IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sandun Chamara Ellangage
19th February 2024

Outline

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

Executive Summary

- The methodologies employed for this project encompass a comprehensive approach to handling and analyzing data. The process commenced with Data Collection through API, utilizing structured interfaces to gather relevant information. Additionally, Data Collection with Web Scraping was implemented, enabling the extraction of valuable data from web sources.
- Subsequently, Data Wrangling techniques were applied to clean, transform, and organize the gathered data, ensuring its suitability for analysis. Exploratory Data Analysis (EDA) was conducted using both SQL queries and Data Visualization tools, providing insights into patterns, trends, and relationships within the dataset.
- Furthermore, the exploration extended into Interactive Visual Analytics with Folium, offering a dynamic and engaging means to interpret geospatial data. The results of these analyses were documented, showcasing key findings and patterns.
- The project culminated in the application of Machine Learning for predictive analytics. Leveraging a Decision Tree classifier, the model demonstrated efficacy in making predictions based on the available dataset.

Introduction

- Our capstone project focuses on predicting the success of Falcon 9 first stage landings, a crucial factor in SpaceX's cost-effective rocket launches. With a launch cost of 62 million dollars compared to competitors exceeding 165 million dollars, the ability to forecast first stage landings holds strategic importance in the aerospace industry.
- Our primary goal is to determine the success of Falcon 9 first stage landings. This prediction directly influences launch costs and becomes essential information for competitors bidding against SpaceX. Accurate forecasting provides a competitive edge in the rocket launch market.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping
- Perform data wrangling
  - Implementing One-Hot Encoding for categorical fields and addressing null values and irrelevant columns ensures data preparedness for effective machine learning modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Constructed and assessed Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT) models to identify the optimal classifier.

Data Collection

- Obtained SpaceX launch data through the SpaceX REST API, providing comprehensive information on launches, rockets, payload details, launch specifications, and landing outcomes.
- Utilized BeautifulSoup for web scraping Wikipedia, serving as an alternative data source to gather Falcon 9 launch data, encompassing details on rockets, payloads, launch specifications, and landing outcomes.

Data Collection – SpaceX API

Make GET Request to SpaceX API:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/da
```

Get JSON Content:

```
from pandas import json_normalize

if response.status_code == 200:
    json_content = response.json()
```

Normalize JSON into DataFrame:

```
df = json_normalize(json_content)
```

Filter the dataframe to only include Falcon 9 launches

```
data_falcon9 = launch_df[launch_df['BoosterVersion'] == 'Falcon 9']
print(data_falcon9)
```

Save DataFrame to CSV:

```
data_falcon9.to_csv('data_falcon9.csv', index=False)
```

- GitHub-Data Collection-Space X API

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

```
response = requests.get(static_url)
```

Extract all column/variable names from the HTML table header

```
html_tables = soup.find_all('table')
```

Create a data frame by parsing the launch HTML tables

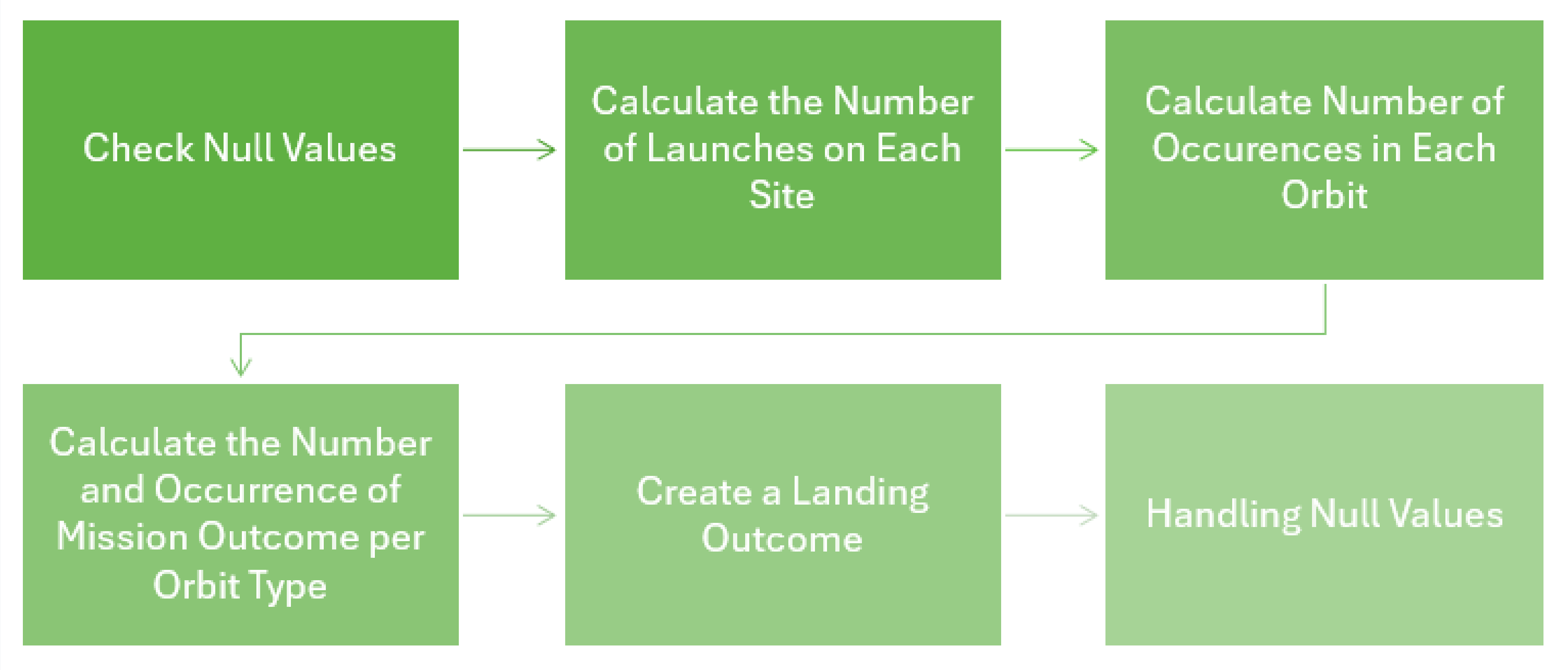
```
launch_dict = {
    'Flight No.': [],
    'Launch site': [],
    'Payload': [],
    'Payload mass': [],
    'Orbit': [],
    'Customer': [],
    'Launch outcome': [],
    'Version Booster': [],
    'Booster landing': [],
    'Date': [],
    'Time': []
}
```

Save DataFrame to CSV:

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

- GitHub-Web_Scraping-Space X API

Data Wrangling



- GitHub-Data_Wrangling

EDA with Data Visualization

- The visualization of SpaceX data has been successfully accomplished through the adept use of various charts. Scatter plots have been instrumental in uncovering correlations. Bar charts have visually depicted the distribution of successful launches across different launch sites or Orbots, facilitating easy comparison. This diverse set of charts has allowed us to gain a comprehensive understanding of the multifaceted aspects of SpaceX's remarkable journey in space exploration of Falcon 9.

- GitHub-Data Visualization

EDA with SQL

- Displaying the unique launch sites in the space mission.
- Listing 5 records where launch sites start with the string 'CCA.'
- Showing the total payload mass carried by boosters launched by NASA (CRS).
- Displaying the average payload mass carried by the booster version F9 v1.1.
- Listing the dates when successful landing outcomes on drone ships were achieved.
- Listing the names of boosters with successful ground pad landings and payload masses between 4000 and 6000.
- Providing a count of total successful and failed mission outcomes.
- Listing the names of booster versions that carried the maximum payload mass.
- Displaying records with month names, successful landing outcomes on the ground pad, booster versions, and launch sites.

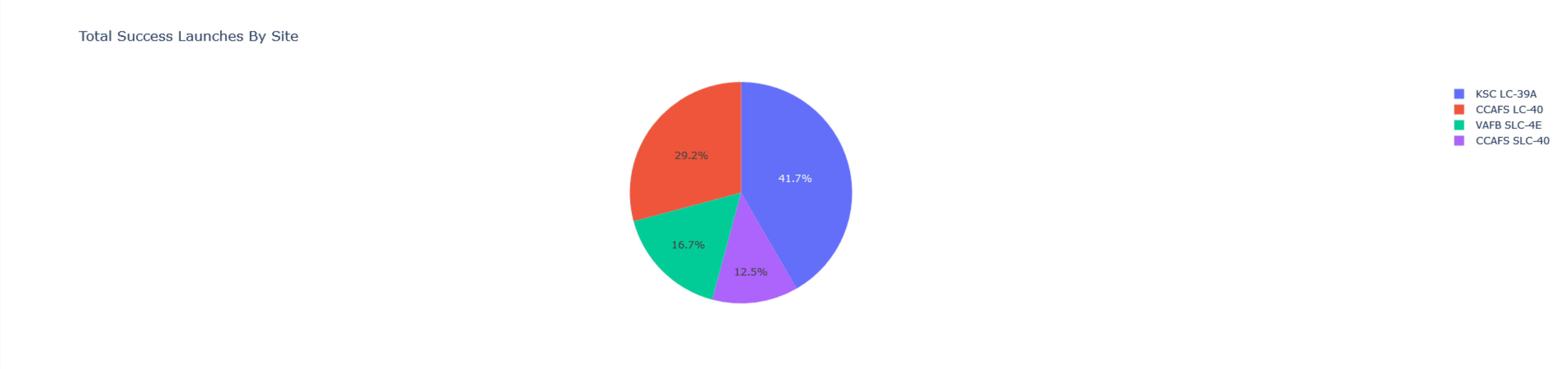
- GitHub-SQL

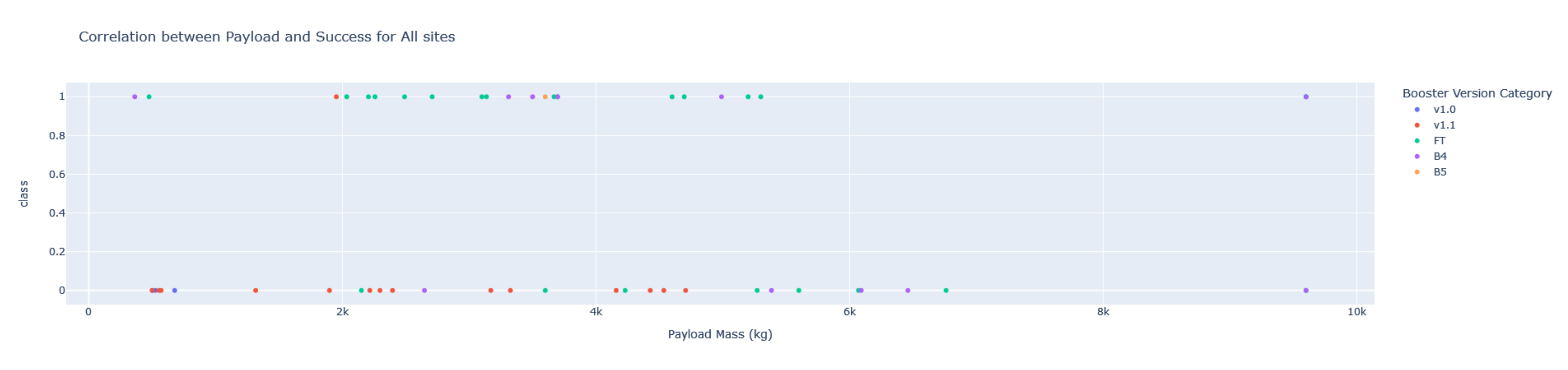
Build an Interactive Map with Folium

- Map markers, color-coded and organized into clusters, have been incorporated onto the map to facilitate the identification of an optimal location for constructing a launch site. The color-labeled markers within these clusters are designed to make it straightforward to discern launch sites that exhibit relatively high success rates.

- GitHub-Folium

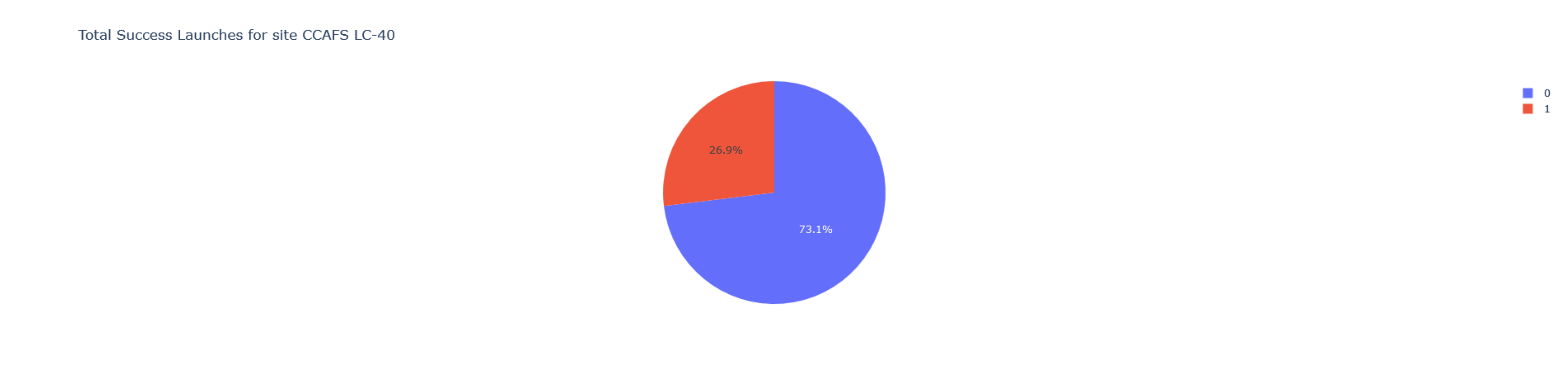
Build a Dashboard with Plotly Dash

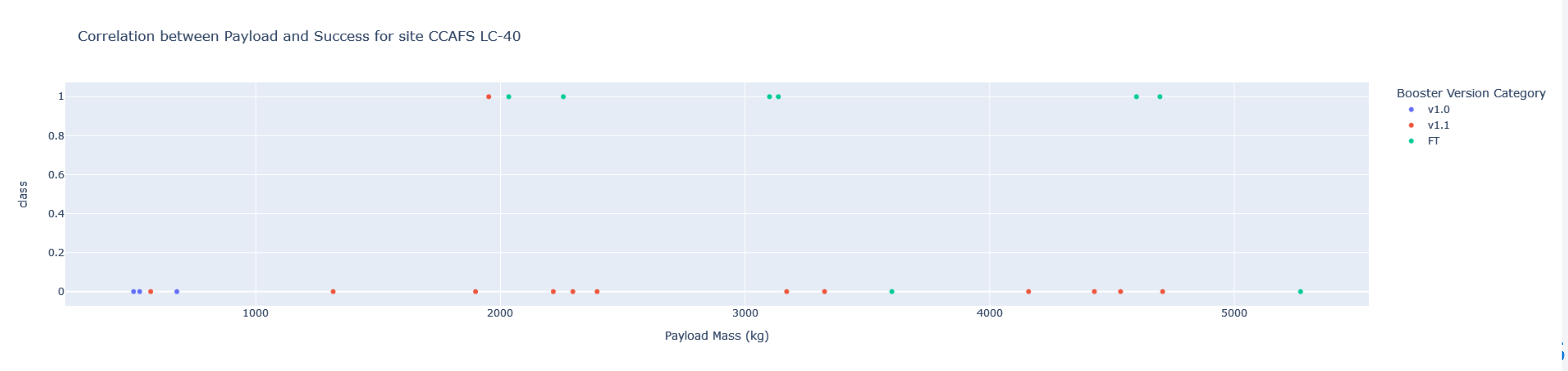




- GitHub-DashApp

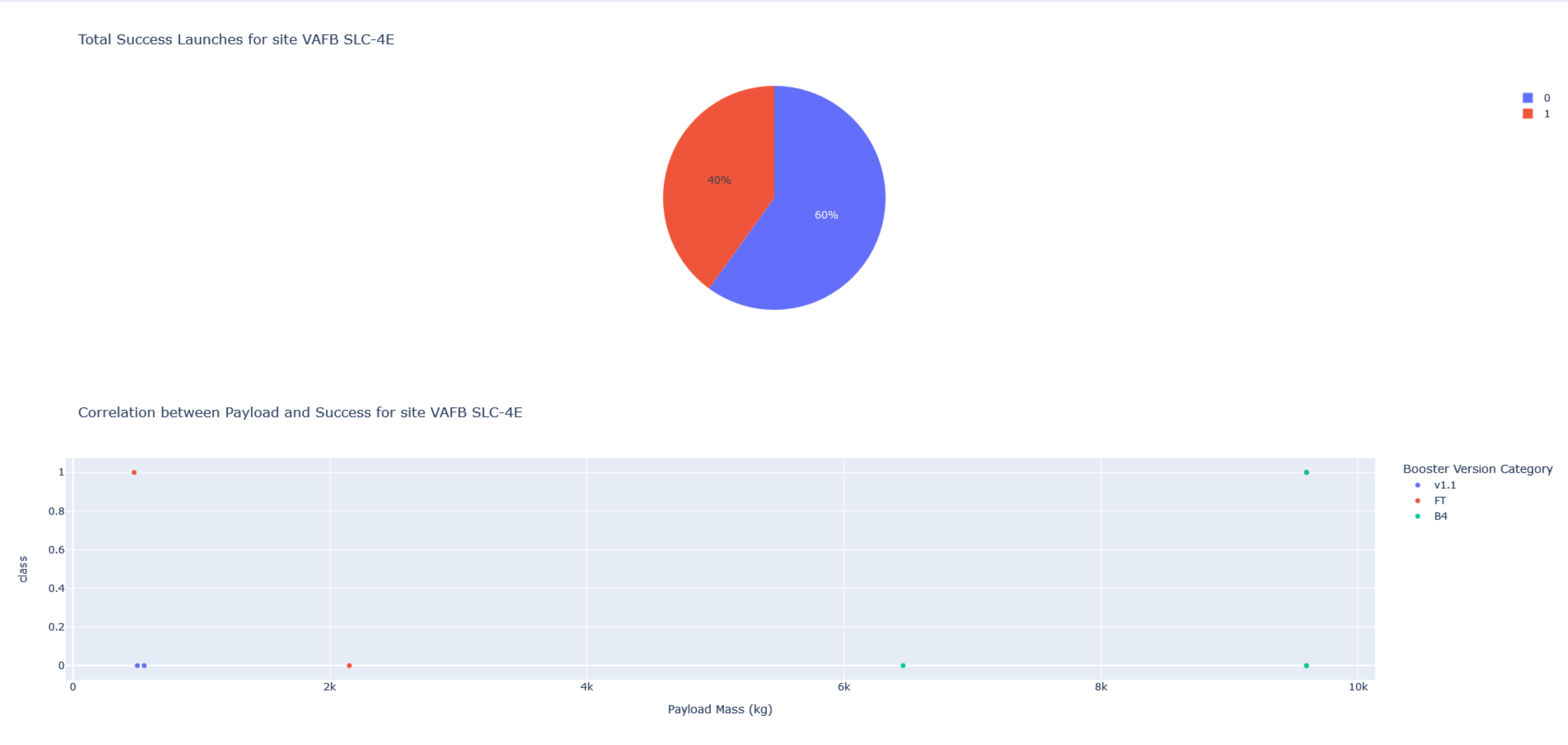
Build a Dashboard with Plotly Dash





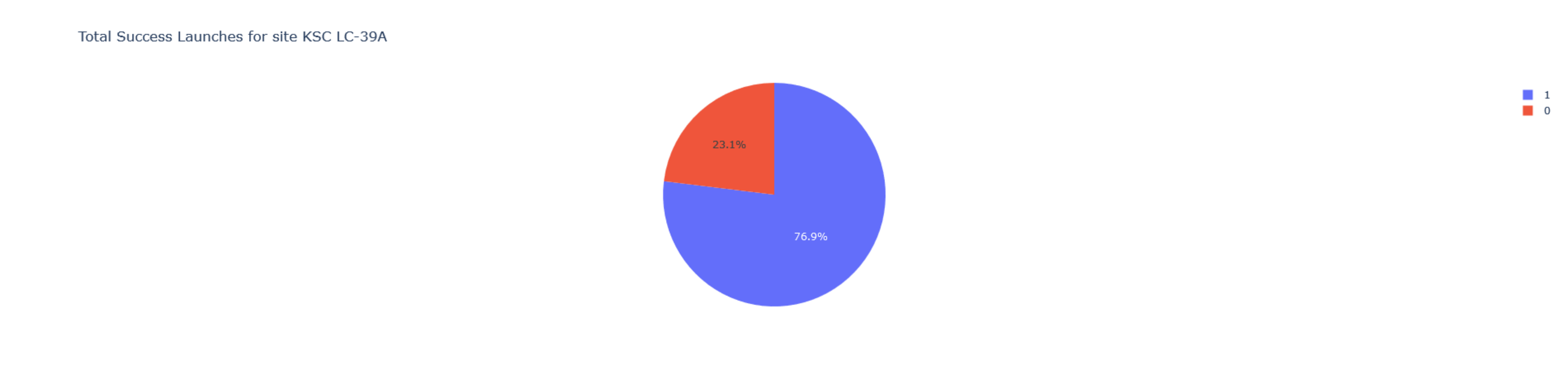
- GitHub-DashApp

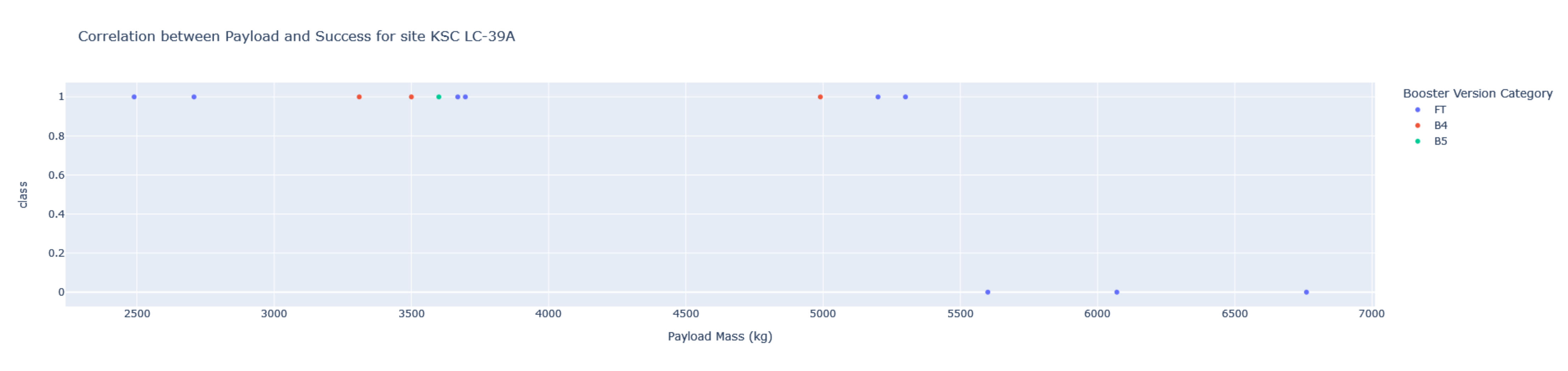
Build a Dashboard with Plotly Dash



- GitHub-DashApp

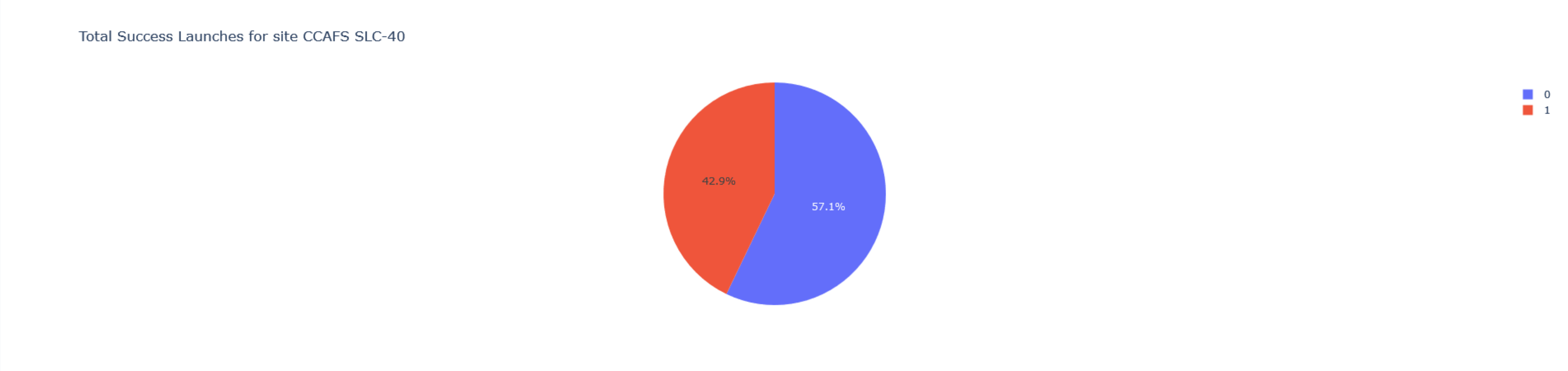
Build a Dashboard with Plotly Dash

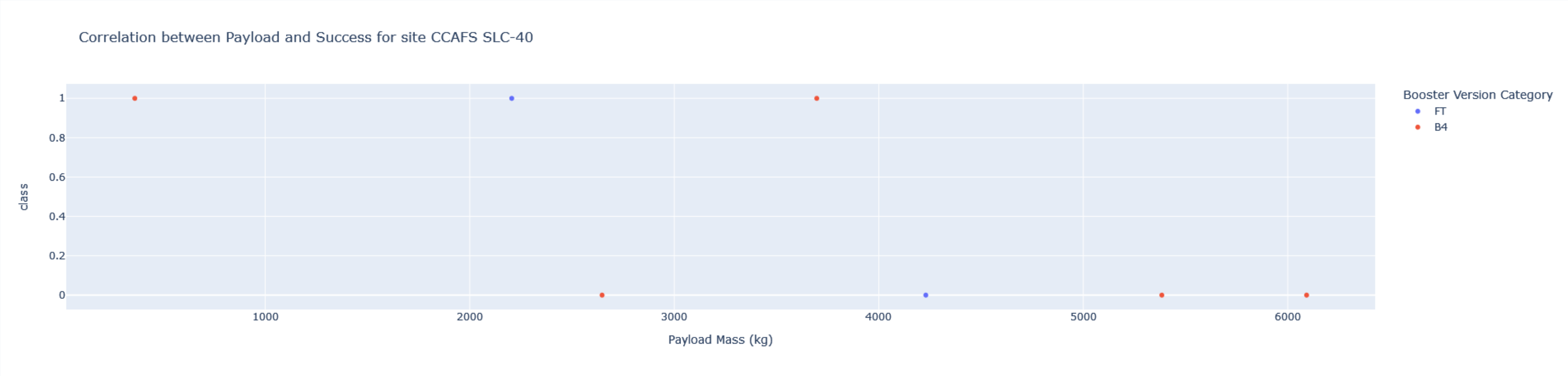




- GitHub-DashApp

Build a Dashboard with Plotly Dash

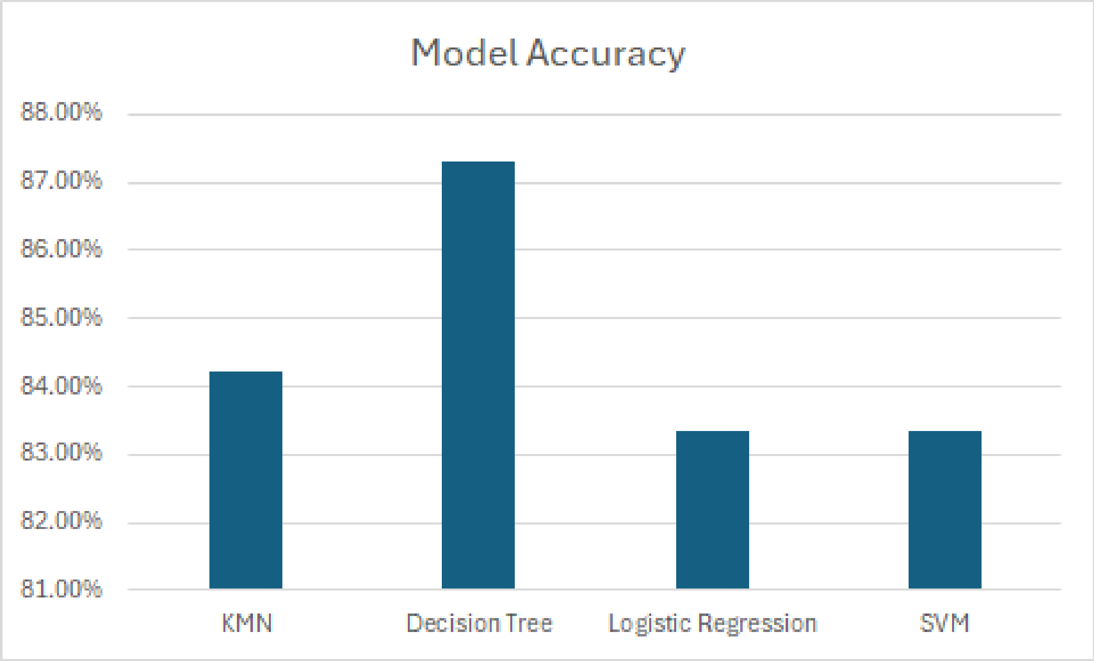


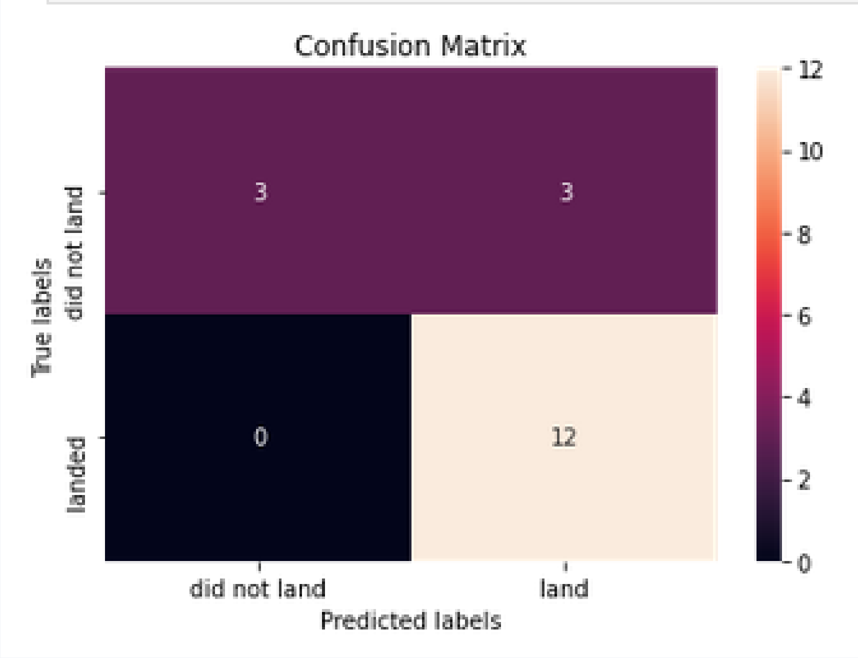


- GitHub-DashApp

Predictive Analysis (Classification)

- The SVM, KNN, Decision Tree and Logistic Regression models achieved commendable accuracy levels exceeding 80%, with Decision Tree outshining the others by demonstrating superior performance at over 87%.





- GitHub-ML

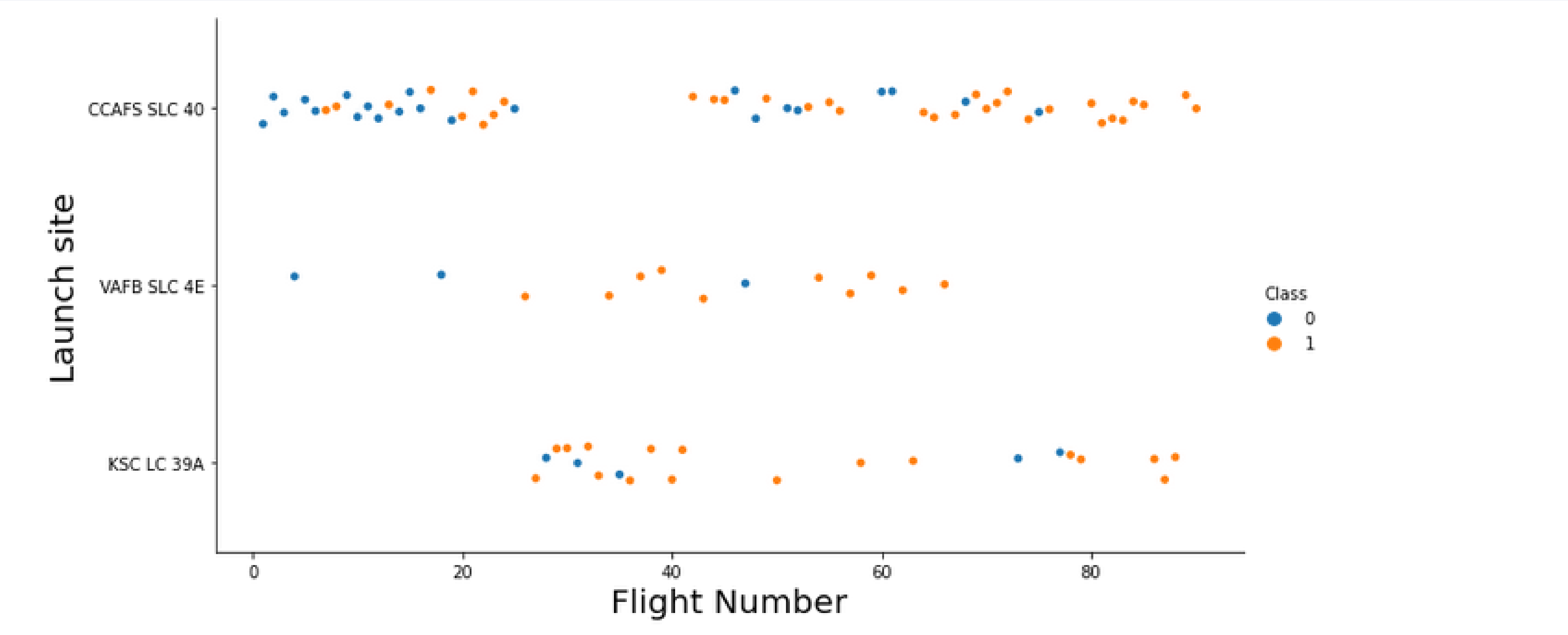
Results

- The SVM, KNN, and Logistic Regression models demonstrate superior prediction accuracy on this dataset.
- Models with lower weighted payloads outperform those with heavier payloads.
- The success rates of SpaceX launches exhibit a direct correlation with the duration of time, indicating an improvement in launch perfection over the years.
- Among all the sites, KSC LC 39A stands out with the highest number of successful launches.
- The orbit types GEO, HEO, SSO, and ES L1 show the most favorable success rates.

Section 2

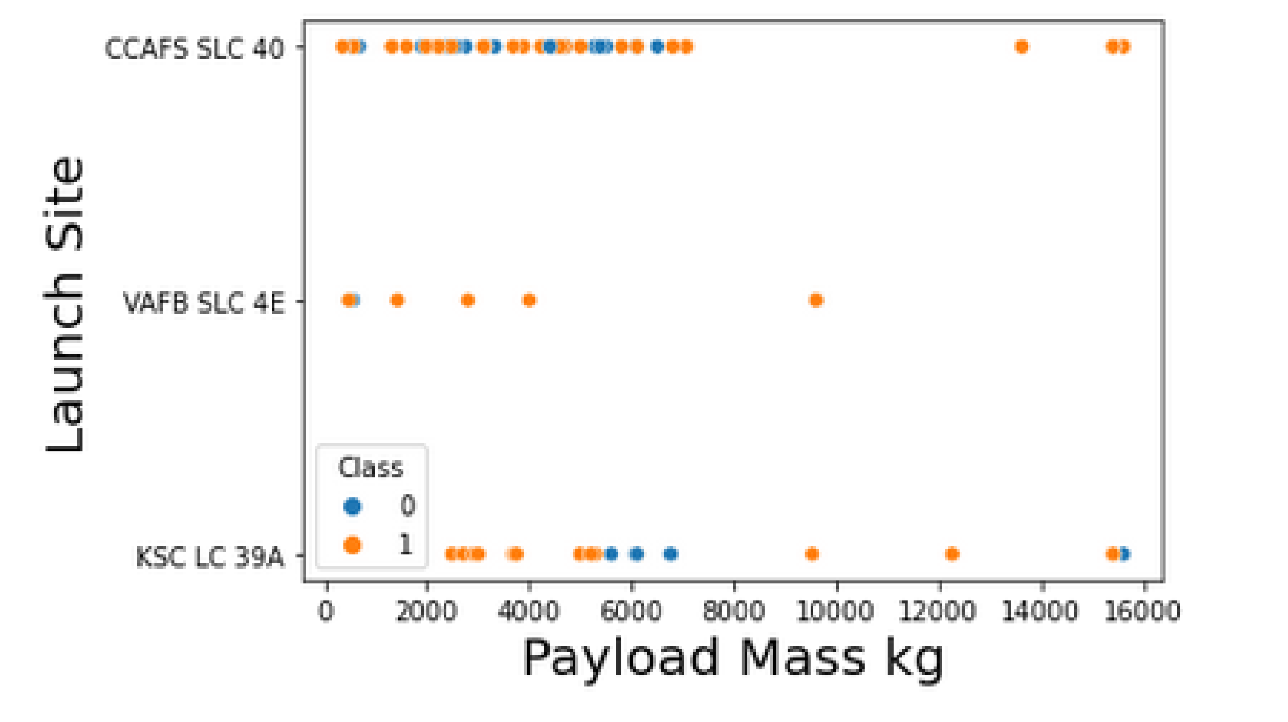
Insights drawn from EDA

Flight Number vs. Launch Site



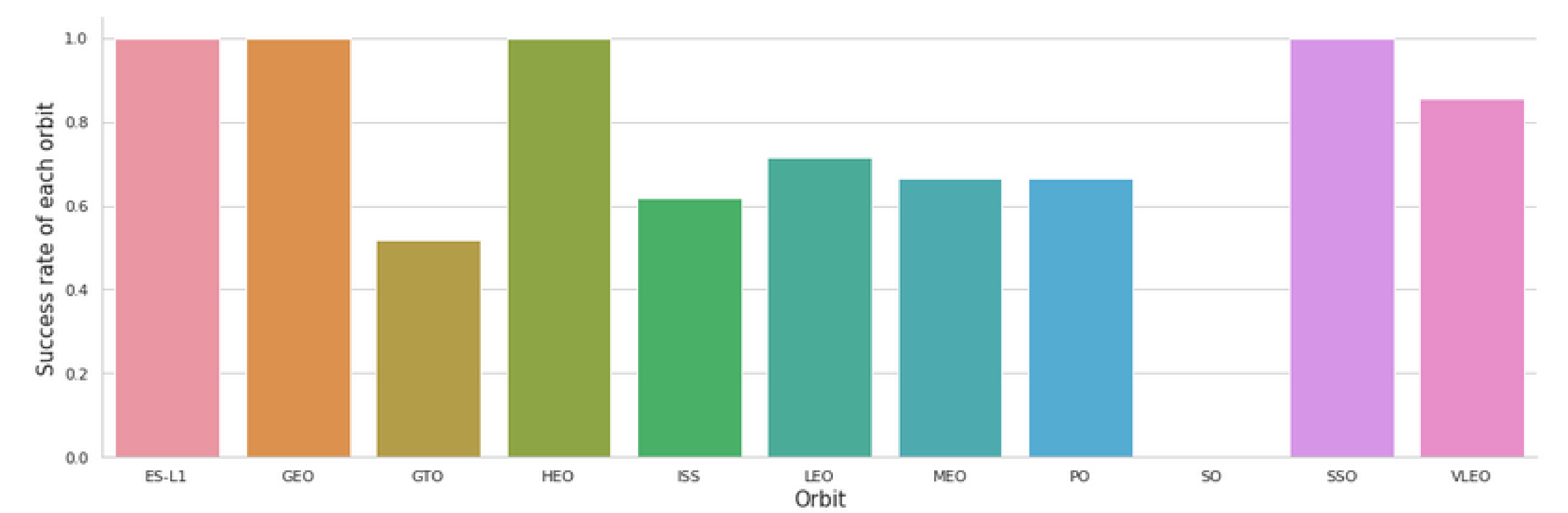
The number of launches from the CCAFS SLC 40 site significantly surpasses those from other launch sites.

Payload vs. Launch Site



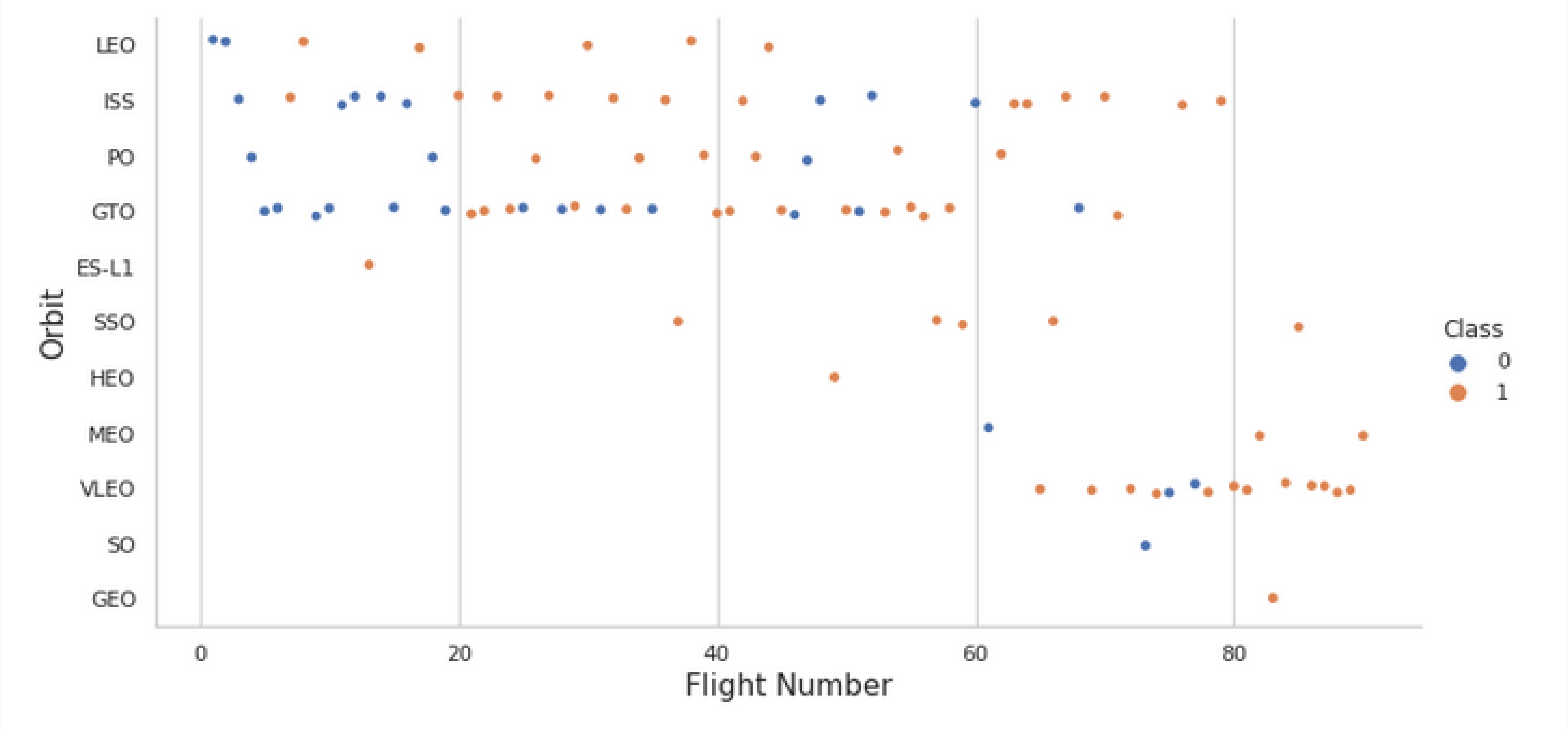
CCAFS SLC 40 has seen the launch of the majority of lower-mass payloads, whereas KSC LC 39A has been primarily employed for higher-mass payloads.

Success Rate vs. Orbit Type



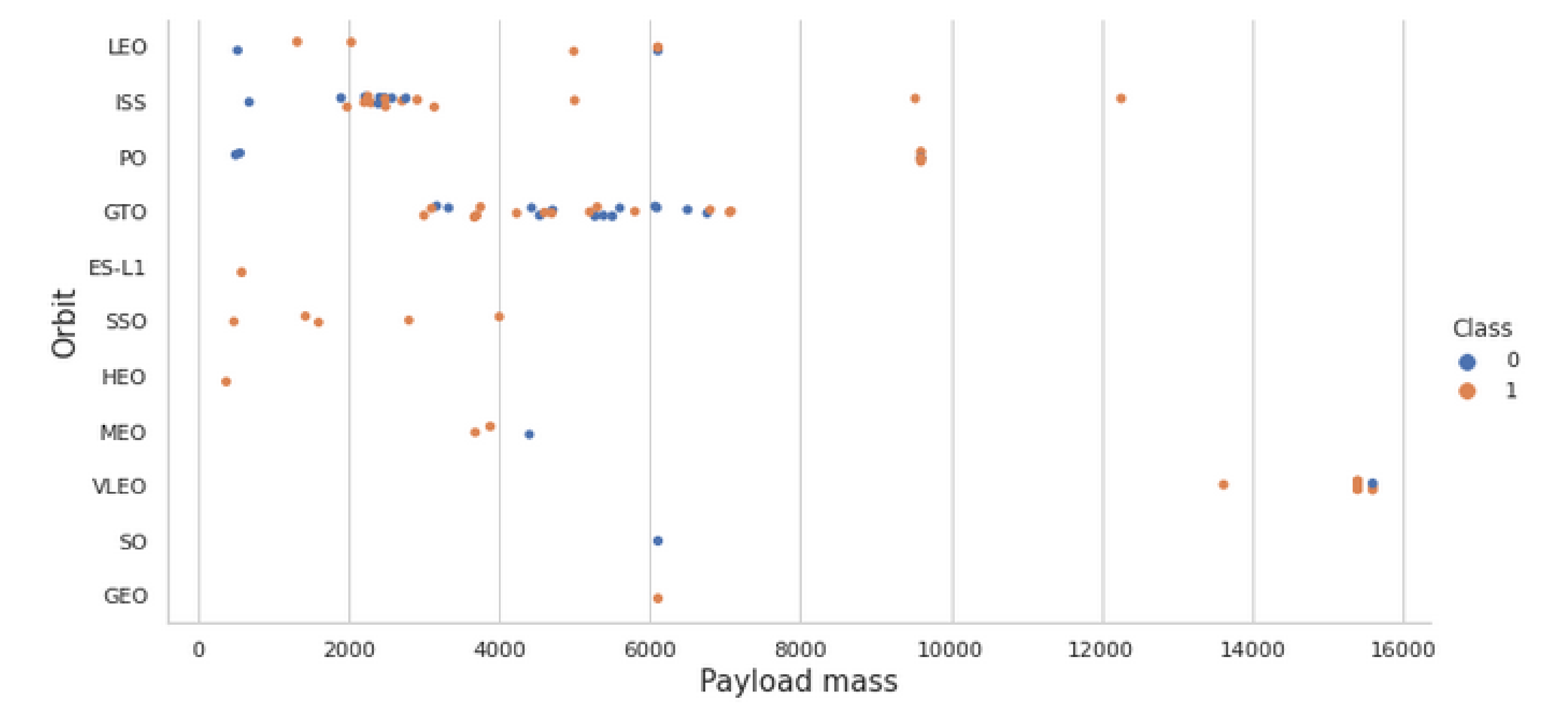
ES-L1, GEO, HEO, and SSO orbit types exhibit some of the highest success rates.

Flight Number vs. Orbit Type



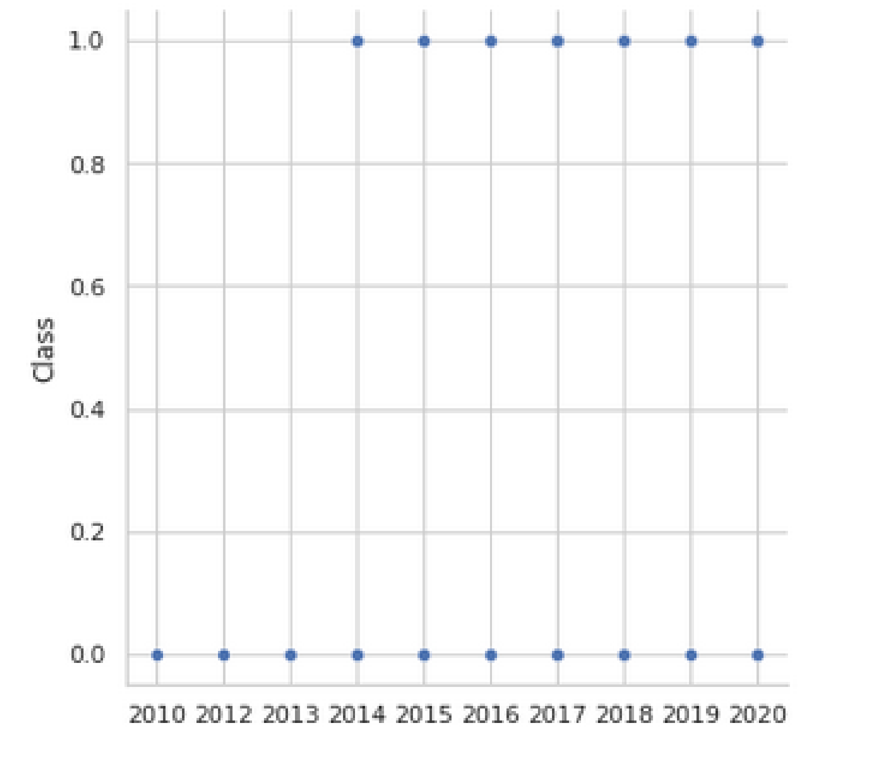
There is a noticeable trend of transitioning to Very Low Earth Orbit (VLEO) launches in recent years. In the initial stages, the preferred options were Geostationary Transfer Orbit (GTO), Polar Orbit (PO), International Space Station (ISS), and Low Earth Orbit (LEO).

Payload vs. Orbit Type



A robust correlation is evident between the International Space Station (ISS) and payloads within the range of approximately 2000, and similarly, a correlation exists between Geostationary Transfer Orbit (GTO) and payload values within the range of 4000-8000.

Launch Success Yearly Trend



The launch success rate has experienced a notable increase since 2014, likely attributed to advancements in technology and the assimilation of valuable lessons.

All Launch Site Names

```
SELECT DISTINCT LaunchSite FROM SpaceX
```

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

Launch Site Names Begin with 'CCA'

- `SELECT * FROM SpaceX WHERE LaunchSite LIKE 'CCA%'LIMIT 5`

	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

```
SELECT  SUM(PayloadMassKG) AS Total_PayloadMass FROM SpaceX WHERE Customer LIKE
'NASA (CRS)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

```
SELECT AVG(PayloadMassKG) AS Avg_PayloadMass FROM SpaceX WHERE BoosterVersion =
'F9 v1.1'
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

```
SELECT MIN(Date) AS FirstSuccessfull_landing_date FROM SpaceX WHERE LandingOutcome
LIKE 'Success (ground pad)'
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT BoosterVersion FROM SpaceX WHERE LandingOutcome = 'Success (drone ship)'
AND PayloadMassKG > 4000 AND PayloadMassKG < 6000
```

	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

```
SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
```

The total number of successful mission outcome is:

	successoutcome
0	100

```
SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
```

The total number of failed mission outcome is:

	failureoutcome
0	1

Boosters Carried Maximum Payload

```
SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
```

	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

```
SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
```

	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

```
SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
```

	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

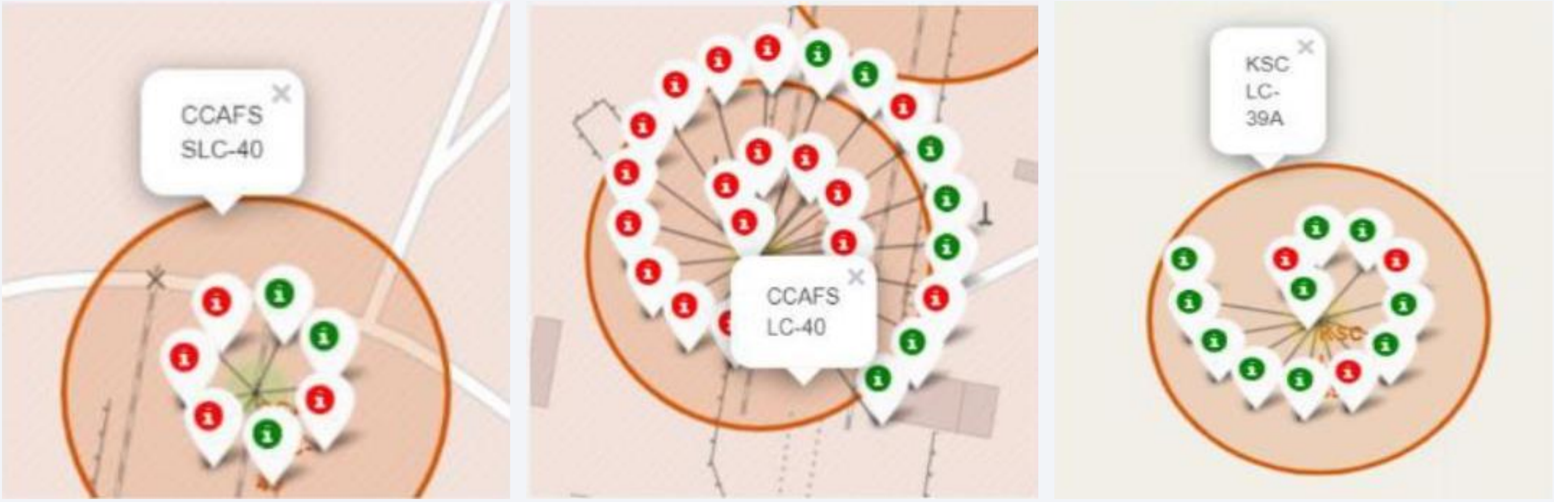
Launch Sites Proximities Analysis

Launch Sites – View

The launch sites for SpaceX strategically located in Cape Canaveral Space Force Station in Florida, Kennedy Space Center in Florida, and Vandenberg Space Force Base in California.

Launch Sites with Colour Labels

Florida Launch Sites



Green Marker – Successful Launch
Red marker – Failed Launch

Launch Sites with Colour Labels

California Launch Site

Green Marker – Successful Launch
Red marker – Failed Launch

Launch Sites Distance to Landmarks

Launch Sites are Close Proximity to railways ? No

Launch Sites are Close Proximity to highways ? No

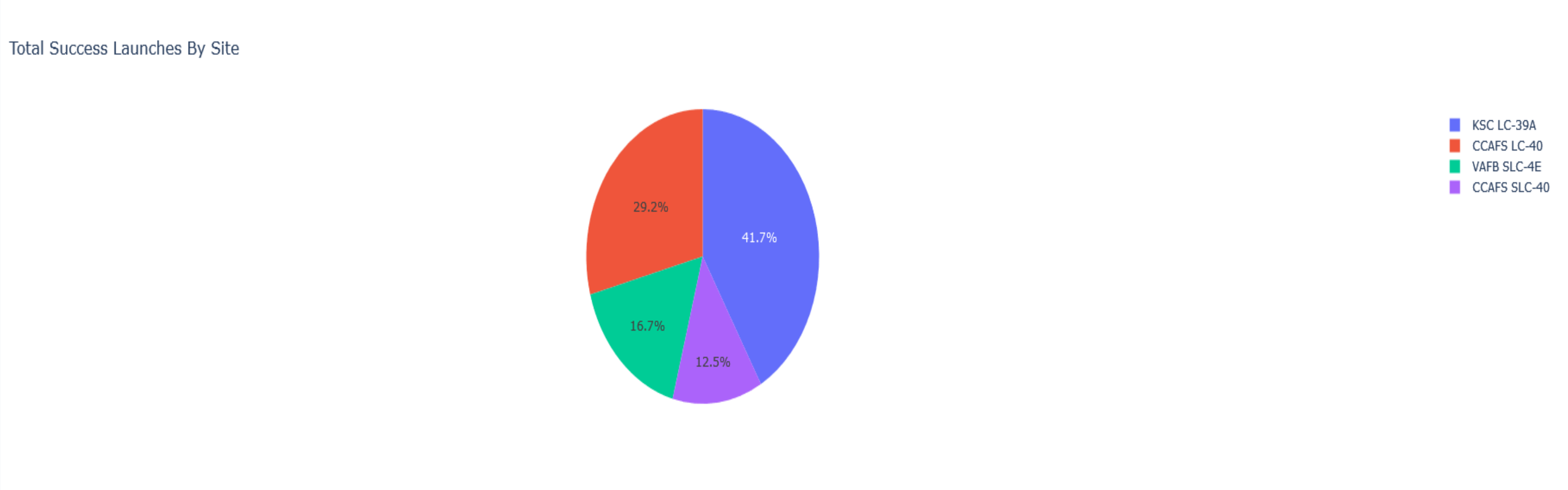
Launch Sites are Close Proximity to Coastline ? Yes

Launch Sites and appropriate Distance to Cities ? Yes

Section 4

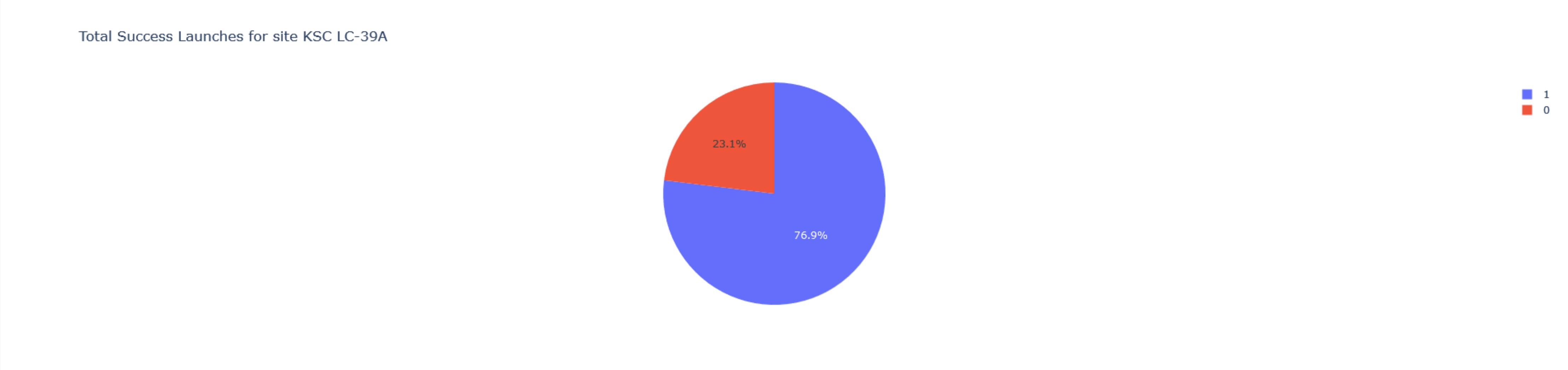
Build a Dashboard with Plotly Dash

Success % by Launch Site



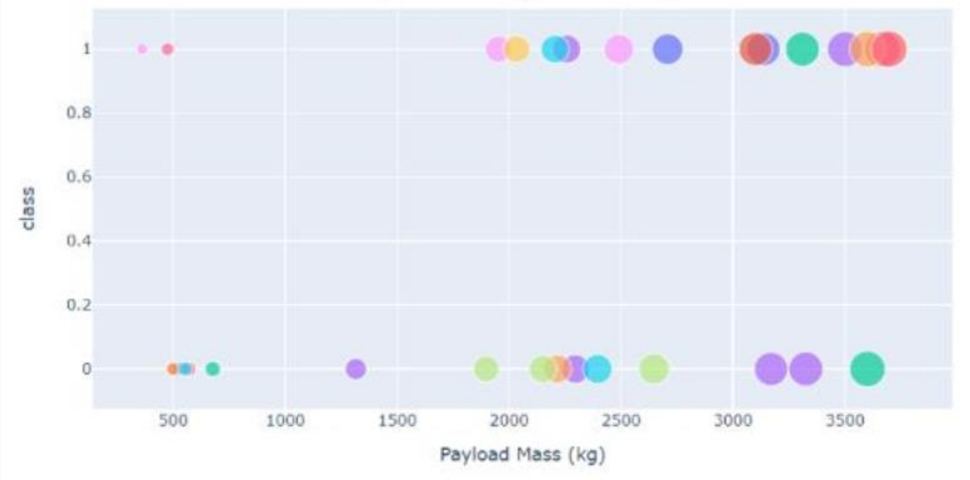
KSC LC-39A was the most successful site for launches out of all the sites

Total Success Launches for Site KSC LC-39A

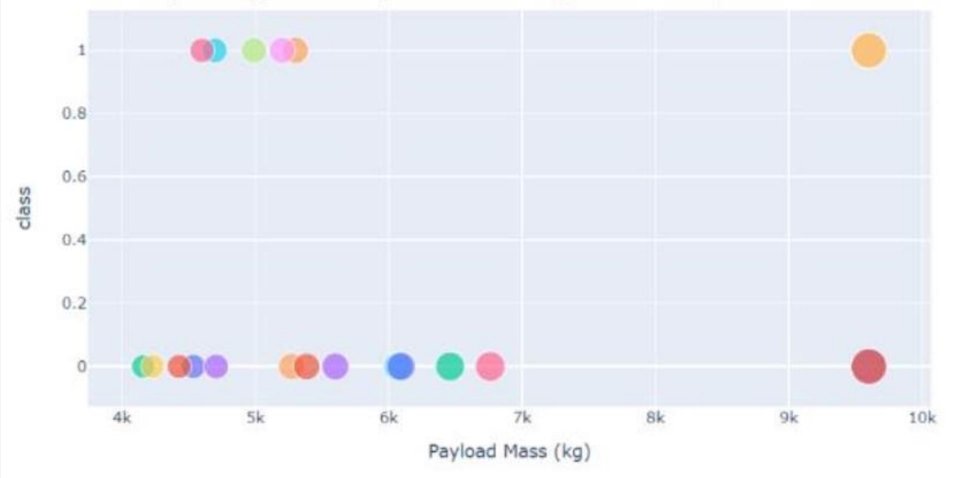


KSC LC-39A achieved a 76.9% Success Rate

Payload vs. Launch Outcome



Lower Payload



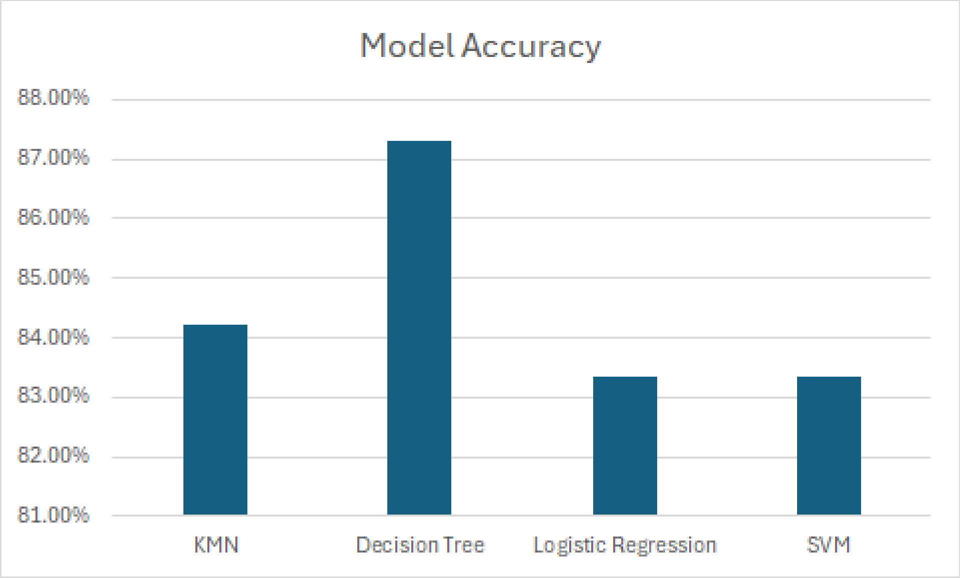
Higher Payload

It can be observed that lower payloads had more success than the higher payloads

Section 5

Predictive Analysis (Classification)

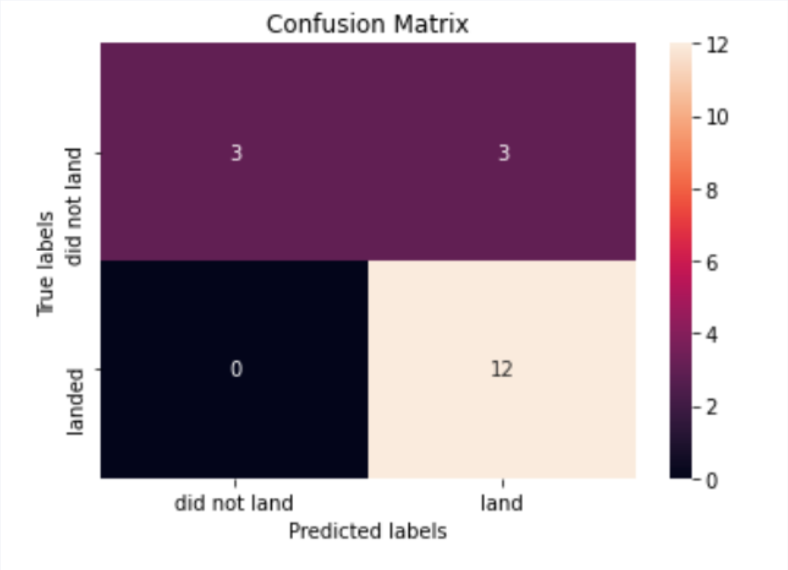
Classification Accuracy




```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
bestalgorithm = max(algorithms, key=algorithms.get)
print('The Best Algorithm from the performed algorithms is',bestalgorithm,'with a score of',algorithms[bestalgori
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
The Best Algorithm from the performed algorithms is Tree with a score of 0.8857142857142856
Best Params is : {'criterion': 'entropy', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_sam
ples_split': 5, 'splitter': 'random'}
```

Confusion Matrix



- The confusion matrix of the decision tree classifier indicates its ability to differentiate between various classes. However, a notable issue lies in the occurrence of false positives, specifically instances where the classifier incorrectly identifies unsuccessful landings as successful ones.

Conclusions

- A positive correlation is observed between the number of flights at a launch site and the corresponding success rate.
- The launch success rate exhibited an upward trend from 2013 to 2020.
- Orbits such as ES-L1, GEO, HEO, SSO, and VLEO demonstrated notably high success rates.
- Among all launch sites, KSC LC-39A recorded the highest number of successful launches.
- The Decision Tree classifier emerges as the most effective machine learning algorithm for this specific task.

Thank you!