Data Science Capstone Project

Chris Vaisnor

August 23, 2021

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary

- This analysis was created to predict successful Falcon 9 booster landings at SpaceX.
- Data was collected via the SpaceX open-source API, and web scraping public data.
- Exploratory data analysis was performed with SQL queries and Folium visualizations.
- While multiple machine learning models were fit and evaluated, a Decision Tree classifier was able to produce the highest accuracy score on the test set.

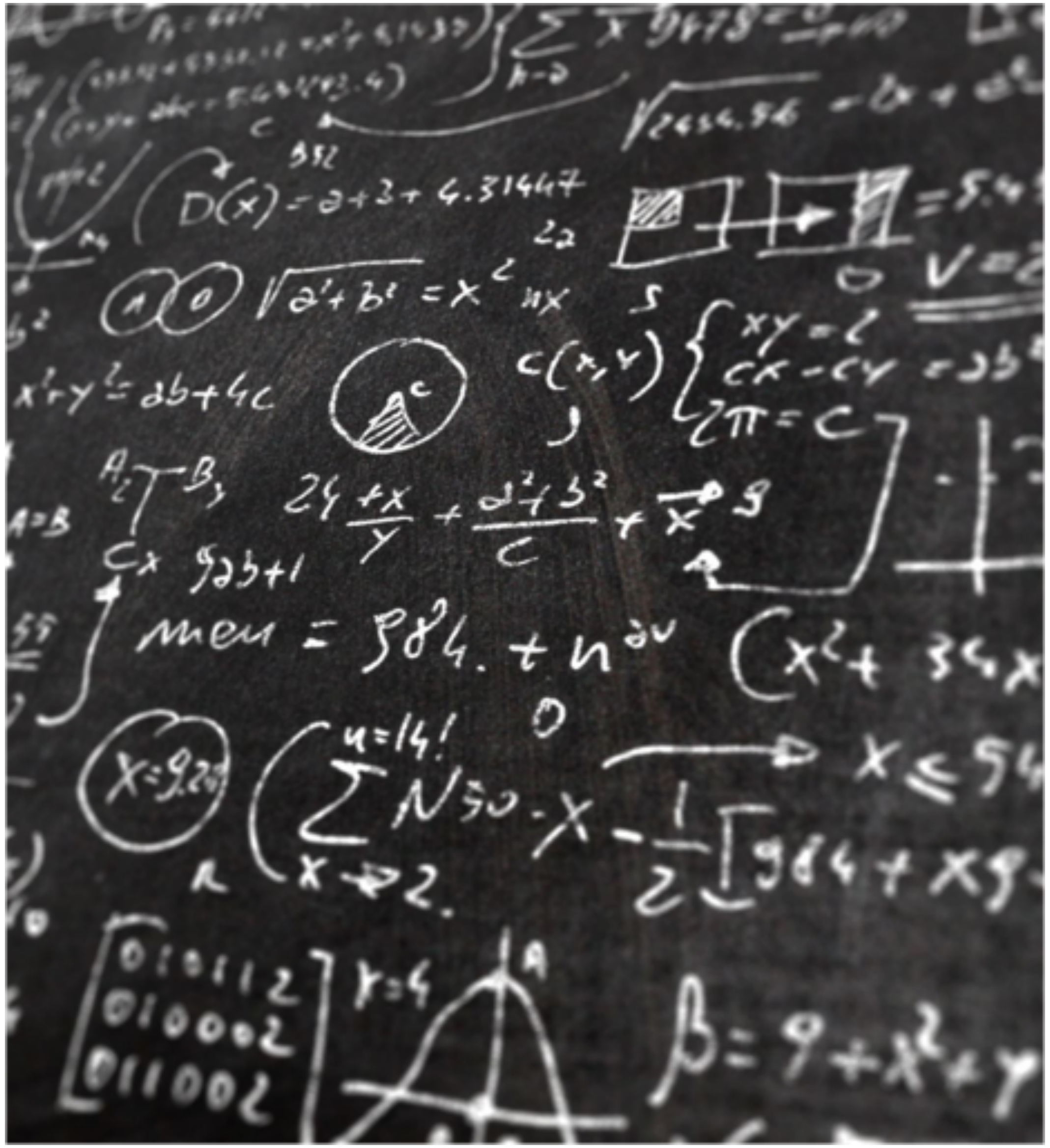
Introduction

The ability to reuse boosters greatly reduces average cost-per-launch, which translates to cheaper prices for companies looking to send their own satellites into space.

If a new company were to join the reusable-booster industry, a proper evaluation of the cost criteria and success factors would be helpful for the new company.

This analysis hopes to answer questions about the factors that contribute to successful landings.

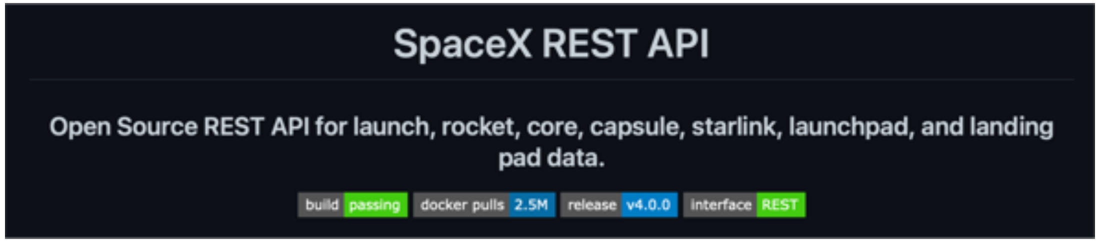
Methodology



Methodology - Overview

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis with classification models

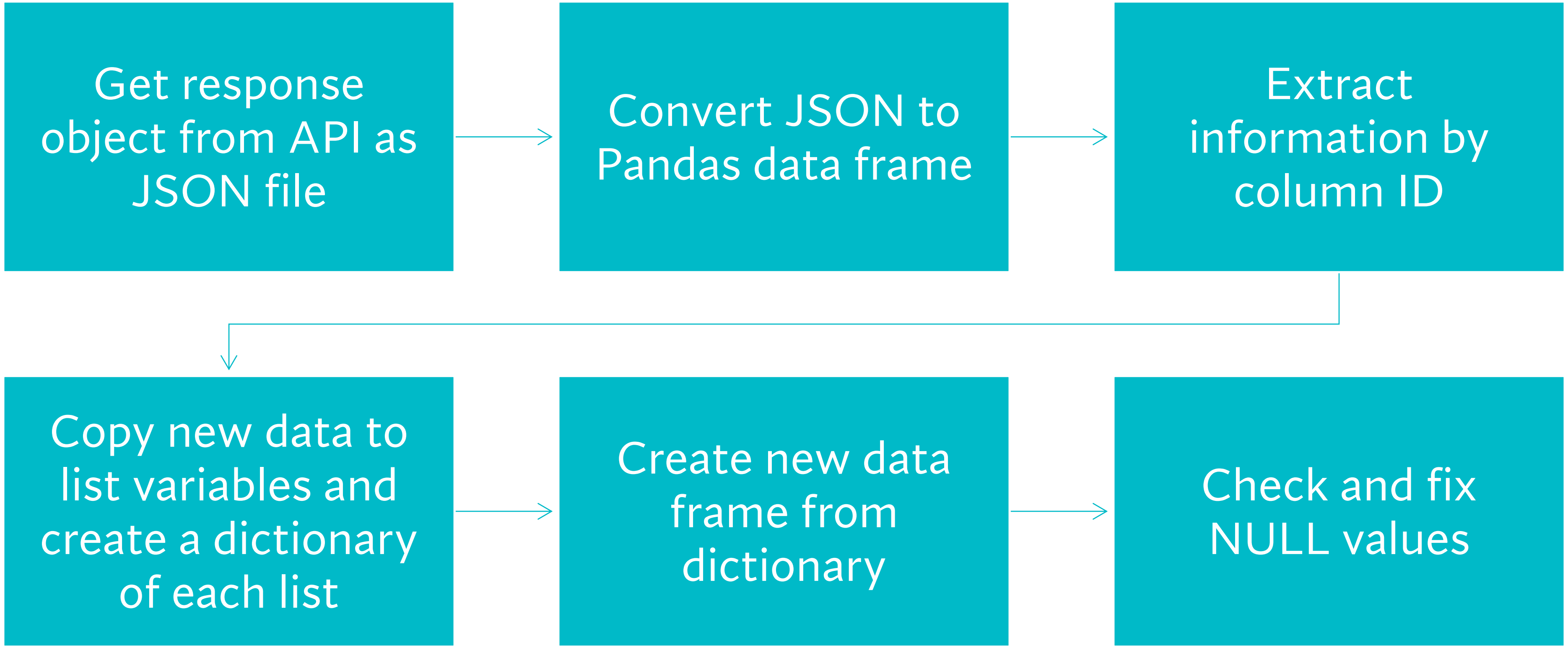
Data Collection



- There are two methods I used to collect historical launch data:
  - A Representational State Transfer (REST) API for SpaceX data
  - Traditional web scraping on Wikipedia

WIKIPEDIA
The Free Encyclopedia

Data Collection – SpaceX API



GitHub Link: [Data Collection API]

Data Collection – Web Scraping

1. Get HTML response
2. Create BeautifulSoup object from HTML text
3. Use find_all() to locate table with launch data
4. Locate column names
5. Parse data
6. Create Pandas data frame from parsed data

GitHub Link: [Data Collection with Web Scraping]

Data Wrangling

- There were originally eight difference landing outcomes, so I organized and separated them into binary variables. (1 indicated a successful landing, 0 indicated a failure landing)

```
df.groupby('Outcome')['FlightNumber'].count()

Outcome
False ASDS      6
False Ocean     2
False RTLS      1
None ASDS       2
None None      19
True ASDS      41
True Ocean      5
True RTLS      14
Name: FlightNumber, dtype: int64
```

1. Identify any remaining missing values
2. Group by launch site and number of flights
3. Calculate occurrence of each orbit type
4. Check number and types of landing outcomes
5. Create new 'Class' column: Success=1 Failure=0

GitHub Link: [Data Wrangling]

EDA with Data Visualization

I created a five scatterplots using Seaborn:

- Payload Mass (kg) and Flight Number
- Launch Site and Flight Number
- Launch Site and Payload Mass (kg)
- Orbit and Flight Number
- Orbit and Payload Mass (kg)

I also made a bar chart to compare landing success (0-1) to Orbit type.

These charts are all used to help visual any patterns in the data
that are difficult to notice when they are plain text.

GitHub Link: EDA with Visualization

EDA with SQL

I made SQL queries for the following:

- Unique launch site locations
- Every launch from a Cape Canaveral site
- Total payload mass (kg) carried by boosters for NASA Commercial Resupply Services (CRS)
- Average payload mass (kg) carried by booster version 1.1
- Date for the first successful ground pad landing
- Boosters with successful drone-ship landing and a payload mass (kg) between 4000 and 6000
- Total count of successful and failure landings
- Boosters that have carried maximum payload mass (kg)
- Count of successful landing types between 2010-06-04 and 2017-03-20

GitHub Link: [EDA with SQL]()

Interactive Map with Folium

Using Folium, a Python library for interactive maps:

- I marked locations of SpaceX launch sites
- At each launch site, I added colored markers detailing landing outcomes
- At each launch site, I noted areas of interest and their distance to the site
    - Railways, Cities, Coastlines, Highways

These markers are all added to help visualize relationships of landing outcomes and factors around a site

GitHub Link: Data Visualization with Folium

Dashboard with Plotly Dash

I created an interactive dashboard that includes:

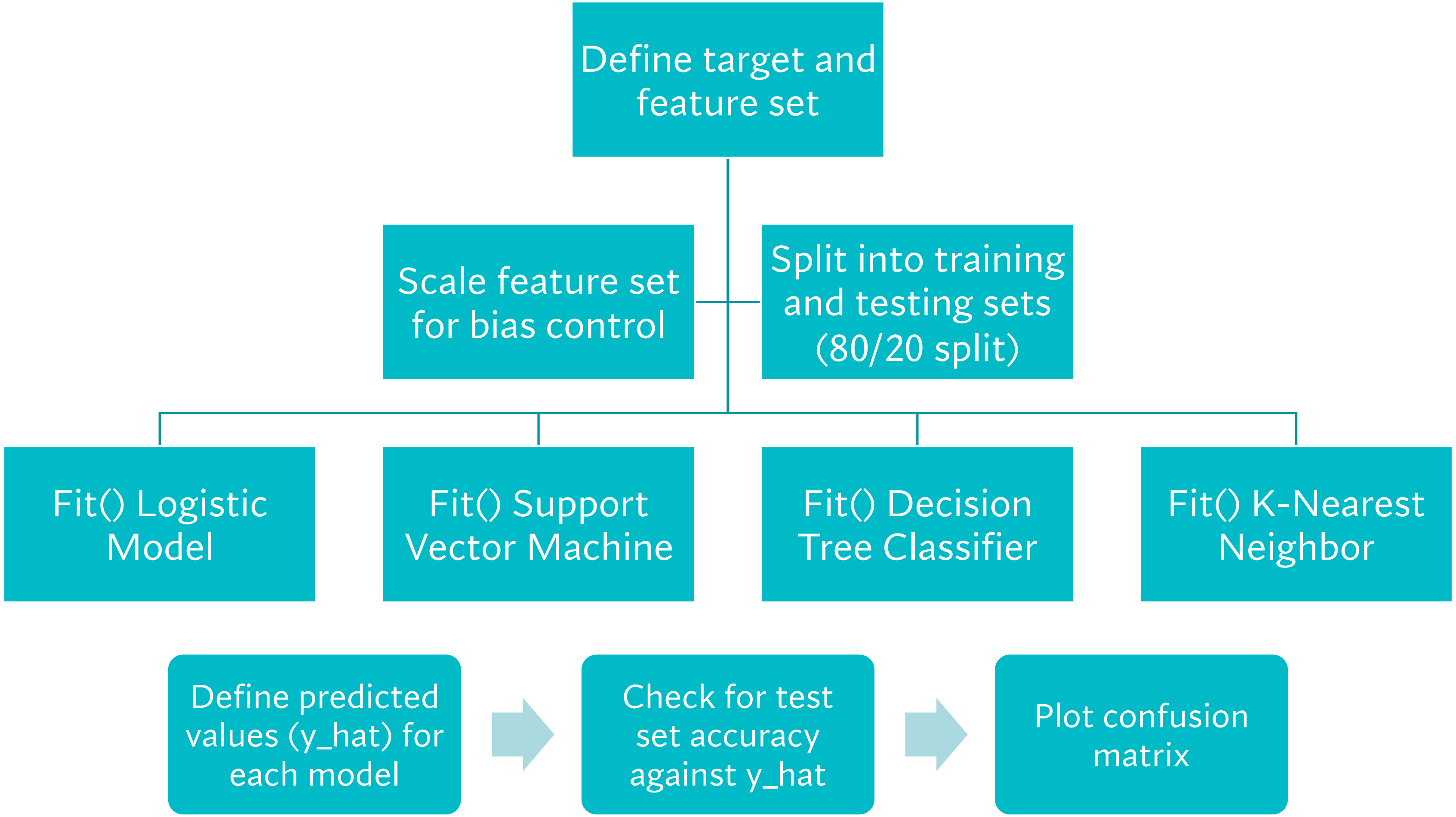
- Pie chart showing success rate of each site
- Scatter plot of success rate by booster version and payload mass (kg)
- A slider to specify payload mass range

These interactive visualizations allow quantitative information to be viewed and processed for pattern recognition.

GitHub Link: [SpaceX Dash App]

Predictive Analysis (Classification)

- I used GridSearchCV() to find most optimal hyperparameters for each model.
- Cross validation was set to 10 folds.



GitHub Link: [Machine Learning Predictions]

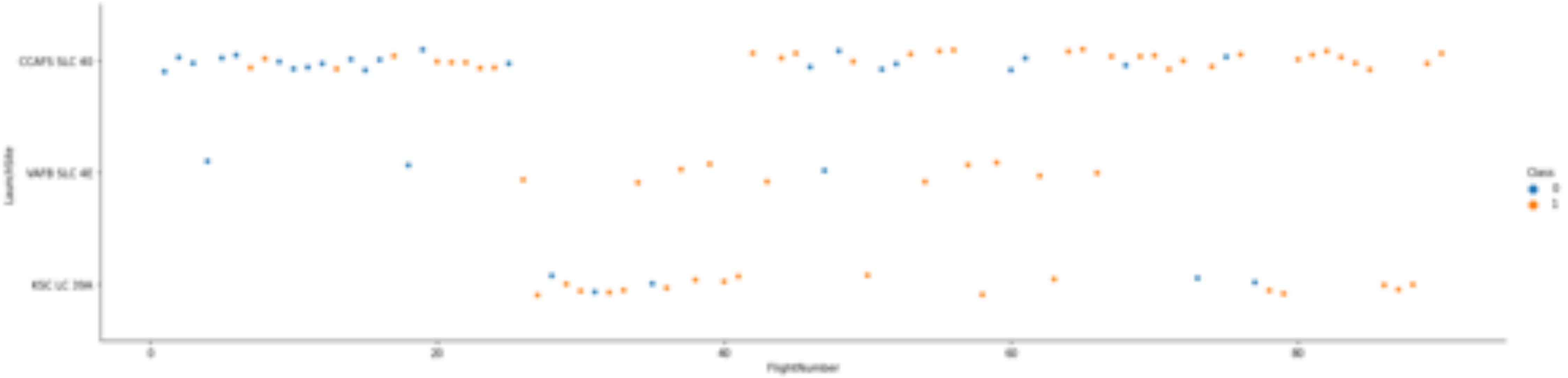
Results:

- Exploratory Data Analysis (EDA) Results
- Interactive Analytics Demo (Screenshots
- Predictive Analytics Results

EDA with Visualization

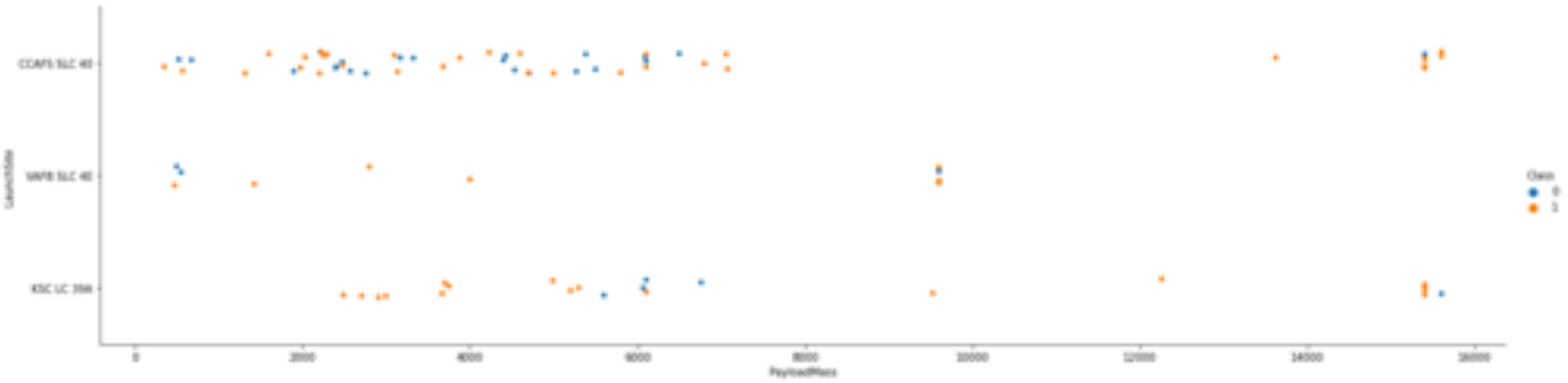
Flight Number vs. Launch Site

- From the scatter plot, we can see how launch site KSC LC-39A had a relatively successful landing rate with compared to site CCAFS SLC-40.
- We can also observe that site VAFB SLC-4E had a lesser concentration of launches.

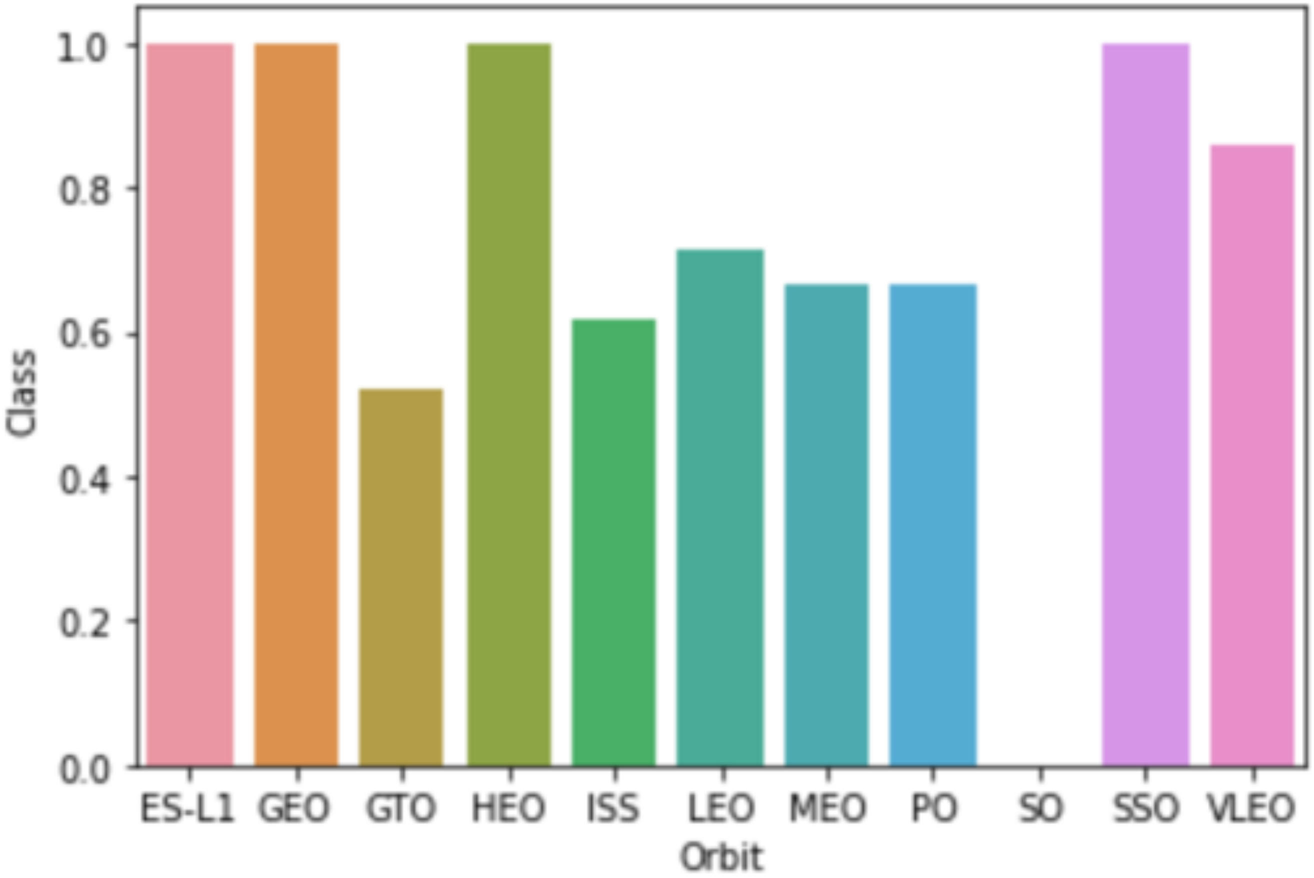


Payload vs. Launch Site

- Payloads over 8000kg are more proportionally successful in landing.
- Most of the payloads are between 0kg and 7000kg, and KSC LC-39A had the highest success rate with those payloads.



Success Rate vs. Orbit Type

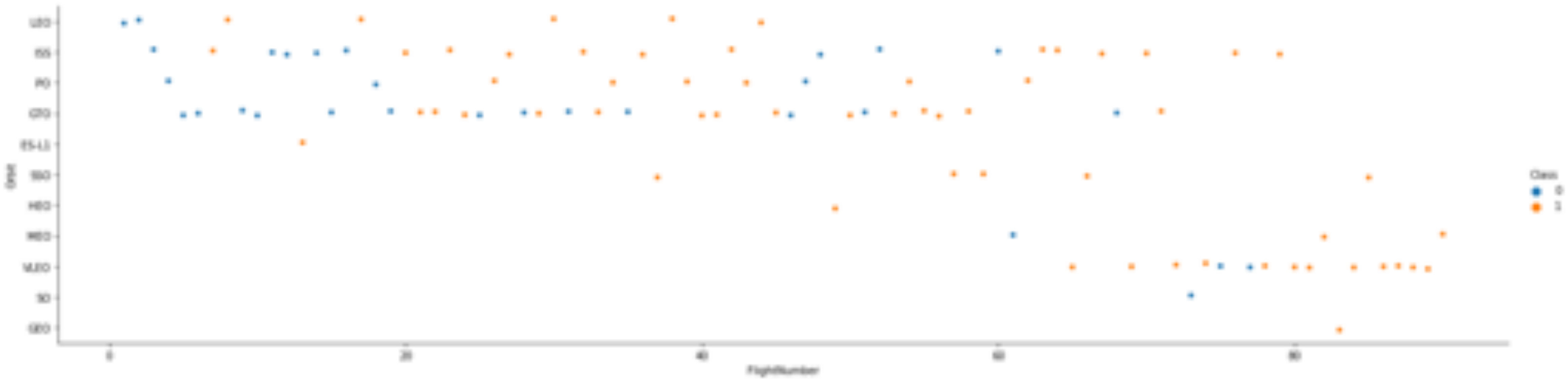


A visual analysis reveals 100% success rates with orbits:

- Lagrange point, ES-L1. This is a point in space with neutral gravity relative to the Earth and Sun.
- Geocentric orbit, GEO. Altitude of 22,236 miles above the equator following Earth's rotation.
- Highly Elliptical Orbit, HEO.
- Sun-synchronous orbit, SSO. This is a polar orbit around Earth defined at local mean solar time.

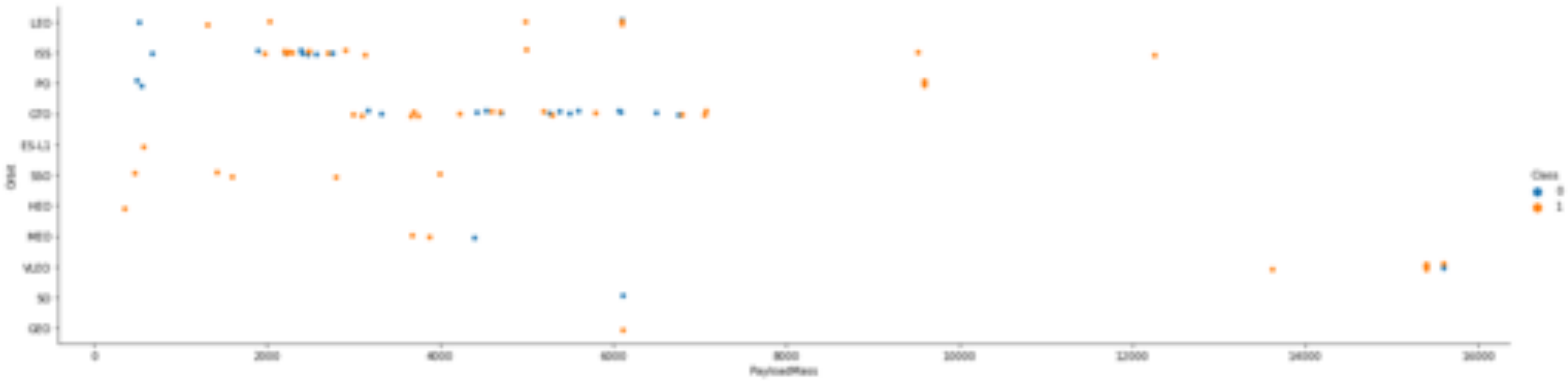
Flight Number vs. Orbit Type

- In Low Earth Orbit (LEO), landing success increased with flight number.
- In Very Low Earth Orbit (VLEO) there were two consecutive landing failures, otherwise success was consistent.



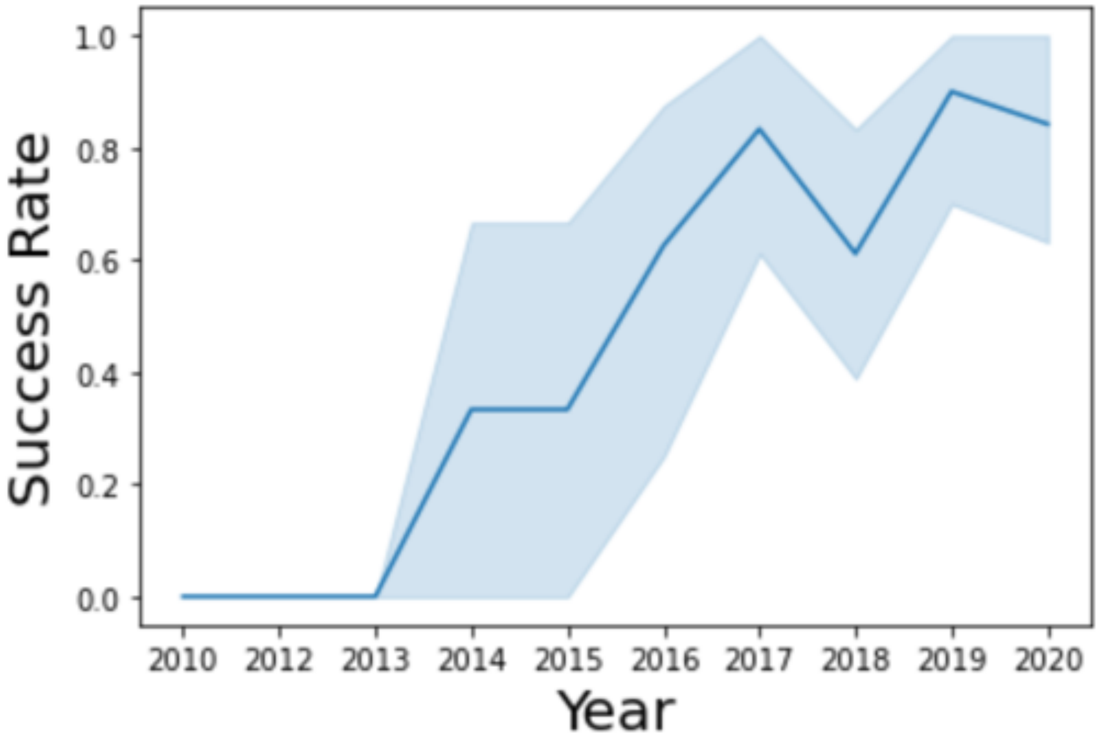
Payload vs. Orbit Type

- As noted previously, there is a relatively high success rate above payloads of 7000kg.
- There appears to be 100% success rate with low payload, Sun-synchronous orbits (SSO).



Launch Success Yearly Trend

- As shown by the line graph, success rate continued to increase until the year 2020.
- The shaded area shows the rest of the variation for that specific year.



EDA with SQL

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

All Launch Site Names

- There are four unique launch sites used for all Falcon 9 Booster versions.

Launch Site Names Begining With ‘CCA’

- These are five launches from sites beginning with CCA. This is the abbreviation of Cape Canaveral.

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

- NASA is a routine customer of SpaceX and sent several resupply missions.
- The total payload launched by SpaceX for NASA sums to 45,596 kilograms. (~100k pounds)

(Falcon 9 Payload Area)

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Of the many versions of Falcon 9, version 1.1 was found to have an average payload mass of 2,928 kilograms. (~6200 pounds)

booster_version	avg_payload_mass
F9 v1.1	2928

First Successful Ground Landing Date

- A major accomplishment for SpaceX was their first ground pad landing. This occurred five years after they began the project.

first_successful_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload Between 4000 and 6000

- These are the Falcon 9 booster versions that have success landing on the drone ship within this payload range.

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

- There was struggle with landing success throughout the analyzed period, but it is important to note that almost all mission outcomes were successful for their customers.
- In other words, SpaceX can reliably launch payloads to their intended orbit with nearly 100% success.

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

Boosters Carrying Maximum Payload

- These are all Falcon 9 Booster versions that carried the maximum payload allowed.

payload_mass__kg_	booster_version
15600	F9 B5 B1048.4
15600	F9 B5 B1049.4
15600	F9 B5 B1051.3
15600	F9 B5 B1056.4
15600	F9 B5 B1048.5
15600	F9 B5 B1051.4
15600	F9 B5 B1049.5
15600	F9 B5 B1060.2
15600	F9 B5 B1058.3
15600	F9 B5 B1051.6
15600	F9 B5 B1060.3
15600	F9 B5 B1049.7

Success Count Between 2010-06-04 and 2017-03-20

- Over a nearly seven-year time span from June 2010 to March 2017, SpaceX performed eight successful landings with Falcon 9 boosters. Five on drone ships and three on ground pads.

landing__outcome	COUNT
Success (drone ship)	5
Success (ground pad)	3

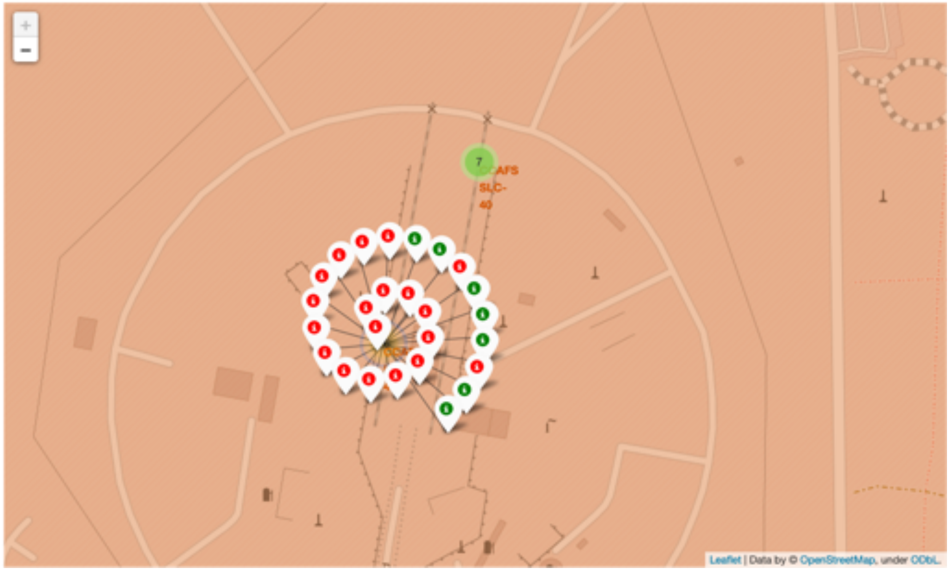
Interactive Map with Folium

Map of Launch Sites

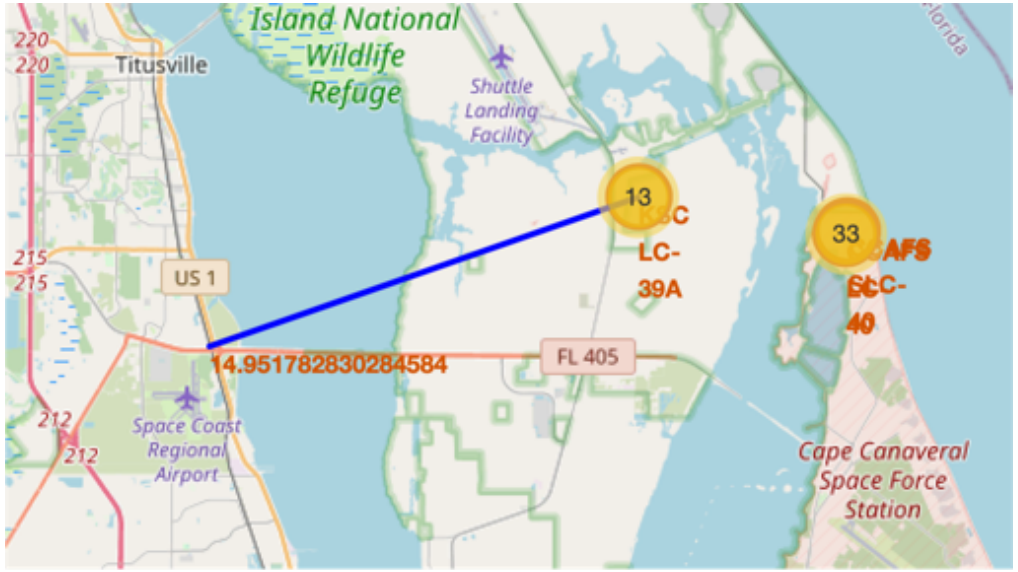
- The launch sites are located on the southern half of each United States coast.
- These are nearing the sections of the country closest to the equator.
- This is likely due to moderate climate and immediate ocean access.

Launch Site Outcomes

- This is a visual representation of the landing outcomes corresponding to each launch site.
- The exact ground pad and drone ship coordinates are not specified in the data, so landing status is marked at the launch locations.
    - Green = Success
    - Red = Failure



Launch Site Proximity

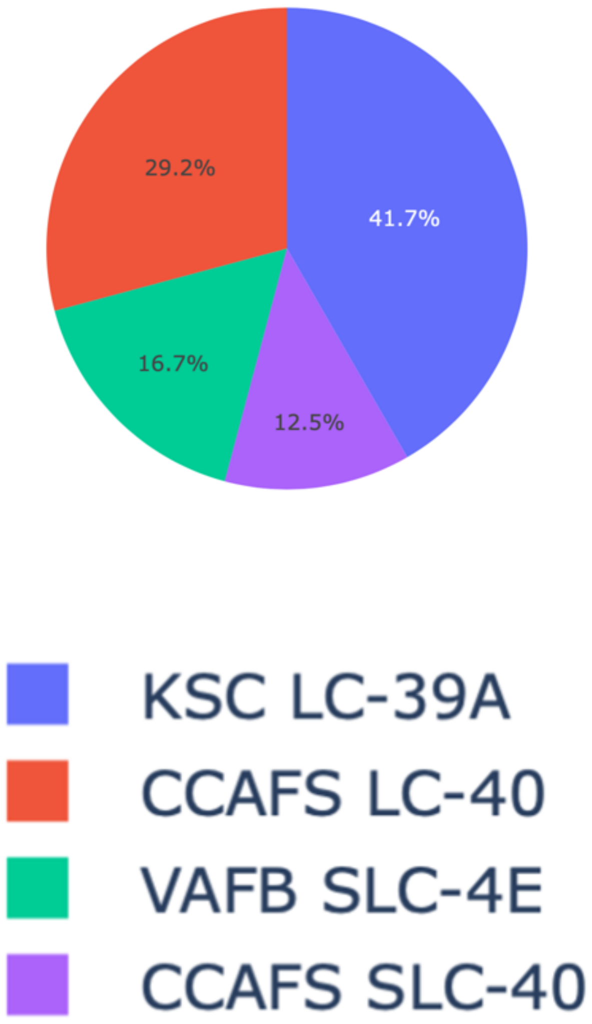


- This is to outline the launch site proximity to geographical features.
- The four sites are all within a single kilometer of water, but nearly 15 kilometers from major highways.

Dashboard with Plotly Dash

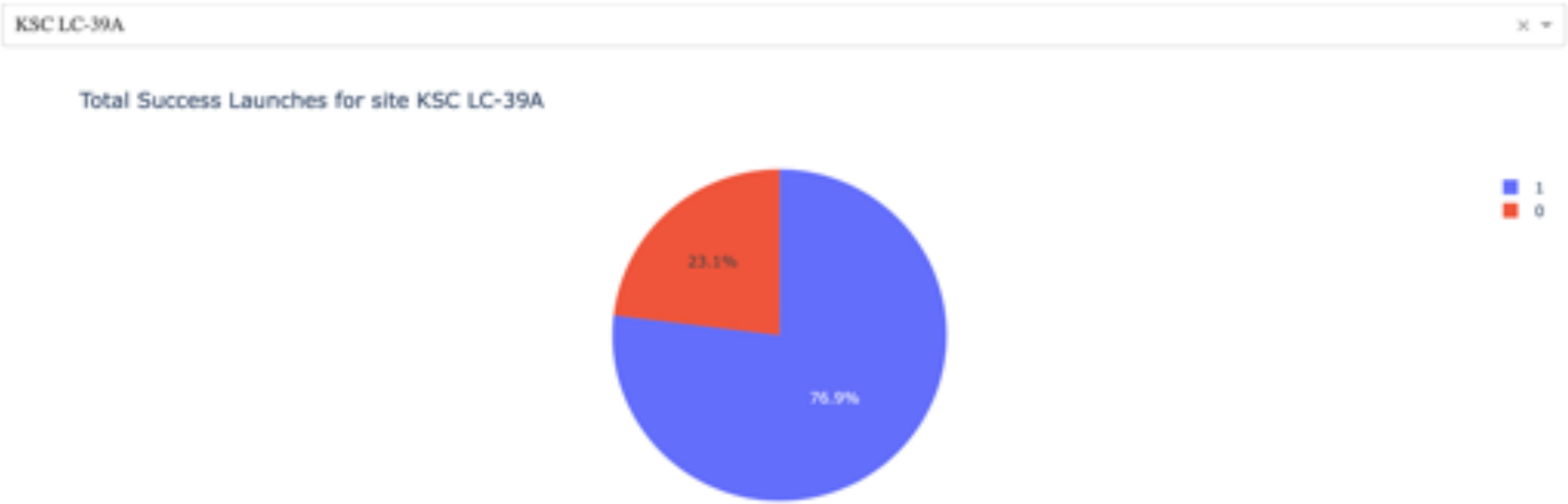
Launch Site Success

- This pie chart is used as a visual indicator of success per site. A pie chart is useful at showing relative comparisons of these locations.
- Sites KSC LC-39A and CCAFS LC-40 make up over 70% of successes.



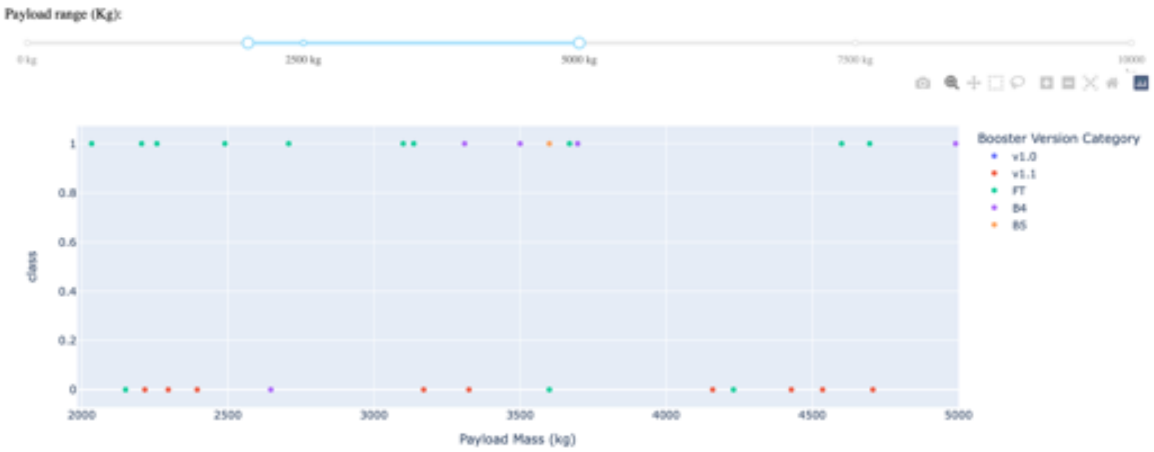
Most Successful Site

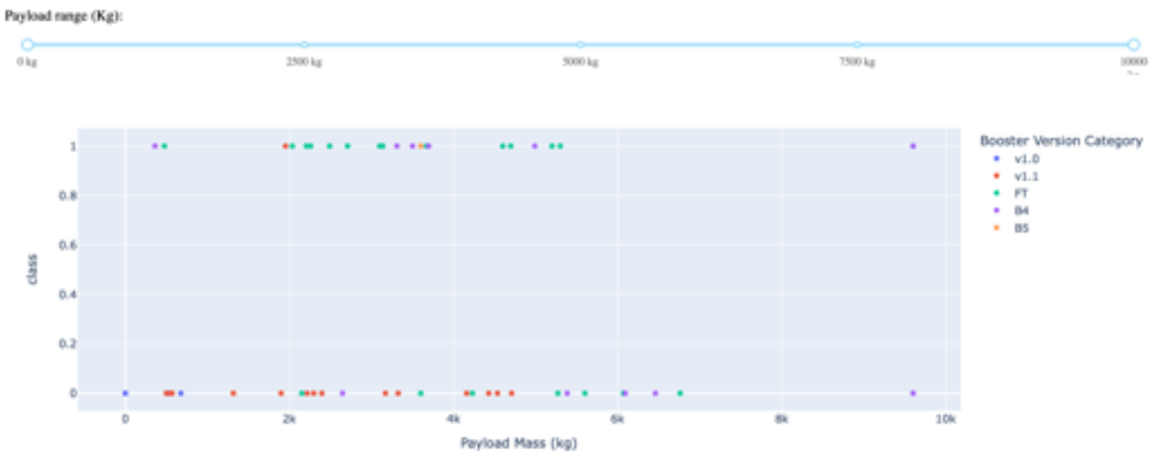
- Site KSC LC-39A had the most success relative to the other sites.
- Checking the actual success rate of KSC LC-39A shows that over 76% of landings succeeded!



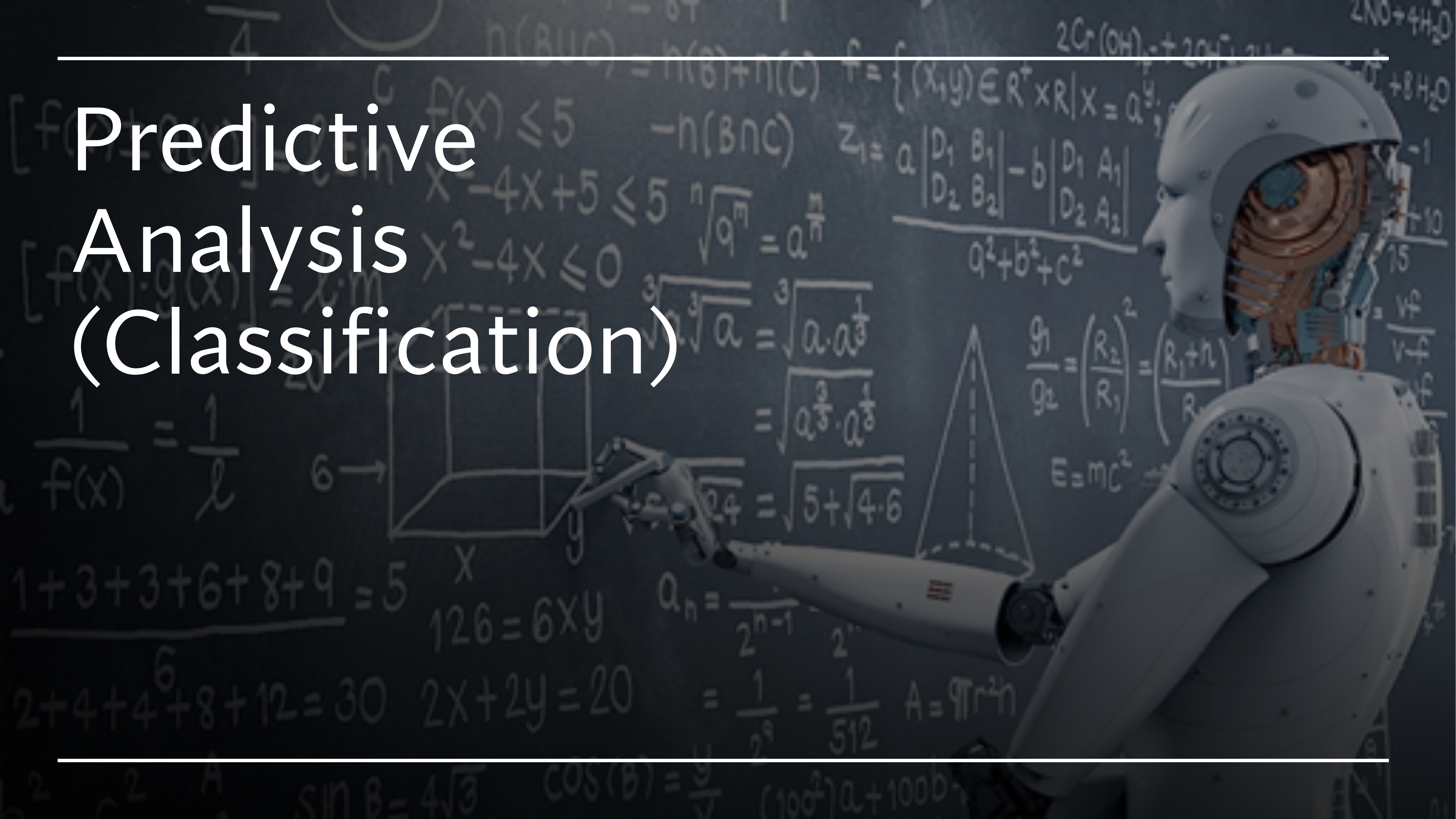
Payload Range and Success by Booster Version

- These scatter plots show the landing outcome and Payload Mass (kg) with each booster version.
- Between 2000kg and 5000kg:
  - Falcon 9 version FT had a good success rate
  - Falcon 9 version v1.1 failed every landing





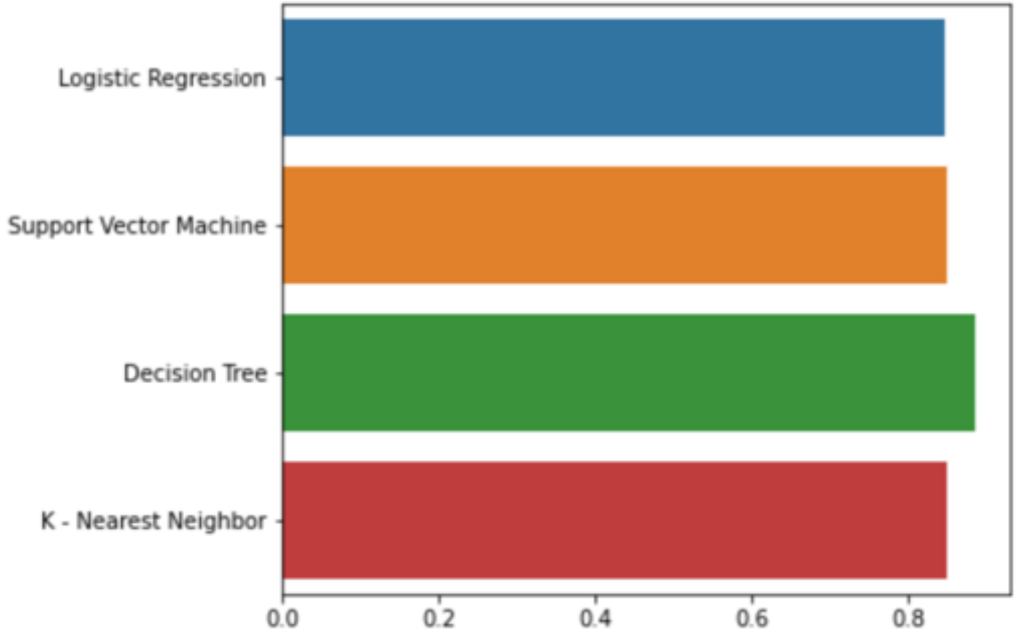
Predictive Analysis (Classification)



Classification Accuracy

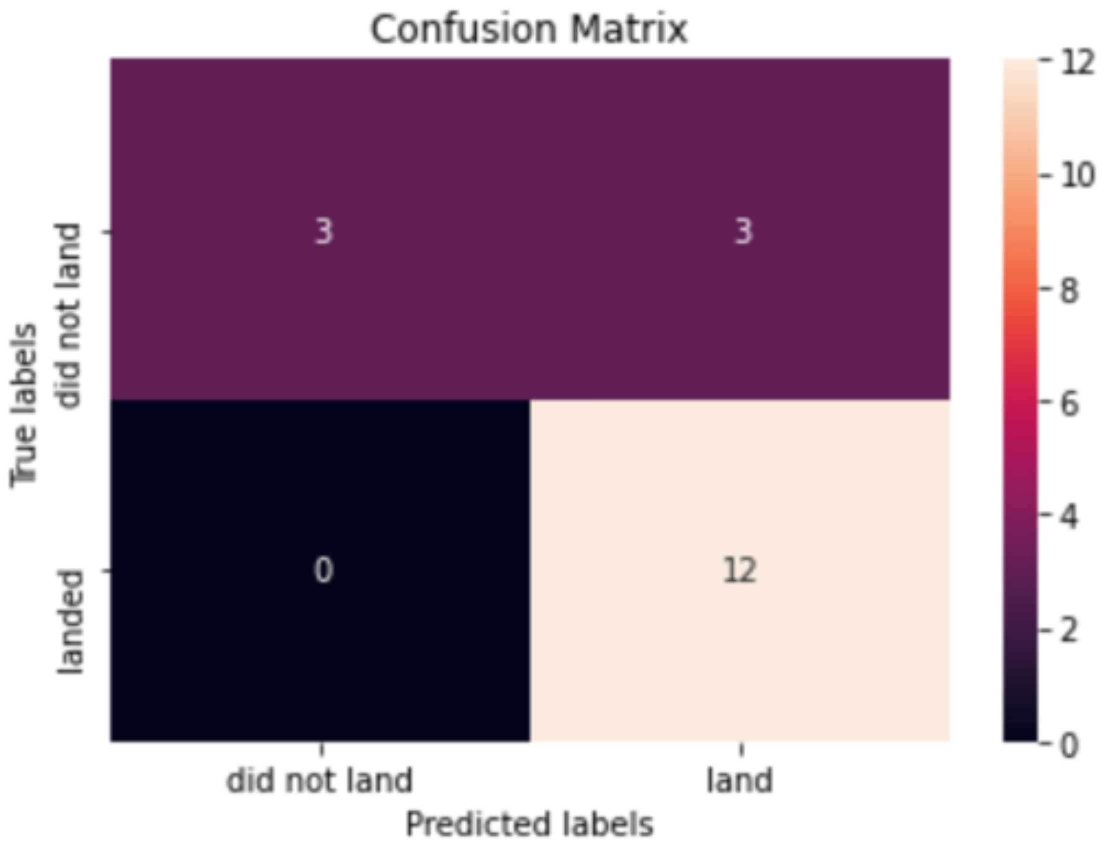
- Among the four models trained on the training set, the Decision Tree model was able to produce the highest level of accuracy.

Models	Scores
Logistic Regression	0.846429
Support Vector Machine	0.848214
Decision Tree	0.885714
K - Nearest Neighbor	0.848214



Confusion Matrix for Decision Tree

- This confusion matrix displays vital information regarding the True Positive and False Positive outcomes.
- The decision tree was able to correctly predict the outcome of a landed booster with 100% accuracy.
- However, it's accuracy on determining the outcome of boosters that did NOT land is 50%.
- This is considered a Type 2 Error.



Conclusion

- As the number of flights increased, the proportion of successful landings increased.
- Launches from KSC LC-39A only occurred after flight number 25 and that site had the greatest success rate.
- The Decision Tree classifier produced the greatest accuracy, but due to the number of false positives, more tuning is needed.
- The limited observations in the data set also prevents higher accuracy. As more landings are performed, model accuracy should increase.

Appendix

All code can be found in my IBM Data Science Capstone repository on GitHub.

Repository Link: [IBM Data Science Capstone]