IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rafael Jeronymo
August 14, 2025

Outline

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

Executive Summary

- This project explores SpaceX Falcon 9 launch data
- The goal is to understand what factors affect successful landings and to build simple models to predict outcome
- The methodology I used was the collection of data, cleaning it, analyzing it with SQL and visualizations, creating dashboards and testing machine learning

Introduction

- The case study of SpaceX launches is relevant because it brings some questions that are important for both the company and organizations interested in this data
- The main question would be: Can we predict if a Falcon 9 first stage will land successfully?
- This is important because successful landings lower costs and make space more accessible.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Data was cleaned for missing values and formatted dates
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis through Logistic Regression, Decision Trees and KNN

Data Collection

- Data was collected using the SpaceX API to pull launch records and web scraping from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”
- Data collection process using API:
  - Request rocket launch data from SpaceX API
  - Parse the data using the GET request
  - Decode the response content as Json and turn it into a Pandas dataframe
- Data collection process using Web Scraping:
  - Request the wiki page from its URL as an HTTP response
  - Create a BeautifulSoup object from the HTML response
  - Extract all column/variables from the HTML
  - Create a data frame by parsing the launch HTML tables

Data Wrangling

- Data was processed for missing values and replaced with the mean
- Data wrangling process
  - Calculate the mean for the PayloadMass
  - Use the mean to replace the missing values with the mean calculated
  - Create a landing outcome label from Outcome column
  - Create a list where zero is bad outcome and 1 is otherwise

EDA with Data Visualization

- Below are the charts used and why they were used
  - **Flight Number vs Launch Site**: showed launch frequency and success by site
  - **PayloadMass vs Launch Site:** explored effect of payload size on outcomes
  - **Success Rate by Orbit Type:** bar chart revealed which orbits had higher success
  - **Flight Number / Payload vs Orbit:** scatter plots to see trends across missions
  - **Launch Success Rate by Year:** line chart showing improvement over time

EDA with SQL

- Display the names of the unique launch sites
- Display 5 records where launch site begins with 'CCA'
- Display the total payload mass carried by boosters launched by NASA
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster versions that have carried the maximum payload mass
- Display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Use circles for the success/failed launches for each site on the map
- Draw lines to show the distances between a launch site to its proximities

- These objects helped visualize launch locations, outcomes, and geographical, factors that may affect mission success.

Build a Dashboard with Plotly Dash

- Pie Chart
  - Showed launch success rates for all sites
- Pie Chart per Site
  - Highlighted site with highest success ratio
- Scatter Plot
  - Explored how payload size affects success
- Dropdown Filter
  - Let users select specific launch sites

Why: these interactions made the dashboard dynamic, helping explore success patterns by site and payload

Predictive Analysis (Classification)

- Models used: Logistic Regression, SVM, Decision Tree, KNN
- Process (Flowchart)
  - Raw Data → Feature Selection → Train/Test Split → Model Training → Model Evaluation → Best Model
- Best model: Decision Tree with training accuracy of 88%

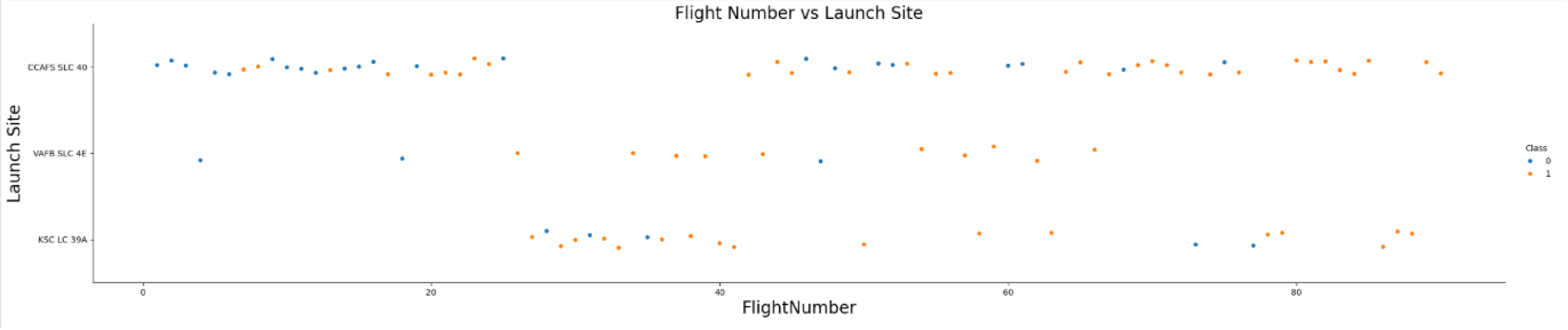
Results

- Exploratory data analysis results
  - Success rates varied by orbit type and launch site
  - Payload range (2000-6000kg) showed higher success probability
  - Success rates improved steadily over the years
- Interactive analytics
  - Folium maps revealed site proximity to coasts and infrastructure
  - Dash allowed filtering by payload and site for deeper insights
- Predictive analysis results
  - Logistic Regression, KNN, Decision Tree and SVM
  - Best performing model: Decision Tree

Section 2

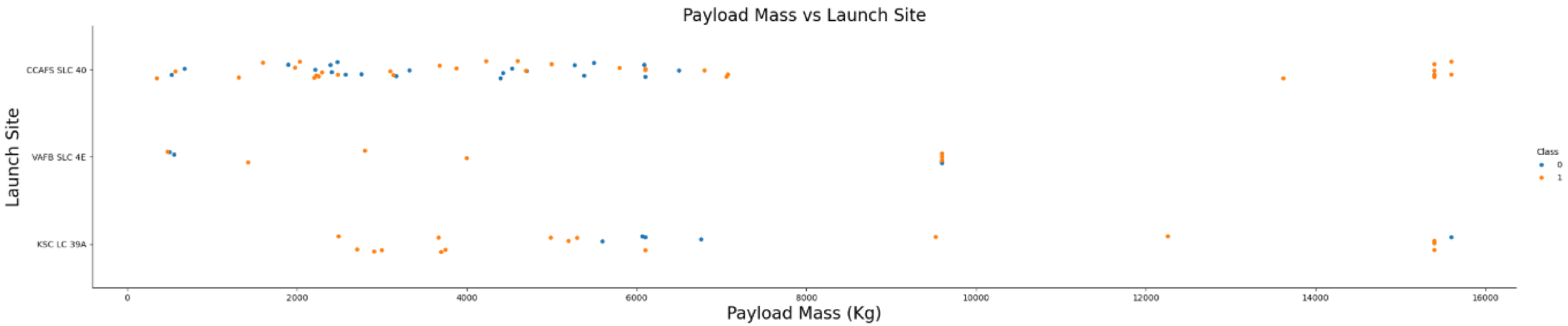
Insights drawn from EDA

Flight Number vs. Launch Site



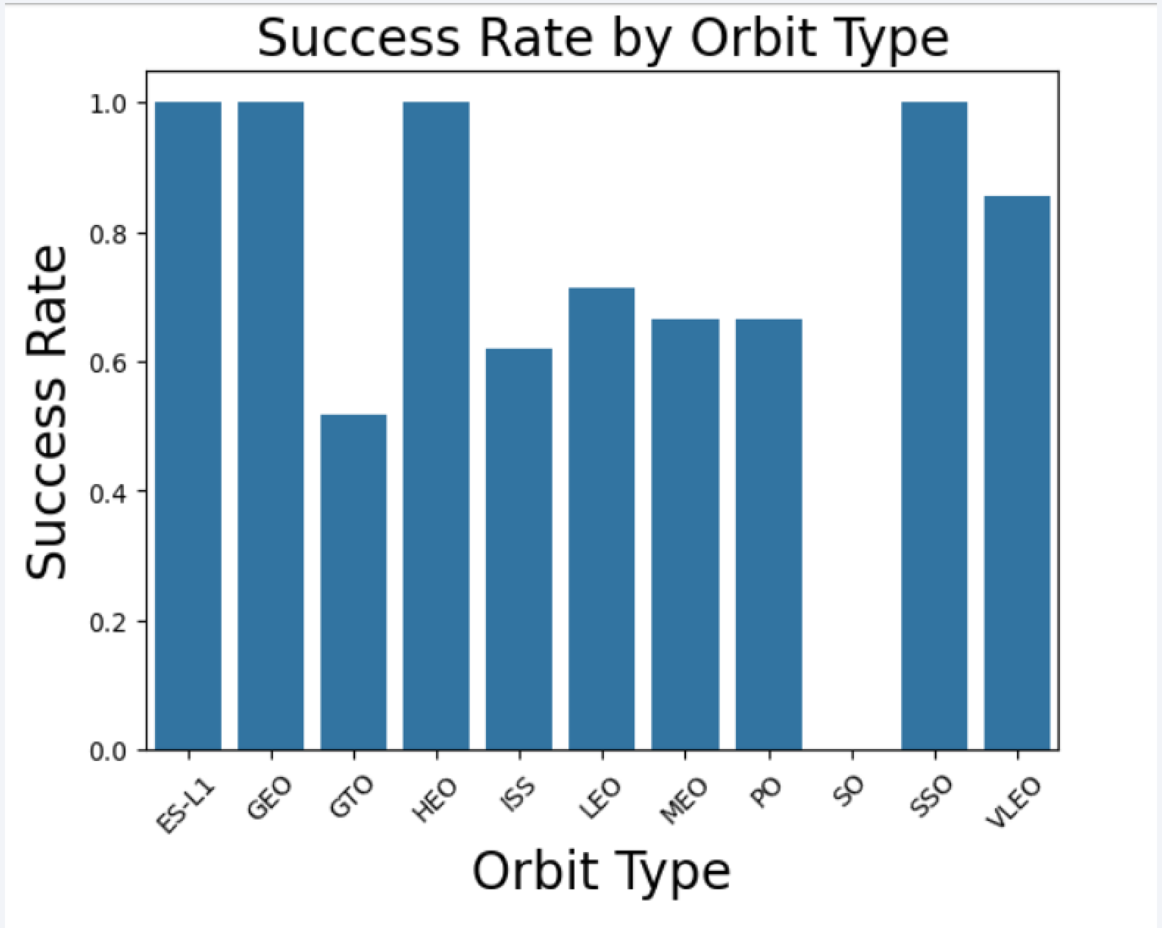
- Scatter plot shows launches grouped by site (CCAFS, VAFB, KSC)
- Early flights had more failures
- Over time, success rate increased at each site
- KSC and CCAFS sites had more frequent launches compared to VAFB

Payload vs. Launch Site



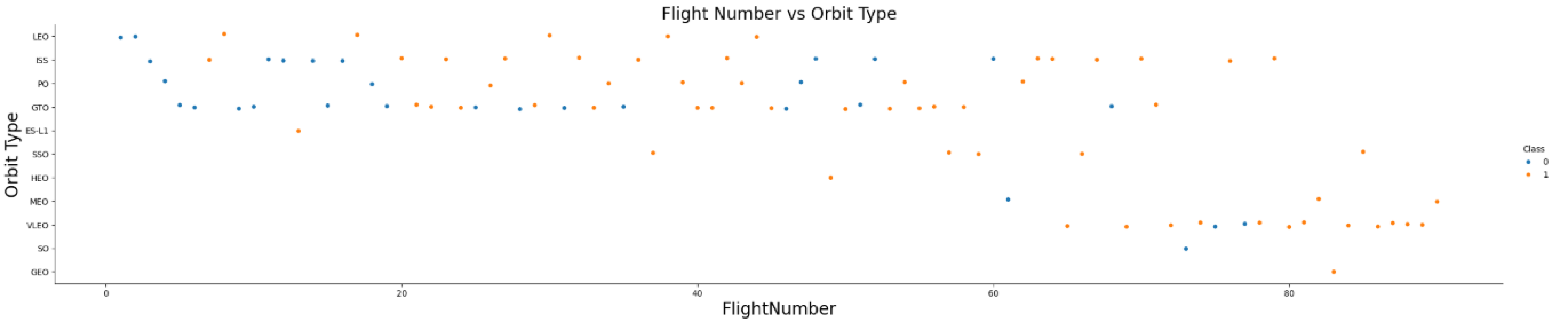
- Scatter plot compares payload size with launch site outcomes
- Most payloads between 2000-6000kg had higher success rates
- Very heavy payloads (> 10,000kg) showed some success but less frequent
- CCAFS and KSC handled the widest range of payloads, while VAFB had fewer launches

Success Rate vs. Orbit Type



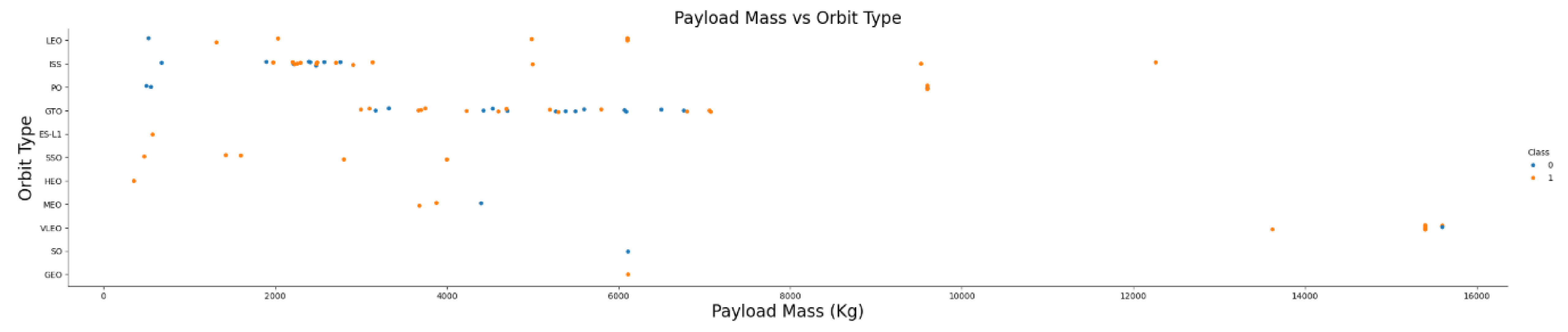
- Some orbits like ES-L1, GEO, HEO and SSO had almost 100% success rates
- Others like GTO and ISS showed lower reliability
- It shows how orbit type influences landing success probability

Flight Number vs. Orbit Type



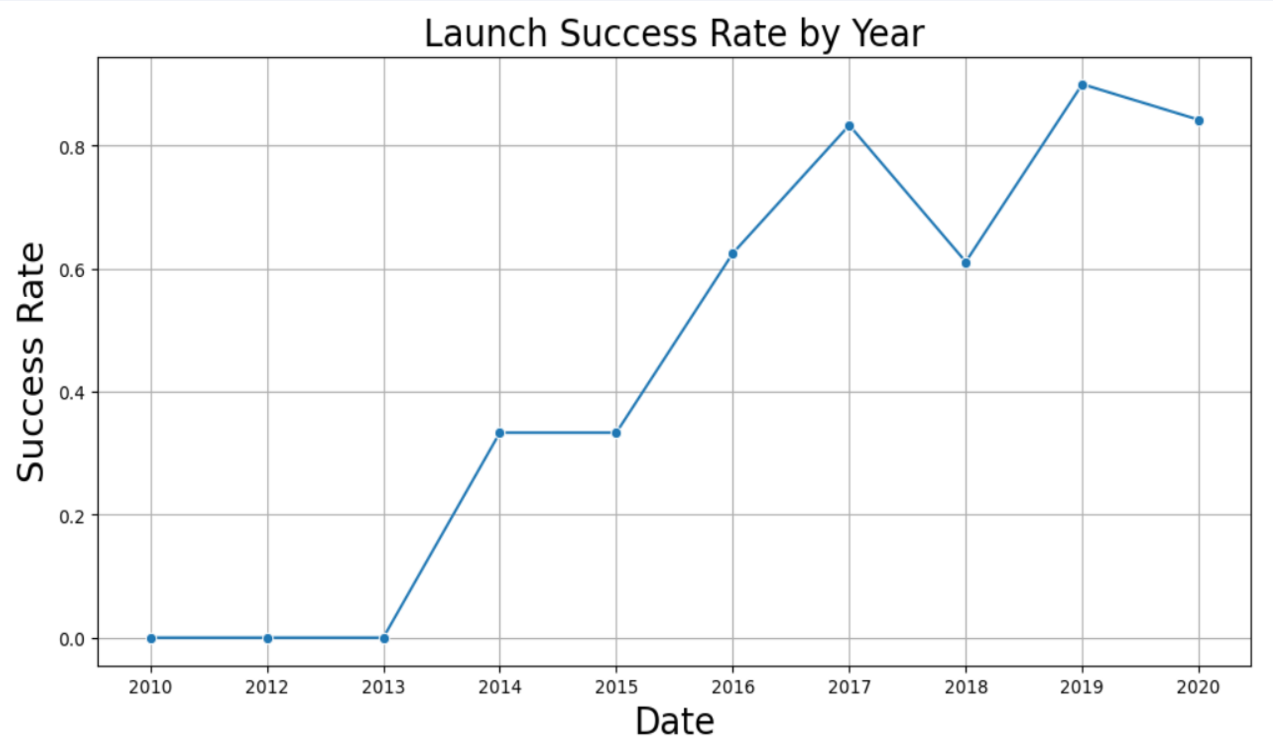
- Scatter plot shows missions grouped by orbit type across flight history
- Early missions had more failures across all orbits
- Over time, success rates improved across most orbit types
- GTO and ISS missions stood out with mixed results, while LEO and SSO showed more consistent successes later

Payload vs. Orbit Type



- Scatter plot shows how payload mass relates to orbit type and landing outcome
- Medium payloads (2000-6000 kg) had higher success rates
- LEO and SSO orbits supported a wide range of payloads and showed more consistent successes

Launch Success Yearly Trend



- Early launches (2010-2013) had no successful landings
- Success rates improved significantly starting in 2014
- By 2017-2019, rates reached above 80-90%, showing strong reliability improvements
- This trend shows how fast SpaceX progressed in landing technology

All Launch Site Names

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

- This is the result of the query showing the four distinct SpaceX launch sites

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

- This is the result of the query showing 5 records where launch sites begin with ‘CCA’

Total Payload Mass

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

```
%%sql
SELECT SUM("Payload_Mass_Kg_") AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)'
```

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

Total_Payload_Mass
0.0

- In this dataset, no payload mass values were recorded for NASA CRS launches.

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG("Payload_Mass_Kg_") AS Average_Payload
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1'
```

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

Average_Payload
0.0

- The dataset returned no payload mass for this booster version.

First Successful Ground Landing Date

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

Hint:Use min function

```
%%sql
SELECT MIN(Date) AS First_Successful_Ground_Landing
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

First_Successful_Ground_Landing
2015-12-22

- The first successful ground landing date was 2015-12-22.

Successful Drone Ship Landing with Payload 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_" > 4000
    AND "Payload_Mass_Kg_" < 6000
```

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

Booster_Version

- The dataset did not contain any launches for this criteria

Total Number of Successful and Failure Mission Outcomes

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

- The dataset shows SpaceX achieved the majority of missions successfully, with only 1 failure (in flight).

Boosters Carried Maximum Payload

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007
F9 v1.1 B1003
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1
F9 v1.1

- The list of boosters is quite large so the screenshot shows some of them.

2015 Launch Records

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

- In 2015, two boosters failed to land successfully on drone ships, and both were launched from the site CCAFS LC-40

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

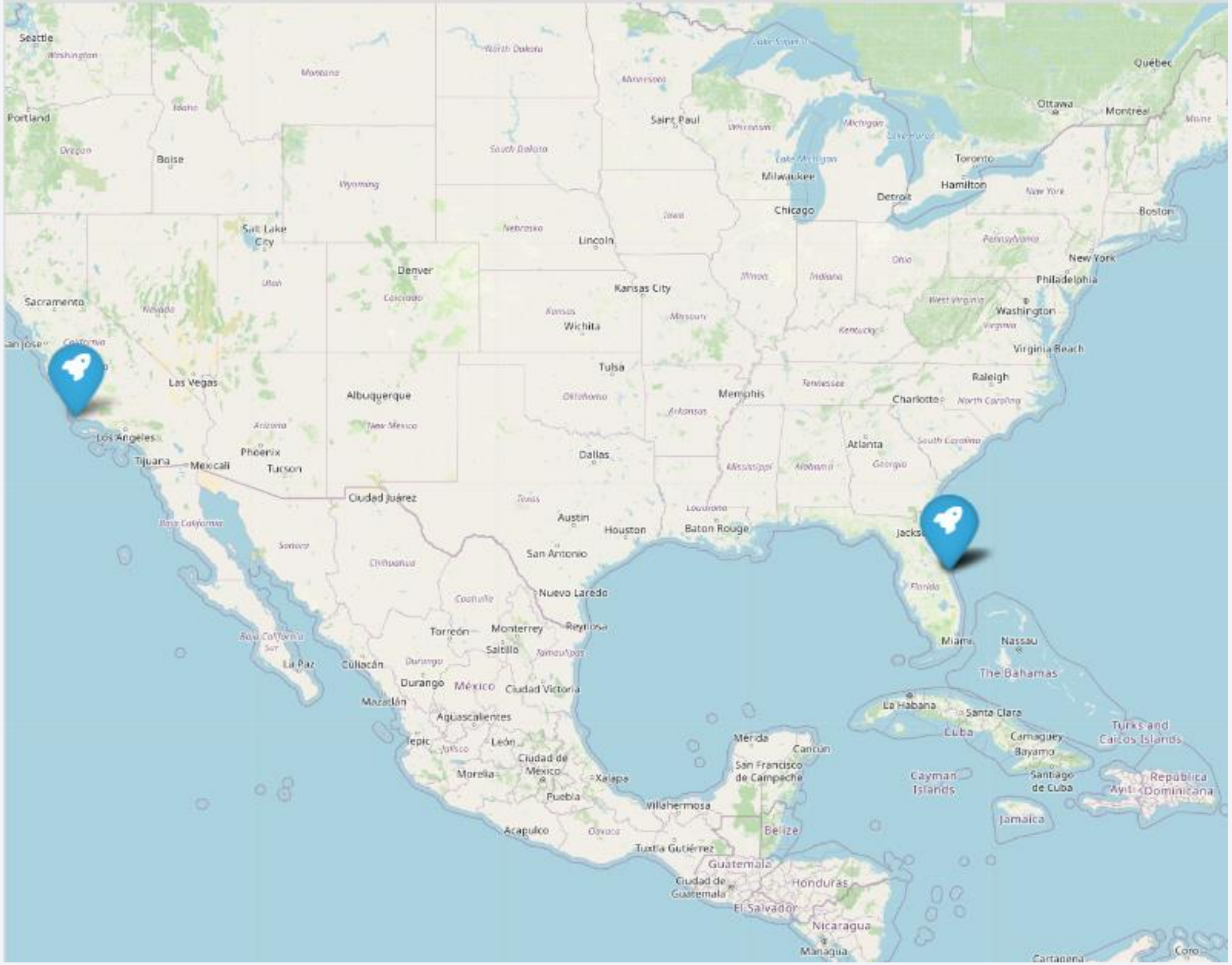
Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3

- Between 2010 and 2017, drone ships landings had more failures, while ground pad landings achieved a higher success ratio.

Section 3

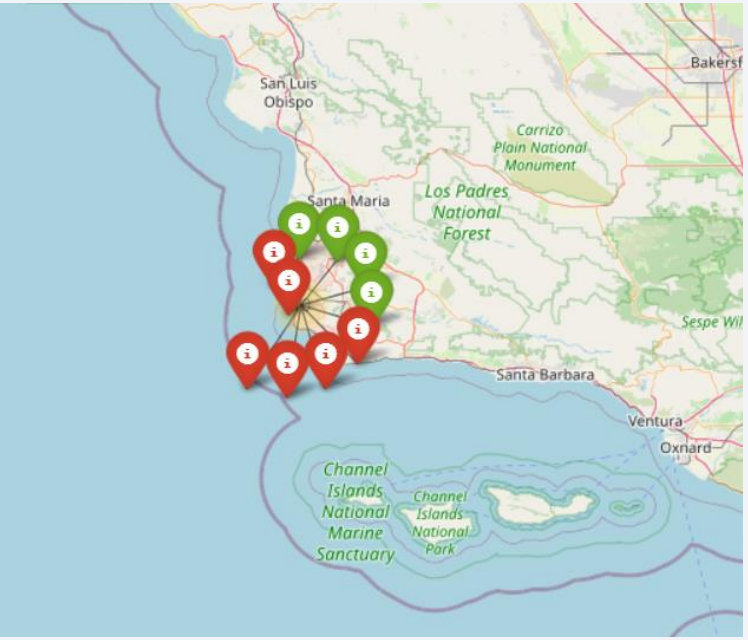
Launch Sites Proximities Analysis

All Launch Sites



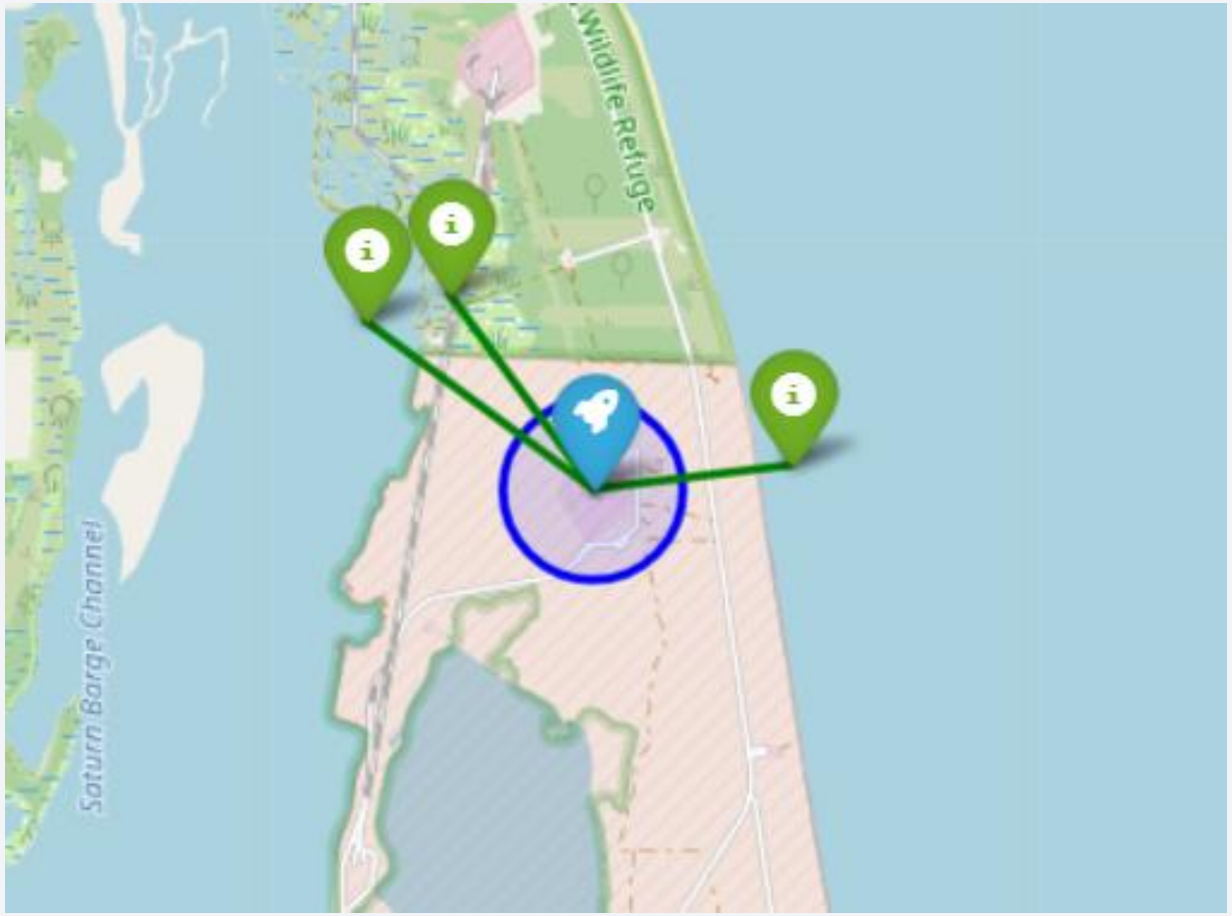
- The map helps visualize SpaceX's geographic strategy in choosing coastal launch sites.

Launch Outcomes



- It shows performance differences across sites
- Green for success and red for failure

Site Proximity Analysis

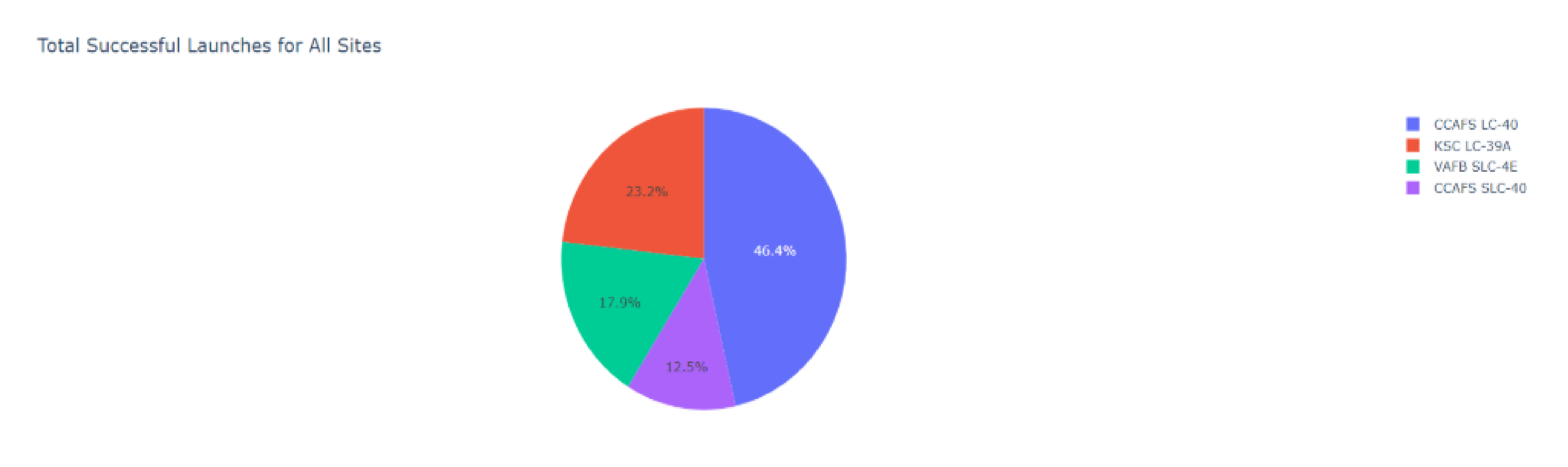


- Distance to nearest Railroad: 1.58Km
- Distance to nearest Coastline: 1.12Km
- Distance to nearest Highway: 1.36Km

Section 4

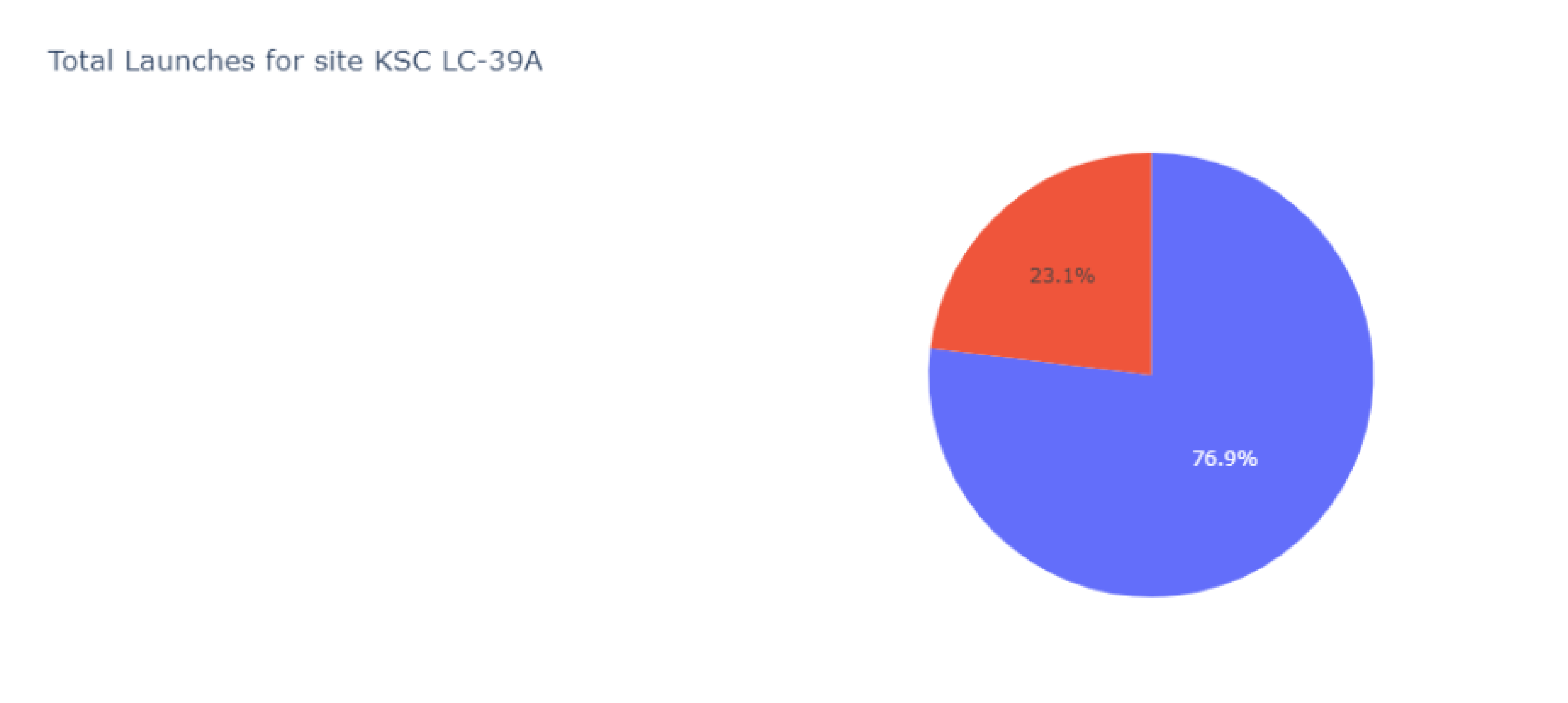
Build a Dashboard with Plotly Dash

Launch Success by Site



- The pie charts shows the overall reliability across all sites

Highest Performing Site



- The site KSC LC-39A had the highest rate of success

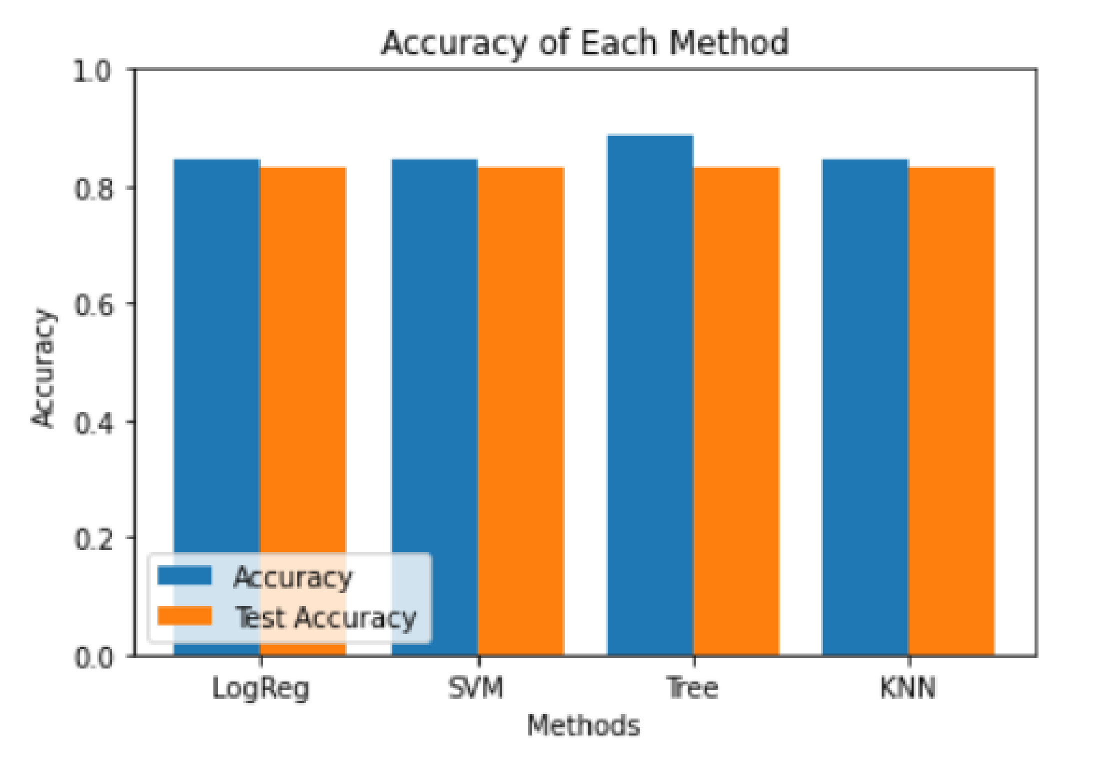
Payload vs Launch Outcome

- It shows how payload range and site relate to success; reveals ranges with higher success probability

Section 5

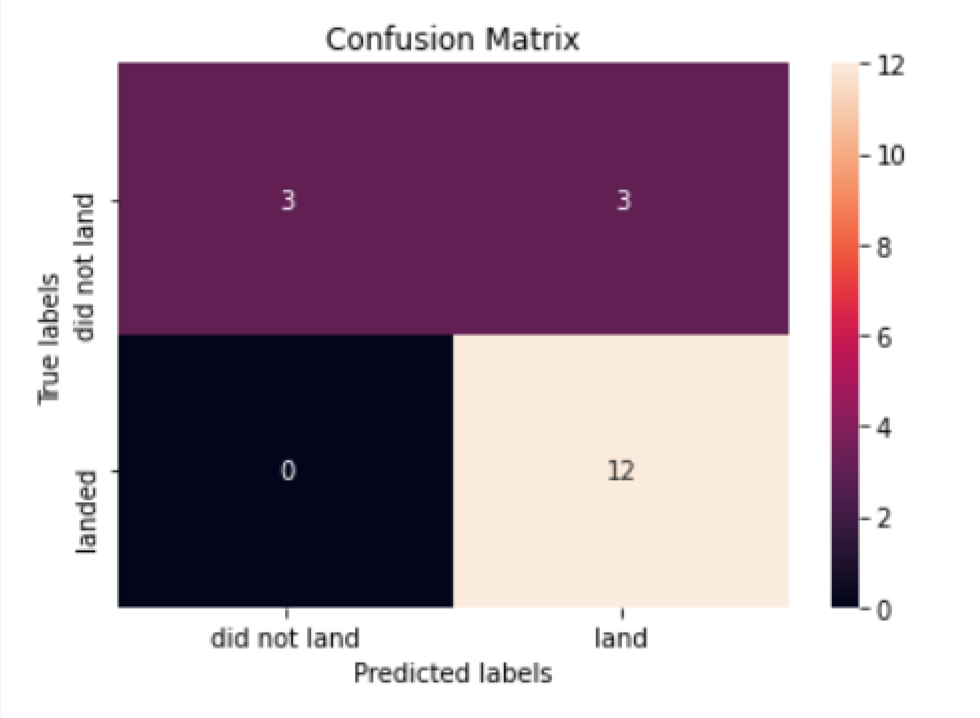
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree model gave the best prediction performance.

Confusion Matrix



- True positives: 12 correctly predicted landings
- True negatives: 3 correctly predicted failures
- False positives: 3 predicted landing but failed
- False negatives: 0 all landings were captured correctly

Conclusions

- SpaceX launch success rates have steadily improved over the years
- Success varies by orbit type and payload mass (best results between 2000-6000kg)
- Interactive tools such as Folium and Dash helped visualize launch location, outcomes, and patterns
- Machine learning models predicted landing success, with Decision Tree giving the best accuracy
- Overall, data analysis shows how SpaceX increased reliability and efficiency in booster landings

Thank you!