IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Taufik Khan

20th Jan 2025

Outline

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

Executive Summary

- Summary of methodologies

Data collection

Data wrangling

EDA with Data Visualization

EDA with SQL

Interactive map with Folium

Building a Dashboard with Plotly Dash

Predictive analysis (Classification)

- Summary of all results

EDA Results

Interactive analytics proof with screenshots

Predictive analysis results

Introduction

- Project background and context

This project is based off the experiment to predict if SpaceX will use the First Stage of their rockets. SpaceX is the most successful company of the commercial space age, their main goal is to make space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of $62MM USD ther providers cost goes up to $165MM USD each, if SpaceX can reuse the first stage, their savings can boost up.

If we can determine if the first stage will land, we can determine the cost of a launch, based on public information and ML Models, we can predict if SpaceX will reuse the first Stage.

- Problems you want to find answers

How does 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 landing increase over the years?

What is the best algoritm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

The main data was collected using SpaceX Rest API and WebScrapping tools from Wikipedia

- Perform data wrangling

Filter the data

Dealing with missing values

Using One Hot Encoding to prepare the data to a binary classification.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

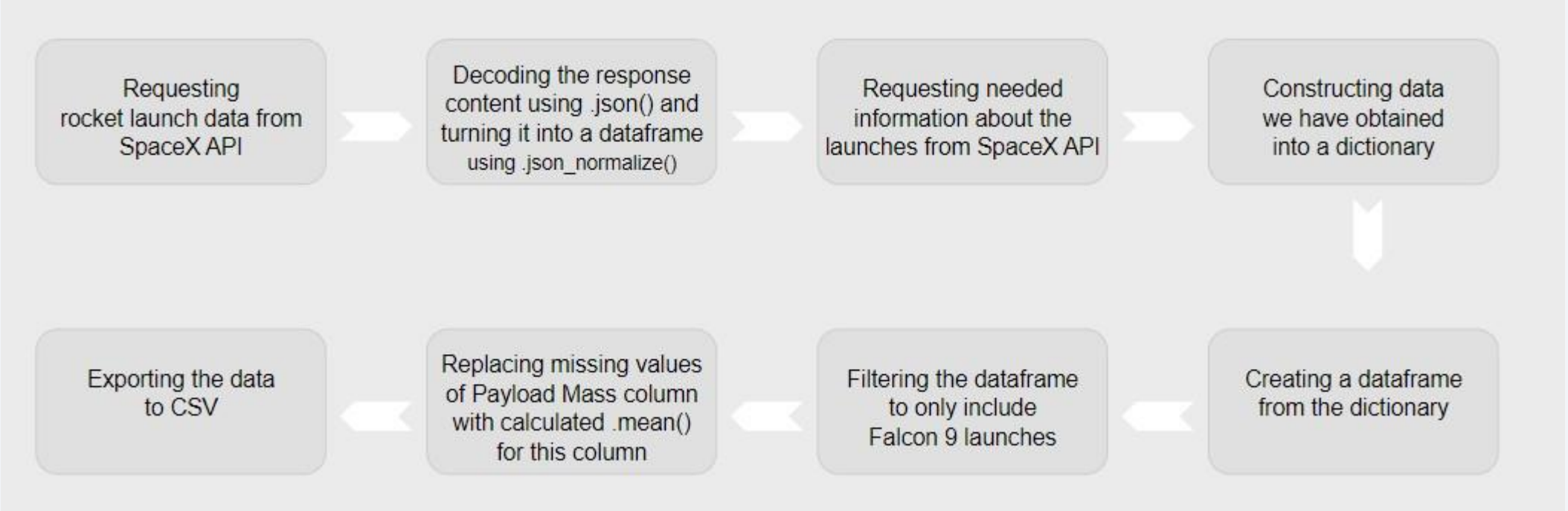
The process of data collection involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX Wikipedia entry, using both methods to get complete information about the launches for a more detailed analysis.

Using the REST API we obtain the following Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Gridfins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

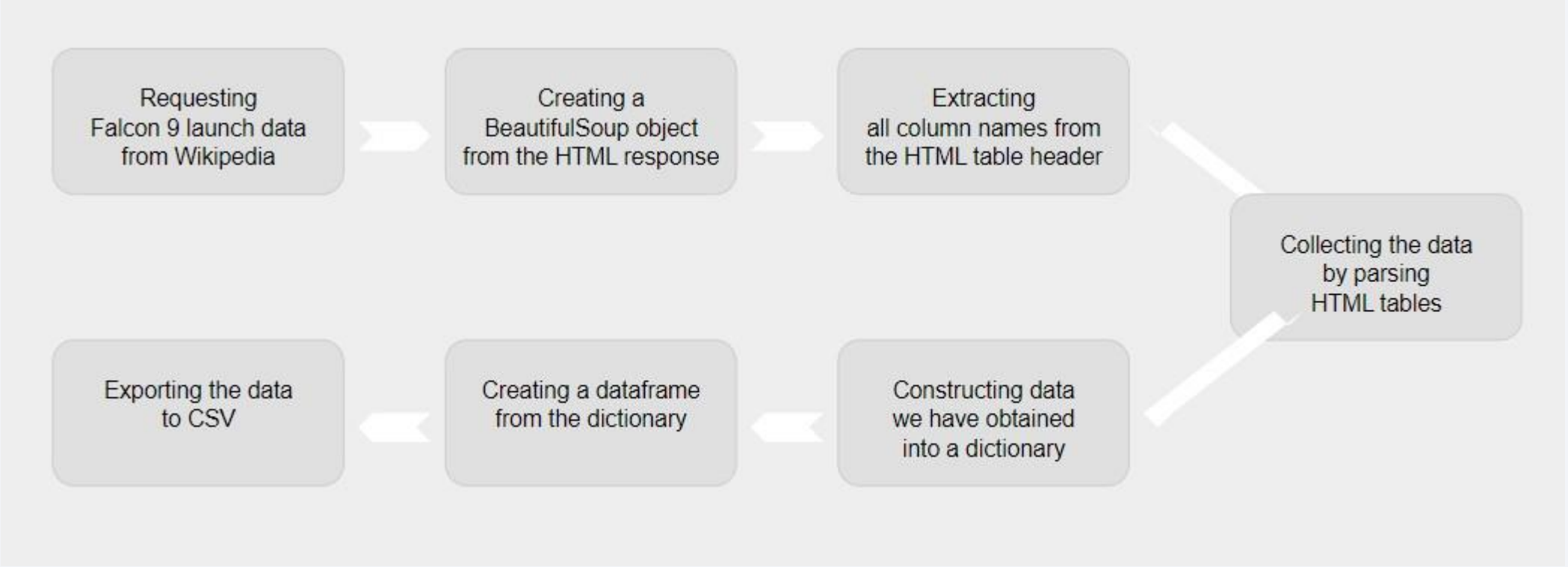
Using Wikipedia Web Scraping:

Flight N°, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date Time.

Data Collection - SpaceX API

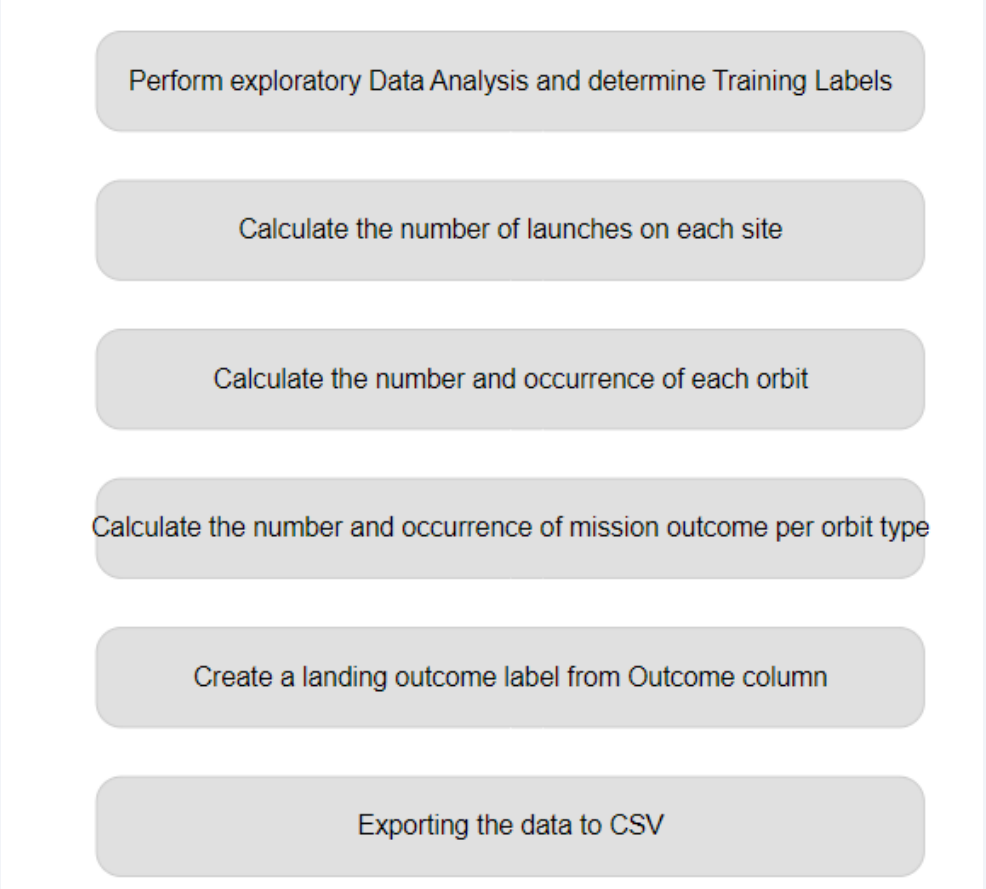


Data Collection - Scraping



Data Wrangling

- In our dataset, we track the success and failure rates of SpaceX rocket landings across different scenarios. Each landing attempt is categorized based on its intended landing location and outcome. Ocean landings indicate attempts to land in specific ocean regions, where a successful landing is marked as 'True Ocean' and a failed attempt as 'False Ocean'.
- For ground-based returns, we use the Return to Launch Site (RTLS) category. A 'True RTLS' indicates that the rocket successfully landed back on its designated ground pad, while a 'False RTLS' represents a failed landing attempt at the pad location.
- The third category involves landings on Autonomous Spaceport Drone Ships (ASDS). These are mobile landing platforms positioned in the ocean. A 'True ASDS' means the rocket successfully landed on the drone ship, while a 'False ASDS' indicates a failed landing attempt on these mobile platforms.
- To simplify our analysis and prepare the data for machine learning applications, we convert these detailed categorical outcomes into straightforward binary labels. We assign '1' to any successful landing attempt, regardless of the landing type, and '0' to any unsuccessful landing. This standardization helps us maintain consistency in our dataset while preserving the essential success/failure information for each launch.



EDA with Data Visualization

- In our Exploratory Data Analysis (EDA), we have generated various visualizations to better understand the relationships between key variables:
- We visualized seven main relationships: Flight Number versus Payload Mass, Flight Number versus Launch Site, Payload Mass versus Launch Site, Orbit Type versus Success Rate, Flight Number versus Orbit Type, Payload Mass versus Orbit Type, and the Success Rate Yearly Trend.
- For each type of visualization, we chose the most appropriate format based on the nature of the data:
- Scatter plots help us identify potential correlations and patterns between continuous variables. These relationships, when significant, can be particularly valuable as features in our machine learning models.
- Bar charts are our tool for visualizing and comparing discrete categories. They help us understand the distribution and relationship between categorical variables and their associated values in a clear and direct way.
- Line charts are specifically used to visualize temporal trends, allowing us to observe the evolution and patterns over time in our data series.

EDA with SQL

In our Exploratory Data Analysis using SQL, we executed several strategic queries to extract meaningful insights from our space mission database:

We began with foundational queries to understand our launch sites, including identifying unique launch locations and specifically examining sites with the 'CCA' designation.

For payload analysis, we focused on key metrics: calculating the total payload mass for NASA's Commercial Resupply Services (CRS) missions and analyzing the average payload capacity of the Falcon 9 v1.1 booster.

Our landing success analysis included several components:

- Identifying the historic milestone of the first successful ground pad landing
- Examining successful drone ship landings with payload masses between 4000 and 6000 units
- Compiling overall mission success and failure statistics

We conducted detailed performance evaluations by:

- Identifying booster versions that achieved maximum payload capacity
- Analyzing failed drone ship landings in 2015, including comprehensive details about booster versions and launch sites
- Creating a chronological ranking of landing outcomes between June 2010 and March 2017, focusing on both drone ship failures and ground pad successes

These queries helped us build a comprehensive understanding of SpaceX's launch and landing performance across different parameters and time periods.

Build an Interactive Map with Folium

In creating our interactive visualization map using Folium, we implemented three main components to analyze SpaceX launch sites and their characteristics:

First, we established base location markers across all launch facilities. We began by placing a marker for the NASA Johnson Space Center as our reference point, using precise latitude and longitude coordinates. We then expanded this to include all SpaceX launch sites, with each marker featuring a circular icon, interactive popup information, and text labels. This mapping helped visualize the geographical distribution of sites relative to the equator and coastlines.

For performance analysis, we implemented a color-coded marking system. Using a Marker Cluster approach, we designated successful launches in green and failed launches in red. This visual representation allows quick identification of launch sites with higher success rates and helps reveal any potential geographical patterns in launch outcomes.

To understand the infrastructure context, we added proximity analysis features. Using the Kennedy Space Center LC-39A as a case study, we created colored line overlays to display distances to key infrastructure elements. These measurements include connections to railways, highways, coastlines, and nearby urban centers, providing insight into the logistical considerations of launch site placement.

Build a Dashboard with Plotly Dash

In developing our interactive SpaceX launch analysis dashboard using Plotly Dash, we implemented several key visualization components to provide comprehensive insights:

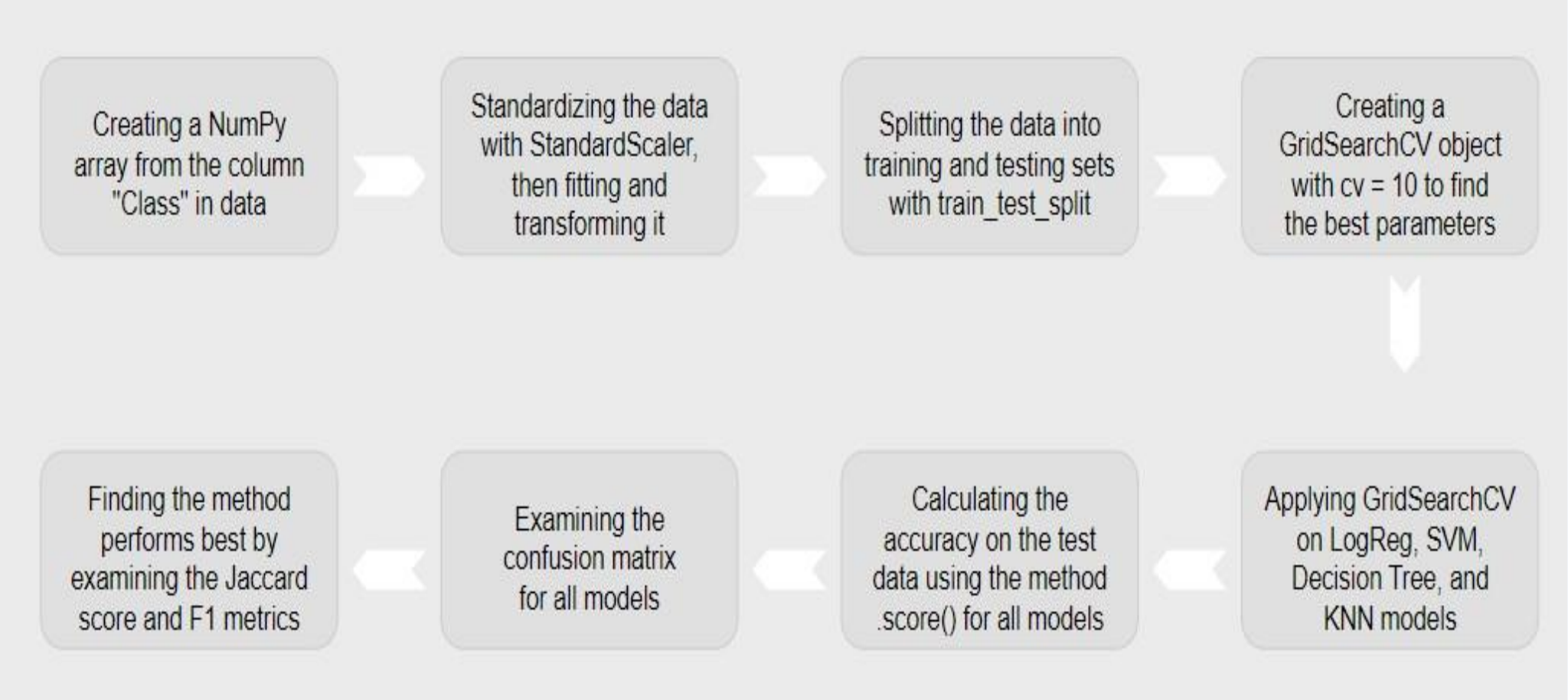
We created an intuitive user interface starting with a Launch Site selection feature. This dropdown menu serves as the primary control element, allowing users to filter and focus the analysis on specific launch locations of interest.

For success rate visualization, we implemented a dynamic pie chart system. This visualization adapts based on user selection, displaying either an overall success rate across all launch sites or detailed success-to-failure ratios for individually selected locations. This provides both broad and focused perspectives on launch performance.

To enable payload analysis, we incorporated an interactive range slider. This feature allows users to filter launches based on specific payload mass ranges, facilitating the investigation of how payload mass might influence mission outcomes.

The dashboard's analytical capabilities culminate in a sophisticated scatter plot that visualizes the relationship between payload mass and success rates across different booster versions. This correlation analysis helps identify potential patterns between payload capacity and mission success, offering valuable insights for future mission planning.

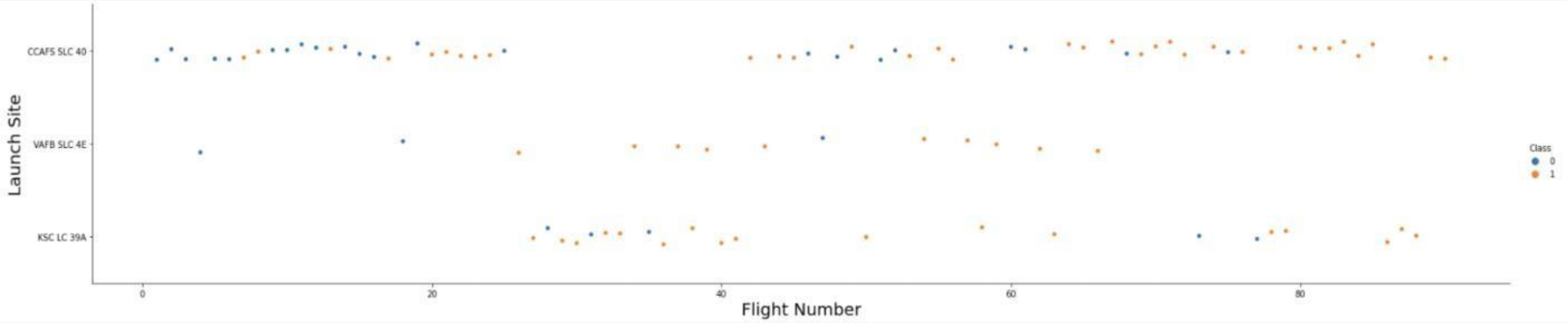
Predictive Analysis (Classification)



Section 2

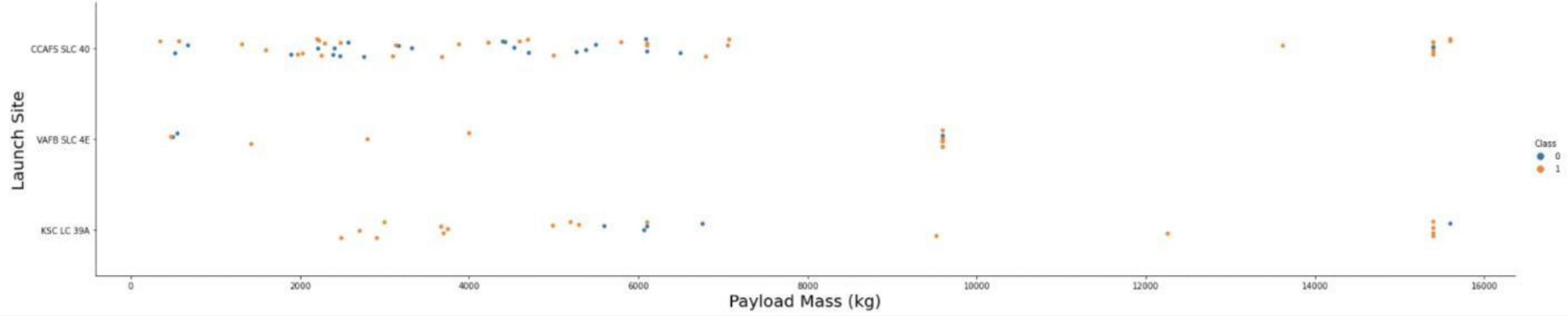
Insights drawn from EDA

Flight Number vs. Launch Site



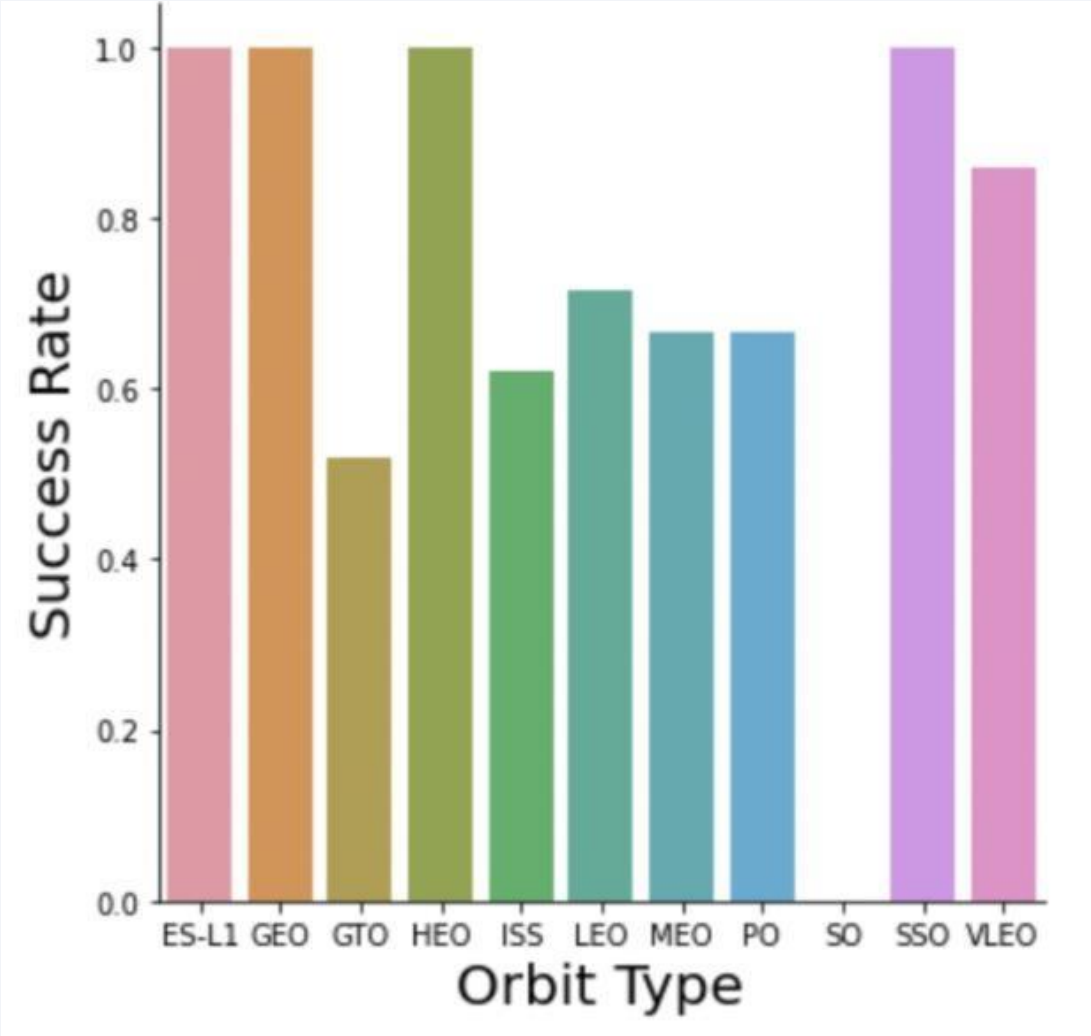
- This scatter plot visualizes SpaceX launches across different launch sites (Y-axis) over time using Flight Number (X-axis). Orange dots (1) represent successful launches while blue dots (0) indicate failures. The pattern shows CCAFS SLC 40 as the most frequently used site, while later flight numbers tend to have more successful launches across all sites, suggesting improved reliability over time.

Payload vs. Launch Site



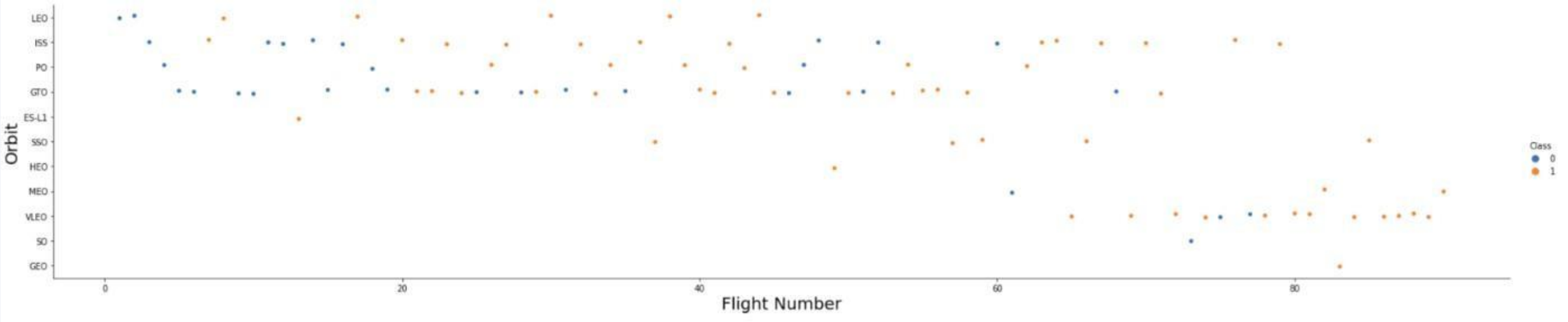
- This scatter plot shows the relationship between Payload Mass (kg) on the X-axis and Launch Sites on the Y-axis for SpaceX missions. Orange dots indicate successful launches while blue dots show failures. The distribution suggests that different launch sites handle varying payload ranges, with some sites showing better success rates with certain payload masses.

Success Rate vs. Orbit Type



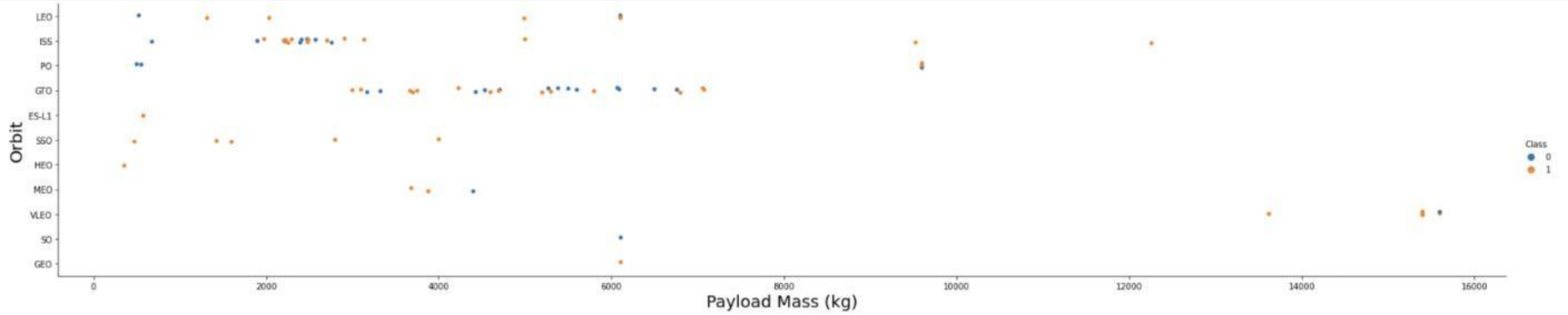
- This bar chart displays success rates (0-1 scale) across different orbit types for SpaceX launches. ES-L1, GEO, HEO, and SO orbits show the highest success rates (near 100%), while orbits like GTO and ISS missions show moderate success rates (around 50-70%). This visualization helps identify which orbit types have been historically more successful for SpaceX missions.

Flight Number vs. Orbit Type



- This scatter plot shows the distribution of SpaceX launches by orbit type (Y-axis) across different flight numbers (X-axis). Blue dots represent failed launches (0) and orange dots successful ones (1). The visualization indicates that certain orbits like LEO and ISS are more frequently targeted, and there's a trend toward higher success rates (more orange dots) in later flight numbers.

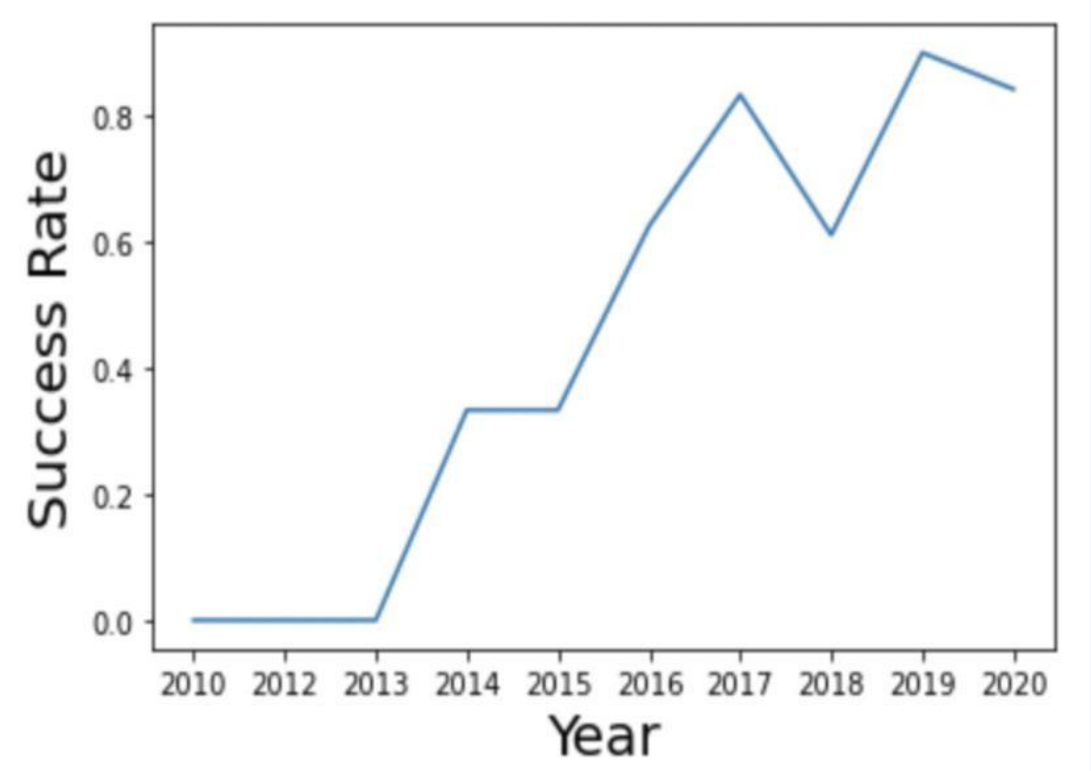
Payload vs. Orbit Type



- This scatter plot displays the relationship between Payload Mass (kg) and Orbit Types for SpaceX launches. The data shows different payload requirements for various orbit types, with LEO and ISS missions spanning a wide range of payload masses. Success (orange) and failure (blue) indicators suggest that heavier payloads don't necessarily correlate with higher failure rates across orbit types.

Launch Success Yearly Trend

- The launch sucess rate kept increasing since 2013 till 2020, with a brief decrease between 2017 - 2018



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

Out[4]:

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

- This chunk of code in SQL displays the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

This chunk of code displays 5 records where the launch site begin with the string “CCA”

Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

Out[6]:

total_payload_mass
45596

This chunk of code displays the total payload mass carried by the boosters launched by the NASA (CRS)

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

Out[7]:

average_payload_mass
2534

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

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

Out[8]:

first_successful_landing
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
```

Out[9]:

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

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

Total Number of Successful and Failure Mission Outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

Out[11]:

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

- The boosters versions that have carried the maximum payload mass

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
         where date between '2010-06-04' and '2017-03-20'
         group by landing__outcome
         order by count_outcomes desc;
```

Out[13]:

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

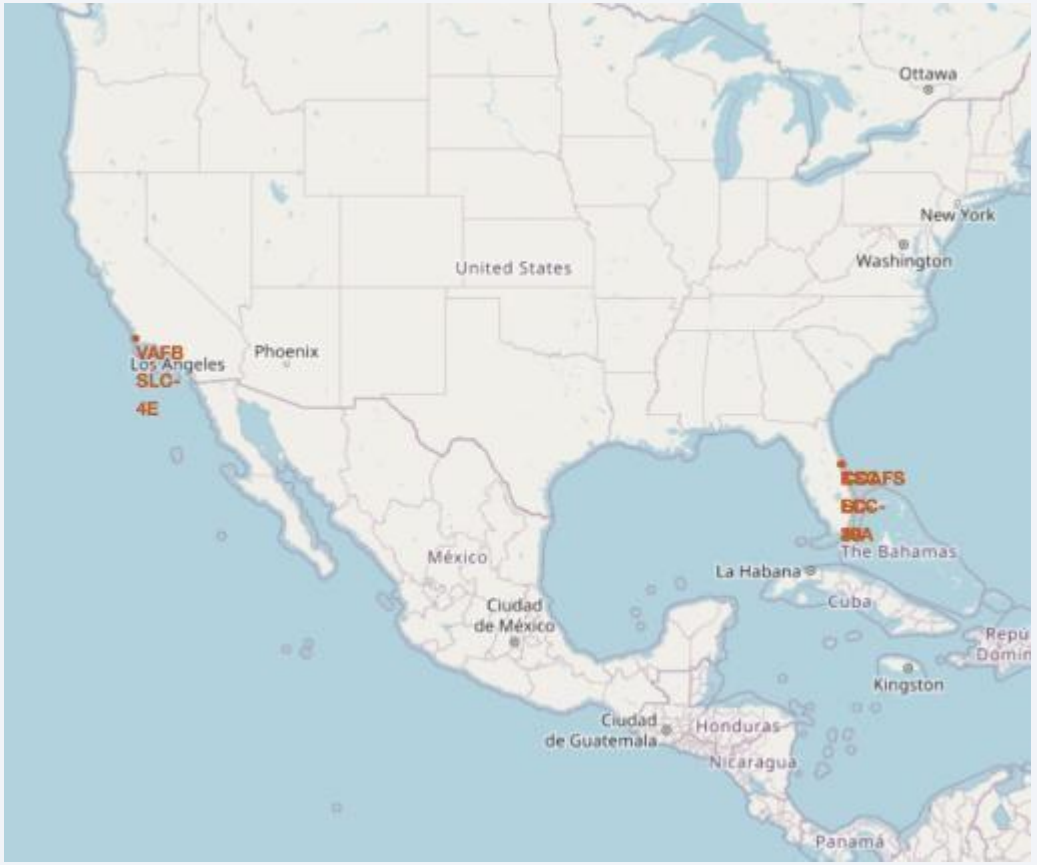
- Count of landing outcomes (Failure: drone ship, Success: Ground pad) between 2010 and 2017

Section 3

Launch Sites Proximities Analysis

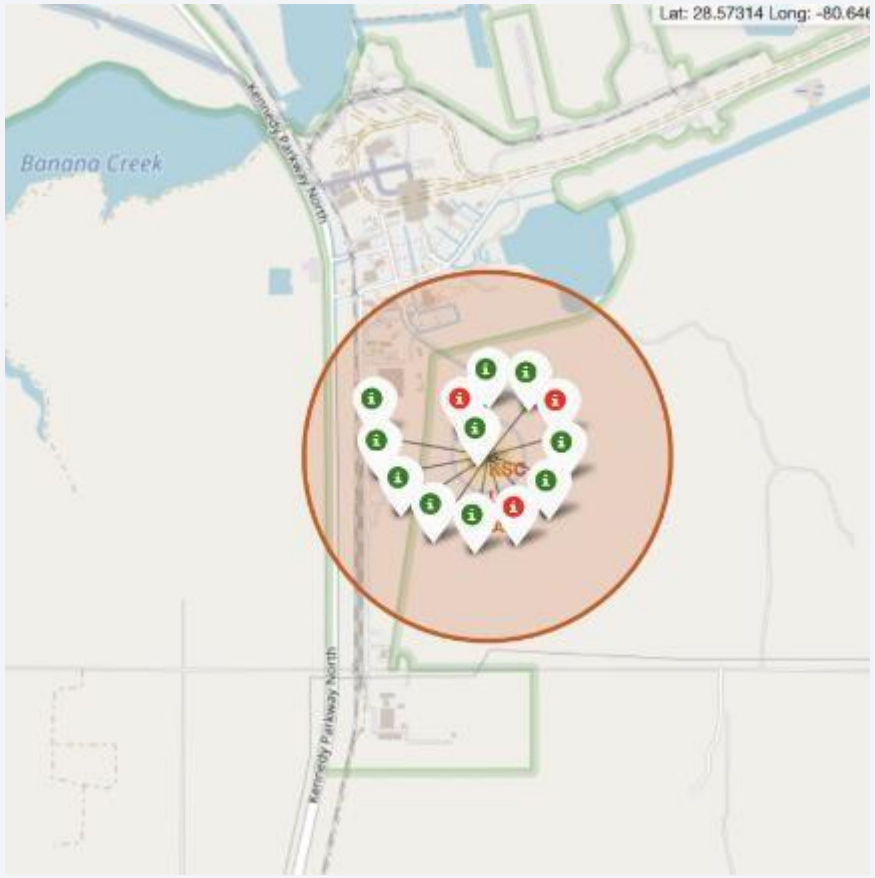
Global Map Markers

- This image explains the strategic placement of SpaceX launch sites. Launch sites are positioned near the equator to take advantage of Earth's faster rotational speed (1670 km/hour), which provides a natural boost to achieve orbital velocity. Additionally, coastal locations are chosen for safety reasons, as launching over water minimizes risks to populated areas in case of debris or launch failures.



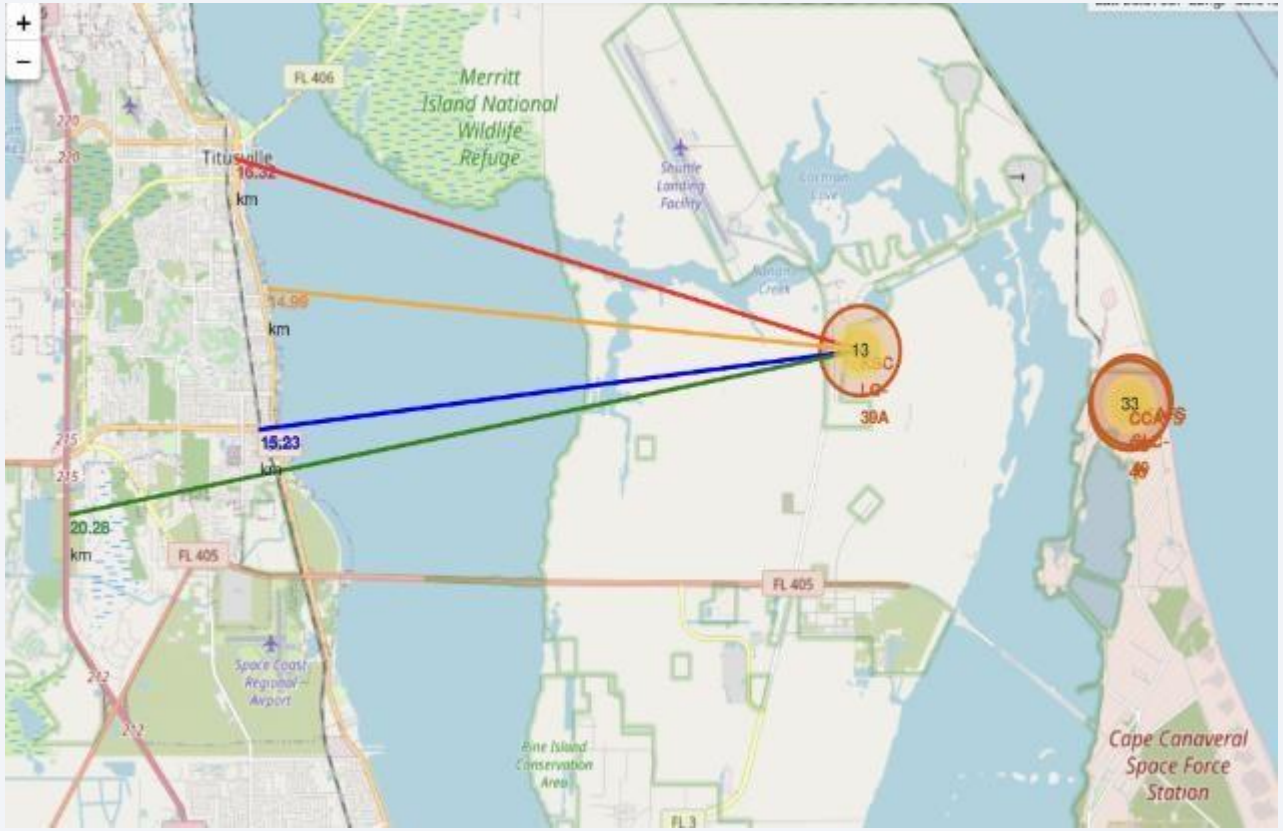
Colored labeled launch records on a map

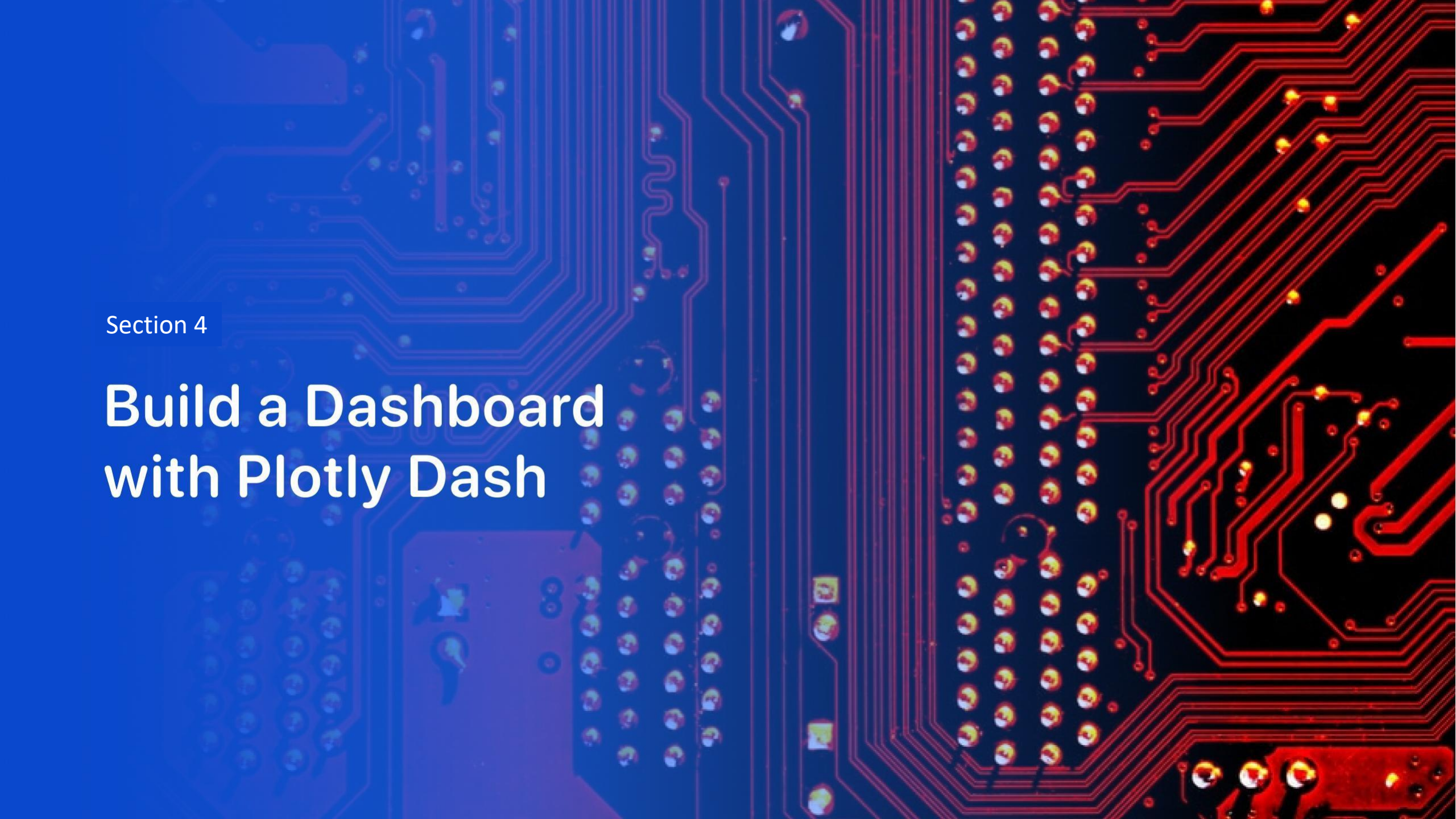
- This map shows launch outcomes at the KSC LC-39A site using color-coded markers: green for successful launches and red for failures. The cluster of markers reveals that this particular launch site maintains a notably high success rate, with green markers (successful launches) significantly outnumbering red ones (failed launches), indicating it's one of SpaceX's most reliable launch facilities.



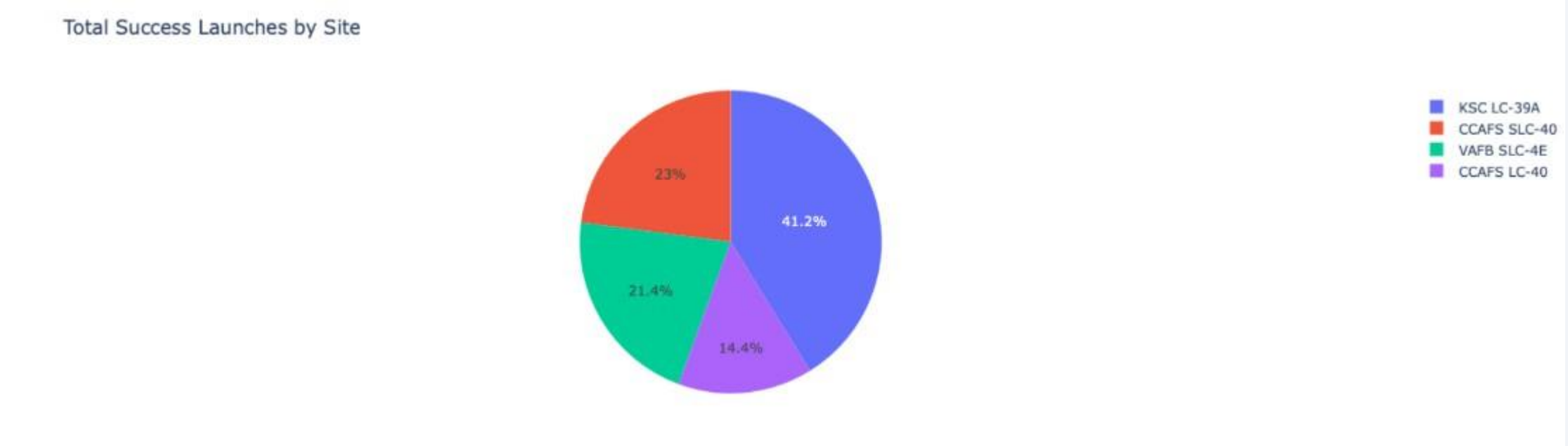
Distances from launches sites and its proximities

- This map displays proximity analysis for the KSC LC-39A launch site, showing its strategic positioning relative to key infrastructure: 15.23 km from railway, 20.28 km from highway, 14.99 km from coastline, and 16.32 km from the nearest city (Titusville). The distance measurements emphasize safety considerations, as a failed launch could cover 15-20 km in seconds, highlighting the importance of maintaining safe distances from populated areas.



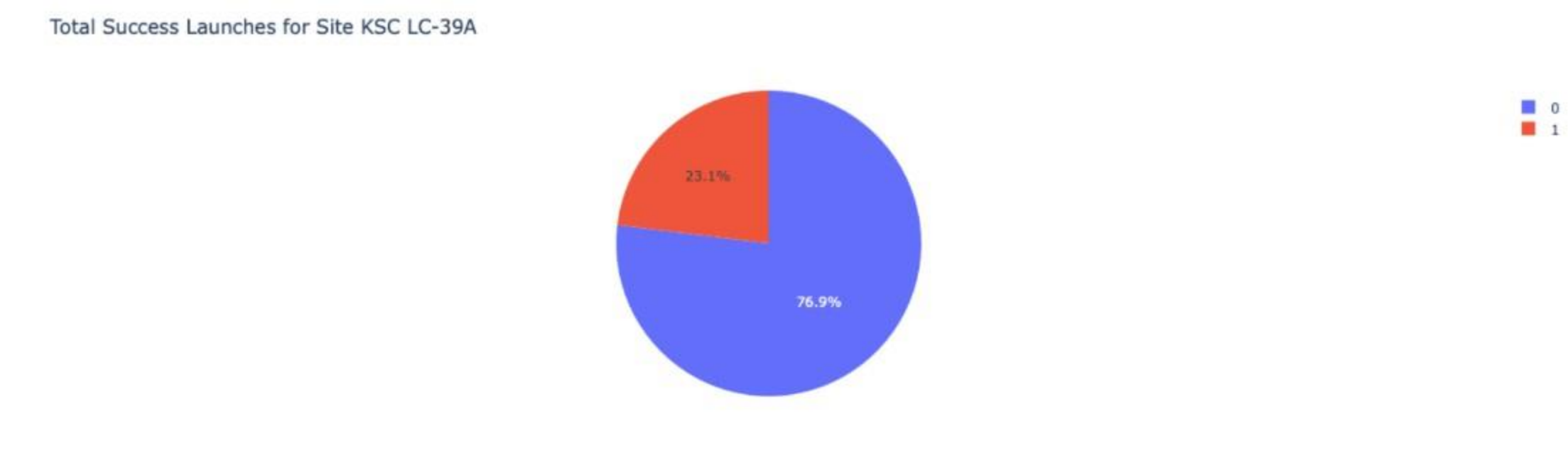

Section 4
Build a Dashboard with Plotly Dash

Total success launches by site



- The chart shows that KSC LC-39A has the most successful launches compared to the other sites

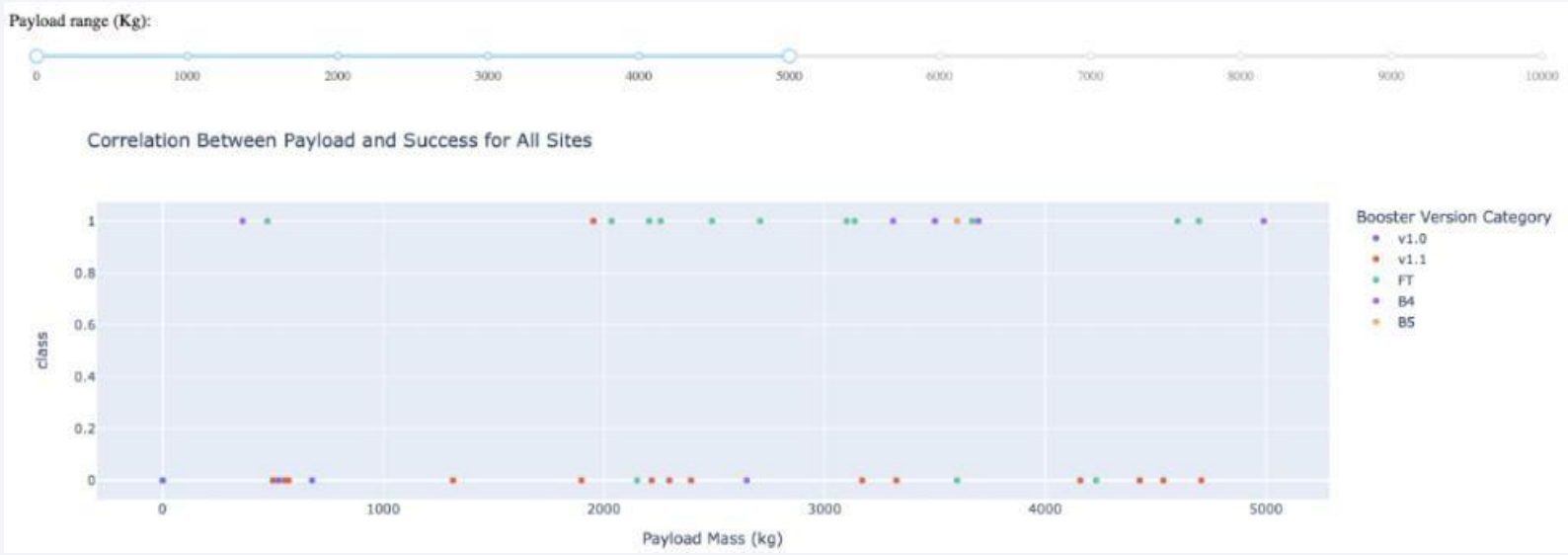
Total success launches for Site KSC LC-39A



- KSC LC-39A has the highest launch success rate with 10 successful launches and only 3 failed landings

<Dashboard Screenshot 3>

- The payloads between 2000 and 5500 kg have the highest success rate



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The model evaluation metrics reveal interesting performance patterns across both test and full datasets. While the test set (n=18) shows identical performance across all models (Jaccard=0.80, F1=0.89, Accuracy=0.83), the full dataset evaluation demonstrates more differentiated results.

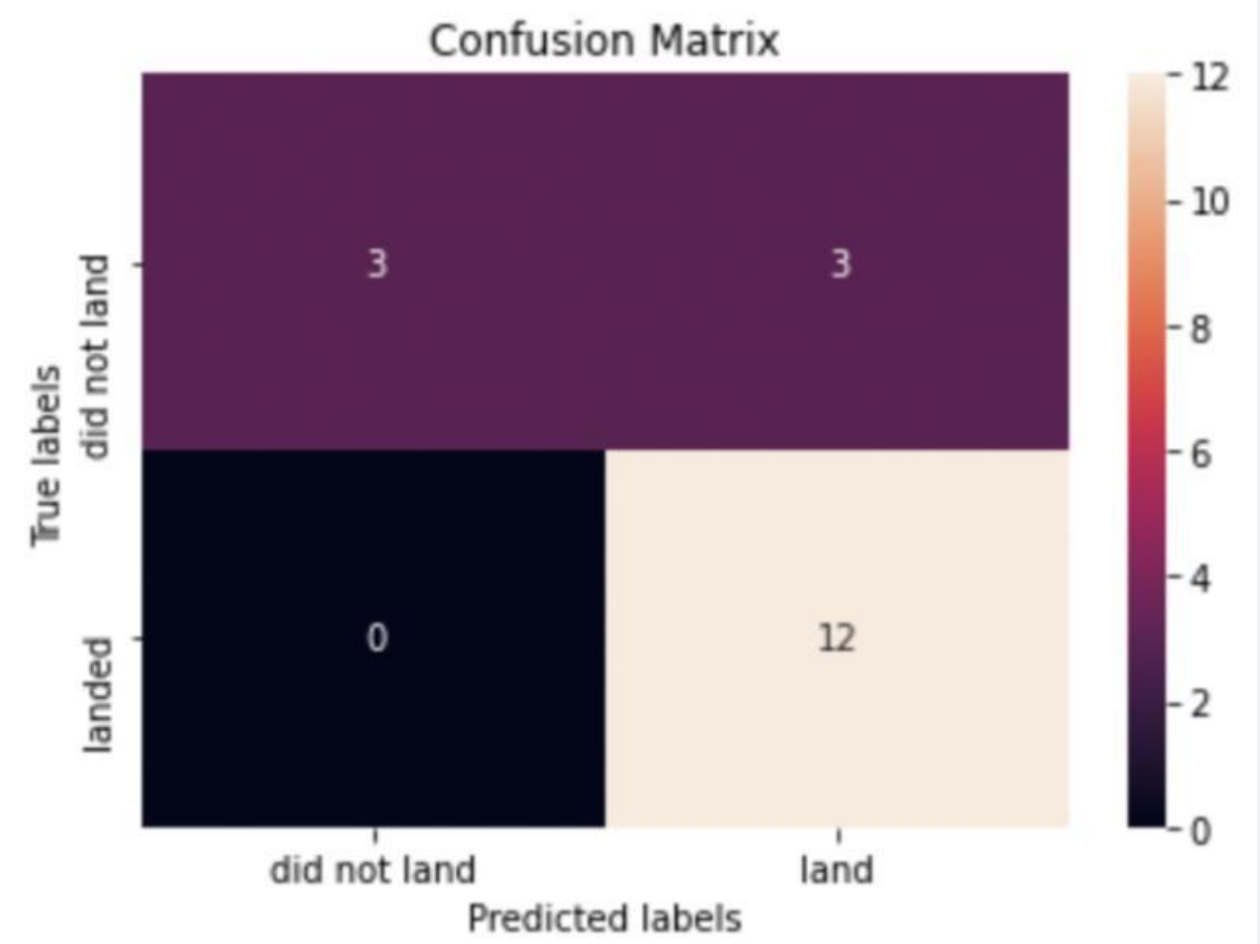
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

- The Decision Tree classifier emerges as the optimal model with superior metrics (Accuracy=0.911, F1=0.938, Jaccard=0.882), outperforming both the SVM (Accuracy=0.878, F1=0.916) and traditional LogReg (Accuracy=0.867, F1=0.909) approaches.
- The homogeneous test set scores can be attributed to the limited sample size, highlighting the importance of comprehensive evaluation on larger datasets for robust model selection.

	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

- Logistic regression can distinguish between classes, however, the major problem is false positives.



Conclusions

Based on the comprehensive analysis presented in the report, here are the key conclusions:

Data accessibility: The combination of SpaceX API and web scraping provided a robust dataset for analysis of launch outcomes

Launch site performance: KSC LC-39A emerged as the most successful launch site with a 76.9% success rate and highest number of successful launches

Payload optimization: Launches carrying payloads between 2000-5500 kg showed the highest success rates across all sites

Model performance: The Decision Tree classifier proved most effective for predicting launch outcomes, achieving 91.1% accuracy and outperforming other models (SVM, LogReg, KNN)

Launch success trend: SpaceX demonstrated significant improvement in launch reliability over time, with success rates increasing steadily from 2013 to 2020

Safety considerations: Launch sites are strategically positioned near coastlines and maintain safe distances (15-20 km) from populated areas and infrastructure

Thank you!