IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ian McMeeking
August 26, 2022

Outline

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

Executive Summary

Data was scraped from Wikipedia and the SpaceX API on the launch characteristics and results of many Falcon 9 launches. This data was cleaned and then analyzed using numerous data visualizations, SQL queries, and classification algorithms to determine factors that impact the success of the Falcon 9's reusable first stage booster.

It was found that **launch date** was a strong predictor of a successful landing, with later years having higher success rates. **Heavier payloads** and launches at the **Kennedy Space Center** also correlated with higher success rates, but this may be due to the fact that heavier payloads and launches at KSC occurred more frequently in later years. Some of the most successful rocket boosters are those of the **FT category**, which have a high success rate with medium-sized payloads ranging from 1,000 to 6,000 kg. A number of classification algorithms (**KNN, logistic regression, and SVM**) were found to predict the results of launches with a high accuracy of 80-85%.

Introduction

As private companies become more and more involved in spaceflight, SpaceX has emerged as an industry leader. One of SpaceX's important innovations is the Falcon 9 rocket, a two-stage rocket with a reusable first stage. Successful repeated use of the first stage leads to huge cost reductions, so it is important for SpaceX and competitors to understand how successfully the first stage can be reused and what factors influence its success. Using data from the SpaceX API and Wikipedia, the following questions were investigated:

- ➢What factors influence the success of the first stage?
  Examples include launch date, payload mass, orbit type, launch location, booster version, etc…

- ➢How accurately can the success of the first stage be predicted?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API, web scraping Wikipedia
- Perform data wrangling
  - Handling missing data
  - One-hot encoding of categorical variables
  - Coding launch outcomes as successes and failures
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Develop models with KNN, decision tree, logistic regression, and SVM algorithms
  - Optimize models with grid search
  - Compare models with accuracy scores and confusion matrices

Data Collection

Data was collected from two key sources:

1. The SpaceX API, and
2. Wikipedia

Data Collection – SpaceX API

- **Initial API Call**
  - Request data on all Falcon rocket launches
- **Secondary API Calls**
  - Use ID codes from initial call to request more information on rocket, payload, launchpad, and cores
- **Filter and Clean Results**
  - Filter out launches that did not use Falcon 9 rocket
  - Fill in missing payload masses with the mean

GitHub:
1. Data Collection API Lab

Data Collection - Scraping

Request Page by URL

- Get the HTML for the Wikipedia page on Falcon 9 launches
- Create a Beautiful Soup object to parse

Extract Tables and Headers

- Find all tables of rocket launch data
- Extract headers to set up columns of a Pandas DataFrame

Populate Dataframe

- Extract data from table cells to populate Pandas DataFrame

GitHub:
2. Data Collection With Web Scraping

Data Wrangling

Survey Data Types & Values

- Check the number of missing values and data type for each column
- Check number of launches per site and orbit type

Identify Types of Landings

- View all values of landing_class column
- Categorize each value as success/failure

Create Target Variable

- Based on landing_class, create new column of binary success or failure values (1 or 0)

GitHub:
[3. Data Wrangling]

EDA with SQL

- Survey launch sites, payload masses, and booster versions
- Check dates and booster versions of successful and failed landings
- View the most common landing outcomes

GitHub:
[4. Exploratory Data Analysis With SQL]

EDA with Data Visualization

- Categorical scatterplots were used to visualize how flight number, payload mass, and launch site interact with landing outcomes
- A bar chart was used to visualize the success rate of launches with different orbit types
- A line chart was used to show the trend of more successful launches in later years

GitHub:
5. Exploratory Data Visualization

Build an Interactive Map with Folium

- Circles and markers were added to a Folium map showing the locations of the 4 launch sites
- Clusters show each successful and failed launch at each site
- Annotated lines show the distance from launch sites to nearby features, specifically coastlines and cities

GitHub:
6. Launch Site Locations (Folium)

Build a Dashboard with Plotly Dash

- Using a dropdown, data can be viewed for a single launch site or all launch sites
- A pie chart shows the proportion of successful launches to easily compare each launch site
- A scatter plot shows the payload mass, booster version category, and outcome of each launch
  - A range slider can be used to set the range of payload masses displayed

GitHub:
7. Interactive Dashboard (Plotly Dash)

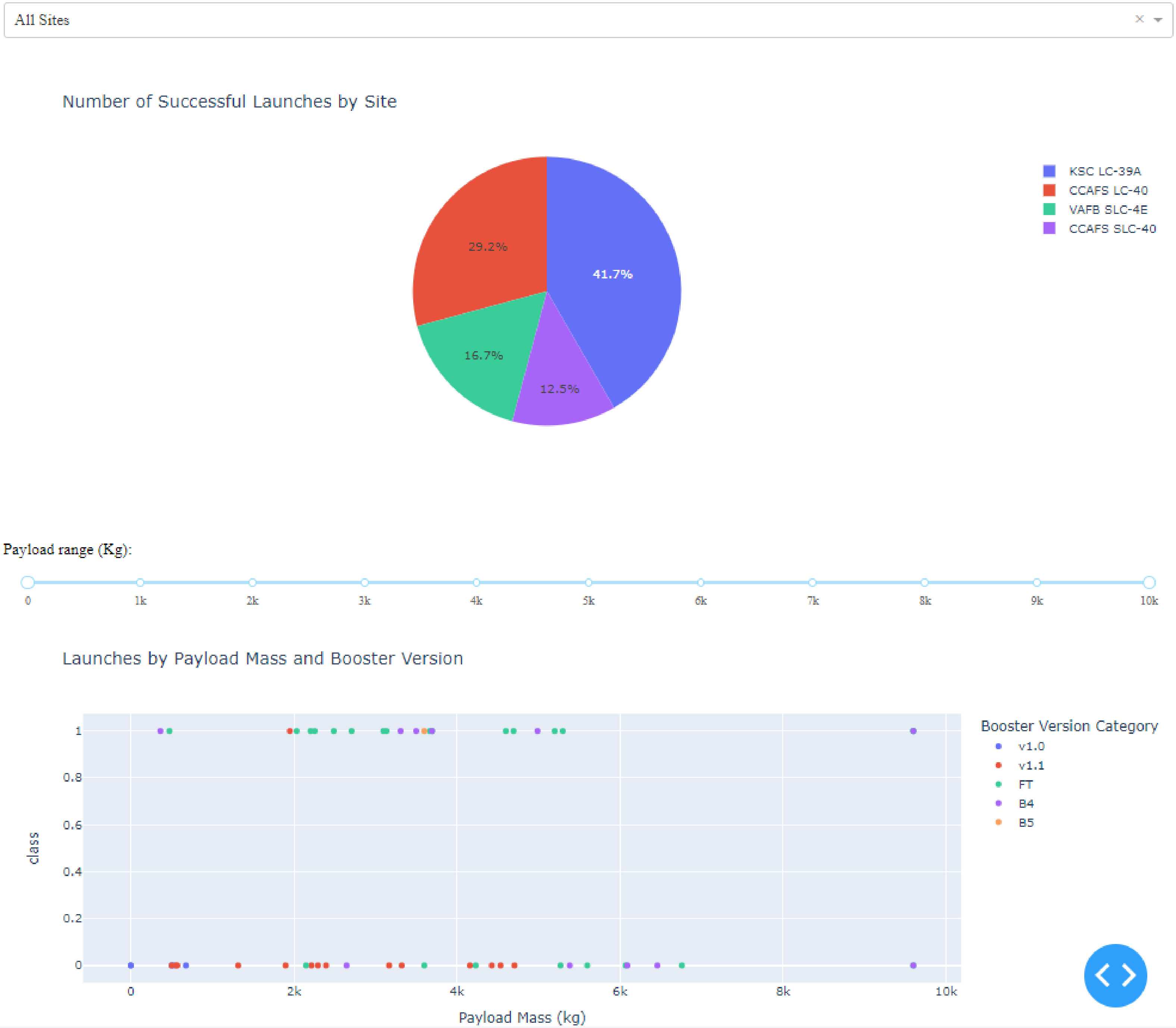
Predictive Analysis (Classification)

Standardize and Split Data	Develop Models	Compare Models
• Standardize each feature variable • Split the data into a training set and testing set	• Employed KNN, logistic regression, SVM, and decision tree algorithms • Grid search was used to select optimal hyperparameters	• Accuracy scores and confusion matrices were used to compare model performance

GitHub:
8. Machine Learning Predictions

Results

- Exploratory Data Analysis:
    - Success rate of launches improved from year to year
    - Heavier payloads and launches at Kennedy Space Center correlate with successful landings
    - The FT boosters had a high success rate
- Predictive Analysis:
    - Models can predict with high accuracy (80-85%) whether past launches were successful
    - 3 algorithms (logistic regression, SVM, KNN) had nearly identical performance

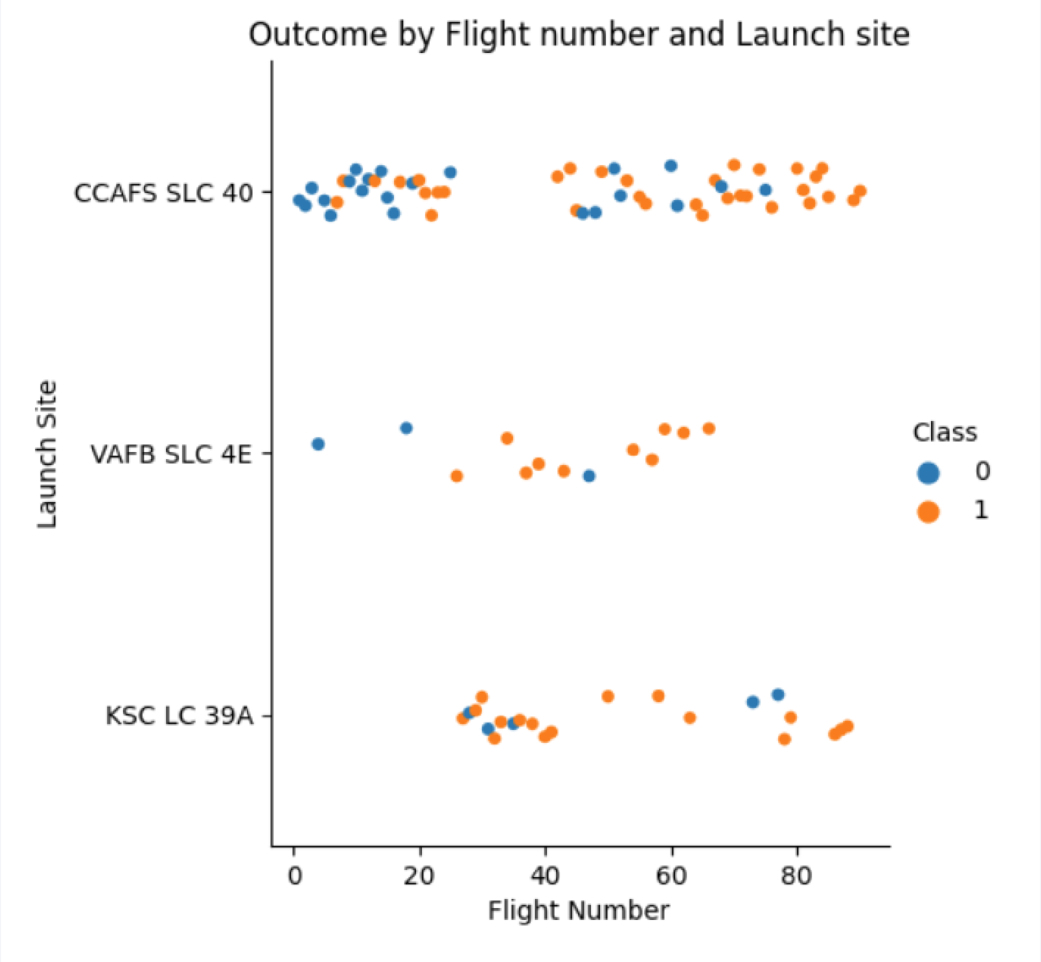


Section 2

Insights drawn from EDA

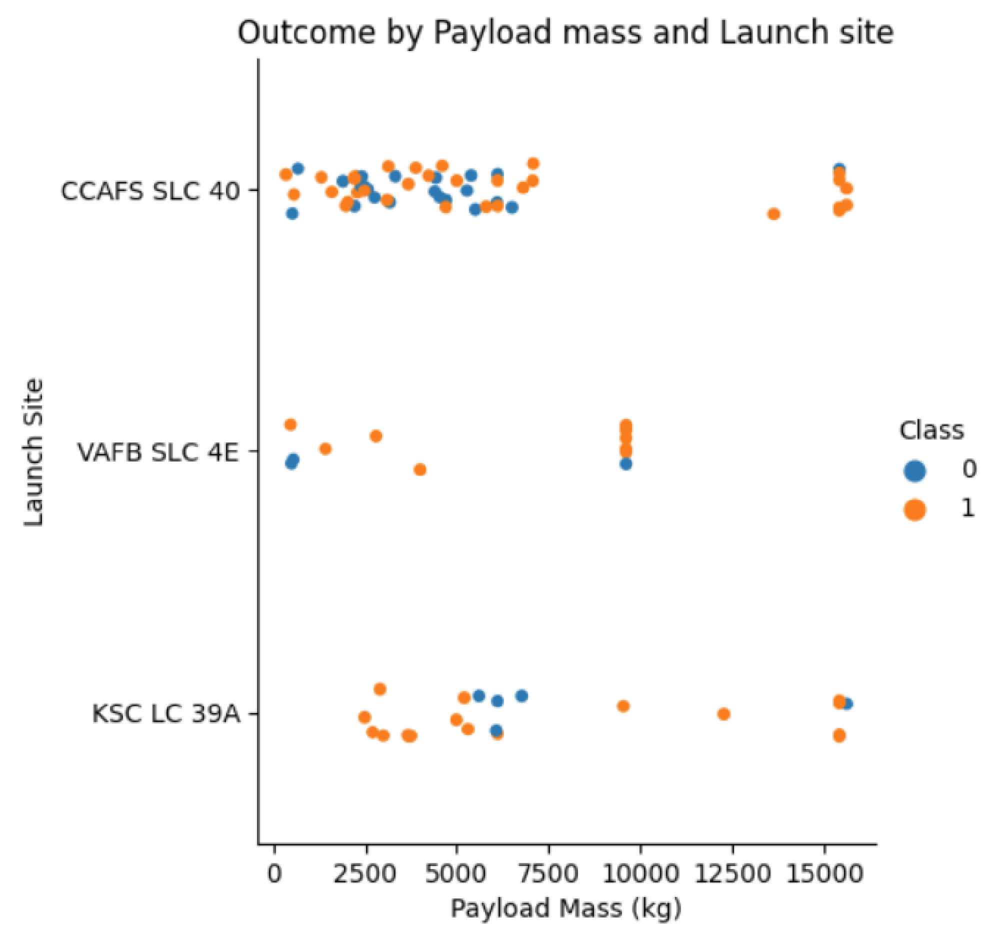
Flight Number vs. Launch Site

- Higher flight numbers correlated with a higher success rate at CCAFS SLC 40 and at VAFB SLC 4E
- CCAFS SLC 40 had the lowest success rate of all launch sites
- The proportion of successful landings is greater for higher flight numbers



Payload vs. Launch Site

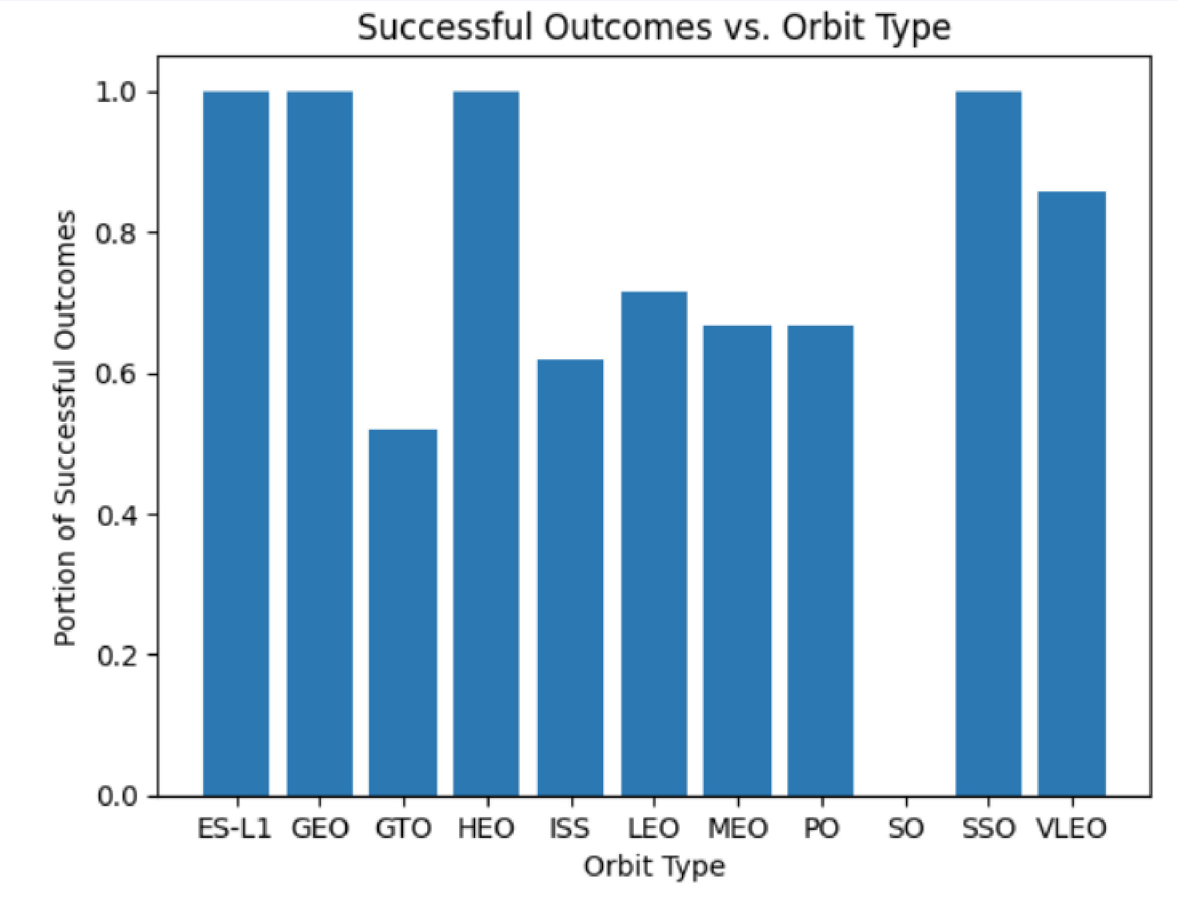
- Higher payloads were correlated with more successful launches, especially at CCAFS SLC 40
- VAFB SLC 4E had the fewest launches and, on average, the lightest payloads



Success Rate vs. Orbit Type

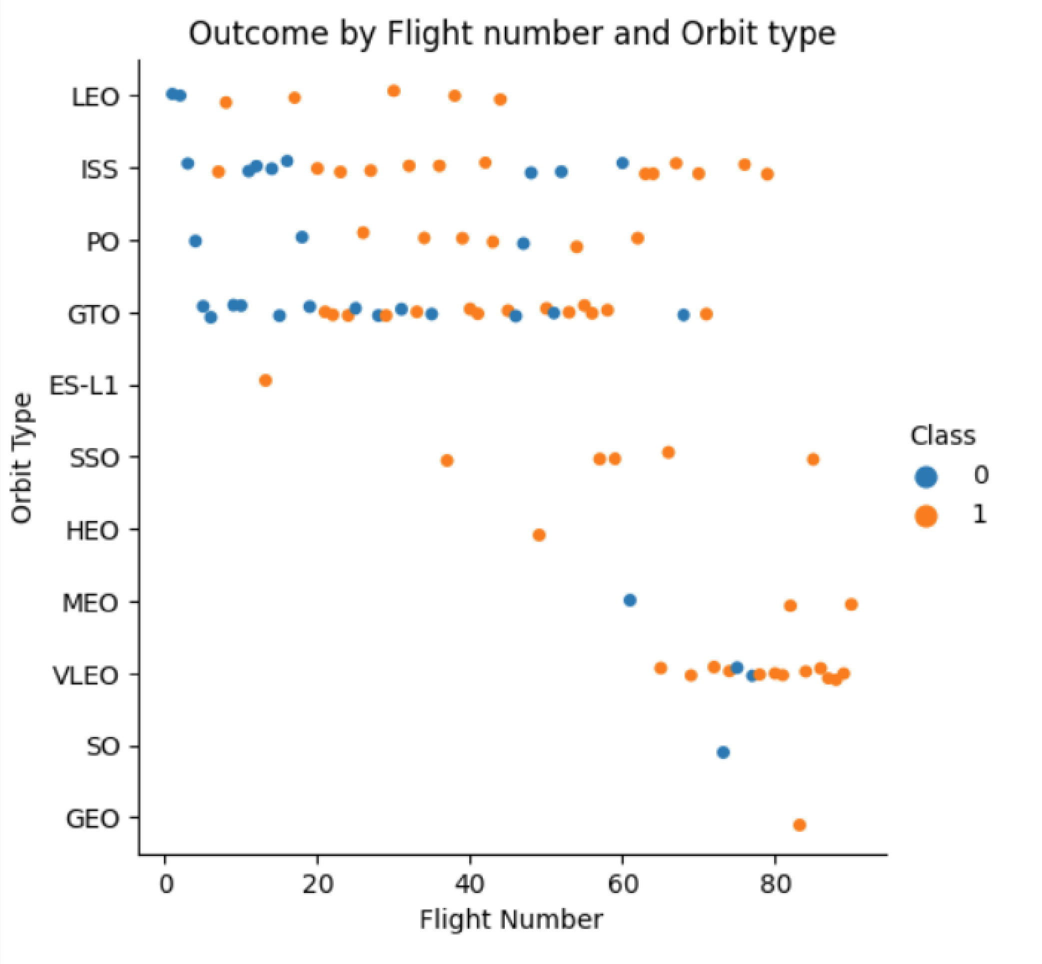
- Launches to ES-L1, GEO, HEO, and SSO all had 100% success rates
- There were no successful launches to SO

**notably, for most of these orbit types (ES-L1, GEO, HEO, and SO), there was only one attempted launch*



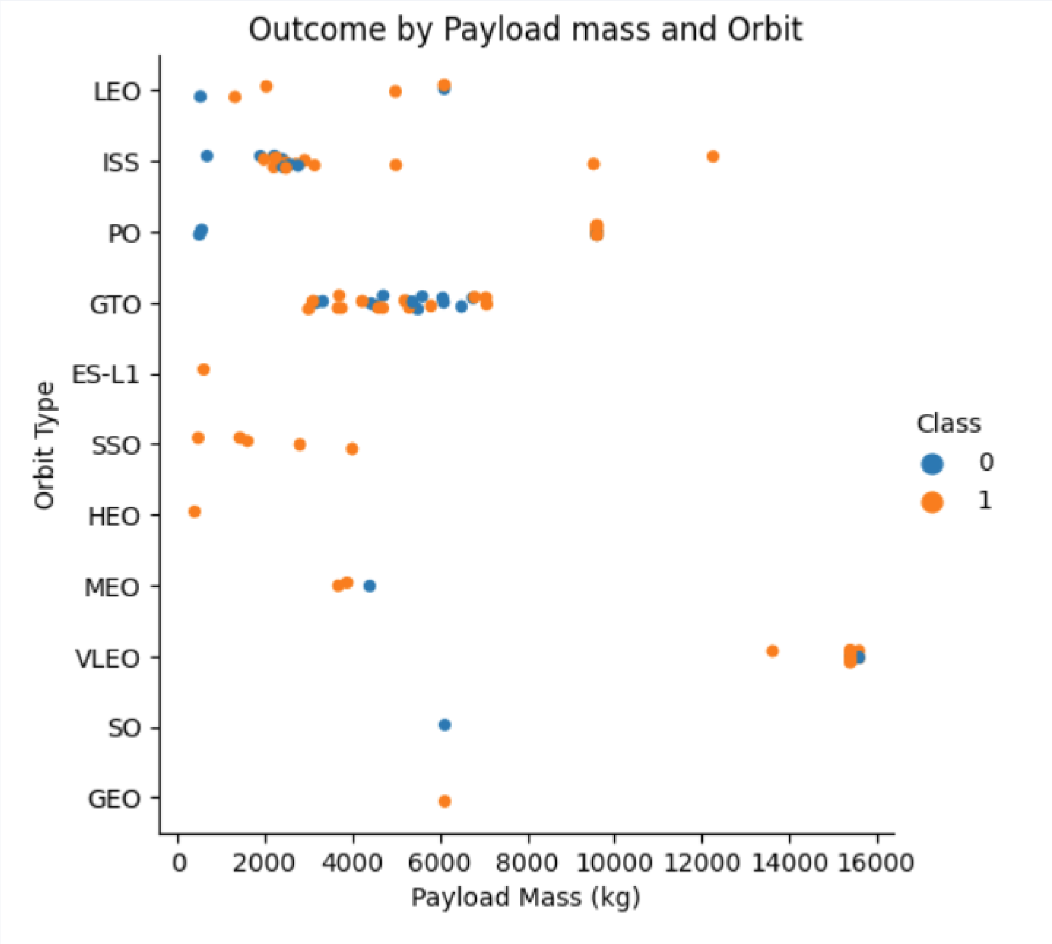
Flight Number vs. Orbit Type

- Early launches (flight number < 50) mostly aimed for 4 types of orbit: LEO, ISS, PO, and GTO
- Recently (flight number > 65), VLEO has been a popular and successful destination



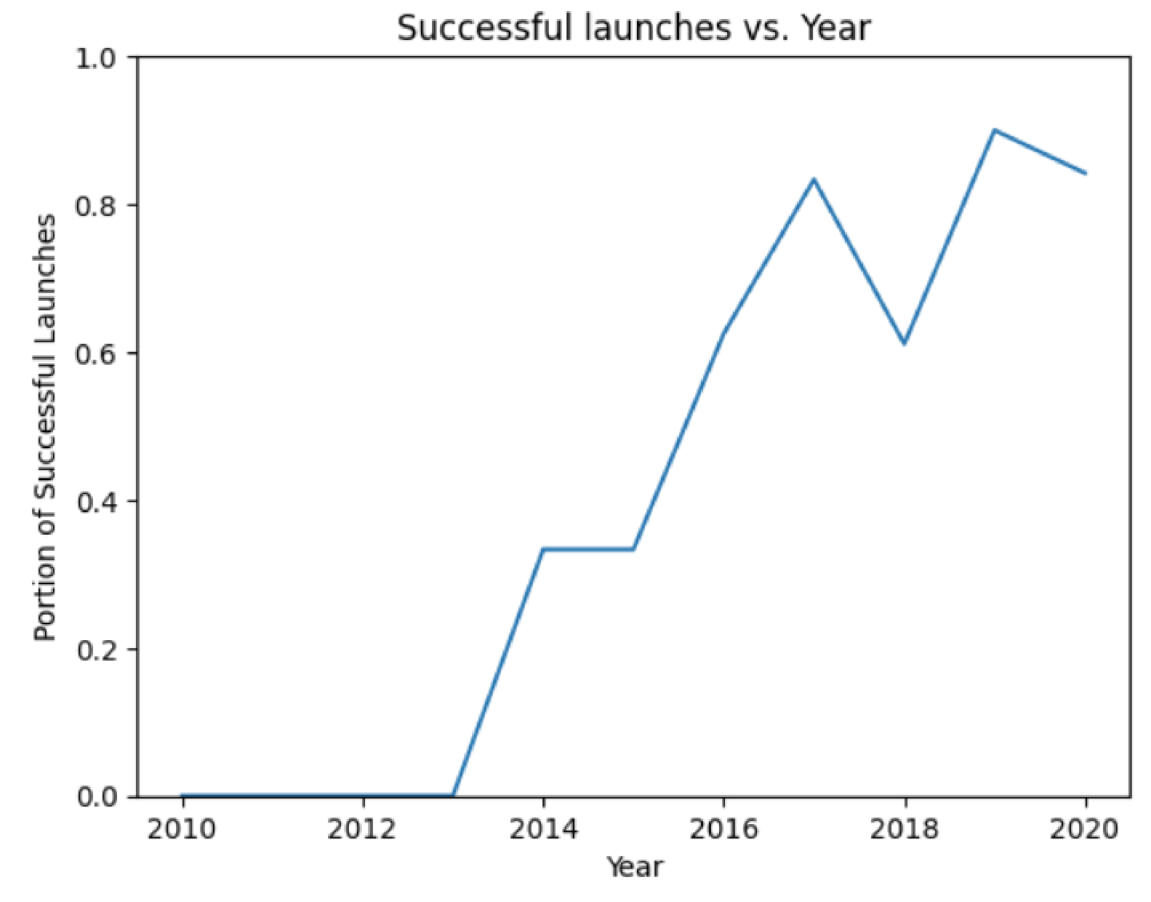
Payload vs. Orbit Type

- The heaviest payloads (> 13,000 kg) have all been launched to VLEO
- Except for launches to GTO, nearly all payloads over 6,000 kg have had successful outcomes



Launch Success Yearly Trend

- The success rate increased dramatically from 0% in 2010 to over 80% in 2020



All Launch Site Names

- The four unique launch sites are:

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

Launch Site Names Begin with 'CCA'

- The first 5 launches from sites beginning with `CCA` are:

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

Total Payload Mass

- In kg, the total payload mass carried by boosters from NASA is:

total
45596

Average Payload Mass by F9 v1.1

- The v1.1 boosters carried a light average payload:

booster_version	average
F9 v1.1	2928

First Successful Ground Landing Date

- The first successful landing on a ground pad was in December, 2015:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Variants of the FT booster (and no other boosters) were used successfully to land payloads between 4,000-6,000 kg on drone ships:

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Only 1-2 missions were unsuccessful (column 2 shows the tally):

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Only variants of the B5 booster have carried the maximum payload mass:

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

2015 Launch Records

- The two failed launches were launched at CCAFS LC-40 with v1.1 boosters:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- No attempt was made to land a number of launches between 2010-06-04 and 2017-03-20, and about half of landing attempts involved drone ships:

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

Section 3

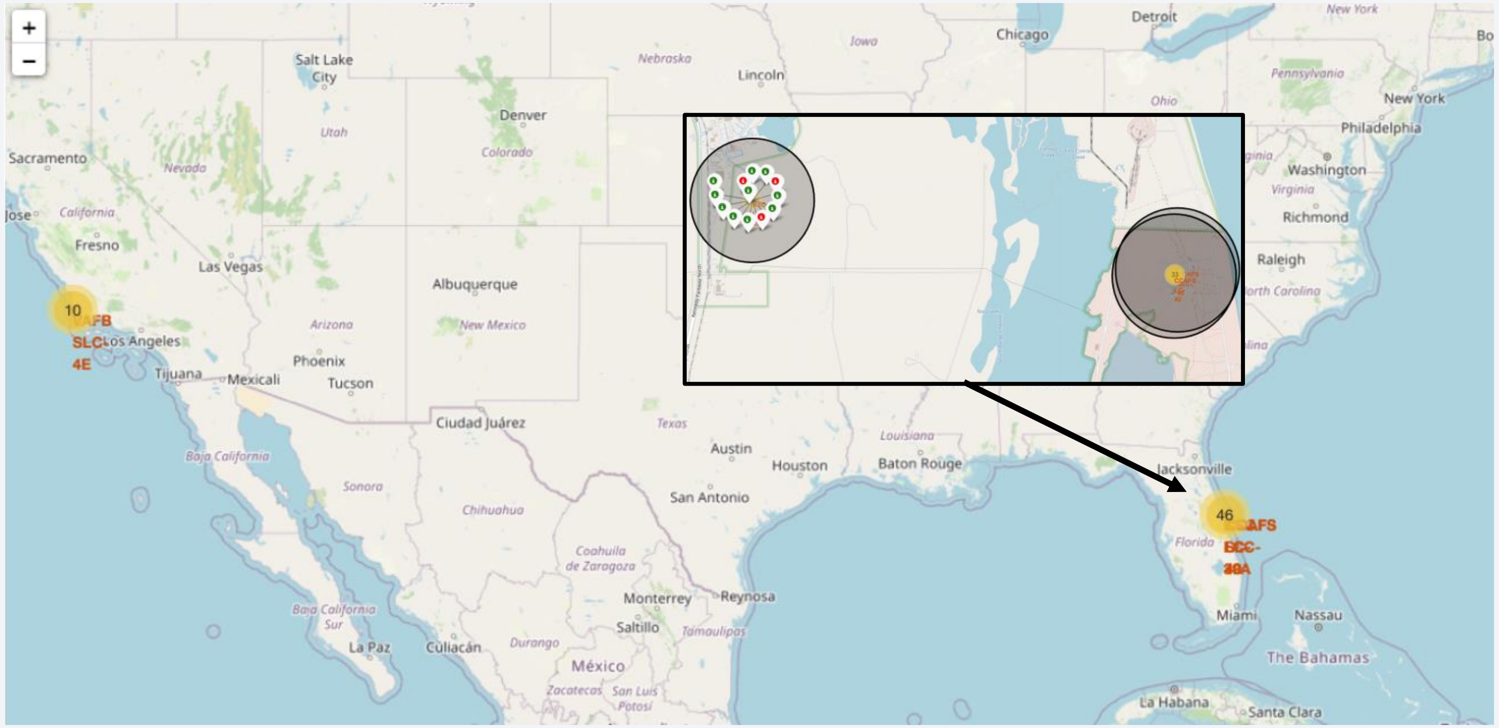
Launch Sites Proximities Analysis

Folium - SpaceX Launch Site Locations

- There are four launch sites:
  - 1 in southern California
  - 3 clustered in Florida
- All sites are in the southern US along the coast

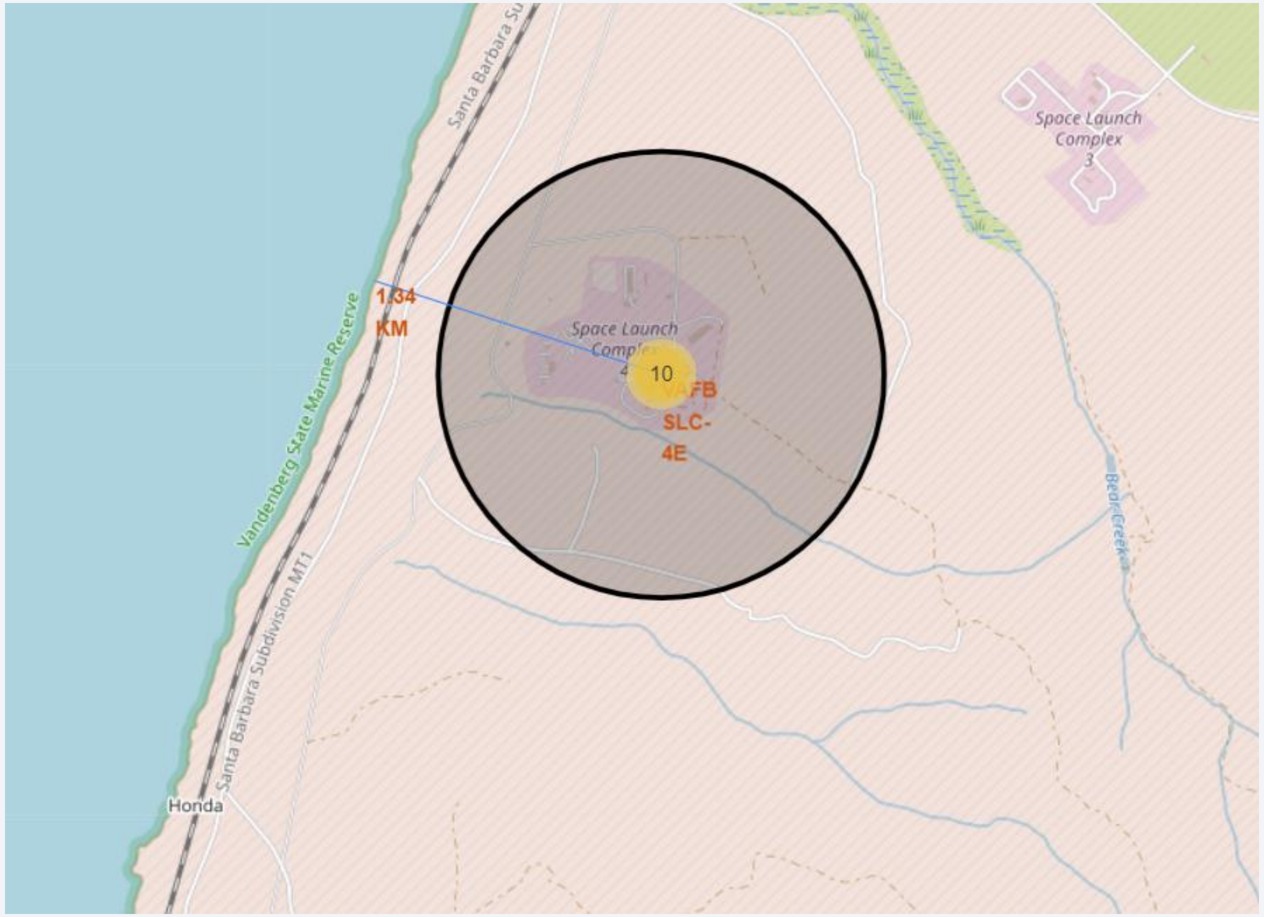
Folium - Successful and Failed Launch Landings

- The large majority of launches occurred at the launch sites in Florida
- The site with the highest success rate was Kennedy Space Center



Folium – Distances to Nearby Features

- All launch sites are near the ocean
  - For example, VAFB SLC-4E is about 1.34 kilometers from the Pacific Ocean

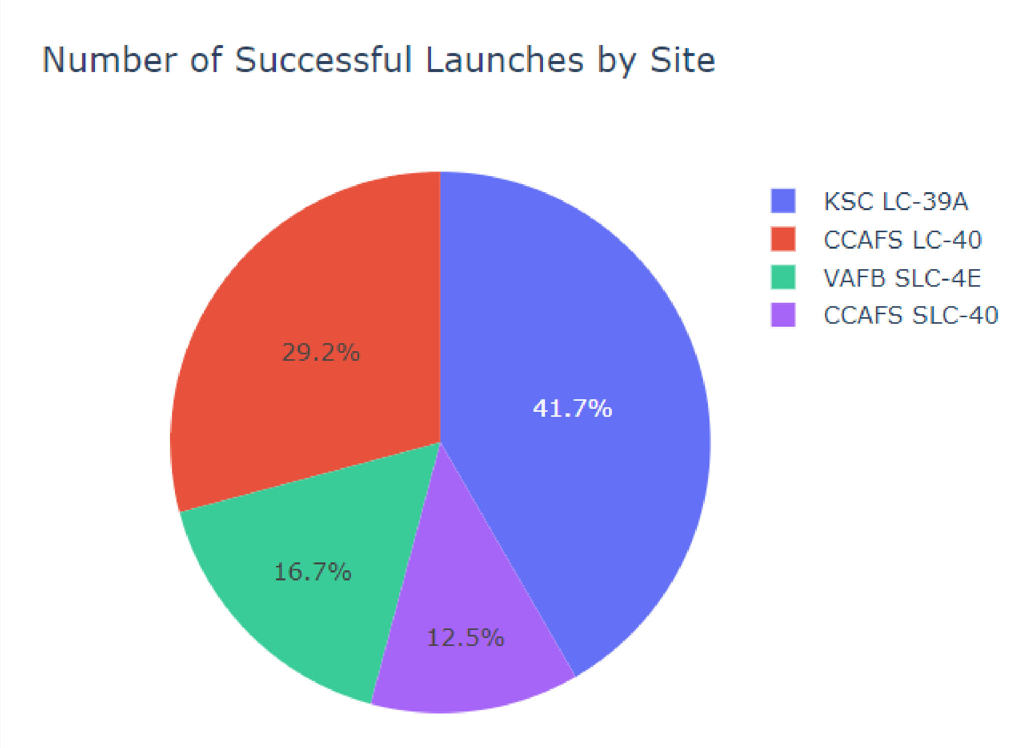


Section 4

Build a Dashboard with Plotly Dash

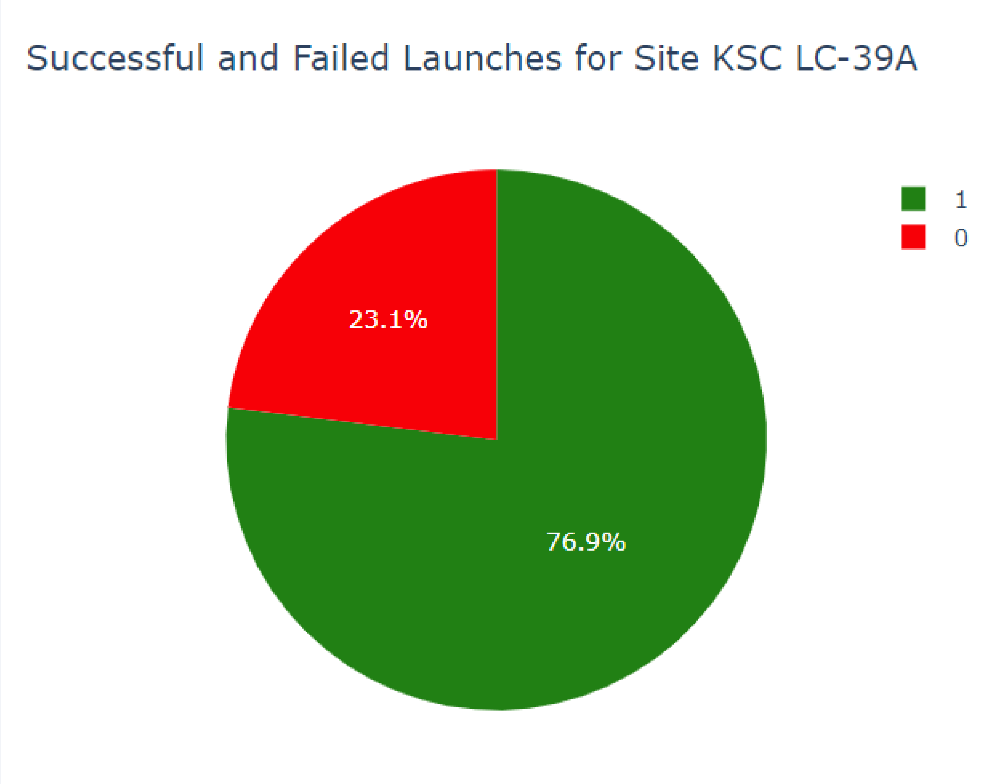
Dash – Successful Launches by Site

- The largest number of successful launches occurred at KSC LC-39A
- The smallest number of successful launches occurred at CCAFS SLC-40



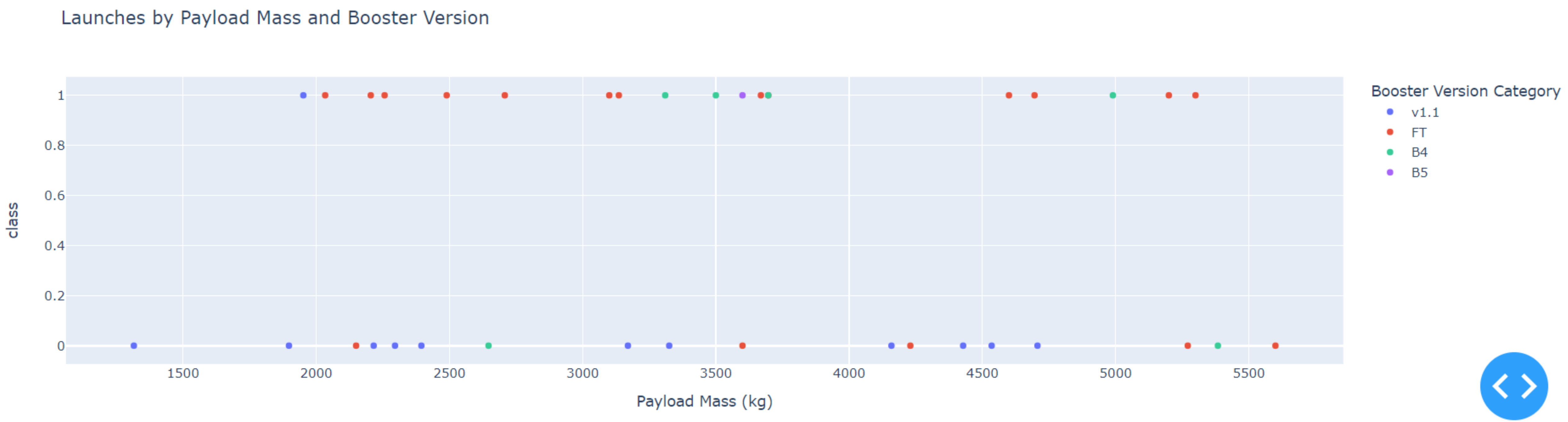
Dash – The Most Successful Site

- By far, the site with the highest success rate was KSC LC-39A (Kennedy Space Center)
- Over ¾ of the launches at KSC had successful landings, compared to less than ½ at all other sites



Dash – The Most Successful Payloads & Booster

- Most successful landings had payloads between 1000 and 6000 kg
- Of these successful landings, most used a FT booster

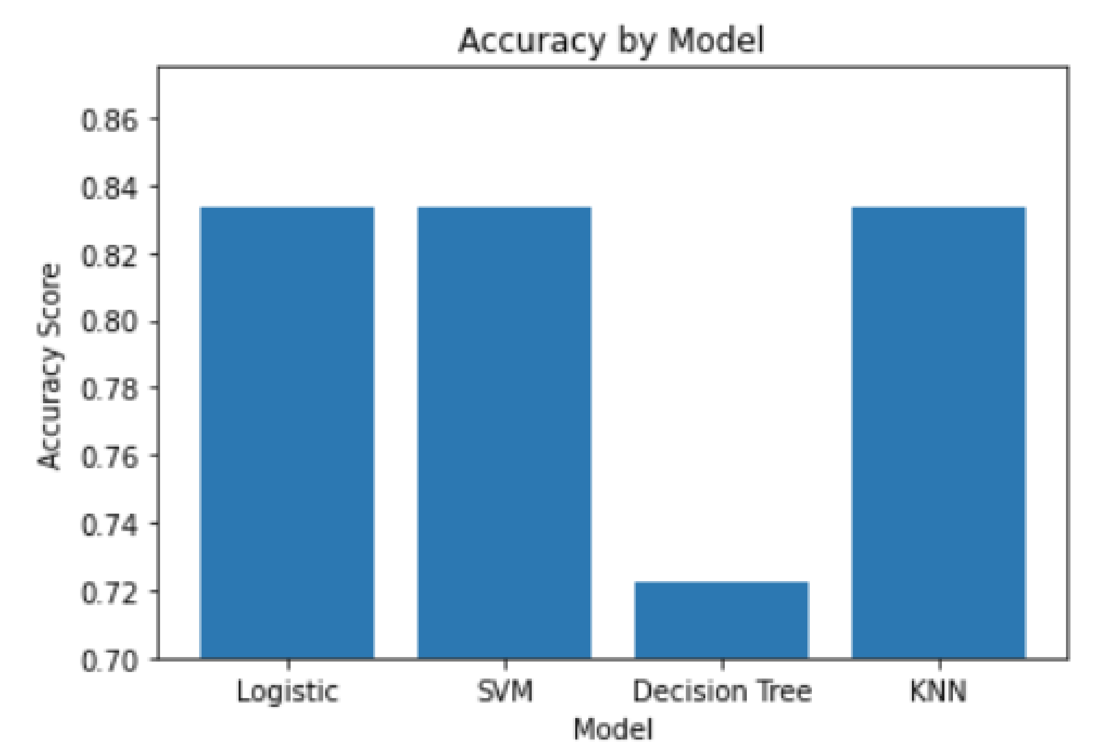


Section 5

Predictive Analysis (Classification)

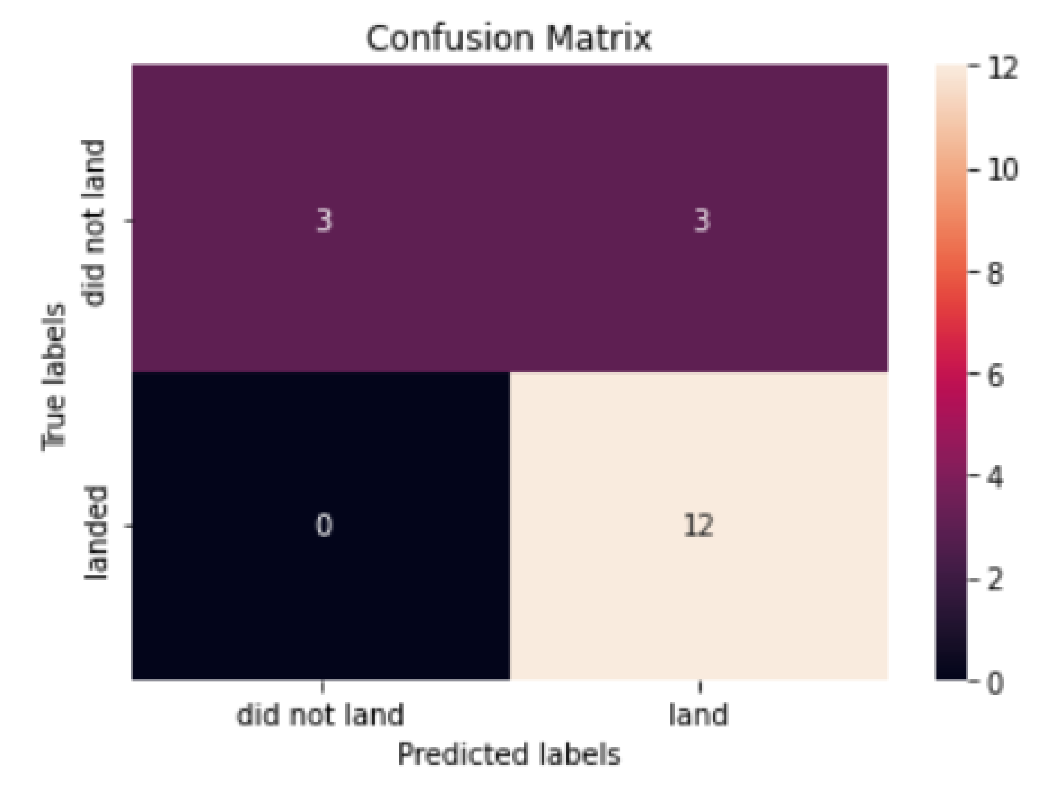
Classification Accuracy

- All models other than the decision tree had identical accuracy, correctly predicting 15/18 outcomes
- In addition, they had virtually identical performance on the training data



Confusion Matrix

- The confusion matrix for the logistic regression, support vector machine, and k-nearest neighbor models is shown
  **as noted before, all models performed identically on the test set*
- The models predicted perfectly on test samples where the launch landed
- However, the models performed poorly (50% accuracy) on test samples where the launch did not land



Conclusions

- **Launch date** was the strongest predictor of a successful landing. Successful landings increased from 0% in 2010 to over 80% in 2020.
- **Launch site** and **payload mass** were also strong predictors. Kennedy Space Center has a successful track record, and heavier payloads have been launched with a high success rate.
- **FT boosters** have been used very successfully, especially with medium-sized payloads (1,000-6,000 kg)
- Classification algorithms can predict the success of a launch with a high degree of accuracy (80-85%)
    - Logistic regression, support vector machines, and k-nearest neighbor all accomplished this

Thank you!