Winning Space Race with Data Science

Ethan DeJongh
6-29-2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

This study examines data pertaining to SpaceX rocket launches and develops a model for predicting the success or failure of the landing of the reusable first stage. Exploratory Data Analysis (EDA) tools, including Python visualization tools and SQL queries, were employed to analyze how independent variables such as launch site and fuel mass correlated with success rate. Several machine learning algorithms were then developed for future predictions.

The EDA analysis shows that success rate is strongly influenced by launch site, flight number, orbit type, and payload mass. All four of the machine learning algorithms performed equally well, making successful predictions for 83% of launches in a test data sample, with all of the incorrect predictions being false positives.

Introduction

- One of the key innovations of SpaceX is the ability to land and reuse the first stage of its rockets. This greatly reduces the cost of operations and makes space travel a much more viable business. Successful landing is not guaranteed, however. As a competing company, it is imperative that we employ data science to study the variables that predict the success or failure of this approach, so that we may learn from the existing data when developing our own rockets.
- The main questions of this study are: what are the key variables in determining landing success rate? Can we use machine learning to predict whether a given landing will be successful?

Section 1

Methodology

Methodology

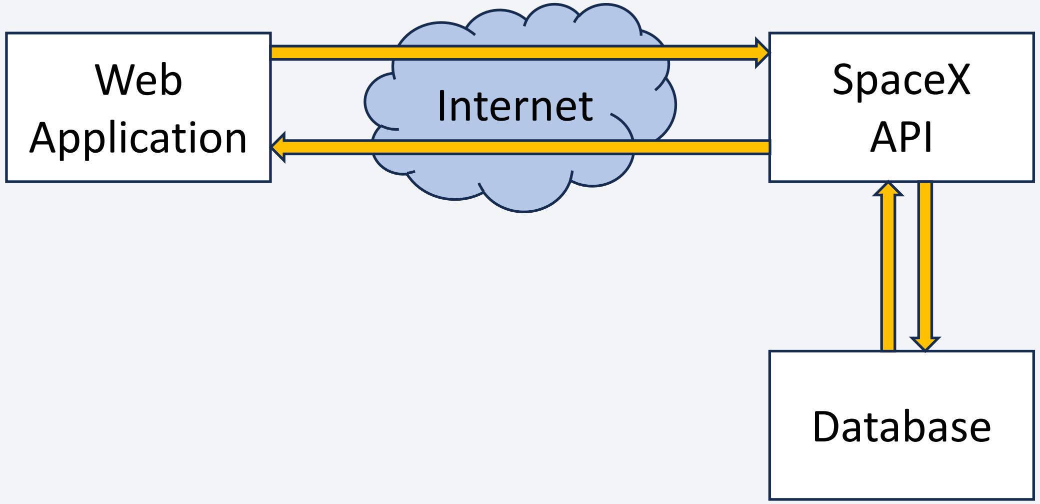
Executive Summary

- Data collection methodology:
  - Data was collected directly from SpaceX using API calls from Python, and from Wikipedia using web scraping
- Perform data wrangling
  - Useful variables were compiled into a single dataframe, along with a binary target variable representing success or failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - By splitting the data into a training and testing set, we could test four classification models – logistic regression, decision tree, Support Vector Machine (SVM), and k-nearest neighbor – and evaluate their accuracy.

Data Collection – SpaceX API

- Data was obtained from SpaceX by making an HTTP request to the API using Python's "requests" library.
- The successful request was decoded in json format and converted to a pandas dataframe:

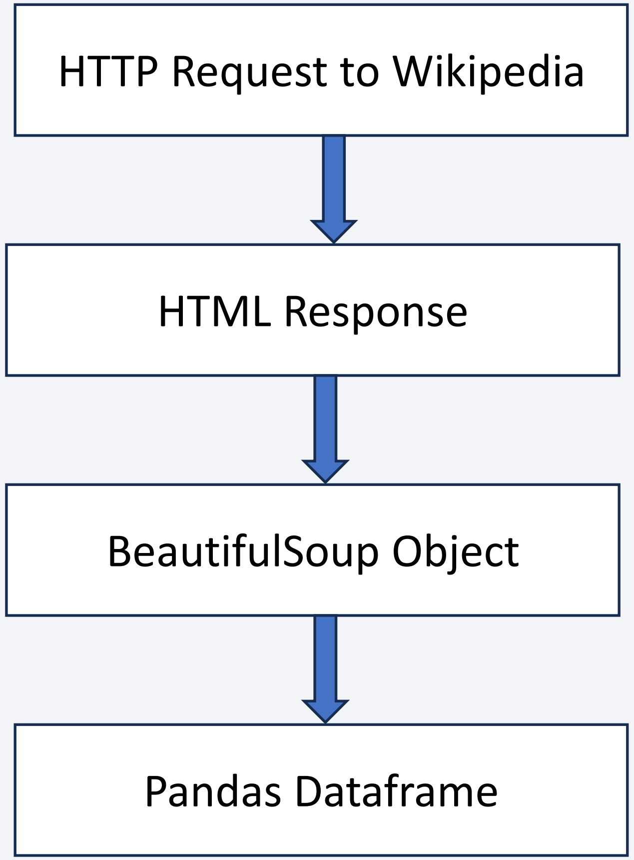
```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```



[Link to Data Collection Notebook]

Data Collection - Scraping

- Another source of data was the Wikipedia page for Falcon 9 rocket launches: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- An HTTP request was made to the website, and the text of the HTML response was saved in a BeautifulSoup object.
- The text was parsed to extract relevant data from the tables on the website and save it to a Pandas dataframe.



Link to Webscraping Notebook

Data Wrangling

- The first task in data wrangling was to filter the dataframe to include only Falcon 9 rocket launches.
- The next step was deal with missing values. Some rows in the dataframe were missing Payload Mass, and this was dealt with by inserting the mean value.
- Finally, the range of reported landing outcomes were simplified into a binary variable, indicating whether the landing was successful or not:

Outcome

True ASDS
None None
True RTLS
False ASDS
True Ocean
False Ocean
None ASDS
False RTLS

Class

0
1

[Link to Wrangling Notebook]

EDA with Data Visualization

To explore the relationships between variables, scatter plots were created for the following combinations, with the points color-coded for landing success or failure:

- Flight Number vs. PayLoad
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Flight Number vs. Orbit Type
- PayLoad vs. Orbit Type

In addition, a bar chart showed the success rate for each orbit type, and a line plot displayed success rate over time.

[Link to Data Visualization Notebook]

EDA with SQL

The following SQL queries were performed for additional EDA:

- Names of unique launch sites
- 5 records where launch sites begin with 'CCA'
- Total payload mass from boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1
- Date of first successful ground pad landing
- names of the boosters which have success in drone ship and have payload mass of 4000-6000
- Total number of successful and failure mission outcomes
- Names of booster versions which have carried the maximum payload
- List of records displaying 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.

[Link to SQL Notebook]

Build an Interactive Map with Folium

- Folium circles with pop-up labels were used to visualize the locations of launch sites on the map.
- Color-coded marker clusters were added to each launch site to visualize the number of successful and failed landings.
- Lines were drawn to display the distances between launch sites and various surrounding features such as coast lines and highways. This provides more context when evaluating the data by launch site.

Link to Folium Notebook

Build a Dashboard with Plotly Dash

The Plotly dashboard consists of two inputs and two visual outputs.

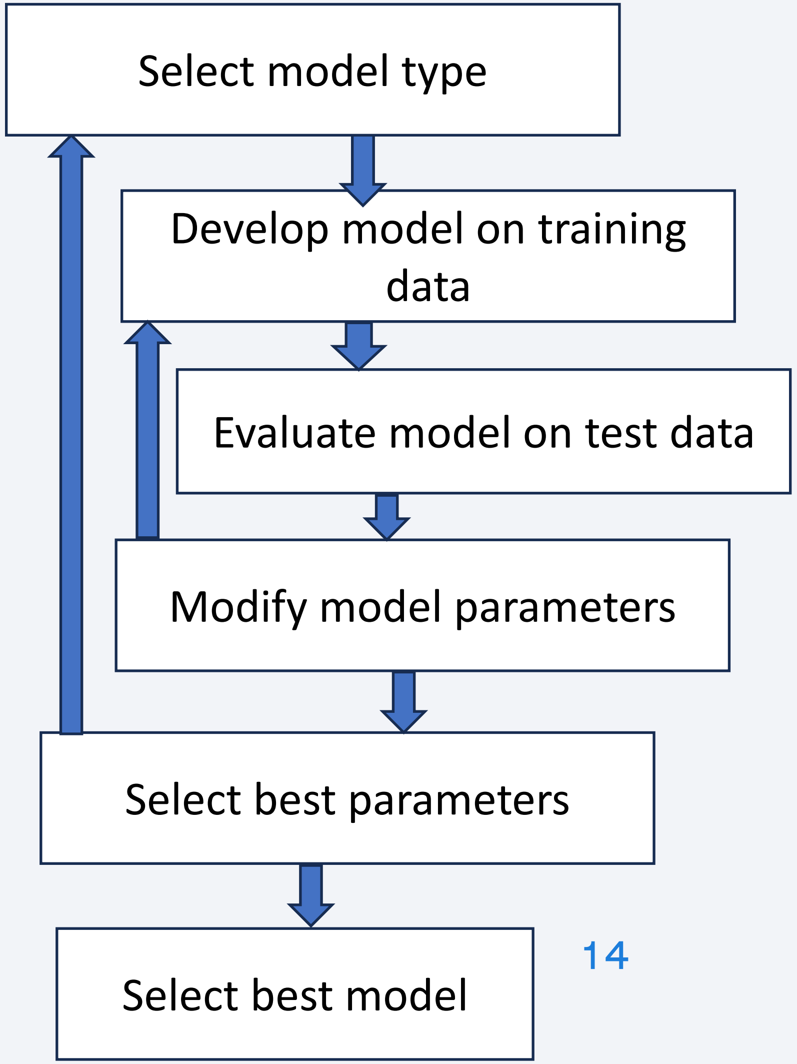
- A dropdown menu allows the user to select all launch sites or an individual site.
- A pie chart displays the landing success rate for the selected site, or the distribution of successful landings for all sites.
- A range slider allows the user to restrict the results to a specific range of payload mass.
- A scatter plot displays payload mass vs. success class, with points color-coded for booster version in order to explore the relationship between these variables.

Link to Dashboard File

Predictive Analysis (Classification)

- Four different classification models were tested:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbor
- Data were normalized and split into training and test sets.
- A grid search was performed on each model to find the best-performing parameters.
- A confusion matrix displayed the test results for each model.

[Link to Classification Notebook]



Results

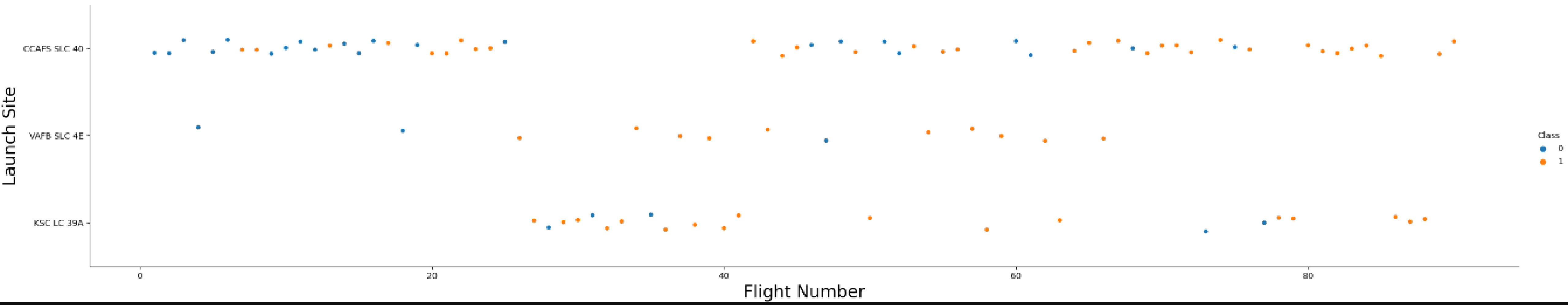
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

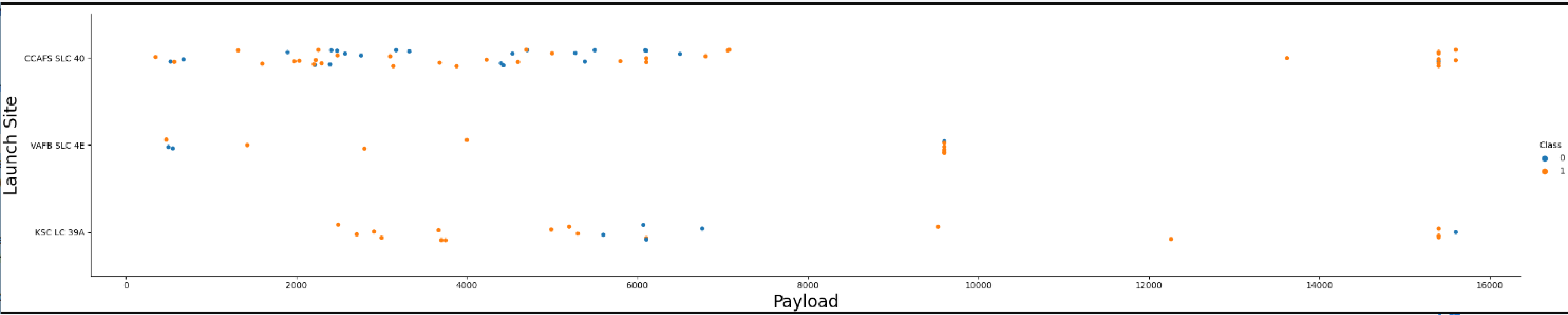
Flight Number vs. Launch Site

- Success rate seems to increase with flight number, regardless of launch site
- Most early flight numbers launched from CCAPS SLC 40
- The lower success rate at site CCAPS SLC 40 may be more due to flight number than any site-specific characteristics.



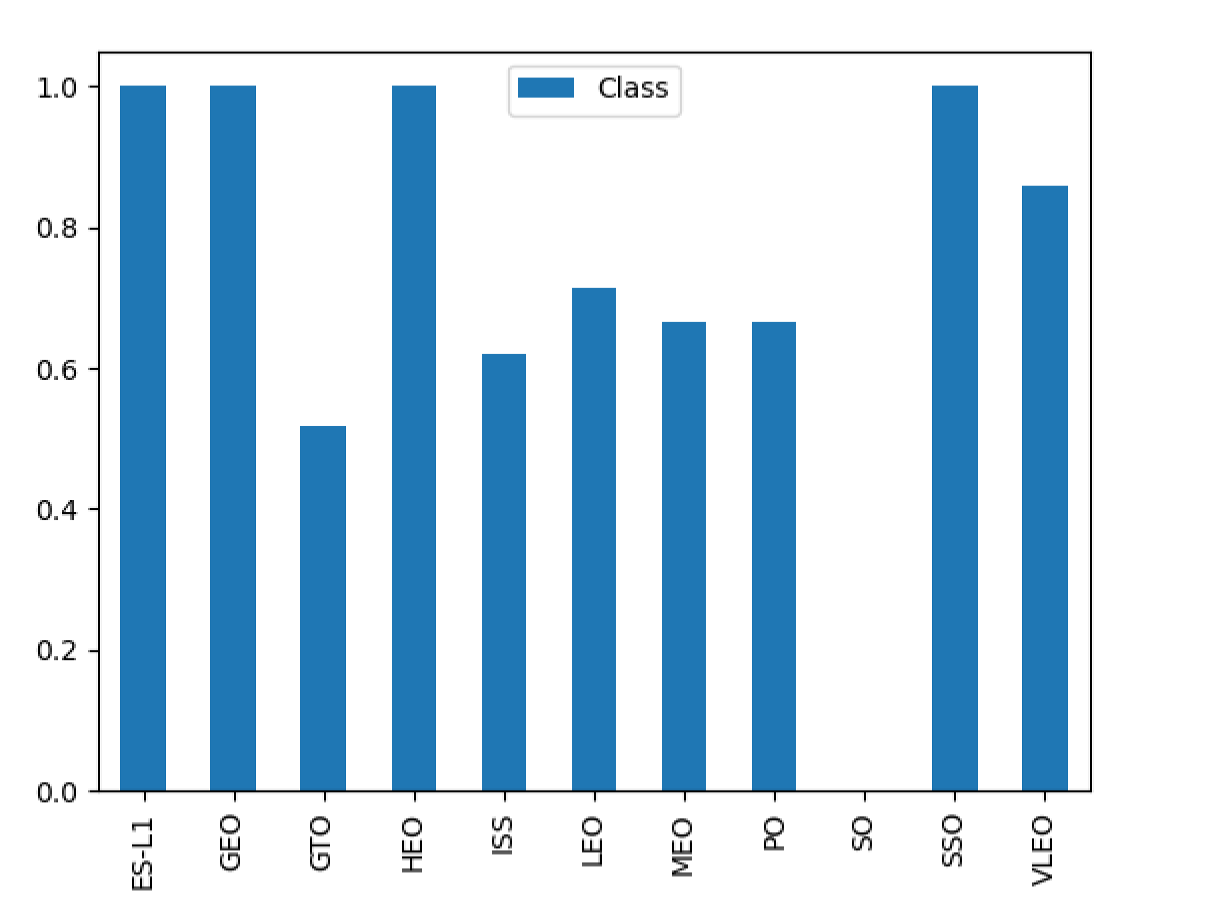
Payload vs. Launch Site

- For the first two launch sites, success rate appears to increase with payload.
- For the KSC site, the relationship is less clear.



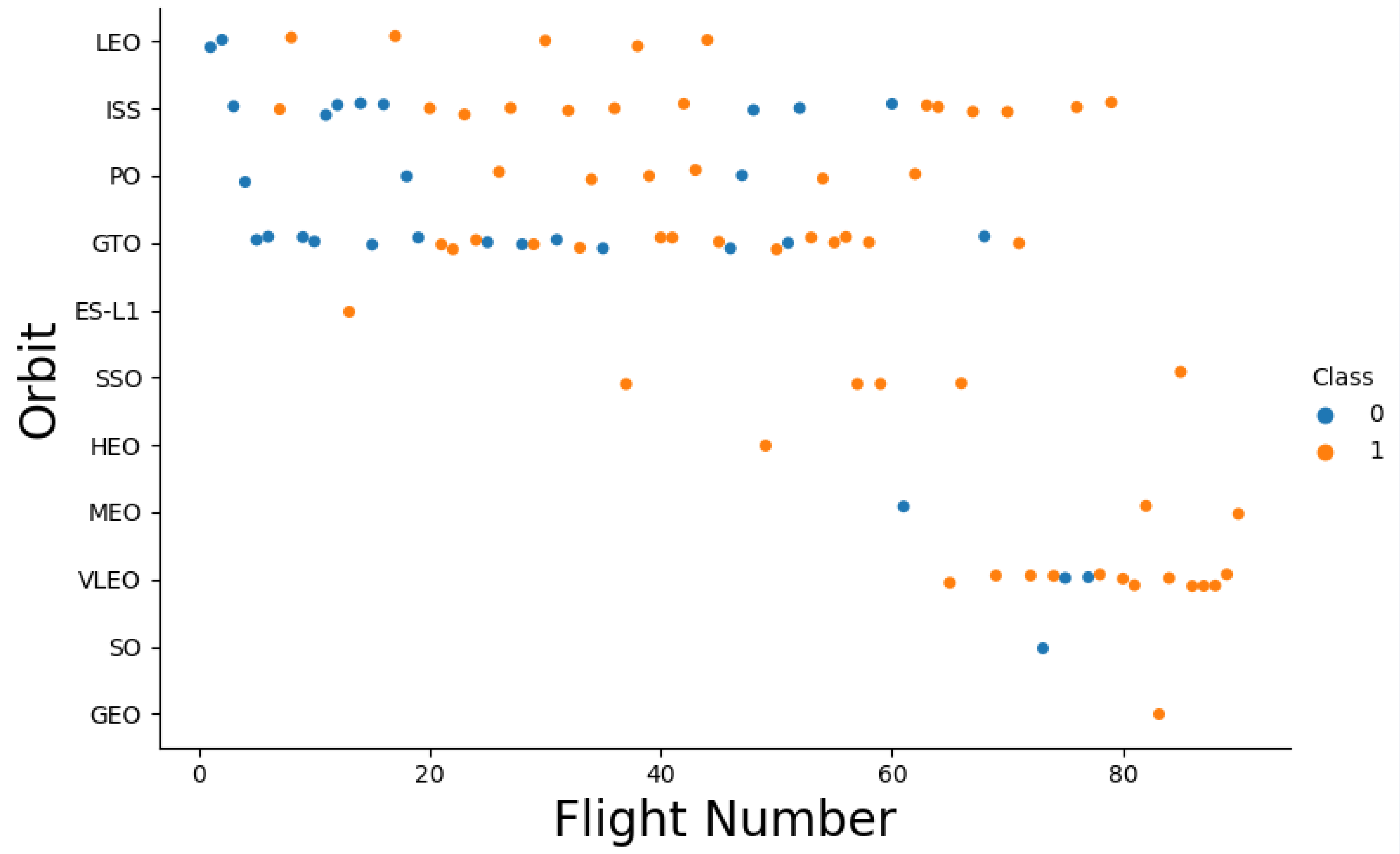
Success Rate vs. Orbit Type

- Four orbit types – ES-L1, GEO, HEO, and SSO – have 100% success rates.
- One orbit type – SO – has a 0% success rate.
- The other types are all at least 50%.



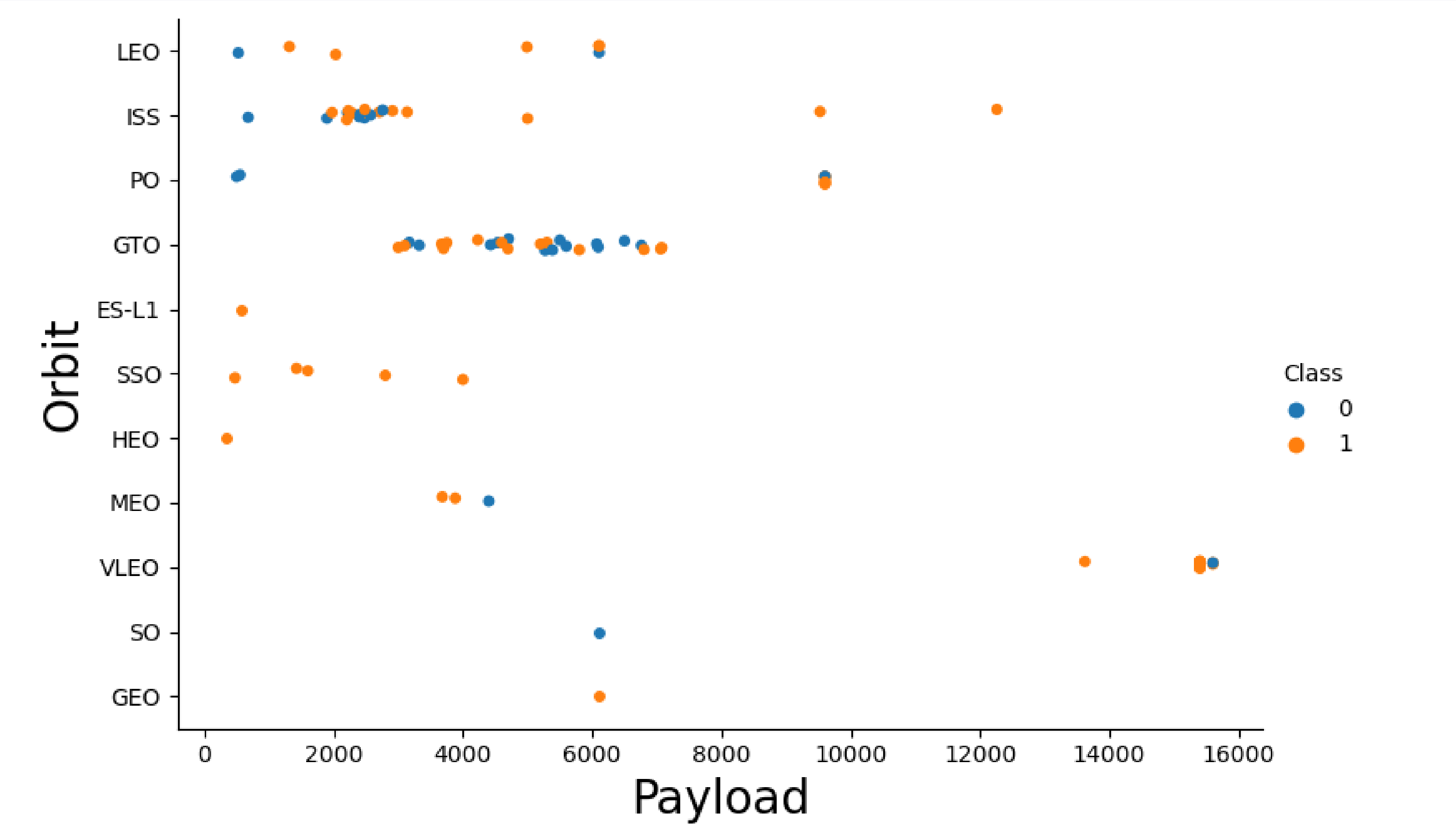
Flight Number vs. Orbit Type

- This plot adds context to the previous chart by showing that 3 of the orbit types with 100% success and the one with 0% success had only one flight each.
- The SSO and VLEO orbits have high success rates but were not used for early flights.



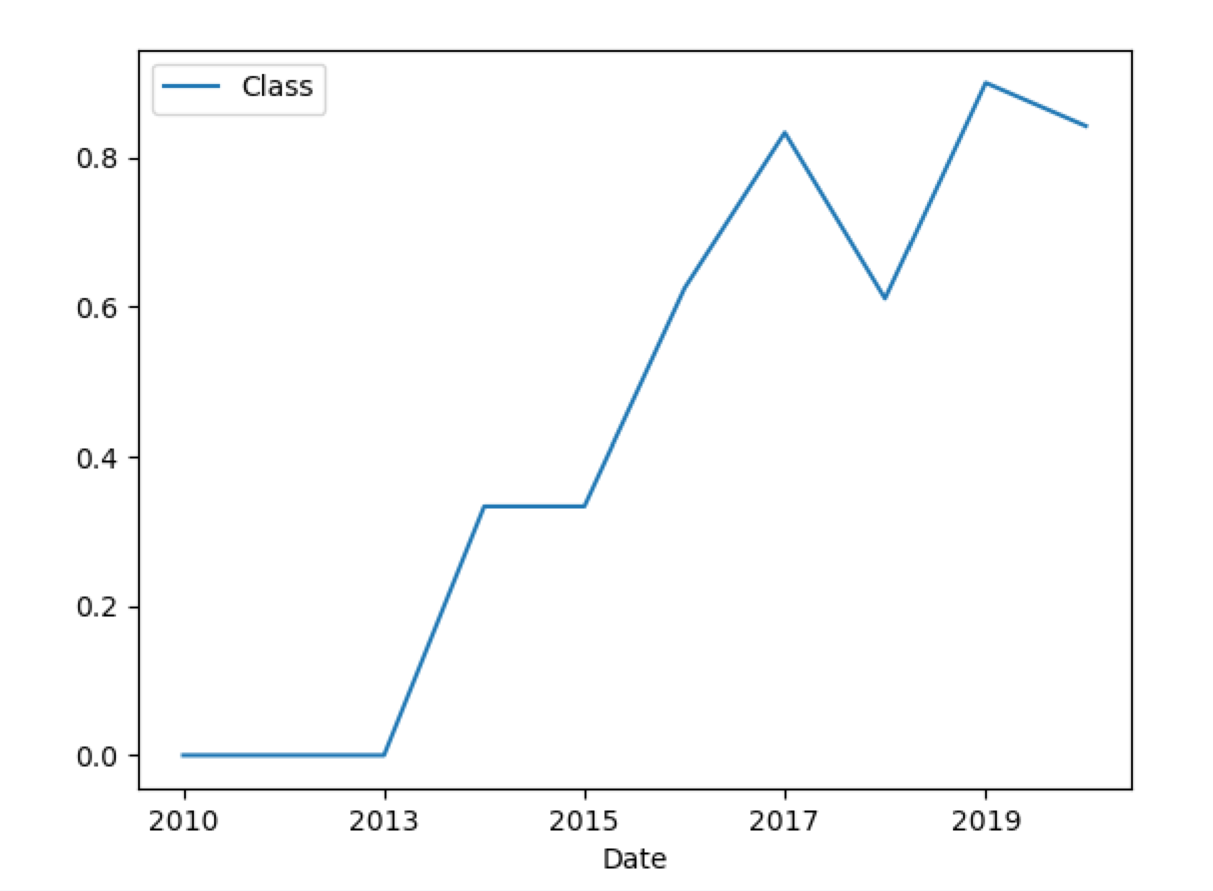
Payload vs. Orbit Type

- This plot indicates that some orbit types such as SSO and MEO require small payloads, while VLEO requires a large payload.
- Orbit types LEO, ISS, and PO have a large range of payloads, with higher success at larger payloads.



Launch Success Yearly Trend

- Success rate has clearly improved over time.
- There is a noticable dip in 2018 that should be investigated.



All Launch Site Names

- This query reveals four unique launch sites, including two from Cape Canaveral.

```
%%sql
select distinct(Launch_Site) from spacextbl;
```

```
* 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'

- These first five records are all from the CCAFS LC-40 site
- All five failed to land

```
%%sql
select * from spacextbl
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_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
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	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- The total payload carried by boosters from NASA is 45,596 kg.

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

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

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.

```
%%sql
select avg(PAYLOAD_MASS__KG_) from spacextbl
where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 12-22-2015.

```
%%sql
select Date from spacextbl
where Landing_Outcome = 'Success (ground pad)'
limit 1;
```

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

Date
22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four booster versions have successfully landed on a drone ship and had payload mass of 4000-6000 kg.

```
%%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.
```

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

Total Number of Successful and Failure Mission Outcomes

- In total, there were 100 successes to 1 failure.
- Rate of success on mission outcome is much higher than rate of success on landing the rocket.

```
%%sql
select Mission_Outcome, count(*) from spacextbl
where (Mission_Outcome like '%Success%') or (Mission_Outcome like '%failure%')
group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- A subquery was needed to determine the maximum payload.
- 12 distinct booster versions carried the maximum.

```
%%sql
select distinct(Booster_Version) from spacextbl
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl);
```

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

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

2015 Launch Records

- There are 7 launch records for the year 2015.
- All launches for this year used the site CCAFS LC-40.
- The first five landing attempts were failures, followed by two successes.

```
%%sql
select substr(Date, 4, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from spacextbl
where Date like '%2015%'
order by Month;
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- There are more successes than failures during this period.
- In most failure cases, a landing wasn't attempted.

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

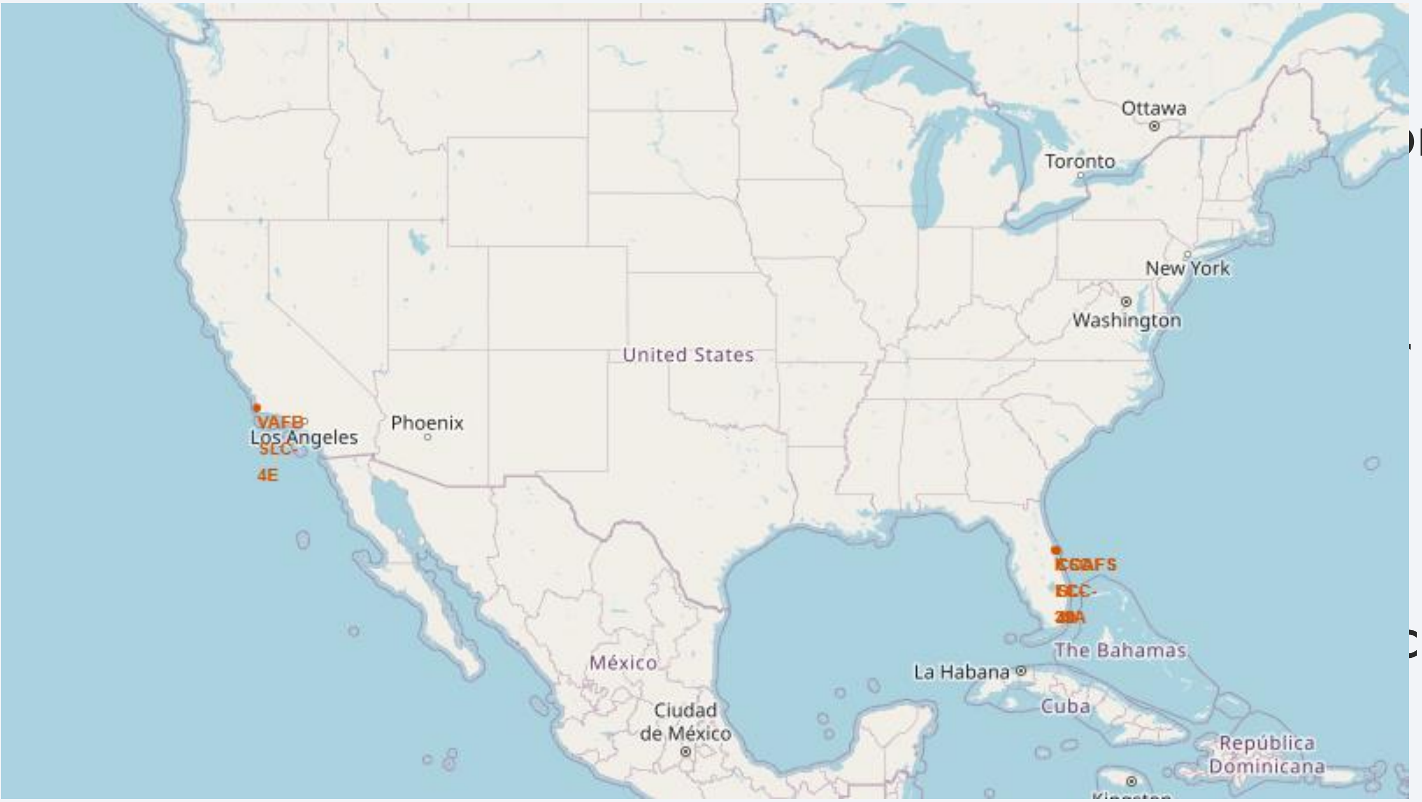
```
 * sqlite:///my_data1.db
Done.
```

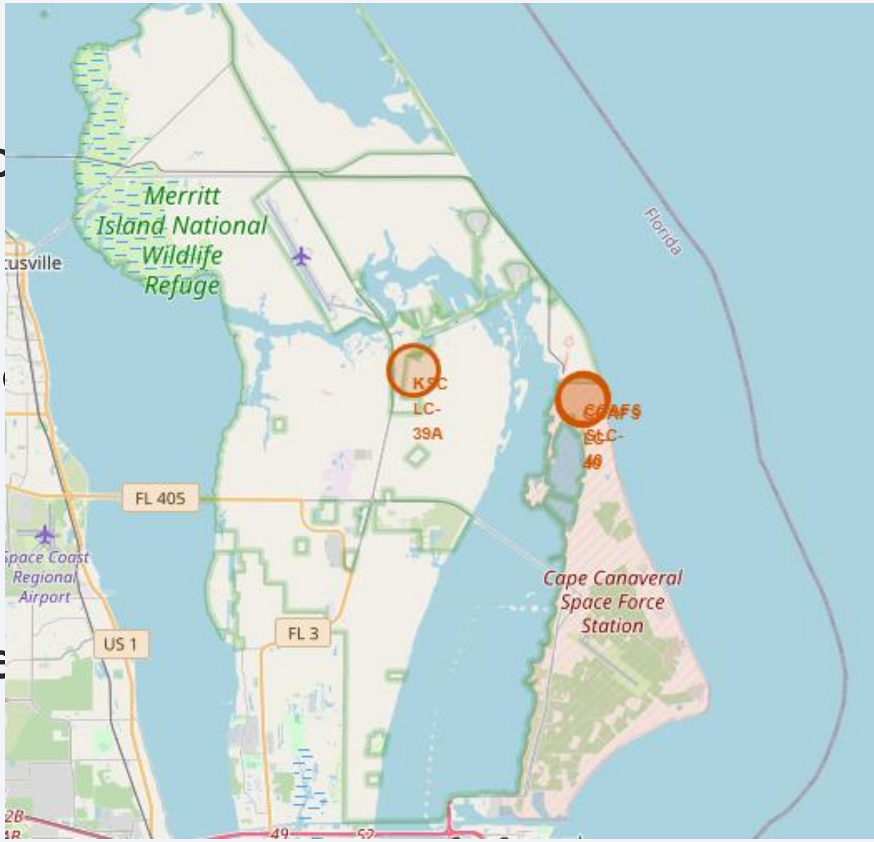
Landing_Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

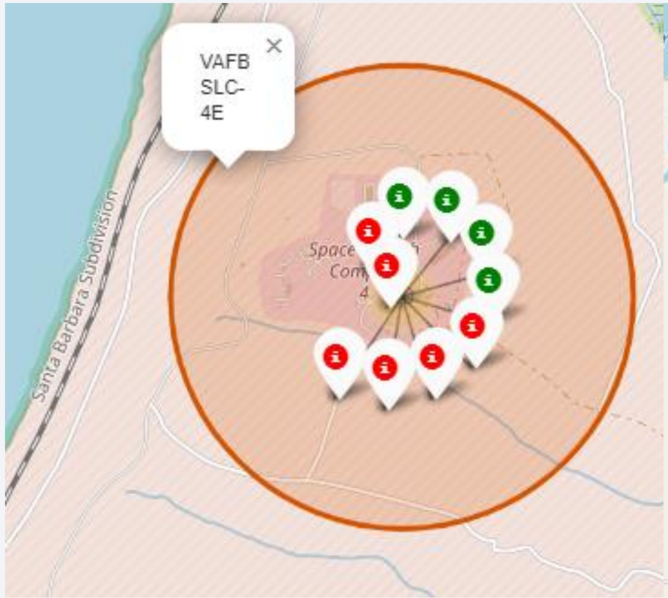
Launch Site Locations

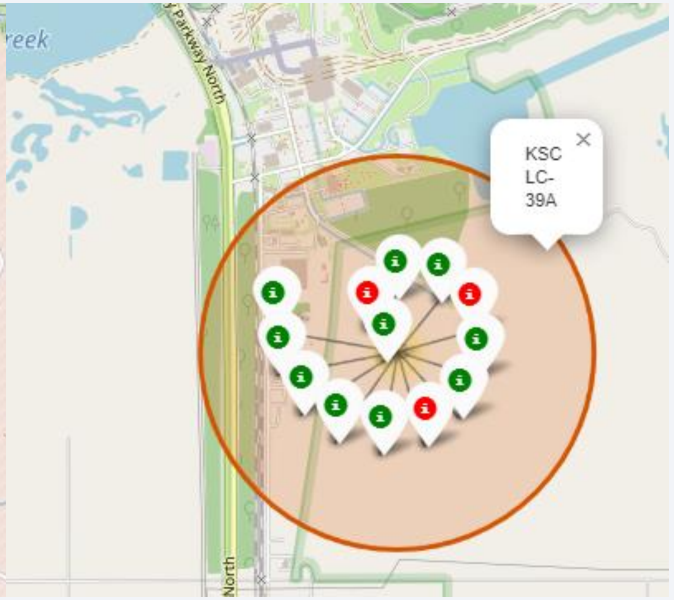


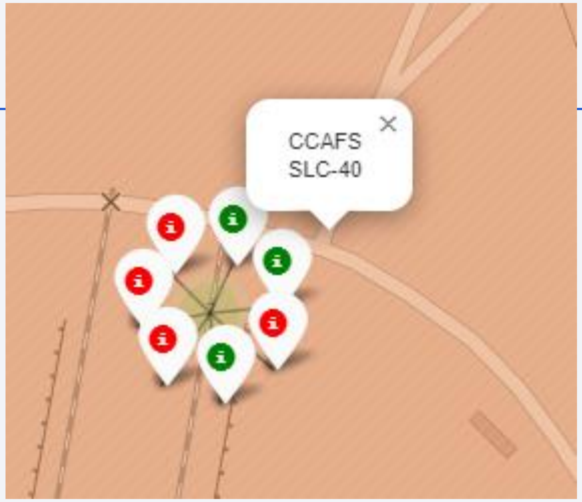


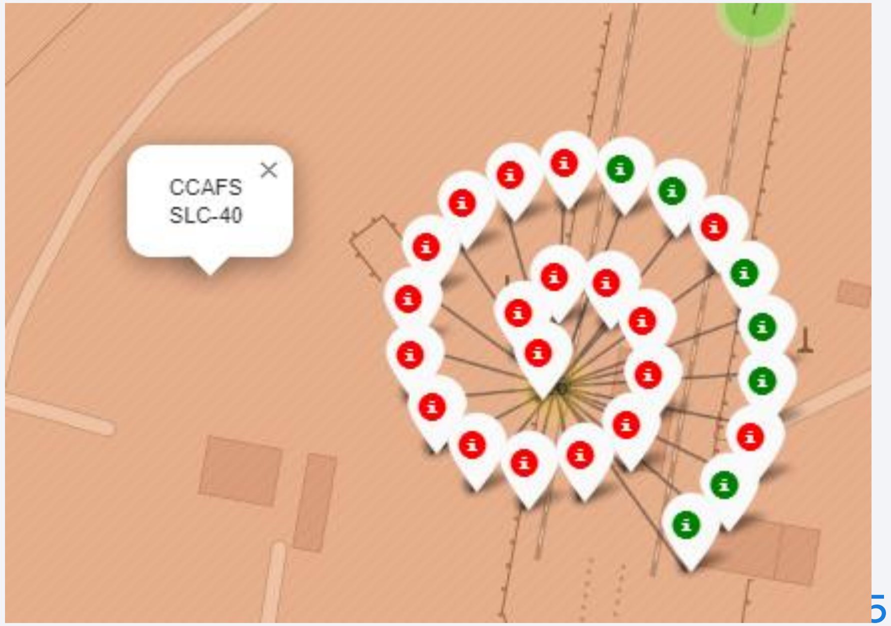
- All four launch sites are located on a coast near the southern end of the United States.
- Three of four sites are on the east coast, while only one is on the west coast.

Color-coded launch outcomes





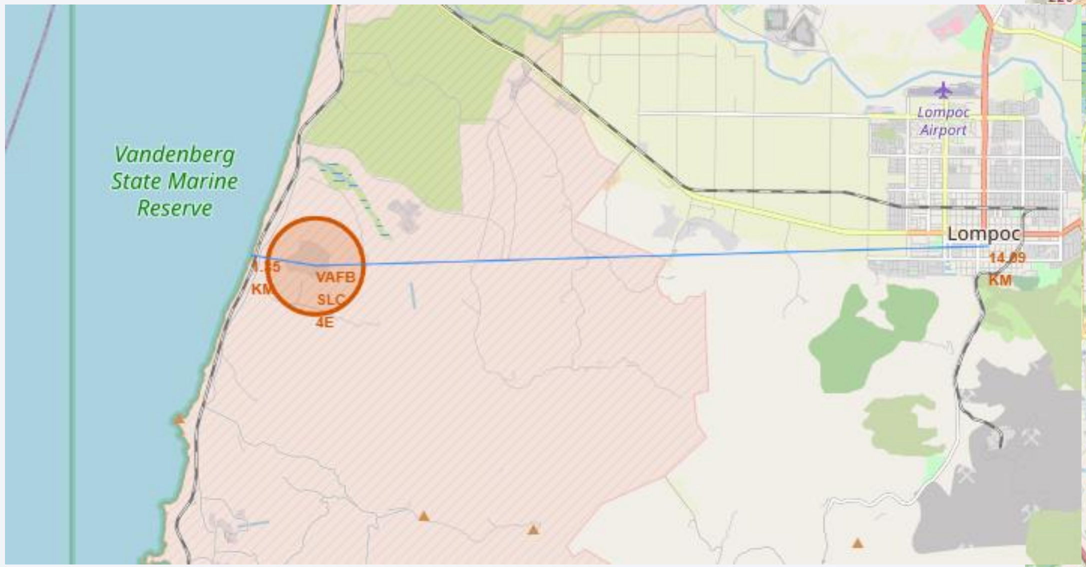


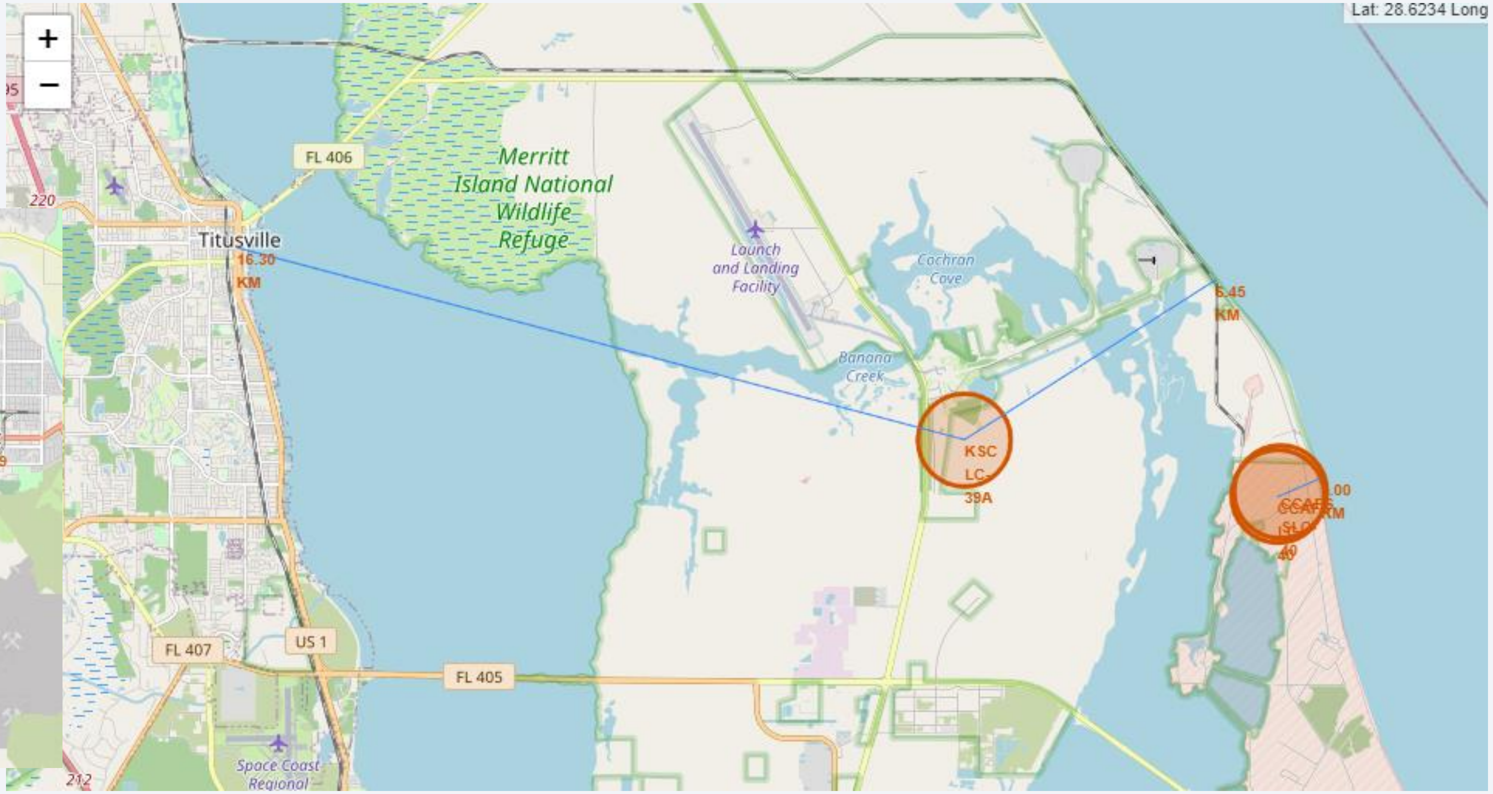


- Folium markers visualize the number of successful and failed landings at each site
- The KSC site is clearly the most successful.

Distance from coastlines and cities

- All of the sites are a similar distance from the closest city
- KSC stands out as being farthest from the coast. This could explain its landing high success rate, as more difficult landings aren’t attempted here.



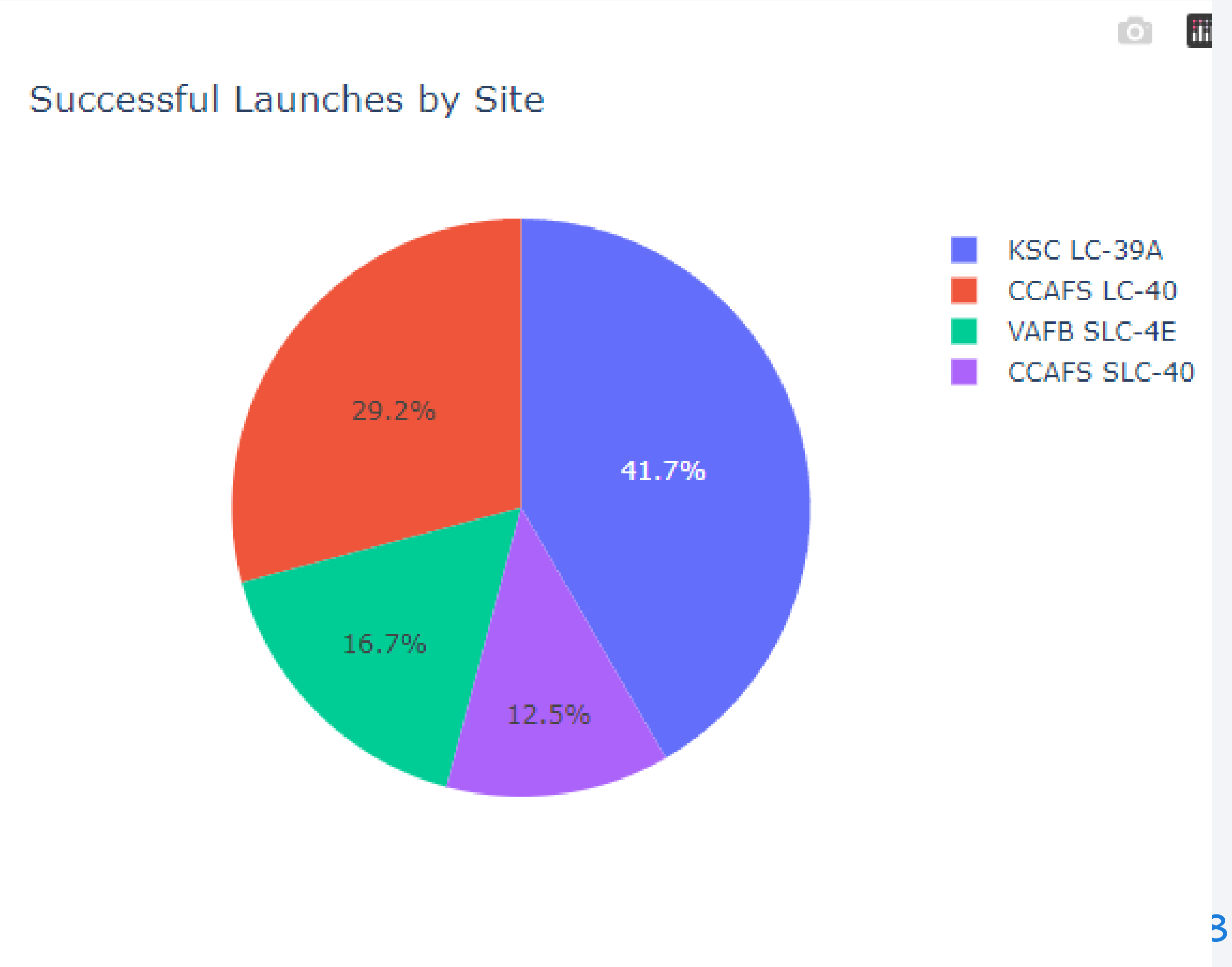


Section 4

Build a Dashboard with Plotly Dash

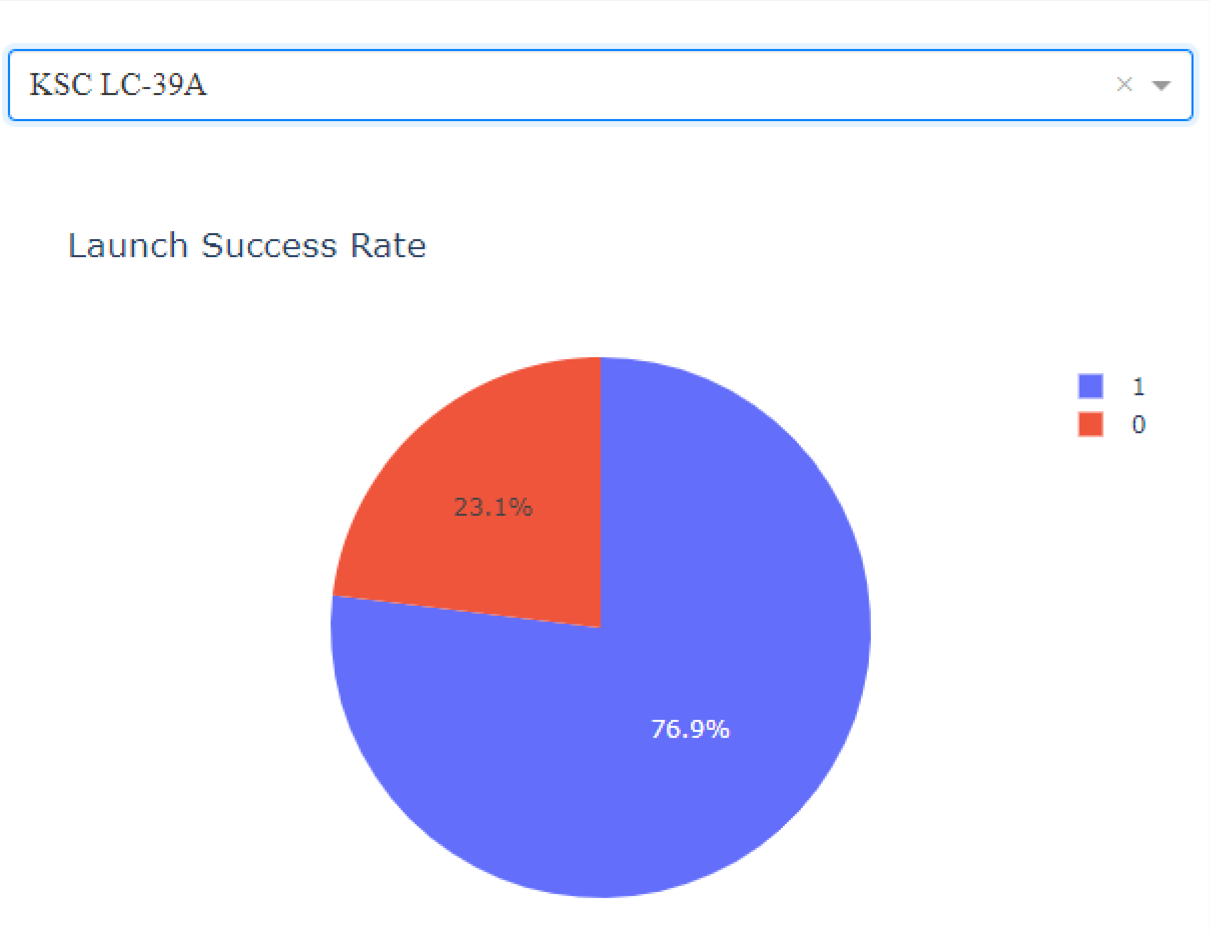
Dashboard Results: Successful Launches by Site

- A plurality of successful launches came from the KSC LC-39A site.
- The fewest number of successful launches came from CCAFS SLC-40.

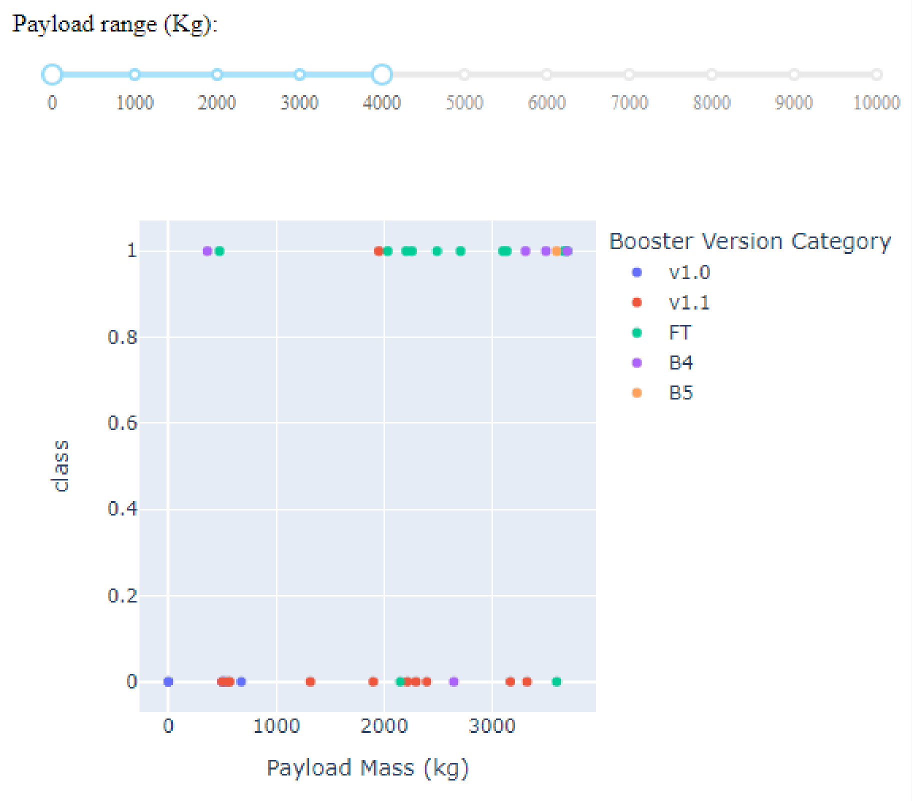


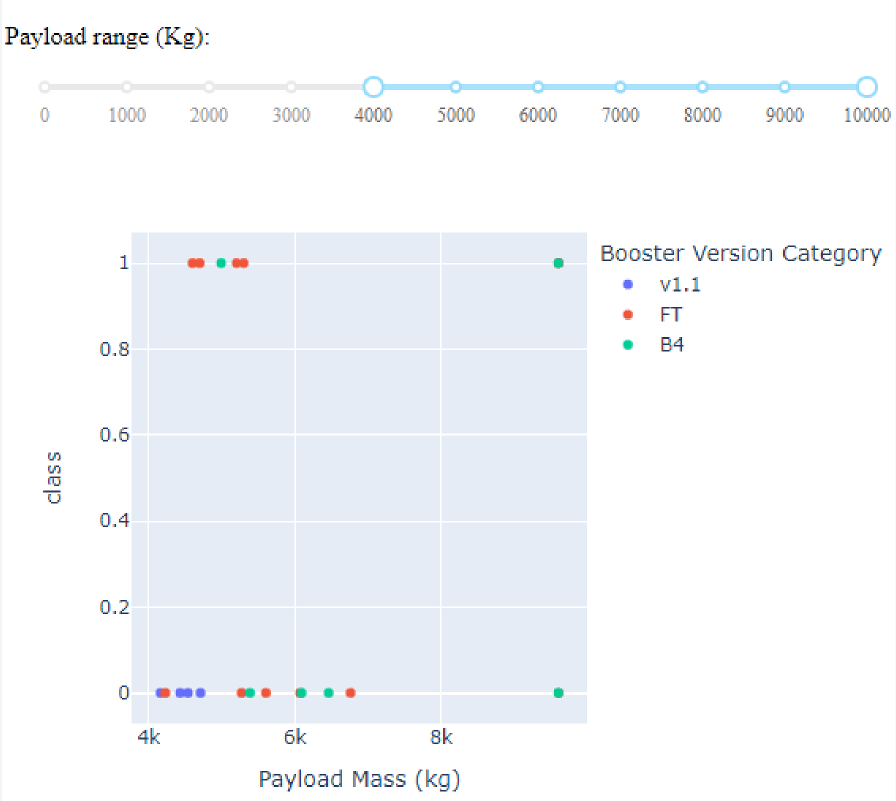
Highest Success Rate

- KSC LC-39A had the highest success rate among the four sites.



Outcome vs. Payload for Different Booster Versions





- These plots show that success rate is much higher for payloads under 4000 Kg.
- Only two booster versions – FT and B4 – have had successful outcomes for payloads above 4000 Kg.

Section 5

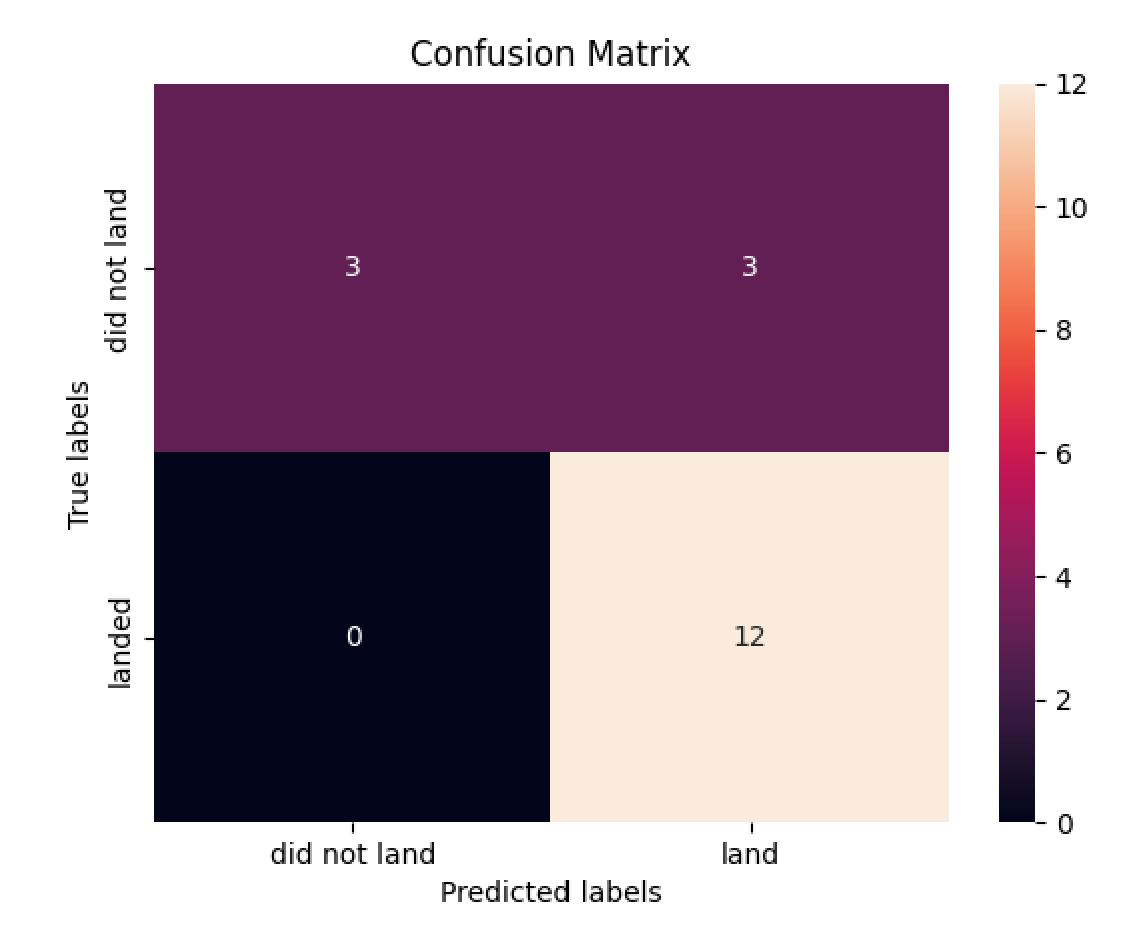
Predictive Analysis (Classification)

Classification Accuracy

- All four classification models achieved an identical accuracy of 83%, correctly predicting 15/18 outcomes from the test data.

Confusion Matrix

- All four models resulted in the same confusion matrix.
- The models correctly predicted 100% of successful landings, but only 50% of failed landings, suggesting that false positives are currently the largest issue.



Conclusions

- Exploratory data analysis shows that the success rate of SpaceX rocket landings has improved over time.
- There is a strong correlation between payload mass and success rate, with larger payloads being less likely to be successful.
- Rockets launched from KSC LC-39A are much more likely to successfully land than others.
- No particular machine learning algorithm stands out as more effective than others.
- Classification models are reasonably successful, but struggle with false positives. The dataset is still relatively small, so more data will allow us to evaluate the models more effectively.

Appendix

- Multiple Folium circles and markers were defined using a for loop and the “iloc” function to retrieve launch site labels and coordinates from the dataframe:

```
# Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)
# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
circles = []
markers = []
for i in range(0,launch_sites_df.shape[0]):
    nasa_coordinate = [launch_sites_df.iloc[i][1], launch_sites_df.iloc[i][2]]
    circles.append(folium.Circle(nasa_coordinate, radius=1000, color='#d35400', fill=True).add_child(folium.Popup(launch_sites_df.iloc[i][0])))

    markers.append(folium.map.Marker(
        nasa_coordinate,
        # Create an icon as a text label
        icon=DivIcon(
            icon_size=(20,20),
            icon_anchor=(0,0),
            html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>' % launch_sites_df.iloc[i][0],
            )
        ))

site_map.add_child(circles[0])
site_map.add_child(circles[1])
site_map.add_child(circles[2])
site_map.add_child(circles[3])
site_map.add_child(markers[0])
site_map.add_child(markers[1])
site_map.add_child(markers[2])
site_map.add_child(markers[3])
```

Thank you!