Applied Data Science Capstone

Predicting Falcon 9 launch outcomes

Assad Khan

2023-09-18

Table of contents

Executive Summary

In this presentation, we will be aiming to predict whether the first stage of the SpaceX Falcon 9 will land successfully or not by using the following methodologies:

Summary of methodologies:

1. Data Collection
2. Data Wrangling
3. EDA (Exploratory Data Analysis)
4. Interactive Data Visualization
5. Machine Learning for Predictive Analysis (Classification models)

Summary of results:

1. EDA (Exploratory Data Analysis) results
2. Data Visualization results
3. Predictive Analysis results

Introduction

Background

SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars; while other providers cost upward of 165 million dollars each, SpaceX can reuse the first stage, this is what enables them to carry out launches at a significantly lower cost. Therefore, if we can determine if the first stage will successfully land, we can determine the cost of a launch. This information could then be used by alternate companies who may choose to bid against SpaceX for a rocket launch.

Open questions:

- How does payload mass, launch site, orbit, and number of flights affect the landing success?
- Do landings become more successful over time?
- Which binary classification model best predicts the outcome of the landing?

Methodology

Executive Summary

1) Data collection

- Collected data via the SpaceX REST API
- Web scraping the Falcon 9 launches page on Wikipedia.

2) Data wrangling

- Data filtering
- Handling missing values
- Applying one hot encoding to prepare the data for binary classification

3) Exploratory data analysis (EDA) & Visualization

- Using SQL to explore the data & answer questions
- Visualize relationships & patterns using Pandas an Matplotlib

4) Interactive data visualization

- Generate an interactive map to analyze the launch site proximity with Folium
- Build an interactive dashboard with Plotly Dash

5) Build models

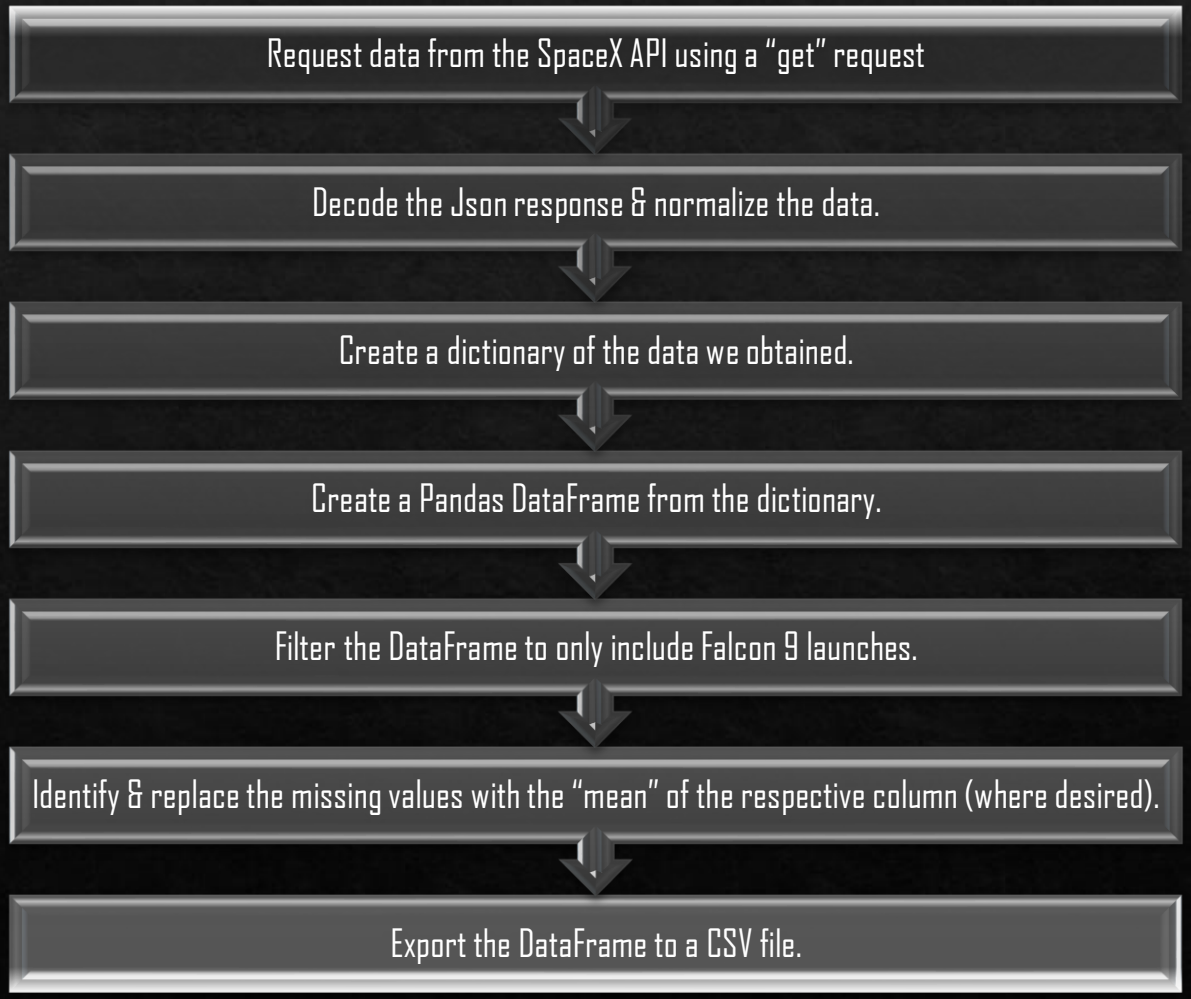
- Build and train different classification models to find the most accurate one to use for our predictions

Data Collection

Section 1

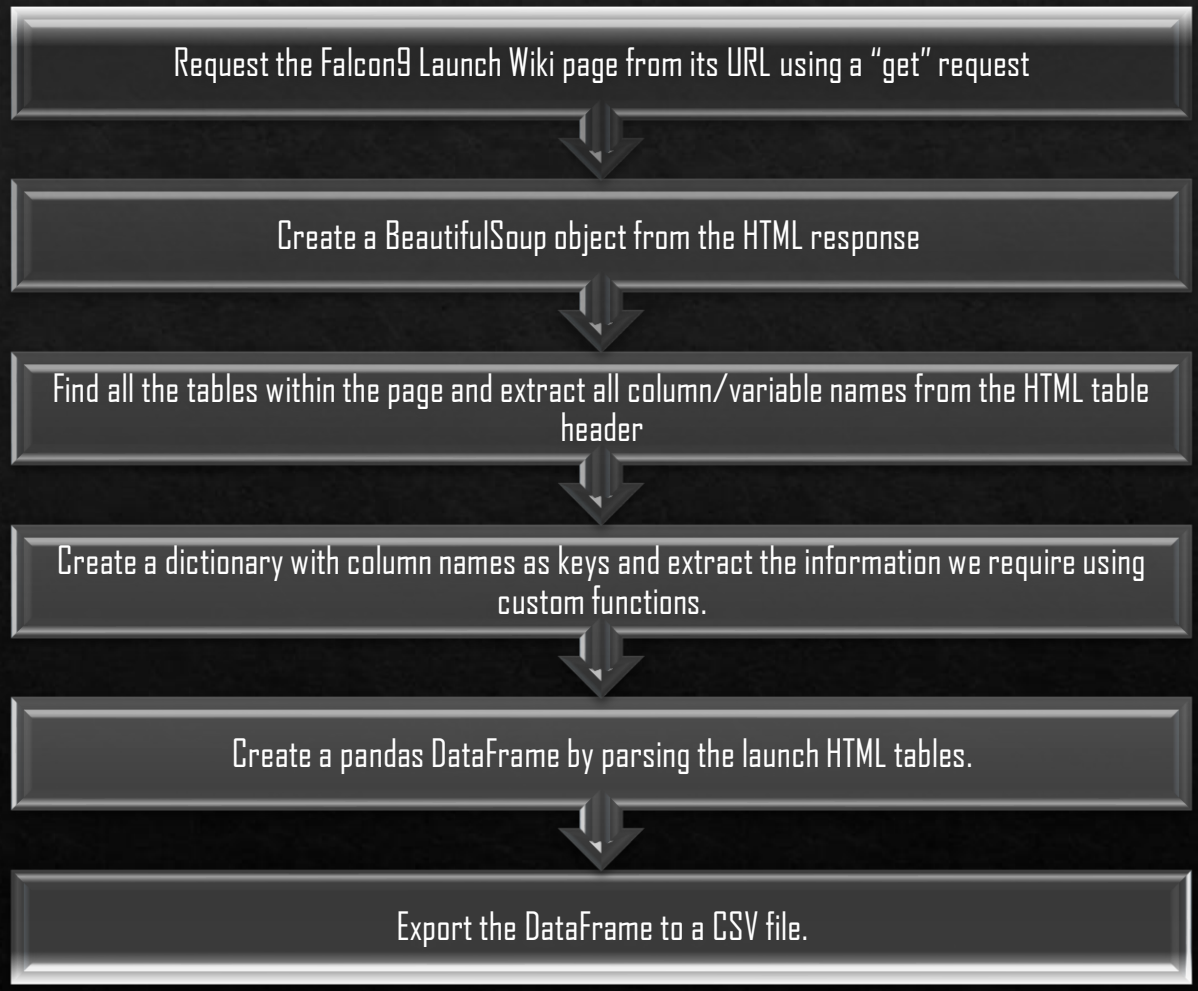
Data Collection – SpaceX REST API

- The flow chart on the right depicts the flow of using the SpaceX API to extract the data we needed for us to move onto the data wrangling stage and transform the data for further use.
- The completed notebook of the entire process used can be found at the following link: Github: Data Collection via API



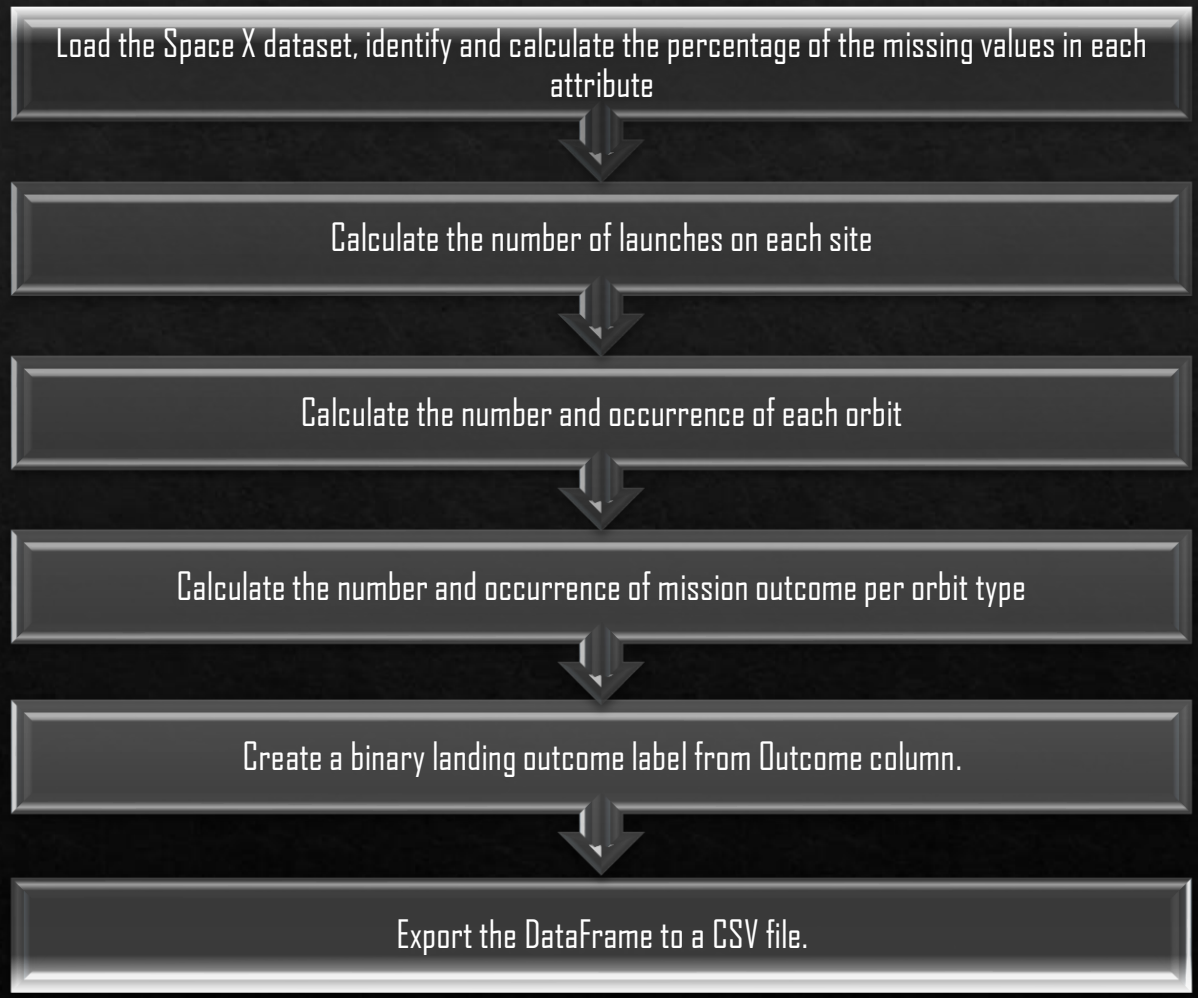
Data Collection – Web Scraping

- The flow chart on the right depicts the flow of using Web scraping to extract the data we needed from the Falcon 9 launch records page on Wikipedia.
- The completed notebook of the entire process used can be found at the following link: GitHub: Web scraping using BeautifulSoup



Data Wrangling

- The objective was to perform exploratory data analysis & determine training labels on the data we have collected so far.
- The flow chart on the right provides the step-by-step process that was used to achieve a result set that allowed us to further analyze and visualize the data.
- The completed notebook of the entire process used can be found at the following link: GitHub: Data wrangling

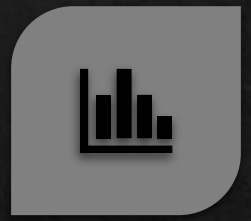


EDA with Data Visualization

LINE CHARTS

A line chart was used to visualize the relationship between:

- The yearly average success rate



BAR CHARTS

A bar chart was used to visualize the relationship between:

- The success rate of each orbit type

SCATTER PLOTS

Scatter plots were used to visualize the relationship between:

- Flight number & launch site
- Flight number & orbit type
- Payload & launch site
- Payload & orbit type

The completed notebook of the entire process used can be found at the following link: GitHub: EDA - Data Visualization

EDA with SQL

SQL queries were executed to gather the following insight from the data:

- 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, using 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 the year of 2015
- 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

The completed notebook of the entire process used can be found at the following link: GitHub: EDA with SQL

Building an Interactive Map with Folium (Geospatial Analysis)

Objectives

Mark all launch sites on the map:

- Added the NASA Johnson Space Centre on the map as the center location of the map via its latitude & longitude coordinates and used folium.Circle to add a highlighted circle area with a text label on the NASA JSC specific coordinate.
- Added markers to all launch sites similarly to NASA JSC, this enables us to see whether the sites are in proximity of the Equator and Coasts.

Mark the successful and failed launches for each site on the map:

- Created a folium MarkerCluster object which allowed the launch sites to be colored Green for successful launches and Red for failed launches
- The allows us to visibly recognize which launch sites have higher success rates

Calculate the distances between a launch site to its proximities:

Plotted distance lines to answer the following questions using folium.PolyLine

- Are launch sites near railways?
- Are launch sites near highways?
- Are launch sites near coastline?
- Do launch sites keep certain distance away from cities?

The completed notebook of the entire process used can be found at the following link: GitHub: Interactive visuals with Folium

Building an interactive dashboard with Plotly Dash

Objectives

Add a dropdown list to enable Launch Site selection:

- Added a dropdown list to enable a user to select a specific launch site.

Add a pie chart to show the total successful launches count for all sites:

- Created a pie chart to allow users to see percentage wise successful/unsuccessful launch rates

Add a slider to select payload range:

- Provided functionality to be able to select payload range

Add a scatter chart to show the correlation between payload and launch success:

- Displayed a scatter plot to allow users to see correlation between payload and launch success with the possibility of filtering the booster version

The completed notebook of the entire process used can be found at the following link: GitHub: Building a dashboard with Plotly Dash

Predictive Analysis (Binary Classification)

MODEL DEVELOPMENT

Perform exploratory Data Analysis and determine Training Labels:

- Create a column for the class
- Standardize the data
- Split into training & test data
- Select algorithm
- Create GridSearchCV object & a dictionary of parameters
- Fit the object to the parameters
- Train the model

MODEL EVALUATION

For each model:

- Apply GridSearchCV
- Check the tuned hyperparameters by using best_params_
- Calculate the accuracy by using the score & best_score_
- Plot and assess the confusion matrix

IDENTIFY BEST MODEL

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression:

- Review model scores
- Select the model that has the best Jaccard_Score, F1_Score & Accuracy.

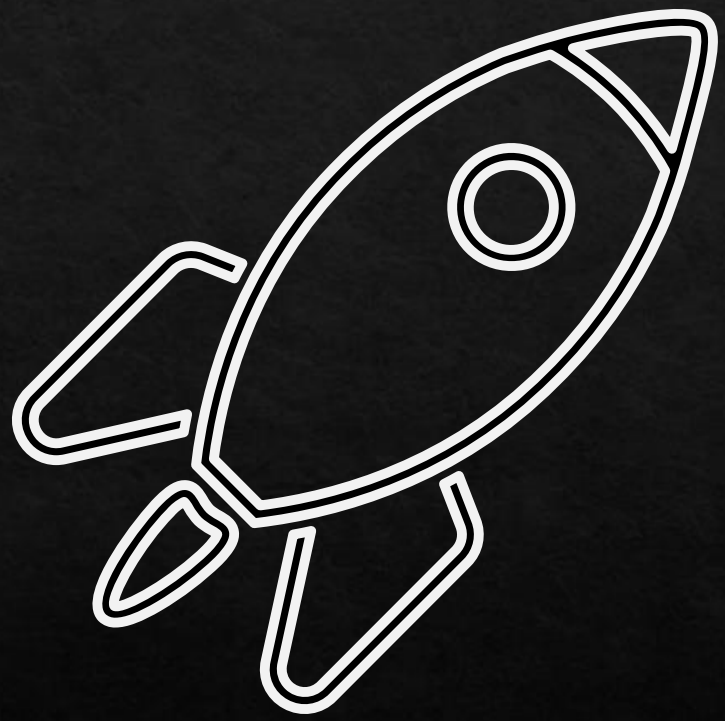
The completed notebook of the entire process used can be found at the following link: [GitHub: Predictive analysis]

Results

Exploratory data analysis results

Interactive visual analytics results demonstrated in screenshots

Predictive analysis results



EDA

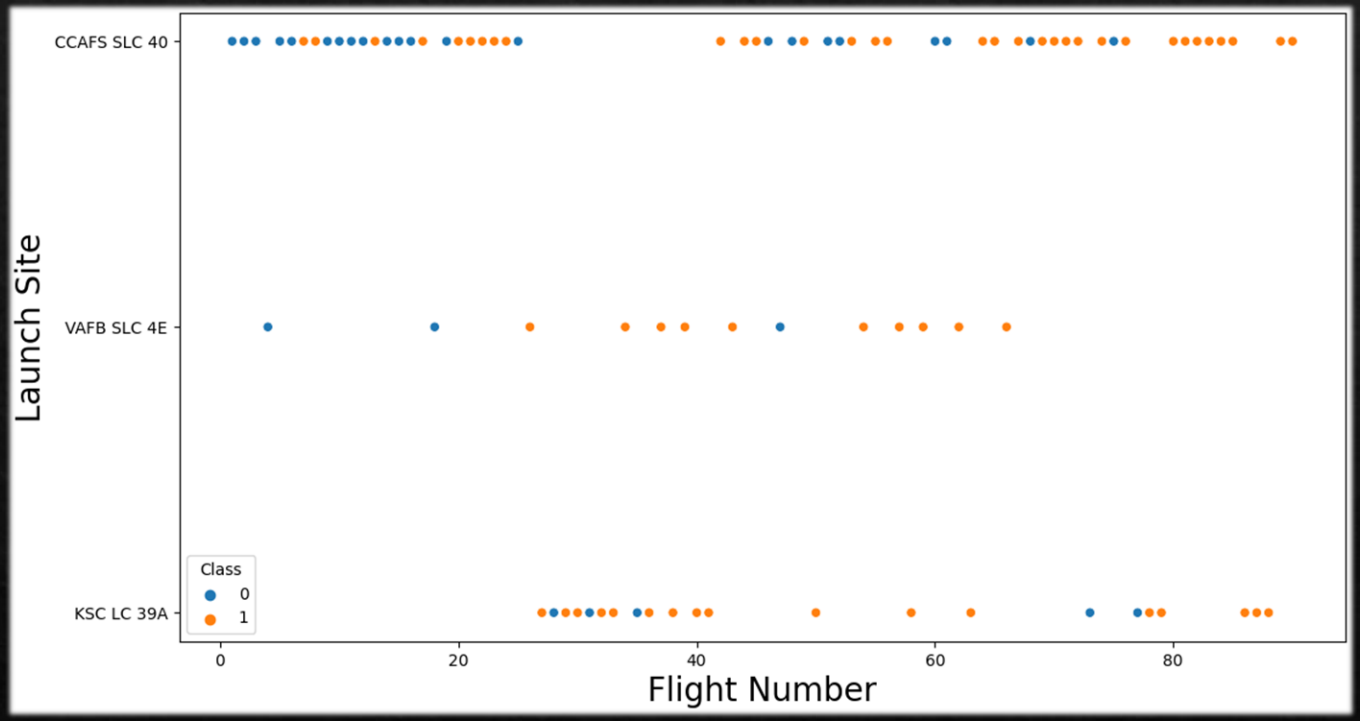
Insights

Visualization

Section 2

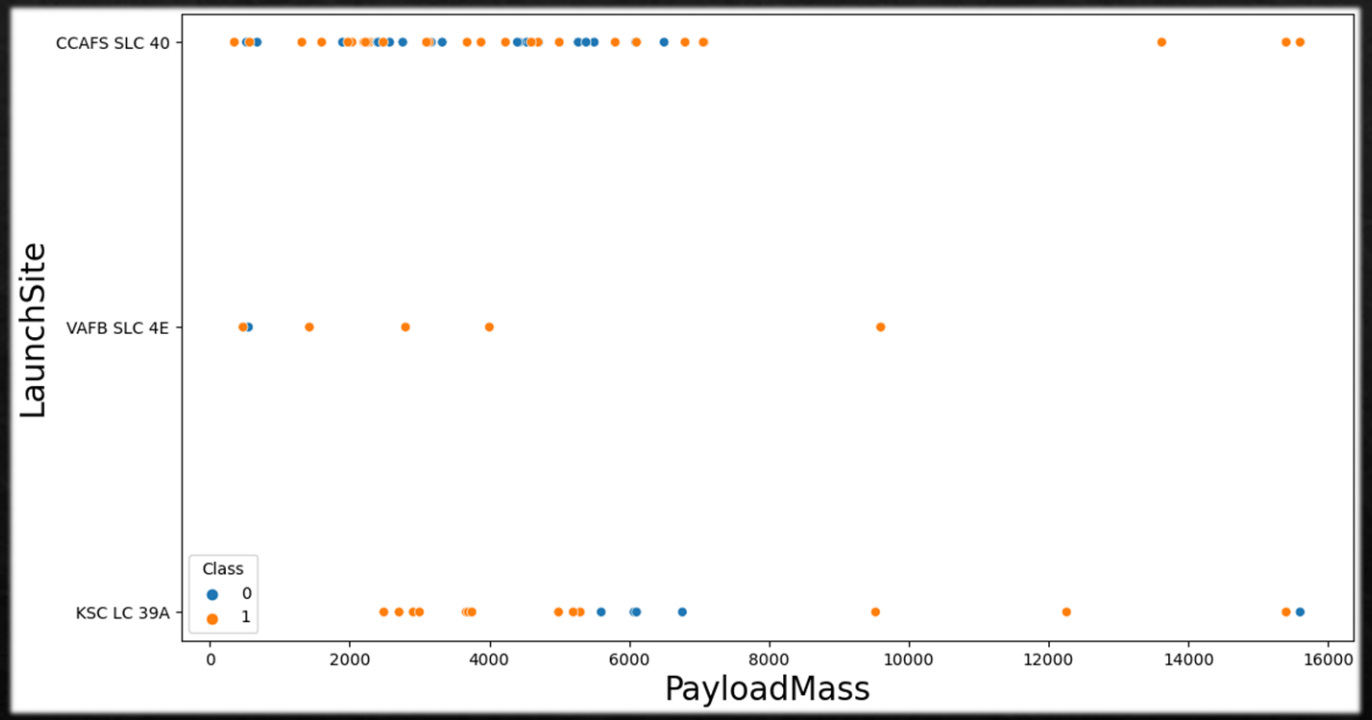
Flight Number vs. Launch Site

- Class 0 depict unsuccessful landings while Class 1 depict successful landings
- Most of the early flights were launched from CCAFS SLC 40 and were unsuccessful, almost 50% of all launches are from this launch site
- From flight number 20 onwards, there seems to be more success overall
- The success rate looks like it improves over time for all launch sites and is at 100% for CCAFS SLC 40 since flight number 75 onwards
- VAFB SLC 4E and KSC LC 39A have higher success rates



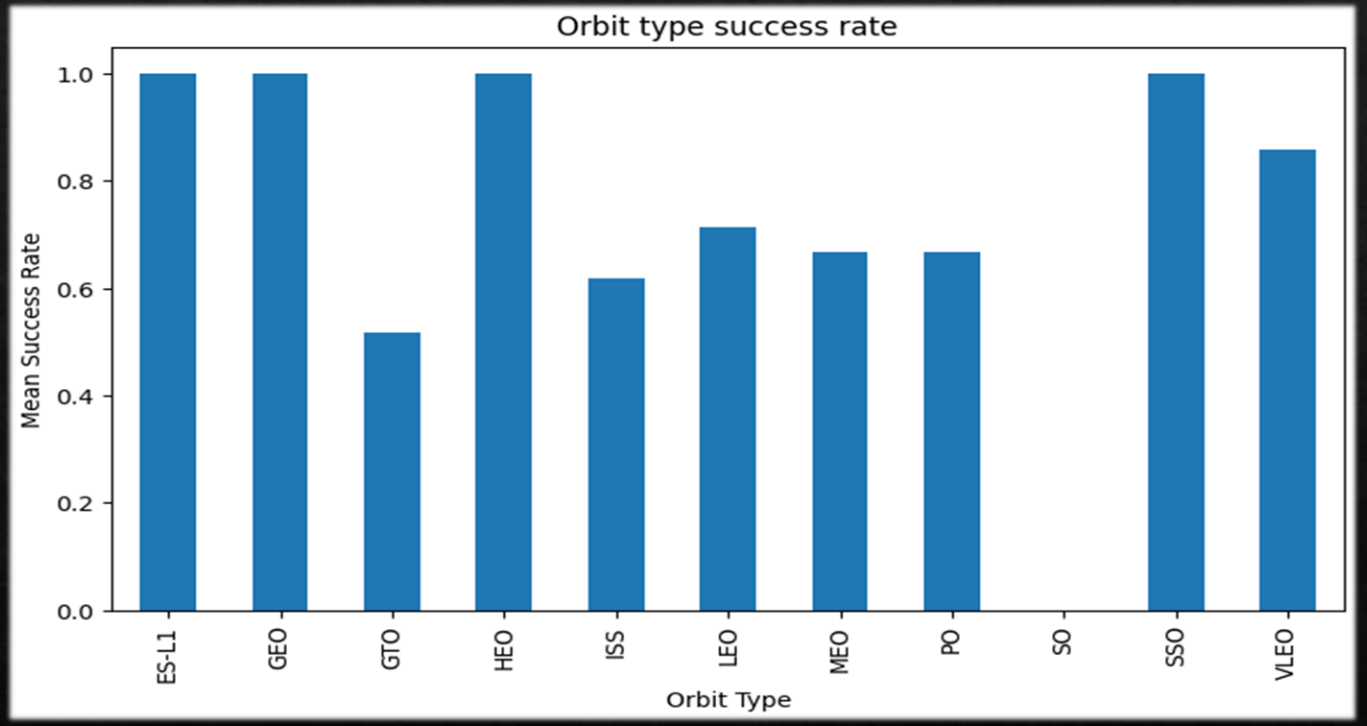
Payload vs. Launch Site

- Launches over 7000kg seem to be mostly successful, however the number of launches is low in comparison to the total number of launches
- More weight seems to strengthen the probability of a successful launch across all sites
- VAFB SLC 4E has a 100% success rate when the payload is above 1000kg
- KSC LC 39A has a 100% success rate when the payload is less than 5500kg

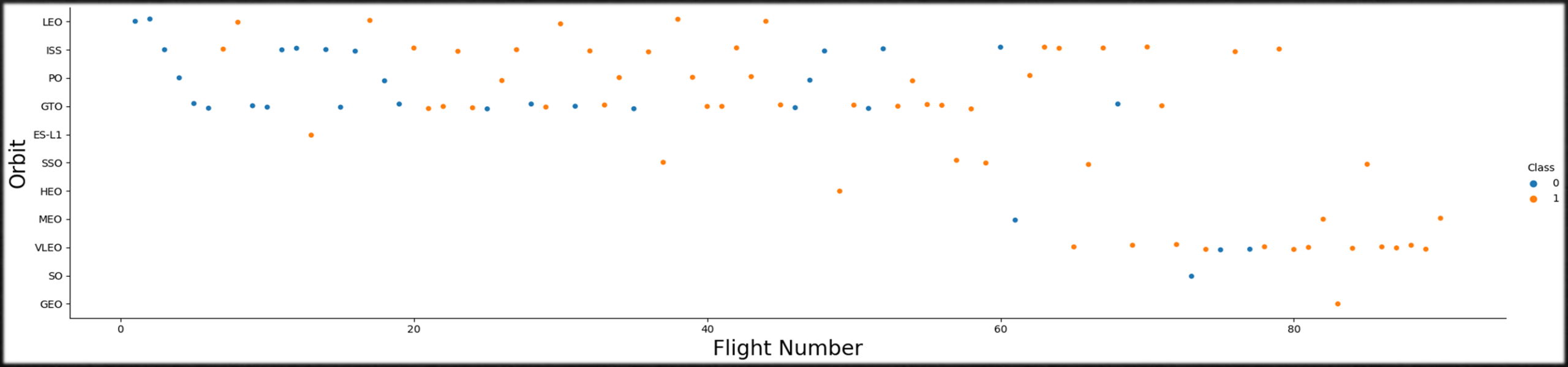


Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO & SSO all have a 100% success rate
- Orbits VLEO, LEO, MEO, PO, ISS & GTO have success rates between 50% and 85%
- Orbit SO has a 0% success rate

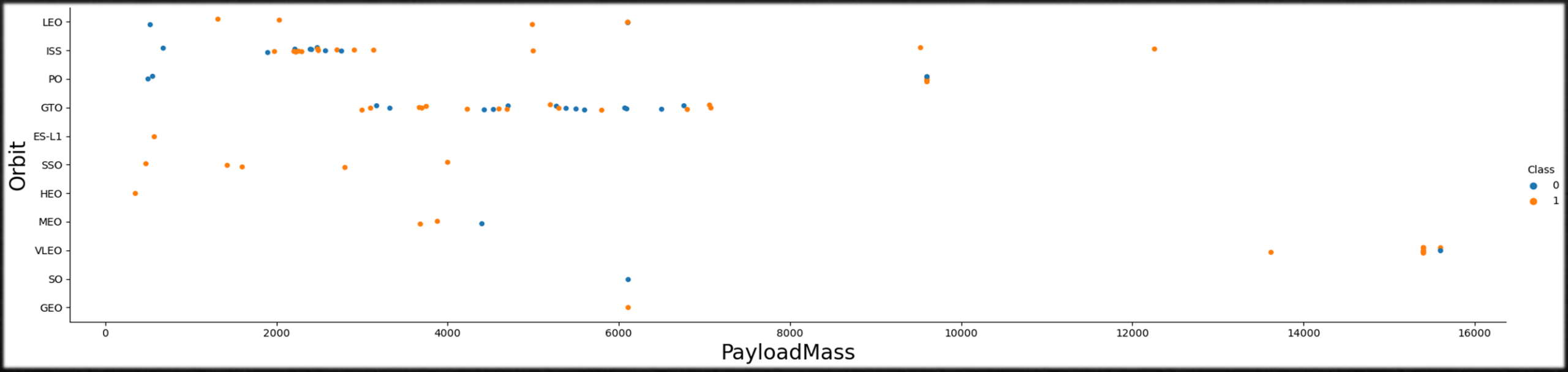


Flight Number vs. Orbit Type



- Class 0 depict unsuccessful landings while Class 1 depict successful landings
- Most of the early flights LEO, ISS, PO & GTO were unsuccessful, and show significant improvement after flight number 20
- LEO, ISS, PO & GTO are the most frequented orbits
- The success rate looks like it improves over time for all orbits
- ES-L1, SSO and HEO have 100% success rates

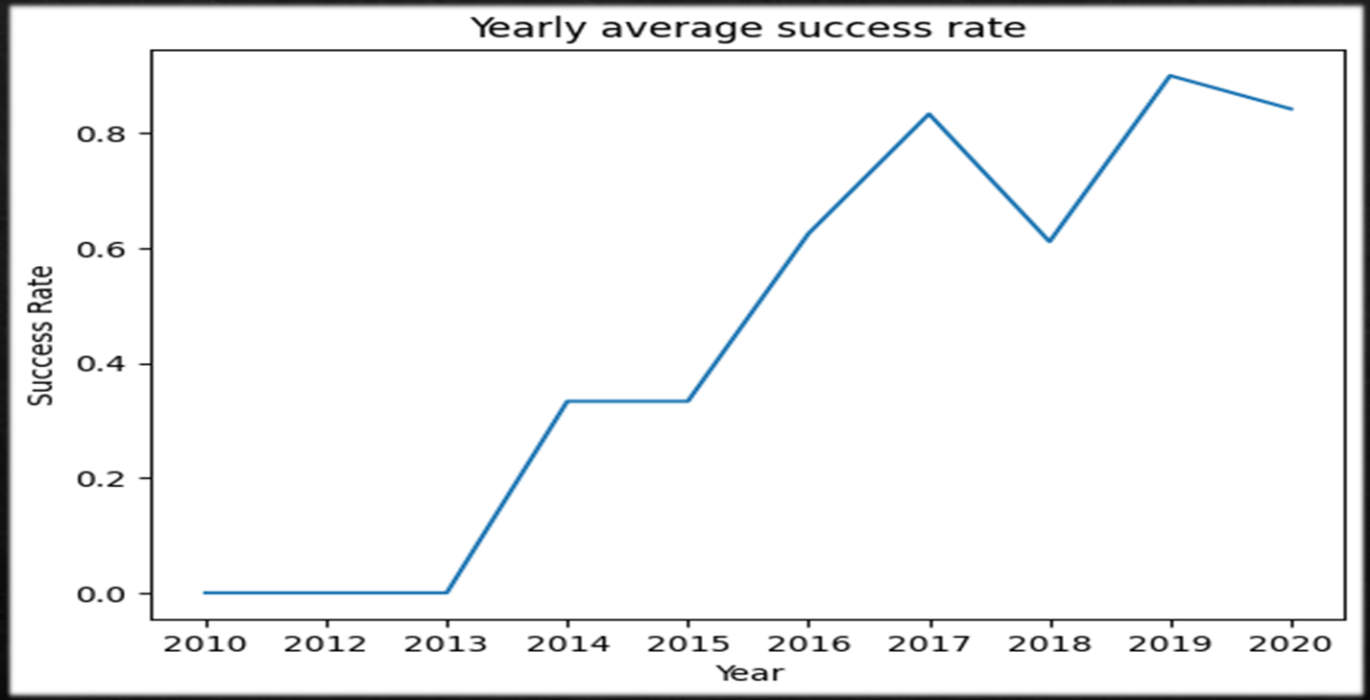
Payload vs. Orbit Type



- Class 0 depict unsuccessful landings while Class 1 depict successful landings
- SSO has performed at 100% success thus far and has reached loads up to 4000kg
- LEO, PO & ISS have performed better with higher payloads
- GTO has a mixed success rate and does not show a predictable pattern
- VLEO (very low earth orbits) are heavier

Launch Success Yearly Trend

- The success rate shows improvement year on year overall
- The success rate improved from 2013-2017 and 2018-2019
- There was a decrease in the success rate from 2017-2018 and 2019-2020



EDA

Insights

SQL

All Launch Site Names

Find the names of the unique launch sites

In [7]:
```
%%sql

SELECT DISTINCT(Launch_Site)
FROM SPACEXTBL;
```

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

Out[7]:

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

Using the "distinct" operator allowed us to filter the result set to unique launch sites

Launch Site Names Begin with 'CCA'

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

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

```
%%sql

SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

[No Title]

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

Out[8]:

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

Using the "limit" operator allowed us to filter the result set to 5 launch sites that contain 'CCA'

Total Payload Mass

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

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

In [9]:
```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total_Payload_KG_NASA(CRS)'
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

Out[9]:

Total_Payload_KG_NASA(CRS)
45596

Using "sum" allowed us to filter the result set to the total payload carried by NASA (CRS)

Average Payload Mass by F9 v1.1

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

In [10]:
```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average_Payload_KG_F9v1.1'
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1';
```

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

Out[10]:

Average_Payload_KG_F9v1.1
2928.4

Using "avg" allowed us to filter the result set to the average payload carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was achieved.

In [11]:
```
%%sql

SELECT MIN(DATE) AS First_Successful_Ground_Pad_Landing
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[11]:

First_Successful_Ground_Pad_Landing
2015-12-22

Using "min" allowed us to filter the result set to the first successful landing was achieved

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

In [18]:
```
%%sql

SELECT DISTINCT(Booster_Version)
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Out[18]:

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

Using "distinct", "and" & "between" allowed us to filter the result set to the desired data

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [31]:

```
%%sql

SELECT
    CASE
        WHEN Mission_Outcome LIKE 'Success%' THEN 'Successful'
        WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
    END AS Final_Status,
    COUNT(*) AS Outcomes

FROM SPACEXTBL

GROUP BY Final_Status;
```

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

Out[31]:

Final_Status	Outcomes
Failure	1
Successful	100

Using a case statement, we were able to categorize the outcome and label it as successful or failed.

Boosters Carried Maximum Payload

List the names of the booster versions which have carried the maximum payload mass (with a subquery)

In [39]:
```
%%sql

SELECT DISTINCT(Booster_Version) AS 'booster_versions which have carried the maximum payload mass'
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[39]:

booster_versions which have carried the maximum payload mass
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

Using a subquery in the “where” clause, we were able to list the boosters that carried the maximum payload.

2015 Launch Records

List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in 2015.

In [71]:
```
%%sql

SELECT substr(Date,6,2) AS "Month", substr(Date,1,4) AS "Year", Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND "Year" = "2015";
```

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

Out[71]:

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Using "substr" we were able to extract the month and display it along with the year when the landings failed in 2015.

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 [97]:

```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome) AS 'Count'
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY 'Count' DESC;
```

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

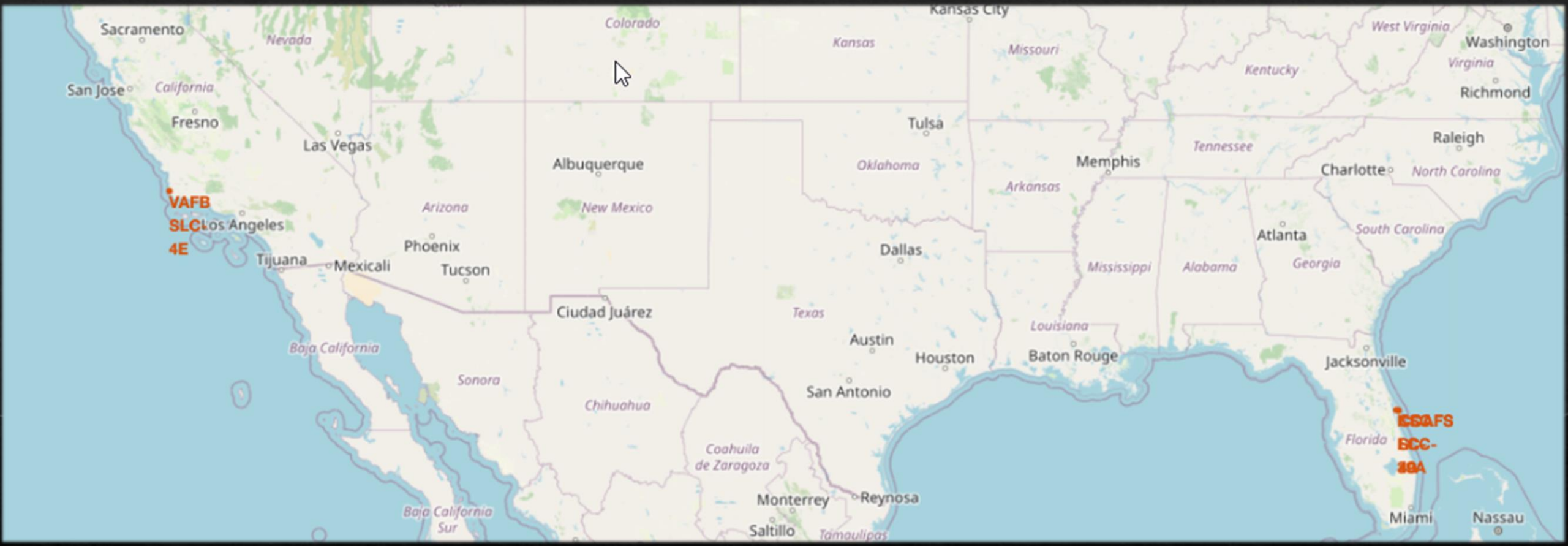
Out[97]:

Landing_Outcome	Count
Uncontrolled (ocean)	2
Success (ground pad)	5
Success (drone ship)	5
Precluded (drone ship)	1
No attempt	10
Failure (parachute)	1
Failure (drone ship)	5
Controlled (ocean)	3

Launch Sites
Proximities Analysis

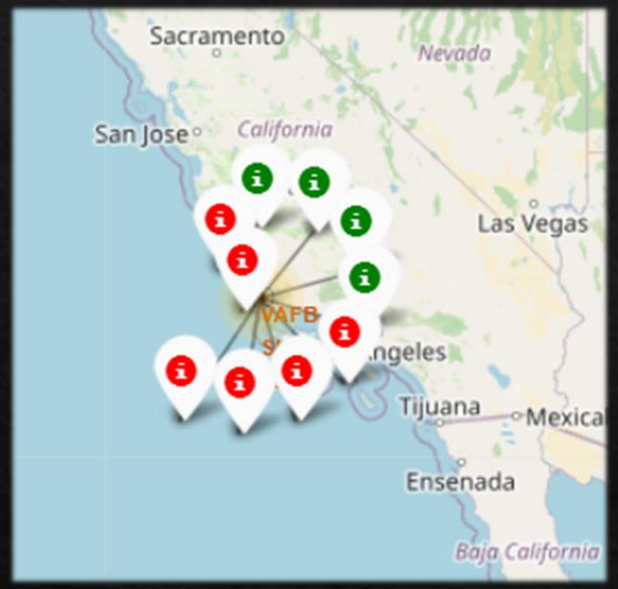
Section 3

Launch Site Locations



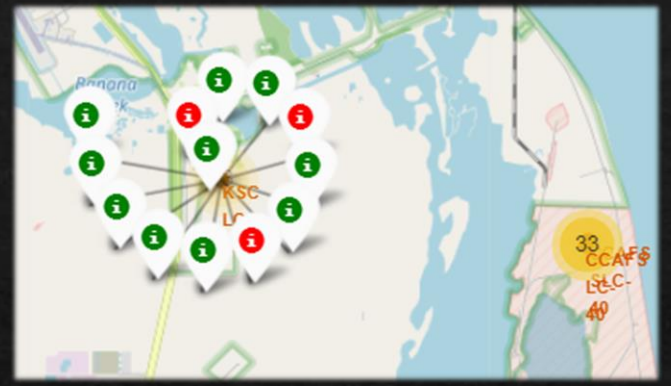
We can see that the launch sites are located within the United States and located around coastal areas

Launch Outcomes



California launch site
VAFB SLC 4E

California launch sites
CCAFS SLC-40
CCAFS LC-40



Florida launch site
KSC LC-39A

The Red markers represent unsuccessful launches while the Green markers represent successful launches

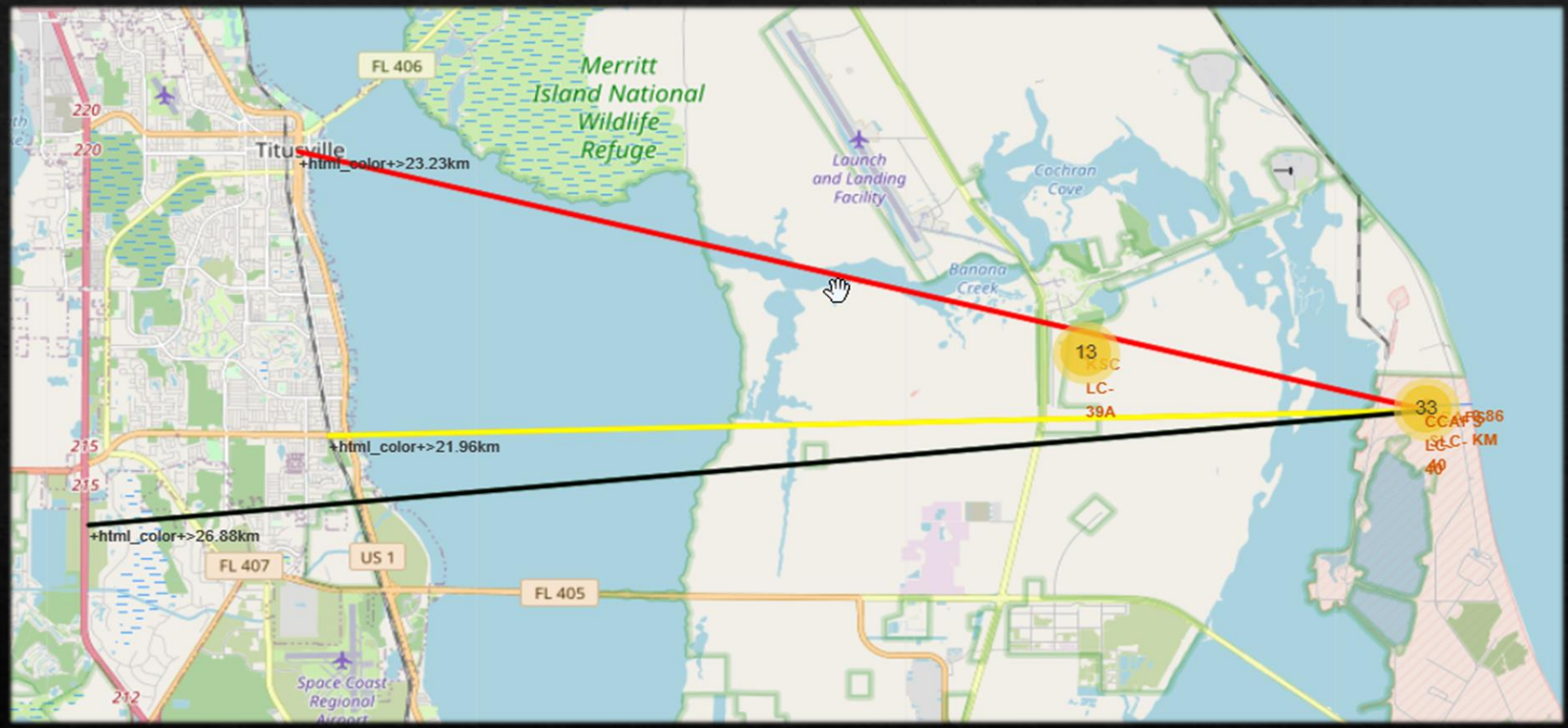
Distances between a launch site to its proximities

Are launch sites in close proximity to railways? No

Are launch sites in close proximity to highways? No

Are launch sites in close proximity to coastline? Yes

Do launch sites keep certain distance away from cities? Yes

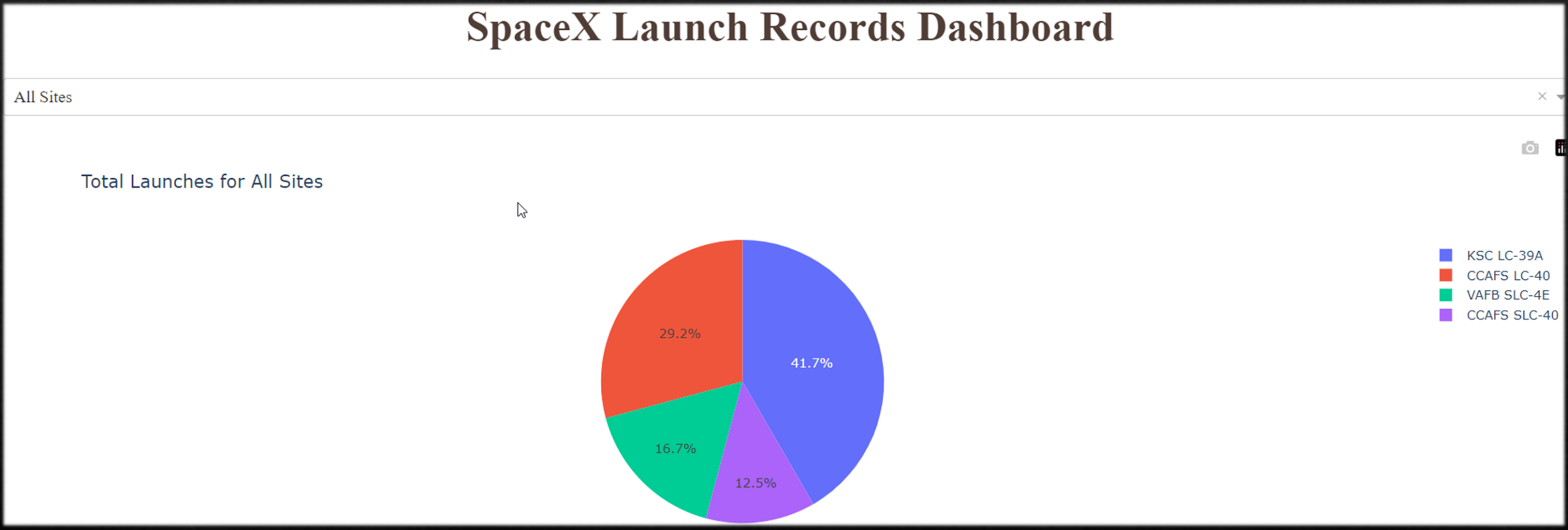


Building a Dashboard

Plotly Dash

Section 4

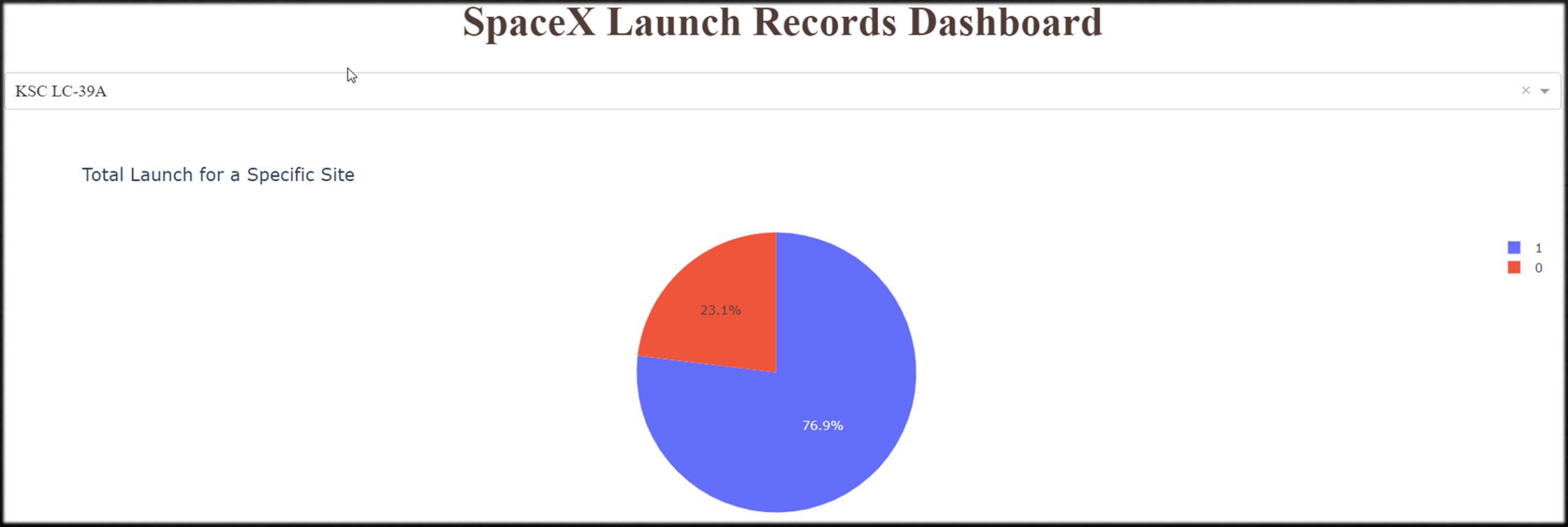
Site Wise Launch Success



KSC LC-39A is the most successful site at 41.7%

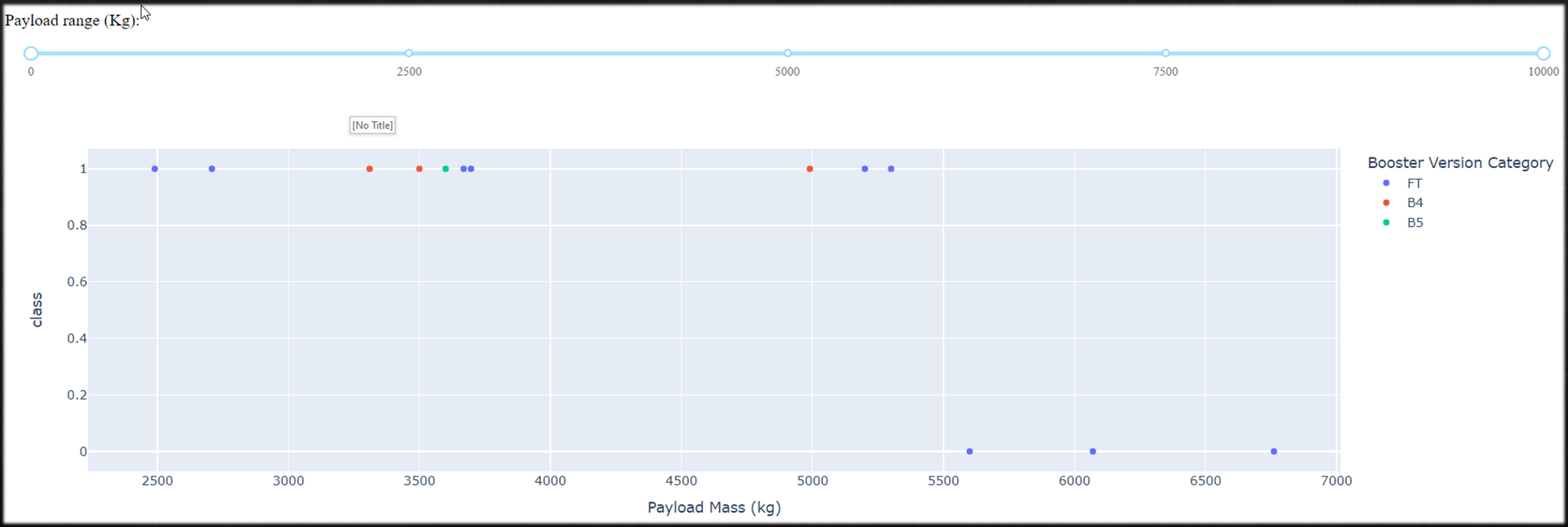
CCAFS SLC-40 is the least successful site at 12.5%

Launch site with highest launch success ratio



KSC LC-39A has a 76.9% success rate and a 23.1% failure rate.

Payload vs. Launch Outcome



Lighter payloads seem to have more success

Predictive Analysis

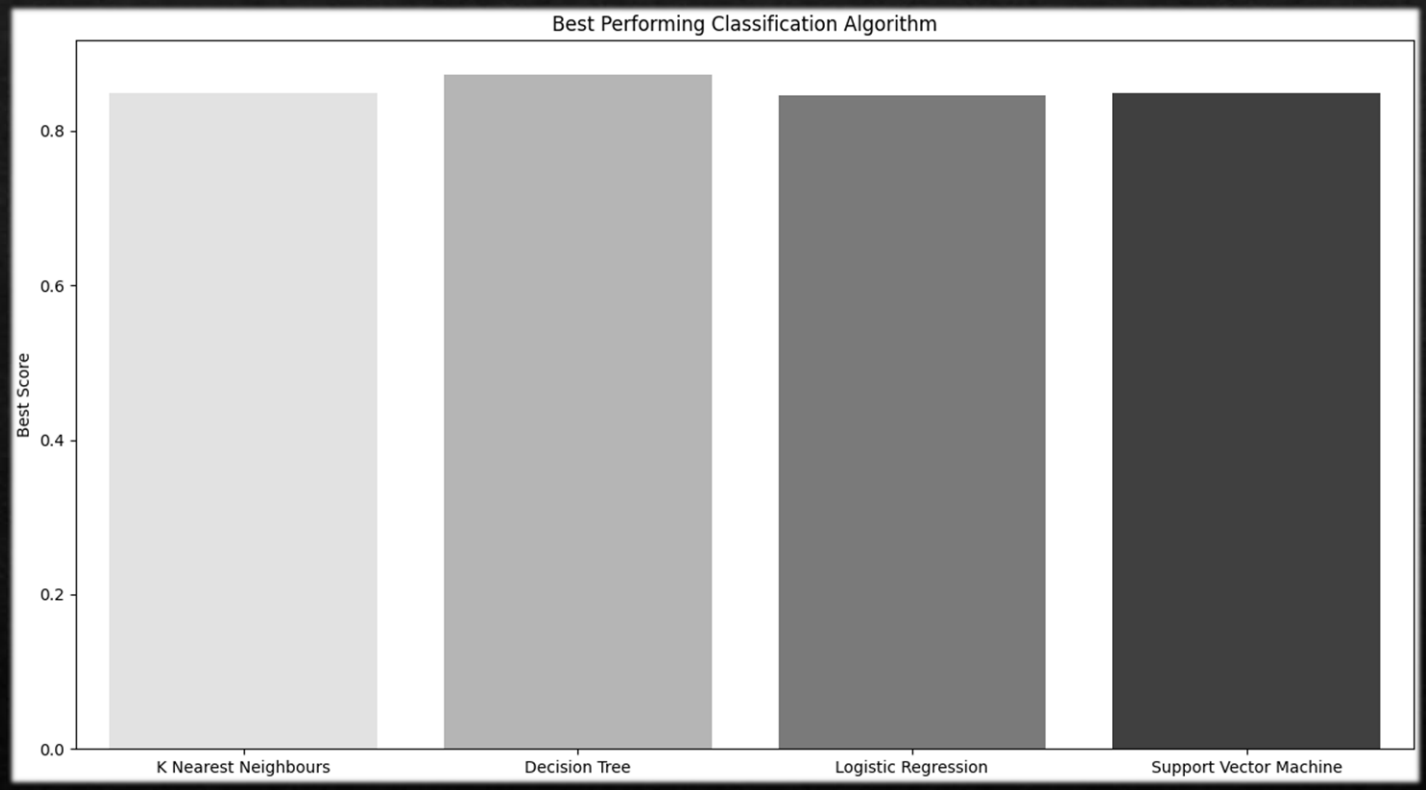
Classification

Section 5

Classification Accuracy

The below mentioned classification algorithms were used to discover which one has the highest accuracy.

	Algorithm	Score
0	K Nearest Neighbours	0.848214
1	Decision Tree	0.873214
2	Logistic Regression\t	0.846429
3	Support Vector Machine	0.848214



The decision tree model performed the best and had an accuracy of 87%

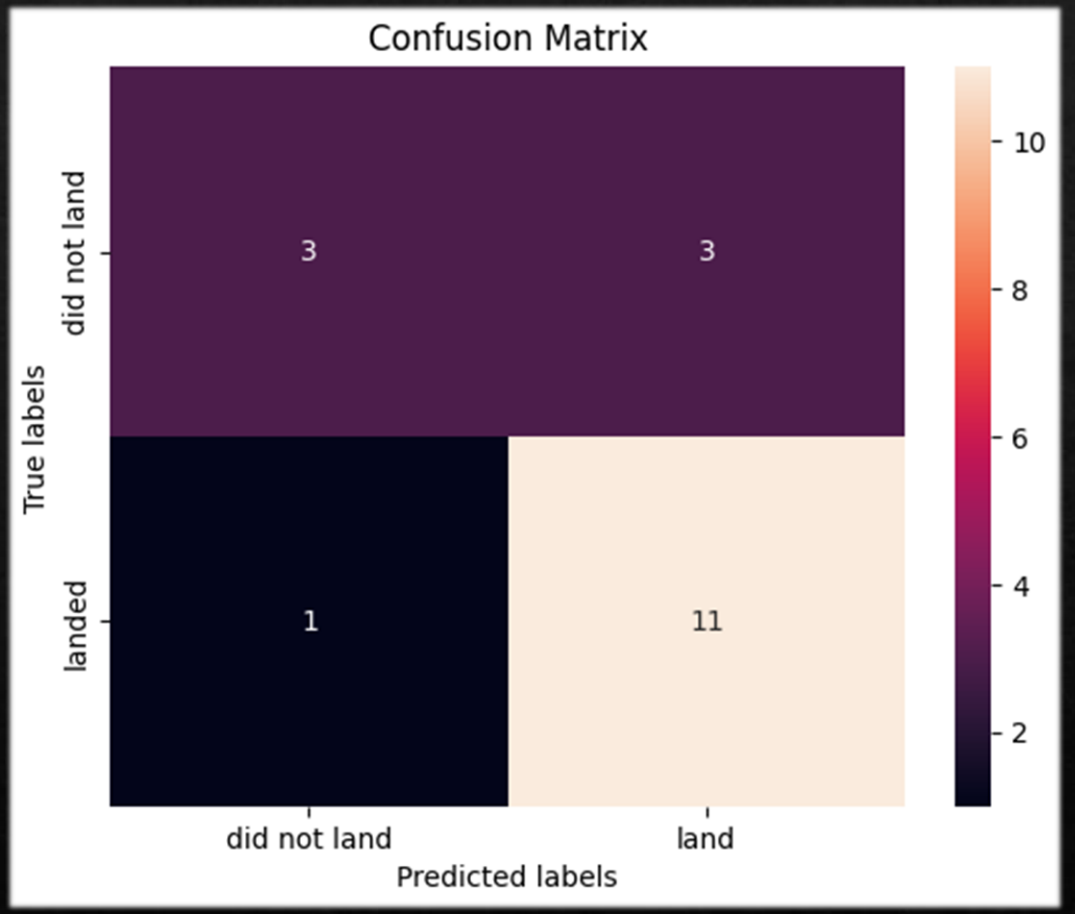
Confusion Matrix

The confusion matrix for the best performing model (decision tree)

11 Truly Positive landings vs 1 False Negative landing
3 Truly Negative failed landings vs 3 False Positive failed landings

All the other algorithms tested predicted 12 Truly Positive landings and hence were not as accurate.

This means that the decision tree predicted 14 correct outcomes vs 13 by the rest.



Conclusions

- The success rate of launches increases with the number of flights carried out
- Orbits ES-L1, GEO, HEO & SSO all have a 100% success rate
- LEO, ISS, PO & GTO are the most frequented orbits
- The total number of successful missions is 100 vs 1 mission failure
- All launch sites are located within the United States and located around coastal areas
- KSC LC-39A is the most successful site at 41.7%
- Lighter payloads seem to have more success
- The decision tree model was the best performing classification model and had an accuracy of 87%

Appendix

- All of the notebooks that were used to complete this project can be found at my GitHub repository at the following link:

IBM Data Science Capstone - SpaceX - Assad Khan

Thank You!