IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Suman
20-11-2025

Outline

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

Executive Summary

- **Objective :** The goal of this project was to analyze SpaceX launch data to identify patterns and factors that contribute to mission success.

- **Approach:** We collected data via the SpaceX API and web scraping, performed extensive cleaning and exploratory analysis, built interactive maps and dashboards, and developed a predictive model for launch success.

- **Key findings:** Our analysis revealed that [e.g., launch site, payload mass, orbit type] are significant factors. The final predictive model achieved an accuracy of 91% and decision tree is best model.

- **Conclusion:** This project demonstrates the potential of data science to uncover valuable insights in the aerospace industry.

Introduction

- Project background and context
- SpaceX has revolutionized the space industry with its reusable rockets.
- Understanding launch success factors is crucial for cost reduction and mission planning.

- Problems you want to find answers
- Which launch sites have the highest success rates?
- How does payload mass and orbit type affect success?
- What is the trend of success over the years?
- Can we reliably predict the outcome of a launch?

Section 1

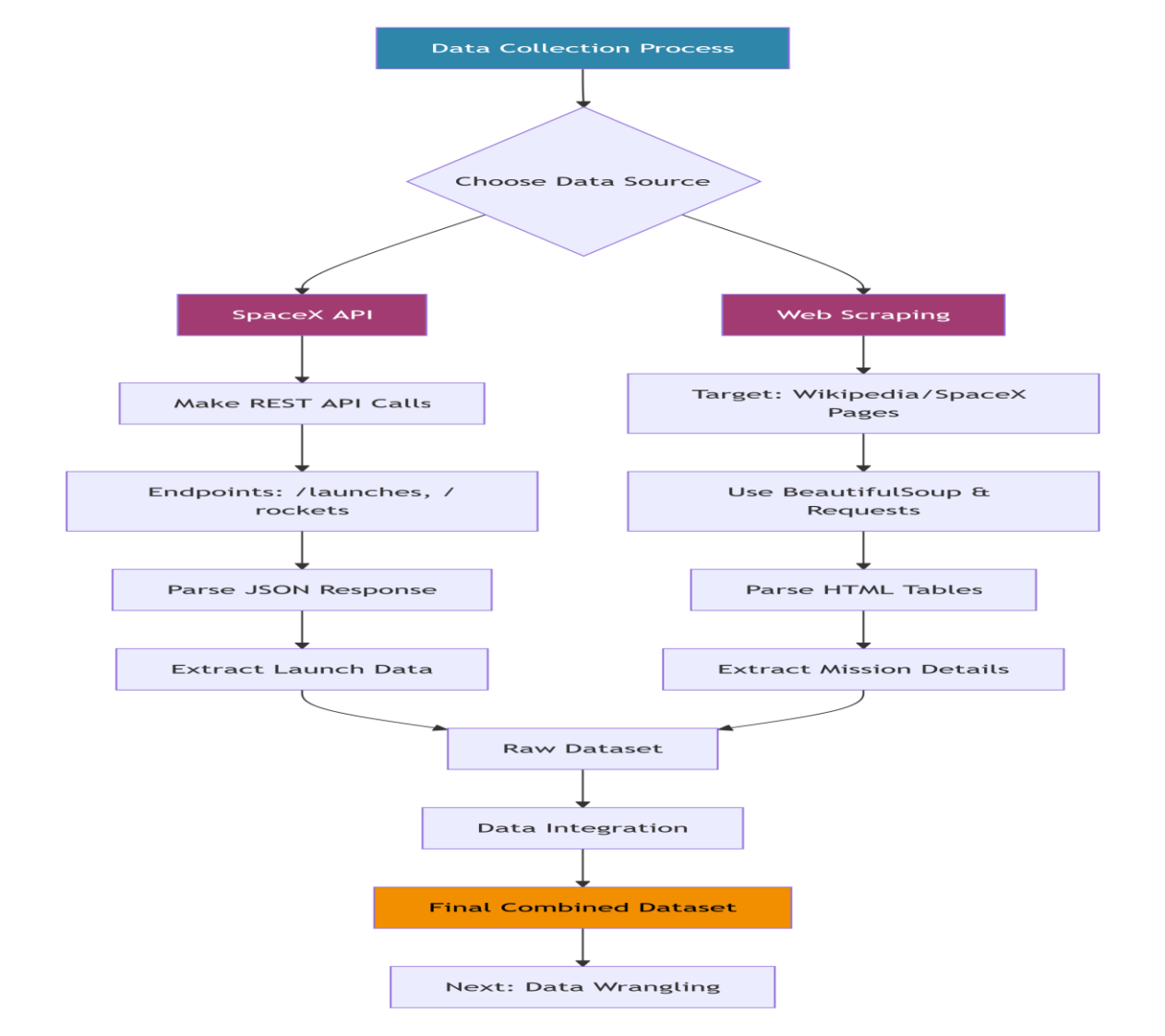
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Use SpaceX REST API to fetch launch data.
  - Send API request using python (requests). Received data in json format converted into pandas dataframe.
- Perform data wrangling
  - Selected only relevant field for analysis. Handled missing values. Removed duplicates and standardized column names.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - **Build:** Split data train multiple model(LR, KNN, SVM, Decision tree), Use **GridSearchCV** to find the best hyperparameters.
  - **Evaluate**: Use accuracy, confusion matrix, precision, recall, F1-score.
  - **Compare** models and pick the best one for prediction.

Data Collection

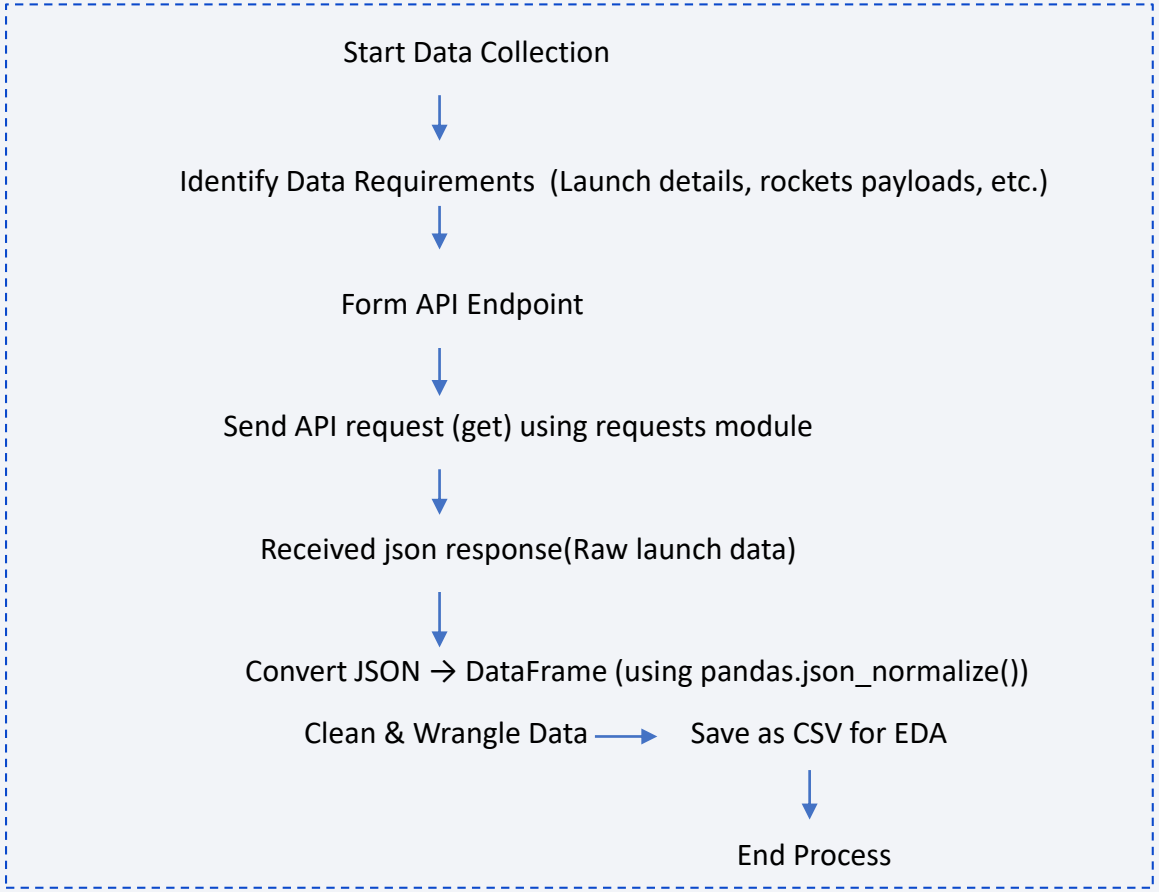


Data Collection – SpaceX API

- The dataset was collected using a combination of API requests and web scraping.
  The SpaceX API was used to retrieve launch records in JSON format, which were then converted into a pandas DataFrame.

GitHub URL:

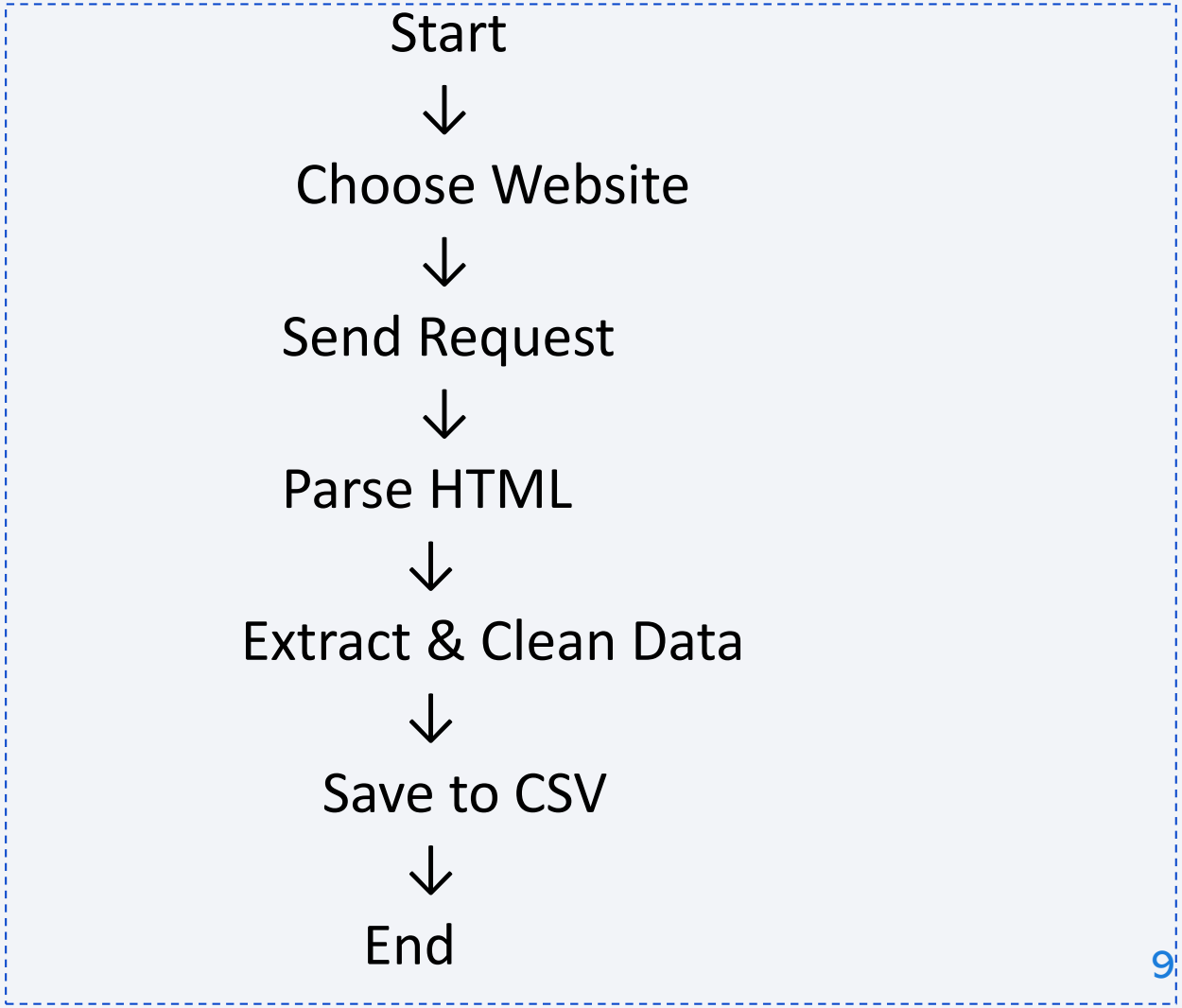
https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection – Web-Scraping

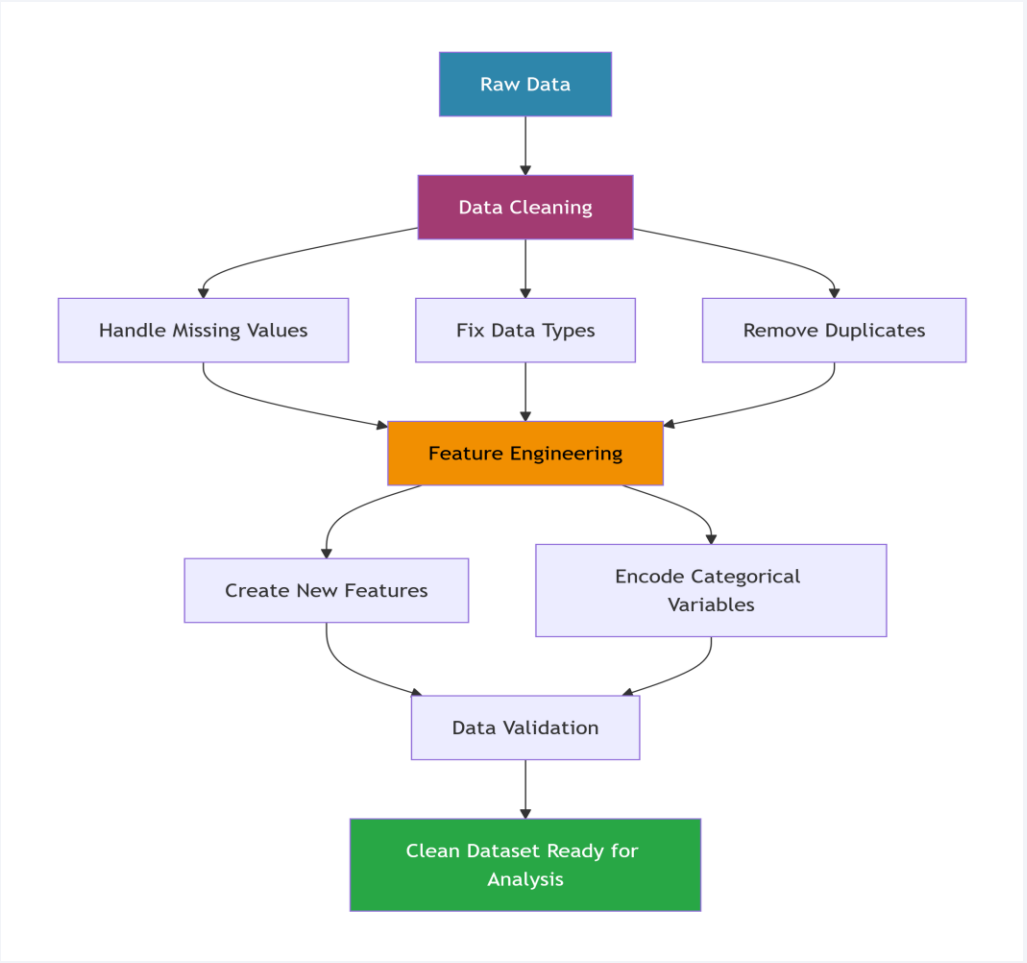
- Identified the official website containing the required SpaceX launch data.
- Inspected the webpage structure (HTML elements, tags, tables).
- Sent HTTP GET requests using Python's **requests** library.
- Parsed the returned webpage HTML using **BeautifulSoup**
- Extracted required data fields such as launch date, payload, mission outcome, and booster information.
- Cleaned the extracted data . Stored cleaned data in a **CSV file** for further analysis.
- GitHub URL:

https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/jupyter-labs-webscraping(2).ipynb



Data Wrangling

- Data cleaning
- Data transformation
- Feature Engineering
- Type conversion
- Missing value handling
- Outlier treatment
- Data Integration
- Dataset preparation
- **Github url:** https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/labs-jupyter-spacex-Data%20wrangling(3).ipynb



EDA with Data Visualization

- Bar Charts

Success rate by orbit type

Launch outcomes (Success vs Failure) :To compare **frequencies** easily.To identify which launch sites are used most often. To check if some sites have **higher success rates**.

- Scatter plot

Payload vs. Launch Outcome

Payload mass vs. Flight number

Orbit vs. FlightNumber

Launch site vs. payload mass

Orbit vs. Payload mass: To observe **relationships and correlations**.

To detect patterns (e.g., heavier payloads → more failures?).

- Github url: https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/edadataviz(5).ipynb

EDA with SQL

- **Identified unique launch sites** to understand all possible SpaceX launch locations.
- **Filtered launch sites starting with 'CCA'** to focus on Cape Canaveral launch operations
- **Calculated total payload mass for NASA (CRS) missions**, showing how much payload NASA launches through SpaceX.
- **Computed average payload mass for booster version F9 v1.1**, useful for performance comparison.
- **Found the date of the first successful ground-pad landing**, showing the milestone achievement.
- **Retrieved booster names with successful drone-ship landings** and payloads between 4000–6000 kg, useful for analyzing heavy-payload missions.
- **Counted successful vs failed miss**
- **Extracted 2015 records showing month name, failed drone-ship landings, booster version, and launch site**, focusing on time-based failure analysis. **ion outcomes** to measure overall mission performance.
- Github-url : https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/jupyter-labs-eda-sql-coursera_sqllite(4).ipynb

Build an Interactive Map with Folium

- Added **Circle Markers** to show success vs failure of launches (green/red).
- Added **Regular Markers with Popups** for launch info at each site.
- Used **MarkerCluster** to combine overlapping markers and keep the map clean.
- Drew **PolyLines** to show distances between launch sites and nearby points.
- Added **Distance Labels** (DivIcon) to display proximity measurements in KM.
- Placed **Markers** for coastline, highway, railway, and city coordinates to support distance calculations.
- GitHub URL : https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/lab_jupyter_launch_site_location(6%60).ipynb

Build a Dashboard with Plotly Dash

- Plots Added

Success Pie Chart: Shows total successes by site or success vs failure for a selected site.

Success–Payload Scatter Plot: Shows how payload mass relates to launch success.

- Interactions Added

Launch Site Dropdown:Filters charts by selected launch site.

Payload Range Slider:Filters scatter plot based on payload mass range.

- Why These Plots & Interactions Were Added

1. To compare launch performance across different launch sites.
2. To understand how payload mass affects success or failure.
3. To identify trends across booster versions.
4. To make the dashboard interactive, user-controlled, and analytical.

Github url: https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/Dashboard.png

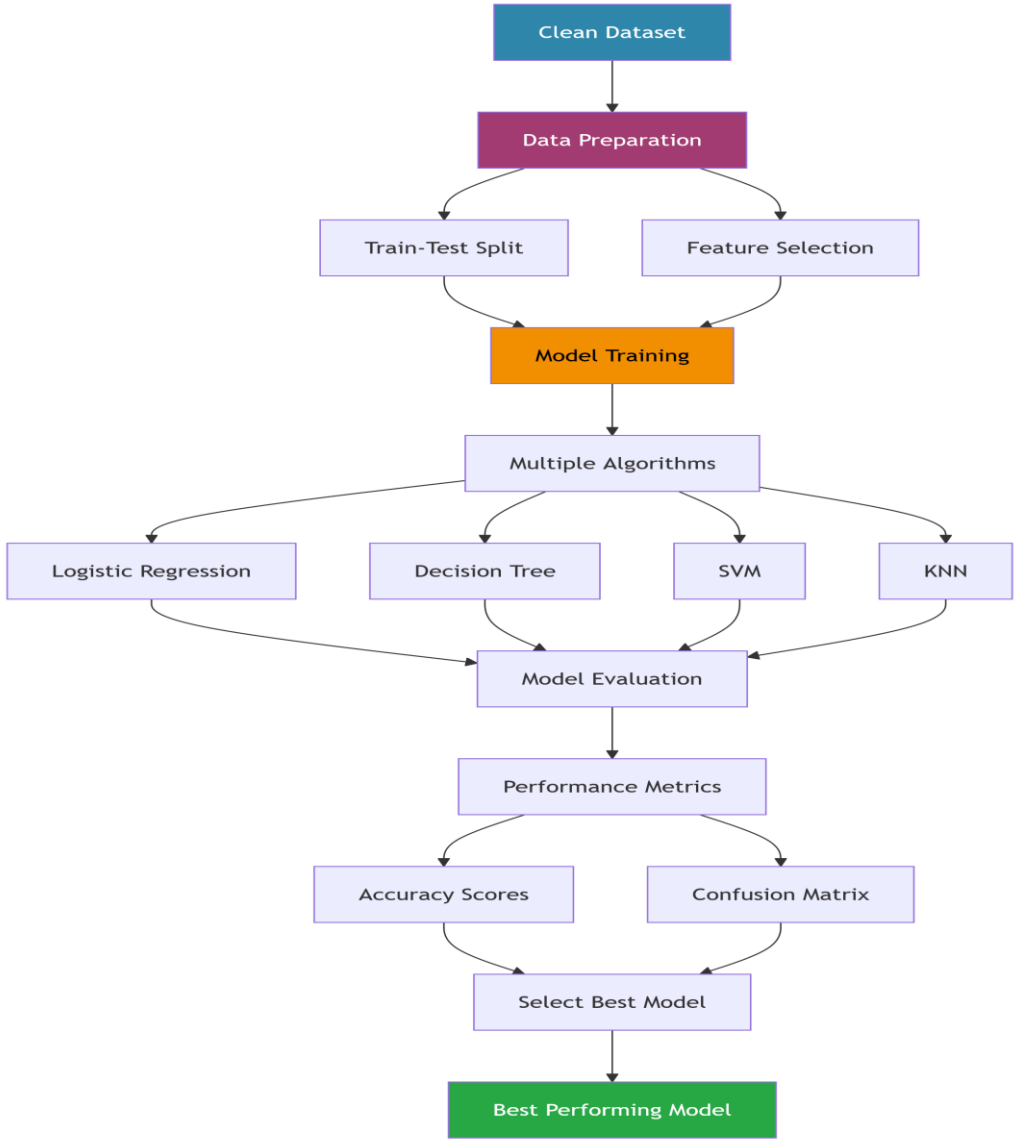
Predictive Analysis (Classification)

Model Development Summary

- Cleaned and prepared dataset for machine learning.
- Split data into training and testing sets.
- Built multiple classification models (LR, KNN, DT, SVM).
- Evaluated models using accuracy, confusion matrix & cross-validation.
- Tuned each model using GridSearchCV.
- Compared both default and optimized models.
- Selected the best-performing model based on accuracy & generalization.

Github url : https://github.com/Suman437596/Data-collection-SpaceX-API/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Flow chart of Predictive analysis



Results

- **Exploratory Data Analysis provided the following insights**

1. The success rate of launches has **steadily increased from 2013 to 2020**, showing major improvements in booster technology.
2. Pay **mass** does not strongly affect landing success, but very high payloads slightly reduce success rates.
3. FT booster versions show the **highest success rates**, while older versions (v1.0, v1.1) show more failures.
4. CCAFS SLC-40 and Kennedy LC-39A are among the most frequently used launch sites.

- **Interactive Analytics Demo**

1. The interactive dashboard allowed users to explore relationships dynamically:
2. A payload slider helped analyze how success rate varies across payload ranges.
3. A launch site dropdown displayed filtered visualizations showing success/failure distribution per site.
4. Interactive scatter plots visualized how booster version and payload influenced landing outcome.

- **Predictive Analysis Results**

Predictive analysis was performed using multiple classification models:

1. Trained models included **KNN, Decision Tree, SVM, and Logistic Regression.**
2. Accuracy scores were:

a) Decision Tree: 0.91 (best)

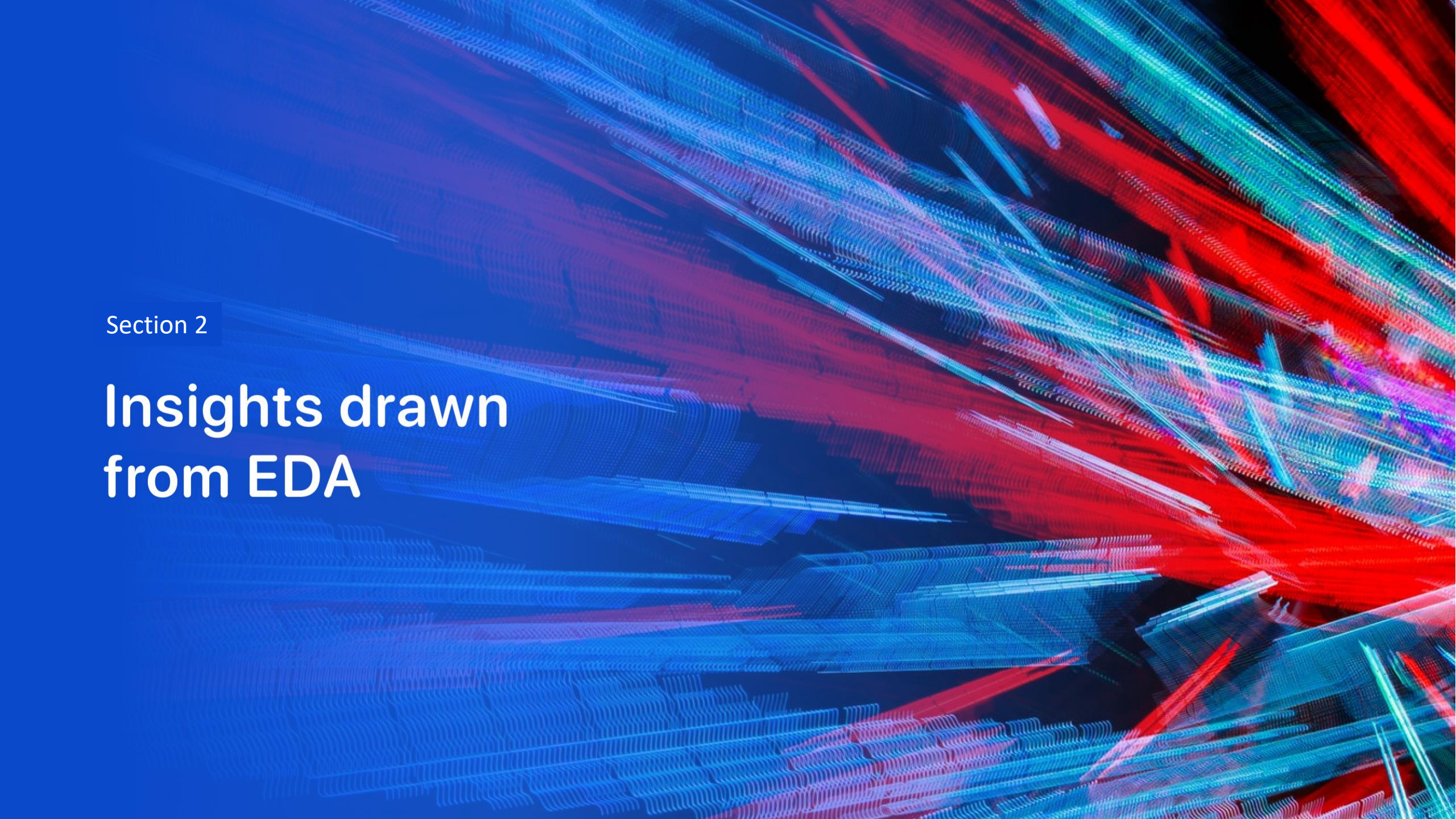
b) KNN: 0.84

c) SVM: 0.82

d) Logistic Regression: 0.82

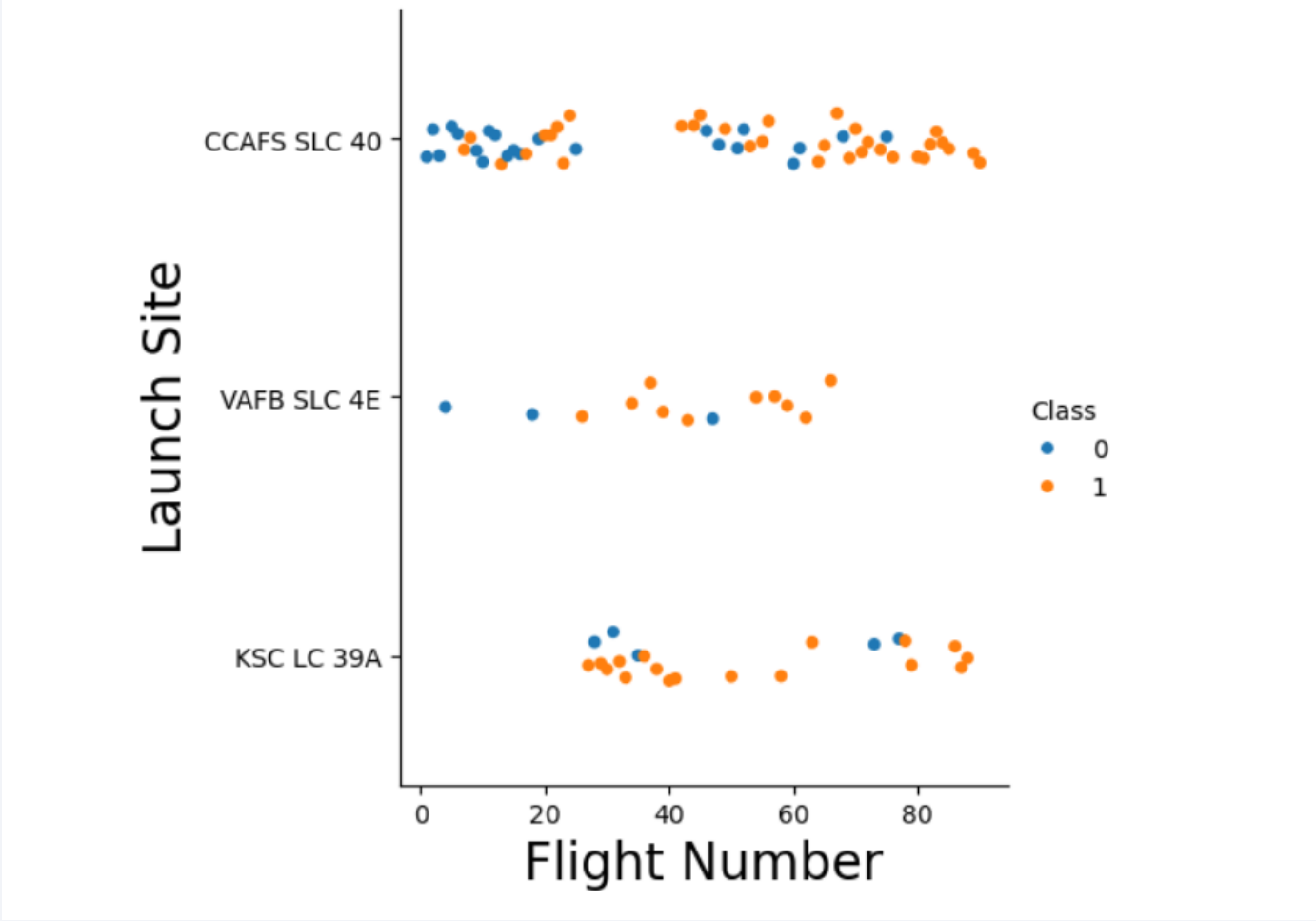
3. Hyperparameter tuning using **GridSearchCV** further improved the decision tree's stability.

4. Confusion matrix showed the model could **accurately identify most successful landings**, confirming high reliability.


Section 2
Insights drawn from EDA

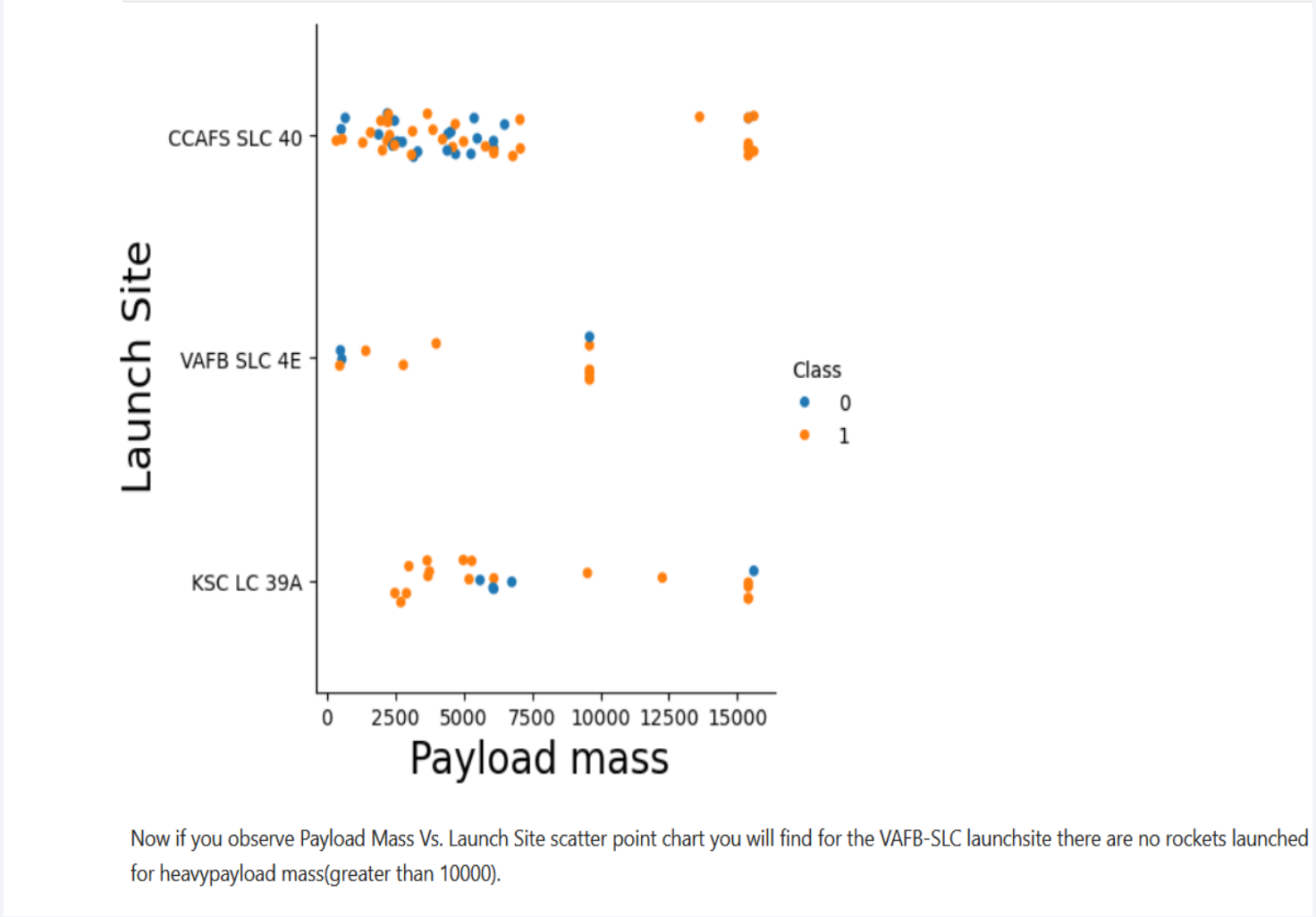
Flight Number vs. Launch Site

- **CCAFS SLC 40** has the most extensive launch history, with flights spanning from early missions to recent ones.
- **KSC LC 39A** started operations later but has seen consistent launch activity.
- **VAFB SLC 4E** has relatively fewer launches compared to the Florida based-site .
- The increasing flight number over time at all sites indicate SpaceX's growing launch cadence and operational experience.



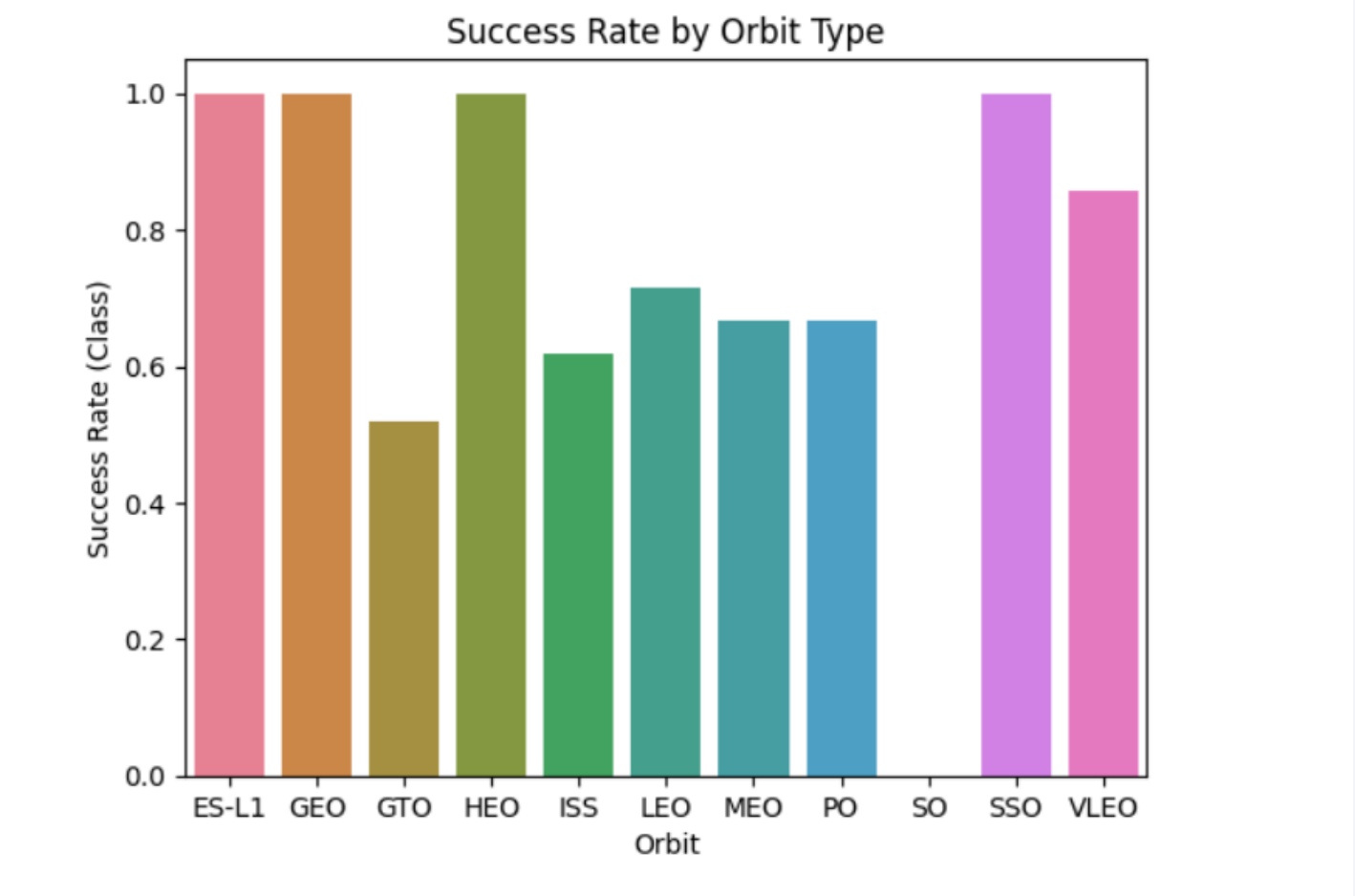
Payload vs. Launch Site

- **KSC LC 39A** handles the heaviest payloads.
- **CCAFS SLC 40** handles medium to heavy payloads.
- **VAFB SLC 4E** is limited to lighter payloads only.
- VAFB SLC 4E has no rockets launched with heavy payload mass (greater than 10,000) - this appears to be a payload capacity limitation at this site.
- Heavier payload missions show mixed success rates across all sites.
- Lighter payloads tend to have more consistent success, particularly at VAFB SLC 4E



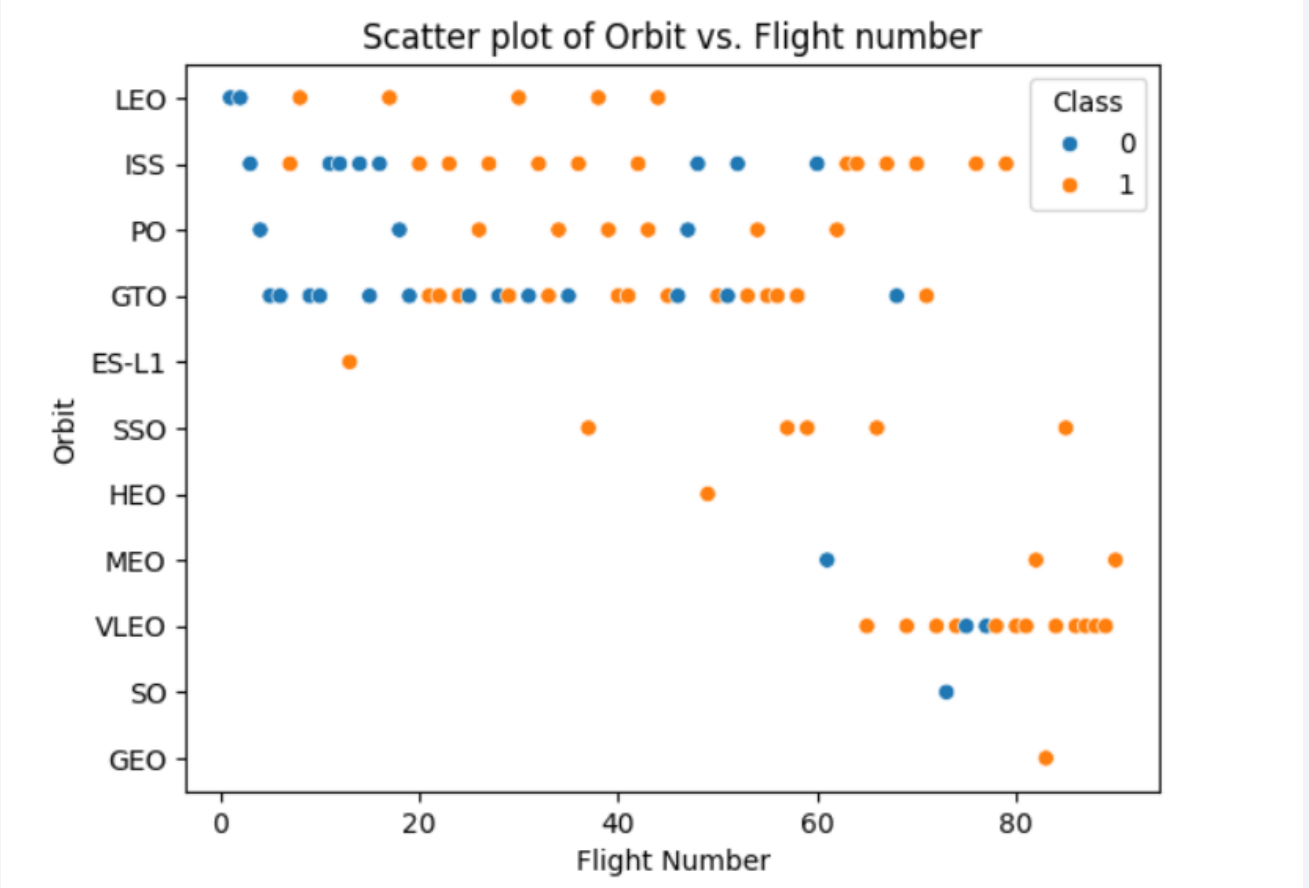
Success Rate vs. Orbit Type

- ES-L1 GEO, GTO, HEO, ISS, LEO, MEO, SO, SSO, VLEO all have 100% success rates. This indicates SpaceX has mastered launches to these common orbit types.
- PO stands out as the only orbit with significant failure rate. Polar Orbit missions have a success rate below 50%.
- Polar Orbit launches present unique technical challenges or higher risk factors. SpaceX may need to focus on improving reliability specifically for Polar Orbit missions.



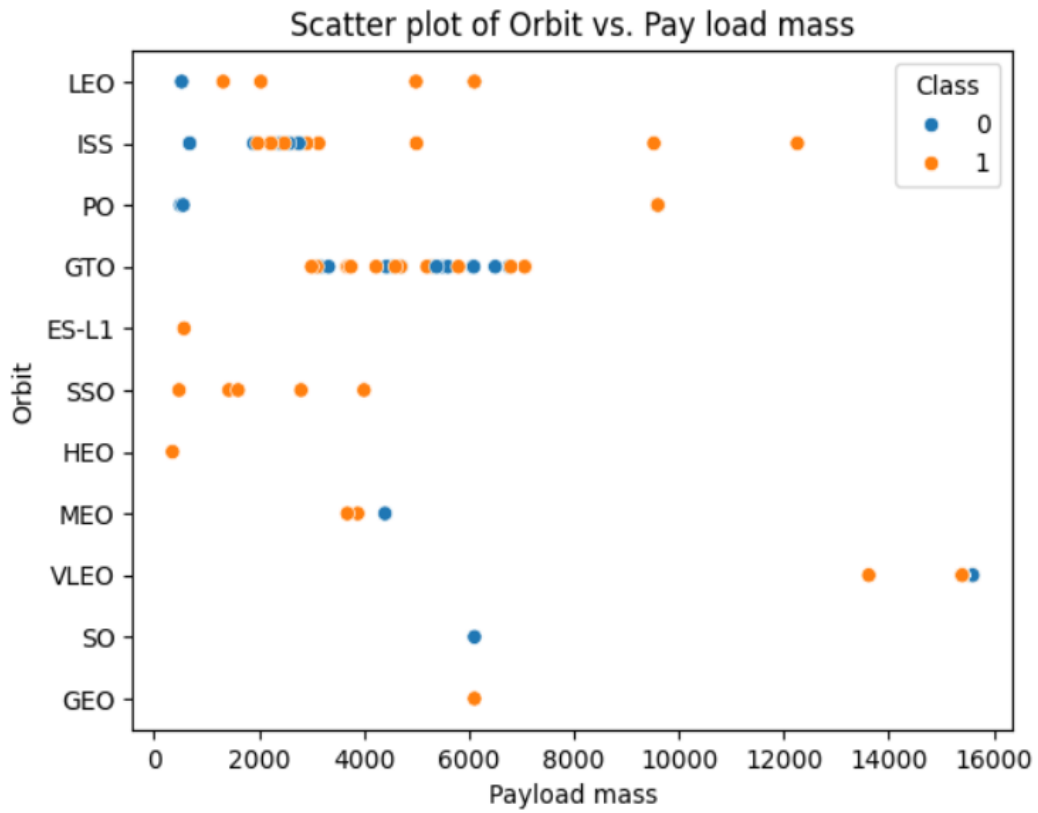
Flight Number vs. Orbit Type

- LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

- GTO handles the heaviest payloads.
- ES-L1, GEO, and LEO also handle significant payload weights.
- ISS ,PO , SSO typically handle medium-range payloads.
- VLEO, SO, MEO, HEO generally handle lighter payloads.
- PO, LEO, and ISS show strong success rates even with heavy payloads.
- GTO missions show mixed results both successes and failures across the payload range, making it difficult to predict outcomes based on mass alone.

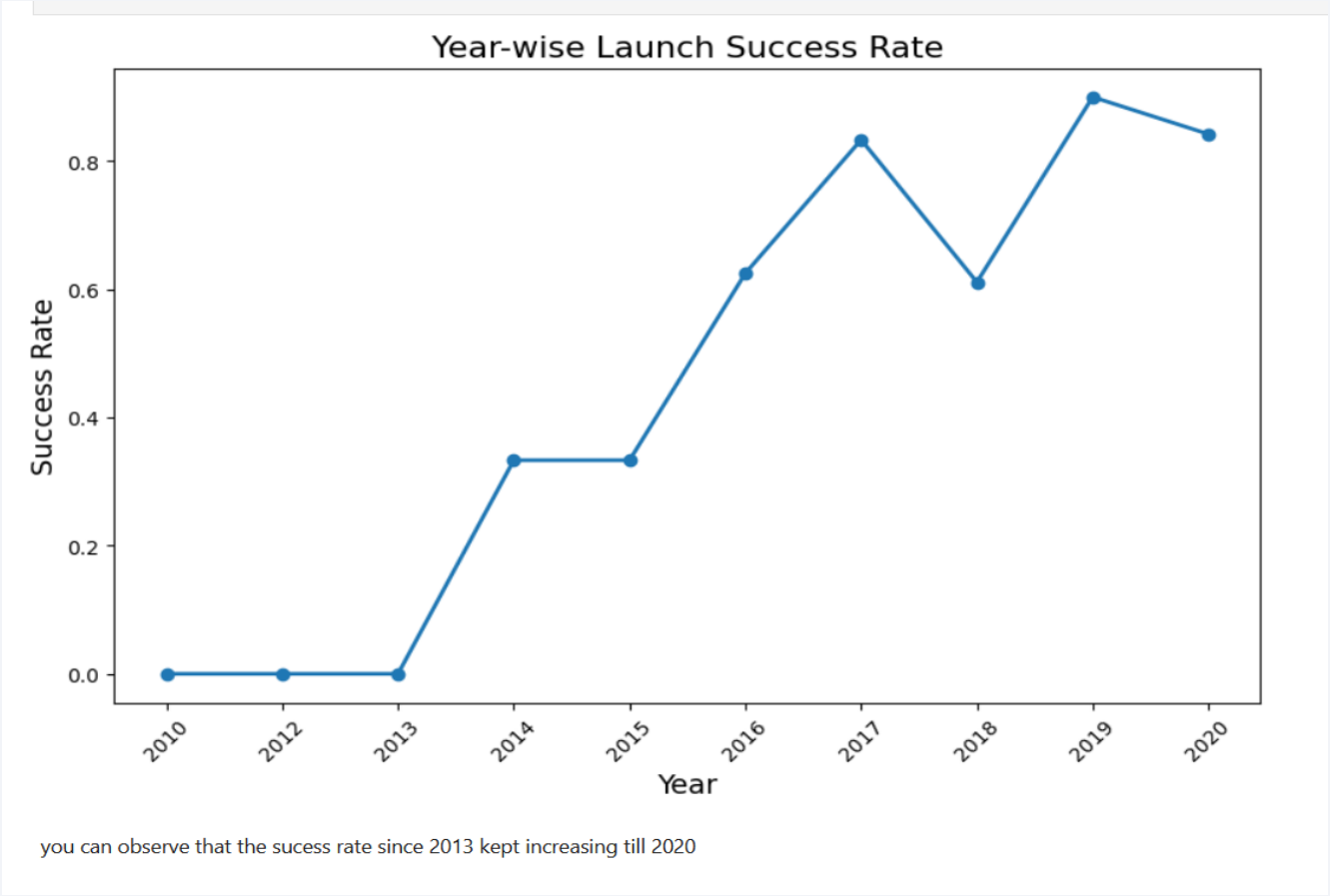


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Success rate steadily increased from 2013–2020.
- Evidence of maturing reusable booster technology.
- Increased reliability in engine, avionics, and landing systems.
- More launches but fewer failures → strong operational efficiency.
- Growing confidence from clients (NASA, private companies).



All Launch Site Names

- Query: %sql select distinct Launch_Site from SPACEXTBL;
- Result:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC 39-A
4. CCAFS SLC-40

- **Explanation**: This query retrieves all unique launch site names from the dataset using the DISTINCT keyword, ensuring no duplicates appear in the result.

Launch Site Names Begin with 'CCA'

- Query: %sql select * from SPACEXTBL where Launch_Site like "CCA%" Limit 5;
- Result:

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_O
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No

- **Explanation:** This query retrieves first 5 launch site names where launch site start with “CCA” from the dataset using the Like keyword, ensuring no other launch site name appear in the result.

Total Payload Mass

- Query: %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like "%NASA (CRS)%";
- Result:

  SUM(PAYLOAD_MASS__KG_)

  48213

- **Explanation:** This query calculates the total payload mass carried by boosters for missions commissioned by NASA (CRS). The SUM() function aggregates all payload masses for records where the customer name contains “NASA (CRS)”.

  The total payload transported is **48,213 kg.**

Average Payload Mass by F9 v1.1

- Query: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = "F9 v1.1";
- Result:

  avg(PAYLOAD_MASS__KG_)

  2928.4

- **Explanation**: This query calculates the average payload mass carried by boosters version “F9 v1.1” . The AVG() function aggregates all payload masses for records where the booster version is “F9 v1.1”.

The average payload mass is **2928.4**

First Successful Ground Landing Date

- Query: %sql select Date from SPACEXTBL where Mission_Outcome = "Success" and Landing_Outcome = "Success (ground pad)";
- Result: Date

  2015-12-22

  2016-07-18

  2017-02-19

  2017-05-01

  2017-06-03

  2017-08-14

  2017-09-07

  2017-12-15

- **Explanation:** This query identifies the breakthrough dates when SpaceX successfully combined mission success with ground pad landings. The results highlight key milestones in reusable rocket technology, starting with [“2015-12-22”] as the pioneering achievement."

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: %sql select Booster_Version from SPACEXTBL where Landing_Outcome like "%drone%" AND PAYLOAD_MASS__KG_ between 4000 and 6000;
- Result: Booster_Version

  F9 FT B1020

  F9 FT B1022

  F9 FT B1026

  F9 FT B1021.2

  F9 FT B1031.2

- **Explanation:** This query identifies the reliable booster versions used for medium-payload missions requiring drone ship landings. Results show that F9 FT B1020 and few other have successfully handled payloads between 4,000-6,000 kg while achieving precise ocean landings, making them ideal for cost-effective satellite deployment missions.

Total Number of Successful and Failure Mission Outcomes

- Query: %sql select Mission_Outcome,Count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome;
- Result:

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

- **Explanation**: This query provides the fundamental success metrics for SpaceX's launch history. The results show that out of 101 missions, 99 were successful while 1 failed, giving SpaceX an overall success rate of approximately 98%, demonstrating their remarkable reliability in space launch operations

Boosters Carried Maximum Payload

- Query: %sql select Booster_Version ,PAYLOAD_MASS__KG_ from SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(select Max(PAYLOAD_MASS__KG_) from SPACEXTBL);
- Result:

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

Explanation: This query identifies the champion booster version responsible for SpaceX's record-breaking heavy-lift missions. The results show that the F9 B5 B1048.4 carried the maximum 15600 kg, demonstrating the upper limit of SpaceX's current lift capability and highlighting their most powerful rocket configuration for satellite deployment."

2015 Launch Records

- Query: %sql select substr(Date, 6,2)as Month, Booster_Version from SPACEXTBL where Landing_Outcome = "Failure (drone ship)" and substr(Date, 0, 5) ="2015";
- Result:

Month	Booster_Version
01	F9 v1.1 B1012
04	F9 v1.1 B1015

- **Explanation**: This query pinpoints the challenging periods in 2015 when SpaceX experienced drone ship landing failures. The results show that first month F9 v1.1 B1012, fourth month F9 v1.1 B1015 had difficulties with ocean-based recoveries, highlighting the technical hurdles SpaceX faced during this crucial development phase of their reusable rocket program. These failures provided valuable data that eventually led to the high success rates seen in later years."

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

- Query: %sql SELECT landing_outcome,COUNT(*) AS outcome_count FROM SPACEXTBL WHERE date >= '2010-06-04' AND date <= '2017-03-20' GROUP BY landing_outcome ORDER BY outcome_count DESC;

- Result:

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

Explanation: This query reveals the landing outcome patterns during SpaceX's crucial development years from 2010-2017. The results show that top 5 outcome from ‘Success (drone ship)’ was the most frequent outcome, followed by Success(ground pad).


Section 3
Launch Sites
Proximities Analysis

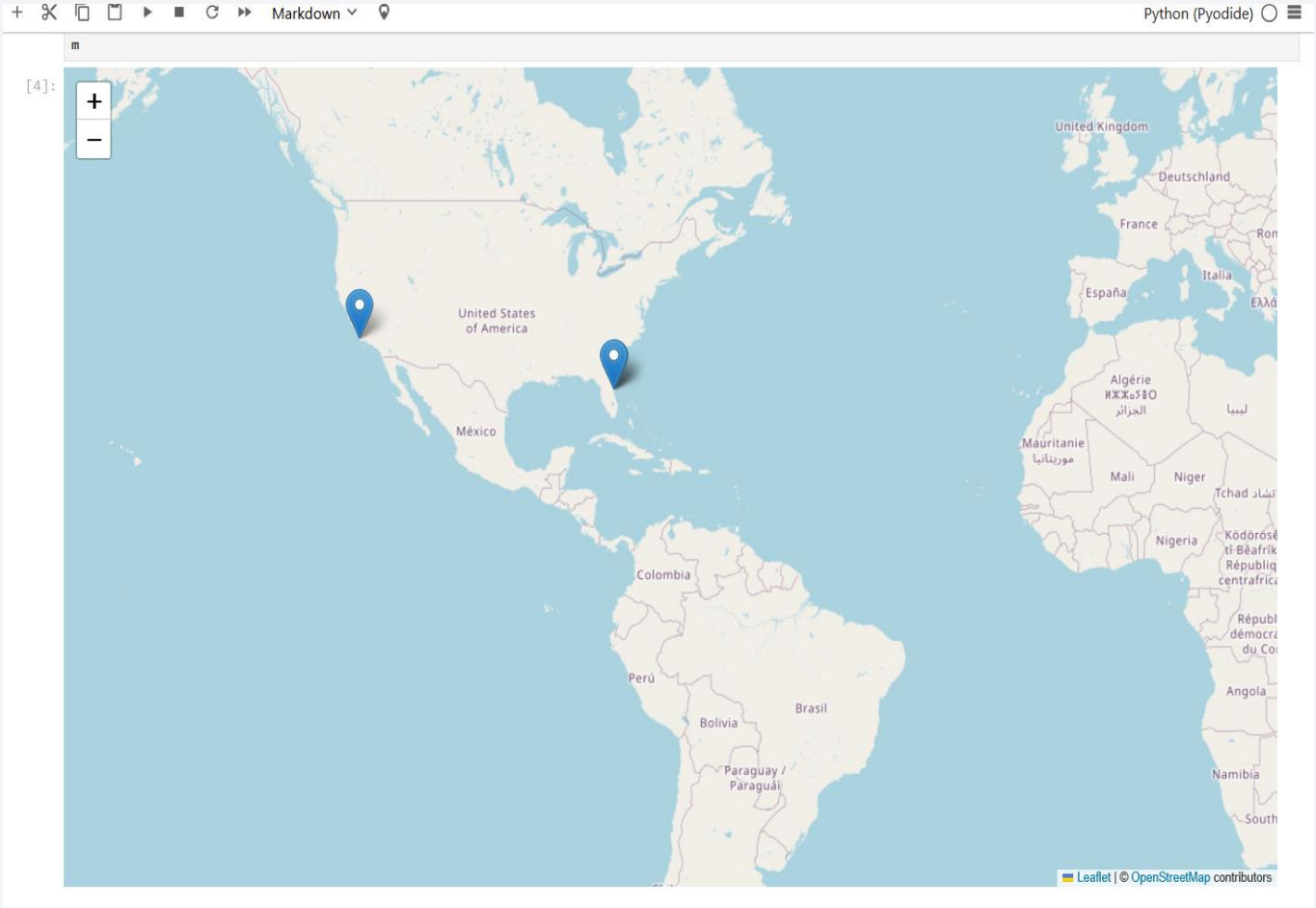
Map of launch site

First launch site: CCA FS Located directly on the east Florida coastline.

Rockets launch eastward over the Atlantic Ocean. Minimizes risk to populated land areas.

Ideal for missions requiring orbital trajectories.

Second launch site: On the *west coast.* Used for polar orbits. Also close to coastline to launch over the Pacific Ocean.



Map of success/failed launches for each site

This Folium map visualization displays SpaceX launch sites with color-coded markers indicating mission outcomes
Green marker = Successful launches
Red marker = Failed launches.
This visualization helps identify which launch sites have historically been most successful, supporting future mission planning and site selection decisions.

Distance of launch site to its proximities

[15]:

	Launch Site	Distance to City (km)	Distance to Highway (km)	Distance to Coastline (km)	Distance to Railway (km)
0	CCAFS LC-40	19.084658	17.244010	17.011743	18.190018
1	CCAFS SLC-40	19.190875	17.349198	17.116774	18.295982
2	KSC LC-39A	20.429538	18.881545	18.690939	19.634662
3	VAFB SLC-4E	3836.909629	3836.997235	3837.010688	3836.851835

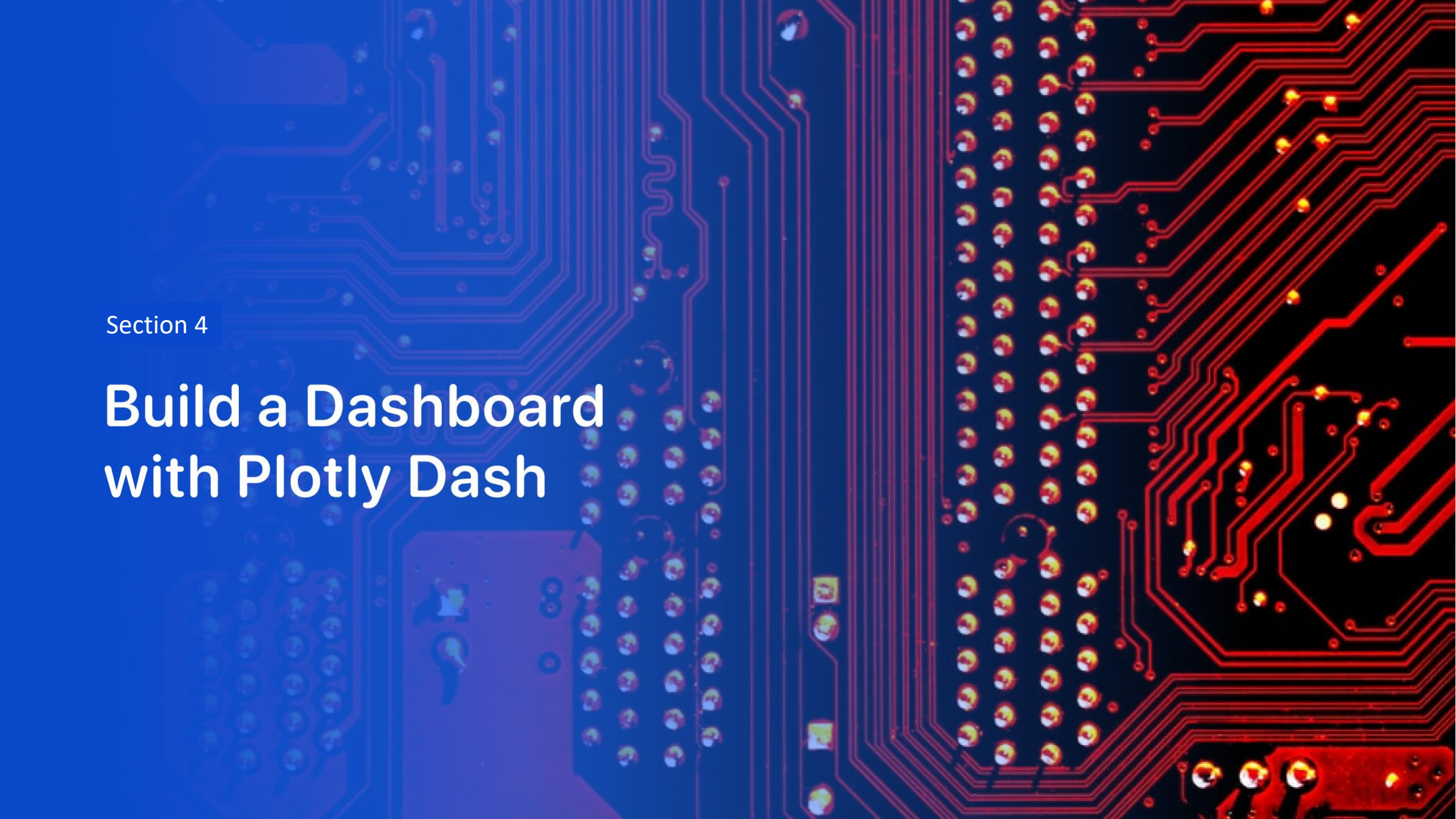
Key findings:

- **Florida Launch Sites (CCAFS & KSC):** Close proximity to all infrastructure (17-21 km.

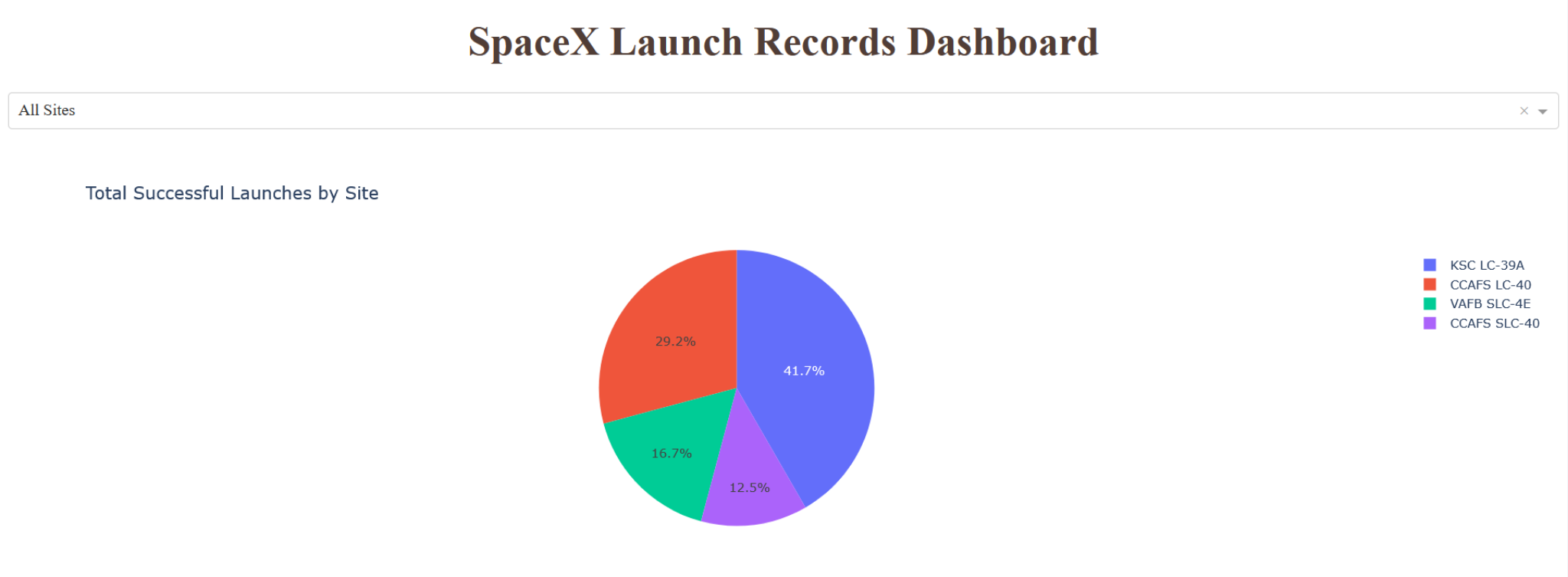
CCAFS LC-40 & SLC-40: Approximately 19 km from city, 17 km from highway/coastline, 18 km from railway.

KSC LC-39A: Slightly more distant (~20 km from city, 19 km from railway).

- **California Launch Site (VAFB SLC-4E):** Extremely remote location**.** Significant geographical isolation compared to Florida sites


Section 4
Build a Dashboard with Plotly Dash

Success count of all site



This dashboard visualization provides a clear breakdown of launch success rates across SpaceX's different launch facilities:

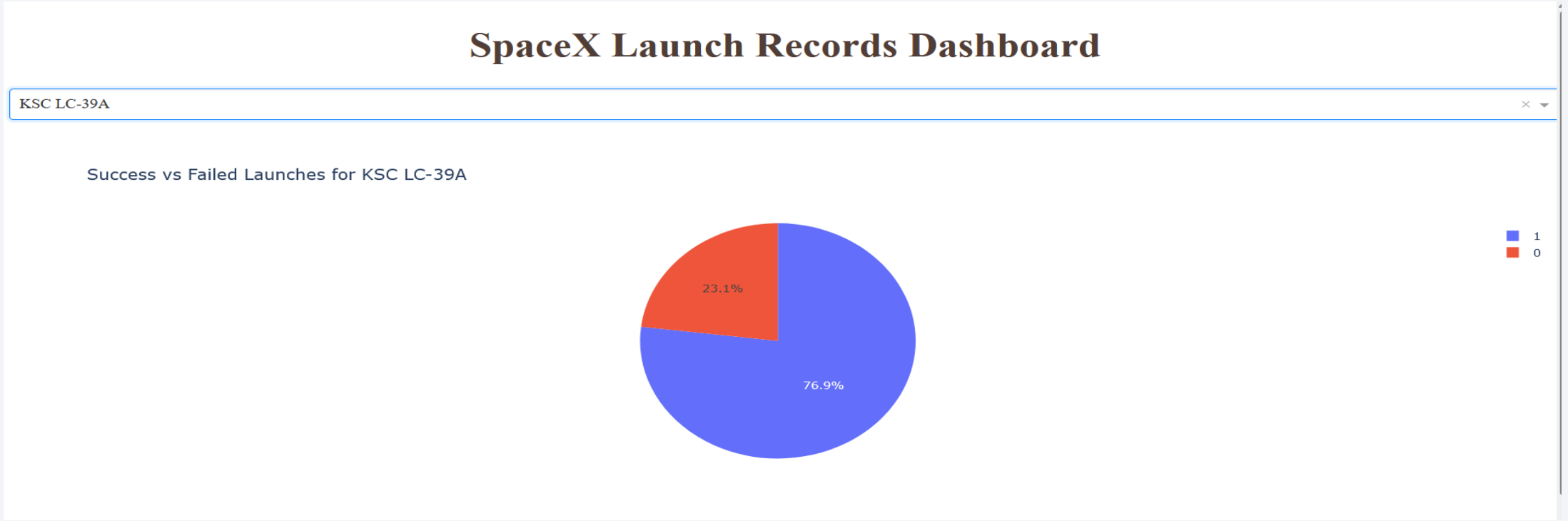
KSC LC-39A:41.7% (Highest rate)

CCAFS LC-40: 29.2%

VAFB SLC-4E:16.7%

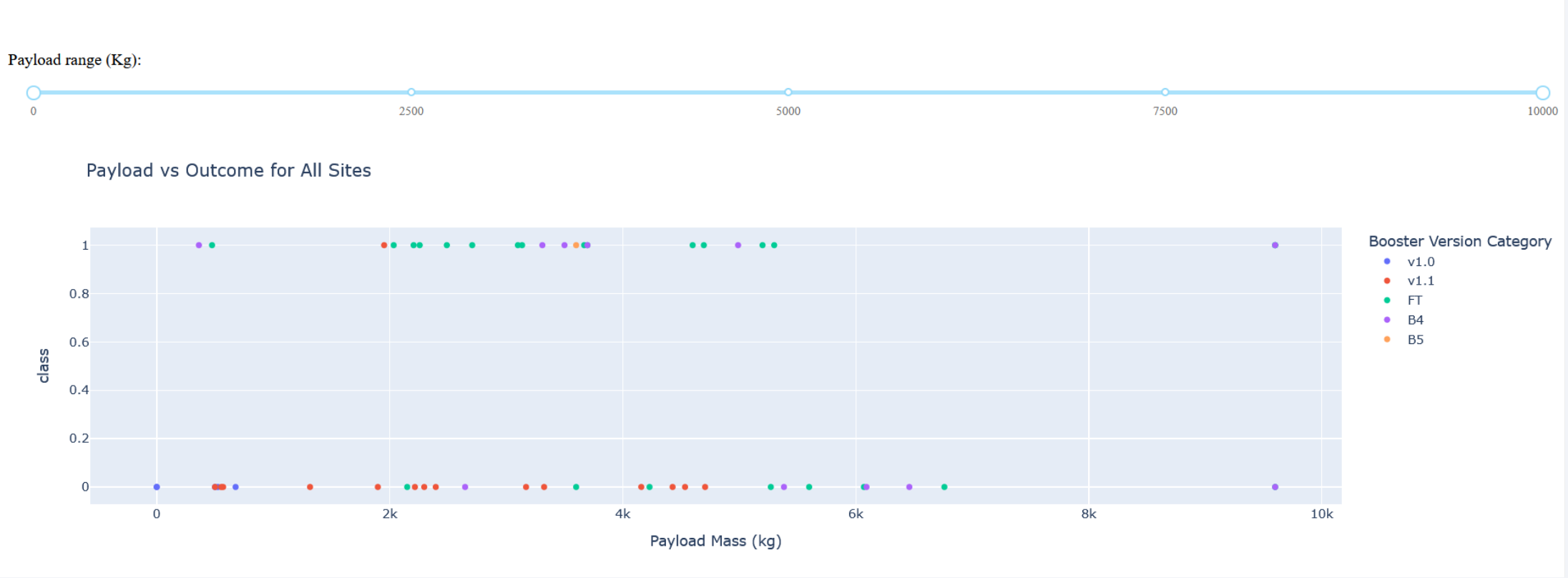
CCAFS SLC-40:12.5%(Lowest rate)

Highest success rate of KSC LC-39A



KSC LC-39A has a success rate 76.9% and failed rate 23.1% . This is highest rate among the all site.

Scatter plot of all site



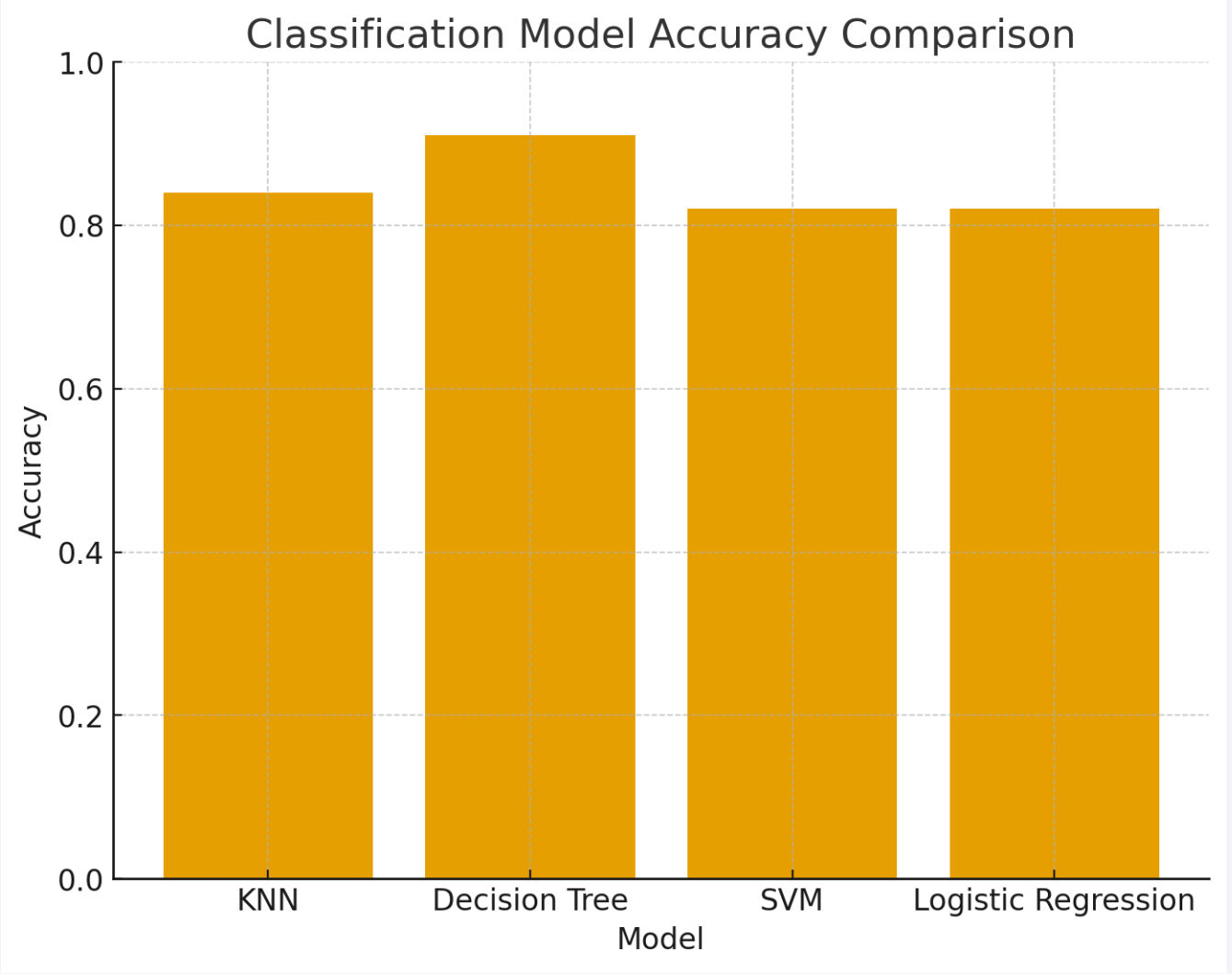
No strong correlation between payload and success.
Most high-payload launches are successful.
No clear payload threshold for failure.

Section 5

Predictive Analysis (Classification)

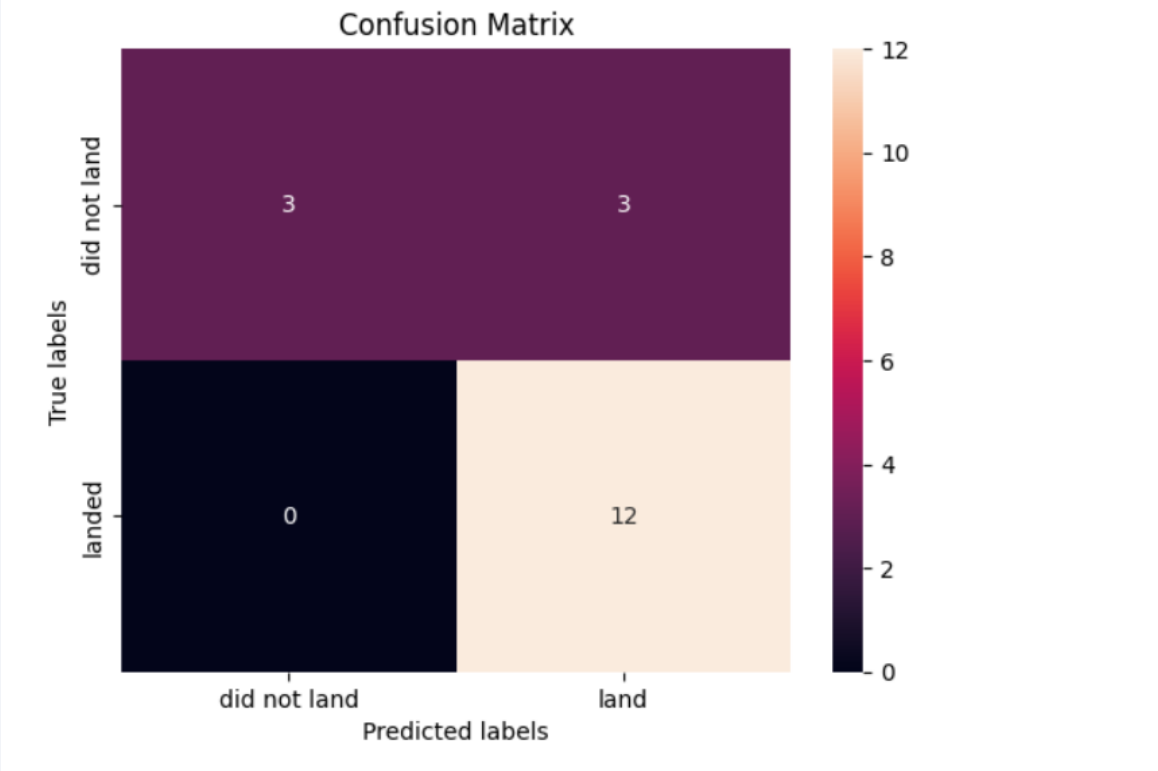
Classification Accuracy

All these model shows the how well each model performed.
Accuracy of Decision tree is 0.91 which is more than the rest of model. Although the rest of model has accuracy around 0.82-0.84



Confusion Matrix

- This is the best confusion matrix out of all confusion matrix that evaluates the performance of the classification model in predicting rocket landing outcomes.
- The Decision Tree performs very well, with the majority of predictions correct.
- The largest error is False Positives (predicting landing when it didn't land).
- Since FN = 0, the model is very good at recognizing landings—important if predicting mission success



Conclusions

- ✓ The Exploratory Data Analysis (EDA) revealed **steady improvement in SpaceX launch success rates**, especially after 2015.
- ✓ **Booster version FT** consistently showed the **highest landing success**, while older versions had more failures.
- ✓ Payload analysis showed that **heavy payloads slightly reduce landing success**, but overall impact is moderate.
- ✓ Geographic analysis confirmed that **most launch sites are near coastlines**, optimizing safety and recovery.
- ✓ The interactive dashboard enabled **real-time exploration** of payload, site performance, and booster versions, improving interpretability.
- ✓ Multiple machine learning models were built to **predict landing success** — Decision Tree performed best with **91% accuracy**.
- ✓ Overall, the combined EDA, interactive analytics, and predictive modeling provide a **holistic understanding** of SpaceX launch performance and future mission success probabilities.

Thank you!