IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrew Rose
29 October 2023

Outline

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

Executive Summary

- In order to determine if a rocket business competing with SpaceX would be viable, are tasked with using data science to see if we can predict a successful first stage landing.
- Through collecting and web-scraping available data from SpaceX API and the wikipedia page on Falcon 9 launches, we're able to create a usable dataset.
- The dataset is data-wrangled by filtering, filling gaps and converting categorical to numerical data for ease of data analysis.
- We analyze the data through several means including, plot visualization and SQL queries. A dashboard app and Folium map are created to give external user access to some of the data. Possible indicators for successful landings are identified, including payload mass, launch site, flight number and booster type.
- Using predictive analysis, we identify no single classification model more accurate than the others. We do get an accuracy of 0.83, indicating good prediction of successful landings, with some false positives for failed landings. Obtaining a reliable model will be a necessary stepping stone in competing with SpaceX.

Section 1

Methodology

Introduction

- In the rapidly evolving space launch industry, the cost of rocket launches plays a pivotal role in decision-making for both space agencies and commercial entities.
- SpaceX has disrupted the industry by offering Falcon 9 rocket launches at a fraction of the cost compared to rival companies.
- SpaceX's ability to reuse the first stage of its Falcon 9 rockets is a game-changer, allowing them to offer launches at $62 million, in stark contrast to competitors whose costs soar upward of $165 million per launch.
- In this presentation we will be evaluating the success rate of Falcon 9 first stage landings. In doing so we will answer if a competing company can challenge the rocket launch cost efficiency that SpaceX boasts.

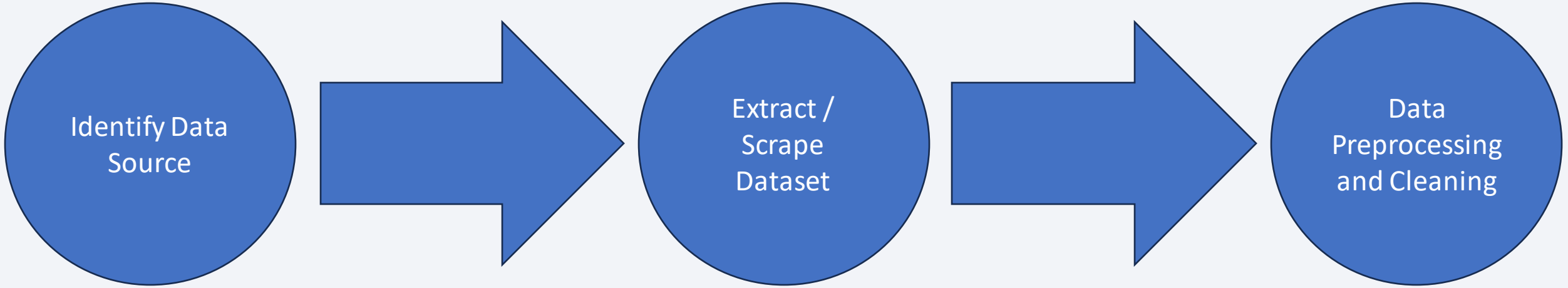
Methodology

Executive Summary

- Data collection methodology:
  - Web Scraping: Extract Falcon 9 and metrics from SpaceX API and HTML tables on the Falcon 9 Wikipedia page.
- Perform data wrangling
  - Data Transformation and Conversion: Parse and organize extracted data into a dataframe. Fill missing data in the dataframe with the average value where appropriate. Transform categorical data into numerical where appropriate.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification Models: Build multiple regression models including logistic regression, support vector machine (SVM), decision tree classifier, and k-nearest neighbors (KNN). Identify the best-fit parameters for each model. Evaluate the predictive success of landing using test data.

Data Collection

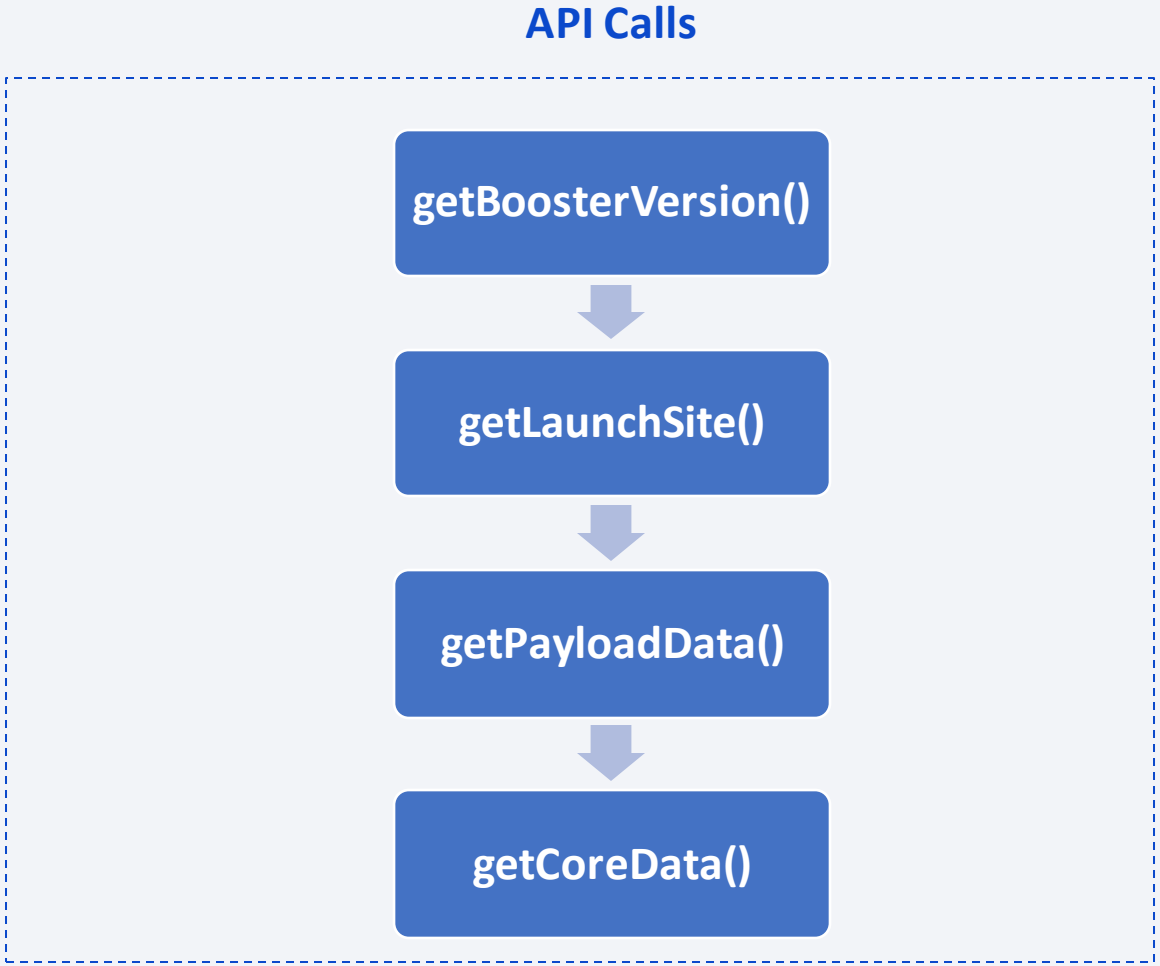
- Data was sourced and extracted from SpaceX API (api.spacexdata.com)
  - Additional historical data was web-scrapped from the Falcon 9 Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Data was preprocessed and cleaned.
  - Removed extraneous variables and metrics



Data Collection – SpaceX API

- Defining calls to SpaceX API:
  1. getBoosterVersion(data): Call API to append booster name using the rocket IDs in the dataset
  2. getLaunchSite(data): Call API to append Launch Site, longitude and lattitude name using the launchpad IDs in the dataset
  3. getPayloadData(data): Call API to append Payload Mass data using the payload IDs in the dataset
  4. getCoreData(data): Call API to append landing outcome, landing type, number of flights, gridfins, reused cored , legs used landing pad, block of the core, times the core has been reused and the core serial number
- API Requests
  - Dataset was requested and parsed from the API using the defined calls to the SpaceX API result was store into a dataframe
- Data Collection Notebook:

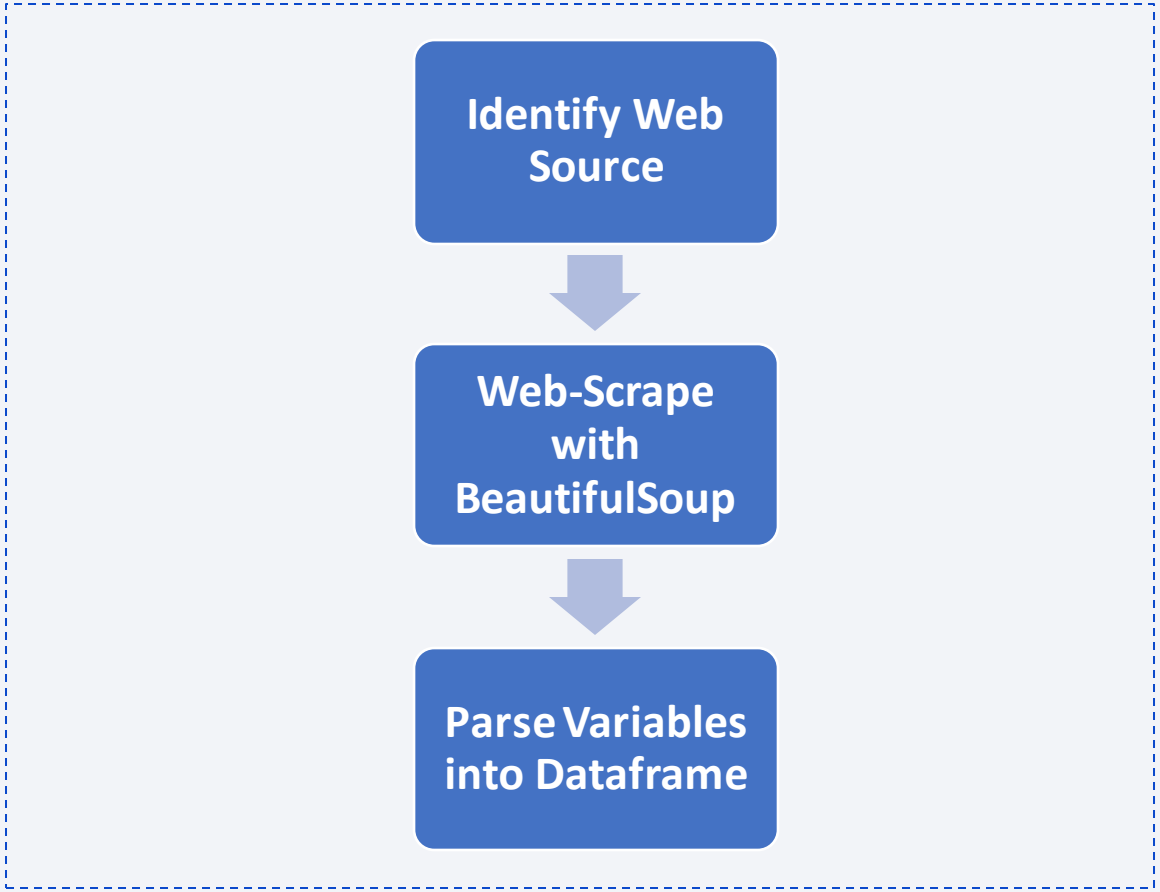
https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

- Web-Scrapping
  - Web-scraped Falcon 9 launch records from the HTLM table on Wikipedia using BeautifulSoup and extracted variables.
- Parsed Launch HTML tables and created dataframe with the following variables:
  1. Flight No.
  2. Launch site
  3. Payload
  4. Payload Mass
  5. Orbit
  6. Customer
  7. Launch outcome
  8. Version Booster
  9. Booster Landing
  10. Date
  11. Time
- Web-scraping Notebook:

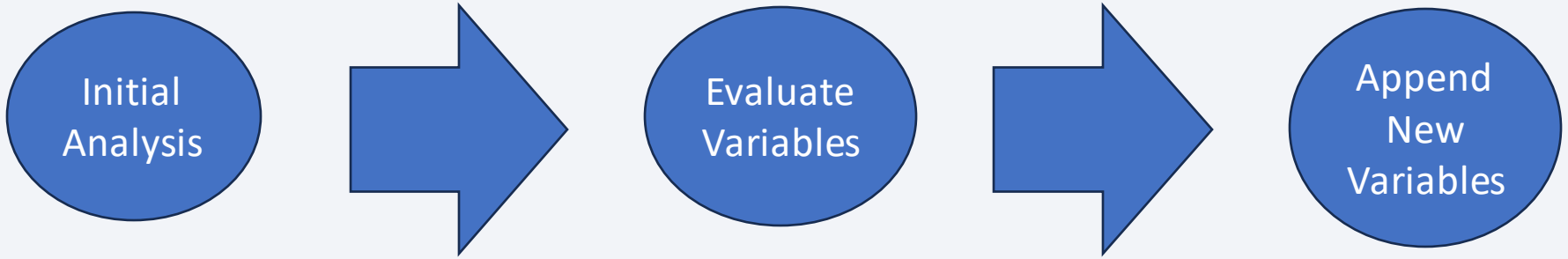
https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data filtering
  - Removed Falcon 1 launches to limit scope to Falcon 9 only
- Handling Missing Values
  - Filled missing payload mass with average mass
- Converted categorical variables to numerical
  - Appended a Outcome numerical column based on the "landing outcome" categorical results
- Data Wrangling Notebook:

  https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

The following plots were created for visual data analysis:

- Flight Number vs. Payload Mass Scatter Plot
  - Examine how the flight number and payload mass are related to the success of landings.
- Flight Number vs. Launch Site Scatter Plot
  - Understand if the choice of launch site and number of launches has any influence on the success of landings.
- Payload Mass vs. Launch Site Scatter Plot
  - Examine if there is any correlation between the launch site and the payload mass with success of landings.
- Success Rate of Each Orbit Type Bar Chart
  - Identify which orbits have higher success rates.
- Flight Number vs. Orbit Type Scatter Plot
  - Examine how the flight number relates to the orbit type and its success.
- Payload Mass vs. Orbit Type Scatter Plot
  - Understand the correlation between payload mass, orbit type, and the success of landings.
- Launch Success Yearly Trend Line Chart
  - Identify the trend in launch success rates over time.

EDA with Data Visualization Notebook:

https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Using SQLLite, the following queries were generated:

- Display unique launch sites
- Display 5 records where launch sites start with '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 and 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 display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

EDA with SQL notebook:

https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following map objects were created and added to a Foliumn Map used to visualize launch sites for SpaceX:

- **Circle objects** were created to highlight specific launch sites, including the NASA Johnson Space Center's launch sites.
- **Marker objects** were used to add markers with icons to the map to represent number of launches at the launch site and their success status. In the notebook, they are used to represent launch sites, with each marker showing the launch site's name as an icon. The icon for launch sites contains the name.
- **Lines and Distance markers** were created to indicate the distance between launch sites and specific proximities, such as coastlines, cities, railways, and highways.
- A **Mouse Position object** was added to the map to show the coordinates (latitude and longitude) of the mouse pointer when hovering over the map.

Folium Interactive Map Notebook:

https://github.com/acr66/testrepo/blob/249a3710dcfc7221e6ef2d873606de3fc9d2e87e/Capstone%20Project/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

The following plots and interactions were added to an interactive Dashboard App:

- Dropdown list to select Launch sites. By default All Sites is displayed and presents data from all launch sites. Individual sites may also be selected.
- When "All Sites" is selected, a pie chart shows the total successful launches count for all sites.
- When a specific launch site is selected, a pie chart displays percentage of successful and failed launches for that site.
- A scatter plot displays the payload mass by launch success, color-coded by booster type.
- User is able to control a slider controls the range of payload mass to include in the scatter plot

Plotly Dash Python Script:

https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/spacex_dash_app.py

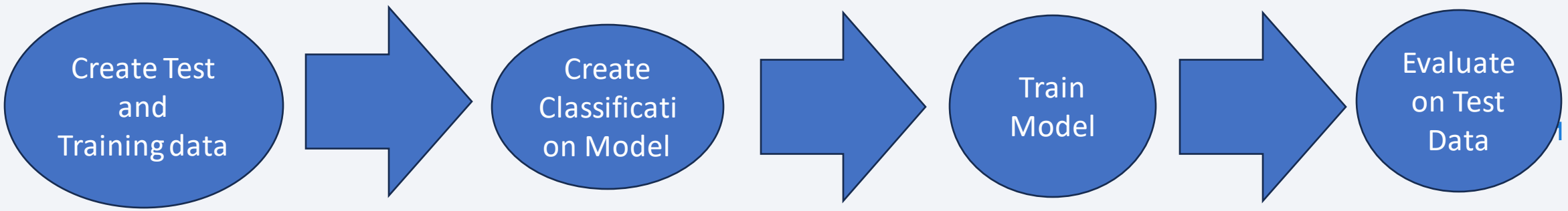
Predictive Analysis (Classification)

The following steps were completed to identify the best performing Machine Learning classification model:

- Create a Training Label
- Standardize the Data
- Split Data into Training and Test Sets
- Created several classification models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, K Nearest Neighbors (KNN)
- Identify the best fitting parameters
- Evaluate the accuracy of the models on the test data and plot the confusion matrix
- Determine the Best Performing Model

Link to Predictive Analysis Notebook:

https://github.com/acr66/testrepo/blob/ca161ee7f973a2639a24b760b7b336013ddee8bc/Capstone%20Project/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

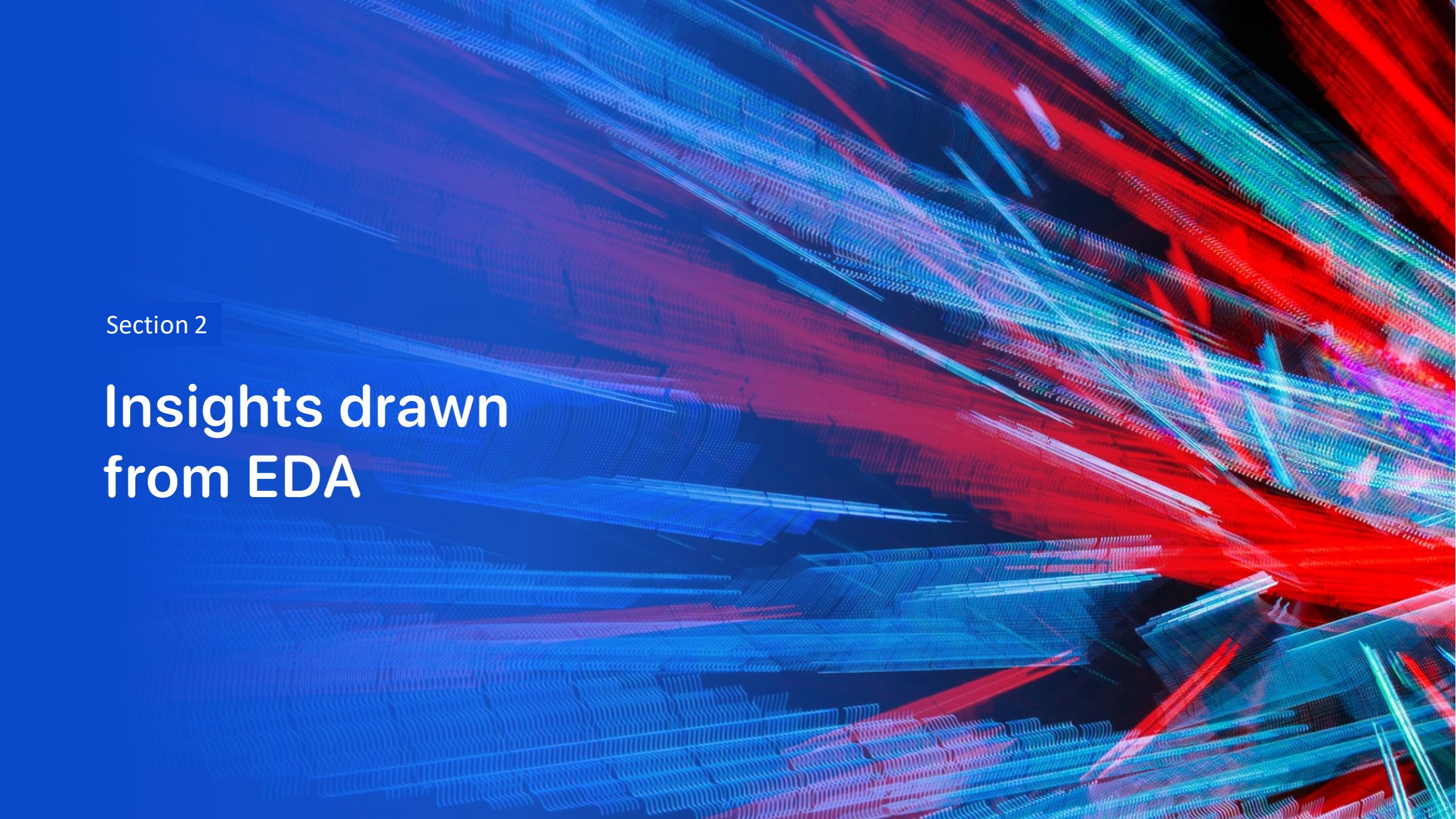


Results

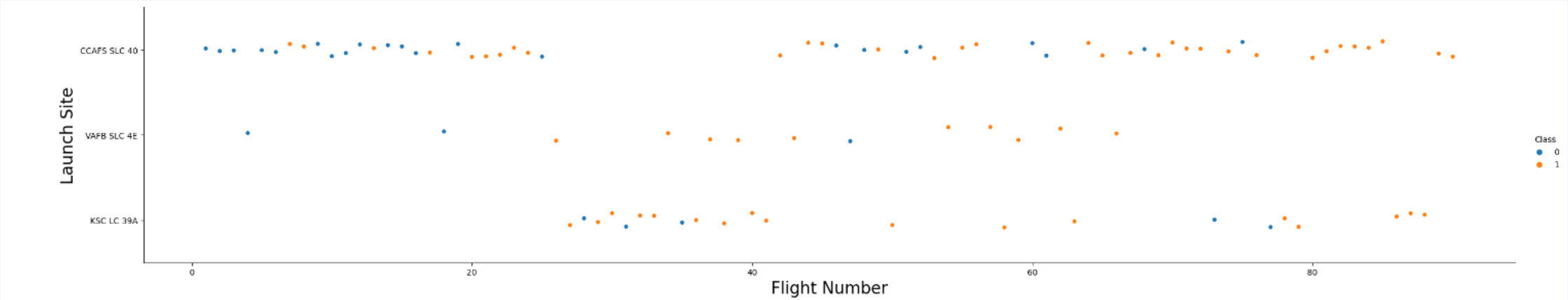
- Through exploratory data analysis, we identified several variables that could be indicative of a successful first stage landing. These variables include: payload mass, launch site, flight number and booster type.
- Interactive analytics were developed in the form of a Folium map and dashboard app continue plots of the indicative variables. Screenshots of these applications are on Section 3 and Section 4.
- Through predictive analysis, we built and evaluated Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbor Regression classification models for predicting successful landings on train and test data. While no single model was determined to be more accurate than another, we do find the accuracy score of our model is 0.83 which is a good first pass in predicting landing success, thus indicating that competitor to SpaceX is viable.

Section 2

Insights drawn from EDA

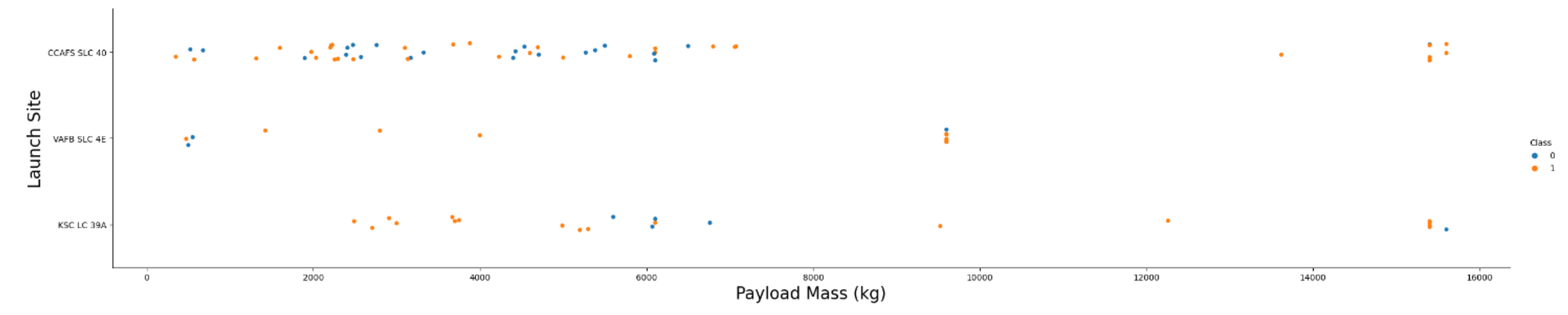


Flight Number vs. Launch Site



- Successful landings are indicated by the colored dot for each data point (Blue for Failure, Yellow for Success)
- Early flight numbers took place predominantly at CCAFS SLC-40, which is also where the majority of launches take place.
- Different launch sites have different success rates. CCAFS SLC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Many of the early flights at CCAFS SLC 40 where failures, however the other sites do not share the same trend.

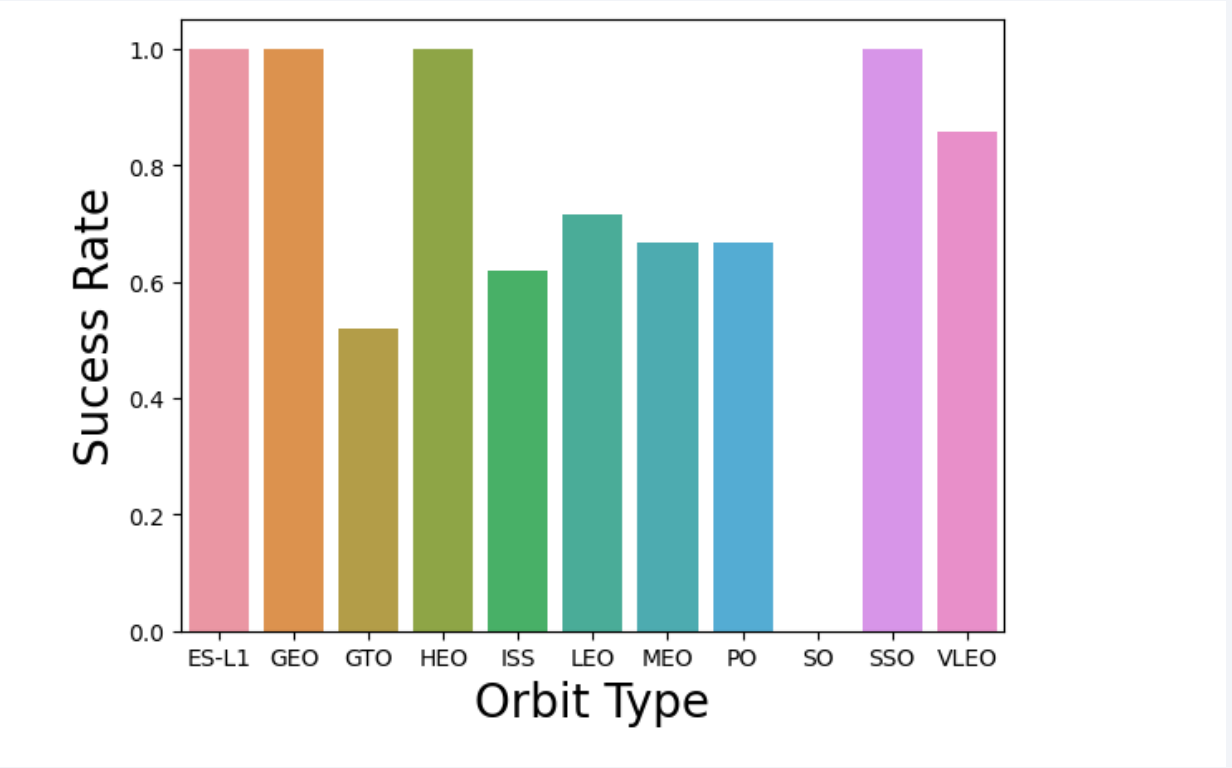
Payload vs. Launch Site



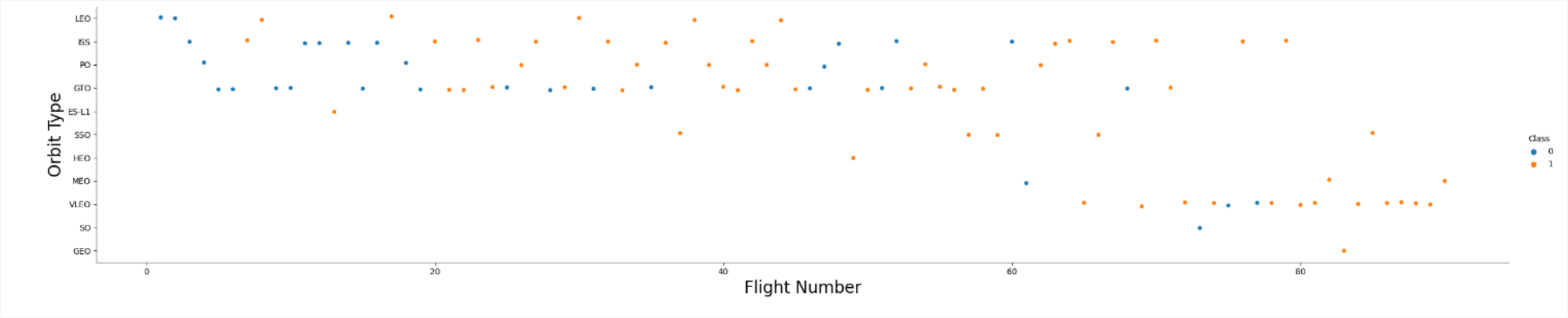
- Successful landings are indicated by the colored dot for each data point (Blue for Failure, Yellow for Success)
- Most launched payload mass is below 8000 kg.
- VAFB SLC 4E does not have any payloads greater than 10000 kg.
- For CCAFS SLC-40, payloads over 6000 kg tend to be more successful. However, there doesn't appear to be strong correlation for payloads and success at other sites.

Success Rate vs. Orbit Type

- ES-L1, GEO and SSO orbits have perfect launch records at SpaceX
- The worst launch record is the SO orbit which has no succesful launches.
- GTO, ISS, LEO, MEO and PO orbit types have a success rate between 0.5 and 0.7, close to the average success rate across the board.

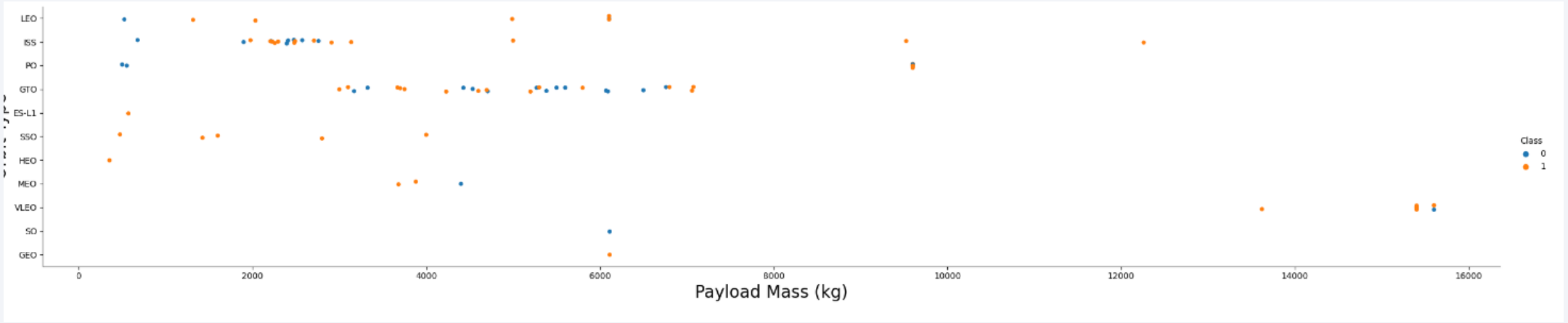


Flight Number vs. Orbit Type



- Successful landings are indicated by the colored dot for each data point (Blue for Failure, Yellow for Success)
- Majority of flights were using the ISS, GTO and VLEO orbits.
- There are less than 3 flights for the ES-L1, HEO, MEO, SO and GEO orbits
- LEO orbit success appears to be related to the number of flights. However, there is no relationship between flight number when in the GTO orbit.

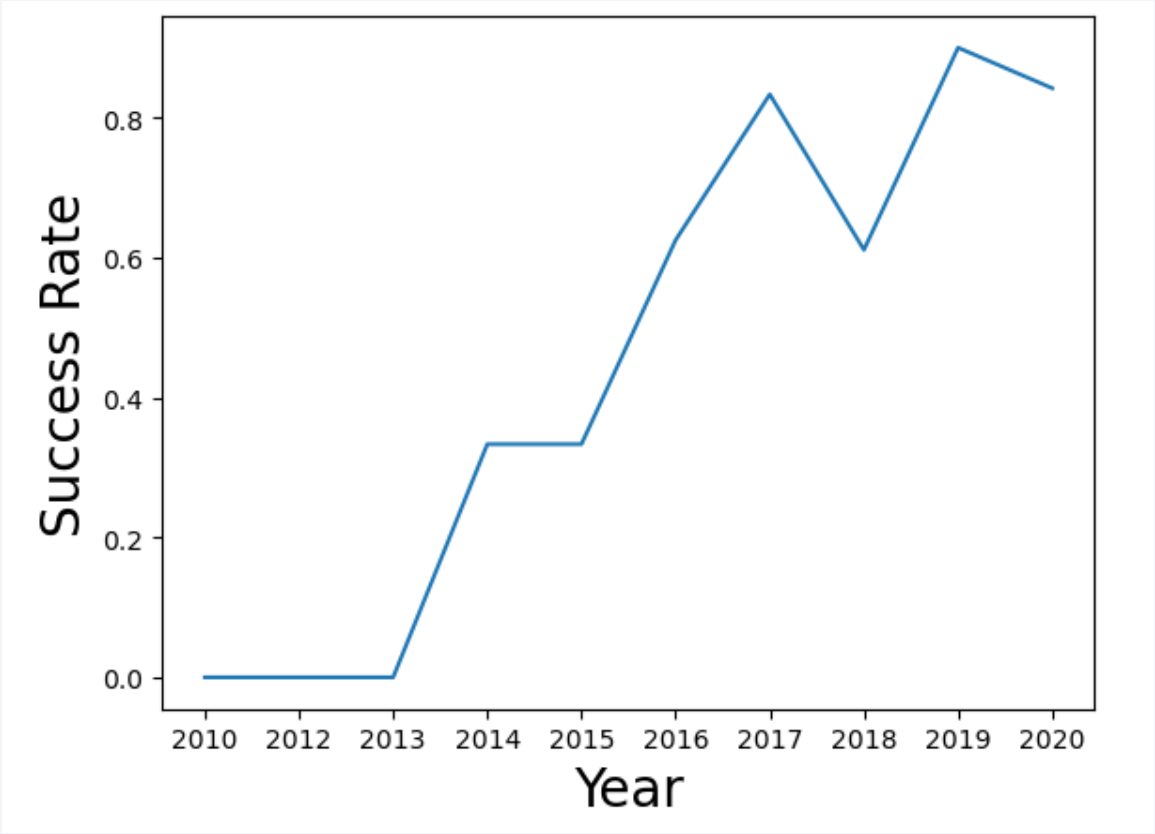
Payload vs. Orbit Type



- Successful landings are indicated by the colored dot for each data point (Blue for Failure, Yellow for Success)
- Payloads greater than 8000 kg used ISS, PO and VLEO orbits
- Heavy payloads are more likely to succeed for Polar, LEO and ISS orbits. GTO however doesn't follow this trend.

Launch Success Yearly Trend

- Success rate for the Falcon 9 generally increased over time from 2013 to 2020



All Launch Site Names

- The SQL query below identified the names of the unique launch sites for the Falcon 9
- The launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E

```
[72]: %sql select Launch_Site from SPACEXTABLE group by Launch_Site
 * sqlite:///my_data1.db
Done.
```

[72]:

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

Launch Site Names Begin with 'CCA'

- The SQL query below displays the first 5 records where launch sites begin with `CCA`

```
[29]: %sql select * from SPACEXTABLE where Launch_Site LIKE 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

[29]:

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

Total Payload Mass

- The SQL query below calculates the total payload carried by boosters from NASA
- The total payload is 99980 kg

```
[19]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer LIKE 'NASA%'
 * sqlite:///my_data1.db
Done.
[19]: sum(PAYLOAD_MASS__KG_)
                      99980
```

Average Payload Mass by F9 v1.1

- The SQL query below calculates the average payload mass carried by booster version F9 v1.1
- The average payload is 2535 kg

```
[22]: %sql select  AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version LIKE 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
[22]: AVG(PAYLOAD_MASS__KG_)
      2534.6666666666665
```

First Successful Ground Landing Date

- The SQL query below identifies the dates of the first successful landing outcome on ground pad
- The first successful landing outcome on ground pad happened on 22 Dec 2015

```
[25]: %sql select  MIN(Date) from SPACEXTABLE where Landing_Outcome LIKE 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
[25]: MIN(Date)
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SQL query below lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The booster names are F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2

```
[26]: %sql select Booster_Version from SPACEXTABLE where Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
 * sqlite:///my_data1.db
Done.
```

[26]:

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

Total Number of Successful and Failure Mission Outcomes

- The SQL query below calculates the total number of successful and failure mission outcomes
- There are 100 successful mission outcomes and 1 failure

```
[38]: %%sql
      select Mission_Outcome, SUM(1)
      from SPACEXTABLE
      group by Mission_Outcome
```

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

[38]:

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

Boosters Carried Maximum Payload

- The SQL query below lists the names of the booster which have carried the maximum payload mass (15600 kg)

[46]:
```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTABLE
where PAYLOAD_MASS__KG_ == (
    select Max(PAYLOAD_MASS__KG_)
    from SPACEXTABLE
    )
```

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

[46]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- The SQL query below lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where Landing_Outcome LIKE 'Failure (drone ship)' AND substr(Date, 1, 4) = '2015'
```

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

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

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

- The SQL query below ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

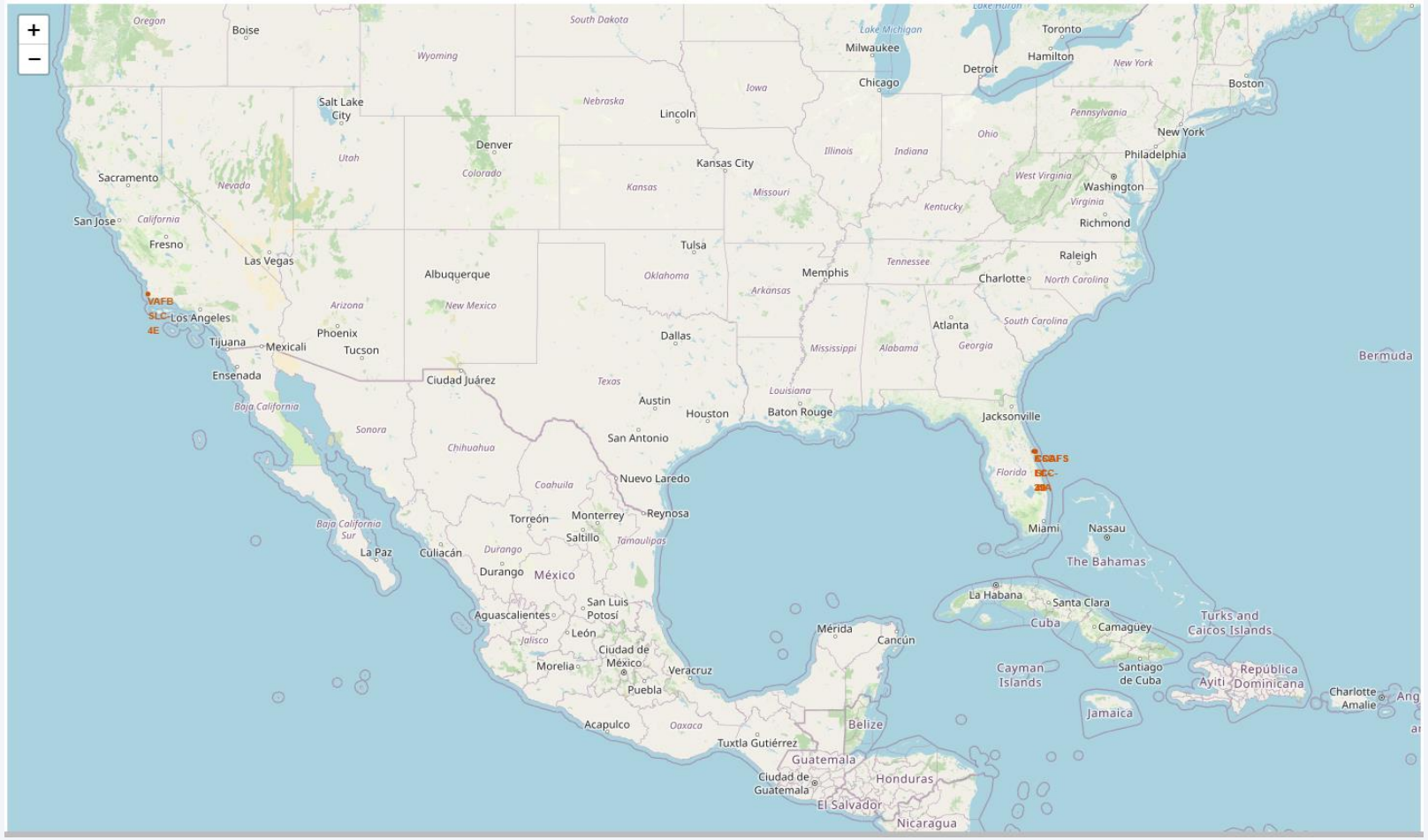
```
%%sql
select Landing_Outcome,count(Landing_Outcome)
from SPACEXTABLE
where Date > '2010-06-04' AND Date < '2017-03-20'
group by Landing_Outcome
order by count(Landing_Outcome) Desc
```

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

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

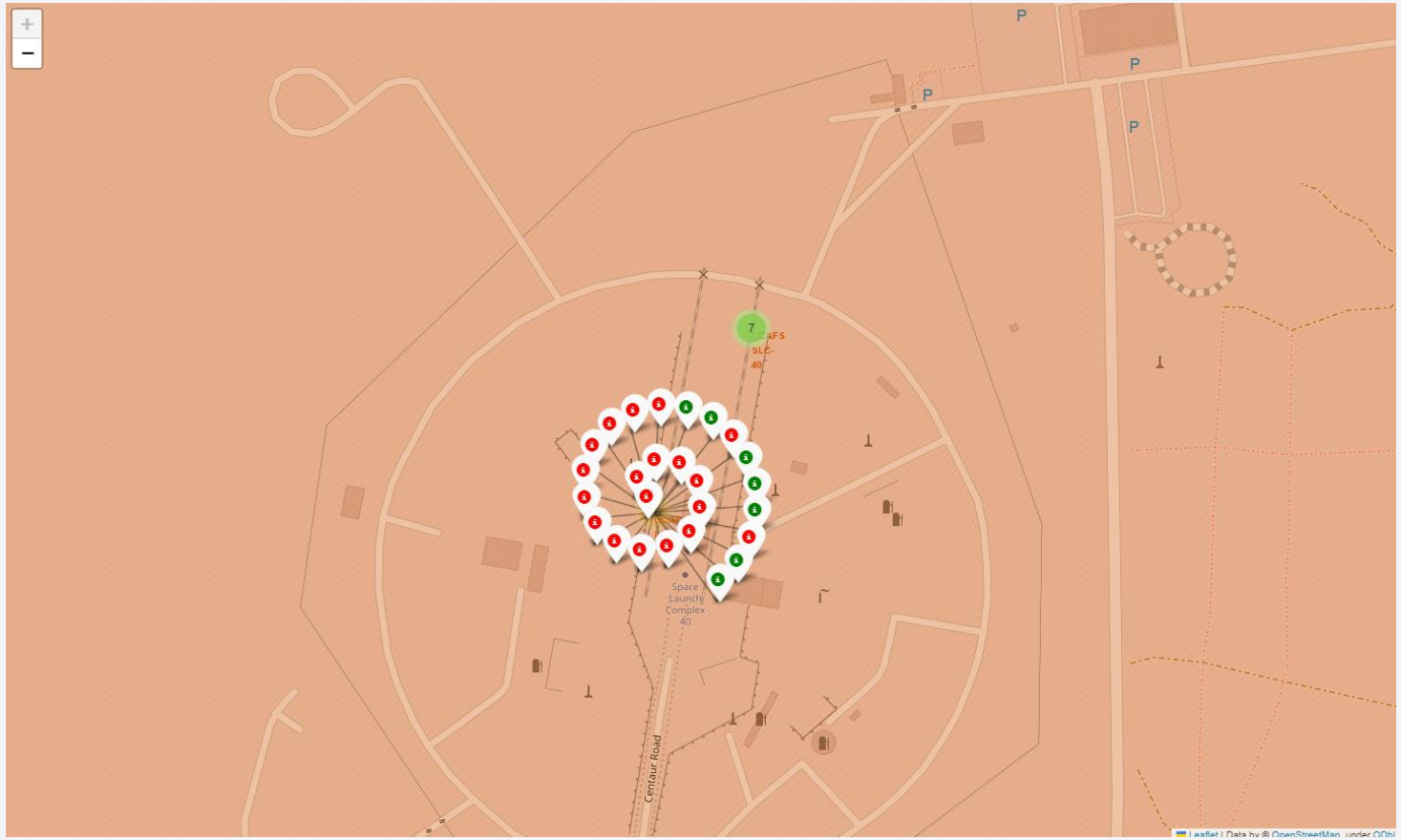

Section 3
Launch Sites
Proximities Analysis

Falcon 9 Launch Sites



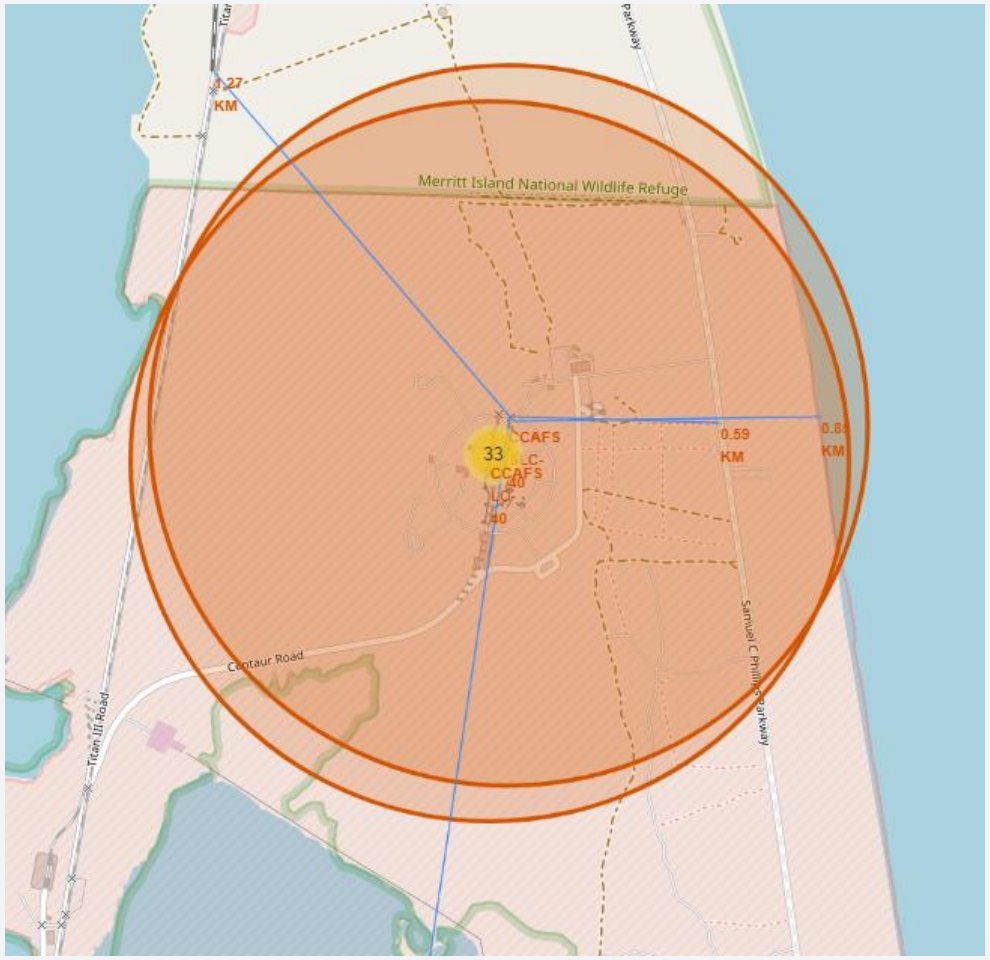
As the map indicates, 1 launch site is in California and 2 launch sites are in Florida. It is notable that these sites are in coastal regions of the US.

Close-up of a Launch Site

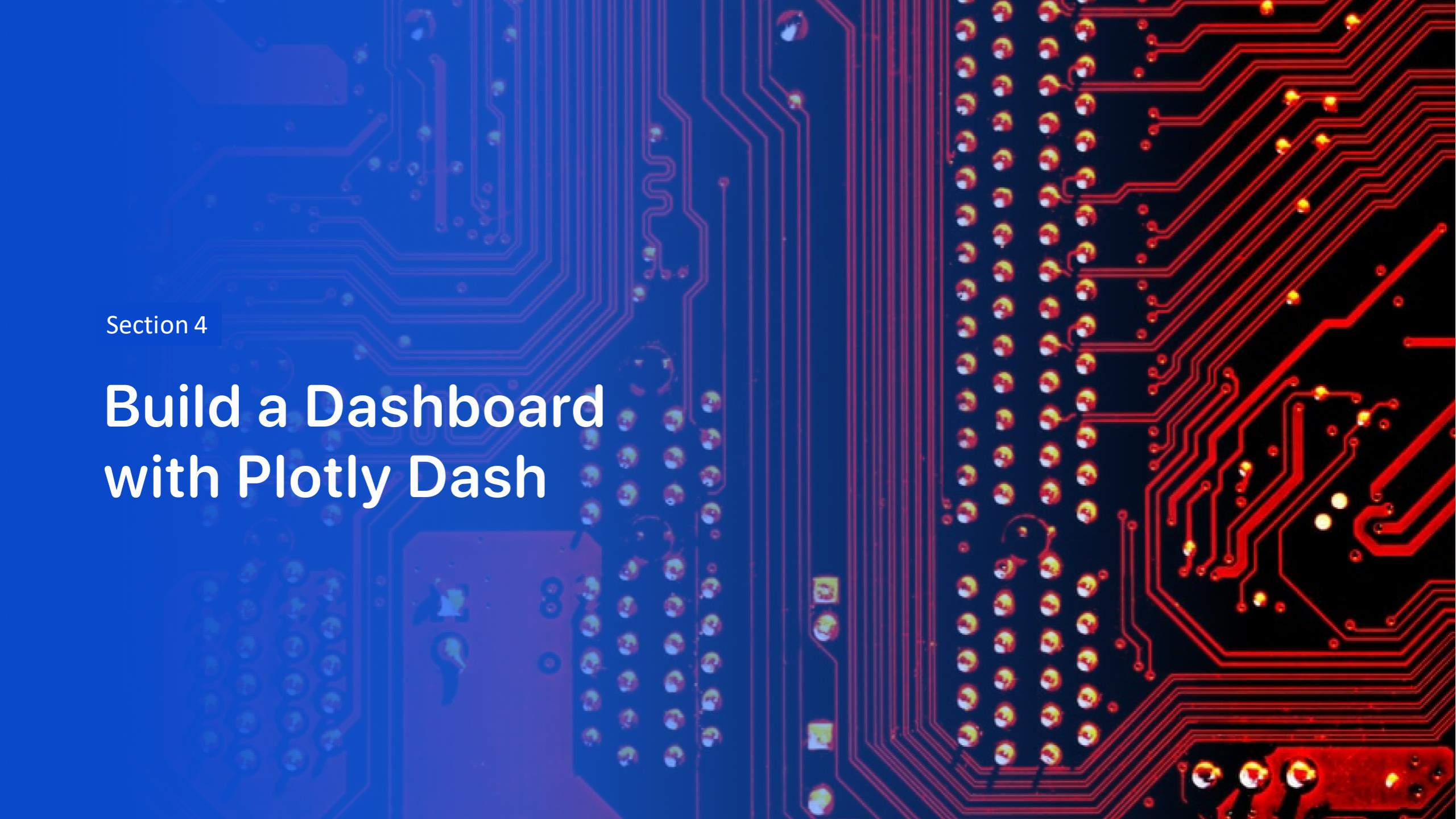


The map indicates number of launches and the landing outcome for this particular site (red is fail, green is success). Many flights happen at each launch site, with variable rates of success.

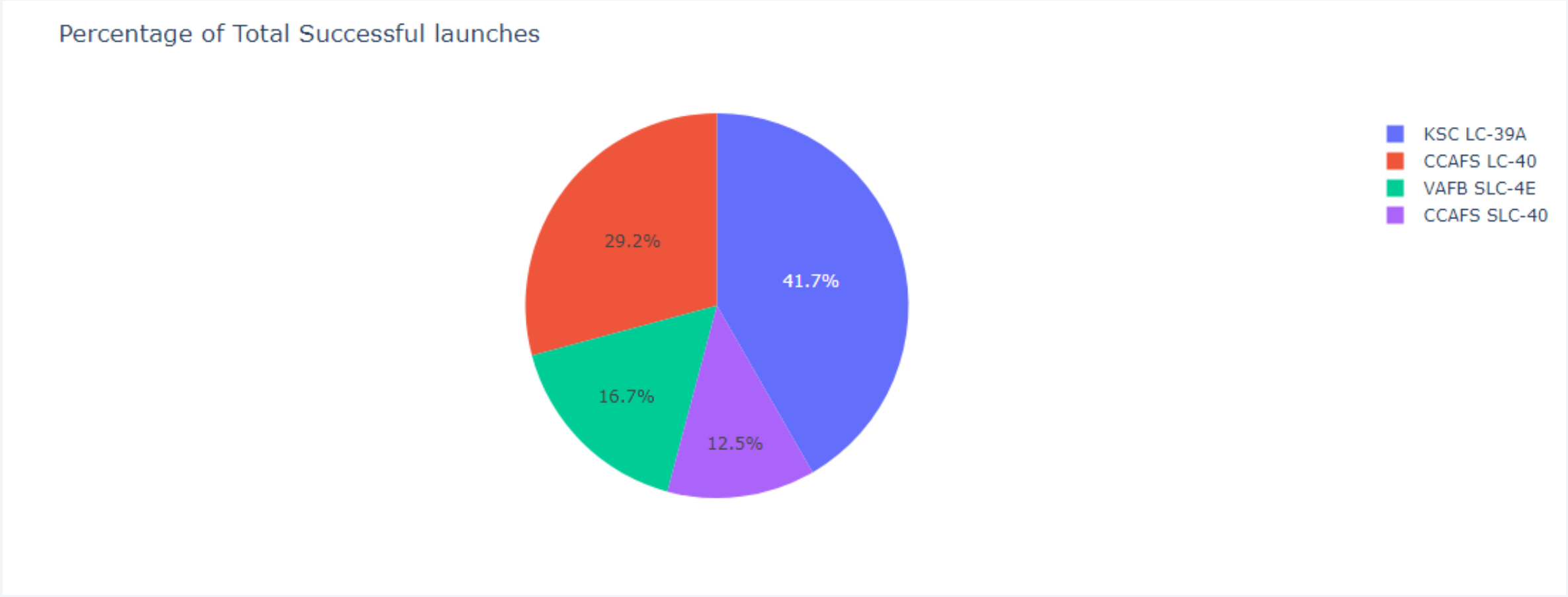
Proximity of Infrastructure and Coastline to a Launch Site



As previously mentioned, the launch sites are close to US coastline. By plotting distance to railway, highways and cities, we can draw other potential conclusions on the impact of proximity to infrastructure on the success of landings.

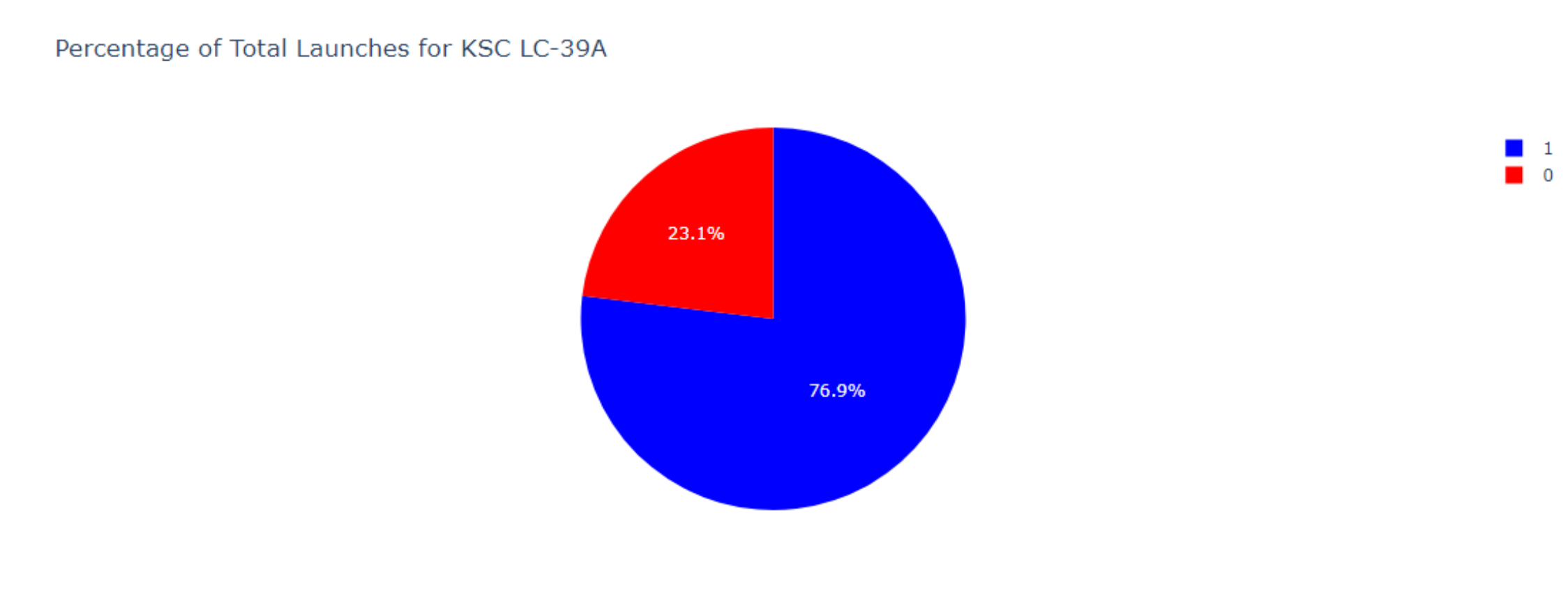

Section 4
Build a Dashboard with Plotly Dash

Percentage of Total Successful Launches by Launch Site



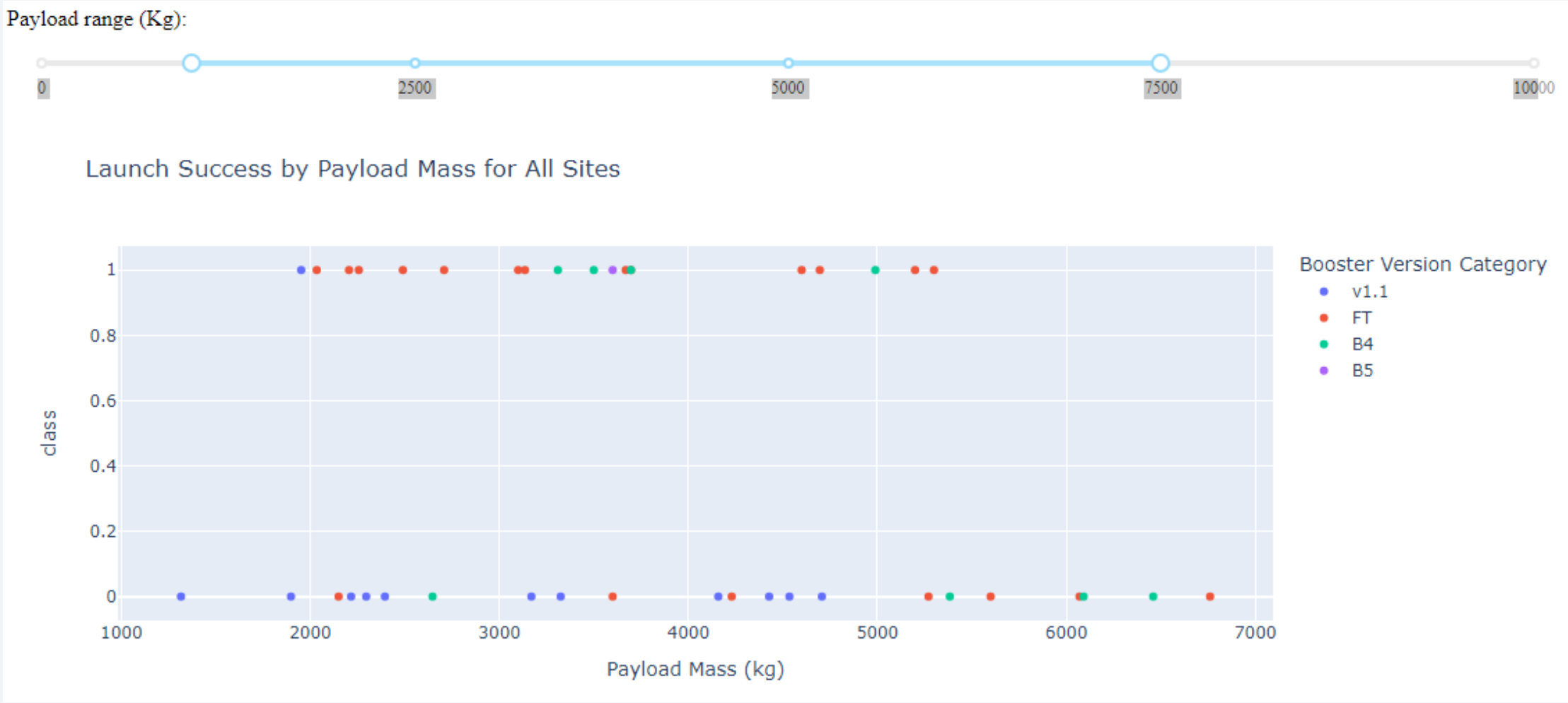
This screenshot from the Dashboard app shows the percentages of successful landings pie chart for when All Sites is specified. Site KSC LC-39A contains the grates percentage of successful landings.

Success (Blue) and Failure (Red) Percentages of Total Launches at KSC LC-39



Screenshot shows the successful landings pie chart for when site KSC LC-39A is specified (Blue = success, red = fail). This site contained the greatest percentage of successful launch percentage at 76.9%.

Launch Success by Payload Mass for All Sites with Payload between 1000 and 7500 kg



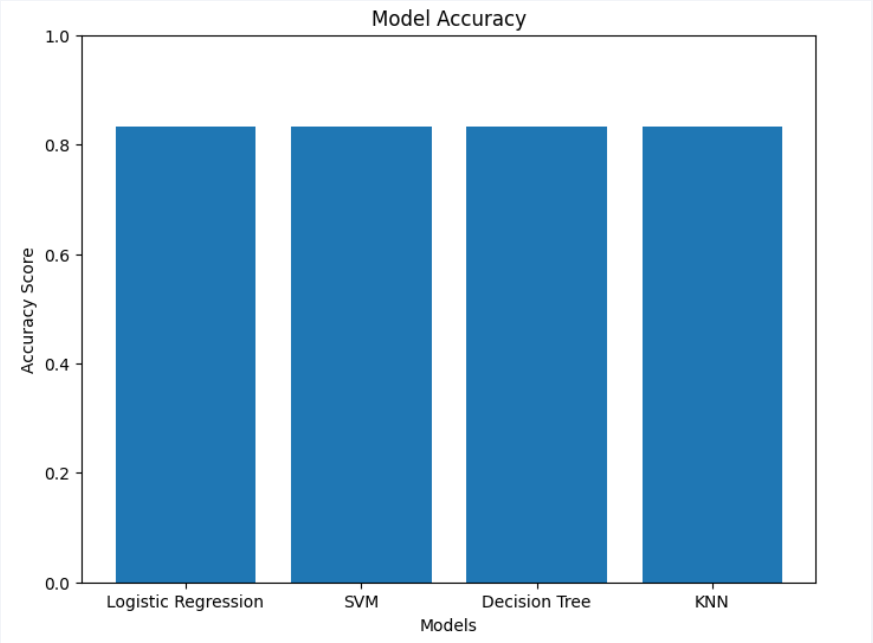
The screenshot depicts the scatter plot and payload range slider. The user is able to use the slider to set the payload range for the Successful landing vs. Payload mass scatter plot. Additionally, the plot is color-coded by booster type.

Section 5

Predictive Analysis (Classification)

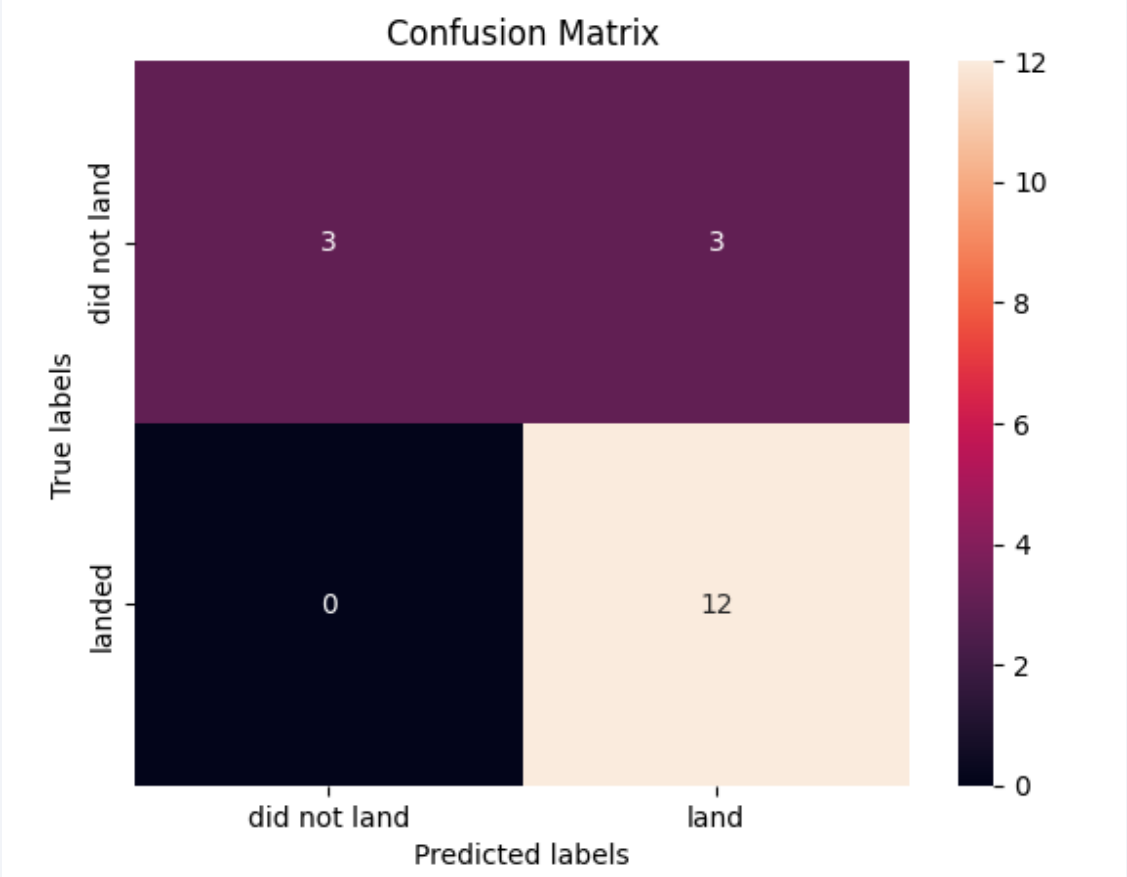
Classification Accuracy

- The bar chart visualizes the accuracy score of the Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbor classification models.
- Based on the results, no model was found to perform better than the rest based on classification accuracy, however the decision tree model had the highest best score based on the best hyperparameters combination at 0.891



Confusion Matrix

- The confusion matrix for the best performing model can distinguish between the different classes.
- However, there are still a significant number of false positives as 3 launches that were predicted to land successfully, did not land.



Conclusions

Not having a particular model stand out has several possible implications on the dataset:

- There are no strong discriminative patterns that can be captured by the different regression models.
- Data could be distributed to the extent that no distinct clusters or classes can be found.

We do find that every model tested had an accuracy of 0.83 which indicates it is a good model for predicting successful landings. The model is not perfect however; there are a significant number of false positives present that categorized a failed landing as a success.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!