IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Syeda Fatima Sarfaraz
20/06/2024

Outline

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

Executive Summary

- *Summary of methodologies*
  - The methodologies used were: Data Collection, Data Wrangling, EDA with Data Visualisation and SQL, using Folium and Plotly Dash to build interactive maps and dashboards, and finally Predictive Analysis with ML.
- *Summary of all results*
  - Screenshots, charts and tables will be provided as output of all these methodologies.

Introduction

SpaceX: A Leading Force in Cost-Effective Space Launch

Within the commercial space sector, SpaceX has emerged as a preeminent leader. The company has revolutionized space travel by significantly reducing launch costs. Public information regarding SpaceX's Falcon 9 launch vehicle indicates a price point of approximately 62 million USD, substantially lower compared to competitor offerings which can exceed 165 million USD per launch. This substantial cost advantage is primarily attributable to SpaceX's successful implementation of first-stage reusability. Consequently, the ability to predict a successful first-stage landing would allow for a more precise determination of overall launch costs. By leveraging publicly available data and employing machine learning models, this analysis aims to forecast the likelihood of SpaceX recovering and reusing the first stage of the Falcon 9 launch vehicle.

Through this we hope to answer the following questions:

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

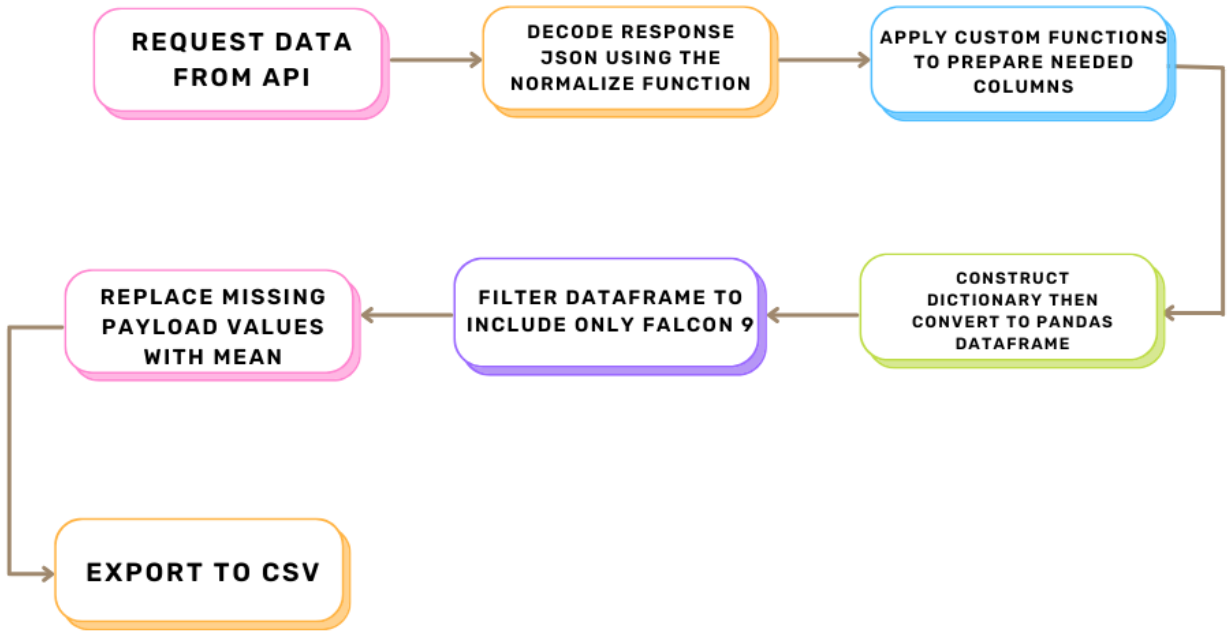
Executive Summary

- Data collection methodology:
  - Using SpaceX's REST API
  - Web scraping from Wikipedia
- Perform data wrangling
  - Filtered Data
  - Dealt with missing values and then used One Hot Encoding to prep the dataset
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tunes, evaluated classification models

Data Collection

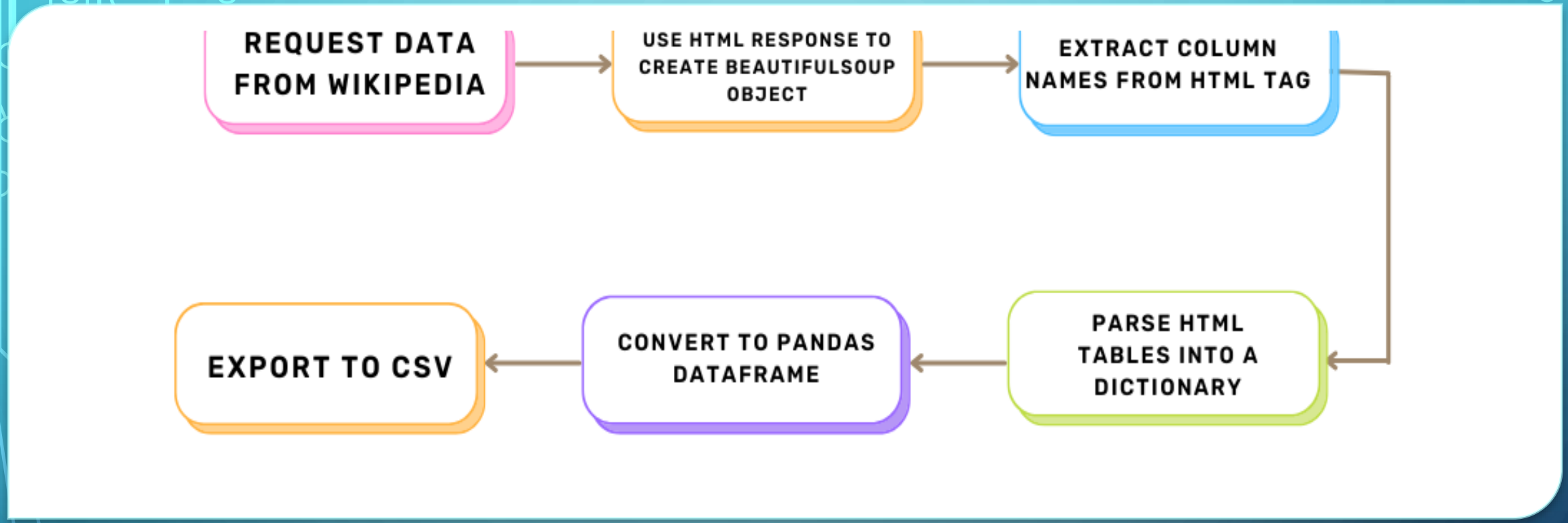
The data collection process for this analysis employed a multifaceted approach, leveraging two primary methodologies:

1. **SpaceX REST API:** This programmatic interface provided access to a rich dataset of SpaceX launch information. Through a series of API requests, we were able to acquire the following data columns pertaining to each launch: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, and Latitude.
2. **Web Scraping of SpaceX Wikipedia Entry:** To obtain a more comprehensive picture of each launch, we supplemented the API data with information extracted from a table within SpaceX's Wikipedia entry. Web scraping techniques were employed to harvest the following data columns: Flight No., Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time.



DATA COLLECTION – SPACEX API

HTTPS://GITHUB.COM/FSSARFARAZ/IBM-APPLIED-DATA-SCIENCE-CAPSTONE-PROJECT/BLOB/MAIN/JUPYTER_LABS_SPACEX_DATA_COLLECTION_API.IPYNB



DATA COLLECTION – SCRAPING

HTTPS://GITHUB.COM/FSSARFARAZ/IBM-APPLIED-DATA-SCIENCE-CAPSTONE-PROJECT/BLOB/MAIN/JUPYTER-LABS-WEBSCRAPING.IPYNB

DATA WRANGLING

Booster Landing Outcomes: The dataset encompasses various booster landing scenarios: - Successful landings: "True Ocean," "True RTLS," "True ASDS" - Unsuccessful landings: "False Ocean," "False RTLS," "False ASDS"

For model training, these outcomes are transformed into binary labels: "1" signifies successful landing, "0" signifies unsuccessful landing.

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA WITH DATA VISUALIZATION

The analysis utilized various chart types to explore relationships within the data.

- Scatter Plots: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type. These visualizations aim to identify potential correlations between the listed variables, which could be leveraged for machine learning model development.
- Bar Charts: Orbit Type vs. Success Rate. This chart focuses on comparing success rates across different orbital categories.
- Line Charts: Success Rate Yearly Trend. This visualization depicts the trend in launch success rate over time, allowing for an assessment of potential temporal patterns.

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/edadataviz.ipynb

EDA WITH SQL

- Unique Launch Sites: Identified the distinct launch site names used throughout the space missions.
- Launch Sites Starting with 'CCA': Retrieved the details of the first 5 launch missions conducted from sites whose names begin with the string "CCA."
- Payload Mass by Customer (NASA CRS): Calculated the total payload mass carried by boosters launched for NASA's CRS program.
- Average Payload Mass by Booster Version (F9 v1.1): Determined the average payload mass carried by the F9 v1.1 booster version.
- First Successful Ground Pad Landing: Identified the date of the first successful mission where the booster landed on a ground pad.
- Successful Drone Ship Landings (4000kg < Payload < 6000kg): Retrieved the names of boosters that successfully landed on a drone ship while carrying a payload mass between 4,000 kg and 6,000 kg.
- Mission Outcome Counts: Summarized the total number of successful and failed missions.
- Maximum Payload Mass by Booster Version: Identified the booster versions responsible for carrying the heaviest payloads.
- Failed Drone Ship Landings (2015): Listed the failed landing outcomes in a drone ship for the year 2015, including the corresponding booster versions and launch site names.
- Landing Outcome Ranking (2010-06-04 to 2017-03-20): Ranked the frequency of various landing outcomes (e.g., Failure (drone ship), Success (ground pad)) in descending order for launches occurring between June 4th, 2010, and March 20th, 2017.

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

Interactive Map Features:

- **Launch Site Locations:** Markers with circles, popup labels, and text labels were incorporated for all launch sites. These markers utilize the respective latitude and longitude coordinates, effectively pinpointing their geographical locations and proximity to the equator and coastlines.
- **Launch Outcome Visualization:** Success and failed launches were distinguished through color-coded markers (green for success, red for failure). Marker clusters were also employed to visually highlight launch sites with notable success rates.
- **Proximity Analysis:** Distances between a designated launch site (e.g., KSC LC-39A) and its surrounding features were depicted using colored lines. These features could include railways, highways, coastlines, and the closest city, providing a clear understanding of the launch site's infrastructure and accessibility.

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb

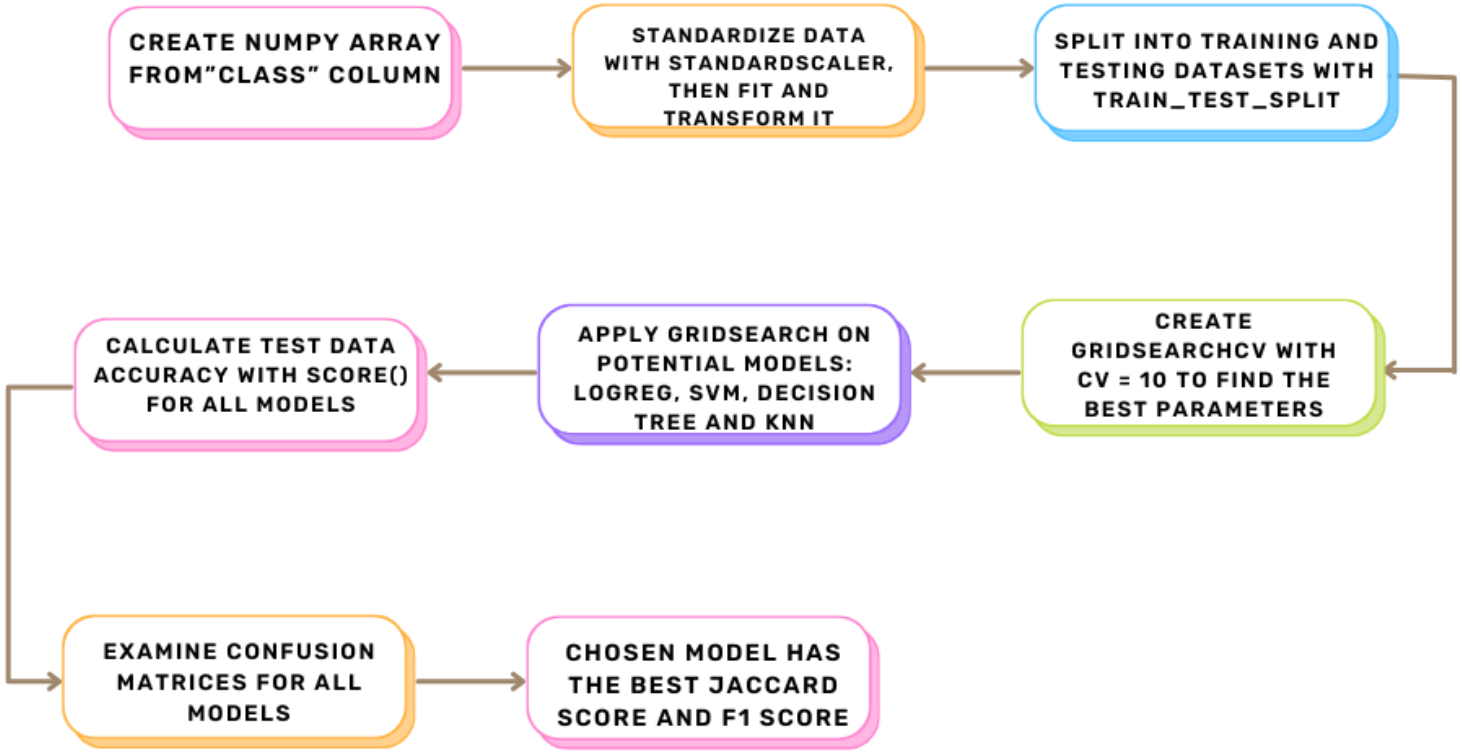
BUILD A DASHBOARD WITH PLOTLY DASH

- **Interactive Launch Site Selection:** A dropdown list was implemented to facilitate the selection of a specific launch site for analysis.
- **Dynamic Pie Chart:** A pie chart displays launch success data. When "All Sites" is selected, the chart presents the total successful launches across all sites. Upon selecting a specific launch site, the chart dynamically updates to show the success and failure breakdown for that chosen site.
- **Payload Mass Range Selection:** A slider allows users to define a desired payload mass range for further investigation.
- **Payload Mass vs. Success Rate Scatter Chart:** A scatter chart visually depicts the correlation between payload mass and launch success rate for the various booster versions employed. This visualization helps identify potential relationships between payload weight and launch outcomes.

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/spacex_dash_app.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

https://github.com/fssarfaraz/IBM-Applied-Data-Science-Capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



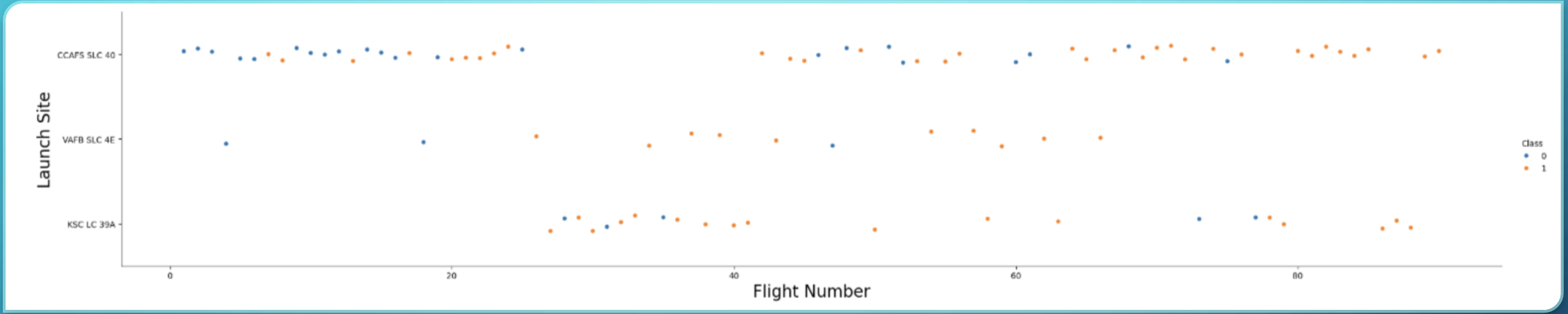
Results

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

Section 2

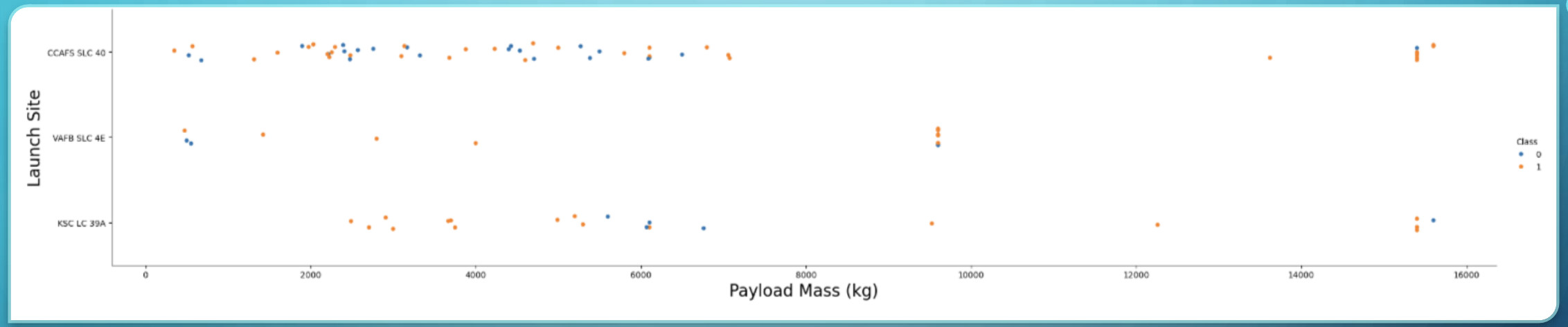
Insights drawn from EDA

FLIGHT NUMBER VS. LAUNCH SITE



- The analysis reveals a potential correlation between flight number (indicating launch order) and success rate. Early missions appear to have exhibited a higher incidence of failure, while later missions demonstrate a greater likelihood of success.
- CCAFS SLC 40 emerges as the prominent launch site, accounting for approximately half of all launches within the dataset.
- VAFB SLC 4E and KSC LC 39A exhibit comparatively higher success rates when compared to other launch sites.

PAYLOAD VS. LAUNCH SITE

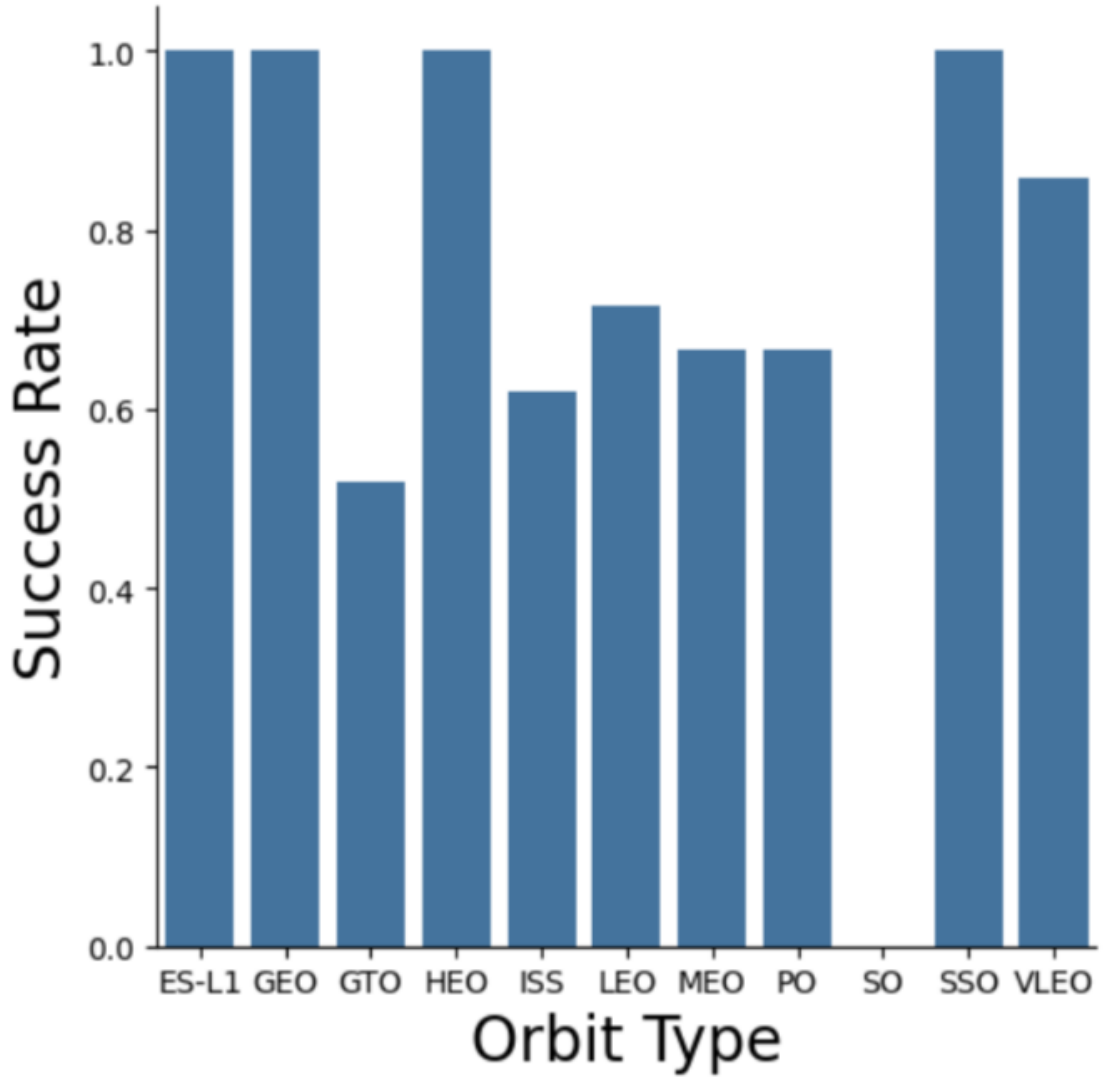


- The analysis suggests a potential positive correlation between payload mass and launch success rate across various launch sites. Launches exceeding 7,000 kg in payload mass appear to have a higher success rate overall.
- Notably, KSC LC-39A exhibits a 100% success rate for missions carrying payloads under 5,500 kg.

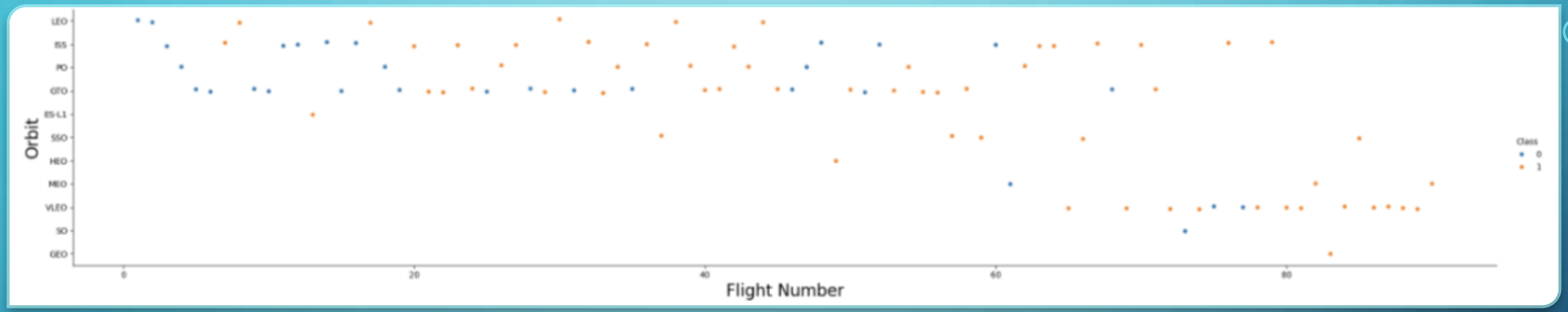
Success Rate vs. Orbit Type

The analysis revealed varying success rates across different orbital categories:

- **100% Success Rate:** Orbits designated ES-L1, GEO, HEO, and SSO exhibited a perfect success record within the dataset.
- **0% Success Rate:** Launches targeting SO orbit did not achieve success in any instances.
- **Moderate Success Rates (50% - 85%):** GTO, ISS, LEO, MEO, and PO orbits demonstrated success rates ranging from 50% to 85%.

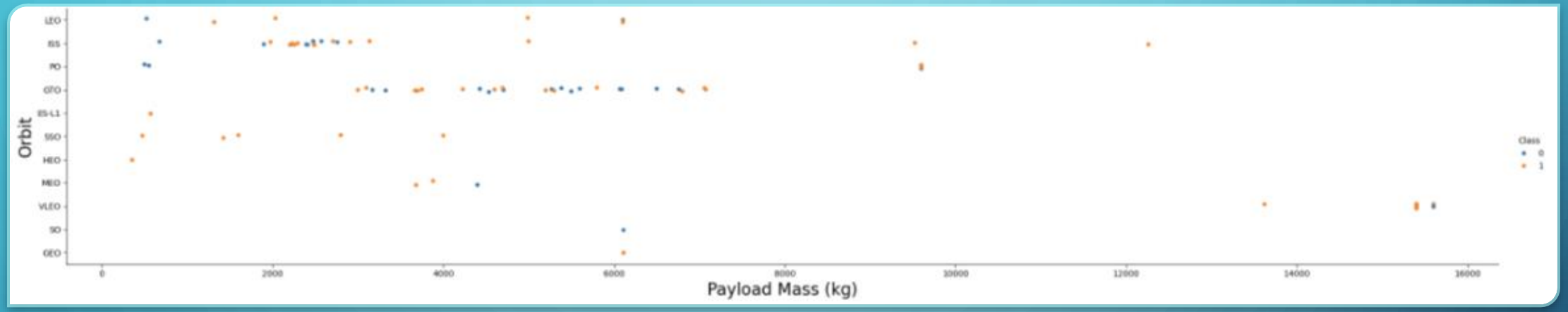


FLIGHT NUMBER VS. ORBIT TYPE



For launches aiming for Low Earth Orbit (LEO), there appears to be a correlation between a higher flight number (indicating later launches) and increased success. Conversely, no apparent relationship between flight number and success rate was observed for missions targeting Geosynchronous Transfer Orbit (GTO)

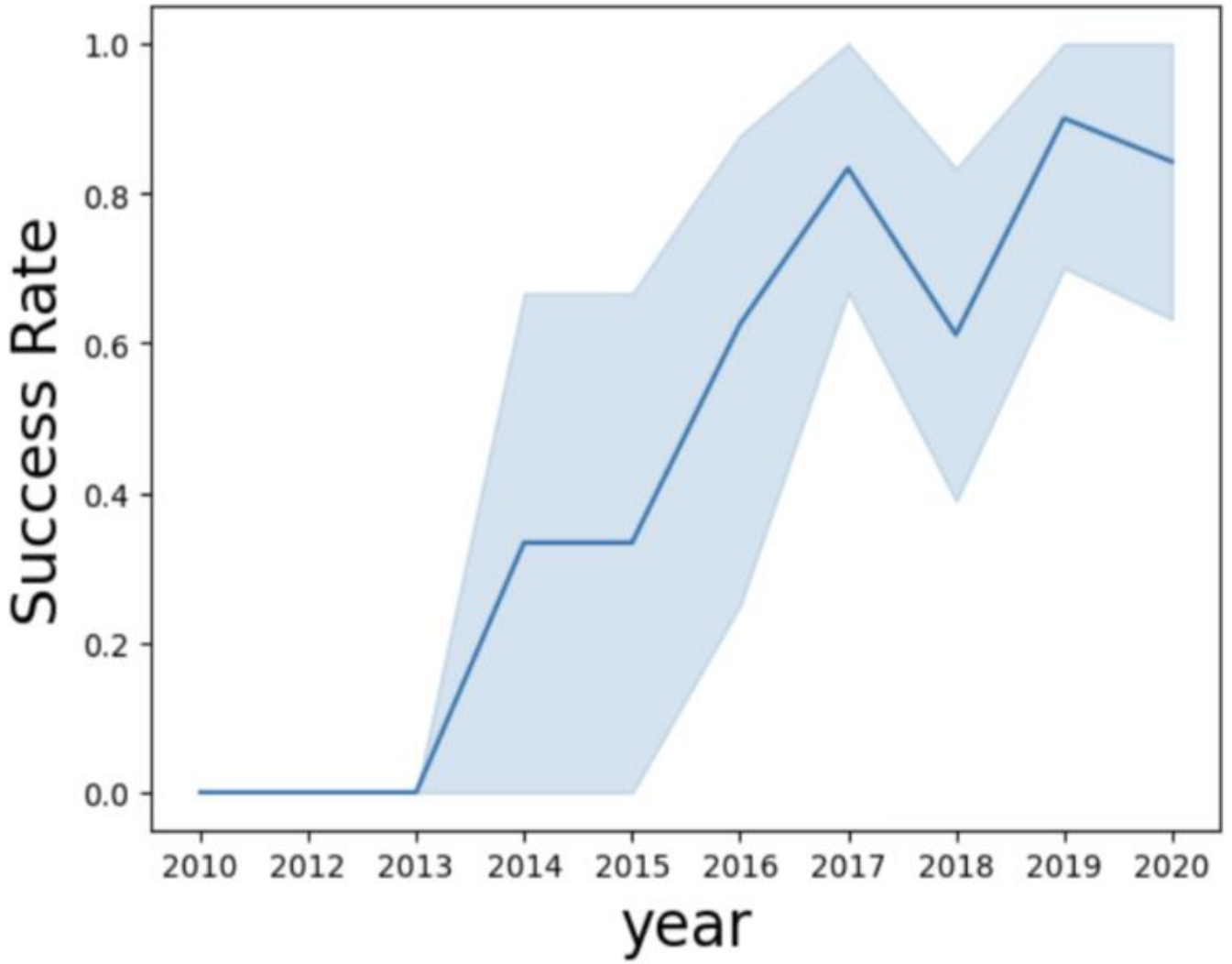
PAYLOAD VS. ORBIT TYPE



For GTO missions, heavier payloads appear to have a negative impact on success. Conversely, heavier payloads might be associated with a positive influence on success rates for both GTO and Polar LEO (specifically ISS) missions

Launch Success Yearly Trend

The success rate rose steadily but not linearly from 2013.



All Launch Site Names

Displays the names of the unique launch sites. We can see that there are four (4).

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

```
%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

Launch Site Names Begin with 'CCA'

Displays five (5) records where the launch site name begins with the string 'CCA'

Display 5 records where launch sites begin with the string '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

Total Payload Mass

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

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

```
%sql select sum(payload_mass__kg_) as totalMass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

totalMass
45596

Average Payload Mass by F9 v1.1

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

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

```
%sql select avg(payload_mass__kg_) as averageMass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

averageMass
2534.6666666666665

First Successful Ground Landing Date

Listing the date when the first successful landing outcome in ground pad was achieved.

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

Hint:Use min function

```
%sql select min(date) as firstSuccessfulLanding from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Listing the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000.

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

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
```

```
 * 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

Listing the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

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

Boosters Carried Maximum Payload

Listing the names of the booster versions which have carried the maximum payload mass.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
9]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

9]:

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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
%%sql select substr(Date, 6, 2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and substr(Date, 0, 5) = '2015';
```

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

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

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

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.

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.

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

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

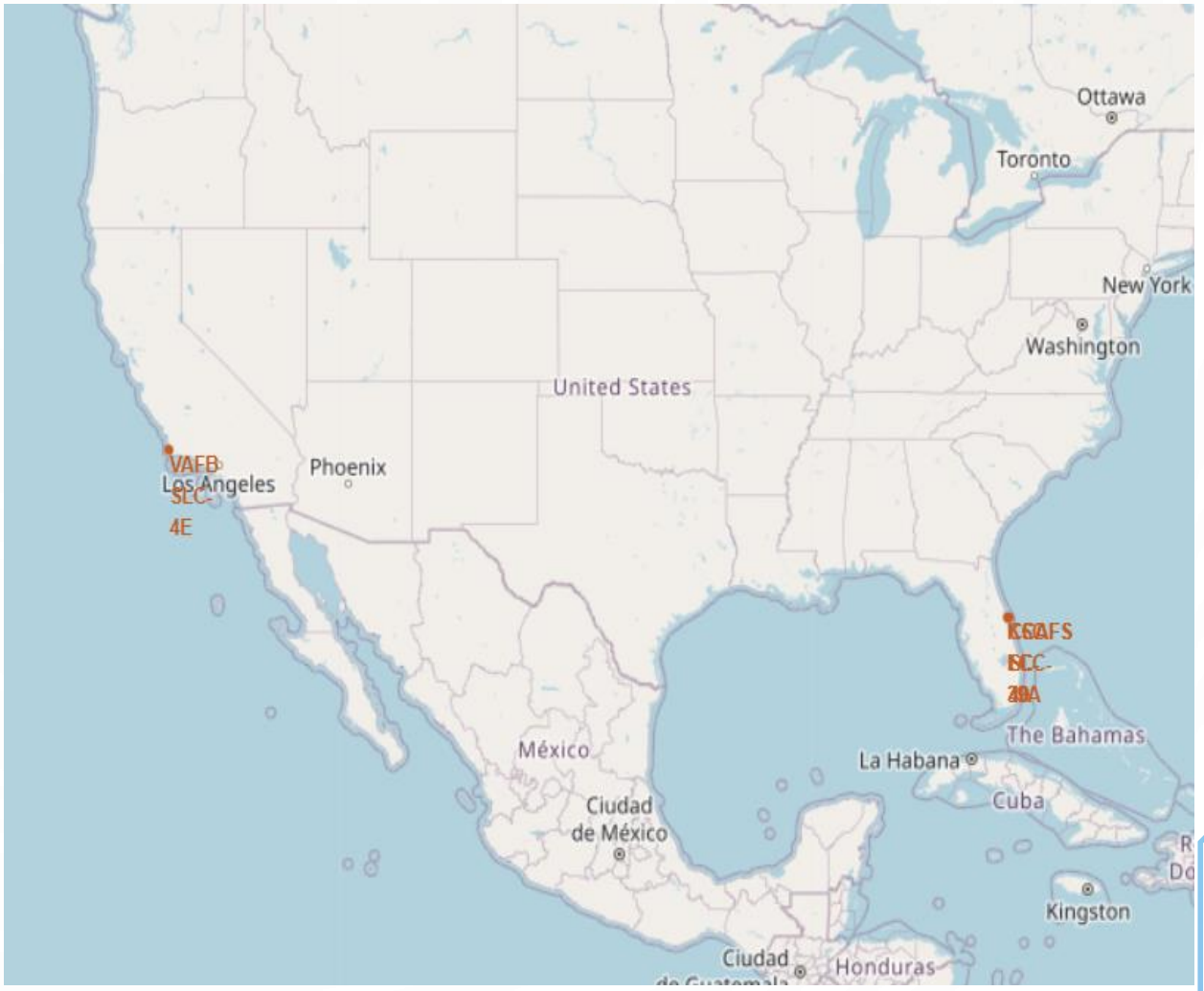
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

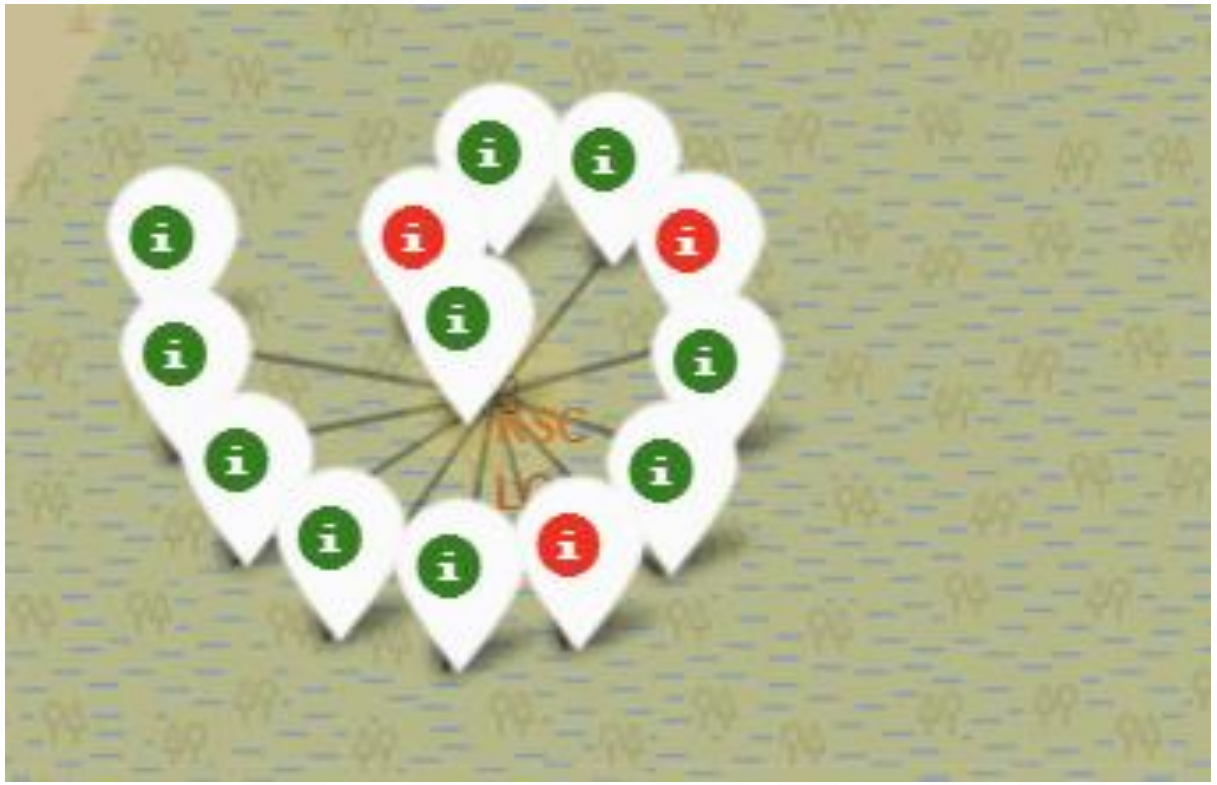
All Launch Sites on a Map

- The analysis reveals a concentration of launch sites near the Earth's equator. At the equator, the Earth's surface has the highest linear velocity. A spacecraft launched from the equator inherits this eastward velocity, providing an initial boost in orbital speed and reducing the fuel required to achieve orbit.
- The clustering of launch sites close to coastlines offers a safety advantage. In the event of a launch mishap, debris dispersion is directed away from populated areas and towards the ocean, minimizing potential risks on land.



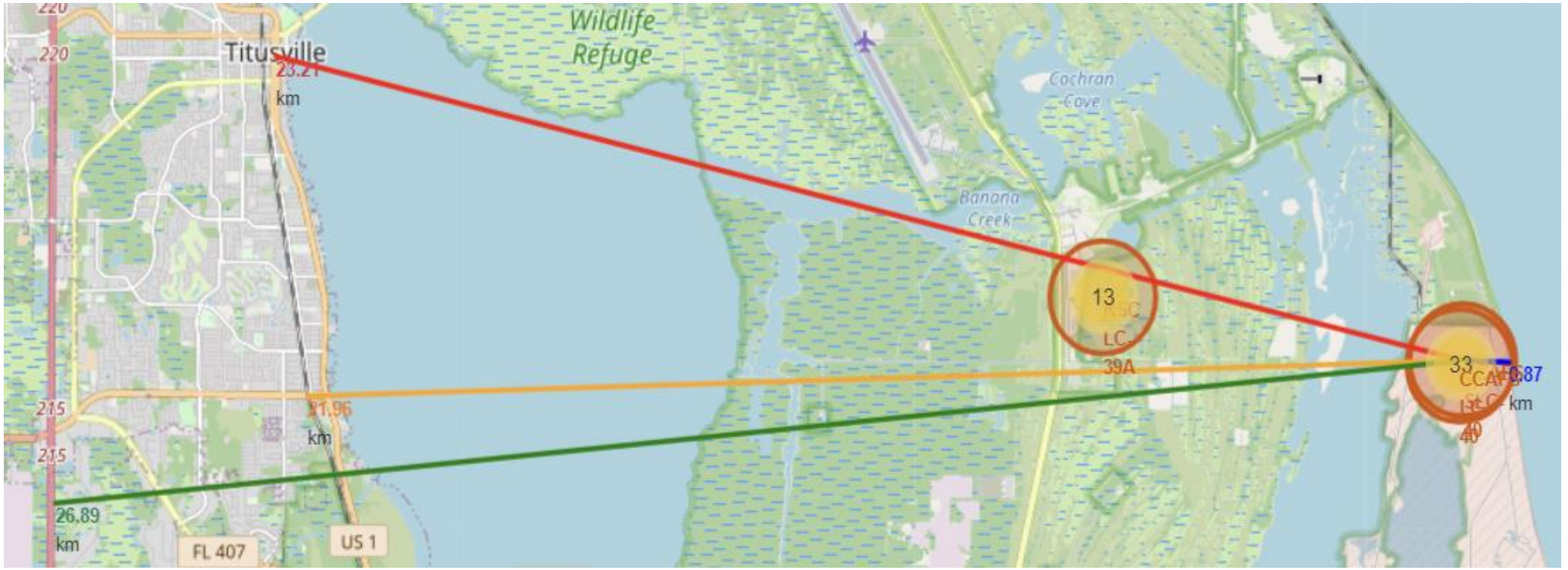
Colour-Coded Launch Outcomes

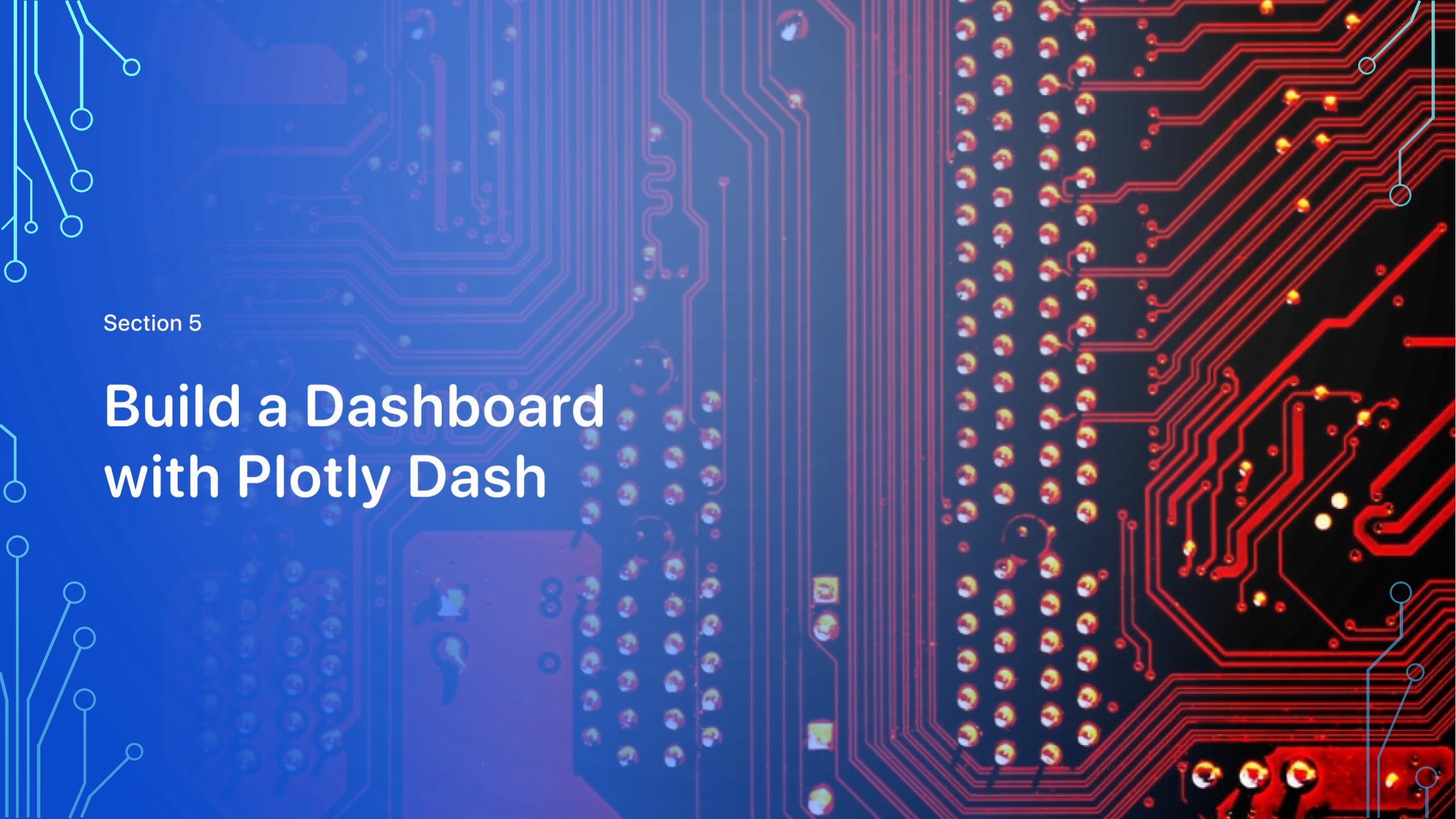
The color-coded markers (green for success, red for failure) facilitate the identification of launch sites with high success rates. A cursory glance reveals that Launch Site KSC LC-39A exhibits a particularly high concentration of green markers, indicating a strong performance record.



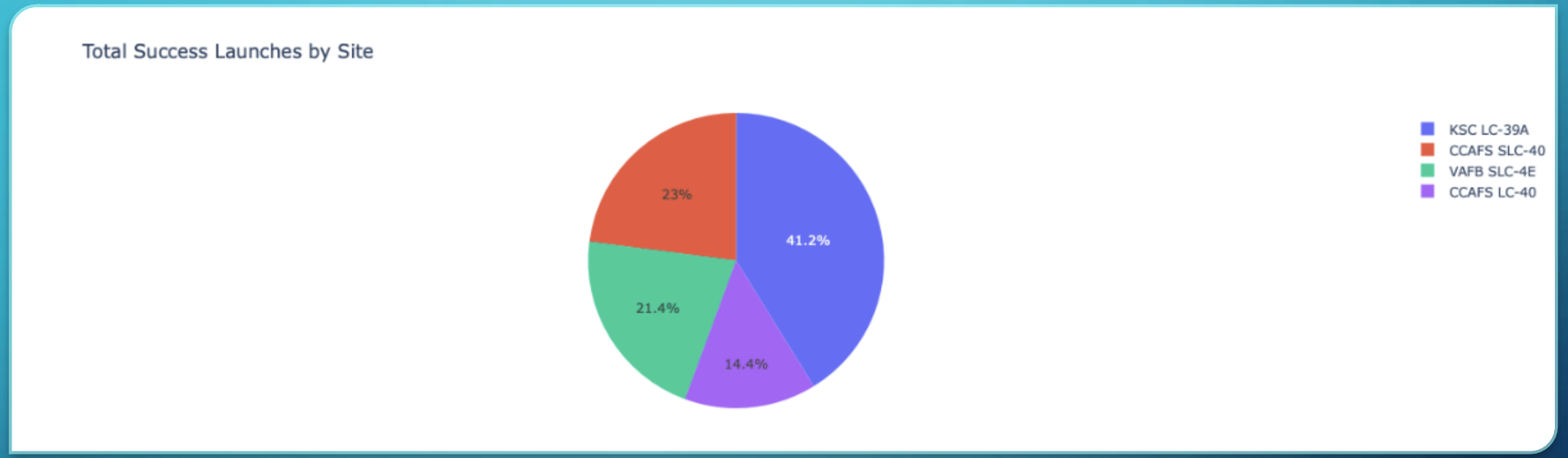
Distance from Launch Site CCAFS SLC-40

This launch site is between 20 and 30 kms away from each of the closest railways, cities and highways. It is only 0.87 km away from the coastline.



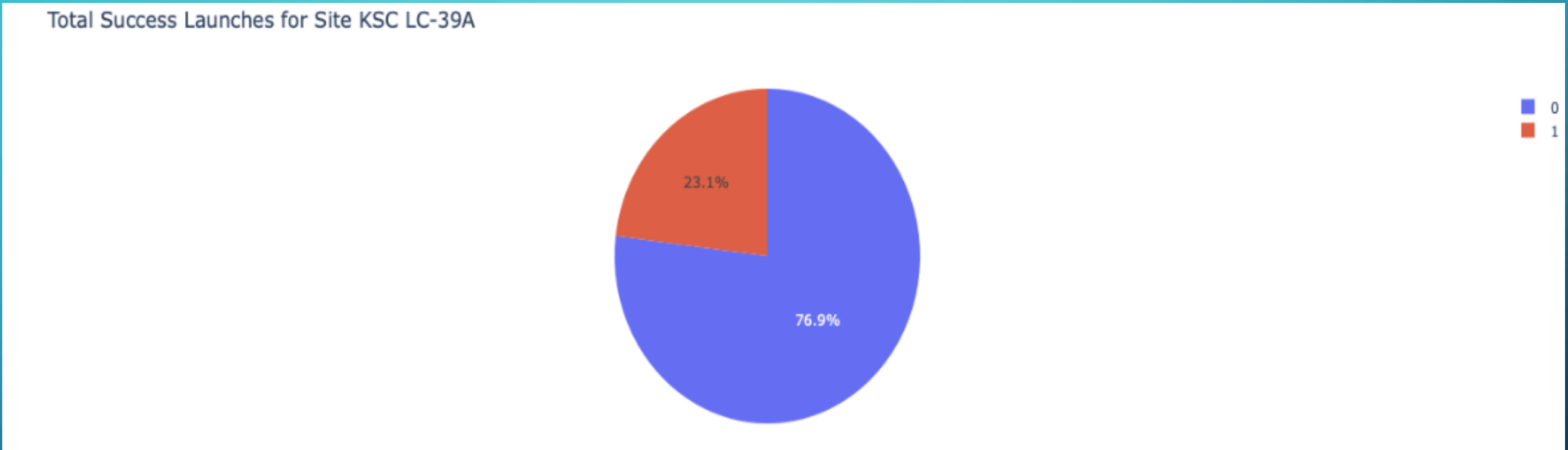

Section 5
Build a Dashboard with Plotly Dash

LAUNCH SUCCESSES FOR ALL SITES



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

SITE WITH HIGHEST SUCCESS RATIO



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

PAYLOAD VS. LAUNCH OUTCOME

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 6

Predictive Analysis (Classification)

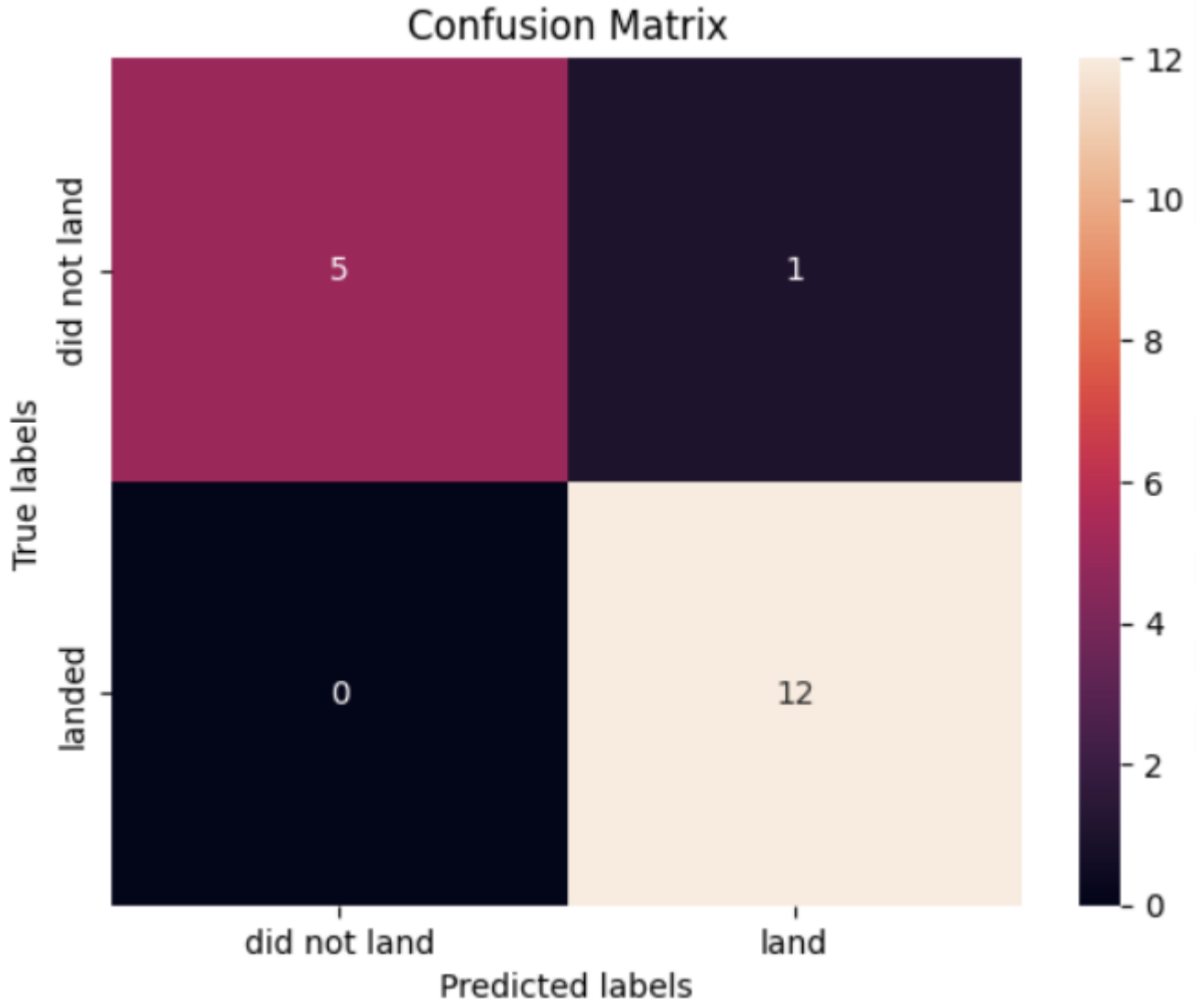
Classification Accuracy

Based on the following scores, we can see that the Decision Tree performed best. It had the highest Jaccard and F1 scores, as well as the highest accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.863636	0.819444
F1_Score	0.909091	0.916031	0.926829	0.900763
Accuracy	0.866667	0.877778	0.900000	0.855556

Confusion Matrix of Decision Tree

False positives are usually the biggest problem but with this model, we can see that there is only 1 false positive.



Conclusions

- The analysis identified the decision tree algorithm as the most suitable model for this dataset based on its performance metrics.
- A trend emerged, suggesting that launches with lower payload mass tend to achieve higher success rates compared to those carrying heavier payloads.
- The launch sites exhibit a strategic clustering near the equator, capitalizing on the Earth's rotational velocity for launch efficiency. Additionally, their proximity to coastlines minimizes potential risks associated with launch mishaps.
- The analysis revealed a potential upward trend in launch success rate over the years, suggesting improvements in technology and launch procedures.
- KSC LC-39A distinguished itself by achieving the highest overall success rate among all launch sites within the dataset.
- Notably, orbits designated ES-L1, GEO, HEO, and SSO exhibited a perfect success record in the data analyzed.

Appendix

Thank you ☺

Thank you!