IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anastasia
March 3, 2023

Outline

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

Executive Summary

Space Y is a new company to the commercial rocket market. In this market, launching new rockets is costly. American spacecraft manufacturer SpaceX offers a competitive cost through its Falcon 9, a rocket that makes space travel commercially available. It does so at a relatively inexpensive cost ($62m), compared to regular rockets ($165m/rocket), by reusing its first stage. Stage 1 is the more critical stage, thereby providing insightful launch cost information

To understand SpaceY’s competitive positioning against SpaceX, Falcon 9 data was assessed to predict whether the rocket will land successfully to understand cost. For this, Falcon 9 data was collected and analyzed various data science methods. Data visualization was also leveraged to create data-driven decisions. Finally, machine learning models were built find the optimal prediction model.

Introduction

- New spacecraft manufacturer company seeking to enter and compete against SpaceX
- Predict whether Falcon 9 will land in Stage 1 to understand cost
- Understand the landscape like launch sites and launch factors like payload and orbit
- Data analysis to understand and predict Falcon 9 launch success accuracy and dependent variables

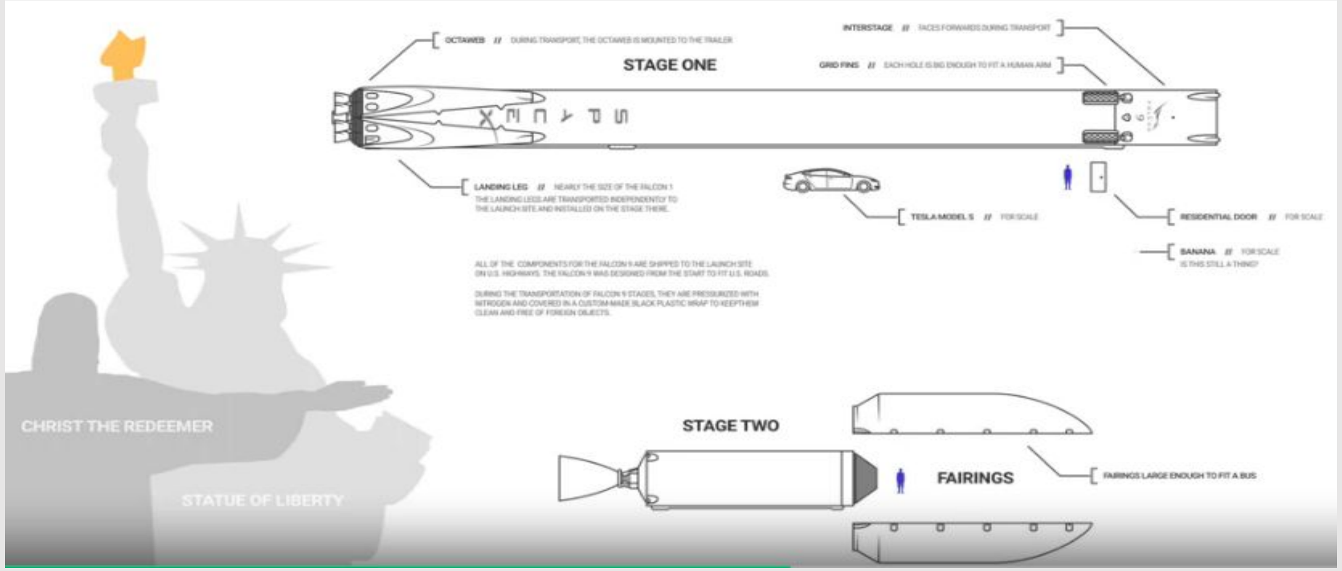


Image displays stages in Falcon 9 launch

Section 1

Methodology

Methodology

2 Data collection

SpaceX REST API & webcraping Wikipdia page to leverage data for analysis

2 Data wrangling

Hot encoding to transform data for future manipulation

2 Exploratory data analysis

Exploratory data analysis (EDA) using data visualization and MatploltLib and SQL

2 Visualization analysis

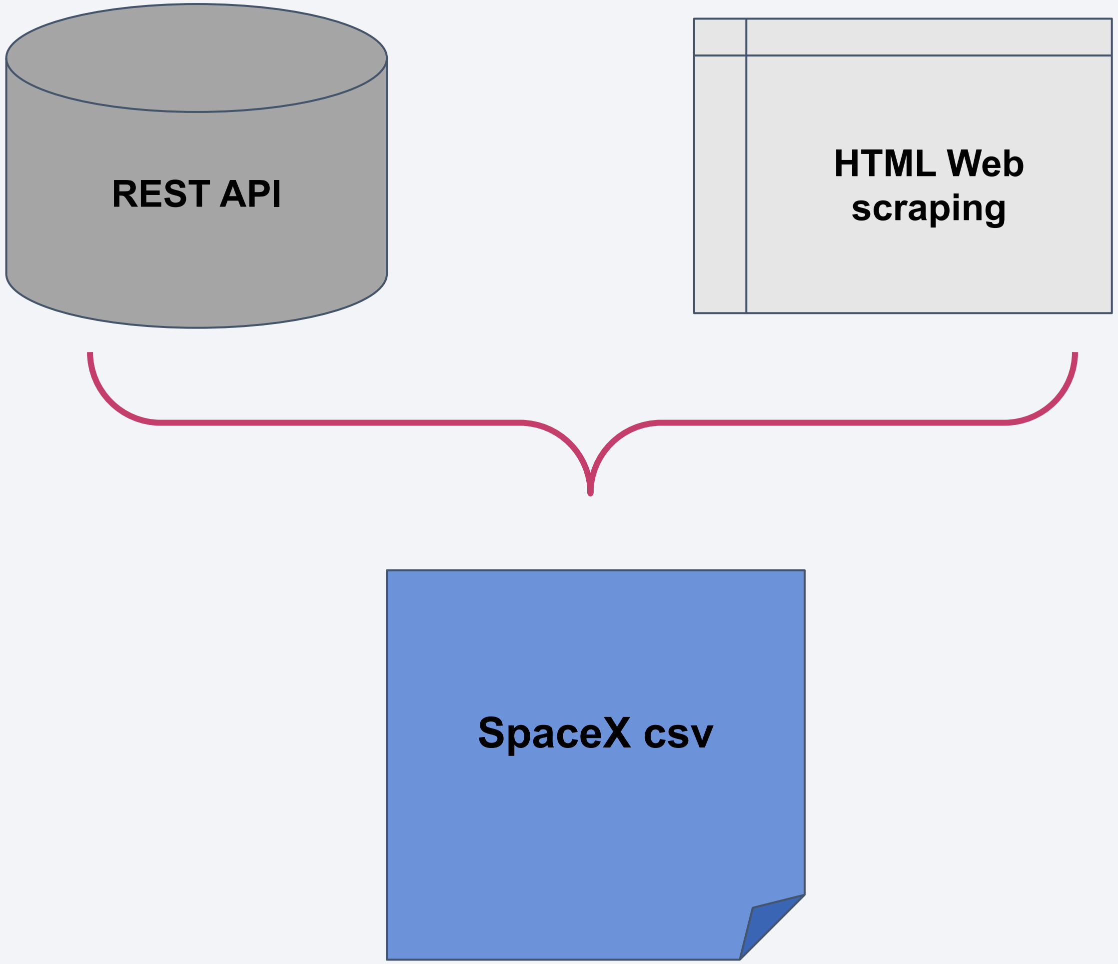
Performed using Folium and Plotly Dash to create bar charts, scatter plots, and pie charts

2 Prediction

Iterate on classification models: Logistic Regression (LR), k-nearest (KNN), support vector machine (SVM), and Decision Tree (DT)

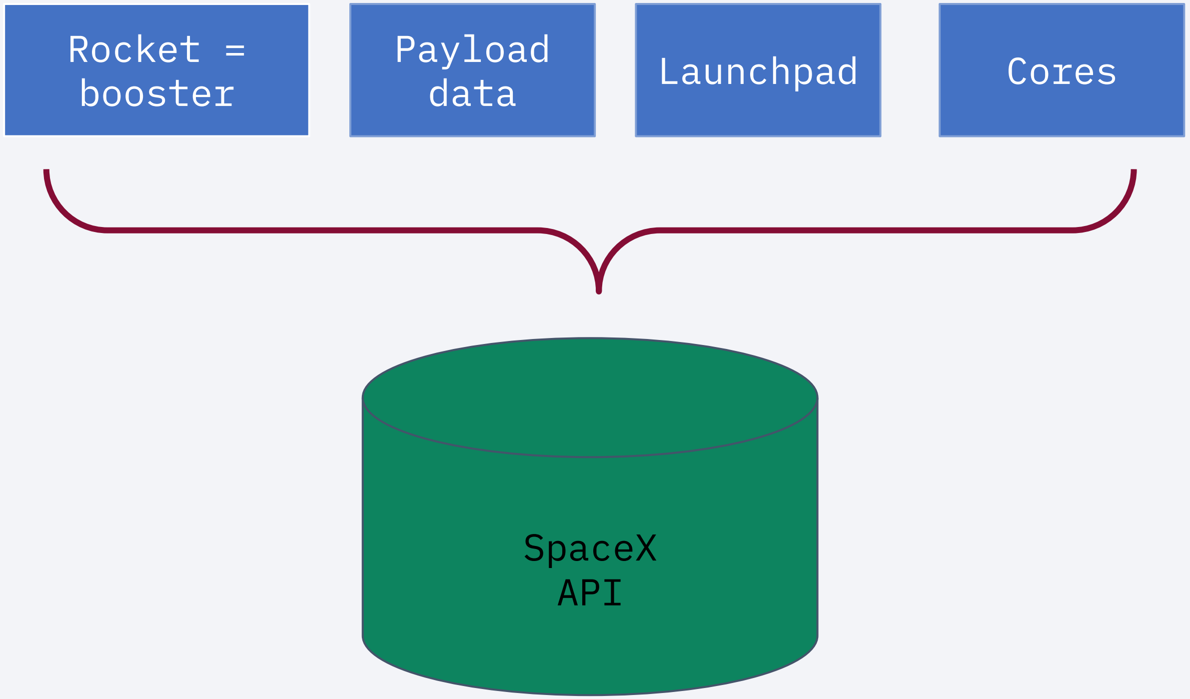
Data Collection

The data used for analysis and prediction was collected from SpaceX API and Web scraping to create the final csv file.



SpaceX API

Collected SpaceX rocket data gathered from API extracted from URL: api.spacexdata.com/v4/launches/past

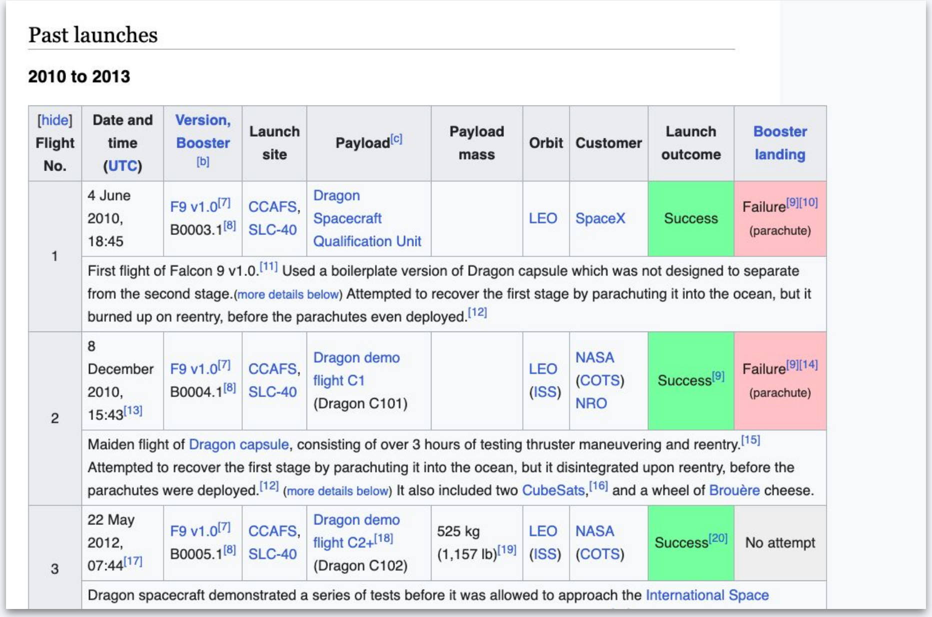


- Request data from REST API using GET request
- API returns information as JSON to create Pandas dataframe
- Prepare data via updating columns and row
- Filtered for Falcon 9 launches
- Explored columns and values
- Manage missing data by populated with payload mean
- Export data as .csv

Notebook

Web Scraping

Collected SpaceX data from webscraping a [Wikipedia page](https://en.wikipedia.org).



Past launches

2010 to 2013

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space								

Table for web scraping

HTML from website

Requested HTML GET response from Wikipedia page

Beautiful Soup

Extract data using package in order to scrape HTML

Verified tables

Checked column and variables, and got column names.

Converted data

Converted data into Pandas dataframe for analysis: named launch_dict

Populated dataframe

Populated launch_dict with launch data

CSV

Exported data as .csv file

[Notebook](#)

Data Wrangling

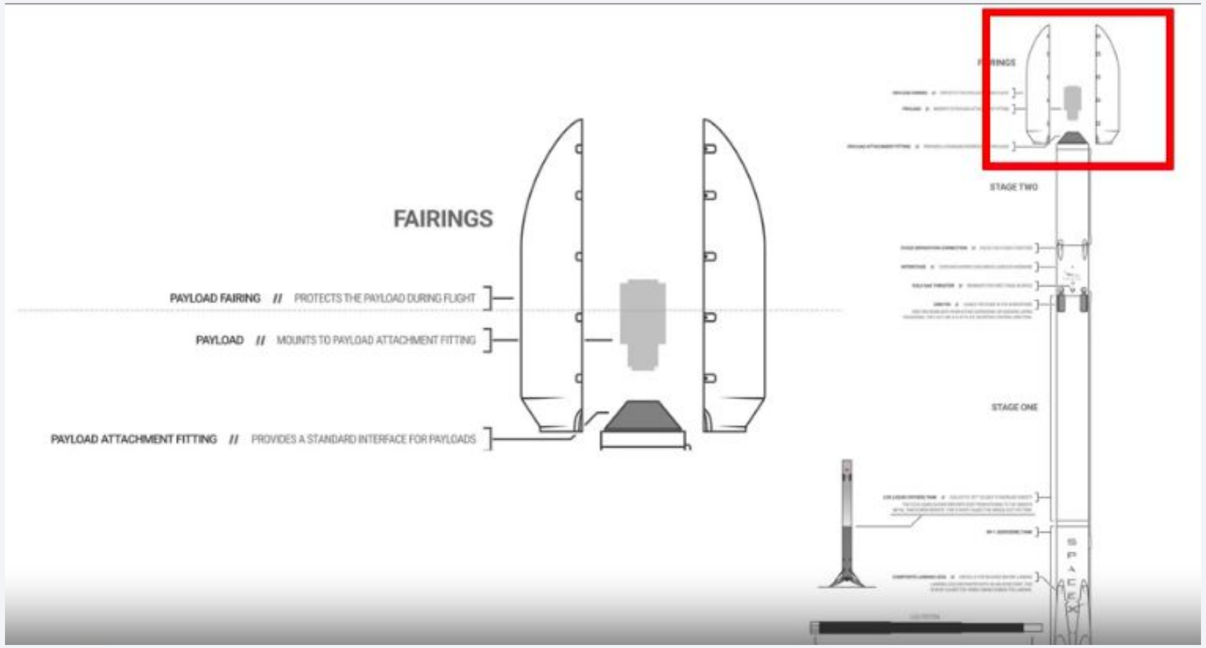
Load data	Data exploration	Launch site	Orbit	Mission outcome	Outcome	Outcome
Loaded SpaceX data set from collection phase.	Calculated percentage of missing values for each attribute (e.g. date and orbit) and found data type for each. This was done to understand data quality and to get bearings.	Calculated launches for each site: Cape Canaveral's 40 VAFB SLC 4E , Vandenberg Air Force Base Space Launch Complex 4E (SLC-4E), and Kennedy Space Center Launch Complex 39A KSC LC 39A	Calculated the number and frequency of each orbit because each launch aims to an dedicated orbit.	Calculated outcome of each mission per orbit type.	Created launch outcomes labels with values 0 (bad outcome) and 1 (good outcome) to understand success rates.	Exported data as .csv for future analysis

Notebook

EDA with Data Visualization: Variables

Assessed the following variables to understand trends and relationships:

- FlightNumber - Indicates launch attempts
- Payload indicates - indicates launch outcome
- Payload Mass - weight of spacecraft component
  - Enclosed in the fairings.
- Orbit type Indicates



Highlighted box shows payload

[Notebook]

EDA with Data Visualization: Charts Types

Scatter plot - Leveraged given numerous numerical data to analyze how, and to what extent, two variables relate as independent and dependent variables. Variables with continuous data suitable:

- Flight numbers
- Payload
- Launch Site

Bar chart - Leveraged to show categorical,discrete data when there was minimal units to compare

- Displayed success rate by orbit type given limit orbit types (8)

EDA with Data Visualization: Exploration

Generated the following scattered point plot charts for a multitude of categorical data including:

- FlightNumber vs. PayloadMass to understand how launch attempts impact the launch outcome
- Flight Number and Launch Site to understand success rates across launch sites
- Payload and Launch Site to understand if there is any relationship between these two variables
- Payload vs. Orbit scatter to understand if there is any relationship between these two variables

[Notebook]

EDA with SQL: Queries

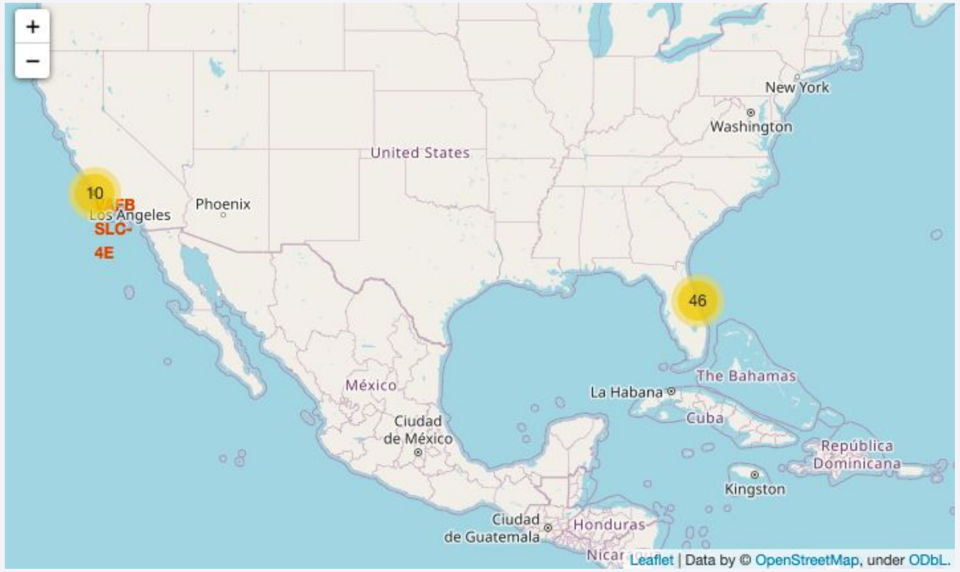
Executed the following queries:

- Unique launch site names
- Launch sites beginning with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass by booster version F9 v1.1
- Date from first successful landing outcome in ground pad
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Complete names of the booster versions that have carried the maximum payload mass
- List of booster versions with their month that failed (drone ship) in 2015
- List of total successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Notebook

Interactive Folium Map: Objects

- folium.Marker() - Indicates launch records (class = 0 or 1)
    - Green for success
    - Red for fail
- marker_cluster() - Grouped launches per site (via launch coordinate)
- markt_color() - Stores color based on class
- folium.Circle() - Icon signifies each launch site
- folium.Icon - Allows us to customize our marker
- folium.PolyLine() - Creates a vector used for understanding distance
- folium.plugins.Anth() - Overlays polyline on a map



[Notebook](#)

Interactive Map with Folium: Details

- Initialized map centered on Houston, the site of NASA Johnson Space Center to start
- Added each launch site's location as a circle (r = 1,000) to represent the coordinate for easy identification
- Added text label for each launch site to enhance readability
- Created markers for all launch records to indicate success
    - Green = Successful
    - Red = Failure
- Created colored cluster to group launch sites with multiple launch records to simplify results and increase readability
- Calculated distance between launch sites and proximity (e.g. highways or coastline) to explore any relationship
- Added marker with calculator distance between launch site and proximity for evaluation
- Create a line to from calculated distance (km) between coastline and a launch site for experimentation
    - Added label for readability

[Notebook]

Interactive Dashboard

Created an interactive dashboard titled **SpaceX Launch Records Dashboard** to explore data dynamically and visually.

Success rates were specifically assessed by launch sites and for a range of payload ranges across booster versions.

As drop down for launch site allowed to toggle view success rates on a granular level. Similarly, a slide for payload mass allowed to us to access success rates across a multiple of ranges.

Imported packets

Pandas, Dash, and Plotly to build in a dashboard in a sandbox.

Load SpaceX data

Loaded and read SpaceX dash .csv to leverage .

Prepared dashboard skelton

Loaded sie launch and payload data into outline while putting placeholders in for a pie chart, slider, and scatterplot.

Call-back for launch site

Called back data to create a drop down for sites to assess success rates by launch.

Call-back for payload

Exported as .csv fil

Launch

Launched dashboard in sandbox.

Predictive Analysis (Classification)

Build model

Compare models

- **Prepare data**: Loaded data into dataframe
- **Classificationpreperation**: Created a NumPy array from column class
- **Standardized data**: Transformed data for manipulation
- **Split data**: Divided data into test and train sets
- **Test models**: LR, SVM, DT, and KNN
- **Fit models**: Build GridSarch CV() and fit each model to find best parameters
- **Paramters**: Find best parameters for model
- **Accuracy**: Calculated model accuracy using score.
- **Confusion matrix**: UNderstand model erors
- **Compare**: Evaluate results

[Notebook]

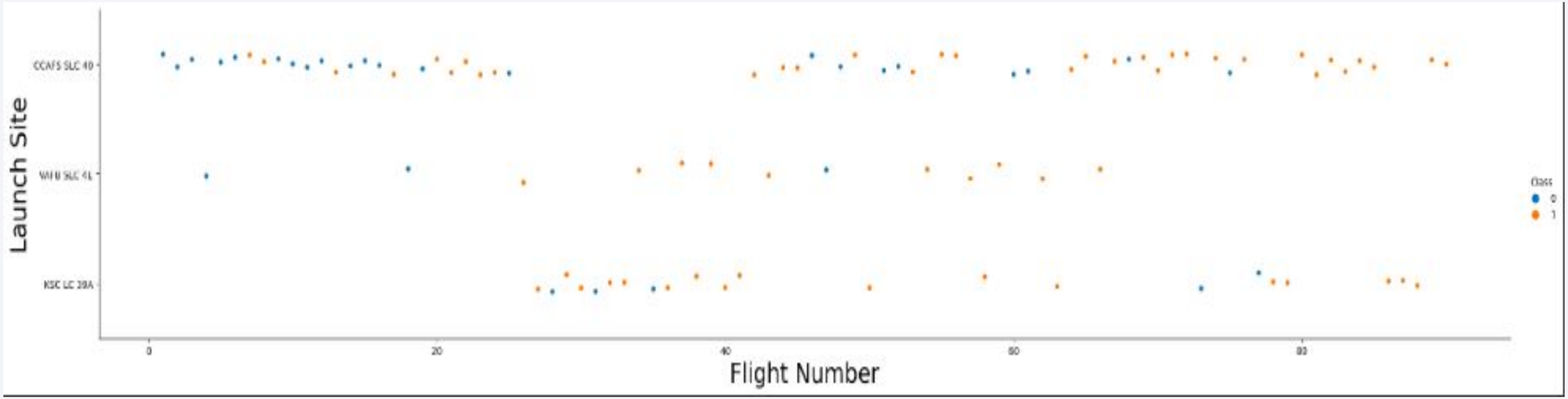
Repeated these steps for four models

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

As flight number increases, there are more successfully launches esp. at CCAFS-40 after 40 flights. Success starts at VAFB SLC 4E and KSC LC 30A begins after 20 flights.

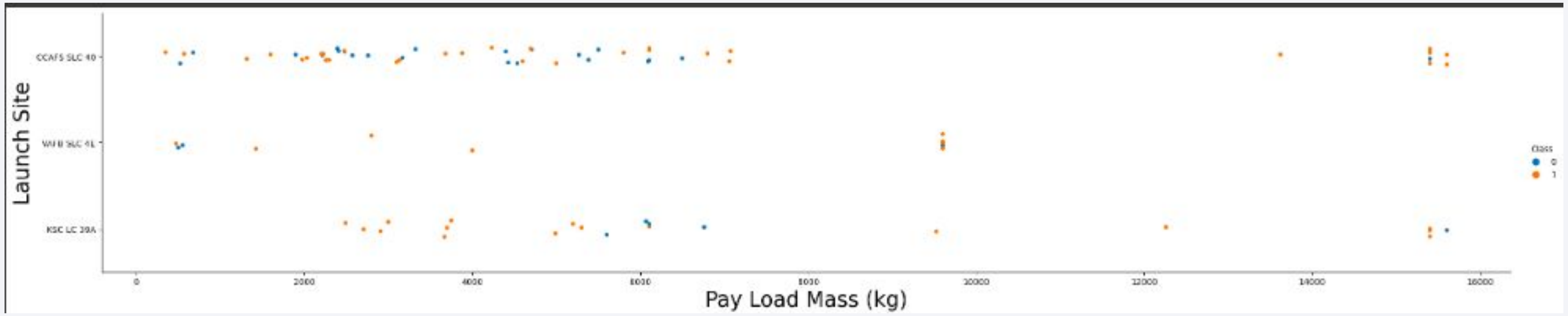


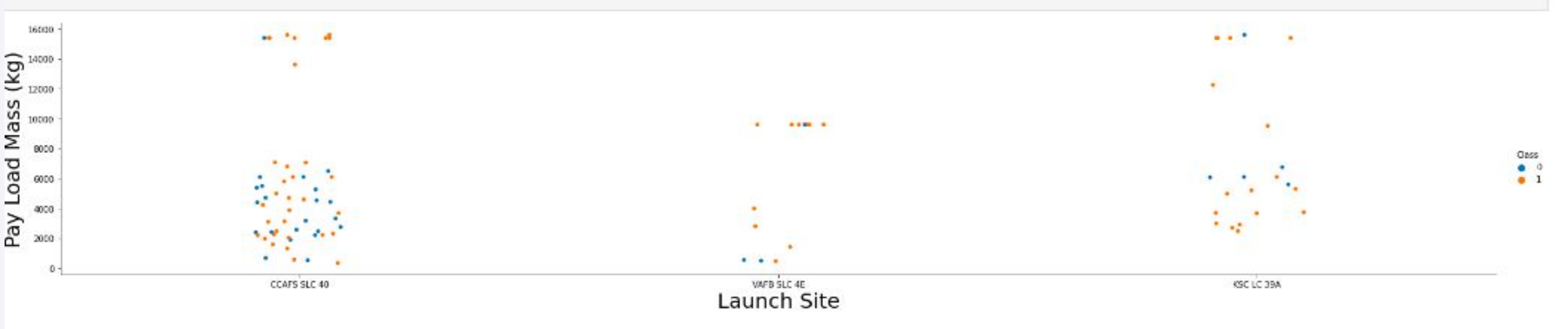
Notebook

Payload vs. Launch Site

As Pay Load increases, the success rate increases with considerable variance per site..

- CCAFS-40 struggle as Pay Load mass approaches 8,000 kg - 12,000 kg
- VAFB SLC 4E struggle with a payload above 10,000 kg
- KSC LC 39A struggle with a payload just below 6,000 kg - 12,000 kg
- At 10,000 kg VAFB SLC 4E shows some success with an outlier as well for KSC LC 39A at this size





[Notebook]

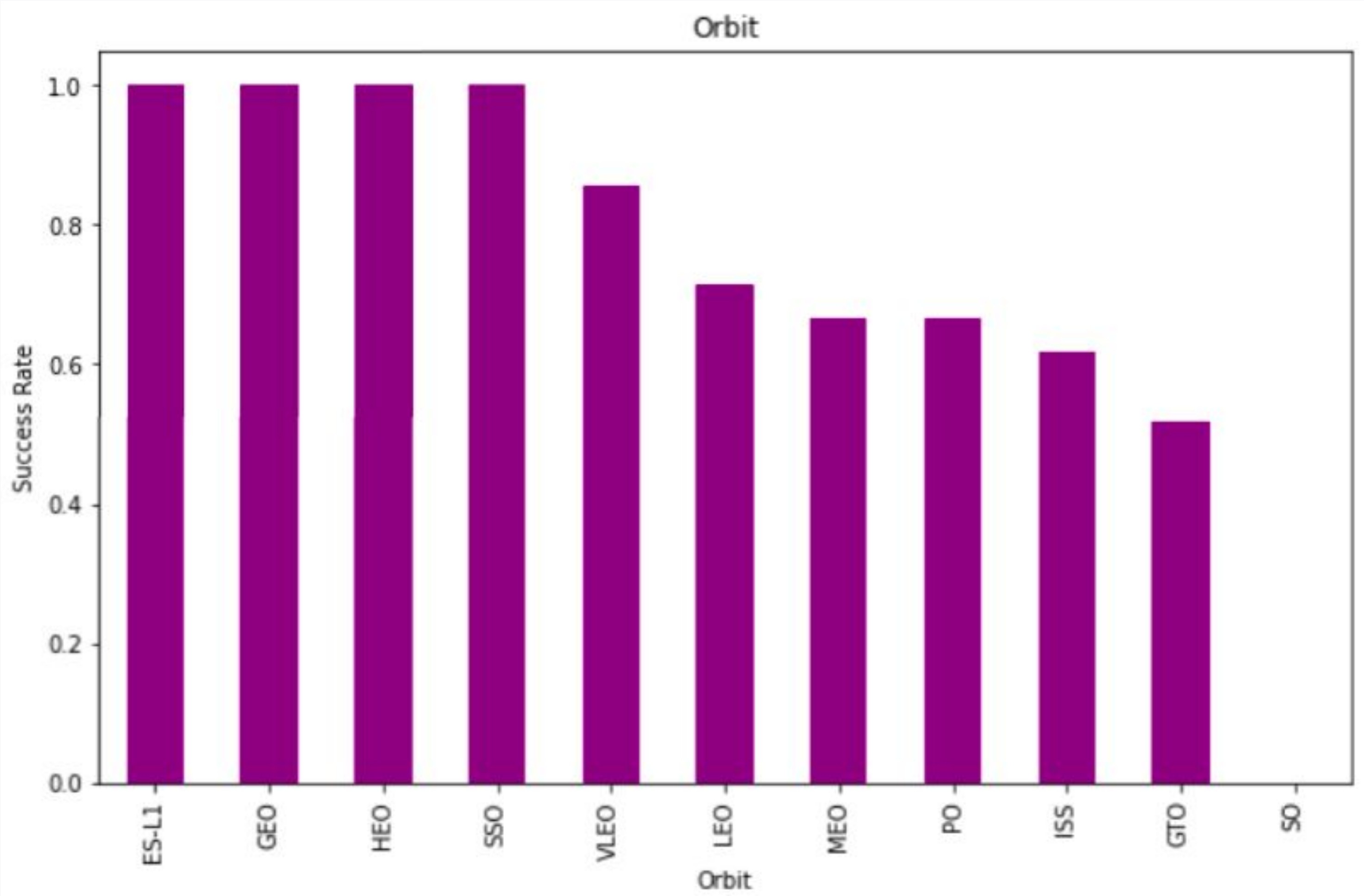
Success Rate vs. Orbit Type

Four orbit types show similar high-ranking success rate. in descending order:

- ESC-1
- GEO
- HEO
- SSO

These four suggest that Falcom's 9 first stage will successfully land (approx. 100%).

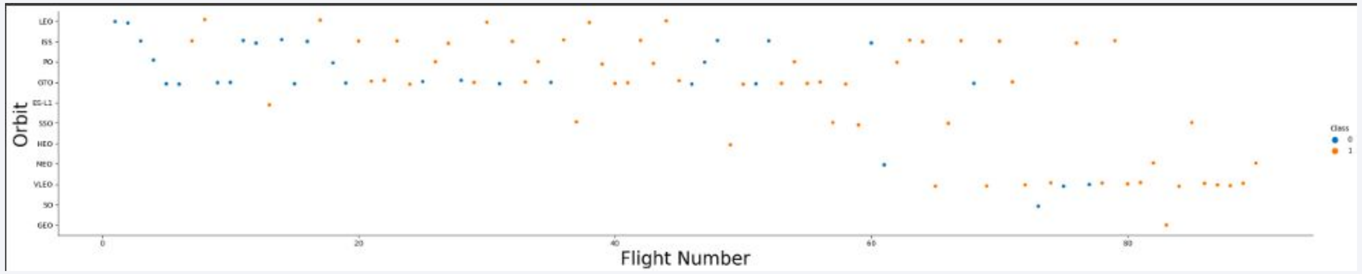
SO has a success rate of 0%.

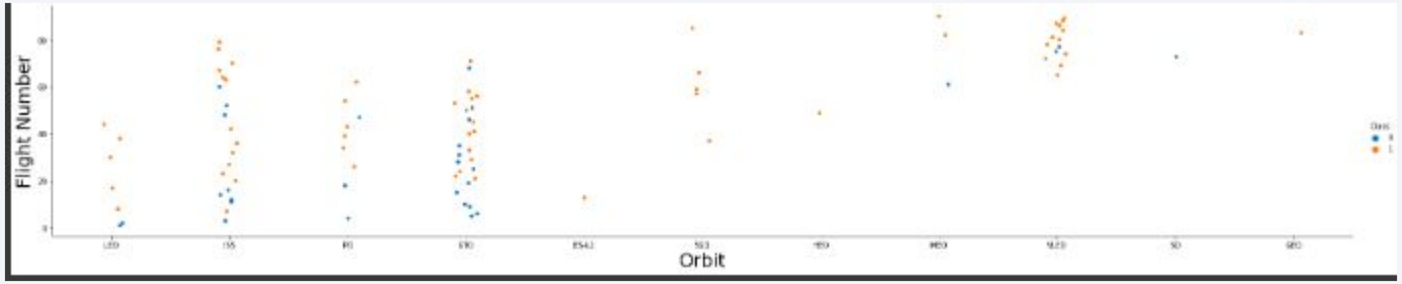


[Notebook]

Flight Number vs. Orbit Type

Non-clear trends across orbits suggesting a weak relationship.





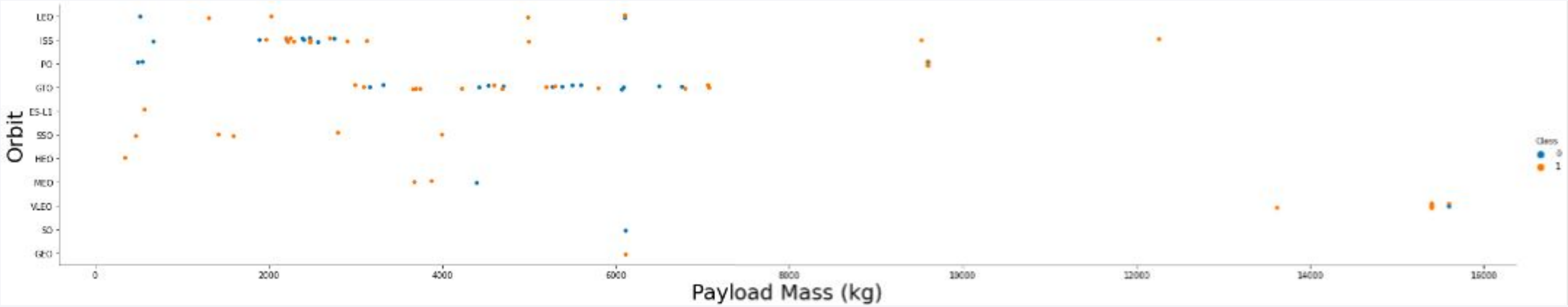
- Most flights struggle across orbits before 20 flights, thereafter LEO, ISS, and PO have more success
- After 60 flights, MEO has successful flights
- All flight for MEO, VLEO, SO, and GEO orbits struggle before approx. 50 flights
- We can observe more successful flights after 60 flights, notably for VLEO

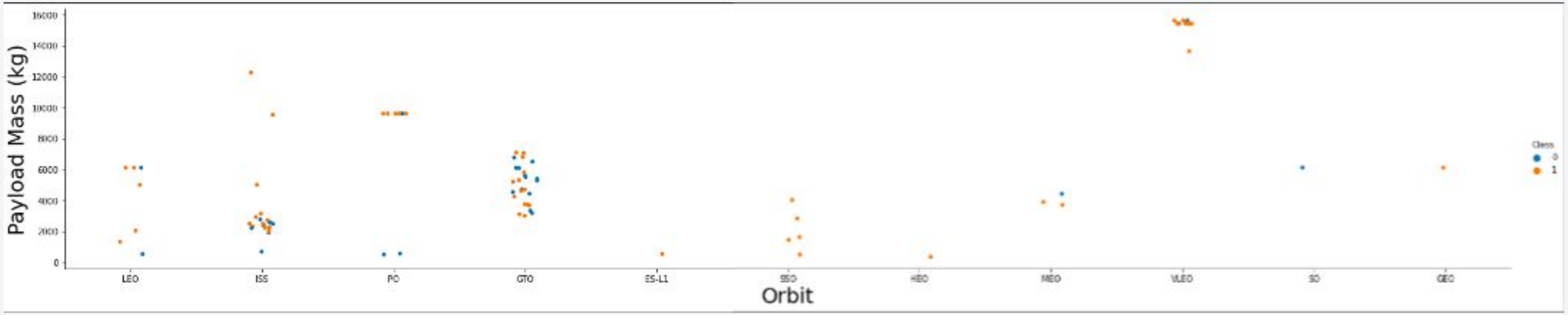
Notebook

Payload vs. Orbit Type

Non-clear trends again with orbits and payload given lack of ubiquity, except for lack of success at extreme end of the payload range.

Notebook

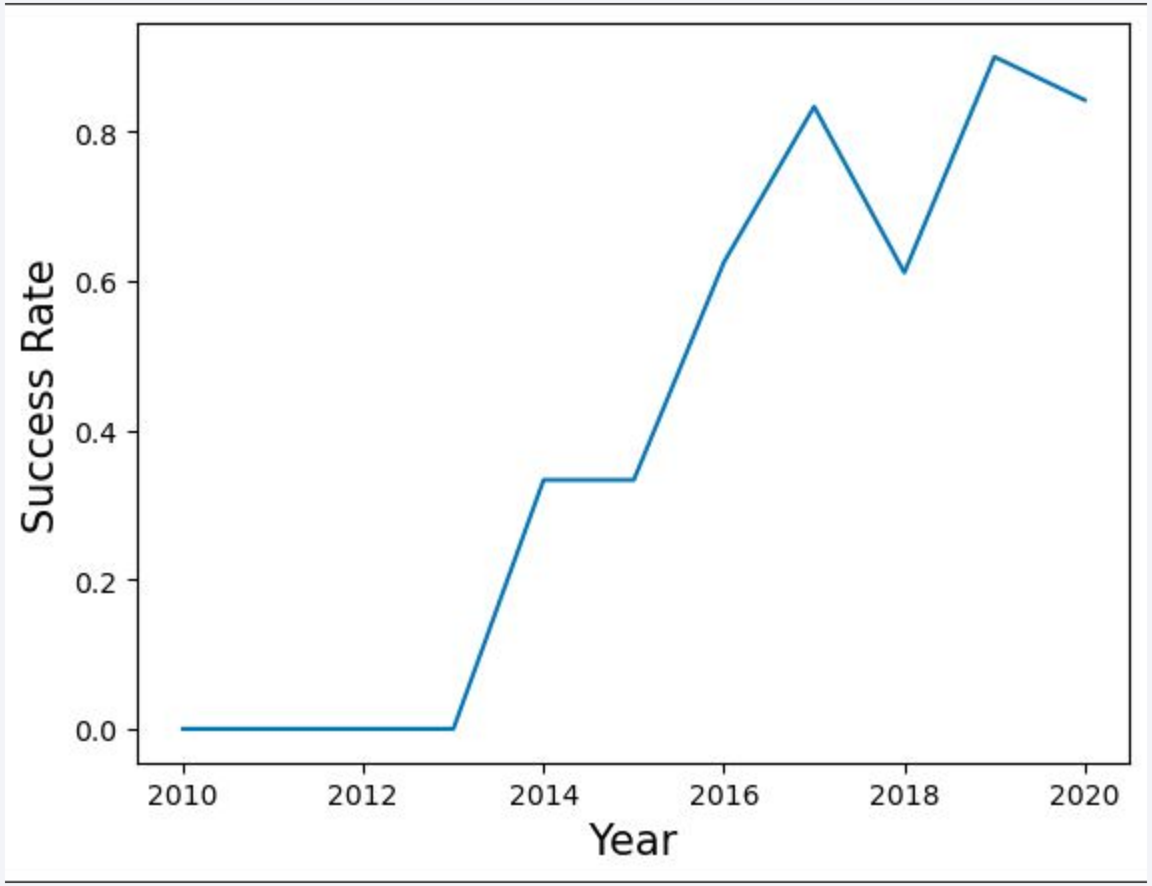




Launch Success Yearly Trend

The success rate has increased overtime since 2012 except in 2016.

The latest data from 2020, suggests a downward trend, but still above than before.



Notebook

All Launch Site Names

Falcon 9 launches from the following unique launch sites:

Query

```
%sql SELECT DISTINCT(LAUNCH_SITE)
from SPACEXTBL;
```

Result

Launch site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Notebook

Launch Site Names Begin with 'CCA'

Query

```
%sql SELECT LAUNCH_SITE
from SPACEXTBL
where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

Launch Site Names

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

Notebook

Total Payload Mass

The max payload the Falcon 9 can carry is 45,596 kg

Query

```
%sql SELECT SUM (PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)' ;
```

Result

45,576

Average Payload Mass by F9 v1.1

The Falcon 9, with booster F9 v1.1, carries a payload of 2,928.4 kg on average.

Query

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where
BOOSTER_VERSION = 'F9 v1.1' ;
```

Result

2,928.4 kg

First Successful Ground Landing Date

The first successful ground landing occured on 01-05-2017.

Query

```
%sql SELECT MIN(DATE), "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)";
```

Result

MIN(DATE)	Landing _Outcome
01-05-2017	Success (ground pad)

Notebook

Successful Drone Ship Landing with Payload between 4000 and 6000

The majority of flights with successfully drone ship landings and with a payload between 4,000 and 6,000 kg have a booster version F9 FT B10##.

Query

```
%sql SELECT "BOOSTER_VERSION", "Landing _Outcome"  FROM SPACEXTBL WHERE ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000) AND ("Landing _Outcome" = "Success (drone ship)");
```

Result

Booster_Version	Landing _Outcom
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

Notebook

Total Number of Successful and Failure Mission Outcomes

The majority of flights are successful at 99%.

Query

%sql SELECT MISSION_OUTCOME, COUNT(*) As Total from SPACEXTBL group by MISSION_OUTCOME

Result

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

Notebook

Boosters Carried Maximum Payload

The maximum payload each booster version can carry is 15,600 kg.

Query

```
%sql SELECT distinct BOOSTER_VERSION, PAYLOAD_MASS__KG_

from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Result

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Notebook

2015 Launch Records

In 2015, booster versions F9 v1.1 B1012 and F9 v1 B1015 failed (in drone ship) during the January and April respectively.

Query

```
%%sql  SELECT substr(Date, 4, 2) AS Month,booster_version,"Landing _Outcome"

FROM SPACEXTBL WHERE  "Landing _Outcome"

='Failure (drone ship)' and substr(Date,7,4)='2015'
```

Result

Month	Booster_Version	Landing _Outcome
01	F9 v1.1 B1012	Failure (drone
04	F9 v1.1 B1015	Failure (drone

Notebook

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

Between 2010-06-04 and 2017-03-20, the most successful landing outcomes occurred on 07-08-2018.

Query

%%sql SELECT "Date","Landing _Outcome", COUNT("Landing _Outcome") AS 'Total'

FROM SPACEXTBL WHERE "DATE" BETWEEN '04-06-2010' AND '20-03-2017'

GROUP BY "Landing _Outcome"

ORDER BY COUNT ("Landing _Outcome") DESC

Result (see right)

Date	Landing _Outcome	Total
07-08-2018	Success	20
08-10-2012	No attempt	10
08-04-2016	Success (drone ship)	8
18-07-2016	Success (ground pad)	6
10-01-2015	Failure (drone ship)	4
05-12-2018	Failure	3
18-04-2014	Controlled (ocean)	3
04-06-2010	Failure (parachute)	2
06-08-2019	No attempt	1

Notebook

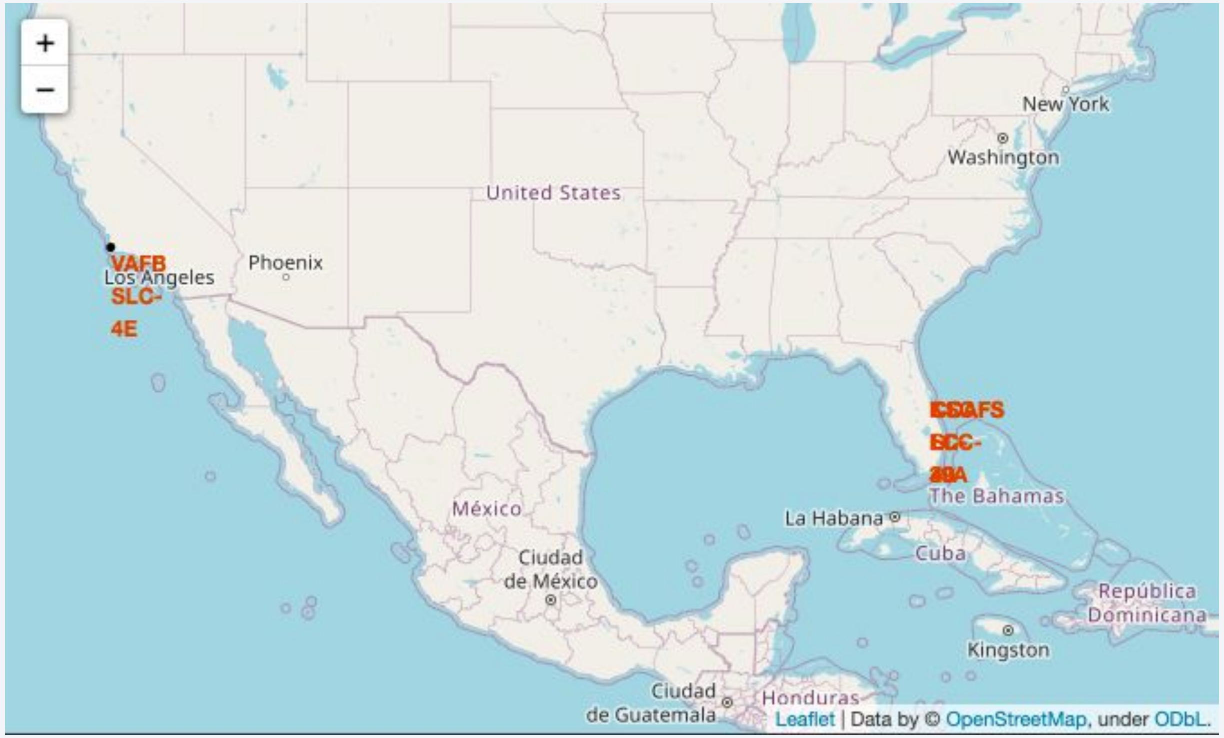
Section 3

Launch Sites Proximities Analysis

Launch sites location

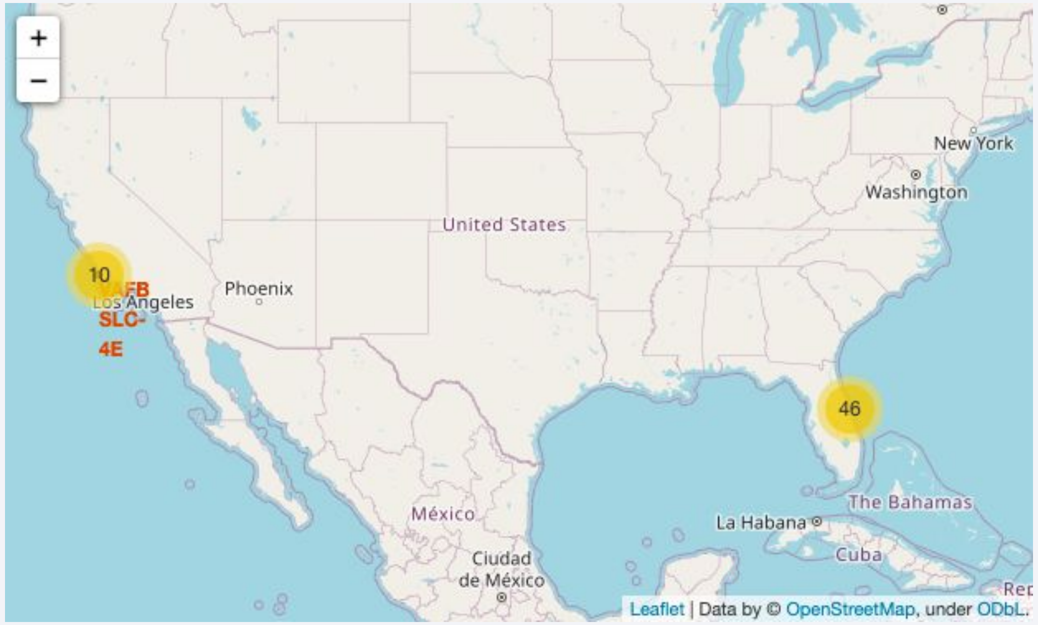
Falcon 9’s four launch sites are located along the coasts including:

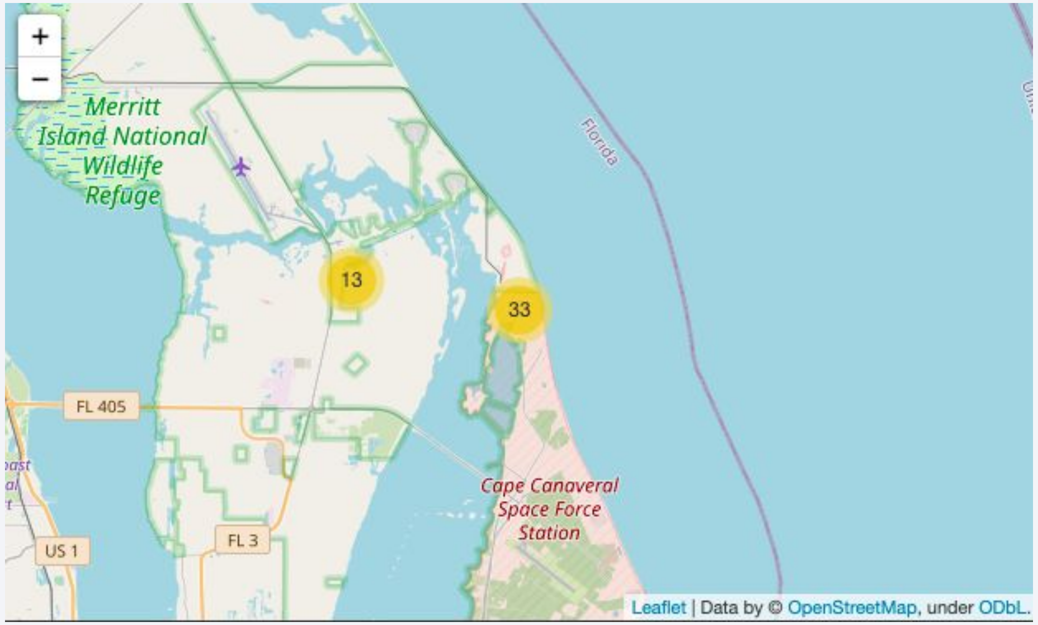
- KSC LC-39A (Merritt Island, FL)
- CCAFS LC-40 (Cape Canaveral ,FL)
- VAFB SLC-4E (Santa Barbara, CA)
- CCAFS SLC-40 (Cape Canaveral ,FL)



Launch site: Launches

The yellow cluster indicates flight launches. Cape Canaveral launches the most flights.

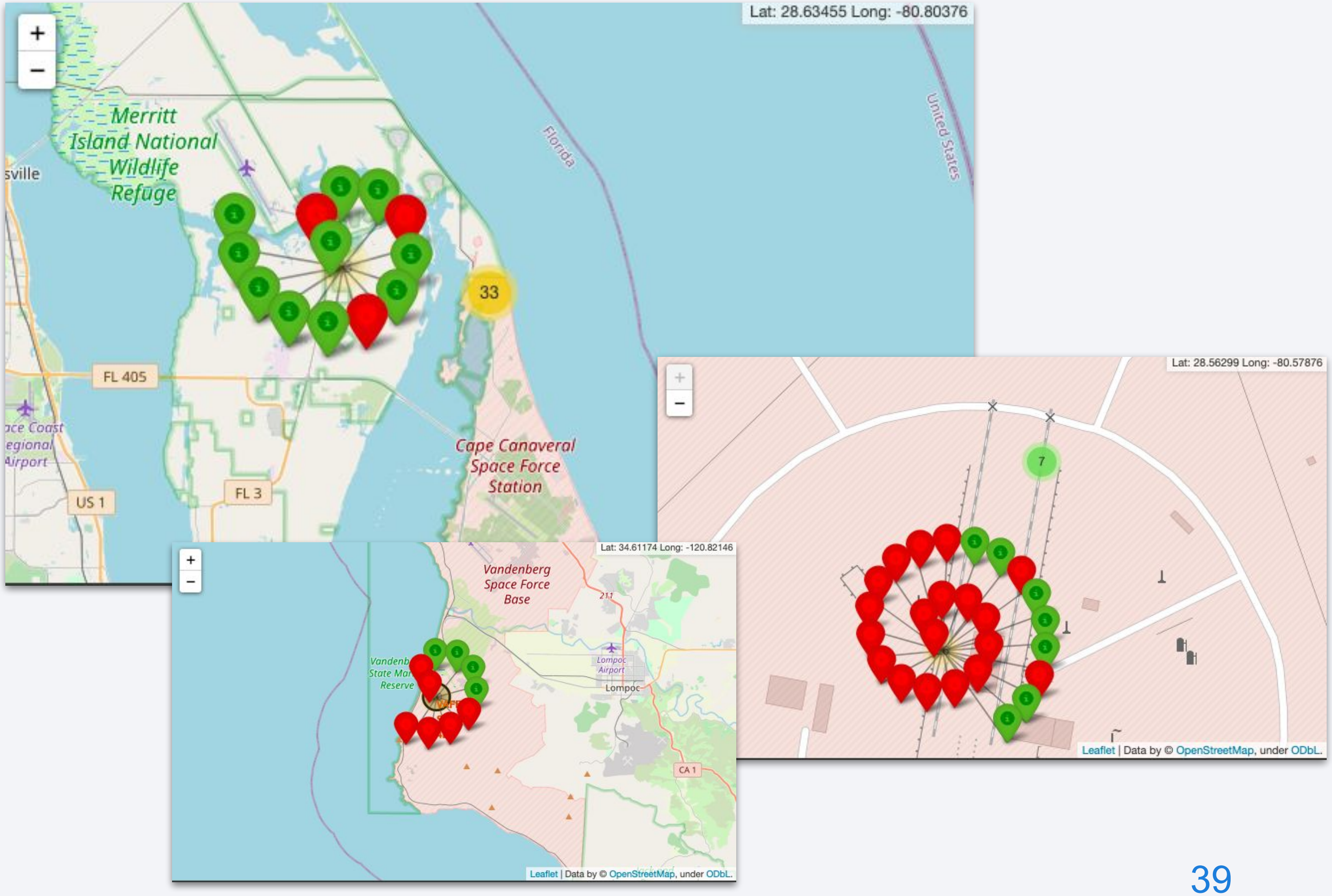




[Notebook]

Launch site

Launch site KSC LC-39A shows a majority successful flights.

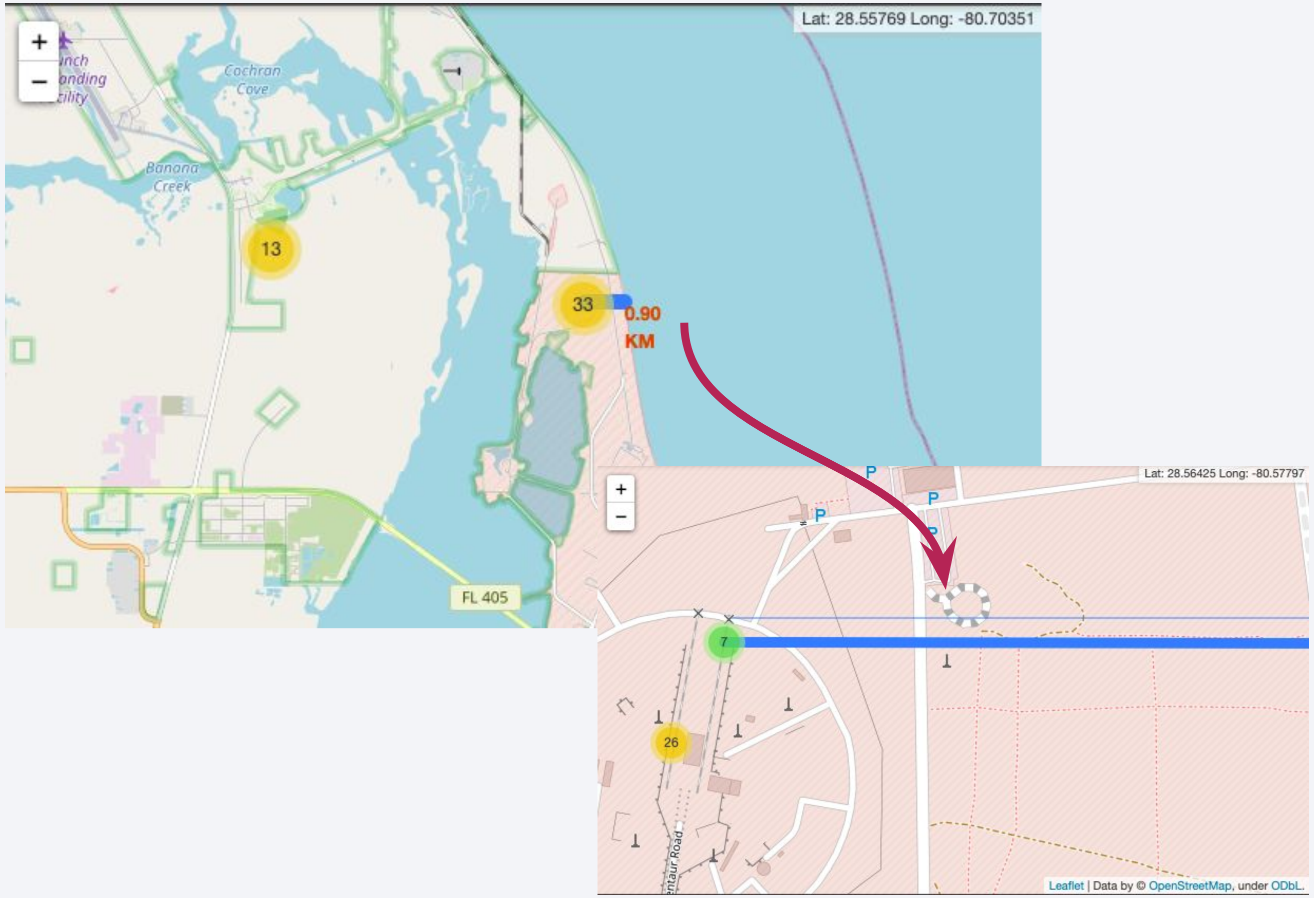


[Notebook](#)

Launch site: Proximity

Launch sites appear to be located close to points of transportation and access like the coastline, highway, or a rail station. This is likely to make the transportation of material easy and for safety.

For example, the blue line shows the distance between CAFS launch sites and to be only 0.90 km and the Florida coast.



Notebook

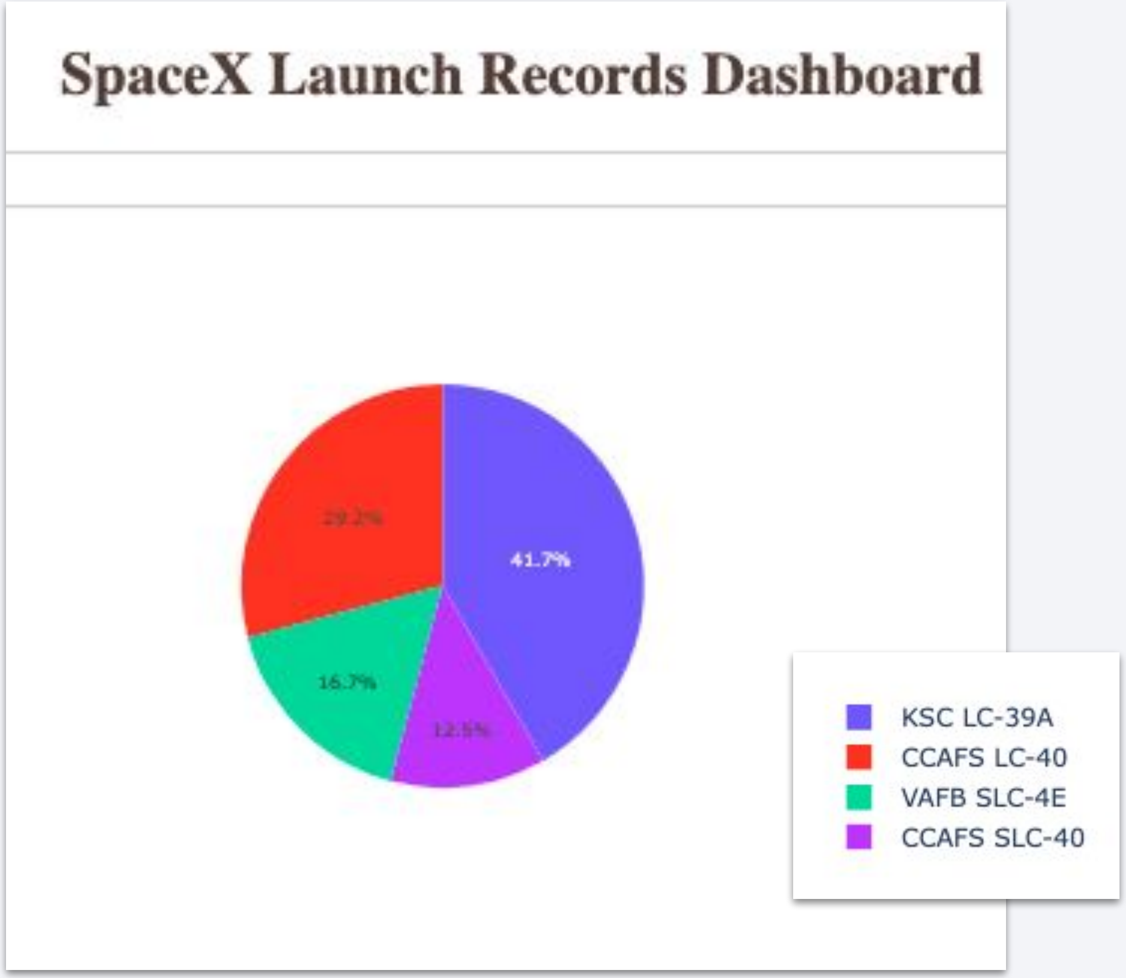
Section 4

Build a Dashboard with Plotly Dash

Total launches per site

Launches per site in descending order:

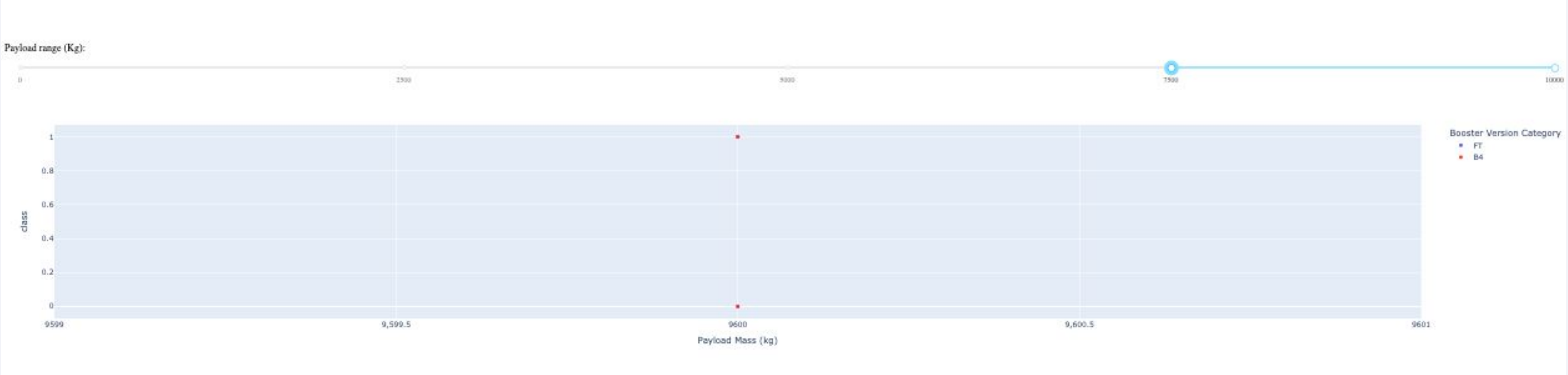
- KSC LC-39A (41.7%)
- CCAFS LC-40 (29.2%)
- VAFB SLC-4E (16.7%)
- CCAFS SLC-4 (12.5%)



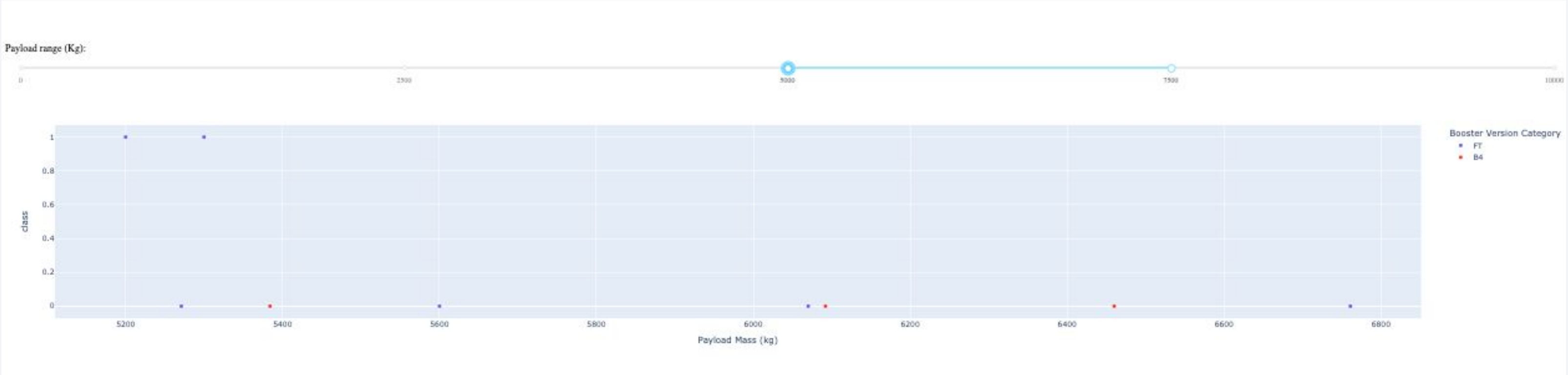
Payload range and launch outcomes

Less successful launches as payload reached higher range (7,500 kg - 10,000kg) of mass

- Only observe boosters versions FT and B4 have a high success rate

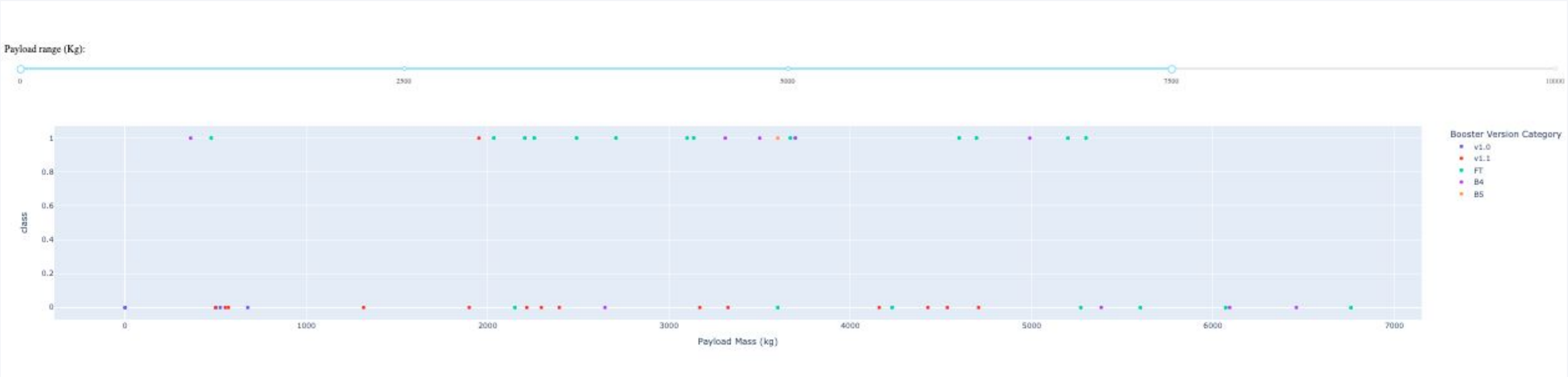


There is even less success between 5,000 and 7,500 kg for booster versions FT and V4



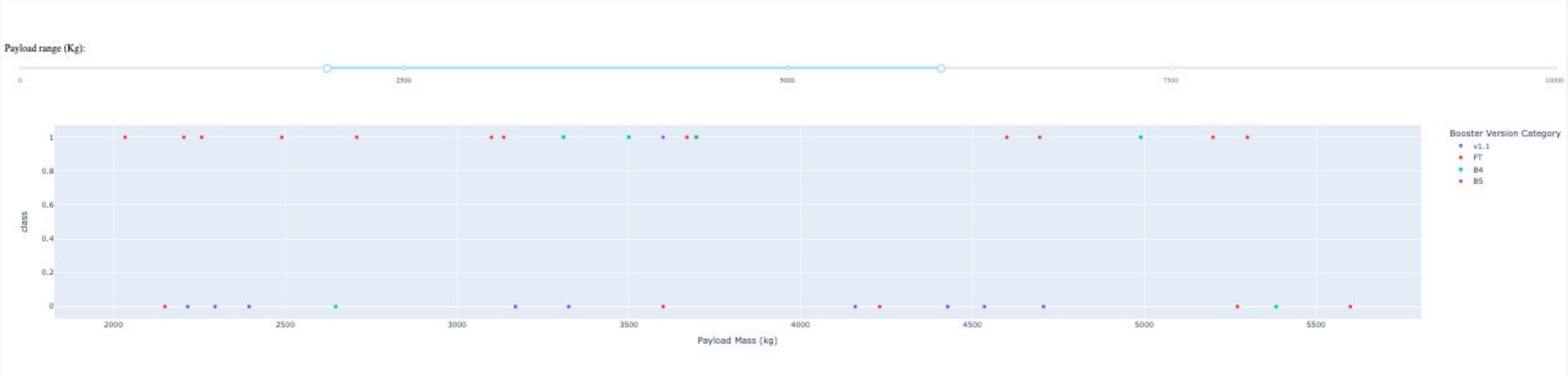
Payload range and launch outcomes

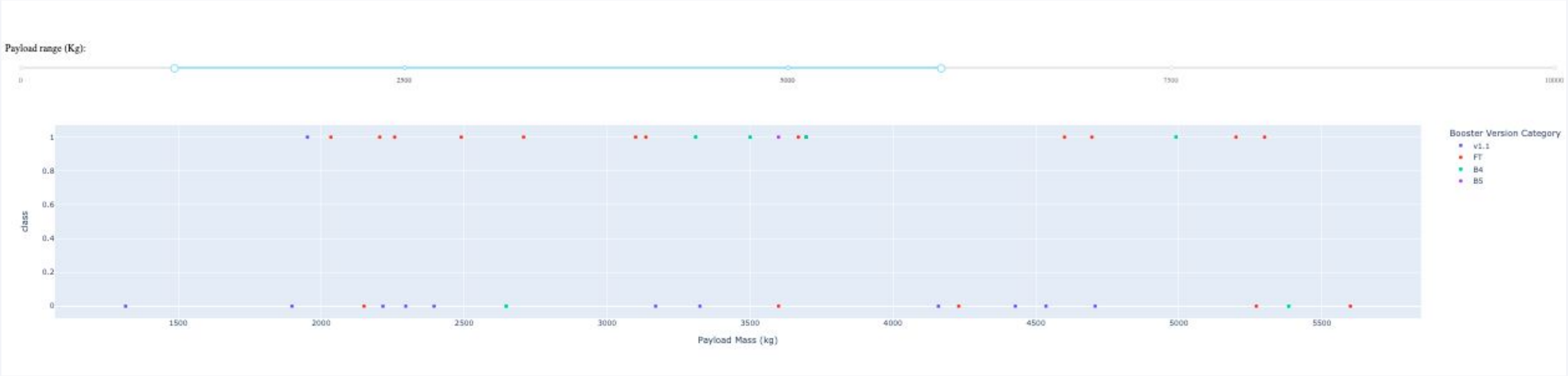
By comparison we see more successful launches across booster version up until 7,500 kg



Payload range and launch outcomes

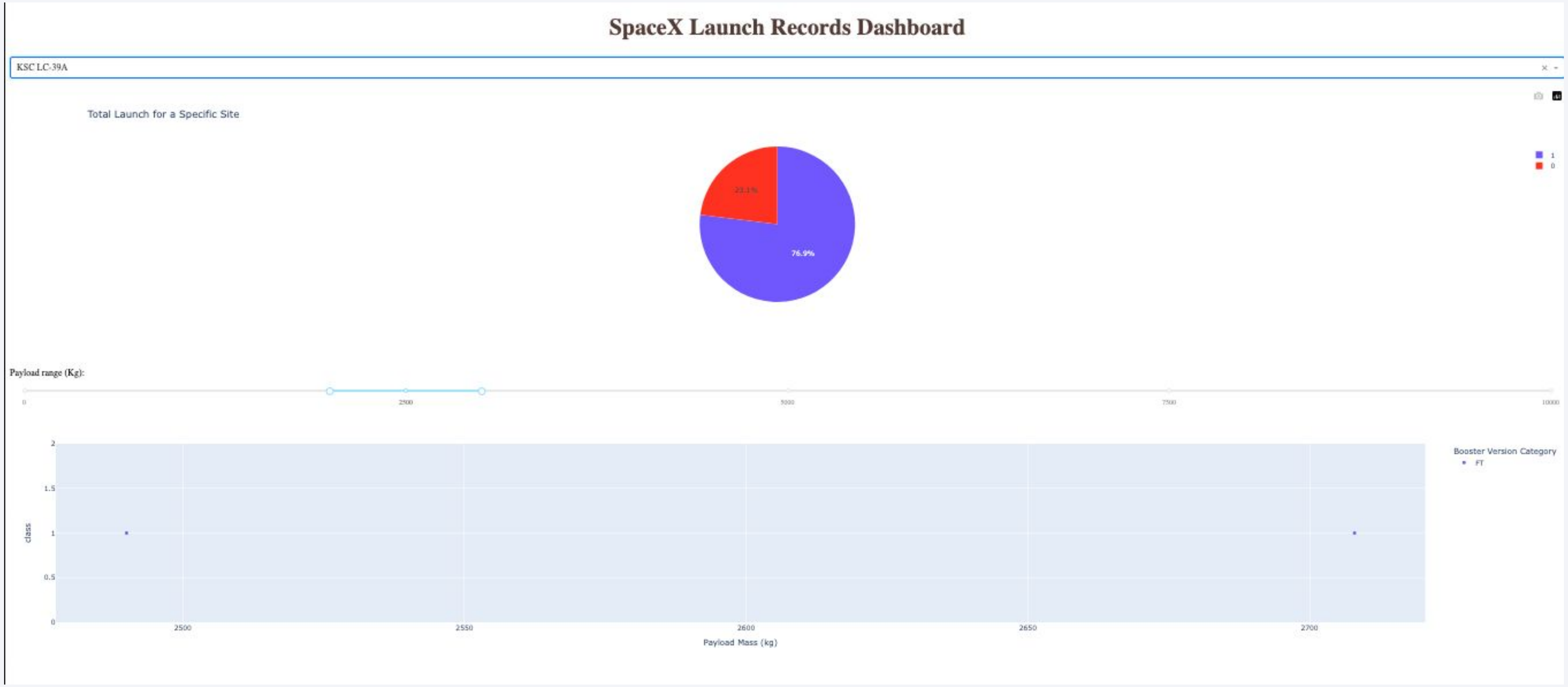
Success rates vary across boosters between approximately 2,000 kg 6,000 kg





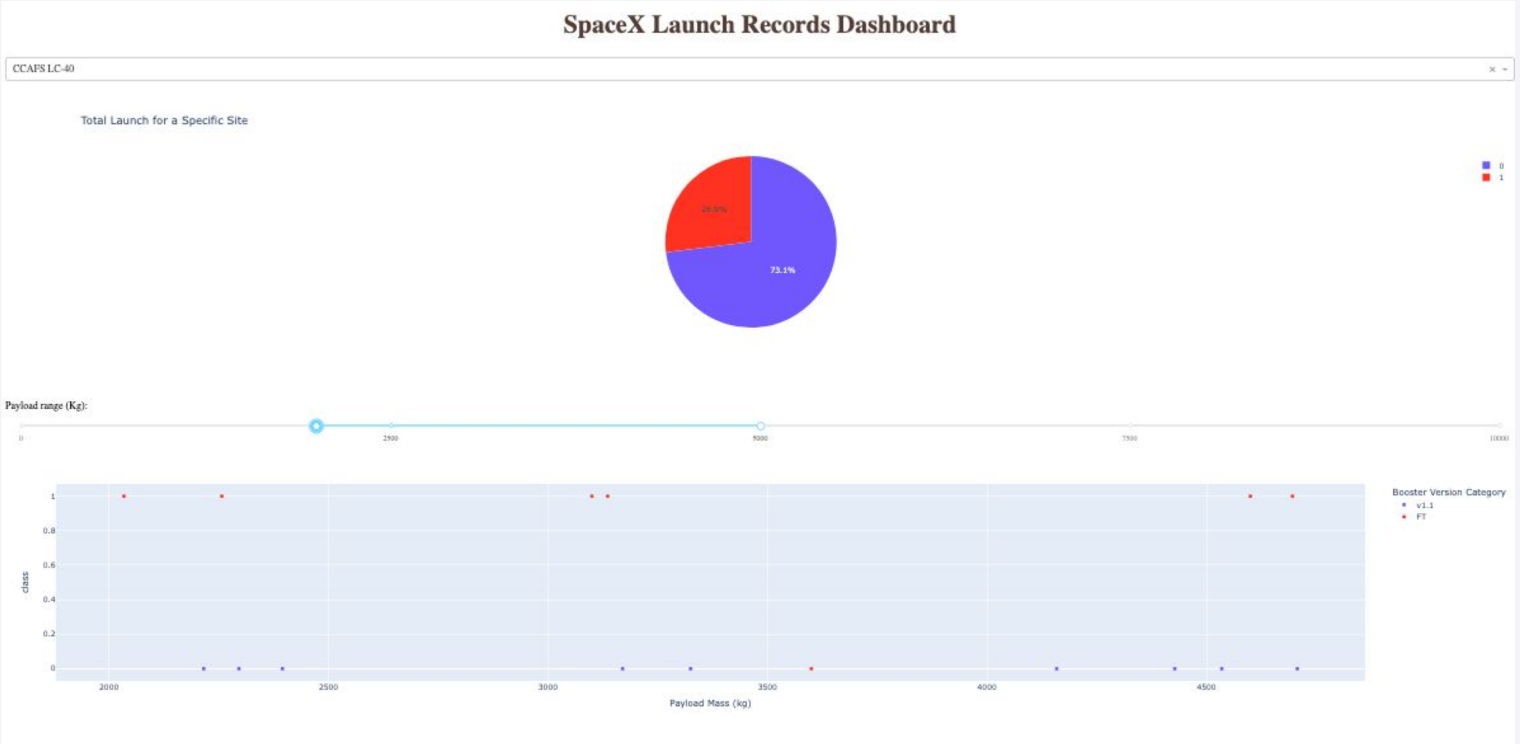
Successful launches for KSC-39A

KSC-39A has the highest success rate with a 74.6% success rate, a payload range around 2,500 kg with booster version FT performing the best.



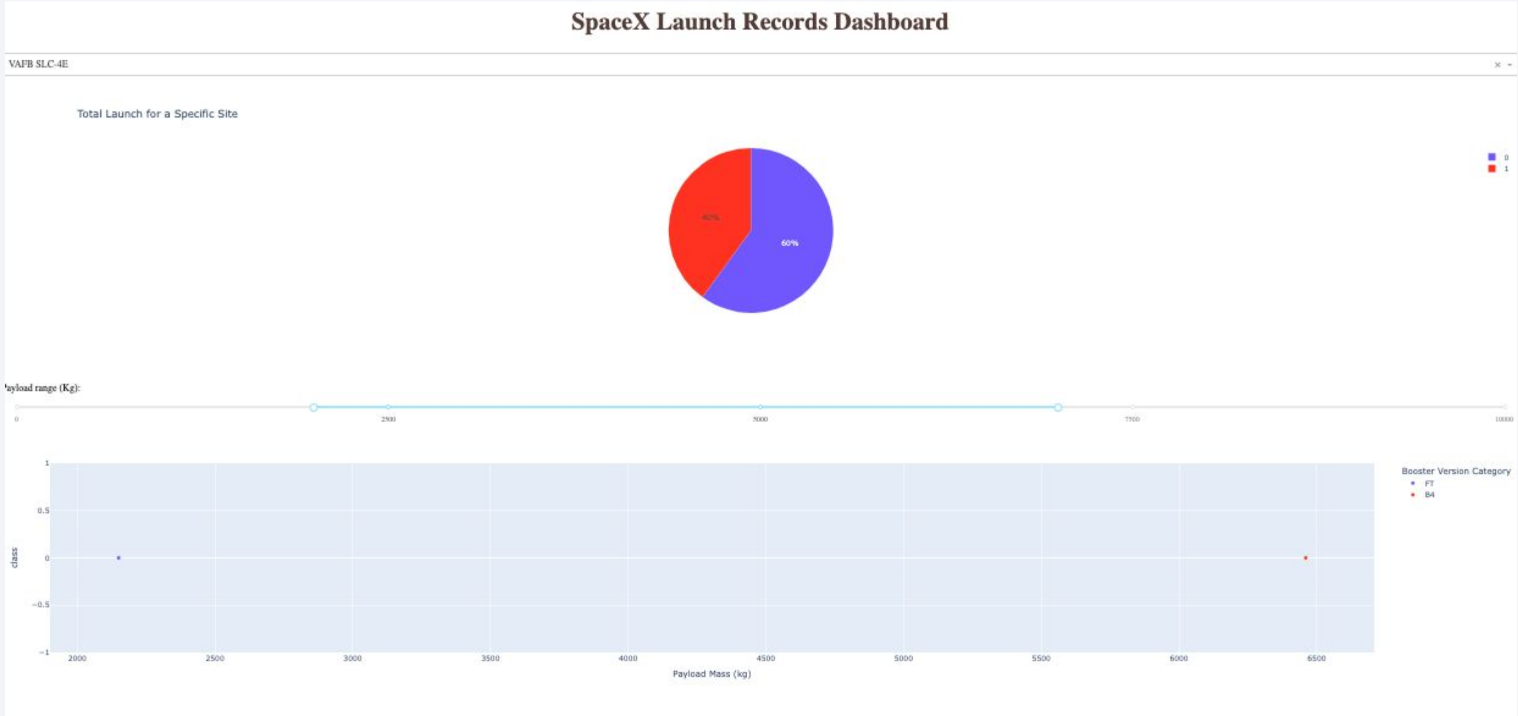
Successful launches for CCFS LC-40

CCSF-LC40 has the highest success rate with a 73.1% success rate, a payload range around 2,000 kg - 5,000 with booster version FT and v1.1. performing the best.



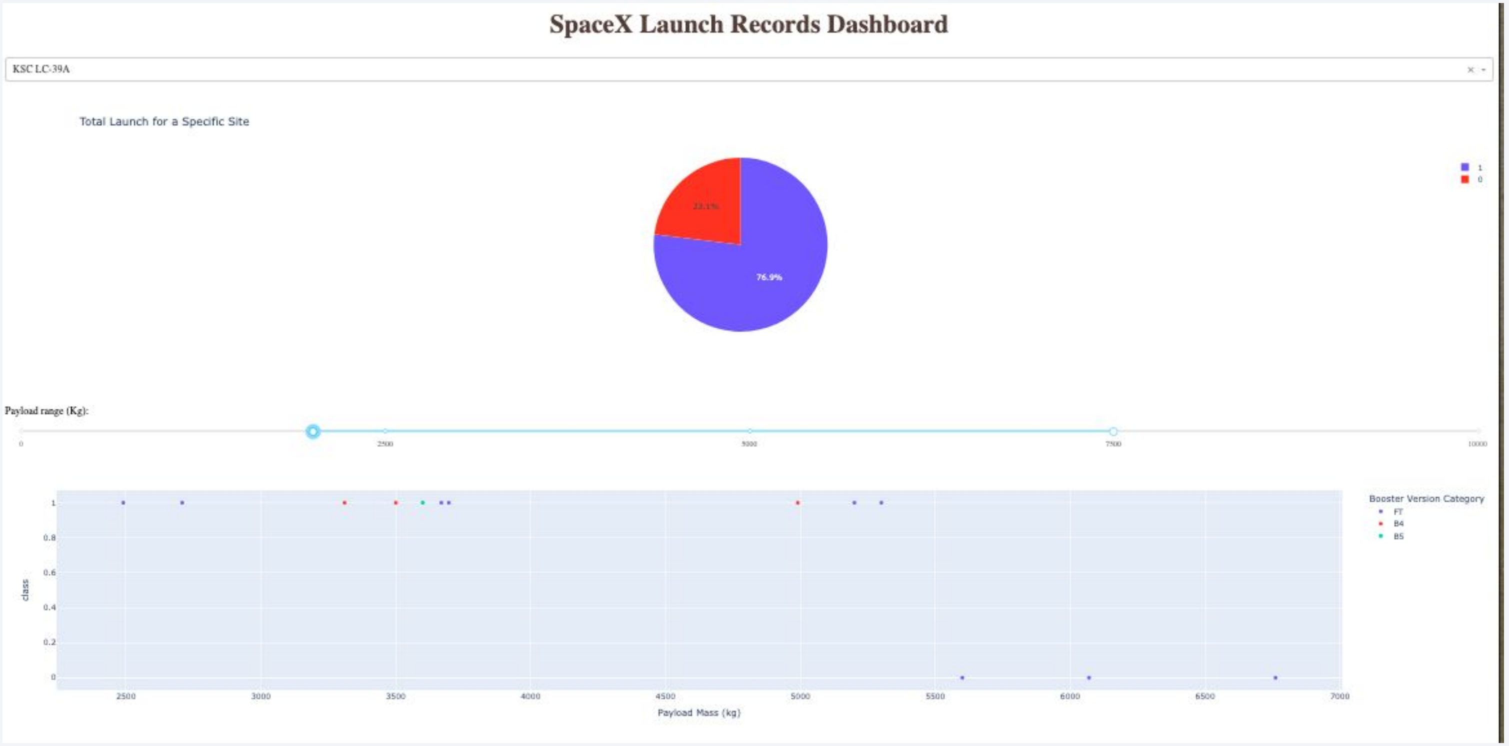
Successful launches for VAFB-SLC-4E

VAFB-SLC-4E has the highest success rate with a 60.0% success rate, a payload range around 2,000 kg - 7,000 with booster version FT and B4 performing the best.



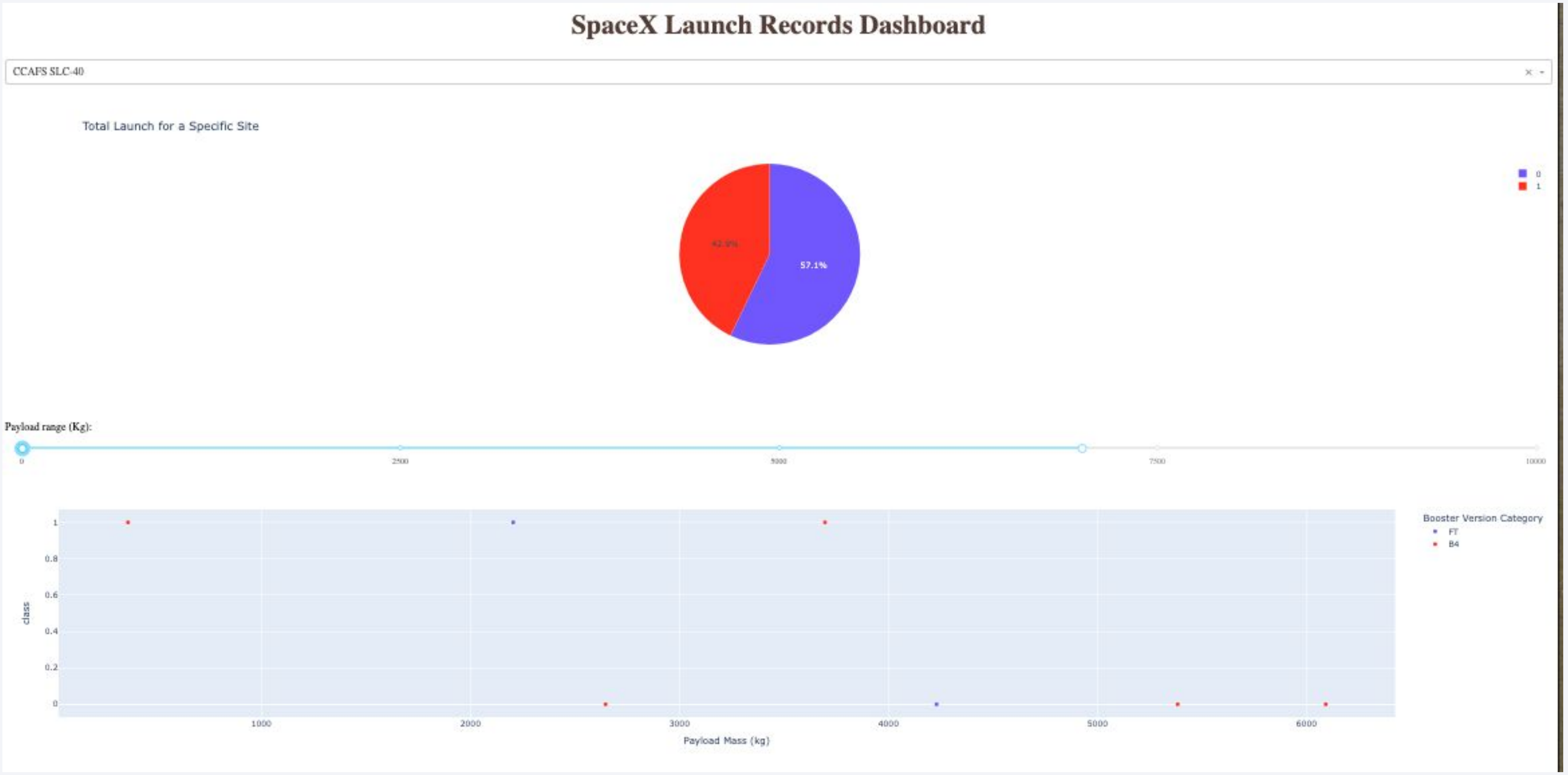
Successful launches for KSC-LA-39A

KSC-LA-39A has the highest success rate with a 76.9% success rate, a payload range around 2,000 kg - 7,500 with booster version FT, B4 and B5 performing the best.



Successful launches for CCAFS-SLC-40

CCAFS-SLC-40 has the highest success rate with a 57.1% success rate, a payload range around 0 kg - 7,000 with booster version FT and B4 performing the best.

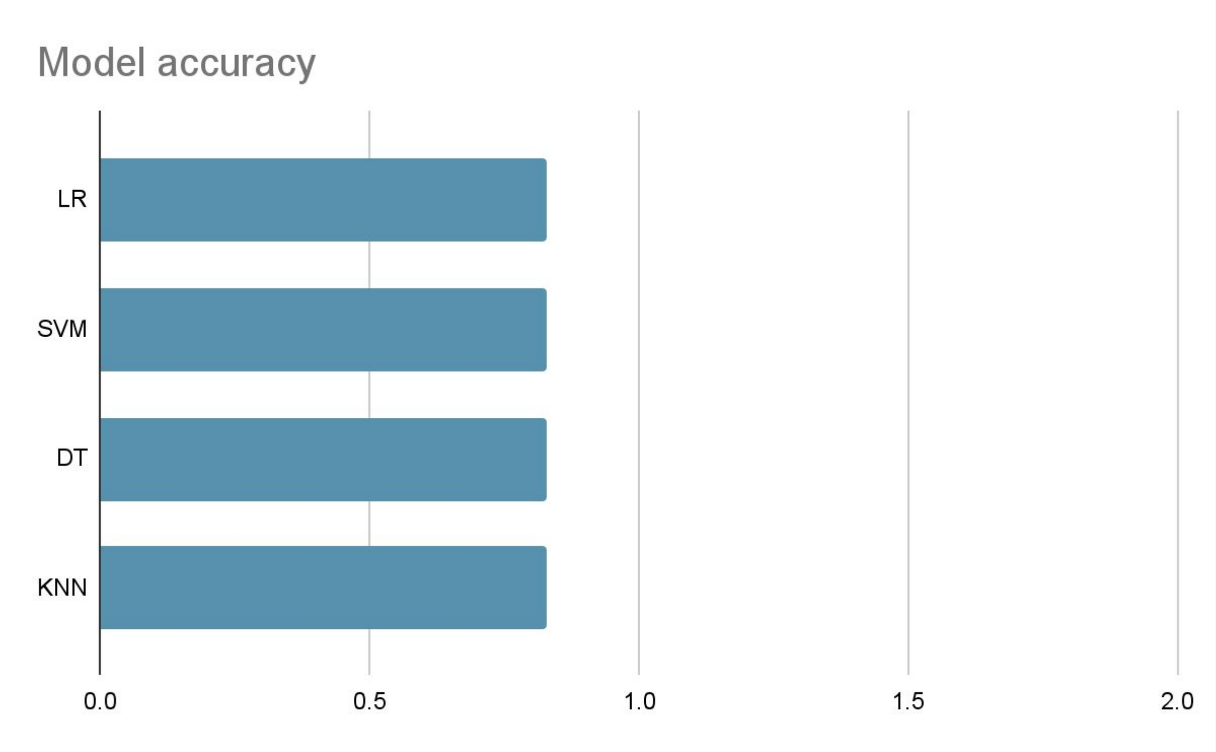


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All four models predicted the data with 83% accuracy
- Equal score means there is no clear winner

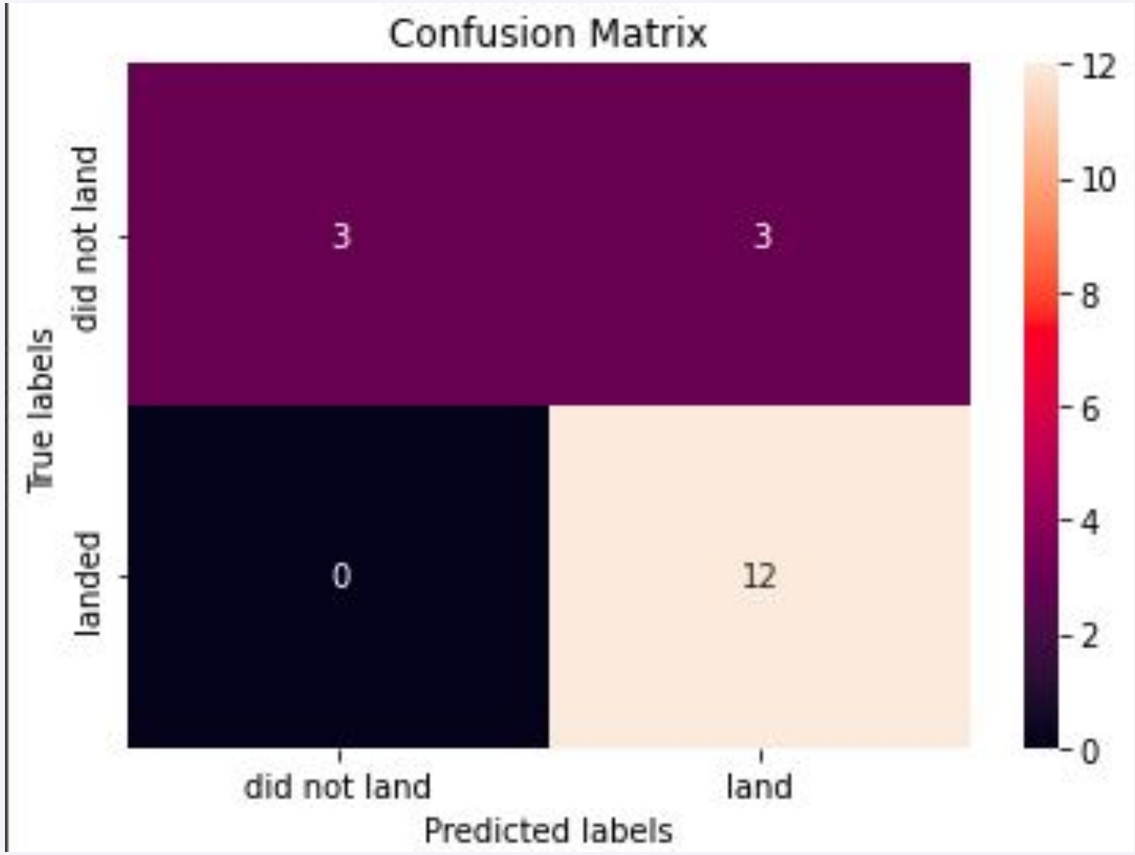


Notebook

Confusion Matrix

The 3 and 12 in the true positive and true negative boxes indicate our models can distinguish between the different classes.

The 3 in the 0 indicates a problem with false positives (Type 1 error).



[Notebook]

Conclusions

- Falcon 9 launches from four sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40, all located on the coast and close to points of transportation
- Success rates decrease as payload mass reaches higher-end, with diversity in mid-range
- The four orbits have high success rates: ESC-1, GEO, HEO, and SSO
- Most launches from the KSC-LC-39A Falcon 9 site
- No definitive classification model winner

Discussion

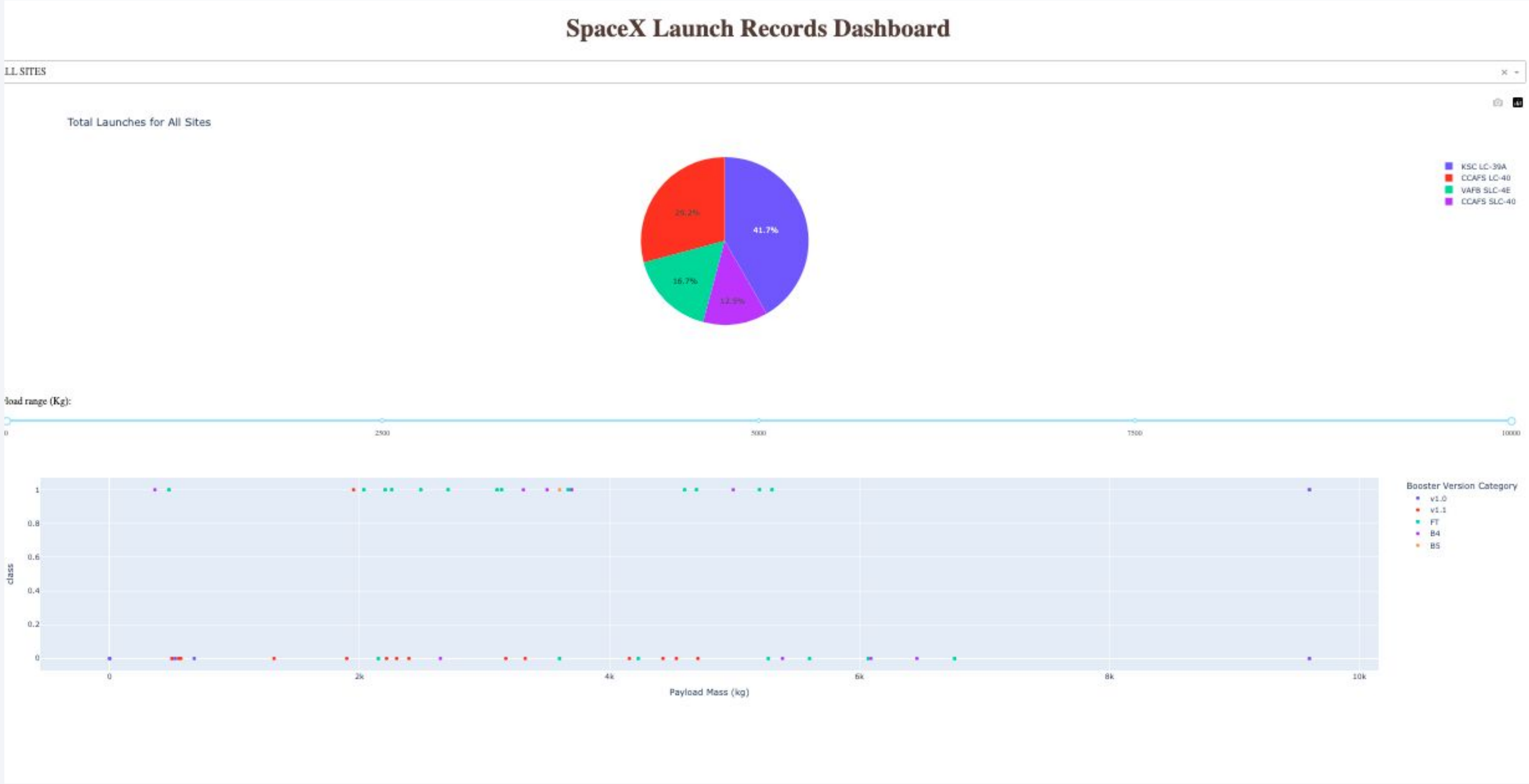
- Further model trainings needed to determine best classification model
- Orbit type and payload showed no clear trend suggesting evaluation of relationship
- Relationship with in mid-range payload varies suggesting more analysis
- Downward trend in successful launches since 2020 should be an area of focus to understand source

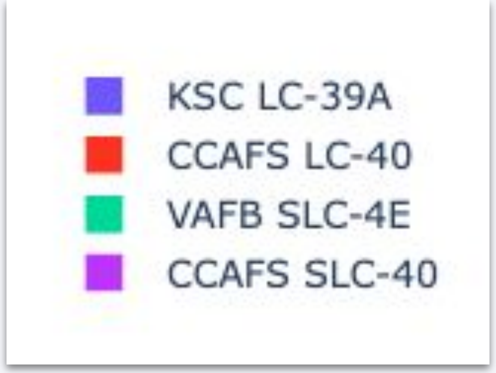
Appendix A: GitHub Notebook

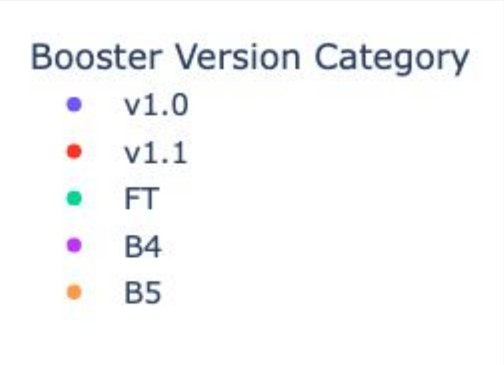
Find all notebooks: https://github.com/anastasia-reyes/capstone

- Week 1 Data Collection API: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_1_FINAL_Data_collection_API.ipynb
- Week 1 Data Collection web scraping: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_1_FINAL_Collecting_Data.ipynb
- Week 1 Data Wrangling: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_1_FINAL_Wrangling_Data.ipynb
- Week 2 EDA:
- https://github.com/anastasia-reyes/capstone/blob/master/WEEK_2_FINAL_EDA_Visual.ipynb
- Week 3 Dashboard: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_3_Dashboard_FINAL.ipynb
- Week 3 Folio: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_3_Visualization_Folio.ipynb
- Week 4 ML: https://github.com/anastasia-reyes/capstone/blob/master/WEEK_5_FINAL_ML_.ipynb

Appendix B Dashboard screenshots







Appendix: Google Collab

To access notebooks outside GitHub, find them within the folder:
https://drive.google.com/drive/u/0/folders/1fqDoznIv9qPgrsJNP81vDLTC7W2ZVXgo

Thank you!