



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Terrence F. Muir
September 19, 2022



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion



Executive Summary

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

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. SpaceX achieves significant savings through the reuse of the first stage. If we can determine the success rate of first stage landings, we can determine the cost of a launch. This information is essential for companies that wish to bid against SpaceX.

I will use data science to analyze, visualize and assess variables inclusive of launch process, launch site, payload mass of rocket, orbit to:

- determine the success rate of first stage landings
- Predict the success of first stage landings
- Assess operating conditions required to ensure a successful program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

The data was collected using SpaceX rest API and data web scraping on Wikipedia webpages.

- Perform data wrangling

This was achieved through the use of Pandas, NumPy; One Hot Encoding data fields for machine learning, data cleaning of null/irrelevant values and columns, data normalization and standardization.

- Perform exploratory data analysis (EDA) using visualization and SQL

This was achieved through the use of libraries such as seaborn and matplotlib for visualization and SQL.

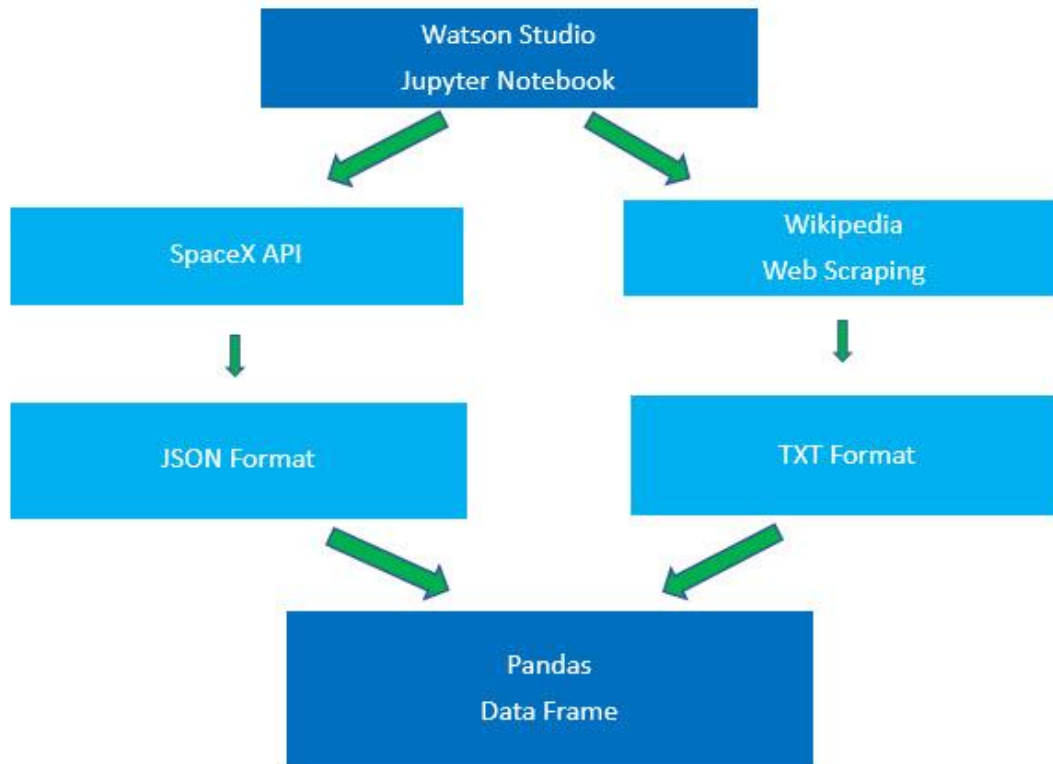
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Logistic Regression, KNN, SVM and Decision Tree classification models were built, tuned and evaluated.


Data Collection

Data was collected from:

- SpaceX API - Open Source REST API. This source provided data about launches, rockets used, core, payload delivered, capsule, launchpad, landing pad data, launch site locations, launch/landing specifications and landing outcome.
- Wikipedia - Free online encyclopedia. Provided great data via web scraping using BeautifulSoup.

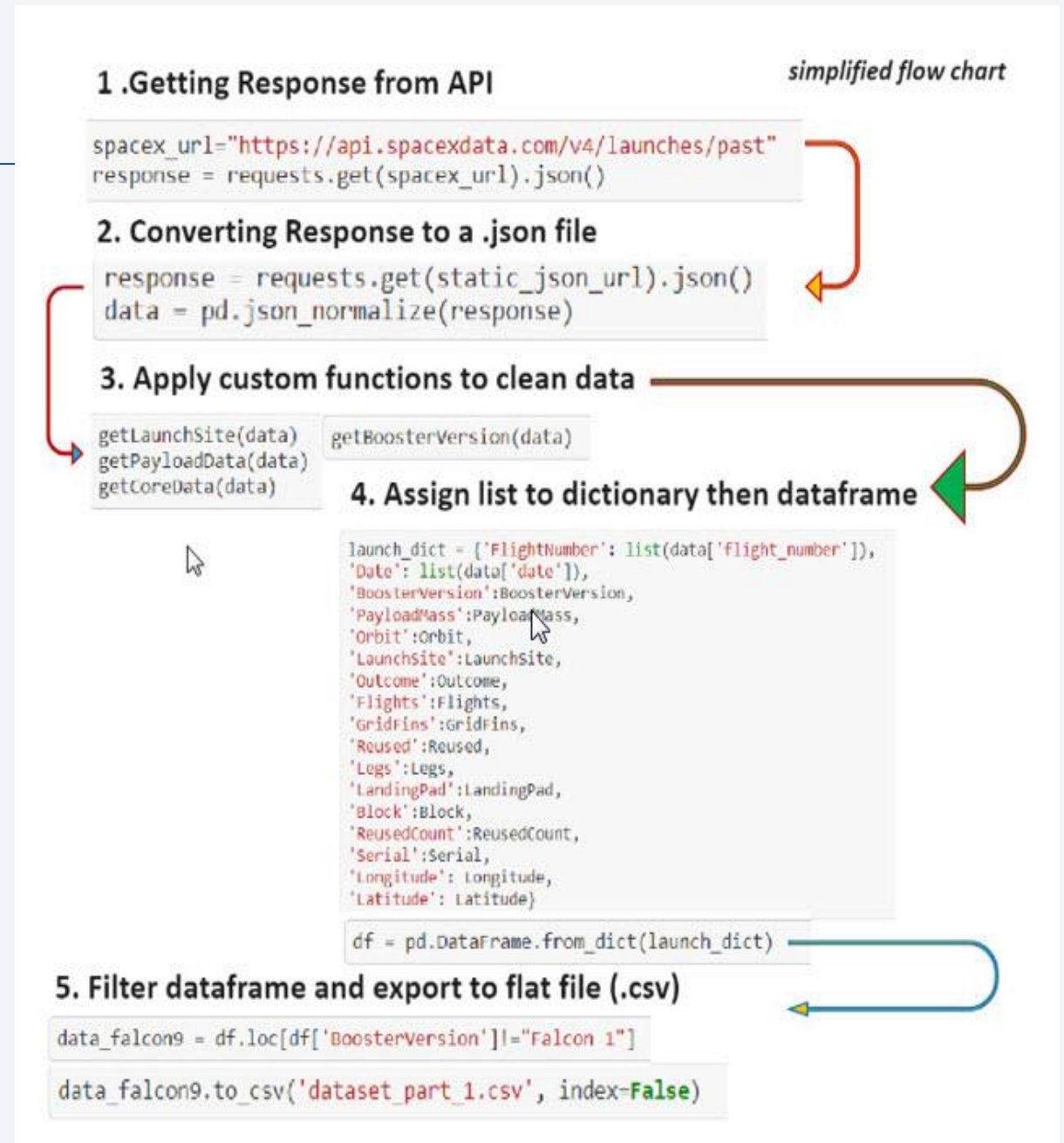


Data Collection – SpaceX API

- Data collection with SpaceX REST calls 

- GitHub URL of the completed SpaceX API calls notebook:

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb>



Data Collection - Scraping

- Web scraping process ➡
- GitHub URL of the completed web scraping notebook:

<https://github.com/flightdesigns/BM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb>

simplified flow chart

1. Getting Response from HTML

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

2. Creating BeautifulSoup Object

```
soup = BeautifulSoup(page.text, 'html.parser')
```

3. Finding tables

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

4. Getting column names

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

5. Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelevant column
del launch_dict['Date and time ( )']

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

6. Appending data to keys (refer) to notebook block 12

```
In [12]: extracted_row = 0
#Extract each table
for table_number,table in enumerate:
    # get table row
    for rows in table.find_all("tr"):
        #check to see if first table
```

7. Converting dictionary to dataframe

```
df = pd.DataFrame.from_dict(launch_dict)
```

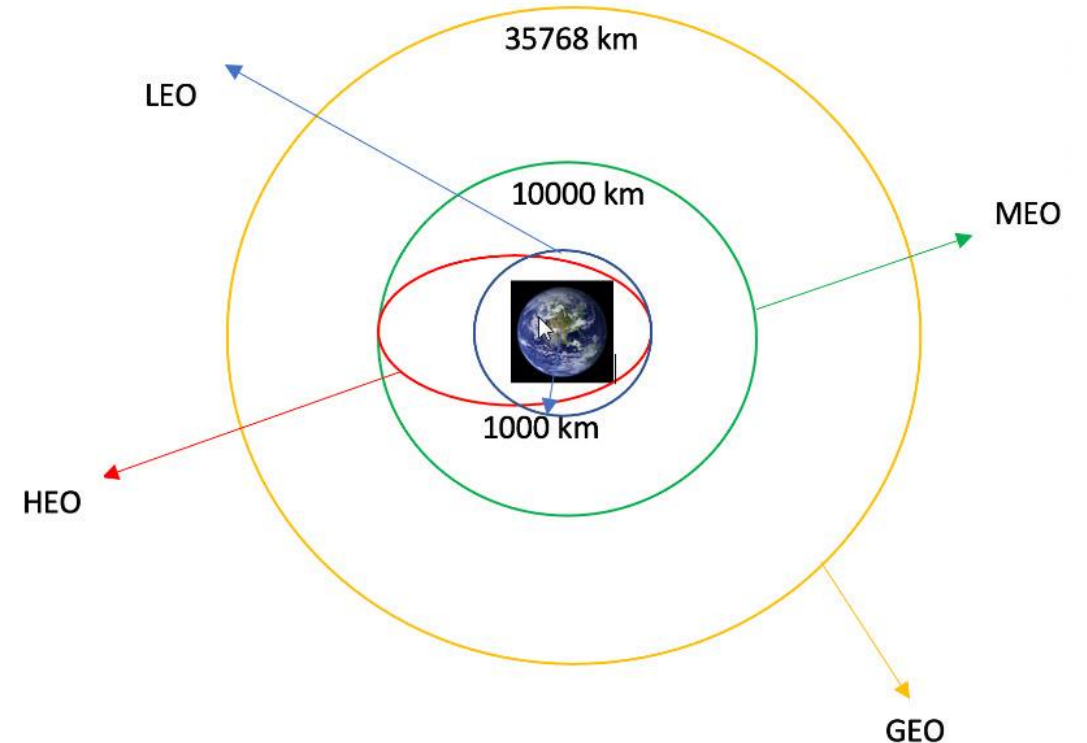
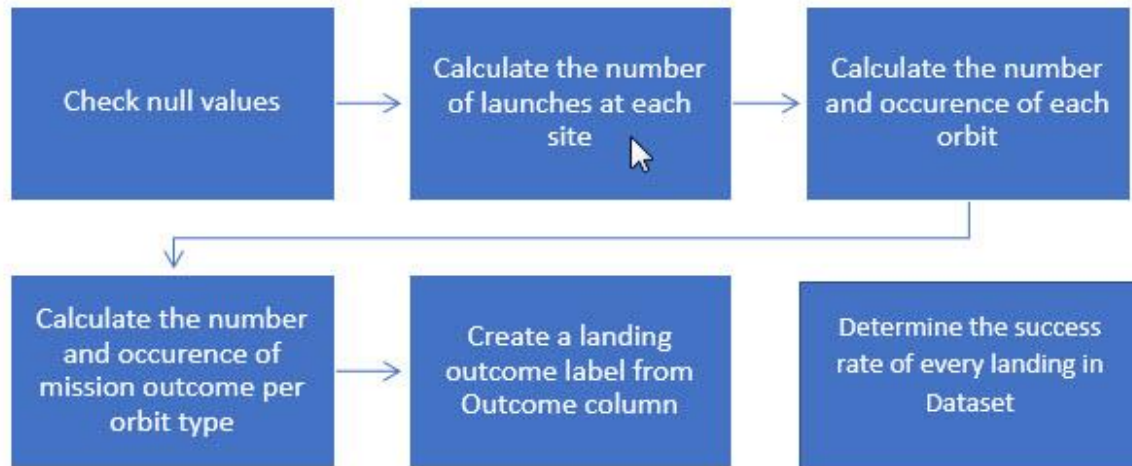
8. Dataframe to .CSV

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

Data Wrangling

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb>

Perform Exploratory Data Analysis on Dataset



EDA with Data Visualization

- GitHub URL of my completed EDA with data visualization notebook:
- <https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Exploratory%20Analysis%20with%20Visualization.ipynb>

Scatter Plots

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. They show how much one variable is affected by another. The relationship between two variables is called correlation. Scatter plots were used for:

- Flight Number vs Payload
- Flight Number vs Launch Site
- Payload vs Launch Site
- Orbit vs Flight Number
- Payload vs Orbit Type
- Orbit vs Payload Mass

Bar Graphs

A bar graph is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. This was used for:

- Mean vs Orbit

Line graph

A line graph is a type of chart used to show information that changes over time. They clearly show data variables and trends and are useful for predictions about results of data not yet recorded. This was used for:

- Success Rate vs Year

EDA with SQL

GitHub URL of my completed EDA with SQL notebook:

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Exploratory%20Analysis%20Using%20SQL.ipynb>

SQL Queries Performed

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the 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 but 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. Used a subquery
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for the in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Build an Interactive Map with Folium

- We took the latitude and longitude coordinates at each launch site and added a circle marker around each launch site with a label.
- 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. Questions asked:
 - Are launch sites in close proximity to railways and highways? No
 - Are launch site in close proximity to coastlines? Yes
 - Do launch sites keep certain distance away from cities? Yes

GitHub URL of my interactive map with Folium:

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics.ipynb>

Build a Dashboard with Plotly Dash

- We added a dropdown list to enable Launch Site selection including the following options - All Sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC- 39A
- We added a pie chart to show the total successful launches count for all sites
- We added a slider to select payload which ranges from 0 -10000
- We added a scatter chart to show the correlation between payload and launch success

GitHub URL of my completed Plotly Dash lab:

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Dashboard.py>

Predictive Analysis (Classification)

The below machine learning stages was used to built, evaluate, improve, and determine the best performing classification model:

- Import the required libraries
- Load the cleaned data
- Standardizing the data to prevent the bias
- Splitting the data into 20% for testing data and 80% training data
- Initializing 4 different classification algorithms:
 - Logistic Regression (LR)
 - Support Vector Machine (SVM)
 - Decision Tree (DT)
 - K nearest neighbours (KNN)
- Using GridSearchCV technique to find the best parameters
- Using Evaluation techniques such as Confusion matrix , F1 score, Jaccard Score for the purpose of using the best model among the algorithms above.

GitHub URL of my completed predictive analysis lab:

<https://github.com/flightdesigns/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb>

Results

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

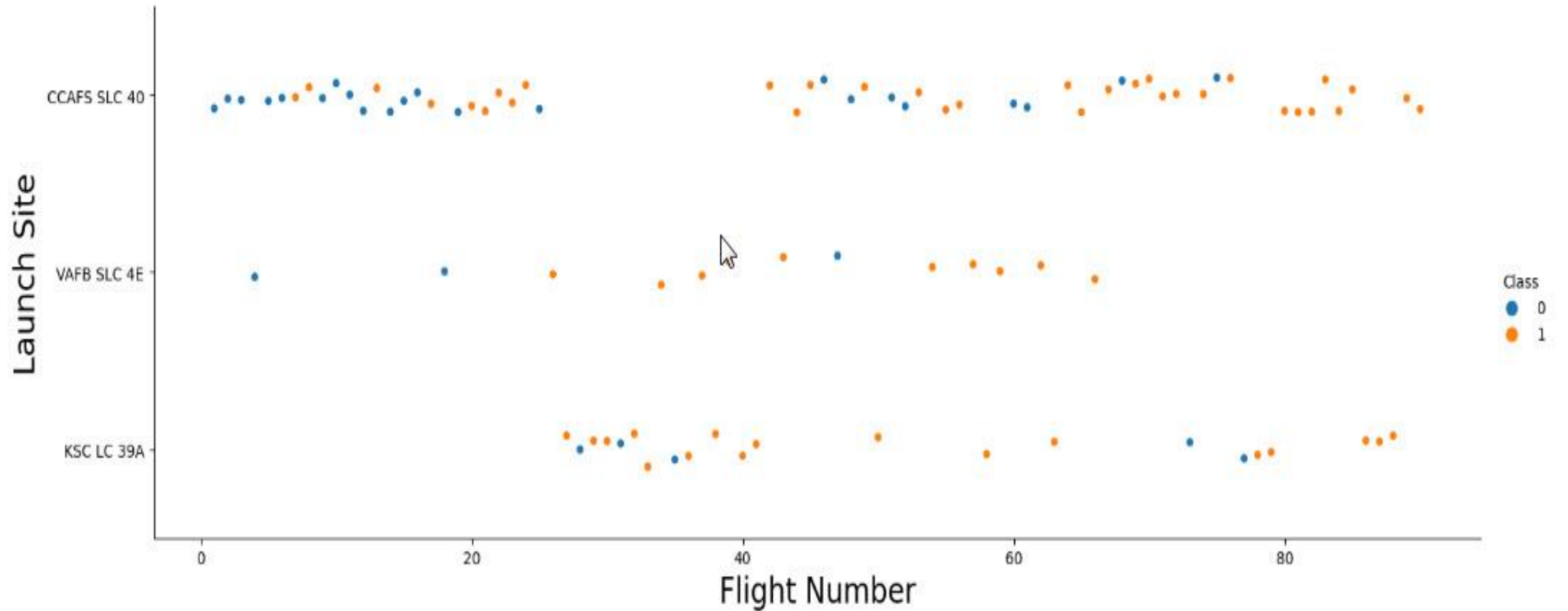


The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

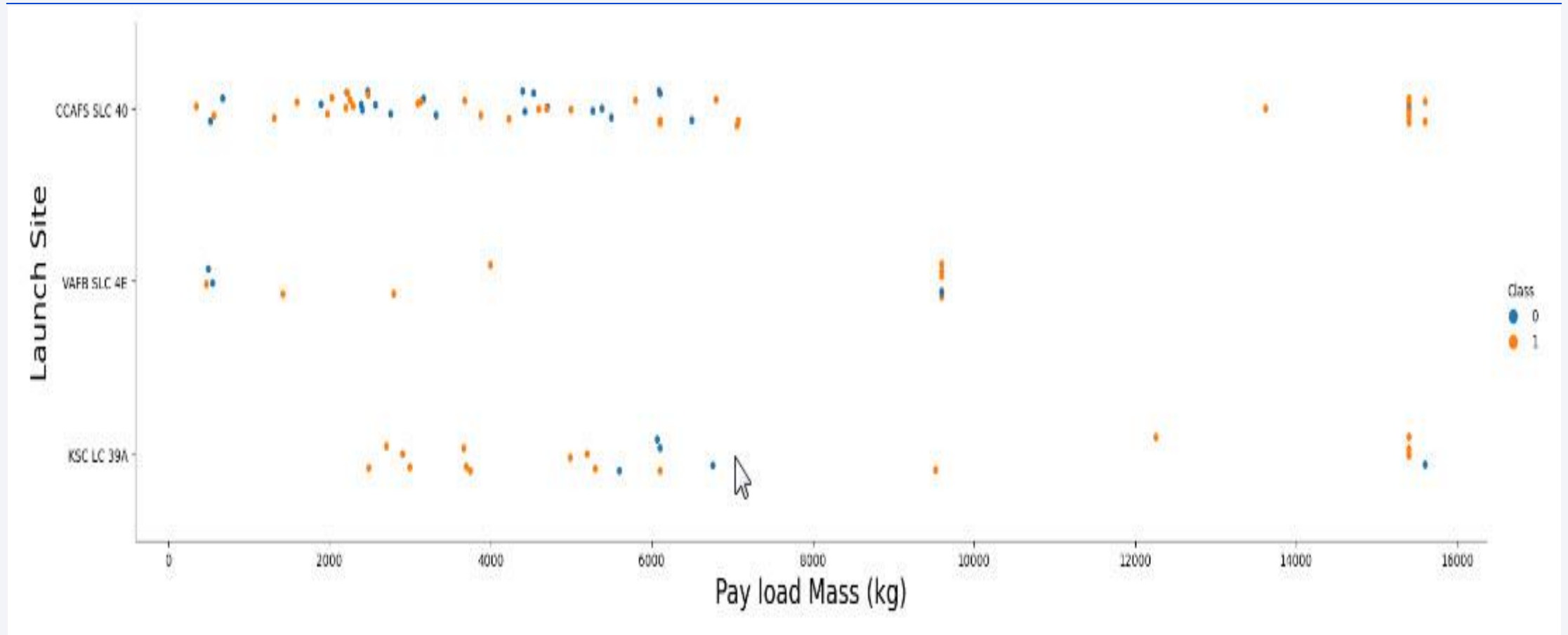
Insights drawn from EDA

Flight Number vs. Launch Site



CCAFS SLC 40 is the most used Launch Site with significantly higher launches than other sites.

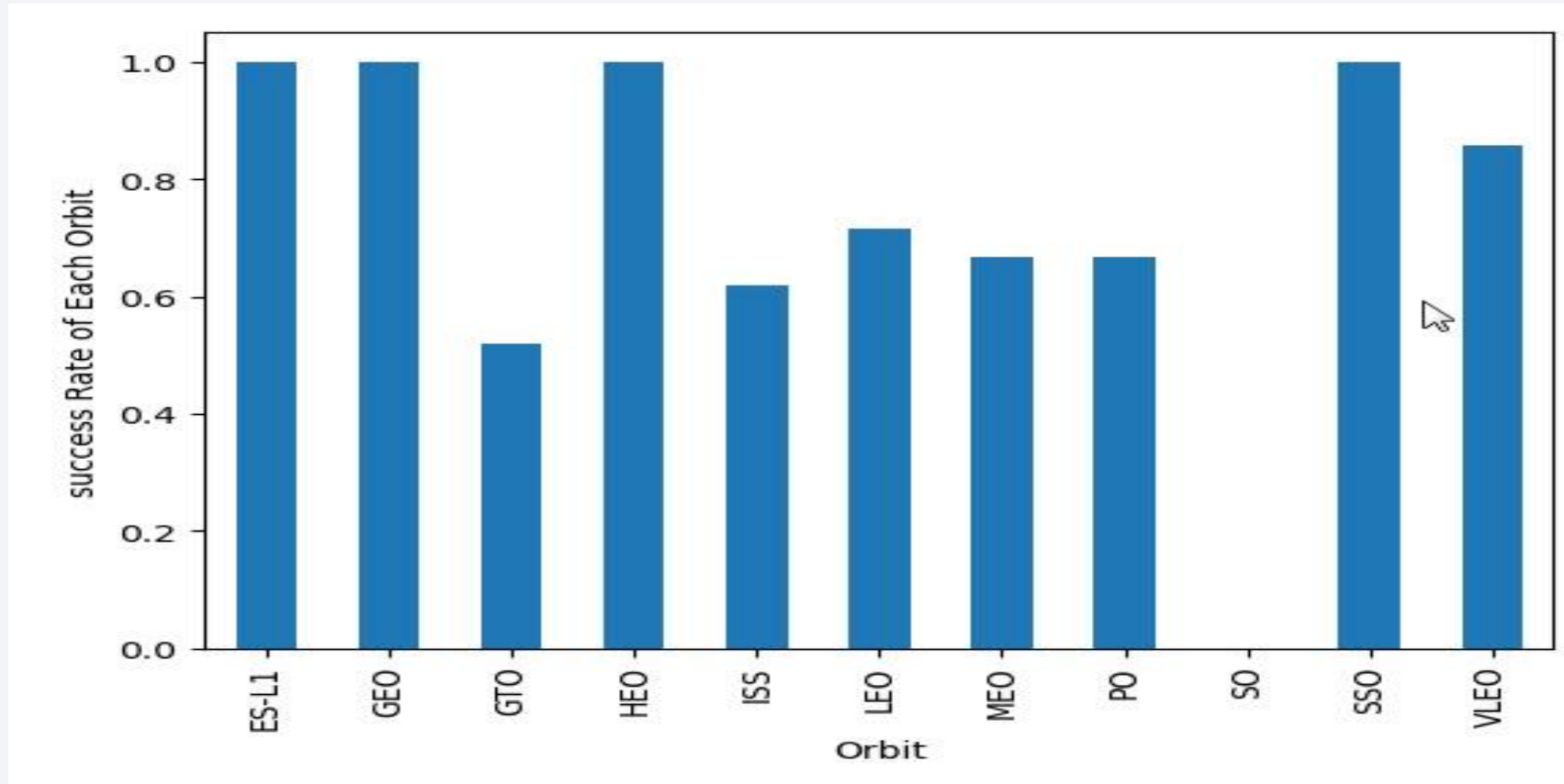
Payload vs. Launch Site



An increase in frequency of flights and Payload Mass improves the success rate.

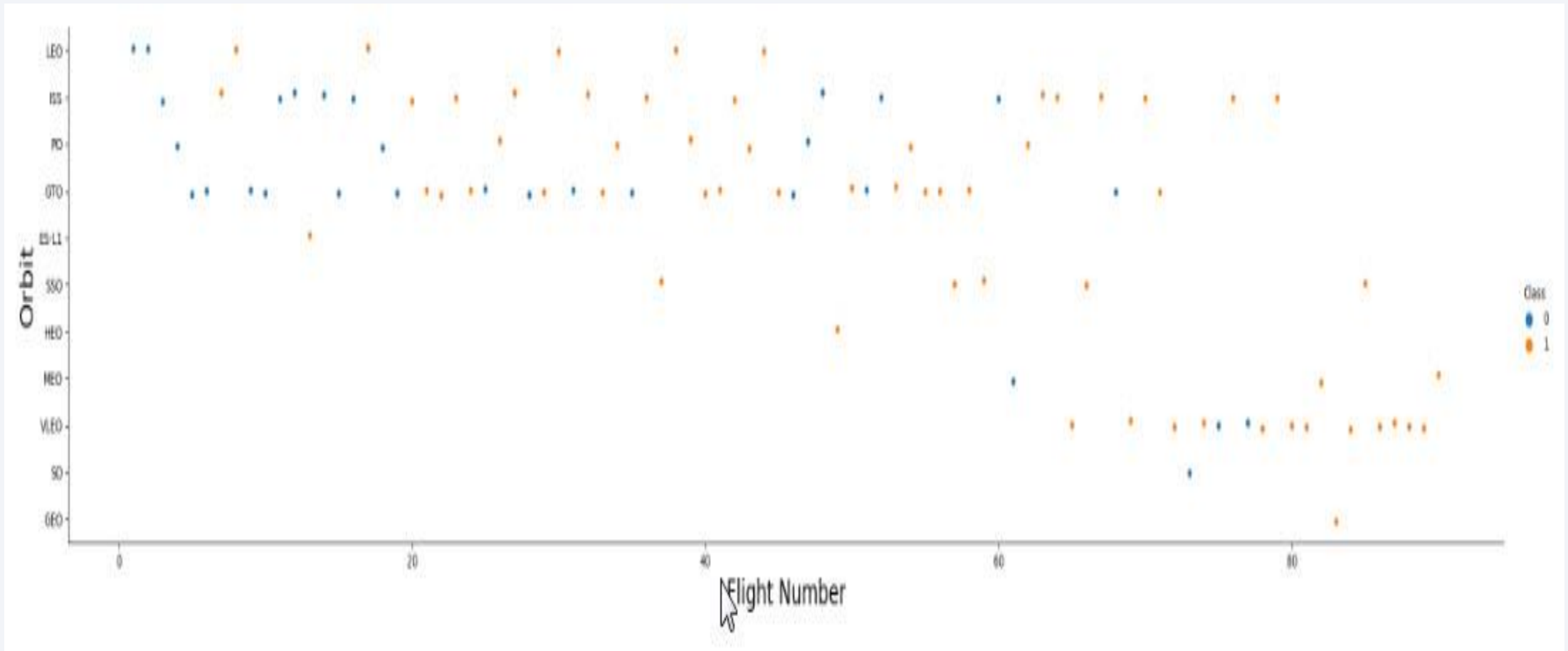
CCFAS SLC 40 remains the most active and most successful launch site.

Success Rate vs. Orbit Type



ES-L1, GEO, HEO and SSO Orbit Types have the highest success rate

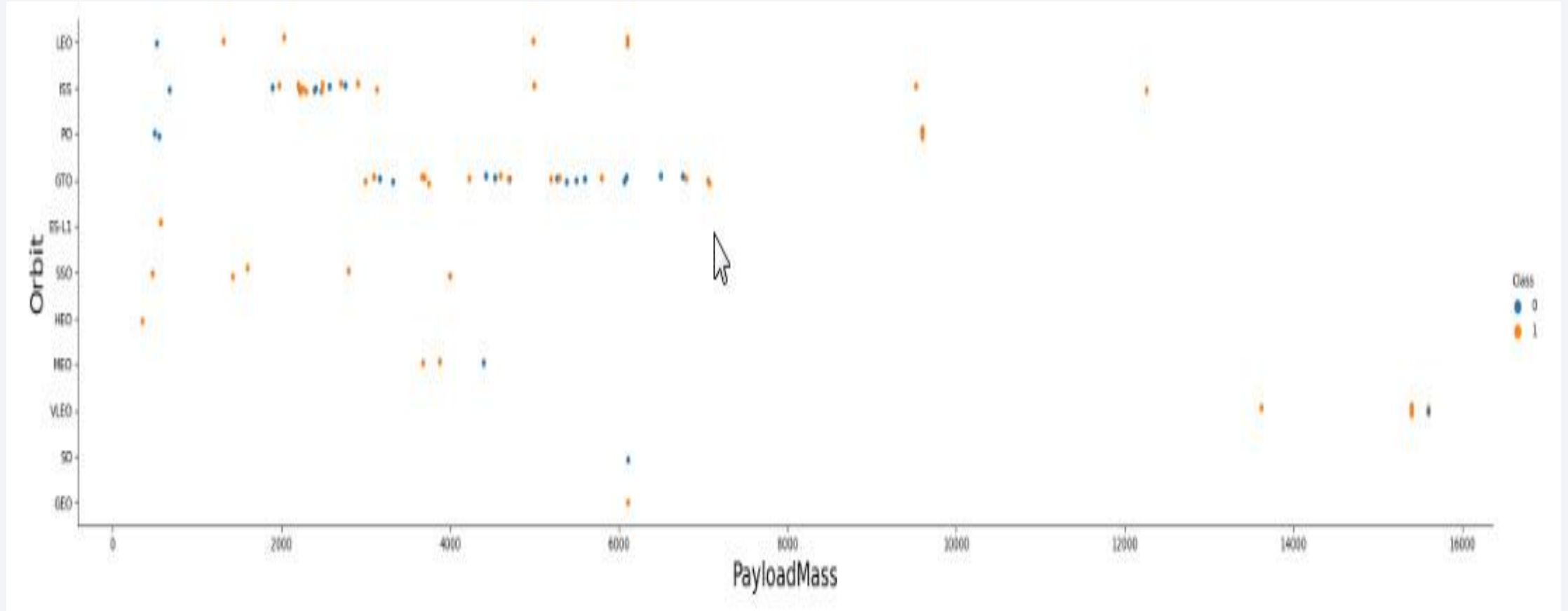
Flight Number vs. Orbit Type



In the LEO orbit, increased flights improves the success rate.

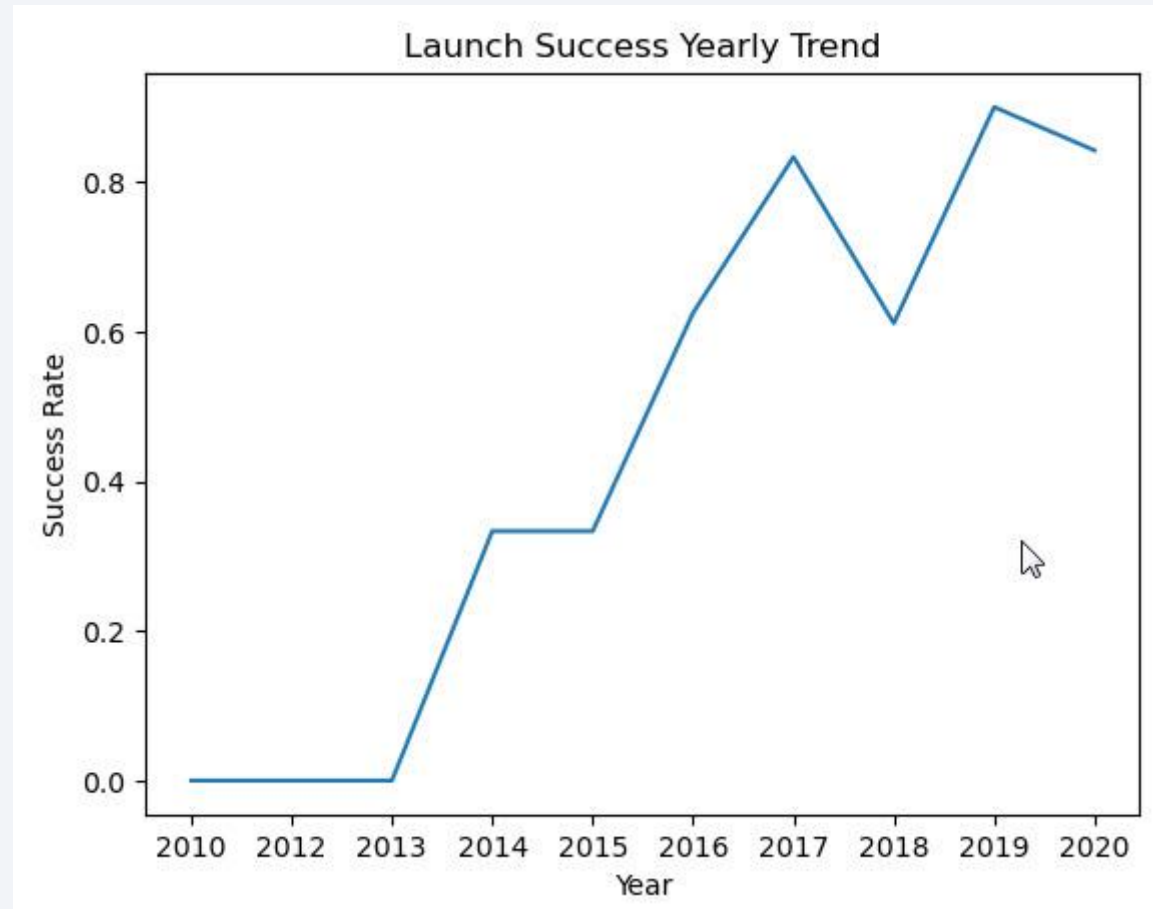
There is no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



Heavy payloads have a negative influence on GTO orbits and positive impact on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



The success rate has increased consistently from 2013 to 2020.

All Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

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

Done.

Launch_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

There are 4 unique launch sites

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

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

<u>sum(payload_mass__kg_)</u>
45596

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

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

```
Done.
```

avg_mass_F9

2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

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

```
MIN("DATE")
```

```
01-05-2017
```

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) as outcome from SPACEXTBL GROUP BY mission_outcome
```

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

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

- Total number of successful mission outcomes – 100
- Total number of failure mission outcomes – 1

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

```
Done.
```

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

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and subst
```

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

```
Done.
```

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' GROUP BY "LANDING _OU"
```

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

Done.

Landing_Outcome	COUNT("LANDING _OUTCOME")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky with stars and a view of the Earth's surface from space. The Earth's surface is mostly dark blue, with a thin layer of white clouds. A curved horizon line separates the dark sky from the Earth's surface. In the lower right, there are bright, glowing yellow and orange lights, likely representing city lights or industrial activity. The overall image has a high-contrast, cinematic quality.

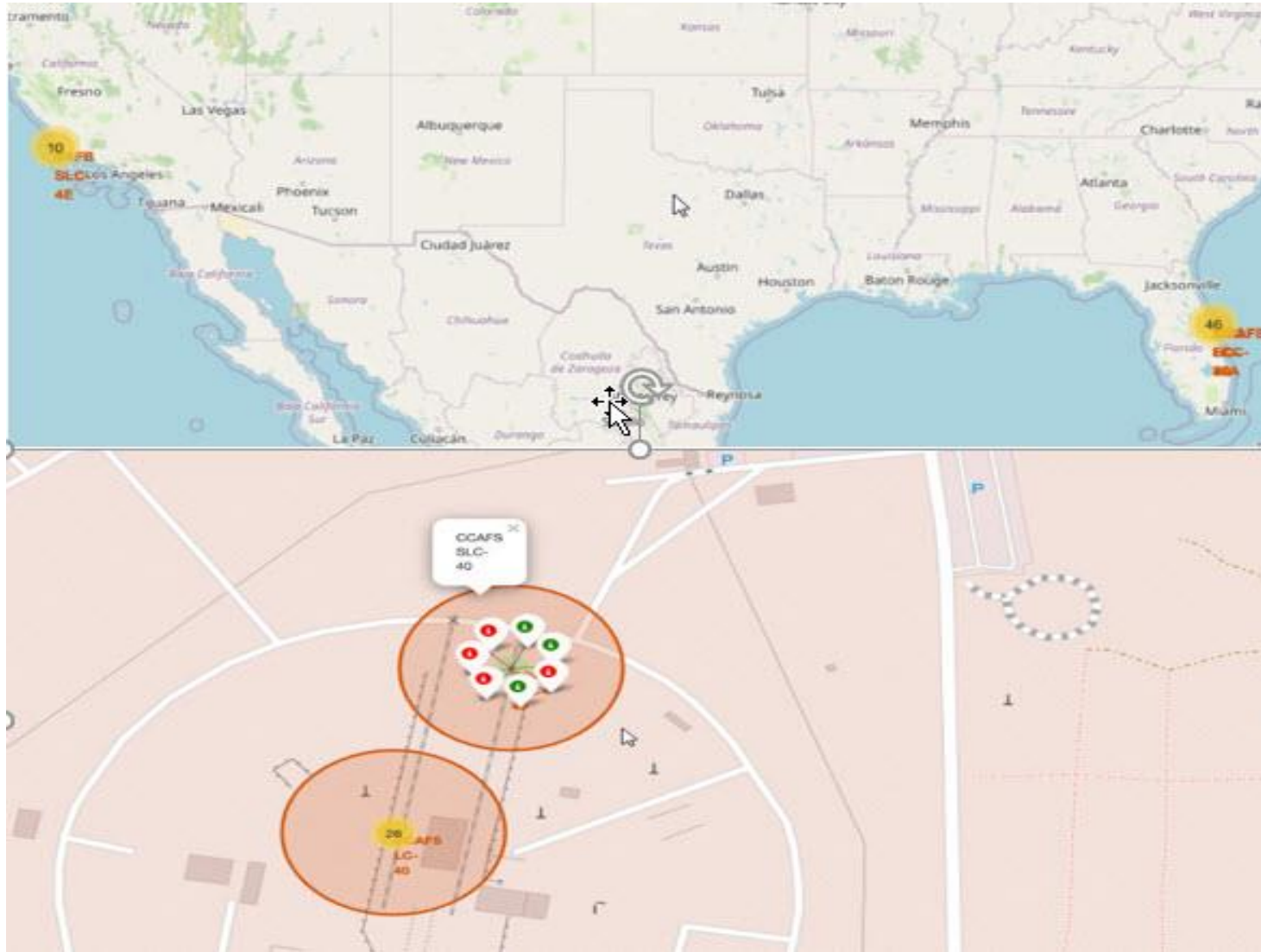
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites



SpaceX Launch Sites are located in the United States, along the California and Florida Coastlines



Launch Location Success Rate

Launch Site Proximity Map

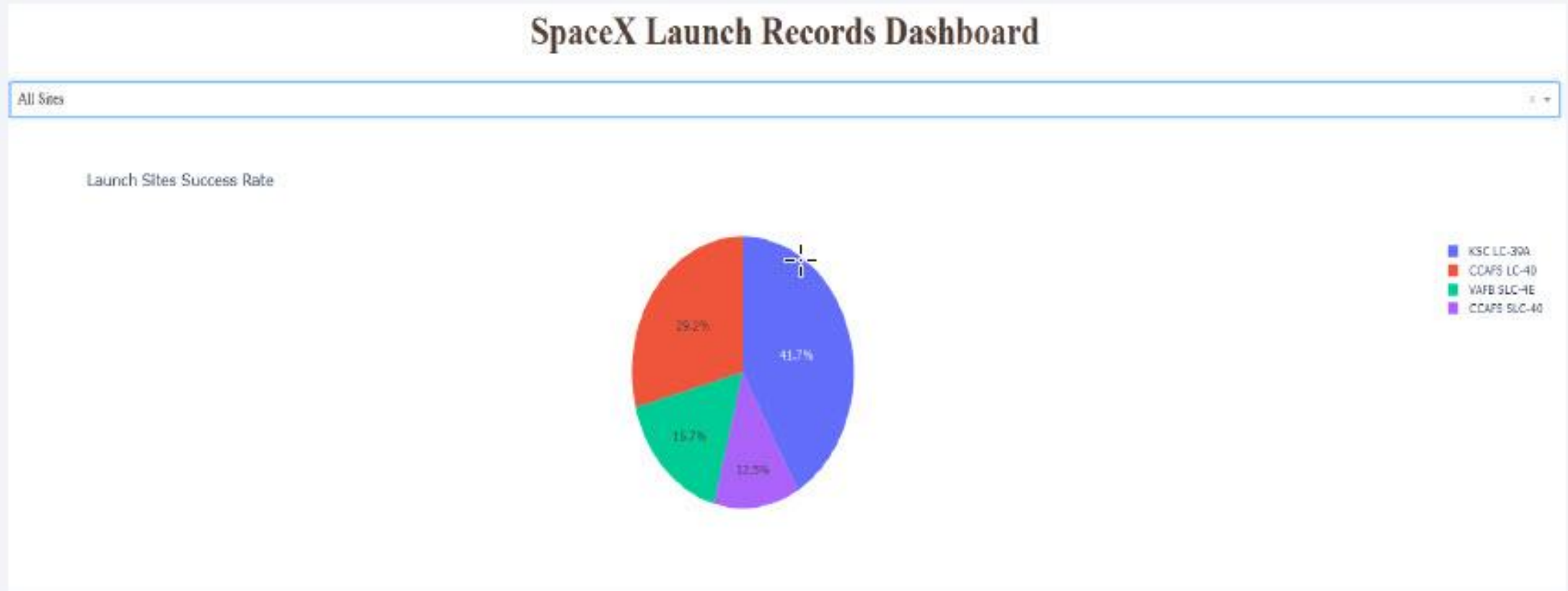




Section 4

Build a Dashboard with Plotly Dash

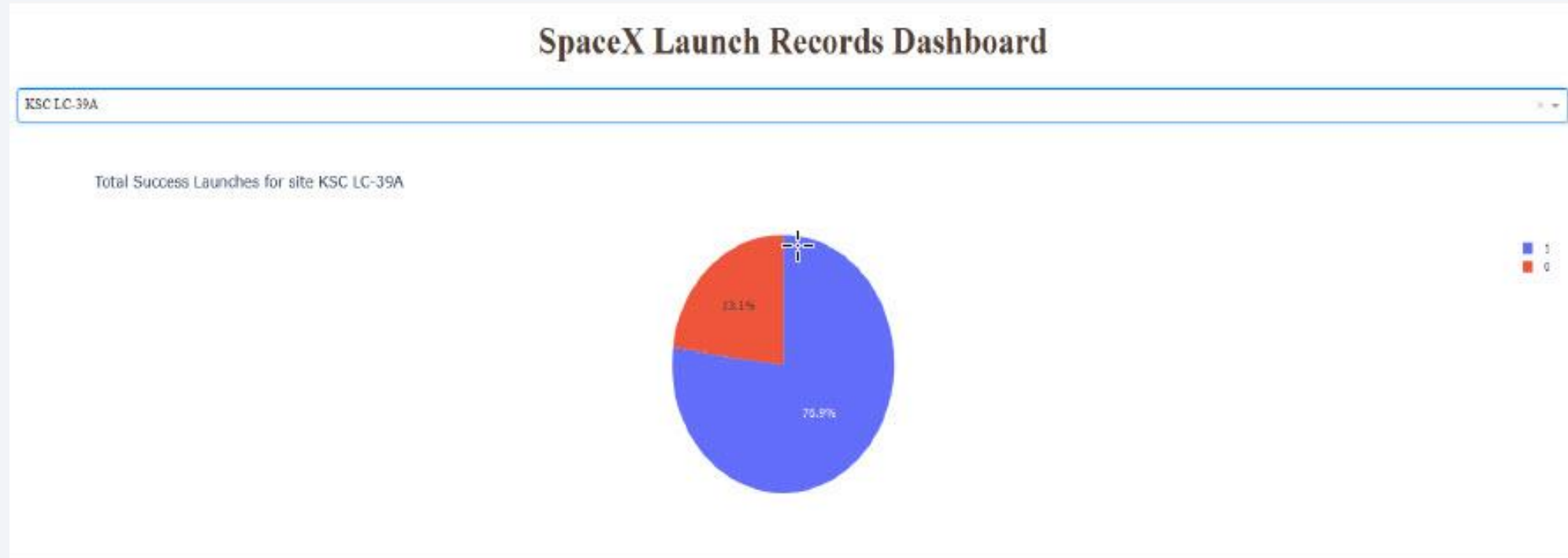
Dashboard – Launch Success Count of Sites



KSC LC-39A is the most successful launch site.

CCAFS SLC-40 is the least successful launch site.

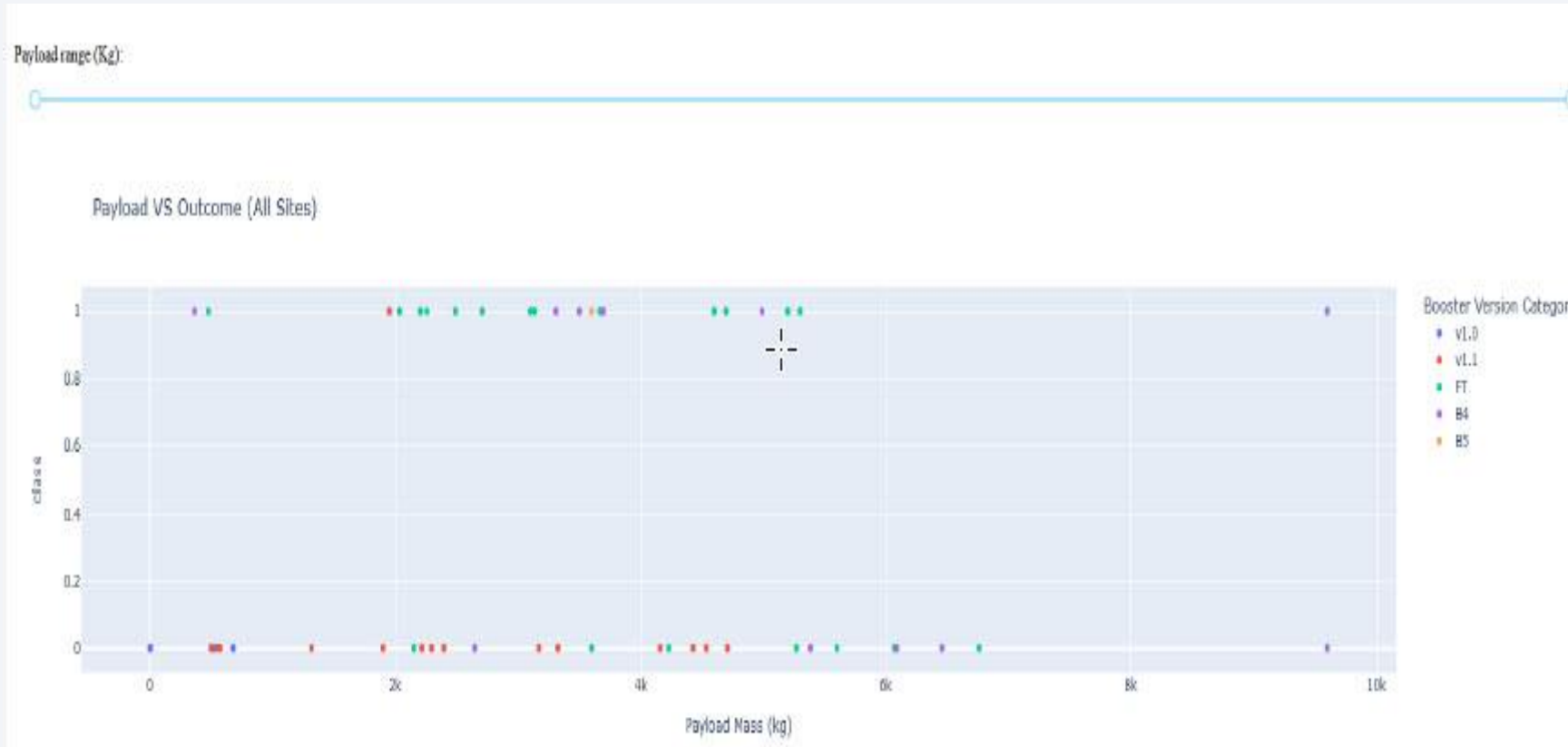
Dashboard – Launch Success for KSC LC-39A



Total Success Launches for site KSC LC-39

- KSC LC-39A with 76.9% successful missions
- KSC LC-39A with 23.1% Failed missions

Dashboard – Payload vs Launch Outcome

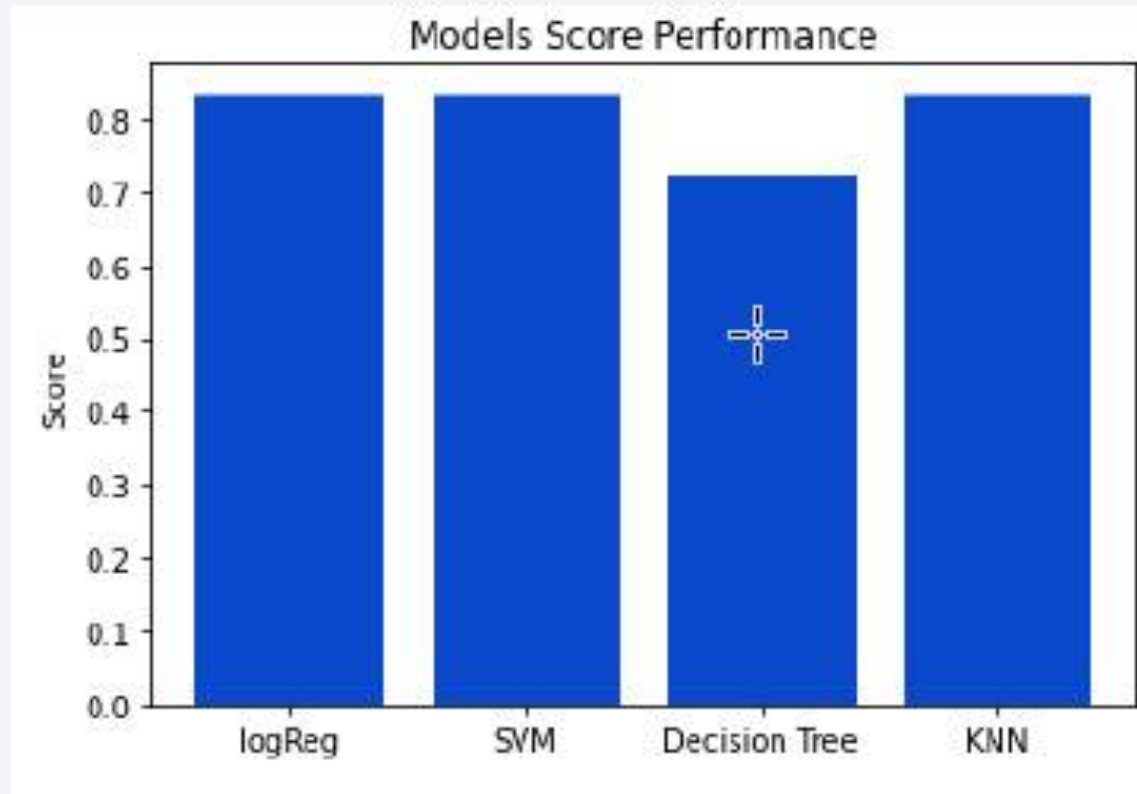


The success rate of low weighted payloads (0 – 4000kg) is higher than the heavy weighted payloads (>4000 – 10000kg)

Section 5

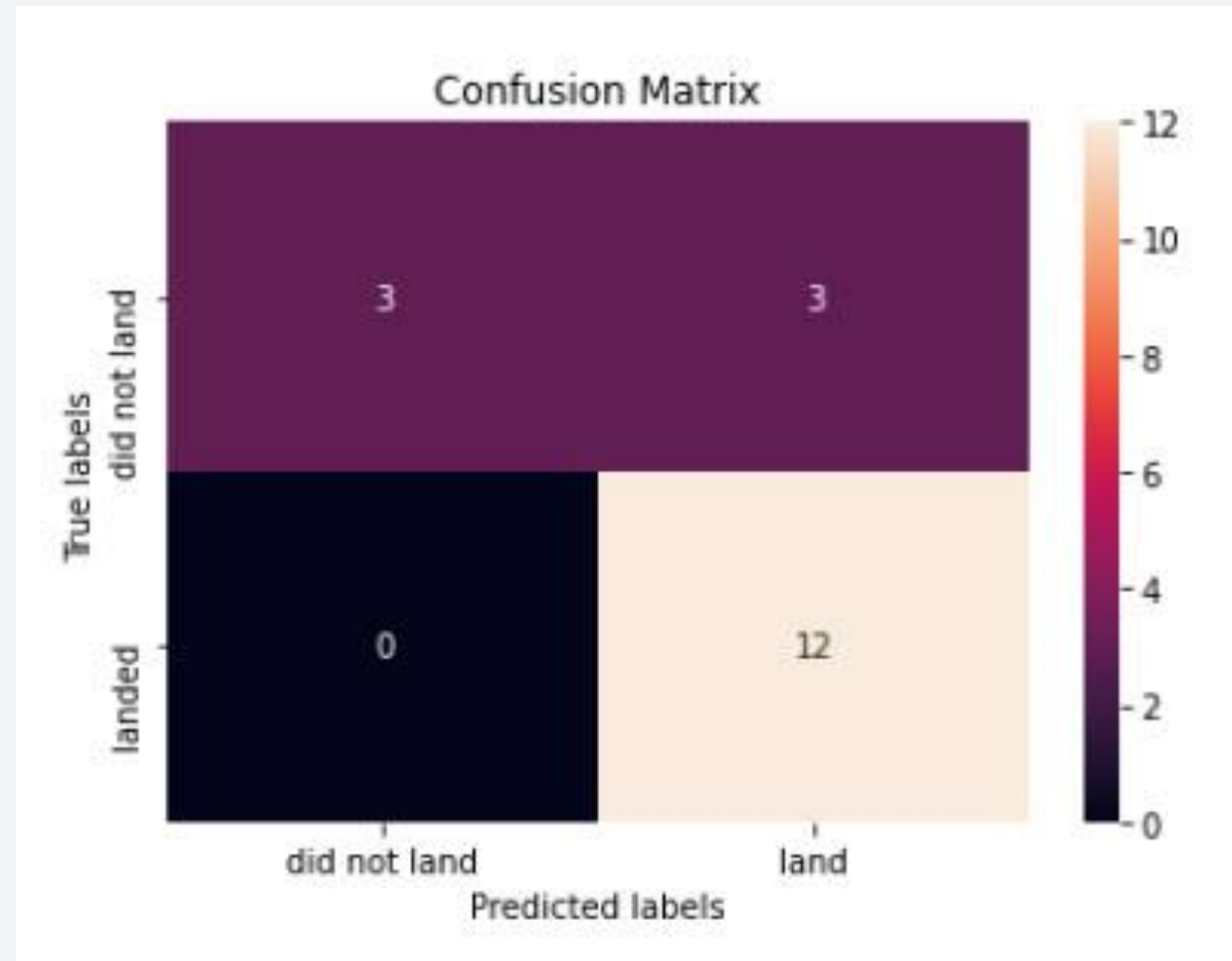
Predictive Analysis (Classification)

Classification Accuracy



Logistic Regression, SVM and KNN achieved similar classification accuracy with a performance score of 0.8. Decision Tree has the lowest score.

Confusion Matrix



Logistic Regression , SVM and KNN have the same confusion matrix and results.

Conclusions

- **SpaceX launch success rate improves yearly**
- **Low weighted payloads perform better than the heavier payloads**
- **KSC LC-39A has the highest success rate**
- **Orbit ES-L1, GEO, HEO and SSO has the highest success rate**
- **Logistic Regression, SVM and KNN models have the best prediction accuracy for the dataset used for this project**

Thank you!

