IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ASAD ALI
15 JUNE 2025

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

SUMMARY OF METHODOLOGIES

SUMMARY OF ALL RESULTS

Introduction

- SpaceX, a private aerospace company, significantly reduces launch costs by reusing the Falcon 9 first stage. Each Falcon 9 launch costs approximately $62 million, which is much lower than traditional providers charging over $165 million. A key factor in this cost efficiency is the successful recovery and reuse of the rocket's first stage. Predicting whether a Falcon 9 first stage will land successfully can provide critical insight into launch cost estimation and risk assessment. This information is valuable for companies or agencies considering competing with or contracting SpaceX for launches. The project involves collecting launch data via the SpaceX API and preparing it for machine learning modeling aimed at predicting landing outcomes.

- Problems I Wanted to Find Answers for:

  - Can we predict whether the Falcon 9 first stage will successfully land using available launch data?

  - What launch parameters influence landing success the most?

  - How reliable is the current SpaceX first stage landing system based on historical data?

  - What patterns or trends exist in successful vs. unsuccessful landings?

Section 1

Methodology

Methodology

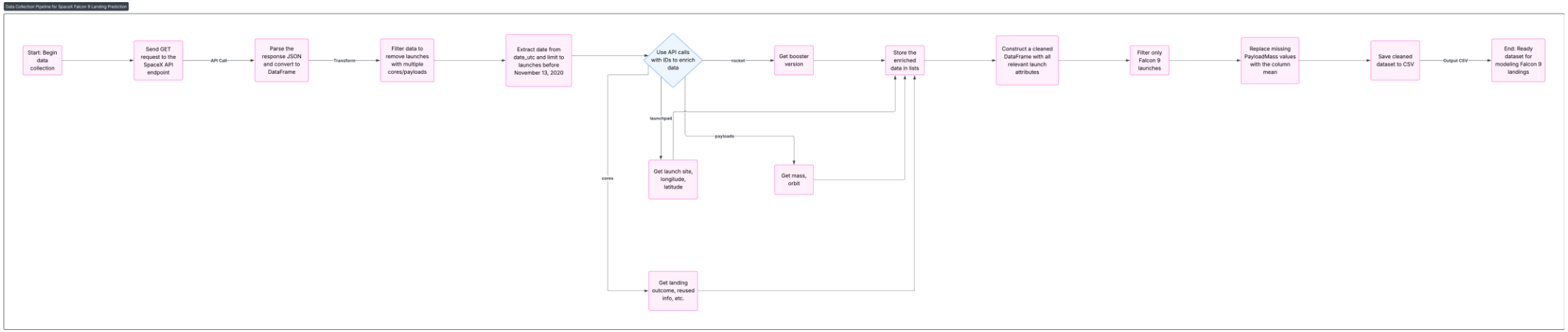
- Executive Summary
- Data collection methodology: The data was collected from the SpaceX public API using HTTP GET requests. The main endpoint (/v4/launches/past) provided launch data, and related details (e.g., rocket type, launchpad, payload, and core) were fetched using their respective IDs via additional API calls.
- Perform data wrangling
    - The JSON response was normalized into a Pandas DataFrame.
    - Only single-core, single-payload Falcon 9 launches before Nov 13, 2020, were kept.
    - IDs were replaced with meaningful values (e.g., booster names, orbit type).
    - Missing PayloadMass values were filled with the column mean; LandingPad missing values were kept as None.
    - The final dataset was saved as dataset_part_1.csv.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Build: Use features like payload mass, orbit, site, etc., to train models like logistic regression or random forest.
    - Tune: Apply grid search or cross-validation to optimize hyperparameters.
    - Evaluate: Use metrics like accuracy, precision, recall, and confusion matrix to assess model performance

Data Collection

The data collection process utilized a combination of API requests to the SpaceX REST API and web scraping from a table on SpaceX's Wikipedia page. Both methods were necessary to gather complete and comprehensive information about the launches for in-depth analysis.

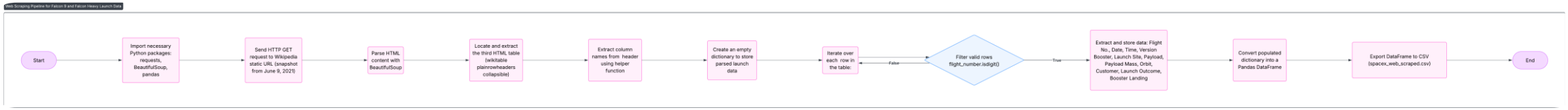
Data fields retrieved from the SpaceX REST API include:
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

Data fields obtained via Wikipedia web scraping include:
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.



Data Collection – SpaceX API

https://github.com/asadali2468/DS_Capstone_Project/blob/main/data-collection-api.ipynb



Data Collection – Scraping

https://github.com/asadali2468/DS_Capstone_Project/blob/main/webscraping.ipynb

Data Wrangling

- The dataset includes various instances where the booster landing was unsuccessful. In some cases, a landing was attempted but failed due to accidents. For example, "True Ocean" indicates a successful landing in a designated ocean region, while "False Ocean" indicates an unsuccessful landing attempt in the same region. Similarly, "True RTLS" represents a successful landing on a ground pad (Return To Launch Site), whereas "False RTLS" indicates a failed attempt. "True ASDS" means the booster successfully landed on a drone ship (Autonomous Spaceport Drone Ship), and "False ASDS" means the landing attempt on the drone ship was unsuccessful.
- These outcomes are converted into training labels for the model, where a value of “1” indicates a successful landing and “0” indicates a failure.

https://github.com/asadali2468/DS_Capstone_Project/blob/main/data%20wrangling.ipynb

EDA with Data Visualization

Various charts were created to visualize the data, including:

- **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**
- **Yearly Trend of Success Rate**
- Scatter plots were used to explore relationships between numerical variables. If patterns or correlations are present, these can be valuable for informing machine learning models.
- Bar charts were used to compare discrete categories, helping illustrate how specific categorical variables relate to measured values.
- Line charts were employed to show trends over time, making them especially useful for analyzing time series data.

https://github.com/asadali2468/DS_Capstone_Project/blob/main/edadataviz.ipynb

EDA with SQL

- Executed several SQL queries, including:
- Retrieving the unique launch site names involved in the space missions
- Selecting 5 records where the launch site names start with the prefix 'CCA'
- Calculating the total payload mass delivered by boosters launched by NASA (CRS)
- Computing the average payload mass for the booster version F9 v1.1
- Finding the date of the first successful landing on a ground pad
- Identifying boosters that successfully landed on drone ships and carried payloads between 4000 and 6000 units
- Counting the total number of missions with successful and failed outcomes
- Listing booster versions associated with the maximum payload masses carried
- Extracting failed drone ship landings along with their booster versions and launch sites for missions during 2015
- Ranking and counting landing outcomes (e.g., Failure on drone ship, Success on ground pad) between June 4, 2010, and March 20, 2017, sorted in descending order

https://github.com/asadali2468/DS_Capstone_Project/blob/main/eda-sql_sqllite.ipynb

Build an Interactive Map with Folium

- **Launch Site Markers:**
- Plotted a marker for NASA Johnson Space Center using its latitude and longitude, including a circle, popup label, and text label to indicate the starting location.
- Plotted markers for all launch sites with circles, popup labels, and text labels based on their geographic coordinates to visualize their locations and proximity to the equator and coastlines.
- **Launch Outcome Markers by Site:**
- Added color-coded markers to represent launch outcomes—green for successful launches and red for failed ones—using Marker Clustering to highlight sites with higher success rates.
- **Distances from Launch Site to Nearby Features:**
- Drew color-coded lines to represent distances from the KSC LC-39A launch site (as an example) to nearby features such as the railway, highway, coastline, and the nearest city.

- https://github.com/asadali2468/DS_Capstone_Project/blob/main/interactive%20map%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Site Selection Dropdown:

- Implemented a dropdown menu to allow users to select a specific launch site.

Success Launches Pie Chart (All Sites or Selected Site):

- Created a pie chart displaying the total number of successful launches across all sites, or a comparison of successful vs. failed launches when a specific site is selected.

Payload Mass Range Slider:

- Added a slider widget to filter the data based on a selected range of payload mass.

Scatter Plot: Payload Mass vs. Launch Success by Booster Version:

- Developed a scatter plot to illustrate the relationship between payload mass and launch outcome, categorized by different booster versions.

- https://github.com/asadali2468/DS_Capstone_Project/blob/main/spacex-dash-app.py

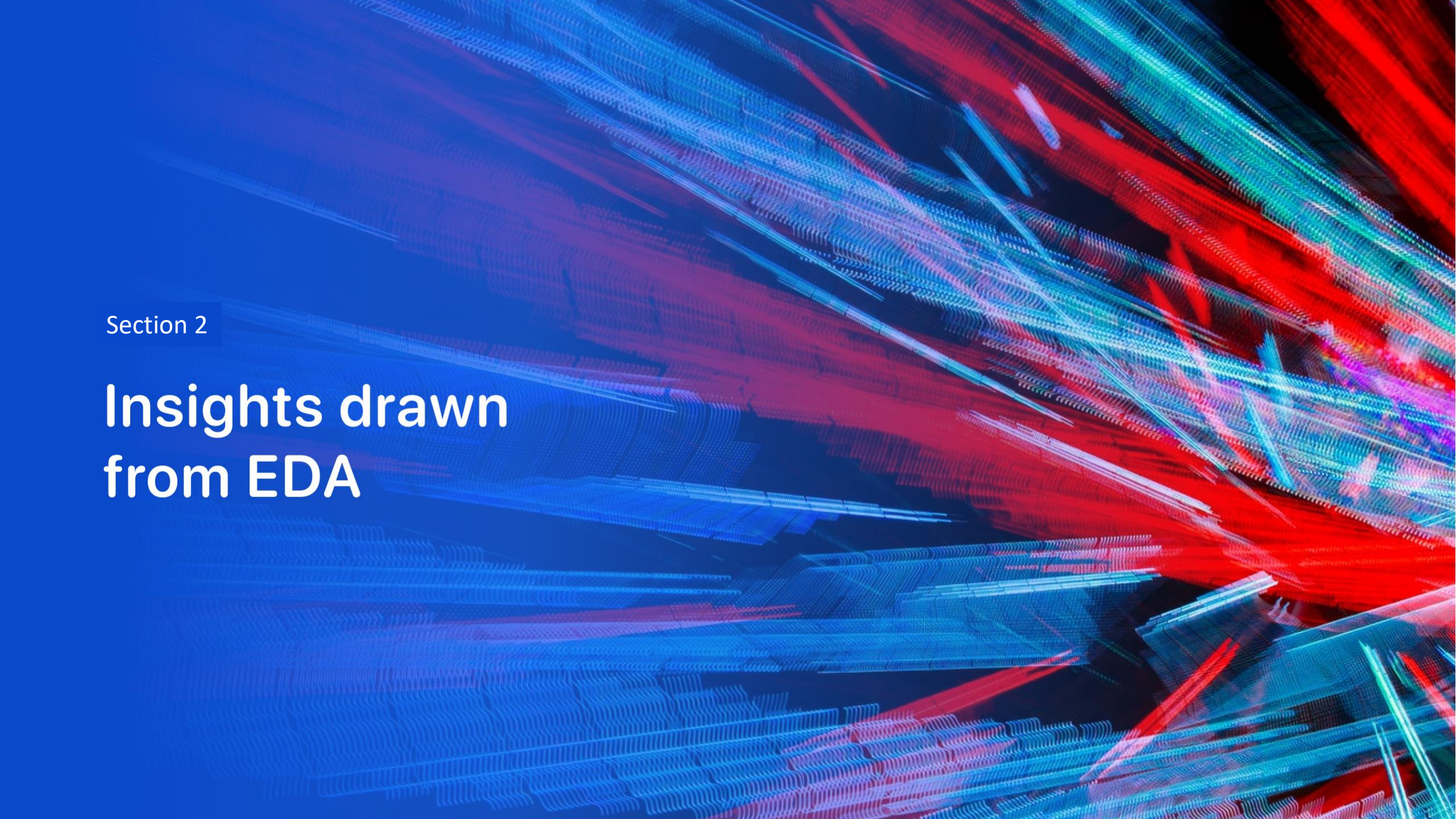
Predictive Analysis (Classification)

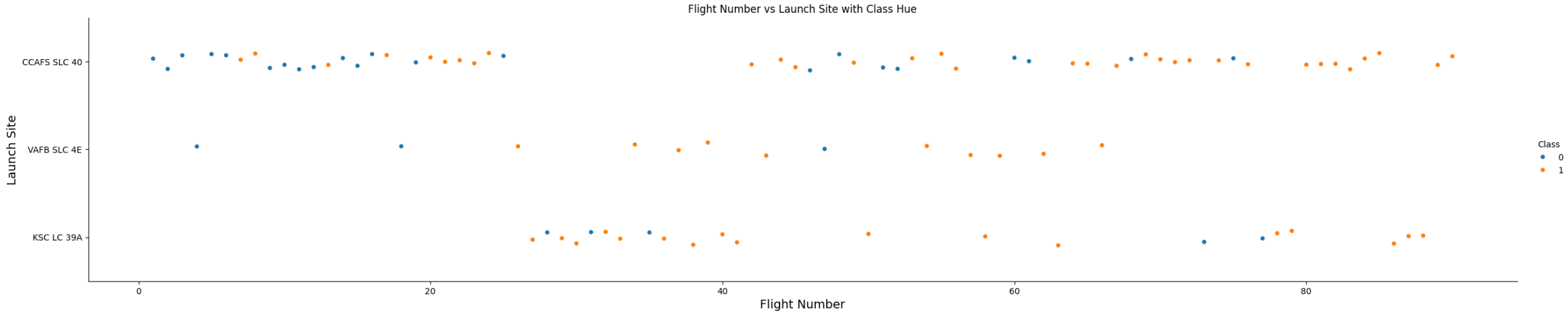
Performed data preparation by creating a binary target for landing success, standardizing features, and splitting data into training and testing sets (80/20). Built and tuned Logistic Regression and SVM models. Logistic Regression (C=0.01, L2) achieved 84.6% CV accuracy; SVM (C=1.0, gamma=0.0316, sigmoid kernel) slightly outperformed with 84.8%. Both models reached 83.3% test accuracy. Confusion matrix revealed false positives as the main error. SVM had marginally better CV results; feature scaling was essential, especially for SVM.

https://github.com/asadali2468/DS_Capstone_Project/blob/main/Machine%20Learning%20Prediction.ipynb

Results

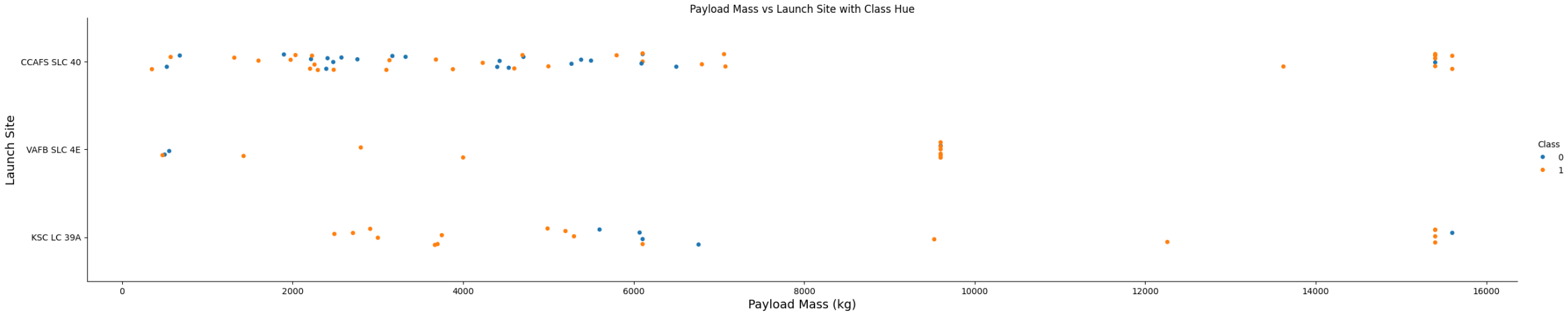
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

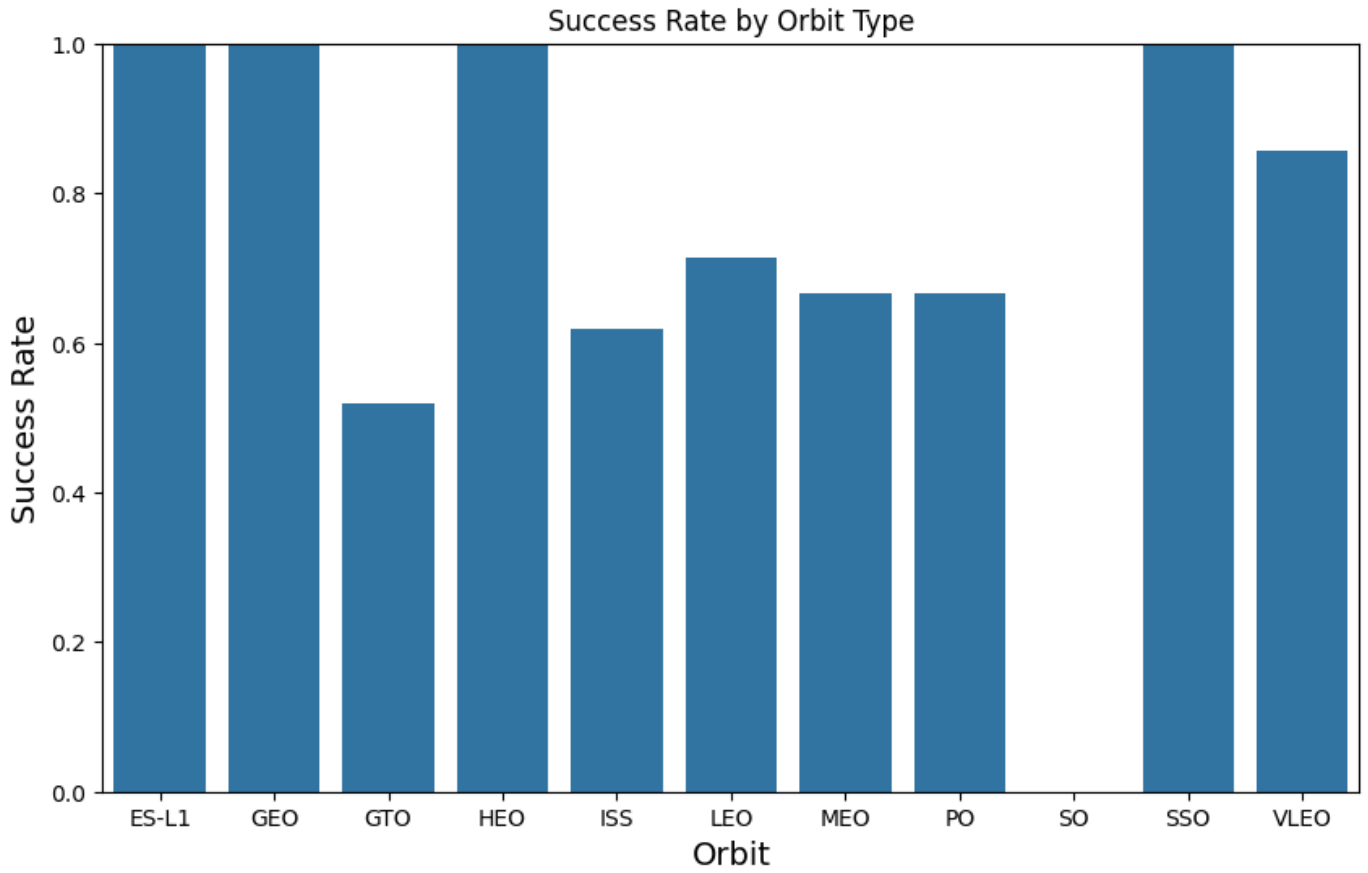


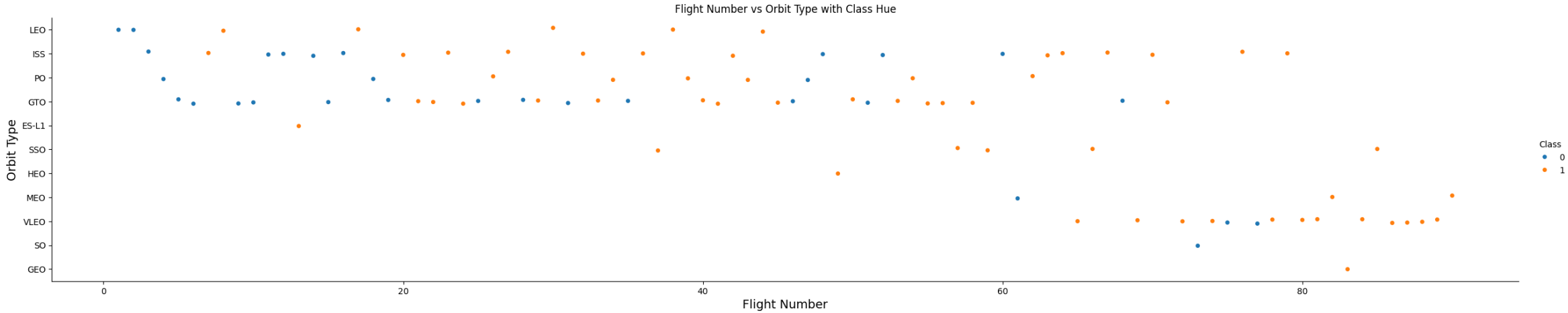
Payload vs. Launch Site

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

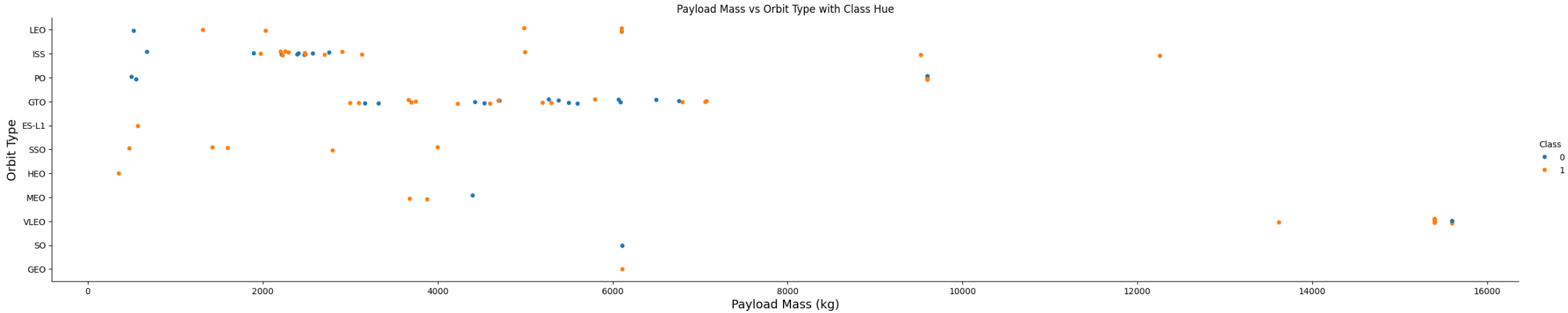
- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO





Flight Number vs. Orbit Type

- 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.

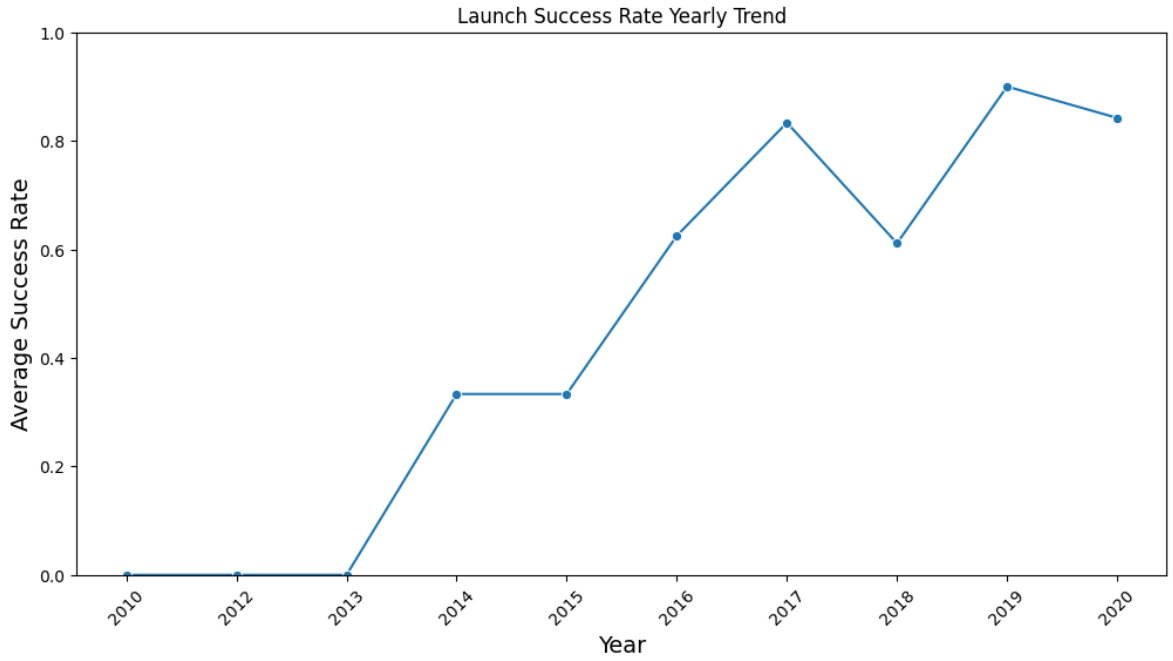


Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020.



Displaying the names of the unique launch sites in the space mission.

All Launch Site Names

```
[10]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

[10]:

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

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

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

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

Displaying 5 records where launch sites begin with the string 'CCA'.

```
[12]: %sql SELECT SUM("Payload_Mass__kg_") AS total_payload_mass FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';

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

[12]:

total_payload_mass
48213

Total Payload Mass

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

```
[13]: %sql SELECT AVG("Payload_Mass__kg_") AS avg_payload_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';

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

[13]:

avg_payload_mass
2928.4

Average Payload Mass by F9 v1.1

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

```
[14]: %sql SELECT MIN("Date") AS first_successful_ground_pad FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success%' AND "Landing_Outcome" LIKE '%ground pad%';

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

[14]:

first_successful_ground_pad
2015-12-22

First Successful Ground Landing Date

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

```
[17]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success%' AND "Landing_Outcome" LIKE '%drone ship%' AND "Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_" <
 * sqlite:///my_data1.db
Done.
```

[17]:

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

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 greater than 4000 but less than 6000.

[16]:
```
%sql SELECT "Mission_Outcome", COUNT(*) AS count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

[16]:

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

Total Number of Successful and Failure Mission Outcomes

Listing the total number of successful and failure mission outcomes.

```
[17]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);

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

[17]:

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

Boosters Carried Maximum Payload

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

```
[18]: %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" LIKE '%Failure%' \
      AND "Landing_Outcome" LIKE '%drone ship%';
```

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

[18]:

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

2015 Launch Records

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

```
[28]: %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.
```

[28]:

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

Rank Landing Outcomes 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 date 2010-06-04 and 2017-03-20, in descending order

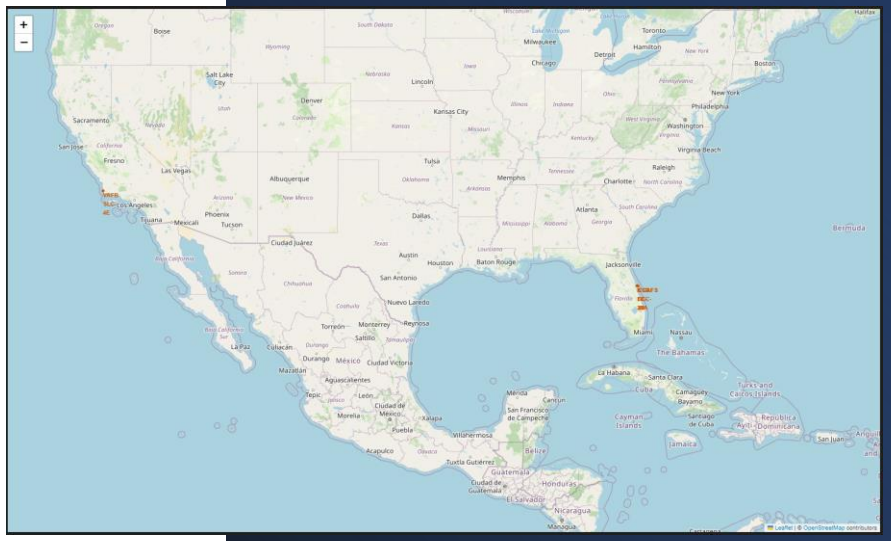
Section 3

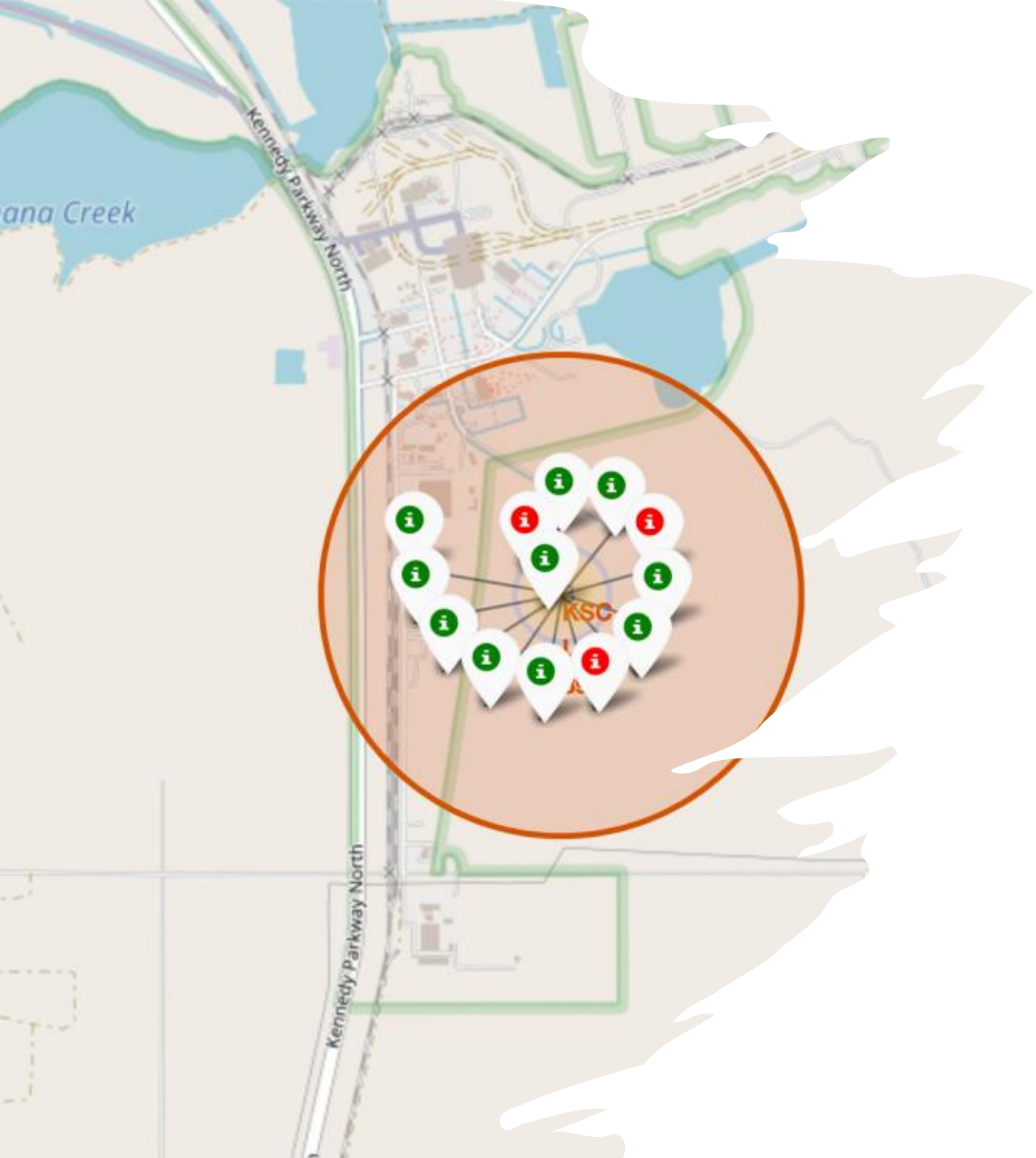
Launch Sites Proximities Analysis

All launch sites’ location markers on a global map

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



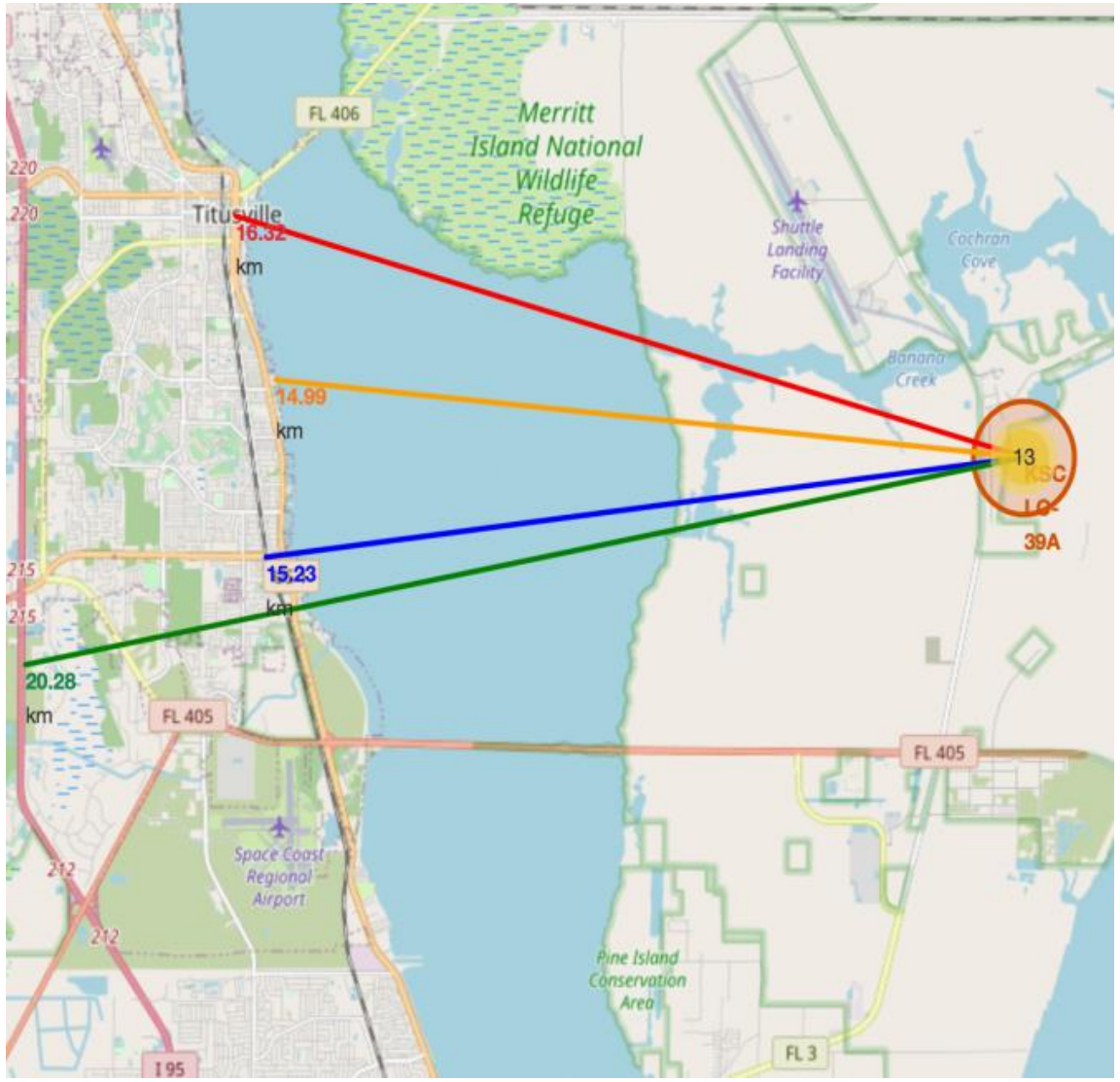


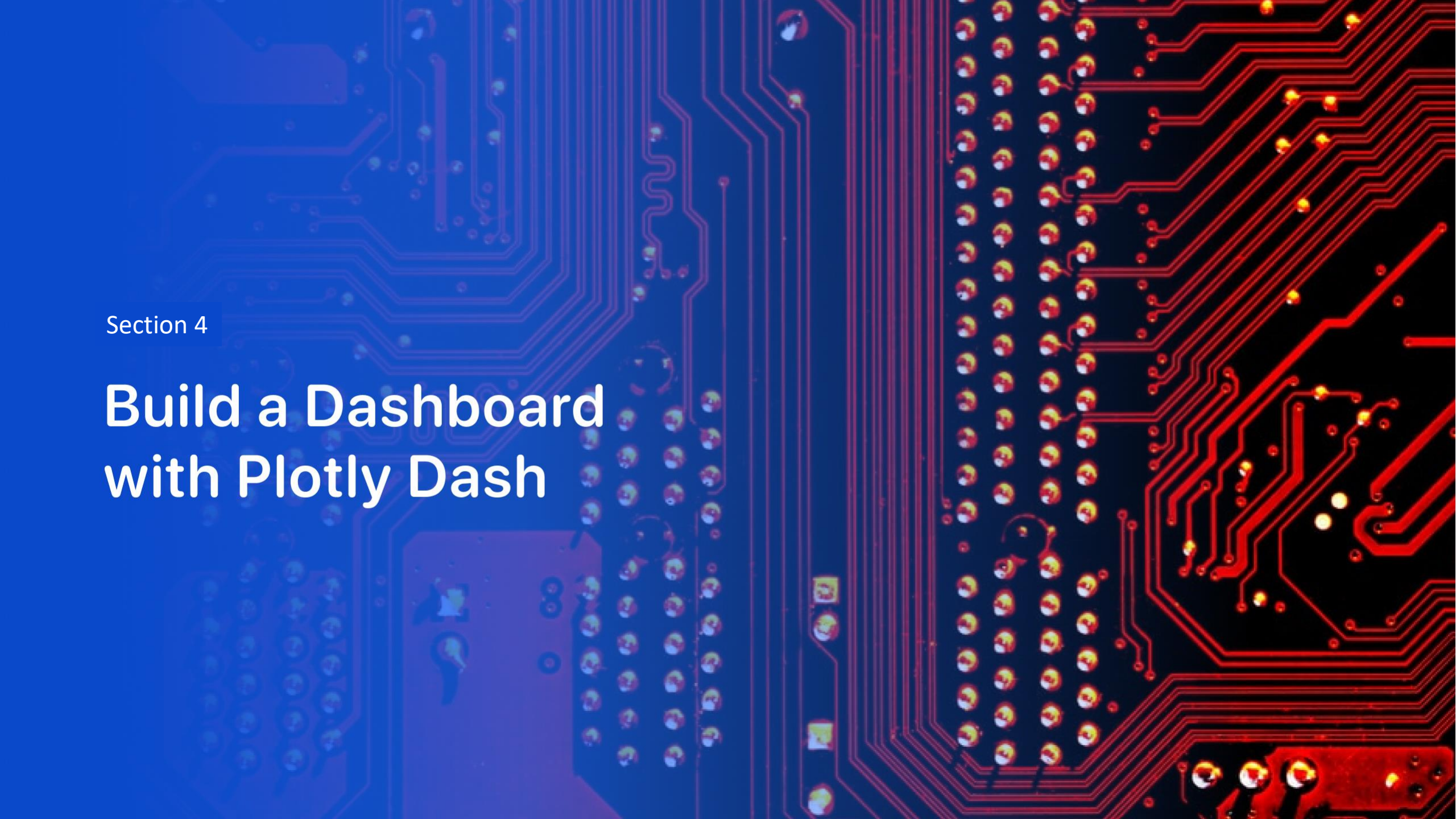
Colour-labeled launch records on the map

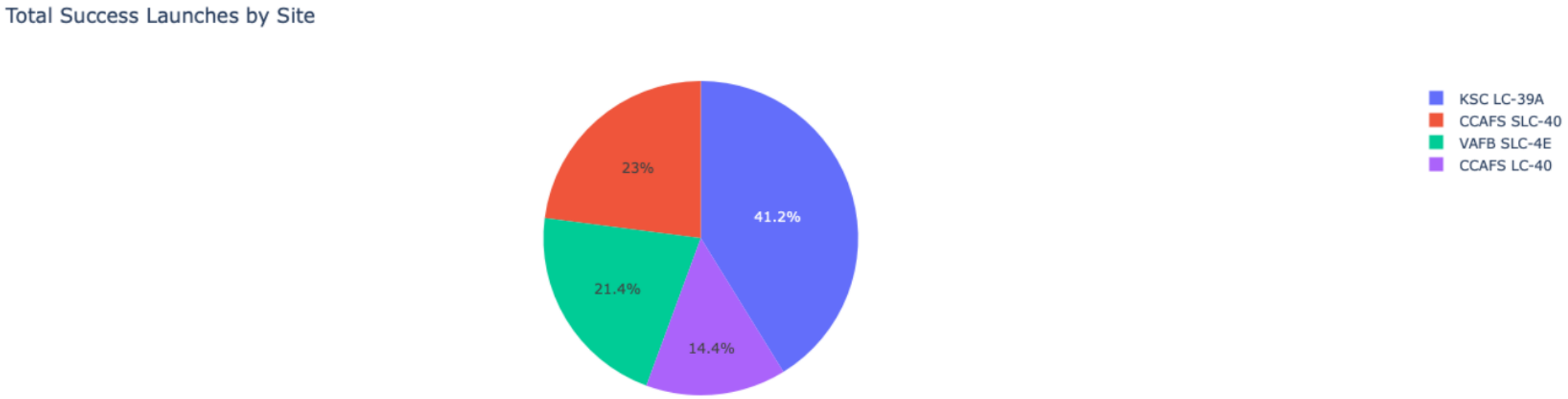
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- - Green Marker = Successful Launch
- - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

Distance from the launch site KSC LC-39A to its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is: - relative close to railway (15.23 km) - relative close to highway (20.28 km) - relative close to coastline (14.99 km) • Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km). • Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

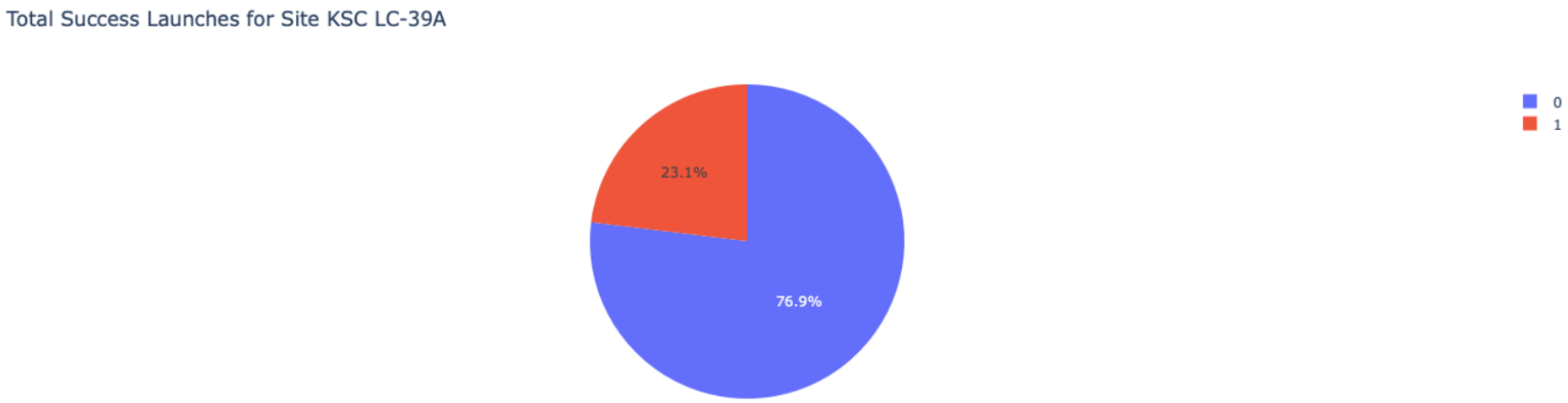



Section 4
Build a Dashboard with Plotly Dash



Launch success count for all sites

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



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

Launch site with highest launch success ratio

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

Payload Mass vs. Launch Outcome for all sites

Section 5

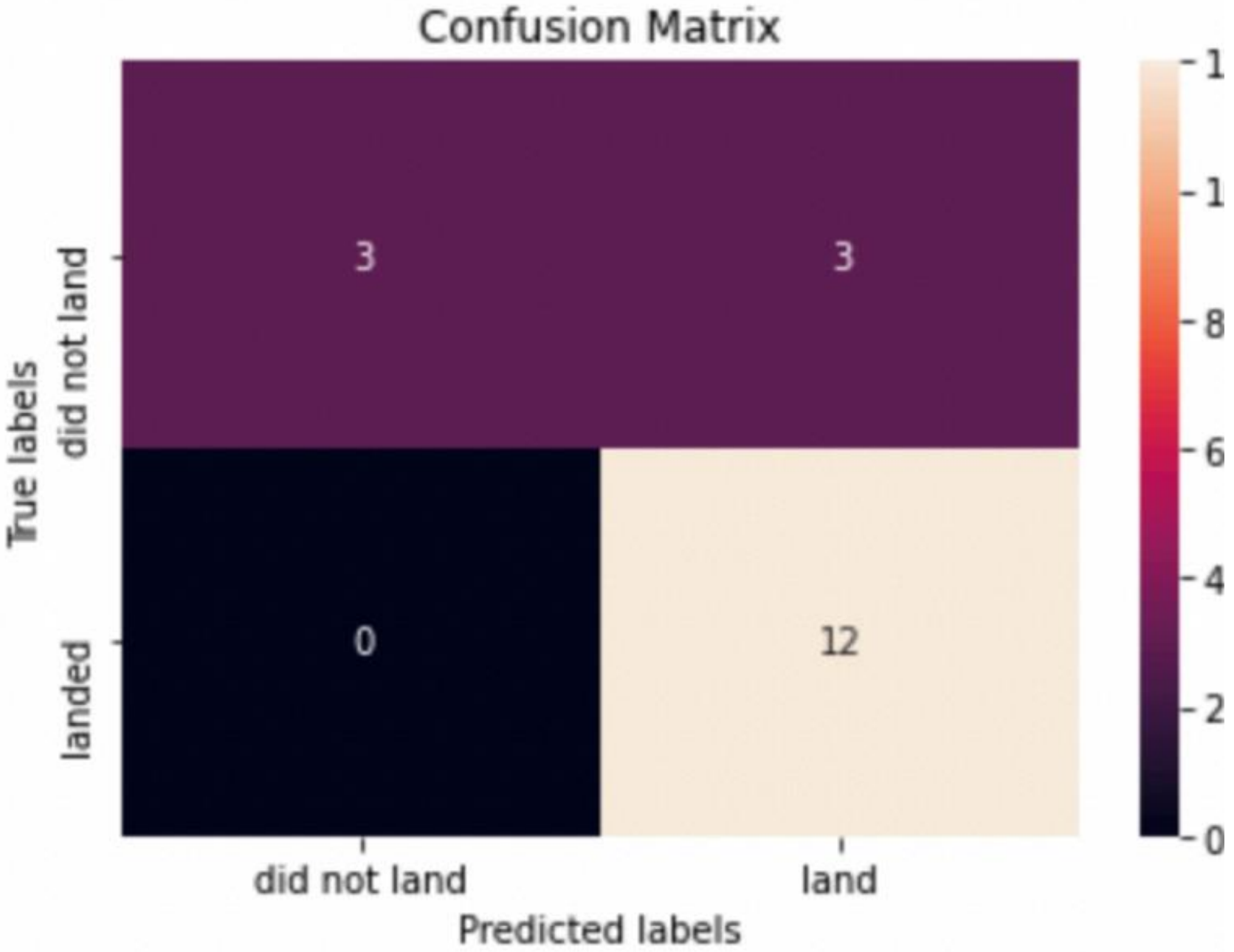
Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

The Decision Tree Model proves to be the most effective algorithm for this dataset.

Launches involving lower payload masses tend to yield more favorable outcomes compared to those with higher payloads.

The majority of launch sites are located near the Equator, with all positioned very close to coastlines.

Launch success rates have shown a consistent upward trend over time.

Among all the launch sites, KSC LC-39A records the highest success rate.

The orbits ES-L1, GEO, HEO, and SSO have achieved a perfect 100% success rate.

Thank you!