IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Hristo Darlyanov

31/03/2023

# Outline

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

# Executive Summary

- Summary of methodologies

- Data Collection through API

- Data Collection with Web Scraping

- Data Wrangling

- Exploratory Data Analysis with SQL

- Exploratory Data Analysis with Data Visualization

- Interactive Visual Analytics with Folium

- Machine Learning Prediction

- Summary of all results

- Exploratory Data Analysis result

- Interactive analytics in screenshots

- Predictive Analytics result

# Introduction

The goal is to assess whether Space Y, a new company, can successfully compete with established player Space X in the space launch market. To do so, we need to find answers to key questions, such as:

-How can we accurately estimate the total cost for launches, taking into account the likelihood of successful landings of the first stage of rockets?
-What is the optimal location for Space Y to conduct their launches in order to maximize their chances of success?

By answering these questions, we can gain a better understanding of the potential viability of Space Y in the space launch market and identify key strategies to help the company succeed.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data collection through SpaceX API
  - Data collection through Wikipedia page web scraping
- Perform data wrangling
  - Encoding by applying one-hot to categorical features
- 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 sets were collected from Space X API: https://api.spacexdata.com/v4/rockets/

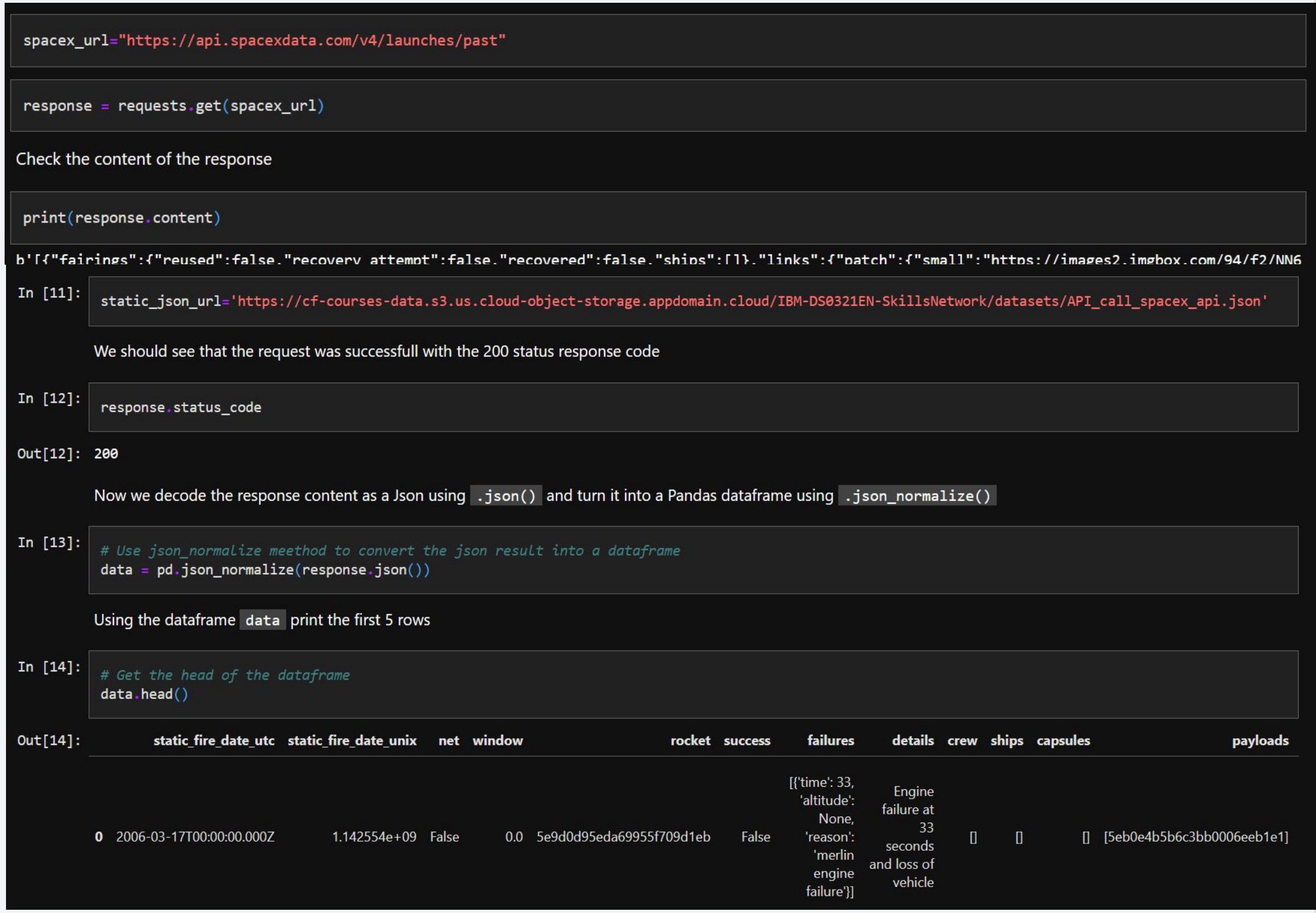
and from Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches, using web scraping

# Data Collection – SpaceX API

- Data was collected through HTTP het requests using the requests python library. After data collection, some data wrangling was applied to the data.

- Full process:

-https://github.com/hristo-darlyanov/SpaceY/blob/master/jupyter-labs-spacex-data-collection-api.ipynb



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Check the content of the response

```
print(response.content)
```

```
b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/94/f2/NN6
```

```
In [11]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
In [12]: response.status_code

Out[12]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
In [13]: # Use json_normalize meethod to convert the json result into a dataframe
         data = pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

```
In [14]: # Get the head of the dataframe
         data.head()
```

Out[14]:

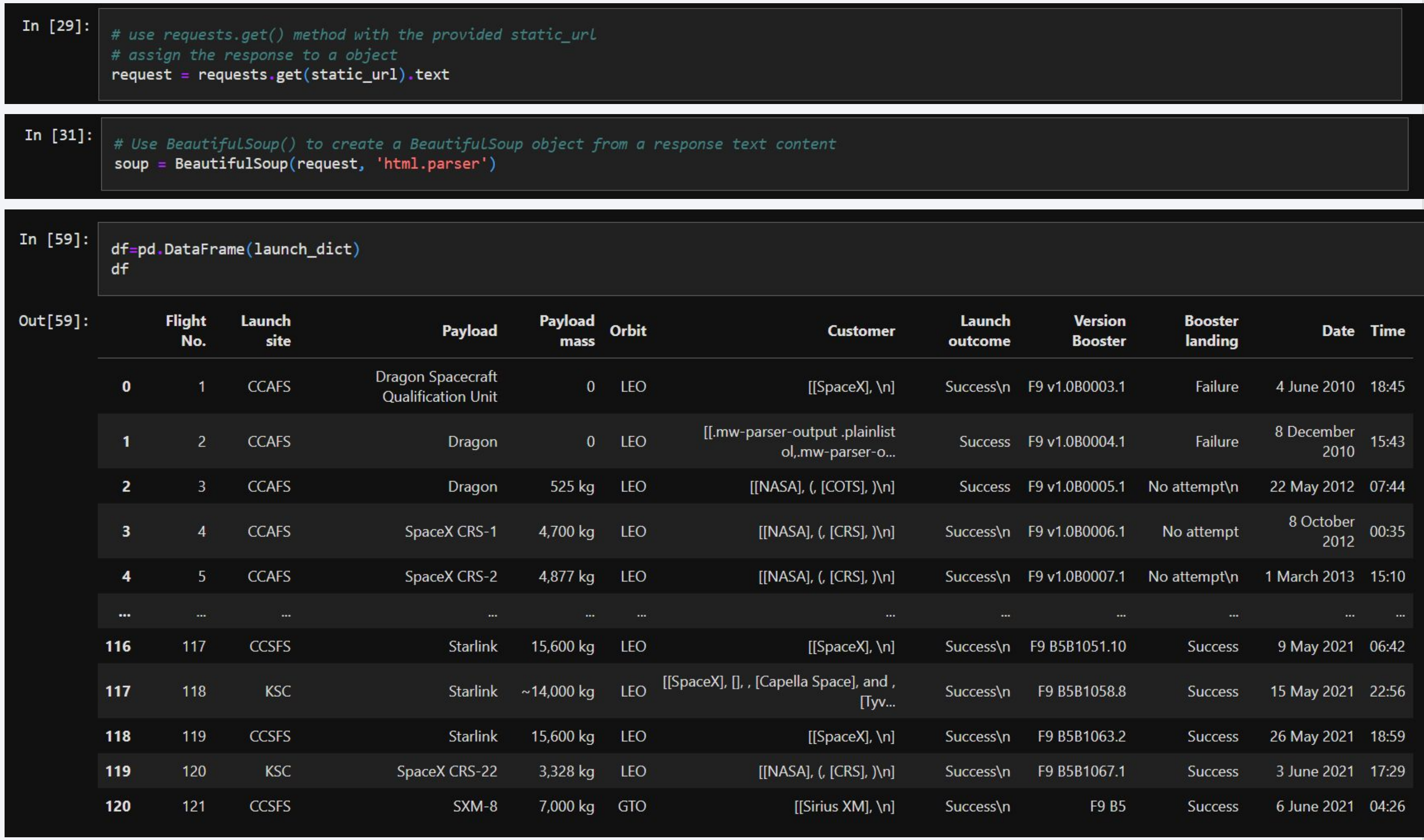
| | static_fire_date_utc | static_fire_date_unix | net | window | rocket | success | failures | details | crew | ships | capsules | payloads |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2006-03-17T00:00:00.000Z | 1.142554e+09 | False | 0.0 | 5e9d0d95eda69955f709d1eb | False | [{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}] | Engine failure at 33 seconds and loss of vehicle | [] | [] | [] | [5eb0e4b5b6c3bb0006eeb1e1] |

# Data Collection - Scraping

- Web scraping was done with the BeautifulSoup library in python. After data collection the data was parsed into a data frame.

- Full process here:

https://github.com/hristo-darlyanov/SpaceY/blob/master/jupyter-labs-webscraping.ipynb



```
In [29]: # use requests.get() method with the provided static_url
         # assign the response to a object
         request = requests.get(static_url).text
```

```
In [31]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
         soup = BeautifulSoup(request, 'html.parser')
```

```
In [59]: df=pd.DataFrame(launch_dict)
         df
```

Out[59]:

| | Flight No. | Launch site | Payload | Payload mass | Orbit | Customer | Launch outcome | Version Booster | Booster landing | Date | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | CCAFS | Dragon Spacecraft Qualification Unit | 0 | LEO | [[SpaceX], \n] | Success\n | F9 v1.0B0003.1 | Failure | 4 June 2010 | 18:45 |
| **1** | 2 | CCAFS | Dragon | 0 | LEO | [[.mw-parser-output .plainlist ol,.mw-parser-o... | Success | F9 v1.0B0004.1 | Failure | 8 December 2010 | 15:43 |
| **2** | 3 | CCAFS | Dragon | 525 kg | LEO | [[NASA], (, [COTS], )\n] | Success | F9 v1.0B0005.1 | No attempt\n | 22 May 2012 | 07:44 |
| **3** | 4 | CCAFS | SpaceX CRS-1 | 4,700 kg | LEO | [[NASA], (, [CRS], )\n] | Success\n | F9 v1.0B0006.1 | No attempt | 8 October 2012 | 00:35 |
| **4** | 5 | CCAFS | SpaceX CRS-2 | 4,877 kg | LEO | [[NASA], (, [CRS], )\n] | Success\n | F9 v1.0B0007.1 | No attempt\n | 1 March 2013 | 15:10 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **116** | 117 | CCSFS | Starlink | 15,600 kg | LEO | [[SpaceX], \n] | Success\n | F9 B5B1051.10 | Success | 9 May 2021 | 06:42 |
| **117** | 118 | KSC | Starlink | ~14,000 kg | LEO | [[SpaceX], [], , [Capella Space], and , [Tyv... | Success\n | F9 B5B1058.8 | Success | 15 May 2021 | 22:56 |
| **118** | 119 | CCSFS | Starlink | 15,600 kg | LEO | [[SpaceX], \n] | Success\n | F9 B5B1063.2 | Success | 26 May 2021 | 18:59 |
| **119** | 120 | KSC | SpaceX CRS-22 | 3,328 kg | LEO | [[NASA], (, [CRS], )\n] | Success\n | F9 B5B1067.1 | Success | 3 June 2021 | 17:29 |
| **120** | 121 | CCSFS | SXM-8 | 7,000 kg | GTO | [[Sirius XM], \n] | Success\n | F9 B5 | Success | 6 June 2021 | 04:26 |

# Data Wrangling

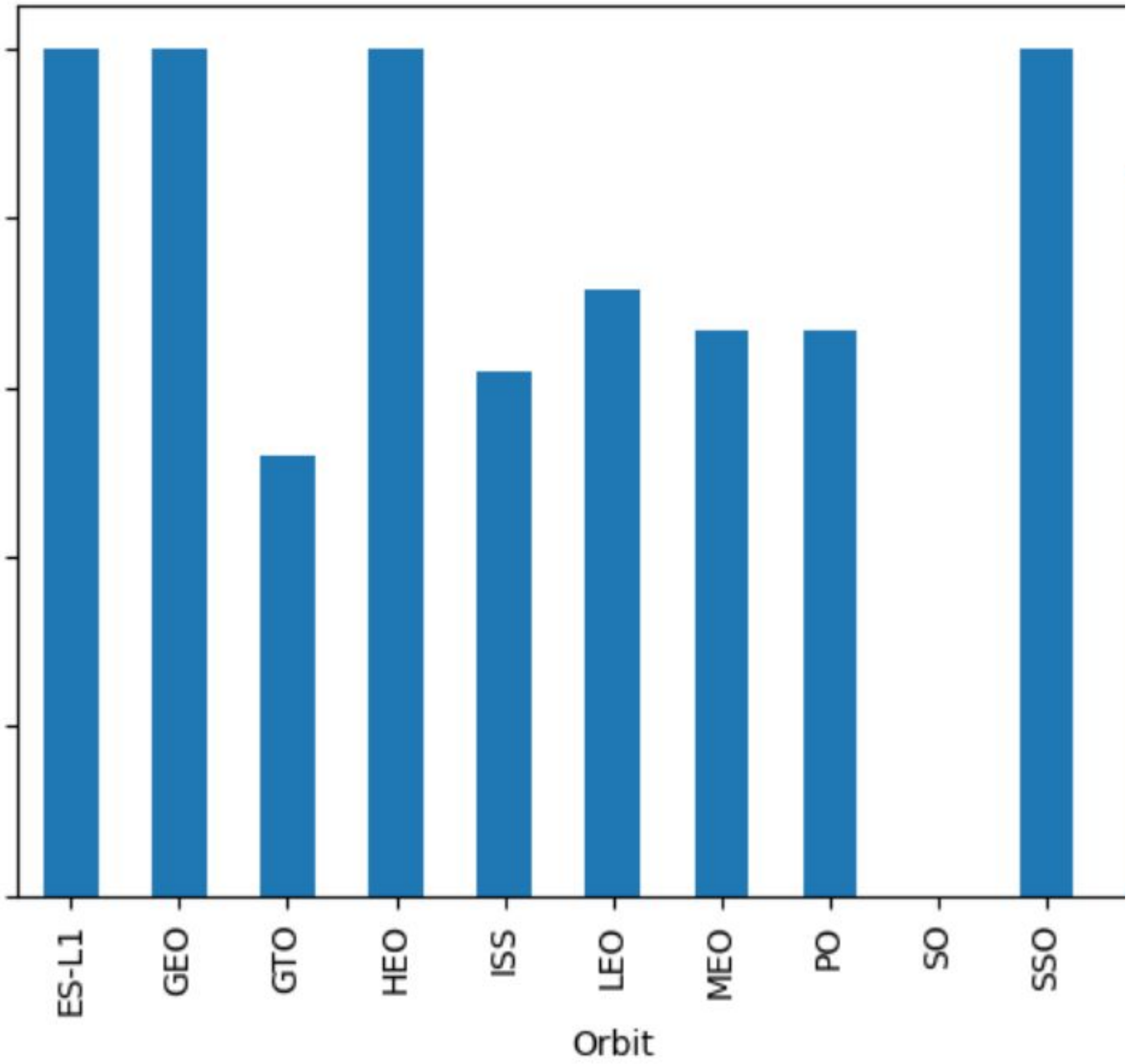
- Performing exploratory data analysis and determined the training labels.
- Calculating the number of launches at each site, and the number and occurrence of each orbits
- Creating landing outcome label from outcome column and exported the results to csv.

Full process here:

https://github.com/hristo-darlyanov/SpaceY/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

# EDA with Data Visualization

- Exploring the data by visualizing the relationship between:
- flight number and launch Site
- Payload and launch site
- Success rate of each orbit type
- Flight number and orbit type
- The launch success yearly trend.
- Full process here: https://github.com/hristo-darlyanov/SpaceY/blob/master/jupyter-labs-eda-dataviz.ipynb



# EDA with SQL

The performed SQL queries:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA’
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was acheived.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranked 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
- Full process here : https://github.com/hristo-darlyanov/SpaceY/blob/master/EDA%20SQL.ipynb

# Build an Interactive Map with Folium

- By using folium maps I plotted launch sites and tracked their outcomes. I marked each launch site on the map and added objects like markers, circles, and lines to indicate whether each launch was a success or a failure. You also assigned a class to each launch outcome, with 0 representing failure and 1 representing success.
- Using these markers and color-labeled clusters, I was able to identify which launch sites have a higher success rate than others. Additionally, I calculated the distances between each launch site and its surrounding area to answer questions like whether the site is located near railways, highways, or coastlines, and whether there is a certain distance between the launch site and nearby cities.
- Full process here: https://github.com/hristo-darlyanov/SpaceY/blob/master/lab_jupyter_launch_site_location.ipynb

# Build a Dashboard with Plotly Dash

- Within the dashboard, I plotted pie charts that showed the total number of launches for specific launch sites. This allowed me to see which sites had the most launches overall.
- In addition to the pie charts, I also created scatter graphs that showed the relationship between launch outcome and payload mass (measured in kg) for different booster versions. This allowed me to examine whether there was a correlation between these two variables and how it differed across different types of boosters.
- Full process here: https://github.com/hristo-darlyanov/SpaceY/tree/master/Dash

# Predictive Analysis (Classification)

-I started by loading the data for my machine learning project using numpy and pandas, transformed the data as needed and split it into training and testing sets.

-Next, I built several different machine learning models and used GridSearchCV to tune their hyperparameters. This allowed me to find the best settings for each model and optimize their performance.

-In order to evaluate the performance of each model, I used accuracy as the metric. Based on this metric, I improved my models through feature engineering and further algorithm tuning.

-Ultimately, I was able to identify the best performing classification model.

Full process here:
https://github.com/hristo-darlyanov/SpaceY/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

# Results

After performing exploratory data analysis on the Space X launch data, we discovered a number of interesting findings. First, we observed that Space X has used four different launch sites for their missions. Additionally, we found that the company's initial launches were directed towards Space X itself and NASA.

We also calculated that the average payload of the F9 v1.1 booster was 2,928 kg. It took five years after the first launch for the first successful landing outcome to occur, and we noted that many Falcon 9 booster versions successfully landed in drone ships with payloads above the average.

Interestingly, nearly 100% of mission outcomes were successful, with only two booster versions failing to land in drone ships in 2015 (F9 v1.1 B1012 and F9 v1.1 B1015). However, we observed that the number of successful landing outcomes improved over time.

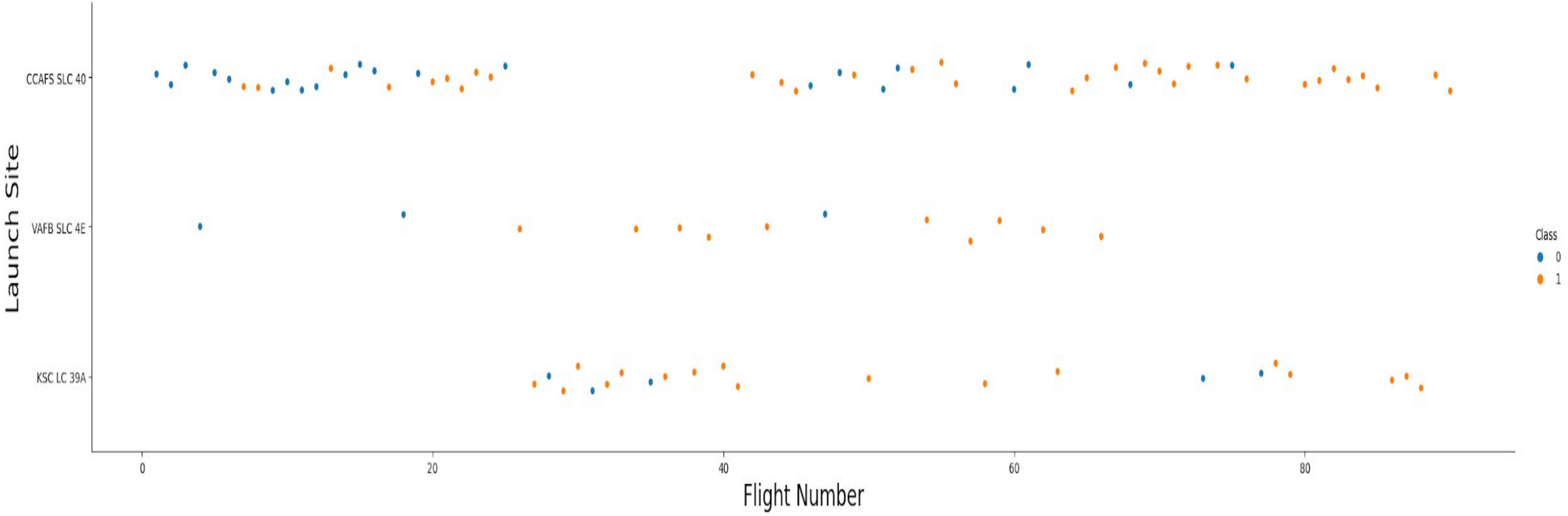
Based on Predictive Analysis, it has been determined that the Decision Tree Classifier is the most effective model for predicting successful landings. This model has demonstrated an accuracy rate of over 87%, and has also been able to accurately predict successful landings for test data with an accuracy rate exceeding 94%. This suggests that the Decision Tree Classifier has the potential to be a highly reliable tool for predicting the outcome of landings, and may be useful in a variety of contexts where successful landings are critical.

Section 2

# Insights drawn from EDA

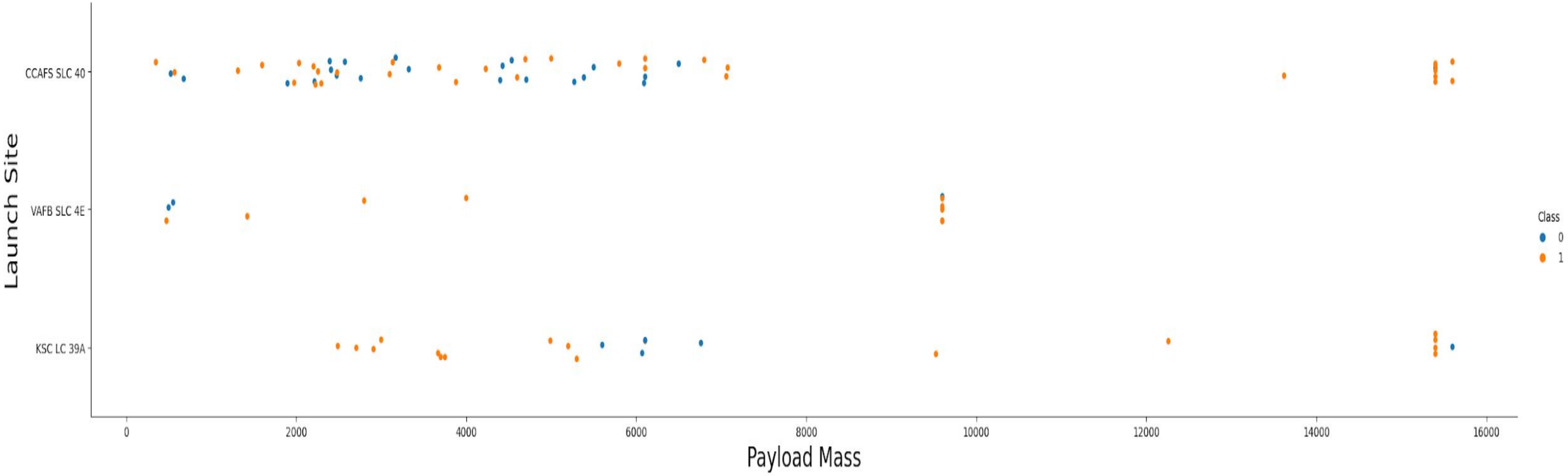
# Flight Number vs. Launch Site

- Based on the given plot, it is evident that CCAF5 SLC 40 is currently the best launch site, as it has recorded the highest number of successful launches. The second-best launch site is VAFB SLC 4E, followed by KSC LC 39A in third place. Moreover, the plot also shows that the overall success rate of launches has improved over time.



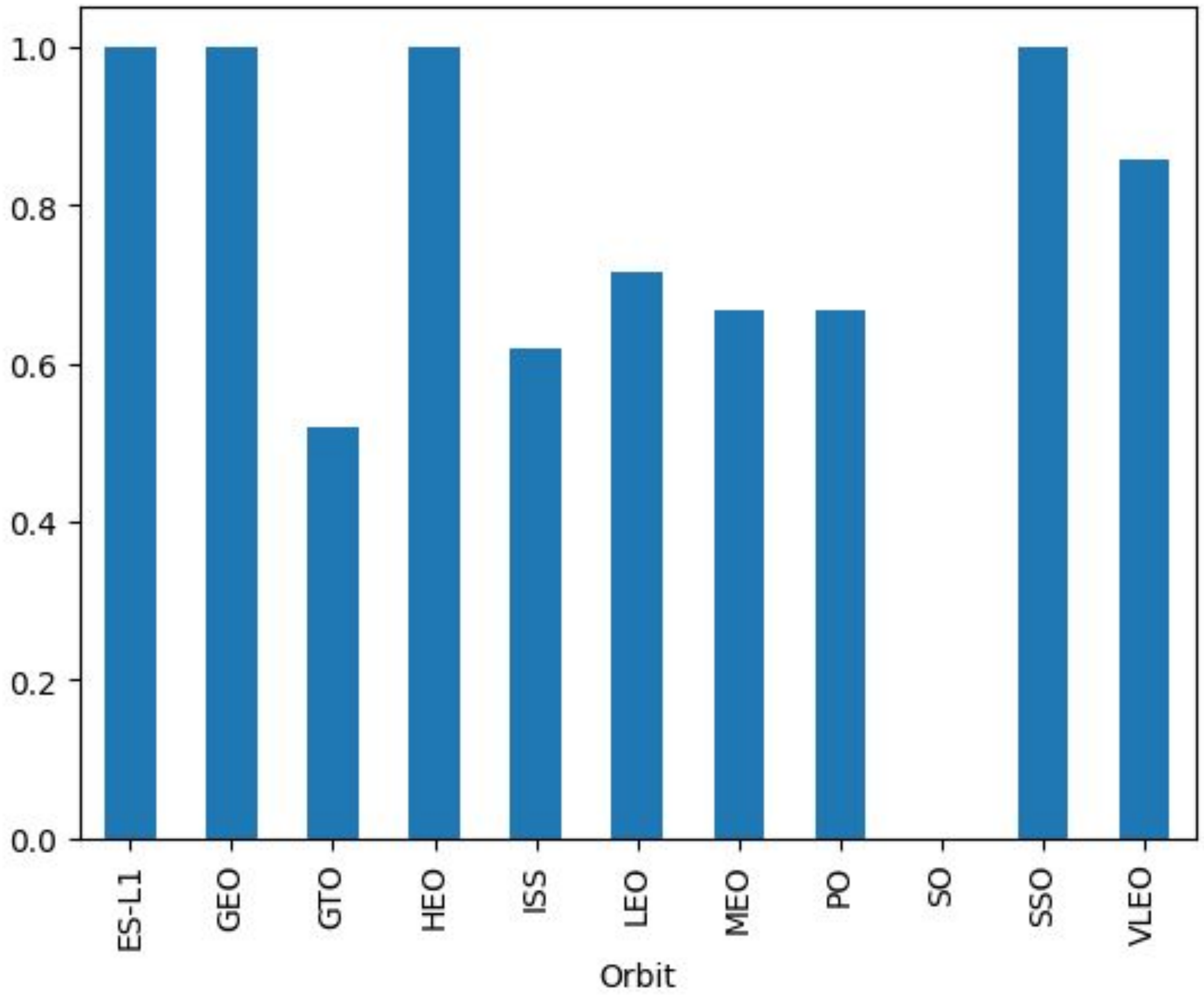
# Payload vs. Launch Site

- According to the data, payloads weighing over 9,000kg have a high success rate. Additionally, it appears that launch sites such as CCAFS SLC 40 and KSC LC 39A are capable of handling payloads over 12,000kg successfully.



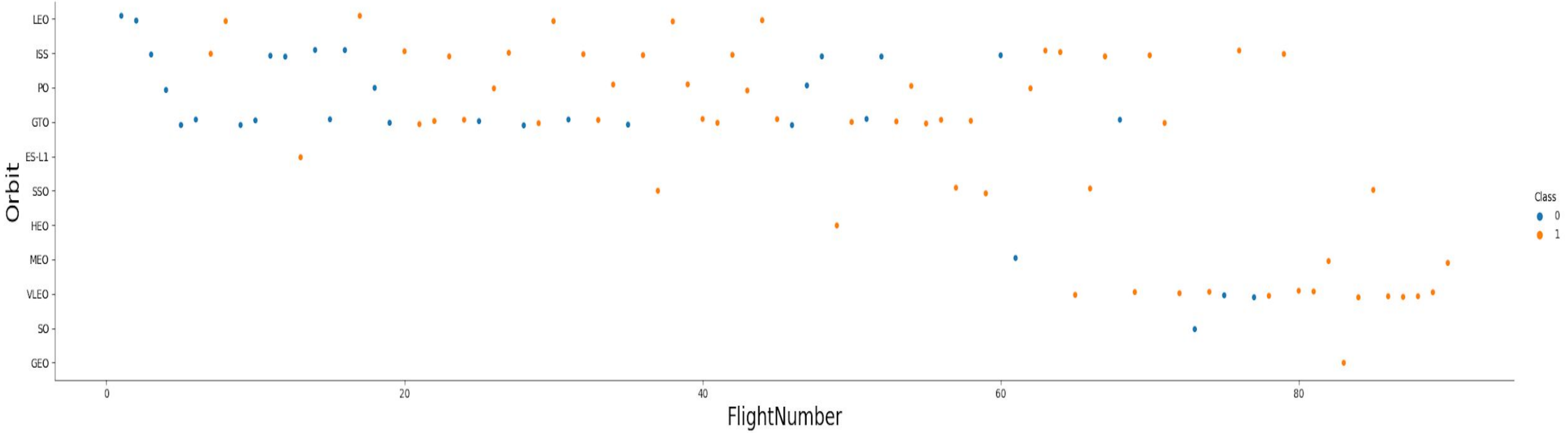
# Success Rate vs. Orbit Type

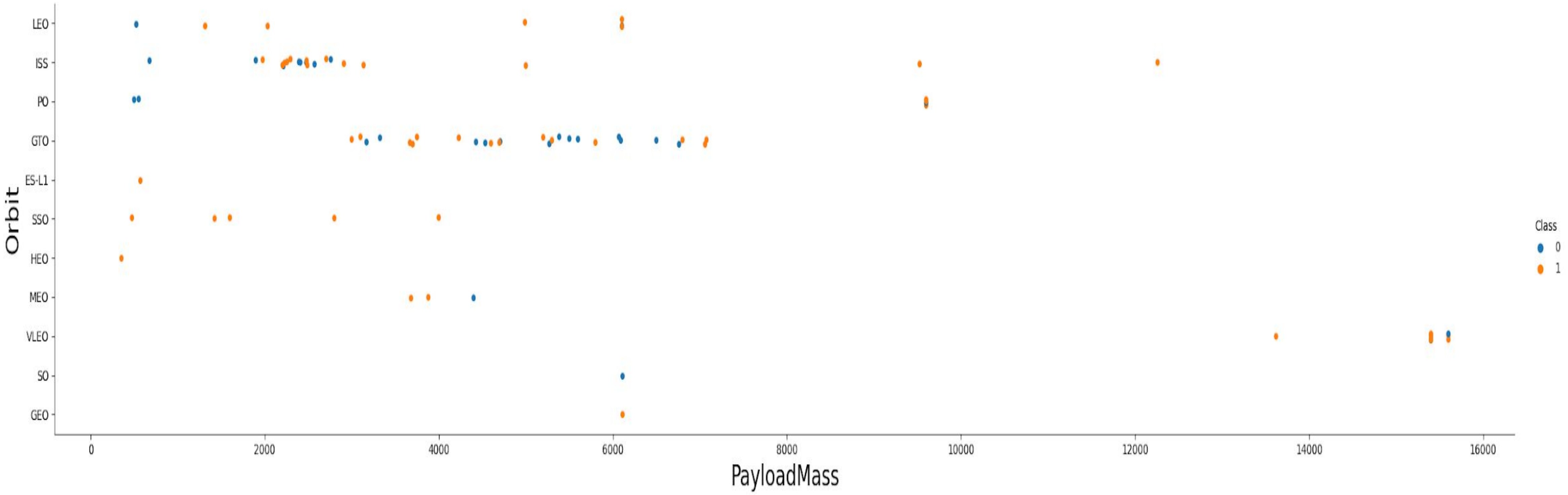
- The plot reveals that the highest success rates are achieved for launches into the following orbits: ES-L1, GEO, HEO, and SSO. Additionally, the success rates for VLEO (Very Low Earth Orbit) are above 80%, and for LFO (Low Earth Orbit) they are above 70%.



# Flight Number vs. Orbit Type

- The data indicates that over time, the success rate of launches into all types of orbits has improved. Additionally, the frequency of launches into VLEO orbit has increased, which suggests that it may be a promising new business opportunity.



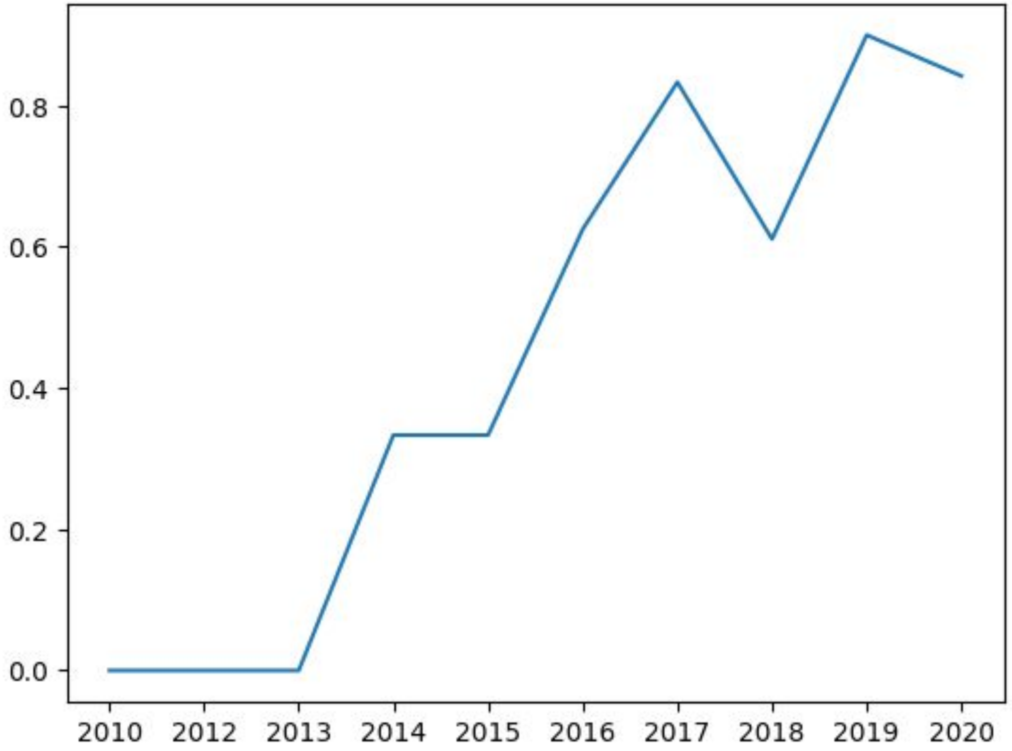


# Payload vs. Orbit Type

The plot suggests that there is no significant correlation between the payload and success rate for launches into the GTO orbit. The orbit of the International Space Station (ISS) has a wide range of payloads and a good success rate. However, there are relatively few launches into the SO and GEO orbits.

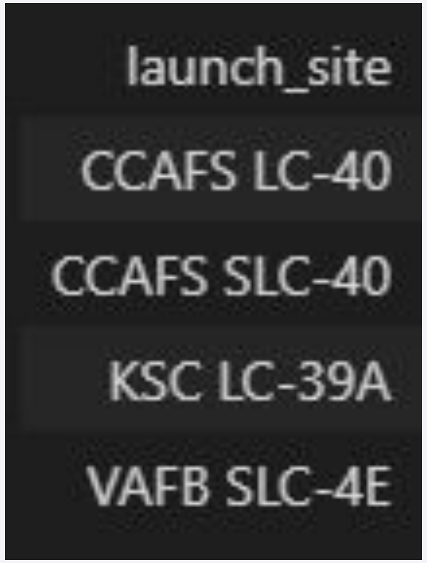
# Launch Success Yearly Trend

- Based on the data, the success rate of launches began to increase in 2013 and continued to improve until 2020. It appears that the first three years were a period of adjustments and technology improvements to achieve the higher success rates in the following years.

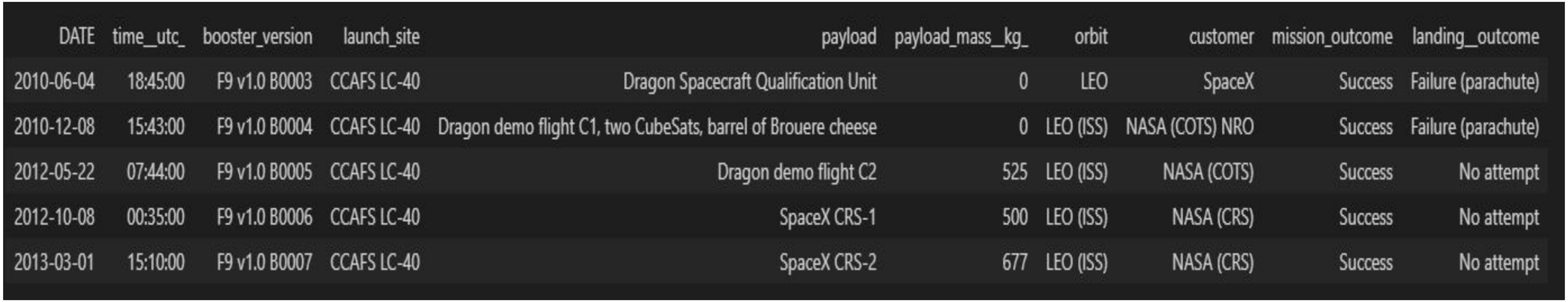


# All Launch Site Names

- The data indicates that there are four unique launch sites, which have been identified by selecting unique values of "launch_site" from the dataset.



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



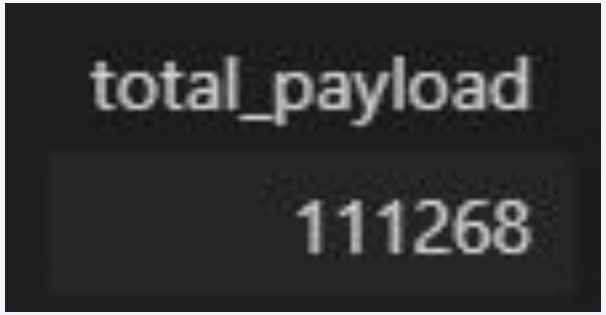
| DATE | time_utc_ | booster_version | launch_site | payload | payload_mass_kg_ | orbit | customer | mission_outcome | landing_outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | 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 |

# Launch Site Names Begin with 'CCA'

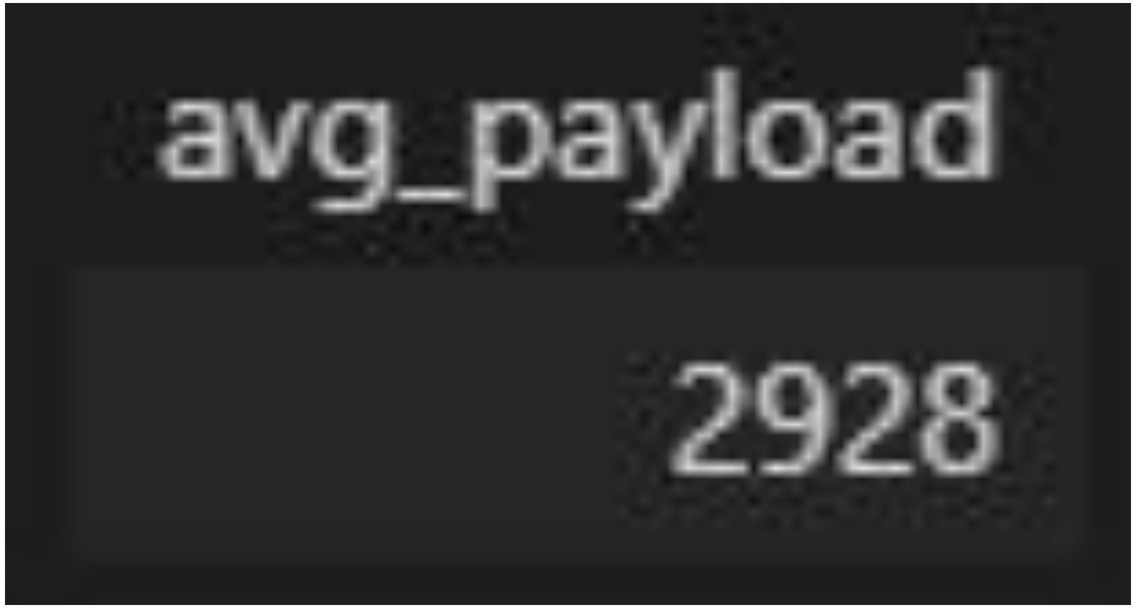
- From the dataset, there are five records where the launch site names begin with "CCA". These launches took place at Cape Canaveral, Florida.

# Total Payload Mass

The dataset provides information on the total payload carried by boosters from NASA. Additionally, the total payload carried by NASA can be calculated by summing all payloads whose codes contain "CRS", which is a program code used by NASA.



| total_payload |
|---|
| 111268 |



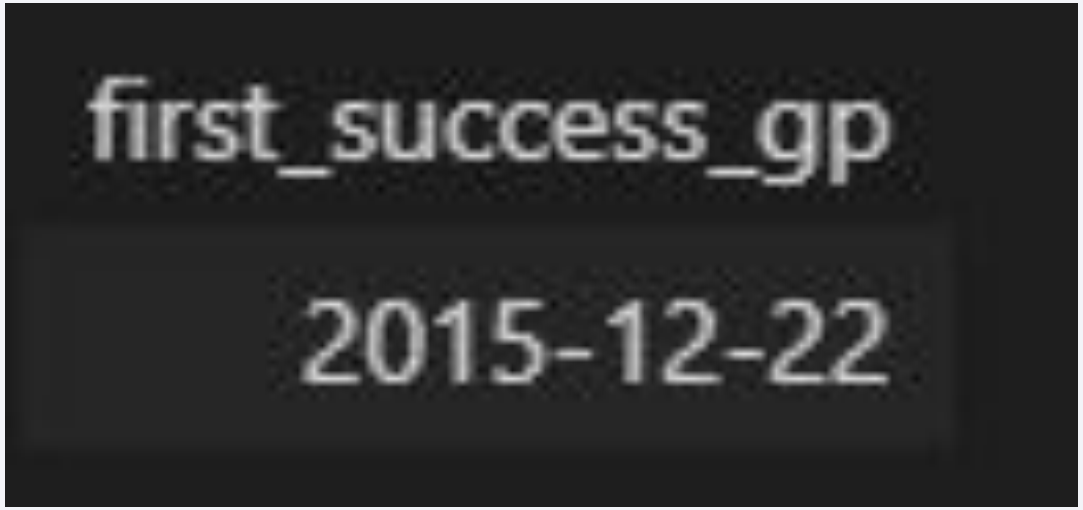
| avg_payload |
|---:|
| 2928 |

# Average Payload Mass by F9 v1.1

- The dataset allows for the calculation of the average payload mass carried by the F9 v1.1 booster version. By filtering the data to only include launches using this booster version and then calculating the average payload mass, it was found to be 2,928 kg.

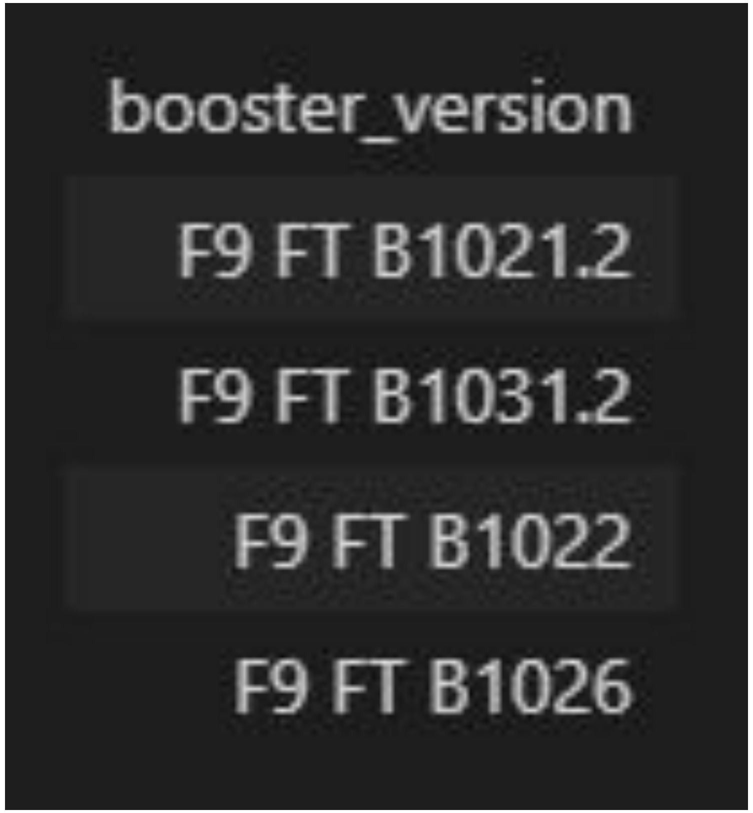
# First Successful Ground Landing Date

- The dataset can be used to determine the first successful landing outcome of a booster on a ground pad. By filtering the data to only include successful landings on a ground pad and obtaining the minimum date value, it was found that the first occurrence happened on December 22, 2015.



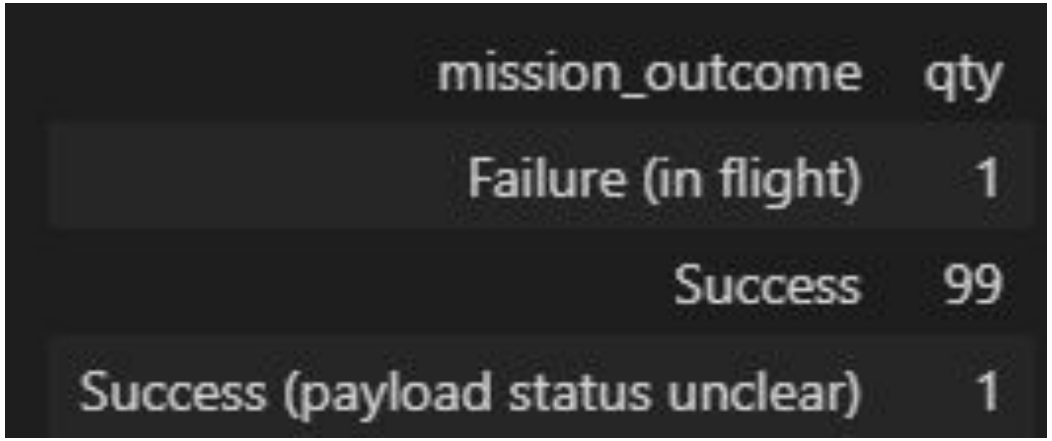
| first_success_gp |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000



| booster_version |
|---|
| F9 FT B1021.2 |
| F9 FT B1031.2 |
| F9 FT B1022 |
| F9 FT B1026 |

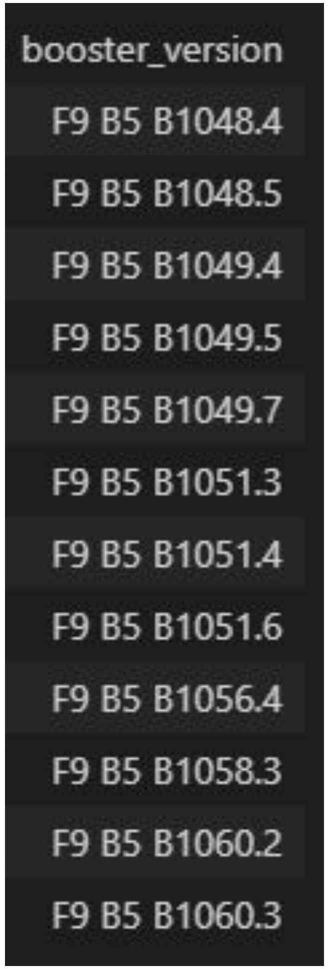
- Using the dataset, it is possible to identify booster versions that have successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000. By selecting distinct booster versions that meet these criteria, the resulting list contains four different booster versions.



| mission_outcome | qty |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Total Number of Successful and Failure Mission Outcomes

- By grouping the mission outcomes and counting the number of records in each group, it is possible to obtain a summary of the number of successful and failed mission outcomes from the dataset.

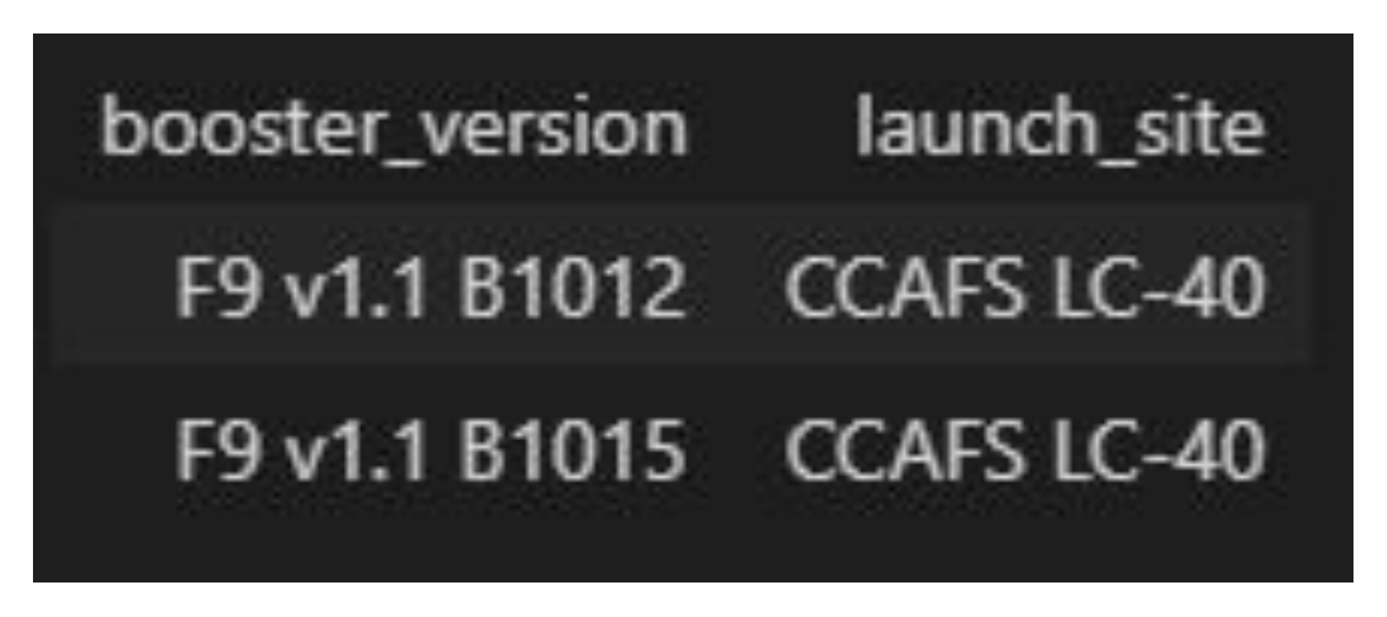


| booster_version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1048.5 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1049.7 |
| F9 B5 B1051.3 |
| F9 B5 B1051.4 |
| F9 B5 B1051.6 |
| F9 B5 B1056.4 |
| F9 B5 B1058.3 |
| F9 B5 B1060.2 |
| F9 B5 B1060.3 |

# Boosters Carried Maximum Payload

- The dataset provides information on the boosters that have carried the maximum payload mass recorded. These are the boosters that have been able to lift the heaviest payloads according to the available data.

# 2015 Launch Records

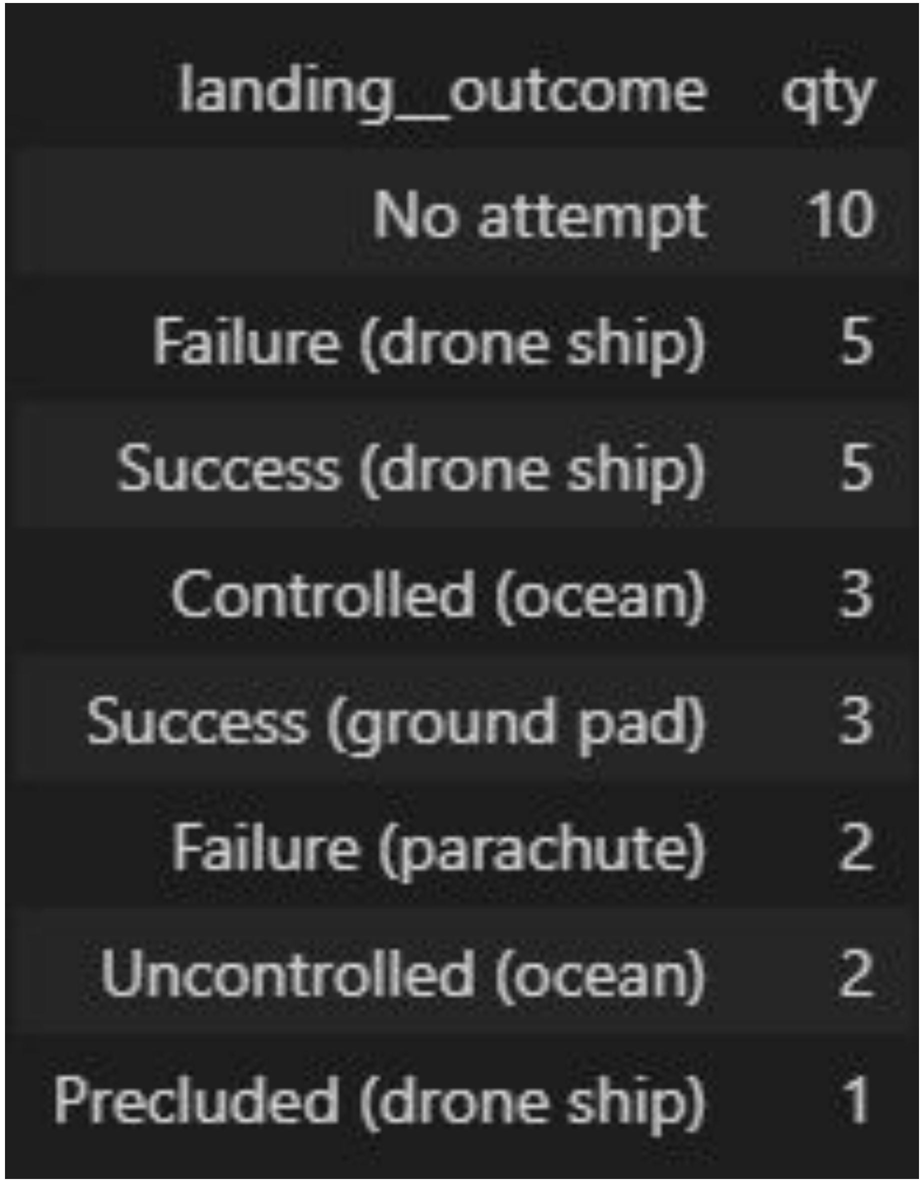


| booster_version | launch_site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

- The dataset contains information on failed landing outcomes on drone ships, as well as the corresponding booster versions and launch site names for the year 2015. The list provided indicates that there were only two occurrences of this type in that year.

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

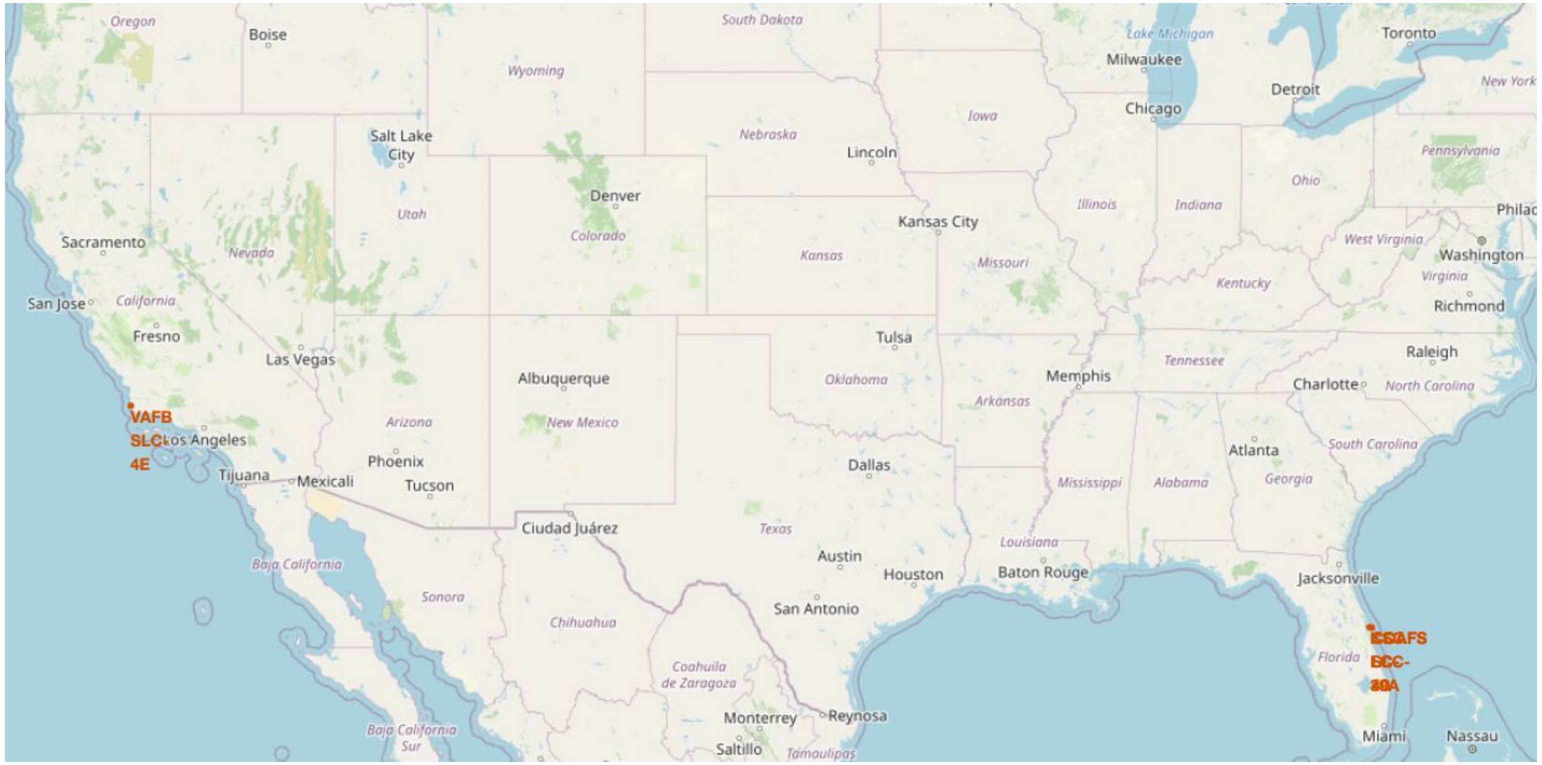
- The dataset allows for a ranking of all landing outcomes that occurred between the dates of June 4, 2010 and March 20, 2017. However, it is important to note that the data indicates the presence of "No attempt" landing outcomes, which should also be taken into account in the analysis.



| landing__outcome | qty |
|---|---|
| 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

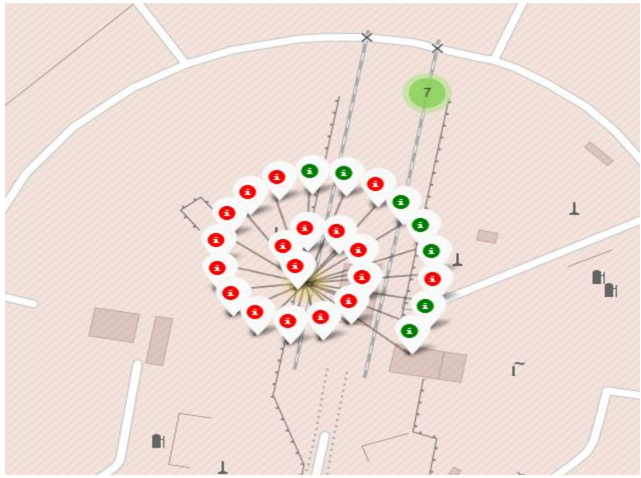


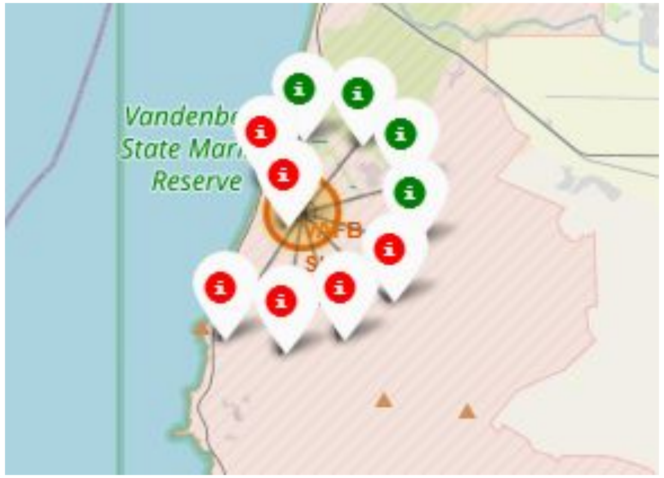
# Launch Sites

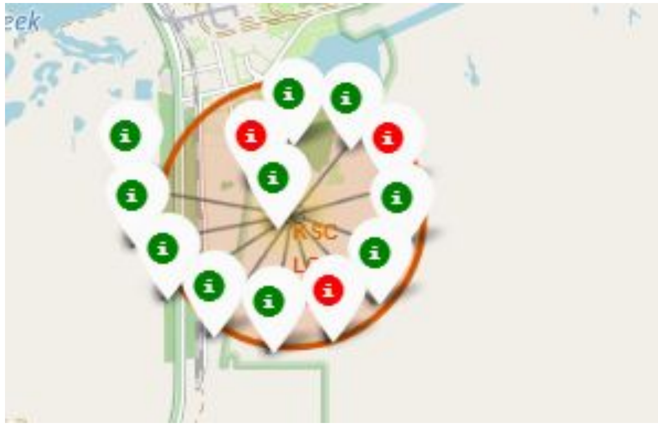
- The launch sites in the dataset are located near the sea, most likely for safety reasons, but they are also relatively close to roads and railroads for logistical purposes.

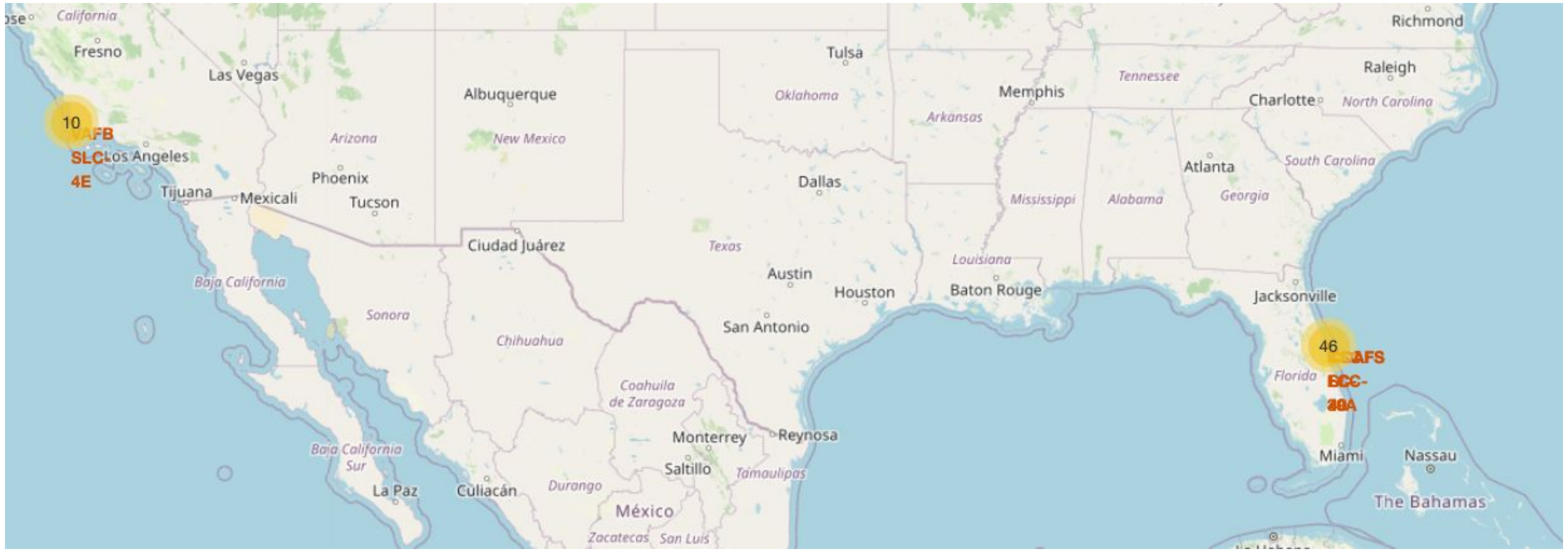
# Site launch outcome

- Green markers indicate successful and red ones indicate failure.



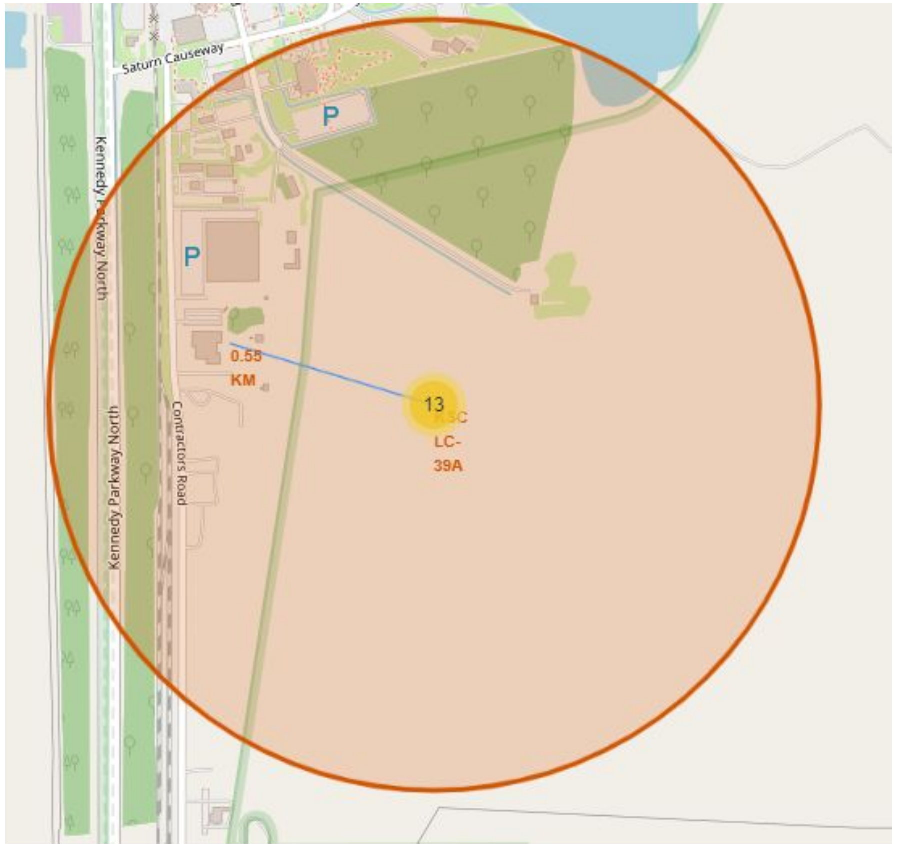






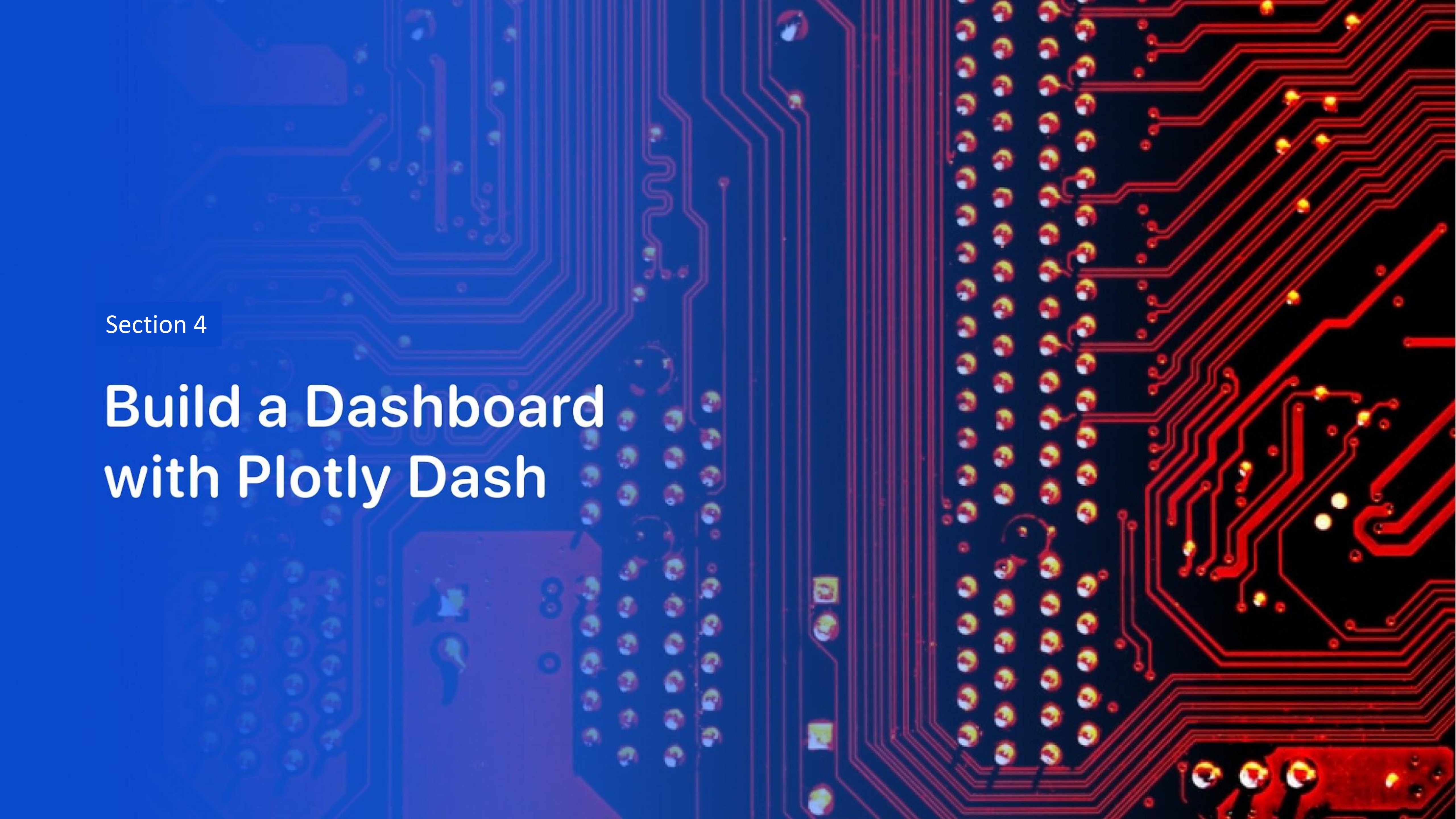
# Launch site proximites

- The launch site known as KSC LC-39A is located in an area that has good logistics advantages. It is situated near a railroad and a road, which makes it easily accessible for transportation of materials and equipment. Additionally, the site is relatively far away from areas with human habitation, which makes it a safer location for launching activities.



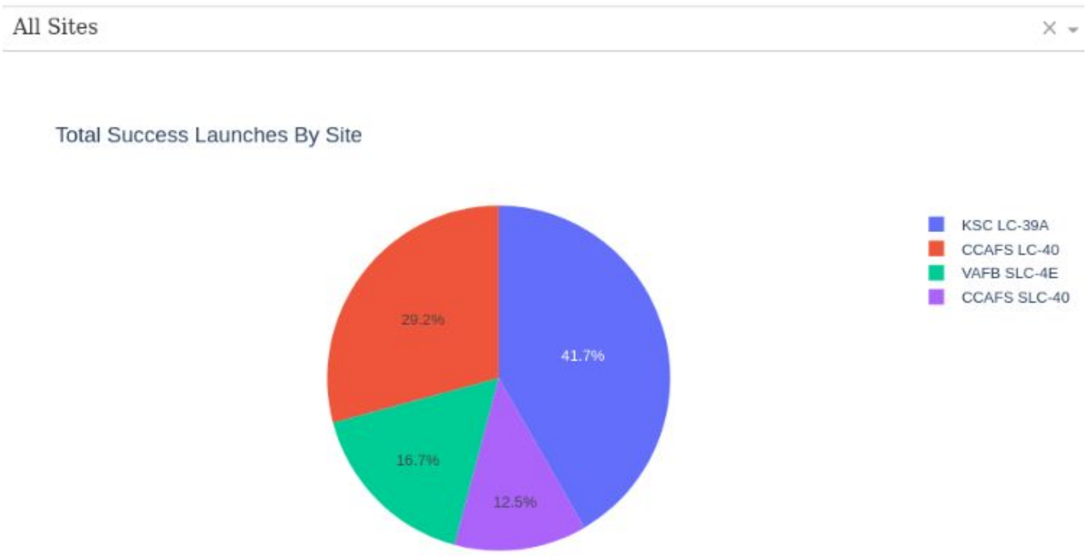
Section 4

# Build a Dashboard with Plotly Dash



# Launch site success

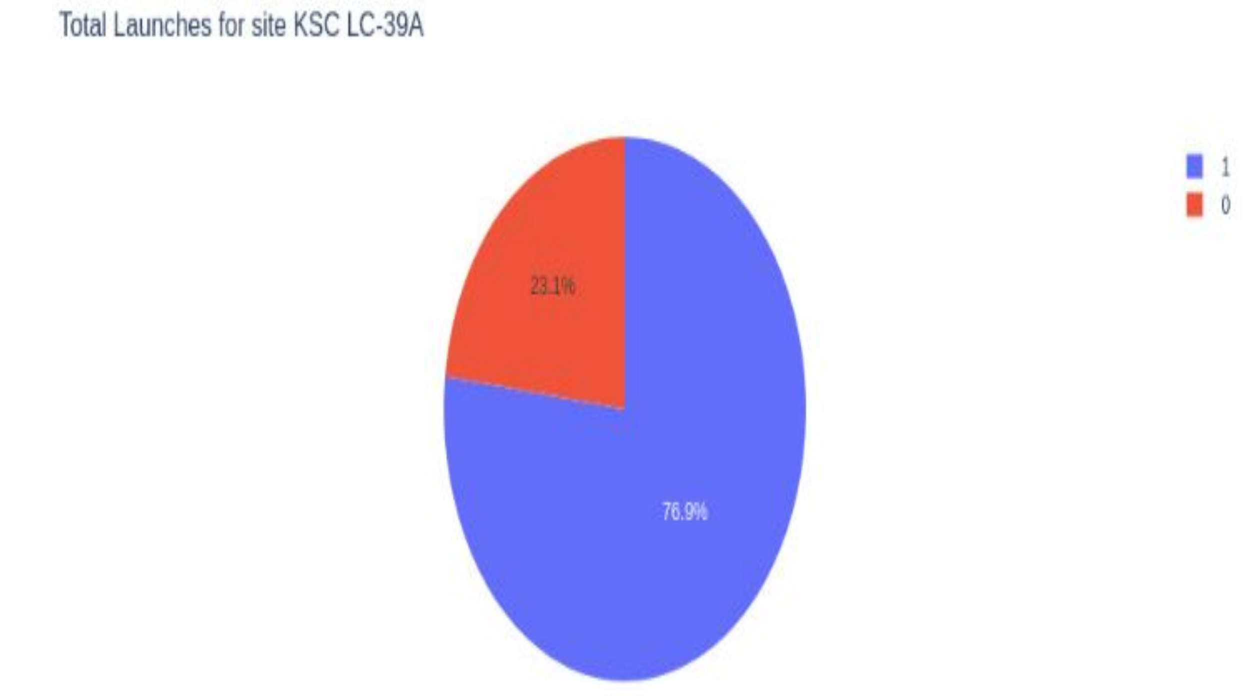
- Success of a launch is different for each launch site.



All Sites

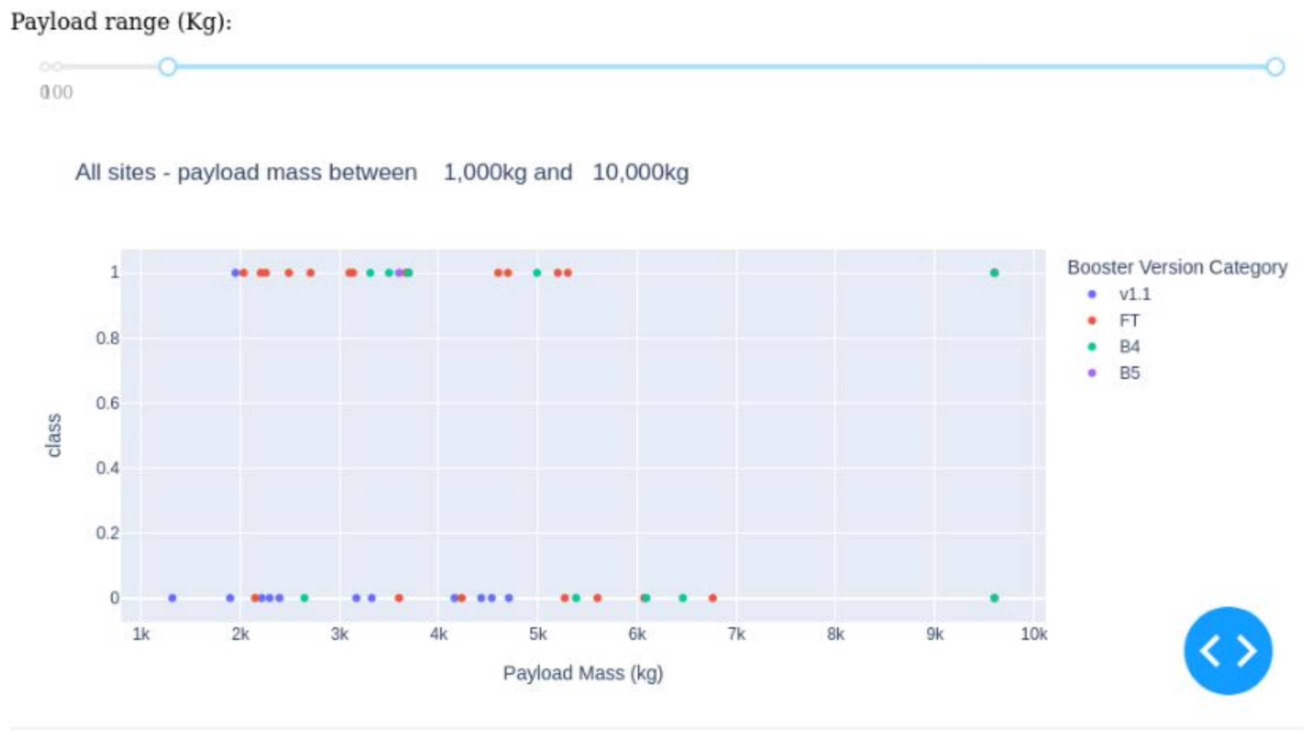
# Launch success rate for KSC LC-39A

- 77% of launches are successful.
- 23% of launches are unsuccessful.



# Launch succession versus Payload mass

- According to data analysis, the most successful combination for payloads weighing under 6,000 kilograms is to use FT boosters. In other words, when FT boosters are used as part of the launch system, the likelihood of a successful launch increases significantly for payloads that are smaller than 6,000 kilograms.



Payload range (Kg):

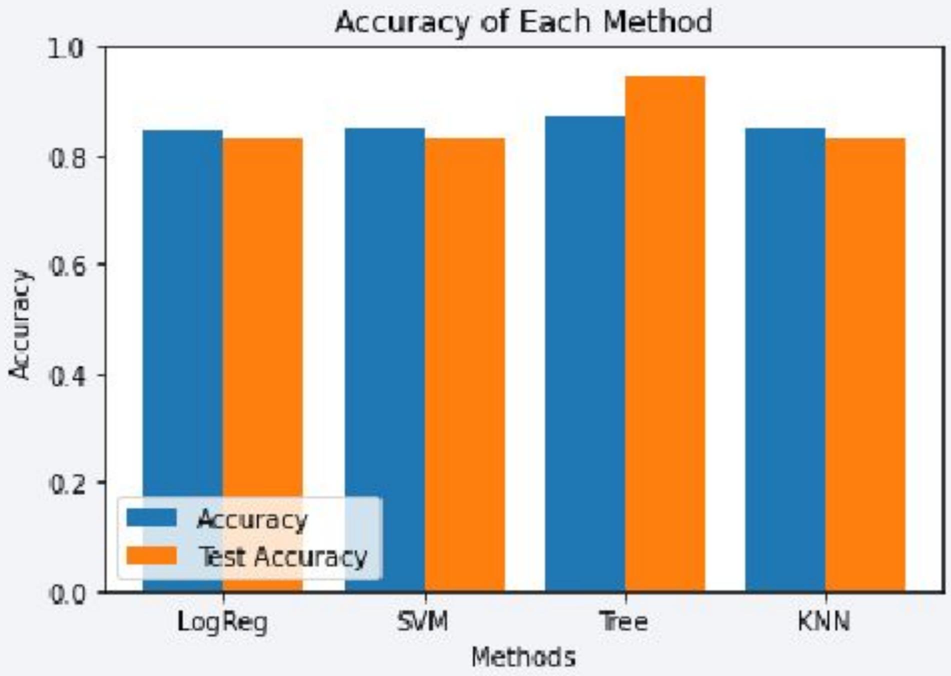
All sites - payload mass between 1,000kg and 10,000kg

Section 5

# Predictive Analysis (Classification)

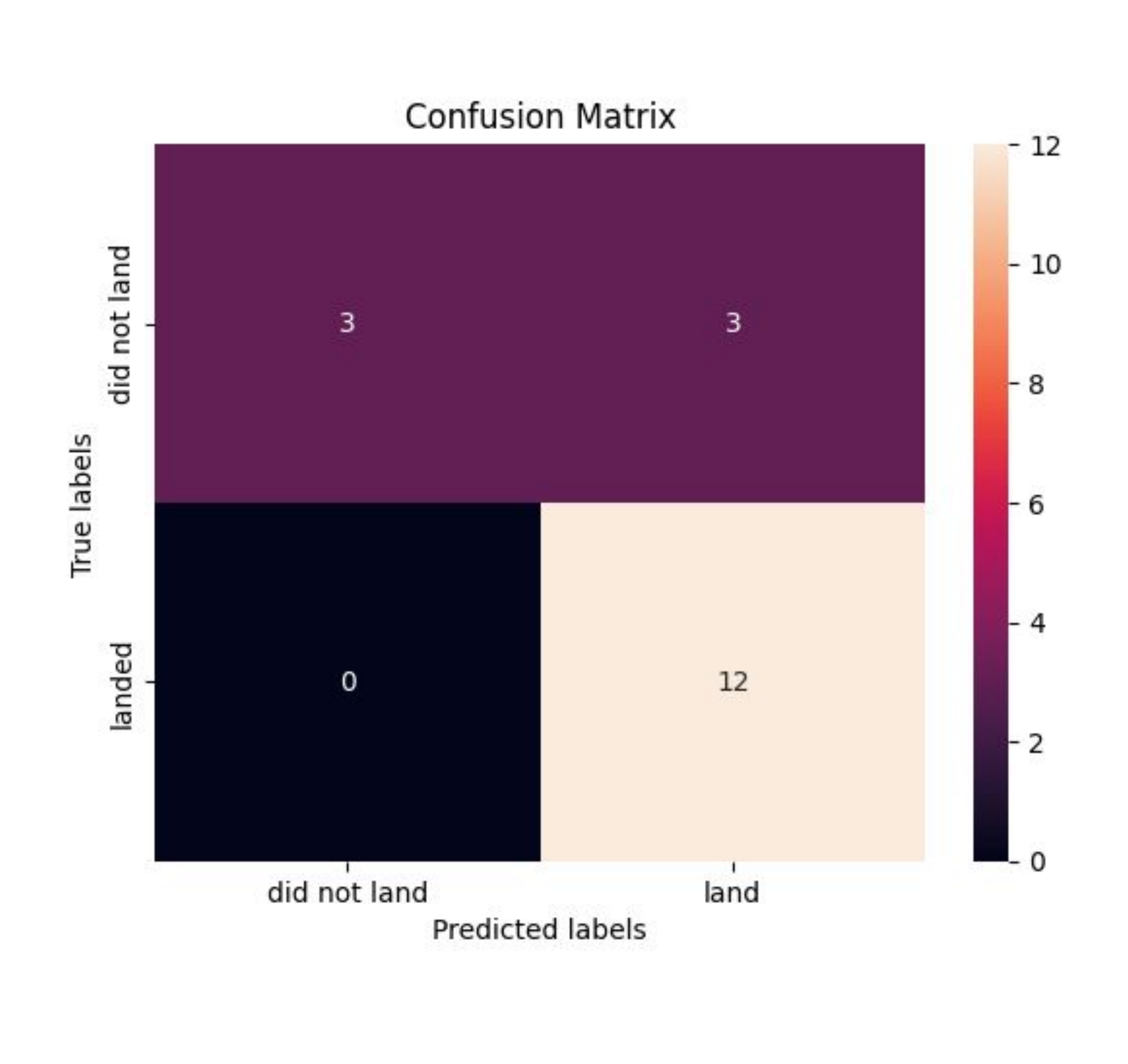
# Classification Accuracy

- Four different classification models were tested to determine their accuracy in predicting the outcomes. It was found that the Decision Tree Classifier had the highest classification accuracy out of all the models. Specifically, its accuracy was 87%, indicating that it was the most reliable and accurate model for the task.



# Confusion Matrix

- The accuracy of the Decision Tree Classifier has been demonstrated through the use of a confusion matrix. This matrix reveals that the model has a high number of true positive and  a balanced number of true negative and false negative, which indicates its ability to correctly identify and classify instances in the dataset.



# Conclusions

- Through the analysis of various data sources, a number of conclusions were reached about the launch process. It was determined that the best launch site is KSC LC-39A and that launches with payloads over 7,000 kilograms are considered to be less risky. While most mission outcomes are successful, the likelihood of a successful landing seems to improve over time as rocket and launch processes evolve.

- To further improve the success rate of landings and increase profits, a Decision Tree Classifier can be used to predict whether or not a landing will be successful. By using this tool, it is possible to make more informed decisions about launch procedures and optimize the process to ensure the greatest likelihood of success.

# Appendix

- Full project here: https://github.com/hristo-darlyanov/SpaceY

Thank you!