Winning Space Race with Data Science

Jonathan Presto
November 7, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected using Web Scraping tool (BeautifulSoup) and SpaceX REST API.
  - Performed data extraction/wrangling, exploratory data analysis (EDA) and data visualization.
  - Built a machine learning classification model to predict success/fail of launch.
- Summary of all results
  - We successfully scraped data from SpaceX Wikipedia site and API calls.
  - EDA provided insights to help select most useful features.
  - Evaluated 4 classification algorithms (Logistic regression, SVM, Decision Trees, KNN) and concluded that all algorithms performed equally well on the test dataset with an accuracy score of 83.3%.
  - Based on the accuracy metric, we can predict that SpaceX launches will reuse its first launch 83% of the time, where each launch from reused components will cost $62 million.

Introduction

- The main objective of this project is to evaluate the capabilities of the new company Space Y to compete with Space X.
- The problems we would like to answer are:
  - Determine the price for each Space X's Falcon 9 launch.
  - Create interactive dashboards that will gain insights on potential features.
  - Using public information, we want to predict if SpaceX will reuse the first stage on subsequent launches so we can estimate the total cost for launches. We would like to know the total cost for Space X launches in order to provide Space Y some insights to budget their launch programs competitively.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Using the SpaceX REST API, we extracted data from https://api.spacexdata.com/v4/
  - Data was also sourced from the Wikipedia site using Python web-scraping tools: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Dataset was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features. The features were also standardized (scaled) to ensure coefficients were stable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - GridSearch method was used for hyperparameter tuning to yield best performing model on the cross-validation set. A confusion matrix was used to evaluate the performance on the test dataset.
  - Evaluated 4 classification algorithms (Logistic regression, SVM, Decision Trees, KNN) and concluded that all algorithms performed equally well on the test dataset with an accuracy score of 83.3%.

Data Collection

- Using the Space X REST API, we ingested JSON files sourced from https://api.spacexdata.com/v4/
- Data from the Wikipedia site https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 was scraped using Python packages `BeautifulSoup4` and `Requests`.

Data Collection – SpaceX API

Provision Tools:

1. SpaceX REST API
2. Python libraries: requests, pandas, numpy, datetime

Request and parse the SpaceX launch data using the GET request.

Filter the dataset to only include Falcon 9 launches.

Replace missing values with the mean of its data column.

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

Provision Tools:

1. Python libraries: requests, BeautifulSoup, re, pandas, unicodedata

Using Falcon9 Launch HTML page as input, submit a HTTP GET request, then create a BeautifulSoup object.

Extract column/variable names from the HTML table header

Parse the launch HTML tables, then create a Pandas DataFrame from it.

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

Preliminary EDA

- Identify and calculate percentage of missing values
- Understand column datatypes

Summarizations

- Calculate number of launches at each site
- Calculate number and occurrence of each orbit
- Calculate number and occurrence of mission outcome per orbit type

Label Creation

- Create a label for landing outcome for supervised model training

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Scatterplots, Barplots and Line charts helped visualize the relationship between 2 features, and how that particular relationship correlated with launch outcome (Class).

- Summary of Scatterplots:
  1. FlightNumber vs PayLoadMass
  2. FlightNumber vs LaunchSite
  3. PayLoadMass vs LaunchSite
  4. FlightNumber vs Orbit Type
  5. PayloadMass vs Orbit Type

- Summary of Barplots/Line chart:
  1. Orbit Type vs Success Rate
  2. Year vs Success Rate

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- Summary of SQL queries you performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used to build a Folium Map:
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, such as NASA Johnson Space
  - Marker clusters indicate groups of events in each coordinate, like launches in a launch site
  - Lines indicate distances between two coordinates

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

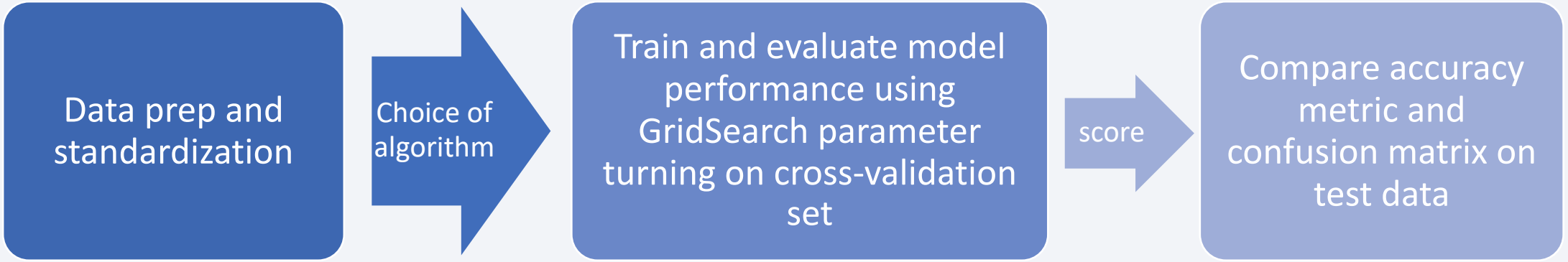
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site by outcome
  - Payload range by outcome by booster type
- This combination allowed us to quickly analyze the relationship between payloads, outcomes and launch sites, to help identify where the best place is to launch according to payloads

Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- After preprocessing and standardizing the dataset, we experimented with 4 classification algorithms: logistic regression, support vector machine (SVM), decision tree and K-nearest-neighbor.
- For each algorithm, we performed GridSearch to find the most optimal parameters using cross-validation technique on the training set.
- We compared overall accuracy of the 4 models against the test dataset to come up with the best model.



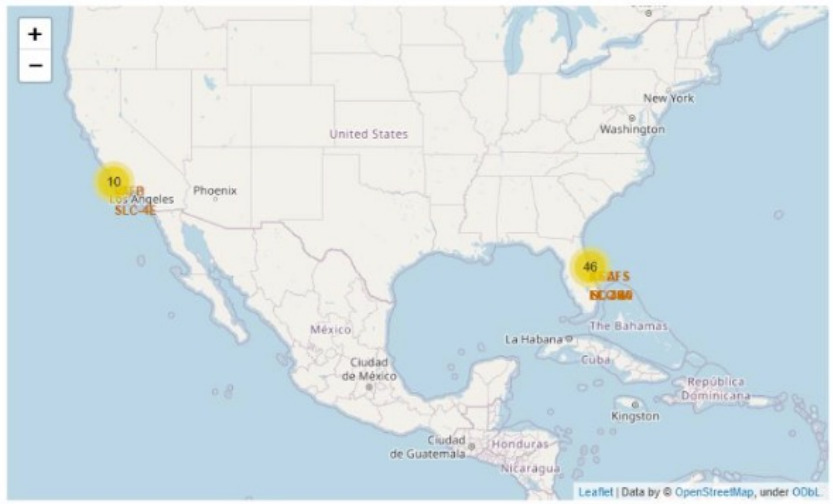
Source code: https://github.com/jonpresto/IBM-Applied-Data-Science-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites
  - The average payload of F9 v1.1 booster is 2928 kg
  - The first success landing outcome happened in 2015, five years after first launch
  - Many Falcon 9 booster versions successfully landed in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed to land in drone ships in 2015; F9 v1.1 B1012 and F9 v1.1 B1015
  - Landing outcomes improved in subsequent years

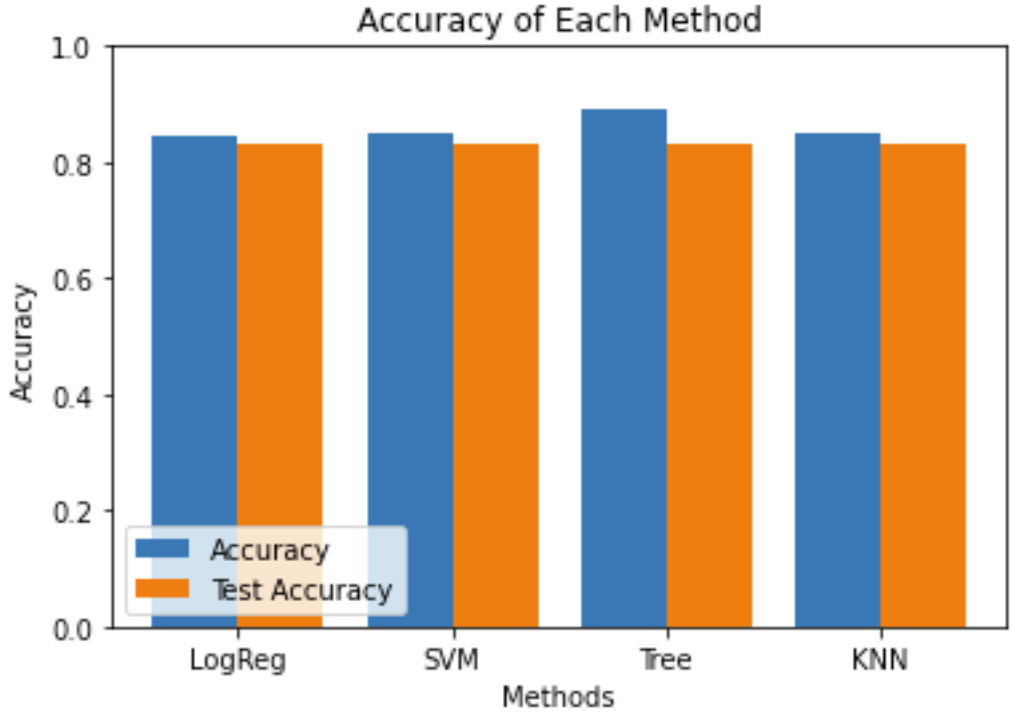
Results

- Using interactive analytics, were able to ascertain that launch sites tend to occur in places near the ocean and favorable weather conditions
- Most launch sites occur at Florida coast

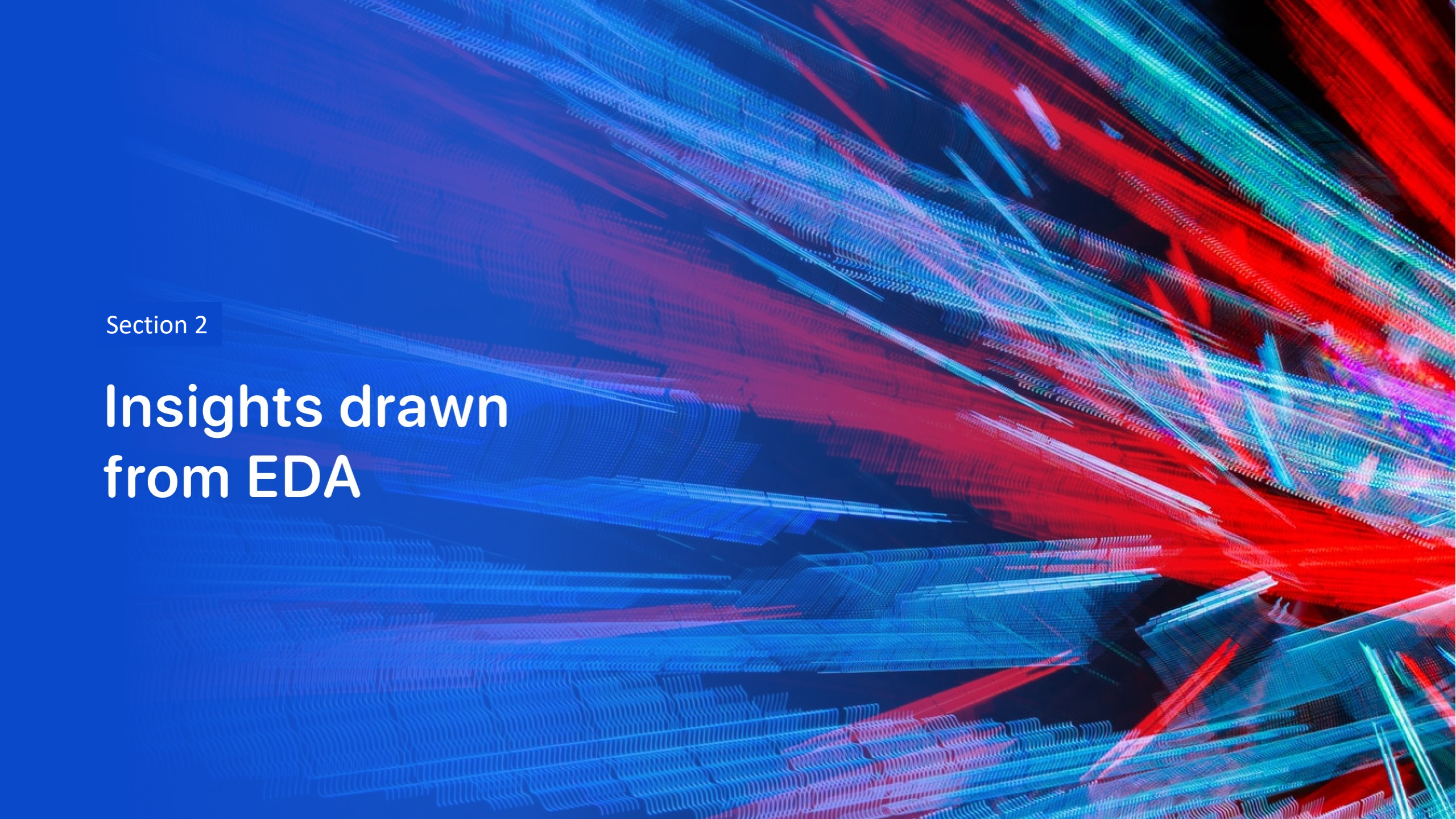


Results

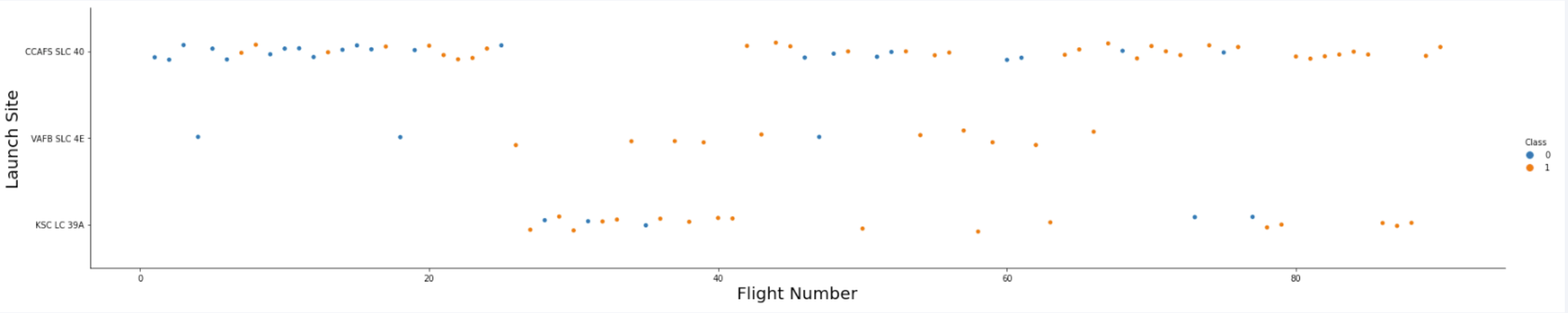
- Predictive analytics suggest that all four models perform equally well on the test dataset, having an accuracy of 83.333%



Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333

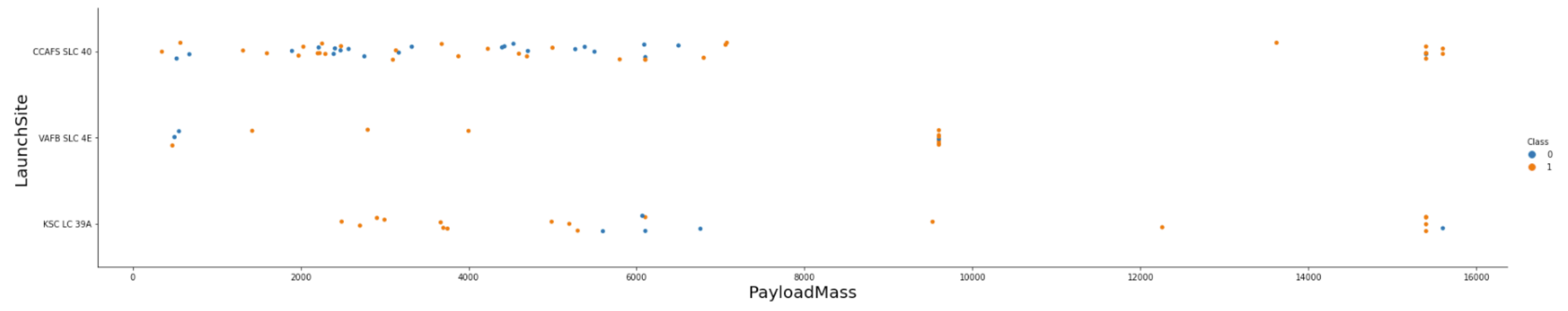

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The best launch site happens to be CCAF5 SLC40, where most recent launches were successful

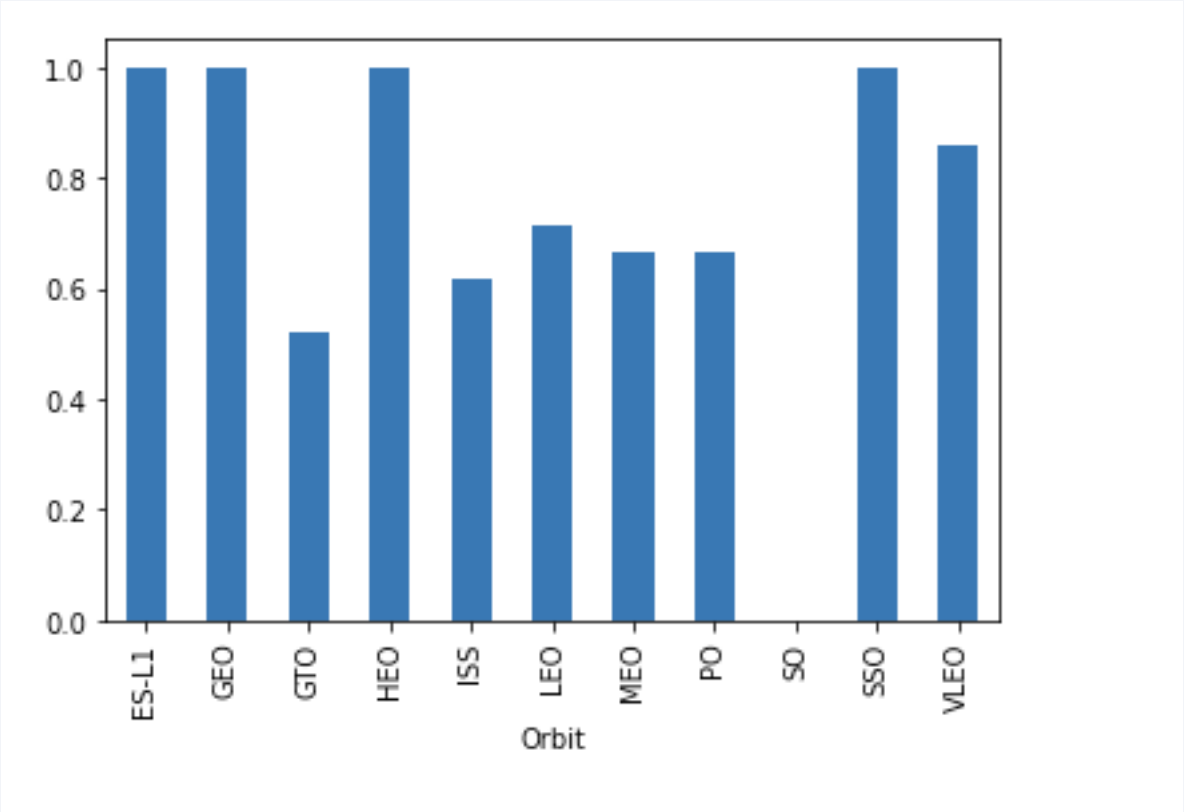
Payload vs. Launch Site



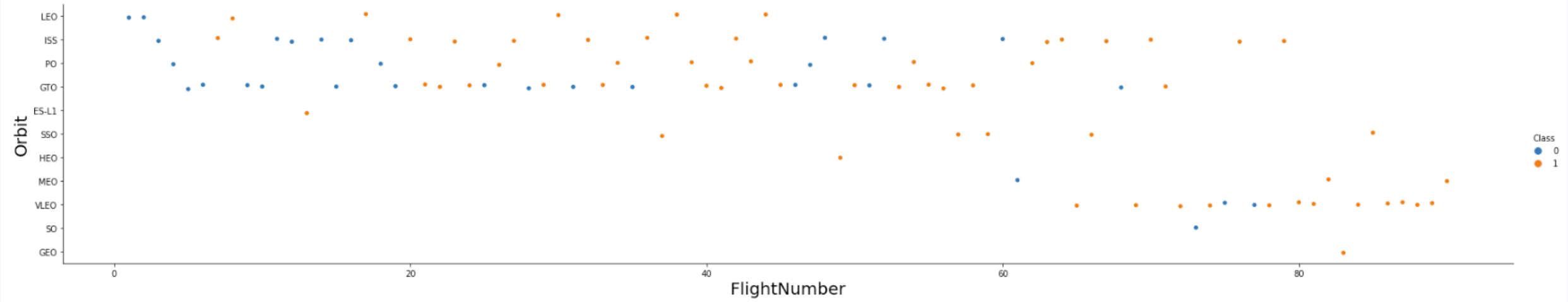
- KSC LC-39A (10 successful launches) had the largest successful launches
- Payload ranges with the highest launch success rate were: 3K-4K, 4.5K-5.5K
- Payload ranges with the lowest launch success rate were: 1K-2K, 4K-4.5K, 5.5K-7K

Success Rate vs. Orbit Type

- The highest success rates are those to orbit:
  - ES-L1
  - GEO
  - HEO
  - SSO

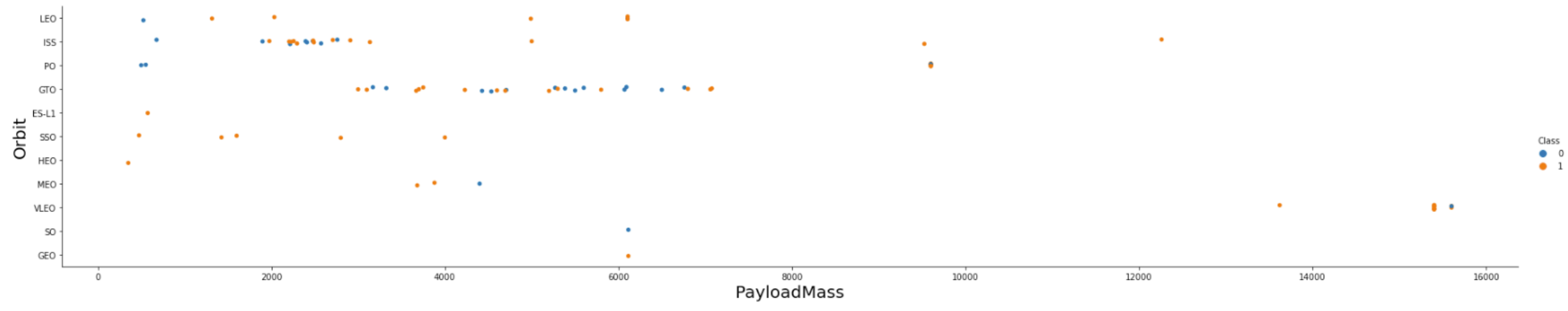


Flight Number vs. Orbit Type



- VLEO and ISS orbits appear to show increasing success rate in recent launches

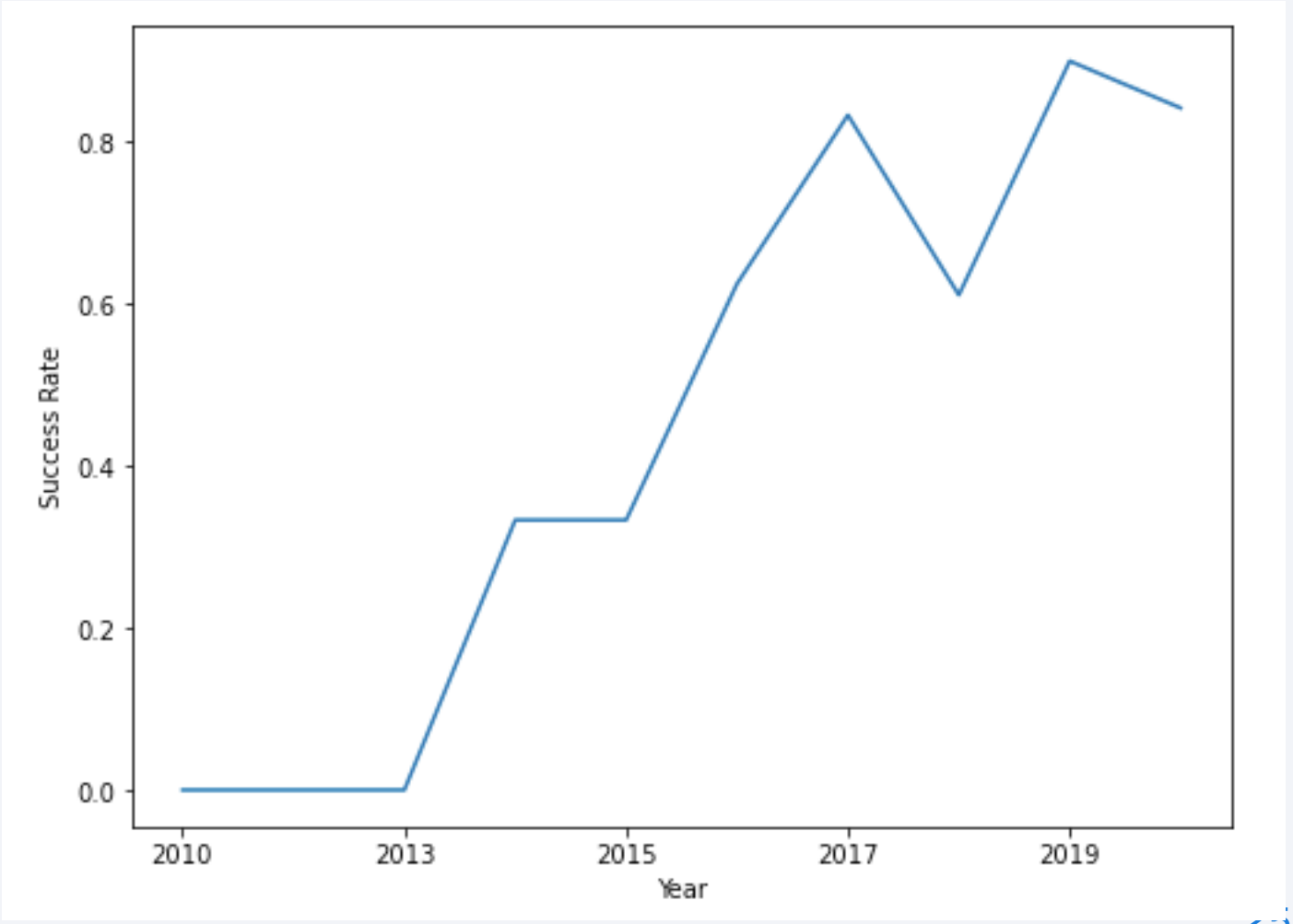
Payload vs. Orbit Type



- With heavy payloads, the successful landing is more prevalent in Polar, LEO and ISS
- Cannot see any improvement in successful landings with increasing payload for GTO

Launch Success Yearly Trend

- We observe increasing success rates from 2013 to 2020
- This is indicative of adopting lessons learned from earlier launches



All Launch Site Names

- There are four launch sites extracted from the dataset
- They were obtained by using a SELECT statement with DISTINCT option:

In [6]:
```
%%sql

SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

```
 * ibm_db_sa://glj70712:***@1bbf73c5-d84a-
Done.
```

Out[6]:

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

Launch Site Names Begin with 'CCA'

- Here are 5 examples where launch sites begin with `CCA`
- They were obtained using SELECT statement with WHERE filter clause:

```
%%sql

SELECT * FROM SPACEX
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

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

- The total payload carried by boosters from NASA was 45,596KG
- The calculation was executed by summing all payloads where customer equals ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version was F9 v1.1 was 2,928KG
- The calculation was obtained by averaging payload mass after filtering for the booster version using WHERE clause statement

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened on December 22, 2015
- We obtained the date by applying the MIN function on the date field after applying the WHERE filter clause for only records having successful landing outcomes

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were:

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- We queried this using SELECT DISTINCT statement on booster version and applied the WHERE filter clause on the payload range of interest

Total Number of Successful and Failure Mission Outcomes

- The number of successful and failure mission outcomes

Mission Outcome	Count
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- The numbers were obtained by grouping on mission outcome then taking the count of records in each bucket

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

Booster Version	Booster Version
F9 B5 B1048.4	F9 B5 B1051.4
F9 B5 B1048.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1056.4
F9 B5 B1049.5	F9 B5 B1058.3
F9 B5 B1049.7	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1060.3

- These boosters were obtained using a WHERE clause that contained a subquery of an inner select statement where the payload equaled the MAX payload.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015:

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Query was obtained by filtering where landing outcome was equal to ‘Failure (drone ship)’ and year of date was 2015

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

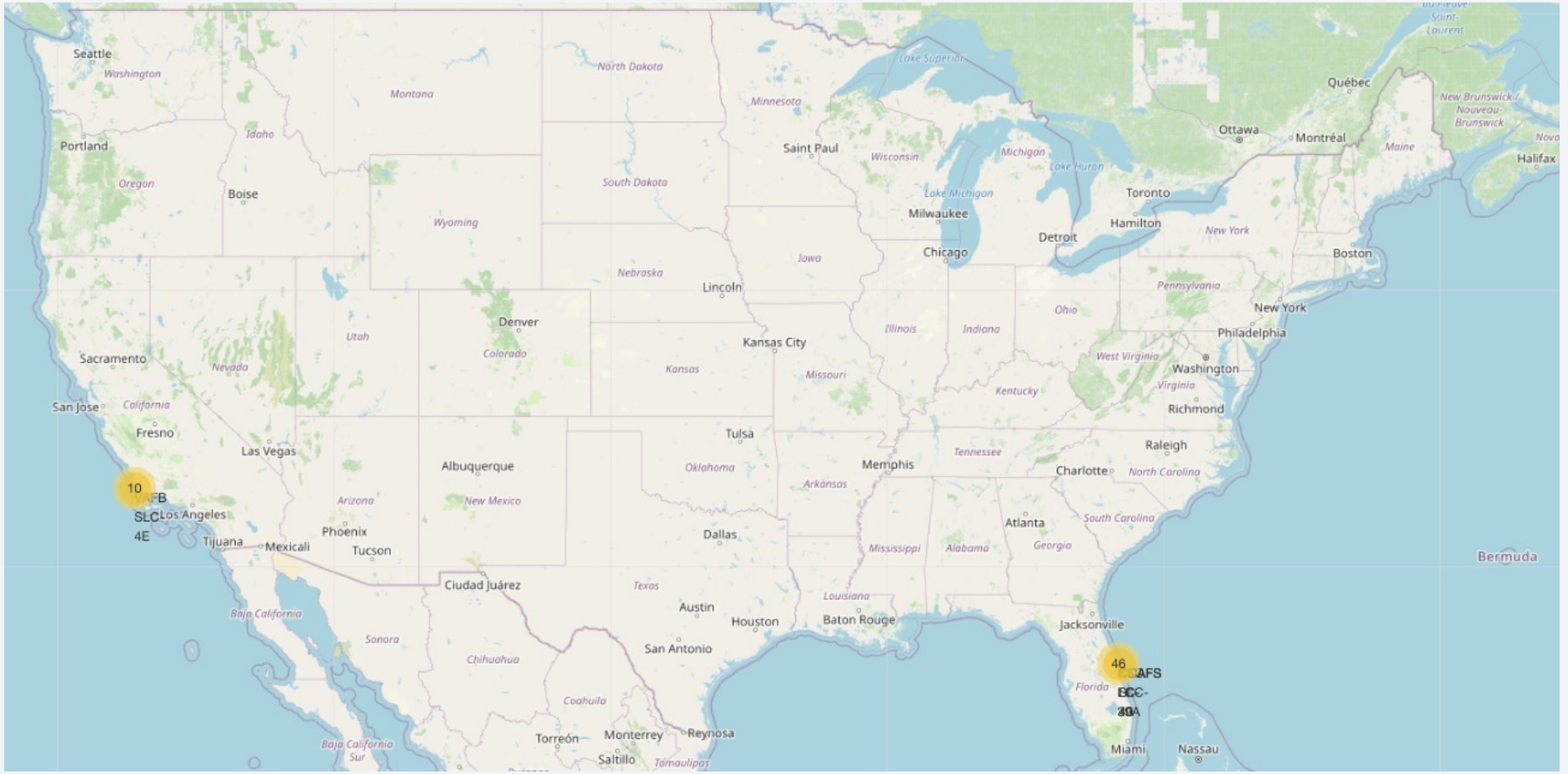
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- We grouped the “date-filtered data” by landing outcome then aggregated by count in descending order

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


Section 3
Launch Sites
Proximities Analysis

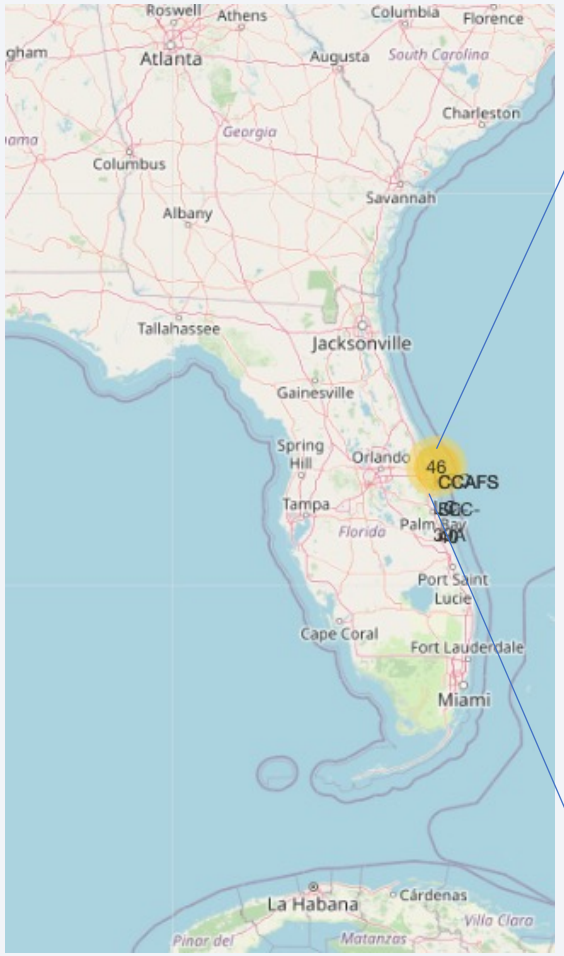
All Launch Sites

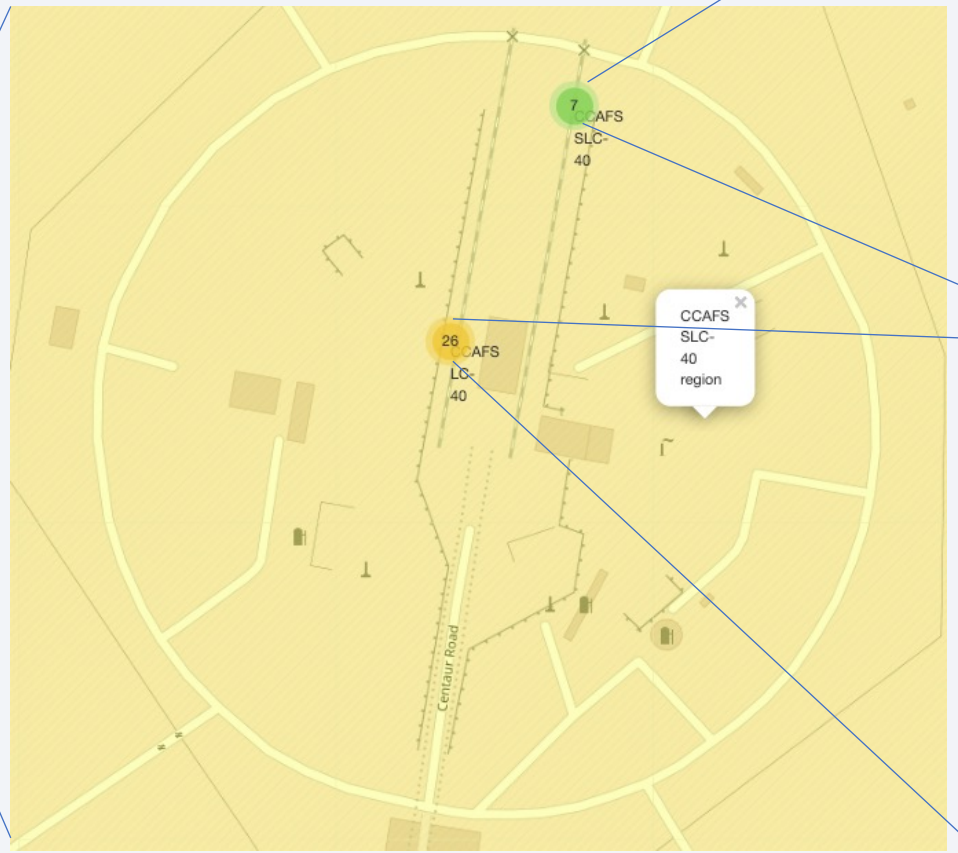
- Launch sites below are near sea, away from large populations, and reasonable distance from roads and railroads to minimize potential harm to infrastructure

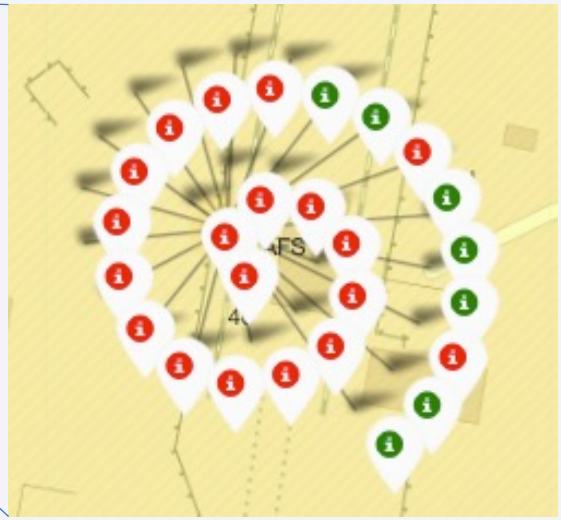


Launch Outcomes by Site

- Example of landing outcomes at site CCAFS SLC-40
- Green markers indicate success and red indicate failure

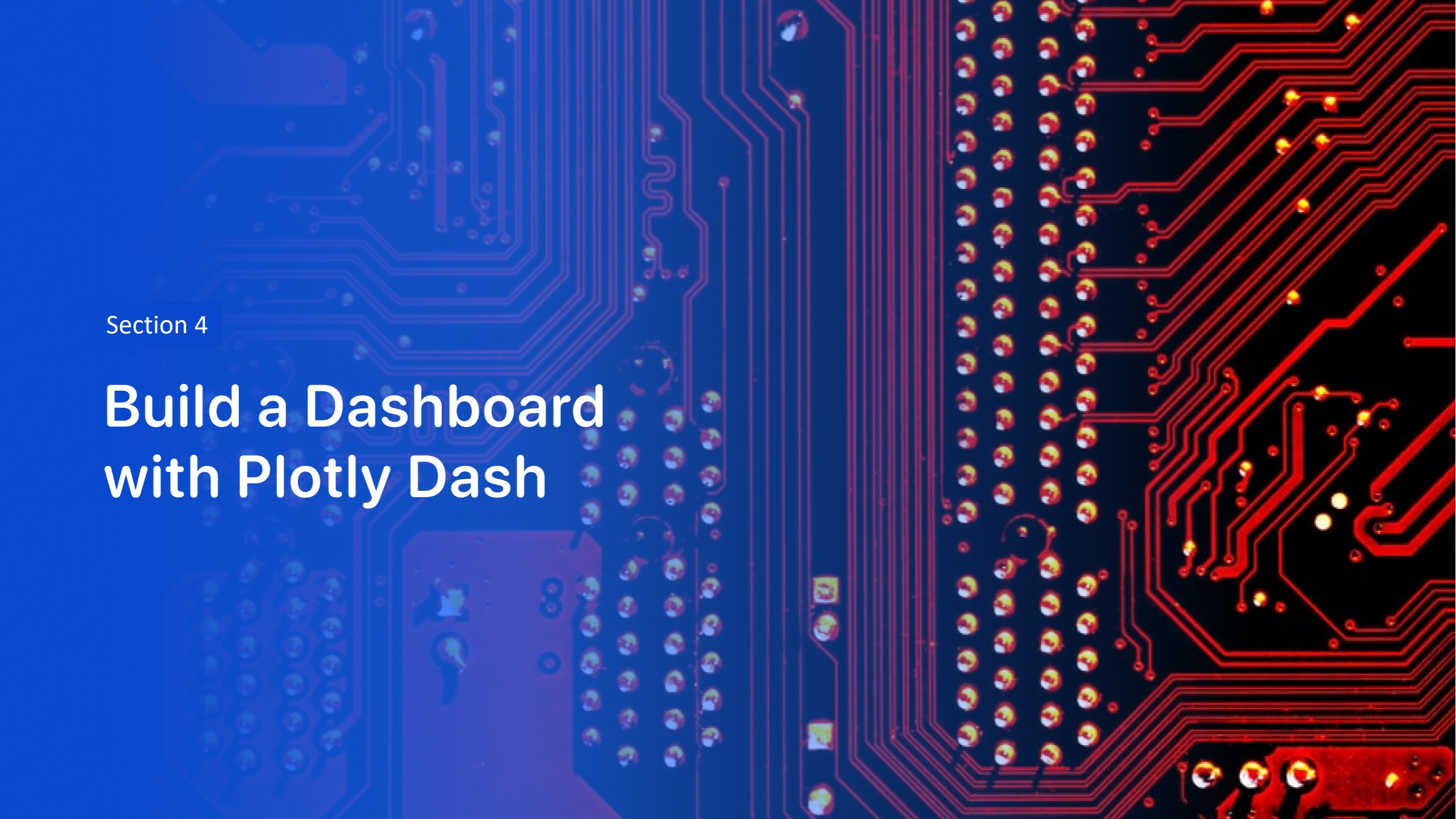






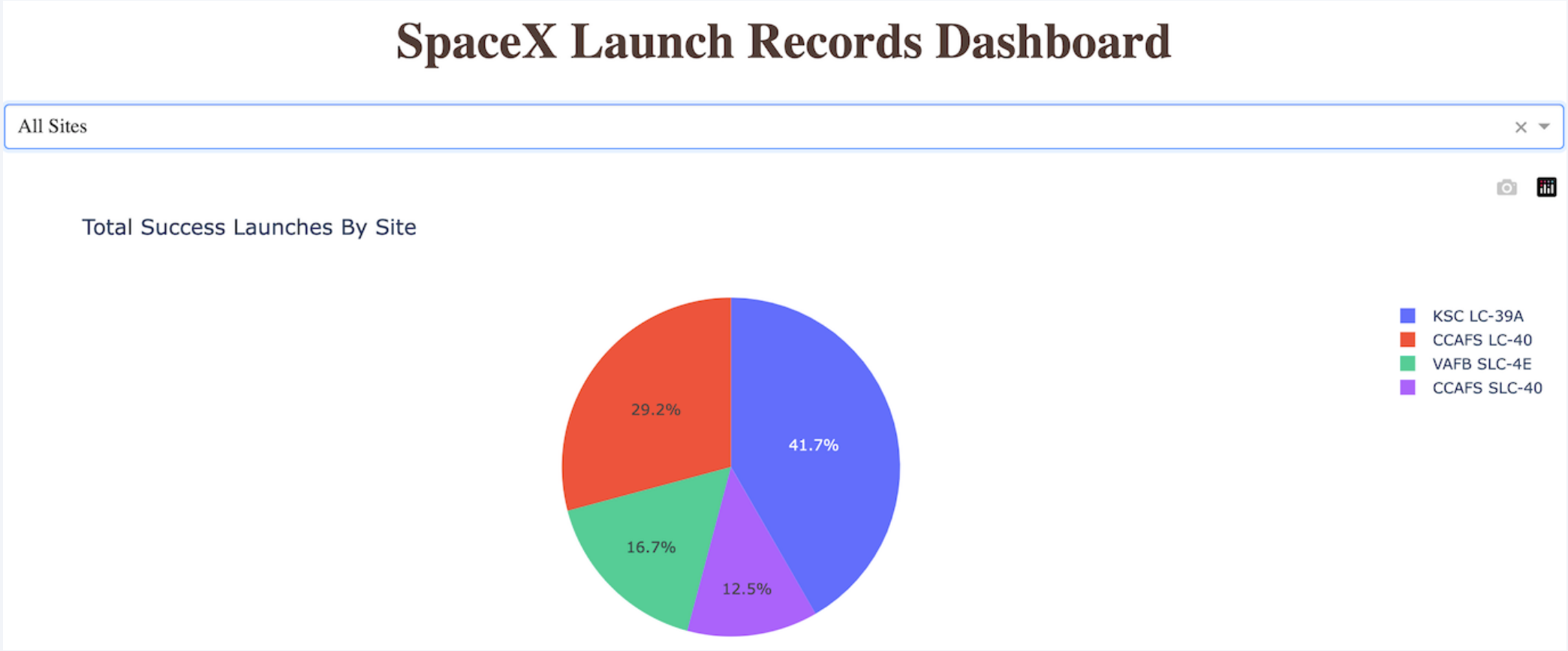
Logistics and Safety

- Launch site CCAFS SLC-40 has good logistics aspects being near coast line and railway lot, and relatively far from inhabited areas


Section 4
Build a Dashboard with Plotly Dash

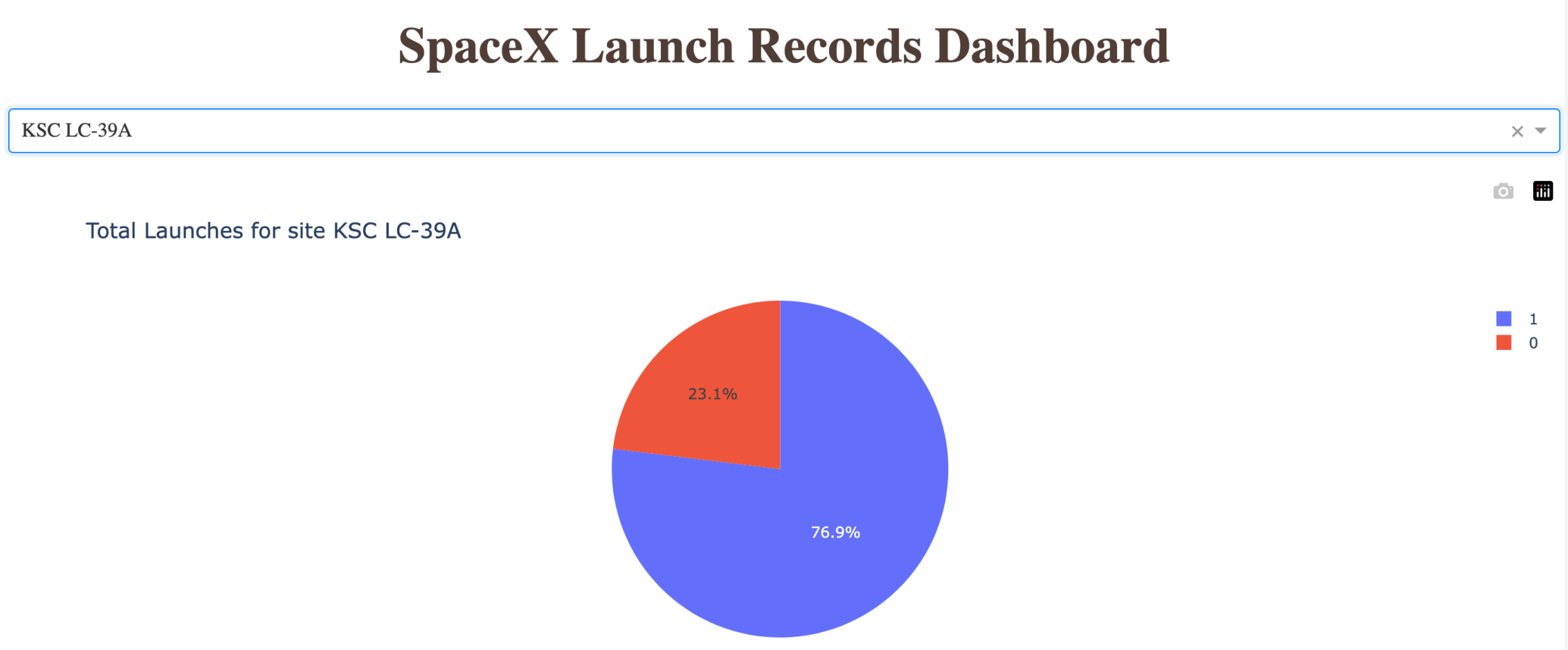
Launch Success Count for all Sites

- Most successful launches occurred at site KSC LC-39A



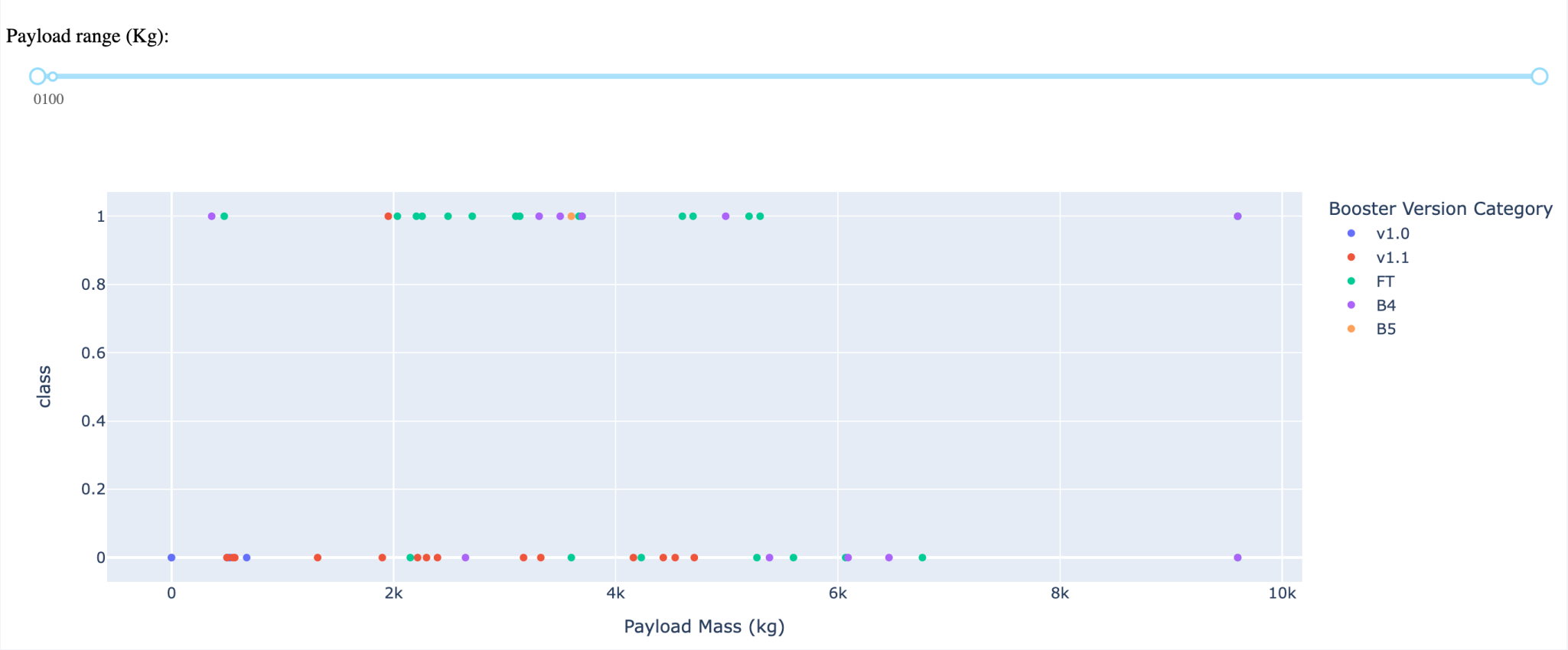
Launch Site with Highest Success

- 76.9% success rate at site KSC LC-39A



Payload vs Launch Outcome

- Payload ranges (KG) with the highest launch success rate: 3K-4K, 4.5K-5.5K
- Payload ranges (KG) with the lowest launch success rate: 1K-2K, 4K-4.5K, 5.5K-7K



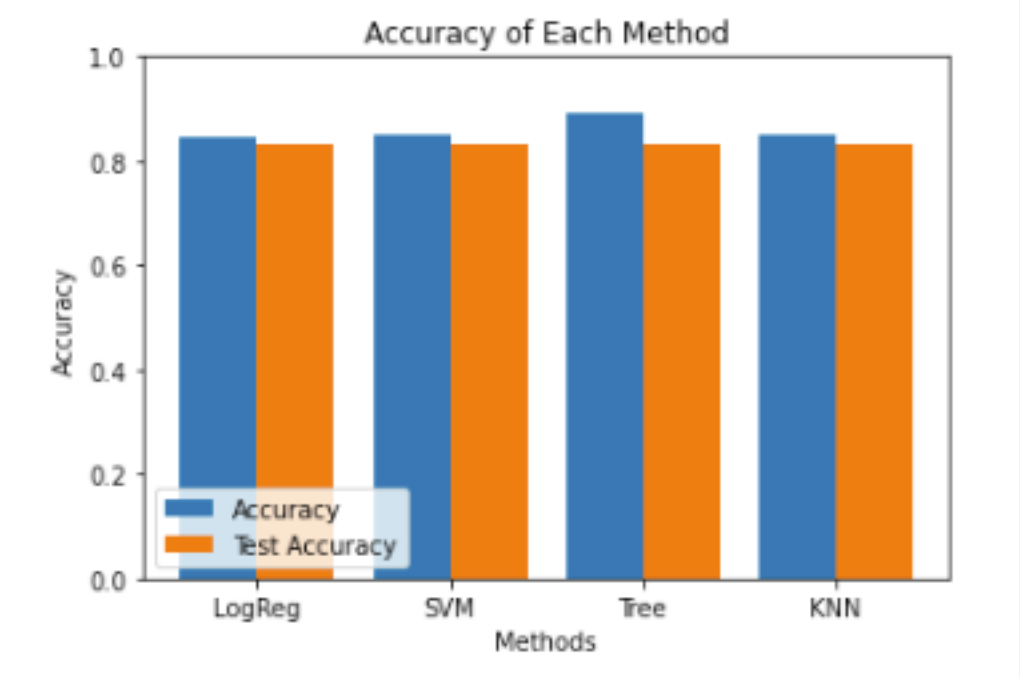
Section 5

Predictive Analysis (Classification)

Classification Accuracy

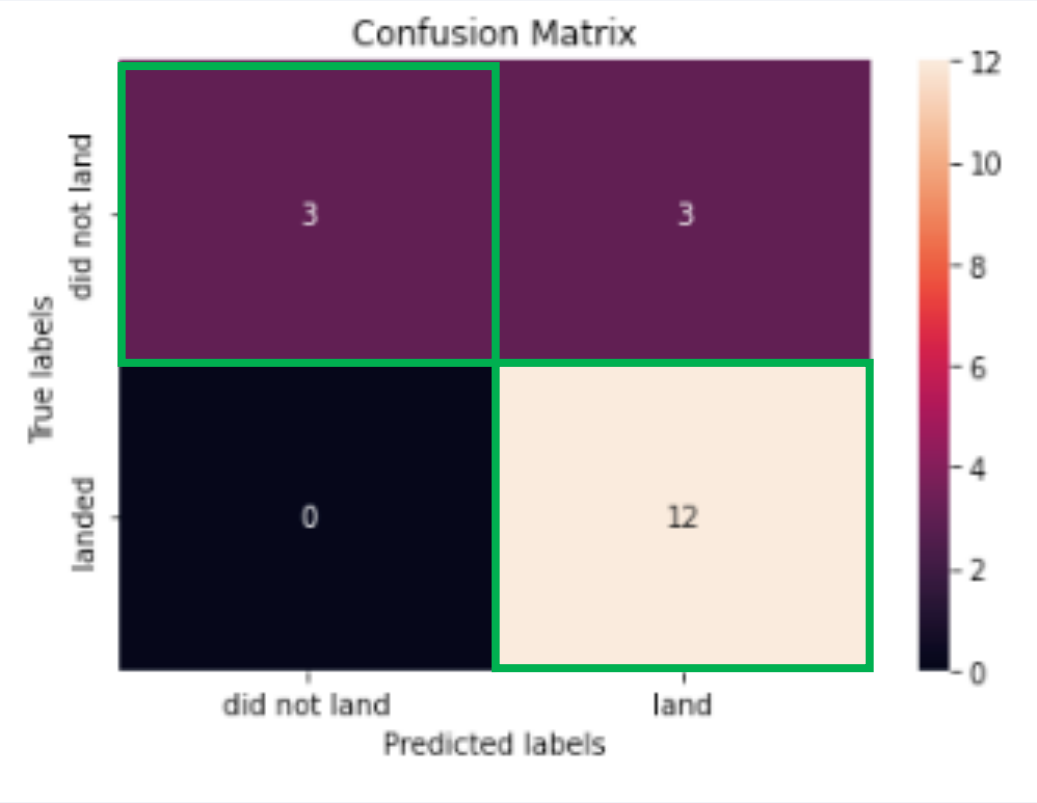
- Four classification models were evaluated by accuracy on training (blue) and test set (orange)

- All 4 models performed equally well on the test set (83.33% accuracy), but the decision tree performed the highest on the training set (88.93% accuracy)



Confusion Matrix of Decision Tree Classifier

- Numbers in cells True Positive and True Negatives are relatively higher compared to cells in False Positive and False Negatives.



Conclusions

- Different data sources where analyzed looking into various factors available from public data (e.g., launch site, payload, booster type, orbit type, etc.)
- KSC LC-39A had the highest success rate for launches
- Payloads in the range 3K-4K, 4.5K-5.5K had the highest success rate
- F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate having13 successful launches.
- Most mission outcomes where successful, but most successful occurrences occurred in more recent years; this makes sense as engineers improve processes from past failures.
- Although the decision tree classifier performed the best in the training set, we should reconsider evaluating all 4 algorithms in future with more historical data and base our decision against the test set; ensure that the models are not overfitted or not underfitted.

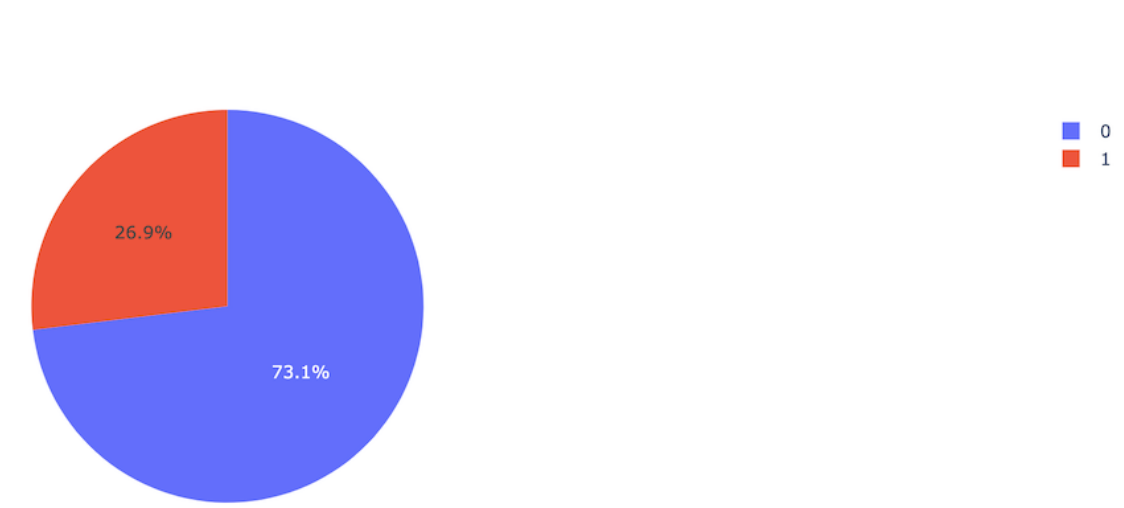
Appendix

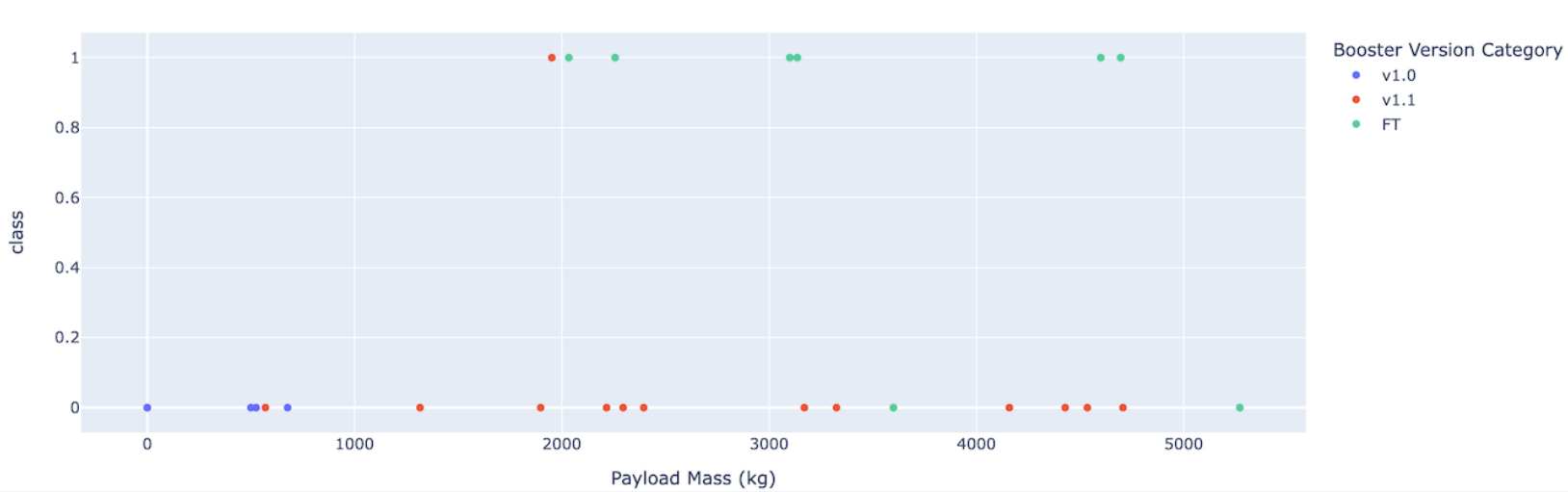
Preview of HTML table from Wikipedia site

Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
	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)
1	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]								
	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)
2	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.								
	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
3	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	
4	CRS-1 was successful, but the secondary payload was inserted into an abnormally low orbit and subsequently lost. This was due to one of the nine Merlin engines shutting down during the launch, and NASA declining a second reignition, as per ISS visiting vehicle safety rules, the primary payload owner is contractually allowed to decline a second reignition. NASA stated that this was because SpaceX could not guarantee a high enough likelihood of the second stage completing the second burn successfully which was required to avoid any risk of secondary payload's collision with the ISS.[26][27][28]								

Appendix

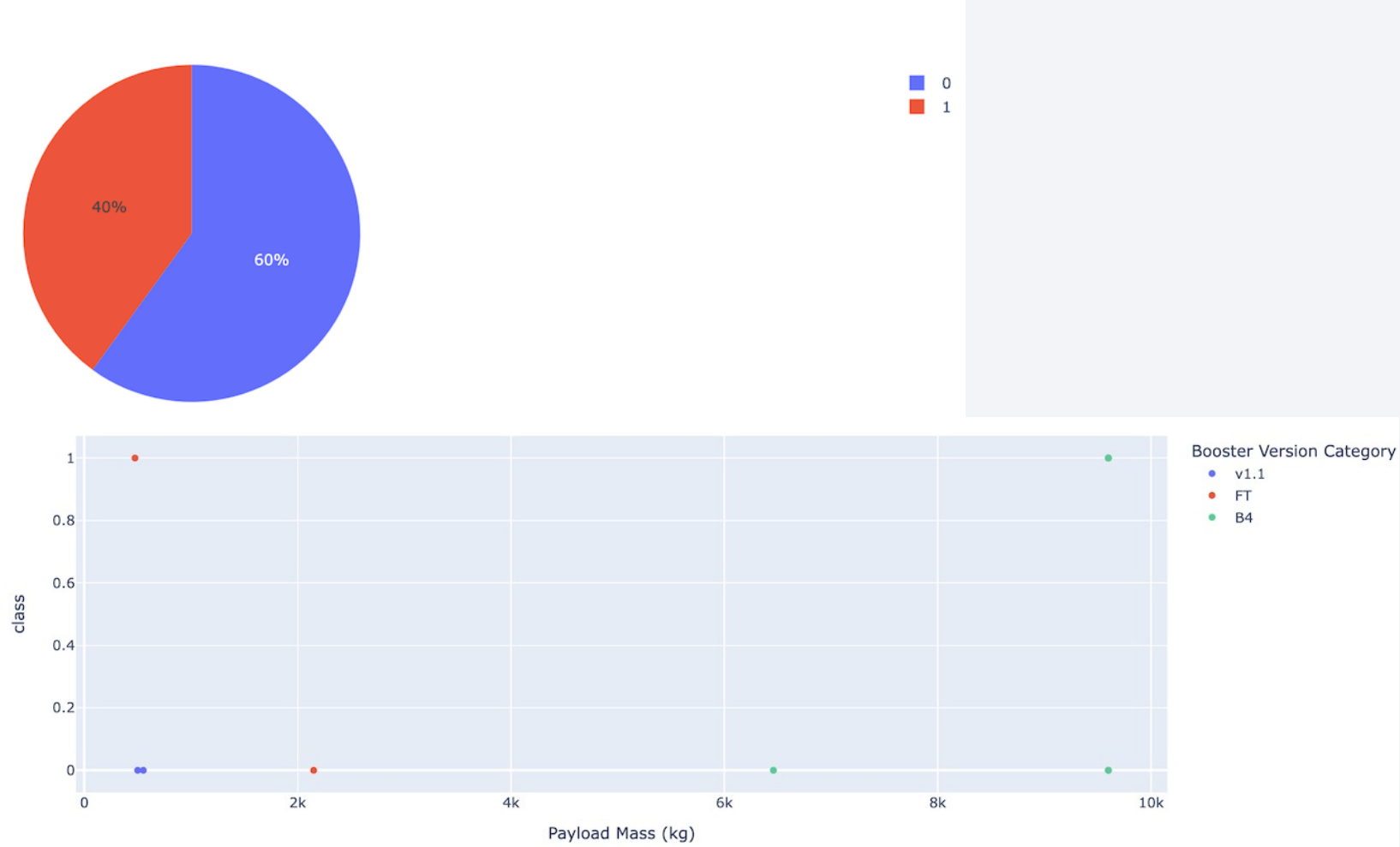
Launch Records Dashboard: CCAFS LC-40





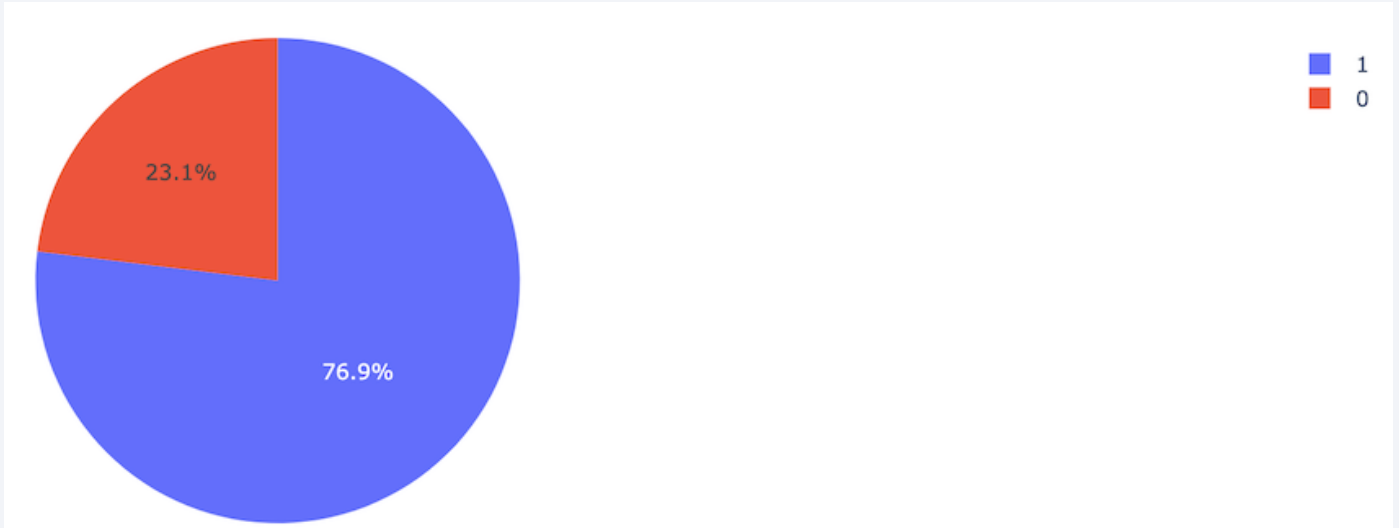
Appendix

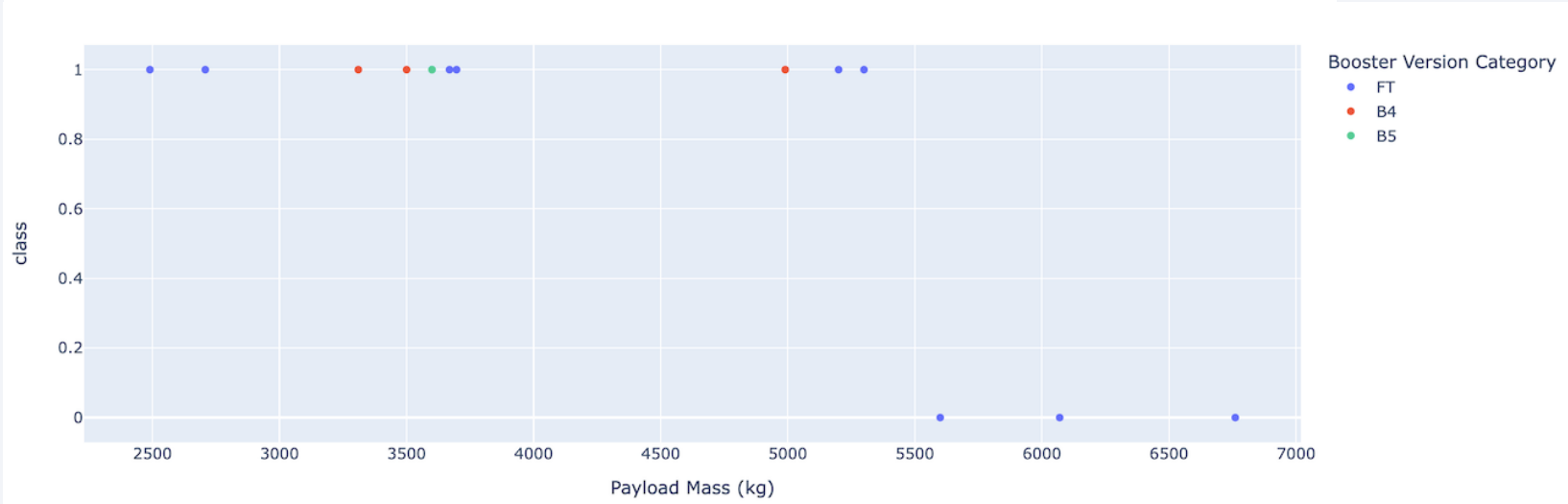
Launch Records Dashboard: VAFB SLC-4E



Appendix

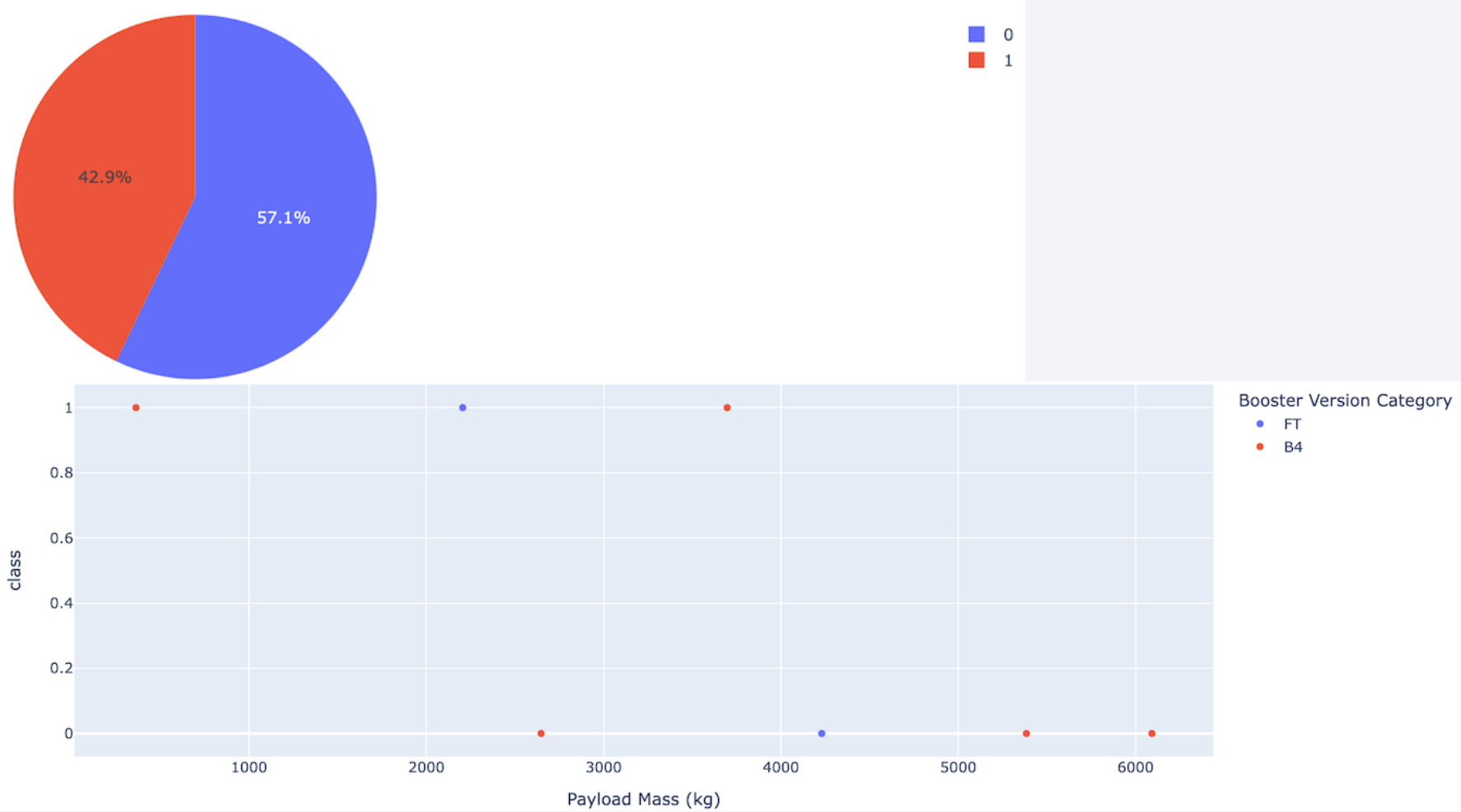
Launch Records Dashboard: KSC LC-39A





Appendix

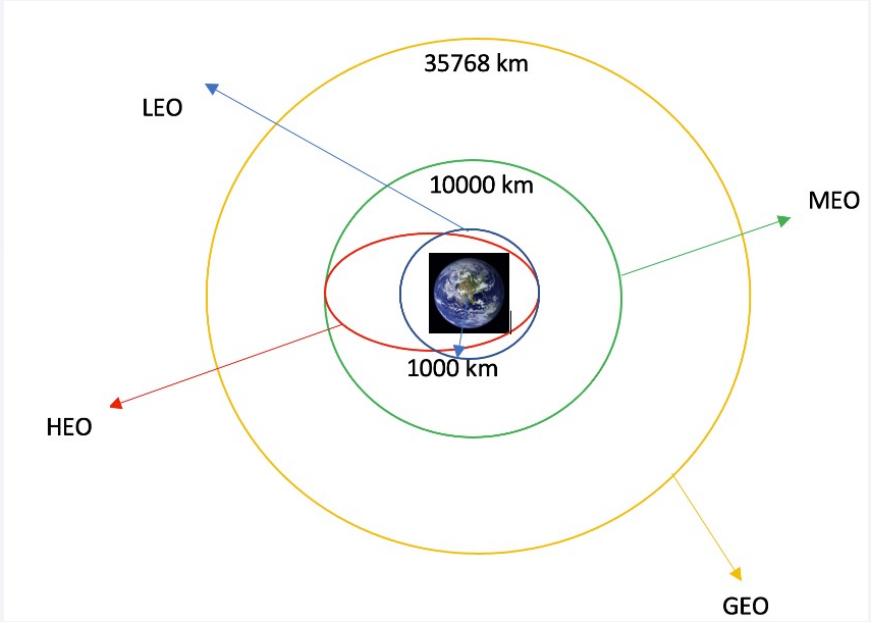
Launch Records Dashboard: CCAFS SLC-40



Appendix

Each launch aims to a dedicated orbit. Some common orbit types listed below:

- **LEO**: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].
- **VLEO**: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].
- **GTO** A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .
- **SSO (or SO)**: It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .
- **ES-L1** :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .
- **HEO** A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].
- **ISS** A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]
- **MEO** Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]
- **HEO** Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]
- **GEO** It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]
- **PO** It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11]



Source: https://en.wikipedia.org/wiki/List_of_orbits

Thank you!