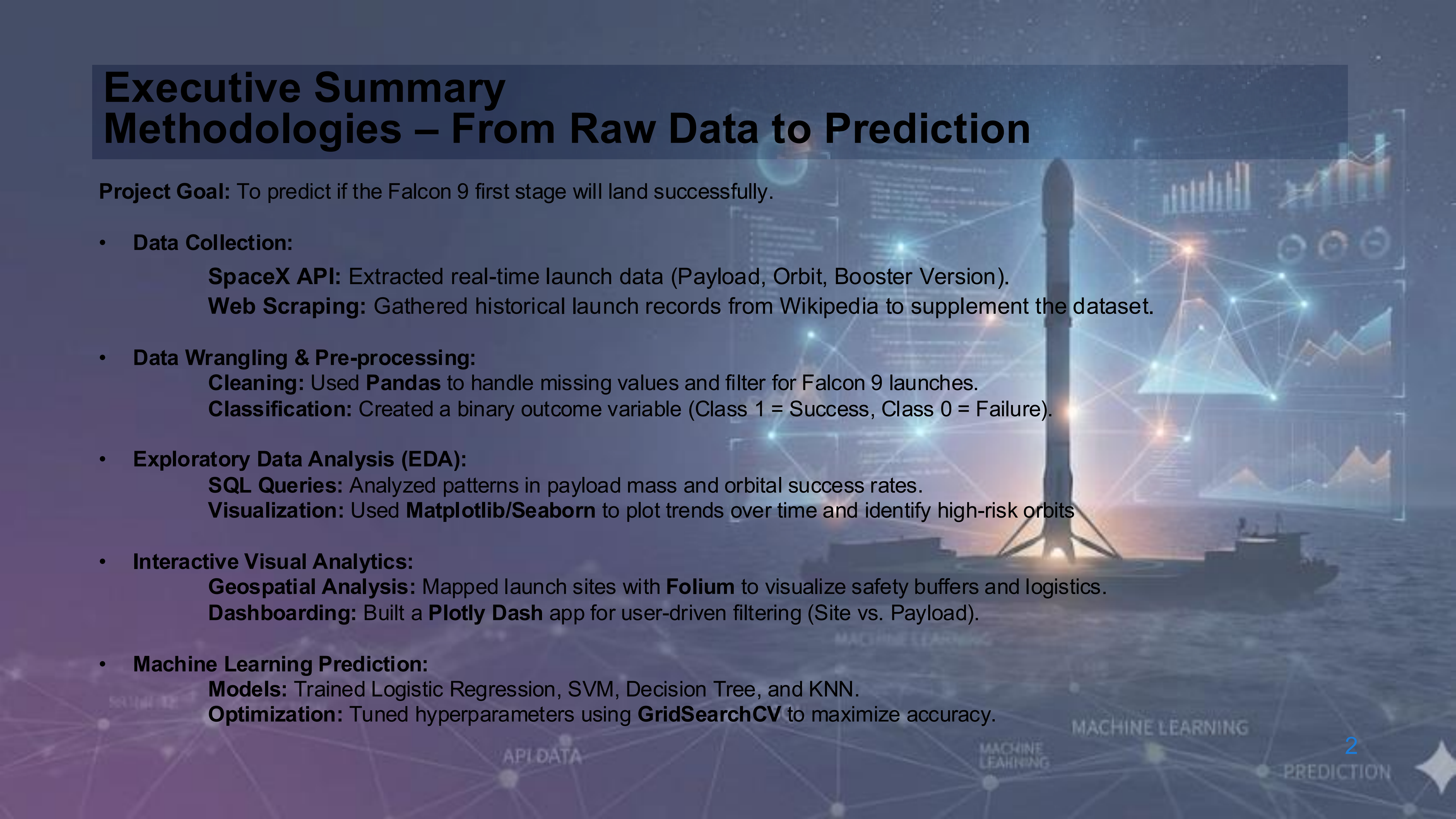
IBM Developer
SKILLS NETWORK

George Monappallil
December 16 2025

Executive Summary
Methodologies – From Raw Data to Prediction



Project Goal: To predict if the Falcon 9 first stage will land successfully.

- **Data Collection:**
  - **SpaceX API:** Extracted real-time launch data (Payload, Orbit, Booster Version).
  - **Web Scraping:** Gathered historical launch records from Wikipedia to supplement the dataset.

- **Data Wrangling & Pre-processing:**
  - **Cleaning:** Used **Pandas** to handle missing values and filter for Falcon 9 launches.
  - **Classification:** Created a binary outcome variable (Class 1 = Success, Class 0 = Failure).

- **Exploratory Data Analysis (EDA):**
  - **SQL Queries:** Analyzed patterns in payload mass and orbital success rates.
  - **Visualization:** Used **Matplotlib/Seaborn** to plot trends over time and identify high-risk orbits

- **Interactive Visual Analytics:**
  - **Geospatial Analysis:** Mapped launch sites with **Folium** to visualize safety buffers and logistics.
  - **Dashboarding:** Built a **Plotly Dash** app for user-driven filtering (Site vs. Payload).

- **Machine Learning Prediction:**
  - **Models:** Trained Logistic Regression, SVM, Decision Tree, and KNN.
  - **Optimization:** Tuned hyperparameters using **GridSearchCV** to maximize accuracy.

Executive Summary
Methodologies – From Raw Data to Prediction

1. **Operational Insights (EDA & SQL)**

   Success Rate Evolution: Success rates have improved significantly over time. The "Full Thrust" (FT) and "Block 5" booster versions stabilized the program, with the first successful drone ship landing occurring on April 8, 2016.

   Launch Site Performance: KSC LC-39A (Kennedy Space Center) is the most productive site, accounting for the highest volume of successful commercial launches (approx. 42%).

2. **Visual & Geospatial Findings**

   Location Strategy: All launch sites are located on coastlines for safety, but extremely close to railway lines for the logistical transport of heavy boosters.

   Payload Reliability: The Falcon 9 is highly reliable across the board but shows a distinct "sweet spot" for commercial payloads between **2,000kg and 5,500kg**. Surprisingly, heavy payloads (>8,000kg) also have a near-perfect success record.

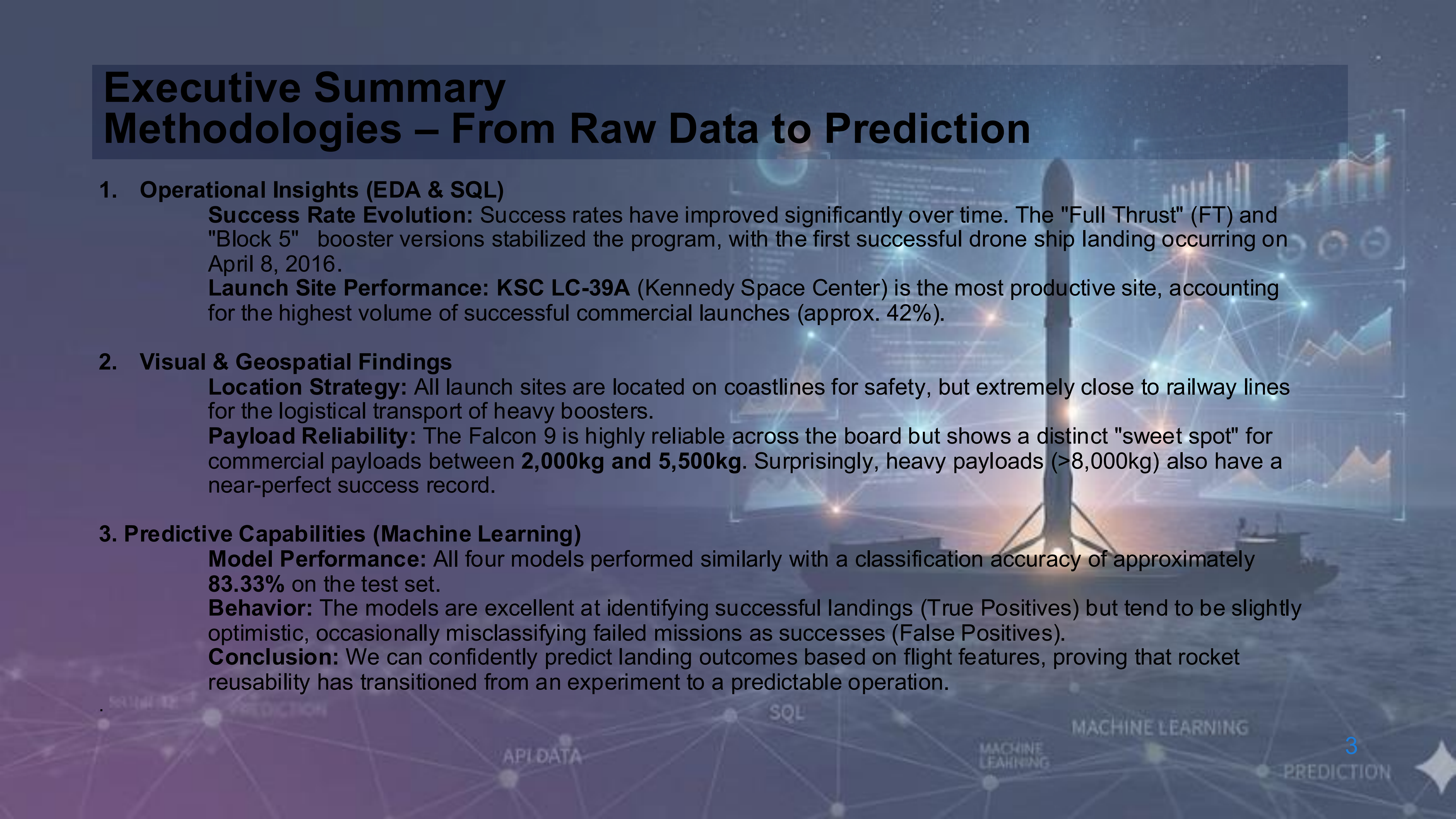
3. Predictive Capabilities (Machine Learning)

Model Performance: All four models performed similarly with a classification accuracy of approximately **83.33%** on the test set.

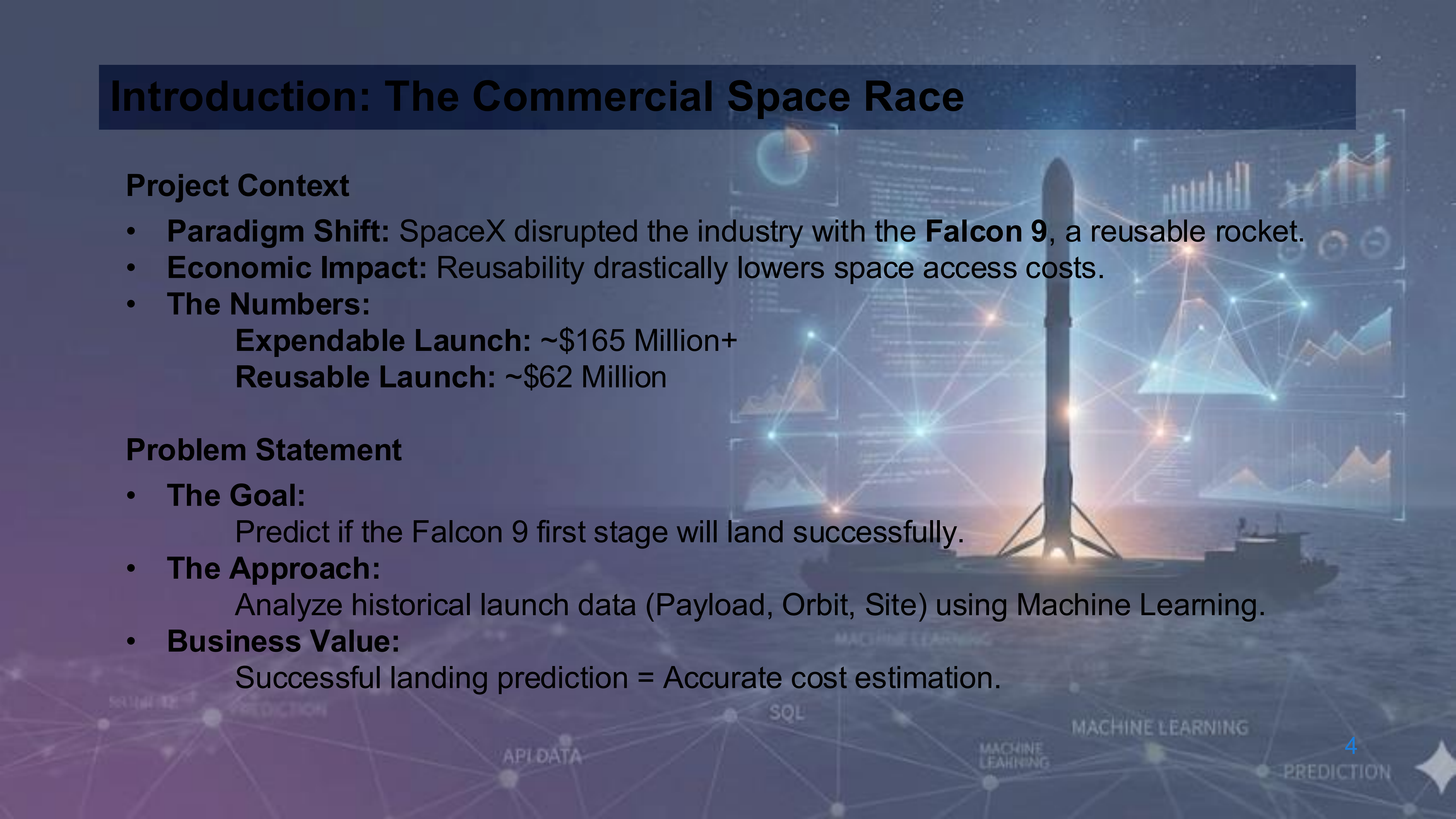
Behavior: The models are excellent at identifying successful landings (True Positives) but tend to be slightly optimistic, occasionally misclassifying failed missions as successes (False Positives).

Conclusion: We can confidently predict landing outcomes based on flight features, proving that rocket reusability has transitioned from an experiment to a predictable operation.

.



Introduction: The Commercial Space Race



Project Context

- **Paradigm Shift:** SpaceX disrupted the industry with the **Falcon 9**, a reusable rocket.
- **Economic Impact:** Reusability drastically lowers space access costs.
- **The Numbers:**
  - **Expendable Launch:** ~$165 Million+
  - **Reusable Launch:** ~$62 Million

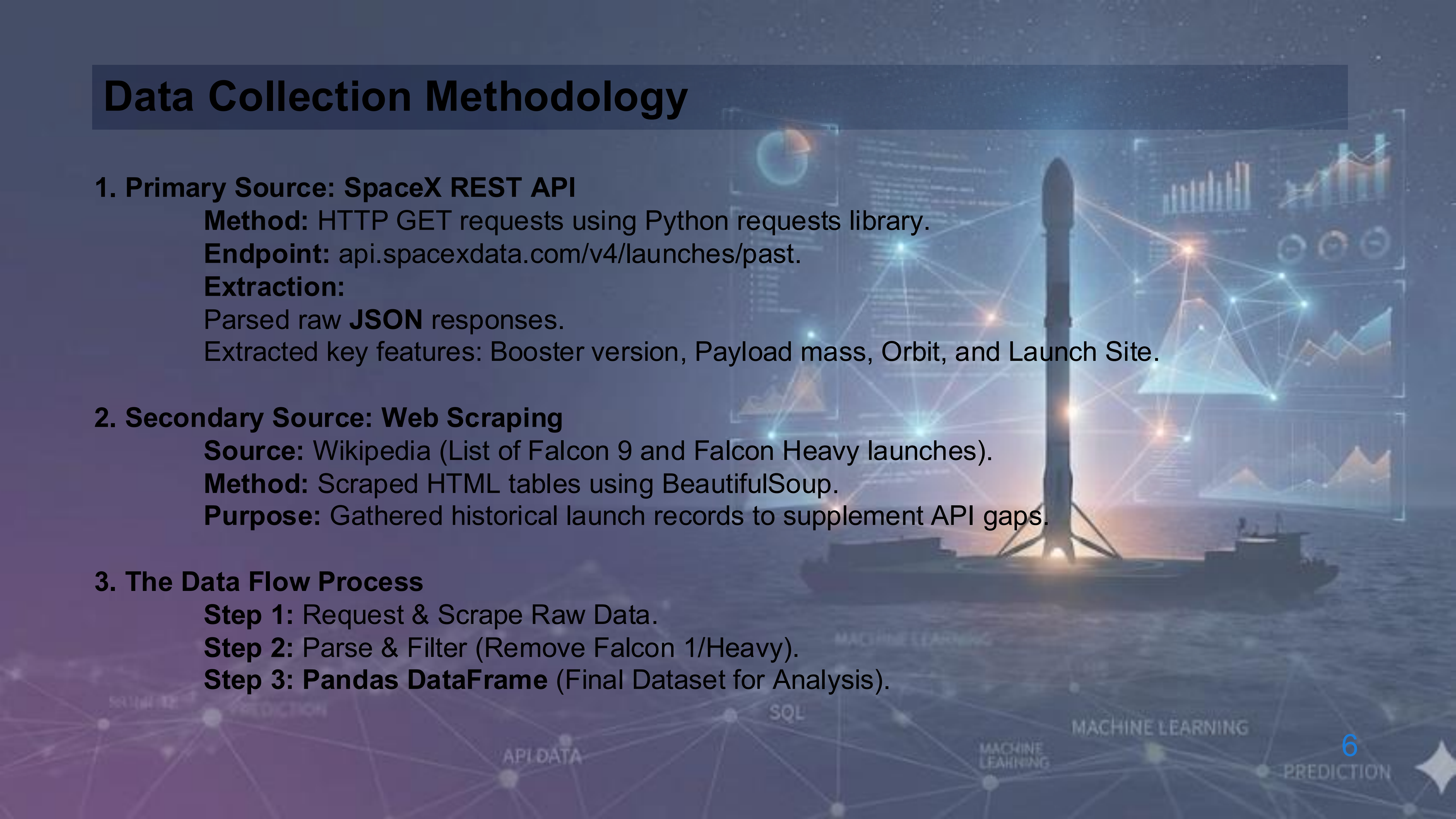
Problem Statement

- **The Goal:**
  Predict if the Falcon 9 first stage will land successfully.
- **The Approach:**
  Analyze historical launch data (Payload, Orbit, Site) using Machine Learning.
- **Business Value:**
  Successful landing prediction = Accurate cost estimation.

Section 1

MACHINE LEARNING
PREDICTION
SQL
MACHINE LEARNING
API DATA
MACHINE LEARNING
PREDICTION

Data Collection Methodology



1. Primary Source: SpaceX REST API

Method: HTTP GET requests using Python requests library.
Endpoint: api.spacexdata.com/v4/launches/past.
Extraction:
Parsed raw **JSON** responses.
Extracted key features: Booster version, Payload mass, Orbit, and Launch Site.

2. Secondary Source: Web Scraping

Source: Wikipedia (List of Falcon 9 and Falcon Heavy launches).
Method: Scraped HTML tables using BeautifulSoup.
Purpose: Gathered historical launch records to supplement API gaps.

3. The Data Flow Process

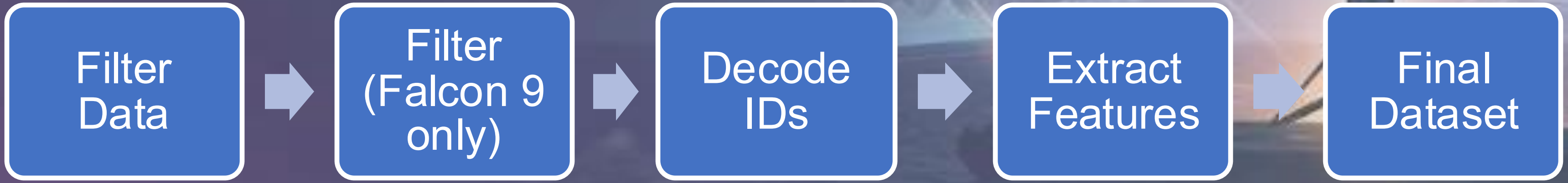
Step 1: Request & Scrape Raw Data.
Step 2: Parse & Filter (Remove Falcon 1/Heavy).
Step 3: Pandas DataFrame (Final Dataset for Analysis).

Data collection – SpaceX API

1. The Extraction Process

Source: Public SpaceX API (api.spacexdata.com/v4/launches/past).
Method: Utilized Python requests library to send GET calls.
Data Format: Responses received in **JSON** and normalized into a **Pandas DataFrame**.

2. Data Enrichment Flow The raw API data required specific filtering and decoding. We implemented the following logic:



GitHub URL:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/10.2/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb#:~:text=SpaceX_Machine,-Learning%20Prediction_Part_5.ipynb

Data collection – Scraping

1. Objective & Source

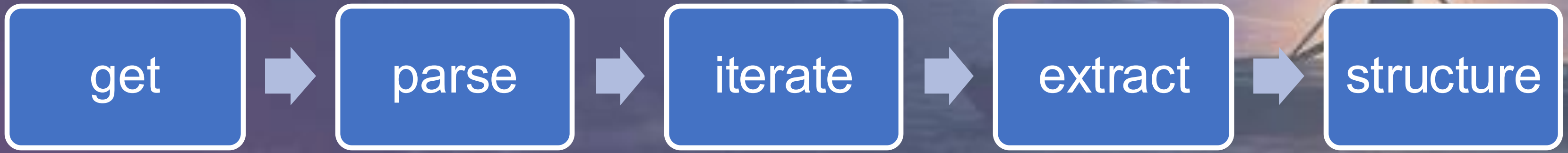
Goal: Supplement API data with historical launch records (dates, outcomes, payload details).

Source: Wikipedia – "List of Falcon 9 and Falcon Heavy launches".

2. Technical Approach

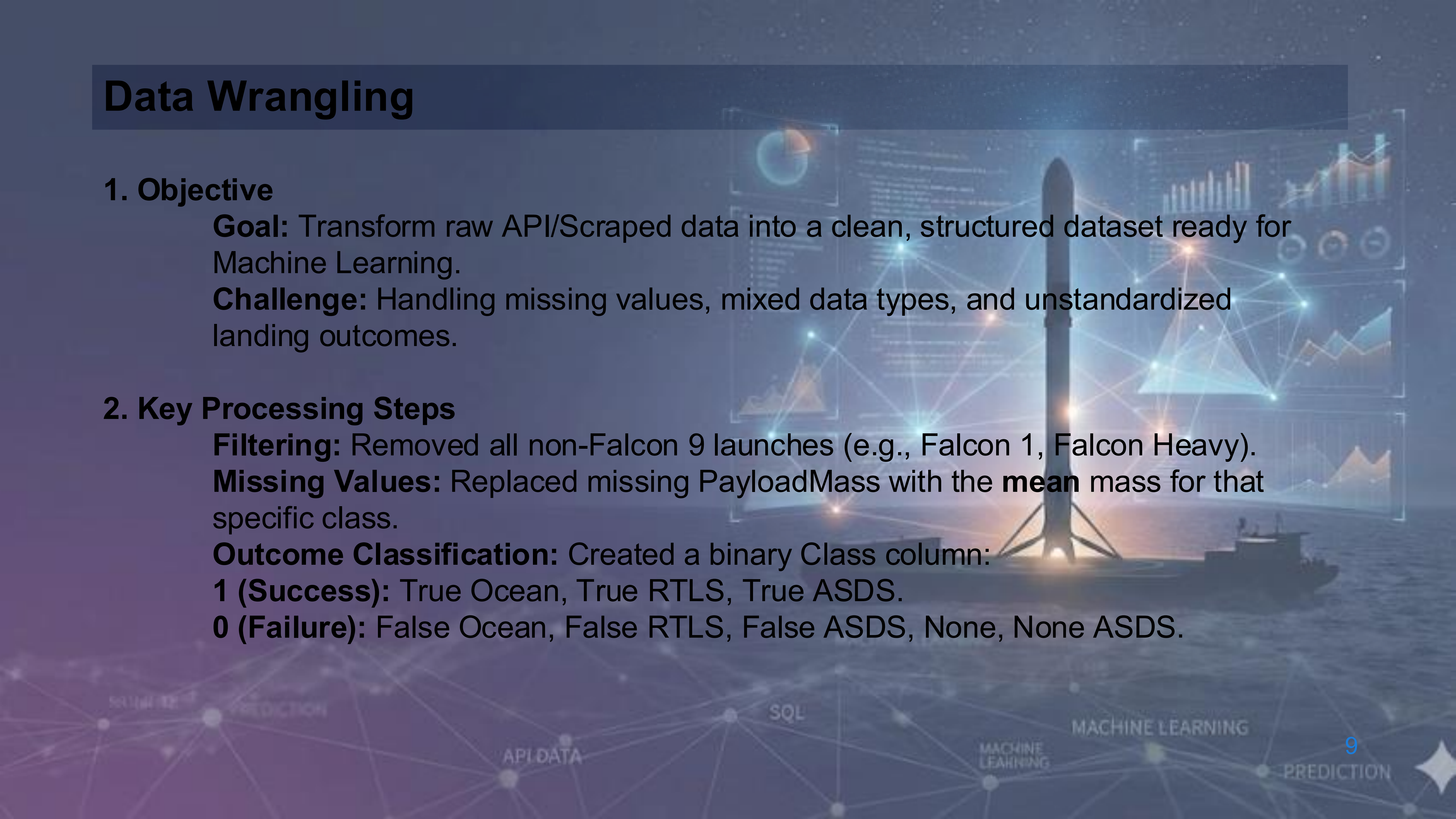
Tools: Python requests (for fetching HTML) and BeautifulSoup (for parsing tags).

Target: Extracted data specifically from HTML tables with the class wikitable plainrowheaders.



GitHub URL:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/9.4/9.4-Web-Scraper.ipynb#:~:text=9.4%2DWeb%2D-,Scraper.ipynb

Data Wrangling



1. Objective

Goal: Transform raw API/Scraped data into a clean, structured dataset ready for Machine Learning.

Challenge: Handling missing values, mixed data types, and unstandardized landing outcomes.

2. Key Processing Steps

Filtering: Removed all non-Falcon 9 launches (e.g., Falcon 1, Falcon Heavy).

Missing Values: Replaced missing PayloadMass with the **mean** mass for that specific class.

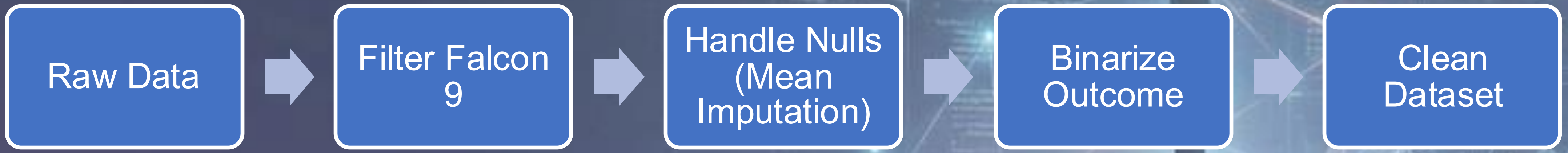
Outcome Classification: Created a binary Class column:

1 (Success): True Ocean, True RTLS, True ASDS.

0 (Failure): False Ocean, False RTLS, False ASDS, None, None ASDS.

Data wrangling (contd.)

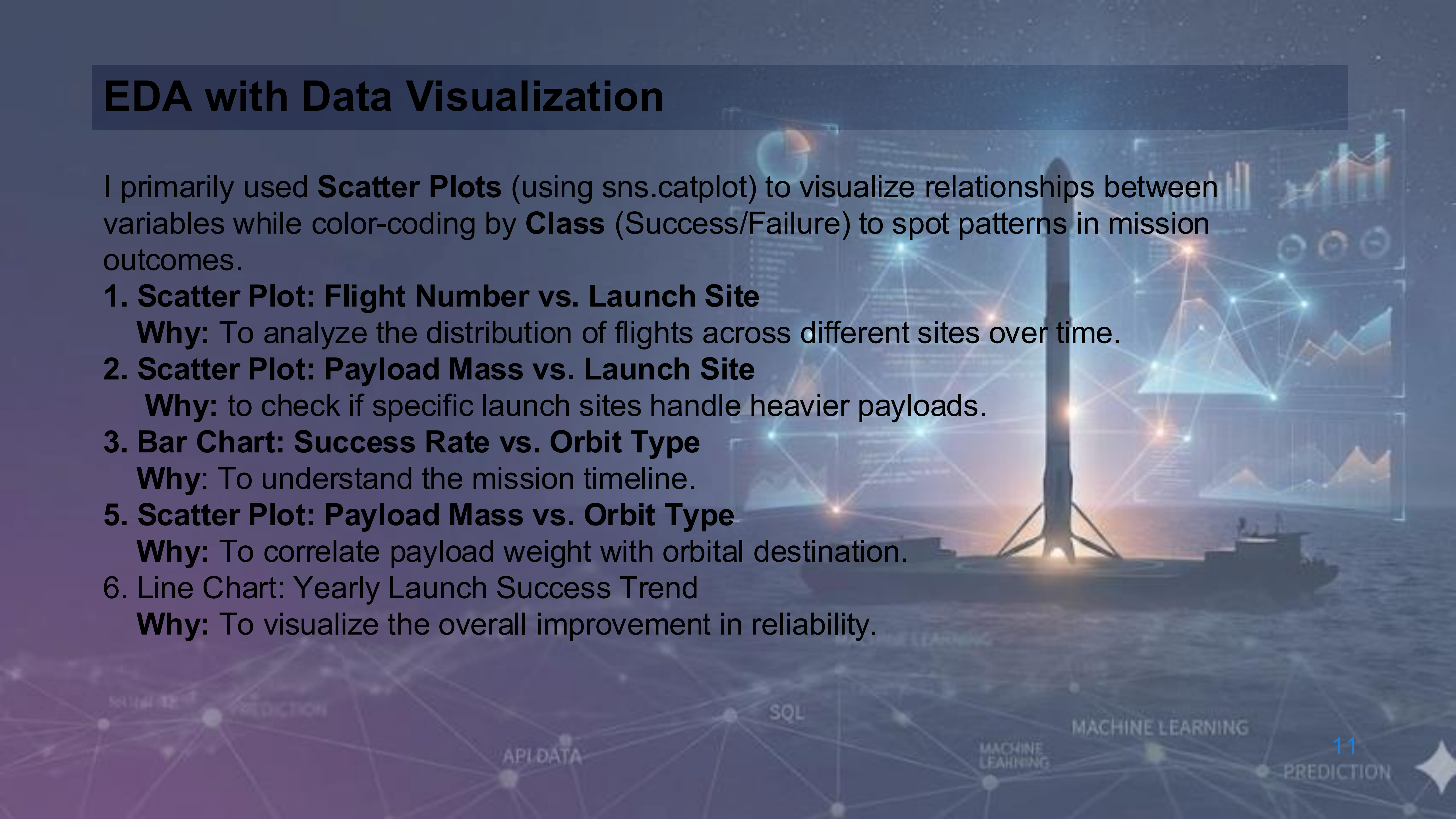
3. The Workflow



Raw Data → Filter Falcon 9 → Handle Nulls (Mean Imputation) → Binarize Outcome → Clean Dataset

GitHub URL:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/9.4/9.4-Web-Scraper.ipynb#:~:text=9.4.Data%2D-,Wrangling.ipynb

EDA with Data Visualization



I primarily used **Scatter Plots** (using sns.catplot) to visualize relationships between variables while color-coding by **Class** (Success/Failure) to spot patterns in mission outcomes.

1. Scatter Plot: Flight Number vs. Launch Site
Why: To analyze the distribution of flights across different sites over time.

2. Scatter Plot: Payload Mass vs. Launch Site
Why: to check if specific launch sites handle heavier payloads.

3. Bar Chart: Success Rate vs. Orbit Type
Why: To understand the mission timeline.

5. Scatter Plot: Payload Mass vs. Orbit Type
Why: To correlate payload weight with orbital destination.

6. Line Chart: Yearly Launch Success Trend
Why: To visualize the overall improvement in reliability.

EDA with Data Visualization (contd.)

GitHub Location:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/tree/main/9.5

EDA with SQL

Objective

Used SQL to query the SpaceX dataset and extract key performance indicators (KPIs) regarding launch sites, payload masses, and landing outcomes.

- **Distinct Values Analysis:**
  - Executed SELECT DISTINCT to identify the unique launch sites (e.g., CCAFS LC-40, KSC LC-39A) defined in the database.
- **Payload Aggregation:**
  - Used SUM to calculate the total payload mass transported by NASA (CRS) missions.
  - Used AVG to determine the average payload mass carried by specific booster versions (e.g., F9 v1.1).
- **Milestone Identification:**
  - Used MIN(Date) to pinpoint the exact date of the **first successful landing** on a ground pad (2015-12-22).
- **Complex Filtering:**
  - Utilized WHERE clauses with BETWEEN logic to identify successful drone ship landings for payloads within a specific range (4,000kg – 6,000kg).
- **Outcome Ranking:**
  - Applied COUNT and GROUP BY to summarize the total number of successful missions versus failed missions.
  - Ranked the landing outcomes (Success/Failure) between the years 2010 and 2017 to observe trends over time.

EDA with SQL (contd.)

GitHub Location:

URL:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/tree/main/9.5

Build an Interactive Map with Folium

Summary of Map Objects

- **Launch Site Markers:**
  I initialized a map centered on the US and added red circles to represent the four major launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

- **MarkerCluster Object:**
  Instead of individual pins for every single launch, this groups launches at the same location into a single cluster icon that expands when clicked.

- **Launch Outcome Markers:**
  I added individual markers for each launch record. I customized the icon colors:
  Green Icon: Represented a successful landing (Class=1).
  Red Icon: Represented a failed landing (Class=0).

- **MousePosition Widget:**
  I added a standard MousePosition plugin to the map, which displays the precise Latitude and Longitude coordinates in the top-right corner as you hover over the map.

- **Proximity Markers (Points of Interest):**
  I manually identified and placed markers for key infrastructure near the launch sites.
  Coastline: The closest point to the ocean.
  City: The closest city center (e.g., Titusville or Lompoc).
  Railway: The nearest railroad track.
  Highway: The nearest major highway.

- **PolyLines:**
  I drew lines connecting the **Launch Site** to each of the **Proximity Markers** (Coast, City, Railway, Highway) to visualize the distance. ***Note:*** gave each proximity marker line a different color for better visualization

Build an Interactive Map with Folium (contd.)

GitHub Location:

URL:
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/10.1/10.1-visualization.ipynb

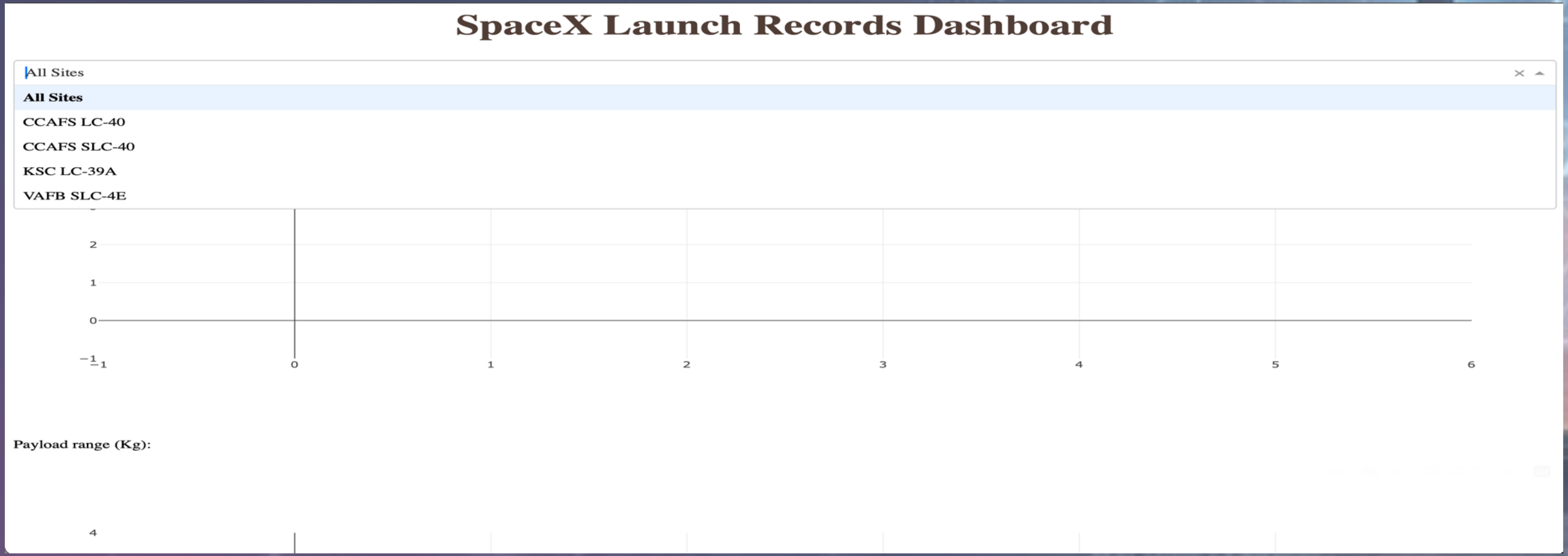
Build a Dashboard with Plotly Dash

Summary of Plots & Interactions

- **Input**: Launch Site Dropdown

  What it is: A menu allowing the user to select "All Sites" or a specific site (e.g., CCAFS LC-40, KSC LC-39A).

  Interaction: Changing this selection instantly updates both the Pie Chart and the Scatter Chart.



SpaceX Launch Records Dashboard

Payload range (Kg):

Why I Added It: Static charts often only show fleet-wide averages, which can hide site-specific issues. This interaction enables **drill-down analysis**, allowing stakeholders to compare a single site's performance against the global average to identify underperforming locations.

Build a Dashboard with Plotly Dash (contd.)

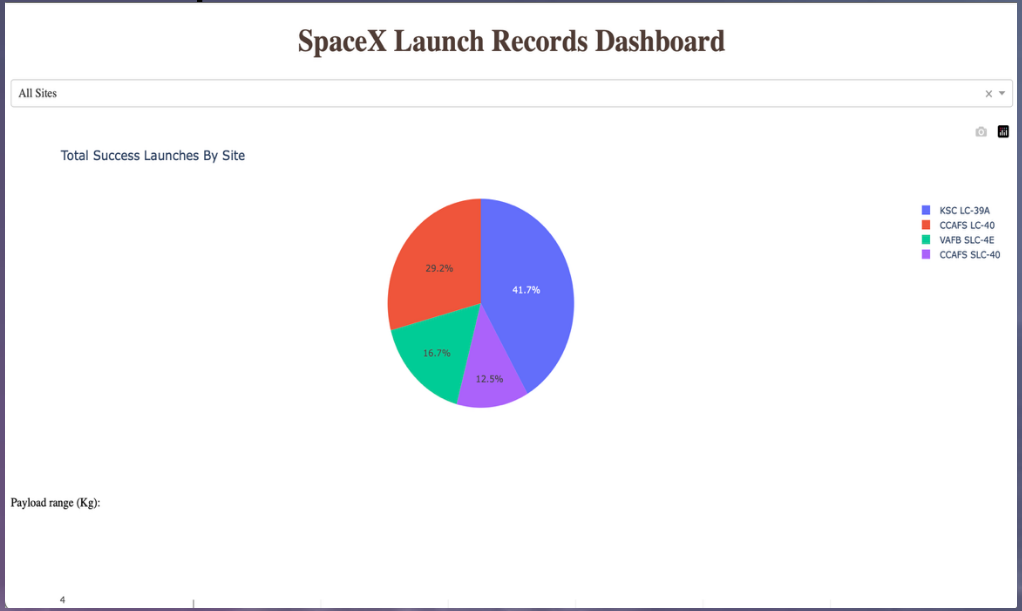
Summary of Plots & Interactions

- **Plot:** Success Pie Chart

  What it shows:

  - If "All Sites" is selected: It displays the total number of successful launches contributed by each launch site. A menu allowing the user to select "All Sites" or a specific site (e.g., CCAFS LC-40, KSC LC-39A).

  - **If a "Specific Site" is selected:** It switches to show the ratio of **Success (1)** vs. **Failure (0)** for that specific site only.

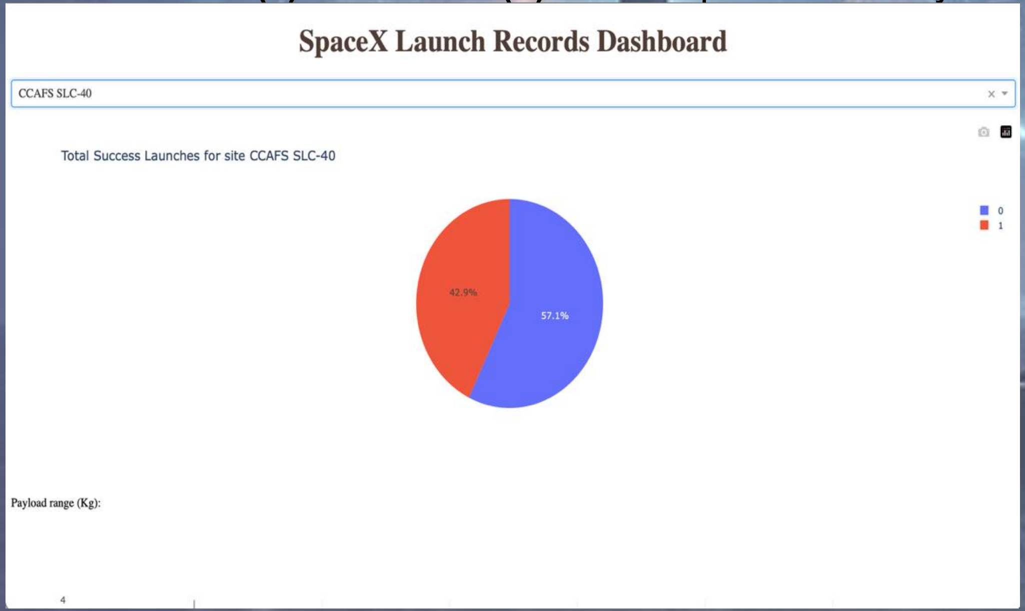


SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

Payload range (Kg):



SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

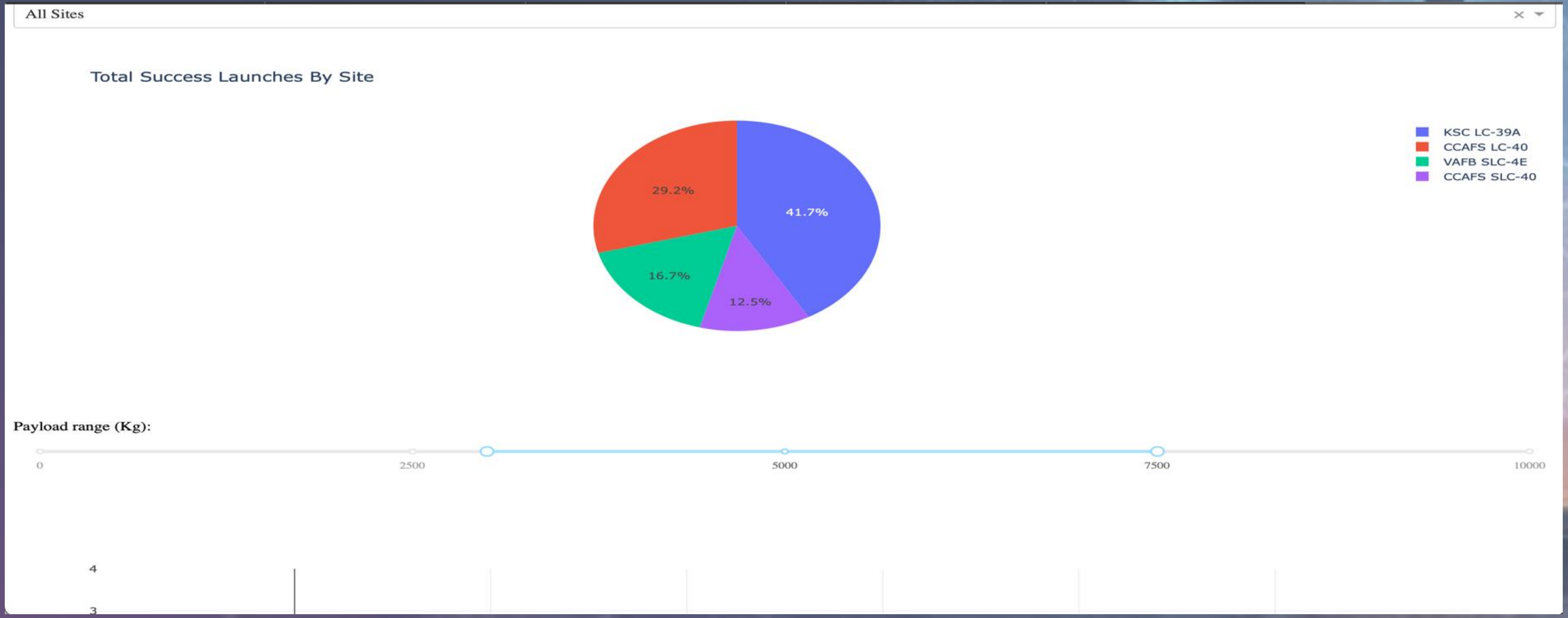
Payload range (Kg):

Why I Added It: To distinguish between **Volume** and **Efficiency**. The "All Sites" view answers *"Which site is busiest?"*, while the "Specific Site" view answers *"How reliable is this location?"*. This distinction is critical for risk assessment.

Build a Dashboard with Plotly Dash (contd.)

Summary of Plots & Interactions

- **Input:** Payload Range Slider
  What it is: A dual-handle slider allowing users to filter launches by their Payload Mass (from 0 kg to 10,000 kg).



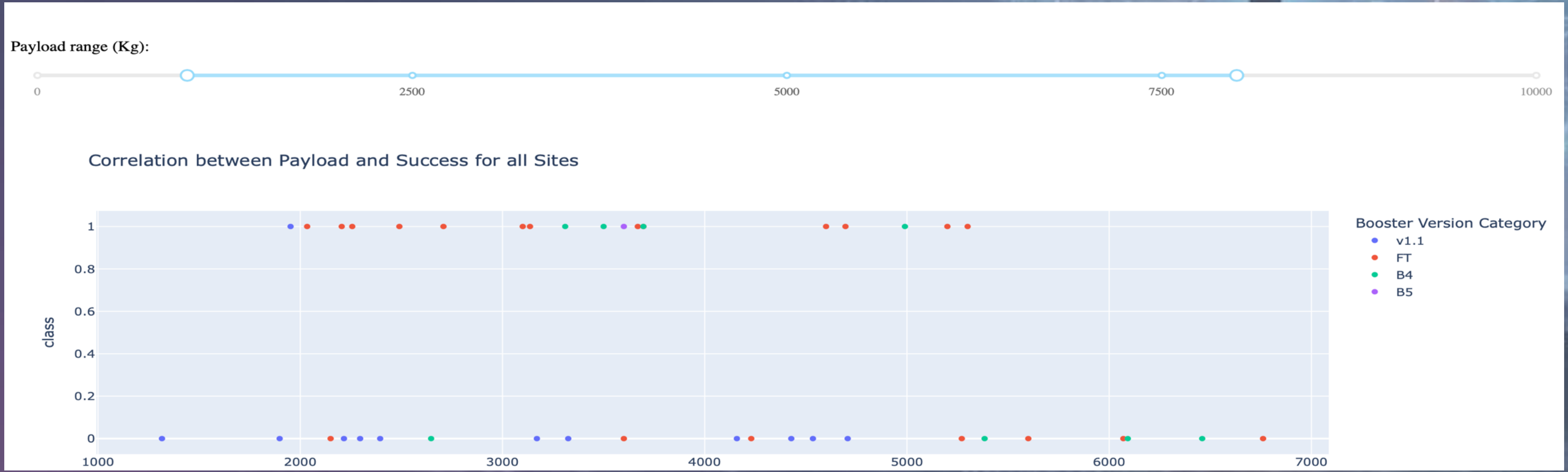
Why I Added It: To test the hypothesis that **heavier payloads are riskier**. By filtering for specific weight classes, users can isolate heavy missions to see if the failure rate increases as the payload approaches the rocket's maximum capacity.

Build a Dashboard with Plotly Dash (contd.)

Summary of Plots & Interactions

- **Plot:** Payload vs. Outcome Scatter Chart

  What it shows: A scatter plot with **Payload Mass (kg)** on the x-axis and **Class (Success/Failure)** on the y-axis. The points are color-coded by **Booster Version** (e.g., v1.1, FT, B5).



Why I Added It: To visualize **technological evolution**. Color-coding by Booster Version (e.g., v1.1 vs. FT) allows us to see if newer hardware has solved the problems that older boosters faced with specific payload weights.

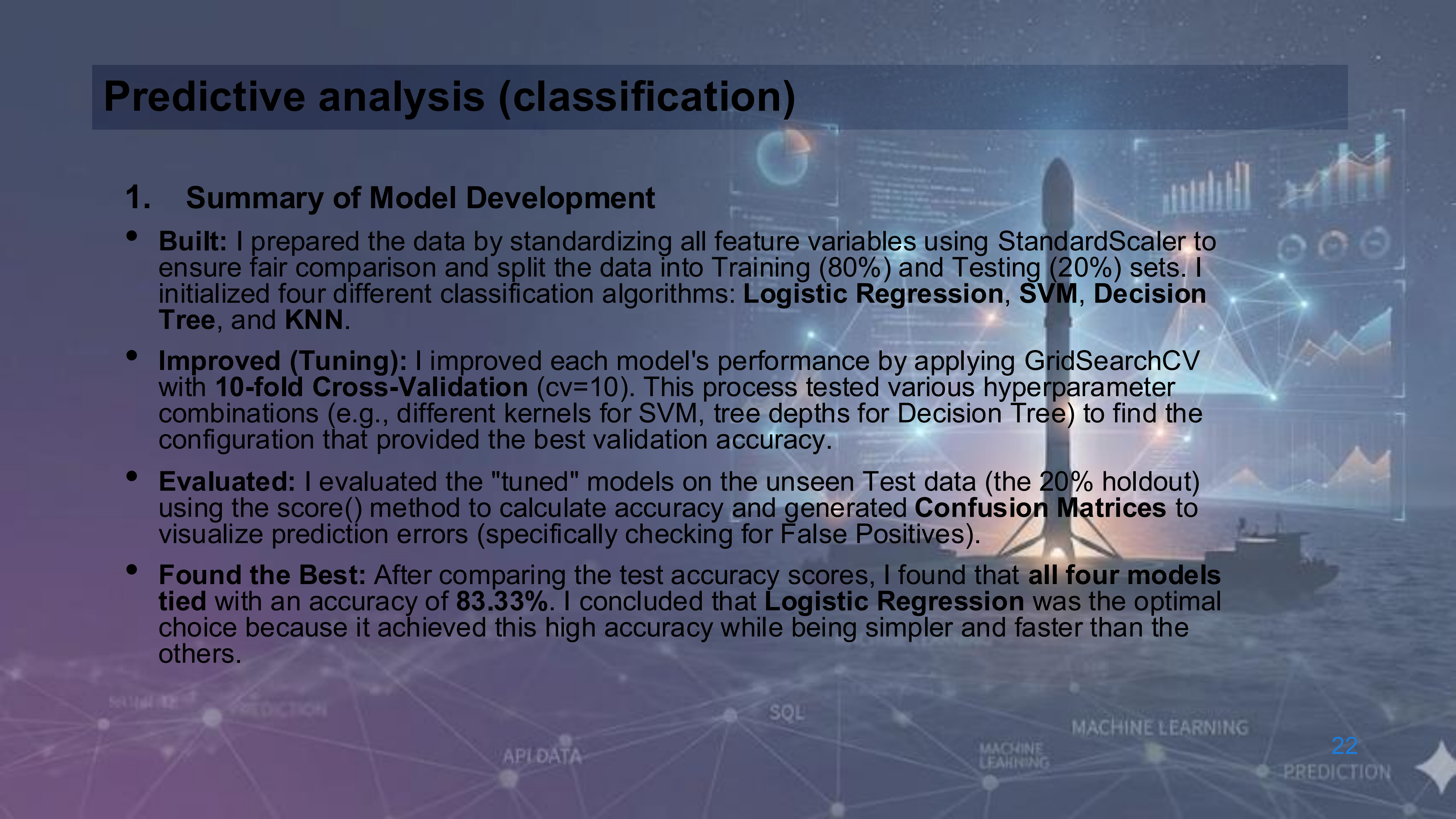
Build a Dashboard with Plotly Dash (contd.)

GitHub URL
https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/tree/main/10.1#:~:text=10.1

Predictive analysis (classification)

1. Summary of Model Development

- **Built:** I prepared the data by standardizing all feature variables using StandardScaler to ensure fair comparison and split the data into Training (80%) and Testing (20%) sets. I initialized four different classification algorithms: **Logistic Regression**, **SVM**, **Decision Tree**, and **KNN**.
- **Improved (Tuning):** I improved each model's performance by applying GridSearchCV with **10-fold Cross-Validation** (cv=10). This process tested various hyperparameter combinations (e.g., different kernels for SVM, tree depths for Decision Tree) to find the configuration that provided the best validation accuracy.
- **Evaluated:** I evaluated the "tuned" models on the unseen Test data (the 20% holdout) using the score() method to calculate accuracy and generated **Confusion Matrices** to visualize prediction errors (specifically checking for False Positives).
- **Found the Best:** After comparing the test accuracy scores, I found that **all four models tied** with an accuracy of **83.33%**. I concluded that **Logistic Regression** was the optimal choice because it achieved this high accuracy while being simpler and faster than the others.



Predictive analysis (classification) (contd.)

2. Development Process Flow

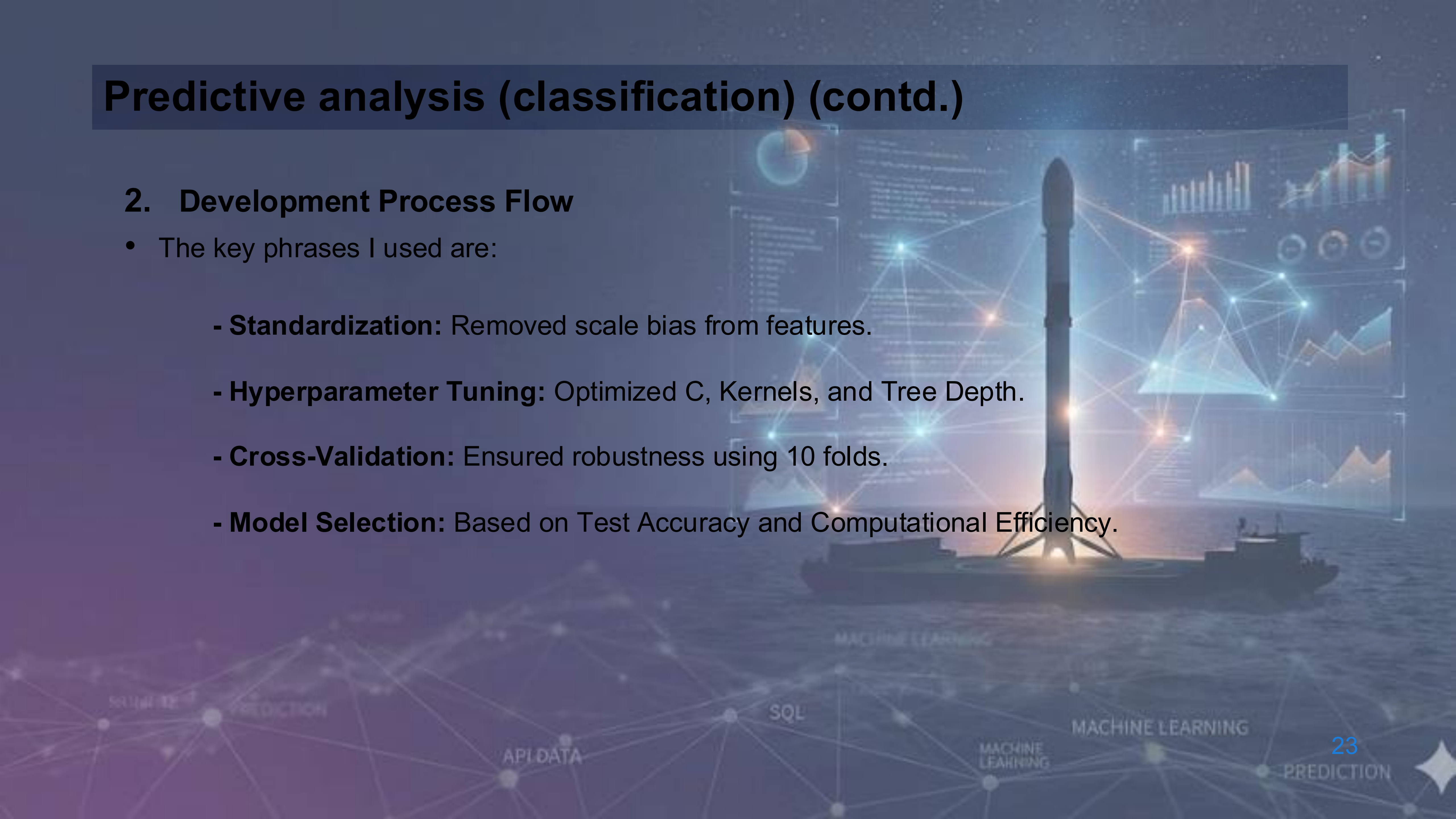
- The key phrases I used are:

  - **Standardization:** Removed scale bias from features.

  - **Hyperparameter Tuning:** Optimized C, Kernels, and Tree Depth.

  - **Cross-Validation:** Ensured robustness using 10 folds.

  - **Model Selection:** Based on Test Accuracy and Computational Efficiency.



Predictive analysis (classification) (contd.)

2. Development Process Flow (contd.)

- **The Process Flowchart is below**

  A [Standardize Data(StandardScaler)] **--> B**[Split Data(Train 80% / Test 20%)]

  B --> C{Hyperparameter Tuning(GridSearchCV cv=10)}

  C -->|Logistic Regression| **D**[Best Parameters Found]

  C -->|SVM| **D**

  C -->|Decision Tree| **D**

  C -->|KNN| **D**

  D --> E[Evaluate on Test Set(Accuracy & Confusion Matrix)]

  E --> F[Compare & Select Best Model(All Tied at 83.33%)]

Predictive analysis (classification) (contd.)

2. Development Process Flow (contd.)

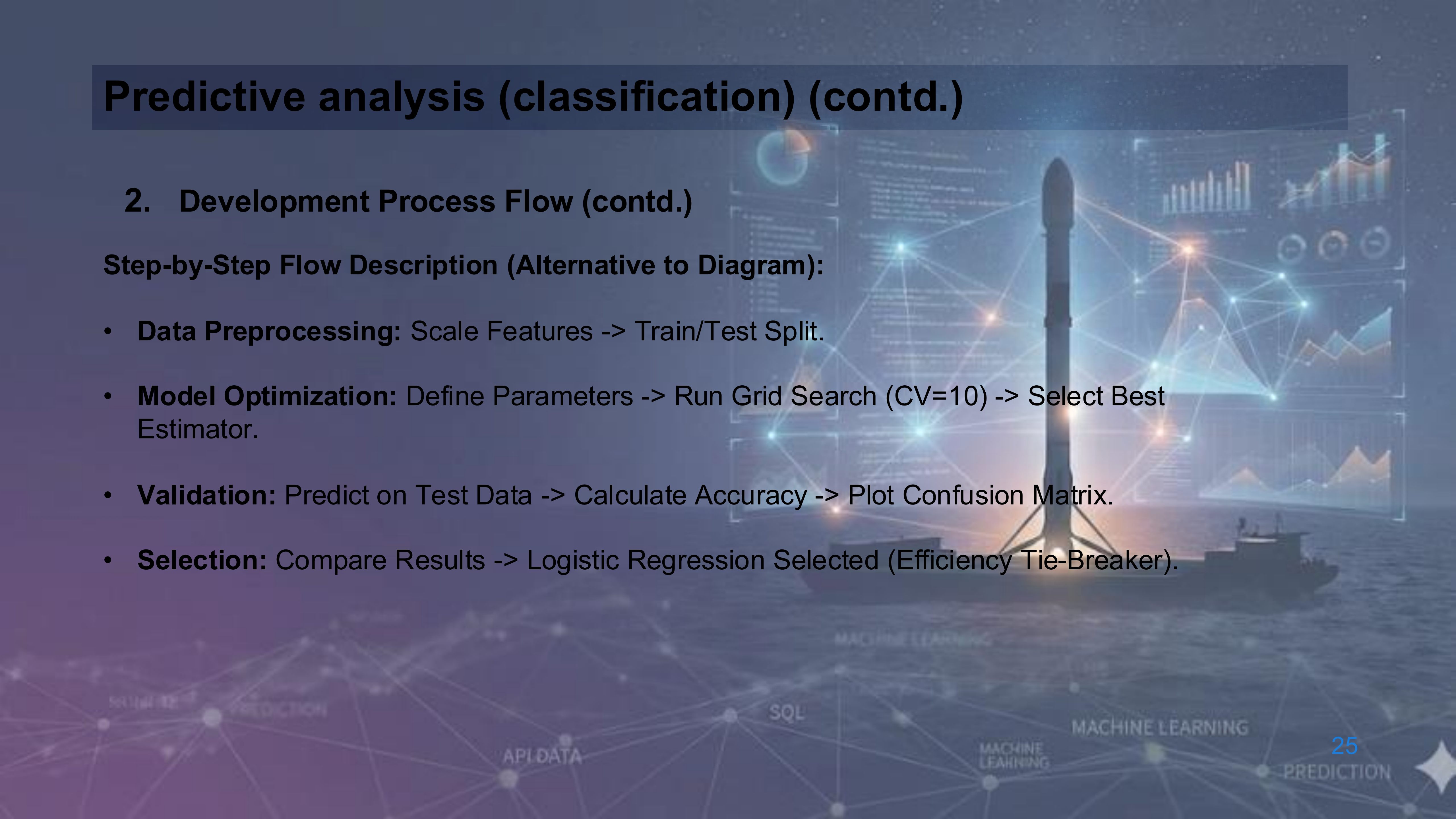
Step-by-Step Flow Description (Alternative to Diagram):

- **Data Preprocessing:** Scale Features -> Train/Test Split.

- **Model Optimization:** Define Parameters -> Run Grid Search (CV=10) -> Select Best Estimator.

- **Validation:** Predict on Test Data -> Calculate Accuracy -> Plot Confusion Matrix.

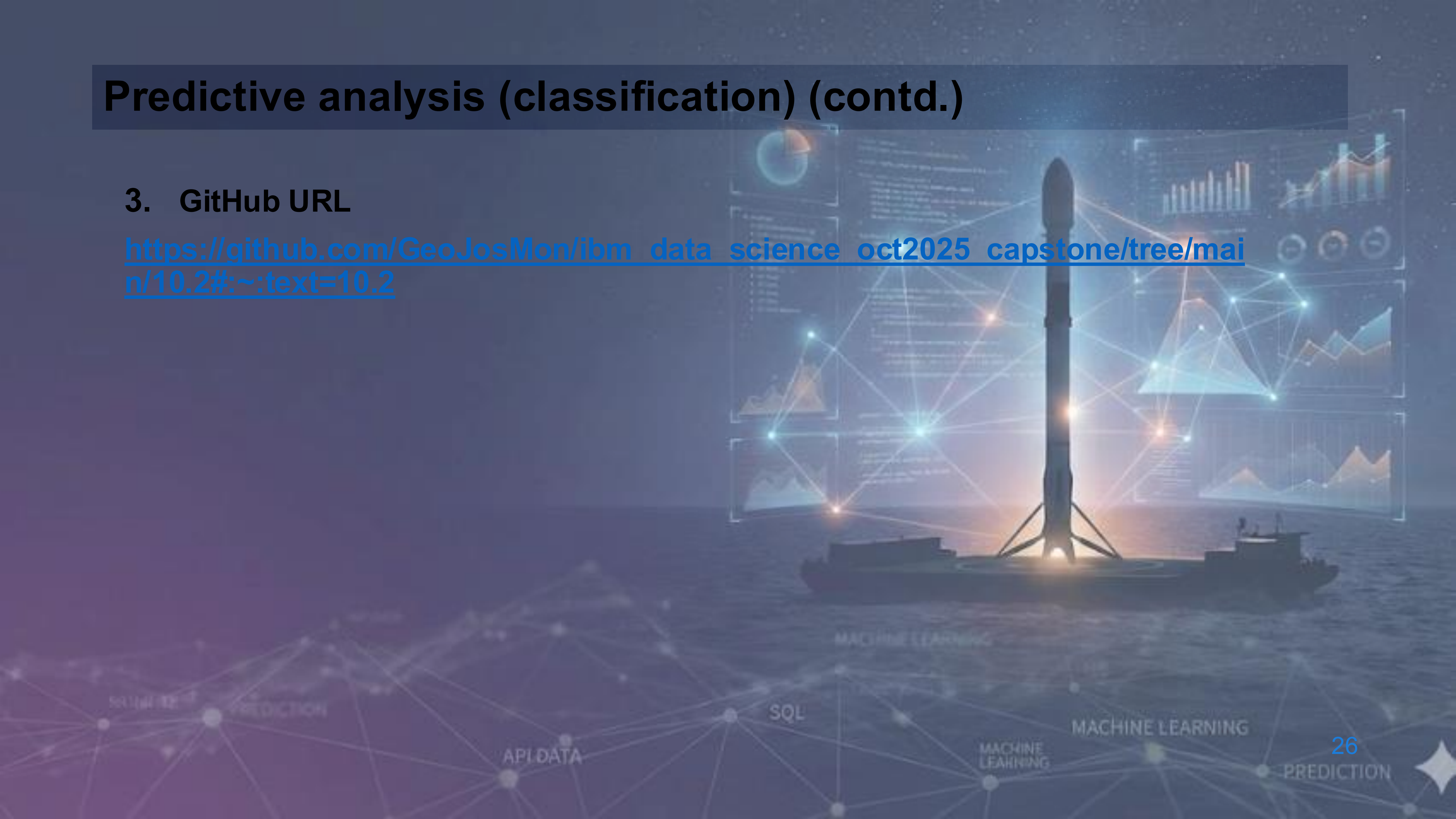
- **Selection:** Compare Results -> Logistic Regression Selected (Efficiency Tie-Breaker).

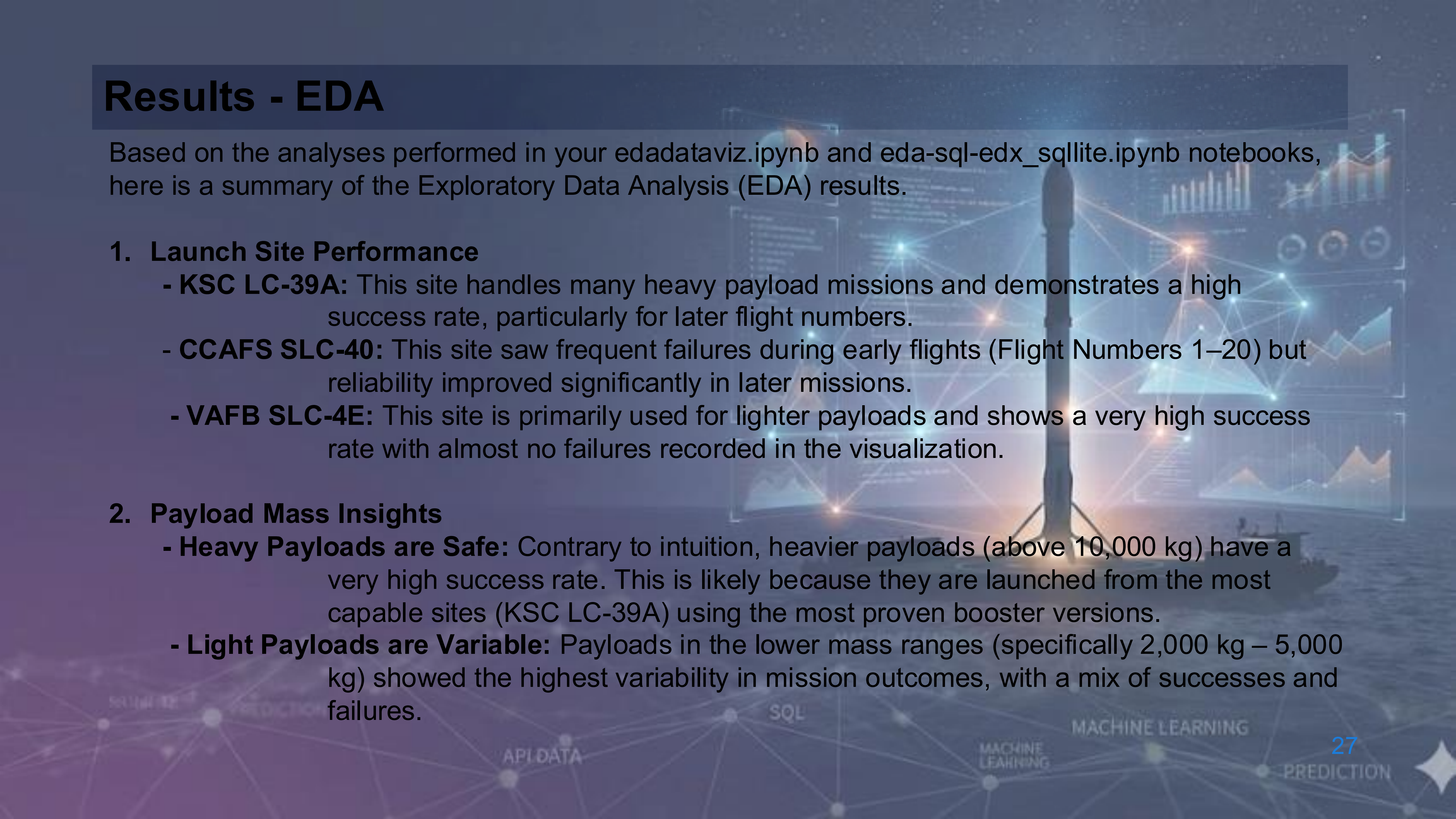


Predictive analysis (classification) (contd.)

3. GitHub URL

https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/tree/main/10.2#:~:text=10.2





Results - EDA

Based on the analyses performed in your edadataviz.ipynb and eda-sql-edx_sqllite.ipynb notebooks, here is a summary of the Exploratory Data Analysis (EDA) results.

1. **Launch Site Performance**
   - **KSC LC-39A:** This site handles many heavy payload missions and demonstrates a high success rate, particularly for later flight numbers.
   - **CCAFS SLC-40:** This site saw frequent failures during early flights (Flight Numbers 1–20) but reliability improved significantly in later missions.
   - **VAFB SLC-4E:** This site is primarily used for lighter payloads and shows a very high success rate with almost no failures recorded in the visualization.

2. **Payload Mass Insights**
   - **Heavy Payloads are Safe:** Contrary to intuition, heavier payloads (above 10,000 kg) have a very high success rate. This is likely because they are launched from the most capable sites (KSC LC-39A) using the most proven booster versions.
   - **Light Payloads are Variable:** Payloads in the lower mass ranges (specifically 2,000 kg – 5,000 kg) showed the highest variability in mission outcomes, with a mix of successes and failures.

Results – EDA (contd.)

3. **Orbit and Mission Trends**

   Top Performers: Orbits such as **ES-L1, GEO, HEO, and SSO** achieved a **100% success rate**.

   Low Performer: The **SO** (Sun-Synchronous Orbit) was the only orbit type with a 0% success rate (based on the limited data available).

   Mission Distribution:

   - LEO & ISS: These orbits are used consistently throughout the program's history and typically carry heavier payloads.

   - GTO: Geostationary Transfer Orbit missions became more frequent as the flight numbers increased, indicating SpaceX's shift toward commercial satellite launches.

4. **Temporal Trends (Time)**

   Reliability Growth: The success rate has steadily increased since 2013, stabilizing at a high level in recent years.

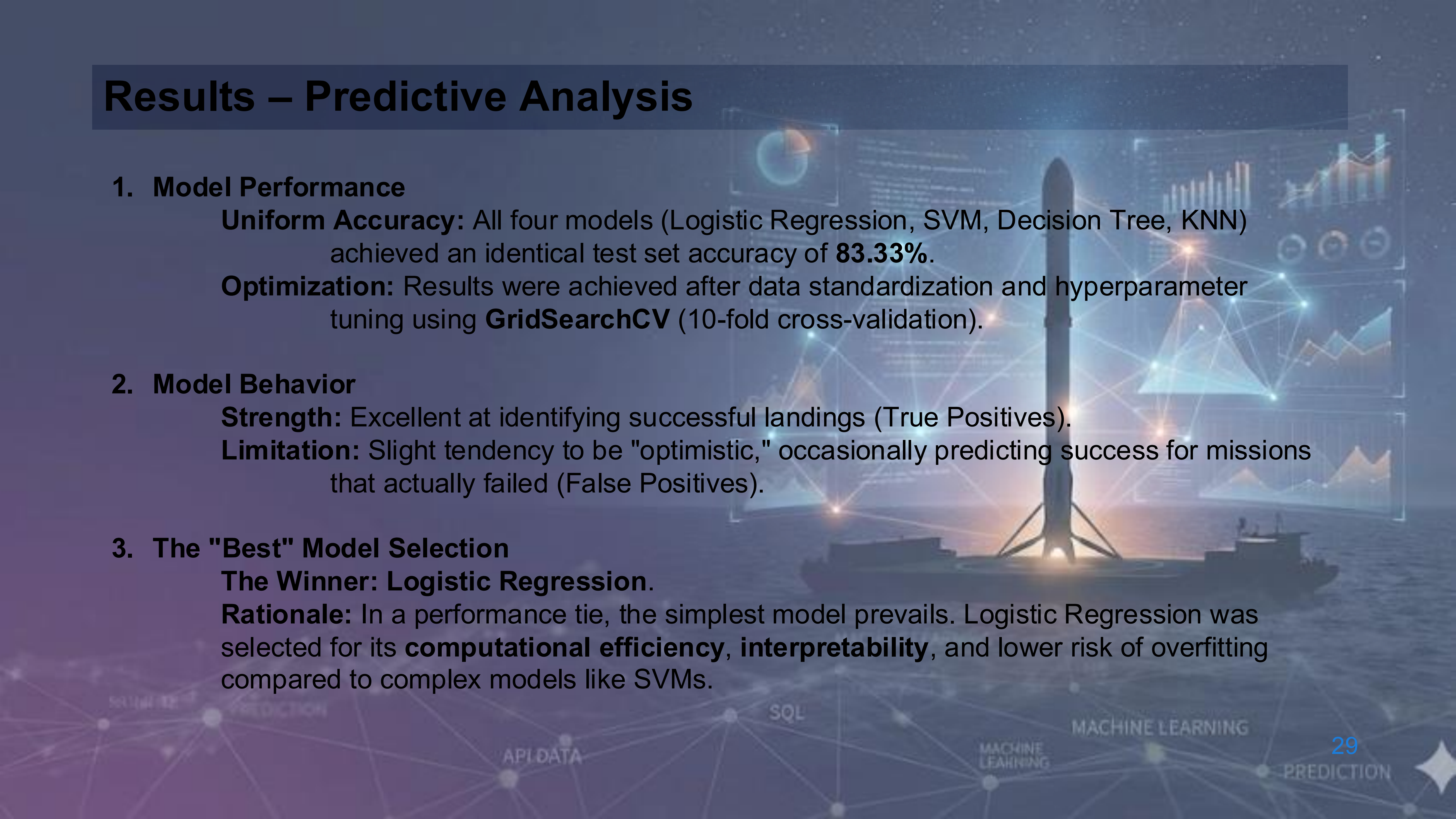
   First Success: SQL analysis identified that the first successful landing on a ground pad occurred on **2015-12-22**.

Results – Predictive Analysis

1. **Model Performance**
   - **Uniform Accuracy:** All four models (Logistic Regression, SVM, Decision Tree, KNN) achieved an identical test set accuracy of **83.33%**.
   - **Optimization:** Results were achieved after data standardization and hyperparameter tuning using **GridSearchCV** (10-fold cross-validation).

2. **Model Behavior**
   - **Strength:** Excellent at identifying successful landings (True Positives).
   - **Limitation:** Slight tendency to be "optimistic," occasionally predicting success for missions that actually failed (False Positives).

3. **The "Best" Model Selection**
   - **The Winner: Logistic Regression**.
   - **Rationale:** In a performance tie, the simplest model prevails. Logistic Regression was selected for its **computational efficiency**, **interpretability**, and lower risk of overfitting compared to complex models like SVMs.



Results –Predictive Analysis (contd.)

3. **Error Analysis (Confusion Matrix)**

   Consistent Behavior: All models displayed a similar error pattern.

   Strength: High recall on **True Positives**—excellent at correctly identifying successful landings.

   Weakness: The primary error source is **False Positives** (predicting success when the rocket actually crashed).

   Insight: The models are slightly "optimistic," occasionally underestimating specific failure conditions.

4. **The Critical Role of Preprocessing**

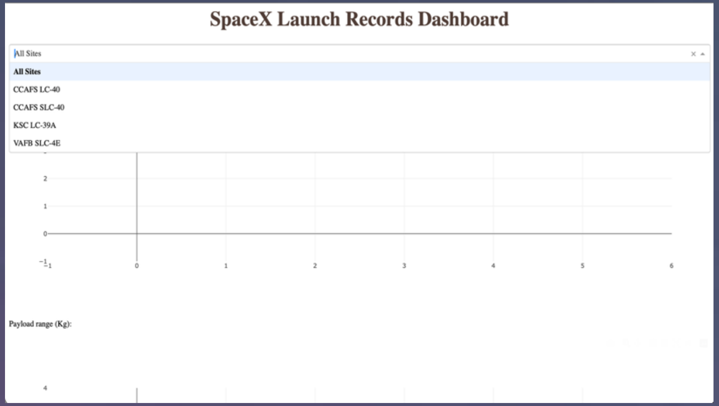
   Key Factor: High accuracy was driven by robust data wrangling, specifically **Standardization**.

   Why it Matters: Without scaling, large numerical values (like **Payload Mass** in kg) would overpower binary features (like **Orbit**).

   Impact: This step was essential for the performance of distance-based algorithms like **SVM** and **KNN**.

Results

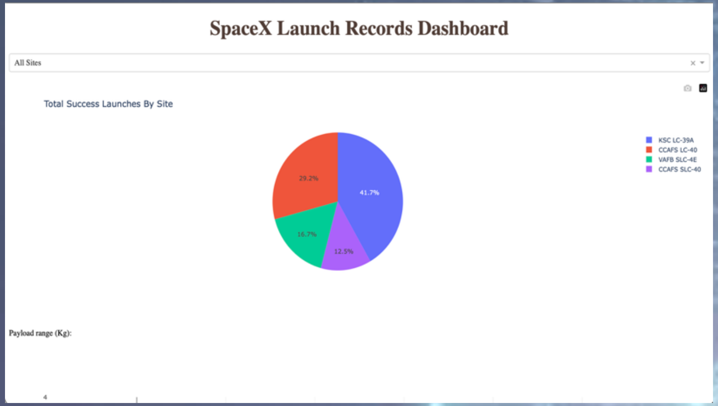
Below are the interactive analytics demo in screenshots



SpaceX Launch Records Dashboard

All Sites

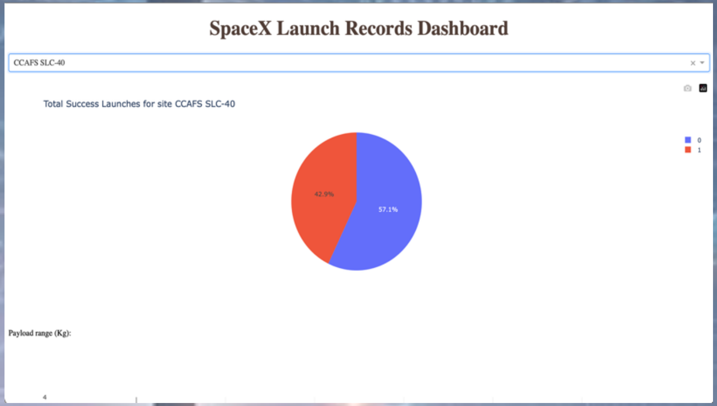
Payload range (Kg):



SpaceX Launch Records Dashboard

All Sites

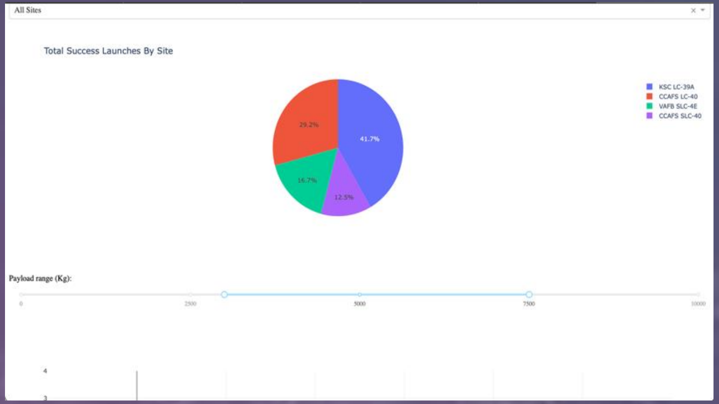
Payload range (Kg):



SpaceX Launch Records Dashboard

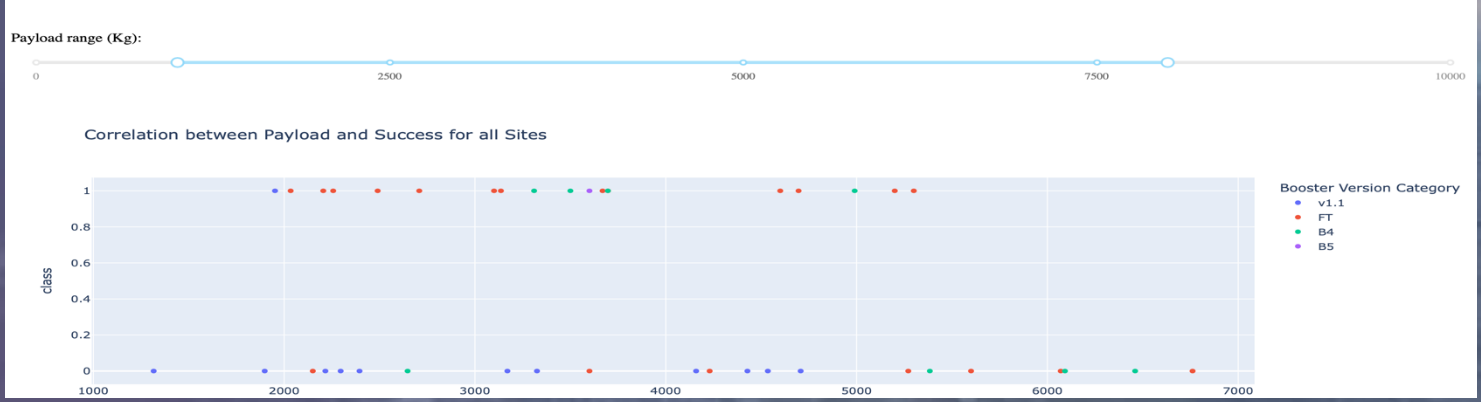
CCAFS SLC-40

Payload range (Kg):



All Sites

Payload range (Kg):

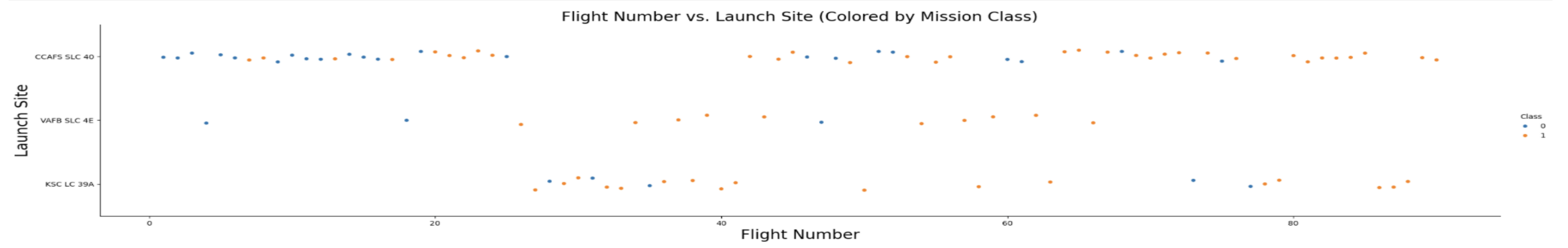


Payload range (Kg):

Section 2

MACHINE LEARNING
PREDICTION
SQL
MACHINE LEARNING
API DATA
MACHINE LEARNING
PREDICTION

Flight Number vs. Launch Site



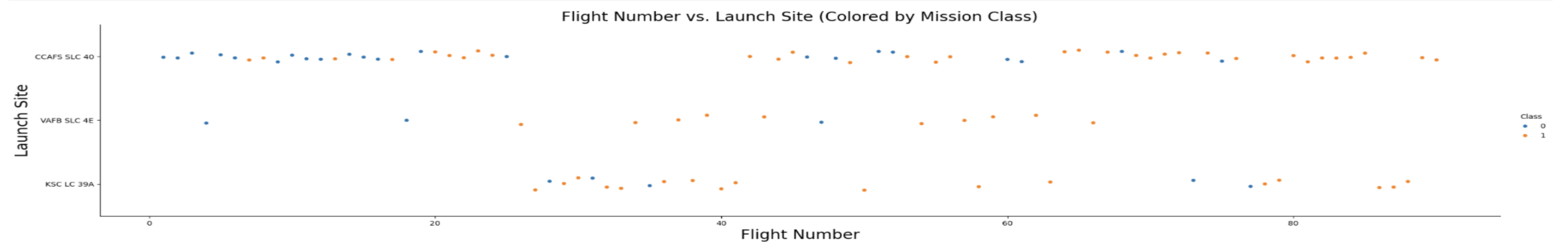
Scatter Plot of Flight Number vs. Launch Site

Plot Overview

Created using sns.catplot to map the chronological sequence of missions across the different launch facilities.

- **X-Axis:** Flight Number (representing the timeline/order of launches).
- **Y-Axis:** Launch Site (CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E).
- **Color (Hue):** Class (0 = Failure, 1 = Success).

Flight Number vs. Launch Site



Scatter Plot of Flight Number vs. Launch Site

Key Insights & Observations

1. Site Analysis: CCAFS SLC-40

Visual Trend: The data reveals a distinct cluster of failures (red dots) specific to this launch site.

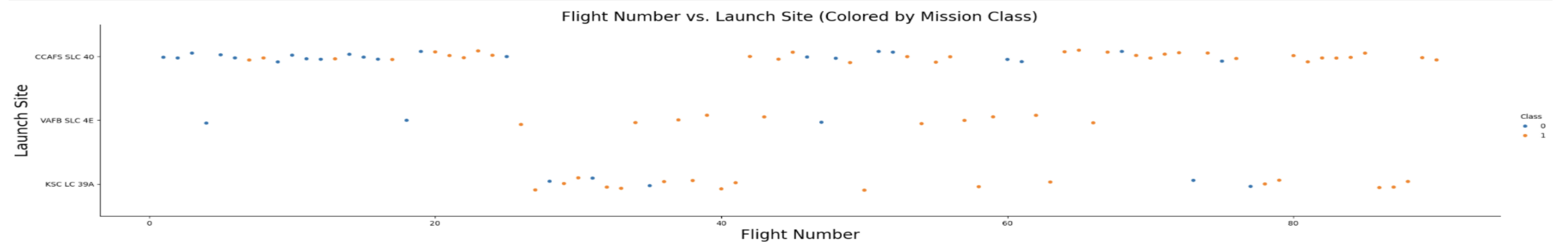
Critical Period: Failures are heavily concentrated between **Flight Numbers 1 and 20**.

2. Strategic Interpretation

Developmental Phase: This cluster represents the initial experimental era of the Falcon 9 program.

Technology Refinement: High failure rates in early flights correlate with the rapid iteration and testing of landing technologies.

Flight Number vs. Launch Site



Scatter Plot of Flight Number vs. Launch Site

Key Insights & Observations (contd.)

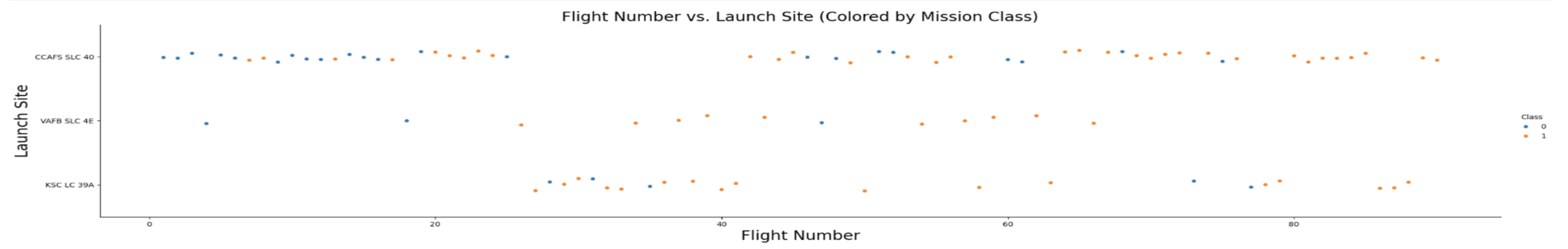
3. **Infrastructure Expansion**

   Late Adopters: Sites **VAFB SLC-4E** and **KSC LC-39A** show no activity during early flight numbers.
   Strategic Growth: These pads were activated only after the Falcon 9 technology was proven, marking a shift from experimentation to commercial scaling.

4. **Reliability Growth**

   The Trend: As Flight Numbers increase (moving right on the X-axis), success indicators become dominant across all sites.
   Maturity: This confirms a direct correlation between program longevity and technical reliability, regardless of the launch pad used.

Flight Number vs. Launch Site



Scatter Plot of Flight Number vs. Launch Site

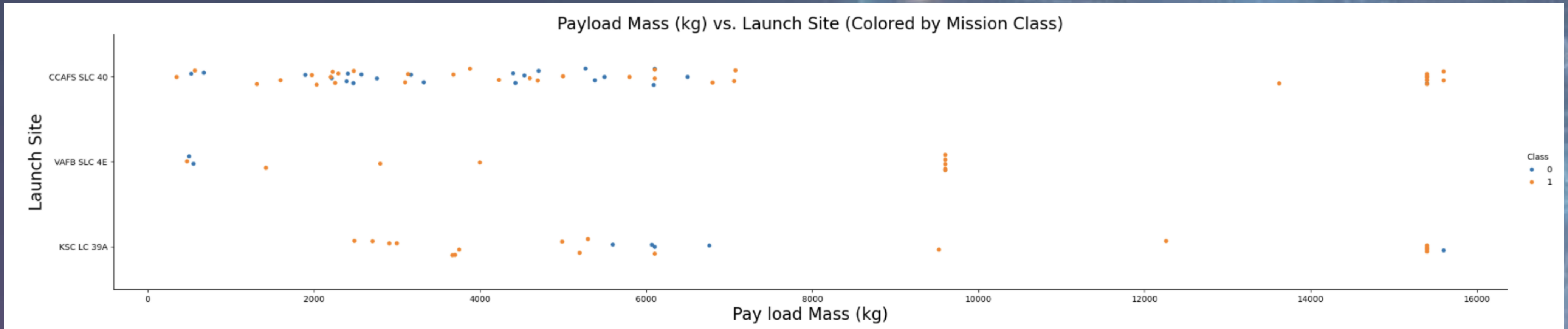
Key Insights & Observations (contd.)

- **The Struggle:** The data visually confirms a high failure rate during the initial phase at **CCAFS**, marking the experimental beginning of the program.

- **The Lesson:** SpaceX leveraged data from these early flights to refine their technology.

- **The Result:** The trend shifts dramatically to high reliability across **all launch sites** in later missions, proving the effectiveness of their iterative engineering approach.

Payload vs. Launch Site



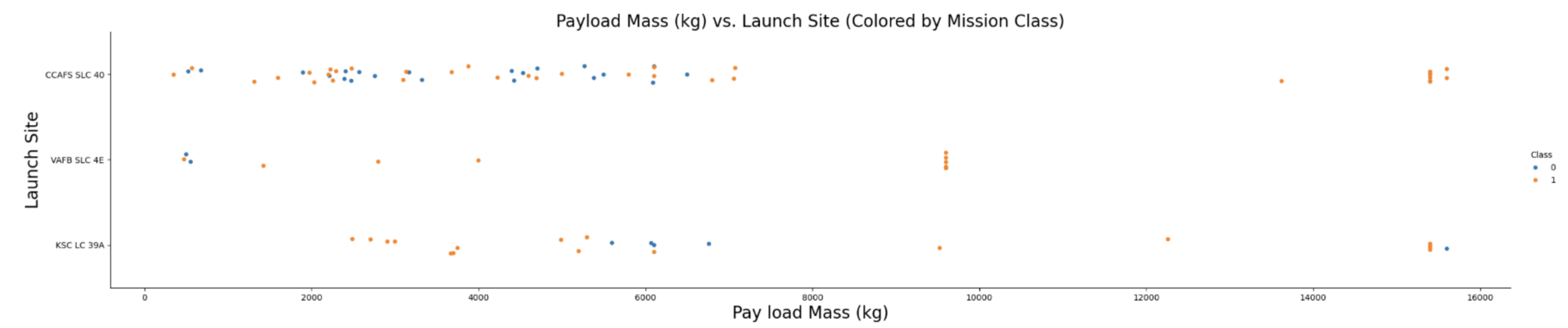
Scatter Plot of Payload vs. Launch Site

Plot Overview

Created using sns.catplot to visualize how different launch sites handle varying payload weights and how that relates to mission success.

- **X-Axis:** Payload Mass (kg)
- **Y-Axis:** Launch Site (CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
- **Color (Hue):** Class (0 = Failure, 1 = Success)

Payload vs. Launch Site (contd.)



Scatter Plot of Payload vs. Launch Site

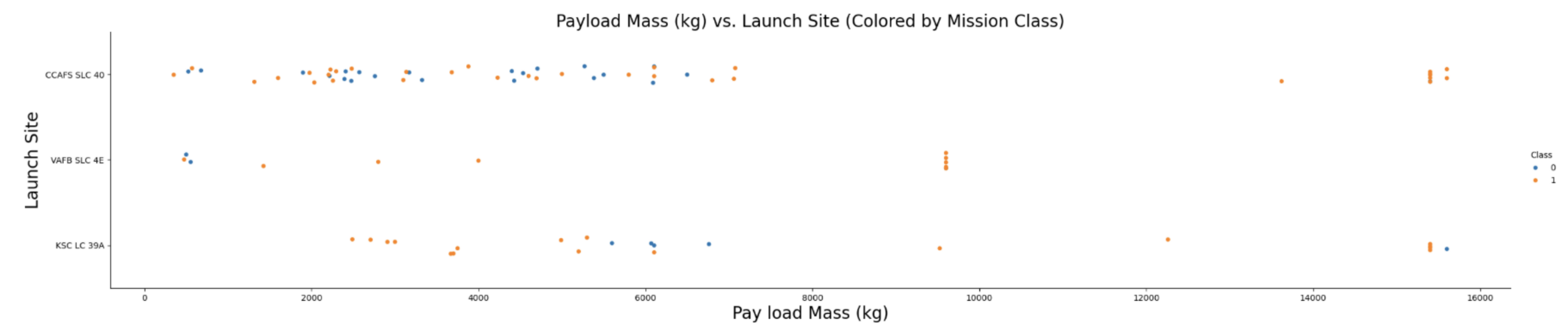
Key Insights & Observations

1. **Site Specialization (Heavy vs. Light):**

   - VAFB SLC-4E: no dots on the far right of the chart for this site, indicating it is **not used for heavy payloads** (typically capping out under 10,000 kg).

   - KSC LC-39A & CCAFS SLC-40: data points extending much further to the right (up to ~15,600 kg), confirming they are the primary facilities for **heavy-lift missions**.

Payload vs. Launch Site (contd.)



Scatter Plot of Payload vs. Launch Site

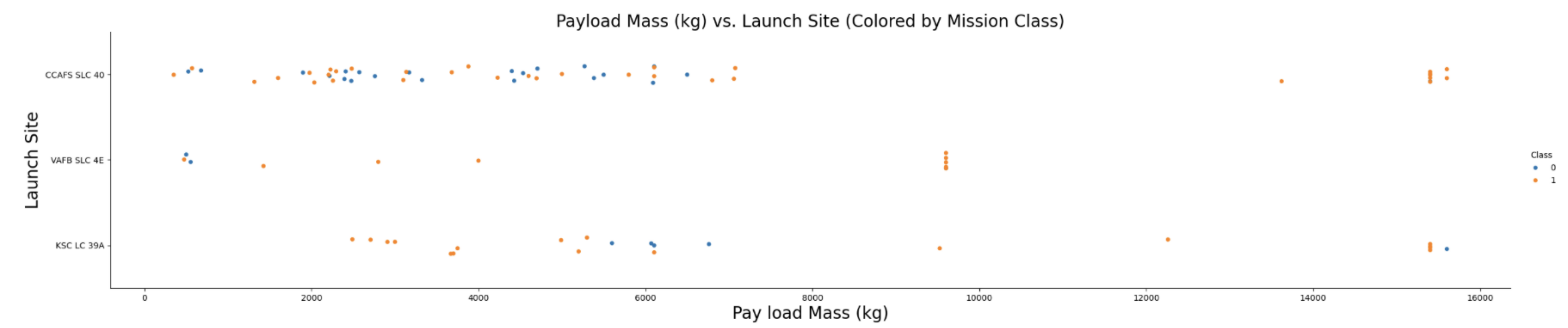
Key Insights & Observations (contd.)

2. **Success Correlation:**

   Heavy Payloads are Safe: For payloads exceeding 10,000 kg (mostly at KSC LC-39A and CCAFS SLC-40), the markers are predominantly **Class 1 (Success)**. This suggests that heavy payloads are launched using the most proven and reliable configurations.

   Mid-Range Variability: majority of failures (Class 0) occur in the mid-range payload category (approx. 2,000 kg to 6,000 kg) rather than the heaviest category.

Payload vs. Launch Site (contd.)



Scatter Plot of Payload vs. Launch Site

Key Insights & Observations (contd.)

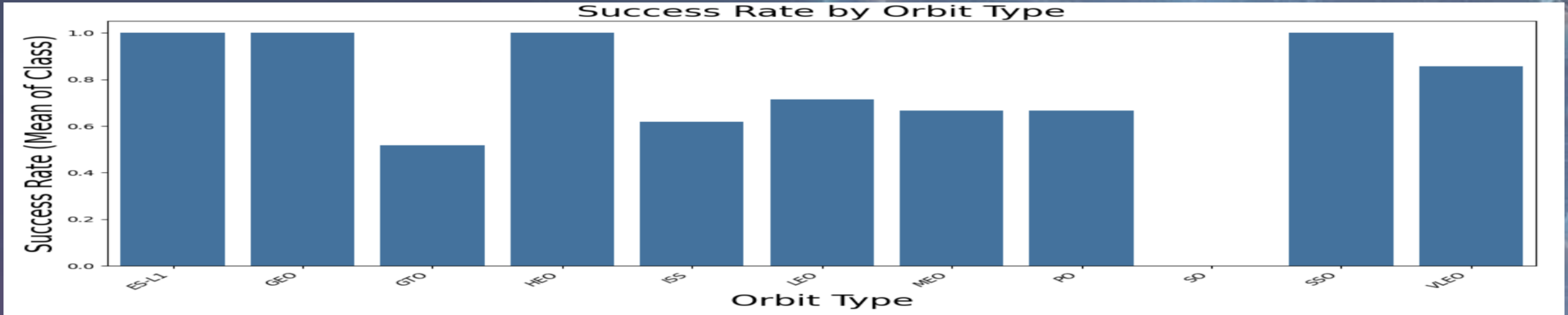
- **Heavier ≠ Riskier**

  The Assumption: Common logic suggests that heavier payloads increase the risk of failure.
  The Reality: The data disproves this, showing that mission success is not negatively correlated with payload mass.

- **Strategic Adaptation**

  Site Optimization: SpaceX has successfully adapted specific launch sites to handle **maximum load missions** reliably.
  Outcome: The program maintains high safety standards even when pushing the rocket's lift capacity to the limit.

Success Rate vs. Orbit Type



Bar Chart of Success rate vs. Orbit type

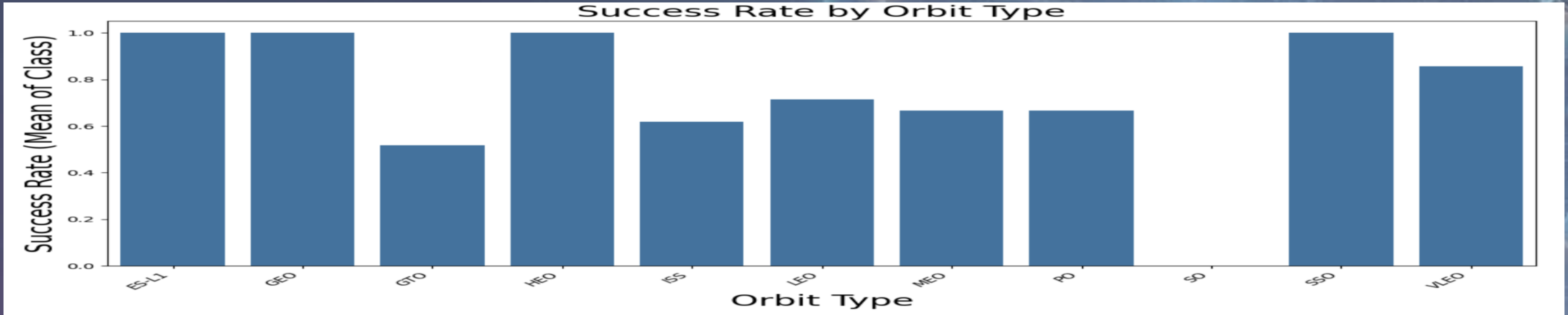
Plot Overview

Created by grouping the data by the Orbit column and calculating the mean of the Class column (where 1 is success and 0 is failure).

X-Axis: Orbit Type (e.g., LEO, GTO, ISS, SO).

Y-Axis: Success Rate (ranging from 0.0 to 1.0, representing 0% to 100%).

Success Rate vs. Orbit Type (contd.)



Bar Chart of Success rate vs. Orbit type

Key Insights & Observations

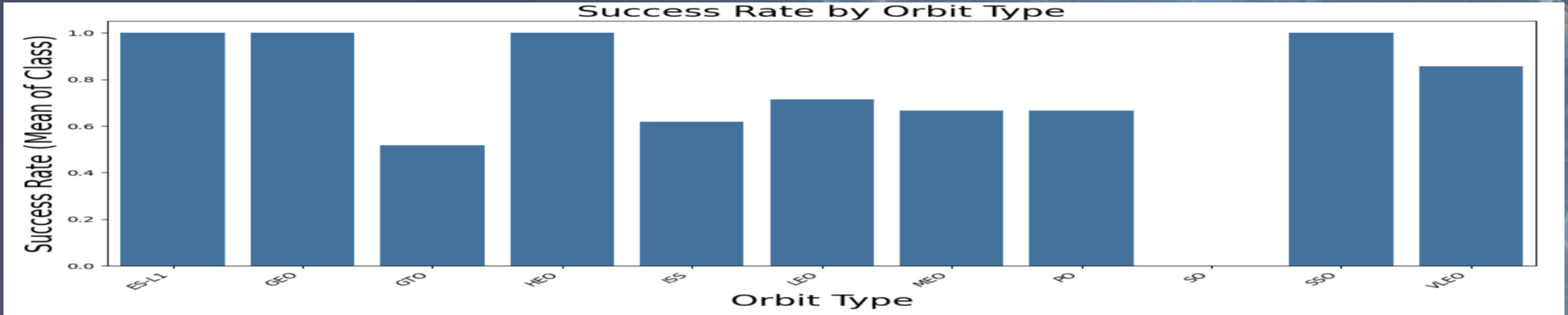
1. **Perfect Reliability (100%):**

   The chart reveals that four specific orbits achieved a **100% success rate**: **ES-L1** (Lagrange Point), **GEO** (Geostationary), **HEO** (High Earth Orbit), and **SSO** (Sun-Synchronous Orbit).

2. **The "SO" Outlier:**

   The **SO** (Sun-Synchronous Orbit) stands out with a **0% success rate**. Although this could most likely due to a very small sample size rather than an inherent inability to reach that orbit.

Success Rate vs. Orbit Type (contd.)



Bar Chart of Success rate vs. Orbit type

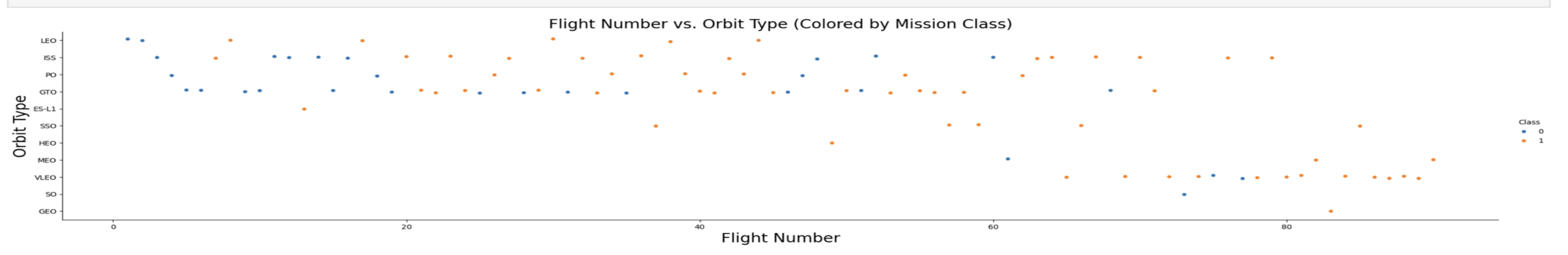
Key Insights & Observations (contd.)

3. **High-Volume Orbits:**

   GTO (Geostationary Transfer Orbit), which is the most common orbit for commercial satellites, typically shows a success rate around **50-60%** in this specific cut of the data. This lower rate reflects the difficulty of these missions and the fact that many early developmental flights targeted GTO.

This chart helps categorize risk. It identifies that while "routine" LEO missions are common, SpaceX has demonstrated perfect reliability on some of the most difficult orbital trajectories (like GEO and ES-L1).

Flight Number vs. Orbit Type



Scatter Plot of Flight Number vs. Orbit type

Plot Overview

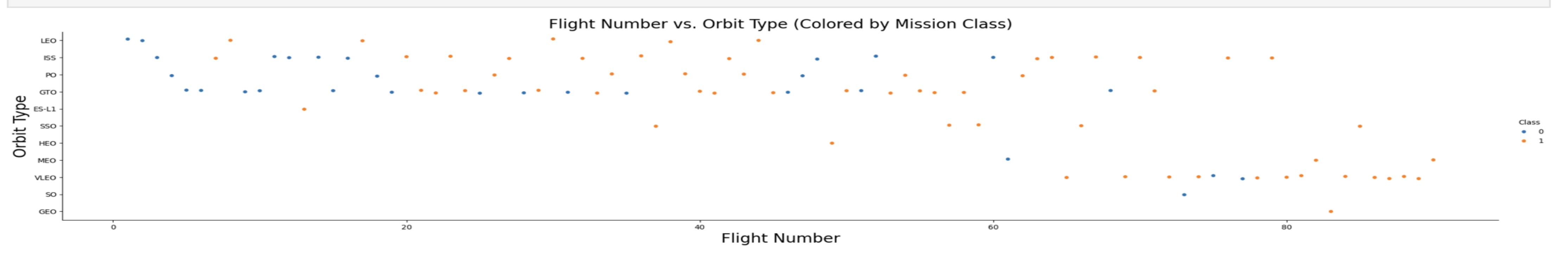
This chart uses sns.catplot to visualize the timeline of missions to different orbital destinations.

X-Axis: Flight Number (Chronological order of launches).

Y-Axis: Orbit Type (e.g., LEO, GTO, ISS, PO).

Color (Hue): Class (0 = Failure, 1 = Success).

Flight Number vs. Orbit Type



Scatter Plot of Flight Number vs. Orbit type

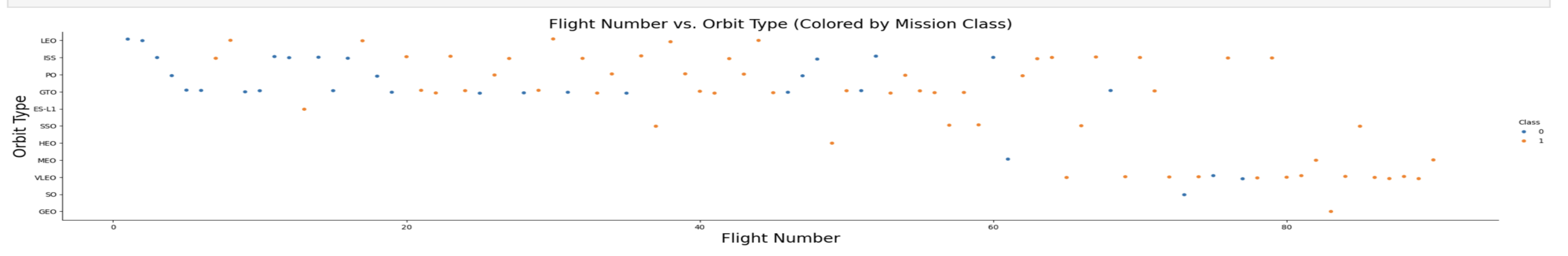
Key Insights & Observations

1. The Core Missions:

- **Consistent Presence:** Low Earth Orbit (LEO) and International Space Station (ISS) missions appear consistently across the entire flight history, from the first launch to the present.

- **Program Foundation:** These frequent resupply and satellite deployment missions serve as the "bread and butter" of the Falcon 9 program.

Flight Number vs. Orbit Type (contd.)



Scatter Plot of Flight Number vs. Orbit type

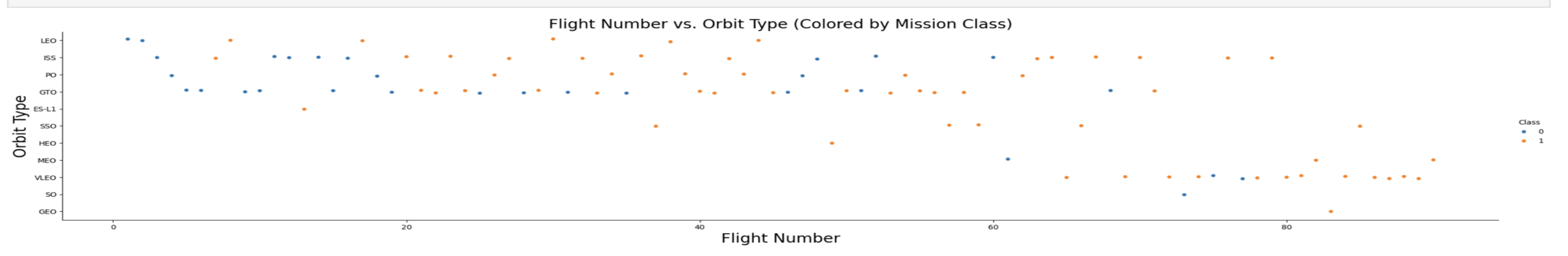
Key Insights & Observations (contd.)

2. Commercial Expansion

- **Market Shift:** As Flight Numbers increase, there is a visible surge in Geostationary Transfer Orbit (GTO) missions.

- **Strategic Evolution:** This trend marks SpaceX's aggressive expansion from government-focused LEO missions to the lucrative private commercial satellite market.

Flight Number vs. Orbit Type (contd.)



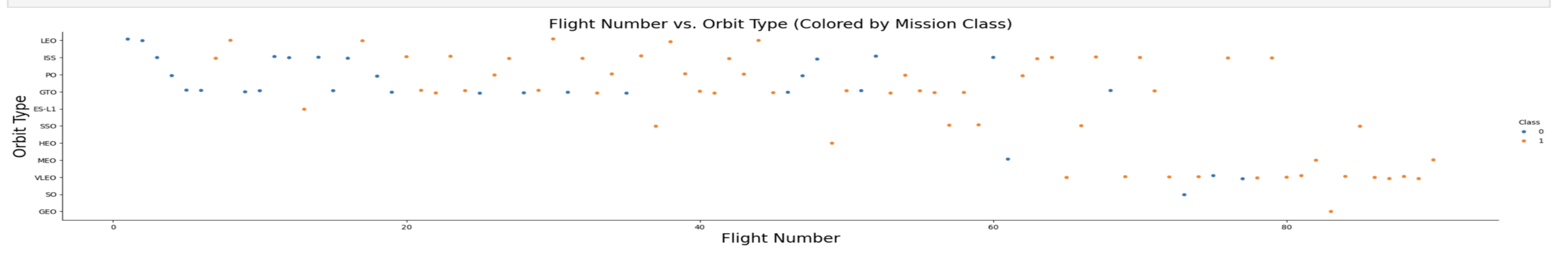
Scatter Plot of Flight Number vs. Orbit type

Key Insights & Observations (contd.)

3. Polar Capability

- **Late Emergence:** Data points for Polar Orbits (PO) appear exclusively at higher flight numbers, visible on the far right of the timeline.
- **Milestone Achievement:** This trend marks a specific capability milestone achieved later in the rocket's operational history.
- **Strategic Expansion:** Correlates with the specific activation of the **VAFB** launch site to reach these specialized polar trajectories.

Flight Number vs. Orbit Type (contd.)



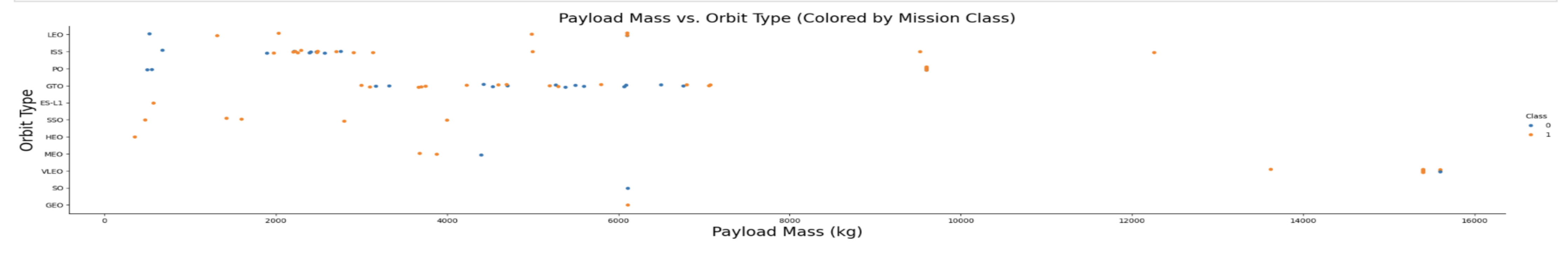
Scatter Plot of Flight Number vs. Orbit type

Key Insights & Observations (contd.)

This plot demonstrates **Capability Expansion.**

- **Beyond the Basics:** The data proves SpaceX did not limit operations to "easy" low-altitude missions.
- **Strategic Pivot:** As experience grew (higher flight numbers), the program successfully transitioned to complex, high-energy trajectories.
- **Full Spectrum Operations:** The successful execution of **GTO** (Geostationary Transfer Orbit) and **Polar** missions confirms mastery of diverse orbital requirements.

Payload Mass and Orbit type



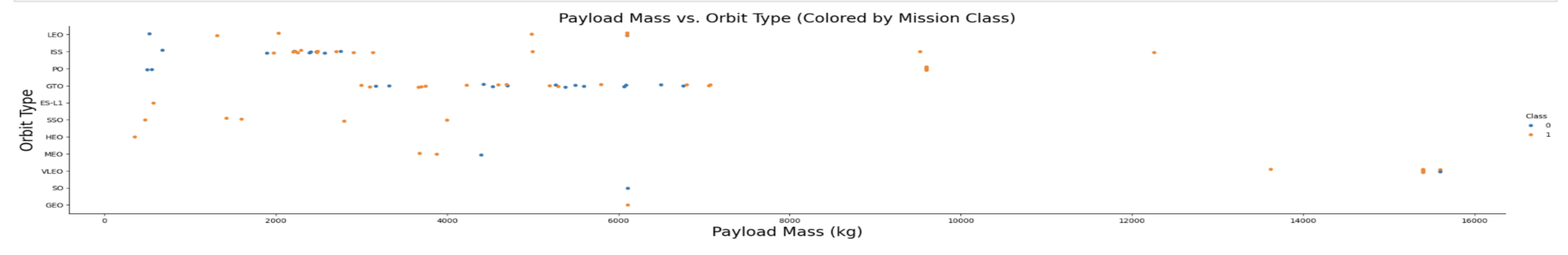
Scatter Plot of Payload Mass and Orbit type

Plot Overview

This chart uses sns.catplot to explore the relationship between how heavy a satellite is and where it is going.

X-Axis: Payload Mass (kg).
Y-Axis: Orbit Type (e.g., LEO, GTO, ISS, PO).
Color (Hue): Class (0 = Failure, 1 = Success).

Payload Mass and Orbit Type (contd.)



Scatter Plot of Payload Mass and Orbit type

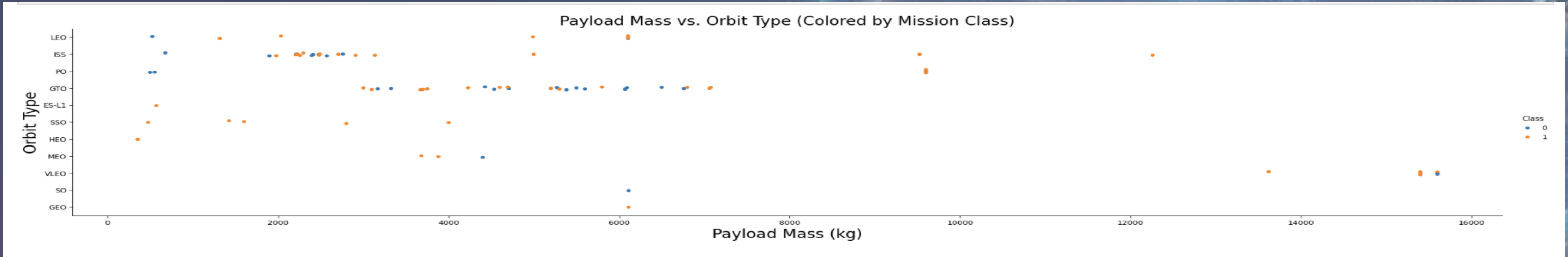
Key Insights & Observations

1. **The Physics of Spaceflight (Mass vs. Distance):**

   - **LEO & ISS (Heavy):** You will see data points for **LEO** (Low Earth Orbit) and **ISS** (International Space Station) stretching far to the right, often reaching **high payload masses**.
   - **GTO (Lighter):** In contrast, **GTO** (Geostationary Transfer Orbit) missions are clustered in the **lower to medium mass range** (typically under 6,000 kg).

Explanation:
This perfectly illustrates the "Rocket Equation" trade-off: It takes significantly more fuel to get to a high-energy orbit like GTO. Therefore, the rocket cannot carry as much cargo as it can to a low orbit like LEO.

Payload Mass and Orbit Type (contd.)

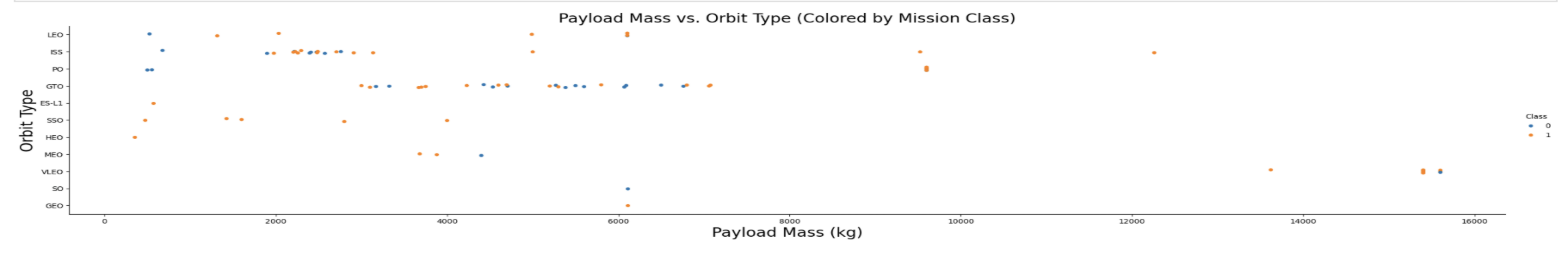


Scatter Plot of Payload Mass and Orbit type

Key Insights & Observations (contd.)

2. **Mission Distinctiveness:**
   - **Polar & SSO:** These orbits show a very distinct, narrow cluster of payload masses. This suggests that missions to these specific orbits are often standardized observation satellites of a similar size/class.
3. **Success Distribution:**
   - The plot shows that even with heavier payloads to LEO/ISS, the success rate (green dots) remains high, confirming that the Falcon 9 is highly capable of heavy-lift missions to low altitudes.

Payload Mass and Orbit Type (contd.)

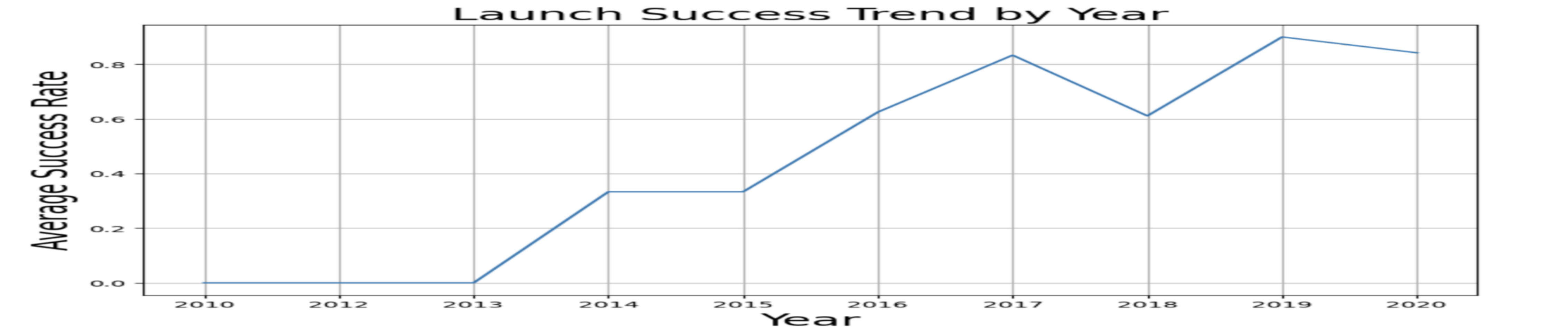


Scatter Plot of Payload Mass and Orbit type

Key Insights & Observations (contd.)

This plot demonstrates **Operational Constraints**. It visually explains why every launch does not carry the maximum payload—the destination (Orbit) dictates the weight limit.

Launch Success Trend by Year



Launch Succes Tred by Year

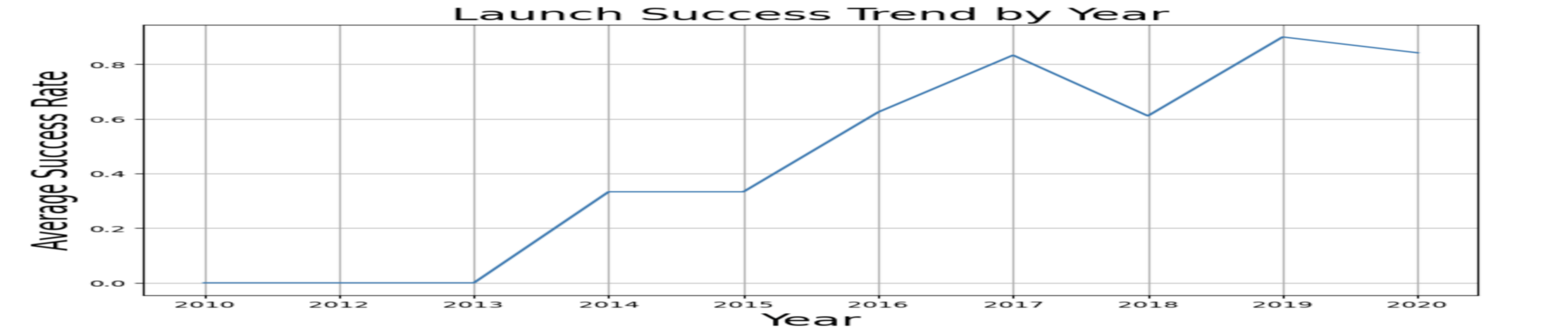
Plot Overview

This visualization uses a **Line Chart** to display the average success rate of launches over time. **Data Processing:** The script extracts the year from the Date column, groups all launches by that year, and calculates the mean of the Class column (Success Rate).

X-Axis: Year (2013, 2014, etc.).

Y-Axis: Average Success Rate (0.0 to 1.0).

Launch Success Trend by Year (contd.)



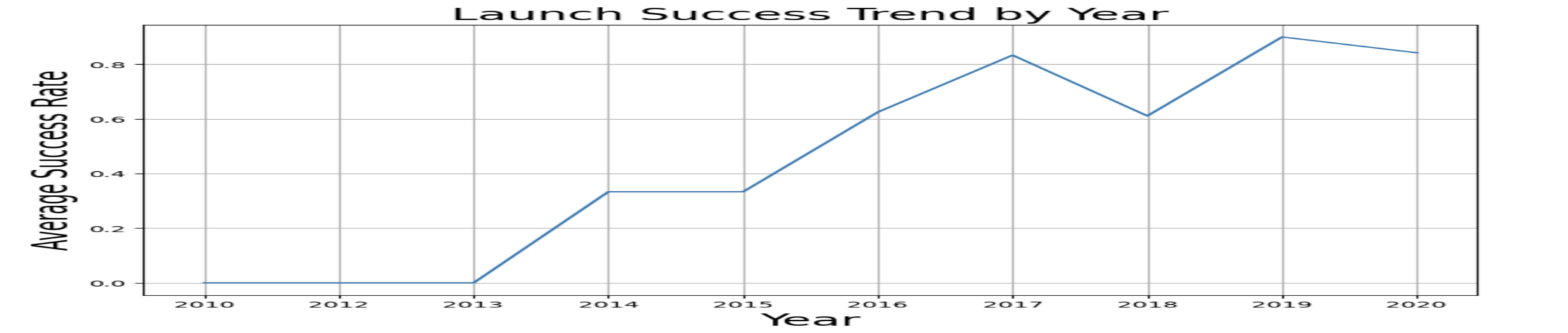
Launch Succes Tred by Year

Key Insights & Observations

1. **The "Learning Curve":**
   Early Volatility (2013–2015): The line likely starts lower or shows fluctuations in the early years. This visually represents the experimental phase where SpaceX was actively testing and refining its landing technology, leading to inconsistent results.

2. **Steady Improvement:**
   The Upward Trend: From around 2016 onwards, the line shows a distinct upward trajectory. This correlates with the introduction of the "Full Thrust" (FT) version of the Falcon 9, which significantly improved performance and reliability.

Launch Success Trend by Year (contd.)



Launch Succes Tred by Year

Key Insights & Observations (contd.)

3. **Stabilization (Maturity):**
   High Reliability (2017+): In the most recent years of the dataset, the line **stabilizes** near the top (close to 1.0 or 100%). This indicates that the Falcon 9 system has matured into a highly reliable operational vehicle.

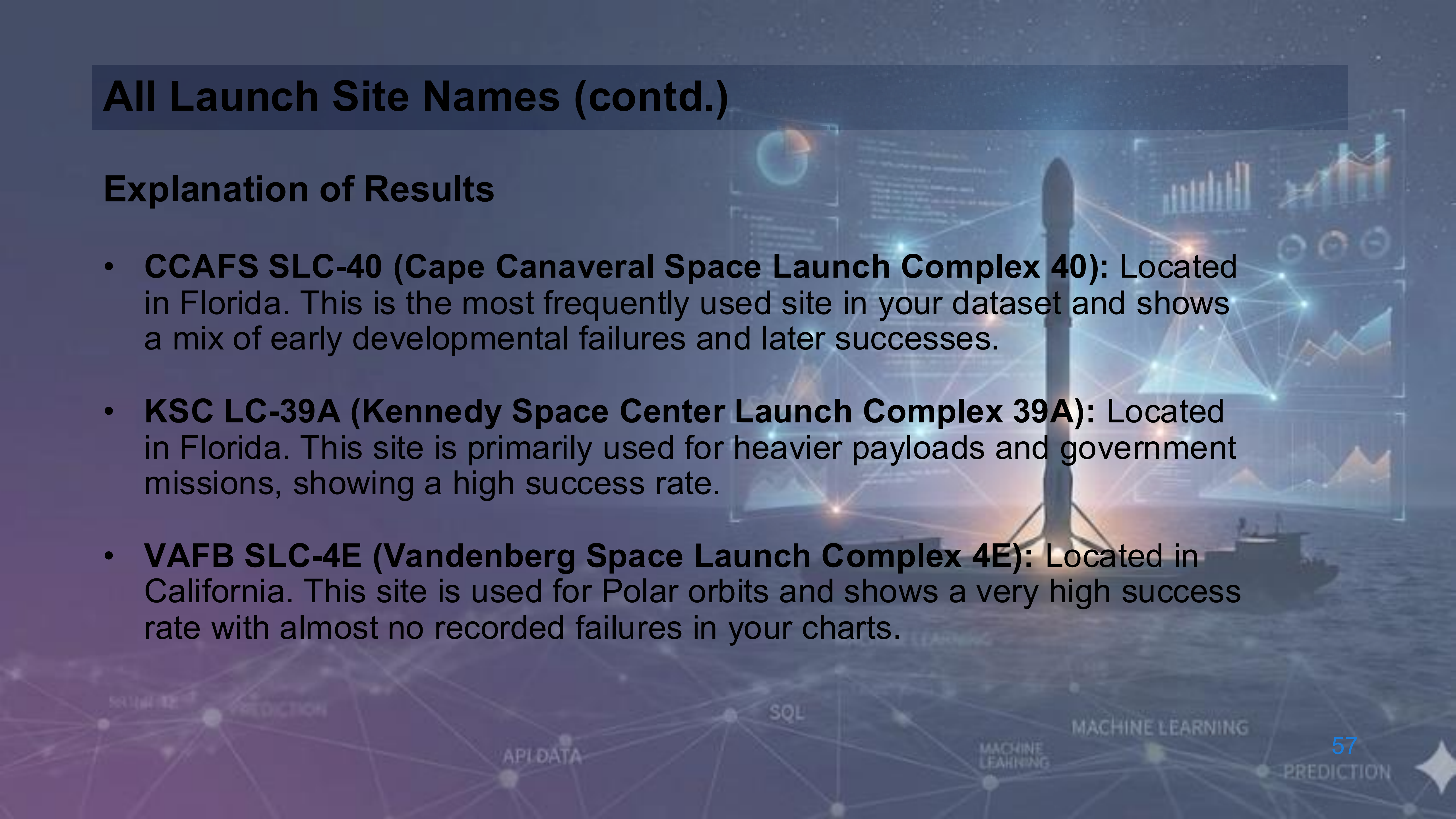
This chart answers the stakeholder question: **"Is the risk decreasing?"** It proves that the failures observed in the earlier scatter plots are largely historical. The trend line demonstrates that the probability of success today is significantly higher than it was at the start of the program..

All launch site names

```
SQL>   SELECT
       DISTINCT Launch_Site
       FROM SPACEXTBL;
```

Unique Launch Sites

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



All Launch Site Names (contd.)

Explanation of Results

- **CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40):** Located in Florida. This is the most frequently used site in your dataset and shows a mix of early developmental failures and later successes.
- **KSC LC-39A (Kennedy Space Center Launch Complex 39A):** Located in Florida. This site is primarily used for heavier payloads and government missions, showing a high success rate.
- **VAFB SLC-4E (Vandenberg Space Launch Complex 4E):** Located in California. This site is used for Polar orbits and shows a very high success rate with almost no recorded failures in your charts.

Launch Site names beginning with “KSC”

```
SQL>    SELECT *
        FROM SPACEXTBL
        WHERE Launch_Site LIKE 'KSC%’
        ORDER BY DATE ASC
        LIMIT 5;
```

Date	Time	Booster Version	Launch Site	Payload	Payload Mass (kg)	Orbit	Customer	Landing Outcome
2017-02-19	14:39:00	F9 FT B1031.1	**KSC LC-39A**	**SpaceX CRS-10**	2490	LEO (ISS)	NASA (CRS)	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	**KSC LC-39A**	**EchoStar 23**	5600	GTO	EchoStar	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	**KSC LC-39A**	**SES-10**	5300	GTO	SES	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	**KSC LC-39A**	**NROL-76**	5300	LEO	NRO	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	**KSC LC-39A**	**Inmarsat-5 F4**	6070	GTO	Inmarsat	No attempt

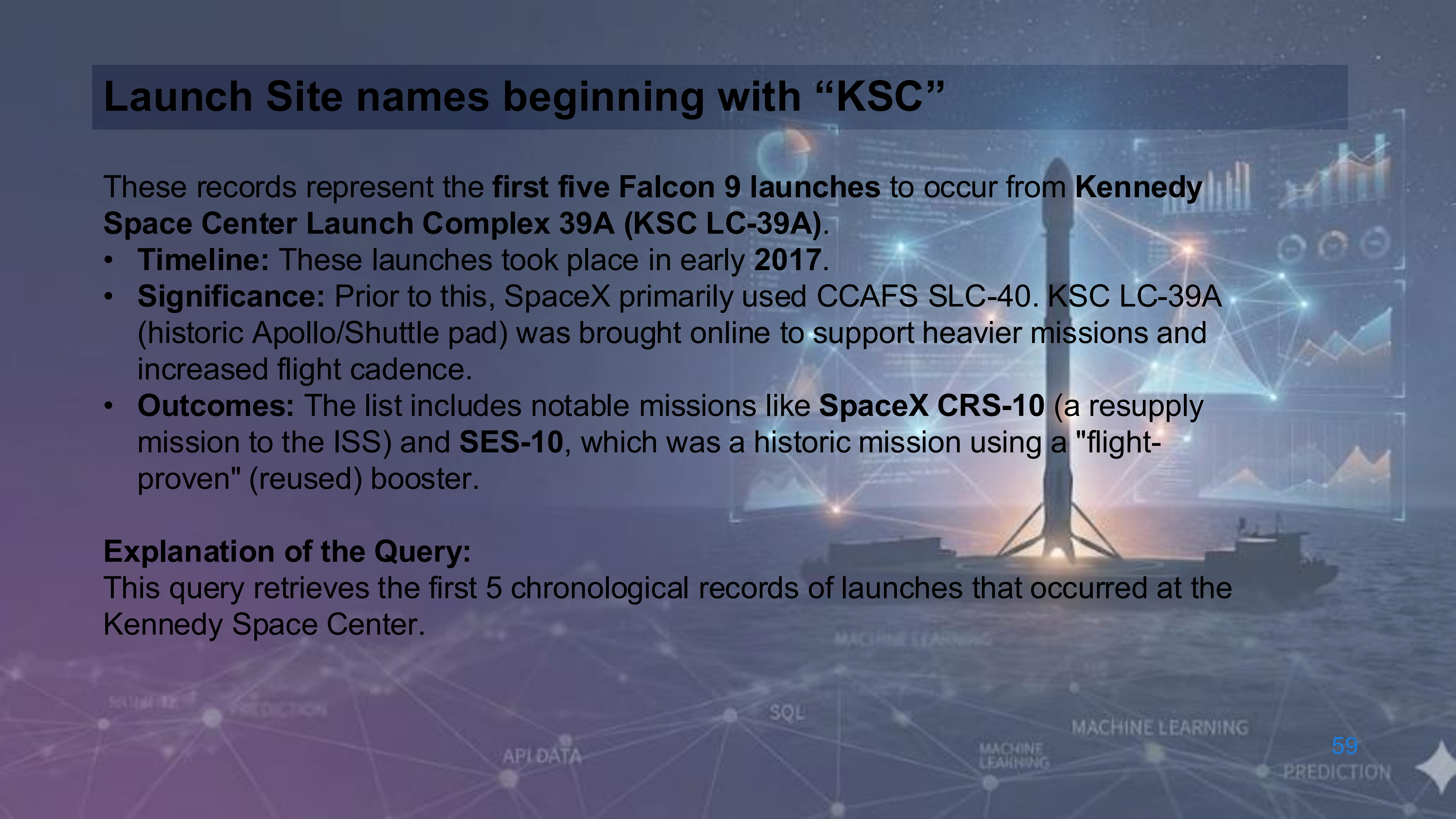
Launch Site names beginning with “KSC”

These records represent the **first five Falcon 9 launches** to occur from **Kennedy Space Center Launch Complex 39A (KSC LC-39A)**.

- **Timeline:** These launches took place in early **2017**.
- **Significance:** Prior to this, SpaceX primarily used CCAFS SLC-40. KSC LC-39A (historic Apollo/Shuttle pad) was brought online to support heavier missions and increased flight cadence.
- **Outcomes:** The list includes notable missions like **SpaceX CRS-10** (a resupply mission to the ISS) and **SES-10**, which was a historic mission using a "flight-proven" (reused) booster.

Explanation of the Query:

This query retrieves the first 5 chronological records of launches that occurred at the Kennedy Space Center.



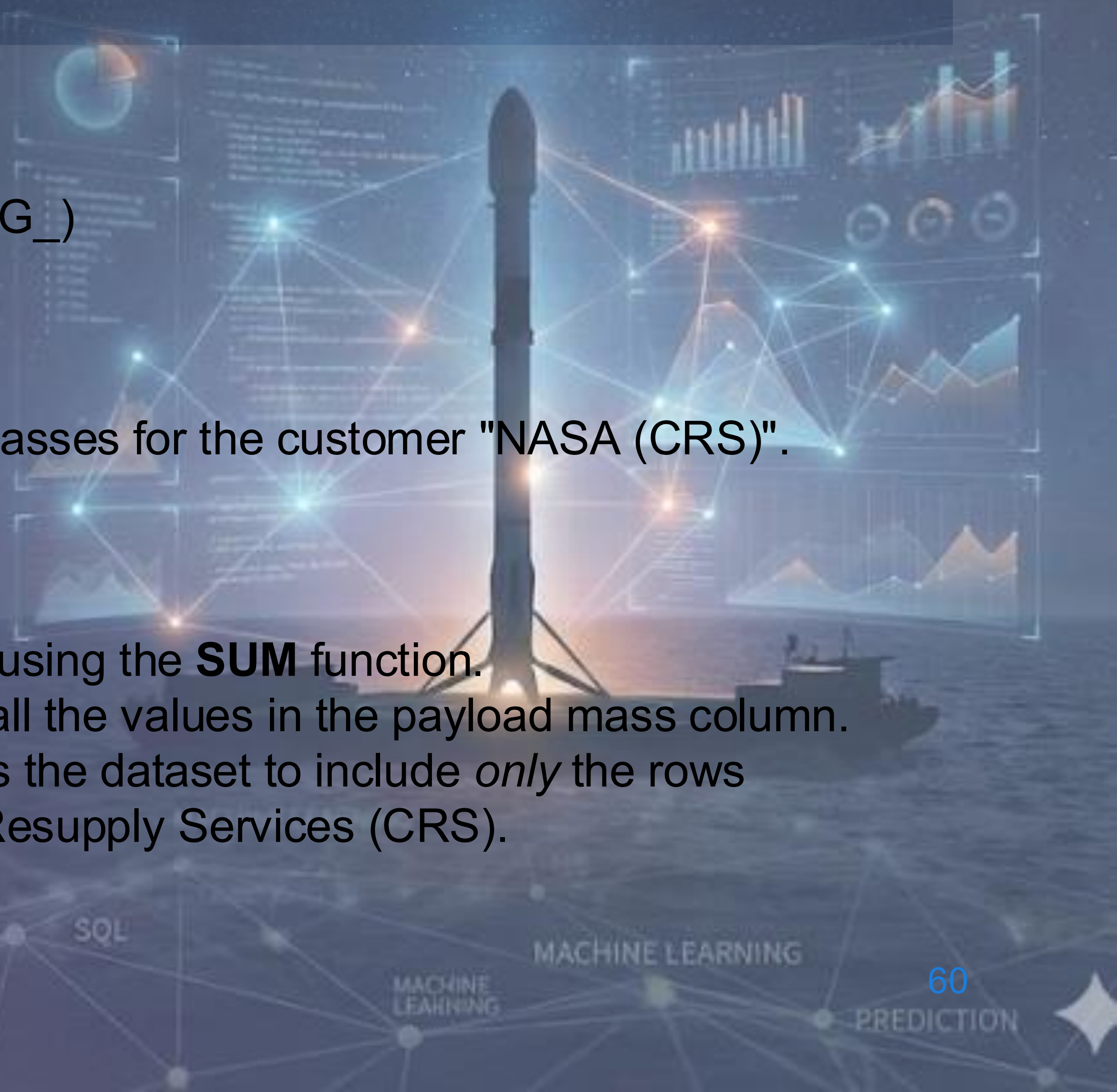
Total Payload Mass

```
SQL>   SELECT SUM(PAYLOAD_MASS__KG_)
       FROM SPACEXTBL
       WHERE Customer = 'NASA (CRS)’;
```

The query calculates the sum of all payload masses for the customer "NASA (CRS)".
Total Payload Mass: 45,596 kg.

Explanation of the Query
To get this result, I performed an aggregation using the **SUM** function.

- **SUM(PAYLOAD_MASS__KG_)**: Adds up all the values in the payload mass column.
- **WHERE Customer = 'NASA (CRS)'**: Filters the dataset to include *only* the rows where the client was NASA's Commercial Resupply Services (CRS).



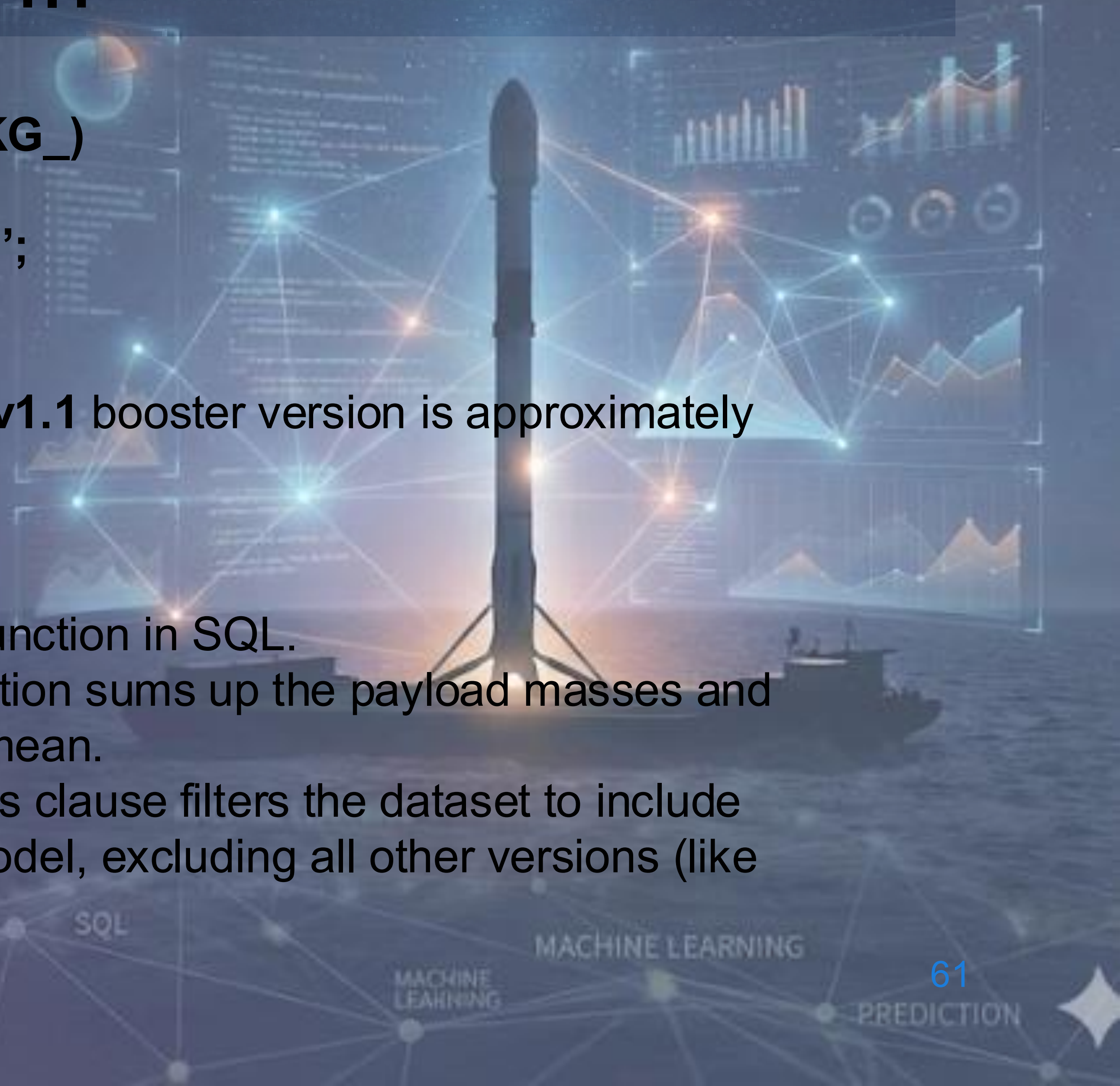
Average Payload mass by F9 v1.1

```
SQL>   SELECT AVG(PAYLOAD_MASS__KG_)
       FROM SPACEXTBL
       WHERE Booster_Version = 'F9 v1.1’;
```

Average Payload Mass Result:
The average payload mass carried by the **F9 v1.1** booster version is approximately **2,928.4 kg**.

Explanation of the Query:
To calculate this, I used the **AVG** aggregate function in SQL.

- **AVG(PAYLOAD_MASS__KG_)**: This function sums up the payload masses and divides by the count of records to find the mean.
- **WHERE Booster_Version = 'F9 v1.1'**: This clause filters the dataset to include *only* the missions flown by the "F9 v1.1" model, excluding all other versions (like v1.0 or FT).



First successful ground landing date

```
SQL>    SELECT MIN(Date)
        FROM SPACEXTBL
        WHERE "Landing _Outcome" = 'Success (drone ship)’;
```

Date of First Successful Landing
The first successful landing on a drone ship occurred on **2016-04-08** (April 8, 2016).

Explanation of Results:
The Event: This date marks the historic moment when the Falcon 9 booster (mission **SpaceX CRS-8**) successfully landed on the drone ship *"Of Course I Still Love You"* in the Atlantic Ocean.

The Query Logic: To find the "first" occurrence, I used the **MIN()** function on the Date column. I filtered the records to look only for outcomes that exactly match the string **'Success (drone ship)'**.

Successful drone ship landing with payload between 4,000kg and 6,000kg

```
SQL>   SELECT Booster_Version
       FROM SPACEXTBL
       WHERE "Landing _Outcome" = 'Success (drone ship)’
                 AND PAYLOAD_MASS__KG_ > 4000
                 AND PAYLOAD_MASS__KG_ < 6000;
```

Booster Names

The boosters that successfully landed on a drone ship with a payload between 4,000 kg and 6,000 kg are:

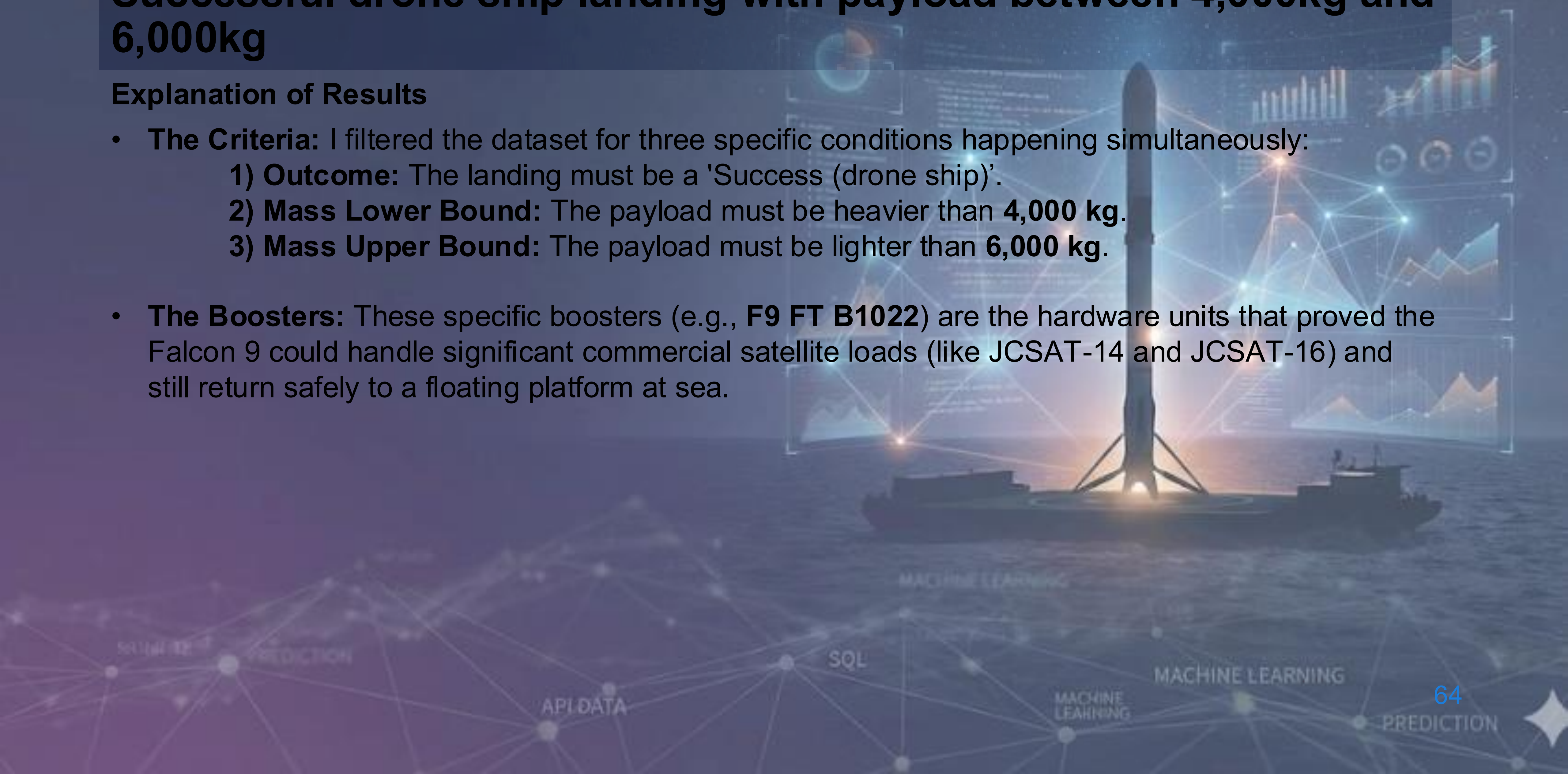
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Successful drone ship landing with payload between 4,000kg and 6,000kg

Explanation of Results

- **The Criteria:** I filtered the dataset for three specific conditions happening simultaneously:
  - **1) Outcome:** The landing must be a 'Success (drone ship)’.
  - **2) Mass Lower Bound:** The payload must be heavier than **4,000 kg**.
  - **3) Mass Upper Bound:** The payload must be lighter than **6,000 kg**.

- **The Boosters:** These specific boosters (e.g., **F9 FT B1022**) are the hardware units that proved the Falcon 9 could handle significant commercial satellite loads (like JCSAT-14 and JCSAT-16) and still return safely to a floating platform at sea.



Total number of successful and failed mission outcomes

```
SQL>   SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Count
       FROM SPACEXTBL
       GROUP BY "Mission_Outcome";

OUTPUT:        [('Failure (in flight)', 1),
               ('Success', 98),
               ('Success ', 1),
               ('Success (payload status unclear)', 1)]
```

Calculation of Mission Outcomes

The dataset records **100 successful missions** and **1 failed mission** in total.

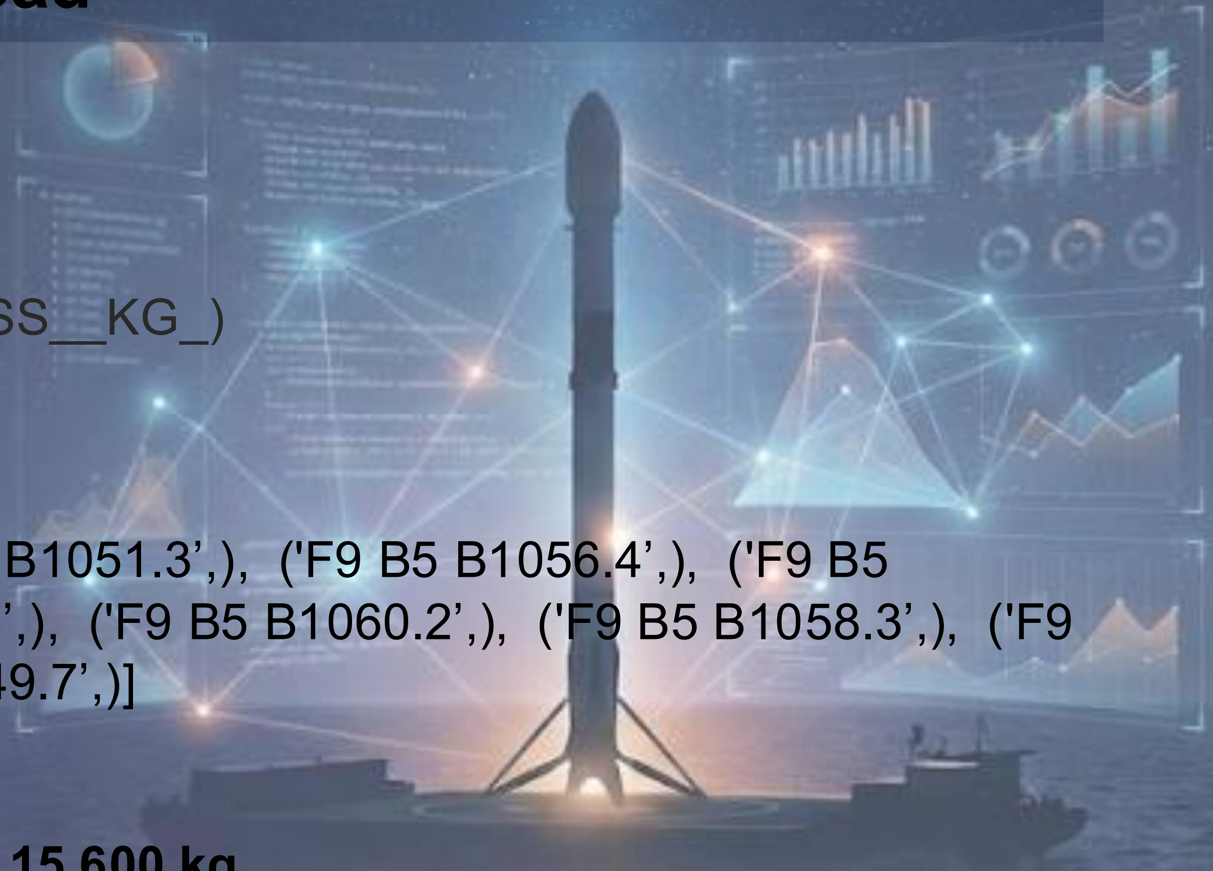
- **Success:** 98
- **Success (payload status unclear):** 1
- **Success (with extra formatting):** 1
- **Failure (in flight):** 1

Total number of successful and failed mission outcomes

Explanation of Results

- Vast majority of missions in this dataset were successful.
- Only one mission is recorded as an in-flight failure.

Boosters carried maximum payload



```
SQL>    SELECT Booster_Version
        FROM SPACEXTBL
        WHERE PAYLOAD_MASS__KG_ = (
                SELECT MAX(PAYLOAD_MASS__KG_)
                FROM SPACEXTBL);
```

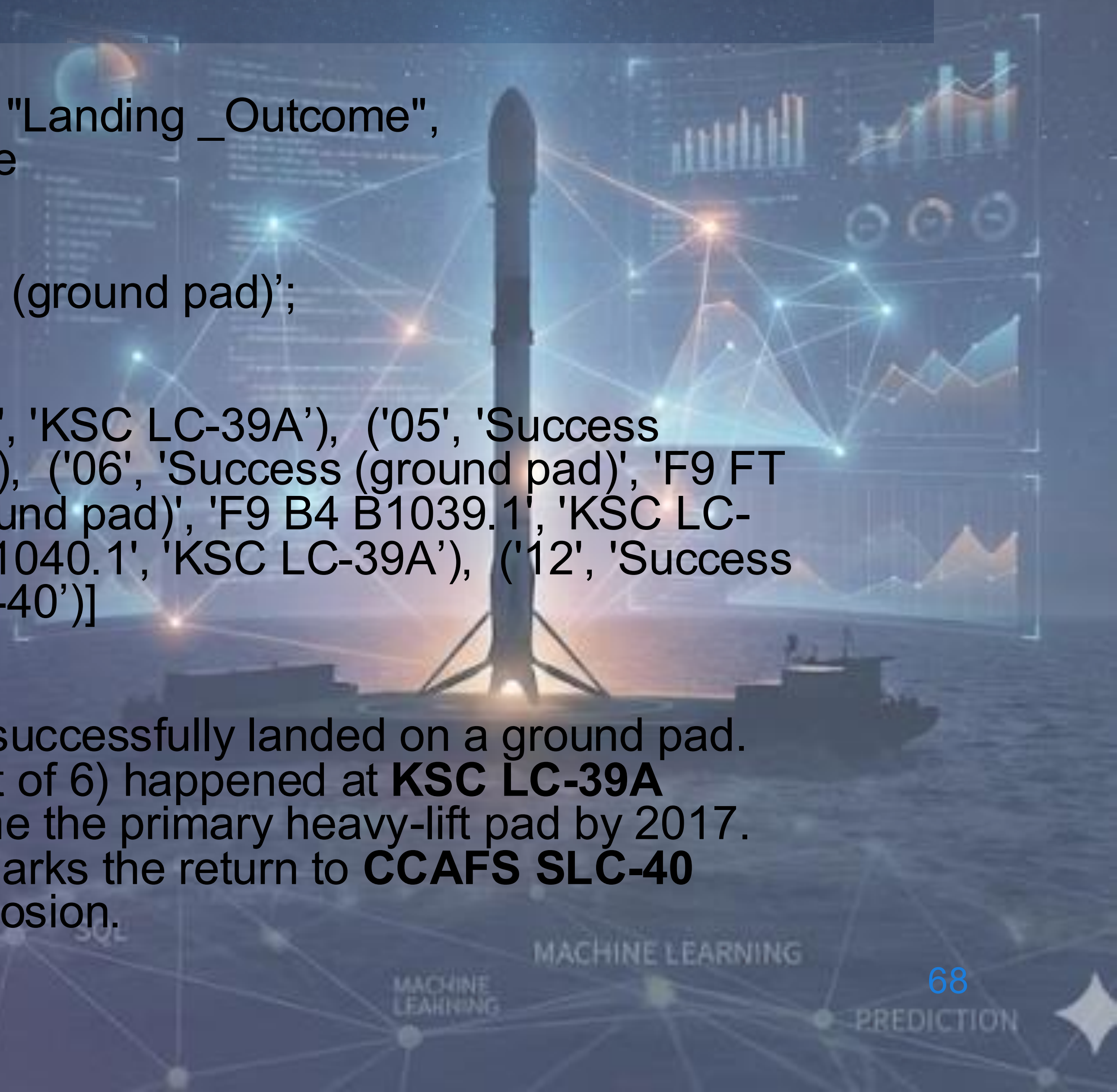
OUTPUT:

```
[('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’,)]
```

Explanation of Results

- Max payload mass recorded in the dataset is **15,600 kg**.
- Corresponds to the **Starlink** missions, where SpaceX launches batches of internet satellites.
- The boosters listed above are all **Falcon 9 Block 5** versions (indicated by "F9 B5"), which are the latest and most powerful iteration, capable of lifting this maximum load repeatedly.

2015 Launch Records



```
SQL>    SELECT substr(Date, 6, 2) as Month, "Landing _Outcome",
                  Booster_Version, Launch_Site
        FROM SPACEXTBL
        WHERE substr(Date, 1, 4)='2017’
        AND "Landing _Outcome" = 'Success (ground pad)’;
```

OUTPUT:

```
[('02', 'Success (ground pad)', 'F9 FT B1031.1', 'KSC LC-39A’),  ('05', 'Success
(ground pad)', 'F9 FT B1032.1', 'KSC LC-39A’),  ('06', 'Success (ground pad)', 'F9 FT
B1035.1', 'KSC LC-39A’),  ('08', 'Success (ground pad)', 'F9 B4 B1039.1', 'KSC LC-
39A’),  ('09', 'Success (ground pad)', 'F9 B4 B1040.1', 'KSC LC-39A’),  ('12', 'Success
(ground pad)', 'F9 FT  B1035.2', 'CCAFS SLC-40’)]
```

Explanation of Results

I found 6 records for 2017 where the mission successfully landed on a ground pad.

- **The Trend:** Notice that most of these (5 out of 6) happened at **KSC LC-39A** (Kennedy Space Center), which had become the primary heavy-lift pad by 2017.
- **The Return:** The final entry in December marks the return to **CCAFS SLC-40** after it was repaired following the 2016 explosion.

2017 Launch Records

```
SQL>    SELECT substr(Date, 6, 2) as Month, "Landing _Outcome", Booster_Version, Launch_Site
        FROM SPACEXTBL
        WHERE substr(Date, 1, 4)='2017’
        AND "Landing _Outcome" = 'Success (ground pad)’;
```

OUTPUT:

```
[('02', 'Success (ground pad)', 'F9 FT B1031.1', 'KSC LC-39A’),  ('05', 'Success (ground pad)', 'F9 FT
B1032.1', 'KSC LC-39A’),  ('06', 'Success (ground pad)', 'F9 FT B1035.1', 'KSC LC-39A'),
 ('08', 'Success (ground pad)', 'F9 B4 B1039.1', 'KSC LC-39A’),  ('09', 'Success (ground pad)', 'F9 B4
B1040.1', 'KSC LC-39A’),  ('12', 'Success (ground pad)', 'F9 FT  B1035.2', 'CCAFS SLC-40’)]
```

Explanation of Results

I found 6 records for 2017 where the mission successfully landed on a ground pad.

- **The Trend:** Notice that most of these (5 out of 6) happened at **KSC LC-39A** (Kennedy Space Center), which had become the primary heavy-lift pad by 2017.
- **The Return:** The final entry in December marks the return to **CCAFS SLC-40** after it was repaired following the 2016 explosion.

Rank landing outcomes between 2010-06-04 and 2017-03-20

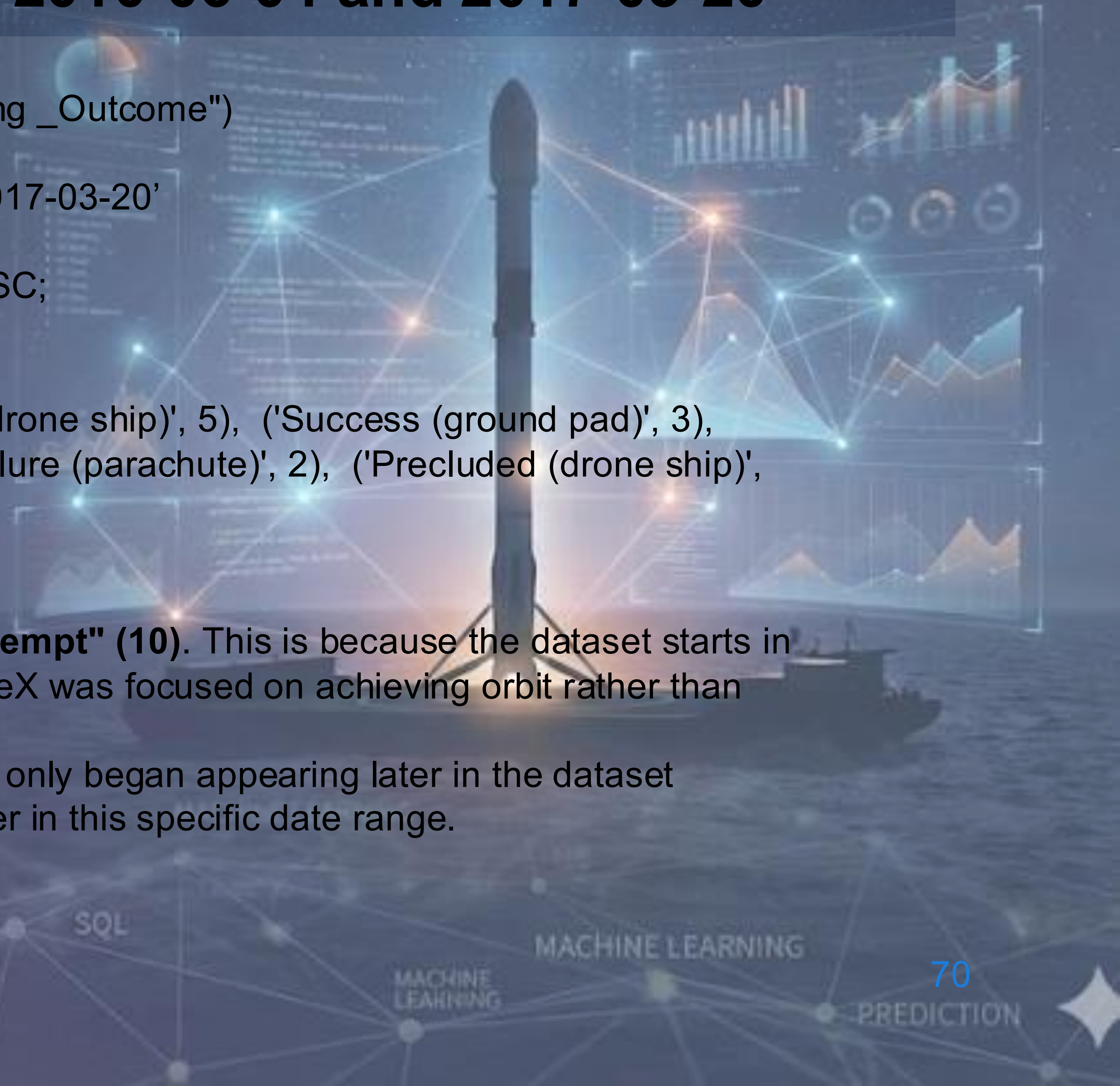
```
SQL>    SELECT "Landing _Outcome", COUNT("Landing _Outcome")
        FROM SPACEXTBL
        WHERE Date BETWEEN '2010-06-04' AND '2017-03-20’
        GROUP BY "Landing _Outcome"
        ORDER BY COUNT("Landing _Outcome") DESC;
```

OUTPUT:

```
[('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 of Results

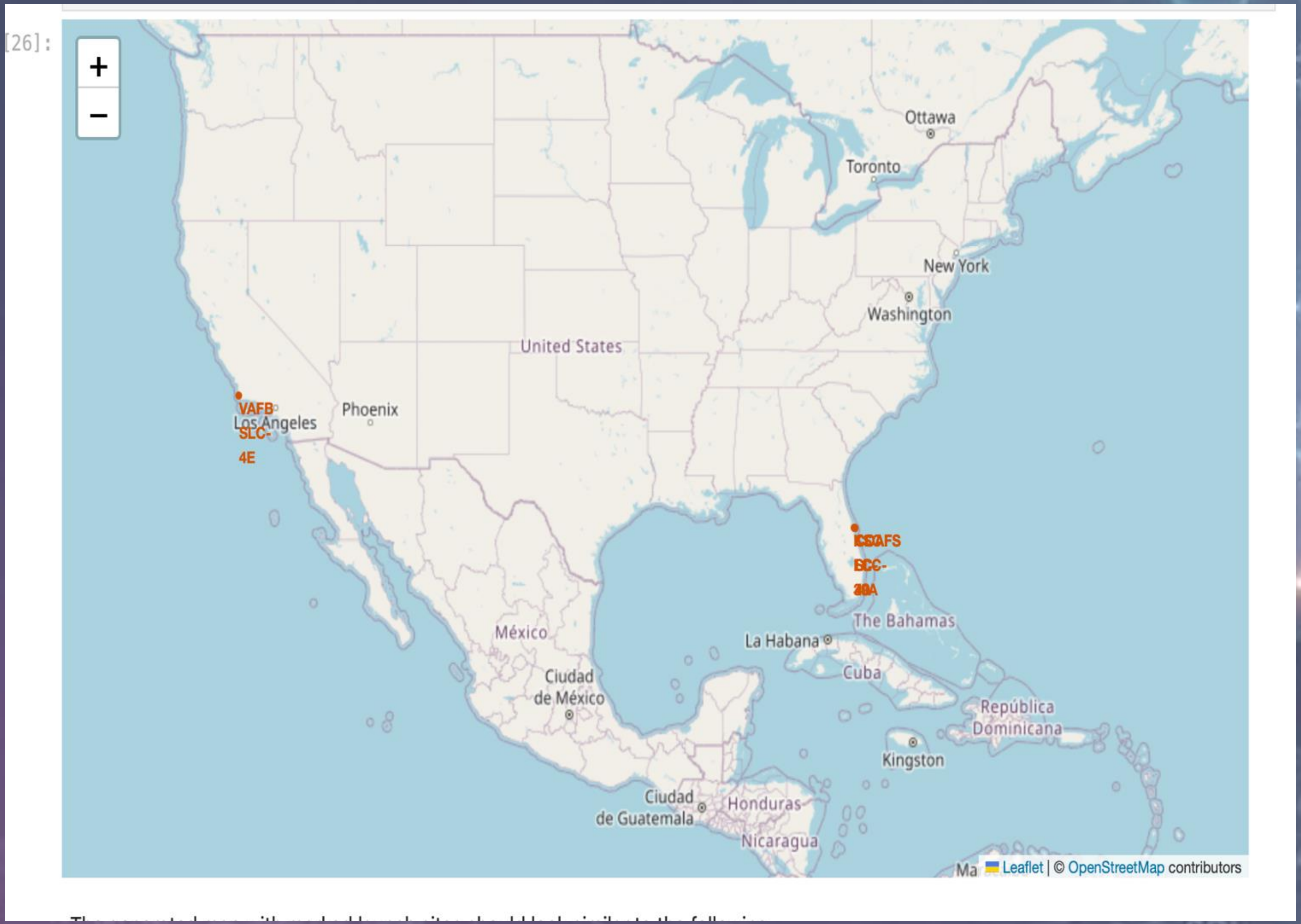
- most common outcome during this period was **"No attempt" (10)**. This is because the dataset starts in 2010, covering the early Falcon 9 v1.0 era when SpaceX was focused on achieving orbit rather than recovering the booster.
- Successful landings (both drone ship and ground pad) only began appearing later in the dataset (2015–2016), which explains why their counts are lower in this specific date range.



Section 3

PREDICTION
API DATA
SQL
MACHINE LEARNING
MACHINE LEARNING
MACHINE LEARNING
PREDICTION

Strategic Distribution of Launch Sites

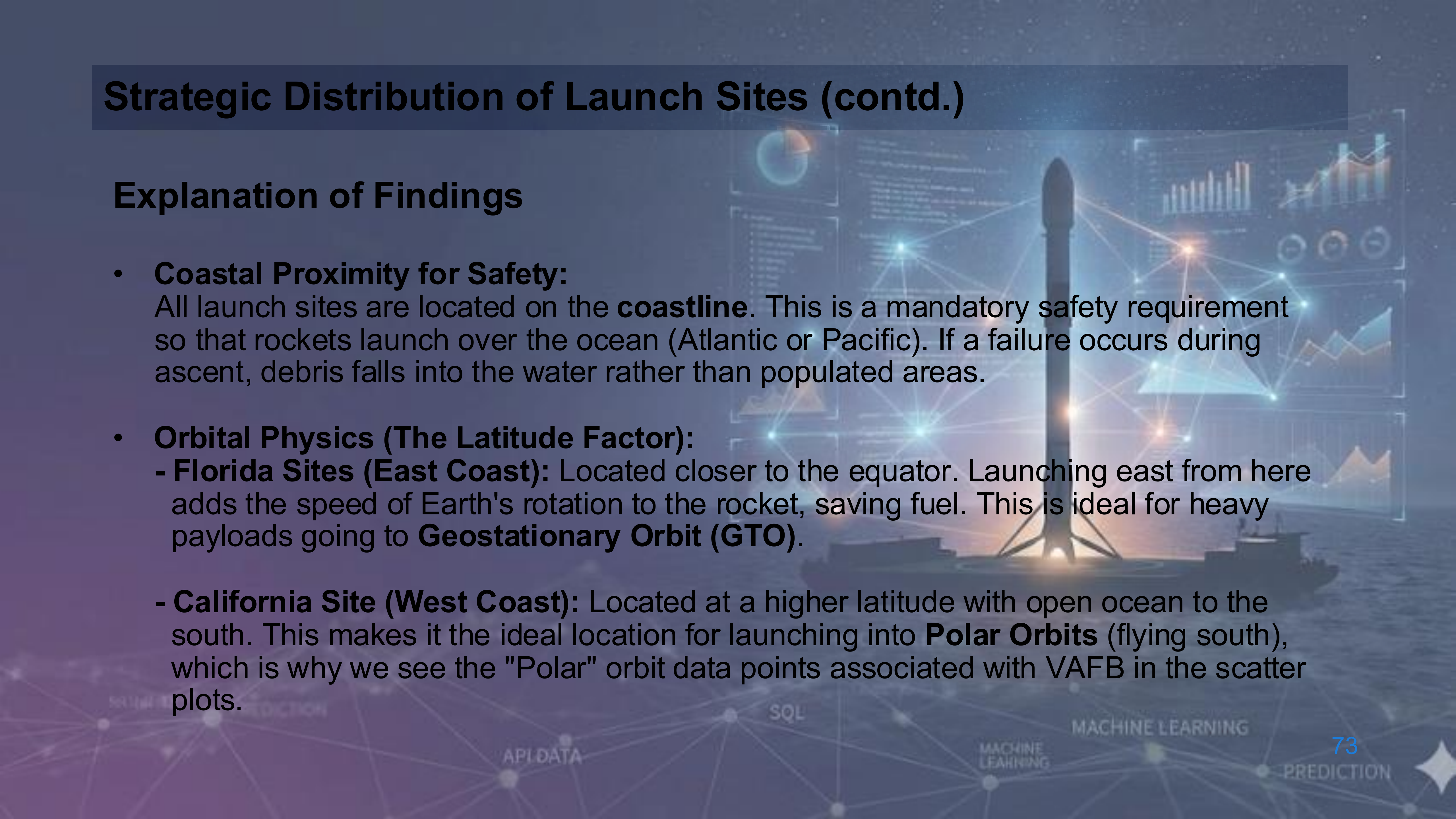


What does the 'Strategic Distribution of Launch Sites' map display?:

- **Geographic Spread:** The map visualizes the locations of SpaceX's primary launch facilities across the United States.

- **The Markers:**
  - West Coast (California): One site (**VAFB SLC-4E**) located at Vandenberg Air Force Base.

  - East Coast (Florida): Three sites (**KSC LC 39A**, **CCAFS LC-40**, **CCAFS SLC-40**) clustered closely together at Cape Canaveral/Kennedy Space Center.



Strategic Distribution of Launch Sites (contd.)

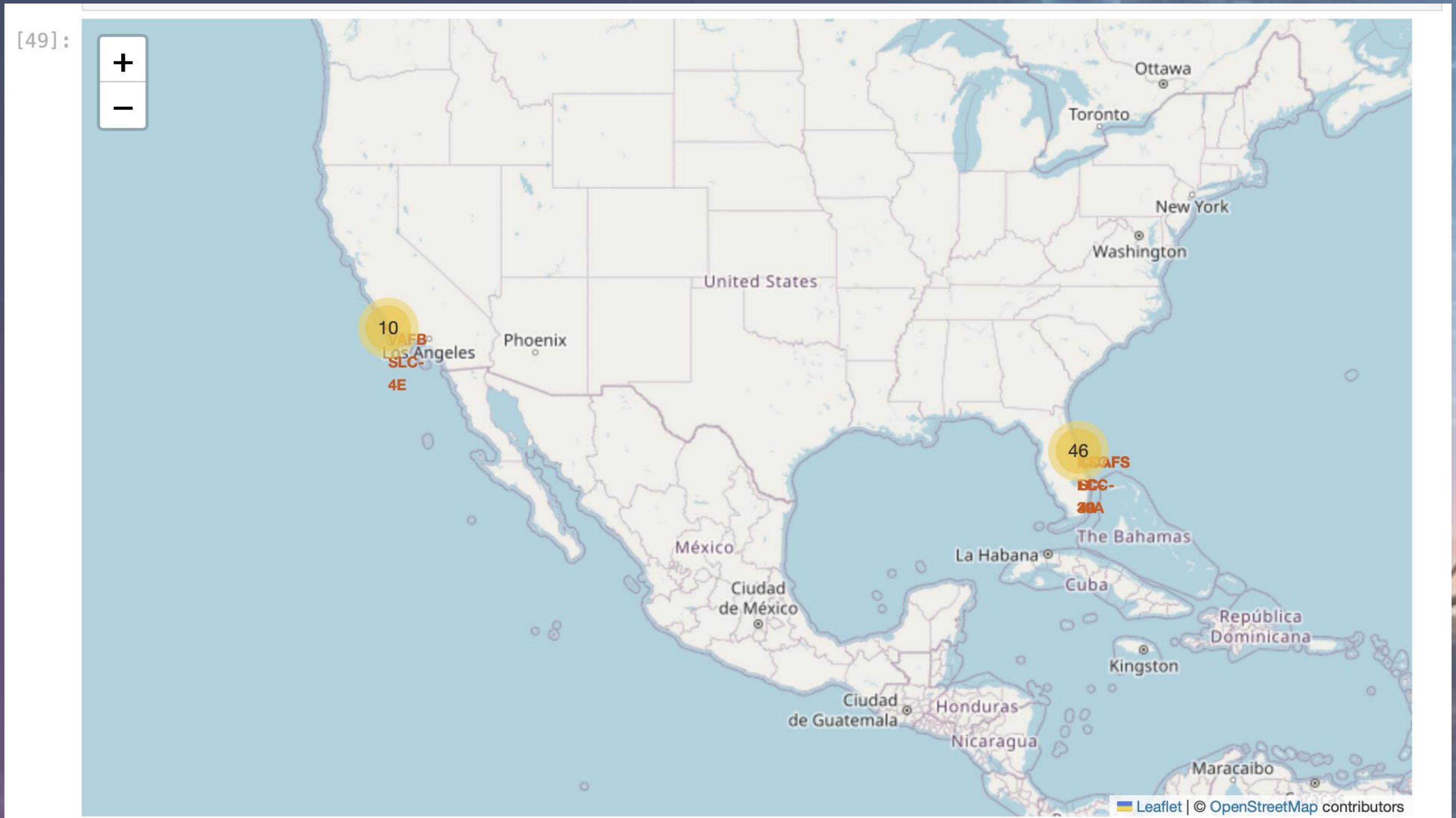
Explanation of Findings

- **Coastal Proximity for Safety:**
  All launch sites are located on the **coastline**. This is a mandatory safety requirement so that rockets launch over the ocean (Atlantic or Pacific). If a failure occurs during ascent, debris falls into the water rather than populated areas.

- **Orbital Physics (The Latitude Factor):**
  - Florida Sites (East Coast): Located closer to the equator. Launching east from here adds the speed of Earth's rotation to the rocket, saving fuel. This is ideal for heavy payloads going to **Geostationary Orbit (GTO)**.

  - California Site (West Coast): Located at a higher latitude with open ocean to the south. This makes it the ideal location for launching into **Polar Orbits** (flying south), which is why we see the "Polar" orbit data points associated with VAFB in the scatter plots.

Launch Site Reliability: Success vs. Failure Analysis



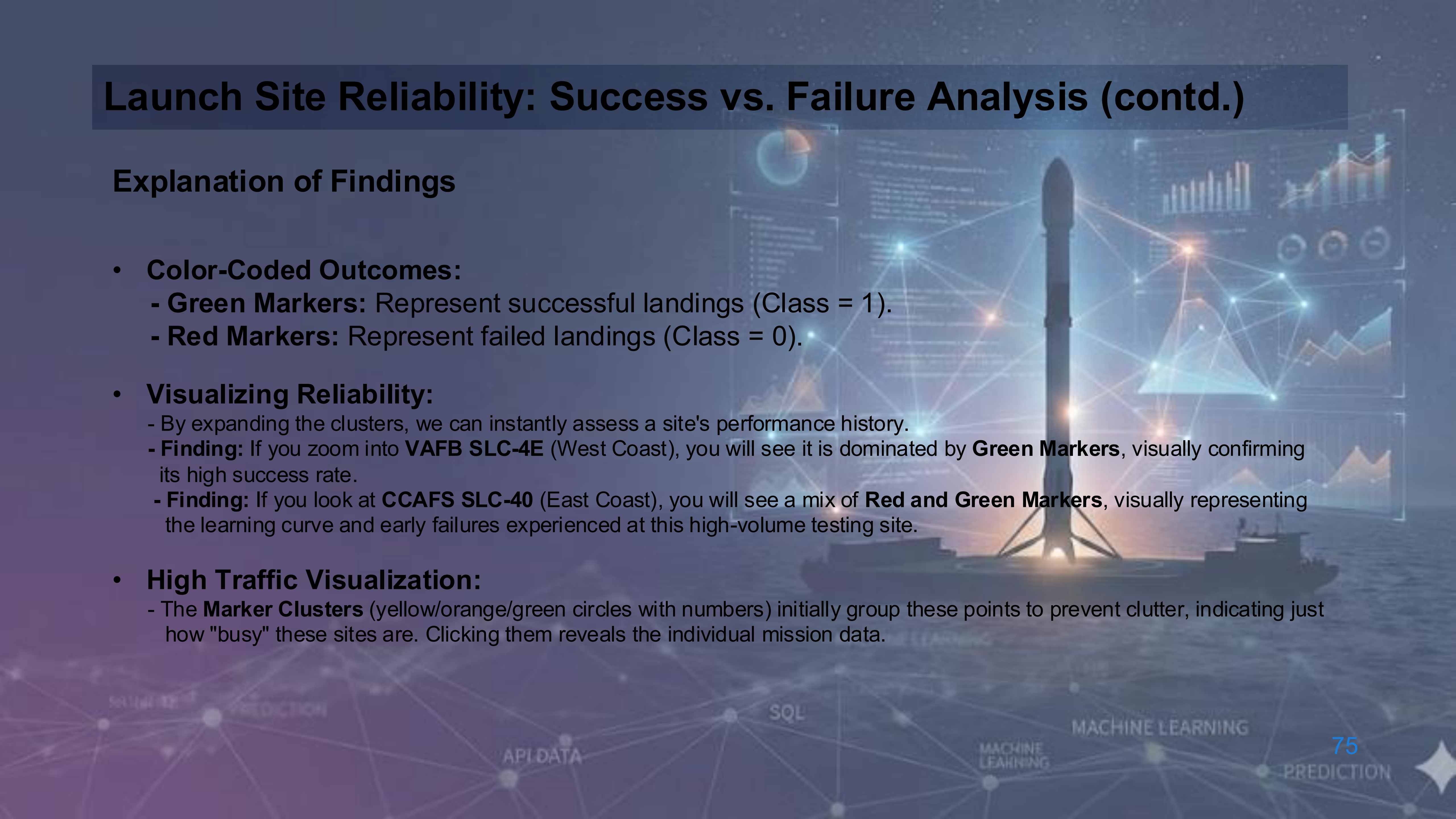
Launch Site Reliability: Success vs. Failure Analysis (contd.)

Explanation of Findings

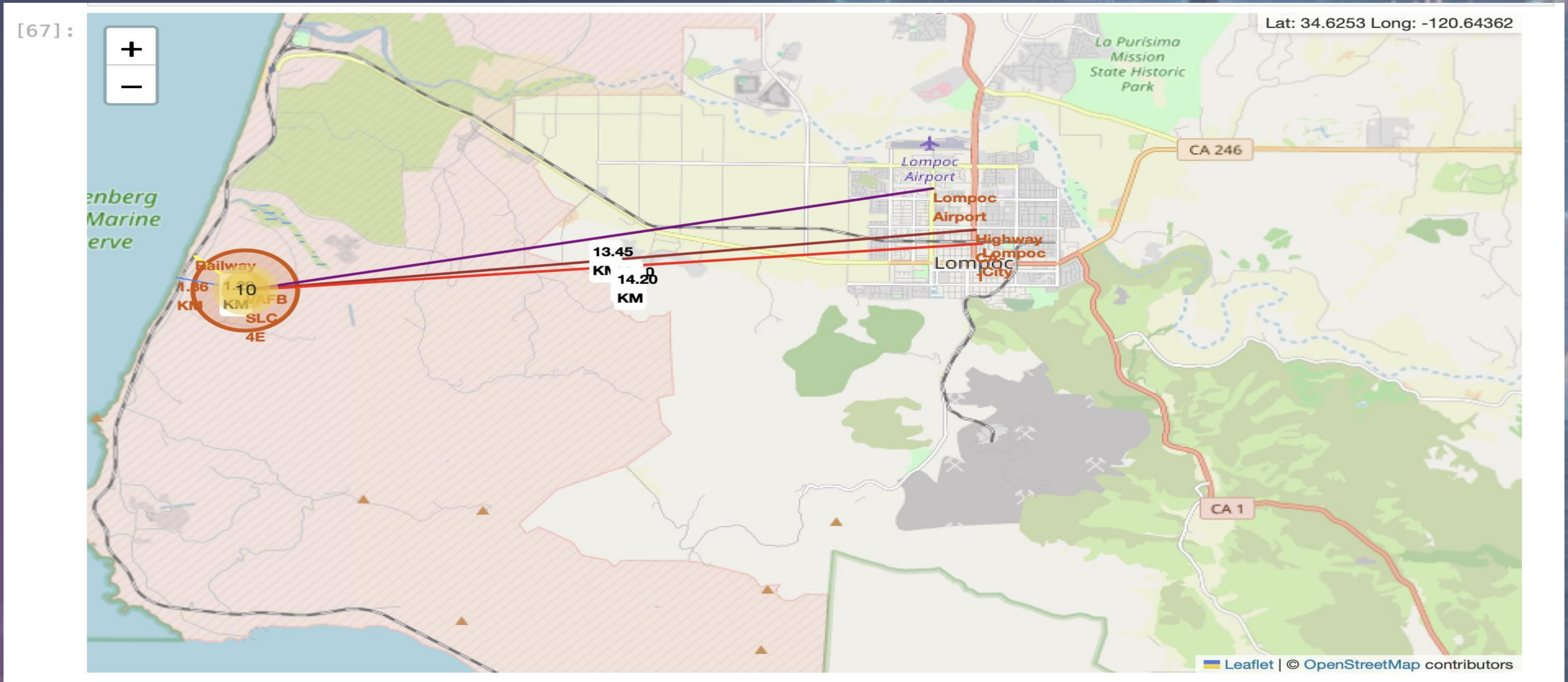
- **Color-Coded Outcomes:**
  - Green Markers: Represent successful landings (Class = 1).
  - Red Markers: Represent failed landings (Class = 0).

- **Visualizing Reliability:**
  - By expanding the clusters, we can instantly assess a site's performance history.
  - Finding: If you zoom into **VAFB SLC-4E** (West Coast), you will see it is dominated by **Green Markers**, visually confirming its high success rate.
  - Finding: If you look at **CCAFS SLC-40** (East Coast), you will see a mix of **Red and Green Markers**, visually representing the learning curve and early failures experienced at this high-volume testing site.

- **High Traffic Visualization:**
  - The **Marker Clusters** (yellow/orange/green circles with numbers) initially group these points to prevent clutter, indicating just how "busy" these sites are. Clicking them reveals the individual mission data.



Operational Logistics: Proximity to Critical Infrastructure



[67]:

Lat: 34.6253 Long: -120.64362

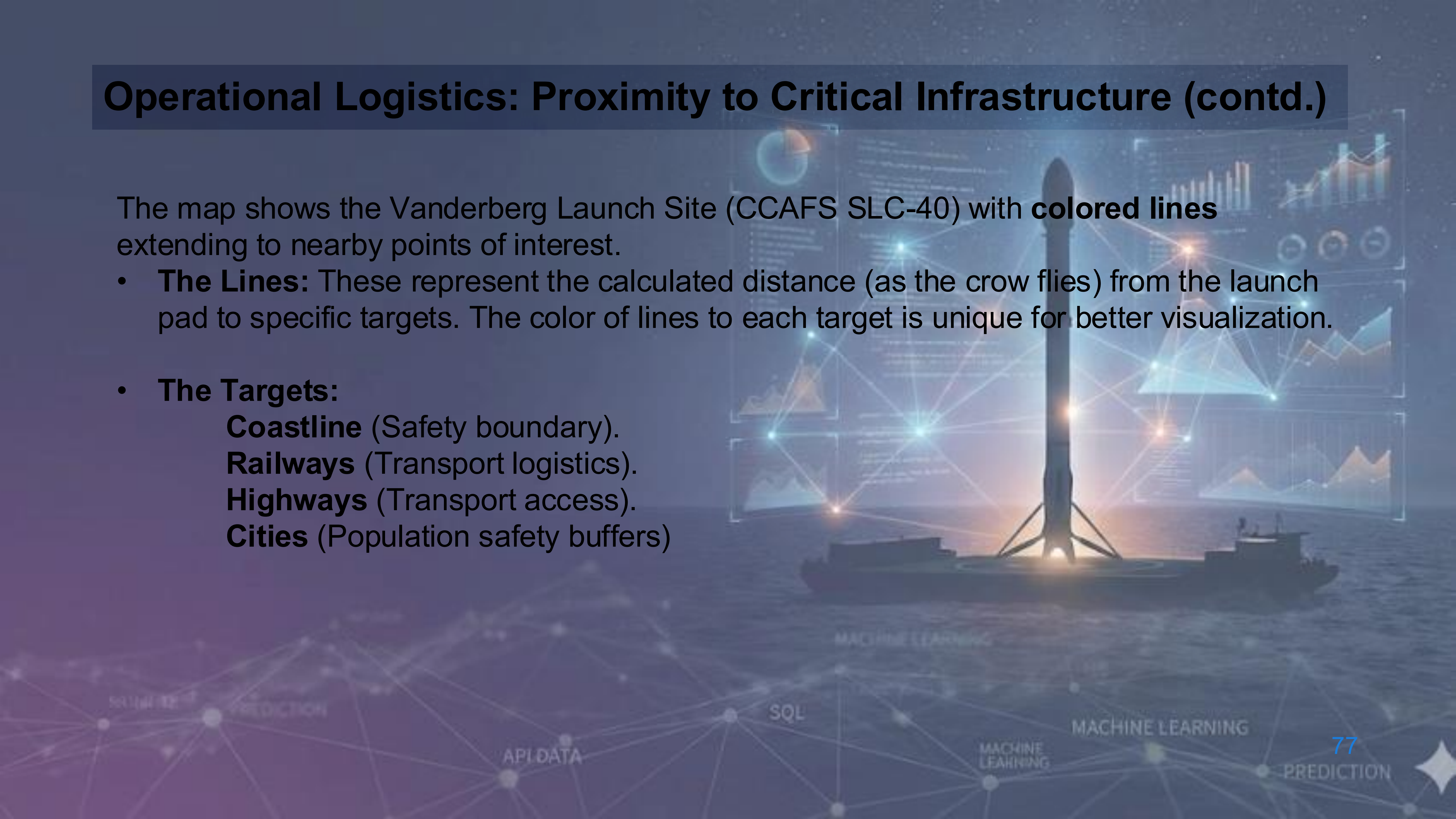
Operational Logistics: Proximity to Critical Infrastructure (contd.)

The map shows the Vanderberg Launch Site (CCAFS SLC-40) with **colored lines** extending to nearby points of interest.

- **The Lines:** These represent the calculated distance (as the crow flies) from the launch pad to specific targets. The color of lines to each target is unique for better visualization.

- **The Targets:**

  Coastline (Safety boundary).
  Railways (Transport logistics).
  Highways (Transport access).
  Cities (Population safety buffers)



Operational Logistics: Proximity to Critical Infrastructure (contd.)

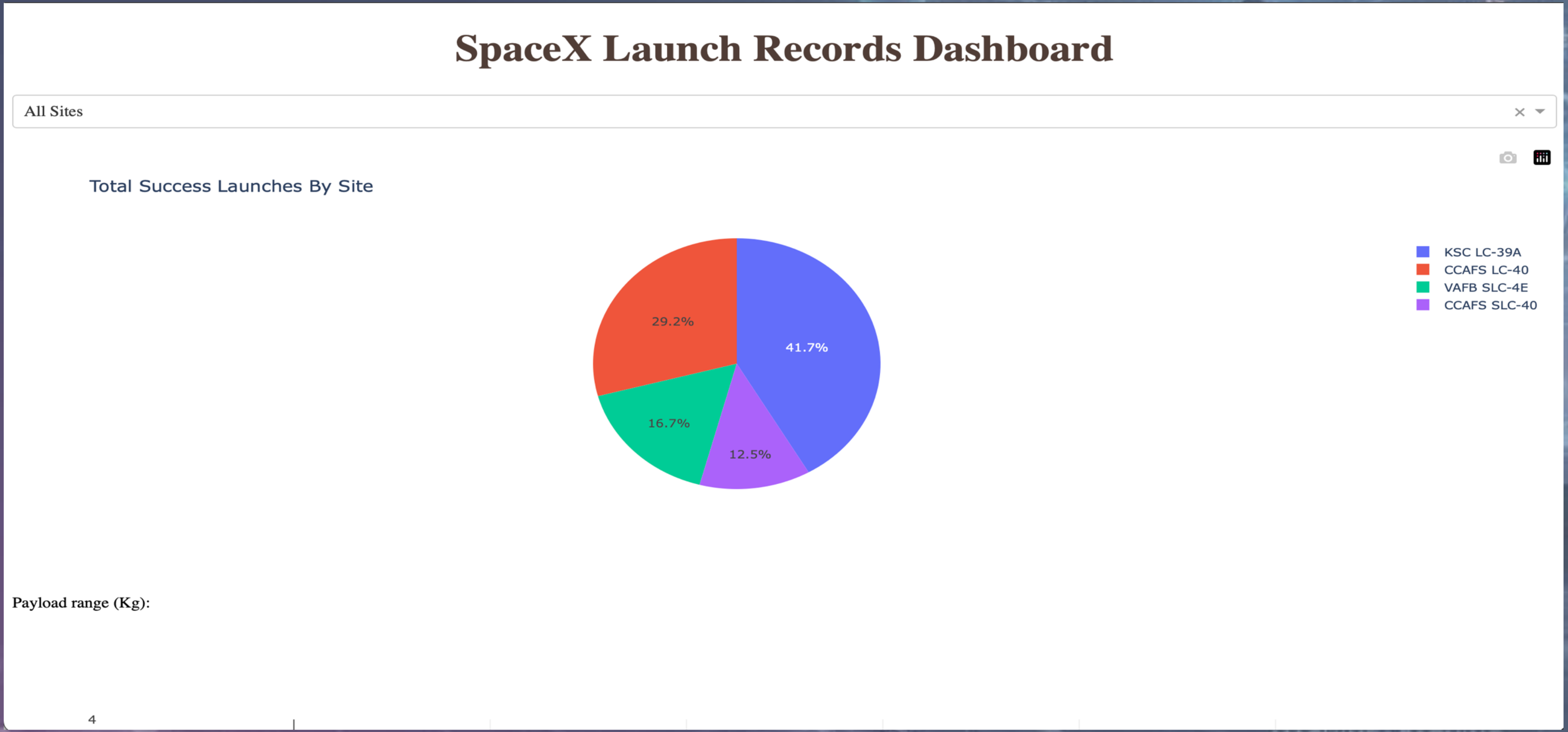
Explanation of Findings

- **Safety First (The Coastline):**
  Finding: The distance to the coastline is extremely short (often **< 1 km**).
  Why: Rockets launch over the ocean. Being right on the coast minimizes the risk to people and property if a launch fails during the initial ascent.
- **Logistical Necessity (The Railway):**
  Finding: The lines show the launch sites are very close to **railway tracks** (e.g., the NASA Railroad).
  Why: Falcon 9 boosters are massive. They cannot easily be transported on standard roads. Direct rail access is crucial for moving heavy hardware to the pad.
- **Public Safety Buffer (Cities):**
  Finding: In contrast to the coast and railways, the distance to the nearest city (like Titusville or Cape Canaveral) is significantly larger.
  Why: This visualizes the safety exclusion zone required to protect the general population from noise and potential hazards.

Section 4

MACHINE LEARNING
PREDICTION
SQL
API DATA
MACHINE LEARNING
MACHINE LEARNING
PREDICTION

Total Success Launches by Site



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Total Success Launches by Site (contd.)

Explanation of Findings

- The slices represent the different Launch Sites (e.g., KSC LC-39A, VAFB SLC-4E).
- The size of the slice represents the **percentage of total successful launches** that occurred at that specific site.

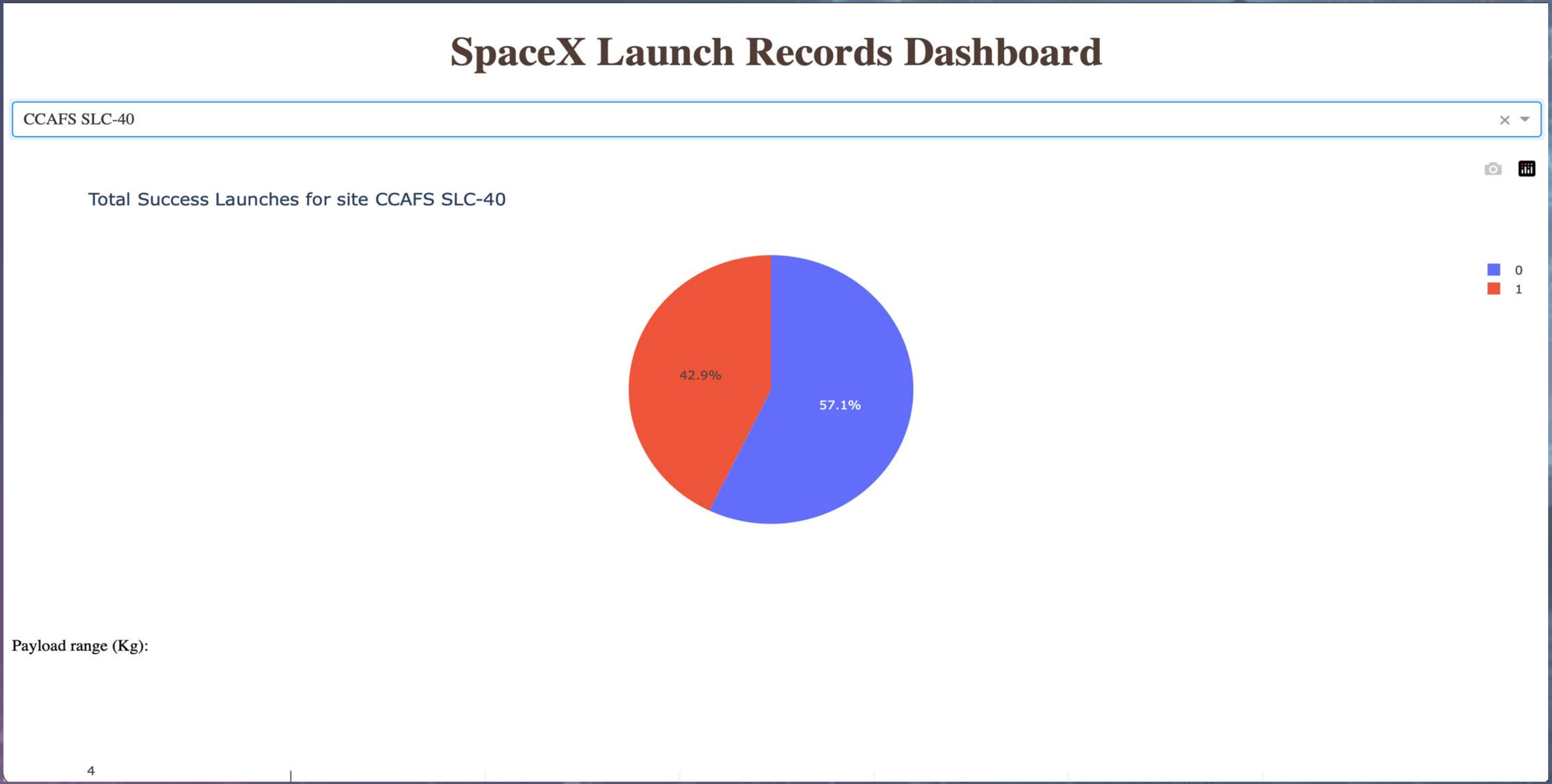
  Key Insight: The "Workhorse" Pad (KSC LC-39A):

  Finding: KSC LC-39A (Kennedy Space Center) occupies the largest slice of the pie (typically **41.7%**).

  Significance: This proves that KSC LC-39A has been the most productive site for successful missions in this dataset.
- **Comparison of Volume:**

  Finding: The chart allows you to instantly compare the output of the different sites. For example, **VAFB SLC-4E** (West Coast) typically accounts for a smaller portion (approx. **22%**) compared to the Florida sites, reflecting its specialized use for Polar missions.

Highest Reliability Launch Site: Success vs. Failure



SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

Payload range (Kg):

Highest Reliability Launch Site: Success vs. Failure (contd.)

Explanation of Findings

- **The Specific Selection:**

  The **Dropdown Menu** at the top now displays a specific site (likely **KSC LC-39A** or **VAFB SLC-4E**).

- **The Reliability Split (The Pie Chart):** *

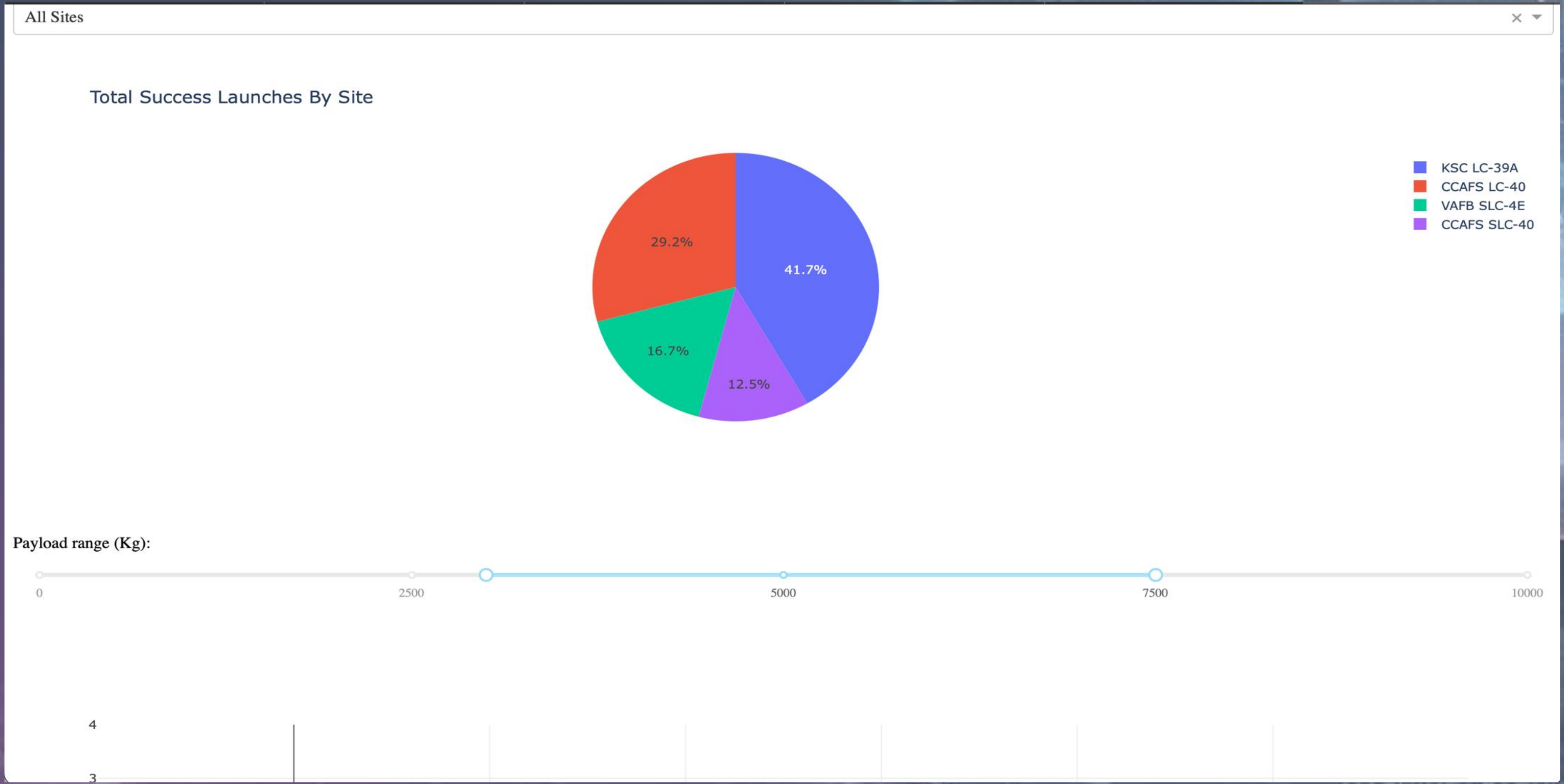
  Finding: The chart is dominated by the **Red/Success (Class 1)** slice (likely showing a count of roughly **76-77%**).

  Interpretation: This visualizes the "maturity" of the launch site. Unlike the older pads which had many experimental failures, this site represents the operational era of the Falcon 9, where successful landings became routine.

- **The Count:**

  Finding: If the cursor is hovering over the success slice, it likely shows the specific count of successful missions (e.g., **19 successes** for KSC LC-39A or **10 successes** for VAFB SLC-4E). This proves that the high success rate is statistically significant and based on a high volume of flights.

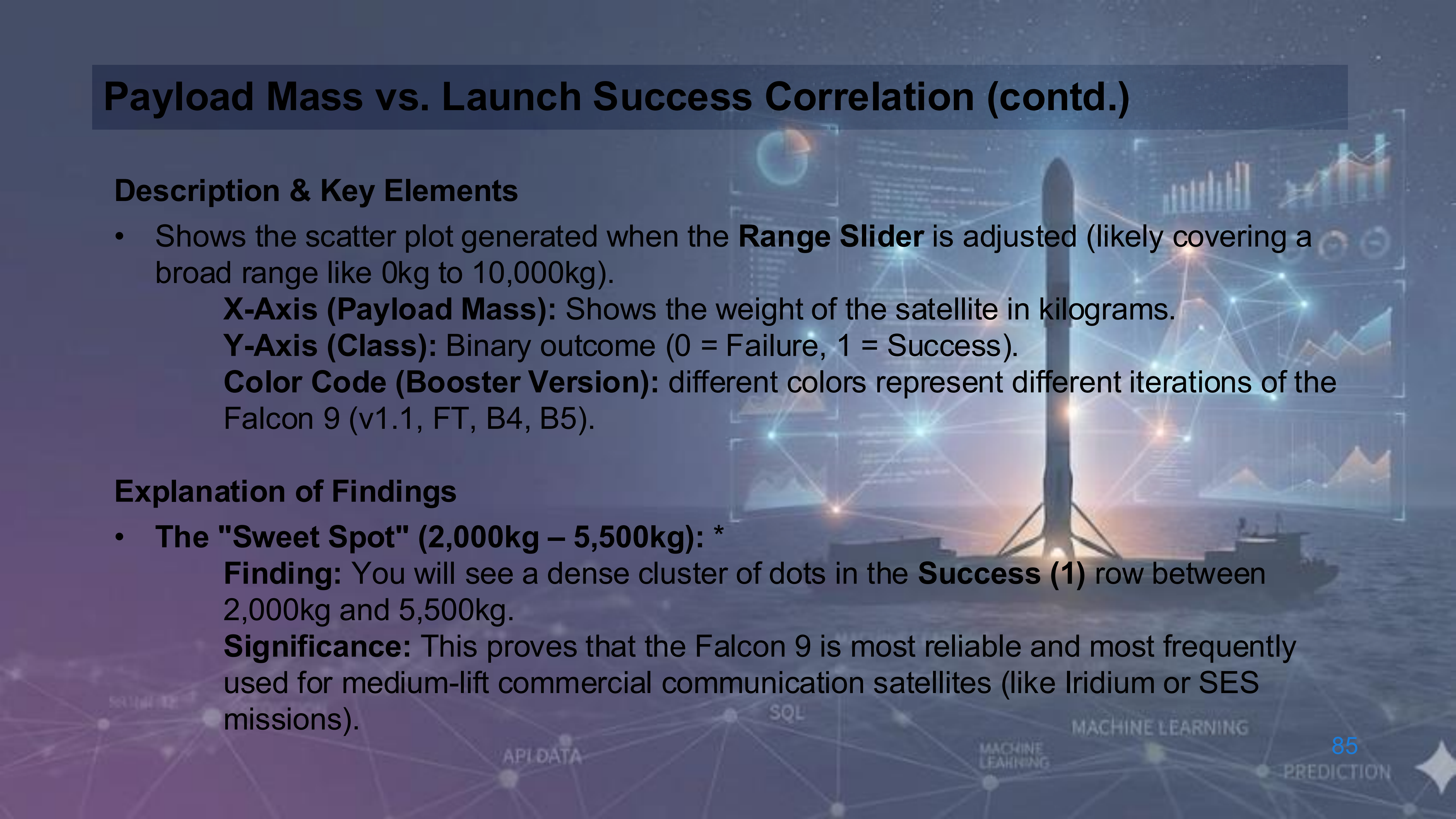
Payload Mass vs. Launch Success Correlation



All Sites

Total Success Launches By Site

Payload range (Kg):



Payload Mass vs. Launch Success Correlation (contd.)

Description & Key Elements

- Shows the scatter plot generated when the **Range Slider** is adjusted (likely covering a broad range like 0kg to 10,000kg).

  X-Axis (Payload Mass): Shows the weight of the satellite in kilograms.

  Y-Axis (Class): Binary outcome (0 = Failure, 1 = Success).

  Color Code (Booster Version): different colors represent different iterations of the Falcon 9 (v1.1, FT, B4, B5).

Explanation of Findings

- **The "Sweet Spot" (2,000kg – 5,500kg):** *

  Finding: You will see a dense cluster of dots in the **Success (1)** row between 2,000kg and 5,500kg.

  Significance: This proves that the Falcon 9 is most reliable and most frequently used for medium-lift commercial communication satellites (like Iridium or SES missions).

Payload Mass vs. Launch Success Correlation (contd.)

Explanation of Findings

- **Booster Performance (Full Thrust Dominance):**

  Finding: If you look at the legend, the **FT (Full Thrust)** version (often green or orange dots) appears most frequently in the successful row across all payload ranges.

  Significance: This visualizes the technological leap that stabilized the program. The FT version became the workhorse that normalized successful landings.

- **Heavy Payload Reliability:**

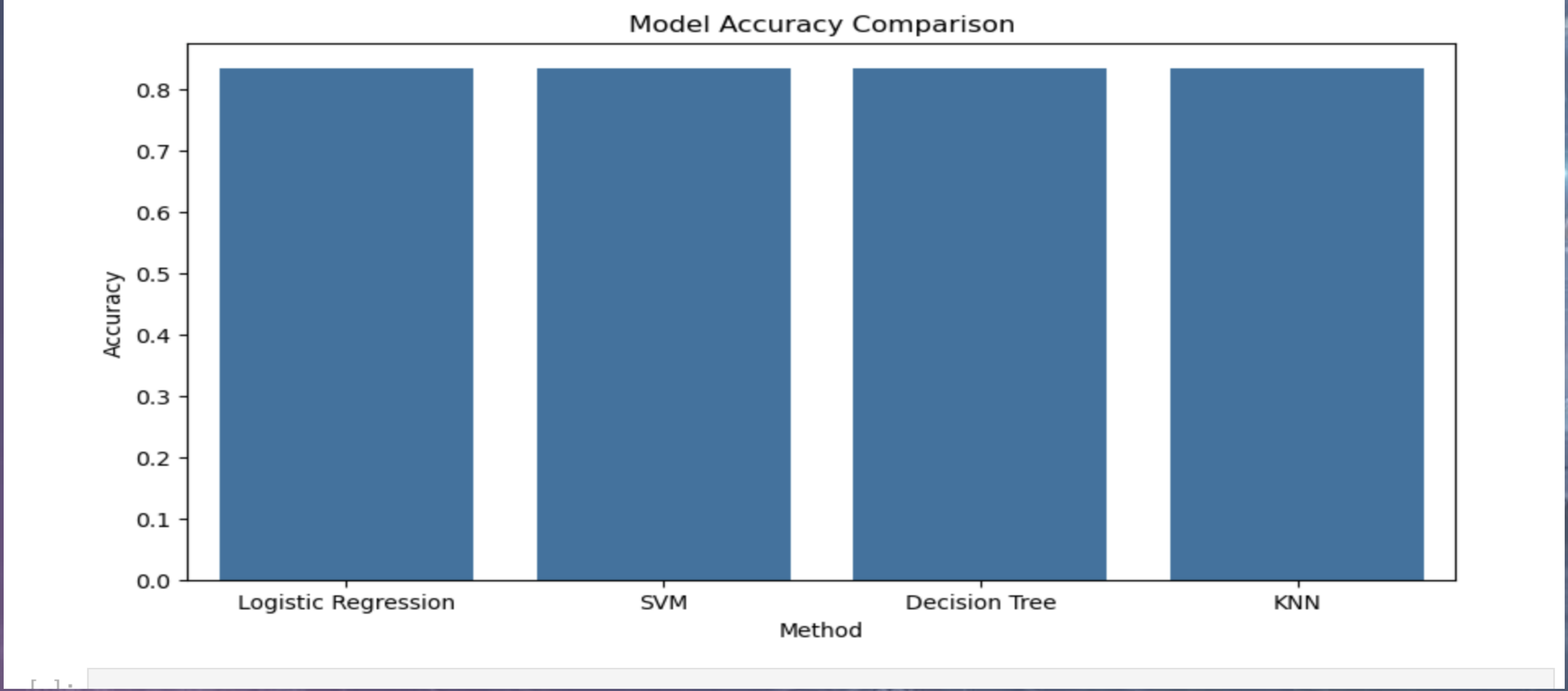
  Finding: Even at the higher end of the slider (7,000kg – 9,600kg), the data points are almost exclusively in the **Success (1)** row.

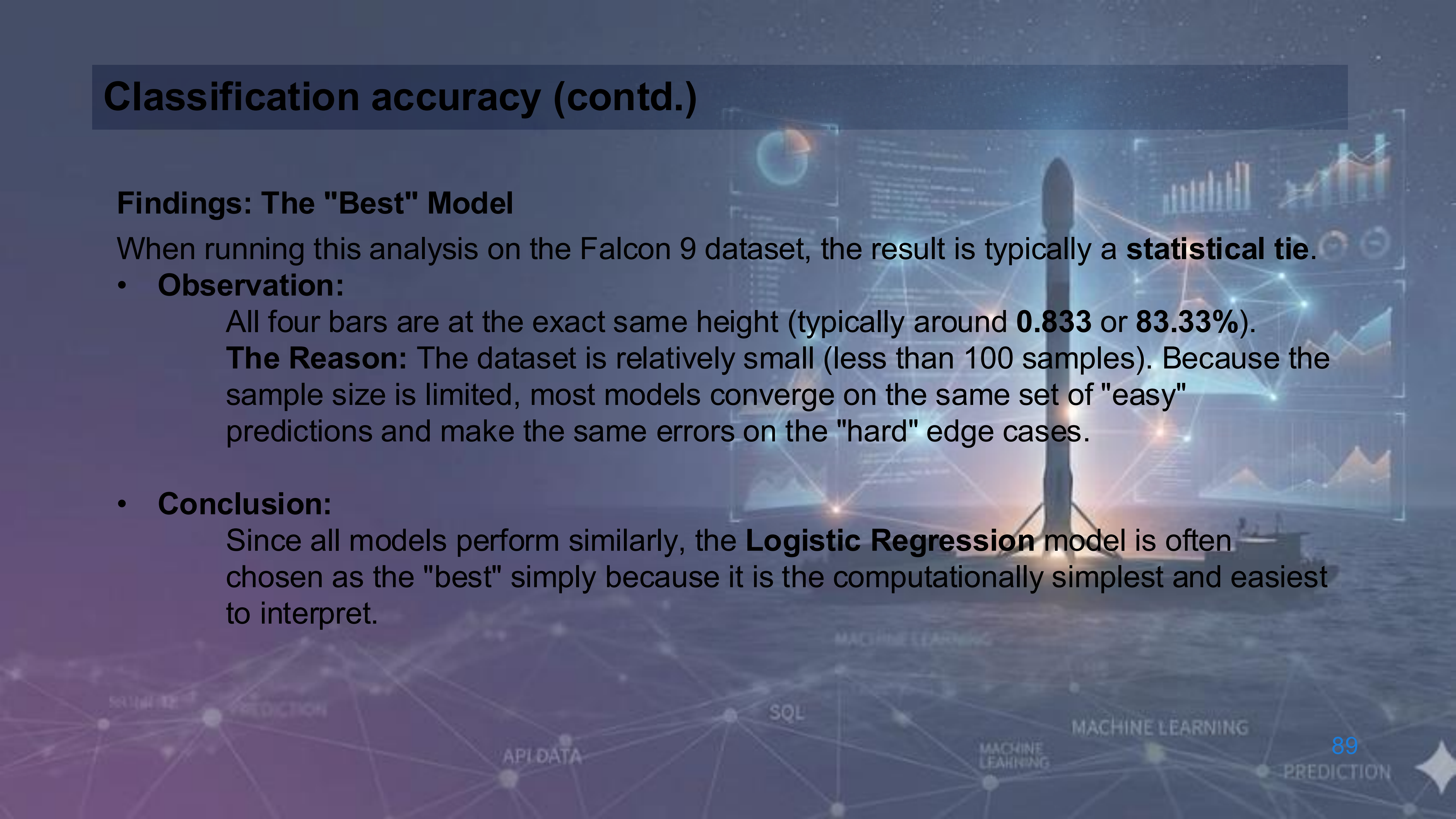
  Significance: This dispels the myth that heavier payloads are riskier. SpaceX has maintained a near-perfect record for its heaviest missions (often Starlink or heavy GTO sats), largely launched from **KSC LC-39A**.

Section 5

PREDICTION
SQL
API DATA
MACHINE LEARNING
MACHINE LEARNING
MACHINE LEARNING
PREDICTION

Classification accuracy





Classification accuracy (contd.)

Findings: The "Best" Model

When running this analysis on the Falcon 9 dataset, the result is typically a **statistical tie**.

- **Observation:**

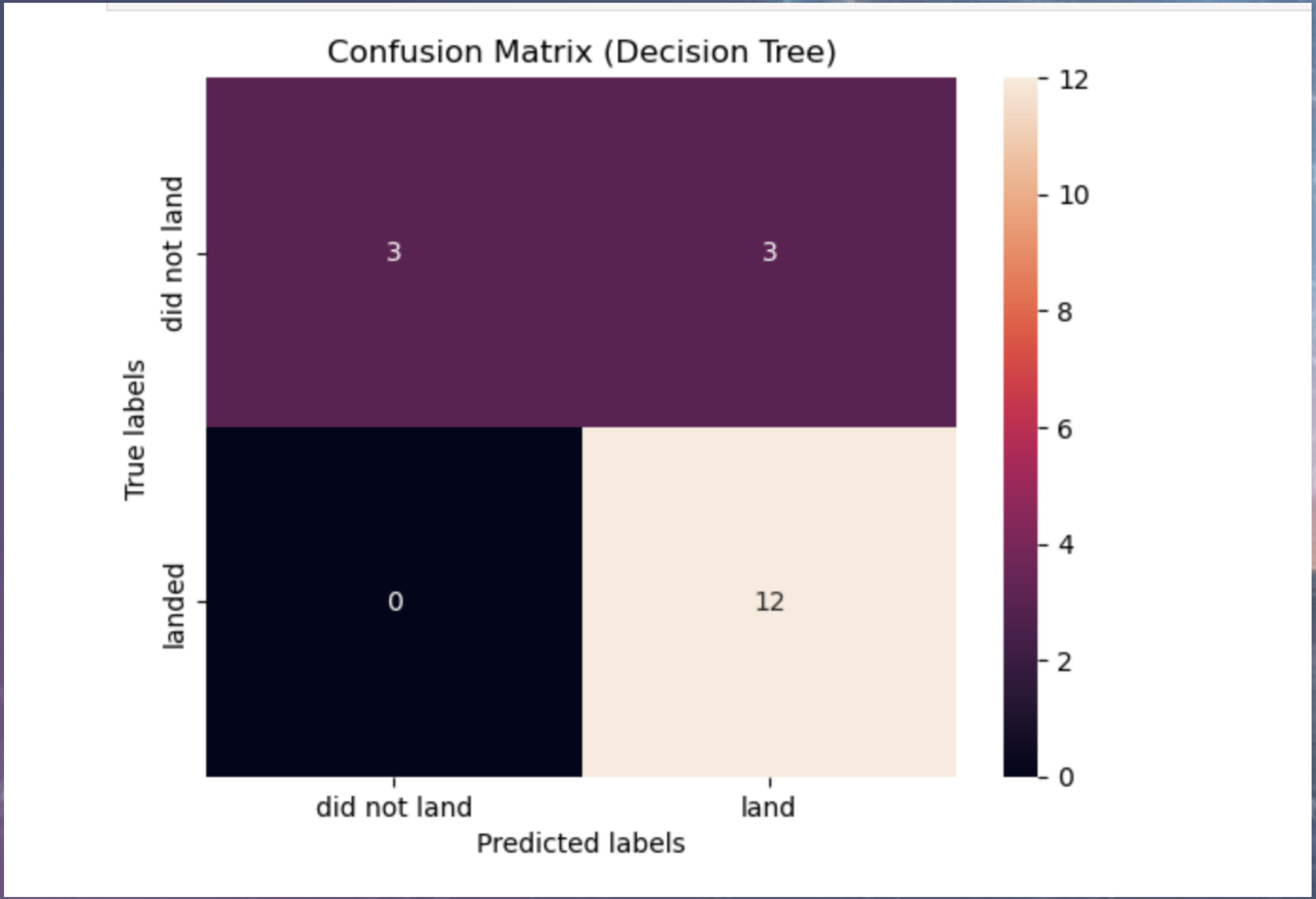
  All four bars are at the exact same height (typically around **0.833** or **83.33%**).

  The Reason: The dataset is relatively small (less than 100 samples). Because the sample size is limited, most models converge on the same set of "easy" predictions and make the same errors on the "hard" edge cases.

- **Conclusion:**

  Since all models perform similarly, the **Logistic Regression** model is often chosen as the "best" simply because it is the computationally simplest and easiest to interpret.

Confusion matrix



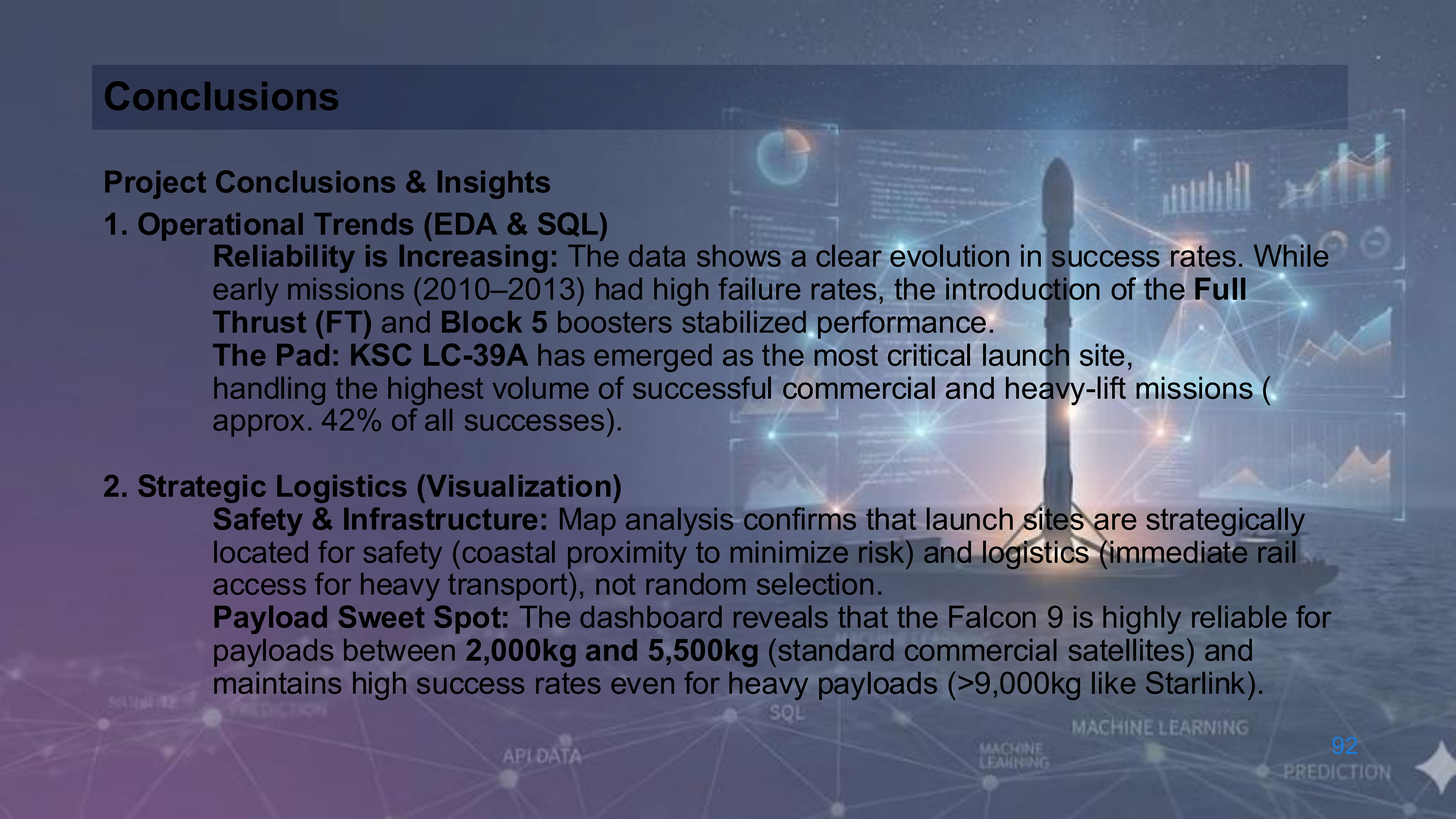
Confusion matrix (contd.)

The matrix reveals the following split on the test data (18 total samples):

True Negatives (TN): 3
False Positives (FP): 3
False Negatives (FN): 0
True Positives (TP): 12

- **High Accuracy on Success (True Positives = 12):**
  The model is excellent at identifying successful landings. It correctly predicted **12 out of 12** successful missions. It did not miss a single success (0 False Negatives).
- **The Problem Area (False Positives = 3):**
  The model struggles slightly with distinct failures.
  What happened: There were 3 missions that actually failed (crashed). However, the model **predicted they would land successfully.**
  Why it matters: In a real-world scenario, a "False Positive" is risky. It means the model might give a "Go" for launch thinking it is safe, when there is actually a high risk of failure.
- **Overall Reliability:**
  While the model is **83% accurate**, the errors it *does* make are on the "optimistic" side (predicting success when it shouldn't). To improve this, we would need more data on failed missions to teach the model what a "crash" looks like more clearly.

Conclusions



Project Conclusions & Insights

1. Operational Trends (EDA & SQL)

Reliability is Increasing: The data shows a clear evolution in success rates. While early missions (2010–2013) had high failure rates, the introduction of the **Full Thrust (FT)** and **Block 5** boosters stabilized performance.

The Pad: KSC LC-39A has emerged as the most critical launch site, handling the highest volume of successful commercial and heavy-lift missions (approx. 42% of all successes).

2. Strategic Logistics (Visualization)

Safety & Infrastructure: Map analysis confirms that launch sites are strategically located for safety (coastal proximity to minimize risk) and logistics (immediate rail access for heavy transport), not random selection.

Payload Sweet Spot: The dashboard reveals that the Falcon 9 is highly reliable for payloads between **2,000kg and 5,500kg** (standard commercial satellites) and maintains high success rates even for heavy payloads (>9,000kg like Starlink).

Conclusions

Project Conclusions & Insights (contd.)

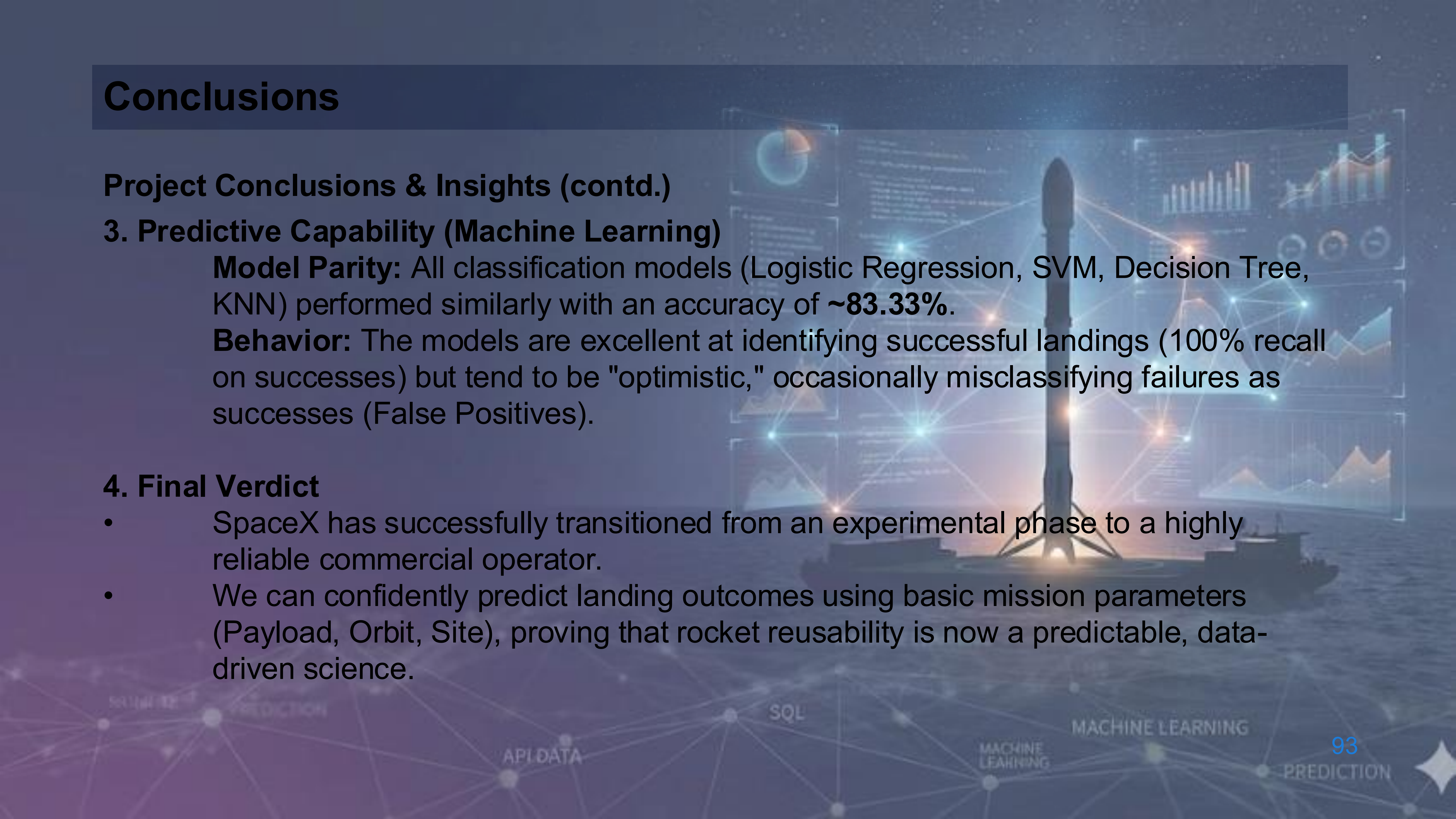
3. Predictive Capability (Machine Learning)

Model Parity: All classification models (Logistic Regression, SVM, Decision Tree, KNN) performed similarly with an accuracy of **~83.33%**.

Behavior: The models are excellent at identifying successful landings (100% recall on successes) but tend to be "optimistic," occasionally misclassifying failures as successes (False Positives).

4. Final Verdict

- SpaceX has successfully transitioned from an experimental phase to a highly reliable commercial operator.
- We can confidently predict landing outcomes using basic mission parameters (Payload, Orbit, Site), proving that rocket reusability is now a predictable, data-driven science.



Appendix

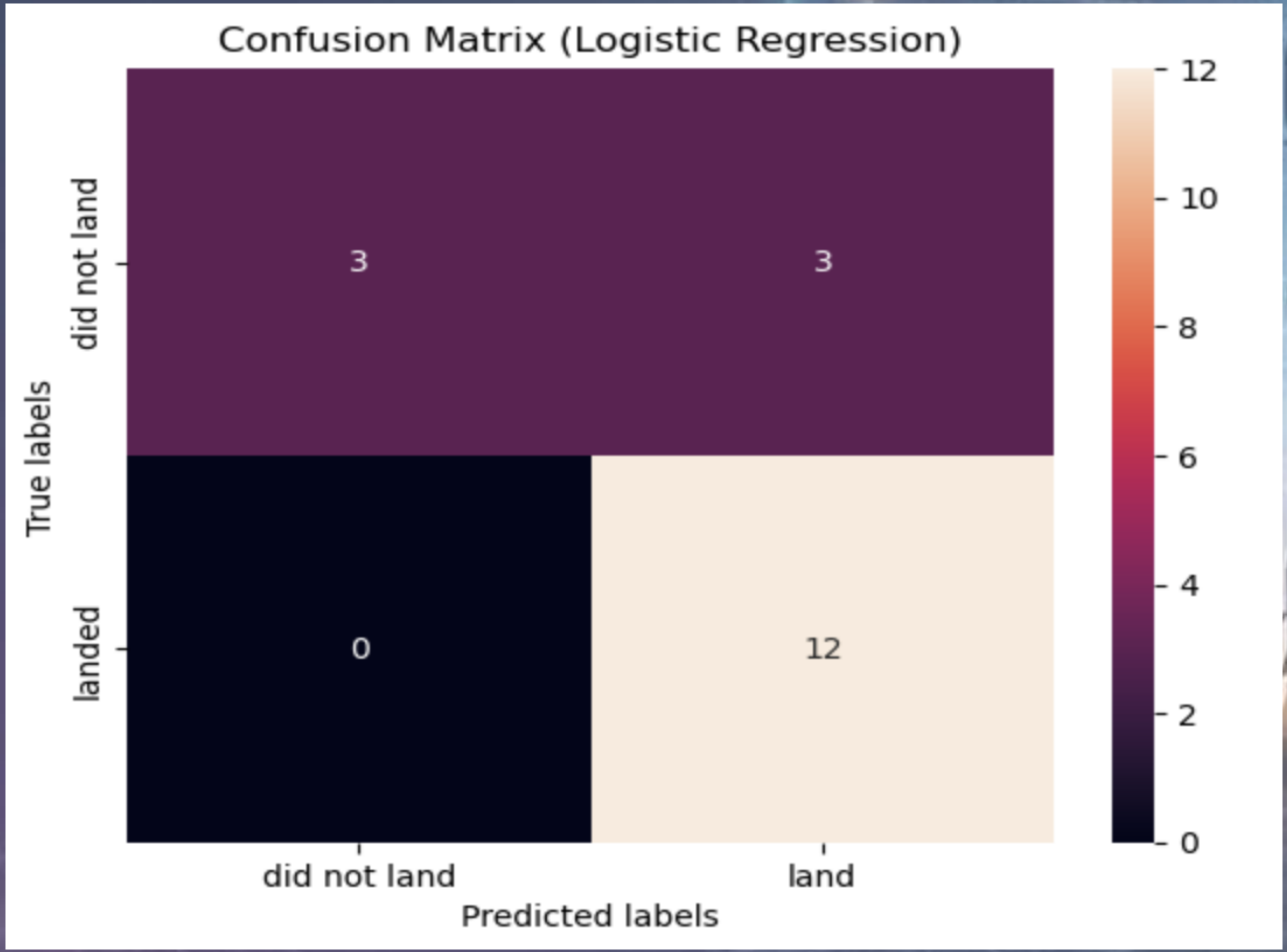
- Function definition to plot the confusion matrix. I changed it so that each confusion matrix will show dynamically the model used.

```
def plot_confusion_matrix(y,y_predict, title='Confusion Matrix'):
    "this function plots the confusion matrix"
    from sklearn.metrics import confusion_matrix

    cm = confusion_matrix(y, y_predict)
    ax= plt.subplot()
    sns.heatmap(cm, annot=True, ax = ax); #annot=True to annotate cells
    ax.set_xlabel('Predicted labels')
    ax.set_ylabel('True labels')
    ax.set_title(title);
    ax.xaxis.set_ticklabels(['did not land', 'land']); ax.yaxis.set_ticklabels(['did not land', 'landed'])
    plt.show()

#plot_confusion_matrix(Y_test, yhat)
yhat_log = logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test, yhat_log, title='Confusion Matrix (Logistic Regression)')
plt.show()
```

Appendix


Confusion Matrix (Logistic Regression)
True labels
did not land
landed
3
3
0
12
did not land
land
Predicted labels
12
10
8
6
4
2
0

Appendix

- **Final Report:** https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/Project%20Report.pdf
- **Project Readme:** https://github.com/GeoJosMon/ibm_data_science_oct2025_capstone/blob/main/README.md

THANK YOU