APPLIED DATA SCIENCE CAPSTONE

A STUDY OF SPACE RACE

NAME: WENHAO CHEN

DATE: JAN 30TH 2026

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Summary of Methodologies
  - 1. Data Collection:
    - Accessed SpaceX launch data from Wikipedia and spacex.
  - 2. Data Cleaning & Preparation:
    - Understand, clean, and preprocess the data.
    - Stored data in database and performed SQL queries.
    - Conducted exploratory data analysis.
    - Created new features and standardized the data.
  - 3. Interactive Visualizations:
    - Built an interactive using Folium.
    - Built an interactive dashboard with Plotly Dash.
  - 4. Model Building & Evaluation:
    - Implement multiple models.
    - Tuned hyperparameters with GridSearchCV.
    - Evaluated models using test data accuracy.

INTRODUCTION

- **Project Background and Context**
  - This capstone project focuses on predicting the successful landing of the Falcon 9 first stage. SpaceX offers rocket launches at substantially lower costs than other providers, primarily because it can recover and reuse the first stage. By accurately predicting landing success, we can better estimate launch costs and generate valuable insights for companies competing with SpaceX in launch service bids.
- **Questions we focus on**
  - What factors influwence the successful landing of the Falcon 9 first stage?
  - How accurately can machine learning models predict landing outcomes?
  - Which machine learning model performs best in predicting landing success?

METHODOLOGY

- Data collection:
  - SpaceX Rest API
  - Web Scrapping
- exploratory data analysis (EDA)
  - Pandas, numpy
  - SQL
  - Matplotlib, seaborn
- interactive visual analytics
  - Folium
  - Plotly
- classification models(scikit-learn)
  - Logistic Regression
  - SVM
  - Decision Trees,
  - KNN

DATA COLLECTION

- Step 1:
  - Fetch data from spaceX
- Step 2:
  - Web Scraping Wikipedia
- Step 3:
  - Data Wrangling

DATA COLLECTION STEP 1:

connect to theSpaceX API.

Endpoint: `https://api.spacexdata.com/v4/launches/past`

Parse API Response

Convert API response from JSON to a Python dictionary.

Extract relevant fields: launch date, launch site, payload mass, rocket type, outcome.

Store Data Locally

Convert data into Pandas DataFrame

Export to csv file.

STEP 2
SCRAPING

Initiate Web Scraping

Target page `https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches`

Parse HTML Content

Use `BeautifulSoup` to parse the HTML content.

Extract the HTML table containing Falcon 9 launch records.

Converting data

Convert the extracted HTML table into a pandas DataFrame.

Clean and format the DataFrame, ensuring data consistency.

Export data to csv file

DATA WRANGLING

Data Cleaning

- Identify and fill missing values in the dataset.(mean value)

Explore the data

- Identified column data types and categorized attributes as numerical or categorical.
- Visualized the number of launches per launch site to understand launch distribution.
- Analyzed the distribution of orbit types within the dataset.
- Explored launch outcomes and grouped them by binary result (successful vs. failed landings)

Identify labels

- Examined launch outcomes and converted them into a binary classification format (success or failure).
- Standardized outcome labels to ensure consistency across the dataset.

Identify Target

- Stored the processed launch outcome as the “Class” label.
- Used the Class attribute as the target variable for training and evaluating machine learning models.

PREDICTIVE ANALYSIS METHODOLOGY

- **1. Problem Definition**
  - Define the objective as predicting the successful landing of the Falcon 9 first stage.
  - Formulate the task as a **binary classification problem** (success vs. failure).
- **2. EDA**
- **3. Label & Target Definition**
  - Convert landing outcomes into a binary **Class** label.
  - Define **Class** as the target variable for model training.
- **4. Model Selection**
  - Split data into training and testing sets.
  - Train multiple machine learning models (e.g., Logistic Regression, Decision Tree, Random Forest, SVM).
  - Compare models to identify the best performer.
  - Apply cross-validation to reduce overfitting.
  - Tune hyperparameters for optimal performance.
- **9. Model Evaluation**
  - Evaluate models using accuracy, precision, recall, F1-score, and confusion matrix.
  - Compare results across models.

DATA VISUALIZATION

To gain insights into the importance of each variable, multiple visualizations were created to explore relationships between key features and launch outcomes:

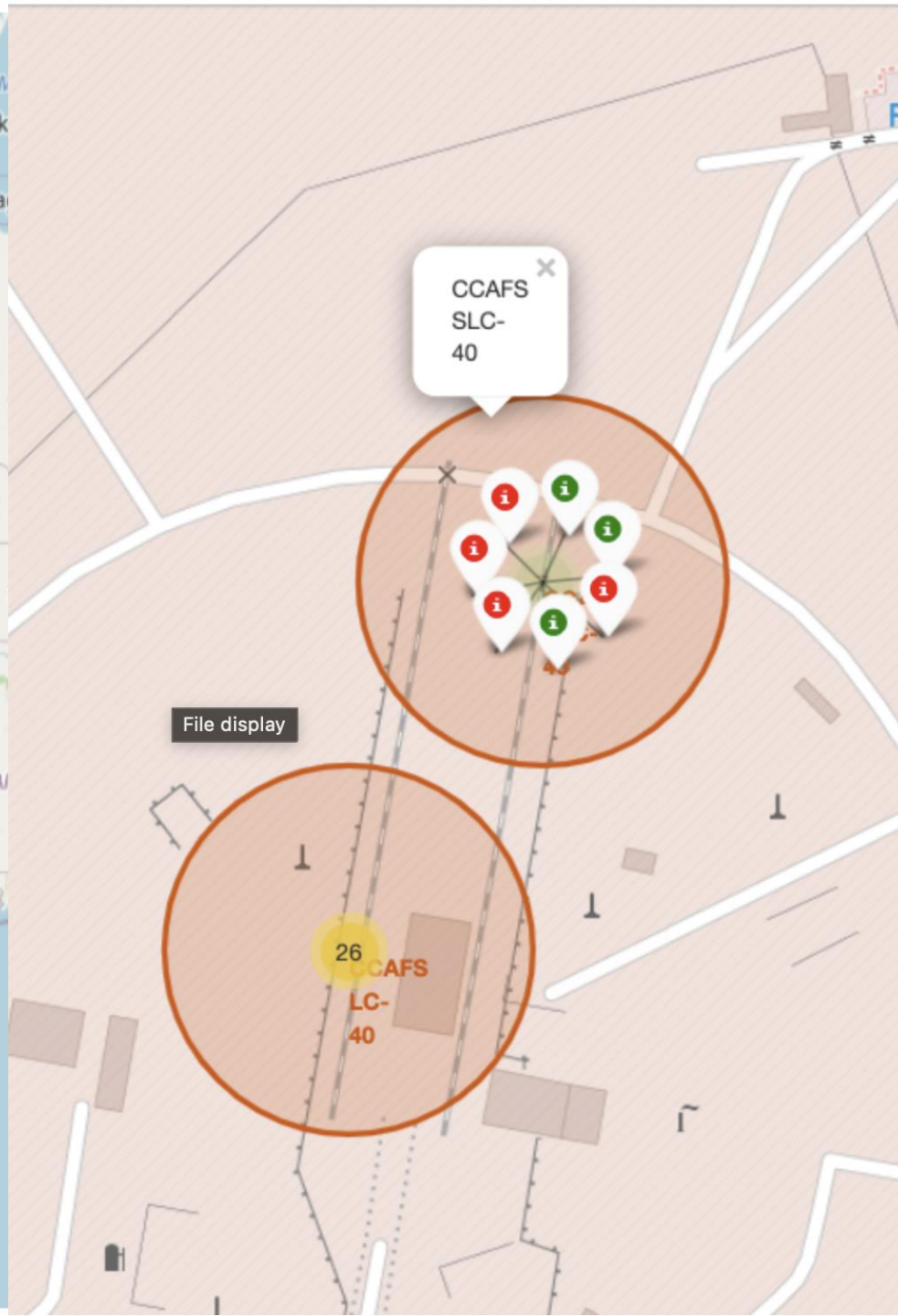
- Analyzed the relationship between **Flight Number and Outcome** to observe performance changes over successive launches.
- Examined **Flight Number and Launch Site** to identify operational patterns across different launch locations.
- Explored **Payload Mass and Launch Site** to assess how payload size varies by launch facility.
- Investigated **Orbit Type and Outcome** to determine whether certain orbits are associated with higher landing success rates.
- Visualized **Flight Number and Orbit** to study mission progression and orbit usage over time.
- Analyzed **Payload Mass and Orbit** to understand how payload requirements differ across orbit types.
- Evaluated the **Yearly Success Rate** to identify long-term trends in Falcon 9 landing performance.
- **Feature Engineering**
  - Expanded categorical variables (such as launch site and orbit type) into **dummy variables** using one-hot encoding.
  - Prepared the dataset for machine learning models by transforming categorical features into numerical representations.
- **Data Preparation**
  - Converted numerical columns to **float64** data type to ensure compatibility with machine learning algorithms.
  - Verified data consistency and finalized the dataset for model training and evaluation.

SQL QUERIES USED

- Show each unique launch site
- Show 5 records where launch site names begin with 'CCA'
- Display the total payload mass carried by boosters launched by 'NASA (CRS)'
- Display average payload mass carried by booster version F9 v1.1
- List the date of the first successful ground landing outcome
- List the booster versions with successful outcomes landing on the drone ship with payloads between 4000kg and 6000kg.
- List the number of successful and failed mission outcomes
- List all of the booster versions that carried the max payload mass
- List the month name, outcome, booster version, and launch site for missions with failure outcomes landingon a drone ship in 2015.
- Show the distribution of outcomes between June 4th, 2010 and March 20th, 2017

INTERACTIVE MAP WITH FOLIUM

- Identified geographical patterns in the data by visualizing launch-related information on a map.
- Marked **all launch sites** to provide an overview of their geographic distribution.
- Displayed **successful and failed launches** to examine spatial differences in launch outcomes.
- Visualized the **distances between launch sites and nearby landmarks** to explore potential geographical or operational influences.



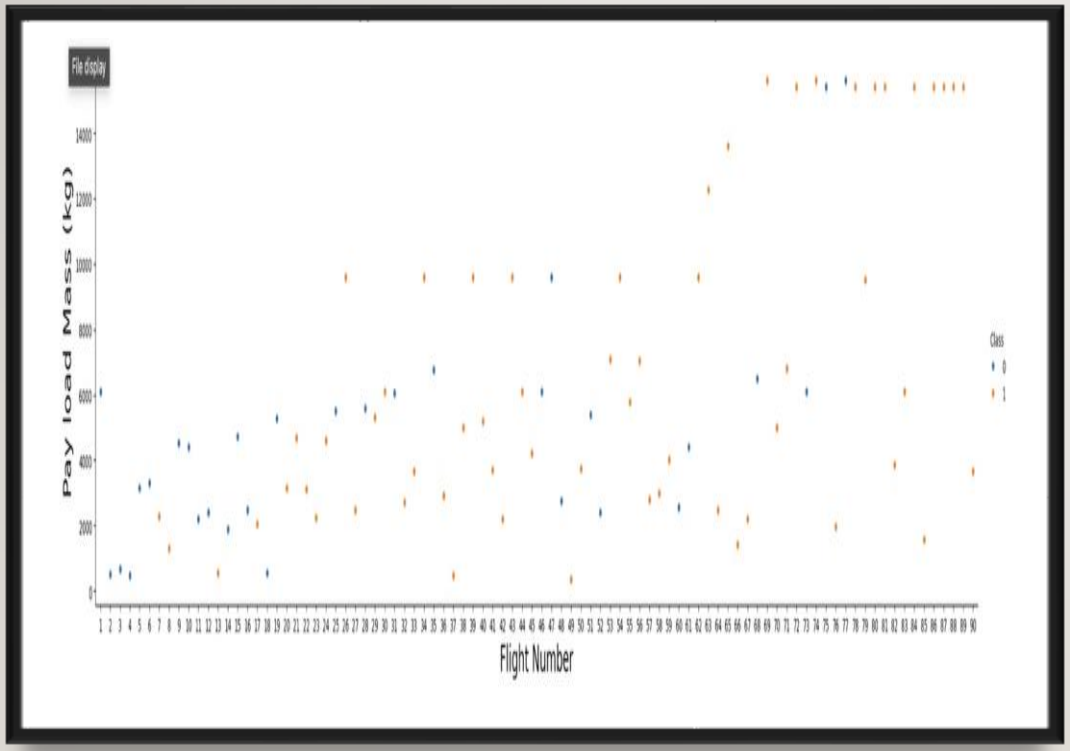
BUILD A DASHBOARD WITH PLOTLY DASH

- To enable interactive exploration of the data, a **Plotly Dash dashboard** was developed with the following components:
- A **Launch Site dropdown selector**, which dynamically updates:
- **Pie Chart**
  - When **all sites** are selected, the chart displays the distribution of successful launches across all launch sites.
  - When a **specific launch site** is selected, the chart shows the proportion of successful versus failed launches for that site.
- **Scatter Plot**
  - When **all sites** are selected, the plot visualizes launch outcomes based on payload mass and booster version across all sites.
  - When a **specific launch site** is selected, the plot displays payload mass versus booster version outcomes for the selected site only.
  - A **Payload Mass range slider** that allows users to filter data points in the scatter plot, enabling focused analysis on specific payload ranges.

PREDICTIVE ANALYSIS (CLASSIFICATION)

1. Load data
2. Apply StandardizedScaler on X
3. Convert Y to numpy array
4. Split training and testing data
5. Use GridSearchCV to test hyperparameters for multiple models:
   - Logistic Regression
   - SVC
   - Decision Tree Classifier
   - K Neighbors Classifier

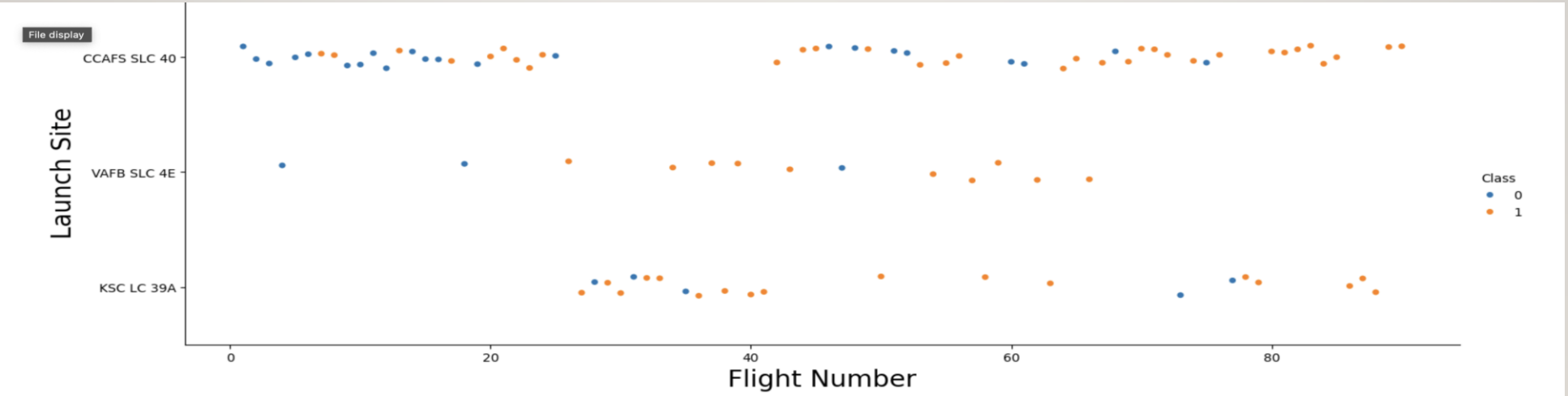
EDA RESULT FROM VISUALIZATION FLIGHT NUMBER VS PAY LOAD MASS



- The scatter plot shows that landing success improves significantly as the flight number increases, indicating the impact of technological maturity and operational experience. While early launches exhibit a higher failure rate, later missions achieve higher success rates even with heavier payloads. This suggests that flight number is a strong predictor of landing success, whereas payload mass alone does not determine the outcome.

EDA RESULT FROM VISUALIZATION FLIGHT NUMBER VS LAUNCH SITE

- **Mixed Outcomes at Major Launch Sites:** Both CCAFS SLC 40 and KSC LC 39A exhibit a combination of successful and unsuccessful landings, suggesting that landing success is influenced by factors beyond the launch site alone.
- **Consistent Activity Across Flight Numbers:** Launches at all sites occur across a wide range of flight numbers, indicating sustained operational activity over time without a clear site-specific trend in

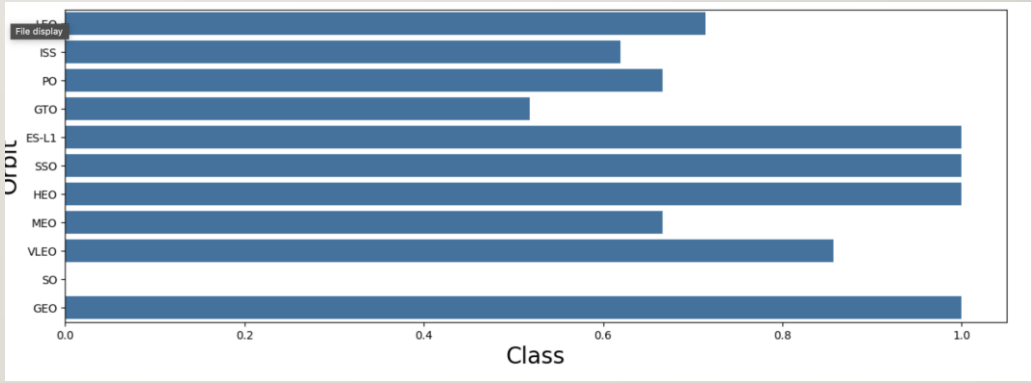


EDA RESULT FROM VISUALIZATION PAYLOAD VS LAUNCH SITE

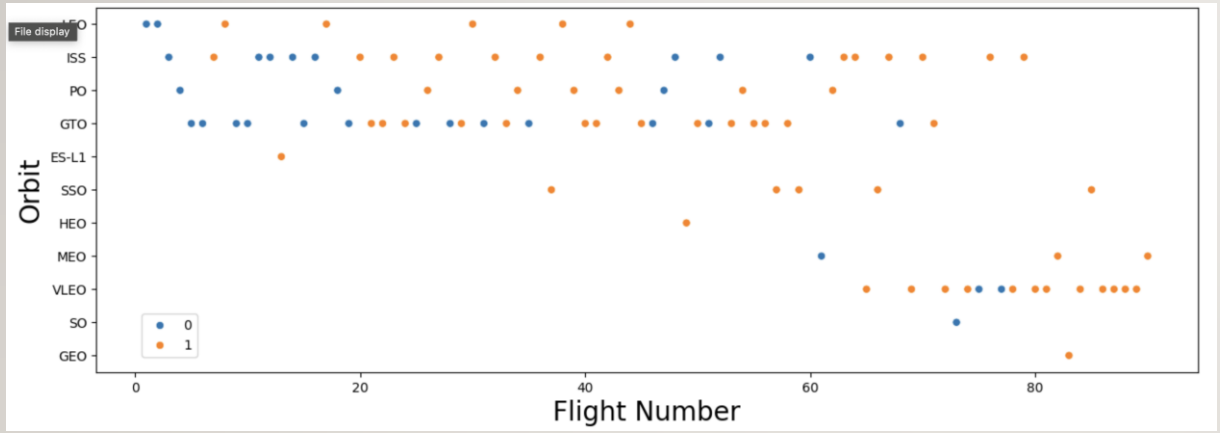
- For every launch site the higher the payload mass, the higher the success
- • Most of the launches with payload mass over 7000 kg were successful.
- • KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

EDA RESULT FROM VISUALIZATION SUCCESS RATE OF EACH 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



EDA RESULT FROM VISUALIZATION FLIGHTNUMBER AND ORBIT TYPE

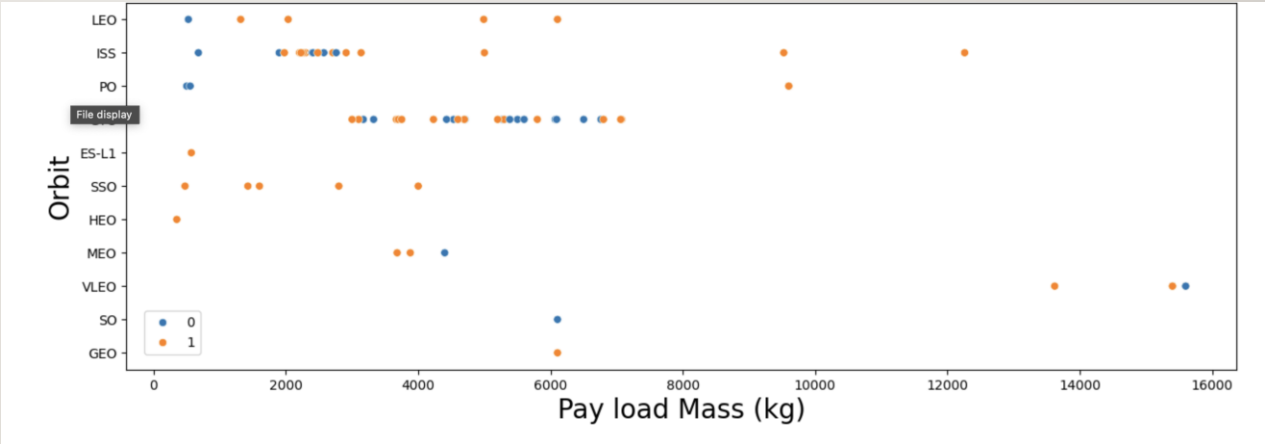


In the LEO orbit the Success appears related to the number of flights;

on the other hand, there seems to be no relationship between flight number when in GTO orbit.

EDA RESULT FROM VISUALIZATION PAYLOAD AND 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 annual launch success rate has shown a significant improvement from 2013 onwards, reaching over 80% by 2020.
- Despite a dip in 2018, the overall trend indicates increasing reliability and success in Falcon 9 launches over the years.

SQL RESULT
LAUNCH SITE RELATIVE TASK

- 1.There is 4 unique launch site.
- 2. we inquire 5 record from launch site name start with ‘CCA’

```
%%sql

SELECT DISTINCT Launch_Site from SPACEXTABLE;
```

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

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

File display

```
%%sql

SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

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

SQL
PAY LOAD MASS AND DATA

Done.

LaunchDate
2015-12-22

File display

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

File display

TOTAL_PAYLOAD
45596

12]:

AVG_PAYLOAD_MASS
2534.6666666666665

- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.

SQL
TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

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

File display

Mission_Status	COUNT(*)
Failure	1
Success	100

SQL
NAMES OF THE BOOSTER_VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS.

File display

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

SQL

RECORDS WHICH WILL DISPLAY THE MONTH NAMES, FAILURE LANDING_OUTCOMES IN DRONE SHIP ,BOOSTER VERSIONS, LAUNCH_SITE FOR THE MONTHS IN YEAR 2015

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

Month	Landing_Outcome	Booster_Version	Launch_Site	Date
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

SQL
RANK THE COUNT OF LANDING OUTCOMES (FAILURE OR SUCCESS) BETWEEN THE DATE 2010-06-04 AND 2017-03-20

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

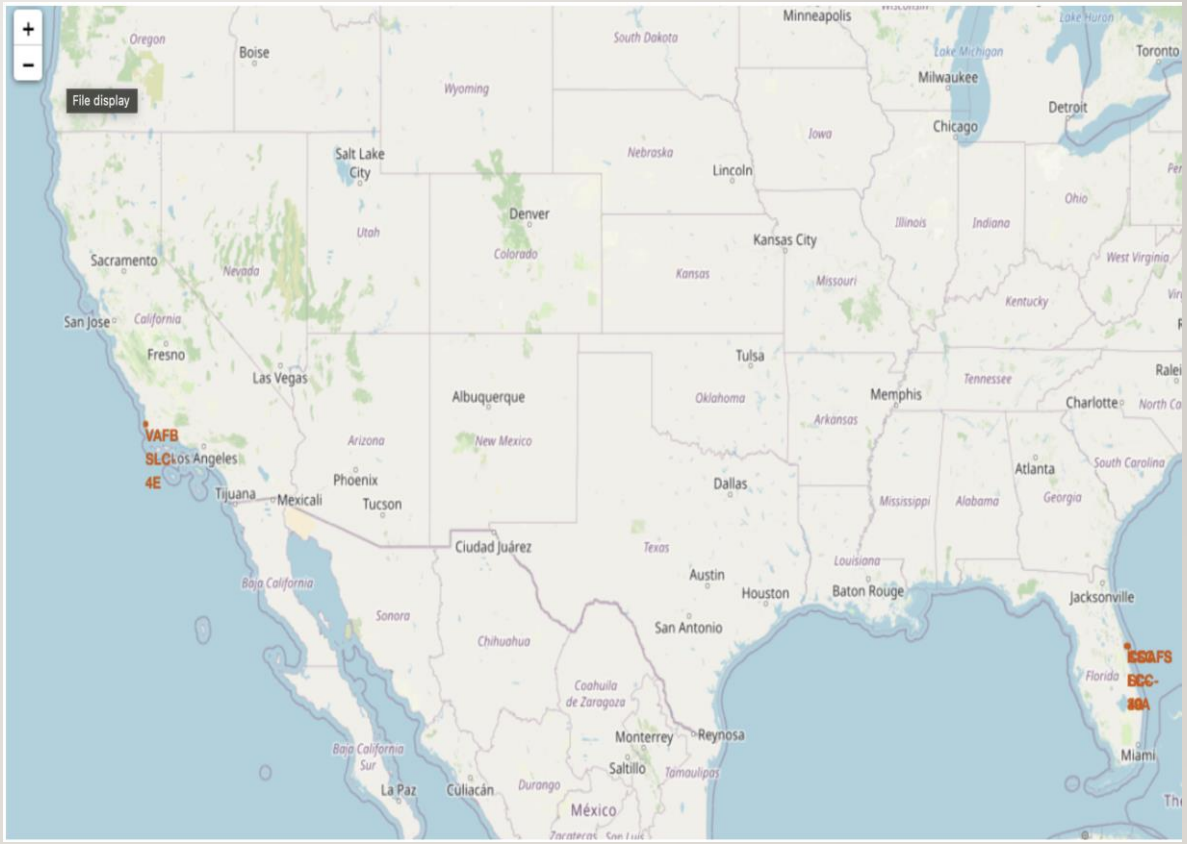
3]:

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

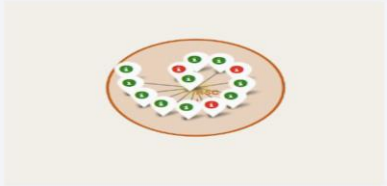
File display

INTERACTIVE MAP WITH FOLIUM ALL LAUNCH SITES

- **1. Are all launch sites in proximity to the Equator?**
  - Not all launch sites are located close to the Equator. While the Florida launch sites—Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40) and Kennedy Space Center Launch Complex 39A (KSC LC-39A)—are relatively closer to the Equator, Vandenberg Air Force Base (VAFB SLC-4E) is situated at a much higher latitude of approximately 34.63° North.
- **2. Are all launch sites in very close proximity to the coast?**
  - All identified launch sites are located near coastal regions. The Florida-based sites (CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A) are positioned along the Atlantic coastline, while Vandenberg Air Force Base (VAFB SLC-4E) is located near the Pacific coast in California.



Vandenberg Space Launch Complex 4 (CA)
VAFB SLC-4E



Kennedy Space Center (FL)
KSC LC 39A

Cape Canaveral (FL)
CCAFS-LC40

Cape Canaveral (FL)
CCAFS-SLC40

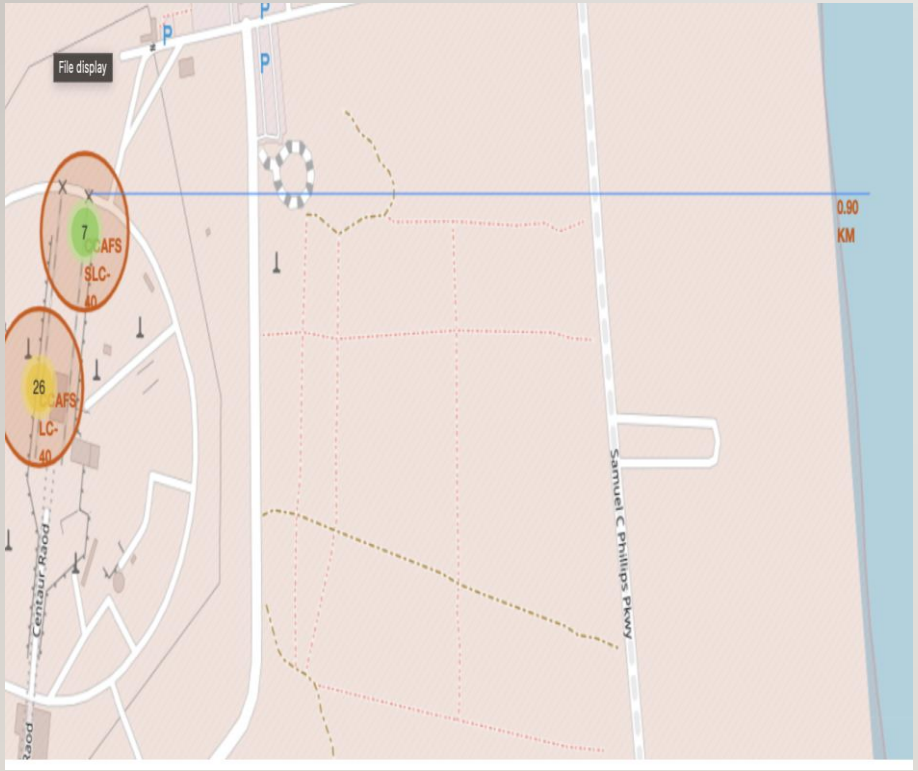
Launch Site	class	
CCAFS LC-40	0	19
	1	7
CCAFS SLC-40	0	4
	1	3
KSC LC-39A	0	3
	1	10
VAFB SLC-4E	0	6
	1	4

Table: Synthesis of launches outcomes
Class 0= failure
Class 1= success

INTERACTIVE MAP WITH FOLIUM
MARK THE SUCCESS/FAILED LAUNCHES FOR EACH SITE ON THE MAP

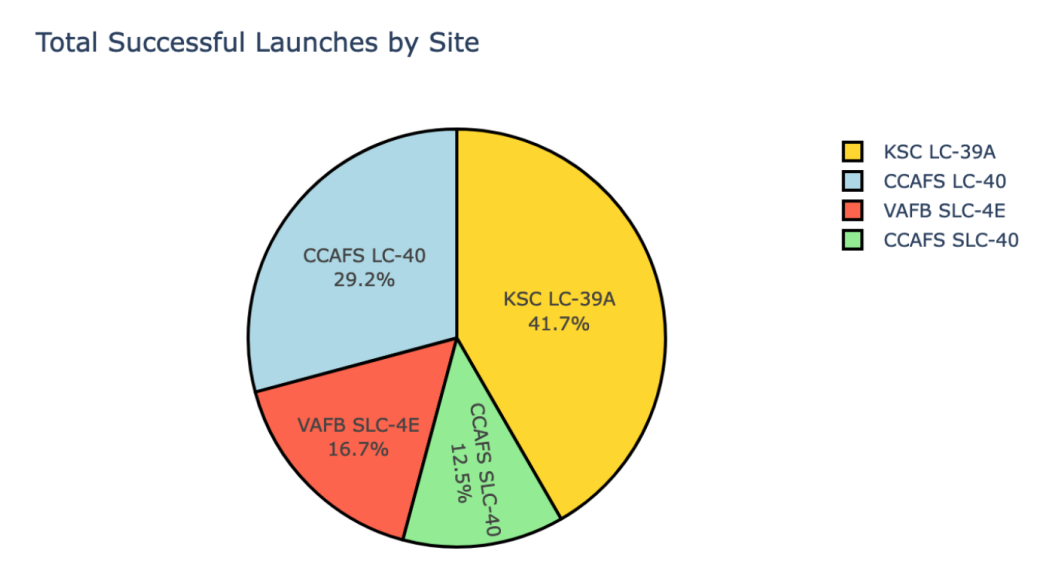
INTERACTIVE MAP WITH FOLIUM

CALCULATE THE DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES



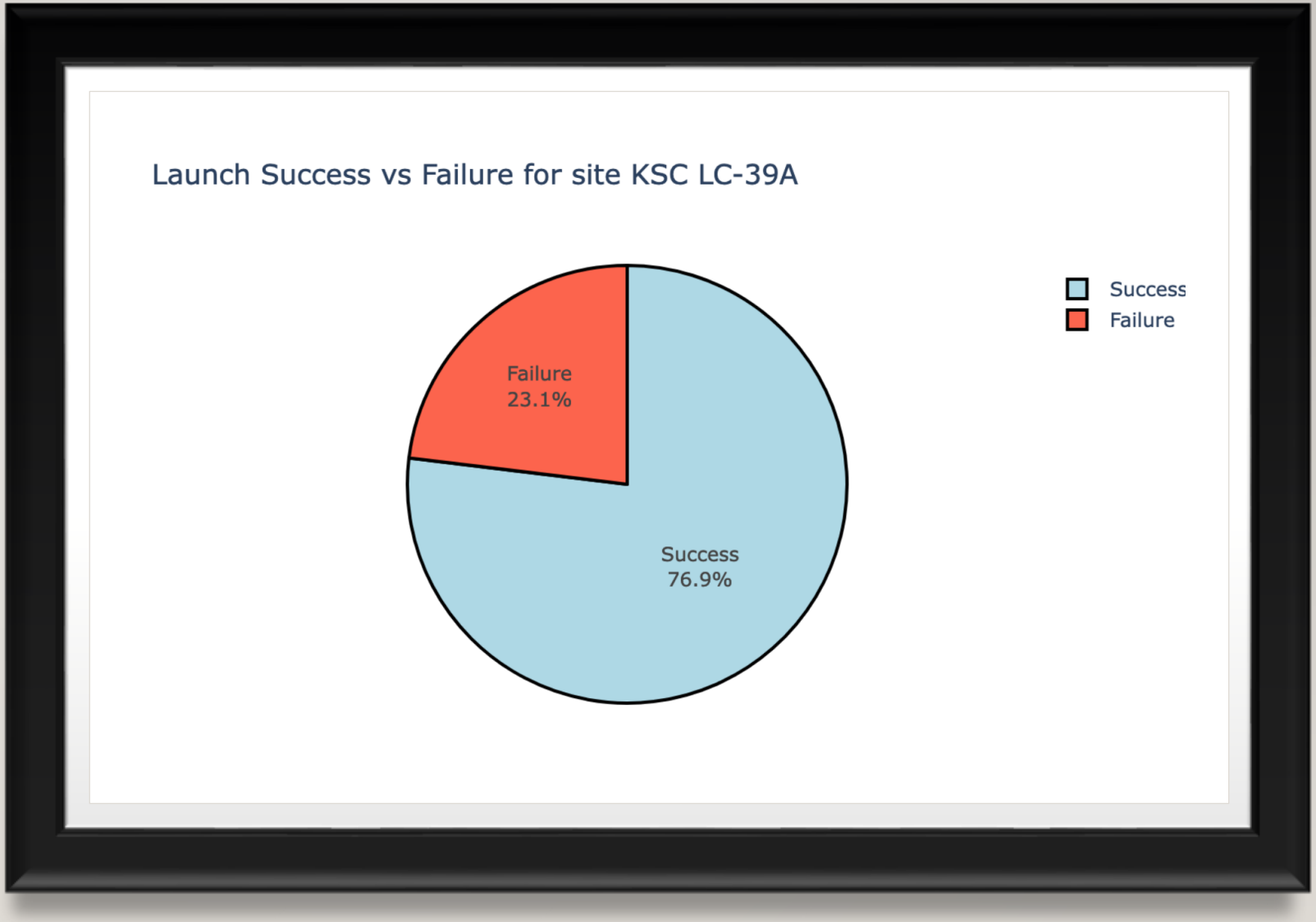
- This plot visually represents the distance between the CCAFS SLC-40 launch site and the nearest coastline. The calculated distance is approximately 0.51 kilometers, as indicated by the marker. The added PolyLine illustrates the straight-line distance, emphasizing the close proximity of the launch site to the coast. Such coastal positioning is typical for launch sites, as it supports over-water flight trajectories and safe booster recovery operations while minimizing risk to populated areas.

BUILD A DASHBOARD WITH PLOTLY DASH LAUNCH SUCCESS COUNT FOR ALL SITES

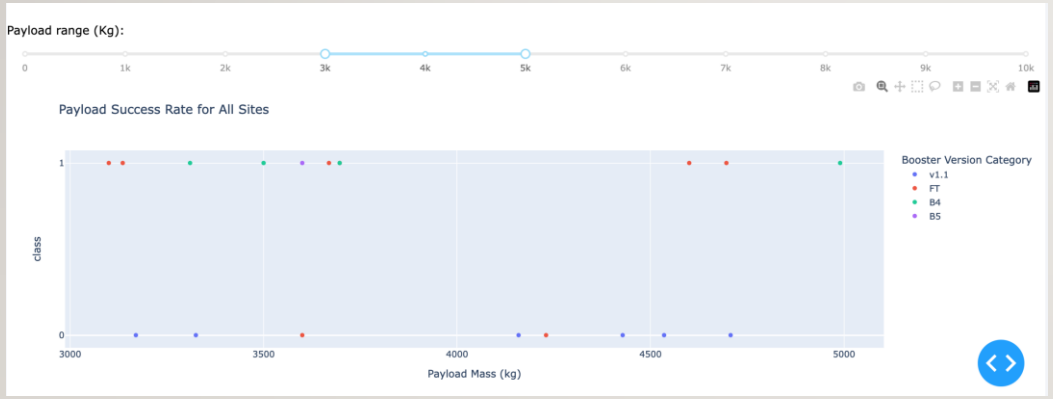


- KSC LC-39A experienced the highest proportion of successful landings, followed by CCAFS LC-40.
- VAFB SLC-4E and CCAFS SLC-40 the lowest.

BUILD A DASHBOARD WITH PLOTLY DASH LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

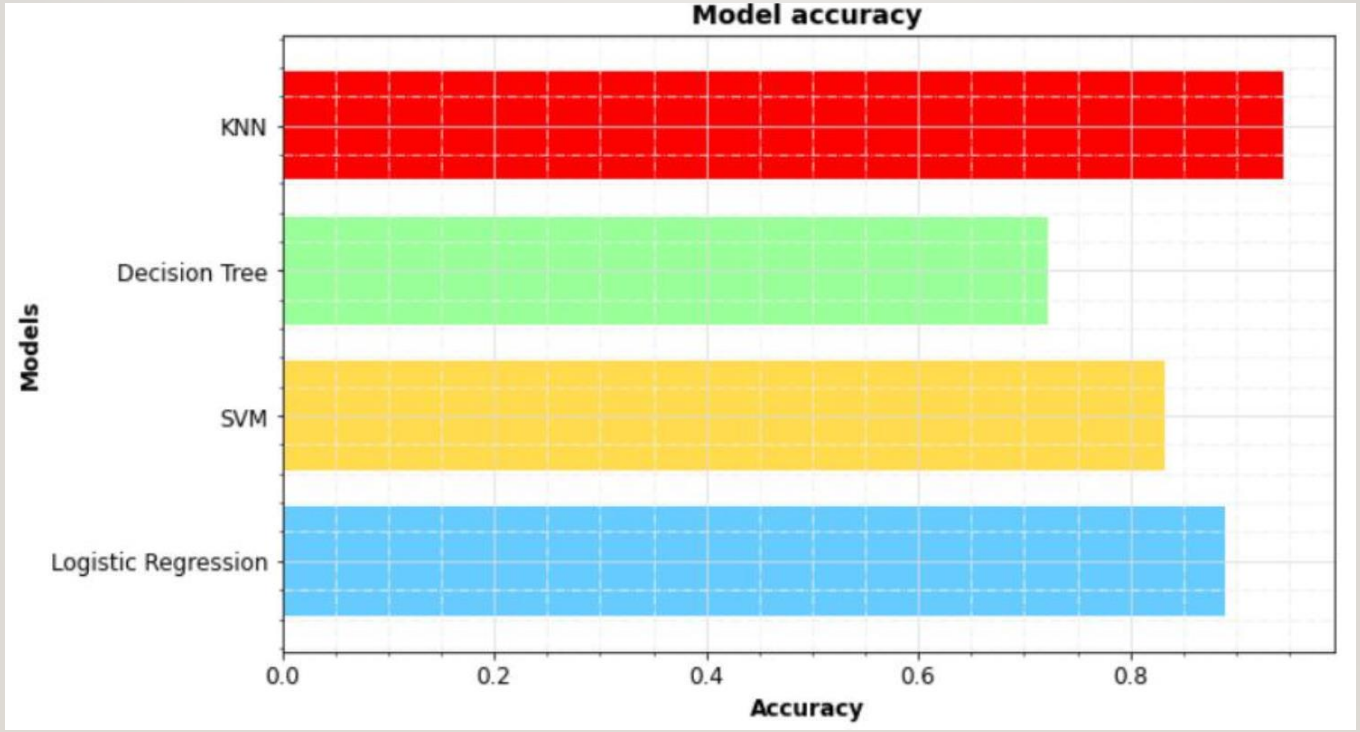


BUILD A DASHBOARD WITH PLOTLY DASH PAYLOAD

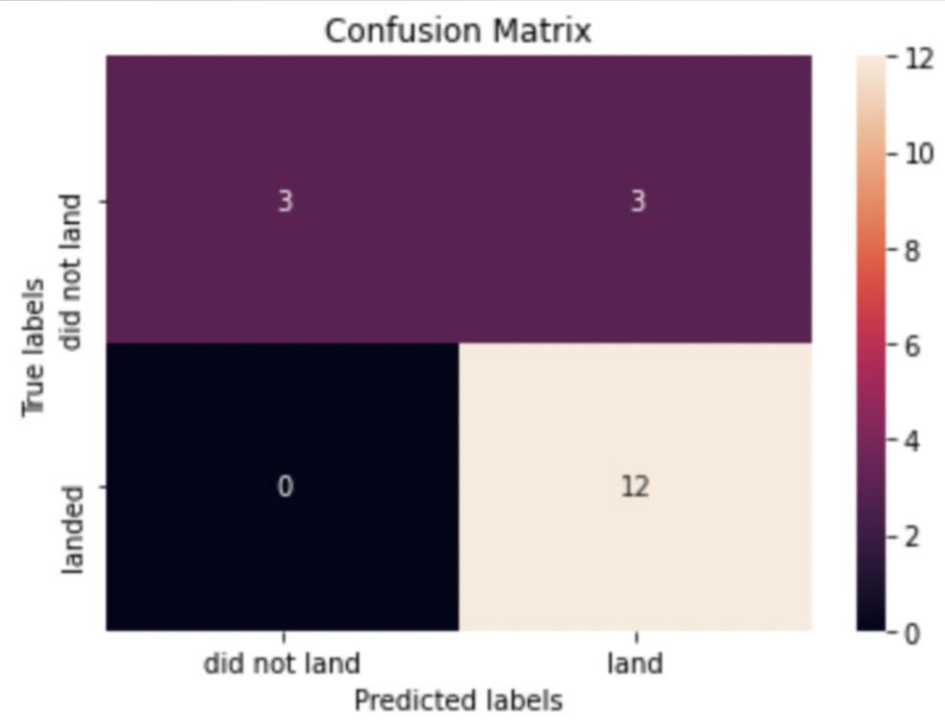


- V1.0 and v1.1 are early launchers with low reliability.
- Landing legs, were pioneered on the Falcon 9 v1.1 version, but that version never landed intact.
- They were phased out in 2015.
- FT: “Full Thrust” is the next generation and has the highest success rate for payload mass under 6 tons. Including with
- “drone landing” (see details in next slide).
- Many FT flights are done with reused launchers. And show good reliability.
- Heavy payload are “high risk”.

CLASSIFICATION ACCURACY



CONFUSION MATRIX



CONCLUSION

- Landing success rates show a clear upward trend over time across all analyzed factors, indicating continuous operational improvements and technological advancements. Analysis by orbit type reveals that certain orbits—such as ES-L1, SSO, HEO, and GEO—consistently achieve higher landing success rates. Launch site also emerged as a strong predictive factor, with KSC LC-39A demonstrating the highest performance, followed closely by CCAFS LC-40. Among the machine learning models evaluated, several achieved acceptable predictive performance; however, the **Decision Tree Classifier** delivered the best overall results, exhibiting high accuracy, precision, and recall in predicting landing outcomes.