IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX has revolutionized the space launch industry by offering Falcon 9 launches at $62 million per mission, significantly undercutting competitors who charge upwards of $165 million. This cost advantage stems primarily from SpaceX's ability to recover and reuse the first stage of their rockets. Understanding whether a first stage will land successfully is therefore crucial for estimating launch costs, particularly valuable for companies considering competitive bids against SpaceX. Our project aims to develop a machine learning pipeline to predict first stage landing success.

- Problems you want to find answers
  - Identify and analyze the key factors that influence successful rocket landings
  - Understand the relationships and interactions between different variables that affect landing success rates
  - Determine the optimal operating conditions required to maximize the probability of successful landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection Process

- Key Steps:
  - API Data: Sent GET requests to SpaceX API, decoded JSON response, and converted to pandas dataframe using .json_normalize().
  - Data Cleaning: Handled missing values and inconsistencies.
  - Web Scraping: Scraped Falcon 9 launch records from Wikipedia using BeautifulSoup, parsed HTML tables, and converted to pandas dataframe.
  - Integration: Combined API and scraped datasets for analysis.

[GET Request → SpaceX API]

[Decode JSON → Pandas Dataframe]

[Clean Data]

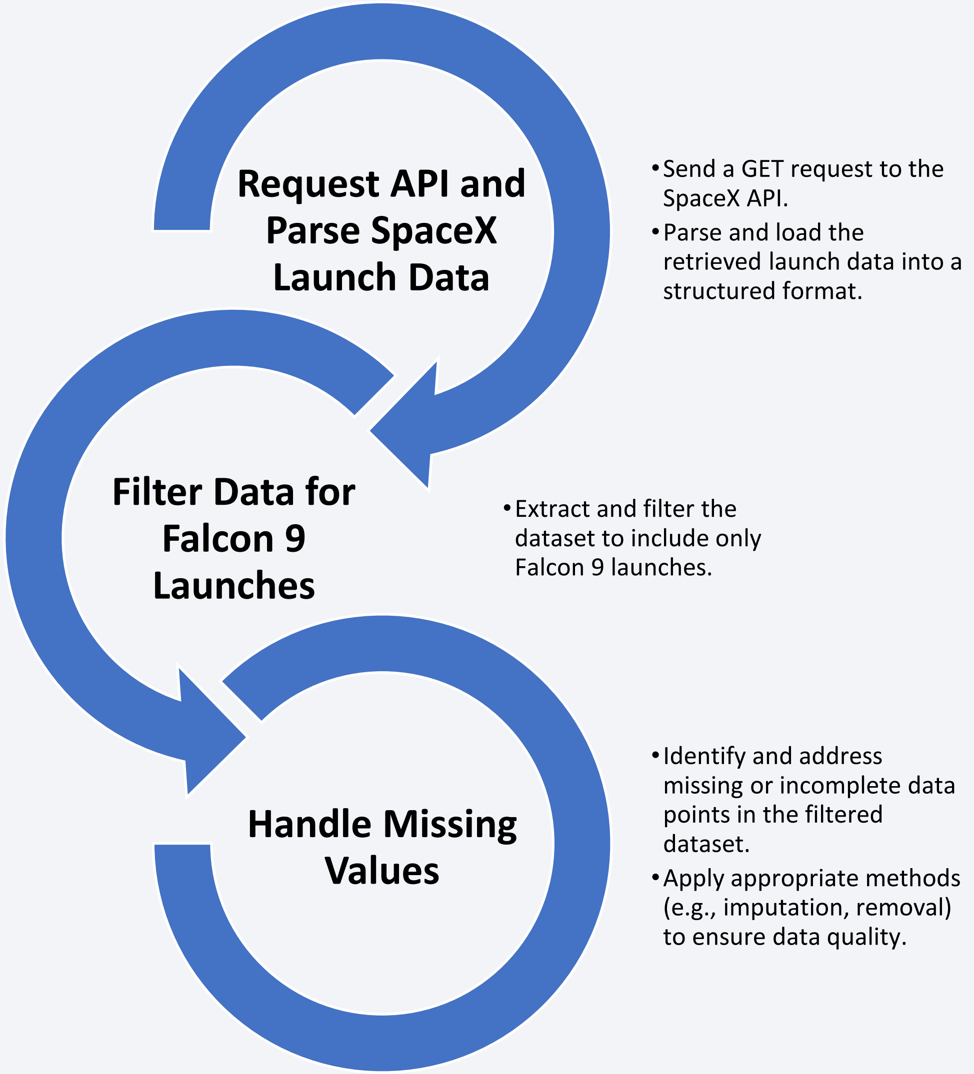
[Scrape Wikipedia → Parse HTML Table → Pandas Dataframe]

[Combine Datasets]

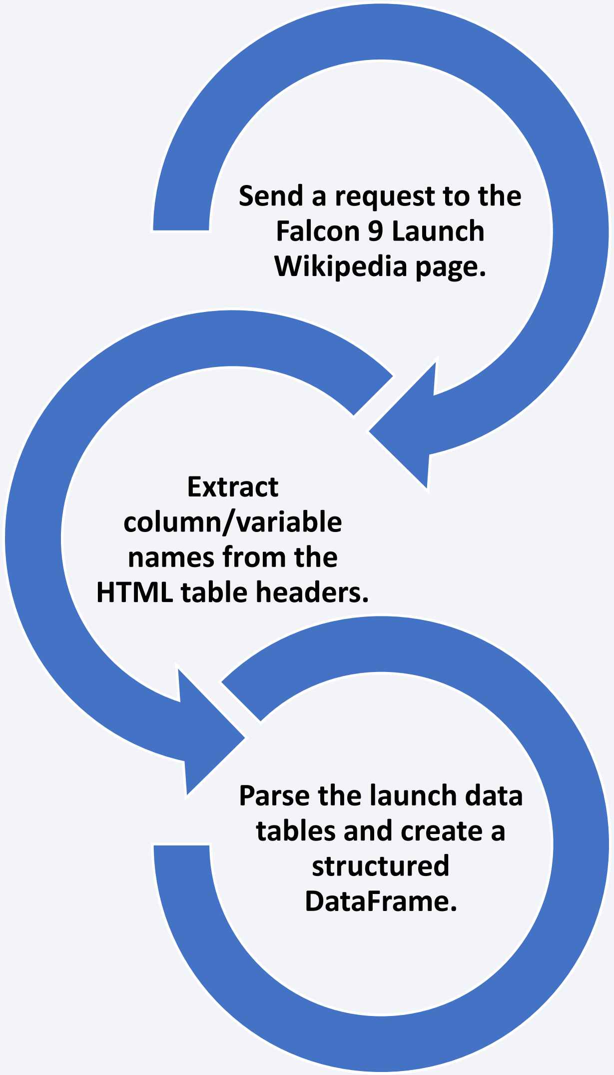
Data Collection – SpaceX API

- SpaceX provides a public API that allows data retrieval and utilization.
- The API was accessed following the outlined flowchart, and the retrieved data was then stored.

📌 The source code is available on GitHub [here]



Data Collection - Scraping



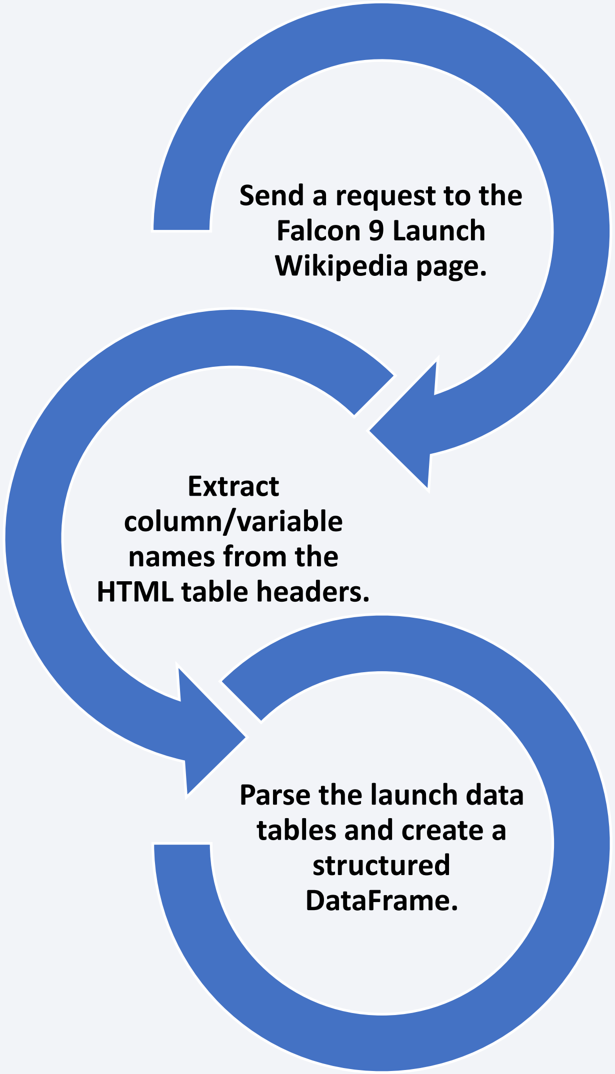
- SpaceX launch data can also be retrieved from Wikipedia.
- The data is downloaded following the outlined flowchart and then stored.

📌 The source code is available on GitHub [here]

Data Wrangling

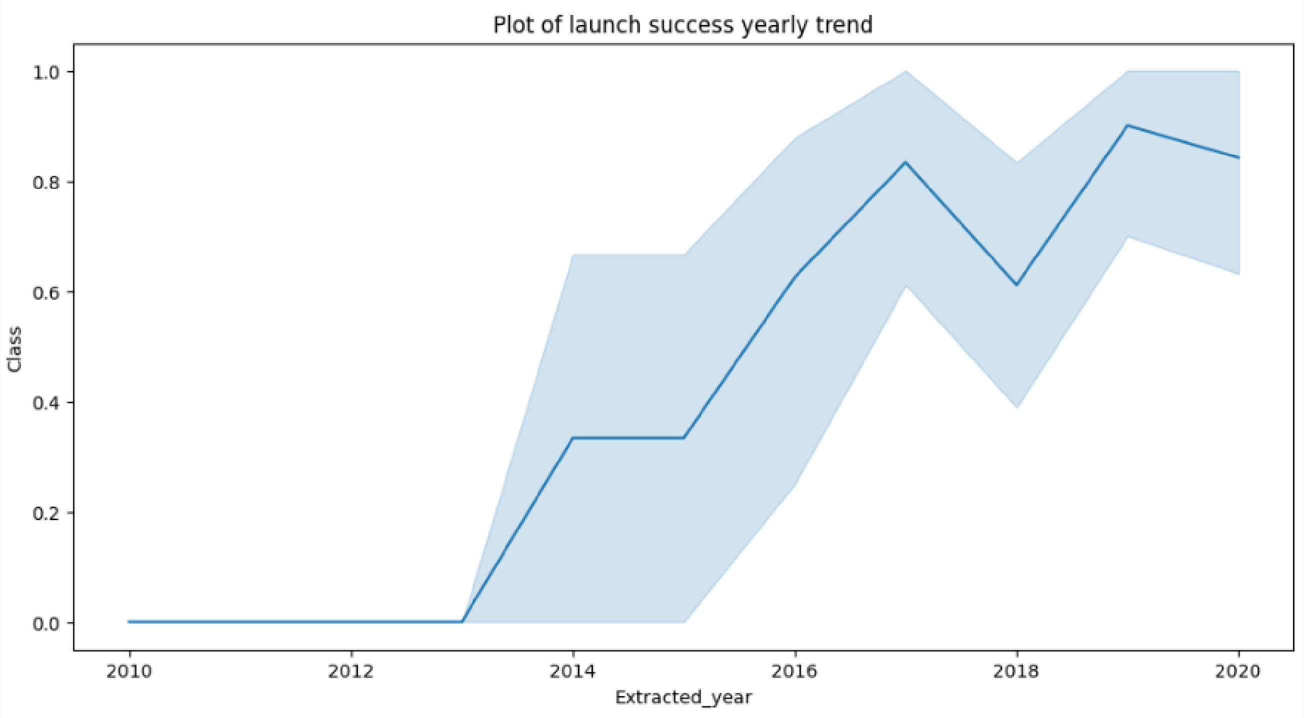
- An initial Exploratory Data Analysis (EDA) was conducted on the dataset.
- Summaries were generated for launches per site, orbit occurrences, and mission outcomes per orbit type.
- Lastly, the landing outcome label was derived from the Outcome column.

📌 The source code is available on GitHub [here]



EDA with Data Visualization

- The EDA with data visualization explored relationships between key launch factors.
  - **Flight Number vs. Launch Site**
  - **Payload Mass vs. Launch Site**
  - **Success Rate vs. Orbit Type**
  - **Flight Number vs. Orbit Type**
  - **Payload Mass vs. Orbit Type**
  - **Yearly Success Trend**



📌 The source code is available on GitHub [here]

EDA with SQL

The following SQL queries were executed:

- Display the unique launch sites in the space mission.
- Show 5 records where launch sites start with the string 'CCA’.
- Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Find the average payload mass carried by booster version F9 v1.1.
- List the date of the first successful landing outcome on the ground pad.
- List the names of boosters that succeeded in drone ships with payload masses between 4000 and 6000.
- Count the total number of successful and failed mission outcomes.
- Identify booster versions that carried the maximum payload mass using a subquery.
- Retrieve records showing the month names, failed landing outcomes on drone ships, booster versions, and launch sites for the months in 2015.
- Rank landing outcomes (e.g., Failure on drone ship or Success on ground pad) between 2010-06-04 and 2017-03-20 in descending order.

📌 The source code is available on GitHub [here]

Build an Interactive Map with Folium

- Markers: Markers are used to indicate specific points on the map, such as launch sites or important locations. They help users easily identify key spots.
- Circles: Circles are used to highlight specific areas, such as around a particular launch site, making it visually clear where a certain event or region is located.
- Lines: Lines can be used to indicate distances or paths between two points, such as the route from one launch site to another, or the distance between launch locations and key landmarks.
- Marker Clusters: These group multiple markers that are located close to each other, making the map cleaner and more organized. This is particularly helpful in cases like launches in multiple sites where markers could overlap.

📌 The source code is available on GitHub [here]

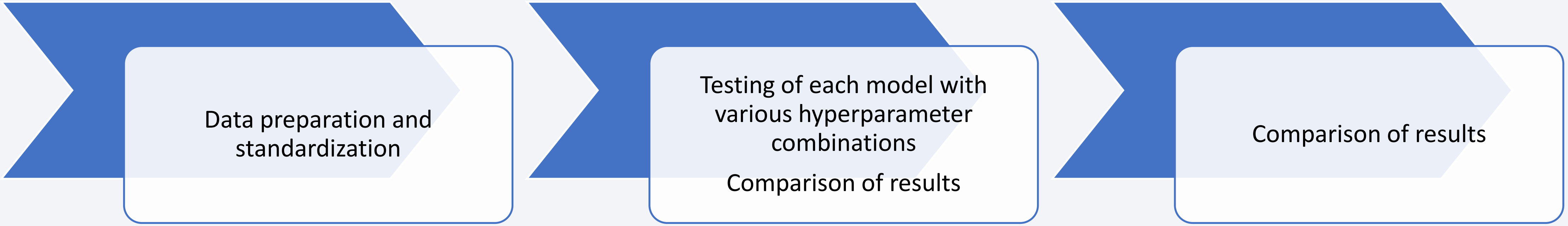
Build a Dashboard with Plotly Dash

- The dashboard includes:
  - Success Pie Chart: Percentage of successful launches by site.
  - Payload-Outcome Scatter Plot: Correlation between payload range and launch success.
  - Dropdown & Slider: Filter by site and payload.
- These visualizations help analyze payload distribution across launch sites and determine the best site for different payloads. These visualizations help analyze payload distribution across launch sites and determine the best site for different payloads.

📌 The source code is available on GitHub [here]

Predictive Analysis (Classification)

- Four classification models were evaluated: logistic regression, support vector machine, decision tree, and k-nearest neighbors.



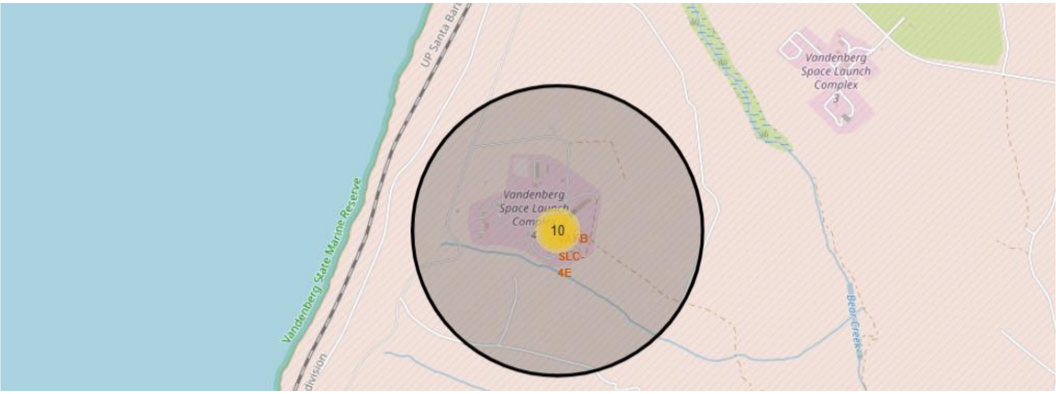
📌 The source code is available on GitHub [here]

Results

- Exploratory data analysis results
  - SpaceX operates from four distinct launch sites, with the majority of launches occurring at East Coast locations.
  - The first launches were conducted for SpaceX and NASA.
  - The average payload capacity of the Falcon 9 v1.1 booster is 2,928 kg.
  - The first successful landing occurred in 2015, five years after the initial launch.
  - Several versions of the Falcon 9 booster successfully landed on drone ships with payloads exceeding the average.
  - Nearly 100% of mission outcomes were successful.
  - In 2015, two Falcon 9 v1.1 booster versions, B1012 and B1015, failed to land on drone ships.
  - Landing success improved over the years.
  - Interactive analytics revealed that launch sites are typically located in safe areas, often near the sea, and are supported by strong logistical infrastructure.

Results

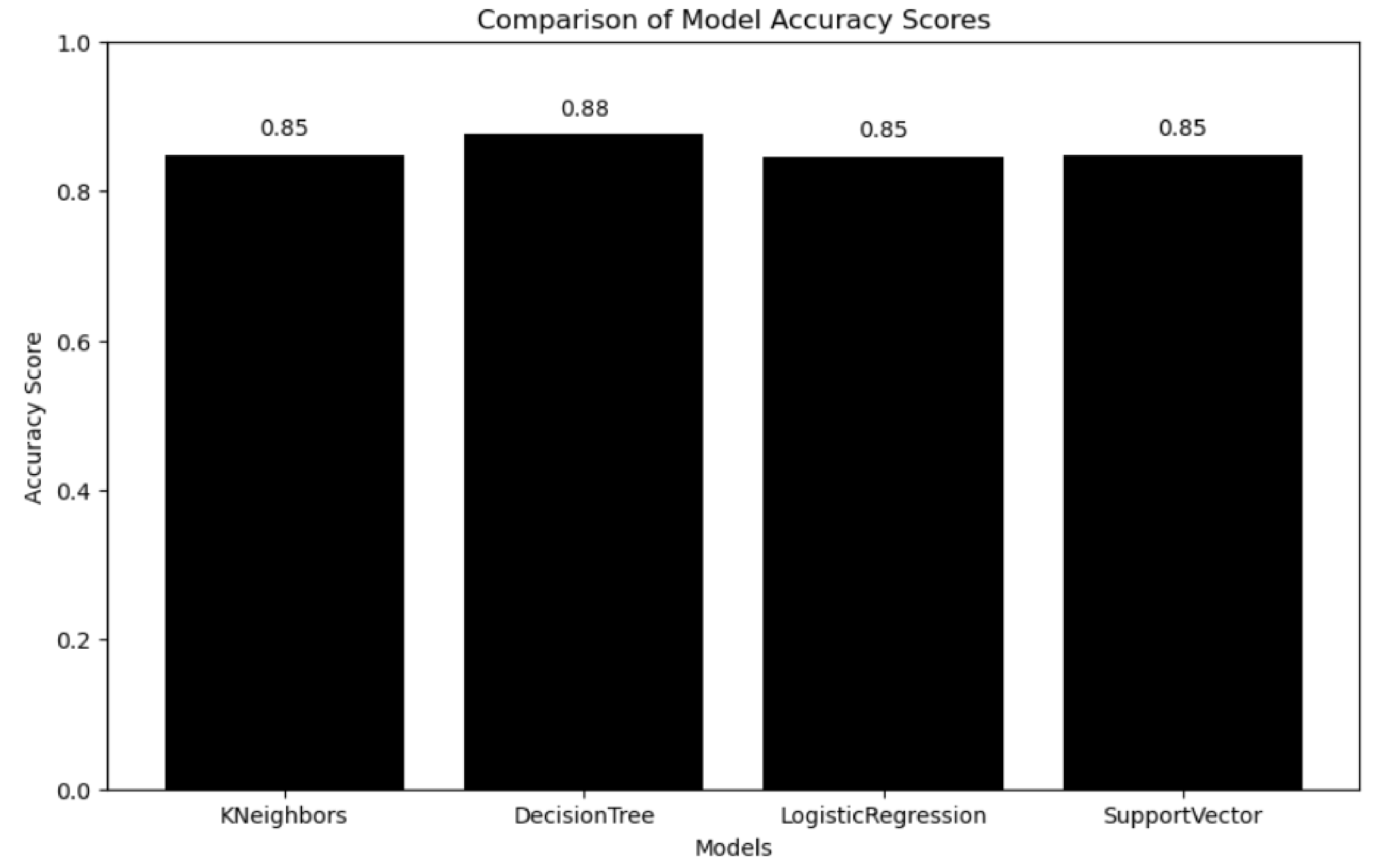

San Jose
California
Fresno
Las Vegas
10
Los Angeles
Tijuana
Mexicali
Arizona
Phoenix
Tucson
Albuquerque
New Mexico
Ciudad Juárez
Baja California
Sonora
Chihuahua
Coahuila
Baja California Sur
La Paz
Culiacán
Durango
Torreón
Monterrey
Saltillo
Tamaulipas
Nuevo Laredo
Reynosa
Texas
Austin
San Antonio
Houston
Dallas
Tulsa
Oklahoma
Arkansas
Memphis
Louisiana
Baton Rouge
Mississippi
Alabama
Tennessee
Kentucky
Atlanta
Georgia
South Carolina
Charlotte
North Carolina
Raleigh
Virginia Beach
Jacksonville
Florida
46
Miami
Nassau
The Bahamas


Vandenberg Space Launch Complex 3
Vandenberg Space Launch Complex
10
SLC
4E
Vandenberg State Marine Reserve


13
33
Kennedy Parkway North

Results

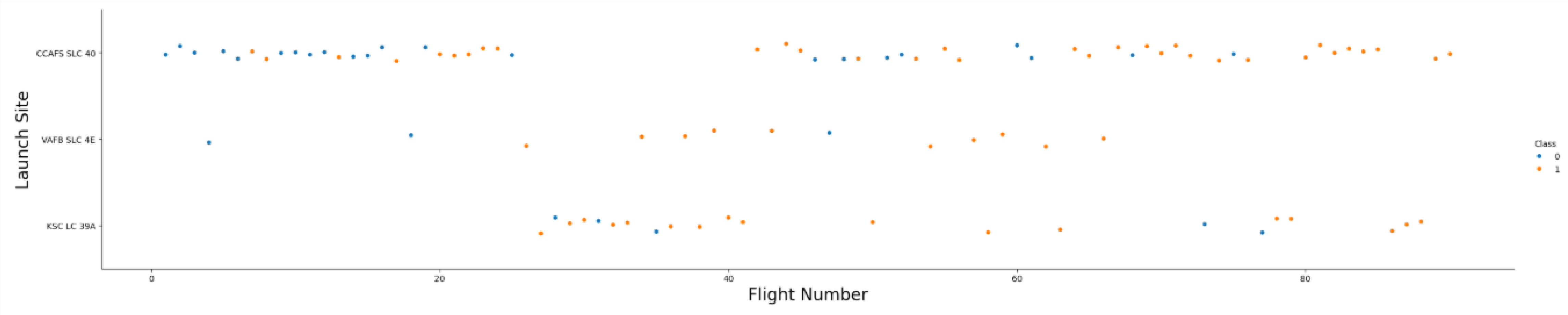
Predictive analysis revealed that the Decision Tree Classifier is the top-performing model for predicting successful landings, with an accuracy of 88%, outperforming other models such as KNeighbors, Logistic Regression, and Support Vector (all at 85%)



Section 2

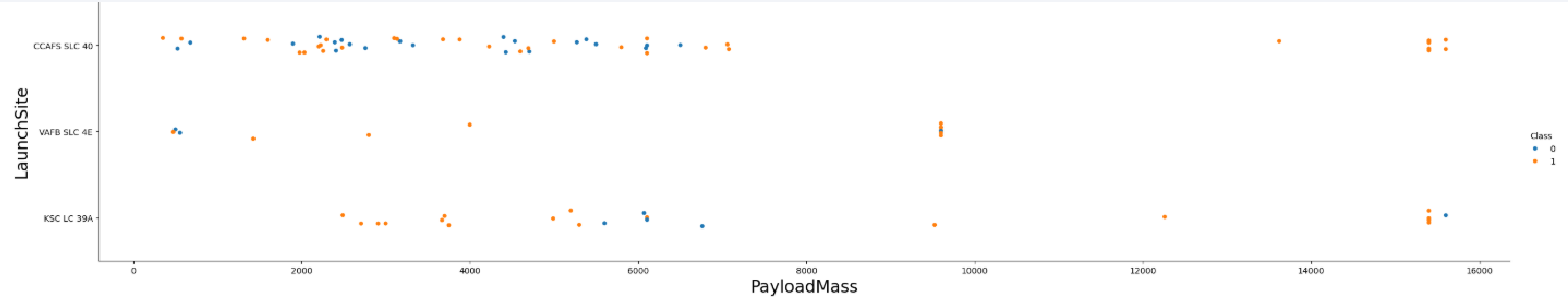
Insights drawn from EDA

Flight Number vs. Launch Site



- - The plot confirms that the best launch site currently is CCAF5 SLC 40, where most recent launches have been successful.
- - In second place is VAFB SLC 4E, followed by KSC LC 39A in third.
- - Additionally, the plot shows that the overall success rate has improved over time.

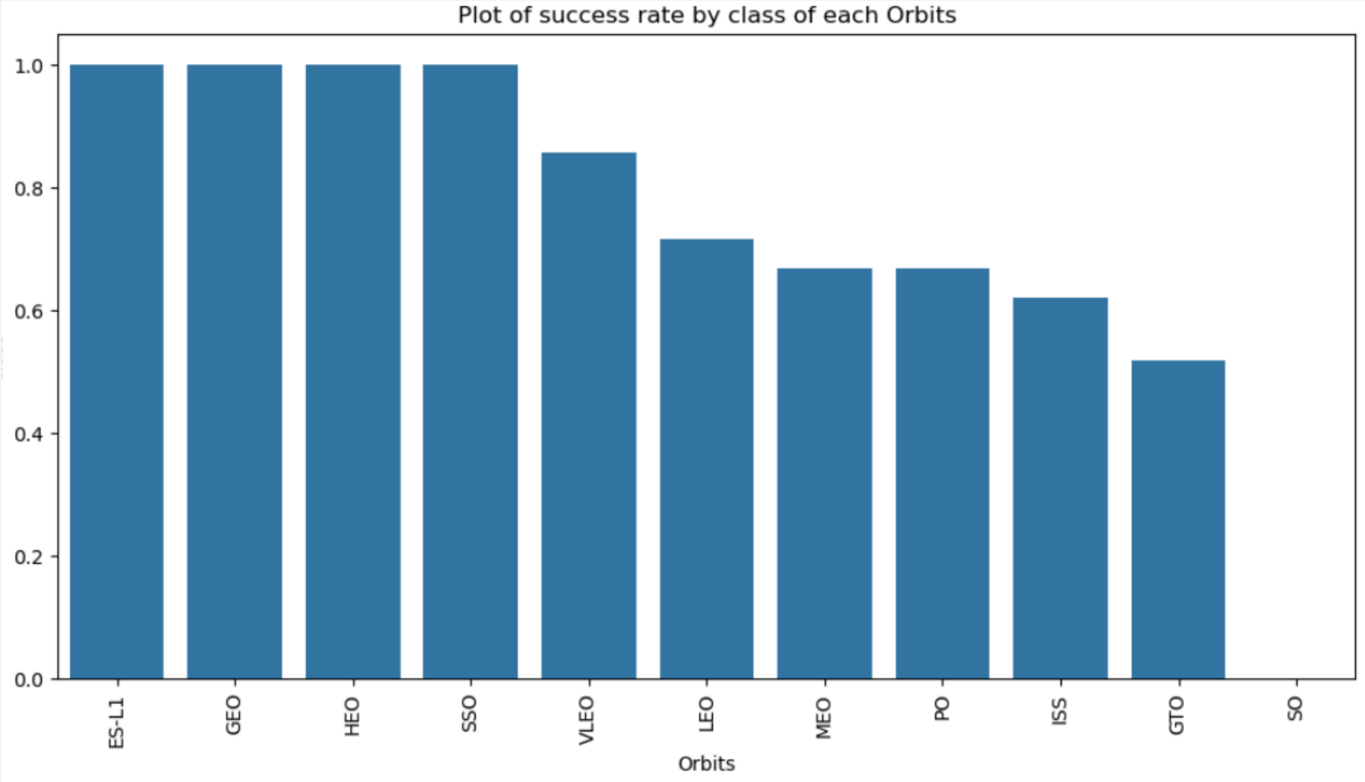
Payload vs. Launch Site



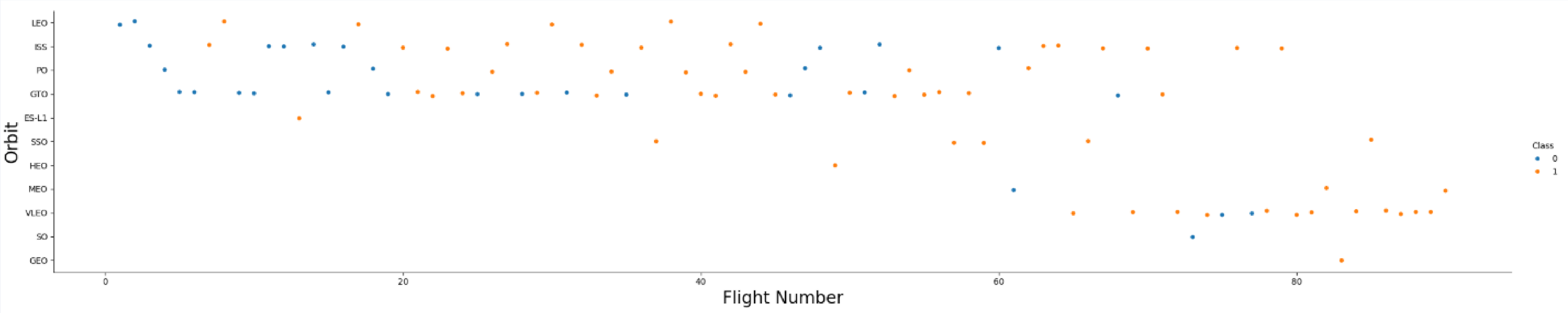
- Payloads exceeding 9,000kg (roughly the weight of a school bus) demonstrate a high success rate.
- Payloads above 12,000kg appear to be feasible only at the CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The highest success rates occur for the following orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Next in line are:
  - VLEO (above 80%)
  - LFO (above 70%)

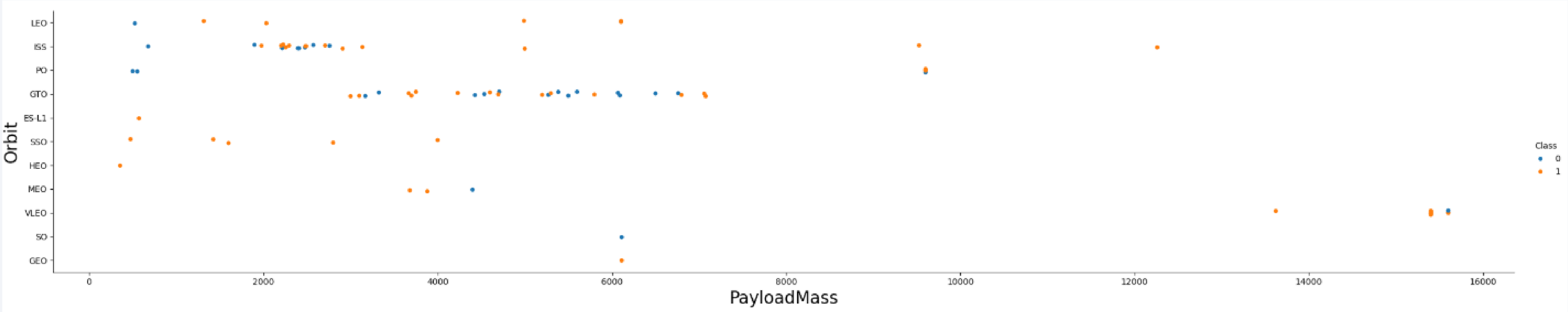


Flight Number vs. Orbit Type



- It appears that the success rate has improved over time across all orbits.
- The VLEO orbit seems to present a new business opportunity, driven by its recent increase in frequency.

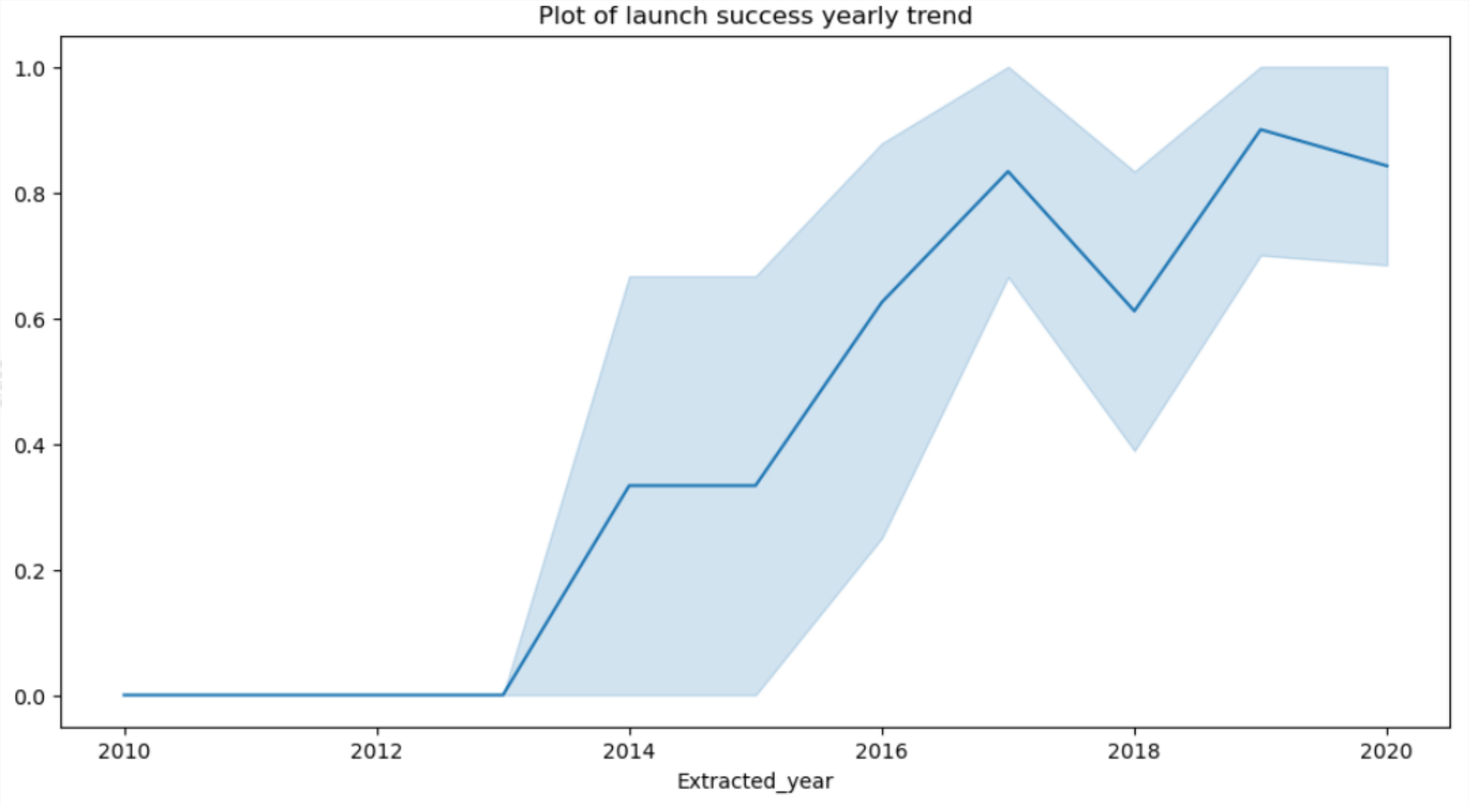
Payload vs. Orbit Type



- It seems there is no correlation between payload and success rate for the GTO orbit.
- The ISS orbit offers the widest range of payloads and a high success rate.
- There have been few launches to the SO and GEO orbits.

Launch Success Yearly Trend

- The success rate began increasing in 2013 and continued to rise until 2020.
- It appears that the first three years were a period of adjustment and technological improvement.



All Launch Site Names

- According to the data, there are four launch sites.

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- These are identified by selecting unique occurrences of the "launch_site" values from the dataset.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

Total Payload Mass

- Total payload carried by boosters from NASA

Total Payload Mass
45596

- The total payload was calculated by summing all payloads with codes containing "CRS," which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

Avg Payload Mass
2928.4

- By filtering the data by the booster version and calculating the average payload mass, we obtained a value of 2,928 kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

First Successful landing date
2015-12-22

- By filtering the data by a successful landing outcome on the ground pad and selecting the minimum date, we can identify the first occurrence, which took place on 2015/12/22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- By selecting distinct booster versions based on the filters above, the result consists of these 4 versions.

Total Number of Successful and Failure Mission Outcomes

- The number of successful and failed mission outcomes can be determined by counting occurrences of each outcome in the dataset.

Mission Outcome	Occurrences
Success	100
Failure	1

- By categorizing mission outcomes and tallying the records for each group, we derived the summary above.

Boosters Carried Maximum Payload

- Boosters that have carried the maximum payload mass, with a payload mass of 15,600 kg for all of them.

Booster Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

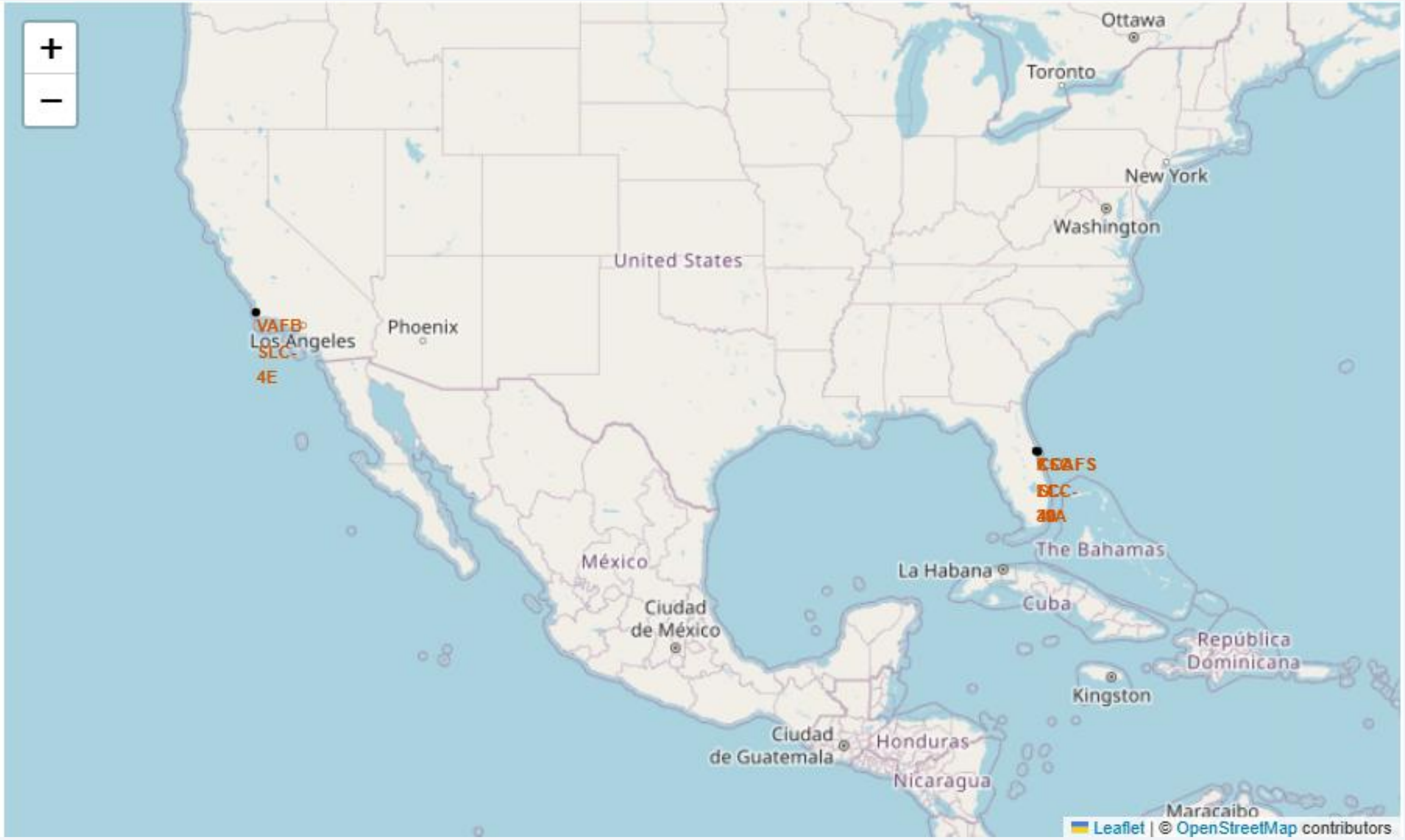
Landing Outcome	Occurrences
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This data perspective highlights that "No attempt" must be considered.

Section 3

Launch Sites Proximities Analysis

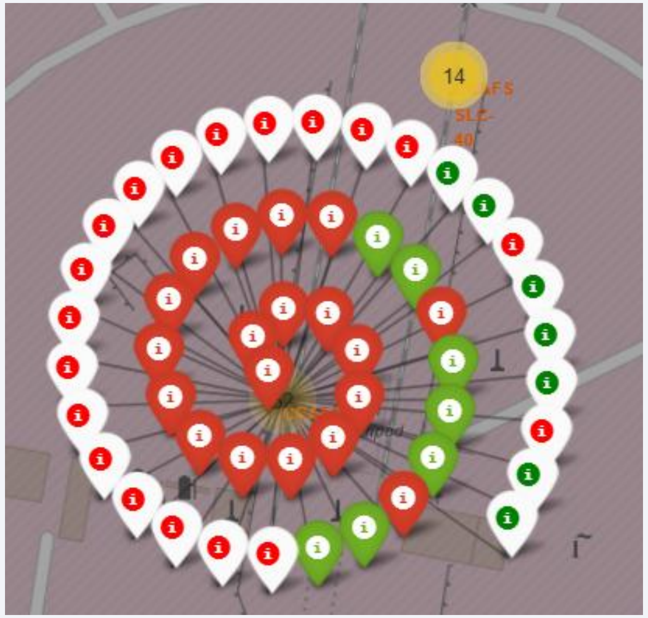
All launch sites



- Launch sites are located near the sea, likely for safety reasons, while still being relatively close to roads and railroads.

Launch outcomes categorized by site.

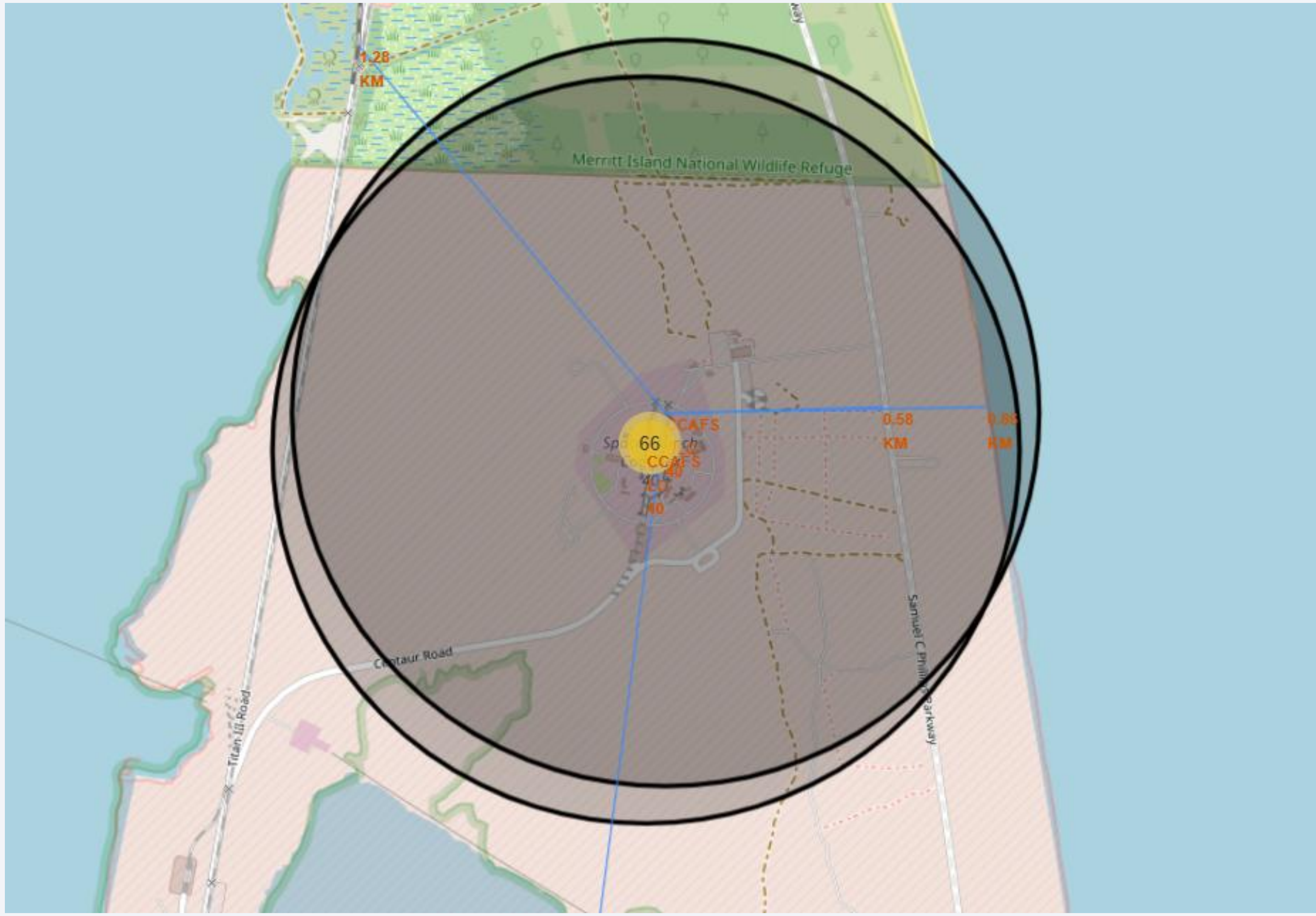
- Example of launch outcomes from the CCAFS SLC-40 launch site.



- Green= **Successful Launch**
- Red= **Failed Launch**

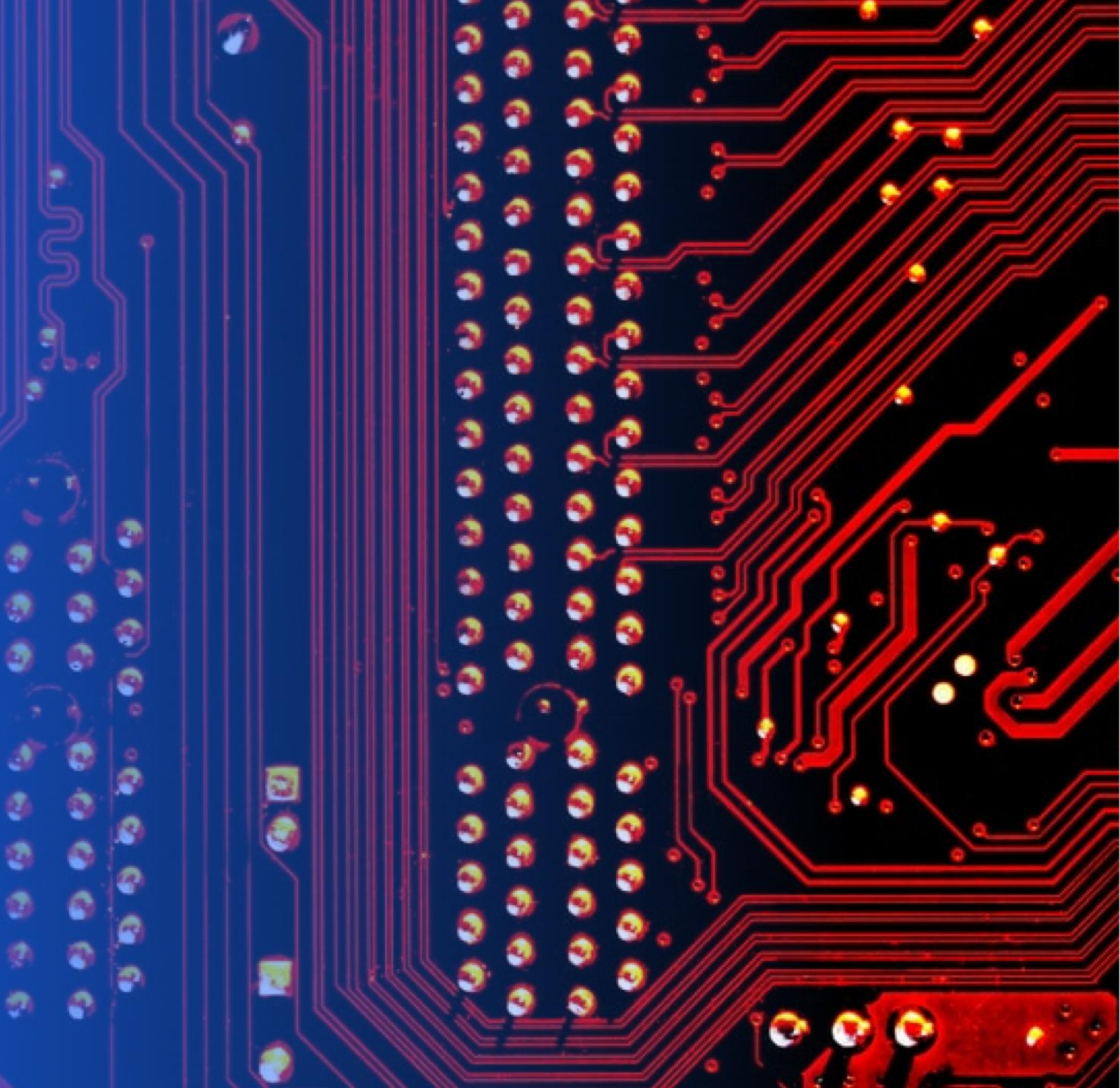
Distance to Launch Site

The obtained results show that all launch sites are located at a safe distance from railway lines and cities.



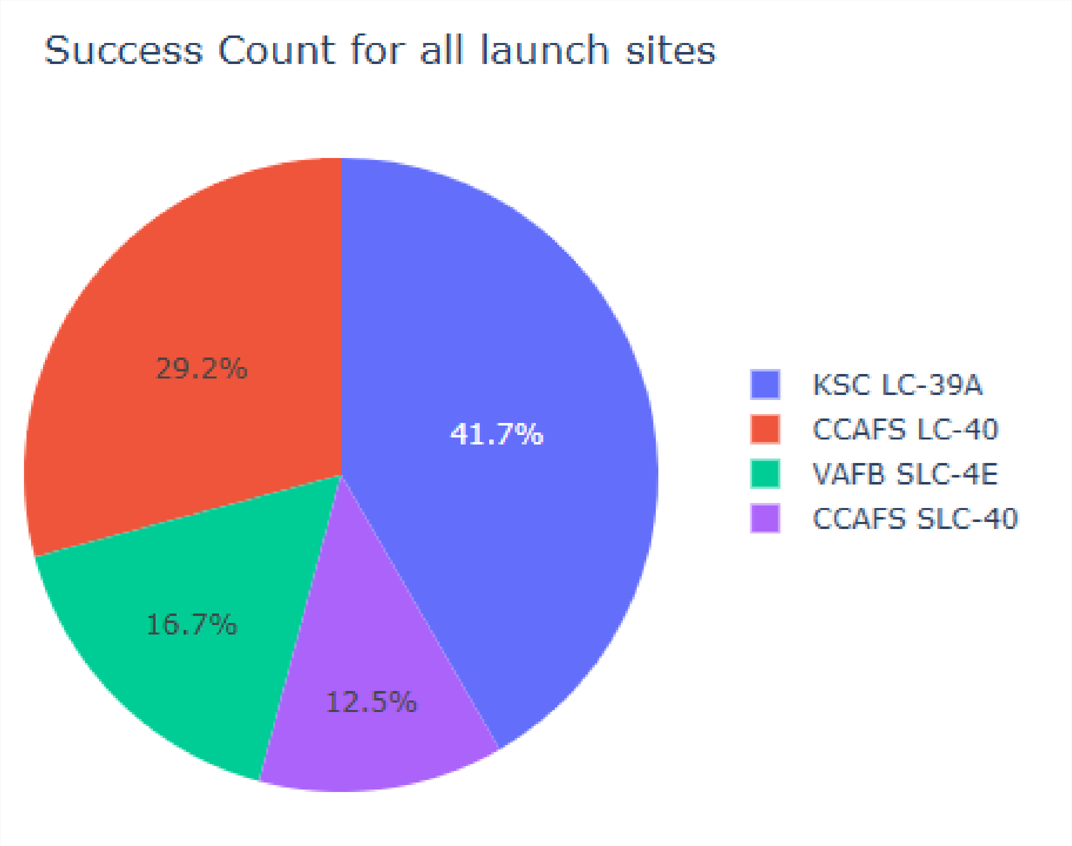
Section 4

Build a Dashboard with Plotly Dash



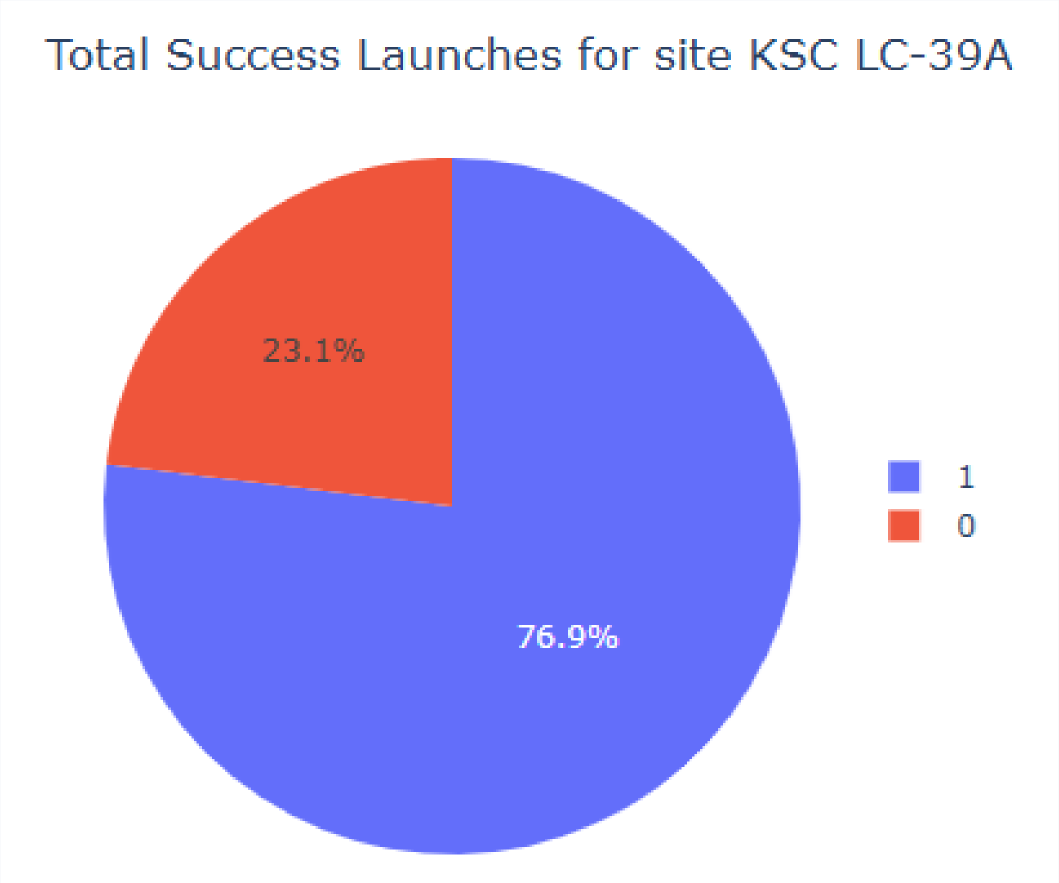
Success Rates of Rocket Launches Across Different Launch Sites

- The highest success rates for launches were recorded at:
  - KSC LC-39A with a 41.7% success rate.
  - CCAFS LC-40 with a 29.2% success rate.

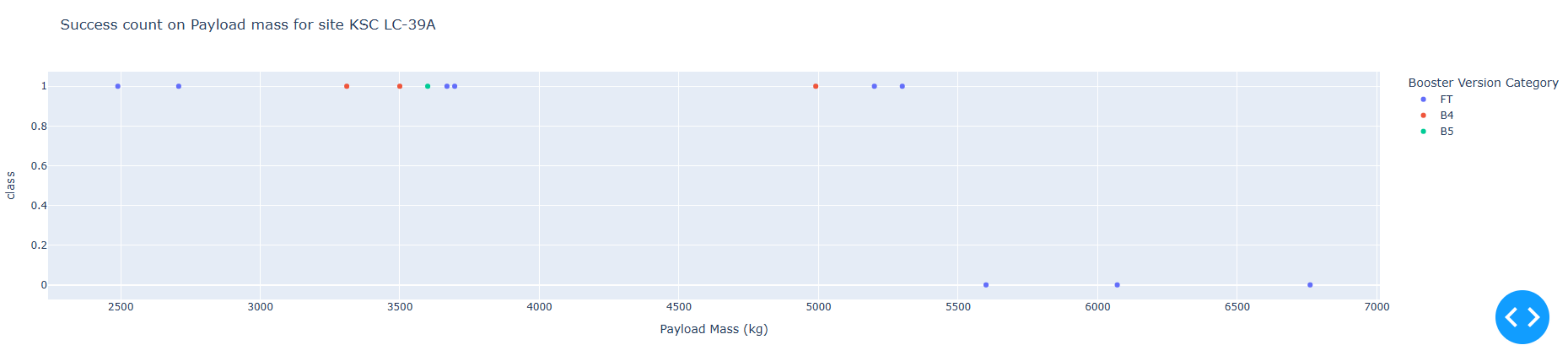


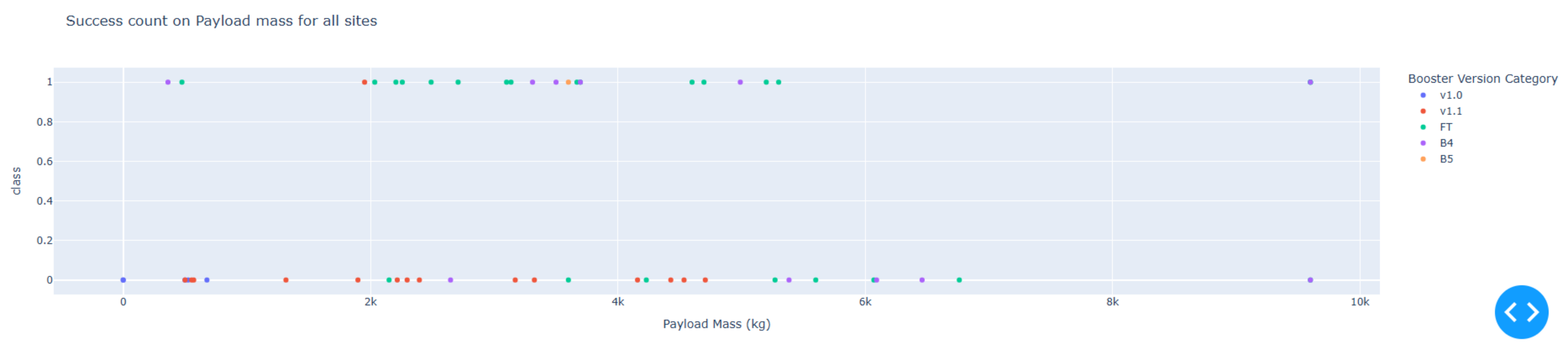
Total Success Launches for site KSC LC-39A

- Site KSC LC-39 success rate is 76.9%



Payload vs. Launch Outcome Scatter Plot



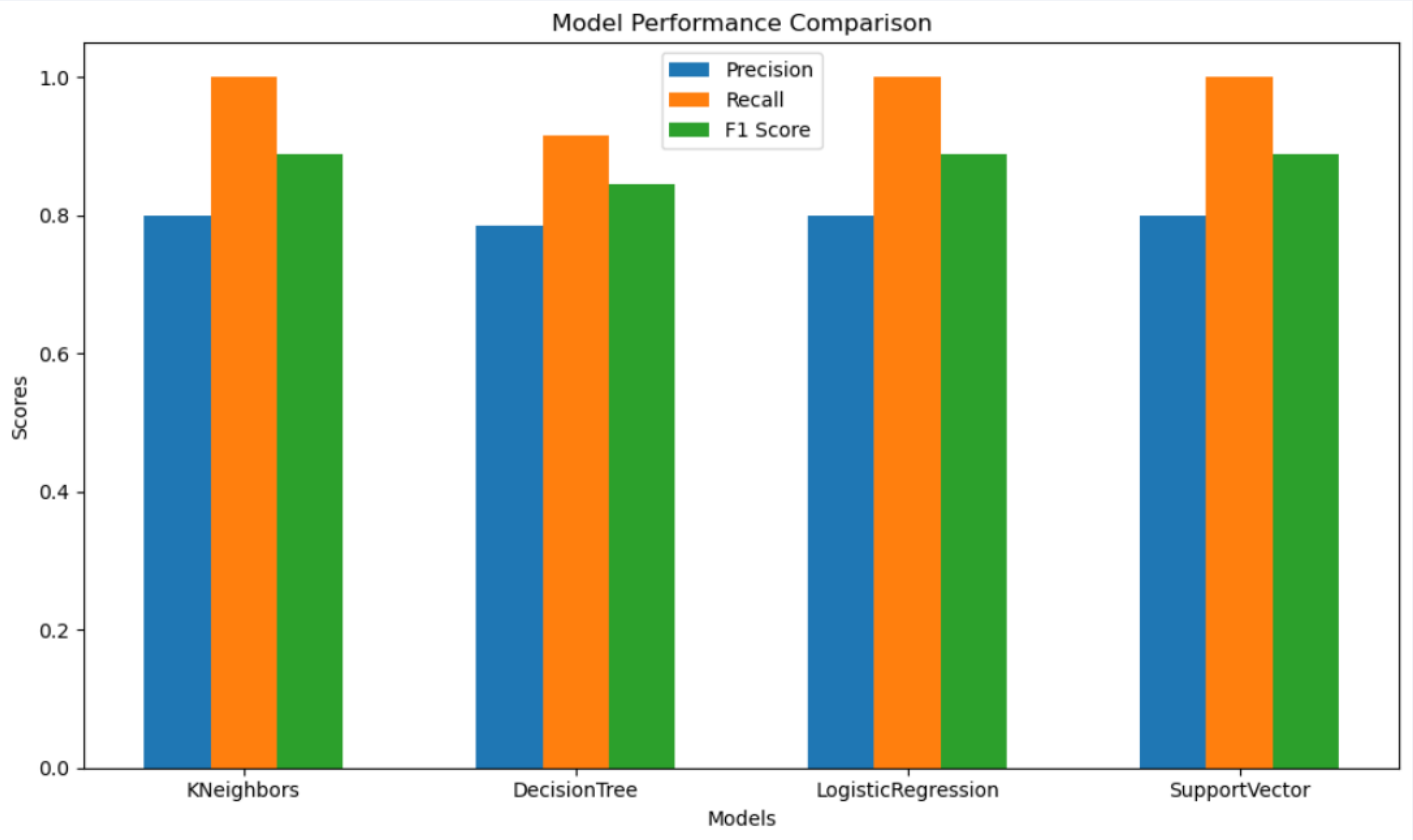


Section 5

Predictive Analysis (Classification)

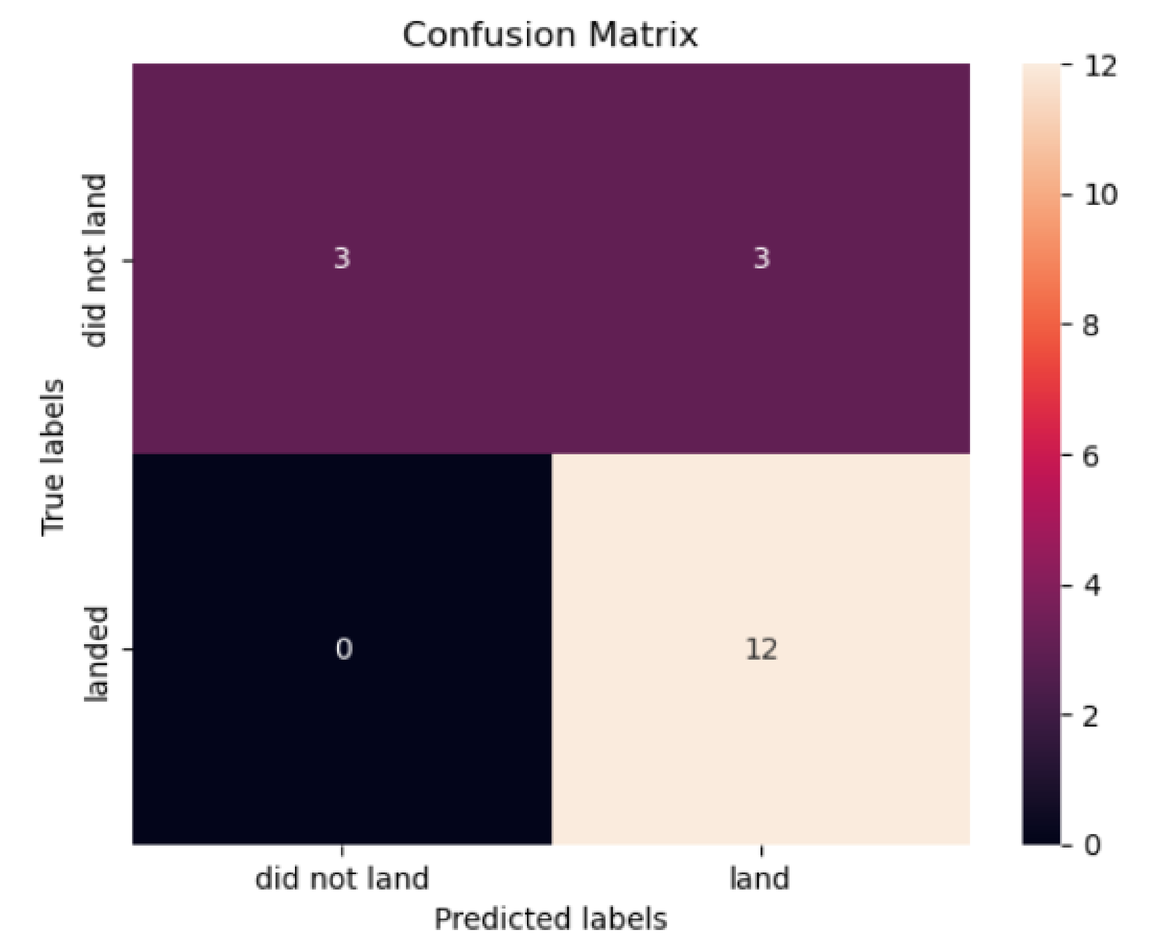
Classification Accuracy

KNeighbors, LogisticRegression, and SupportVector all have high precision (0.80), perfect recall (1.00), and an F1 score of 0.89. DecisionTree has slightly lower precision (0.79) and an F1 score of 0.85. Despite this, it achieved the highest score during cross-validation, with its optimal parameters contributing to its best performance.



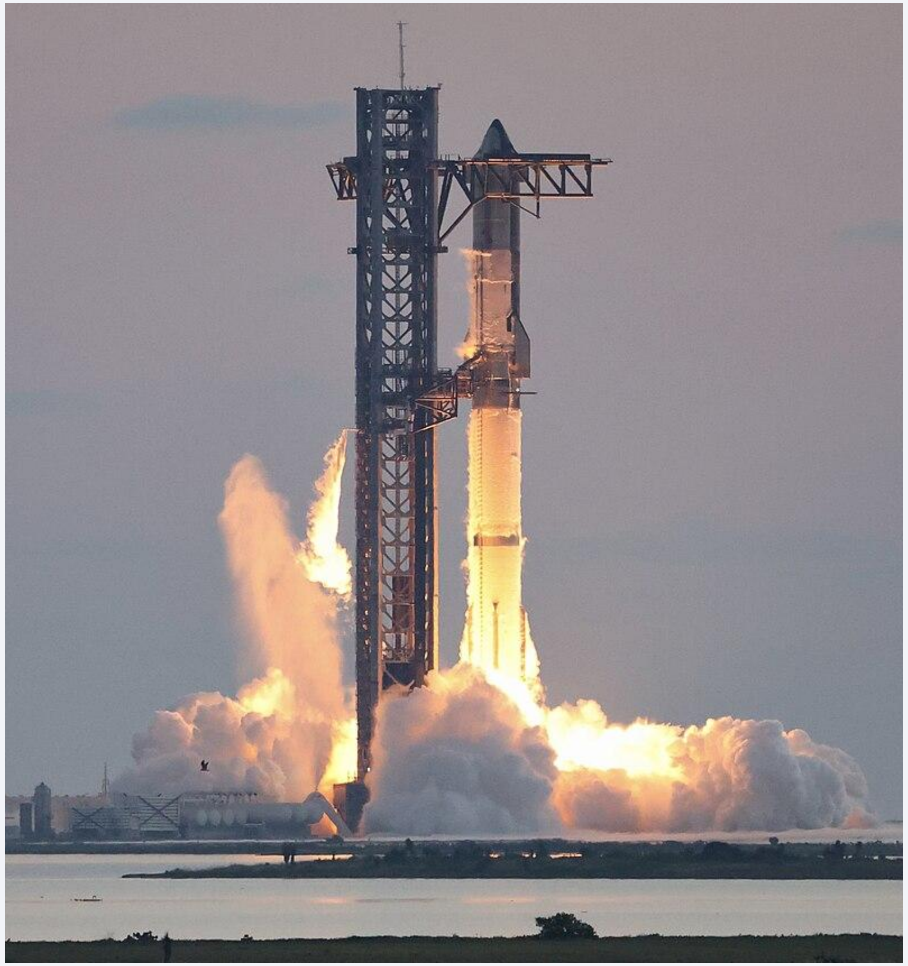
Confusion Matrix

The confusion matrix for the decision tree classifier reveals that the model is capable of distinguishing between the various classes. However, the primary issue lies in the occurrence of false positives, where unsuccessful landings are incorrectly classified as successful by the model.



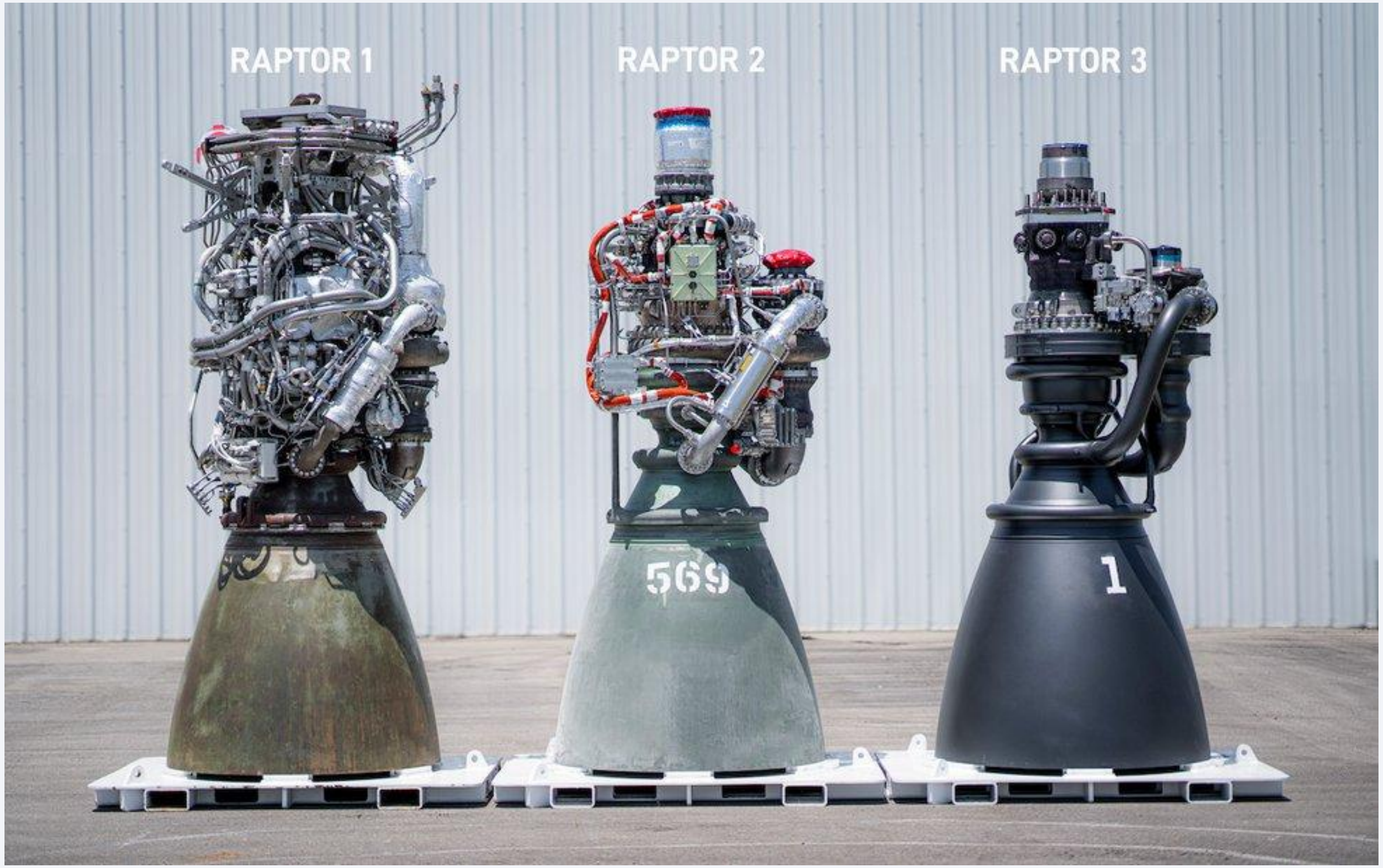
Conclusions

- Based on the analysis, we can draw the following conclusions:
  - Launch sites with a higher number of flights tend to exhibit a greater success rate.
  - The launch success rate showed a consistent upward trend from 2013 through 2020.
  - Orbits such as ES-L1, GEO, HEO, SSO, and VLEO demonstrated the highest success rates.
  - Among all launch sites, KSC LC-39A recorded the highest number of successful launches.
  - The decision tree classifier outperformed other machine learning algorithms for this specific task.



Appendix

- 📂 [**GitHub**]



Thank you!