IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luis Angel Quezada
15 09 2025

Content

Executive Summary

This presentation is part of the **Applied Data Science Capstone** for the IBM data science professional certificate.

In this capstone project, the behavior of Space X spacecraft is observed, analyzed, and predicted. In that way, the data is collected, cleaned, visualized, analyzed, and used for create a model to predict the success (or failure) of the rocket launches.

The project was made with Python. We used libraries such as Pandas, Numpy, Folium, Dash and Skitlearn for the different stages of the process.

All the documentation is in the next GitHub repository: https://github.com/luisangelquezada88-netizen/Applied-Data-Science-Capstone-IBM/tree/9f176e1143aee17651c9d3747f0934d58f7dbbad

Introduction

Project background and context

SpaceX has become the most successful company of the commercial space era by significantly reducing the cost of space travel. A key factor in this cost reduction is SpaceX's ability to reuse the rocket's first stage. Therefore, if we can predict whether the first stage will successfully land, we can also estimate the actual cost of a launch.

Using publicly available data and machine learning models, this project aims to predict whether SpaceX will be able to reuse the first stage of the rocket.

Key questions to address

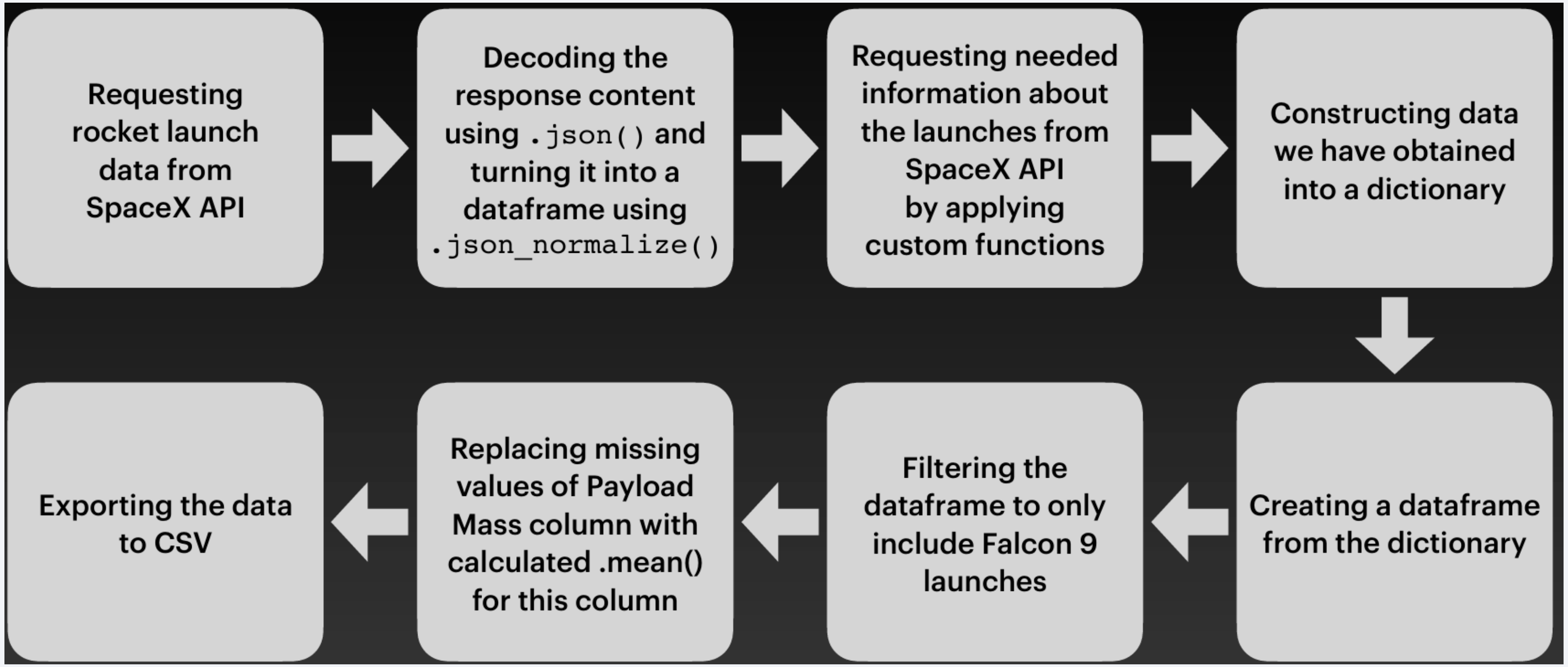
- How do variables such as payload mass, launch site, number of flights, and orbit type affect the success of the first stage landing?
- Has the success rate of landings improved over the years?
- What is the most effective algorithm for binary classification in this context?

Section 1

Methodology

Data Collection

The data was collected throw the use of API's and web scraping for extract the data. Later, the data is transformed to a data frame that we can manipulate, hosted in a CSV file.



Data Wrangling

In this stage, we verify the quality of the data. Furthermore, we execute queries with pandas.

```
df.isnull().sum()/len(df)*100
```

```
FlightNumber       0.000000
Date               0.000000
BoosterVersion     0.000000
PayloadMass        0.000000
Orbit              0.000000
LaunchSite         0.000000
Outcome            0.000000
Flights            0.000000
GridFins           0.000000
Reused             0.000000
Legs               0.000000
LandingPad        28.888889
Block              0.000000
ReusedCount        0.000000
Serial             0.000000
Longitude          0.000000
Latitude           0.000000
dtype: float64
```

```
df.dtypes
```

```
FlightNumber        int64
Date               object
BoosterVersion     object
PayloadMass       float64
Orbit              object
LaunchSite         object
Outcome            object
Flights             int64
GridFins             bool
Reused               bool
Legs                 bool
LandingPad         object
Block             float64
ReusedCount         int64
Serial             object
Longitude         float64
Latitude          float64
dtype: object
```

EDA with Data Visualization and SQL

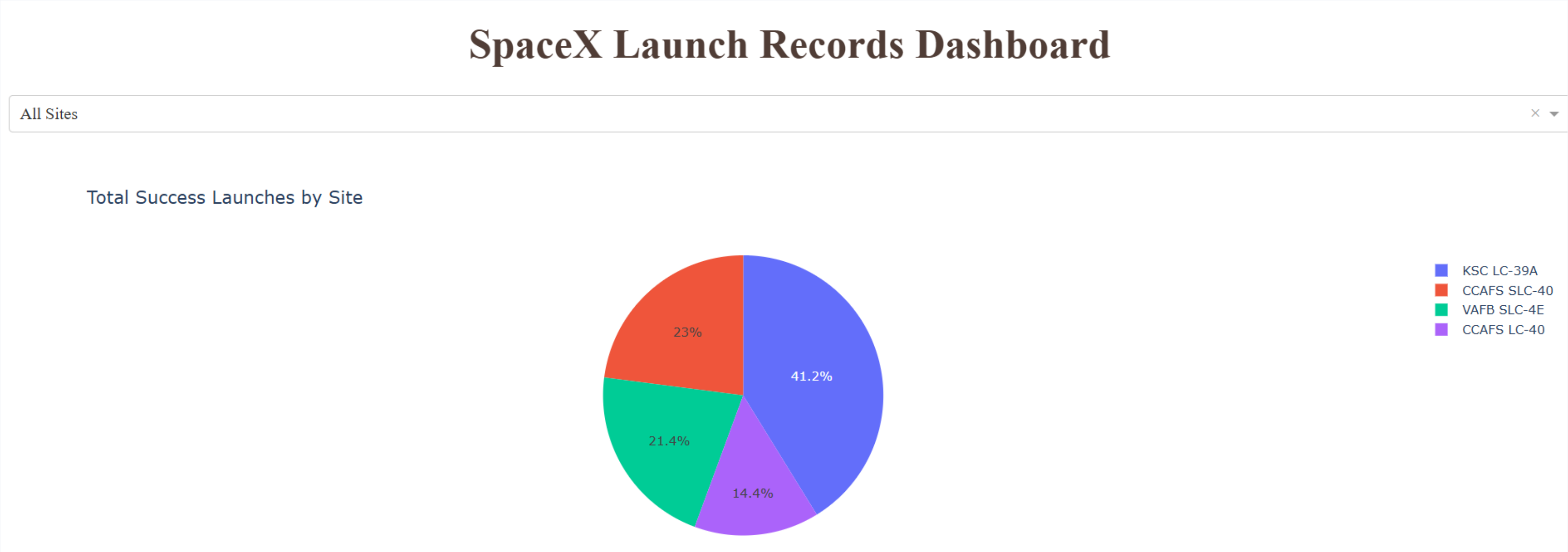
For the Exploratory Data Analysis, we can use some charts and grafics (data visualization) to see the relationship between some variables. We use scatterplots, barplots, pie plots, line charts, etc.

We can use SQL to do useful queries in the structured data.

Interactive visualization

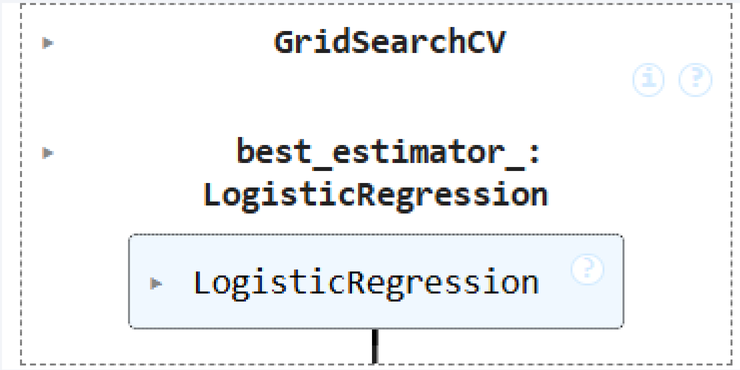
We can visualize the ubication of the launch sites, using the Folium library to create interactive maps with the geographical information of the project. In that way, we use markers, clusters, lines and other elements to do a basic geographical analysis.

The interactive visualization comprises the design of a Dashboard. Using the Dash library, we created an interactive dashboard with the information of the success of failure of the launches by launchsite.



Predictive Analysis (Classification)

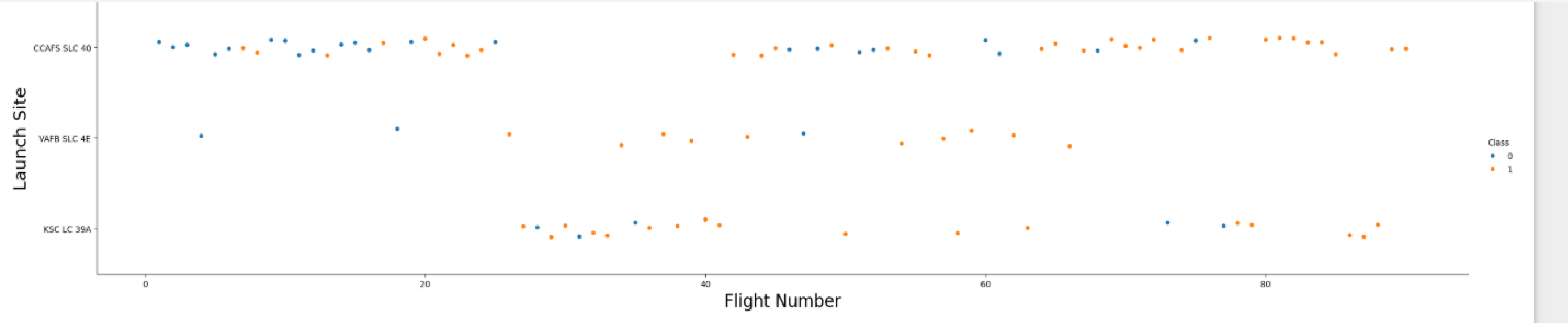
In this stage, a model is designed to predict the success or failure of the launches. In that way, it's necessary to use Machine Learning Algorithms, dividing the data between training and test set, evaluating the best hyperparameter, choosing the more precise method and others.



Section 2

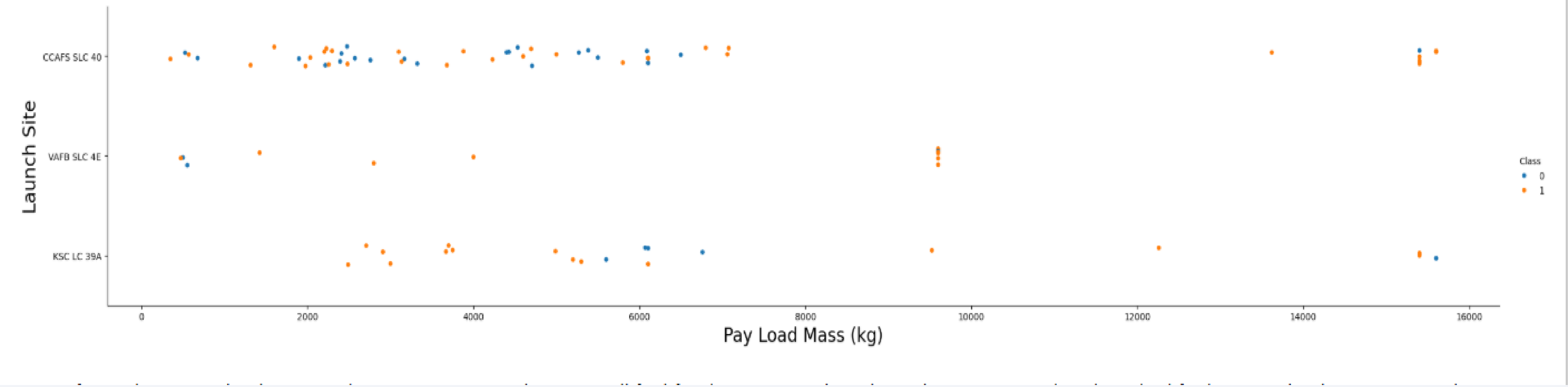
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

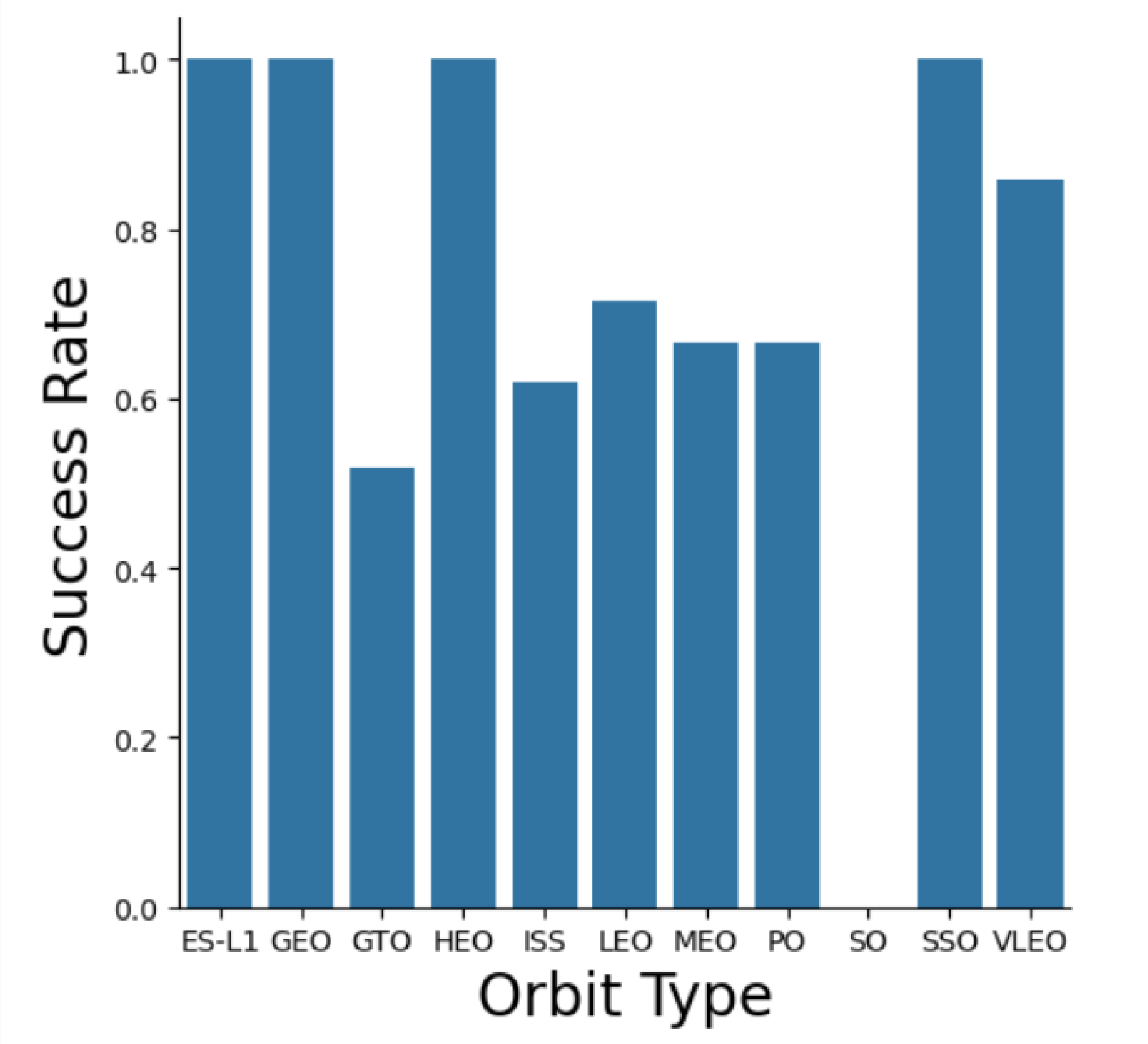
Payload vs. Launch Site



- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

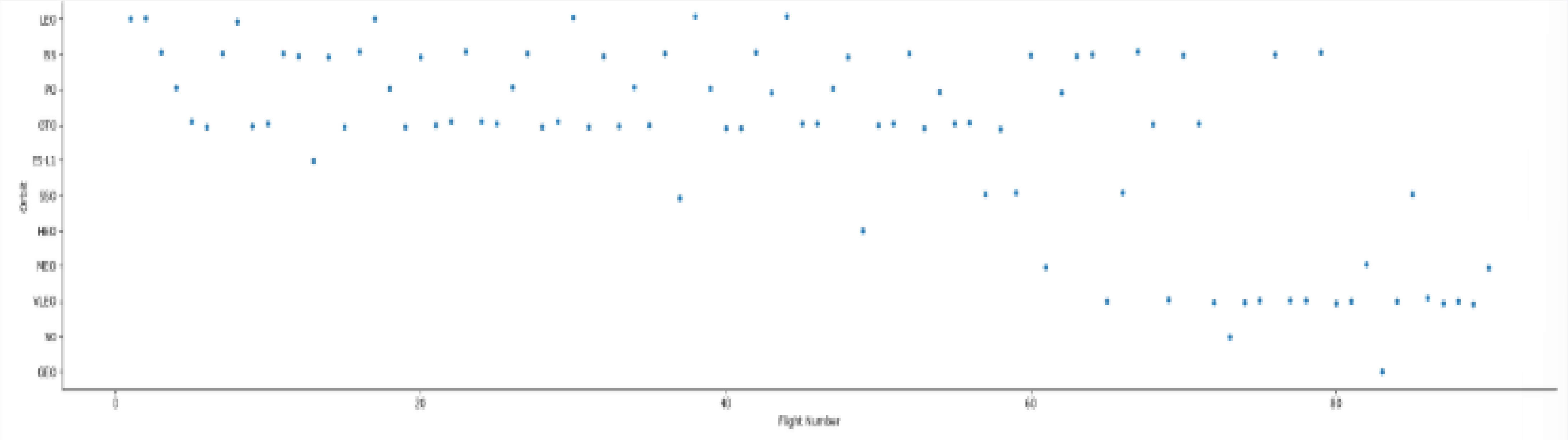
Success Rate vs. Orbit Type

- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO



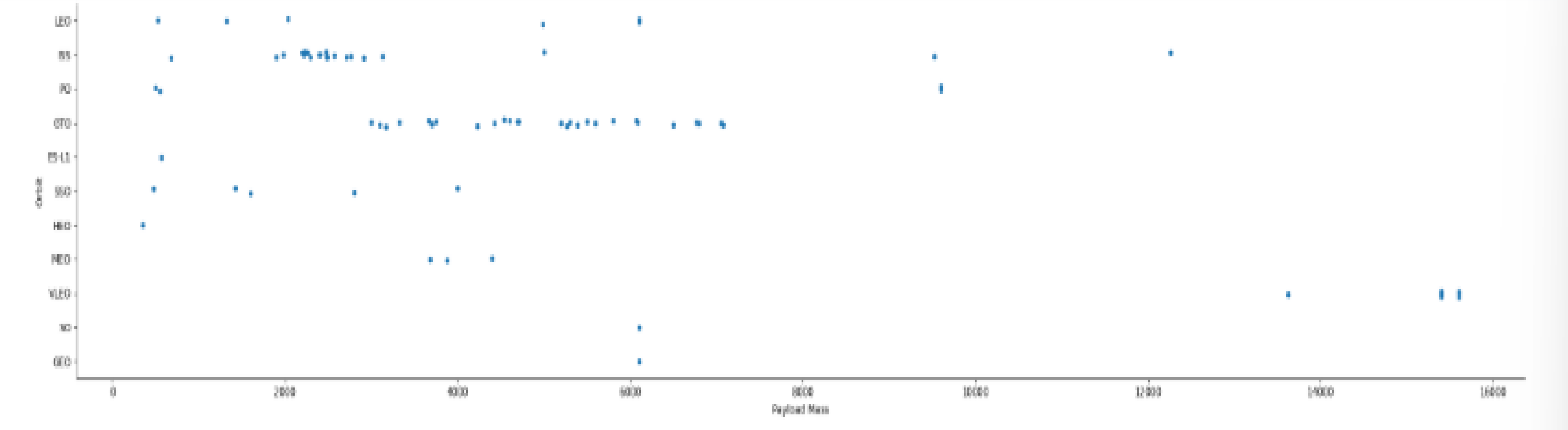
Flight Number vs. Orbit Type

In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



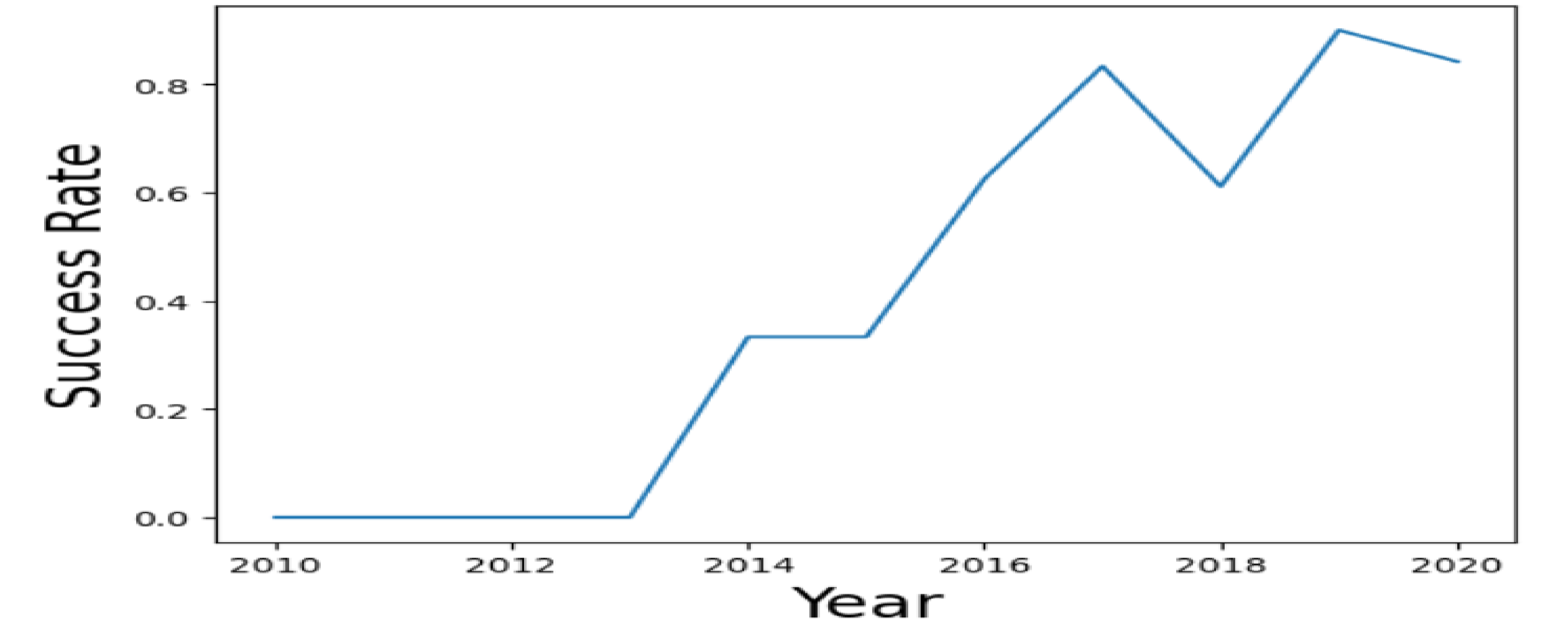
Payload vs. Orbit Type

Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020.



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct Launch_Site from spacex
```

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

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
select Launch_Site from spacex
where Launch_Site like 'CCA%'
limit (5)
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
select sum (PAYLOAD_MASS__KG_) from spacex
where Customer = 'NASA (CRS)'
```

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

sum (PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql
select AVG (PAYLOAD_MASS__KG_) from spacex
where Booster_Version like 'F9 v1.1%'
```

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

AVG (PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

Display average payload mass carried by booster version F9 v1.1

```
%%sql
select AVG (PAYLOAD_MASS__KG_) from spacex
where Booster_Version like 'F9 v1.1%'
```

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

AVG (PAYLOAD_MASS__KG_)
2534.6666666666665

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql select distinct Booster_Version from spacex
where Landing_Outcome = 'Success (drone ship)'  and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///spacex.db
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

List the total number of successful and failure mission outcomes

```
%%sql
select Mission_Outcome, count(*) as total_number from spacex
group by Mission_Outcome;
```

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

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

All the booster versions that have carried the maximum payload mass.

2015 Launch Records

The failed landing outcomes in drone ship with their booster versions and launch site names for in year 2015

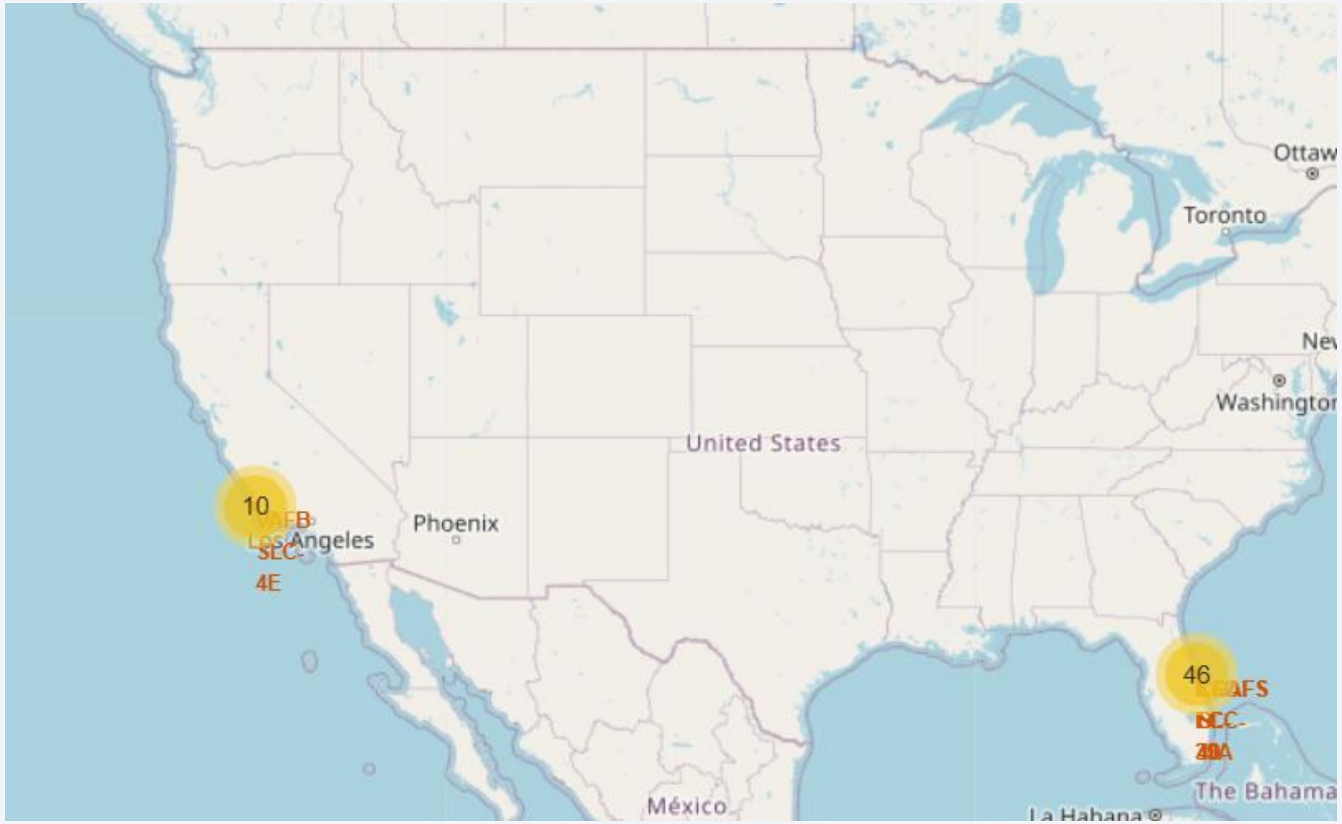
Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Section 3

Launch Sites Proximities Analysis

Launch sites

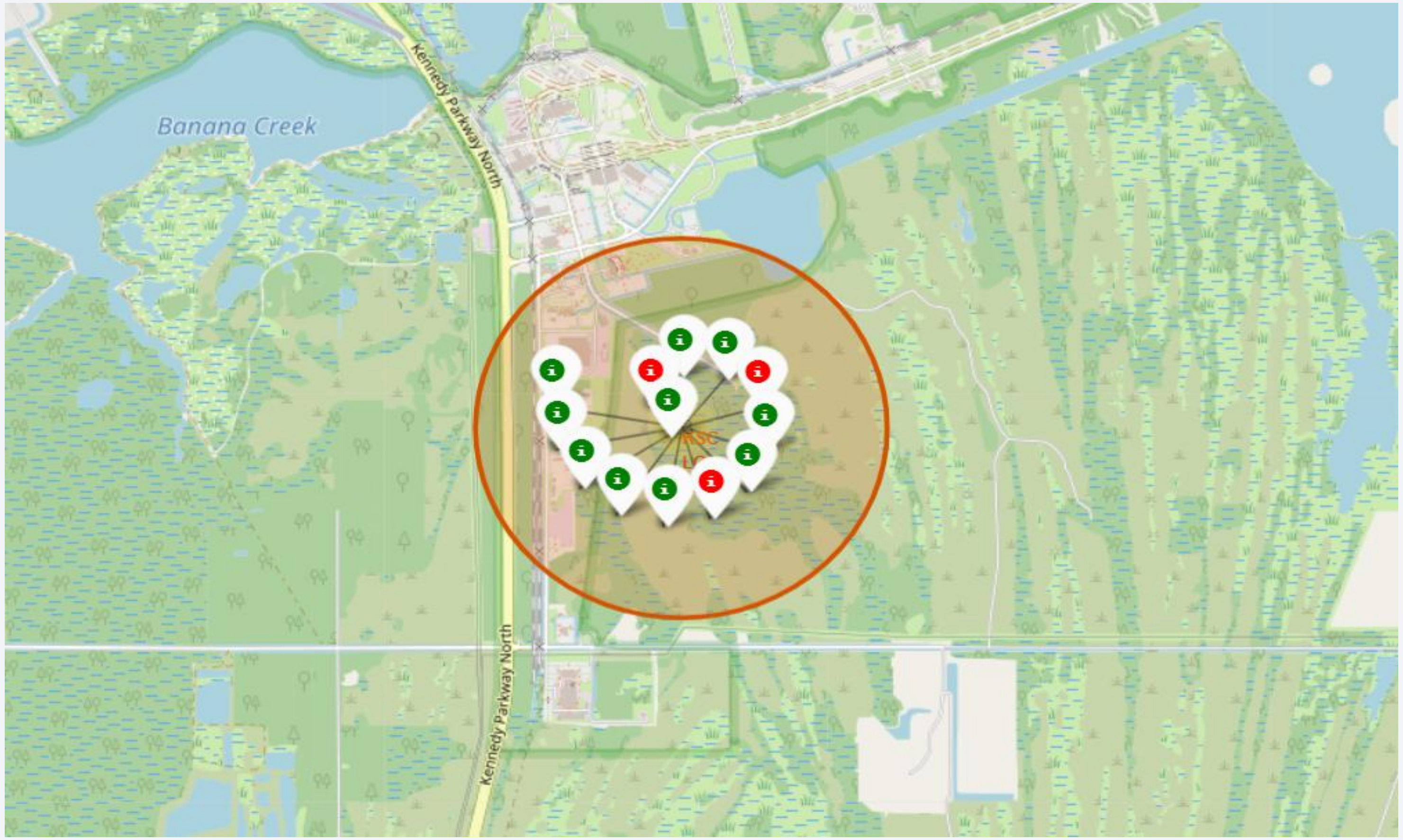
An interactive map was constructed to see the geographical ubication of the launch sites.



Success/failure of the launches

Green: Success

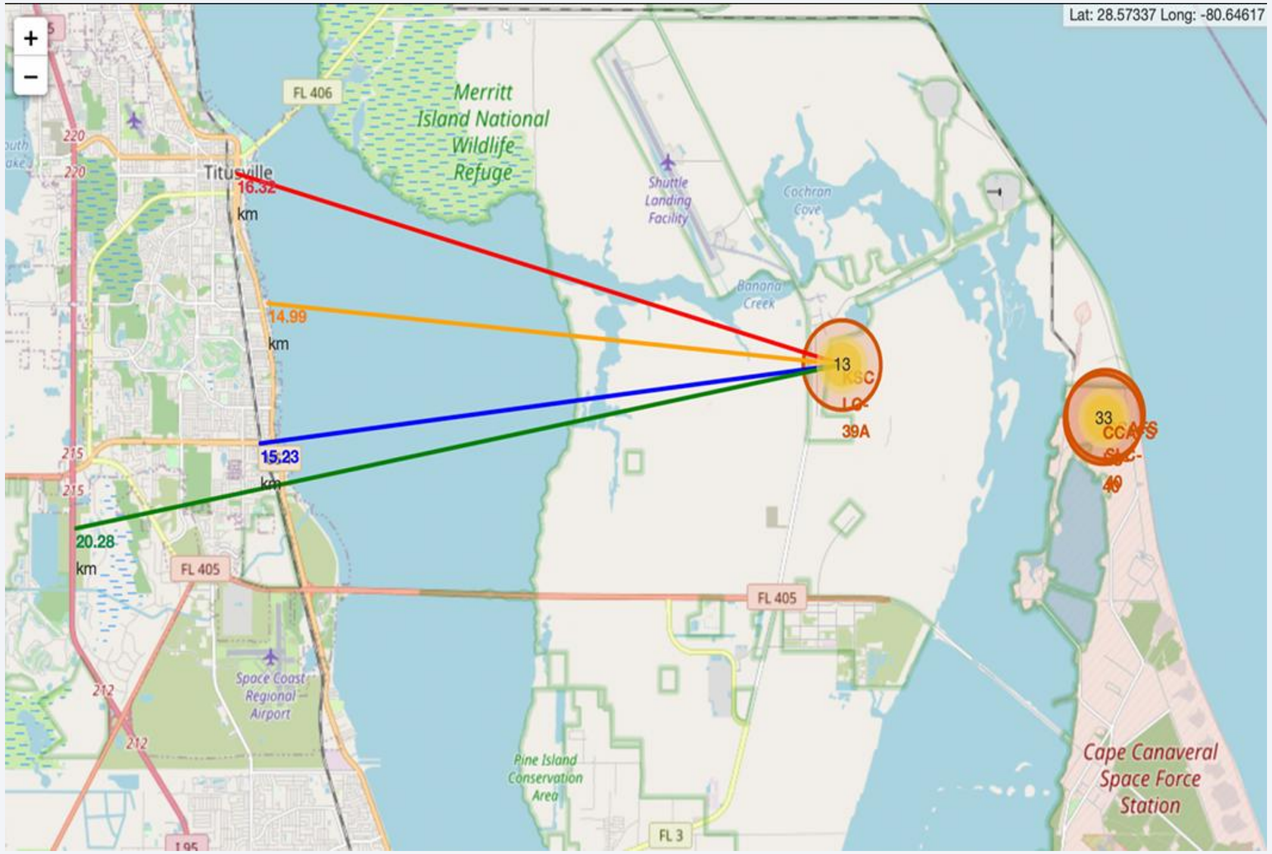
Red: Failure



Proximities

The Launch Site **KSC LC-39A**:

- Is 15.23 km from the railway
- Is 20.28 km from the highway
- Is 14.99 km from the coastline

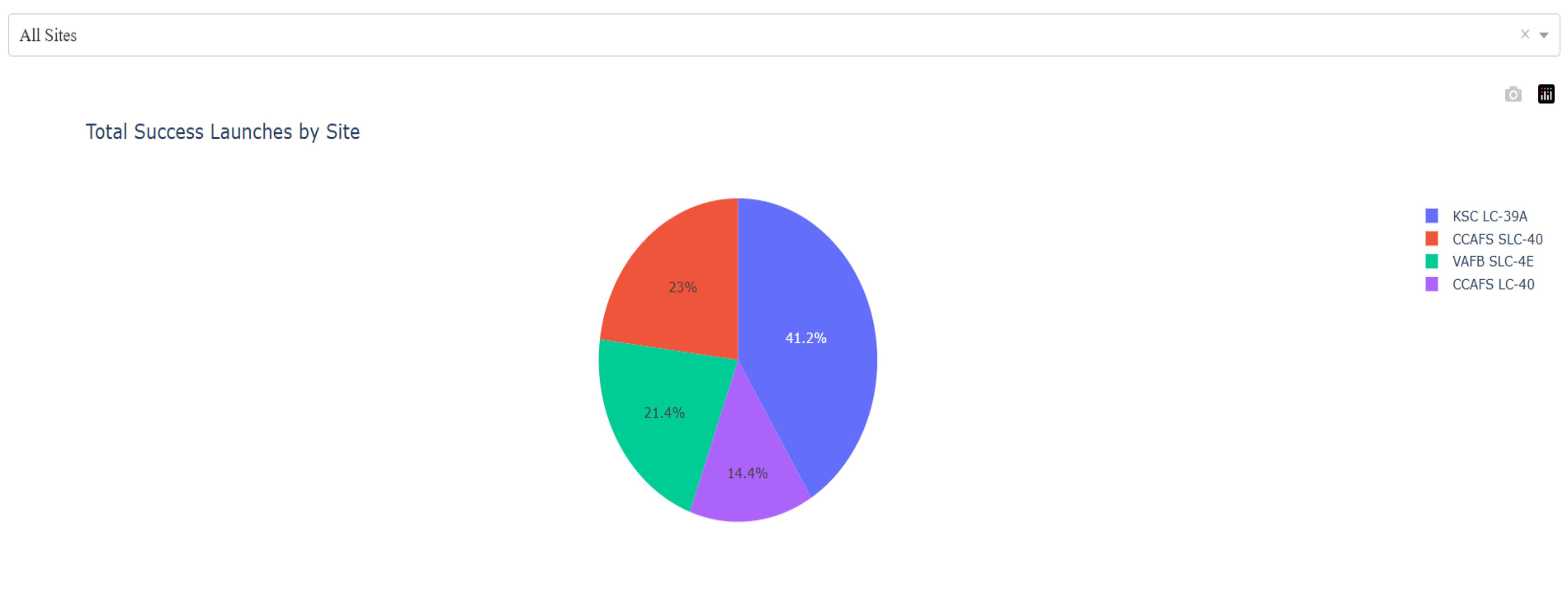


Section 4

Build a Dashboard with Plotly Dash

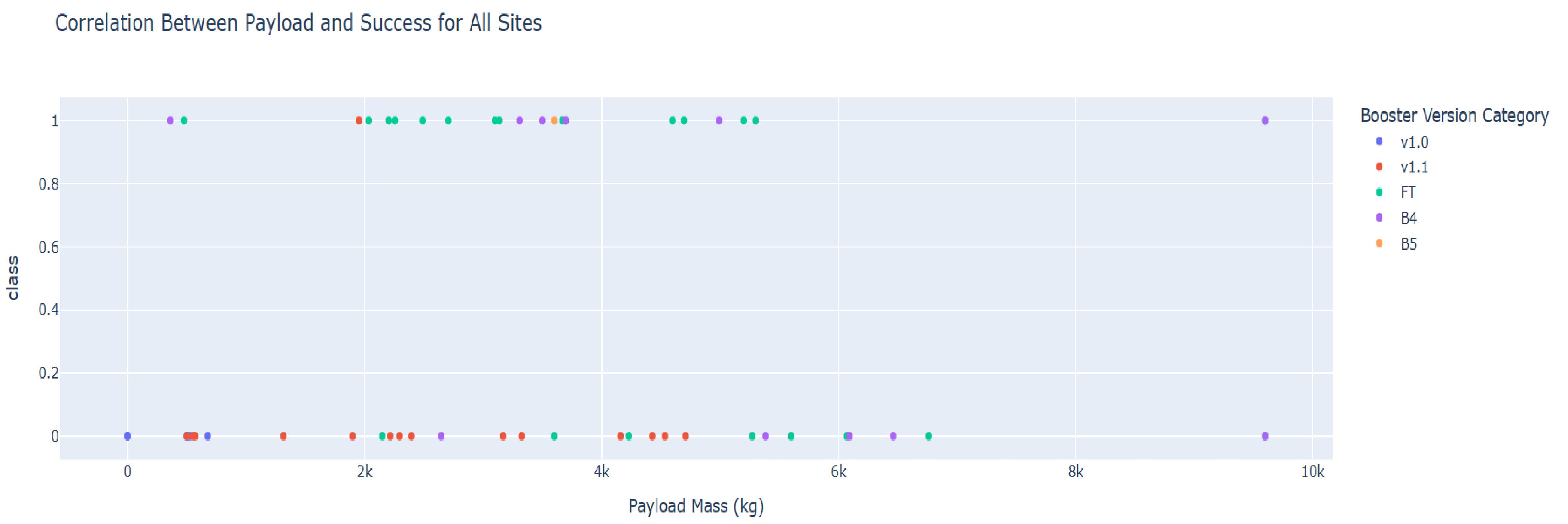
Interactive Dashboard

KSC LC 39-A is the launch site with the greatest number of successful launches.



Correlation between Success and Payload Mass

The charts show that payloads between 2000 and 5500 kg have the highest success rate.



Section 5

Predictive Analysis (Classification)

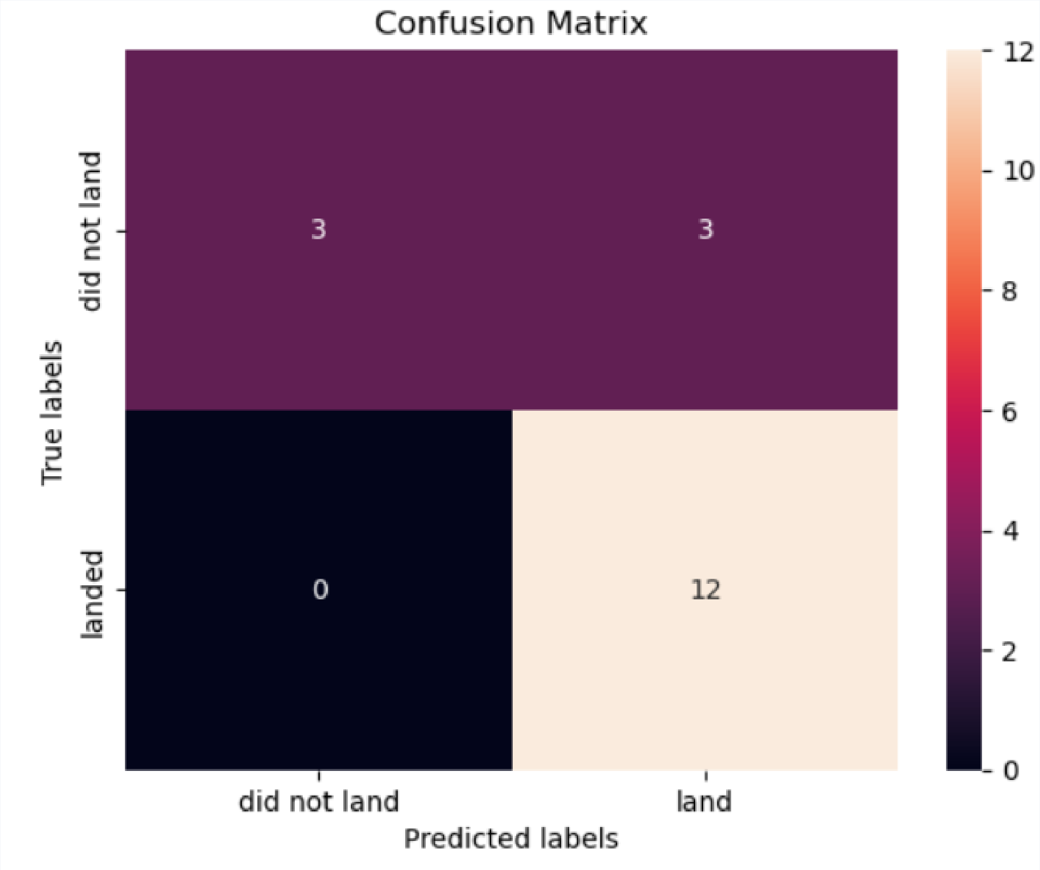
Classification Accuracy

The most accurate model is decision tree.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Based on the confusion matrix, the most important problem are the False Positives (FN)



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!