IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bernard Prata Meireles Vieira Fernandes
02-JUL-2023

Table of Contents

Executive Summary

- Bspace is a new commercial rocket launch provider who wants to bid against SpaceX.
- SpaceX advertises launch services starting at $62 million for missions that allow some fuel to be reserved for landing the 1st stage rocket booster, so that it can be reused.
- Given mission parameters such as payload mass and desired orbit, the models produced in this report were able to predict the first stage rocket booster landing successfully with an accuracy level of 83.3%.
- As a result, BSpace will be able to make more informed bids against SpaceX by using 1st stage landing predictions as a proxy for the cost of a launch.

Introduction: BACKGROUND

- This report has been prepared as part of the Applied Data Science Capstone course.
- In this capstone, I take the role of a data scientist working for a new rocket company called
- Bspace With the help of the data science findings and models in this report, SpaceY will be able to make more informed bids against SpaceX for a rocket launch.

Introduction: BUSINESS PROBLEM

- SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars when the first stage of their rockets can be reused.
- Sometimes SpaceX will sacrifice the first stage due to mission parameters such as payload, orbit, and customer.
- Therefore, this report aims to accurately predict the likelihood of the first stage rocket landing successfully as a proxy for the cost of a launch.

Section 1

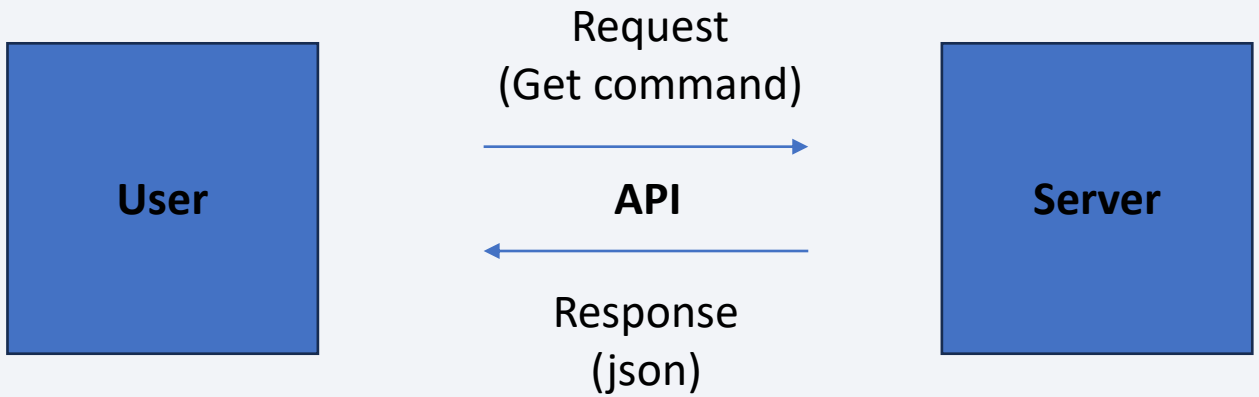
Methodology

Methodology

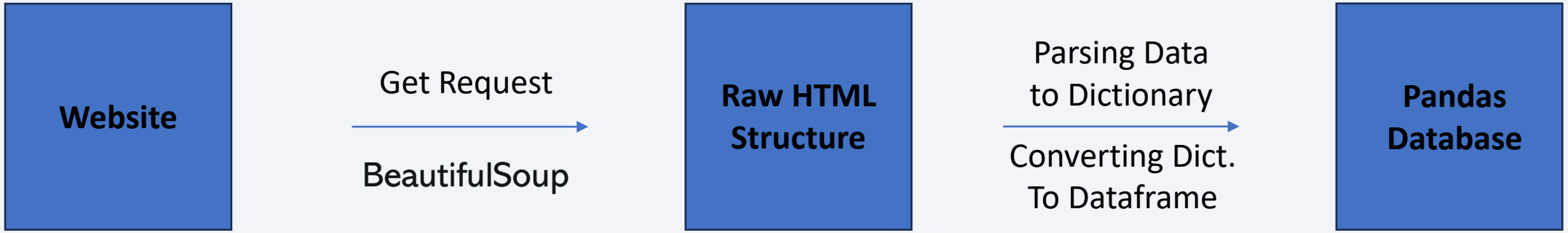
- Data collection methodology:
  - HTTP requests to get data from an Open-Source REST API for SpaceX
  - Web Scraping Wikipedia for a list of the Launches
- Perform data wrangling
  - Prepare the data for training supervised models through Exploratory Data Analysis (to find patterns), data cleaning (dealing with missing values, etc.) and labeling data (for predictive models)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing the data to fit into different types of prediction models we determine the best Hyperparameters for each and then evaluate the best one based on the accuracy on the test data and the results from the Confusion Matrix.

Data Collection

- The data sets were collected using get requests to the SpaceX API and using web scrapping with BeautifulSoup on a Wikipedia page.
- API:



- Web Scrapping:



Data Collection – SpaceX API

- Acquired historical launch data from Open Source REST API for SpaceX
- Requested and parsed the SpaceX launch data using the GET requests
- Filtered the dataframe to only include Falcon 9 launches

- GitHub Link

Data Collection - Scraping

- Requested a Falcon9 launch records HTML table from a Wikipedia page
- Parsed the table and converted it into a Pandas data frame

- GitHub Link

Data Wrangling

- Describe how data were processed:
- Replaced missing payload mass values from classified missions with mean
- Explored data to determine the label for training supervised models
- Created a landing outcome training label from 'Outcome' column

- GitHub Link

EDA with Data Visualization

Used Matplotlib and Seaborn visualization libraries to understand how the variables influence the outcome (Feature Engineering).

- FlightNumber x PayloadMass †
- FlightNumber x LaunchSite †
- Payload x LaunchSite †
- Orbit type x Success rate
- FlightNumber x Orbit type †
- Payload x Orbit type †
- Year x Success rate

† = with Class overlayed (1st stage booster landing outcome)

- GitHub Link

EDA with SQL

Ran SQL queries to display and list information about

- Launch sites
- Payload masses (Average and Total)
- Booster versions
- Mission outcomes
- Booster landings

- GitHub Link

Build an Interactive Map with Folium

Objects created on the folium map:

- Launch site locations – Close to the coast and to the Equator Line
- Marked the success/failed launches for each site – To determine which have high success rates
- Calculate the distances between a launch site to its proximities (cities, railroads and highways) – Close to railways and highways for logistics and personnel reasons. Away from cities to minimize risk to the population and property damage.

- GitHub Link

Build a Dashboard with Plotly Dash

- Used Python interactive dashboarding library called Plotly Dash to enable exploration and manipulation of the data in an interactive way
- User Interactions
  - Range slider to specify payload amount
  - Drop-down menu to choose between all sites and individual launch sites
- Pie chart showing success rates:
  - Total by site (Color coded by launch site)
  - For each individual site
- Scatter chart showing payload mass vs. landing outcome (Color coded by booster version)

- GitHub Link

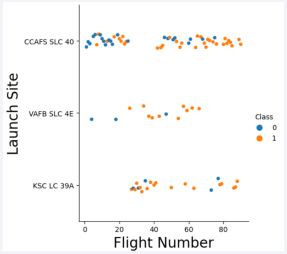
Predictive Analysis (Classification)

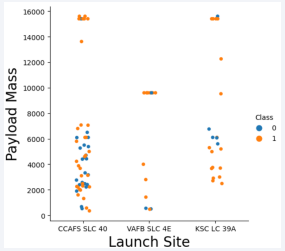
- Loaded the data frame created throughout the whole course
- Split the data into training data and test data
- Fit the training data to various model types (Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors Classifier)
- Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each mode
- Evaluated accuracy of each model using test data to select the best model

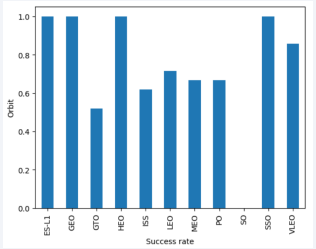
- GitHub Link

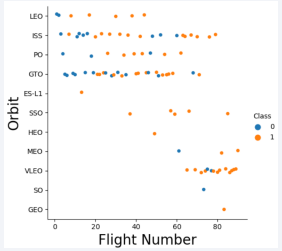
Results

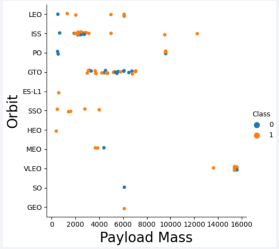
- Exploratory data analysis results

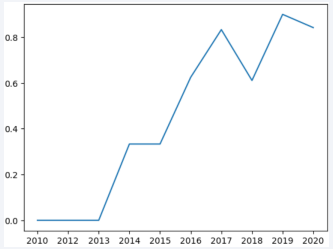




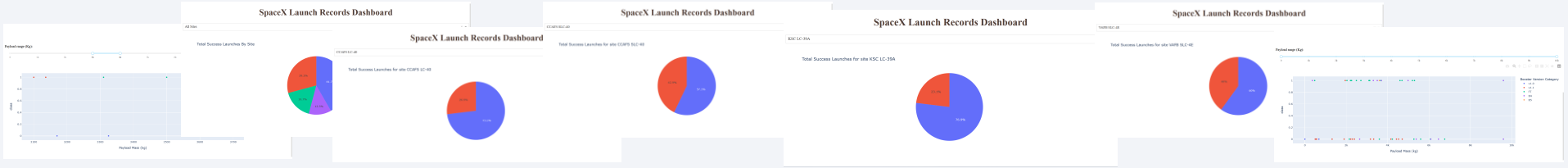




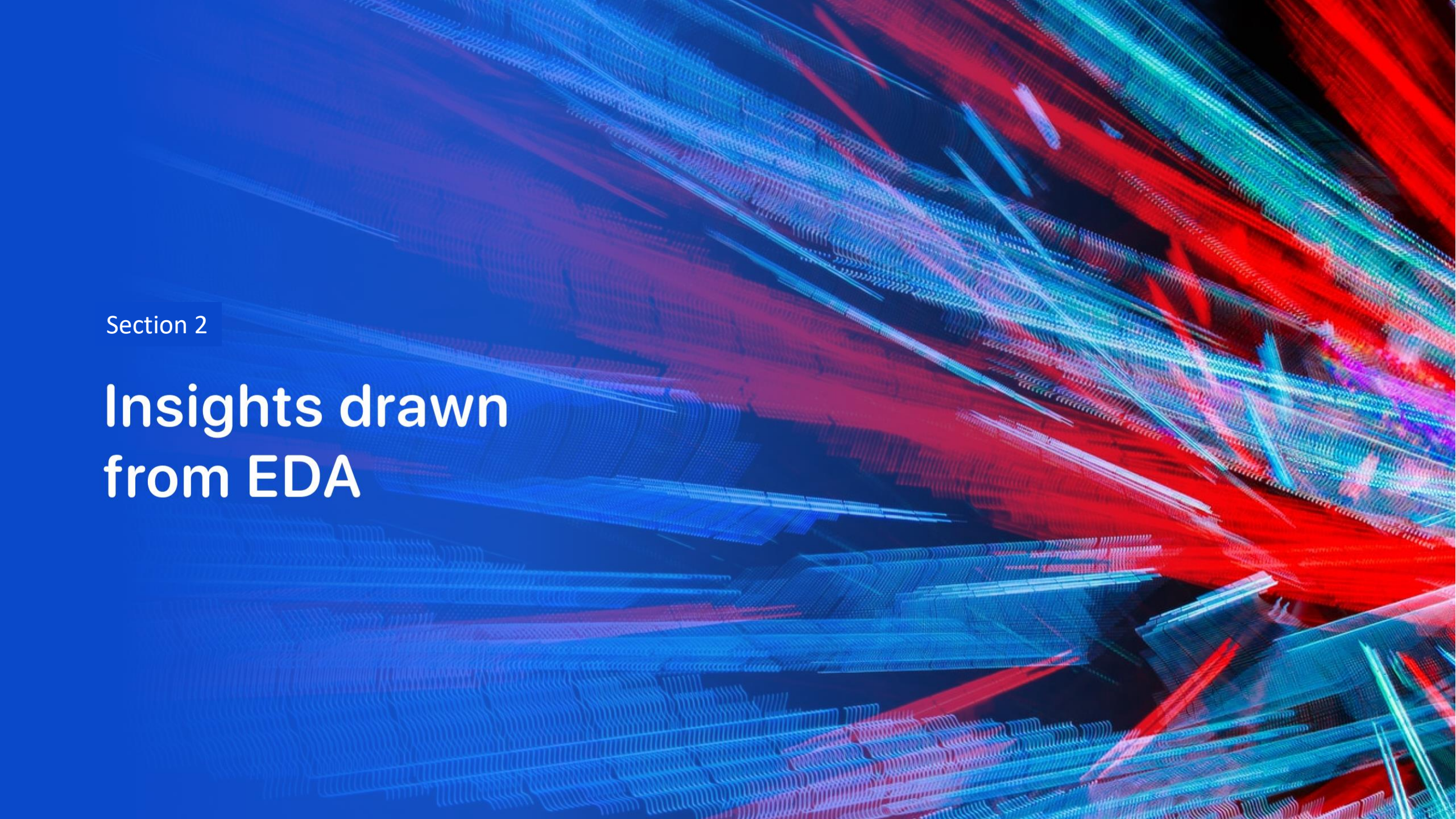




- Interactive analytics demo in screenshots



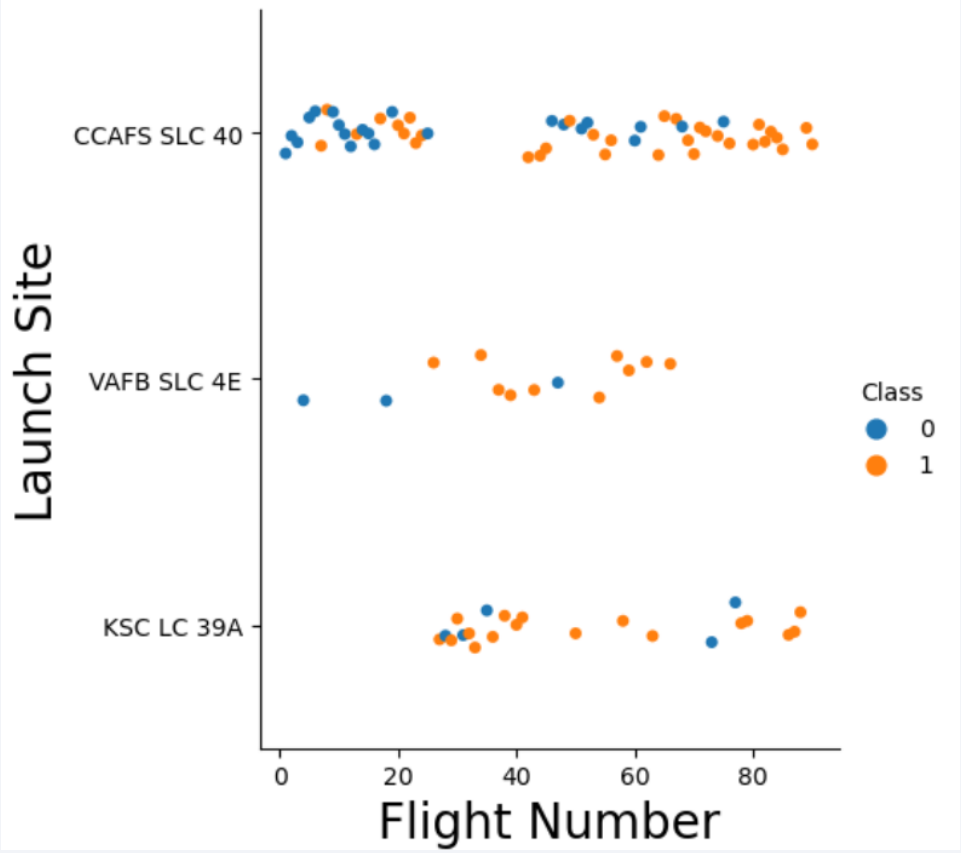
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

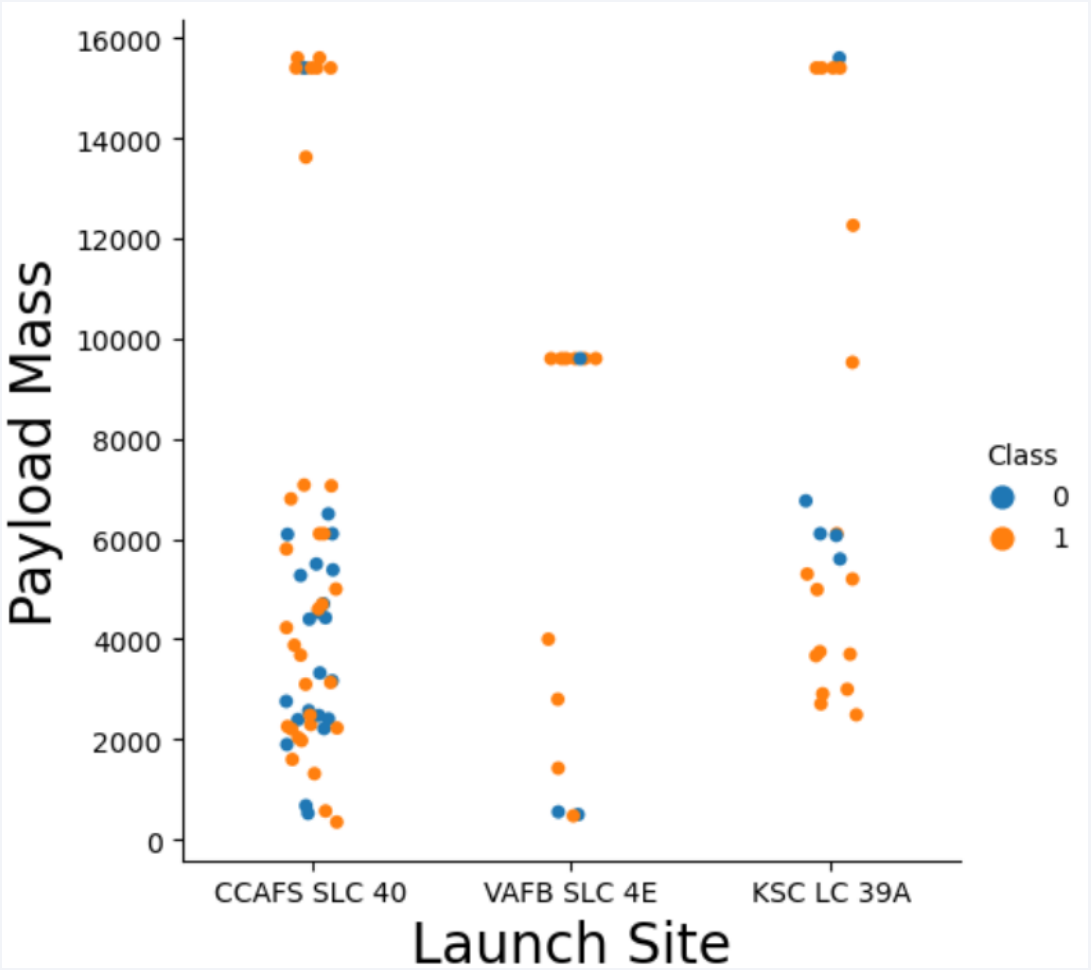
- CCAFS SLC 40 has the most unsuccessful landings (total and percentage)
- KSC LC 39A Has the least percentual unsuccessful landings
- VAFB SLC 4E Has the least total unsuccessful landings

- As flight number increases, the number of unsuccessful landings seams to decrease



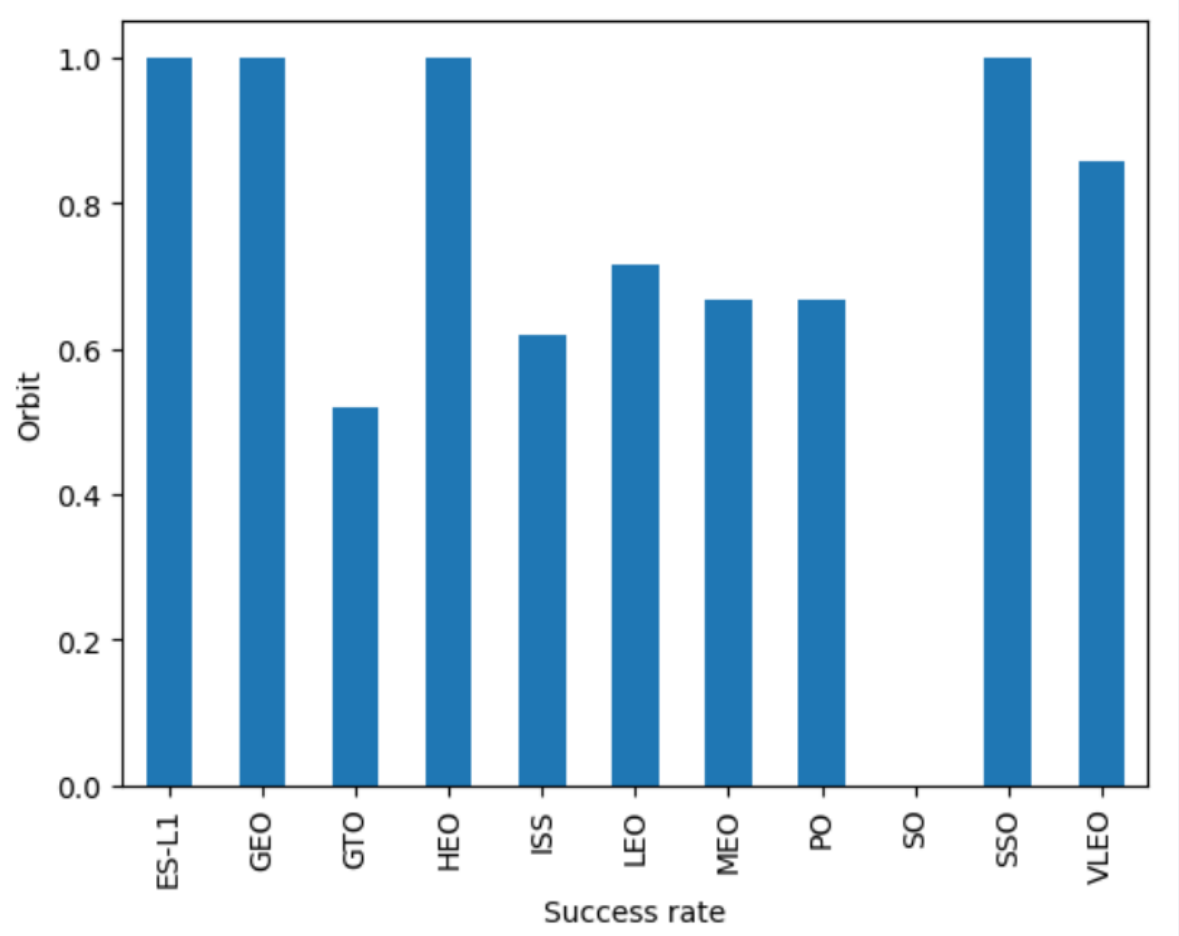
Payload vs. Launch Site

- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)



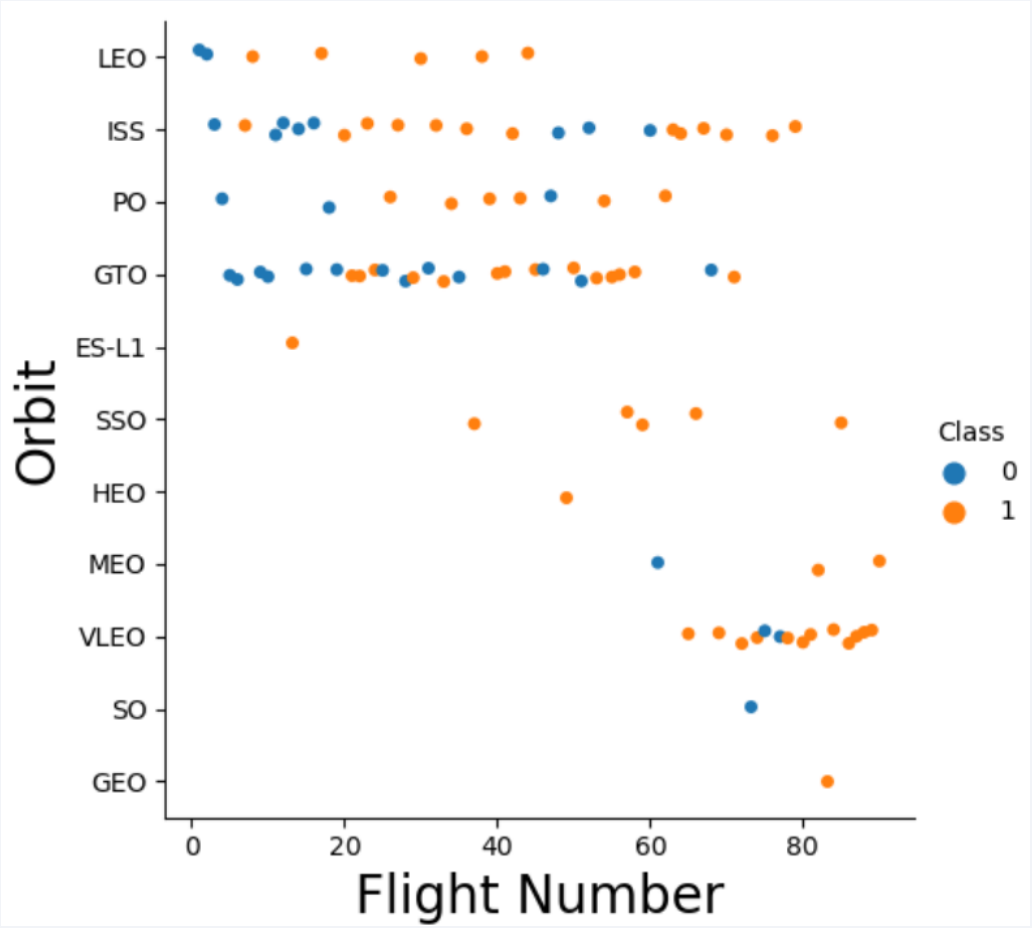
Success Rate vs. Orbit Type

- All orbit types except 'SO' have had successful 1st stage landings



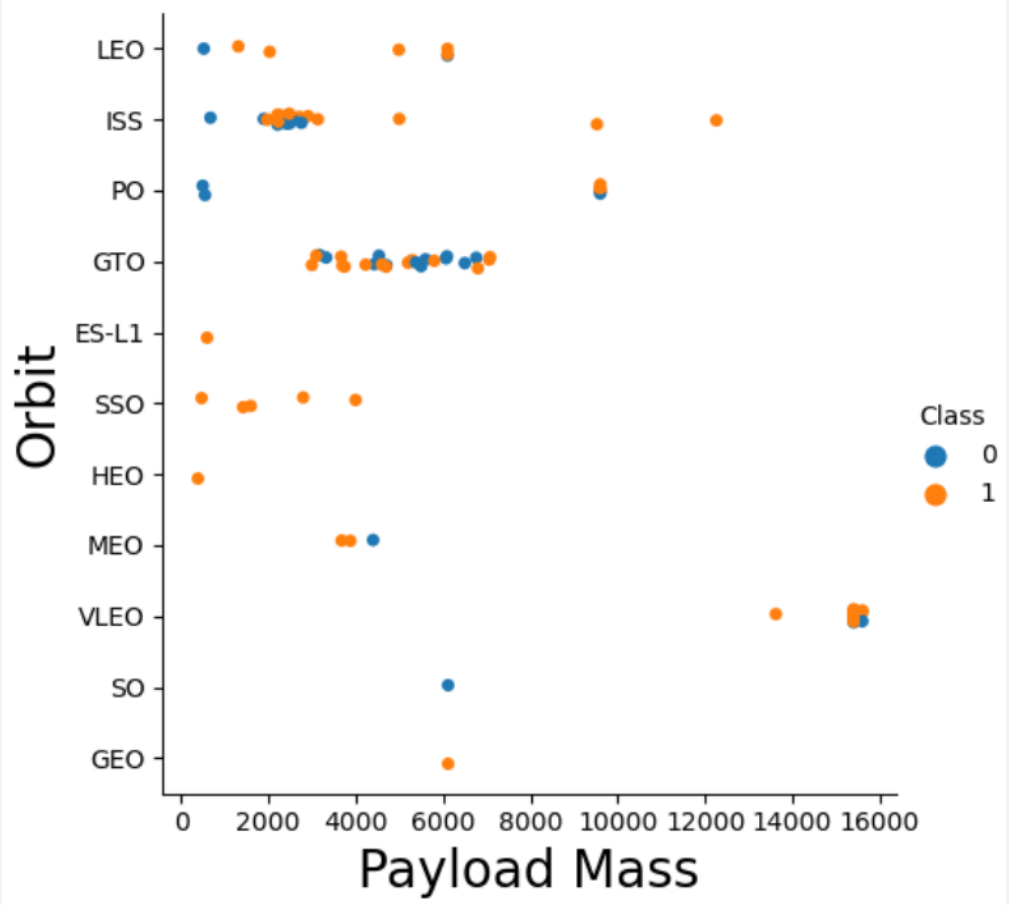
Flight Number vs. Orbit Type

- Flight number positively correlated with1st stage recovery for all orbit types
- Higher flight numbers seem to favor the lower list of orbits in the y axis.



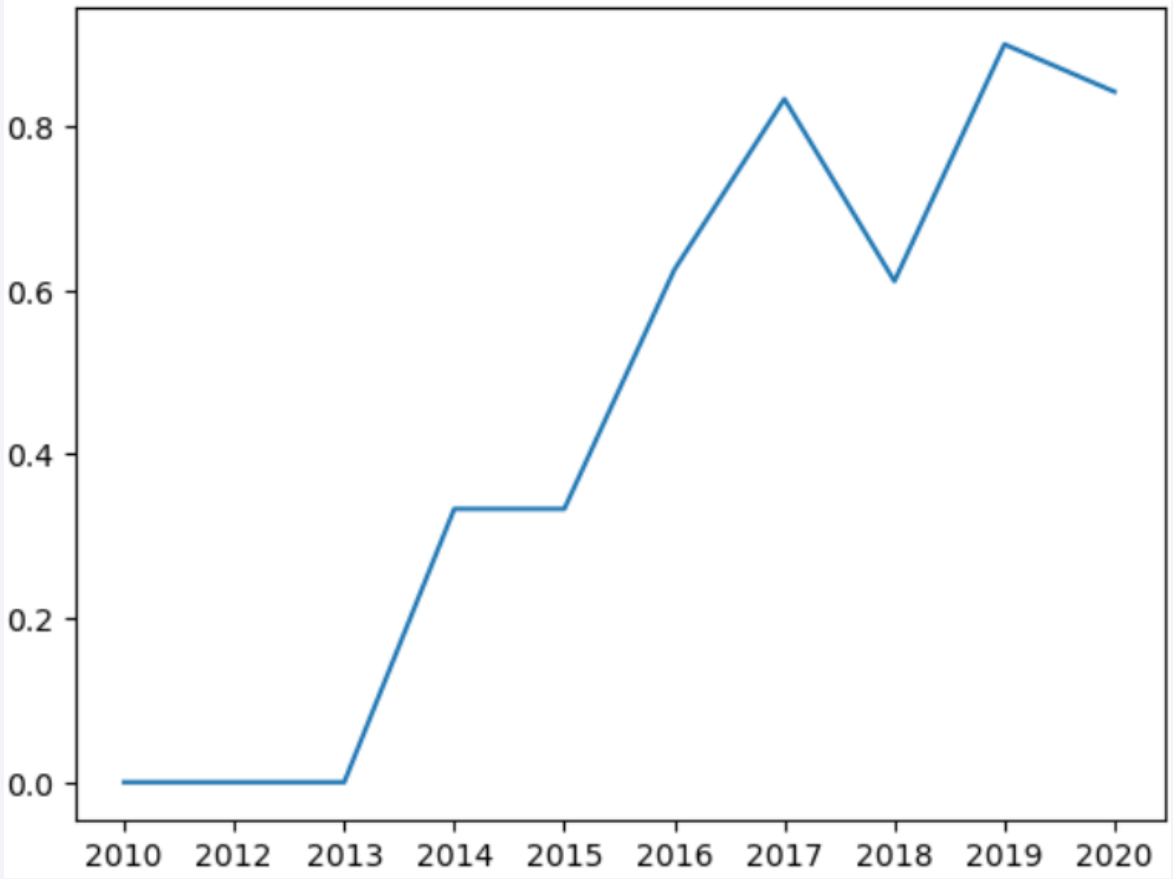
Payload vs. Orbit Type

- GTO orbit has a high mission failure rate
- SSO, ES-L1, HEO and GEO orbit have no failed missions.
- VLEO only has heavy payload missions
- ES-L1, HEO and GEO seam to be low demand orbits to SpaceX



Launch Success Yearly Trend

- As time went by, mission success rate increased.



All Launch Site Names

In [10]:

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

In [11]:
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

In [12]:

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

In [13]:

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

In [14]:

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [15]:

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

In [16]:

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:

The total number of successful mission outcome is:

	successoutcome
0	100

The total number of failed mission outcome is:

	failureoutcome
0	1

Boosters Carried Maximum Payload

In [17]:

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

In [18]:

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

In [19]:

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

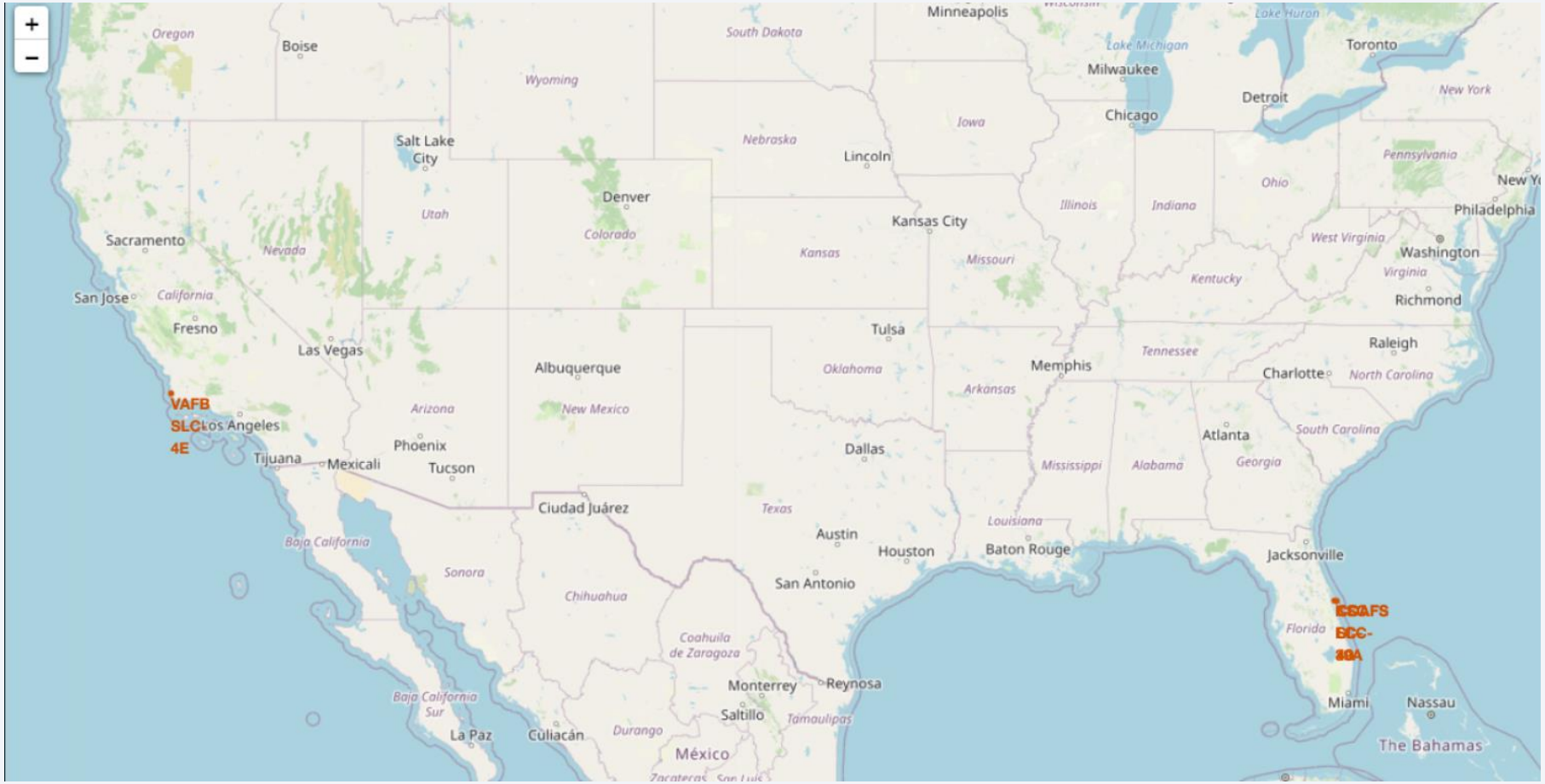
Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

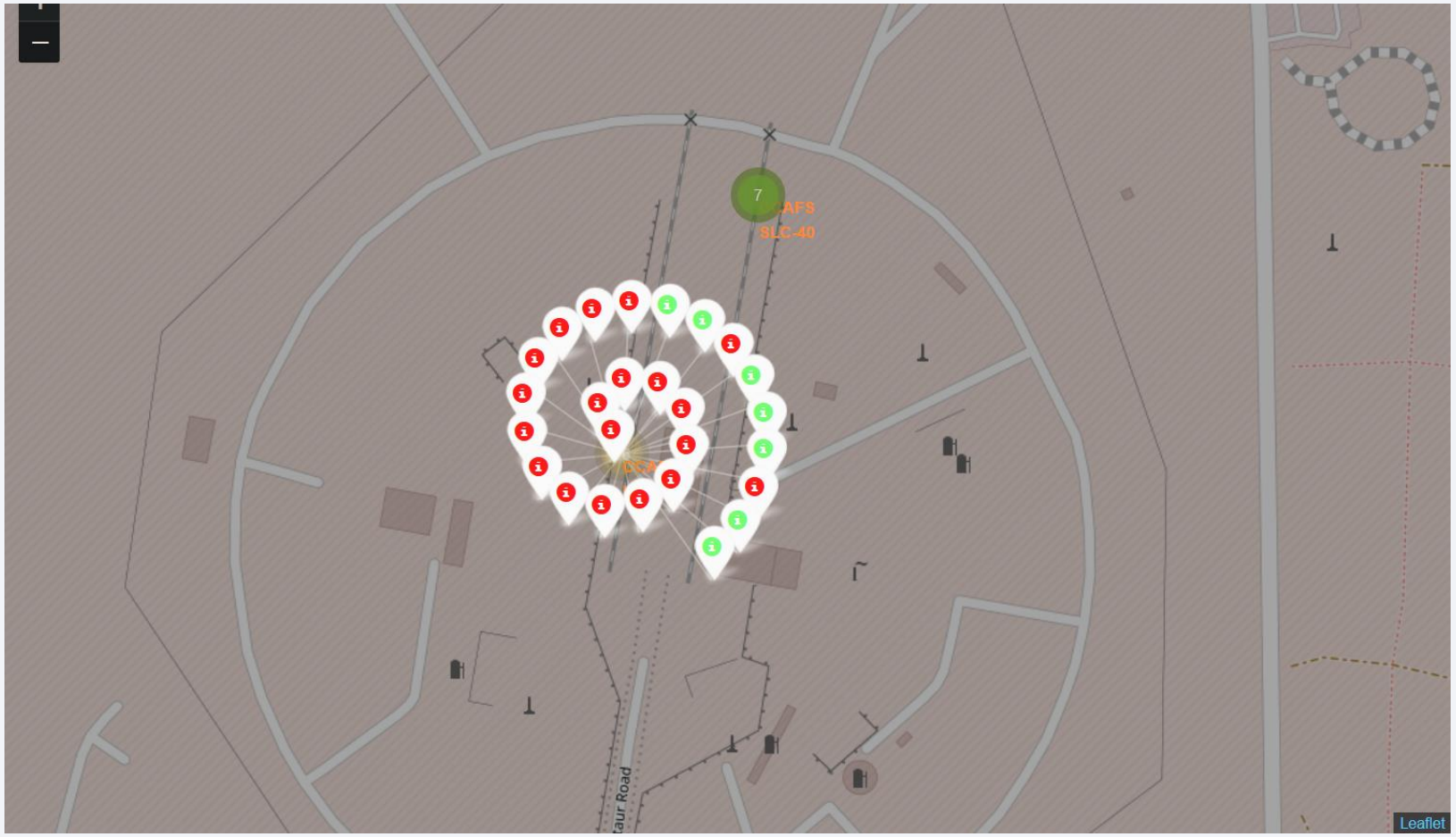
Launch Sites Proximities Analysis

All Launch Sites



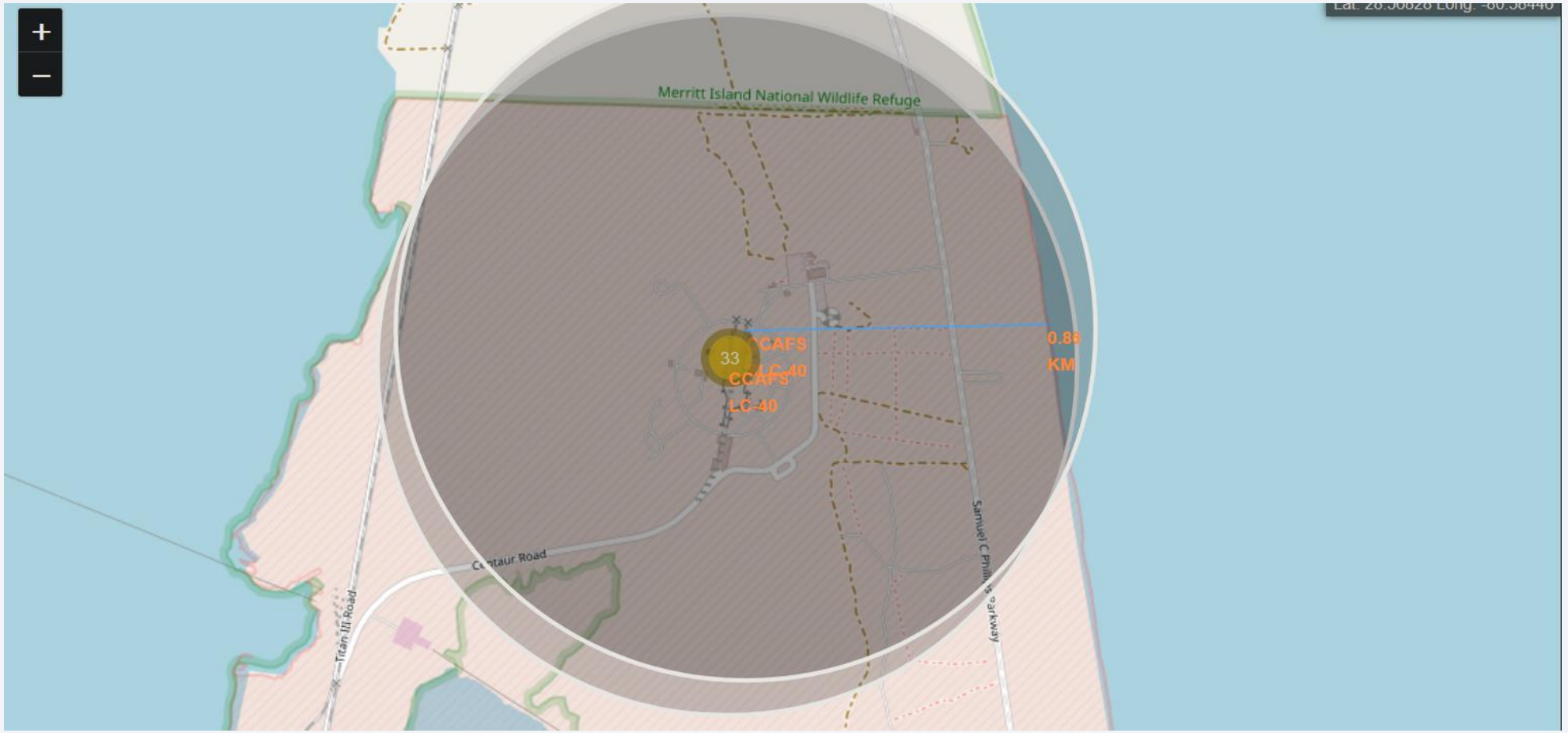
- All launch sites are close to the cost (for safety reasons) and to the equator line (it takes less fuel to get into space from the equator)

Launch Outcomes

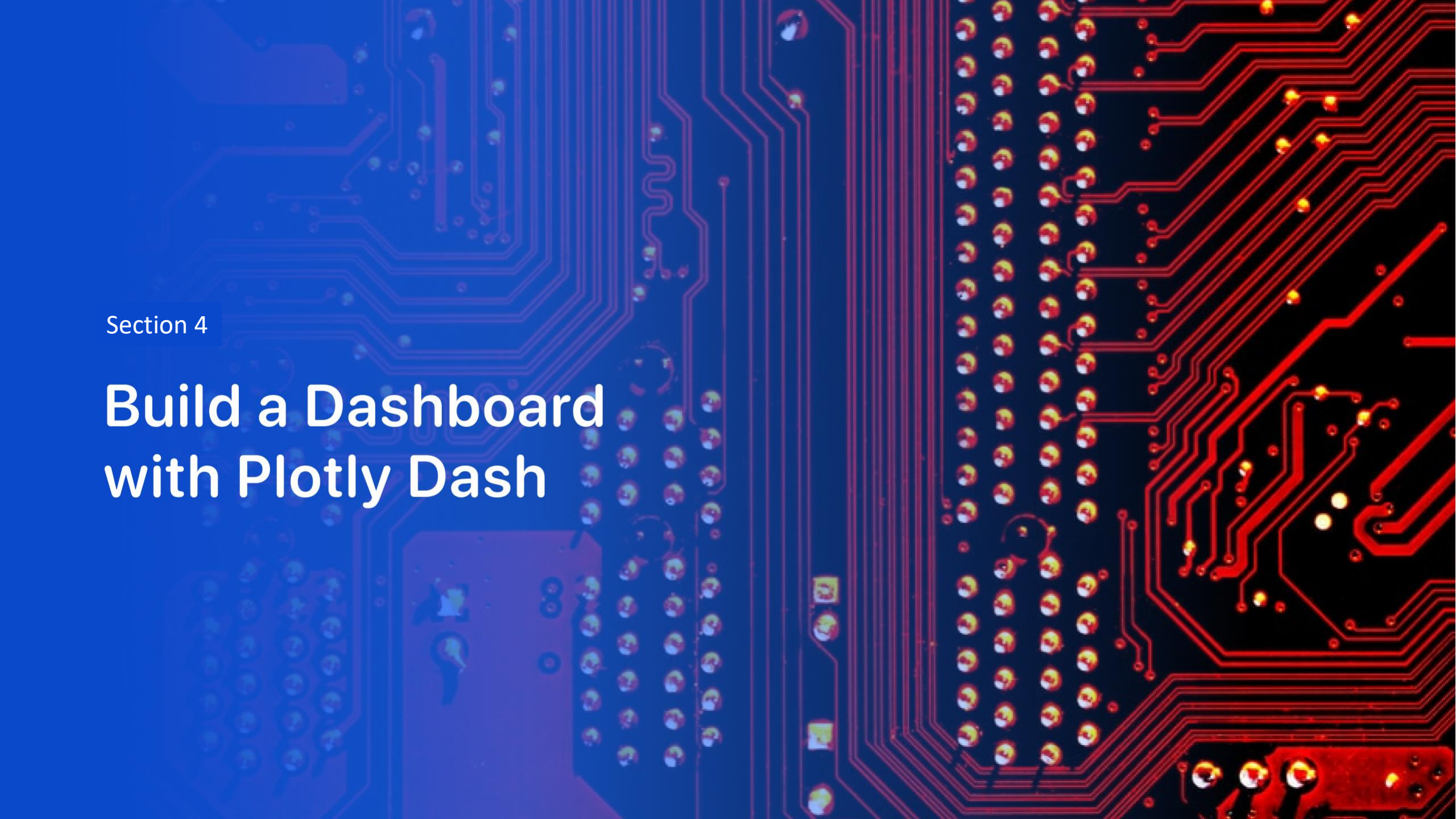


- Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely KSC LC39A

<Folium Map Screenshot 3>

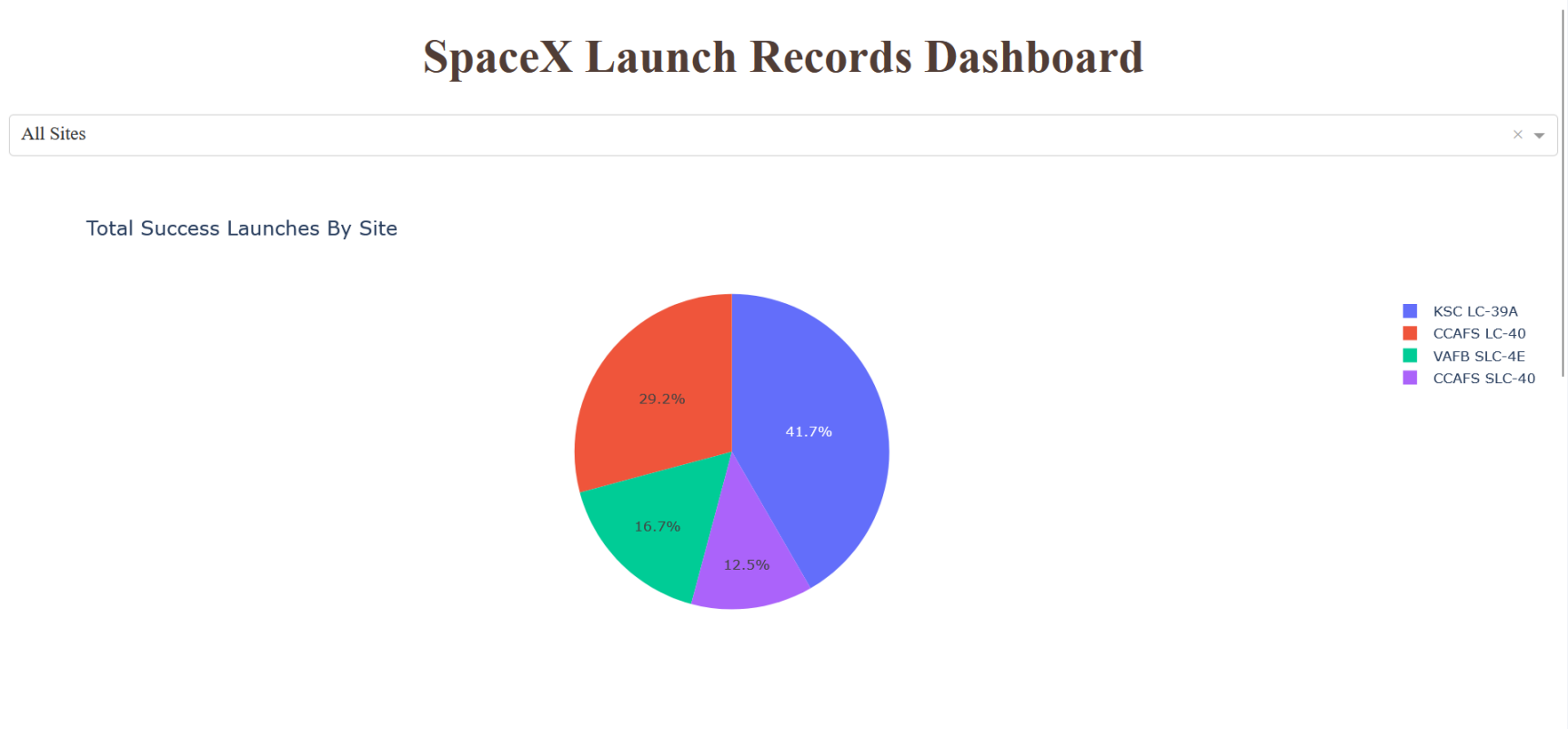


- Close to railways and highways for logistics and personnel reasons. Away from cities to minimize risk to the population and property damage.


Section 4
Build a Dashboard with Plotly Dash

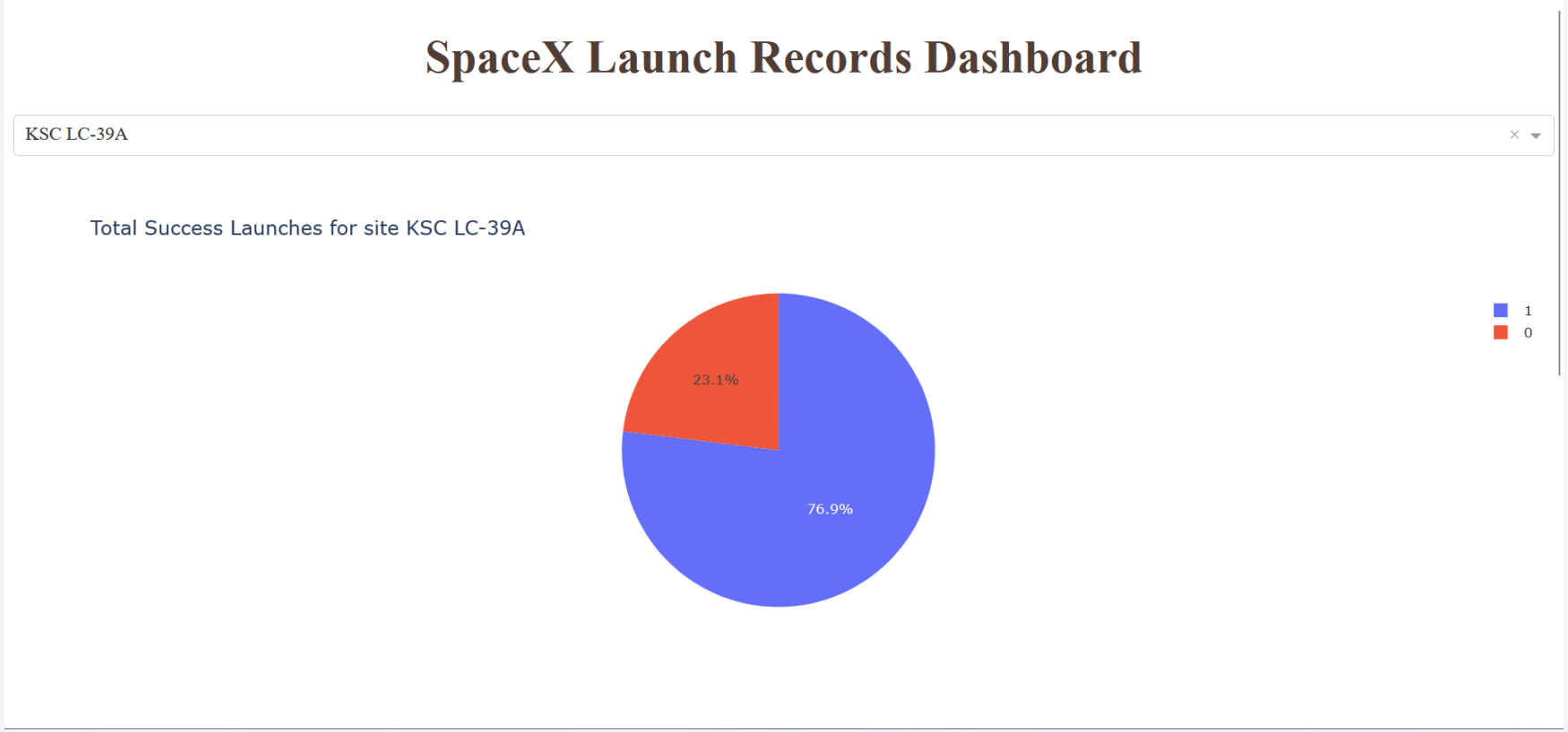
Launch success count for all sites

- KSC LC-39ª has the most successful launches

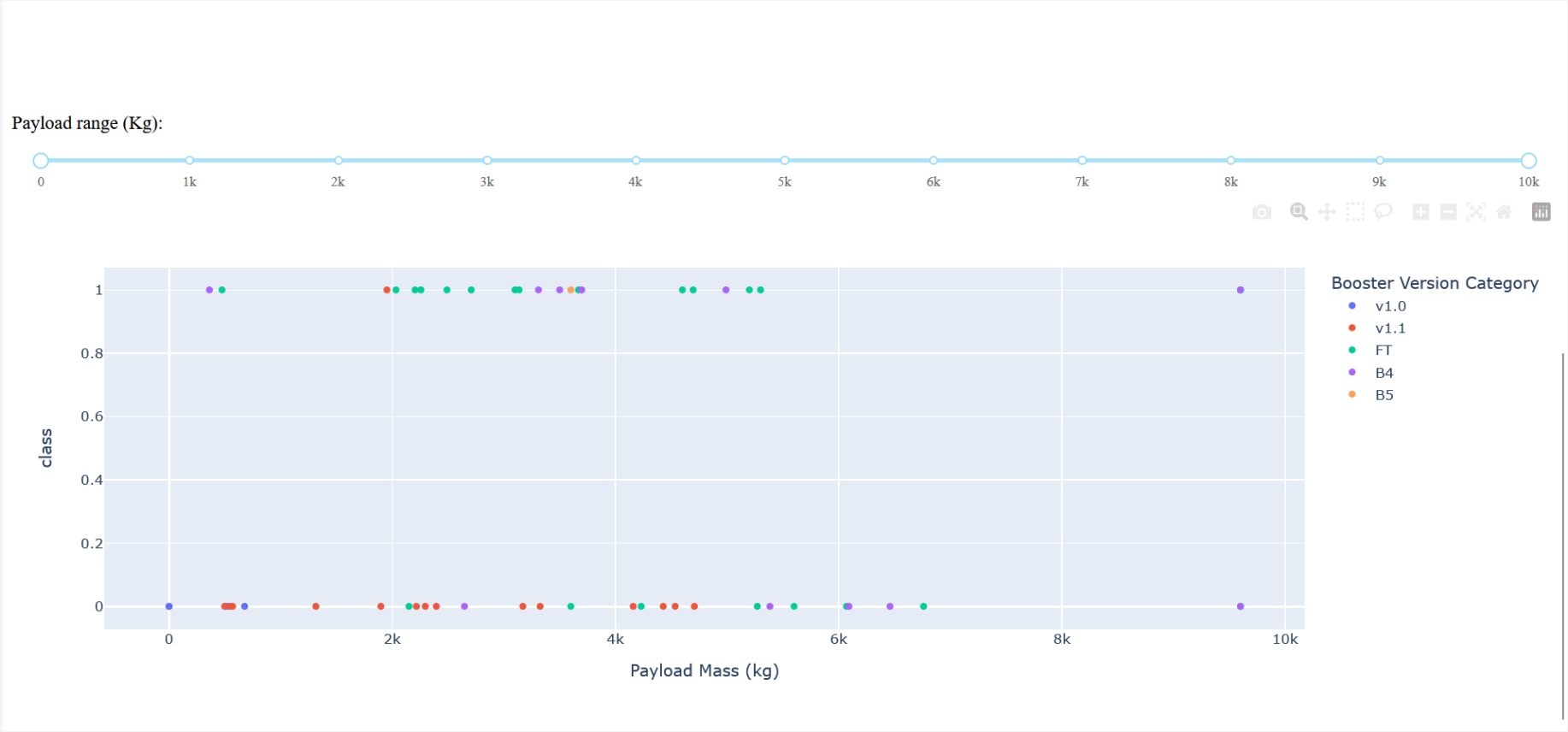


Highest launch success ratio site

- It has 23,1% of failure rate despite being the most successful site

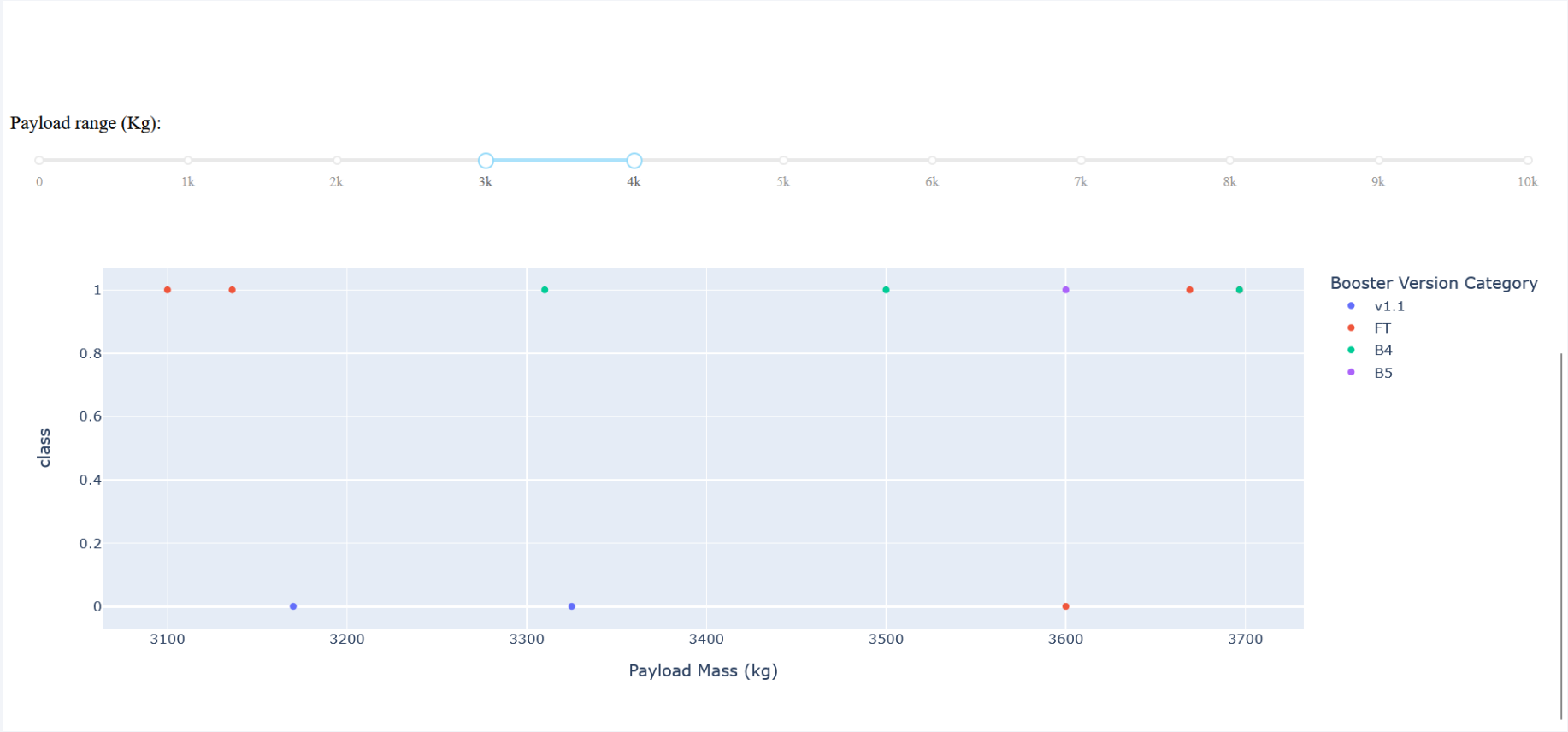


Launch Outcome Scatter Plot for All Sites



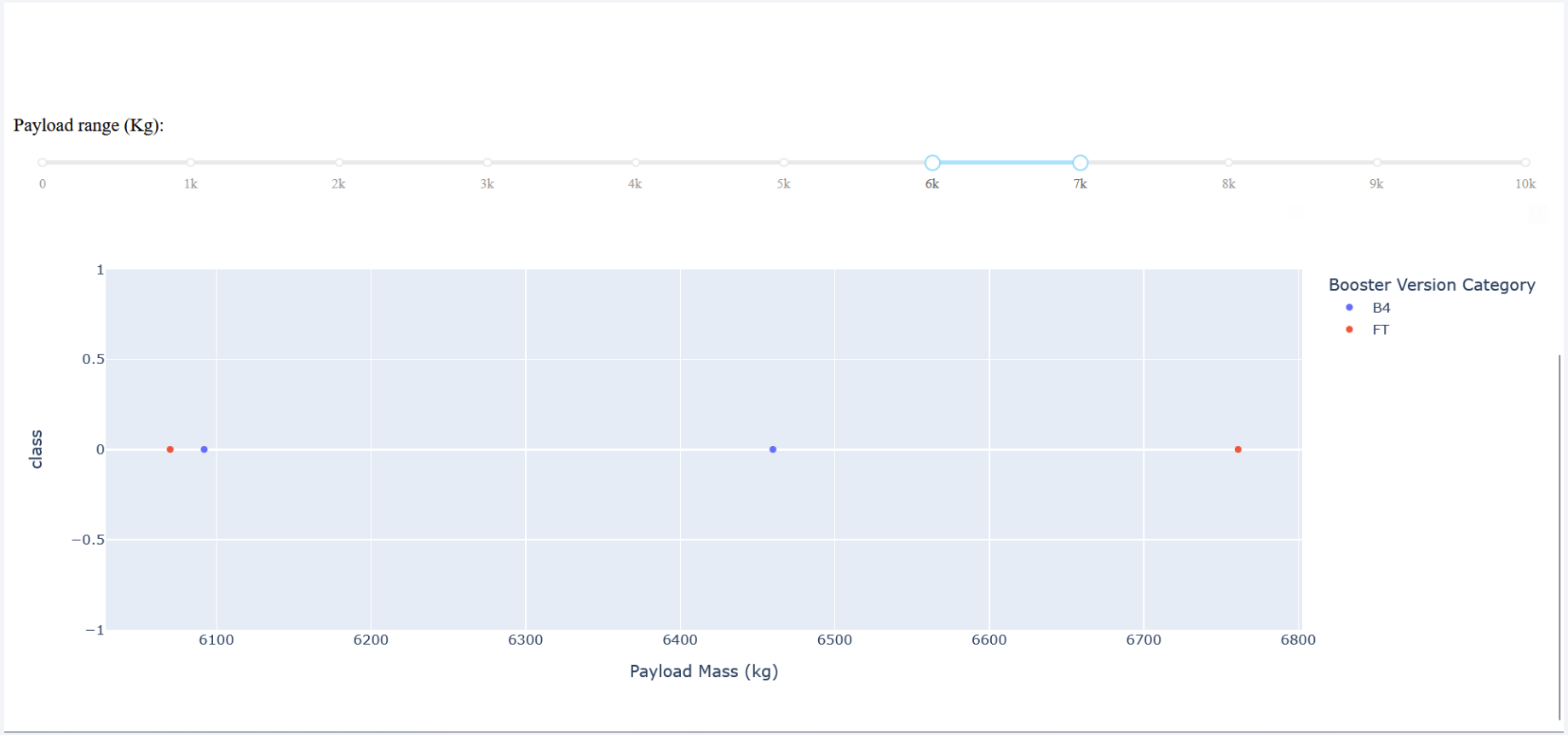
- Booster with the most successful landings is the ‘FT’ booster
- Booster with the most unsuccessful landings is the ‘v1.1’ booster

Launch Outcome Scatter Plot - Highest Success Rate



- Highest successful landing rate payload range (3-4k)

Launch Outcome Scatter Plot - Lowest Success Rate

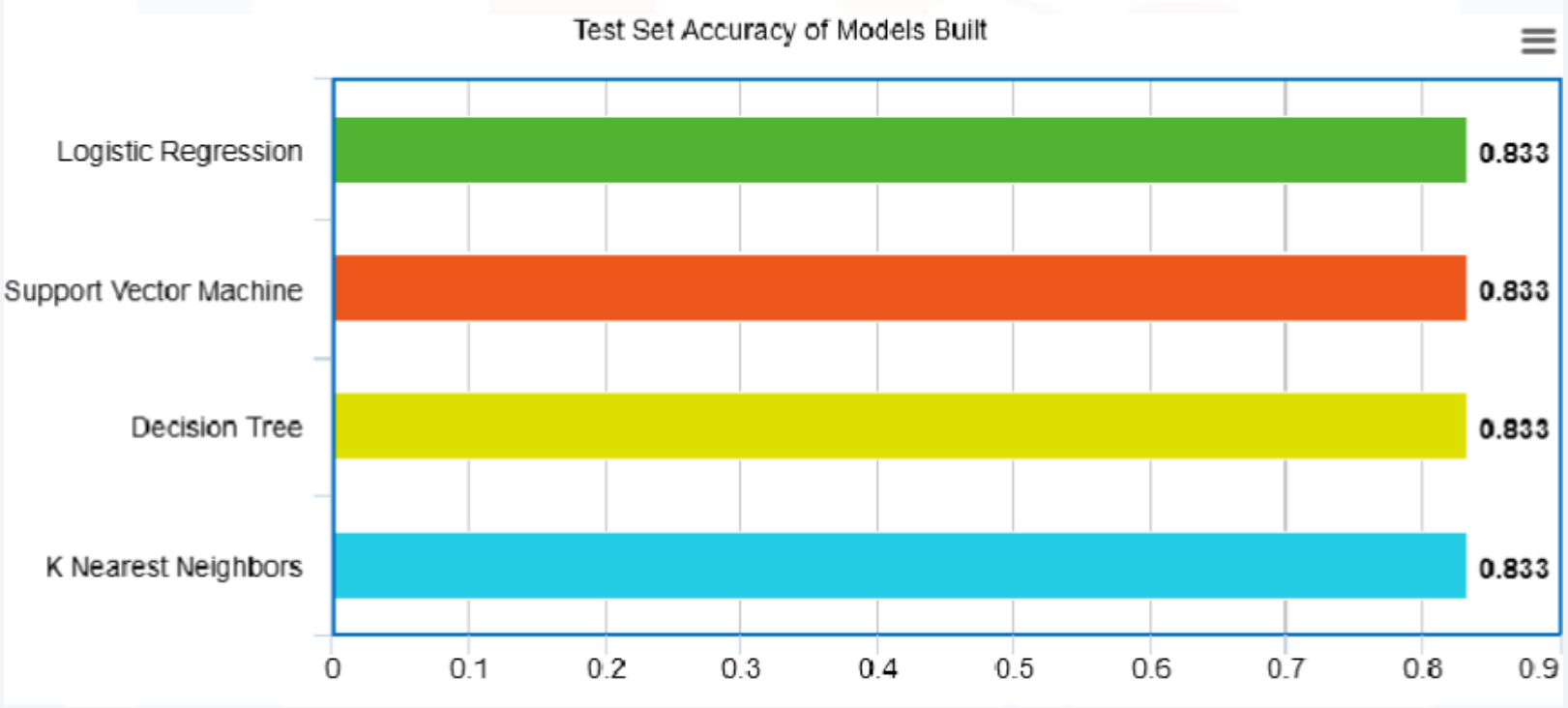


- Lowest successful landing rate payload range (6-7k)

Section 5

Predictive Analysis (Classification)

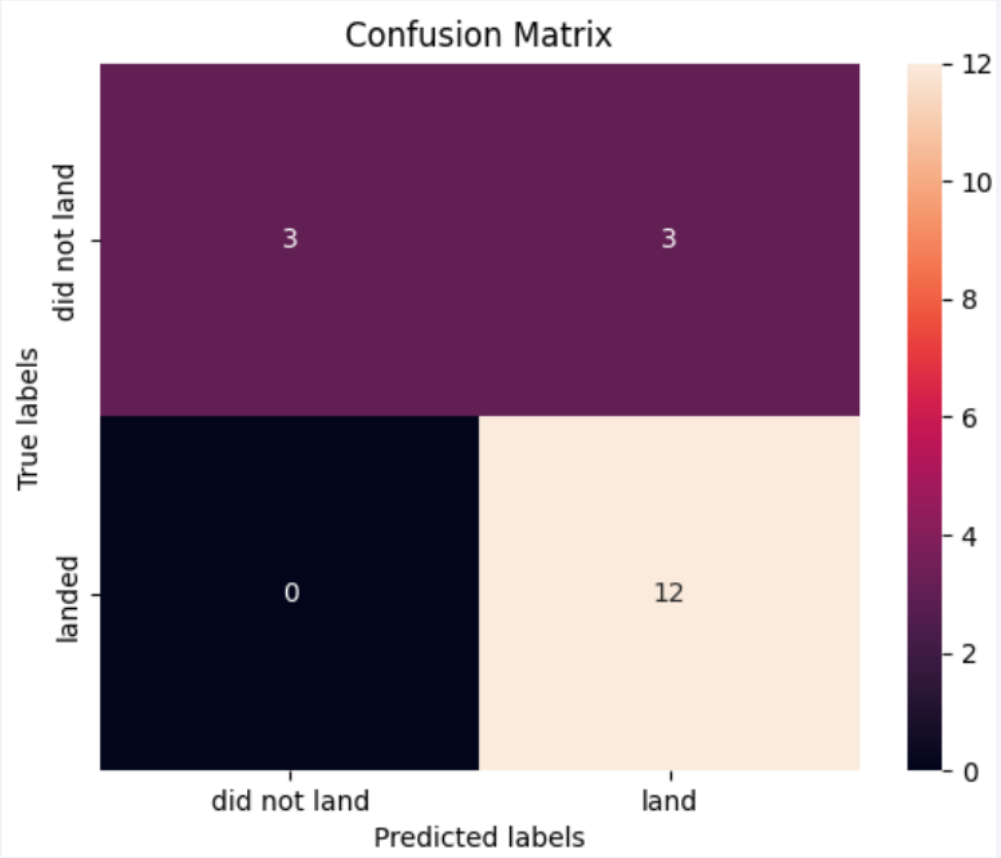
Classification Accuracy



- Each of the four models built came back with the same accuracy score, 83.33%

Confusion Matrix

- The confusion matrices of the best performing models (4-way-tie) are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set



Conclusions

- Using the models from this report Bspace can predict when SpaceX will successfully land the 1st stage booster with 83.3% accuracy
- SpaceX public statements indicate the 1st stage booster costs upwards of $15 million to build
- This will enable Bspace to make more informed bids against SpaceX, since they will have a good idea when to expect the SpaceX bid to include the cost of a sacrificed 1st stage booster
- With a list price of $62 million per launch, sacrificing the $15+ million 1st stage, would put the SpaceX bid at upwards of $77 million

Thank you!