IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

P.L. YIP
February 2025

Outline

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

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 result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
  - SpaceX's reusable rocket technology offers significant cost savings. This project aims to build a machine learning pipeline predicting first-stage landing success using pre-existing data. Accurate predictions can inform cost estimates and provide valuable insights for companies competing with SpaceX.
- Problems you want to find answers
  - Which machine learning model best predicts rocket landing success?
  - What are the key factors influencing successful landings?
  - Can we identify specific operational or environmental conditions that significantly impact the likelihood of a successful landing and reuse of the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

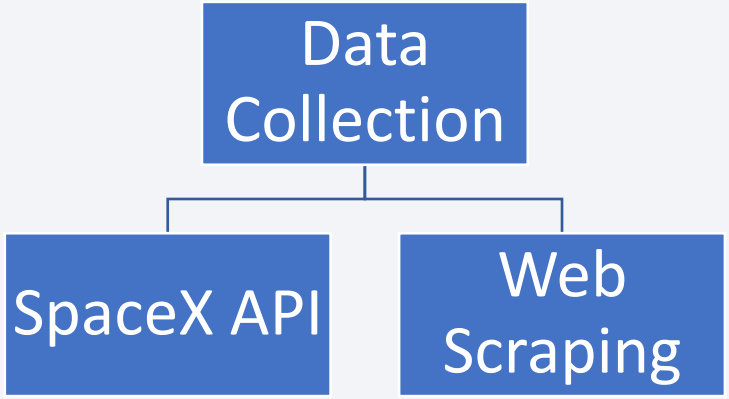
The data sets for the Falcon 9 First Stage Landing Prediction were collected from two main sources:

1.SpaceX API:

- Utilized an Open Source REST API that provides data on launches, rockets, cores, capsules, Starlink, launchpads, and landing pads.
- Made GET requests to the SpaceX API to retrieve relevant information for analysis.
- Extracted data from the API responses to gather details about launches, rockets, cores, payloads, and landing outcomes.

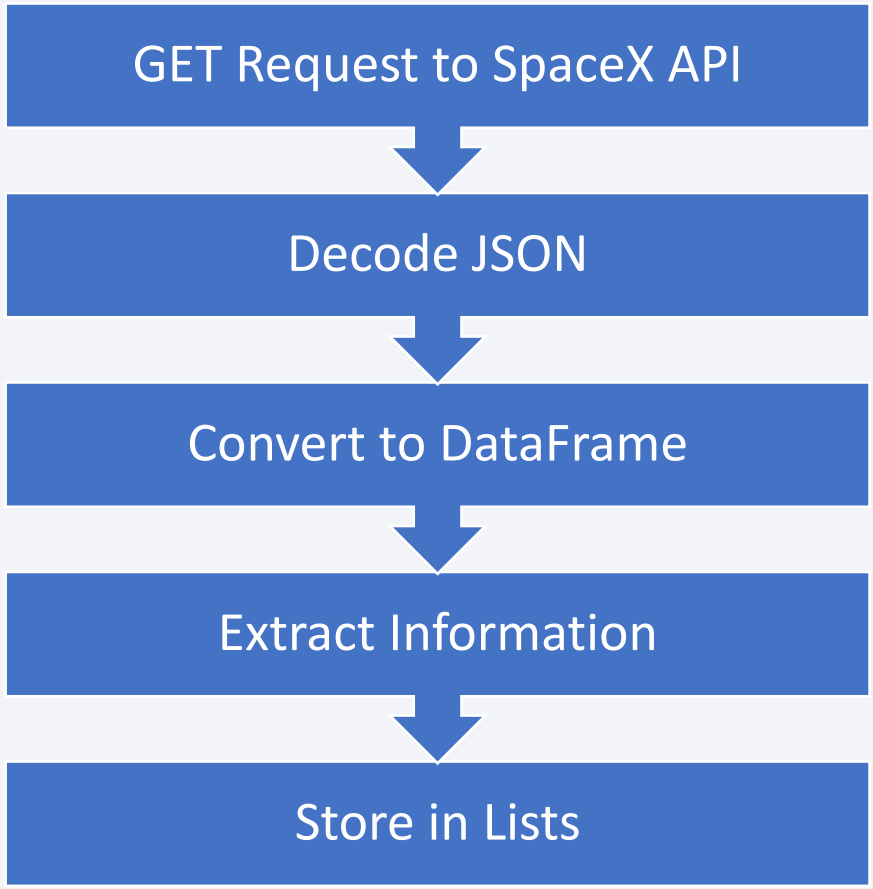
2.Web Scraping from Wikipedia:

- Accessed Wikipedia, a free online encyclopedia maintained by volunteers worldwide and hosted by the Wikimedia Foundation.
- Used web scraping techniques to extract data from the Falcon 9 launch Wiki page on Wikipedia.
- Extracted information such as flight numbers, dates, times, versions, launch sites, payloads, payload masses, orbits, customers, launch outcomes, and booster landings from the Wikipedia page.



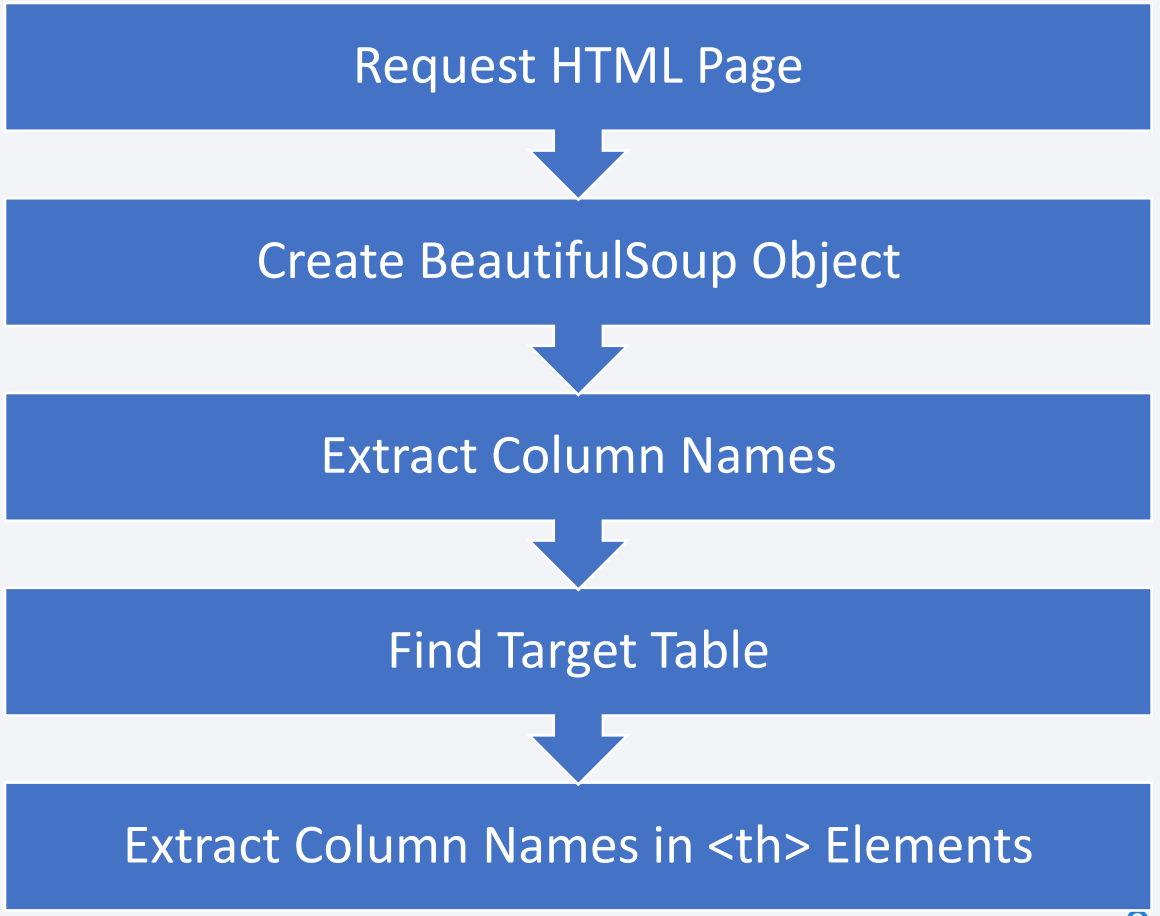
Data Collection – SpaceX API

- Data Collection Process
  - Make a GET request to the SpaceX API
  - Decode the response content as JSON
  - Convert the JSON results into a Pandas dataframe using json_normalize()
  - Extract information about rockets, payloads, launchpads, and cores
  - Store the extracted data in lists for further processing
- GitHub of the completed SpaceX API calls notebook: Here



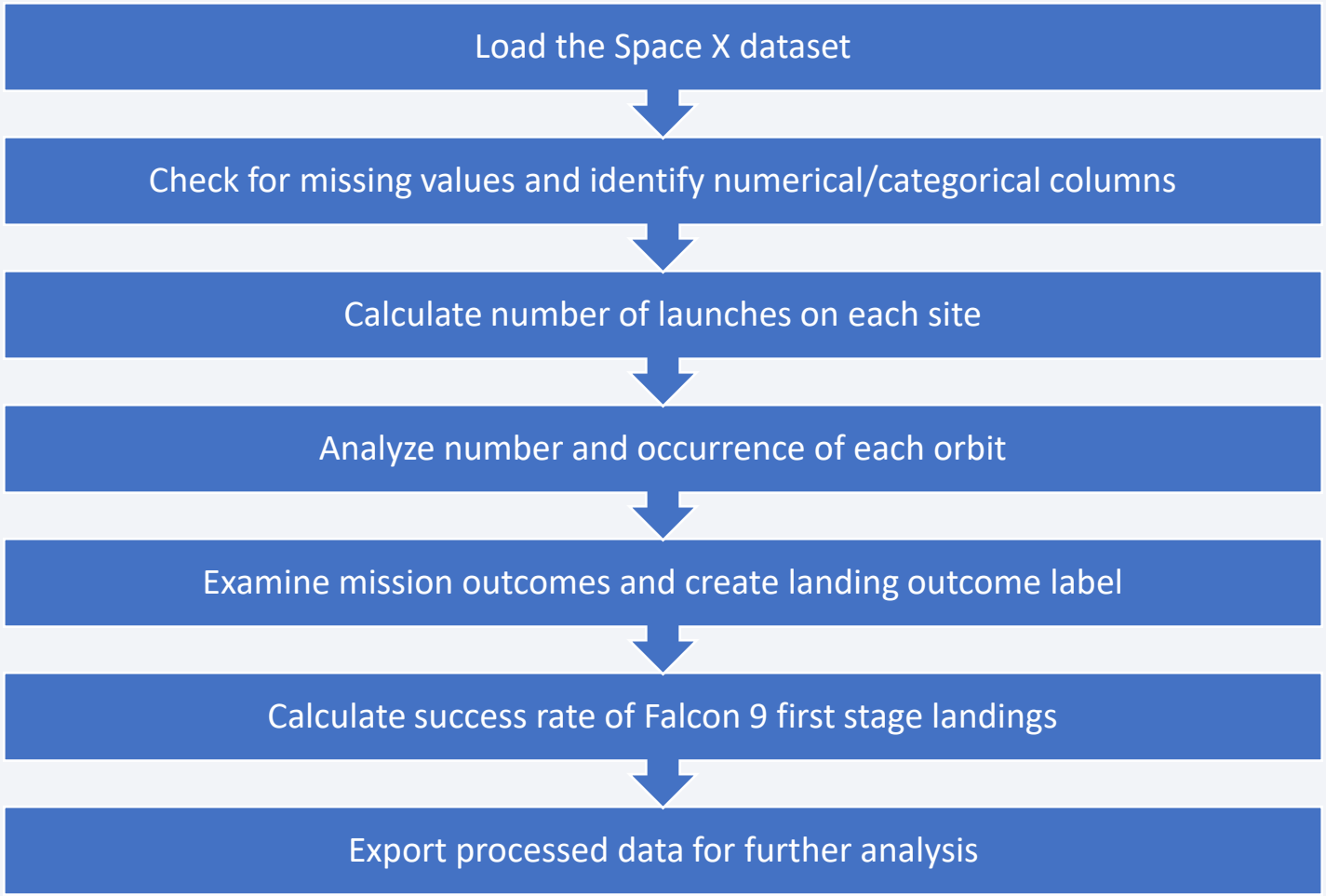
Data Collection - Scraping

- Web scraping process
  - Request the Falcon9 Launch Wiki page from its URL using the requests.get() method
  - Create a BeautifulSoup object from the HTML response content
  - Extract column/variable names from the HTML table header using find_all function in BeautifulSoup
  - Start from the third table which contains the actual launch records
  - Find the specific table containing the column names in <th> elements
- GitHub of the completed web scraping notebook: Here



Data Wrangling

- Data processing
  - The data processing involved loading the dataset, identifying missing values, determining numerical and categorical columns, calculating the number of launches on each site, analyzing the number and occurrence of each orbit, examining the mission outcomes, creating a landing outcome label, calculating success rate, and exporting the processed data.
- GitHub of the completed data wrangling related notebooks: Here



EDA with Data Visualization

- Charts used to explore factors affecting Falcon 9's landing success.
  - **Flight Number vs. Launch Site**: Scatter plot showing site usage and success rates.
  - **Payload vs. Launch Site**: Scatter plot to relate payload mass to launch site success.
  - **Success Rate by Orbit Type**: Bar chart identifying best (ES-L1, GEO, HEO, SSO) and worst (GTO) orbits.
  - **Flight Number vs. Orbit Type**: Scatter plot to find any correlation.
  - **Payload vs. Orbit Type**: Scatter plot showing heavy payloads succeed in Polar, LEO, ISS orbits.
  - **Yearly Trend**: Line chart showing success rate increases from 2013-2017.
  - These insights informed feature engineering, converting categorical data for predictive modeling.
- GitHub of the completed EDA with data visualization notebook : Here

EDA with SQL

- The notebook performs various SQL queries on a SpaceX dataset
  - Find unique launch sites.
  - Show 5 launch sites starting with 'CCA'.
  - Calculate total payload mass for NASA (CRS).
  - Calculate average payload mass for F9 v1.1 boosters.
  - Find the date of the first successful ground pad landing.
  - List boosters with successful drone ship landings and specific payload mass.
  - Count successful and failed mission outcomes.
  - List booster versions carrying the maximum payload mass.
  - Show failure landing outcomes in 2015.
  - Rank landing outcomes between 2010 and 2017.
- GitHub URL of the completed EDA with SQL notebook: Here

Build an Interactive Map with Folium

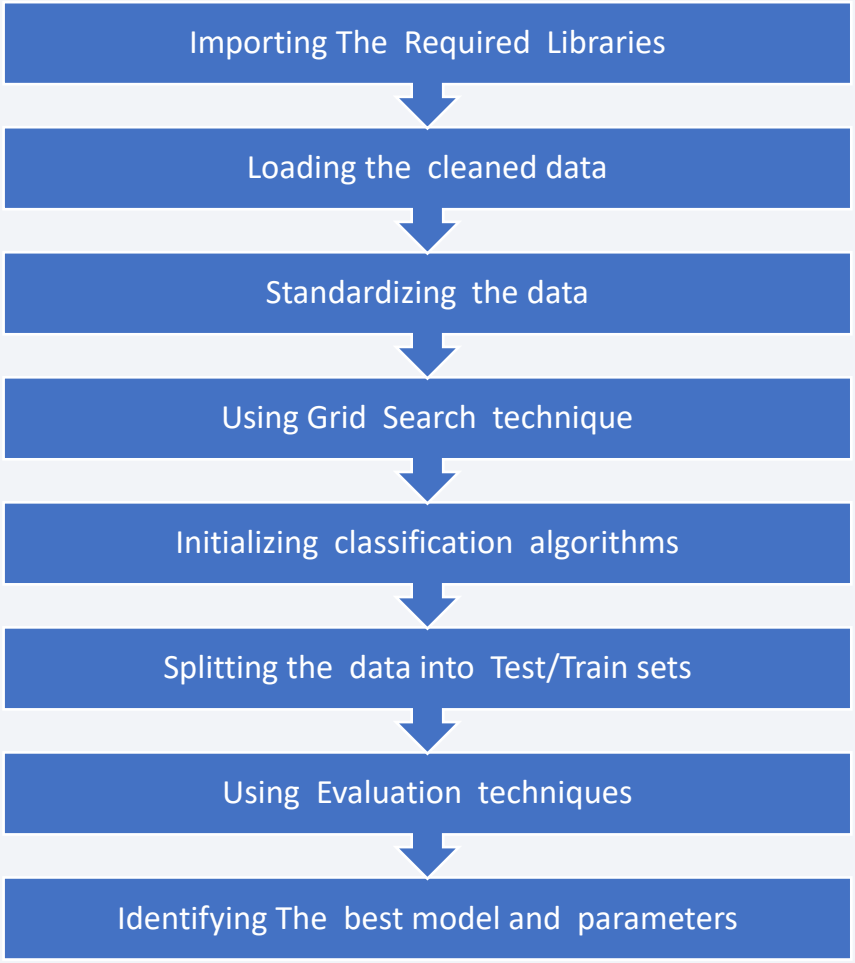
- The Folium map for SpaceX launch sites includes:
  - Markers: Indicate launch site locations with popups displaying site names, enhancing identification.
  - Circles: Highlight areas around each site (1000-meter radius), visually representing proximity and safety zones.
  - Mouse Position: Allows users to see coordinates of any point on the map, aiding in proximity analysis.
  - Reasons for Adding These Objects:
  - Enhanced Visualization: Clear representation of launch site locations and areas.
  - Interactive Analysis: Quick access to information via popups without cluttering the map.
  - Geographical Patterns: Helps identify relationships between launch outcomes and locations, informing strategic decisions for future launches.
- GitHub URL of the completed interactive map with Folium map: Here

Build a Dashboard with Plotly Dash

- **The dashboard includes:**
  - Dropdown List: Allows selection of launch sites ("All Sites," "CCAFS LC-40," etc.).
  - Pie Chart: Displays success rates. Shows overall rates for all sites, or success vs. failure counts for a selected site.
  - Range Slider: Filters data by payload mass (0-10000 kg).
  - Scatter Chart: Visualizes the correlation between payload mass and launch success. Shows all sites or data for the selected site.
- **Reasons:**
  - Dropdown: Enables interactive data filtering by launch site.
  - Pie Chart: Provides a clear visual representation of launch success rates, aiding quick comparisons between sites.
  - Range Slider: Allows users to explore the impact of different payload mass ranges on launch outcomes.
  - Scatter Chart: Reveals the relationship between payload mass, booster version, and success, providing insights into factors influencing mission success.
- GitHub URL of the completed Plotly Dash lab: Here

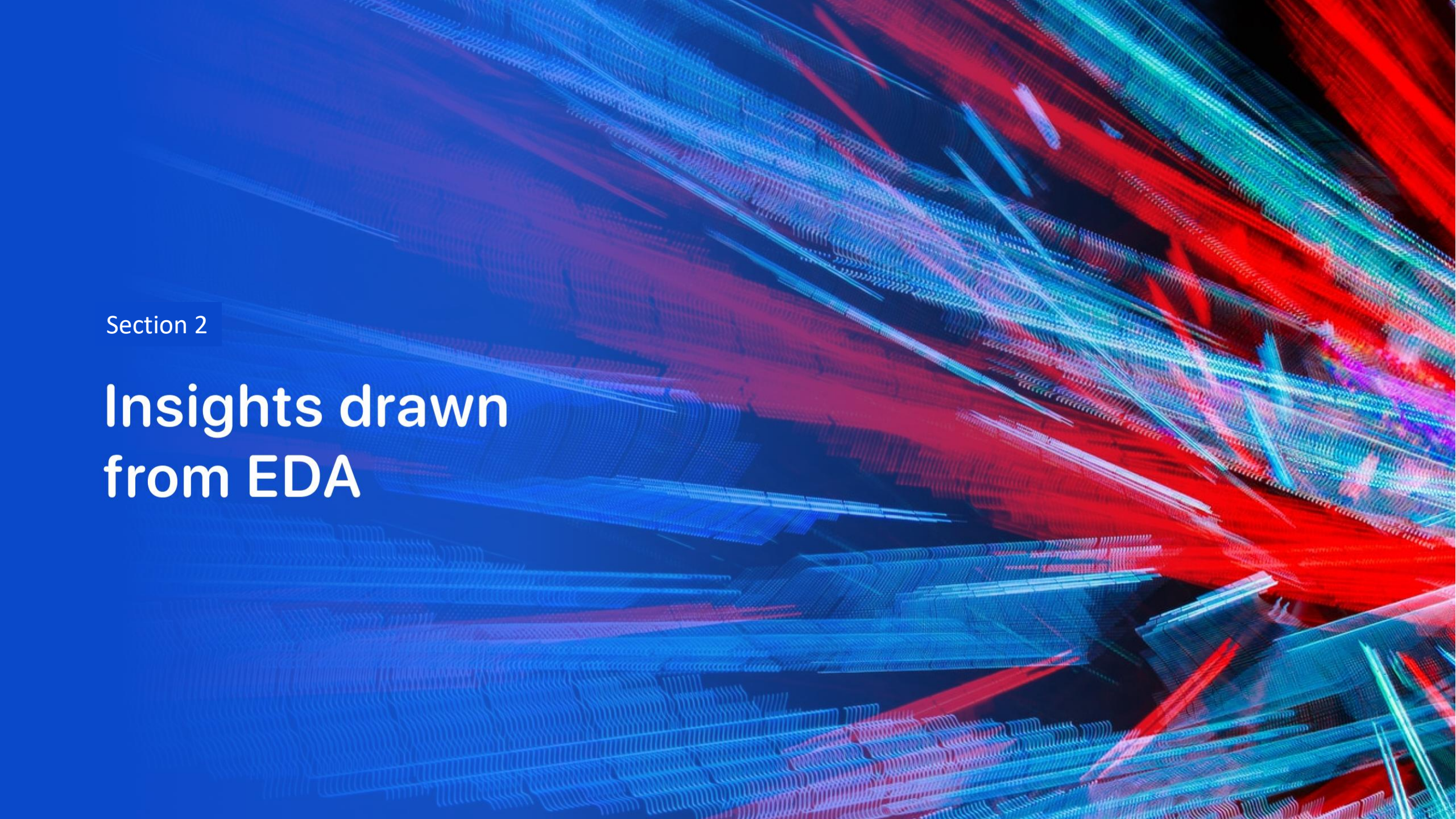
Predictive Analysis (Classification)

- The model development involved: Data Prep (NumPy array, StandardScaler, train/test split), Model Building & Tuning (Logistic Regression, SVM, Decision Tree, KNN, GridSearchCV for hyperparameters), Model Evaluation (accuracy on test data, confusion matrices), and Model Selection (Jaccard, F1 scores). Logistic Regression, SVM, and KNN performed best, with Decision Tree underperforming.
- GitHub URL of the completed predictive analysis lab: Here

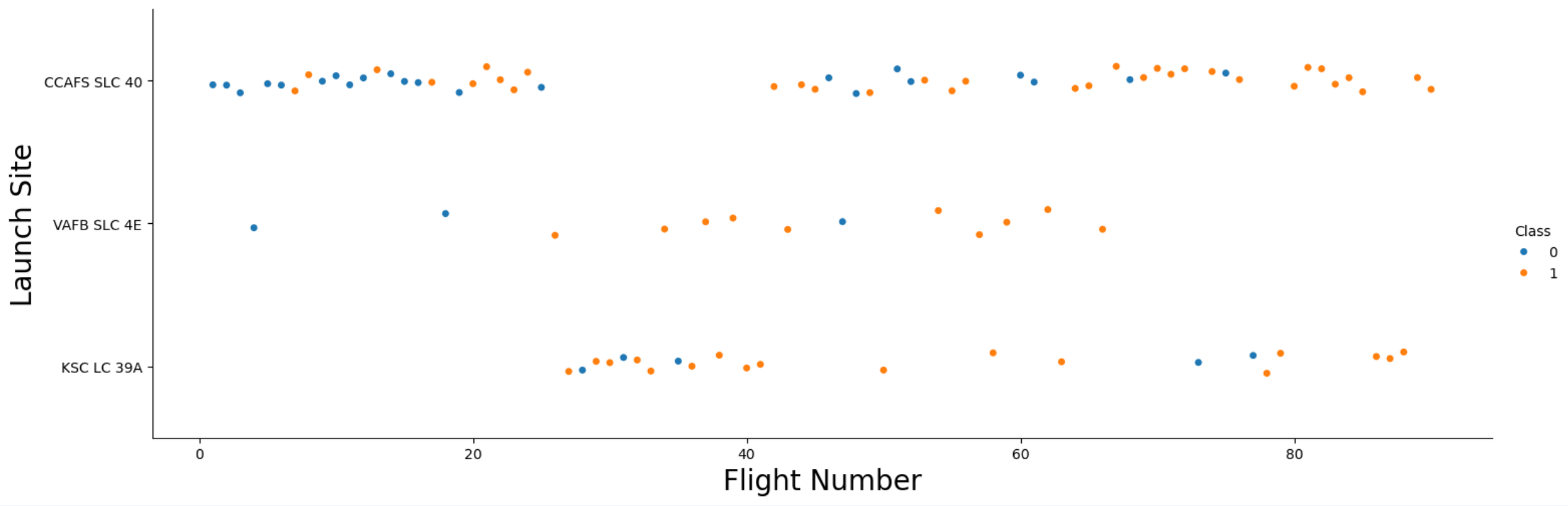


Results

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

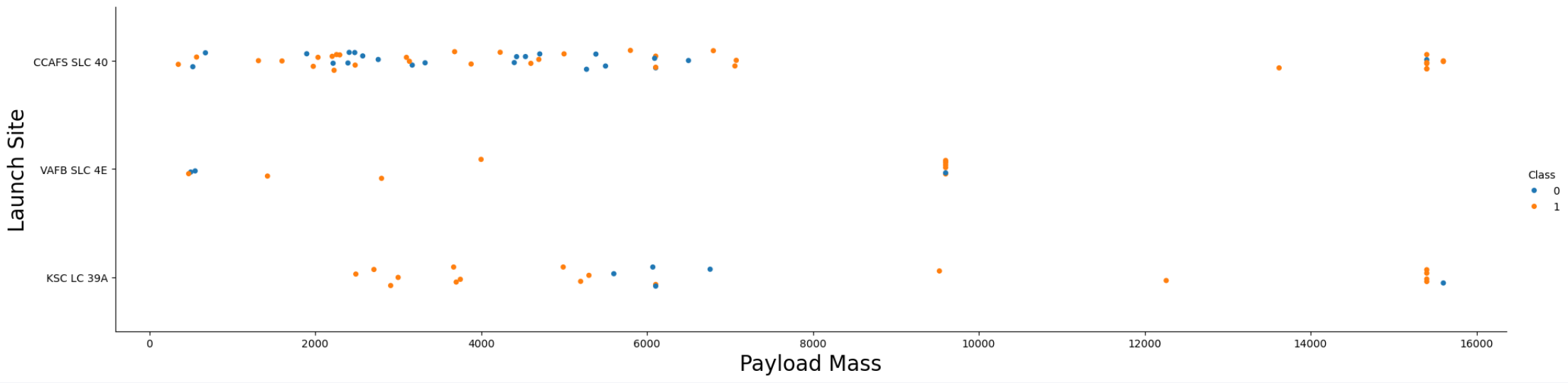

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



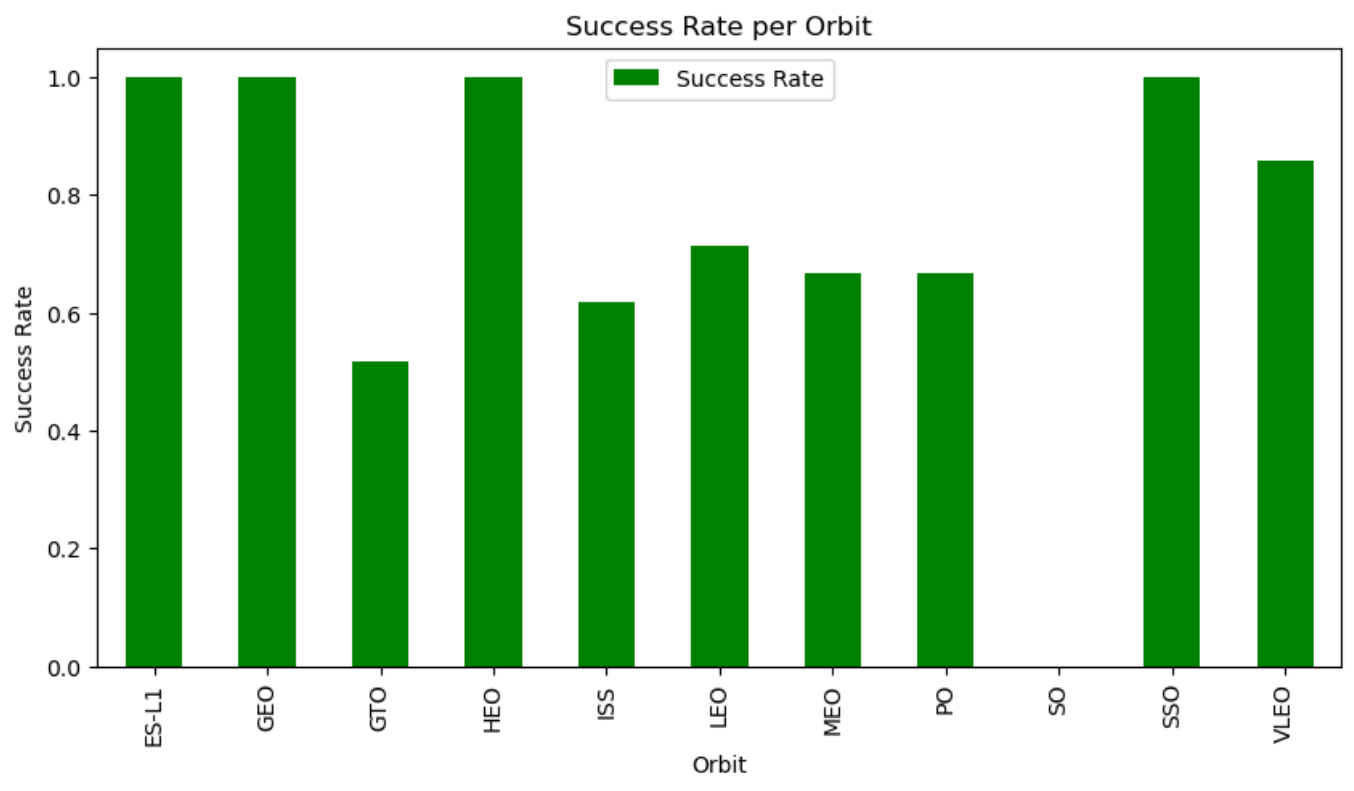
- 1- CCAFS SLC 40 : is the most usable site for launching SpaceX's rockets and it has 55 trials, 33 of them are successful and 22 of them are failed # 60% success rate
- 2- VAFB SLC 4E : is the least usable site for launching SpaceX's rockets and it has 13 trials, 10 of them are successful and 03 of them are failed # 77% success rate
- 3- VAFB SLC 4E : is a moderate site in terms of launching SpaceX's rockets and it has 22 trials, 17 of them are successful and 05 of them are failed # 77% success rate

Payload vs. Launch Site



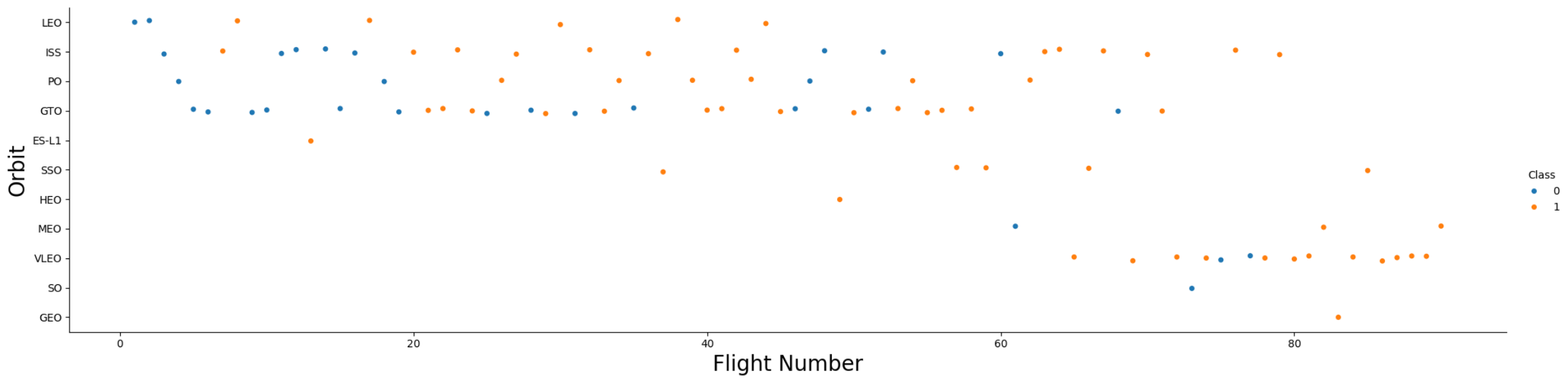
- There is no strong relationship between the payload mass and the success of first stage return since there are approximately equivalent numbers of failed and successful trial

Success Rate vs. Orbit Type



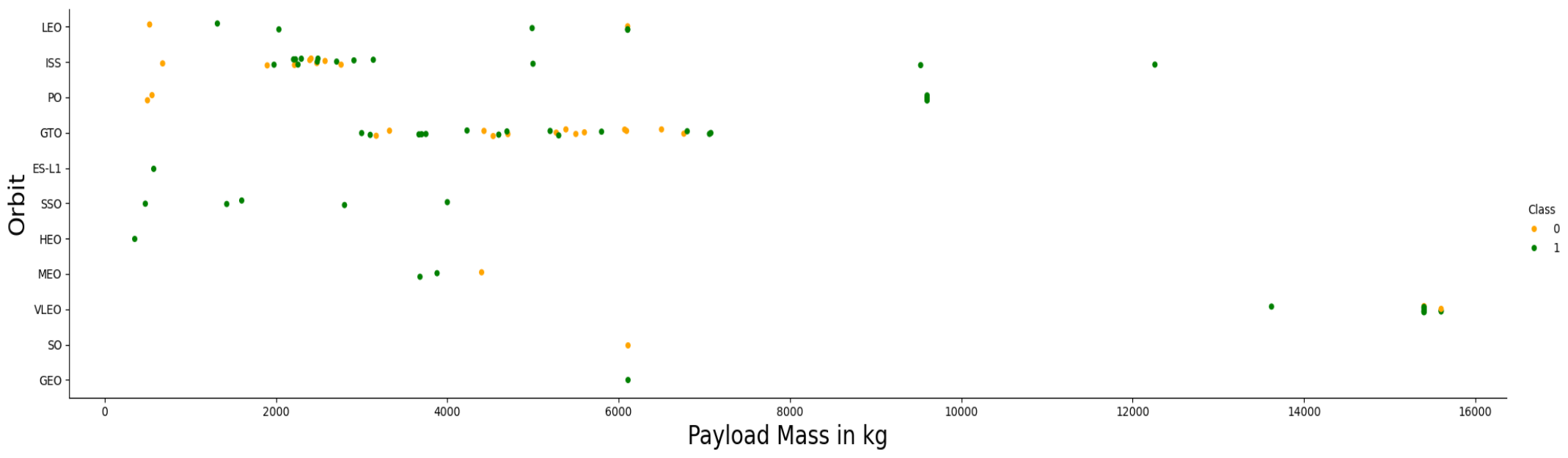
- The best orbits in terms of successful first stage return are ['ES-L1', 'GEO', HEO, SSO]
- and the worst orbit is 'GTO' , which we have to understand why it is the worst to avoid the failure of first stage return.

Flight Number vs. Orbit Type



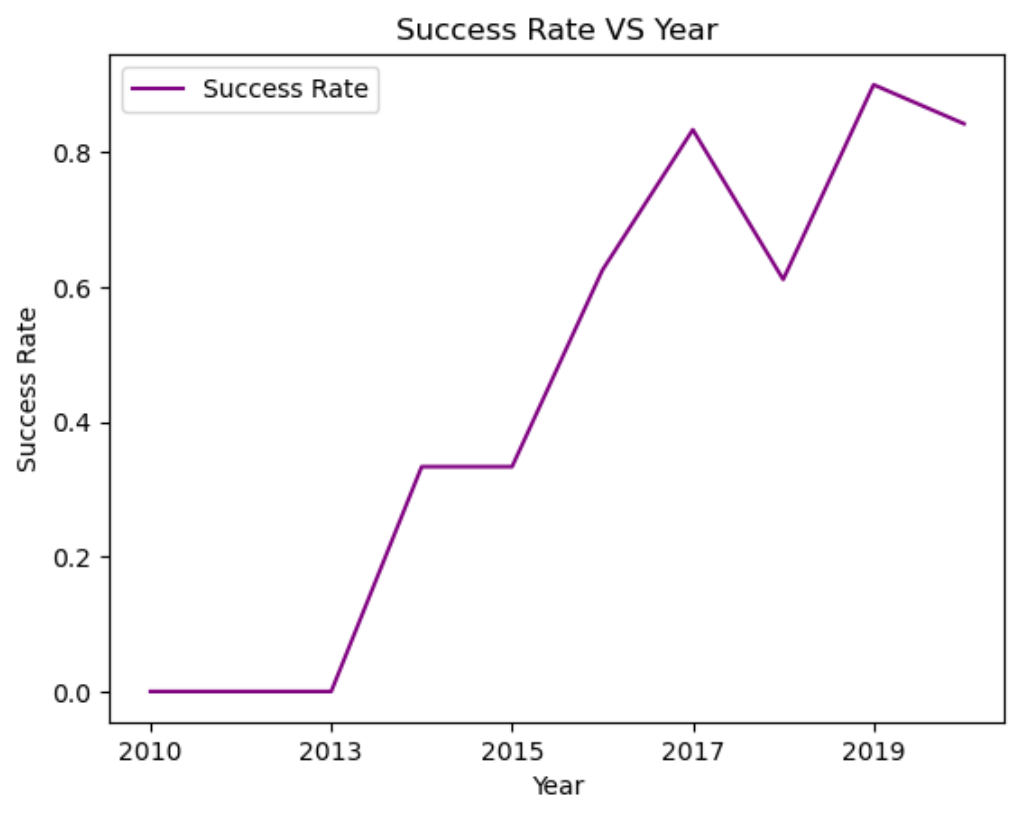
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

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

- 4 distinct sites for rockets launches.

Launch Site Names Begin with 'CCA'

```
In [21]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[21]:

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

- We used the query above to display 5 records where launch sites begin with
- `CCA`

Total Payload Mass

```
In [22]: %sql select sum(payload_mass__kg_) from SPACEXTABLE where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[22]: sum(payload_mass__kg_)
         45596
```

- The total amount of payload that moved to the outer space by NASA through SpaceX rockets equals vto 45596 Kg

Average Payload Mass by F9 v1.1

```
In [23]: %sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTABLE where booster_version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[23]:
```

avg_mass_F9
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928 kg

First Successful Ground Landing Date

```
In [25]: %sql select min(DATE) from SPACEXTABLE where landing_outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[25]: min(DATE)
         2015-12-22
```

- Date of the first successful landing outcome on ground pad was in 22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

In [26]:
```
%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.
```

Out[26]:

Booster_Version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

- The boosters which have success in drone ship landing with payload between 4000 and 6000 kg are :
  - F9 FT B1029.1
  - F9 FT B1036.1
  - F9 B4 B1041.1

Total Number of Successful and Failure Mission Outcomes

```
In [28]: %sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
 * sqlite:///my_data1.db
Done.
```

Out[28]:

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

- Total number of successful and failure mission outcomes are 100 and 1 respectively

Boosters Carried Maximum Payload

```
In [30]: %sql select distinct booster_version from SPACEXTBL\
         where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

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

Out[30]:

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

In [32]:
```
%sql select landing_outcome, booster_version, launch_site from SPACEXTBL\
where (landing_outcome = 'Failure (drone ship)' and date like '2015%')
```

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

Out[32]:

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

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

In [33]:
```
%sql select landing_outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL\
where DATE between '2010-06-04' and  '2017-03-20' group by landing_outcome\
order by count(landing_outcome) desc
```

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

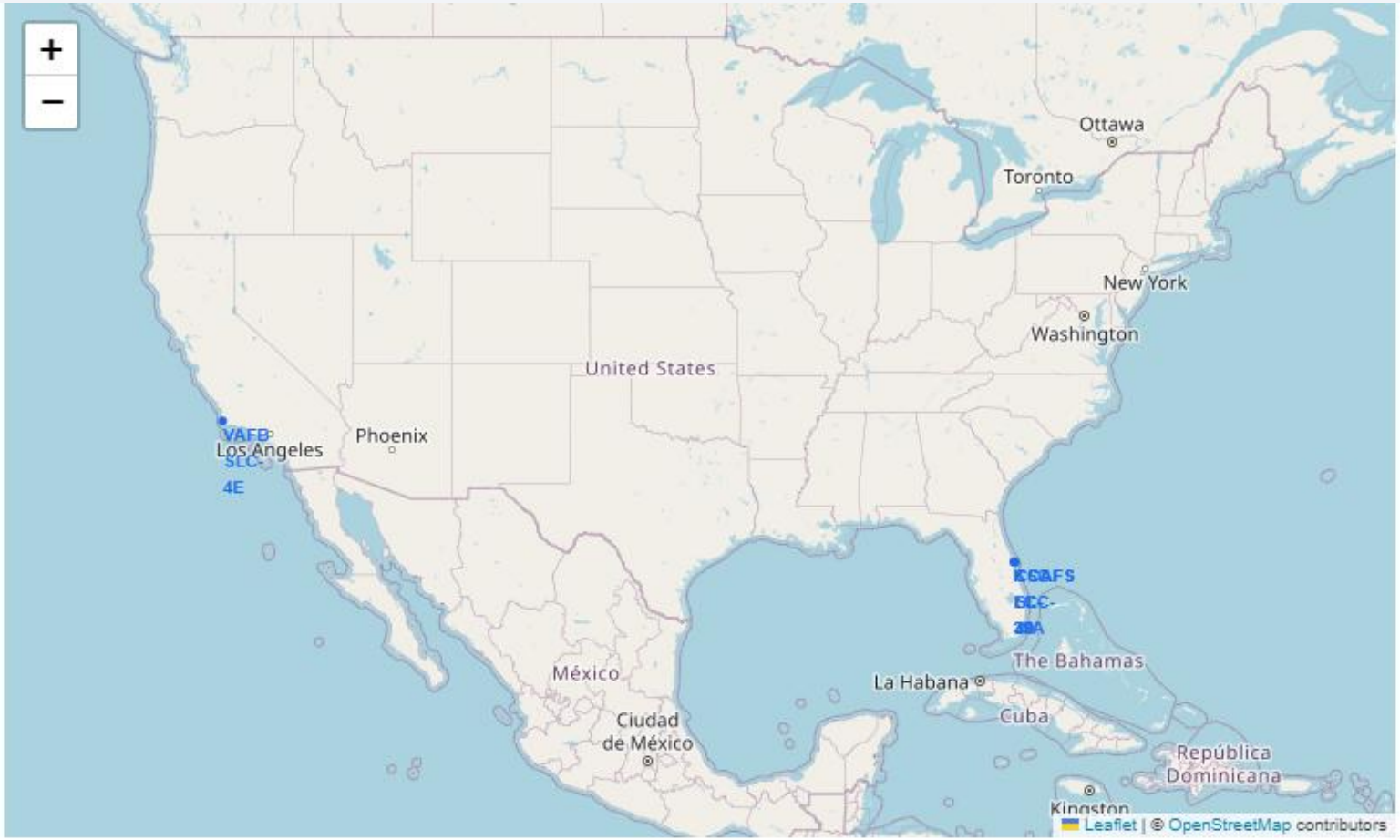
Out[33]:

Landing_Outcome	counts_of_landing_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

Section 3

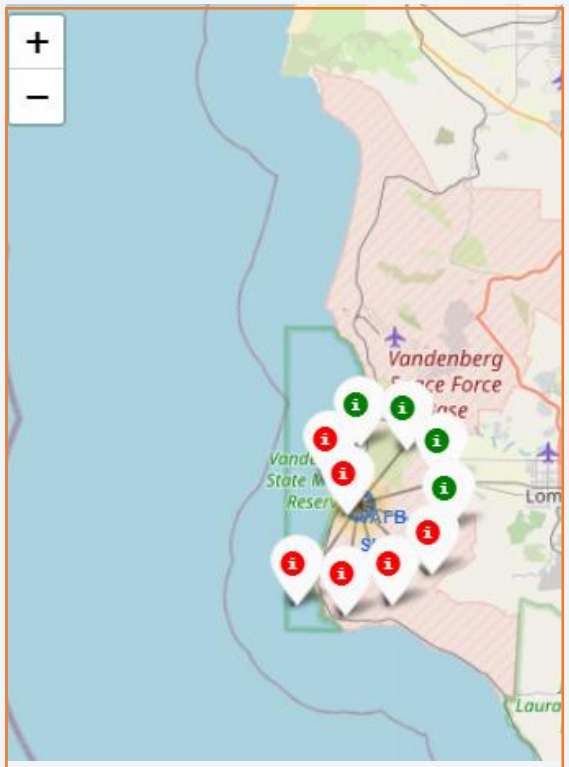
Launch Sites Proximities Analysis

All Launch Sites

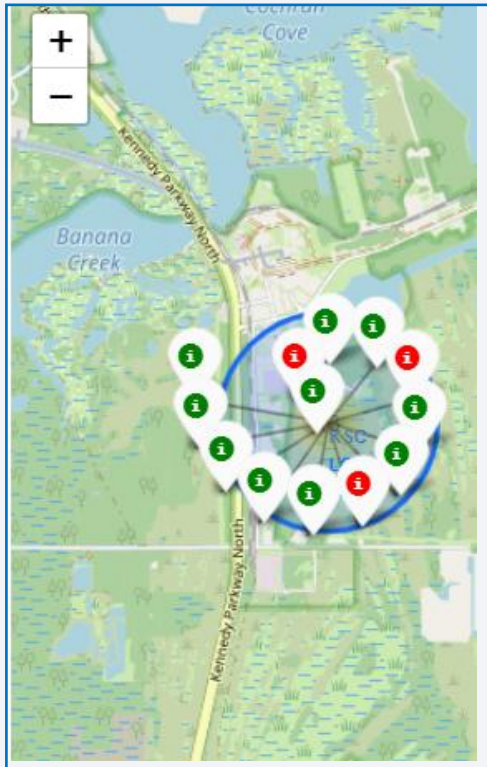


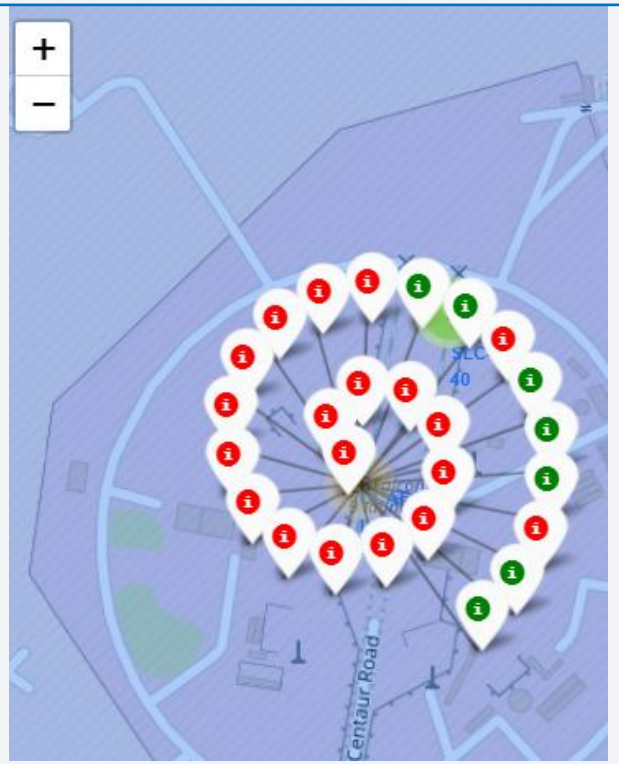
- All launch sites are in proximity to the Equator line
- All launch sites are in very close proximity to the coast

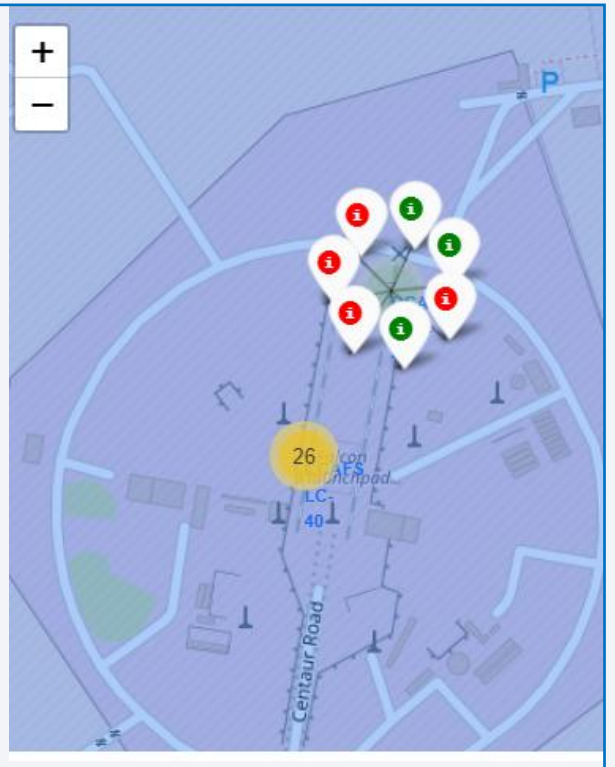
The Success/Failed Launches for Each Site



- California



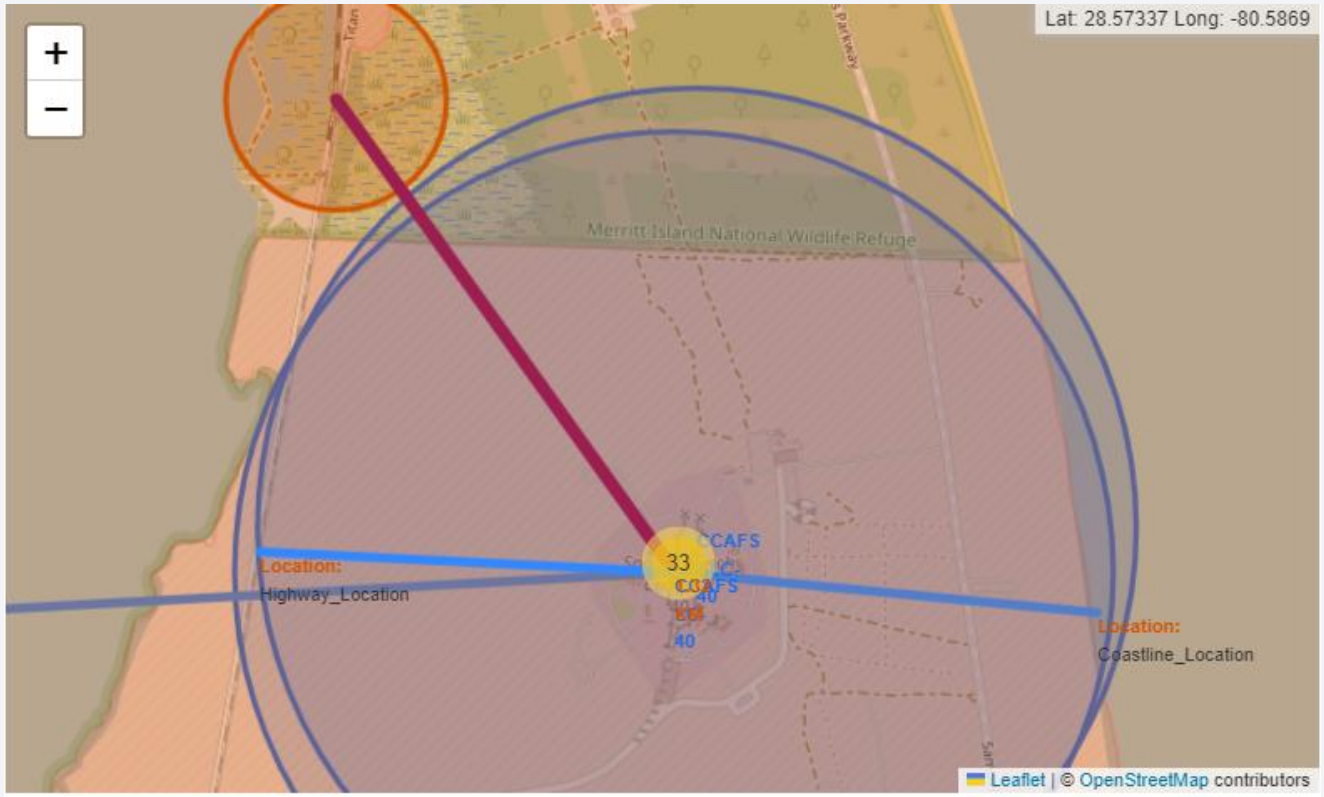




- Florida

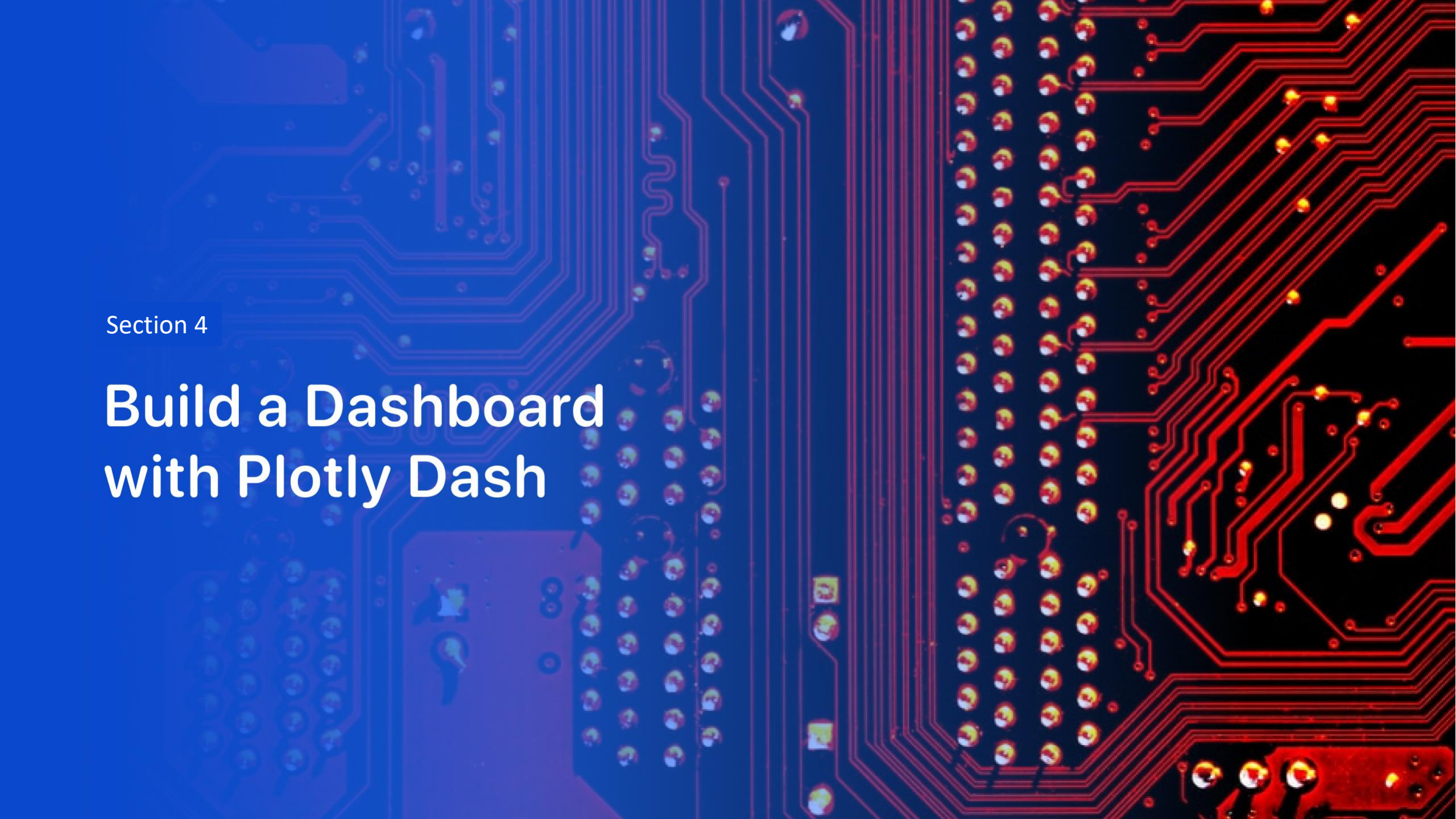
- Green Marker = Successful Return
- Red Marker = Failed Return

The Distances Between a Launch Site to Its Proximities

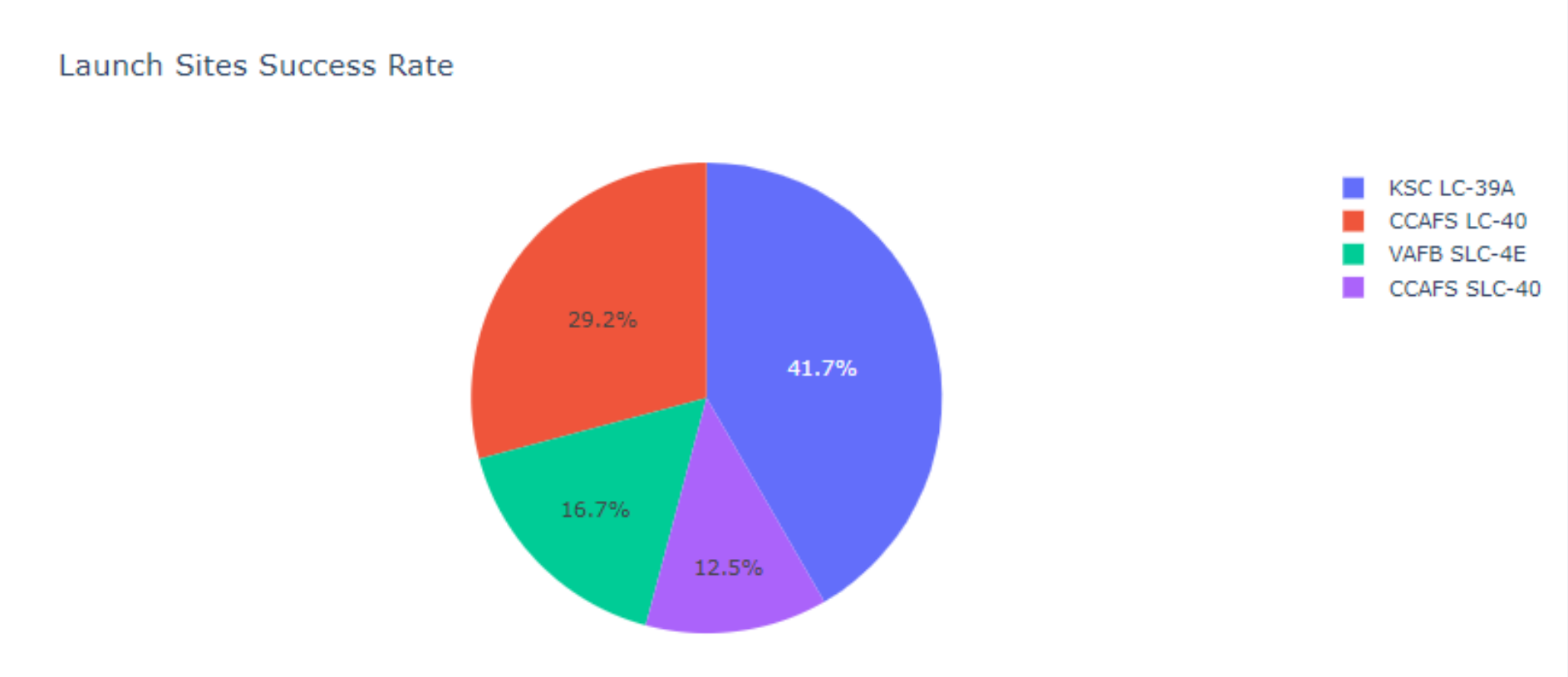


The distances between the launch site (CCAFS LC-40) to its proximities

- Orlando City Distance ≈ 78.8 Km,
- Coastline Distance ≈ 0.97 Km,
- Highway Distance ≈ 0.95Km

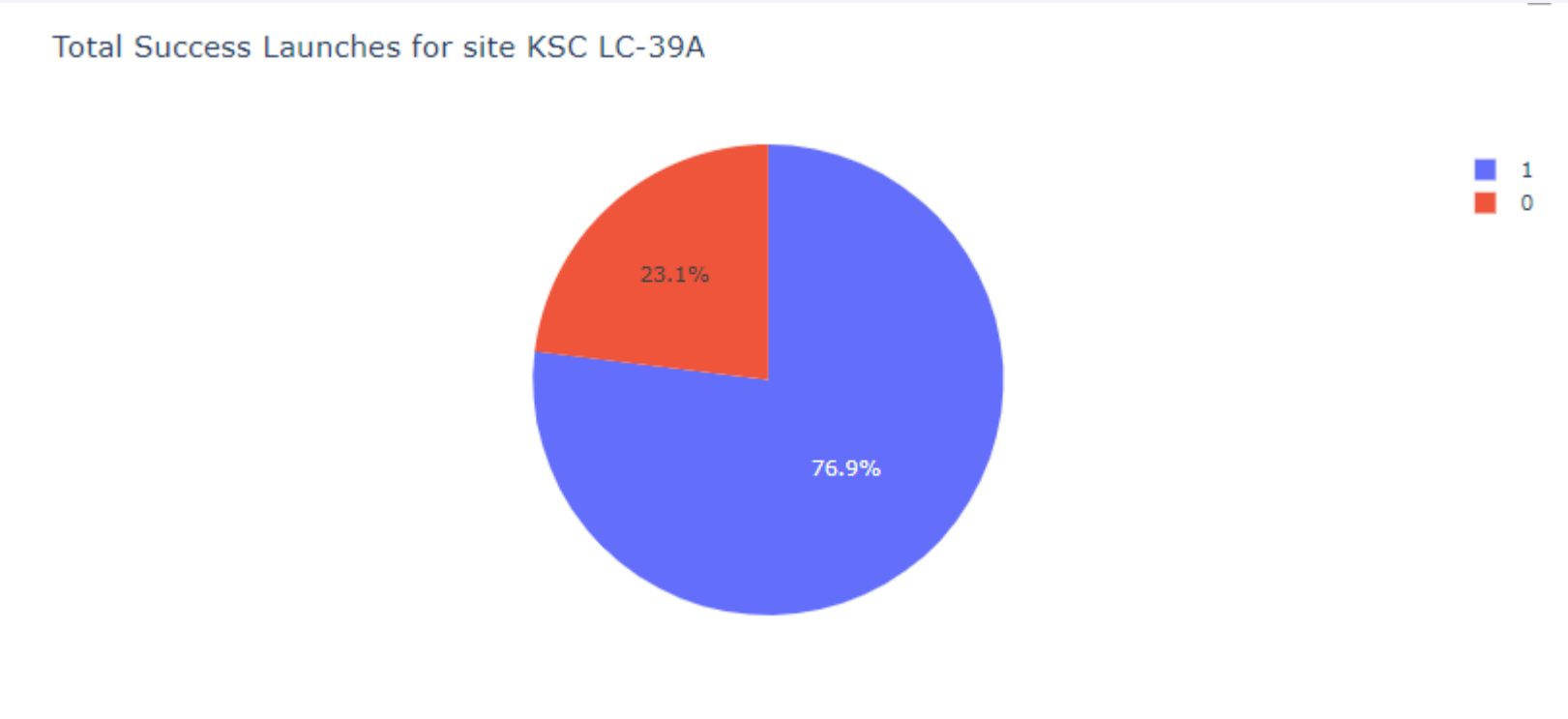

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



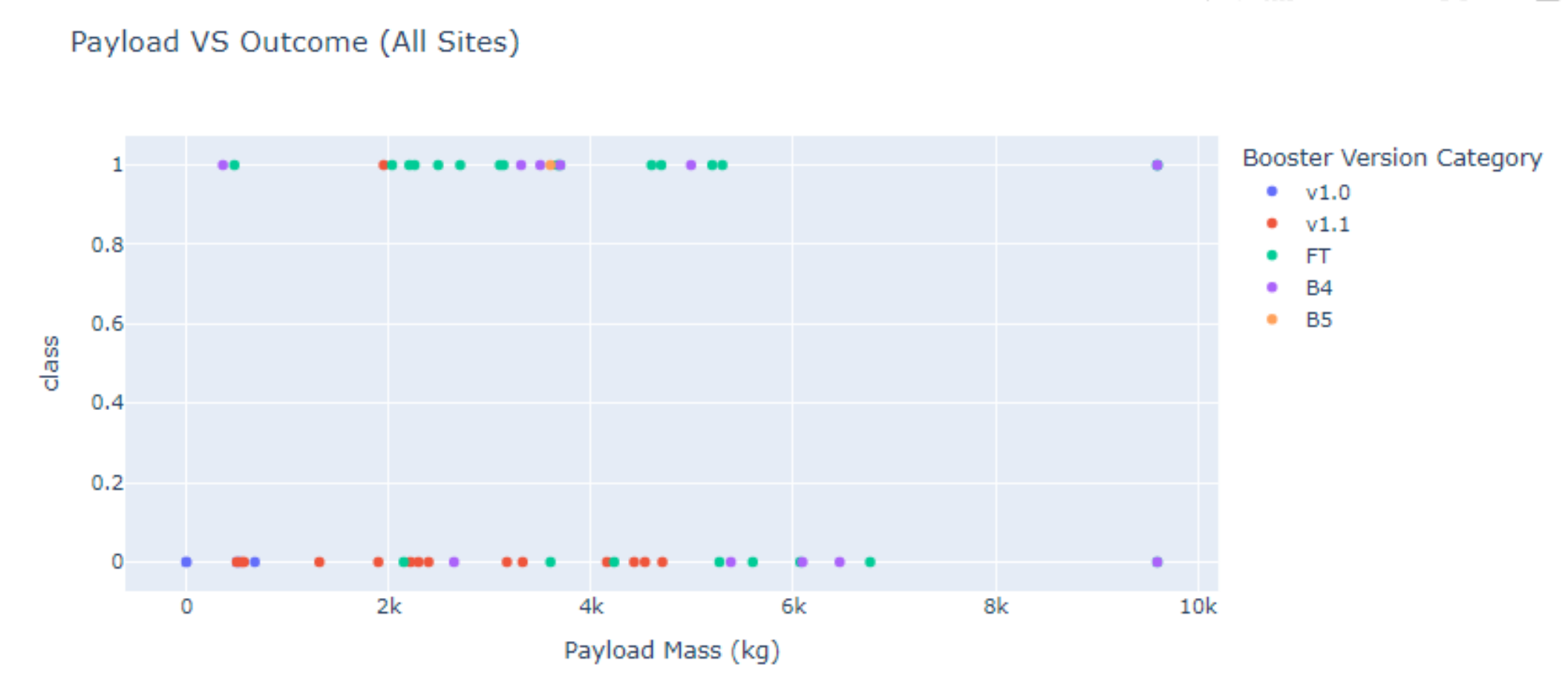
- KSC LC-39A with 41.7% has the highest successful launches

Launch success for KSC LC 39A



- • KSC LC-39A with 76.9% successful launches
- • KSC LC-39A with 23.1% failed launches

<Dashboard Screenshot 3>



- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

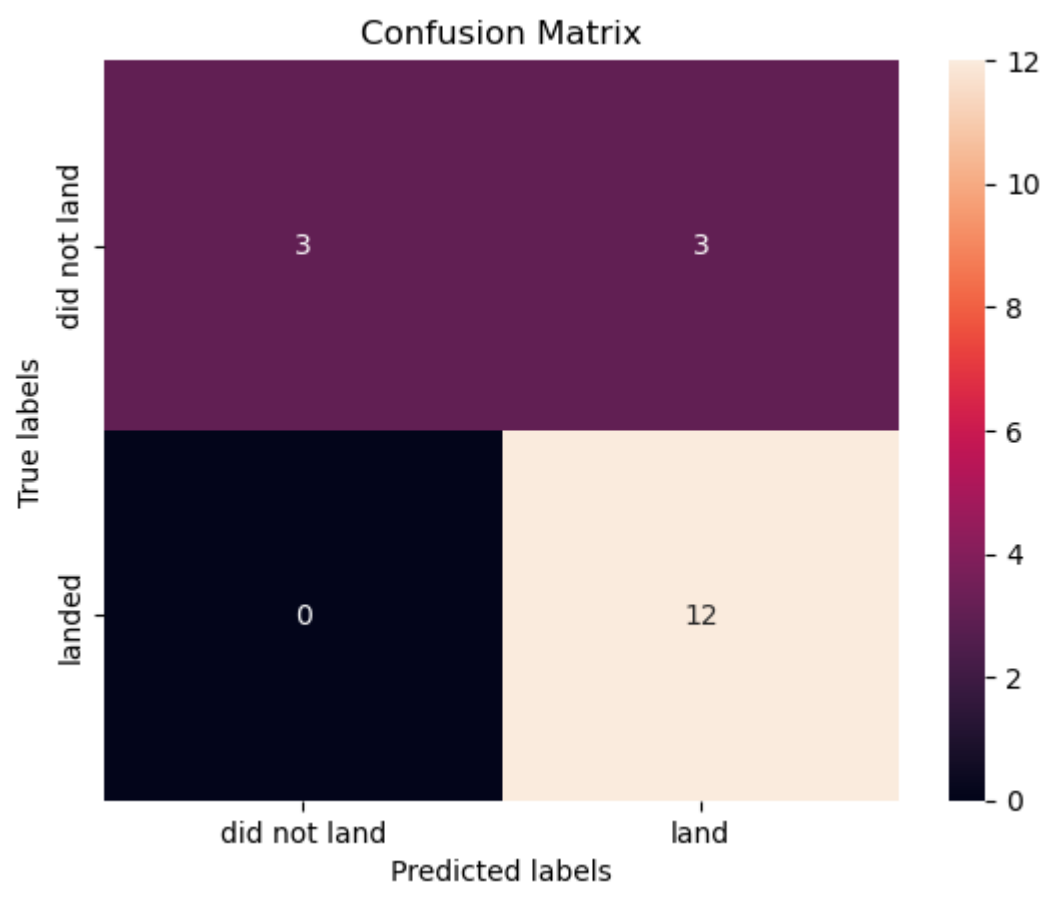
Predictive Analysis (Classification)

Classification Accuracy

	Jaccard Score	F1 Score	Accuracy
Logistic Regression	0.8000	0.777778	0.833333
SVM	0.8000	0.777778	0.833333
Decision Tree	0.6875	0.629630	0.722222
KNN	0.8000	0.777778	0.833333

- **Logistic Regression**, **SVM**, and **KNN**: All three models have identical and the highest Jaccard Score (0.80), F1 Score (0.777778), and Accuracy (0.833333).
- **Decision Tree**: This model shows lower performance across all three metrics compared to the others.

Confusion Matrix



- The confusion matrix for the decision tree classifier indicates its ability to effectively distinguish between different classes.
- A significant issue identified is the presence of false positives, where unsuccessful landings are incorrectly classified as successful.

Conclusions

- **Launch Site Flight Volume**: A higher flight volume at a launch site correlates with a greater success rate.
- **Temporal Trend**: Launch success rates exhibited an upward trend from 2013 to 2020.
- **Optimal Orbits**: ES-L1, GEO, HEO, SSO, and VLEO orbits demonstrated the highest success rates.
- **Leading Launch Site**: KSC LC-39A recorded the most successful launches among all sites.
- **Best Performing Algorithm**: The Decision Tree classifier proved to be the most effective machine learning algorithm for this particular task.

Appendix

- SpaceX API URL
- Wikipedia

Thank you!