APPLIED DATA SCIENCE CAPSTONE PROJECT

Gareth Mouter

12/06/2025

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Data methodologies employed:

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) analysis using SQL
- EDA for Data visualization
- Building an interactive map with Folium
- Building a dashboard with Plotly Dash
- Predictive analysis using machine learning

Results:

- EDA results
- Interactive map and dashboard
- Predictive analysis results

INTRODUCTION

Project purpose:

With the proliferation of private space companies over the last 20 years, one has emerged as an industry leader; SpaceX. One of the principle reasons for this success has been its Falcon 9 rocket system, capable of running mission for $62 million, significantly cheaper than its competitors whose missions typically run for $165 million. The Falcon 9 system can achieve this low cost by safely landing and reusing its first stage, reducing costs.

The goal of this project to determine the cost for a launch by using predictive analysis to determine whether or not the first stage will land safely.

Questions we will explore during the project:

- How do various variables (for example, payload mass, number of launches, launch site etc), impact the success rate
- Are successful landings becoming more prevalent over time
- Which algorithm is best suited for binary classification in this instance

METHODOLOGY

Data collection methods employed:

- Using SpaceX Rest API
- Web scraping data from Wikipedia

Data wrangling methods employed:

- Data filtered
- Missing values dealt with
- Preparing data for binary classification using one hot encoding

EDA performed using SQL and data visualisation

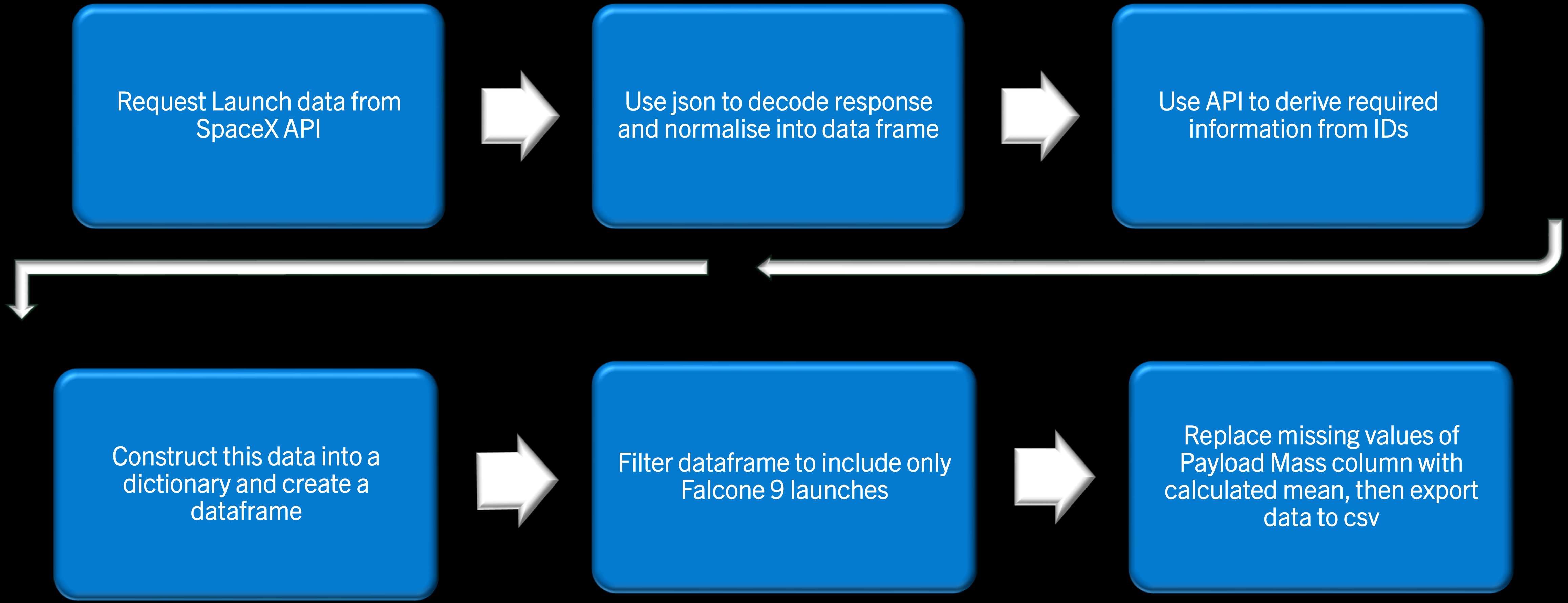
Created Folium based map and Plotly dashboard for additional data visualisation and analytics

Classification models used for predictive analysis:

- Logistic Regression, Decision Tree, k-Nearest Neighbours
- These were refined and tuned before calculating best performing model

DATA COLLECTION

SpaceX - API



Request Launch data from SpaceX API

Use json to decode response and normalise into data frame

Use API to derive required information from IDs

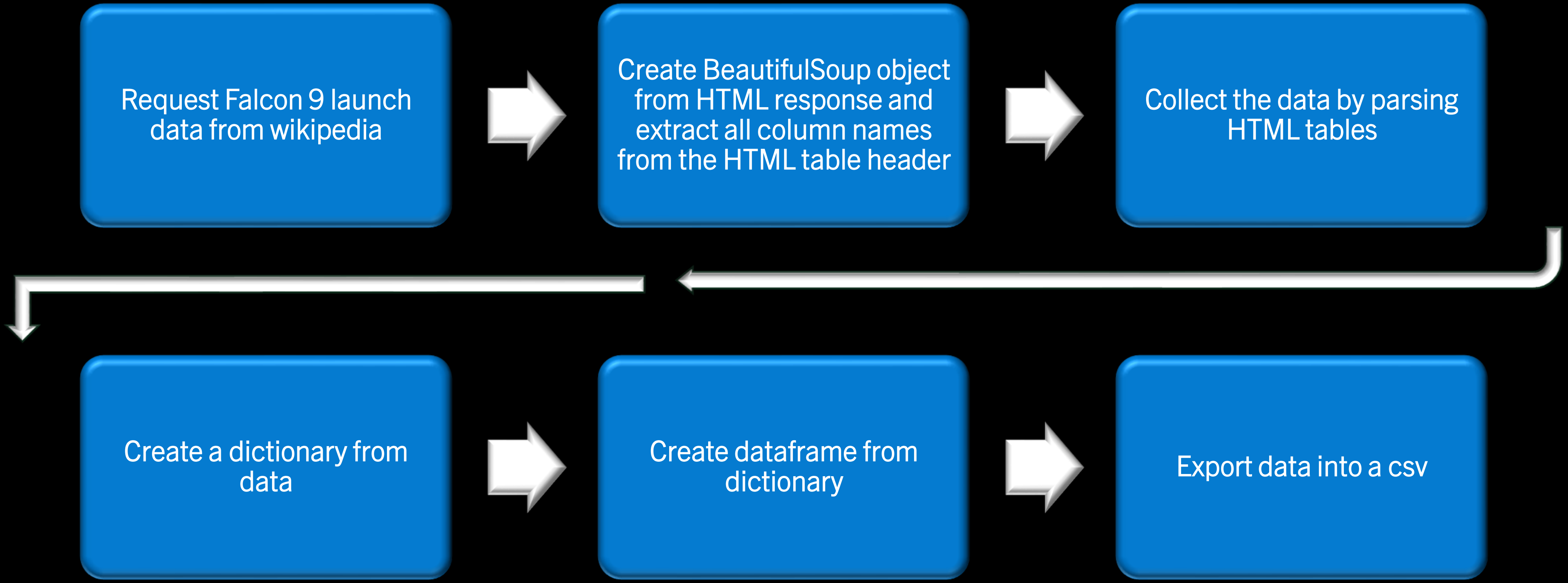
Construct this data into a dictionary and create a dataframe

Filter dataframe to include only Falcone 9 launches

Replace missing values of Payload Mass column with calculated mean, then export data to csv

DATA COLLECTION

Web Scraping



Request Falcon 9 launch data from wikipedia

Create BeautifulSoup object from HTML response and extract all column names from the HTML table header

Collect the data by parsing HTML tables

Create a dictionary from data

Create dataframe from dictionary

Export data into a csv

DATA WRANGLING

- Within the data there are multiple demarcations for successful or failed landings.
- These pertain to various circumstances, such as whether the landing took place on a barge or on land
- In order to simplify things, we have mostly converted these into 1 and 0 where 1 is a successful landing and 0 is not.
- Value counts was used to calculate the number of outcomes of the various different types
- Using ‘set’, a variable called ‘bad_outcomes’ was created which pulled together the various type of failed landings.
- Then, a list was created wherein if an outcome was within ‘bad_outcomes’, it was given a 0. Otherwise and it was given a 1

- The origin launch pad of each launch was also tallied, along with the purpose of the mission. For example, aiming for LEO or HEO

EDA WITH SQL

Various methods sql queries were employed to help gather useful information for the project. The information gleaned was as follows:

- Names of unique launch sites
- R records where the launch sites began with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version Falcon v1.1
- Date for first successful ground pad landing
- Names of boosters with success in drone ship with payload mass between 4-6000kg
- Total number of successful and failed mission outcomes
- Names of booster versions which have carried max payload
- Failed drone landings, alongside booster versions and launch site names for the months in 2015
- Rank of count of landing outcomes between 06/04/2010 and 20/03/2017

EDA- DATA VISUALIZATION

This primarily involved the creation of charts illustrating several trends and associations. These were as follows:

- Flight Number v Payload Mass
- Flight Number v Launch Site
- Orbit Type v Success Rate
- Flight Number v Orbit Type
- Payload Mass v Orbit Type
- Success Rate Yearly Trend

In addition OneHotEncoder was applied to create dummy variable to categorical columns

INTERACTIVE FOLIUM MAP

Mark all launch sites on a map:

- Added circular marker, pop-up label and test label of NASA Johnson Space Centre using its lat and long co-ordinates as a start location
- Repeated this with other launch sites, showing their geographical locations and proximity to equator

Colour Markers of launch outcomes for each launch site:

- Green markers (success) and red markers (failure) added using marker cluster to denote successes and failures for each launch site

Distances between a launch site to its proximities:
- Coloured lines added to show distances between the launch site (e.g CAFS LC-40) and its proximities such as railway, highway, closest city and coastline

PLOTLY DASH DASHBOARD

Launch sites dropdown list:

- Dropdown list to enable launch site selection added

Pie chart showing successful launches:

- Pie chart added which shows total successful launches count for all sites and the success /failure counts for individual sites

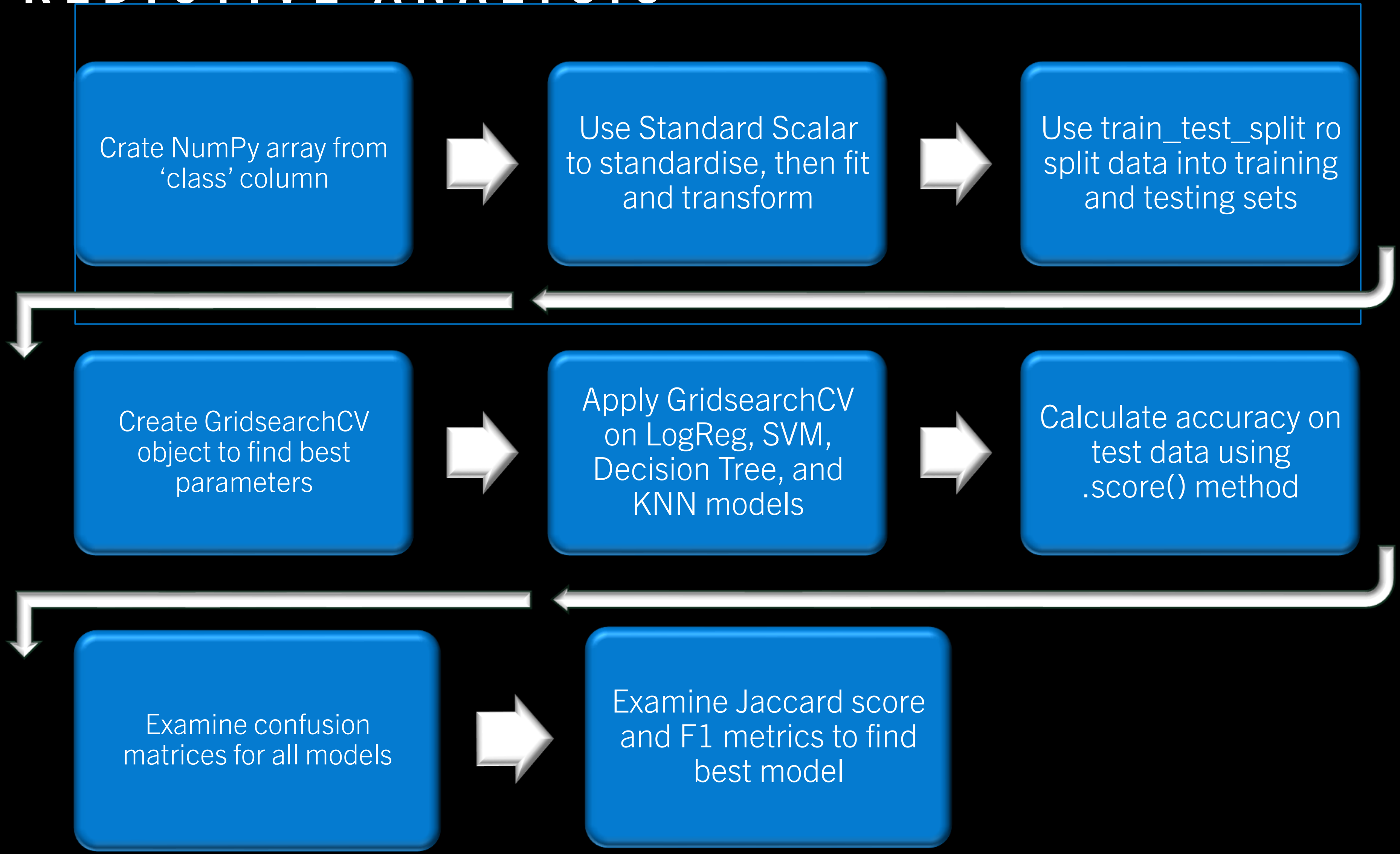
Slider of payload mass range:

- Slider added to allow selection of payload range

Scatter chart of payload mass vs success rate for different booster versions:

- Scatter chart added to show correlation between payload and launch success

PREDICTIVE ANALYSIS



Crate NumPy array from 'class' column

Use Standard Scalar to standardise, then fit and transform

Use train_test_split ro split data into training and testing sets

Create GridsearchCV object to find best parameters

Apply GridsearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculate accuracy on test data using .score() method

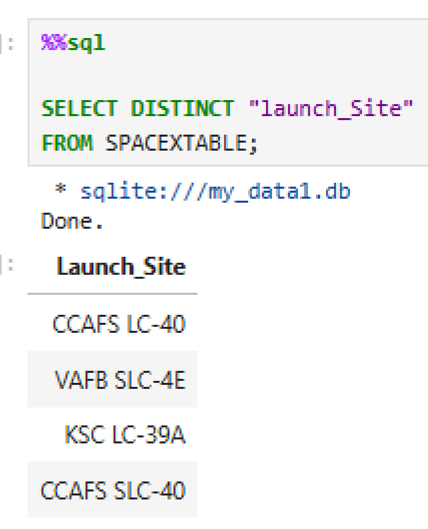
Examine confusion matrices for all models

Examine Jaccard score and F1 metrics to find best model

RESULTS

EDA with SQL

Names of unique launch sites in the space mission

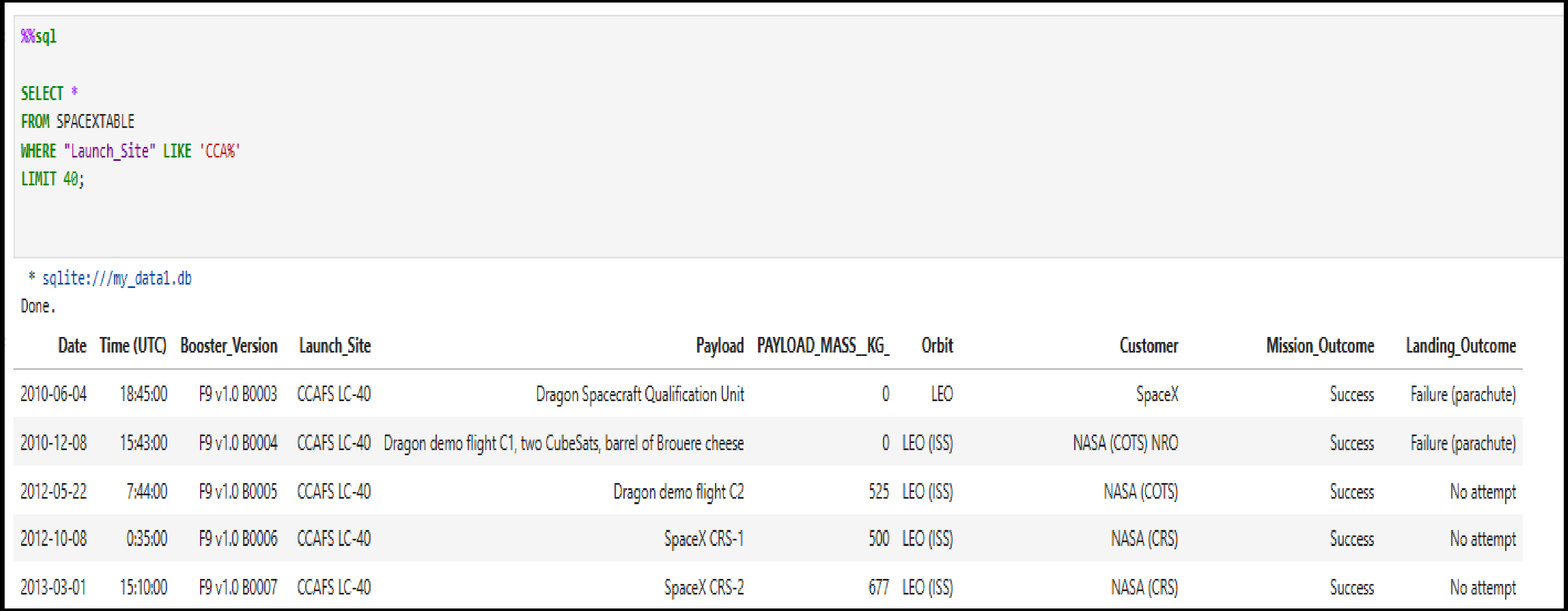


```
%%sql

SELECT DISTINCT "launch_Site"
FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Display records where launch sites begin with the string ‘CCA’

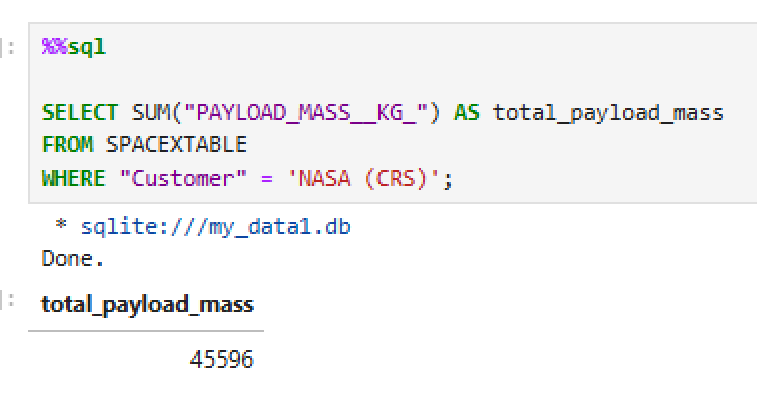


```
%%sql

SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 40;
```

```
 * sqlite:///my_data1.db
Done.
```

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

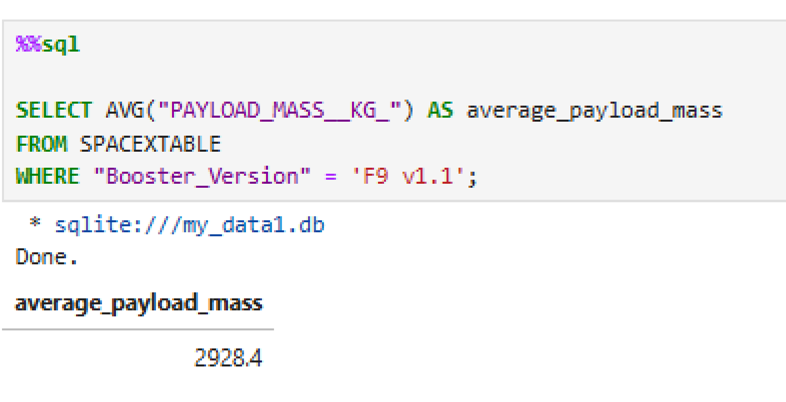
Total payload mass carried by boosters launched by NASA (CRS)



```
%%sql

SELECT SUM("PAYLOAD_MASS__KG_") AS total_payload_mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
total_payload_mass
             45596
```

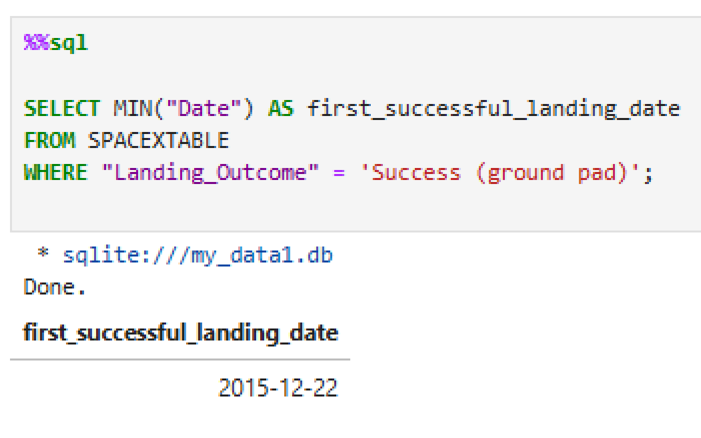
Average payload mass carried booster version F9 v1.1



```
%%sql

SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload_mass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
average_payload_mass
              2928.4
```

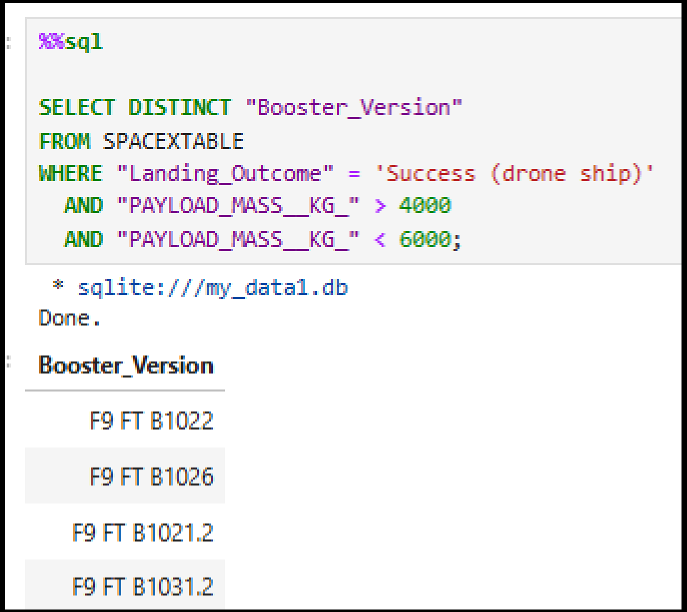
Date when first successful landing outcome on ground pad was achieved



```
%%sql

SELECT MIN("Date") AS first_successful_landing_date
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
first_successful_landing_date
                   2015-12-22
```

Names of boosters which have success in drone ship and have a payload mass greater than 4000 but less 6000

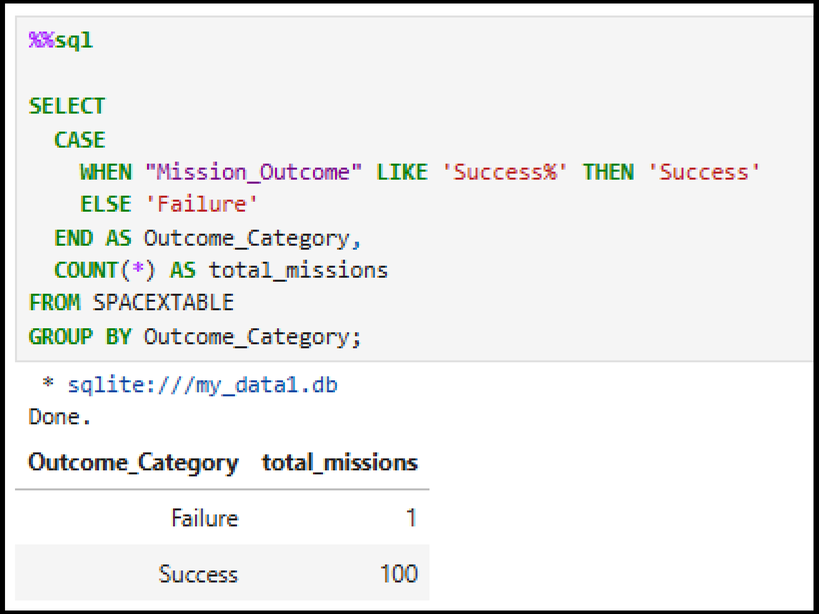


```
%%sql

SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
  AND "PAYLOAD_MASS__KG_" > 4000
  AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total number of successful and failed mission outcomes

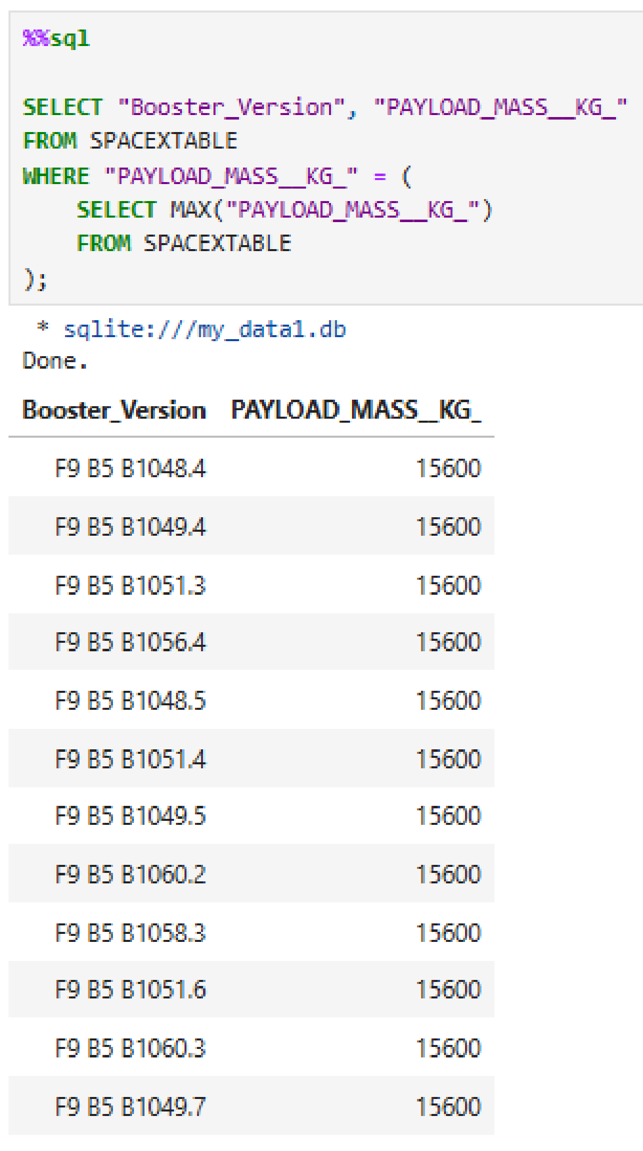


```
%%sql

SELECT
  CASE
    WHEN "Mission_Outcome" LIKE 'Success%' THEN 'Success'
    ELSE 'Failure'
  END AS Outcome_Category,
  COUNT(*) AS total_missions
FROM SPACEXTABLE
GROUP BY Outcome_Category;
```

```
 * sqlite:///my_data1.db
Done.
```

Outcome_Category	total_missions
Failure	1
Success	100

All booster versions that have carried max payload

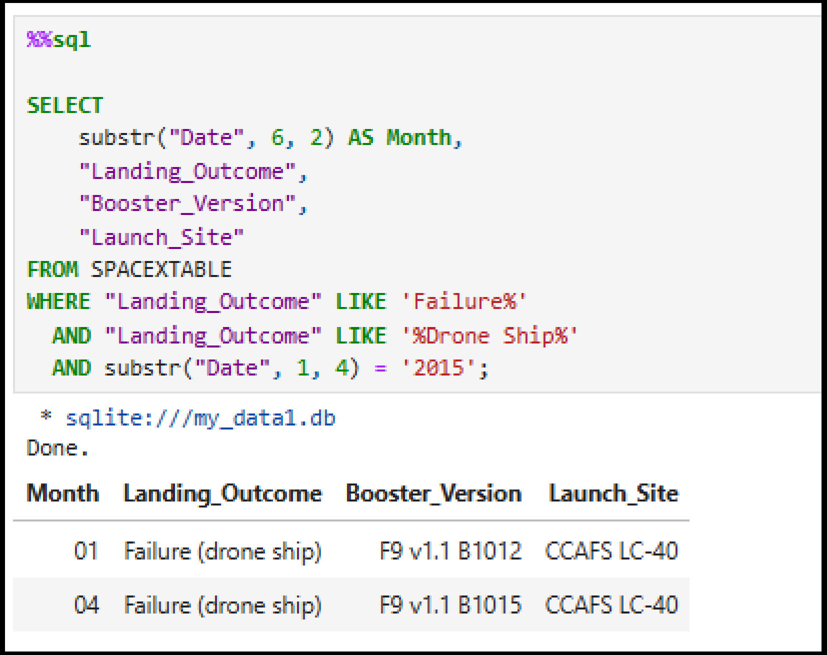


```
%%sql

SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015

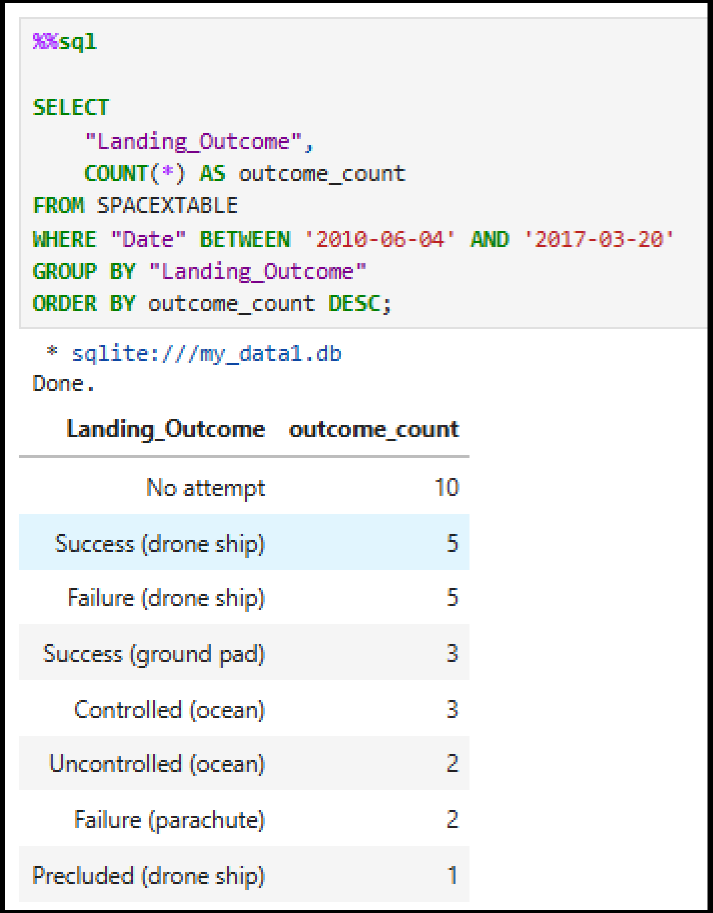


```
%%sql

SELECT
    substr("Date", 6, 2) AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Failure%'
  AND "Landing_Outcome" LIKE '%Drone Ship%'
  AND substr("Date", 1, 4) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

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

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.



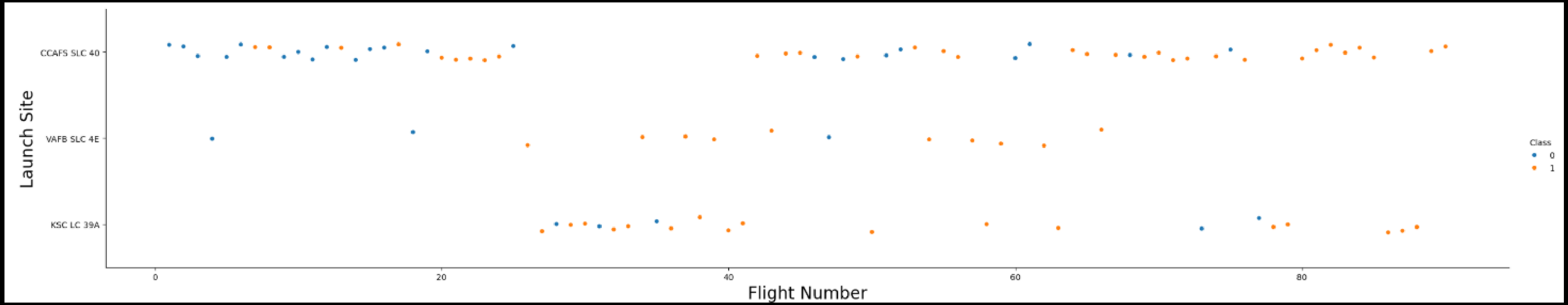
```
%%sql

SELECT
    "Landing_Outcome",
    COUNT(*) AS outcome_count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

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

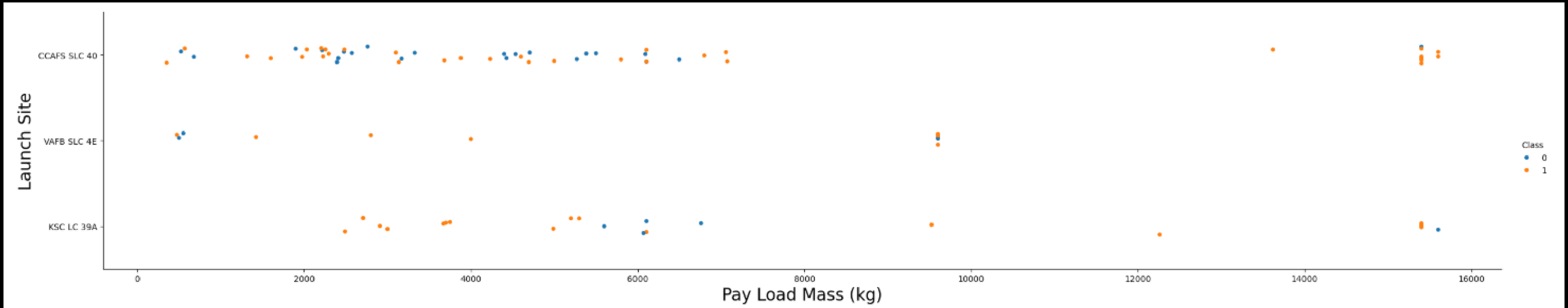
EDA WITH VISUALIZATION

Flight Number vs. Launch Site



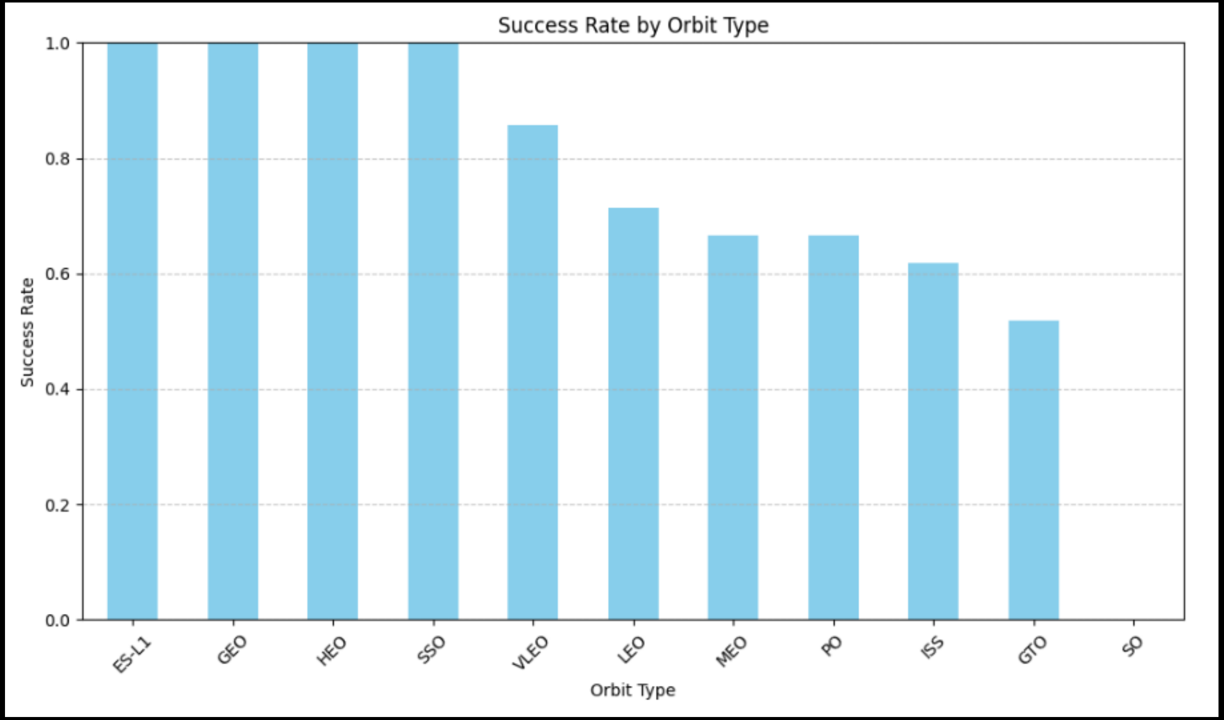
- A clear trend of failures to successful outcomes can be seen
- CCAFS has largest total number flights and the lowest success rate

Payload Mass vs Launch Site



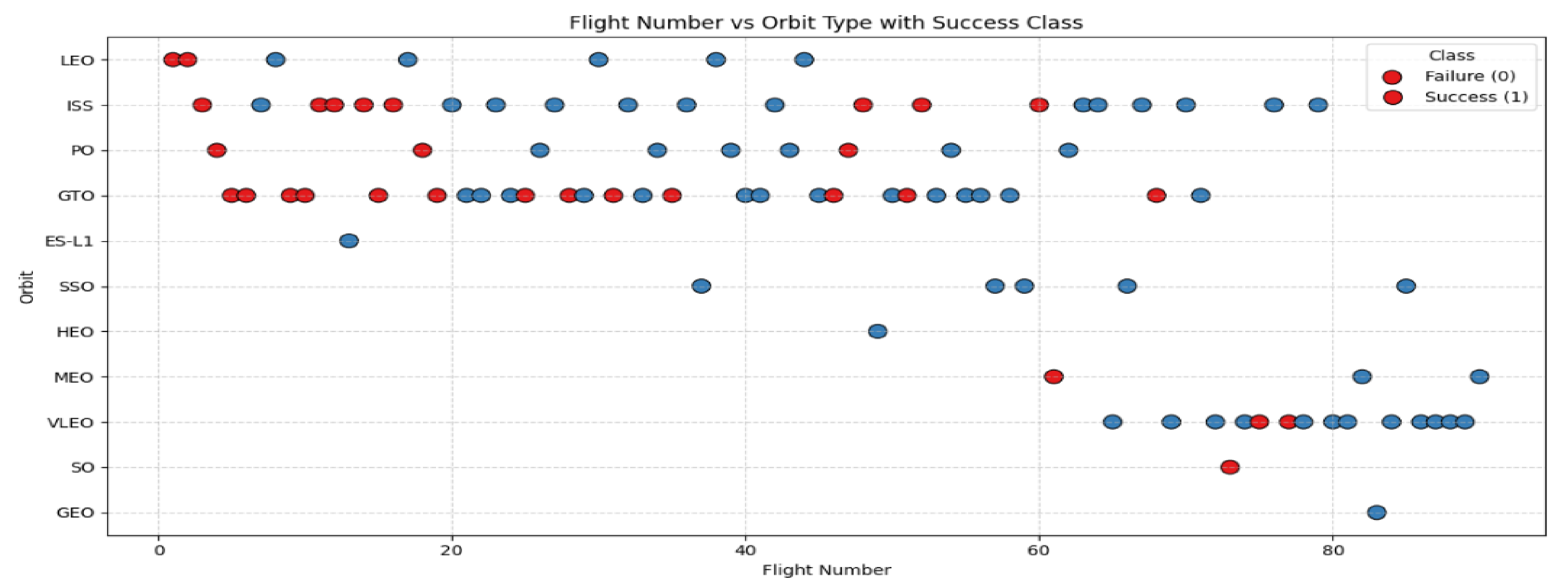
- The higher the payload mass, the greater the chance of successful outcome
- Payload launches over 8000kg have a very high success rate
- No failures under 5500kg for KSC

Success Rate by Orbit Type



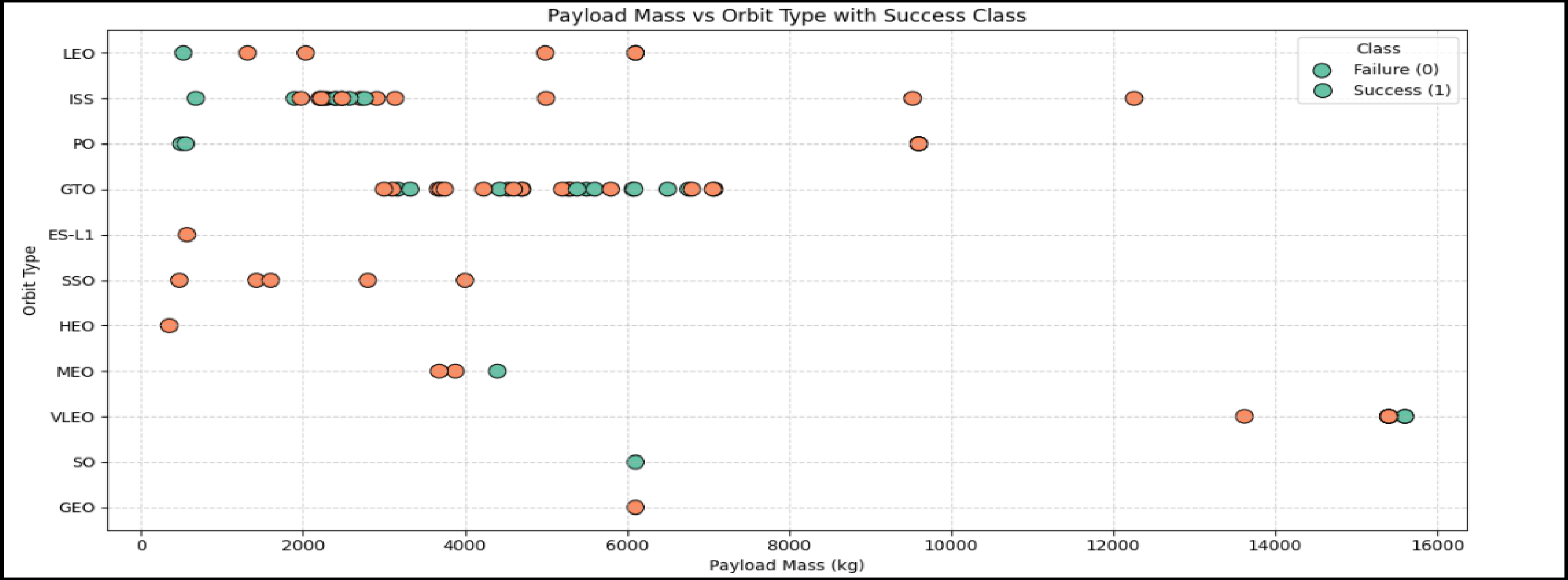
- ES-L1, GEO, HEO and SSSO all have a 100% success rate
- SO is the only orbital type with no successful outcomes

Flight Number vs Orbit Type with Success Class



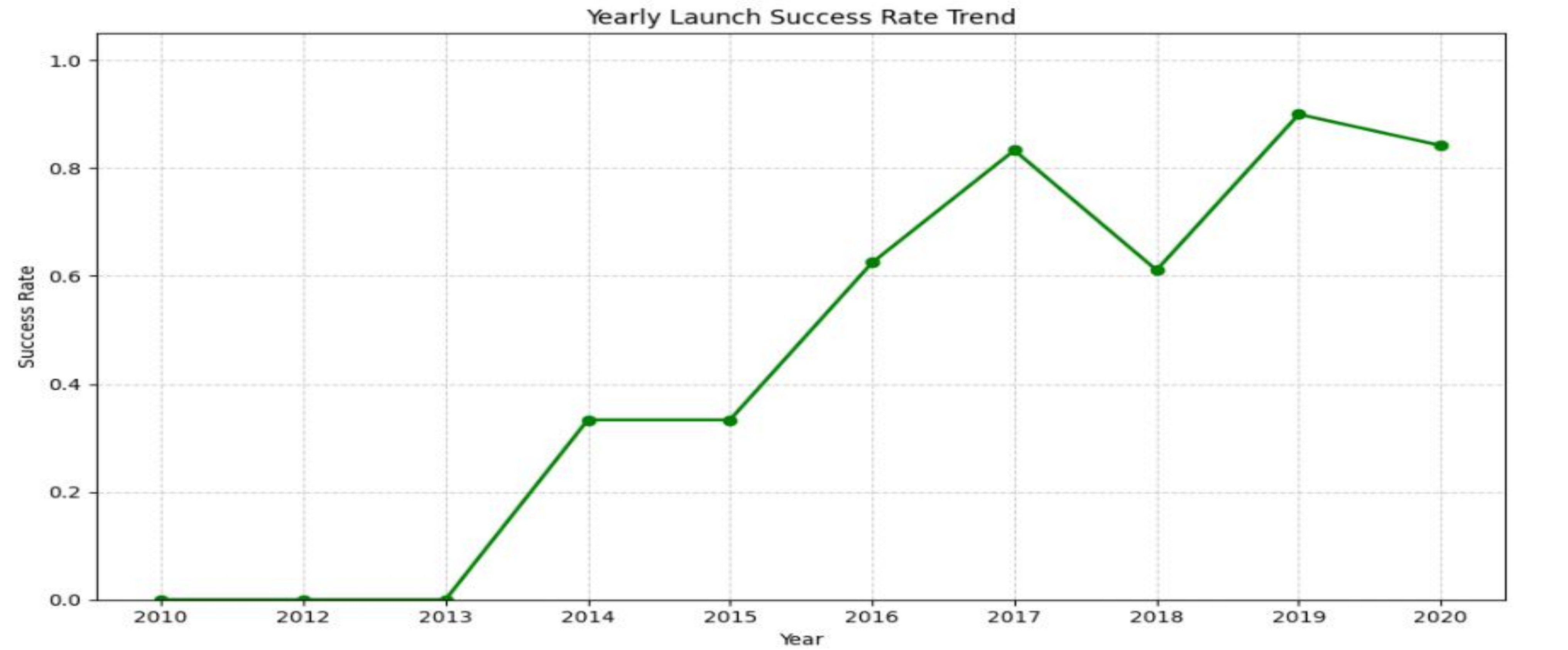
The number of flights for the most part increases success outcome, though this is more pronounced in some orbits than in others

Payload Mass vs Orbit Type with Success Class



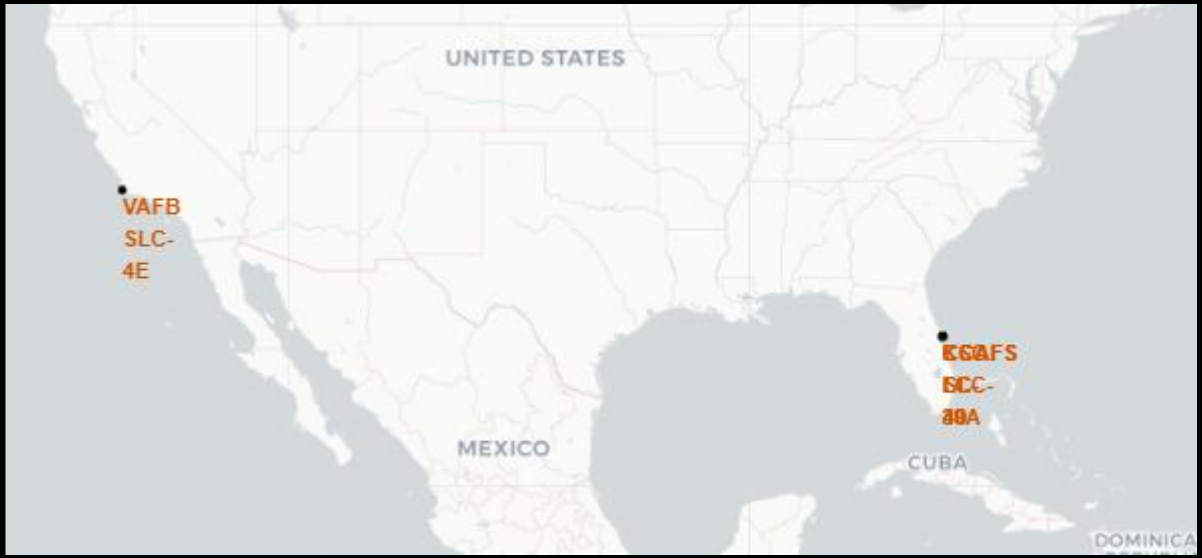
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Yearly Launch Success Rate Trend



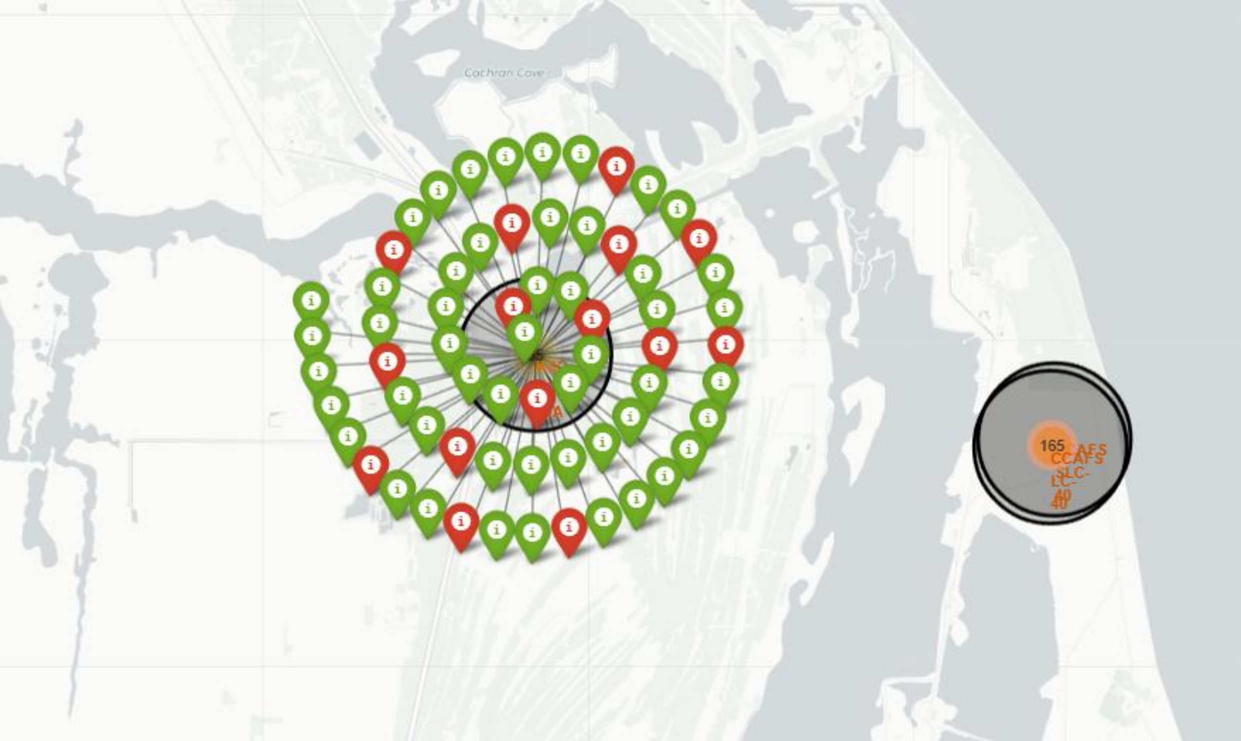
- Success rate is on a positive trend, though may be slowing as of 2018, more data required going forward to confirm

INTERACTIVE MAP WITH FOLIUM



- We can see that all launch sites are in the south of US near the equator. The purpose of this is to take advantage of the rotation of the earth at the equator (where it is quickest) and so less delta V is required to achieve orbit, reducing fuel requirements and boosting payload capacity
- They are also on the coast, particularly clustered in Florida, where they can launch over the Atlantic

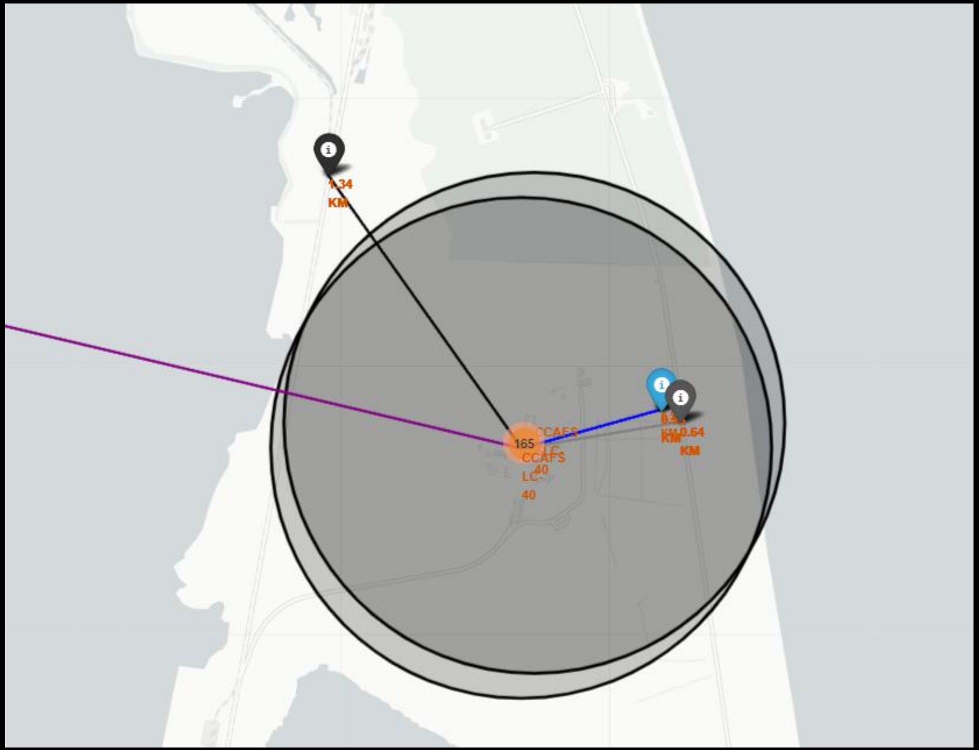
COLOUR LABELLED LAUNCH RECORDS



Green indicates a successful outcome, whilst red denotes failure

The launch site pictured is KSC

DISTANCE FROM CCAFS LC-40 TO LOCAL PROXIMITIES



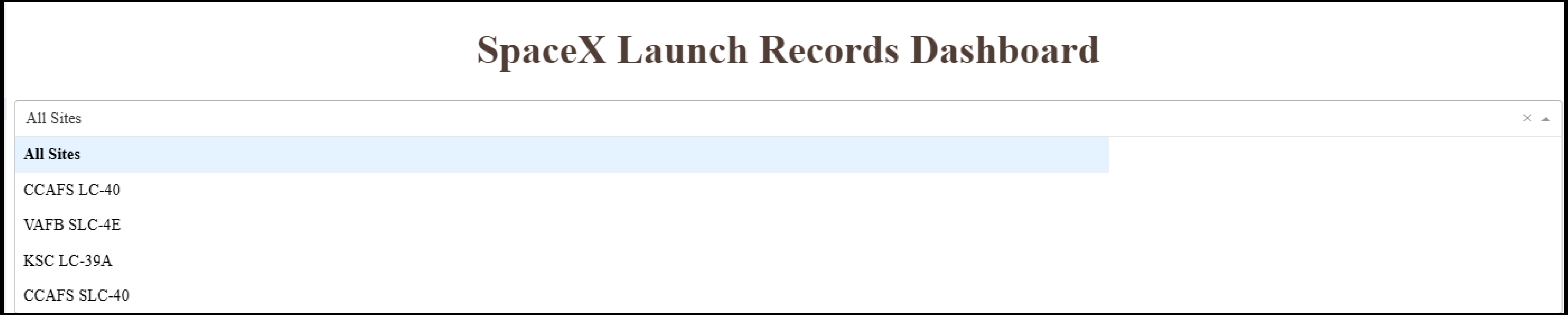
This map allows us to see the distance of CCAFS LC-40 to local points of interest such as:

- Railway
- Highway
- Coastline
- Nearest city

This is useful for estimating threat potential to local populations, as well as amenity availability for workers and transportation of materials etc.

BUILD A DASHBOARD WITH PLOTLY DASH

SpaceX Launch Records Dashboard



SpaceX Launch Records Dashboard

All Sites

All Sites

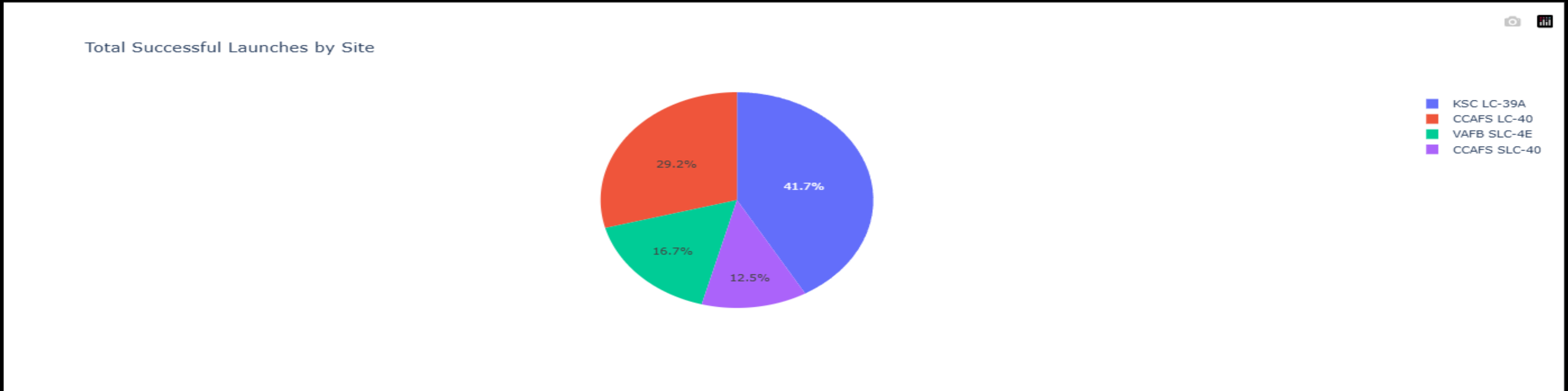
CCAFS LC-40

VAFB SLC-4E

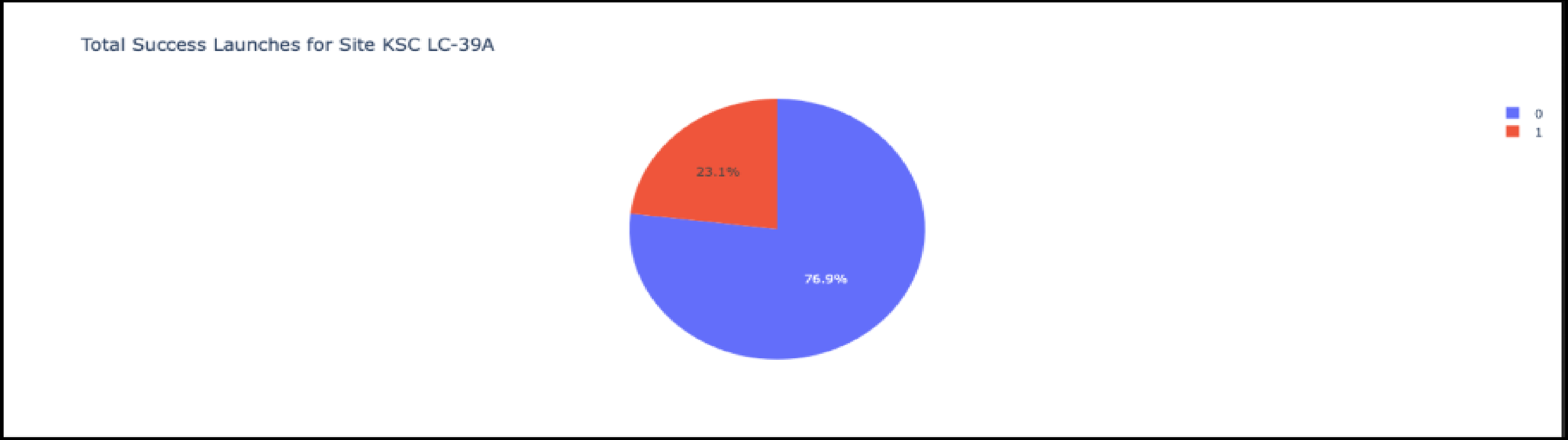
KSC LC-39A

CCAFS SLC-40

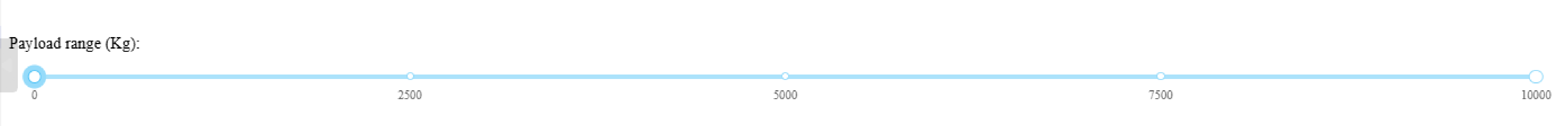
PIE CHART FOR COUNT OF LAUNCHES



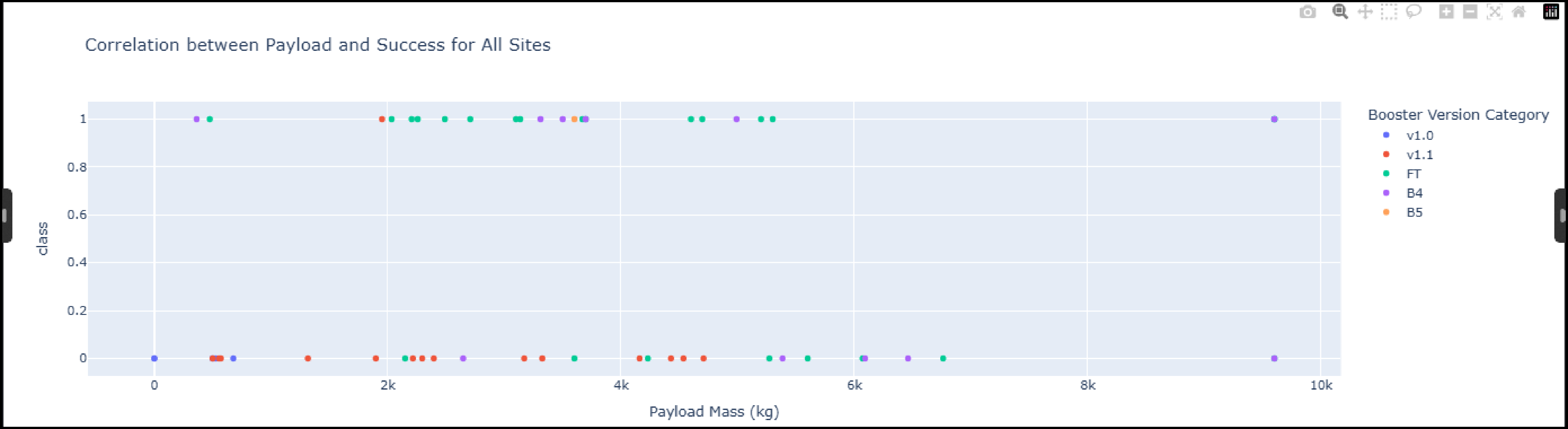
- The above is for all sites. Individual sites could also be selected via the dropdown, as seen below



PAYLOAD RANGE SLIDER

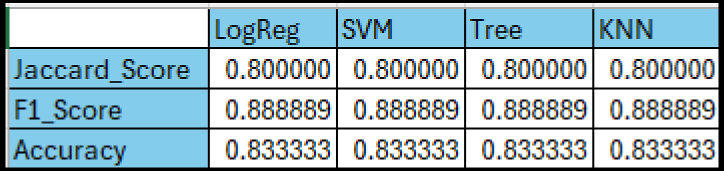


PAYLOAD MASS VS. LAUNCH OUTCOME



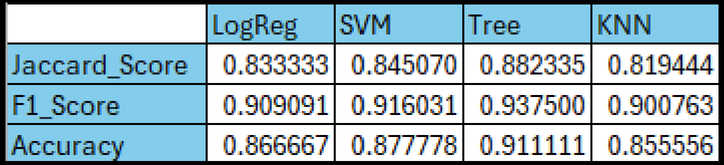
PREDICTIVE ANALYSIS

Scores and Accuracy of Test Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

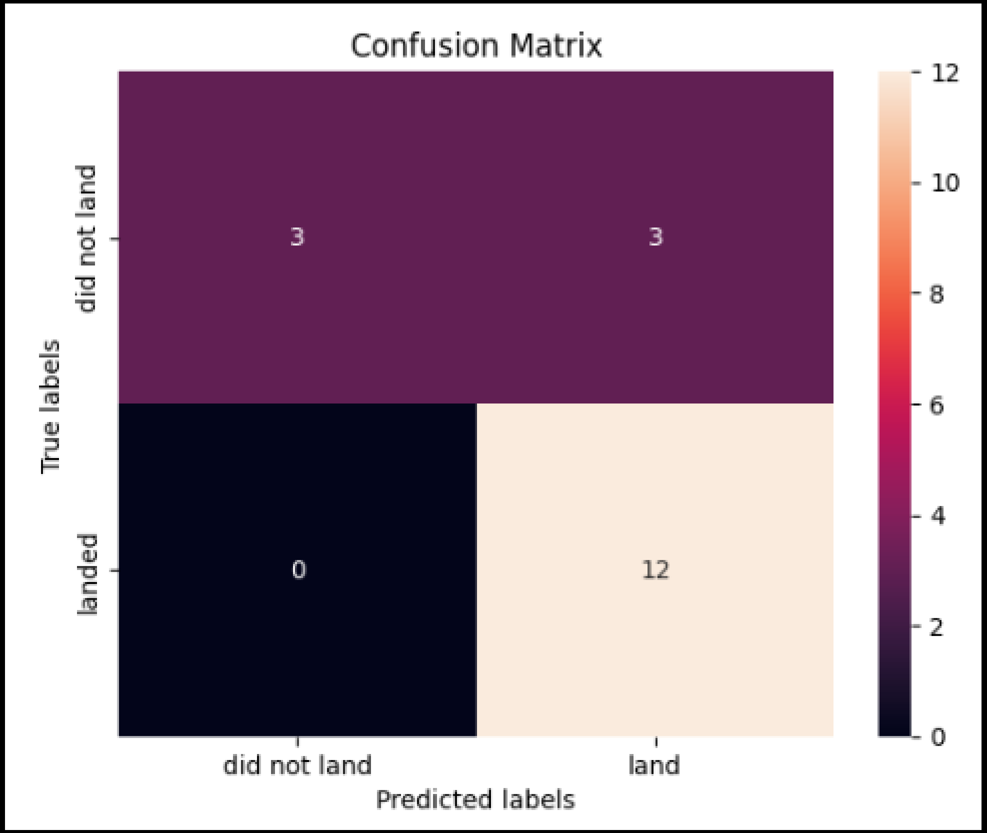
Scores and Accuracy of Entire Data Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882335	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

- As can be seen, using the test data, a best method cannot be confirmed
- This is most likely due to small sample size
- When applied to the entire data set, the Tree method has both the highest scores and accuracy

CONFUSION MATRIX



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.
Overview:
True Positive - 12 (True label is landed, Predicted label is also landed)
False Positive - 3 (True label is not landed, Predicted label is landed)

CONCLUSION

The main takeaways from this study are:

- A high payload mass indicates a greater chance of success
- Differing orbital goals also has an impact on the likely success of a landing
- Successes have become more frequent over time
- Most launch sits are close to the equator and coast
- The Decision Tree Model provided the best results for predicting landing success

APPENDIX

I would like to give special thanks to the coursera and IBM team, as well as the fellow student who marked this presentation