IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael Saulon B
6/5/2023

Outline

- Executive Summary
- Introduction
- Methodology / Outline
- EDA and Observations
- Results
- Conclusion
- Appendix

Executive Summary

- **Goal**:
  - Determine the factors that best determine if the first stage of a launch will be successful

- **Methodologies used**
  - Data used for prediction and analysis is collected through the SpaceX API and some web-scraping.
  - Data will be filtered such that only records of Falcon 9 launches are shown.
  - Exploratory Data Analysis was done through a combination of matplotlib/seaborn, Folium, and Plotly Dash. Some SQL was also done to perform analysis on various queries. Flight number, payload mass, orbit type, and launch site, among other factors will be analyzed in relation to each other.
  - Choosing from a logistic regression, a support vector machine, a decision tree, and a k-nearest neighbors model, find the algorithm that predicts a successful launch the most efficiently.

- **Result summary**
  - Certain orbits and ranges for payload mass have a better chance for success
  - Launch site KSC LC-39A has the most successful launch rate out of any launch site.
  - A Decision Tree algorithm is the best model in predicting launch success overall.

Introduction – Our Goal

- One key aspect that makes SpaceX stand out in the space travel market is the **relatively low cost of their rocket launches**. While **other providers** charge upwards of **165 million dollars** per launch, **SpaceX** advertises Falcon 9 rocket launches on their website for only **62 million dollars** which is contributed by the company's ability to reuse the first stage of their rockets.
-
- The main objective:
  - ***Determine the factors that best determine if the first stage of a launch will be successful***
    - By efficiently predicting a successful launch, we can better determine the cost of the launch.

Section 1

Methodology

Methodology Outline

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

- Data Collection using the SpaceX API

Requested rocket launch data using the SpaceX API to receive info on each launch, which was then formatted into JSON which was then normalized to a DataFrame.

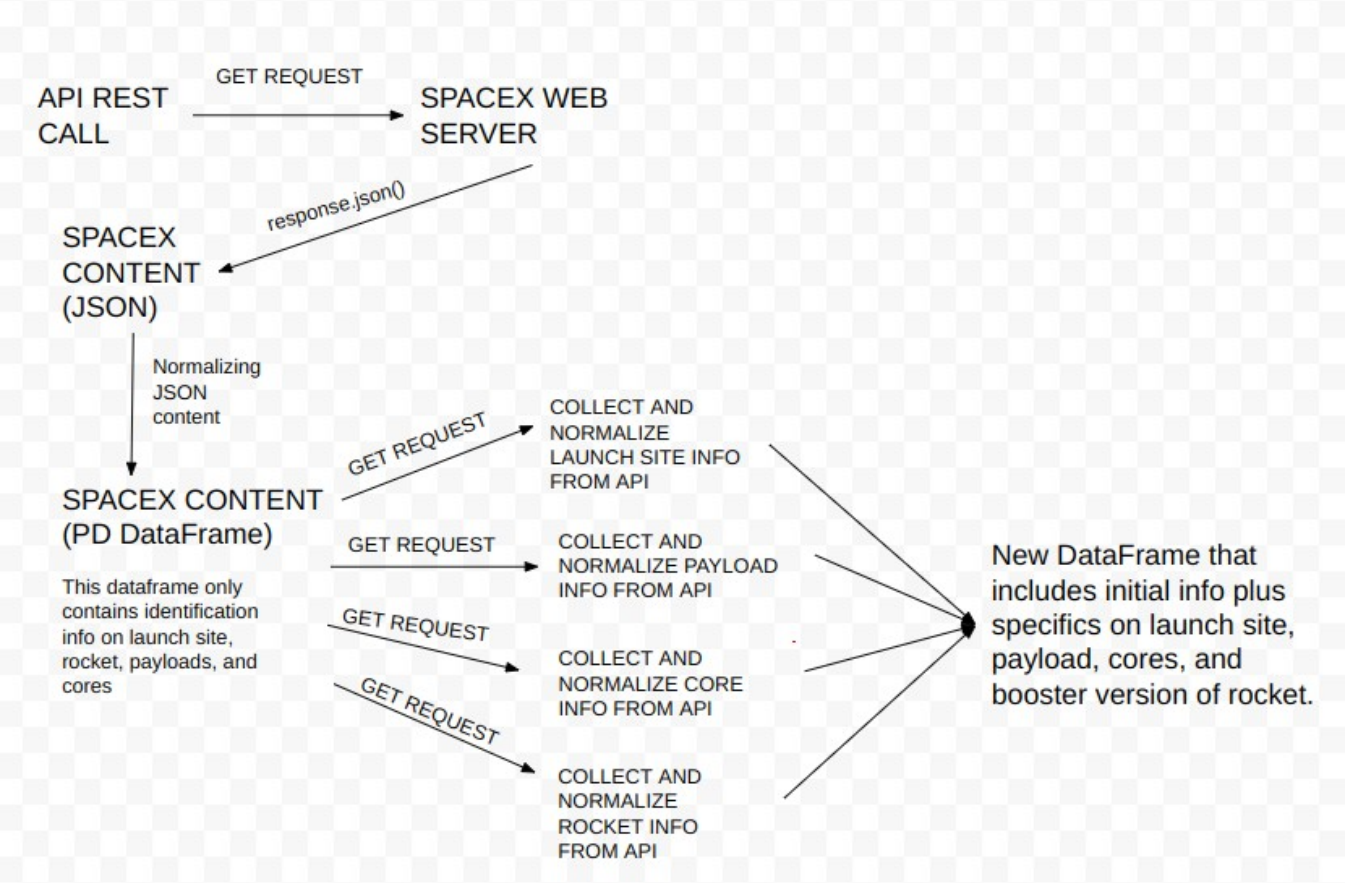
Info received on rockets, payloads, cores, and launch site only had their respective ID so additional GET requests were made to the API for each category.

Afterwards, a new DataFrame was made that better specifies info on the launch's site (latitude/longitude/site name), payload (name/mass), rocket (booster version), and cores.

- Data Collection from web scraping:

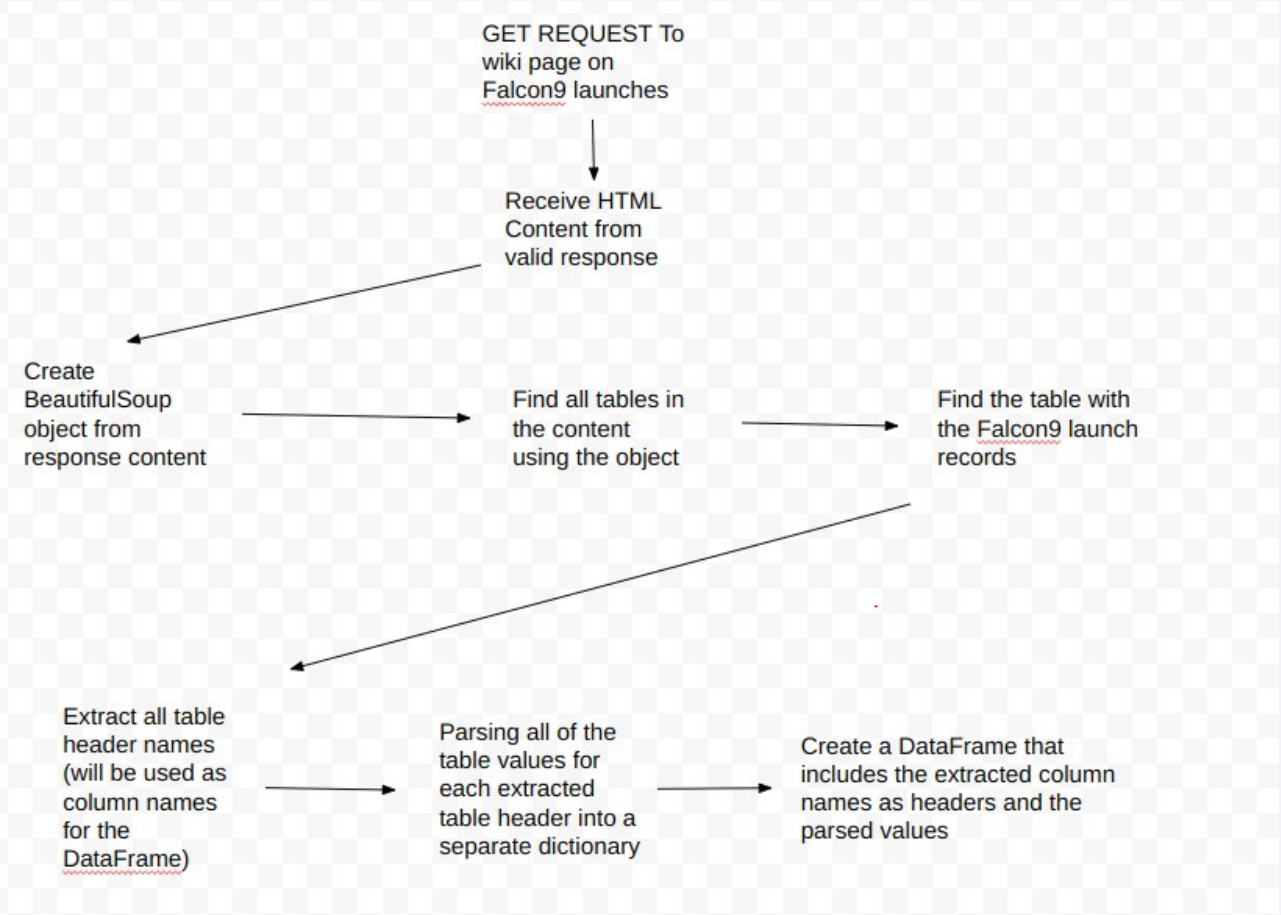
Several Falcon 9 launch records were fetched and scrapped from a Wikipedia page using BeautifulSoup that includes a table of past launches for Falcon 9 and Falcon Heavy (see references for wiki page).

Data Collection Flowchart – SpaceX API



https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_data_collection_api_1.ipynb

Data Collection Flowchart - Scraping



https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_data_webscraping_2.ipynb

Data Wrangling

Describe how data were processed

1) Collect spreadsheet of data that was collected using the SpaceX API
2) Filtered launch data to only include Falcon 9 launches.
3) Payload masses of some launches were missing which were dealt with by substituting those missing values with the mean payload mass.

https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_data_wrangling_3.ipynb

EDA with Data Visualization

The following charts were plotted for this analysis:

- **Relationship between Flight Number and Launch Site**
  - To determine if a launch site has a higher tendency to succeed or fail over time
- **Relationship between payload mass and launch site**
  - To check if a launch site favors a certain range of mass for their launches and to see if it affects the outcome of the launch
- **Success Rate by Orbit**
  - To see if certain orbits are more favorable to succeed.
- **Relationship between Flight Number and Orbit type**
  - To visualize if launches began favoring certain orbits in later flights and if those choices were favorable for the launch
- **Relationship between Payload and Orbit type**
  - To understand if various orbits favored lighter or heavier masses for their launches and to see how favorable the choices are for launch.
- **Launch Success Yearly Trend Line**
  - To get a grasp on how favorable the launches are over the years

https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_eda_matplotlib_5.ipynb

EDA with SQL

The following SQL queries were performed in this analysis:

- Find all distinct launch site names
- 5 records where launch sites began with ‘CCA’
- Find the total payload mass carried by boosters launched by NASA
- Finding the average payload mass carried by booster F9 v1.1
- Finding the date where the first successful landing outcome in a ground pad was achieved
- Finding the names of boosters which had success in drone ships and have payload masses between 4000kg and 6000kg
- Listing the total number of possible successful and failing mission outcomes
- Listing the names of booster versions that held the maximum payload mass
- Listing the records which displays month, failure in drone ship landings, booster versions, and launch sites in 2015
- Ranking the count of landing outcomes between April 6, 2010 and March 20, 2017.

https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_eda_sql_4.ipynb

Build an Interactive Map with Folium

- Markers and circles were added to represent each launch site and to mark down their locations
- Added a cluster of markers for each launch site that contains all of their successful and failed launches (color-coded) to give a better perspective on each site’s launch history
- Added lines between each launch site and their nearest coastlines, and railways to explore patterns or relationships between proximity of certain landmarks and launch success.

https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_folium_map_6.ipynb

Build a Dashboard with Plotly Dash

The following graphs were added to the dashboard for this analysis:

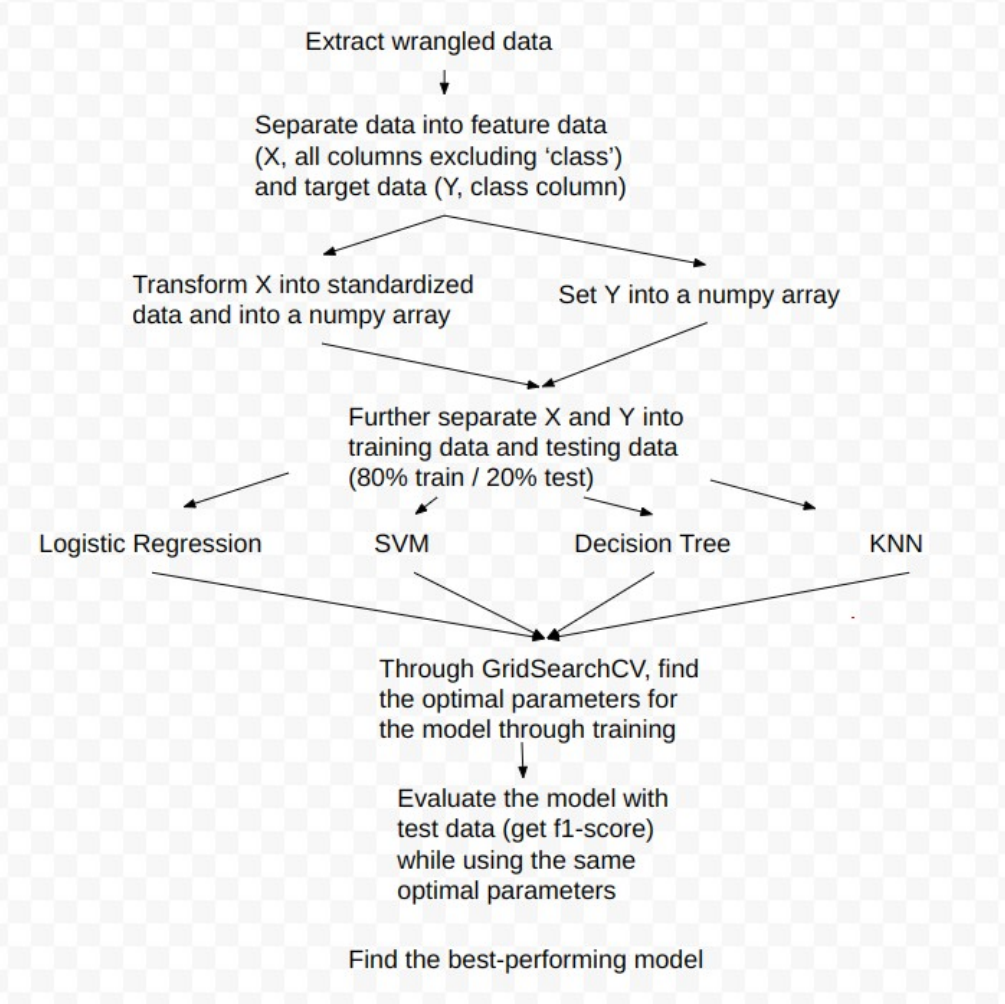
- Pie chart showing percentage of successful launches between all launch sites
  - Visualize relative success between launch sites

- Pie chart showing percentage of successful launches within each launch site
  - Visualize individual success within a launch site.

- Scatter-plot of correlation between Payload Mass and Successful Launches
  - To better understand the ranges of mass that leads to more and less successful outcomes.
  - To determine how booster versions fare in various ranges of mass.

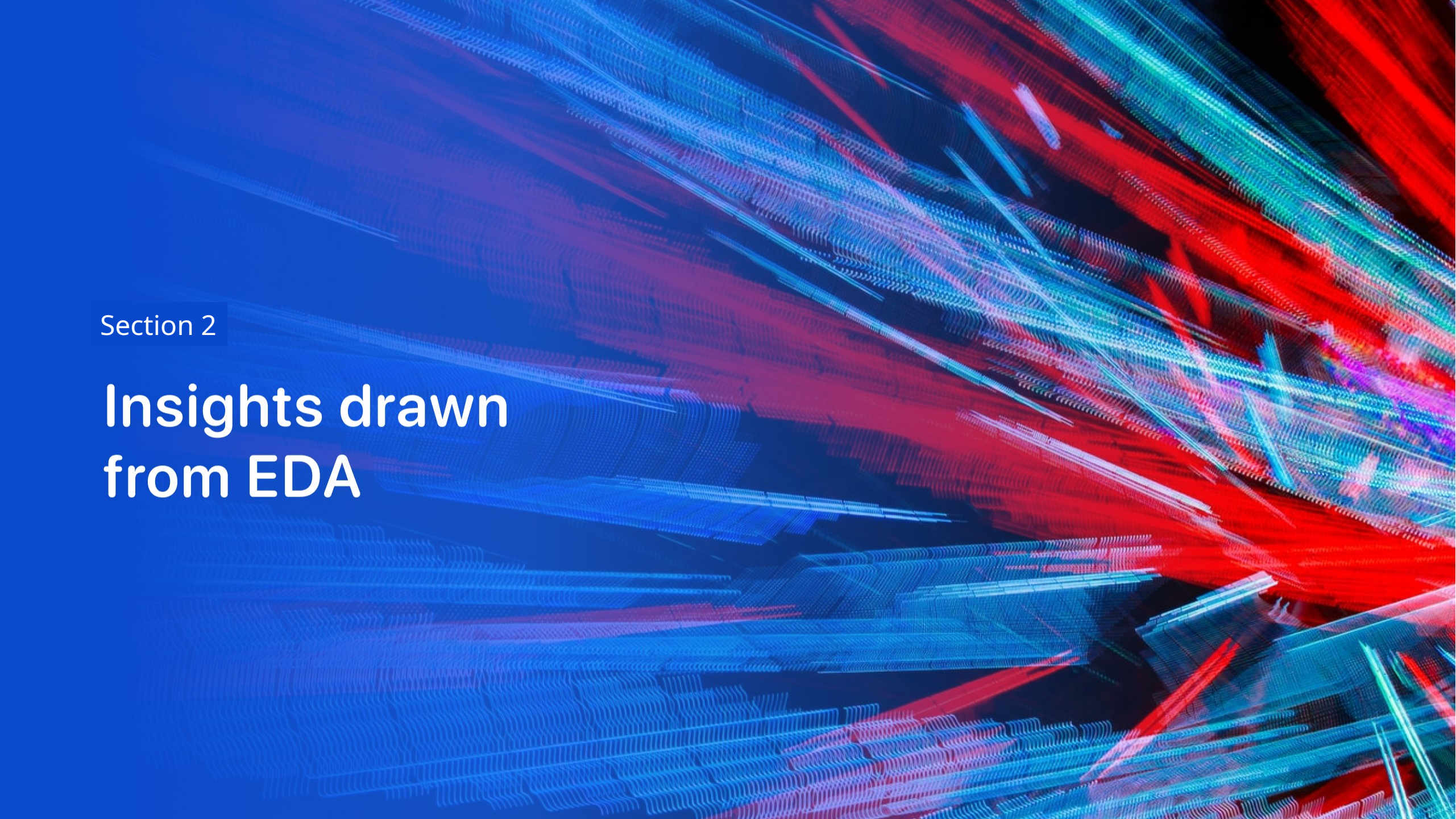
https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_dash_app.py

Predictive Analysis Flowchart (Classification)

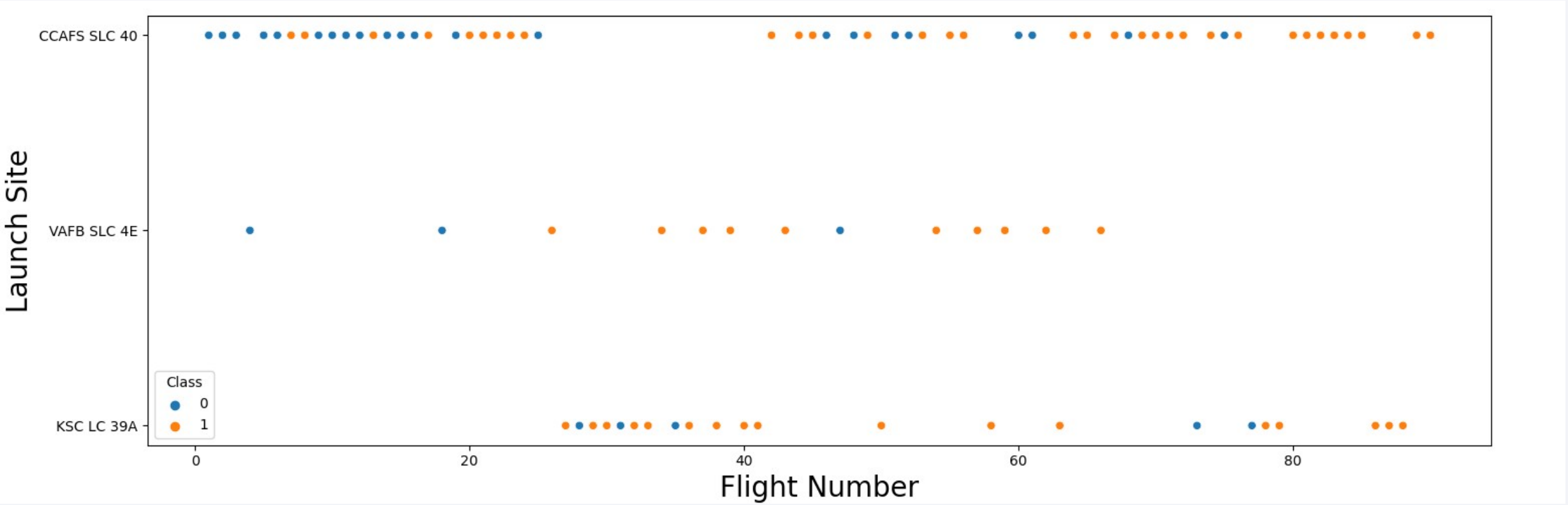


- Four classification models were developed to find the best at predicting launches: LR, SVM, DT, and KNN.
- F1 Score will be the primary metric that will be used to compare model performance. If scores are tied, test accuracy and training accuracy will be used as tiebreakers.
- Training data and test data are constant through all models.

https://github.com/MSB46/DataProjects/blob/main/IBM_Capstone/spacex_data_prediction_8.ipynb

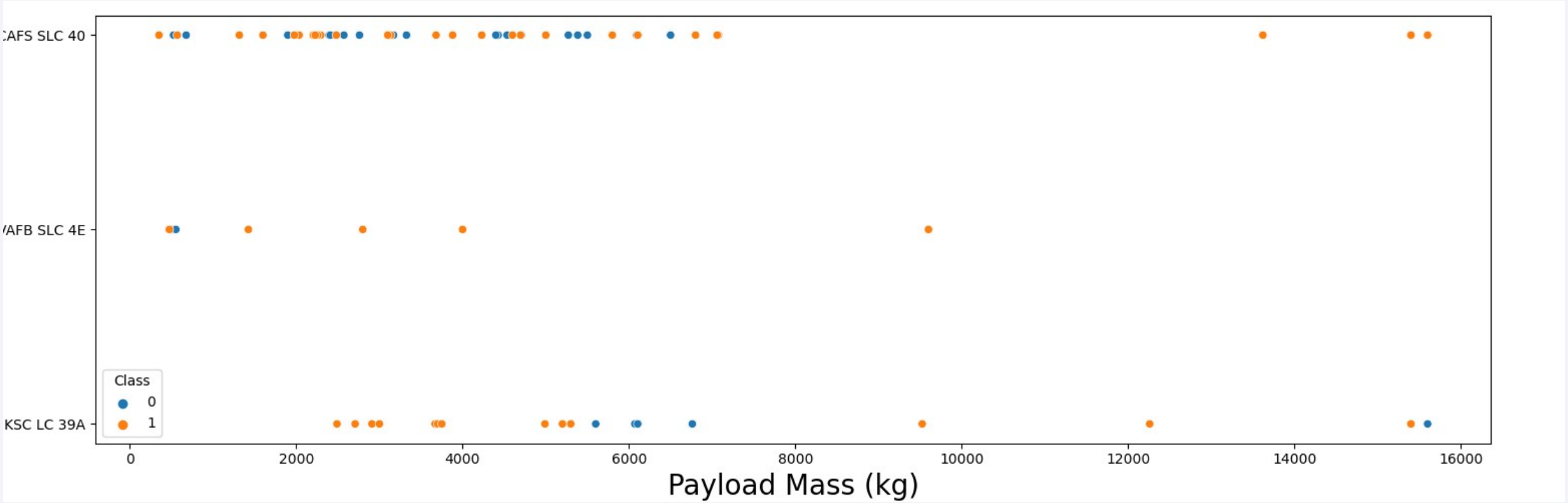

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



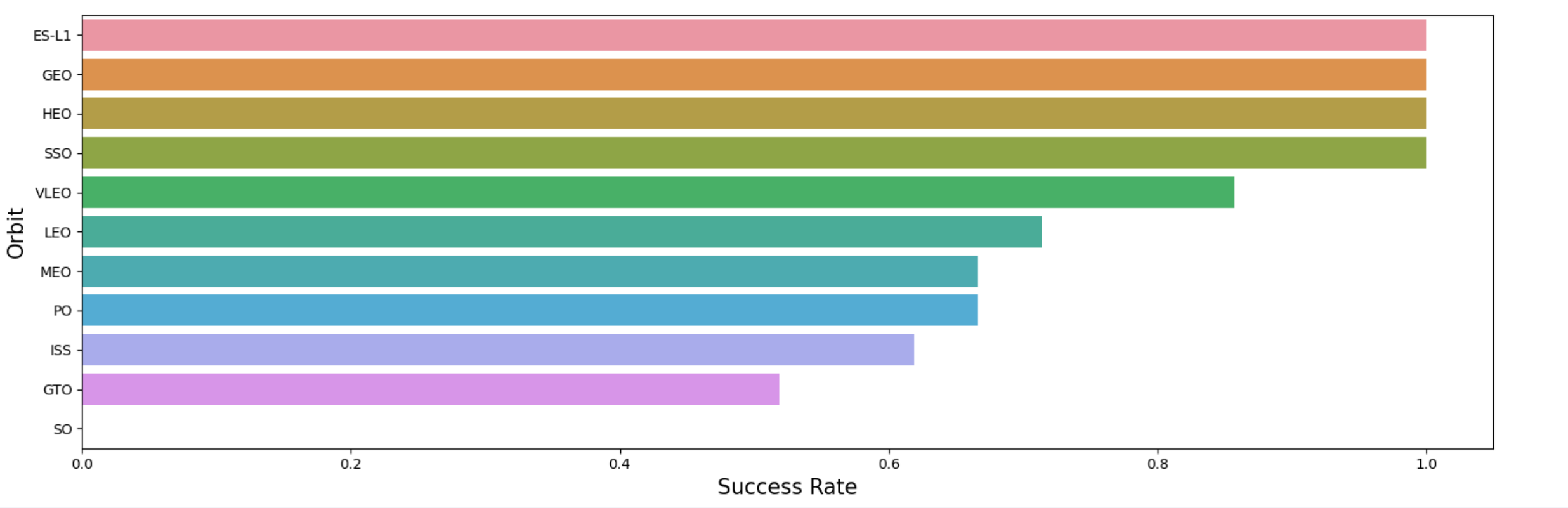
- For each launch site, successful launches have been happening more frequently in later flights
- 100% launches after flight 80 are successful
- The earlier the flight, the higher chance for failure and vice versa.

Payload vs. Launch Site



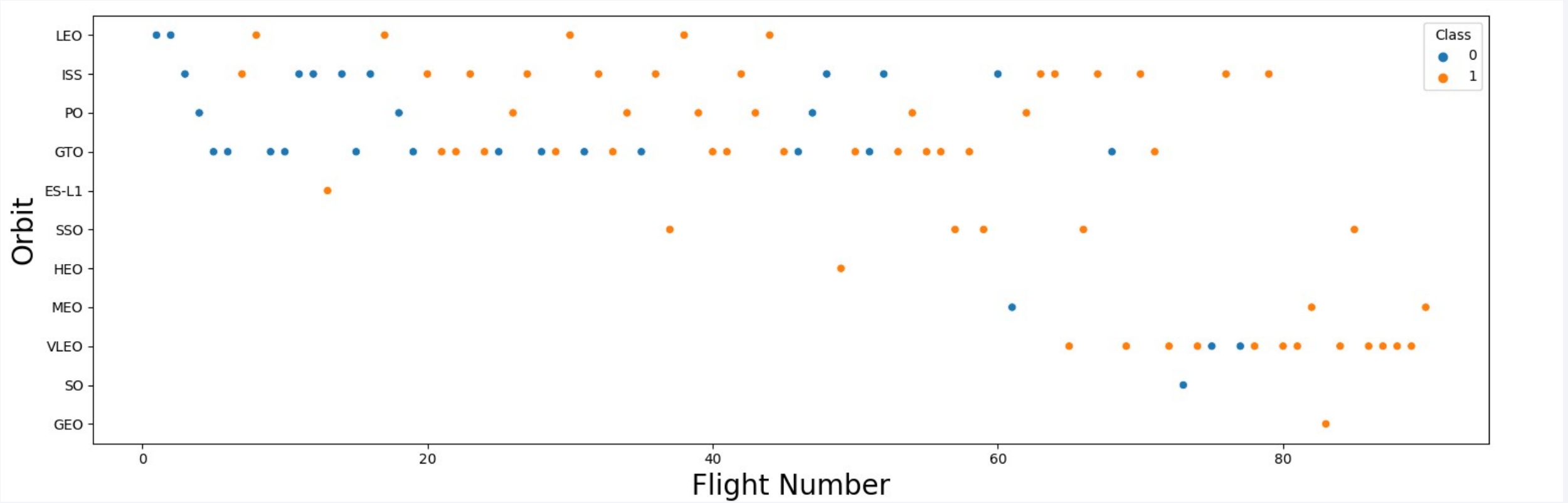
- Site VAFB SLC 4E has shown a lot of success despite using typically low masses (approx. 500-4000 kg) for their launches.
- Site KSC LC 39A is extremely successful in the 2000-5200 kg range
  - Between 5500 and 7000kg has the opposite occuring however.
- For site CAFS SLC 40, Not much of a pattern going on between 0-8000kg. However, the few masses past 8000kg have been pretty successful.
- Certain ranges of payload mass have a higher success rate than others throughout all launch sites
  - (primarily 2000 – 5200 kg and at least 10000 kg)

Success Rate vs. Orbit Type



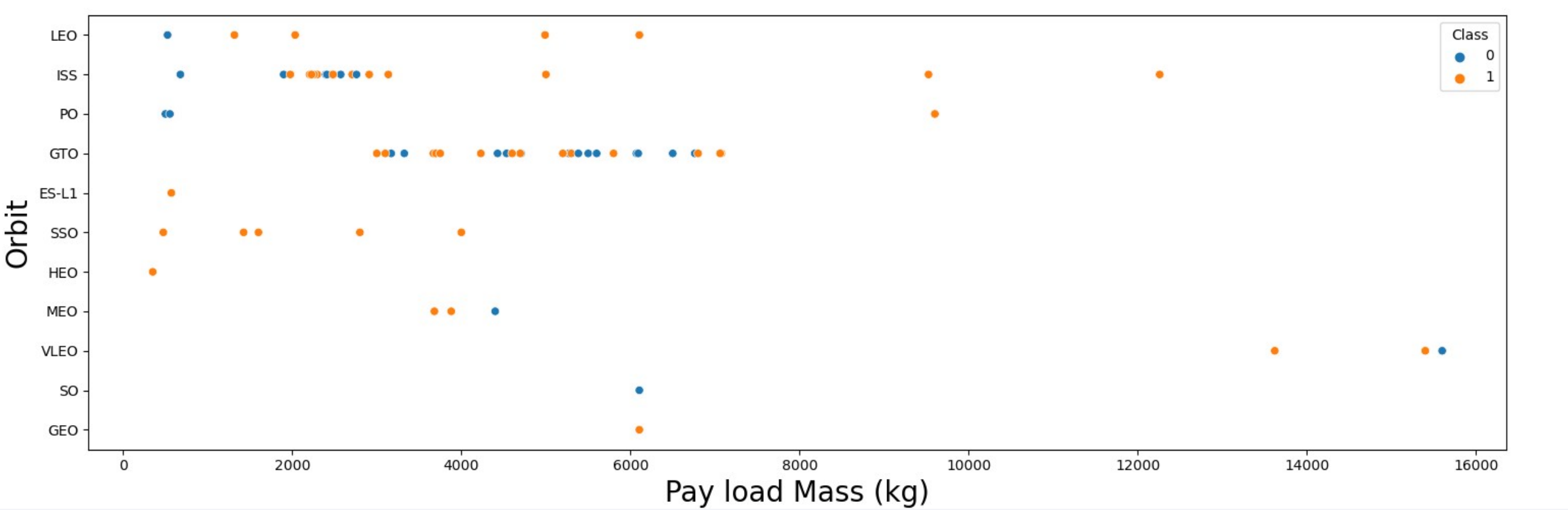
- However, at least 3 orbits managed to receive a 100% success rate: ES-L1, GEO, GEO, and SSO
  - SO (which is seen here as a separate orbit despite being the same as SSO) is the only orbit type with a 0% success rate which means SSO might have a lower success rate than the other three 100% orbits.
- VLEO is the second highest at a little over 83%.

Flight Number vs. Orbit Type



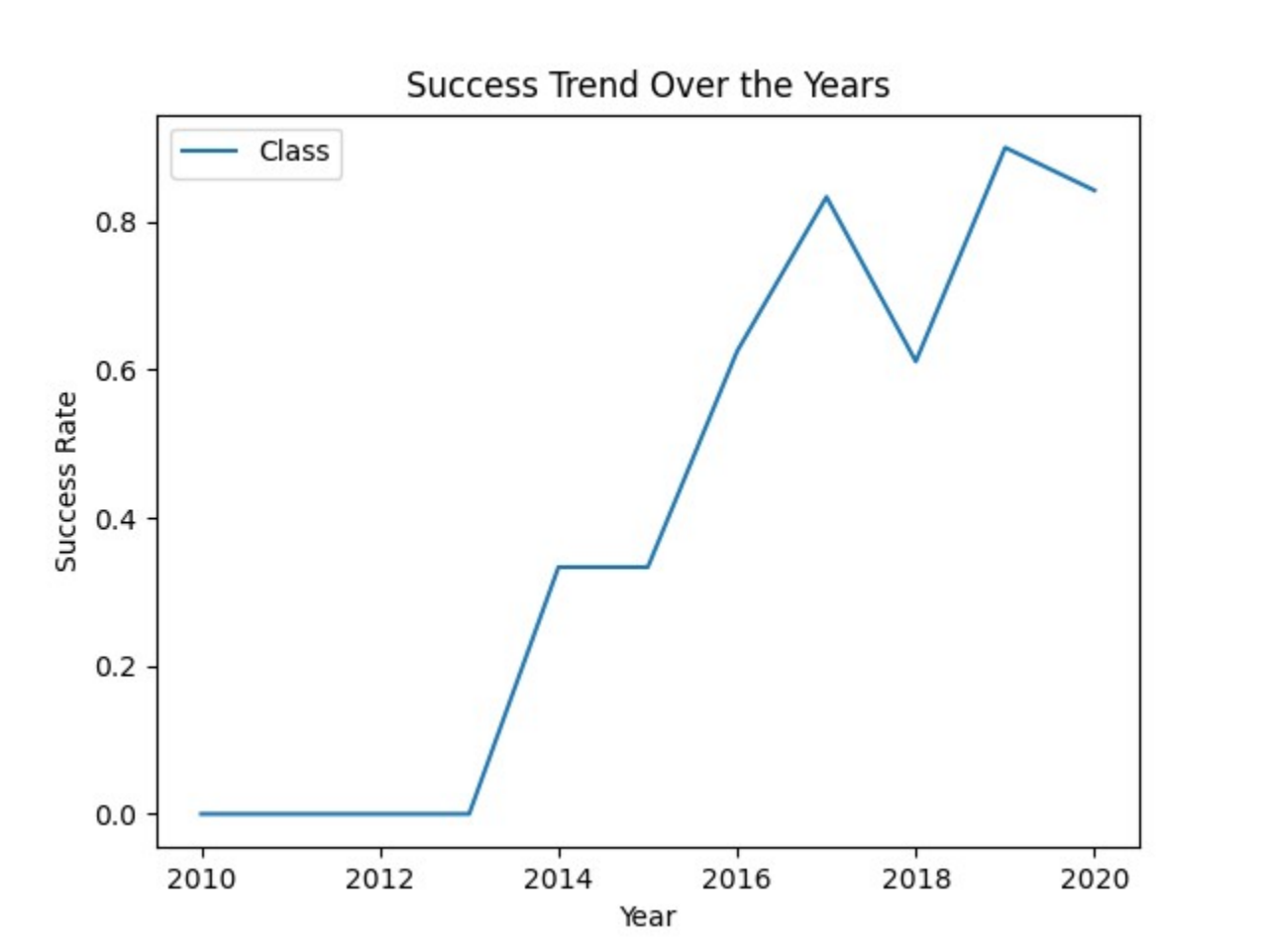
- LEO has only failed launches in the first two flights, the launches after those two are all successful.
- ES-L1, HEO, and GEO despite having 100% successful launch rates, all only have one recorded flight. Need more information to draw more solid conclusions on these orbits.
- The launches labeled SSO has had 5 launches and all of them were successful while the launch labeled as SO has only had one failed launch
  - When counting both, the final success rate is 5/6 or approx. 83.33% which is on par with the VLEO orbit.

Payload vs. Orbit Type



- The payload mass used and its success may vary on the type of orbit considered for the launch.
  - For example, SSO/SO seems to do best when using less mass (500 – 4000 kg) while some orbits like ISS and PO sees more success when using a bigger mass for its launch (> 4000kg)
  - Some orbits like GTO are hard to tell if the payload mass plays a role in its success.

Launch Success Yearly Trend



- There are two periods where the success rate stays constant between years: 2010 – 2013 and 2014 – 2015
- There are two periods where the success rate drops: 2017 - 2018 and 2019 – 2020.
- There are three periods where the success rate rises: 2013-2014, 2015 – 2017, and 2018 – 2019
- Overall, the success rate has been consistently increasing since 2010. Although it tends to drop between years every now and then.

Query - All Launch Site Names

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

There are only four distinct launch sites.

Two of the four have the same prefix (CCAFS).

Query - Launch Site Names Begin with 'CCA' (5 examples)

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success

Of these 5, there are some consistencies besides the launch site, all boosters use the F9 series and LEO is the orbit used for the launch. All examples here are successful too.

Query - Total Payload Mass Carried by NASA

total_payload_mass
99980.0

When considering all of the NASA launches on record, the total amount of mass used for payloads is almost 100,000 kg.

Query - Average Payload Mass by Booster F9 v1.1

avg_payload_mass
2534.6666666666665

The average payload mass used by the F9 v1.1 booster is approximately 2534.67 kg.

Query - First Successful Ground Landing Date

first_date
01/08/2018

The first successful ground landing in a launch takes place in 2018, during the start of the second week in January

Query - Successful Drone Ship Landing by Payload/Booster using a mass between 4000 and 6000 kg

Payload	Booster_Version
JCSAT-14	F9 FT B1022
JCSAT-16	F9 FT B1026
SES-10	F9 FT B1021.2
SES-11 / EchoStar 105	F9 FT B1031.2

There are 4 distinct payload and booster combinations that are associated with successful drone ship landings using a mass between 4000 and 6000 kg.

All payloads have either a prefix of JCSAT or SES and all boosters are associated with the Falcon 9 FT series.

Query - Total Number of Successful and Failure Mission Outcomes

failure_outcomes	success_outcome
1	3

There are three different ways to succeed in a launch mission while there is only one way to fail.

Query - 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

There are 12 distinct boosters versions that held the maximum payload mass ever used in a launch.

All boosters are apart of the Falcon9 B5 series.

Query – Records of Failed Drone Ship Landings in 2015

month	Booster_Version	Launch_Site	Landing_Outcome
10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

There are only two launches that had failed drone landings in October and April 2015 respectively.

Both took place in the launch site CCAFS LC-40.

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

Landing_Outcome	count
No attempt	5
Failure (drone ship)	2
Success (ground pad)	1
Success (drone ship)	1
Failure (parachute)	1
Controlled (ocean)	1

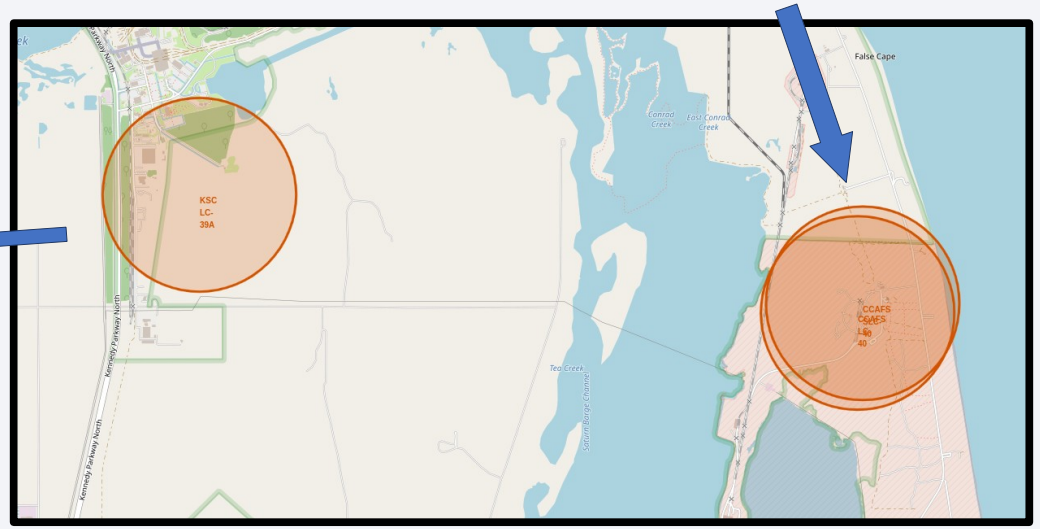
Between June 4, 2010 and March 20, 2017, most launches didn’t attempt a landing. There were two failed drone ship landings making it the second most common outcome during the time period.

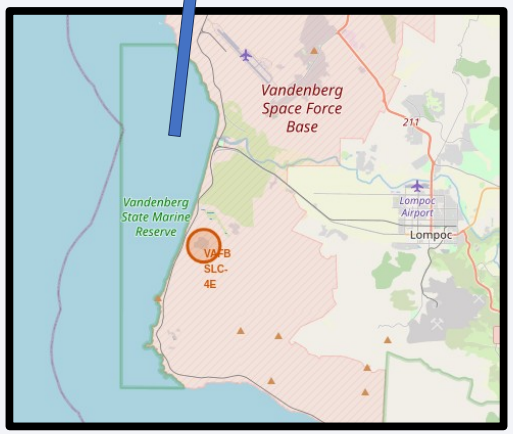
Lastly there is one successful ground pad landing, one successful drone ship landing, one failed parachute landing, and one controlled ocean landing.

Section 3

Launch Sites Proximities Analysis

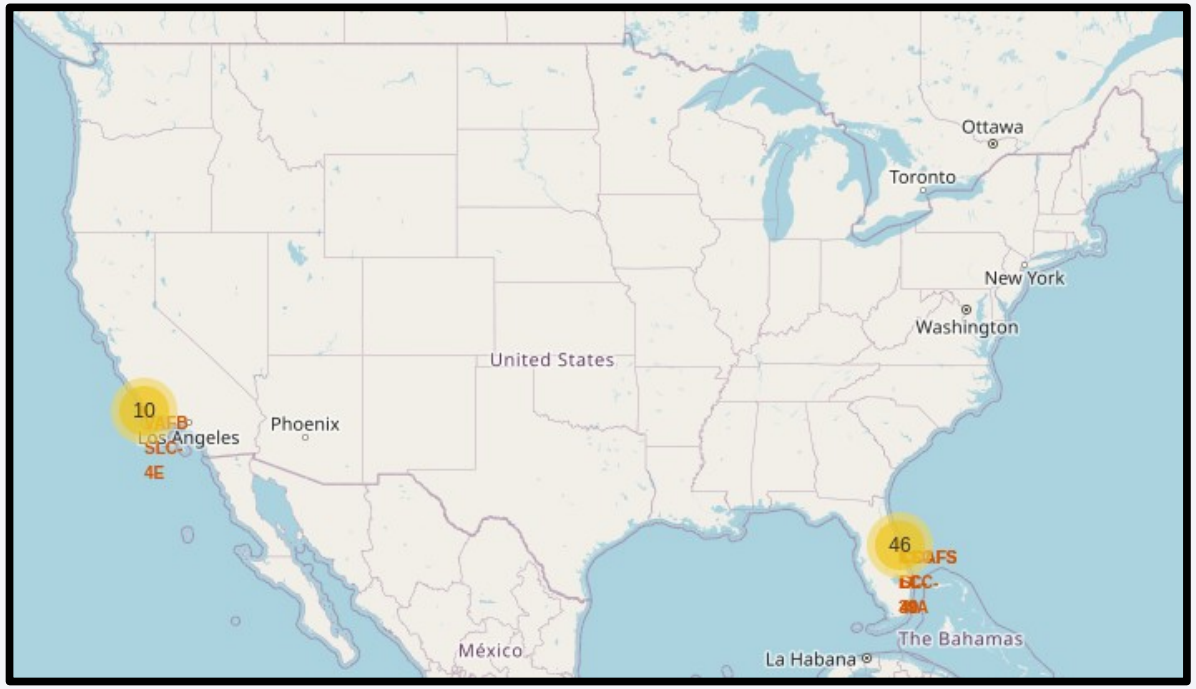
Location of Launch Sites

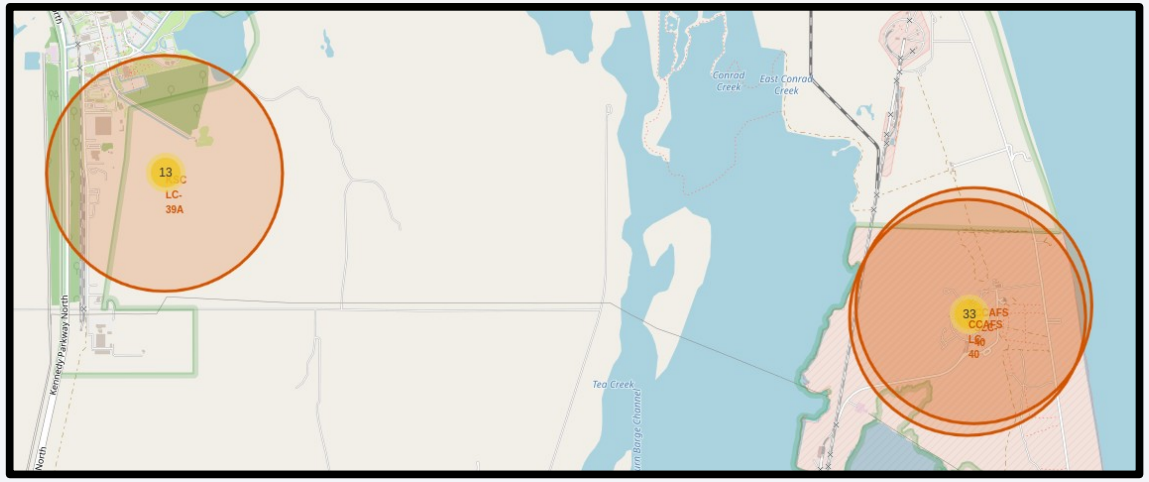


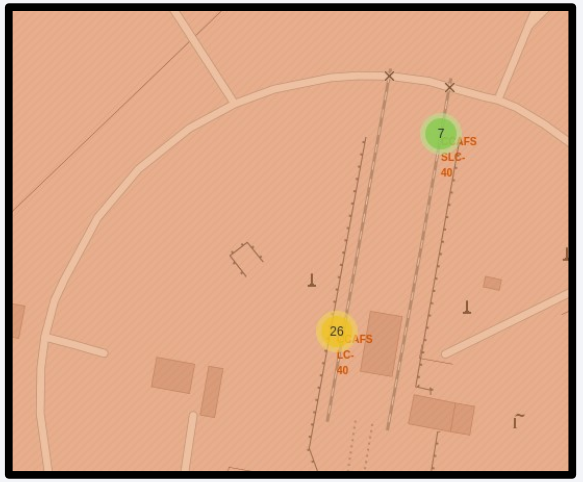


- One launch site is located in California, the rest are located in Florida.
- However two of the launch sites from the latter are extremely close to each other (CCAFS LC-40 and CCAFS SLC-40).
- Almost all sites have very close proximity to an ocean.
  - KSC LC 39A is the only launch site not near a coast

Launch Count By Launch Site

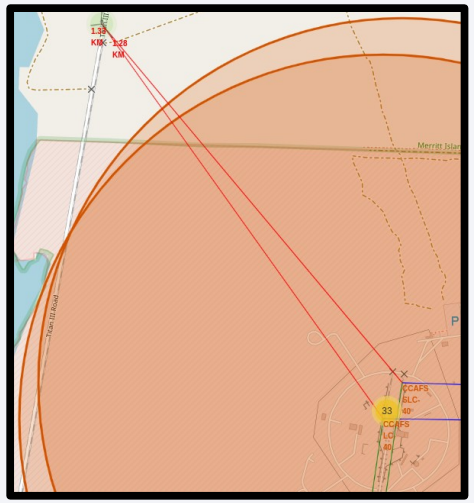




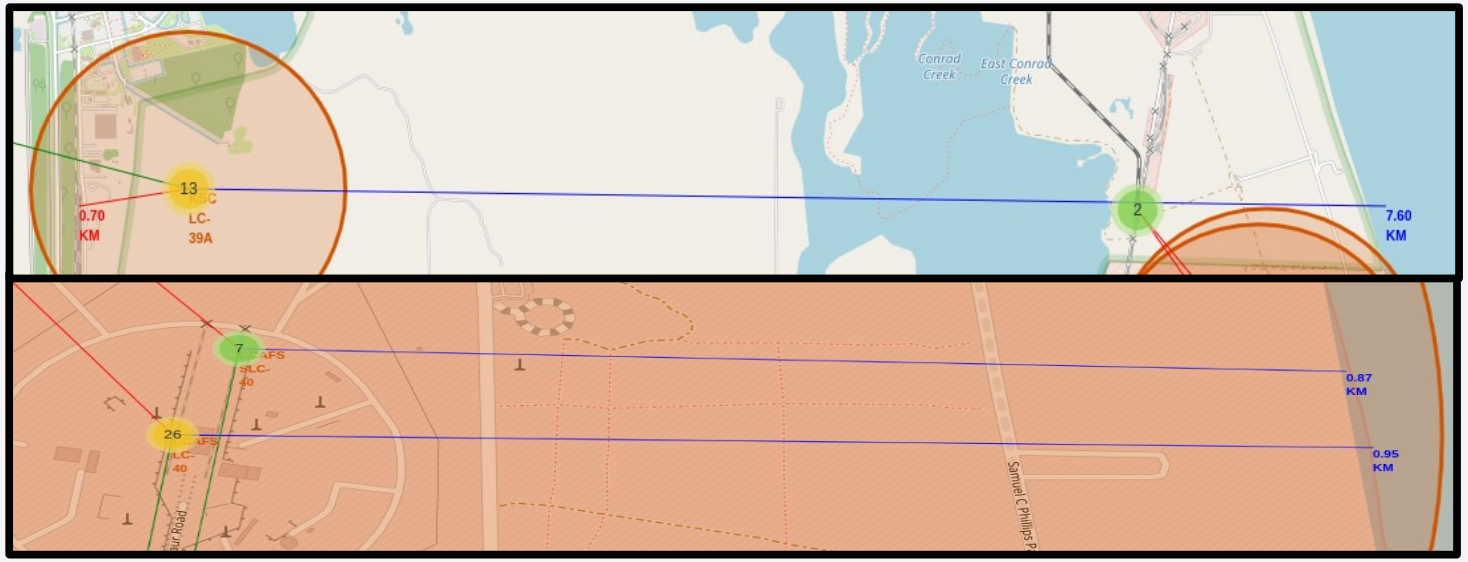


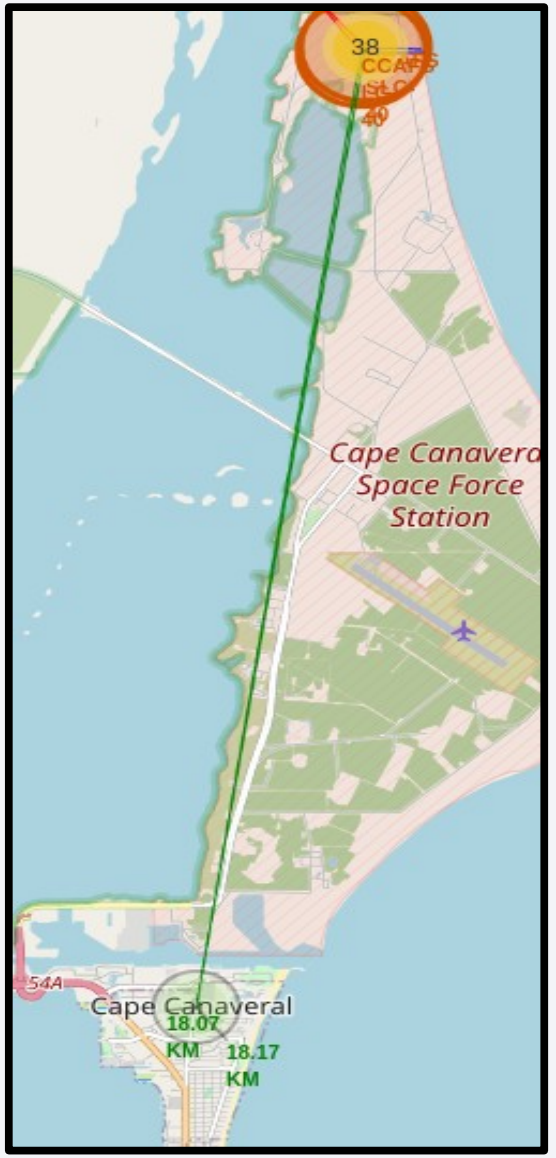
- Because there is only one launch site in California, SpaceX has a higher count of launches by the east coast than the west coast by more than 400%.
- The CCAFS LC 40 launch site has the most launches of any site (26 launches). The SLC 40 variant, on the other hand, has the least amount of launches (7 launches).

Launch Site Distance From Proximities

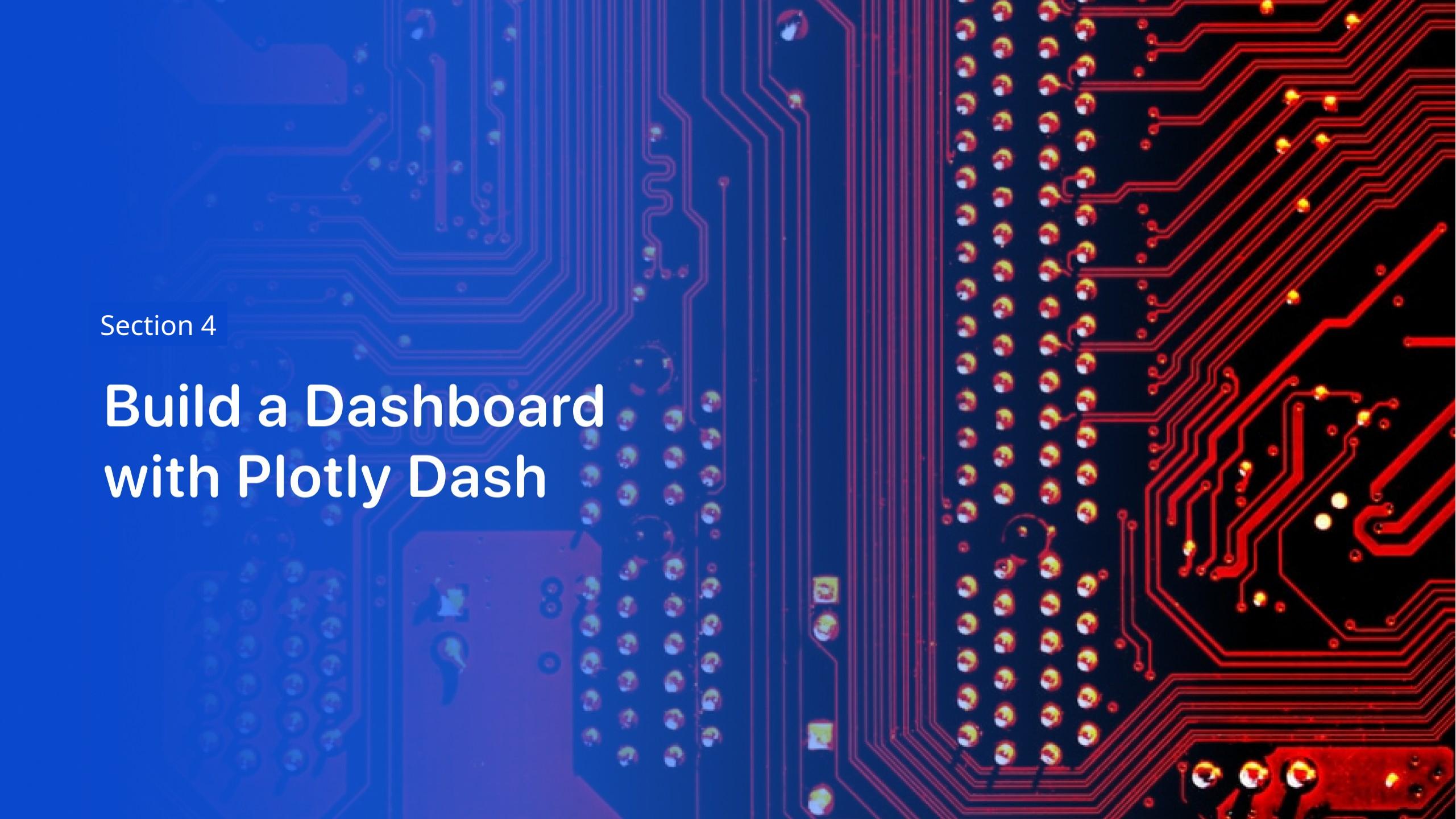


- All launch sites are typically **not far from an ocean** (no more than 2 km away).
  - The **KSC LC 39A** launch site is the **only exception** being **more than 7 km away** from the nearest coast. This launch site's closest access to a body of water is a creek.
- The proximity between sites and railways are also **relatively close** but they tend to have a **slightly larger distance than a launch site to the nearest coastline**
- However, **there is always a massive distance between cities and launch sites** with the CCAFS launch sites having nearly more than 20 times the distance between the nearest coastline and the sites. (18.12km vs 0.91km average).

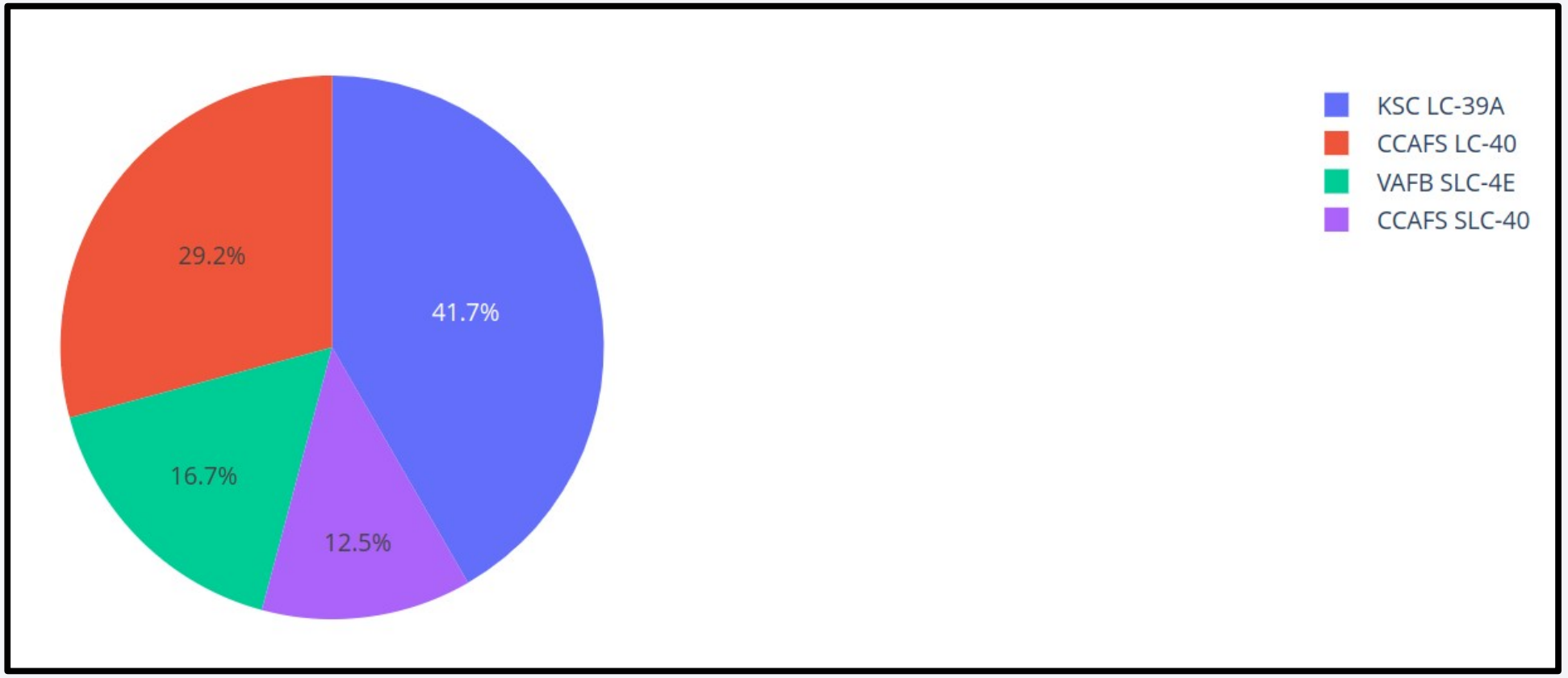




- Line color code
- Blue: Distance to nearest coastline
- Red: Distance to nearest railway
- Green: Distance to nearest city

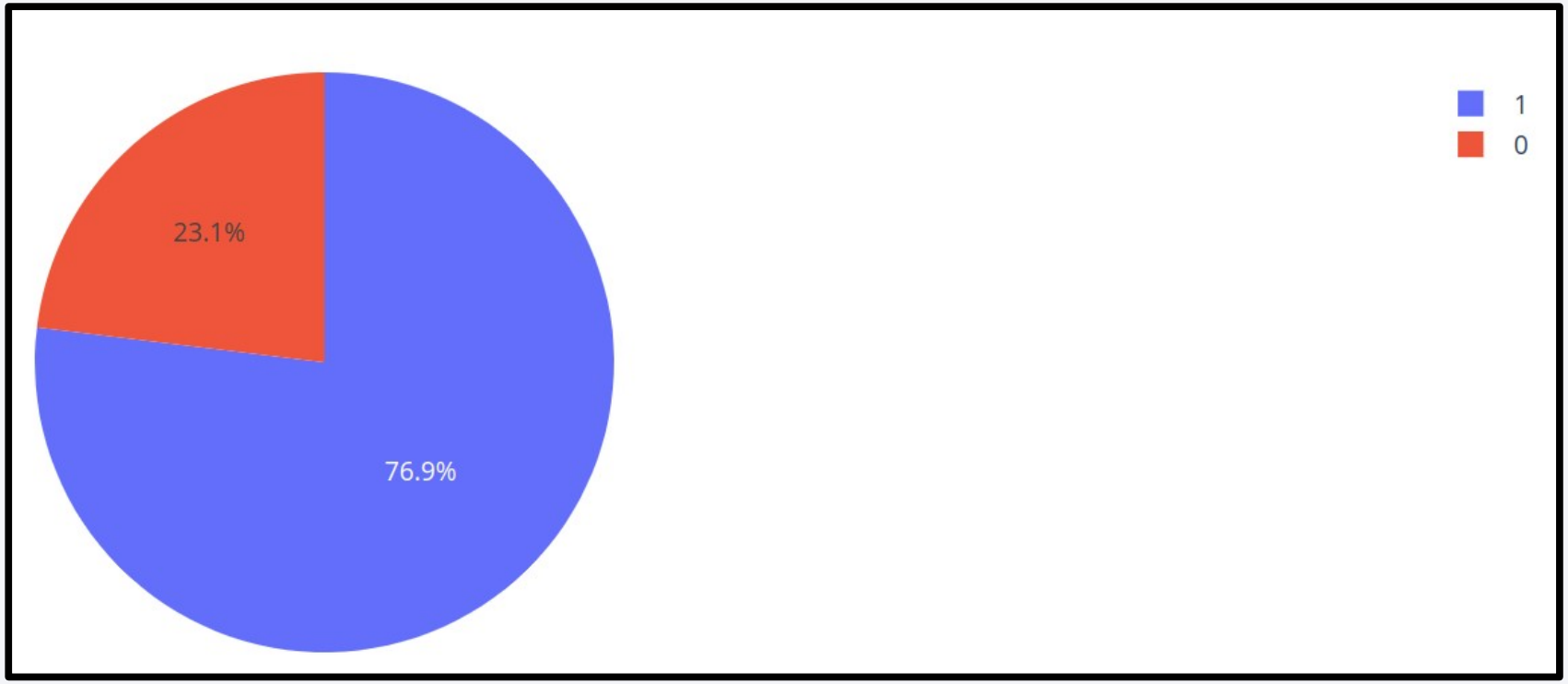

Section 4
Build a Dashboard with Plotly Dash

Relative Percentage of Successful Launches by Launch Site



Site KSC LC-39A holds the most successful launches at 41.7%. CCAFS LC-40 has the second most at 29.2%, VAFB SLC-4E is third at 16.7% and CCAFS SLC-40 at last place with only 12.5% of all successful launches. If we count the CCAFS launch sites as a singular site, it would be tied with KSC LC-39A for the most successful launches.

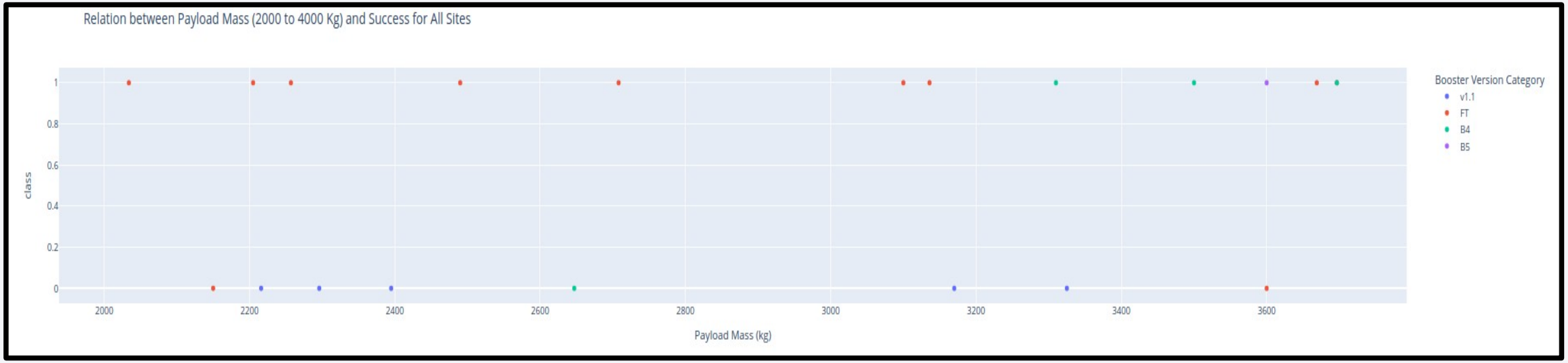
Successful Launches Within Site KSC LC-39A



- 1 – Successful Launch
- 0 – Failed Launch

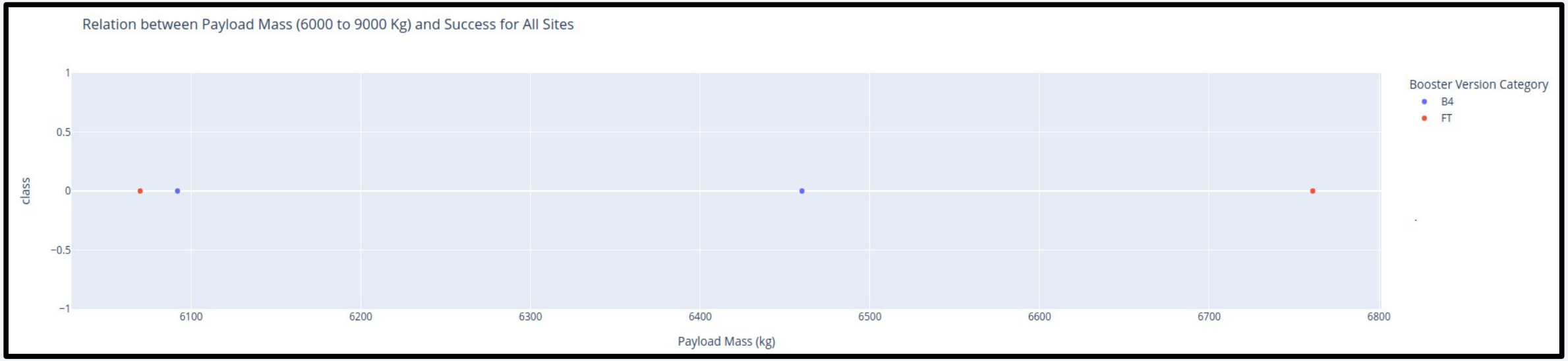
When it comes to the launch site that has the most success, it holds a success rate of 76.9%. In other words, a little over ¾ of launches that come from launch site KSC LC-39A are successful.

Payload Range with Highest Success



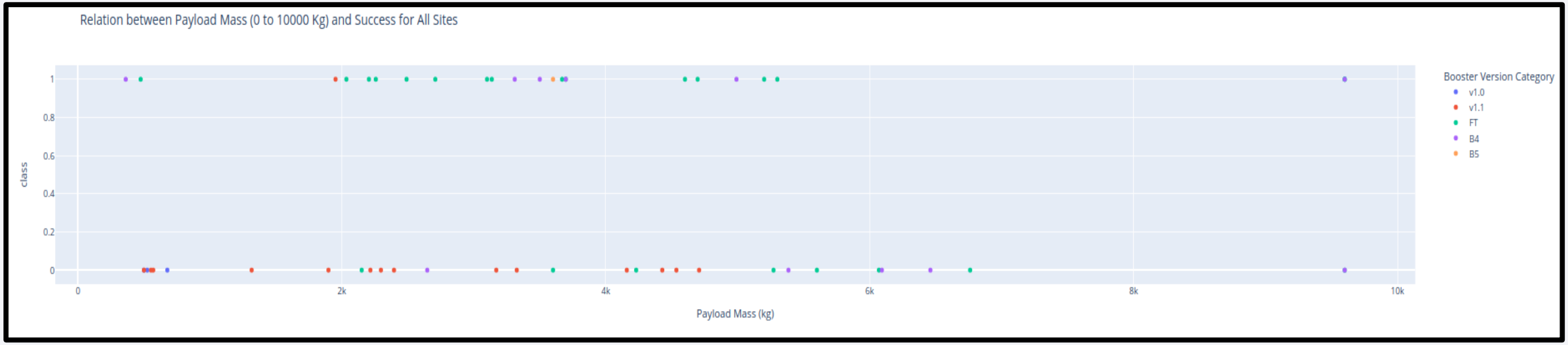
- Launches more than 2000kg and less than 4000kg tend to have more successful launches than not.
  - V1.1 is responsible for most failed launches in this range
  - Not enough info for B5 despite having only successful one launch in this range.
- FT has the highest success rates out of any boosters with B4 as a not-as-close second.

Payload Range with Lowest Success



- Setting the range between 6000kg and 9000kg shows the lowest possible success at 0%.
- Half of the failed launches come from the B4 booster while the other half is from FT.

Payload Mass vs Success (All Masses)

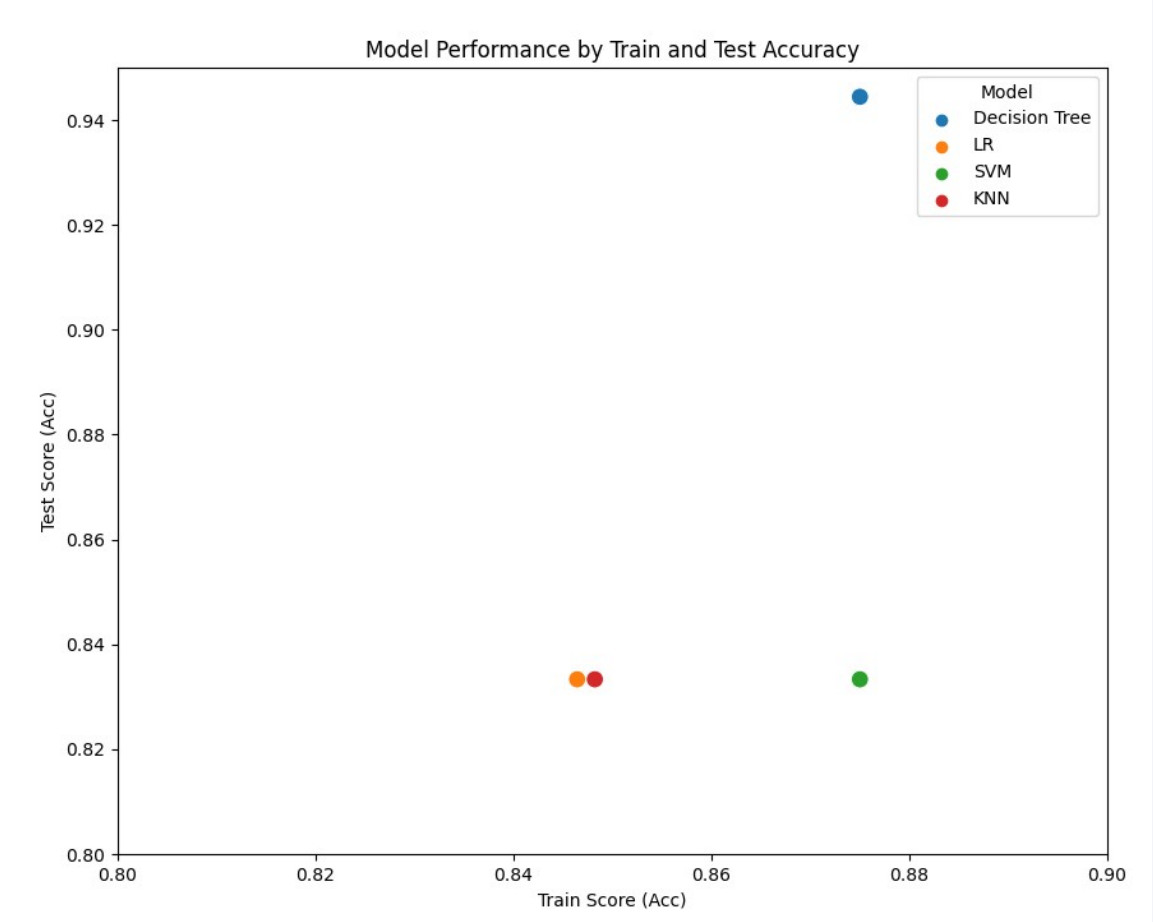


- While FT and B4 are the only boosters to be in the range with the lowest success, they are also the only boosters prominent in the range with the highest success rate.
  - This exemplifies the idea that there's no "one-size-fits-all" range and that certain payload ranges perform better when using certain boosters.

Section 5

Predictive Analysis (Classification)

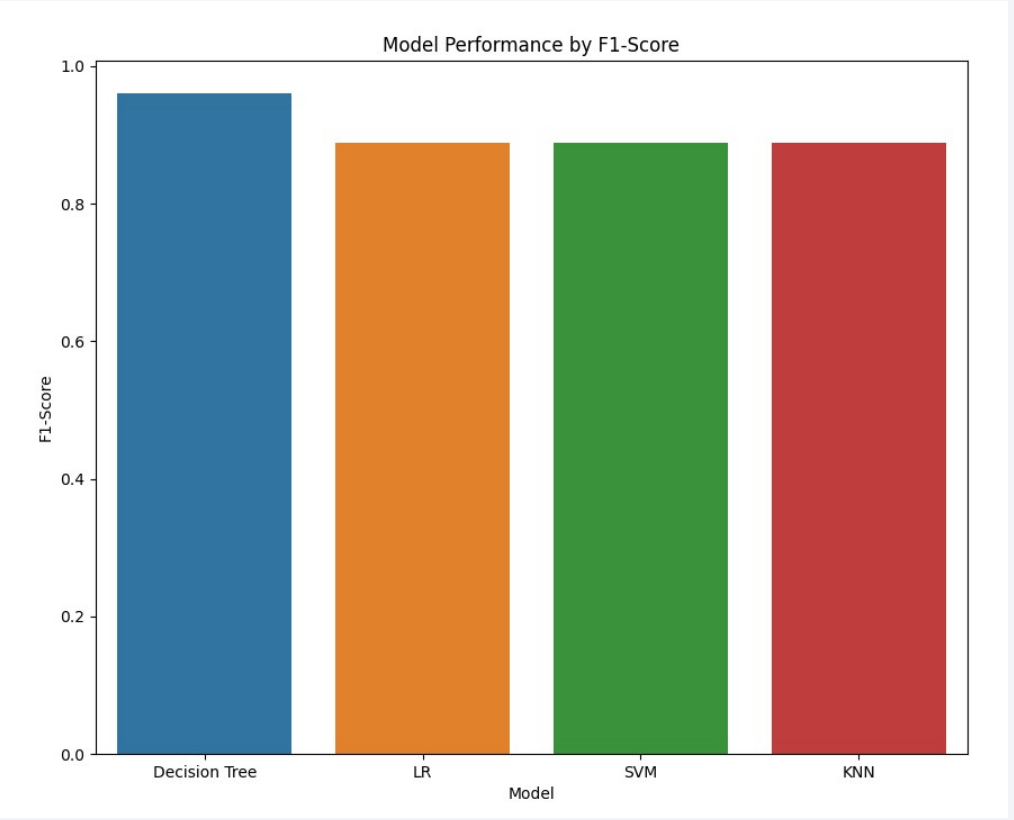
Classification Accuracy Results



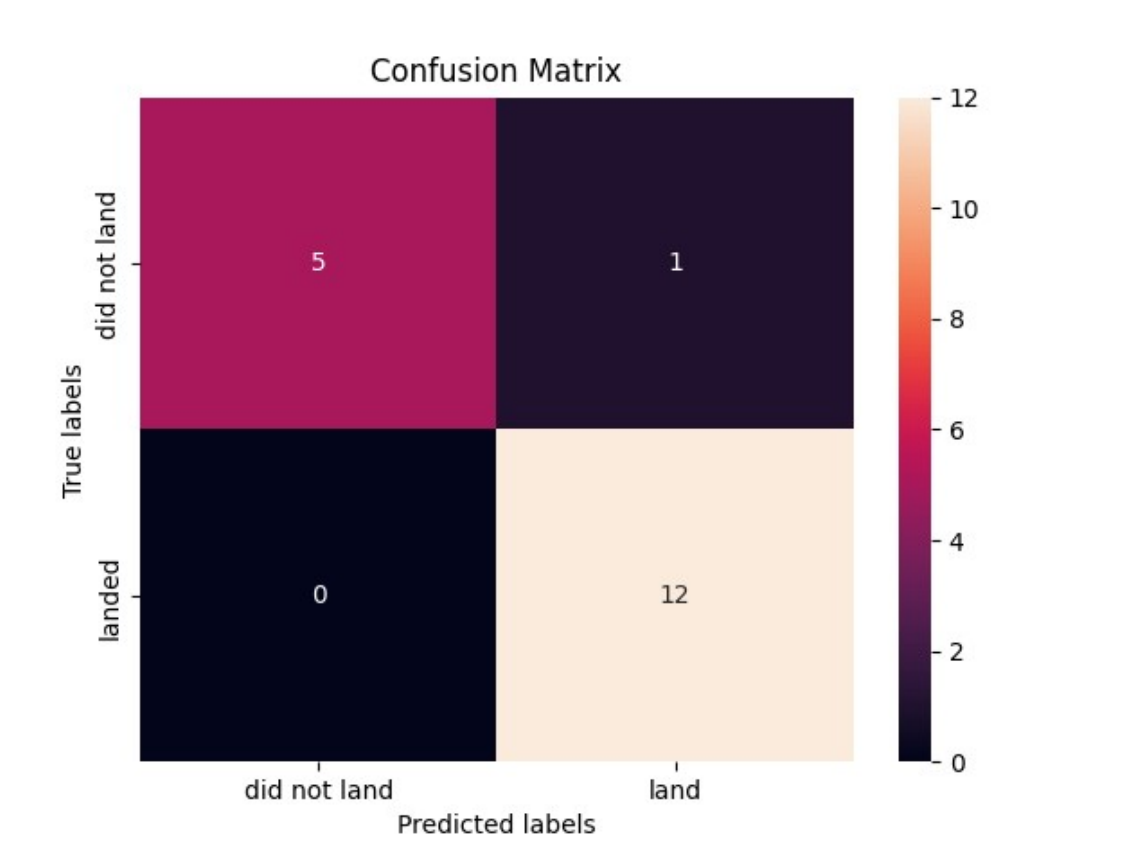
- Decision Tree, while tied with SVM for the highest train accuracy, still holds the hightest test accuracy.
- The SVM model, has a similar performance to the Logistic Regression and the KNN model when it comes to test accuracy.

Classification F1-Score Results

- All models have performed fairly well at the least.
- Overall however, the decision tree model has the slight edge over the others.
  - DT F1 = 0.96
  - F1 of other models = 0.888889



Decision Tree Confusion Matrix



- The decision tree model does a perfect job predicting when a rocket's launch will not end in succession.
  - Although, none of the other models ever encountered a false negative either.
- While the DT model encountered a single false positive. It's still the most effective in predicting successful landings, correctly predicting a successful landing 12 times out of 13.
  - Consequently, this model is the best model to deal with false positives, as every other model encountered at least 3 false positives during evaluation.

Results

Exploratory data analysis results:

- Launch sites typically perform better at later flights than earlier ones
- Each launch site has a similar optimal range of payload mass (either between 2000 and 4000 or more than 10000 kg).
- Some orbit types are more effective than others, namely SSO/SO and VLEO
  - Some orbits (ES-L1, GEO, HEO) are technically better as they have never failed a launch but those orbits have only one launch so we need more info to verify their effectiveness.

Map and Dashboard observations:

- 3 out of the 4 launch sites are in the east coast (Florida) with the exception being in California.
- 3 out of the 4 launch sites are near a coastline (less than 2km away).
  - The exception is > 7km away from the nearest coastline.
- The launch site furthest away from a coastline is also the most successful site, holding 41.7% of the successful Falcon 9 launches.
- There is always a massive distance from launch sites and cities but launch sites are typically close to railways and coastlines.

Predictive analysis results

- Out of the four models developed, the decision tree model performed the best with an F1-Score of 0.96. However, the other models aren't too far from first place all with an F1-Score of approx. 0.88
- Every other model besides the decision tree model tends to struggle slightly more when it comes to detecting false positives. In other words, other models are more prone to predicting a launch will be successful when it doesn't in reality.

Conclusions

- All launch sites tend to do **better at later launches** despite the occasional failure.
- **ES-l1, GEO, HEO** while being the only orbits to have a **100% success rate,** need more information to confirm their effectiveness.
  - More exploration with these orbit types is encouraged.
  - **SSO** and **VLEO** are overall effective orbits to use for a launch as they both have an approx. 83% success rate.
- Despite being the only launch site that is **more than 2 km away from a coastline**, **KSC LC-39A** has the **highest success rate** out of the four main launch sites.
  - Further study and exploration on this launch site is desired to find any other notable differences from other sites
- A payload mass **between 2000 kg and 4000 kg** with a **FT or B4** booster would be an **optimal** setting for a successful launch. A mass more than 10000 kg may also work.
  - If using either type of booster, avoid going between 4000 and 9000 kg if possible.
- A **decision tree** algorithm would be the **most ideal choice for predicting** a successful rocket launch.
  - Exploration with other tree-based models is encouraged to potentially further improve performance.

Appendix

All assets and datasets used can be found on my GitHub page at https://github.com/MSB46/DataProjects/tree/main/IBM_Capstone

Thank you!