IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Clifford Nyoni
21/05/25
https://github.com/Cliffnyoni

Outline

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

Executive Summary

- **Summary of Methodologies**
- -Data Collection
- -Data Wrangling
- -EDA with SQL
- -EDA with Visualisation
- -Building an Interactive map with folium
- -Building a Dashboard with Plotly Dash

Results Summary

- -EDA Results
- -Interactive Analytics
- -Predictive Analysis

Introduction

- **Project background and context.**
- **-**SpaceX promotes its Falcon 9 rocket launches on its website, highlighting a cost of 62millionperlaunchsignificantlylowerthancompetitors,whochargeupwardsof62millionperlaunch—significantlylowerthancompetitors,whochargeupwardsof165 million. This substantial cost reduction is largely due to SpaceX's ability to reuse the Falcon 9's first stage, a key advantage over traditional rocket providers that rely on expendable vehicles.

- **Problems you want to find answers**
- **-**By determining whether the first stage of a Falcon 9 rocket will successfully land, we can estimate the cost of a launch. To achieve this, we will train a machine learning model using historical SpaceX launch data to predict the likelihood of a successful first-stage landing. This prediction will help assess potential launch costs, as reusability significantly reduces expenses compared to expendable rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- -Use API requests on Spacex REST API.
- - Webscrape SpaceX Wikipedia entry.
- -Combine into dataframe to get complet dataset.
- Perform data wrangling:
- -Data cleaning (e.g. replacing missing payload mass with its mean; removing irrelevant columns)
  -Create a landing outcome label
  -Data transformation (One Hot Encoding for categorical features, Standardization of numerical 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:
- -Linear Regression, K Nearest Neighbors, Support Vector Machine, and Decision Tree models have been built and evaluated for the best classifier

Data Collection

-Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX’s Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

-**Data Columns are obtained by using SpaceX REST API:** FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

-**Data Columns are obtained by using Wikipedia Web Scraping:** Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date and Time .

Data Collection – SpaceX API

- We fetched data from the SpaceX API via a GET request, cleaned the retrieved dataset, and applied foundational data wrangling and formatting techniques.

1. Get request for rocket launch data using API

In [6]:
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

In [7]:
```
response = requests.get(spacex_url)
```

2. Use json_normalize method to convert json result to dataframe

In [12]:
```
# Use json_normalize method to convert the json result into a dataframe

# decode response content as json
static_json_df = res.json()
```

In [13]:
```
# apply json_normalize
data = pd.json_normalize(static_json_df)
```

3. We then performed data cleaning and filling in the missing values

In [30]:
```
rows = data_falcon9['PayloadMass'].values.tolist()[0]

df_rows = pd.DataFrame(rows)
df_rows = df_rows.replace(np.nan, PayloadMass)

data_falcon9['PayloadMass'][0] = df_rows.values
data_falcon9
```

Data Collection - Scraping

- Using BeautifulSoup, we scraped and parsed Falcon 9 launch records from a webpage, then transformed the extracted table data into a pandas DataFrame for further processing.

1. Apply HTTP Get method to request the Falcon 9 rocket launch page

In [4]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

In [5]:
```
# use requests.get() method with the provided static_url
# assign the response to a object
html_data = requests.get(static_url)
html_data.status_code
```

Out[5]:
```
200
```

2. Create a BeautifulSoup object from the HTML response

In [6]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(html_data.text, 'html.parser')
```

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

In [7]:
```
# Use soup.title attribute
soup.title
```

Out[7]:
```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

3. Extract all column names from the HTML table header

In [10]:
```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header() to get a column name
# Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called column_names

element = soup.find_all('th')
for row in range(len(element)):
    try:
        name = extract_column_from_header(element[row])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

4. Create a dataframe by parsing the launch HTML tables
5. Export data to csv

Data Wrangling

We conducted exploratory data analysis (EDA) to examine the dataset and establish training labels. Our analysis included:

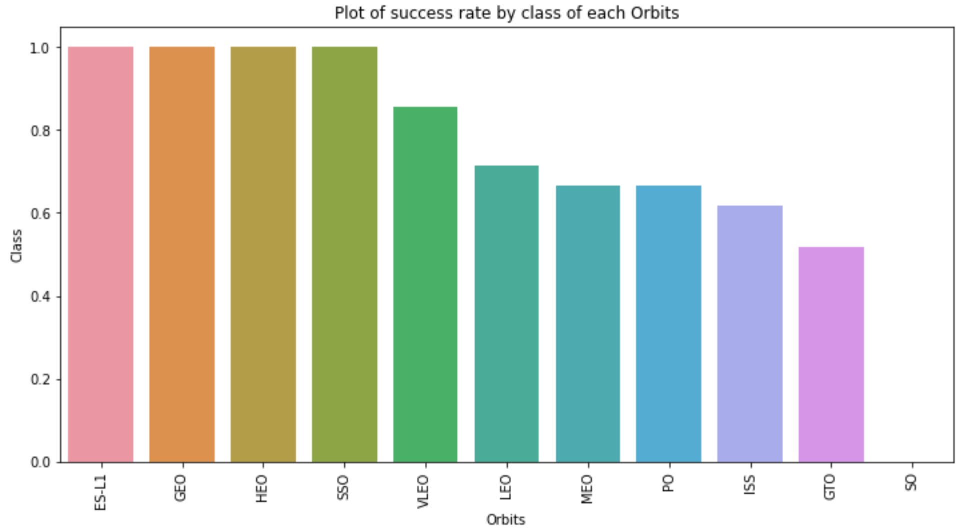
Calculating launch frequencies by site

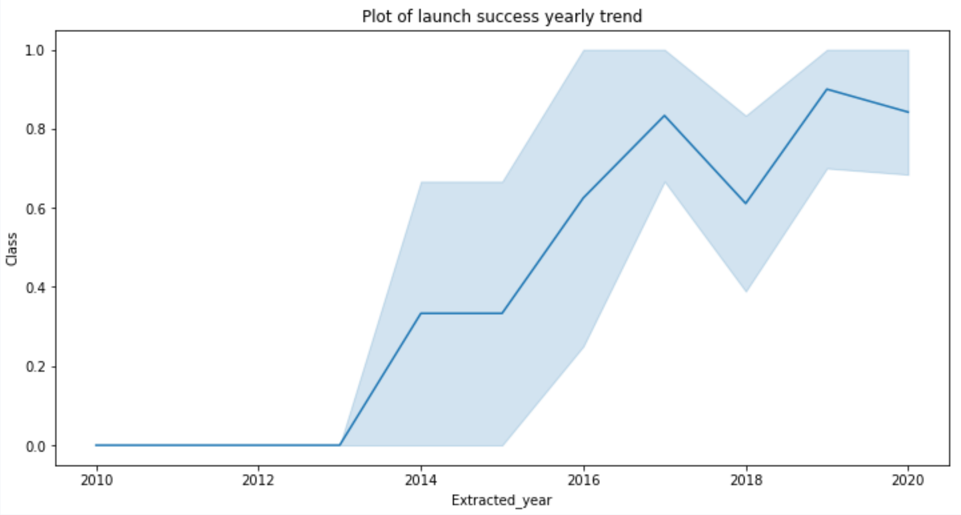
Determining orbit type distributions (both counts and occurrence rates)

Creating landing outcome labels derived from the outcome column

The processed results were then exported to a CSV file for further use.

EDA with Data Visualization





-To gain insights into the dataset, we visualized key relationships, including: The distribution of flight numbers across launch sites, The correlation between payload mass and launch site, The success rate of different orbit types, The trend in flight numbers by orbit type over time, The yearly trend in launch success rates.

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.

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.

Build a Dashboard with Plotly Dash

We developed an interactive dashboard using Plotly Dash to visualize space launch data. The dashboard features:

Pie Charts: Visualizing the distribution of total launches by launch site, providing an overview of each site's contribution to space missions.

Scatter Plots: Illustrating the relationship between mission outcome (success/failure) and payload mass (in kilograms), with data points segmented by booster version. This visualization helps identify potential patterns between payload capacity and mission success across different rocket variants.

The dashboard enables users to interactively explore launch statistics and analyze trends across different launch sites and vehicle configurations.

Predictive Analysis (Classification)

-We began by loading and preprocessing our dataset using NumPy and Pandas. After performing necessary data transformations, we split the dataset into training and testing sets to enable proper model evaluation.

-We then developed multiple machine learning models and systematically optimized their hyperparameters using GridSearchCV for comprehensive parameter tuning. Throughout the model development process, we used accuracy as our primary evaluation metric.

Results

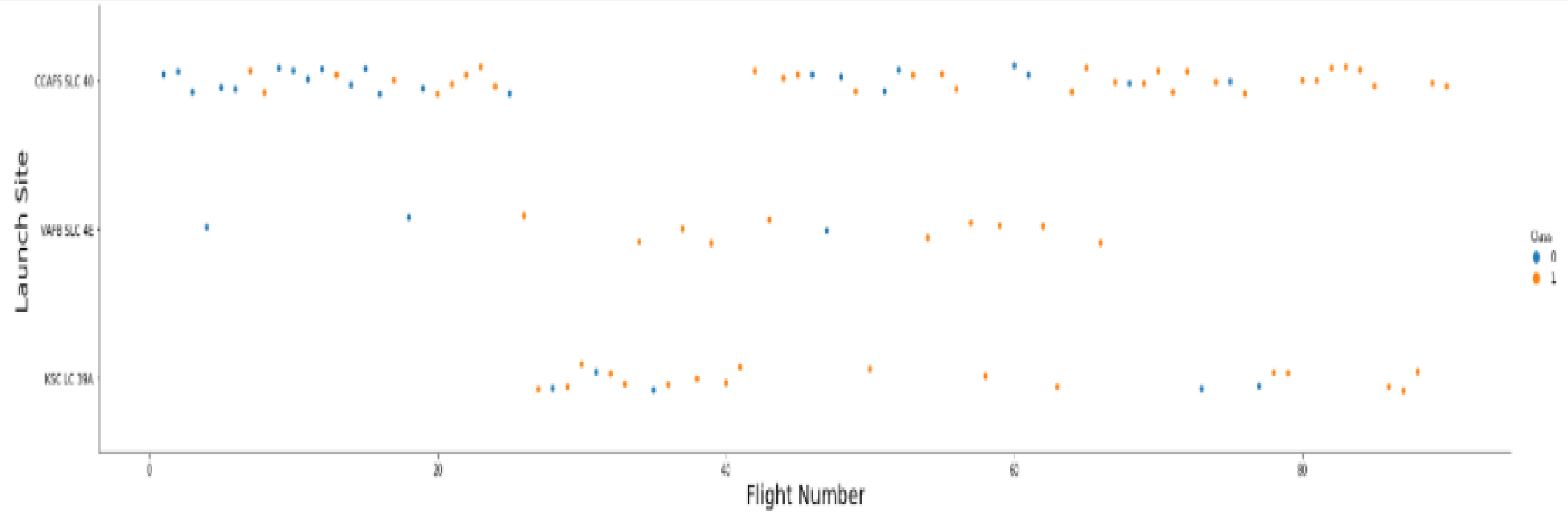
- Exploratory data analysis results
- Launch Sites Proximities Map
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

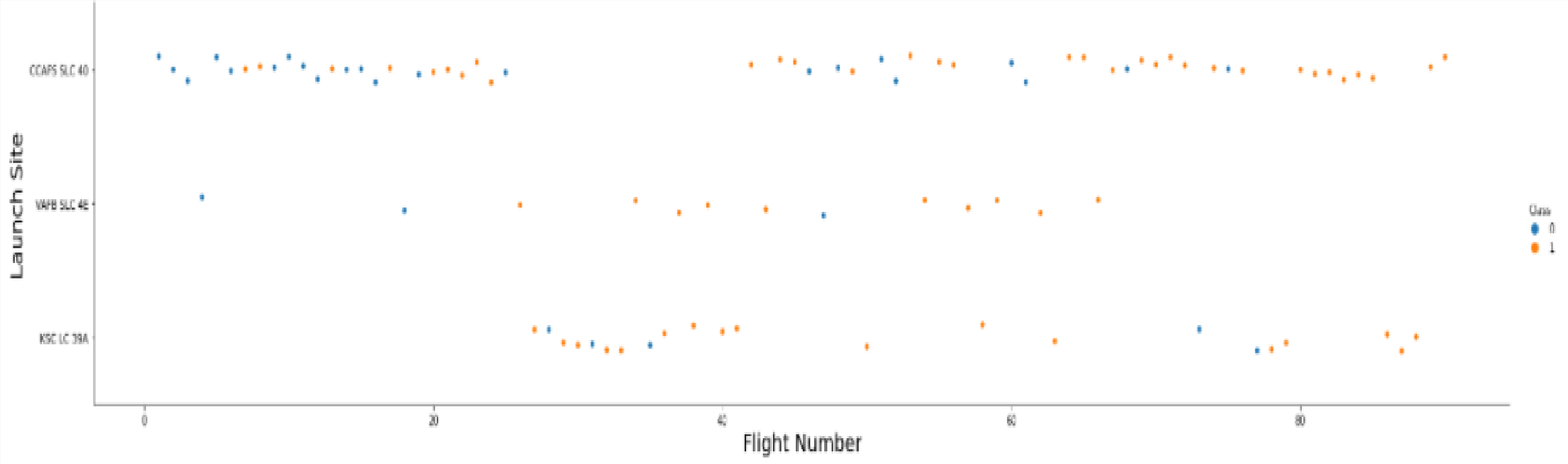
Flight Number vs. Launch Site

-The plot revealed that launch sites with higher flight numbers tend to have greater success rates.



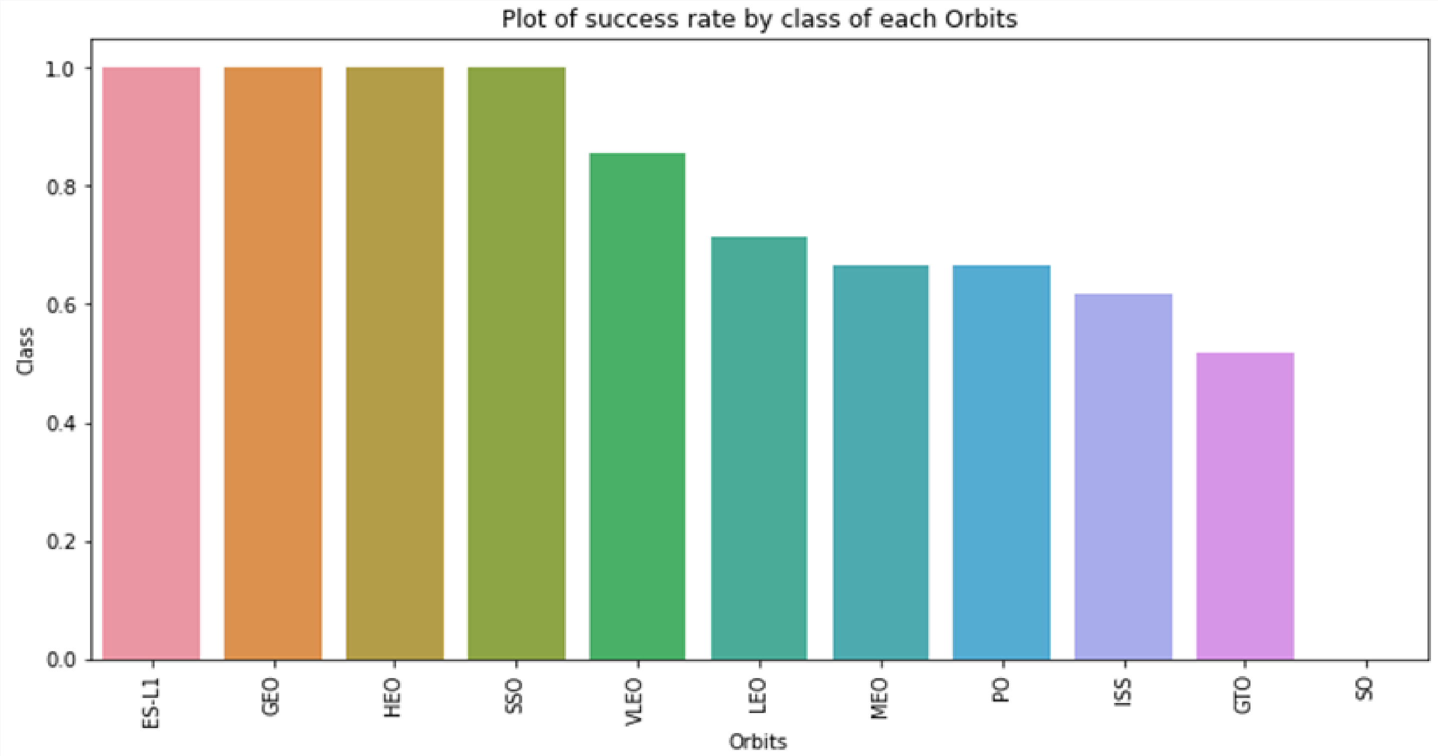
Payload vs. Launch Site

-Heavier payloads at CCAFS SLC 40 tend to result in more successful launches.



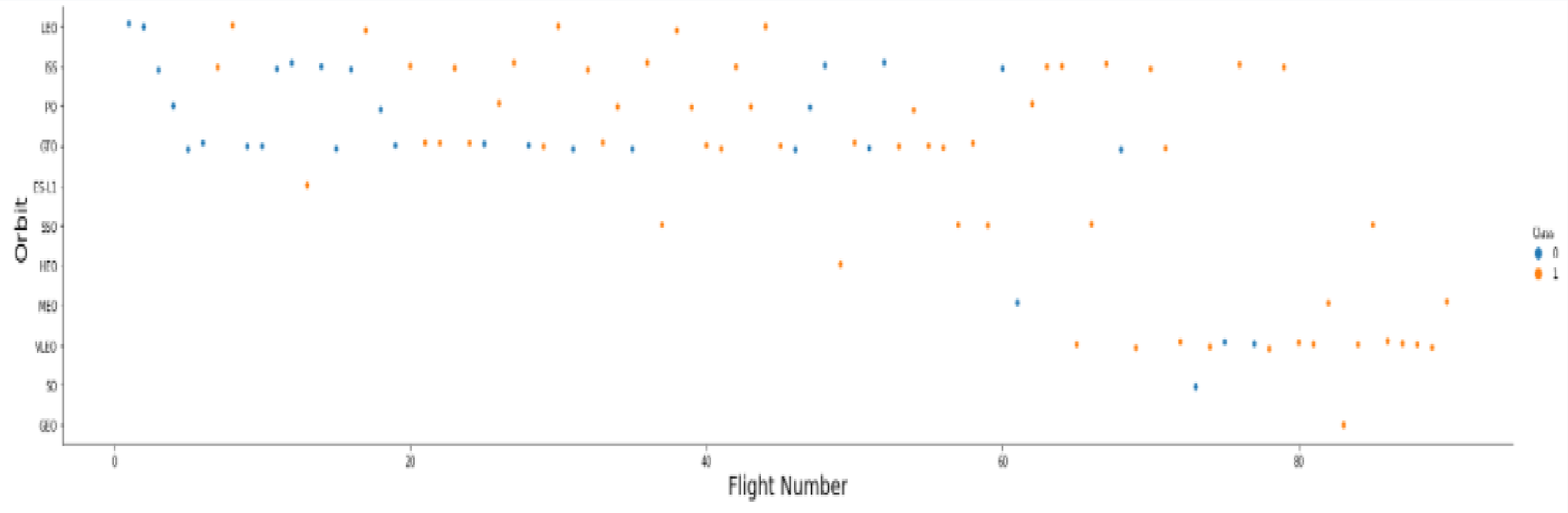
Success Rate vs. Orbit Type

-According to the plot, ES-L1, GEO, HEO, SSO, and VLEO demonstrated the greatest success rates.



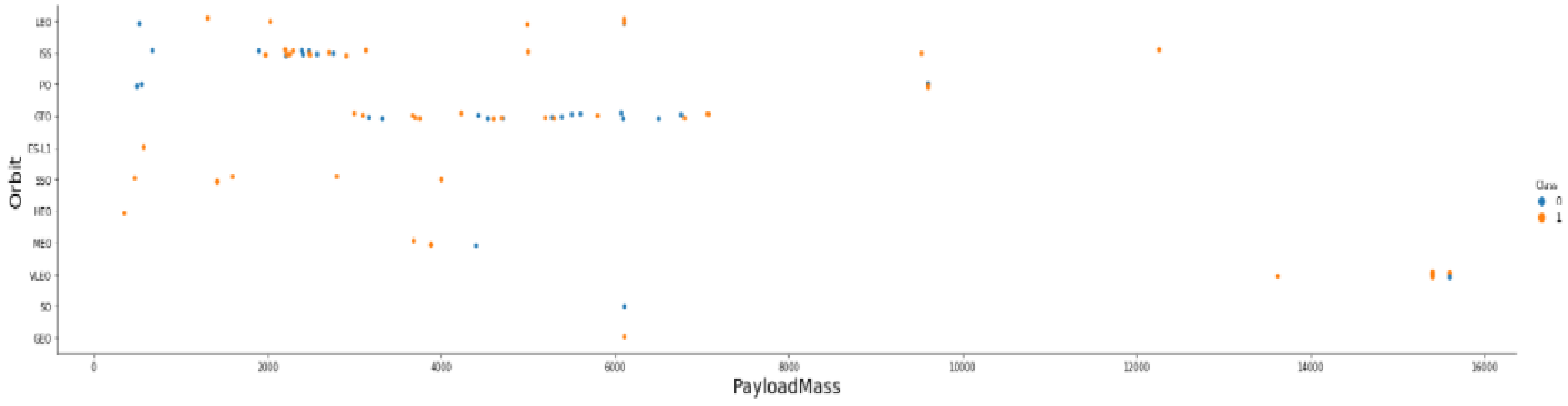
Flight Number vs. Orbit Type

-A clear cluster of failures (class=0) appears in lower flight numbers, reflecting SpaceX's early years between 2010 and 2013 (see Launch Success Yearly Trend).
- As flight numbers grew, missions to Low Earth Orbit (LEO) demonstrated increasingly consistent success.



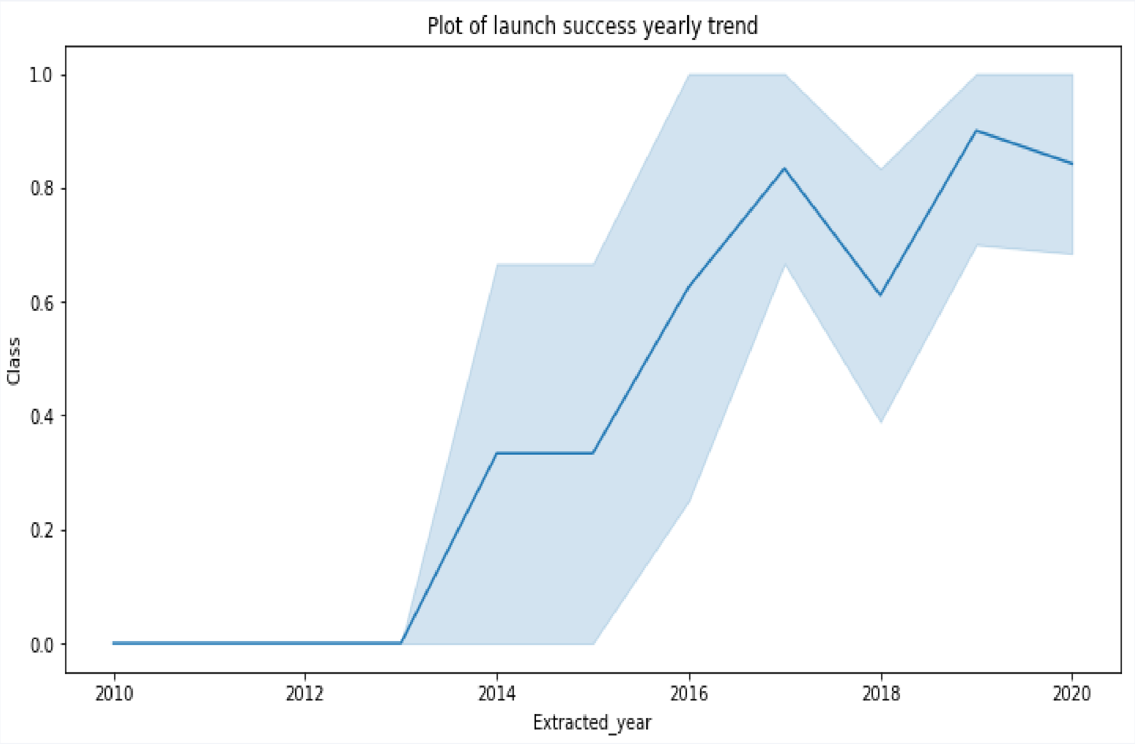
Payload vs. Orbit Type

-Heavy payloads (exceeding 8,000 kg) had the highest landing success rates when using the ISS, PO, and VLEO orbits. No other orbits supported payloads above this weight.
-In contrast, all five successful landings in the SSO orbit involved lighter payloads, each roughly 4,000 kg or less.



Launch Success Yearly Trend

-The rate of successful landings has generally increased after 2013, reaching a maximum value of 90% in 2019.



All Launch Site Names

Cape Canaveral Launch Complex, Vandenberg Space Force Base, Kennedy Space Center, Cape Canaveral Space Launch Complex

Out[10]:

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

Launch Site Names Begin with 'CCA'

-First five launches from Cape Canaveral. All landings failed or were not attempted.

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

-The total mass of payloads carried by boosters from NASA was 45,596 Kg.

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

-The average payload mass carried by booster version F9 v1.1 was 2928.4 Kg.

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

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

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

-Boosters Successfully Landed on Drone Ship (Payload: 4,000–6,000 kg)

Out[15]:

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

Total Number of Successful and Failure Mission Outcomes

-Table showing mission outcomes categorized by success and failure.

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

Out[16]:

	failureoutcome
0	1

Boosters Carried Maximum Payload

-List of the names of boosters which have carried the maximum payload mass (15,600 Kg)

Out[17]:

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

2015 Launch Records

-List of launches with failed landings in drone ship during 2015, including their booster versions and launch site names.

Out[18]:

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

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

-Ranking of the count of landing outcomes between June 4, 2010 and March 3, 2017, in descending order.

Out[19]:

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

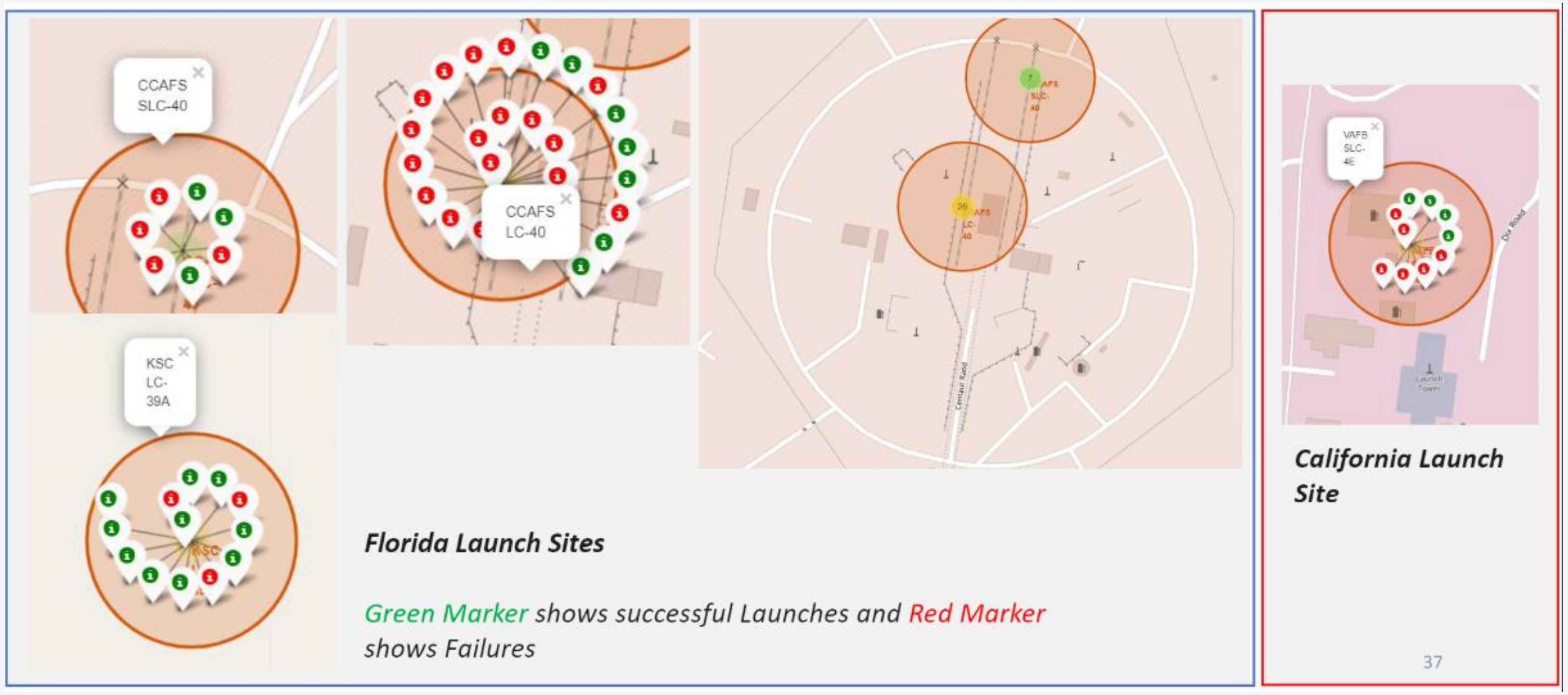
Section 3

Launch Sites Proximities Analysis

LAUNCH SITE MAP

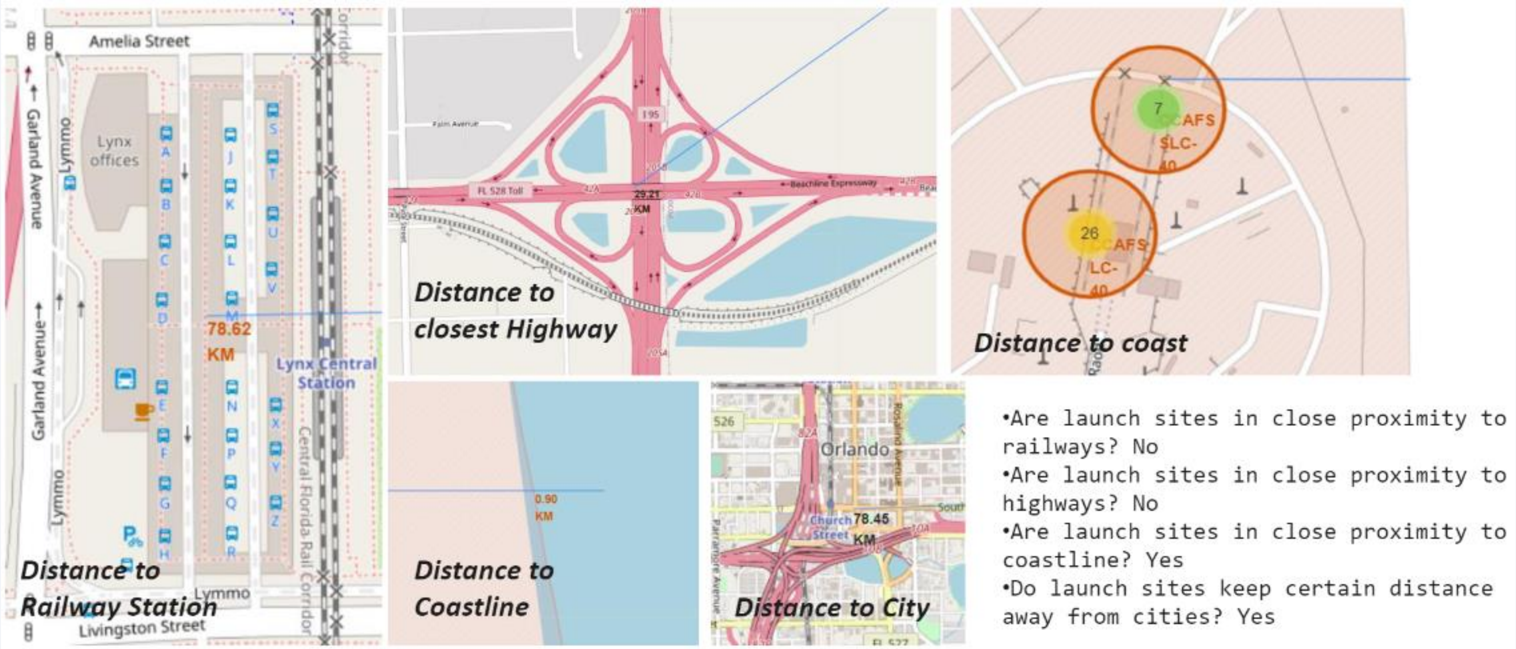
-All launch sites are located near coastal areas.

Color Label Launch Outcomes



Launch Site Proximity to Transportation

Map showing launch site distance from transport hubs.

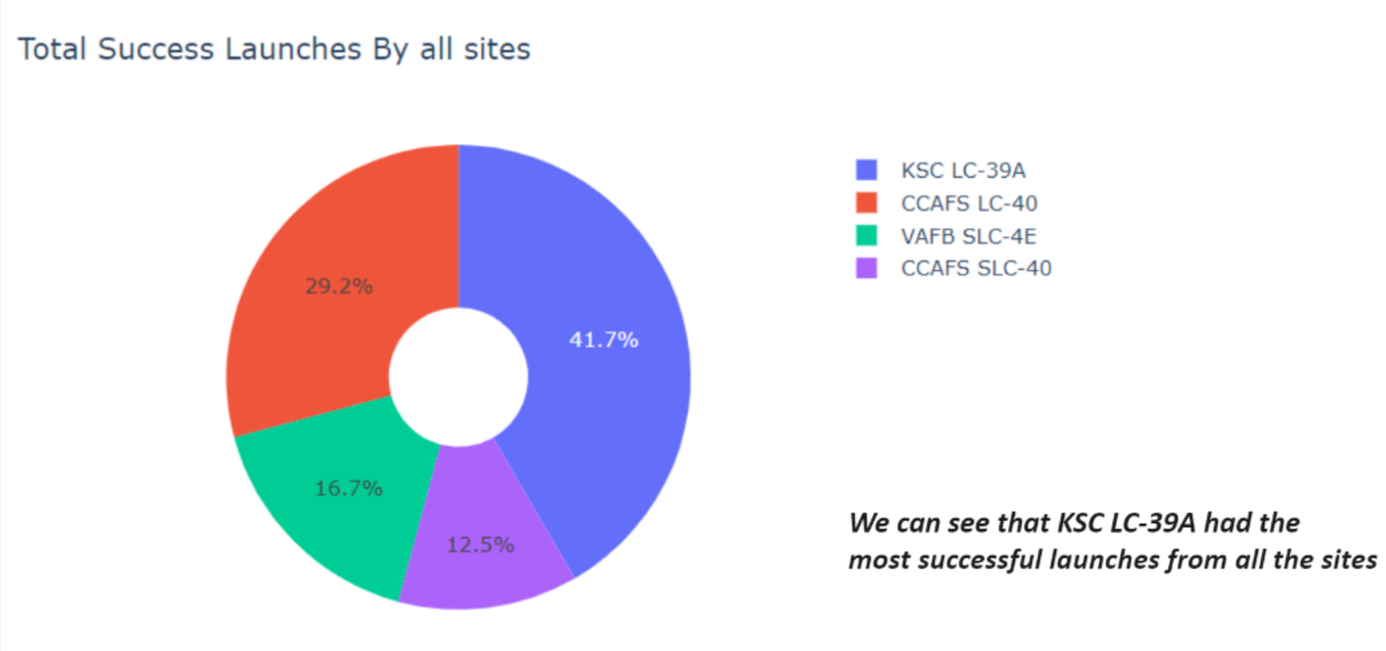


Section 4

Build a Dashboard with Plotly Dash

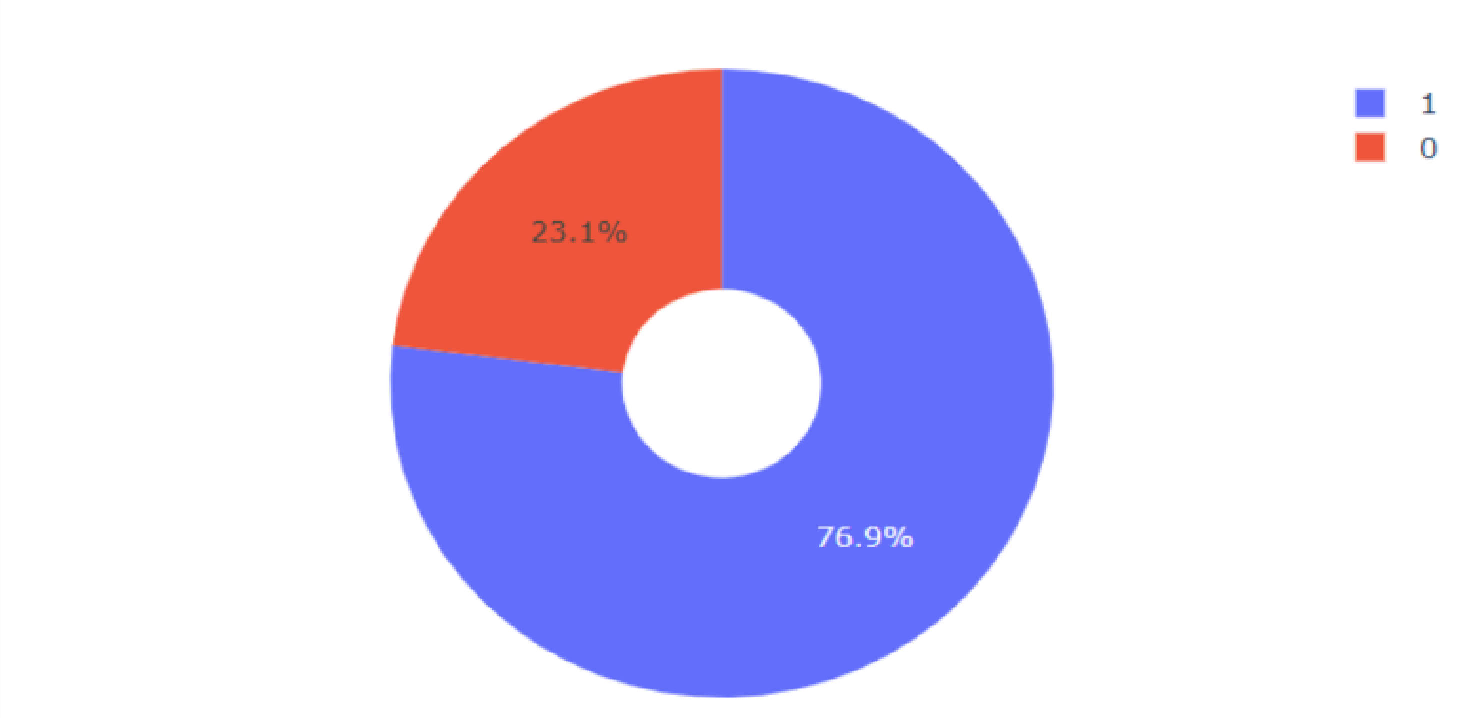
Launch Success Rate

-The Kennedy Space Center (KSC) achieved the highest success rate for landings, accounting for 41.7% of all successful landings. Interestingly, when combining the success rates of Cape Canaveral's two launch sites—the older CCAFS LC-40 and the newer CCAFS SLC-40—its overall share matches that of KSC.



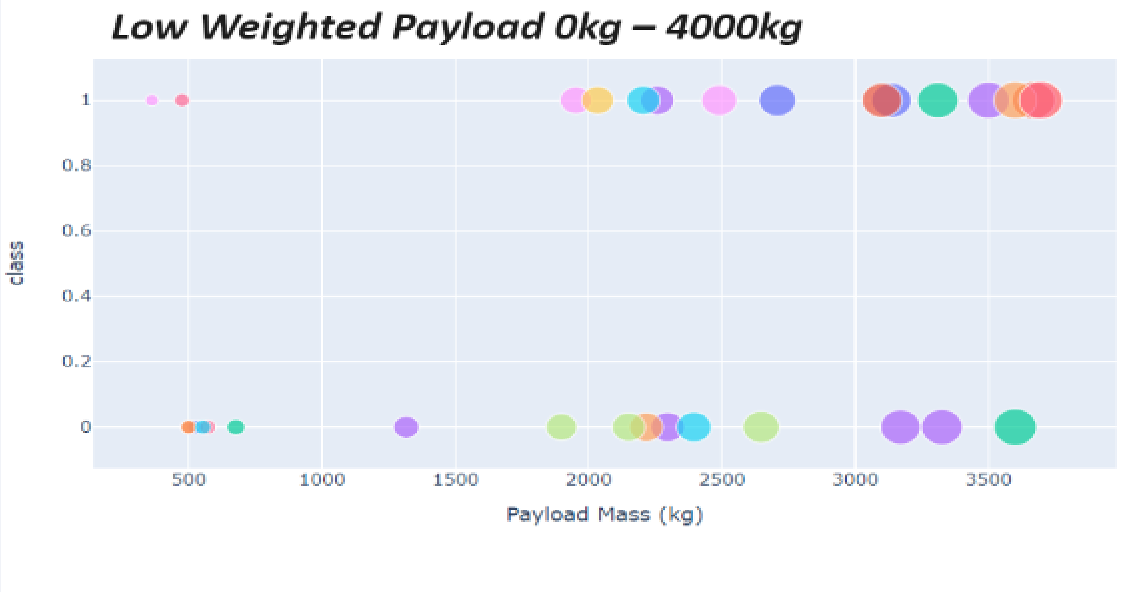
Launch Site with the Highest Success Rate for Rocket Landings.

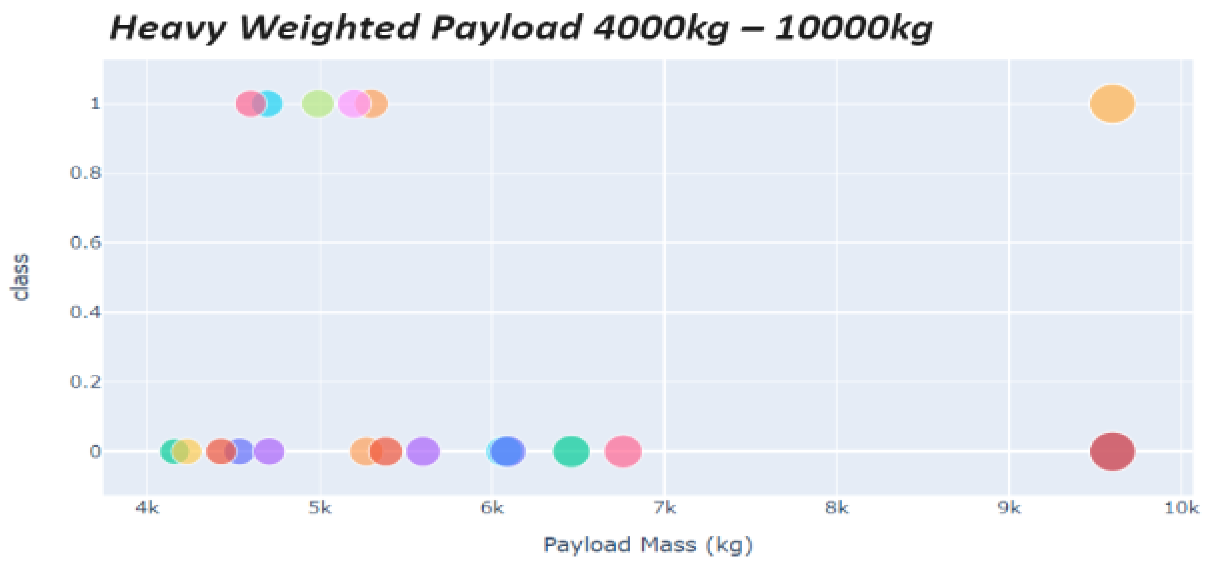
-The Kennedy Space Center leads in both overall success rate and landing success, with 76.9% of its launches landing successfully.



Payload Mass vs Launch Outcome

-The scatter plots on the left reveal that the FT booster had a higher success rate than other boosters, particularly for payloads between 2,000 and 4,000 Kg—a range with the highest number of successful landings.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

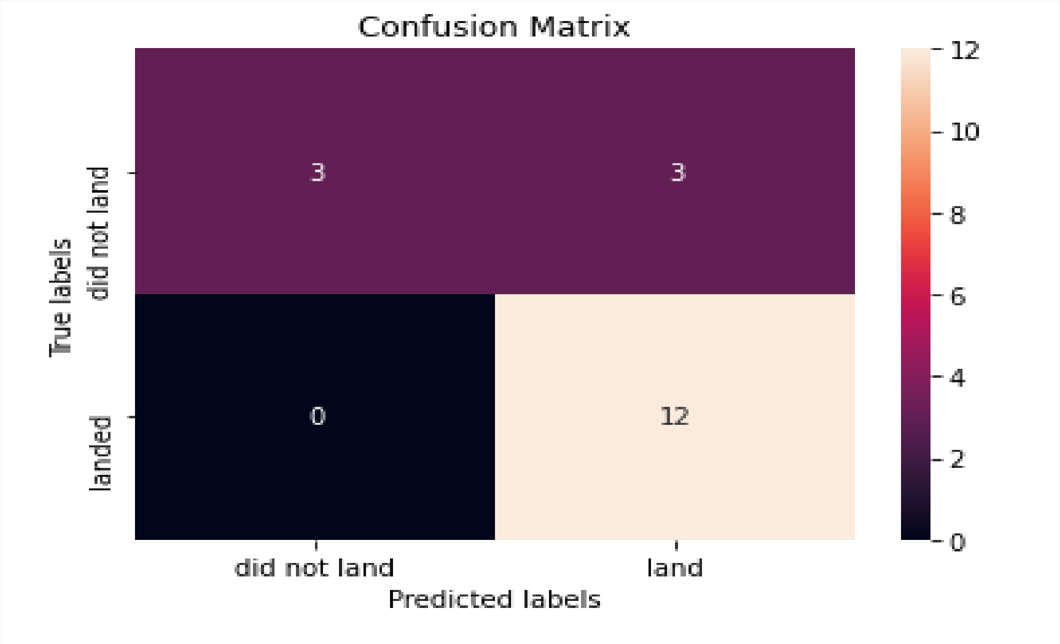
-The decision tree classifier achieves the highest classification accuracy among all models.

```
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

-When evaluated on the test data, all four models generated identical confusion matrices, each showing 3 false positives and 0 false negatives. This resulted in a consistent accuracy of 83.33% across all models.To better distinguish between the models' performance, a larger test dataset may be required.



Conclusions

SpaceX's Progress in Reusable Rocket Technology
SpaceX has made substantial strides toward cost-effective space travel, with its first-stage rocket landing success rate steadily increasing since the program's inception. Since 2019, the landing success rate has exceeded 90%, demonstrating remarkable reliability.

Payload and Orbit Performance
Successful landings have been achieved with progressively heavier payloads, though the most frequently and successfully used payload range remains 2,000–4,000 kg.

Landings have been attempted across 11 different orbits, with success on 10 of them. However, several orbits have only been tested once. Very Low Earth Orbit (VLEO) has the highest success rate, with over a dozen attempts.

Predictive Model Performance
Among the four predictive models evaluated, the Decision Tree Classifier achieved the highest built-model accuracy at 87.5%. When tested, all models performed equally, with an accuracy of 83.3%.

Future Outlook
As SpaceX continues to launch new missions, additional data could yield deeper insights and further refine predictive models in future analyses.

Appendix

Thank you!