IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Oliveira
03FEB2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of methodologies

- SpaceX Data Collection using SpaceX API
- SpaceX Data Collection with Web Scraping
- SpaceX Data Wrangling
- SpaceX Exploratory Data Analysis using SQL
- Space-X EDA DataViz Using Python Pandas and Matplotlib
- Space-X Launch Sites Analysis with Folium/Interactive Visual Analytics and Plotly Dash
- SpaceX Machine Learning Landing Prediction

Summary of all results

- EDA results
- Interactive Visual Analytics and Dashboards
- Predictive Analysis(Classification)

Introduction

- Project background and context

- On its website, SpaceX promotes Falcon 9 rocket launches at a price of $62 million, a significant contrast to other providers whose costs soar to at least $165 million per launch. The substantial savings stem largely from SpaceX's innovative practice of reusing the first stage. Consequently, by assessing the likelihood of a successful first stage landing, one can gauge the overall cost of a launch. This knowledge becomes valuable for potential competitors seeking to bid against SpaceX in the rocket launch market.

- Problems you want to find answers

- In this capstone project, our objective is to forecast the successful landing of the Falcon 9 first stage by analyzing data derived from the Falcon 9 rocket launches as presented on SpaceX's official website.

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

- Description of how SpaceX Falcon9 data was collected.
  - Initially, data was gathered through the SpaceX API, a RESTful API, by executing a GET request to the SpaceX API URL. To facilitate this, a set of helper functions were defined to interact with the API efficiently, extracting pertinent information using identification numbers present in the launch data. Subsequently, the rocket launch data was retrieved and parsed using the GET request, with the response content decoded as a JSON result. This JSON result was then converted into a Pandas data frame to ensure uniformity and ease of analysis.
  - Additionally, to enhance the dataset, web scraping was employed to acquire historical Falcon 9 launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." Using BeautifulSoup and request Libraries, the Falcon 9 launch HTML table records were extracted from the Wikipedia page. The retrieved HTML data was then parsed, and the relevant information was converted into another Pandas data frame for comprehensive data integration and analysis.

Data Collection – SpaceX API

- The data collection process involved utilizing the SpaceX API, a RESTful API. This was achieved by initiating a GET request to the SpaceX API, followed by the retrieval and parsing of SpaceX launch data through the same GET request. The response content, obtained in JSON format, was then decoded and converted into a Pandas data frame.
- Here is the GitHub link to the finished notebook containing the SpaceX API calls:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

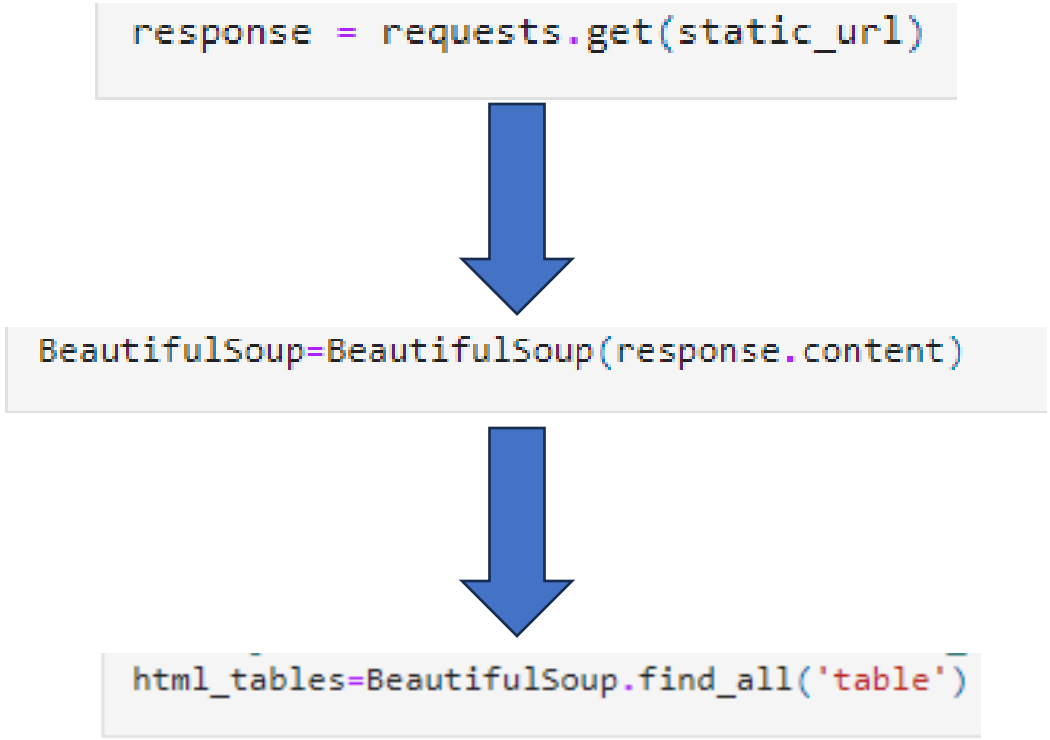
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
json_data=response.json()
data=pd.json_normalize(json_data)
```

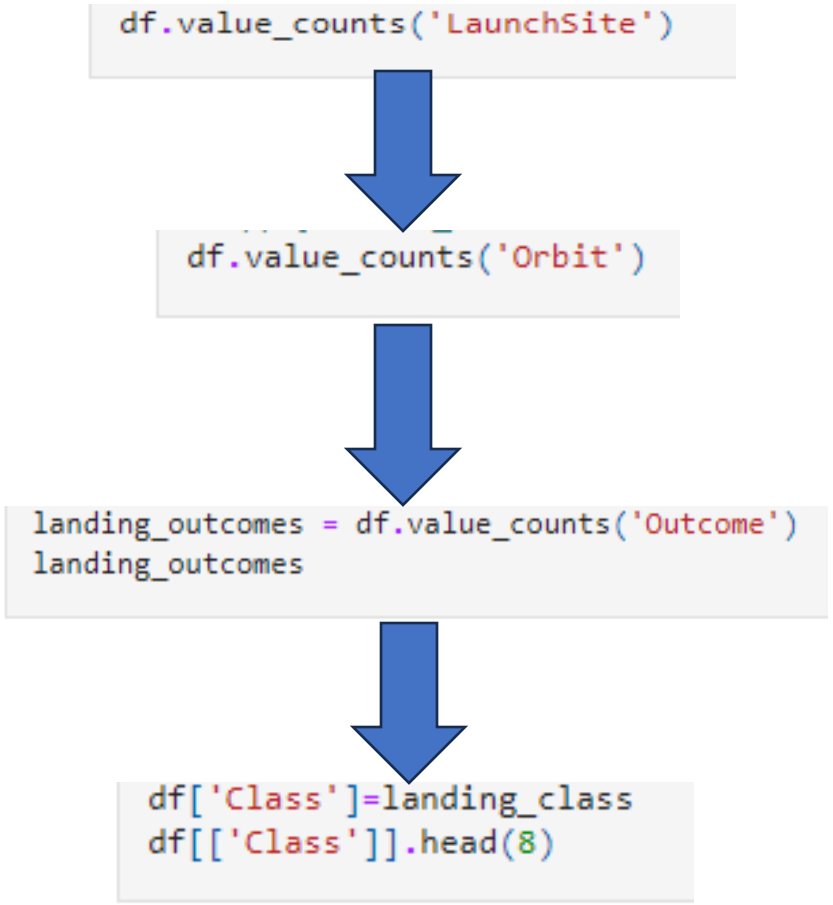
Data Collection - Scraping

- Conducted web scraping to gather historical Falcon 9 launch records from Wikipedia, employing BeautifulSoup and requests. After the extraction of the data, a data frame was created by parsing the launch HTML.
- Here is the GitHub link to the finished notebook containing the SpaceX Web Scrapping:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Prior to engaging in data wrangling, following the acquisition and creation of a Pandas DataFrame from the gathered data in the SpaceX API notebook, a series of steps were taken. The data was filtered based on the Booster Version column to retain only Falcon 9 launches. Subsequently, attention was directed towards addressing missing values in the Landing Pad and Payload Mass columns. For the Payload Mass column, the missing data points were substituted with the mean value of the column to ensure a more complete and representative dataset.
- During the data wrangling phase, several essential tasks were accomplished:
  - Calculated the count of launches on each site.
  - Computed the number and frequency of each orbit.
  - Determined the count and frequency of mission outcomes corresponding to the orbits.
  - Formulated a landing outcome label derived from the Outcome column.
- Here is the GitHub link to the finished notebook containing the SpaceX Data Wrangling:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- In the exploratory data analysis, the following actions were undertaken using Pandas, Seaborn, and Matplotlib:
  - Utilized scatter plots to visually represent relationships between:
    - Flight Number and Launch Site
    - Payload Mass and Launch Site
    - Flight Number and Payload Mass
    - Flight Number and Orbit Type
    - Payload and Orbit Type
  - Employed bar charts to visualize the success rate of each orbit type.
  - Utilized a line plot to depict the yearly trend in launch success, offering a comprehensive view of launch outcomes over time.
- OneHotEncoder was applied to the columns 'Orbits', 'LaunchSite', 'LandingPad', and 'Serial'.
- Here is the GitHub link to the finished notebook containing the SpaceX Data Visualization:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The following SQL queries were performed for EDA:
- Retrieve the names of unique launch sites in the space mission.

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

- Display five records where launch sites begin with the string 'CCA.'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

- Present the total payload mass carried by boosters launched by NASA (CRS).

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

- Show the average payload mass carried by booster version F9 v1.1.

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

EDA with SQL (cont.)

- Provide the date when the first successful landing outcome on the ground pad was achieved.

```
%sql select min(date) from SPACEXTABLE where Landing_Outcome = 'Success'
```

- List the names of boosters that achieved success on a drone ship and had a payload mass greater than 4000 but less than 6000.

```
%sql select "Booster_Version" from SPACEXTABLE where Mission_Outcome = 'Success' and PAYLOAD_MASS__KG_>4000 and PAYLOAD
```

- Display the total number of successful and failure mission outcomes.

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE GROUP BY Mission_Outcome;
```

- List the names of booster versions that have carried the maximum payload mass, utilizing a subquery.

```
%sql select booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

EDA with SQL (cont.)

- Provide records displaying the month names, failure landing outcomes in a drone ship, booster versions, and launch sites for the months in the year 2015.

```
%sql SELECT Landing_Outcome, COUNT(*) AS OutcomeCount FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
```

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

```
%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, booster_version, Launch_Site FROM SPACEXTABLE WHERE substr(Da
```

- Here is the GitHub link to the finished notebook containing the SpaceX EDA SQL:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

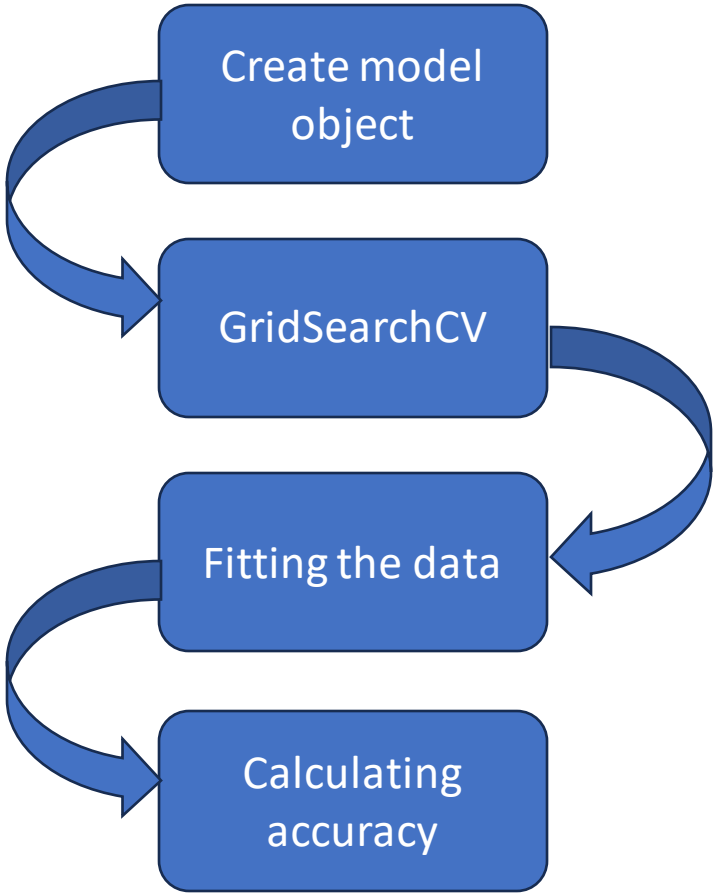
- To mark the launch sites, circles and marker clusters were created.
- Markers were color-labeled in order to identify which launch sites have higher success rate.
- Lines were also created to see the distance between the launch sites and the coastline, highways or cities
- Here is the GitHub link to the finished notebook containing the SpaceX Folium Map:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Created an interactive dashboard application using Plotly Dash, incorporating the following features:
- Integrated a Launch Site Drop-down Input Component to facilitate site selection.
- Implemented a callback function to dynamically render a success-pie-chart based on the chosen launch site from the dropdown.
- Introduced a Range Slider to enable payload selection.
- Incorporated a callback function to dynamically render the success-payload-scatter-chart scatter plot based on the chosen payload range from the slider.
- Here is the GitHub link to the finished notebook containing the SpaceX Dash:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Performed testing using various machine learning models, including Support Vector Machines (SVM), Classification Trees, k-Nearest Neighbors, and Logistic Regression, following these steps:
- Created an object for each algorithm and assigned a set of parameters for each model.
- Established a GridSearchCV object with cv=10 and fit the training data into it to find the best hyperparameters.
- Presented the GridSearchCV object for each model, displaying the best parameters using the data attribute best_params_ and showcasing the accuracy on the validation data through the data attribute best_score_.
- Utilized the score method to calculate the accuracy on the test data for each model and generated confusion matrices for visual representation of test and predicted outcomes.
- Compiled a table displaying the test data accuracy scores for each method, enabling a comparison to determine the best-performing model among SVM, Classification Trees, k-Nearest Neighbors, and Logistic Regression.
- Here is the GitHub link to the finished notebook containing the SpaceX Predictive Analysis:
  - https://github.com/JoaoOliveira2707/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

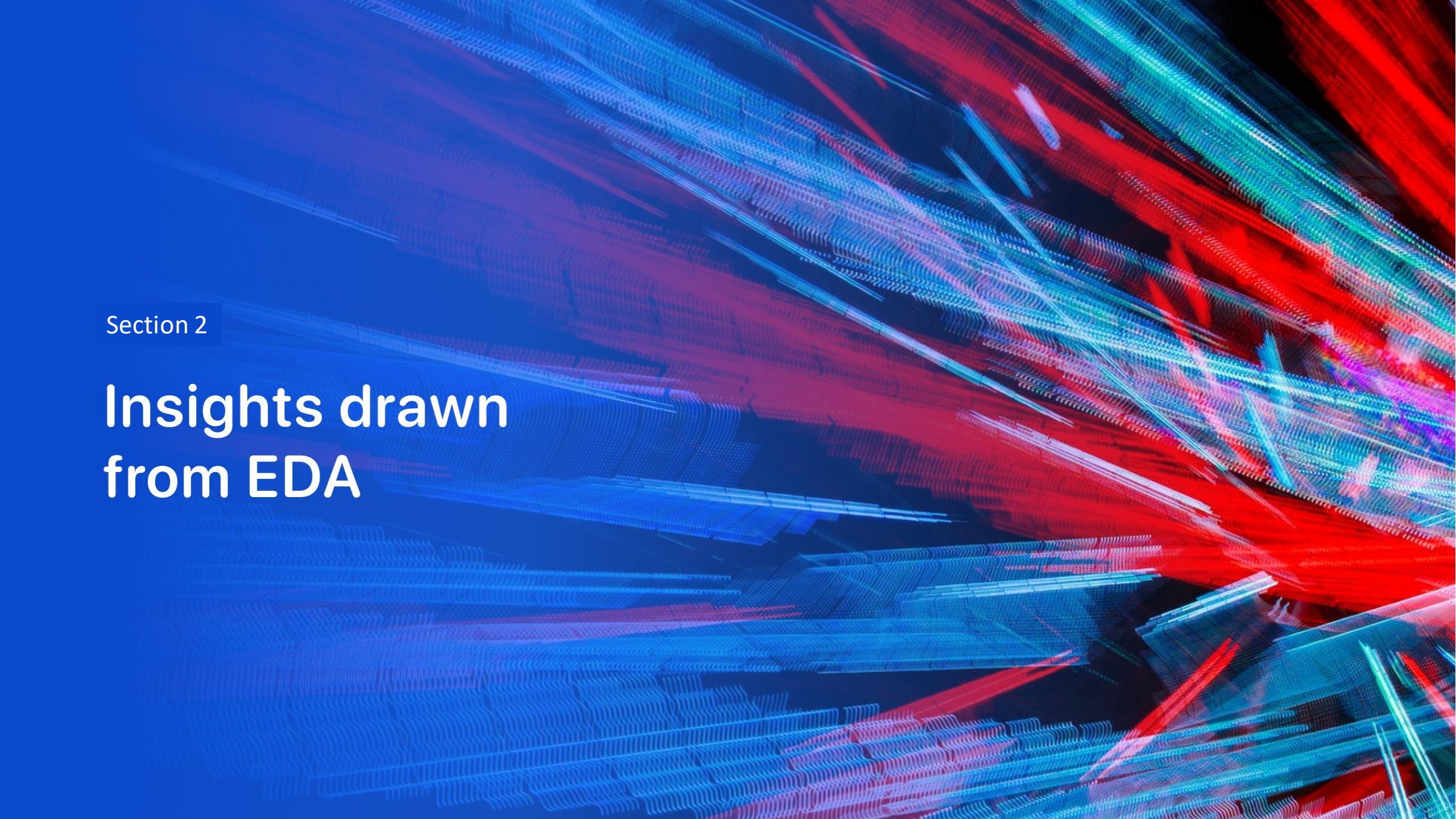
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

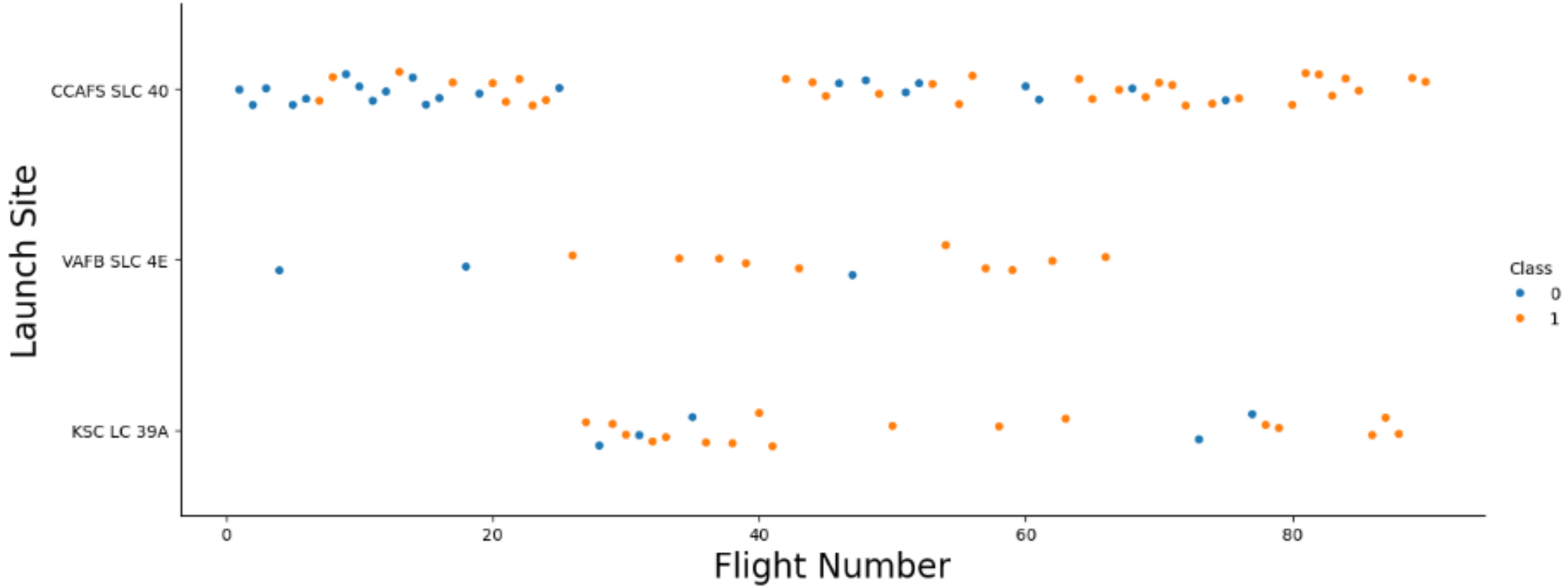
Section 2

Insights drawn from EDA



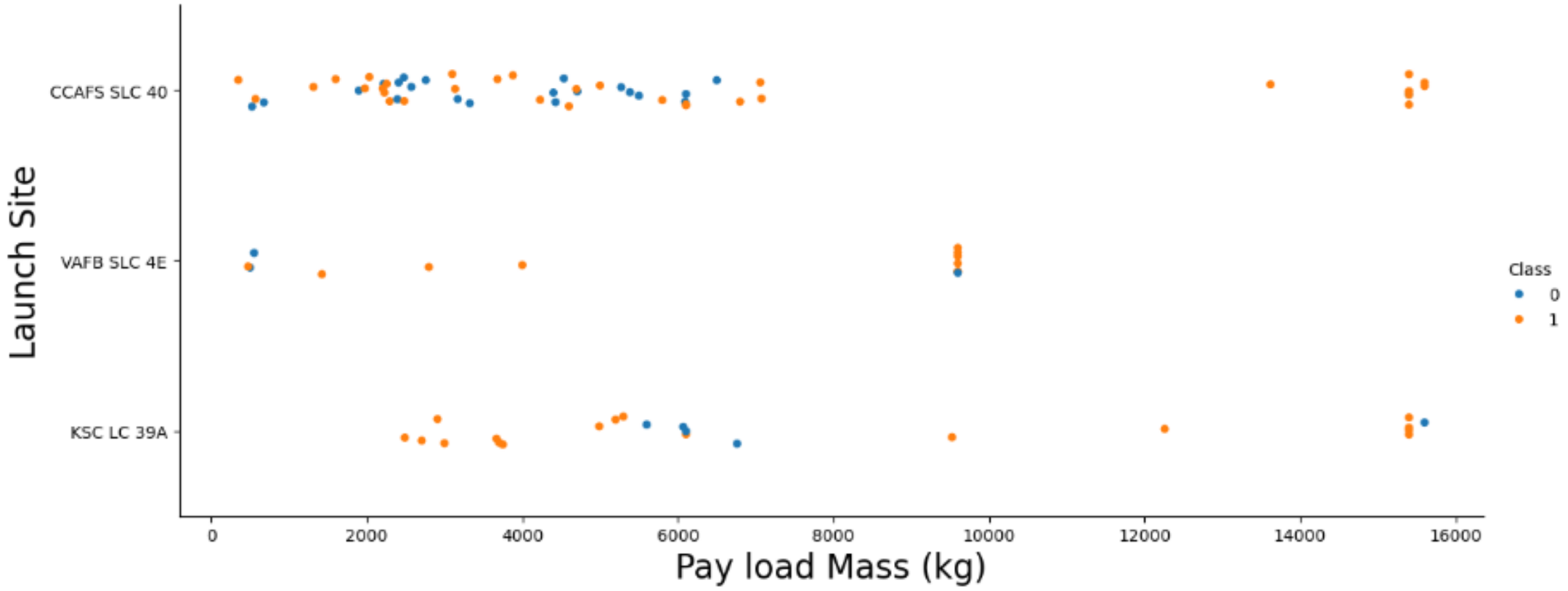
Flight Number vs. Launch Site

- In this plot, it is observed that earlier flights had a lower success rate. In contrast, later flights exhibited a higher success rate. About half of the launches took place at the CCAFS SLC 40 launch site. The VAFB SLC 4E and KSC LC 39A launch sites showed higher success rates. Overall, the data suggests that newer launches tend to have a higher success rate.



Payload vs. Launch Site

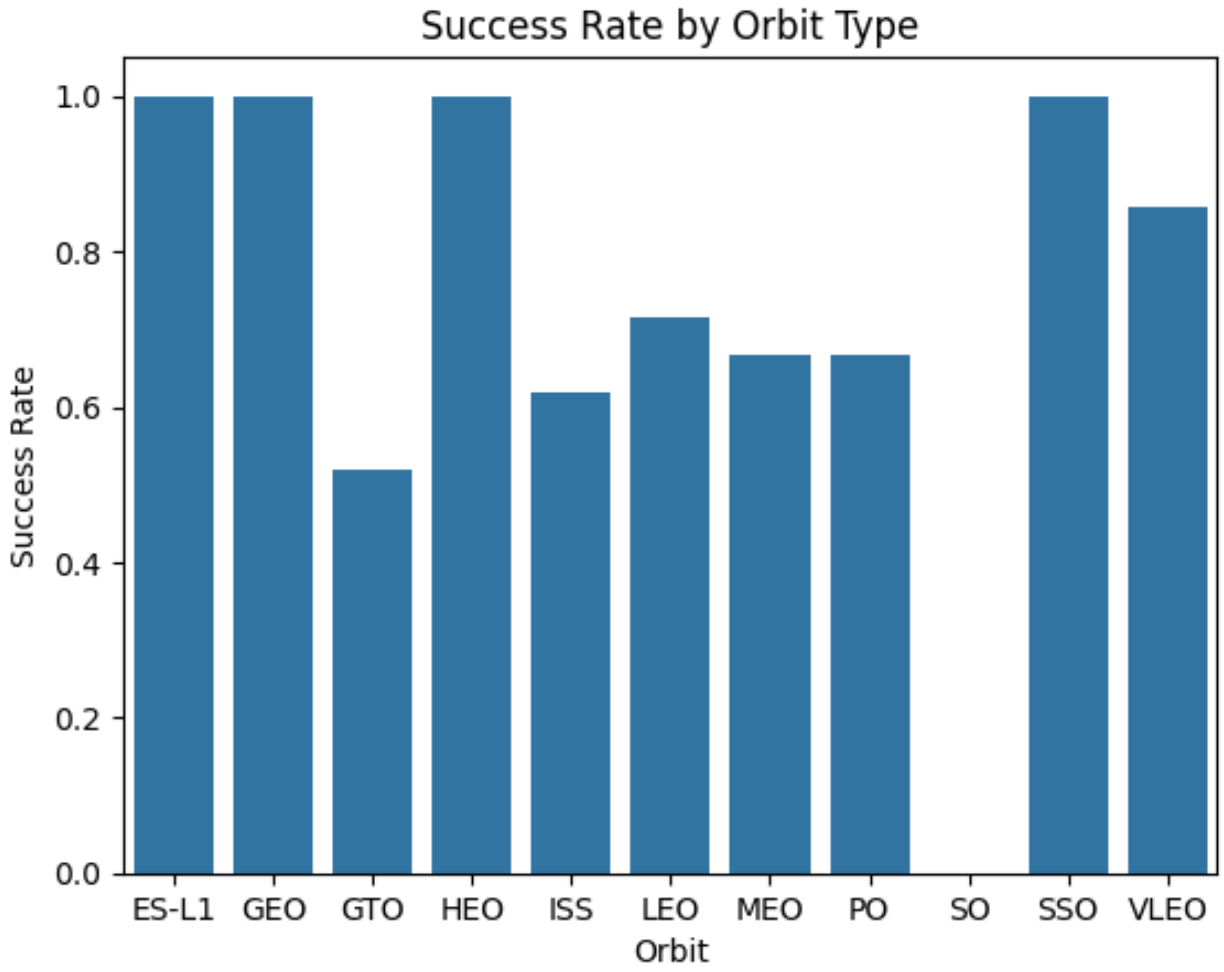
- In this plot, it is evident that a higher payload mass (kg) correlates with a higher success rate. Launches with a payload exceeding 7,000 kg tend to be more successful. Oddly, KSC LC 39A exhibits a 100% success rate for launches with a payload less than 5,500 kg. Additionally, VAFB SLC 4E has not conducted any launches with a payload greater than approximately 10,000 kg.



Success Rate vs. Orbit Type

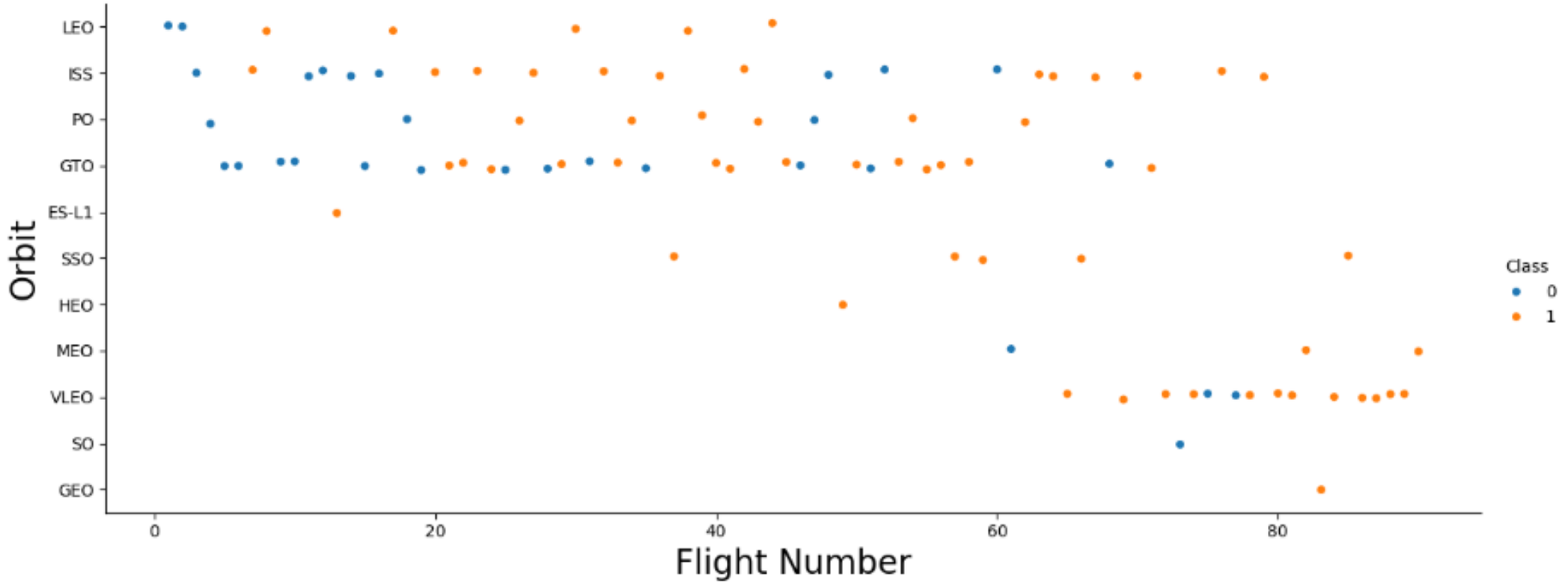
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO

- It must be noticed that the orbit types with 50%-80% Success Rate are the only ones with more than one occurrence, with the exception of SSO.



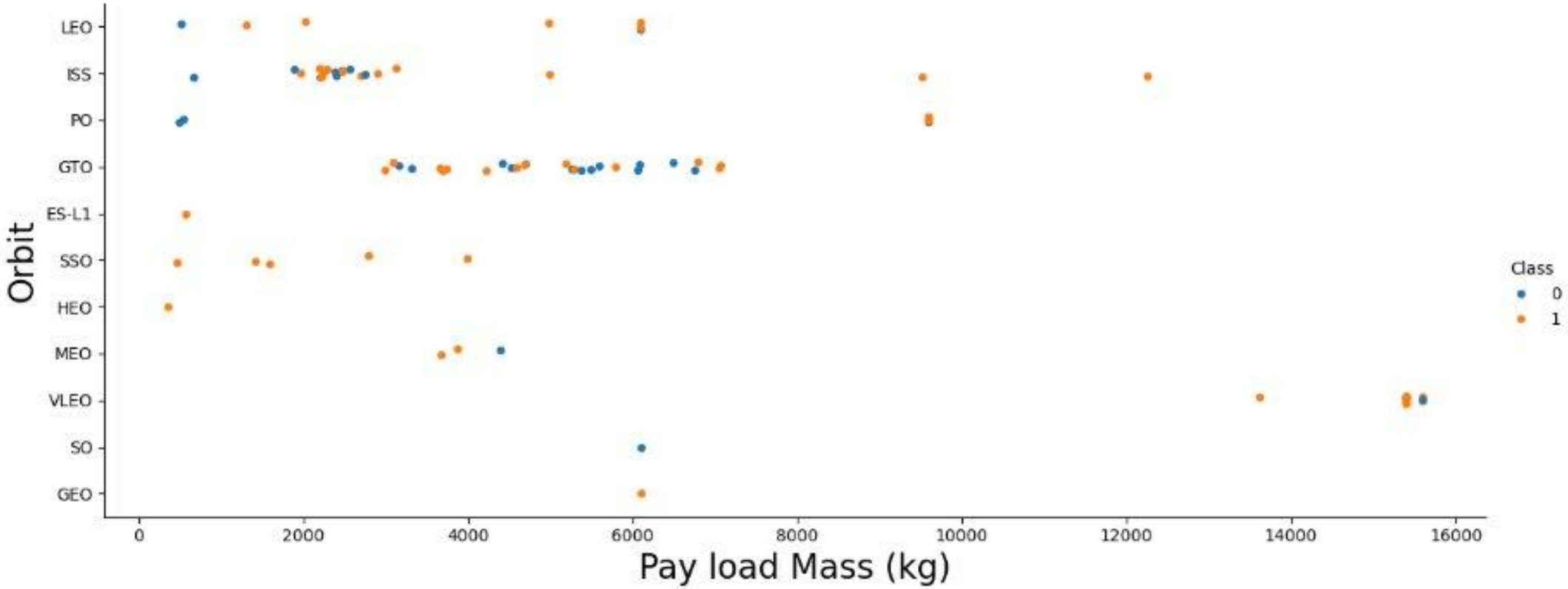
Flight Number vs. Orbit Type

- In this plot, a notable trend is observed where the success rate tends to increase with the number of flights for each orbit. This relationship is particularly evident in the Low Earth Orbit (LEO). However, it's important to note that the Geostationary Transfer Orbit (GTO) appears to not conform to this trend, showing a different pattern in the relationship between success rate and the number of flights.



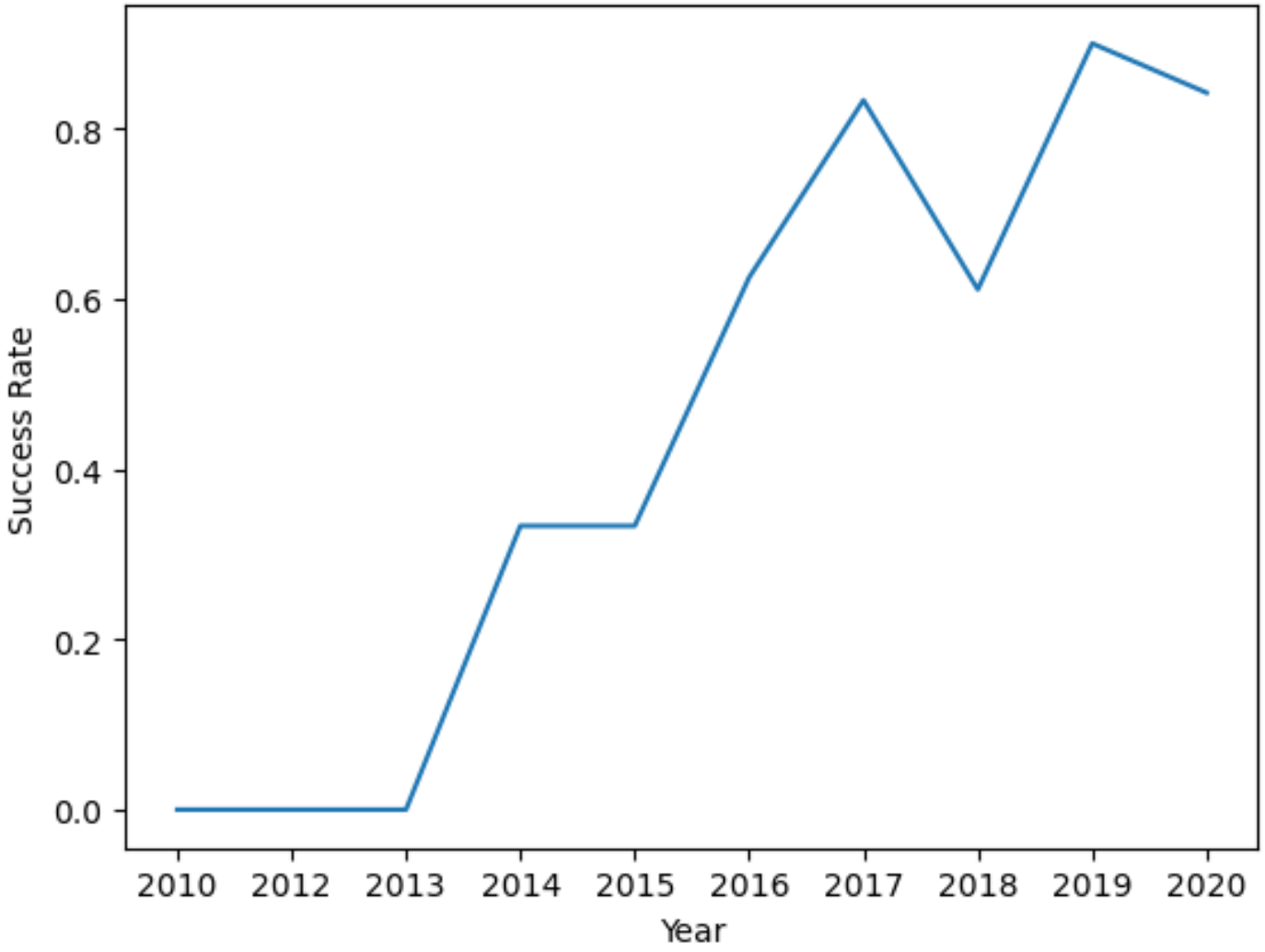
Payload vs. Orbit Type

- In this plot, it is indicated that heavy payloads exhibit better success rates in Low Earth Orbit (LEO), as well as in orbits associated with the International Space Station (ISS) and Polar Orbits (PO). However, the Geostationary Transfer Orbit (GTO) shows mixed success with heavier payloads.



Launch Success Yearly Trend

- In this plot, several temporal trends in the success rates are observed:
- There was an improvement in the success rate during the periods 2013-2017 and 2018-2019.
- Conversely, a decrease in the success rate is observed from 2017-2018 and from 2019-2020.
- Despite these fluctuations, the overall trend indicates an improvement in the success rate since 2013.



All Launch Site Names

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

- There are 4 Launch sites in both East and West Coast Regions

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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
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

- Limit 5 was used to extract the first five records.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- The total payload carried by boosters from NASA is 45596 Kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4 Kg

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

- The date of the first successful landing outcome on ground pad is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select "Booster_Version","Payload" from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' \
and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000
```

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Here are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%sql select Mission_Outcome, count(*) as Total_Missions from SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- There are 12 Boosters which have carried the maximum payload mass:

```
%sql select booster_version from SPACEXTABLE\
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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, 6, 2) AS Month, Landing_Outcome, booster_version, Launch_Site \
FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

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

Here are the failed landing outcomes in drone ship, their booster versions, and launch site names regarding the year 2015

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

- 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

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

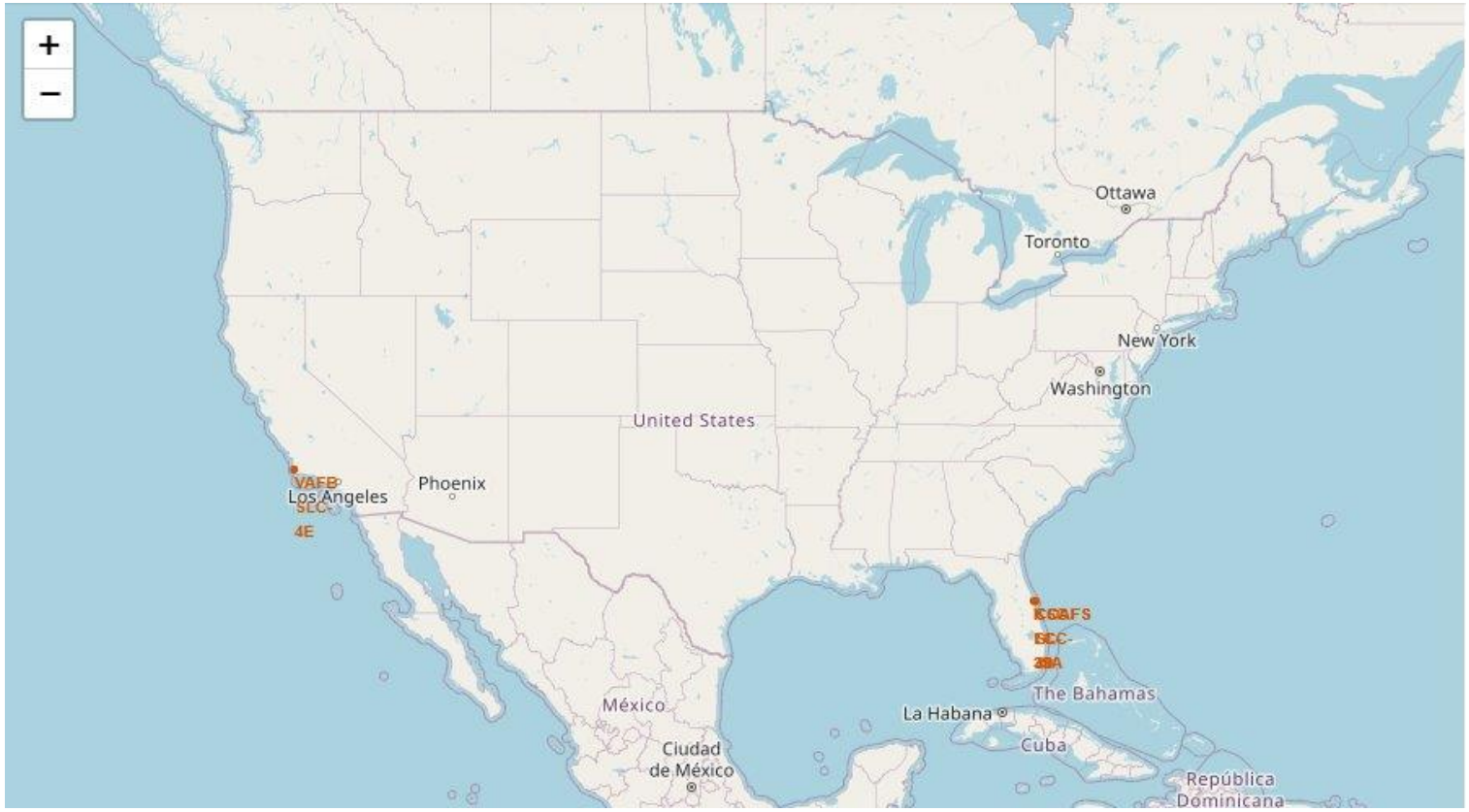
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	OutcomeCount
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

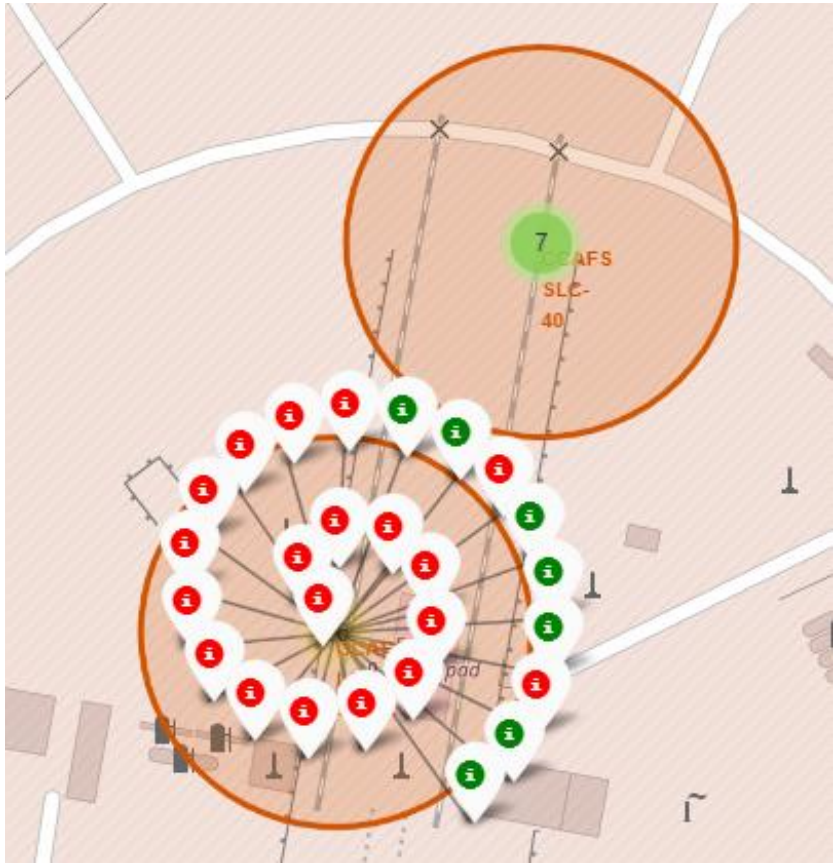
Launch Sites

- The **proximity** of a launch site to the **equator** plays a crucial role in space launches. Launch sites closer to the equator benefit from the **Earth's rotational speed**, providing a natural boost for rockets aiming for equatorial orbits. This phenomenon is particularly **advantageous for prograde orbits**. Rockets launched from near-equatorial sites receive additional **assistance from the Earth's rotation**, reducing the need for extra fuel and boosters and, consequently, contributing to **cost savings** in the launch process.



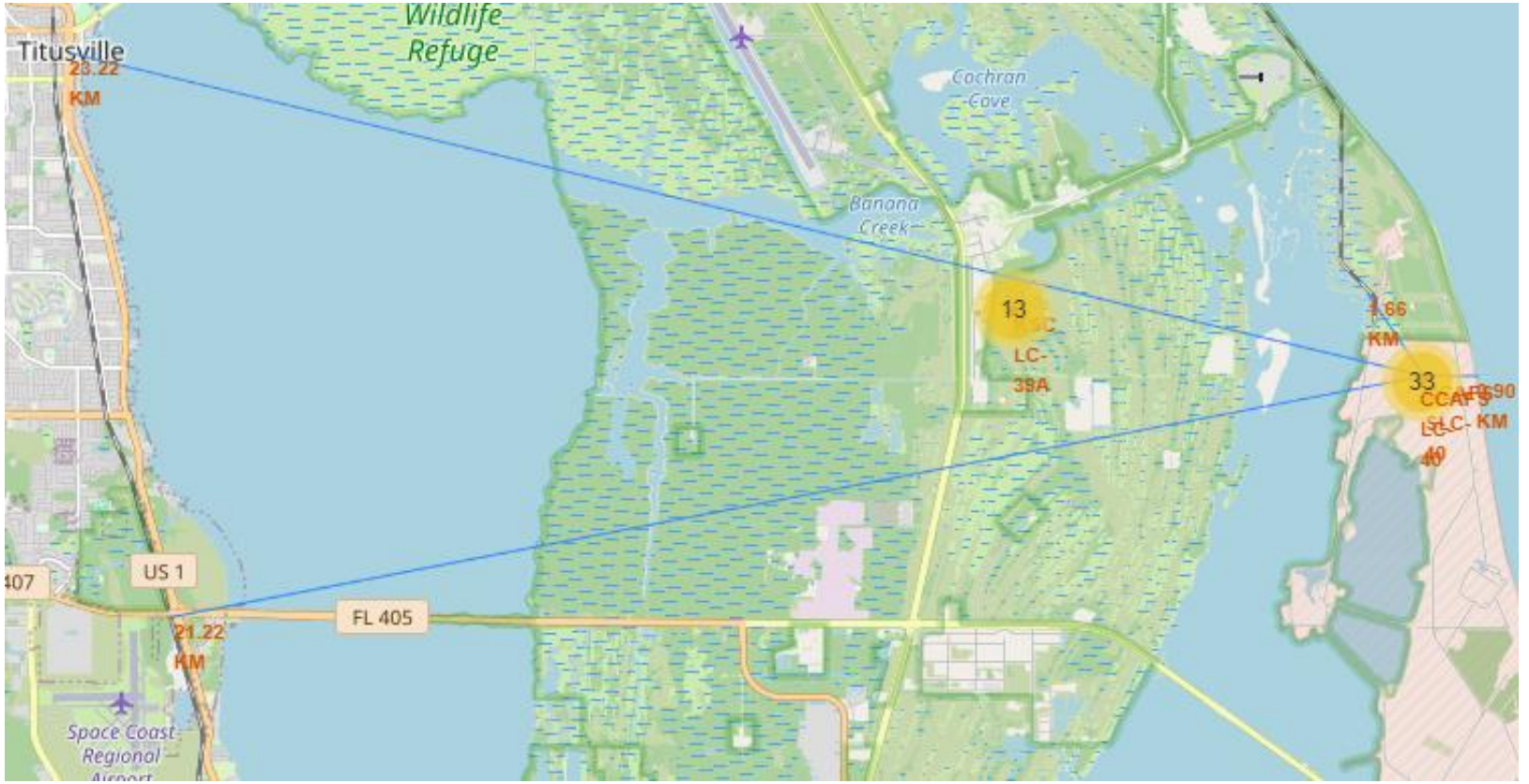
Launch Outcomes

- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS LC-40 has a 7/26 success rate



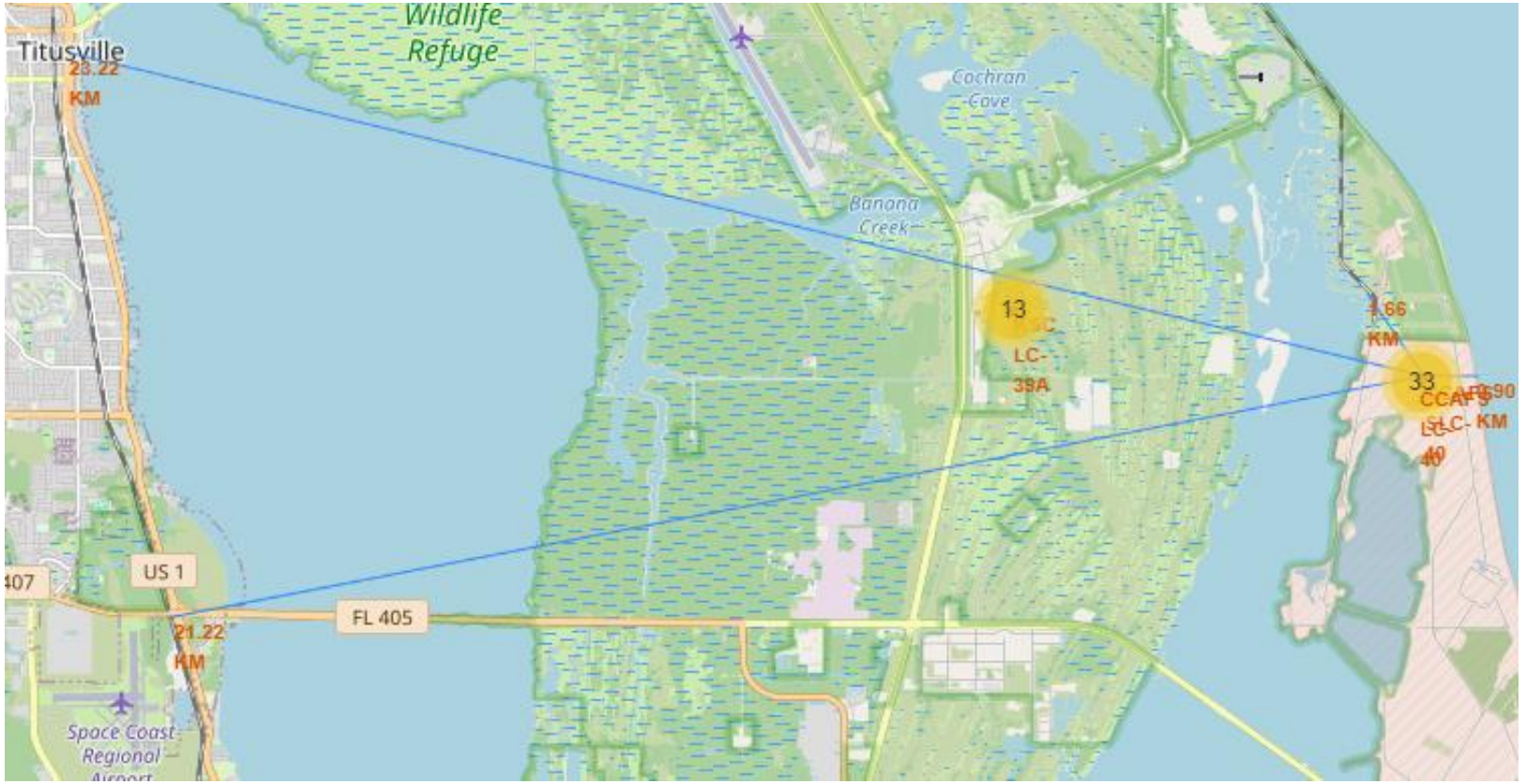
Distance to Proximities to CCAFS SLC-40

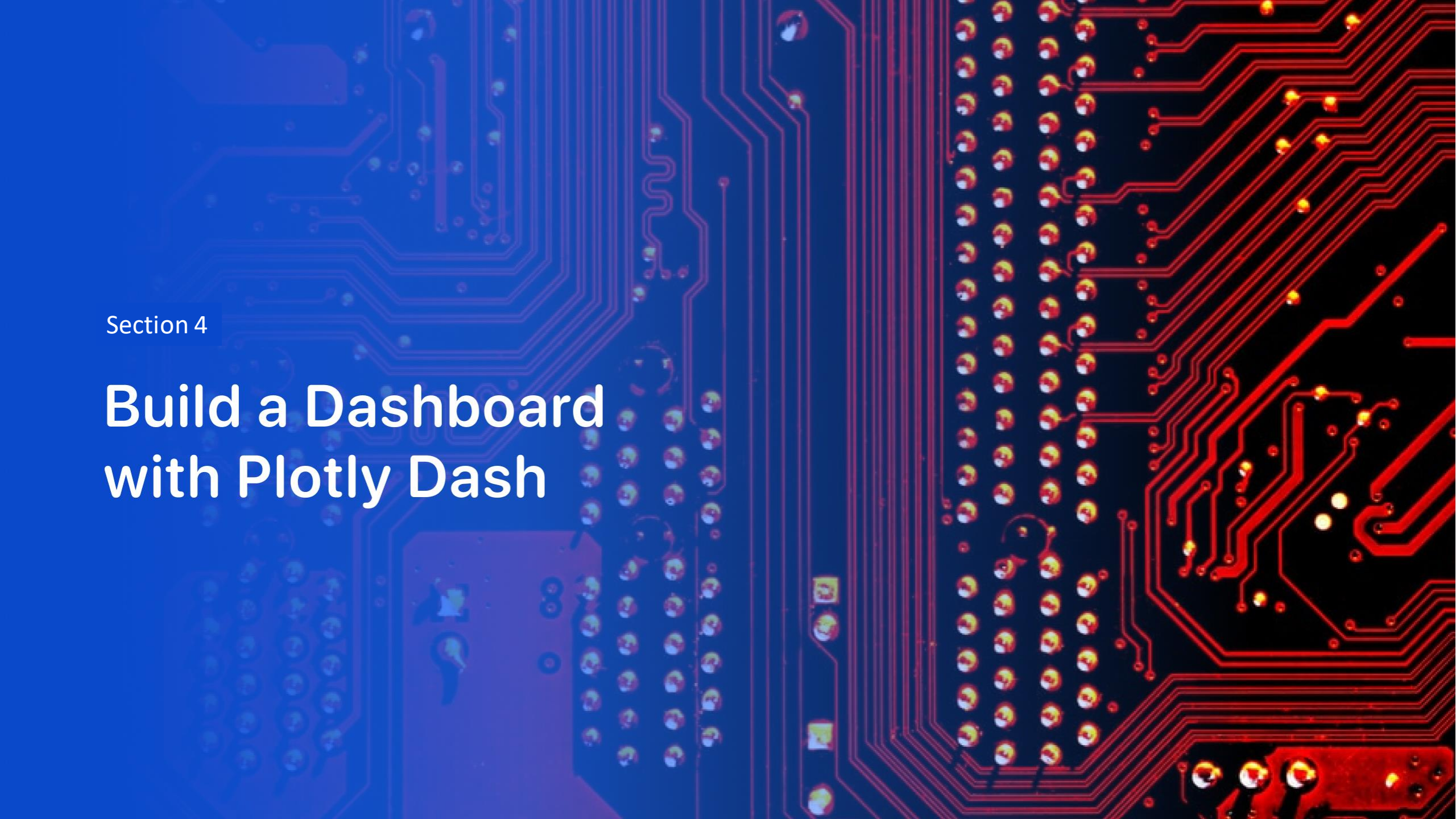
- 0.90 km from nearest coastline
- 1.66 km from nearest railway
- 23.22 km from nearest city
- 21.22 km from nearest highway



Distance to Proximities to CCAFS SLC-40 (cont.)

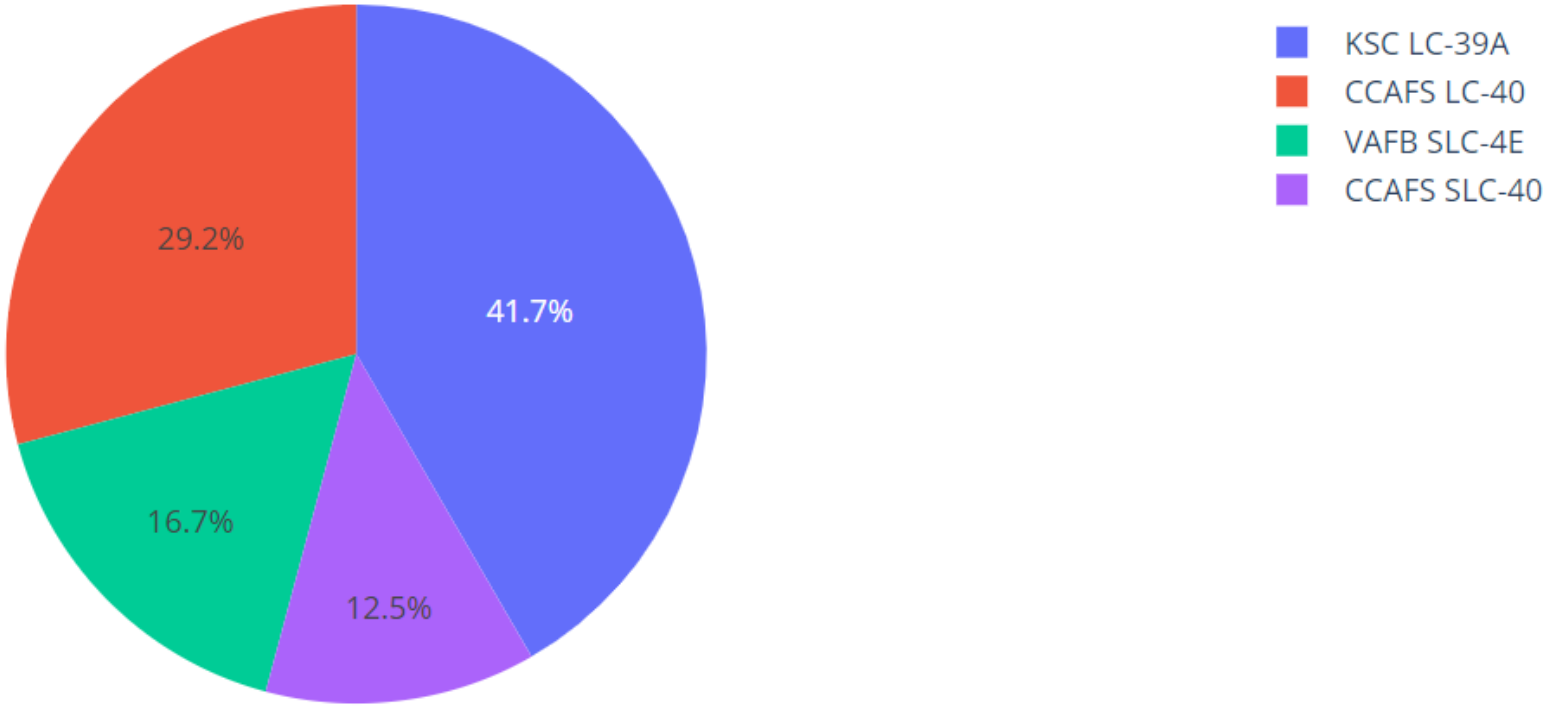
- Coasts: Prevent debris from launches falling on people or property.
- Safety / Security: Establish exclusion zones for site safety.
- Transportation/Infrastructure: Balance proximity to support services with safety, avoiding damage from potential launch failures.




Section 4
Build a Dashboard with Plotly Dash

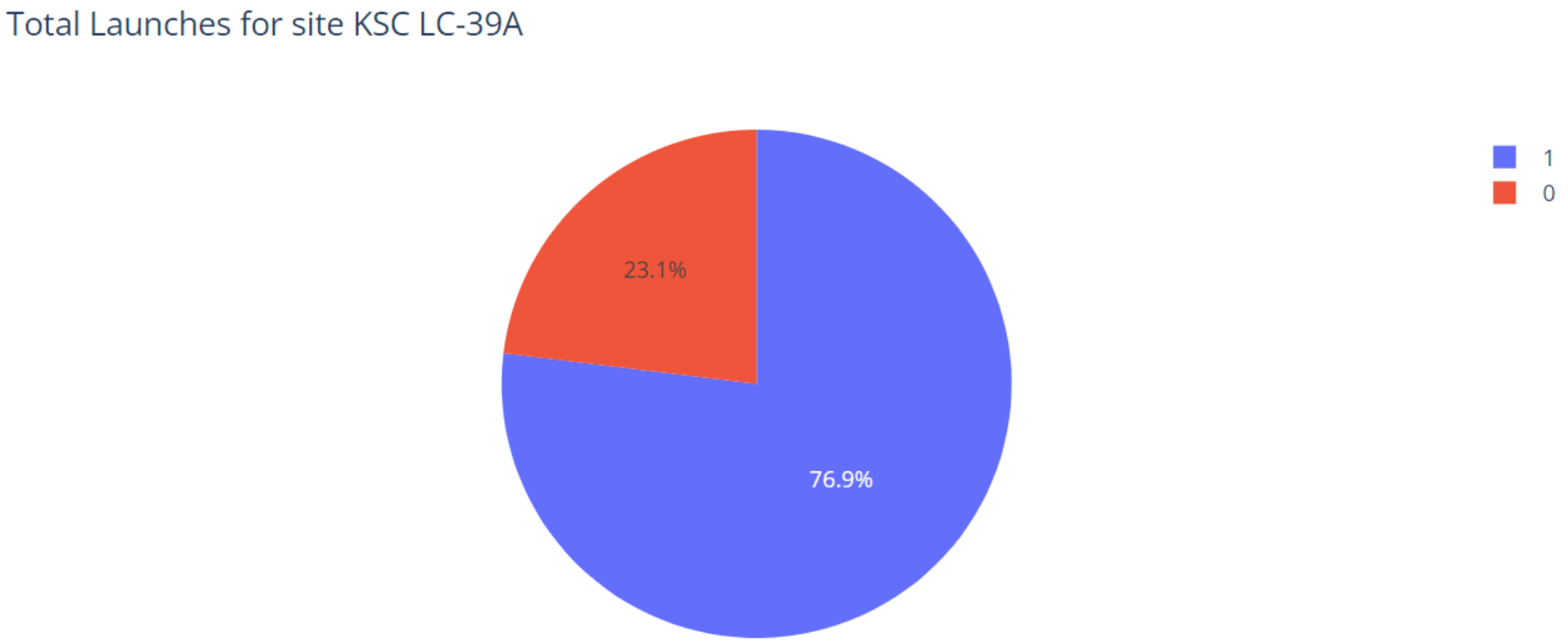
Launch Success by Site

- KSC LC-39A has the most successful launches amongst launch sites (41.7%) while CCAPS SLC-40 has the least (12.5%).



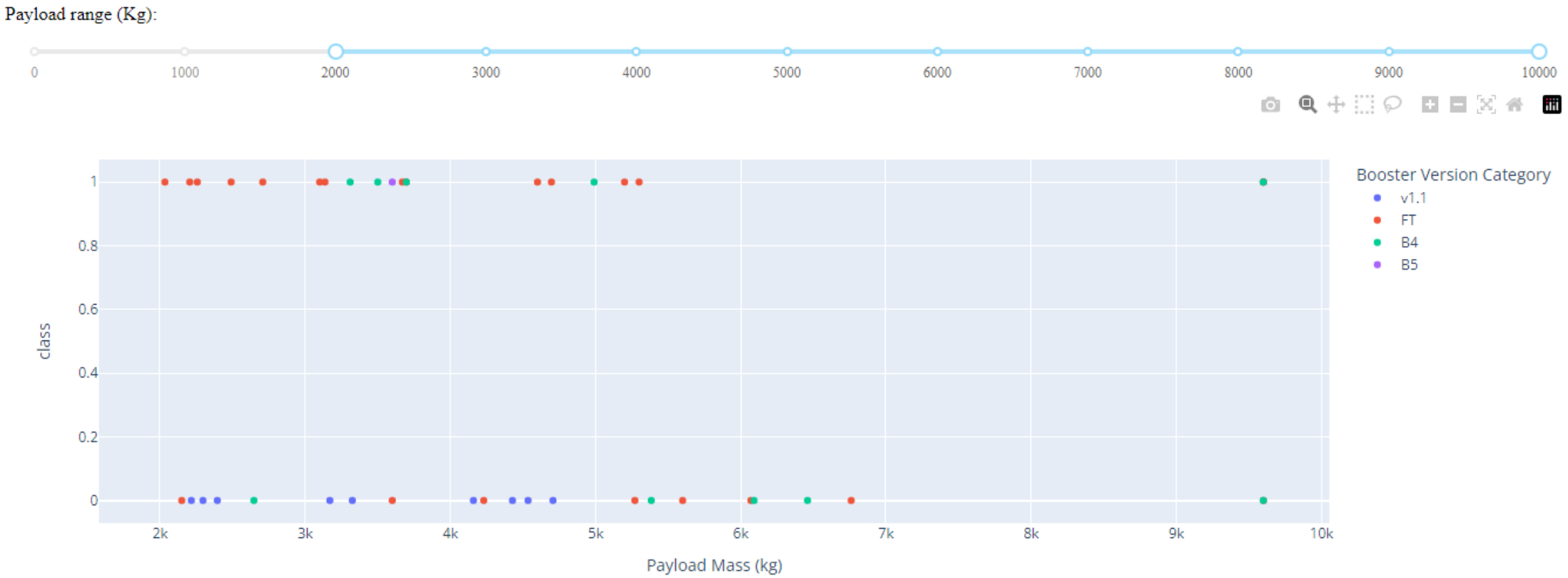
Launch Success (KSC LC-29A)

- KSC LC-39A boasts the highest success rate among launch sites, standing at 76.9%.



Payload Mass and Success

- A higher success rate is observed within the payload range of 2000-6000 kg. Notably, both FT and B5 exhibit larger success rates in this payload category.

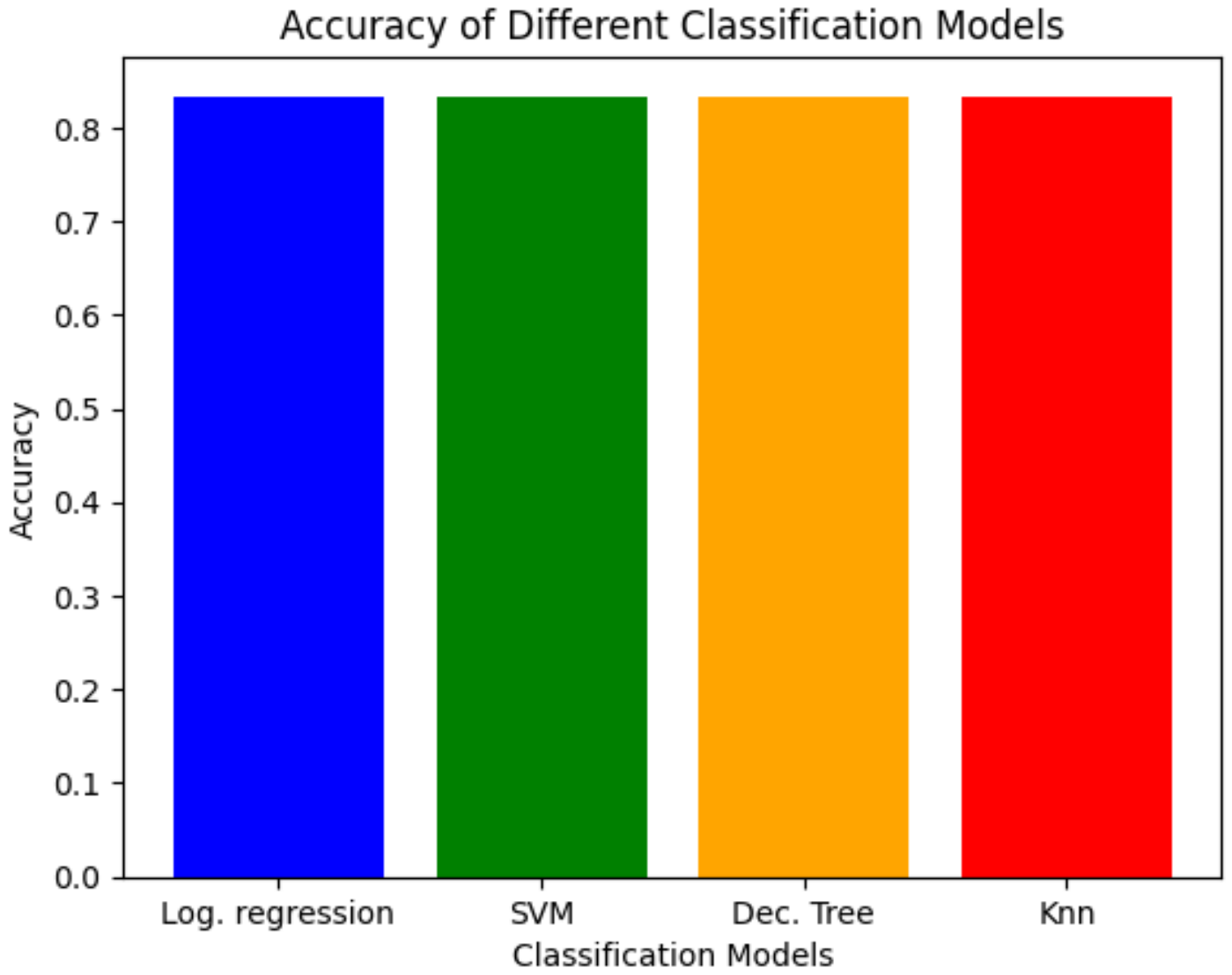


Section 5

Predictive Analysis (Classification)

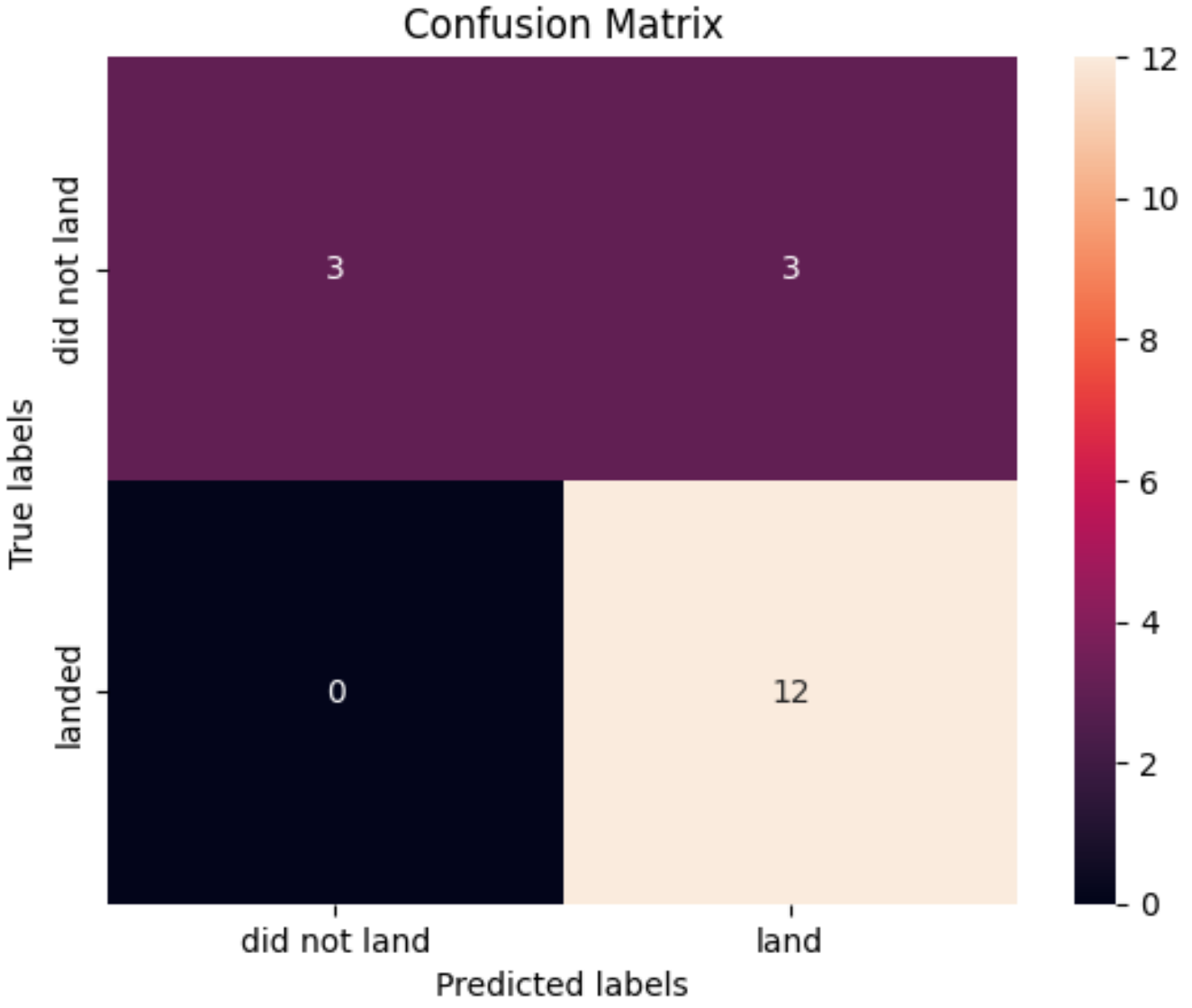
Classification Accuracy

- All methods demonstrate equal performance on the test data, with each achieving an accuracy of 0.833333.



Confusion Matrix

- All four classification models share identical confusion matrices, effectively distinguishing between different classes. However, a common challenge across all models is the occurrence of false positives, which stands out as a major issue.



Conclusions

Payload Mass: Across all launch sites, there is a positive correlation between higher payload mass (kg) and a higher success rate.

Launch Success: The overall success rate of launches has shown an increasing trend over time since 2013.

Equator: Launch sites' proximity to the equator provides an additional natural boost due to Earth's rotational speed, contributing to cost savings in fuel and boosters.

Coast: All launch sites are situated close to the coast, enhancing safety by minimizing the risk of debris falling on populated areas.

Model Performance: The models exhibited the same accuracy values.

In the future, models such as XGBoost and Random Forests should be considered for improved insights, as well as PCA to enhance model accuracy.

Thank you!