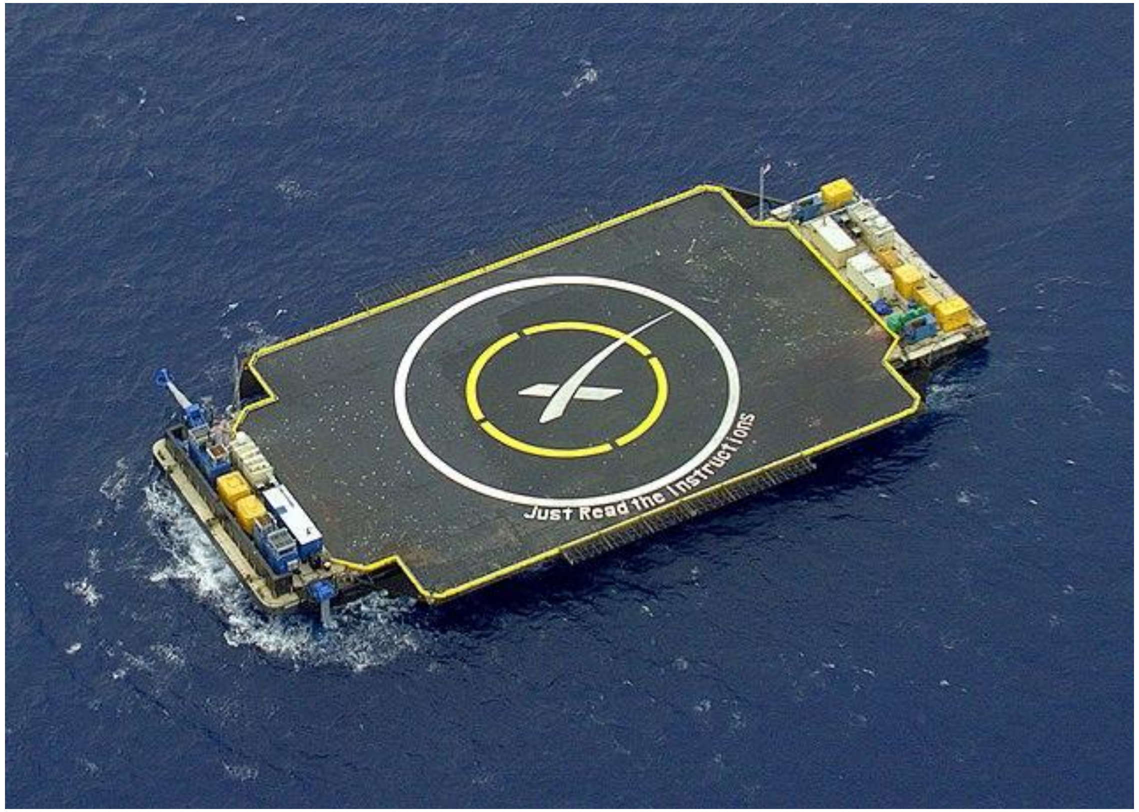
SpaceX booster landing prediction

DIGBO Pierre-E
17/06/2023

Content



Executive Summary

Summary of the methodologies

- Data collection from the SpaceX REST API and web scraping wiki pages about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Data wrangling: converting “raw” data into another format more suitable for use.
- Exploratory data analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Data visualization to explore and manipulate data in an interactive and real-time way.
- Model development and model evaluation using machine learning to determine if the first stage of Falcon 9 will land successfully.

Results

- There is a relationship between the payload, the orbits and the success of the landing.
- SSO orbit has a 100% of success rate for a payload below 6000 kg.
- The launch success rate gets better with time.
- Most of the launch sites are close to the equator and are in close proximity to the coastline.
- Among all the prediction methods: it is the decision tree that slightly performs better.

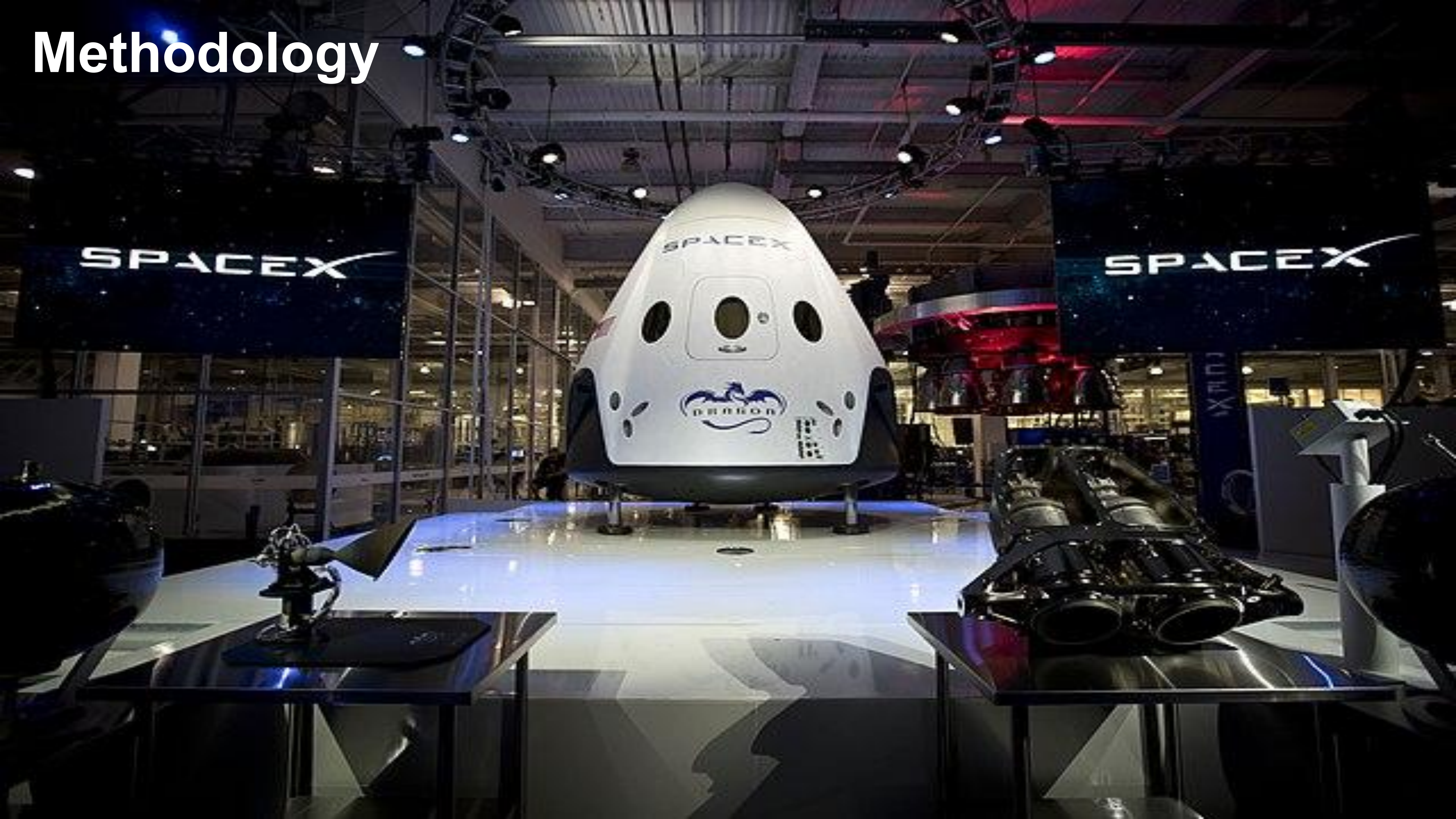
Introduction

Project background and context

Space X has made one of the biggest innovations in the aerospace industry in recent years by being able to land its booster that counts for a big part of the total cost of launching a rocket and hence making it much cheaper. Launching a rocket usually costs $165M while Space X is able to decrease this cost to $65M.

Predicting if a first stage will land successfully can help to determine the total cost of a launch. This information can therefore be used to compete and bid against Space X.

Methodology



Data collection with API

- Data request from the Space X API (https://api.spacexdata.com): Request and parse the SpaceX launch data using the GET request.
- Turning the JSON response into a pandas dataframe
- Getting information about the launches (rocket, payloads, launchpad, and cores) using the IDs given for each launch.
- Filtering the dataframe to only include Falcon 9 launches

https://github.com/b4tos4i/capst/blob/main/Data%20collection%20API.ipynb

Data collection with web scraping

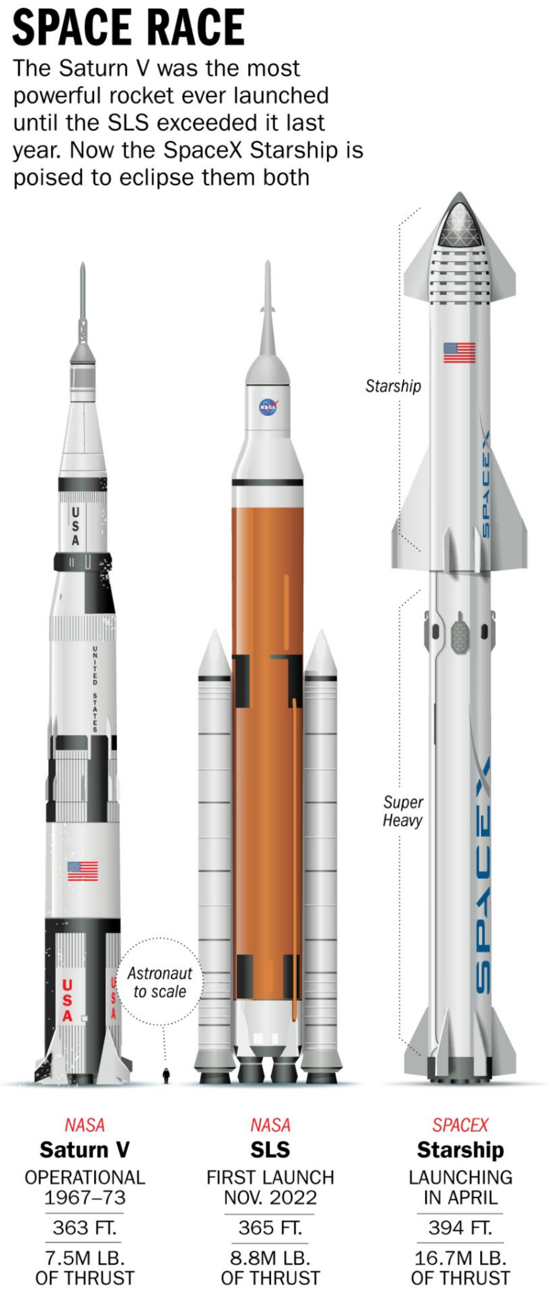
- Request the Falcon9 Launch Wiki page from its URL by performing a HTTP GET method
- Creation of a BeautifulSoup object from the HTML response
- Extraction of all column/variable names from the HTML table header
- Creation of a data frame by parsing the launch HTML tables

https://github.com/b4tos4i/capst/blob/main/Data%20collection%20webscraping.ipynb

Data wrangling

- Identifying and calculating the percentage of the missing values in each attribute
- Calculation of the number of launches on each site
- Calculation of the number and occurrence of each orbit
- Calculation of the number and occurence of mission outcome per orbit type
- Conversion of the outcomes into training labels with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful.

https://github.com/b4tos4i/capst/blob/main/Data%20wrangling.ipynb



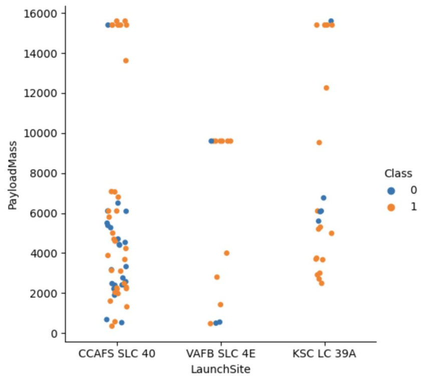
EDA with data visualization

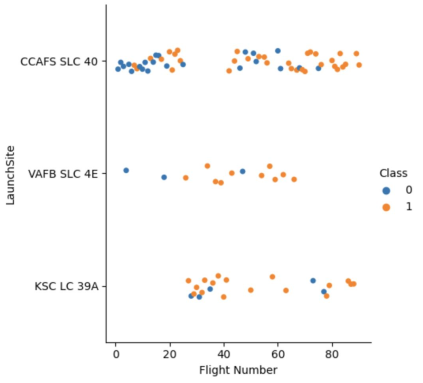
- Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.

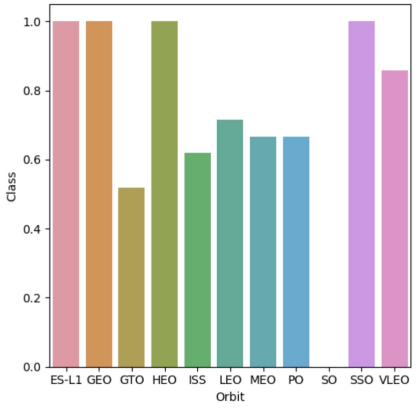
- **Visualization of the relationship** between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - FlightNumber and Orbit type
  - Payload and Orbit type

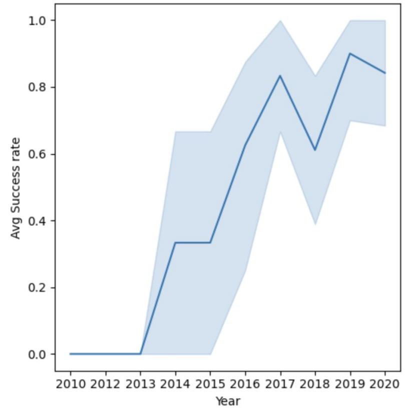
  by using **scatter plots**.
- Visualization of the **comparison** between
  - success rate of each orbit type

  by using **bar charts**.
- Visualization of the launch success yearly **trend** by using a **line chart**.









https://github.com/b4tos4i/capst/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- 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. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/b4tos4i/capst/blob/main/EDA%20with%20SQL.ipynb

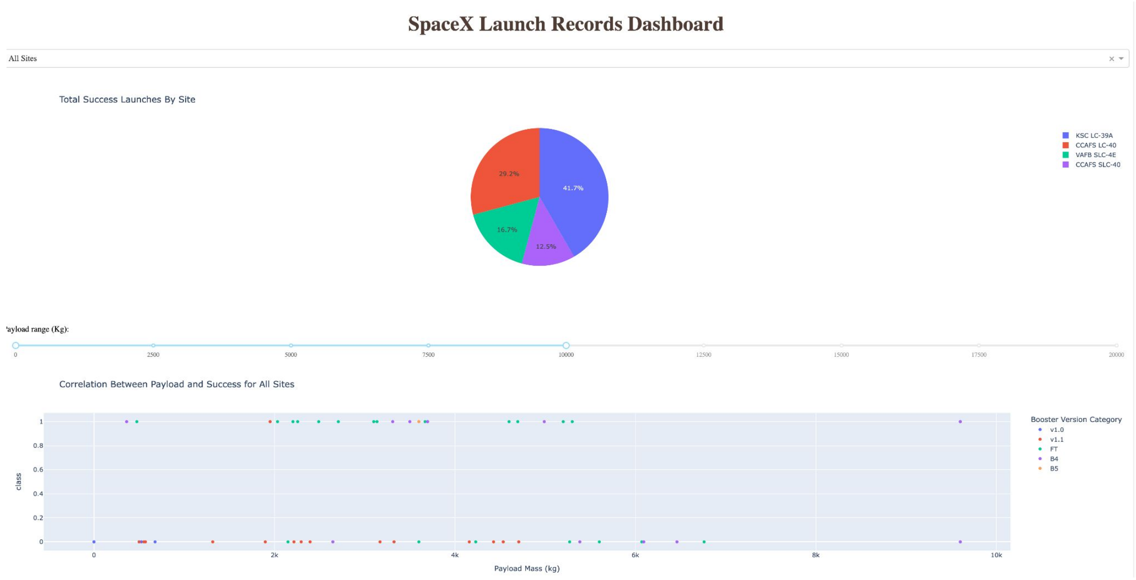
Interactive map with folium

- Marking of all launch sites on a map using a dataframe with their latitude and longitude and highlighting the name of the launch site on the map.
- Marking of the **success/failed launches** for each site on the map using **color-labeled marker clusters**, green being the successful launches, red being the unsuccessful ones.
- Calculation of the **distances** between a launch site to its proximities (city/railway/highway) and highlight of this distance by adding a **colored line**.

https://github.com/b4tos4i/capst/blob/main/Interactive%20visual%20analytics%20with%20Folium.ipynb

Dashboard with Plotly Dash

- Adding a **dropdown list** to enable Launch Site selection
- **Pie chart** to show the total successful launches count for all sites
- **Slider** to select payload range
- **Scatter chart** to show the correlation between payload and launch success



https://github.com/b4tos4i/capst/blob/main/Interactive%20Visual%20Analytics%20with%20Plotly%20dash.py

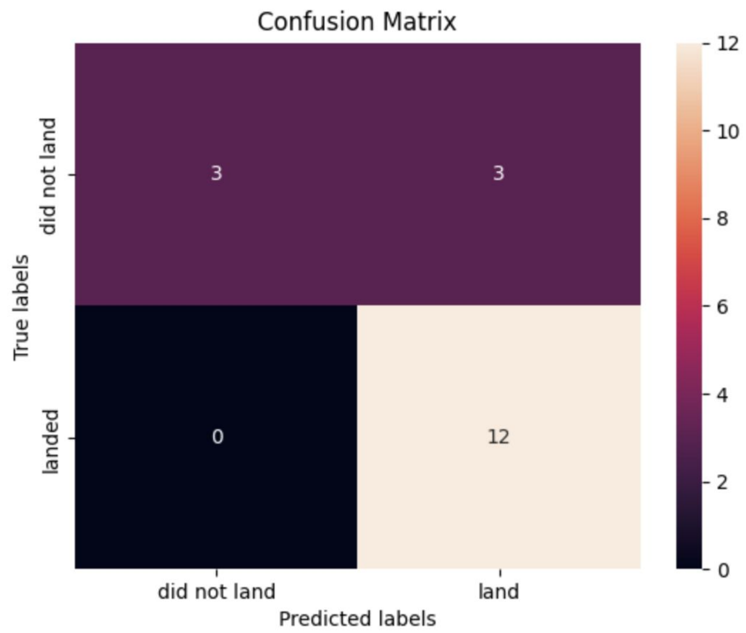
Predictive analysis

- Creation a **NumPy array** from the column Class in data.
- **Standardize** and **fit** the data.
- **Split** of the data into **training and test**
- Creation of a **GridSearchCV object** after having created a **logistic regression**, **SVM**, **decision tree classifier**, and **KNN** object.
- Finding the method that performs by using:
  - the **accuracy** with **.score()** method.
  - a **confusion matrix**.

https://github.com/b4tos4i/capst/blob/main/Machine%20Learning%20Prediction.ipynb

In [26]:
```
lr_score = logreg_cv.score(X_test, Y_test)
print("score :", lr_score)
```

```
score : 0.8333333333333334
```



Results

Exploratory data analysis results

- There is a relationship between the payload, the orbits and the success of the landing.
- SSO orbit has a 100% of success rate for a payload below 6000 kg.
- The launch success rate gets better with time.

Interactive analytics

- Most of the launch sites are close to the equator and are in close proximity to the coastline.
- But also far from city centers while being accessible by highways and railways.

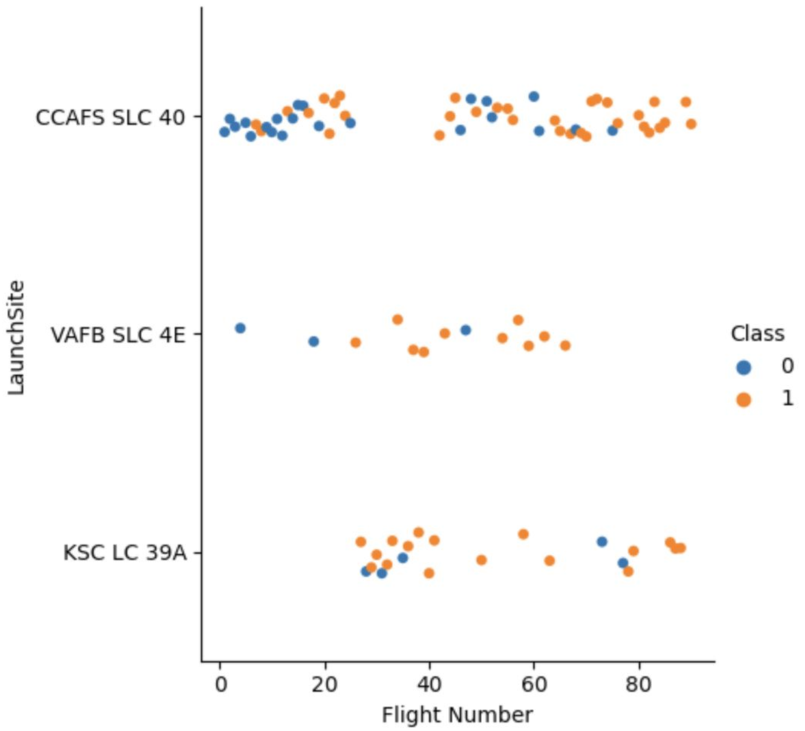
Predictive analysis

- Among all the prediction methods: it is the decision tree that slightly performs better.

Insights drawn from EDA

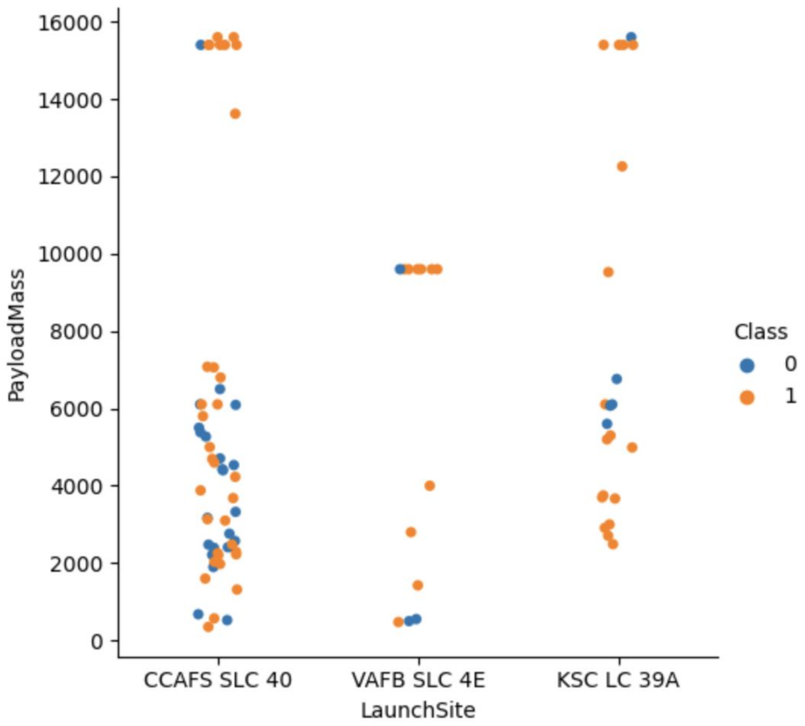
Flight number vs launch site

- KSC LC-39A and VAFB SLC 4E have a better success rate than CCAFS LC-40.
- The success of a landing is correlated to the number of flights. The more launches are done the more landing are successful.



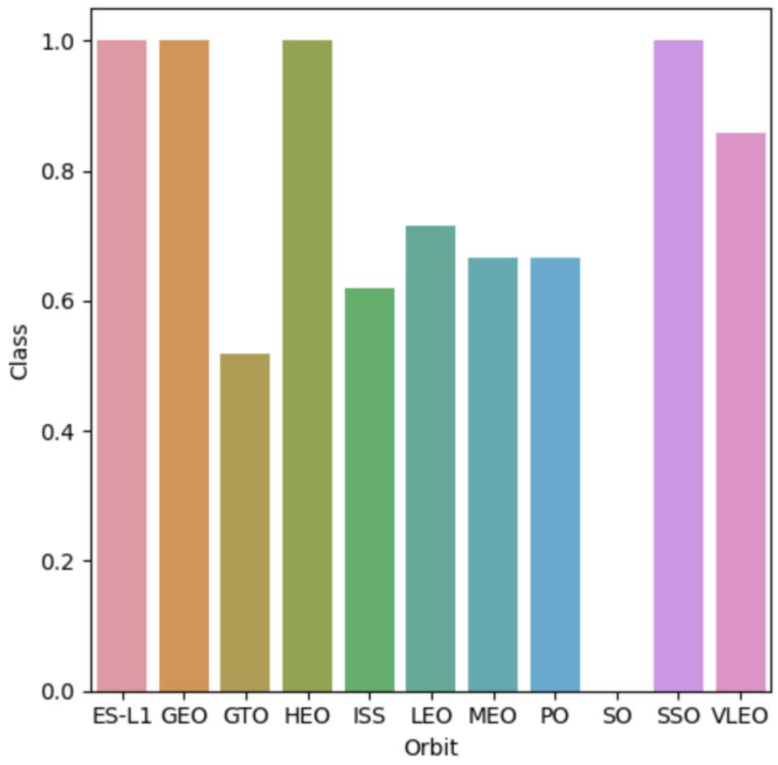
Payload vs launch site

- There is a correlation between the pay load mass and the success rate: the heavier it is (above 10000kg), the higher the success rate.
- VAFB SLC 4E doesn't launch above 10000kg.
- KSC LC 39A has a 100% success rate below 5500kg



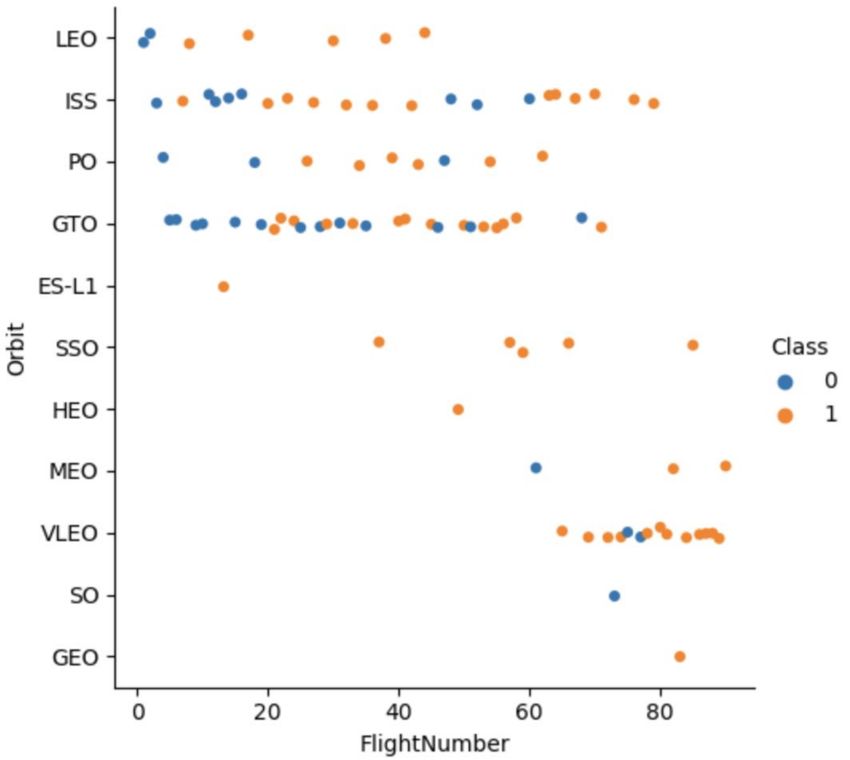
Success rate vs orbit type

- ES-L1, GEO, HEO, SSO have a 100% success rate.
- GTO has the lowest success rate.



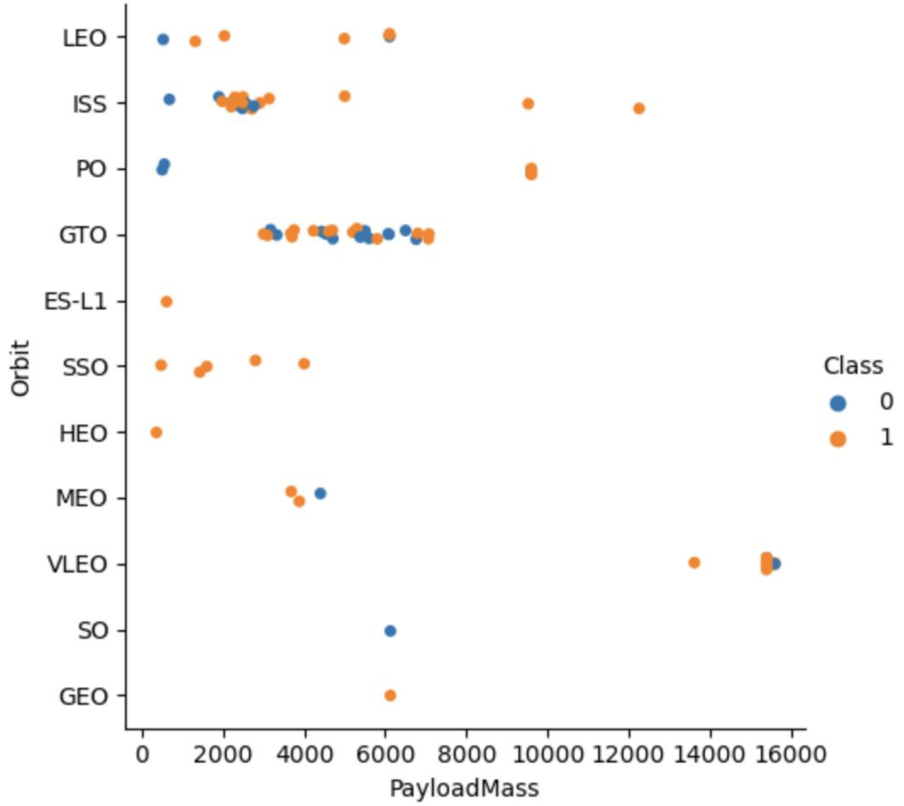
Flight number vs orbit type

- From flight number 75, it is a 100% success rate
- SSO orbit has the highest success rate (100%) for several flights.
- The success rate increases with the number of flights for the orbits VLEO and ISS.



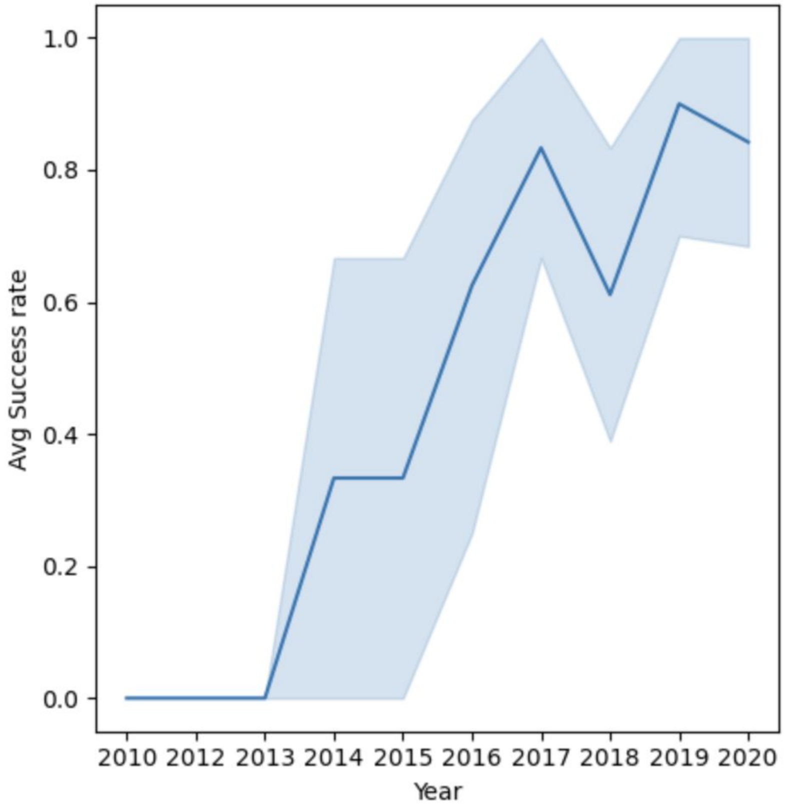
Payload vs orbit type

- SSO orbit has the highest success rate and is used for payload mass of less than 4000 kg.
- 100% of success rate is achieved for orbit ISS and a payload mass from 4000 kg.
- VLEO seems to be the only orbit with the highest success rate for payload mass of more than 14000kg.



Launch success yearly trend

- Overall the success rate improved with time
- The highest improvement period was from 2013 to 2017.



All launch site names

There are 4 launch sites:

a. CCAFS LC-40
b. VAFB SLC-4E
c. KSC LC-39A
d. CCAFS SLC-40

In [9]:
```
%sql select distinct launch_site from spacextbl;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch site names starting with CCA (5 records)

In [10]:

```
%sql select * from spacextbl \
    where launch_site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total payload mass carried by boosters from NASA

In [11]:

```
%sql select sum(payload_mass__kg_) from spacextbl \
     where customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

sum(payload_mass__kg_)
45596.0

Average payload mass by F9 v1.1

```
In [12]: %sql select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.
```

Out[12]:

avg(payload_mass__kg_)
2928.4

First successful ground landing date

```
[16]: %sql select min(date) from SPACEXTBL where landing_outcome = "Success (ground pad)"

 * sqlite:///my_data1.db
Done.
```

[16]:

min(date)
01/08/2018

The first successful ground landing was on 1st August 2018

Successful drone ship landing with payload between 4000 and 6000

```
[15]: %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000

 * sqlite:///my_data1.db
Done.
```

[15]:

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

The best boosters for drone ship landing with a payload mass between 4000 kg and 6000 kg are:

1. F9 FT B1022
2. F9 B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

Total number of successful and failure mission outcomes

In [16]:
```
%sql select mission_outcome, count (mission_outcome) from spacextbl group by mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Mission_Outcome	count (mission_outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

There is a total of 99 successful mission.

Booster carried maximum payload

[17]:
```
%sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

Failed drone ship landing records in 2015

[18]:
```
%sql select *,  substr(Date, 4, 2) as 'month names' from spacextbl where landing_outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	month names
01/10/2015	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	10
14/04/2015	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	04

The failed drone ship mission in 2015 used:

- the CCAFS LC-40 launch site
- the boosters F9v1.1B1012 and F9v1.1B1015
- for a respective payload mass of 2395 kg and 1898 kg.

Rank landing outcomes 2010-06-04 and 2017-03-20

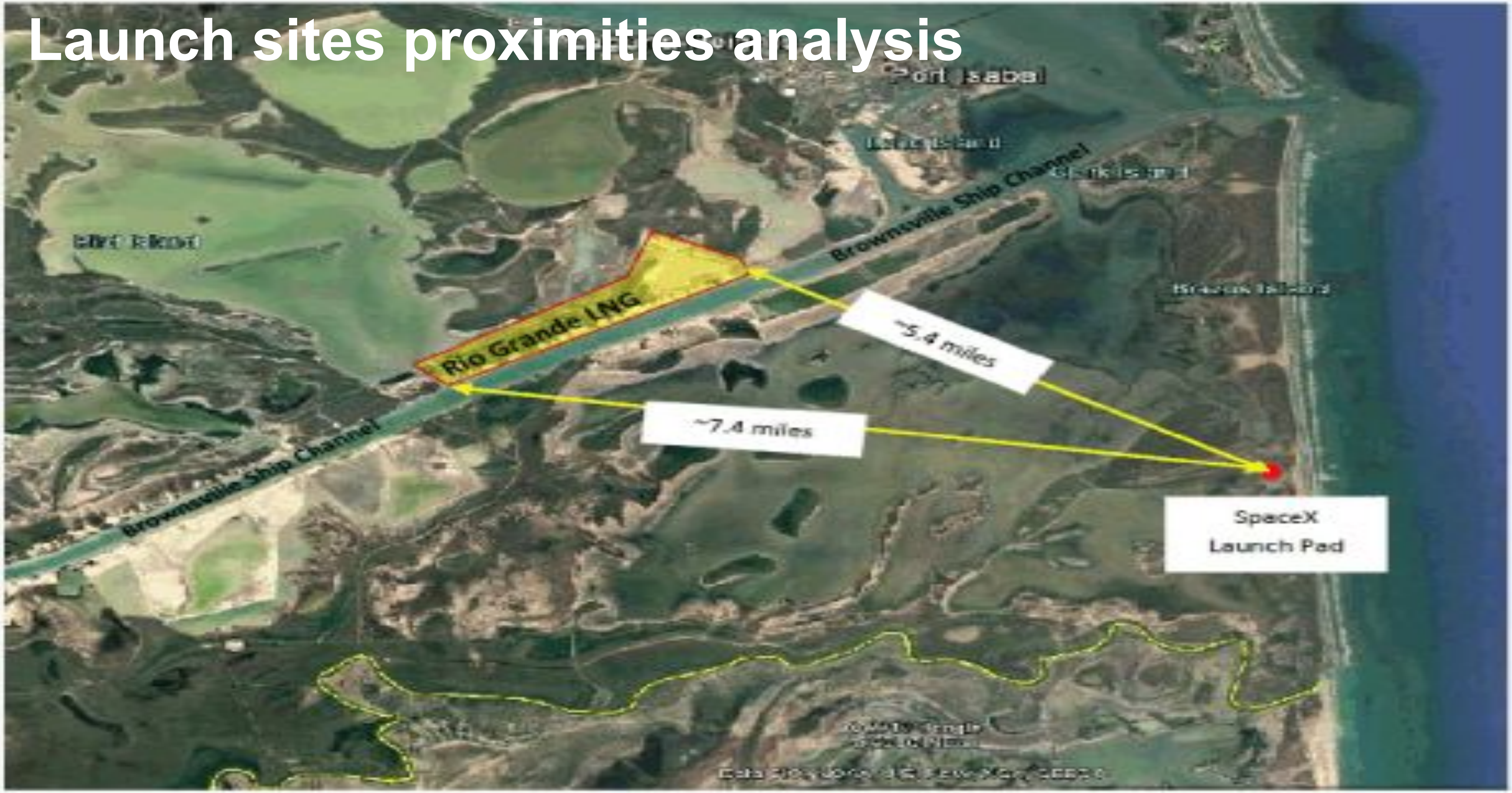
[125]:
```
%%sql select Landing_Outcome, count(Landing_Outcome) as count
from spacextbl
where Date between '04-06-2010' and '20-03-2017'
and (Landing_Outcome like 'Success%')
group by Landing_Outcome
order by count desc
```

```
 * sqlite:///my_data1.db
Done.
```

[125]:

Landing_Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	7

Launch sites proximities analysis

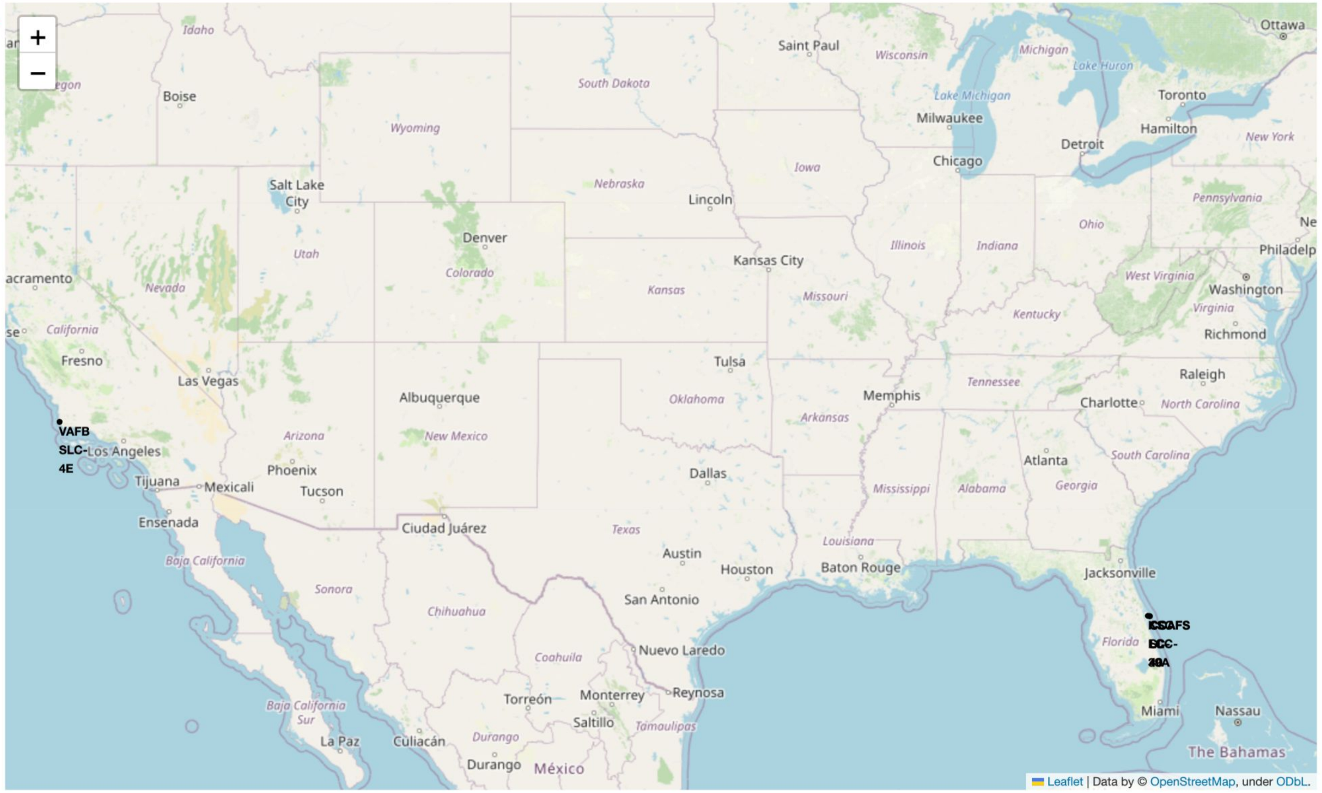


Source: Rio Grande LNG LLC

Launch sites’ location

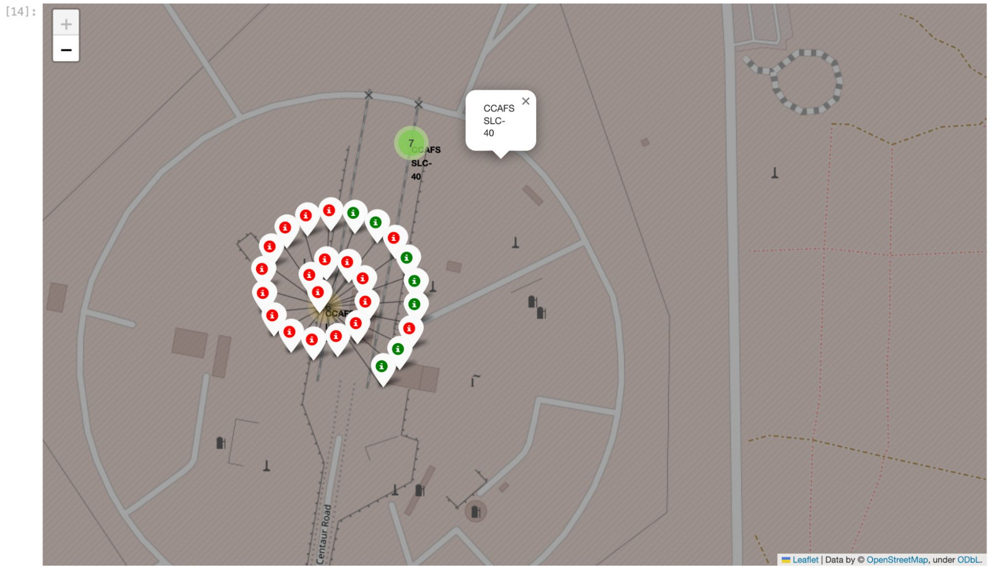
All the launch sites are close to the equator (relative to the US territory) and on the coastline.

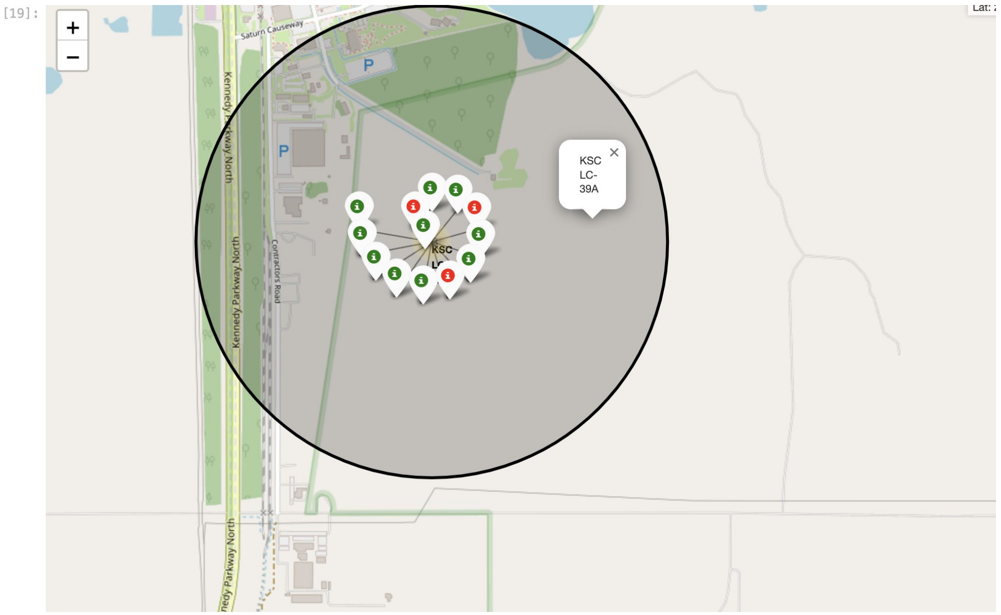
[9]:



Launch outcomes

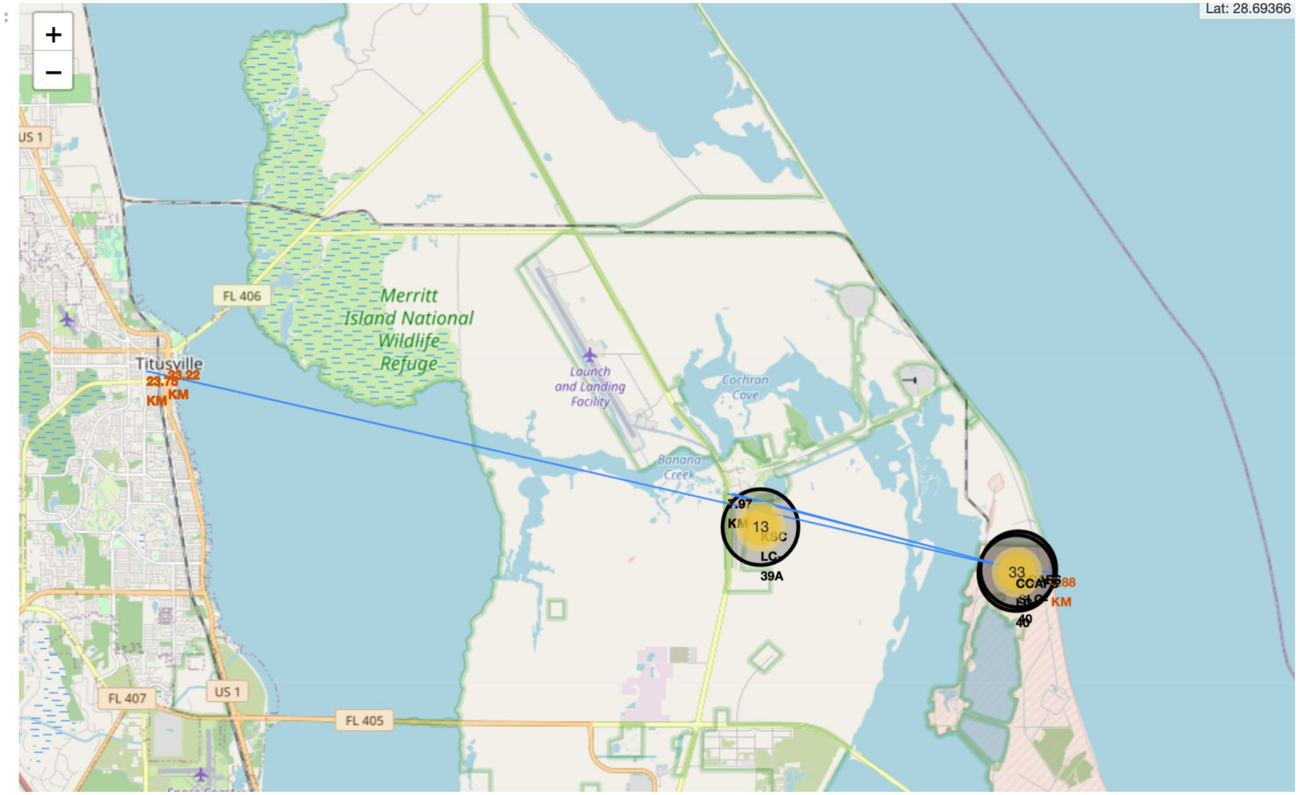
KSC LC-39A has the highest success rate (10/13 => 77%).





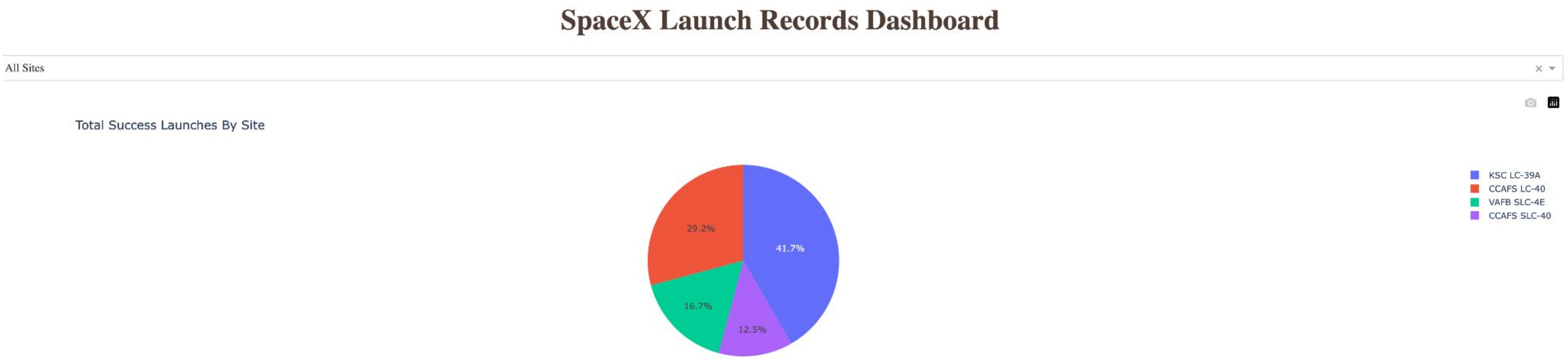
Launch sites surroundings

Launch sites try to be relatively far from city centers (probably by safety) while being close to coastlines, and still maintain an access via railway and highway for supplies and staff to access the sites.



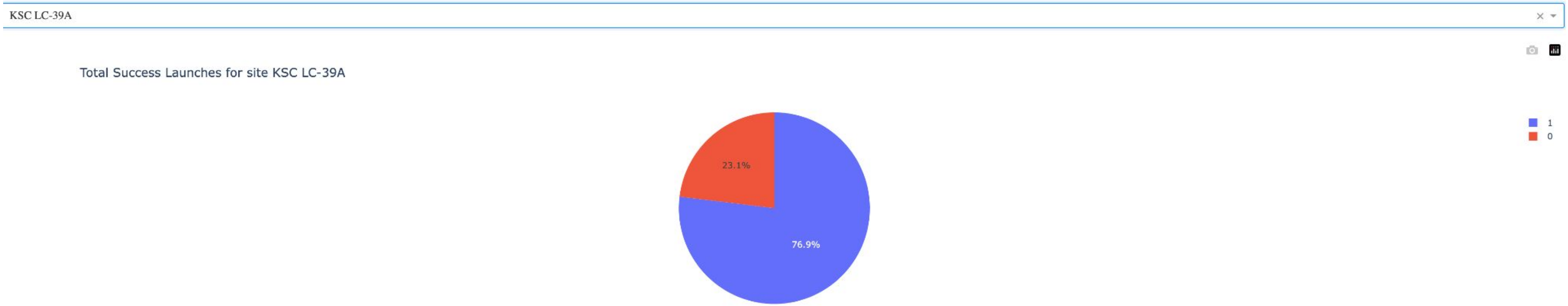
Build a dashboard with Plotly Dash

Launch success count for all sites



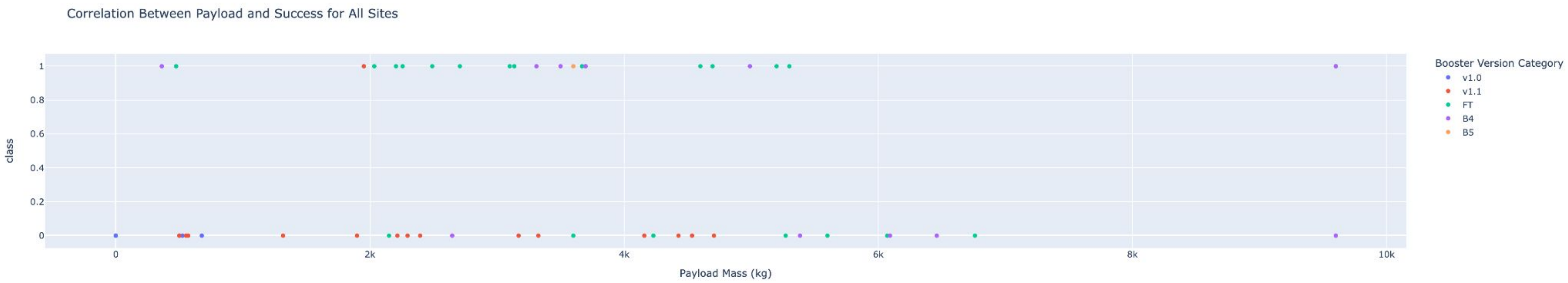
KSC LC-39A represents 41.7% of the total successful launches.

Launch site with highest launch success ratio



KSC LC-39A has almost 77% of success rate.

Correlation between Payload and success for all sites



Most of the successes come from a payload mass between 2000 kg and 5500kg.

Predictive analysis

Models performance

Logistic regression, support vector machine, decision tree, and K nearest neighbours have the **same accuracy**: **0.833**.

In [26]:
```
lr_score = logreg_cv.score(X_test, Y_test)
print("score :", lr_score)
```
```
score : 0.8333333333333334
```

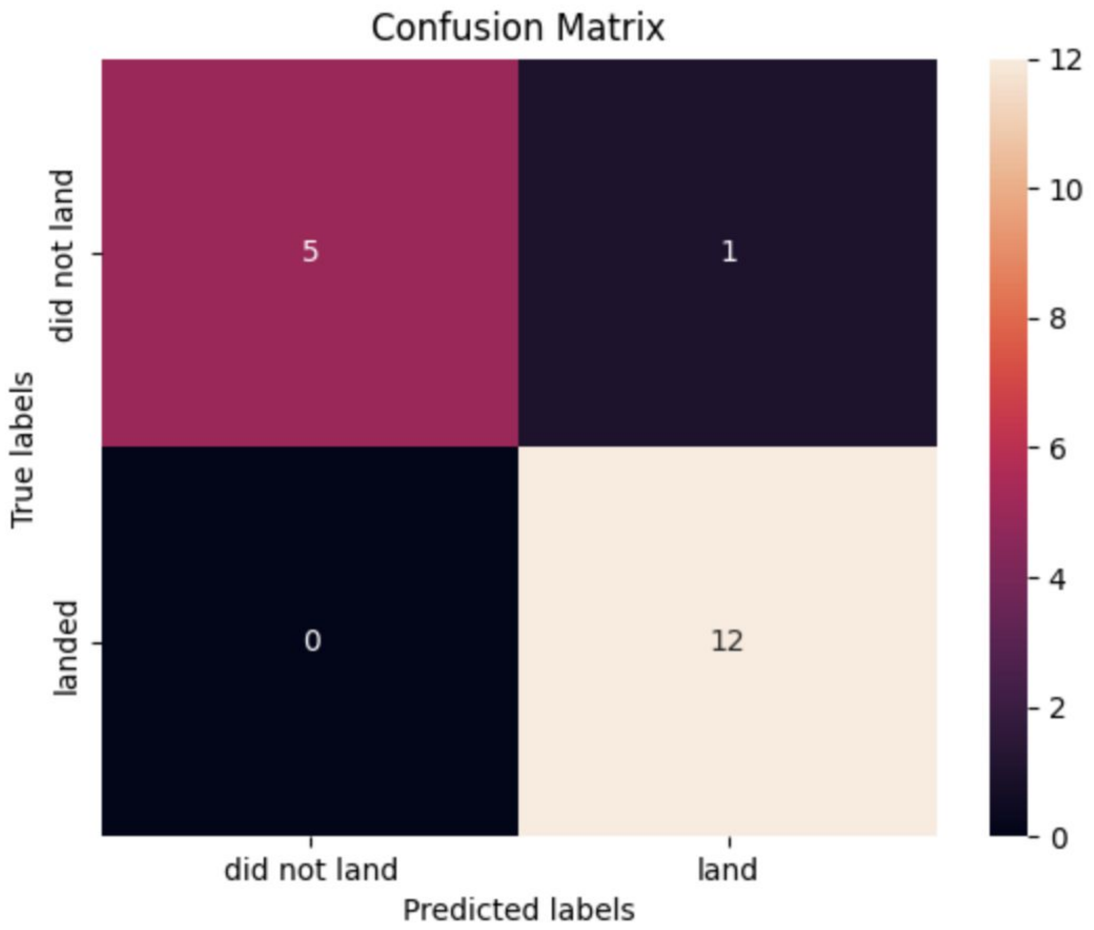
In [33]:
```
svm_cv_score = svm_cv.score(X_test, Y_test)
print('score :',svm_cv_score)
```
```
score : 0.8333333333333334
```

In [46]:
```
tree_cv_score = svm_cv.score(X_test, Y_test)
print("score :",tree_cv_score)
```
```
score : 0.8333333333333334
```

In [54]:
```
knn_cv_score = knn_cv.score(X_test, Y_test)
print('knn score: ', knn_cv_score)
```
```
knn score:  0.8333333333333334
```

Confusion matrix

The **decision tree classifier** has only one false positive (1 “didn’t land” from the true labels counted as “landed” from the predicted label, while the other models have 3 false positive) and appears to be the **best performing model**.



Conclusion

Space X has completely disrupted the aerospace industry making space more accessible from an economic point of view to various stakeholders will now being able to reduce their costs. The company has now set a new norm in terms of launch boosters.

But the analysis of data helped us know that several other parameters such as the orbit and payload mass are also important parameters to consider in order to have the maximum success rate and hence decrease costs significantly.