Winning Space Race with Data Science

Nakui Creary
10-24-2023

Outline

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

Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone.git

Executive Summary

- Summary of methodologies
    1. Data Collection through API
    2. Data Collection with Web Scraping
    3. Data Wrangling
    4. Exploratory Analysis using SQL
    5. Exploratory Analysis using Pandas and Matplotlib
    6. Interactive Visual Analytics and Dashboard
    7. Predictive Analysis
- Summary of all results
    1. Exploratory analysis results
    2. Interactive Analytics
    3. Predictive analysis results

Introduction

SpaceX promotes Falcon 9 rocket launches on its website at a price of $62 million, significantly undercutting other providers whose costs can exceed $165 million per launch. The primary source of these savings is SpaceX's ability to recycle the first stage, making it crucial to predict first-stage landings to estimate launch costs and provide valuable data for potential competitors bidding against SpaceX. The project's main objective is to develop a machine learning pipeline for predicting the successful landing of the first stage.

Problems you want to find answers

1. What are the key factors influencing the successful landing of a rocket?
2. How do different features interact to affect the success rate of a landing?
3. What operational conditions must be met to guarantee a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features and dropping unnecessary data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluation of classification models to ensure the best results

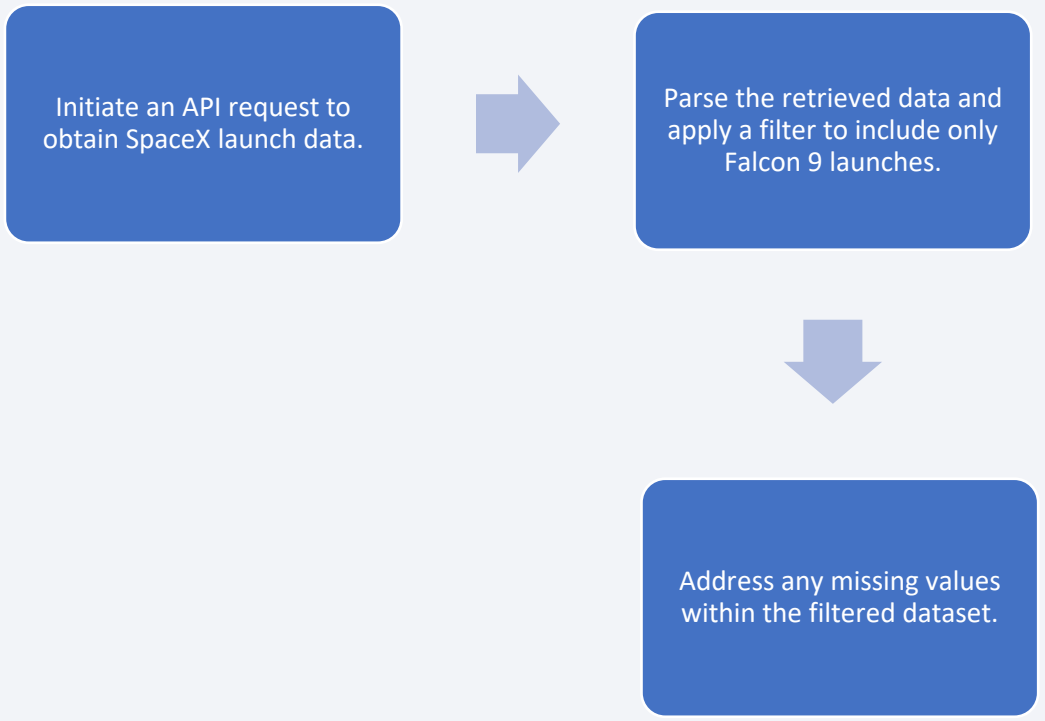
Data Collection

The data collection process comprised a dual approach, incorporating API requests from SpaceX's REST API and web scraping to extract data from a table within SpaceX's Wikipedia entry. This combined methodology was essential to gather comprehensive information about the launches, enabling a more in-depth analysis.

- Data was collected from Space X API (https://api.spacexdata.com/v4/rockets/) and Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping.
- The data was retrieved via SpaceX REST API which included FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- The data was collected through Wikipedia web scraping which encompassed Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API

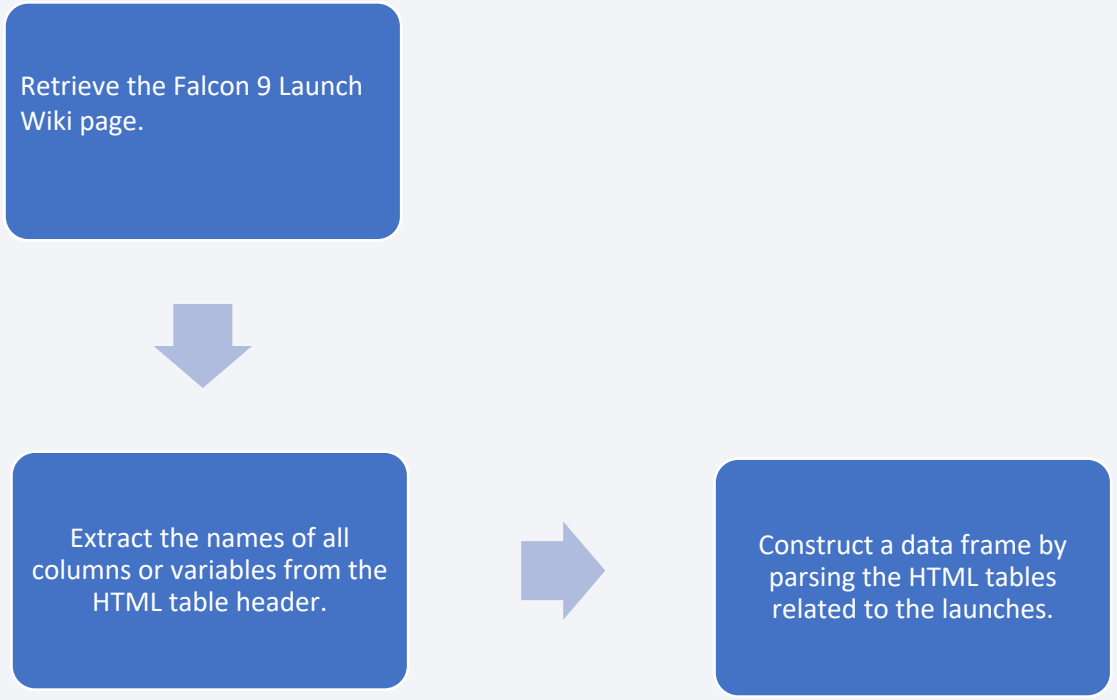
- SpaceX provides a publicly accessible API that allows users to retrieve data for subsequent utilization. This API has been employed following the depicted flowchart, and the obtained data is stored for future use.



Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/ff37003fa340a9e857e1adeb675cb3cf6d4b89ec/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

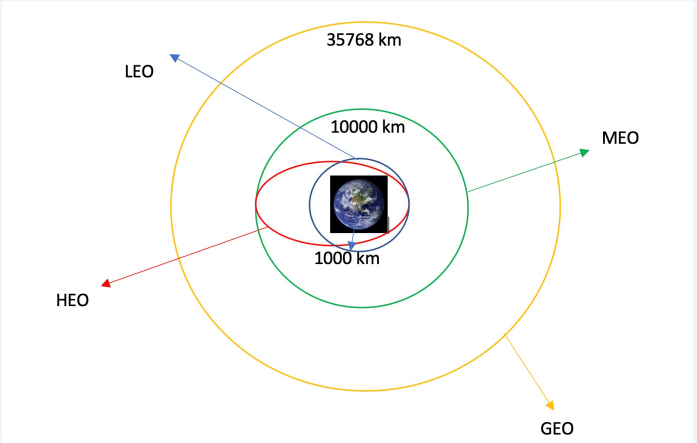
- Information about SpaceX launches can be sourced from Wikipedia. The data retrieval process follows a specific flowchart, and the acquired data is subsequently saved for preservation.



Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/ff37003fa340a9e857e1adeb675cb3cf6d4b89ec/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

- Our process included exploratory data analysis and the identification of training labels. We also conducted an analysis to count the launches at each site and quantify the number and frequency of different orbits. Additionally, we generated a landing outcome label based on the information in the outcome column and exported the resulting data.



Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/c59bcb3996d2963f04eae82143ca33d524c6beea/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

The following charts were created:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, Success Rate Yearly Trend

- Scatter charts prove valuable for examining connections or correlations between two numeric variables.
- Bar charts are employed to make comparisons between a numerical value and a categorical variable. The choice between horizontal or vertical bar charts depends on the data's scale and presentation.
- Line charts typically display numerical values on both axes and are commonly applied to illustrate how a variable changes over time.

Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/bbf9cc12309065840ff72eb3ce83db45257a42f6/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(1).ipynb

EDA with SQL

The SQL queries performed on the data:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/bbf9cc12309065840ff72eb3ce83db45257a42f6/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following markers and visual elements were implemented:

Markers for All Launch Sites:

- Included markers with circles, popup labels, and text labels for NASA Johnson Space Center, using its latitude and longitude coordinates as the starting location.
- Added markers with circles, popup labels, and text labels for all launch sites, showcasing their geographic positions in relation to the Equator and coastlines.

Colored Markers for Launch Outcomes:

- Introduced colored markers to distinguish between successful (Green) and failed (Red) launches, utilizing Marker Clusters to highlight launch sites with relatively high success rates.

Distance Indicators between Launch Sites and Their Proximities:

- Integrated colored lines to visually represent distances between the launch site KSC LC-39A (for example) and its nearby features such as railways, highways, coastlines, and the closest city.

Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/bbf9cc12309065840ff72eb3ce83db45257a42f6/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

Key additions and features in the project include:

Launch Sites Dropdown List:

- Implementation of a dropdown list for easy selection of launch sites.

Pie Chart Displaying Success Launches:

- Introduction of a pie chart to depict the overall count of successful launches across all sites.
- In the case of a specific launch site selection, the chart showcases the success and failure counts for that site.

Slider for Payload Mass Range:

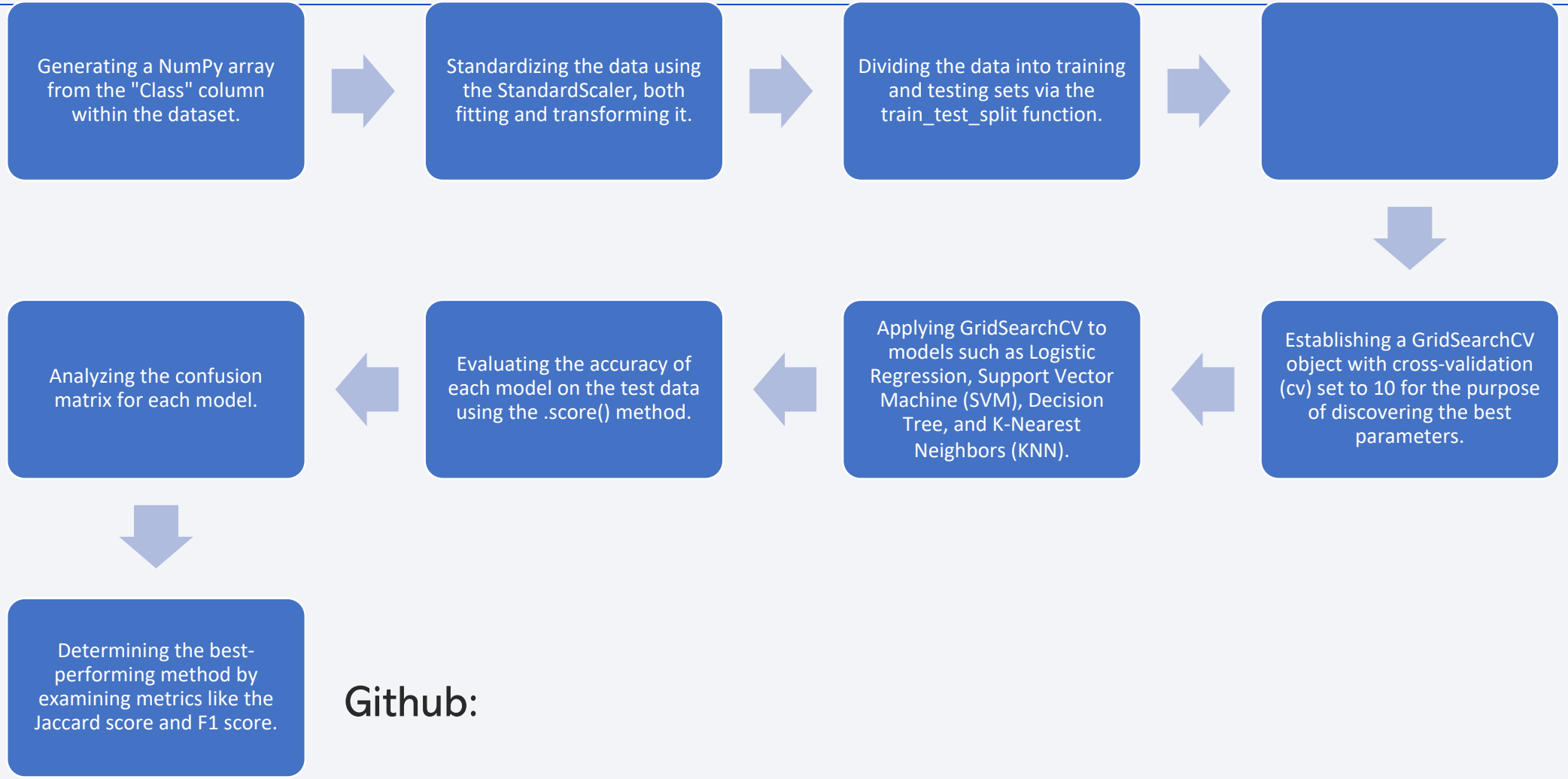
- Inclusion of a slider tool for selecting a specific payload mass range.

Scatter Chart Depicting Payload Mass vs. Success Rate:

- Development of a scatter chart that illustrates the relationship between payload mass and success rate, particularly in the context of different booster versions.

Github: https://github.com/nakuicreary/Applied-Data-Science-Capstone/blob/bbf9cc12309065840ff72eb3ce83db45257a42f6/spacex_dash_app%20(1).py

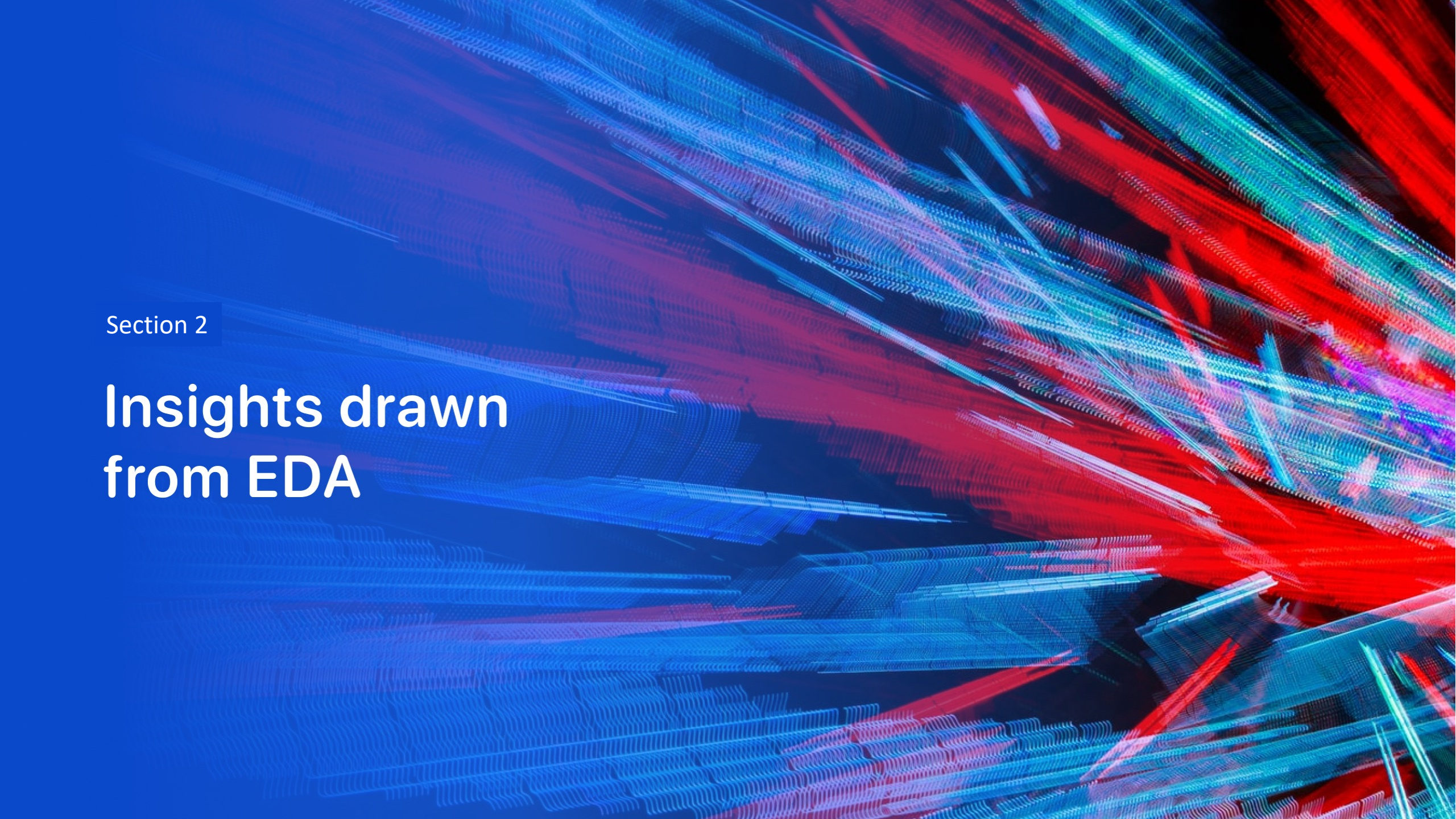
Predictive Analysis (Classification)



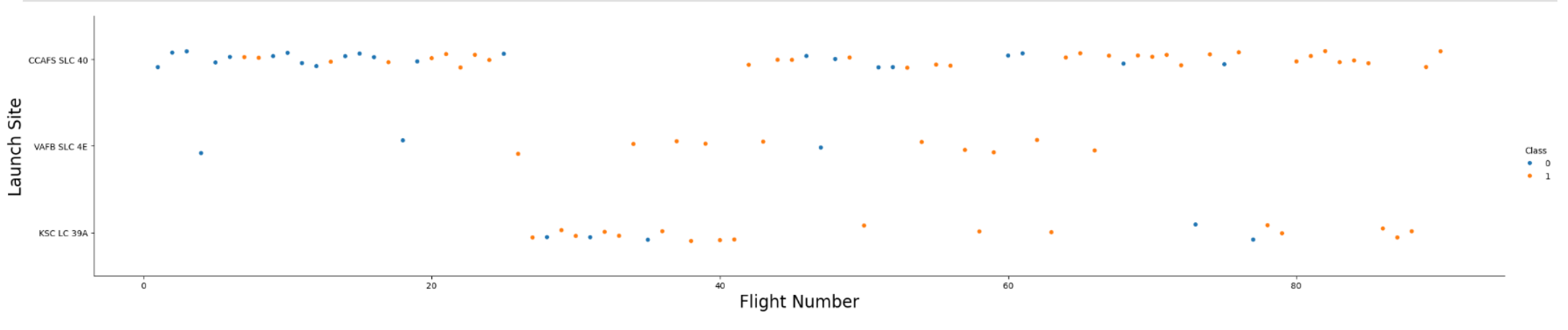
Github:

Results

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

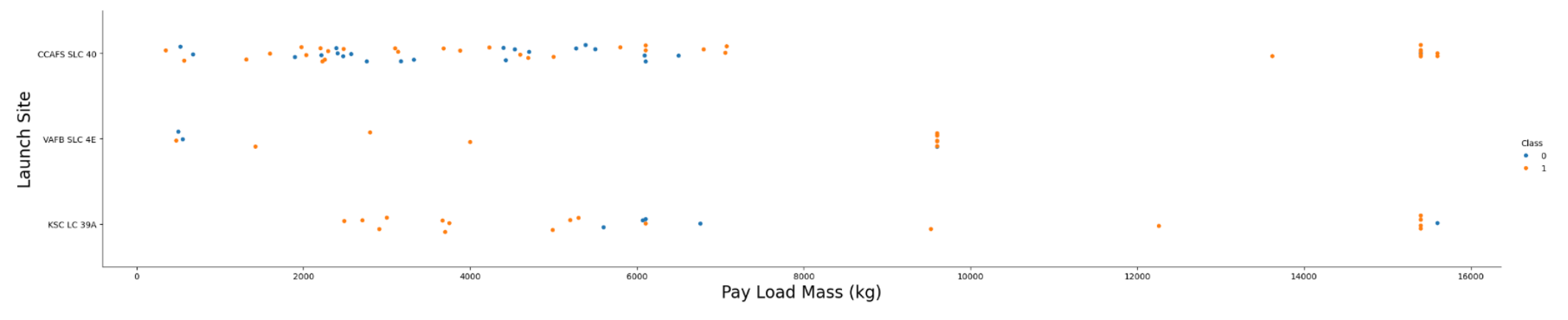

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Based on the depicted plot, we can observe that the most successful launch site at present is CCAF5 SLC 40, with most recent launches ending successfully. Following closely, VAFB SLC 4E holds the second position, while KSC LC 39A secures the third spot in terms of success rates. Additionally, there is a noticeable overall improvement in the success rate over time.

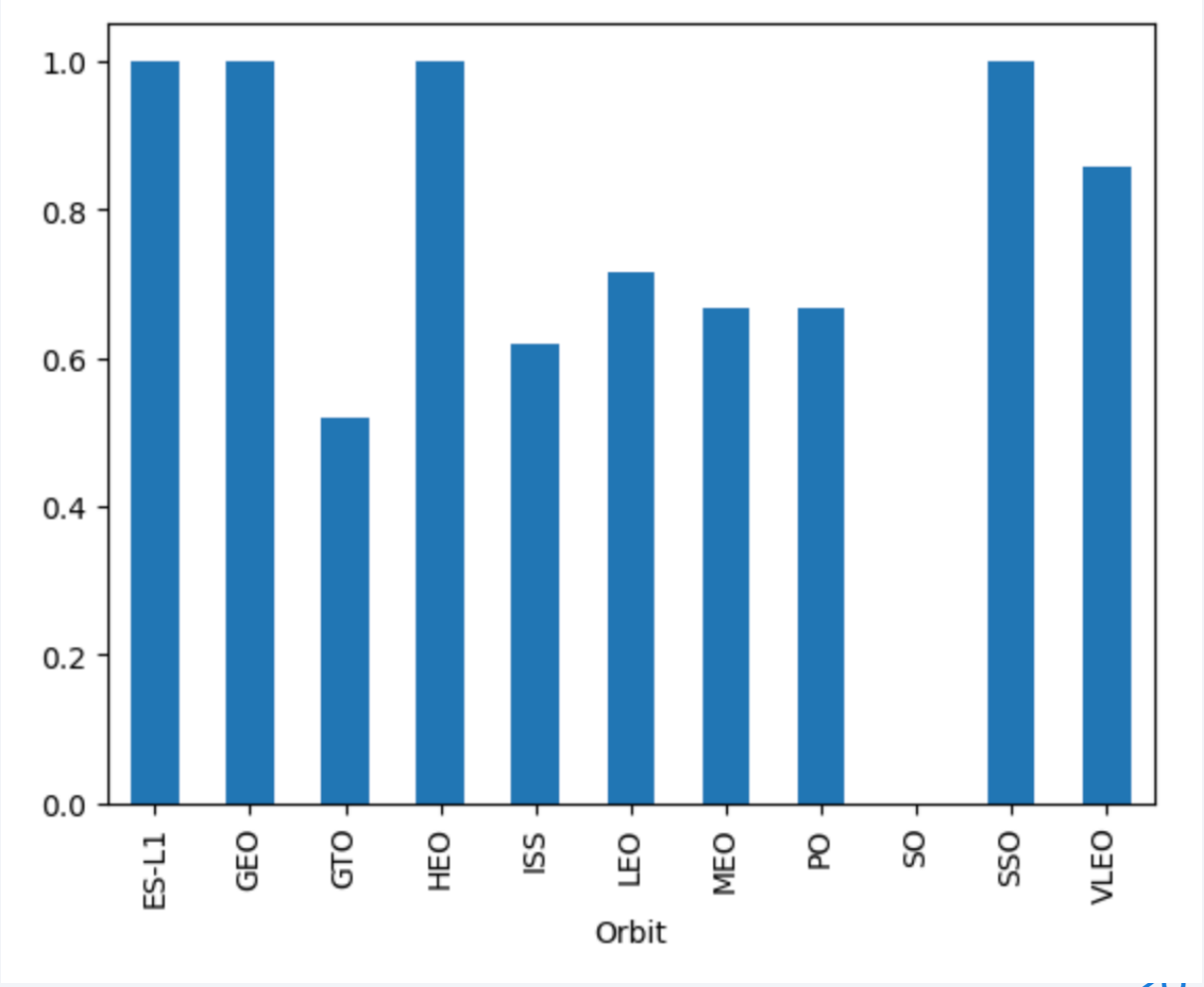
Payload vs. Launch Site



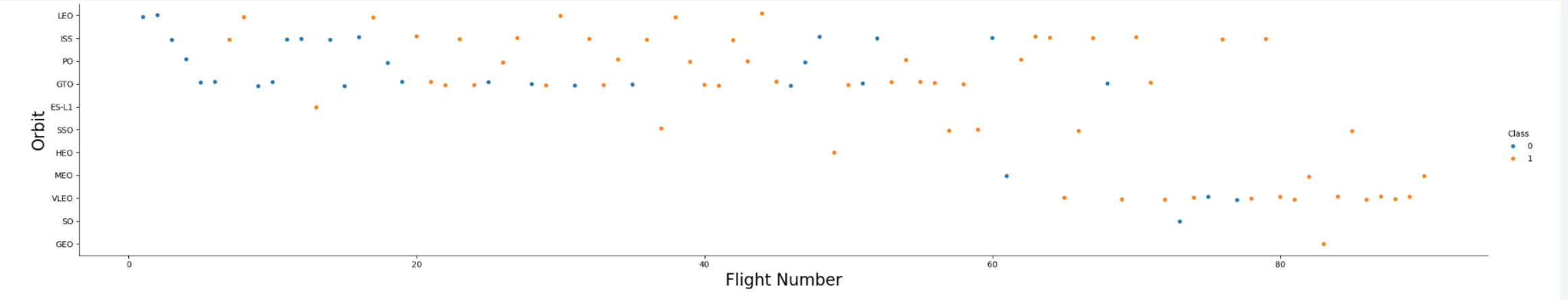
Across all launch sites, a consistent trend is evident: a higher payload mass is associated with a higher success rate. Furthermore, it's notable that most launches with a payload mass exceeding 7000 kg achieved success. Moreover, KSC LC 39A boasts a remarkable 100% success rate for payloads under 5500 kg as well.

Success Rate vs. Orbit Type

- The plot reveals that ES L1, GEO, HEO, SSO, and VLEO exhibit the highest success rates.

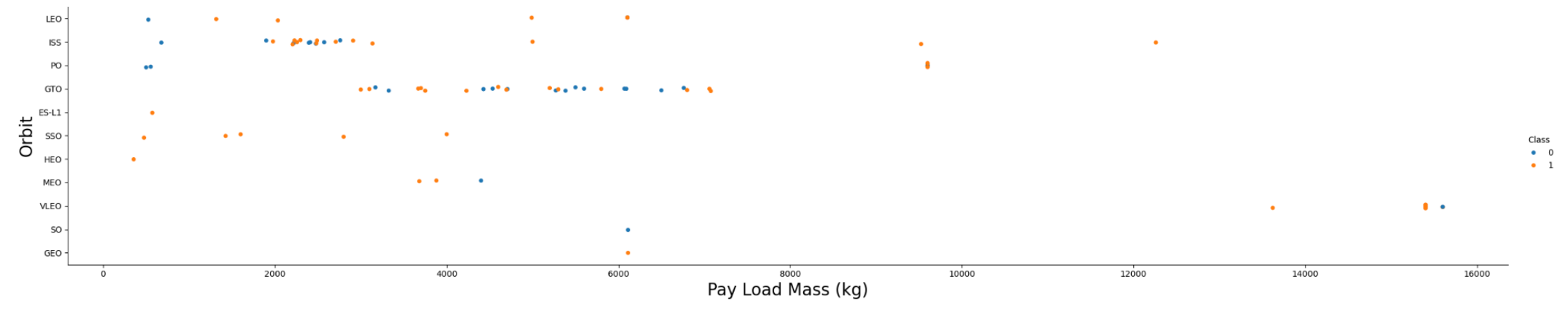


Flight Number vs. Orbit Type



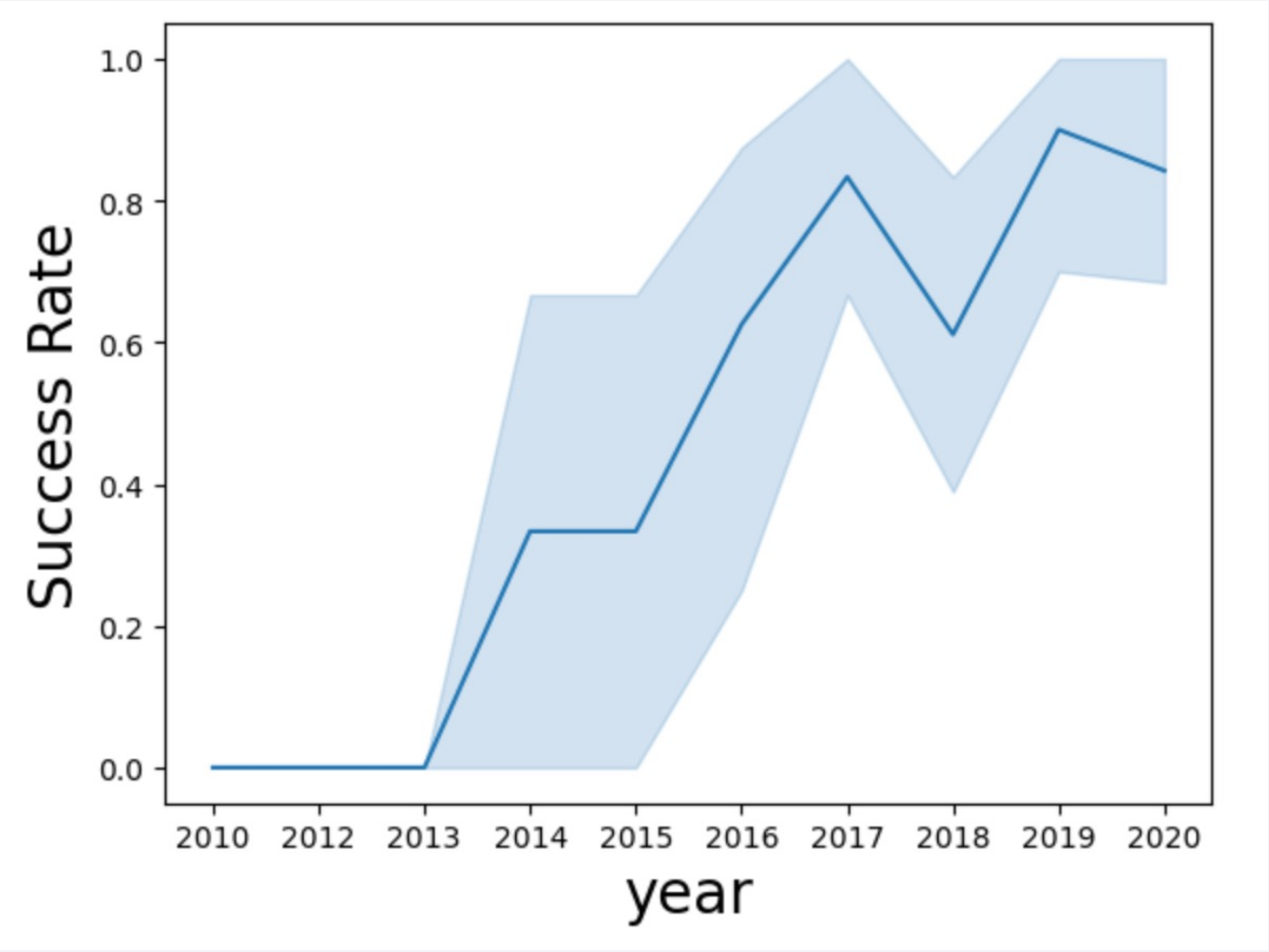
- The depicted plot illustrates the relationship between Flight Number and Orbit type. It becomes apparent that, in the LEO orbit, success rates are influenced by the number of flights, while in the GTO orbit, there is no discernible connection between flight number and the orbit's success.

Payload vs. Orbit Type



It's noticeable that for heavy payloads, there is a higher frequency of successful landings in PO, LEO, and ISS orbits.

Launch Success Yearly Trend



From the graph, it's evident that the success rate consistently rose from 2013 to 2020.

All Launch Site Names

- Showing the distinct names of the space mission launch sites.

In [8]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- Showing 5 entries where the launch sites have names starting 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_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (I
2010-08-12	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 (I
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Presenting the cumulative payload mass transported by NASA's (CRS) launched boosters.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE
```

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

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- Showing the mean payload mass transported by booster version F9 v1.1.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- Enumerating the date of the initial successful landing on a ground pad.

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Succe
```

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

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Enumerating the names of boosters that achieved success on a drone ship and carried a payload mass exceeding 4000 but less than 6000.

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 400
```

```
 * 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

- Enumerating the total count of both successful and unsuccessful mission outcomes.

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME OR
```

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

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

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT M
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Here are the boosters that transported the highest recorded payload mass in the dataset.

2015 Launch Records

- Enumerating the unsuccessful landing results on drone ships, along with their corresponding booster versions and launch site names for the months within the year 2015.

```
%sql SELECT substr(Date, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND sul
```

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

Month	Booster_Version	Launch_Site
04	F9 v1.1 B1015	CCAFS LC-40
10	F9 v1.1 B1012	CCAFS LC-40

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

- Sorting the number of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

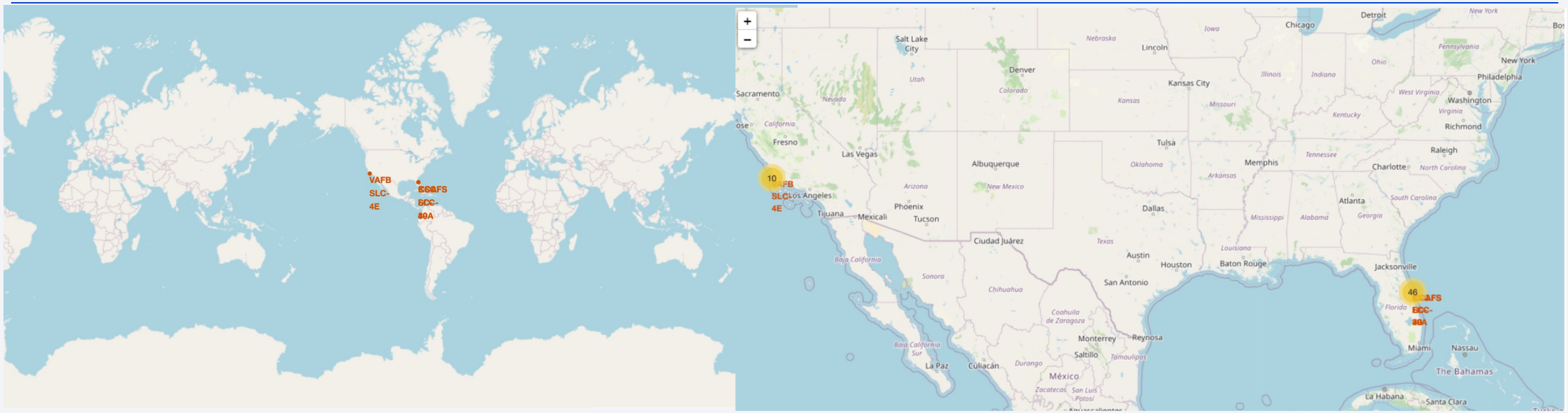
```
%sql select landing_outcome, count(*) as count_outcomes from SPACEXTBL where date between '2010-06-04' and '2017-03-20' group by landing_ou
```

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

Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

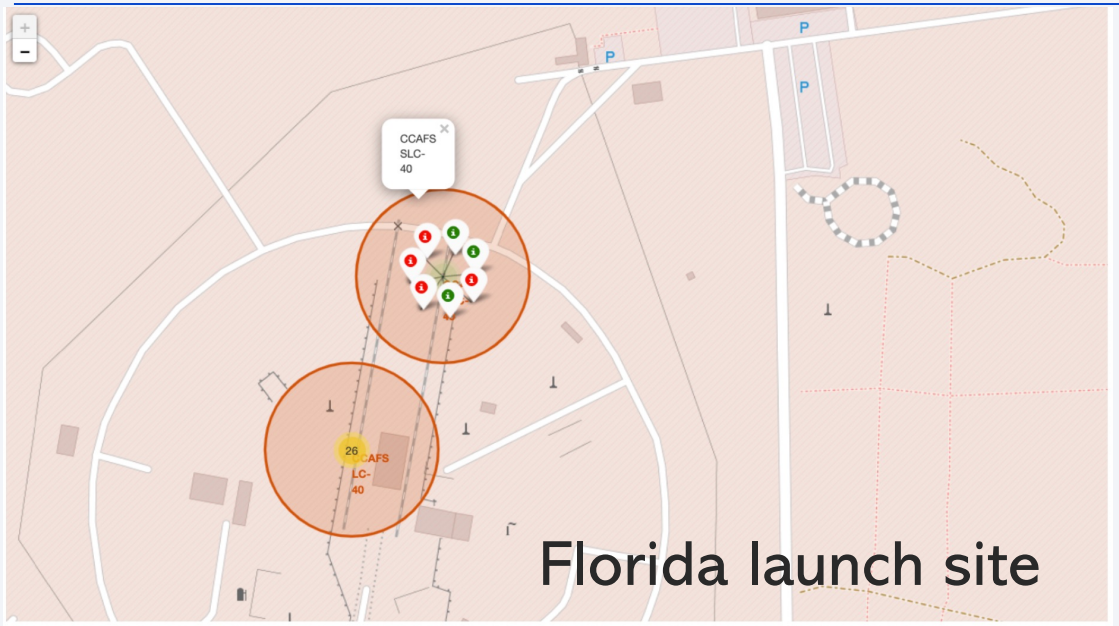

Section 3
Launch Sites
Proximities Analysis

Map markers for all global launch sites

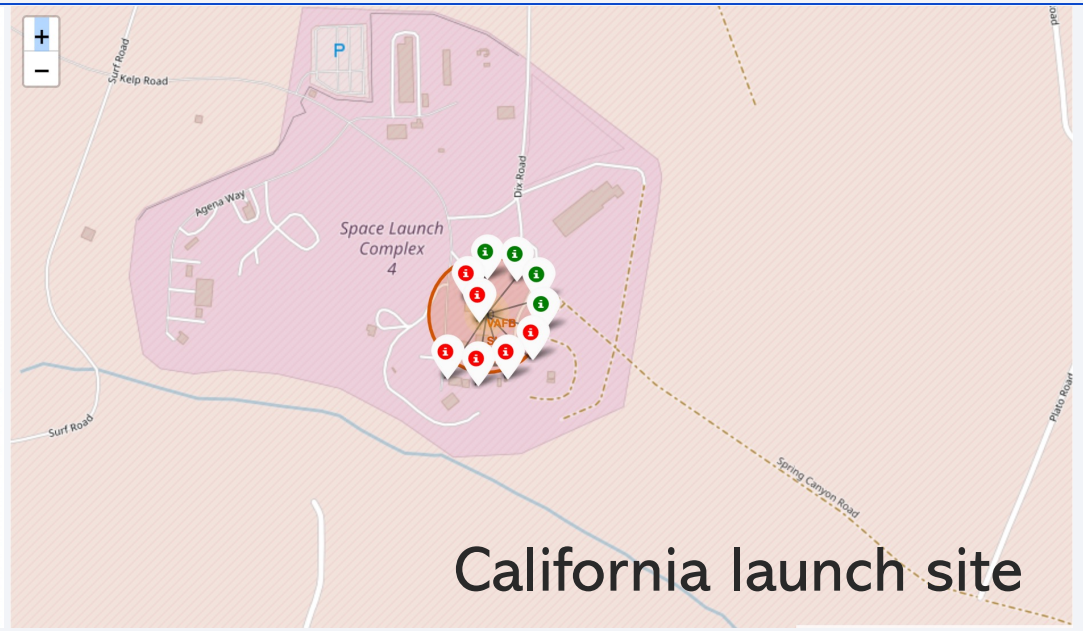


SpaceX launch sites are in Florida and California

Markers on display indicating launch sites with labels



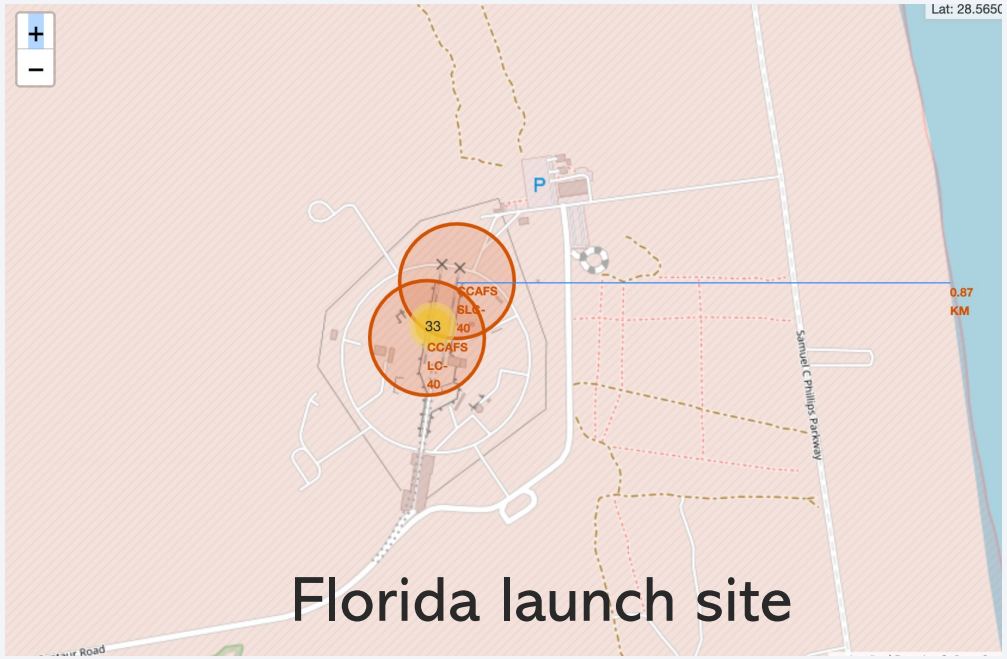
Florida launch site



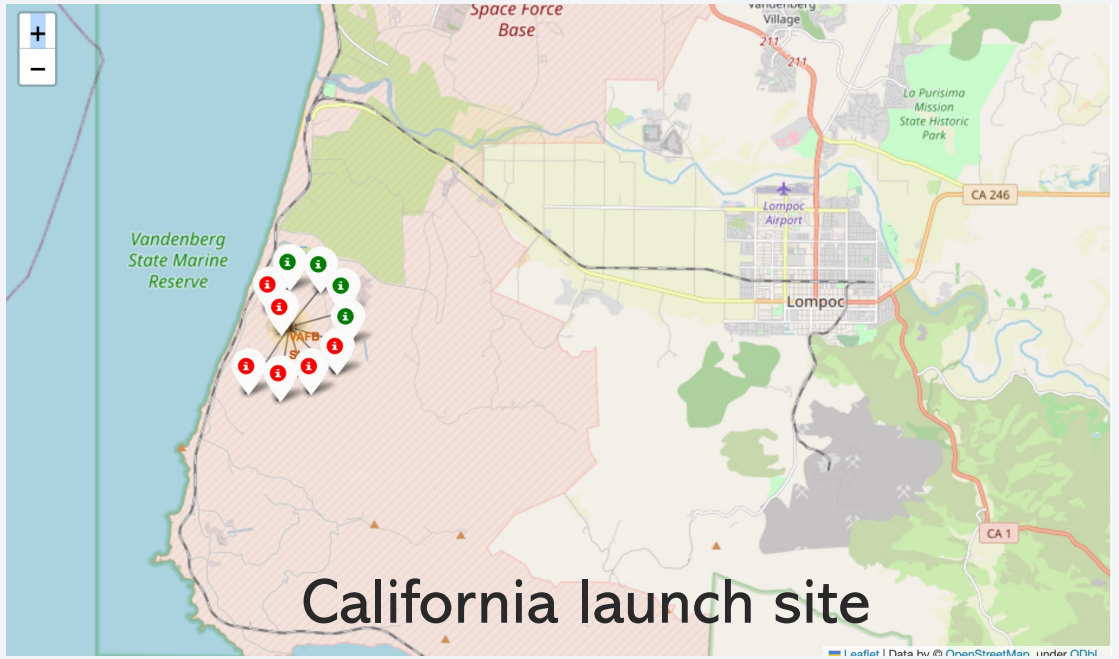
California launch site

Green Markers show successful launches and Red Markers show failed launches

Distance from launch sites to notable landmarks

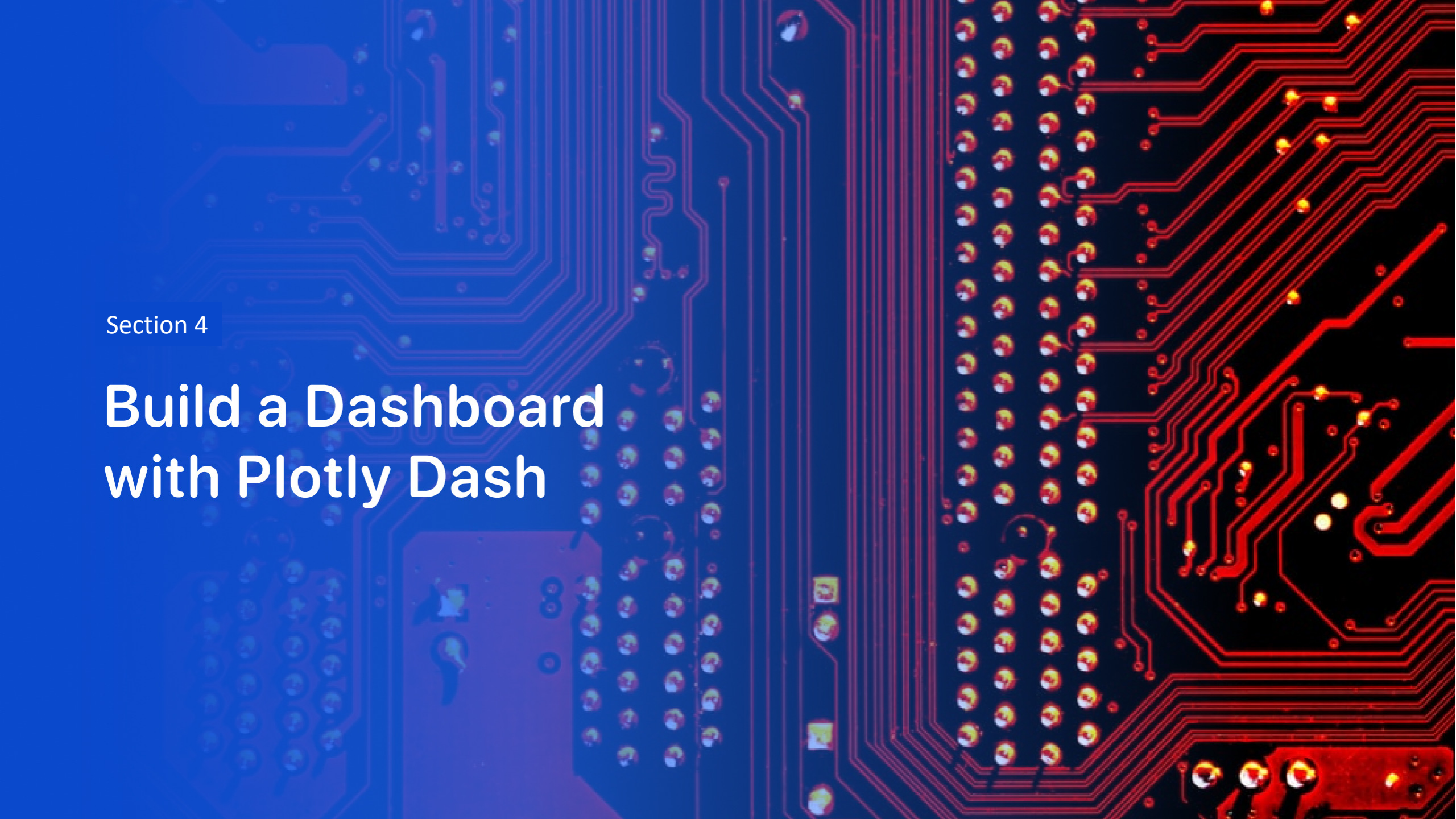


Florida launch site

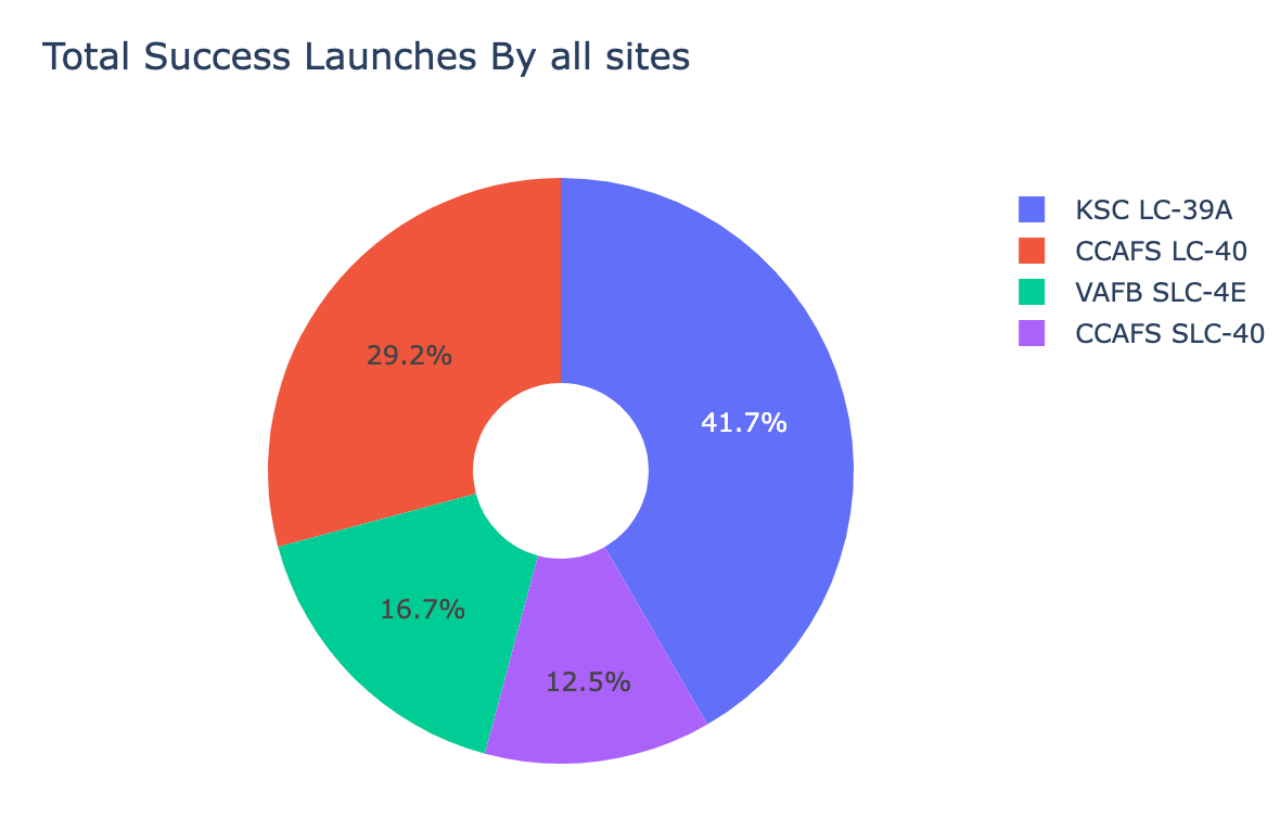


California launch site

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to railways? Yes
- Do launch sites keep certain distance away from cities? Yes

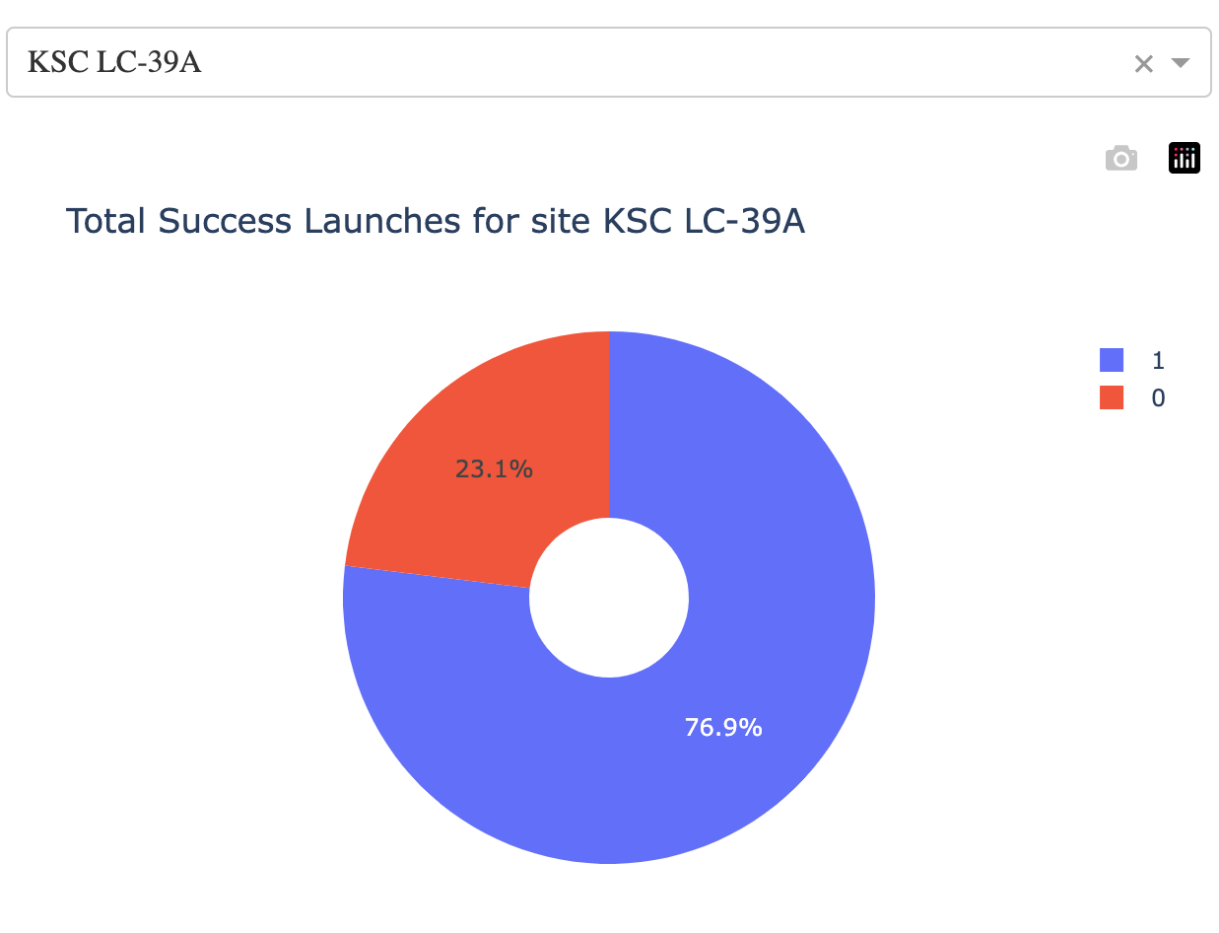

Section 4
Build a Dashboard with Plotly Dash

Pie chart representing the success percentages attained by each launch site



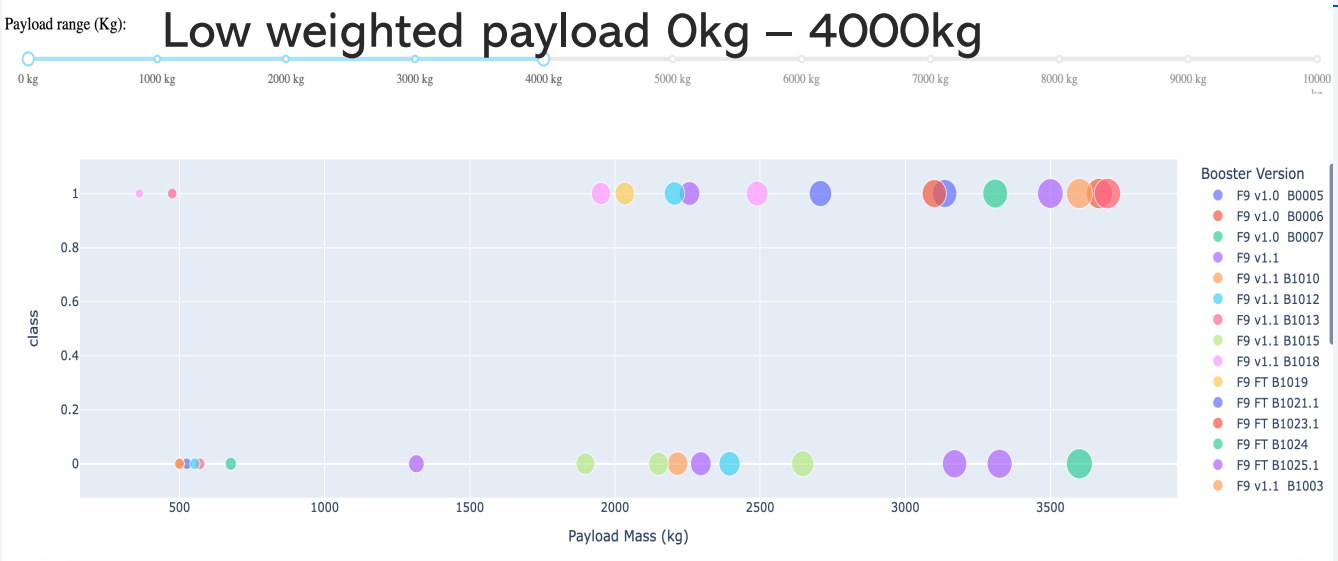
- Among all the launch sites, KSC LC-39A has the highest number of successful launches.

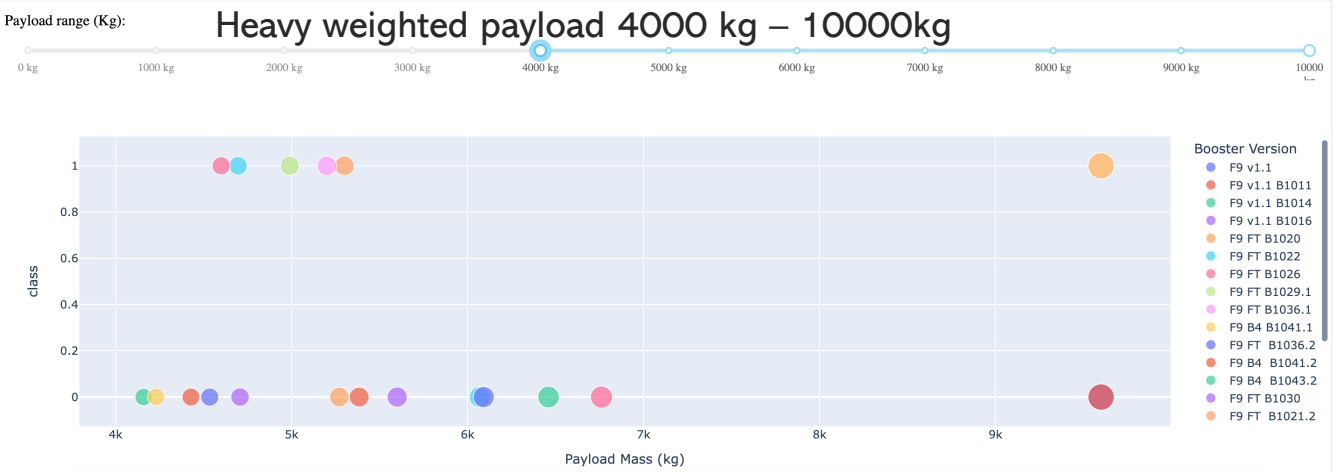
Pie chart illustrating the launch site with the greatest success rate for launches



- KSC LC-39A recorded a 76.9 percent success rate and a 23.1 percent failure rate.

Payload vs Launch outcome scatter plot for all sites



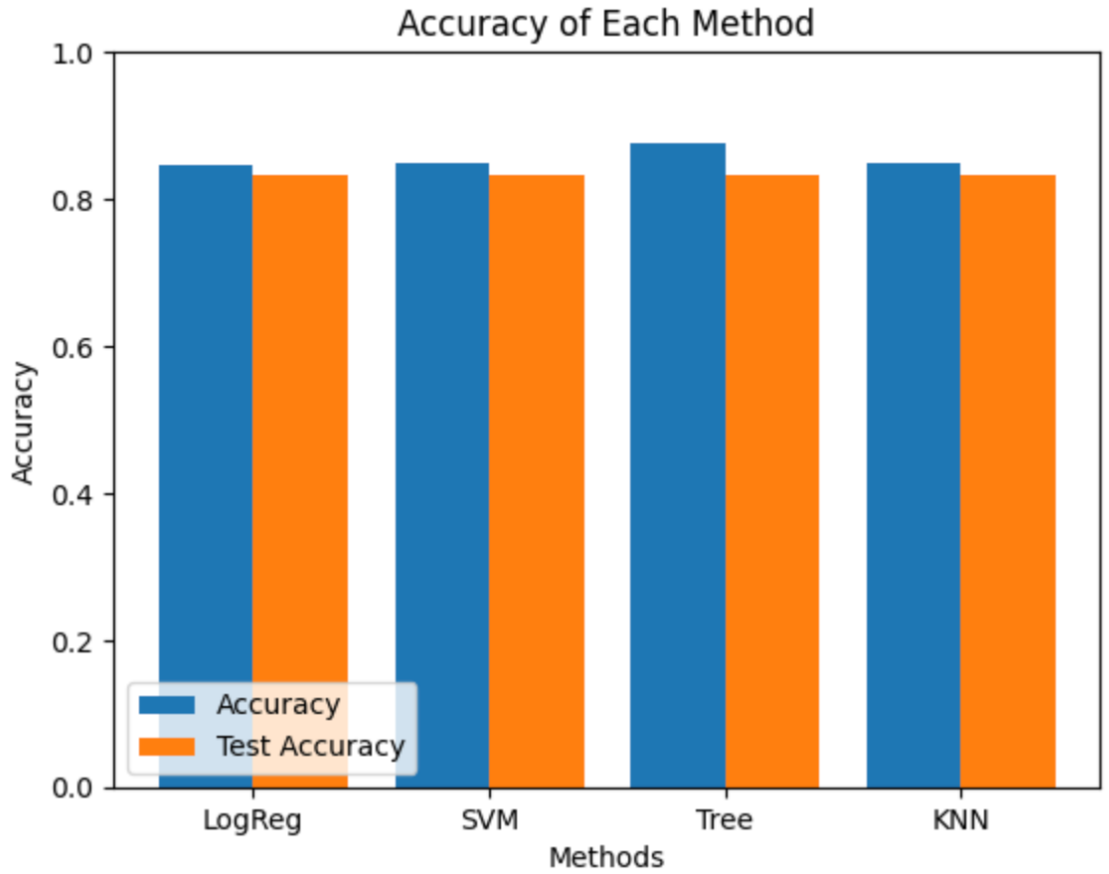


The success rates for lighter payloads surpass those for heavier ones

Section 5

Predictive Analysis (Classification)

Classification Accuracy

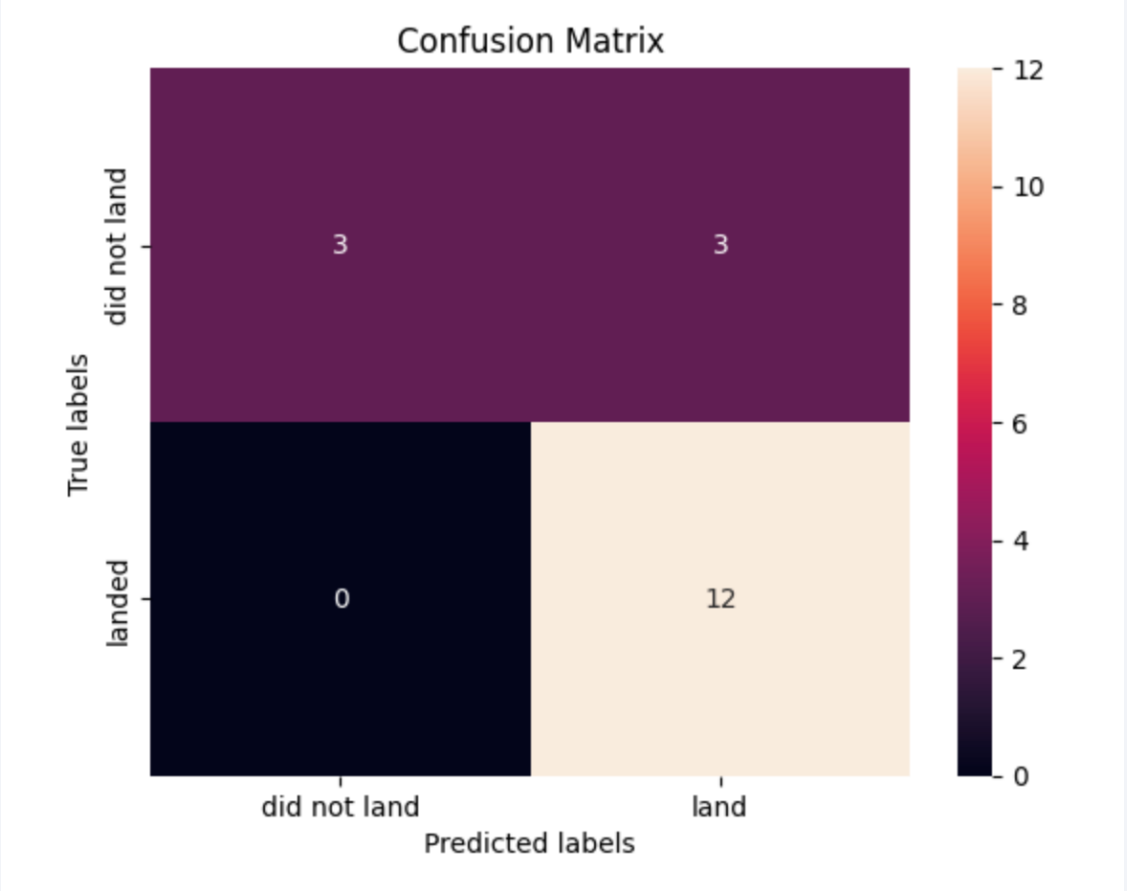


The decision tree classifier boasts the highest accuracy among the models.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.83333
KNN	0.84821	0.83333

Confusion Matrix

The confusion matrix associated with the decision tree classifier clearly demonstrates the model's ability to differentiate between distinct classes. The primary challenge arises from the occurrence of false positives, where unsuccessful landings are incorrectly classified as successful landings by the classifier.



Conclusions

Based on the data analysis, we can conclude that:

- The Decision Tree is the best machine learning algorithm for this data
- Launch success rate stated to increase in 2013 to 2020
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate
- KSC LC-39A had the highest success rate of the launches from all the sites
- Launches with a smaller payload mass provided better results than launches with larger payload mass

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!