IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francesco Cardinale
15/01/2024

Outline

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

Executive Summary

- **Summary of Methodologies**
  - Data Collection: Acquisition of historical data relating to past launches, sites and Falcon 9
  - Data Wrangling: Cleaning and transforming data for analysis
  - Exploratory Data Analysis (EDA): Understanding dataset characteristics
  - Visual Analytics: Leveraging Maps and Dashboard for interactive analysis
  - Predictive Analysis: Implementing Machine learning models for landing predictions
- **Summary of all Results**
  - Temporal Trend Analysis: Success rates have improved since 2013
  - Launch Site Impact: Varying success rates at different launch sites
  - Attribute Correlation: Identifying attributes influencing successful landings
  - Visual Analytics Dashboard: Interactive tool for stakeholders

Introduction

- **Project background and context**

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; Other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- **Problems we want to find answers**

  - Determine if the Falcon 9 first stage will land successfully
  - Impact on launch cost and overall efficiency
  - Explore influential features for predicting successful landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Ensure data quality through cleaning, handling missing values, and one-hot encoding
- 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 models with various algorithms, tune hyperparameters using GridSearch, and evaluate performance by scoring with suitable metrics

Data Collection

- SpaceX API Data Collection:
  - Perform a GET request to SpaceX API for launch data.
  - Filter data for Falcon 9 launches.
  - Address missing values by replacing with the calculated mean for PayloadMass.
- Web Scraping Data Collection:
  - Request Falcon 9 Launch Wiki page.
  - Use BeautifulSoup to parse HTML and extract launch records.
  - Extract column/variable names from HTML table header.
  - Create a DataFrame with parsed launch HTML tables.

Data Collection – SpaceX API

SpaceX API Calls

- Initiate GET request to SpaceX API for launch data.
- Distinct API calls for targeted data extraction, like *booster names*, *launch site* details, *payload* information and *core landing* specifics

Filter for Falcon 9 Launches

- Parse and filter the DataFrame to include only *Falcon 9* launches

Handling Missing Values

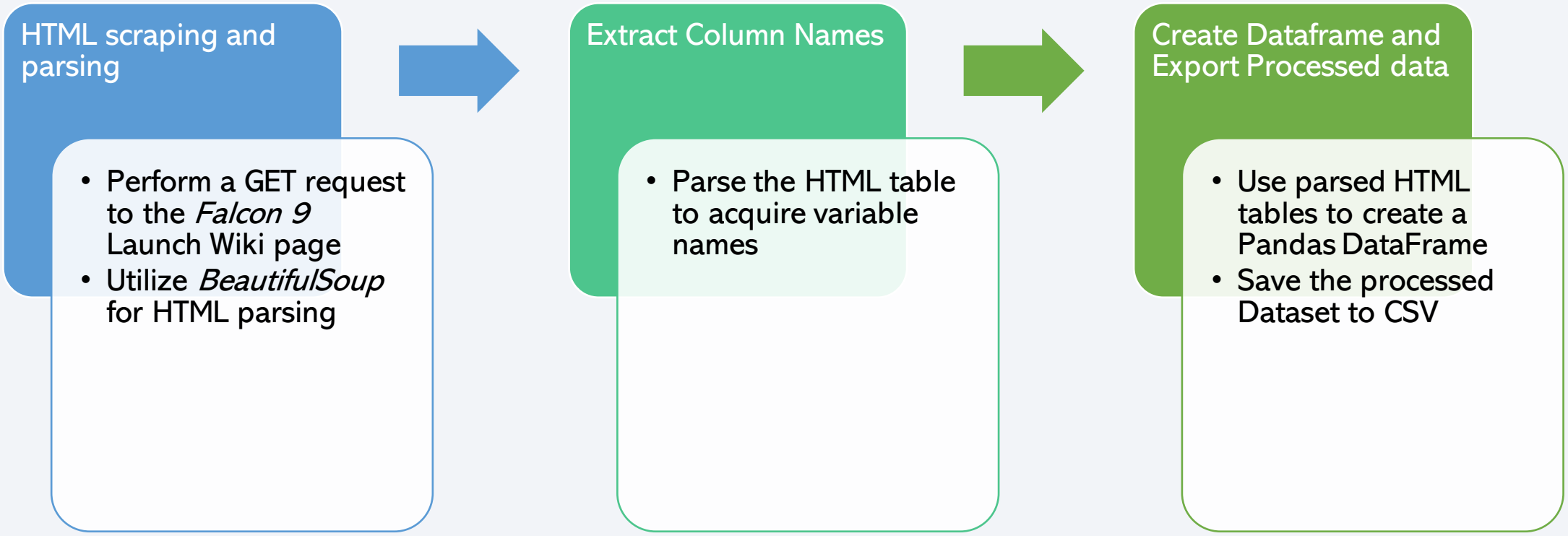
- Calculate mean for *PayloadMass*
- Replace NaN values with calculated mean

Export Processed Data

- Save the processed dataset to CSV

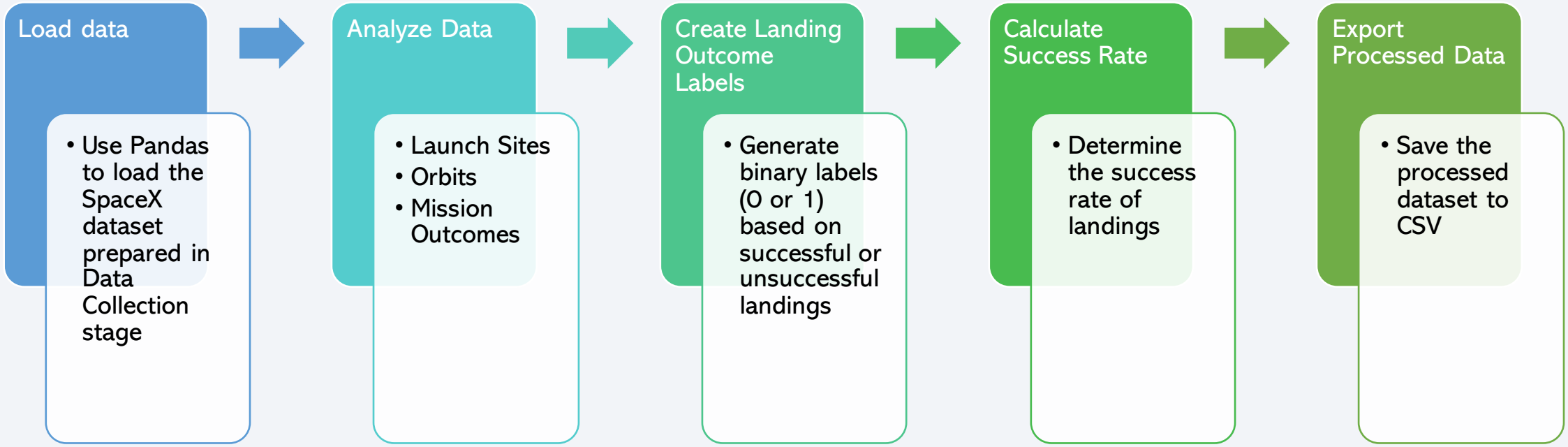
GitHub URL: Data Collection SpaceX API Notebook

Data Collection - Scraping



GitHub URL: Data Wrangling SpaceX Notebook

Data Wrangling



GitHub URL: Data Collection SpaceX WebScraping Notebook

EDA with Data Visualization

- **FlightNumber vs. PayloadMass:** Scatter plot showing that higher FlightNumber correlates with a higher chance of success, while heavier payloads tend to have a lower success rate.
- **FlightNumber vs. LaunchSite:** Scatter plot revealing the relationship between FlightNumber and LaunchSite, indicating different success rates at various launch sites.
- **Payload vs. LaunchSite:** Scatter plot demonstrating that, for the VAFB-SLC launch site, there are no launches for heavy payloads (above 10000).
- **Orbit Success Rate:** Bar chart displaying success rates for different orbits.
- **FlightNumber vs. Orbit:** Scatter plot indicating the relationship between FlightNumber and Orbit, highlighting success patterns in LEO orbits.
- **Payload vs. Orbit:** Scatter plot illustrating the payload's impact on success rates in different orbits.
- **Launch Success Yearly Trend:** Line chart depicting the increasing trend in average success rates from 2013 onwards.

GitHub URL: EDA Data Visualization SpaceX Notebook

EDA with SQL

- **Dataset Overview**: SpaceX dataset captures payload records from historic missions.
- **SQL Queries**: Executed queries to identify unique launch sites, filter records by launch site prefix, and calculate payload metrics.
- **Mission Outcomes**: Analyzed mission outcomes, determining total payload mass for NASA (CRS) and average payload mass for F9 v1.1 boosters.
- **Landing Success**: Explored successful landing dates, boosters achieving drone ship success with specific payload mass ranges.
- **Temporal Trends**: Investigated trends from 2015, listing booster names with failure outcomes and ranked landing outcomes between 2010-2017.

GitHub URL: EDA with SQL Notebook

Build an Interactive Map with Folium

- **Launch Site Markers**
  - Purpose: To visually represent the exact locations of each launch site.
  - Reasoning: Essential for understanding the spatial distribution of launch facilities.
- **Marker Clusters**
  - Purpose: Highlight success/failure outcomes at each site.
  - Reasoning: Provides an immediate overview of success rates at each launch site. Clustering prevents map clutter.
- **Distance Markers and Polylines**
  - Purpose: Show distances between launch sites and key features (coastline, highway, railroad, city).
  - Reasoning: Offers insights into strategic placement considerations, such as safety distances from cities and proximity to transportation infrastructure.

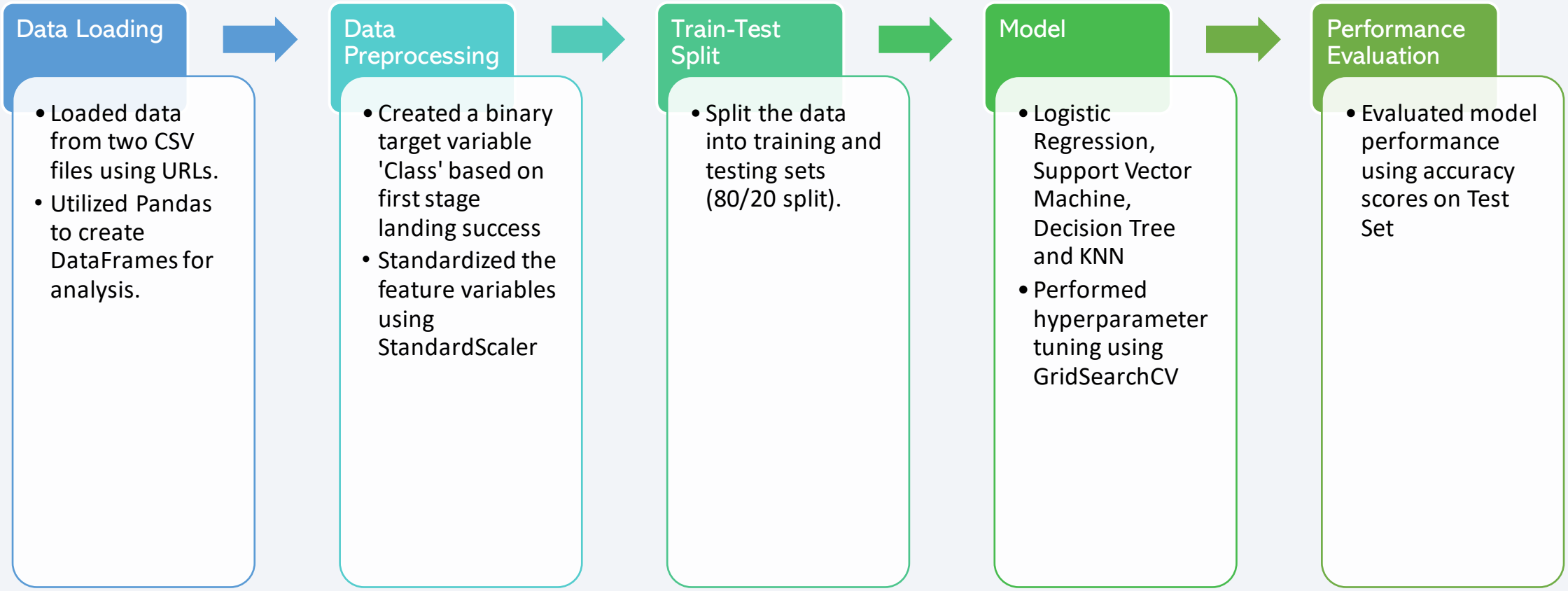
GitHub URL: Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- **Launch Site Drop-down**
  - Purpose: Select different launch sites
  - Reasoning: Analyze success counts and rates by launch site
- **Callback Function for success-pie-chart**
  - Purpose: Render a dynamic pie chart showing launch success counts
  - Reasoning: Visually compare success counts for all or a selected site
- **Range Slider for Payload (dcc.RangeSlider)**
  - Purpose: Choose payload ranges
  - Reasoning: Explore correlation between payload mass and launch success
- **Success Payload Scatter Chart**
  - Purpose: Generate scatter plot showing payload mass vs. launch success, color-labeled by booster version
  - Reasoning: Explore correlations between payload, booster versions, and launch outcomes

GitHub URL: Dashboard with Plotly Dash

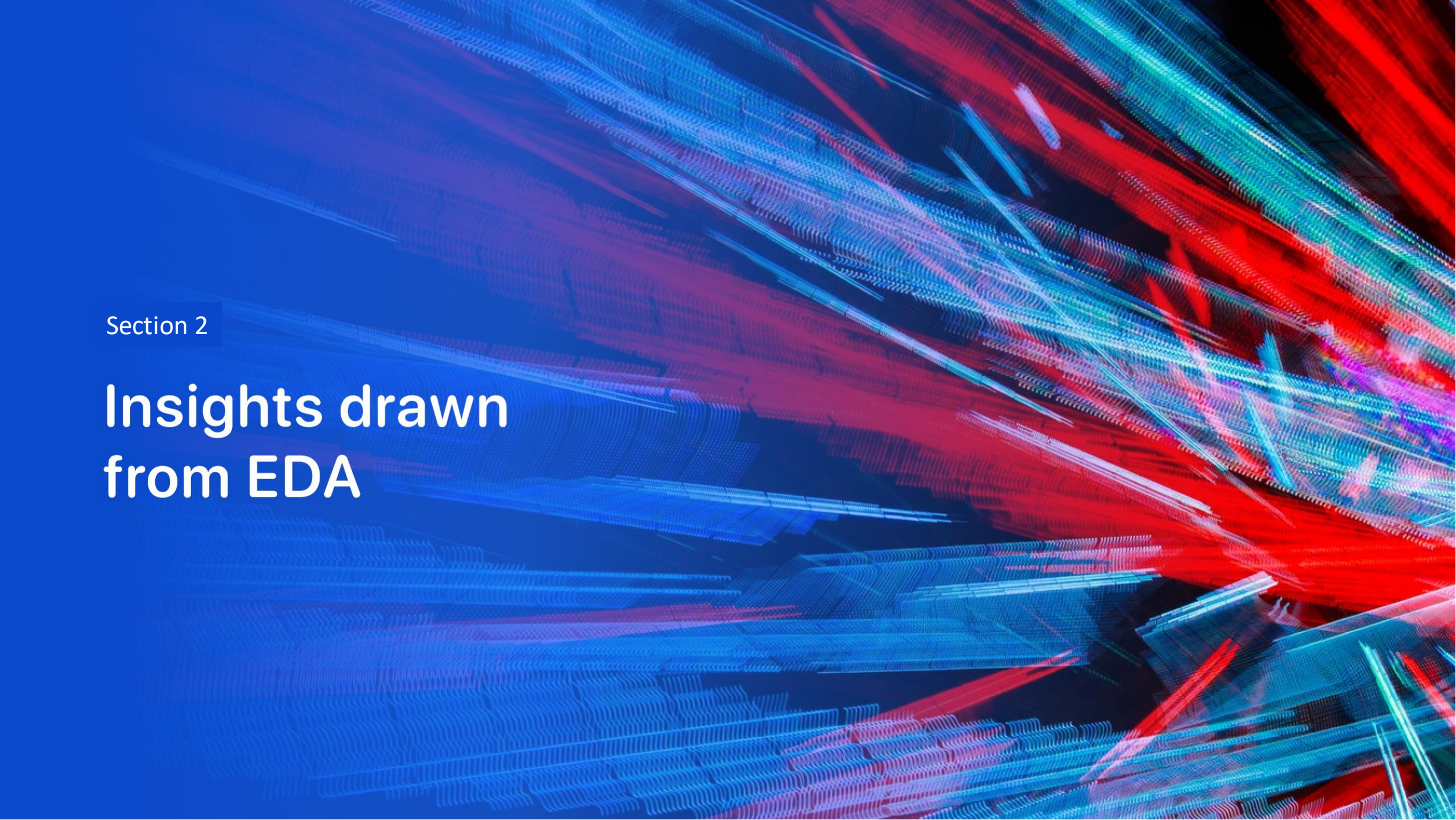
Predictive Analysis (Classification)



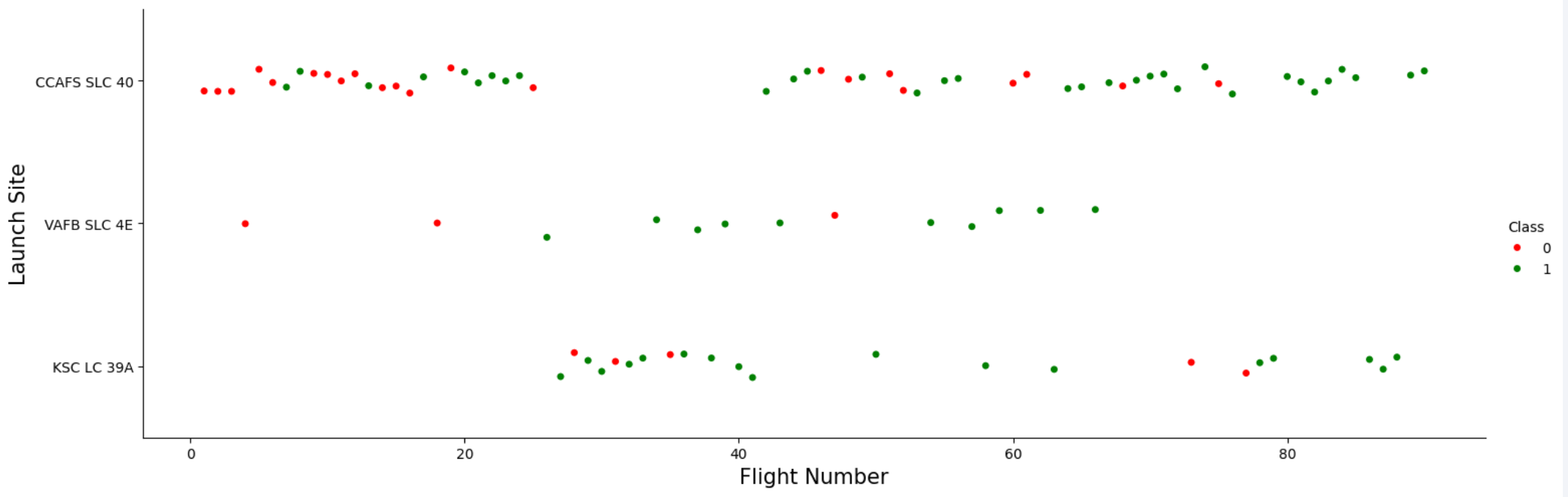
GitHub URL: [Predictive Analysis Notebook]

Results

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

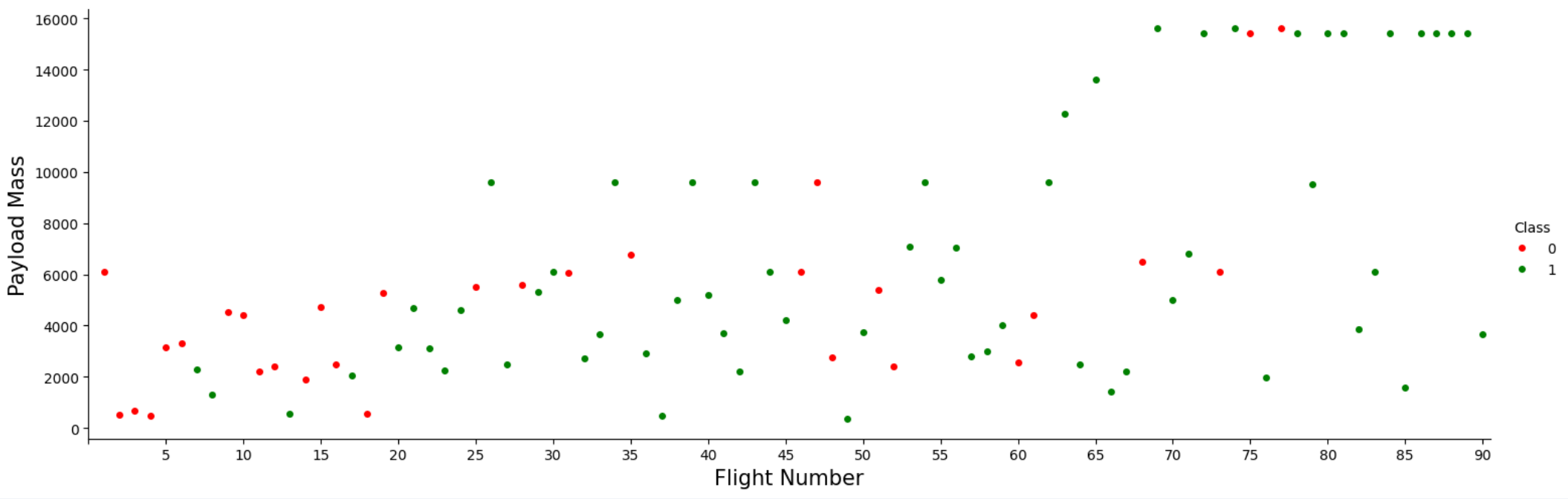

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



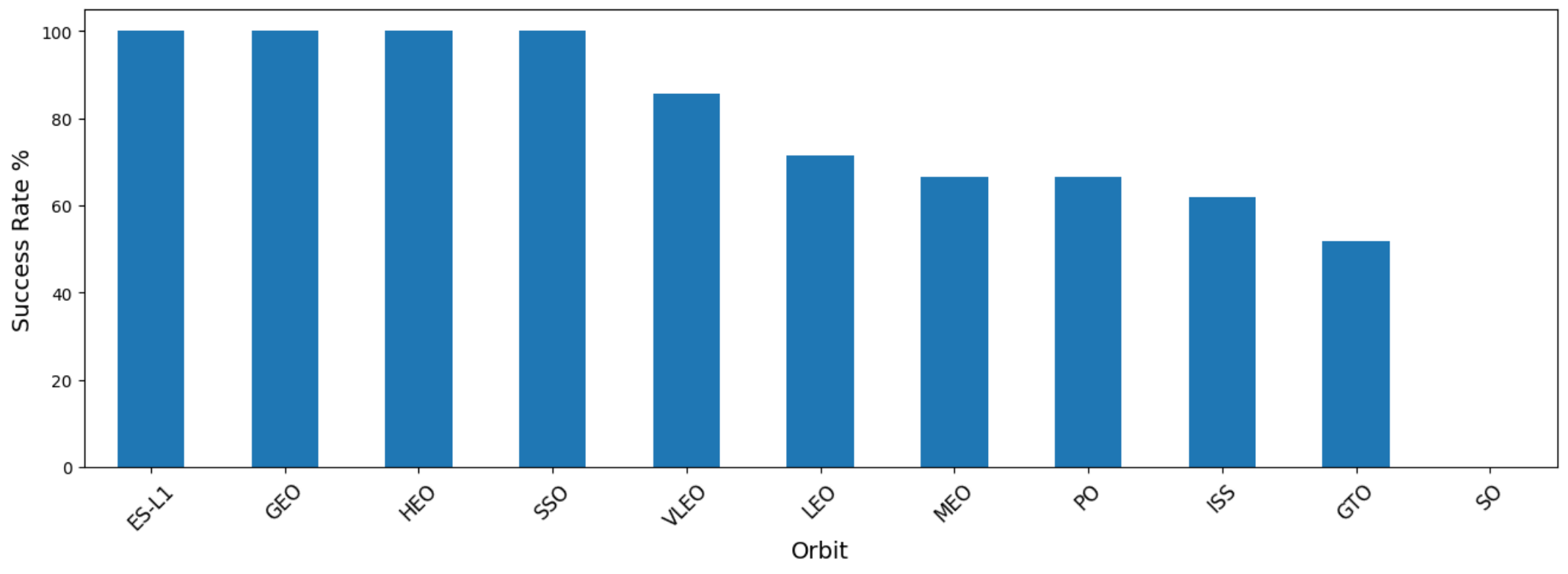
- In the early flights, there is a cluster of unsuccessful launches primarily from CCAFS SLC 40
- Over time, as Flight Number increases, the success rate improves
- CCAFS SLC 40 has a higher concentration of unsuccessful launches compared to other sites

Payload vs. Launch Site



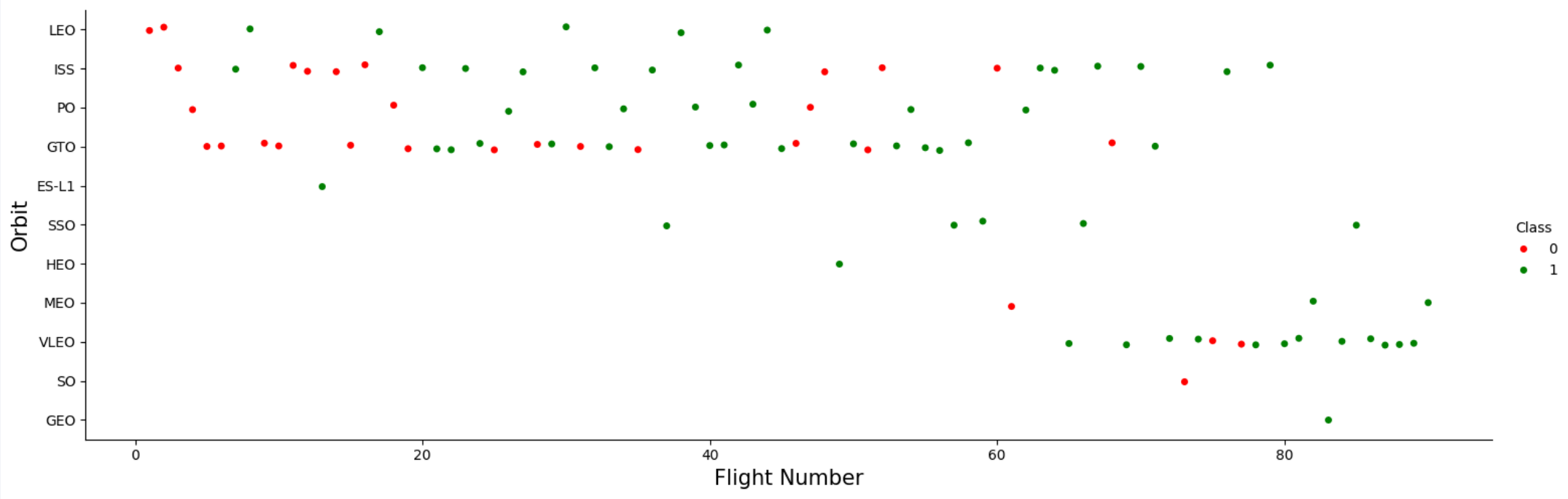
- Payload mass varies widely, ranging from a few hundred to over 15,000 kg.
- Unsuccessful launches are scattered across different payload masses.
- There is a concentration of unsuccessful launches with lower payload masses.

Success Rate vs. Orbit Type



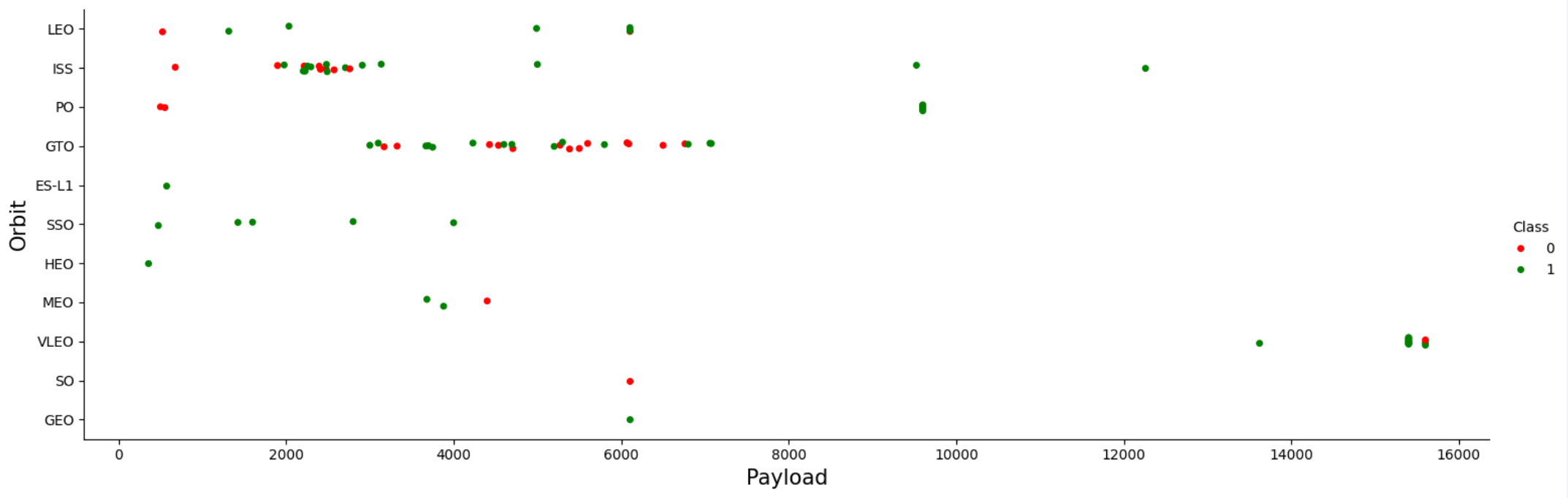
- ES-L1, GEO, HEO, and SSO orbits boast a perfect 100% success rate
- VLEO, LEO, MEO, PO, ISS, and GTO exhibit success rates ranging from 51.85% to 85.71%
- GTO experiences a lower success rate, while Suborbital missions show 0% success

Flight Number vs. Orbit Type



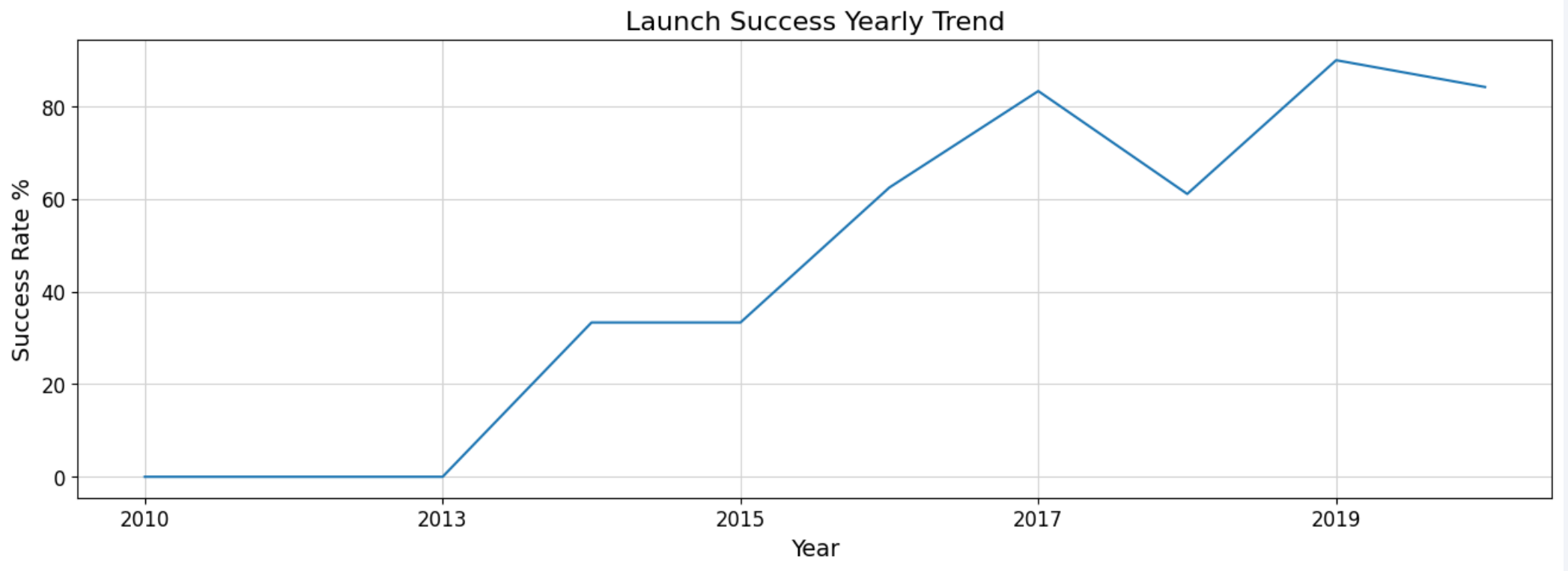
- Success in LEO orbit appears to correlate positively with the flight number
- No clear relationship is observed between flight number and success in GTO orbit
- Various orbits like ISS, PO, HEO, SSO, VLEO, MEO, GEO show mixed patterns, with success influenced by factors beyond the flight number

Payload vs. Orbit Type



- Polar (SSO), LEO, ISS: High payload masses correlate with success
- GTO Orbit: Payload mass alone doesn't distinctly predict success
- Overall: Payload impact varies by orbit

Launch Success Yearly Trend



- 2010-2014: Low and erratic success rates, with a slight increase in 2014
- 2015-2017: Significant improvement, with a steady increase each year
- 2018-2020: Success rates consistently high, peaking in 2019

All Launch Site Names

```
SELECT DISTINCT Launch_Site FROM  SPACEXTABLE
```

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

- The query retrieves distinct launch site names from the SPACEXTABLE, identifying the unique locations where SpaceX launches have taken place.
- The result lists four unique launch sites, namely CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

```
SELECT * FROM  SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 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	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

- The SQL query selects all columns from the table where the *Launch_Site* column begins with the characters 'CCA'. The LIMIT 5 clause ensures that only the first 5 records meeting this condition are returned.
- In the provided result, it's possible to see details of launches from launch sites starting with 'CCA', which, in this case, corresponds to the 'CCAFS LC-40' launch site.

Total Payload Mass

```
SELECT sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'
```

sum(PAYLOAD_MASS__KG_)
45596

- The SQL query is used to calculate the total payload mass (in kilograms) of all SpaceX launches where the customer is specified as 'NASA (CRS)'. The function aggregates the payload masses for all records that meet the condition in the where clause.
- The result of the query is 45596, indicating the total payload mass (in kilograms) for all SpaceX launches associated with NASA's Commercial Resupply Services (CRS) program.

Average Payload Mass by F9 v1.1

```
SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- The SQL query is used to calculate the average payload mass (in kilograms) carried by booster versions that match the pattern 'F9 v1.1%' in the Booster_Version column. The *avg* function computes the average payload mass for all records that meet the condition specified in the where clause.
- The result of the query is approximately 2534.67, indicating the average payload mass (in kilograms) for all SpaceX launches where the booster version is categorized as 'F9 v1.1%'.

First Successful Ground Landing Date

```
SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

min(Date)
2015-12-22

- The SQL query is used to find the earliest date when the first successful landing outcome on a ground pad was achieved. The *min* function is applied to the Date column, selecting the minimum (earliest) date from the records that meet the condition specified in the WHERE clause.
- The result of the query is '2015-12-22', indicating that the first successful landing outcome on a ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND
(PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

- The SQL query is designed to retrieve the names of boosters that achieved success in a drone ship landing and had a payload mass greater than 4000 kg but less than 6000 kg.
- The result of the query is a list of booster versions that successfully landed on a drone ship and carried a payload within the specified mass range.

Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, count() FROM SPACEXTABLE GROUP BY trim(Mission_Outcome)
```

Mission_Outcome	count()
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The SQL query is designed to provide the total number of successful and failure mission outcomes.
- The result of the query is a summary of the count of each unique mission outcome. This information provides an overview of the distribution of mission outcomes in terms of success and failure

Boosters Carried Maximum Payload

```
SELECT distinct Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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 SQL query is designed to retrieve the names of booster versions that have carried the maximum payload mass.
- The result of the query is a list of distinct booster versions that have carried the maximum payload mass. The list includes the booster versions such as *F9 B5 B1048.4*, *F9 B5 B1049.4* and others, indicating these boosters achieved the maximum payload mass in the dataset

2015 Launch Records

```
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)'
```

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

- The SQL query is designed to retrieve records that display month names, failure landing outcomes in drone ship, booster versions, and launch sites for the months in the year 2015.
- The result shows entries for January (Month 01) and April (Month 04) in the year 2015 where the landing outcome was a failure on the drone ship.

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

```
SELECT Landing_Outcome,
       count()
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
       AND (Landing_Outcome LIKE '%Success%' OR Landing_Outcome LIKE '%Failure%')
GROUP BY Landing_Outcome
ORDER BY count() desc
```

Landing_Outcome	count()
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

- This SQL query is designed to rank the count of landing outcomes, such as "Success (drone ship)" or "Failure (ground pad)," between the date '2010-06-04' and '2017-03-20' in descending order.
- The result of the query is a table that ranks the count of different landing outcomes within the specified date range. The landing outcomes are grouped by their types, and the counts are presented in descending order.

Section 3

Launch Sites Proximities Analysis

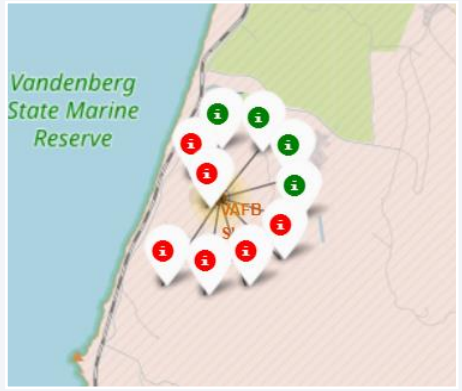
Global Launch Site Distribution

All launch sites are located in the United States:

- Florida Launch Sites:
  - Cape Canaveral Air Force Station Launch Complex 40 (CCAFS LC-40)
  - Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS SLC-40)
  - Kennedy Space Center Launch Complex 39A (KSC LC-39A)
- California Launch Site:
  - Vandenberg Air Force Base Space Launch Complex 4E (VAFB SLC-4E)

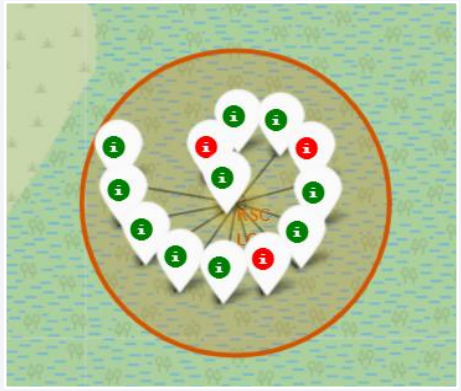
Success/Failed Launches for each Site

VAFB SLC-4E



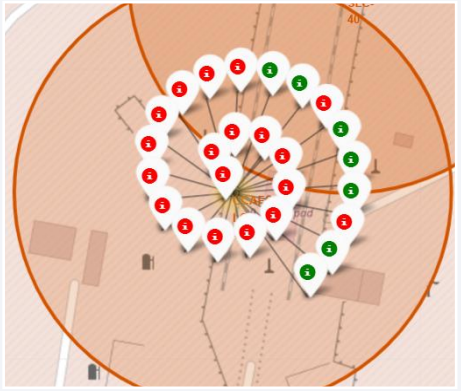
- Successful Launches: 4
- Failed Launches: 6
- Success Rate: 40 %

KSC LC-39A



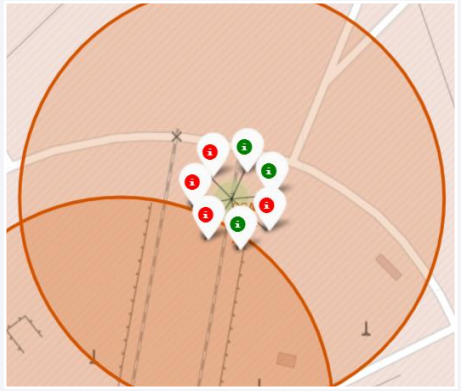
- Successful Launches: 10
- Failed Launches: 3
- Success Rate: 76.9 %

CCAFS LC-40



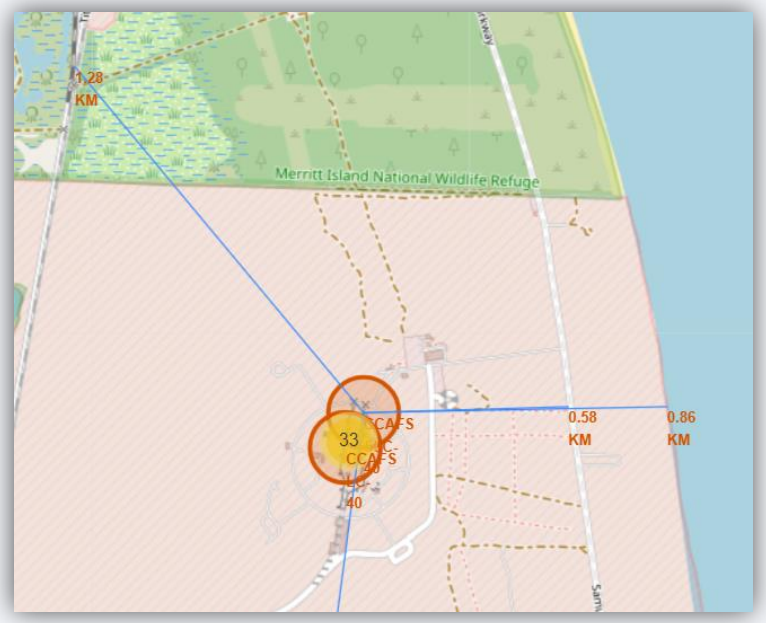
- Successful Launches: 7
- Failed Launches: 19
- Success Rate: 26.9 %

CCAFS SLC-40



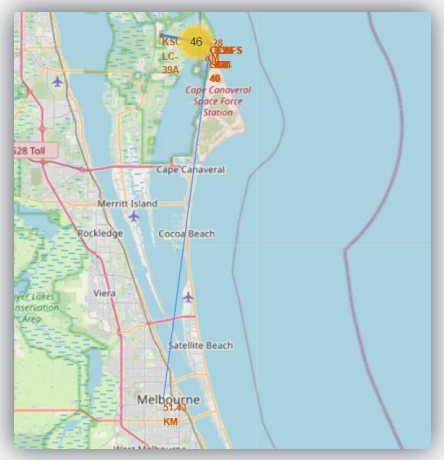
- Successful Launches: 3
- Failed Launches: 4
- Success Rate: 42.9 %

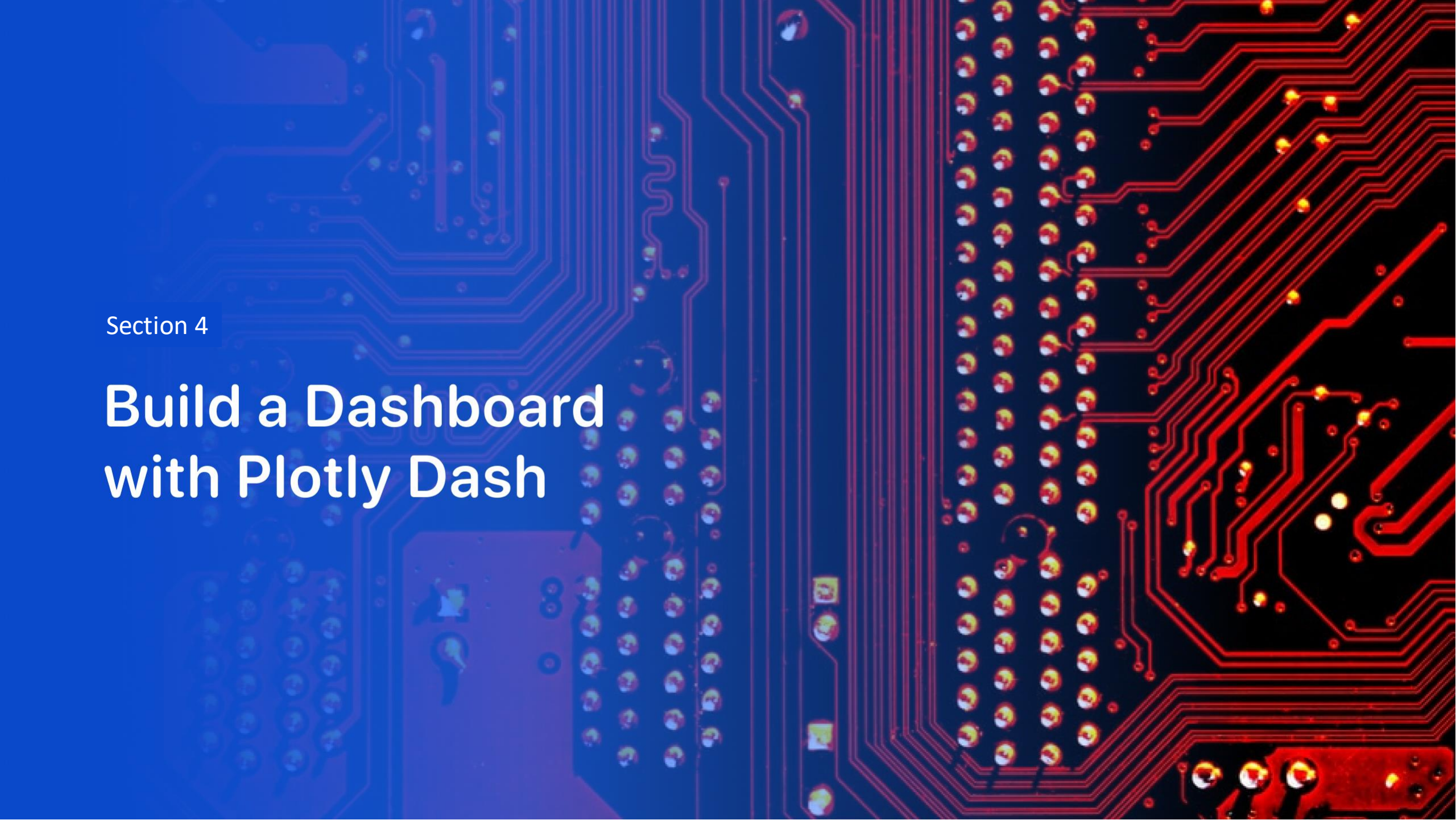
Proximity Analysis for Launch Site CCAFS SLC-40



- **Coastline Safety**: Proximity to coastlines provides safety benefits, allowing for water landings in case of launch aborts and minimizing risks to populated areas from falling debris. (Distance from coastline: 0.86 km)
- **Highway Access**: Launch sites are strategically located near highways for efficient transportation of personnel, equipment, and payloads, supporting the logistical needs of space missions. (Distance from highway: 0.58 km)
- **Railway Efficiency**: Close proximity to railways facilitates the transport of heavy cargo, offering a cost-effective and efficient means of moving large payloads. (Distance from railway: 1.28 km)

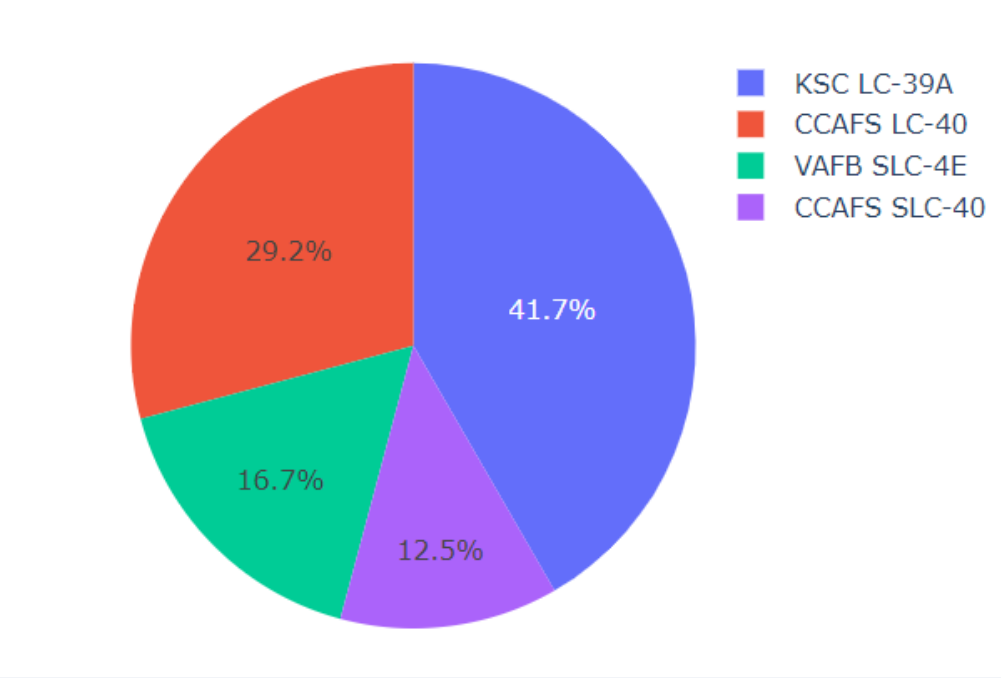
- **Urban Distance**: Launch sites are intentionally situated away from densely populated areas, minimizing risks and ensuring the safety of human life and property in case of launch failures. (Distance from city - Melbourne: 51.43 km)




Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site

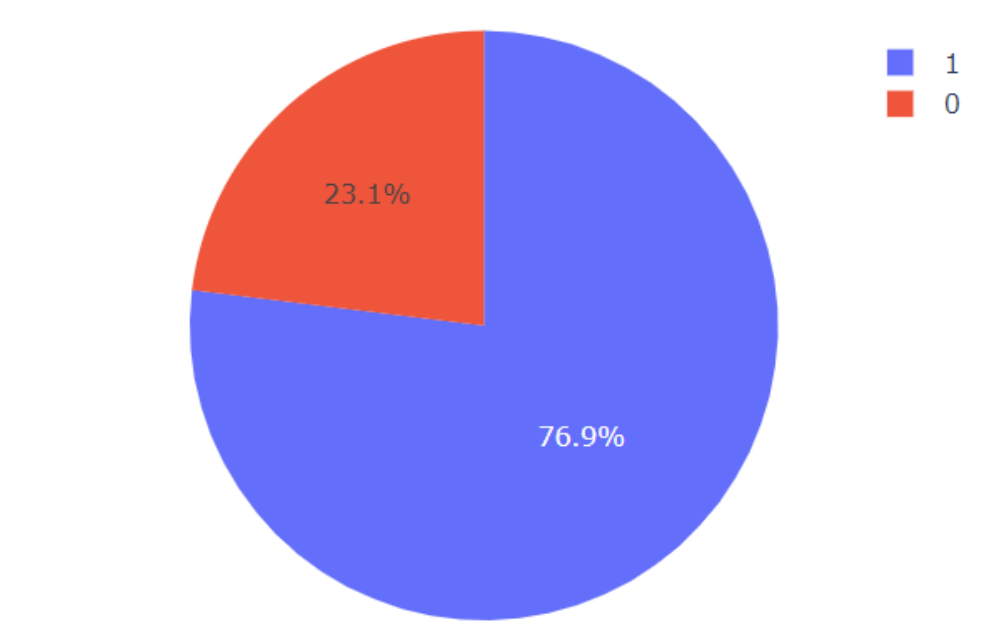
The pie chart illustrates the distribution of successful launches among four launch sites:



- **KSC LC-39A** (41.7%): Dominates with the highest success rate.
- **CCAFS LC-40** (29.2%): Contributes significantly to successful launches.
- **VAFB SLC-4E** (16.7%): Plays a notable role in successful missions.
- **CCAFS SLC-40** (12.5%): Contributes to the overall success distribution.

Total Success Launches for Site KSC LC-39A

The pie chart highlights the launch success ratio for the site with the highest success percentage:



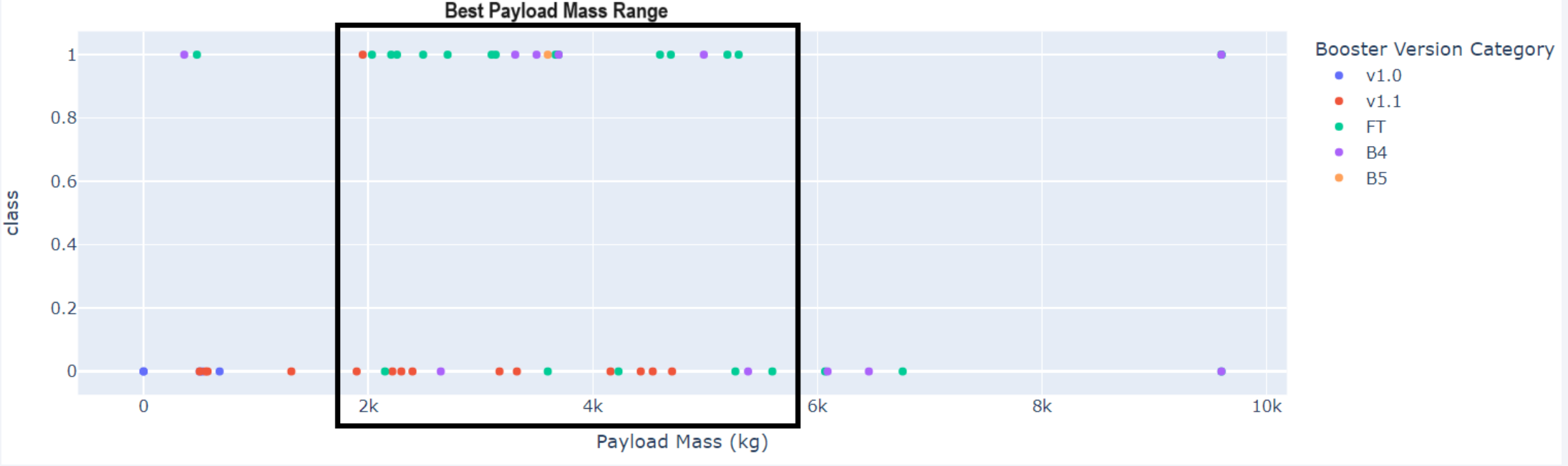
KSC LC-39A

- **Success Rate** (76.9%): Represents the proportion of successful launches
- **Failure Rate** (23.1%): Indicates the percentage of unsuccessful launches from the same site.

This chart emphasizes KSC LC-39A's impressive success rate, demonstrating its effectiveness as a launch site with a majority of missions ending successfully.

Success Rates: Payload Mass and Booster Versions

The scatter plot displays the relationship between Payload Mass, Launch Outcome (Success or Failure), and the associated Booster Version

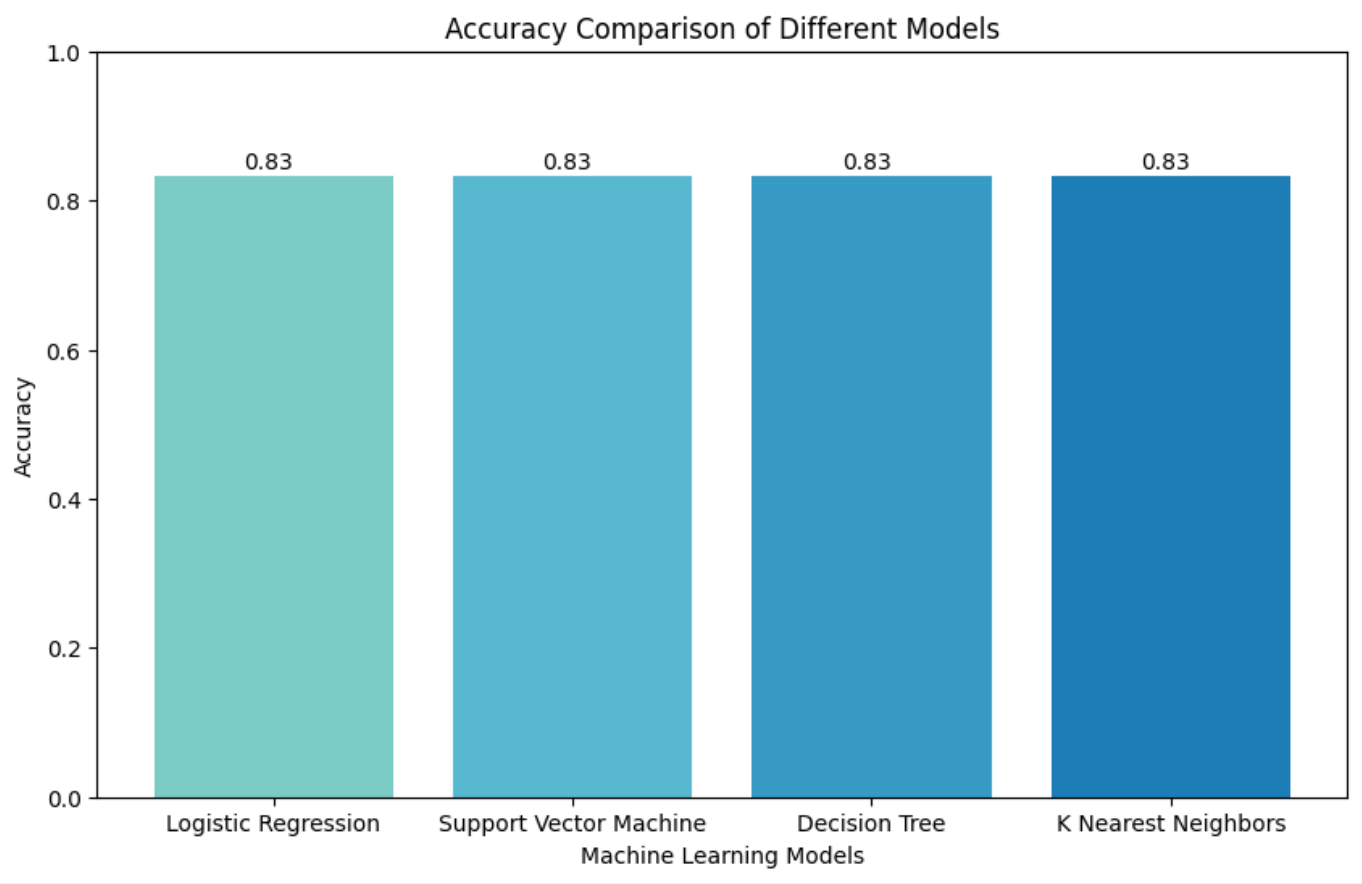


- **Payload Range**: The scatter plot shows that the most successful launches cluster within the payload mass range of approximately 1900 to 5500 kg.
- The **Booster Version Category** "FT" stands out as having the highest success rate

Section 5

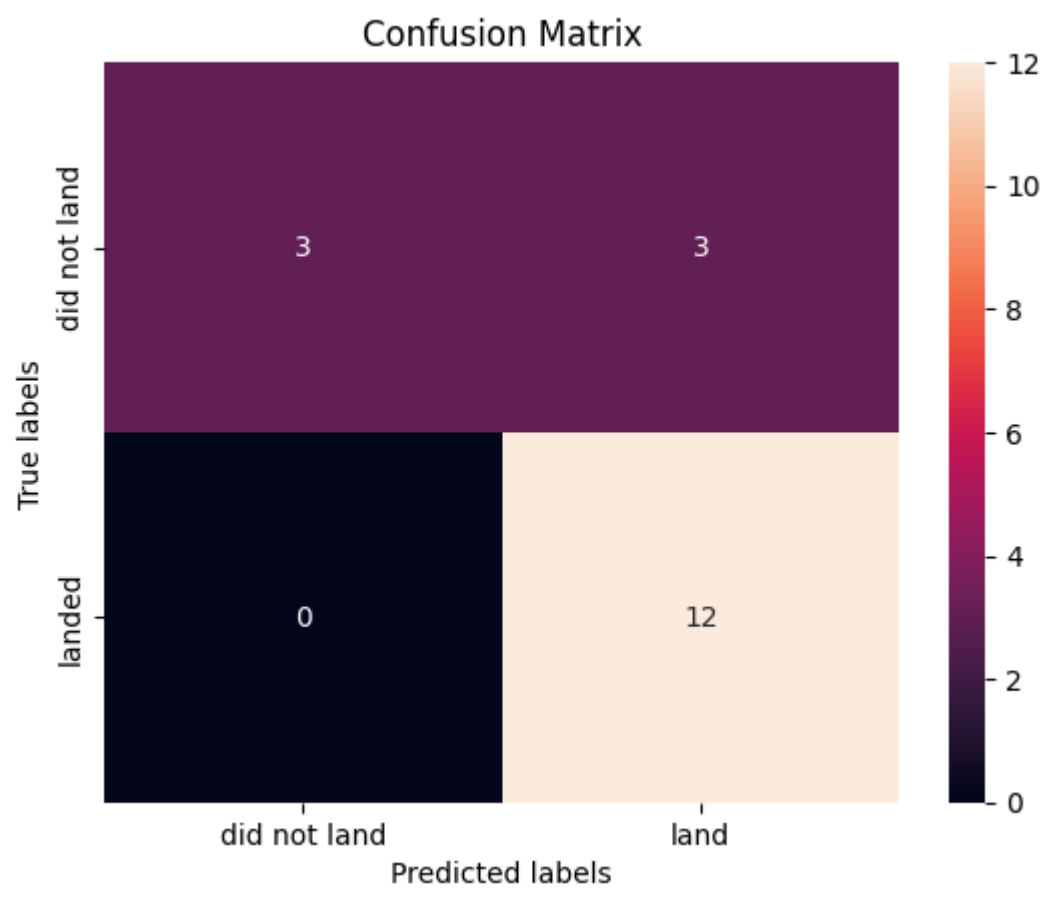
Predictive Analysis (Classification)

Classification Accuracy



The **accuracy** values for all models are the same: **83.33%**

Confusion Matrix



- The models, including Logistics Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors, have collectively correctly identified 3 instances as negative and 12 instances as positive.
- They have made 3 false positive errors, incorrectly predicting 3 instances as positive when they were actually negative.
- They have made 0 false negative errors, meaning they correctly predicted all positive instances.

Conclusions

- **Optimal Payload and Booster Version:**
  - Payloads between 1900 - 5500 kg and booster version "FT" show the highest success rates.
- **Temporal Success Trends:**
  - Success rates steadily increased since 2013, with notable peaks in 2016 and 2019.
- **Strategic Launch Site Positioning:**
  - Launch sites strategically located for efficiency near the equator, coastline, highways, and railways.
- **Geographical Launch Distribution:**
  - Launch sites concentrated in America, specifically in Florida and California.
  - CCAFS LC-40 and KSC LC-39A have more failed launches.
- **Unified Model Accuracy:**
  - All classification models achieved a consistent high accuracy of 83.33%.
  - The models collectively performed well, with minimal false positives and no false negatives.

Conclusions

Optimal Payload and Booster Version

- Payloads between 1900 - 5500 kg and booster version "FT" show the highest success rates

Temporal Success Trends

- Success rates steadily increased since 2013, with notable peaks in 2016 and 2019

Strategic Launch Site Positioning

- Launch sites strategically located for efficiency near the equator, coastline, highways, and railways

Geographical Launch Distribution

- Launch sites concentrated in America, specifically in Florida and California
- CCAFS LC-40 and KSC LC-39A have more failed launches

Unified Model Accuracy

- All classification models achieved a consistent high accuracy of 83.33%
- The models collectively performed well, with minimal false positives and no false negatives

Thank you!