IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Telu Geyasree
18 October 2025

Outline

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

Executive Summary

Summary of methodologies

- This project predicts the landing success of SpaceX Falcon 9 first stages using historical launch data, leveraging Python, SQL, visualization tools, and machine learning.
- Data was collected through the SpaceX public API and Wikipedia web scraping, followed by extensive data cleaning, integration, and feature engineering.
- Exploratory Data Analysis (EDA) was conducted using both SQL queries and Python visualizations, while interactive tools like Folium and a Plotly Dash dashboard were built for deeper insights.
- Several classification models were developed and tuned to predict landing outcomes.

Summary of all results

- Analysis revealed that payload mass, launch site, and orbit type are strong predictors of Falcon 9 landing success.
- The project found a clear trend of increasing landing reliability over time, with some sites and mission types achieving much higher success rates.
- Interactive tools and dashboards made these patterns easily accessible. Among tested models, Decision Tree and Logistic Regression showed the highest accuracy, successfully distinguishing between successful and failed landings. These insights help clarify which factors most impact SpaceX's reusable rocket strategy.

Introduction

Project background and context

- SpaceX Falcon 9 rockets have pioneered reusability, aiming to make spaceflight more affordable.
- Reliable landing of the first stage is critical for cost savings and future missions.
- Data-driven analysis and prediction can help engineers improve landing strategies.
- This project uses historical launch data to uncover trends and driving factors behind landing outcomes.

Problems you want to find answers

- What key factors influence successful Falcon 9 first-stage landings?
- How do launch site, payload mass, booster version, and orbit type affect landing outcome?
- Can machine learning models accurately predict landing success based on available data?
- Which trends or patterns can guide future mission planning and rocket development?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Launch data was extracted using the SpaceX public API, providing flight, payload, landing outcome, and site details.
- Additional attributes and mission outcomes were scraped from Wikipedia using Python web scraping tools.

Perform data wrangling

- The collected datasets were merged and cleaned.
- Data wrangling steps included handling missing fields, renaming columns for clarity, and engineering new features such as outcome labels and payload categories.

Perform exploratory data analysis (EDA) using visualization and SQL

- SQL queries were used to filter, aggregate, and summarize launch and landing statistics.
- Visualization libraries (Matplotlib, Seaborn, Plotly) helped uncover trends, patterns, and relationships in the data.

Methodology

Executive Summary

Perform interactive visual analytics using Folium and Plotly Dash

- Folium maps provided interactive visualization of launch sites and outcome distribution.
- A custom Plotly Dash dashboard enabled users to explore landing success patterns by site, payload, and orbit type.

Perform predictive analysis using classification models

- Multiple classification algorithms (Logistic Regression, Decision Tree, SVM, KNN) were trained and evaluated.
- Hyperparameters were tuned and cross-validation used for reliable model selection.
- Model performance was optimized, and the best classifier achieved high landing success prediction accuracy.

Data Collection

Data Collection and Integration Workflow for SpaceX Launch Information

STEP 1
API Request

Fetch launch data from SpaceX API and store locally

STEP 2
Web Scraping Wikipedia

Extract and parse launch data from Wikipedia HTML table

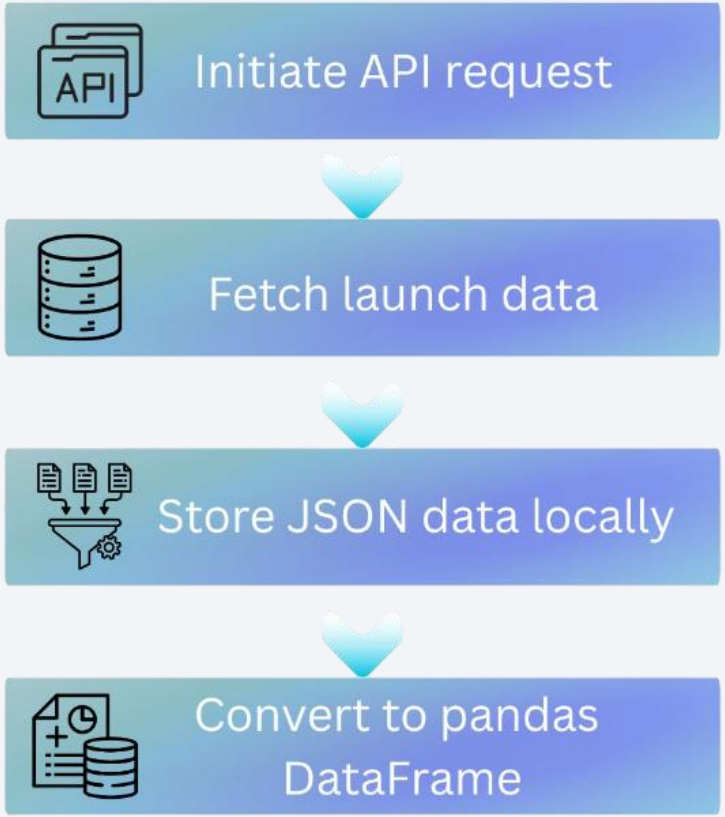
STEP 3
Data Integration

Merge SpaceX API and Wikipedia datasets into a final integrated dataset

Data Collection – SpaceX API

- Accessed SpaceX REST API to fetch Falcon 9 launch data with Python (requests library)
- Retrieved mission details: launch site, payload, rocket info, landing outcome
- Saved all response data as raw JSON, then processed into pandas DataFrame
- Ensured all historical launches included for completeness
- Used direct API calls for up-to-date, structured data

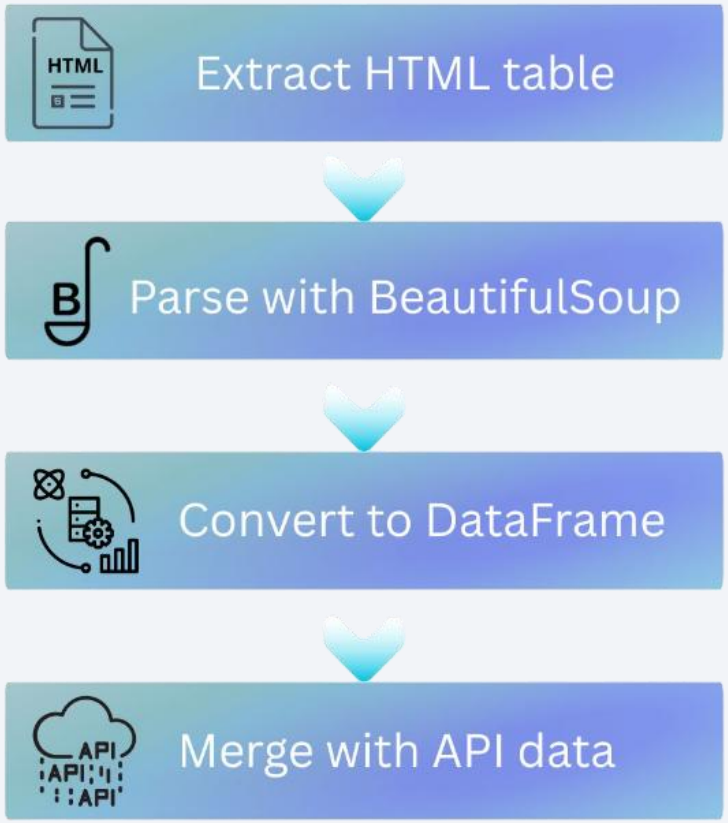
- GITHUB URL - 1_Spacex_API_Data_Collection.ipynb



Data Collection - Scraping

- Used Python (requests + BeautifulSoup) to access SpaceX Falcon 9 Wikipedia page
- Located and extracted HTML launch tables containing mission details unavailable in the API
- Parsed and transformed table data to pandas DataFrame
- Merged scraped data with API data for enhanced accuracy

- GITHUB URL - 2_Data_Collection_WebScraping.ipynb



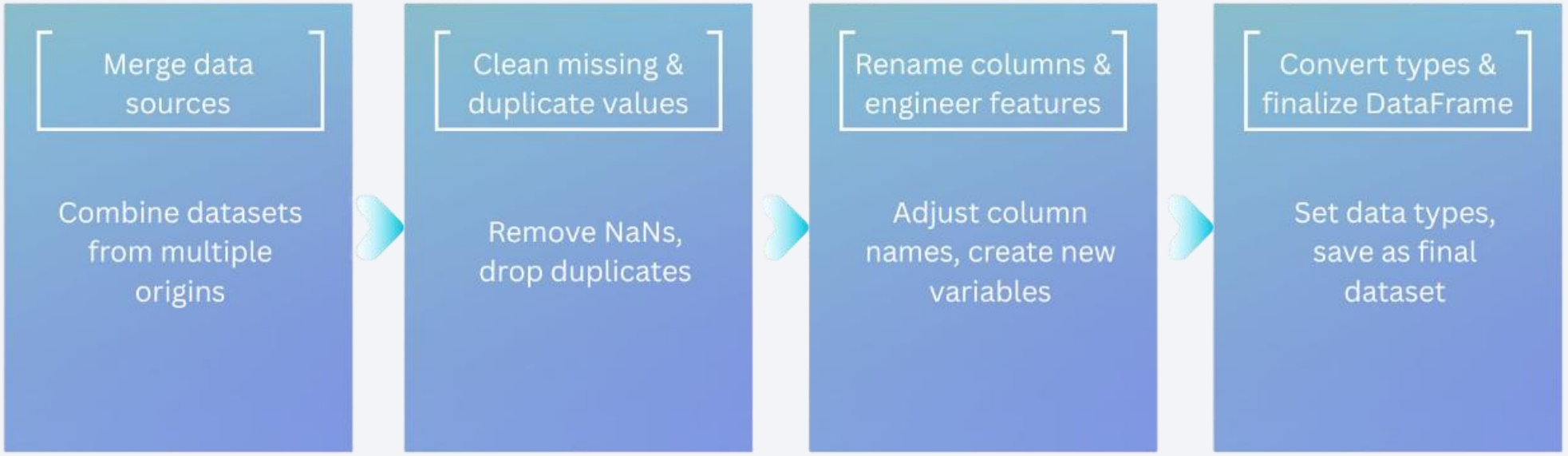
Data Wrangling

- Combined SpaceX API and Wikipedia datasets, ensuring all relevant launch data was included.
- Checked for missing values, filled or removed NaNs, and dropped duplicate entries.
- Renamed columns for clarity and standardized formatting.
- Filtered out irrelevant records, keeping only valid launch information.
- Created new columns for landing outcome categories and payload mass bins.
- Converted column types to numeric, date, and categorical where needed.
- Double-checked DataFrame consistency before proceeding to EDA.

- GITHUB URL – 3_Data_Wrangling.ipynb

Data Wrangling

Data Wrangling Workflow: Cleaning and Transforming Raw Launch Data



EDA with Data Visualization

- Plotted bar charts for landing outcomes by launch site.
- Used scatter plots to explore correlation between payload mass and landing success.
- Created pie charts to visualize proportions of successful vs failed landings.
- Used time series plots to show landing success over years.
- Plotted interactive Folium maps to visualize the geographic distribution of launch sites.
- Visualized booster version categories to study their effect on landing outcomes.
- Compared success rates across different orbits for further feature analysis.
- Each chart revealed trends, relationships, and data quality issues, helping guide machine learning model building.

- GITHUB URL - 5_EDA_with_Visualization.ipynb

EDA with SQL

- Extracted unique launch sites.
- Filtered launches above/below certain payload thresholds.
- Counted successful and failed landings per year.
- Grouped missions by orbit and calculated average payload mass.
- Ranked launch sites by number of successful landings.
- Used SQL queries to analyze booster version impact on success rates.
- Joined tables to enrich launch records with site, payload, and outcome info.
- Identified gaps, inconsistencies, and cleaned data for better modeling.
- SQL findings directly influenced which features and columns were used for machine learning.

- GITHUB URL - 4_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Added markers for each launch site to pinpoint their exact geographic locations on the map.
- Drew circles around launch sites to visualize safety zones and operational proximities.
- Connected related sites and nearby features with lines to show possible logistic or spatial relationships.
- The interactive map helps users visually explore SpaceX launch activity, site distribution, and safety considerations.
- Each object (marker, circle, line) enhances geographic context, enabling better understanding of mission planning and risk areas.

- GITHUB URL - 6_Interactive Visual Analytics with Folium.ipynb

Build a Dashboard with Plotly Dash

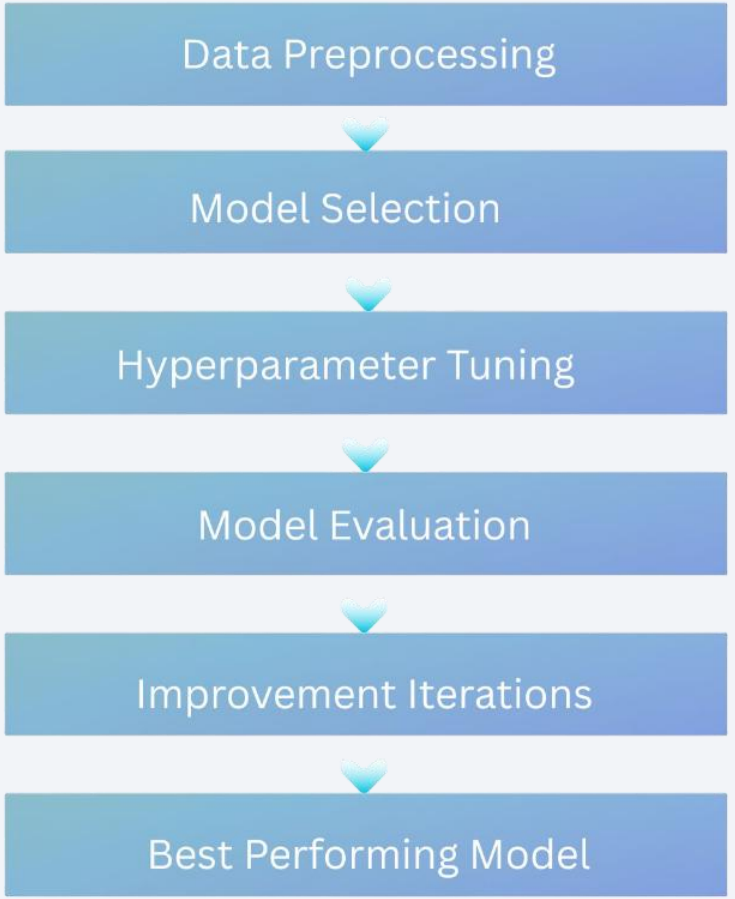
- Created interactive charts including pie charts displaying successful vs failed launches and scatter plots showing payload mass effect on launch success.
- Implemented dynamic dashboard controls such as launch site dropdown menus and payload range sliders for user-driven filtering and comparative analysis.
- Enabled detailed exploration of SpaceX launch trends, empowering stakeholders to make data-driven decisions.
- Incorporated responsive design for easy accessibility and mobile-friendly viewing.
- Added annotations and tooltips on charts to improve interpretability and highlight key data points.

- GITHUB URL - spacex_dash_app.py

Predictive Analysis (Classification)

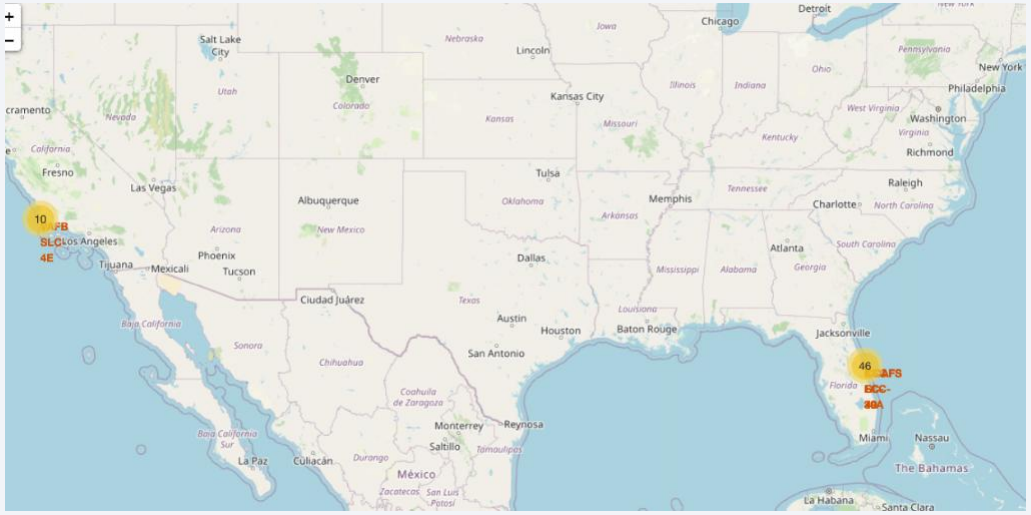
- Developed multiple classification models to predict SpaceX launch success using mission features.
- Standardized and split the dataset for robust training and testing.
- Evaluated various algorithms, tuning hyperparameters for optimal accuracy and reliability.
- Used metrics like F1-score and confusion matrix to compare model performance.
- Selected and improved the best model through iterative validation and adjustment

- GITHUB URL - 7_SpaceX_Machine_Learning_Prediction.ipynb



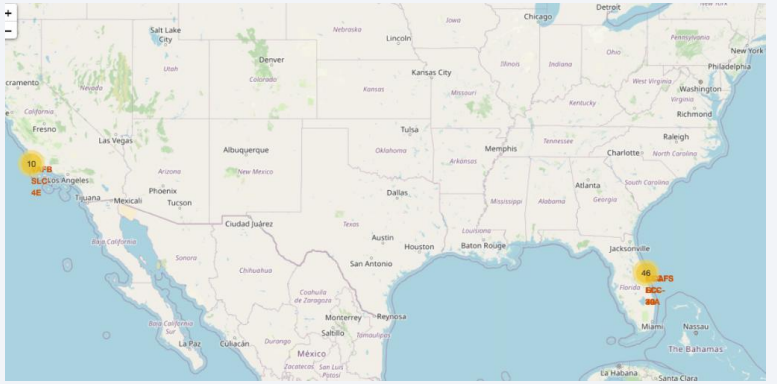
Results

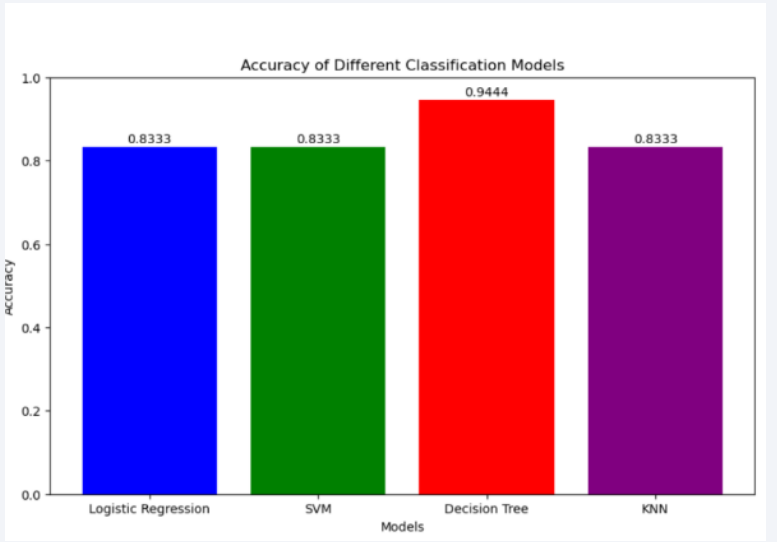
- Results of Exploratory data analysis results
- SpaceX operates from four main launch sites, with KSC LC-39A having the highest number of successful launches. East coast locations are most frequently used, and mission success rates have improved over the years.
- Heavier payloads and advanced booster versions have shown consistent success in recent years.
- Interactive analytics demo in screenshots,
- Predictive analysis results



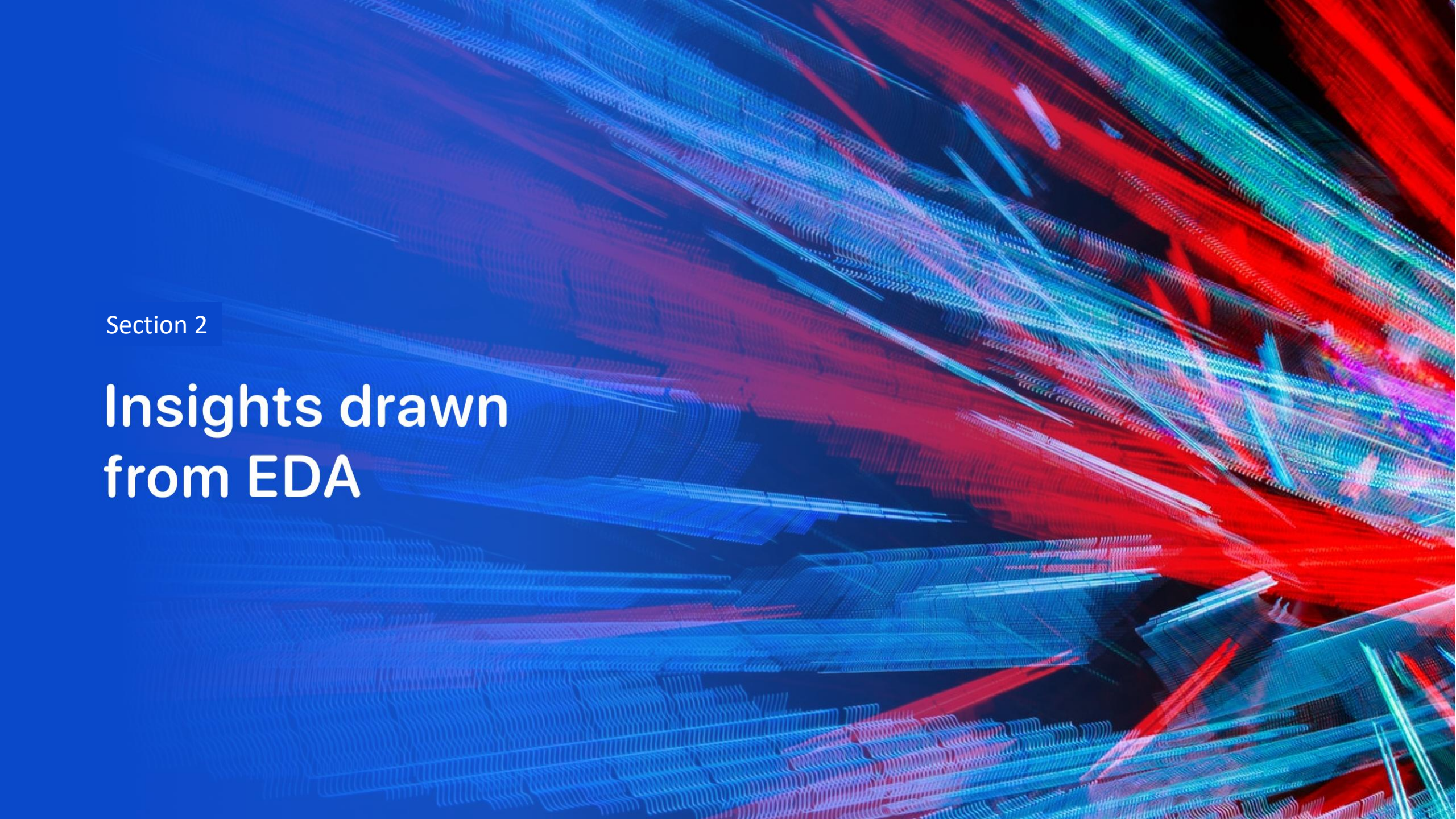
Results

- Results of Exploratory data analysis results
- SpaceX operates from four main launch sites, with KSC LC-39A having the highest number of successful launches.
- East coast locations are most frequently used, and mission success rates have improved over the years.
- Heavier payloads and advanced booster versions have shown consistent success in recent years.
- Interactive dashboards and maps reveal the geographic distribution and outcomes of launches.





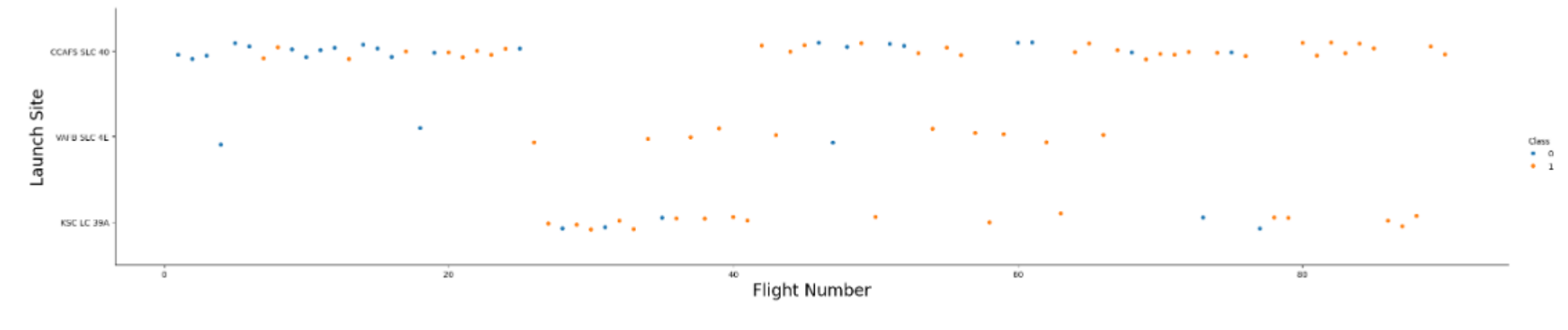
- Predictive analysis results
- Predictive analysis found Decision Tree Classifier to be the most accurate, with over 87% training and 94% test accuracy in predicting landing success


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- This plot explores how flight numbers correspond with launch sites.

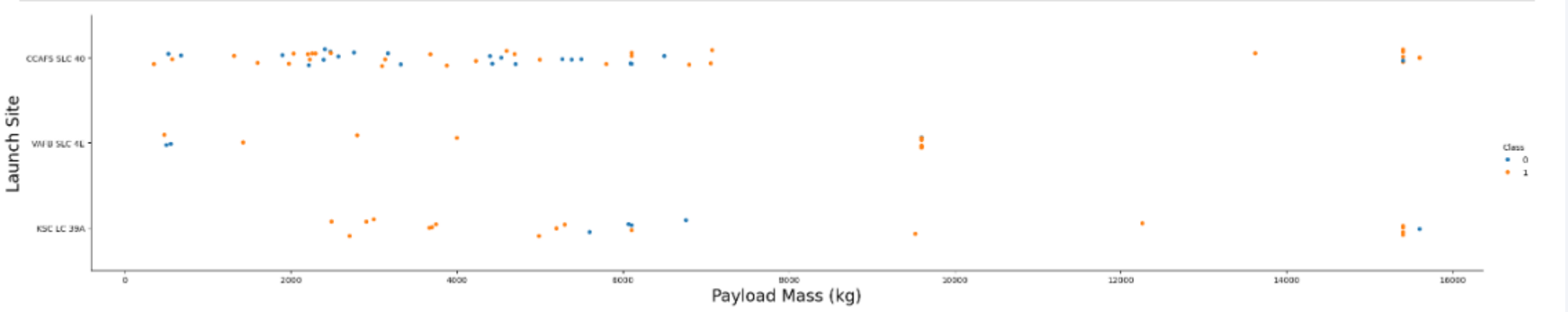
- It demonstrates the usage frequency of SpaceX's major launch facilities over time, highlighting regular sites for launch operations and success patterns.



Payload vs. Launch Site

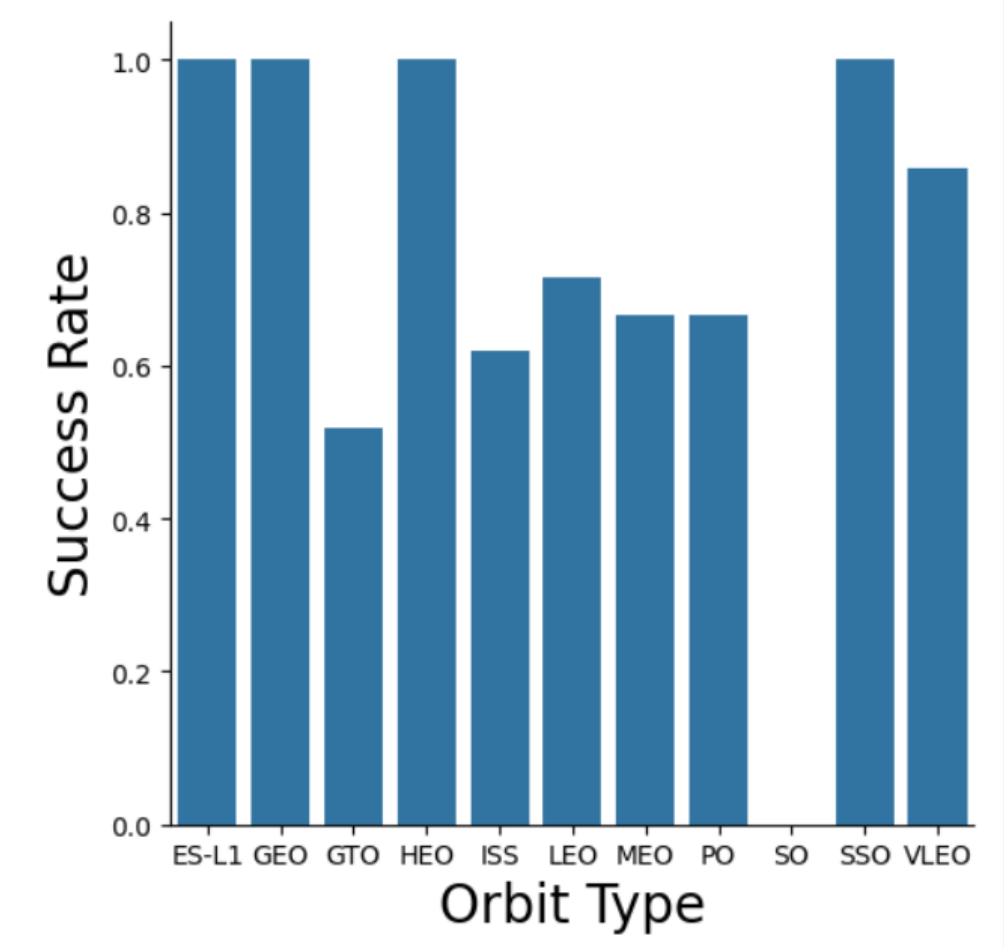
- The scatter plot analyzes the relationship between payload mass (kg) and launch sites.

- Payloads are distributed across multiple launch sites, showing no single site dominates heavy payload launches, and successful missions (Class 1) are widely represented.



Success Rate vs. Orbit Type

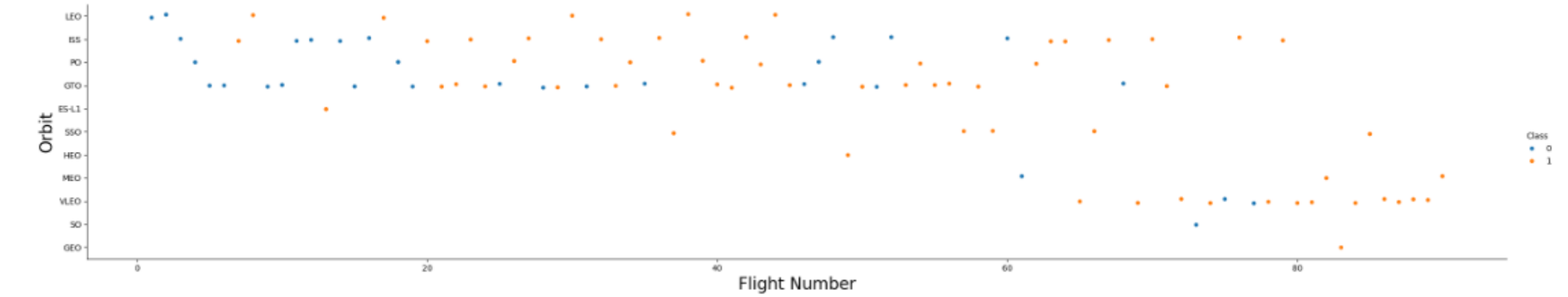
- The bar chart shows the success rate for each orbit type.
- GEO and GTO orbits have the highest success rates, while some orbits such as ISS show moderate performance. This analysis helps identify which missions are most reliable based on destination orbit.



Flight Number vs. Orbit Type

- This scatter plot visualizes each flight's number along the x-axis and mission orbit type along the y-axis.

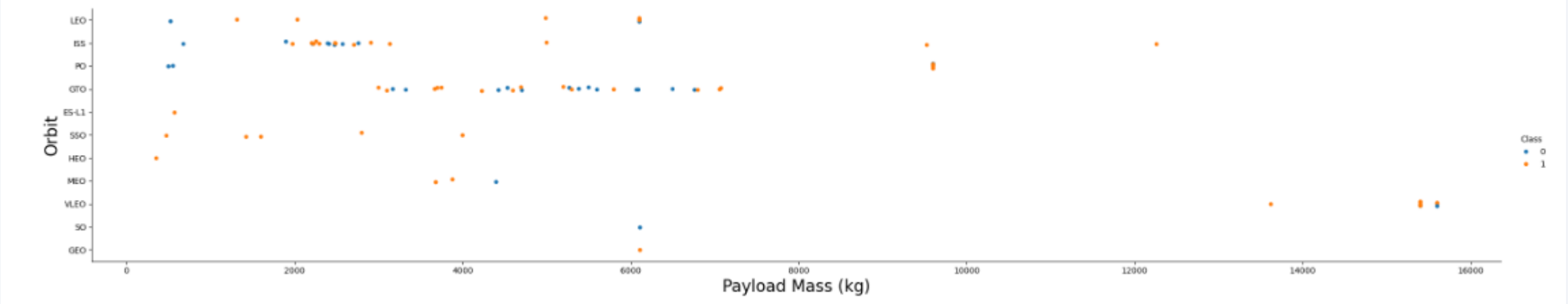
- It helps track SpaceX's mission diversity and highlights how certain orbit types are favored at different launch stages throughout Falcon 9's history.



Payload vs. Orbit Type

- The chart shows payload mass (kg) against orbit type.

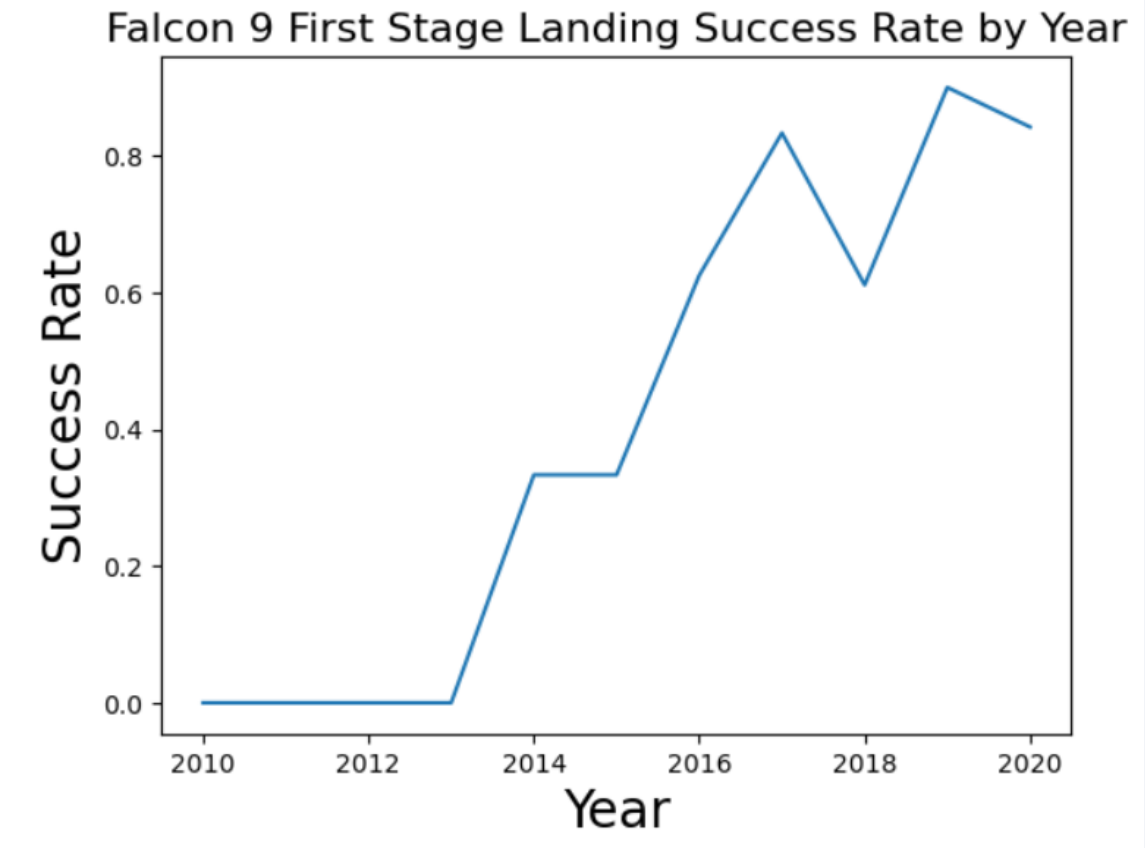
- Distribution reveals which orbits tend to receive heavier payloads and points out that high-mass payloads generally target specific orbits, aiding mission planning.



Launch Success Yearly Trend

- The line chart displays the yearly average success rate of Falcon 9 first stage landings from 2010 to 2020.

- You can observe that the success rate remained low until 2014, then steadily increased, peaking near 2018–2019, showing significant improvement in SpaceX's landing technology.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "launch_site" FROM SPACEXTBL;
```

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

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

- The query lists all unique launch site names from the dataset.
- This gives a quick reference to the facilities used for Falcon 9 launches, helping contextualize subsequent analyses by site.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT name FROM sqlite_master WHERE type='table';
```

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

name
SPACEXTBL
SPACEXTABLE

```
%sql select * from "SPACEXTBL" 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

- This query finds the first five launch records where the site name starts with "CCA".
- It demonstrates how to filter and preview data for specific launch sites, which often helps in analyzing site-based performance and trends.

Total Payload Mass

- The query sums total payload mass for all NASA (CRS) missions.

- A total of 0 suggests either the data subset excludes NASA or payload data for these missions wasn't recorded, highlighting a gap in the records.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PayloadMass") FROM SPACEXTBL WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

SUM("PayloadMass")
0.0

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried on missions by booster version F9 v1.1.

- A null or missing value suggests no matching records, perhaps due to data limitations or versioning issues.

Task 4

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

```
%sql SELECT AVG("PayloadMass") FROM SPACEXTBL WHERE "BoosterVersion" = 'F9 v1.1';
```

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

AVG("PayloadMass")
None

First Successful Ground Landing Date

- This query identifies the earliest date of a successful landing outcome on a ground pad.

- The result, December 22, 2015, marks a milestone for SpaceX as their first ground pad landing, a significant technical achievement.

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query lists booster versions with successful drone ship landings and payloads in the 4000–6000 kg range.

- It highlights which boosters were able to land successfully under these specific challenging payload conditions.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXTBL where landing_Outcome = 'Success (drone ship)' and payload_mass__kg_
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The query calculates the total successes and failures across all mission outcomes.

- A high total success number compared to failures shows improvement in SpaceX's mission reliability.

Task 7

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query lists all booster versions that have carried the maximum recorded payload mass.
- The results show several flights with boosters recorded under the ‘BoosterVersion’ column, highlighting the reliability and capacity of these specific boosters for heavy payload missions.

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%sql SELECT "BoosterVersion" FROM SPACEXTBL WHERE "PayloadMass" = (SELECT MAX("PayloadMass") FROM SPACEXTBL);
```

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

"BoosterVersion"
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion
BoosterVersion

2015 Launch Records

- The query lists failed drone ship landings, their booster versions, and site names for launches in 2015.

- This pinpoints the months and specific launches where drone ship landings did not succeed, useful for year-over-year reliability analysis.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT SUBSTR("Date", 6, 2) AS "Month", "BoosterVersion", "LaunchSite", "Landing_Outcome" FROM SPACEXTBL WHEF
```

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

Month	"BoosterVersion"	"LaunchSite"	Landing_Outcome
01	BoosterVersion	LaunchSite	Failure (drone ship)
04	BoosterVersion	LaunchSite	Failure (drone ship)

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

- The SQL query ranks all landing outcomes within the selected dates by their occurrence count.
- The most frequent outcome was "No attempt", followed by equal counts of "Success (drone ship)" and "Failure (drone ship)". This provides insight into which outcomes were most common during SpaceX's earlier launches.

Task 10

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.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS "OutcomeCount" FROM SPACEXTBL WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
```

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

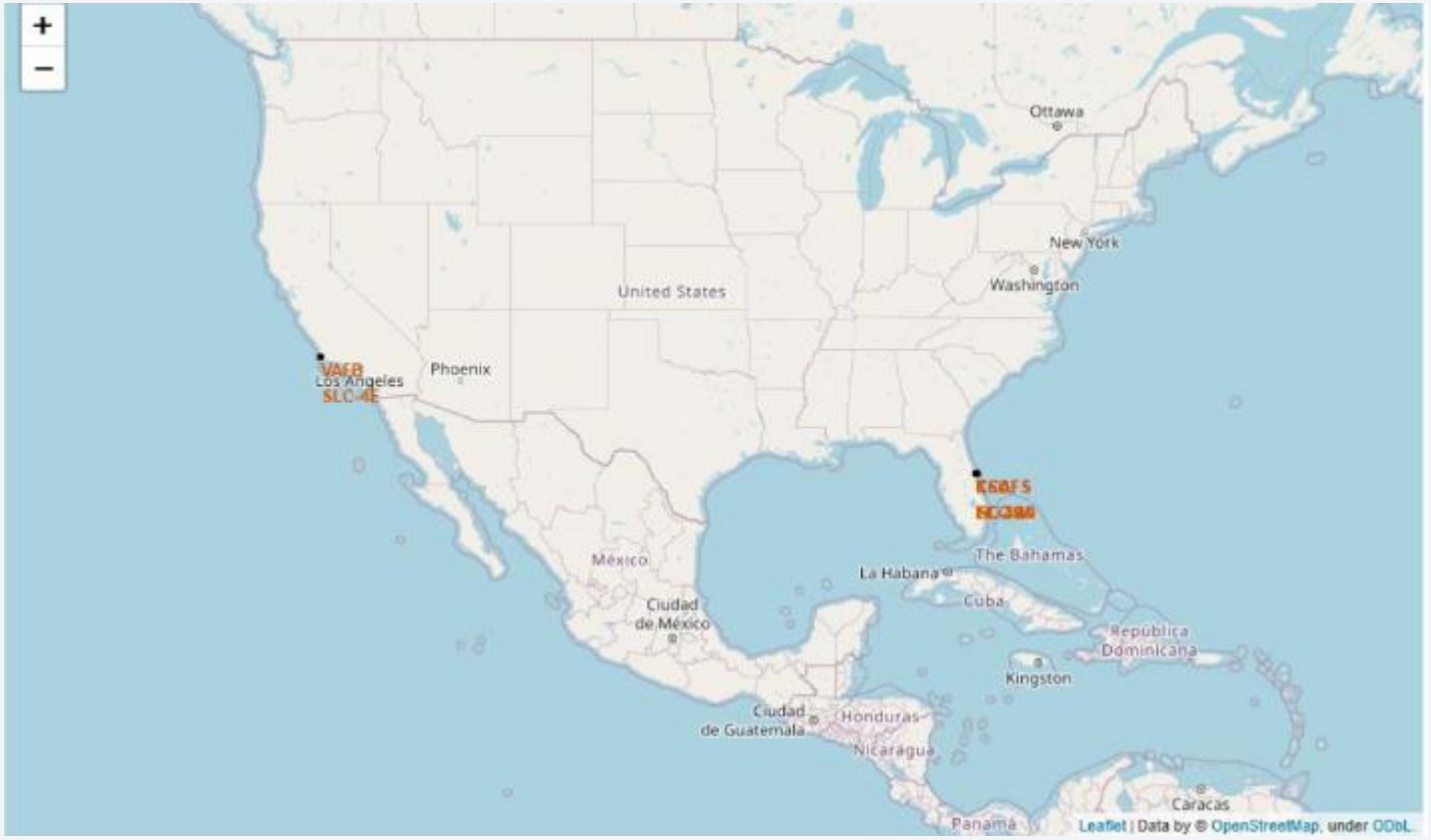
Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites

Launch sites are near sea, probably by safety, but not too far from roads and railroads.



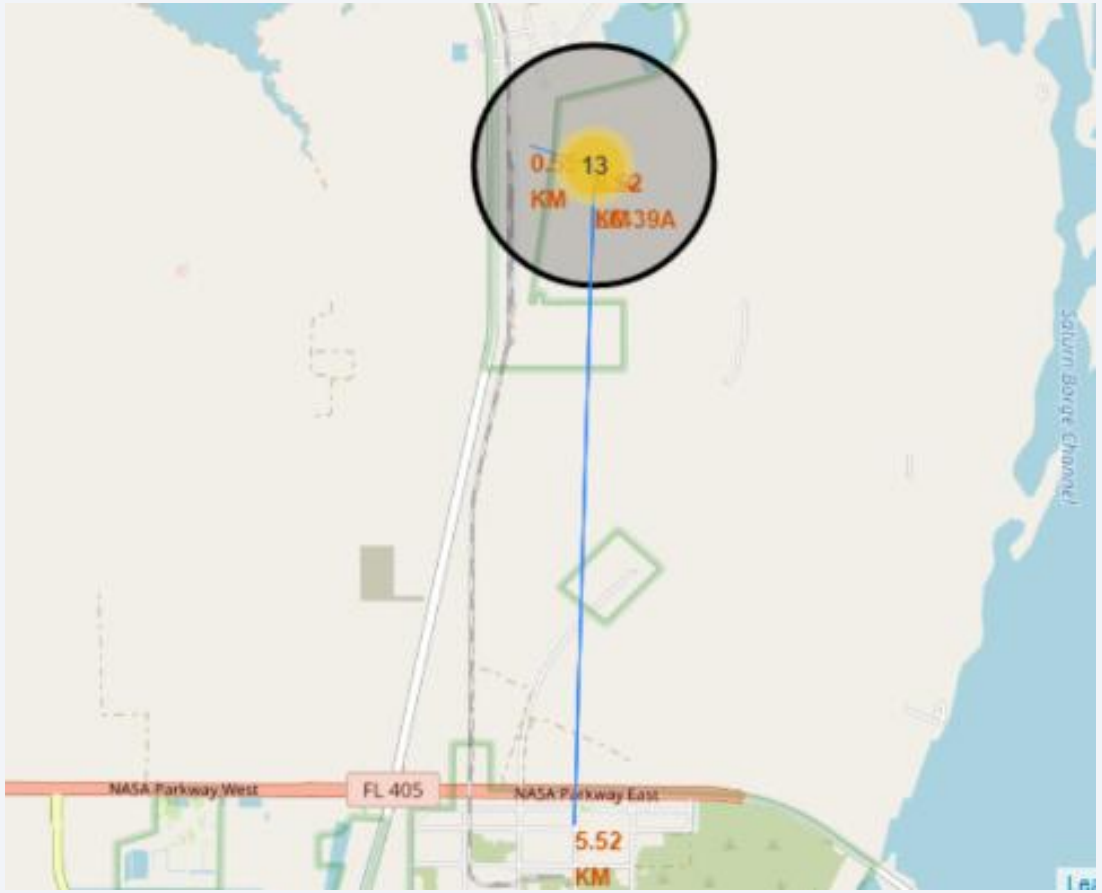
Launch Outcomes by Site

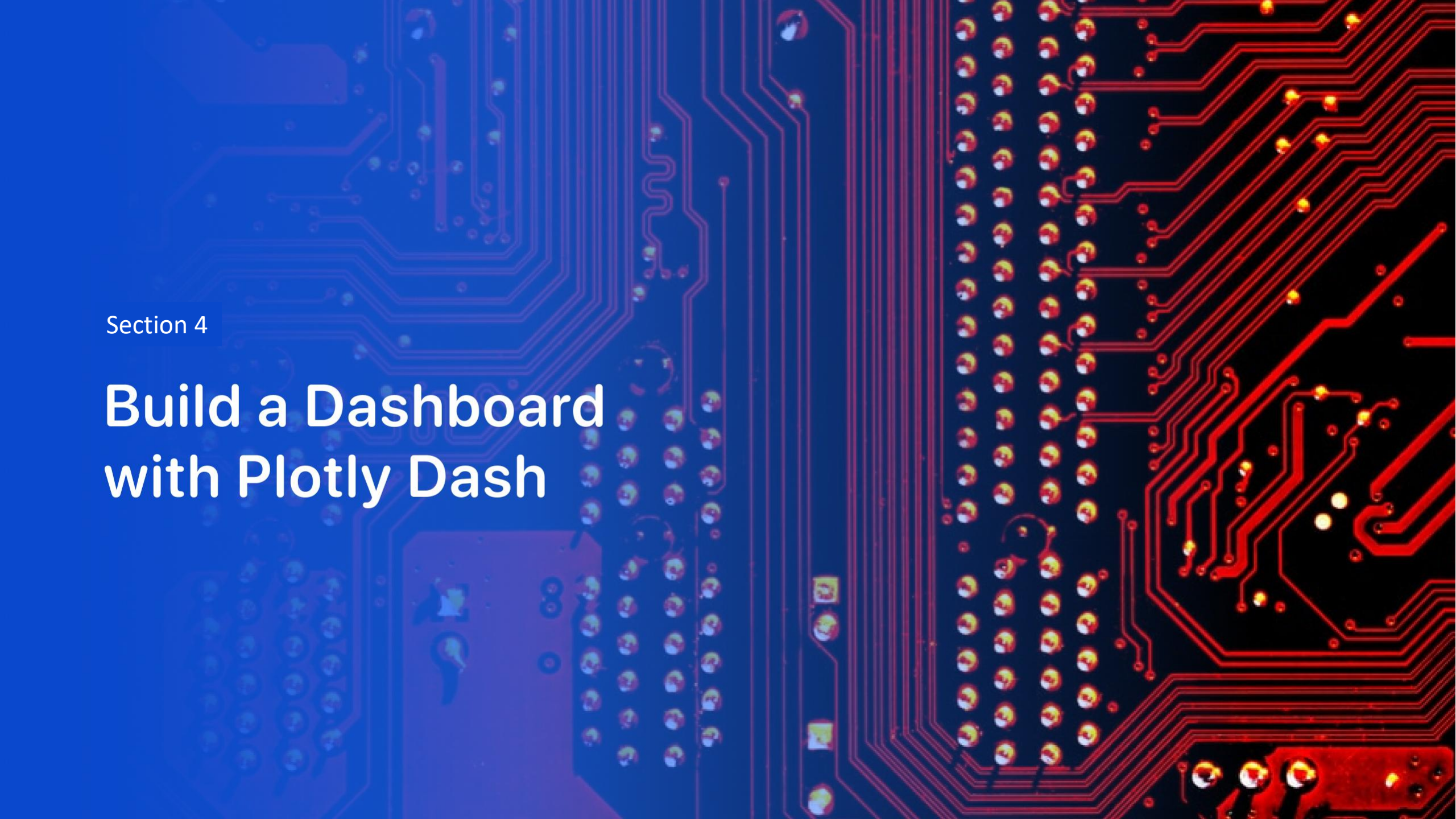
Example of KSC LC-39A launch site launch outcomes

Green markers indicate successful and red ones indicate failure

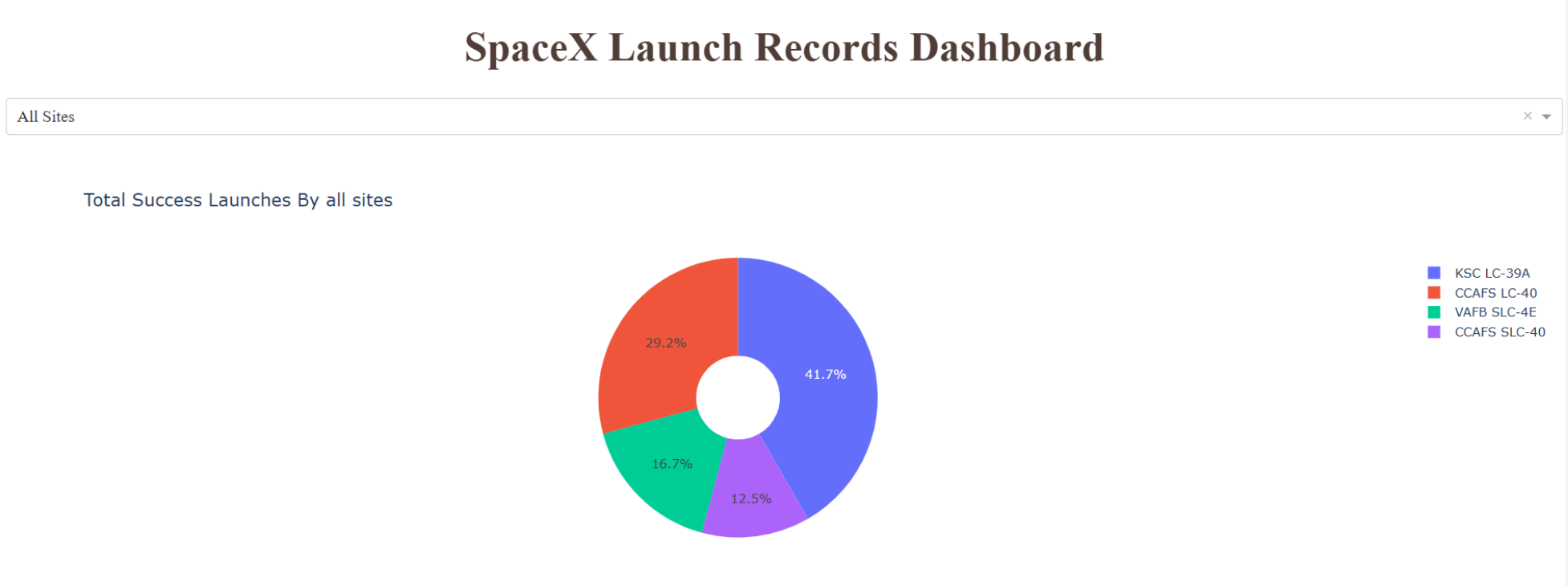
Logistics and Safety

Launch site KSC LC-39A has logistics aspects, being near railroad and road and relatively far from inhabited areas.



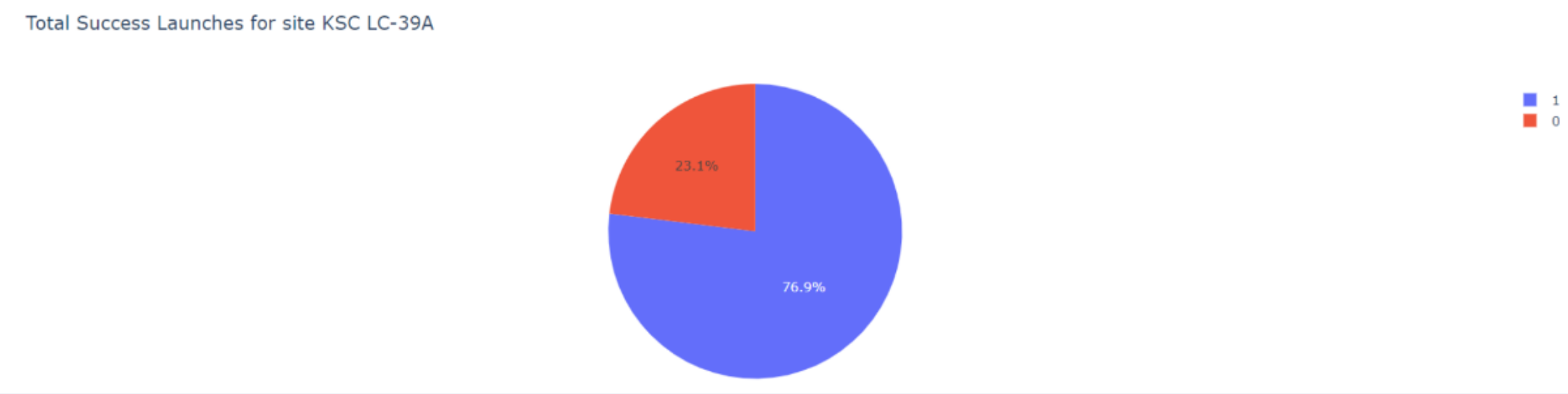

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



- The place from where launches are done seems to be a very important factor of success of missions
- The percentages help quickly compare the relative performance and utilization of each site.

Launch Success Ratio for KSC LC-39A



- The majority of launches from KSC LC-39A have been successful, with a 76.9% success rate.
- This demonstrates the strong reliability of this launch site for SpaceX missions.

Payload vs. Launch Outcome

- Most successful and failed launches are distributed across a range of payload masses.
- There is no clear relationship between payload mass and launch success; both classes occur throughout the payload range.

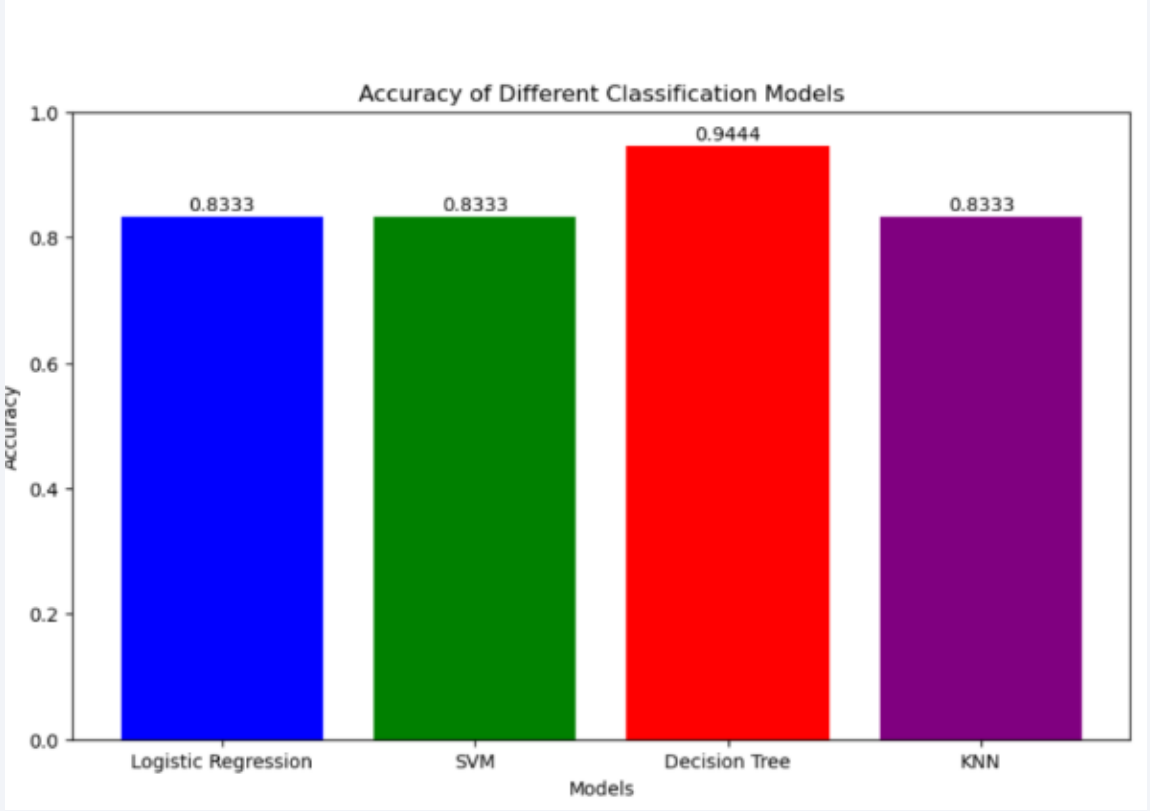
Section 5

Predictive Analysis (Classification)

Classification Accuracy

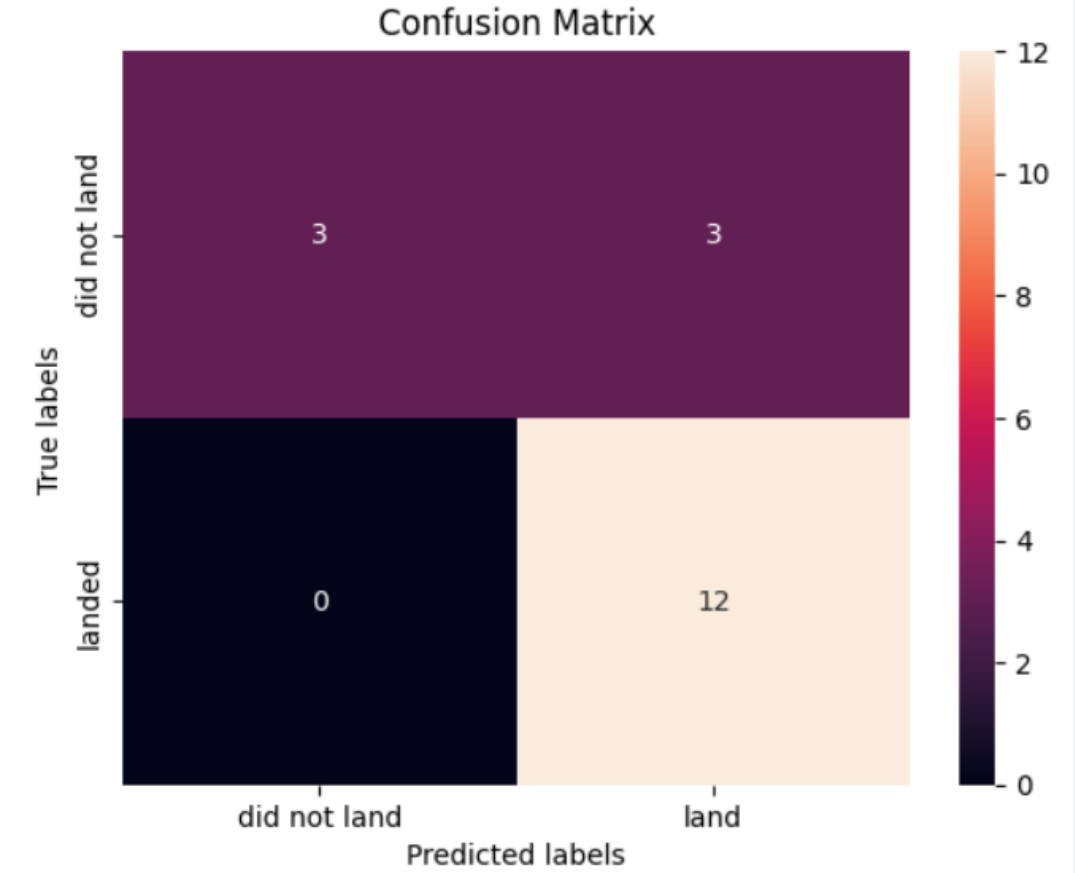
- The Decision Tree model achieved the highest accuracy of 94.4%, outperforming Logistic Regression, SVM, and KNN, each of which had an accuracy of 83.3%.

- This demonstrates that the Decision Tree is the most effective model for predicting SpaceX landing outcomes using the available features.



Confusion Matrix

- The confusion matrix shows the model correctly predicted all actual landings and made only three false positives for non-landings.
- This indicates strong performance, especially in predicting successful landings, with only minor misclassifications among non-landed cases.



Conclusions

- This study utilized exploratory data analysis and interactive visualization techniques to uncover key insights into SpaceX launch operations. Four primary launch sites were identified, with KSC LC-39A demonstrating the highest launch success rate. Launches with lower payload masses generally yielded better results, though factors like booster versions and launch site conditions also played significant roles.

- The analysis showed that most launch sites are near the equator and coastal areas, optimizing safety and logistics. Importantly, success rates have progressively improved over the years, reflecting advancements in technology and operational procedures.

- Predictive modeling identified the Decision Tree Classifier as the most effective tool, achieving over 94% accuracy in test data. This model offers a robust method to forecast launch success, potentially aiding in strategic planning and operational efficiency.

- Overall, the combination of data-driven insights and predictive analytics provides a valuable framework to enhance decision-making and contribute to the ongoing success of reusable rocket technology at SpaceX.

Appendix

- For complete code, data, and all project assets, see:
  GitHub Repository: Applied-DataScience-Capstone-Project

Special Thanks to :

Instructors
Coursera
IBM

Thank you!