

Winning Space Race with Data Science

Santiago Cortes
18/11/2025



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Strategic Conclusions
- Appendix

Executive Summary

This project applies data science to analyze and predict rocket launch results, leveraging both public data sources and machine learning techniques. Key steps included data collection, exploratory data analysis, interactive visualization, and predictive modeling, all aimed at supporting decision-making for space launch enterprises

Introduction

The goal is to assess whether a new enterprise, Space Y, can compete effectively with SpaceX. Two main business questions guided this study:

- How to estimate launch costs by accurately predicting whether rockets' first stages will land successfully.
- What are the optimal launch site locations?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection: SpaceX's public API and Wikipedia's launch records via web scraping.
- Data Wrangling: Cleaning procedures included managing missing values, creating outcome labels, and summarizing feature statistics.
- Exploratory Data Analysis (EDA): Performed with SQL queries, bar plots, scatterplots, and other visualizations.
- Interactive Analytics: Used Folium maps for site analysis, and Dash dashboards for launch data.
- Predictive Modeling: Prepared and split data into training/test sets, evaluated four classifiers: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors.

Data Collection

- SpaceX API: Fetched structured launch records filtered to Falcon 9 missions.
- Wikipedia Scraping: Extracted tables from the Falcon 9/Heavy launch history for richer feature description.

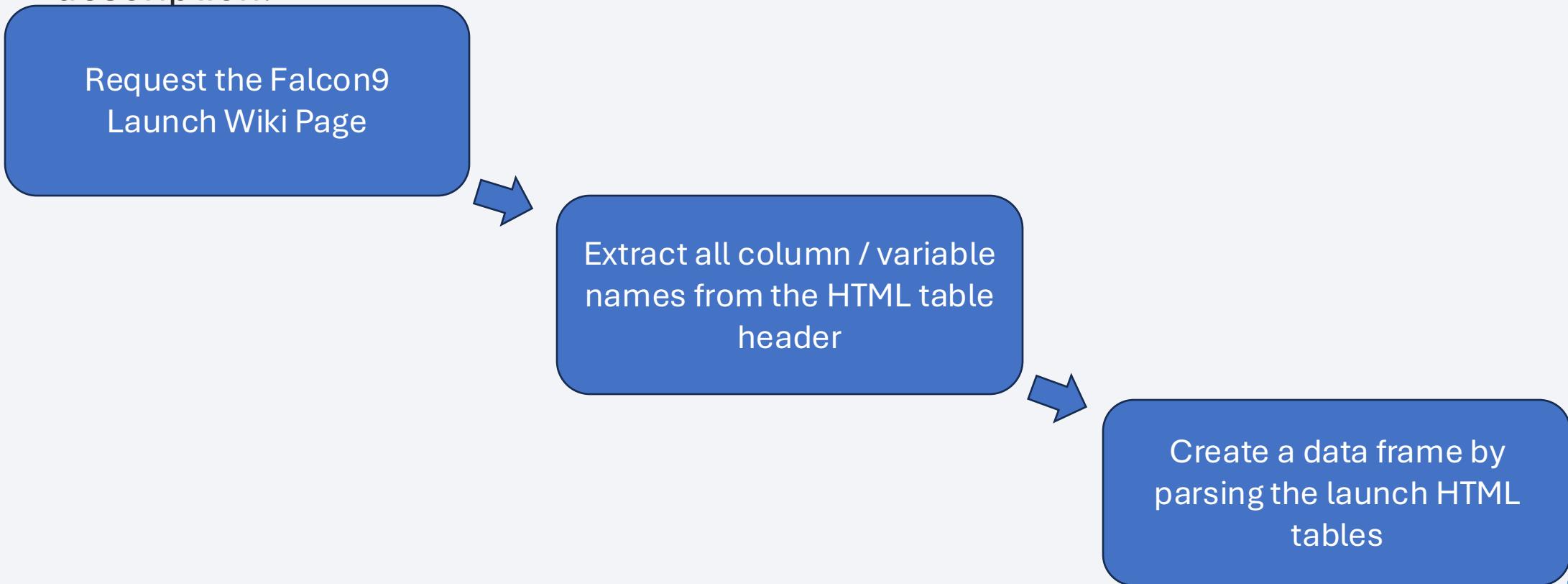
Data Collection – SpaceX API

- Fetched structured launch records filtered to Falcon 9 missions.



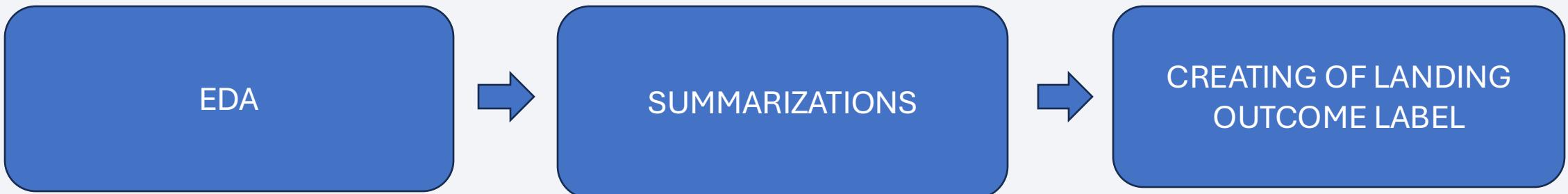
Data Collection - Scraping

- Extracted tables from the Falcon 9/Heavy launch history for richer feature description.



Data Wrangling

- Initial EDA revealed site, orbit, payload, and outcome statistics.
- Created a binary landing outcome label to enable prediction (success/failure).
- SQL queries were used for custom aggregation, e.g. site trends, payload analysis.



<https://github.com/santiago-cortes14/Capstone-Data-Science-Coursera-Santiago-Cortes/blob/main/Data%20Wrangling.ipynb>

EDA with Data Visualization

- Several charts were created to better understand launch data and highlight key trends:
 - Scatter plots illustrated the relationship between flight number, payload mass, and launch success across different launch sites. These visualizations helped us observe which sites were most frequently used and where the highest payloads were managed.
 - Bar charts were used to compare the success rate for different orbit types, quickly revealing which orbits had consistently higher probabilities of successful missions.
 - Line plots tracked launch success rates year over year, providing a clear picture of improvement over time and operational consistency.
- Additional scatter plots explored the connection between flight number and orbit, as well as between payload mass and orbit type, to discover site-specific and mission-specific patterns in outcomes.
- These visualizations were essential for spotting outliers, confirming expected patterns (such as higher payloads concentrated at specific launch pads), and directing further modeling efforts.

EDA with SQL

- A series of SQL queries were designed and executed to extract essential statistics and insights from the SpaceX dataset:
 - Identified unique launch sites and summarized their respective mission counts.
 - Filtered records to examine trends specifically for launches taking place at Cape Canaveral (CCA).
 - Calculated total and average payload mass for specific missions, such as those carried out by NASA CRS and certain Falcon 9 booster versions.
 - Determined the date of the first successful ground-landing, highlighting key milestones in SpaceX's operational history.
 - Listed boosters that succeeded with drone ship landings and handled heavy payloads.
 - Generated overall counts of successful and failed missions, revealing reliability trends.
 - Used subqueries to find which booster types managed the heaviest payloads.
 - Isolated failed drone ship landings in 2015, including booster version and launch site, for incident analysis.
 - Ranked landing outcome counts between two dates to study historical performance patterns.
- These queries helped establish a quantitative foundation for both site selection and model training

Build an Interactive Map with Folium

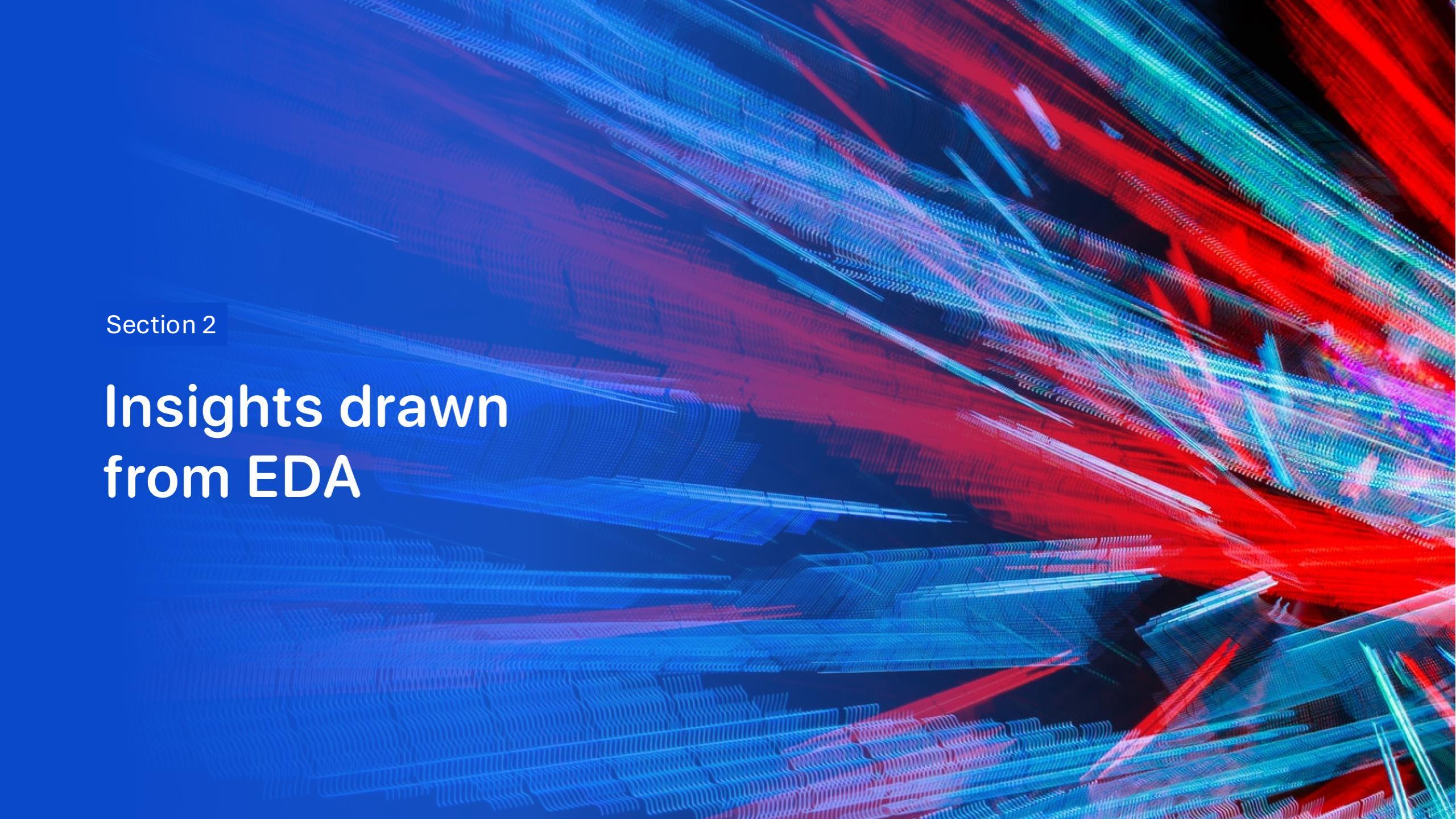
- Folium Map: Displayed launch locations, proximity to logistics (roads, railways), and outcomes.
- Dash Dashboard: Highlighted percentage of successful launches, payload-to-outcome relationships.

Build a Dashboard with Plotly Dash

- An interactive dashboard was built using Plotly Dash to allow for dynamic exploration of the SpaceX launch dataset.
- Features included:
 - A dropdown menu enabling users to select a specific launch site or view data for all sites.
 - A pie chart summarizing launch successes by site, dynamically updated based on the chosen site.
 - A slider controlling the selectable payload mass range, allowing for focused analysis of payload outcomes.
 - A scatter plot displaying the relationship between payload mass and launch success, with points colored by booster version and filtered by site and payload range.
- Custom callback functions ensured that all visualizations updated in real-time according to user inputs.
- The dashboard provided rapid insights into which sites and payload masses were most strongly associated with successful launches, supporting deeper operational understanding.

Predictive Modeling Results

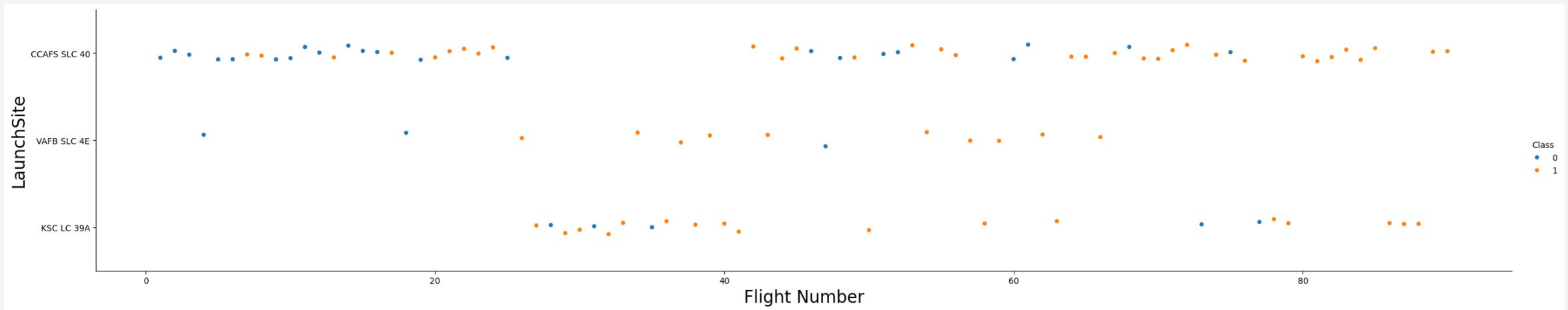
- Compared four classification models:
 - Logistic Regression
 - Support Vector Machine
 - Decision Tree (best performing)
 - K-Nearest Neighbors
- Key Result: Decision Tree Classifier achieved the best performance among all models tested:
 - Training Accuracy: 87%
 - Test Accuracy: 80%
 - Cross-validation Score: Consistent across folds This demonstrates strong generalization capability without overfitting.
- Confusion Matrix: Showed high rates of correct predictions for landing outcomes.

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D wireframe or a network of data points. The overall effect is futuristic and dynamic, suggesting concepts like data flow, digital communication, or complex systems.

Section 2

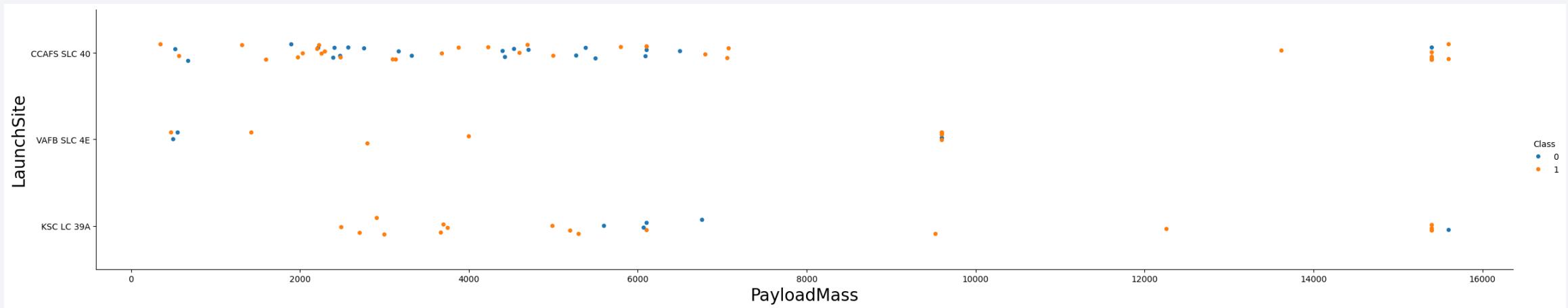
Insights drawn from EDA

Flight Number vs. Launch Site



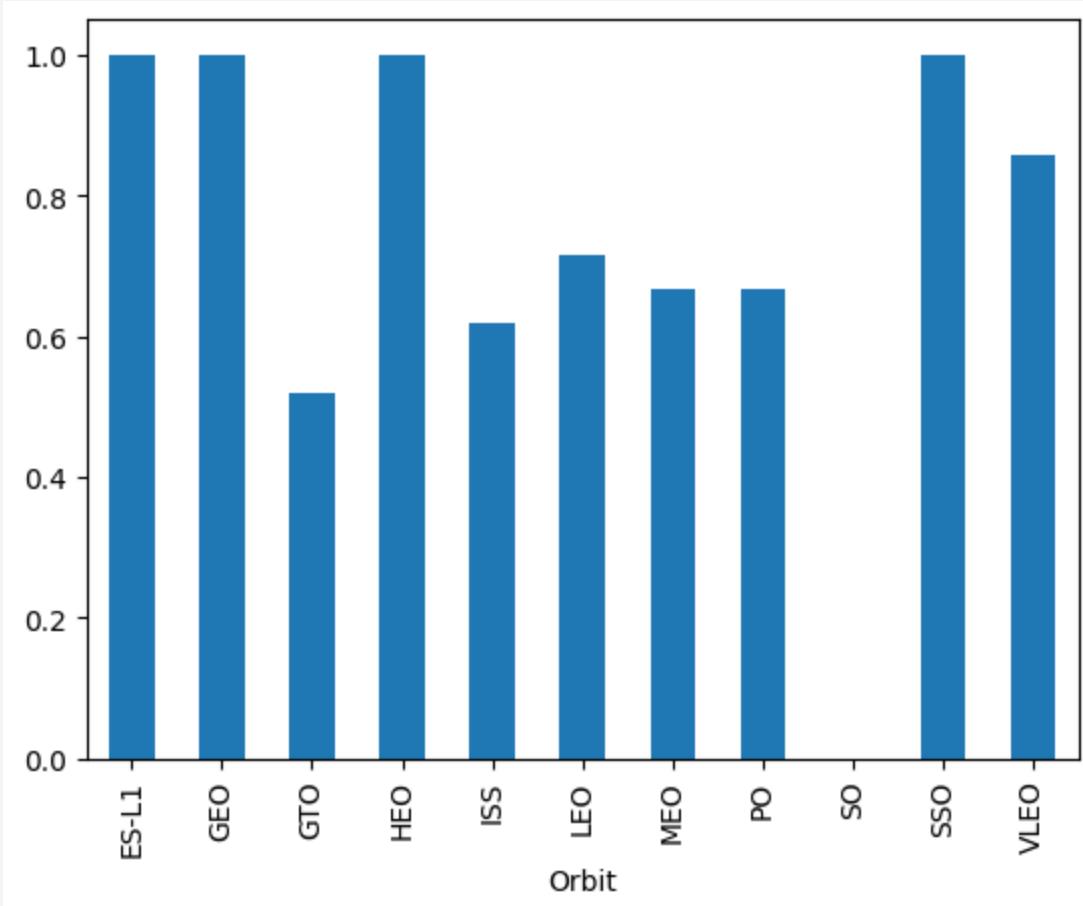
- This scatter plot demonstrates the relationship between flight number and launch site, with mission outcomes highlighted.
- Insight: As the number of launches increases at each site, the probability of a successful landing tends to improve, showcasing the role of iterative experience and operational learning in performance gains.

Payload vs. Launch Site



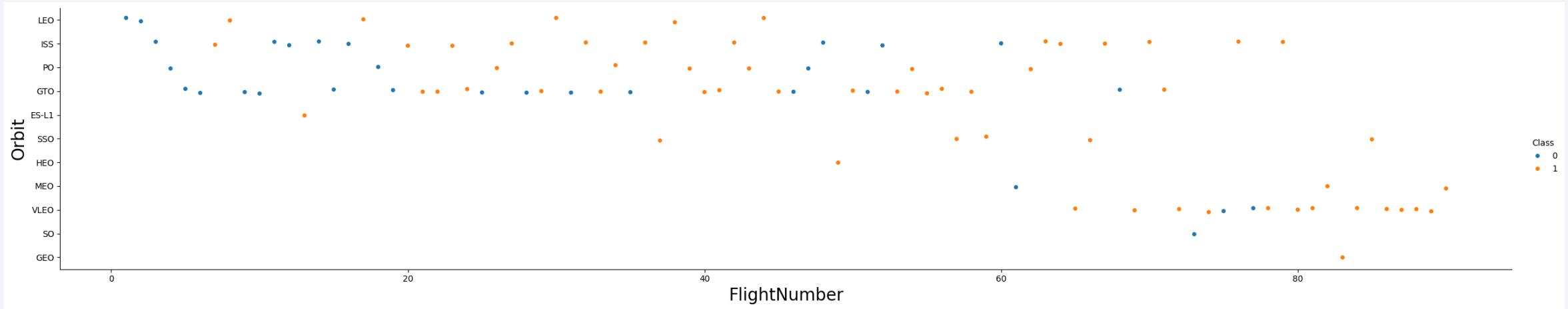
- In this visualization, payload mass is plotted against the launch site, colored by success and failure.
- Observation: Heavier payloads are typically handled at select locations like CCAFS SLC-40 and KSC LC-39A, confirming these sites' suitability for high-capacity missions.

Success Rate vs. Orbit Type



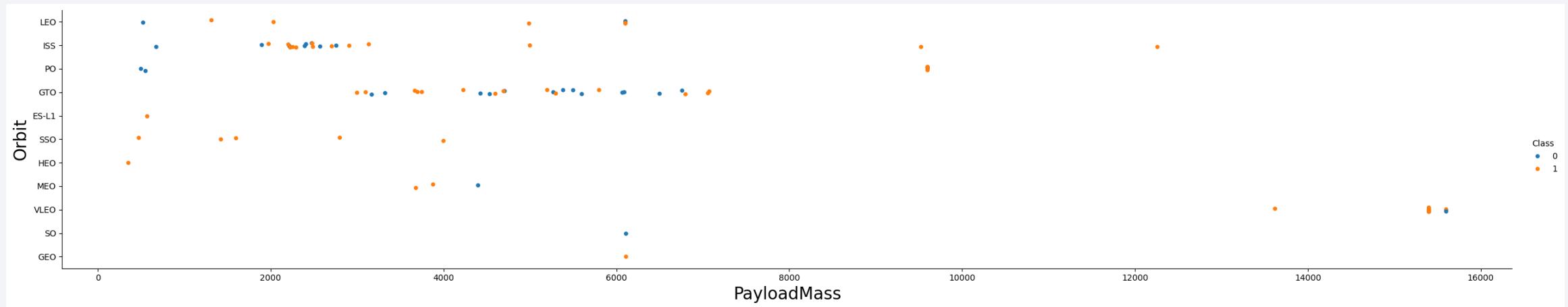
- This bar chart compares success rates across various orbital destinations.
- Key finding: Missions targeting orbits such as ES-L1, GEO, SSO, and HEO show nearly perfect success rates, while GTO missions have noticeably lower reliability.
- Understanding which orbits are more challenging informs planning and risk assessment.

Flight Number vs. Orbit Type



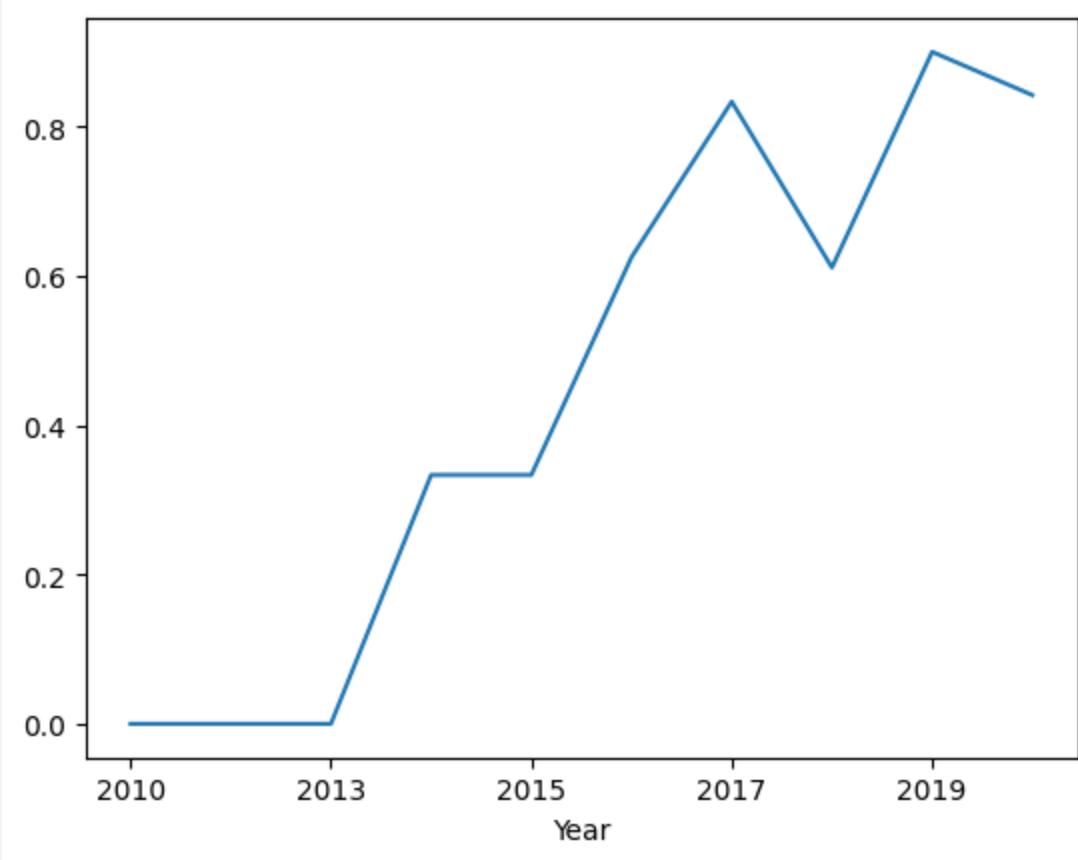
- This scatter plot illustrates flight number against orbit type, colored by class.
- Conclusion: Later flights—representing greater operational experience—consistently target a broader range of orbits and show higher success rates, especially for typically more risky orbits.

Payload vs. Orbit Type



- Here we see a scatter plot of payload mass versus orbit type, with outcome status indicated.
- Finding: High-payload launches succeed most frequently to orbits like SSO and GEO, while GTO missions involving large payloads experience more failures. This highlights orbit-specific risk factors related to payload.

Launch Success Yearly Trend



- A line plot charts the launch success rate over time.
- Interpretation: Steady improvement year after year reflects technological advances and process optimization, underscoring SpaceX's growing reliability and experience.

All Launch Site Names

Regarding to the data, there are 4 launch sites.

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

They are obtained by selecting unique occurrences of "launch_site" values from the dataset.

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with CCA

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

Total Payload Mass

The total payload carried by boosters from NASA

Total Payload (kg)
111.268

Calculated by summing all payloads whose codes contain 'CRS', which corresponds to NASA

Average Payload Mass by F9 v1.1

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

Avg Payload (kg)
2.928

Filtering data by the booster version above and calculating the average payload mass

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad

Min Date
2015-12-22

By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Selecting distinct booster versions according to the filters above.

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Group mission outcomes and counting records for each group

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

Booster Version (...)	Booster Version
F9 B5 B1048.4	F9 B5 B1051.4
F9 B5 B1048.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1056.4
F9 B5 B1049.5	F9 B5 B1058.3
F9 B5 B1049.7	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1060.3

These are the boosters which have carried the maximum payload mass registered

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

This list has only two occurrences

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

Ranking of 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

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

This view of data alert us that "No attempts" must be taken in account

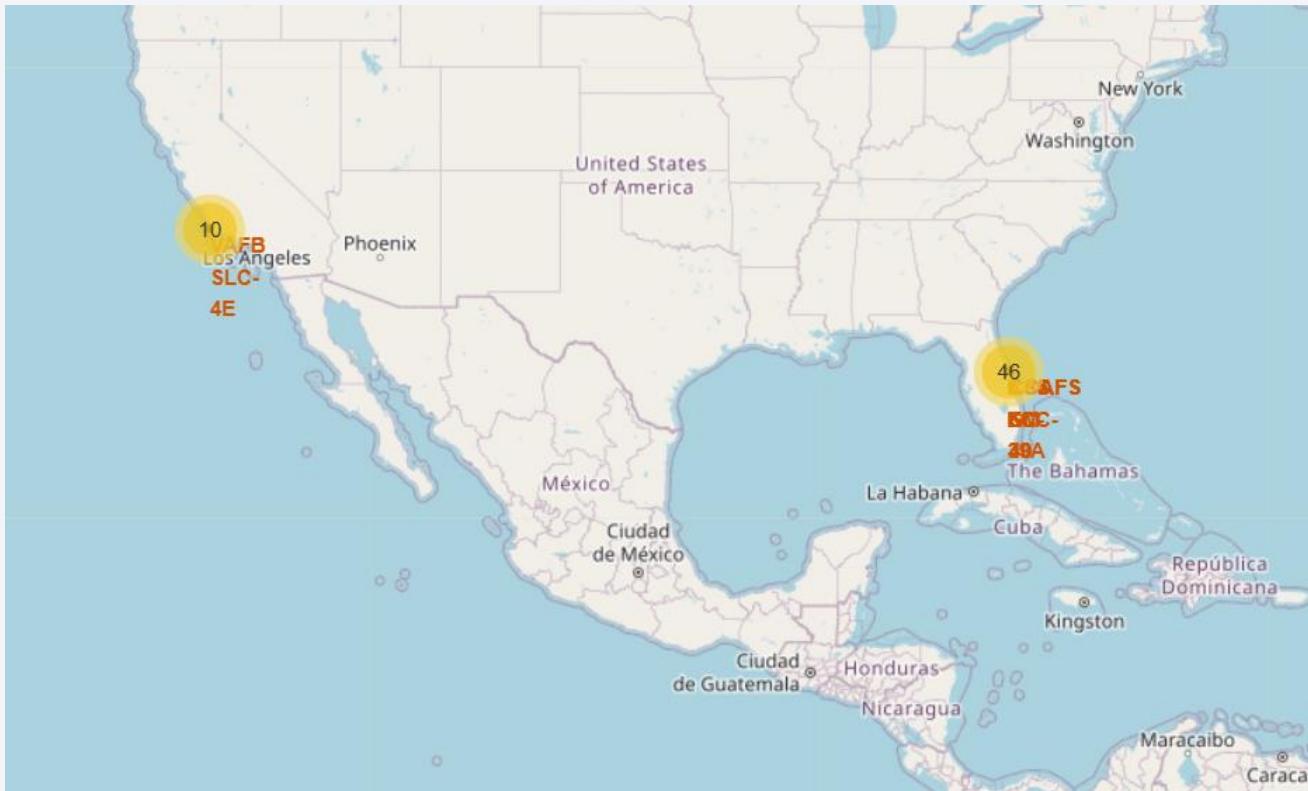
The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth's horizon against a dark blue sky. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper right, there are bright green and yellow bands of light, likely the Aurora Borealis or Australis. The overall atmosphere is dark and mysterious.

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 KSCLC-39A launch site launch outcomes

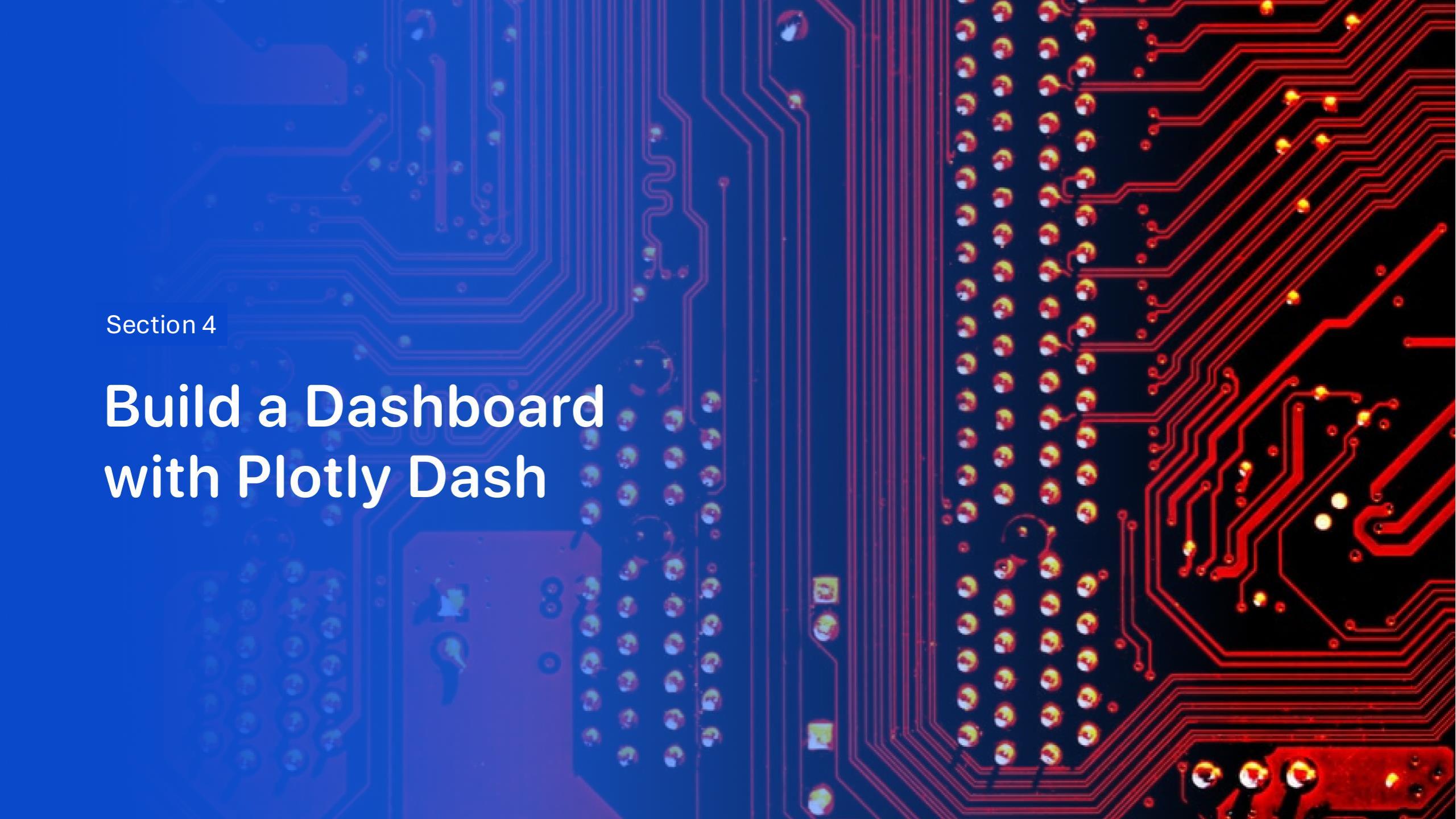


- Green markers indicates successful
- Red ones indicate failure

Logistics and safety

Launch site KSCLC-39A has good logistics, being near railroad and road. Relatively far from inhabited areas.

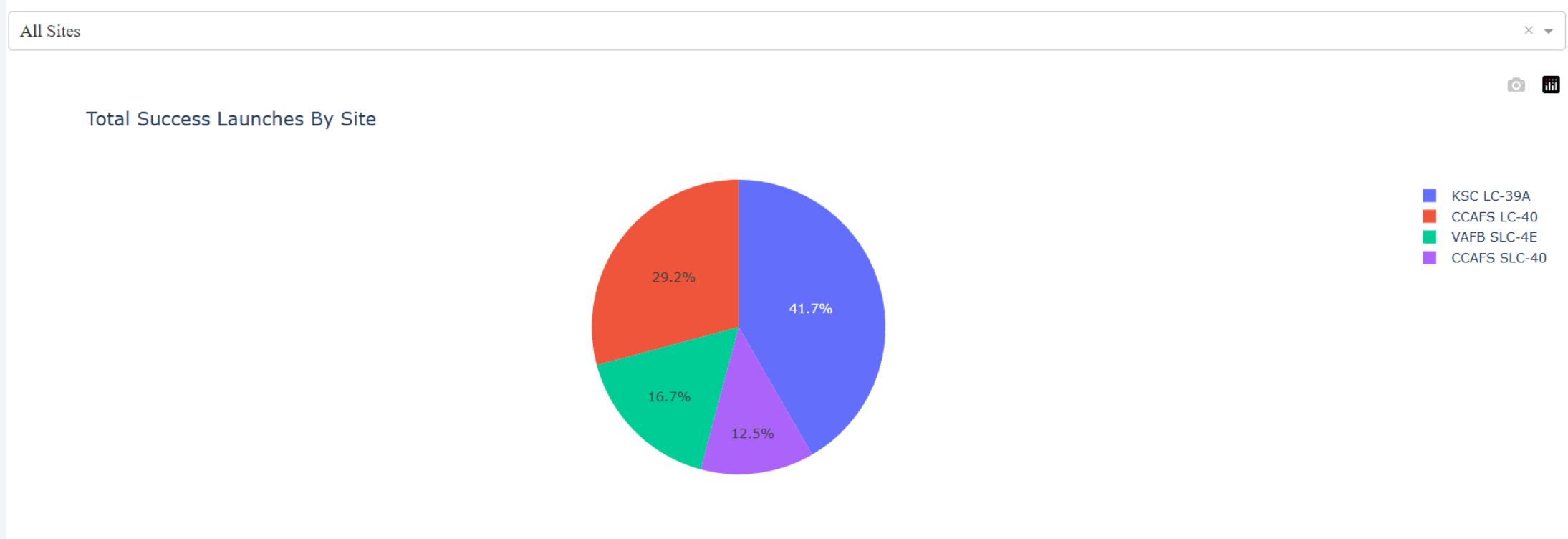


The background of the slide features a close-up photograph of a printed circuit board (PCB). The left side of the image has a blue color overlay, while the right side has a red color overlay. The PCB itself is dark blue/black with numerous red and blue printed circuit lines. Numerous small, circular gold-colored components, likely surface-mount resistors or capacitors, are visible. A few larger blue and red components are also present.

Section 4

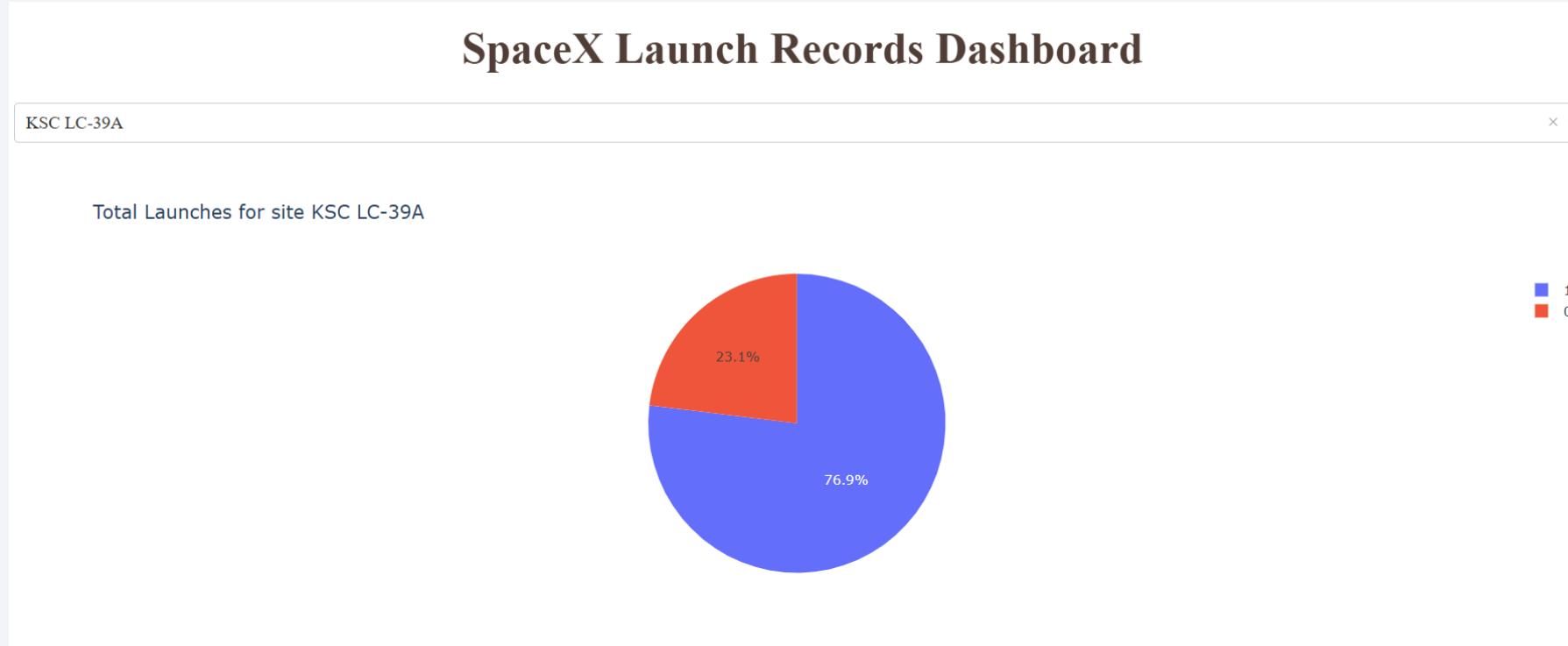
Build a Dashboard with Plotly Dash

Launch Success Distribution Across All Sites



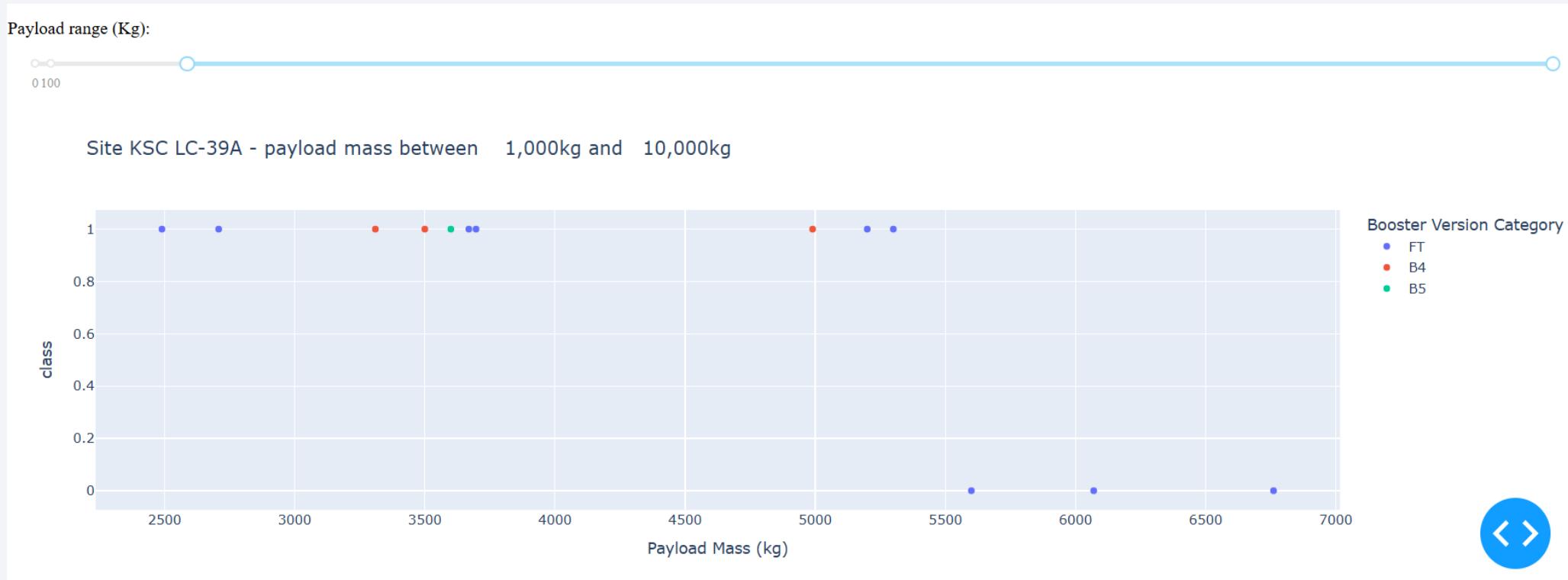
The pie chart reveals the distribution of successful launches across our operational sites

Success Rate Analysis for Top Performing Site



The metrics demonstrate 76.9% successful launches at this facility

Payload Mass Impact on Launch Success



Payloads between 0 kg and 6000 kg and FT Boosters show the highest success rates

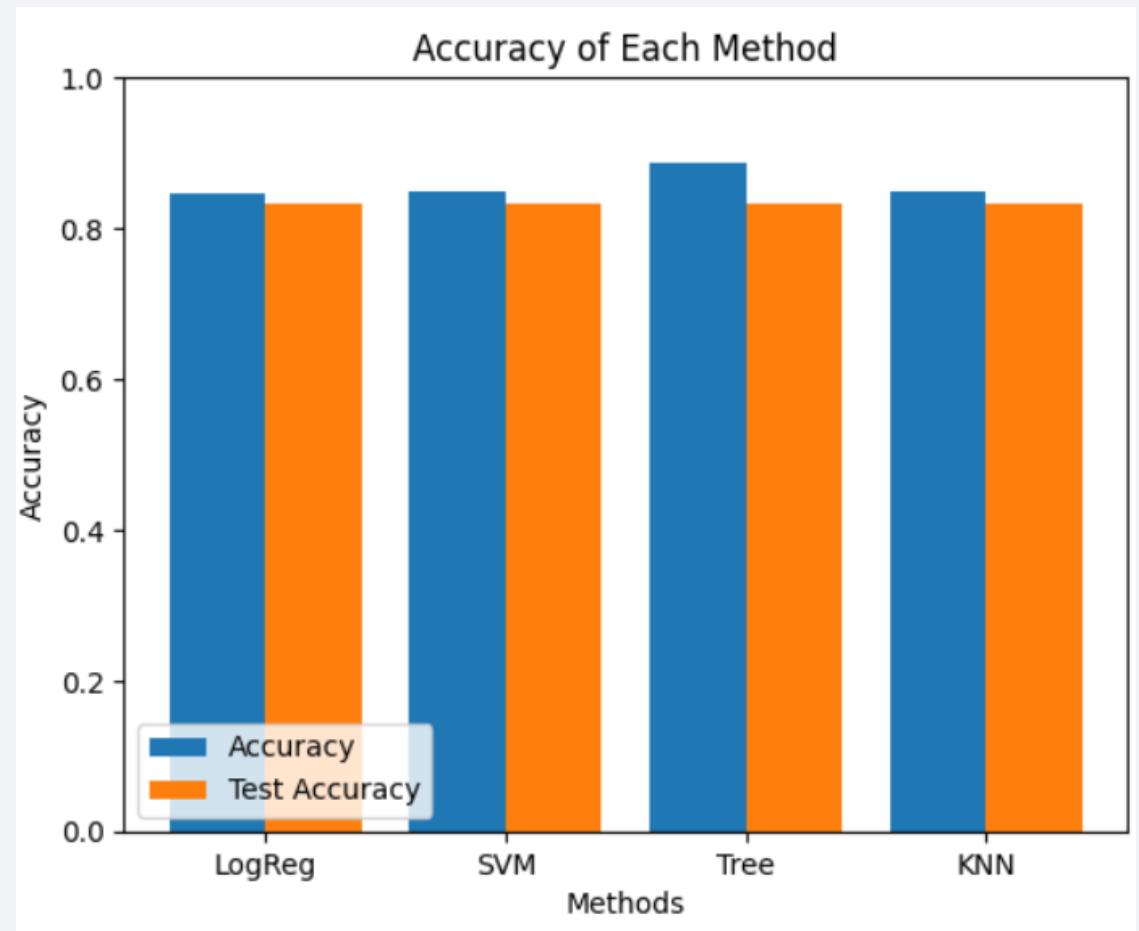
The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines in shades of blue and yellow, creating a sense of motion and depth. The lines curve from the top left towards the bottom right, with some lines being more prominent than others. The overall effect is reminiscent of a tunnel or a high-speed journey through a digital space.

Section 5

Predictive Analysis (Classification)

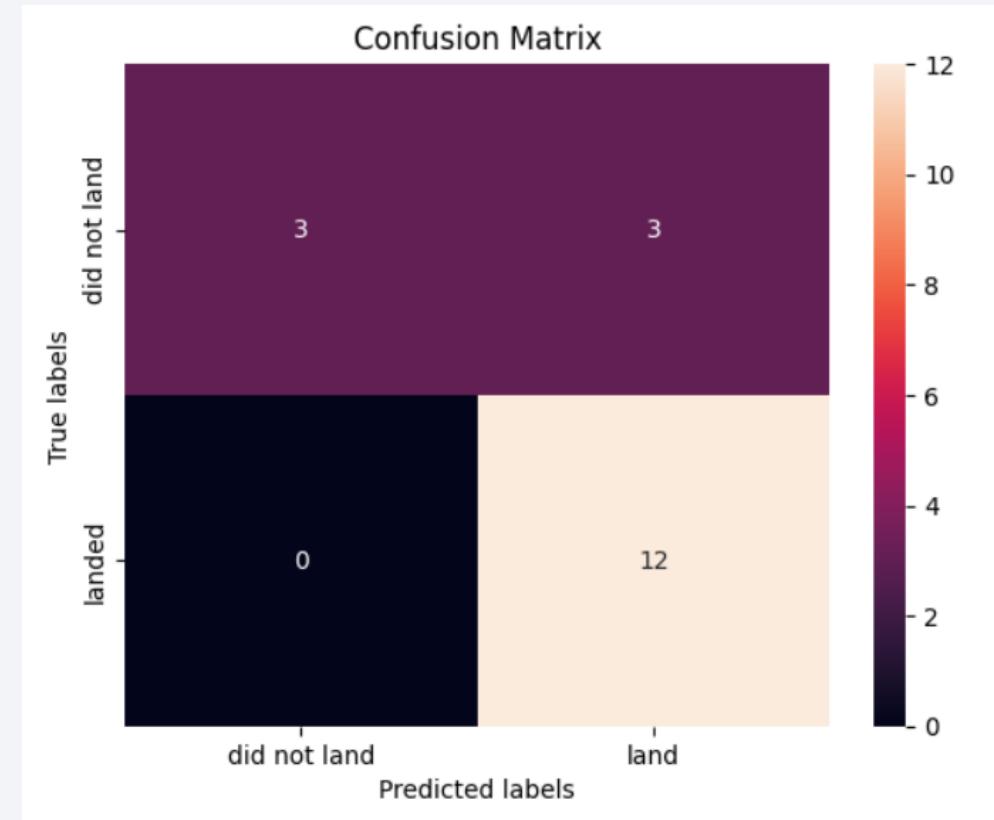
Classification Accuracy

- Four classification models were developed and evaluated using 10-fold cross-validation
- Decision Tree Classifier achieved the highest accuracy at 87% for training data



Confusion Matrix

- The confusion matrix validates the Decision Tree's classification performance
- True Positives: 12 - correctly predicted successful landings
- True Negatives: 3 - correctly predicted failed landings
- False Positives: 3 - cases incorrectly predicted as successful
- False Negatives: [0] - cases incorrectly predicted as failures



Strategic Conclusions

- Site Selection Impact: CCAFS SLC-40 and KSC LC-39A demonstrate superior performance metrics, making them preferred facilities for mission-critical launches
- Payload Optimization: Payloads between 0 kg and 6000 kg achieve the highest success rates (as shown in the dashboard analysis), particularly with FT Boosters and B5 variants.
- Technology Evolution: Success rates correlate strongly with booster version advancement; newer B5 variants significantly outperform legacy F9 v1.0-v1.1 versions.
- Predictive Value: The Decision Tree classification model provides actionable insights for real-time mission risk assessment, enabling data-driven decision-making in launch operations
- Continuous Improvement: The upward trend in success rates demonstrates organizational learning; predictive models should be retrained quarterly with new mission data

Appendix

- GitHub Repositories: [Include links to your completed notebooks]
 - Data Collection (API & Web Scraping)
 - Data Wrangling & Preprocessing
 - EDA with SQL & Visualization
 - Folium Interactive Maps
 - Plotly Dash Dashboard
 - Predictive Analysis & Model Training
- Key Artifacts:
 - Python notebooks with complete code cells and outputs
 - SQL query scripts for exploratory analysis
 - Dashboard application source code
 - Model training and evaluation scripts
- Tools & Technologies:
 - Programming Language: Python 3.x
 - Data Analysis: Pandas, NumPy
 - Visualization: Matplotlib, Seaborn, Plotly, Folium
 - Machine Learning: Scikit-learn
 - Dashboard Framework: Plotly Dash
 - Database: SQL

Thank you!

