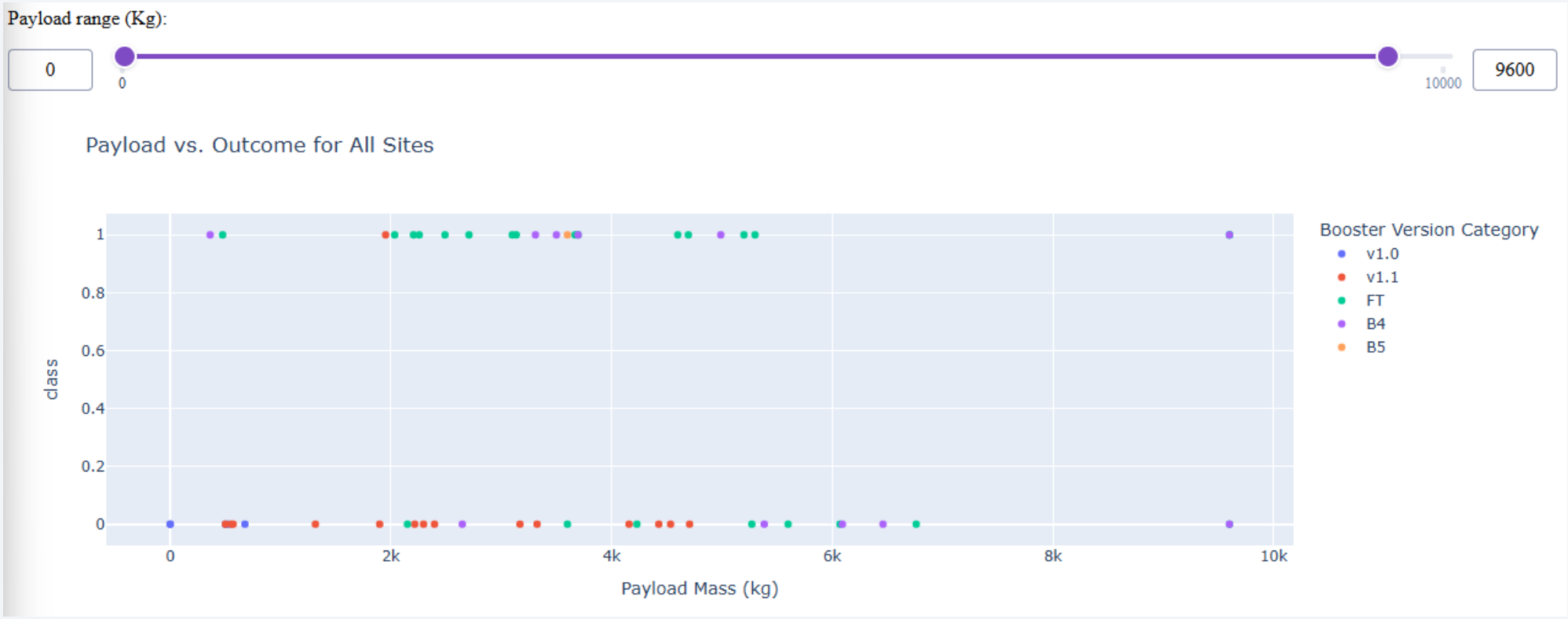
- KSC LC-39A is the leading location by success rate.
- CCAFS LC 40, VAFB SLC-4E, CCAFS SLC-40 follow in terms of success rate

Leading Location by Success Rate Analysis



- KSC LC-39A is the leading location by success rate
- This location reports a 76.9% success rate with various payload masses and orbit types being tested.

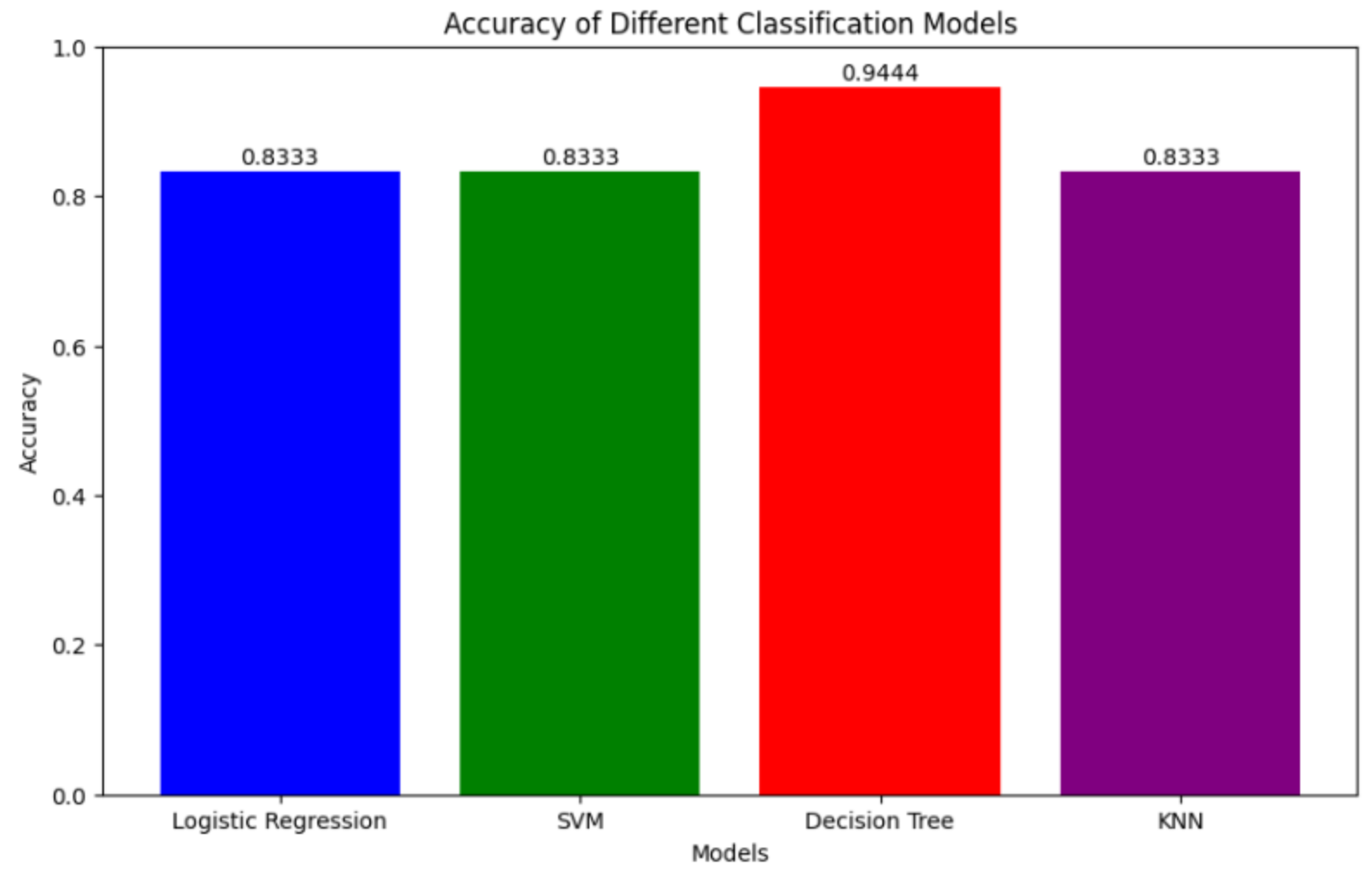
Payload vs. Launch Outcome Scatter Plot



Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- This confusion matrix shows that the Decision Tree returned no false negatives with enables us to feel secure in its future findings
- False positives are limited and overall shows that this is a strong choice
- Decision Tree model was able to have a 94.44 % accuracy

Conclusions

- This confusion matrix shows that the Decision Tree returned no false negatives with enables us to feel secure in its future findings. False positives are limited and overall shows that this is a strong choice. Decision Tree model was able to have a 94.44 % accuracy
- Payload mass was shown to not have a overreaching effect on the success of landings.
- Location data and orbit type are strong factors in being able to predict if a launch will land successfully.

Thank you!