IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nikolas Daniel Vincenti
25/09/2025

Outline

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

Executive Summary

- **Methodologies:**
- Data collection via SpaceX API and Wikipedia web scraping
- SQL-based exploratory analysis with 10+ complex queries
- Interactive visualization using Folium maps and Plotly Dash
- Machine learning classification with 4 algorithms: Logistic Regression, SVM, Decision Tree, and KNN
- **Key Results:**
- Achieved 83.33% prediction accuracy using Logistic Regression, SVM, and KNN
- Success rates improved from 2013-2020, showing clear learning curve
- CCAFS SLC-40 is the most active launch site with highest success volume
- LEO and ISS orbits show highest landing success rates
- Payload mass 4000-6000kg range optimal for drone ship landings

Introduction

- **Project Background:** SpaceX revolutionized space industry through reusable rocket technology, reducing launch costs from $165M+ (competitors) to $62M. Landing success directly impacts cost competitiveness and market position.

- **Business Problem:** Predict Falcon 9 first stage landing success to enable accurate cost estimation for competitive analysis. Successful landings enable rocket reuse, creating significant cost advantages in commercial space market.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API (94 launches), Wikipedia scraping, JSON/HTML parsing
- Perform data wrangling
  - Missing value imputation, binary labels, data cleaning, feature selection
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - GridSearchCV, 4 algorithms, train-test split, accuracy comparison

Data Collection

- **Data Sources:**
- **SpaceX API:** 94 historical launch records with comprehensive mission details
- **Wikipedia Scraping:** Additional launch data from "List of Falcon 9 and Falcon Heavy launches"
- **Time Period:** 2010-2020 launch history
- **Data Collection Process:**
- API requests to SpaceX endpoints for launches, rockets, payloads, and cores
- Web scraping with BeautifulSoup for HTML table extraction
- Data validation and cleaning for consistency
- Integration of multiple data sources into unified dataset

REPO: https://github.com/NikVince/ibm-data-scientist-professional-certificate-2025/tree/main/modules/course-10-applied-data-science-capstone/completed_assignements

Data Collection – SpaceX API

- **API Data Extraction Process:**
- **GET Request** → SpaceX REST API endpoints
- **Parse JSON** → Extract launch, rocket, payload, core data
- **Helper Functions** → getBoosterVersion(), getLaunchSite(), getPayloadData(), getCoreData()
- **Data Integration** → Combine related information using rocket/payload/core IDs
- **Export** → Clean structured dataset (dataset_part_1.csv)
- **Key Features Extracted:** Flight number, date, booster version, payload mass, orbit, launch site, landing outcome, grid fins, reuse status, core serial numbers

Data Collection - Scraping

- **Web Scraping Process:**
- **Target URL** → Wikipedia Falcon 9 launches page (static snapshot)
- **HTML Parsing** → BeautifulSoup4 to extract table data
- **Data Cleaning** → Remove references, normalize Unicode, handle missing values
- **Feature Extraction** → Date/time, booster versions, launch sites, payloads, outcomes
- **Export** → Supplemental dataset (spacex_web_scraped.csv)
- **Technical Challenges:** Complex HTML structure, Wikipedia reference links, inconsistent formatting, multiple nested tables

Data Wrangling

- **Data Processing Steps:**
- **Missing Value Analysis** → Identified gaps in PayloadMass and LandingPad columns
- **Data Imputation** → Mean substitution for PayloadMass missing values
- **Binary Classification** → Success (1): True ASDS/RTLS/Ocean; Failure (0): False/None outcomes
- **Feature Engineering** → Created Class variable for machine learning
- **Data Validation** → 90 records with consistent formatting
- **Output:** Clean dataset with binary target variable ready for analysis (dataset_part_2.csv)

EDA with Data Visualization

- **Visualization Analysis:**
- **Flight Number vs Launch Site:** Scatter plot showing success improvement over time
- **Payload Mass vs Launch Site:** VAFB SLC-4E limited to lighter payloads (<10,000kg)
- **Success Rate by Orbit:** Bar chart showing LEO, ISS, VLEO highest success rates
- **Yearly Trend Analysis:** Line plot demonstrating continuous improvement 2013-2020
- **Feature Engineering:** One-hot encoding for categorical variables (Orbit, LaunchSite, LandingPad, Serial)
- **Key Insights:** Clear learning curve effect, site specialization by payload, orbit-specific success patterns

EDA with SQL

- **Key SQL Queries Performed:**
- **Launch Site Analysis:** SELECT DISTINCT "Launch_Site" → 3 unique sites identified
- **CCA Site Records:** SELECT * WHERE "Launch_Site" LIKE 'CCA%' → CCAFS SLC-40 dominance
- **NASA Payload Mass:** SUM(CAST("Payload_Mass__kg_" AS INTEGER)) for NASA (CRS)
- **F9 v1.1 Performance:** AVG("Payload_Mass__kg_") by booster version
- **Landing Milestones:** MIN("Date") for first successful ground pad landing
- **Success Rates:** Mission outcome distribution and landing outcome rankings
- **Performance Analysis:** Booster versions with maximum payload capacity

Build an Interactive Map with Folium

- **Interactive Map Features:**
- **Launch Site Markers:** Circles (1000m radius) and custom icons for CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A
- **Success/Failure Visualization:** Green markers (successful landings), red markers (failed landings) with clustering
- **Distance Analysis:** Calculated proximities to coastlines, railways, highways using Haversine formula
- **Geographic Insights:** All sites coastal (ocean access), infrastructure proximity, optimal launch trajectories
- **Key Finding:** CCAFS SLC-40 distance to coastline: 0.58km, enabling efficient ocean landing operations

Build a Dashboard with Plotly Dash

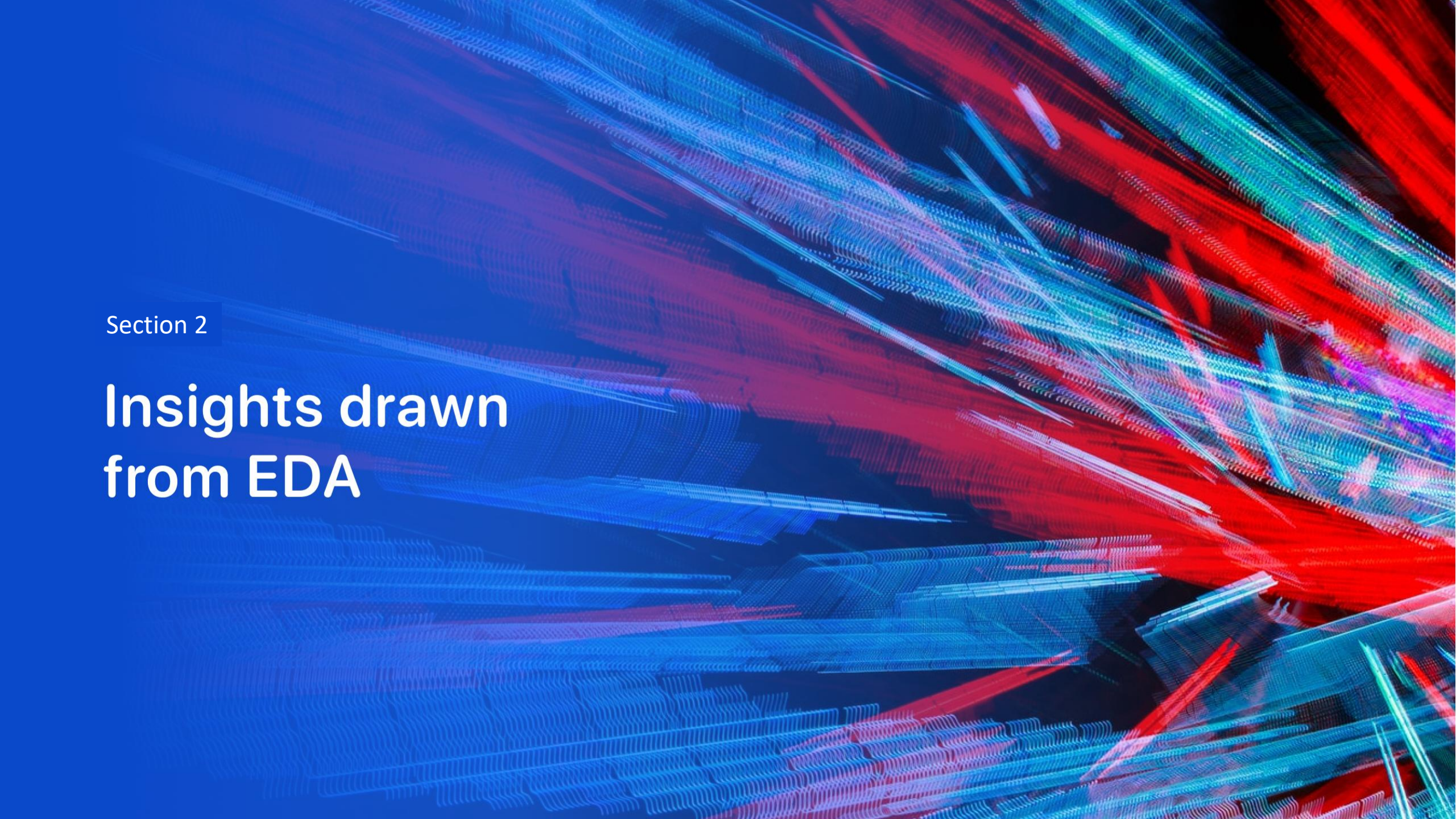
- **Dashboard Components:**
- **Interactive Pie Chart:** Launch success distribution across all sites
- **Site-Specific Analysis:** Dropdown filter for individual launch site performance
- **Payload Range Slider:** Dynamic filtering by payload mass (0-10,000kg)
- **Success Rate Visualization:** Real-time calculation of success ratios
- **Scatter Plot Integration:** Payload vs outcome correlation analysis
- **Interactivity:** Users can filter by launch site and payload range to identify optimal launch conditions and success patterns

Predictive Analysis (Classification)

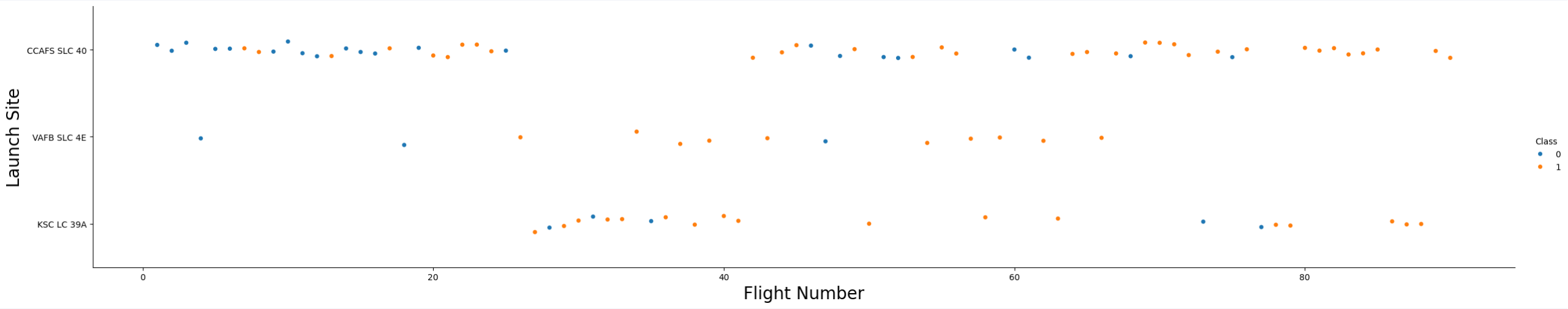
- **Model Development Process:**
- **Data Preparation:** StandardScaler normalization, 80/20 train-test split
- **Algorithm Selection:** Logistic Regression, SVM, Decision Tree, KNN
- **Hyperparameter Tuning:** GridSearchCV with 10-fold cross-validation
- **Model Evaluation:** Accuracy comparison and confusion matrix analysis
- **Best Model Selection:** Logistic Regression (interpretability + performance)
- **Optimal Parameters:** LogisticRegression(C=0.01, penalty='l2', solver='lbfgs')

Results

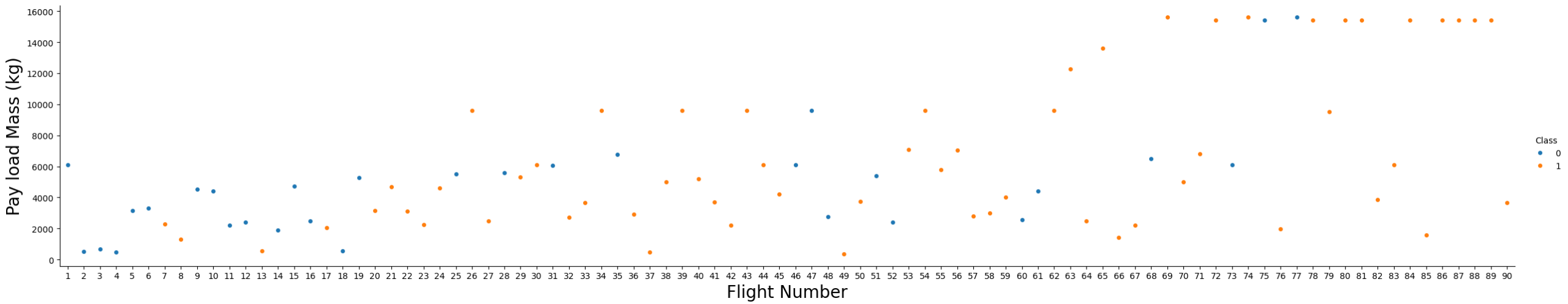
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

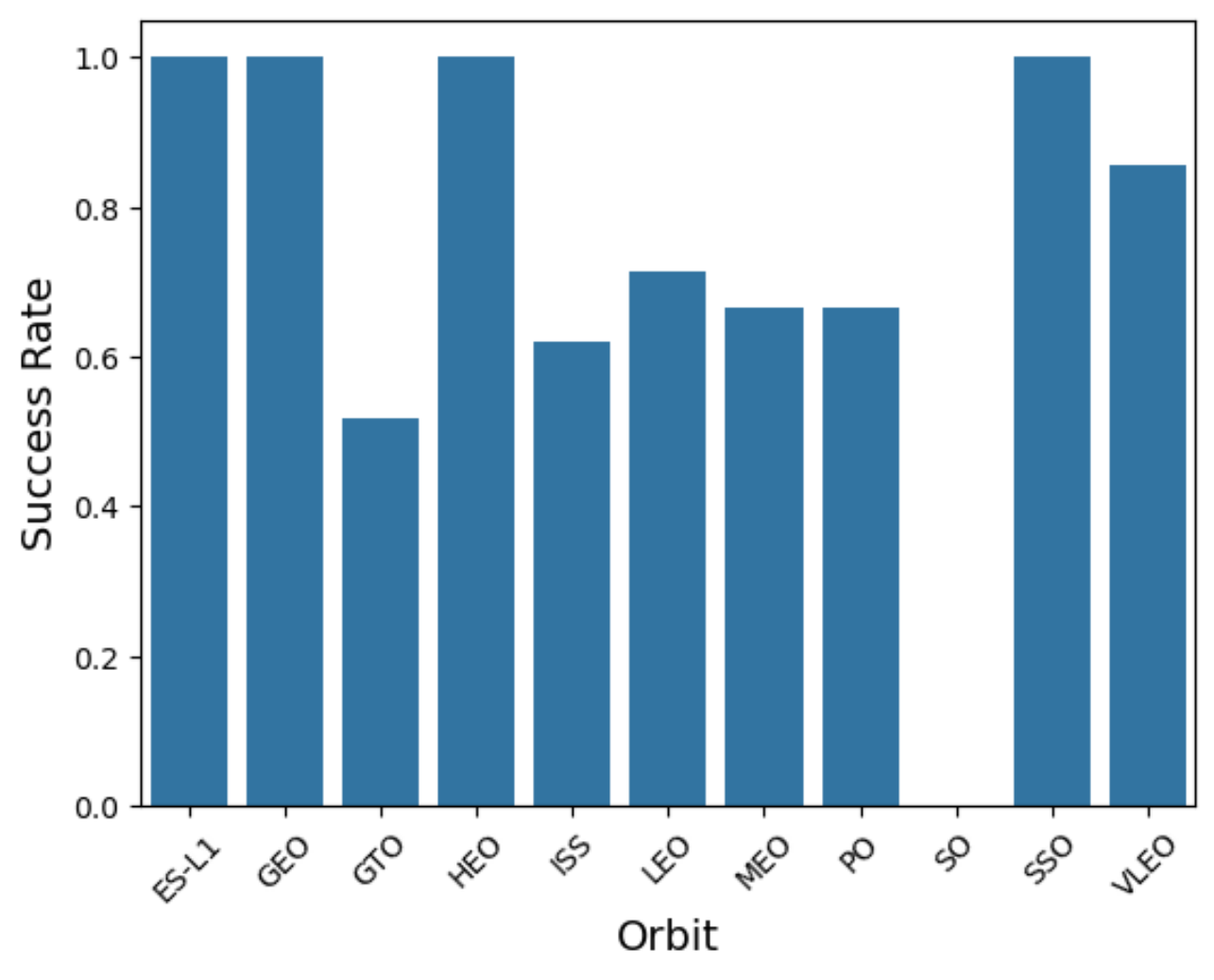
Flight Number vs. Launch Site



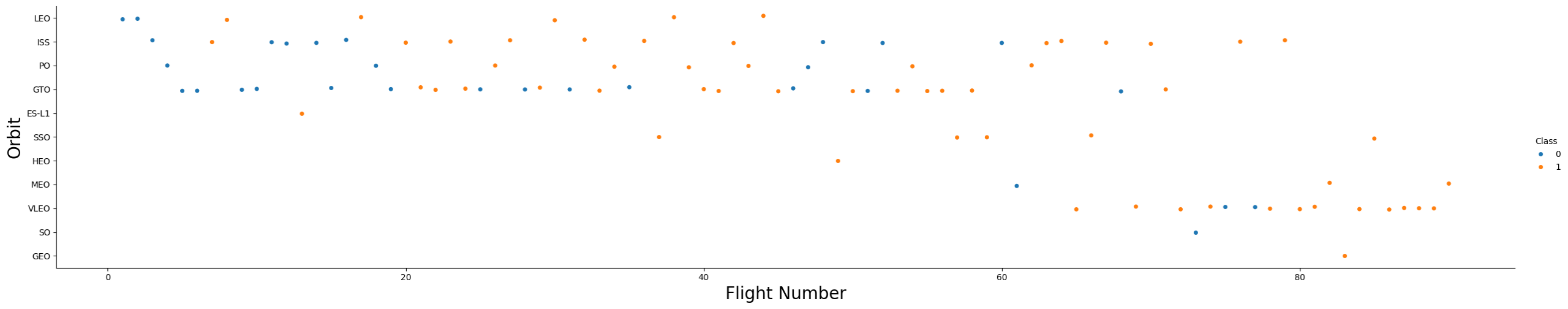
Payload vs. Launch Site



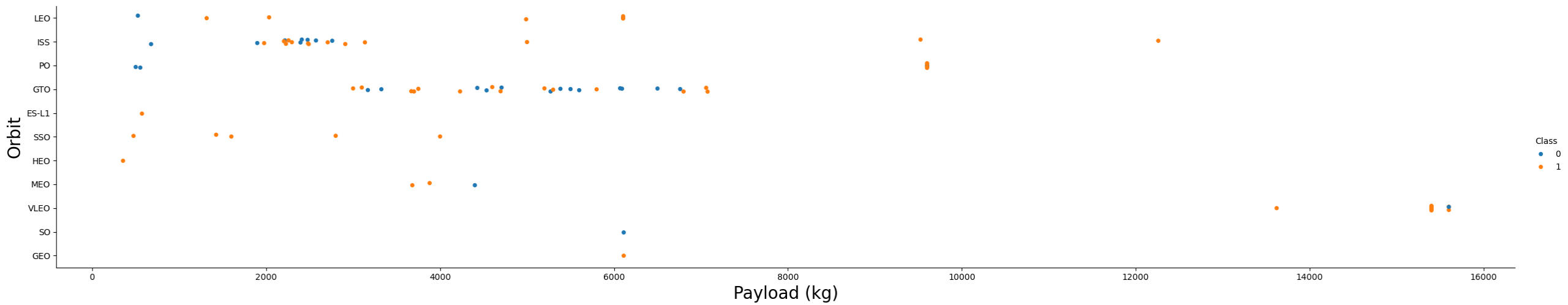
Success Rate vs. Orbit Type



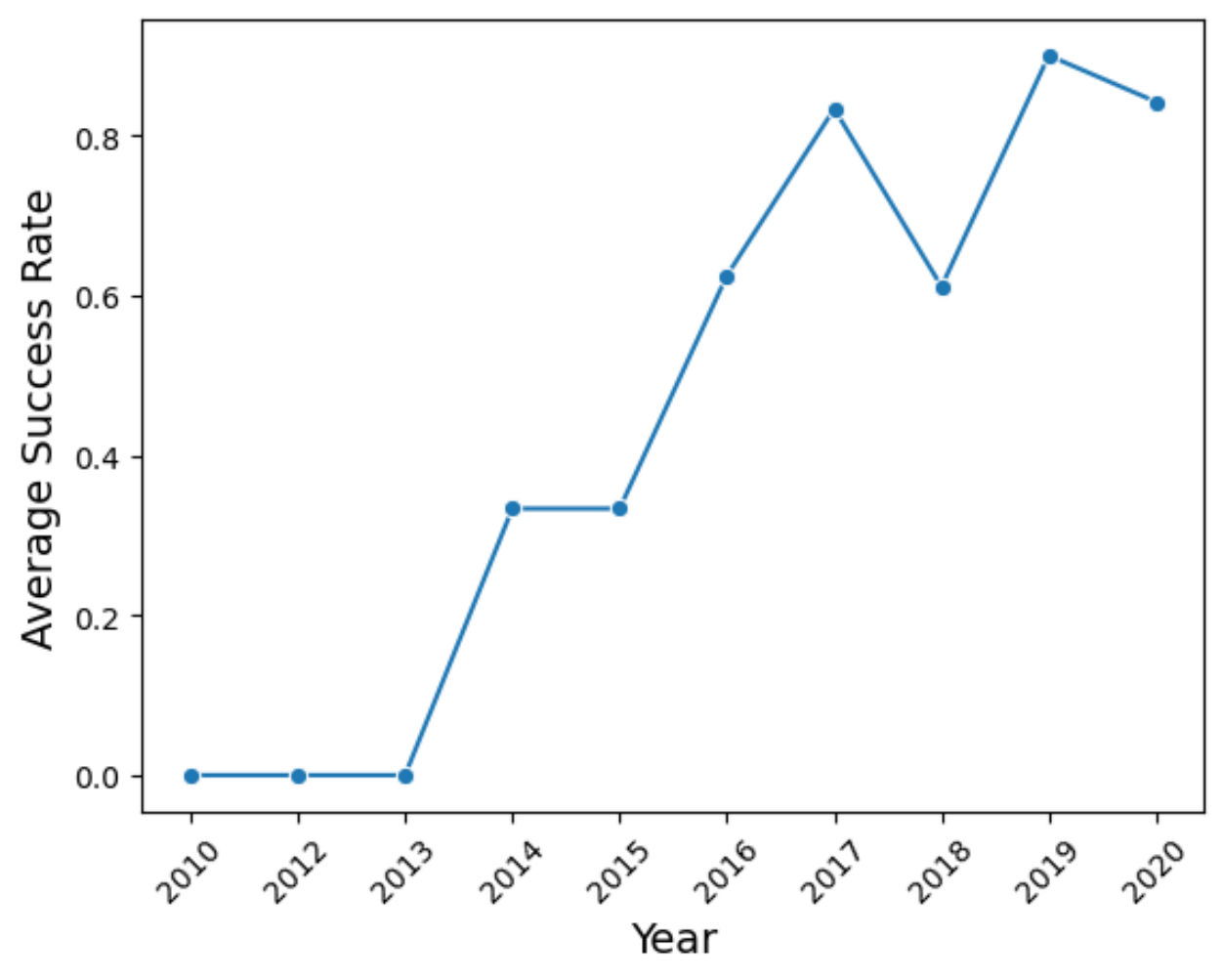
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

SQL Query: SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
Results:

- CCAFS SLC-40 (Cape Canaveral Space Force Station)
- VAFB SLC-4E (Vandenberg Air Force Base)
- KSC LC-39A (Kennedy Space Center)
- **Explanation:** Three operational launch sites, each serving specific mission types and orbital requirements

Launch Site Names Begin with 'CCA'

SQL Query: SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;

- **Results:** 5 sample records from CCAFS SLC-40
- **Finding:** CCAFS SLC-40 is the primary SpaceX launch facility, handling majority of Falcon 9 missions including commercial, NASA, and military payloads

Total Payload Mass

SQL Query: SELECT SUM(CAST("Payload_Mass__kg_" AS INTEGER))
FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)’;

- **Result:** NASA CRS missions total payload mass calculation
- **Business Value:** Quantifies SpaceX's cargo capacity contribution to NASA's Commercial Resupply Services program for ISS operations

Average Payload Mass by F9 v1.1

SQL Query: SELECT AVG(CAST("Payload_Mass__kg_" AS INTEGER)) FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1’;

- **Result:** F9 v1.1 average payload performance metrics
- **Technical Insight:** Demonstrates payload capacity evolution across booster versions, informing performance benchmarks

First Successful Ground Landing Date

SQL Query: SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)’;

- **Result:** Historical milestone date identification
- **Significance:** Marks breakthrough achievement in rocket reusability technology, revolutionary moment for space industry cost reduction

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query: SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "Payload_Mass__kg_" BETWEEN 4000 AND 6000;

- **Results:** Booster versions achieving optimal performance in medium-heavy payload range
- **Technical Finding:** Identifies sweet spot for drone ship landing success, balancing payload capacity with landing feasibility

Total Number of Successful and Failure Mission Outcomes

SQL Query: SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Mission_Outcome";

- **Results:** Mission success/failure distribution
- **Reliability Metrics:** Overall mission success rate calculation for business reliability assessment and customer confidence

Boosters Carried Maximum Payload

SQL Query: SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);

- **Results:** Booster versions with maximum payload capacity
- **Performance Analysis:** Identifies highest-performing rocket configurations for heavy-lift mission planning

2015 Launch Records

SQL Query: SELECT "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Date" LIKE '2015%' AND "Landing_Outcome" = 'Failure (drone ship)';

- **Results:** Failed drone ship landings in 2015 with booster and site details
- **Historical Context:** Documents early reusability program challenges, showing learning curve during technology development phase

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

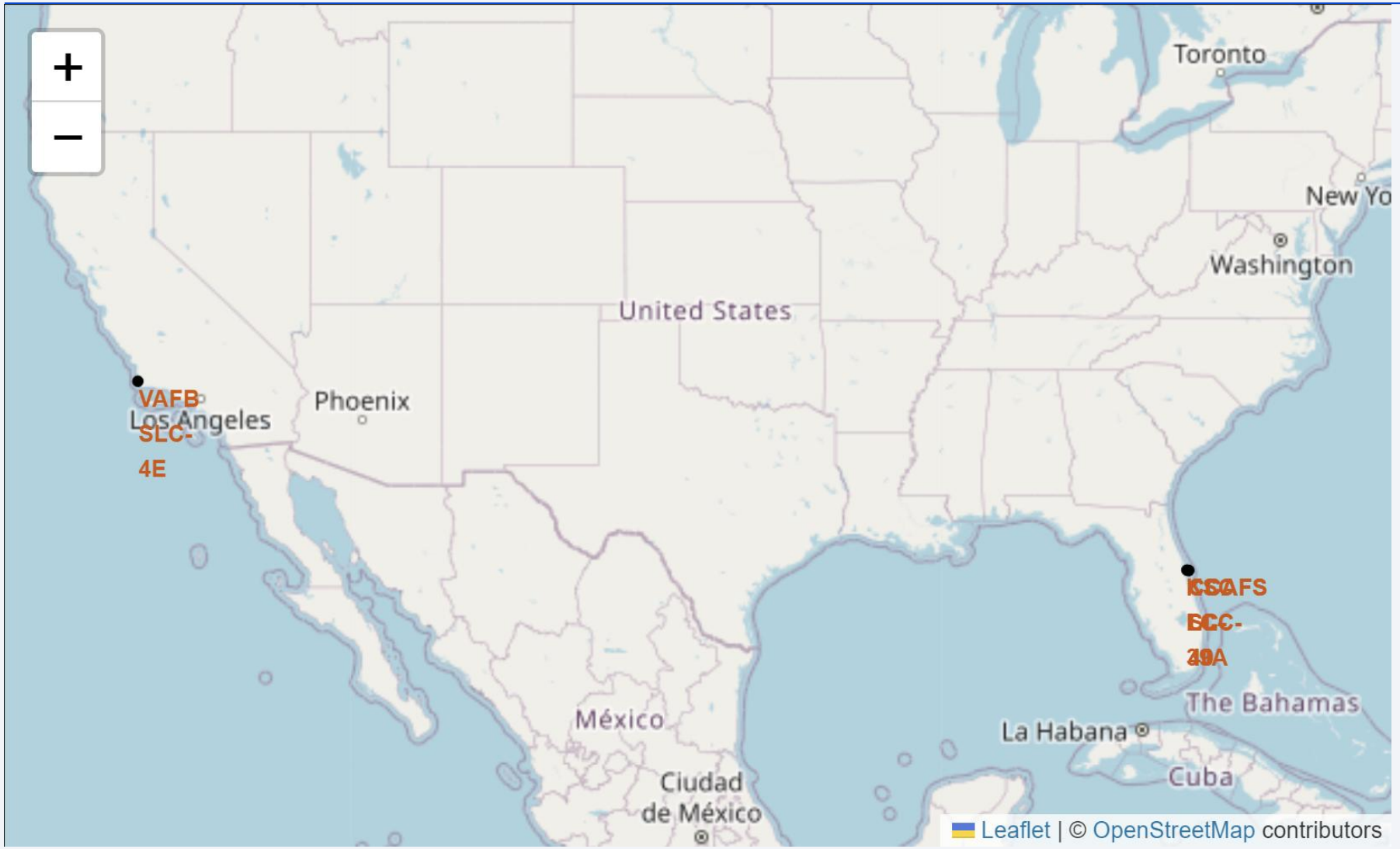
SQL Query: SELECT "Landing_Outcome", COUNT(*) FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY COUNT(*) DESC;

- **Results:** Landing outcome frequency ranking during early SpaceX period
- **Evolution Analysis:** Shows progression from failures to successes, quantifying technology improvement trajectory

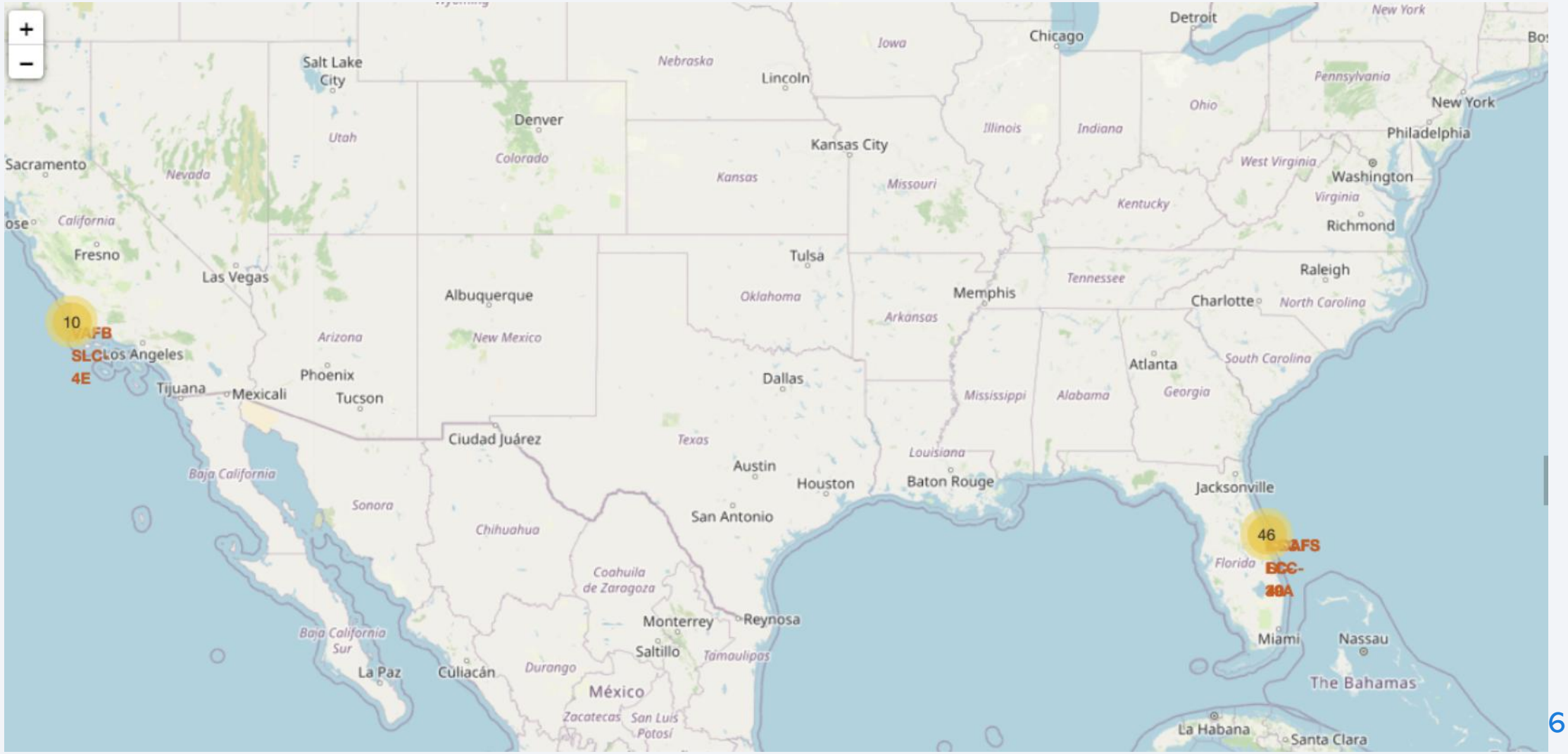
Section 3

Launch Sites Proximities Analysis

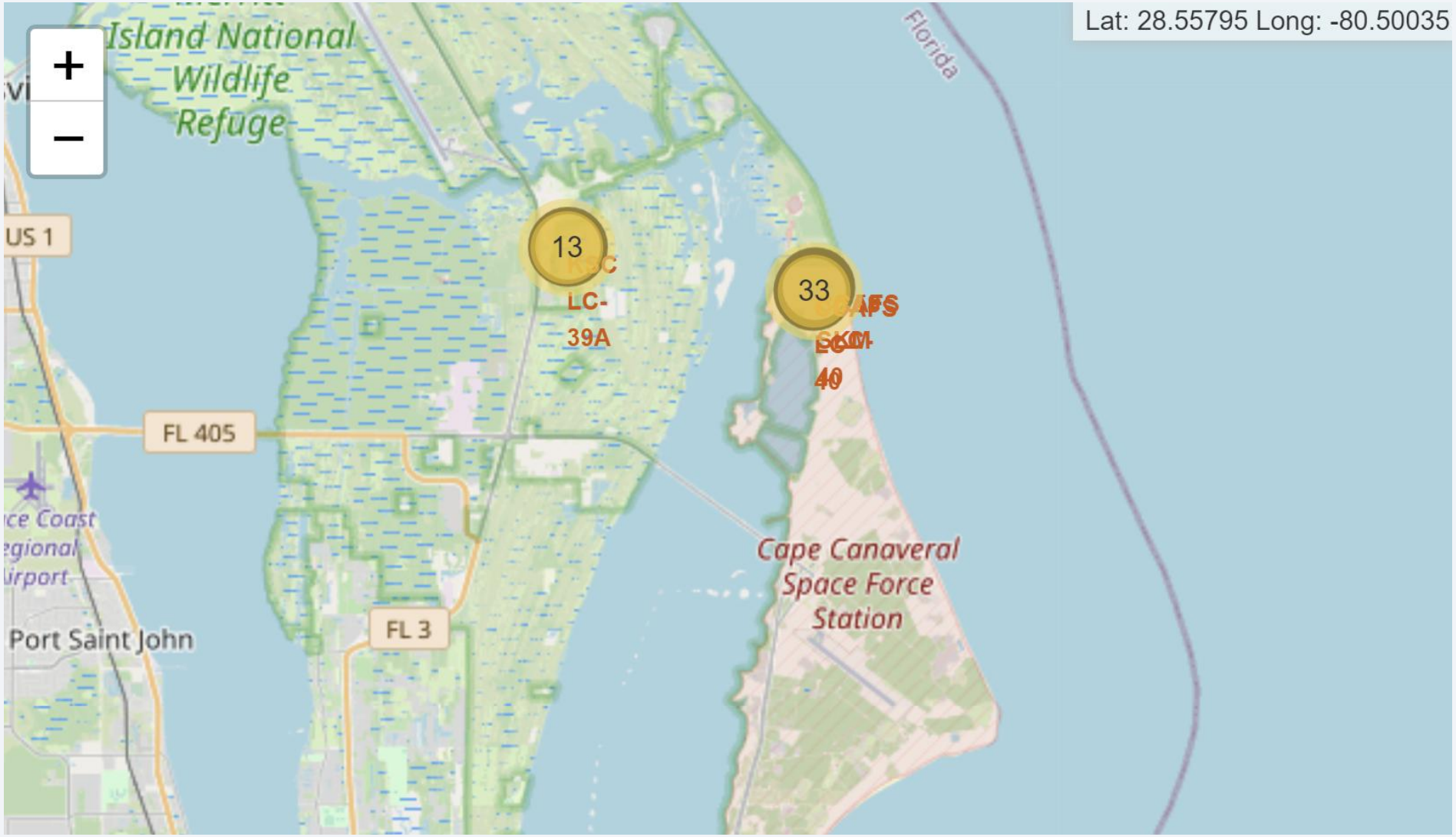
Starting Map

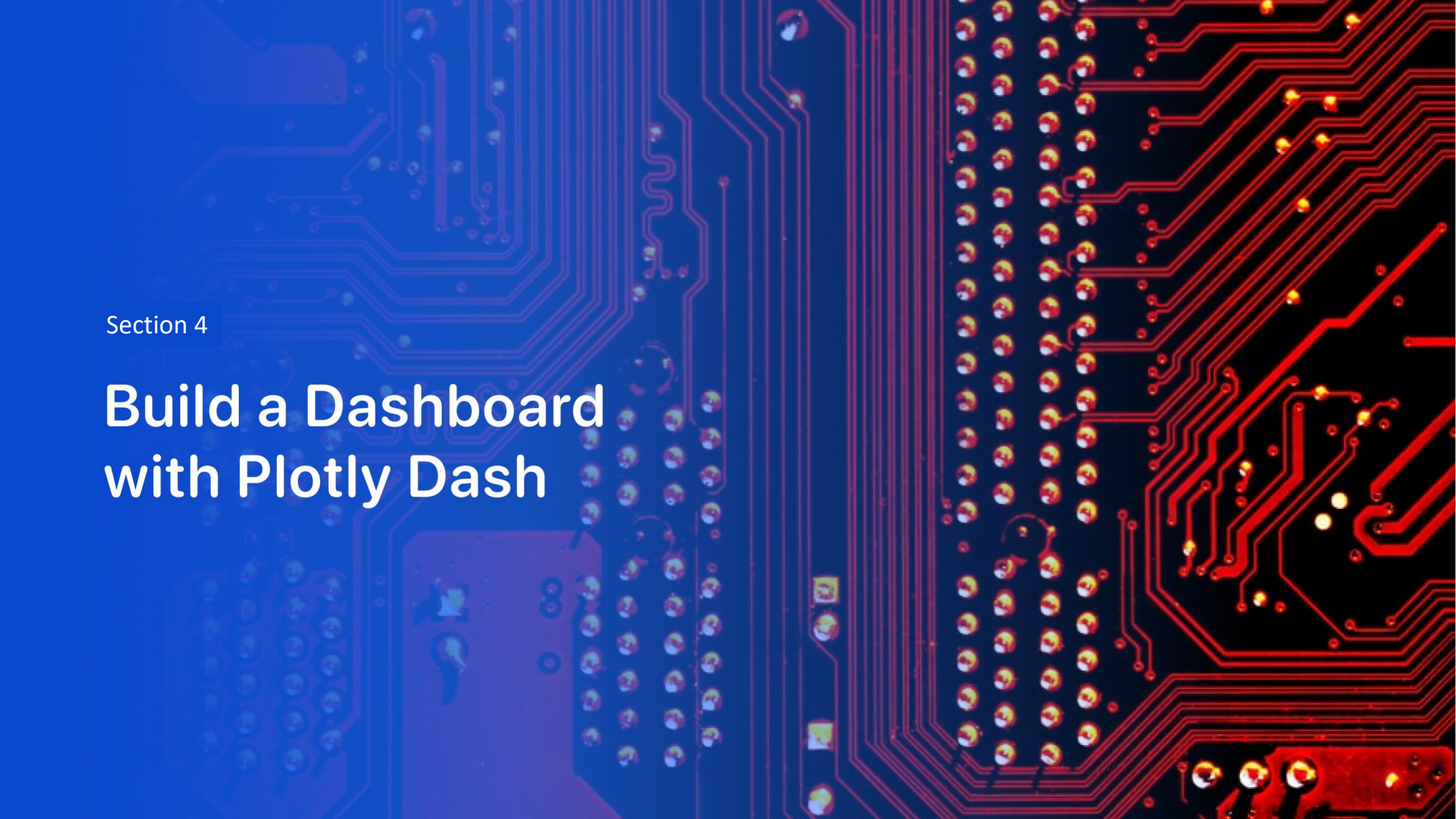


Map After Adding Main Location Markers

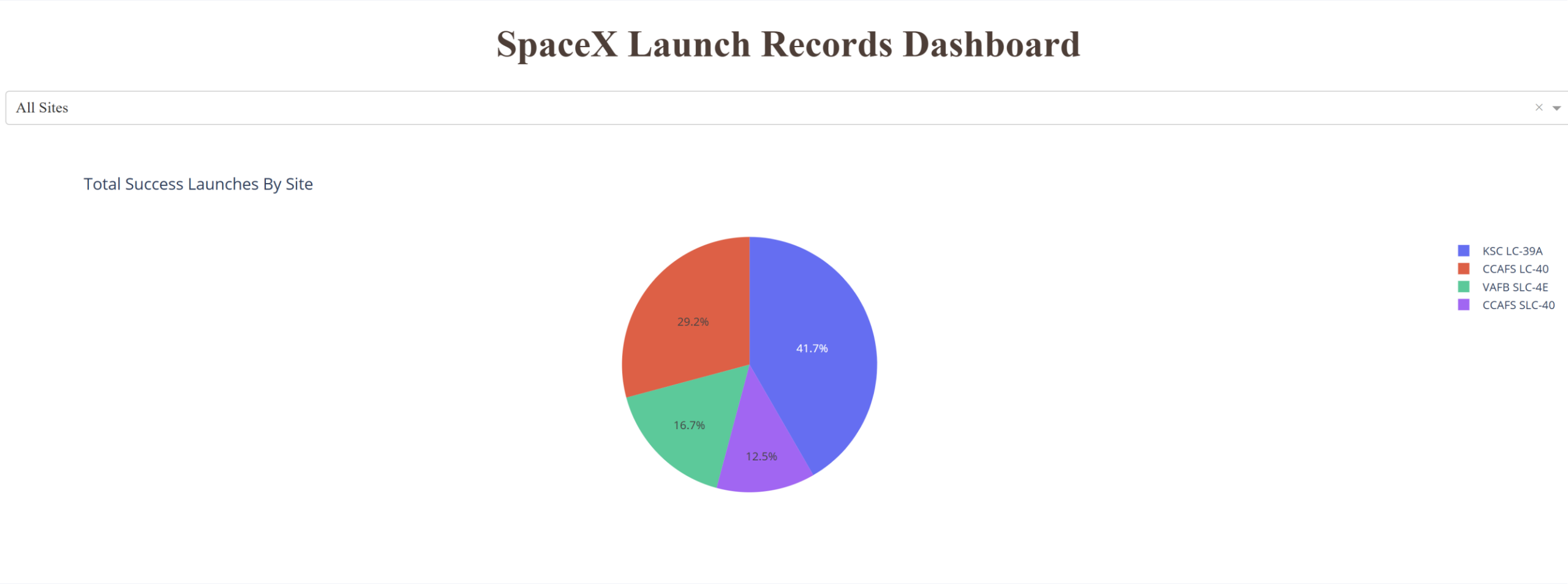


Zoomed in Map

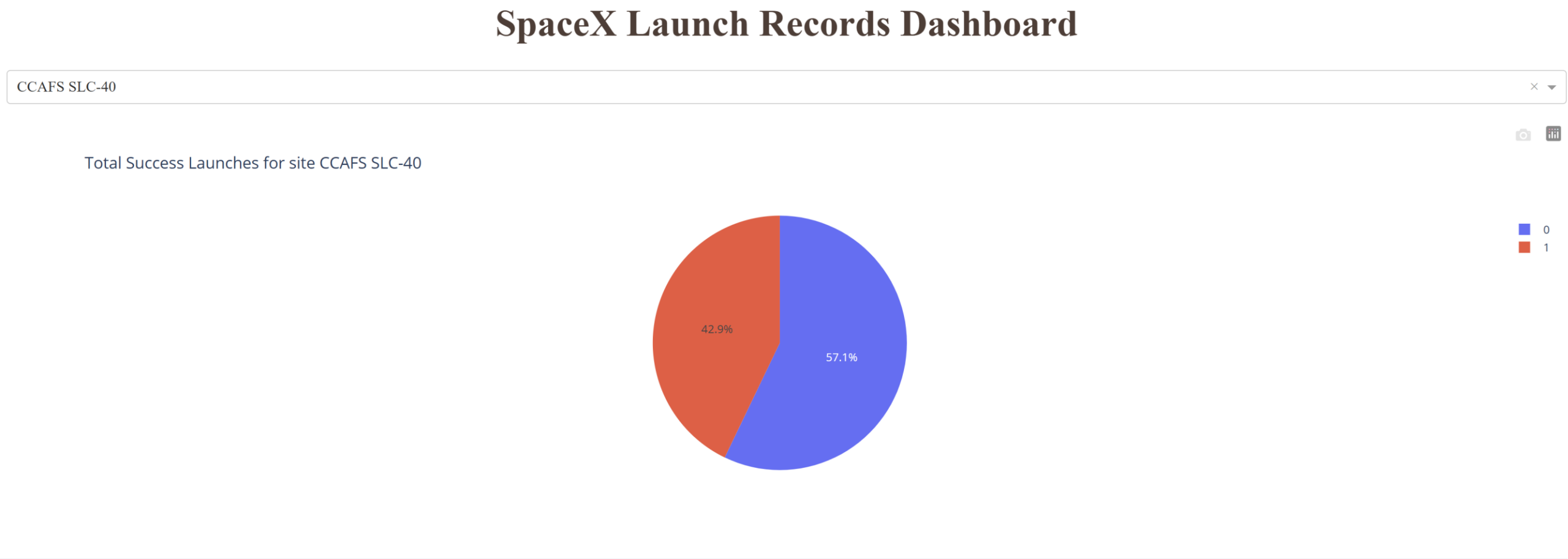



Section 4
Build a Dashboard with Plotly Dash

Dashboard showing ‘All Sites’



Dashboard showing only ‘CCAFS SLC-40’



Dashboard Payload Range with Slider

Section 5

Predictive Analysis (Classification)

Classification Accuracy

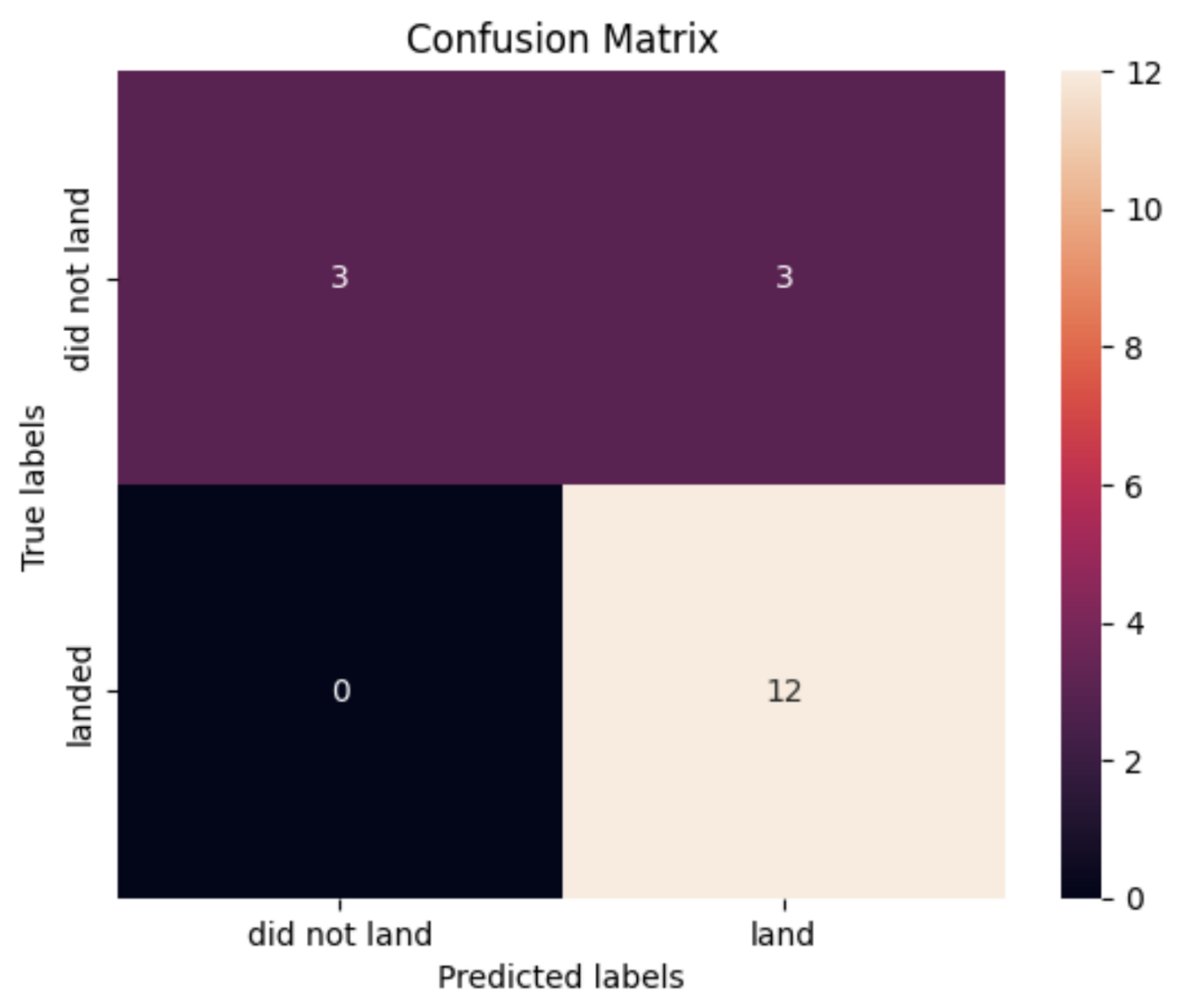
Method scores:

- Logistic Regression: 0.8333
- SVM: 0.8333
- Decision Tree: 0.6111 KNN: 0.8333

Best performing method:

- Logistic Regression with accuracy: 0.8333

Confusion Matrix



Conclusions

- **Predictive Accuracy**: Achieved 83.33% landing success prediction using multiple ML algorithms
- **Success Evolution**: Clear improvement from ~20% (2013) to ~90% (2020) success rates
- **Site Specialization**: VAFB SLC-4E optimized for lighter payloads, CCAFS SLC-40 handles full range
- **Optimal Conditions**: LEO/ISS orbits and 4000-6000kg payload range maximize landing success probability
- **Business Impact**: $62M vs $165M+ cost advantage through successful landing prediction and reusability
- **Geographic Factors**: Coastal proximity essential for landing operations, infrastructure access critical
- Technology Maturation: Demonstrated learning curve validates continuous improvement in reusability systems

Appendix

Technical Assets:

- **Datasets:** 4 processed CSV files (API, web scraped, wrangled, engineered features)
- **SQL Database:** SQLite with 10+ analytical queries
- **Python Notebooks:** 7 complete analysis workflows with code and outputs
- **Interactive Maps:** Folium visualizations with distance calculations
- **Dashboard:** Plotly Dash application with filtering capabilities
- **ML Models:** 4 trained classification algorithms with hyperparameter optimization
- **Performance Metrics:** Confusion matrices, accuracy scores, cross-validation results

Thank you!