IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

By Arhan Sawhney
11/28/24

Outline

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

Executive Summary

Methodology

The methodology used in this report contains data from *SpaceX REST API* containing:

- Launch data
- Rocket data
- Core data
- Capsule data
- Starlink data
- Launchpad & landing pad data

Solution

Using a predictive ML model with the highest accuracy for successful future Falcon 9 landings from SpaceX (whether the first stage will be reused)

Introduction

- SpaceX is another spacecraft manufacturing/technology corporation, with their most successful rocket being the Falcon 9.
- This can be attributed to the Falcon 9's reusability and cost, where other manufacturers are **2.6x** more expensive.
- The questions we want to acquire from this report is:
  - What is the **best model** for predictive analysis for *future F9 landings*?
  - What **booster versions** of the F9 have the *highest success* rate?
  - Which **launch site** has the *highest success* rate for the F9?
  - Which **launch site** the *highest failure* rate for the F9?
  - What **payload range** has the *highest success* rate for the Falcon 9?
  - What **type of landing** outcome had the *highest success*?

Section 1

Methodology

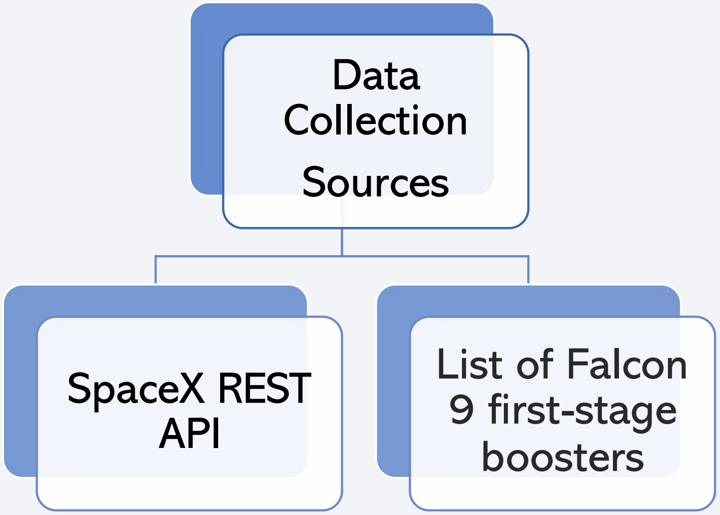
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using the SpaceX REST API (https://api.spacexdata.com/v4/launches/past) in JSON format, and Wikipedia tables showing the previous list of launches (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - First, we filter only F9 launches, get rid of missing values and format the landing results to a new column if a landing was successful (1) or a failure (0).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A classification model is built using the scikit-learn library in Python, then the best parameters are automatically chosen and we split the training data into test data to evaluate the accuracy.

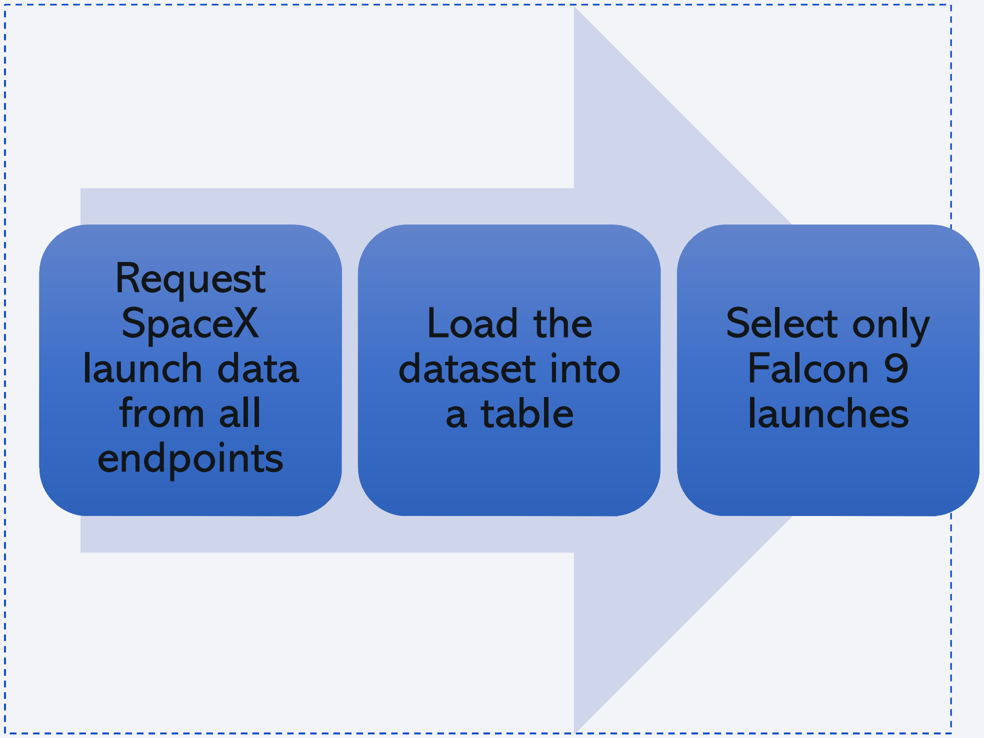
Data Collection

- The SpaceX Rest API (https://api.spacexdata.com/v4/) contains many different endpoints for different additional information, important to our findings such as:
  - the booster name
  - location of the launch site
  - payload mass
  - core information
- From Wikipedia, we web-scrape "List of Falcon 9 first-stage boosters" table that gives us additional information about
  - the payload used
  - the specific booster version for each serial number.



Data Collection – SpaceX API

- Using the SpaceX API, we send a **request** to the website for our desired information (in this case, the rocket, launchpad, payload & cores) to determine the outcome and reusability of each flight.
- After that, the dataset we need is loaded into a table. Only the Falcon 9 launches are required here for this research.
- This will later help us find key trends in the data, perform **data wrangling** and answer the initial questions stated in the introduction.



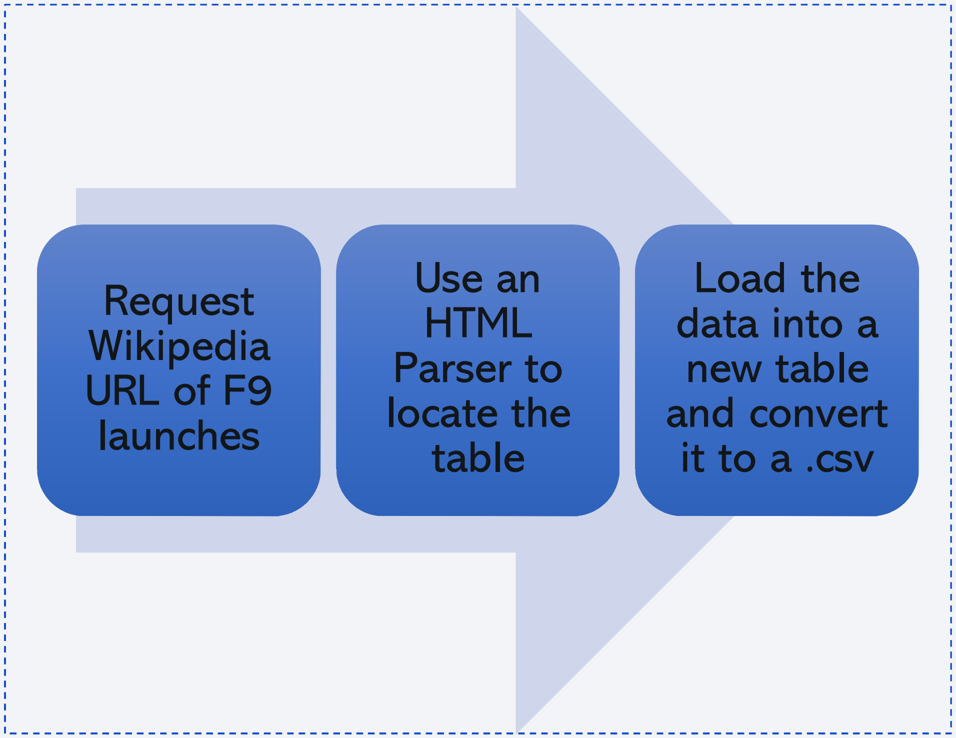
Request SpaceX launch data from all endpoints

Load the dataset into a table

Select only Falcon 9 launches

Data Collection - Scraping

- First, we send a request to the server for scraping the data.
- We then find the table we want to use for this project by parsing the HTML structure of the website.
  - Here, we are using the BeautifulSoup library in Python to achieve this.
- The table gets loaded into a new table, and finally we convert the dataset to a .csv for future use.



Request Wikipedia URL of F9 launches

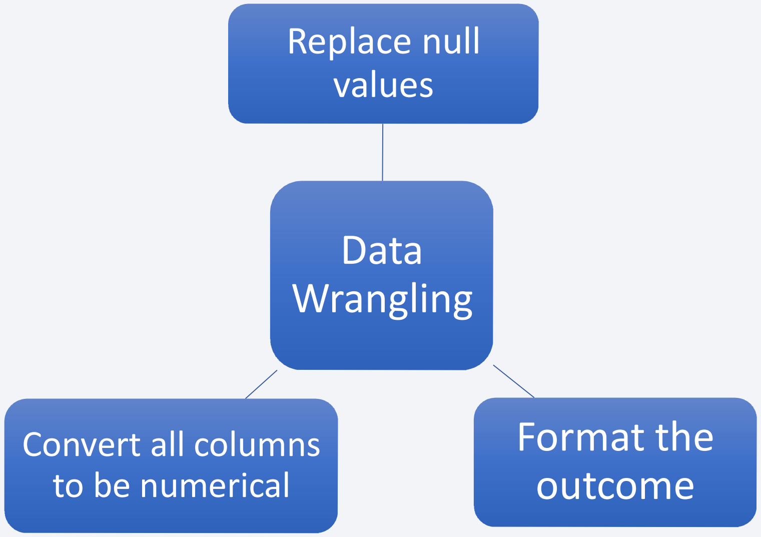
Use an HTML Parser to locate the table

Load the data into a new table and convert it to a .csv

Data Wrangling

- Now that we have collected our data, we need to clean it for exploratory data and predictive analysis.

1. Since we need observations with non-missing values, we replace all missing values for the mean payload mass by taking the mean per launch.
2. Our mission outcome comes in multiple categories, so for simplicity we can convert the result into a binary outcome where 1 means that the first stage landed successfully and 0 means that the second stage did not land successfully.
3. One-hot encoding will also be used on the orbit, launch-site, landing pad and the serial number of the flight for predictive analysis.



EDA with Data Visualization

1. The relationship between the flight number and launch site is used to map the success overtime for each launch site.
2. The relationship between payload mass and launch site is used to figure out what payload mass range is the safest for each site.
3. The success rate for each orbit type is used to find the orbits with the highest success rate.
4. The relationship between the flight number and orbit type is used to visualize the success of each orbit over time.
5. The relationship between payload mass and orbit type is used to figure out the optimum payload mass range for each orbit.
6. The launch success rate per year shows the success rate of launches over time.

EDA with SQL

The SQL queries that were performed:

- The names of each launch site
- Each launch performed in the Cape Canaveral Space Force Station
- The total payload mass carried by boosters launched by NASA and the average payload mass carried by booster version F9 v1.1
- The date of the first successful landing in ground pad
- Booster names which had successful drone ship landings and a payload mass between 4000 and 6000 kg.
- Number of successful and failed mission outcomes
- All booster versions carrying the maximum payload mass
- Record of month, booster version and launch site where the landing outcome is a drone ship failure in 2015.
- Count of each landing outcome between the date 2016-06-04 and 2017-03-20 (descending order).

Build an Interactive Map with Folium

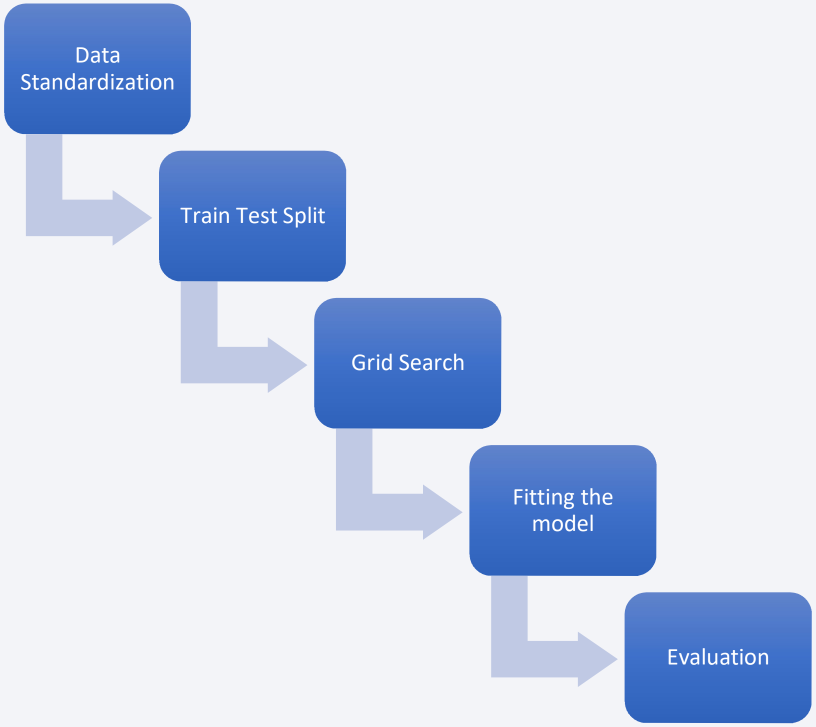
- Markers were added to each launch site, so it is easy to locate.
- A marker cluster was added to show the landing outcomes (successful or not) for each launch site.
- Lines were added to show the proximity of Kennedy Space center to its' coastline (5.14km) and the Vandenberg Space Launch complex to the nearest city of Lompoc (14.1km)

Build a Dashboard with Plotly Dash

- A pie chart representing the total success of launches, to determine which launch site has the highest amount of success for all sites and the success rate for each site
- A scatter plot showing the correlation between payload and success of each launch, to find the optimum payload range for each booster version, by all sites or for each site.

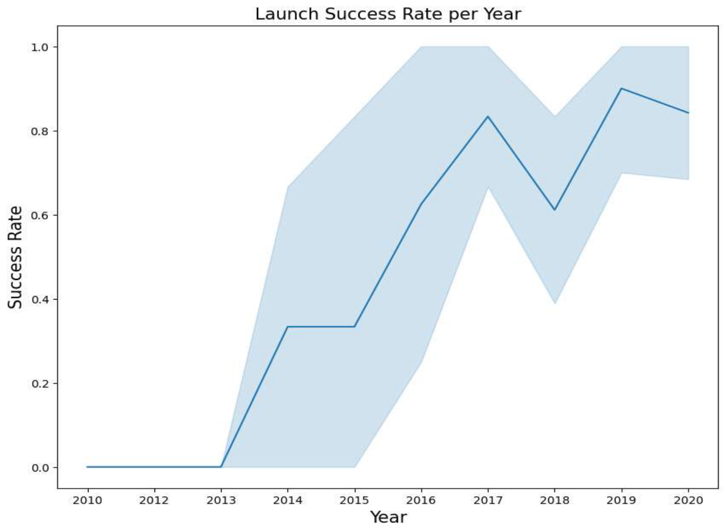
Predictive Analysis (Classification)

- Each predictive model was built using the scikit-learn library, where we load and standardize the data for the model.
- Then the dataset is split for training and testing. We will be comparing the performance of classification algorithms using the test set compared to the actual values.
- The best hyperparameters for each model is automatically selected to find the best parameters when fitting our model on the training set.
- Finally, we evaluate the accuracy of each model using our test data and by plotting a confusion matrix.



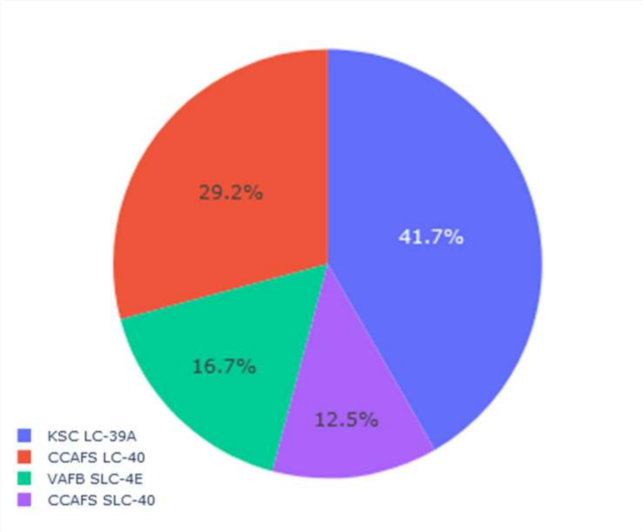
Results

Exploratory Data Analysis Results



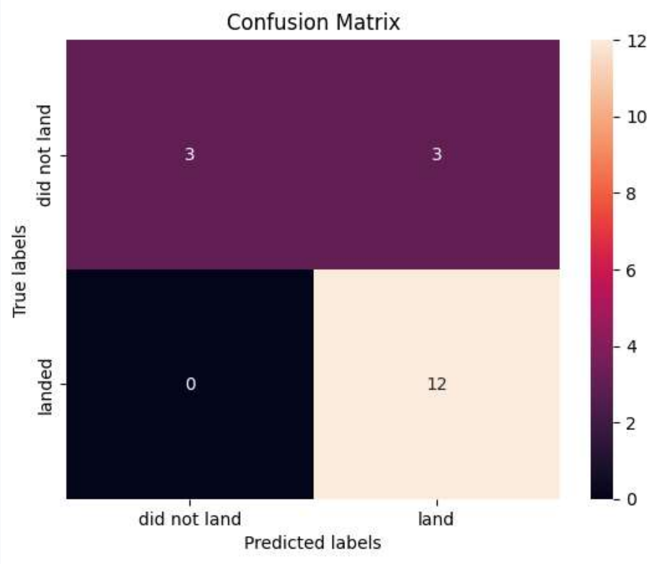
Launch Success Rate per Year

Interactive Analytics Results

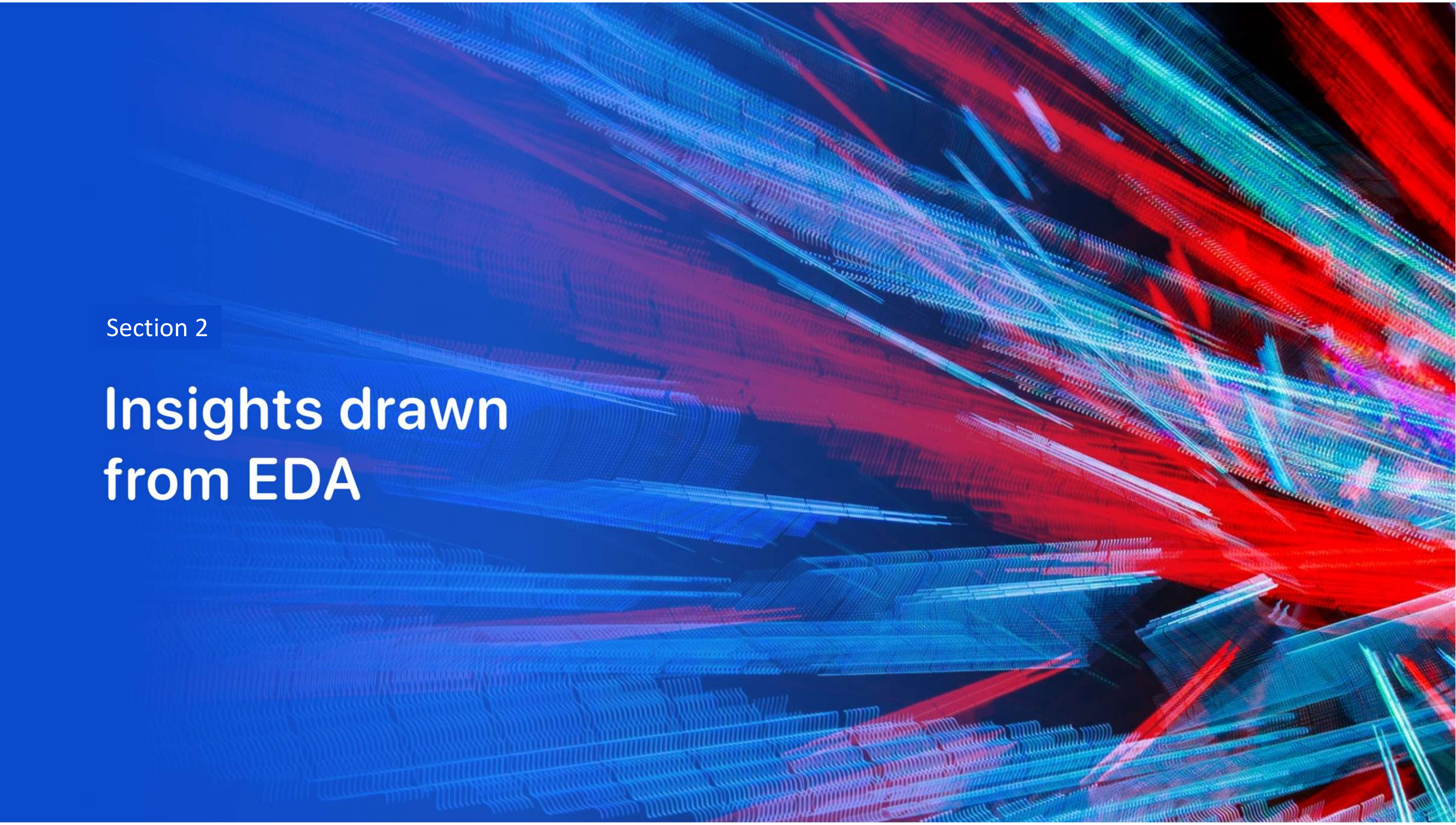


Total Success of Launches by Site

Predictive Analysis Results



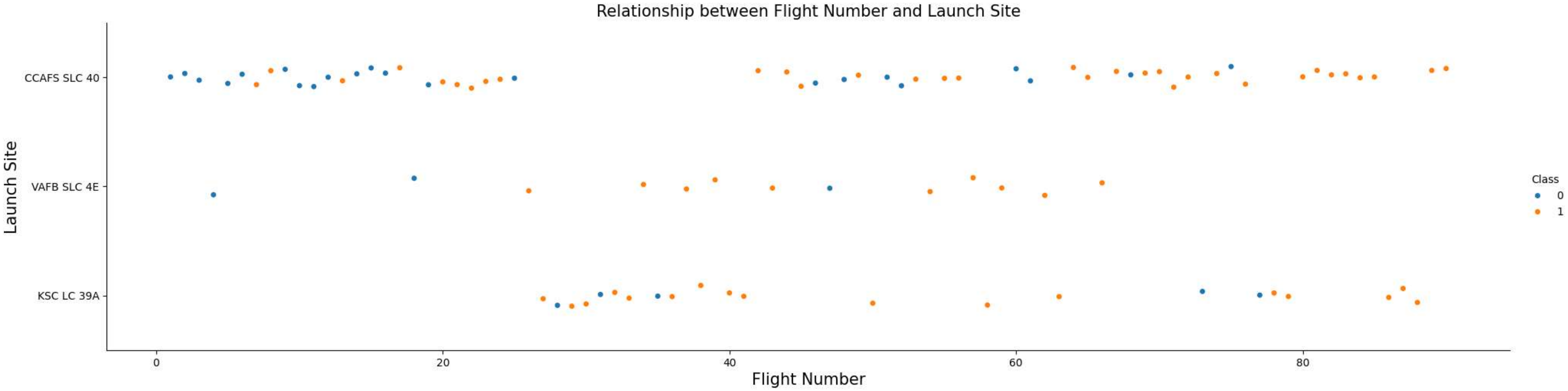
Decision Tree Classifier (Best Performing Model)



Section 2

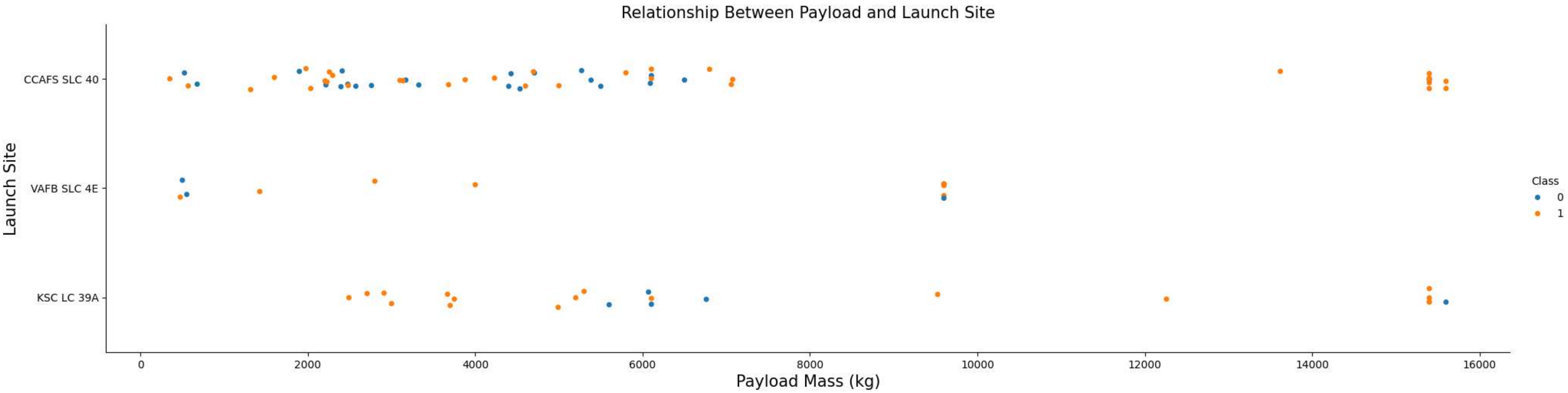
Insights drawn from EDA

Flight Number vs. Launch Site



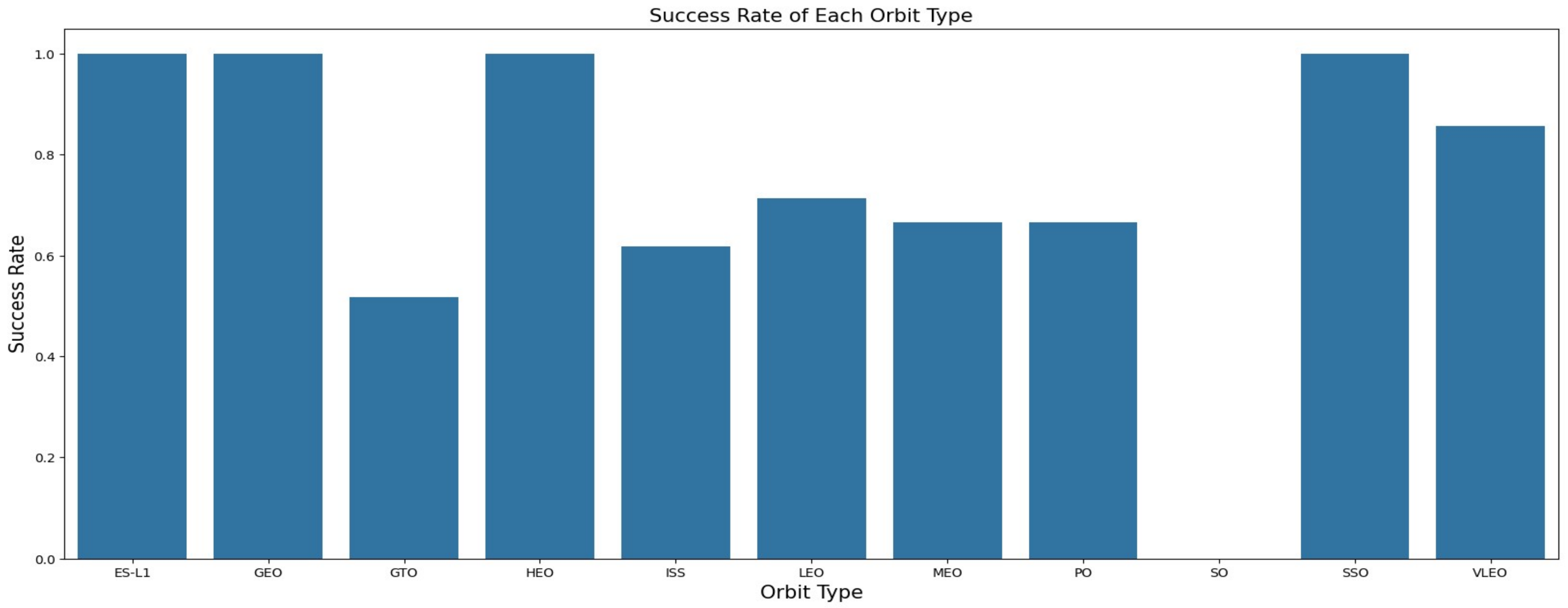
Here, we can see that CCAFS SLC-40 has the most launches and recently they have had more successful launches, whereas VAFB SLC 4E has the least launches. KSC LC-39A hasn't had any flights until around flight number 30 but seems to have the highest success rate over time.

Payload vs. Launch Site



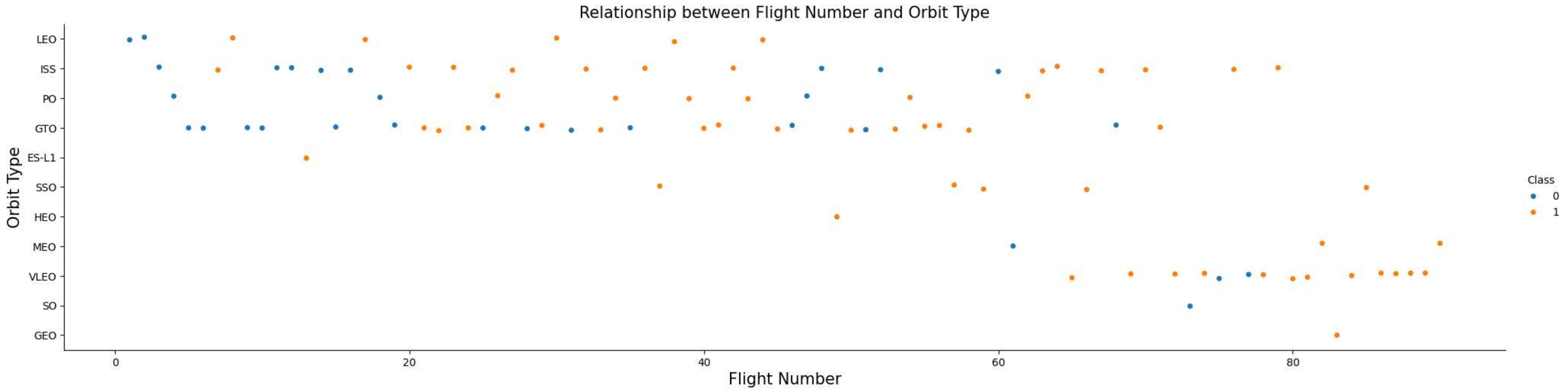
It’s hard to make out the optimum payload for CCAFS SLC-40 as it is nuanced but seems to handle high payloads very well. Here we can notice that KSC LC-39A has an optimum payload range of 2300 to around 3800 kilograms of weight.

Success Rate vs. Orbit Type



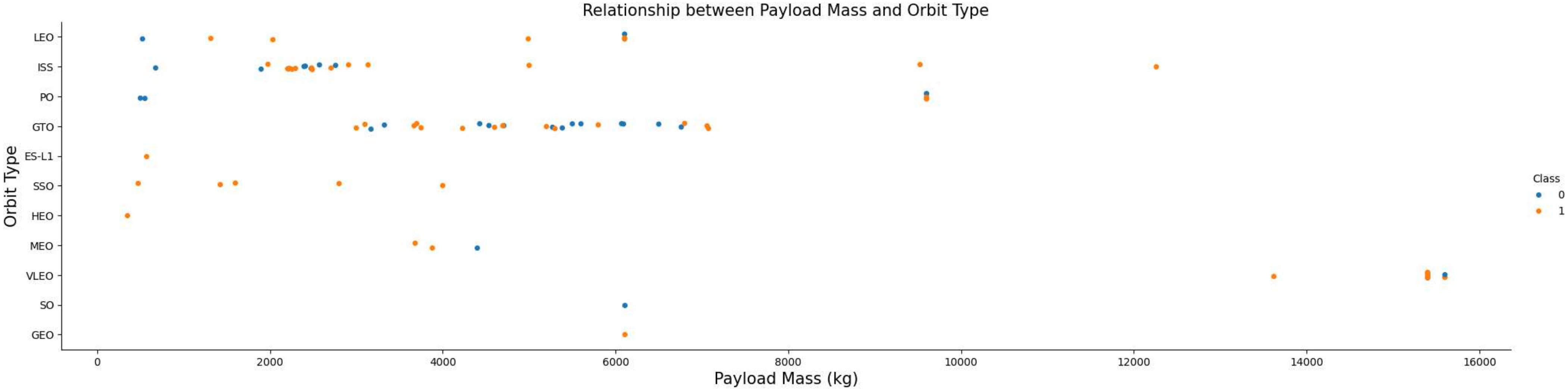
Here, we can see that the highest success rates include highly elliptical orbit, sun-synchronous orbit, circular geosynchronous orbit and the Lagrange point orbit.

Flight Number vs. Orbit Type



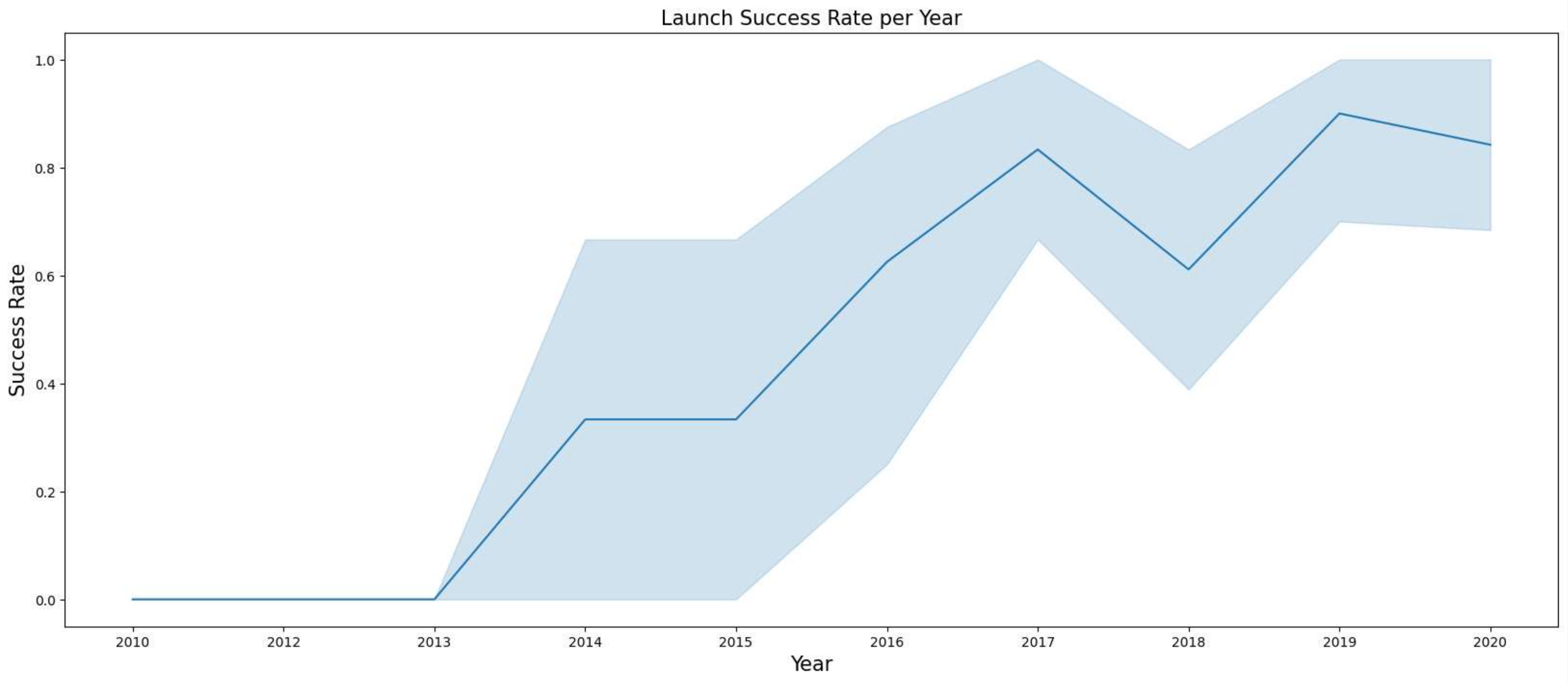
The Low Earth Orbit seem to be correlated with the flight number, appearing in most records. The Very Low Earth Orbit has a very high success rate, only appearing much later and lastly the geostationary transfer orbit seems to have no correlation.

Payload vs. Orbit Type



The Low Earth Orbit allows for a higher payload, along with the International Space Station and Very Low Earth Orbit, unlike the geostationary transfer orbit which has both successful and non-successful landings. The Sun-Synchronous orbit has a lower payload threshold (up to 4000kg), though no failed landings are present.

Launch Success Yearly Trend

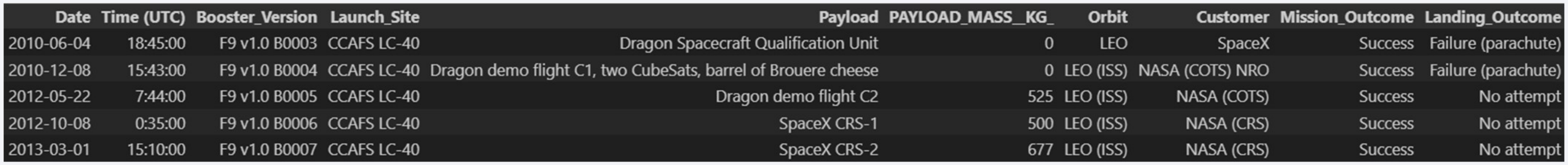


The success rate starts increasing in 2013 to 2017, decreasing in 2018 and then staying around 75-80%.

All Launch Site Names

- All the launch site names for Falcon 9 launches include:
  - Cape Canaveral Space Launch Complex 40
  - Vandenberg Space Launch Complex 4E
  - Kennedy Space Center Launch Complex 39A
  - Cape Canaveral Launch Complex 40

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

- The earliest five records at the Cape Canaveral Space Launch Complex show that most of them were carried out by NASA all in Low Earth Orbit and failed landing outcomes from 2010 to 2013.

Total Payload Mass

- The total payload mass carried by boosters from NASA (CRS) is 45,596kg in total.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 version 1.1 is 2928.4kg.

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and a payload mass between 4000 and 6000 is the Falcon 9 Full-Thrust B1022, B1026, B1021.2 & B1031.2.

Total Number of Successful and Failure Mission Outcomes

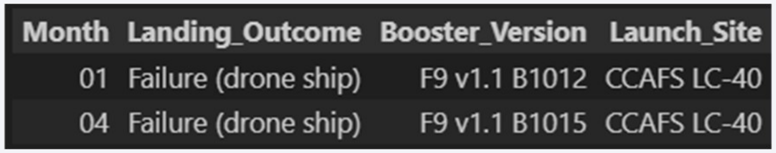
- There is only 1 failed outcome out of 101 mission outcomes, the rest being successful.

Boosters Carried Maximum Payload

- The names of the booster versions which have carried the maximum amount of payload are the Falcon 9 Block 5 B1048.4, B1049.4, B1051.3, B1056.4, B1048.5, B1051.4, B1049.5, B1060.2, B1058.3, B1051.6, B1060.3 and the B1049.7.

2015 Launch Records

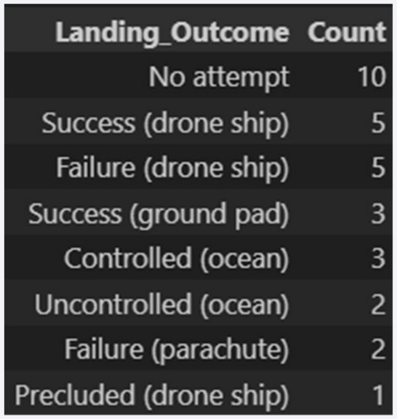
- The failed landing outcomes (drone ship) happened at the Cape Canaveral Launch Complex 40 in January and April respectively using the booster v1.1.



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

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

- The landing outcomes between 2010-06-04 and 2017-03-20 show that most of the landing outcomes had no attempt, other than that the drone ship both had the highest number of successes and failures

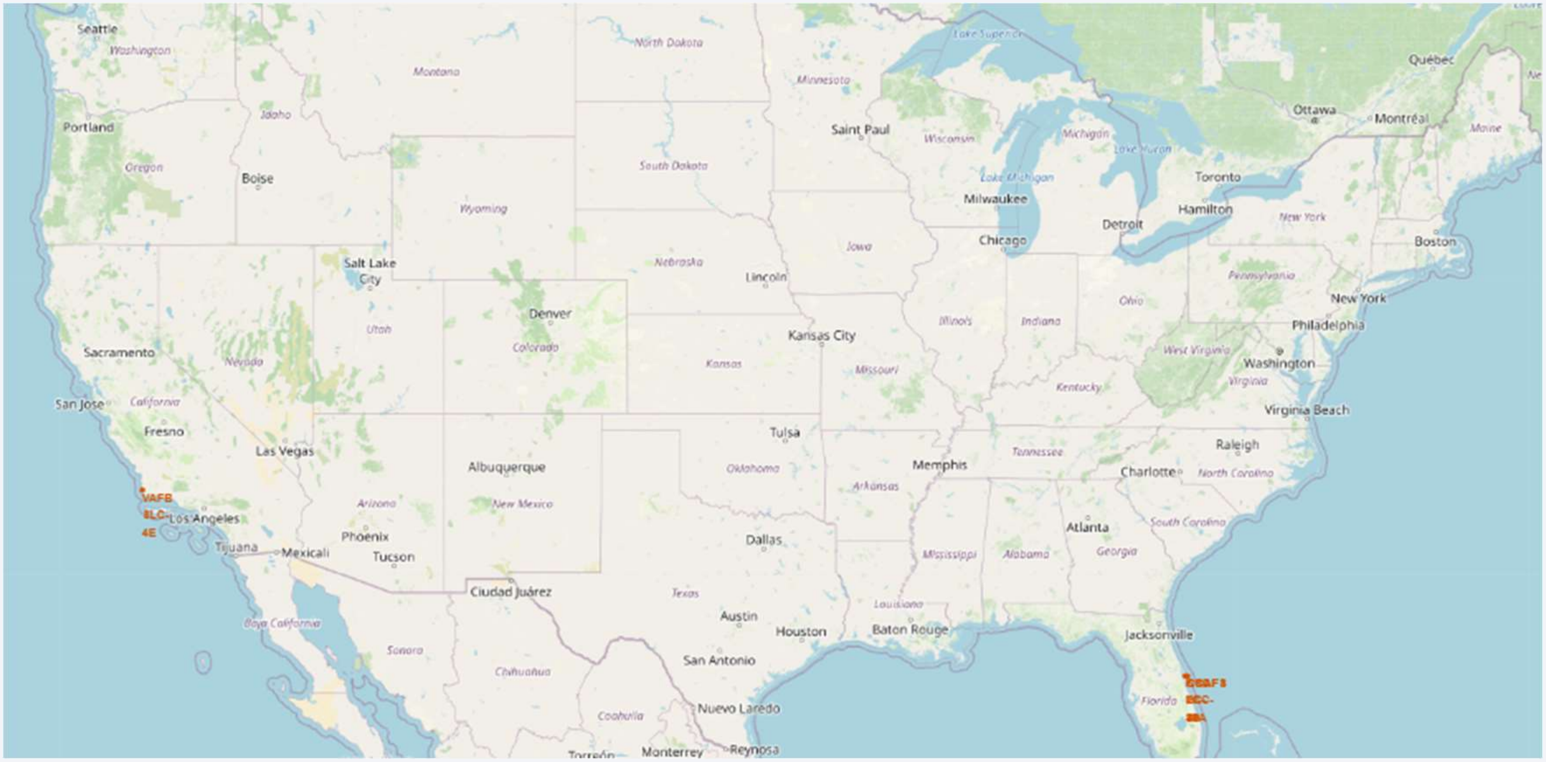


Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

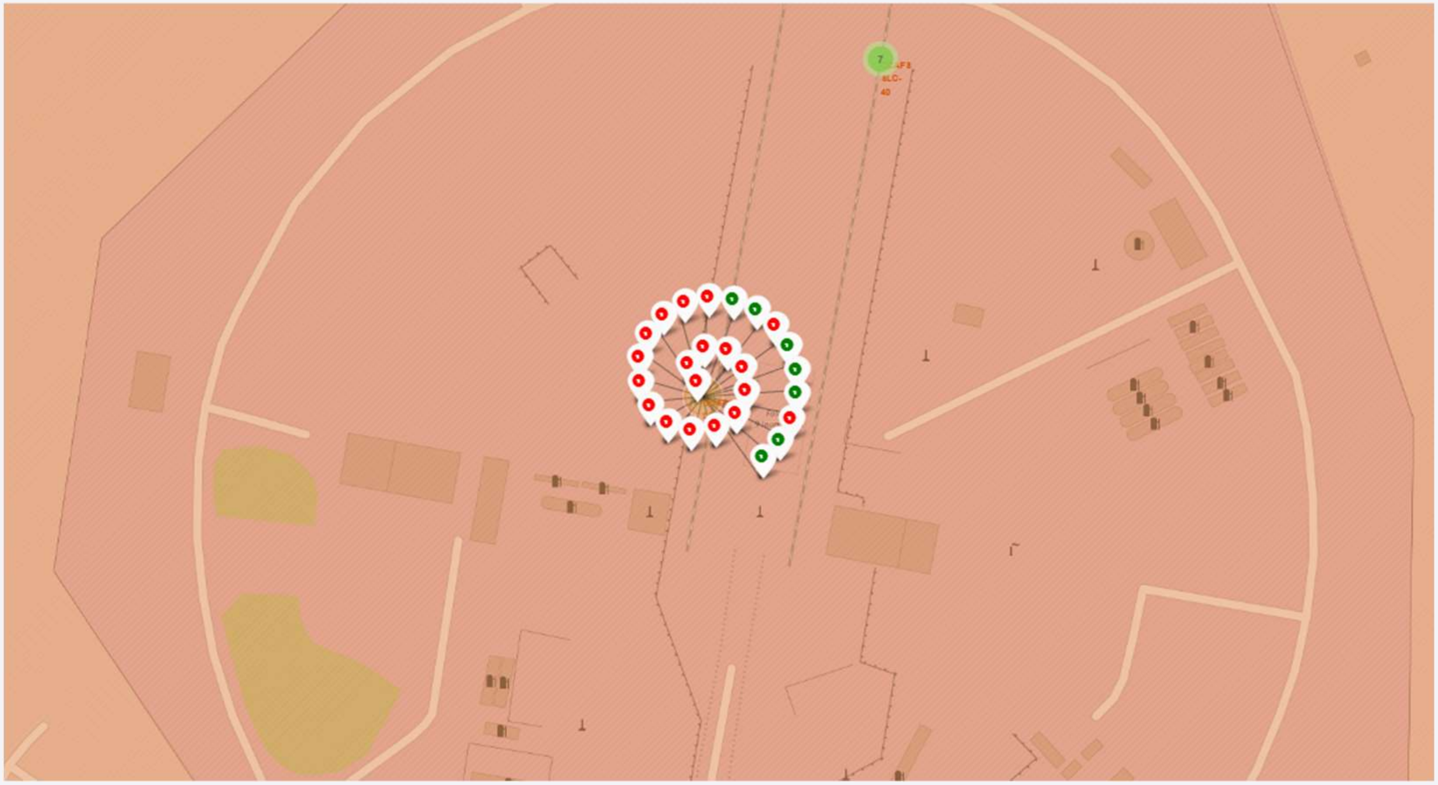
Launch Sites Proximities Analysis

Location of each Launch Site in the U.S.



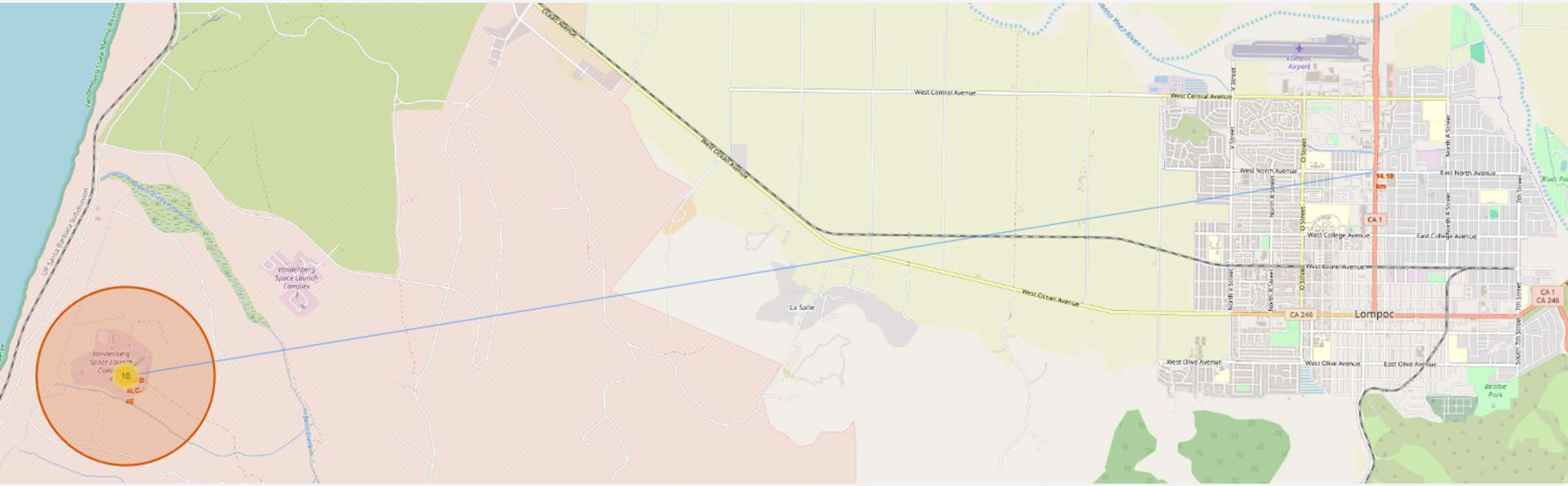
Here we can see that most of the launch sites in Florida, and the other launch site being in California.

Launch Results for Cape Canaveral SLC-40

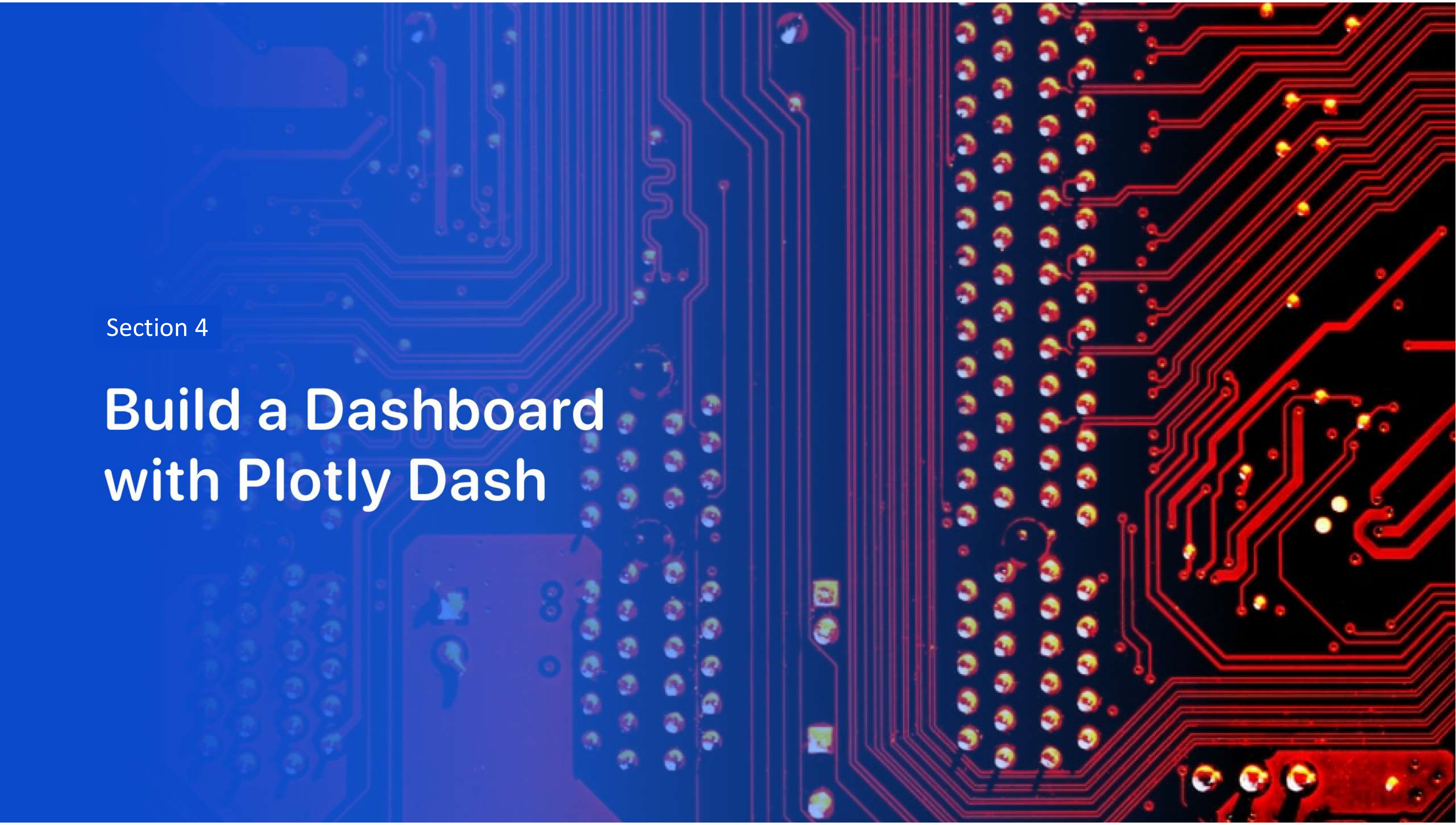


Most of the launch results for CCAFS SLC-40 indicate a failure, with only 7 having been successful.

Closest City to the Vandenberg SLC-4E



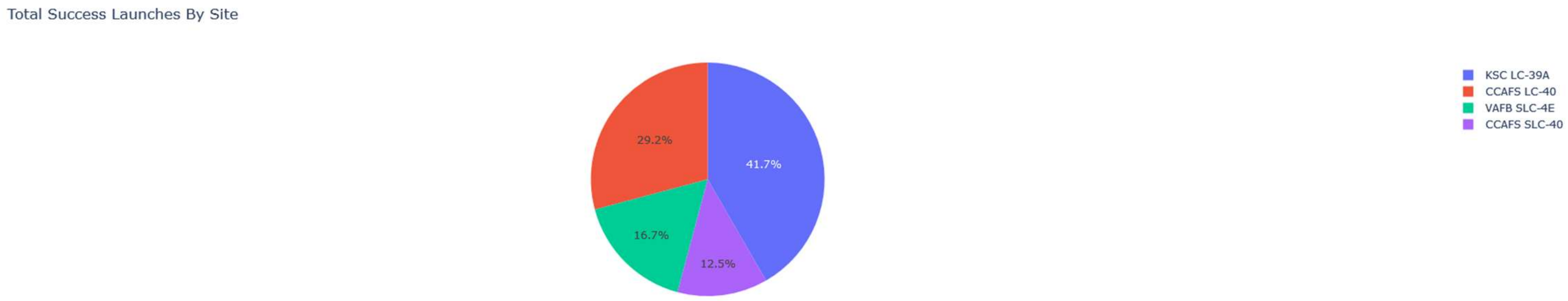
The closest city to VAFB SLC-4E is approximately 14 kilometers away, showing that the launch site is built in a remote environment (in this case, a marine reserve)



Section 4

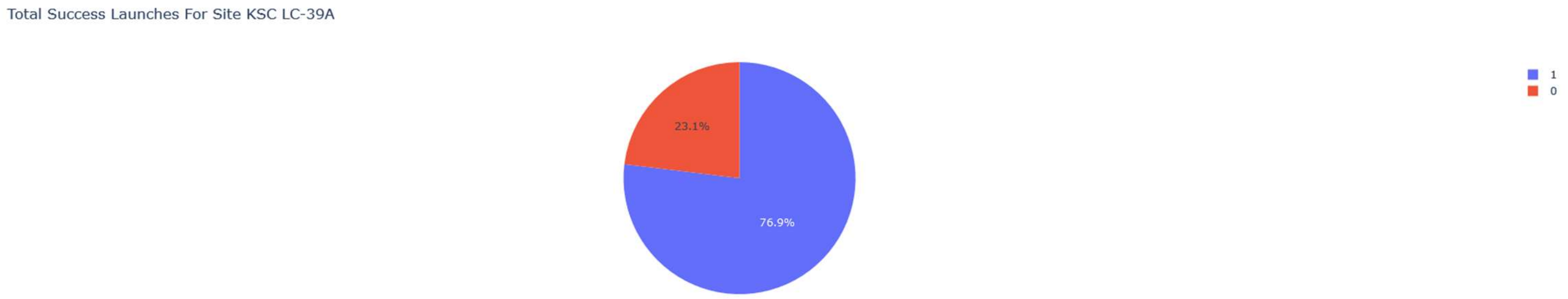
Build a Dashboard with Plotly Dash

Total Success of Launches for All Sites



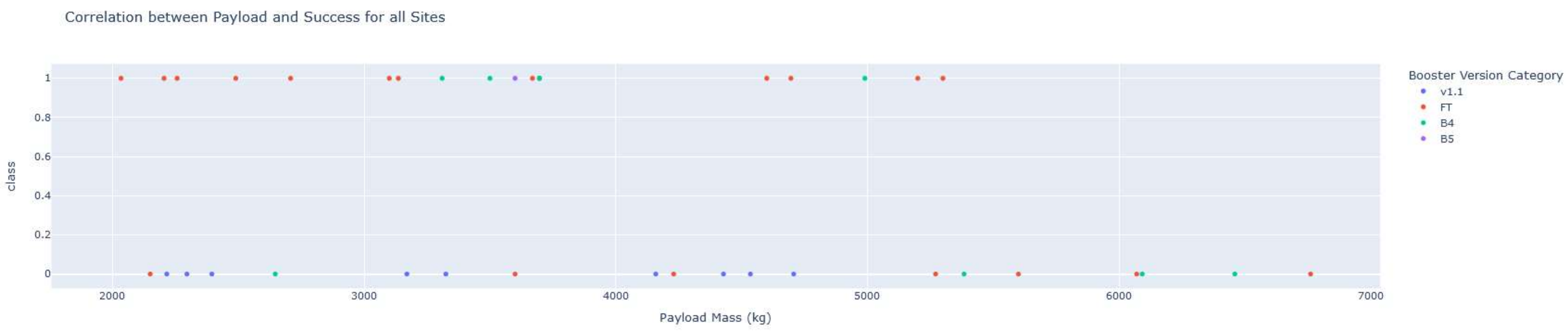
The Kennedy Space Center Launch Complex-39A has the **highest success rate** out of all launches at 41.7%, whereas the Cape Canaveral Space Launch Complex 40 has the **lowest** (12.5%).

Launch Site with the highest Success Rate



The Kennedy Space Center Launch Complex-39A has the highest success ratio at 76.9%, compared to a 23.1% chance of failure.

Payload vs Launch Outcome for all Sites



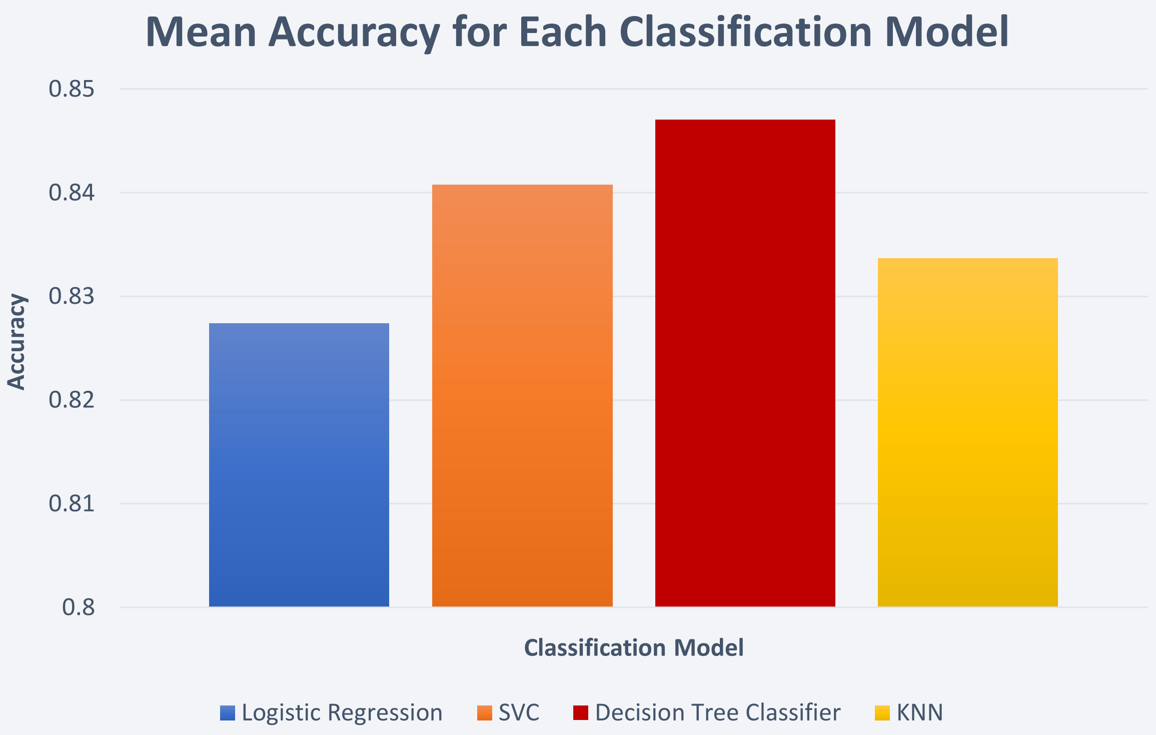
The payload mass which has the highest success rate is between 2000 and around 3700kg, with the most successful booster version being the Falcon 9 Full Thrust. The least successful booster version is v1.1.

Section 5

Predictive Analysis (Classification)

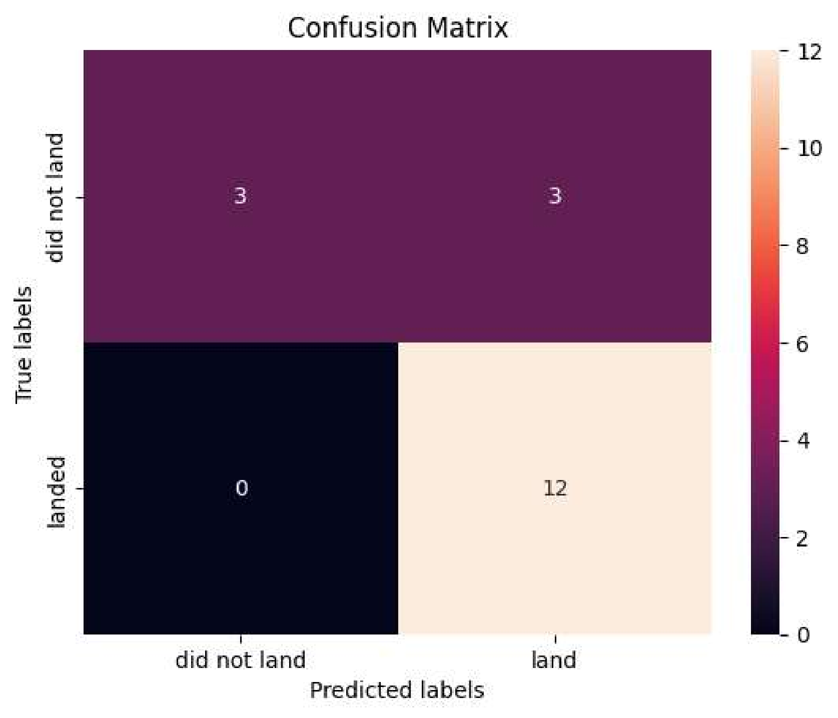
Classification Accuracy

The best performing classification model for predictive analysis on this dataset is the Decision Tree Classifier at an accuracy rate of around 0.845%.



Confusion Matrix of Decision Tree Classifier

- The confusion matrix of the Decision Tree Classifier shows that out of 18 records, 15 were labelled correctly at an accuracy of 0.83%. This is not anything special, since most classification models were able to achieve the same test accuracy.
- That being said, the decision tree classifier had the highest training accuracy around 0.86% with the optimum hyperparameters.



Conclusions

- Going back to the questions that were first discussed in the presentation, we now know that:
  - The best predictive model to determine future outcomes for Falcon 9 landings is the Decision Tree Classifier.
  - The Falcon 9 Full-Thrust booster version has the highest success rate.
  - The launch site with the highest success rate for the Falcon 9 is the KSC LC-39A (Kennedy Space Center) at 76.9%
  - The launch site with the highest failure rate for the Falcon 9 is the CCAFS SLC-40 (Cape Canaveral) is 42.9%
  - The payload range with the highest success rate for the Falcon 9 is around 2000-4000kg.
  - The landing outcome with the highest success rate is a success by drone ship.

Appendix

- For the Decision Tree Classifier, the most optimal hyperparameters used in this report are criterion set to entropy, max_depth set to a value of 10, max_features set to sqrt, min_samples_leaf set to a value of 1, min_samples_split set to 10, and splitter set to random.

Thank you!