IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daria
October 14, 2023

Outline

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

Executive Summary

Methodologies:

- Data collection through API and web scraping
- Exploratory data analysis with:
    - Pandas and Numpy
    - SQL
    - Visualizations (Matplotlib and Seaborn)
- Interactive visual analysis using Folium and Dash
- Predictive Analysis (Classification)

Key results:

- The factors that affect the outcome of the rocket launch were analysed, as well as the relationship between these factors
- The historical trends of rocket landings were analysed
- Spatial analysis of launch sites was performed
- Interactive visualizations and a dashboard were developed
- Best model for launch success prediction was determined (Decision Tree)

Introduction

In this capstone project, we predict if the Falcon 9 first stage will land successfully, by addressing the following questions:

- What factors affect the outcome of the rocket launch, and what is the relationship between these factors?
- What are the historical trends of rocket landings?
- Where are launch sites located (geography, proximity to other objects) and which launch site is the most successful?
- Which model predicts the success of the launch most accurately?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:** Data was collected using SpaceX API and web scraping from Wikipedia
- **Perform data wrangling:** Data processing involved transforming all collected data into pandas dataframes and application of one-hot encoding to categorical features
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models:** Constructing and evaluating classification models, including Logistic Regression, SVM, Decision Tree, and K Nearest Neighbors, including hyperparameter tuning and assessing model performance using accuracy scores and confusion matrices.

Data Collection

Data collection workflow:

1. GET request to access SpaceX API data
2. Transformation of the response content into a JSON format utilizing the .json() function and converting it into a Pandas dataframe with .json_normalize()
3. Cleaning procedures
4. Conducting web scraping from Wikipedia using BeautifulSoup. The aim was to extract launch records from an HTML table, parse the table, and translate it into a Pandas dataframe for subsequent analysis.

Data Collection – SpaceX API

The process is shown on the flowchart

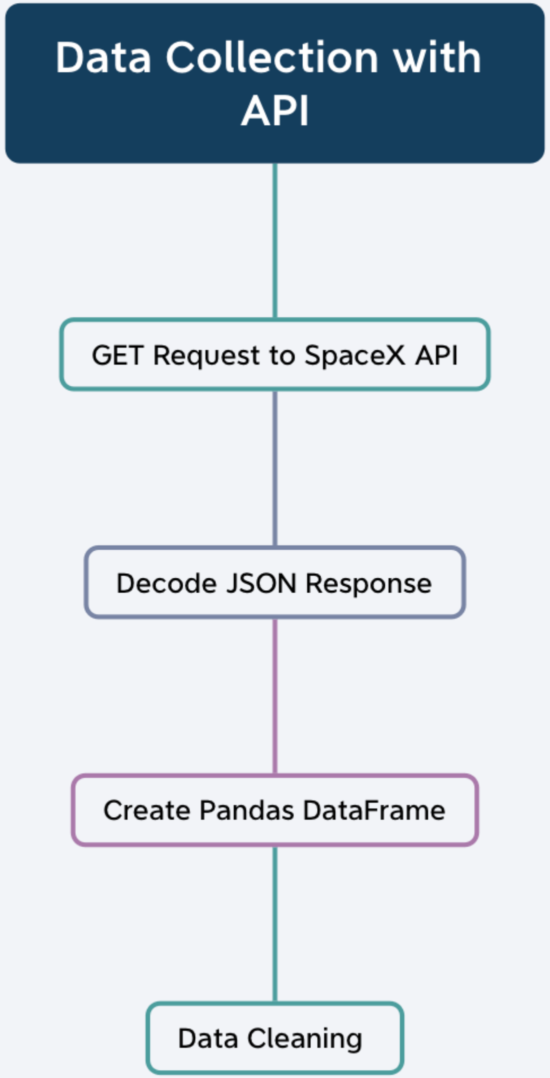
CLICK HERE to see the Notebook on GitHub

Example of the code:

```
# Calculate the mean value of PayloadMass column
payload_mean = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, payload_mean, inplace=True)

data_falcon9.isnull().sum()
```



Data Collection - Scraping

The process is shown on the flowchart

CLICK HERE to see the Notebook on GitHub

Example of the code:

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

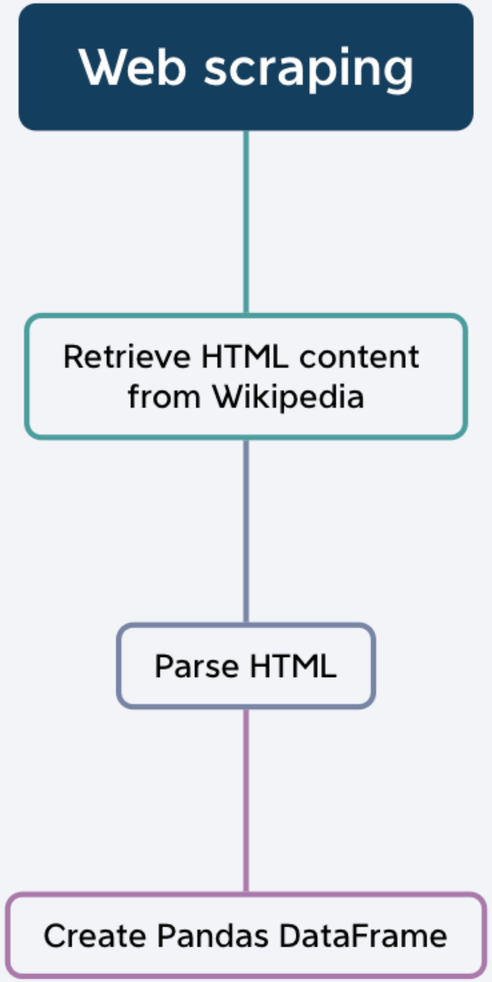
Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.content, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```



Data Wrangling

Four tasks successfully performed

CLICK HERE to see the Notebook on GitHub

Example of the code:

```
# Apply value_counts() on column LaunchSite
launch_site_counts = df['LaunchSite'].value_counts()
print(launch_site_counts)
```

Task 1: Calculate Launch Counts by Site

a. Retrieve LaunchSite data.

b. Use value_counts() to calculate the number of launches by site.

Task 2: Calculate Orbit Counts

a. Retrieve Orbit data.

b. Use value_counts() to count the occurrences of each orbit.

Task 3: Calculate Landing Outcomes for Orbits

a. Retrieve Outcome data.

b. Use value_counts() to count landing outcomes for orbits.

Task 4: Create Landing Outcome Labels

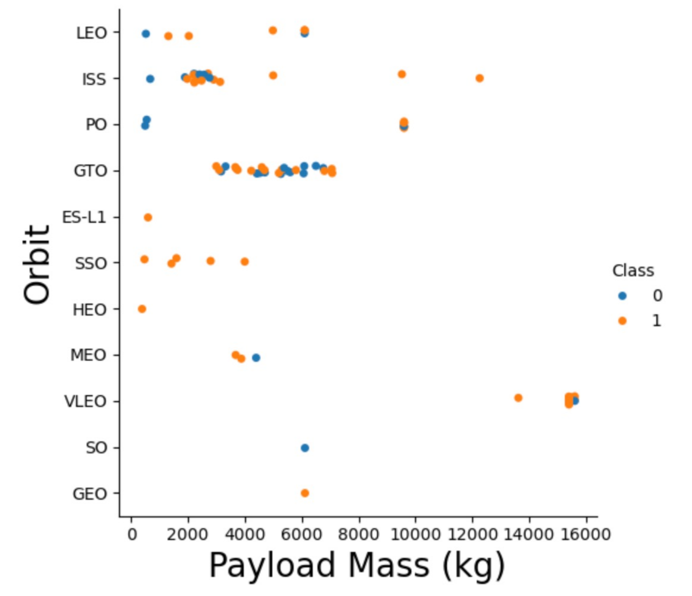
a. Retrieve Outcome data.

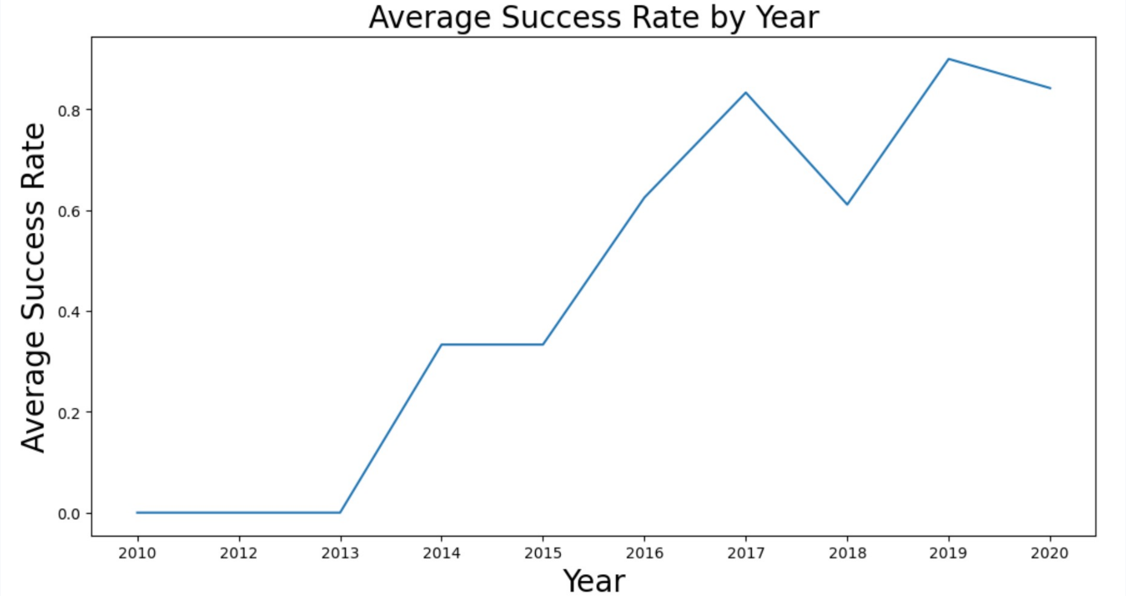
b. Create a landing_class label (0 for bad outcomes, 1 otherwise).

EDA with Data Visualization

Visualizing the relationship between flight number and launch site, payload mass and launch site, success rate of each orbit type, flight number and orbit type and the launch success yearly trend.

[CLICK HERE](#) to see the Notebook on GitHub





EDA with SQL

SQL queries performed:

- Launch sites analysis: the names of the unique launch sites, records where launch sites begin with 'CCA'
- Payload mass analysis: the total payload mass carried by boosters launched by NASA (CRS), average payload mass carried by booster version F9 v1.1, the names of the boosters that carry certain payload mass, etc.
- Landing outcomes analysis: the date when the first successful landing outcome in ground pad was achieved, the total number of successful and failure mission outcomes, rank the count of landing outcomes in 2010-2017, etc.

CLICK HERE to see the Notebook on GitHub

Build an Interactive Map with Folium

Objects added on map:

- Markers to represent the locations of all the launch sites on the map. Each marker is placed at the geographical coordinates of a launch site
- Circles to mark the locations on the map where the launch outcome was successful
- Lines were drawn on the map to illustrate the distances between the launch site and its proximities

CLICK HERE to see the Notebook on GitHub

Build a Dashboard with Plotly Dash

Plots/graphs and interactions added to a dashboard:

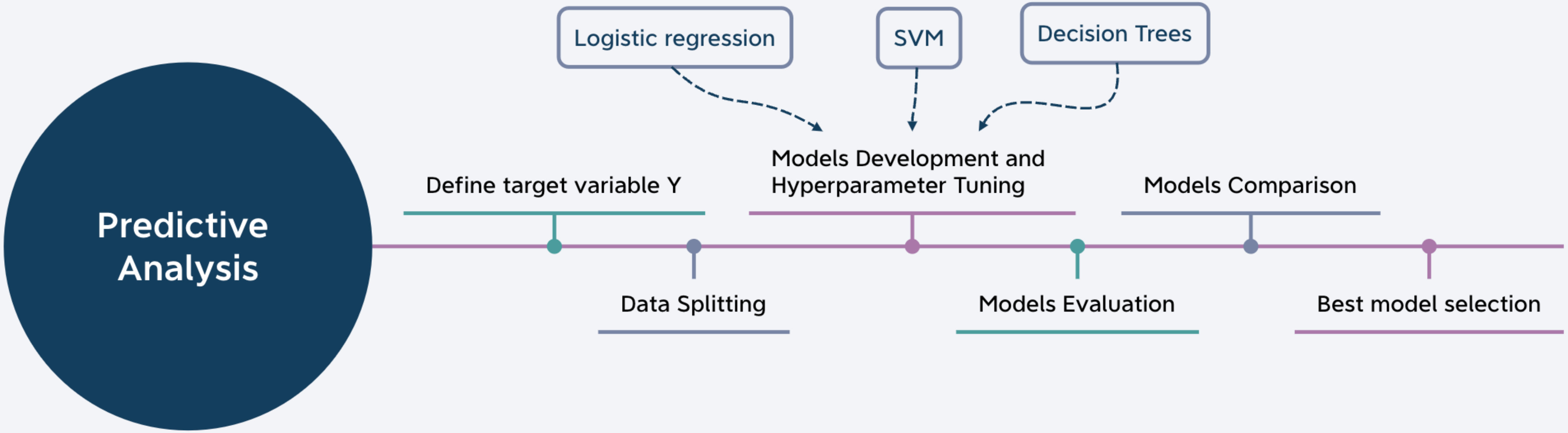
- Launch Site Dropdown allows users to select a launch site from a list
- Success-Pie-Chart to display the success rate of launches for the selected launch site.
- Payload Range Slider to allow users to specify a payload range they are interested in. This interactive component helps users filter launches based on payload mass.
- Success-Payload-Scatter-Chart that displays the relationship between payload mass and launch success for the selected launch site. It helps users explore whether there is any correlation between payload mass and launch outcomes

CLICK HERE to see the Notebook on GitHub

Predictive Analysis (Classification)

The workflow is presented on the chart

CLICK HERE to see the Notebook on GitHub



Results

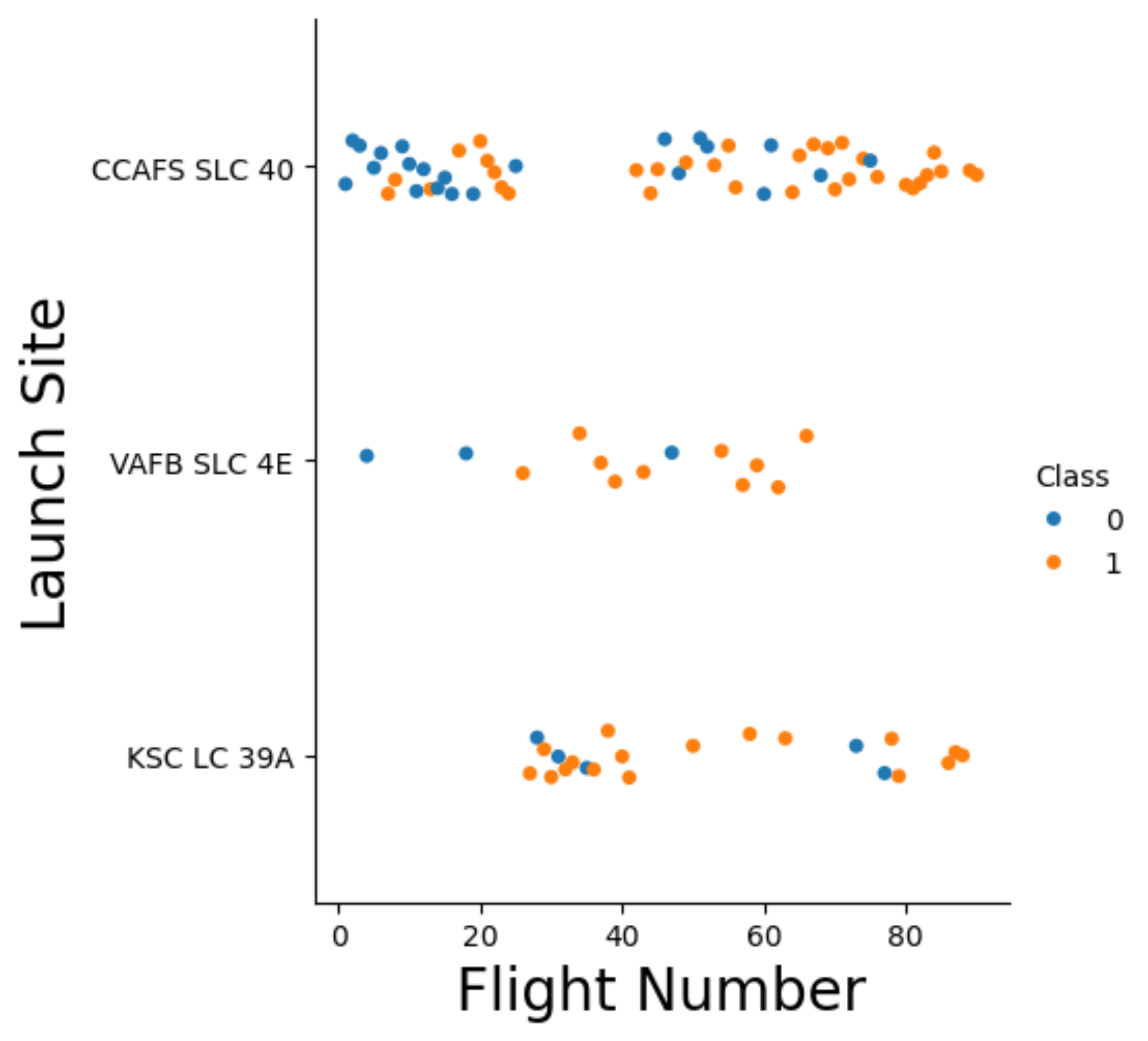
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

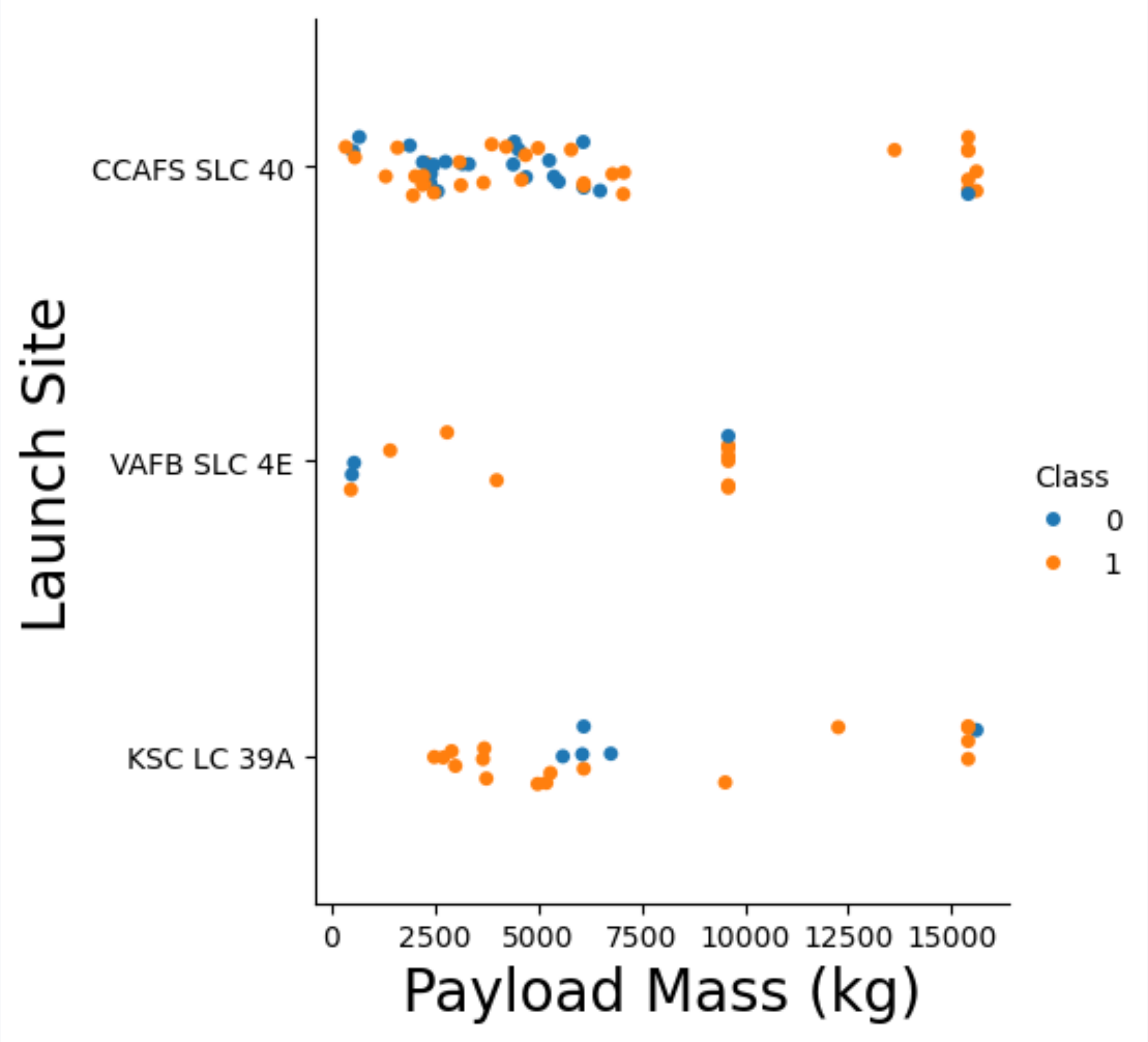
Flight Number vs. Launch Site

- The biggest number of flights was from CCAFS SLC 40 - however, a significant amount of them wasn't successful
- KSC LC 39A and VAFB SLC 4E have fewer number of flight, but the success rate is greater



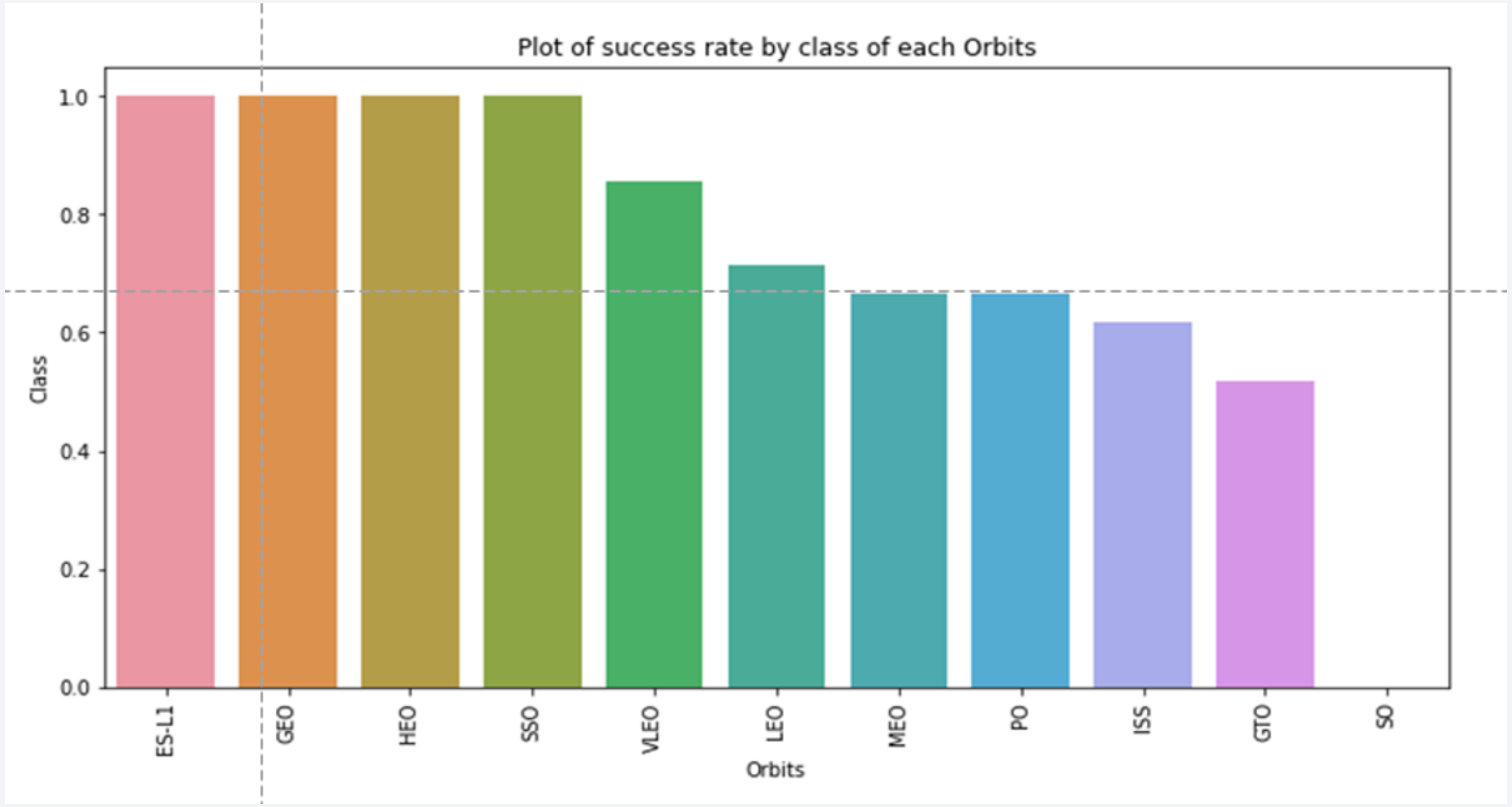
Payload vs. Launch Site

- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).
- There seems to be a positive correlation between the payload mass and the success of the mission, but more analysis is needed



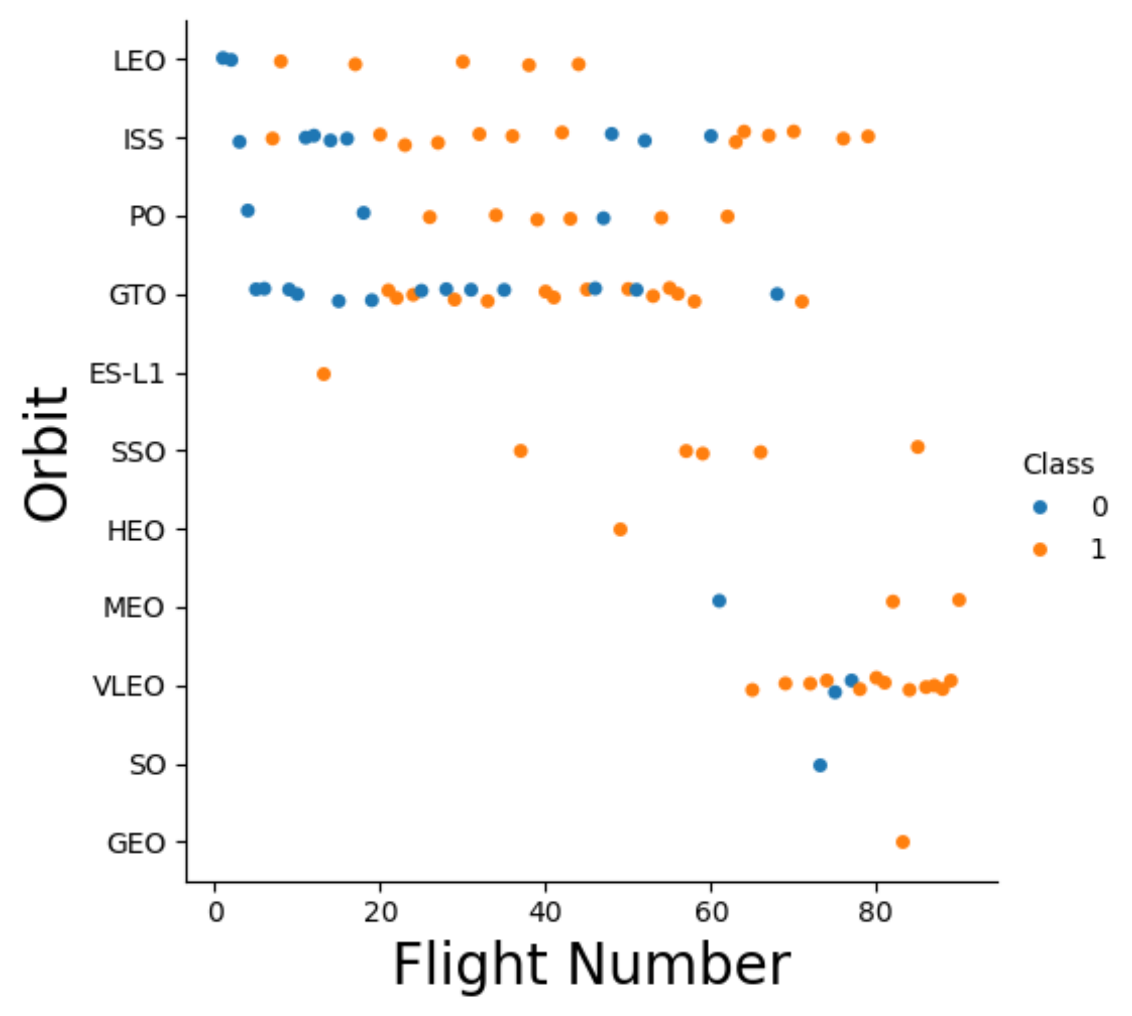
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO, and VLEO have the biggest success rate



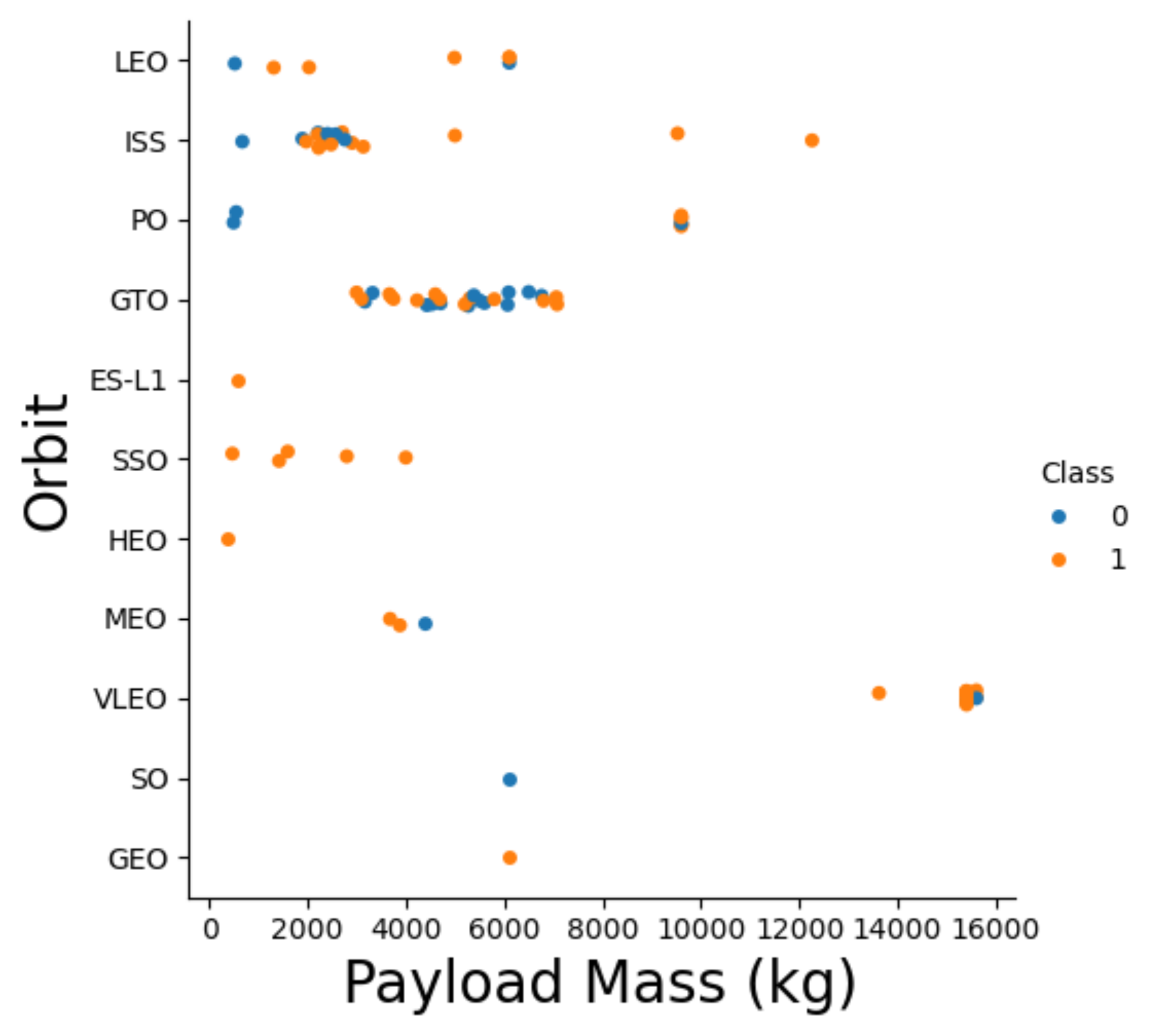
Flight Number vs. Orbit Type

- In the LEO orbit the success appears to be related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



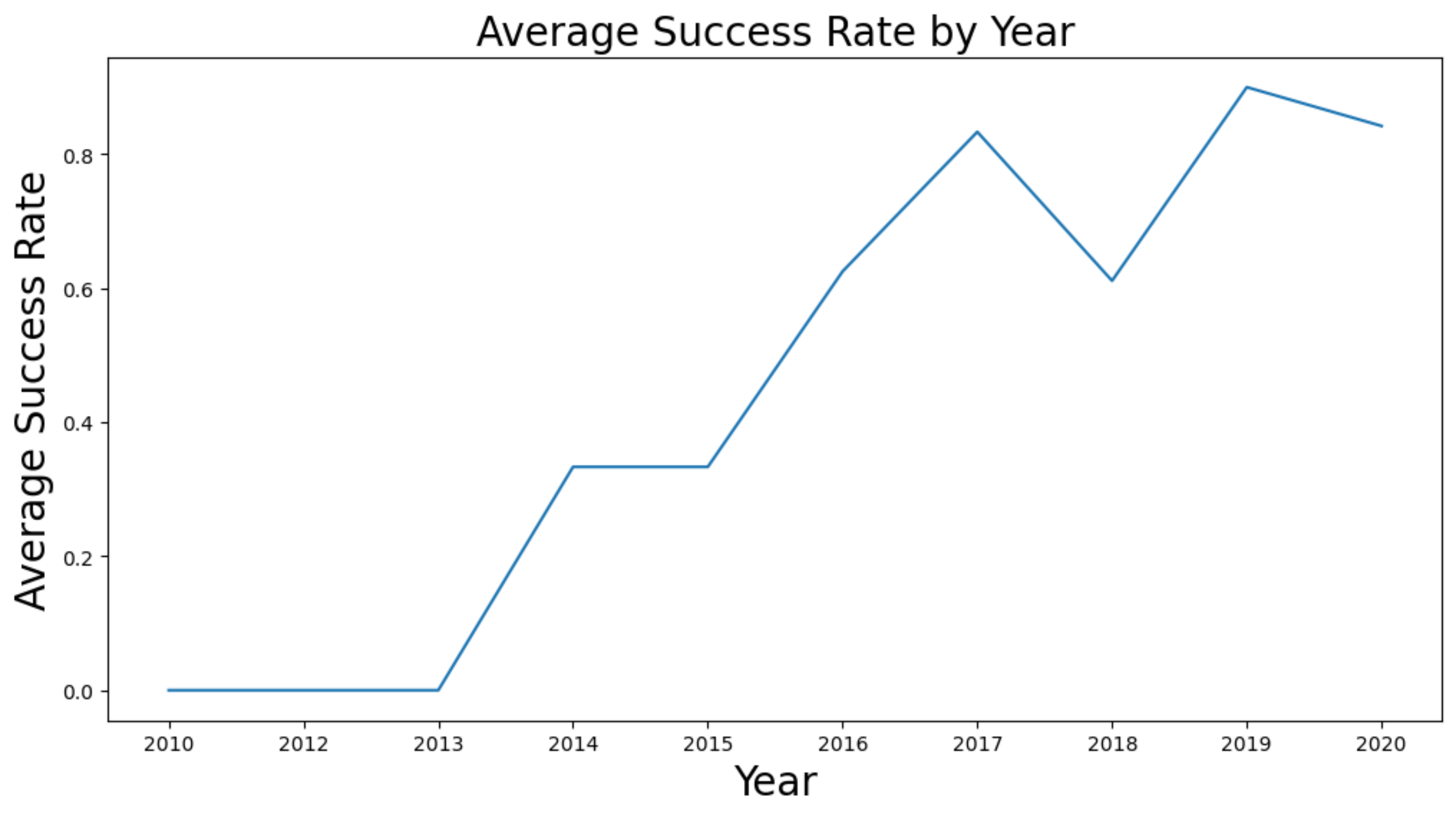
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020



All Launch Site Names

- The key word **DISTINCT** was used to show only unique launch sites from the SpaceX data.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

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 total payload carried by boosters from NASA was calculated using sum() function

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 was calculated using avg() function

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was calculated using min(Date)

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- I used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

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

Total Number of Successful Mission Outcomes

- The total number of successful mission outcomes

Mission_Outcome	count(*)
Success	22
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The booster that have carried the maximum payload was determined using a subquery in the **WHERE** clause and the **MAX()** function.

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

- In 2015, there were 2 failed landing_outcomes in drone ship
- A combination of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions was used to filter for failed landing outcomes in drone ship, their booster versions, and launch site names

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

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

- I selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- I applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order

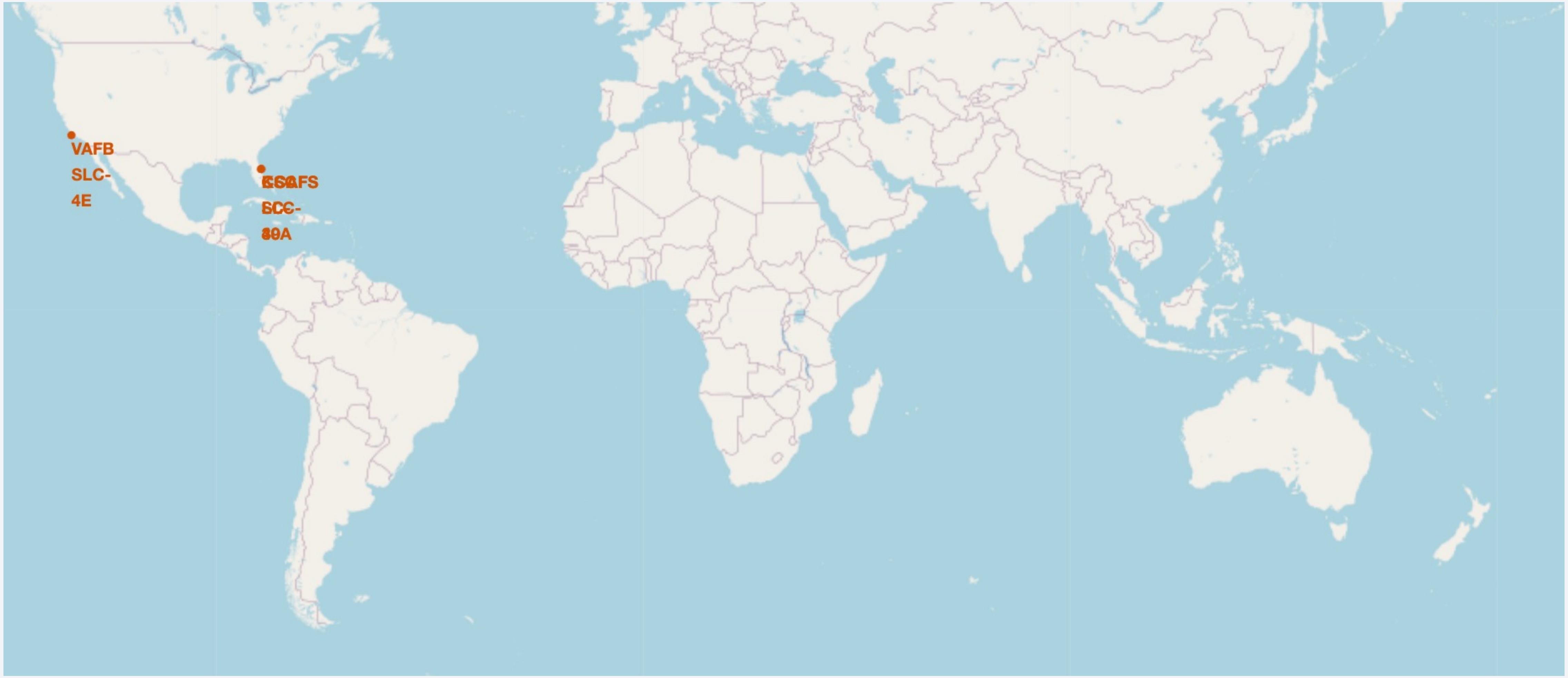
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

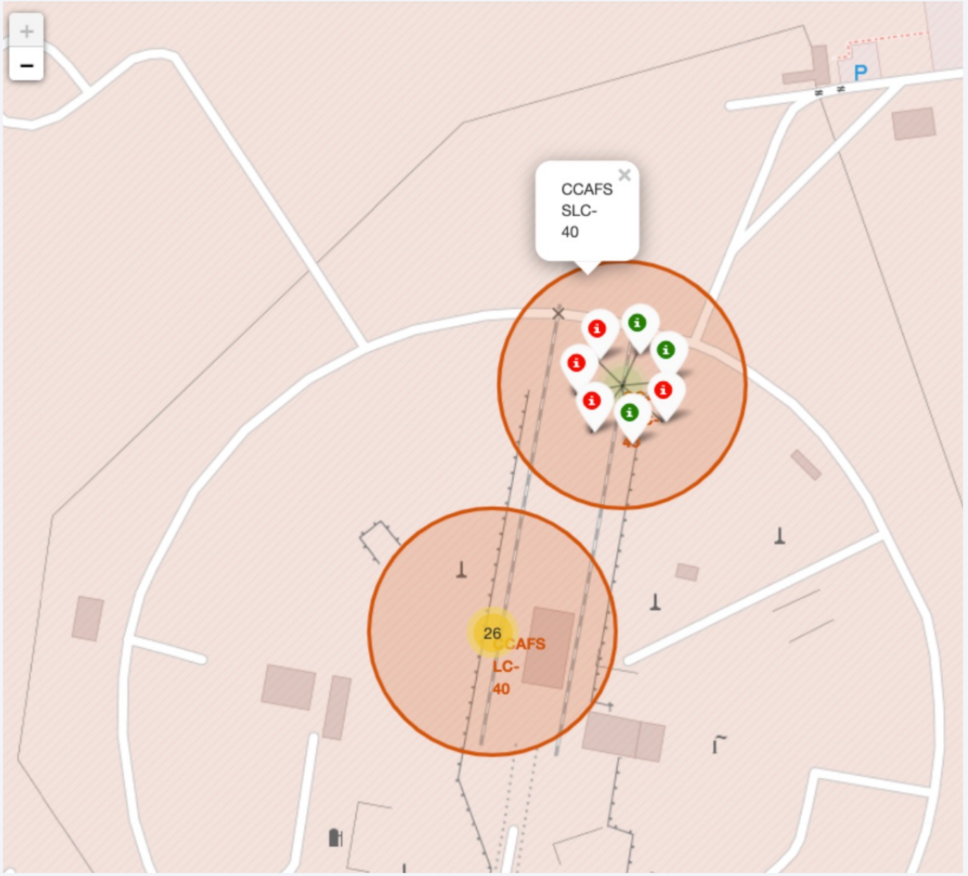
Launch Sites locations

All launch sites are located in the US, on the coasts of Florida and California



Lauch Outcomes Markers

Green Markers show successful launches and Red Markers show failures



Proximity to Railway

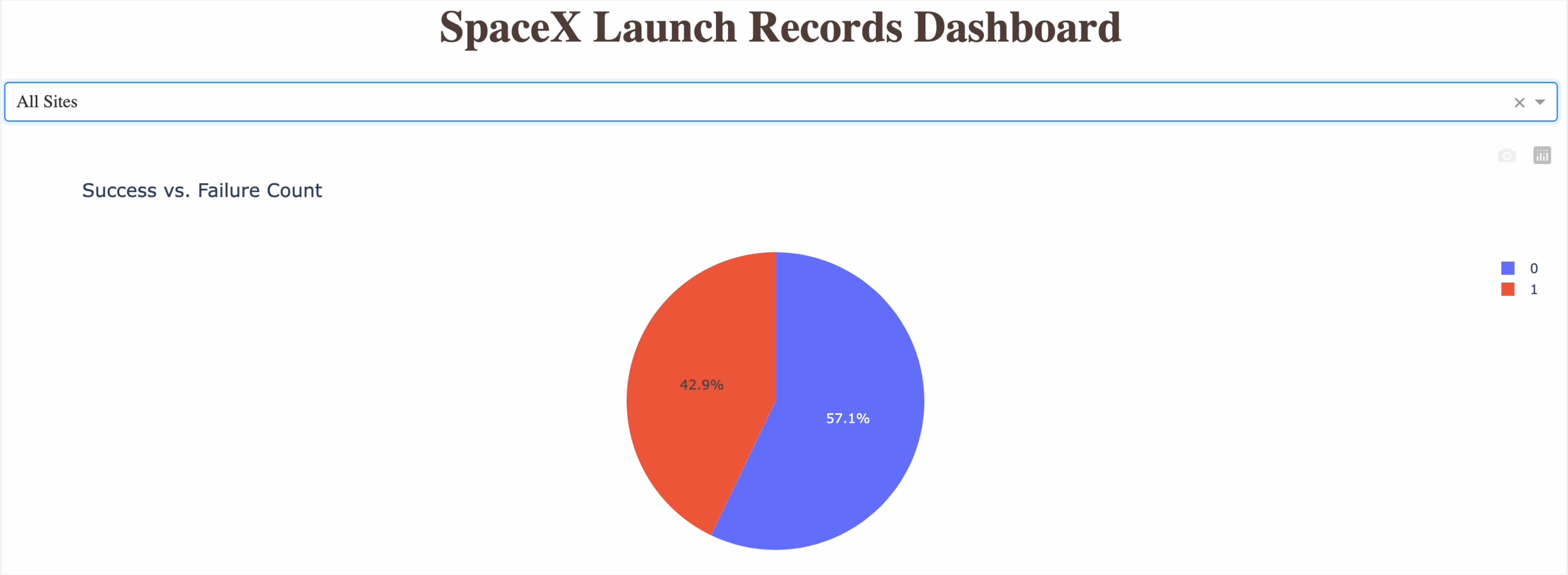
- Florida launch site is located in close proximity to railway

Section 4

Build a Dashboard with Plotly Dash

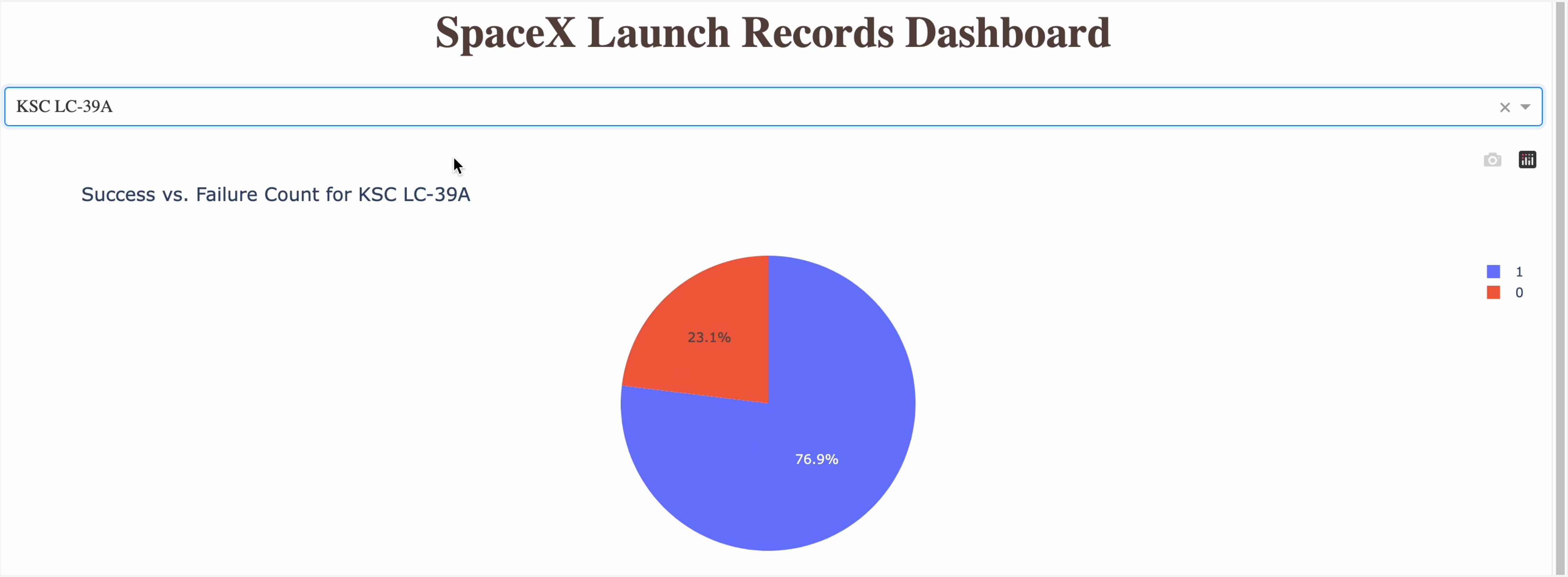
Dashboard: Launch success for all sites

- In total, there is 57.1% success on all launch sites



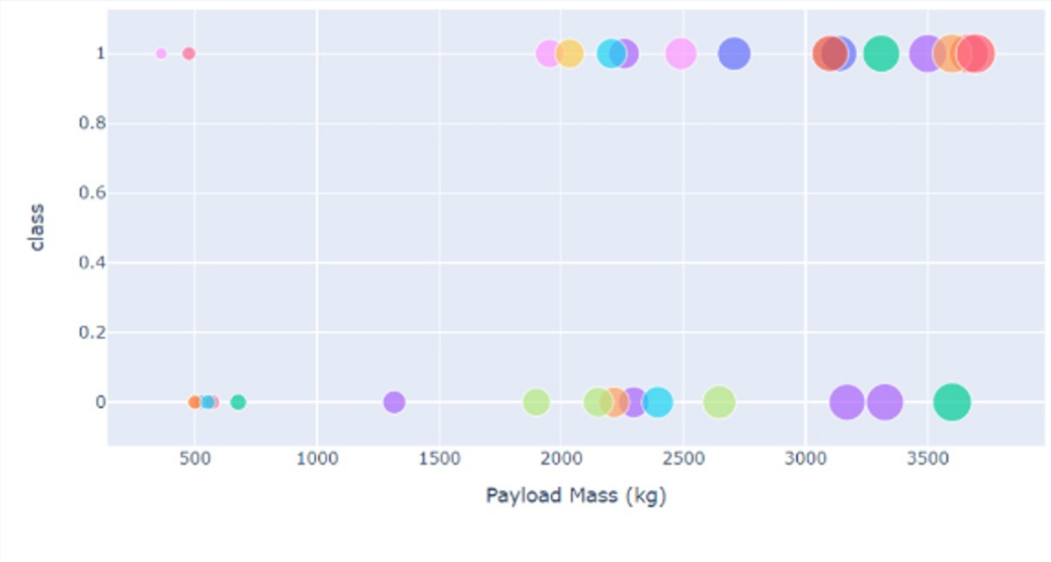
Pie chart showing the Launch site with the highest launch success ratio

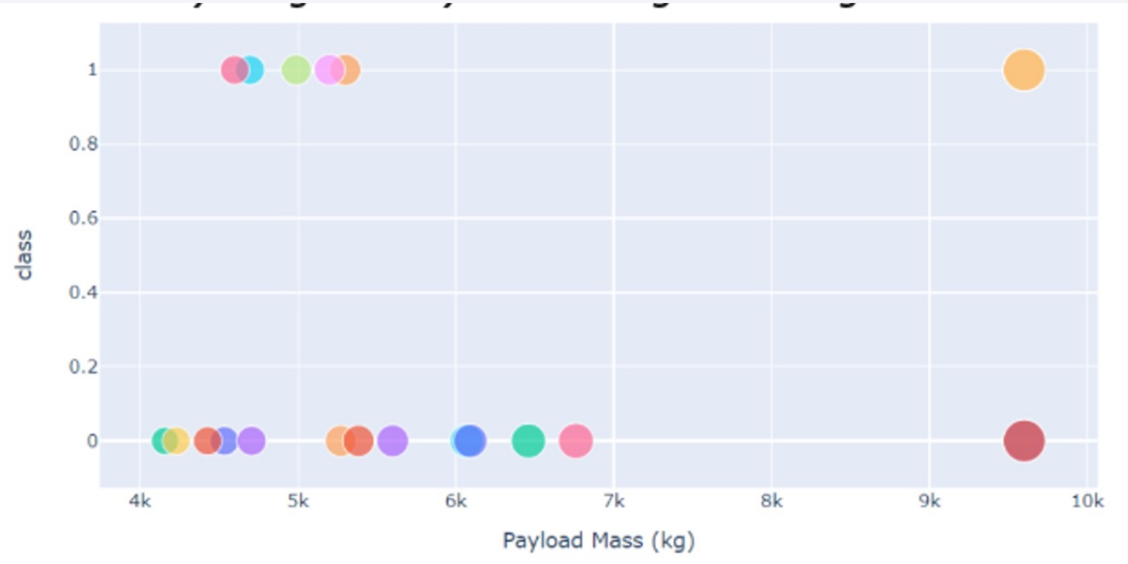
- KSC LC-39A achieved 76.9% success rate



Payload vs. Lauch Outcome (all sites)

- Low weighted payloads are more successful



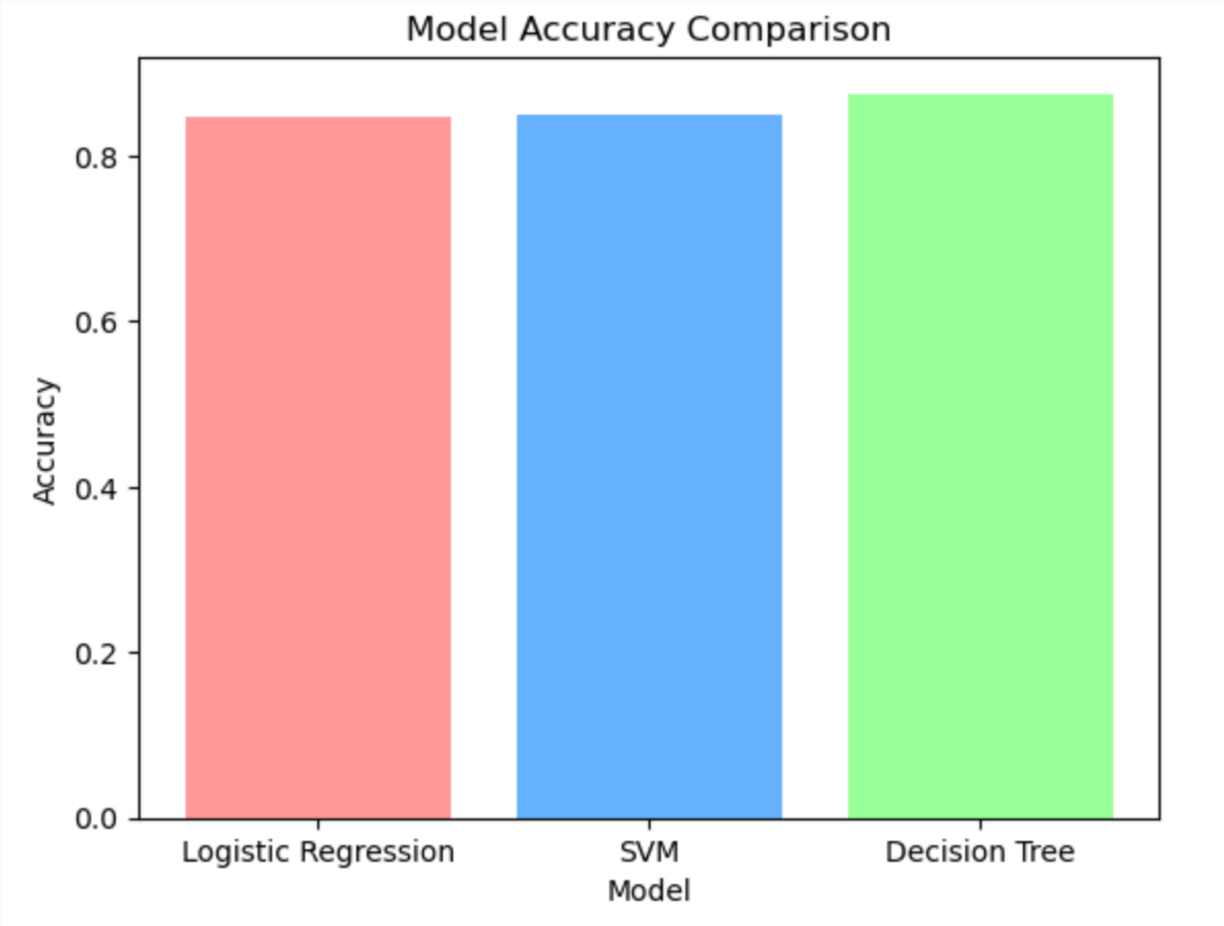


Section 5

Predictive Analysis (Classification)

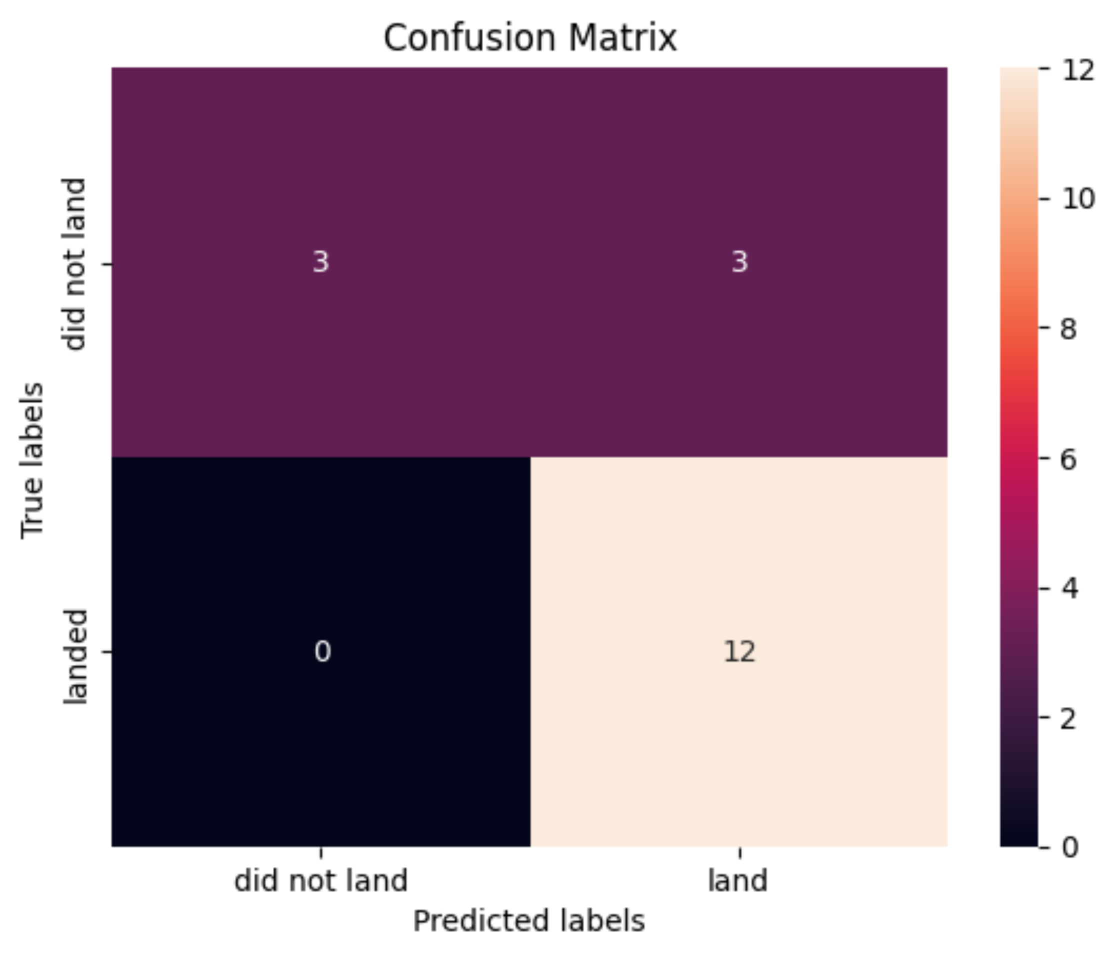
Classification Accuracy

- Decision Tree is the model with the highest classification accuracy of 0.875



Confusion Matrix

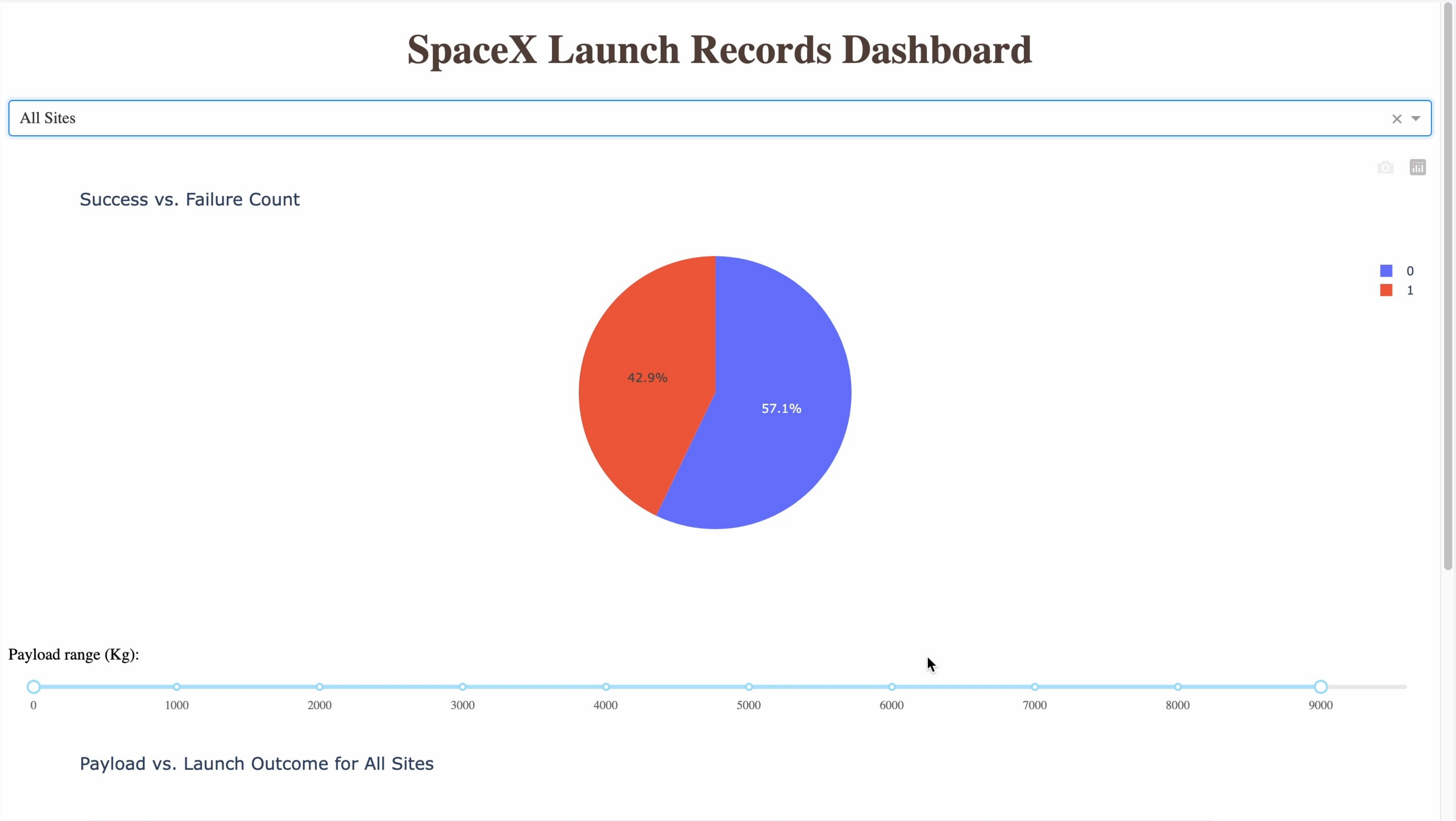
- The model is not perfect, as it incorrectly predicts 50% of unsuccessful landings as false positives



Conclusions

- The factors that predict the outcome of the launch are the Launch Site, Payload and Orbit
- The success rate of launches since 2013 kept increasing until 2020
- KSC LC-39A had the most amount of successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm to predict the success of the launch

Appendix: Dashboard screen recording



Thank you!