IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Ali Ford>
<17/01/2025>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix
- GITHUB URL: https://github.com/DSsuperguy/IBM-Capstone

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Interactive map with Folium
- Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- EDA results
- Interactive analytics results
- Predictive analysis results

Introduction

Leading commercial space company offering relatively affordable launches Falcon 9 launches cost $62M vs. competitors' $165M. Cost savings largely due to reusable first stage Falcon 9's first stage does most work, is largest and most expensive.

Working as data scientist for "Space Y" Goal: Predict launch prices and first stage reusability Method: Using machine learning and public data instead of rocket science. Will create dashboards and analyze SpaceX data

The problem we are trying to solve is determining the cost of rocket launches?
Predicting whether SpaceX's Falcon 9 first stage will successfully land and be reused?

By solving this, we can estimate launch costs accurately and help a new rocket company, "Space Y," compete with SpaceX by leveraging insights into reusability and cost efficiency.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 rest API

Initialize API Request and Data Import

Data Parsing and Initial Cleaning

Extract Detailed Information via API Calls

Construct and Refine Dataset

Filter and Save the Dataset

SpaceX rest API

Setup and Import Libraries

Retrieve and Parse HTML

Extract Table Structure

Scrape and Process Data

Convert to DataFrame and Save

Data Collection – SpaceX API

Import necessary libraries for API interaction and data manipulation.

↓

Fetch launch data from SpaceX API using GET request.

↓

Convert JSON response to a pandas DataFrame.

↓

Clean DataFrame by filtering for single-core and single-payload launches.

↓

Extract detailed information with helper functions and additional API calls.

↓

Assemble detailed data into a new DataFrame, manage missing data.

↓

Filter for Falcon 9 launches, reset flight numbers, and save to CSV.

https://github.com/DSsuperguy/IBM-Capstone

Data Collection - Scraping

Install and import required libraries.

↓

Fetch the static Wikipedia page for launch records.

↓

Parse HTML content using BeautifulSoup.

↓

Extract column names from the table header.

↓

Scrape data from table rows.

↓

Convert scraped data into a pandas DataFrame.

↓

Save the DataFrame to a CSV file.

Data Wrangling

Add 'Class' column to indicate landing success.

↓

Display first 8 rows of 'Class' column to verify data.

↓

Show the first 5 rows of the DataFrame to confirm structure.

↓

Calculate the success rate of landings using the mean of 'Class'.

↓

Interpret the result where 0.6667 means about 67% success rate.

↓

Prepare data for export, ensuring consistency for future labs.

↓

Export the DataFrame to a CSV file for further analysis.

EDA with Data Visualization

Scatter charts were used to visualise the relationships of the following:

Flight Number VS Launch Site
Payload Mass VS Launch Site
Success rate of each orbit type
FlightNumber vs Orbit type
Payload Mass vs Orbit type

A single Plot line chart was used to visualise:

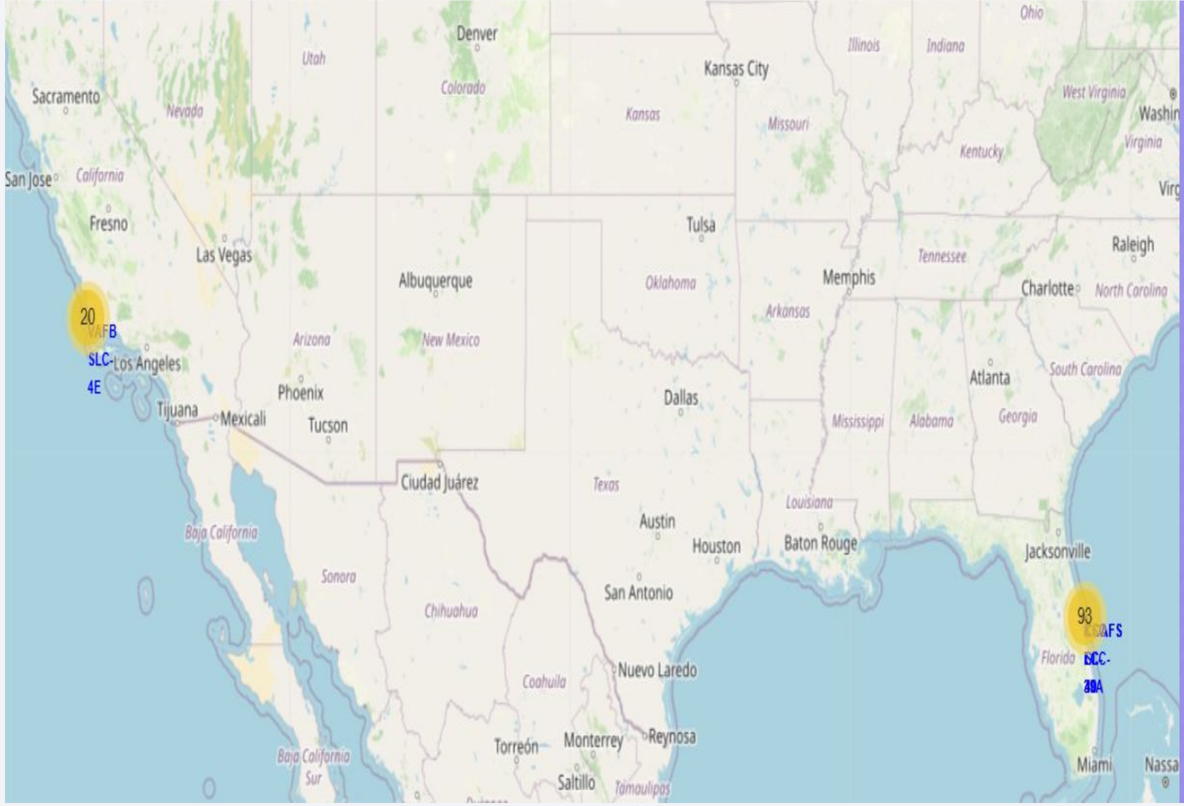
The launch success yearly trend

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Display month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 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.

Build an Interactive Map with Folium

Used to display all the takeoff sites

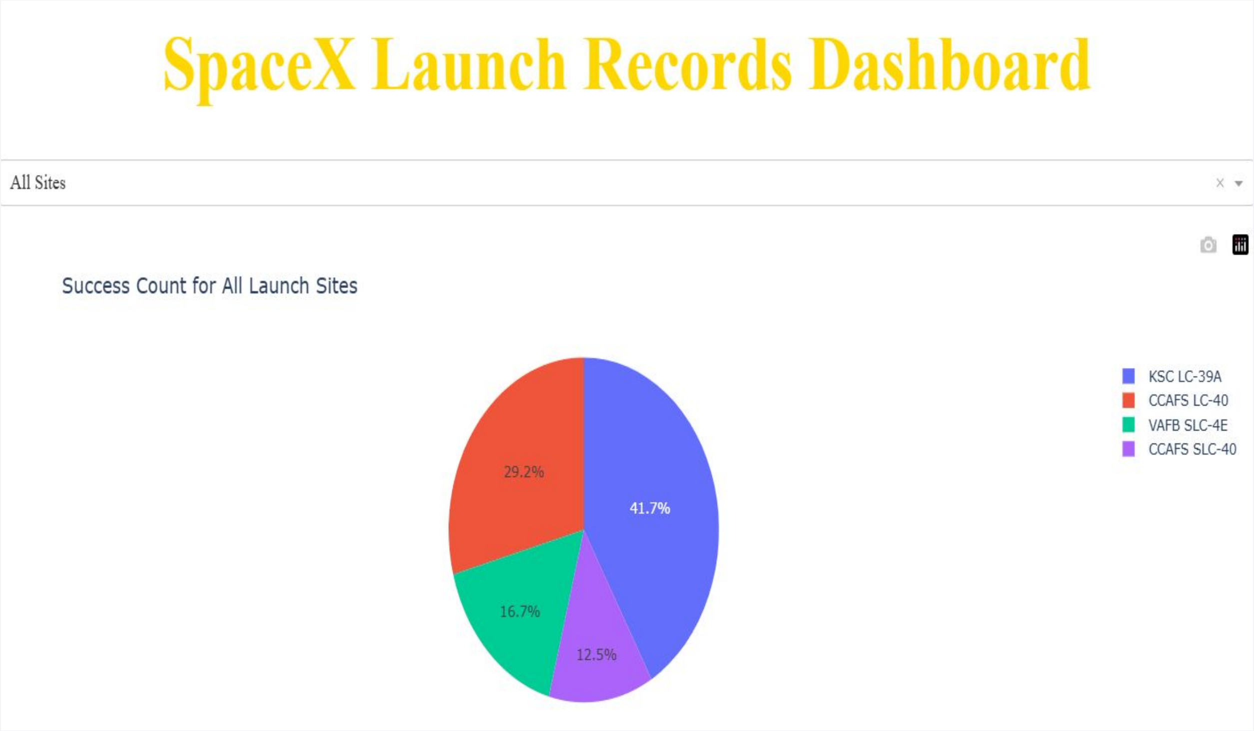


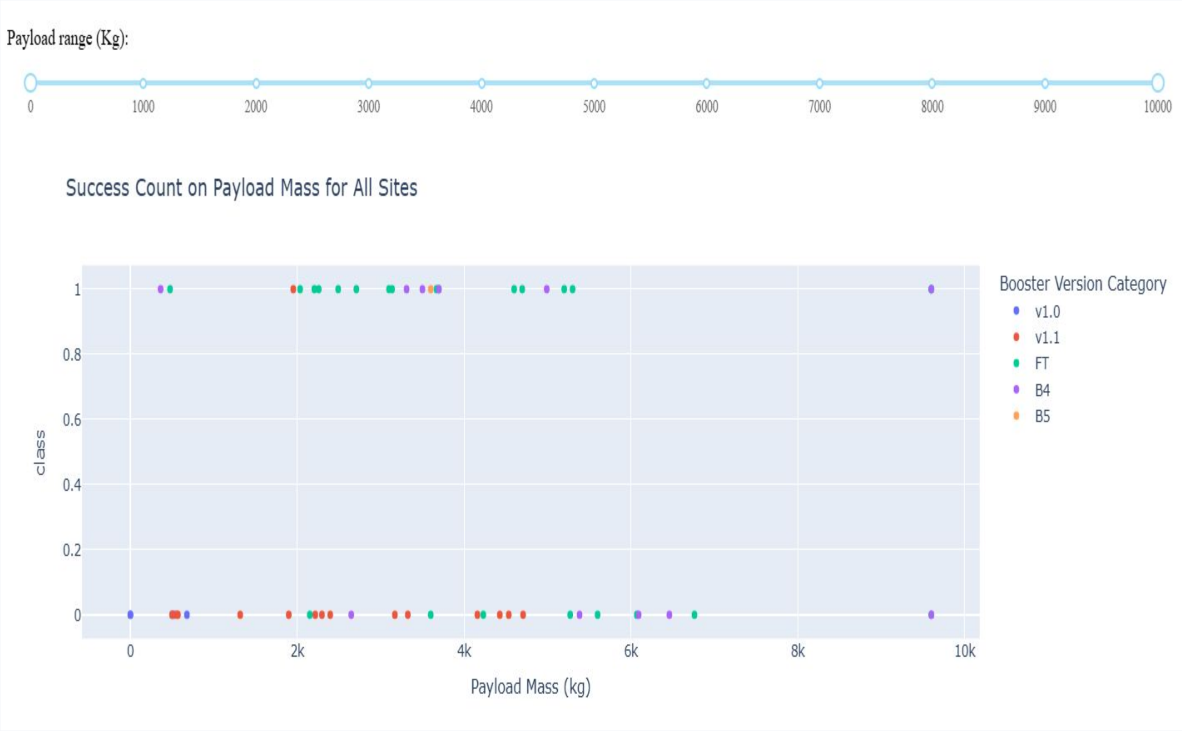
Used to display distance to the nearest coast

Build a Dashboard with Plotly Dash

Dropdown menu with a pie chart to show the success launches of all and or each site.

Scatter plot using a Payload Range Slider to show the successful launches of all or by each site by payload mass.





Predictive Analysis (Classification)

Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) models are created and optimized using the GridSearchCV method to determine the best parameters. Once the optimal parameters are identified, the models are trained on the training dataset.

The accuracy on the test dataset is evaluated for each model. Among them, the Decision Tree model achieves the highest accuracy at 0.87, followed by Logistic Regression, SVM, and KNN, each with an accuracy of 0.84.

“Logistic regression” has an accuracy : 0.8464285714285713
“Support vector machine” has an accuracy of 0.8482142857142856
“Decision Tree” has an accuracy : 0.875
“k nearest neighbors” has an accuracy : 0.8482142857142858

Results

Exploratory data analysis results

- Success rate of launches increases over time
- KSC LC-39A has the highest success rate among landing sites
- GEO,HEO,SSO,ES L1 orbit types has highest success rate.

Interactive analytics demo in screenshots

- All sites are near coastal regions, but still near towns and cities

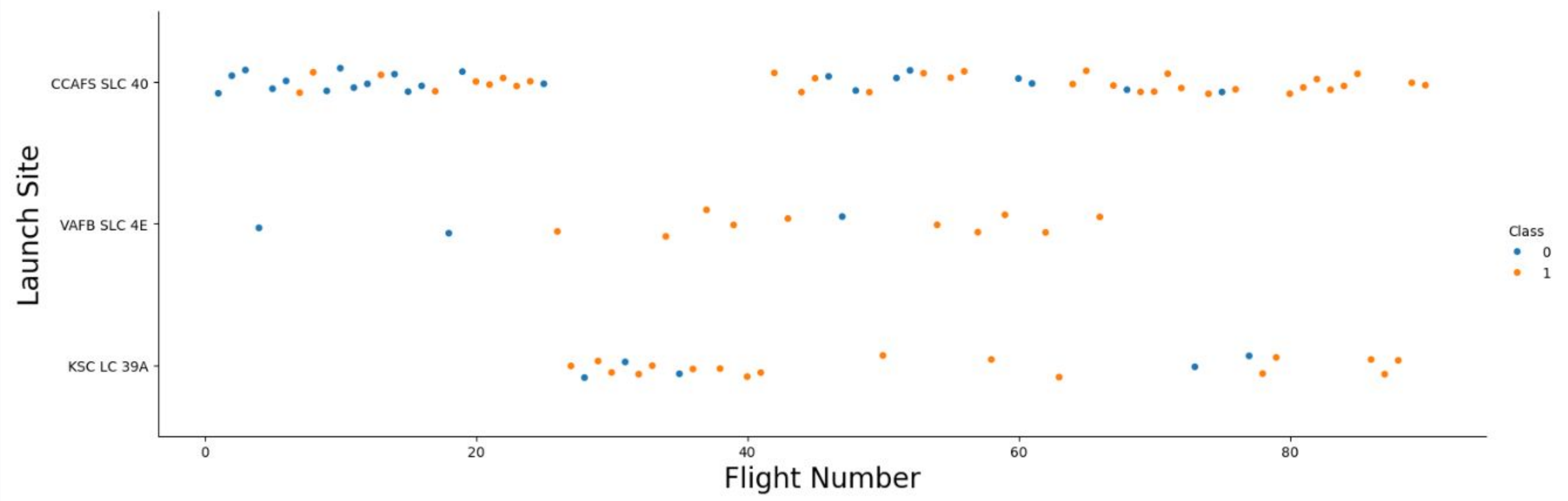
Predictive analysis results

- Decision tree model was the best performing one.

Section 2

Insights drawn from EDA

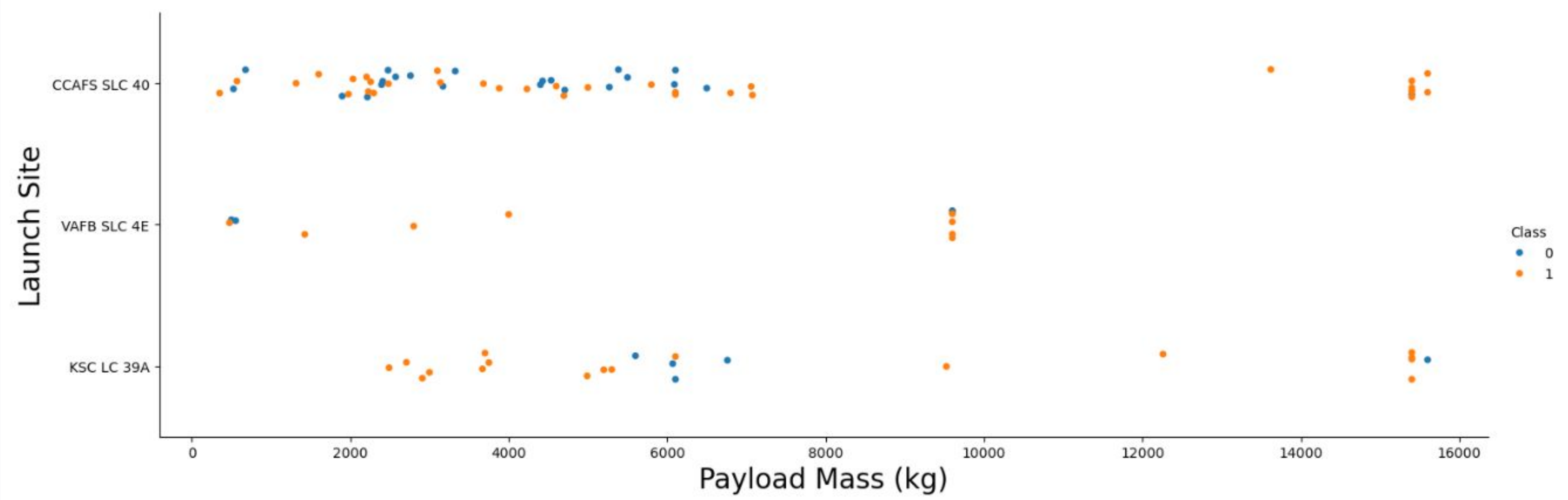
Flight Number vs. Launch Site



We can see both the failed and successful launches. Most of the failures were early on, but become completely outnumbered by the successes, past flight 20.

Although CCAFS SLC 40 has the most total flights, VAFB SLC 4E has the fewest failures.

Payload vs. Launch Site

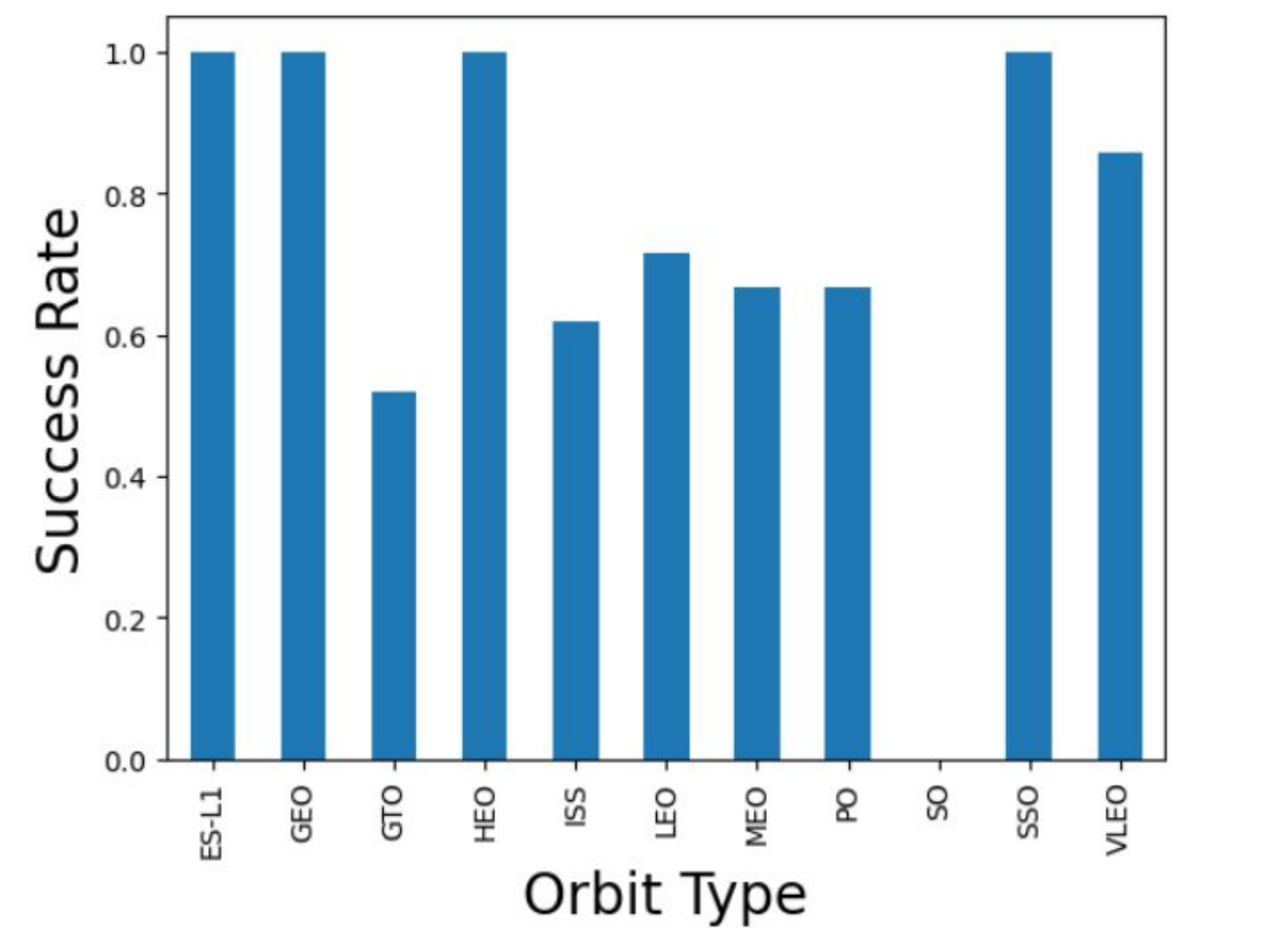


In each launch site the higher the payload mass, the higher the success rate. In addition almost all launches with >7000 payload mass was successful.

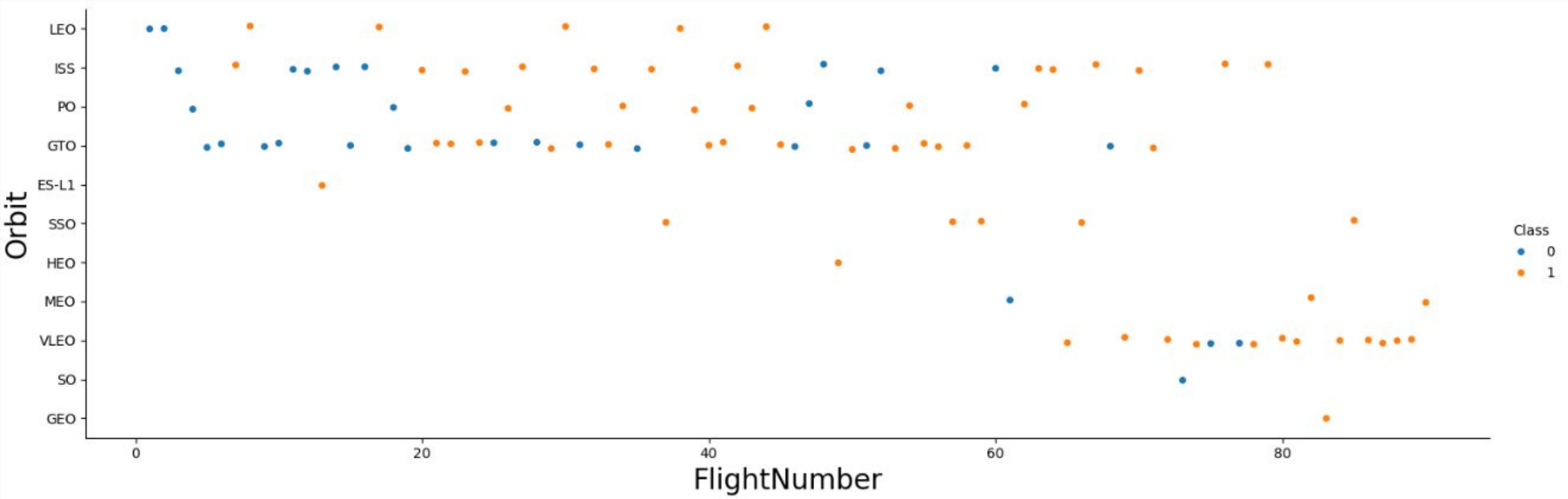
Success Rate vs. Orbit Type

ES-L1, GEO, HEO, and SSO orbits has the highest success rates.

The rest was more mixed. However SO had a 0% success rate.

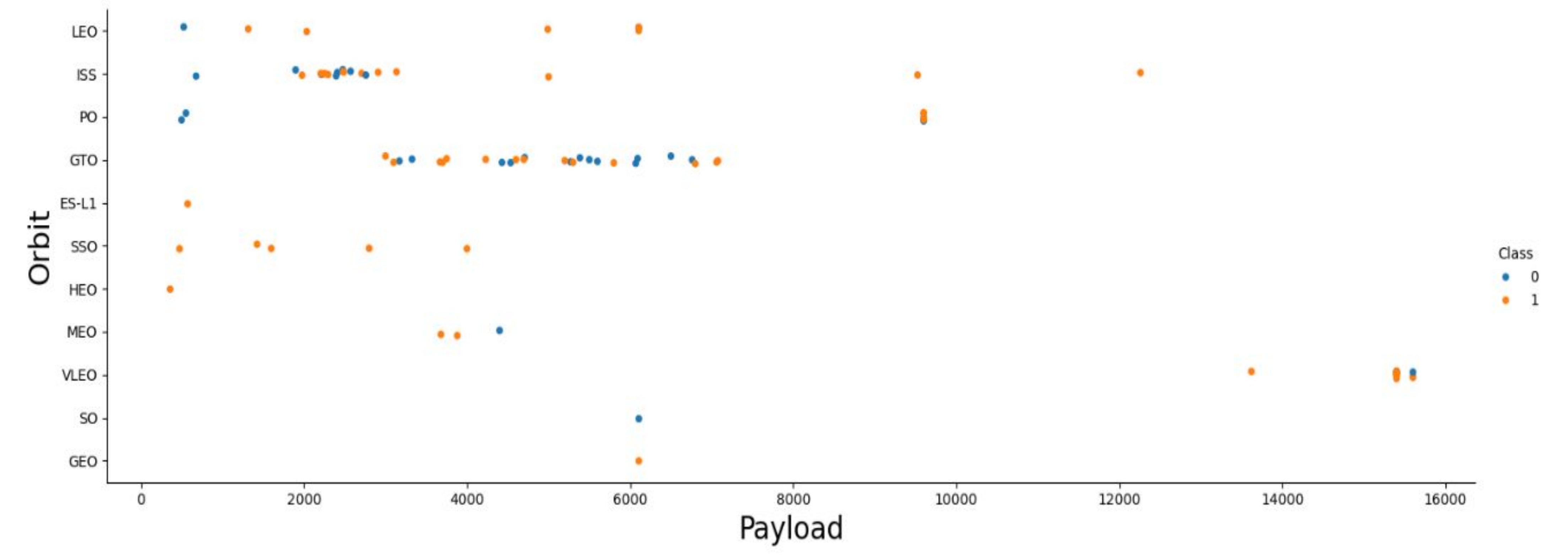


Flight Number vs. Orbit Type



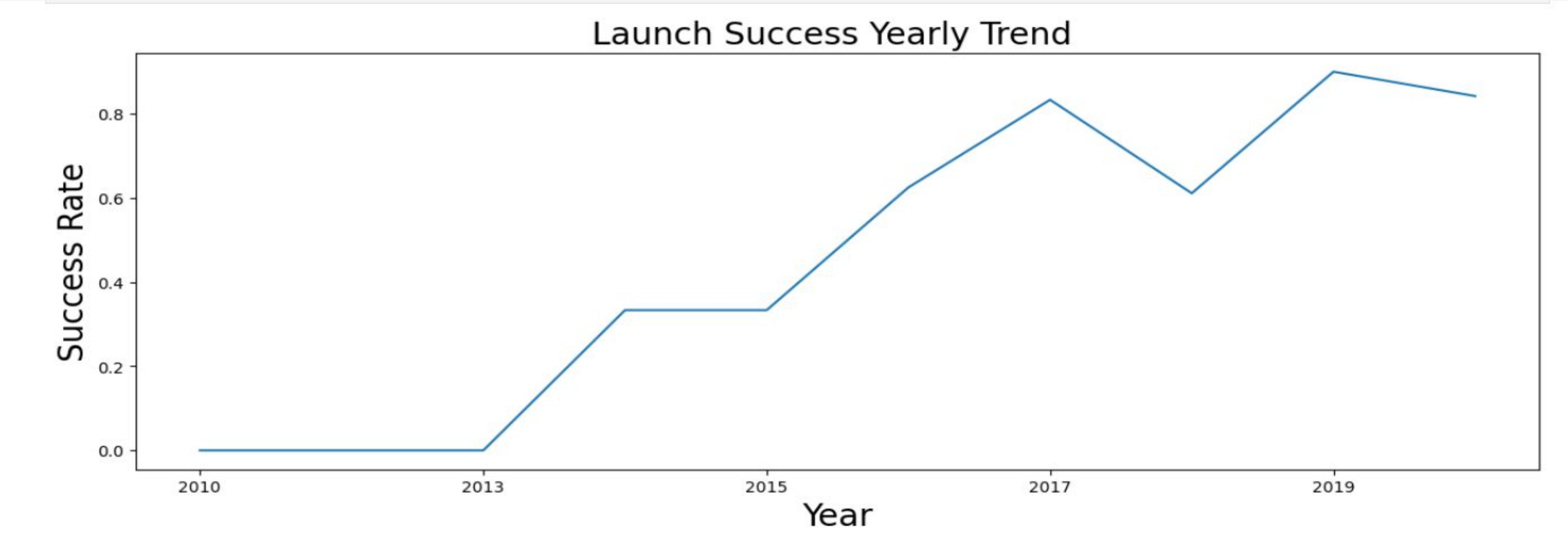
Past flight 50 there are seldom flight\ failures failures, except for SO which has 100% failure rate. Past Flight 60 VLEO is the dominate Orbit type.

Payload vs. Orbit Type



There is no much of an obvious correlation between payload mass and orbit type. There appears to be fewer failures when payload is >8000.

Launch Success Yearly Trend



Overall the Success rate kept on increasing from 2013 to 2020.

All Launch Site Names

```
[10]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;

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

[10]:

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

Retrieved all the unique Launch sites

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL 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	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

Retrieved the first 5 records where launch sites name begins with the string 'CCA'.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

I queried for the total payload mass carried by boosters launched from NASA (CRS) which is 45,596 kg.

Average Payload Mass by 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

Found out the average payload mass carried by booster version F9 v1.1 is 2928.4kg.

First Successful Ground Landing Date

```
%sql select min(Date) as "Min Date" from SPACEXTBL where Landing_Outcome = "Success (ground pad)";
```

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

Min Date
2015-12-22

Retrieved the date of the first successful landing outcome from ground pad..

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4001 and 5999;
```

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

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

Here we got the Booster Verions of which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) as Frequency from SPACEXTBL group by Mission_Outcome;
```

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

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

Here we can see the total number of Successful Outcomes from the missions. All successful except for one.

Boosters Carried Maximum Payload

```
%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

We we can see that there were 12 boosters carrying the maximum payload.

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

Retrieved the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

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

Ordered the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 by Rank in descending order. “No attempt” had the highest count.

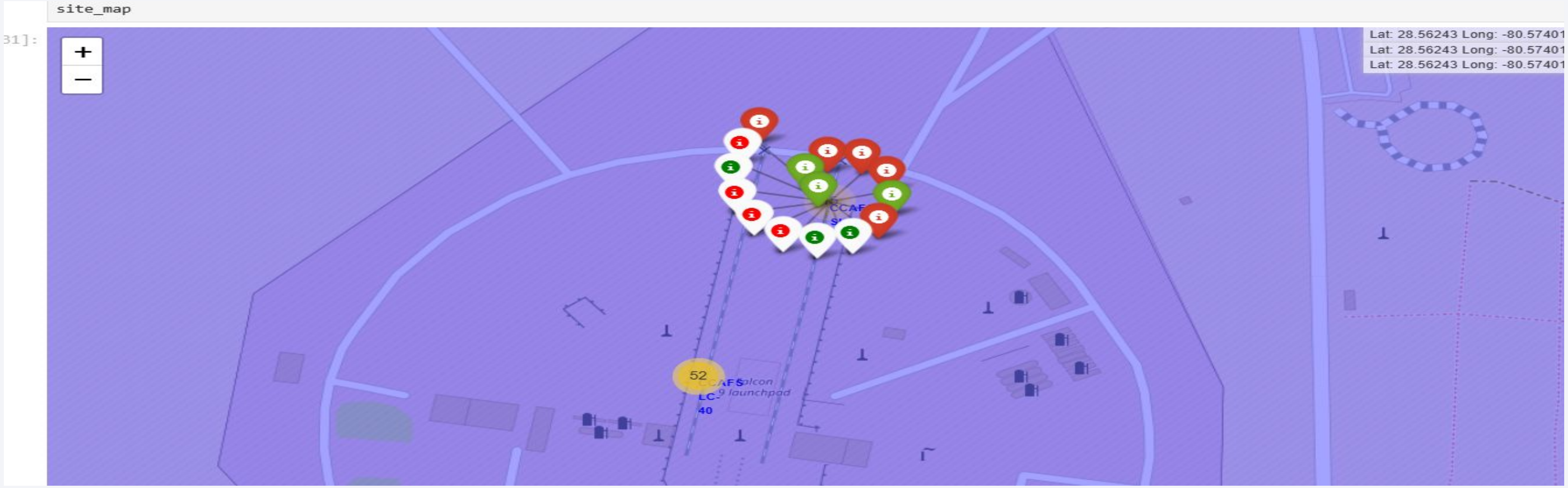
Section 3

Launch Sites Proximities Analysis

All Launch Site locations

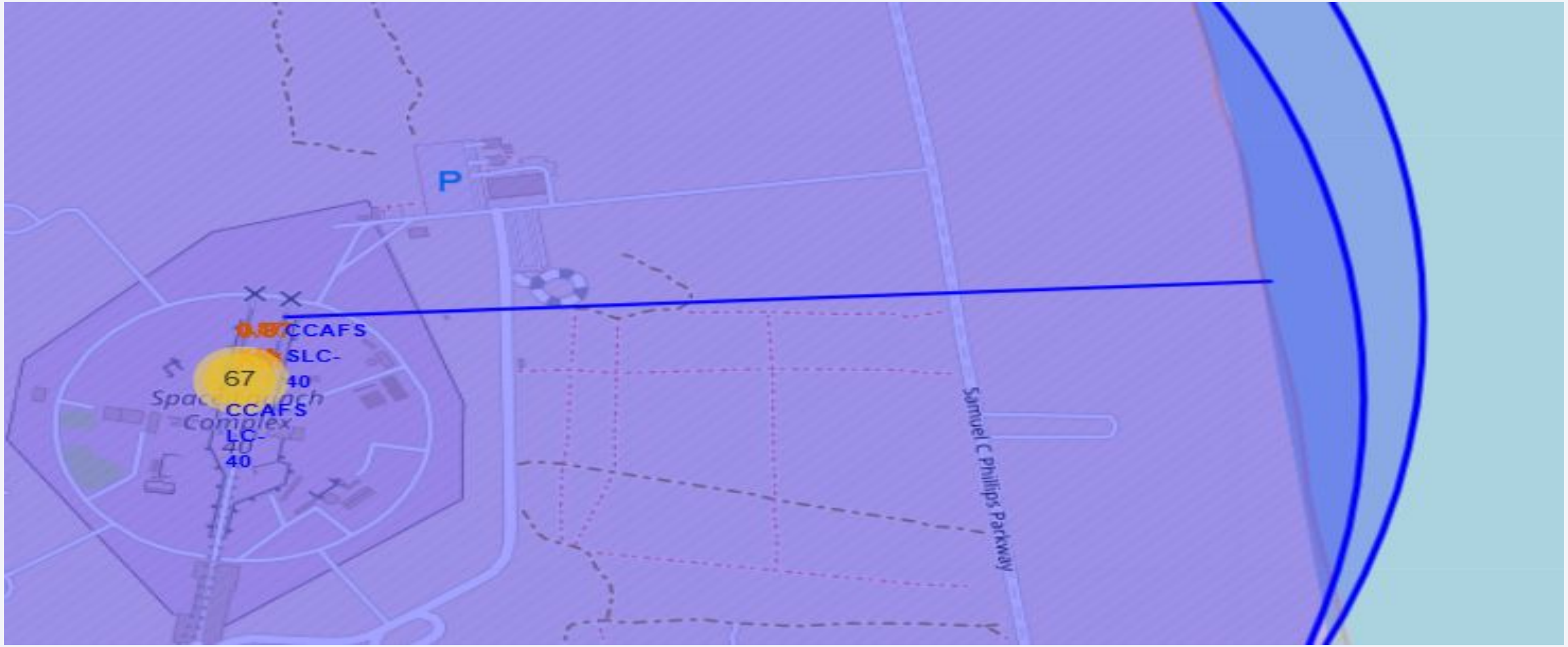
There are four launch sites with CCAFS LC-40 and being right CCAFS SLC-40 next to each other in florida

All success/failed launches for each site on the map



The launch site attempts are labelled by green markers for successful launches, and red markers for unsuccessful ones. The above is for CCAFS SLC-40 for example

Launch Site Proximities

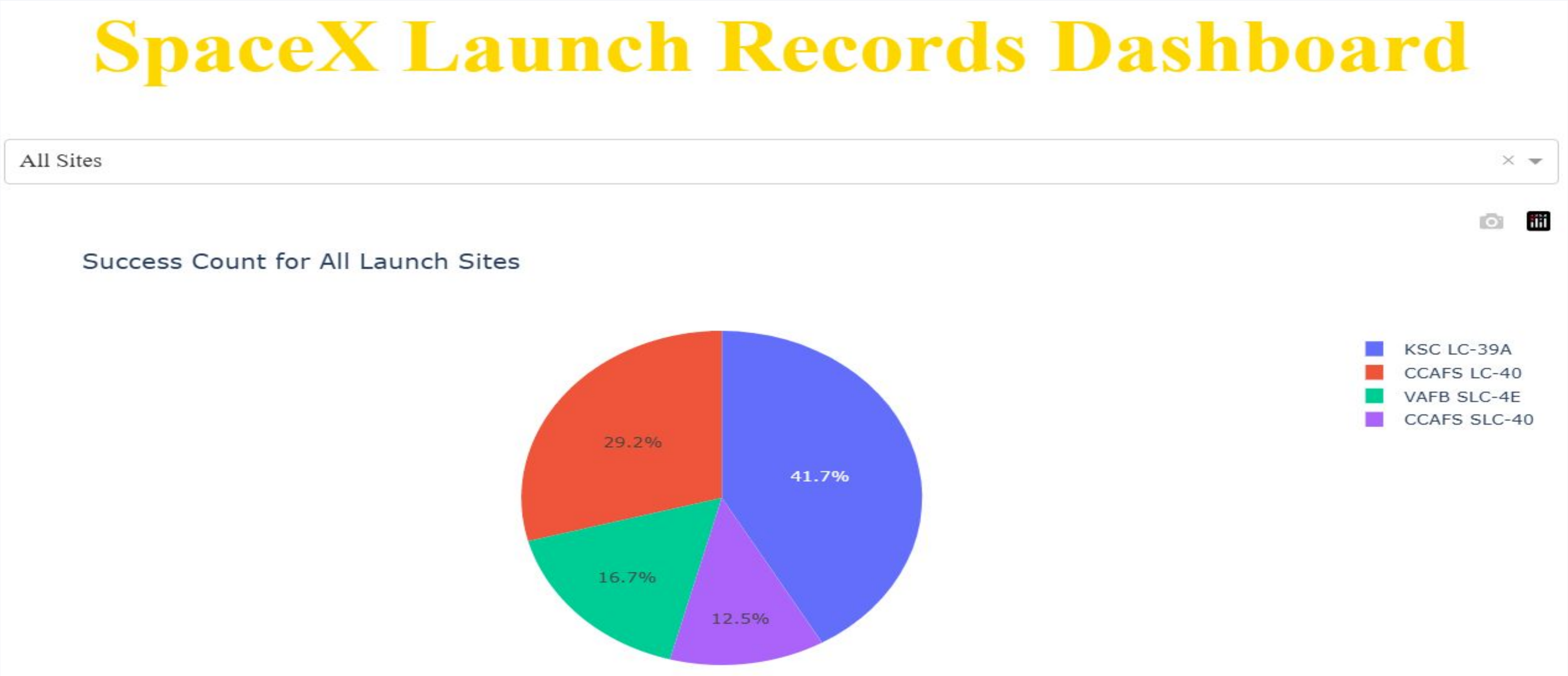


Launch Site CCAFS SLC-40 is 0.87 km from the coastline from the east.

Section 4

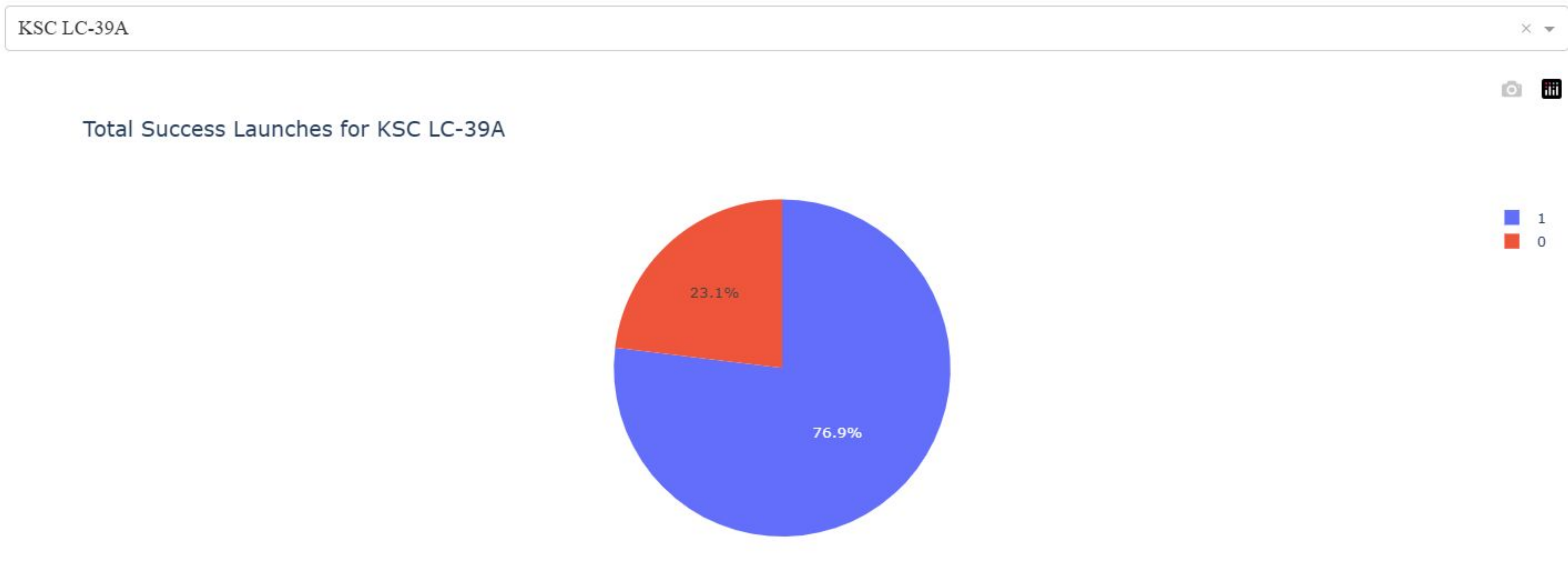
Build a Dashboard with Plotly Dash

Total success launches for all sites



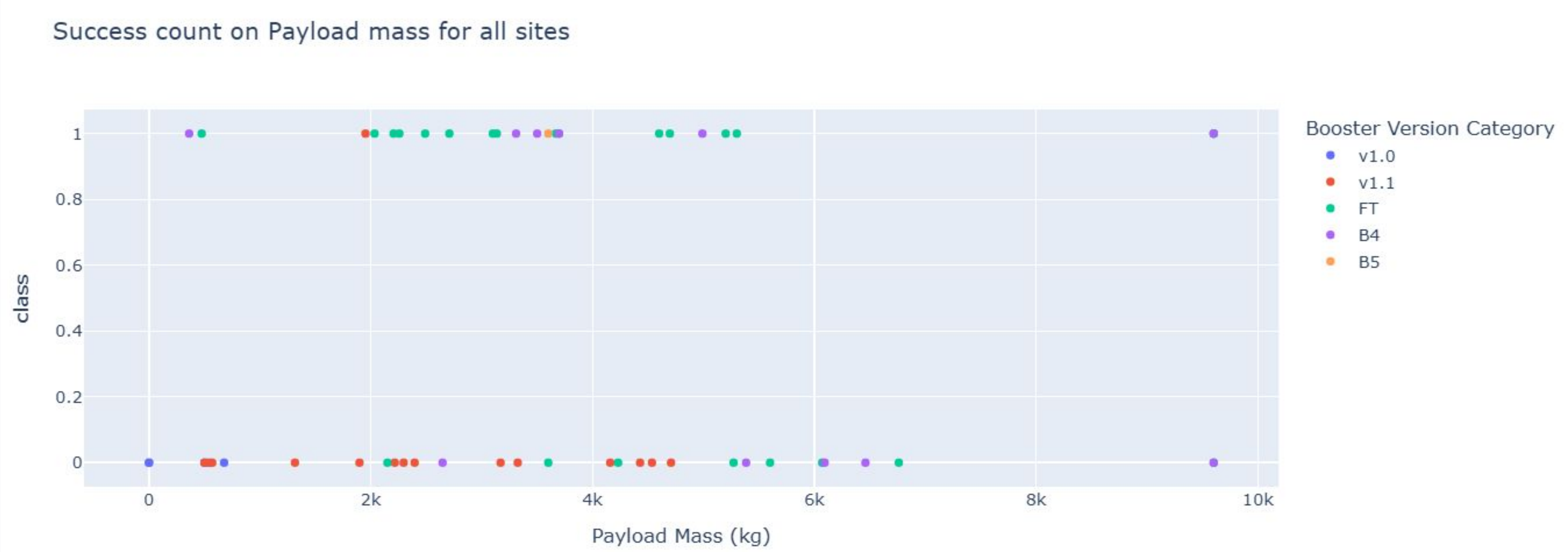
KSC LC-39A with 41.7% is the most successful, whereas CCAFS SLC-40 is the least successful at 12.5%.

Launch site with the highest success ratio



KSC LC-39A has 76.9% successes and only 23.1% failures.

Payload vs. launch outcome

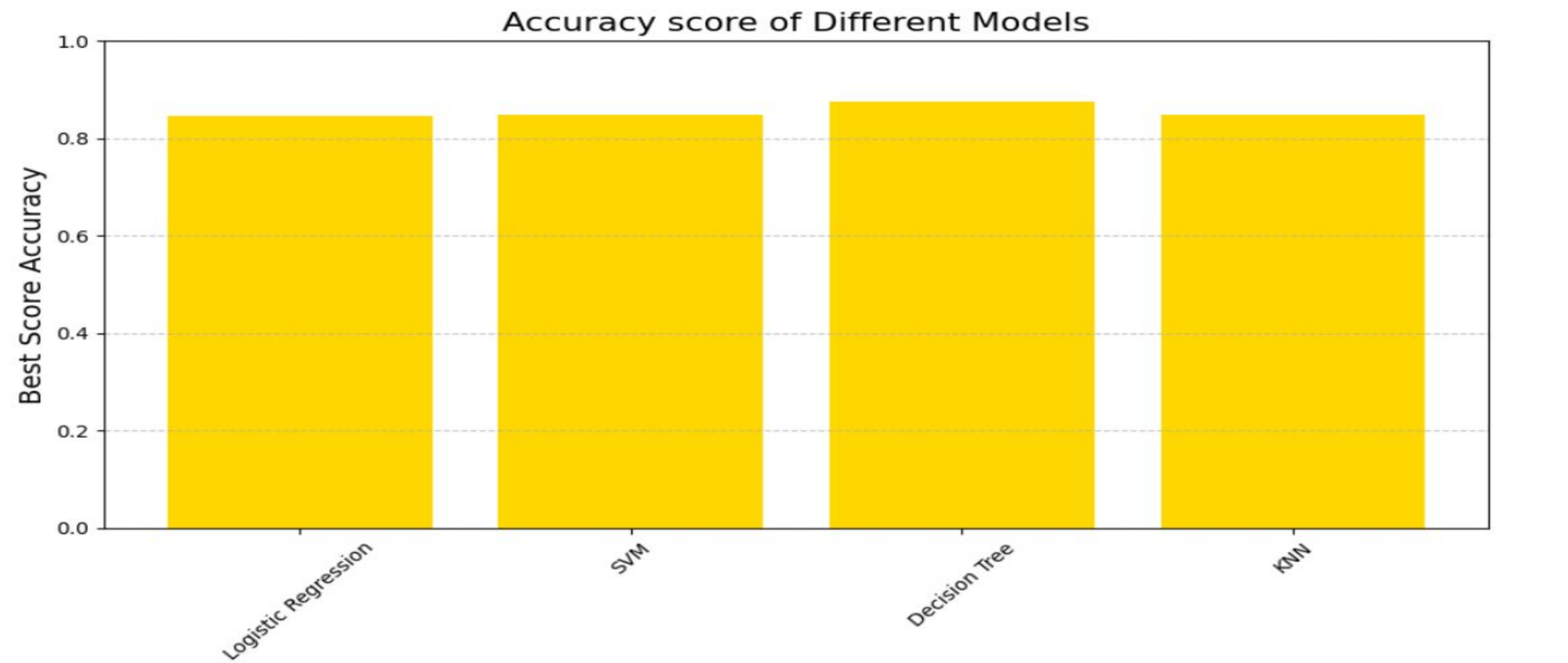


Overall the most successful launches are between 2K and 4K in Payload mass followed by ranges between 4K and 6K. v1.1 and FT appear to be the most common

Section 5

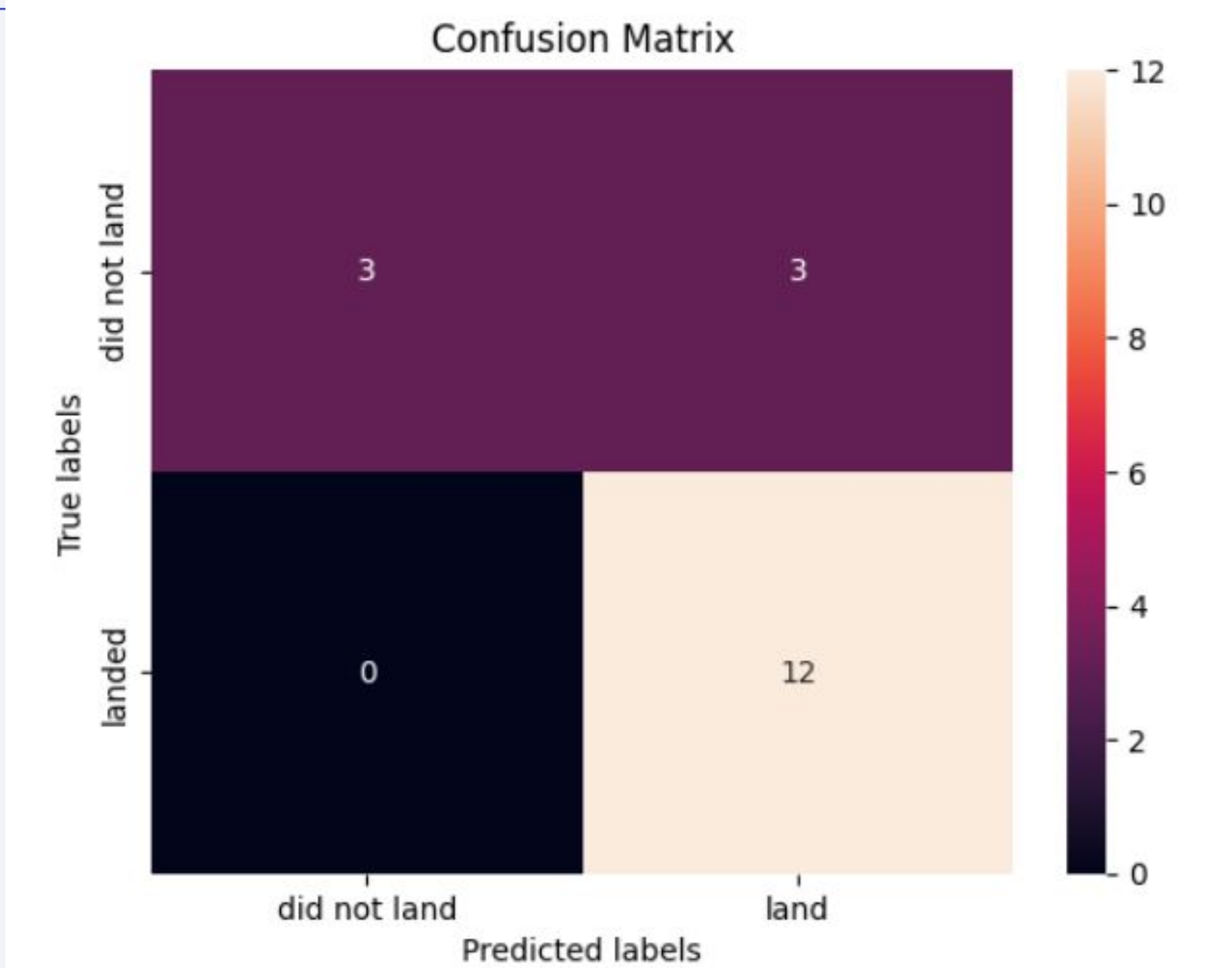
Predictive Analysis (Classification)

Classification Accuracy



The Decision Tree Accuracy was the best at 87.50% of which is the below followed by Logistic Regression Accuracy: 84.64% and then lastly SVM Accuracy and KNN Accuracy with both getting 84.82%

Confusion Matrix for the best model



The Decision Tree model was almost perfect with 3 True Negatives, 0 False negatives and 12 True Positive scores. However it got let down by 3 False positives.

Conclusions

The Success Rate of the launches got higher over time.

The model to use going forwards is the Decision Tree model for predicting the success\ failure of the future launches correctly.

Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

KSC LC-39A had the most successful launches of any sites.

…

Appendix

Github repo: https://github.com/DSsuperguy/IBM-Capstone

Thank you!