IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Rohith Vijaey M C>
<07-12-2025>

Outline

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

Executive Summary

- **Summary of methodologies**
  - ❖ Data collection
  - ❖ Data wrangling
  - ❖ EDA with Data visualization
  - ❖ building an interactive map with Folium
  - ❖ Building a Dashboard with Plotly
  - ❖ Building a Classification model
- **Summary of all results**
  - ❖ Exploratory Data Analysis results
  - ❖ Interactive analytics demo
  - ❖ Predictive analysis results

Introduction

- Project background and context

  SpaceX has become the leading force in the commercial space era by drastically reducing the cost of space travel. The company lists Falcon 9 launches at around 62 million dollars on its website, while other launch providers often charge more than 165 million dollars. A major reason for this price advantage is SpaceX's ability to reuse the rocket's first stage.Because of this, being able to predict whether the first stage will successfully land can help estimate the overall launch cost. Using publicly available data and machine learning techniques, our goal is to forecast the likelihood that SpaceX will recover the first stage.

- Problems you want to find answers
  - How do factors like payload mass, launch location, flight count, and orbital destination influence the likelihood of a successful first-stage landing?
  - Does the frequency of successful landings show an upward trend over time?
  - which machine-learning algorithm is most suitable for performing binary classification for this prediction task?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API -
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best result.

Data Collection

- The data for this project was gathered using a combination of SpaceX's REST API and web scraping from a launch table on SpaceX's Wikipedia page. Using both sources was necessary to compile a complete dataset and enable a more thorough analysis of the missions.
- The SpaceX REST API provided the following fields:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- The Wikipedia web scraping process supplied additional information, including:
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API

Requesting rocket launch data from Spacex API

Decoding the response content using .json and turning into a dataframe.

Requesting required information about the launches from SpaceX API

constructing data we have obtained from Spacex API into a dictionary

Creating a dataftrame from the dictionary

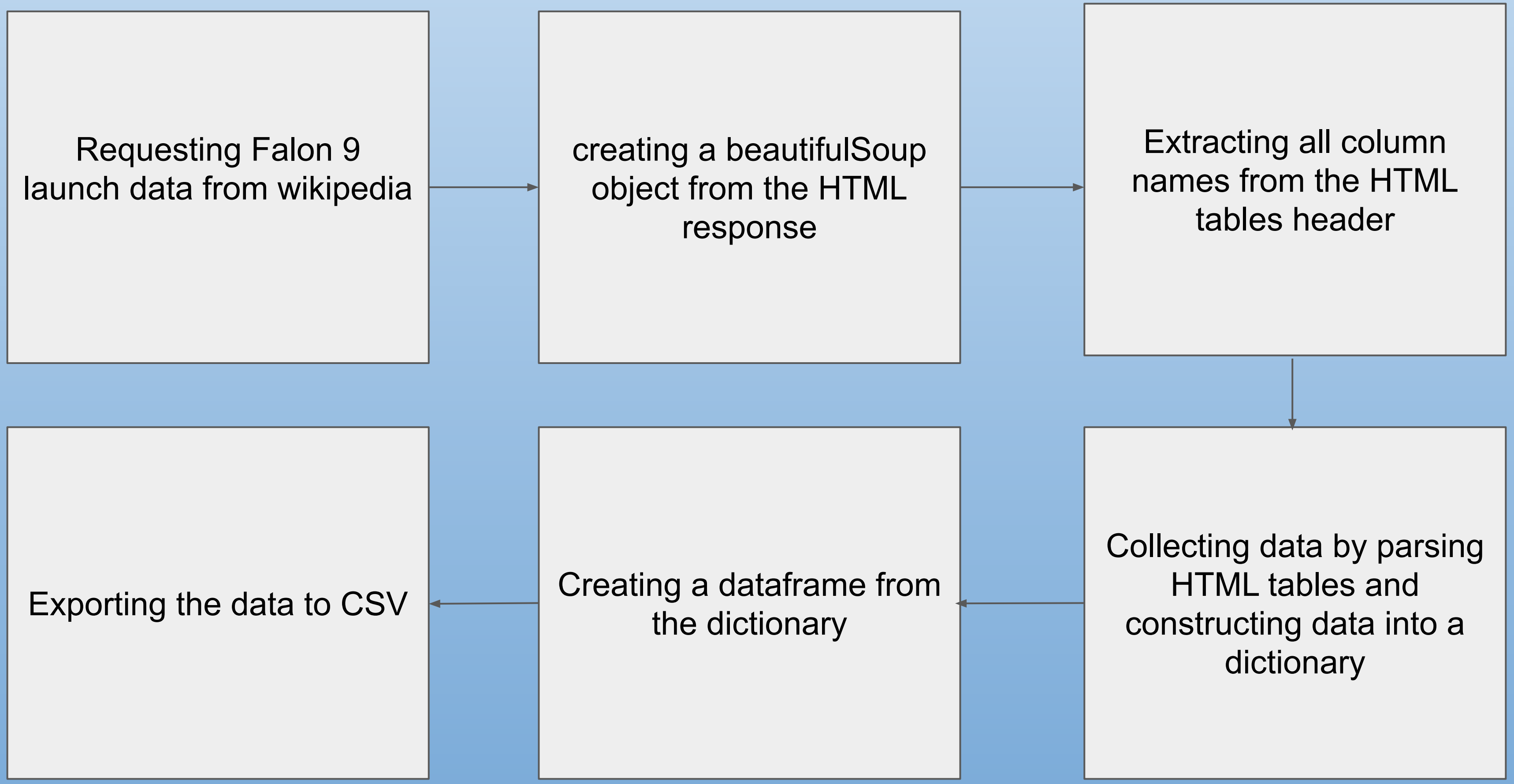
Filtering the dataframe to only include Falcon 9 launches

Replacing the missing values of payload column with mean value

Exporting the data to csv

GitHub URL : Spacex Data Collection API

Data Collection - Scraping



[GitHub URl : We Scraping]

Data Wrangling

A landing attempt could fail due to an accident. For example:

- **True Ocean**: booster successfully landed in a designated ocean region.
- **False Ocean**: booster failed to land in the ocean.
- **True RTLS**: booster successfully landed on a ground pad.
- **False RTLS**: booster failed to land on a ground pad.
- **True ASDS**: booster successfully landed on a drone ship.
- **False ASDS**: booster failed to land on a drone ship.

The landing outcomes are converted into training labels, where 1 indicates a successful booster landing and 0 indicates a failed landing.

Performing EDA and determining training labels

Calculating number of launches on each site

calculate the number and occurrence of mission outcome per orbit

create a landing outcome label from Outcome column

Exporting the data to CSV

GitHub URL : Data Wrangling

EDA with Data Visualization

- A series of charts were generated to explore different relationships in the dataset, including:
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Orbit Type vs Success Rate
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Yearly Success Rate Trend

- Scatter plots were used to visualize how two continuous variables relate to each other. When a clear pattern appears, it can indicate useful features for building machine-learning models.
- Bar charts were applied to compare values across distinct categories, helping highlight how each category differs in terms of the measured metric.
- Line charts were used to illustrate how performance or outcomes change over time, making them effective for analyzing long-term trends or time-series behavior.

GItHub URL : Data Visulaization

EDA with SQL

SQL queries were executed to explore and analyze the SpaceX mission dataset.

- Retrieving the distinct launch site names used across all missions.
- Selecting the first five records where the launch site name begins with “CCA”.
- Calculating the total payload mass transported by boosters for NASA (CRS) missions.
- Determining the average payload mass associated with the booster model F9 v1.1.
- Identifying the earliest date on which a landing was successfully completed on a ground pad.
- Finding boosters that achieved successful drone-ship landings and carried payloads between 4000 and 6000 kg.
- Counting how many missions resulted in successful versus failed outcomes.
- Extracting the booster versions that transported the highest payload mass, based on an aggregate max query.
- Listing failed drone-ship landings in 2015, along with their booster versions and launch site names, using month extraction from the date field.
- Ranking, in descending order, the frequency of different landing outcomes (e.g., Failure (drone ship), Success (ground pad)) for launches between June 4, 2010 and March 20, 2017.

GITHUB URL : EDA with SQL

Build an Interactive Map with Folium

Mapping Launch Sites and Nearby Features

- **Launch Site Markers:**
  The map begins with a reference marker placed at NASA's Johnson Space Center, using its geographic coordinates. Markers with circles, popup descriptions, and text labels were then added for every SpaceX launch site, highlighting their positions and showing how close they are to the equator and nearby coastlines.

- **Outcome-Based Marker Colors:**
  To visualize mission performance, markers were color-coded—green for successful launches and red for failed attempts. A Marker Cluster was used to group these points, making it easier to compare success rates across different launch locations.

- **Distance Visualization to Nearby Infrastructure:**
  Colored lines were drawn from the KSC LC-39A launch site to surrounding features such as the nearest railway line, highway, coastline, and closest city. This helps illustrate the environmental and logistical context of the launch facility.

GITHUB URL : Visual Analysis with Folium

Build a Dashboard with Plotly Dash

Interactive Dashboard Features

- **Launch Site Selection Menu:**
  A dropdown list was included to allow users to choose a specific launch site for focused analysis.

- **Success Rate Pie Charts:**
  A dynamic pie chart was created to display overall successful launch counts across all sites. When a particular site is selected, the chart updates to compare that site’s successful versus failed launches.

- **Payload Mass Range Slider:**
  A slider was added so users can filter data based on a chosen payload mass range, enabling more targeted exploration.

- **Payload vs. Success Rate Scatter Plot:**
  A scatter plot was implemented to visualize how payload mass relates to launch success, grouped by different booster versions, helping reveal potential correlations.

GITHUB URL : Spacex Dashboard

Predictive Analysis (Classification)

Creating NumPy array from the column

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object to find the best parameters

Applying GridSearchCV on LogReg, SVM, decision Tree and KNN model

Calculating the accuracy on the test data using the method. score() for all models

reviewing the confusion matrix for all models

Finding the best model by examining the F1_score metrics

GitHub URL : Machine Learning

Results

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

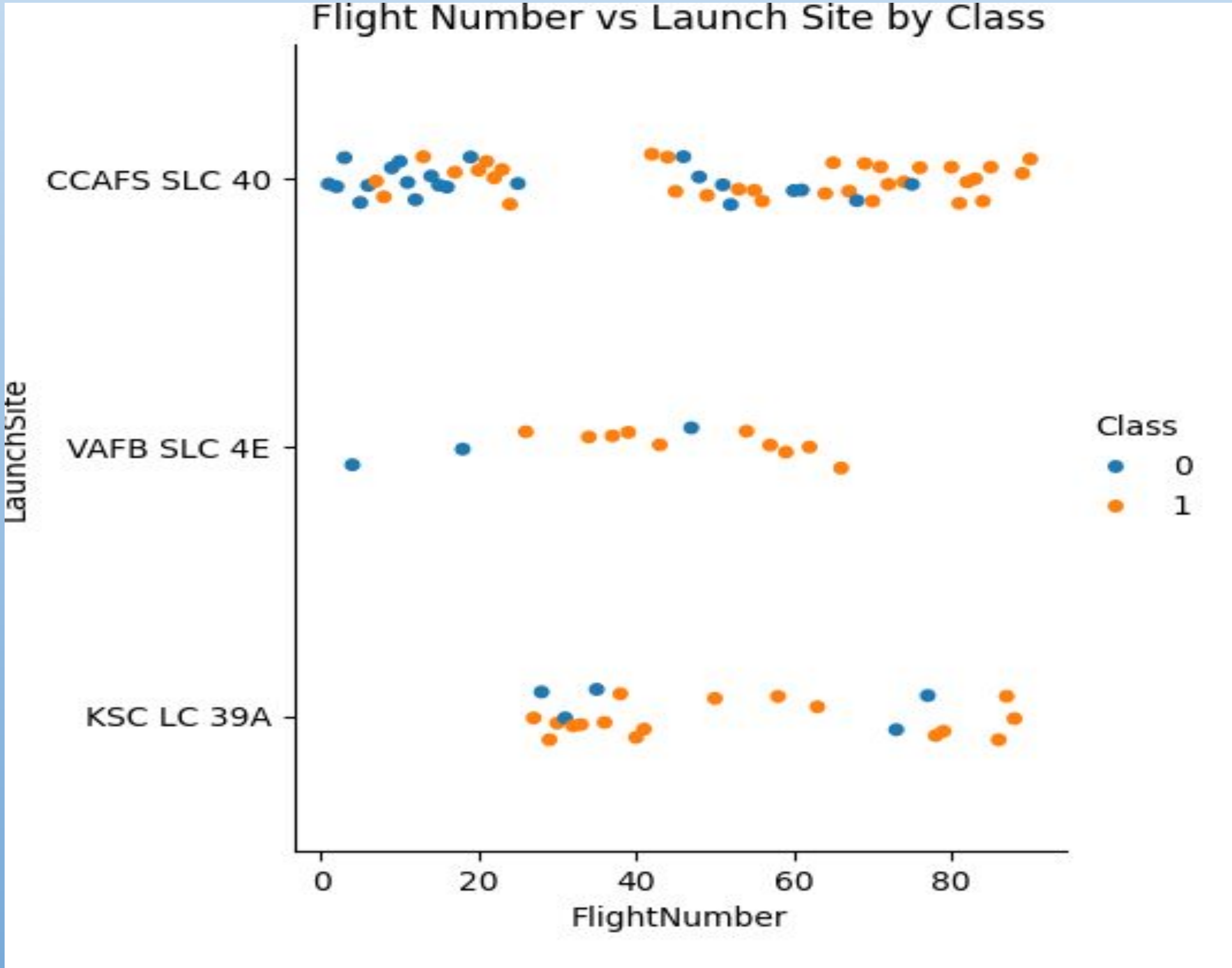
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Insights:

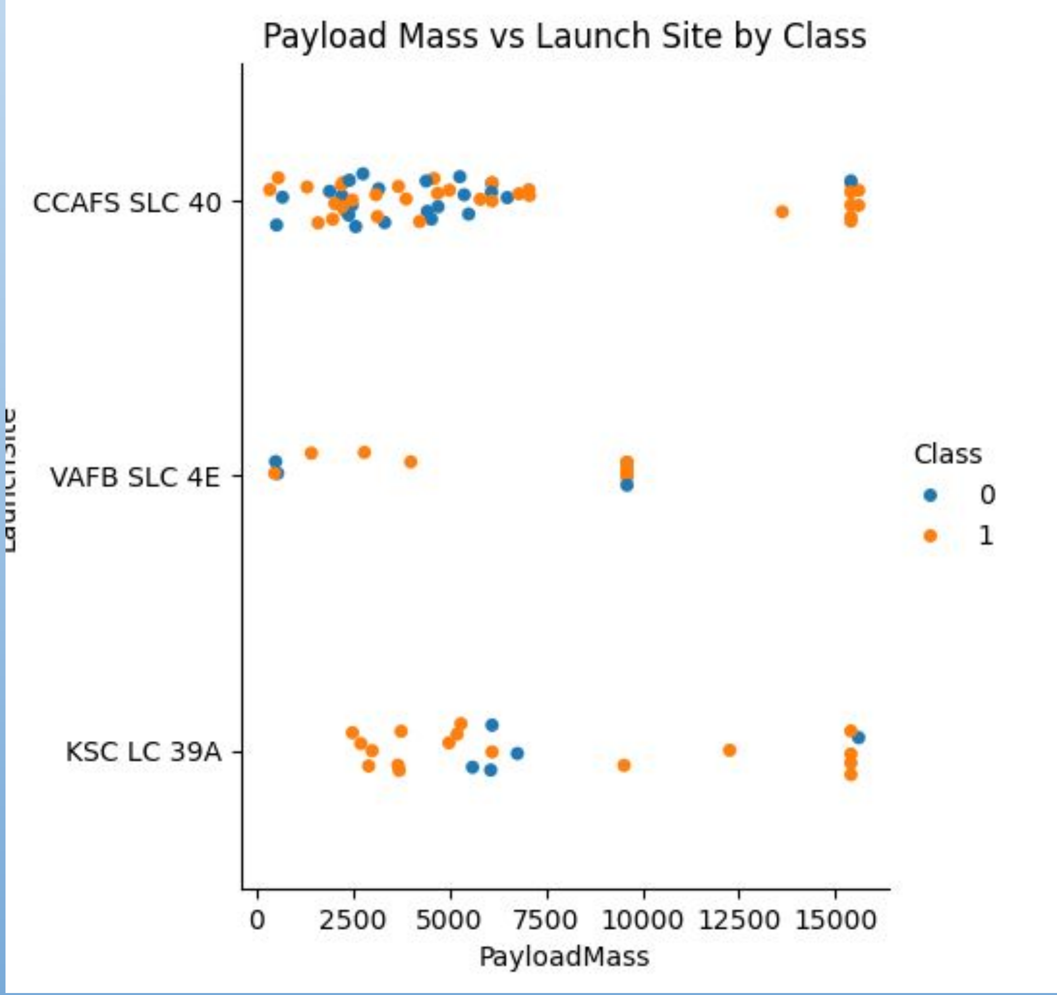
- The first missions in the dataset mostly resulted in failures, while the more recent launches achieved consistent success.
- Nearly half of all launches took place at the CCAFS SLC-40 site.
- VAFB SLC-4E and KSC LC-39A demonstrate noticeably higher success rates compared to other locations.
- Overall, the trend suggests that SpaceX's success rate has steadily improved with each successive launch.



Payload vs. Launch Site

Insights :

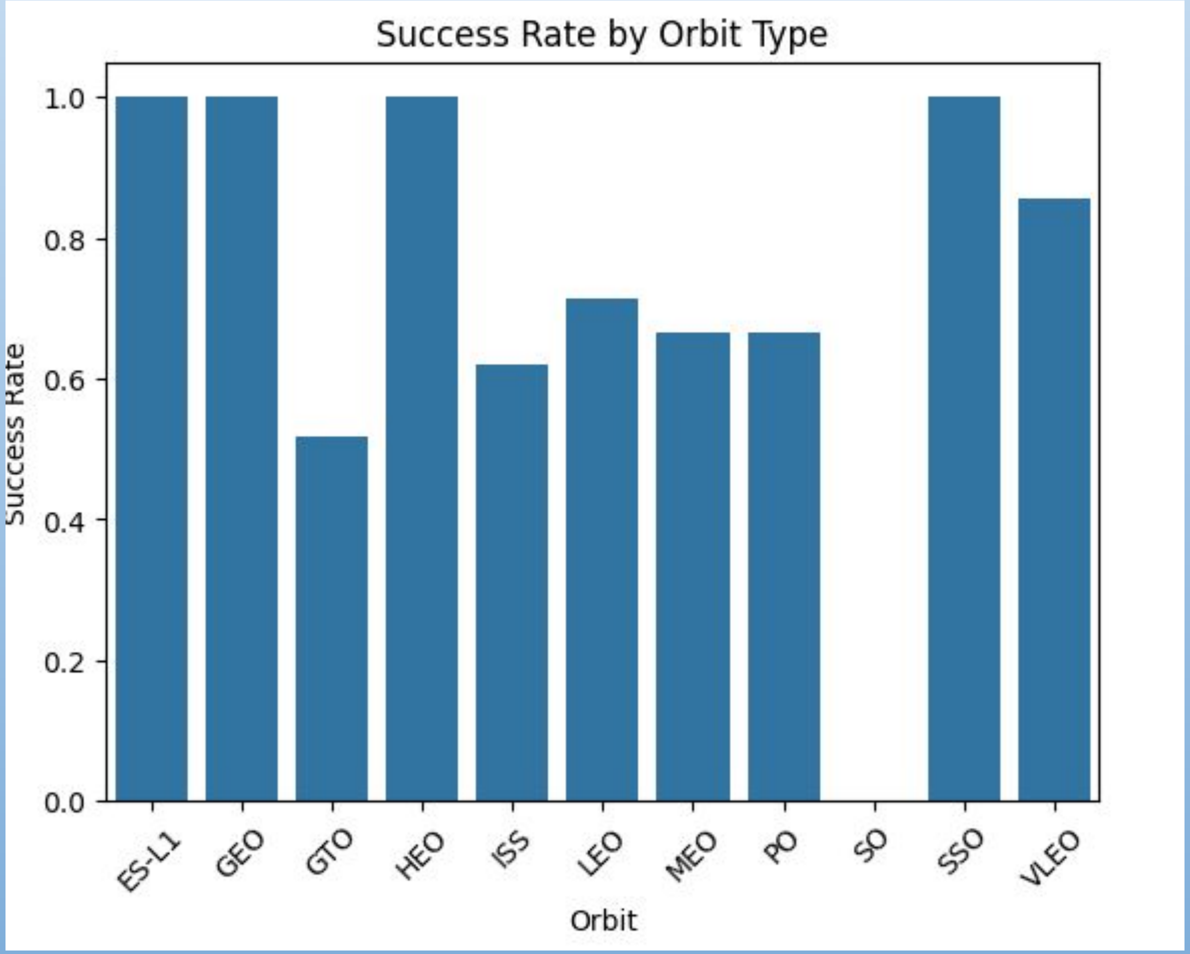
- At all launch sites, missions carrying heavier payloads generally show a higher likelihood of success.
- The majority of launches with payloads exceeding 7000 kg resulted in successful outcomes.
- KSC LC-39A shows perfect performance, achievinga 100% success rate even for payloads below 5500 kg.



Success Rate vs. Orbit Type

Insights:

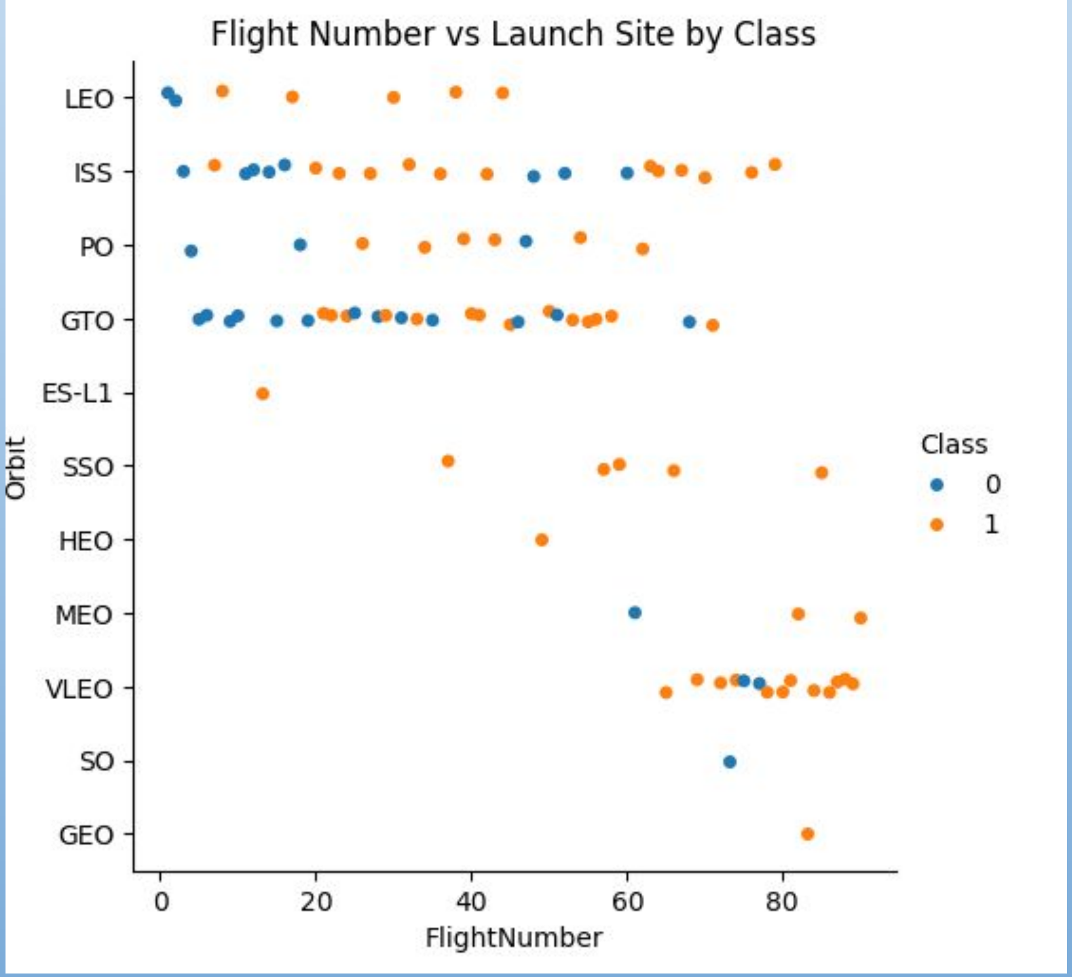
- The orbits that achieved a perfect success rate (100%) include ES-L1, GEO, HEO, and SSO.
- The SO orbit recorded a 0% success rate, with no successful launches.
- Orbits such as GTO, ISS, LEO, MEO, and PO show moderate success rates, ranging roughly between 50% and 85%.



Flight Number vs. Orbit Type

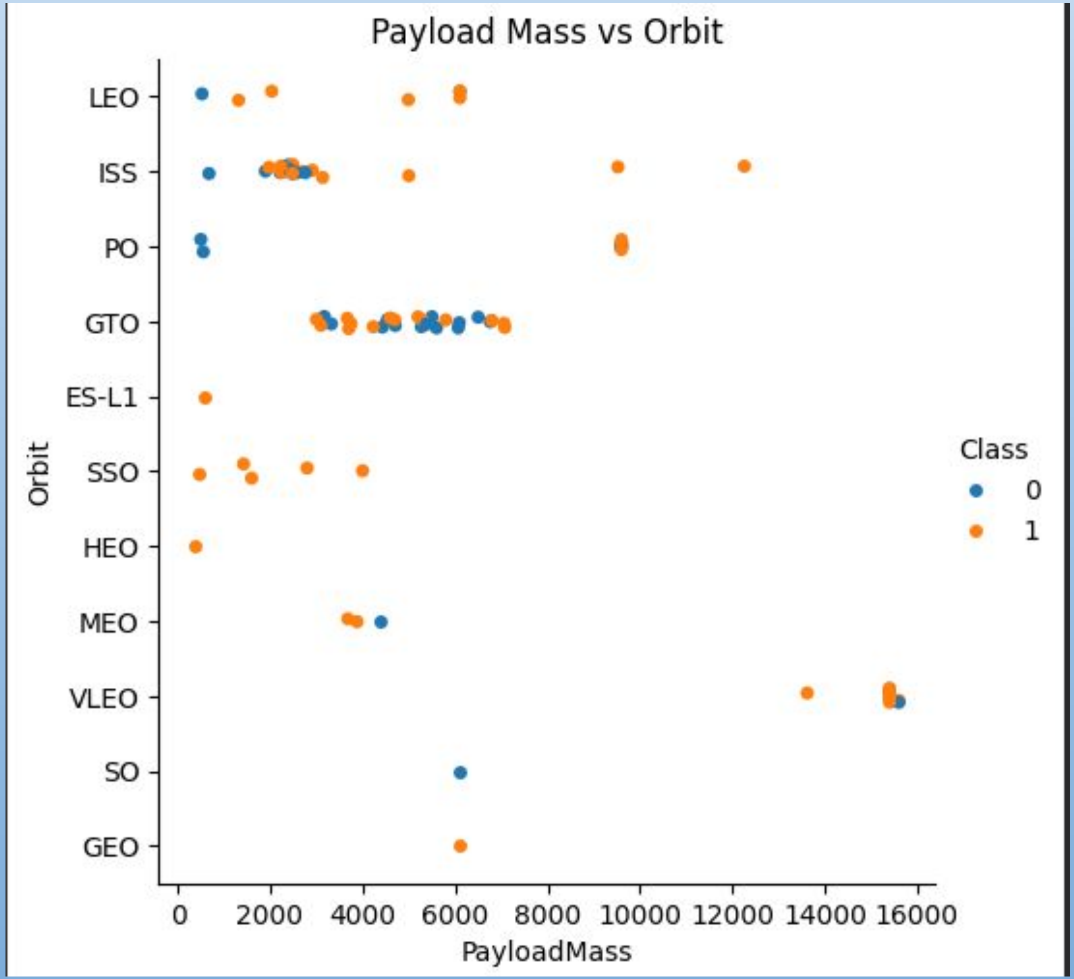
- In **LEO orbit**, the likelihood of success seems to improve as the number of flights increases.

- In contrast, for **GTO orbit**, the success rate does not show any clear connection to how many flights have occurred.



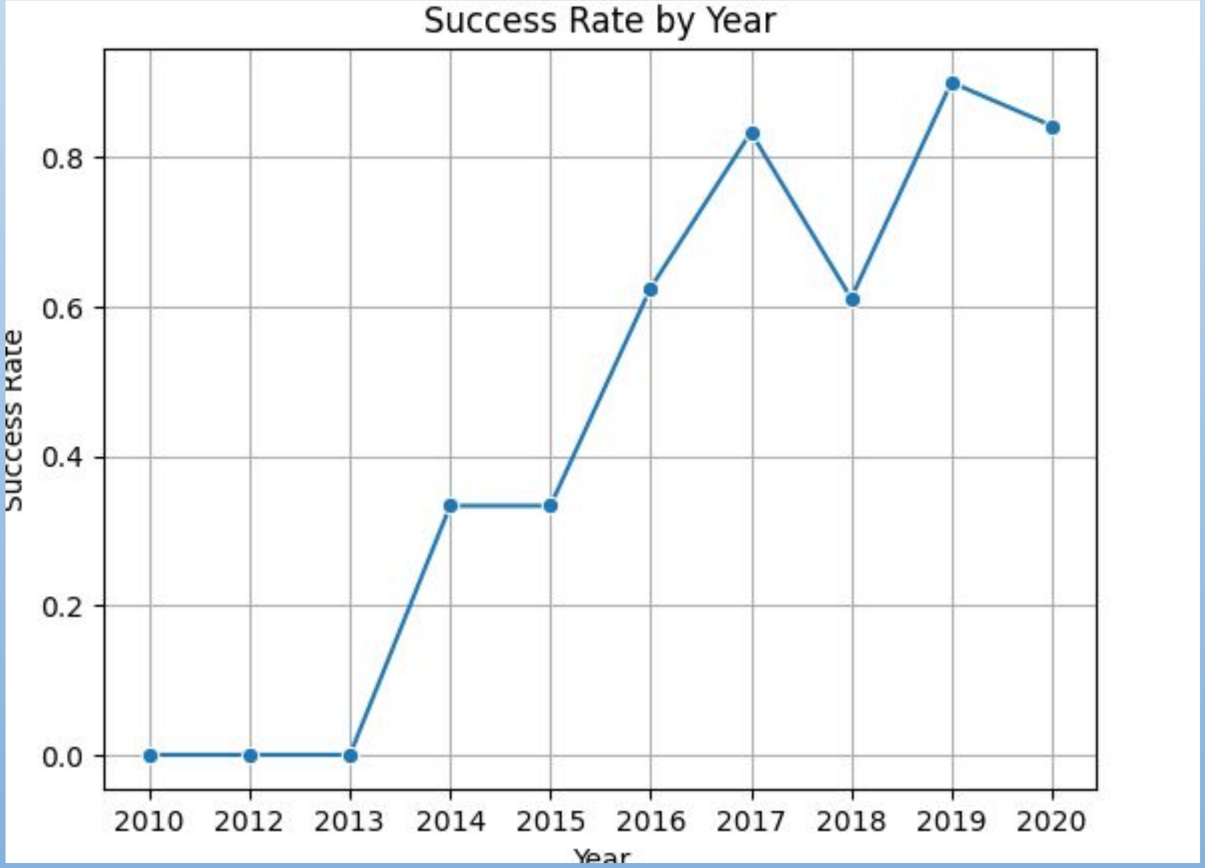
Payload vs. Orbit Type

- Heavier payloads tend to reduce success rates for GTO missions, while they appear to have a positive or supportive effect on launches targeting GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- From 2013 onward, the success rate steadily rose and continued improving up to the year 2020.



All Launch Site Names

- Extracting and presenting the individual launch sites, ensuring only one instance of each site is displayed.

In [12]:

```
%%sql

SELECT DISTINCT "Launch_Site"
FROM SPACEXTBL;
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA'

In [13]:
```
%%sql
SELECT *
FROM SPACEXTBL
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

```
In [19]: %%sql
         SELECT SUM("PAYLOAD_MASS__KG_") AS TotalPayloadMass
         FROM SPACEXTBL
         WHERE "Customer" = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.

Out[19]:
```

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1

```
In [20]: %%sql
         SELECT avg("PAYLOAD_MASS__KG_") AS avgPayloadMass
         FROM SPACEXTBL
         WHERE "Booster_Version" = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

Out[20]:
```

avgPayloadMass
2928.4

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved

```
In [23]: %%sql
         SELECT MIN("Date") AS FirstSuccessfulGroundPadLanding
         FROM SPACEXTBL
         WHERE "Landing_Outcome" = 'Success (ground pad)';

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

Out[23]:

FirstSuccessfulGroundPadLanding
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Retrieving the names of boosters that achieved successful drone-ship landings while transporting payloads weighing more than 4000 kg but under 6000 kg.

```
In [26]: %%sql
         SELECT "BoosterVersion"
         FROM SPACEXTBL
         WHERE "Landing_Outcome" = 'Success (drone ship)'
           AND "PayloadMassKG" > 4000
           AND "PayloadMassKG" < 6000;

          * sqlite:///my_data1.db
         Done.

Out[26]: "BoosterVersion"
```

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes.

In [27]:
```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS Total
FROM SPACEXTBL
GROUP BY "Mission_Outcome";
```

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

Out[27]:

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

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass

In [33]:
```
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTBL
);
```

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

Out[33]:

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

2015 Launch Records

- List the booster versions, launch site names, and failed drone ship landing outcomes for launches that occurred in 2015.

In [35]:

```
%%sql
SELECT
    substr("Date", 6, 2) AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTBL
WHERE
    substr("Date", 1, 4) = '2015'
    AND "Landing_Outcome" LIKE 'Failure (drone ship)%';
```

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

Out[35]:

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

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

- Rank the landing outcomes (e.g., Failure on drone ship, Success on ground pad) by their counts in descending order for launches between 2010-06-04 and 2017-03-20.

In [38]:

```
%%sql
SELECT Landing_Outcome,
    COUNT(*) AS outcome_count,
    RANK() OVER (ORDER BY COUNT(*) DESC) AS outcome_rank
FROM SPACEXTBL
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome
ORDER BY outcome_count DESC;
```

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

Out[38]:

Landing_Outcome	outcome_count	outcome_rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

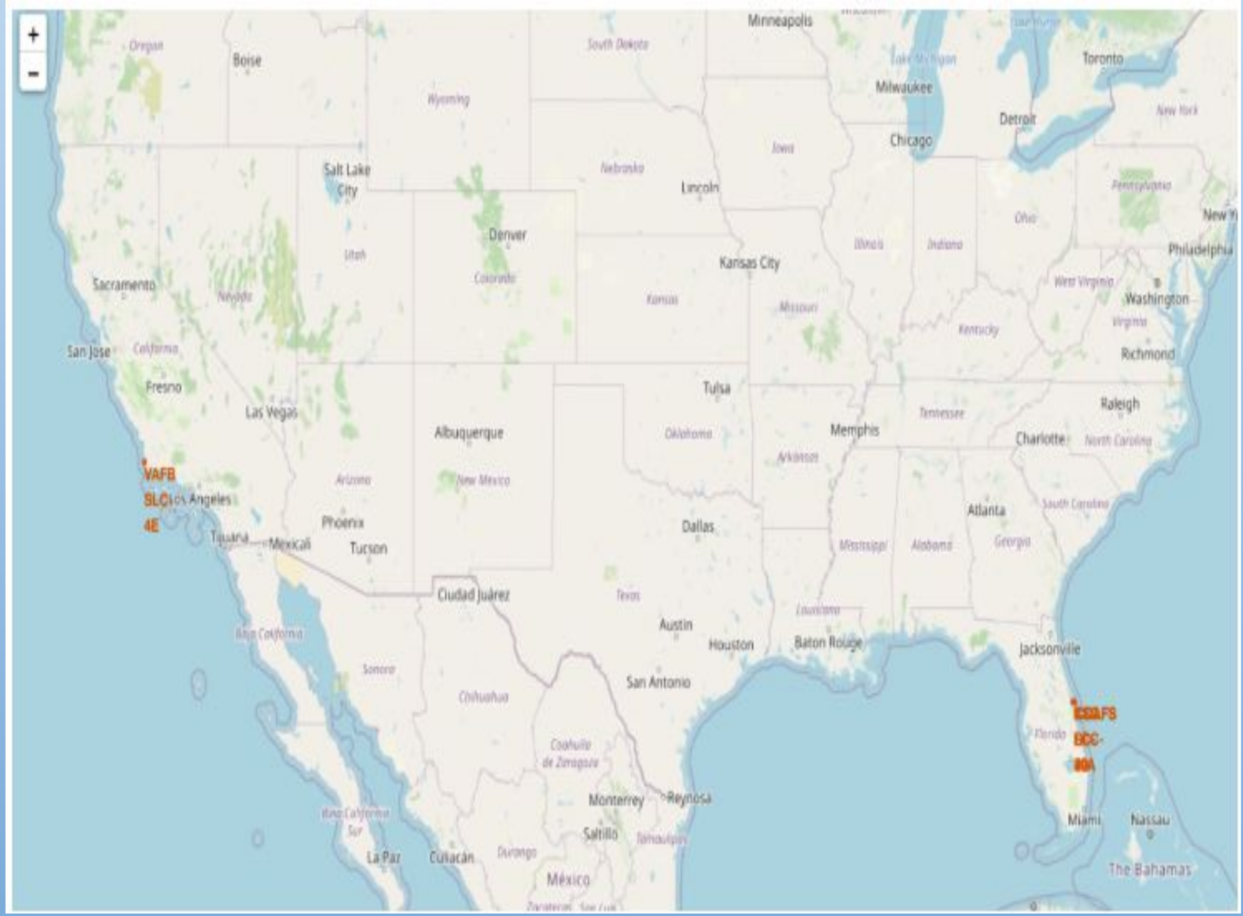
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Most launch sites are located near the Equator, where the Earth's surface moves fastest—about 1670 km/h. A rocket launched from the Equator already has this rotational speed, which, thanks to inertia, helps it achieve the velocity needed to stay in orbit.

- Launch sites are also situated close to the coast so that rockets can be launched over the ocean, reducing the risk of debris falling or explosions near populated areas.



<Folium Map Screenshot 2>

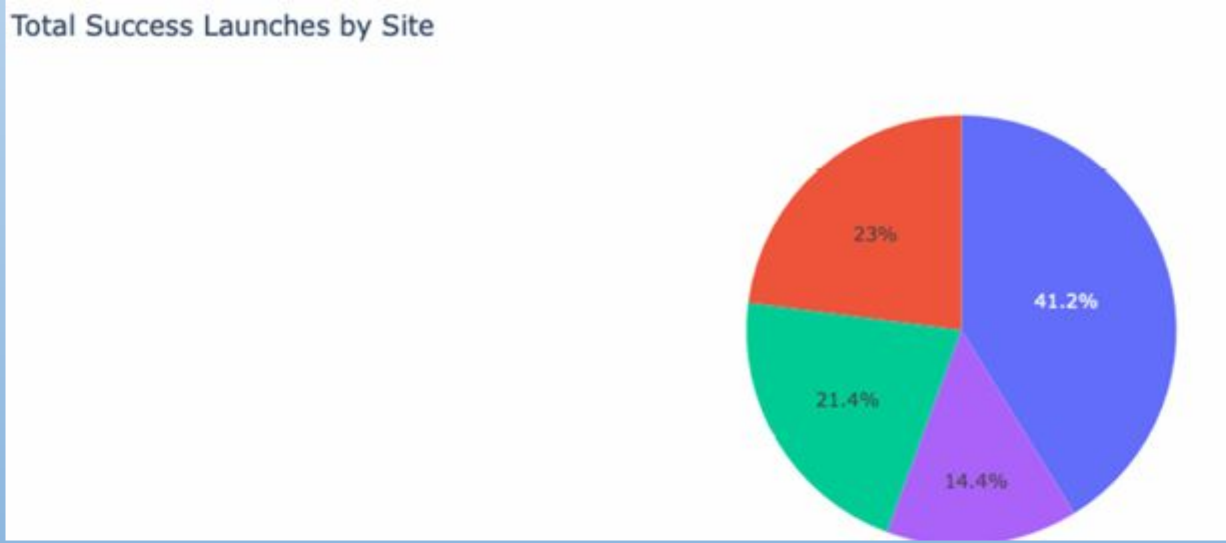
- The color-coded markers make it easy to see which launch sites have higher success rates: green indicates a successful launch, and red indicates a failed launch.

- The launch site KSC LC-39A stands out for having a very high success rate.

Section 4

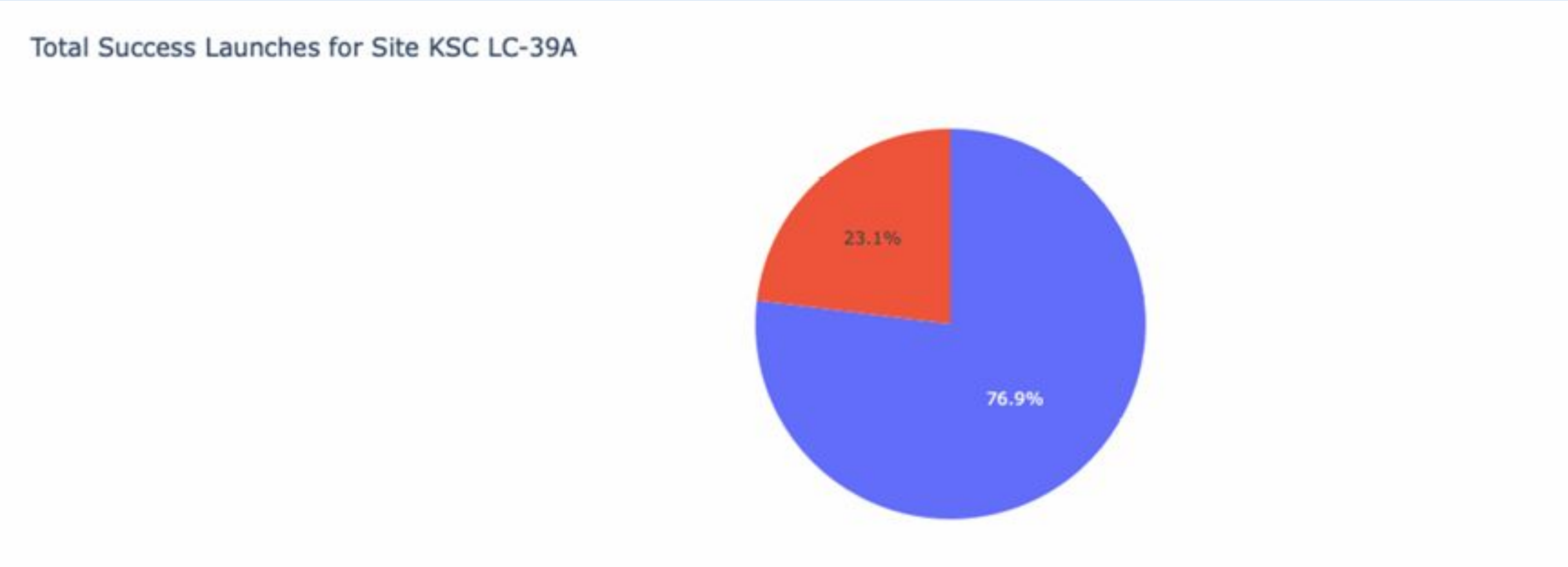
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

<Dashboard Screenshot 2>



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

<Dashboard Screenshot 3>

The charts show that payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Test Set scores alone are insufficient to determine the best-performing method.
- The identical Test Set scores might result from the small sample size (only 18 samples). To get a clearer picture, all methods were evaluated on the entire dataset.
- The full dataset results show that the Decision Tree model performs best, achieving the highest scores and overall accuracy.

Scores and Accuracy of the test Set

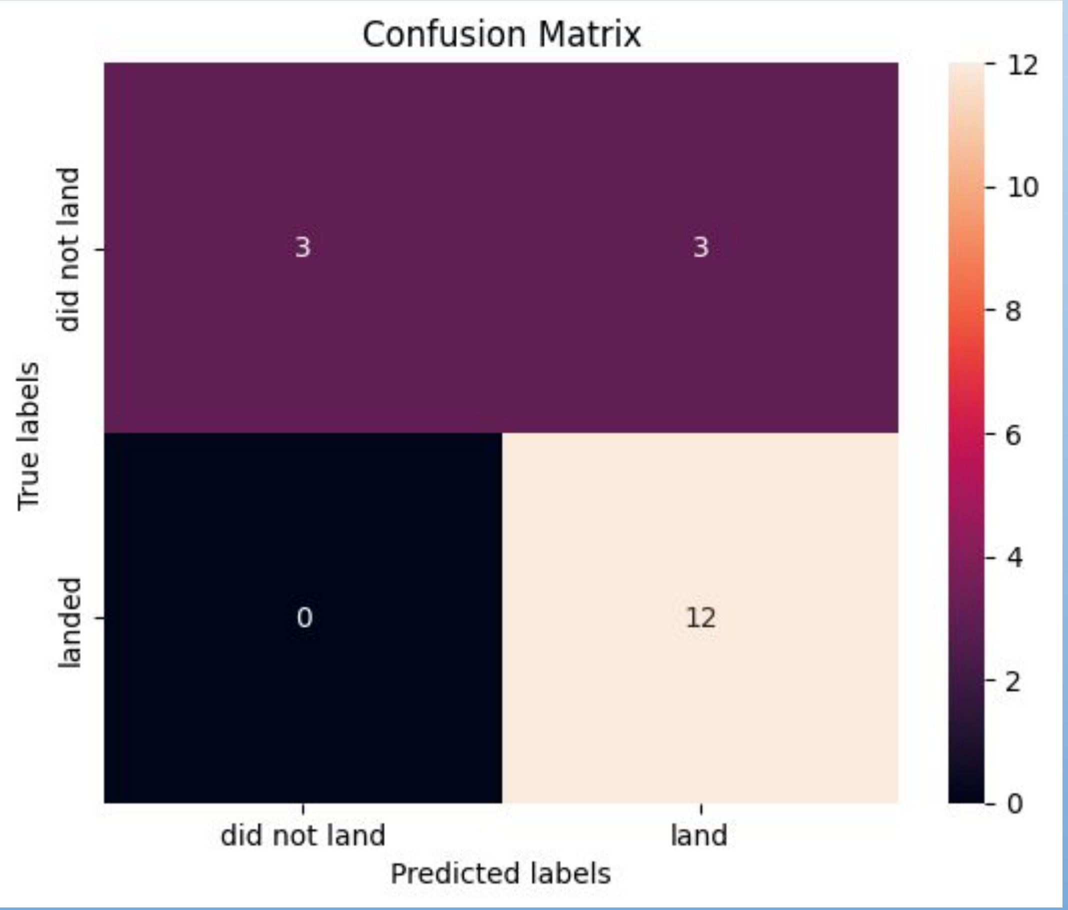
	LogReg	SVM	Tree	KNN
F1_Score	0.888889	0.888889	0.800000	0.888889
Accuracy	0.833333	0.833333	0.666667	0.833333

Scores and Accuracy of the Entire Data set

	LogReg	SVM	Tree	KNN
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.666667	0.855556

Confusion Matrix

- Looking at the confusion matrix, logistic regression is able to differentiate between the classes. However, the main issue lies in a high number of false positives.



Conclusions

- ❖ The Support Vector Machine Learning model is the most effective algorithm for this dataset.
- ❖ Launches with lighter payloads tend to perform better than those with heavier payloads.
- ❖ Most launch sites are located near the Equator, and all are close to the coast.
- ❖ Launch success rates have improved over the years.
- ❖ KSC LC-39A has the highest success rate among all launch sites.
- ❖ Launches to ES-L1, GEO, HEO, and SSO orbits have achieved a 100% success rate.

Thank you!