Winning Space Race with Data Science

Benjamin Hotchner
June 29th, 2022

Outline

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

Executive Summary

An analysis of SpaceX business model; in particular, the range of clientele given the variety of launch sites, payload capacity, booster types, and orbit types available. Lastly, the analysis will analyze the likelihood a given rocket is successfully landed for re-use later.

The methodology is constructed via data collection and preparation for analysis, descriptive data analysis, data visualization with charts and maps, and predictive data analysis.

Introduction

- Project Background: This project is the last step in completing the IBM Data Science certification.
- Problem Statement: What are the capacities and strengths of the Spacex business model? How & to what extent can it be competed with given current facts?

Section 1

Methodology

Methodology

Executive Summary

- Data collection
- Data wrangling
- Performed exploratory data analysis (EDA) using visualization and SQL to describe key factors at Spacex.
- Visual analytics using Folium and Plotly Dash to present a geospatial and interactive presentation of Spacex history and success rate.
- Predictive analysis using classification models, including logistic regression, decision trees, support-vector machines, and confusion matrices.

Data Collection – SpaceX API

- Github URL [Here]

SpaceX API Steps

1. Download data, and input it into columns provided by Coursera. Create a pandas dataframe.
2. Filter Falcon 1 (retired rocket) from dataframe, leaving only Falcon 9 and Falcon Heavy launches.
3. Find out how many missing values exist in the data frame. Replace missing values in Payload Mass with a mean of the existing data for the category.

Data Collection - Web Scraping

- Github URL [Here]

Web Scraping Steps

1. Using BeautifulSoup package, access Wikipedia list of Falcon 9 and Falcon Heavy launches. Create a data frame by extracting title, column names, and row data.

Data Wrangling

- Github URL Here

Data Wrangling Steps

1. Calculate number of missing values in table, and identify whether variables are numerical or categorical.
2. Calculate the number of launches at each site, and the number of launches/orbit type.
3. Calculate the mission outcomes per orbit type.
4. Create a landing outcome variable to count whether the rocket successfully landed or not following launch (important for future analysis)

EDA w/ Data Visualization

- Github URL [Here]

EDA/DV Steps

1. After downloading table and initial exploratory analysis, constructed scatter plots of Payload Mass vs. Flight Number, Flight Number vs. Launch Site, Payload vs. Launch Site, Success Rate vs. Orbit Type, Flight Number vs. Orbit Type, Payload vs. Orbit Type.
2. Followed with a bar chart visualizing launch success as yearly trends.
3. Features Engineering: create dummy variables to categorical columns (Orbits, Launch Site, Landing Pad, and Serial #). Casted all numeric columns to float64 for easier coding later.

Build an Interactive Map with Folium

- Using the Folium package, I created maps to demonstrate a) from where SpaceX launches its rockets, b) success rate at given facilities, and c) the proximity of CCAFS SLC-40 to railroads, highways, the ocean, and cities.
- Github URL Here

Build a Dashboard with Plotly Dash

- Using plotly, I created a dashboard which depicts the success rate of a) SpaceX launches across its facilities (can be conditioned on payload), and b) a comparison of launch outcome depending on the payload capacity.
- This allows us to understand a) the success rate at different facilities, and b) the extent to which payload capacity affects that success rate.
- Github URL Here

Predictive Analysis (Classification)

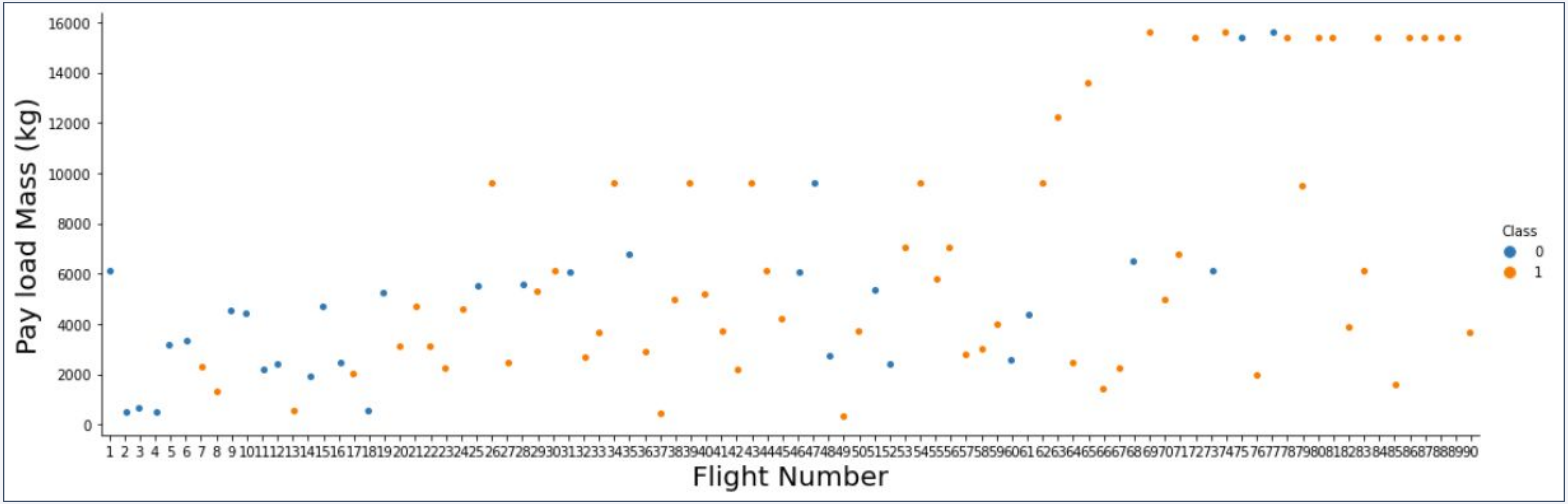
- I built three models to predict the likelihood of a given rocket booster landing following launch. This is important for understanding the likelihood SpaceX can keep its prices down (by extending the life cycle of rockets currently in use).
- The three models are K Nearest-Neighbor, Logistic Regression, and a Decision Tree. As will be shown, the Decision Tree is the best predictor for rocket success or failure.
- Github URL Here

Section 2

Insights drawn from EDA

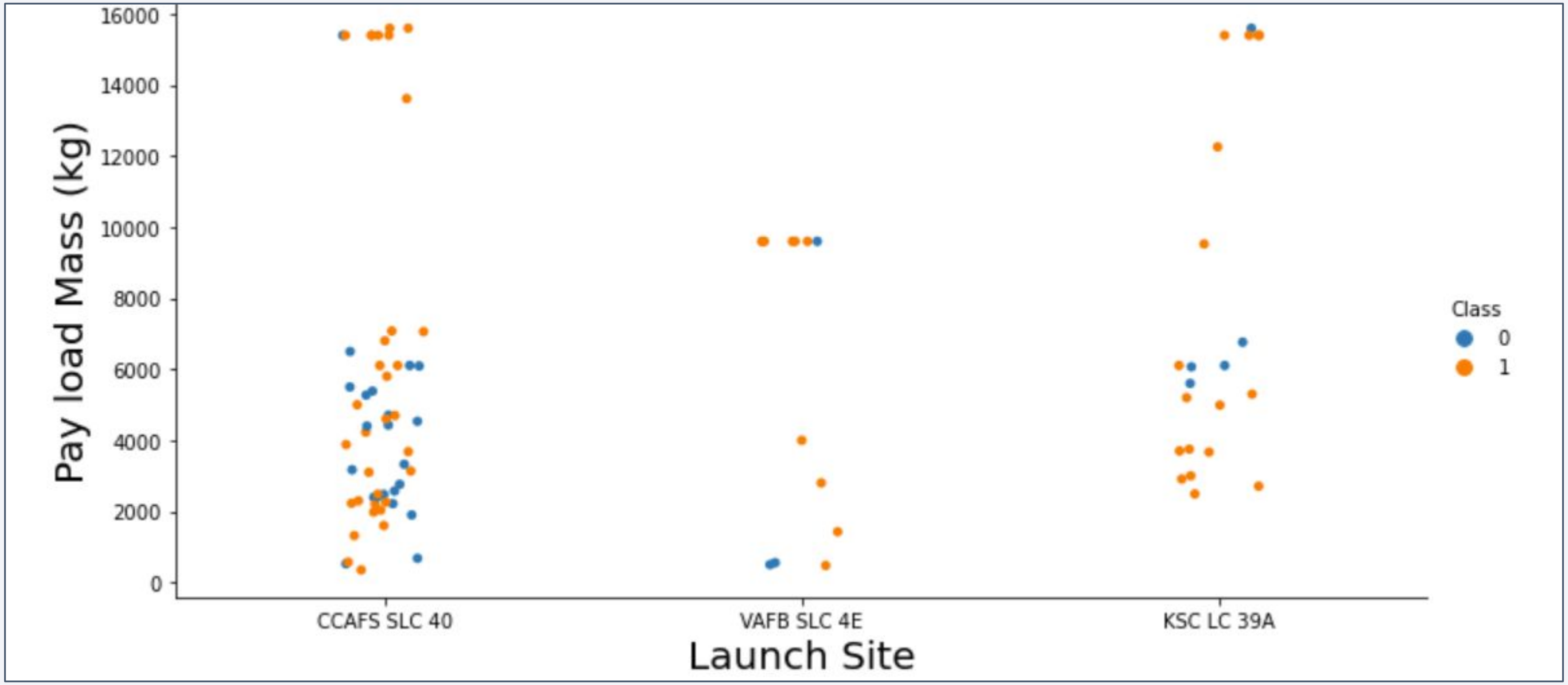
Flight Number vs. Payload

- An increase in payload capacity with time/flight number, suggesting that SpaceX is improving its product continuously. A payload mass above 10000 kg is now the norm.



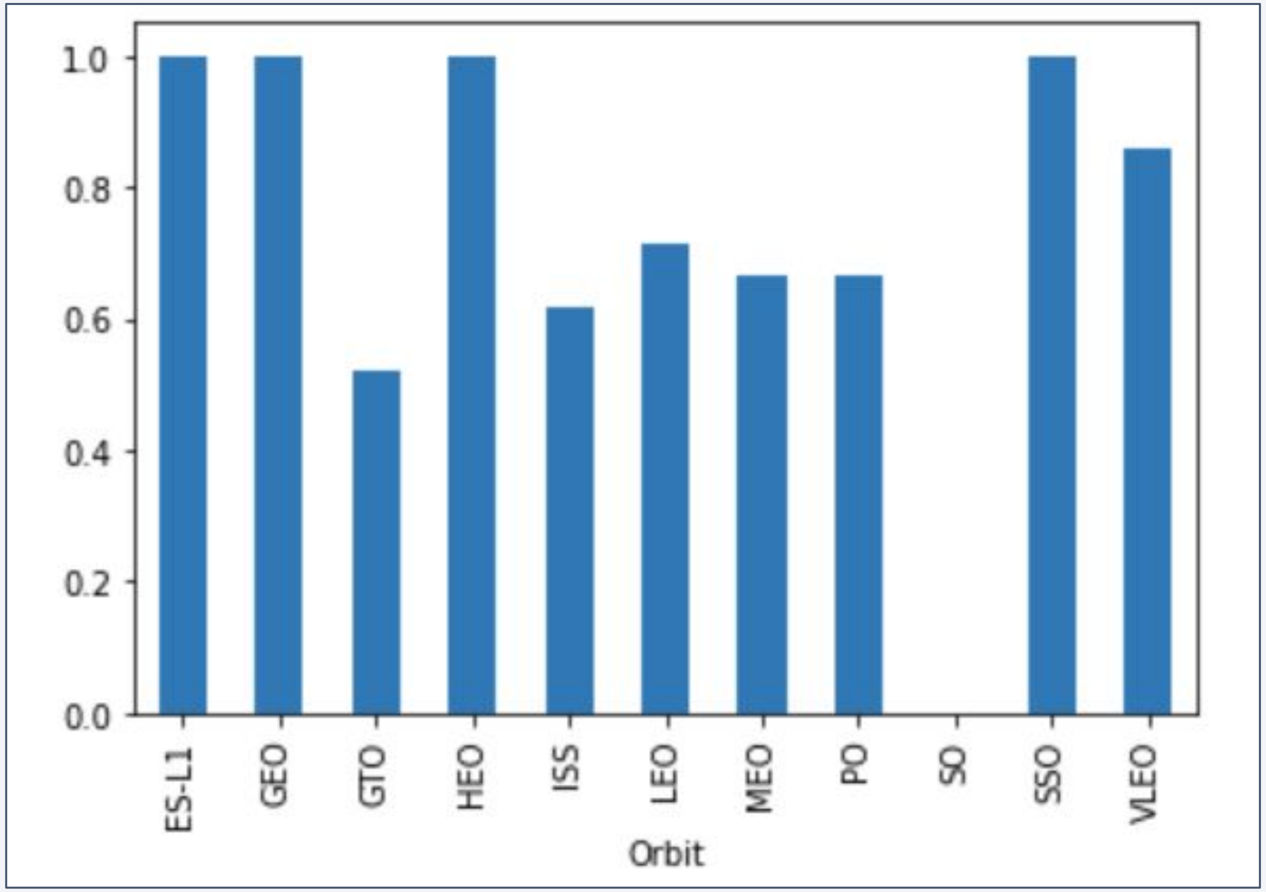
Payload vs. Launch Site

- *CCAFS SLC 40* has the most use and widest range of payloads, with the majority below 8000 kg but a significant number of flights above 12 000 kg.
- VAFB SLC 4E has the least use and smallest payload range, with a roughly equal number of flights less than 5000 kg and 10 000 kg.



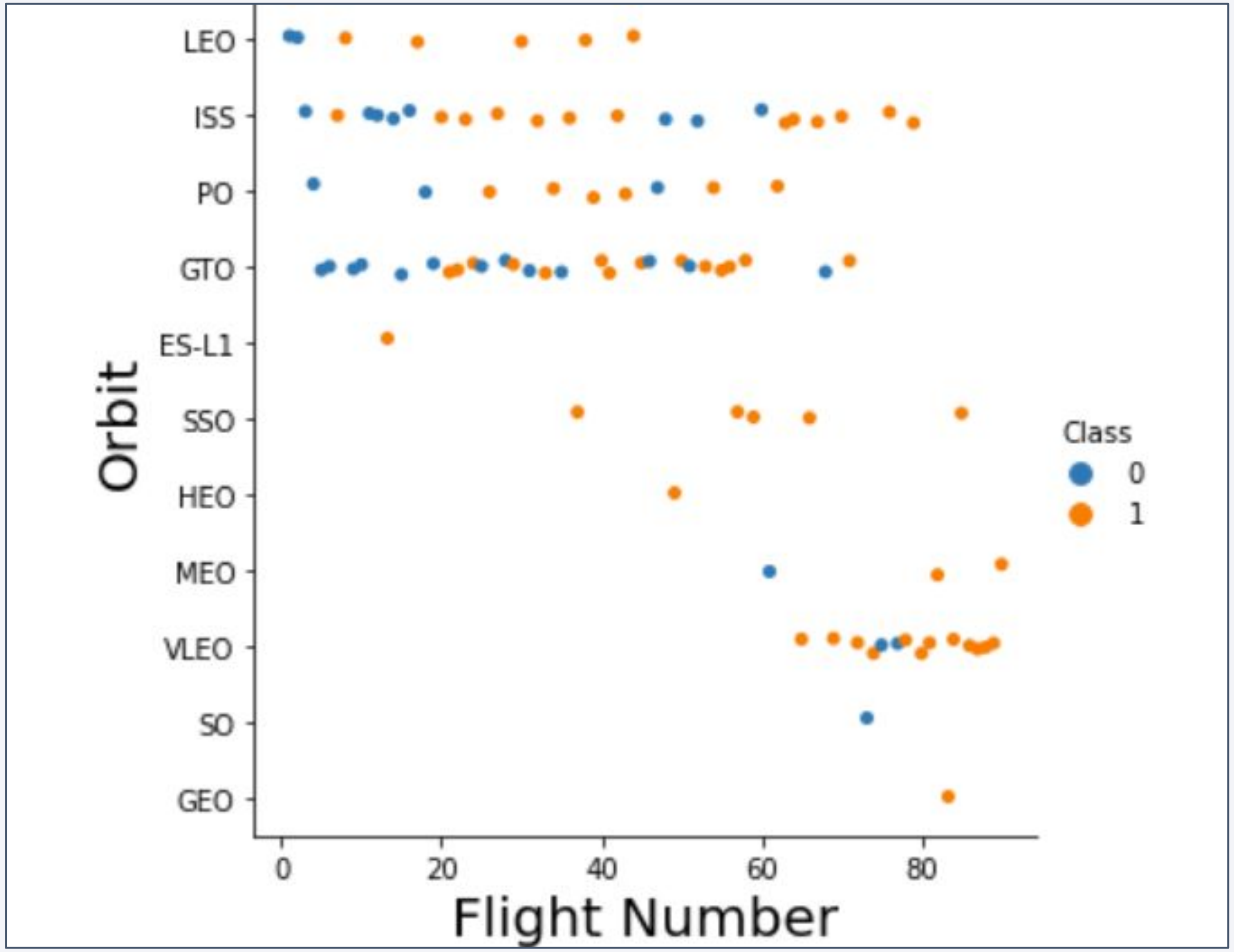
- KSC LC 39A has a medium level of use and range, with payloads ranging from 2000 to 16 000 kg.

Success Rate vs. Orbit Type



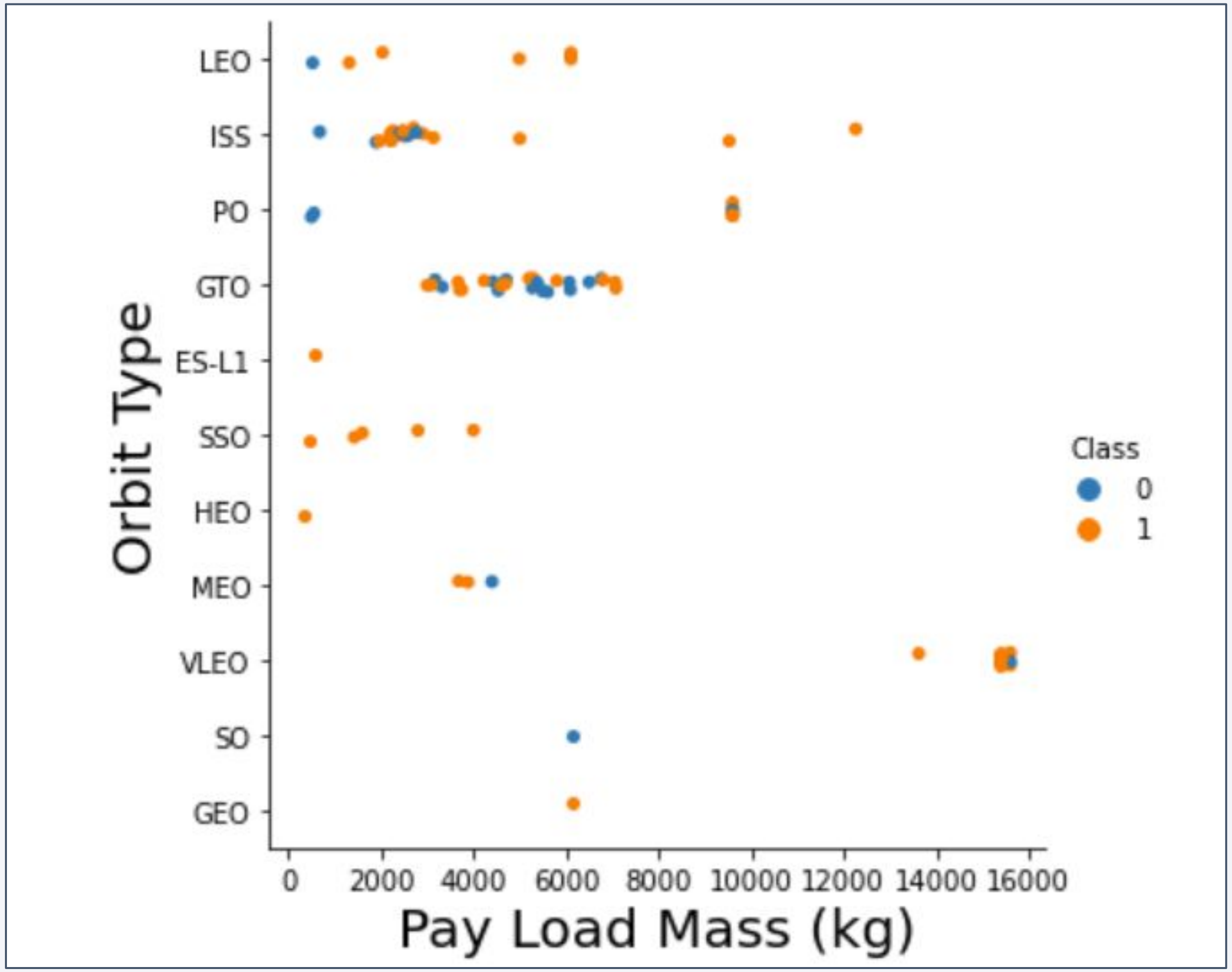
- ES-L1, GEO, HEO, and SSO orbits have the highest success rate (near 100%)
- GTO, ISS, LEO, MEO, PO, SO have the lowest success rate (70% and lower)

Flight Number vs. Orbit Type



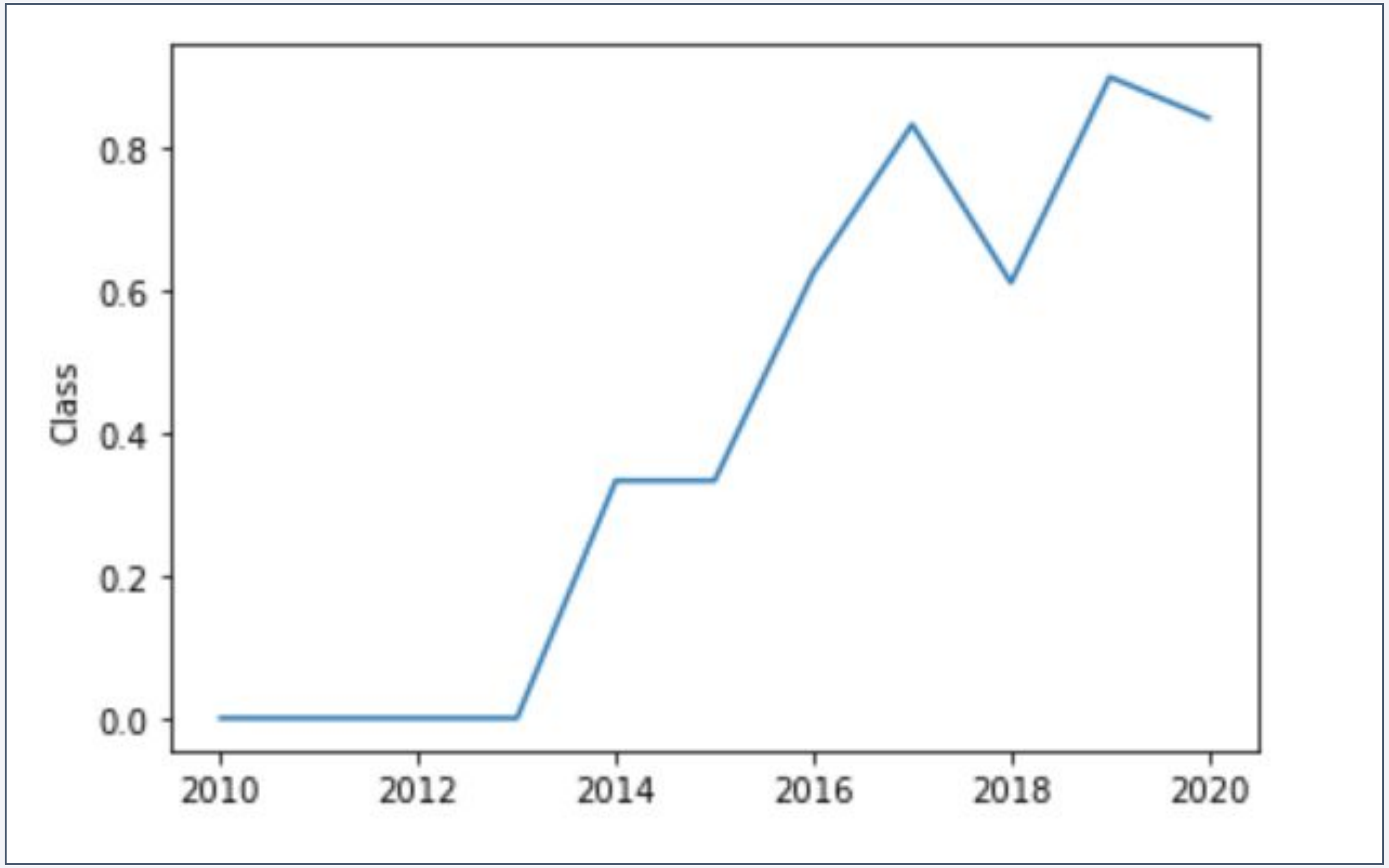
- A shift from LEO, ISS, PO, and GTO orbits to SSO, MEO, VLEO, SO, and GEO orbits after Flight #40.

Payload vs. Orbit Type



- VLEO has the highest average payload capacity (above 14 000 kg)
- LEO, PO, GTO, ES-L1, SSO, HEO, MEO, SO, and GEO have the lowest payload capacity (all less than 10 000 kg)
- ISS has the widest range of payload capacity, from less than 2000 kg to 13 000 kg.

Launch Success Yearly Trend



- Launch success increased markedly in 2013 from 0% to over 80% in 2020, after a small drop in 2018.

All Launch Site Names

```
%sql select distinct LAUNCH_SITE from SPACEXTBL
```

	Launch Site	Lat	Long	Marker_color
0	CCAFS LC-40	28.562302	-80.577356	NAN
1	CCAFS SLC-40	28.563197	-80.576820	NAN
2	KSC LC-39A	28.573255	-80.646895	NAN
3	VAFB SLC-4E	34.632834	-120.610745	NAN

Launch Site Names Begin with 'CCA'

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA%';
```

Answer: 99980 kg

Average Payload Mass by F9 v1.1

%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';

Answer: 2534.67 kg/payload

Github URL [Here]

First Successful Ground Landing Date

%sql SELECT MIN(DATE) FROM SPACEXTBL where 'Landing _Outcome' = '%Success%';

Answer: 22-12-2015

Github URL [Here]

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

- Present your query result with a short explanation here

Github URL Here

Total Number of Successful and Failure Mission Outcomes

%sql SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%'

Answer: 1 Launch

%sql SELECT COUNT(*) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%'

Answer: 100 Launches

Github URL [Here]

Boosters Carried Maximum Payload

%sql select booster_version, MAX(PAYLOAD_MASS__KG_) AS PAYLOAD_MAX from spacextbl

Answer: F9 B5 B1048.4 (with payload max of 15 600 kg)

Github URL Here

2015 Launch Records

%sql select substr(Date,4,2), 'landing _outcomes', Booster_Version, launch_site from spacextbl where substr(Date,7,4)='2015' and 'landing _outcomes' LIKE 'Failure (drone ship)'

Github URL Here

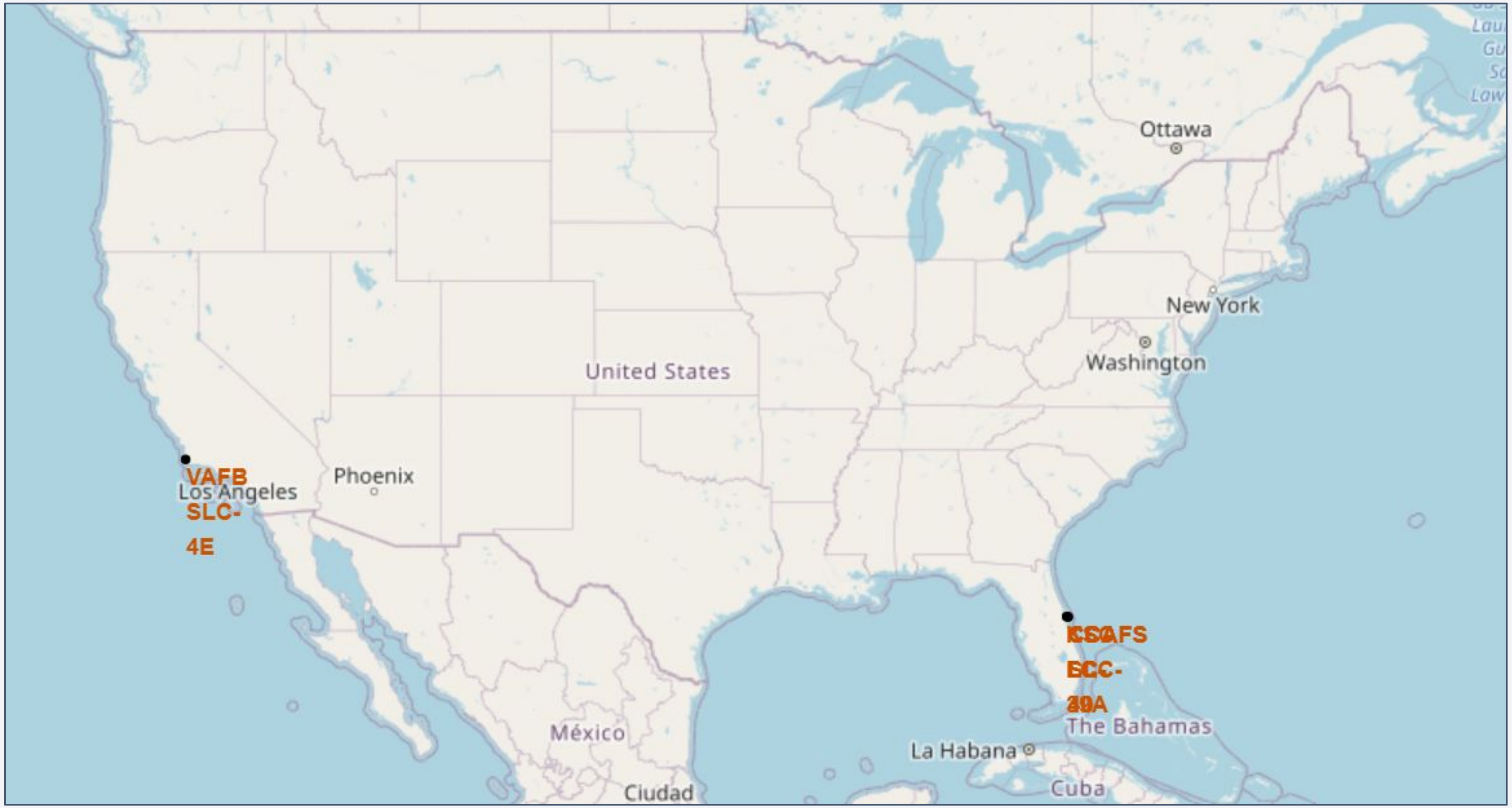
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Github URL Here

Section 3

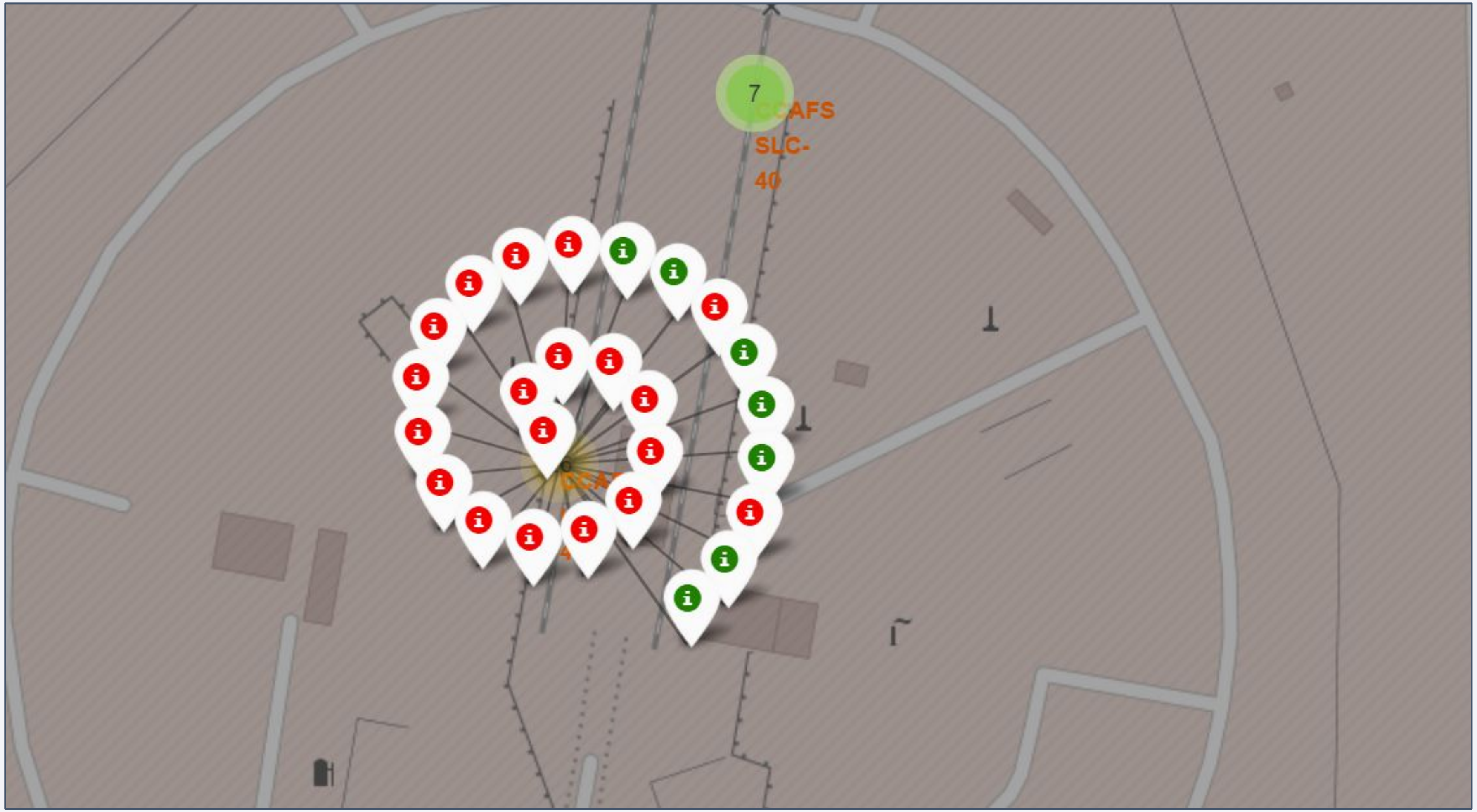
Launch Sites Proximities Analysis

Map 1 of SpaceX Facilities (using Folium)



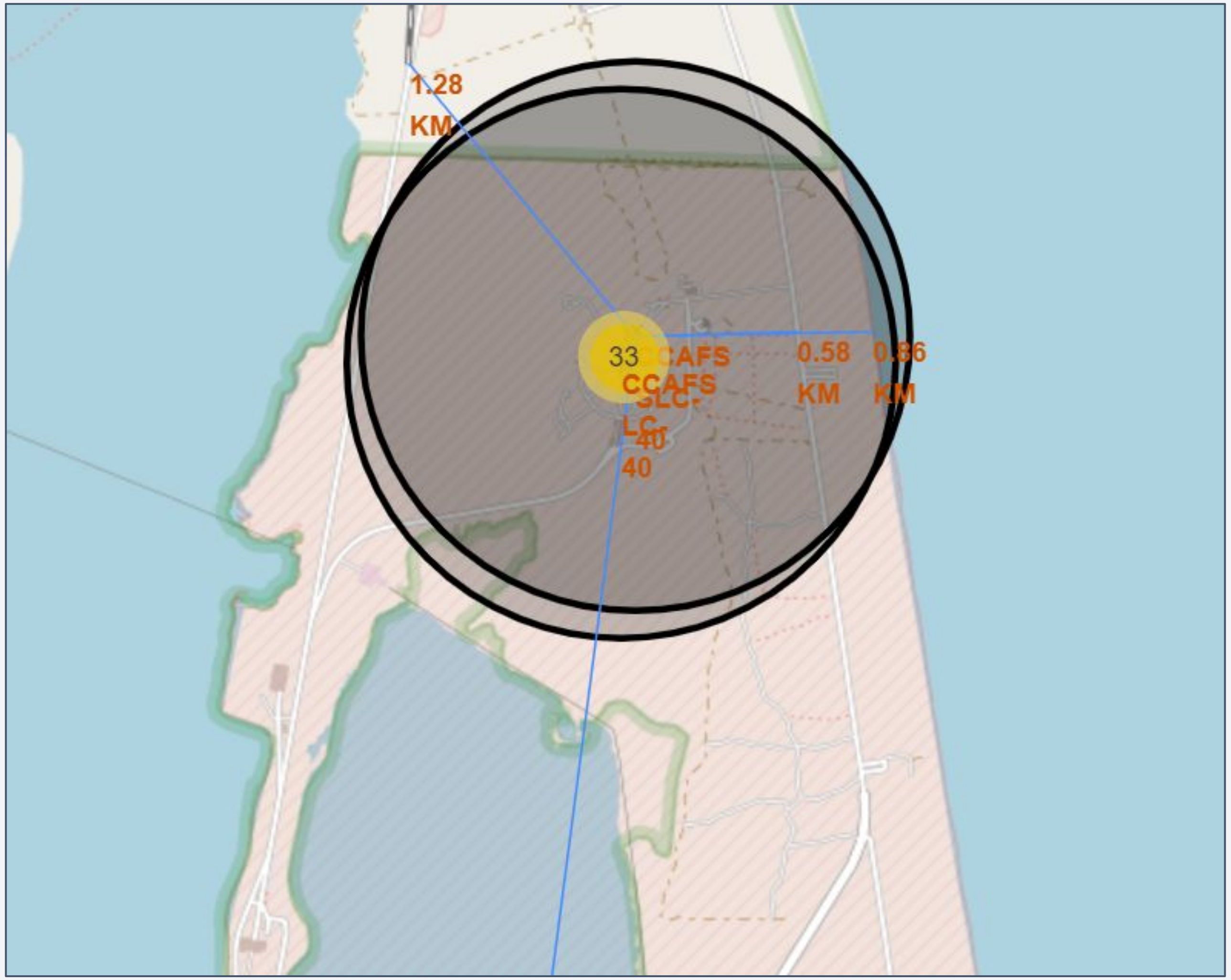
- SpaceX has facilities in southern California and Florida.

Map 2 of SpaceX Facilities (using Folium)



- Flight success rate at CCAFS LC-40, showing that the vast majority of the flights have ended in Failure. Fortunately, things have improved since!

Map 3 of SpaceX Facilities (using Folium)



- CCAFS LC-40 is ubicated in south Florida, and is related to important geographical markers, including:
  - 0.86km from the ocean
  - 1.28km from the nearest railroad
  - 51.43km from the nearest city (Orlando/Melbourne)

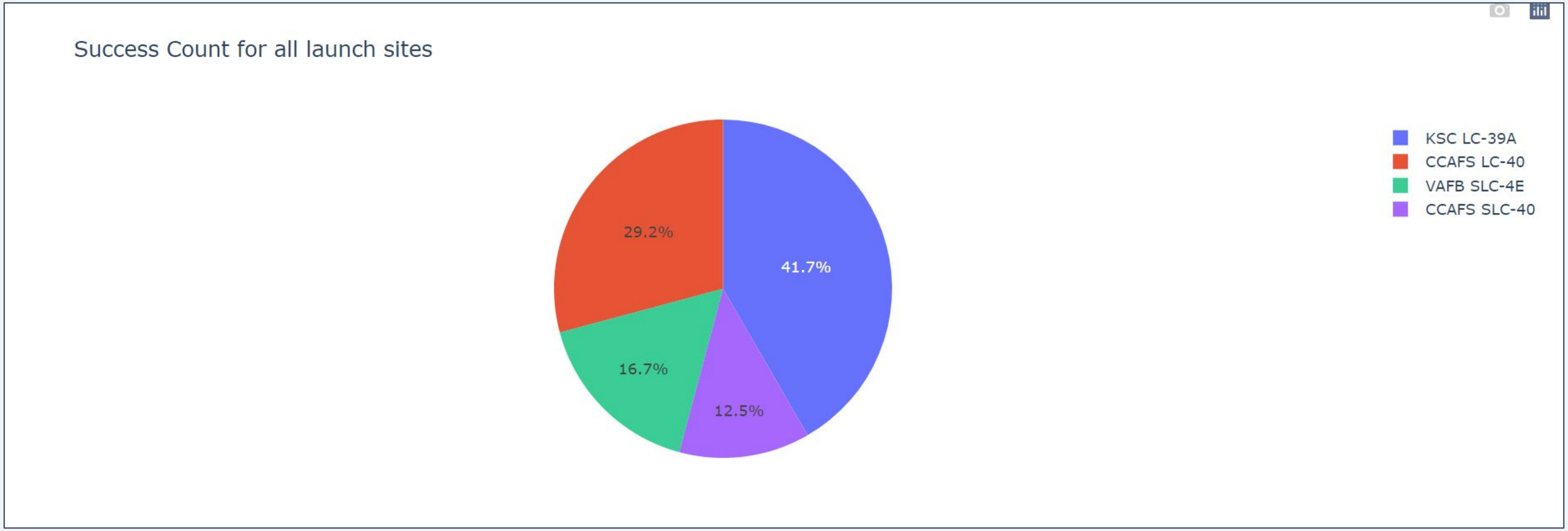
This is important because it minimizes sound pollution to the city, while a) maximizing access to supplies via railroad and b) minimizing rocket risks given rockets launch and land near the ocean.

Section 4

Build a Dashboard with Plotly Dash

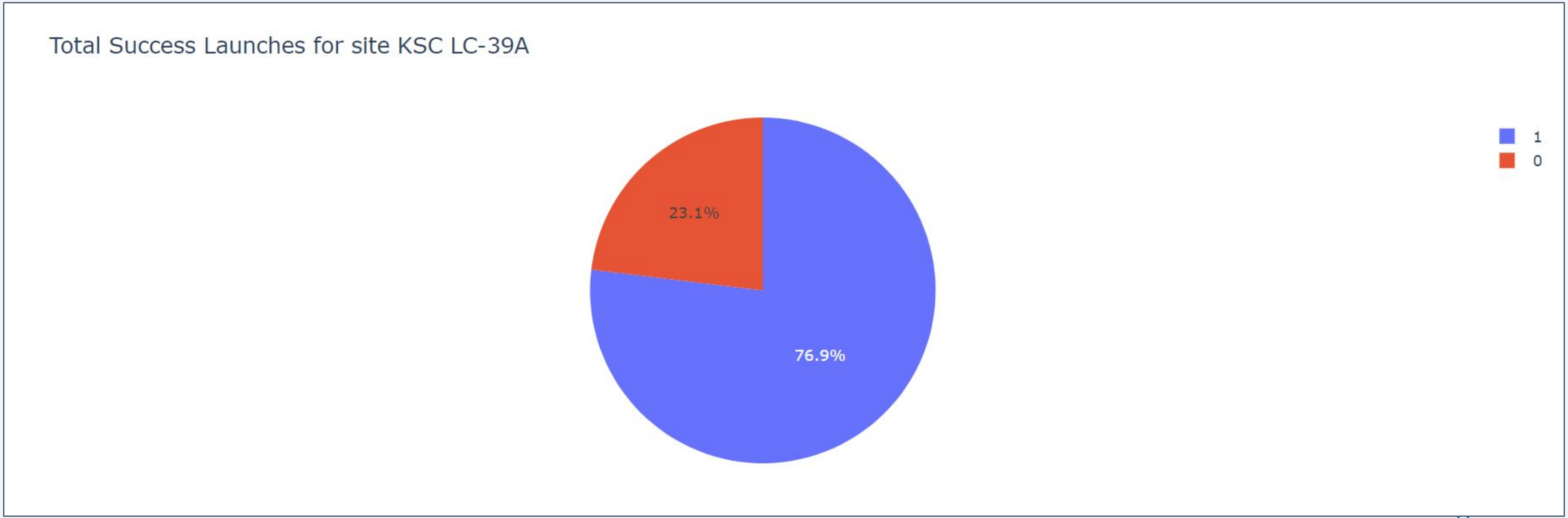
Dashboard - Pie Chart of Launch Success at ALL

Of all successes, where are they distributed by site? KSC LC-39A is the most successful, followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40, respectively.



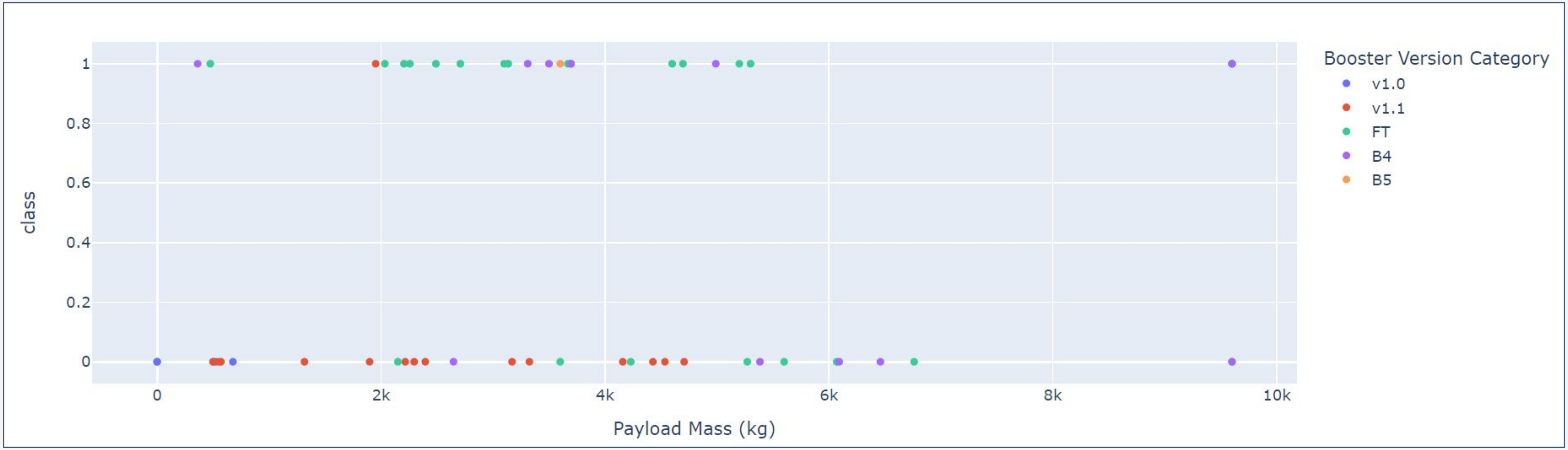
Dashboard - Pie Chart of Launch Success at X

Of all launches at KSC LC-39A, only 76.9% have been successful.



Dashboard - Payload (kg) vs. Launch Outcome at ALL

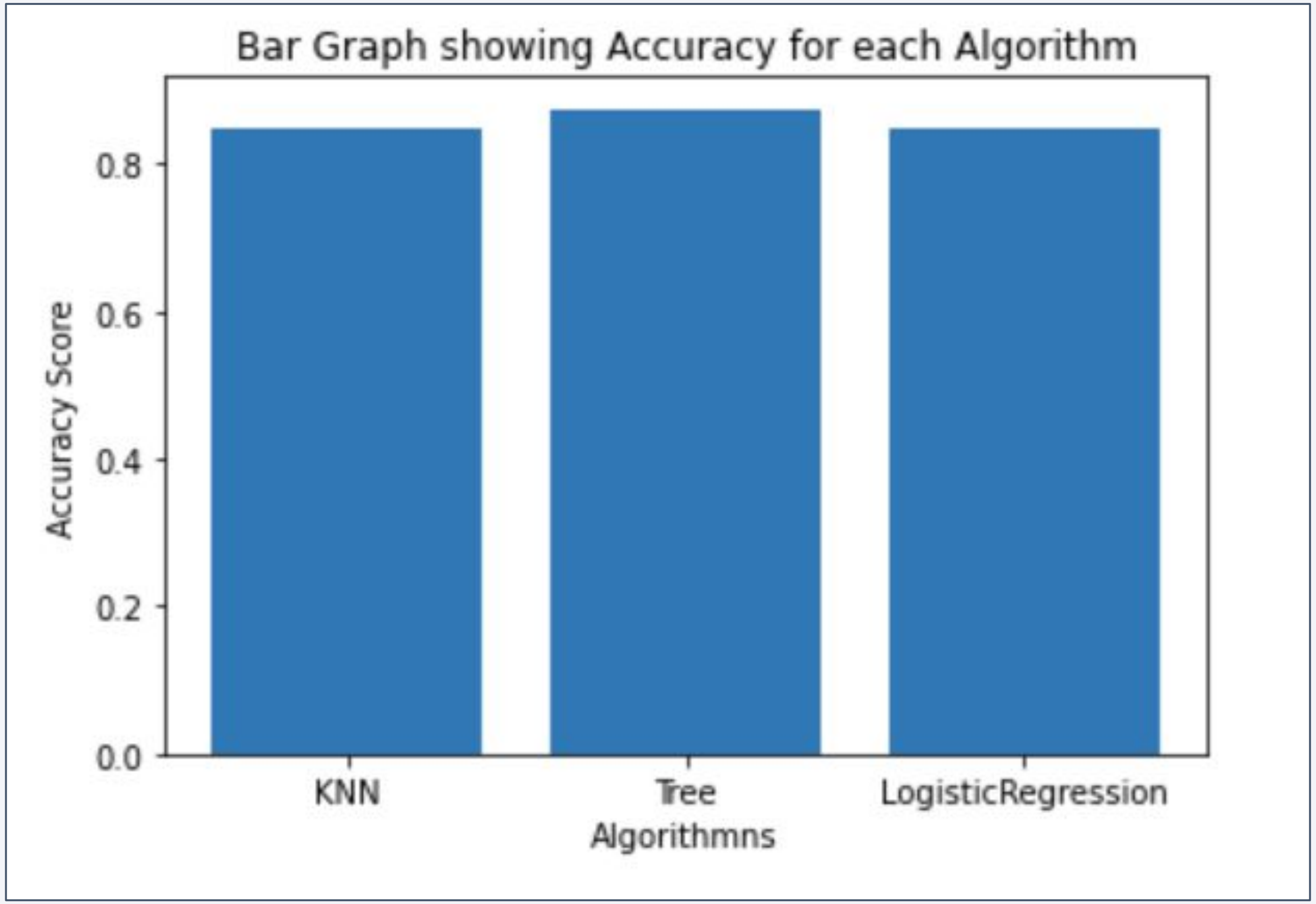
There does not appear to be a clear difference in success rates depending on payload capacity.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

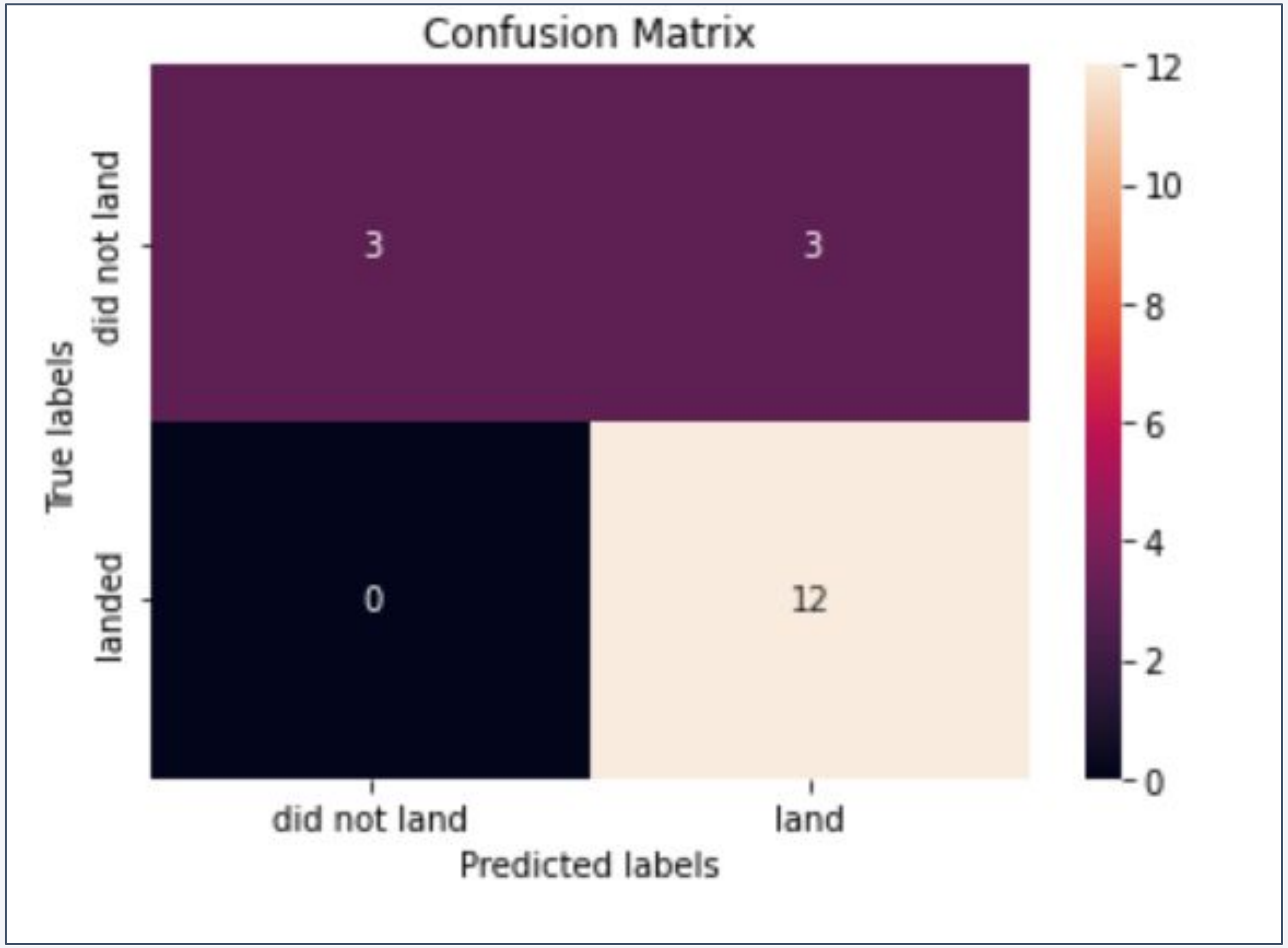


- KNN Accuracy Rate: **84.8%**
- Decision Tree Accuracy Rate: **87.5%**
- Logistic Regression Accuracy Rate: **83.3%**

Confusion Matrix of Decision Tree

- Of the 15 flights predicted to land, 12 landed successfully.
- Of the 3 flights predicted to fail, all 3 failed to land.

According to the sample, the decision tree correctly ascertains the flights that will not land BUT exaggerates the flights that will land. The decision tree therefore has a small bias toward predicting that flights will land.



Conclusions

- SpaceX, since 2010, has built a company able to launch a wide variety of payloads to a wide variety of orbit types. This business model will only improve with time as the technology and expertise improves to increase the likelihood of landing the rockets and therefore saving money on rocket production.
- In order to compete with SpaceX, a given competitor must either develop an expertise/specialty that serves a particular orbit type or payload capacity, OR must match SpaceX across a variety of metrics.

Thank you!