IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sarannya Thelapurath

24/07/2025

Outline

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

Executive Summary

In this capstone project, we aim to predict the successful landing of the Falcon 9 first stage. SpaceX’s cost-effective strategy is largely due to its reusable first-stage boosters. Accurately predicting landing success can significantly impact cost estimations and provide leverage for potential competitors bidding on launch contracts. This project involves collecting launch data from the SpaceX API, performing data wrangling, conducting exploratory and interactive data analysis, and building classification models to predict outcomes.

Introduction

SpaceX has revolutionized space transportation by developing reusable rocket technology, especially with its Falcon 9 launch vehicle. The reusability of the rocket’s first stage is a key factor in lowering launch costs from the industry average of over $165 million to approximately $62 million per launch. Predicting the likelihood of a successful landing of the first stage booster is critical for cost forecasting, mission planning, and strategic decision-making.

- Can we accurately predict whether the Falcon 9 first stage will land successfully?
- What launch factors (payload mass, orbit type, launch site, etc.) influence landing success?
- Which machine learning models are most effective for this classification task?
- How can interactive visualizations help stakeholders explore launch outcomes and trends?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - SpaceX REST API Calls
    - Web Scraping
- Perform data wrangling
    - Data cleaning, Feature Engineering, Encoding and Normalization
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

- **SpaceX REST API Calls**
- . We used the SpaceX public API (https://api.spacexdata.com/v4/launches) to retrieve structured JSON data about Falcon 9 launches.
- The API provided details including launch dates, payload masses, booster versions, landing outcomes, orbit types, and launch sites.
- This automated method ensured up-to-date and consistent access to real-time data
- **Web Scraping**
- To supplement missing fields (e.g., launch pad details, landing pad types, distances to coastlines or roads), we scraped structured tables from reliable websites like Wikipedia.
- Python libraries such as BeautifulSoup and Requests were used to extract and parse HTML tables into pandas DataFrames.
- **Data Integration**
- The data collected from the API and web scraping were merged into a single unified dataset.
- Each record was carefully linked by launch ID, booster version, or launch date to ensure consistency.

Data Collection – SpaceX API

- We used the SpaceX public API to retrieve structured JSON data on Falcon 9 launches. This provided real-time details like launch dates, payloads, boosters, landing outcomes, orbits, and launch sites.

- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- To supplement missing data fields, we scraped structured tables from reliable sources like Wikipedia. Using Python libraries like BeautifulSoup and Requests, we extracted and parsed HTML tables into pandas DataFrames.
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Loaded raw launch data from SpaceX API in JSON format
- Parsed and structured data into Pandas DataFrames
- Cleaned missing or inconsistent fields (e.g., pad types, distances)
- Scraped supplemental data (e.g., from Wikipedia) using BeautifulSoup and Requests
- Calculated landing success rate using df["Class"].mean()
- Exported final dataset as dataset_part_2.csv for future analysis
- Filtered data by date range to ensure consistency across labs
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Bar Chart** – Compared landing success across launch sites
- **Pie Chart** – Showed distribution of landing outcomes
- **Scatter Plot** – Analyzed relationship between payload mass and success
- **Box Plot** – Compared payload mass ranges by launch site
- **Histogram** – Visualized distribution of payload masses
- **Line Chart** – Tracked launch success trends over time

- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Connected to the database and verified the table structure using SELECT * FROM SPACEXTBL LIMIT 5
- Retrieved unique launch sites using SELECT DISTINCT Launch_Site FROM SPACEXTBL
- Counted the total number of successful landings using SELECT COUNT(*) WHERE Landing_Outcome LIKE 'Success%'
- Extracted launch records where payload mass was between specified ranges (e.g., 4000–6000 kg)
- Filtered data for specific booster versions using WHERE Booster_Version = '...'
- Calculated the average payload mass per launch site using GROUP BY Launch_Site
- Identified missions with the maximum payload mass using ORDER BY Payload_Mass__kg_ DESC LIMIT 1
- Ranked landing outcomes within a date range using GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers**: Used to indicate specific locations (e.g., cities or landmarks).
- **Circle** or **CircleMarker**: Placed on a location to represent a region or area with a given radius.
- **Line (PolyLine)**: Used to draw a path or connection between multiple geographic points.
- Markers were added to highlight important locations on the map and provide interactive popups for additional information, such as the name of a city or a data point.
- Circles were used to visualize the magnitude or impact of a location-based attribute — for example, population size or event intensity.
- Lines (PolyLine) were included to represent travel routes, boundaries, or connections between points, which helped illustrate spatial relationships or paths.
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/Folium_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plots and Graphs Added:

- Success Pie Chart – Shows launch success counts by site or success vs. failure for selected site
- Success-Payload Scatter Plot – Displays correlation between payload mass and launch success; colored by Booster Version

Interactive Components:

- Dropdown Menu – Selects a specific launch site or all sites
- Range Slider – Filters data based on payload mass range (0–10,000 kg)

Purpose:

- Enable real-time visual analysis of launch success by site and payload
- Identify trends, such as which sites or payload ranges have higher success rates
- Explore booster performance through color-coded insights
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- **Data Preprocessing:** Clean data, handle missing values, encode categories
- **Train-Test Split:** 80% train, 20% test
- **Model Training:** Logistic Regression, SVM, Decision Tree, KNN
- **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix
- **Hyperparameter Tuning:** GridSearchCV for optimization
- **Best Model: KNN with 100% test accuracy**, outperforming others (SVM, Tree, Logistic Regression at 83.33%)
- https://github.com/Sarannya-Thelapurath/Applied-Data-Science-Capstone/blob/main/Predictive_AnalysisSpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

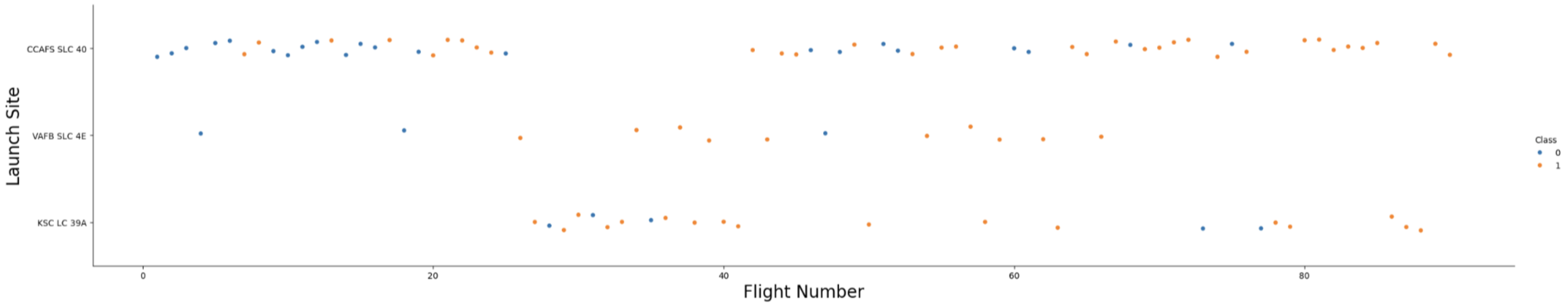
- **EDA:** Cleaned data, visualized features, identified key predictors
- **Interactive Demo:** Dynamic site dropdown & payload slider; pie & scatter charts update live
- **Predictive Results:**
  - Tested models: Logistic Regression, SVM, Decision Tree, KNN
  - KNN best with 100% test accuracy
  - Others ~83% accuracy
  - Used GridSearchCV to tune models

Section 2

Insights drawn from EDA

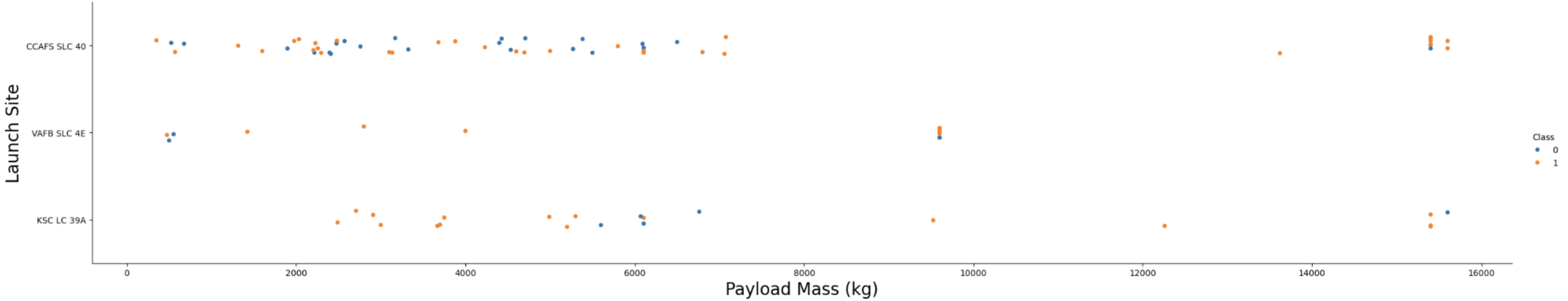
Flight Number vs. Launch Site

- Some launch sites may show more dense flight activity (more dots clustered).
- Sites with mostly green dots (Class=1) indicate higher success rates.
- Sites with red dots (Class=0) indicate failures; spotting these helps identify problem areas or early learning curves.
- Trends in success over flight numbers may show improvement or consistency in certain sites.



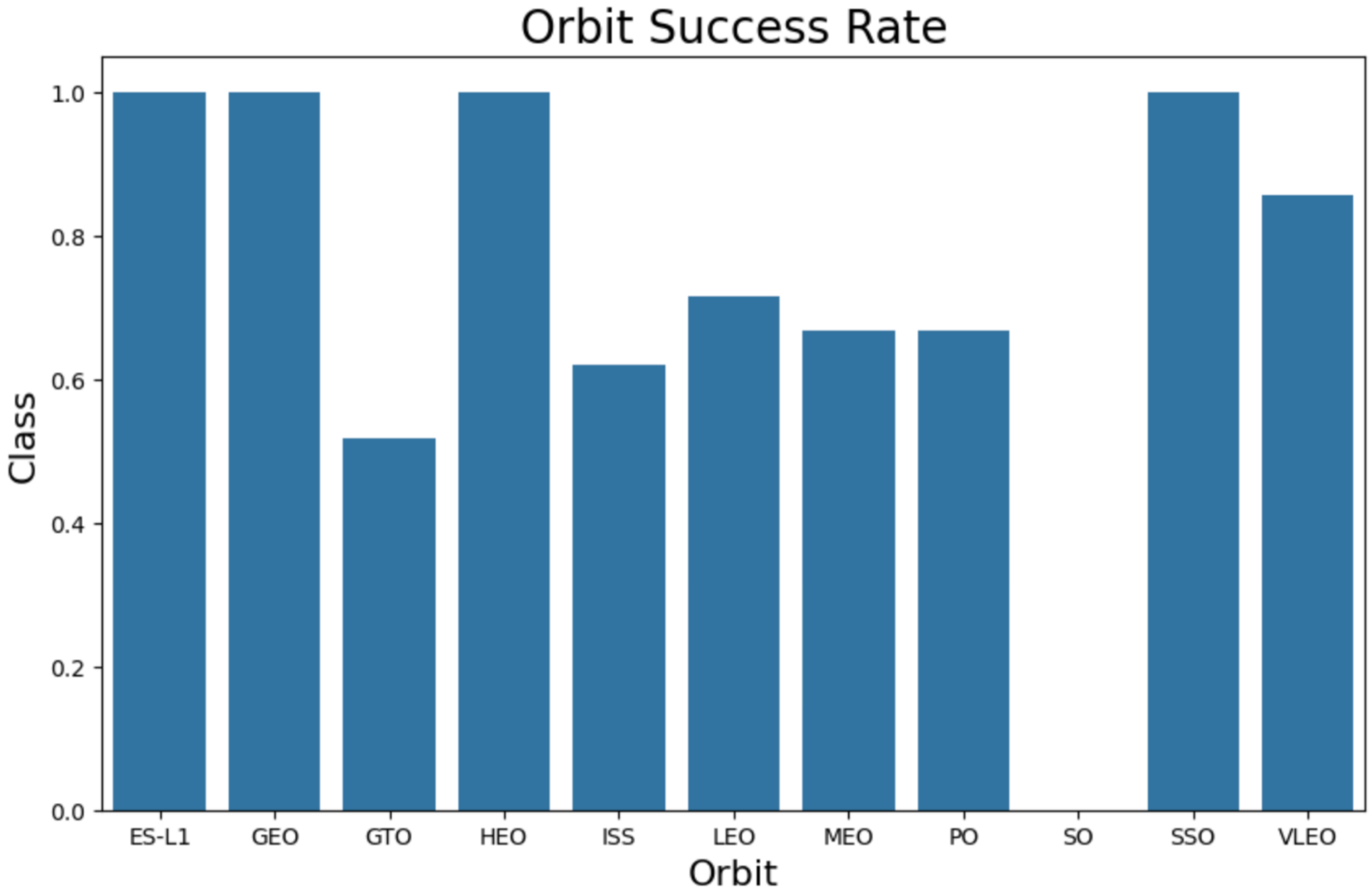
Payload vs. Launch Site

- Payload mass varies across launch sites, reflecting differences in rocket capacity or mission type.
- Clusters of successful launches at specific payload ranges indicate optimal weights for mission success at each site.
- Failures at certain payload masses reveal riskier payloads, helping identify how payload affects success by launch site.



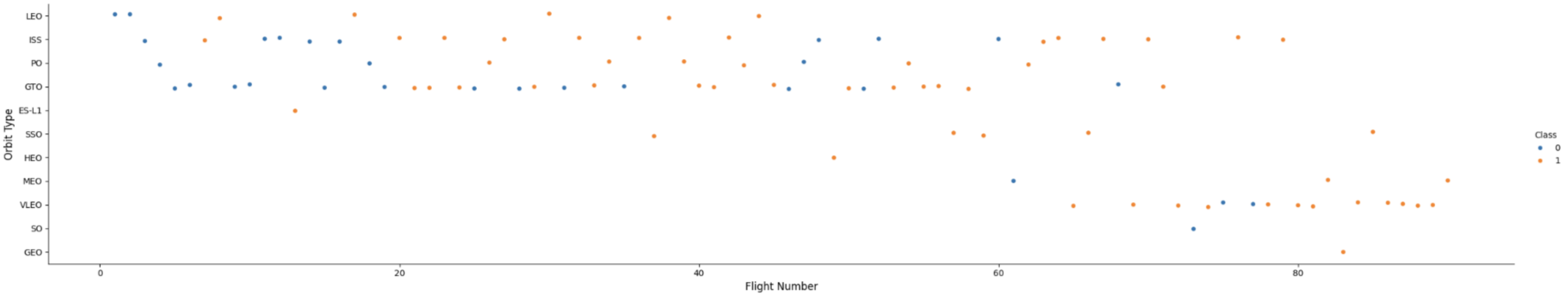
Success Rate vs. Orbit Type

- The plot allows comparison of mission success across different orbit categories.
- Some orbits have consistently higher success rates, suggesting easier or more reliable missions.
- Other orbits may have lower success rates, possibly due to higher mission complexity or technical challenges.
- This visualization helps identify which orbit types are more successful for launches.



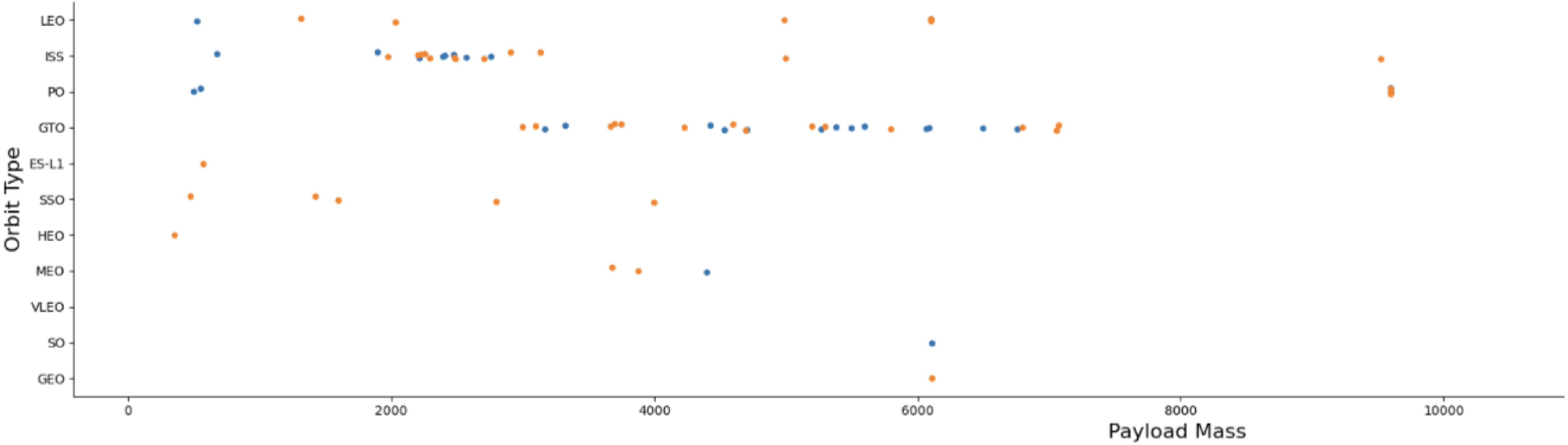
Flight Number vs. Orbit Type

- You can observe trends in success and failure rates over time for different orbits.
- Certain orbits may have more launches (denser points) or higher success rates as flight numbers increase.
- Failures might cluster at earlier flight numbers, showing learning or reliability improvements over time.
- This helps analyze how mission success varies by orbit type and flight number



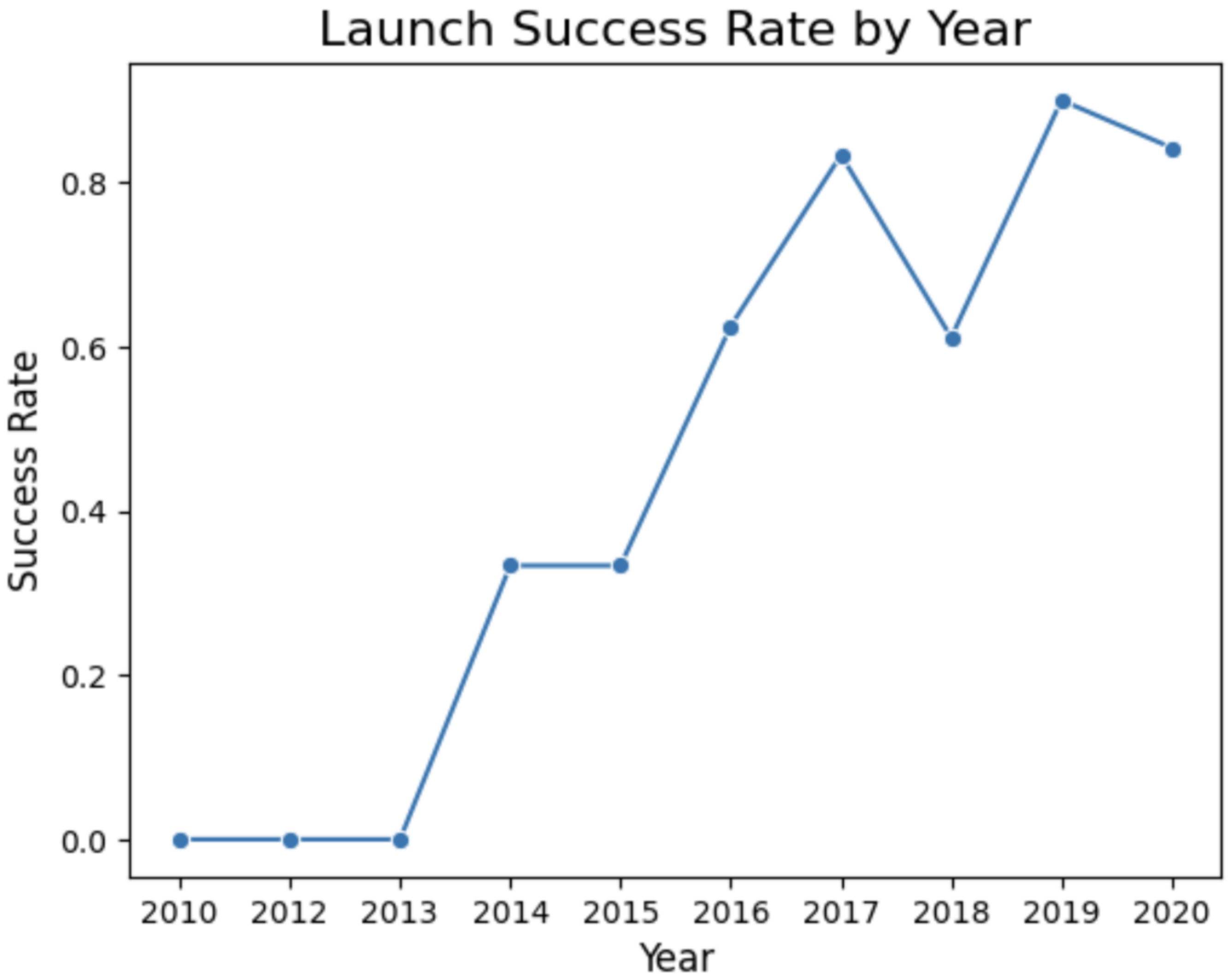
Payload vs. Orbit Type

- Certain payload mass ranges may be more common for specific orbit types.
- Clusters of successful launches at particular payload masses suggest ideal payloads for those orbits.
- Failed launches clustered at specific payload masses indicate riskier payload weights for some orbits.
- Overall, this visual helps assess how payload mass impacts success for different orbits.



Launch Success Yearly Trend

- The plot reveals trends in launch success rates over time.
- An upward trend indicates improving reliability and success in launches as years progress.
- Success rate since 2013 kept increasing till 202
- This visualization helps understand how mission success evolves historically.

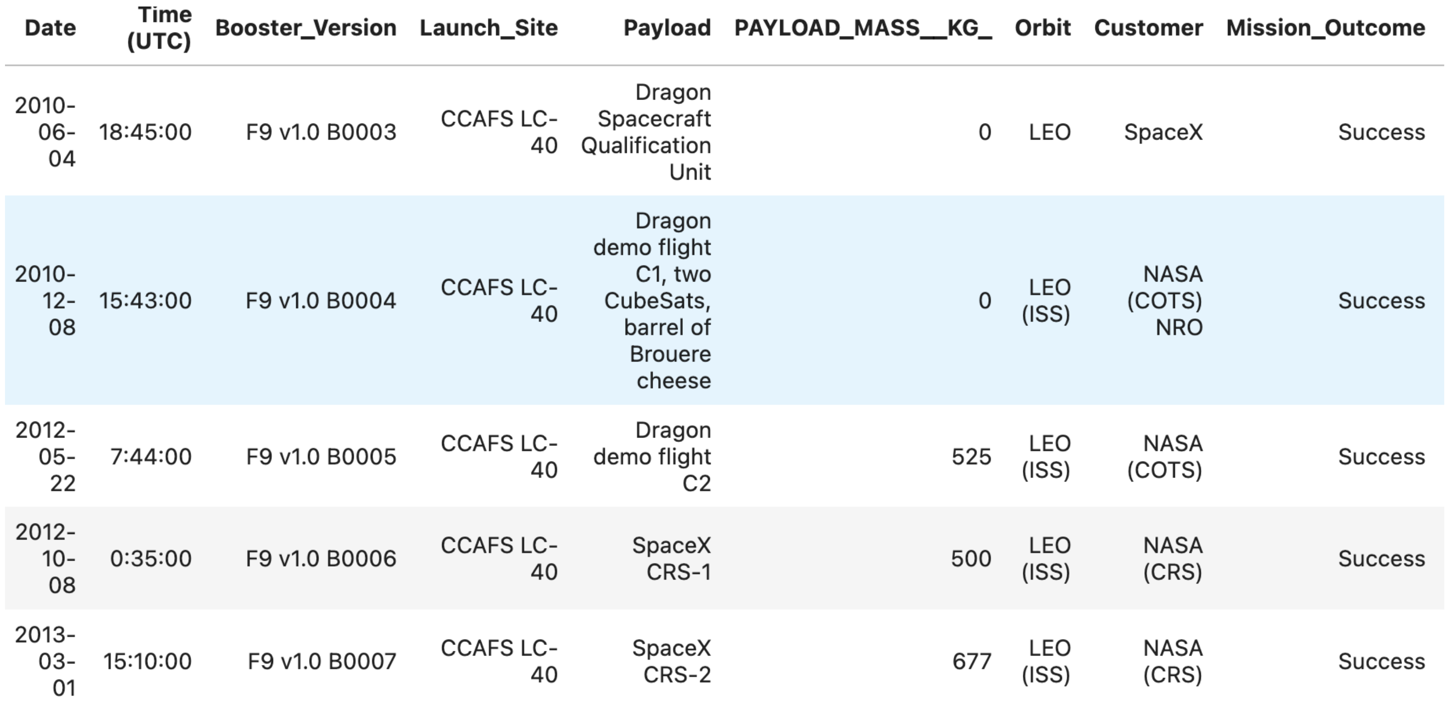


All Launch Site Names

- %sql SELECT DISTINCT "Launch Site" FROM SPACEXTBL
- This SQL query retrieves all unique launch site names from the SPACEXTBL table. The DISTINCT keyword ensures no duplicates in the result, helping identify all different launch locations used for SpaceX missions in the dataset.
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The query retrieves 5 records where the launch site name starts with "CCA," filtering for Cape Canaveral Air Force Station sites.
- This helps focus analysis on launches specifically from those locations by showing a sample of relevant data.



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- %sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_payload FROM SPACEXTBL WHERE "Booster_Version" LIKE '%NASA (CRS)%'
- The query calculates the total payload mass (in kg) carried by boosters labeled with NASA (CRS) in their version name.
- This query sums the payload mass for all launches involving NASA (CRS) boosters, providing insight into the total cargo these boosters have delivered. It helps quantify the contribution of NASA-supported missions in terms of payload weight.

Average Payload Mass by F9 v1.1

- %sql SELECT AVG("PAYLOAD_MASS__KG_") AS AVG_PAYLOAD FROM SPACEXTBL WHERE "Booster_Version" LIKE '%F9 v1.1%’
- The average payload mass carried by boosters with version containing “F9 v1.1” is 2534.67 kg.
- This query calculates the average payload weight for launches using the Falcon 9 version 1.1 boosters, providing insight into the typical cargo capacity handled by this booster type.

First Successful Ground Landing Date

- %sql SELECT MIN(Date) AS First_Landing FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)’
- The first successful landing on a ground pad occurred on 2015-12-22.
- This query finds the earliest date when a rocket successfully landed on a ground pad, marking a key milestone in SpaceX’s reusable launch technology achievements.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_"< 6000
- Boosters with successful landings on drone ships and payloads between 4000 and 6000 kg are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- This query identifies specific booster versions that achieved successful drone ship landings while carrying medium-heavy payloads, highlighting their reliability in these mission conditions.

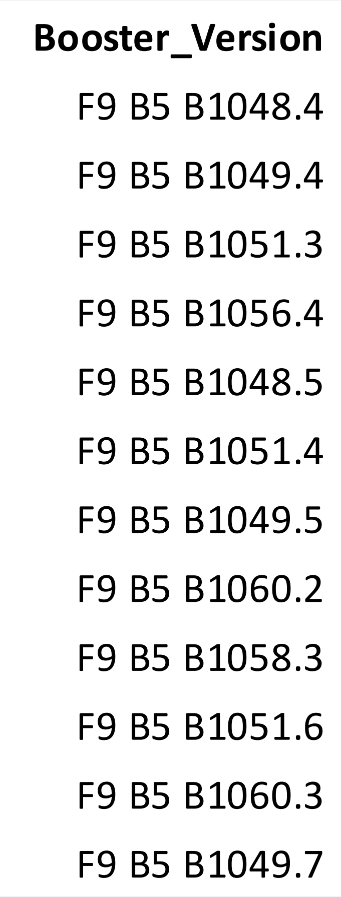
Total Number of Successful and Failure Mission Outcomes

- %sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTBL GROUP BY "Mission_Outcome”
- The query counts the number of missions by their outcome. There were 99 successful missions (including one with unclear payload status) and 1 failure (in flight). This shows a high success rate for the launches in the dataset.

Mission Outcome	**Total Count**
Failure (in flight)	1
Success	99 (98 + 1)
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
- This query identifies the booster versions that have carried the highest payload recorded in the dataset, highlighting the most powerful boosters in terms of pay Capacity.



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

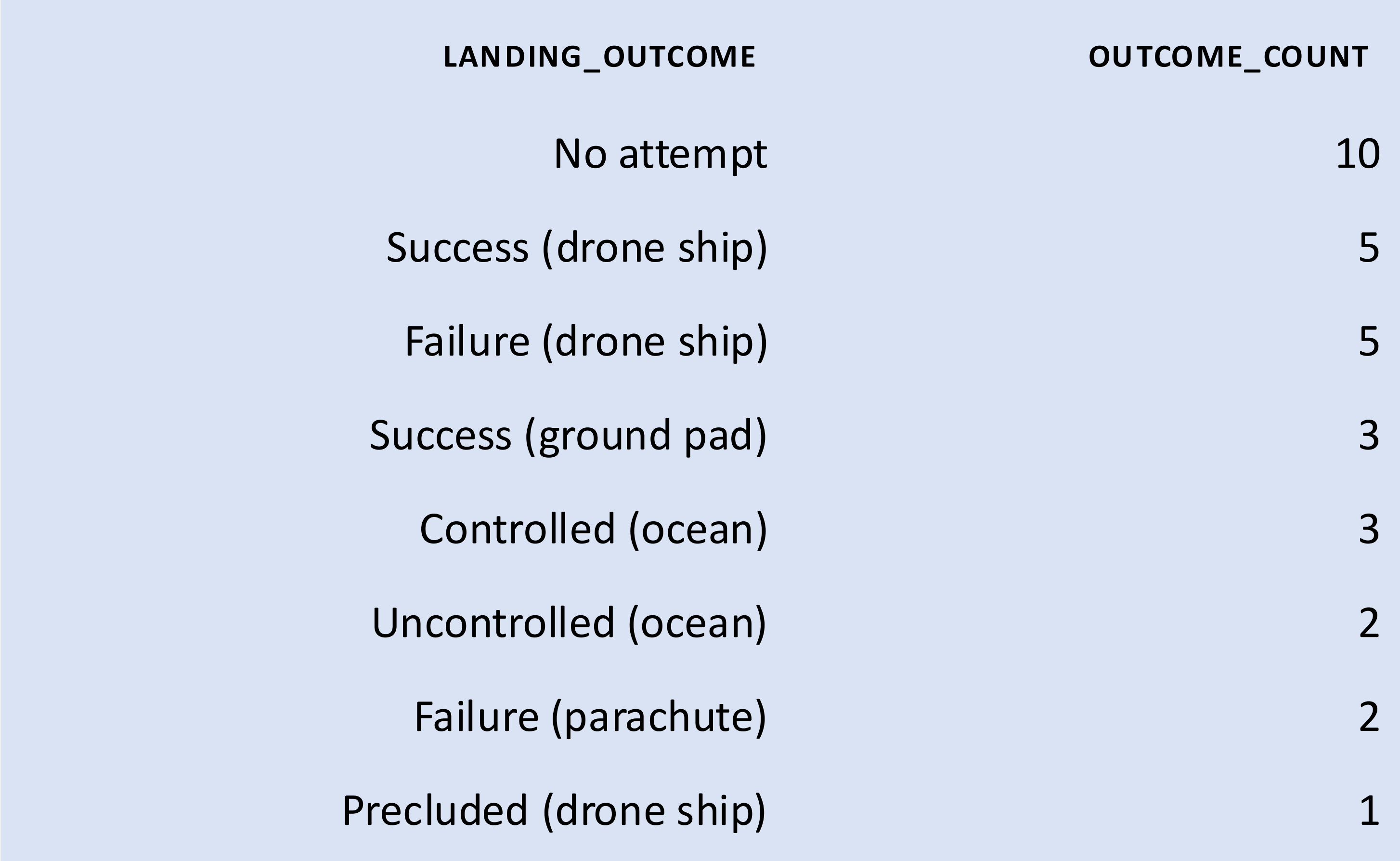
2015 Launch Records

- This query lists all failed landings on drone ships in 2015, showing the month of failure, booster version involved, and the launch site. It highlights two specific failures occurring in January and April at the CCAFS LC-40 site with F9 v1.1 boosters.

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query ranks different landing outcomes between June 4, 2010, and March 20, 2017, by their frequency in descending order. The most common outcome was “No attempt” (10 times), followed by both “Success (drone ship)” and “Failure (drone ship)” (5 times each). This shows the varied success and failure rates of drone ship landings within the given period.



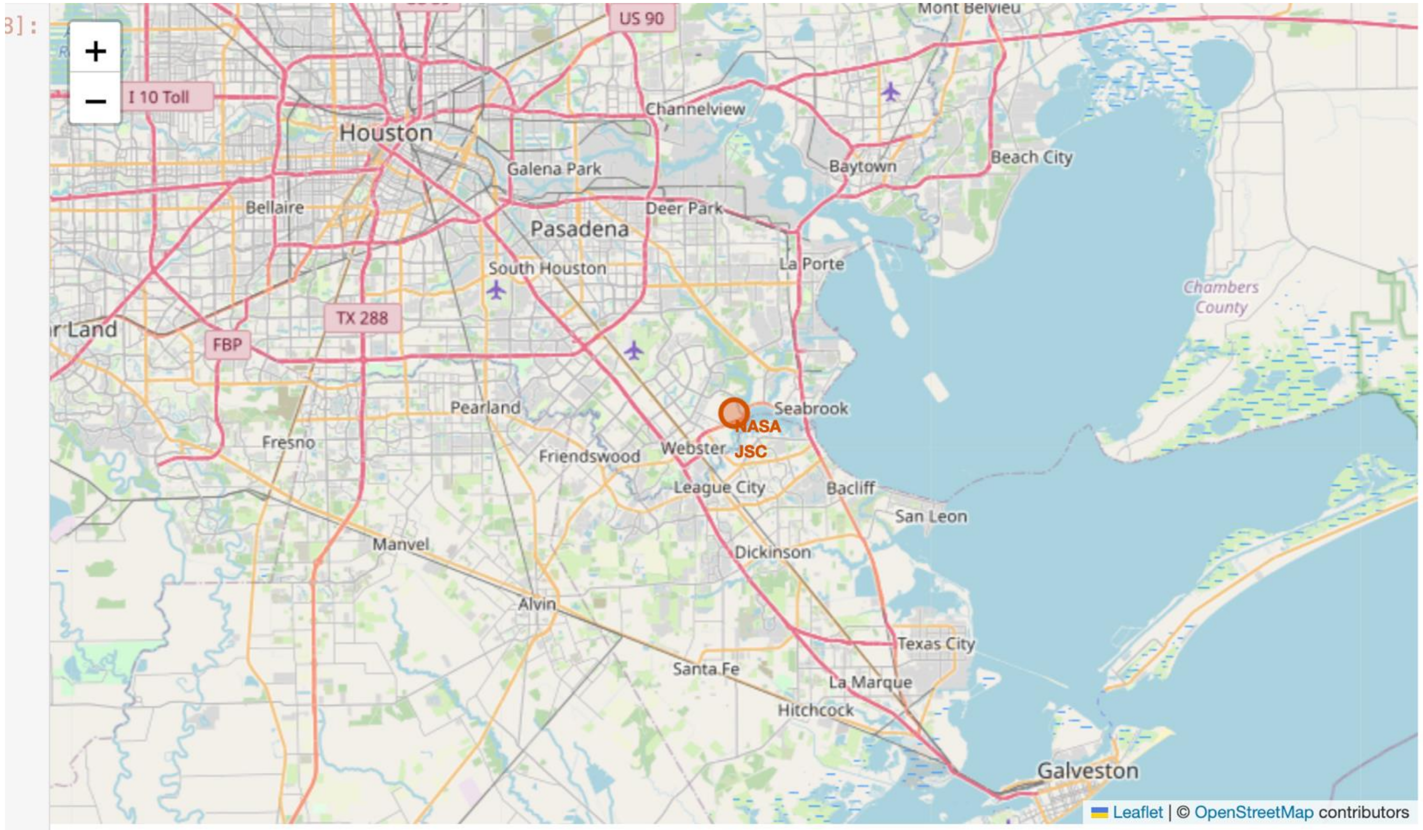
LANDING_OUTCOME	OUTCOME_COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

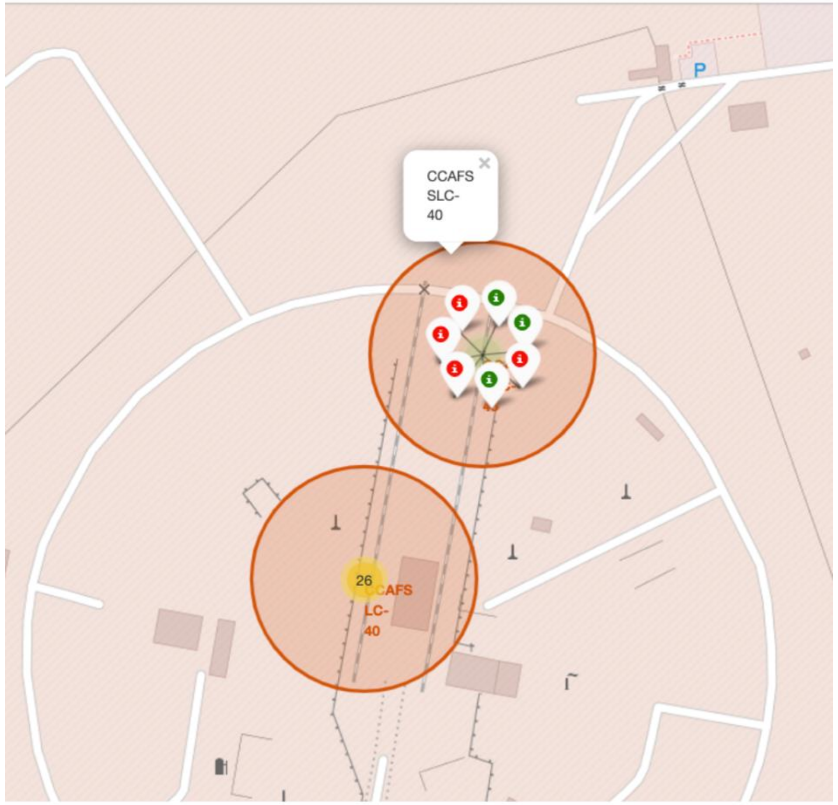
Launch Sites Location Map

- The map displays multiple launch sites as circles, with the NASA Johnson Space Center prominently highlighted in orange as a key reference point.
- It is zoomed in to show detailed spatial relationships around NASA Johnson Space Center, enabling clear visualization of nearby launch sites.
- Interactive markers provide labels or popups for easy site identification, supporting analysis of geographical distribution and related metrics like success rates or proximity.



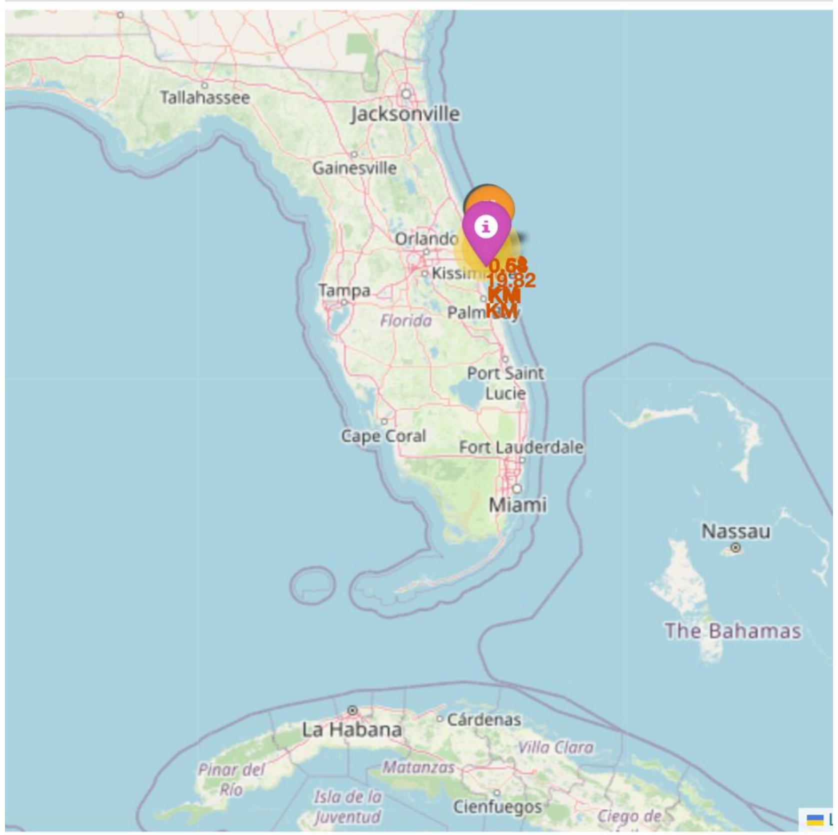
Map Launch Success and Failure per Site

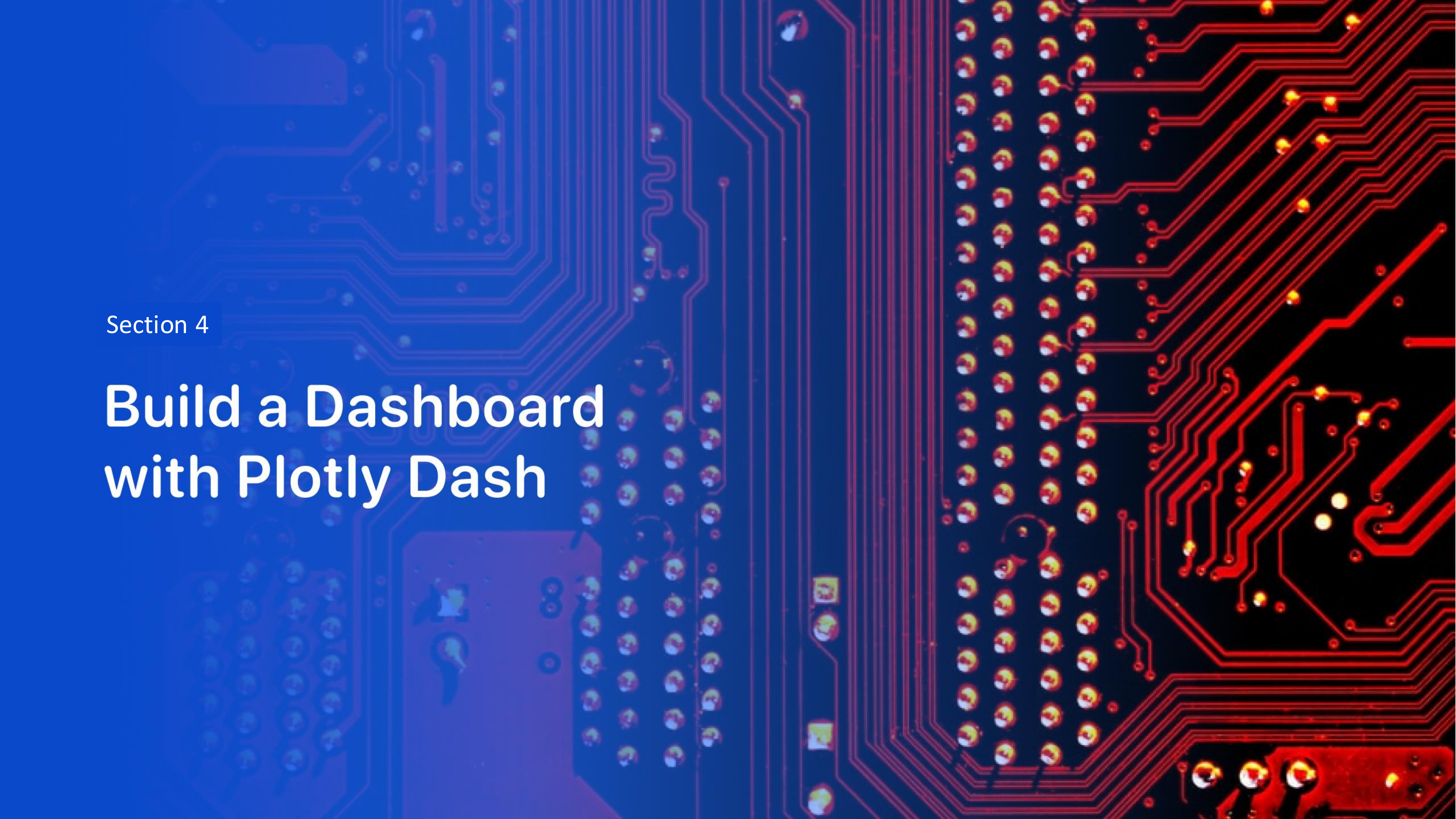
- Color-Coded Markers: Green markers indicate successful launches, while red markers show failures, making it easy to visually assess outcomes at each site.
- Interactive Popups and Clustering: Each marker includes a popup with the launch site name and outcome, and nearby markers are grouped using clustering for better readability.
- Key Insights: The map reveals geographic patterns in launch success and failure, helping identify high- or low-performing sites based on location.



Spatial Distances Among Launch Sites

- Labeled Proximities with Distances: The map shows the launch site (CCAFS LC-40) connected to its nearest city, railway, and highway using colored lines and labeled distance markers for clear spatial reference.
- Color-Coded Icons and Lines: Each proximity type is represented with distinct colors and icons—purple for city, gray for railway, and orange for highway—making the map easy to interpret visually.
- Key Insight: The close distances to infrastructure suggest the launch site is well-positioned for accessibility, logistics, and emergency support planning.



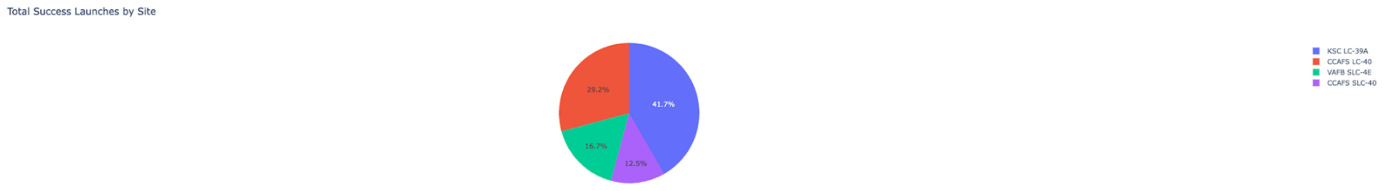


Section 4

Build a Dashboard with Plotly Dash

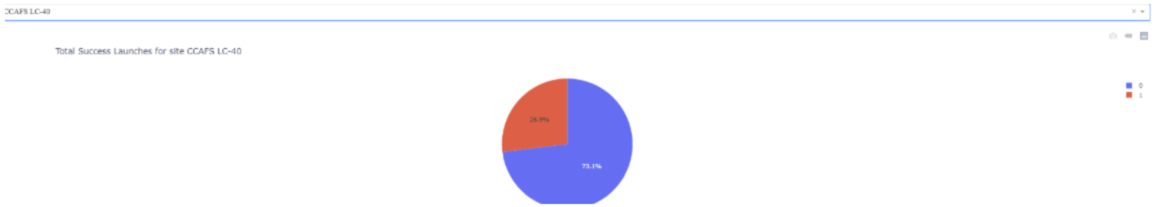
Launch Success Pie Chart by Site

- **Interactive Dropdown Control:** Users select either "All Sites" or a specific launch site from the dropdown, triggering an automatic update of the pie chart.
- **Pie Chart Shows Launch Outcomes:** For "All Sites," the chart displays total successful launches. For a specific site, it visualizes the counts of successes and failures distinctly.
- **Insightful Performance Comparison:** The chart highlights overall launch success rates and helps identify which launch sites perform reliably versus those with more failures, supporting data-driven decisions.



Launch Site with Highest Success Rate

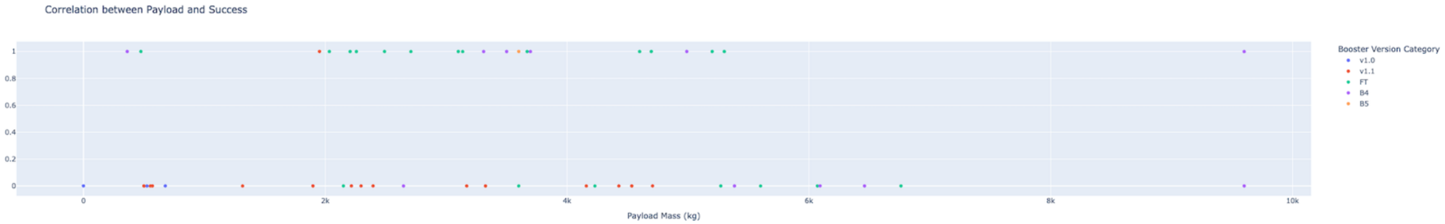
- The pie chart shows the proportion of successful versus failed launches at the selected launch site, with clear color-coded slices.
- A large success slice indicates the site’s high launch reliability and minimal failures.
- This visual helps quickly identify the most dependable launch site and supports informed decisions for future missions.



CCAFS LC-40

Scatter plot of Payload vs. Launch Outcome

- **Interactive Scatter Plot:** The chart updates based on the selected launch site and payload mass range, showing payload on the x-axis and launch success (class) on the y-axis.
- **Booster Version Color Coding:** Each point is color-coded by booster version category, helping identify which boosters are associated with successful or failed launches.
- **Insights on Payload and Success:** The plot reveals correlations between payload mass and launch outcomes, highlighting payload ranges and booster types with higher success rates, both overall and per site.

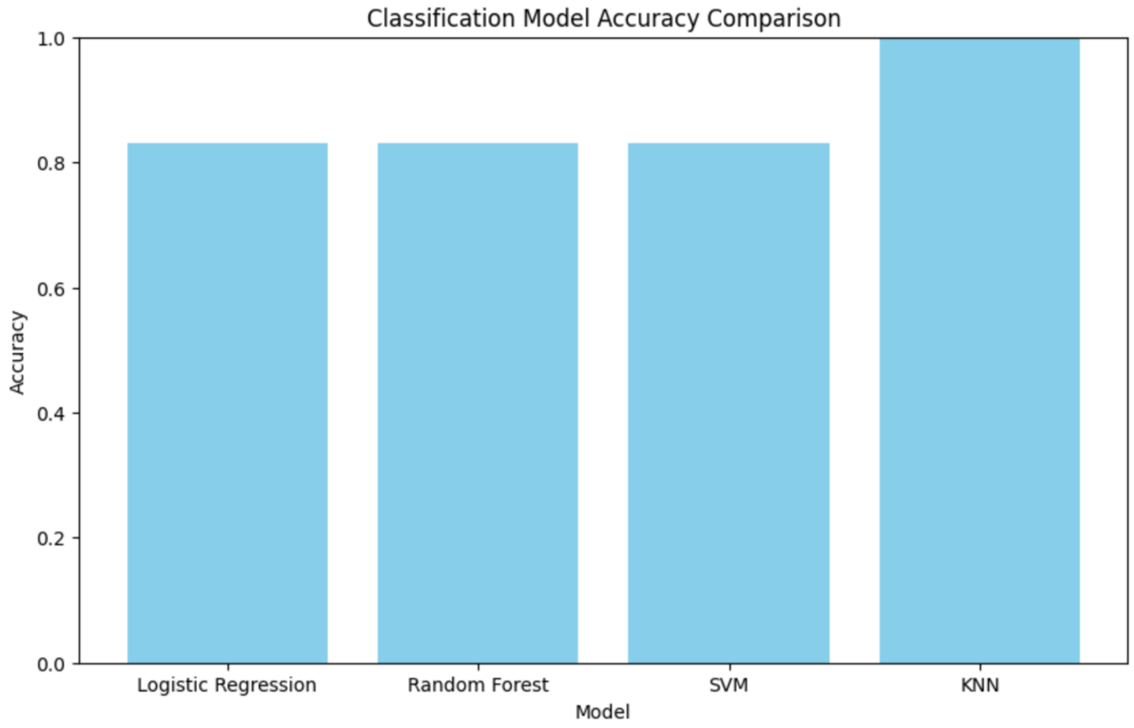


Section 5

Predictive Analysis (Classification)

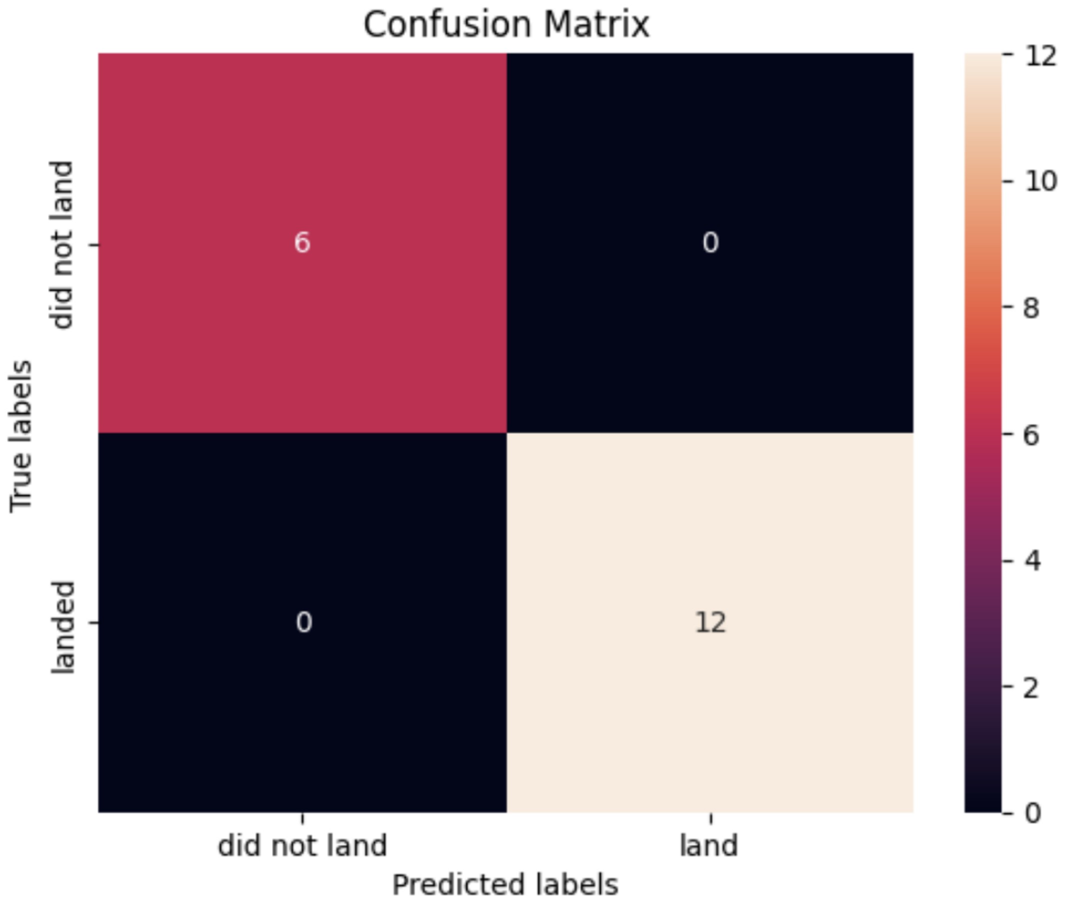
Classification Accuracy

- The best Model is KNN with 100% test accuracy, outperforming others (SVM, Decision Tree, Logistic Regression at 83.33%)



Confusion Matrix

- The KNN model perfectly separates the classes on the test set, suggesting excellent generalization for this data.
- The confusion matrix visually confirms the accuracy score by showing zero classification errors.
- It's a strong indicator of model reliability on unseen data



Conclusions

- High Accuracy Models: Several classification models show strong performance in predicting Falcon 9 first stage landing success.
- Payload and Booster Impact: Payload weight and booster version significantly influence landing outcomes, as visualized in scatter plots.
- Launch Site Matters: Different launch sites exhibit varying success rates, highlighting location as an important factor.
- Best Model Validation: The KNN model achieved perfect test accuracy with zero prediction errors, confirming its reliability on test data.
- Practical Application: These insights can help optimize launch parameters and improve mission planning for successful first stage landings.

Appendix

- Dataset: SpaceX launch records including launch site, payload mass, booster version, and launch outcome (success/failure).
- Models Tested: KNN, Random Forest, SVM, Logistic Regression, Gradient Boosting with hyperparameter tuning via cross-validation.
- Key Visuals:
  - Pie charts showing launch success by site
  - Scatter plots of payload vs. success colored by booster version
  - Bar chart comparing model accuracies
  - Confusion matrix confirming KNN’s perfect accuracy
- Limitations & Future Work:
  - Dataset size and class imbalance may affect model generalization
  - Future work includes testing on larger datasets and exploring deep learning models

Thank you!