Winning Space Race with Data Science

Chris Lynch

August 09, 2025

Outline

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

Executive Summary

- Methodologies:
  - Data Collection through SpaceX API and Web Scraping
  - Data Wrangling and processing
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Visual Analytics
  - Machine Learning Prediction
- Results:
  - Exploratory Data Analysis results
  - Interactive dashboard and maps results
  - Predictive Analytics results

Introduction

- SpaceX drastically cuts launch costs by reusing Falcon 9 first stages
- A single Falcon 9 launch: ~$62 million vs. traditional providers: ~$165 million+
- Successful first stage landings are critical to cost savings and mission success
- Predicting landing success can:
  - Optimize mission planning, aid competitors in bidding strategies, and support risk assessment and cost forecasting
- Objective: Predict whether a Falcon 9 first stage will successfully land
- Answer key questions:
  - What factors most influence successful landings?
  - How accurately can we make landing predictions?
  - Can this model support real-world aerospace business decisions?

Section 1

Methodology

Methodology

Executive Summary

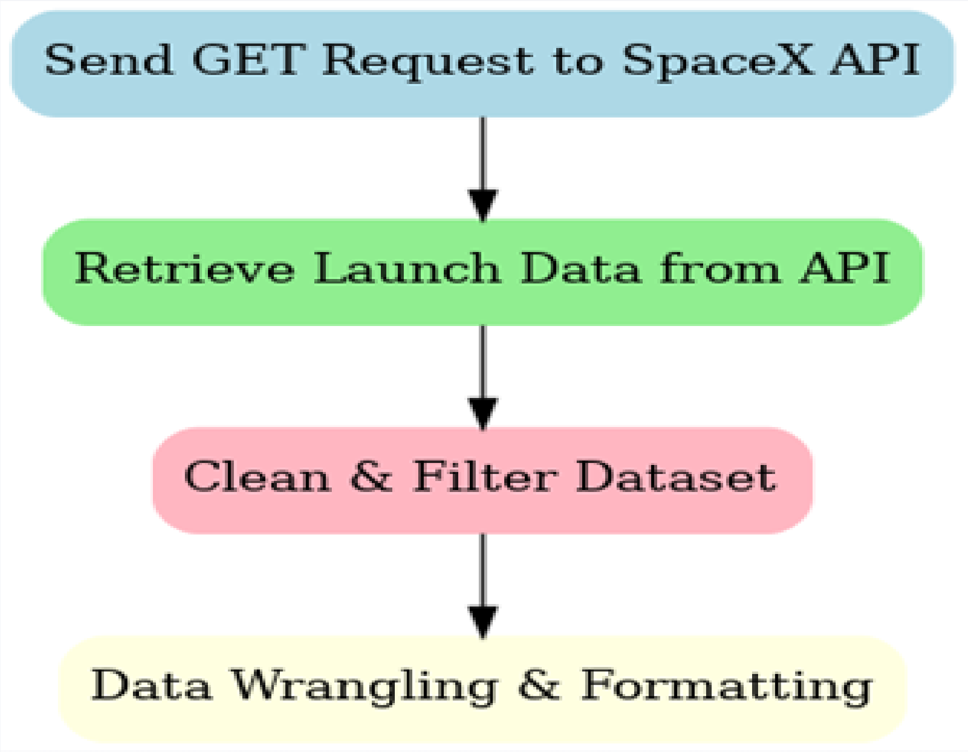
- Data collection methodology:
  - Data collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Data processed using one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Retrieved launch data via GET requests to the SpaceX API
- Parsed JSON responses into Pandas DataFrames using .json() and .json_normalize()
- Cleaned datasets, handled missing values, and performed data validation
- Scraped Falcon 9 launch records from Wikipedia using BeautifulSoup
- Extracted HTML tables, parsed them, and converted to DataFrames for analysis

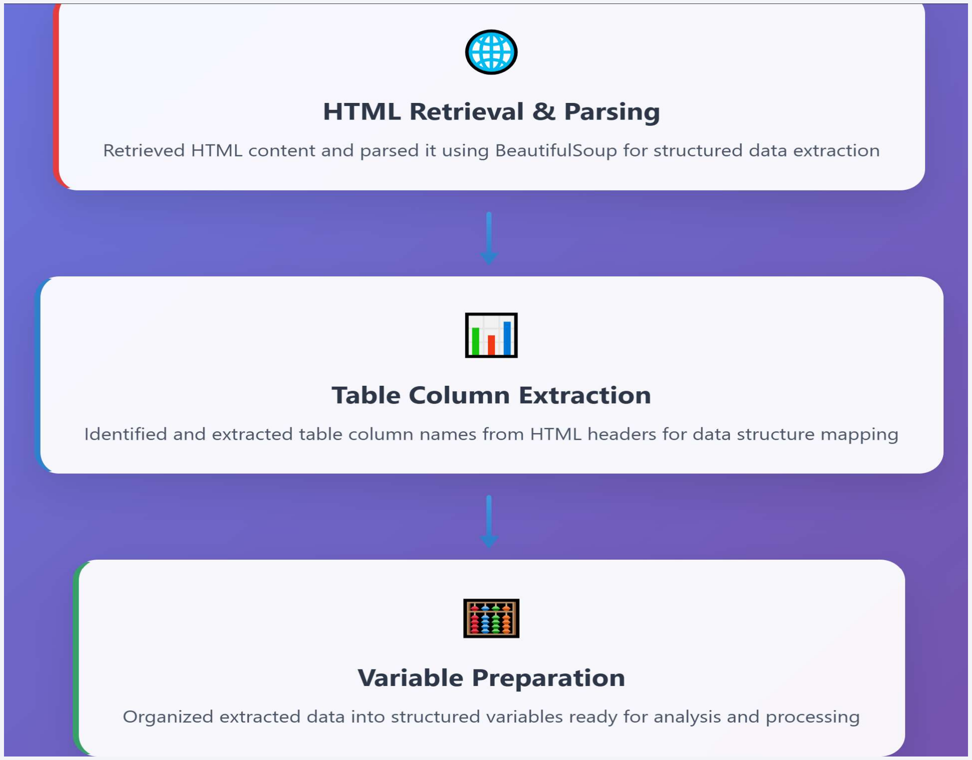
Data Collection – SpaceX API

- Sent a GET request to the SpaceX API to retrieve launch data.
- Cleaned and filtered the retrieved dataset.
- Performed basic data wrangling and formatting for analysis.
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt1%20-%20Data%20Collection.ipynb



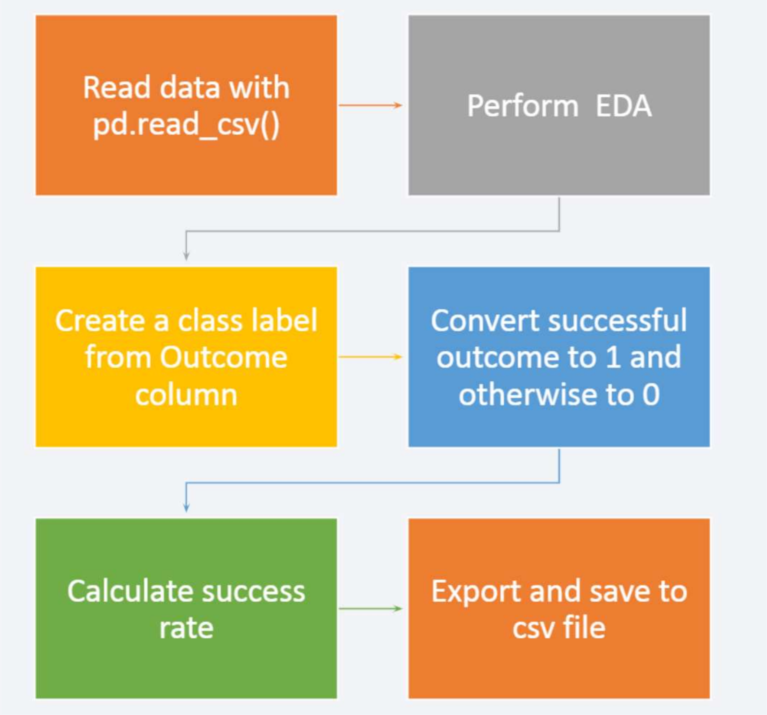
Data Collection - Scraping

- Retrieved Falcon 9 Launch Wikipedia page HTML and parsed it using BeautifulSoup
- Extracted dataset structure by identifying all table column names from HTML headers
- Prepared structured variables for further data processing and analysis
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt2%20-%20Web%20Scraping.ipynb



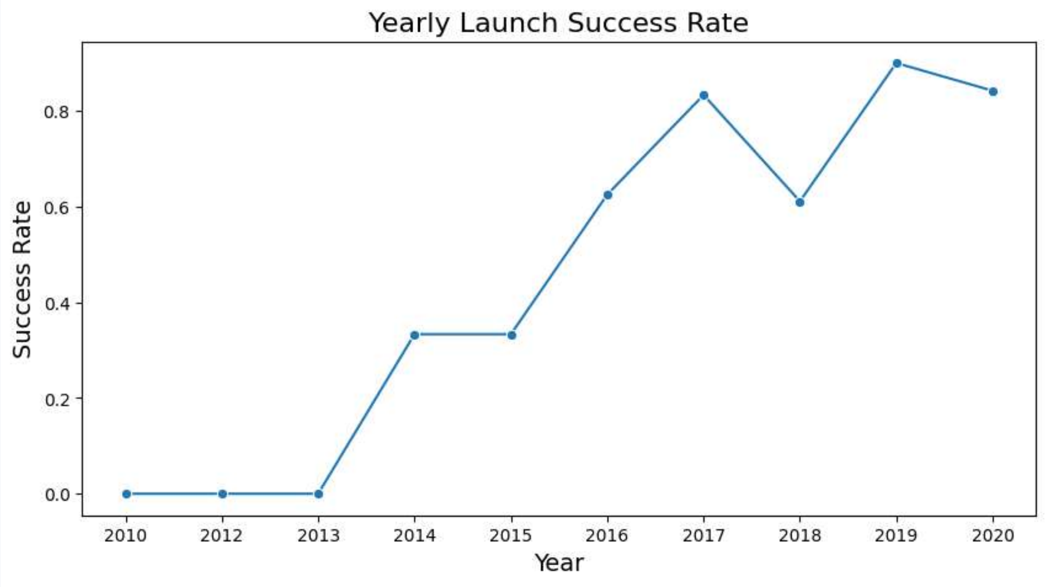
Data Wrangling

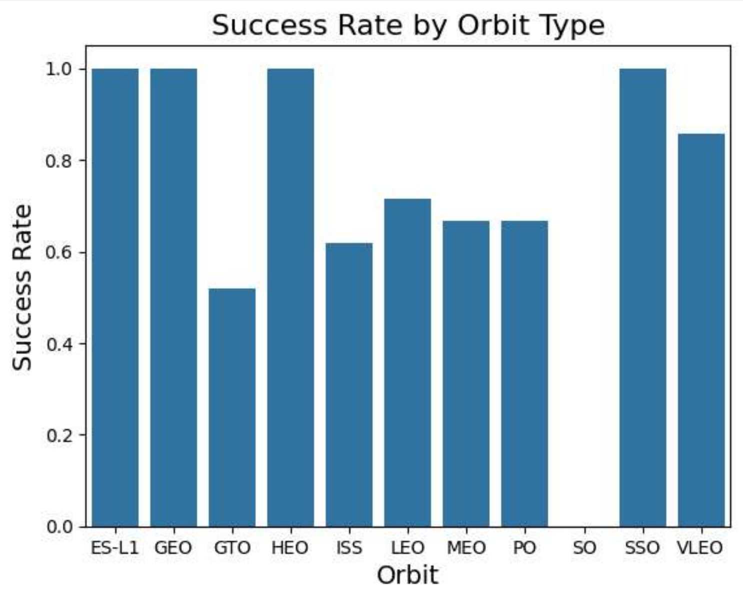
- Exploratory Data Analysis on the data
- Convert outcomes to training labels of 0/1
- Calculate success and export
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt3%20-%20Data%20Wrangling.ipynb



EDA with Data Visualization

- The data was analyzed through visualizations examining key relationships: flight number vs launch site, payload vs launch site, success rates across different orbit types, flight number vs orbit type, and launch success trends over time.





https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt5%20-%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- **SQL queries were executed on the dataset to extract:**
  - Distinct launch site locations
  - Total payload mass for NASA Commercial Resupply Services missions
  - Mean payload capacity for Falcon 9 v1.1 booster
  - First successful ground pad landing date
  - Success and failure mission counts
  - Booster variants with maximum payload capacity
  - Landing outcome frequency distribution
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt4%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- **Interactive mapping visualization implemented using Folium components:**
  - **Circle markers** - Created highlighted circular regions centered on NASA JSC location
  - **Location markers** - Placed individual markers at specific launch site coordinates
  - **Clustered markers** - Grouped successful and failed launch indicators by geographic proximity
  - **Coordinate display** - Enabled mouse hover functionality to show latitude/longitude and calculate coastal distances
  - **Polyline connections** - Drew connected line segments measuring distances from launch sites to coastlines, transportation routes, and urban centers
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt6%20-%20Visual%20Analytics.ipynb

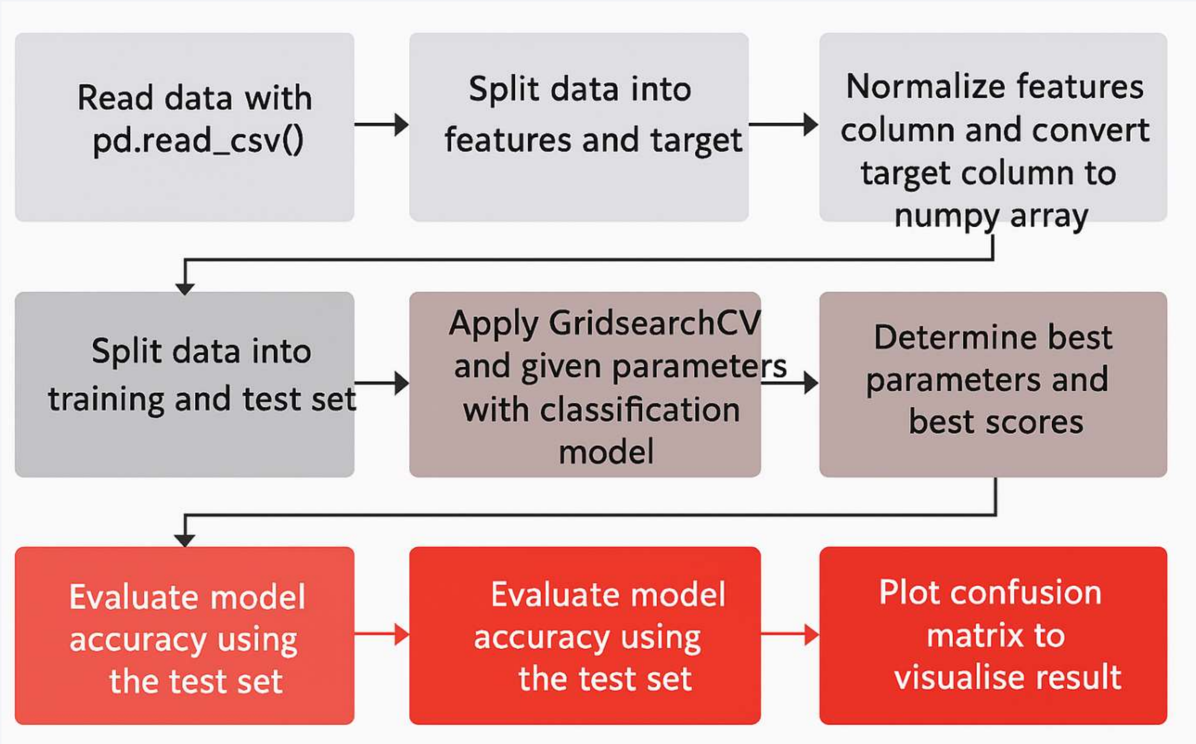
Build a Dashboard with Plotly Dash

- **Interactive dashboard components implemented:**
  - **Launch site selector** - Dropdown menu for filtering by specific launch locations
  - **Success rate visualization** - Dynamic pie chart displaying success percentages for selected launch sites
  - **Payload mass filter** - Range slider control for payload weight selection
  - **Correlation analysis** - Scatter plot revealing relationships between payload mass and mission success rates across chosen sites
- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt7%20-%20SpaceX%20Dash%20App.py

Predictive Analysis (Classification)

- **Classification model development and evaluation process:**
  - **Data preprocessing** - Dataset loaded and separated into feature variables and target variable
  - **Feature scaling** - Feature columns normalized and target variable transformed into NumPy array format
  - **Data partitioning** - Dataset divided into training and testing subsets
  - **Hyperparameter optimization** - GridSearchCV applied across all classification algorithms to identify optimal parameters and performance scores via .best_params_ and .best_score_ attributes
  - **Performance assessment** - Test set accuracy evaluated using .score() method
  - **Result visualization** - Confusion matrix generated to display classification outcomes

Predictive Analysis (Classification)



- https://github.com/chrislynch4/DS-Capstone---Coursera/blob/main/Capstone%20pt8%20-%20Machine%20Learning.ipynb

Results

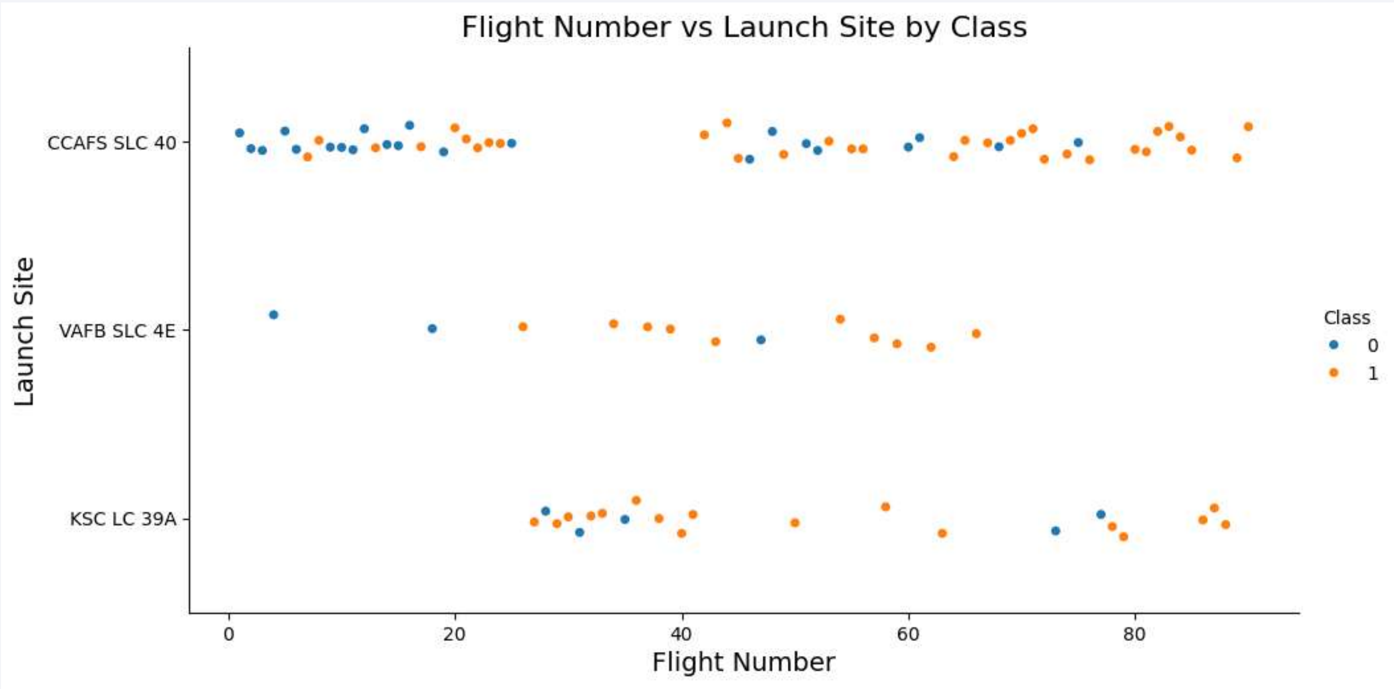
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

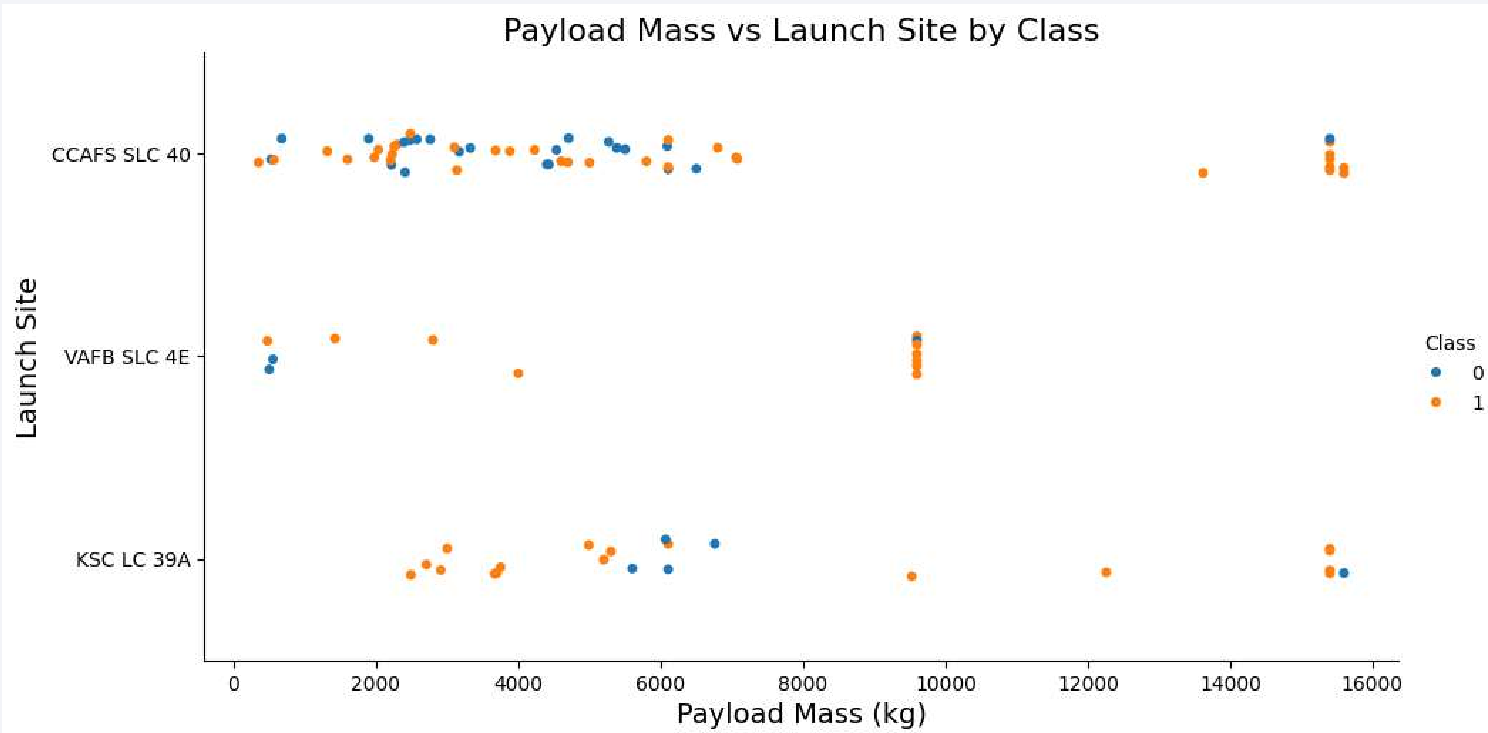
Flight Number vs. Launch Site

- The scatter plot reveals a positive correlation between flight frequency and success rates at launch sites - higher flight volumes generally correspond to improved success rates. CCAFS SLC40, however, demonstrates the weakest adherence to this trend among all launch facilities.



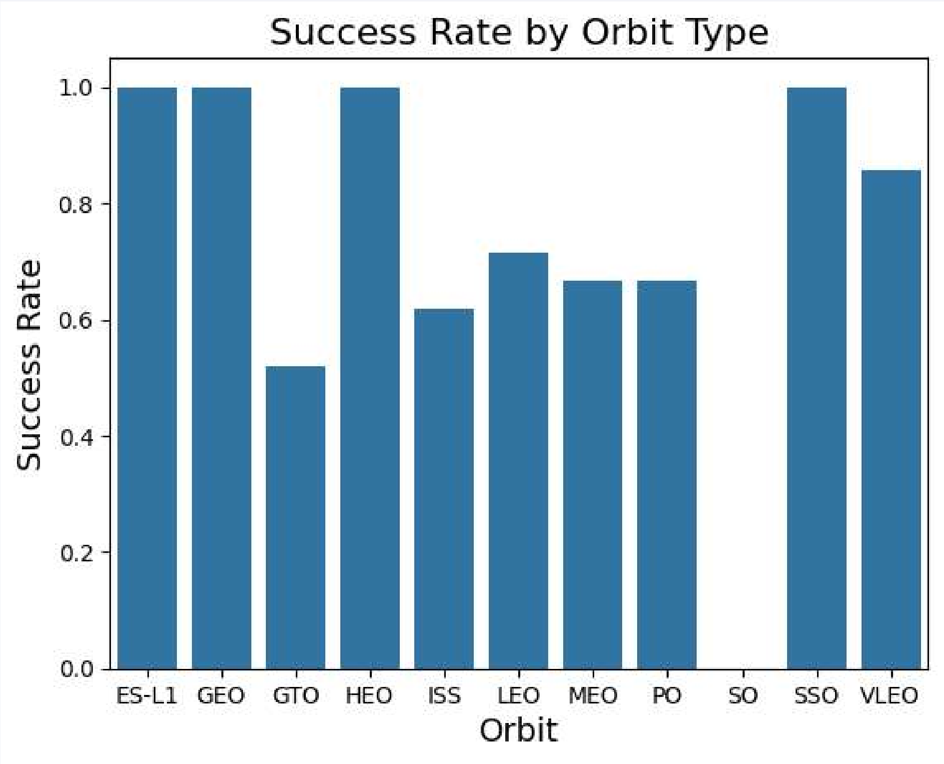
Payload vs. Launch Site

- The scatter plot analysis indicates that payload masses exceeding 7000kg significantly enhance mission success probability. Conversely, no definitive correlation exists between launch site selection and payload mass regarding success rate outcomes.



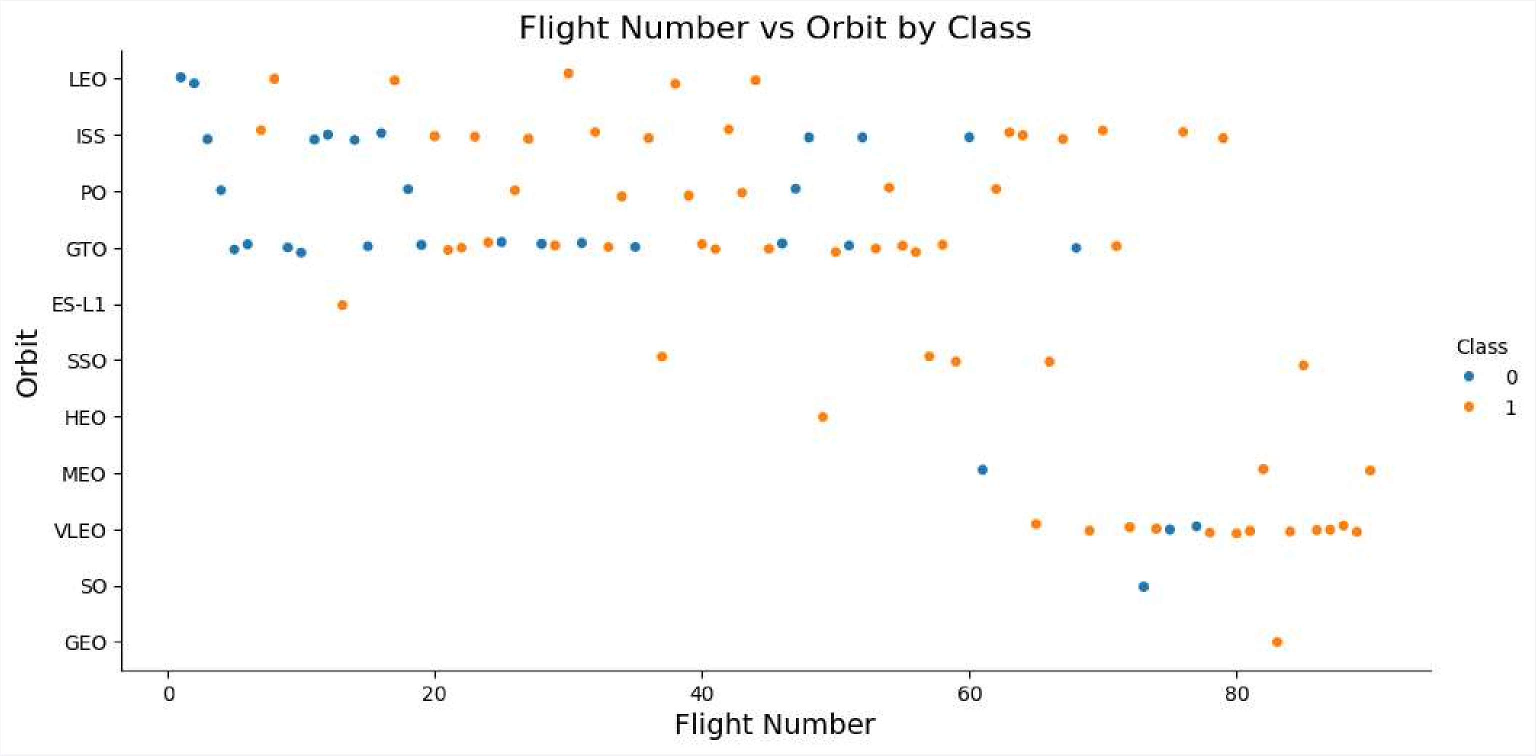
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits demonstrate 100% mission success rates.
- Other orbital destinations exhibit success rates varying from 50% to 90%.
- SO orbit is an exception, recording zero successful missions.



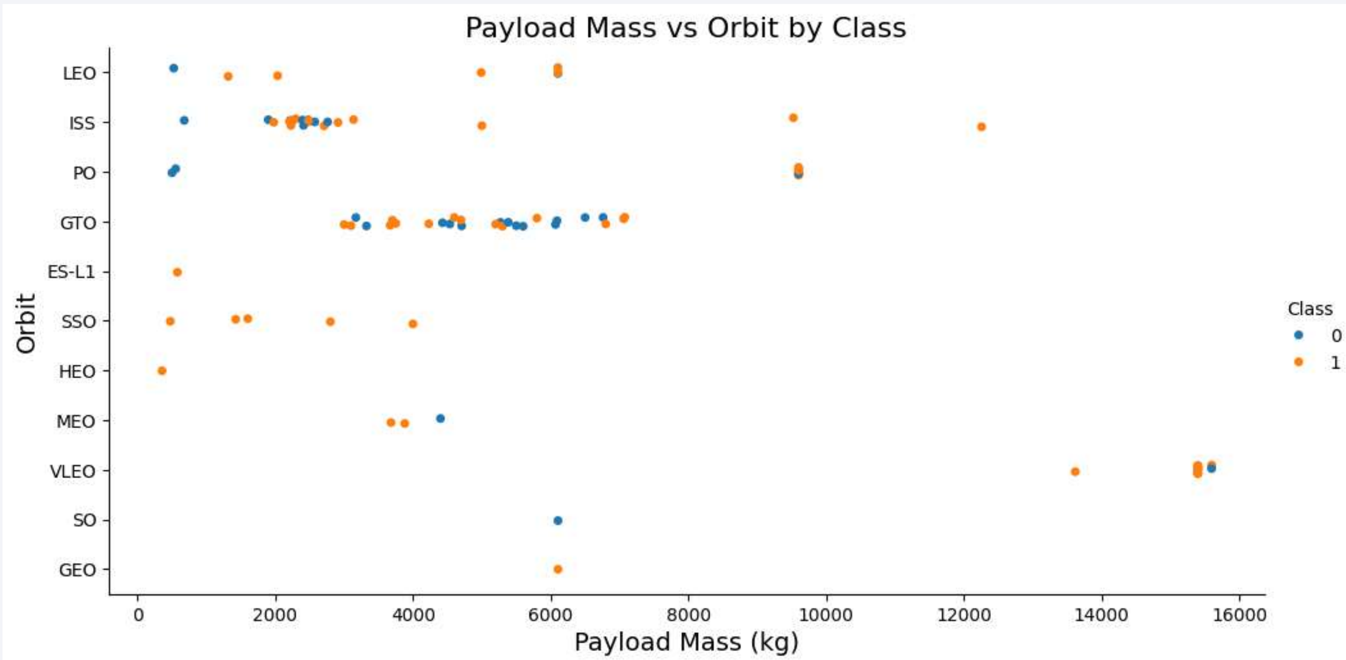
Flight Number vs. Orbit Type

- LEO orbit missions show improved success rates with higher flight sequence numbers.
- GTO orbit missions display no correlation between flight number and success rate.



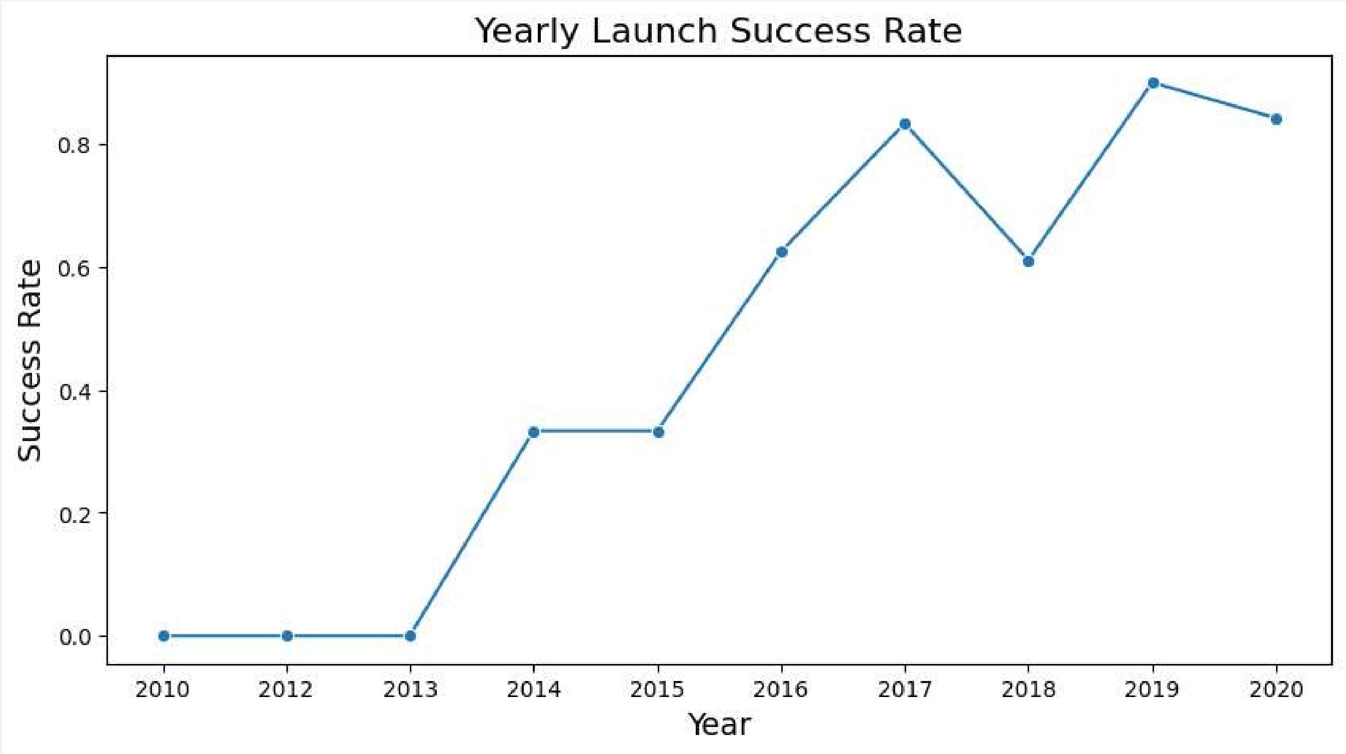
Payload vs. Orbit Type

- Increased payload mass correlates with higher success rates for LEO, PO, and ISS orbital missions.
- GTO orbital missions show no correlation between payload mass and mission success.



Launch Success Yearly Trend

- Success rate has increased over the years overall
- Some drop in 2018



All Launch Site Names

- The DISTINCT keyword was applied to display only unique launch sites from the SpaceX dataset.

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

avg_payload_mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

first_successful_ground_pad_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Landing_Outcome	outcome_count
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

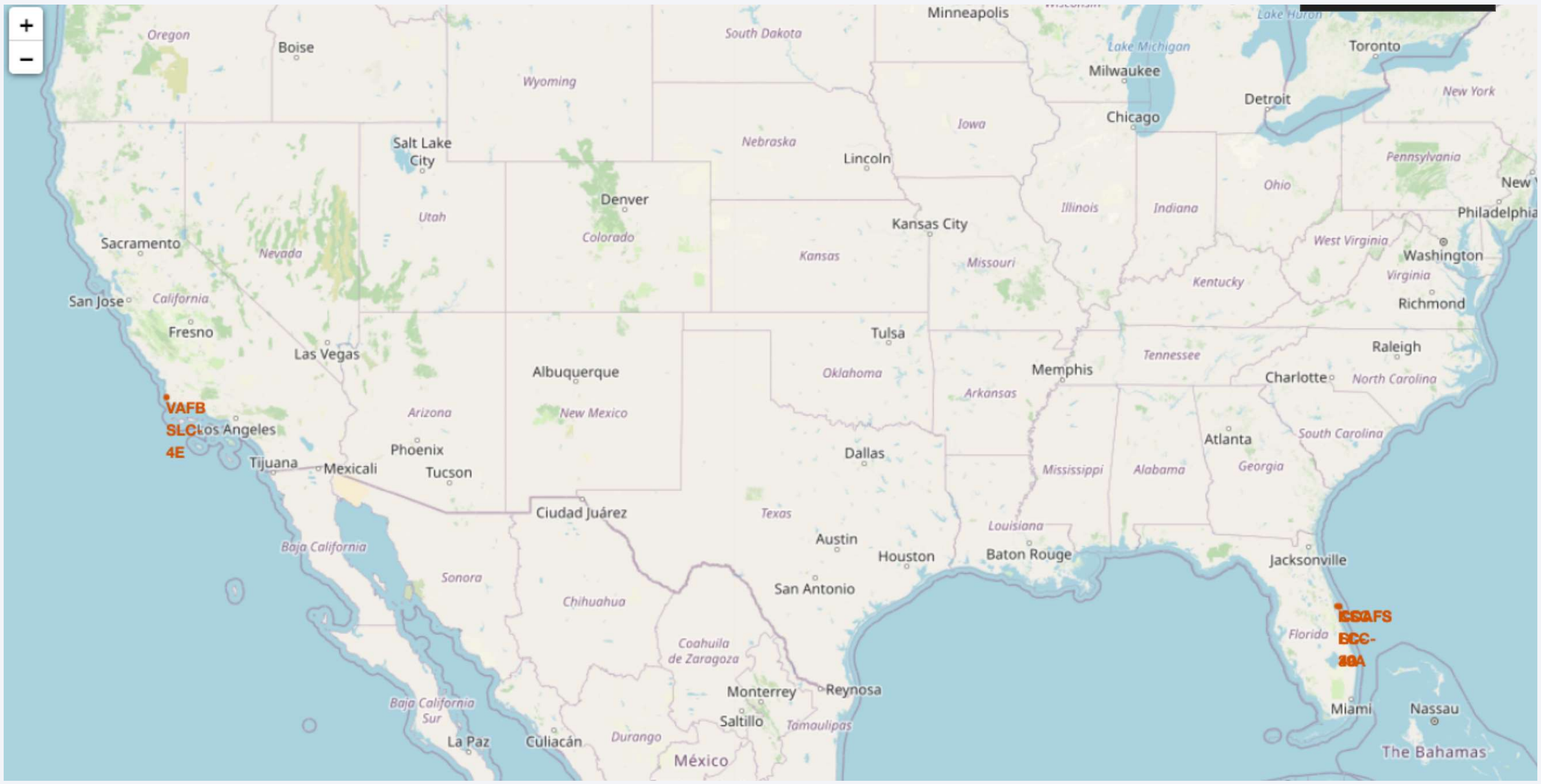
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

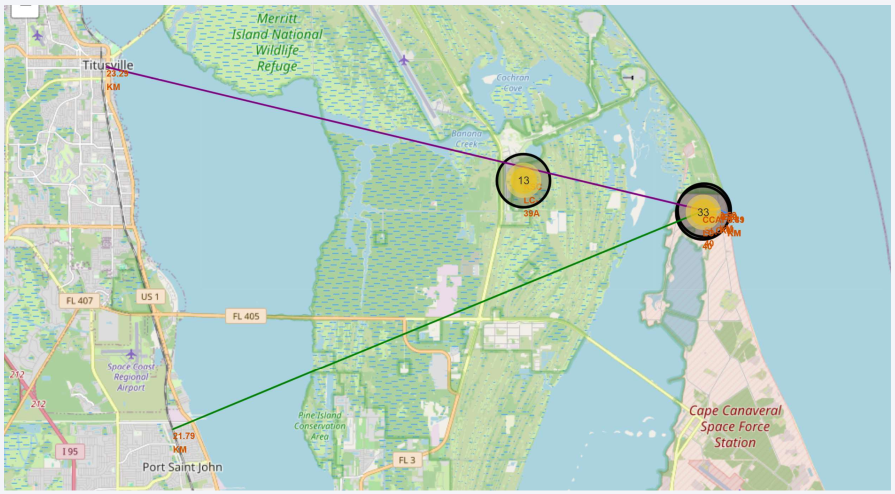
Folium Map: Launch Sites

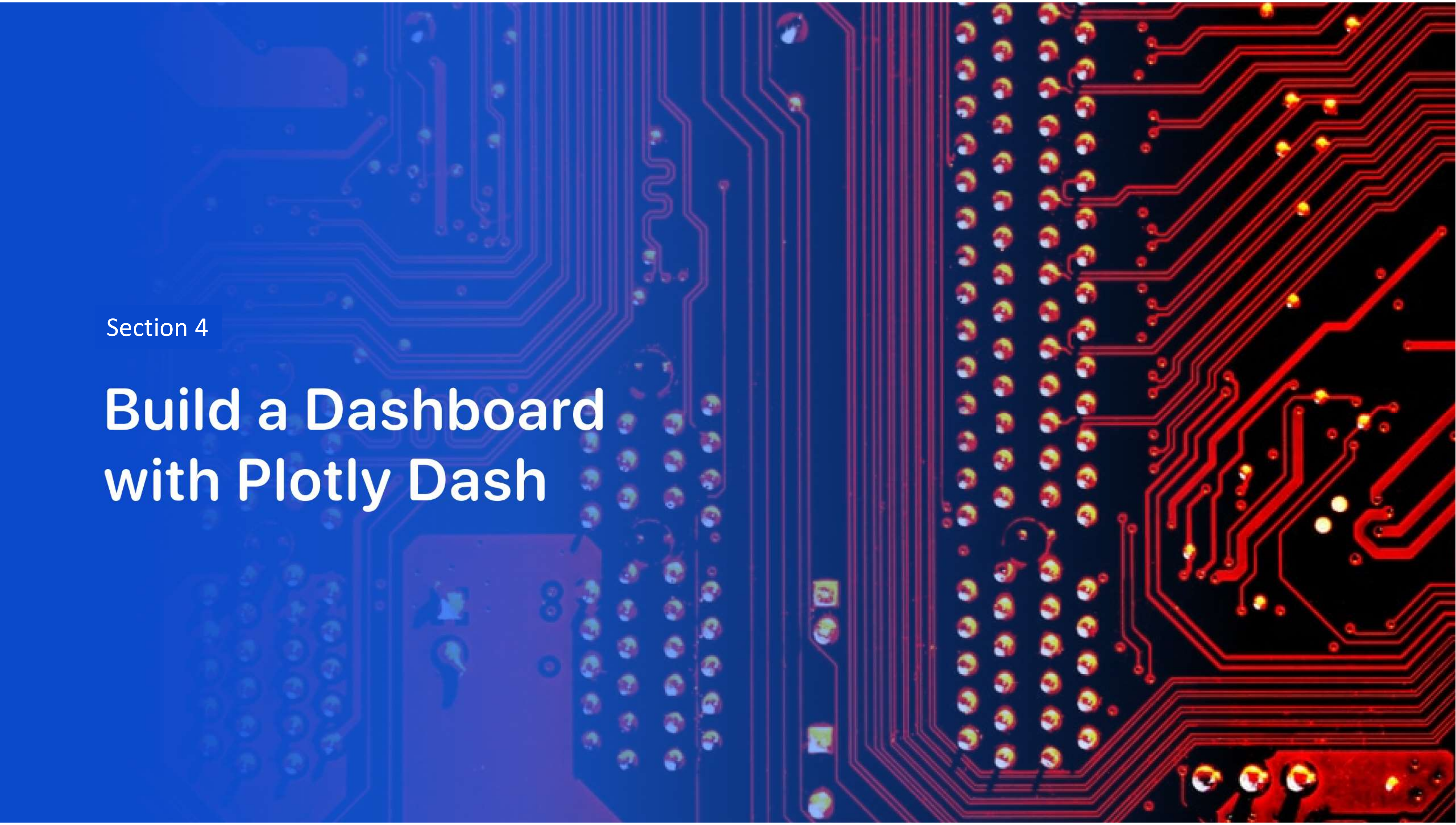
- The launch sites are on the coasts of California and Florida



Folium Map: Launch Outcomes

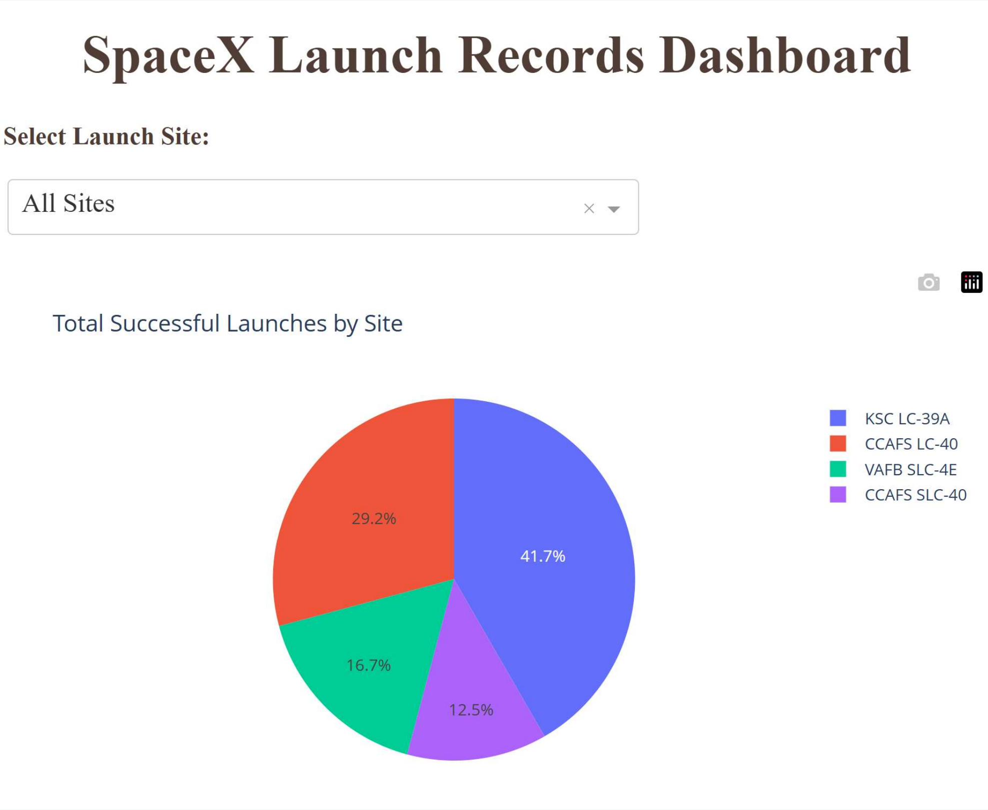
Folium Map: Launch Site Proximities




Section 4
Build a Dashboard with Plotly Dash

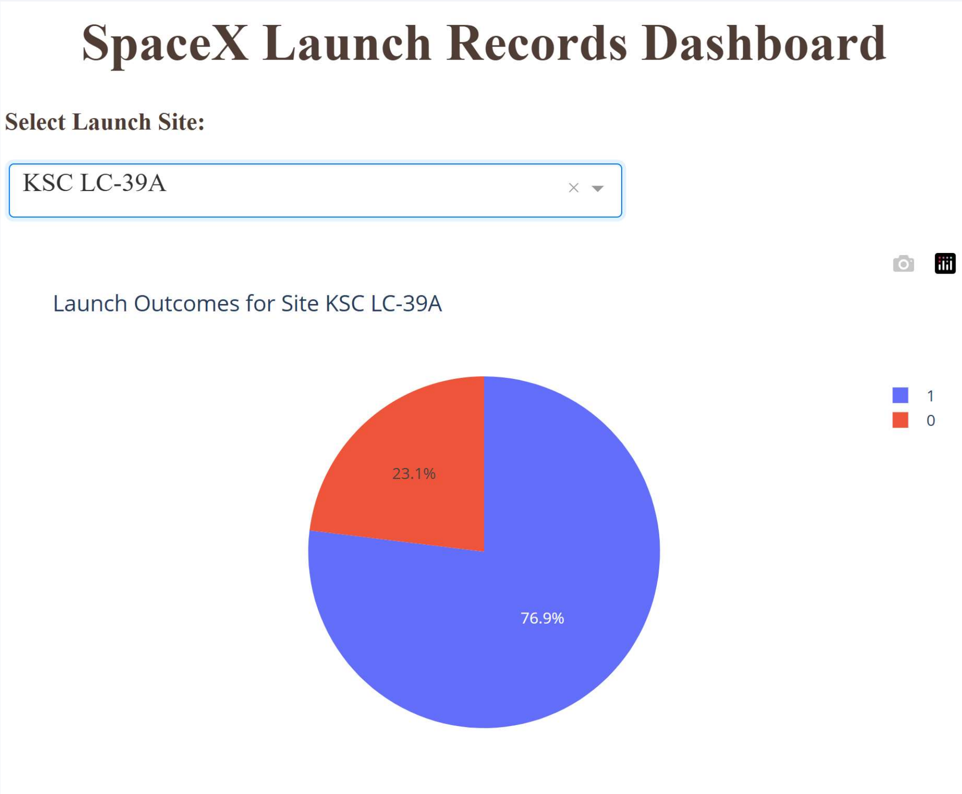
Dashboard: All Sites Success Rate

- As illustrated, KSC LC-39A leads in launch success performance, contributing about 41.7% to the overall success rate among all launch sites



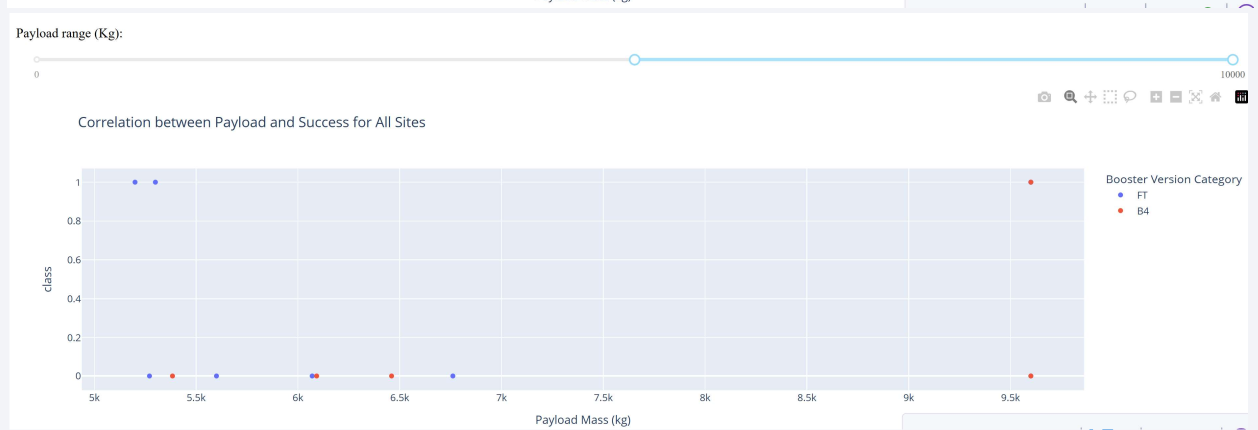
Dashboard: Highest Success Rate

- The data in shows KSC LC-39A's dominance with approximately 76.9% success rate, surpassing other launch facilities:
  - CCAFS LC-40: 73.1%
  - VAFB SLC-4E: 60%
  - CCAFS SLC-40: 57.1%



Dashboard: Payload vs Launch Outcome

- Based on the figures, the FT booster version demonstrates optimal performance within the payload mass range of approximately 700kg to 5,500kg.
- The data also reveals that launches carrying payloads exceeding 5,500kg experience reduced success rates, indicating an inverse relationship between payload weight and mission success probability.

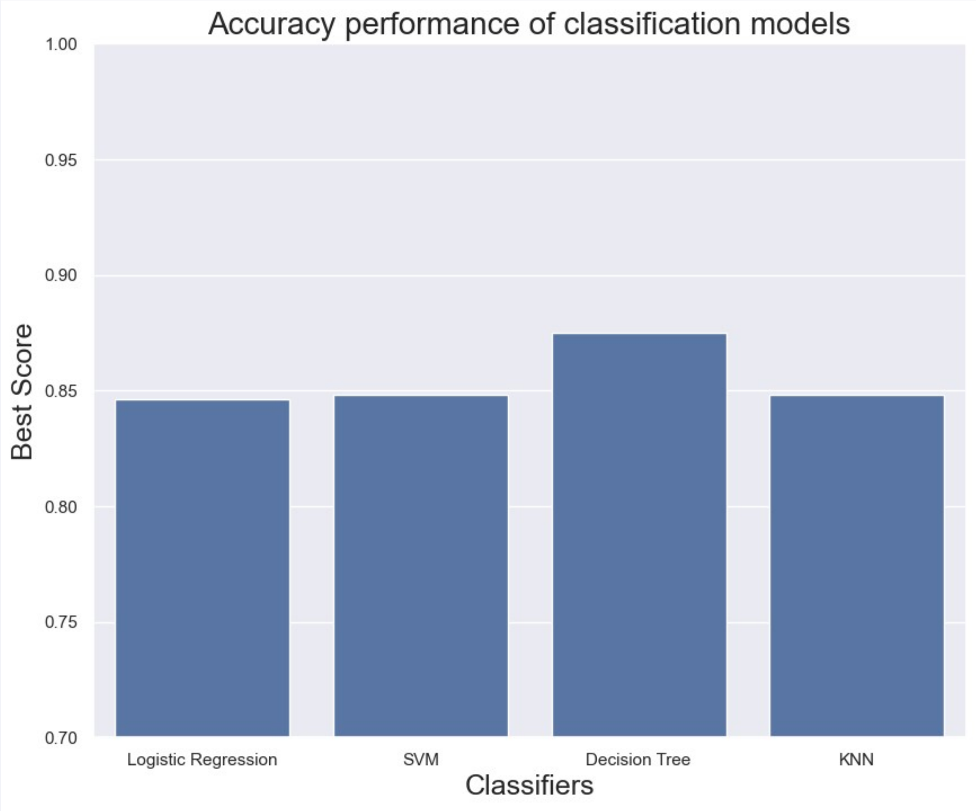


Section 5

Predictive Analysis (Classification)

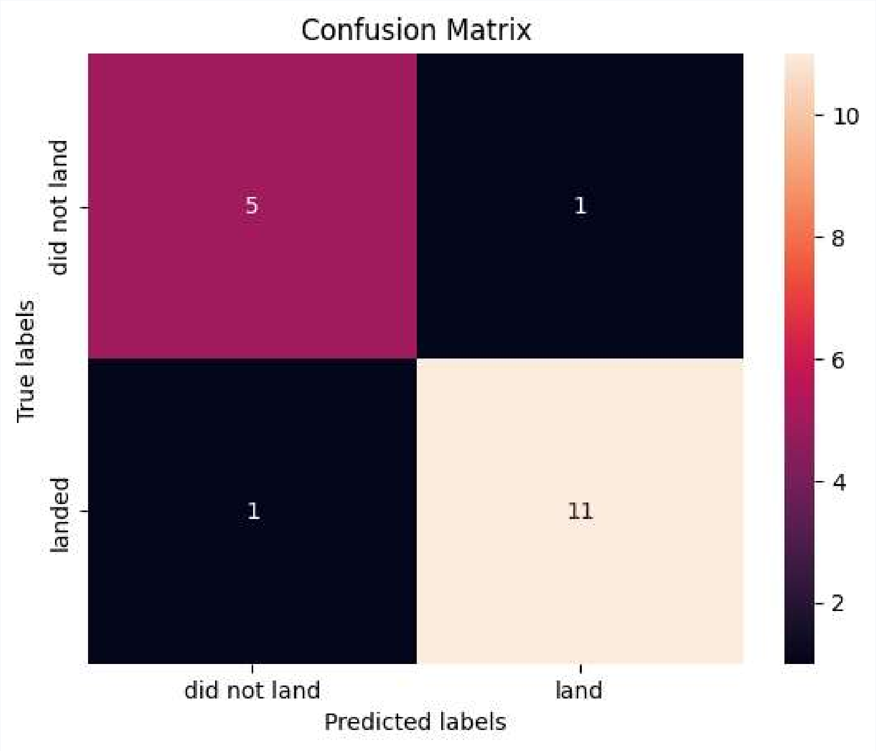
Classification Accuracy

- The Decision Tree classifier performed the best with an accuracy score of 88.8%



Confusion Matrix

- Following the train-test split of the dataset, the test set contained 18 samples for evaluation.
- The decision tree classifier demonstrated strong performance on the test set, achieving 11 correct predictions for successful landings (true positives) and 5 accurate predictions for unsuccessful landings (true negatives).



Conclusions

- As SpaceX's landing success rates continue to improve, launch costs are expected to decrease significantly due
- Several factors contribute to higher mission success rates: lighter payloads demonstrate better performance, specific orbit types (VLEO, ES-L1, GEO, HEO, and SSO) show superior success rates, and newer booster technologies correlate with improved landing outcome**s.**
- Launch facilities are strategically positioned in coastal locations to facilitate booster recovery operations while maintaining safe distances from populated areas and major transportation corridors to reduce potential casualties during mission failures.
- Recent launch data indicates consistently improving success rates over time, with more recent missions achieving higher performance levels.
- Among various machine learning approaches tested, decision tree classifiers achieved optimal performance with approximately 89% accuracy, making them highly effective for predicting landing outcomes.
- To enhance model accuracy and reduce false positive rates, additional post-2021 data should be incorporated for training and validation of predictive models.


Thank you!