
# Winning Space Race with Data Science

Sosthène SEDOGBO

May 05th 2023

# Outline

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

# Executive Summary

The objective of this project is to understand the factors determining landing outcome of space X falcon 9 rockets. To do so, a dataset was constructed using information provided by Space X itself through its API, but also from Wikipedia website. The data was cleaned, treated, and analyzed using tools like pandas, NumPy, matplotlib and scikit-learn.

Key findings include:

- Determining factors of the outcome, include Flight number, Launch Site, Payload Mass (kg) and Booster version.
- The Kennedy Space Center (KSC LC-39A) in Florida site is the most used for falcon 9 launches ( 41,7%)
- But this site is also the one with the high success launches rate (76,9%).
- As the flight number increases, the success rate increases too.
- Trained machine learning models produced a relatively high accuracy (about 83,33%) for predicting the outcome of new attempts of rocket landing.

Overall, this analysis provided information about how to choose the best parameters to have the highest probabilities of launch success and landing success.

# Introduction

Launching a rocket into space is very expensive and only governmental agencies and multi billionaires' businessmen can afford it. In fact, most of providers advertise the launches, approximately 165 million dollars each. However, the famous company Space X succeeded something nobody did before: they can reuse the first stage of a rocket after a launching, but only if the landing is a success.

It's then a crucial question to know if we can predict the landing outcome of a launch so that the cost of a launch can be predictable.

Section 1

# Methodology

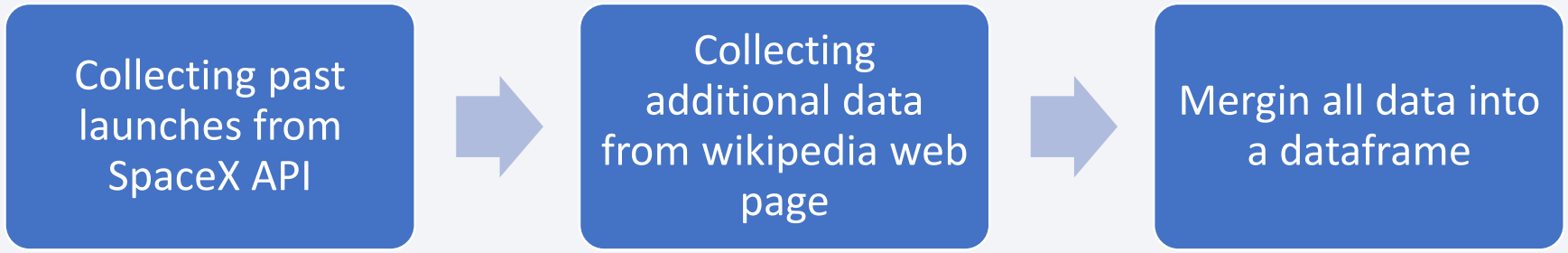
# Methodology

## Executive Summary

- Data collection methodology: A dataset was constructed using information provided by Space X itself through its API, but also from Wikipedia website (check appendix for links).
- Perform data wrangling: The dataset was then filtered to retain only interesting features such as flight number, booster version, launch site… Missing values have been replaced with the mean of their respective features because they were minoritarian. Also, some categorical features have been replaced with numerical or Boolean values.
- Perform exploratory data analysis (EDA) using visualization and SQL: We proceed to analyze the data using visualization tools like scatter plots and bar plots. The insights were found, and we easily selected the potential good features.
- Perform interactive visual analytics using Folium and Plotly Dash: We then draw an interactive map to look at the site's locations from where the Space X rockets were launched. This helped us to see that all the locations were very close to coastline.
- Perform predictive analysis using classification models: We built many different machine learning models such as logistic regression, KNN, or Decision Tree Classifier and we trained them over a training set. The results obtained were quite satisfying: almost 84% accuracy over the testing set for the best model.

# Data Collection

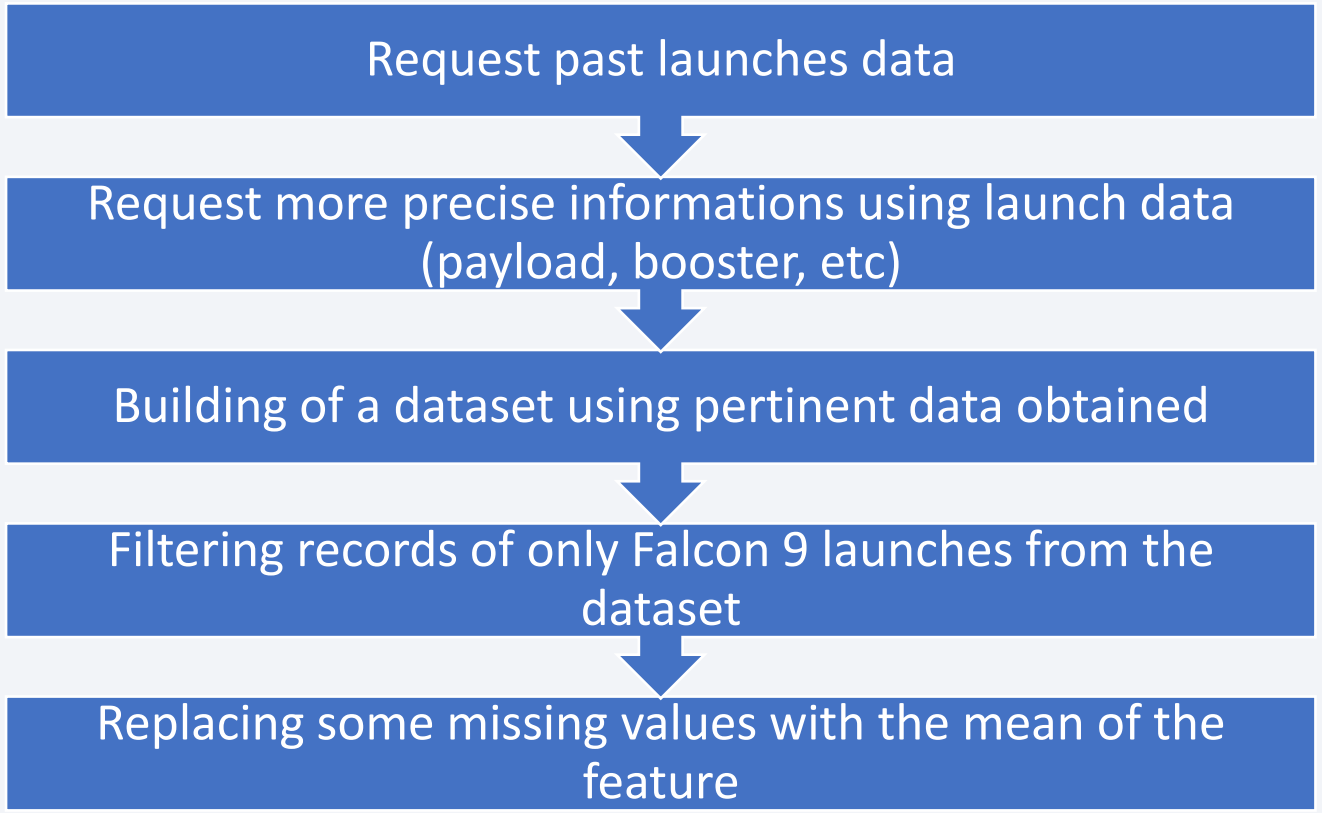
- A dataset was constructed using information provided by Space X itself through its API, but also from Wikipedia website (check appendix for links).



# Data Collection – SpaceX API

Github notebook reference file:

- Name: api_rest.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

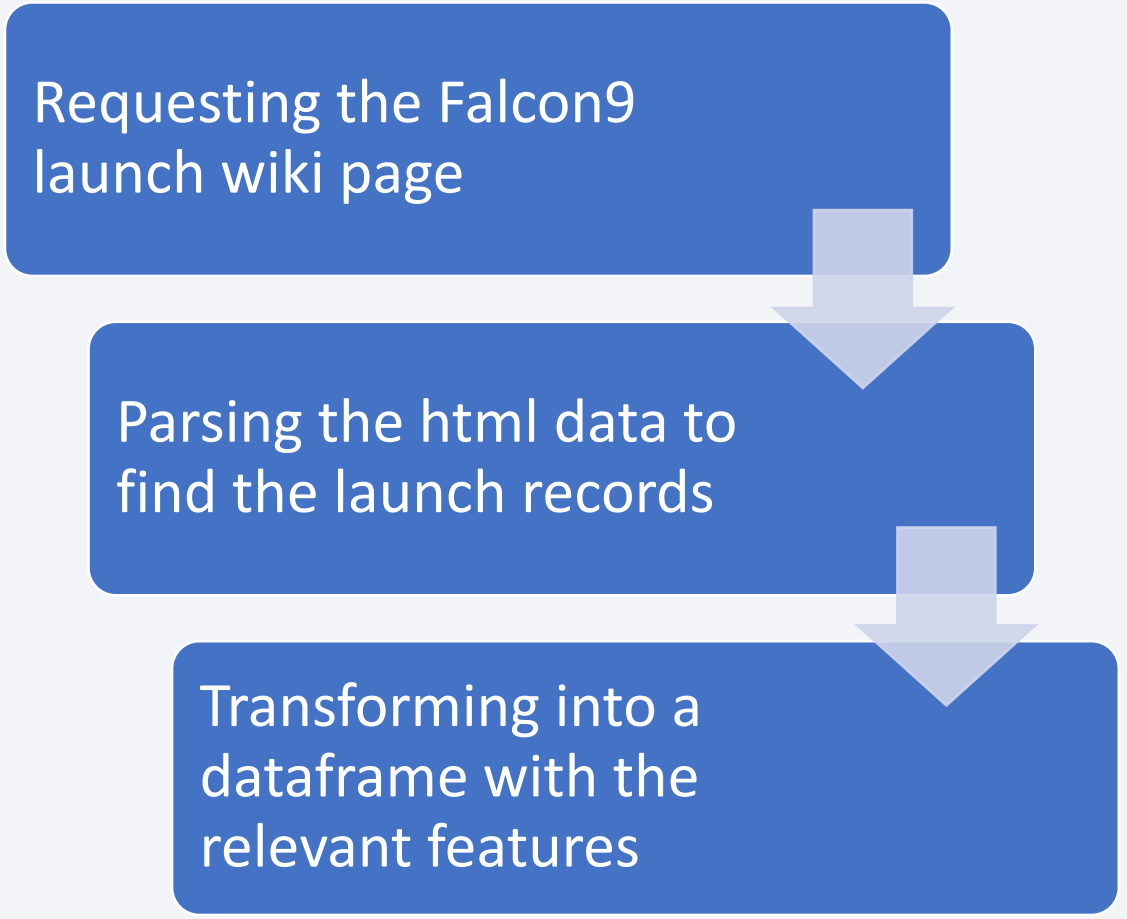



Data collection process using space X API

# Data Collection - Scraping

Github notebook reference file:

- Name: web_scrapping.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*



# Data Wrangling

- ❖ Summary of the dataset
- ❖ Computing the number of launches on each site
- ❖ Creating the “Class” feature , a Boolean one from the “Output” feature

Github notebook reference file:

- Name: data_wrangling.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

# EDA with Data Visualization

❖ We proceed to analyze the relationship between features: using visualization tools like scatter plots for numerical data and bar plots for categorical data. The insights were found, and we easily selected the potential good features.

❖ These plots are the simplest for visualization

Github notebook reference file:

- Name: visualization.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

# EDA with SQL

We performed some SQL queries to test the database in which the data is stored. Like:

- ❖ Select the names of the unique launch sites in the space mission
- ❖ Select 5 records where launch sites begin with the string ‘CCA’
- ❖ Select the total payload mass carried by boosters launched by NASA
- ❖ Select the average payload mass carried by booster version F9 v1.1
- ❖ Select the date when the first successful landing outcome in ground pad was achieved

Github notebook reference file:

- Name: sql.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

# Build an Interactive Map with Folium

An interactive map to look at the site's locations from where the Space X rockets were launched.

Was built using Folium library.

Markers, popup and icons were added to easily notice the locations.

Cluster markers were also used to facilitate visualization of launch success rate on each site.

This helped us to see that all the locations were very close to coastline.

Github notebook reference file:

- Name: analytics.ipynb
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

# Build a Dashboard with Plotly Dash

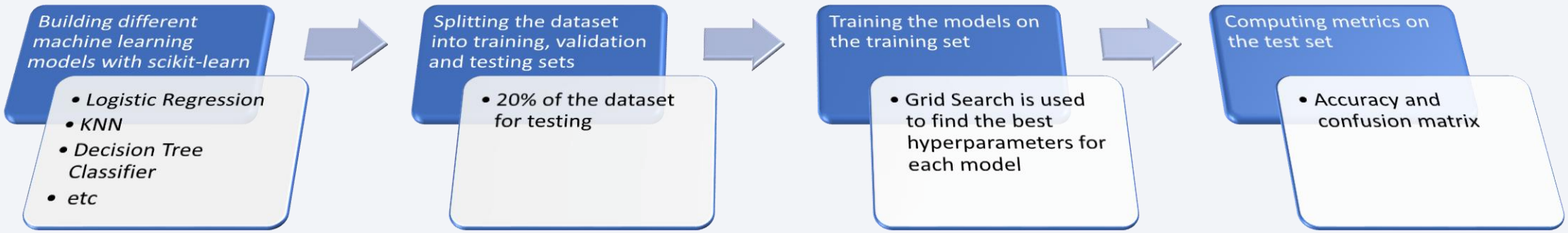
We built a dashboard app using Dash library. This dashboard is composed of:

- a dropdown menu that user can use to choose a launch site
- A pie chart showing the success landing rate for each site or the proportions for all sites.
- A scatter plot showing relationship between payload mass, launch site and booster version.
- A range slider that is used to select a range for the payload mass on the scatter plot.

Github notebook reference file:

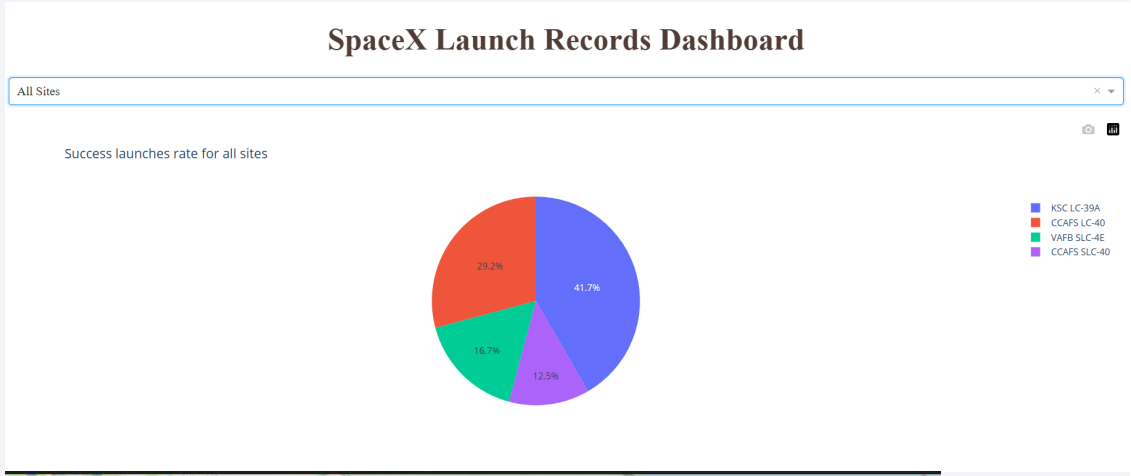
- Name: spacex_dash_app.py
- Link: *https://github.com/TheFiresword/ibm_notebook_grade.git*

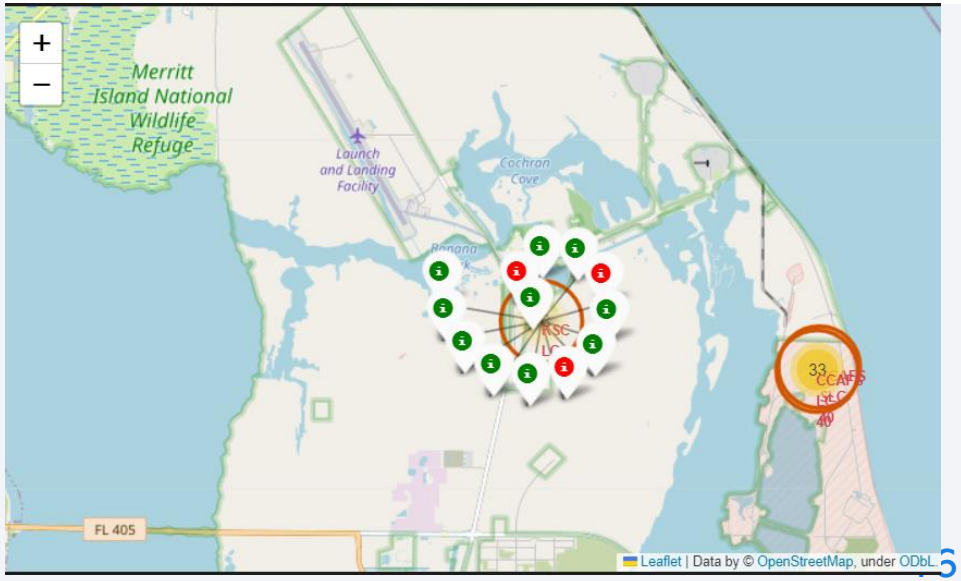
# Predictive Analysis (Classification)

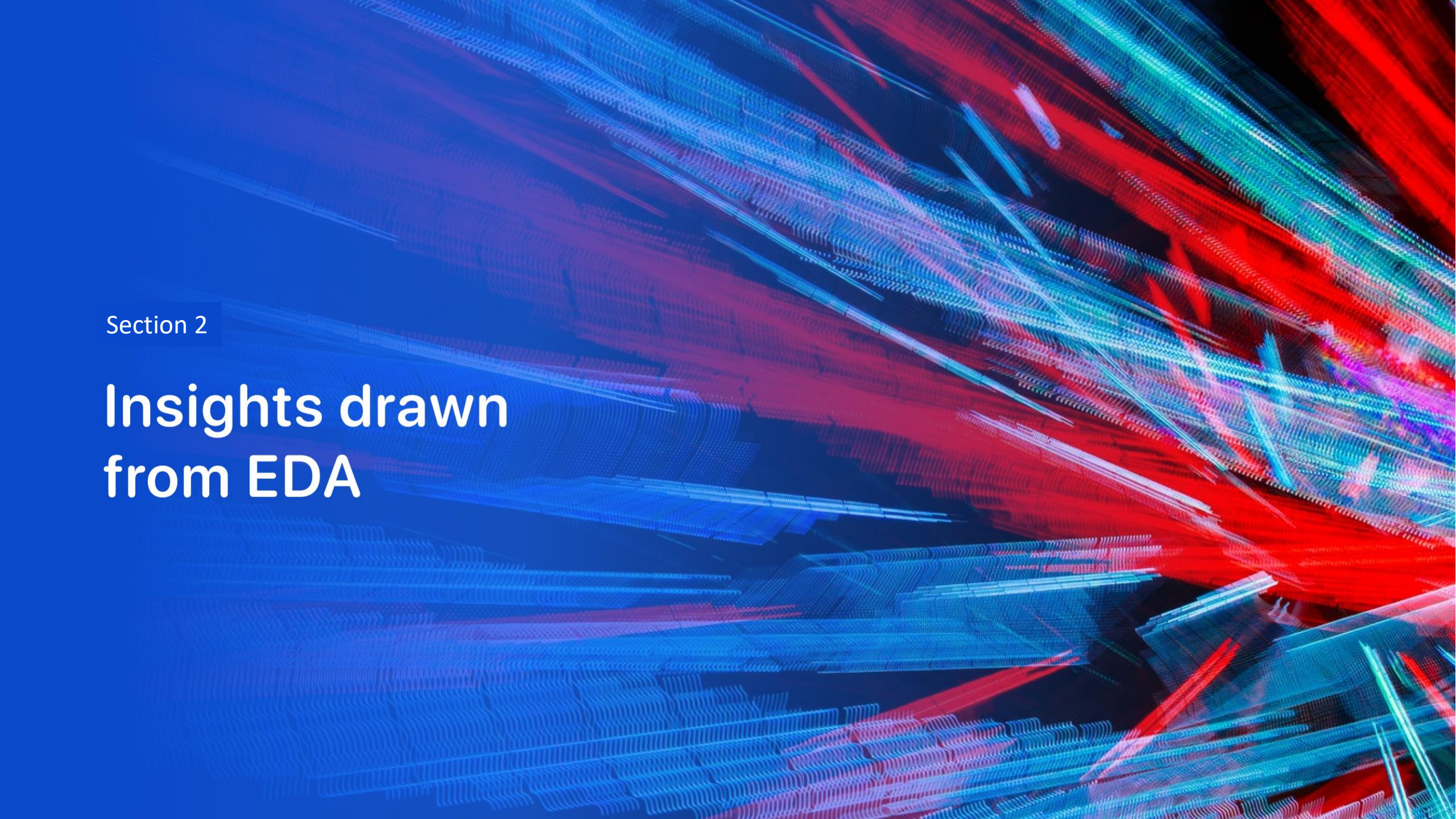


# Results

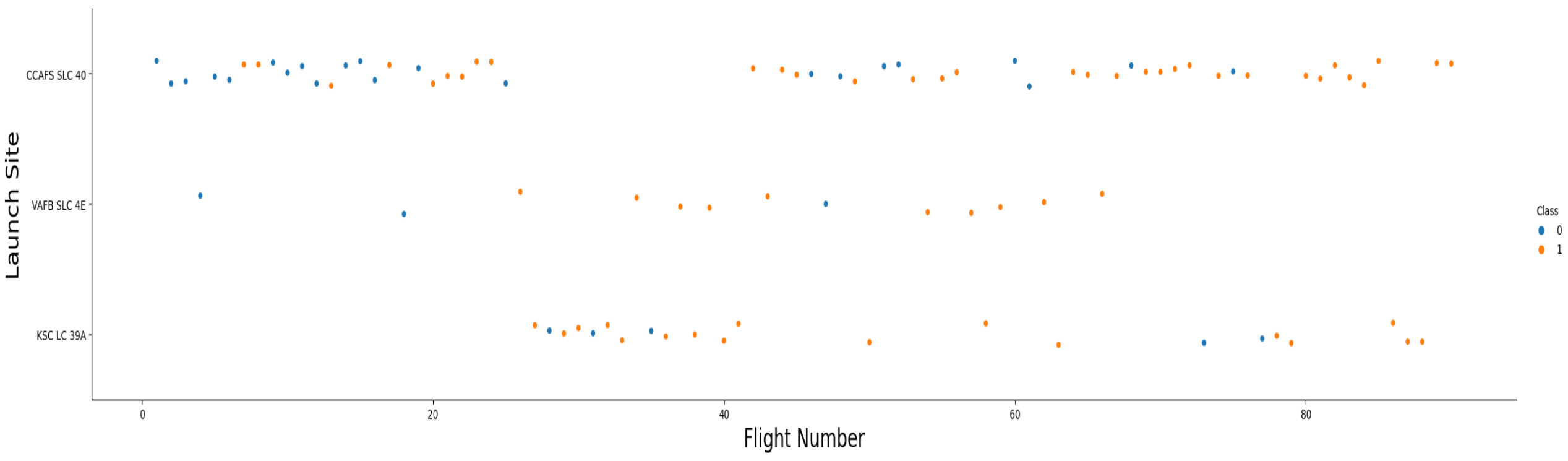
- Determining factors of the outcome, include Flight number, Launch Site, Payload Mass (kg) and Booster version.
- The Kennedy Space Center (KSC LC-39A) in Florida site is the most used for falcon 9 launches ( 41,7%)
- But this site is also the one with the high success launches rate (76,9%).
- As the flight number increases, the success rate increases too.
- Trained machine learning models produced a relatively high accuracy (about 83,33%) for predicting the outcome of new attempts of rocket landing.



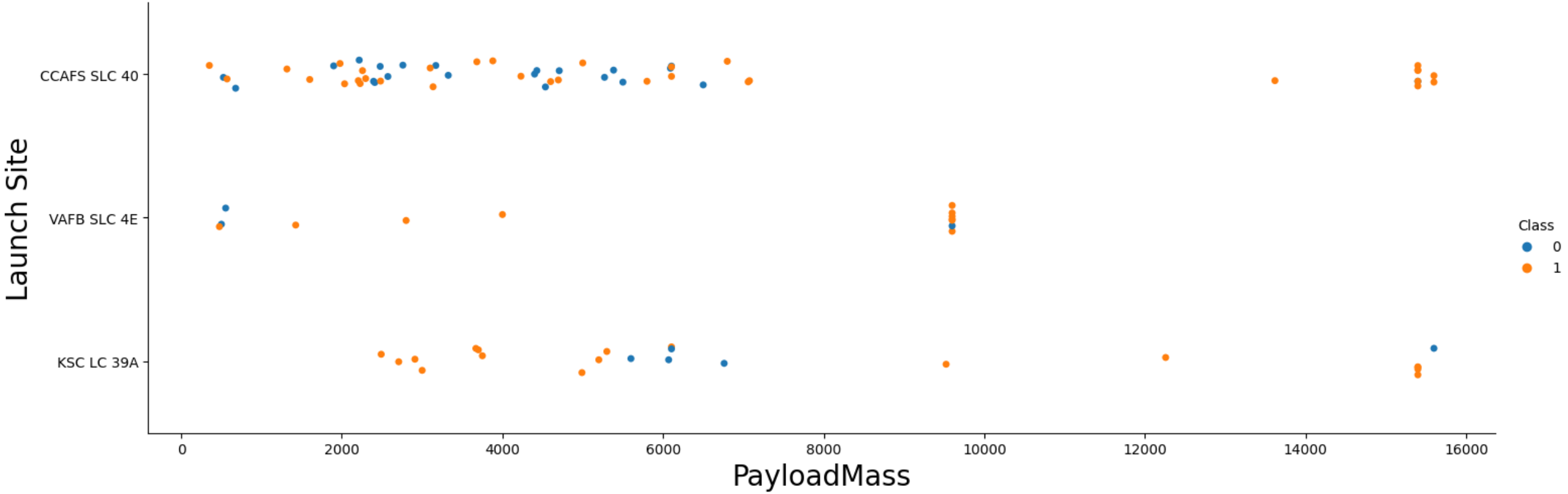

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site



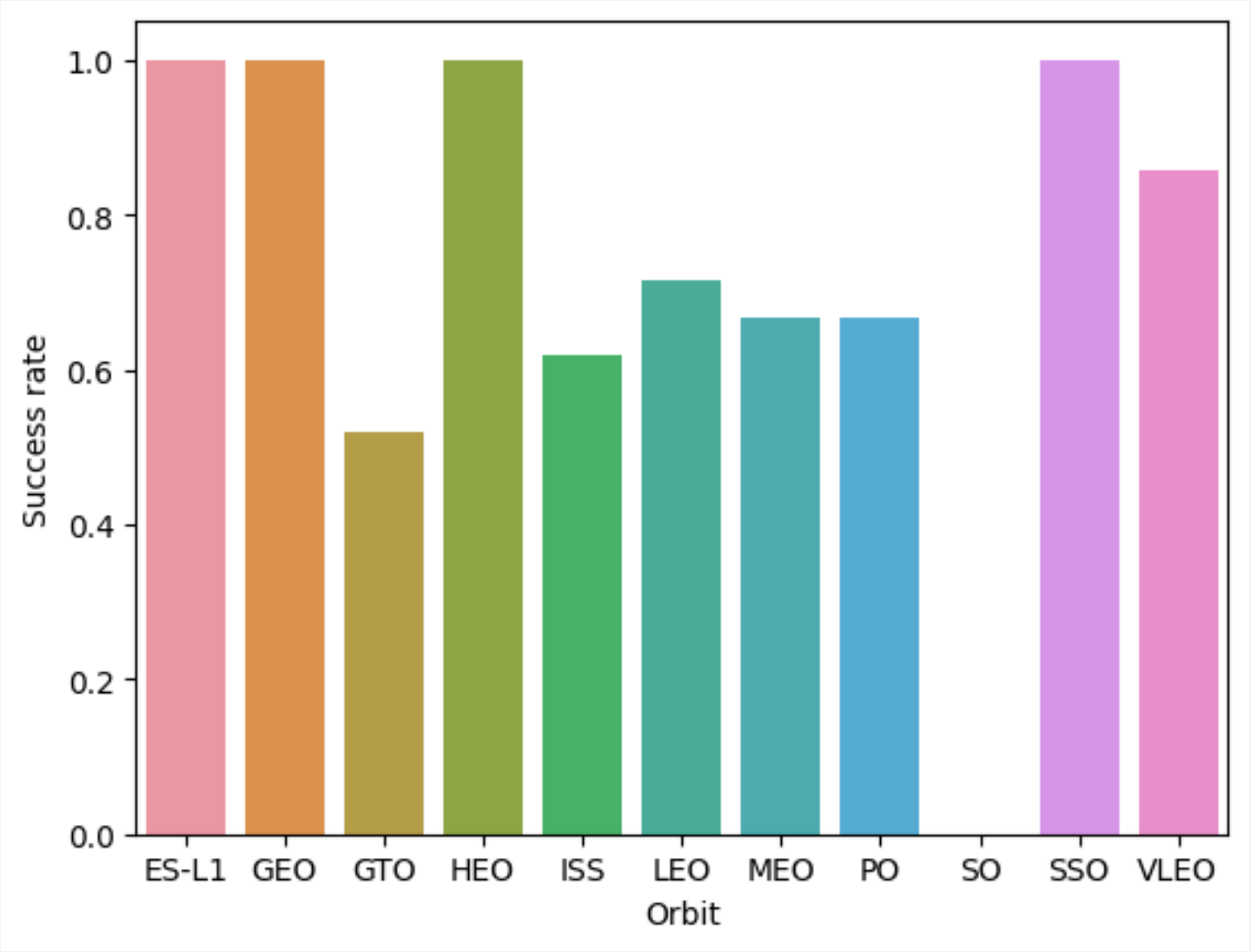

- This plot shows the landing outcome of flights depending on the flight number and the launch site. The orange points represent the successful landing.
- Globally, it shows that as the flight number increases, the successful landing rate increases. We also notice that the CCAFS SLC 40 Site is the most used over thos 80 first flights but has the most failures.

# Payload vs. Launch Site



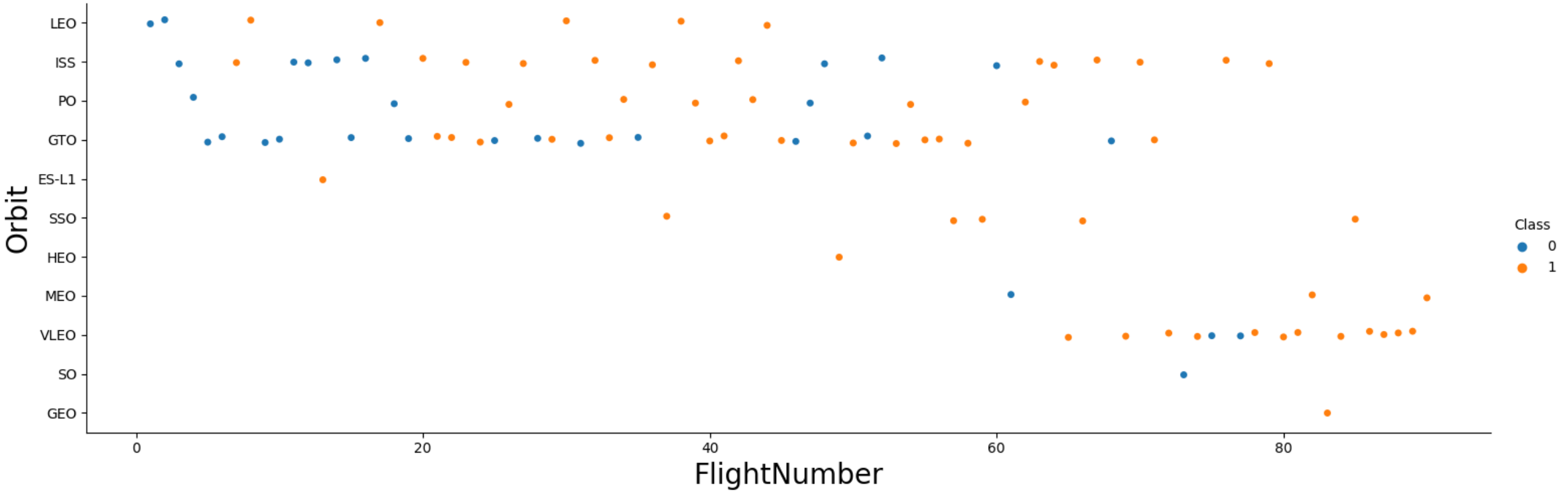

For the VAFB-SLC launch site there are no rockets launched for heavy payload mass
But there is no evident correlation between the 2 features;

# Success Rate vs. Orbit Type



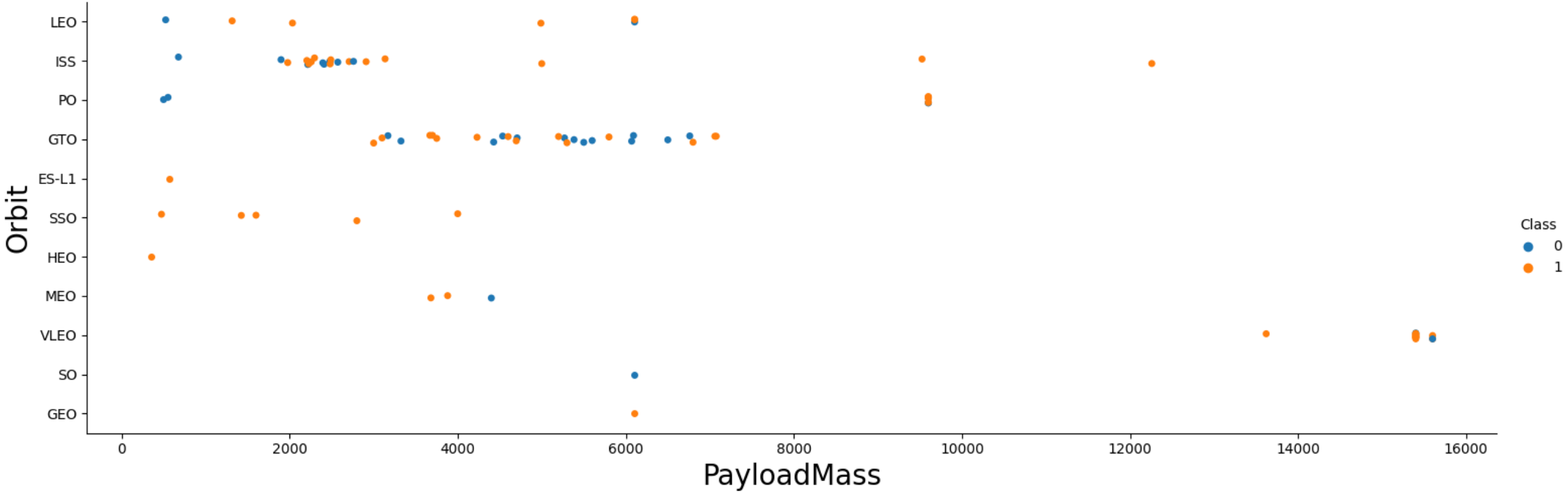

- Choosing ES-L1, or GEO or HEO or SSO orbits assures a succesful landing.
- The GTO orbit is the worst

# Flight Number vs. Orbit Type



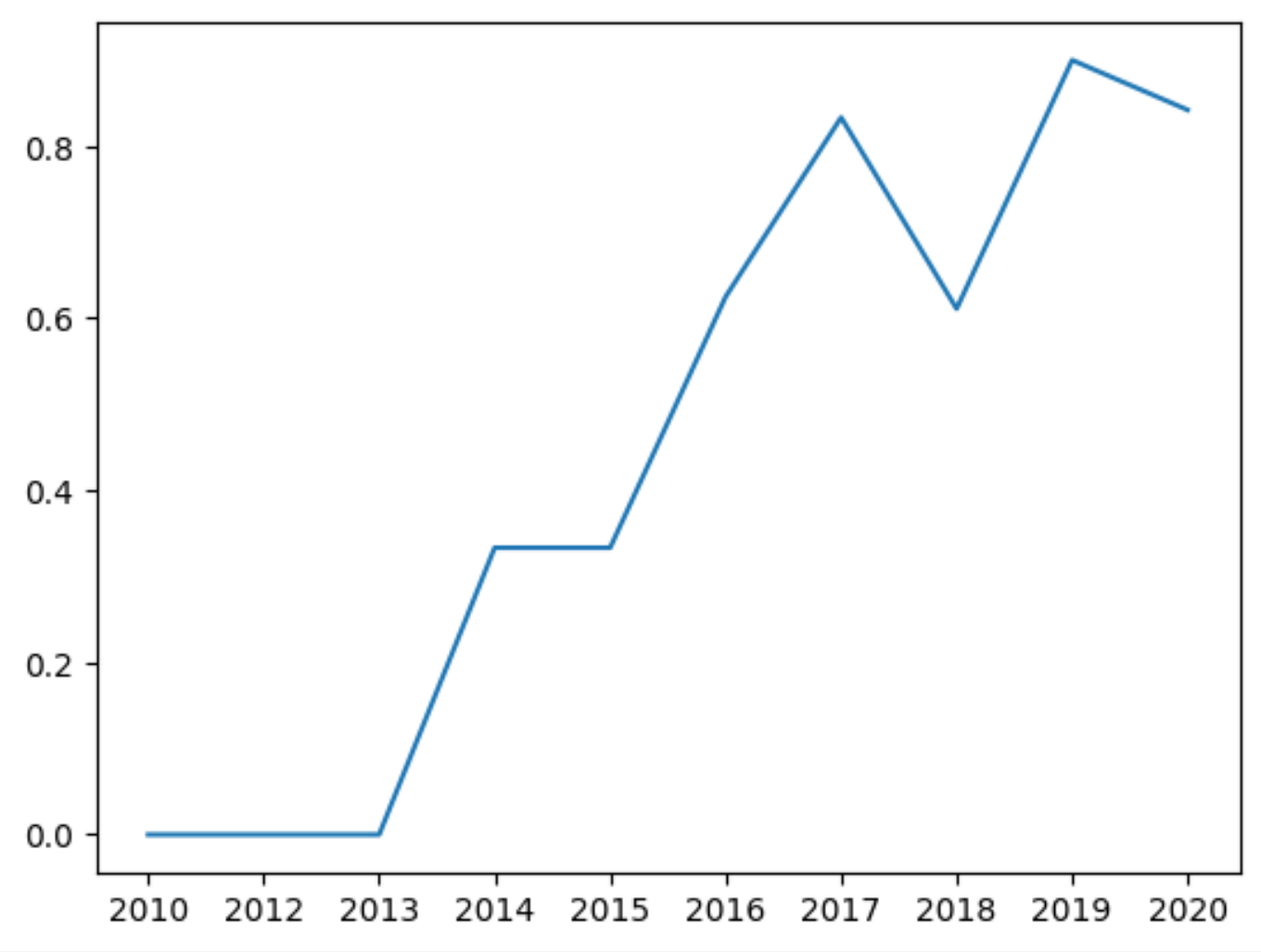

The first flights were sent only on LEO, ISS, PO, GTO, and ES-L1 orbits and there is a high rate of failure. But as much as the flight number increases there are much success landings, especially in the new orbits used.

# Payload vs. Orbit Type




With heavy payloads the successful landing are more for Polar, LEO and ISS.
However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.

# Launch Success Yearly Trend




We can see that globaly, success rate keeps increasing since 2013

# All Launch Site Names

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

By querrying the database, it appeared that there are 4 sites used to launch the rockets.

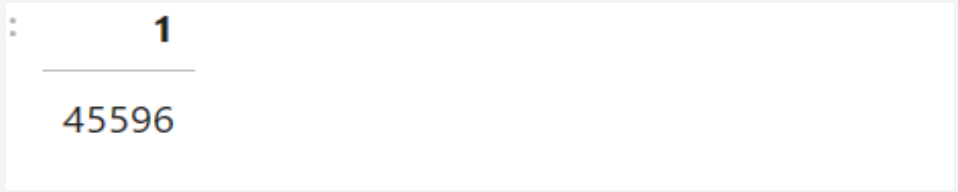
# Launch Site Names Begin with 'CCA'

| DATE | time_utc | booster_version | launch_site | payload | payload_mass_kg_ | orbit | customer | mission_outcome | landing_outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

By querrying the database, we see that the CCAFS LC-40 site is very used

# Total Payload Mass

The total payload mass carried by boosters launched by NASA is:
45596 KG



# Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is : 2928 KG

| 1 |
| --- |
| 2928 |

# First Successful Ground Landing Date

```
%sql SELECT min(DATE) FROM SPACEX WHERE MISSION_OUTCOME = 'Success' AND LANDING_OUTCOME = 'Success (ground pad
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.clou
Done.
```

| 1 |
| --- |
| 2015-12-22 |

The first successful ground landing occured at the end of 2015.
This is quite late knowing that the launches started in 2010 and the first success were obtained  from 2013. So the ground landing seems a bit more complex.

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEX  \
WHERE LANDING_OUTCOME = 'Success (drone ship)'AND PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.datab
Done.
```

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

Those are the booster versions used in rockets that successfully landed on drone ship with a payload between 4000 and 6000

# Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, count(*) as Count \
FROM SPACEX \
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d021
Done.
```

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

Only 1 failure (in flight) was recorded over 102 missions.
That concerns just the launching of the rocket not the landing

# Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX \
WHERE PAYLOAD_MASS_KG_ = (SELECT max(PAYLOAD_MASS_KG_) FROM SPACEX)
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l
Done.
```

| 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 |

Several boosters can carry the maximum payload. This is a good thing

# 2015 Launch Records

```
%sql SELECT LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE \
FROM SPACEX \
WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io9
Done.
```

| landing_outcome | booster_version | launch_site |
|---|---|---|
| Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

In 2015, just 2 landings on drone ship failed.

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

```
%sql \
SELECT LANDING_OUTCOME, count(*) as Count\
    FROM SPACEX \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING_OUTCOME \
    ORDER BY Count DESC
```

```
 * ibm_db_sa://qmx13331:***@21fecfd8-47b7-4937-840d-d791d0218
Done.
```

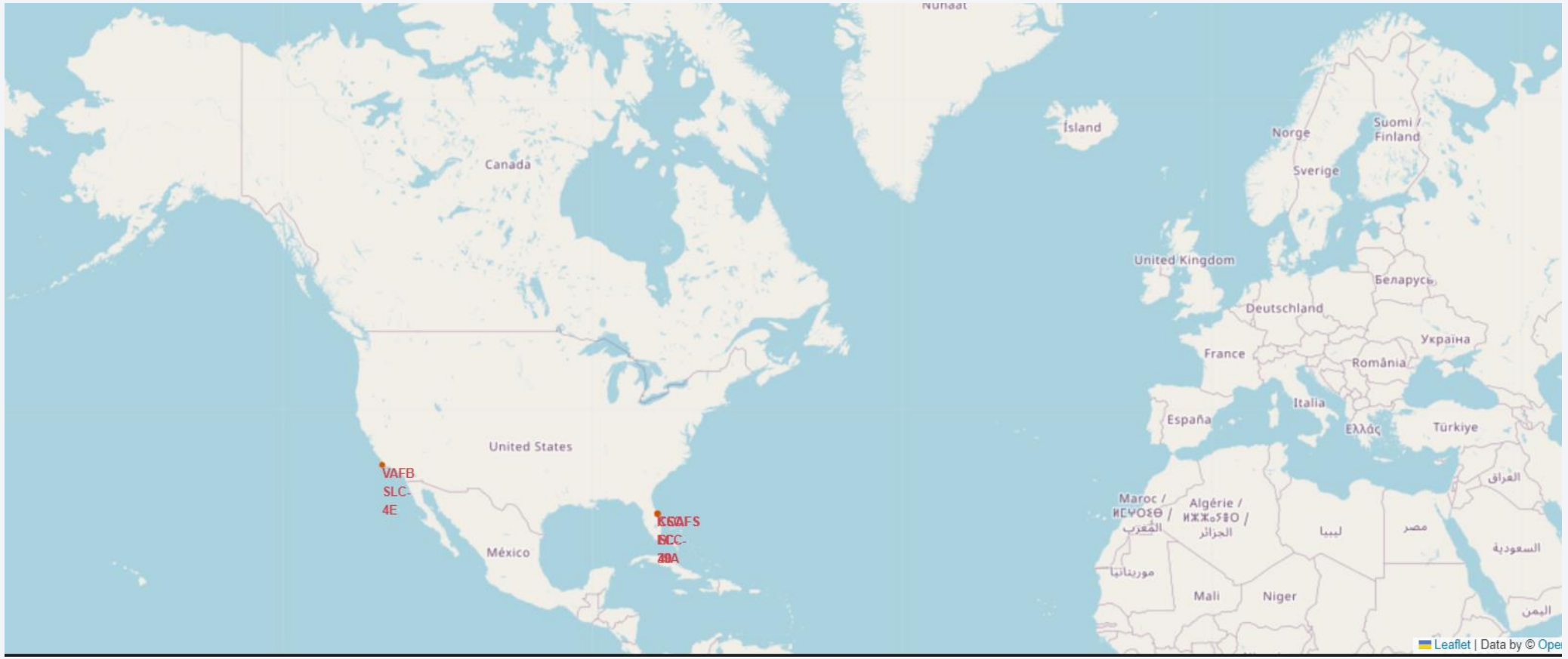
| landing_outcome | COUNT |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Between 2010 and 2017, 31 launching were made. But only 11 were globally a success.

Section 3

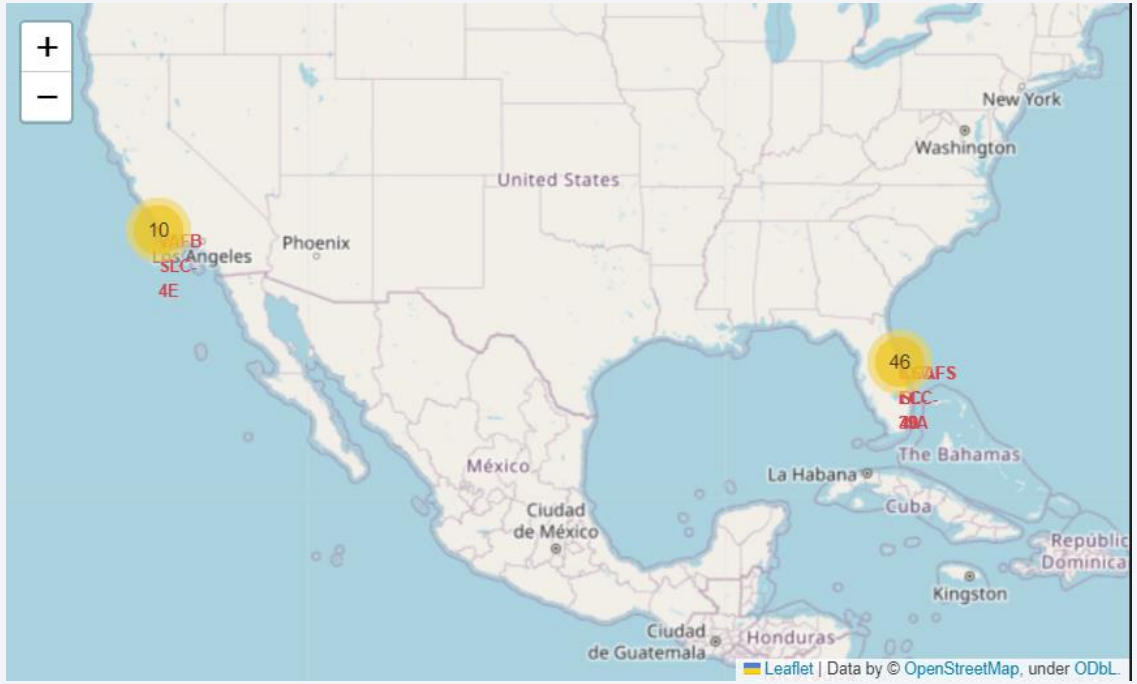
# Launch Sites Proximities Analysis

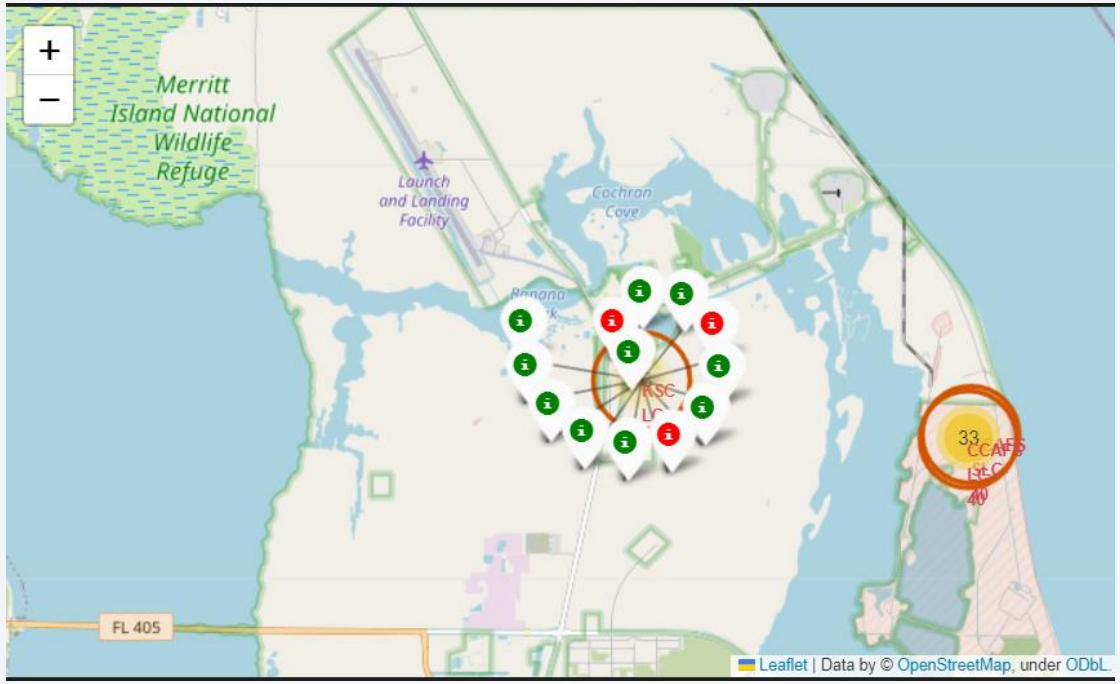
# Launch Sites locations on a map




They are all in the US, close to the Equator line and are very close proximity to the coast

# Launches Outcome on the map

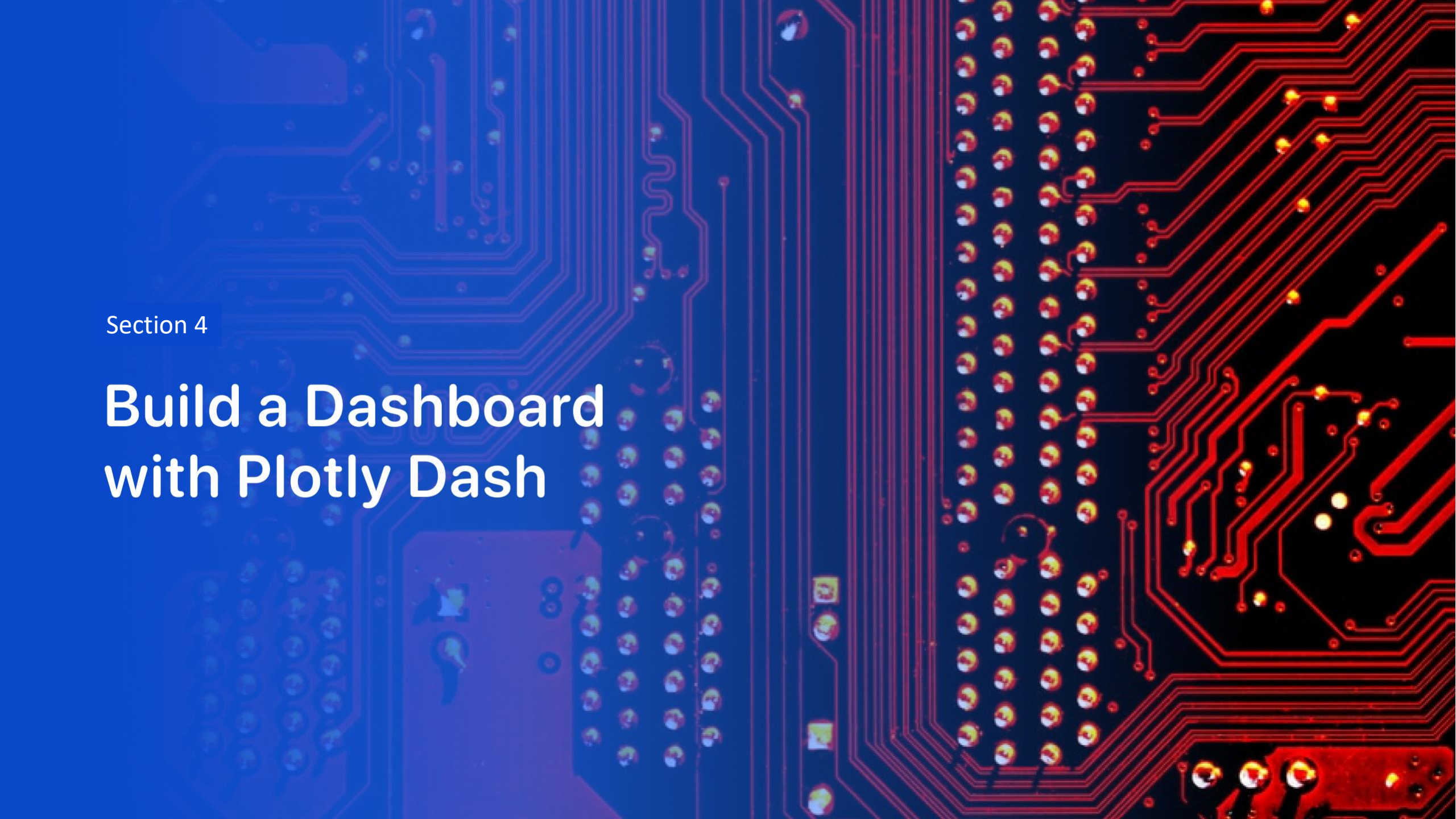


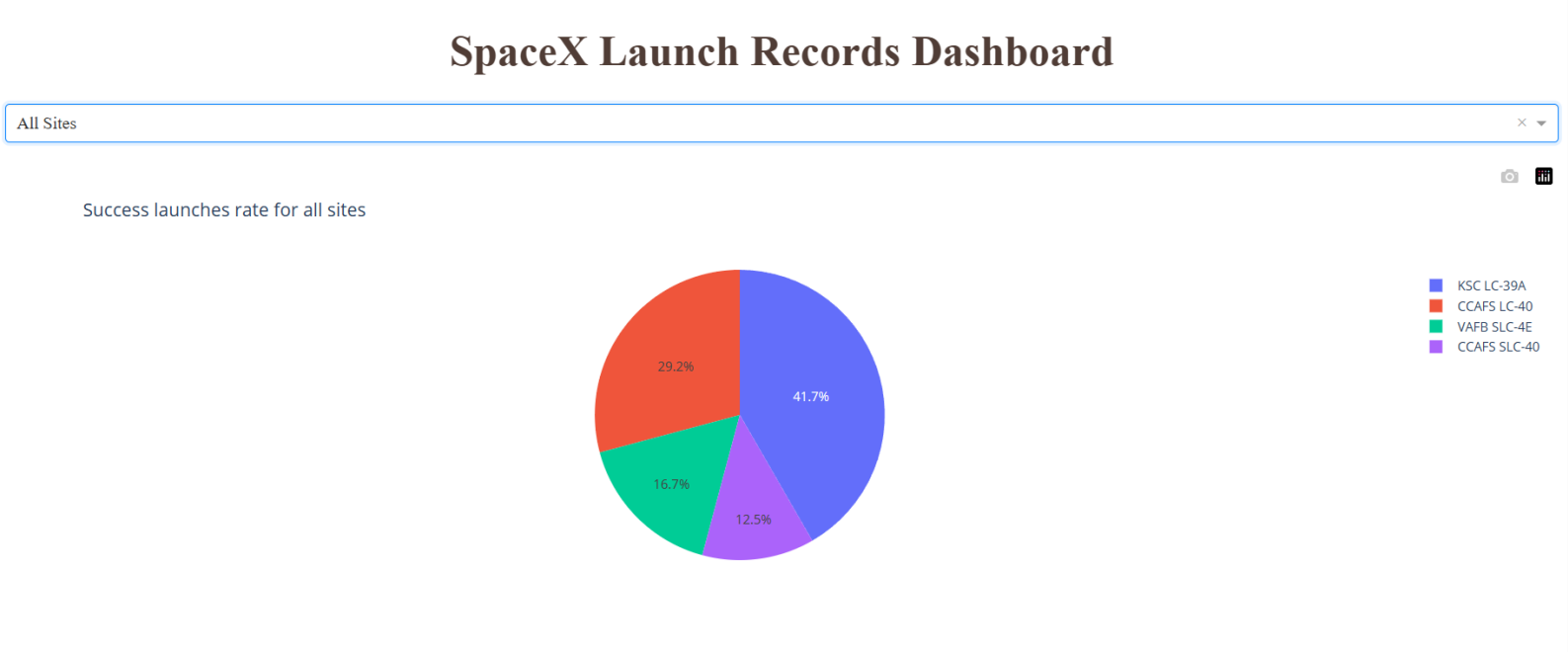

We can easily see that the KSC site is the highest success rate site ( 3 failures over 13 launches)

# Sites proximities


The launch sites are close proximity to coastline, railways, highways and keep certain distance away from cities

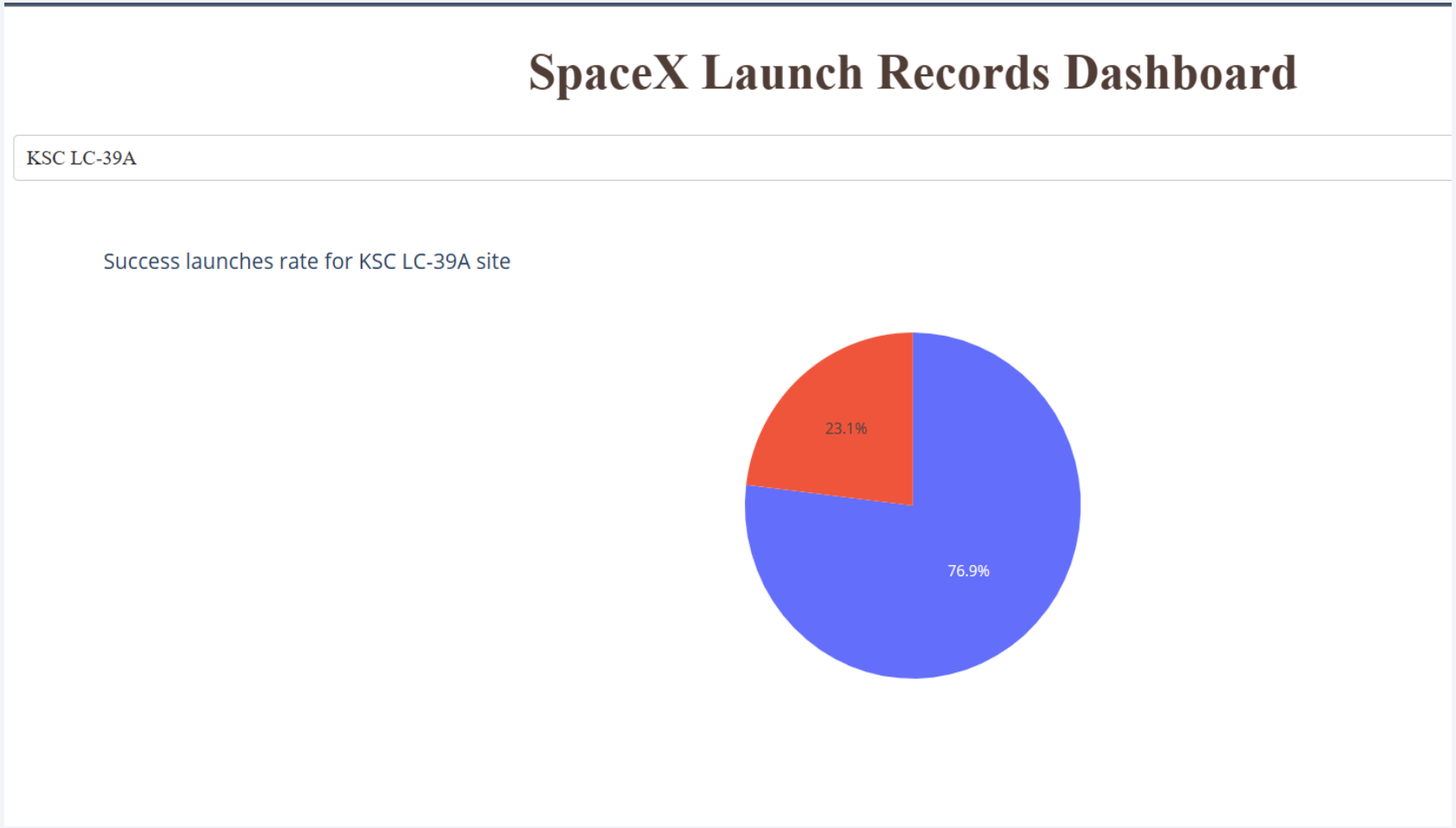

Section 4
Build a Dashboard with Plotly Dash

# Launches rates all sites



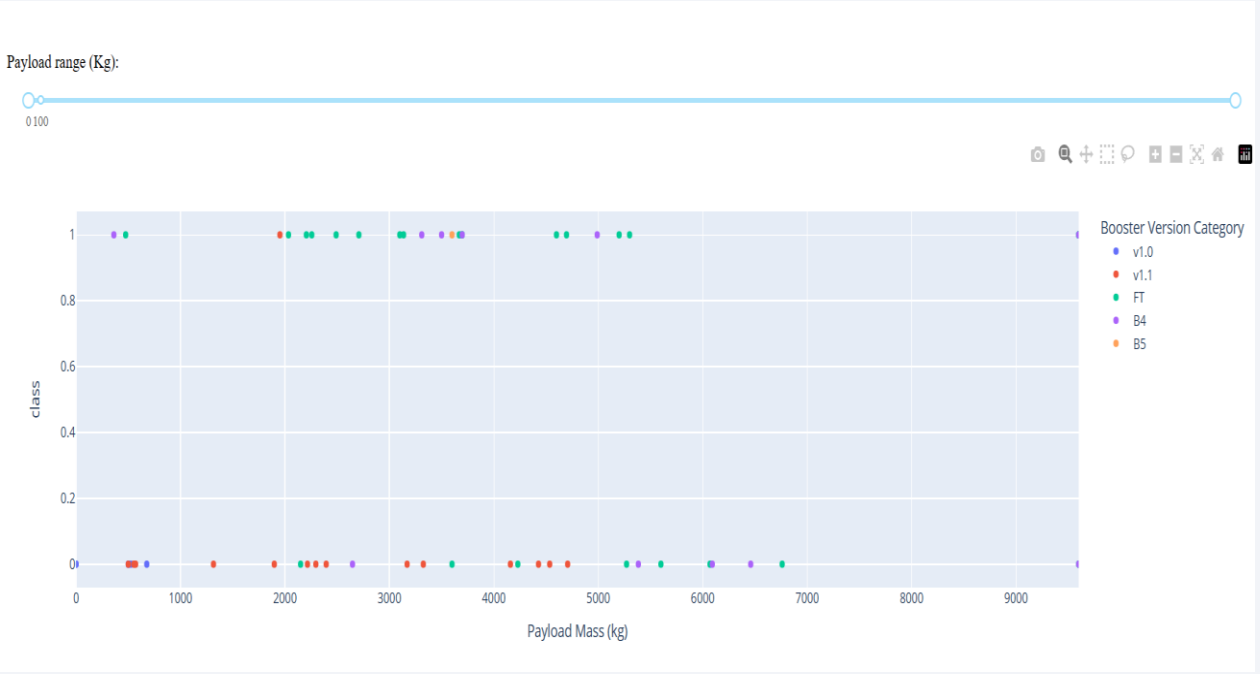

This piechart confirms that the KSC LC site is the most used for launching.

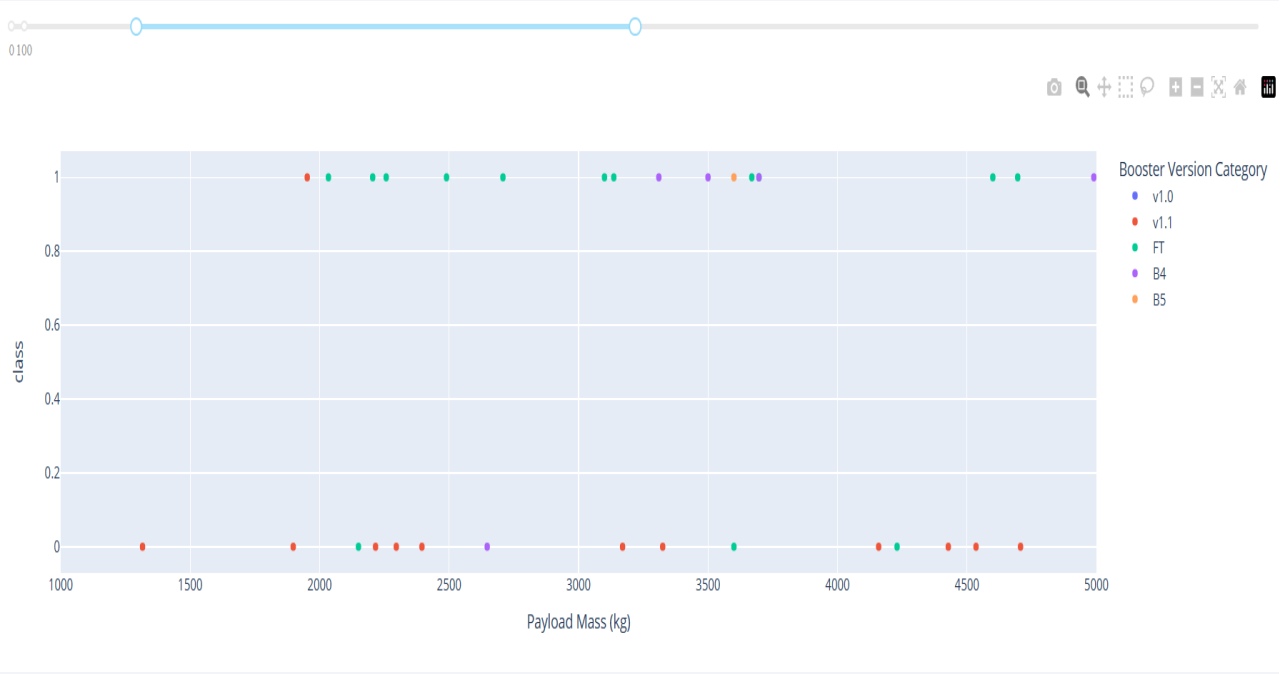
# Highest success rate Site




76,9% of the launches are succeeded on the Kennedy Space Center (KSC LC-39A) in Florida

# Payload vs Launch Outcome



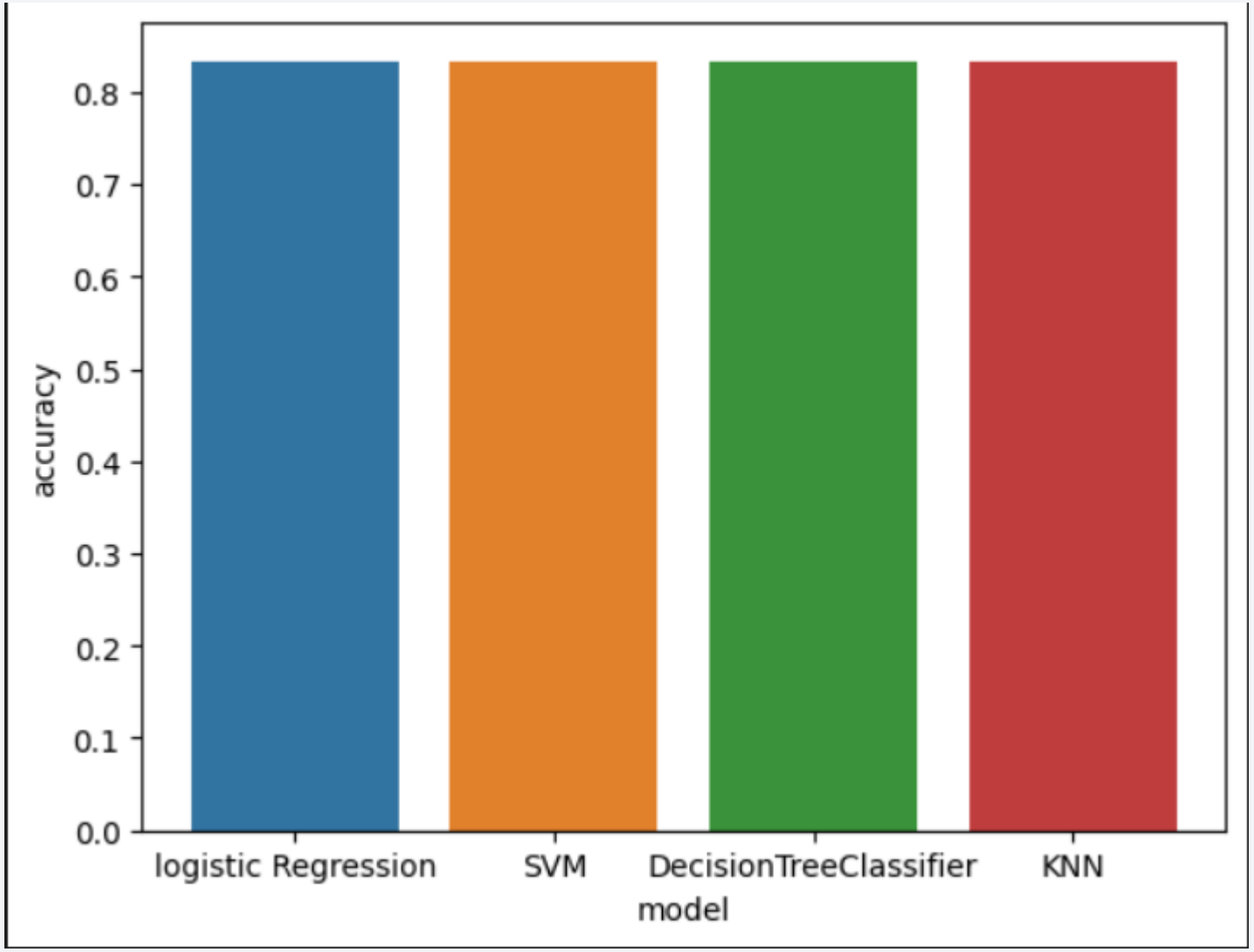

The main observations are:

- v1.1 and FT boosters are the most used but FT ones are the most efficients.
- The payload range with a maximum of success, is [2000 – 5000]kg

Section 5

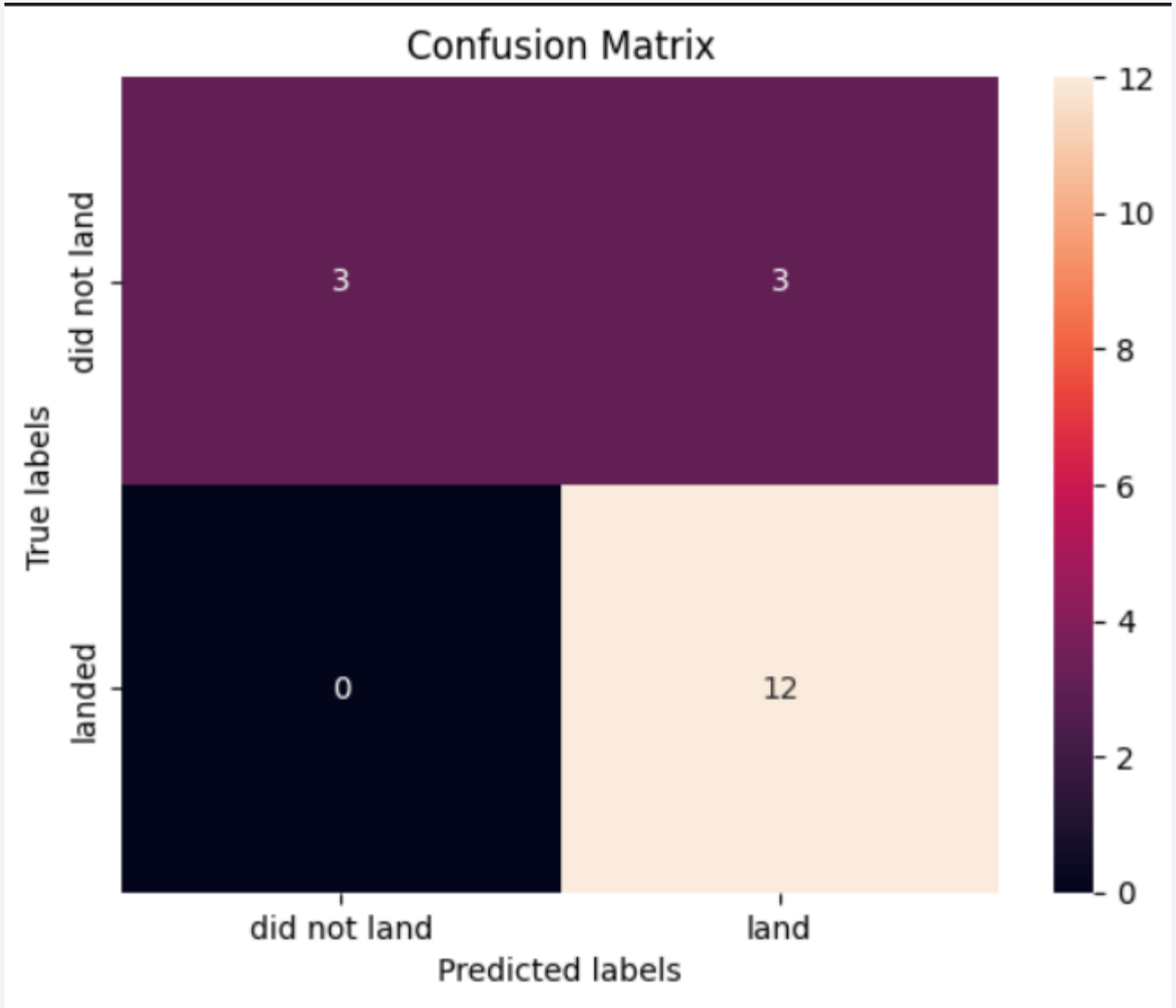
# Predictive Analysis (Classification)

# Classification Accuracy




It seems that all the model builts have an accuracy of 83% on the test dataset. Then, we can choose any to predict future launches outcomes.

# Confusion Matrix




As said previously, we can choose any model.
Examining the confusion matrix of the Decision Tree Classifier, we see that the model can distinguish between the different classes, but the major problem is false positives. It classifies as success 3 launches that were actually failures.

# Conclusions

- Determining factors of the outcome, include Flight number, Launch Site, Payload Mass (kg) and Booster version.
- The Kennedy Space Center (KSC LC-39A) in Florida site is the most used for falcon 9 launches ( 41,7%)
- But this site is also the one with the high success launches rate (76,9%).
- As the flight number increases, the success rate increases too.
- Trained machine learning models produced a relatively high accuracy (about 83,33%) for predicting the outcome of new attempts of rocket landing.

# Appendix

Space X data sources:

- Wiki page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- API: https://api.spacexdata.com/v4/launches/past

# Thank you!