Winning Space Race with Data Science

Timothy R. Zura
10/02/2022

Outline

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

Executive Summary

- Through analysis of launch data from previous SpaceX flights, modeling was explored to predict the successful outcome of first-stage booster landings. Successful landings reduce mission costs as the first stage boosters can then be reused.
- Data pertaining to the orbits, payload mass, launch sites, and booster types of previous SpaceX missions were evaluated for their influence over outcome and to assess their value in the development of machine learning models.
- While landing outcomes were often failures in the early days of the SpaceX booster re-use program, the success rate has steadily improved. Payload mass, orbital type, and booster version provide key indicators as to the success of future missions.
- Adhering to parameters that contribute to successful outcome is key to competing with SpaceX.

Introduction

- SpaceX has offered an alternative to government sponsored space flight. Through successful landing and subsequent reuse of first stage boosters, SpaceX's commercial space flight program has reduced costs for missions.
- By understanding the factors that influence successful first stage booster landings, and being able to predict successful outcomes, other companies, including SpaceY, may have opportunity to compete in the commercialization of space flight.
- What are the factors that contribute to successful landing and re-use of first stage boosters?
- Can models be developed to accurately predict the successful outcome of a first-stage booster landing?
- This knowledge could provide a foundation for setting a competitive strategy.

Section 1

Methodology

Methodology

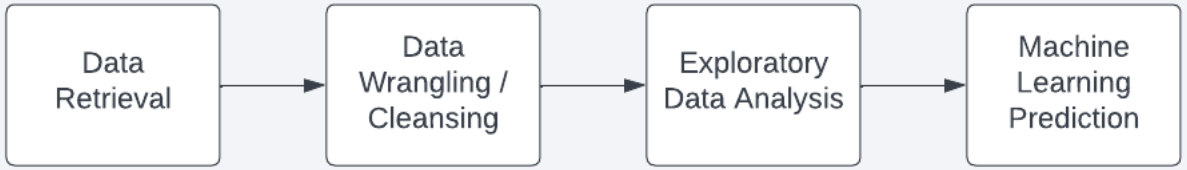
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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
- GitHub URL for research: https://github.com/wrabbling/spacey

Data Collection

- Two main data sources:
  - JSON API provided by SpaceX
  - Web Scraping of Falcon 9 mission information from Wikipedia.

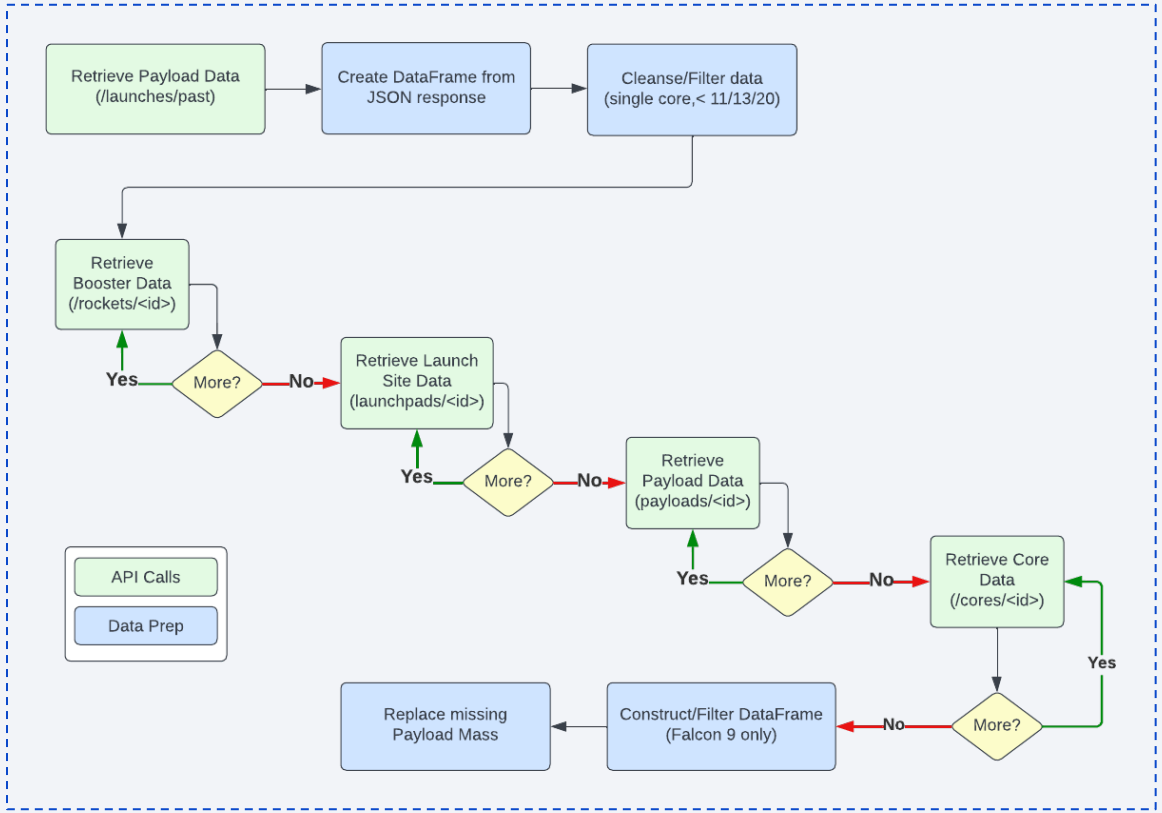
- Data relevant to launch outcome is retrieved, cleaned, explored, and compiled into a useful dataset that can be used to inform models for predicting future landing success.



Data Collection – SpaceX API

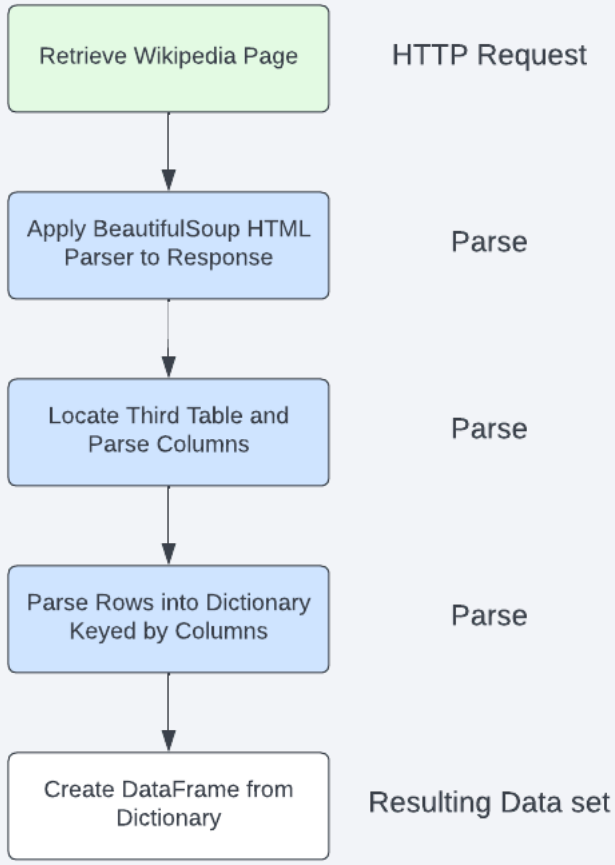
Data Retrieval | Data Wrangling / Cleansing | Exploratory Data Analysis | Machine Learning Prediction

- SpaceX Data API
  - https://api.spacexdata.com/v4
- API Endpoints
  - Past Launch Information: /launches/past
  - Rocket Booster Data: /rockets/<rocket_id>
  - Launch Site Data: /launchpads/<launchpad_id>
  - Payload Data: /payloads/<payload_id>
  - Core Data: /cores/<core_id>
- Convert JSON API responses into a DataFrame containing data for past rocket launches
  - Filter to Falcon 9 rockets and launches prior to 11/13/2020
- GitHub URL
  - https://github.com/wrabbling/spacey/blob/master/Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

- Scraping Wikipedia Falcon 9 Heavy Launch Page
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- BeautifulSoup Python Library
  - https://beautiful-soup-4.readthedocs.io/en/latest/
- HTML table converted to Pandas DataFrame
- GitHub URL
  - https://github.com/wrabbling/spacey/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Data wrangling occurred at various stages.
- API Data Retrieval (See slide 8)
  - https://github.com/wrabbling/spacey/blob/master/Data%20Collection%20API%20Lab.ipynb
  - Past launch data was filtered to launches using single cores and for flights prior to 11/13/2020.
  - Resulting DataFrame removed non-Falcon 9 records.
  - Missing payload mass values replaced with mean() value.
- Web Scraping Data Retrieval (See slide 9)
  - https://github.com/wrabbling/spacey/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb
  - Appropriate table was identified and selected from HTML response.
  - Columns and row values filtered out or substituted for missing values.
- EDA
  - https://github.com/wrabbling/spacey/blob/master/EDA.ipynb
  - Various landing outcome scenarios converted to 1 (success) or 0 (failure).
- EDA with Visualization
  - https://github.com/wrabbling/spacey/blob/master/EDA%20with%20Visualization%20Lab.ipynb
  - OneHotEncoder was used to create numerical representation of categorical columns for better consumption by machine learning algorithms.

EDA with Data Visualization

- Charts plotted to determine effect of different features on landing outcome and to identify trends.
- Charts:
  - Payload Mass vs. Flight Number; overlaid with launch outcome
  - Launch Site vs. Flight Number; overlaid with launch outcome
  - Launch Site vs. Payload Mass; overlaid with launch outcome
  - Orbit Type Success Rate
  - Orbit Type vs. Flight Number; overlaid with launch outcome
  - Orbit Type vs. Payload Mass; overlaid with launch outcome
  - Success Rate Yearly Trend
- https://github.com/wrabbling/spacey/blob/master/EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- Gain a better understanding of available data and exploration opportunity.
- Queries performed for data exploration and analysis:
  - Determine number of launch sites: 4 sites with distinct names
  - Identify launches from Kennedy Space Center.
  - Determine total mass of payload on NASA-launched boosters: 619967 Kg
  - Identify first successful landing on a drone ship: 04/08/2016
  - Identify boosters carrying a certain amount of payload landing successfully on a ground pad.
  - Determine total successful and failed mission outcomes.
  - Identify Booster versions that have carried the maximum payload mass.
  - Query information about successful missions launched in 2017.
  - Rank count of successful landing outcomes by outcome type between 2010 and 2017.
- https://github.com/wrabbling/spacey/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Data Retrieval | Data Wrangling / Cleansing | Exploratory Data Analysis | Machine Learning Prediction

- Folium was used to visualize launch site locations to better understand geographic characteristics of launch sites that support a positive outcome.
- Markers and Circles were added to identify the launch sites coordinates and provided textual information (name).
- Markers were used to indicate the success or failure of each launch at each launch site. Marker Clusters were used to group these markers together in a more readable and functional manner.
- Tying coordinates to mouse position facilitated the ability to measure the distance of the launch sites to various geographic and infrastructure landmarks.
- In scouting for new potential launch sites, it is helpful to understand the characteristics of a new location: Distance from population centers mitigate casualty count from explosions and toxic releases. Distance from highways and railroads affect supply lines and the ability to evacuate.
- https://github.com/wrabbling/spacey/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20Lab.ipynb. (Note, generated maps not showing even though notebook is trusted).

Build a Dashboard with Plotly Dash

- Interactive visualizations to explore data dynamically
  - Pie-Chart illustrating launch site success ratio. All sites or per site.
    - Dropdown provides for selection of launch site.
  - Booster Version overlaid on a scatter chart plotting launch success against payload mass.
    - Dropdown provides selection of launch site.
    - Range slider provides focusing on range of payload mass.
- It is believed that payload mass and booster version affect landing outcome.
  - Launch sites use a different mix of booster versions; further segmentation by launch site provide smore insight into consistency of booster versions on landing outcome.
  - Ability to zoom on data ranges is as important as being able to zoom on maps.
- https://github.com/wrabbling/spacey/blob/master/spacex_dash_app.py

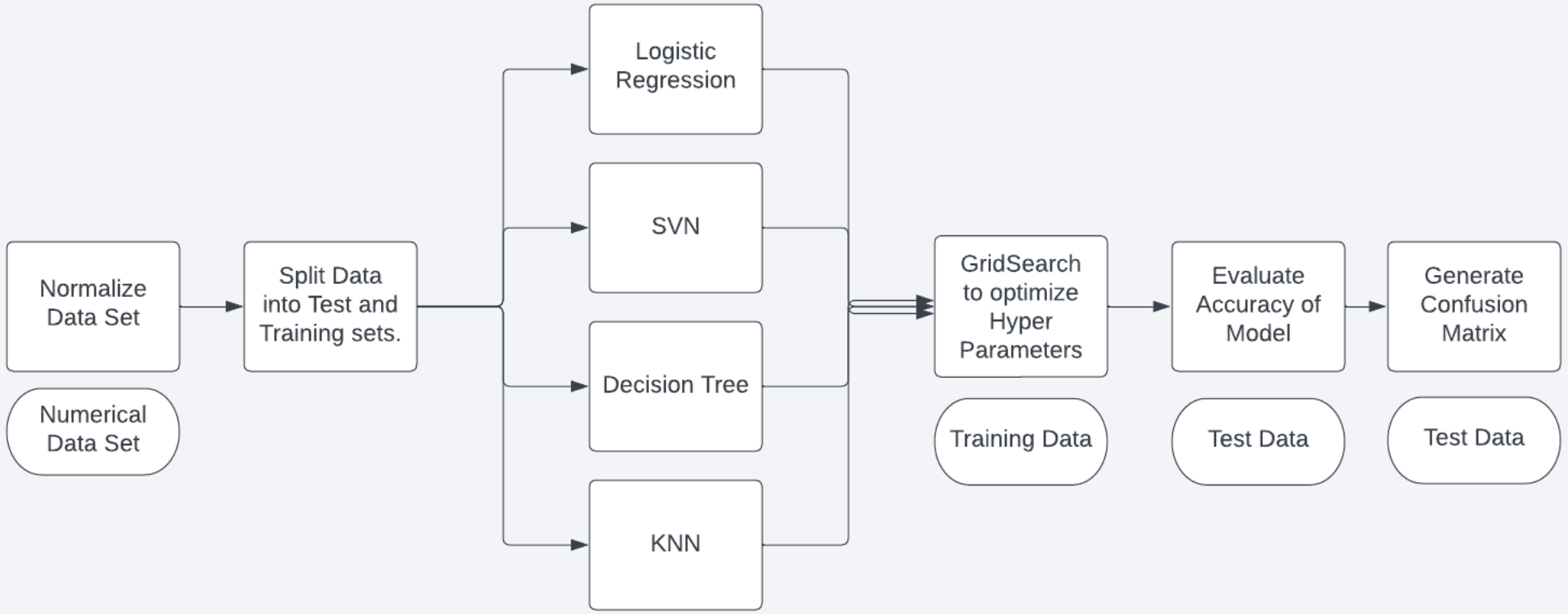
Predictive Analysis (Classification)

- Data evaluated using 4 machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- Data set was normalized using StandardScaler transformation to account for varying scale of feature values.
- Data set was split into test and training sets with test size accounting for 20% of overall set.
- Hyper parameters specific to each model were determined through Grid Search.
- Accuracy of model was determined against test set.
- https://github.com/wrabbling/spacey/blob/master/Machine%20Learning%20Prediction.ipynb

Predictive Analysis (Classification)

Data Retrieval | Data Wrangling / Cleansing | Exploratory Data Analysis | Machine Learning Prediction

Approach to predictive analysis and model evaluation



Results

- Exploratory data analysis results
  - Basic understanding of contents of data set.
  - Identification of features that have relationship to landing outcome.
- Interactive analytics demo in screenshots
  - Dashboard providing ability to dynamically sift through data.
- Predictive analysis results
  - 4 Models trained and evaluated.

Results - EDA

EDA and data wrangling

- Analyzing the dataset using DataFrames provided opportunities to understand and prepare data.
- Create additional ‘outcome’ role, reducing various measures/indicators of success to either 0 or 1.

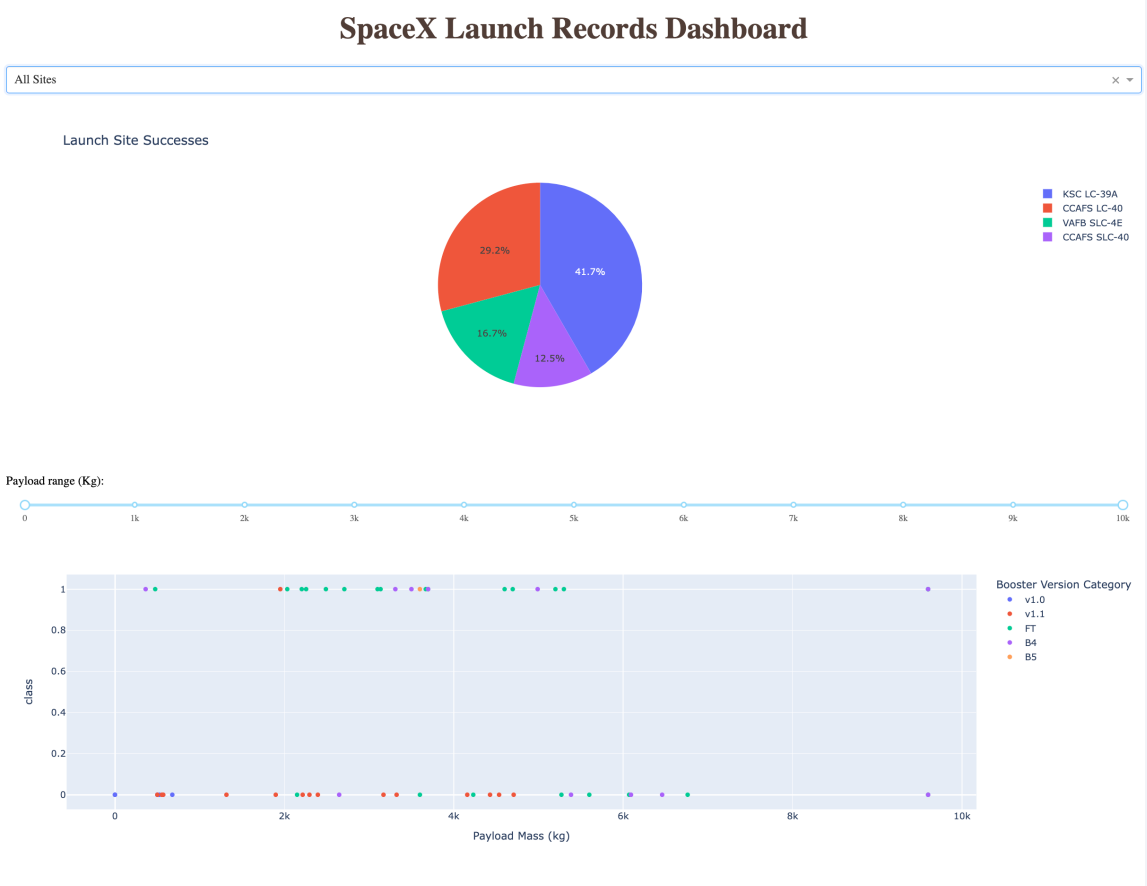
EDA and SQL

- Provided a means of exploring dataset and understanding possible segmentation opportunity.
- Did not directly feed into any predictive modeling, but useful for gaining a bigger context of the dataset.

EDA with visualization

- Create visualization to aid understanding of the interplay between various features and the outcome.
- Helped to see potential trends.

Results – Interactive Analytics



Dashboard provides interactive ability to understand relationship between various features and outcome.

- Launch Sites
- Payload Mass (kg)
- Booster Version

Ability to view data across all launch sites or on a per site basis provides a starting segmentation from which to understand other variables at play.

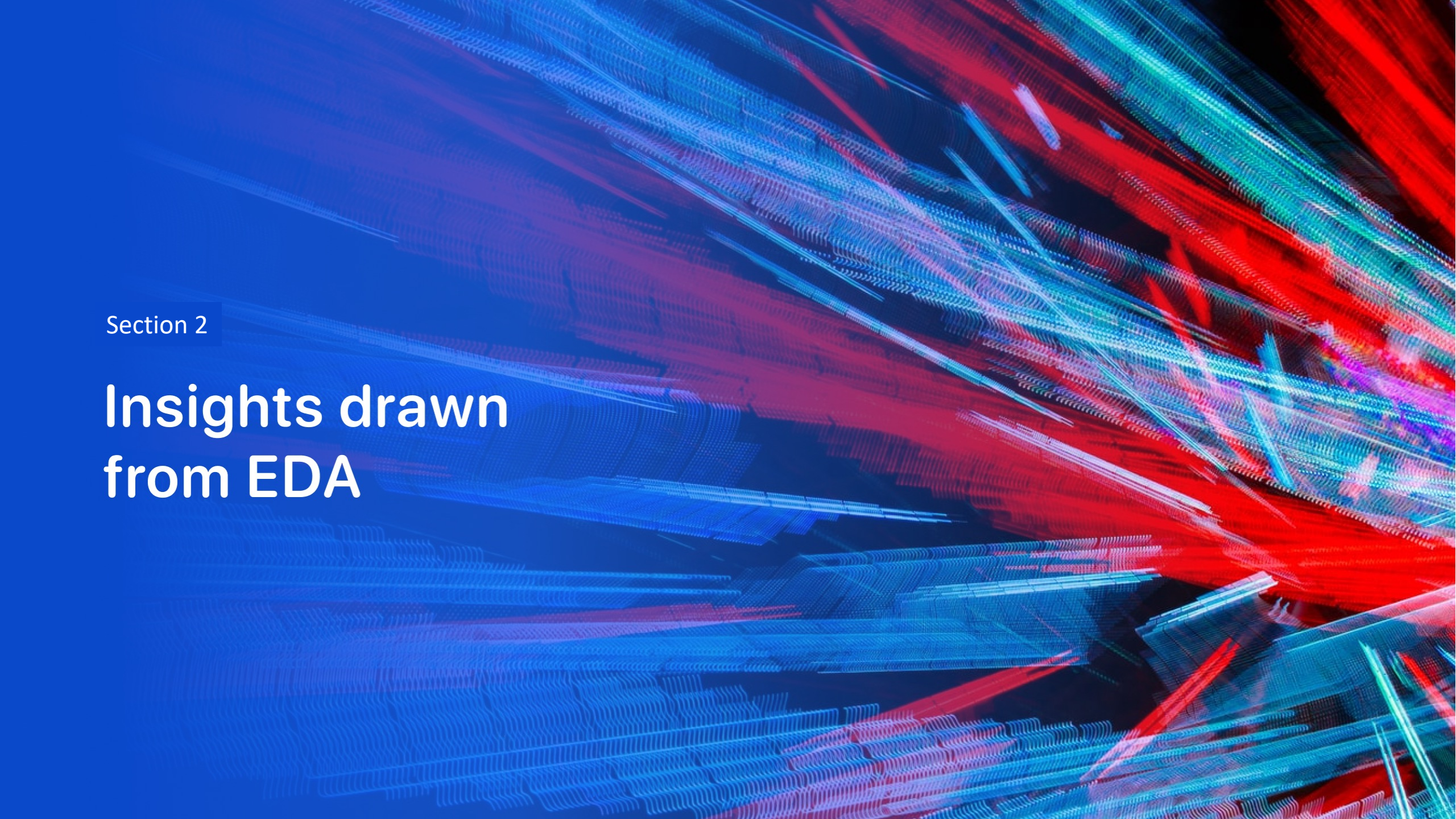
Zoom in/out on payload mass helps distill the noise, understand anomalies, create hypothesis to guide further analysis.

Results – Predictive Analysis

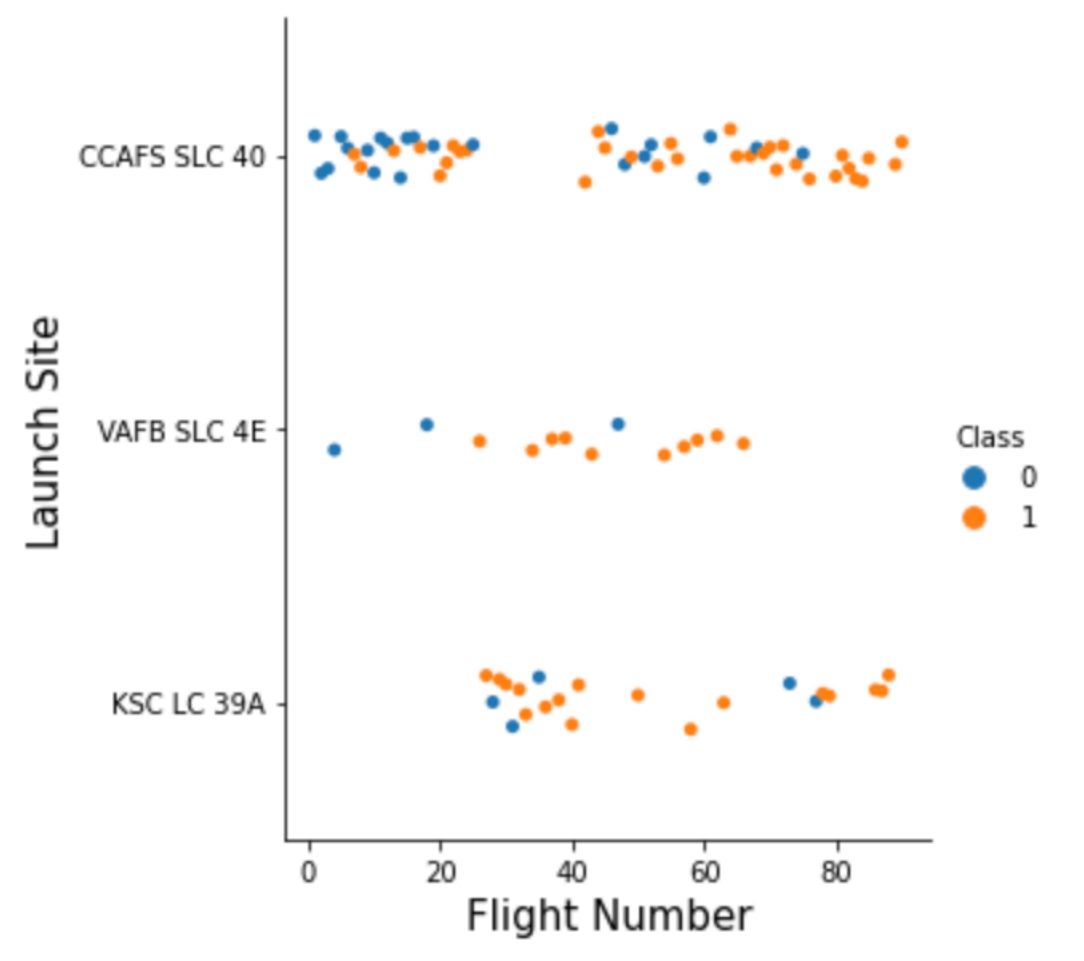
4 Predictive models evaluated :

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-nearest Neighbor (KNN)

All 4 models showed similar success and accuracy at predicting landing outcome

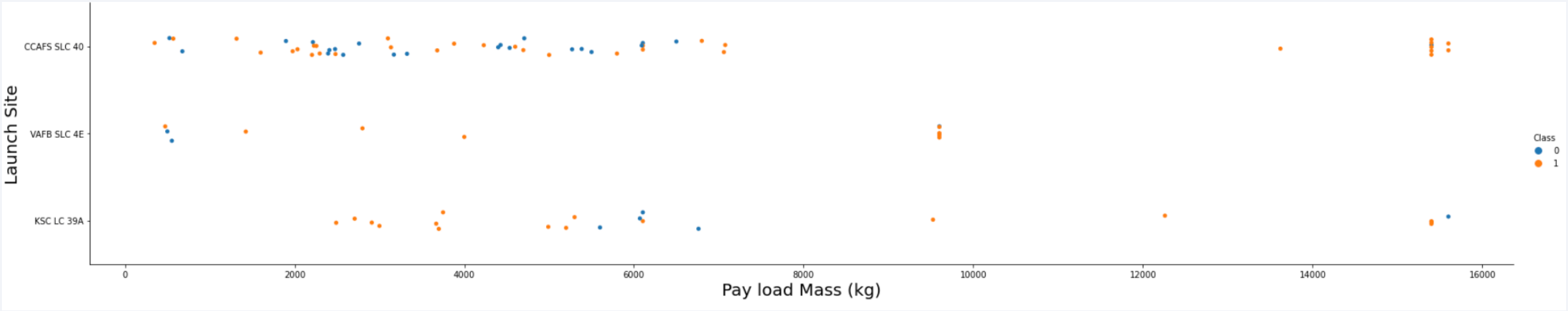

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



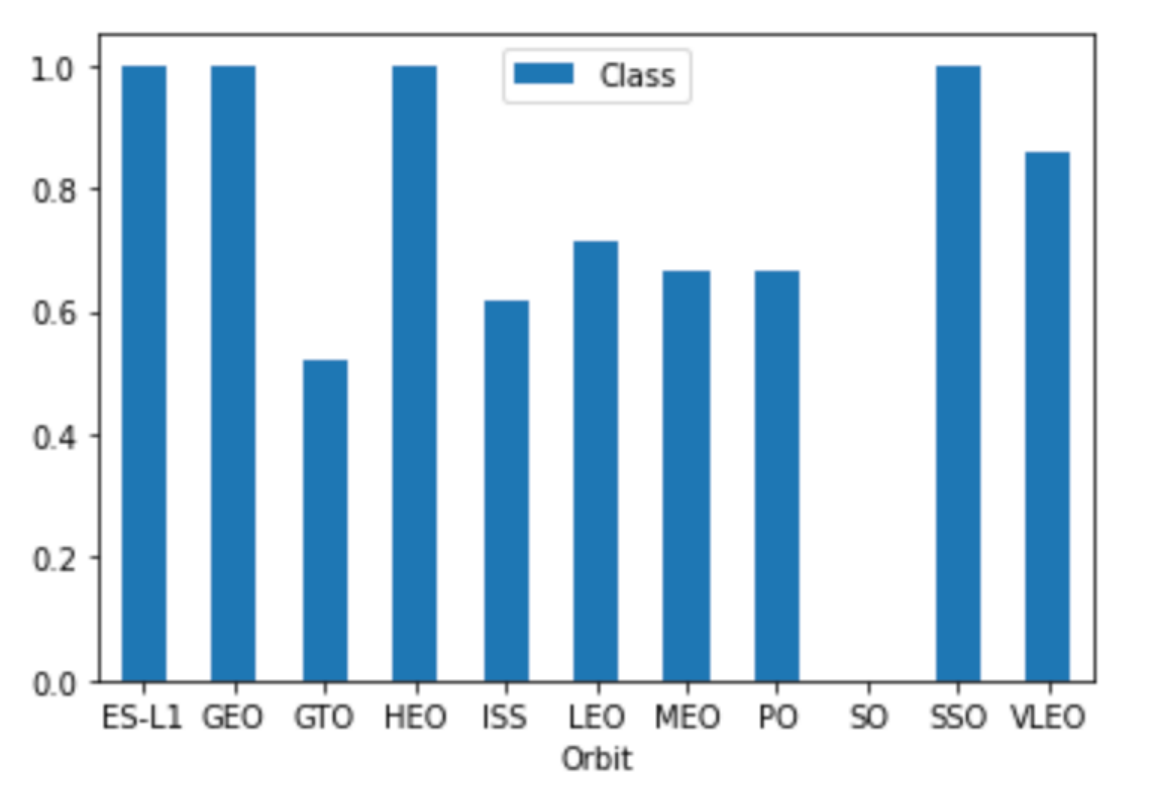
- Blue = failed landing.
- Orange = successful landing.
- The CCAFS SLC 40 launch site may have a lower success rate than other launch sites, but that is potentially misleading.
- There was more risk earlier in the program and CCAFS SLC 40 supported most of the launches.
- The risk of a failed landing outcome has decreased over time. Success rate at all launch sites is high later in the program.

Payload vs. Launch Site



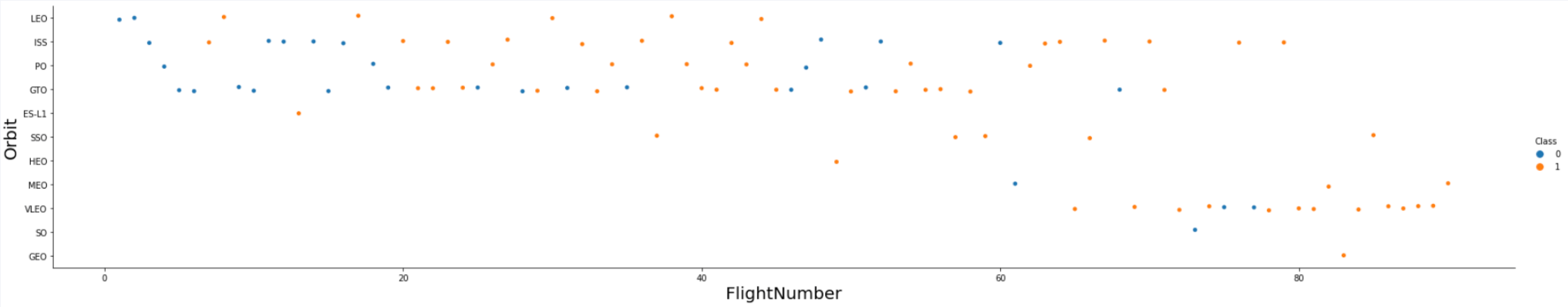
- Heavy payloads (>10,000 kg) do not launch out of Vandenberg Air Force Base.
- Success rate of payloads > 7000 kg appears higher than lighter payloads.
- Determination of success rate needs to consider not just payload, but also the date of the flight. Many of the unsuccessful outcomes shown in this plot may have happened earlier in the program.

Success Rate vs. Orbit Type



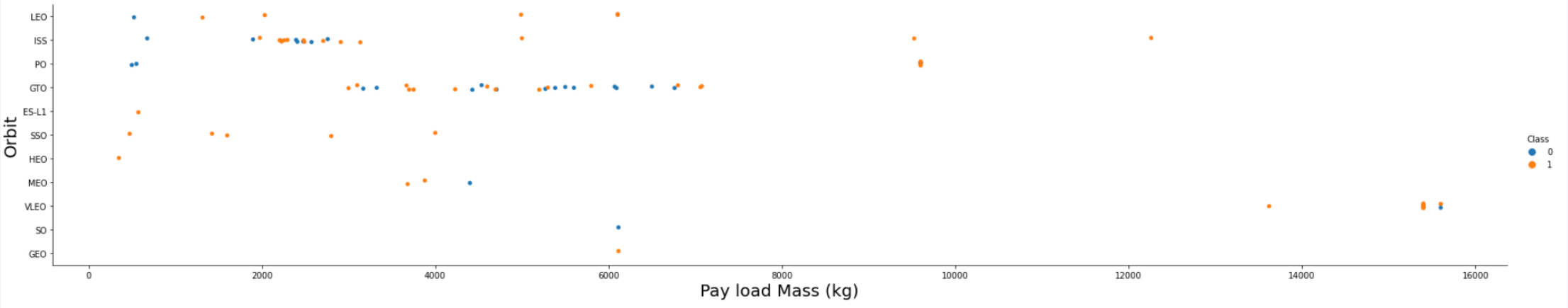
- Success rate is affected by the type of orbit being attempted.
- ES-L1, GEO, HEO, and SSO orbits all have 100% success.
- The flight number may contribute to an orbit type's success rate. Do later flights have a higher success rate because of program maturity?

Flight Number vs. Orbit Type



- The success of GTO orbits does not appear to be affected by flight number (program maturity).
- The success of LEO orbits does appear to improve with later flights, however.
- There is a weak correlation between orbit type and success with flight number.

Payload vs. Orbit Type

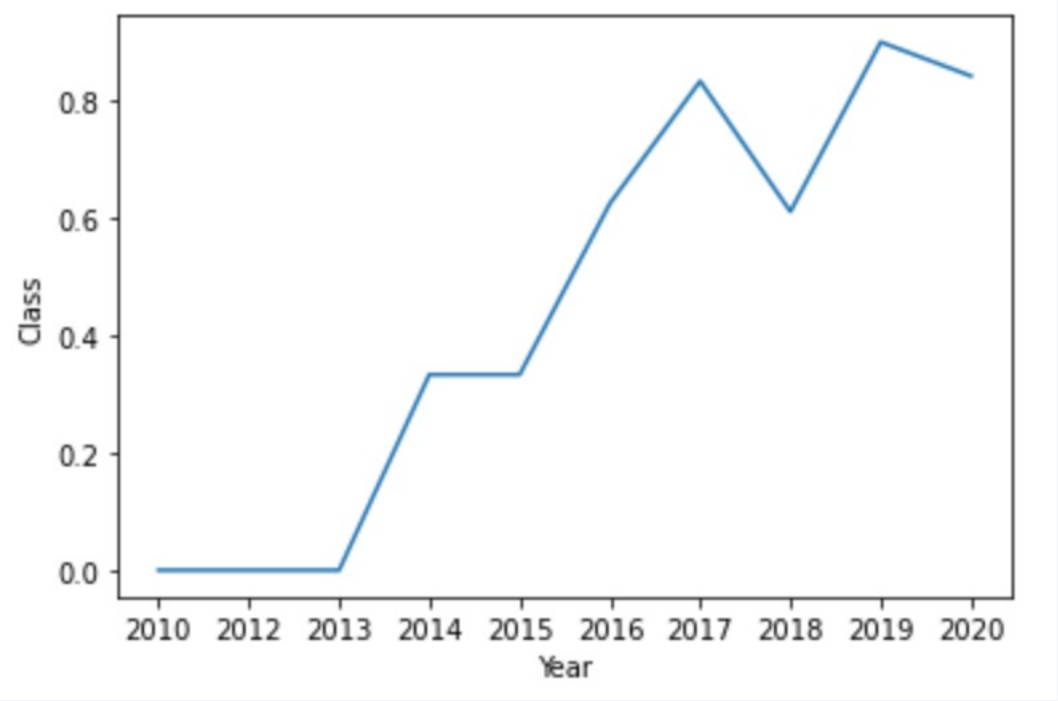


- Heavier payloads seem to have a relationship with positive success for Polar, LEO, and ISS orbit missions.
- GTO orbit missions seem to have a high negative outcome regardless of payload.
- Difficult to determine influence of payload on some orbital types because there is not enough data. (GEO, SO, MEO, HEO, SSO, ES-L1).
- Influence of orbit on outcome is difficult to determine because of lack of data about relationship between payload and orbit as well as conflicting data about flight number and orbit.

Landing Success Yearly Trend

- Positive landing outcome has increased significantly between 2010 and 2020, with a dip in 2018 and in 2020.

- First three years were abysmal.



All Launch Site Names

- Four different launch sites
  - Cape Canaveral Air Force Station : CCAFS LC-40
  - Cape Canaveral Air Force Station : CCAFS SLC-40.
  - Kennedy Space Center : KSC LC-39A
  - Vandenberg Air Force Base : VAFB SLC-4E
- Segmentation by launch site could help determine if it is a significant factor in landing outcome.

Launch Site Names Begin with 'KSC'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- Sample of 5 launches from Kennedy Space Center
- Provides opportunity to further evaluate landing successes across a variety of booster versions, payloads, and orbits.

Total Payload Mass

- Total payload carried by NASA (CRS) : 45596 kg
- It is possible to determine the total payloads and average payloads per customer. Could be interesting to see which customers participate in more successful outcomes based on payload average.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 : 2,928 kg
- Could be interesting to see if successful landing outcomes of various boosters based on payload mass that they carry.

First Successful Ground Landing Date

- The first successful landing outcome on a drone ship occurred on 04/08/2016.

- Knowing first successful landing outcome happened in 2013, it took another 3 years before the first successful landing outcome happened on a drone ship.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1032.1
  - F9 B4 B1040.1
  - F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:
  - Failure (in flight) : 1
  - Success : 99
  - Success (payload status unclear) : 1

- Although successful landing outcome is lower, the success of the missions overall is very high.

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass:
  - 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

- Knowing that heavier payload missions appear to have better landing outcomes, these boosters may play a factor in predicting future successful outcome.

2017 Launch Records

- Records of successful landing outcomes in 2017

February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

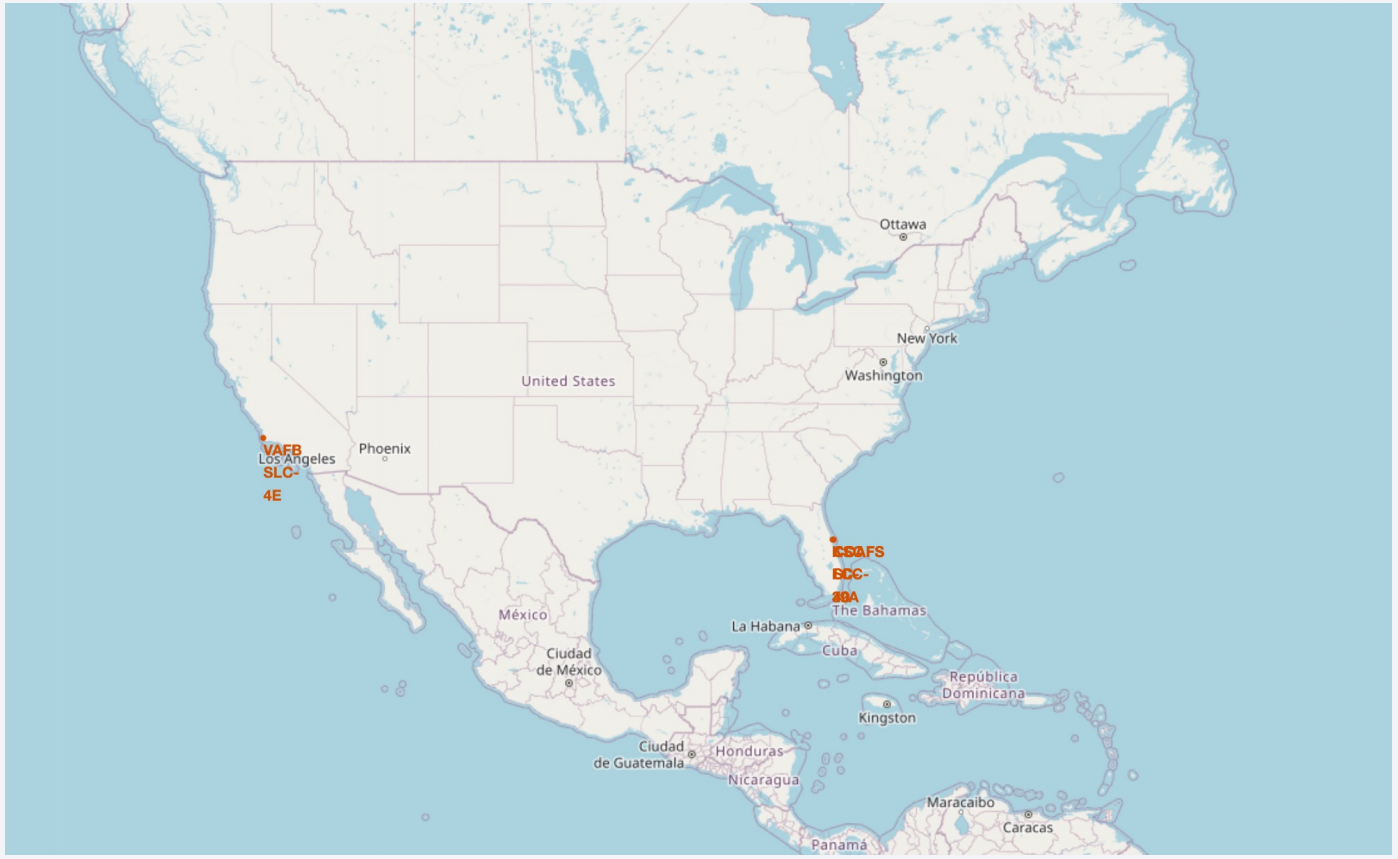
- Ability to look at records for a specific year is useful.

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

- Between 06/04/2010 and 03/20/2017:
  - There were 5 successful landings on drone ships.
  - There were 3 successful landings on ground pads.


Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>



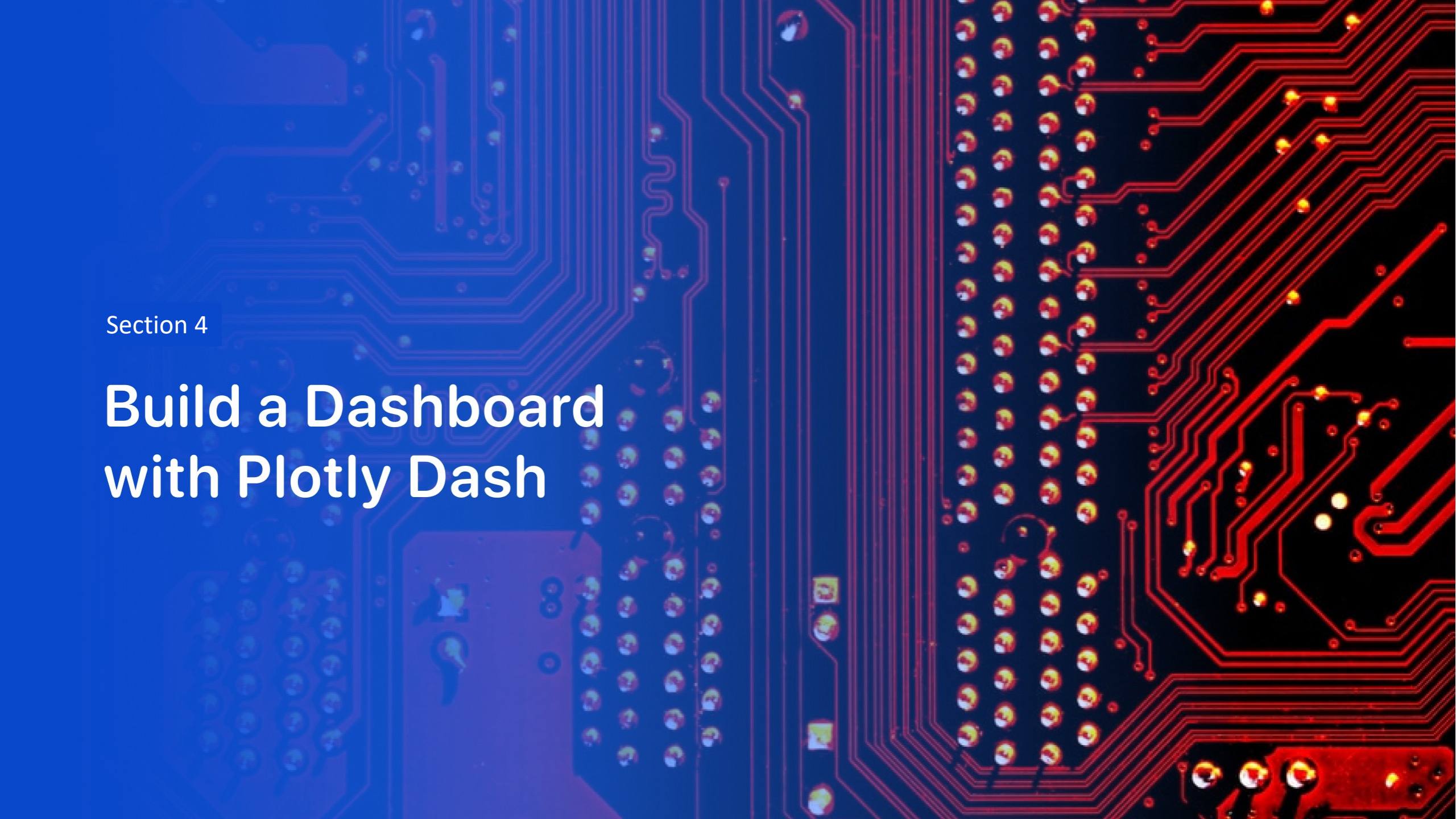
- Four launch sites are located along coastal area in southern half of United States.

Kennedy Space Center Launch Outcomes

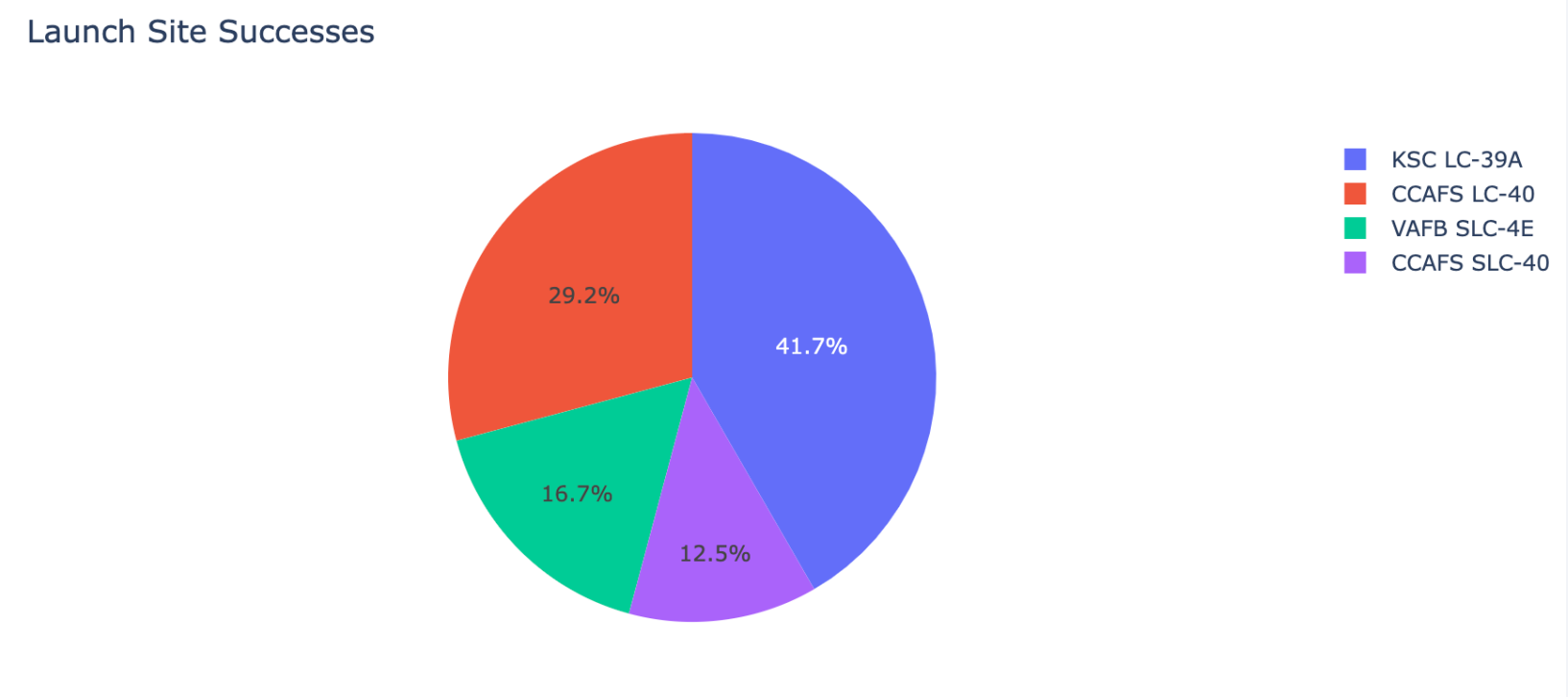
- Green indicates launches with successful landing outcome.
- Red indicates launches with unsuccessful landing outcome.

Vandenberg Launch Site Location Information

- Vandenburg Air Force Base (now, ‘Space Force’) illustrating distance to nearest city, supply infrastructure, and coastline.

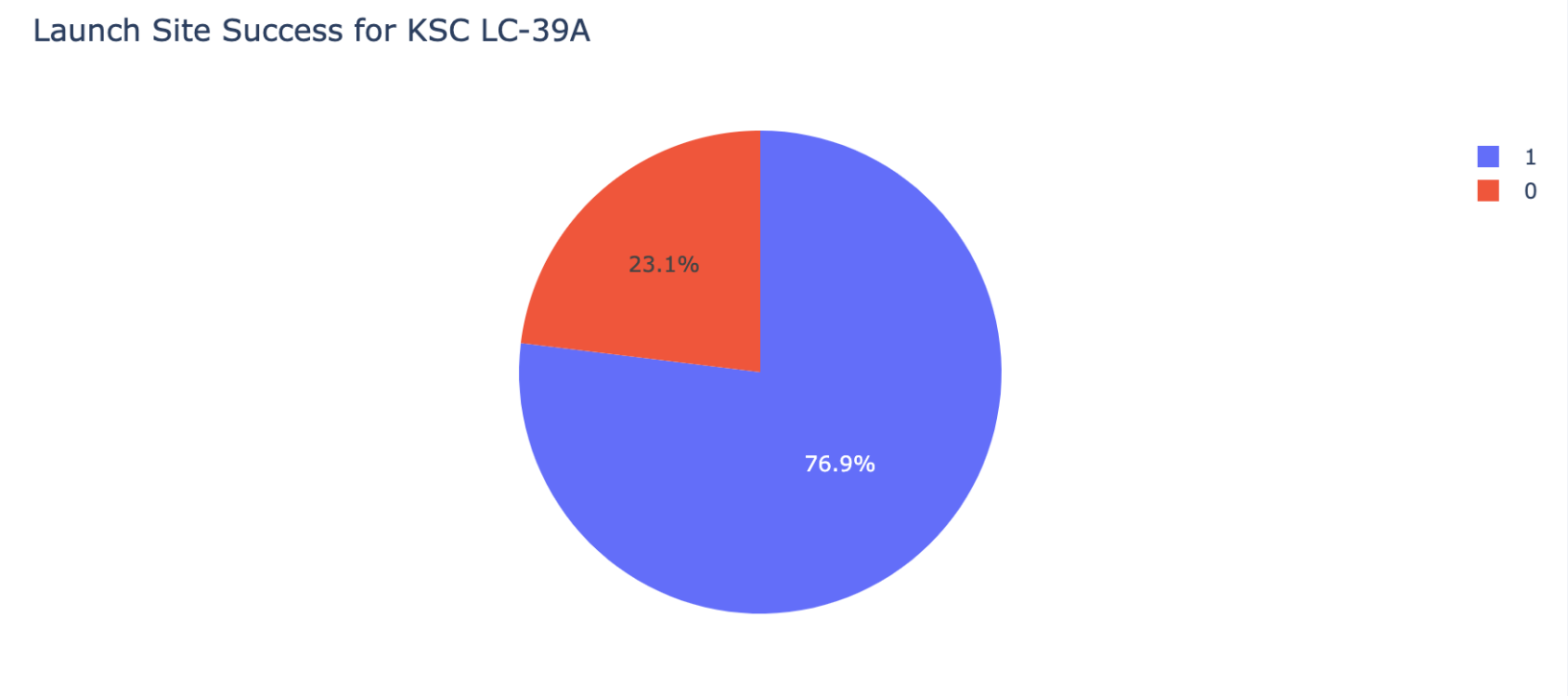

Section 4
Build a Dashboard with Plotly Dash

Landing Outcomes by Launch Location



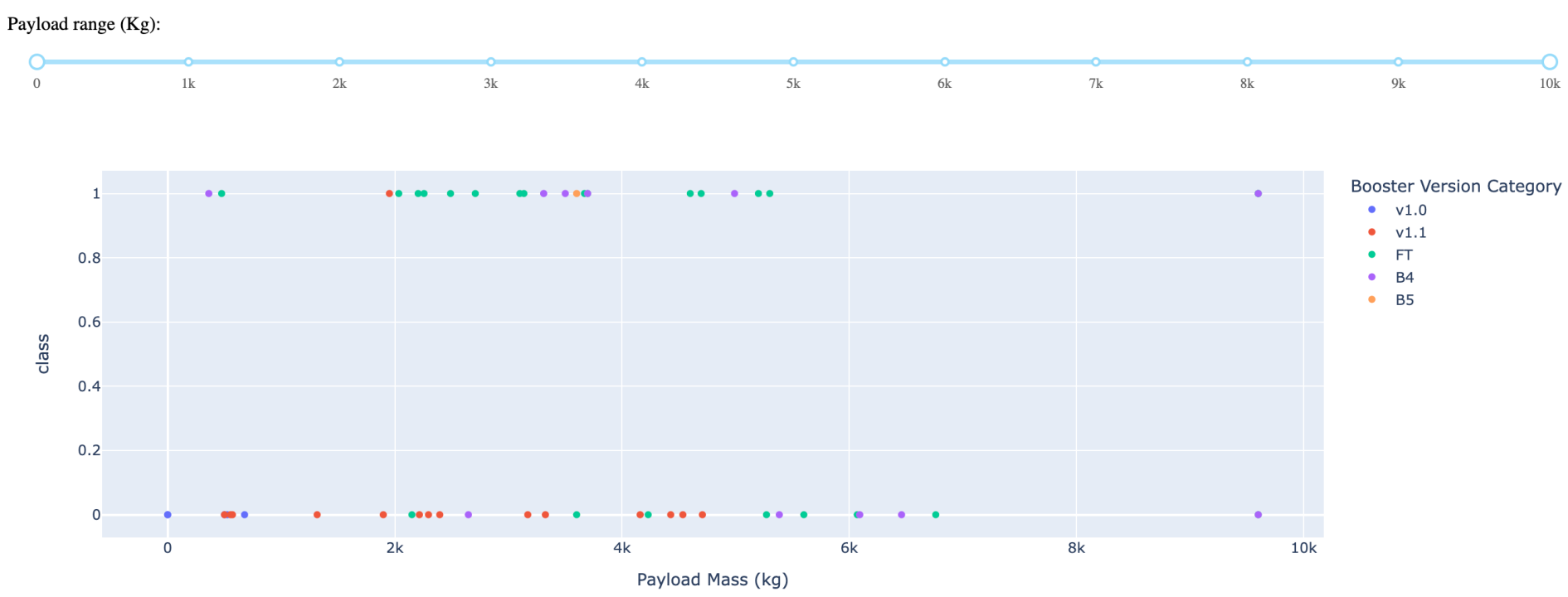
- Successful landing outcomes for each of the 4 launch sites.
- May not tell the whole story, as success rate was low in early stages of the program but have improved.

Launch Site with Most Successful Landing Outcomes



- Kennedy Space Center has 76.9% success ratio for landing outcomes.
- KSC did not participate in earlier flights where failure rate was higher. May have used payloads and orbits more likely to achieve success.

Successful Landing Outcomes vs Payload Mass



- Across all launch sites, high ratio of successful outcomes between 2k and 4k kg.
- Booster Version ‘FT’ appears to be associated with successful landing outcome most often.

Section 5

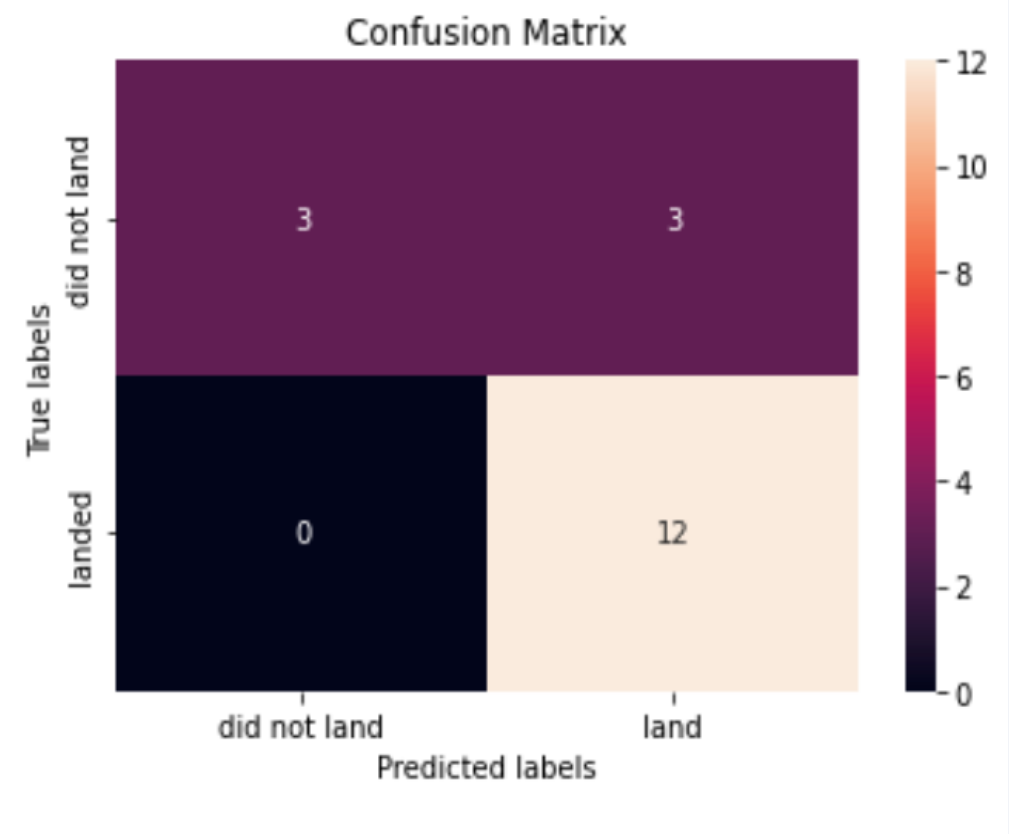
Predictive Analysis (Classification)

Classification Accuracy

- Accuracy against validation data:
  - Logistic Regression: .846
  - Support Vector Machine: .848
  - K-Nearest Neighbors : .848
  - Decision Tree Classifier: .904

- All models have a classification accuracy of .833

Confusion Matrix



- All models resulted in a similar confusion matrix.
- The model does provide some false positives (3 out of 18 test records)

Conclusions

- Landing outcomes can be predicted with high accuracy through consideration of payload, booster version, and orbit type.
- Launch site locations are all displaying success in later years of program. Success rate of some launch locations is low because of their participation in early part of program which was more prone to failure.
- Payloads between 2K kg and 6K kg have higher success rates, as do payloads > 8K kg.
- Consistent accuracy of prediction across multiple models provides confidence in predictions.
- Space Y, assuming it has the funding, technology, and years of trial-and-error that Space X experienced, should focus on the orbits, payloads, and boosters that provide a higher chance of successful landing outcomes to keep costs low.

Appendix

- API Retrieval Methods
  - API requests were timing out, likely due to throttling by the API.
  - Retrieval functions were rewritten. Example below.

```
from requests.adapters import HTTPAdapter
from requests.packages.urllib3.util.retry import Retry

def getBoosterVersion(data):
    RETRY_STRATEGY = Retry(
        total=5,
        backoff_factor=1
    )
    ADAPTER = HTTPAdapter(max_retries=RETRY_STRATEGY)
    request_session = requests.Session()
    for x in data['rocket']:
        if x:
            data_url = "https://api.spacexdata.com/v4/rockets/"+str(x)
            request_session.mount(data_url, ADAPTER)
            response = request_session.get(data_url)
            response = response.json()
            BoosterVersion.append(response['name'])
```

Thank you!