

Winning Space Race with Data Science

<Aayush Pratap Bhuyan>
<22 June 2025>



Outline

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

GitHub: <https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage->

Executive Summary

- Summary of methodologies
 - Data Collection through API
 - Data Collection with Web Scraping
 - Data Wrangling
 - Exploratory Data Analysis with Data Visualization
 - Interactive Visual Analytics with Folium
 - Machine Learning Prediction
- Summary of all results
 - Exploratory Data Analysis Result
 - Interactive analytics in screenshots
 - Predictive Analysis Result

Introduction

- Project background and context
 - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- Problems you want to find answers
 - What factors determine if the rocket will land successfully?
 - The interaction amongst various features that determine the success rate of a successful landing.
 - What operating conditions needs to be in place to ensure a successful landing program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
 - One-hot encoding was applied 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
 - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
 - Data collection was done using get request to the SpaceX API.
 - Next, we decoded the response content as a Json using `.json()` function call and turn it into a pandas dataframe using `.json_normalize()`.
 - We then cleaned the data, checked for missing values and fill in missing values where necessary.
 - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
 - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection - SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- GitHub URL:
<https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage/blob/main/spacex-data-collection-api.ipynb>

```
In [10]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_...
```

We should see that the request was successfull with the 200 status response code

```
In [11]: response.status_code
```

```
Out[11]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
In [15]: # Use json_normalize meethod to convert the json result into a dataframe
if response.status_code == 200:
    json_data = response.json()
    # Normalize the JSON data into a flat table
    data = pd.json_normalize(json_data)
    print(data.head())
else:
    print("Failed to retrieve data")
```

Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- GitHub URL:
<https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/Webscraping.ipynb>

In [5]:

```
# 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

In [6]:

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

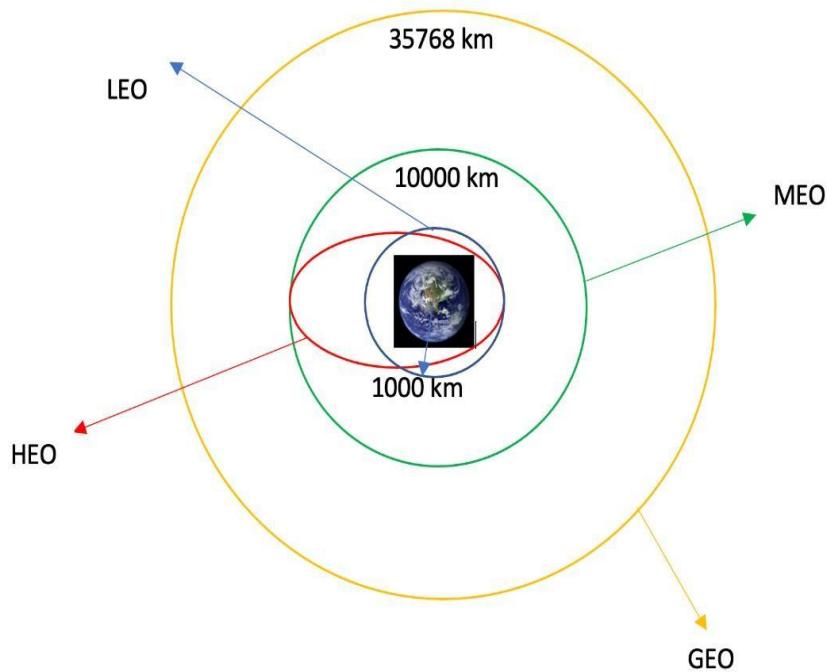
Print the page title to verify if the BeautifulSoup object was created properly

In [7]:

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

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

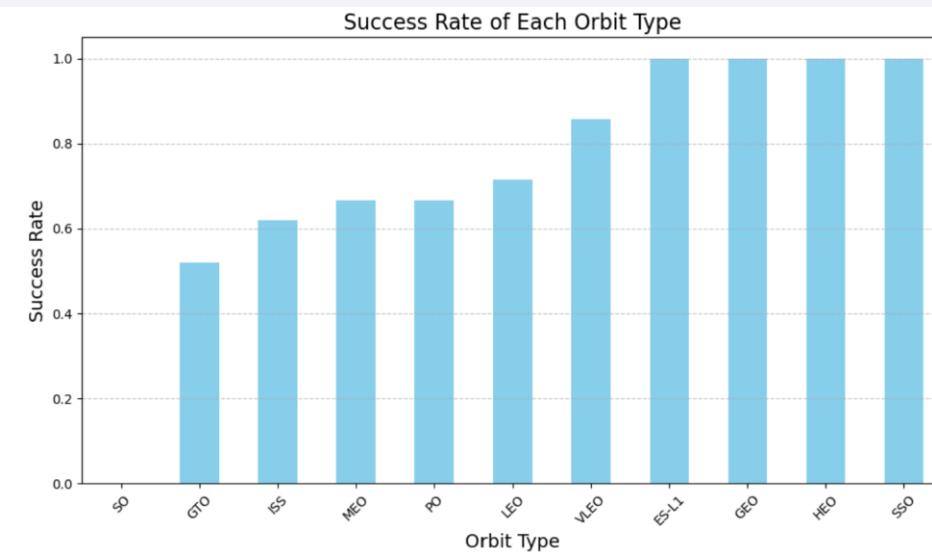
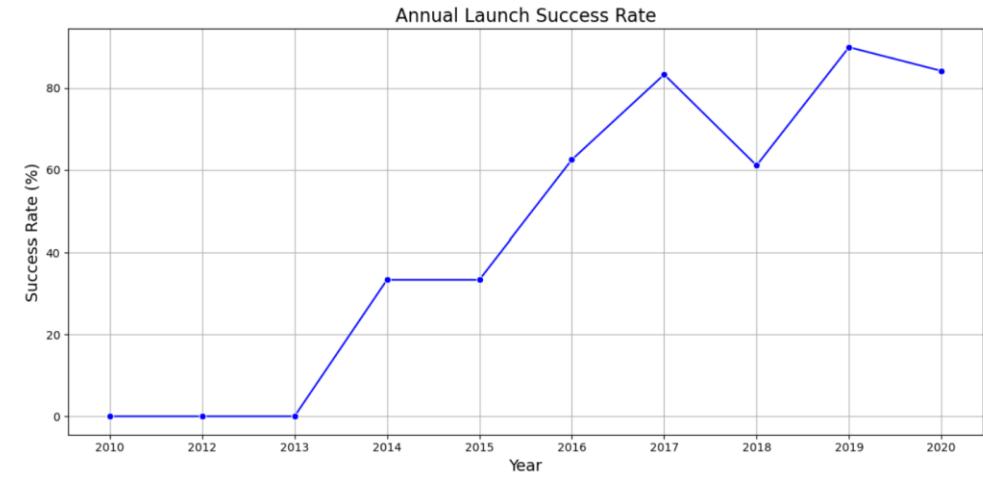
Data Wrangling



- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.
- GitHub URL: <https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/Spacex-Data%20Wrangling.ipynb>

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.
- GitHub URL:
<https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage/blob/main/EDA-dataviz.ipynb>



EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
 - The names of unique launch sites in the space mission.
 - The total payload mass carried by boosters launched by NASA (CRS)
 - The average payload mass carried by booster version F9 v1.1
 - The total number of successful and failure mission outcomes
 - The failed landing outcomes in drone ship, their booster version and launch site names.
- GitHub URL: https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/EDA-SQL_sqlite.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
 - Are launch sites near railways, highways and coastlines.
 - Do launch sites keep certain distance away from cities.
- GitHub URL: <https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/LaunchSiteAnalysis%20-%20Folium.ipynb>

Build a Dashboard with Plotly Dash

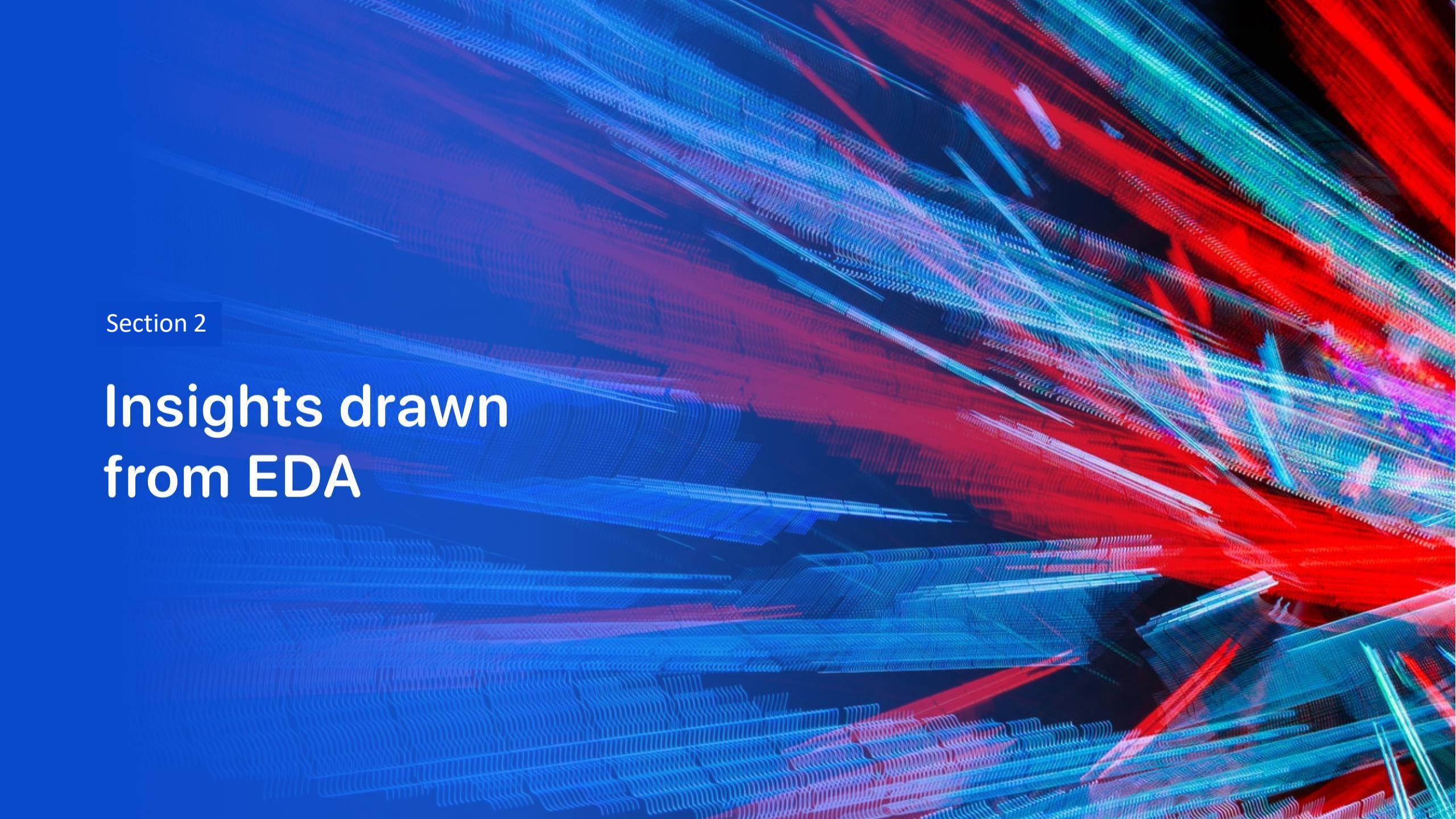
- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- GitHub URL: https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- GitHub URL: <https://github.com/Aayush-Bhuyan/Capstone-Falcon-9---First-Stage-/blob/main/SpaceX-Machine-Learning-Prediction.ipynb>

Results

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

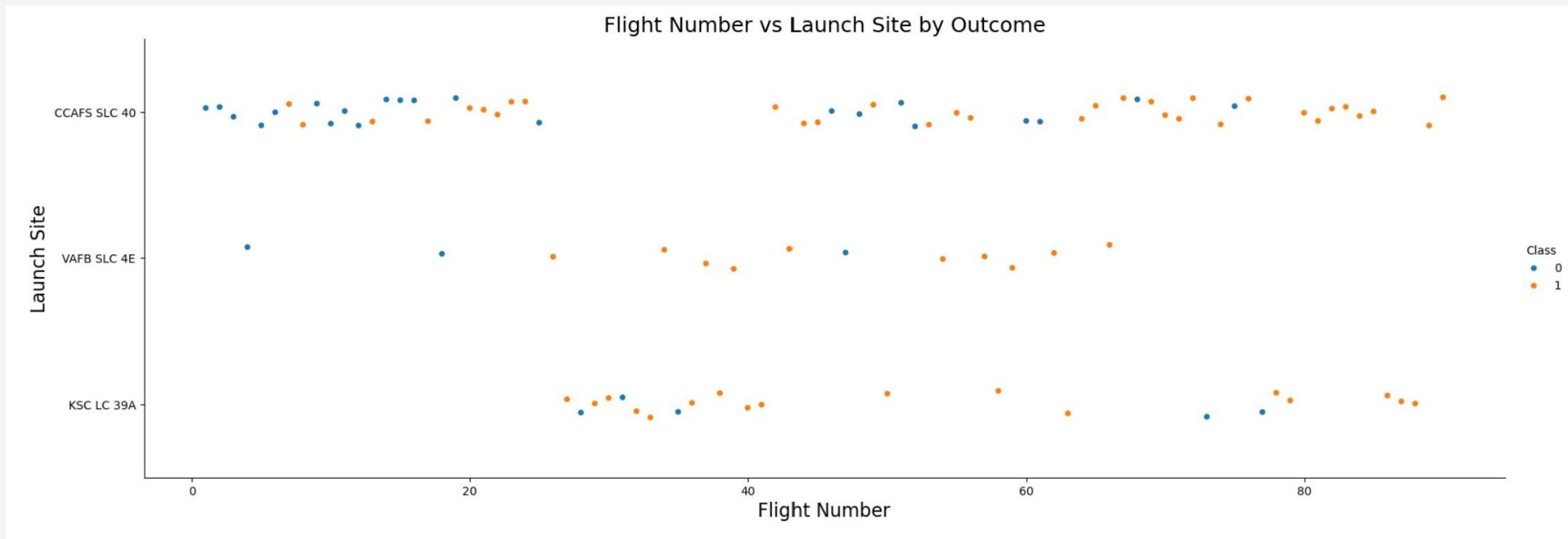
The background of the slide features a complex, abstract pattern of wavy, horizontal lines. These lines are primarily colored in shades of blue, red, and green, creating a sense of depth and motion. They are arranged in several layers, with some lines being more prominent than others. The overall effect is reminiscent of a digital or futuristic landscape.

Section 2

Insights drawn from EDA

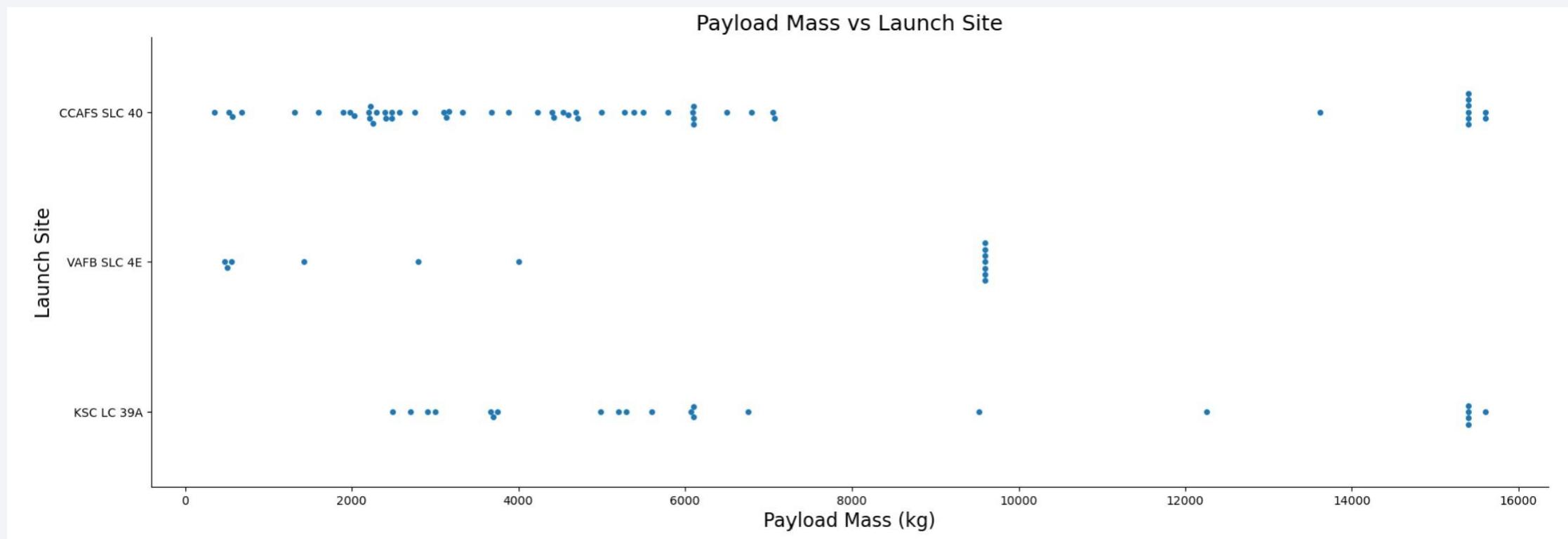
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



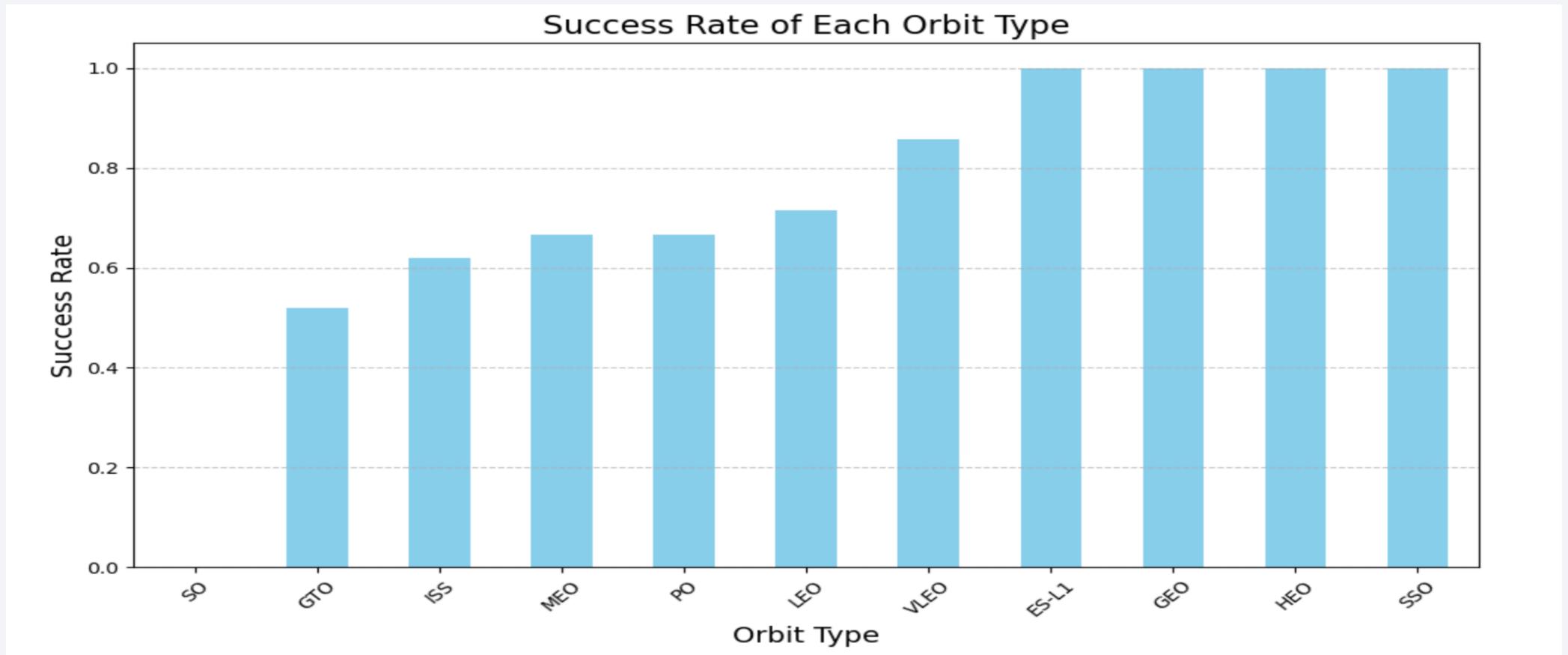
Payload vs. Launch Site

- The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



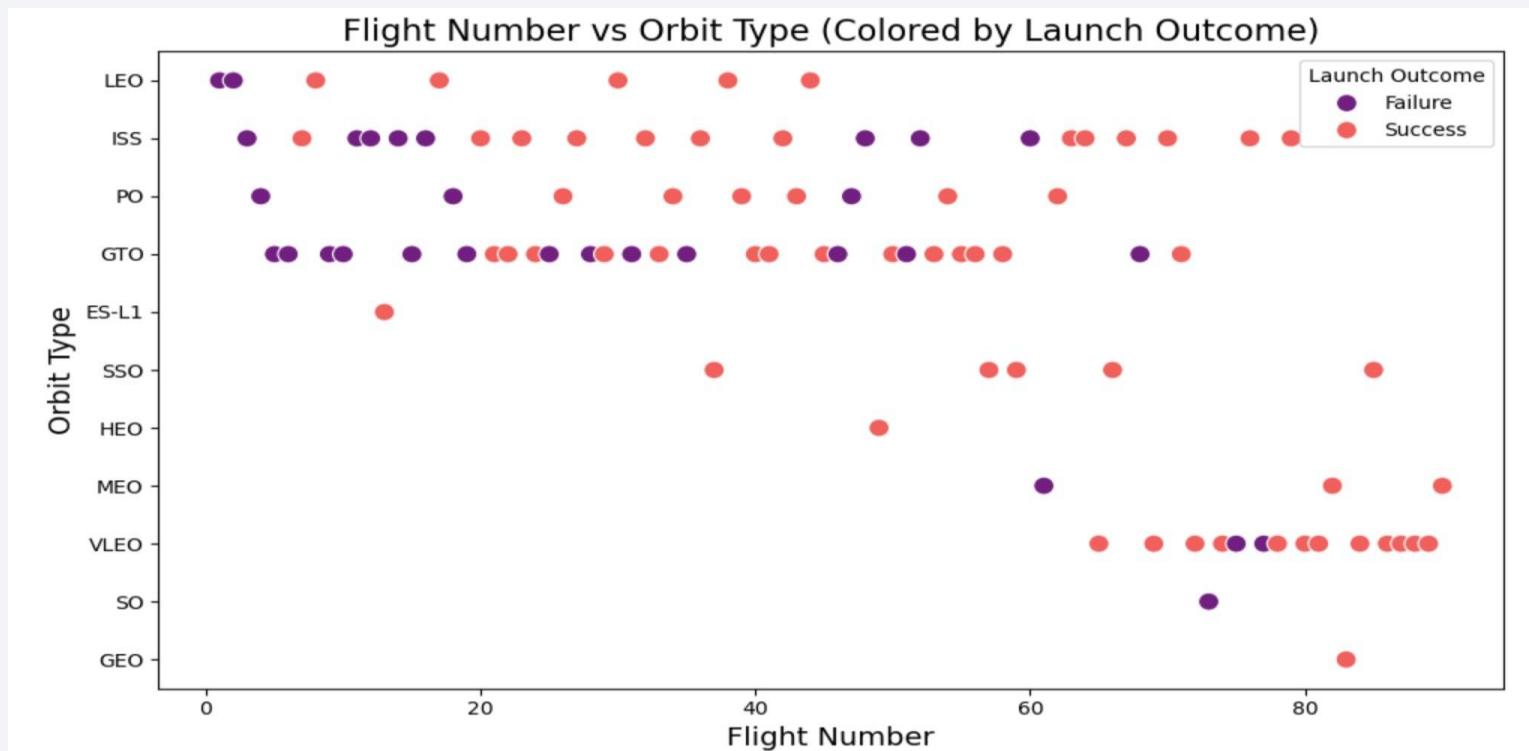
Success Rate vs. Orbit Type

- From the plot, we can see that ESL1, GEO, HEO, SSO, VLEO had the most success rate.



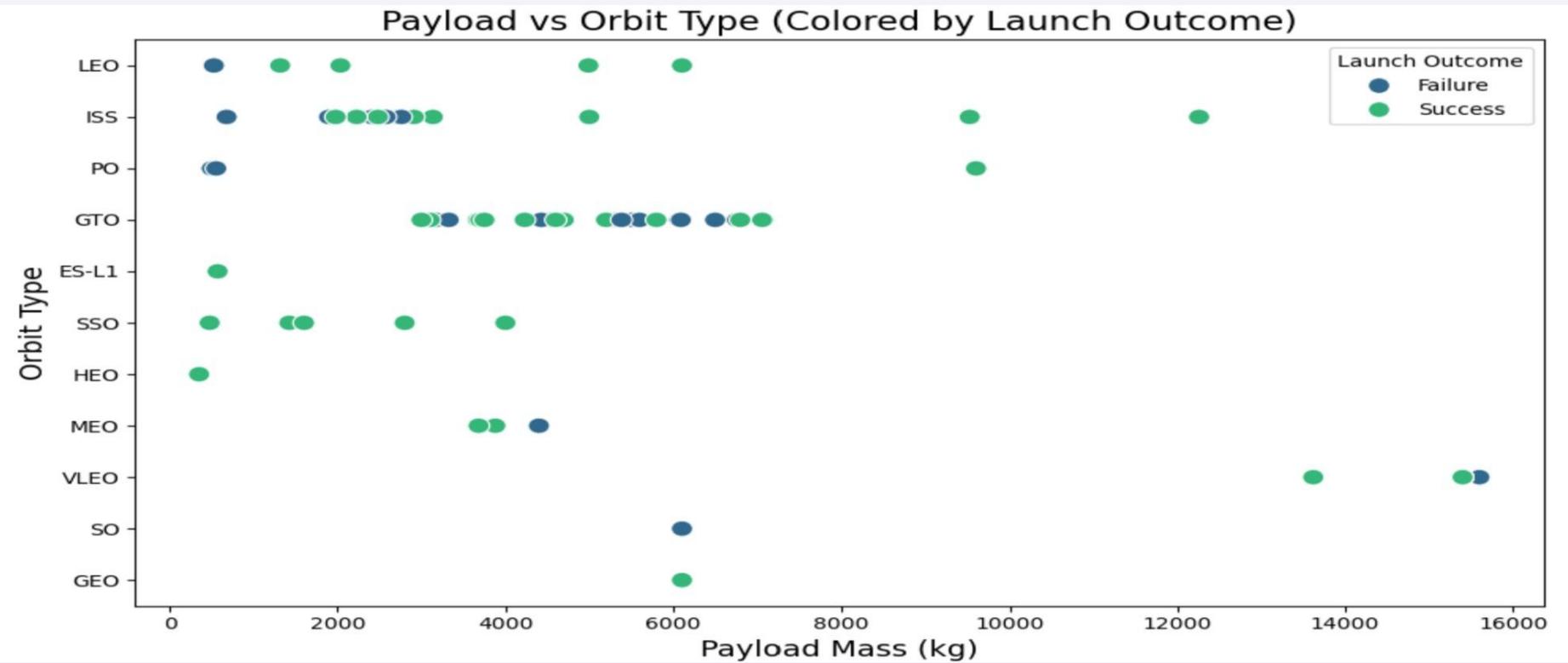
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



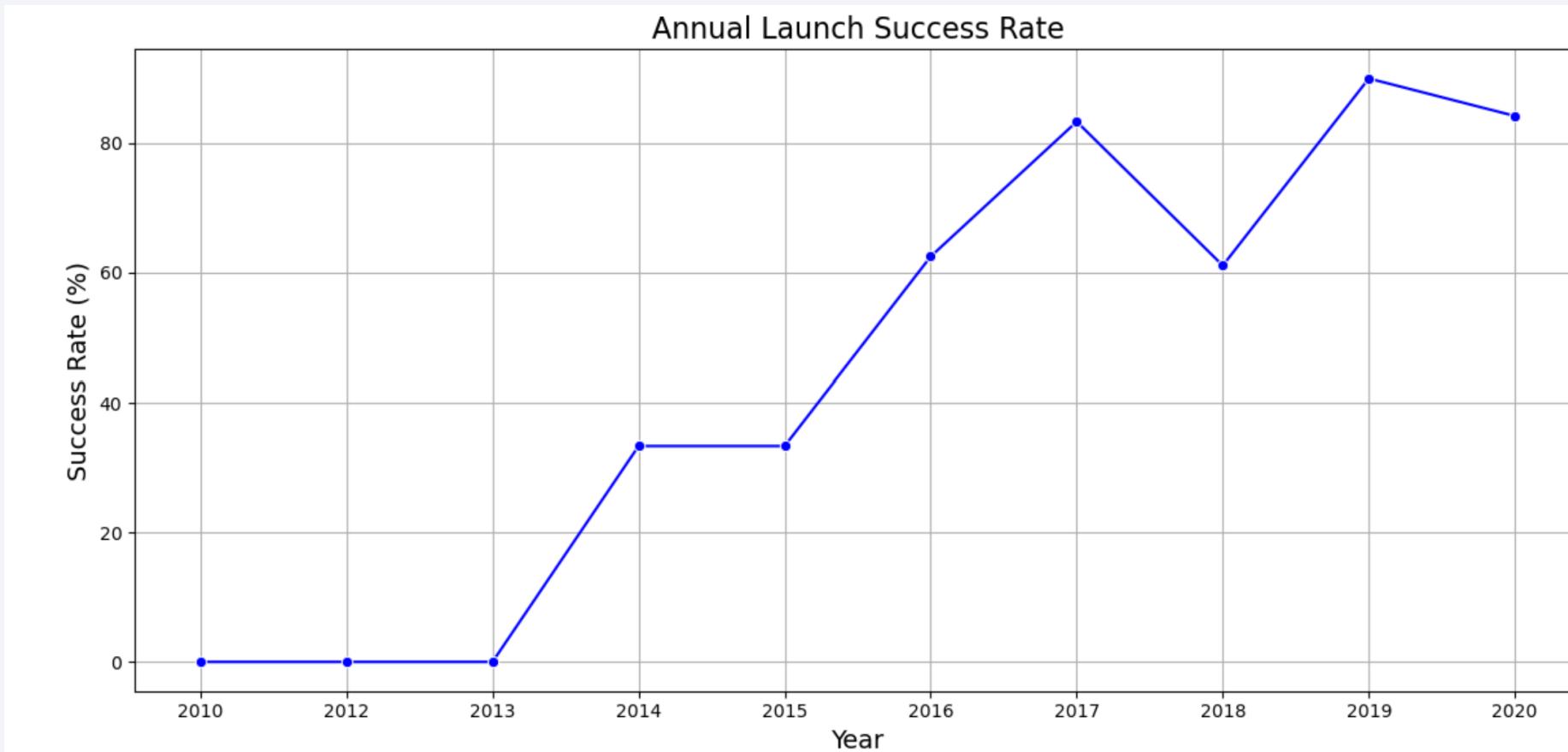
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

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

```
[11]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

```
[11]: Launch_Site
```

```
-----  
CCAFS LC-40
```

```
VAFB SLC-4E
```

```
KSC LC-39A
```

```
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- We used the query above to display 5 records where launch sites begin with `CCA`

```
[12]: %sql SELECT * FROM SPACEXTABLE 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_Outcome |
|------------|------------|-----------------|-------------|---|------------------|-----------|-----------------|-----------------|---------------------|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 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 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
[13]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

```
Done.
```

```
[13]: SUM(PAYLOAD_MASS_KG_)
```

```
45596
```

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
[14]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';  
* sqlite:///my_data1.db  
Done.  
[14]: AVG(PAYLOAD_MASS_KG_)  
2928.4
```

First Successful Ground Landing Date

We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015.

```
[15]: %sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';

* sqlite:///my_data1.db
Done.

[15]: MIN(Date)
-----
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- We 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

```
[16]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000
* sqlite:///my_data1.db
Done.

[16]: Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2
```

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like '%' to filter for WHERE Mission_Outcome was a success or a failure.

```
[17]: %sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTABLE WHERE "Mission_Outcome" IN ('Success', 'Failure') GROUP BY "Mission_Outcome";  
* sqlite:///my_data1.db  
Done.  
[17]: 

| Mission_Outcome | Total |
|-----------------|-------|
| Success         | 98    |


```

Boosters Carried Maximum Payload

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

```
[18]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
* sqlite:///my_data1.db
Done.

[18]: 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

- We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
[19]: %%sql
SELECT
CASE
    WHEN substr("Date", 6, 2) = '01' THEN 'January'
    WHEN substr("Date", 6, 2) = '02' THEN 'February'
    WHEN substr("Date", 6, 2) = '03' THEN 'March'
    WHEN substr("Date", 6, 2) = '04' THEN 'April'
    WHEN substr("Date", 6, 2) = '05' THEN 'May'
    WHEN substr("Date", 6, 2) = '06' THEN 'June'
    WHEN substr("Date", 6, 2) = '07' THEN 'July'
    WHEN substr("Date", 6, 2) = '08' THEN 'August'
    WHEN substr("Date", 6, 2) = '09' THEN 'September'
    WHEN substr("Date", 6, 2) = '10' THEN 'October'
    WHEN substr("Date", 6, 2) = '11' THEN 'November'
    WHEN substr("Date", 6, 2) = '12' THEN 'December'
    ELSE 'Unknown'
END AS "Month_Name",
"Mission_Outcome",
"Booster_Version",
"Launch_Site"
FROM
SPACEXTABLE
WHERE
substr("Date", 0, 5) = '2015';
* sqlite:///my_data1.db
Done.
```

| Month_Name | Mission_Outcome | Booster_Version | Launch_Site |
|------------|---------------------|-----------------|-------------|
| January | Success | F9 v1.1 B1012 | CCAFS LC-40 |
| February | Success | F9 v1.1 B1013 | CCAFS LC-40 |
| March | Success | F9 v1.1 B1014 | CCAFS LC-40 |
| April | Success | F9 v1.1 B1015 | CCAFS LC-40 |
| April | Success | F9 v1.1 B1016 | CCAFS LC-40 |
| June | Failure (in flight) | F9 v1.1 B1018 | CCAFS LC-40 |
| December | Success | F9 FT B1019 | CCAFS LC-40 |

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

- We 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.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

```
[20]: %%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS "Count"
FROM
    SPACEXTABLE
WHERE
    "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    COUNT(*) DESC;
* sqlite:///my_data1.db
Done.

[20]: 

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

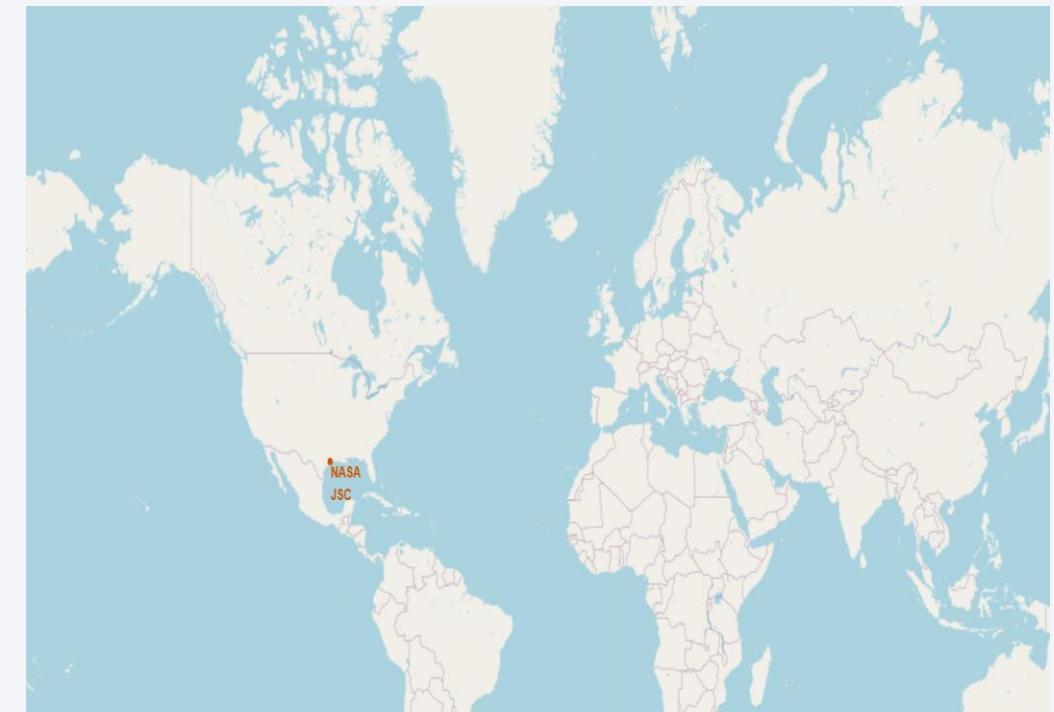
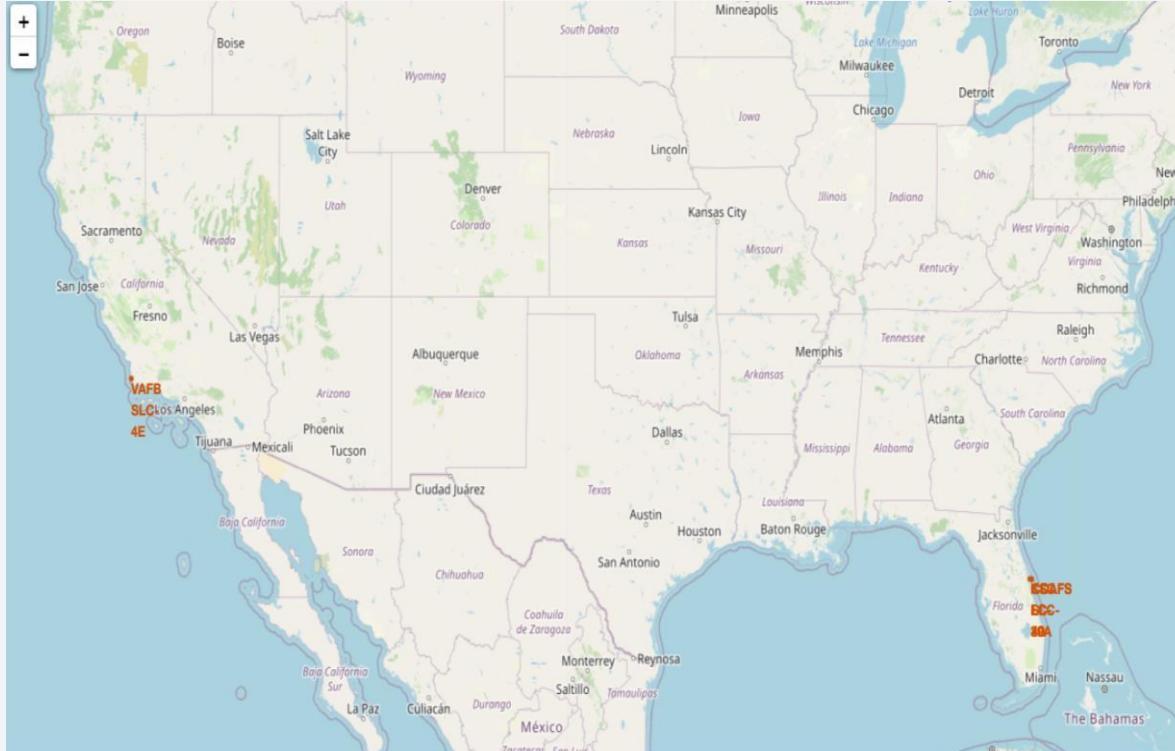

```

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where a large urban area is illuminated. In the upper left quadrant, the green and blue glow of the aurora borealis or a similar atmospheric phenomenon is visible.

Section 3

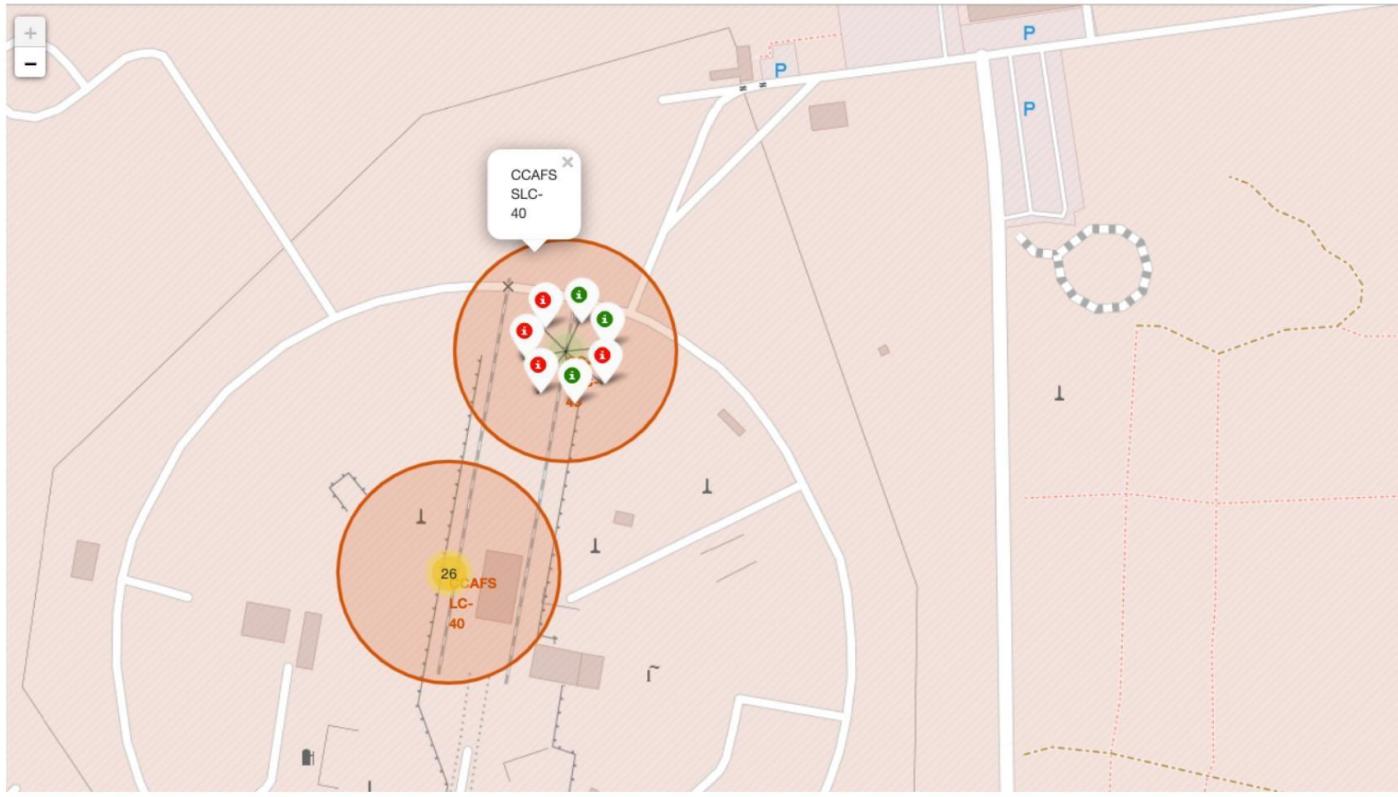
Launch Sites Proximities Analysis

All launch sites global map markers



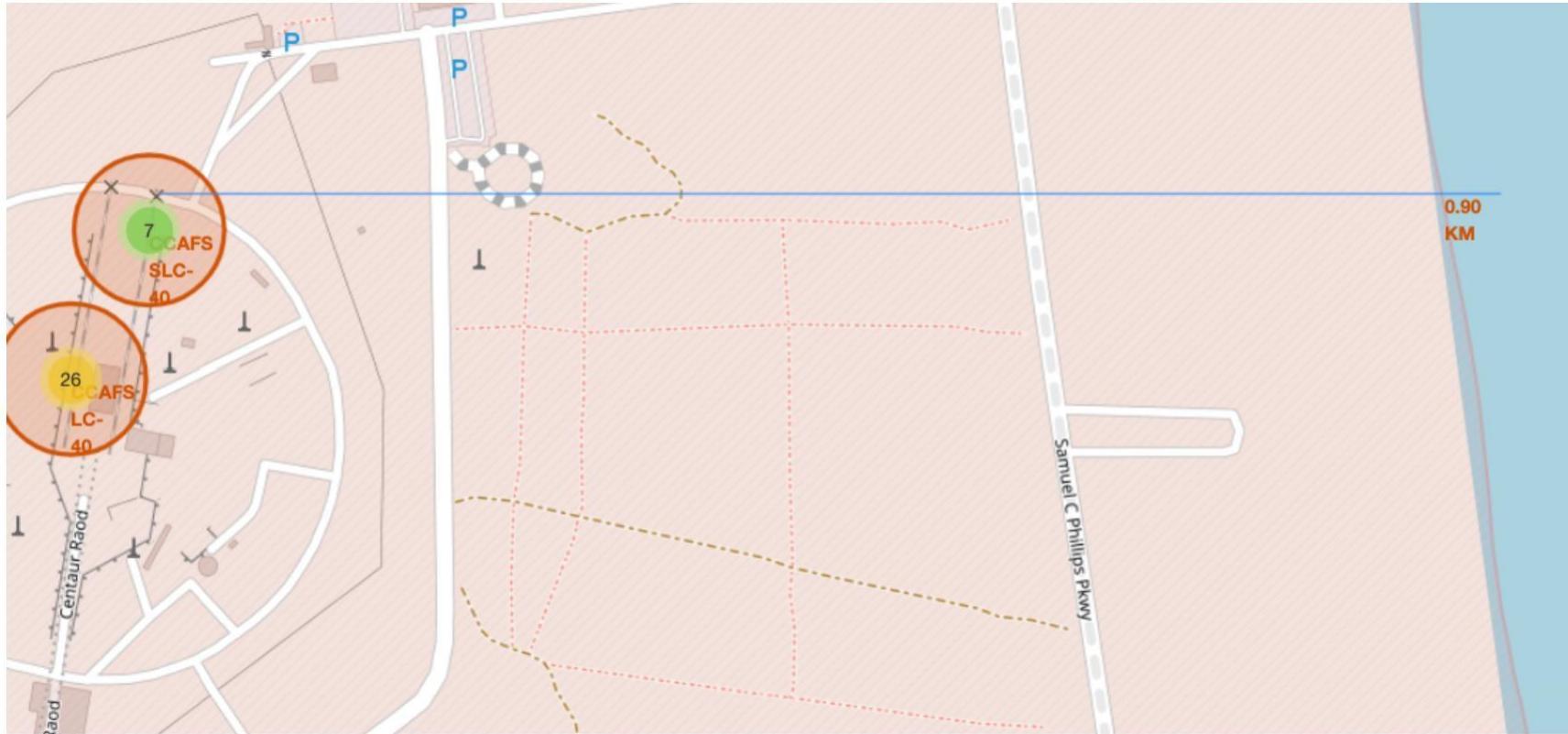
All launch sites are in the USA in Florida and California respectively

Markers showing launch sites with color labels



Launch sites in Florida

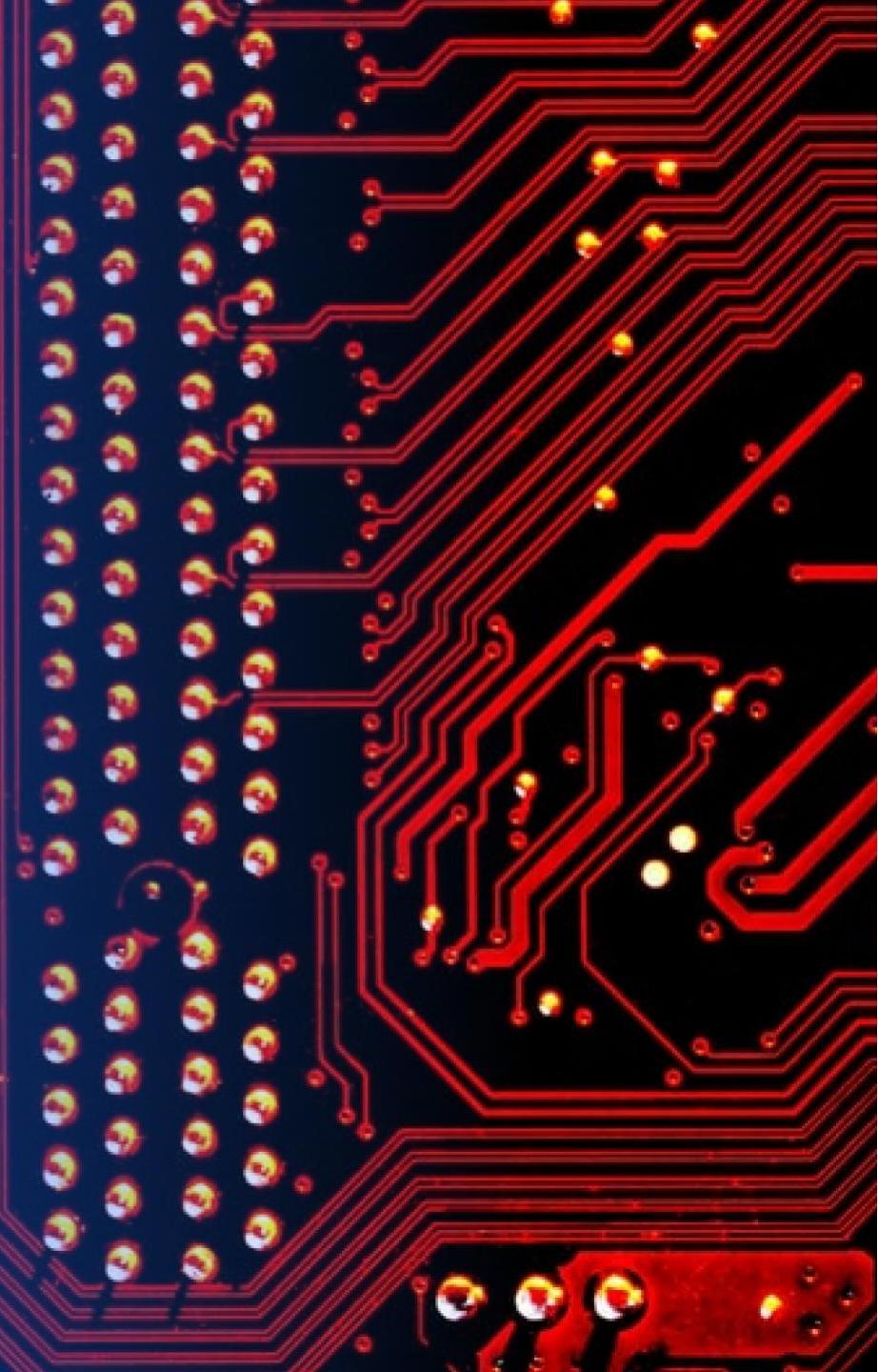
Launch Site distance to landmarks



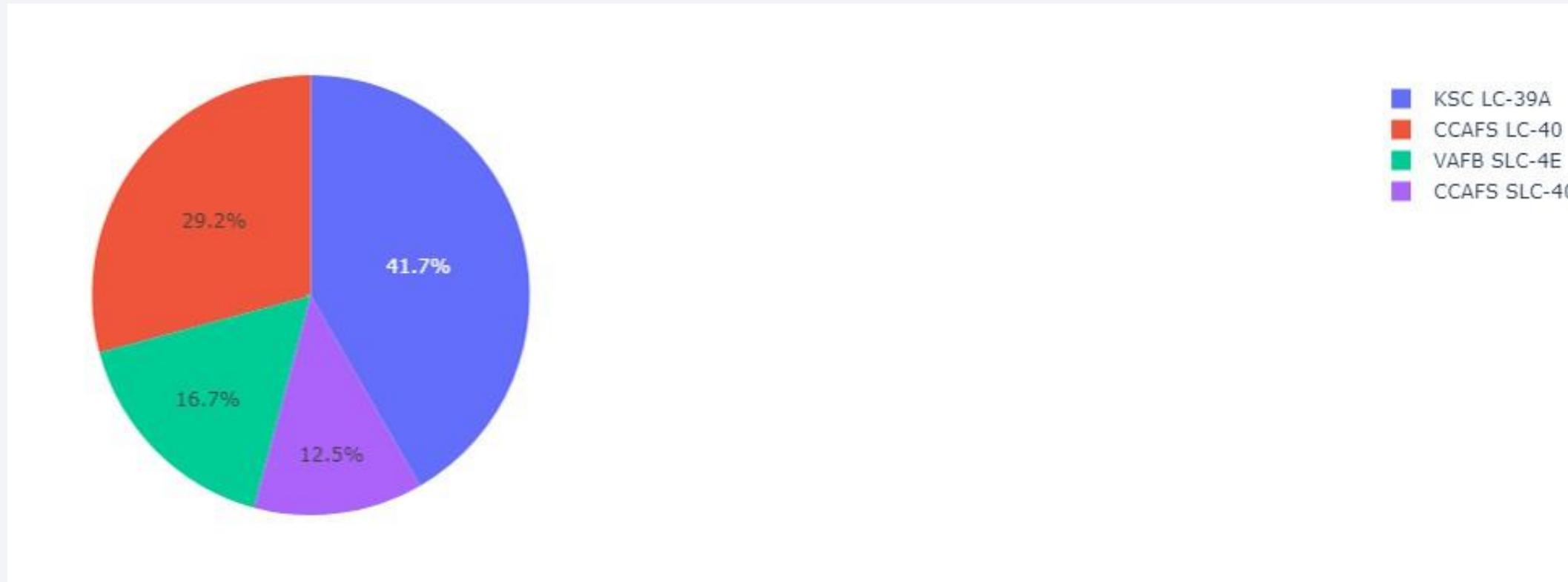
The map shows the distance of the launch sites from the coast

Section 4

Build a Dashboard with Plotly Dash

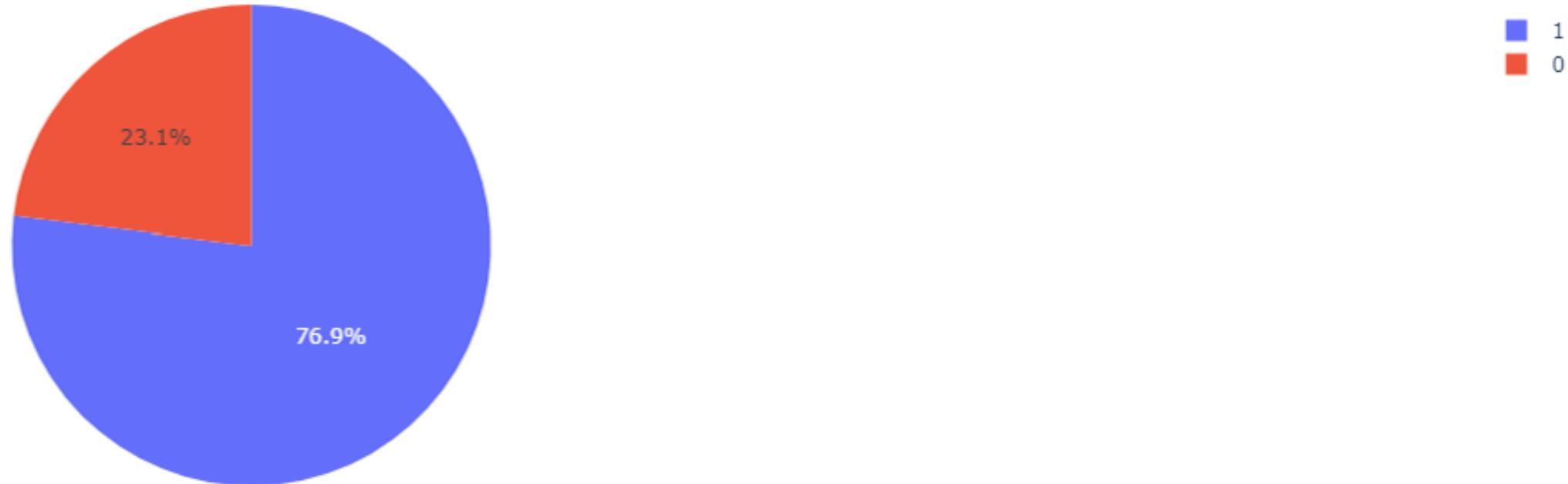


Pie chart showing the success percentage achieved by each launch site



- We can see that KSC LC-39A has the highest success

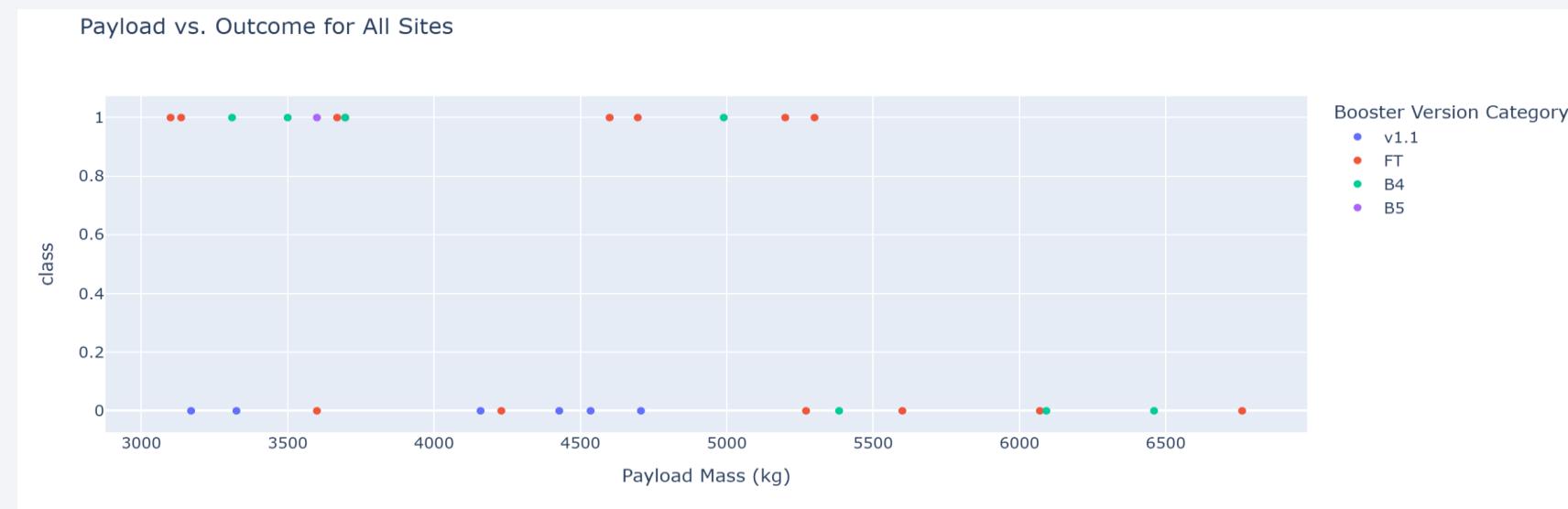
Pie chart showing the Launch site with the highest launch success ratio



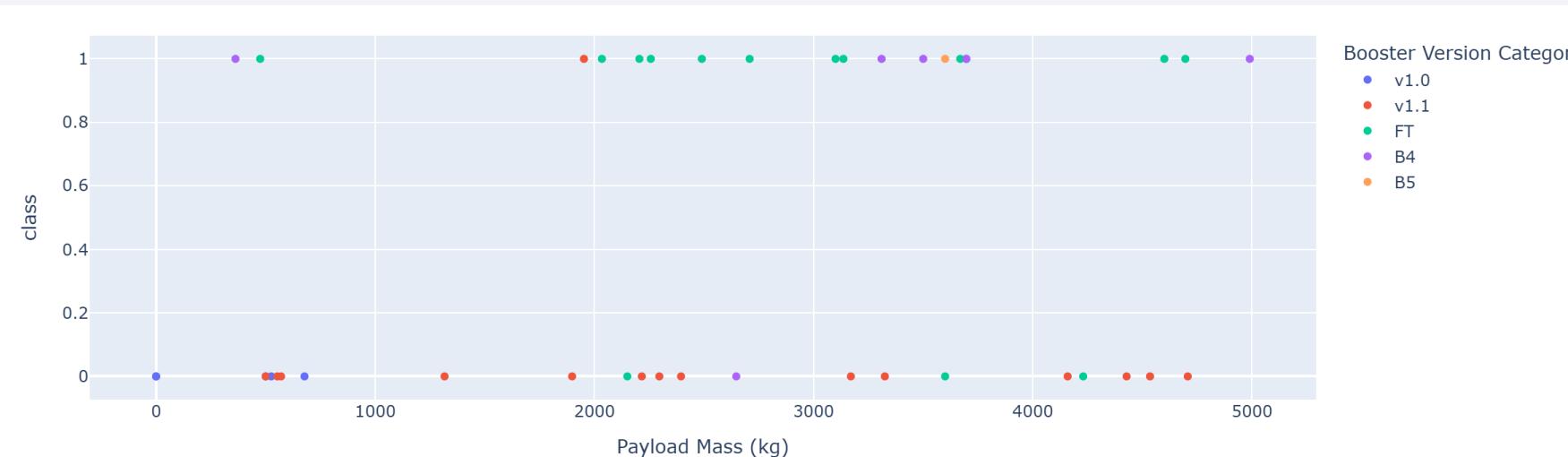
KSC LC-39A achieved a success rate of 76.9%

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

3000 kg – 6500 kg Payload



0 – 5000 kg Payload



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree has the highest accuracy

```
In [45]: import matplotlib.pyplot as plt

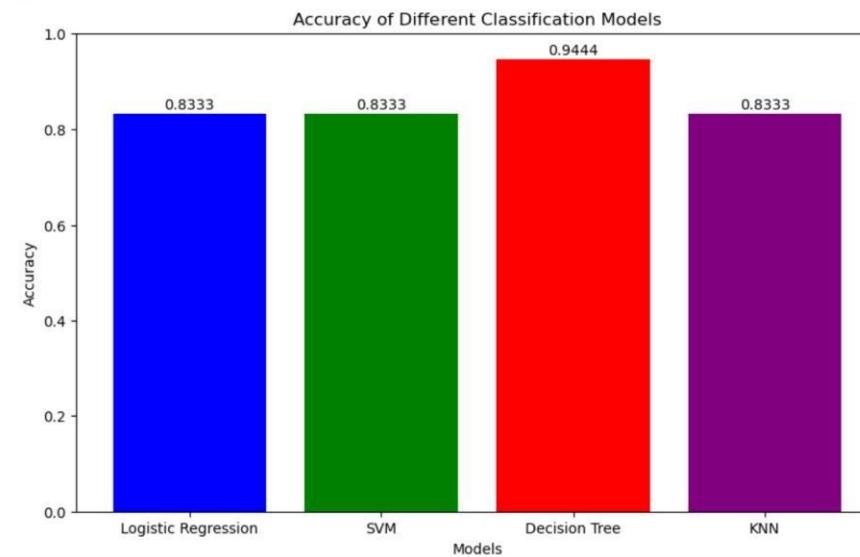
# Model names and their corresponding accuracies
models = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']
accuracies = [0.8333, 0.8333, 0.9444, 0.8333]

# Create a bar chart
plt.figure(figsize=(10, 6))
plt.bar(models, accuracies, color=['blue', 'green', 'red', 'purple'])

# Add title and labels
plt.title('Accuracy of Different Classification Models')
plt.xlabel('Models')
plt.ylabel('Accuracy')
plt.ylim(0, 1) # Setting the y-axis range from 0 to 1

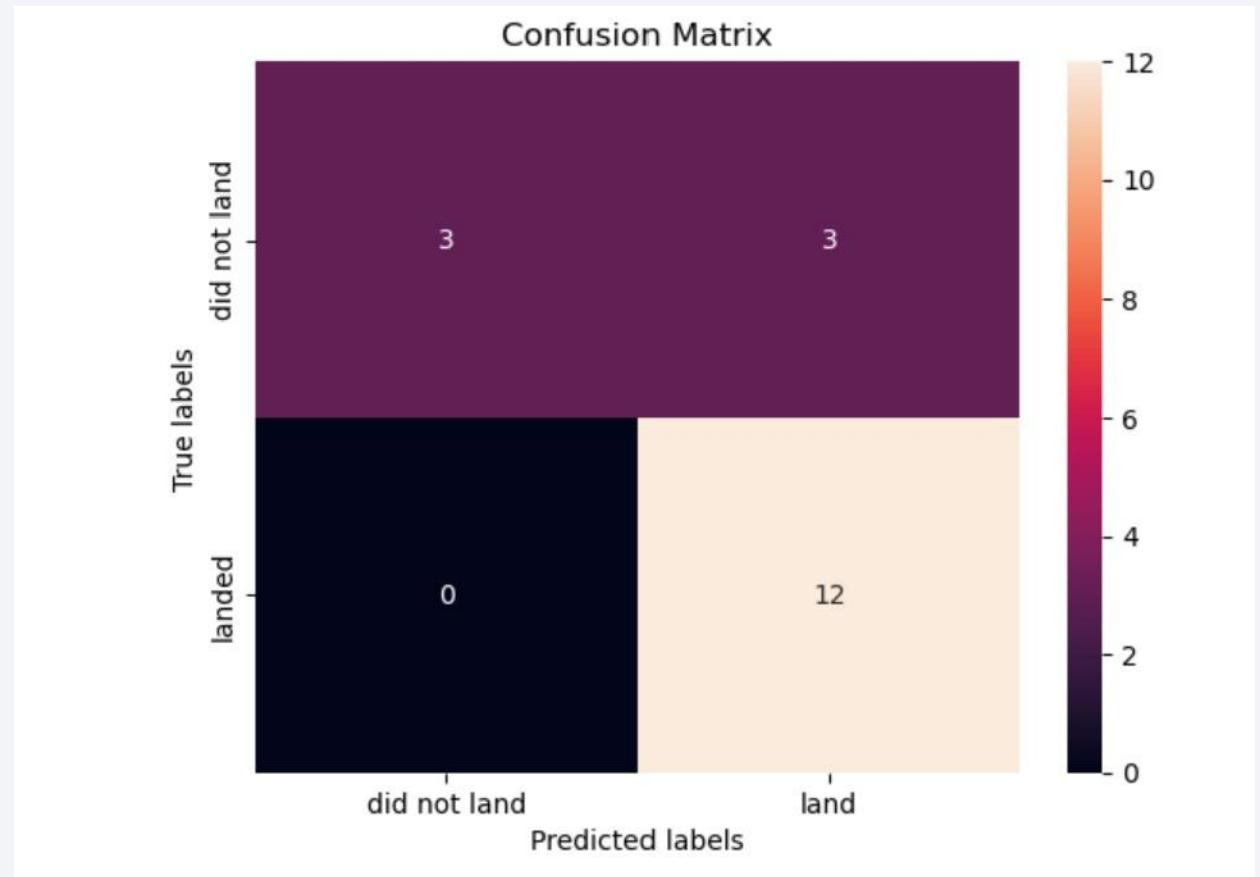
# Add accuracy values on top of the bars
for i, accuracy in enumerate(accuracies):
    plt.text(i, accuracy + 0.01, f'{accuracy:.4f}', ha='center')

# Show the plot
plt.show()
```



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- We can conclude that:
- The larger the flight amount at a launch site, the greater the success rate at a launchsite.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

```
4      import dash_html_components as html
5      import dash_core_components as dcc
6      from dash.dependencies import Input, Output
7      import plotly.express as px
8
9      # Read the airline data into pandas dataframe
10     spacex_df = pd.read_csv("spacex_launch_dash.csv")
11     max_payload = spacex_df['Payload Mass (kg)'].max()
12     min_payload = spacex_df['Payload Mass (kg)'].min()
13
14     # Create a dash application
15     app = dash.Dash(__name__)
16
17     # Create an app layout
18     app.layout = html.Div(children=[html.H1('SpaceX Launch Records Dashboard',
19                                         style={'textAlign': 'center', 'color': '#503D36',
20                                         'font-size': 40}),
21                                         # TASK 1: Add a dropdown list to enable Launch Site selection
22                                         # The default select value is for ALL sites
23                                         dcc.Dropdown(id='site-dropdown',
24                                         options=[
25                                             {'label': 'All sites', 'value': 'ALL'},
26                                             {'label': 'CCAFS LC-40', 'value': 'CCAFS LC-40'},
27                                             {'label': 'VAFB SLC-4E', 'value': 'VAFB SLC-4E'},
28                                             {'label': 'KSC LC-39A', 'value': 'KSC LC-39A'},
29                                             {'label': 'CCAFS SLC-40', 'value': 'CCAFS SLC-40'}],
30                                         value='ALL',
31                                         placeholder="Select a Launch Site here",
32                                         searchable=True
33                                         ),
34                                         html.Br(),
35
36                                         # TASK 2: Add a pie chart to show the total successful launches count for all sites
37                                         # If a specific launch site was selected, show the success vs. failed counts for the site
38                                         html.Div(dcc.Graph(id='success-pie-chart')),
39                                         html.Br(),
40
41                                         html.P("Payload range (kg):"),
42                                         # TASK 3: Add a slider to select payload range
43                                         dcc.RangeSlider(id='payload-slider',
44                                         min=0, max=10000, step=1000,
45                                         marks={0: '0', 100: '100'},
46                                         value=[min_payload, max_payload]),
47
48                                         # TASK 4: Add a scatter chart to show the correlation between payload and launch success
49                                         html.Div(dcc.Graph(id='success-payload-scatter-chart')),
50                                         ])
```

Thank you!

