IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gonzalo Nicolás Martínez Carreras
14/04/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API and Web Scraping;
  - Exploratory Data Analysis (EDA) with Data wrangling, SQL, Data Visualization and Interactive visual analytics with Folium.
  - Machine Learning for Predictive Analysis.
- Summary of all results
  - Exploratory data analysis results
  - Interactive maps and dashboard
  - Predictive results

Introduction

- Project background and context

The objective of the project is to provide information to find out if other companies can compete with SpaceX for the launch of a rocket. This by predicting successful landings of the first stage of the rocket. Since, in principle, the cost difference lies mainly in the reuse of the first stage.

- Problems you want to find answers

Mainly knowing which is the best place to launch. Added to which are the variables that affect the probability of success or failure of the landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was collected from 2 sources:
    - Space X API
    - Web Scraping from Wikipedia
- Perform data wrangling
  - Removing dispensable columns and using classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was divided into training and test sets. They were evaluated using different classification models. Finally, the precision of each model was estimated.

Data Collection

- Data sets were collected from:
  - SpaceX Rest API
  - Wikipedia using webscraping.

Data Collection – SpaceX API

SpaceX offers an API where the data can be used freely and free of charge.
The following steps were followed in order to have the necessary data available and in a clear way.

Flowchart of Rest API:

1. Getting Response from API;
2. Converting to JSON File;
3. Create dictionary;
4. Create data frame;
5. Filter data frame.

https://github.com/gonmartinezc/IBM_Capstone/blob/21b067a870093a3a6b1082b5fe39b417a2b6c4f5/01%20-%20Labs%20Collecting%20Data.ipynb

Data Collection – Scraping

The data was downloaded from Wikipedia, using web scraping, following the steps below:

Flowchart of web scraping

1. Getting response from HTML;
2. Create BeautifulSoup object;
3. Finding all tables;
4. Create dictionary;
5. Add data to keys;
6. Create dataframe.

https://github.com/gonmartinezc/IBM_Capstone/blob/21b067a870093a3a6b1082b5fe39b417a2b6c4f5/01%20-%20Labs%20Collecting%20Data.ipynb

Data Wrangling

- An exploratory data analysis (EDA) was performed, key variables were calculated. String variables were transformed into Dummy type categories.

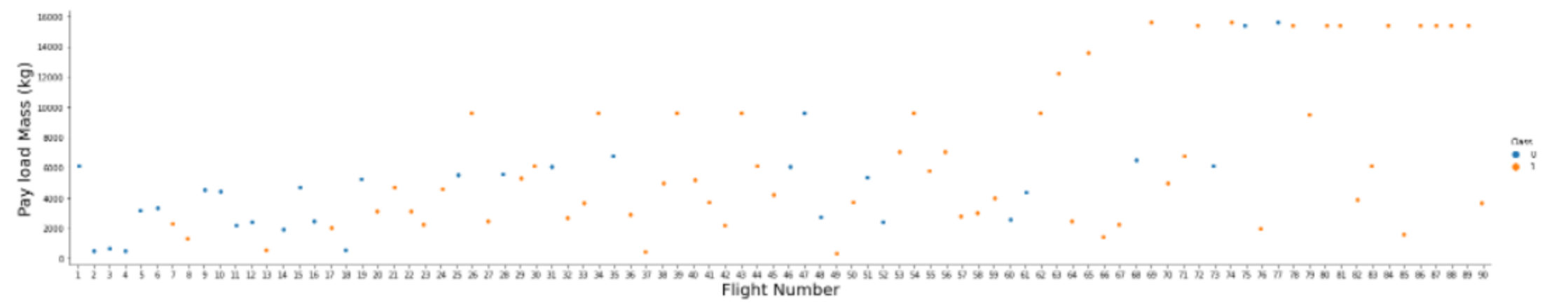
Flowchart of data wrangling

1. The number of launches for each site was calculated;
2. The number and occurrence of each orbit was calculated;
3. The number and occurrence of the mission result by type of orbit was calculated;
4. Created landing result tag.

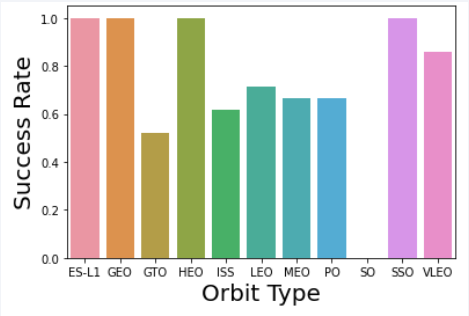
https://github.com/gonmartinezc/IBM_Capstone/blob/f9258c074176fcf4f0f1c308fa324a6af339d27e/02%20%20-Labs%20Data%20wrangling.ipynb

EDA with Data Visualization

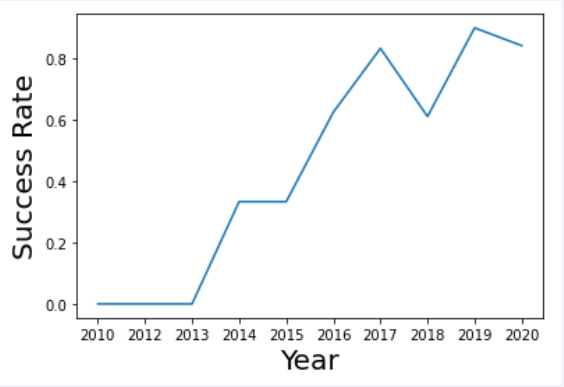
- I used scatter plots to visualize the correlation between some variables.



- With the b
  numerical and categorical variables.



- I used the line graphs to visualize the global behavior of the variables.



https://github.com/gonmartinezc/IBM_Capstone/blob/f9258c074176fcf4f0f1c308fa324a6af339d27e/04%20-%20Labs%20EDA%20Visualization.ipynb

EDA with SQL

- Displaying 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 records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site forthe months in year 2015;
- Rank the count of successful landing_outcomes between the date04-06-2010 and 20-03-2017in descending order.

https://github.com/gonmartinezc/IBM_Capstone/blob/f9258c074176fcf4f0f1c308fa324a6af339d27e/03%20-%20Labs%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Analysis was focused on NASA's Johnson Space Center in Houston using folium map.

- Launch sites are indicated with dots;
- Circles highlight areas around specified coordinates;
- Markup groups were used to highlight events in a coordinate;
- The distance between two specific points on the map are measured by lines.

https://github.com/gonmartinezc/IBM_Capstone/blob/f9258c074176fcf4f0f1c308fa324a6af339d27e/05%20-%20Labs%20Visualization%20Folium%20.ipynb

Build a Dashboard with Plotly Dash

First clarify that there is a drop-down menu that allows you to choose the launch site.

The pie chart shows the success and failure of the launch site.

Also, you can choose a payload mass in a fixed range using the Rangeslider and then with the scatter plot we see the relationship between Success and Payload Mass.

Using this board, we can identify the best place for launches.

https://github.com/gonmartinezc/IBM_Capstone/blob/675d7a591f783955f15a3d76e26d3db303d56a6e/spacex_dash_app.py

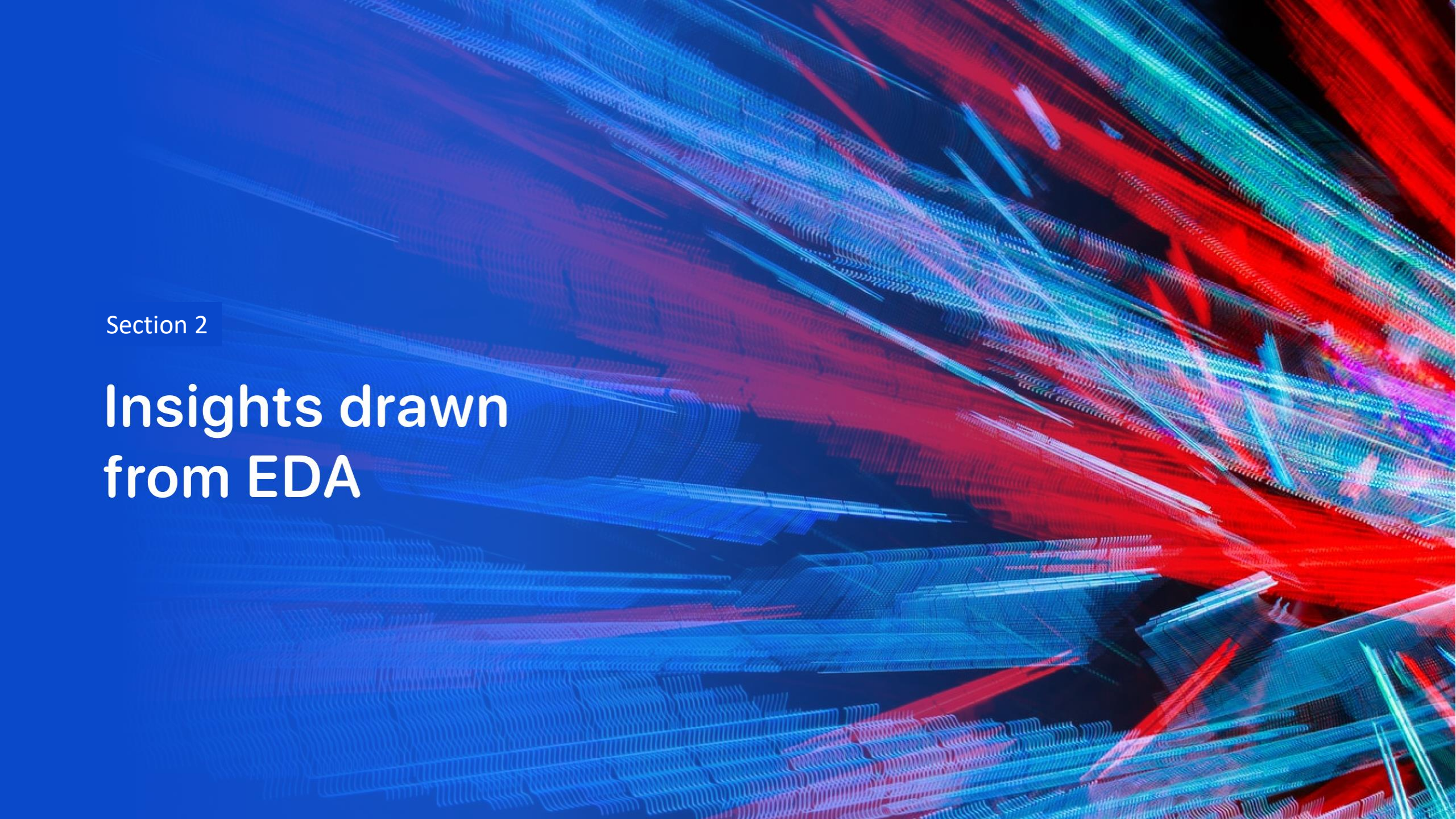
Predictive Analysis (Classification)

- The following classification models were used:
  - Logistic regression;
  - Support vector machine;
  - Decision tree;
  - K-nearest neighbors.
- The following classification models were used:
  - Data preparation;
  - Model preparation;
  - Model evaluation;
  - Model comparison for results.

https://github.com/gonmartinezc/IBM_Capstone/blob/675d7a591f783955f15a3d76e26d3db303d56a6e/06%20-%20Labs%20Predictive.ipynb

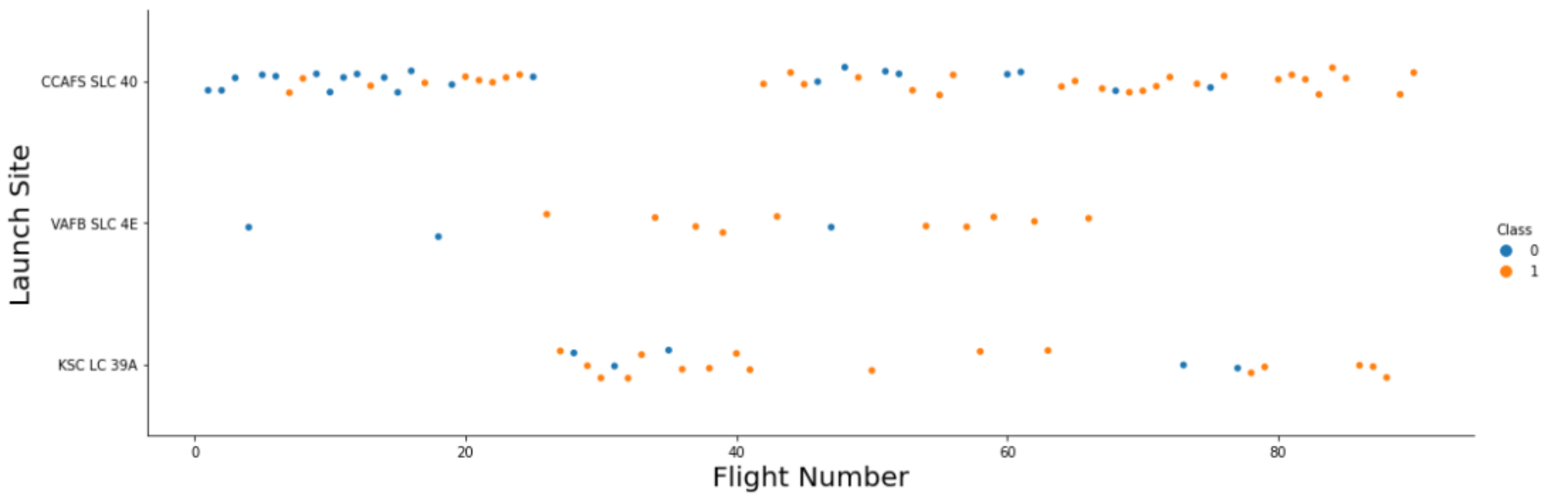
Results

- Exploratory data analysis results
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - The number of landing outcomes became as better as years passed;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed F9 v1.1 B1012 and F9 v1.1 B1015.
- Interactive analytics demo in screenshots
  - Most launches happens at east cost launch sites because it’s a safety place.
  - This improves the probability of success.
- Predictive analysis results
  - All predictive models have a high accuracy, exceeding 80% in all cases. Decision Tree Classifier is the model that best predicted successful landings.


Section 2
Insights drawn from EDA

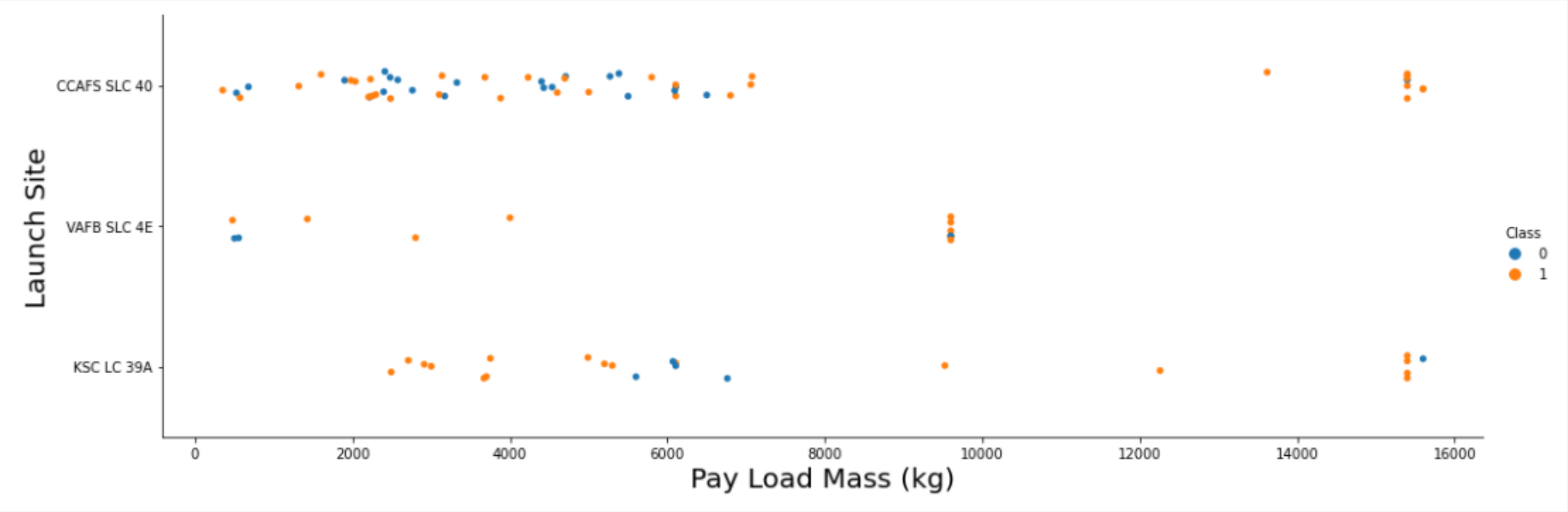
Flight Number vs. Launch Site

This graph allows us to verify that the general success rate is increasing over time, which is expected over time and all the advances that are applied.



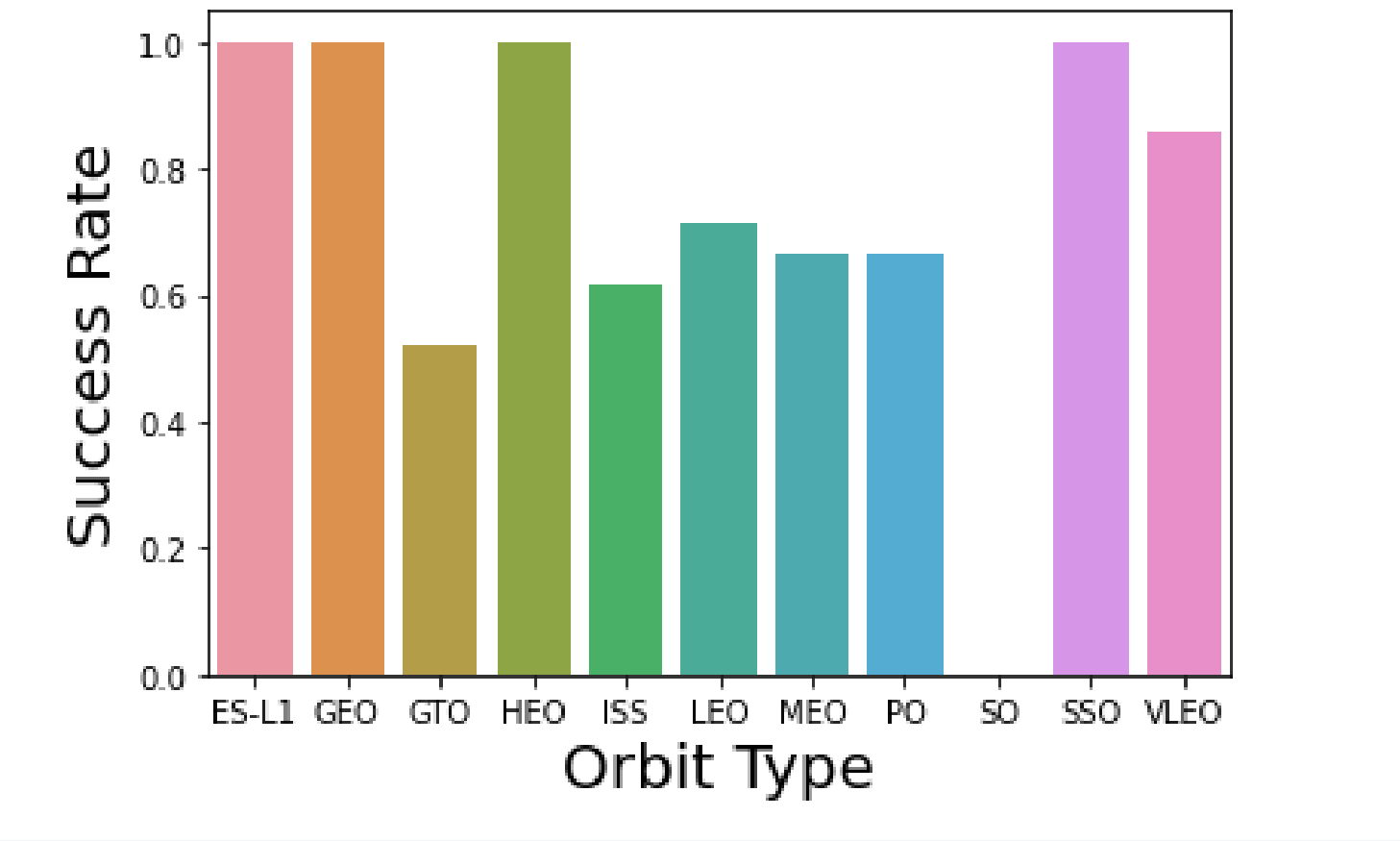
Payload vs. Launch Site

This graphic shows the relationship between launch site and payload. Very heavy payloads can begin to complicate the landing, it seems that the most optimal would be around 9,500 kg.



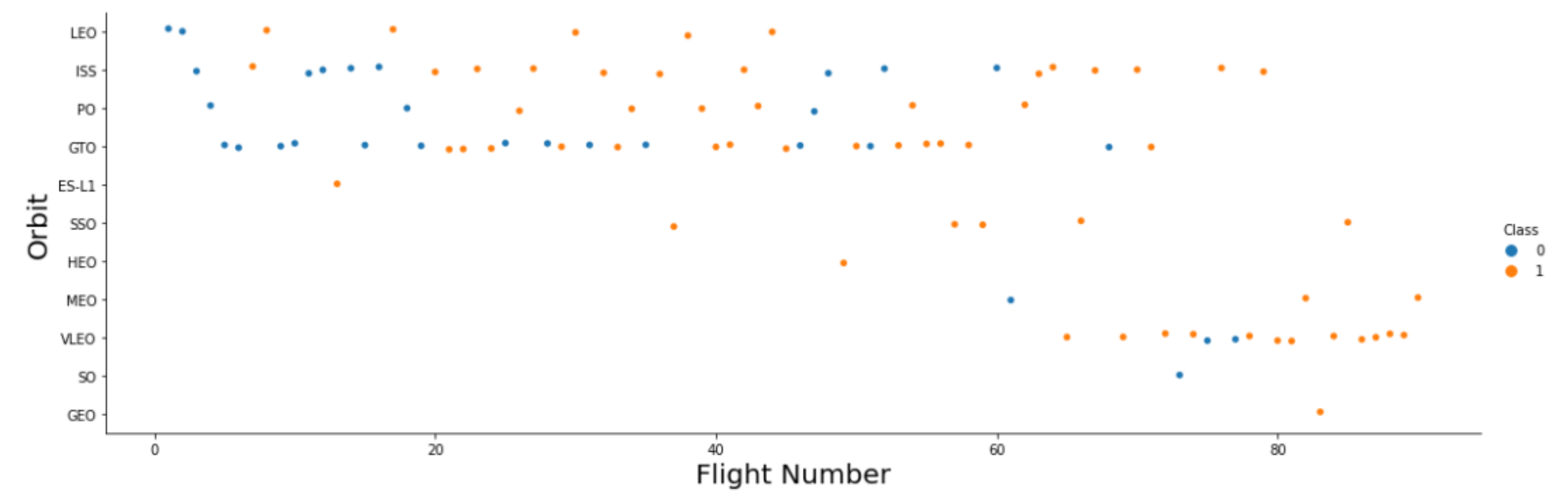
Success Rate vs. Orbit Type

This bar graph allows us to see the probability of success according to the different types of orbits. Where we can see which are those that would give us a greater chance of success.



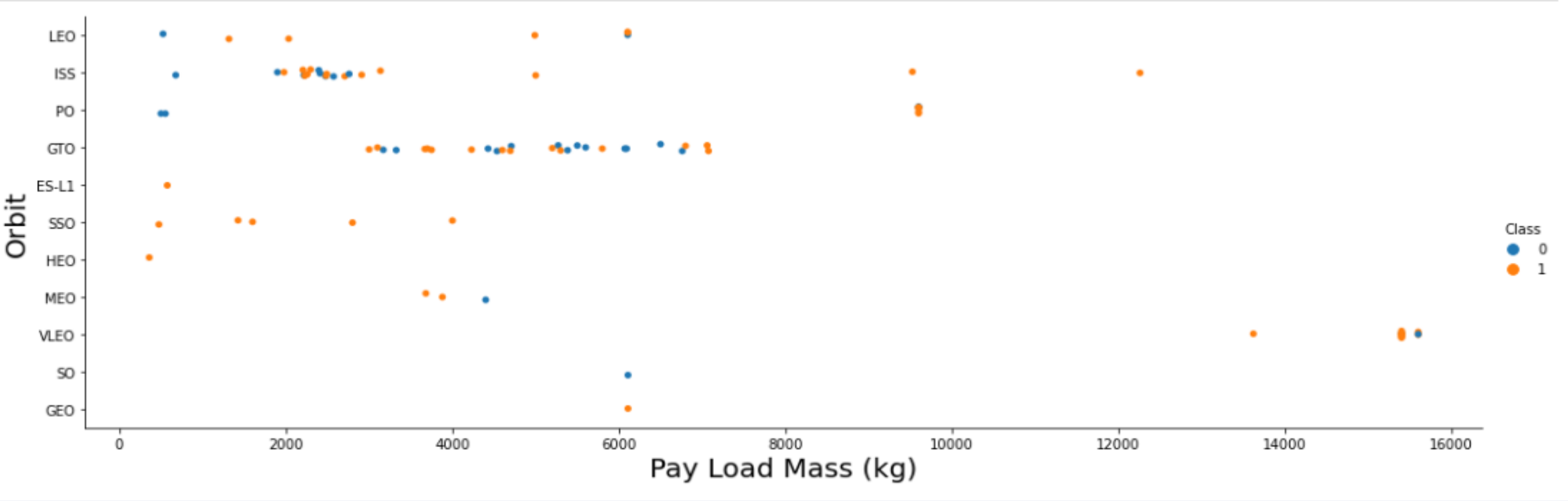
Flight Number vs. Orbit Type

In the first place, we can see how the success rate improves with the number of launches, something logical if we think about the experience gained. At first glance, VLEO is the one that presents the best results.



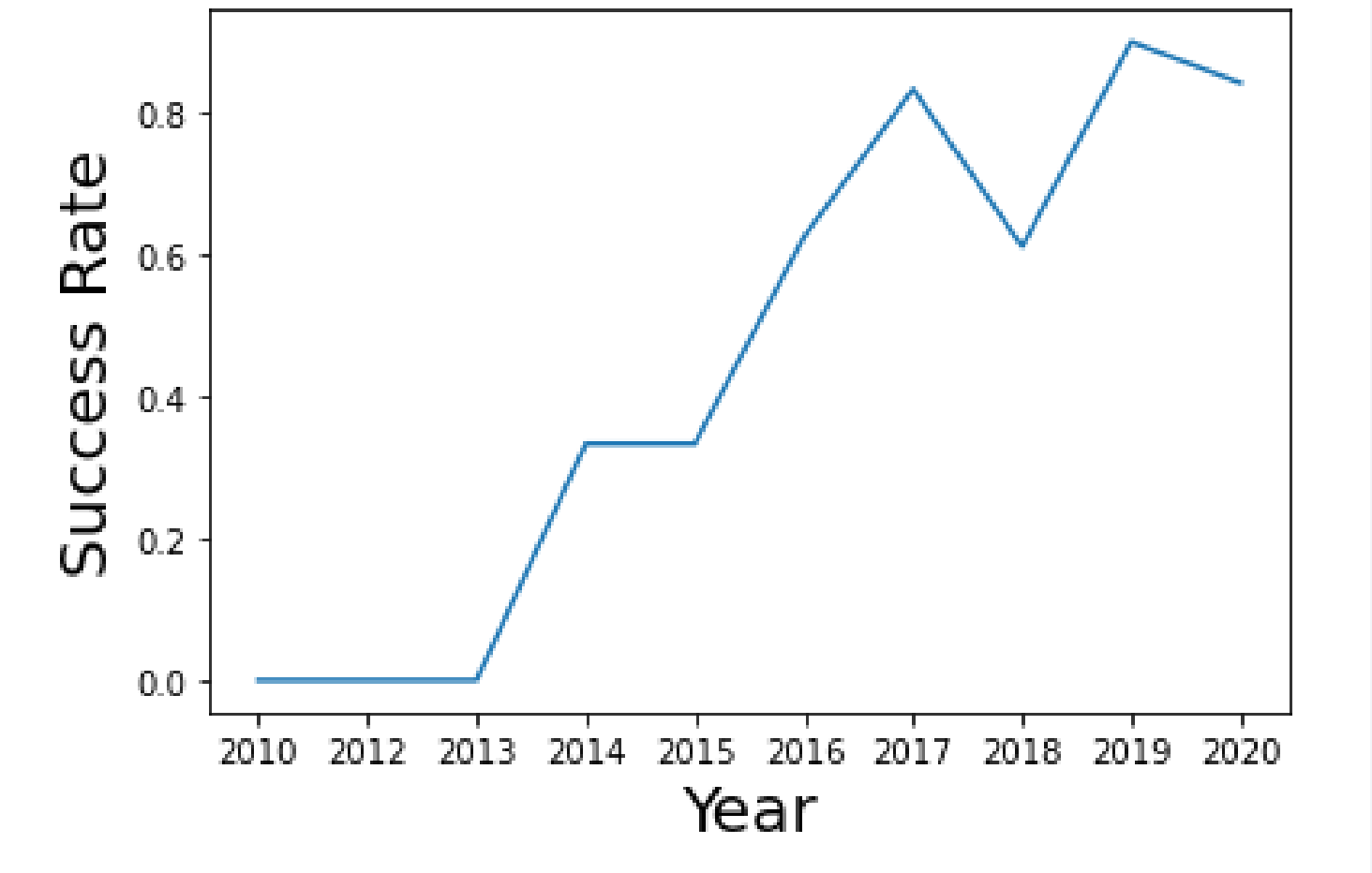
Payload vs. Orbit Type

The most tested orbits, GTO and ISS, do not show a clear relationship between success and payload mass. The rest of the orbits in general have a higher proportion of successes, but with very few samples.



Launch Success Yearly Trend

A clear positive trend can be seen, with large jumps in the early years and plateauing near the present. Logically the passage of time, technological improvement and experience take effect.



All Launch Site Names

In this case, the use of DISTINCT shows us a list of the possible values of launch_site.

```
%sql SELECT Distinct LAUNCH_SITE FROM SpaceXDB
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

In this case to see the first 5 records we use the LIMIT function and with the LIKE clause we filter the CCA substring

```
%sql SELECT * FROM SpaceXDB WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Using the SUM function we can calculate the sum of the payload_mass where the customer is NASA, obtaining 45596 kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SpaceXDB WHERE CUSTOMER='NASA (CRS)'
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

The AVG function shows us the average of the payload masses where booster_version contains F9 v1.1. The result is 2,928.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SpaceXDB WHERE BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

1
2928

First Successful Ground Landing Date

The date when the first successful landing outcome in ground pad was achieved fue el 22/12/2015.

```
%sql SELECT min(DATE) FROM SpaceXDB WHERE LANDING__OUTCOME='Success (ground pad)'
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

WHERE and AND are used to filter the booster version where the landing was successful with a payload mass between 4,000 and 6,000 kg.

```
%sql SELECT BOOSTER_VERSION FROM SpaceXDB WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

The list of successful and Failure Mission outcomes is 101.

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM SpaceXDB WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

1
101

2015 Launch Records

There are 2 failed landing_outcomes in drone ship in year 2015

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, LANDING__OUTCOME AS LANDING__OUTCOME, BOOSTER_VERSION AS BOOSTER_VERSION, LAU
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Between the date 2010-06-04 and 2017-03-20

```
%sql SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SpaceXDB WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND LANDING__OUTCOME LIKE '%Succe
```

```
 * ibm_db_sa://nnk89499:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1


Section 3
Launch Sites
Proximities Analysis

Folium Map - First glance

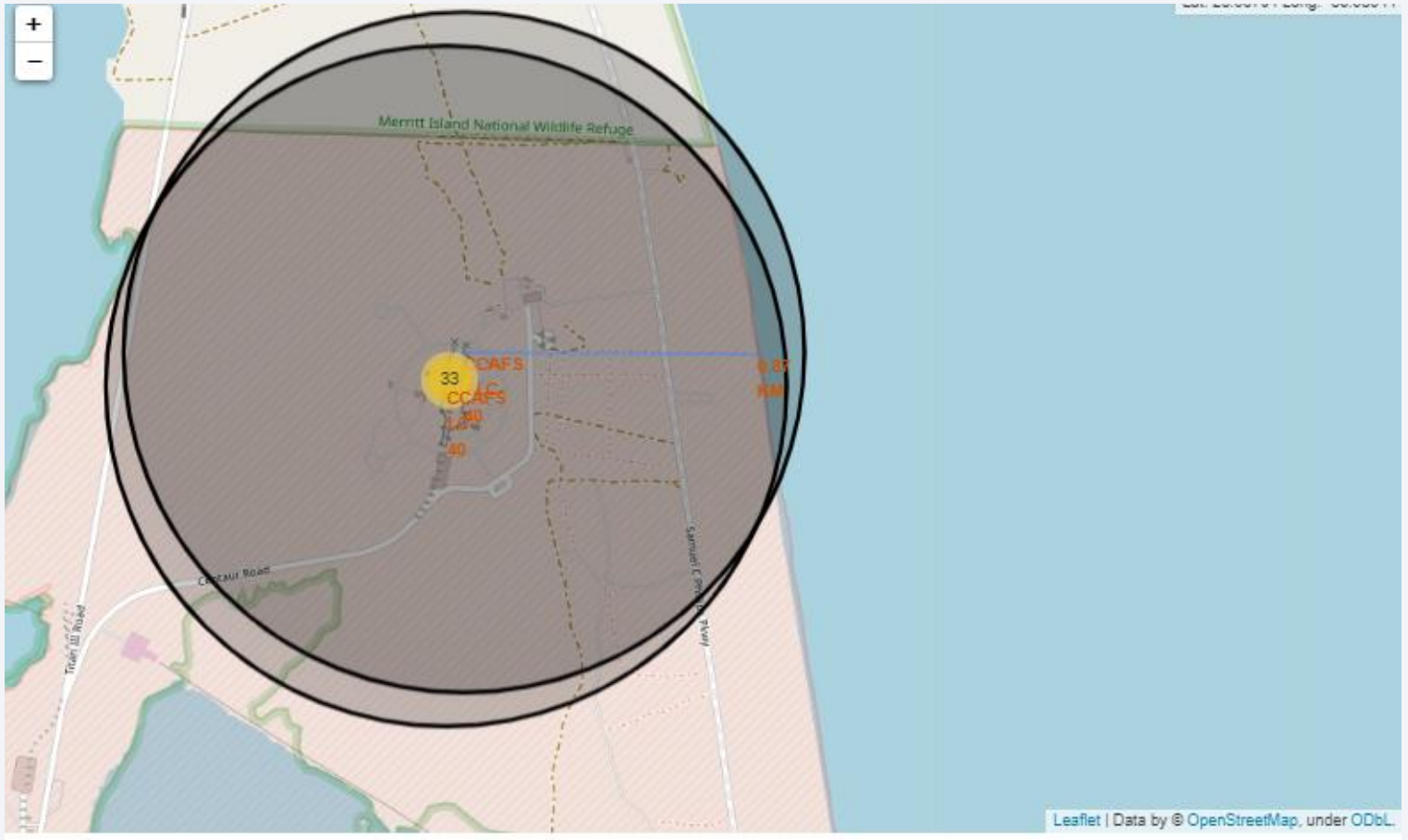
You can see two circles on the map, one on the Pacific coast and one on the Atlantic. For safety, launches near the coast are chosen.

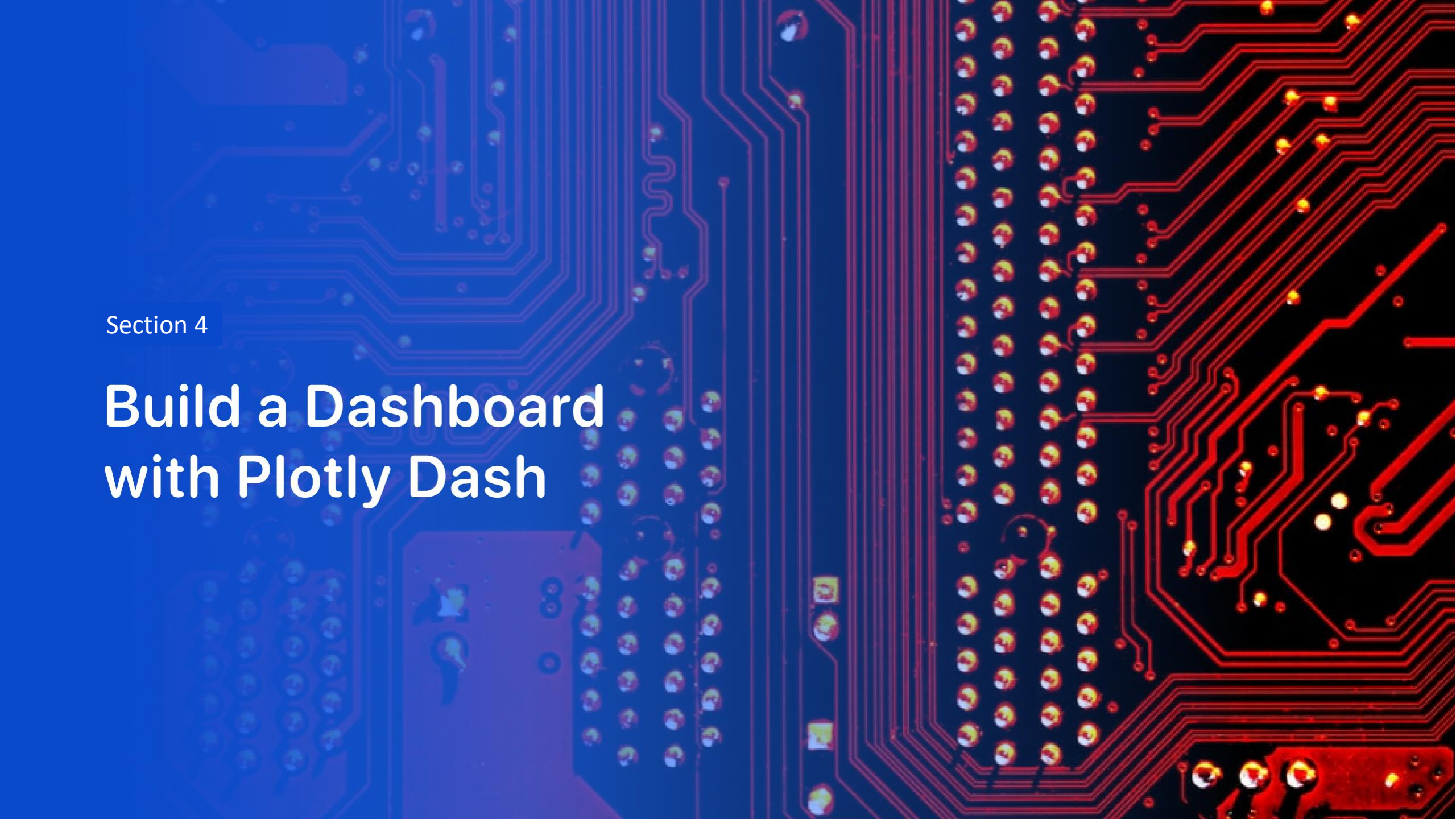
Folium Map – CCAFS LC-40 success or failure

As an example, the launches of CCAFS LC-40 are shown, where the red color means that the launch has failed and the green color is a success. We see the high failure rate in the first graph. Also the CCAFS SLC-40 shows better results.

Folium Map – CCAFS LC-40 distance

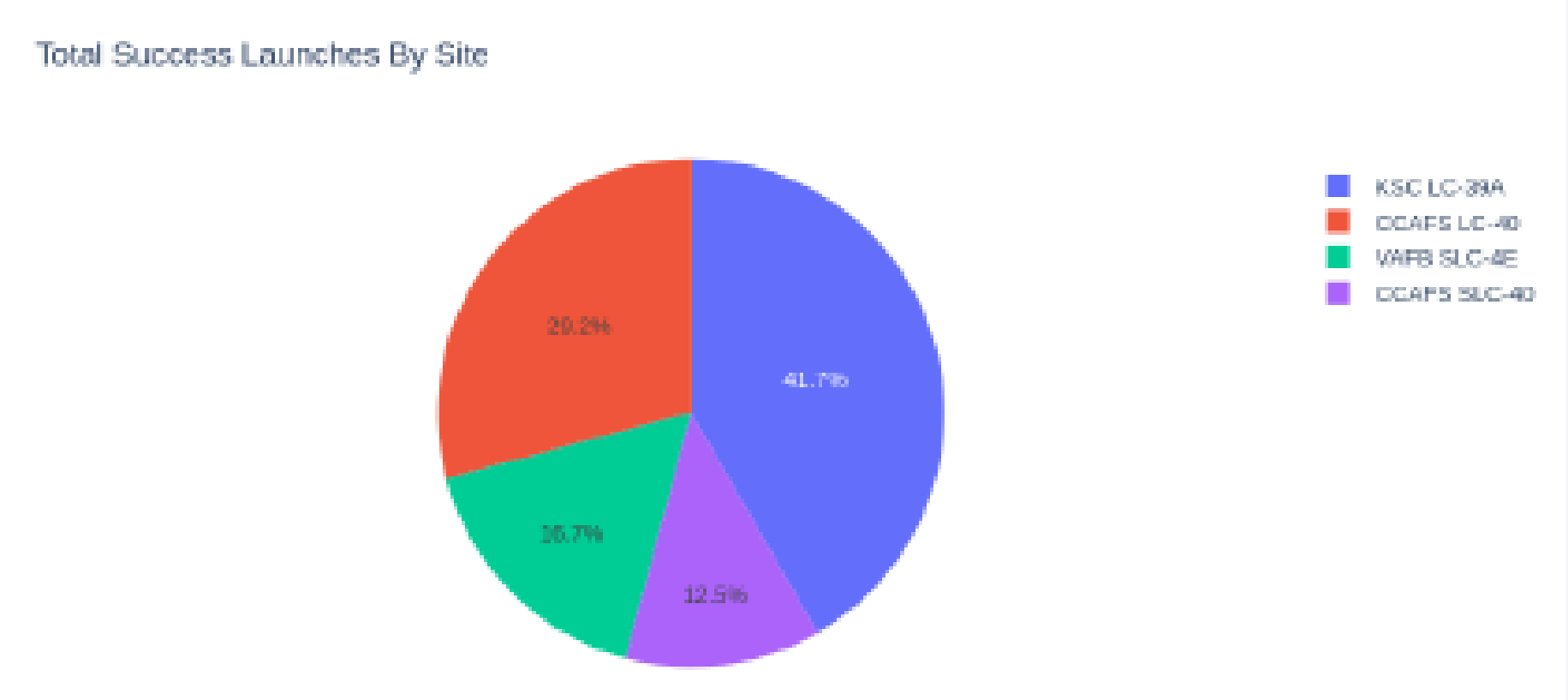
As an example, the distance between CCAFS SLC 40 and the railway is shown. the launch area is relatively close, which can lead to accidents and higher costs.




Section 4
Build a Dashboard with Plotly Dash

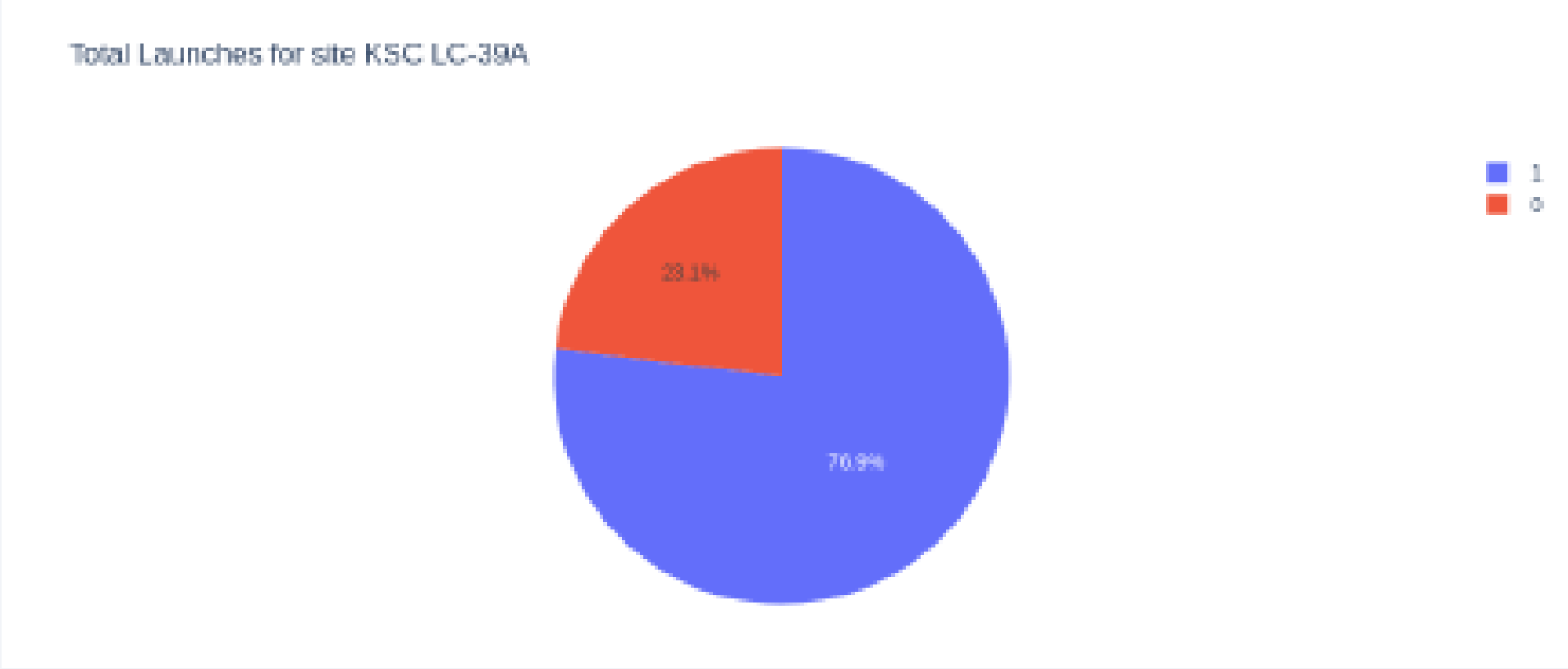
Dashboard – Success by site

Here we can see how successful releases are distributed. 41.7% KSC LC-39A. This does not mean that it is the area with the best results, because in the graph we cannot see the total number of attempts per site.



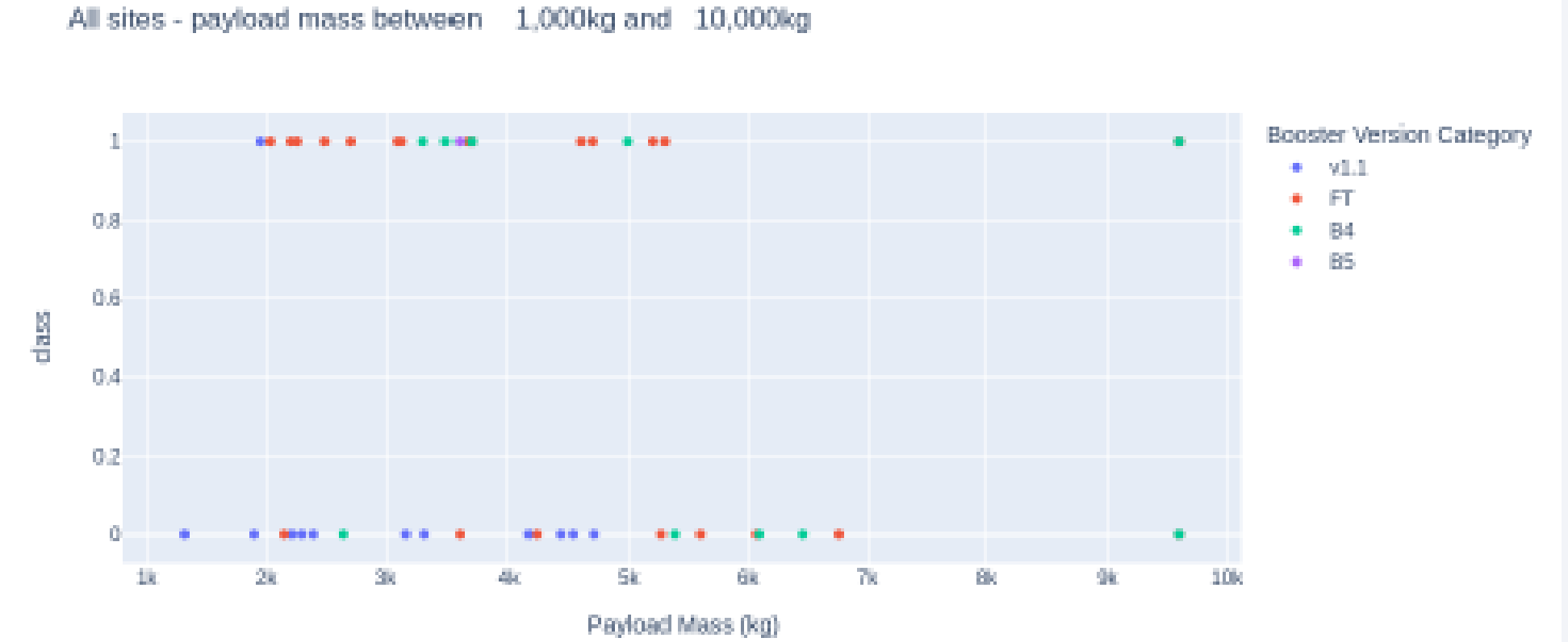
Dashboard KSC LC-39A success or failure

76.9% of KSC LC-39A launches are successful.



Dashboard

We can see how the number of successes begins to decline as the payload mass increases, no successes are seen with more than 5,500 kg. There is no predominant Booster version category, they all have successful and unsuccessful launches.

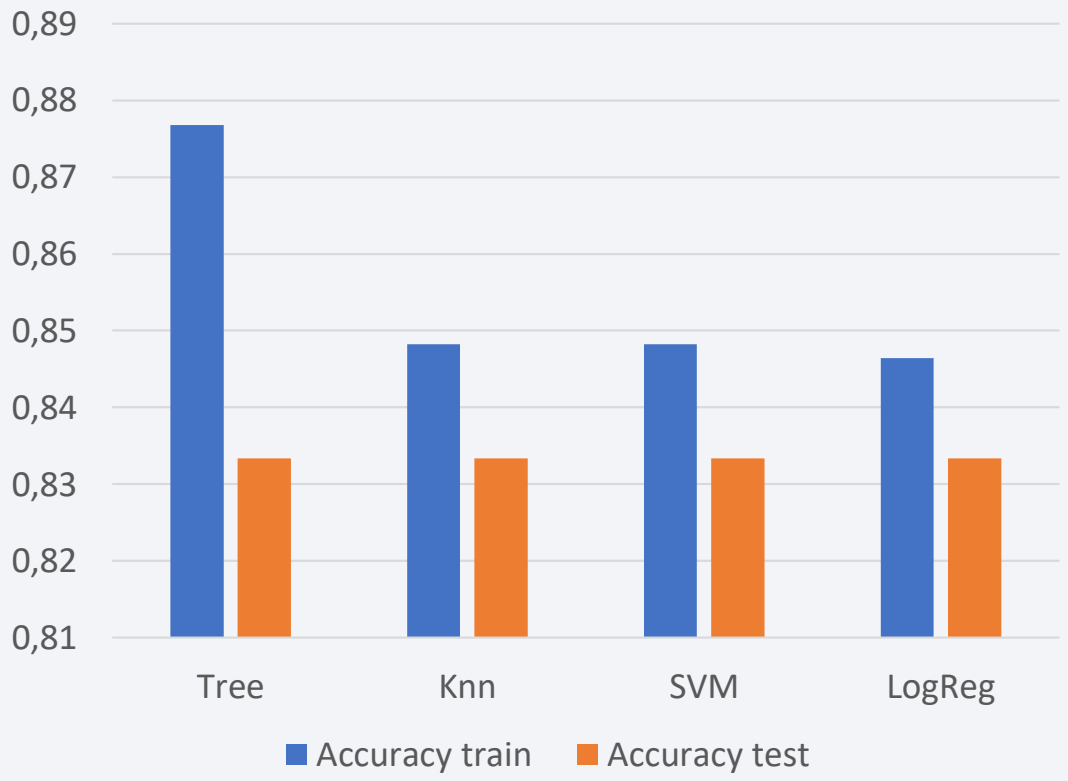


Section 5

Predictive Analysis (Classification)

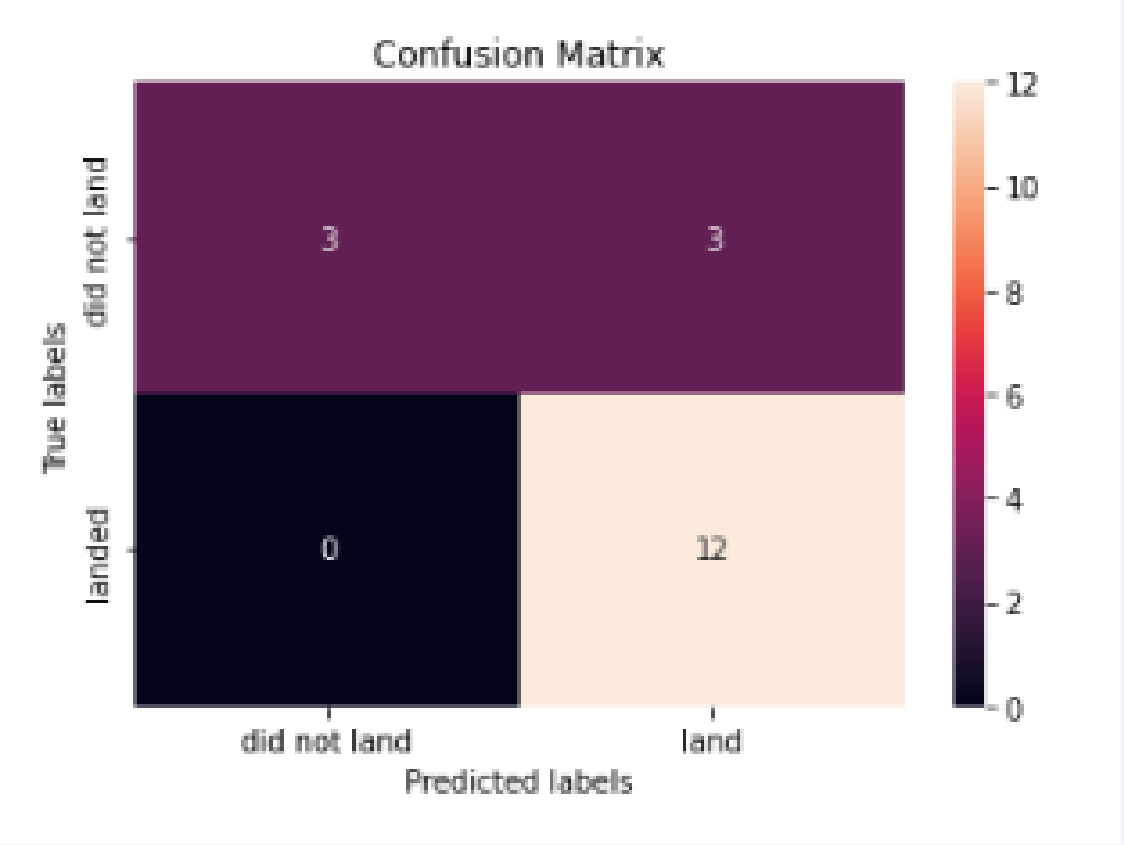
Classification Accuracy

We can see in the graph how all the predictive models gave the same test accuracy, but the decision tree algorithm has a better training accuracy.



Confusion Matrix - Decision Tree Classifier

Decision Tree Classifier's confusion matrix shows a large number of true positives and true negatives, so we can guess that it will be a good predictive model.



Conclusions

- In the first place, as is to be expected, the results have been improving over time, something that is explained by the technological improvement, experience and greater knowledge in general.
- It seems that the KSC LC 39A site is the one that presents the best results, without being able to obtain a cause for this.
- The results vary according to the orbits, a priori the less heavy payload generates better results.
- All predictive models gave equal test accuracy between models, but the decision tree algorithm has better training accuracy.

Thank you!