IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Caleb Hudson
06/03/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with SQL
  - EDA with Visualization
  - Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - EDA Results
  - Predictive Analysis

Introduction

Project Background

- To help SpaceY be a competitor with SpaceX

Problems you want to find answers

- Determine the best chances of successfully landing the first stage rocket booster

Section 1

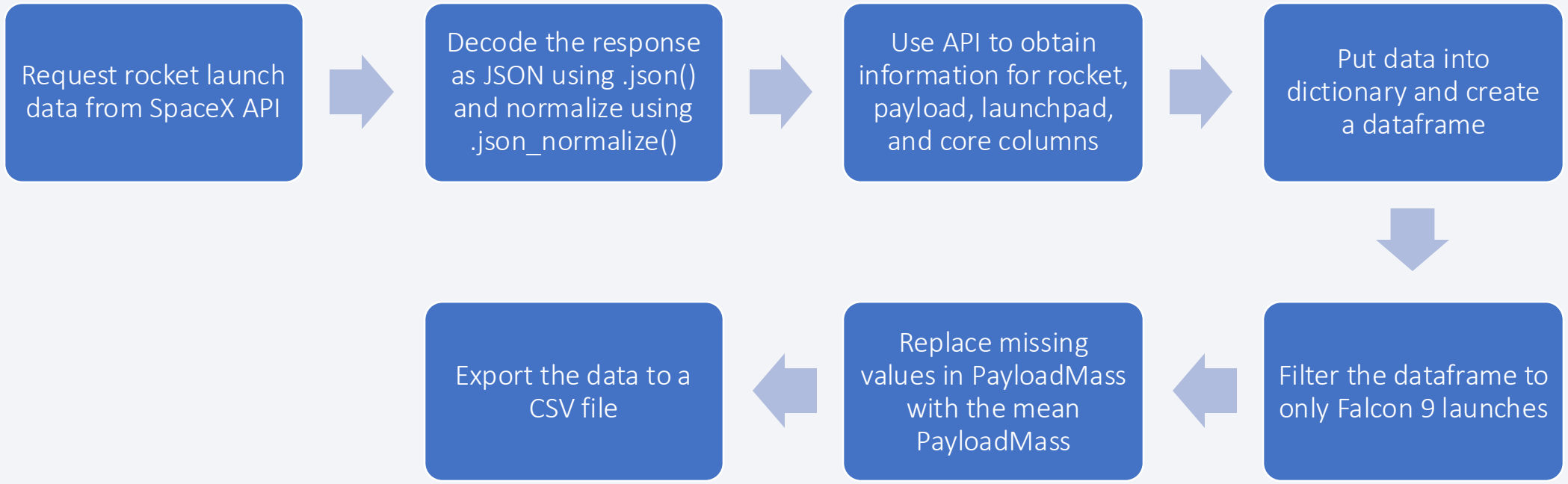
Methodology

Methodology

Executive Summary

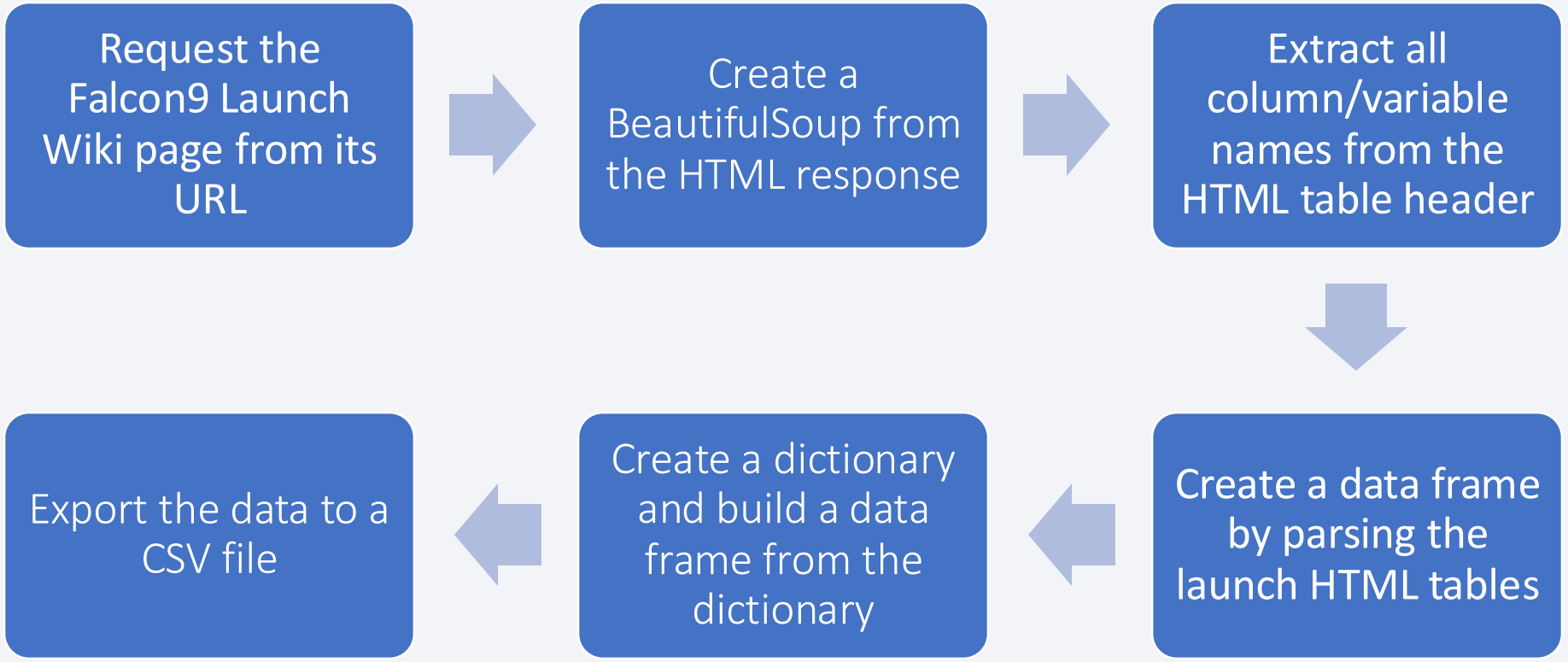
- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from SpaceX Wikipedia page
- Perform data wrangling
  - Data was filtered, missing values were dealt with, and One Hot Encoding was used to prepare binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was normalized, divided into training and testing sets, and run through four classification models

Data Collection – SpaceX API



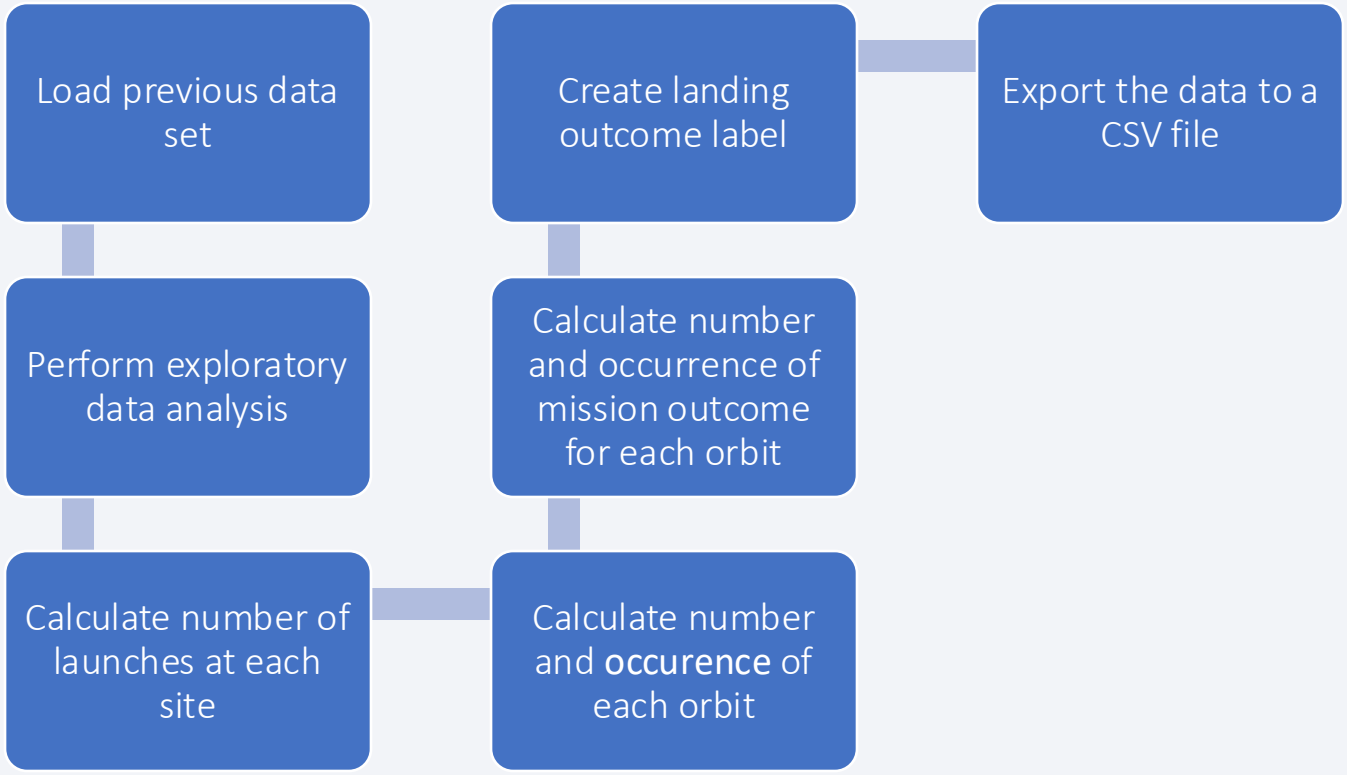
GitHub URL: Data Collection API

Data Collection - Scraping



- GitHub URL: SpaceX Webscrapping

Data Wrangling



Github URL: SpaceX Data Wrangling

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Success rate of each orbit type, Flight Number vs Orbit Type, Payload Mass vs. Orbit Type, and Yearly Success Rate
  - Scatter plots were used to help show potential relationship between two variables.
  - A bar chart was used to help compare categories.
  - A line graph was used to compare success rate over time.

GitHub URL: EDA with Data Visualization

EDA with SQL

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first succesful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List all the booster_versions that have carried the maximum payload mass. Use a subquery.
9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
10. 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.

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Markers for launch sites
  - Added circle markers with text and popup label for each launch site at the latitude and longitude of each site.
- Colored markers for launch outcomes
  - Added clustered colored markers to indicate success or failure of the booster landing for each launch site.
- Distance lines
  - Added colored lines to mark the distance to the closest highway, railway, and coastline to a launch site.

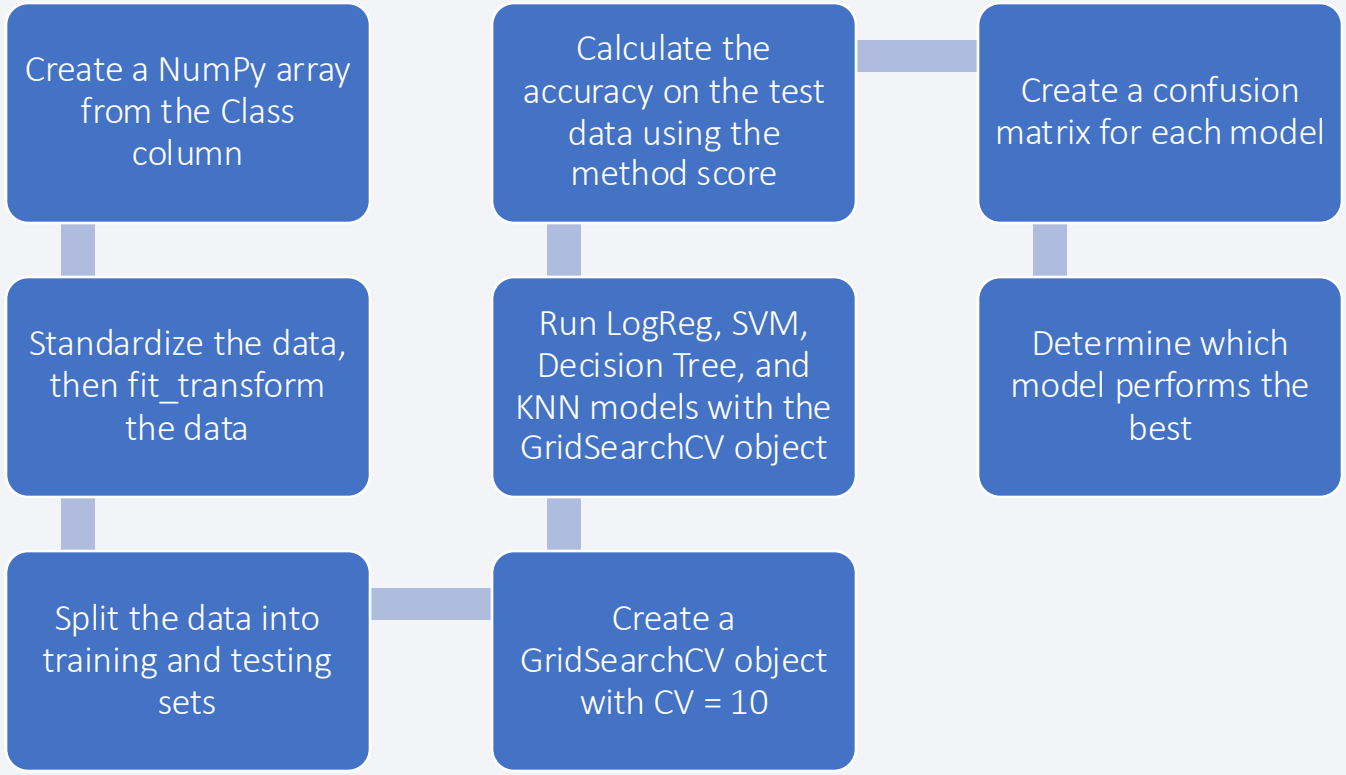
GitHub URL: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Dropdown List
  - Added a dropdown list to allow selection of launch sites.
- Pie Chart
  - Added a pie chart to show successful launches from all launch sites. If a specific lauch site is selected the pie chart will update to show success and failed launches for that launch site.
- Slider
  - Added a slider to for payload mass range.
- Scatter Chart
  - Added a scatter chart to show the correlation between payload mass and launch success

GitHub URL: Dashboard with Plotly Dash

Predictive Analysis (Classification)



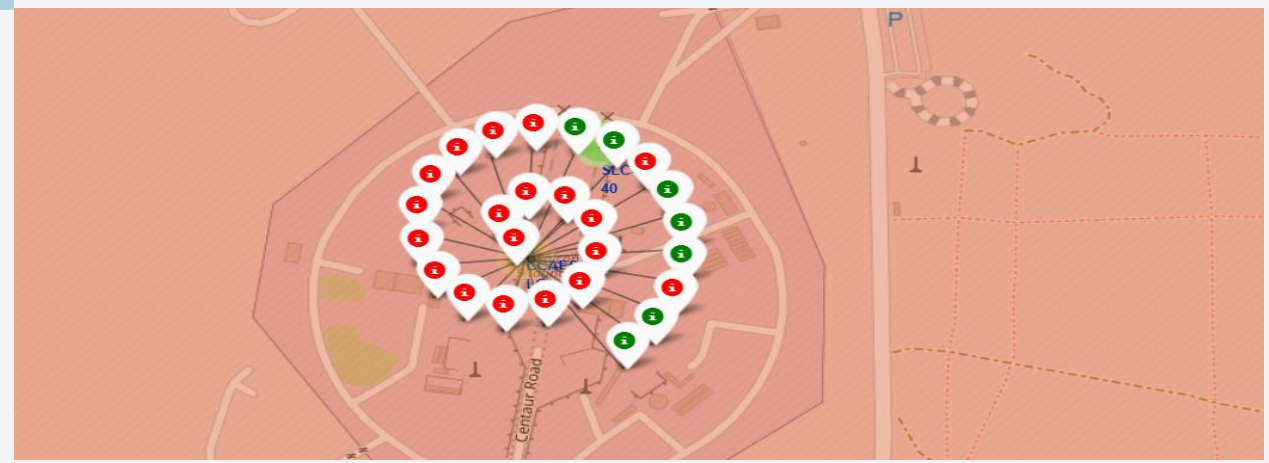
GitHub URL: Prediction Analysis

Results

- Exploratory data analysis results
  - SpaceX uses four launch sites
    - KSC LC-39A launch site has the highest rate of successful landings
  - The first successful landing occurred in 2015
  - Four different booster versions have successfully landed on drone ships
  - Only 1 out of 101 missions was a total failure
  - The success rate from 2013 through 2020 increased

Results

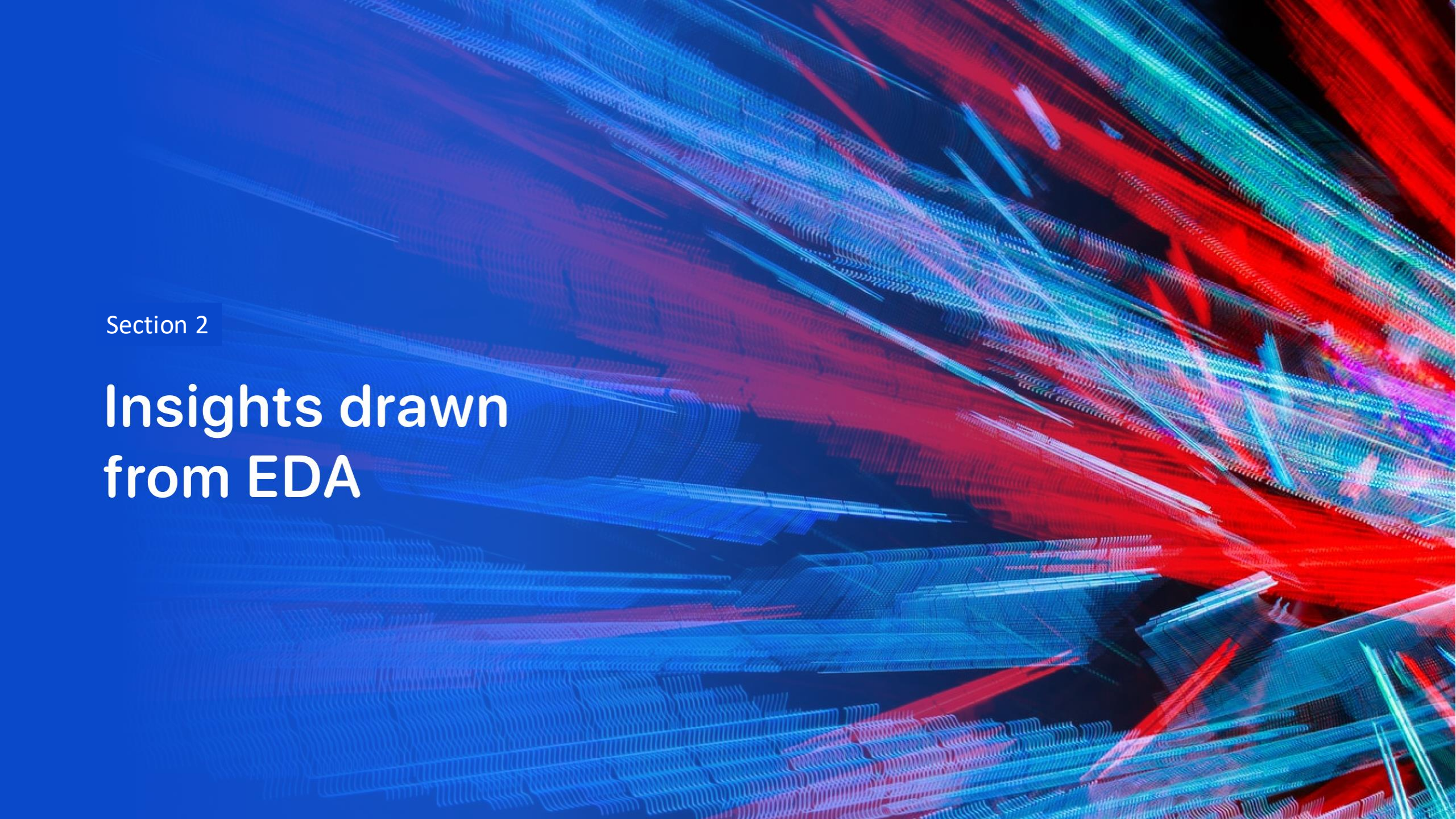
- Interactive analytics demo in screenshots



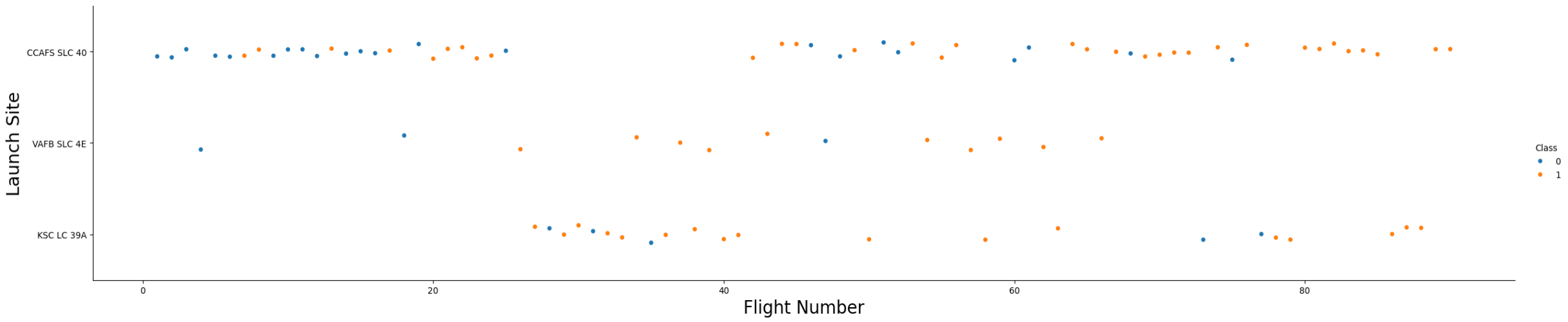
Results

- Predictive analysis results
  - The Decision Tree method provides the highest accuracy

```
The accuracy for Logistic Regression method: 0.8333333333333334
The accuracy for Support Vector Machine method: 0.8333333333333334
The accuracy for Decision Tree method: 0.8888888888888888
The accuracy for K Nearest Neighbors method: 0.8333333333333334
```

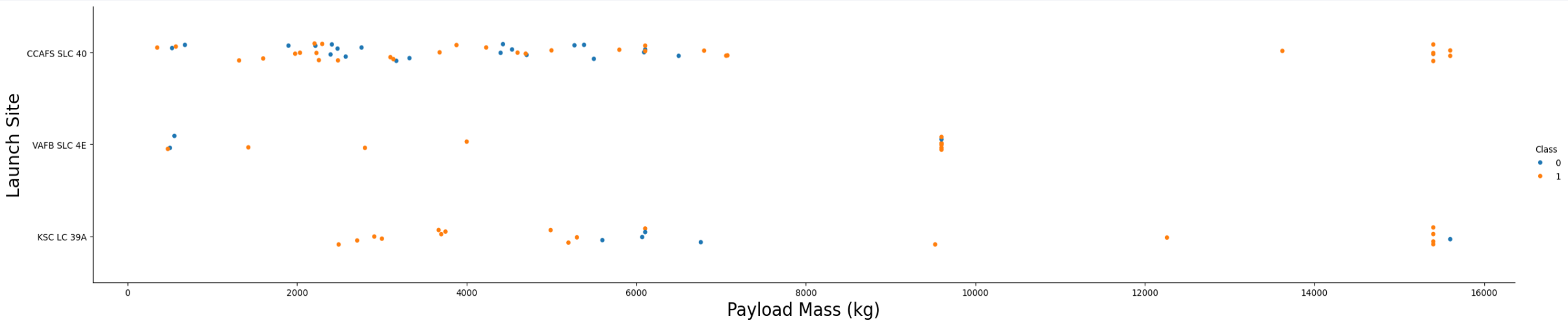

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation:
  - CCAPS SLC 40 has had the most overall launches
  - KSC LC 39A has the highest rate of success
  - The first six attempts were unsuccessful
  - The most recent 13 attempts were successful

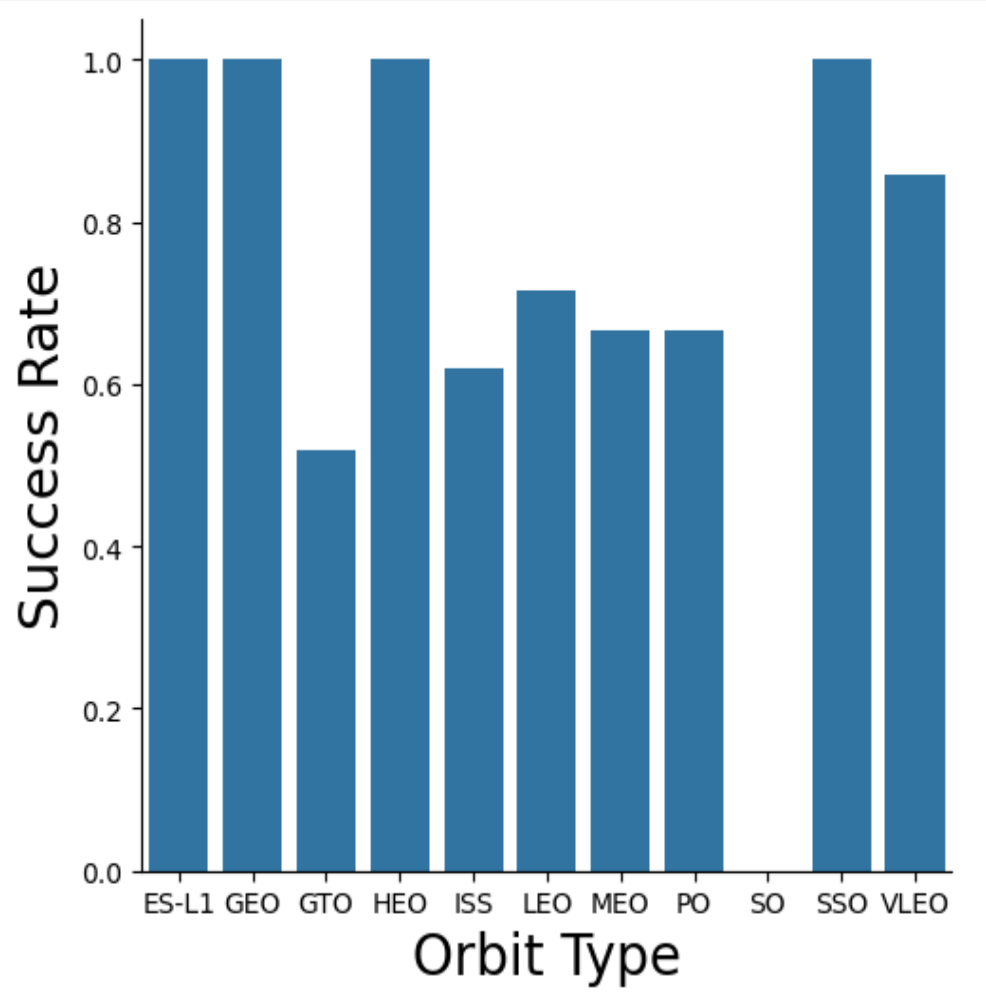
Payload vs. Launch Site



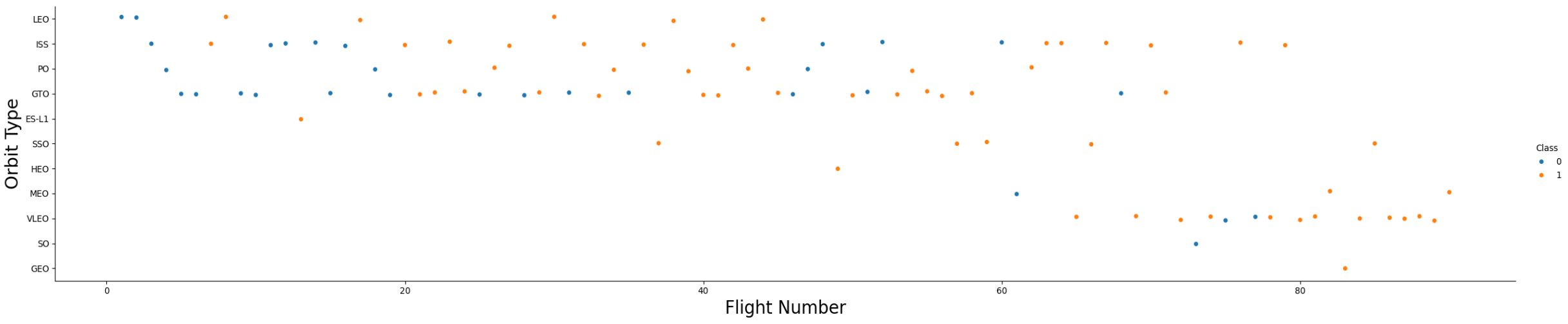
- Explanation:
  - For each launch site higher payload mass launches have a higher success rate
  - VAFB SLC 4E appeas to only be able to launch payloads under 10,000 kg
  - KSC LC 39A has a 100% success rate for payloads less than 5,500 kg

Success Rate vs. Orbit Type

- Explanation:
  - ES-L1, GEO, HEO, and SSO have a 100% success rate
  - SO has had the lowest success rate at 0%
  - All other orbits are between 50-85% success rate

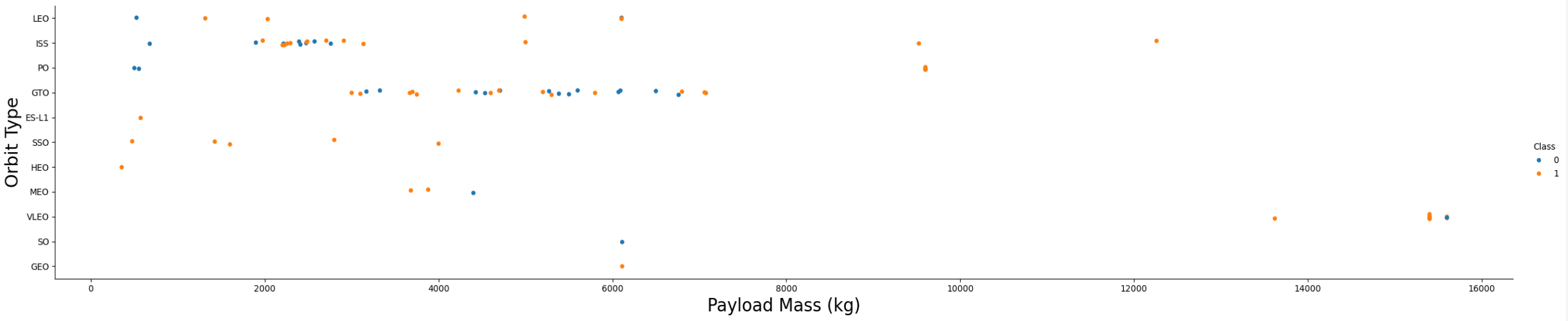


Flight Number vs. Orbit Type



- Explanation:
  - GTO has had the most attempts
  - SO has a 0% success rate but there has only been one attempt
  - VLEO has had the more attempts recently with a high success rate
  - ES-L1, SSO, HEO, MEO, and GEO have all had 5 or fewer attempts

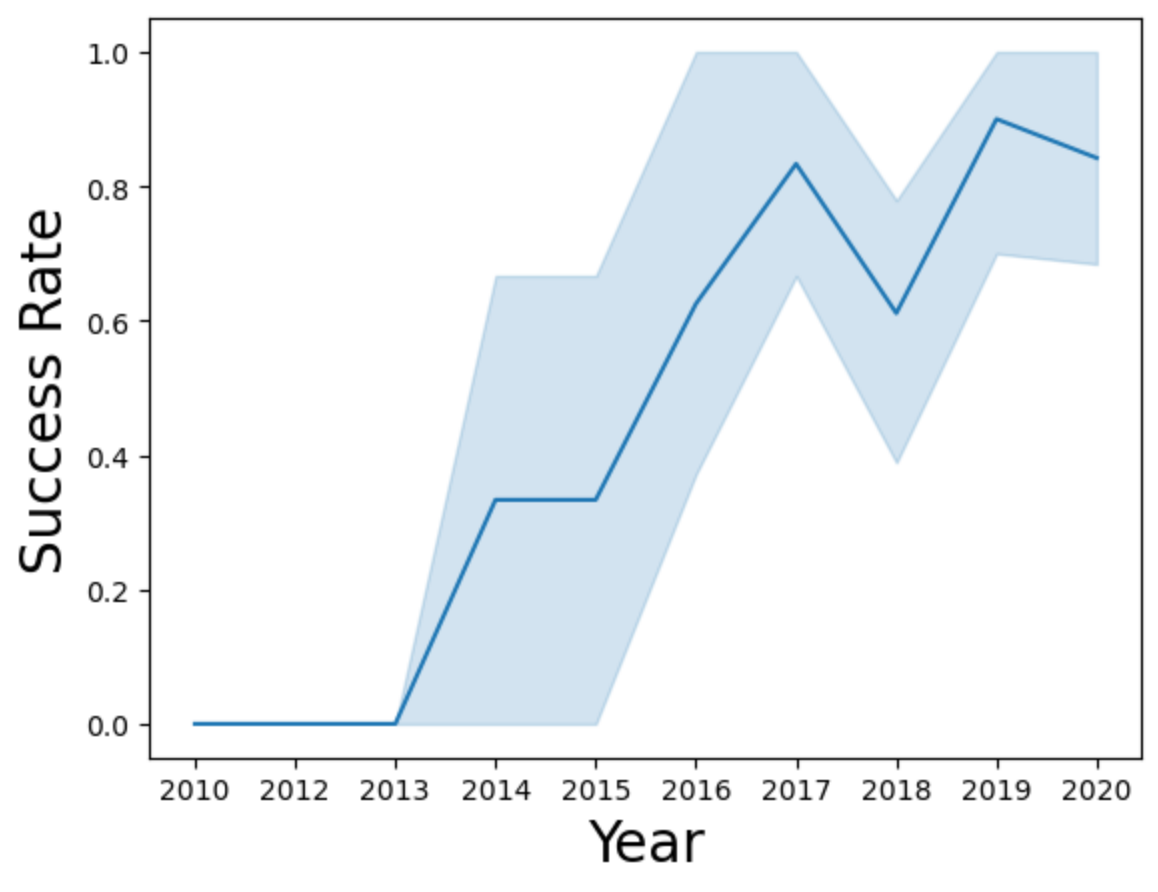
Payload vs. Orbit Type



- Explanation:
  - ISS has the largest range of payload mass
    - Heavier payloads still have success to ISS orbit
  - Payload mass doesn't seem to affect success to GTO orbit

Launch Success Yearly Trend

- Explanation:
  - Generally success has increased yearly since 2013.
  - Since 2016 the success rate has been at least 60%



All Launch Site Names

- Explanation:
  - Unique launch site names

Display the names of the unique launch sites in the space mission

In [10]:
```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE;
```

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

Out[10]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Explanation:
  - 5 records where launch sites begin with `CCA`

```
In [11]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

- Explanation:
  - Calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER LIKE '%NASA (CRS)%';
```

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

Out[12]:

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- Explanation:
  - Calculate the average payload mass carried by booster version F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Out[13]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Explanation:
  - Find the dates of the first successful landing outcome on ground pad

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [14]:
```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[14]:

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Explanation:
  - List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [15]:
```
%%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.
```

Out[15]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Explanation:
  - Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [23]:
```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Out[23]:

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

Boosters Carried Maximum Payload

- Explanation:
  - List the names of the booster which have carried the maximum payload mass

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

In [26]:
```
%%sql
SELECT DISTINCT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Out[26]:

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

- Explanation:
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [27]:
```
%%sql
SELECT substr(Date,6,2) as month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015';
```

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

Out[27]:

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

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

- Explanation:
  - 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

In [33]:

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS TOTAL FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY TOTAL DESC;
```

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

Out[33]:

Landing_Outcome	TOTAL
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

Section 3

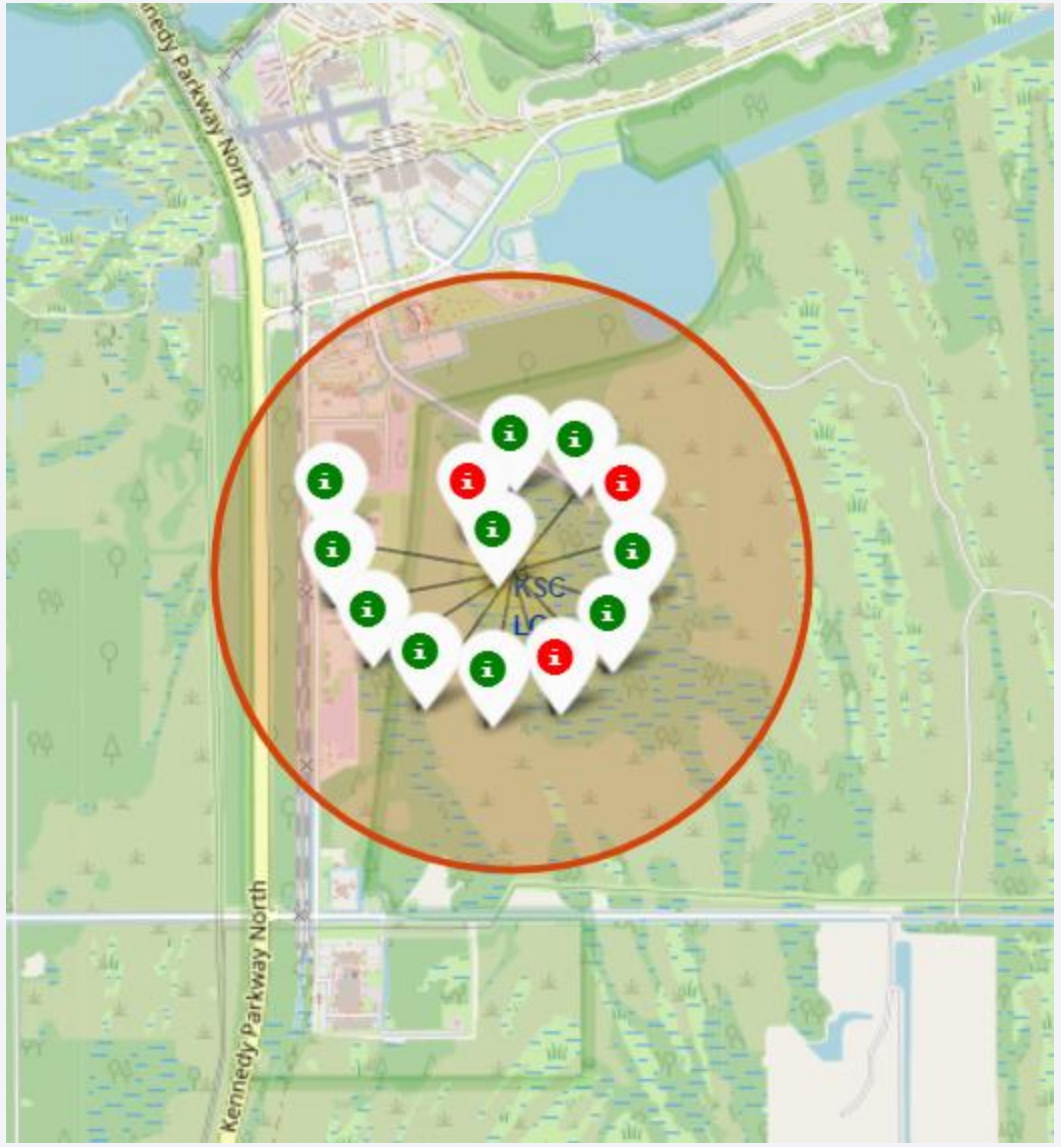
Launch Sites Proximities Analysis

All Launch Sites

- Explanation:
  - All launch sites are located close to coastlines, likely for safety.
  - All launch sites are also in the southern United States. This likely allows launches

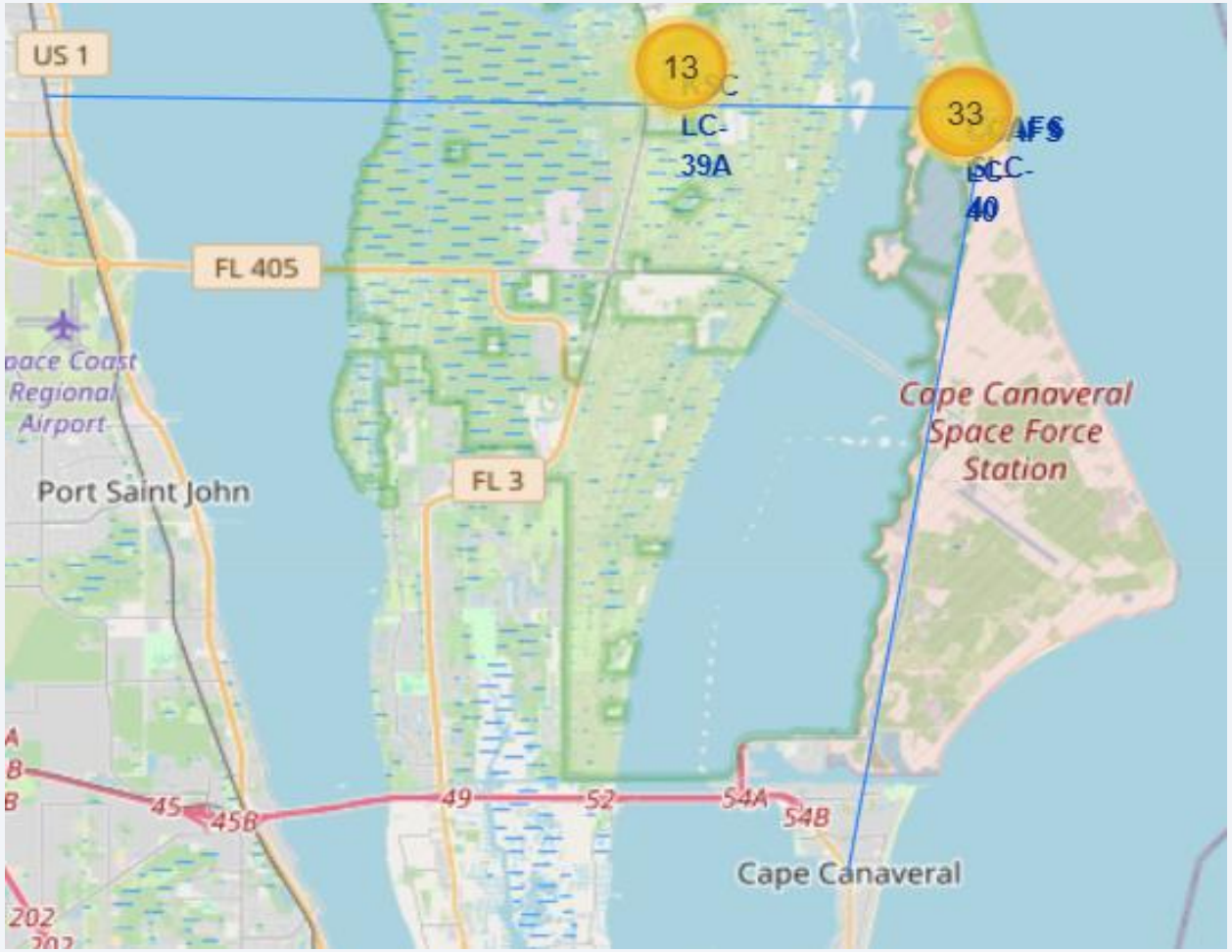
Color Coded Marker Cluster

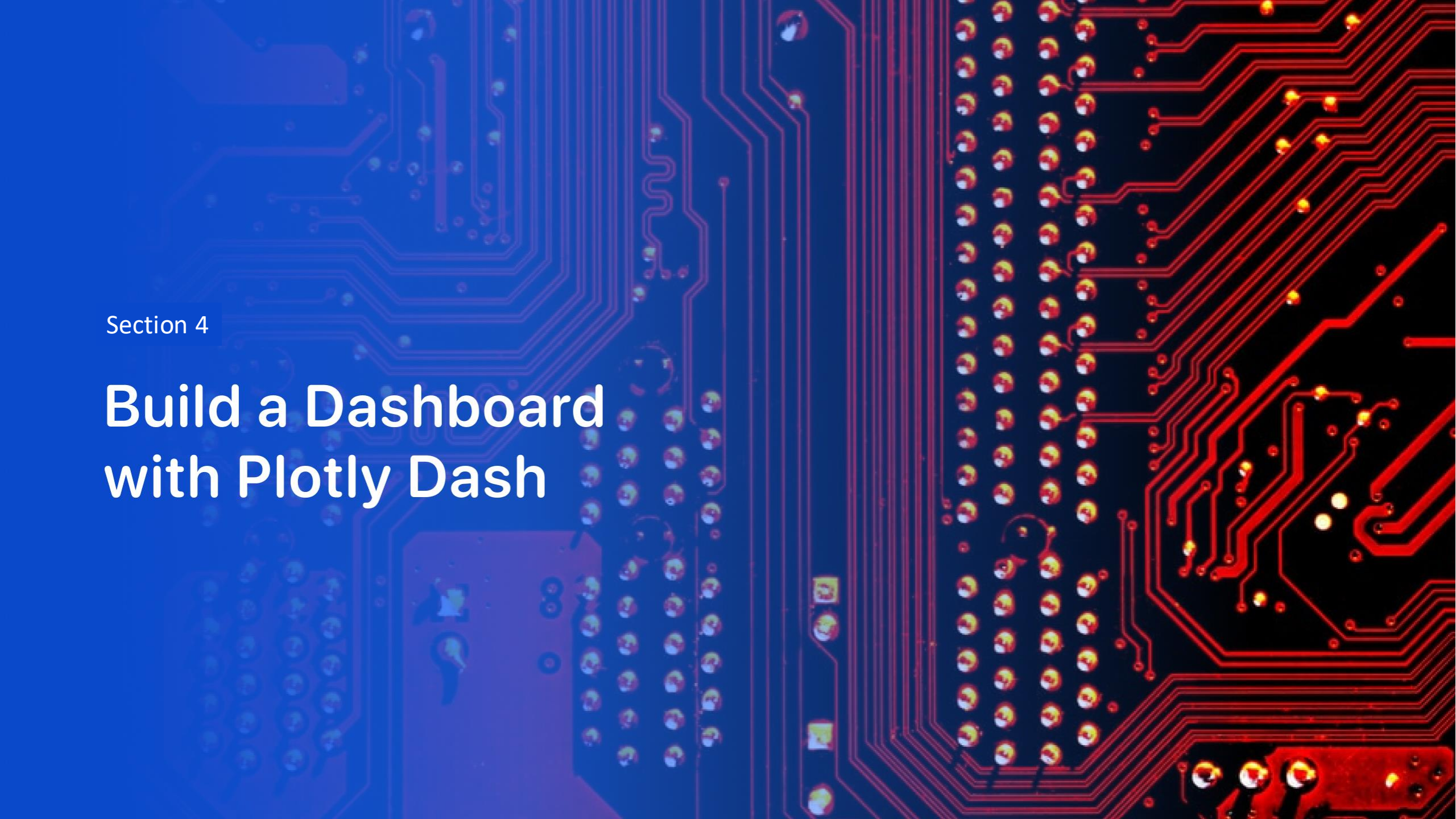
- Explanation:
  - Color coded markers for all launches from launch site KSC LC39A
    - Green = Successful launch
    - Red = Failed launch
  - Overall launches from this site have a high success rate



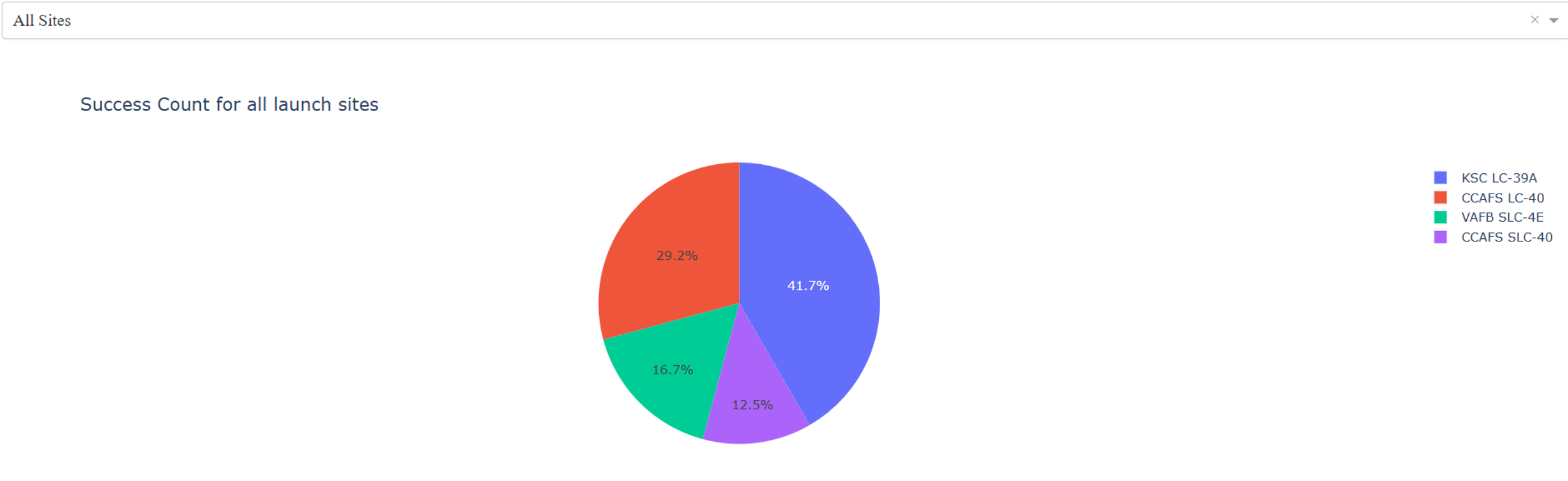
Proximities to Launch Site

- Explanation:
  - From launch site CCAFS SLC-40:
    - Closest city: Cape Canaveral at 19.8 km
    - Closest highway: Kennedy Parkway North at 22.6 km
    - Closest coastline: 0.90 km




Section 4
Build a Dashboard with Plotly Dash

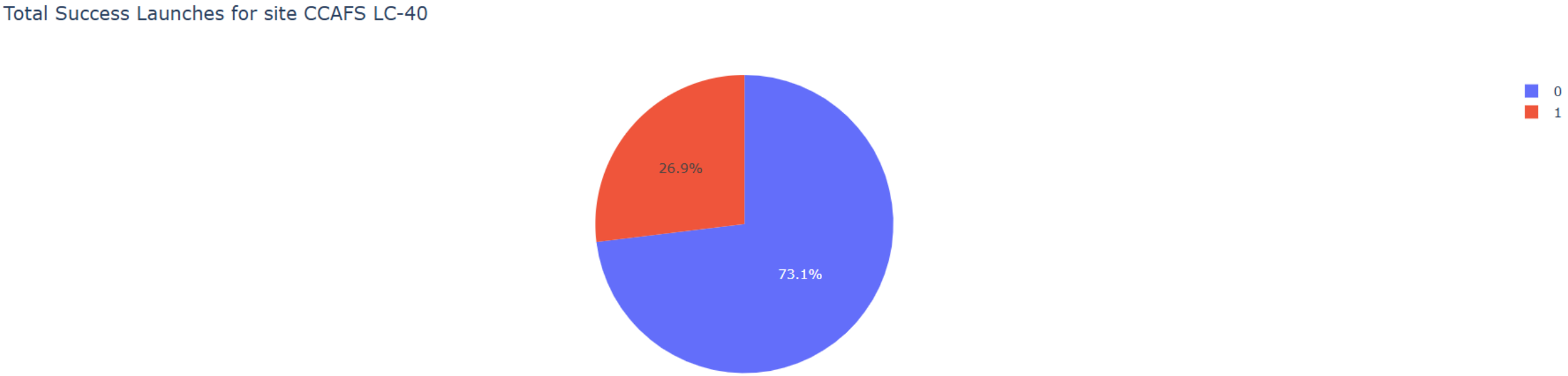
Successful Launches for each Launch Site



- Explanation:
  - Pie chart shows the rate of successful launches at each site

Launch Success Ratio

CCAFS LC-40



- Explanation:
  - Launch success and failure ratio for site CCAFS LC-40

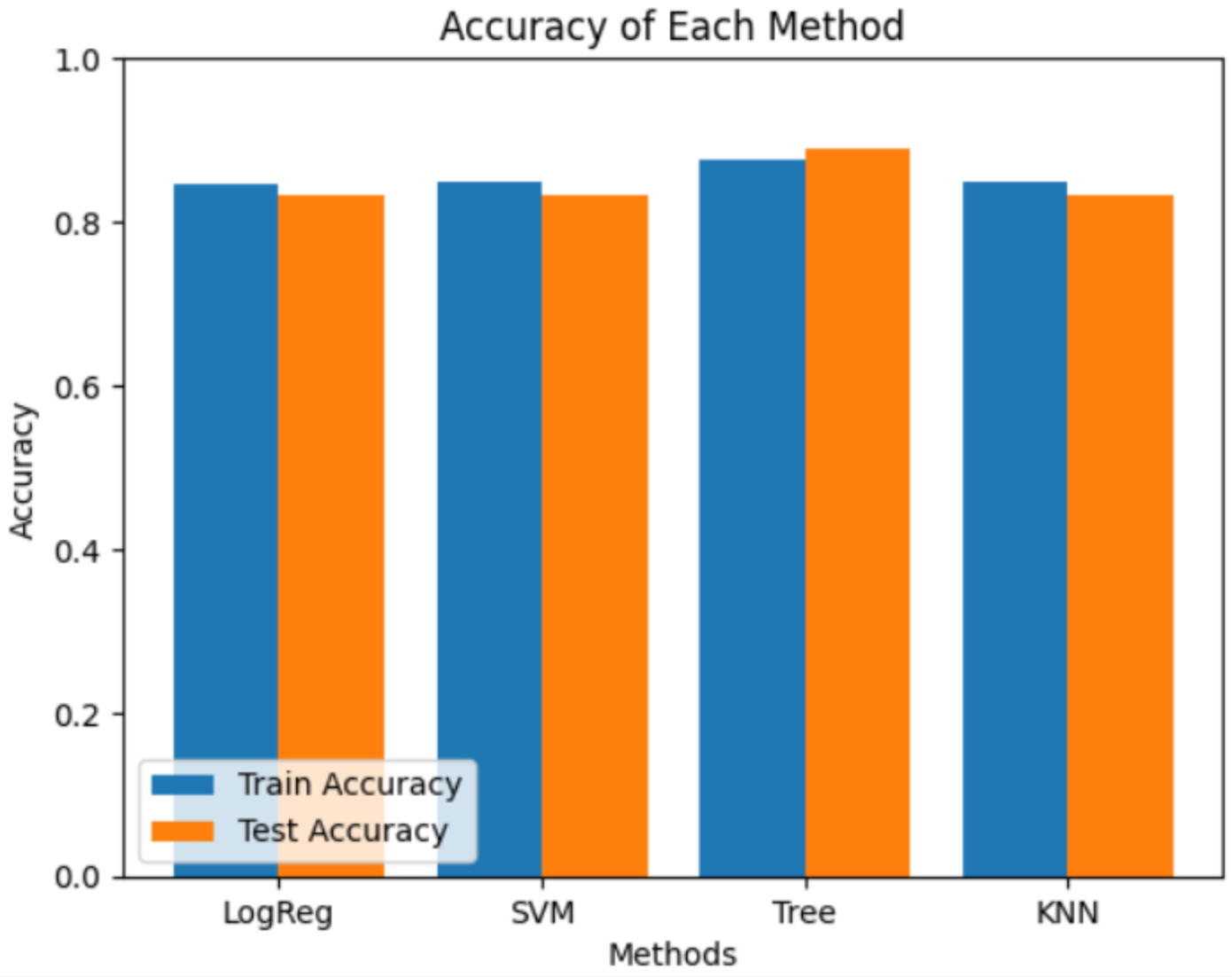
Payload Mass vs. Launch Outcome

- Explanation:
  - Payloads between 2,000 and 5,500 kg have the highest success rate

Section 5

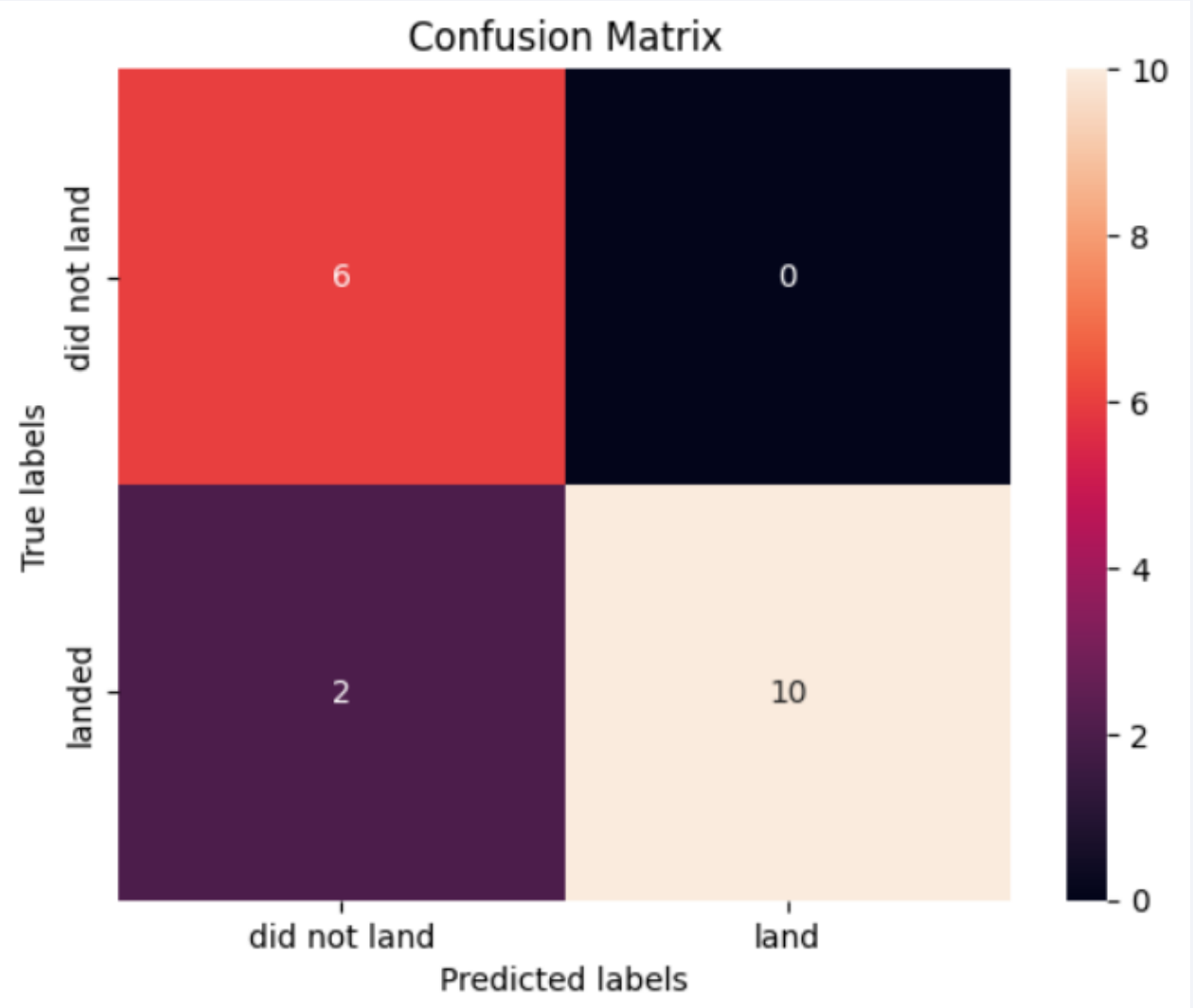
Predictive Analysis (Classification)

Classification Accuracy



- Four classification models where used to test the data:
  - Logistic Regression (LogReg)
  - Support Vector Machine (SVM)
  - Decision Tree (Tree)
  - K Nearest Neighbor (KNN)
- The Decision Tree model had the highest accuracy of the models. The training data was 87.5% accurate and the testing data was 88.89% accurate.

Confusion Matrix



- The confusion matrix helps show the accuracy by showing the high volume of true positives (10) and true negatives (6) the model predicted.
- The model predicted 2 false negatives but 0 false positives.

Conclusions

- Launch site KSC LC 39A had the highest rate off successful launches.
- Successful landings increased beginning in 2013 with a peak of 85% in 2019.
- ES-L1, GEO, HEO, and SSO launch orbits had a 100% success rate.
- Drone ship landings had the highest number off successful landings.
- All launch sites are located in close proximity to a coastline.
- The Decision Tree model performed the best for predicting launch success and failure.

Thank you!