# Winning the Space Race with Data Science

-

IBM Developer Skills Network

# Table of Contents


Executive Summary 3

Introduction 4

Methodology 5

Results

- Insights Drawn from EDA 13
- Plotly Dashboard 23
- Launch Sites Proximities Analysis 27
- Predictive Analysis 30

Conclusion 33

# Executive Summary

Throughout the project, various techniques were employed to analyse data around SpaceX launches. From the initial data collection/wrangling and SQL querying, to visualisation techniques and machine learning models, the data was explored to answer some key questions around factors influencing the SpaceX launches.

Data was collected using REST APIs and processed by removing unhelpful columns (features), and filling missing data with suitable replacements. The new dataset was then analysed using SQL for targeted queries such as maximum payload mass or first successful launch dates. Visualisations were also used for exploratory data analysis, including a Folium map for geographical insights and a Plotly Dashboard to interact with the data. Finally, given this analysis, machine learning models were then used for predictive classification to determine whether the first stage (of a launch will land.

The results concluded that more landing attempts were successful from certain launch sites, most notably KSC LC-39A and typically landing had more success with heavier payload masses – specifically above 8,000kg. Launches improved with success as time went on as techniques improved with experience. Orbit types played a role in landing success, with types like SO never being associated with a successful landing, whilst the ES-L1, GEO, HEO and SSO orbit types never failed. Finally, the most accurate classification model (Decision Tree) had an accuracy of above 80%, so its accuracy could be trusted.

# Introduction

The purpose of this project is to explore the success(es) of SpaceX launches and analyse the extent to which different factors influence the outcome of the launch. By investigating data surrounding previous launches, predictive analysis can be conducted on future launches and their success, offering SpaceX insights into the expected costs associated. This information can then be used to help support a competitive edge for the the company against any upcoming commercial bids.

Context given to the project was detailed in the introduction, which is shown below:

> *SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.*

Therefore, the project aims to uncover which factors have the greatest impact on the success of a launch/land; what kind of conditions, locations, features are required to ensure success; and with the data provided, is it possible to predict whether a launch will be successful using machine learning.

# Methodology

-

Section 1

# Methodology

- Data was initially collected from using a SpaceX REST API and by webscraping specific URLs, notably the Wikipedia page for SpaceX Falcon 9 rockets.
- Extracted data was then processed by converting it from a JSON file into a Pandas dataframe, before missing values were found and corrected, and columns/features that were not useful were removed.
- Once the dataset was of a good quality, initial exploratory data analysis was conducted using a range of techniques including visualisations, SQL queries, Folium analytics and a Plotly Dash application; offering different insights and narratives/views into the data.
- Finally, predictive analysis was conducted using classification models including logistical regression, decision trees and k-nearest neighbours algorithms. Of which, the best method was identified and evaluated using different visualization techniques.

# Data Collection & Wrangling – Description

Initially, a GET request was used to retrieve data from the SpaceX API, ensuring a successful request by checking the status code (Slide 7).

The response was then decoded as Json and normalized into a Padas dataframe so that it is more manageable.

A second subset from that dataframe was taken, excluding any irrelevant features like IDs or names.

The API was used again to fill lists of extra information that is highly relevant to launch success, and then put into lists.

These lists were then combined into a dictionary, which was used to create a final Pandas dataframe.

The number/percentage of missing values was calculated for each variable, and replaced with suitable values that would not invalidate the data (i.e. mean).

Various calculations were made to give a better understanding of the data, including frequency metrics, overview of the data types for each variable, success ratio, etc.

New columns were created as simple representations, including whether the first stage of the launch was successful (Class: 1) or unsuccessful (Class:0).

The entire data collection stage can be found on the Slide 8

https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/SpaceX%20Data%20Collection%20API.ipynb
https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/SpaceX%20Data%20Wrangling.ipynb

# Data Collection & Wrangling – Using APIs

```
import requests
import pandas as pd
import datetime
import numpy as np
```

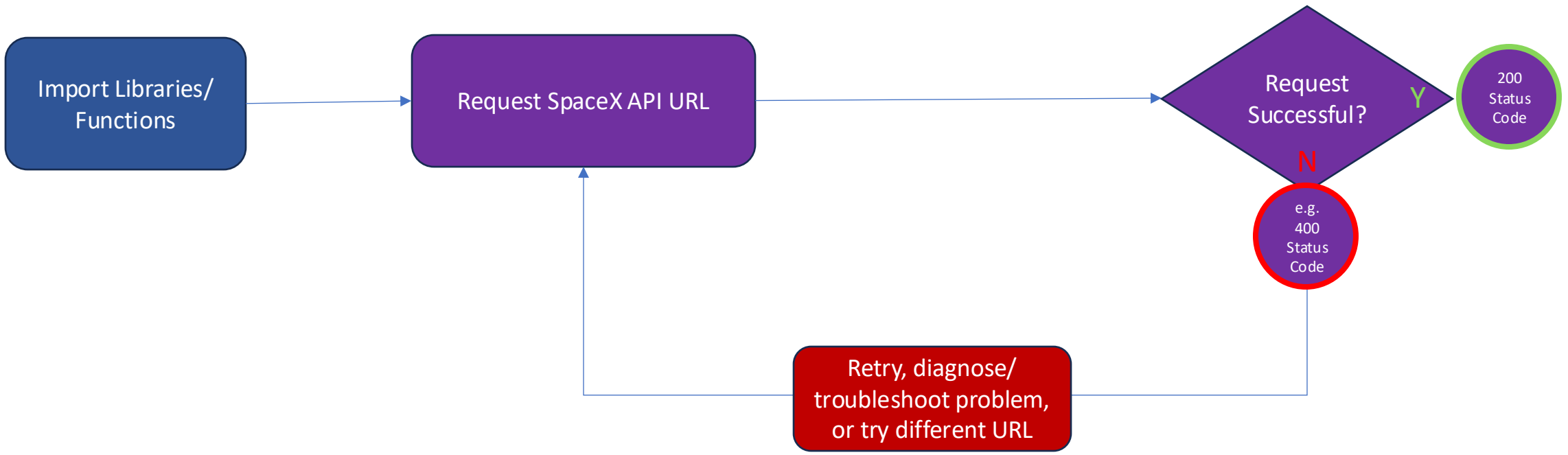
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

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

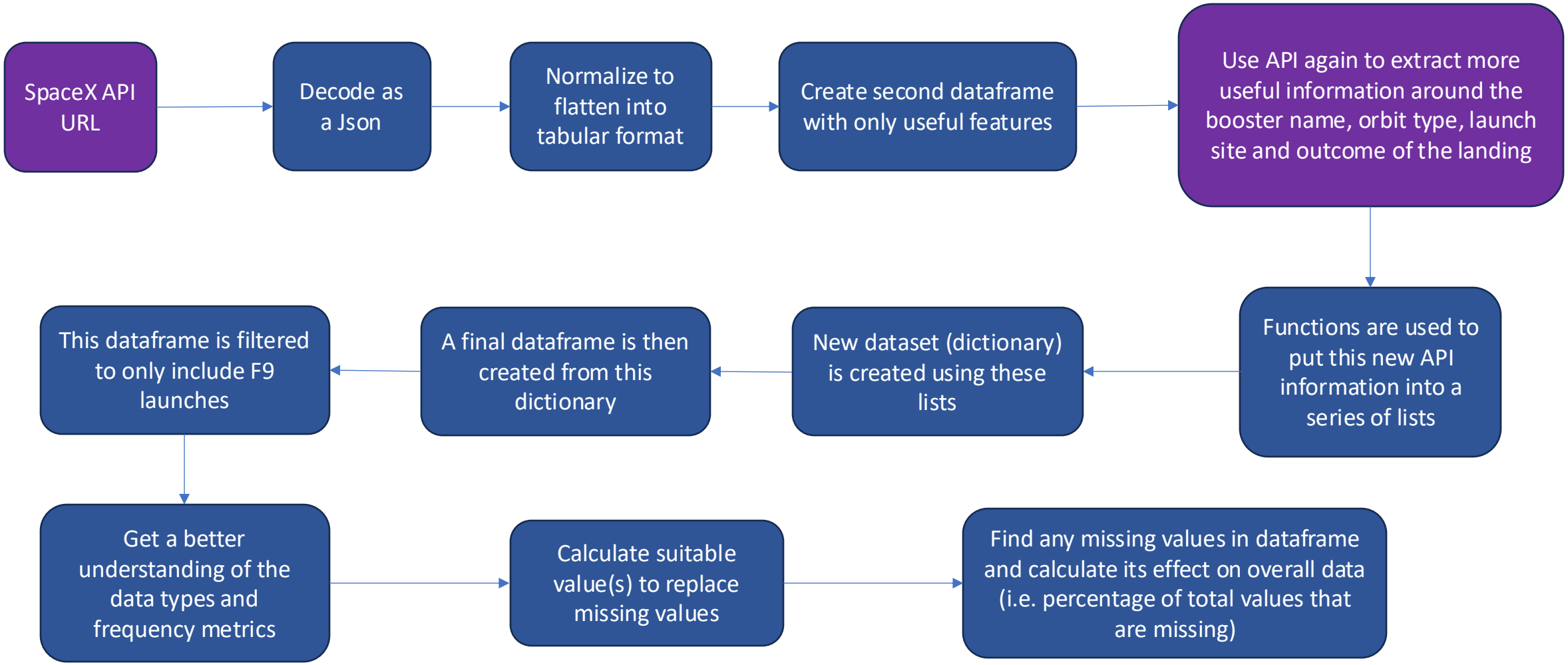
```
static_json_url='https://cf-courses-data.s3.us.cloud-object-stor
```

```
response.status_code
```

```
200
```



# Data Collection & Wrangling – Overview



# EDA with Visualization & SQL

A range of different scatter plots, bar graphs and line plots were created to visually understand the dataset, before extracting meaningful insights and patterns from the visualisations. These visualisations included analysis of payload mass, flight numbers and orbit types by both the launch site and success, as well as analysis into the change in yearly average success of landings.

From these visualisations, it was quick to ascertain the following:

- The CCAFS SLC-40 site had the most success, particularly after the first 25 flights
- On average, the launches with heavier payload masses typically were more successful (8,000kg <)
- Launches were more successful as time went on (i.e. success increased between 2010 and 2020)
- Most launches were from CCAFS SLC-40, but most success on average was had at KSC LC-39A

Using SQL, quick insights were drawn from the data, in queries that were better understood as outputs or tables than visualisations. These include:

- The number of successful and unsuccessful launch (landing) outcomes
- The maximum payload mass
- The date of the first successful mission, by landing outcome
- The companies/customers who supplied the largest amount of payload mass

https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb
https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

# Interactive Folium Map & Plotly Dash Application (Dashboard)

An interactive map was built using Folium to better understand the geographic overview of the launch sites, taking into consideration the relative distance to types of areas.

Distance markers were placed between the launch site that had the fewest markers (CCAFS SLC-40) and the nearest coastline, railway, highway and city (selected to be Titusville).

These distances were then marked on the map, at the end of each of their respective lines.

A Plotly Dash application was also made to interact with the data, and quickly drill into the success of launches by different features, including launch site, payload mass and boosters.

This application was useful as one could freely filter to specific launch sites or booster versions as well as change the overall range of the graph so view more granular details of various features.

https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/Launch%20Site%20Location%20Folium%20Map.ipynb
https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/SpaceX%20Dash%20App.py

# Predictive Analysis (Classification)

```python
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```python
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr, parameters, cv=10).fit(X_train, Y_train)
```

tuned hpyerparameters :(best parameters) {'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'} accuracy : 0.8196428571428571

Data is loaded and split into training and testing data

First model (logistic regression) is built by creating an GridSearchCV object, *logreg_cv* with given parameters, estimator and cross validation

GridSearchCV object outputs the parameters and accuracy on the validation data

Accuracy score is printed

```python
print(logreg_cv.score(X_test, Y_test))
```

Confusion matrix is calculated to determine how the model performed

```python
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Process is repeated for all other types of models

Accuracy and best score accuracy are calculated to determine best model method

https://github.com/alexandermanning23/Data-Science-Capstone/blob/main/SpaceX%20Machine%20Learning%20Prediction.ipynb

# Insights Drawn from EDA

-

Section 2

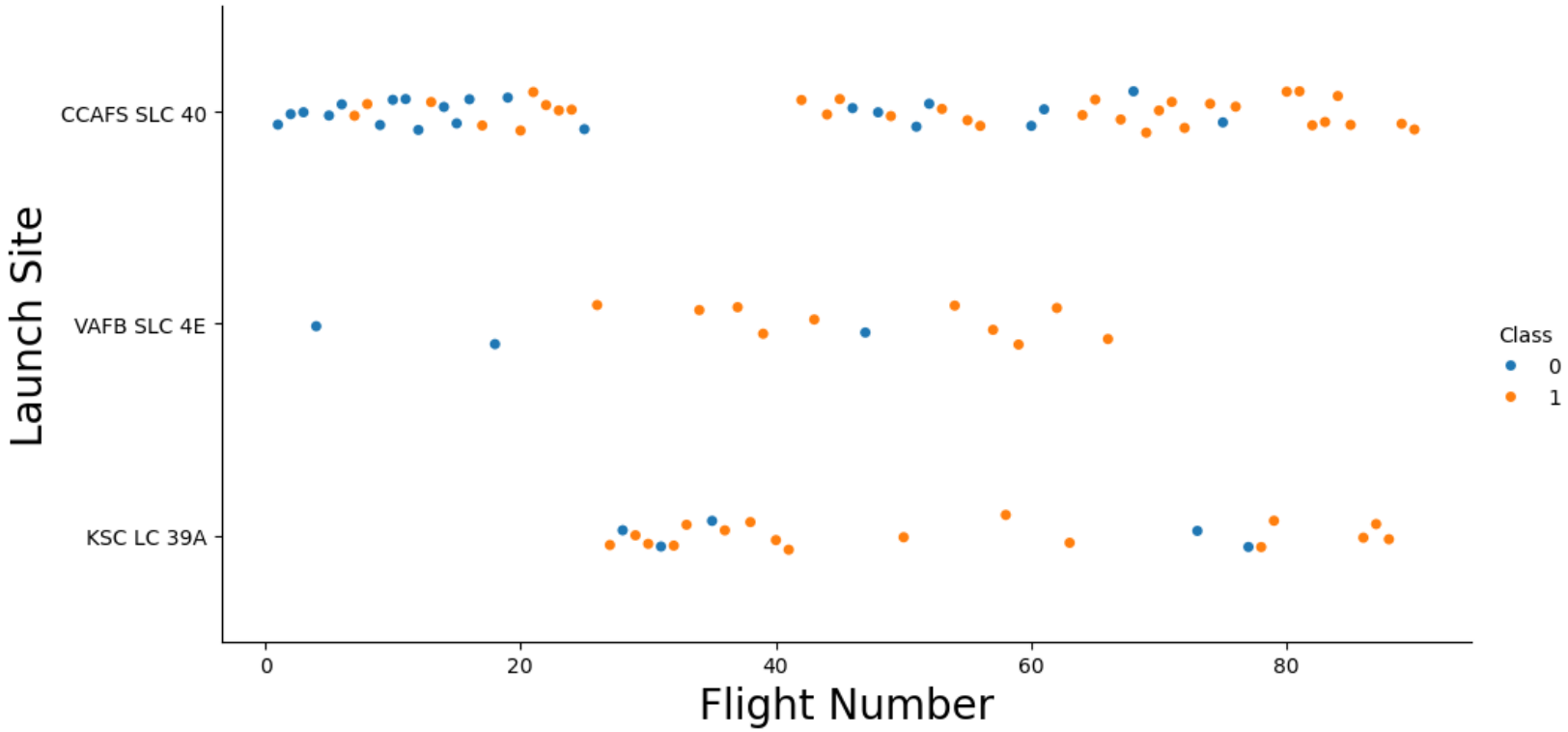
# Flight Number against Launch Site

As one might expect, the later flights (missions with higher Flight Number) tended to be more successful, likely due to improvements and experience.

The frequency of missions by launch site can also be determined, with the CCAFS SLC 40 site having the most launches and VAFB SLC 4E having the fewest.

One can also see that around the 25th flight, KSC LC 39A was used used for the first time, and fairly frequently for over the next 15 flights. It was also at a time where no missions were launched from CCAFS SLC 40 which was previously the most used site.

After this gap between the 25th and 40th Flight Numbers, the CCAFS SLC 40 site was reused again with the same frequency as the first 25 flights, but with a higher success rate. From this, it could suggest that that site underwent maintenance that improved the success ratio of flights.



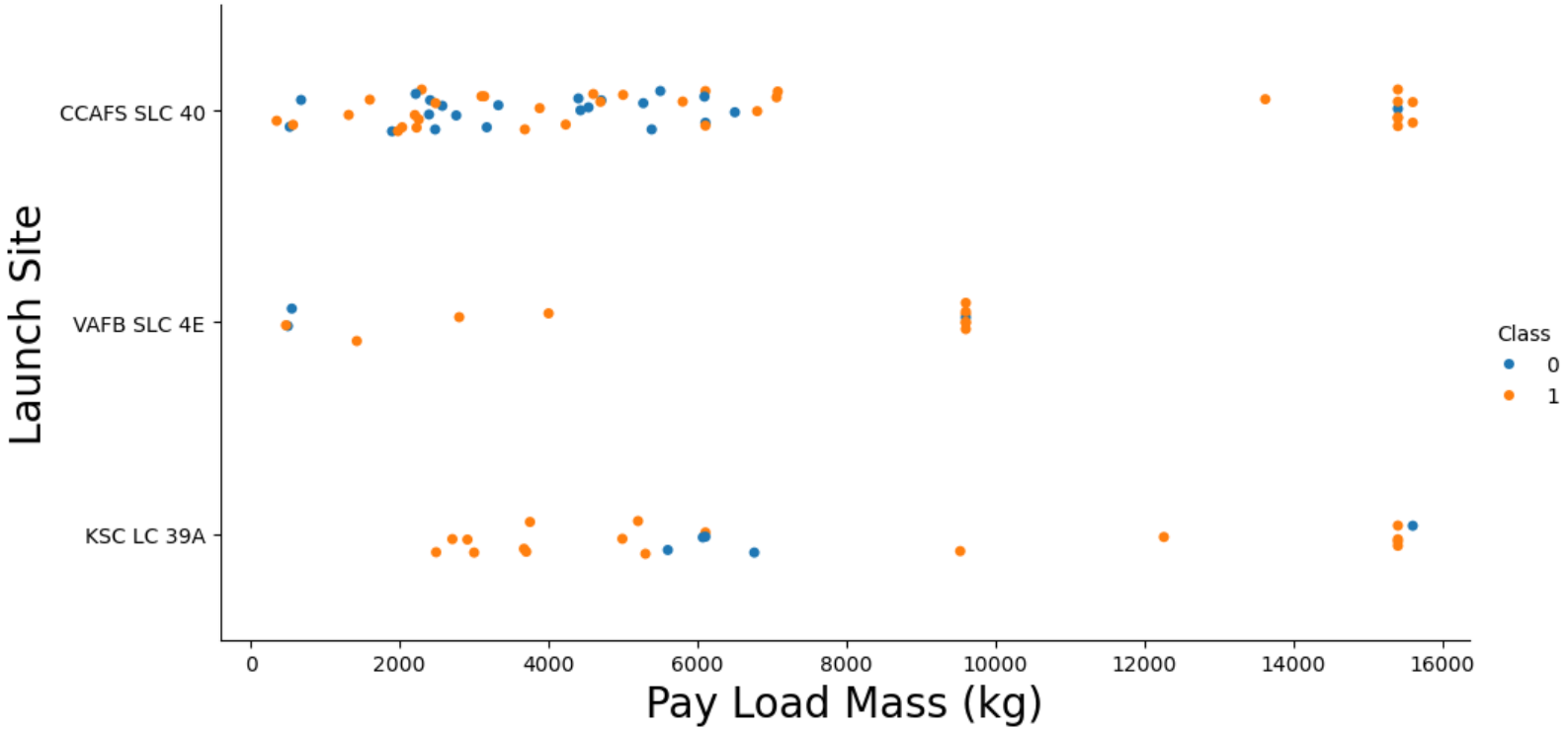
# Payload Mass by Launch Site

From the Payload vs Launch site, one can determine which payloads are most used at each launch site and their respective success.

One can see that the maximum payload used at the VAFB SLC 4E site was 10,000kg, whereas both the other sites exceeded that by a value of over 5,000kg.

Similarly, the KSC LC 39A site used rockets that exclusively launched with a payload of over 2,000kg.

One can also see that after a successful launch, the same payload mass was typically re-used, particularly with the larger masses. For example, at least 6 rockets were launched at around 9,500kg and more than 10 rockets were launched at around 15,500 – 15,600kg.

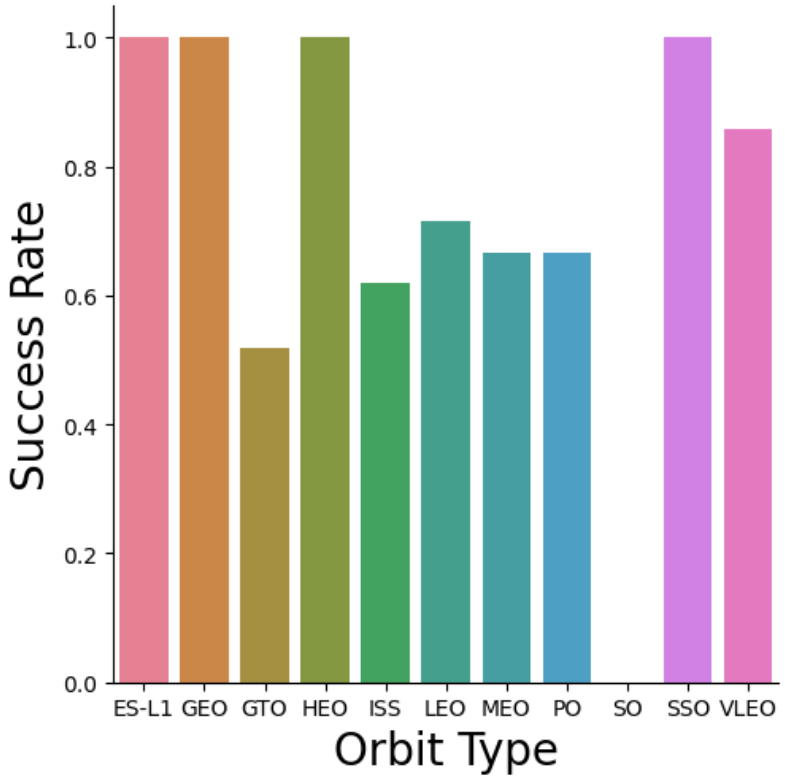


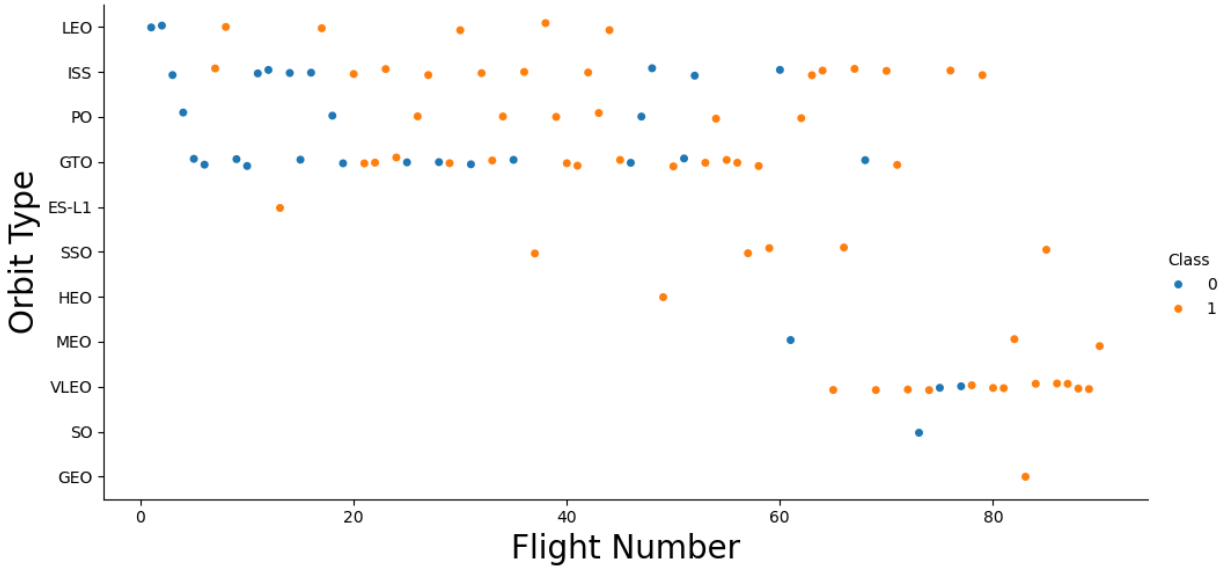
# Success Rate / Flight Number by Orbit Type

From the graph show, one can see that there are clearly more successful and less successful orbit types.

For example, the SO orbit was never successful, whilst the ES-L1, GEO, HEO and SSO orbit types were always successful.

One can also see that all other orbit types had a success percentage of ≥ 50%, with VLEO having a success rate of almost 90%.




When comparing the Flight number to orbit type, one can see that the typically, the later flights were more successful, with clear orbit types being more successful the later the flight number. Those orbits include LEO, PO and MEO.

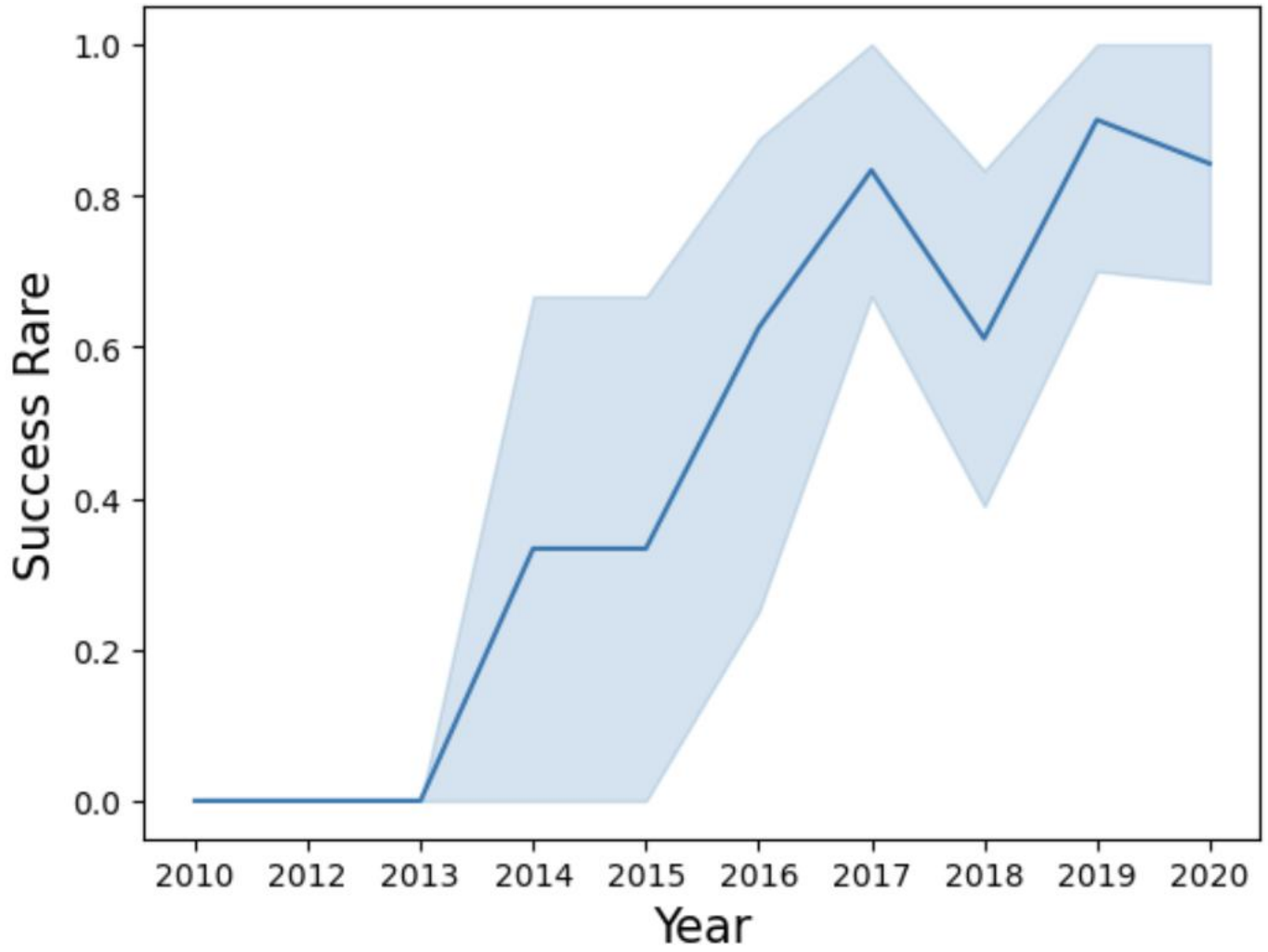
Other orbits like ISS and GTO have more success, but there is much less of an obvious correlation, suggesting that with more flights, there have been no major insights to be gained.

# Launch Success Yearly Trend

This graph shows the yearly average success rate of launches between 2010 and 2020.

As one might expect and/or hope, the overall success rate continued to increase over time, from 2013 until 2020. The only exceptions were between 2014-2015, where the success rate did not change and in 2018 and 2020, which were the only years where the the average success rate actually decreased from the previous year.

One can also see the level of improvement made in those 7 (or 10) years, where the later years (particularly 2017, 2019 and 2020) had a very high average success rate of over 0.8.



# All Launch Site Names

| Launch_Site | count |
|---|---|
| CCAFS LC-40 | 26 |
| CCAFS SLC-40 | 34 |
| KSC LC-39A | 25 |
| VAFB SLC-4E | 16 |

These are all the different unique launch sites, and the number of launches recorded in the data associated with them.
One can quickly and easily see that the CCAFS SLC-40 site was the most used with a total of 34 launches, followed by CCAFS LC-40 and KSC LC-39A with 26 and 25 launches respectively. The least used site was VAFB SLC-4E with only 16 launches.

```
%sql SELECT "Launch_Site", count(*) AS count FROM
SPACEXTABLE GROUP BY "Launch_Site"
```

# Launch Site Names Begin with 'CCA'

Here are five records where the launch site starts with CCA, which includes two of the four possible sites - CCAFS SLC-40 and CCAFS LC-40. Despite both sites starting with CCA, the five records are all related to the latter launch site.

Since these records are ordered (by date), one can see that the CCAFS LC-40 launch site was the first to be used of all the analysed launch sites.

| 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 |

```
%sql select * from SPACEXTABLE
where Launch_Site like 'CCA%' limit 5;
```

# Total Payload Mass

This is the total payload mass carried by all of the boosters provided by NASA.

Using this total, one can drill down to determine whether NASA was the largest supplier of payload mass for SpaceX, regardless of the number of launches or average payload mass they supported/supplied.

| sum(PAYLOAD_MASS__KG_) |
| --- |
| 45596 |

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE
where CUSTOMER = 'NASA (CRS)';
```

# Average Payload Mass by F9 v1.1

| avg(PAYLOAD_MASS__KG_) |
| --- |
| 2928.4 |

Similarly, this is the average payload mass carried by the F9 v1.1 booster. This is another statistical characteristic of the dataset, which can be used in the converse way, to extrapolate outward. For example, to find the total payload carried by the F9 v1.1 booster, the 2928.4kg value can be multiplied by the number of launches using that booster.

It can also be assumed that the F9 v1.1 booster is typically used on rocket launches with smaller payload masses.

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE
where BOOSTER_VERSION = 'F9 v1.1';
```

# First Successful Ground Landing Date

This result is the date of the first successful landing by ground pads.
Using SQL, one can pass dates through statistical metrics like maximum, mean, mode - or in this case - minimum, to find the smallest (earliest) date that fits the criteria.

This can be used to quickly determine which landing techniques were used first or introduced later, or whether they were typically more successfully earlier on, or became more successful with more (failed/unsuccessful) attempts. Similarly, one can establish whether they were successful or not in their most recent launch.

| min(DATE) |
| --- |
| 2015-12-22 |

```
%sql select min(DATE) from SPACEXTABLE
where LANDING_OUTCOME = 'Success (ground pad)';
```

# Successful Drone Ship Landing with Payload between 4000 and 6000

These are all the boosters that have successfully landed on drone ships with a payload mass between 4000kg and 6000kg.

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

The query can easily be revised to change the payload mass range to include heavier or lighter payloads, against each landing outcome.
From this example, one can see there are four different booster versions that successfully landed on drone ships, with the given payload mass range of 4000-6000kg.

```
%sql select BOOSTER_VERSION from SPACEXTABLE where
LANDING_OUTCOME = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

# Total Number of Successful and Failure Mission Outcomes

These are the total numbers of each successful and failed mission outcomes, by their individual landing outcome. One can see that there was only one unsuccessful mission, which failed during the flight, with a lot more launches failing upon their landing, making a total of 10 failed landing attempts.

| MISSION_OUTCOME | Mission_Outcome | Landing_Outcome |
|---|---|---|
| 5 | Success | Controlled (ocean) |
| 3 | Success | Failure |
| 5 | Success | Failure (drone ship) |
| 2 | Success | Failure (parachute) |
| 21 | Success | No attempt |
| 1 | Success | No attempt |
| 1 | Failure (in flight) | Precluded (drone ship) |
| 38 | Success | Success |
| 14 | Success | Success (drone ship) |
| 9 | Success | Success (ground pad) |
| 2 | Success | Uncontrolled (ocean) |

```
%sql select count("Mission_Outcome") as
MISSION_OUTCOME,MISSION_OUTCOME, LANDING_OUTCOME
from SPACEXTABLE
group by "Landing_Outcome";
```

# Boosters Carried Maximum Payload

These are all the booster versions that carries the maximum payload mass (which was found to be 15,600kg as shown with the query at the bottom).

From this list to the right, one can see that twelve boosters in total carried that maximum payload mass

| 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 |

```
%sql select BOOSTER_VERSION from SPACEXTABLE where
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from
SPACEXTABLE);
```

```
%sql select payload_mass__kg_ from
SPACEXTBL order by payload_mass__kg_ desc
LIMIT 1
```

| PAYLOAD_MASS__KG_ |
|---|
| 15600 |

# 2015 Launch Records

| month | Date | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|---|
| 01 | 2015-01-10 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | 2015-04-14 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

These are any failed landing attempts from the year 2015. This can easily be revised for any lading outcome and any year to quickly get an overview of the level of success for a given year.

In this example, one can see there were two unsuccessful landing attempts (by drone ship) in both January and April of 2015

```
%sql select substr(Date,6,2) as MONTH,
DATE,LANDING_OUTCOME,BOOSTER_VERSION,
LAUNCH_SITE from SPACEXTABLE where LANDING_OUTCOME =
'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

This Shows all landing outcomes between 4/6/2010 and 3/20/2017, and ranks them by their frequency.
You can see the most common outcome is for there to be no landing attempted which is twice as high as the second most common occurrence.

The joint second most common occurrences are the success and failure of the drone ship landings, suggesting that despite the high frequency of success, they fail just as often.

| Landing_Outcome | Frequency |
|---|---|
| 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 |

```
%sql select LANDING_OUTCOME, count(*) as Frequency from
SPACEXTABLE
where DATE between '2010-06-04' and '2017-03-20' group by
LANDING_OUTCOME order by Frequency DESC;
```

# Launch Sites Proximities Analysis

-

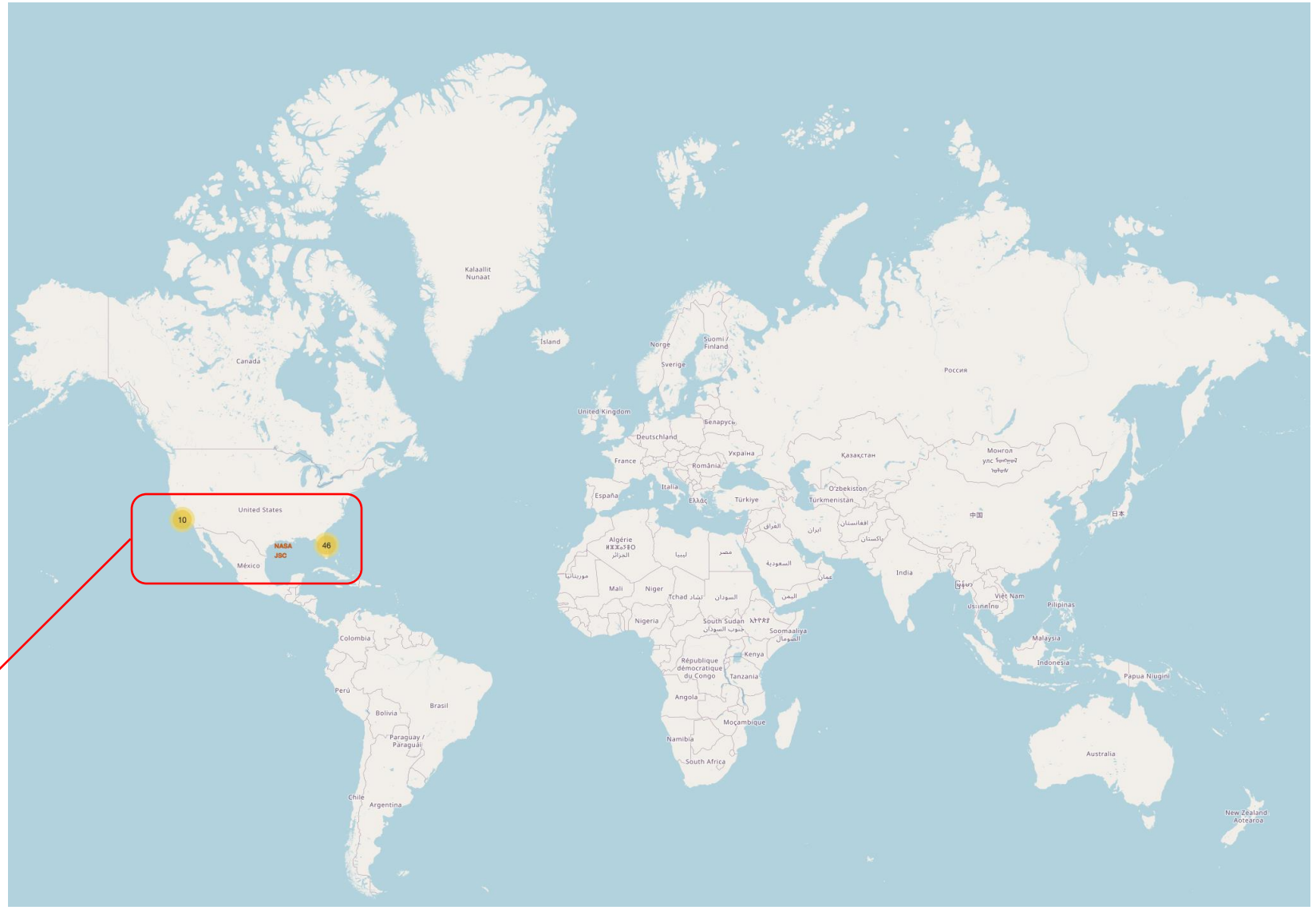
Section 3

# Global Launch Sites Analysed

Despite the vast number of launch sites on the globe, the dataset used for this project only included SpaceX launches.

That is the reason, when looking at the Folium map generated to the right, the only two areas analysed are within the United States.

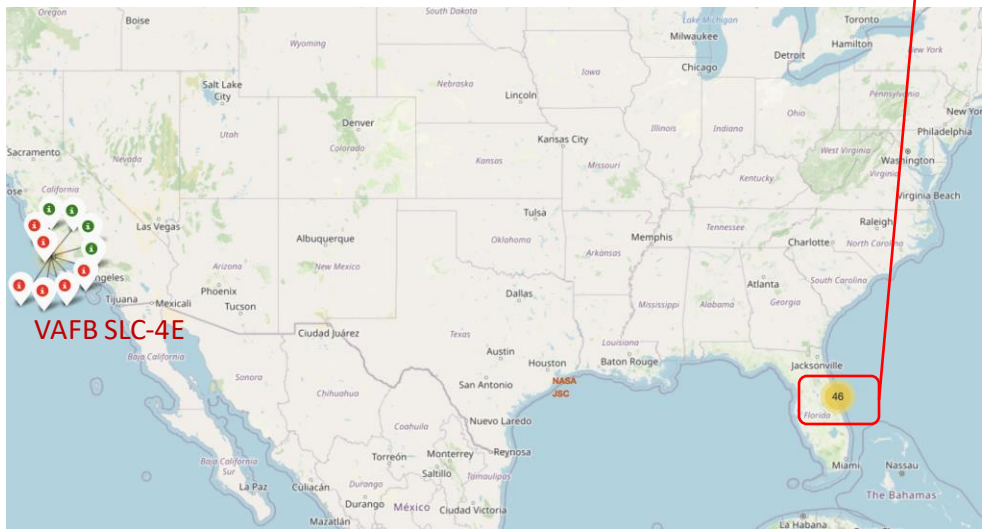
One can also clearly see a total of 56 launches analysed, with 46 launches in the state of Florida, and 10 in the state of California.

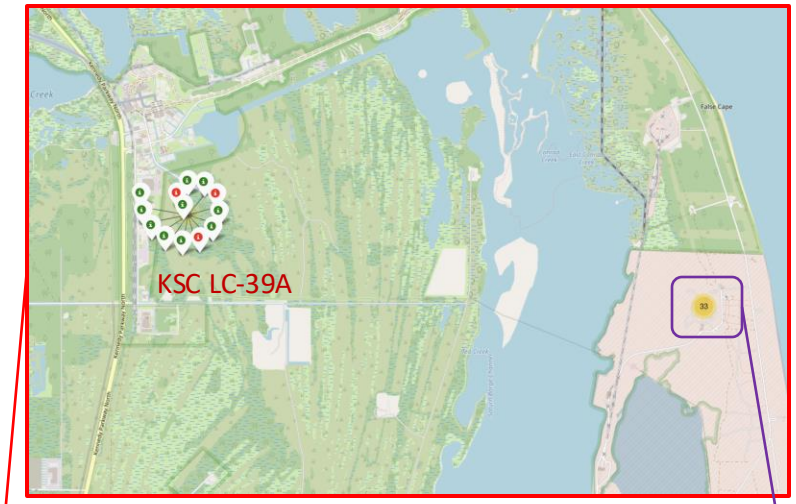



# Visual Analysis of Launch Sites

From these two areas (California and Florida), one can drill down to see their levels of success, shown by colour-coordinating a failed and successful landing by red and green respectively.

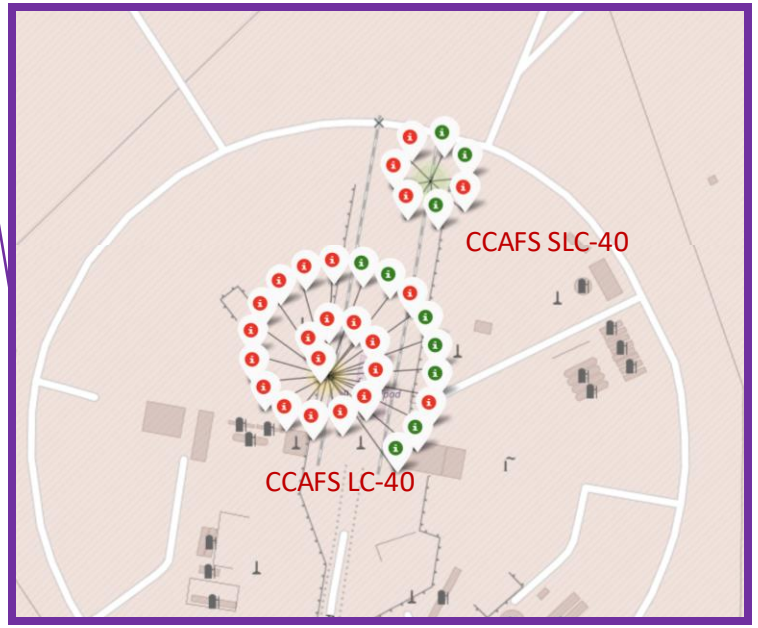
When zooming into the map, and clicking on each of the four locations, the landing attempts fan out to show whether they were successful or not.



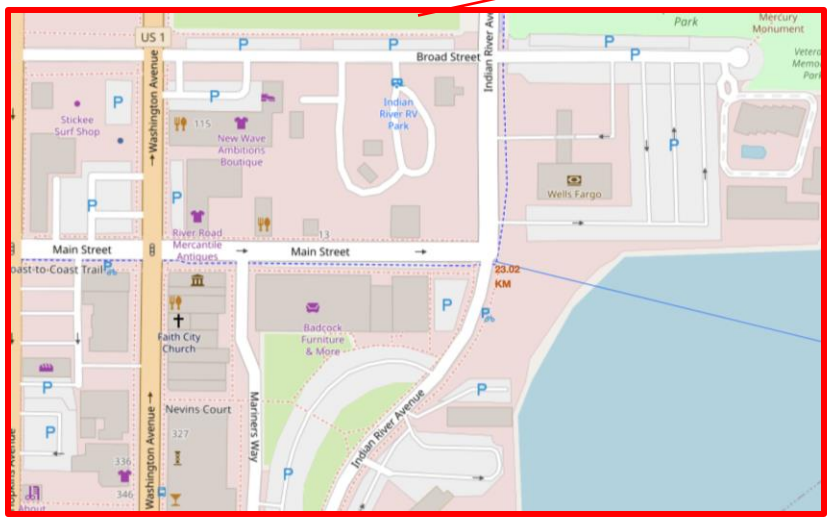

The number 46 in Florida divides into two different areas once zoomed in enough, revealing one launch site and another embedded number of launches (33), which expands again to reveal the two CCAFS LC-40 and CCAFS SLC-40 sites

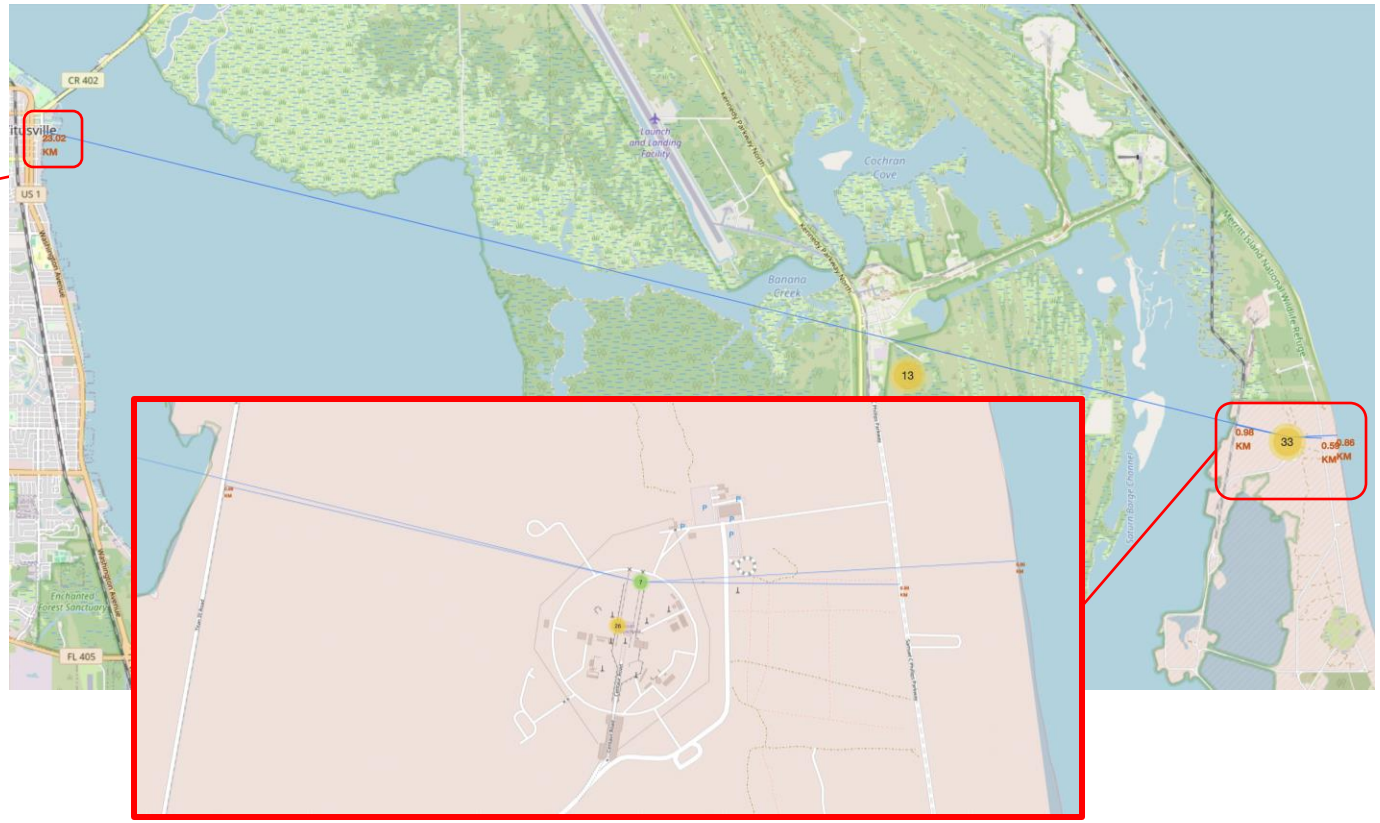
Once all the sites are reviewed, one can quickly see that KSC LC-39A was the site that had the most successful landing outcomes, with only 3 red marks and 10 green marks. On the other hand, CCAFS LC-40 was the least successful site.



# Analysis into Launch Site Distances

Selecting a single launch site, CCAFS SLC-40, one can anaylse the nearest types of areas, including cities, highways, railways and coastlines.




From these maps, one can see that from CCAFS SLC-40, these are the nearest types of areas:

| Area/Landmark | City | Coastline | Railway | Highway |
|---|---|---|---|---|
| Distance from launch site | 23.02 KM | 0.86 KM | 0.98 KM | 0.59 KM |

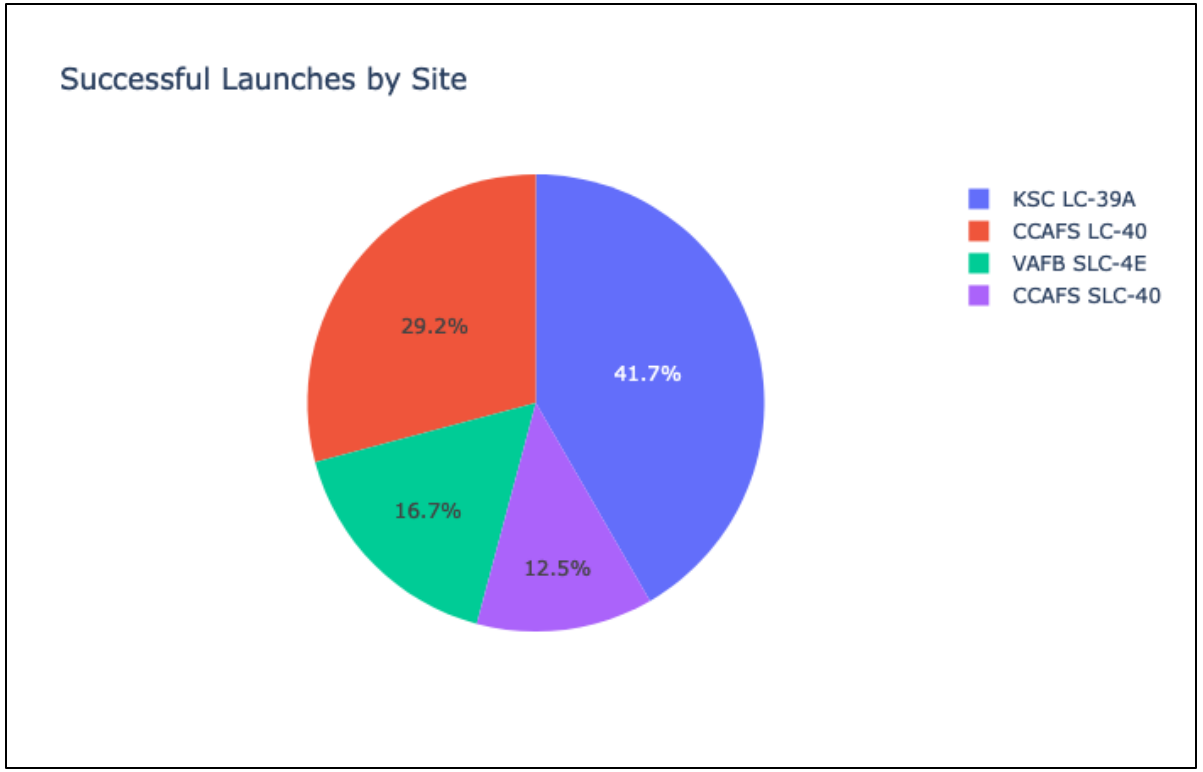
# Build a Plotly Dashboard

-

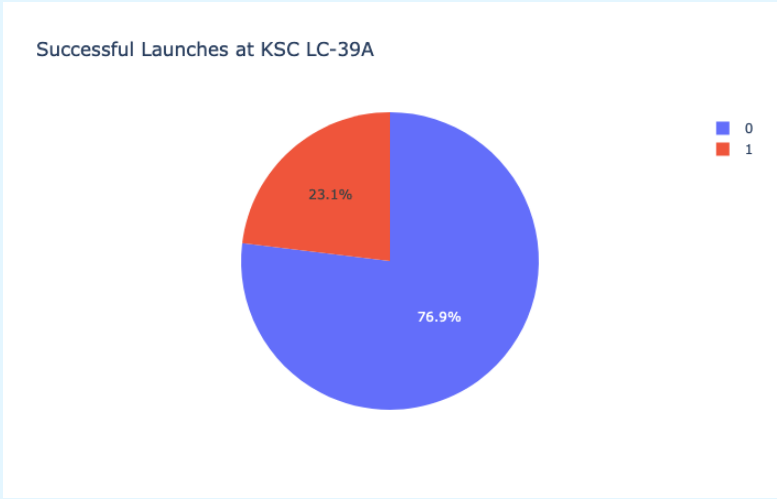
Section 4

# Successful Launches (Landings) by Launch Site

From the pie chart showing the percentage of all successful launches by site, one can see that the most successful site was KSC LC-39A, contributing to over 40% of the total successful launches.

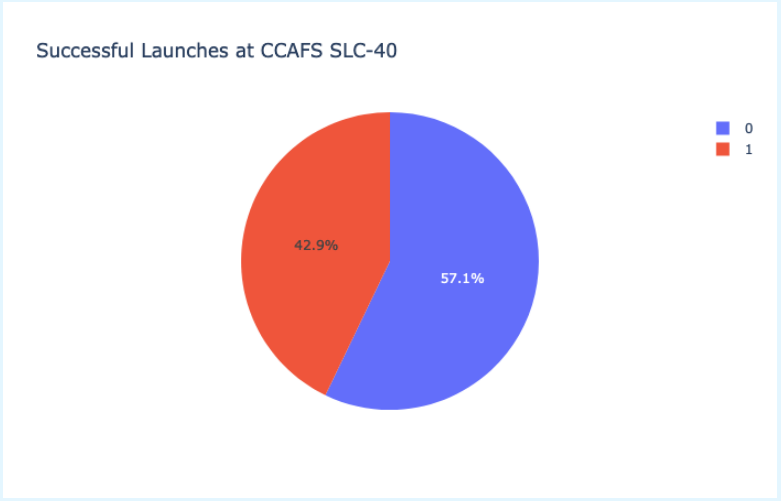



# Launch Sites with Highest / Lowest Launch Successes

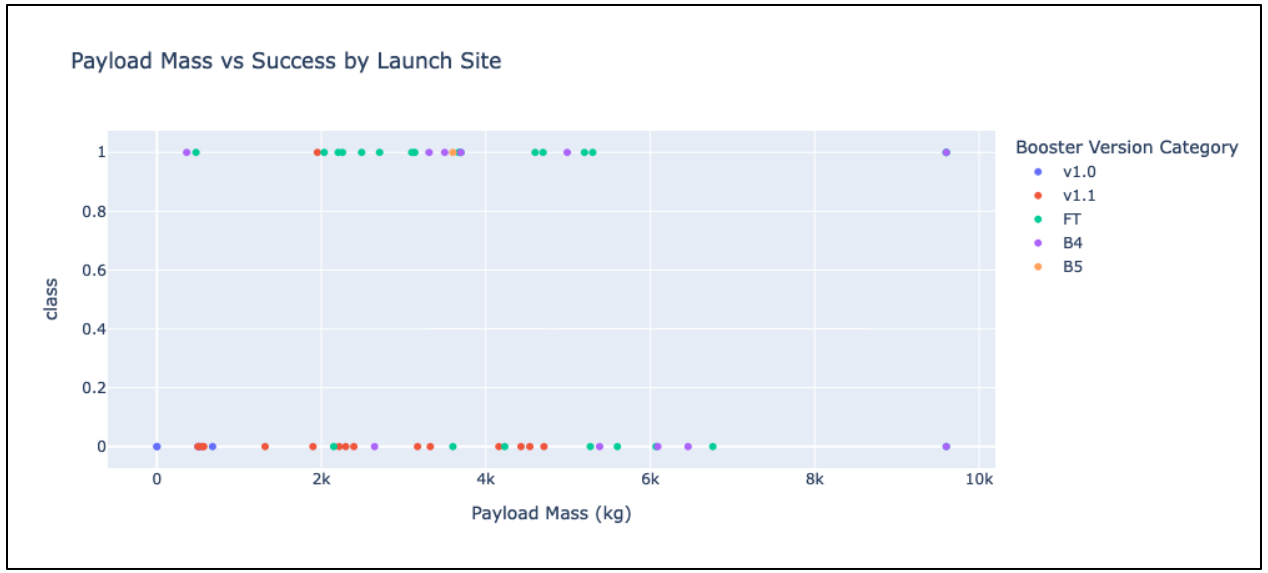



Looking at specific sites with the highest and lowest success percentages, one can see that of all KSC LC-39A launches, 76.9% were successful.

Whereas CCAFS SLC-40 had the lowest percentage of successful launches at 57.1%



# Impact of Payload Mass on Launch Outcome

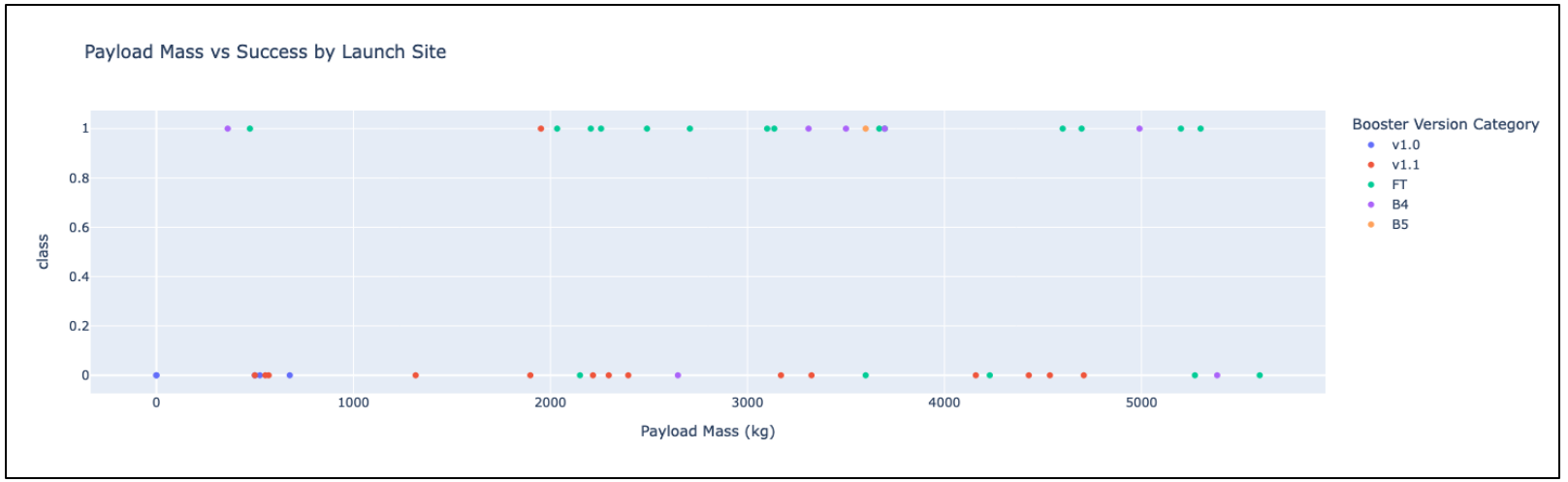



Looking at the impact of payload mass on launch outcome, between 0kg and 10,000kg one can see that the most success was had with payload masses of less than 6,000kg, with only a single success larger than that with the B4 booster (at around 9,600kg).

One can also see launches with payloads of over 5,000kg were only attempted by FT and B4 booster versions

By scaling the range down to 6,000kg, one can get a more granular view of the outcomes.

This reveals that the FT booster typically had the most success and the v1.0 and v1.1 had the least success, with the former being attempted three times with no success at all.



# Predictive Analysis (Classification)

-

Section 4

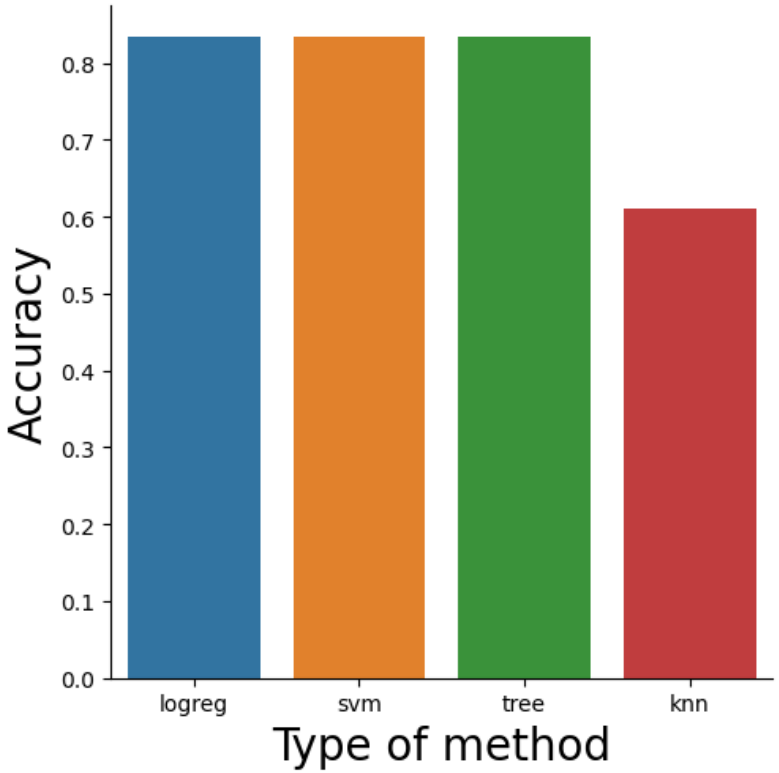
# Classification Accuracy

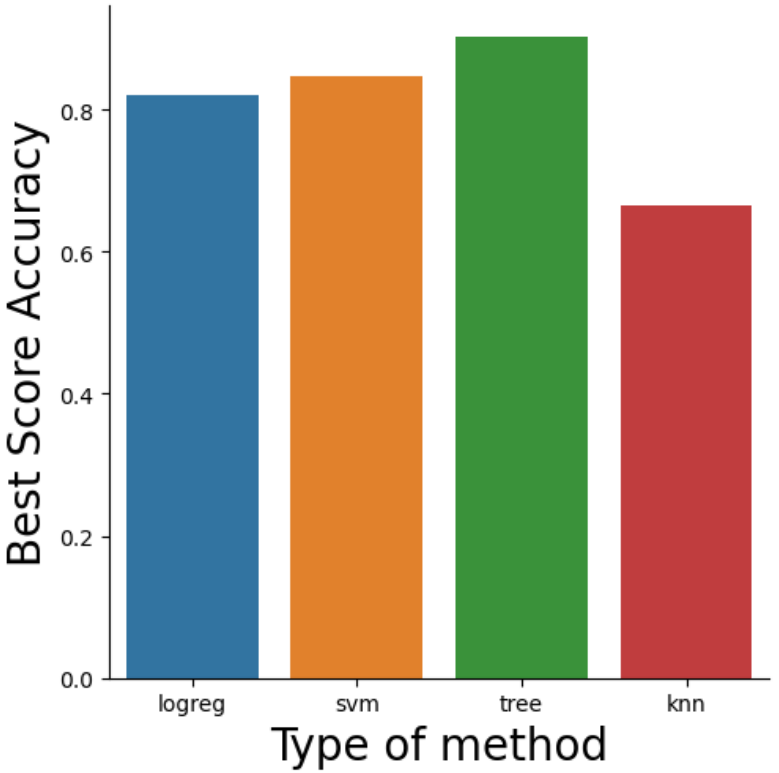
When looking at the predictive analysis, it was important to ascertain which machine learning model method had the most accurate score.

As shown in the graphs, by looking at just the accuracy score, the most accurate type of method is reduced by only one, down to three methods, failing to suggest only one method.

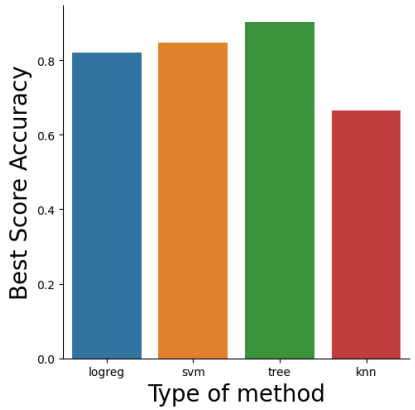
However, upon further analysis by including the best accuracy score, one can clearly see that the decision tree method offers the highest accuracy overall.

| | method | accuracy | best_score_accuracy |
|---|---|---|---|
| 0 | logreg | 0.833333 | 0.819643 |
| 1 | svm | 0.833333 | 0.846239 |
| 2 | tree | 0.833333 | 0.901786 |
| 3 | knn | 0.611111 | 0.664286 |



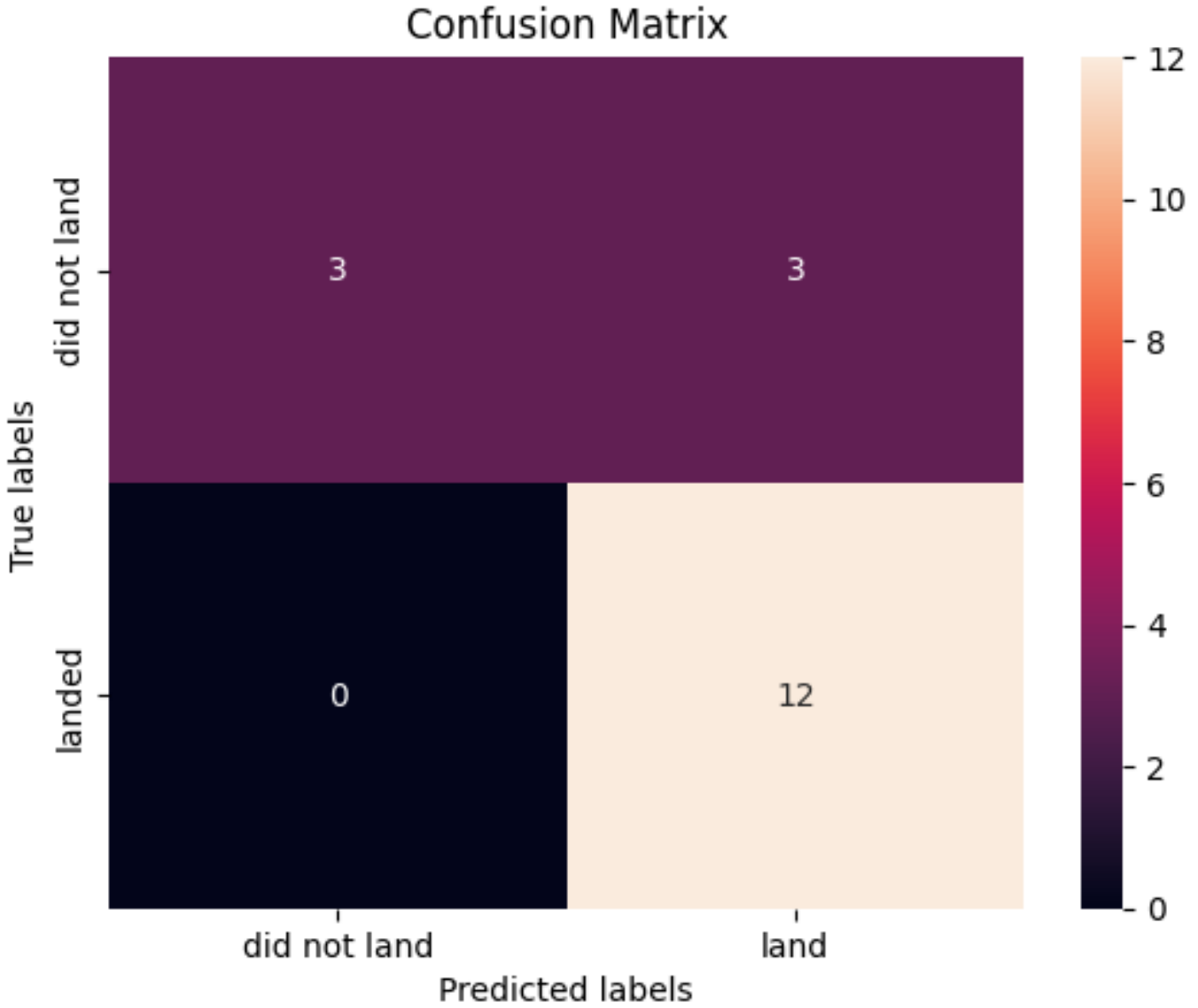



# Confusion Matrix




Since it was the decision tree that had the highest average accuracy score, the confusion matrix on this slide is for that machine learning method.

Looking at the confusion matrix, one can see that the decision tree can determine the different classes, but has a problem around identifying false positives.



# Conclusion

- KSC LC-39A had the highest percentage of successful launches at 76.9%
- CCAFS SLC-40 had the lowest percentage of successful launches at 57.1%
- Typically landing had more success with heavier payload masses – specifically above 8,000kg.
- Launches improved with success as time went on as techniques improved with experience.
- The FT booster version typically had the most success and the v1.0 and v1.1 versions had the least success
- Orbit types like SO were never associated with a successful landing, whilst the ES-L1, GEO, HEO and SSO orbit types never failed.
- The most accurate classification model (Decision Tree) had an accuracy of above 80%, so its accuracy could be trusted.

# Thank You

-

Close