
# Winning Space Race with Data Science

Eduard Tiron
15-02-2023

# Table of Contents

# Executive Summary

- Summary of methodologies:
  - Data collection: API, Web Scraping
  - Data Wrangling, Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis
- Summary of all results:
  - Exploratory Data Analysis results
  - Interactive maps and dashboard
  - Predictive results

# Introduction

- In this project we take the role of a rocket launching company, Space Y, trying to compete with space X.
- We need to predict the success of the second stage recovery in order to reduce costs.
- We need to find the best launching location to increase the probability of a successful second stage recovery.

# Introduction

- Throughout the report, we refer to launches as successful (class 1) or unsuccessful (class 0). This merely refers to whether or not the 2nd stage was recovered for later reuse, but has no bearing on the actual mission of sending the payload to space.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX API
  - Wikipedia (Webscraping): “List of Falcon 9 and Falcon Heavy launches”
- Performed data wrangling
  - Cleaned data.
  - Created a landing outcome label based on outcome data after summarizing and analyzing features.

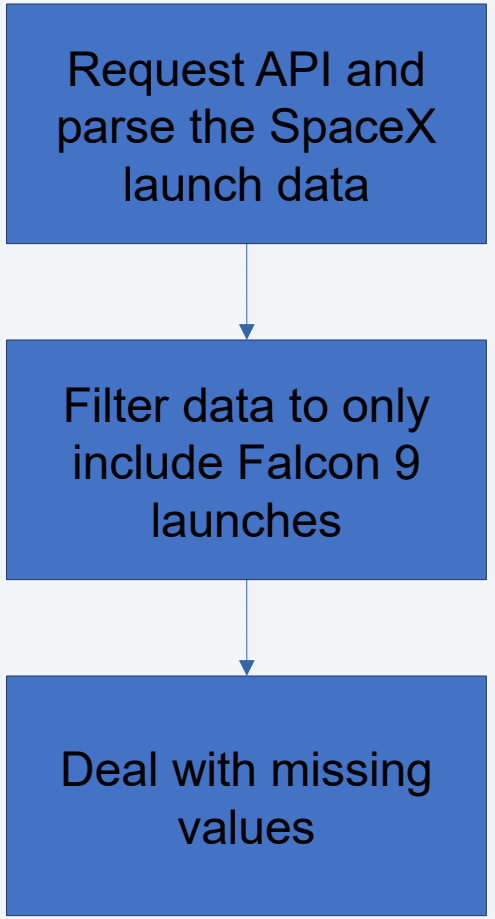
# Methodology

## Executive Summary

- Performed EDA using Python visualization libraries and SQL
- Performed interactive visual analytics with Folium and Plotly Dash
- Performed predictive analysis using many classification models
  - Selected best model based on prediction score on the testing set.
  - Found optimal parameters of each model with with GridSearchCV.

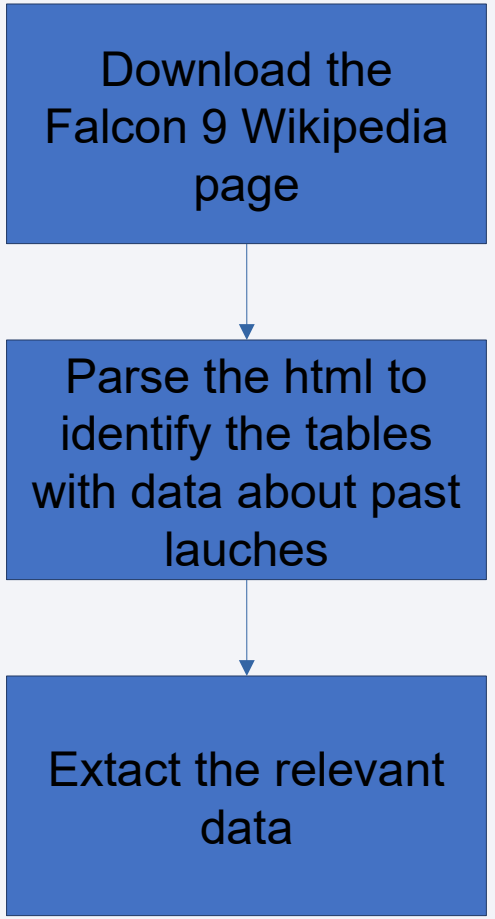
# Data Collection – SpaceX API

- SpaceX offers a public API from where data about past launches can be requested for free.
- Click here to see the relevant Jupyter notebook hosted on GitHub.



# Data Collection - Webscraping

- Data from past launches is also publically available on Wikipedia.
- It is presented in tables intended for human reading, but we can extract the data with Python's library BeautifulSoup.
- Click here to see the relevant Jupyter notebook hosted on GitHub.

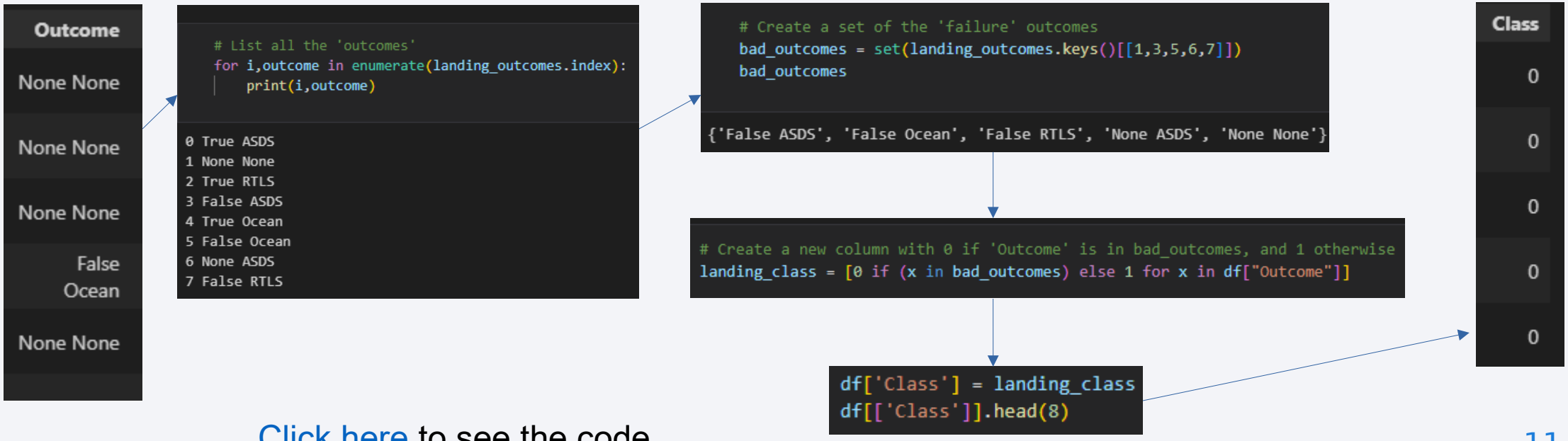


# Data Wrangling

In the raw dataset, the ‘success’ value is mixed with other information:

- True Ocean, True RTLS, True ASDS means the mission was successful.
- False Ocean, False RTLS, False ASDS means the mission was a failure.

We need to extract the True/False string into a categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.




Click here to see the code.

# EDA with Data Visualization

## Scatter Graphs

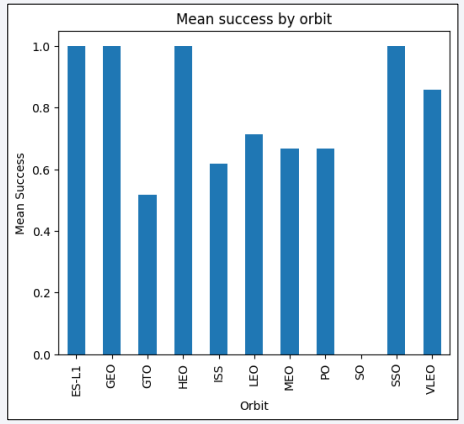
(shows relationship between variables to infer correlations)

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mass

## Bar Graph

(shows the relationship between numeric and categoric variables)

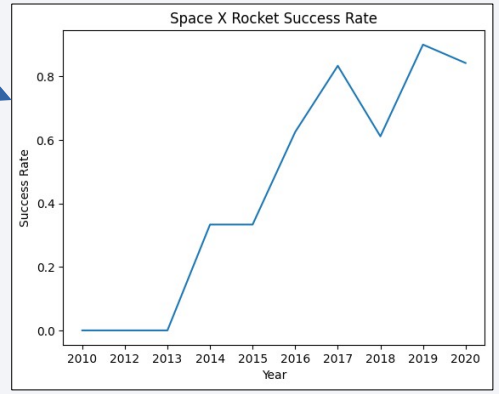
- Success rate vs. Orbit

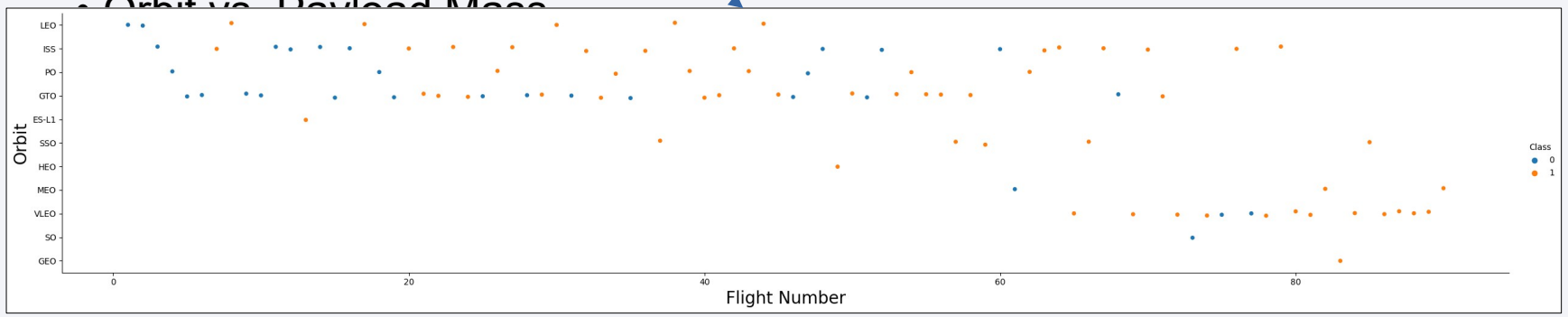



## Line Graph

(shows trends)

- Success rate vs. Year




[Click here](#) to see the code.

# EDA with SQL

**Performed SQL queries:**

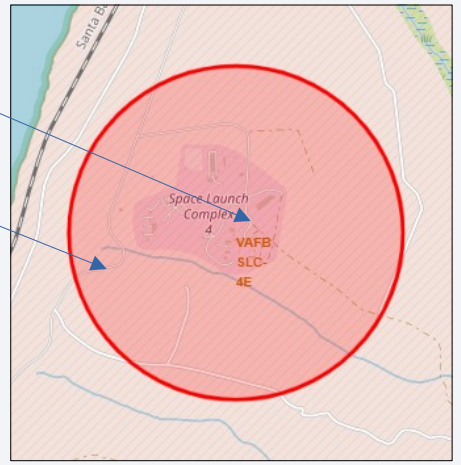
- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string ‘CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass.
- List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.
- Rank the count of landing outcomes (such as Failure at drone ship or Success at ground pad) between the date 2010-06-04 and 2017-03-20 in descending order.

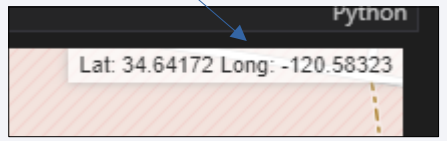
Click here to see the code.

# Build an Interactive Map with Folium

**Graphical entities used with Folium Maps:**

- Markers indicate points like launch sites.
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center.
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site.
- Mouse position is used to show the coordinates of the location under the cursor.
- Lines and small numeric labels are used to indicate distances between two points.





Click here to see the code.

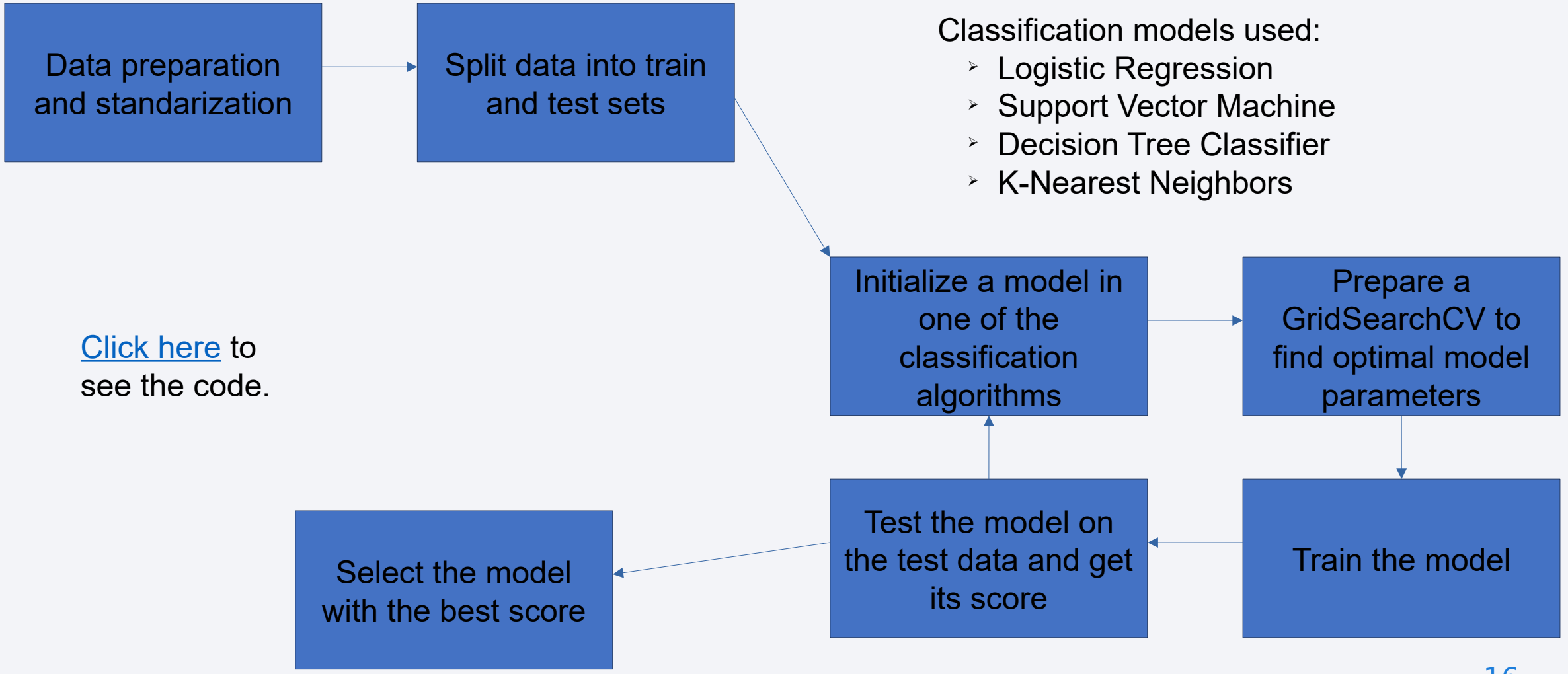
# Build a Dashboard with Plotly Dash

**The final dashboard has the following components:**

- Dropdown: allows a user to select a launch site (or all of them).
- Pie chart: shows the total success and the total failure for the selected launch site.
- Rangeslider: allows a user to select a payload mass in a fixed range.
- Scatter chart: shows the relationship between two variables, in particular Success vs Payload Mass.

Click here to see the code.

# Predictive Analysis (Classification)




Classification models used:

- Logistic Regression
- Support Vector Machine
- Decision Tree Classifier
- K-Nearest Neighbors

Click here to see the code.

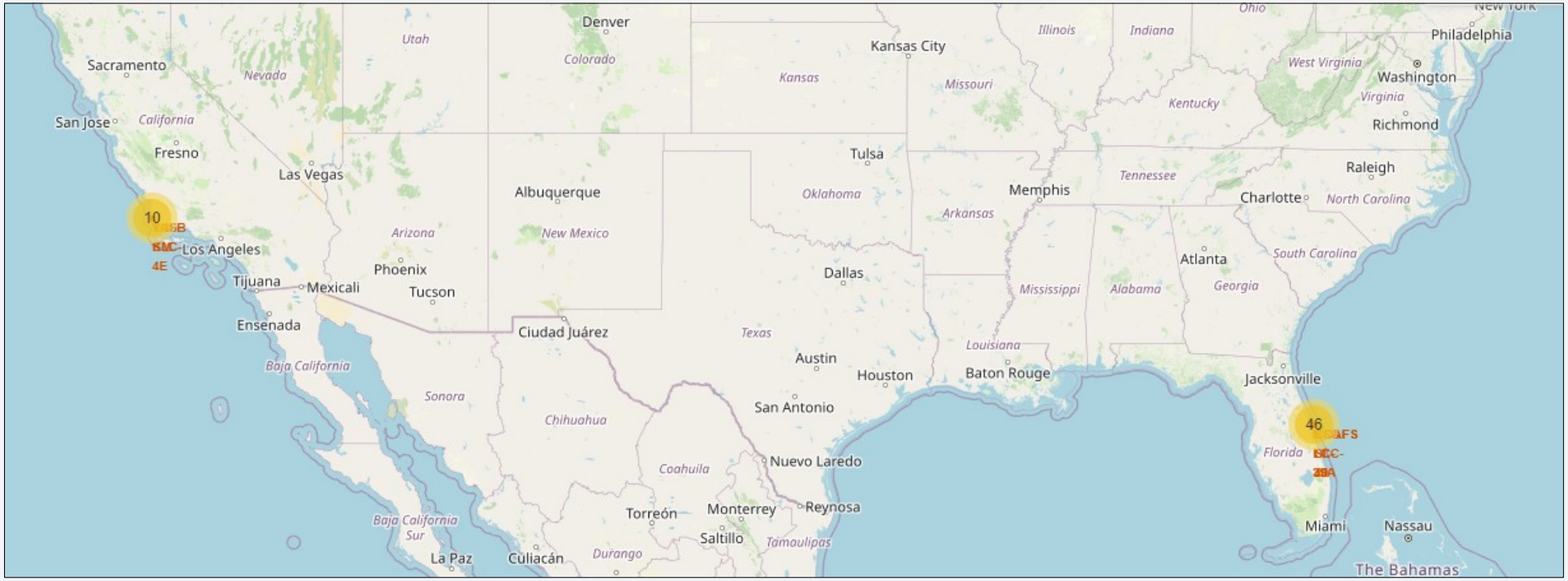
# Results

**Exploratory data analysis results:**

- Space X uses 4 different launch sites.
- The first launches were done to Space X itself and NASA.
- The average payload of F9 v1.1 booster is 2,928 kg.
- The first success landing outcome happened in 2015 fiver year after the first launch.
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
- Almost 100% of mission outcomes were successful.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
- The ratio of successful landing outcomes increased as years passed.

# Results

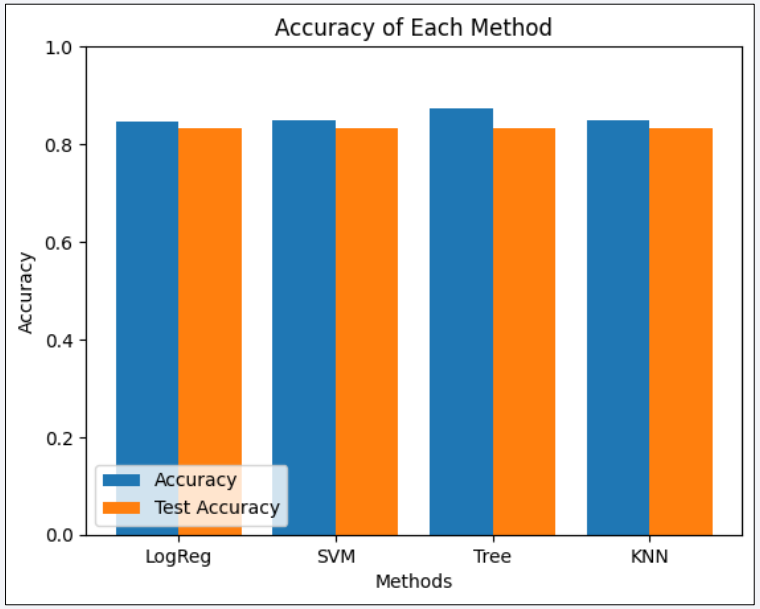
- Interactive mapping revealed  that launch sites are usually in safe places away from population and near the sea.
- They also have good logistic infrastructure around.
- Most launches happens at east cost launch sites.



# Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy of approximately 87%, and a similar test accuracy to other models.

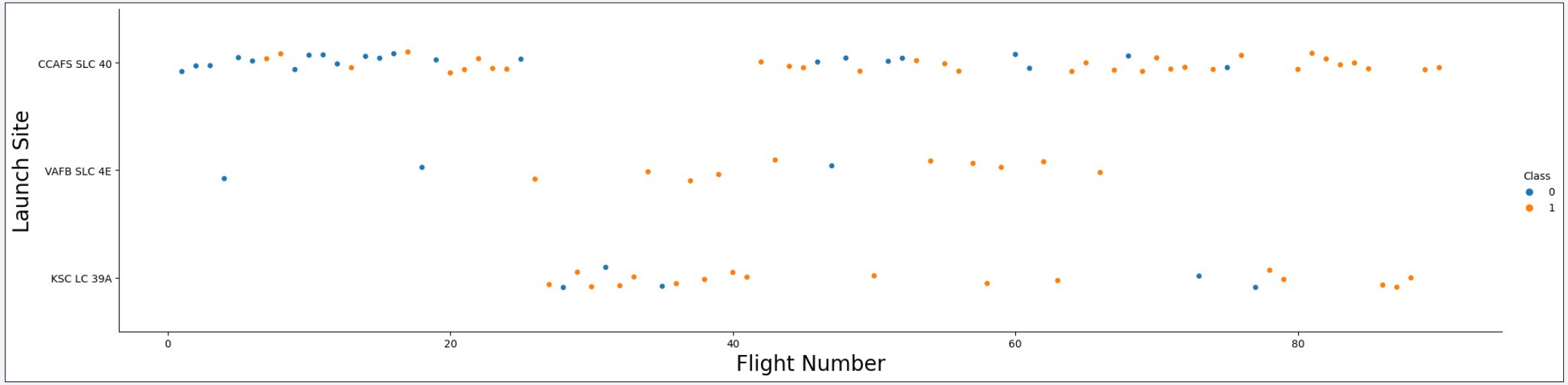
| Model | Accuracy | TestAccuracy |
|---|---|---|
| LogReg | 0.84643 | 0.83333 |
| SVM | 0.84821 | 0.83333 |
| Tree | 0.87321 | 0.83333 |
| KNN | 0.84821 | 0.83333 |



Section 2

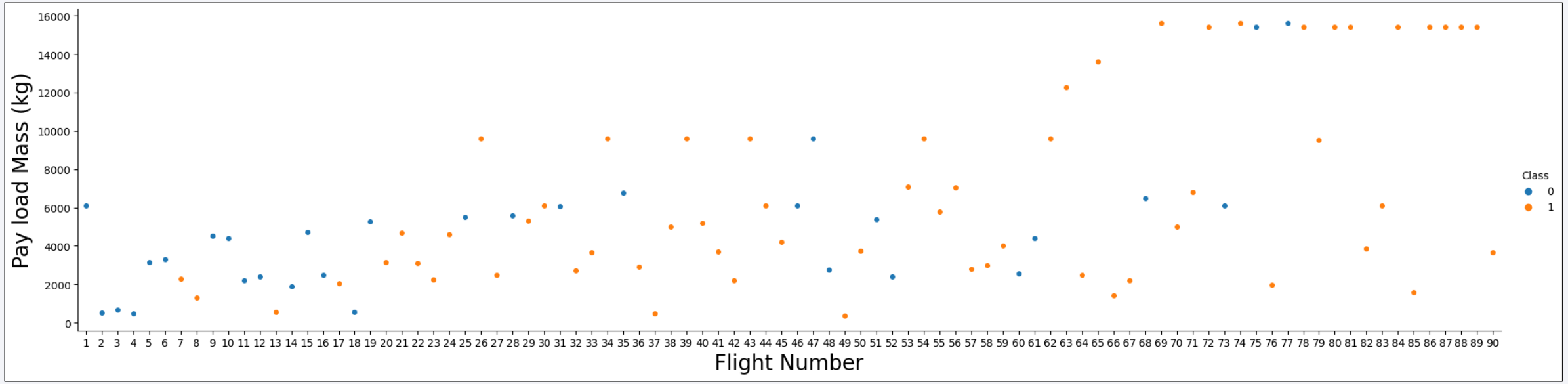
# Insights drawn from EDA

# Flight Number vs. Launch Site



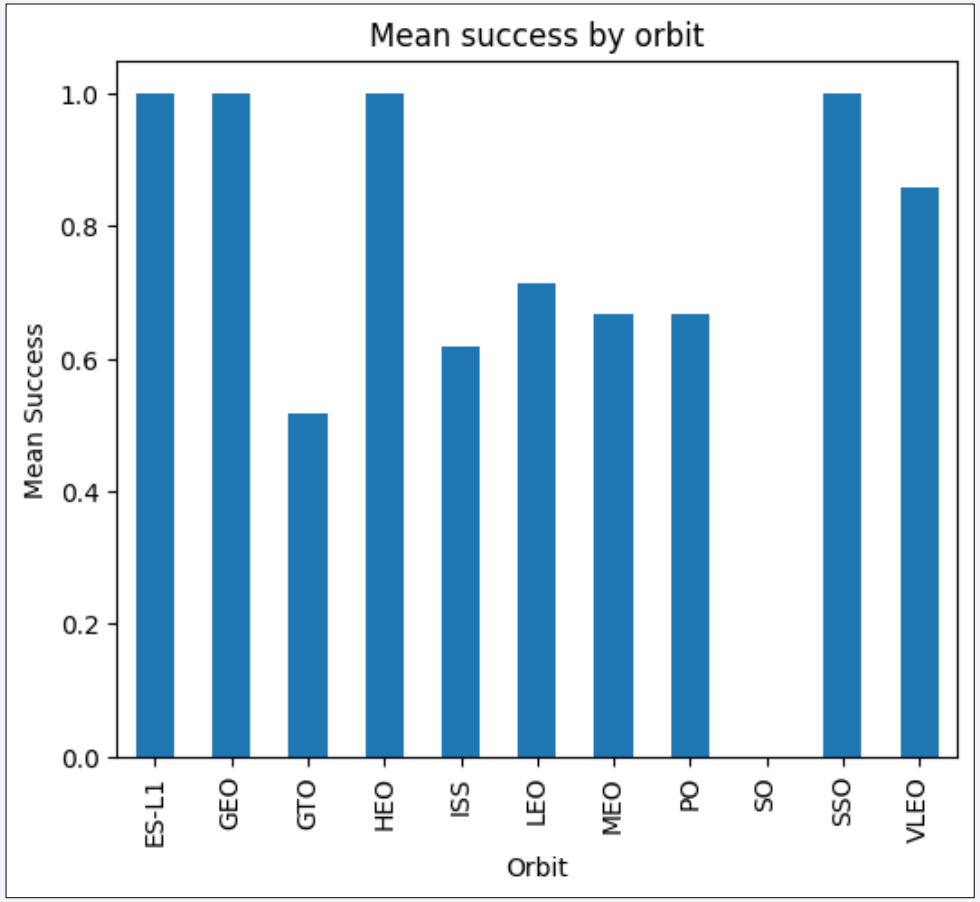

- The best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
- In second place there's VAFB SLC 4E, and in third place KSC LC 39A.
- The general success rate has improved over time for all sites.

# Payload vs. Launch Site




- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

# Success Rate vs. Orbit Type




The biggest success rates (100%) happens when lauching to orbits:

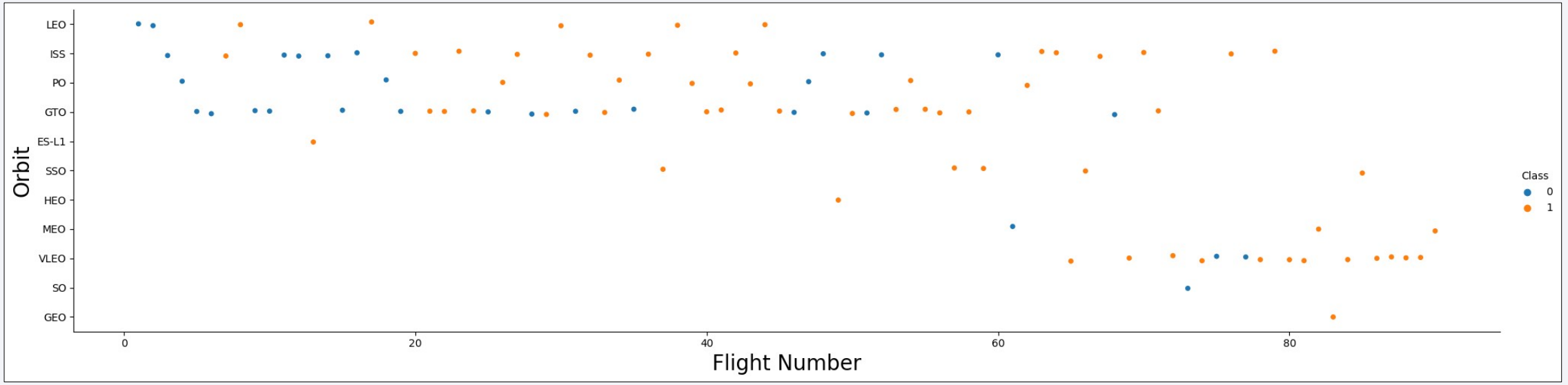
- ES-L1
- GEO
- HEO
- SSO

Followed closely by:

- VLEO (above 80%)

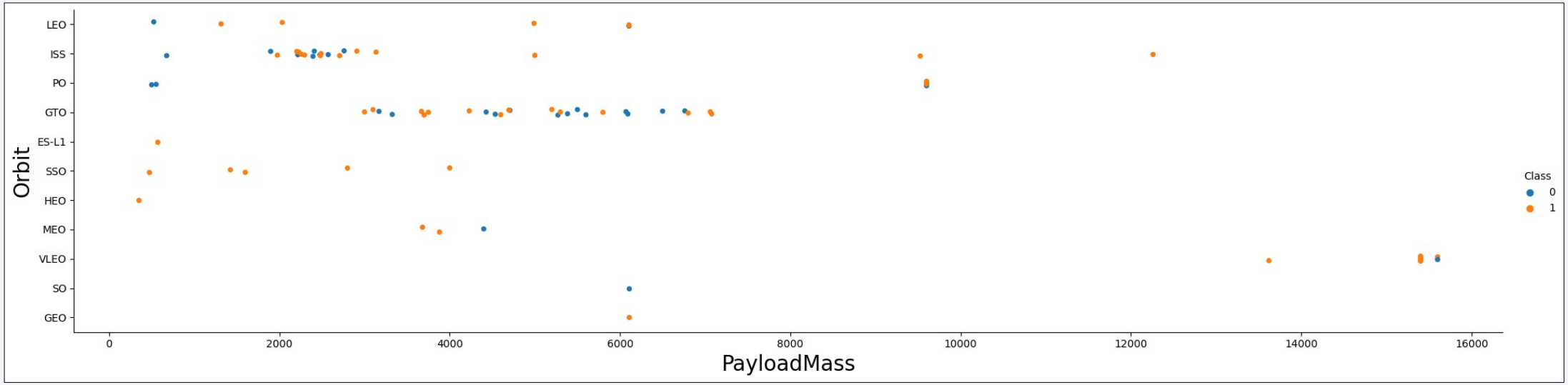
Note that a high success rate might be due to luck if the orbit has been used only a handful of times.

# Flight Number vs. Orbit Type




- Observations:
  - Success rate improved over time regardless of the orbit.
  - VLEO orbit seems a new business opportunity. We saw that it had a lowers success rate than others (80% vs 100%) but it has had many more attempts.

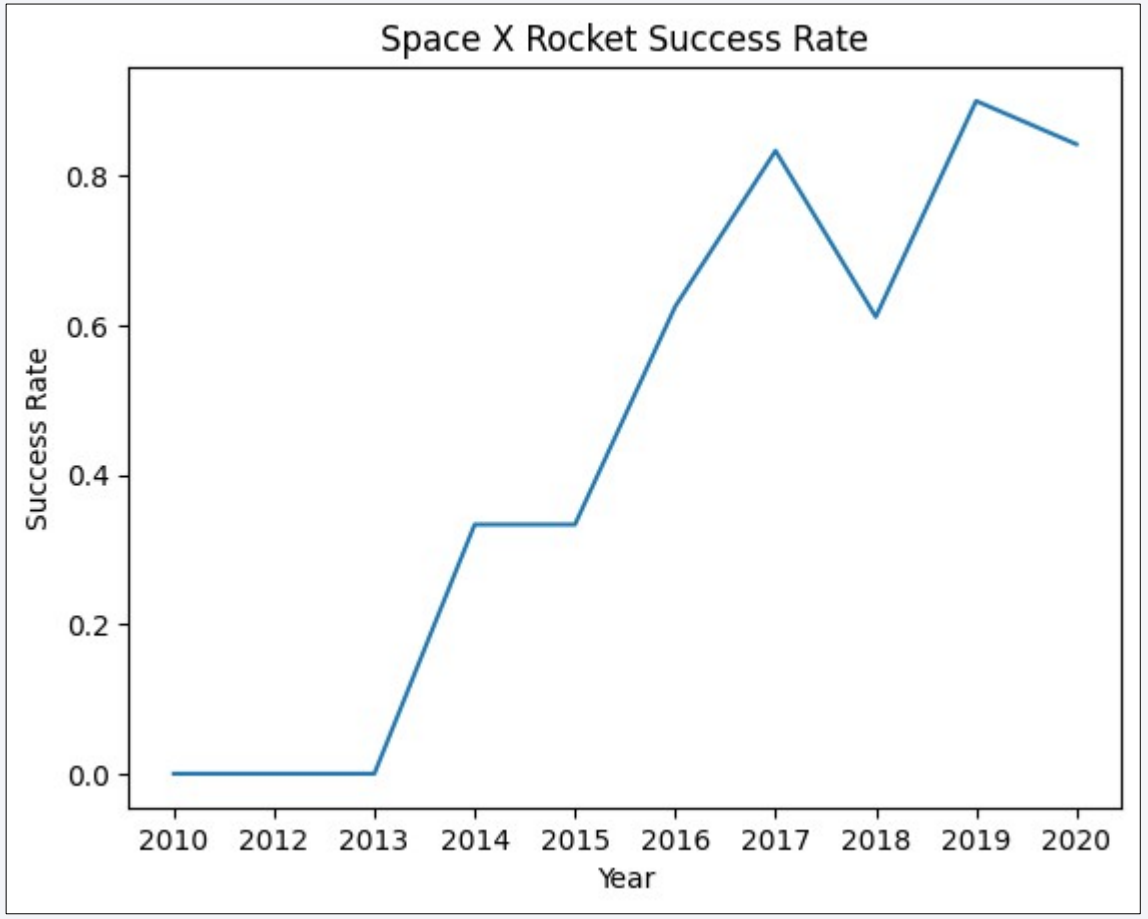
# Payload vs. Orbit Type




Observations:

- There is no significant relationship between payload and success rate to orbit GTO.
- ISS orbit has the widest range of payload and a good rate of success.
- There are very few launches to the orbits SO and GEO, and of a significantly high payload.

# Launch Success Yearly Trend



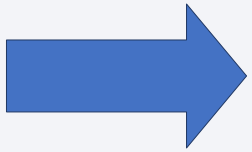

Observations:

- Success rate started increasing in 2013 and kept rising until 2020 (with a dip in 2018).
- It seems that the first three years were a period of experimentation of the technology in its infancy.

# All Launch Site Names

This query returned a list of all used launch sites:

```
%%sql
SELECT DISTINCT(LAUNCH_SITE) FROM SPACE;
```

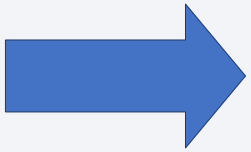


| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

# Launch Site Names Begin with 'KSC'

This query found 5 records where launch sites begin with `KSC`:

```
%%sql
SELECT DATE, TIME__UTC_, BOOSTER_VERSION, LAUNCH_SITE
    FROM SPACE
    WHERE LAUNCH_SITE LIKE 'KSC%'
    LIMIT 5;
```

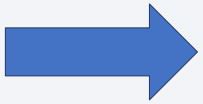


| DATE | time__utc_ | booster_version | launch_site |
|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A |
| 2017-03-16 | 06:00:00 | F9 FT B1030 | KSC LC-39A |
| 2017-03-30 | 22:27:00 | F9 FT B1021.2 | KSC LC-39A |
| 2017-05-01 | 11:15:00 | F9 FT B1032.1 | KSC LC-39A |
| 2017-05-15 | 23:21:00 | F9 FT B1034 | KSC LC-39A |

# Total Payload Mass

This query calculated the total payload carried by boosters from NASA:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload"
    FROM SPACE
    WHERE CUSTOMER = 'NASA (CRS)';
```



| Total Payload |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

This query calculates the average payload mass carried by booster version F9 v1.1:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE"
    FROM SPACE
    WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

| average |
| --- |
| 2534 |

# First Successful Ground Landing Date

This query finds the dates of the first successful landing outcome on ground pad:

```
%%sql
SELECT MIN(DATE) AS "First_Success"
    FROM SPACE
    WHERE LANDING__OUTCOME LIKE 'Success%';
```

| First_Success |
| --- |
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

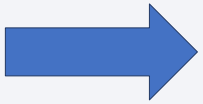
```
%%sql
SELECT BOOSTER_VERSION
    FROM SPACE
    WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
    AND LANDING__OUTCOME = 'Success (ground pad)';
```

| booster_version |
| --- |
| F9 FT B1032.1 |
| F9 B4 B1040.1 |
| F9 B4 B1043.1 |

# Total Number of Successful and Failure Mission Outcomes

This query calculates the total number of successful and failure mission outcomes:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS "Number"
    FROM SPACE
    GROUP BY MISSION_OUTCOME;
```



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

# Boosters Carried Maximum Payload

This query lists the names of the boosters which have carried the maximum payload mass:

```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_
    FROM SPACE
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACE);
```

| 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 |

# 2017 Successful Launch Records

This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2017:

```
%%sql
SELECT MONTHNAME(DATE) AS "Month", BOOSTER_VERSION AS "Booster Version", LAUNCH_SITE AS "Launch Site"
    FROM SPACE
    WHERE LANDING__OUTCOME='Success (ground pad)'
    AND YEAR(DATE)='2017';
```

| Month | Booster Version | Launch Site |
|---|---|---|
| February | F9 FT B1031.1 | KSC LC-39A |
| May | F9 FT B1032.1 | KSC LC-39A |
| June | F9 FT B1035.1 | KSC LC-39A |
| August | F9 B4 B1039.1 | KSC LC-39A |
| September | F9 B4 B1040.1 | KSC LC-39A |
| December | F9 FT B1035.2 | CCAFS SLC-40 |

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

This query ranks the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order:

```
%%sql
SELECT LANDING__OUTCOME AS "Landing Outcome", COUNT(*) AS "Count"
    FROM SPACE
    WHERE DATE BETWEEN '4/06/2010' AND '20/03/2017'
    AND LANDING__OUTCOME LIKE 'Success%'
    GROUP BY LANDING__OUTCOME
    ORDER BY Count DESC;
```

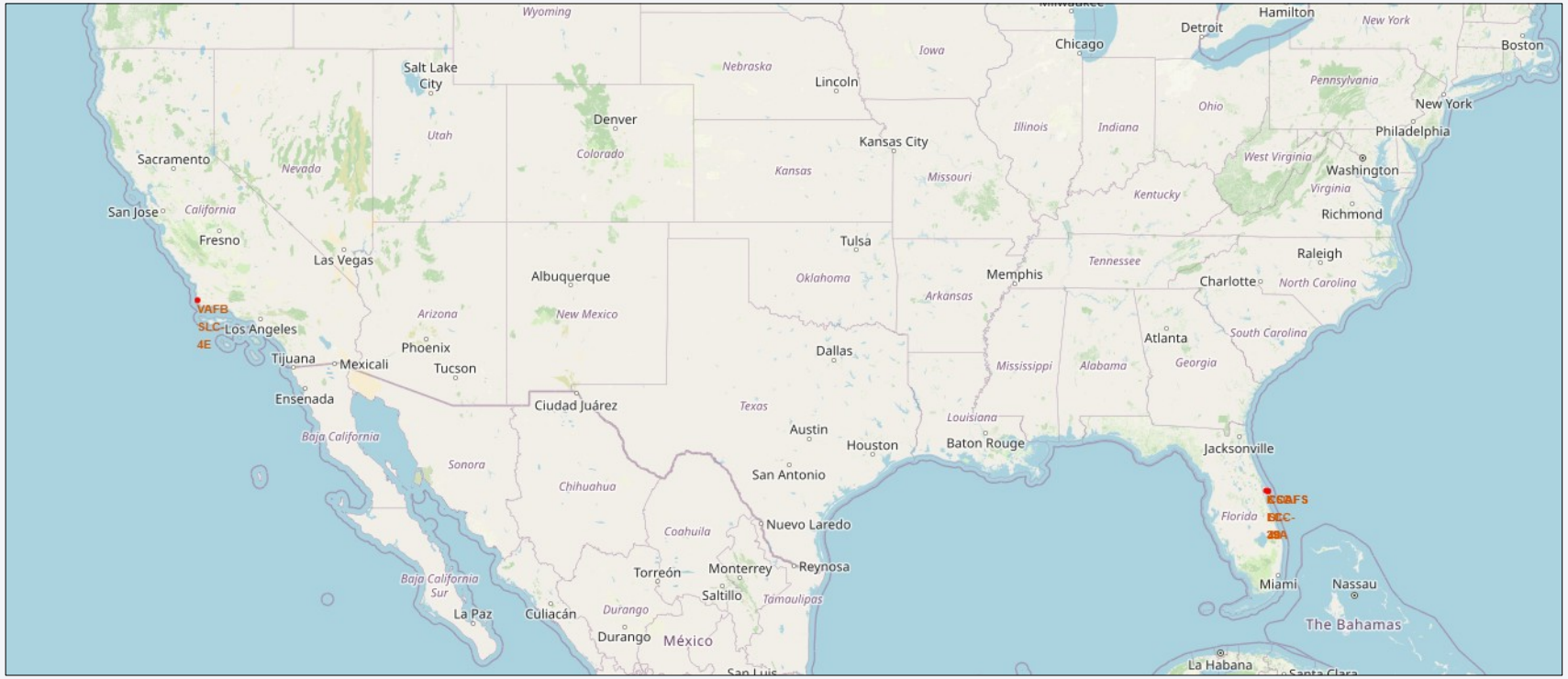
| Landing Outcome | Count |
|---|---|
| Success (drone ship) | 5 |
| Success (ground pad) | 3 |

Section 3

# Launch Sites Proximities Analysis

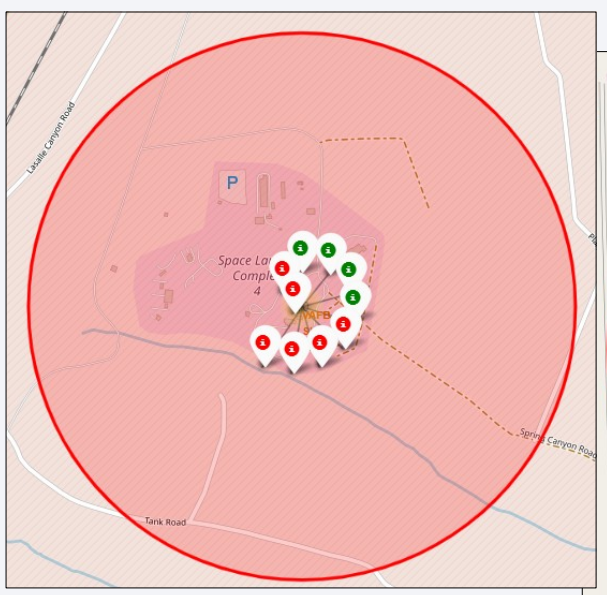
# All launch locations

The following map shows the launch locations used by SpaceX. They are always next to the ocean.

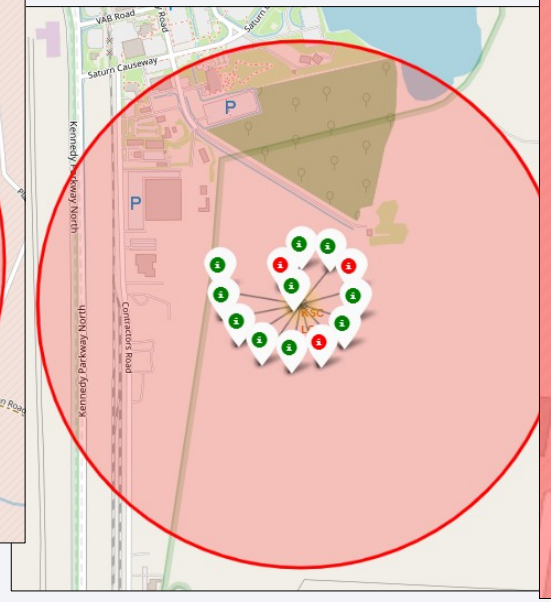


# Launch outcomes by location

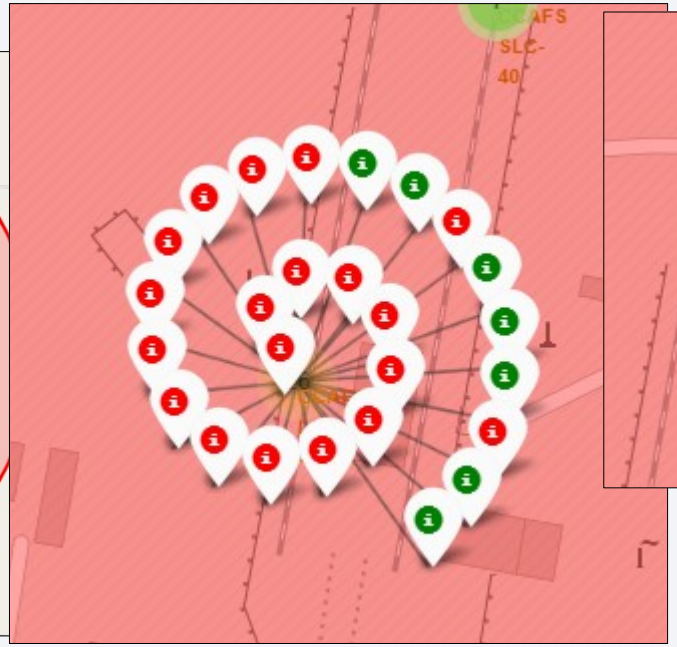
These are the color-labeled launch outcomes on each location (green for success, red for failure).



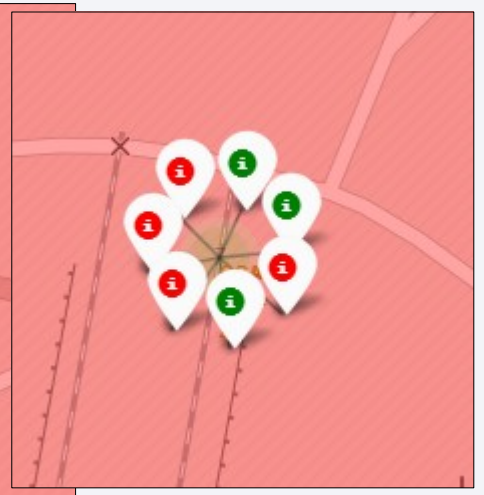

VAFB SLC-4E




KSC LC-39A



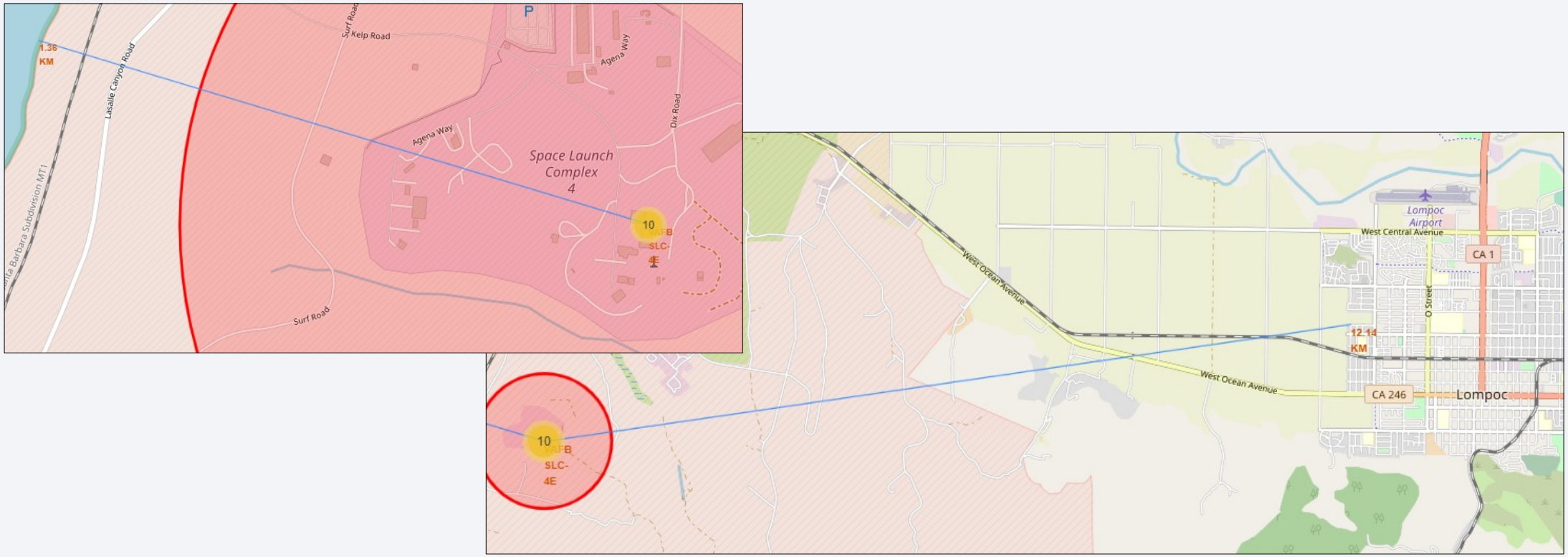

CCAFS LC-40

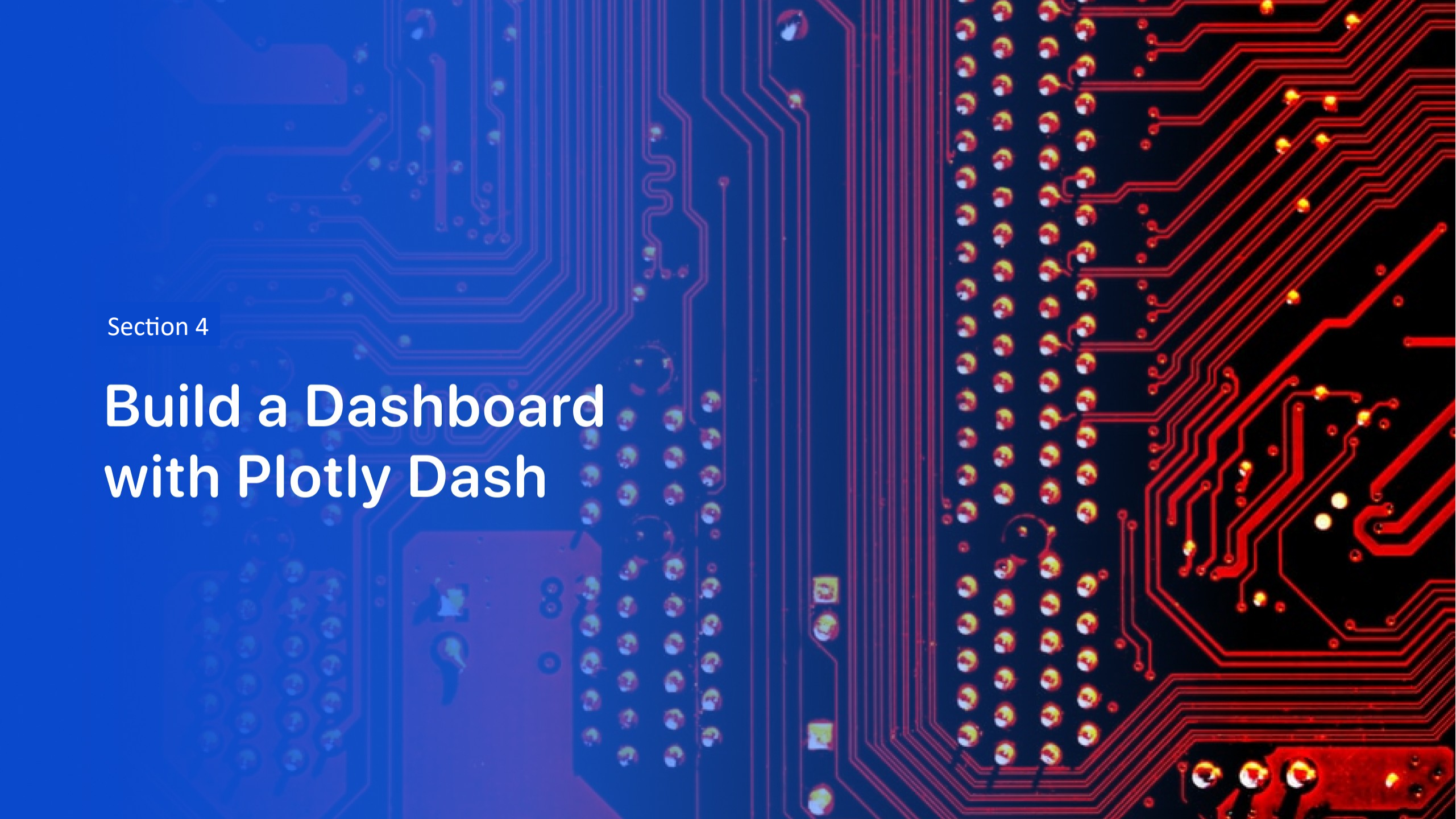


CCAFS SLC-40

# Launch site surroundings

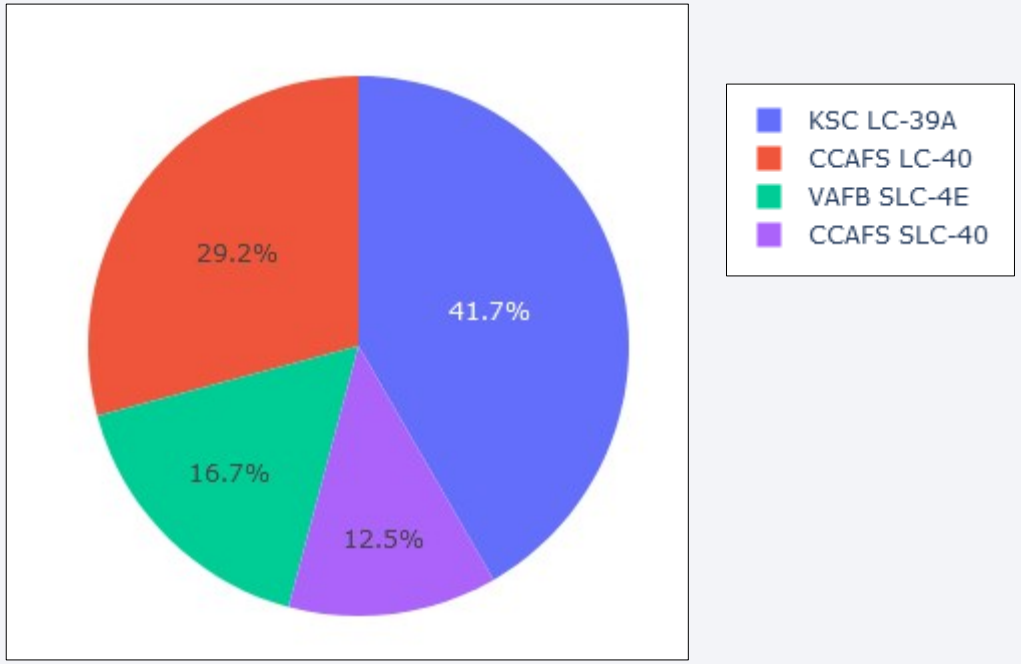
These screenshots show a launch site located close to the ocean and far from the nearest city, for safety of the population. Railroads are also relatively close for logistics. This pattern is true for all launch sites.




Section 4
Build a Dashboard with Plotly Dash

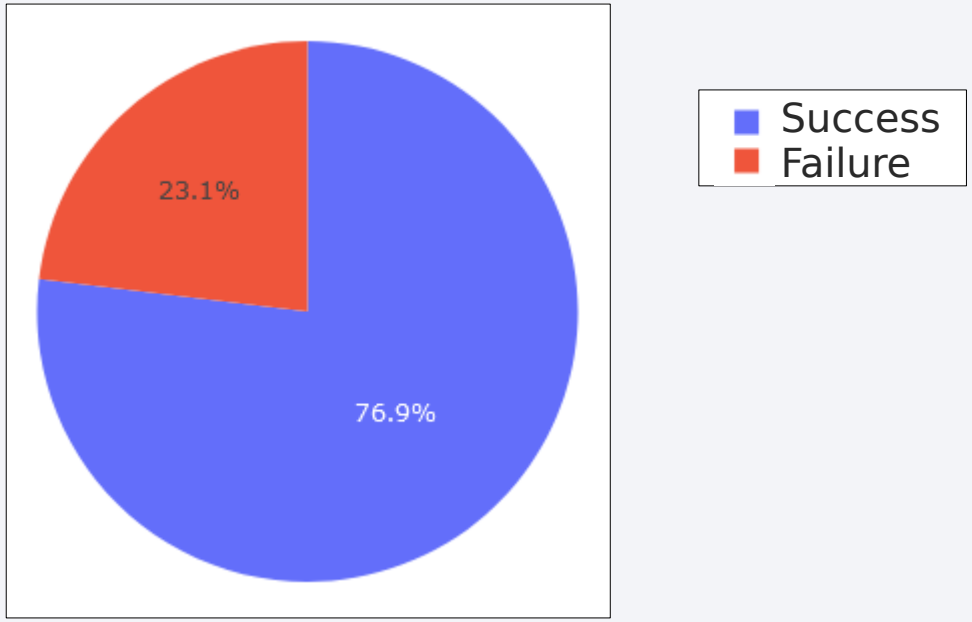
# Successful launch percentage for every site

This piechart shows how all the successful launches are distributed among the different launch sites.



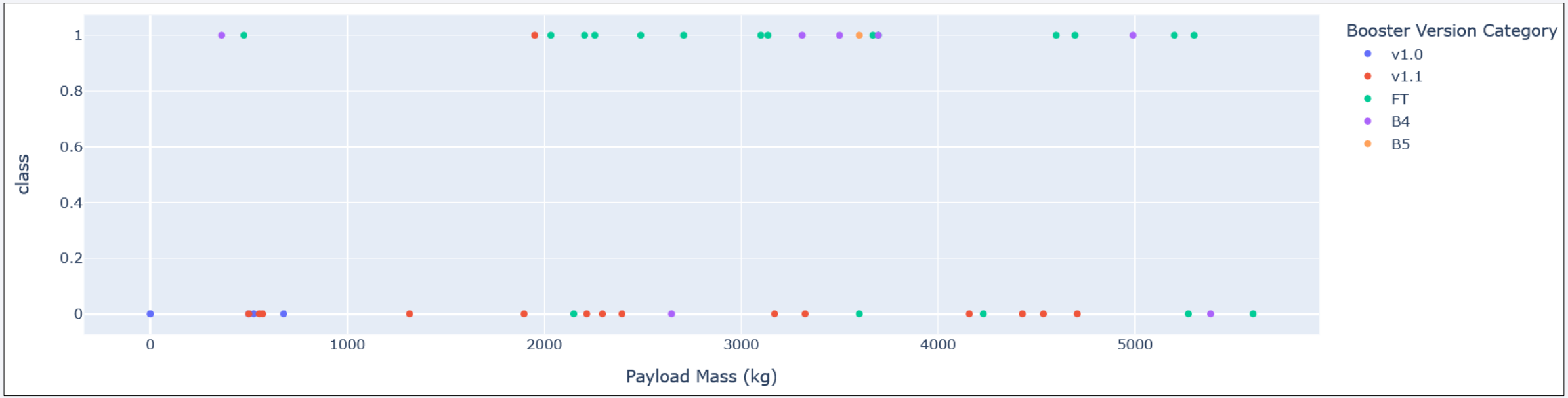
# Launch success ratio for site KSC LC-39A

KSC LC-39A is the site with the highest successful launch ratio.



# Payload Mass vs Launch Outcome (small loads)

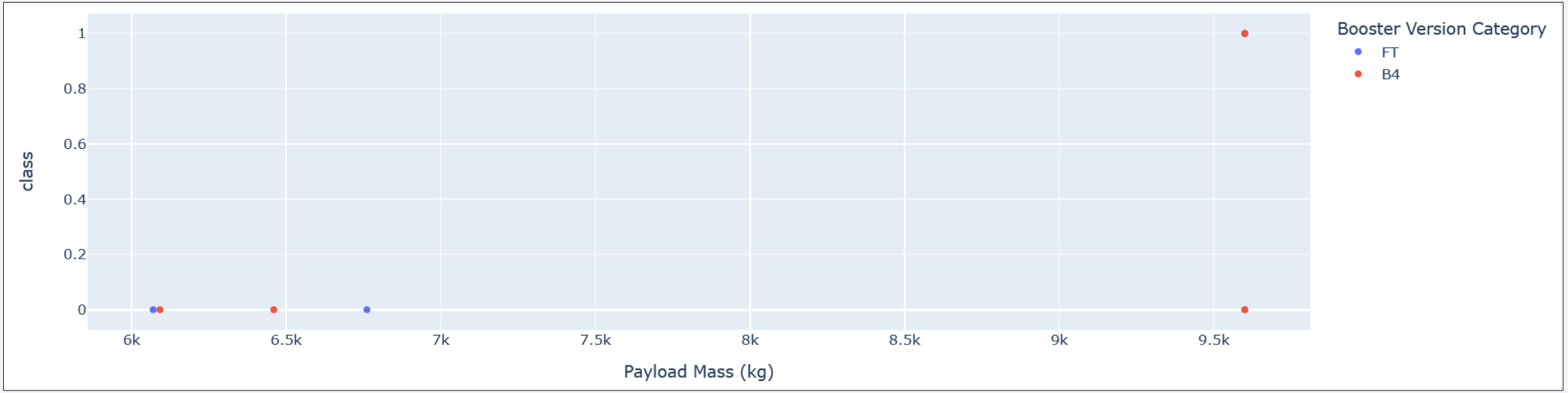
This scatter plot shows Payload vs. Launch Outcome for all sites, with payloads up to 6000 kg.




At these loads, FT is the most reliable booster version.

# Payload Mass vs Launch Outcome (large loads)

This scatter plot shows Payload vs. Launch Outcome for all sites, with payloads from 6000 kg to 10000 kg.




At these loads, there is a low chance of success. Only the B4 booster has achieved one successful launch.

Section 5

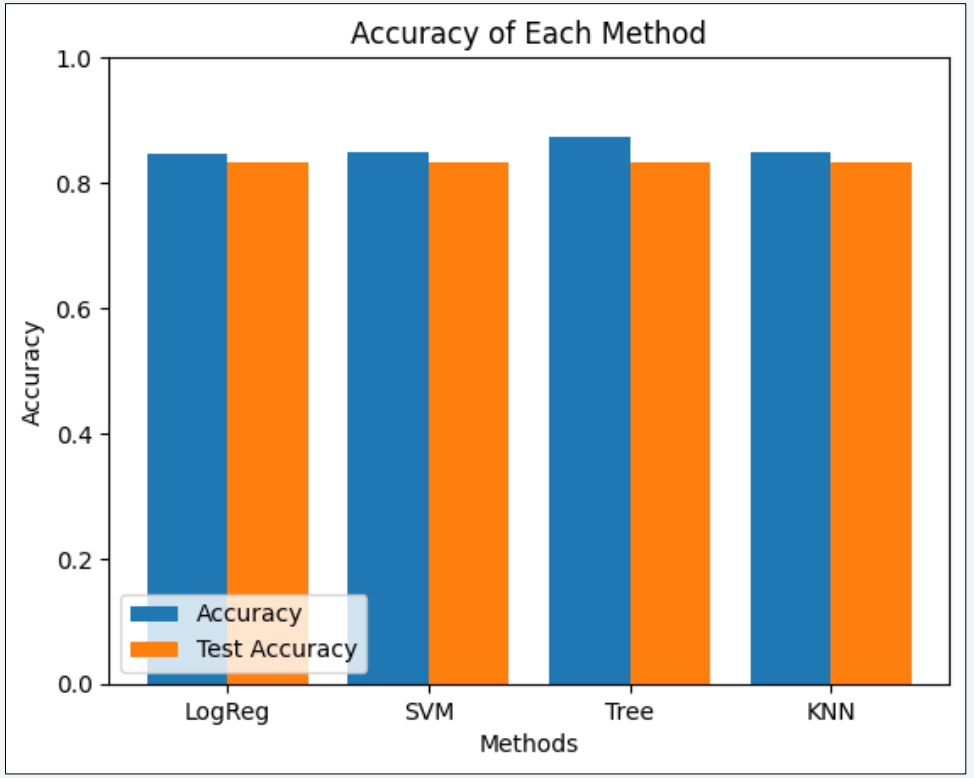
# Predictive Analysis (Classification)

# Classification Accuracy

Four classification models were built and optimized:

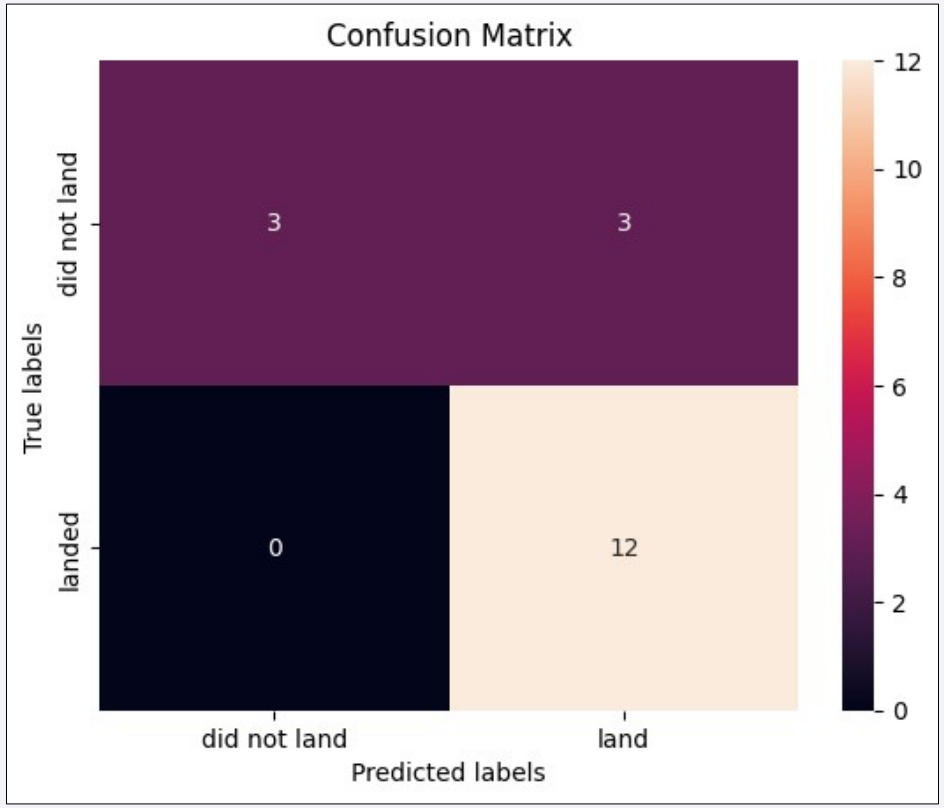
- Logistic regression
- SVM
- Tree classifier
- K-Nearest neighbors

For accuracy on the test set, all methods performed similar. We could repeat the model creation with different random training and testing sets, but if we really need to choose one right now, we would take the decision tree (87% accuracy in training).



# Confusion Matrix

- All the confusion matrices of each model were the same. They all predict negatives well but have a problematic number of false positives.



# Conclusions

- The best launch site is KSC LC-39A.
- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches.
- Successful landing outcomes seem to improve over time. We can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- All the sites are in very close proximity to the coast and logistic facilities, while being far from urban centers.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

# Thank you!