Space Race: Determining SpaceX Rocket Landing Success with ML

Sami Khalayli

Outline

Executive Summary

This project attempts to explore factors that contribute to successful rocket launches.

Methodologies

- **Collect** data using SpaceX's API and from Wikipedia with BeautifulSoup
- **Wrangle** data using Pandas.
- **Perform** Exploratory data analysis (EDA) by visualization and SQL
- **Create** interactive visual analytics using Folium and Plotly Dash
- **Perform** predictive analysis using classification models

Results

Exploratory Data Analysis:

- Success rates vary by site, with some reaching up to 77%.
- No clear relationship between payload mass and launch success.
- Specific orbits (ES-L1, GEO, HEO, SSO) achieved 100% success.

Location Strategy:

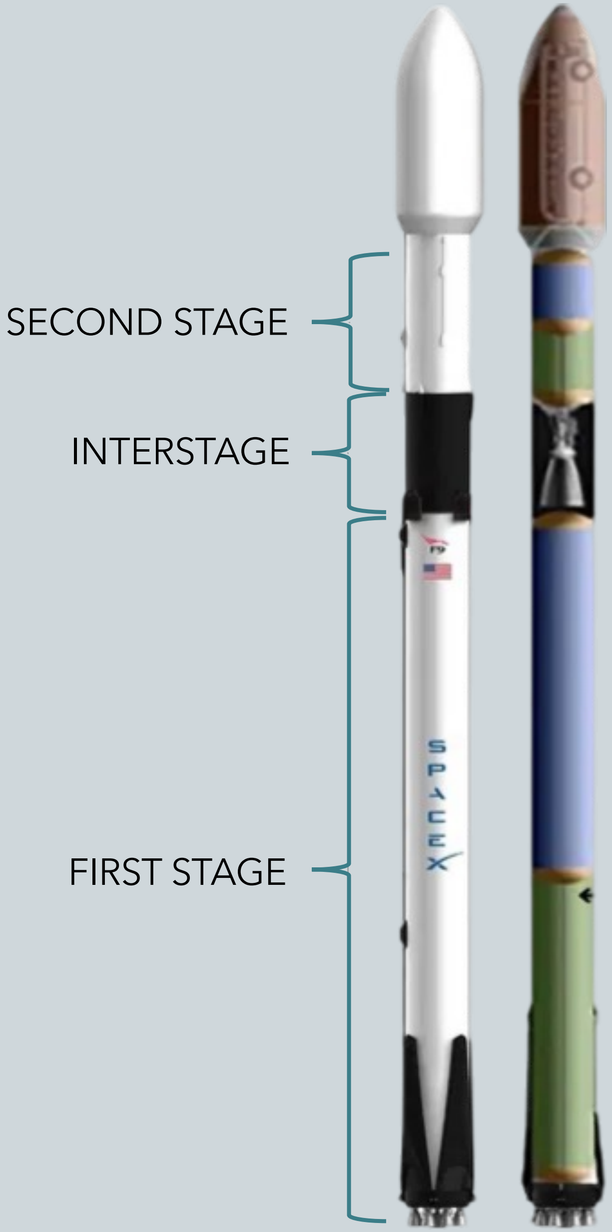
- Launch sites strategically near coasts for safety benefits.
- Eastern sites benefit from Earth's rotation to launch heavier payloads.

Predictive Analysis:

- KNN, Logistic Regression, and SVM showed high accuracies.
- Decision trees overfit and performed poorly on new data.

Introduction

- We are now in a space age where companies are racing to provide affordable space flight.
- However, what is required for a successful rocket launch?
- This project aims to explore the factors that contribute to a successful first stage landing using ML techniques that can be used to compete with established companies such as SpaceX.
- By exploring the probability that the first stage of a SpaceX Falcon9 rocket will land, we could infer the launch cost that will allow new companies to bid against SpaceX.



SpaceX. "Falcon 9 v1.2 or Full Thrust - Block 5 Tech Specs." [Wevolver]

Section 1

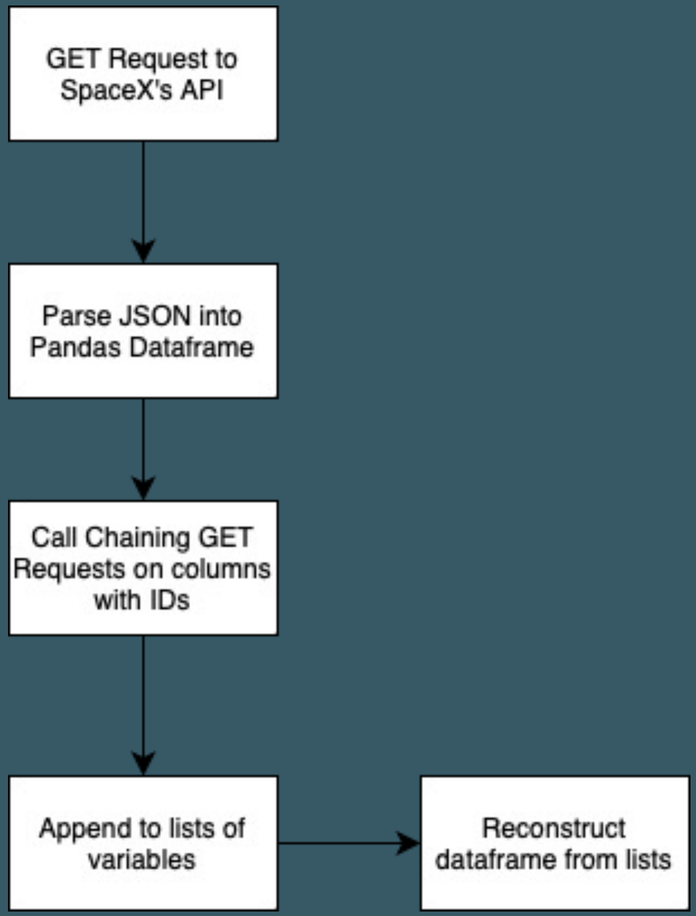
Methodology

Methodology

- Data Collection
  - Datasets were collected using SpaceX's API and webscraping Wikipedia with BeautifulSoup.
- Data Wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Create interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

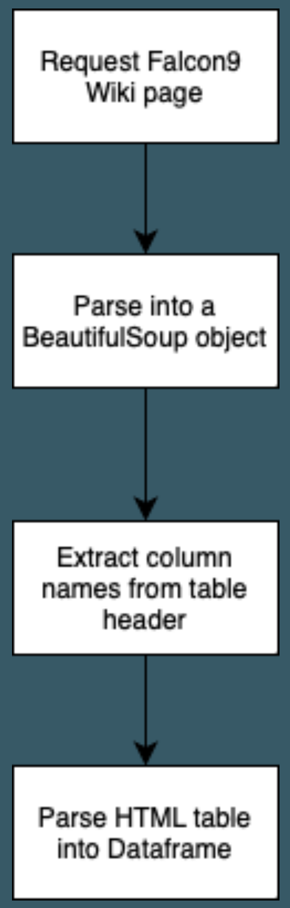
Data Collection – SpaceX API

- Rocket launch data was imported through the API.
- I chained requests to gather the rest of the rocket and payload data using their IDs.
- After the dataframe was reconstructed, missing values of the landing pad were replaced with their mean.
- Link to notebook



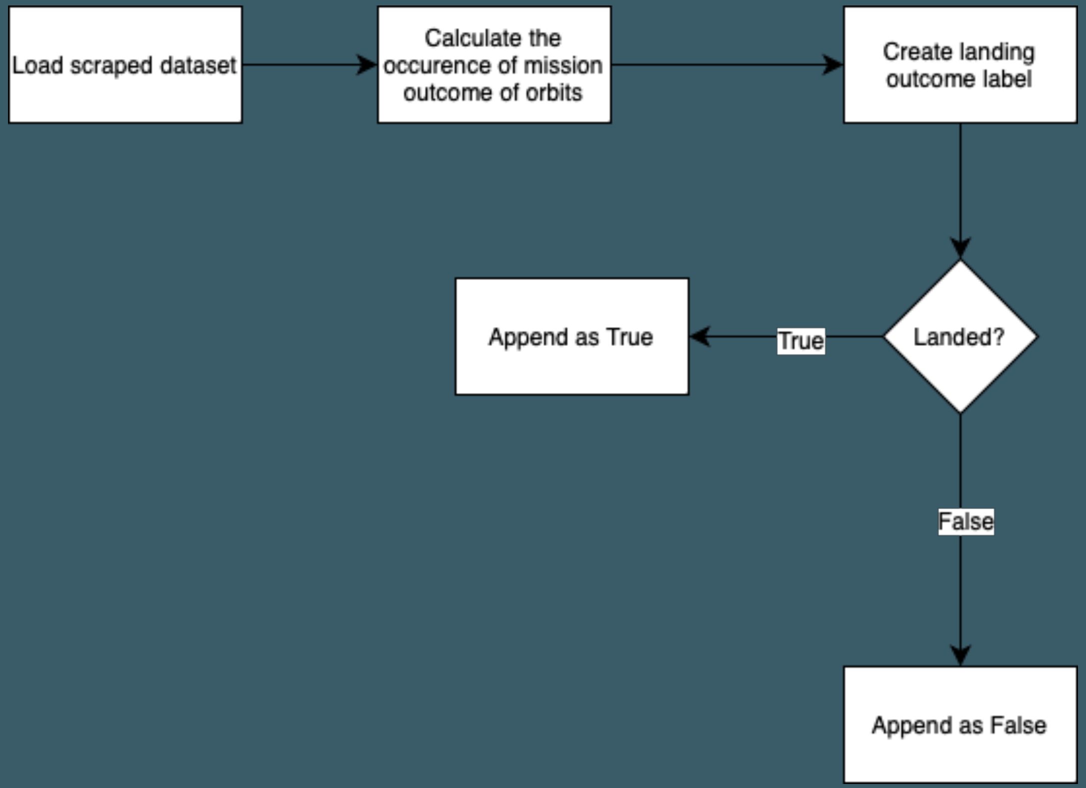
Data Collection - Scraping

- I extracted Falcon9 launch records with BeautifulSoup.
- Parsed the HTML table into a dataframe.
- Link to notebook



Data Wrangling

- Outcome labels of each rocket were constructed through the 8 outcome labels.
- The 8 landing outcome labels which specify how the rocket landed were simplified into True/False.
- Link to notebook



EDA with Data Visualization

I plotted different types of charts to visualize relationships between different variables and features such as:

- **Scatter** plot with Outcome as hue:
  - Flight Number & Launch Site
  - Launch Site & Payload Mass
  - Flight Number & Orbit
  - Payload & Orbit
- **Bar** chart:
  - Success rate of each orbit
- **Line** chart:
  - Launch success yearly trend

Then, I created dummy variables and one hot encoded categorical columns to prepare for future ML steps.

- Link to notebook

EDA with SQL

I began the EDA with the following steps:

1. Create table from dataframe using df.to_sql() method and **remove** blank rows.

2. Perform SQL queries to **display** numerous records**:**

- Names of unique launch sites in the space mission
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

EDA with SQL

3. As well as SQL queries to **uncover new features**:

   - Date of the first successful landing outcome in ground pad
   - Names of successful drone ship boosters with a payload mass between 4000 & 6000 kg
   - Total number of successful and failure mission outcomes
   - Names of the booster versions which carried the maximum payload mass using a **subquery**.
   - Month names, failure landing outcomes on drone ship, booster versions, launch site for the months in year 2015.
   - Count of landing outcomes (Failure/Success) between the date 2010-06-04 and 2017-03-20, in descending order.

- Link to notebook

Build an Interactive Map with Folium

- The following Folium features to visualize launch sites to help answer proximity to railways/coastline/cities questions:
  - **Circles**
  - **Markers**
- Utilized **Marker Clusters** to visualize all launch records with the same coordinates.
  - Launch sites with **success**/**failure** were colored in **green**/**red**
- Added a **Mouse Position** to the map to easily find coordinates of any points of interest.
- Later, I drew a **polyline** and calculated the distance to the closest coastline point using the Haversine formula.
- Link to Notebook

Build a Dashboard with Plotly Dash

I built a Plotly Dash application to perform interactive visual analytics on SpaceX launch data in real-time.

The following charts help answer questions regarding the sites success rate and payload mass influence:

- Pie Total Success Launches by Site
    - Launch Site is selected with a **dropdown**.
- Scatter point Payload Mass & Success with Booster Version as hue
    - Payload Mass range is selected with a **slider**.
- Link to code

Predictive Analysis (Classification)

- Preprocessing Steps:
  - **Load** the dataset
  - **Create** Y label for Class column (success)
  - **Standardize** X using StandardScaler()
  - **Perform** an 80/20 Train/Test split
- Classification Algorithms utilized:
  - **Logistic Regression**
  - **Support Vector Machine**
  - **Decision Tree**
  - **K Nearest Neighbors**

- The best parameters were found using **GridSearchCV** exhaustive search method.
- Confusion matrices were plotted and the best algorithm was found using **np.argmax().**
- Link to notebook

Results

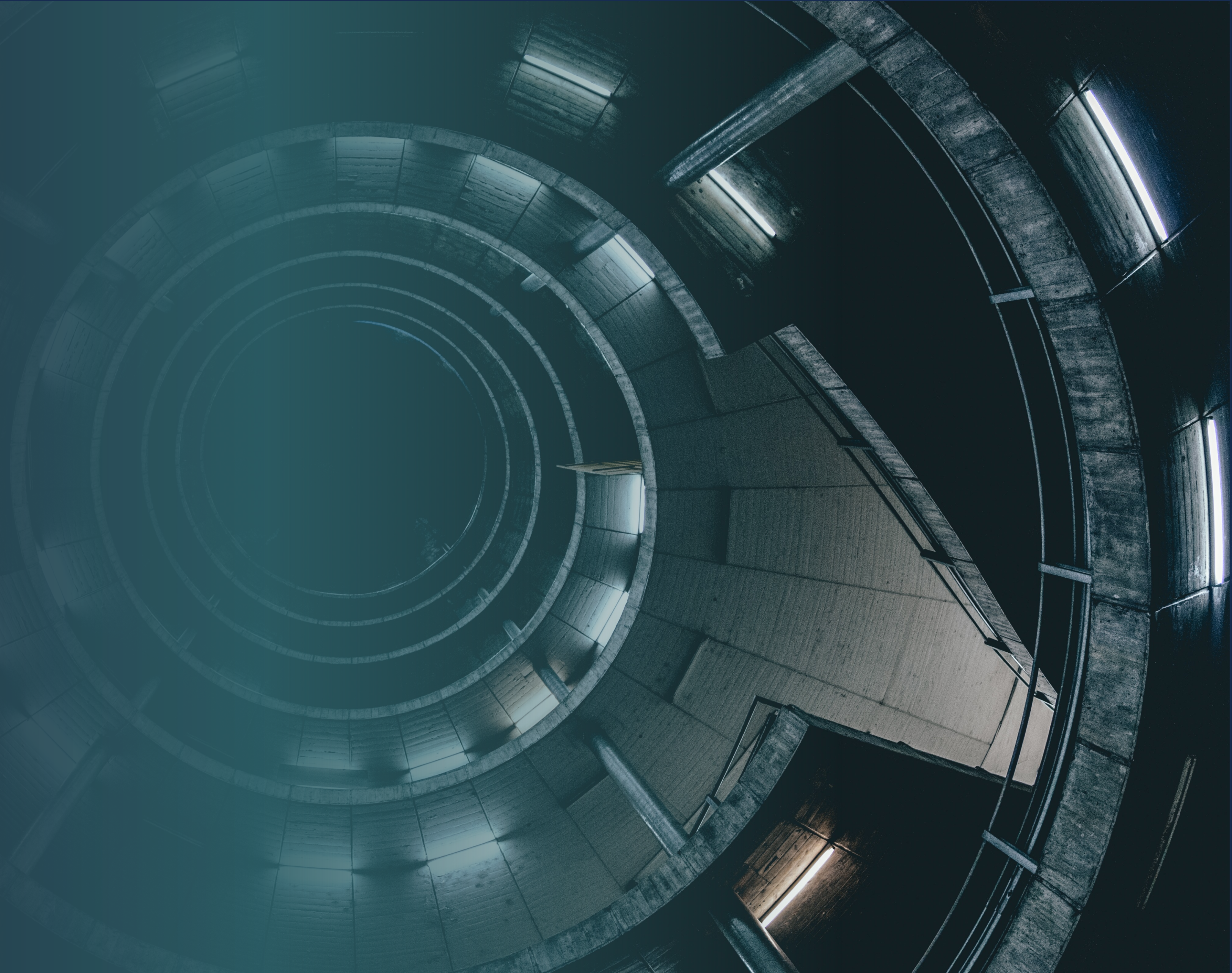
- **Exploratory data analysis**
  - The success rate of launch sites **improved** over time
  - Different launch sites have different success rates
    - CCAFS LC-40, has a success rate of **60%**
    - KSC LC-39A and VAFB SLC 4E has a success rate of **77%**
  - VAFB-SLC site didn't launch any rockets with heavy payload mass **greater than 10,000kg**.
  - There is **no clear relationship** between the payload mass and success outcome.
  - ES-l1,GEO,HEO,SSO orbits had a **100% launch success outcome**.

Results

- **Location Analysis with Folium**
  - VAFB-SLC site was used mainly for **polar orbits** which requires launching north-south direction.
  - The rest of the sites located on the coast on the East made use of **Earth's rotational speed** launching East hence the ability to launch with **greater payloads.**
  - All launch site locations were in **close proximity** to the coast and railways.
    - The decision to be close to the coastline is for safety reasons to increase the chance of the debris falling in the ocean instead of populated land.
- **Predictive analysis results**
- KNN, Logistic regression, SVMs yield the same accuracy whereas Decision trees overfit to training data and lacked in the final scoring.

Section 2

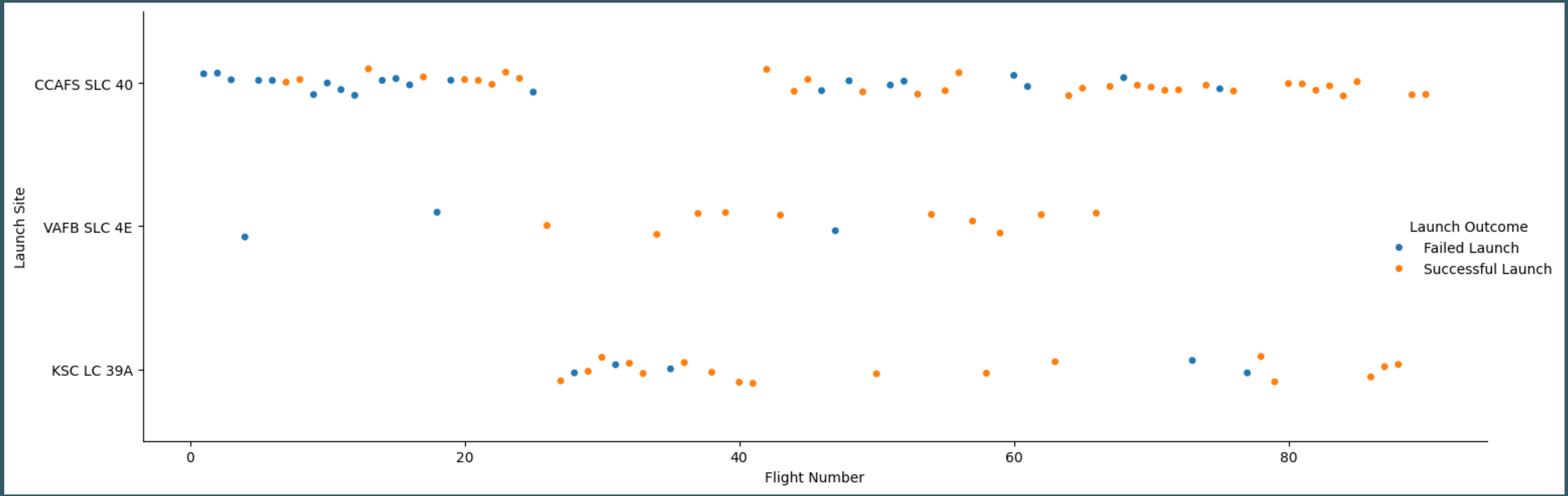
EDA Insights



Freepik/wirestock

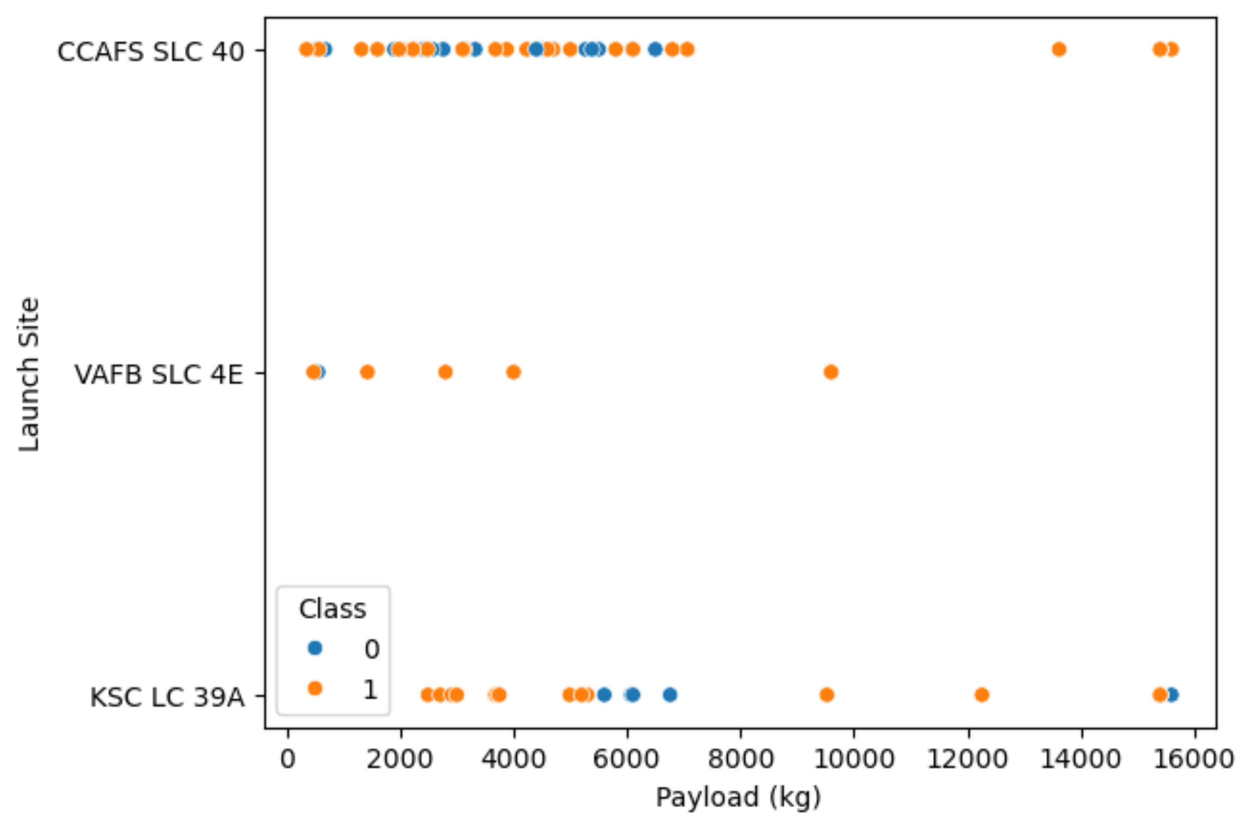
Flight Number vs. Launch Site

- Early flights were carried in **CCAFS SLC 40**
  - They had a very **low** success rate
- **VAFB SLC 4E** and **KSC LC 39A** have the most success rate.
- Majority of Launches were carried from **the CCAFS SLC 40** Launch Site.
- Success rate **increased** with flight numbers.



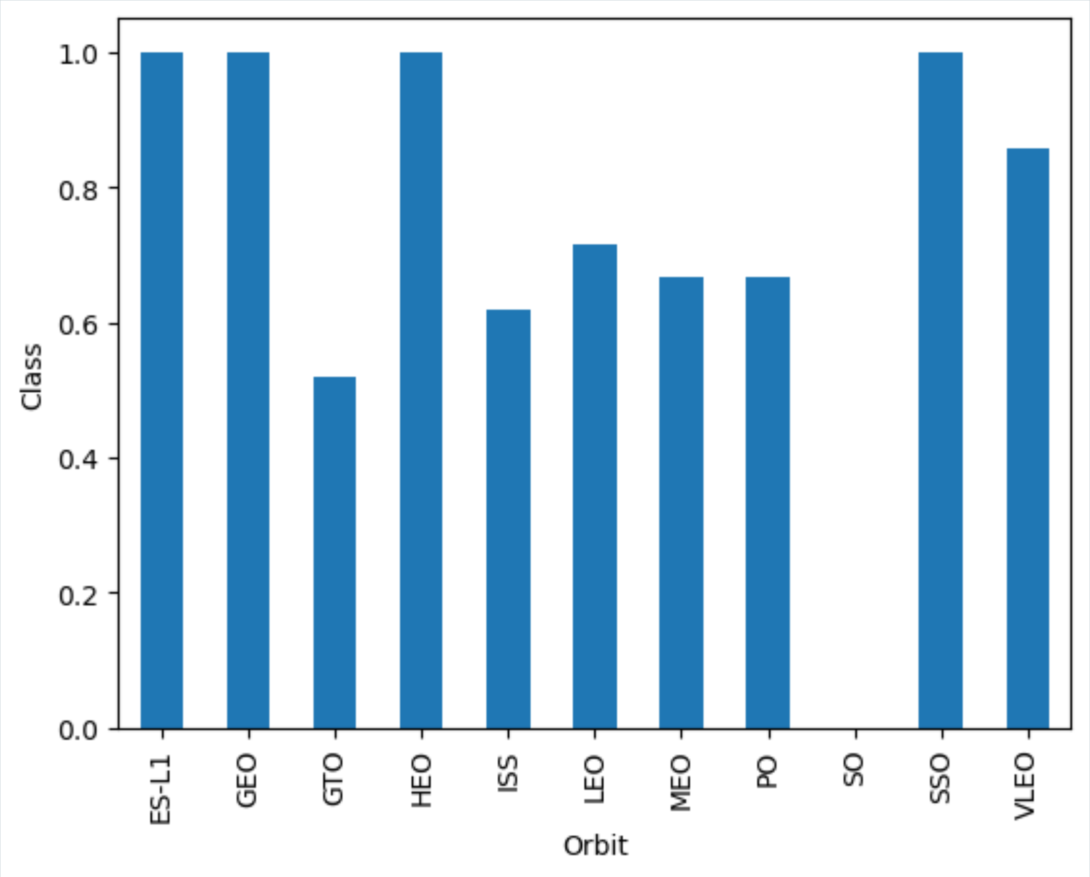
Payload vs. Launch Site

- There is **no clear relationship** between Payload mass and success rate.
- **CCAFS SLC 40** and **KSC LC 39A** are located on the East and they make use of the earth's rotational speed, hence they could launch payloads greater than 10,000 more efficiently.
- Most payloads at KSC between 5600 KG and 7000 KG **failed**.



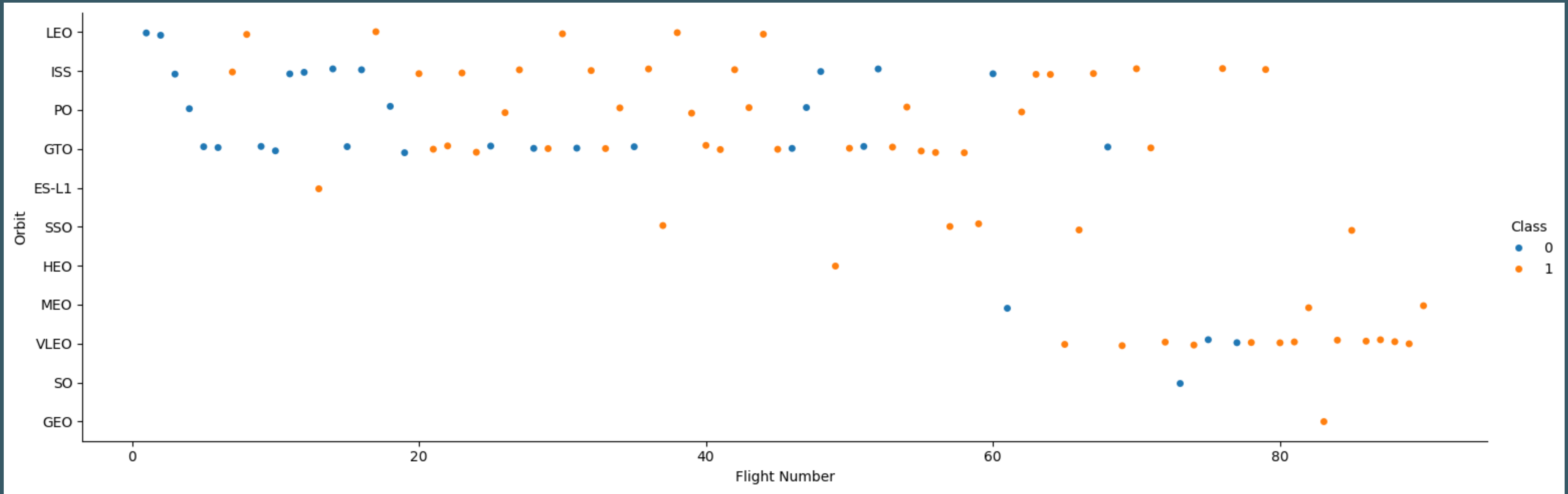
Success Rate vs. Orbit Type

- **ES-L1,GEO,HEO and SSO** all had 100% success rates.
- **GTO, ISS, LEO, MEO, PO, VLEO** had success rates between 50% and 90%
- **SO** had a 0% success rate.



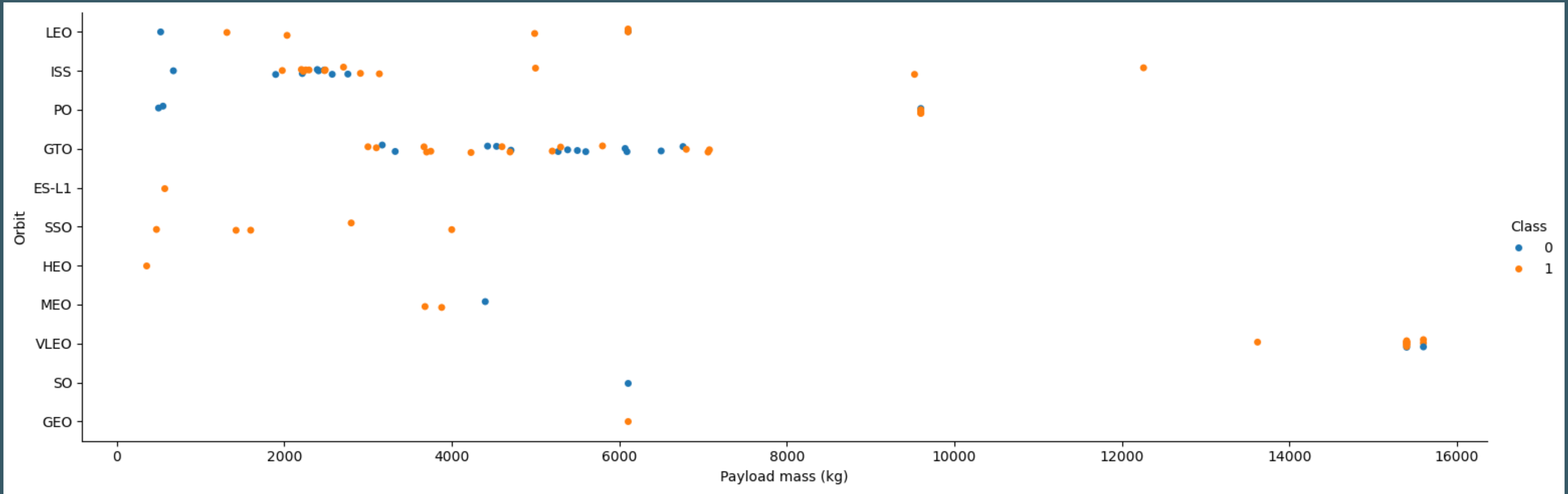
Flight Number vs. Orbit Type

- **VLEO** orbits were only achieved later.
- Majority of launches were in **GTO or ISS** orbits.
- **ES-L1, HEO, SO and GEO** orbit launches were only attempted once.



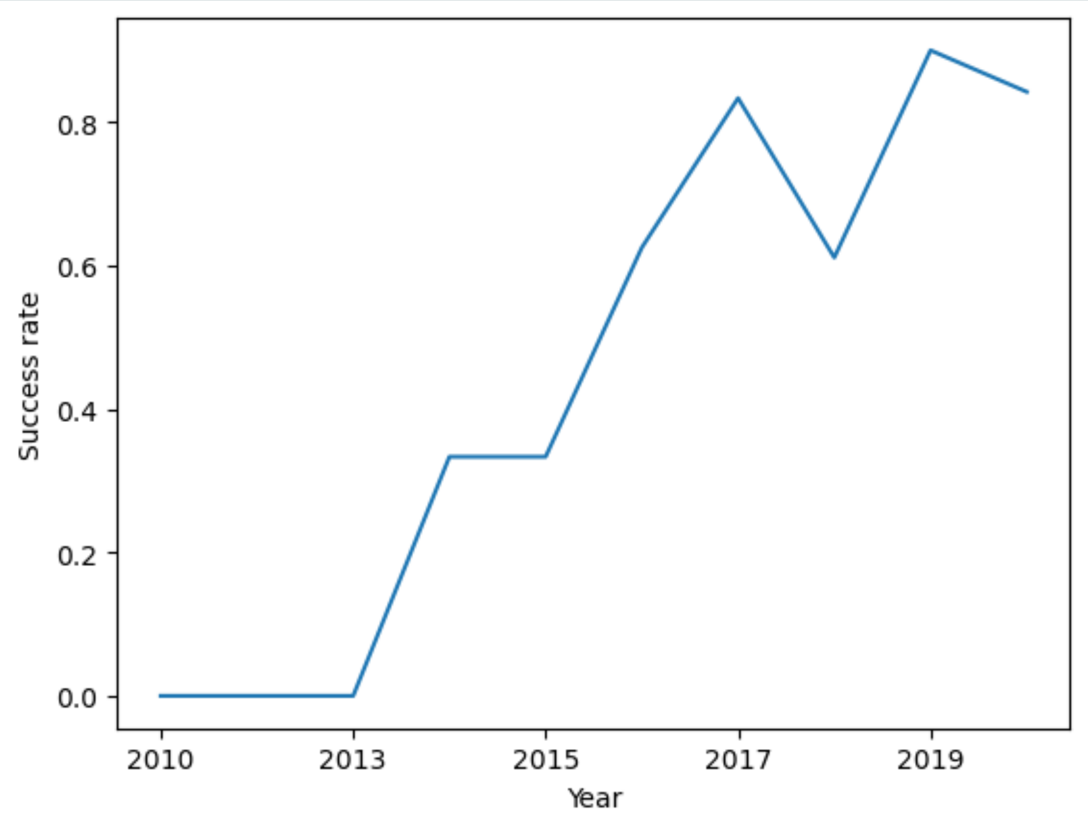
Payload vs. Orbit Type

- Launches in **VLEO** orbit carried the heaviest payloads greater than 13,000kg
- **GTO** payloads were exclusively between 3000 and 7500kg
- **ISS and GTO** typically carried a range of smaller payloads.
- Certain orbits launches hold different payload mass ranges



Launch Success Yearly Trend

- Launch success has drastically improved over the period of **9 years**
- Each year has progressed almost **20%** in success rate.



All Launch Site Names

Query: select distinct Launch_Site from SPACEXTABLE

Launch Sites:

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

Launch Site Names Begin with 'CCA'

Query: select * from SPACEXTABLE where Launch_Site like "CCA%" LIMIT 5

Result:

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

Query: select SUM(Payload_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)’

Total payload carried by boosters from **NASA**

- **45596 KG**

Average Payload Mass by F9 v1.1

Query: select AVG(Payload_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1’

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

- **2928.4 kg**

First Successful Ground Landing Date

Query: select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' Limit 1

The date of the **first successful landing outcome** on ground pad

- **2015-12-22**

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' AND Payload_MASS__KG_ BETWEEN 4000 AND 6000
- Booster Versions:
  - **F9 FT B1022**
  - **F9 FT B1026**
  - **F9 FT B1021.2**
  - **F9 FT B1031.2**

Total Number of Successful and Failure Mission Outcomes and Landing Outcomes

Queries:

- select count(*) as 'SUCCESS' from SPACEXTABLE where Landing_Outcome like 'Success%';
- select count(*) as 'FAIL' from SPACEXTABLE where Landing_Outcome NOT like 'Success%';
- select count(*) as 'SUCCESS' from SPACEXTABLE where Mission_Outcome like 'Success%';
- select count(*) as 'FAIL' from SPACEXTABLE where Mission_Outcome NOT like 'Success%';

Mission Outcomes:

Success: 100 Fail: 1

Landing Outcomes:

Success: 61 Fail: 40

Boosters Carried Maximum Payload

Query: select Booster_Version from SPACEXTABLE WHERE Payload_MASS__KG_ = (select max(Payload_MASS__KG_) from SPACEXTABLE)

Booster versions

- **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 B1060.2**
- **F9 B5 B1058.3**
- **F9 B5 B1051.6**
- **F9 B5 B1060.3**
- **F9 B5 B1049.7**
- **F9 B5 B1049.5**

2015 Launch Records

Query: select substr(Date,6,2), Launch_Site from SPACEXTABLE WHERE substr(Date,0,5) = '2015' AND Landing_Outcome = 'Failure (drone ship)'

Month and Launch site of Drone ship failures in 2015:

- **01 CCAFS LC-40**
- **04 CCAFS LC-40**

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

Query: select count(*),Landing_Outcome from SPACEXTABLE WHERE DATE between '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC

Result:

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

Section 3

Launch Sites Proximities Analysis

Freepik

Launch Sites

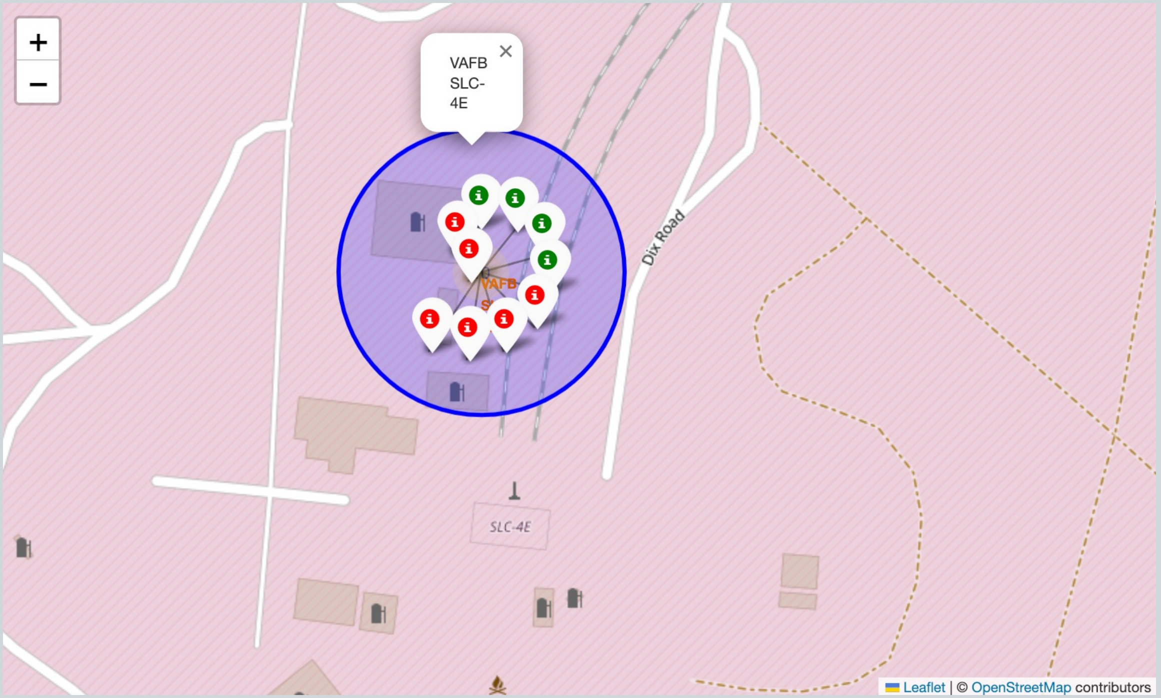
- Sites are located on the **East** and **West**
- Only **VAFB SLC-4E** is located in the West

Launch Outcomes

Launch Outcomes of sites were color labelled

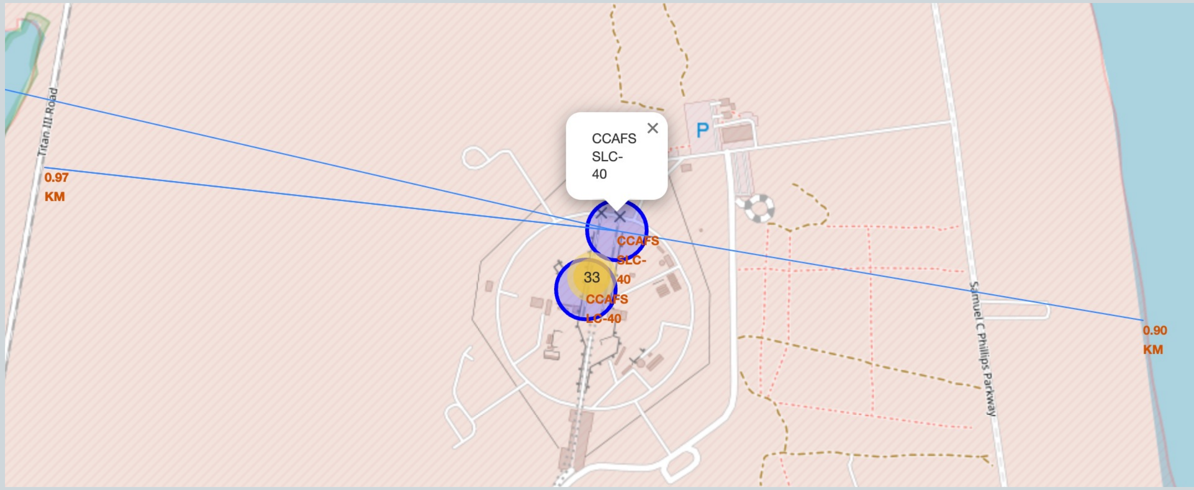
- Green for **success**
- Red for **failure**

- Majority of VAFB site outcomes were **failures.**



Distance to Proximities

- Distance to coastline**: 0.90 KM**
- Distance to rail **0.97 KM**
- Distance to city: **23.26 KM**



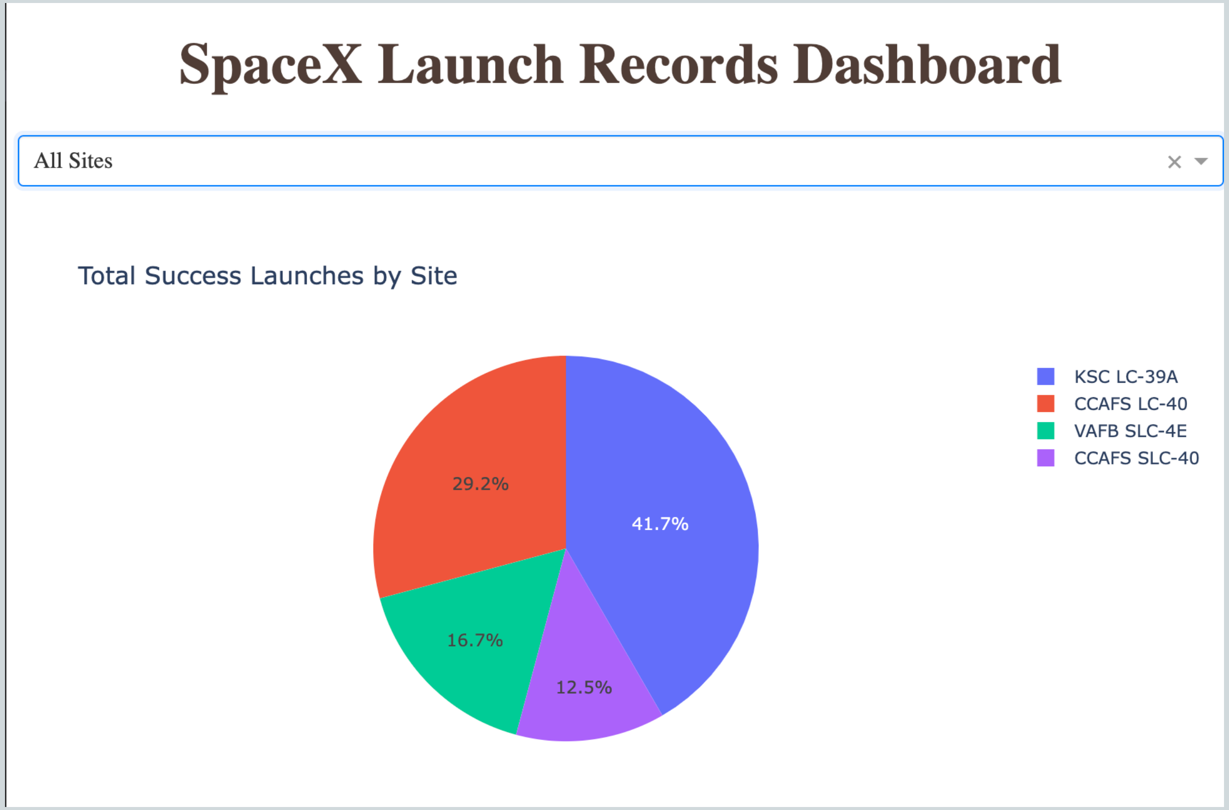
Section 4

Dashboard with Dash

Freepik

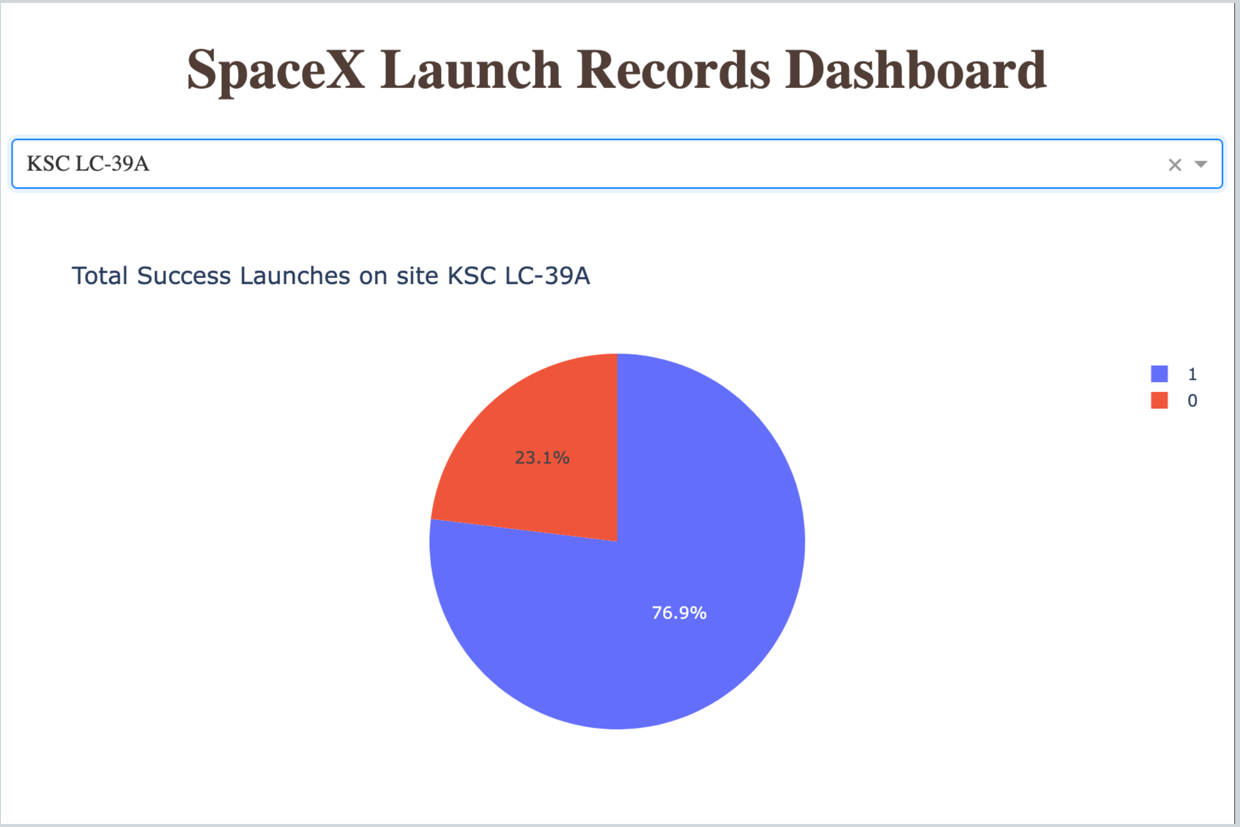
Total Success Launches

- Most successful rockets launched from **KSC LC-39A** with a 41.7% compared to other launch sites.
- The least success contributor was **CCAFS SLC-40** with 12.5% share.



Success Launches

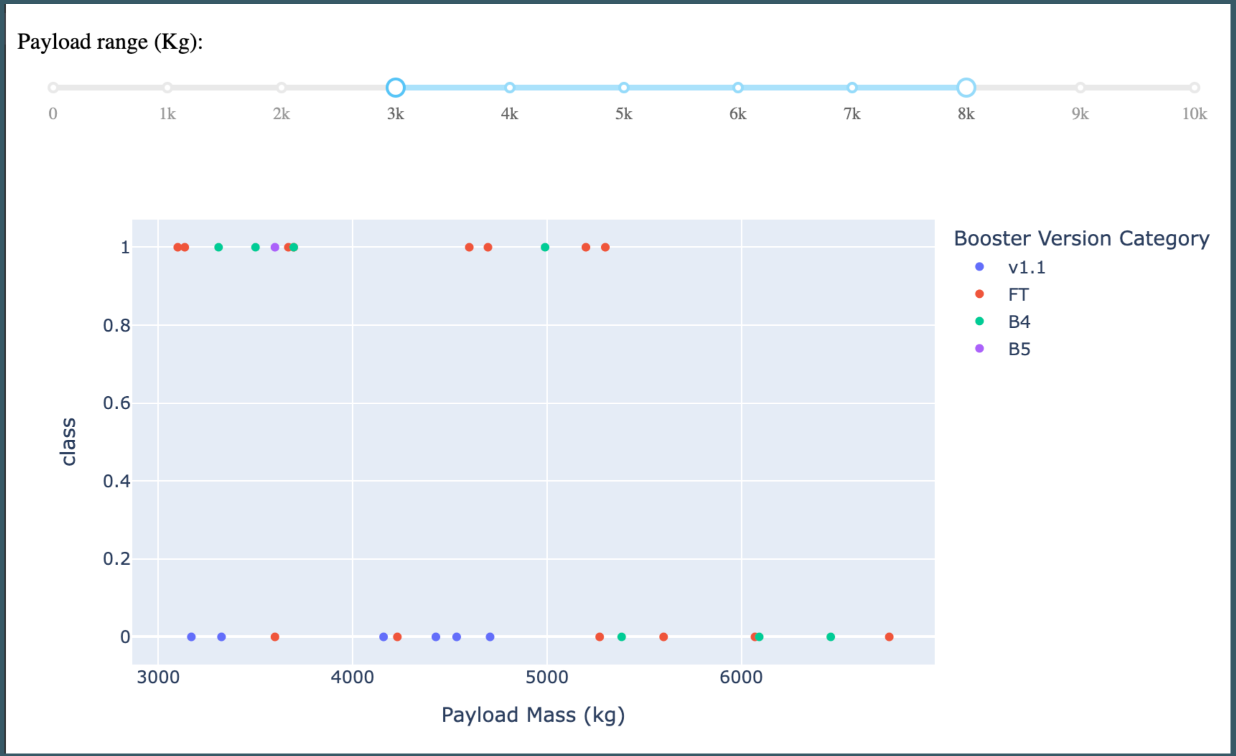
- 76.9% of Launches on site **KSC LC-39A** were successful.



Payload vs. Launch Outcome

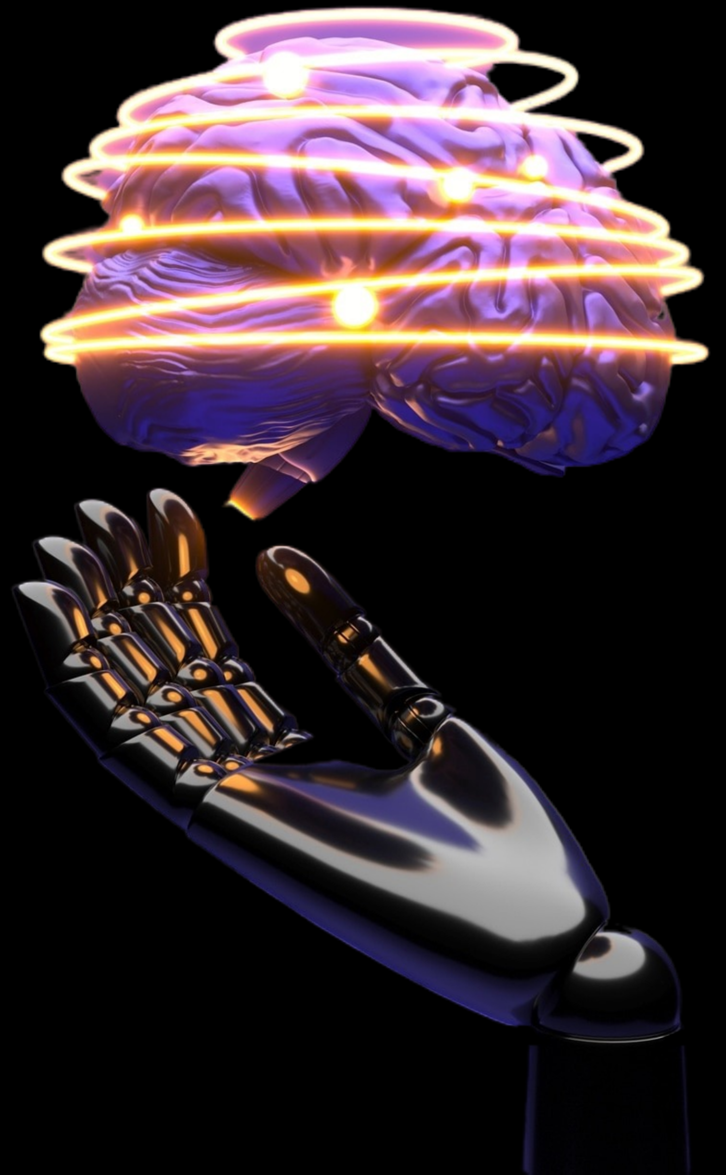
In the range between **3000 and 8000 kg:**

- Launches were **more failures** than successful
- All **v1.1** Booster version launches failed.
- All **B5** booster versions were a success.



Section 5

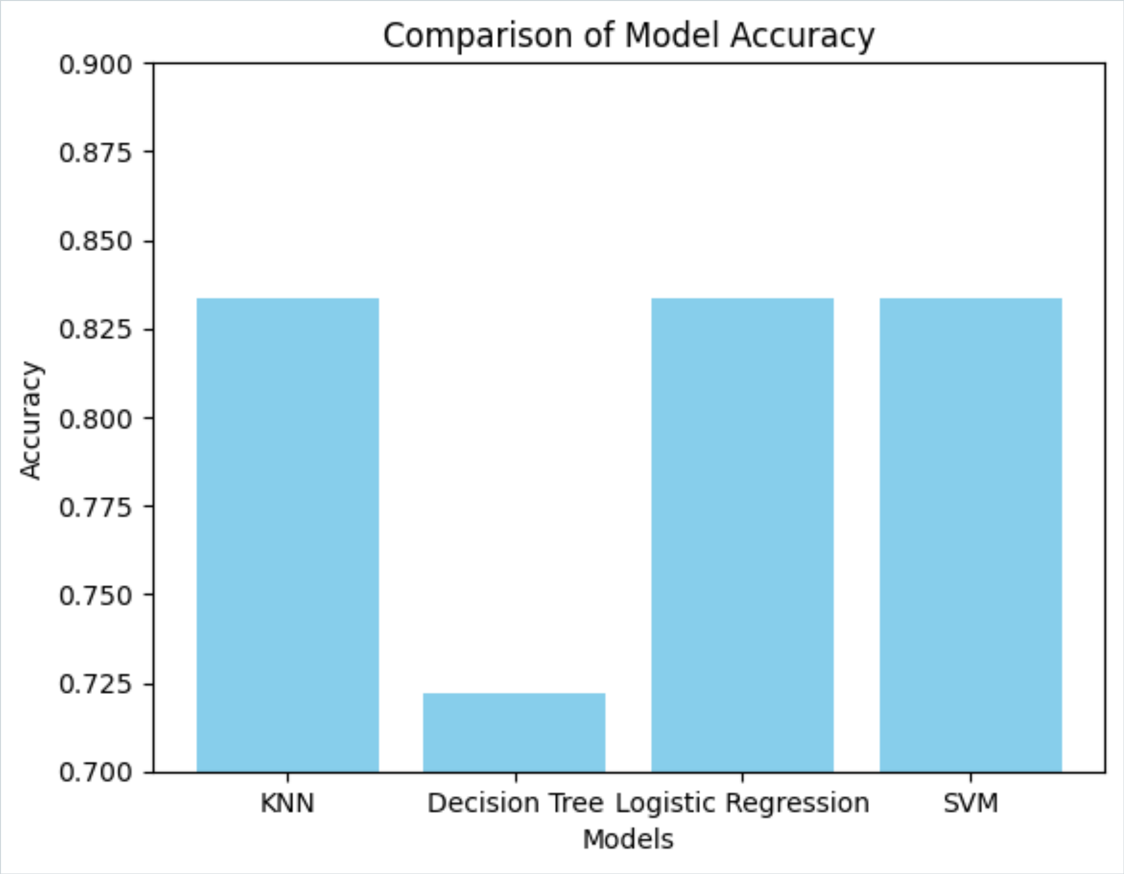
Predictive Analysis



Freepikk

Classification Accuracy

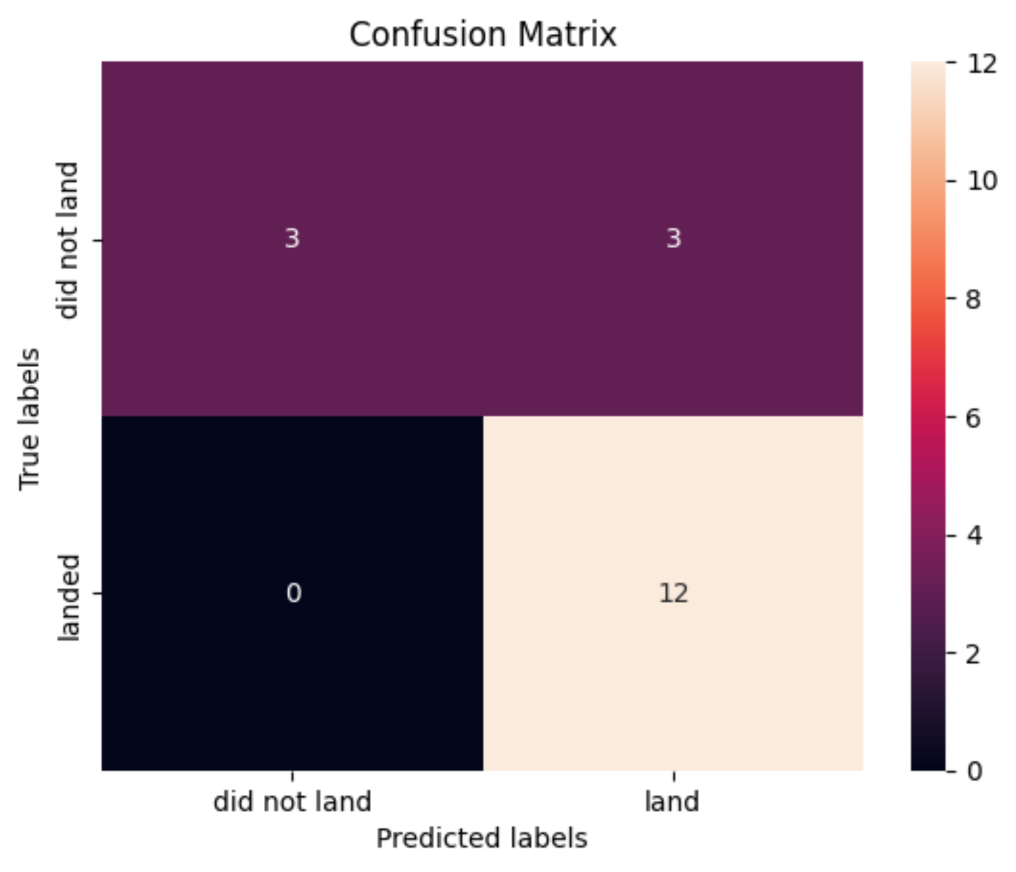
- **KNN, SVM and Logistic Regression** had the highest accuracies and are identical.
- Decision tree **fell behind** at 72.2% accuracy on new unseen data, despite its very high training accuracy.



Confusion Matrix

- **KNN, SVM and Logistic Regression** also had identical confusion matrices.

- They had 0 False Negatives, with **100% recall.**



Conclusions

- **Location Efficiency**: Launch sites are situated close to the coast and railways to boost safety and logistics. Moreover, sites on the east coast benefit from Earth’s rotational speed for to launch heavier payloads more efficiently.
- **Predictive Model Performance**: While KNN, SVM, and Logistic Regression demonstrate high predictive accuracy, Decision Trees could not generalize. Choosing the correct model that is robust against overfitting is important.
- **Orbital Considerations**: The strong success in orbits like GEO and HEO suggests that there are certain technical and operational strategies that are more effective for these missions.
- **Trend of Success Improvement**: There is a clear trend of increasing success rates over the years, which is the result of the evolving technology and space knowledge.
- **Reusability Cost:** There are huge cost savings in reusing the first stage. Since 60.4% of the first stages land back again, there is a clear potential economical benefits in increasing the reliability of the first stage.