Winning Space Race with Data Science

<Name>
<Date>

OUTLINE

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

EXECUTIVE SUMMARY

This capstone project focuses on predicting the successful landing of the SpaceX Falcon 9 first stage using multiple machine learning classification techniques.

The project follows a structured process:

- Data collection, preparation, and formatting
- Exploratory data analysis
- Interactive data visualization
- Machine learning modeling and prediction

Our analysis indicates that several launch features are correlated with mission outcomes (success or failure).

Results suggest that the decision tree algorithm performs particularly well for predicting Falcon 9 first-stage landing success.

INTRODUCTION

The goal of this capstone is to predict whether the Falcon 9 first stage will land successfully.

SpaceX advertises Falcon 9 launches at $62 million per mission, while competitors charge $165 million or more. The cost advantage largely comes from SpaceX's ability to reuse the first stage.

By predicting landing success, we can estimate launch costs—insightful for competitors considering bids against SpaceX.

Not all unsuccessful landings indicate failure; in some cases, SpaceX intentionally performs controlled ocean landings.

Key question: Given features such as payload mass, orbit type, and launch site, can we accurately predict whether the Falcon 9 first stage will land successfully?

Section 1

Methodology

METHODOLOGY

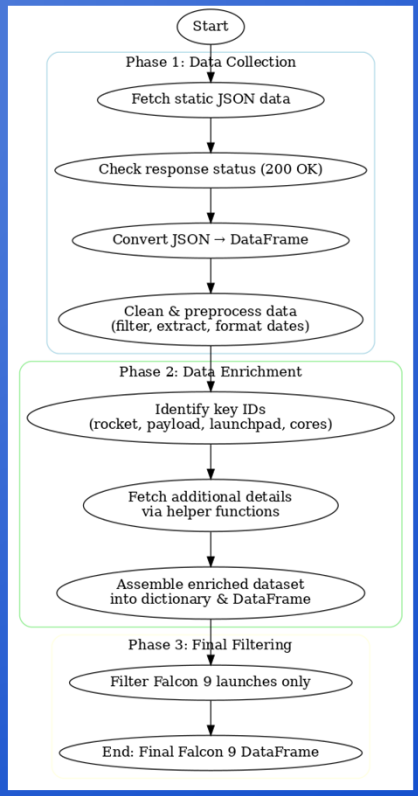
Executive Summary

- Data collection methodology:
  - Sourced from spacedata.com API and Wikipedia
- Perform data wrangling
  - Processed using local Jupyter notebooks in Python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Supervised machine learning models from Scikit-Learn package

Data Collection – SpaceX API

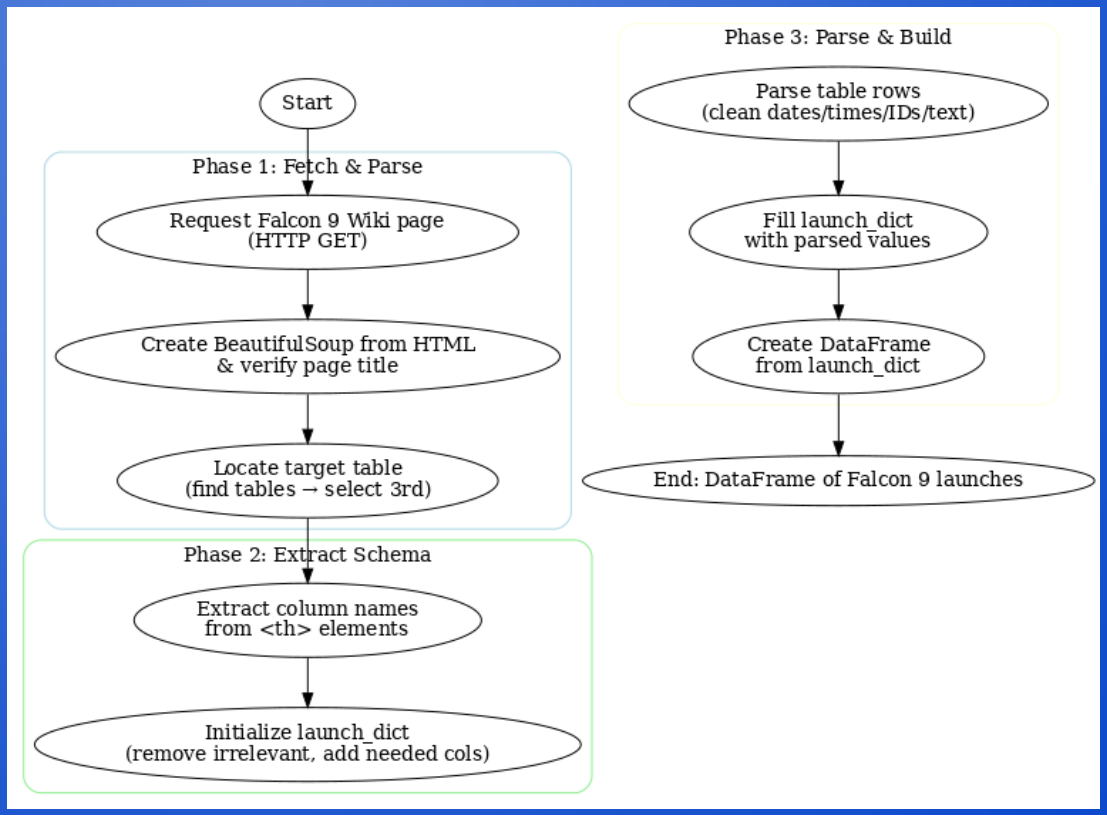
- The API used is https://api.spacexdata.com/v4/rockets/.
- Called by “requests” Python package

  requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
- Code found here: https://github.com/scissory/applied-data-science-capstone/blob/0ac0955b3c9c774e4e400727af61ffc92e43f10b/Data%20Collection%20API.ipynb



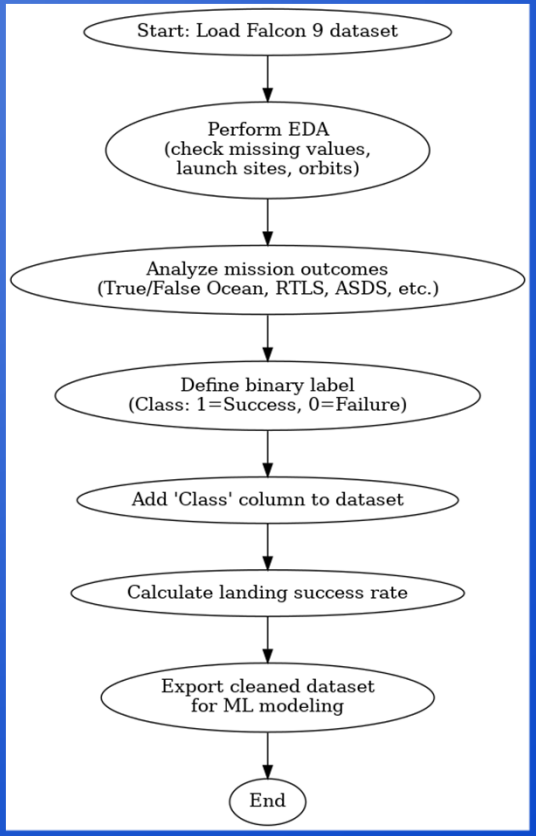
Data Collection - Scraping

- Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Using BeautifulSoup package for web scrapping
- Code found here: https://github.com/scissory/applied-data-science-capstone/blob/0ac0955b3c9c774e4e400727af61ffc92e43f10b/Data%20Collection%20Web%20Scraping.ipynb



DATA WRANGLING

- **Exploratory Data Analysis (EDA):** Loaded and examined the SpaceX Falcon 9 dataset, assessed missing values, and analyzed launches by site and orbit types to understand data distribution.
- **Mission Outcomes & Label Creation:** Categorized landing outcomes (e.g., True/False Ocean, RTLS, ASDS) and simplified them into a binary classification label (1 = successful landing, 0 = unsuccessful).
- **Prepared Dataset for Modeling:** Created a new Class column representing landing success, calculated success rates, and exported the cleaned dataset for use in supervised machine learning tasks.
- Code found here: https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/Data%20Wrangling.ipynb



EDA with Data Visualization

Scatter plots (Flight Number vs. PayloadMass, Launch Site, Orbit):
Used to examine how launch frequency, payload size, and different sites/orbits influence landing success. These help identify trends such as higher success with more flight experience and limitations of specific launch sites with heavy payloads.

Bar chart (Orbit vs. Success Rate):
Plotted to compare average landing success across orbit types. This visualization highlights which orbits have historically higher or lower success rates.

Line chart (Year vs. Success Rate):
Used to show the yearly trend of launch outcomes, demonstrating SpaceX's improvement in landing success over time, especially post-2013.

Code found here:
https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/EDA%20Visualization.ipynb

EDA with SQL

Exploration of launch data: Queries retrieve distinct launch sites, filter records by site prefix, and examine payload/booster information (e.g., total and average payload mass for specific customers or booster versions).

Mission outcome analysis: Queries identify first successful landings, count success vs. failure outcomes, and list failed or successful landing cases by year, site, and booster type.

Advanced filtering and ranking: Queries use subqueries and grouping to find boosters with maximum payload mass and to rank landing outcomes between specific dates.

Code found here:
https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/EDA%20SQL.ipynb

Build an Interactive Map with Folium

Site markers & circles: Added a folium. Circle and text Marker at each launch site to pin exact coordinates and label the site name—this gives quick geographic context and lets you visually compare site locations on the map.

Outcome markers with clustering: Plotted per-launch Markers colored by success (green) or failure (red) and grouped them with MarkerCluster—this reduces clutter at identical coordinates and makes success rates by site easy to spot at a glance.

Distance tools & lines to proximities: Enabled MousePosition to read coordinates, then added distance-labeled Markers (DivIcon) and PolyLines from a launch site to nearby features (coastline, city, railway, highway)—to assess how proximity to infrastructure or coast might relate to launch operations.

Code found here: https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Interactive inputs (filters): A **Launch Site dropdown** (including “All Sites”) to switch between global view and a specific site, and a **Payload range slider (0–10,000 kg)** to focus analyses on chosen payload intervals — both enable quick slicing of the dataset to answer “where/when do we succeed?” questions.

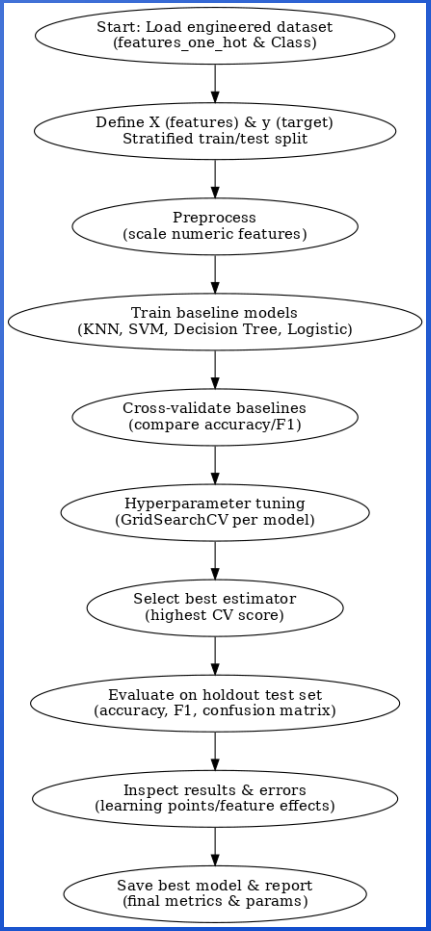
Success distribution pie chart: A **pie chart** that updates via callback from the site dropdown—showing **total successes by site** when “All Sites” is selected, or **success vs. failure counts for the chosen site**—so users can compare which site has the most successes and which has the highest success rate.

Outcome vs. payload scatter plot: A **scatter chart** filtered by both **site** and **payload range**, plotting **Payload Mass (x)** vs. **Launch Outcome/Class (y)** and **coloring by Booster Version**—to visualize how payload correlates with success and to see which **booster versions** perform best across sites and payloads.

Code found here:
https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/dashboard%20app.py

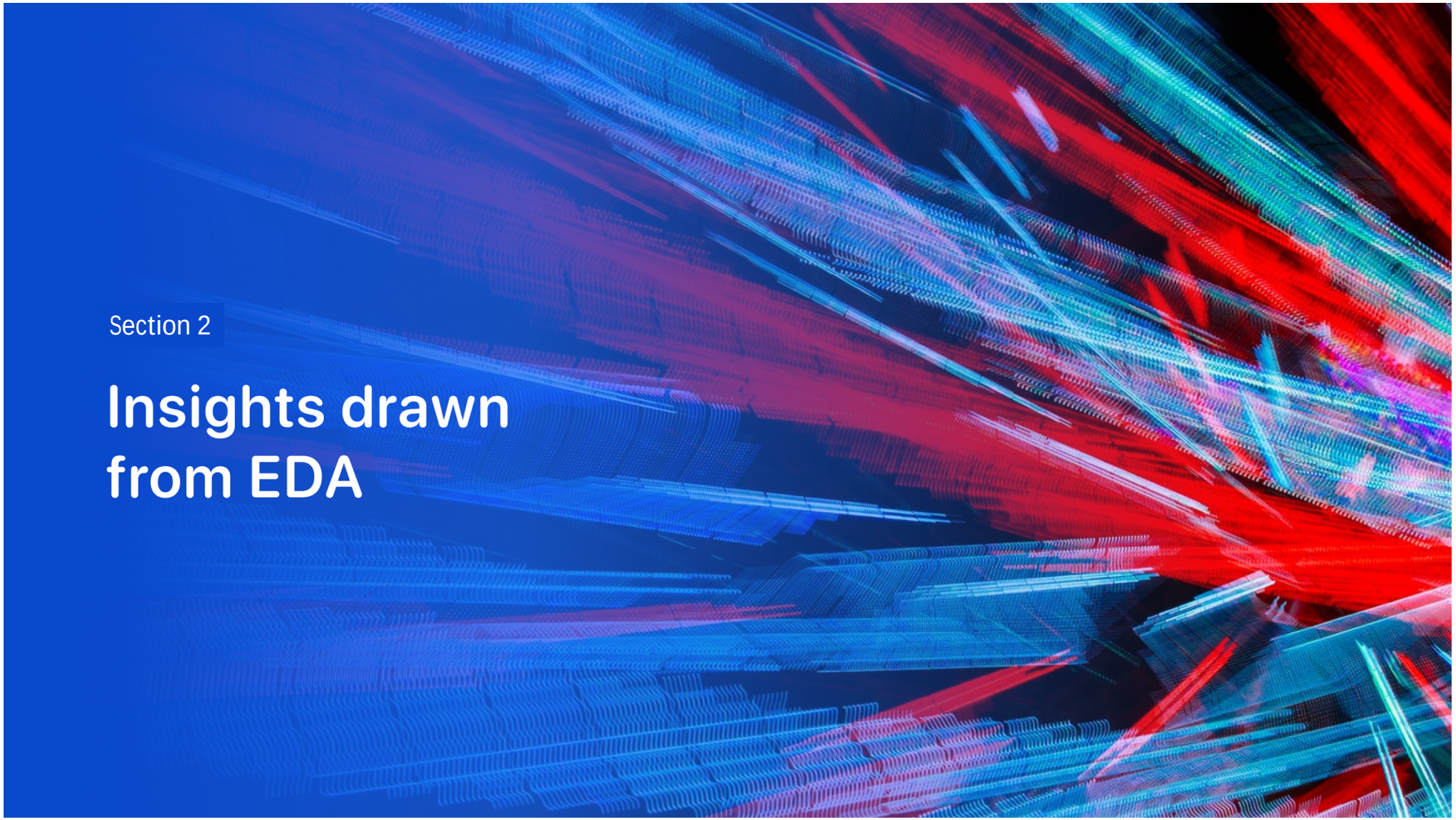
Predictive Analysis (Classification)

- **Build:** Loaded engineered features/labels, performed a stratified train/test split, applied preprocessing (e.g., scaling), and trained multiple baseline classifiers (KNN, SVM, Decision Tree, Logistic Regression).
- **Evaluate & Improve:** Used cross-validation to compare baseline performance, then ran GridSearchCV per model to tune hyperparameters (e.g., k, C, gamma, max_depth) and reduce variance/bias.
- **Select Best:** Chose the best estimator by highest cross-validated score, confirmed on the hold-out test set with metrics (accuracy, F1, confusion matrix), and recorded final parameters/metrics.
- Code found here: https://github.com/scissory/applied-data-science-capstone/blob/e47e4f4d0e8be119f3e5cb0e736579ee4e0d95ee/Machine%20Learning%20Prediction.ipynb



Results

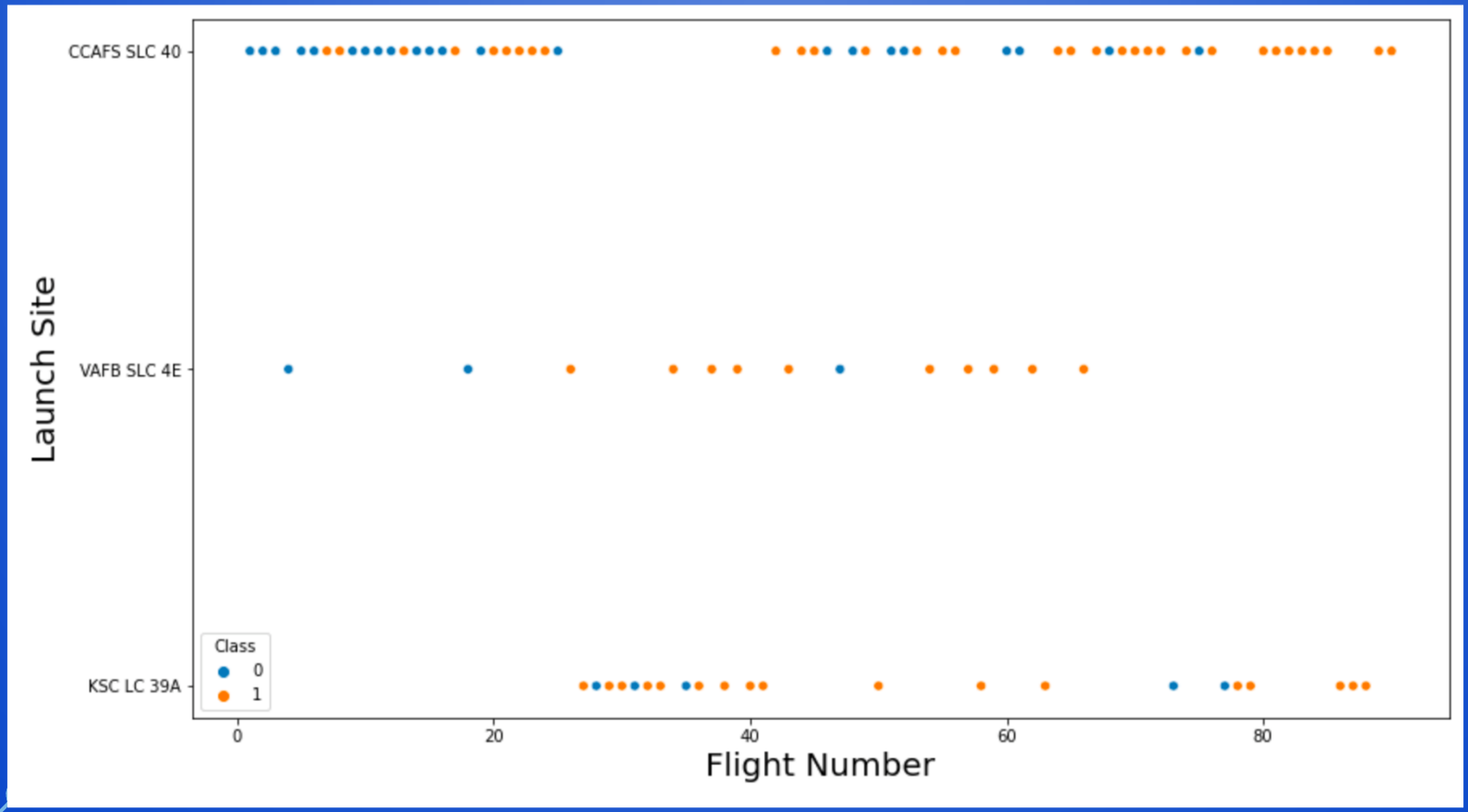
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



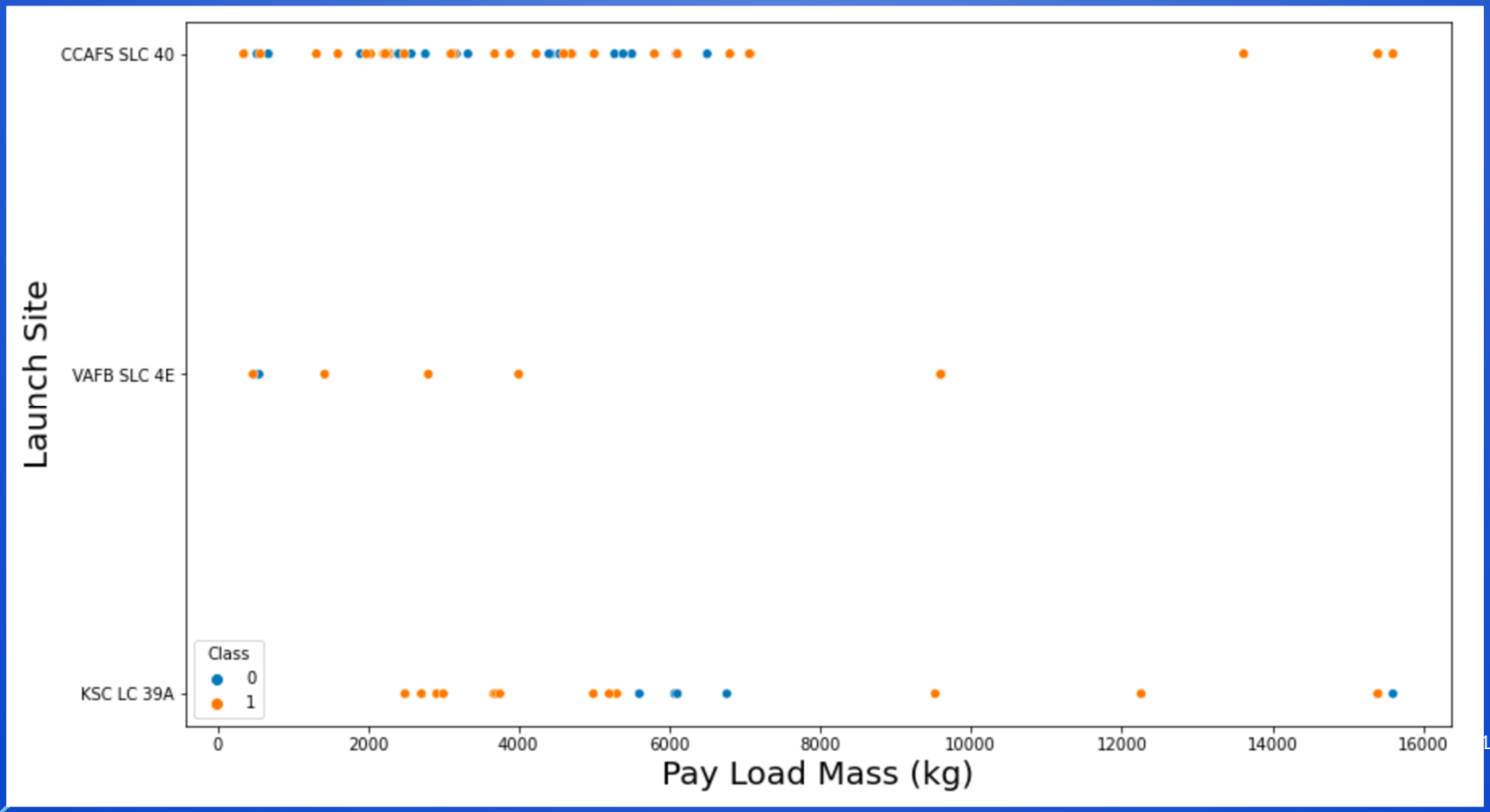
Section 2

Insights drawn from EDA

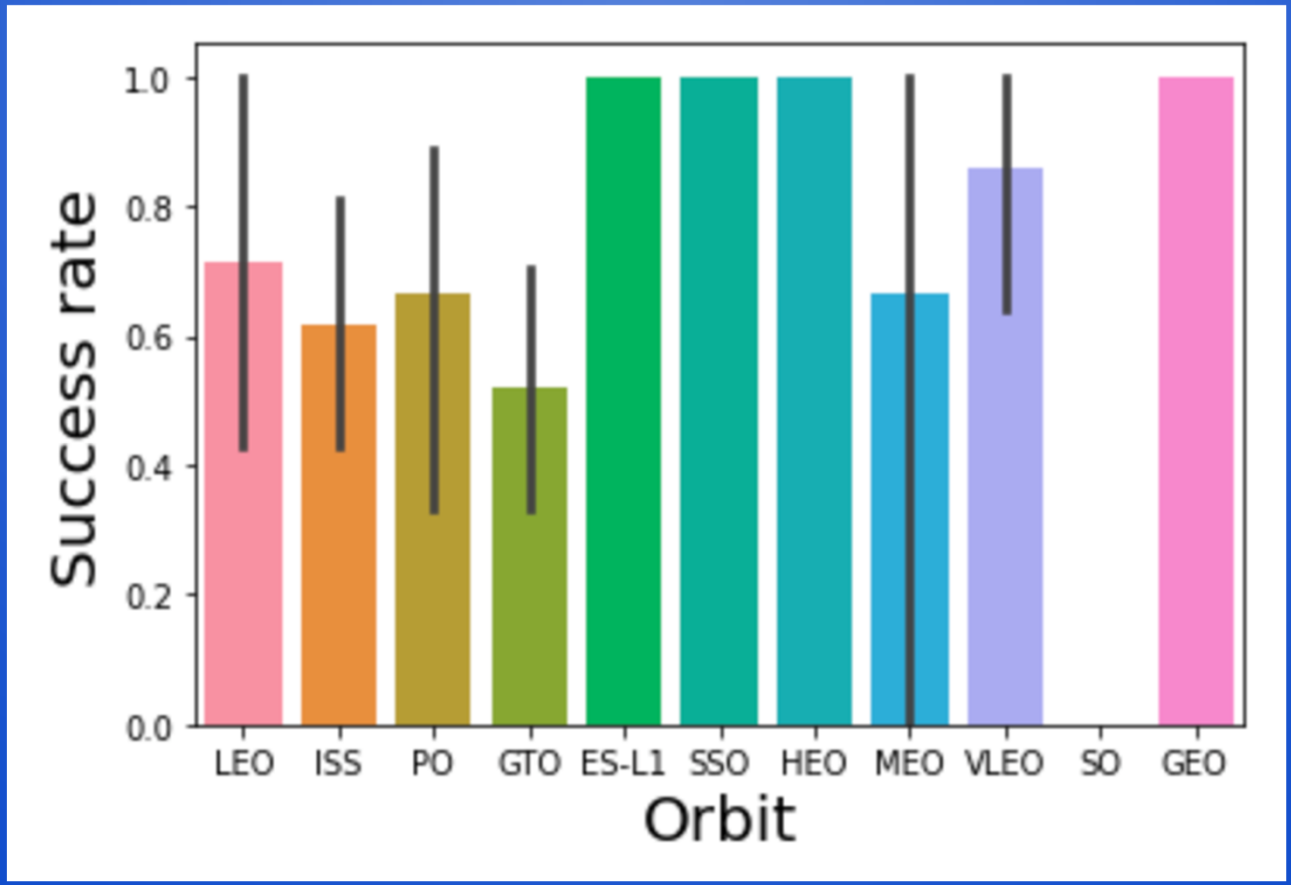
Flight Number vs. Launch Site



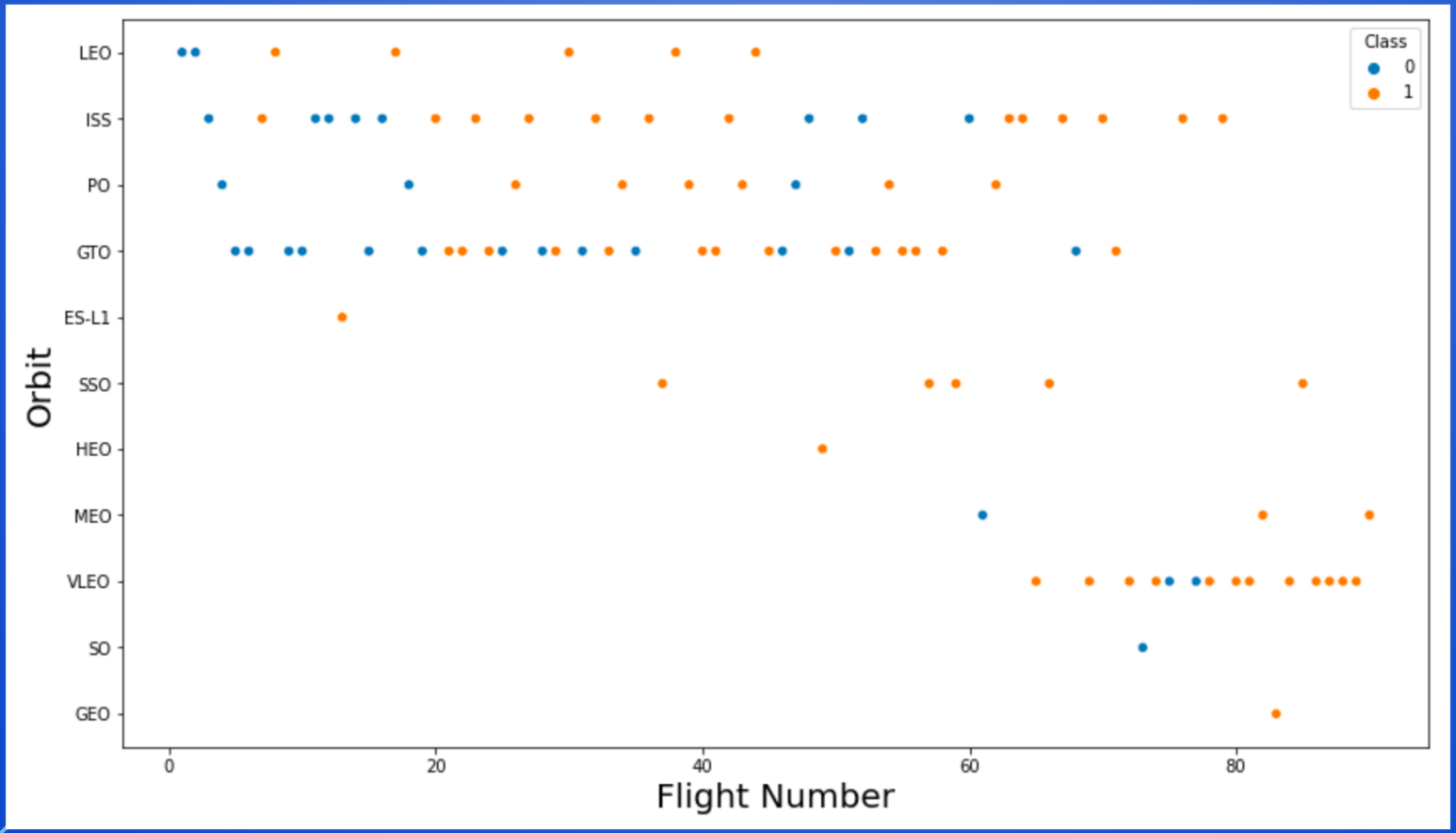
Payload vs. Launch Site



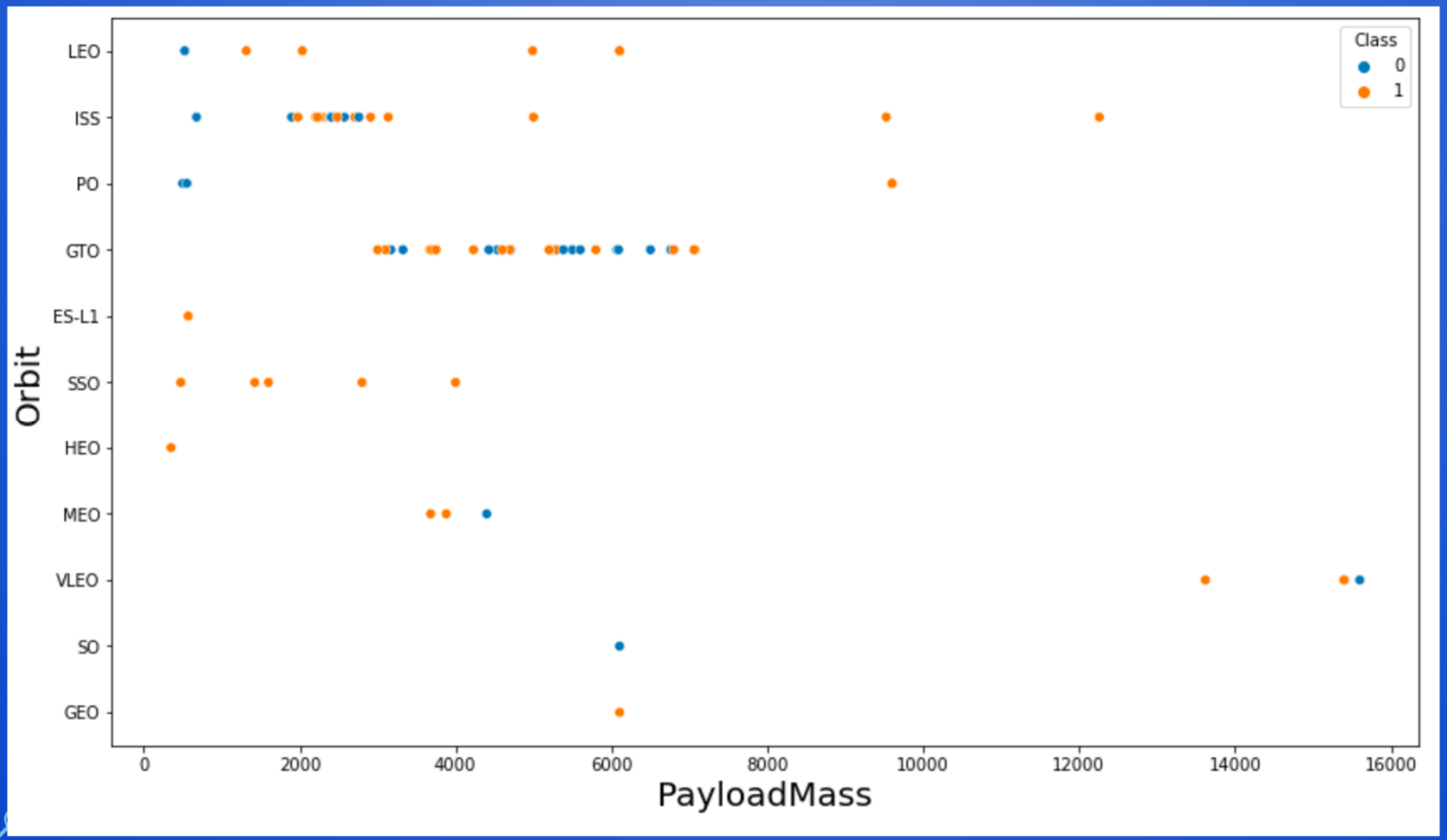
Success Rate vs. Orbit Type



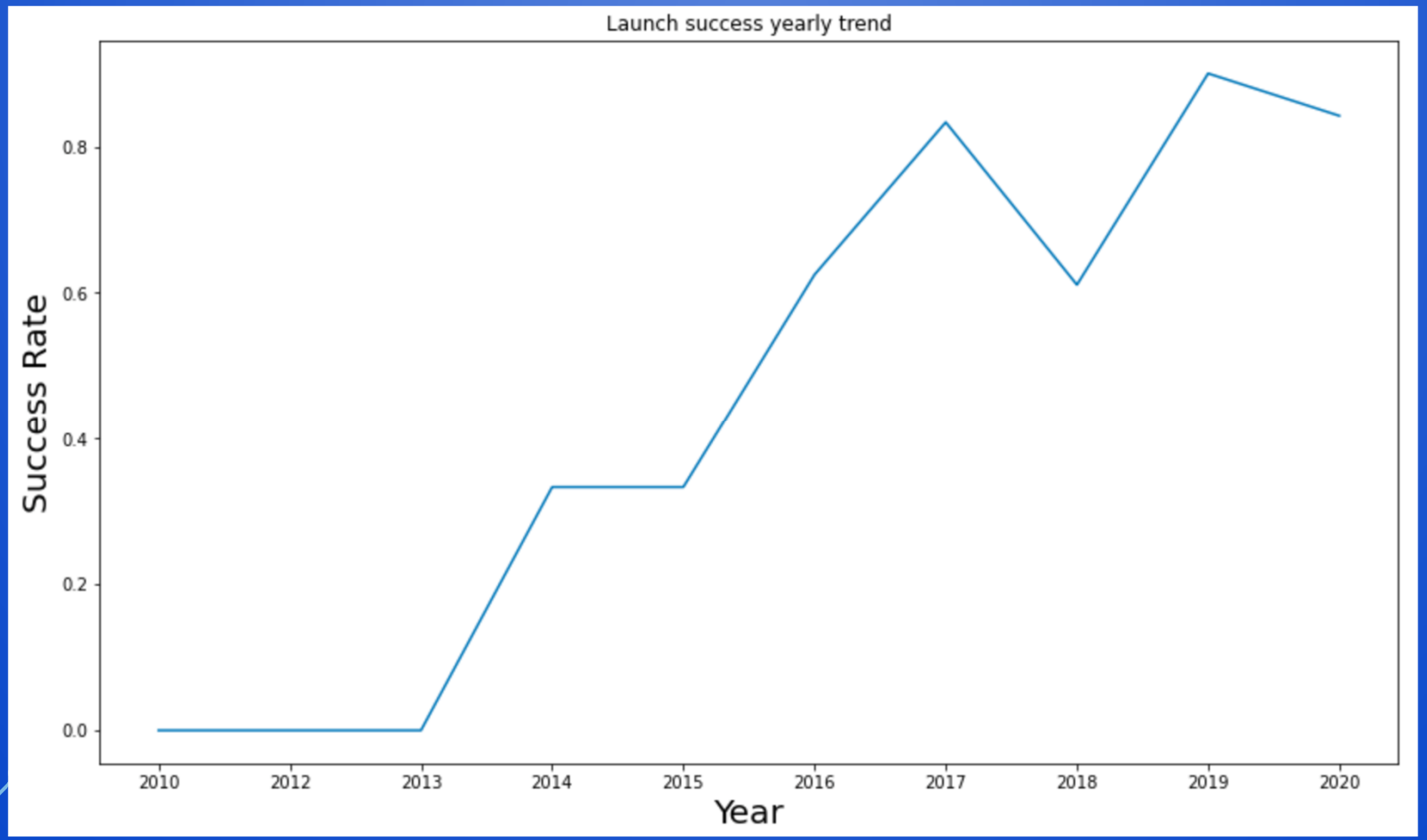
Flight Number vs. Orbit Type



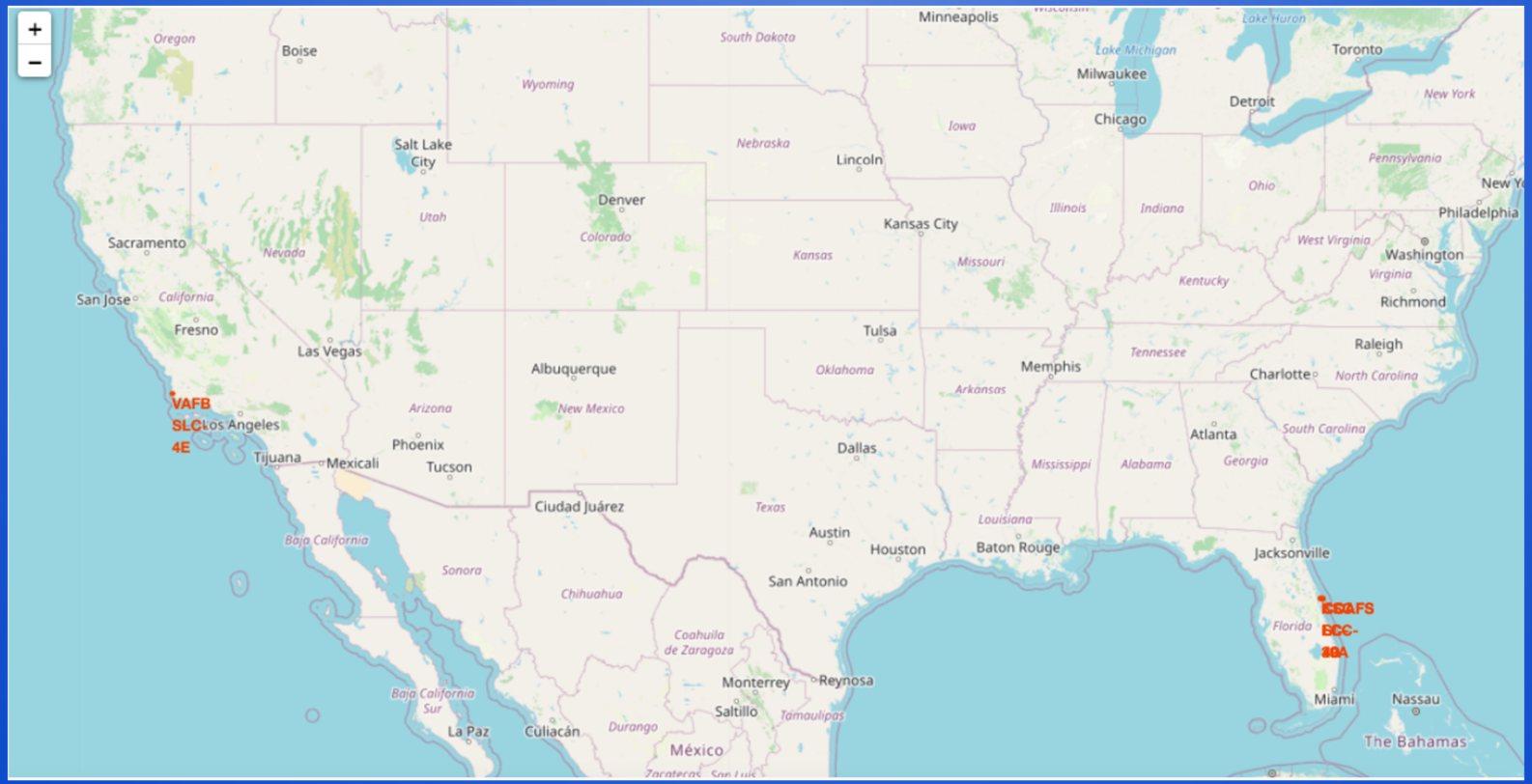
Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names



Launch Site Names Begin with 'CCA'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

Total payload mass by NASA (CRS)
45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1
```

First Successful Ground Landing Date

Date of first successful landing outcome in ground pad
2015-12-22

```
%sql SELECT MIN(DATE) AS "Date of first successful landing outcome in ground pad" FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

Total Number of Successful and Failure Mission Outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

```
%sql SELECT number_of_success_outcomes, number_of_failure_outcomes FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%') success_table, (SELECT COUNT(*) number_of_failure_outcomes FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%') failure_table
```

Boosters Carried Maximum Payload

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTBL);
```

2015 Launch Records

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE year(DATE) = '2015' AND LANDING__OUTCOME = 'Failure (drone ship)';
```

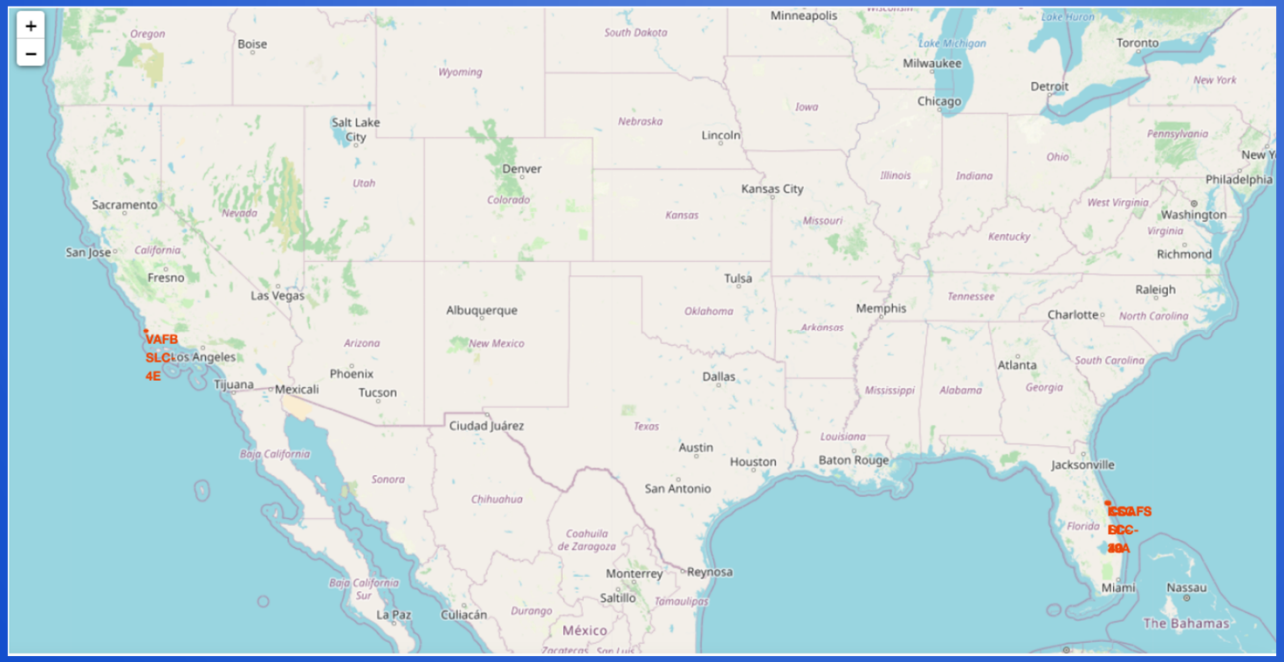
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

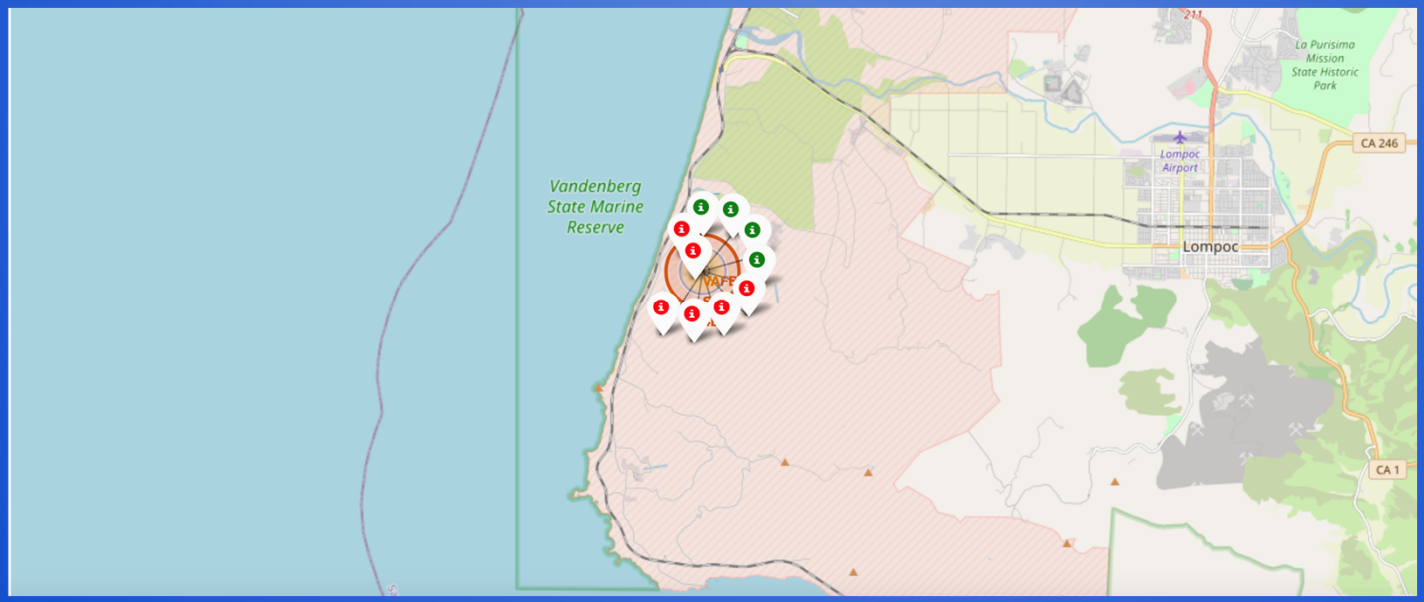
```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE year(DATE) = '2015' AND LANDING__OUTCOME = 'Failure (drone ship)';
```


Section 3
Launch Sites
Proximities Analysis

All Launch Sites

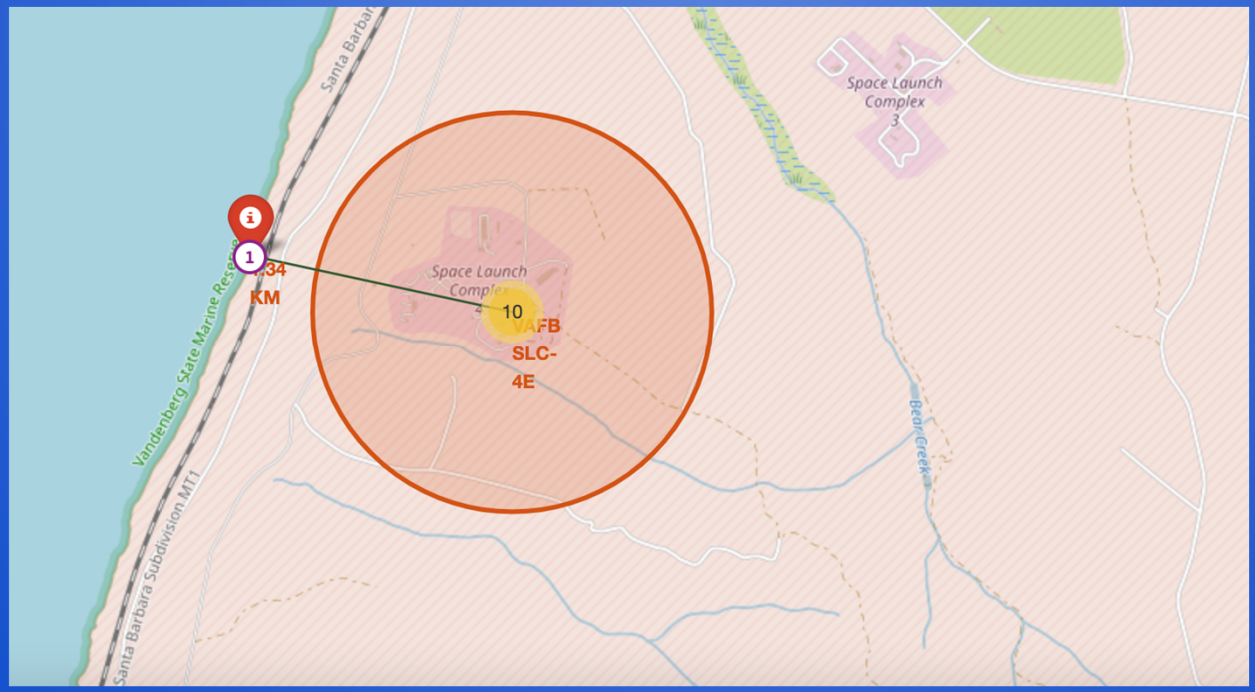


Successful/Failed Launches at Site

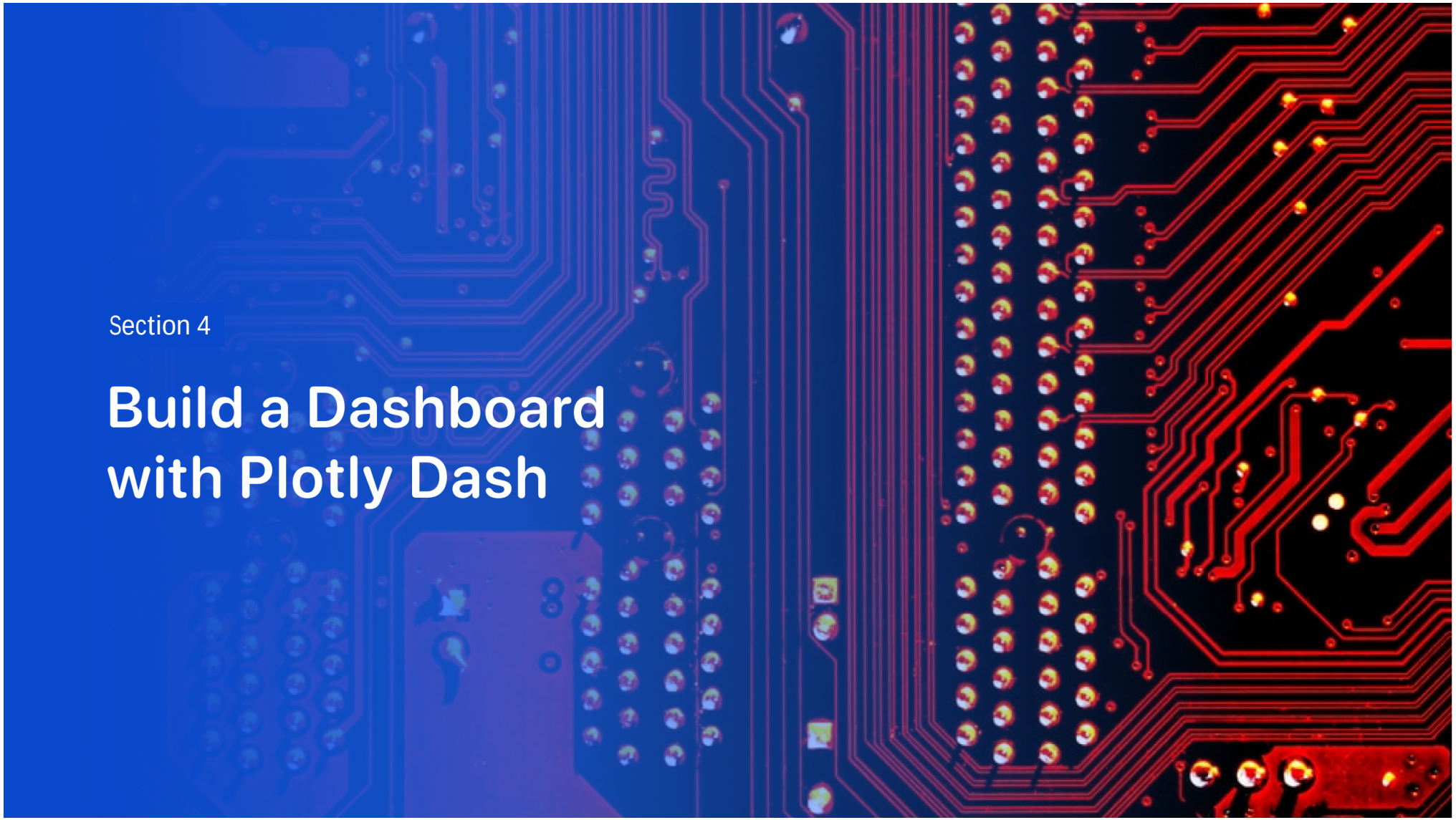


The map shows each launch site with markers indicating successful and failed launches. Zooming into a site reveals color-coded tags: green for success and red for failure.

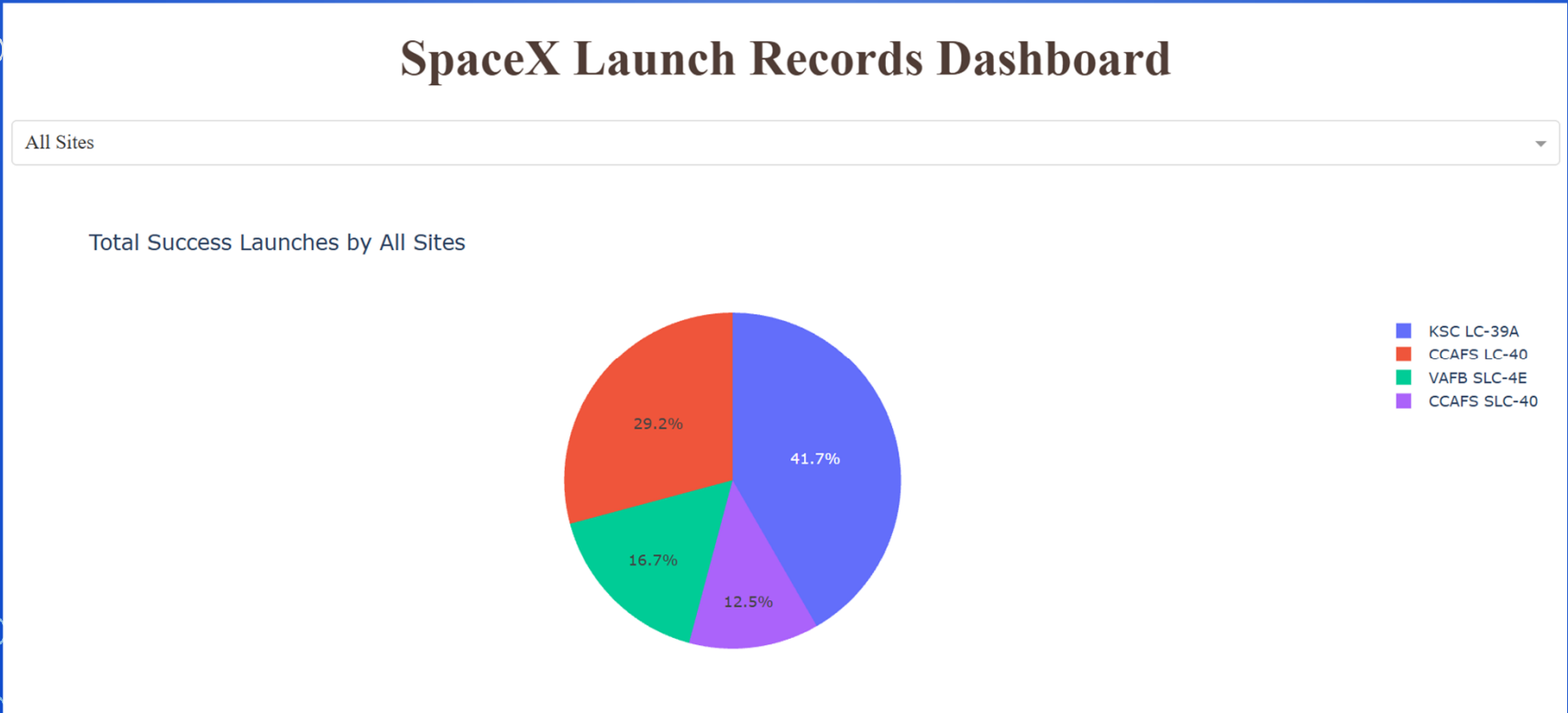
Launch Site Proximity to Cities



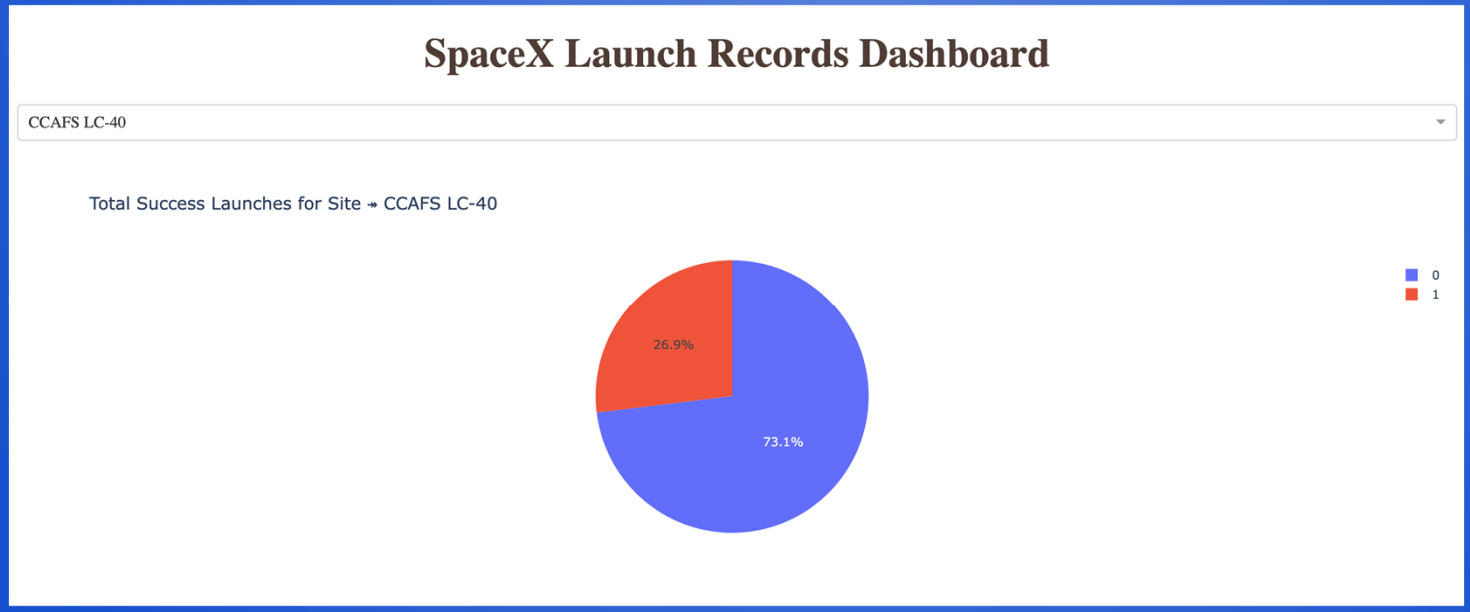
The picture above shows the distance between the VAFB SLC-4E launch site and the nearest coastline


Section 4
Build a Dashboard
with Plotly Dash

Total Launch Successes: All Sites



Site with Highest Launch Success



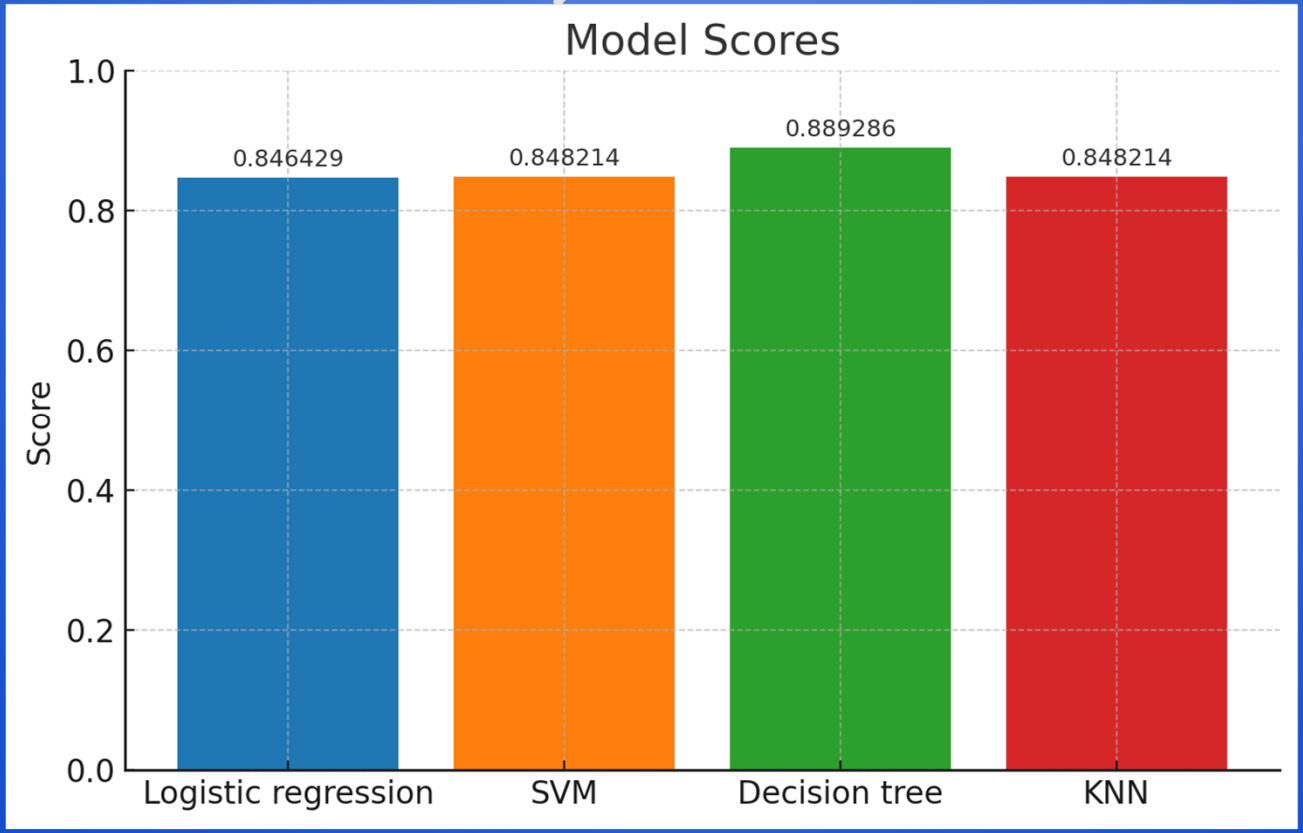
0 = failed launches, 1 = successful launches. CCAFS LC-40 is the most successful site with 73.1% launch success rate

Payload vs Launch Success

Section 5

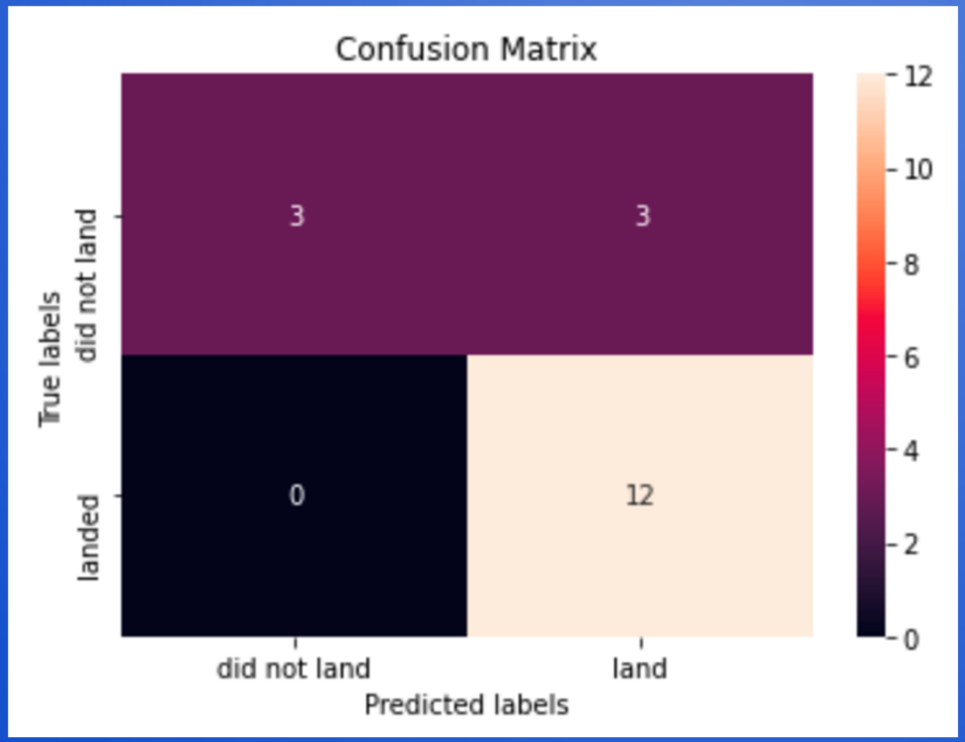
Predictive Analysis (Classification)

Classification Accuracy



Green bar = Decision Tree = Highest Classification accuracy

Confusion Matrix – Best Performing Model – Decision Tree



Confusion Matrices for all models were the same, indicating all model were very close to one another in terms of performance

Conclusions

We can conclude that we can predict Falcon 9 first stage will land successfully with 88% accuracy based on features in the launch data collected

Predictive models all performed very similarly. This could be due to the relatively small data set available for study

Prediction accuracy and model performance could be improved with additional data. Once more data is available, all predictive models should be re-evaluated to see if there is a true difference in the predictive power.