IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aishwarya Bandapelly
16th November 2025

Outline

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

Executive Summary

Executive Summary - Methodologies

- Comprehensive Data Collection
- Data Cleaning and Wrangling
- Exploratory Data Analysis (EDA) with Advanced Visualizations
- SQL-based Data Exploration
- Interactive Mapping with Folium
- Dashboard Development using Plotly Dash
- Predictive Analysis and Classification

Key Results

- Insights from Exploratory Data Analysis (EDA)
- Demonstration of Interactive Analytics via Screenshots
- Predictive Model Results and Performance Metrics

Introduction

Project Background and Context

- SpaceX is a leader in the commercial space industry, revolutionizing space travel by making it more affordable. Its Falcon 9 rocket launches are listed at $62 million per flight, a fraction of the $165 million charged by other providers. This cost reduction is largely due to SpaceX's ability to reuse the rocket's first stage. Predicting whether the first stage will land successfully is crucial to estimating launch costs. Using public data and machine learning models, this project aims to predict the likelihood of SpaceX reusing the first stage.

Key Questions to Address

- How do factors like payload mass, launch site, number of flights, and orbit type influence the success of the first stage landing?
- Has the success rate of landings improved over time?
- Which algorithm is most effective for binary classification in this context?

Section 1

Methodology

Methodology

Executive Summary

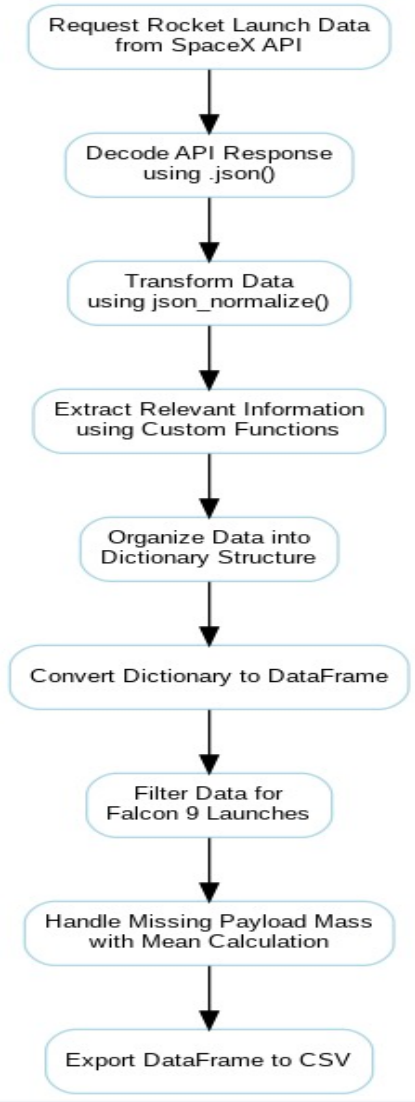
- Data collection methodology:
  - Retrieved data via SpaceX REST API
  - Conducted web scraping from Wikipedia
- Perform data wrangling
  - Filtered and cleaned the data
  - Handled missing values effectively
  - Applied One-Hot Encoding to prepare data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated classification models to achieve optimal results

Data Collection

- The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping data from tables in SpaceX's Wikipedia entry.
- Both methods were essential to obtaining comprehensive information for detailed analysis of SpaceX launches.
- Data Columns Collected via SpaceX REST API - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns Collected via Wikipedia Web Scraping - Flight Number, Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, Time

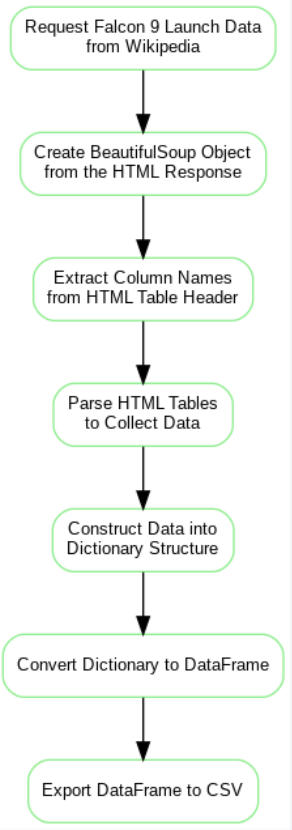
Data Collection – SpaceX API

- Request launch data from the SpaceX REST API.
- Parse API response using .json() and convert to a DataFrame with json_normalize().
- Apply custom functions to extract relevant launch information.
- Organize extracted data into a dictionary structure.
- Transform the dictionary into a DataFrame for analysis.
- Filter data to include only Falcon 9 launches.
- Replace missing payload mass values with the column mean.
- Export the final DataFrame to a CSV file for further use.
- Github link - Data_Collection_Lab1



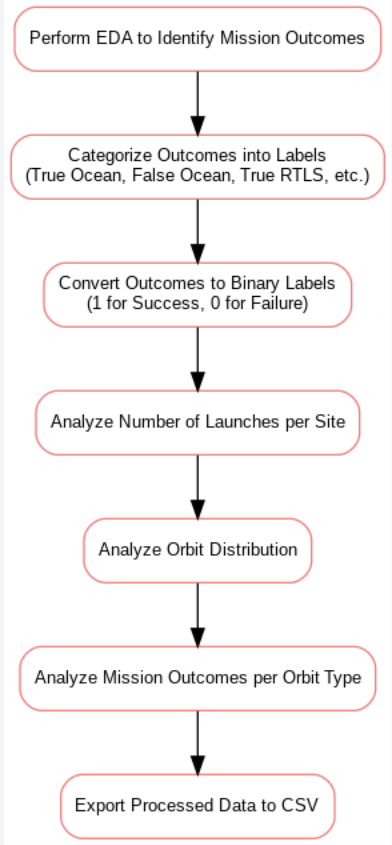
Web Scraping

- Request Falcon 9 launch data from Wikipedia using HTTP requests.Parse the HTML response using BeautifulSoup.Extract column names from the table headers in the HTML.Collect data by parsing rows and cells in the HTML table.Organize the parsed data into a dictionary structure.Convert the dictionary into a pandas DataFrame.Save the DataFrame as a CSV file for further use.
- Github link - Data_WebScraping_Lab2



Data Wrangling

- The data wrangling process involved categorizing mission outcomes into specific labels, such as successful or unsuccessful landings in the ocean, on ground pads, or on drone ships.
- These labels were converted into binary training data, where "1" represents a successful landing and "0" indicates failure.
- Key exploratory analysis was conducted to determine the number of launches per site, the distribution of orbits, and the frequency of mission outcomes by orbit type.The cleaned and labeled data was then prepared and exported as a CSV file for further analysis.
- Github link - EDA_SQL_Lab4



EDA with SQL

- SpaceX utilizes multiple launch sites, including Cape Canaveral (CCA), indicating a strategic approach to mission deployment.
- The total payload mass for NASA's Commercial Resupply Services (CRS) missions was calculated, highlighting SpaceX's role in supporting ISS operations.
- The average payload mass for the F9 v1.1 booster version was determined, providing insights into its operational capacity.
- The date of the first successful ground pad landing was identified, marking a significant achievement in SpaceX's reusability efforts.
- The total number of successful and failed mission outcomes was calculated, offering an overview of SpaceX's launch success rate
- Github link - EDA_SQL_Lab4

Build an Interactive Map with Folium

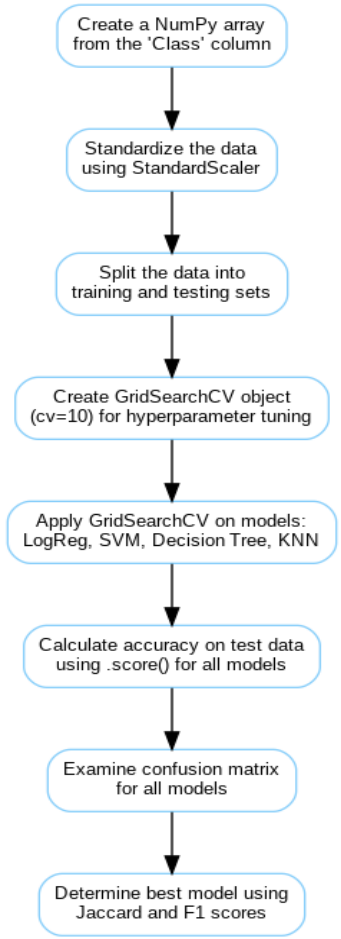
- Added markers for all launch sites using circles, popup labels, and text labels to display their latitude and longitude coordinates.
- Included the NASA Johnson Space Center as the starting location.
- Placed markers to visualize the geographical locations of all launch sites and their proximity to the equator and coastlines.
- Used colored markers (green for success, red for failure) to represent launch outcomes for each site.
- Leveraged marker clusters to identify launch sites with higher success rates.
- Added lines to indicate distances between launch sites (e.g., KSC LC-39A) and nearby features such as railways, highways, coastlines, and cities.
- Github link -VisualAnalytics_Folium_Lab6

Build a Dashboard with Plotly Dash

- Added a dropdown list to select specific launch sites.
- Created a pie chart to visualize the total successful launches, including success vs. failure counts, for all sites or a specific site.
- Implemented a slider to adjust and filter data based on payload mass range.
- Included a scatter chart to analyze the relationship between payload mass and success rate for different booster versions.
- Github link – Spacex_Dash_App_Lab7

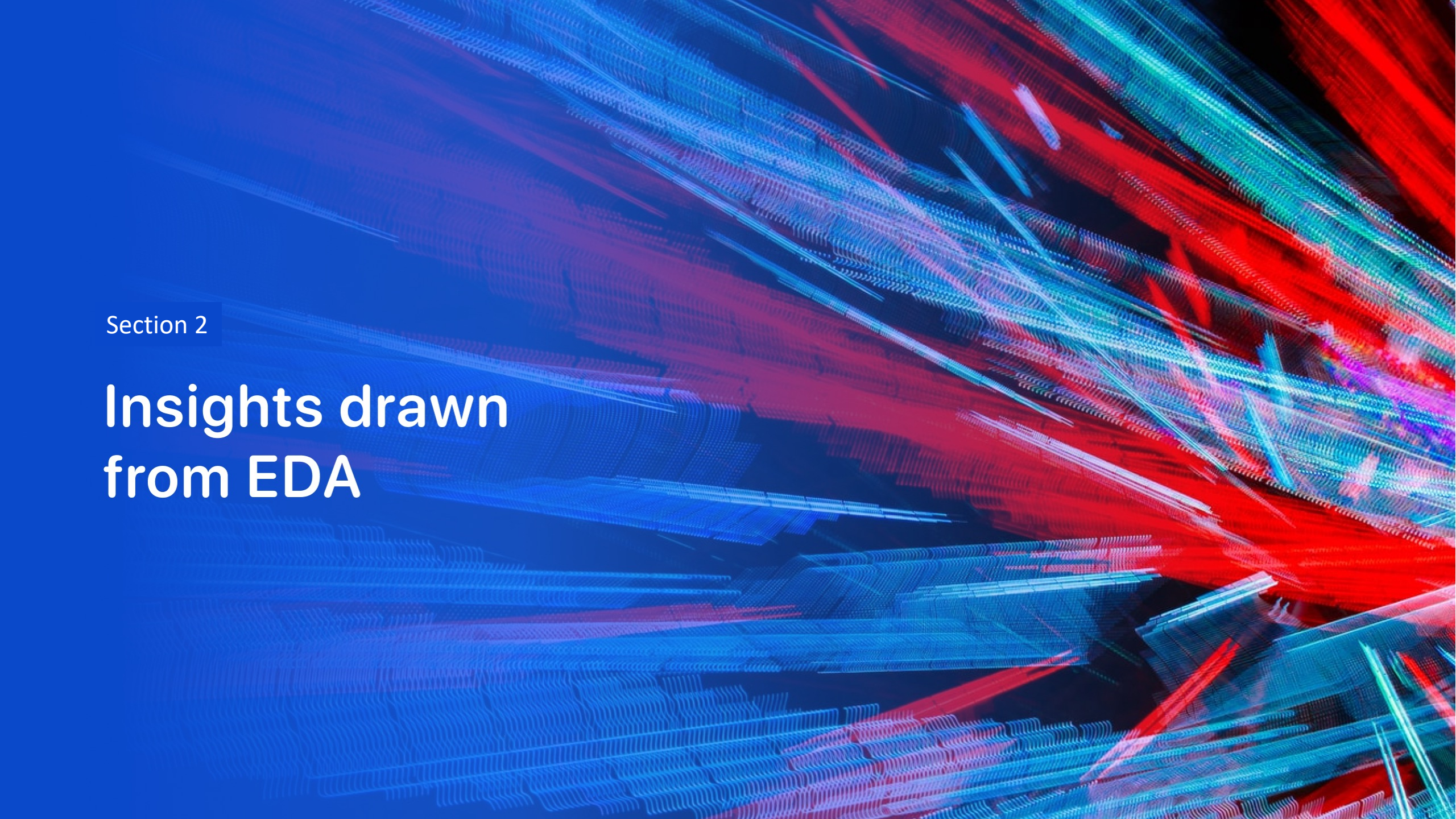
Predictive Analysis (Classification)

- The predictive analysis process begins with preparing the data for classification by converting relevant columns into a NumPy array.
- The data is then standardized using a StandardScaler to improve model performance. The dataset is split into training and testing sets to validate model accuracy.
- A GridSearchCV object is used with cross-validation (cv=10) to determine the best hyperparameters.
- Various machine learning models, such as Logistic Regression, SVM, Decision Tree, and KNN, are trained and evaluated.
- Model performance is assessed using metrics like accuracy, Jaccard Score, F1 Score, and confusion matrices, with the goal of selecting the best-performing method.
- Github link – [Machine-Learning-Prediction_Lab8]

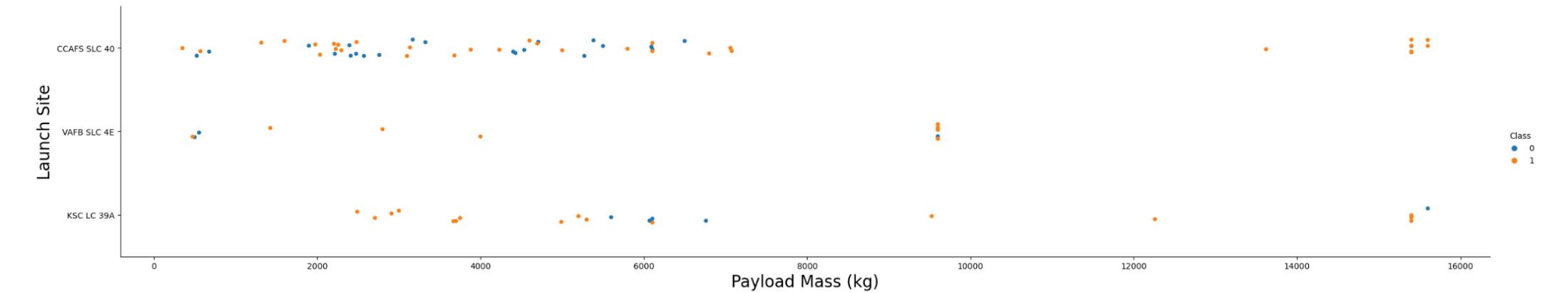


Results

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

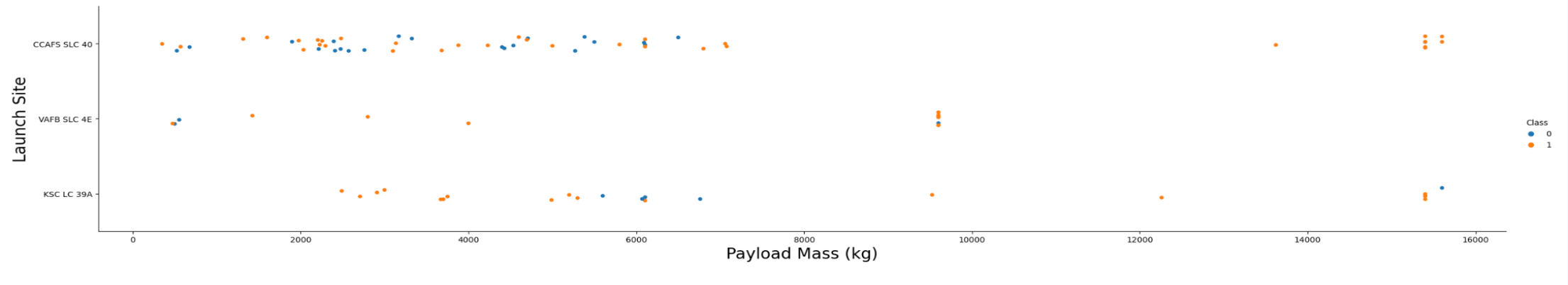

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all resulted in failures, indicating initial challenges in achieving successful landings.
- The CCAFS SLC 40 launch site accounts for approximately half of all recorded launches.
- The VAFB SLC 4E and KSC LC 39A launch sites demonstrate higher success rates compared to other locations.
- The data suggests an increasing trend in the success rate over time, with newer launches achieving better outcomes.

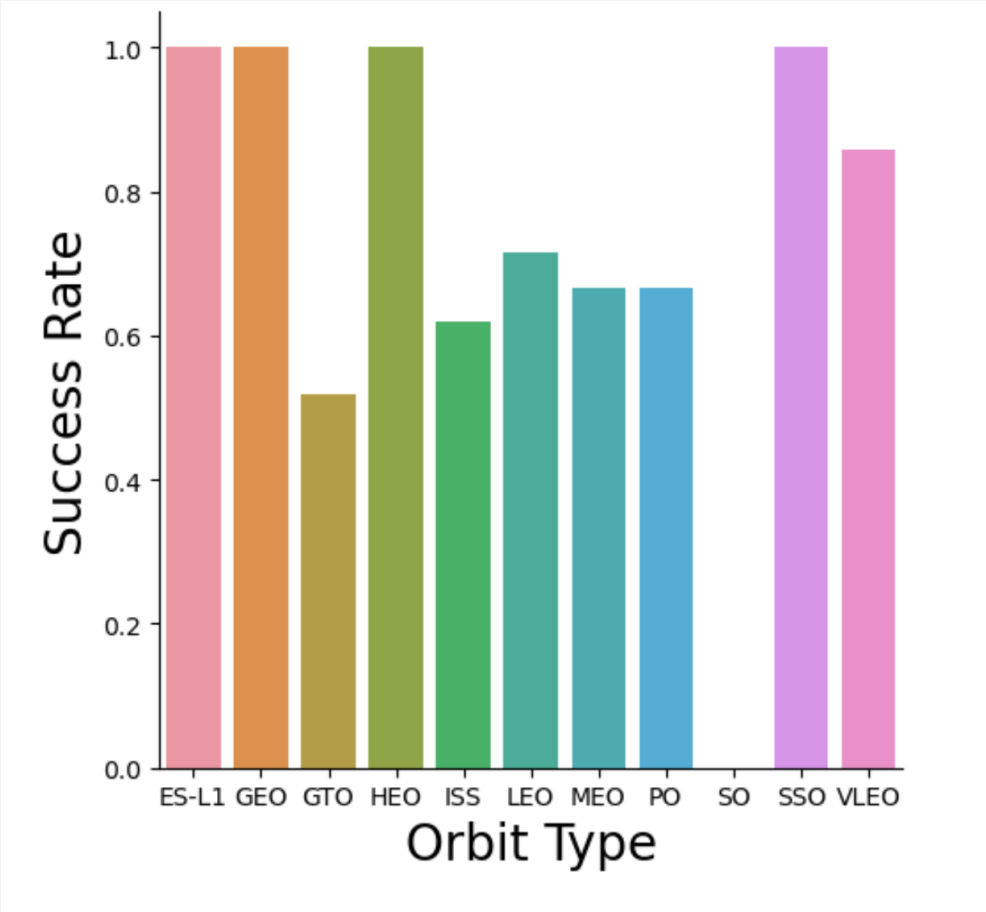
Payload vs. Launch Site



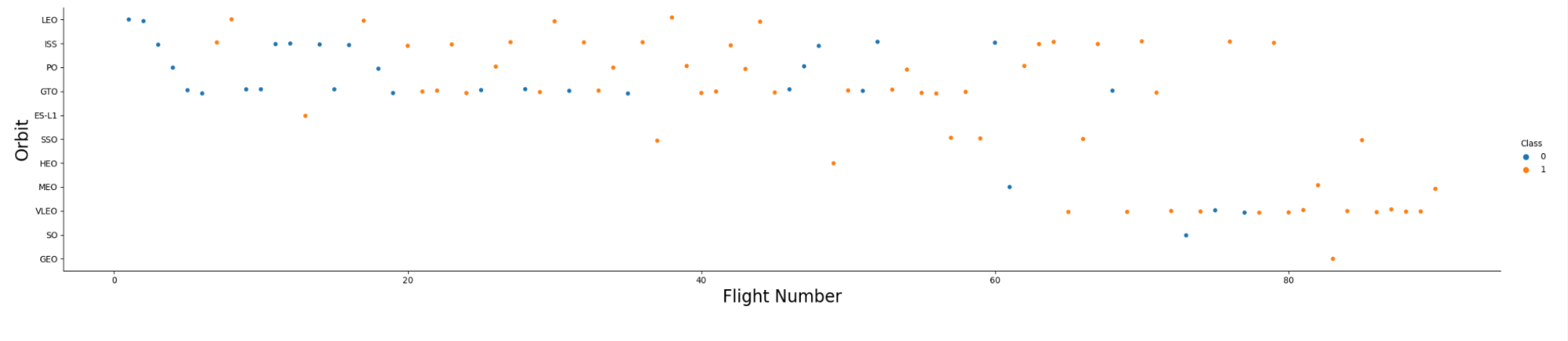
- For every launch site, a higher payload mass generally corresponds to a higher success rate.
- Most launches with payload masses exceeding 7000 kg were successful.
- The KSC LC 39A site achieved a 100% success rate for payload masses under 5500 kg.

Success Rate vs. Orbit Type

- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO.
- Orbit with 0% success rate: SO.
- Orbits with success rates between 50% and 85%: GTO, ISS, LEO, MEO, PO.

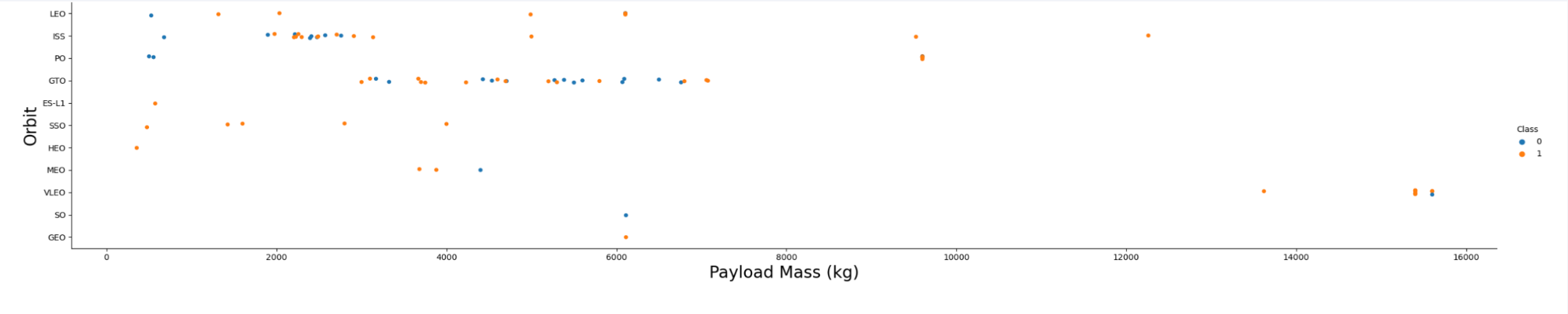


Flight Number vs. Orbit Type



- In the LEO orbit, success appears to correlate with the number of flights, indicating potential improvements or adjustments over time.
- In the GTO orbit, there seems to be no clear relationship between the flight number and the success rate.

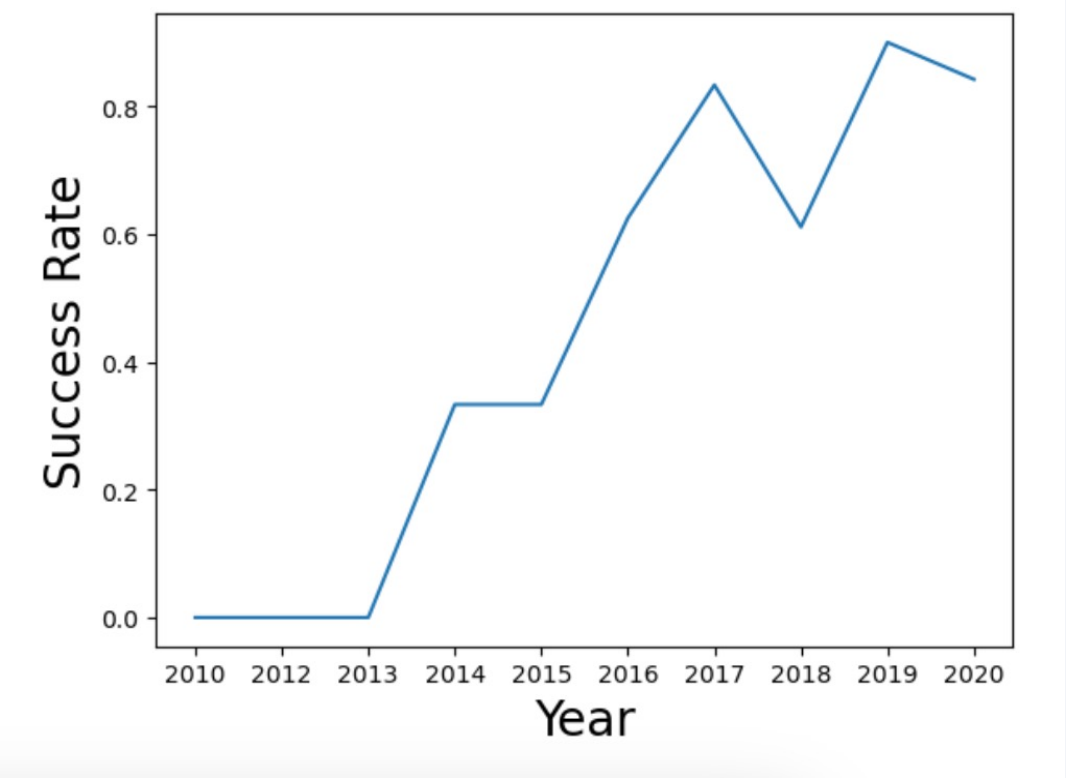
Payload vs. Orbit Type



- Heavy payloads negatively influence success rates in GTO orbits.
- Conversely, heavy payloads positively impact success rates in Polar LEO (ISS) orbits.

Launch Success Yearly Trend

•The success rate of launches has steadily increased from 2013, reaching its peak around 2020.



All Launch Site Names

- This command retrieves the unique names of the launch sites from the SPACEXDATASET table.
- It ensures that each launch site is listed only once, highlighting the distinct locations used for SpaceX launches.

```
In [68]:  1 %sql SELECT DISTINCT launch_site FROM SPACEXDATASET;
 * sqlite:///my_data1.db
Done.
```

Out[68]:

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

Launch Site Names Begin with 'CCA'

- Displays 5 records where launch_site starts with "CCA", filtering for sites like CCAFS LC-40.
- Includes details such as date, booster version, payload, orbit, mission outcome, and landing outcome.

```
1 %sql SELECT * FROM SPACEXDATASET 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

Total Payload Mass

- This query calculates the total payload mass carried by boosters for missions conducted by NASA (CRS).
- The result shows the sum of the payload_mass__kg_ column filtered by the customer name "NASA (CRS)".

```
In [12]:
1 %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[12]:
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried by missions using the booster version F9 v1.1.
- The result shows the mean value of the payload_mass__kg_ column filtered for missions with the specified booster version.

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

In [13]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1
```

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

Out[13]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- This query retrieves the date of the first successful ground landing by selecting the minimum date where the landing__outcome is "Success (ground pad)".
- The result shows the first successful ground landing occurred on 2015-12-22.

```
In [18]:  1 %sql SELECT MIN(date) AS first_successful_landing FROM SPACEXDATASET WHERE landing_outcome = 'Success (ground pa
```

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

Out[18]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query retrieves the booster_version of rockets that achieved a successful drone ship landing with a payload mass between 4000 kg and 6000 kg.
- The result lists the relevant booster versions meeting these criteria.

In [23]:
```
1 %sql SELECT booster_version FROM SPACEXDATASET WHERE landing_outcome = 'Success (drone ship)' AND payload_mass_k
```

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

Out[23]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- This query retrieves the total number of successful and failure mission outcomes by grouping the data based on the mission_outcome column.
- The COUNT(*) function is used to calculate the number of records for each unique mission outcome.

```
In [24]:  1 %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

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

Out[24]:

mission_outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query lists the names of booster versions that carried the maximum payload mass.
- The MAX() function is used in a subquery to determine the maximum payload mass. The outer query retrieves the corresponding booster versions.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [25]:
```
%sql SELECT booster_version FROM SPACEXDATASET WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM SPACE
```

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

Out[25]:

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

- Lists failed drone ship landings in 2015.
- Includes month, date, booster version, launch site, and landing outcome as "Failure (drone ship)".

In [26]:

```
%%sql
SELECT
    strftime('%m', date) AS month,
    date,
    booster_version,
    launch_site,
    landing_outcome
FROM SPACEXDATASET
WHERE landing_outcome = 'Failure (drone ship)'
  AND strftime('%Y', date) = '2015';
```

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

Out[26]:

month	date	booster_version	launch_site	landing_outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranks landing outcomes based on their count in descending order.
- Ranks landing outcomes based on their count in descending order.

In [27]:

```
%%sql
SELECT
    landing_outcome,
    COUNT(*) AS count_outcomes
FROM SPACEXDATASET
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome
ORDER BY count_outcomes DESC;
```

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

Out[27]:

landing_outcome	count_outcomes
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


Section 3
Launch Sites
Proximities Analysis

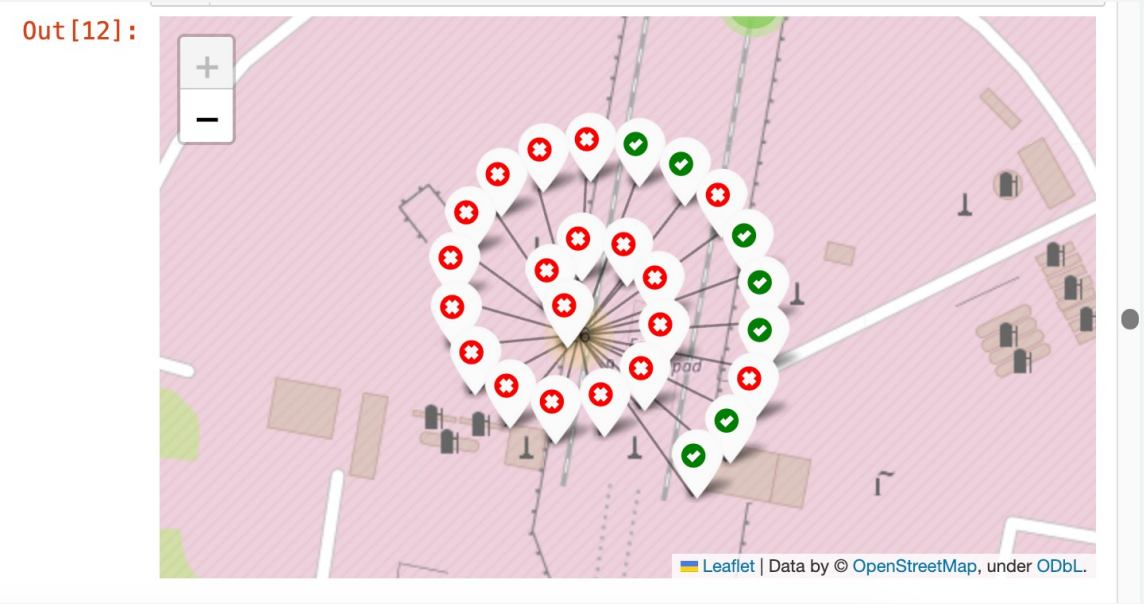
Launch Sites Location Markers on a Global Map

- Most launch sites are located near the equator to leverage the Earth's rotational speed, which is faster at the equator, aiding spacecraft in achieving orbit efficiently.
- Proximity to the coast ensures that rockets launched over the ocean minimize risks of debris or failures impacting populated areas.

Out[7]:

Colour-labeled Launch Records

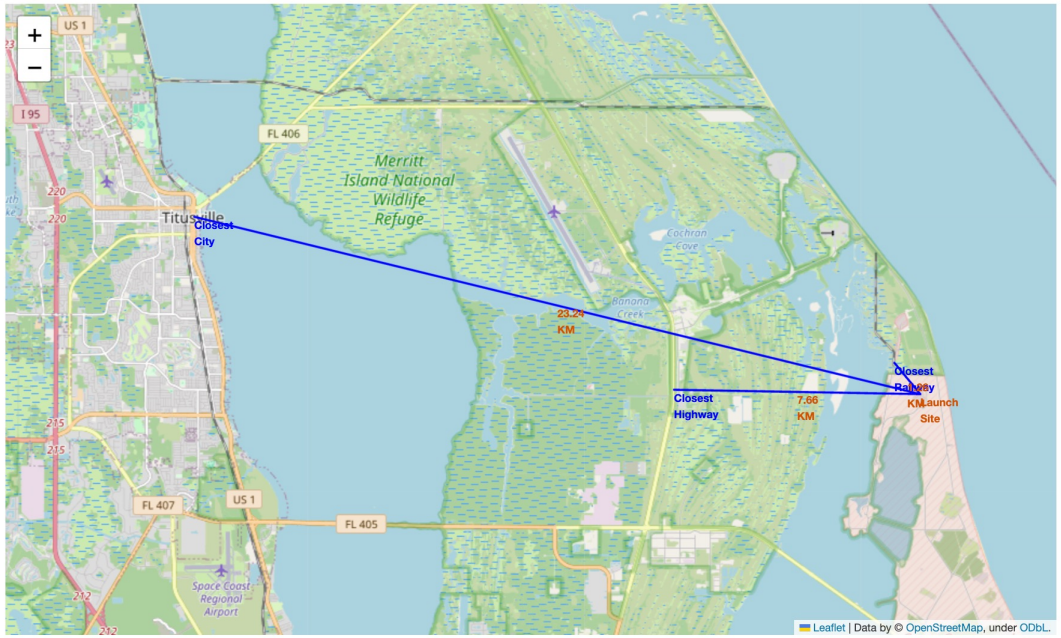
- The color-labeled markers visually represent the success rates of launches at different sites.
- Green Marker: Indicates successful launches.
- Red Marker: Indicates failed launches.
- The KSC LC-39A launch site has a notably high success rate.

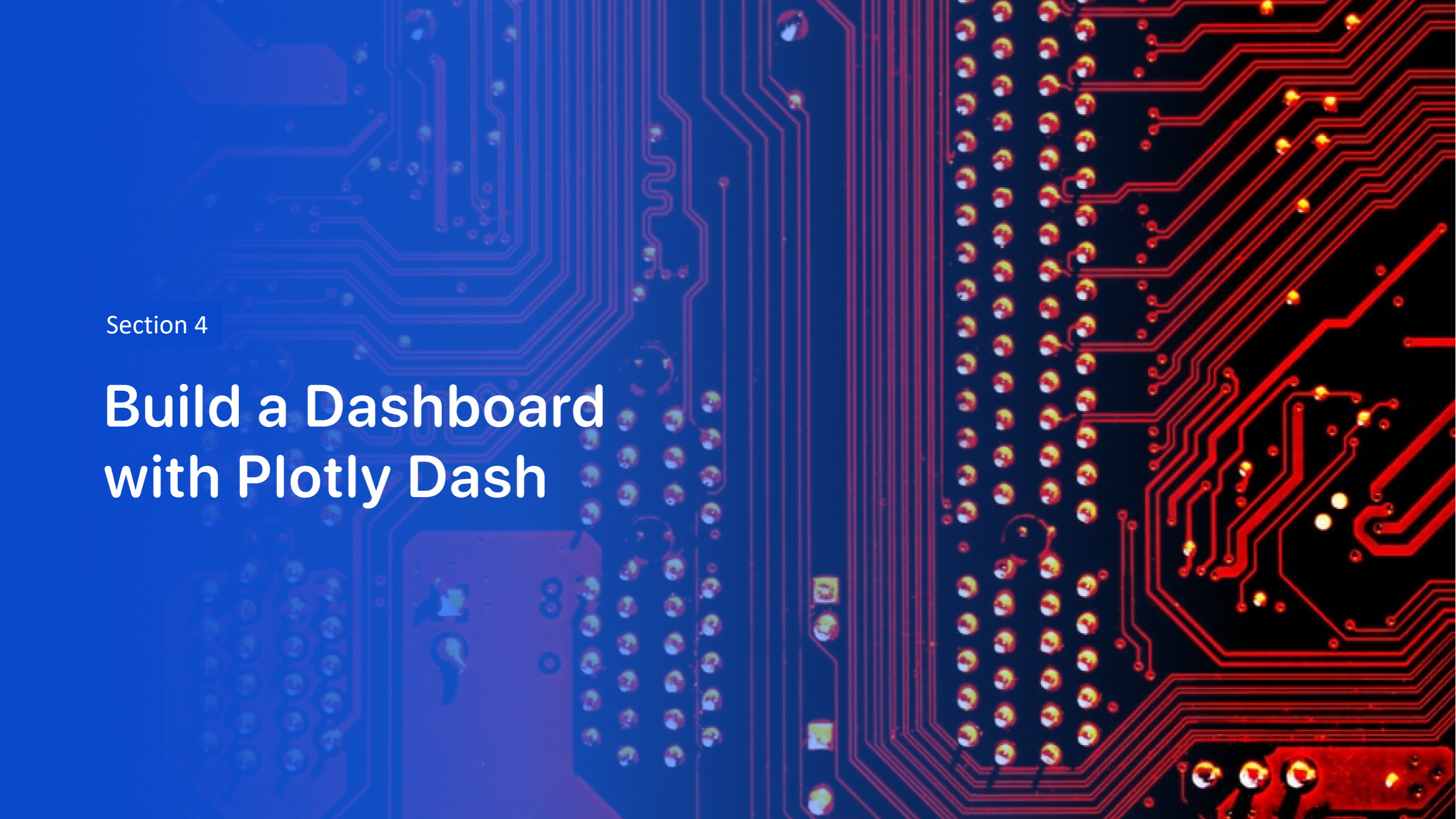


<Folium Map Screenshot 3>

- The launch site KSC LC-39A is located near essential landmarks, such as railways, highways, coastlines, and nearby cities, with distances ranging between 15 to 20 kilometers.
- Due to the high speeds of rockets, any failure could pose potential risks to surrounding areas within this range, emphasizing the importance of strategic launch site placement.

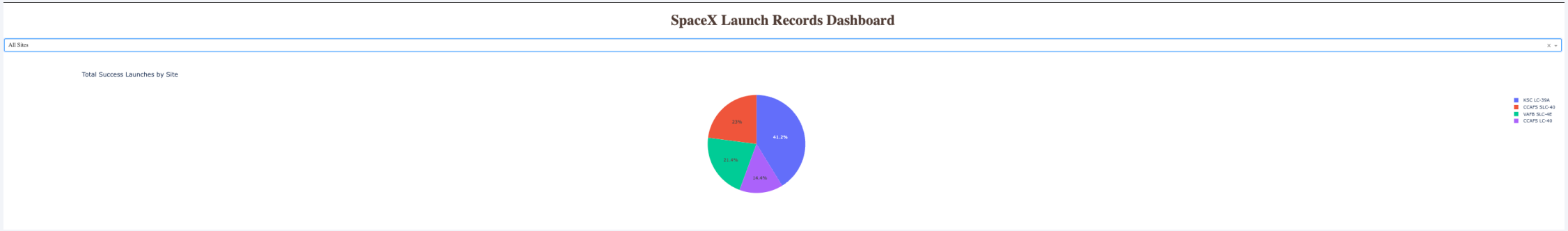
Out[17]:




Section 4
Build a Dashboard with Plotly Dash

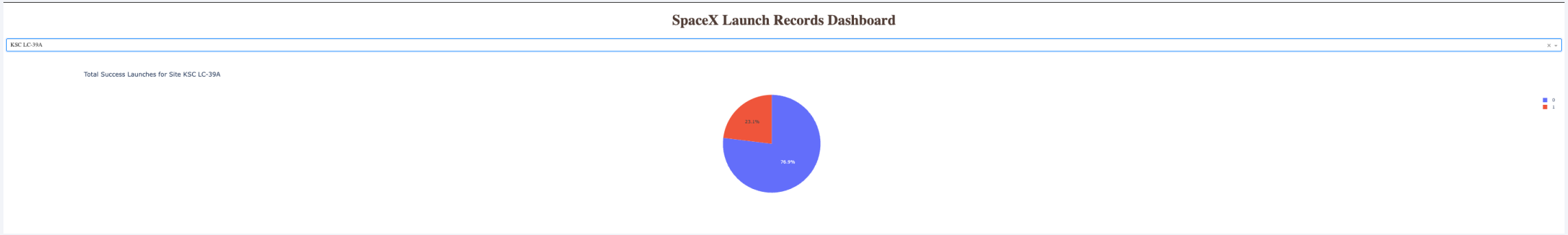
Launch Success Count for all Sites

- The pie chart illustrates that KSC LC-39A accounts for the largest share of successful launches among all launch sites.
- It highlights KSC LC-39A as the most prominent site in terms of launch success rate compared to the other sites.



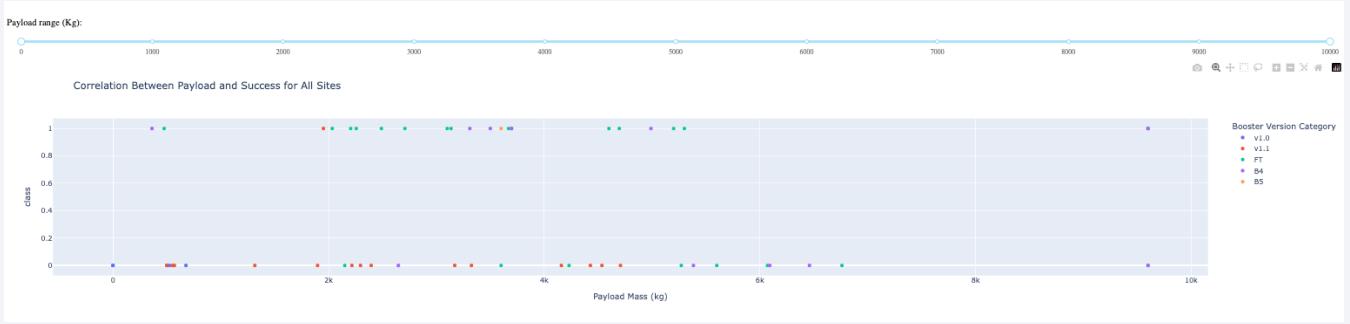
Site with Highest Launch Success Ratio

- KSC LC-39A has the highest launch success ratio at 76.9%, which translates to 10 successful launches out of a total of 13 attempts.
- The remaining 3 launches at this site ended in failure, accounting for 23.1% of the total launches..



Payload Mass vs Launch Outcome

- The charts demonstrate that payloads in the range of 2000 to 5500 kg have the highest success rates across all launch sites.
- This range represents the optimal payload mass for achieving successful launches, indicating favorable conditions for mission outcomes.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

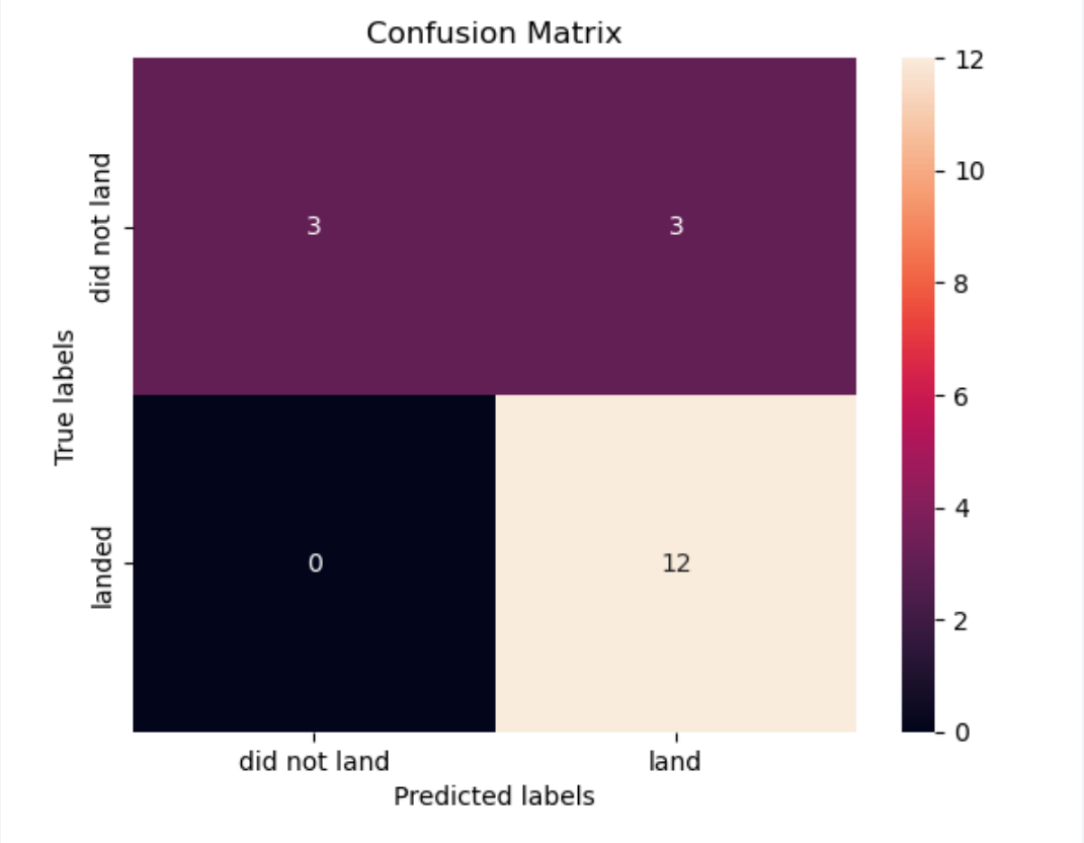
- SVM is the best-performing model, with the highest Jaccard Score (0.845), F1 Score (0.916), and Accuracy (0.878).Decision
- Tree performs the worst, with the lowest scores across all metrics (Jaccard: 0.667, F1: 0.800, Accuracy: 0.667).
- Both Logistic Regression and KNN perform moderately well but are outperformed by SVM in all metrics.

Out[34]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.666667	0.819444
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.666667	0.855556

Confusion Matrix

- chieved the highest performance across all metrics, with a Jaccard Score of 0.845, F1 Score of 0.916, and Accuracy of 0.878, indicating superior predictive capabilities.
- Both models demonstrated moderate performance, each with a Jaccard Score of 0.800, F1 Score of 0.889, and Accuracy of 0.833, suggesting comparable effectiveness in classification tasks.
- Recorded the lowest performance metrics, with a Jaccard Score of 0.667, F1 Score of 0.800, and Accuracy of 0.667, indicating challenges in accurately classifying the data.



Conclusions

- The Decision Tree model performed the best on the dataset when considering both accuracy and F1-score for classification tasks.
- Launches with lower payload mass generally exhibited higher success rates, while heavier payloads were more challenging to launch successfully.
- Launch sites located near the equator and coastlines, like KSC LC-39A, showed higher success rates, aligning with optimal launch conditions.
- The Decision Tree model performed exceptionally well on the entire dataset, making it the most suitable algorithm for this analysis.
- Proximity to the Equator and coastal regions significantly contributes to the success of launch sites due to optimal speed and reduced risks.
- Launch success rates have improved consistently over the years, with KSC LC-39A emerging as a top-performing launch site.

Appendix

Coursera has provided with the following –

- Data Tables and Charts
- Code Snippets
- Survey Instruments
- Interview Transcripts
- Additional Documentations and Links

Thank you!