IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Leonardo Olivastri
11-01-2024

Outline

Executive Summary

- Summary of methodologies
  - Data Collection through API & Web Scraping with BeautifulSoup
  - Data Wrangling with Numpy, Pandas & Matplotlib
  - EDA with SQL, Numpy, Pandas & Matplotlib
  - Visual Analysis with Folium and Dash
  - Machine Learning Predictions using Classification Models
- Summary of all results
  - Launch success has improved a lot since 2013
  - KSC LC-39A has the highest success rate of all the launch sites
  - Orbits GEO, HEO, SSO and ES-L1 have a success rate of 100%

Introduction

- Project background and context

The commercial space age is here: companies are making space travel affordable for everyone, and Perhaps the most successful is SpaceX. SpaceX's accomplishments include:

- Sending spacecraft to the International Space Station
- Starlink, a satellite internet constellation providing satellite Internet access
- Sending manned missions to Space.

One reason SpaceX can do this is that the rocket launches are relatively inexpensive: SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. Other providers cost upwards of 165 million dollars each.

Much of the savings is because SpaceX can sometimes reuse the first stage: therefore, if we can determine if the first stage will land, we can determine the cost of a launch. In fact, sometimes the first stage does not land: sometimes it will crash and other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Space Y would like to compete with SpaceX.

Introduction

- Problems we want to find answers to

The problem is to determine the price of each launch.

We did this by gathering information about Space X and creating dashboards for the team. The main problem is to determine if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, we'll train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

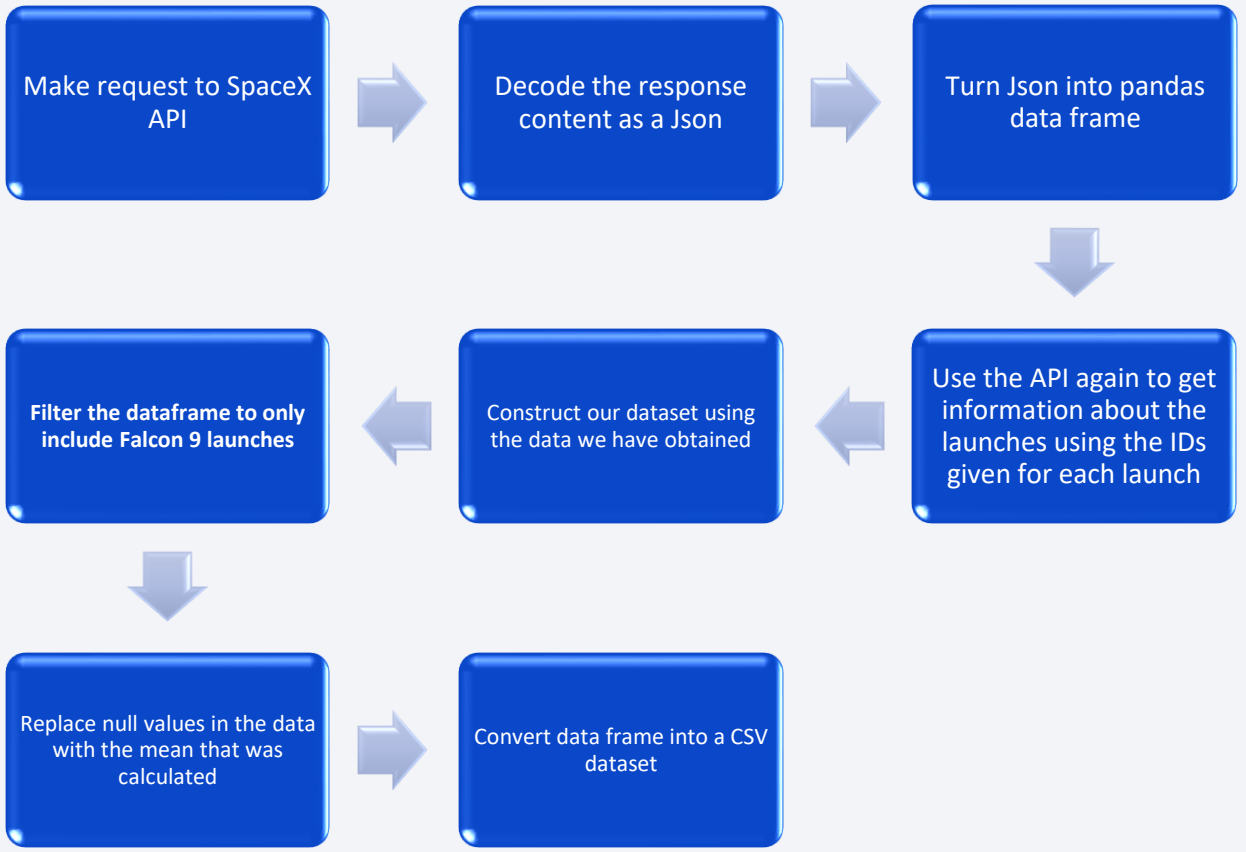
Methodology

Methodology

Executive Summary

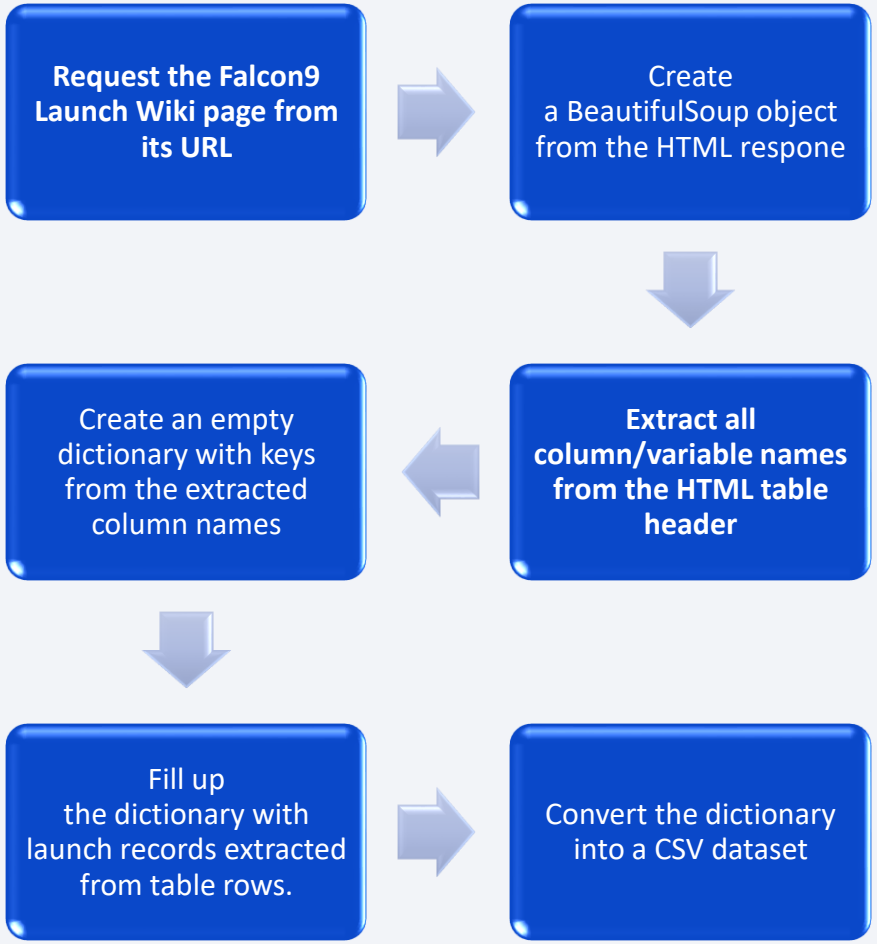
- Data collection methodology:
  - Through SpaceX REST API & BeautifulSoup
- Perform data wrangling
  - Filtering the data, handling missing values and applying one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, K-Nearest Neighbors, Decision Tree & SVM

Data Collection Flow – SpaceX API



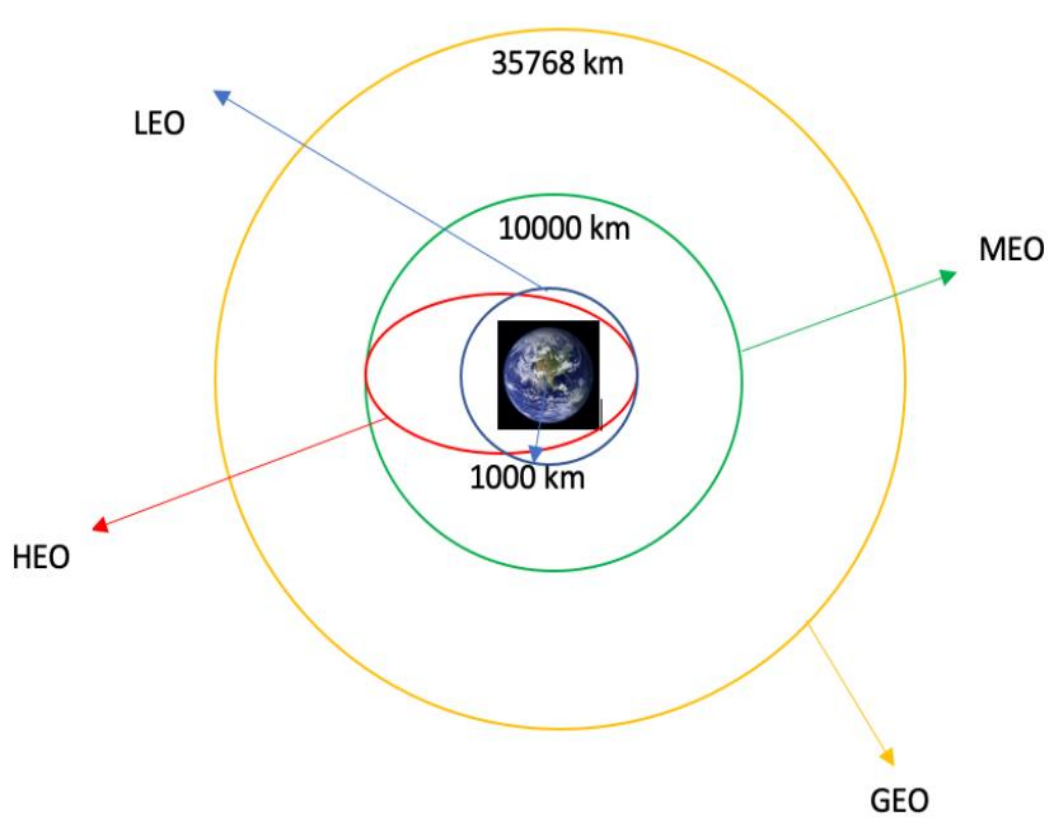
https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/1.Data_Collection_Through_API.ipynb

Data Collection Flow - Scraping



https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/2.Webscraping_With_BeautifulSoup.ipynb

Data Wrangling



The goal in this stage is to find patterns in the data and determine the label for training supervised machine learning models.

In the data set, there are several different cases where the rocket did not land successfully and they may be influenced by a few factors, such as the orbit, the payload mass and so on. For example, *True RTLS* means the rocket successfully landed on a ground pad while *False RTLS* means the rocket unsuccessfully landed on a ground pad.

Those outcomes were converted into Training Labels whereby *1* means the rocket landed successfully while *0* means it was unsuccessful.

https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/3.Data_Wrangling.ipynb

EDA with Data Visualization

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type

- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
    - Mark all launch sites on a map
    - Mark the succeeded launches and failed launches for each site on the map
    - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/2.EDA_With_SQL.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/5.EDA_Visualization_with_Python.ipynb

Build an Interactive Map with Folium

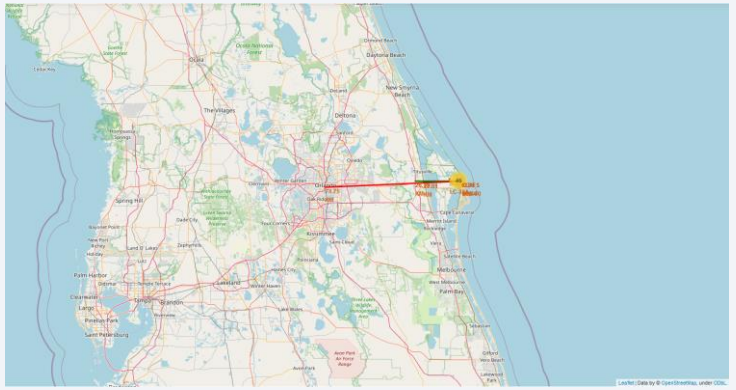
Launch success rate may depend on the location and proximity of a launch site. **Folium Interactive Map was used for visualizing and analyzing SpaceX Launch Sites.**

- Used Interactive mapping library called Folium
- Identified all SpaceX launch sites on a map: Florida, California
- Included longitude and latitude info.
- Identified successful/failed launches for each site on map

Calculated the distance between a launch site (CCAFS_SLC40 in Cape Canaveral, FL) and:

- Closest coastline
- Closest high traffic density railway: Florida East Coast Railway
- Closest high traffic density highway: Interstate I95
- Closest high density urban area: Orlando (FL)

For reference, we added the localization of European Space Agency (ESA) /ArianeEspace Ariane 5 and Soyuz launch pads in Kourou, French Guiana.



CCAFS_SLC40 in Cape Canaveral FL
Coordinates: -80.577°, 28.563°

Ariane launch pad - Kourou in French Guiana
Coordinates: -52.792°, **5.265° (~ Equator)**

https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/6.Maps_with_Folium.ipynb

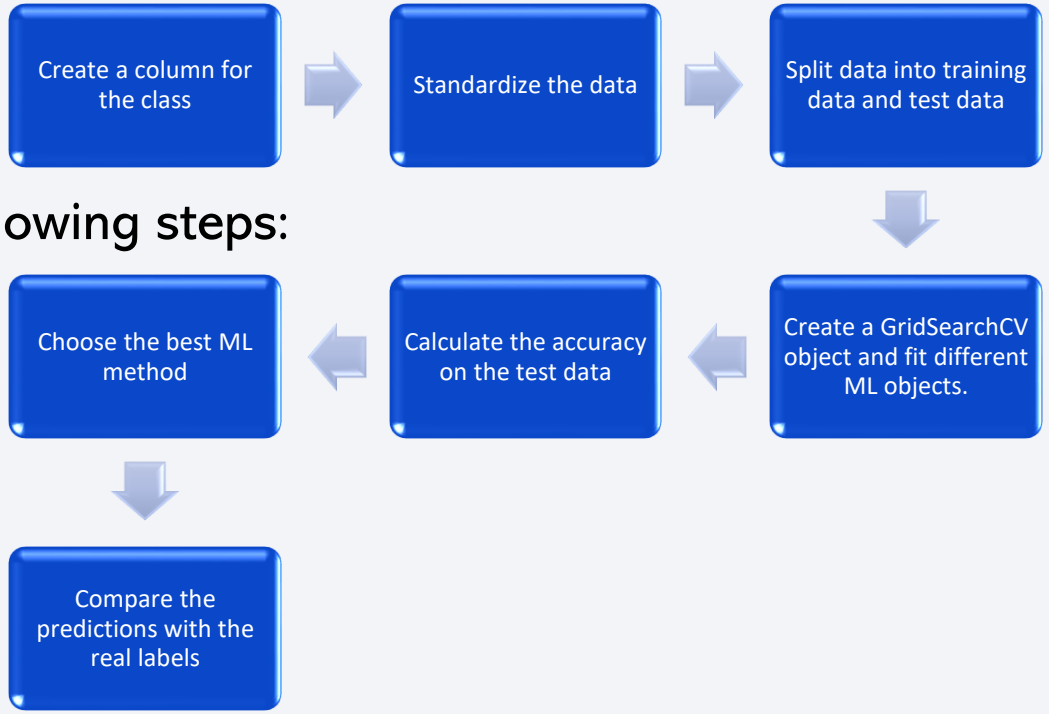
Build a Dashboard with Dash by plotly

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site

https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/7.Dashboard_with_Dash.py

Predictive Analysis (Classification)

- Functions from the Scikit-learn library are used to create our machine learning models.

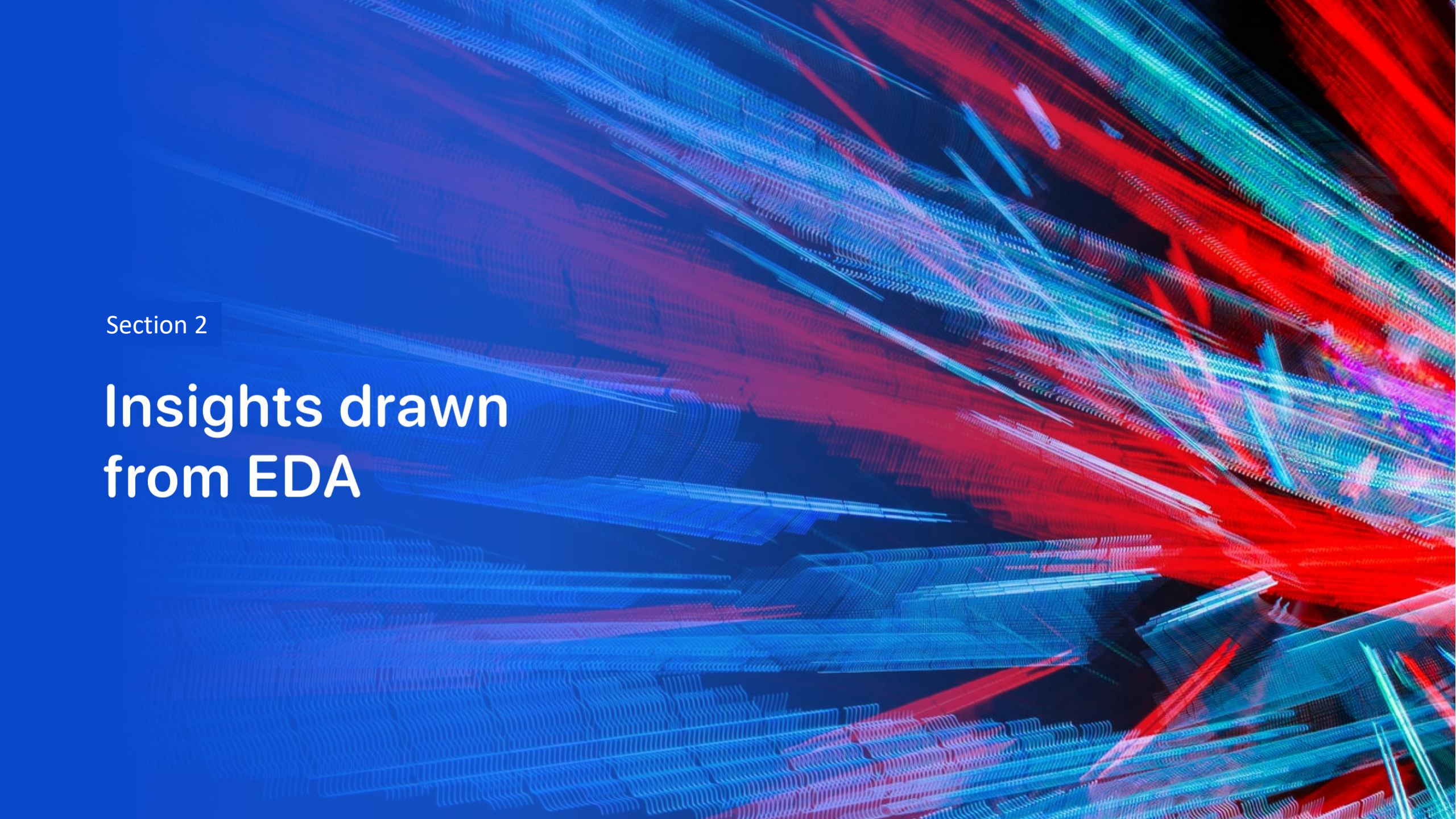


- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Find the best combination of hyperparameters for each model
  - Evaluate the models based on their accuracy scores and confusion matrix

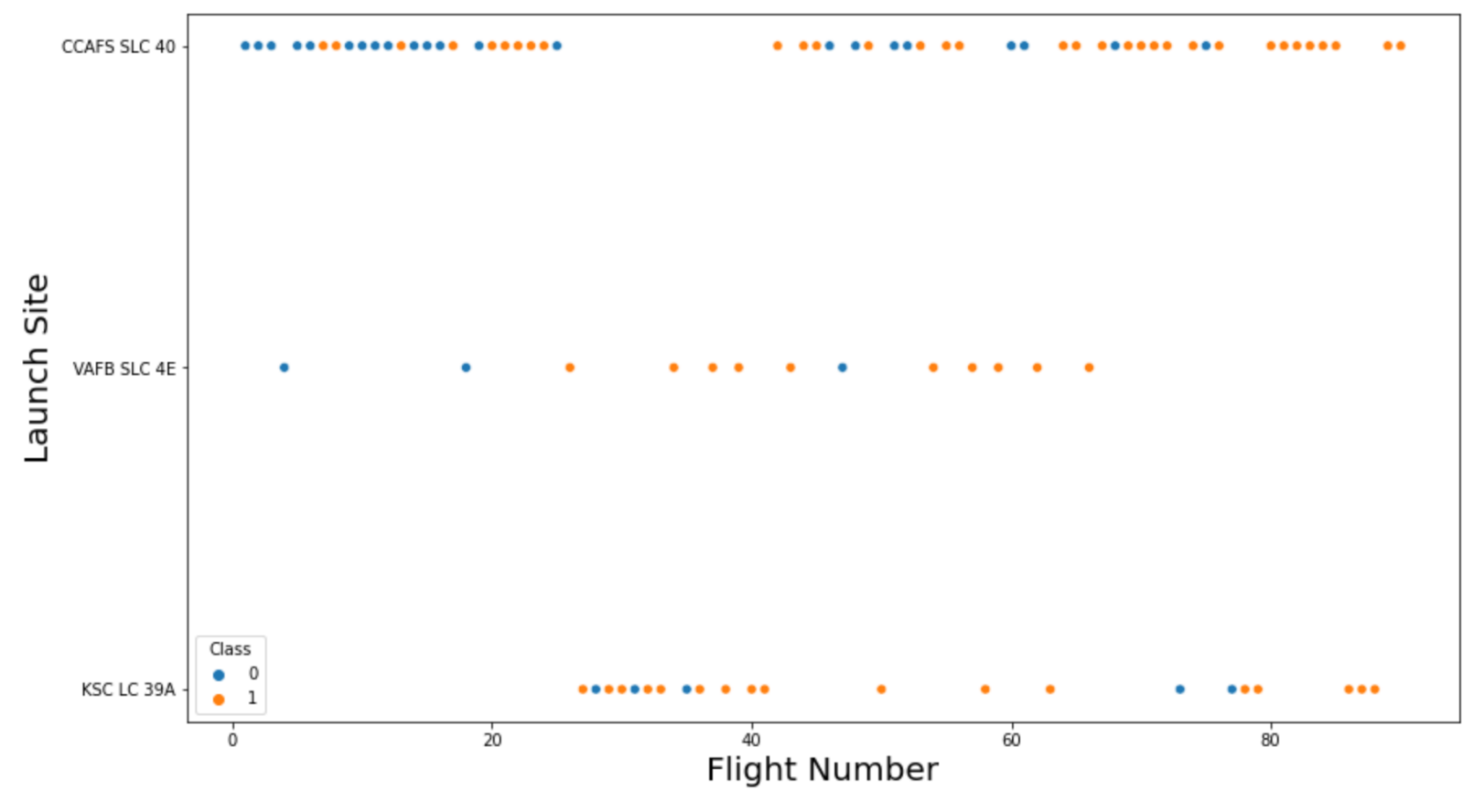
https://github.com/LeonardoOlivastri/Space-Rocket-Landing-Prediction/blob/main/8.Machine_Learning_Prediction.ipynb

Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%. This accuracy can be increased in future projects with more data.

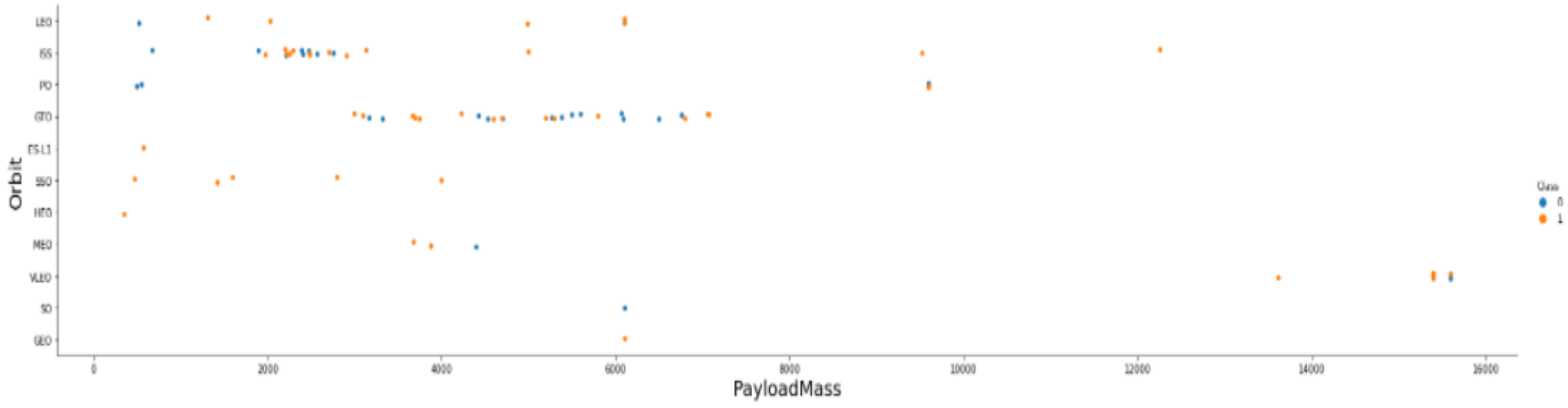

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



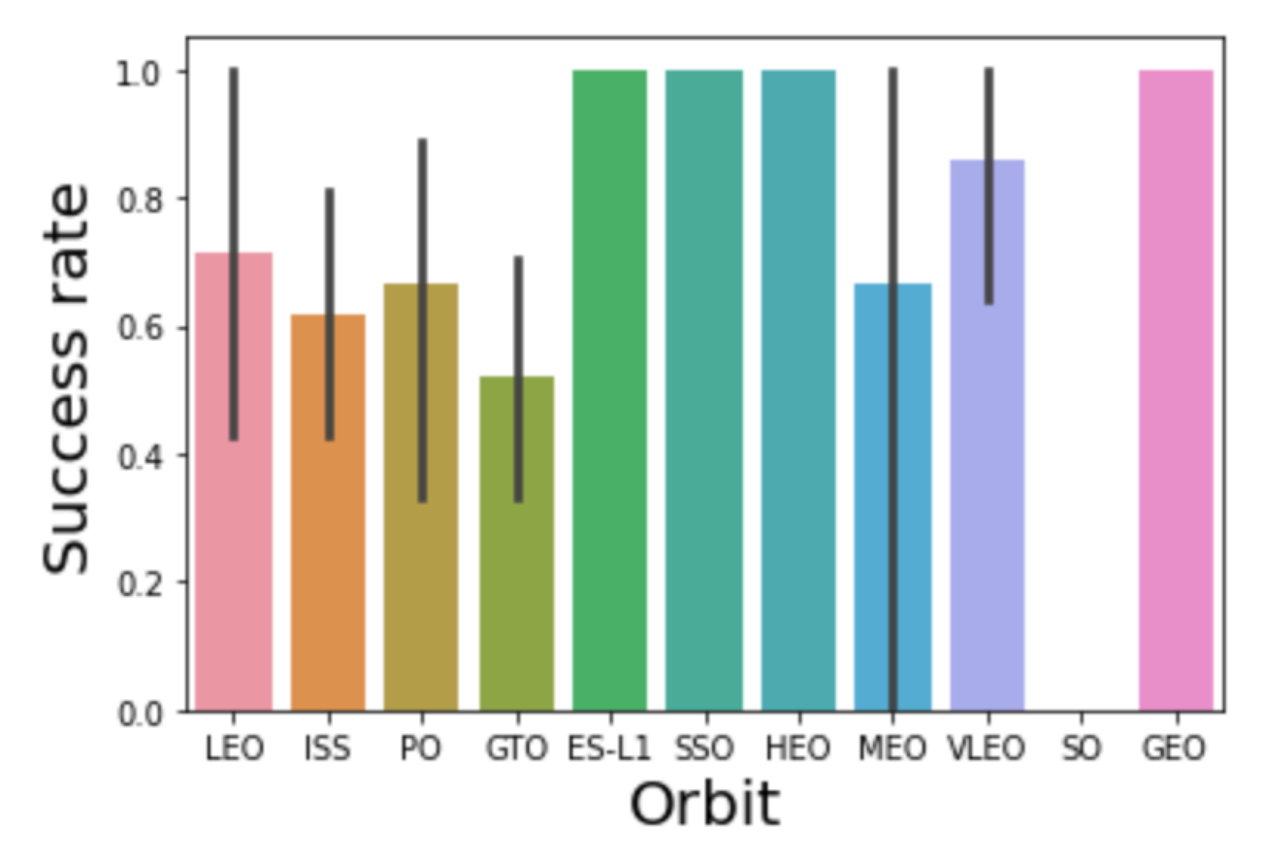
It appears that there were more successful landings as the flight numbers increased. It also seems that launch site **CCAFS SLC 40** had the greatest number of landing attempts while the site **VAFB SLC 4E** had the least number of attempts.

Payload vs. Launch Site



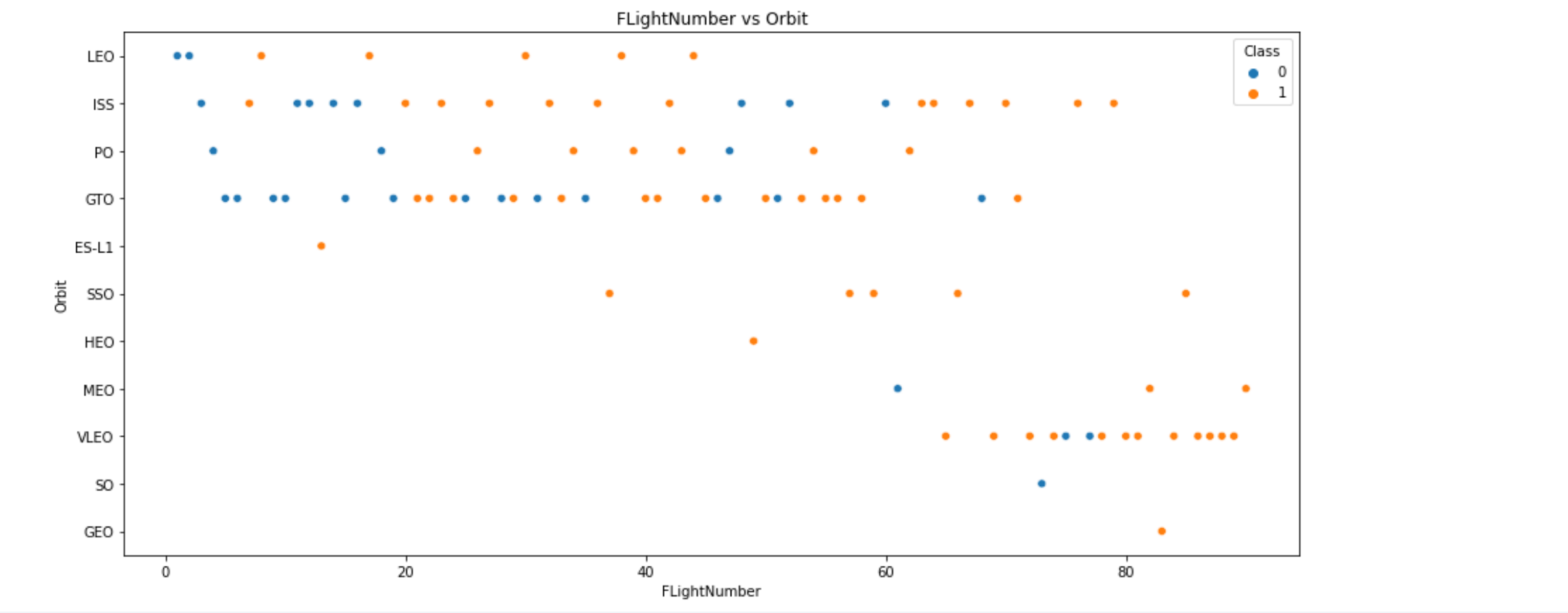
- Given Falcon9 specifications, heavy payloads > 10000 kg are sent to low/medium orbits LEO/MEO only.
- It looks like the percentage of failures is lower for heavy payload. Which would indicate that low orbits are less risky to the success of the mission (recovery of booster).
- Light payloads are not necessarily all sent to GTO/GEO.
- More information is needed for extracting some correlation: success rate v. payload/orbit

Success Rate vs. Orbit Type



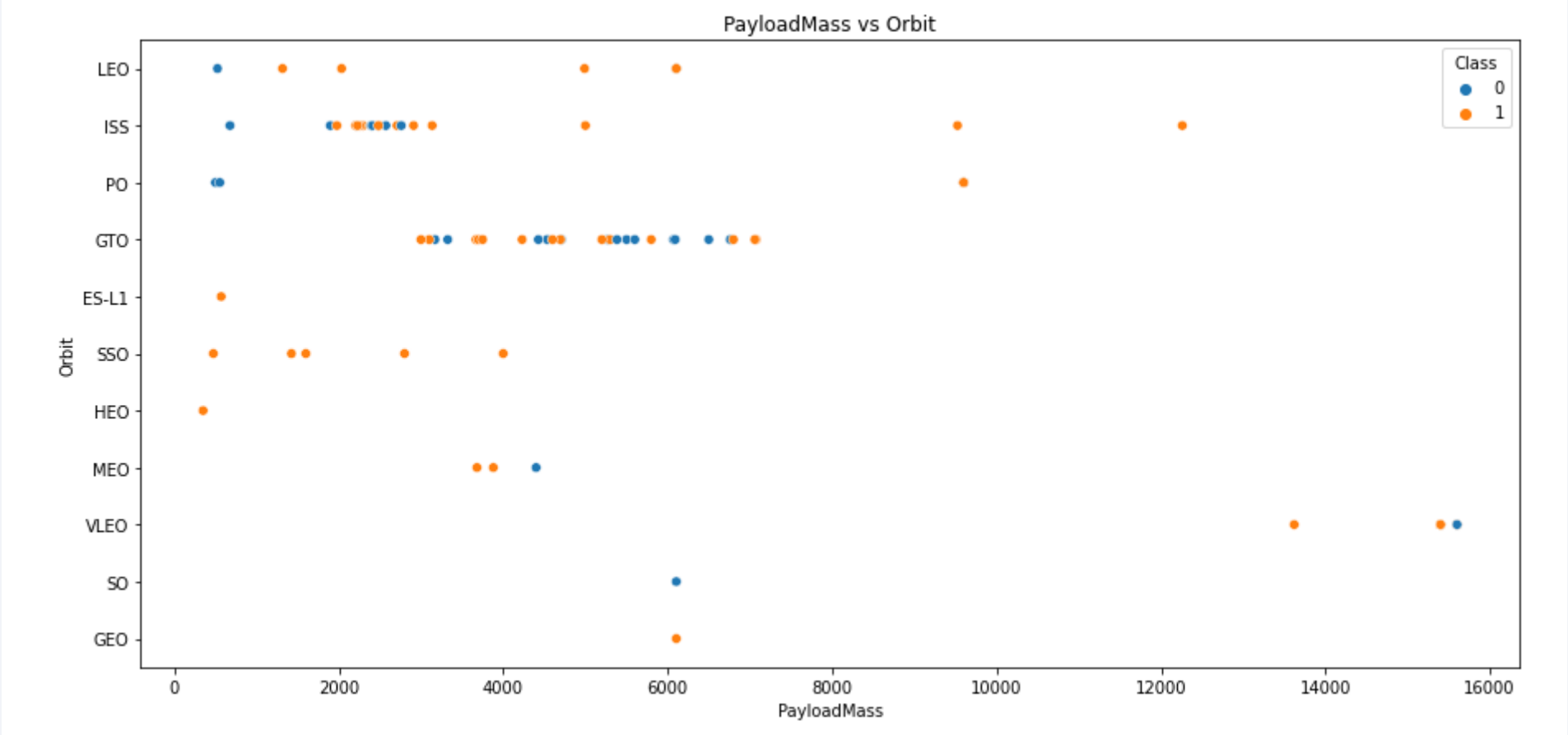
The orbit types **SSO**, **HEO**, **GEO** and **ES-L1** had the highest success rate (100%).

Flight Number vs. Orbit Type



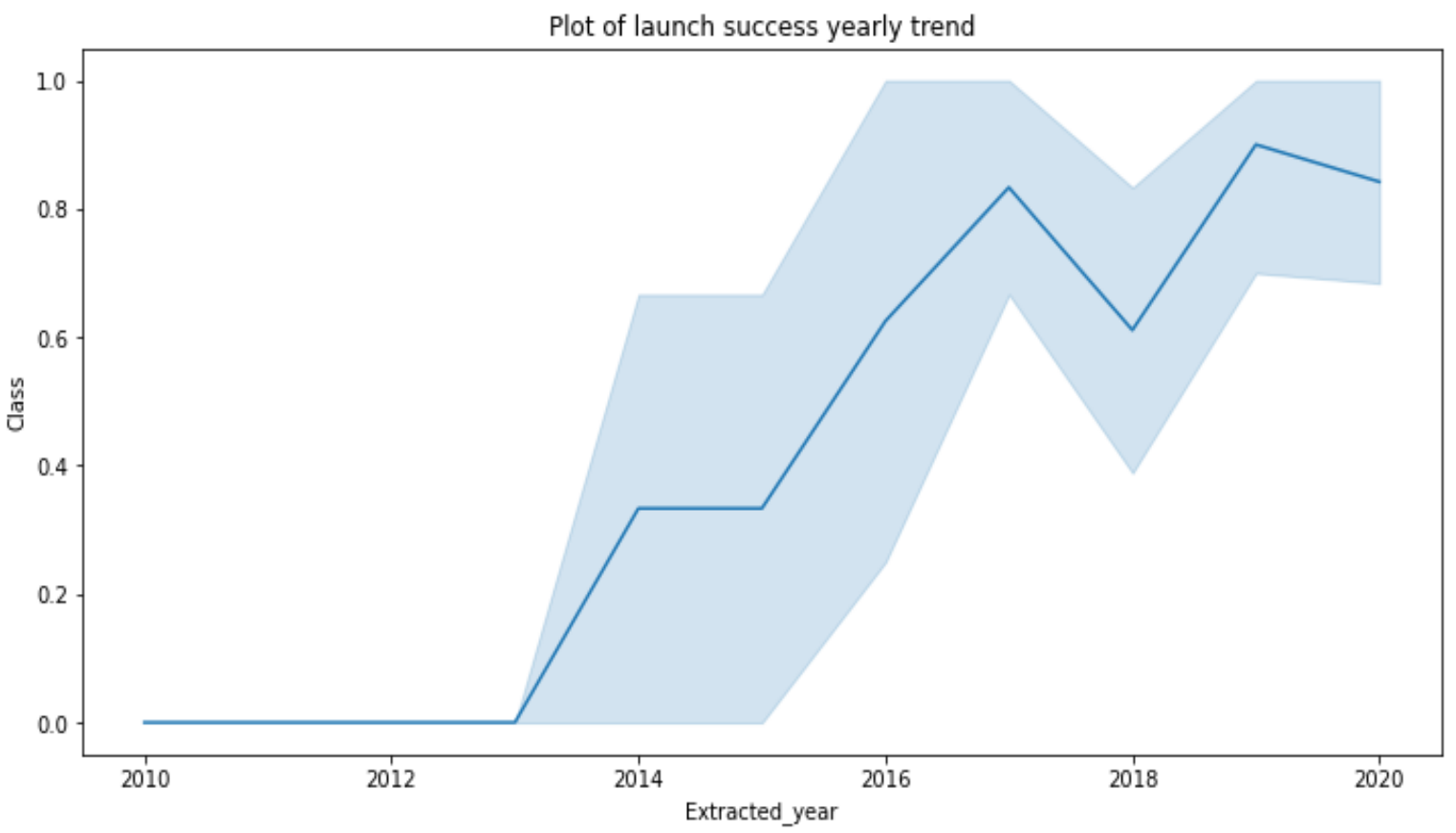
You can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend



From the plot, we can observe that success rate since 2013 kept on increasing till 2020.

All Launch Site Names

```
%%sql

SELECT DISTINCT(Launch_Site)
FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

```
%%sql

SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload mass by NASA (CRS)"
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)"
```

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

Total Payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

```
%%sql

SELECT ROUND(AVG(PAYLOAD_MASS__KG_),2) AS "Average Payload mass carried by booster version F9 v1.1"
FROM SPACEXTABLE
WHERE Booster_Version LIKE "F9 v1.1%"
```

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

Average Payload mass carried by booster version F9 v1.1
2534.67

First Successful Ground Landing Date

```
%%sql

SELECT MIN(Date)
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success%'
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select * from spacextable limit 1
```

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql

SELECT
    (SELECT COUNT(*) FROM spacextable WHERE landing_outcome LIKE 'Success%') AS Success,
    (SELECT COUNT(*) FROM spacextable WHERE landing_outcome LIKE 'Failure%') AS Failure
FROM SPACEXTABLE
LIMIT 1
```

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

Success	Failure
61	10

Boosters Carried Maximum Payload

```
%%sql

SELECT DISTINCT(Booster_Version)
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
ORDER BY Booster_Version
```

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

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

2015 Launch Records

```
%%sql

SELECT SUBSTR(Date, 6, 2) AS Month, Booster_Version, Launch_Site, COUNT(*) AS Num_of_Failures
FROM SPACEXTABLE
GROUP BY Month, Booster_Version, Launch_Site
HAVING SUBSTR(Date, 0, 5)='2015' AND Landing_outcome LIKE "Failure (Drone Ship)"
```

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

Month	Booster_Version	Launch_Site	Num_of_Failures
01	F9 v1.1 B1012	CCAFS LC-40	1
04	F9 v1.1 B1015	CCAFS LC-40	1

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

```
%%sql

SELECT Landing_Outcome, COUNT(*)
FROM SPACEXTABLE
GROUP BY Landing_Outcome
HAVING Date BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY Landing_Outcome DESC
```

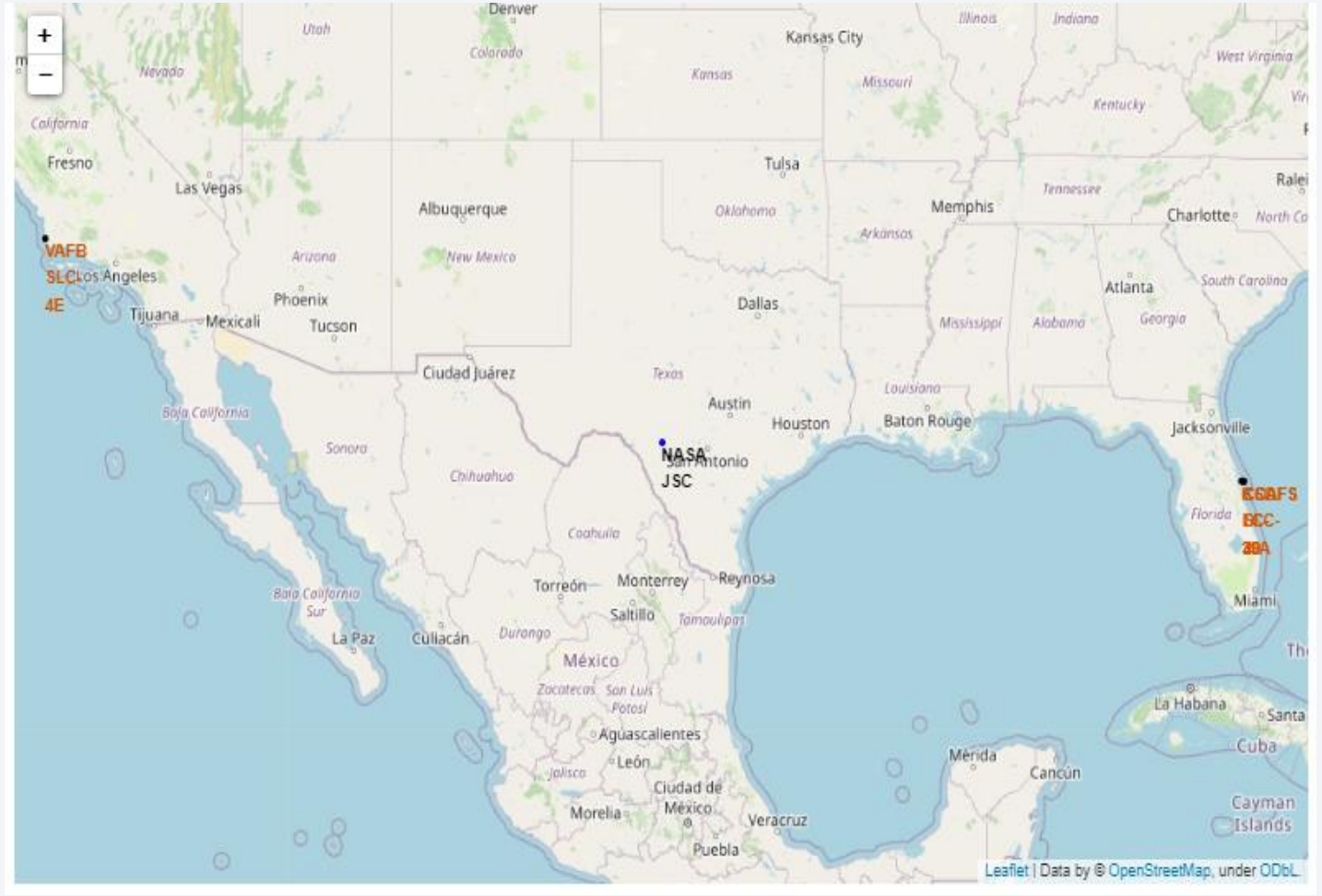
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	COUNT(*)
Uncontrolled (ocean)	2
Success (ground pad)	9
Success (drone ship)	14
Precluded (drone ship)	1
No attempt	21
Failure (parachute)	2
Failure (drone ship)	5
Controlled (ocean)	5

Section 3

Launch Sites Proximities Analysis

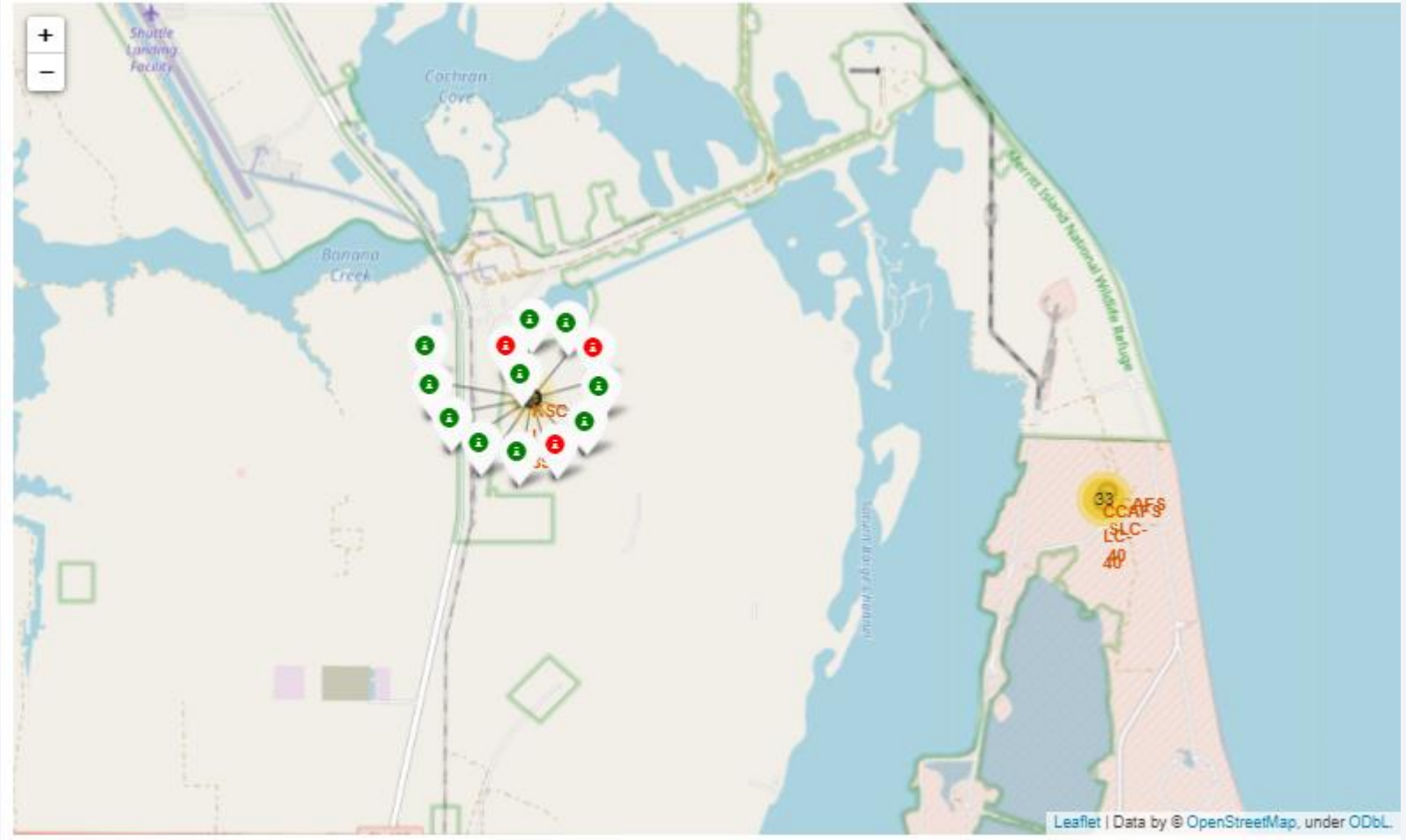
Launch Sites Location



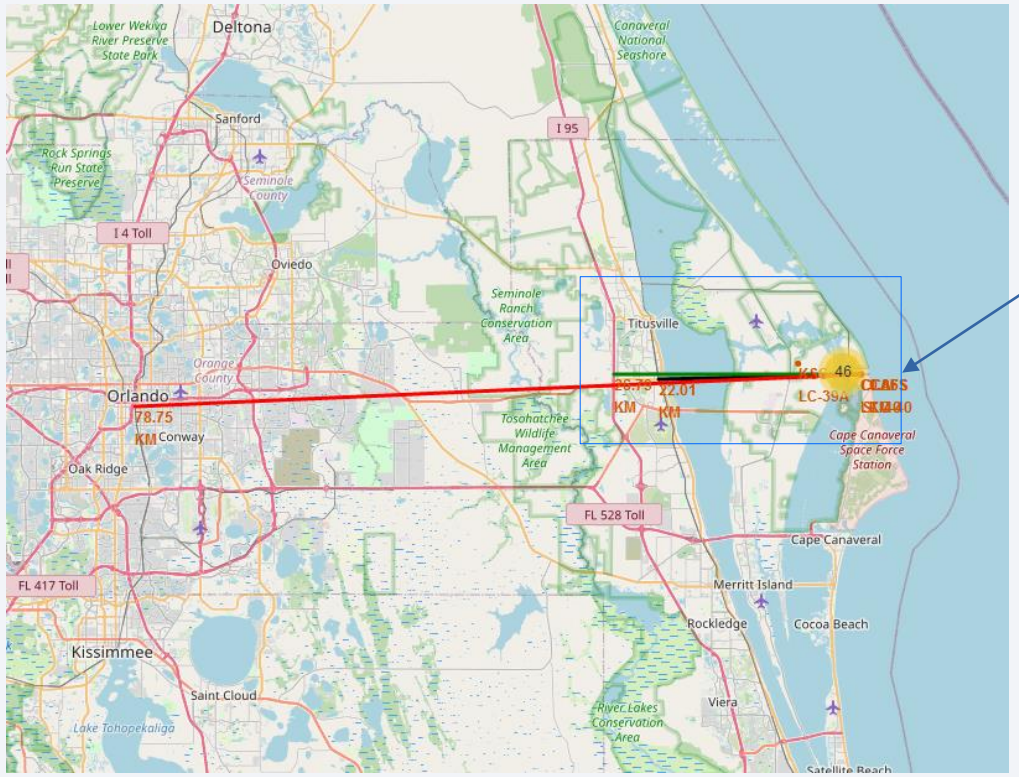
We can see that all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.

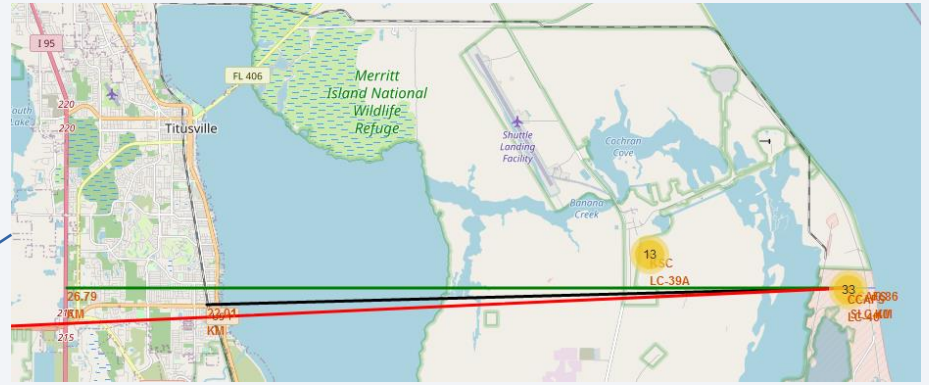
Success Rate of KSC LC 39-A

- The successful launches are represented by a green marker while the red marker represents failed rocket launches.
- It appears that **KSC LC-39A** had the highest success rate of rocket launches compared to other launch sites.



Distances between Launch Sites to surrounding points of interest





Distance from CCAFS_SLC40 to:

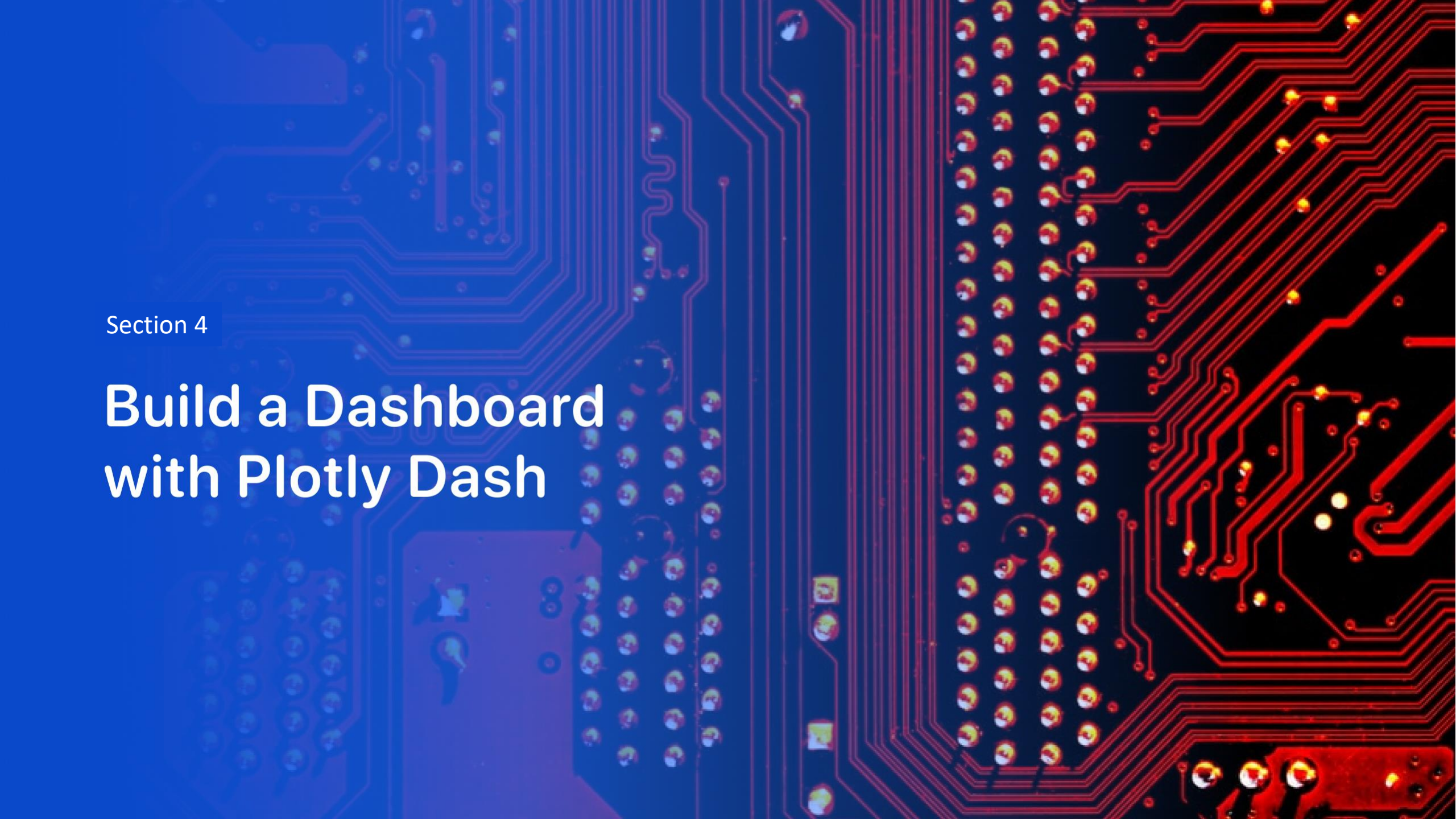
- **Closest coast: ~900 m**
- **Florida East Coast Railway: 22.0 km**
- **Highway I 95: 26.8 km**
- **Orlando: 78.75 km**

Launch sites are close to coasts. For safety issues if launcher is lost in the early stage of the flight.

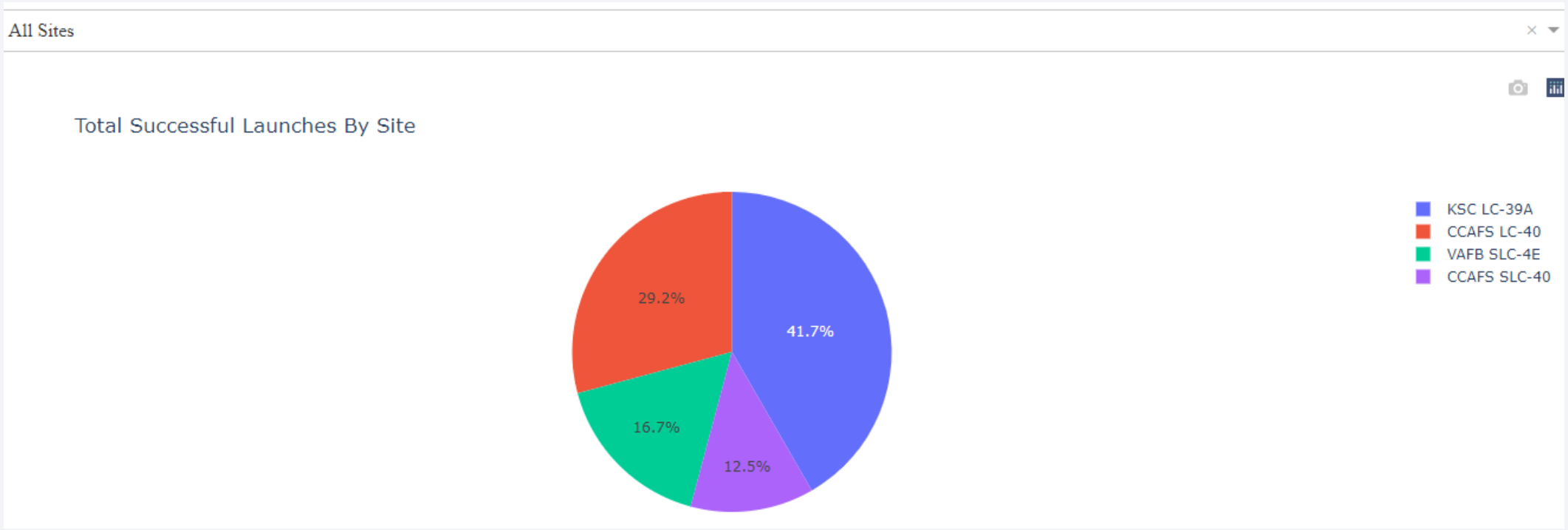
Rockets are launched:

- From West to East over the ocean in Florida.
- North or South bound over the ocean in California. (Polar orbits only)

Launch sites are relatively far from populated areas for protecting population from serious incidents at lift off: explosion on the launch pad.


Section 4
Build a Dashboard with Plotly Dash

Falcon 9: Launch Success for all sites

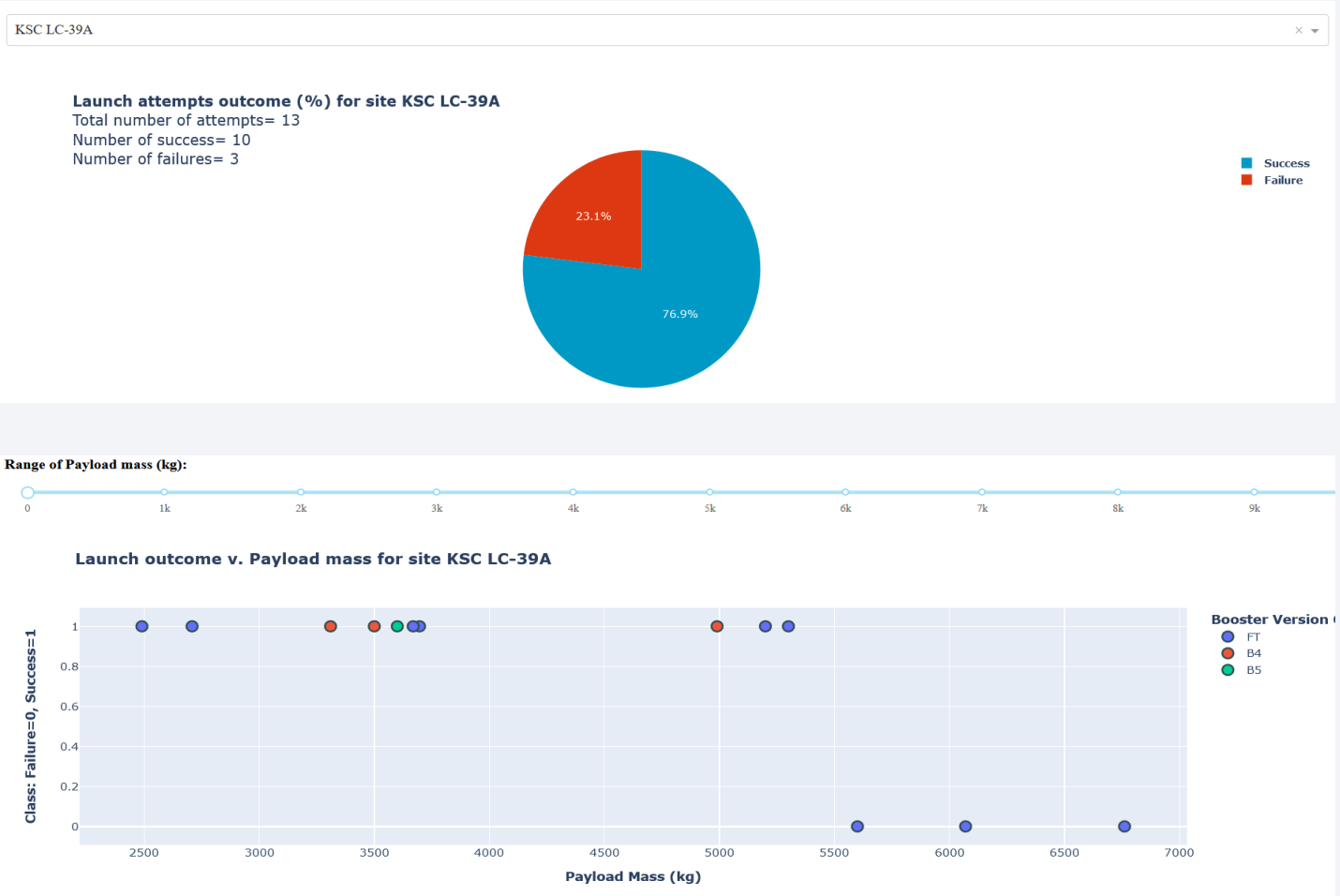


- Site KSC LC-39A has the largest successful launches as well the highest launch success rate.
- More investigation may be needed to determine why KSC LC-39A is the preferred launch site.

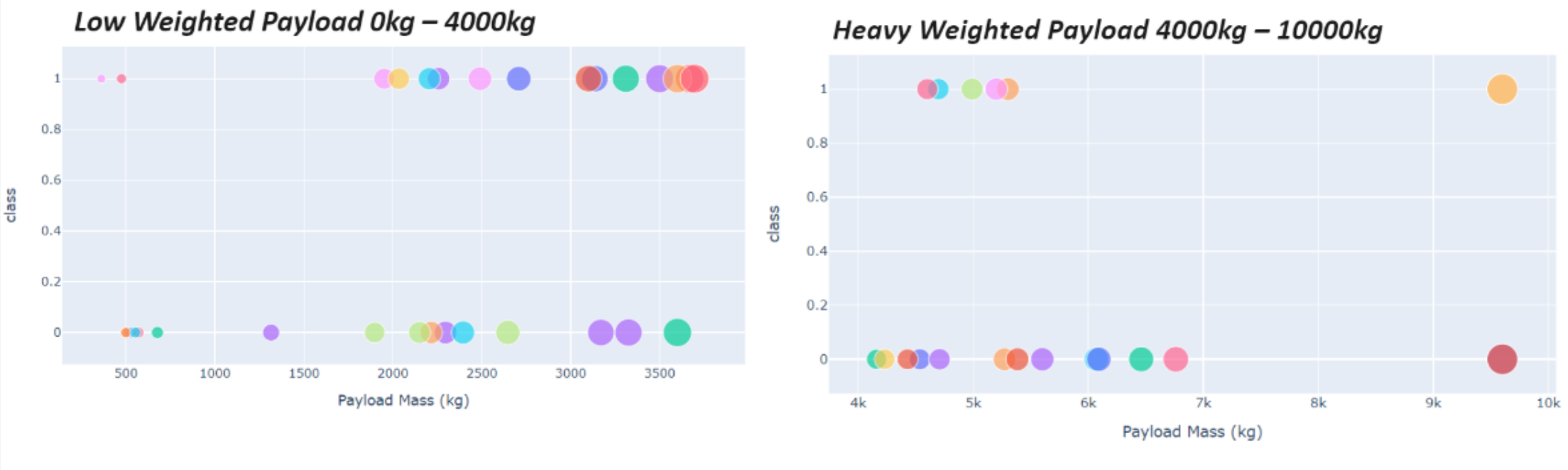
Launch Site with the highest success ratio

KSC LC-39A
Kennedy Space Center in Florida.
13 flights, 10 successful missions.

- Heavy payload are “high risk”
- Success does not seem to depend upon boosters versions with low mass payload <5500kg.
- B5 and FT are the most reused launchers. Data is not sufficient, but may indicates that they are as reliable as 1 time launchers.



Payload Mass vs Success Rate for all sites



- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was also very low between the payload range of 0kg and 2500kg. Perhaps very low masses decrease launch success

Section 5

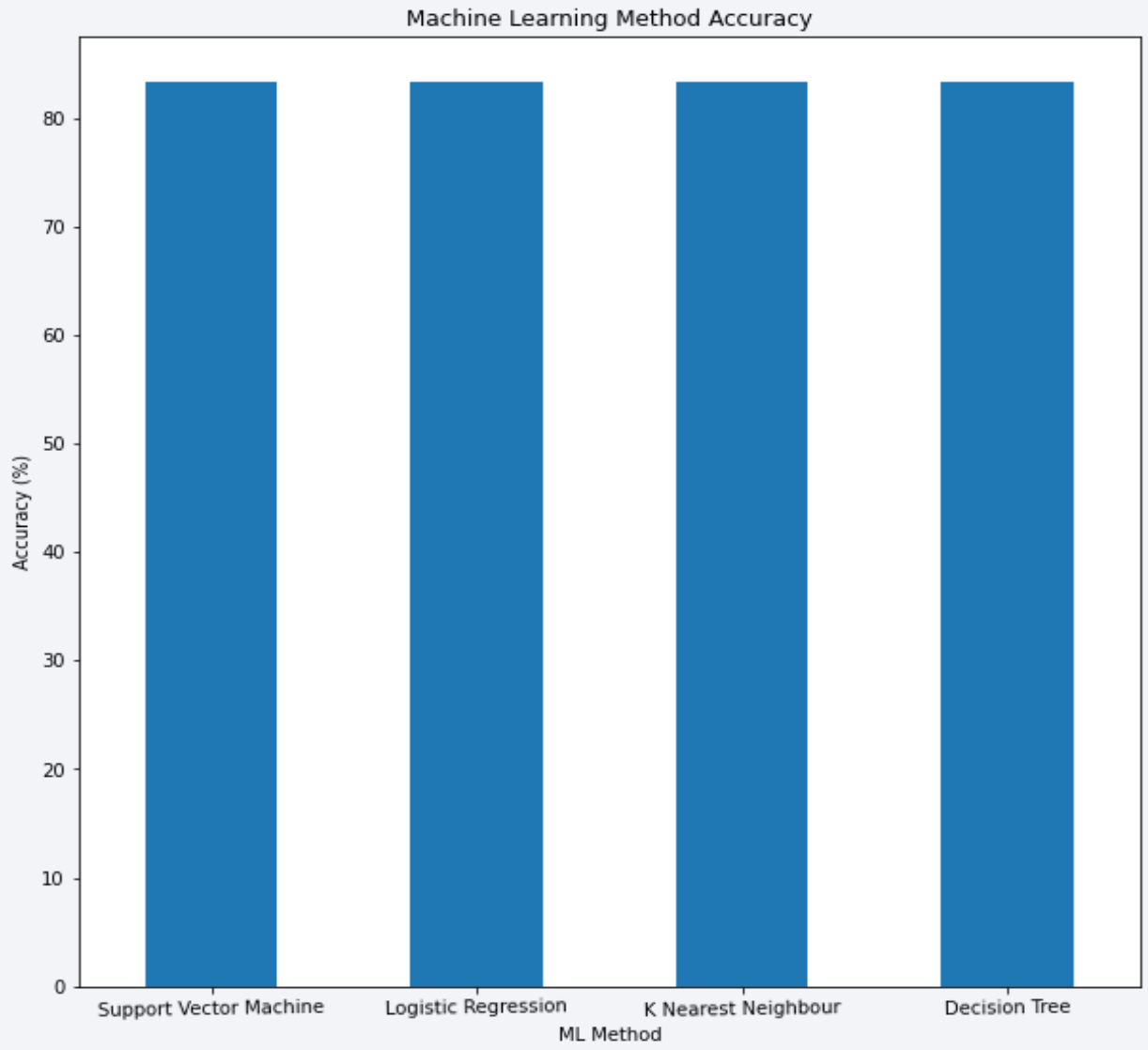
Predictive Analysis (Classification)

Classification Accuracy

For the classification we used 4 models:

- SVM
- Logistic Regression
- KNN
- Decision Tree

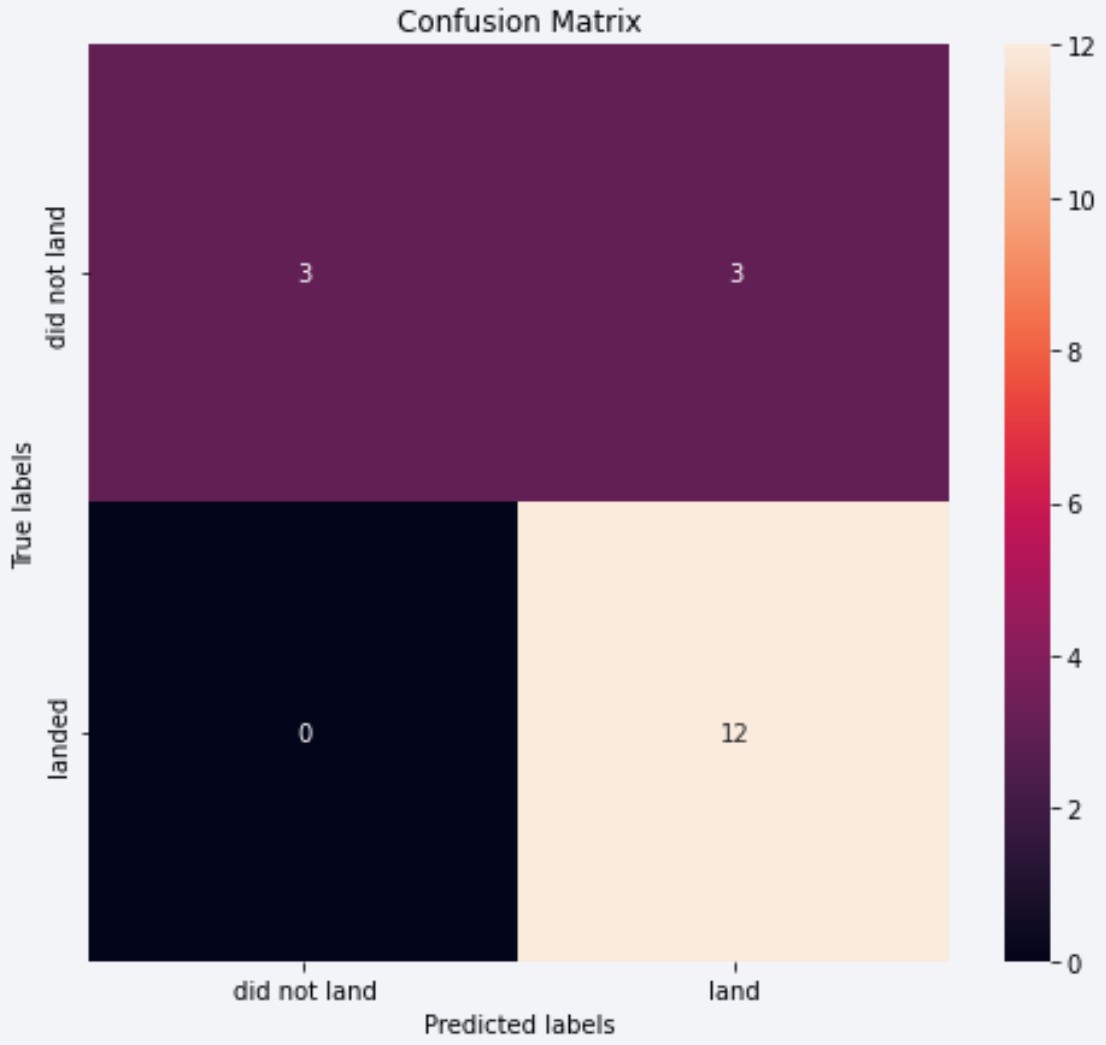
All the methods have an identical accuracy score of 83.33%.



Confusion Matrix

The chart shows the confusion matrix of the Logistic Regression model that was chosen.

The model only failed to accurately predict 3 labels (False Positive).



Conclusions

In order to compete with SpaceX, it was crucial to analyze their data. Through this process, a general picture of their success methods was produced.

- All their launch sites are located near the coast, away from nearby cities. This enabled to them to test their rocket landings without much interference.
- Site KSC LC-39A had the highest launch success rate out of all the launch sites.
- From 2013 onwards, the success rate of rocket landings significantly increased. It was also apparent that landing success increased with flight number

All this data was used to train a machine learning model that is able to predict the landing outcome of rocket launches with 83.33% accuracy.

Thank you!