IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael Paulsen
9/11/2025

Outline

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

Executive Summary

- SpaceX has paved the way for the future of space flight to be not only more efficient but with considerable cost savings compared to competitors like Blue Origin and Virgin Galactic.
- SpaceX Falcon 9 rocket launches cost around $62 million vs. $165+ million for competitors because of the reusability of the first stage booster.
- Continual improvements and experience have greatly increased the landing success rates of the first stage boosters.
- Exploratory Data Analysis (EDA) reveals that the highest success rates are correlated to a combination of the launch site itself, the payload mass and the orbit type, such that we can predict with a accuracy rate of 83% a successful landing of the booster.

Introduction

Can we predict whether a given Falcon 9 rocket launch will have a successful first stage booster landing?

We will determine if the landing will be successful for a given launch, based on the evaluation of a number of factors: launch site, payload mass and orbit type.

Exploratory Data Analysis (EDA) was conducted using a combination of data sources and techniques, including interactive dashboards and predictive analysis with machine learning.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via the provided SpaceX API as well as web scraping historic launch records from Wikipedia. These were then outputted to CSV files for later use.
- Perform data wrangling
  - From the data gathered, analysis was performed on the launch counts from each site, the counts of each orbit type, as well as the landing outcomes to create a column that stores either 0 for failure or 1 for success, which we can use to train a classification model later.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - SQL was used to perform additional EDA, exploring unique launch sites, maximum and average payload mass, number of successful mission outcomes, booster versions, etc.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Using Folium and Plotly Dash, an interactive dashboard was created so that launch sites and payload mass range could be explored interactively to display a scatter plot of successful and failed landings.
- Perform predictive analysis using classification models
  - Four classification models were evaluated using a training and test set of the data to determine the accuracy of the landing results and hopefully determine which model performed best

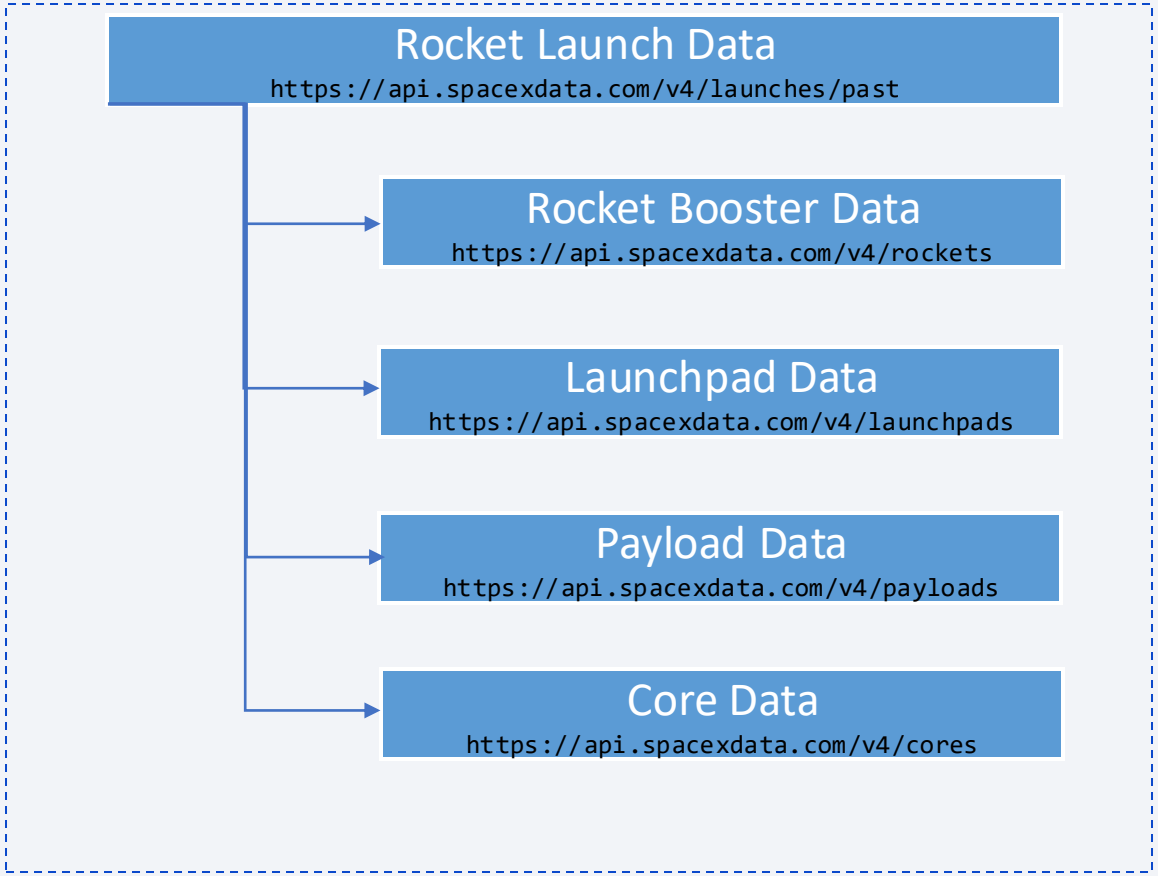
Data Collection

- You will see in the next series of slides how the data was collected

Data Collection – SpaceX API

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

- Using the rocket launch data, collect additional booster, launchpad, payload and core data.
- Combine the data into a single DataFrame and then filter on only Falcon 9 rocket data
- **Fields:** FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Gridfins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Use the average payload mass to fill in any missing PayloadMass values
- Export to CSV



Data Collection - Scraping

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/jupyter-labs-webscraping.ipynb

- Use the Wikipedia page to scrape data using BeautifulSoup:
  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The data we want is in the 3rd table on the page listing: Flight No., Date/time, Booster Version, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Landing Outcome
- Using a number of helper functions, we will extract all the relevant data, and split Date/Time into 2 separate fields
- Store as a DataFrame and export to CSV

For each row of scraped flight data, parse and append:

- Flight No.
- Date, Time
  Using date_time helper function
- Booster Version
  Using booster_version helper function
- Launch Site
- Payload
- Payload Mass
  Using get_mass helper function
- Orbit
- Customer
- Launch Outcome
- Landing Status
  Using landing_status helper function

Data Wrangling

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

- Load the CSV data for the flight data obtained via the SpaceX API
- Calculate the % of missing values in each field. From this we determine that 29% of the LandingPad data is missing
- Also determine which fields are numerical and categorical
- Calculate the number of launches per site
- Calculate the number of each orbit type
- Calculate the number of each landing type
- Calculate the number of each landing outcome
- Create a landing outcome label -- either 0 or 1 -- for failed and successful landings, respectively
- We determine from this that there is a 67% success rate
- Export data with new outcome field to CSV

EDA with Data Visualization

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/edadataviz.ipynb

- Visualize the relationship between PayloadMass and FlightNumber using a scatter plot
- Visualize the relationship between LaunchSite and FlightNumber using a scatter plot
- Visualize the relationship between LaunchSite and PayloadMass using a scatter plot
- Visualize the relationship between Success Rate of each orbit type using a bar plot.
- Visualize the relationship between FlightNumber and Orbit type using a scatter plot
- Visualize the relationship between Payload Mass and Orbit type using a scatter plot
- Visualize the launch success rate by year using a line plot
- Select the features (fields) to predict success rate
- Use one-hot encoding to assign dummy variables to each selected categorical field
- Cast all numeric fields to float
- Export to CSV (as part3)

EDA with SQL

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- View the first 30 records: `SELECT * FROM SPACEXTABLE LIMIT 30;`
- View the first 5 records where launch sites begin with 'CCA': `SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;`
- Display the total payload mass carried by boosters launched by NASA (CRS)
  `SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';`
- Display the average payload mass carried by booster version F9 v1.1
  `SELECT AVG(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';`
- List the date when the first successful landing outcome was achieved
  `SELECT MIN(Date) AS first_landing FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success%';`
- List the names of the boosters which have success in drop ship and have payload mass greater than 4000 but less than 6000
  `SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;`
- List the total number of successful and failure mission outcomes
  `SELECT Mission_Outcome, COUNT(*) AS outcome_count FROM SPACEXTABLE GROUP BY Mission_Outcome;`
- List all the booster versions that have carried the maximum payload mass (using a subquery)
  `SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);`
- List the failure landing outcomes in drone ship along with booster versions and launch site for each month in 2015
  `SELECT substr(Date,6,2) AS monthname, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE`
- `WHERE substr(Date,0,5) = '2015' AND Landing_Outcome = 'Failure (drone ship)';`
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, ranked in descending order
  `SELECT Landing_Outcome, COUNT(Landing_Outcome) AS outcome_count`
- `FROM SPACEXTABLE`
- `WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'`

Build an Interactive Map with Folium

Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/lab_jupyter_launch_site_location.ipynb

- A map was created with a marker for NASA's Johnson Space Center in Houston, for reference and centering the map
- Markers were added for each launch site: 1 on the California coast, and 3 along Florida's east coast, near Cape Kennedy
- Marker clusters were created for each launch site indicating success (green) or failure (red) when zoomed in on a given launch site
- Distances were calculated from a selected launch site (Vandenberg, in this case), to the coast, to the nearest railway and the nearest road using a distance calculator function.

Build a Dashboard with Plotly Dash

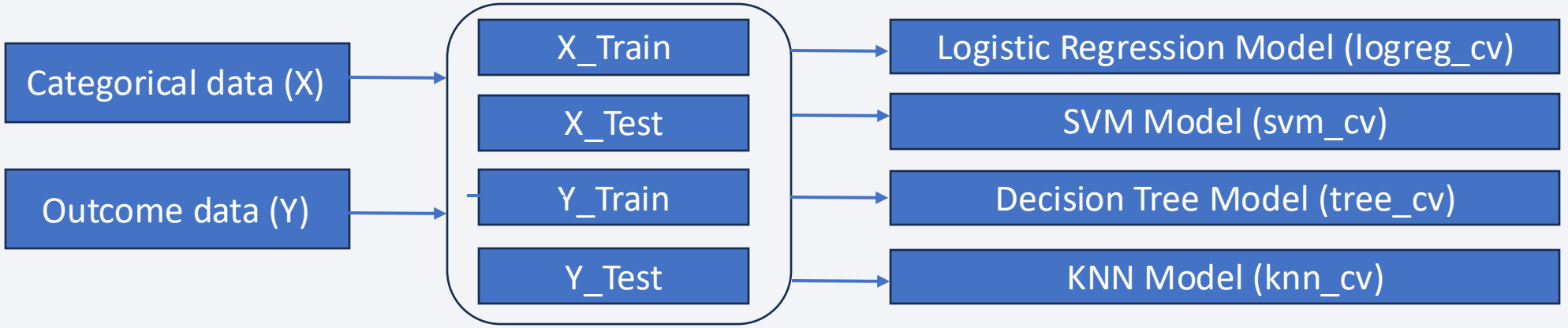
Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/spacex-dash-app.py

- The concept here is to build an interactive dashboard the allows all sites or a single site to be selected while also including a payload mass range selector in order to drill down further on the booster category vs. Payload mass.
- If 'All Sites' are selected, then a pie chart is displayed to display all successful launches by site by %.
- If a single site is selected, then a pie chart is displayed showing the distribution of success to failure (i.e, 2 pie slices).
- The payload mass range is used to interactively display a scatter plot below that shows the Payload Mass distribution by Booster Category for a given site or all sites

Predictive Analysis (Classification)

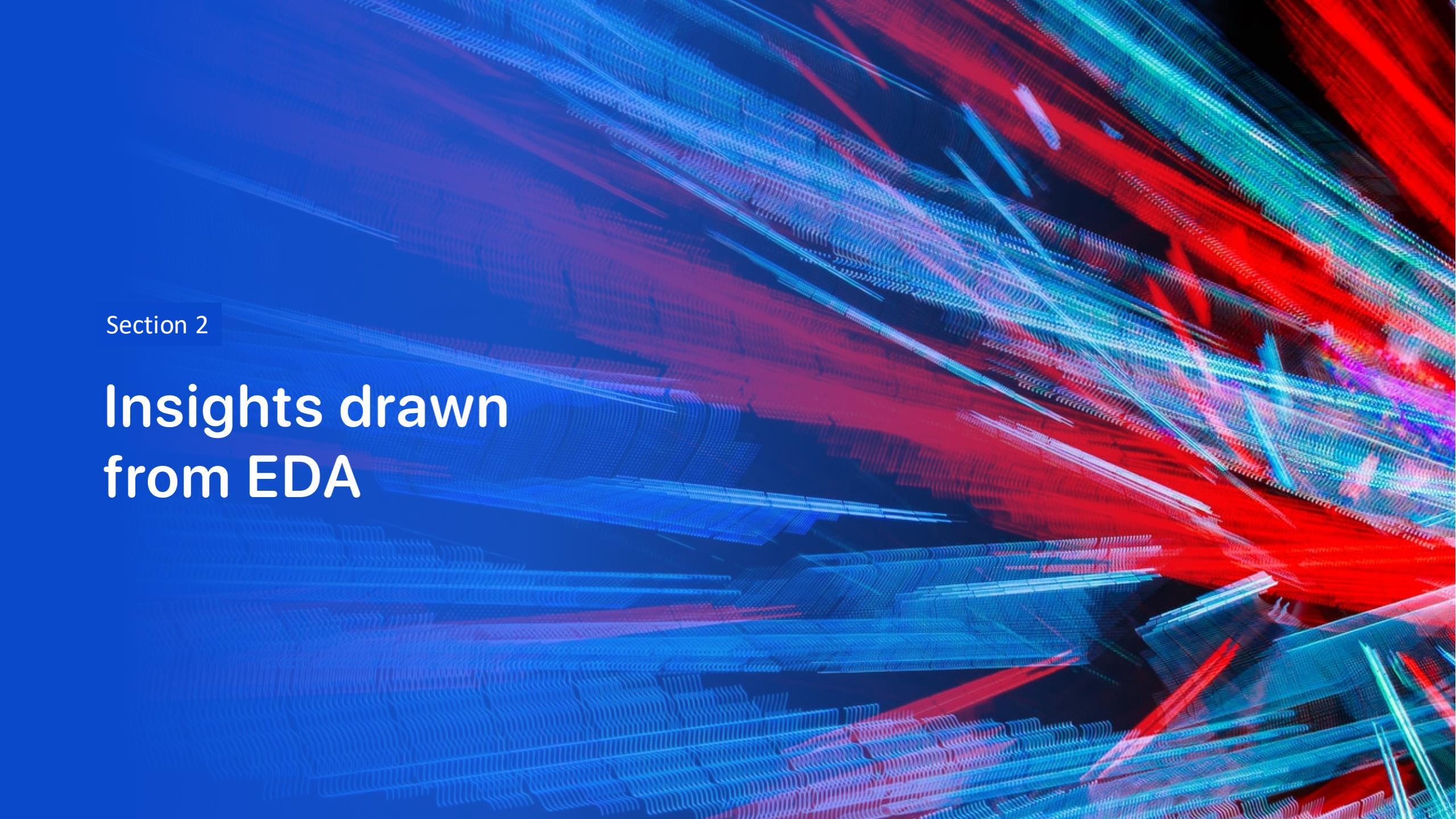
Completed Notebook URL: https://github.com/astralplanecrash/testrepo/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

- The landing outcome data (Class field) was loaded into a NumPy array (Y), while the features were loaded into a separate DataFrame from the 3rd CSV export with the categorical fields (X)
- The categorical data was then standardized using StandardScaler
- Used train_test_split to split X and Y into training and test data
- Implemented a GridSearch and plotted a confusion matrix for each of the 4 models
    1. Logistic Regression
    2. SVM
    3. Decision Tree
    4. K-nearest Neighbor (KNN)
- After evaluating the test accuracy values as well as the best score values, the best performing model was identified

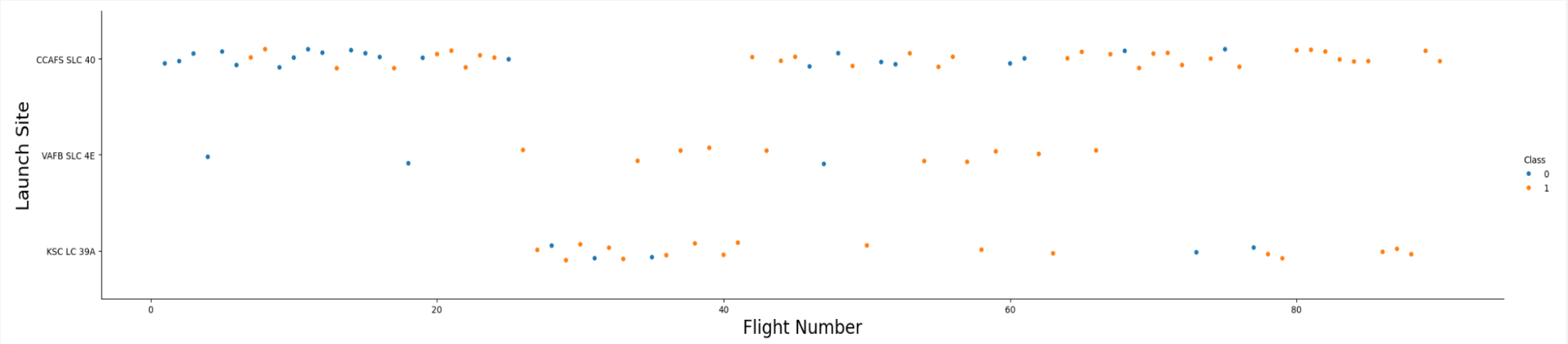


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

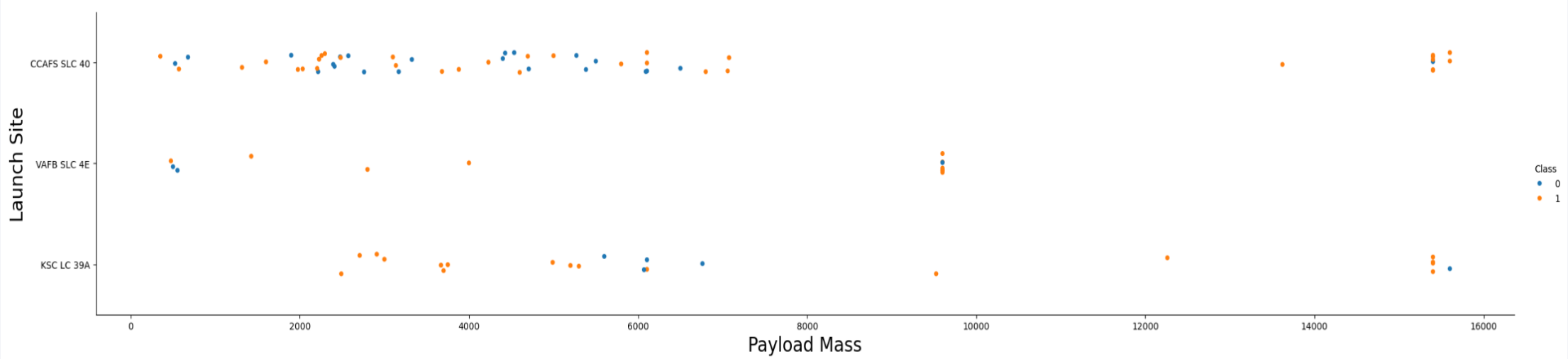

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



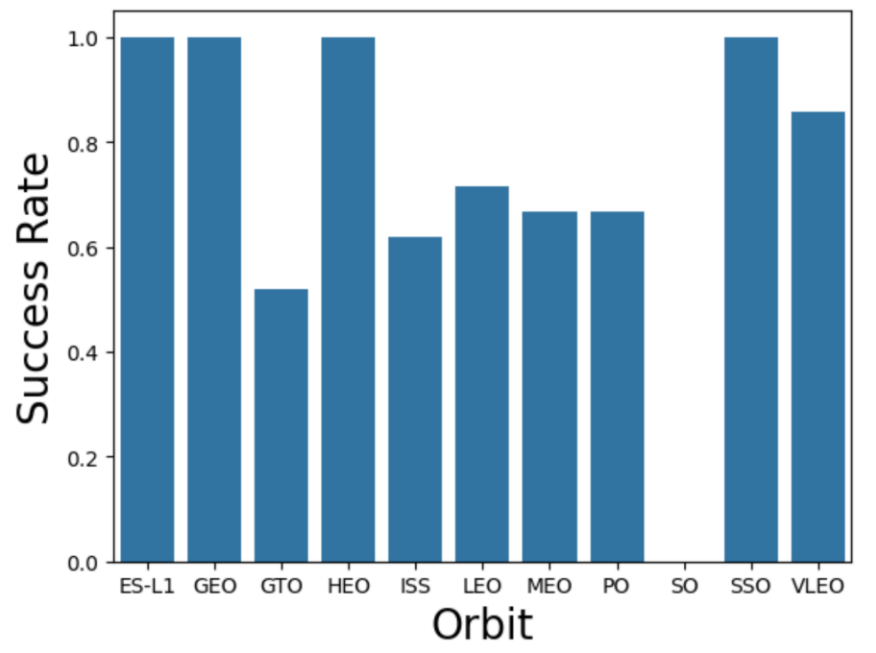
From this scatter plot, you can see that as the flight numbers increase, the success rate of each launch site improves, however CCAFS SLC 40 has a slightly lower success rate for flights numbers greater than 40.

Payload vs. Launch Site



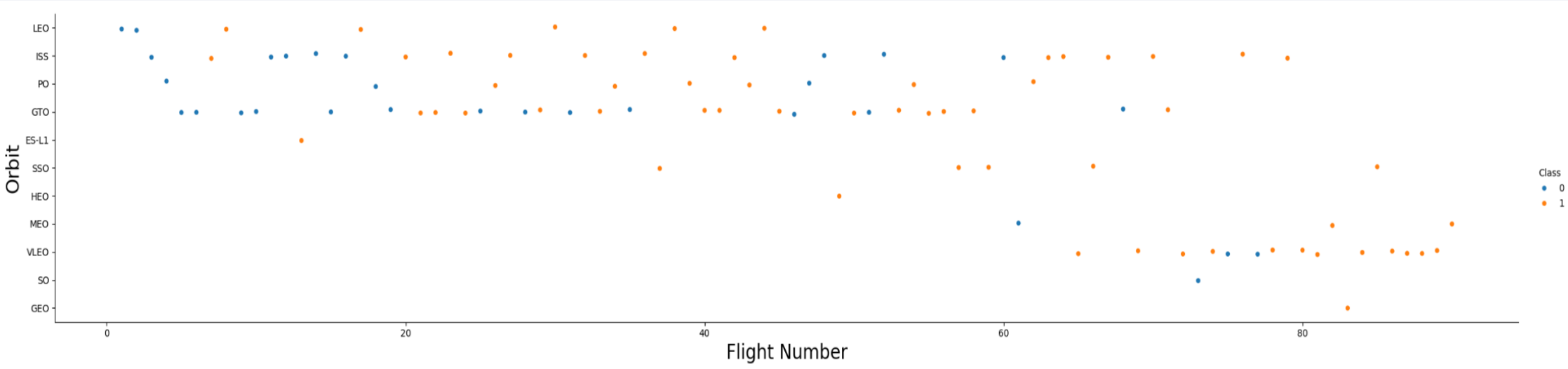
- From this scatter plot, you can see that as payload mass exceeds 8000kg, the success rate of each launch site is pretty high.
- Both VAFB SLC 4E and KSC LC 39A have relatively high success rates with payloads above about 1500kg.
- Site CCAFS SLC 40 has a pretty spotty success rate with payloads < 8000kg.

Success Rate vs. Orbit Type



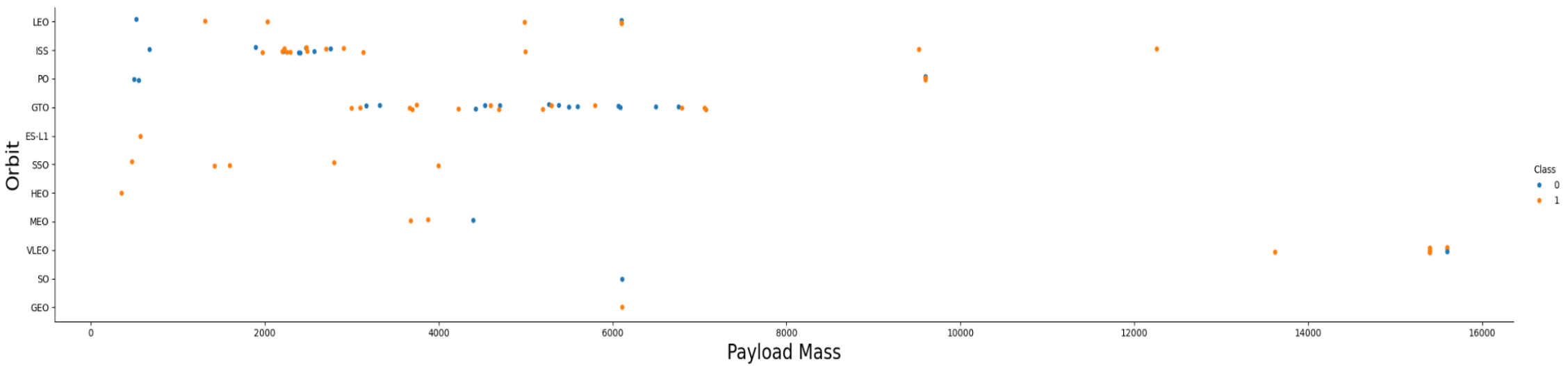
- From this bar plot, you can see that 4 orbit types – ES-L1, GEO, HEO and SSO – have 100% success rate, with VLEO also pretty high at above 80%.

Flight Number vs. Orbit Type



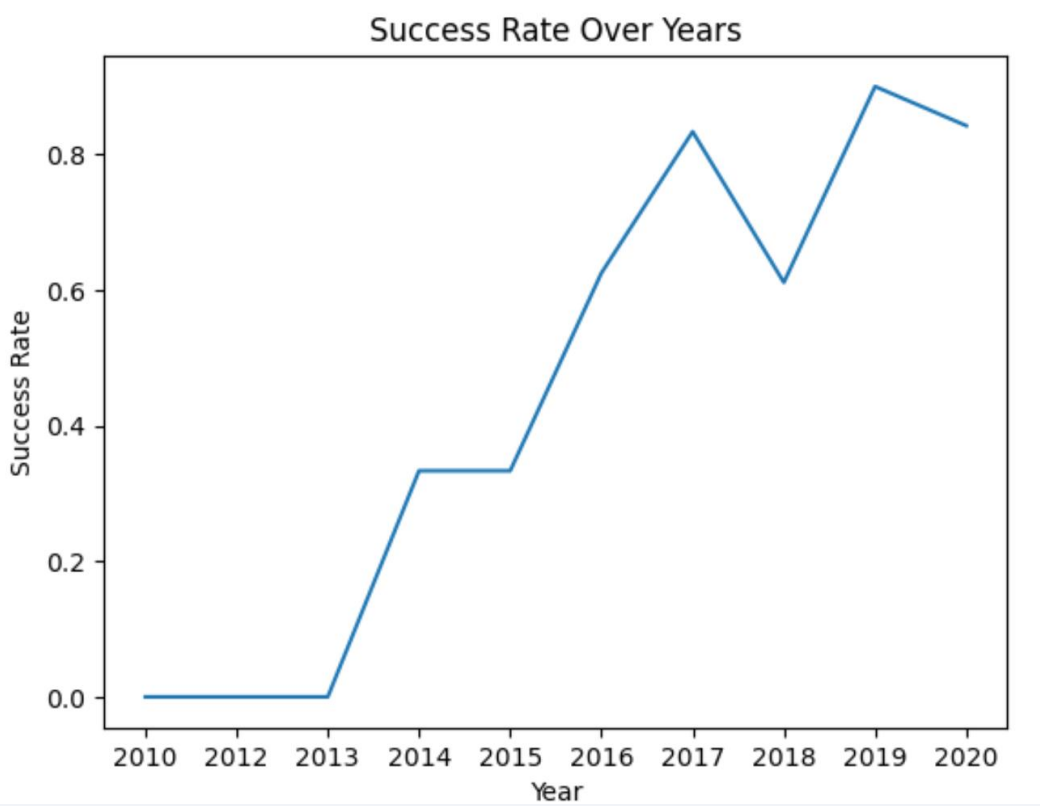
- From this scatter plot, you can see that, for orbit types that have a significant number of launches, success rate is increasing, for the most part.
- Skipping over the first 6 flights, LEO and SSO orbits have a 100% success rate
- VLEO has a pretty steady success rate, as does ISS and PO.

Payload vs. Orbit Type



- From this scatter plot, you can see that almost every launch with a payload above 8000kg has had successful landings
- SSO has a perfect track record with lighter payloads 4000kg or less
- GTO has a very spotty record, all with mid-size payloads

Launch Success Yearly Trend



- Very clearly, the success rate trends upwards starting in 2013
- There is a slight dip in success rates between 2017 and early 2018.
- There is also a slight dip from 2019 to 2020

All Launch Site Names

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

There are 4 unique launch sites

Launch Site Names Begin with 'CCA'

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

Here we are displaying the first 5 records

Total Payload Mass

total_payload
45596

The total payload carried by boosters launched by NASA (CRS) was 45,596kg.

Average Payload Mass by F9 v1.1

total_payload
2534.6666666666665

The total payload carried by booster type F9 v1.1 was 2,535kg.

First Successful Ground Landing Date

first_landing
2015-12-22

The first successful landing was on Dec. 22, 2015, after several years of failures.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Four booster versions match the criteria of drone ship landing where the payload was greater than 4000kg and less than 6000kg.

Total Number of Successful and Failure Mission Outcomes

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

Mission outcomes (a successful launch as opposed to a successful landing outcome) show a greater than 98% success rate, despite having 3 different success categories (unique values) in the database. The 2 'Success' result categories should really be the same, but there may be an errant trailing space on the 2nd one.

Boosters Carried Maximum Payload

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

12 booster types have carried the maximum payload

2015 Launch Records

monthname	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

2 drone ship landing failures occurred in January and April of 2015, involving the same launch site, CCAFS LC-40, and 2 different boosters

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

Landing_Outcome	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

The landing outcome counts (both success and failure) for each landing outcome

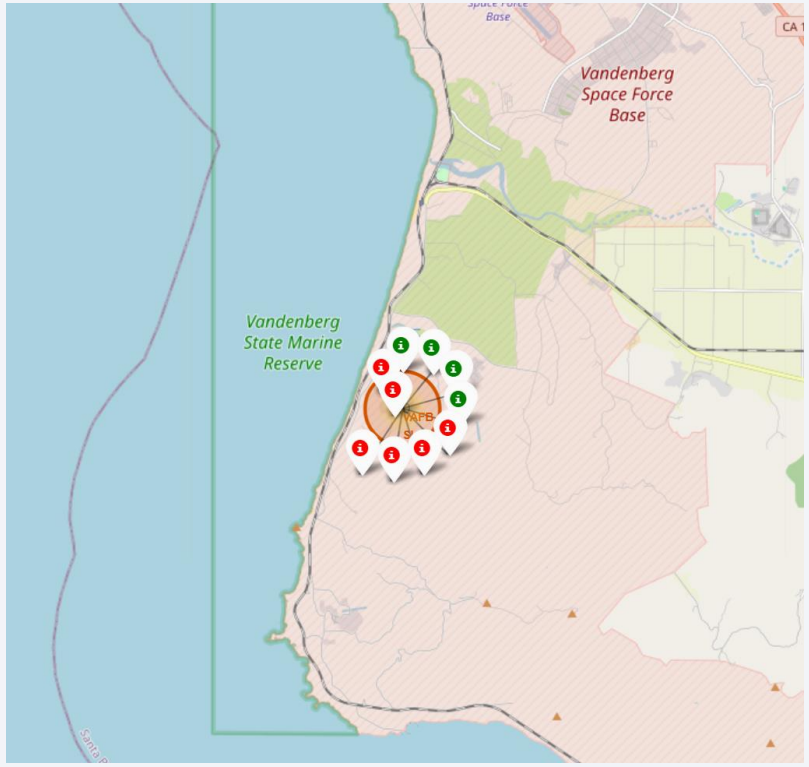
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

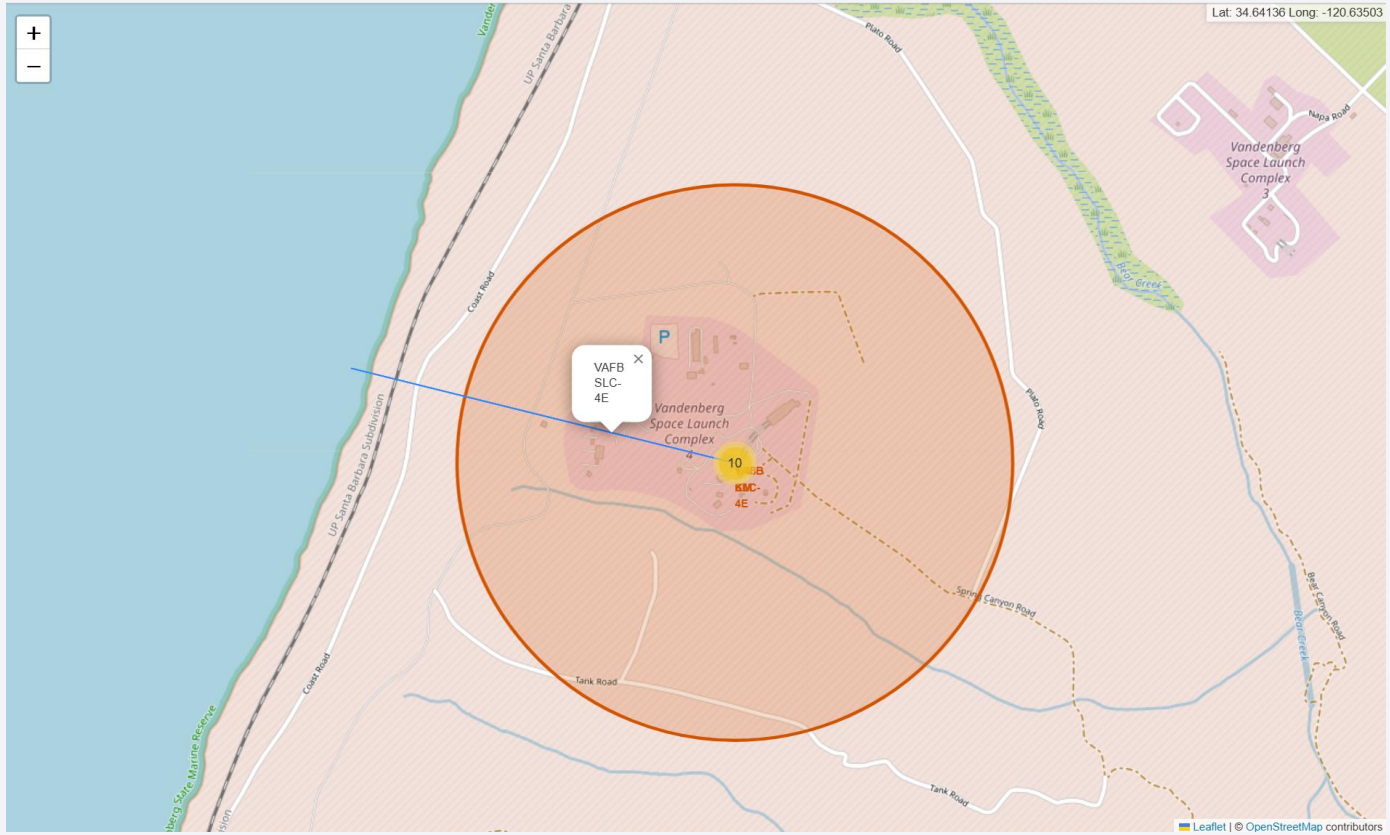
There are a total of 4 sites, 3 of which are in close proximity in Florida (see map on right)

Launch Outcomes for Vandenberg launch site VAFB SLC-4E

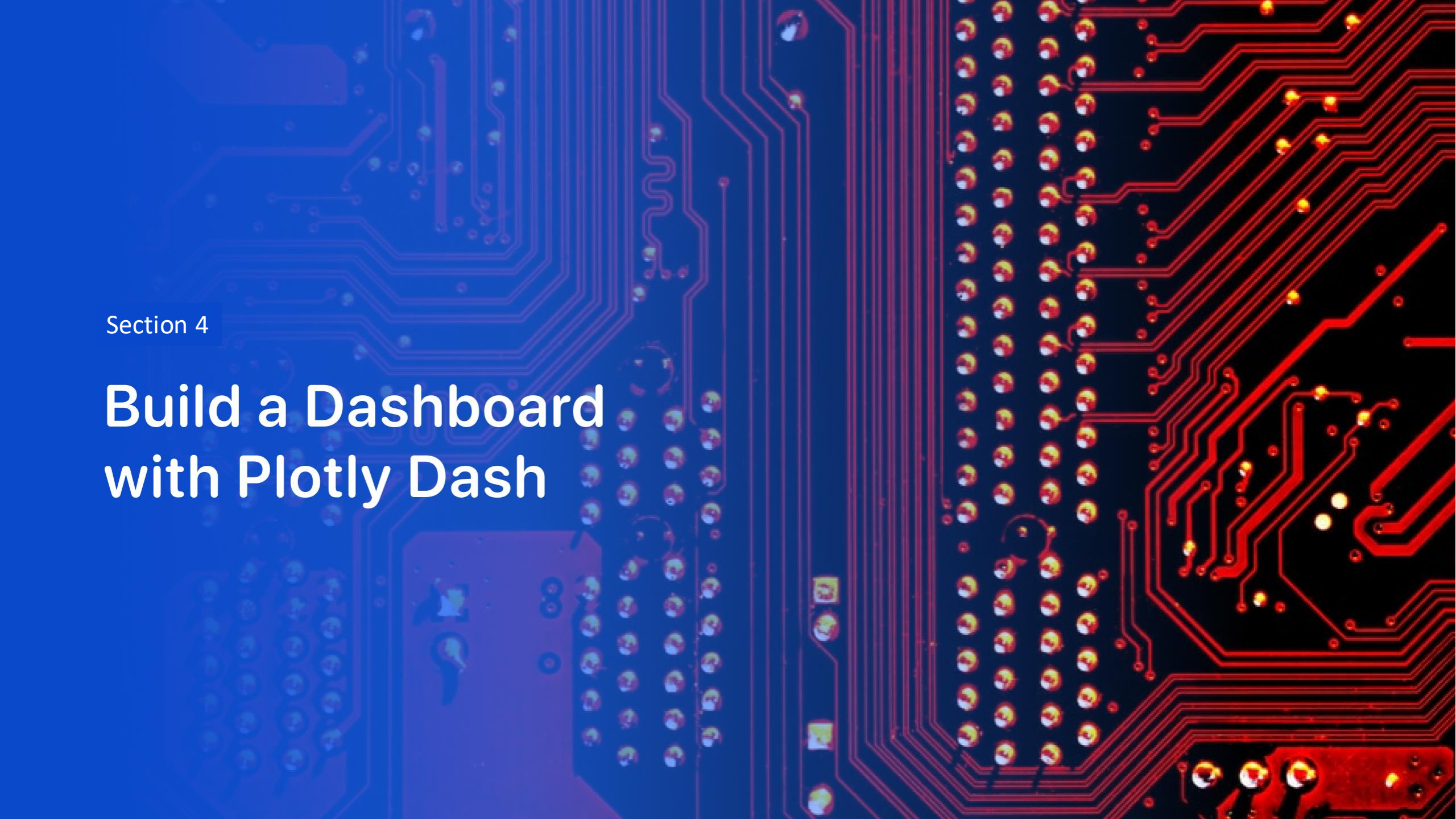


Here, we can see a total of 10 outcomes: 4 successful, 6 failed

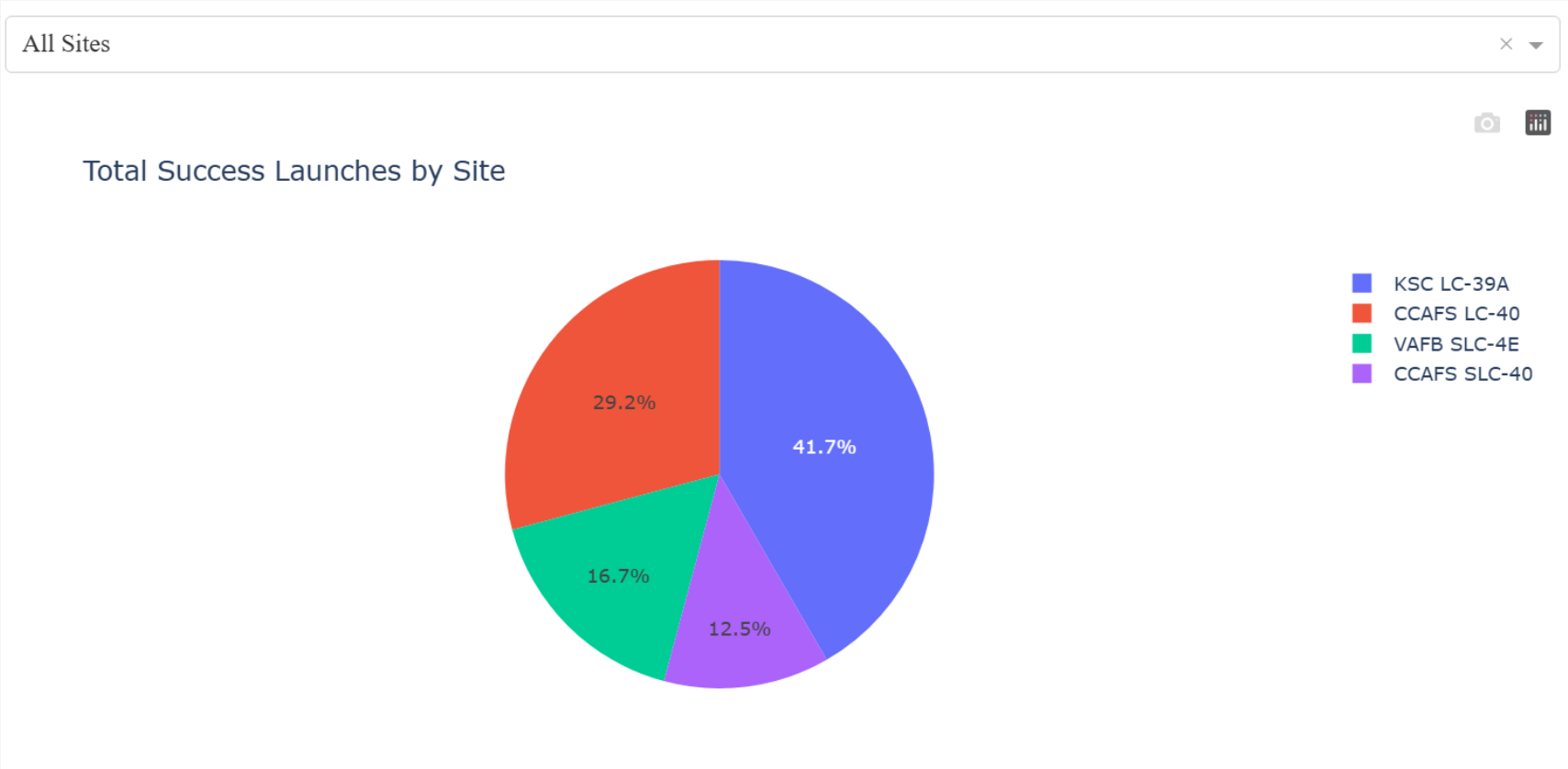
Line indicating distance to the coastline for launch site VAFB SLC-4E



The blue line from the launch site to the coast (Pacific Ocean) is 1.42 km, however the distance label did not display correctly. From this we can deduce that launch sites need to be near a large body of water in the event of catastrophic failure while the rocket is airborne. The launch trajectories direct them over the water, not land.

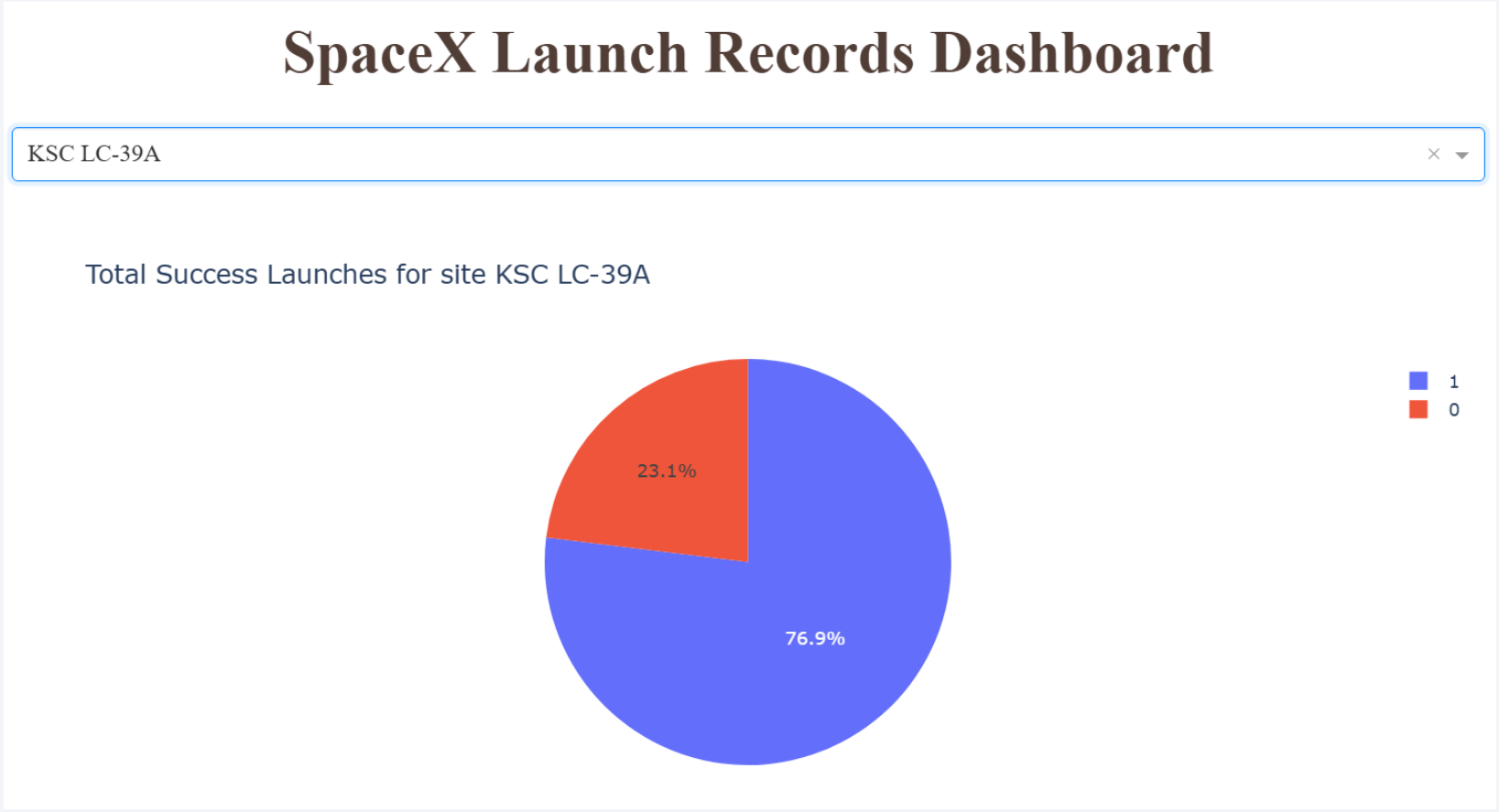

Section 4
Build a Dashboard with Plotly Dash

Launch Successes by Site



NOTE: While this doesn't necessarily tell us which sites have more successful launches, it does indicate the frequency of launches per site. For more data on successful launches, we need to drill down more.

Success Rate for Launch Site KSC LC-39A

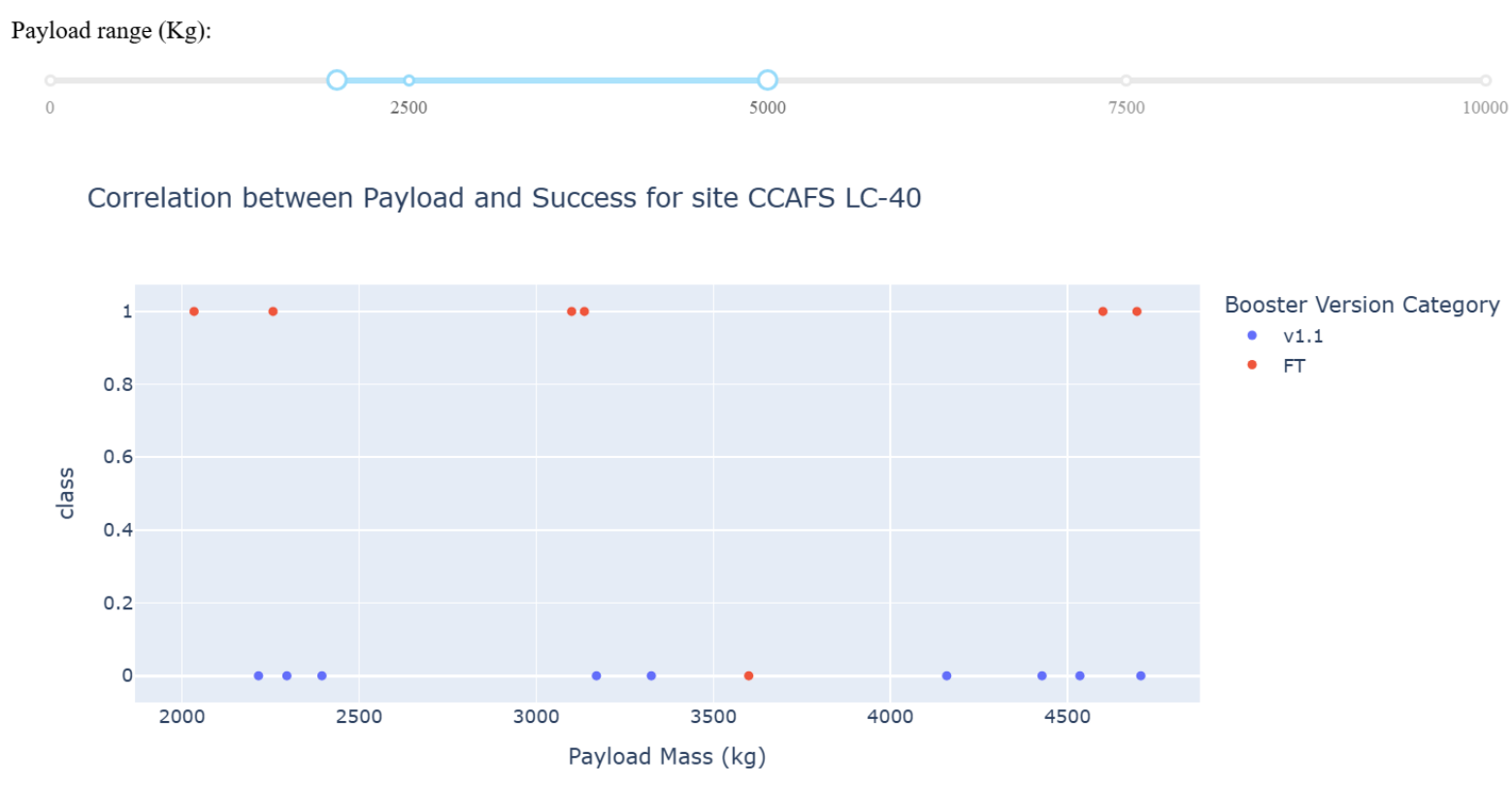


Site KSC LC-39A (Florida) had the highest successful launch rate of 76.9% of all the launch sites.

Correlation between Payload and Success for All Sites

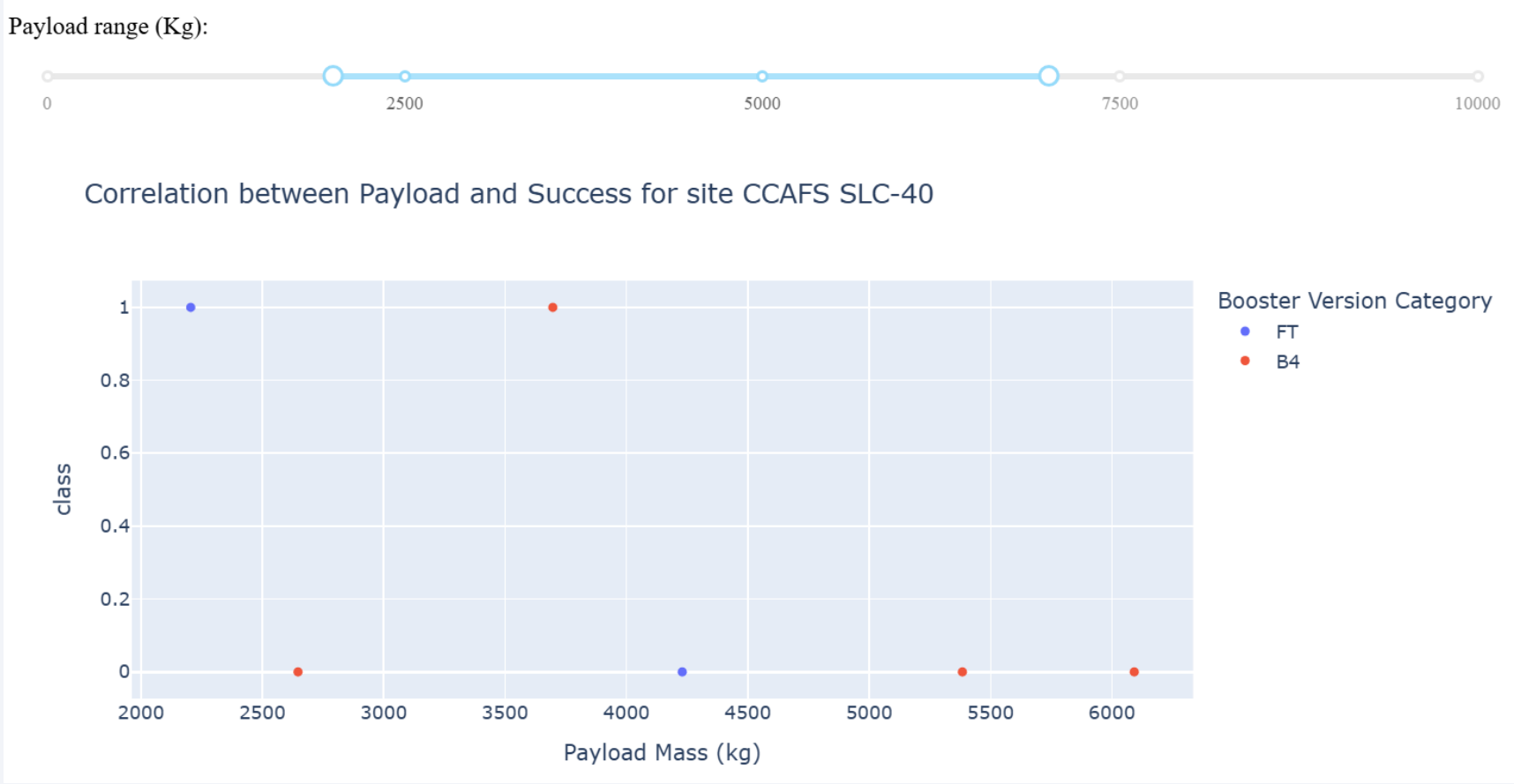
- Here, we can see that most of the successful launches occur with a payload between 2000kg and 6000kg, however there are a number of failures as well.
- The FT booster has the best success rate in the 2000-6000kg payload range
- The v1.1 booster has the poorest success rate overall

Correlation between Payload and Success for site CCAFS LC-40



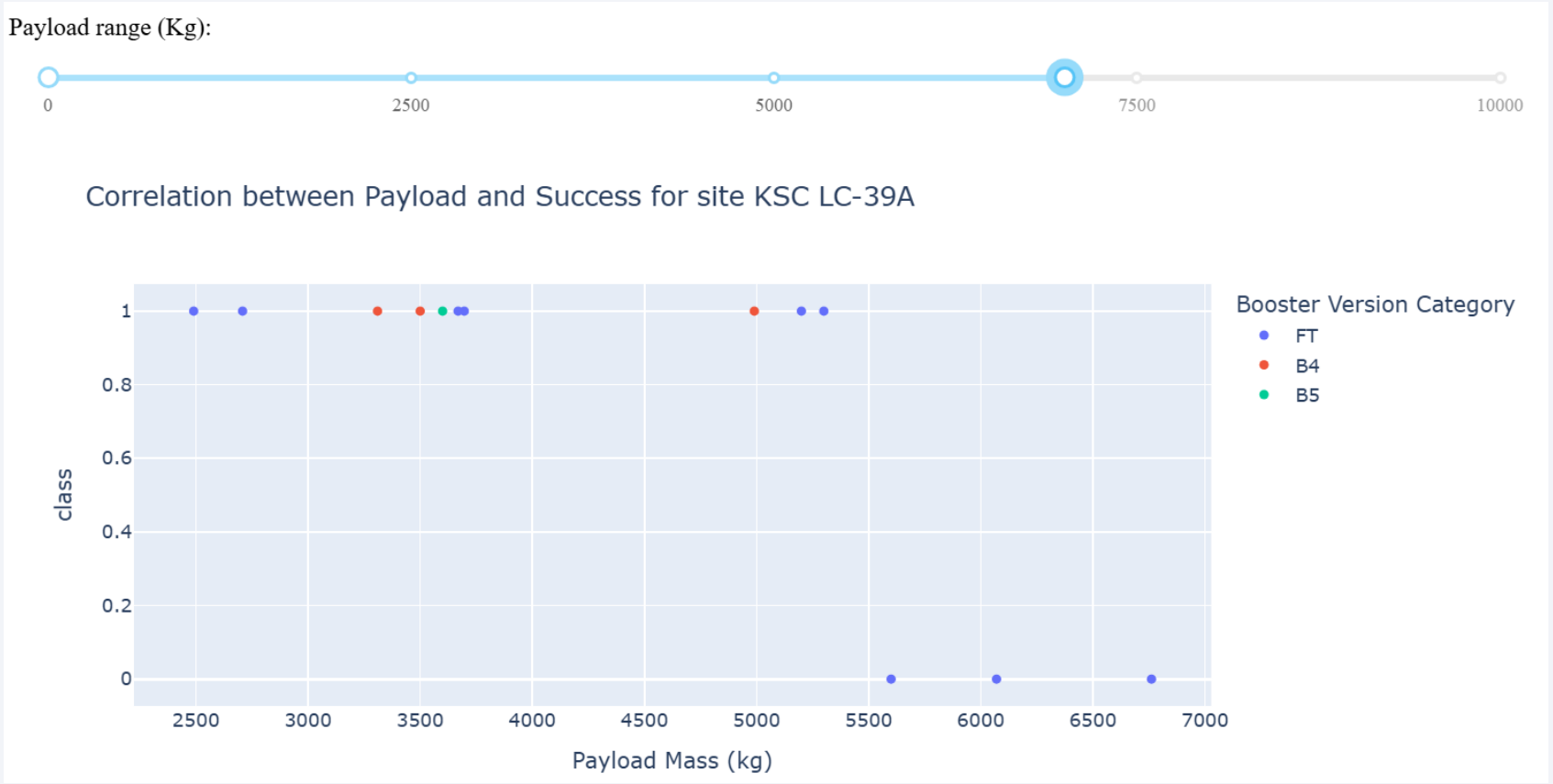
- Here, we have restricted the payload range to 2000kg through 5000kg.
- The FT booster has a high success rate, while the v1.1 has a dismal success rate (0%)

Correlation between Payload and Success for site CCAFS SLC-40



- Here, we have restricted the payload range to 2000kg through 7000kg.
- Neither the FT or B4 booster fared very well in with successful launched, but the data suggests that heavier payloads (above 4000kg) increase the likelihood of failure

Correlation between Payload and Success for site KSC LC-39A



- Here, we restricted ("zoomed in") the payload to 7000kg, since there were no larger payloads.
- You can clearly see that the success rate is 100% for all 3 boosters when the payload is under 5500kg.
- The FT booster failed on all 3 launches with payloads above 5500kg

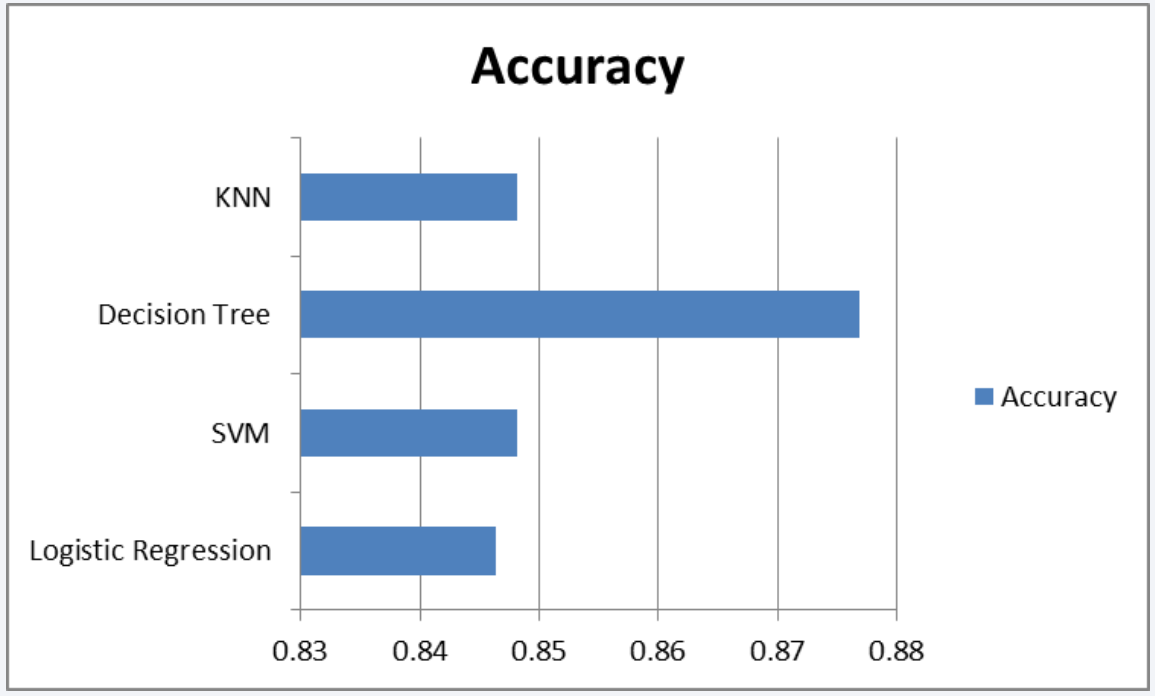
Correlation between Payload and Success for site VAFB SLC-4E

- Here, the payload range is restricted to the high end (5000kg and up)..
- This launch site does not have nearly as many launches as the previous 3.
- We can see that only the B4 booster was used for these heavier payloads, but the success rate is poor.

Section 5

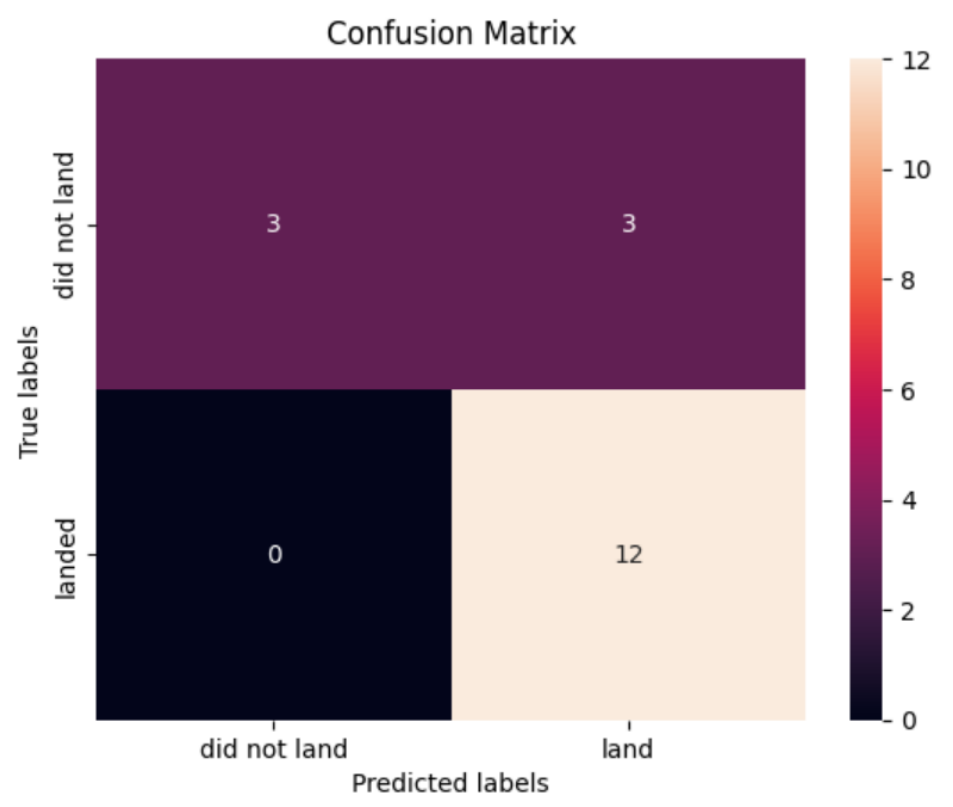
Predictive Analysis (Classification)

Classification Accuracy



- Since the confusion matrices of all 4 models were identical, the only distinguishing factor is the accuracy computed from the models' best_score_ value, shown above.
- From this, we can conclude that the Decision Tree model performed slightly better

Confusion Matrix



As it turned out, all 4 models that were tested resulted in identical confusion matrices where there were 12 true positives and 3 true negatives. The models failed only with 3 false positives – where the rocket did not land but the prediction indicated it would.

The accuracy score of all 4 models was 83.3%.

Conclusions

- Judging from the identical confusion matrices of all 4 models used, we cannot readily favor one model over any of the others
- Only one model had a slightly higher "best score" value of 87.7% and that was the Decision Tree.
- The identical results aren't surprising given that the data set consisted of only 90 launches, of which 80% (72) were used for training and 20% (18) were used for testing.
- As more launches occur over the next few years, the accuracy should only increase.
- While there could potentially be more features we could incorporate (e.g., weather conditions, etc.) to make the predictions more accurate, the current predictions are leaning strong but ultimately not particularly reliable.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!