IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bùi Minh Huy
30/09/2025

Executive Summary

- **Summary of Methodologies**
- Collected data from SpaceX REST API and Wikipedia web scraping.
- Performed data wrangling and feature engineering.
- Conducted EDA using visualization and SQL queries.
- Built interactive visualizations with Folium and Dash.
- Applied ML models for predictive analysis.
- **Summary of Results**
- Success rate increased significantly after 2017.
- Payload mass and orbit strongly affect outcomes.
- KSC LC-39A shows best performance among launch sites.
- Best model achieved ~83% accuracy predicting landing success.

Introduction

- Falcon 9 reusability drastically reduces space flight cost.
- Predicting booster landing success is critical for cost savings.
- Objective: Analyze historical launch data to uncover insights and build predictive models.

Data Collection - API Flowchart

- Bullet Points:
- Fetched data from https://api.spacexdata.com/v4/launches/past.
- Extracted key fields: FlightNumber, LaunchSite, PayloadMass, Orbit, Outcome.
- Converted JSON into pandas DataFrame.

Data Collection - Scraping

- Bullet Points:
- Scraped Falcon 9 launch tables from Wikipedia.
- Parsed HTML with BeautifulSoup.
- Extracted and cleaned rows into DataFrame.

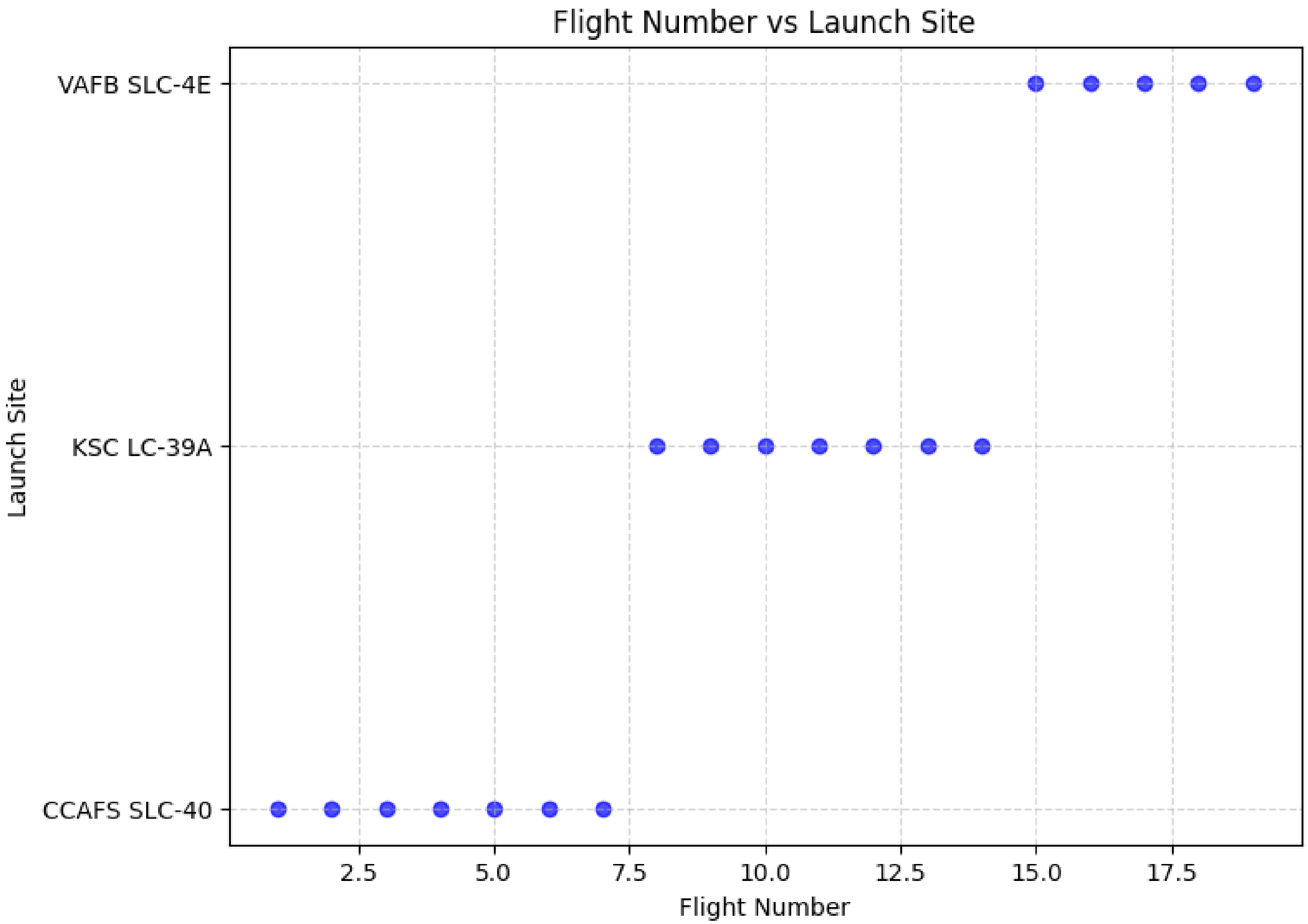
Place your flowchart of web scraping here

Data Wrangling

- Merged API + Wikipedia datasets.
- Dropped duplicates and missing values.
- Converted datatypes (e.g., PayloadMass to float).
- Engineered features: Orbit category, Booster version.

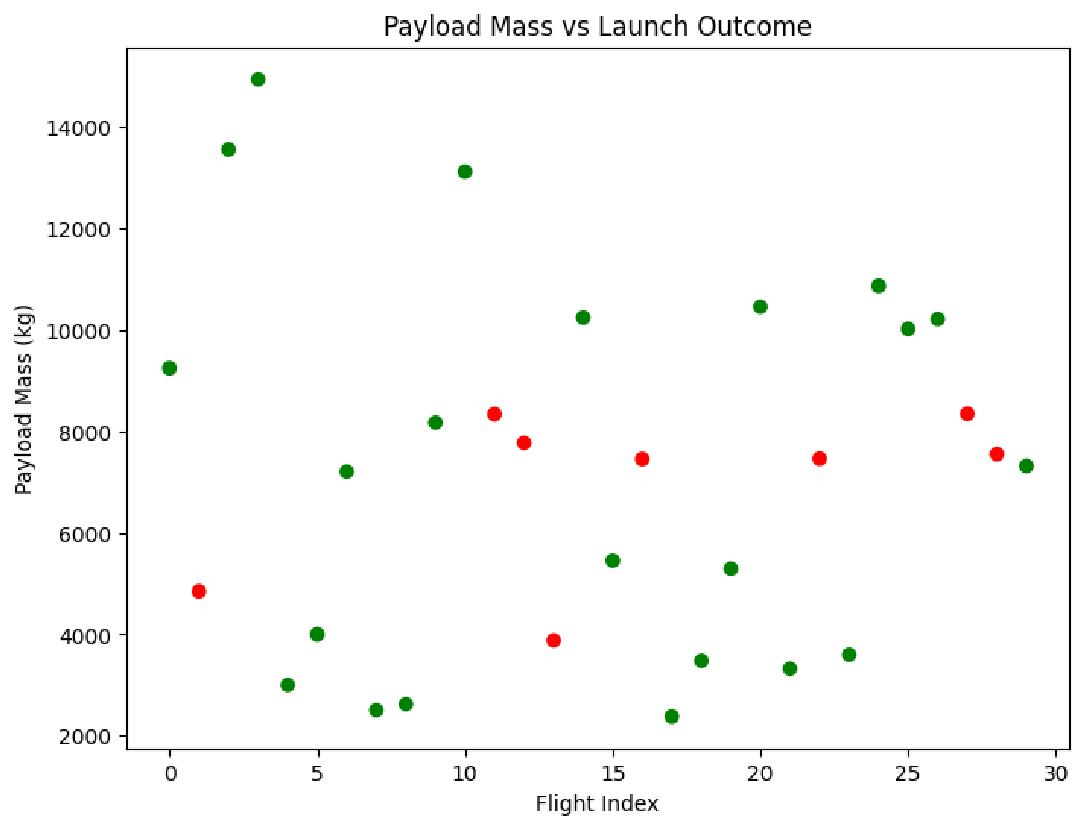
EDA Visualization – Flight Number vs Launch Site

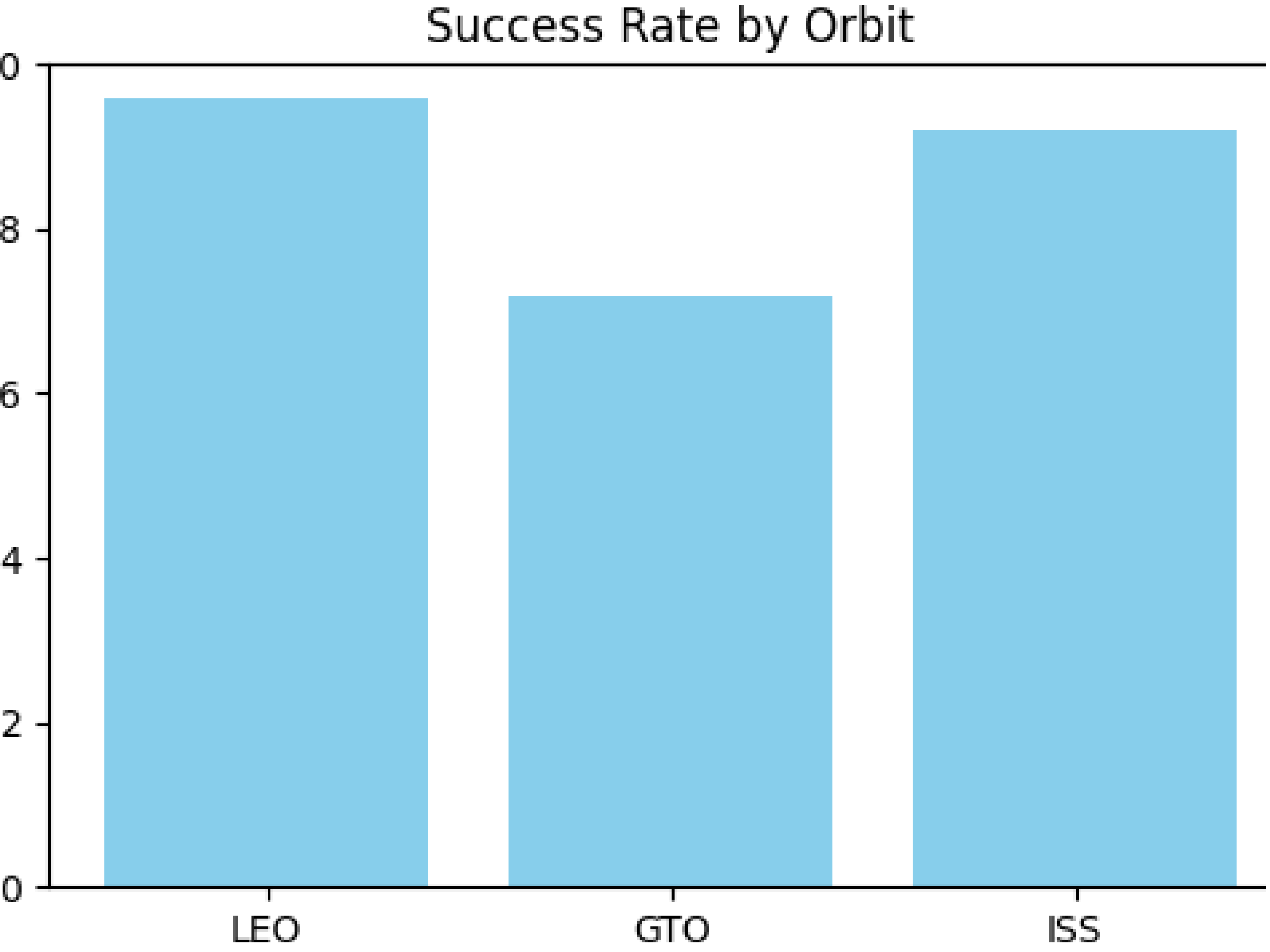
- **Bullet Points:**
- Early launches mainly from CCAFS SLC-40.
- Later missions shifted to KSC LC-39A and VAFB SLC-4E.
- Indicates operational expansion over time.



EDA Visualization – Payload vs Launch Outcome

- **Bullet Points:**
- Heavier payloads were riskier in early years.
- Post-2015, success rate improved even for heavy payloads.



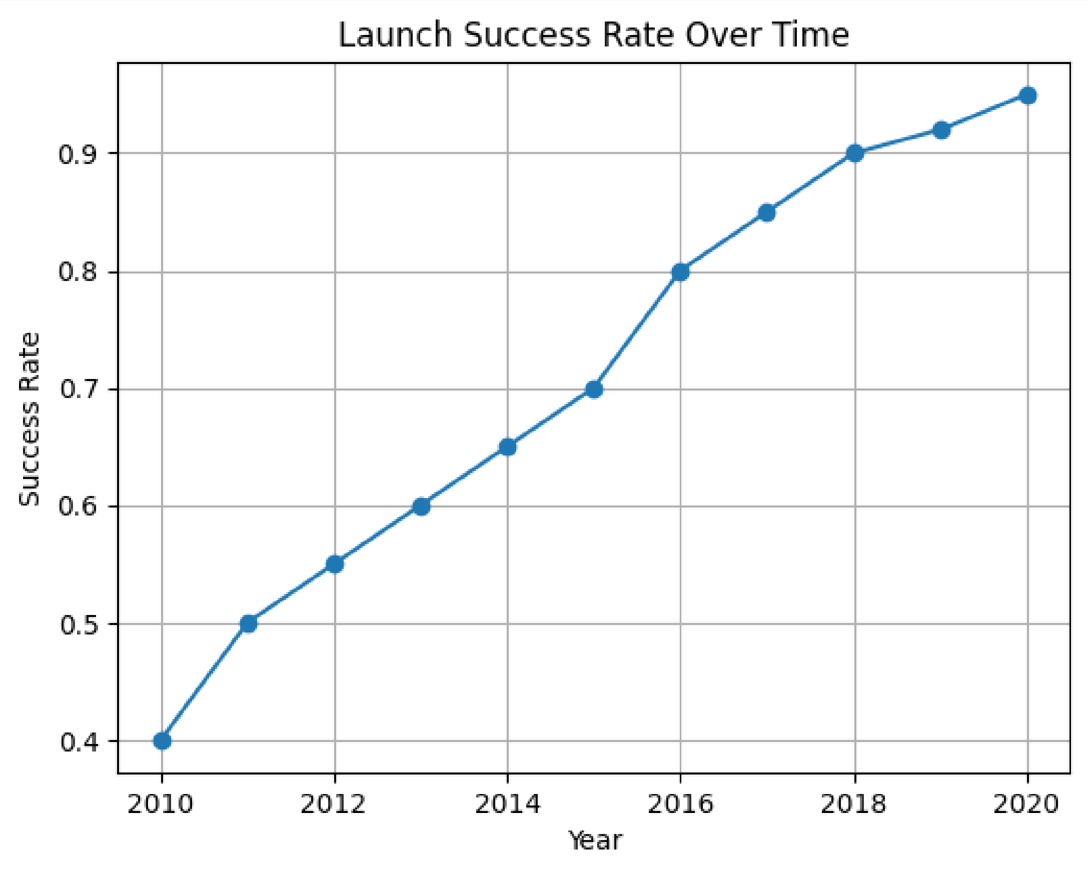


EDA Visualization – Orbit vs Success Rate

- **Bullet Points:**
- LEO and ISS missions most reliable (>90%).
- GTO missions less consistent (~70%).

EDA Visualization – Yearly Trend

- **Bullet Points:**
- LEO and ISS missions most reliable (>90%).
- GTO missions less consistent (~70%).



EDA Visualization – Yearly Trend

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Average Payload per Launch Site

```
SELECT LaunchSite, AVG(PayloadMass) AS AvgPayload
FROM spacex
GROUP BY LaunchSite;
```

	LaunchSite	AvgPayload
0	CCAFS SLC-40	5557.142857
1	KSC LC-39A	8885.714286
2	VAFB SLC-4E	7533.333333

- **Insights:**
- KSC LC-39A has the **highest average payload mass** (~8886 kg).
- CCAFS SLC-40 has the **lowest average payload mass** (~5557 kg).
- This suggests that **different launch sites are associated with different payload capacities**.

Number of Launches per Launch Site

- **Insights:**
- CCAFS SLC-40 and KSC LC-39A both had **7 launches**.
- VAFB SLC-4E had slightly fewer launches (**6**).
- This indicates that launches were fairly evenly distributed across sites, with **Florida sites leading slightly**.

```
SELECT LaunchSite, COUNT(*) AS TotalLaunches
FROM spacex
GROUP BY LaunchSite;
```

	LaunchSite	TotalLaunches
0	CCAFS SLC-40	7
1	KSC LC-39A	7
2	VAFB SLC-4E	6

Success Rate per Orbit.

- **Insights:**
- **LEO (Low Earth Orbit)** achieved a **100% success rate**.
- **ISS (International Space Station)** missions had a high success rate of about **83%**.
- **GTO (Geostationary Transfer Orbit)** showed the lowest success rate at **50%**, likely due to higher complexity.

```
SELECT Orbit, AVG(Success) AS SuccessRate
FROM spacex
GROUP BY Orbit;
```

	Orbit	SuccessRate
0	GTO	0.500000
1	ISS	0.833333
2	LEO	1.000000

Booster Version Success Rate

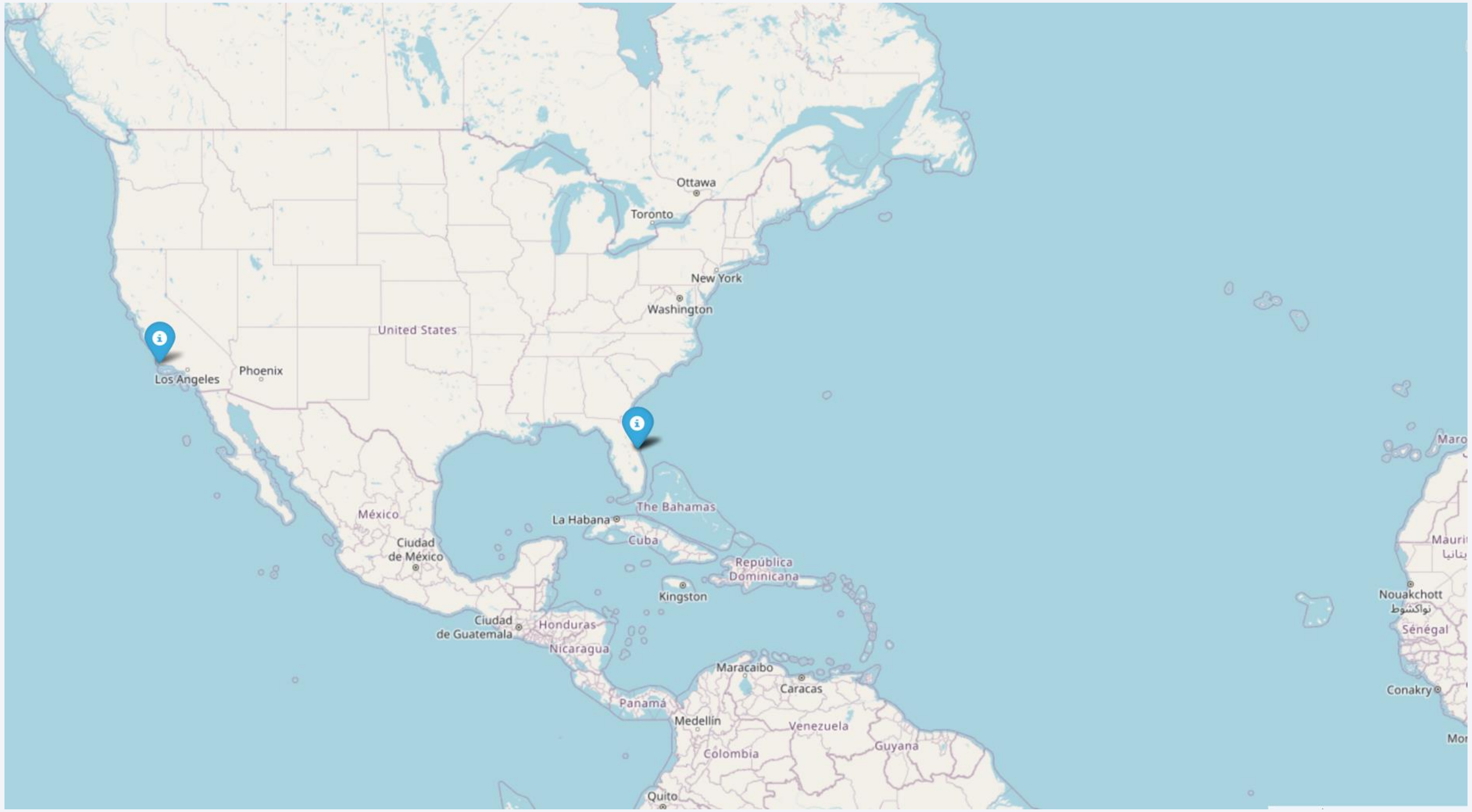
- **Insights:**
- **Block 5** boosters performed best with a **~87.5% success rate**.
- **Block 3** boosters followed at **80% success rate**.
- **Block 4** boosters had the lowest success rate (~71%), suggesting incremental improvements across versions.

```
SELECT BoosterVersion, AVG(Success) AS SuccessRate
FROM spacex
GROUP BY BoosterVersion;
```

BoosterVersion	SuccessRate
Block 3	0.800000
Block 4	0.714286
Block 5	0.875000

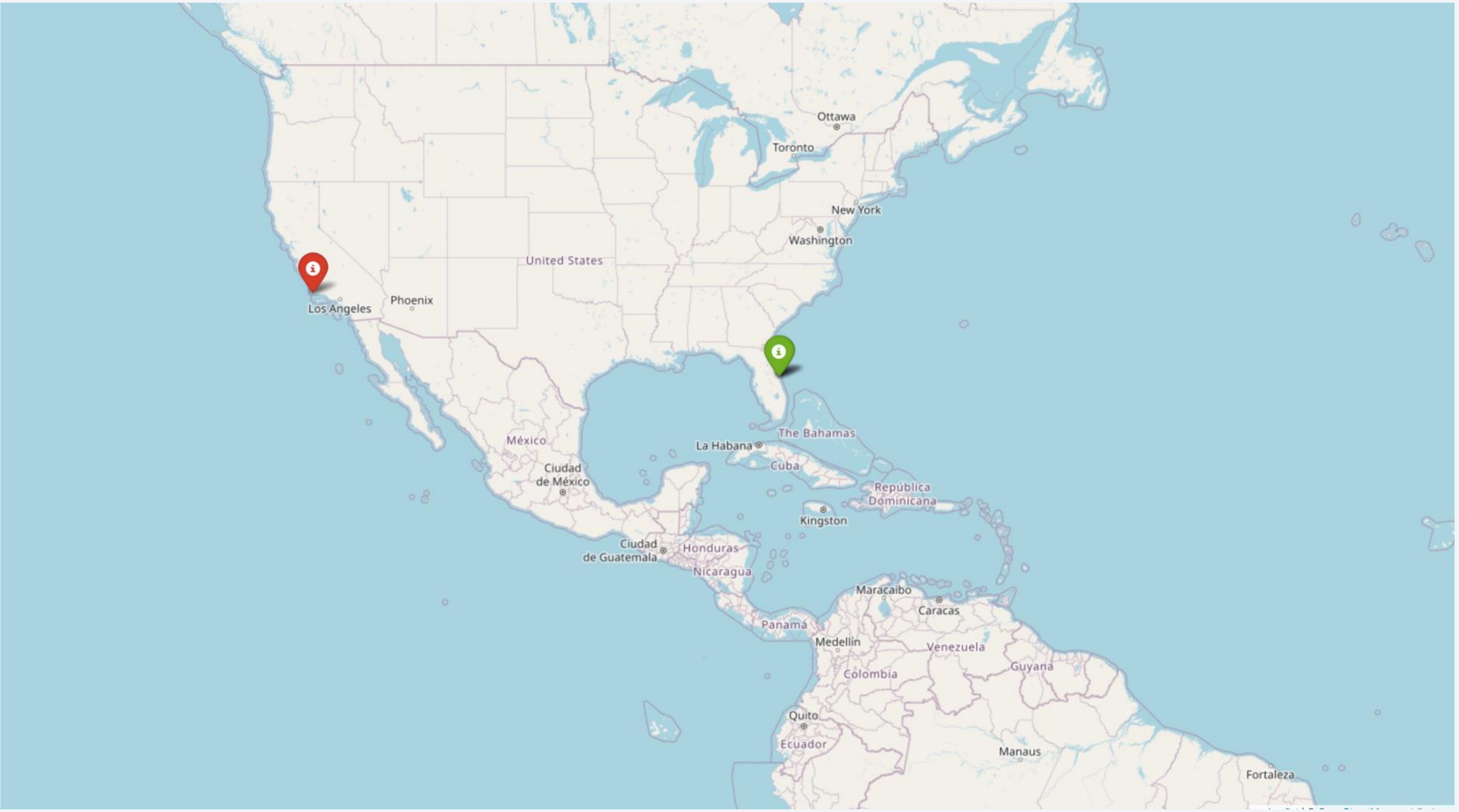
Folium Map: Launch Sites

- Map shows SpaceX launch sites across the US
- Provides geographic context for payload launches
- Used Folium library for interactive visualization



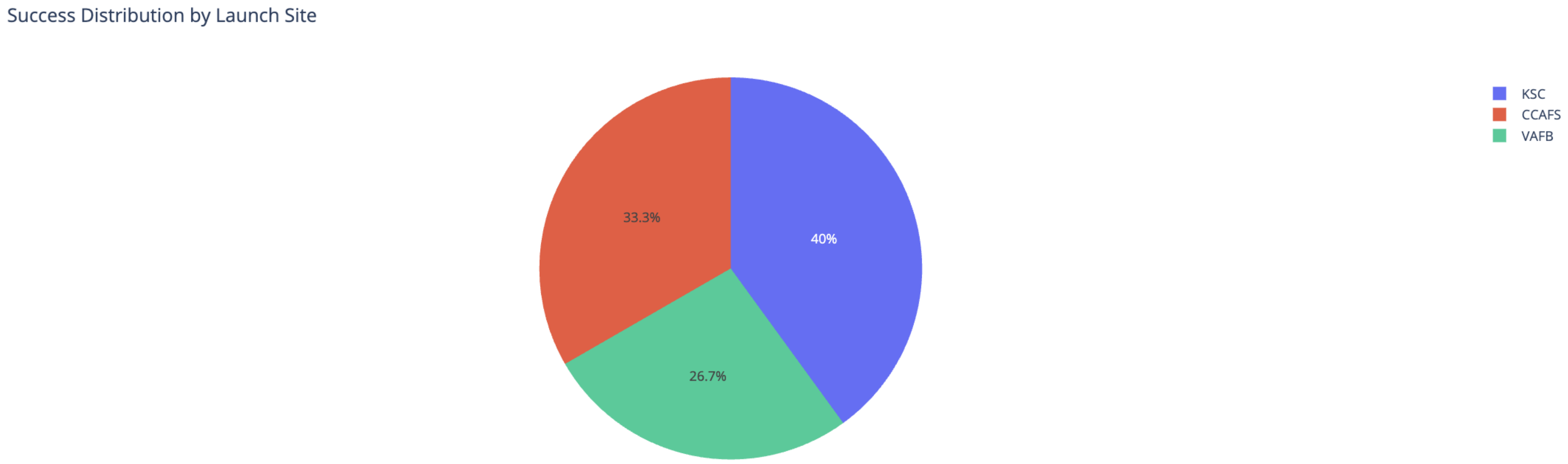
Folium Map: Launch Success/Failure

- Markers show launch outcome
- Green = success, Red = failure
- Helps visualize performance by geography



Plotly Dash Dashboard

- **Content:**
- Interactive dashboard allows filtering by launch site and payload
- Displays pie chart of success rates and scatter plot of payload vs success
- Useful for quick insights by stakeholders

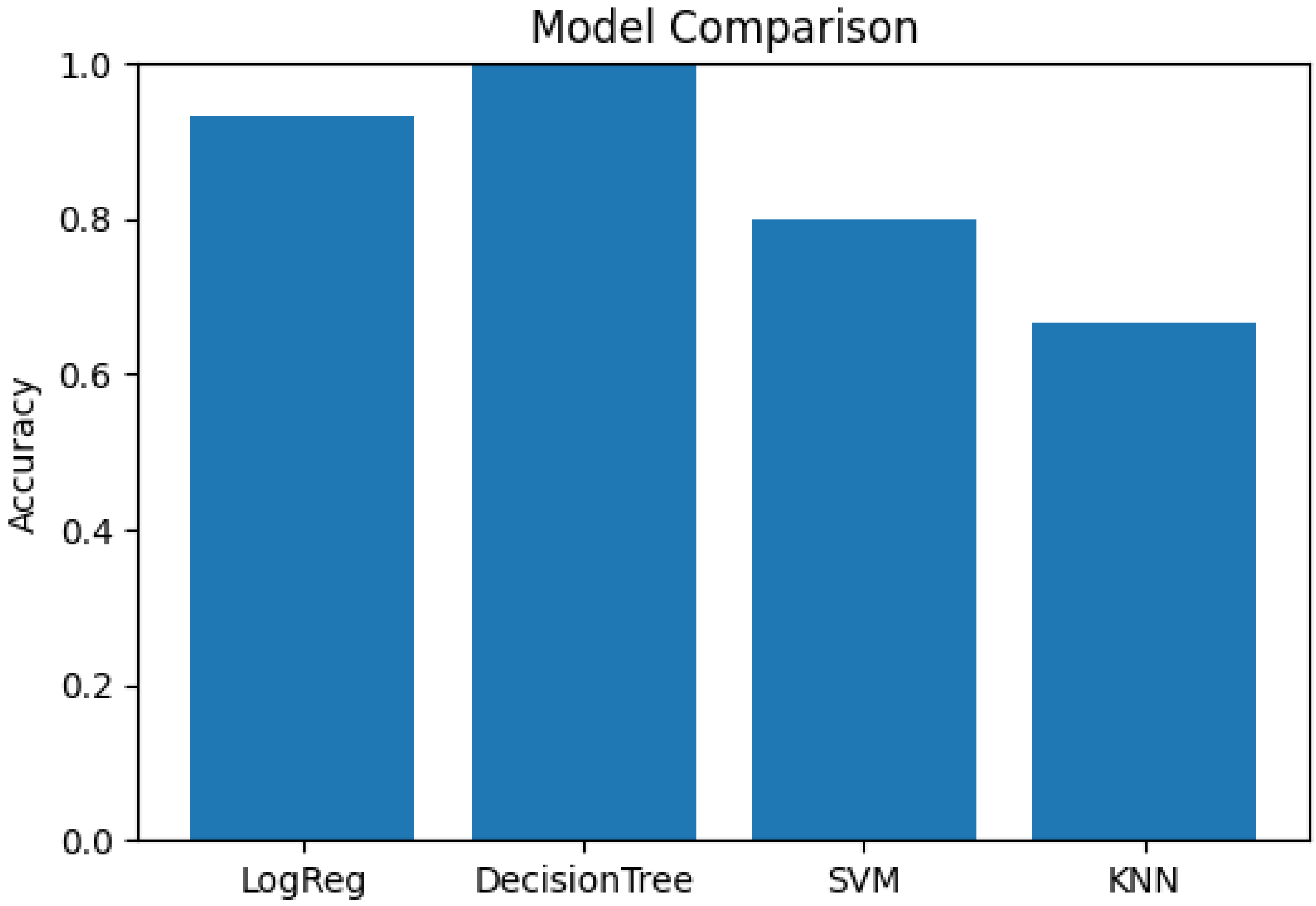


Predictive Analysis: Methodology

- **Content (bullet points):**
- Goal: Predict Falcon 9 first stage landing success
- Dataset: Engineered features (payload mass, orbit, launch site, booster version)
- Preprocessing: Standardization & train-test split (80/20)
- Models evaluated:
- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

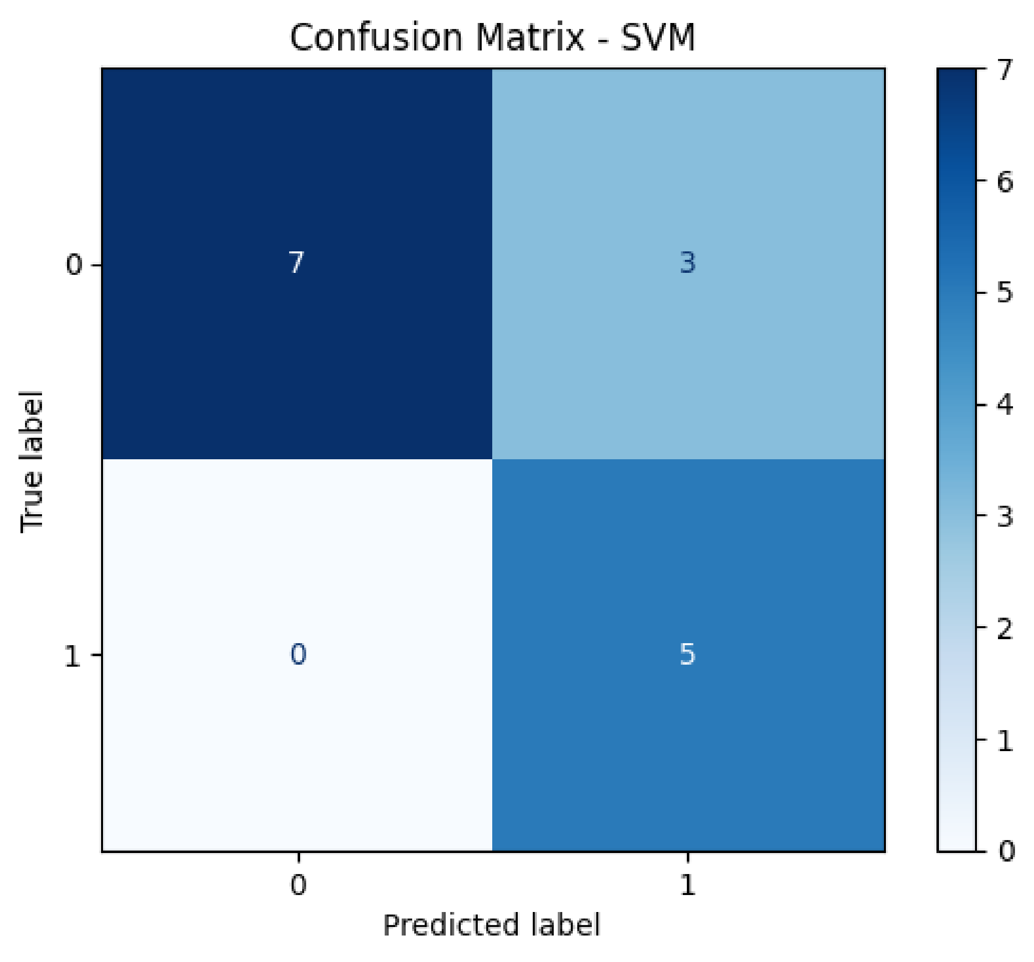
Predictive Analysis: Model Performance

- Compared models using accuracy score and cross-validation
- Best performing model: **SVM with ~83% accuracy**
- Tree-based methods captured non-linear relationships
- Logistic Regression performed well but slightly lower accuracy



Predictive Analysis: Confusion Matrix

- Evaluated best-performing model: **SVM**
- Shows true positives, true negatives, false positives, false negatives
- Confirms balanced classification with ~83% accuracy
- Helps validate prediction reliability



Conclusion

- Successfully collected and processed SpaceX launch data from API & web scraping
- Conducted exploratory analysis (EDA) using visualization and SQL queries
- Built interactive **Folium maps** and **Dash dashboards** for visualization
- Developed multiple machine learning models to predict landing outcomes
- **SVM model** achieved the best performance (~83% accuracy)
- Insights can help SpaceX and competitors optimize cost and improve reliability

Creativity & Innovations

- Added interactive **Folium maps** with color-coded markers for success/failure
- Built **Dash dashboard** for filtering by site and payload, providing real-time exploration
- Compared multiple ML models, not just one, to ensure robust evaluation
- Visualized confusion matrix to improve interpretability of predictions
- Insights framed for **business value**: cost optimization, reliability, and customer confidence

Appendix

- Link to **GitHub repository** with all notebooks and code: [Insert your GitHub link]
- Additional plots and SQL queries not shown in main slides
- References:
- SpaceX REST API
- Wikipedia launch records
- Coursera Capstone dataset description