IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matías López

June - 9, 2022

Outline

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

Executive Summary

During this project we have performed the following tasks:

- Data collection
- Data wrangling
- Data visualization
- Machine learning predictive analysis

We found interesting relationships between many features of the data we used such as the payload mass, the orbit the rocket follows and the launch site. Finally, according to the predictive methods we developed, we conclude that all perform nearly the same.

Introduction

In this project we were interested in determining if SpaceX Falcon 9 first stage rocket will land successfully. This question has a significant importance to the company's savings, as the reuse of the first stage of the rocket would improve the costs of the future launches.

In order to achieve this goal we made some analytical and predictive analysis with the SpaceX data we collected from the company's API. The launch site, the mission outcome, the payload mass used, the booster version model and the date, are some of the aspects to take into consideration when answering this question.

Finding patterns and making predictions with this data may help us to reach our goal and help the company to make data-driven decisions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and Web scraping
- Perform data wrangling
  - Filtering data frames and dealing with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected from two different sources: *(i)* SpaceX API and *(ii)* Wikipedia.

SpaceX API

- Requesting and parsing the launch data using a GET request
- Data was saved as a Pandas data frame

Wikipedia

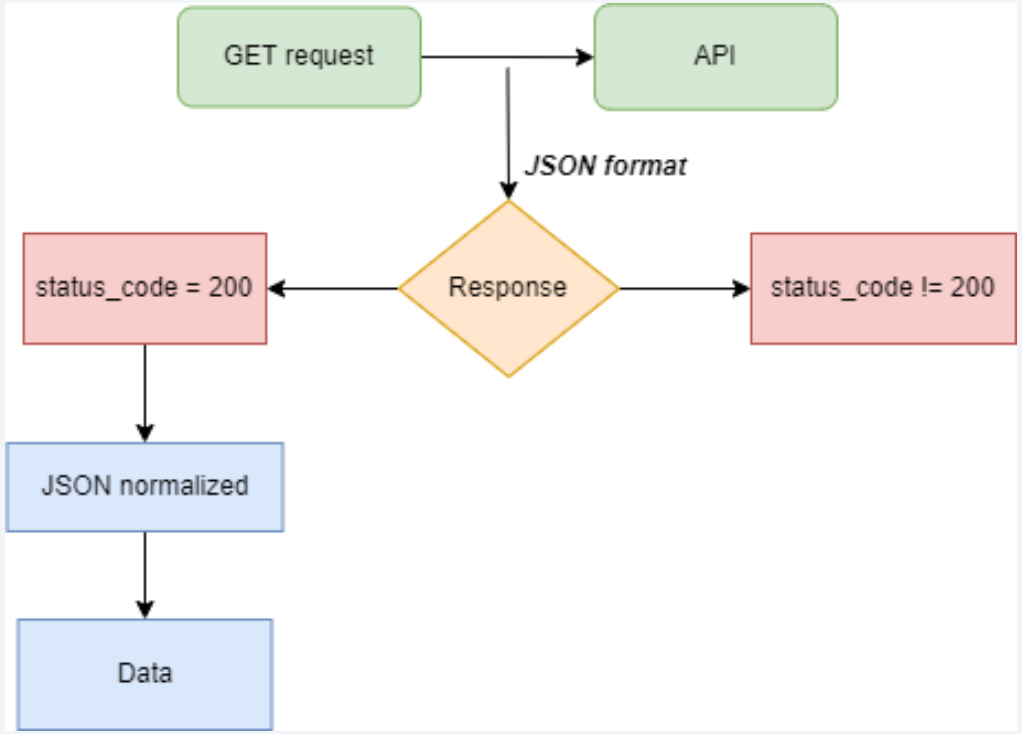
- Requesting the Falcon 9 launch wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

Data Collection – SpaceX API

Steps:

- Call the API with a GET request method
- Normalized the JSON content response if the status code is 200
- The data is saved in a Pandas data frame

- Click on this GitHub link to see the details

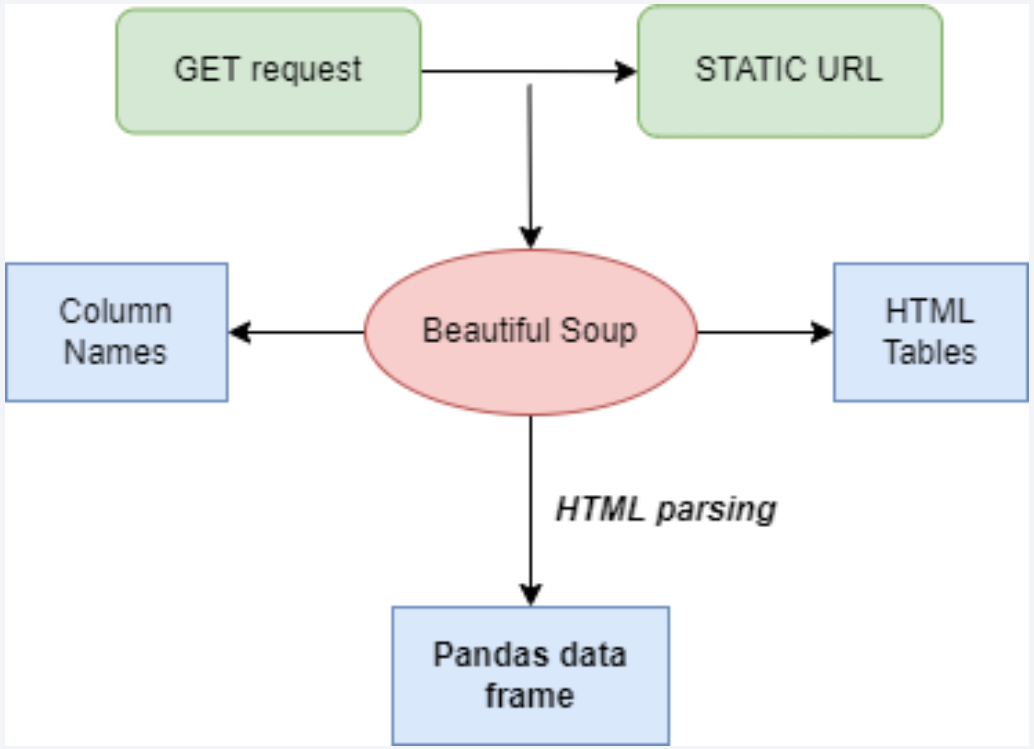


Data Collection - Scraping

Steps:

- Request data from a GET method to this Wikipedia page
- Creating a Beautiful Soup object to manage HTML content
- HTML tables were parsed and saved in a Pandas data frame

- Click on this GitHub link to see the details



Data Wrangling

- We create a subset of the data given by the API only to preserve relevant information of rocket, payloads, launchpad, cores, flight number and date.
- Rows with only one core and payload were considered.
- Date data was converted to datetime format.
- We got some information from these features by calling some Python functions we create.
- All this information was saved in a new Pandas data frame.
- Finally we keep only those records with Falcon 9 as launch, removing all Falcon 1 launches.
- We reset the FlightNumber column and remove those missing values by the mean of the correspondent column (we keep missing values of LandingPads).
- Click on this GitHub link to see the details

EDA with Data Visualization

Regarding data visualization, we plot many data attributes to study their relationships and how they can affect our final conclusion. Some of the plots we create are listed here:

- FlightNumber vs. PayloadMass
- FlightNumber vs LaunchSite
- Payload vs Launch Site
- Orit vs Succes Rate
- FlightNumber vs Orbit
- Payload vs Orbit
- Success Rate vs Year
- Click on this GitHub link to see the details

EDA with SQL

We retrieve some valuable information from the SpaceX dataset using SQL.

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload between 4000 and 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- The count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- Click on this GitHub link to see the details

Build an Interactive Map with Folium

During this project we create an interactive map using Folium with the following features:

- Markers:
  - For each launch site: 4 (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
  - For every success/fail launch per launch site
- Circles: 4 (one per launch site)
- Lines: 2 (from CCAFS LC-40 to the nearest coastline and railway)

These objects have helped us to visualize the location of the launches and their surroundings, and how important is to prepare the landing because of the potential danger nearby.

Click on this GitHub link to see the details

Build a Dashboard with Plotly Dash

To illustrate some of our findings interactively, we create a dashboard using Plotly where we plot 2 different graphs for: *(i)* all launch sites or *(ii)* any of the four specific launch sites.

- Pie chart of the success rate
- Scatter plot of the payload mass vs success rate

The user can select either all sites together or one launch site in particular. Also, there is a slider to handle the range of the payload mass between 0 and 10.000 Kg.

- Click on this GitHub link to see the details

Predictive Analysis (Classification)

- Previous to define the train and test samples we standardize our data.
- We use the 20% of the data for testing with a random state of 2. While the other 80% was used for training.
- The model we considered were: Logistic Regression, Decision Tree, Support Vector Machine and K-Nearest Neighbors.
- For each of these models we found the best parameter through a cross-validation grid search, using a CV = 10 for all the cases.
- We calculate the accuracy for the best parameter and the accuracy for the test model.
- Click on this GitHub link to see the details

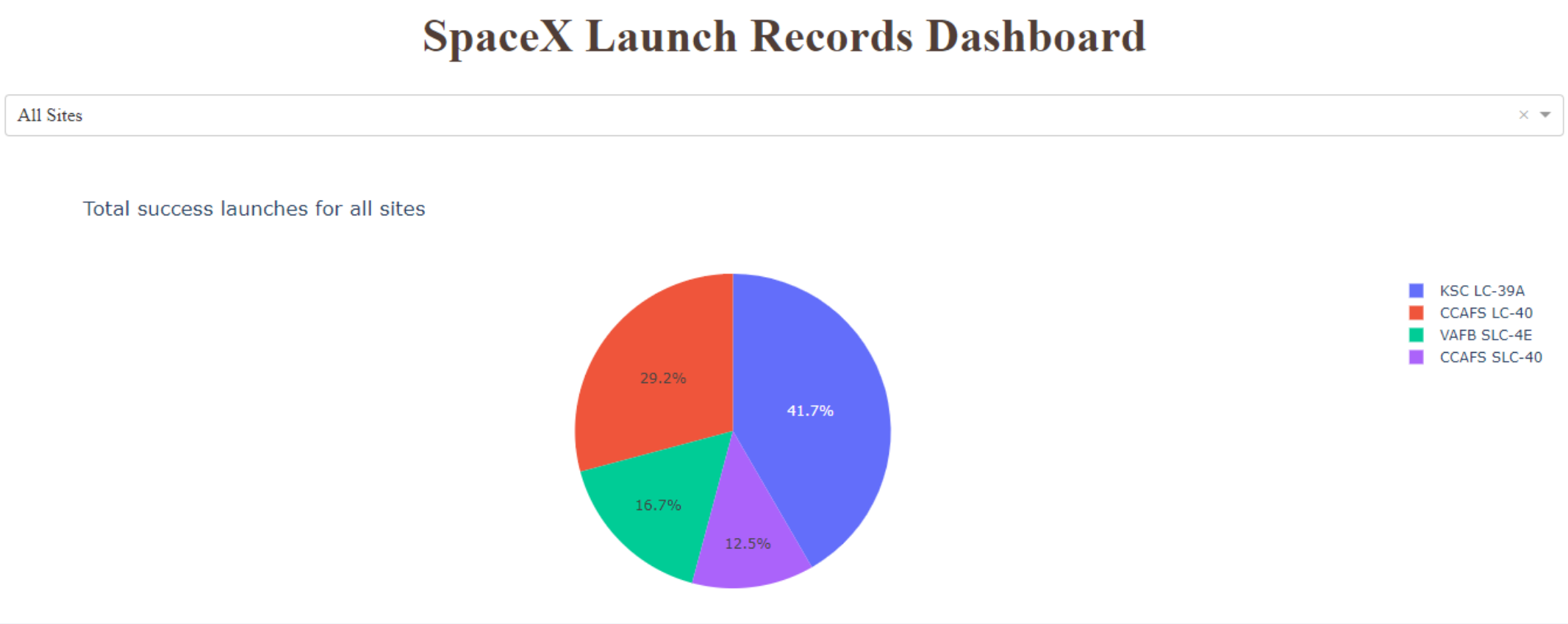
Results

Exploratory data analysis results

- From our data analysis exploration we found that the dataset was almost complete, with only 5 records without payload mass information.
- The maximum payload mass carried by boosters launched by NASA was 45.596 Kg.
- The average payload mass carried by booster version F9 v1.1 was 2928.4 Kg.
- The first successful landing outcome in ground was on 01/03/2013.
- There are only 23 successful landings in drone ships with payload masses between 4000 and 6000 Kg.
- In total, there were 98 successful landings and only 1 failure.
- 12 were the booster versions that have carried the maximum payload mass.
- During 2015, two drone ships landings failed.
- Between 2010 and 2016 there were 8 and 6 successful landing in drone ships and ground pad.

Results

Interactive analytics demo in screenshots



Results

Interactive analytics demo in screenshots

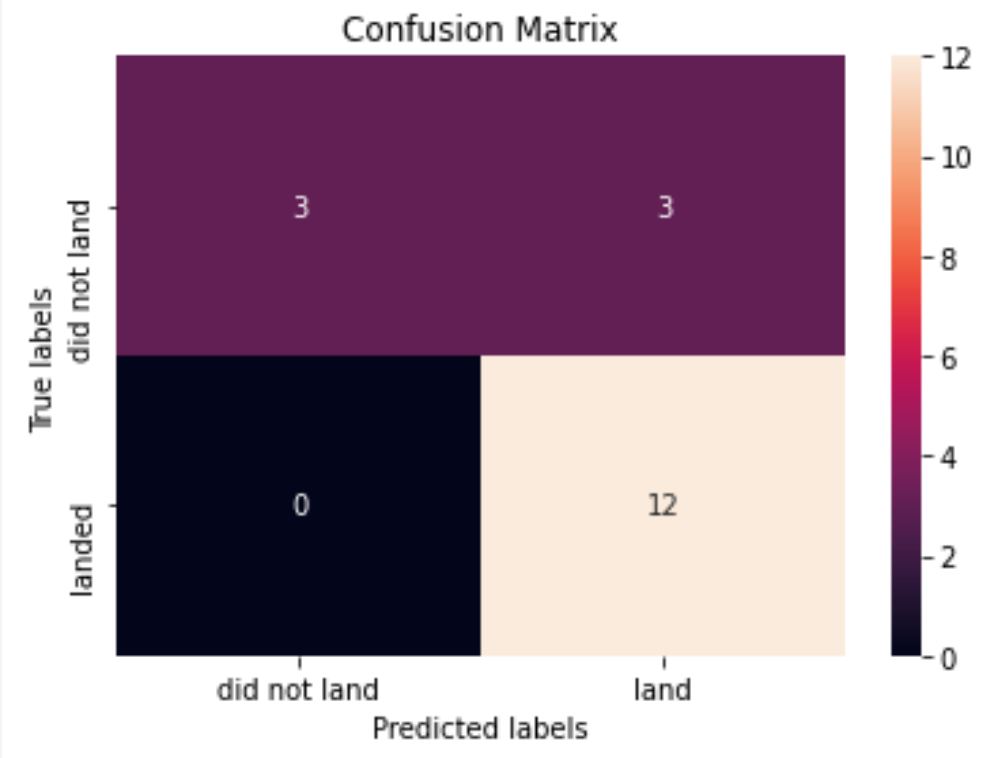
Results

Predictive analysis results

We found the following best parameters for each of the models mentioned before.

Logistic Regression:

- C = 0.01
- Penalty: l2
- Solver: lbfgs
- Accuracy with these parameters: 84.7%
- Accuracy on the test data: 83.3%



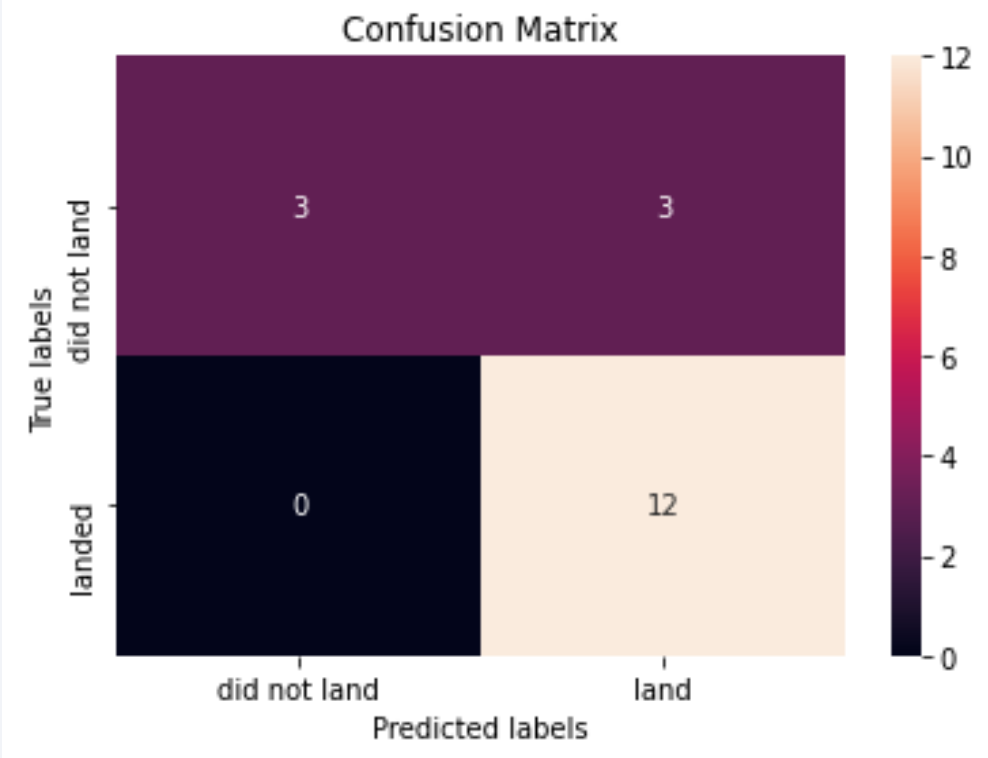
Results

Predictive analysis results

We found the following best parameters for each of the models mentioned before.

Suppport Vector Machine:

- C = 1.0
- Gamma: 0.0316
- Kernel: sigmoid
- Accuracy with these parameters: 84.7%
- Accuracy on the test data: 83.3%



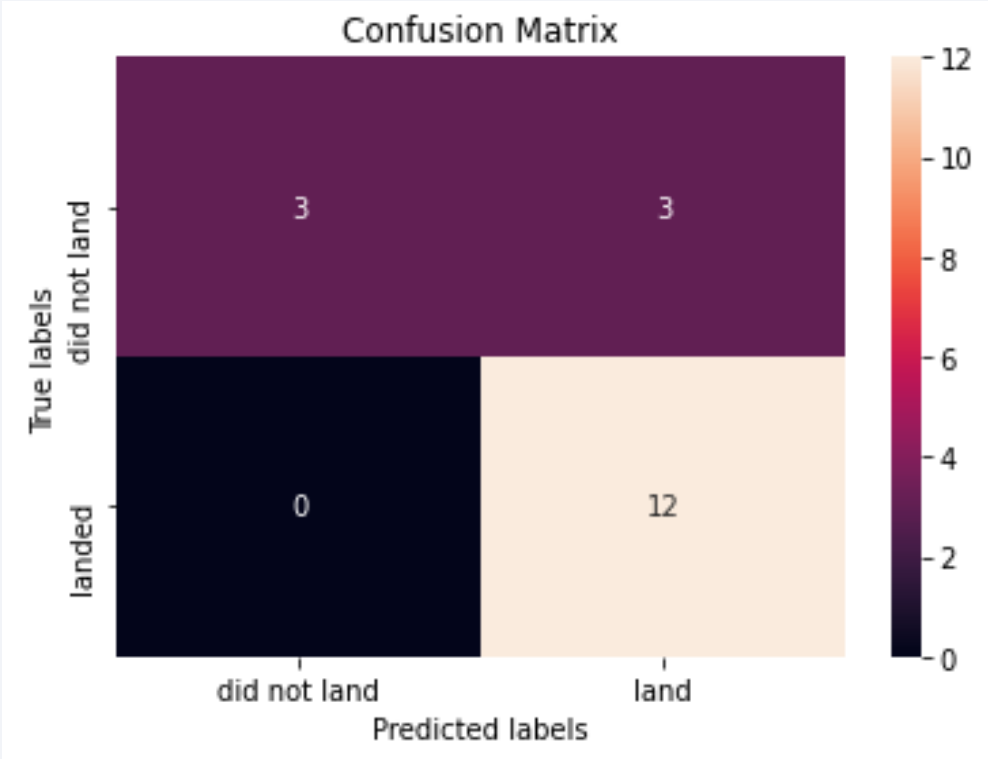
Results

Predictive analysis results

We found the following best parameters for each of the models mentioned before.

Decision Tree:

- Criterion: entropy
- Max depth: 2
- Max features: sqrt
- Min samples leaf: 1
- Min samples split: 2
- Splitter: best
- Accuracy with these parameters: 88.8%
- Accuracy on the test data: 83.3%



Results

Predictive analysis results

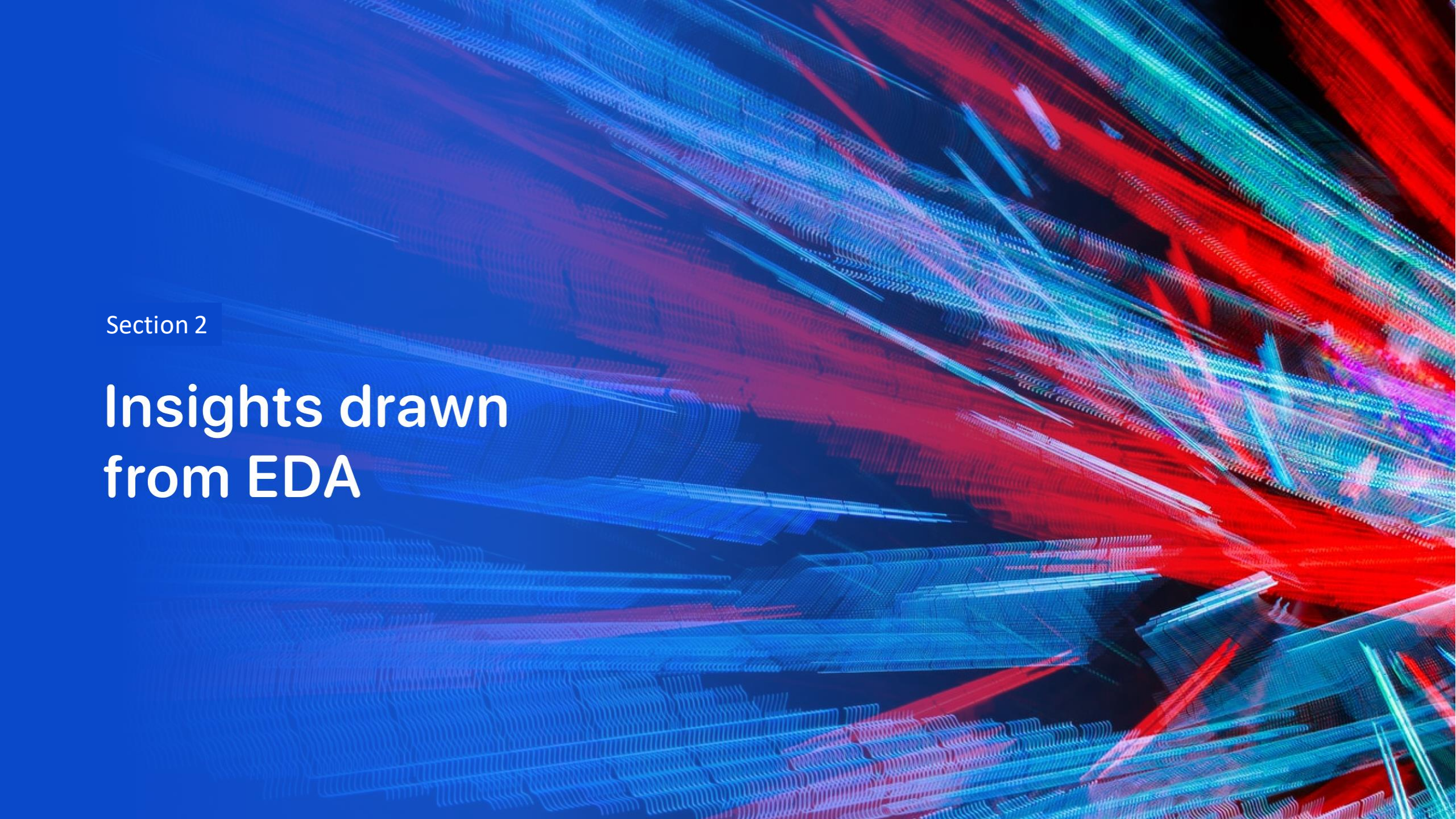
We found the following best parameters for each of the models mentioned before.

K-Nearest Neighbors:

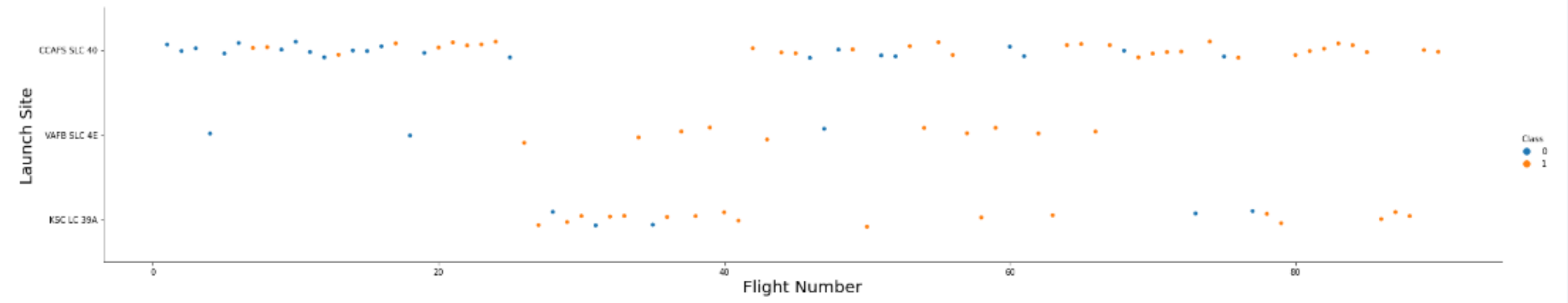
- Algorithm: auto
- N neighbors: p
- P: 1
- Accuracy with these parameters: 84.7%
- Accuracy on the test data: 83.3%

Section 2

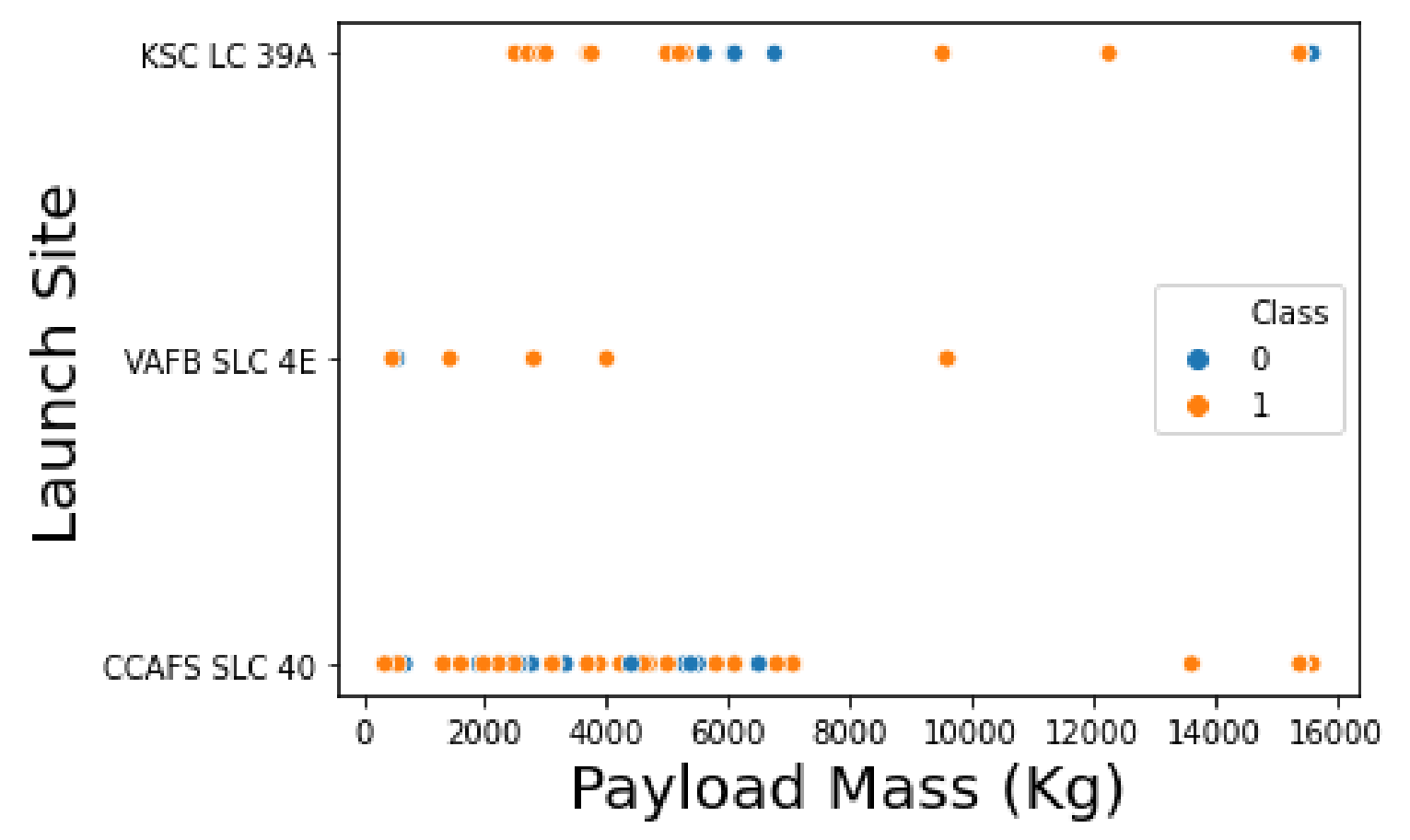
Insights drawn from EDA



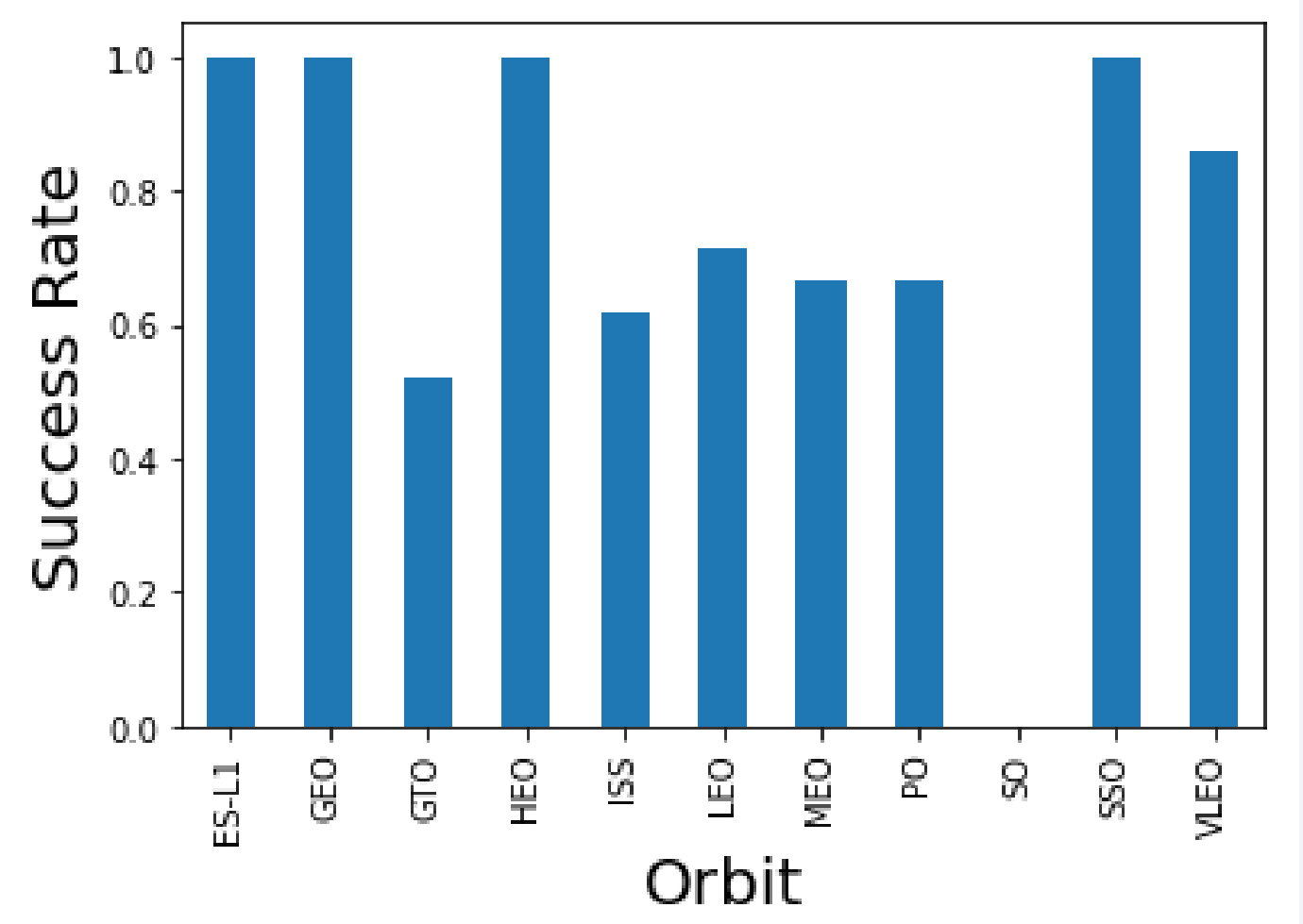
Flight Number vs. Launch Site



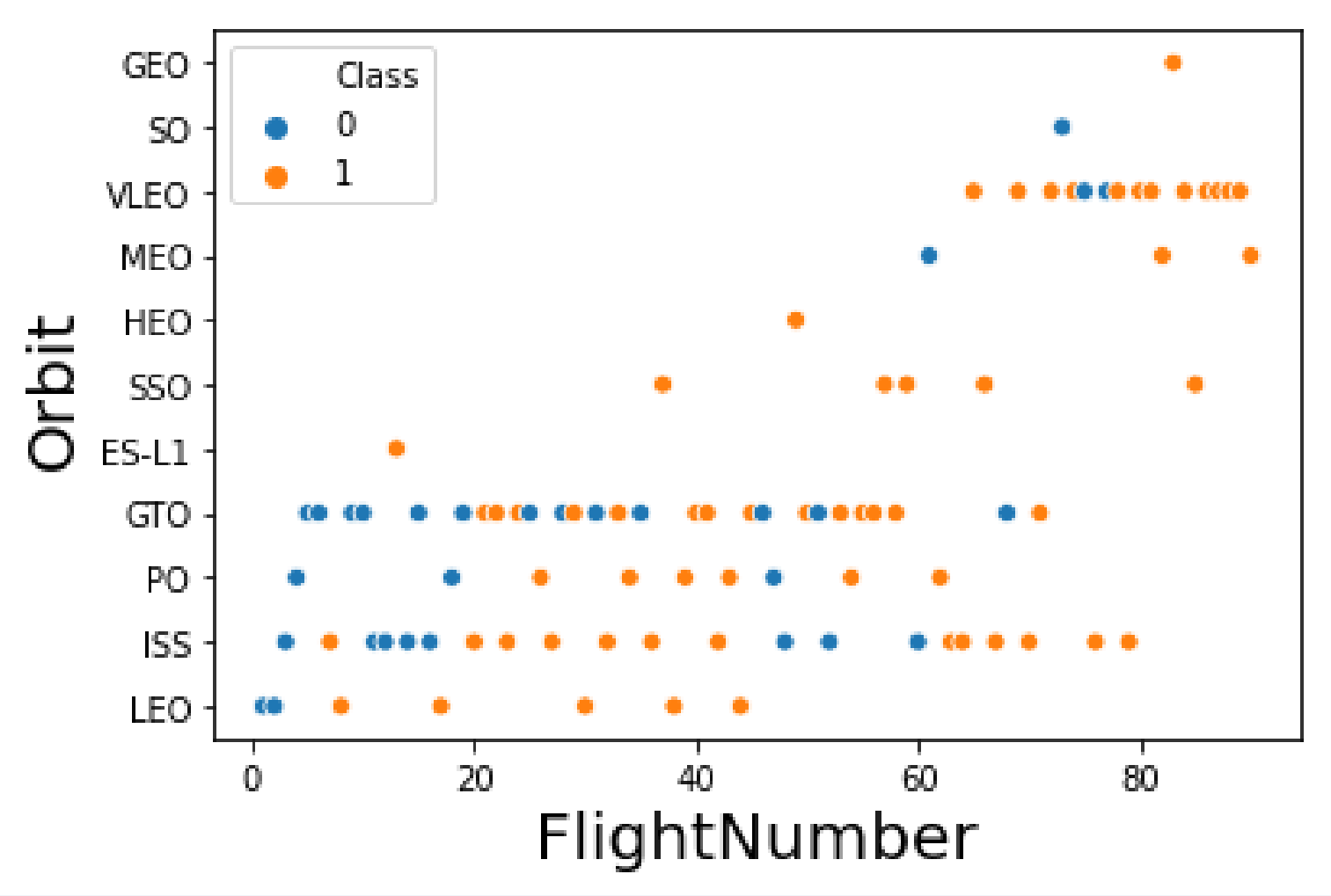
Payload vs. Launch Site



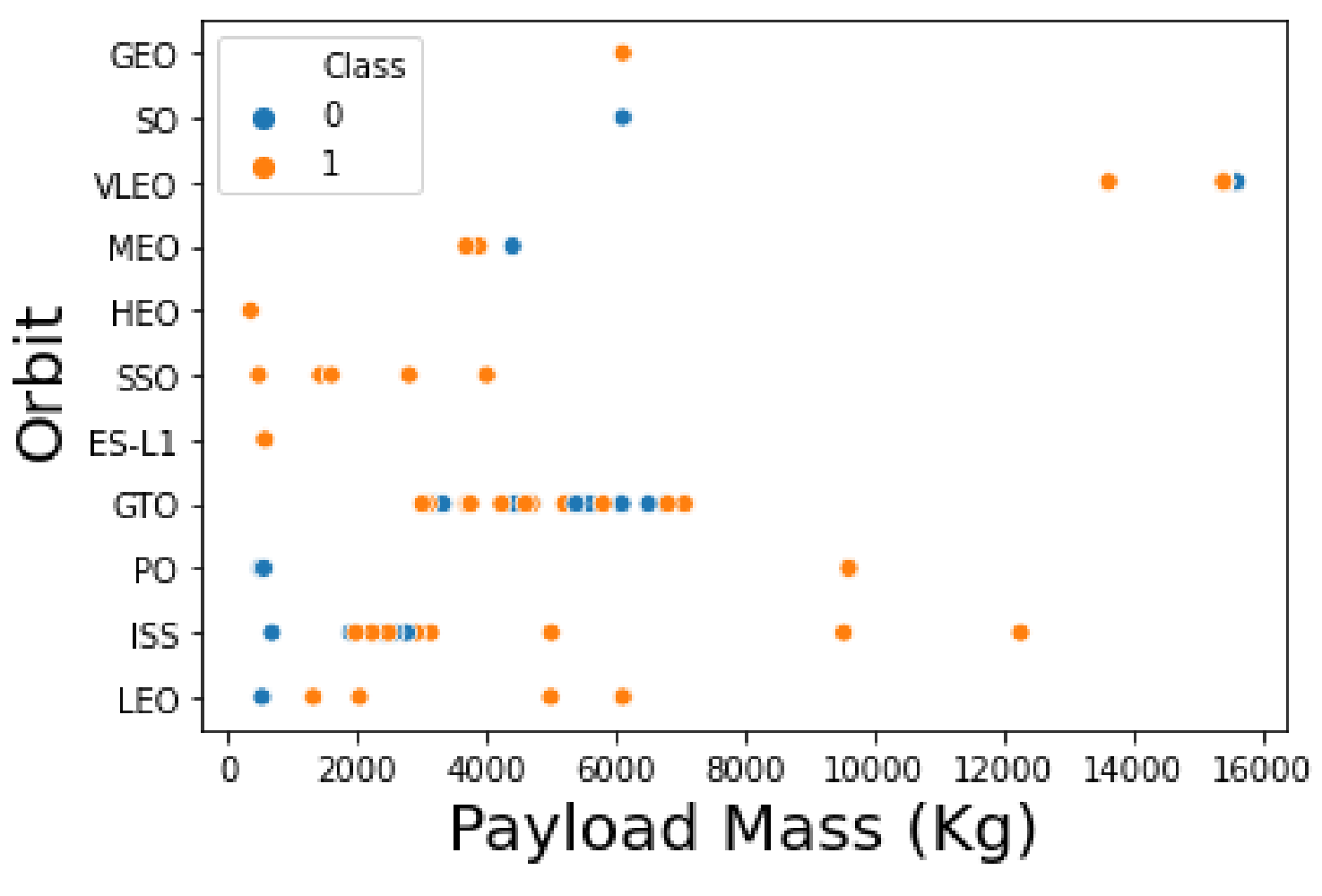
Success Rate vs. Orbit Type



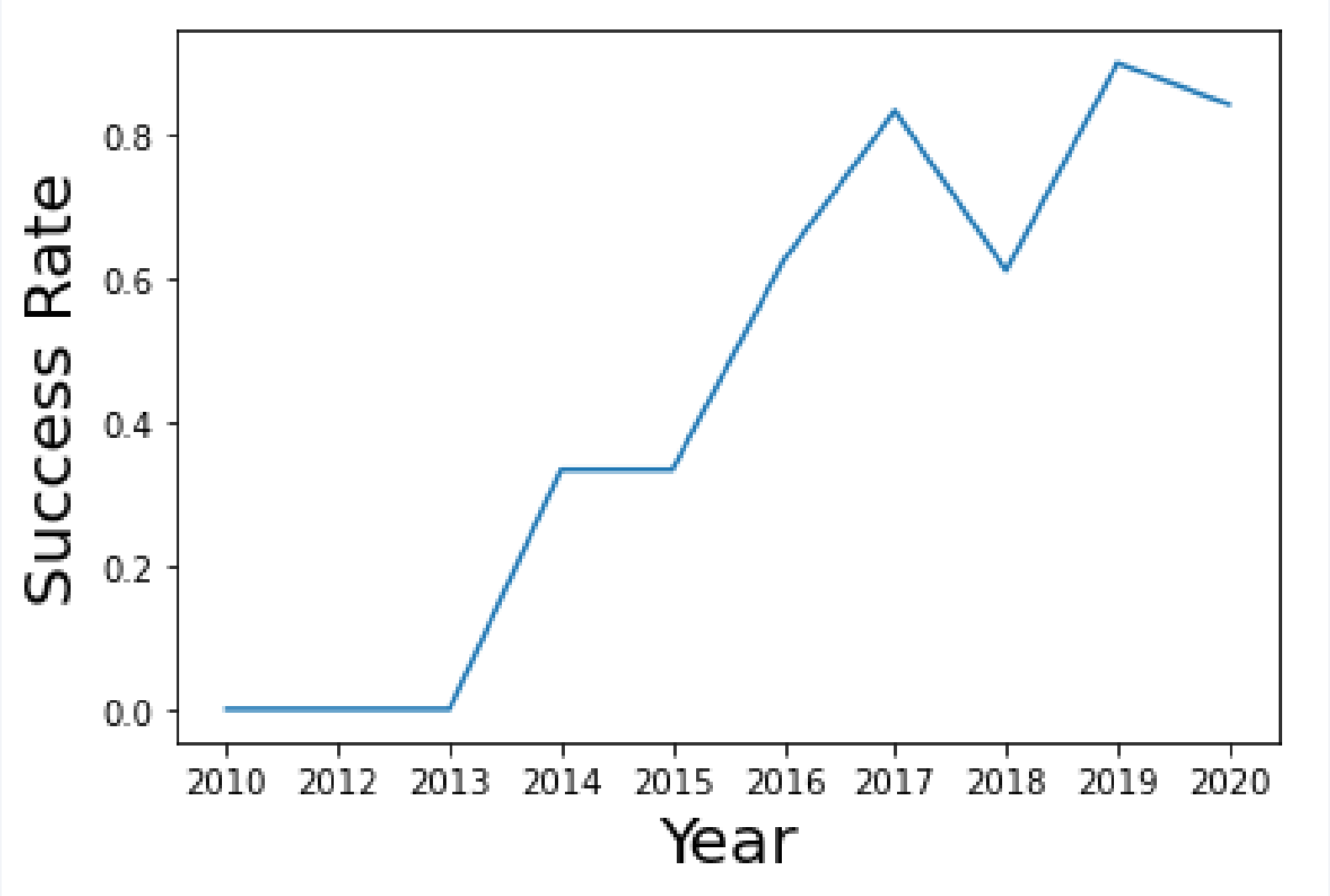
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- Find the names of the unique launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

- Query

```
%%sql

SELECT DISTINCT(LAUNCH_SITE)

FROM SPACEXTBL
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Query

```
%%sql

SELECT *

FROM SPACEXTBL

WHERE LAUNCH_SITE LIKE 'CCA%'

LIMIT 5;
```

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - 45.596 Kg

- Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - 2928.4 Kg

- Query

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - 01/03/2013

- Query

```
%%sql

SELECT MIN(DATE)

FROM SPACEXTBL

WHERE MISSION_OUTCOME = 'Success';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version	Mission_Outcome	PAYLOAD_MASS__KG_
F9 v1.1	Success	4535
F9 v1.1 B1011	Success	4428
F9 v1.1 B1014	Success	4159
F9 v1.1 B1016	Success	4707
F9 FT B1020	Success	5271
F9 FT B1022	Success	4696
F9 FT B1026	Success	4600
F9 FT B1030	Success	5600
F9 FT B1021.2	Success	5300
F9 FT B1032.1	Success	5300
F9 B4 B1040.1	Success	4990
F9 FT B1031.2	Success	5200
F9 B4 B1043.1	Success (payload status unclear)	5000
F9 FT B1032.2	Success	4230
F9 B4 B1040.2	Success	5384
F9 B5 B1046.2	Success	5800
F9 B5 B1047.2	Success	5300
F9 B5B1054	Success	4400
F9 B5 B1048.3	Success	4850
F9 B5 B1051.2	Success	4200
F9 B5B1060.1	Success	4311
F9 B5 B1058.2	Success	5500
F9 B5B1062.1	Success	4311

- Query

```
%%sql

SELECT BOOSTER_VERSION, MISSION_OUTCOME, PAYLOAD_MASS__KG_

FROM SPACEXTBL

WHERE MISSION_OUTCOME LIKE '%Success%' AND (4000 < PAYLOAD_MASS__KG_) AND (6000 >
PAYLOAD_MASS__KG_)
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - Successful: 98
  - Failure: 1

- Query

```
%%sql

SELECT COUNT(*) AS TOTAL_SUCCESS,

        (SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE
'%Failure%') AS TOTAL_FAILURES

FROM SPACEXTBL

WHERE MISSION_OUTCOME = 'Success';
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

- Query

```
%%sql

SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_

FROM SPACEXTBL

WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTBL)
```

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

MONTH	YEAR	Booster_Version	Launch_Site	Landing _Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Query

```
%%sql

SELECT SUBSTRING(DATE,4,2) AS MONTH, SUBSTRING(DATE,7,4) AS YEAR, BOOSTER_VERSION,
LAUNCH_SITE, "Landing _Outcome"

FROM SPACEXTBL

WHERE SUBSTRING(DATE,7,4) = '2015' AND "Landing _Outcome" LIKE '%Failure%';
```

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

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

Count	Landing _Outcome	Date
20	Success	07-08-2018
8	Success (drone ship)	08-04-2016
6	Success (ground pad)	18-07-2016

- Query

```
%%sql

SELECT COUNT("Landing _Outcome") as "Count", "Landing _Outcome", DATE

FROM SPACEXTBL

WHERE "Landing _Outcome" LIKE '%Success%' AND DATE BETWEEN '04-06-2010' AND '20-03-
2017'

GROUP BY "Landing _Outcome"

ORDER BY "Count" DESC
```

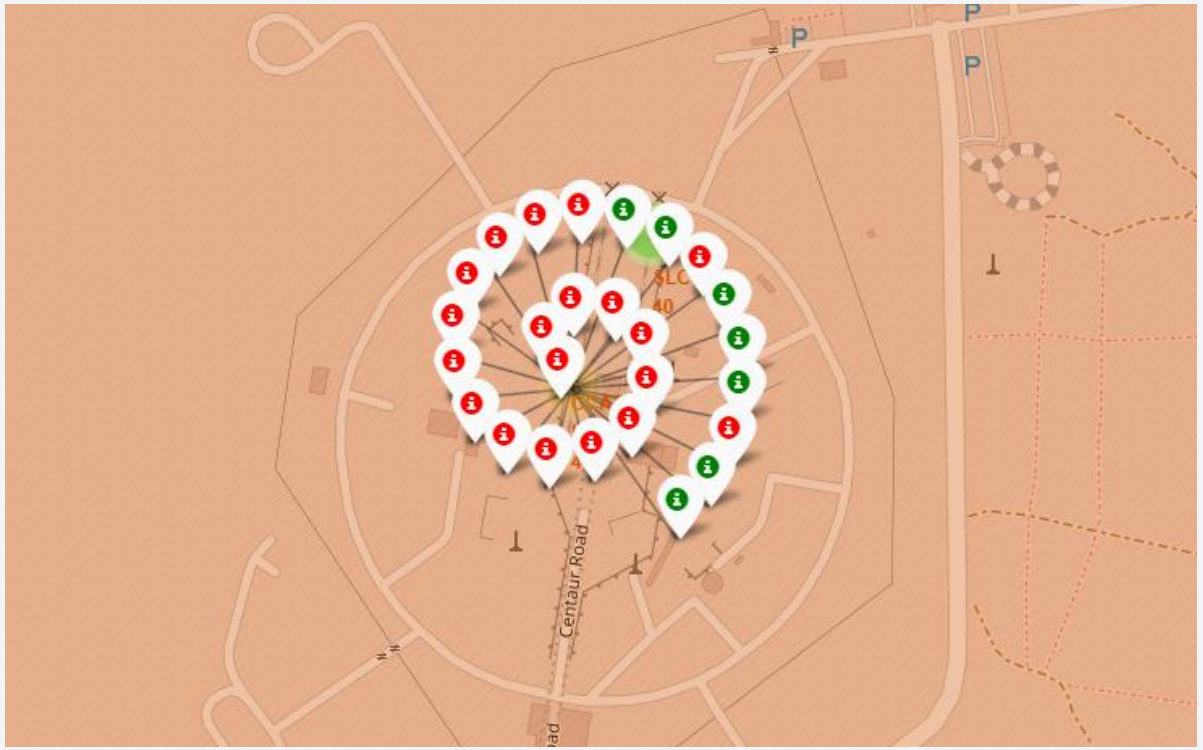

Section 3
Launch Sites
Proximities Analysis

Launch Sites Location Map

Here we can see the location of the 4 different launch sites, where 3 are located on the east coast and the other in the west coast of USA.

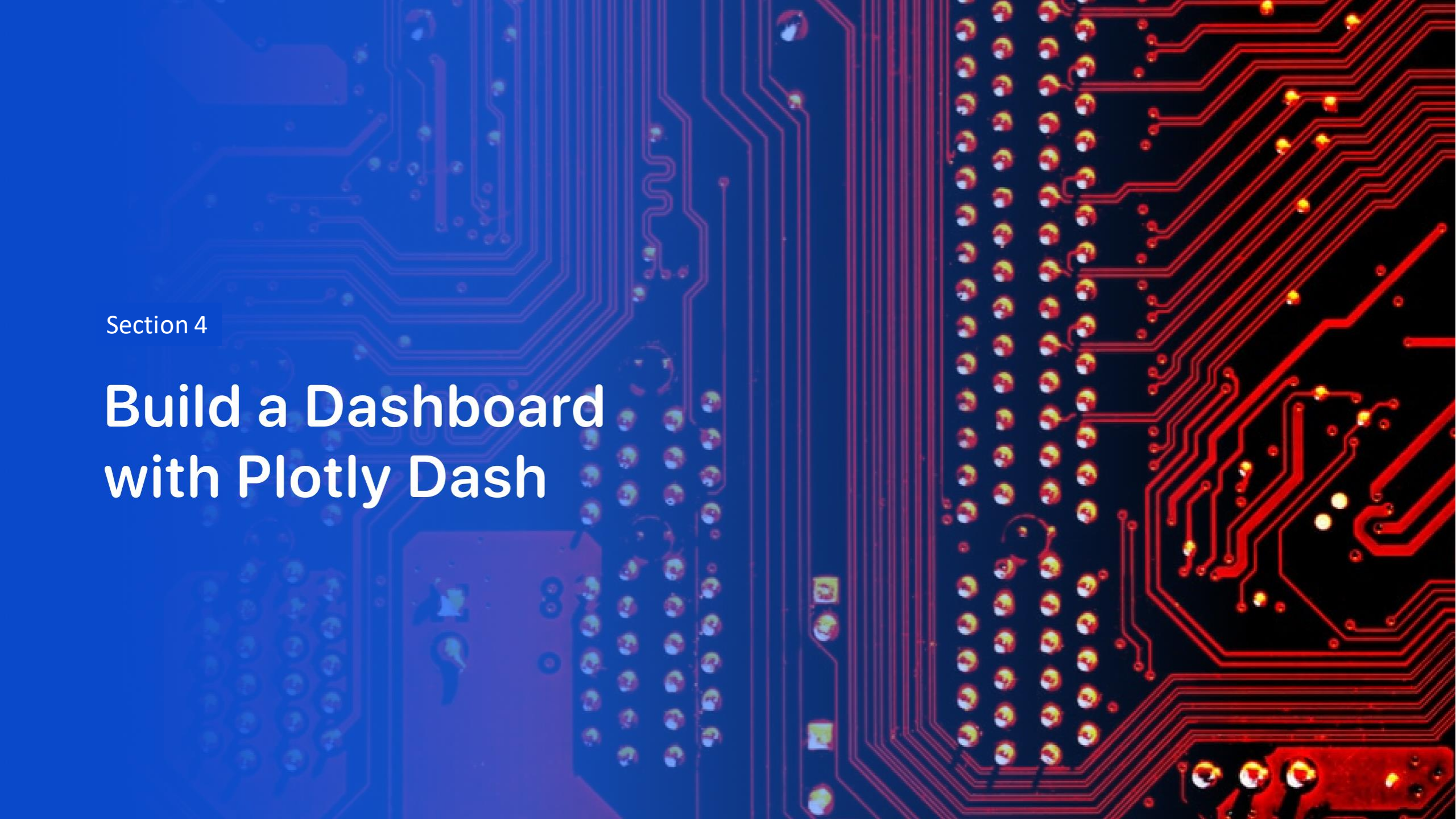
Launch Outcomes Location Map

In this map we visualize the location of the outcomes for an specific launch site (it is applied to the rest of the locations). The green labels correspond to the successful landings while the red ones to the failures.



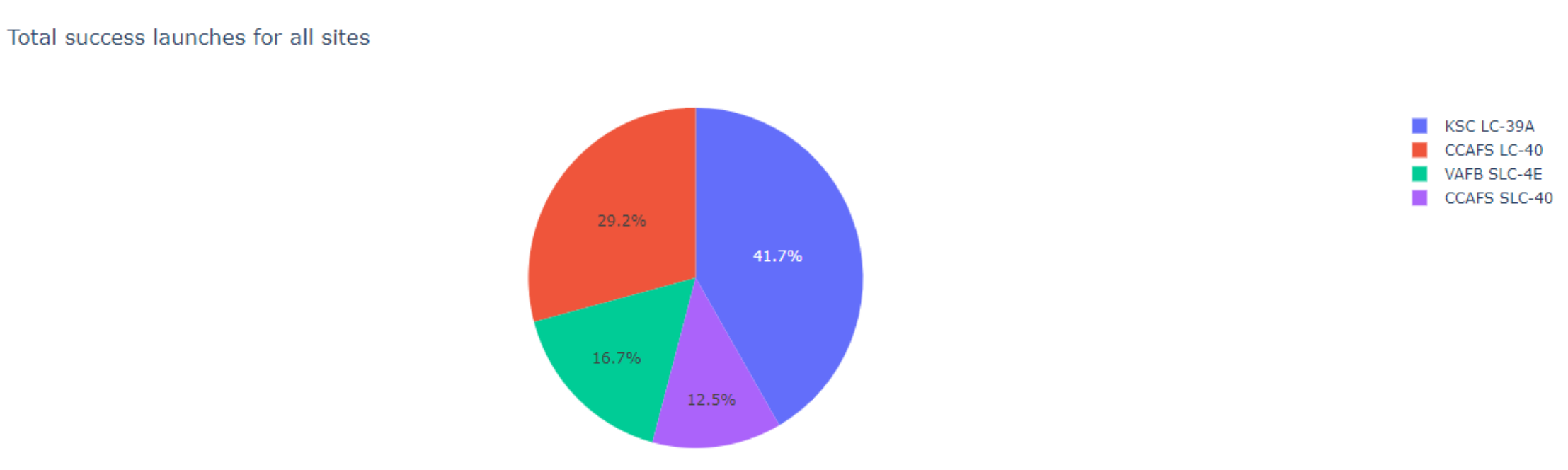
Launch Sites Distance to Coastline Map

This map shows the distance from one of the launch sites to the nearest coastline.


Section 4
Build a Dashboard with Plotly Dash

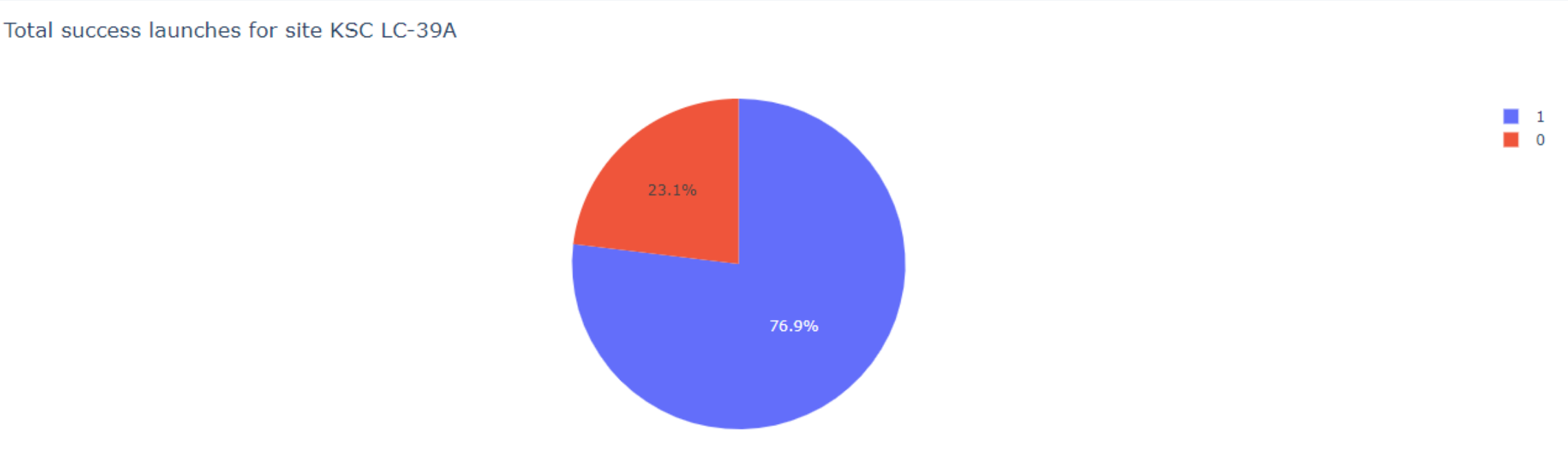
Success Rate for All Launch Sites

This pie chart shows the success launch rate for all the launch sites.



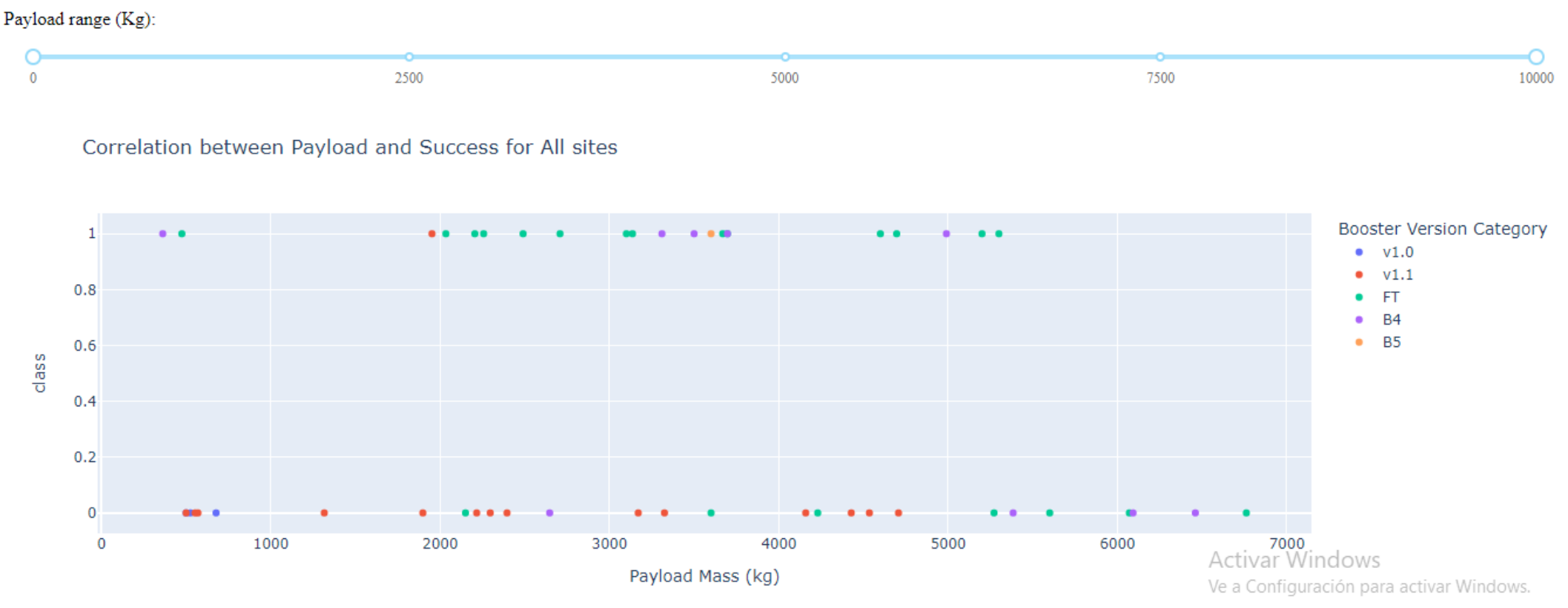
Highest Success Rate

The KSC LC-39 launch site has the highest success launch rate with 76.9%. The "1" label correspond to successful landings while the "0" to the failure ones.



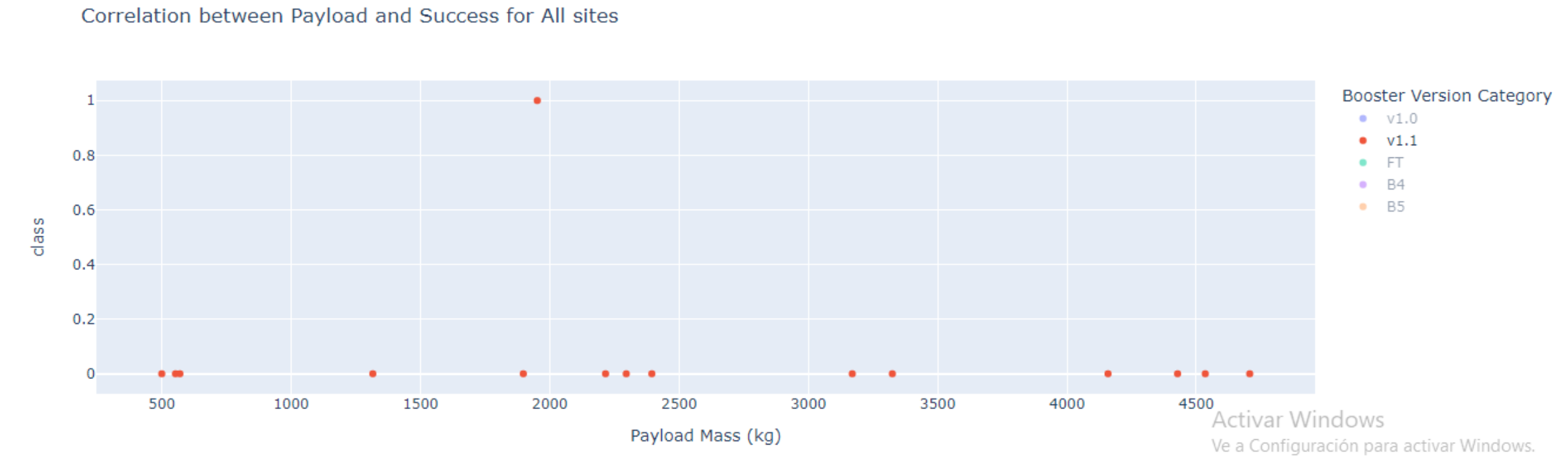
Payload Mass vs Launch Outcome for All Sites

This scatter plot shows the relationship between the payload mass and the launch outcome in the full range, from 0 Kg to 10.000 Kg for different booster versions.



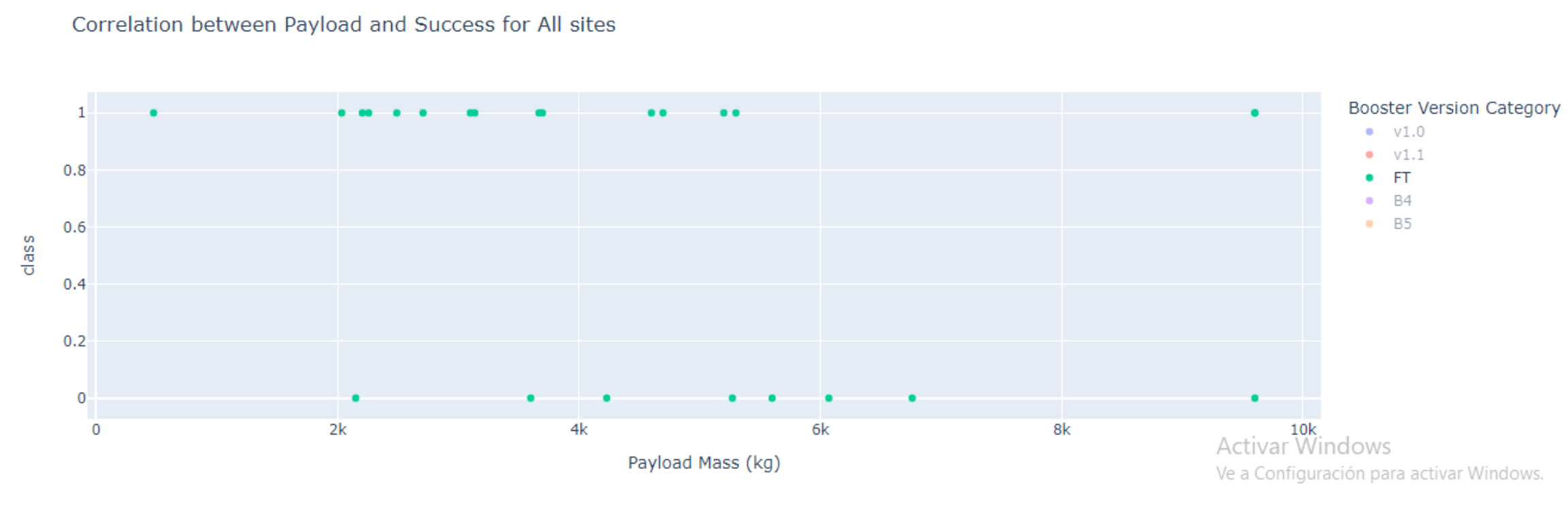
Payload Mass vs Launch Outcome for All Sites

If we filter by the booster version we can see that v1.1 has the worst success rate in the whole range of payload mass.



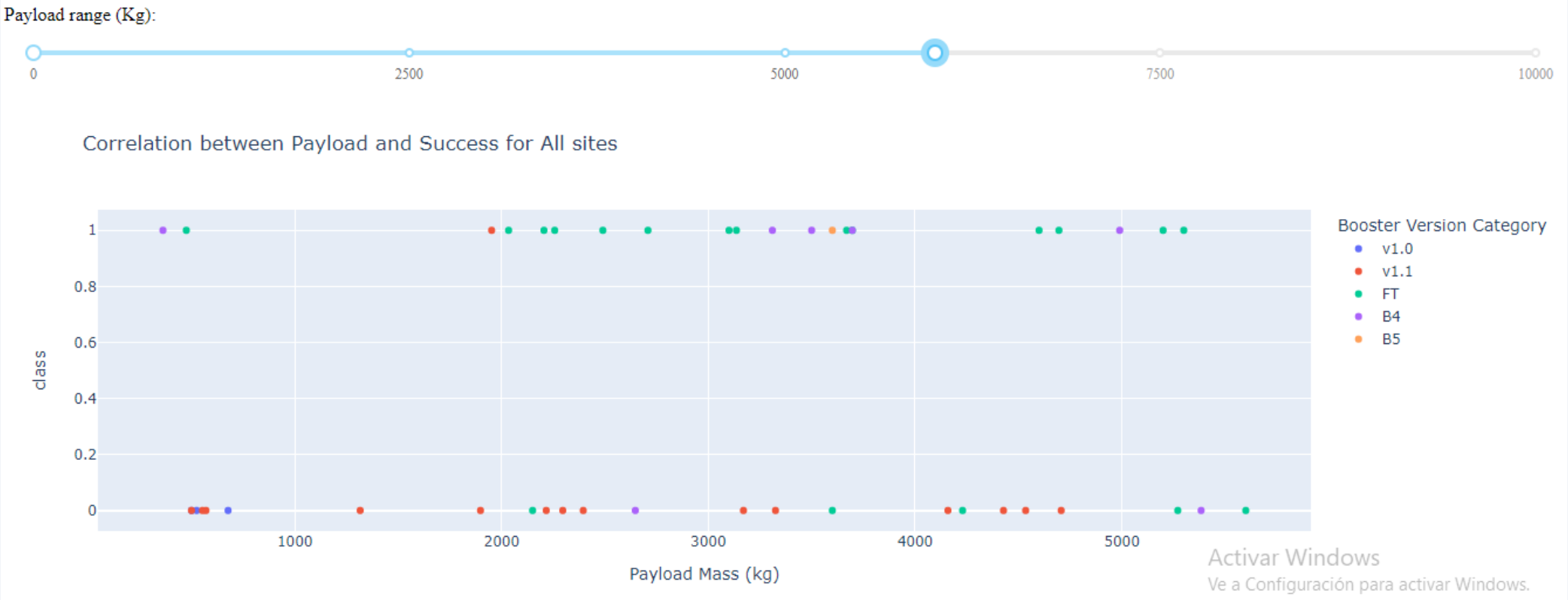
Payload Mass vs Launch Outcome for All Sites

On the other hand, the booster version FT has the highest success rate if we consider the whole range of payload mass.



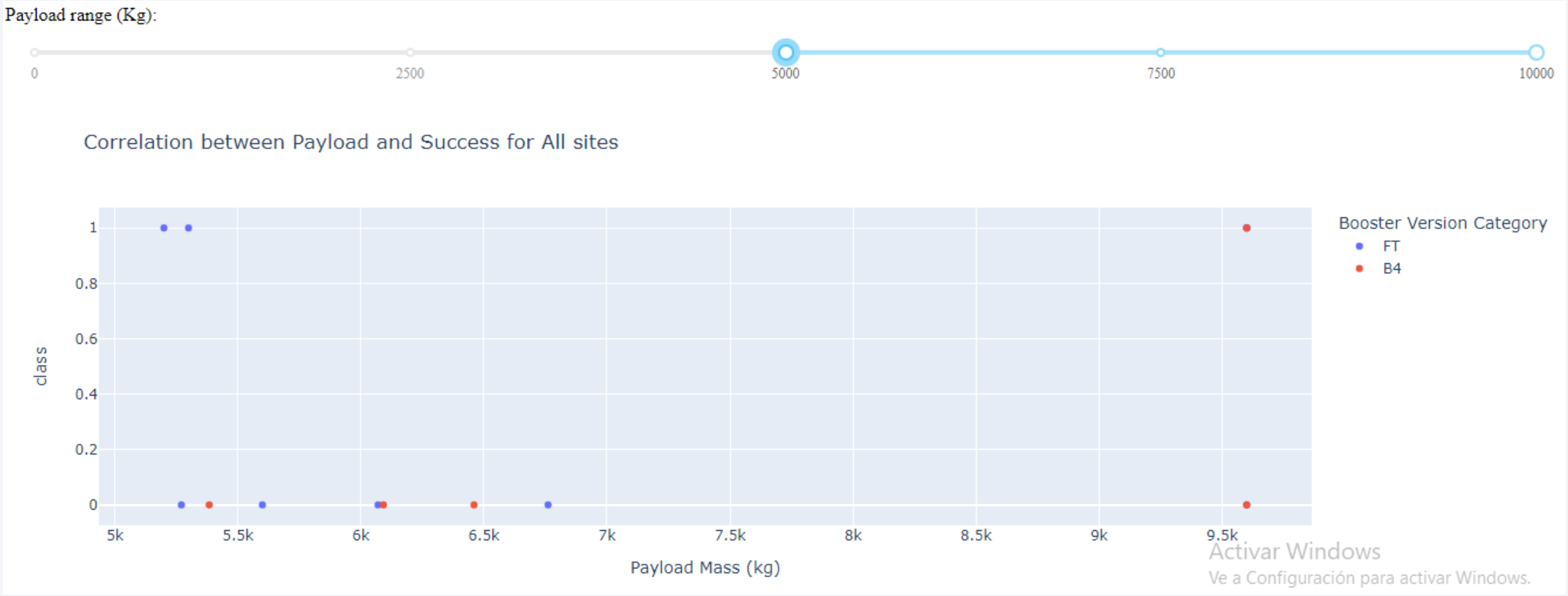
Payload Mass vs Launch Outcome for All Sites

If we filter by the payload mass, we can visualize that most of the successful launches are below of 5000 Kg.



Payload Mass vs Launch Outcome for All Sites

Meanwhile, above 5000 Kg, there are only a few successful launches. Most of the outcomes on this range are failures.

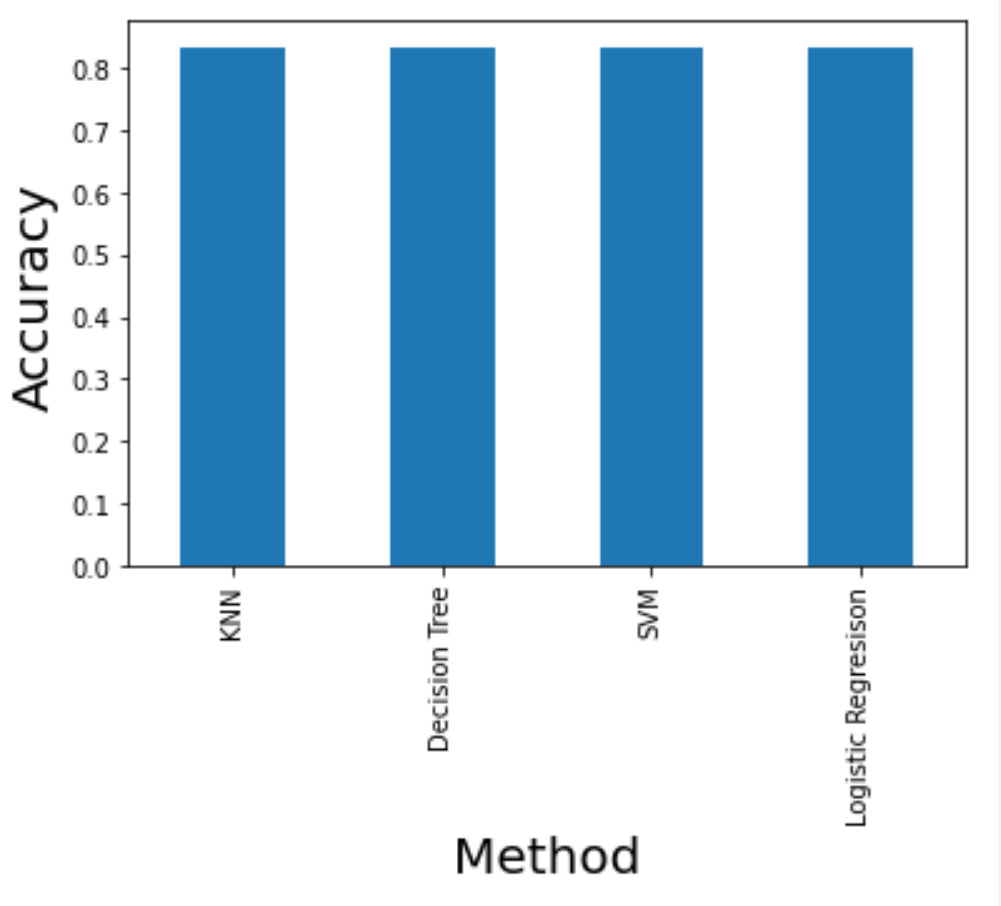


Section 5

Predictive Analysis (Classification)

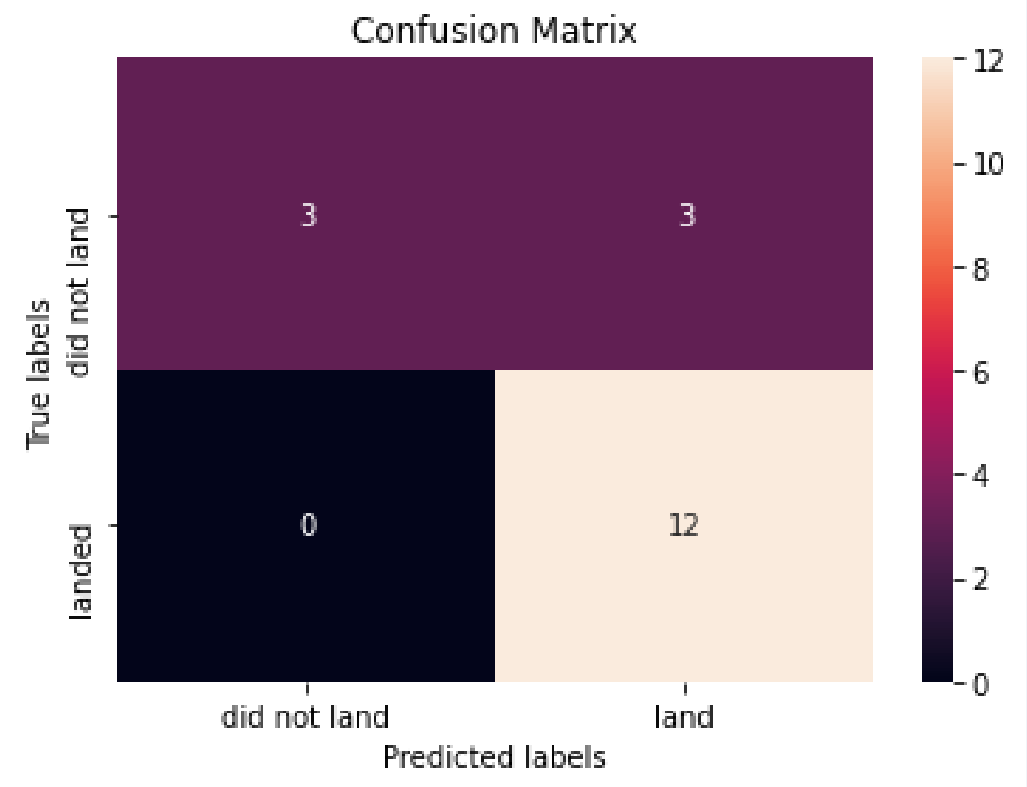
Classification Accuracy

With this bar char we can clearly see that all the methods considered in the analysis perform in the same way.



Confusion Matrix

In this confusion matrix we can see that the model is a good estimator, the accuracy is high when predicting the successful landing outcomes. However, it could be improved to perform better with these 3 false negative values.



Conclusions

- The payload mass feature has a direct impact on the prediction of the mission outcome. The less the mass of the payload, the highest the chance of success landing.
- The methods considered in the predictive analysis perform the same both with the best parameters found and the test data, except from decision tree whose accuracy is higher (88%) than the rest (83%).
- As each launch site has launched different amount of rockets, we cannot relate the flight number with the mission outcome..
- KSC LC-39 launch site and the FT booster version have the highest success rate.

Appendix

Some of the Python libraries used in the project:

- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Numpy
- Folium
- Plotly

Thank you!