IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rui Huang

06.01.2024

Outline

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

Executive Summary

Methodologies

The aim of this study is to identify the determinants of a successful rocket landing. To achieve this objective, the investigation employed the methodologies including:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA) using SQL, Pandas, Matplotlib
- Interactive Visual Analysis using Analytics using Folium, Plotly, Dash
- Predictive Analysis using Machine Learning

Summary

The results were being extracted using the following essential component as shown in subsequent slides.

- **Exploratory Data Analysis (EDA)**
- **Interactive Visual Analysis**
- **Predictive Analysis**

Introduction

Project background

SpaceX, founded by Elon Musk in 2002, is a California-based aerospace company focused on **reducing space transportation costs and advancing space exploration**. Notable achievements include operating Falcon 9 and Falcon Heavy rockets, Dragon and Starship spacecraft, and creating the largest-ever satellite constellation called Starlink.

SpaceX has pioneered various milestones, such as the first private liquid-propellant rocket to reach orbit, launching, orbiting, and recovering spacecraft, sending a spacecraft to the International Space Station, and achieving vertical propulsive landings and reusability of orbital rocket boosters.

Problems statement

- In this capstone project, our objective is to **forecast the successful landing of the Falcon 9 first stage**.
- SpaceX promotes its Falcon 9 rocket launches on its official platform, pricing them at 62 million dollars, significantly lower than other providers whose costs can go up to 165 million dollars per launch.
- The **substantial savings SpaceX achieves are largely attributed to the ability to reuse the first stage**.
- By accurately predicting the first stage landing outcome, we can **effectively estimate the overall launch cost**. This insight becomes crucial for other companies considering bidding against SpaceX for a rocket launch contract.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - **Using SpaceX REST API and web scraping techniques**
- **Perform data wrangling**
  - **By filtering the data, handling missing values and applying hot encoding**
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - **To find best model and parameters**

Data Collection

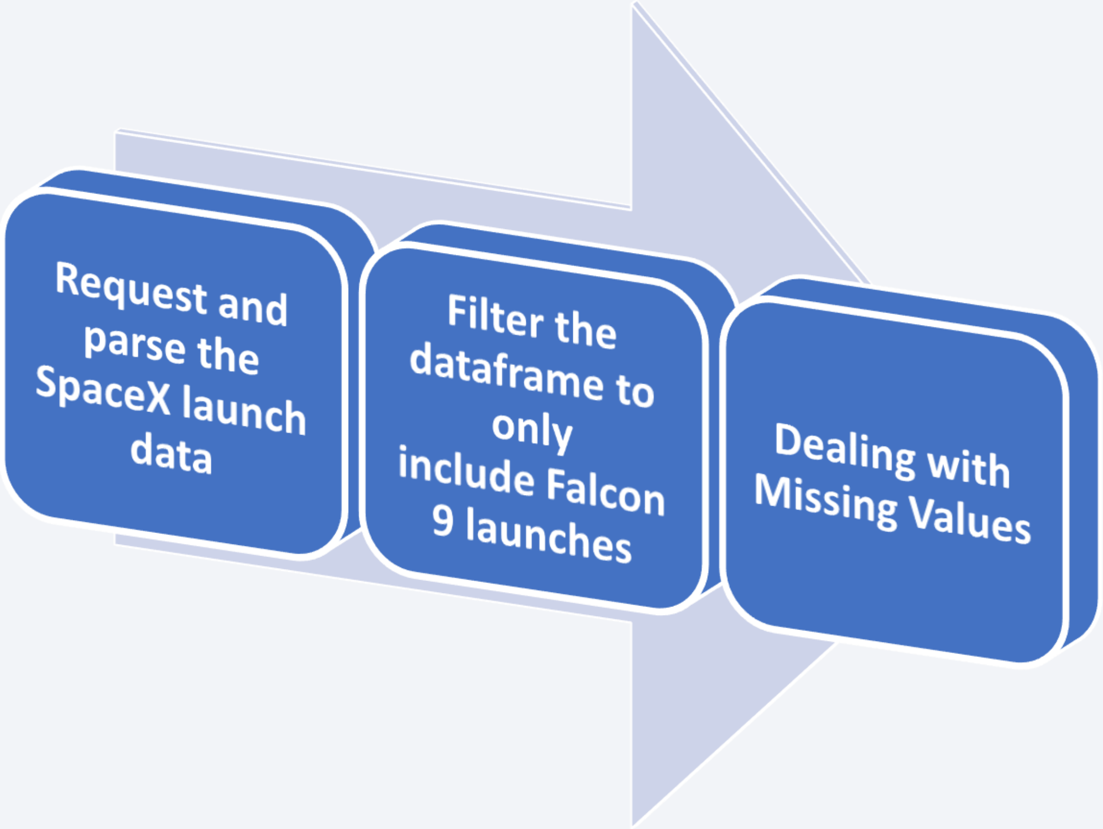
This information may prove valuable if an alternate company intends to bid against SpaceX for a rocket launch. The data will be collected and **ensured to be in the correct format from an API**.

Web scraping will be performed to collect **Falcon 9 historical launch records from Wikipedia** page titled "List of Falcon 9 and Falcon Heavy launches."

Data Collection – SpaceX API

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/1_Data%20Collection%20API.ipynb

CSV file outcome after methodology:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/dataset_part_1.csv


Request and parse the SpaceX launch data

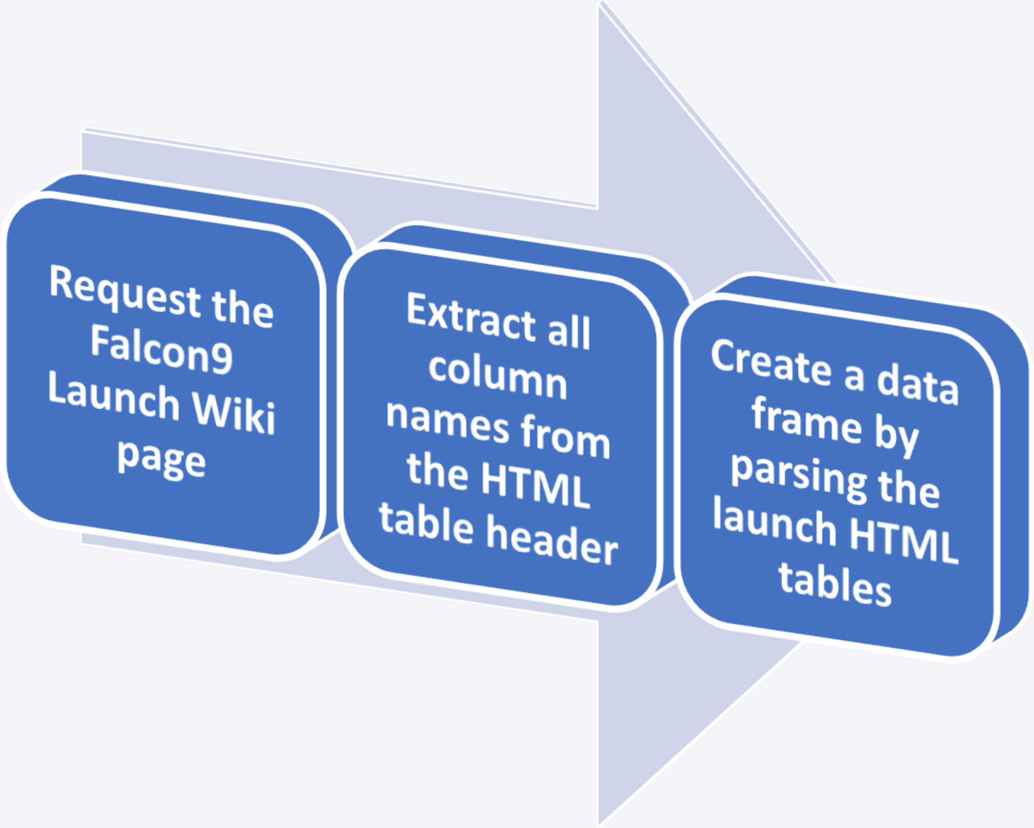
Filter the dataframe to only include Falcon 9 launches

Dealing with Missing Values

Data Collection - Scraping

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/2_Webscraping%20.ipynb

CSV file outcome after methodology:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/spacex_web_scraped.csv

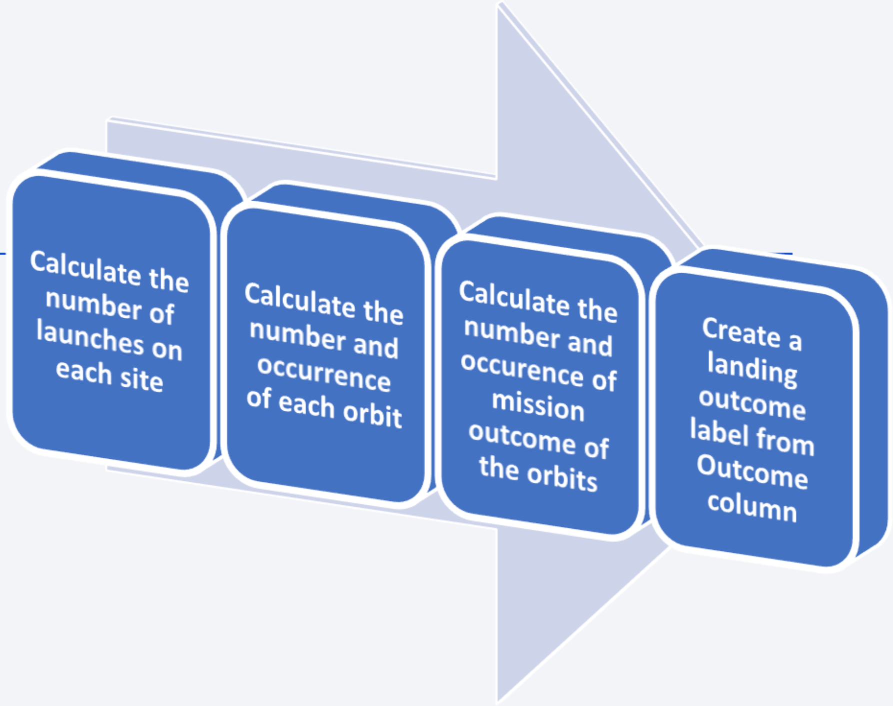


Request the Falcon9 Launch Wiki page

Extract all column names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling



Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

- Exploratory Data Analysis (EDA) will be performed to identify patterns in the data and determine the label for training supervised models.

- Within the dataset, various cases exist where the booster did not land successfully. Instances include attempted landings that failed due to accidents.

- The primary focus of this lab will be to transform these outcomes into training labels, where 1 signifies a successfully landed booster, and 0 signifies an unsuccessful landing.

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/3_Data%20Wrangling.ipynb

CSV file outcome after methodology:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/dataset_part_2.csv

EDA with Data Visualization

The relationship through the chart is visualised
- Scatterplots
- Barplots

Then, create dummy variables to categorical column and cast numeric columns to .float

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/5_EDA_Data%20Visualisation.ipynb

CSV file outcome after methodology:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/dataset_part_3.csv

Payload Mass vs Flight Number

Launch Site vs Flight Number

Launch Site vs Payload Mass

Orbit vs Flight Number

Payload vs Orbit

Year vs Average Success Rate

EDA with SQL

Queries

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA’;
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

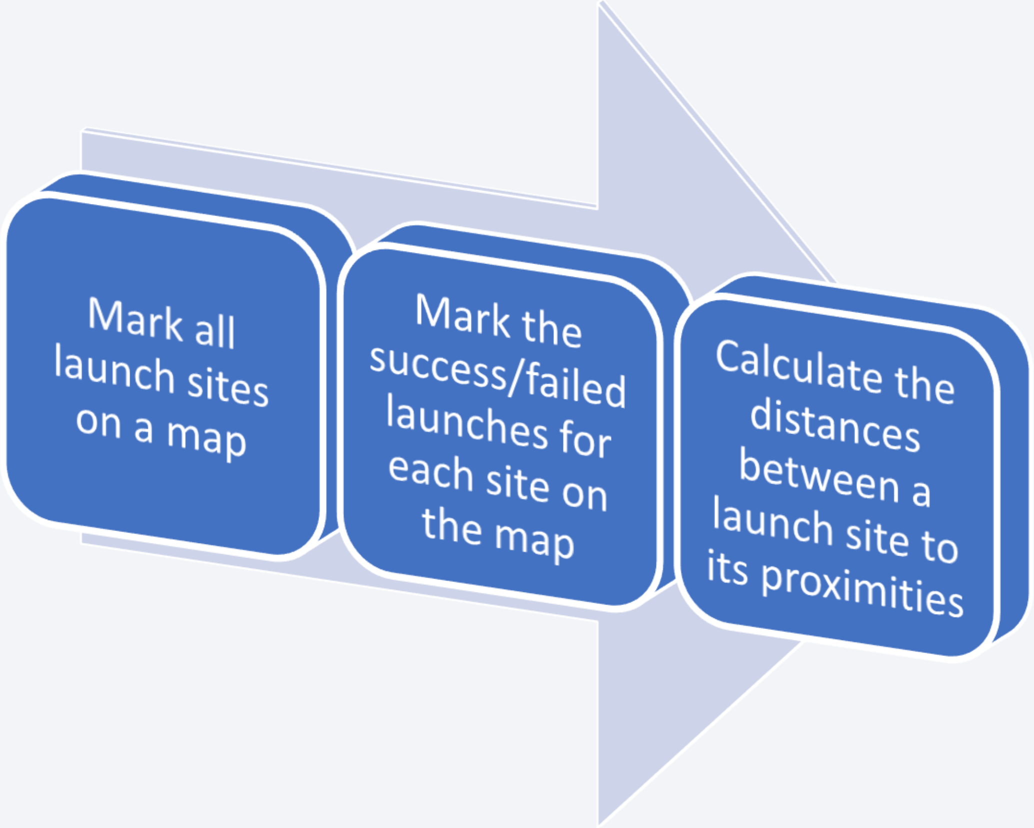
GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/4_EDA_SQLite.ipynb

Build an Interactive Map with Folium

The launch success rate may depend on **location and proximities** of a launch site, i.e., the initial position of rocket trajectories.

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers: points
- Circles: highlighted areas around specific coordinates
- Lines: distances between two coordinates
- Marker clusters: groups of events in each coordinate



Mark all launch sites on a map

Mark the success/failed launches for each site on the map

Calculate the distances between a launch site to its proximities

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/6_Data%20Visulisation_Folium.ipynb

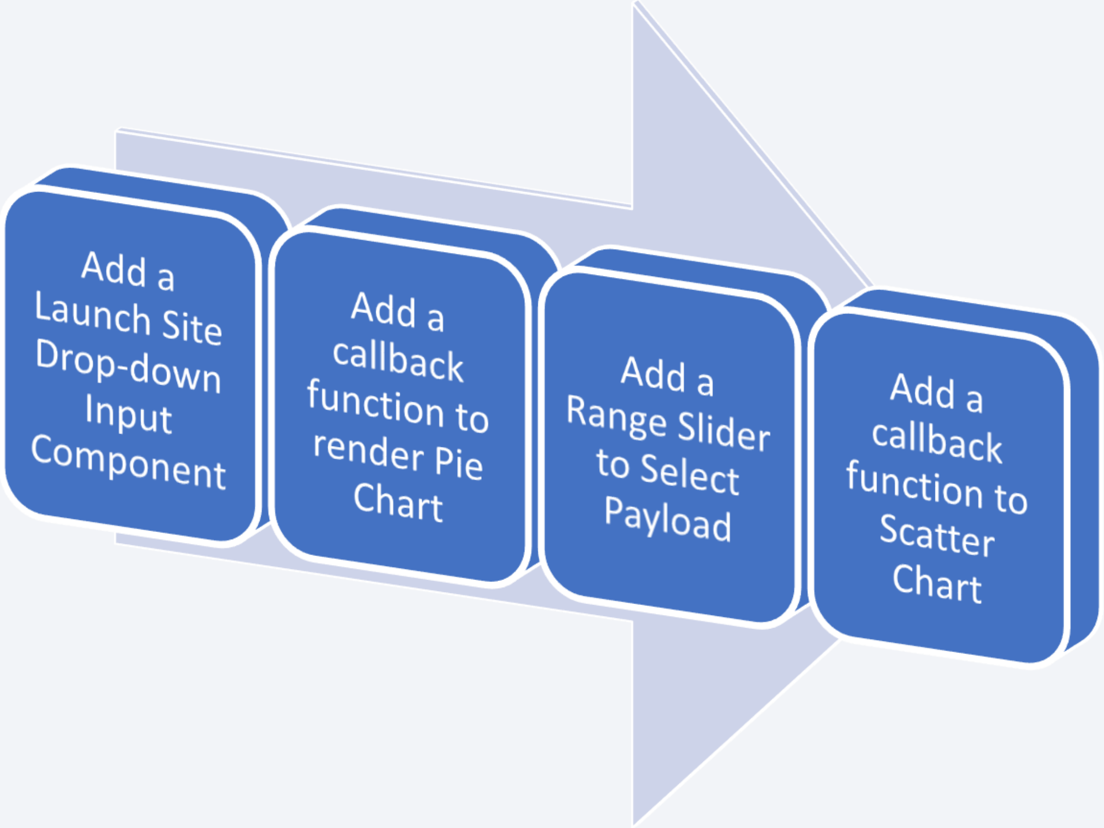
Build a Dashboard with Plotly Dash

A Plotly Dash application will be built for users to engage in interactive visual analytics on SpaceX launch data in real-time.

The dashboard application contains input components, including:

- A dropdown list to interact with a pie chart and a scatter point chart
- A range slider to interact with a scatter point chart

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/7_Data%20Visulisation_Plotly_Dash.py



Add a Launch Site Drop-down Input Component

Add a callback function to render Pie Chart

Add a Range Slider to Select Payload

Add a callback function to Scatter Chart

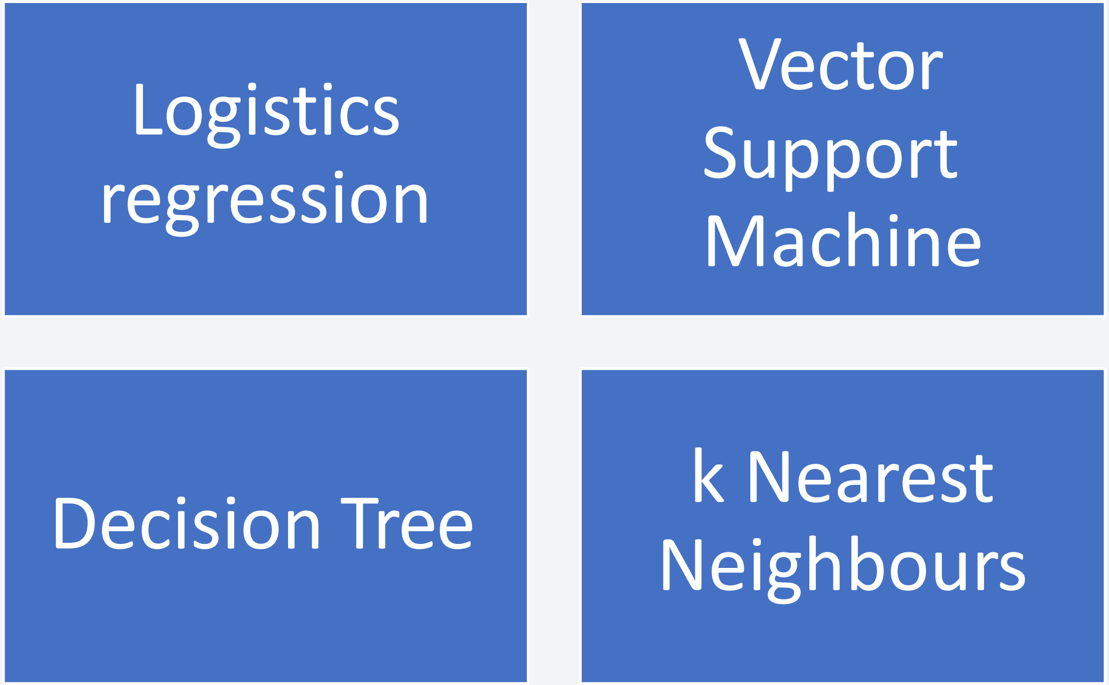
Predictive Analysis (Classification)

Exploratory Data Analysis will be performed

- Create a column for the class
- Standardize the data
- Split into training data and test data

Then, the method which performs best using test data is determined
by finding the best Hyperparameter for SVM, Classification Trees and Logistic Regression

GitHub Reference:
https://github.com/andyngjw/IBM-DataSc-Capstone-SpaceX/blob/main/8_Machine%20Learning.ipynb



Logistics regression

Vector Support Machine

Decision Tree

k Nearest Neighbours

Results

Exploratory data analysis results through graphs

- CCAFS SLC 40 has the highest success rate, compared to KSC LC 39A and VAFB SLC 4E
- The higher the payload mass, the higher the success rate
- ES-L1, GEO, HEO, and SSO orbit has 100% success rate
- Some orbits has higher success rate, but the total flight number in those orbits is less
- The payload does not show significant difference compared to orbit type
- The success rate shows improvement throughout the year starting from 2013 to 2020

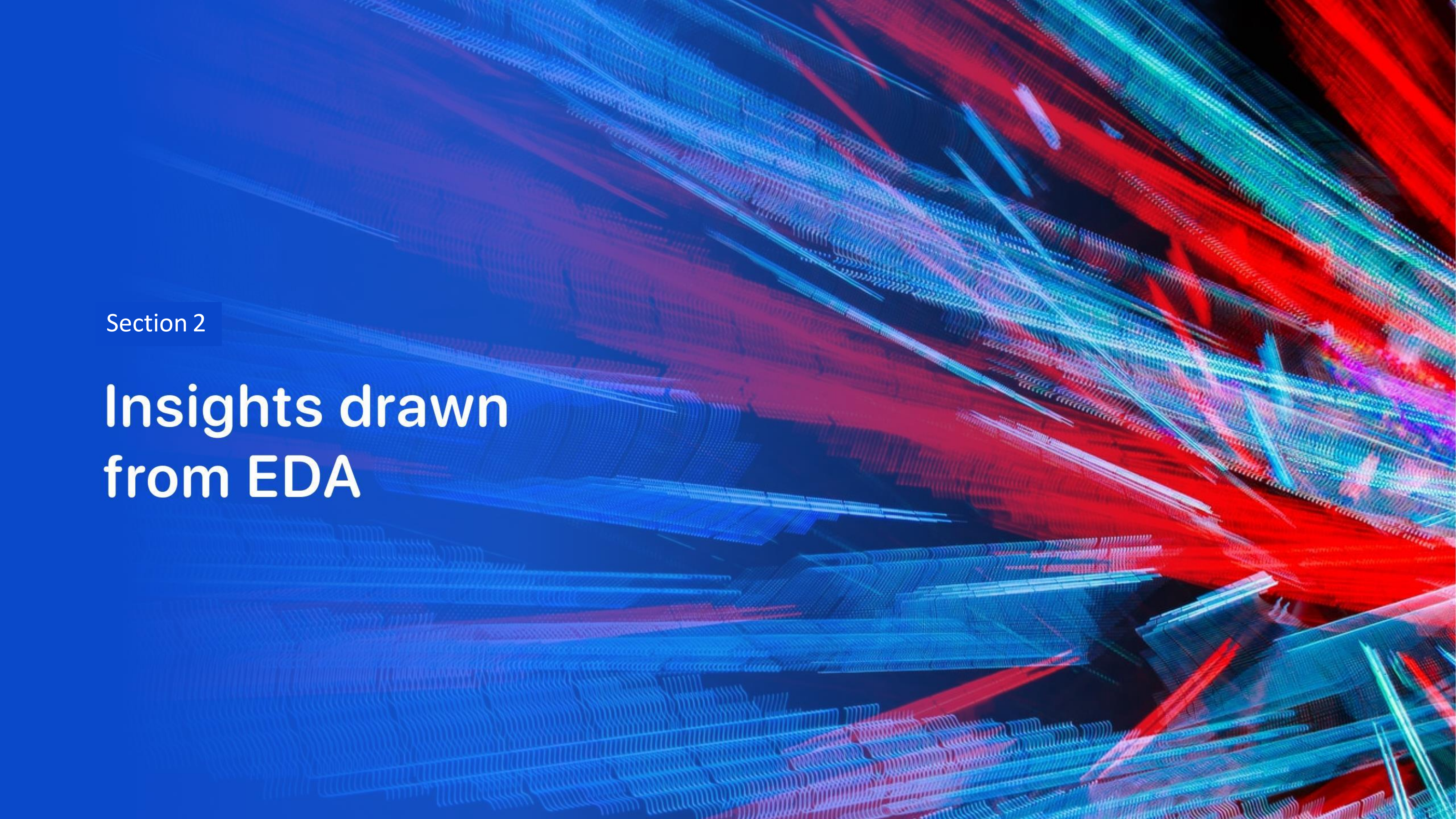
Interactive analytics results

- All launch sites were built in seashore area, and far away from the city, highway, and railway
- This is to ensure the safety of the citizens, and avoid causing further damage when the landing is unsuccessful
- KSC LC-39A has the highest success rate (41.2%), CCAFS LC-40 has the lowest success rate (14.4%)

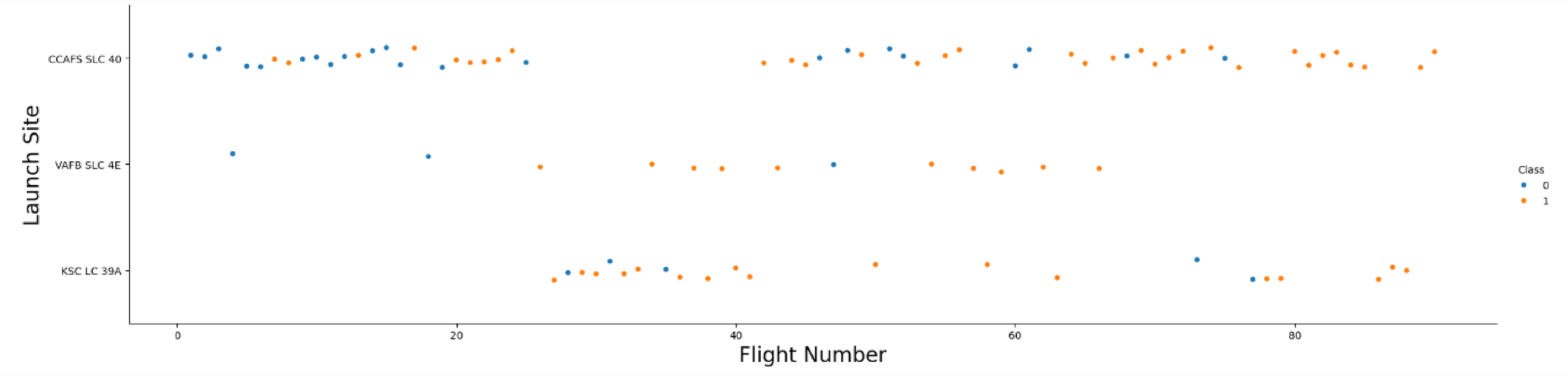
- **Predictive analysis results**
- The Decision Tree Classifier shows the highest accuracy as it shows the highest value of true positive and true negative values

Section 2

Insights drawn from EDA



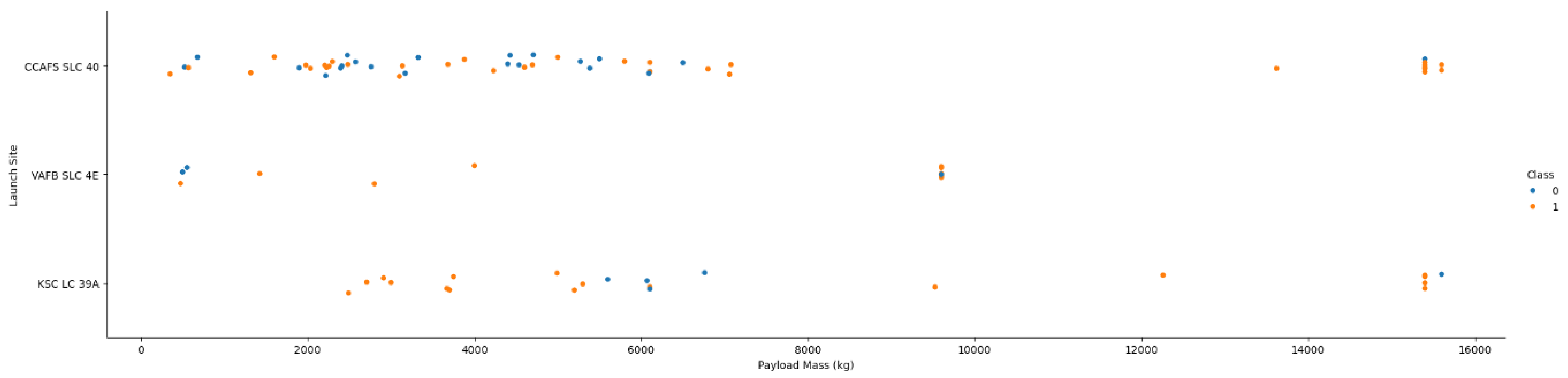
Flight Number vs. Launch Site



Explanation:

- **Blue dots** represents the landing is **not successful**, **Orange dots** represents **successful landing**
- The higher the flight number, the later the flight is launched
- Results show the higher the flight number, the higher the success rate (More orange dots appear)
- **CCAFS SLC 40 has the highest success rate, compared to KSC LC 39A and VAFB SLC 4E**
- It infers the success landing rate is improved over time

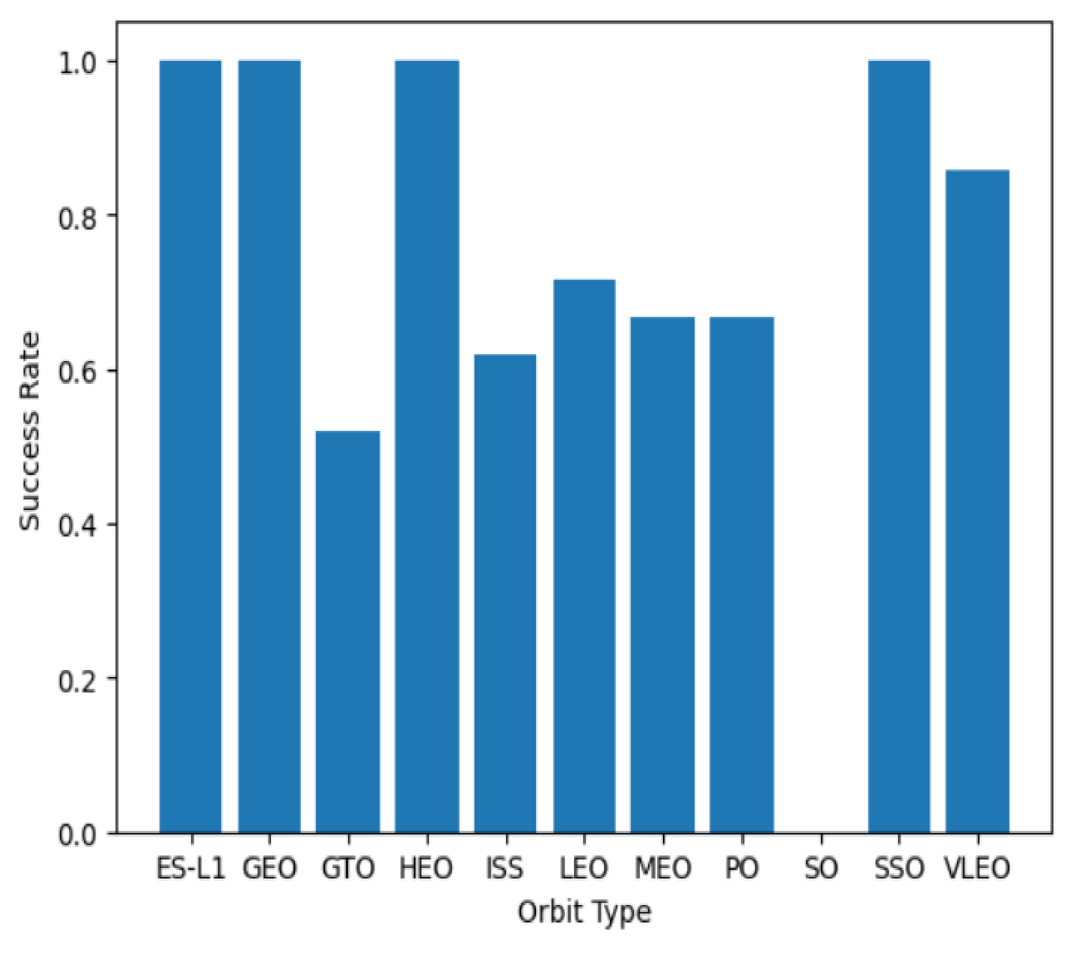
Payload vs. Launch Site



Explanation:

- **Results show the higher the payload mass, the higher the success rate**
- The landing is **almost successful** when the payload mass is **more than 7000 kg**.
- The maximum payload mass for VAFB SLC 4E is about 9000 kg, while for CCAFS SLC 40 and KSC LC 35A, the maximum payload mass can reach to around 15000 kg.

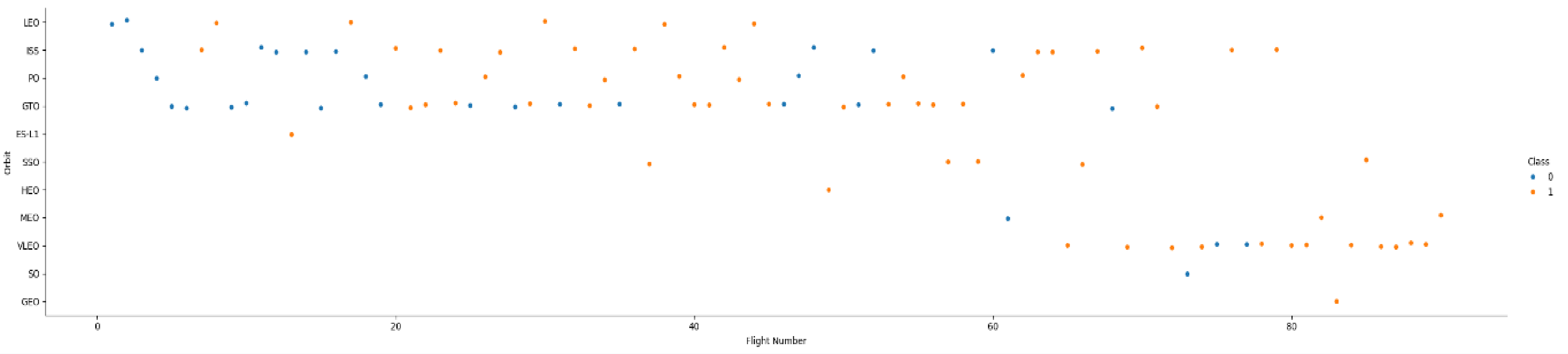
Success Rate vs. Orbit Type



Explanation:

- **100% success rate: ES-L1, GEO, HEO, and SSO**
- >80% success rate: VLEO
- Lowest success rate (0%): SO

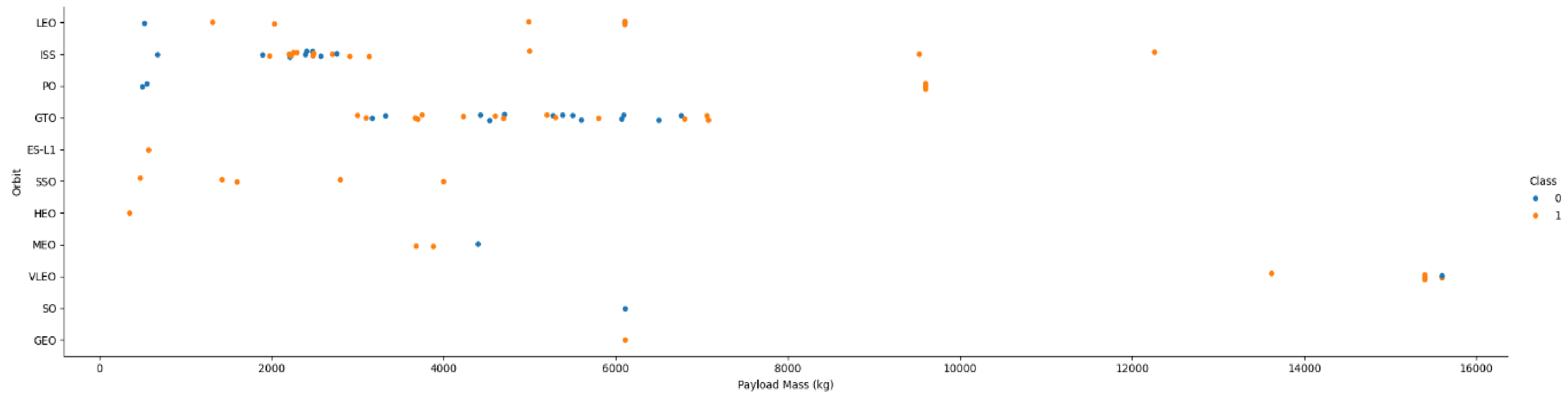
Flight Number vs. Orbit Type



Explanation:

- **Some orbits has higher success rate, but the total flight number in those orbits is less**
- It could be observed that the later the flight, the success rate is improved
- VLEO has a higher frequency in late stage, with high rate of success

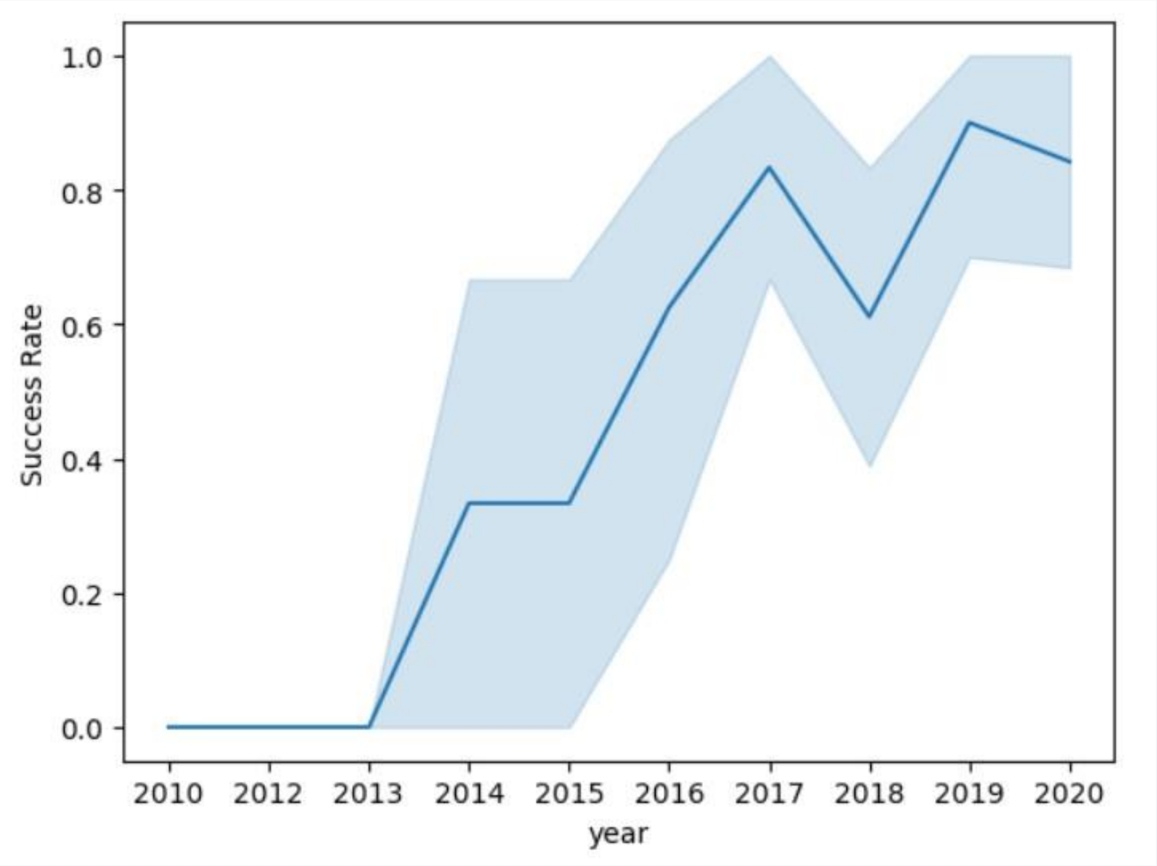
Payload vs. Orbit Type



Explanation:

- **The payload does not show significant difference compared to orbit type**
- The higher payload mass brings higher success rate in LEO and ISS

Launch Success Yearly Trend



Explanation:

- **The success rate shows improvement throughout the year starting from 2013 to 2020**
- The success rate from 2010 to 2013 remains 0%, it could be inferred that the launching is still in development

All Launch Site Names

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Query:
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;

Launch Site Names Begin with 'CCA'

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

Query:

%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

****Only first 5 rows are shown in this query, to show the Launch Site ‘CCAFS-LC 40’**

Total Payload Mass

Total_Payload_Mass_kg
45596

Query:

```
%sql SELECT SUM(PAYLOAD_MASS___KG_) AS Total_Payload_Mass_kg FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

Query:

```
%sql SELECT AVG(PAYLOAD_MASS___KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

MIN(DATE)
2015-12-22

Query:

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Query:

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

Total Number of Successful and Failure Mission Outcomes

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

Query:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

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

Query:

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

2015 Launch Records

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Query:

```
%sql SELECT substr(Date,6,2) AS month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
WHERE [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015’;
```

substr(Date,6,2) AS month → This extracts a substring from the "Date" column starting at the 6th character and taking 2 characters. It is used to get the month part of the date.
substr(Date,0,5) ='2015’ → This extracts a substring from the "Date" column starting at the 0th character (the beginning) and taking 5 characters. It is used to compare the first 5 characters of the date to check if they are equal to '2015'. This condition filters the results to include only those records where the year in the "Date" column is '2015’.

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

Landing_Outcome	**count_outcomes**
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

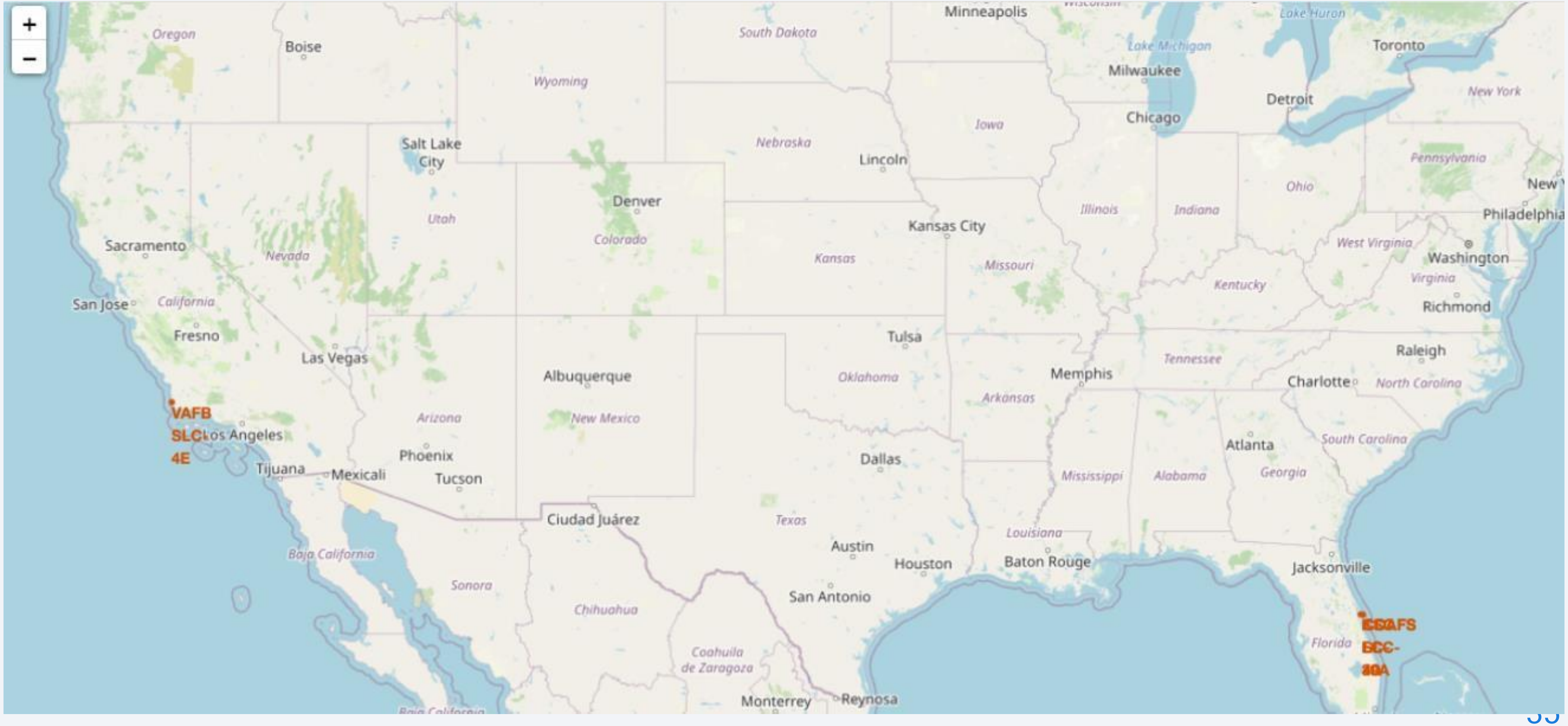
Query:

```
%sql SELECT Landing_Outcome, COUNT(*) as count_outcomes FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY count_outcomes DESC;
```

Section 3

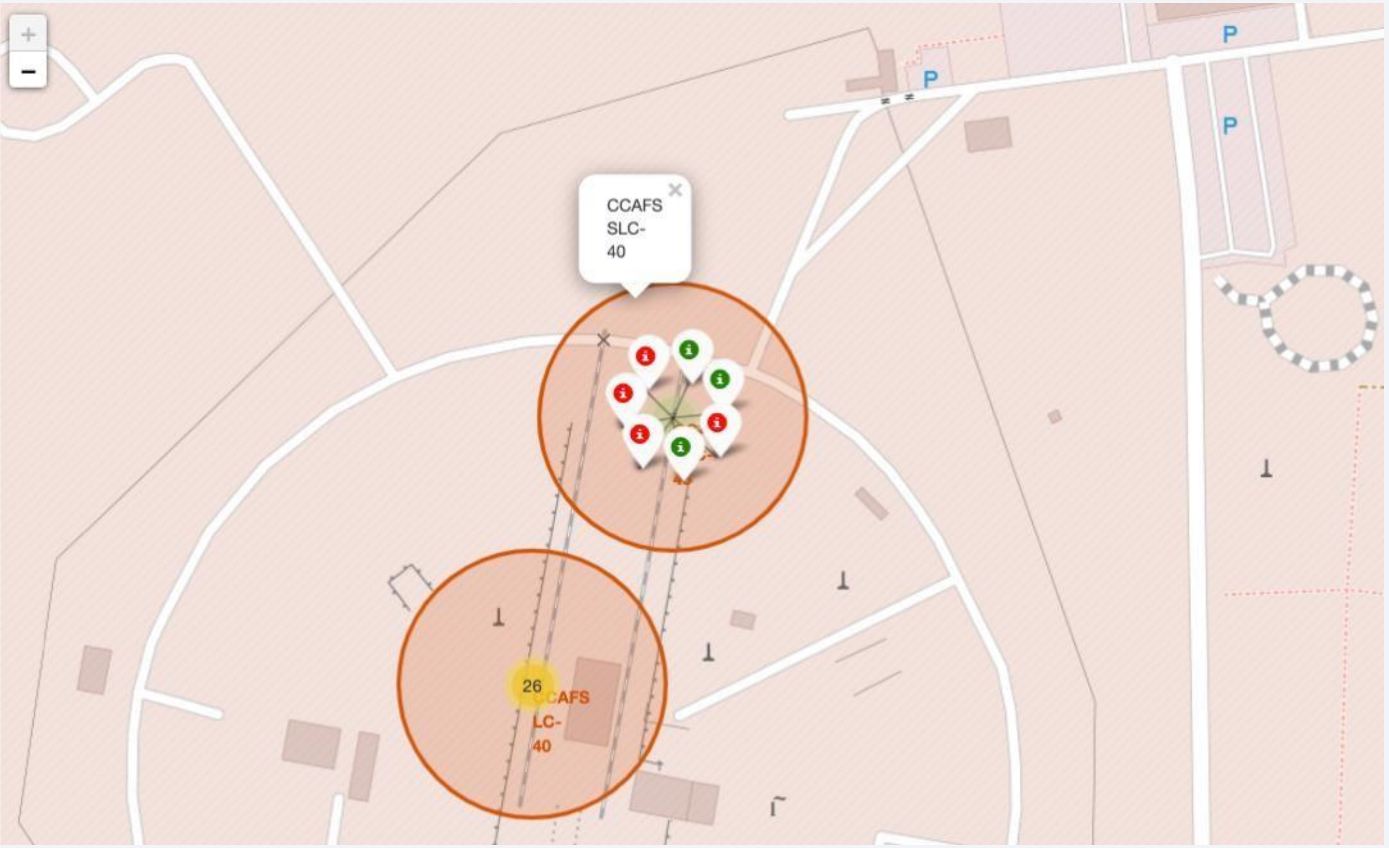
Launch Sites Proximities Analysis

Launch Sites



- All launch sites were in seashore area → for plannings if the landing is not successful

Launch Outcomes by Site

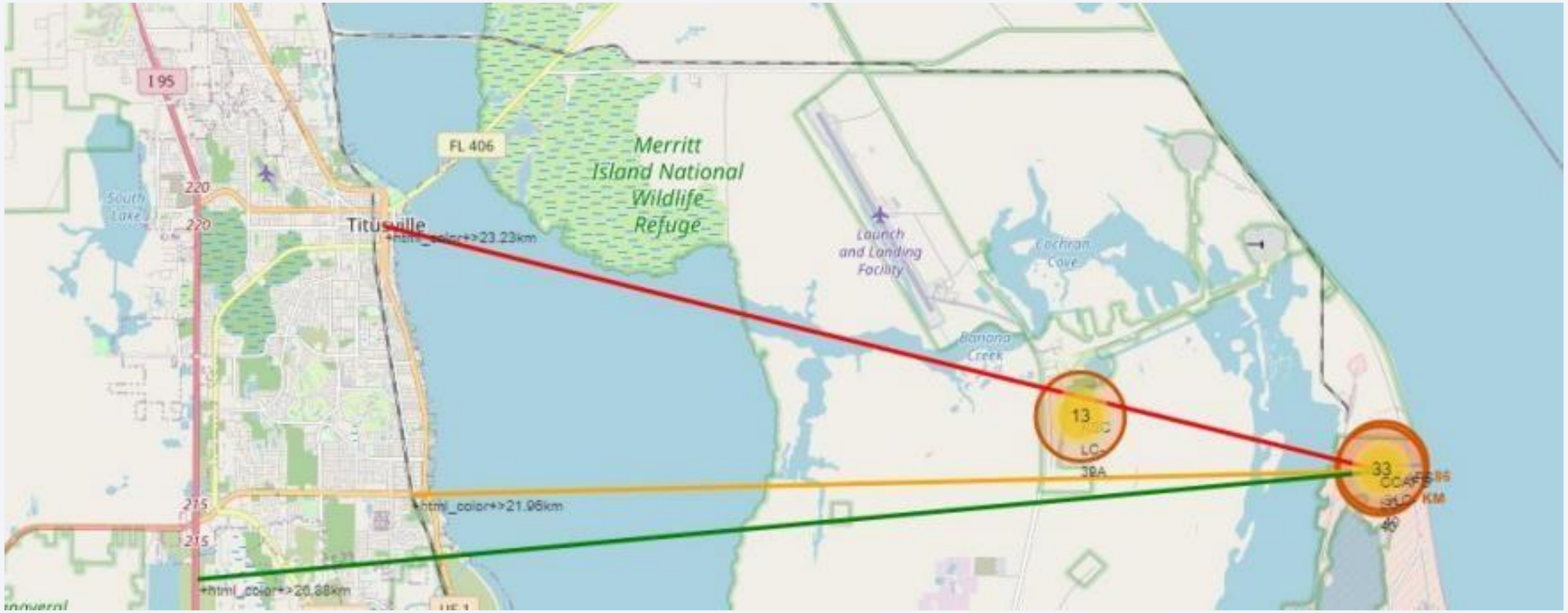


- Showing one of the marker clusters in CCAFS SLC-20, it could be observed that the success rate of this site is 3 per 7.

- Another site could be analysed for the outcome through click the centre of the site, e.g. 26.

Distance to Proximities

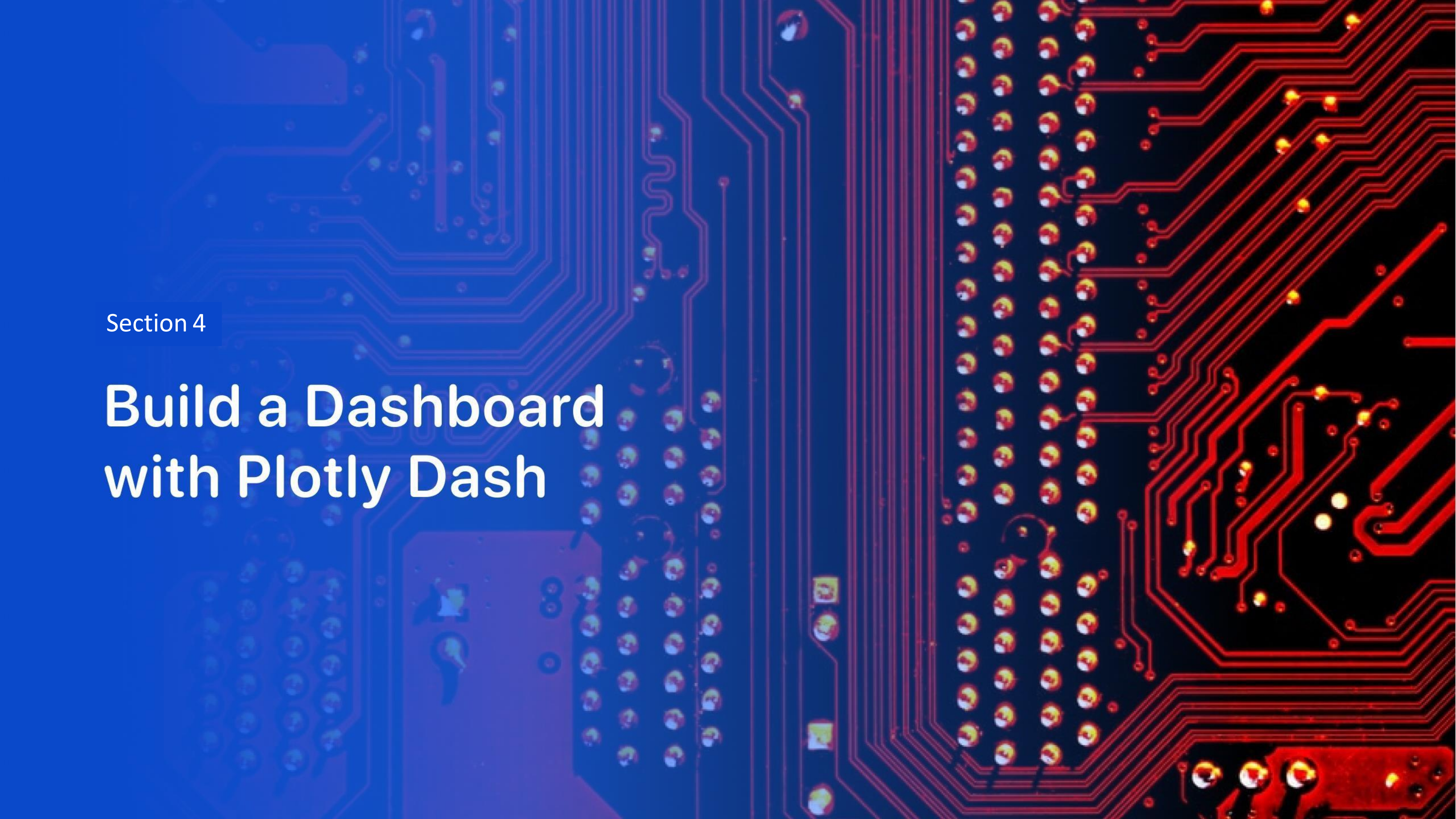
Example: CCAFS SLC-20



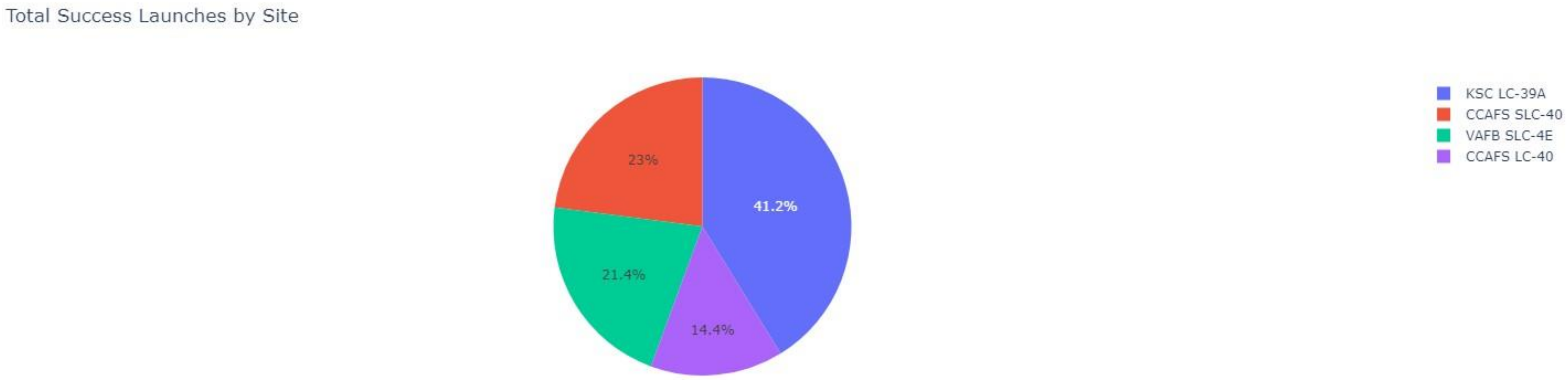
- It is built very near to the coastline, while far away from the city, highway, and railway
- To ensure the safety of the citizens, and avoid causing further damage when the landing is unsuccessful

Section 4

Build a Dashboard with Plotly Dash



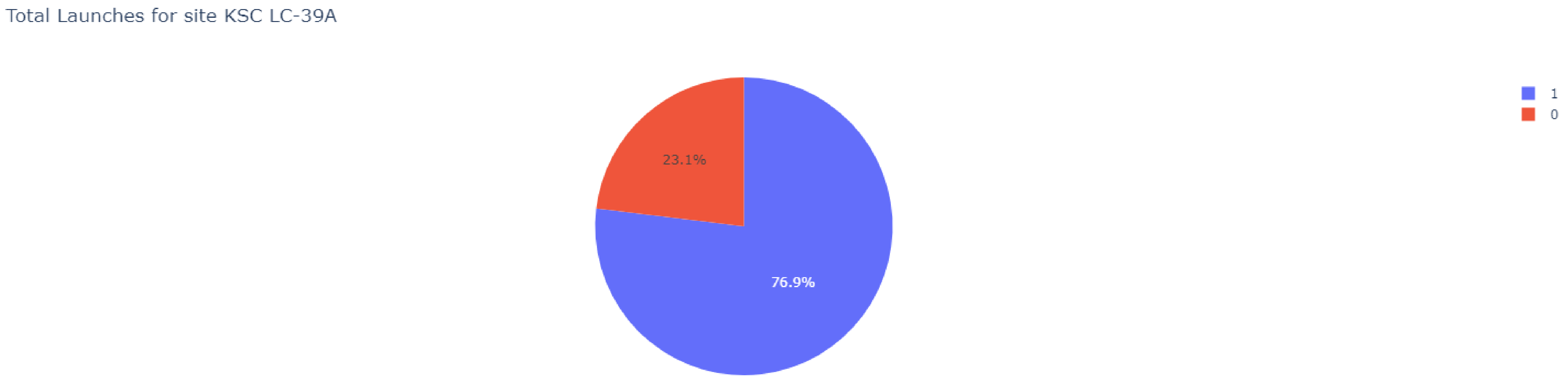
Total Success Launches by Site



By comparing the total success launches by site,

- **KSC LC-39A has the highest success rate (41.2%)**
- **CCAFS LC-40 has the lowest success rate (14.4%)**

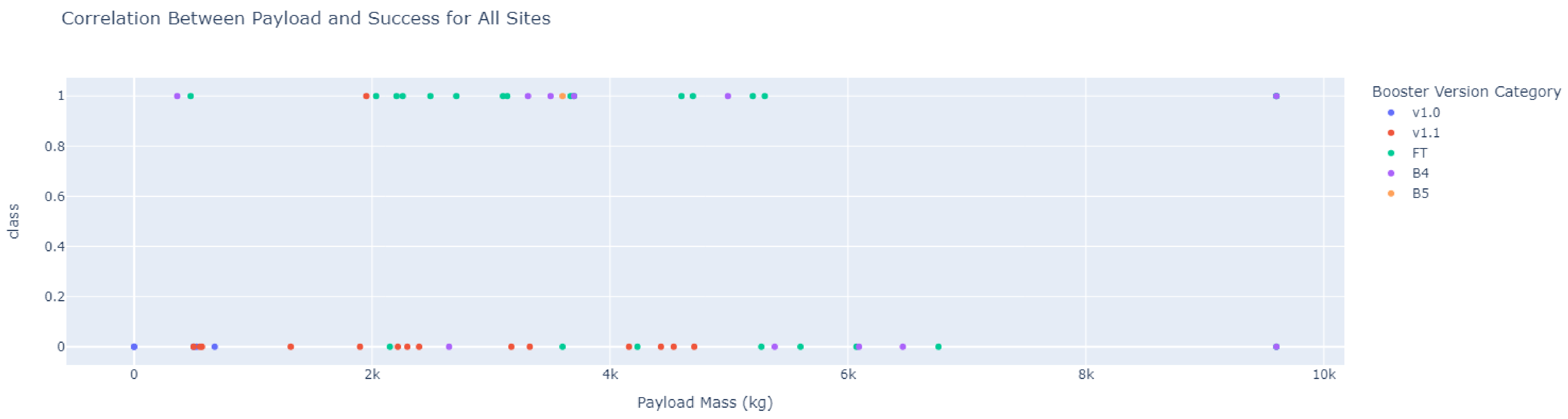
Launch Success Analysis for KSC LC-39A



In total success launches by site KSC LC-39A,

- **The success rate reaches 76.9%**

Payload vs Success Rate

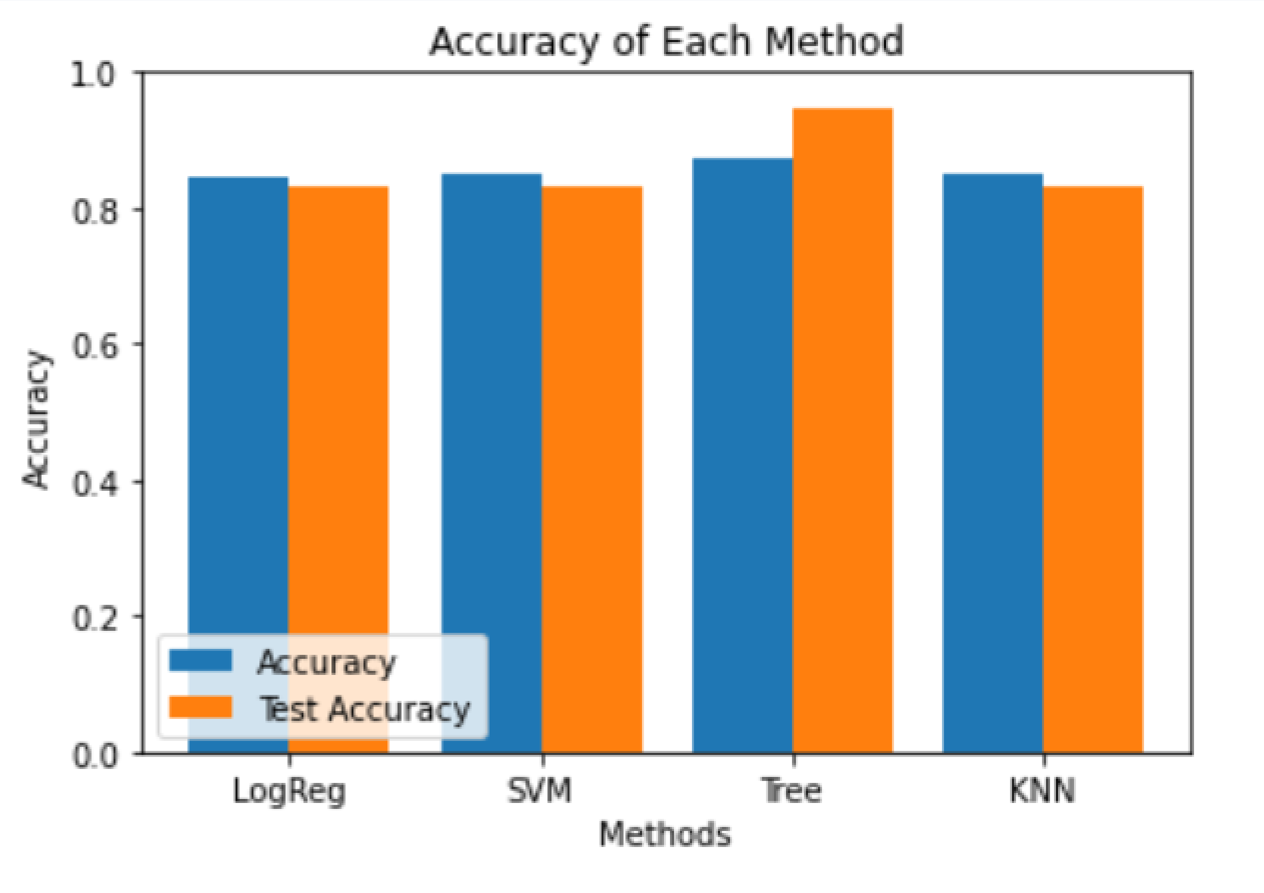


- In the graph, 1 = success, 0 = not success
- **In successful launches,** the payload mass is normally between 2000kg and 6000kg
- **In successful launches,** FT Booster has the highest prevalence, showing the success rate is the highest when FT Booster is used

Section 5

Predictive Analysis (Classification)

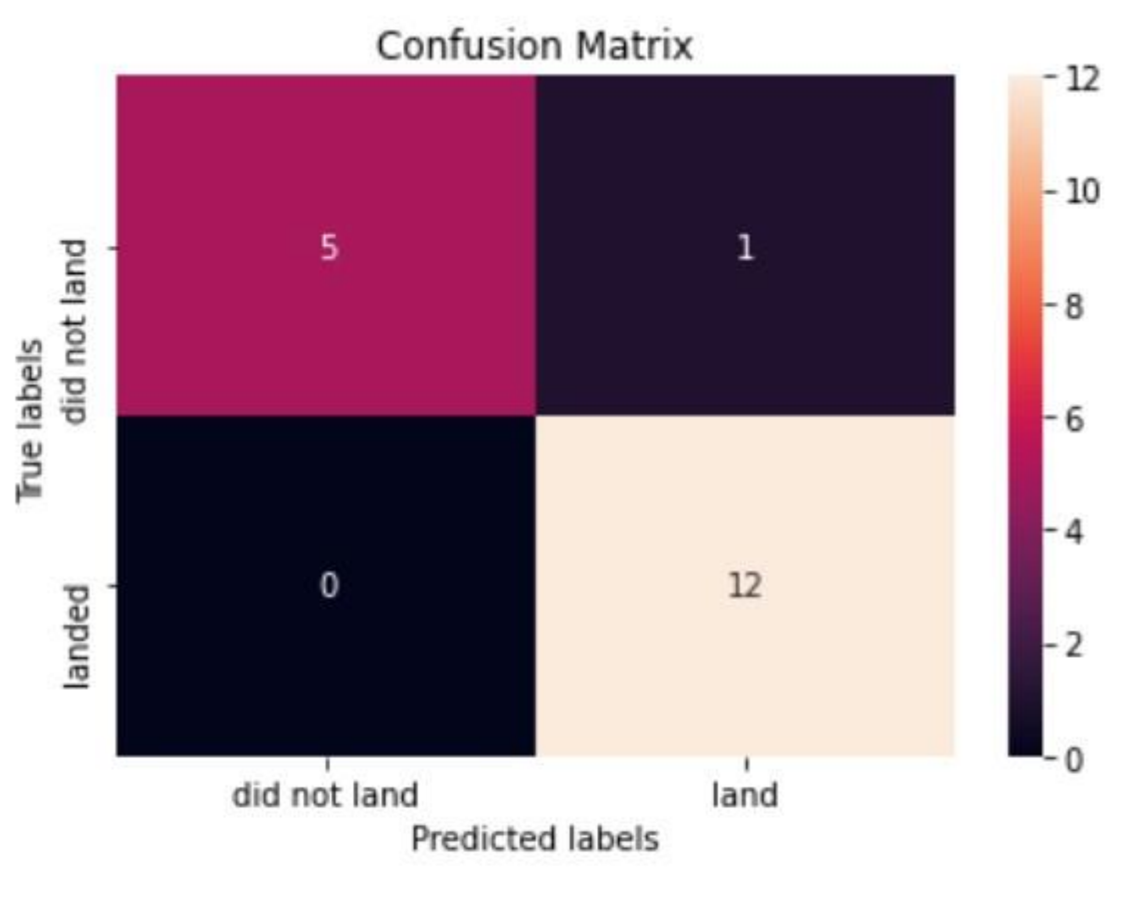
Classification Accuracy



- Through the accuracy of the models, the decision tree model has the highest values among others.
- The graph is plotted based on the table below, by showing accuracy score using each ML method:

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	94.444444
3	Decision Tree	83.333333

Confusion Matrix



Example: Decision Tree Classifier

- The Decision Tree Classifier shows the highest accuracy as it shows the highest value of true positive and true negative values.
- The false positive (Type I) is 1, while the false negative (Type II) is 0.

Conclusions

- The analysis reveals that CCAFS SLC 40 boasts the highest success rate among launch sites, and there is a positive correlation between payload mass and mission success.
- Certain orbits consistently achieve a 100% success rate, but some high-performing orbits have fewer overall flights.
- Over the years, there is a noticeable improvement in success rates. Interactive analytics highlight strategic launch site locations for safety, with KSC LC-39A having the highest success rate and CCAFS LC-40 the lowest.
- The Decision Tree Classifier stands out in predictive analysis, demonstrating the highest accuracy.
- Overall, the findings provide insights into the factors influencing space mission success, encompassing launch site performance, payload characteristics, and orbital considerations.

Appendix

- GitHub Project Repository!

https://github.com/rhuang7/DataScience_IBM

Thank you!