IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

By Flavio Aguirre
23/04/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions

Executive Summary

This project aimed to estimate the economic viability of SpaceX launches by predicting the probability of a successful landing of the Falcon 9 first stage.

The dataset was provided by IBM, and the project followed all stages of the data science lifecycle: from data acquisition via the SpaceX API, through data cleaning and transformation (ETL), to exploratory data analysis and predictive modeling. The development was structured across eight interdependent notebooks.

The final model utilized the K-Nearest Neighbors algorithm and achieved an accuracy of 83%. Multiple influential factors in the landing success were identified, confirming that this outcome can be predicted based on measurable patterns. This approach demonstrates the real-world applicability of machine learning in the aerospace industry.

Introduction

SpaceX has revolutionized the aerospace industry with its reusable Falcon 9 launch system, reducing launch costs from over $165 million to approximately $62 million. This reusability offers a greater economic and technological advantage.

Predicting the landing success of the Falcon 9 first stage is essential for mission planning and cost estimating. It also provides valuable information for stakeholders in space logistics and investment.

This project seeks to build a machine learning model to predict landing outcomes using data provided by IBM and the SpaceX API. By applying the full data science workflow (data acquisition, processing, analysis, and modeling), we identify the key factors behind successful landings and offer a practical predictive tool for the aerospace industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Requests were made to the SpaceX API and web scraping was also used on Wikipedia.
- Perform data wrangling
  - Pandas and Numpy were used for data management.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models
  - Different classification models were implemented, giving KNN as the best algorithm with 83% accuracy.

Data Collection

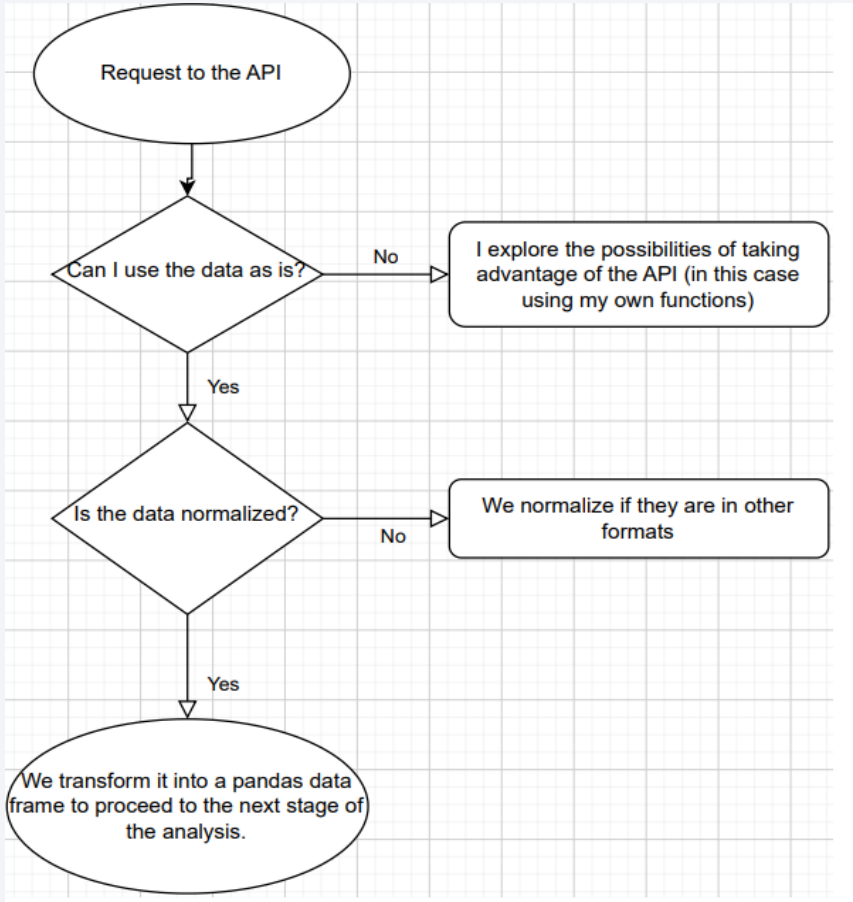
The data for the following final project was collected using two different techniques. One was web scraping, which compiled historical Falcon 9 launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches."
Where we created a web archive of Falcon 9 launch logs with `BeautifulSoup` and extracted an HTML table of Falcon 9 launch logs from Wikipedia:
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

But the dataset we kept to follow the project was the one collected through the SpaceX API, through GET requests. We also performed some basic data manipulation and formatting operations.

Data Collection – SpaceX API

In addition to using the request module, a series of auxiliary functions were defined that allowed the API to be used to extract information using ID numbers from the release data. The work was carried out as outlined in the following flowchart.

You can find the complete step-by-step extraction instructions at:
Link to the notebook



Data Collection - Scraping

In this project, we created a web archive of Falcon 9 launch logs using BeautifulSoup and extracted an HTML table from Wikipedia. The columns were then scraped using webscraping techniques to create a Pandas data frame. As shown here:

Link to the notebook

2020 [edit]

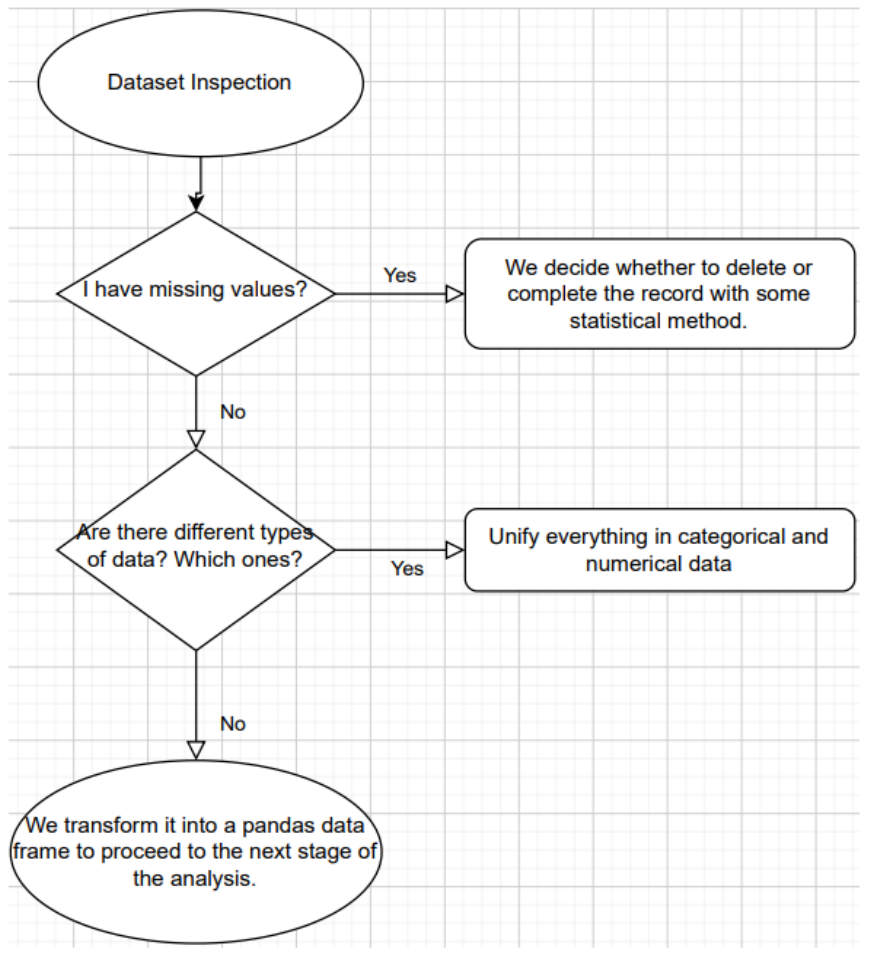
In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
	29 January 2020,	F9 B5 ♺	CCAFS,						Success

Data Wrangling

We analyzed the data to determine the type of data in the dataset and to see if we found any missing, duplicate, or null values. The following flowchart was followed.

You can also view the process at:
Link to the notebook



EDA with Data Visualization

For this approach, three types of graphs were used using pandas, matlplotlib, and seaborn:

- Scatter plot: To visualize the relationship between two numerical variables.
- Bar chart: To compare categorical values.
- Line chart: To observe trends over time.

The full analysis and plots are available here:
Link to the notebook

EDA with SQL

- The SPACEXTABLE table was created by filtering records with non-zero dates.
- Unique launch sites (Launch_Site) were listed.
- The payloads sent by NASA (CRS) were summed.
- The average payload for F9 v1.1 was calculated.
- The date of the first successful landing was obtained.
- Successful and failed landings were counted.
- The booster version with the largest payload was identified.
- Pad landing failures during 2015 were listed with relevant details.

The complete and detailed process here:
Link to the notebook

Build an Interactive Map with Folium

The following **Folium** objects were used to build the map:

- folium.Map: to create the interactive base map.
- folium.Circle: to draw informative circles over specific locations.
- folium.Marker: to visually mark the launch sites.
- folium.MarkerCluster: to group nearby markers and reduce visual clutter.
- MousePosition: to display real-time coordinates as the mouse moves over the map.
- folium.DivIcon: to add custom-styled HTML/CSS text on the map.
- folium.PolyLine: to draw lines connecting locations.
- folium.FeatureGroup: to organize and manage related map elements together.

Consult:

Link to the notebook

Build a Dashboard with Plotly Dash

Components used in the **Dash** web application:

- **dcc.Dropdown**: to select a launch site and display success vs. failure rates in a pie chart.
- **dcc.RangeSlider**: to filter payload range and analyze its relation to success rate.
- **dcc.Graph**: to render the pie and scatter plots.
- **Dash callbacks**: enable dynamic interaction between interface elements.

The following graphics were used:

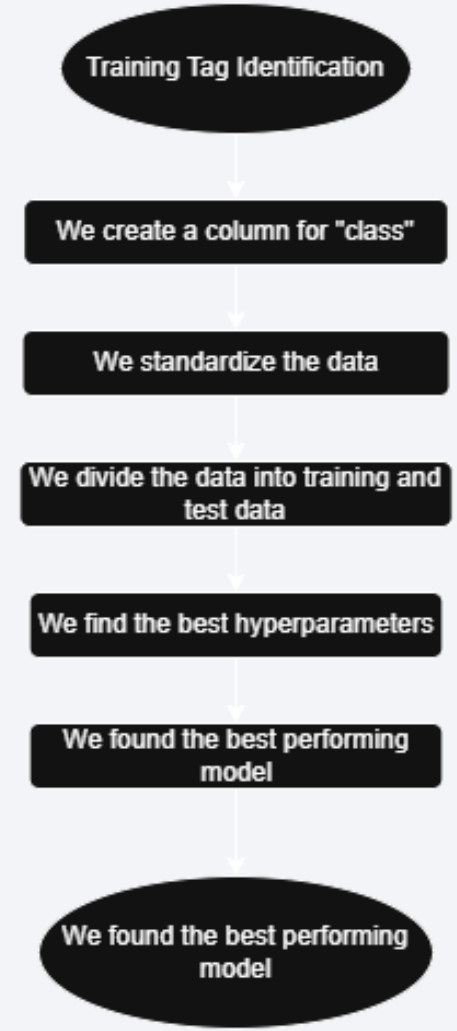
- **Pie Chart**: shows the proportion of successful and failed launches by site.
- **Scatter Plot**: illustrates the relationship between payload and launch outcome.

The complete process in:
Link to the notebook

Predictive Analysis (Classification)

- We performed exploratory data analysis and determined training labels.
- We created a column for "class".
- We standardized the data.
- We split the data into training and test data.
- We found the best hyperparameters for SVM, classification trees, and logistic regression.
- We found the best performing method using test data.

The complete process in:
Link to the notebook

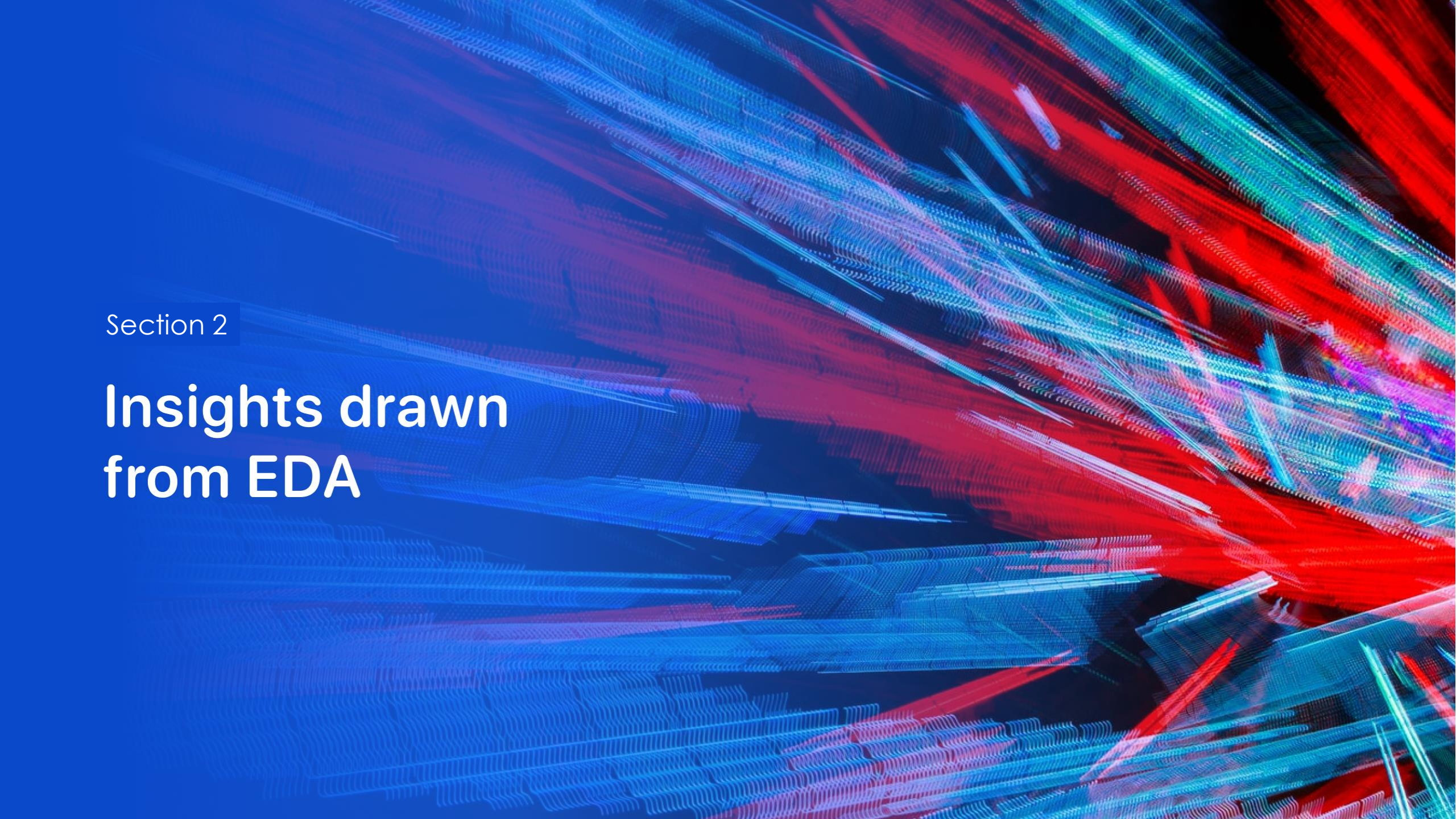


Results

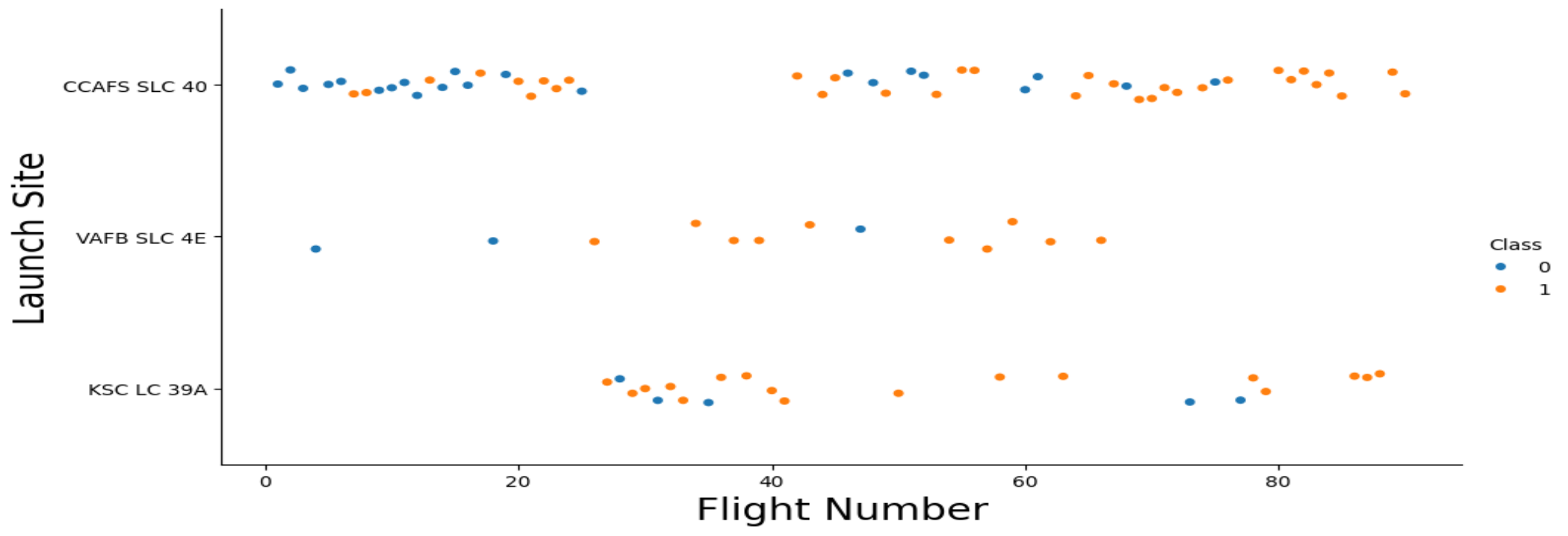
Exploratory data analysis results:
- We were able to corroborate that there are certain parameters that directly influence the success or failure of any landing.

Interactive analysis demonstration in screenshots:
- We were able to visually determine how the launch site directly affects the success rate of any mission.

Predictive analysis results:
- A classification model (KNN) was developed with an accuracy of 83% for predicting the success of the first stage landing of SpaceX rockets.

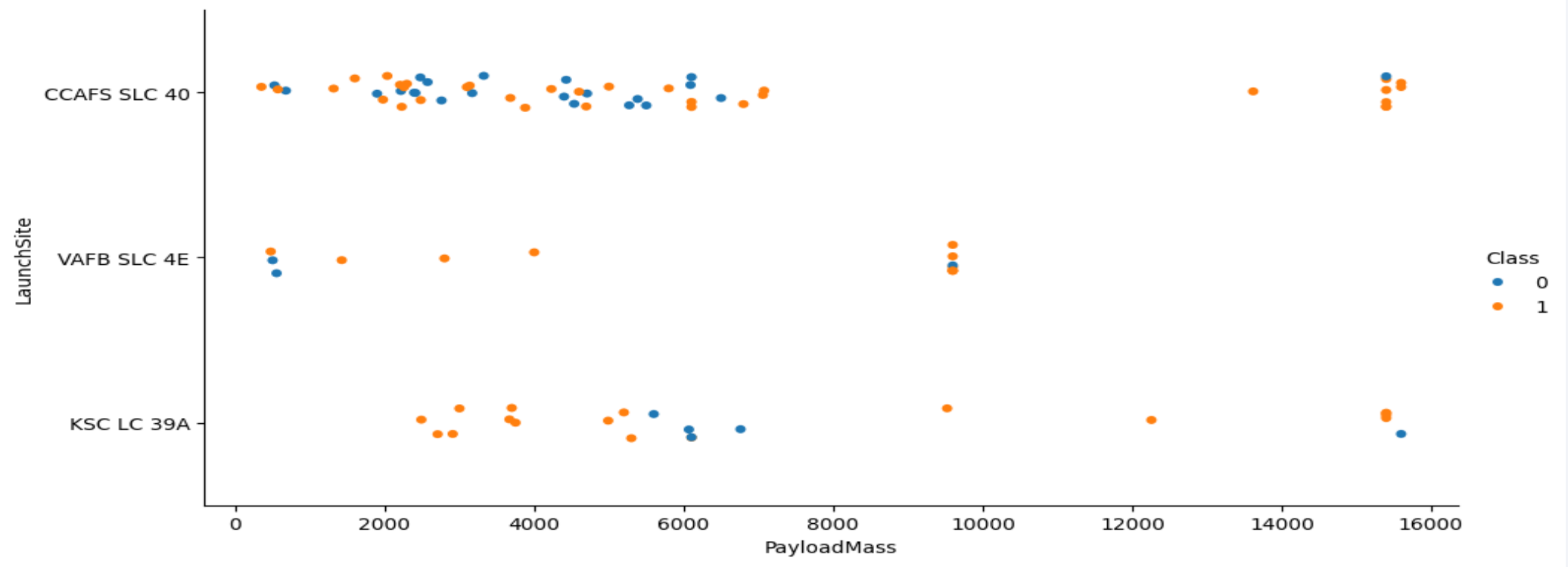

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



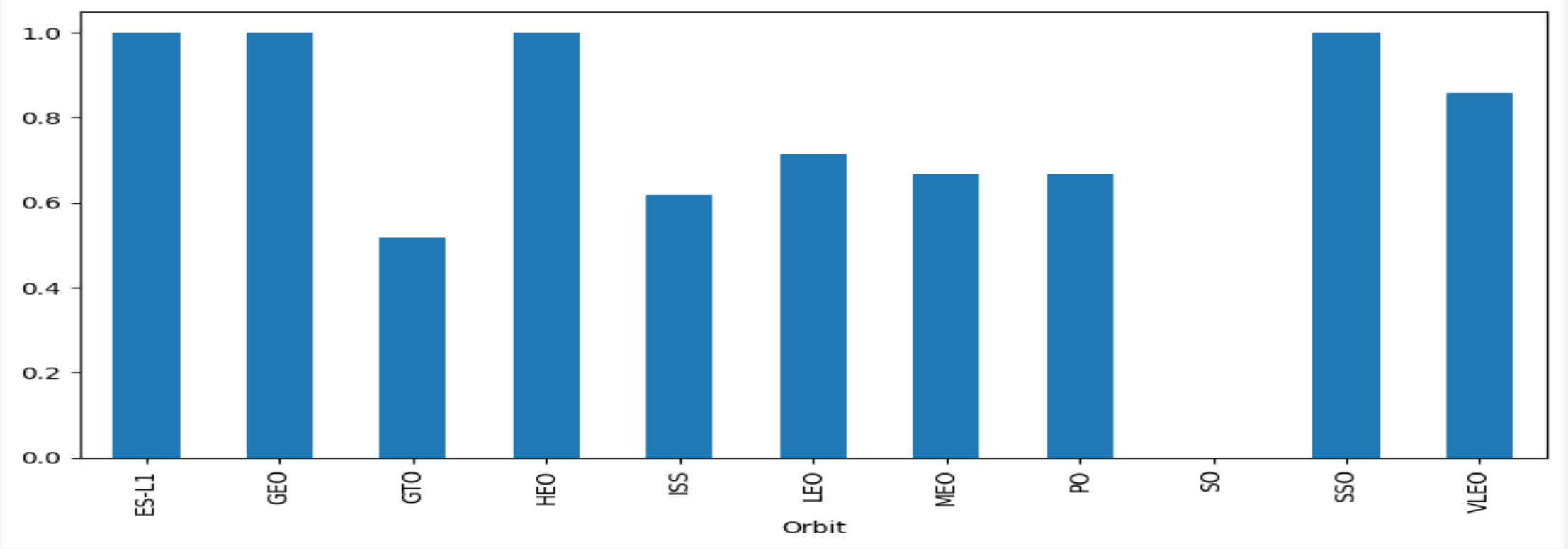
We see how as the flight number increases with respect to the launch site, the success rate of the missions at each launch site increases.

Payload vs. Launch Site



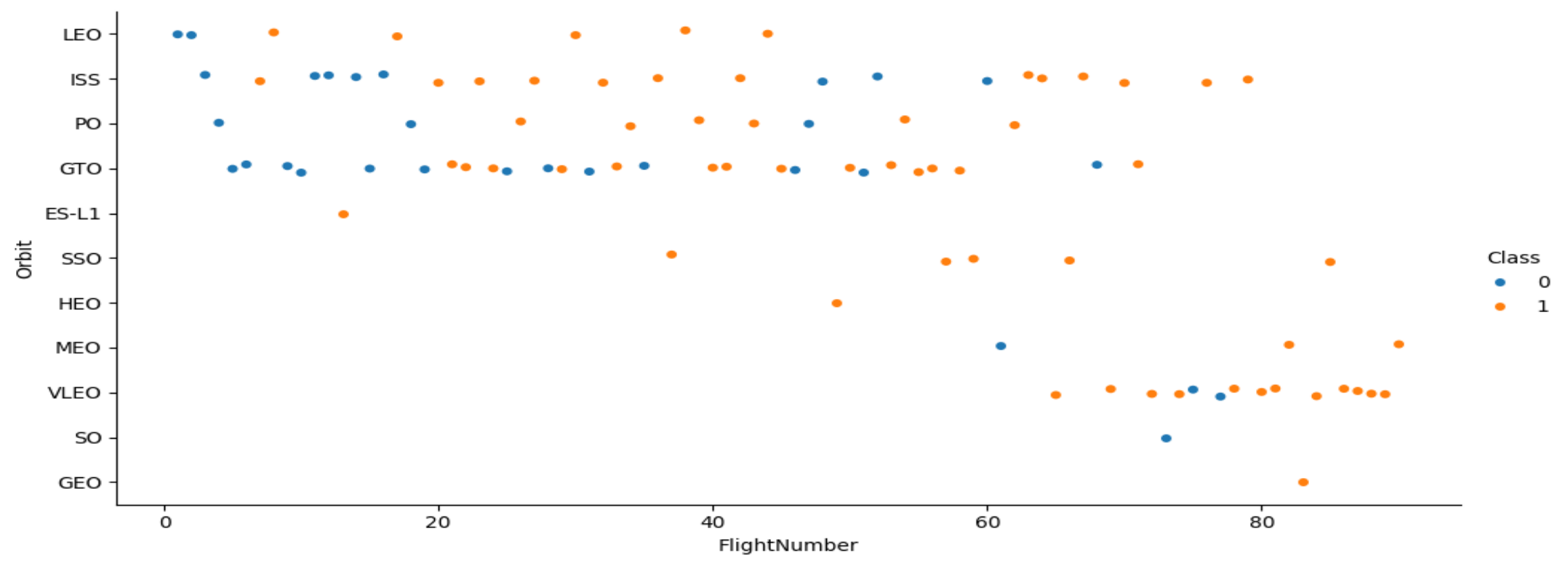
We can see that as the weight of the payload increases relative to the launch site, the number of successful missions increases.

Success Rate vs. Orbit Type



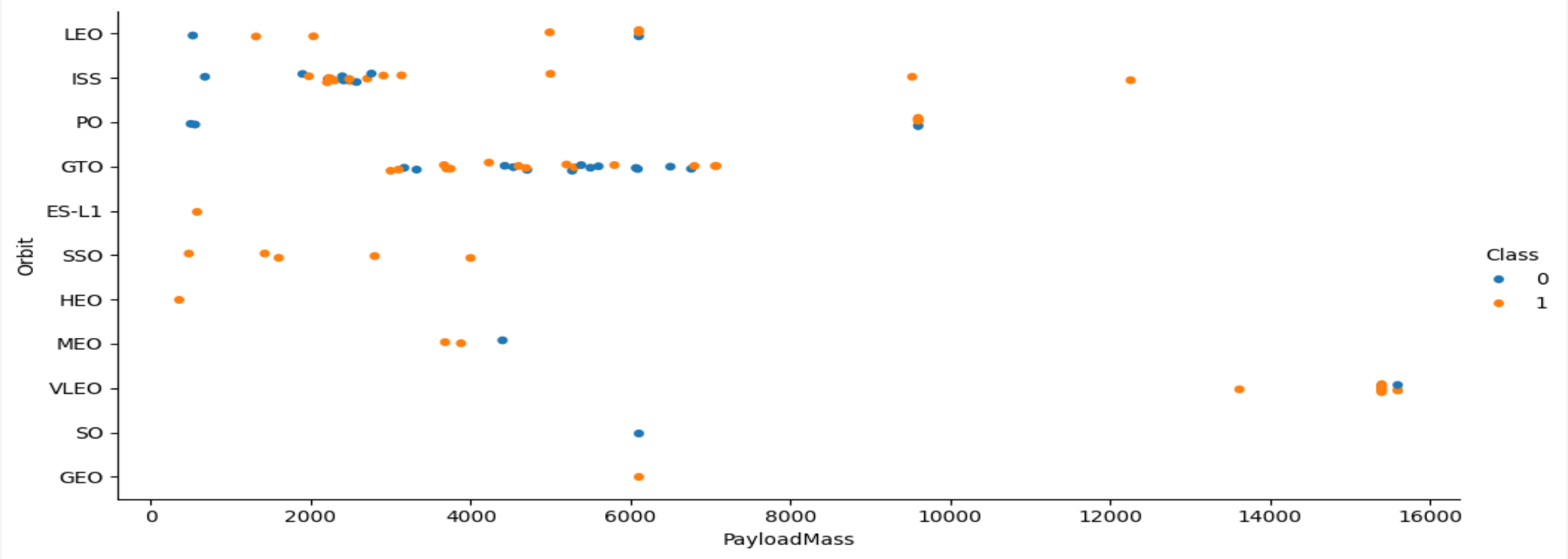
It can be seen that depending on the orbit to which the mission leaves, its success rate increases considerably compared to the others.

Flight Number vs. Orbit Type



We see how some orbits are much more common than others, and how some orbits seem to have a higher success rate. It's worth noting that as the flight number increases, the success rate of first-stage recovery increases.

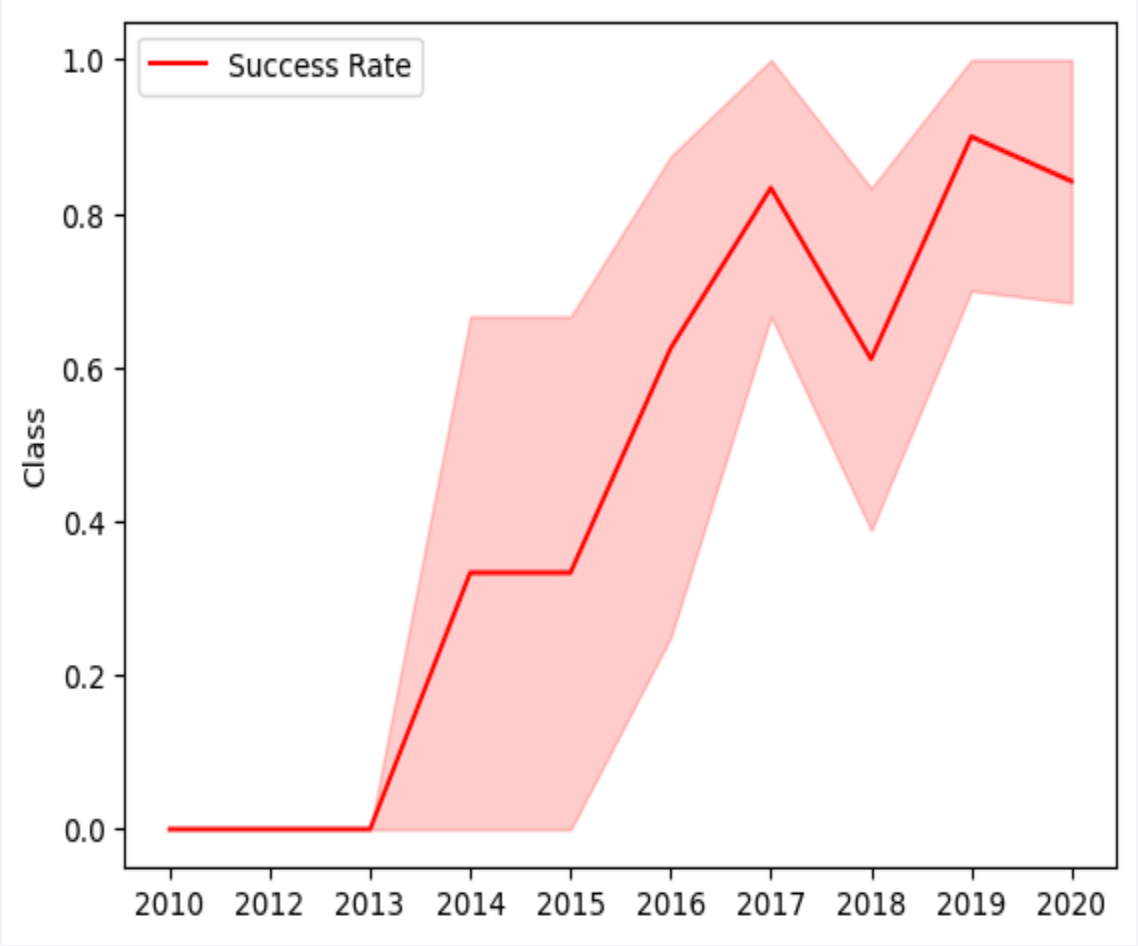
Payload vs. Orbit Type



We distinguish how the mass of the payload varies considerably depending on the orbit. But that in the same way the success rate remains considerable regardless of the weight of the payload, not so in all missions.

Launch Success Yearly Trend

This graph perfectly summarizes SpaceX's evolution toward success in rocket reusability:
-They went from 0% success rate in the early years to maintaining a rate above 85% from 2017 onward.
-It reflects progressive learning, technological consolidation, and continuous improvement.



All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE ;
```

Python

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

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

With this SQL query we can visualize the 4 different launch locations that exist in our Database

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

Python

```
 * 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

With this query we were able to see the first 5 records with launch site CCAFS LC-40

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS 'PAYLOAD IN KG'
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

Python

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

PAYLOAD IN KG
45596

The total weight of all payloads for all missions can be calculated using the SQL query shown in the image. The result is **45596 kg**.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVERAGE PAYLOAD IN KG'
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

Python

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

AVERAGE PAYLOAD IN KG
2928.4

We also see the average payload weight per launch through the SQL query, resulting in **2928.4 kg.**

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) AS 'First successful landing on the platform'
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success%';
```

Python

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

First successful landing on the platform
2015-12-22

We were able to identify the date of the first successful landing on the platform, as shown in the SQL query in the image. The result is December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT
    ROW_NUMBER() OVER (ORDER BY Booster_Version) AS num,
    Booster_Version,
    PAYLOAD_MASS__KG_ AS Weight_kg
FROM SPACEXTABLE
WHERE
    PAYLOAD_MASS__KG_ > 4000
    AND PAYLOAD_MASS__KG_ < 6000
    AND Landing_Outcome LIKE 'Success%';
```

Python

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

num	Booster_Version	Weight_kg
1	F9 B4 B1040.1	4990
2	F9 B4 B1043.1	5000
3	F9 B5 B1046.2	5800
4	F9 B5 B1047.2	5300
5	F9 B5 B1048.3	4850
6	F9 B5 B1051.2	4200
7	F9 B5 B1058.2	5500
8	F9 B5B1060.1	4311
9	F9 B5B1062.1	4311
10	F9 FT B1021.2	5300

Thanks to this SQL query we can list all launches with a payload ranging from 4000 to 6000kg which have been successful.

Total Number of Successful and Failure Mission Outcomes

```
%%sql

SELECT SUM(Landing_Outcome LIKE 'Success%') AS 'Total successful landings',
       SUM(Landing_Outcome LIKE 'Failure%') AS 'Total failed landings'
FROM SPACEXTABLE
```

Python

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

Total successful landings	Total failed landings
61	10

As in the image query, we see that we have 61 successful launches vs 10 that were classified as failures.

Boosters Carried Maximum Payload

As we see in the image below, with that SQL query we were able to list the 12 launches that have carried their maximum load.

```
%%sql
SELECT
    ROW_NUMBER() OVER (ORDER BY Booster_Version) AS num,
    Booster_Version,
    PAYLOAD_MASS__KG_ AS Carga_Maxima_Kg
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

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

num	Booster_Version	Carga_Maxima_Kg
1	F9 B5 B1048.4	15600
2	F9 B5 B1048.5	15600
3	F9 B5 B1049.4	15600
4	F9 B5 B1049.5	15600
5	F9 B5 B1049.7	15600
6	F9 B5 B1051.3	15600
7	F9 B5 B1051.4	15600
8	F9 B5 B1051.6	15600
9	F9 B5 B1056.4	15600
10	F9 B5 B1058.3	15600
11	F9 B5 B1060.2	15600
12	F9 B5 B1060.3	15600

2015 Launch Records

It can be seen how only two missions were classified as failures this year, with a difference of 3 months and at the same launch site.

```
%%sql
SELECT
  substr(Date,6,2) AS month,
  Customer AS customers,
  Booster_Version AS version,
  Launch_Site AS site,
  Landing_Outcome AS outcome,
  PAYLOAD_MASS__KG_ AS payload
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date,0,5)='2015';
```

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

month	customers	version	site	outcome	payload
01	NASA (CRS)	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2395
04	NASA (CRS)	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	1898

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

To conclude this analysis, we list the number of positive outcomes, sorting them by number of landing outcomes (such as failure (drone) or success (ground platform)) between June 4, 2010, and March 20, 2017, in descending order.

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

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

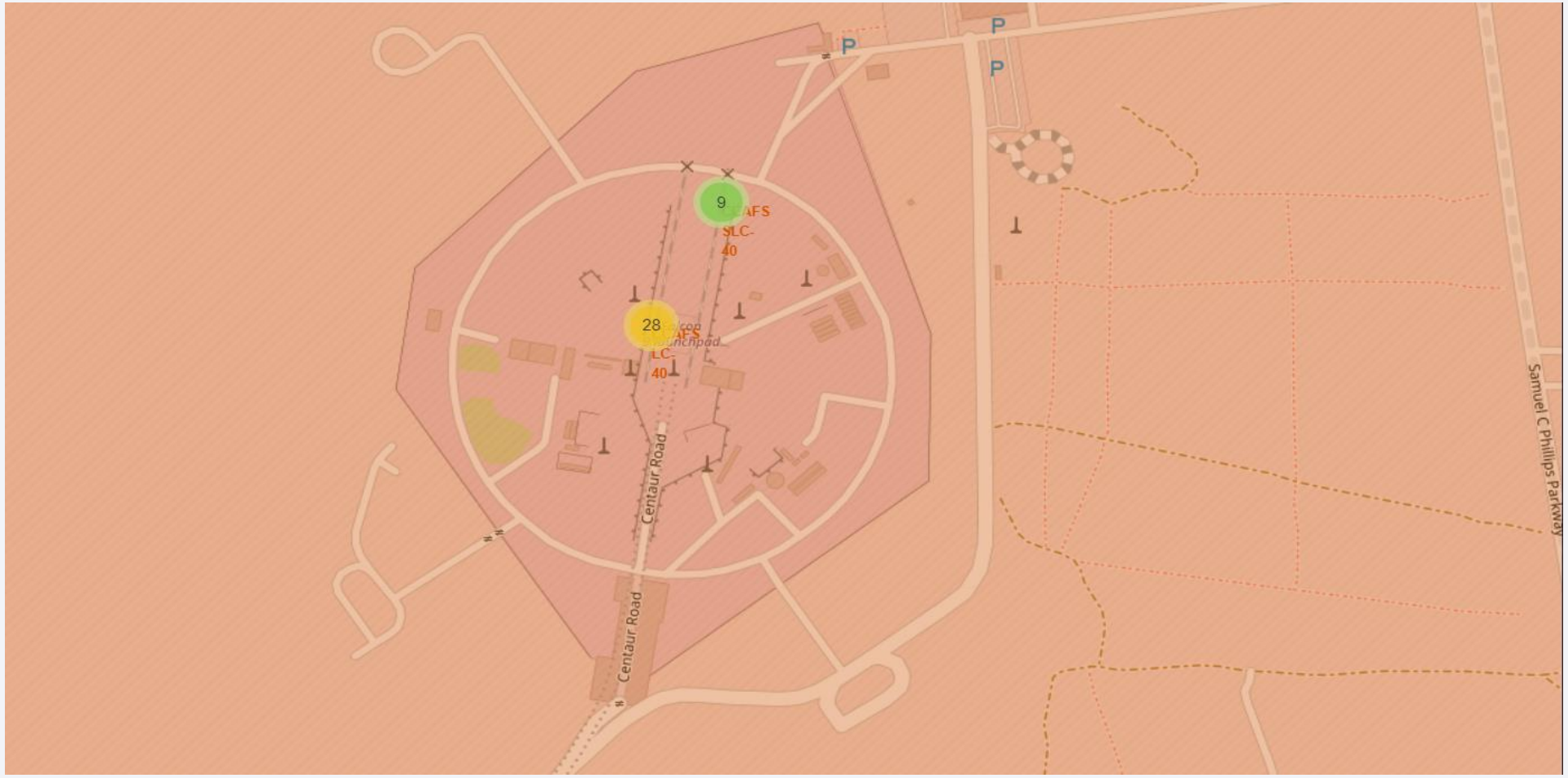
Landing_Outcome	Number_Landings
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

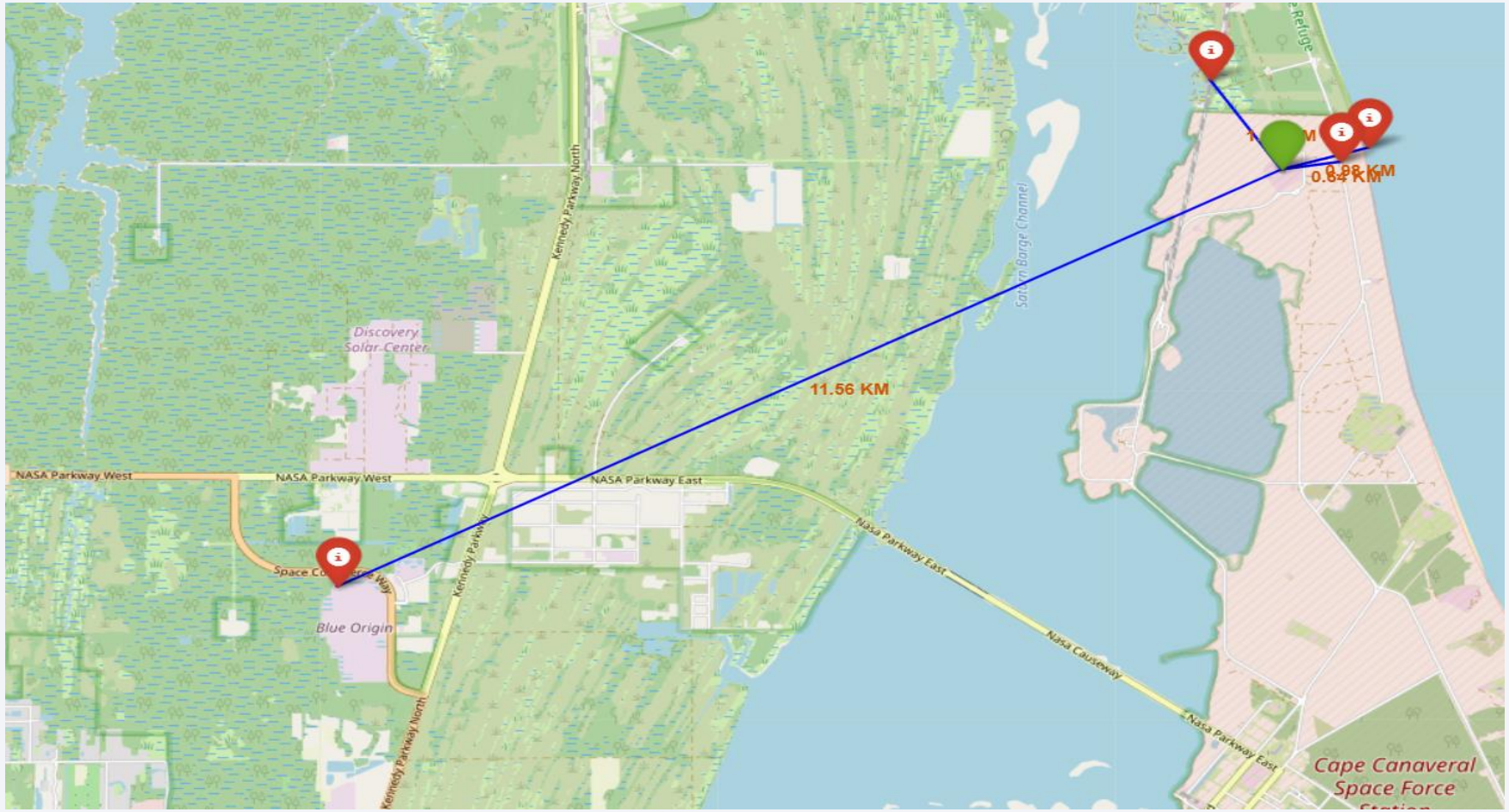
Launch Sites Proximities Analysis

1- All launch sites with Folium

2- The launch results labeled by colors.



3- Launch location and distance to surrounding areas (cities, railways, etc.)



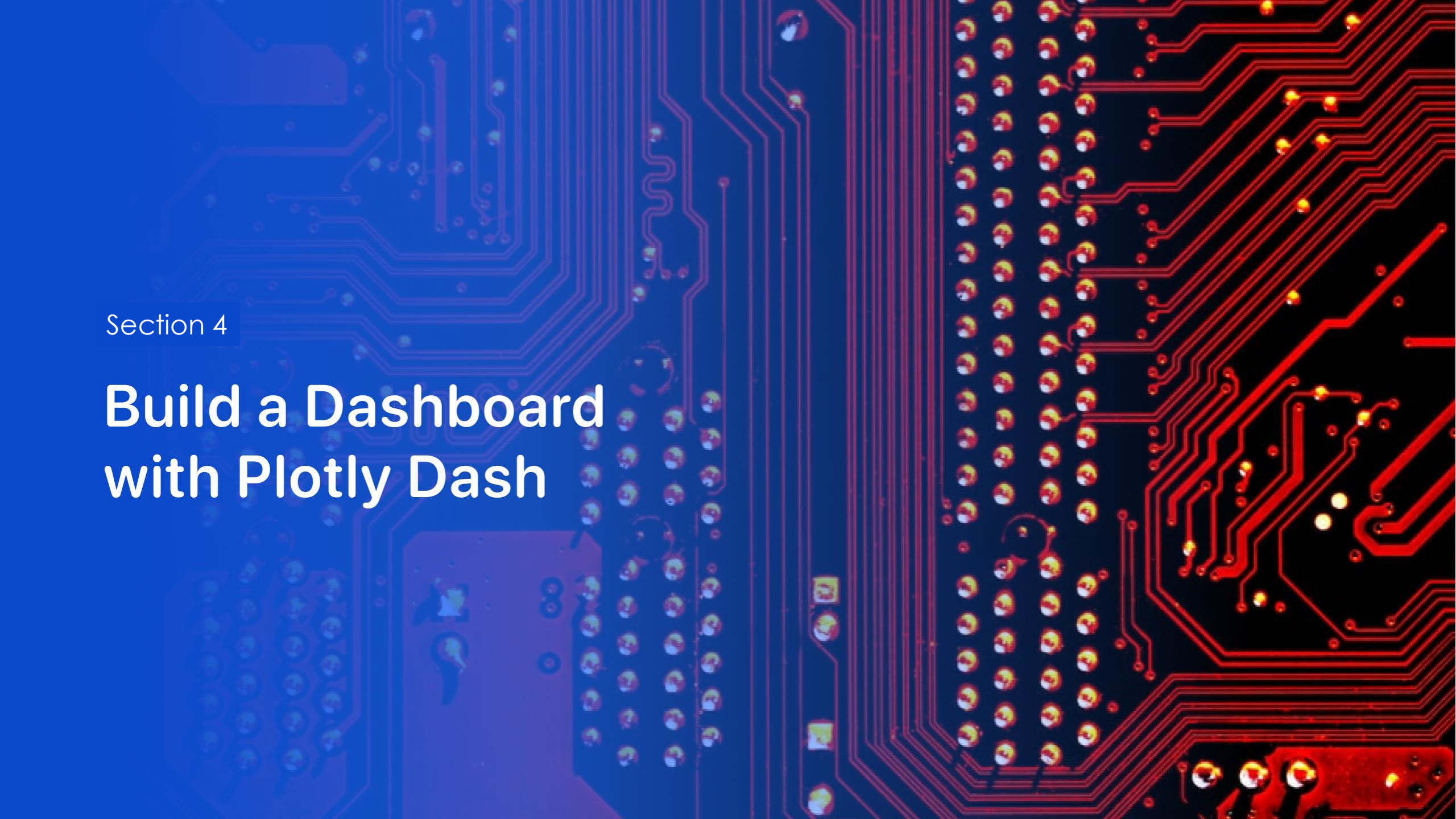
Analysis result

1- From the first image we can see the location of the launch sites on a global map.

2- The second image allows us to note, with colored labels, the number of successful launches in green, and the number of failed launches in orange, along with their respective locations.

3- We see that they are all close to the coasts and railways and far from cities and highways.

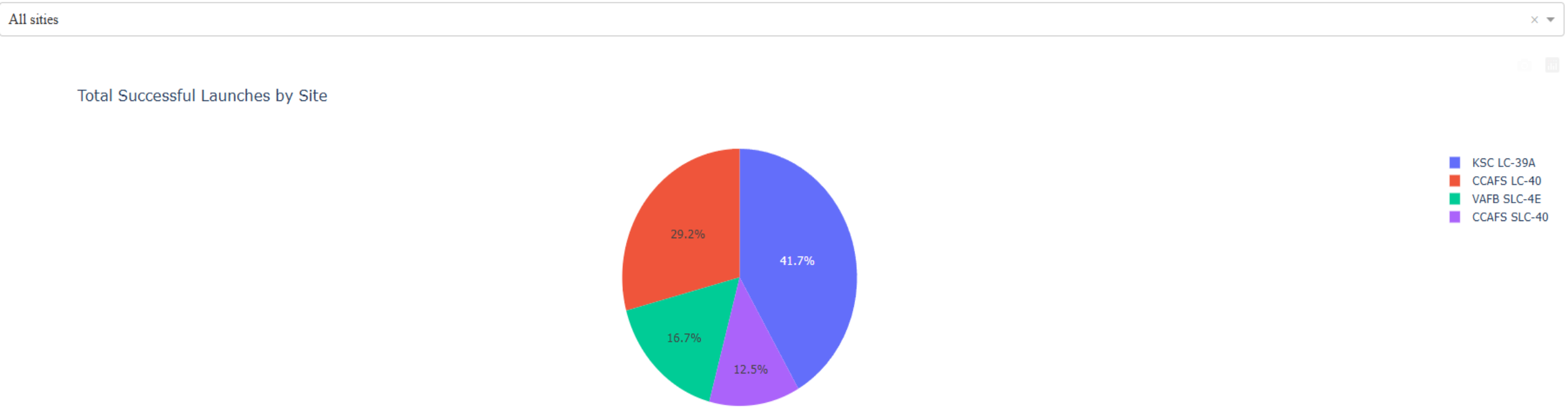
References:
Link to the notebook



Section 4

Build a Dashboard with Plotly Dash

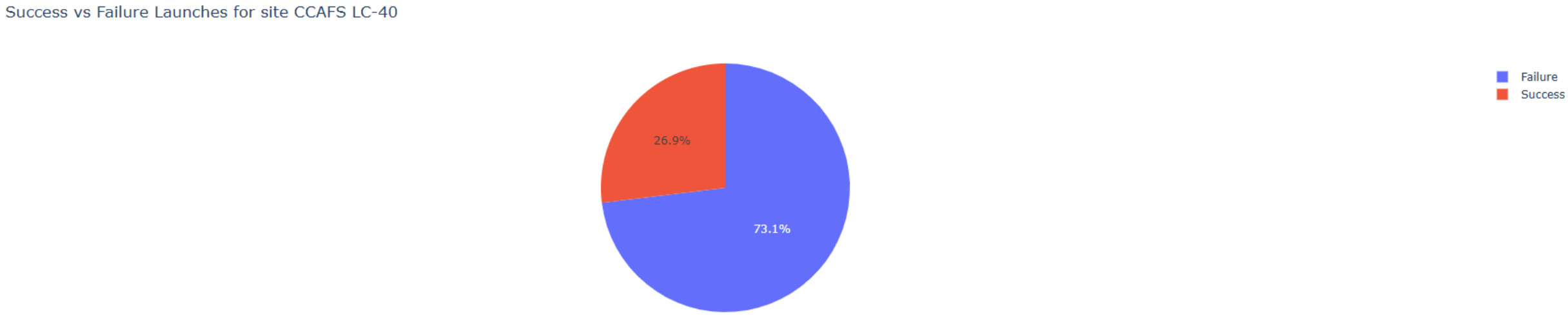
SpaceX Launch Records Dashboard



We can see how the KSC LC-39A launch site has the highest success rate of 41.7% among all launches, while CCAFS SLC-40 has the worst success rate of 12.5%.

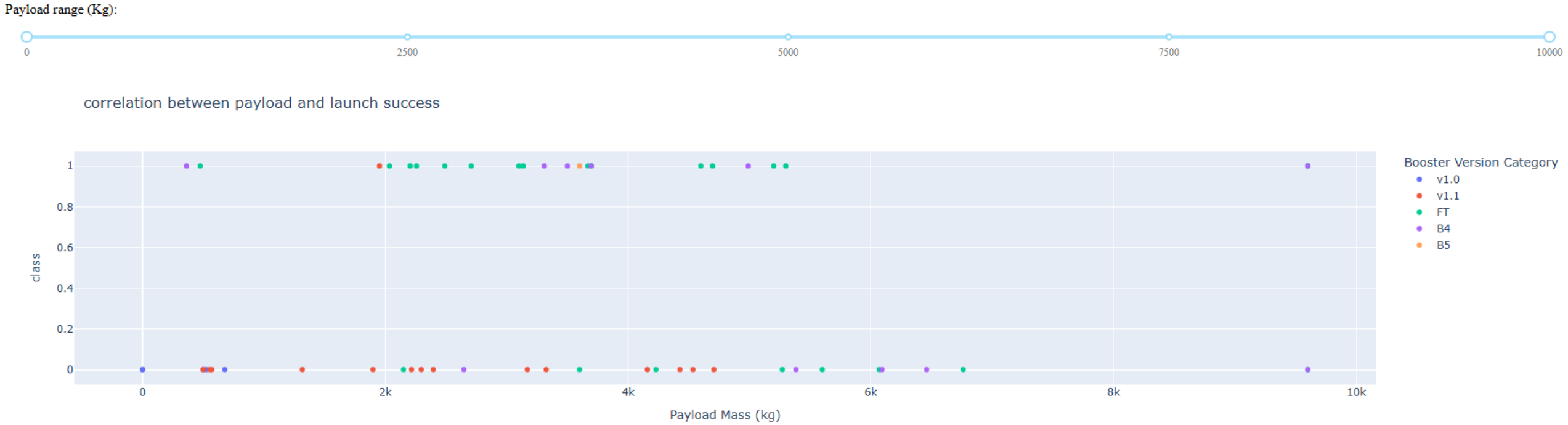
SpaceX Launch Records Dashboard

CCAFS LC-40



We see how CCAFS LC-40 has a high margin of more than 73% and only almost 26% of failures.

SpaceX Launch Records Dashboard



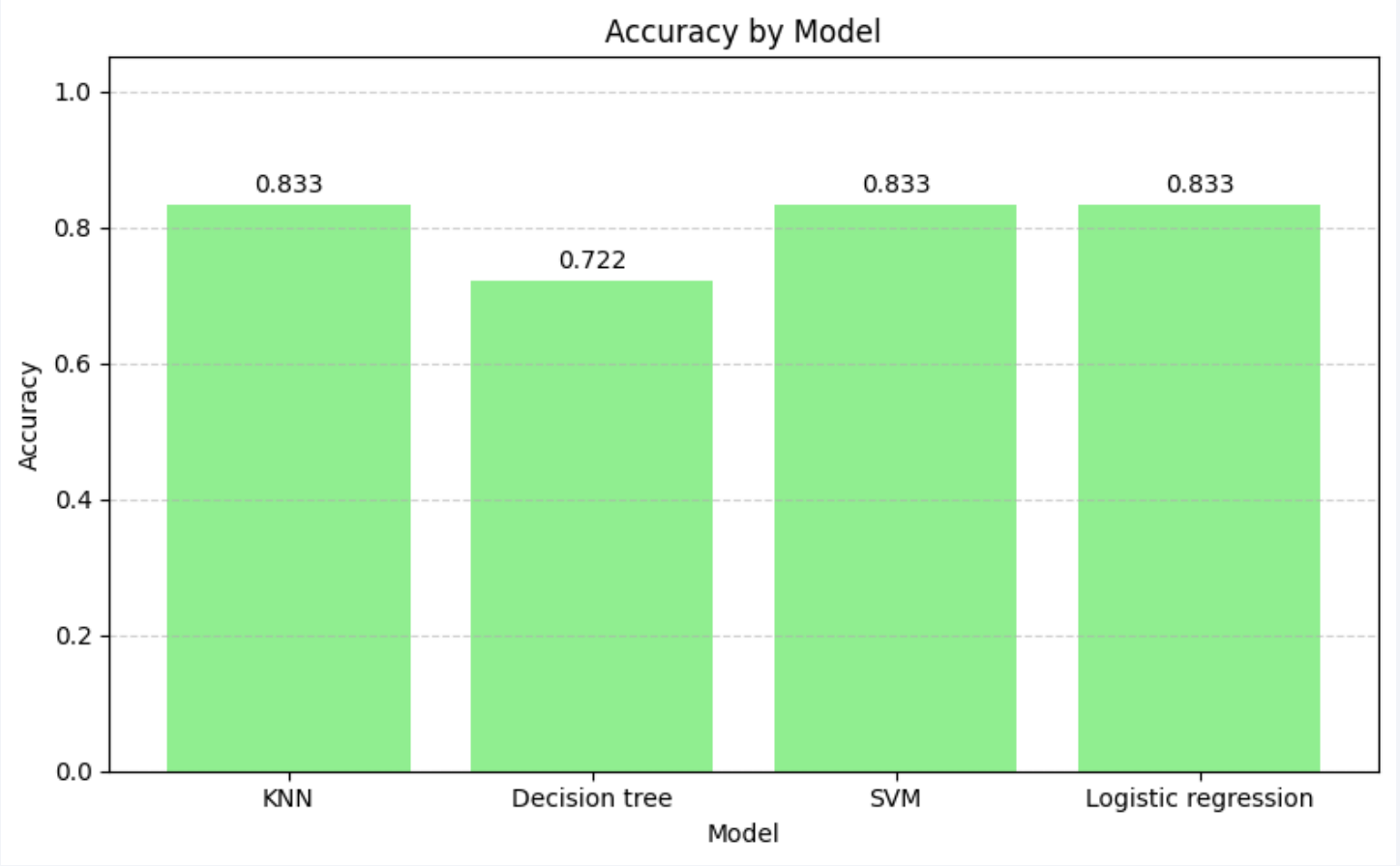
The graph doesn't show a clear relationship between weight and launch success. The "FT" and "B5" versions have more successes, while the "v1.0" and "v1.1" versions have more failures. Even payloads of up to 10,000 kg have been successfully launched.

Section 5

Predictive Analysis (Classification)

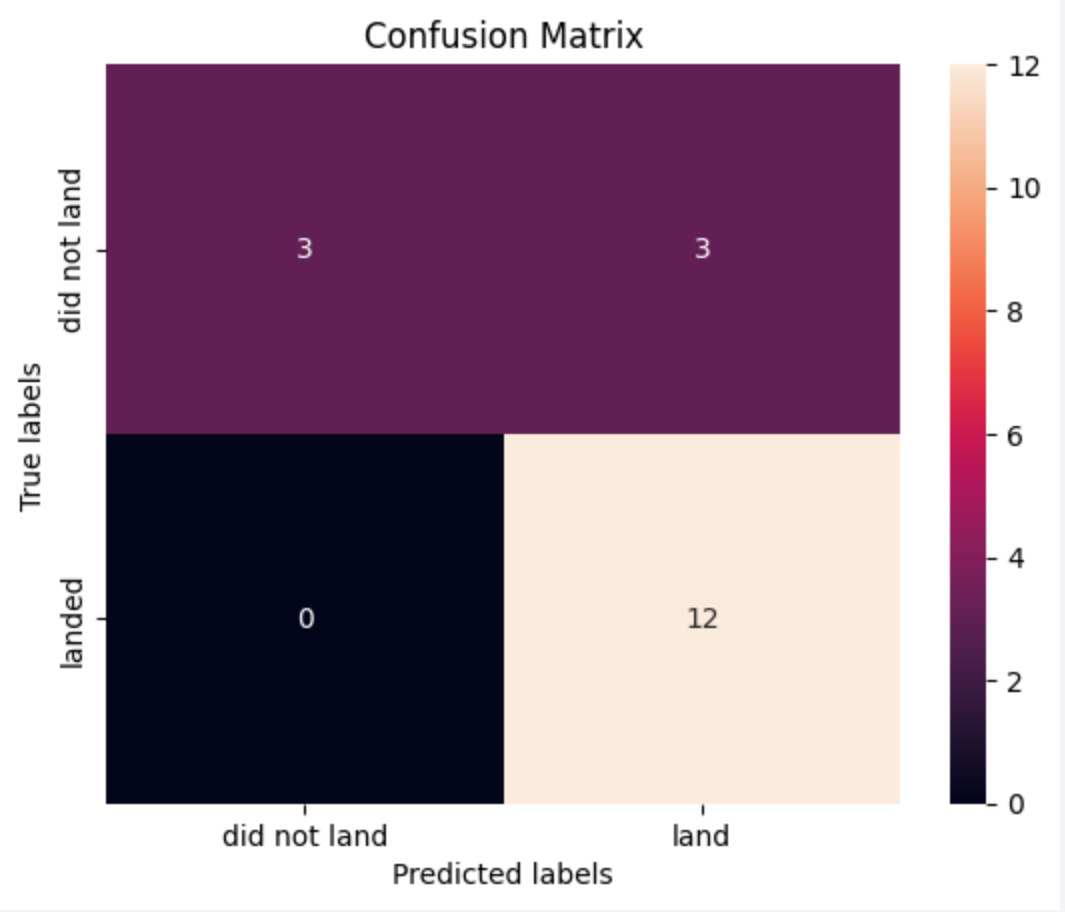
Classification Accuracy

We can see that there are no significant differences between the logistic regression, SVM, and KNN models, with all three models achieving a similar result of 83%. The decision tree performed the worst, with 72% accuracy.



Confusion Matrix

The model is very good at detecting those that do land (12/12 correct).
But it's quite confused about those that didn't land (12/12 correct).
It gets 3 right but 3 wrong.
It has perfect recall (it doesn't miss a single "landed").
But its accuracy isn't perfect, because it sometimes predicts "landed" when it actually didn't.



Conclusions

With 83% accuracy, the KNN model demonstrated that it is possible to predict the landing of reusable rockets with high reliability using measurable parameters and supervised machine learning techniques. Although this work is educational in nature, the results achieved reflect the enormous potential of applying artificial intelligence to real-world problems. This approach not only optimizes critical decisions but also anticipates a future where space exploration and automation advance hand in hand. This is just the beginning: data science doesn't predict the future, it builds it.

Thank you!