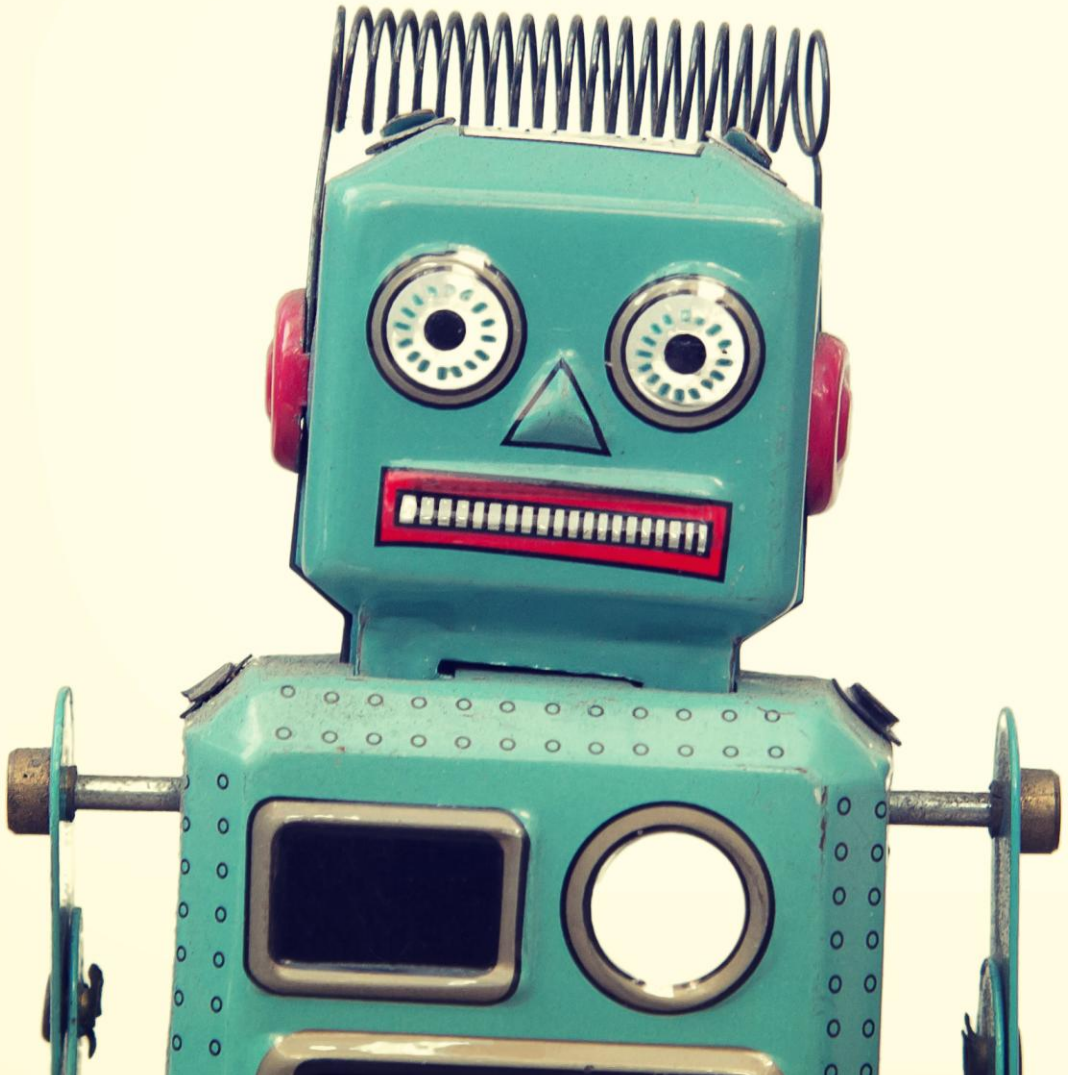


Winning Space Race with Data Science

Viktorija Trubaciute

2024-01-26

Outline

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion
6. Appendix

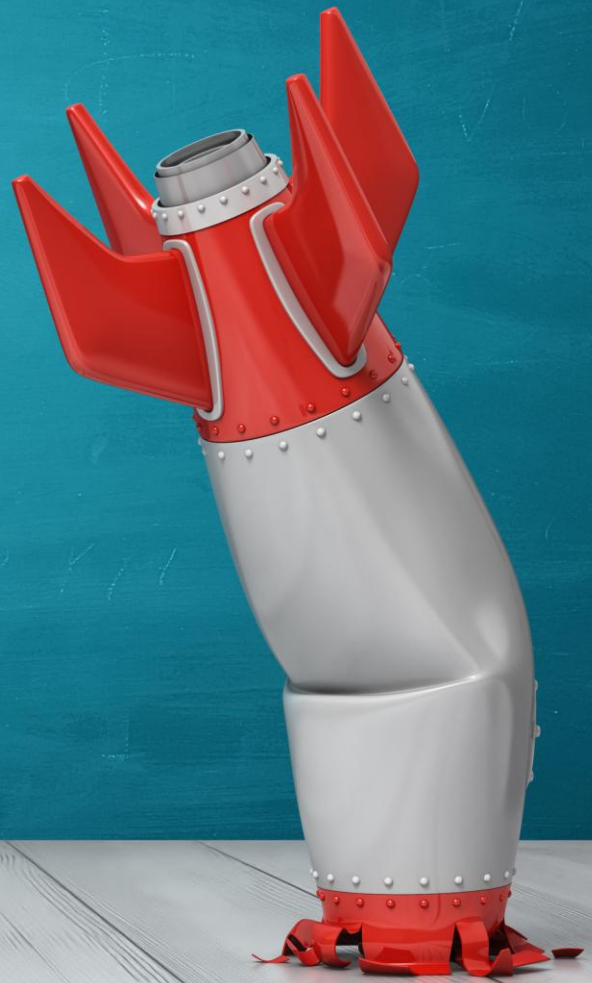
Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

Falcon 9 rocket launches cost SpaceX $62M dollars, while other providers claim costs of $165 and above. This is due to their reusable first stage, therefore this cost is heavily dependent on landing success rate.

We want to predict the success of successful landing and determine what criteria influences success rate, which would allow to create a successful program to compete with SpaceX.



Methodology

Section 1

Methodology

- Data collection:
  - Downloading data using SpaceX API
  - Scraping additional data from Wikipedia
- Data wrangling:
  - Replacing missing values
  - Calculating frequencies
  - Adding new columns
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models: Logistic regression, SVM, decision tree and KNN

1. Data Collection – SpaceX API

1. Downloaded data from SpaceX API
2. Multiple requests were needed. First, launch table, then additional information, using launch IDs
3. Filtered the data, keeping only Falcon 9 launches
4. This was done using information in “BoosterVersion” column
5. Replaced missing payload mass values with their mean
6. Final dataset has missing values only in column “LandingPad”

URL to data collection with API notebook on GitHub

2. Data Collection – Scraping

1. Scraped Falcon 9 Wiki page using BeautifulSoup
2. Relevant information was stored in a HTML table
3. Extracted column names from table headers
4. One column wasn't needed (Date and time), the rest were all relevant
5. Created a dataframe with relevant columns and filled it with scraped data
6. In this step, some parsing was necessary, for which helper functions were used

URL to web scraping notebook on GitHub

3. Data Wrangling

1. Checked missing values and data types (numerical or categorical) in the data
2. Calculated the number of launches for:
    1. Each site
    2. Each orbit
3. Calculated frequencies of all landing outcomes
4. Created a new column for landing outcome (successful or not, 1 or 0)
5. Saved the resulting dataset into a CSV file

URL to data wrangling notebook on GitHub

4. EDA with Data Visualization

- Most charts are scatterplots. They show relationships between:
  - Flight number and Payload Mass
  - Flight number and Launch Site
  - Payload mass and Launch Site
  - Flight number and Orbit
  - Payload mass and Orbit
- One bar chart shows the relationship between success rate and orbit type
- One line chart is used to show how success rate changed over the years

URL to data visualization notebook on GitHub

5. EDA with SQL

SQL is used to select:

- Unique launch sites in the space mission
- 5 records where launch sites begin with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first succesful landing outcome in ground pad was achieved
- Names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Booster versions which have carried the maximum payload mass
- Months, landing outcomes, booster versions and launch sites for failed drone ship landings in 2015
- Landing outcomes, ranked by their count, between 2010-06-04 and 2017-03-20

URL to SQL analysis notebook on GitHub

6. Interactive Map with Folium

1. Created and added objects to a Folium map:
    1. markers for launch sites
    2. labels of launch sites
2. This shows that launch sites are not far from the Equator line and in very close proximity to the coast.
3. Marked launch sites and successful/failed launches (using additional column with values 0 and 1 for launch outcome, then assigning red or green color to the marker accordingly).
4. Calculated distances between launch sites and its proximities, drew lines on the map and discovered that launch sites are in close proximity to railways, highways and coastline, but they keep a certain distance away from cities.

URL to interactive Folium map notebook on GitHub

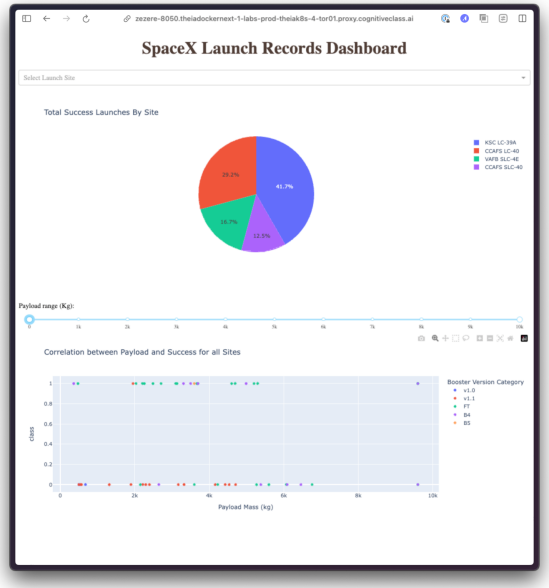
7. Dashboard (Plotly Dash)

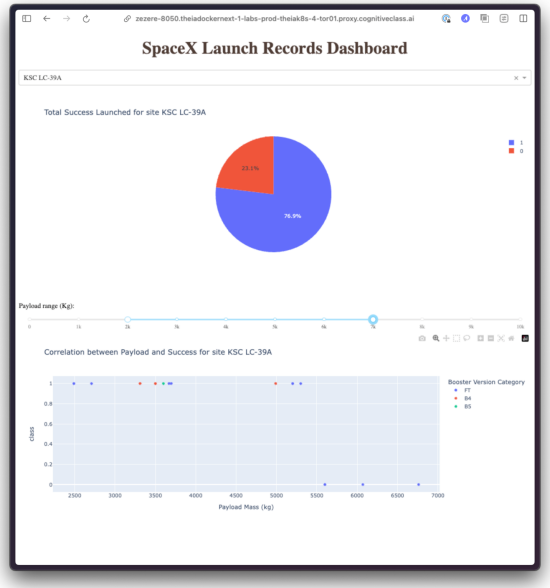
Dashboard consists of two main parts:

1. Dropdown and pie chart, where user chooses a launch site from the dropdown (or all of them) to display the success rate

2. Range slider and scatterplot, where user chooses the range of payload mass to display successful/failed launches, colored by booster version category

URL to Plytly Dash notebook on GitHub





8. Predictive Analysis (Classification)

1. Loaded and transformed the data
2. Assigned target variable Y to column “Class” (0 or 1 for landing outcome)
3. Normalized data using StandardScaler
4. Split data into training and testing sets (X_train, X_test, Y_train, Y_test)
5. Built ant applied different ML models:
    1. Logistic regression
    2. Support vector machine
    3. Decision tree classifier
    4. K nearest neighbors
6. Each ML model was fitted with best parameters using GridSearchCV, run on training dataset and afterwards accuracy was evaluated
7. All models had the same accuracy when evaluated against the test data (0.8333)

URL to predictive analysis notebook on GitHub

Results

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

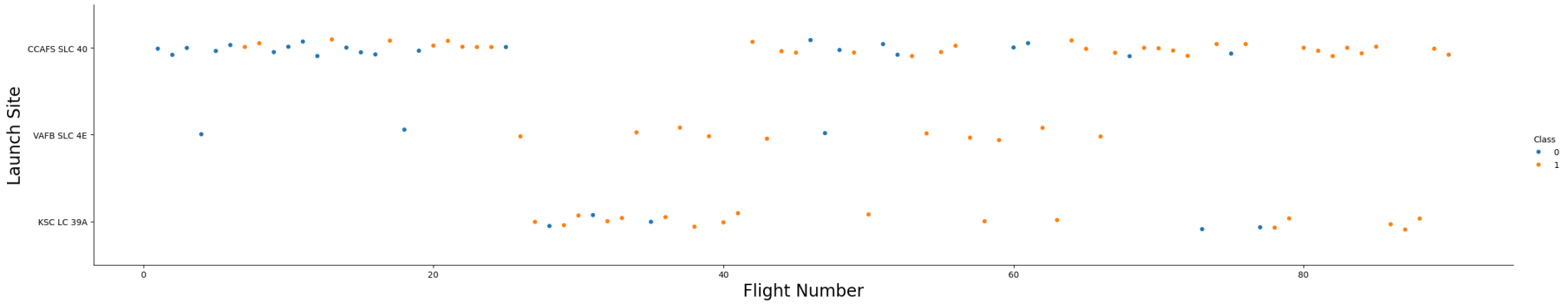
Insights from EDA

Section 2

Flight Number vs. Launch Site

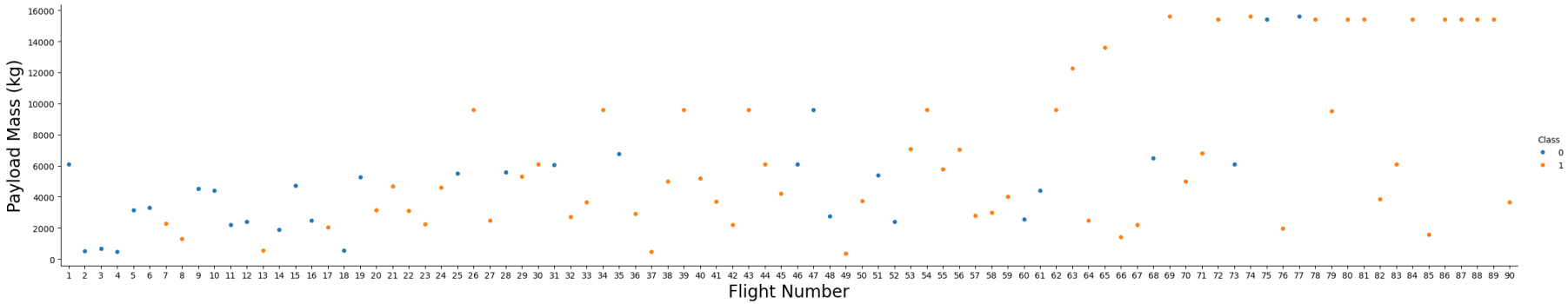
We can see that most initial launches were performed at CCAFS SLC 40. After a pause, this site remained the most “popular” one. Launch site VAFB SLC 4E was used least frequently.

Another observation is the increasing success rate. Note that there are no failures among the most recent launches.



Payload vs. Launch Site

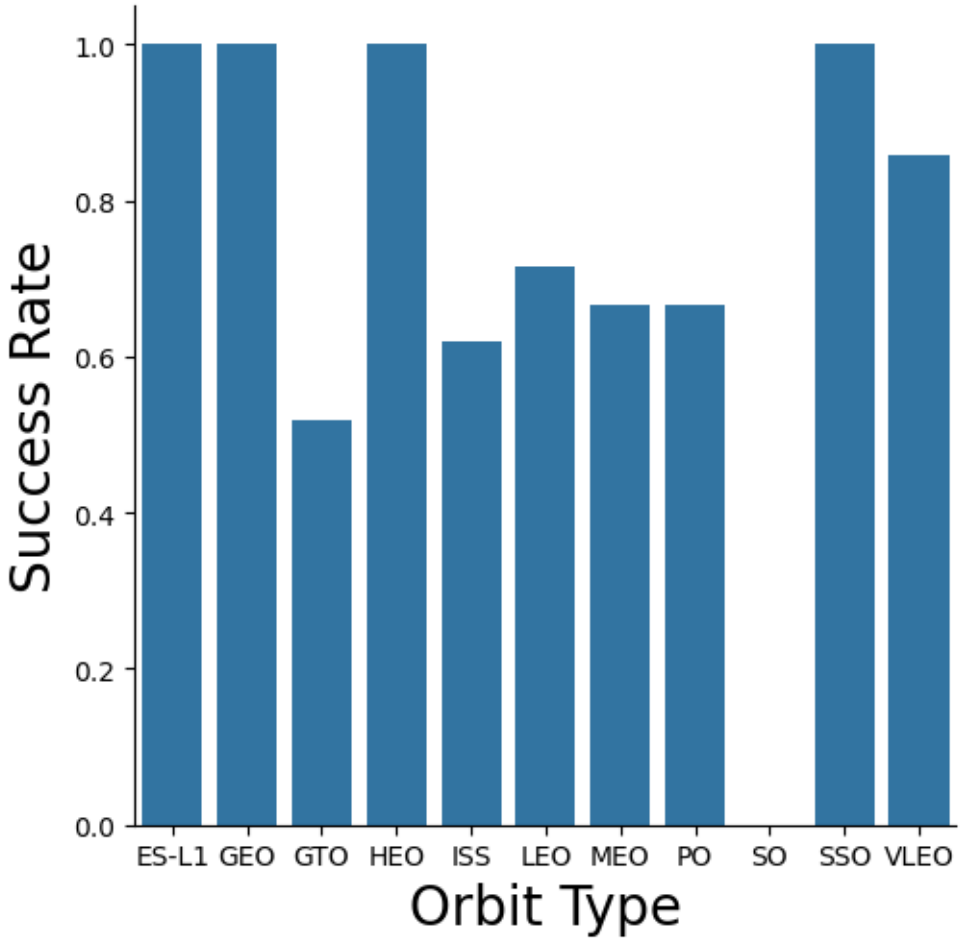
Payload mass has been consistently increasing, as well as the success rates. Only two launches, carrying the highest payload mass, have failed, the rest have succeeded.



Success Rate vs. Orbit Type

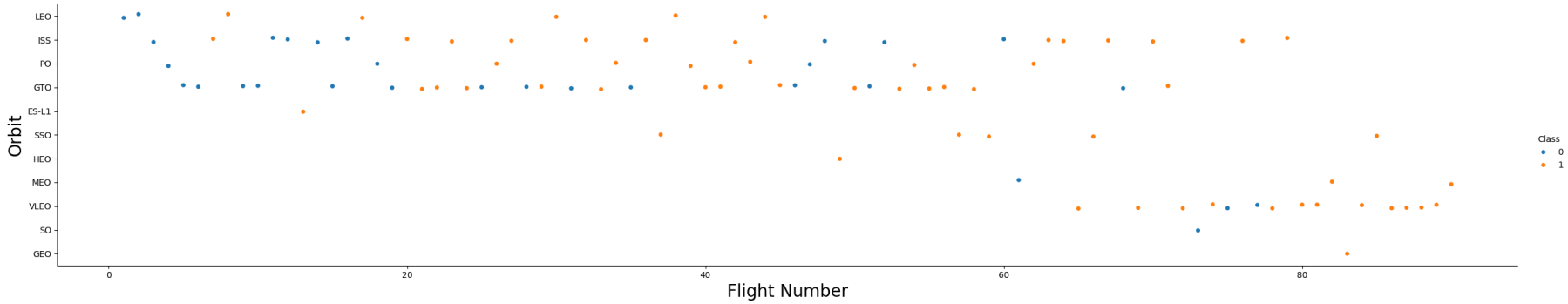
Four orbits have 100% success rate:

- ES-L1
- GEO
- HEO
- SSO



Flight Number vs. Orbit Type

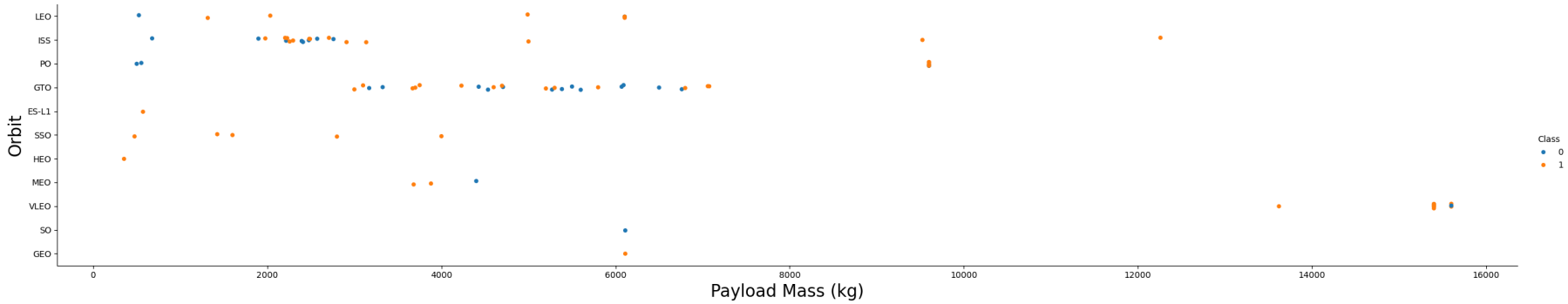
There is a relationship here as well. Earlier rockets were launched into LEO, ISS, PO and GTO orbits. Later – mostly VLEO. The highest orbit, GEO, had one launch and it was successful.



Payload vs. Orbit Type

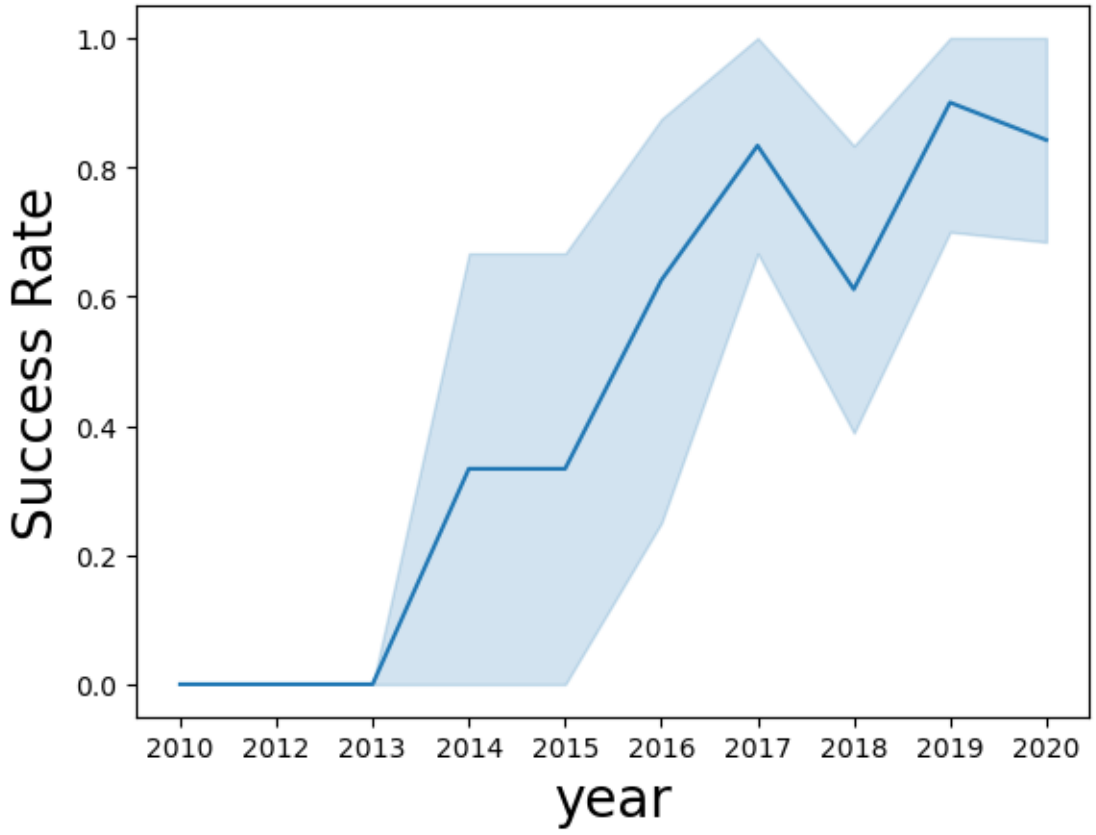
The interesting takeaway here is that heavy payloads better success rate is achieved when launching into LEO, ISS and PO.

Orbit GTO has mixed results.



Launch Success Yearly Trend

Success rate was 0% until year 2013. Since then, it was increasing sharply, with a mild hiccup in 2018.



All Launch Site Names

```
SELECT DISTINCT "Launch_Site" AS "Launch Sites" FROM SPACEXTABLE
```

```
%sql SELECT DISTINCT "Launch_Site" AS "Launch Sites" FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Launch Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begining with “CCA”

5 records where launch sites begin with `CCA`

```
SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%"
LIMIT 5
```

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total payload carried by boosters from NASA

```
SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass"
FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)"
```

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass" FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)"
 * sqlite:///my_data1.db
Done.
```

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1. Two ways: either booster versions named exactly “F9 v1.1”, or versions that have “F9 v1.1” as part of their name.

```
SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass"
FROM SPACEXTABLE WHERE "Booster_Version" LIKE "%F9 v1.1%"
```

```
SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass"
FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"
```

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass" FROM SPACEXTABLE WHERE "Booster_Version" LIKE "%F9 v1.1%"
 * sqlite:///my_data1.db
Done.
```

Average Payload Mass
2534.6666666666665

Some booster versions have "F9 v1.1" as part of their name. If we include those, we get a different number.

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass" FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"
 * sqlite:///my_data1.db
Done.
```

Average Payload Mass
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad

```
SELECT MIN("Date") AS "First Successful Landing" FROM
SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground
pad)"
```

```
%sql SELECT MIN("Date") AS "First Successful Landing" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)"
 * sqlite:///my_data1.db
Done.
```

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
SELECT "Booster_Version", "PAYLOAD_MASS__KG_" AS "Payload
Mass", "Landing_Outcome" AS "Landing Outcome" FROM
SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Success (drone
ship)%" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" AS "Payload Mass", "Landing_Outcome" AS "Landing Outcome" FROM SPACEXTABLE WHERE "Landing_Outcom
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Payload Mass	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

Total number of successful and failure mission outcomes

```
SELECT COUNT(*) AS "Total number of successful outcomes"
FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Success%"
```

```
SELECT COUNT(*) AS "Total number of failed outcomes" FROM
SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Fail%"
```

```
%sql SELECT COUNT(*) AS "Total number of successful outcomes" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Success%"
 * sqlite:///my_data1.db
Done.
```

Total number of successful outcomes
61

```
%sql SELECT COUNT(*) AS "Total number of failed outcomes" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "%Fail%"
 * sqlite:///my_data1.db
Done.
```

Total number of failed outcomes
10

Boosters that Carried Maximum Payload

Names of the boosters which have carried the maximum payload mass

```
SELECT DISTINCT "Booster_Version", "PAYLOAD_MASS__KG_" AS
"Payload Mass" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" =
(SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
```

```
%sql SELECT DISTINCT "Booster_Version", "PAYLOAD_MASS__KG_" AS "Payload Mass" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MAS
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Payload Mass
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	15600

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for year 2015

```
SELECT substr(Date, 6, 2) AS "Month", substr(Date,0,5) AS
"Year", "Landing_Outcome" AS "Landing Outcome",
"Booster_Version" AS "Booster Version", "Launch_Site" AS
"Launch Site" FROM SPACEXTABLE WHERE substr(Date,0,5) =
"2015" AND "Landing_Outcome" = "Failure (drone ship)"
```

```
%sql SELECT substr(Date, 6, 2) AS "Month", substr(Date,0,5) AS "Year", "Landing_Outcome" AS "Landing Outcome", "Booster_Version" AS "Booster Versio
 * sqlite:///my_data1.db
Done.
```

Month	Year	Landing Outcome	Booster Version	Launch Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	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

```
SELECT "Landing_Outcome" AS "Landing Outcome",
COUNT("Landing_Outcome") AS "Total Count" FROM SPACEXTABLE
WHERE "Date" BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY
"Landing_Outcome" ORDER BY COUNT("Landing_Outcome") DESC
```

```
%sql SELECT "Landing_Outcome" AS "Landing Outcome", COUNT("Landing_Outcome") AS "Total Count" FROM SPACEXTABLE WHERE "Date" BETWEEN "2010-06-04" AN
```

```
 * sqlite:///my_data1.db
Done.
```

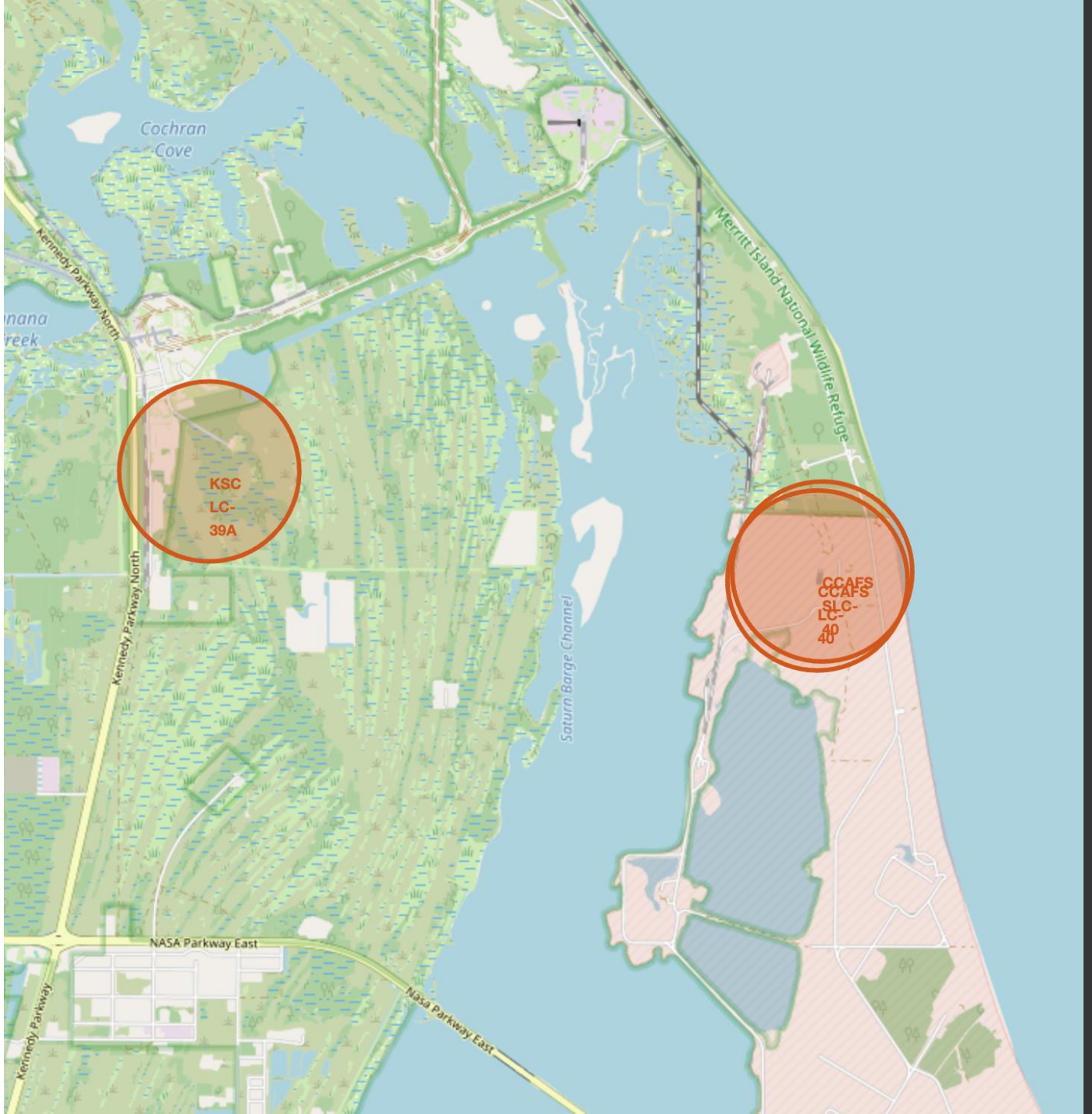
Landing Outcome	Total Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Launch Sites Proximities Analysis

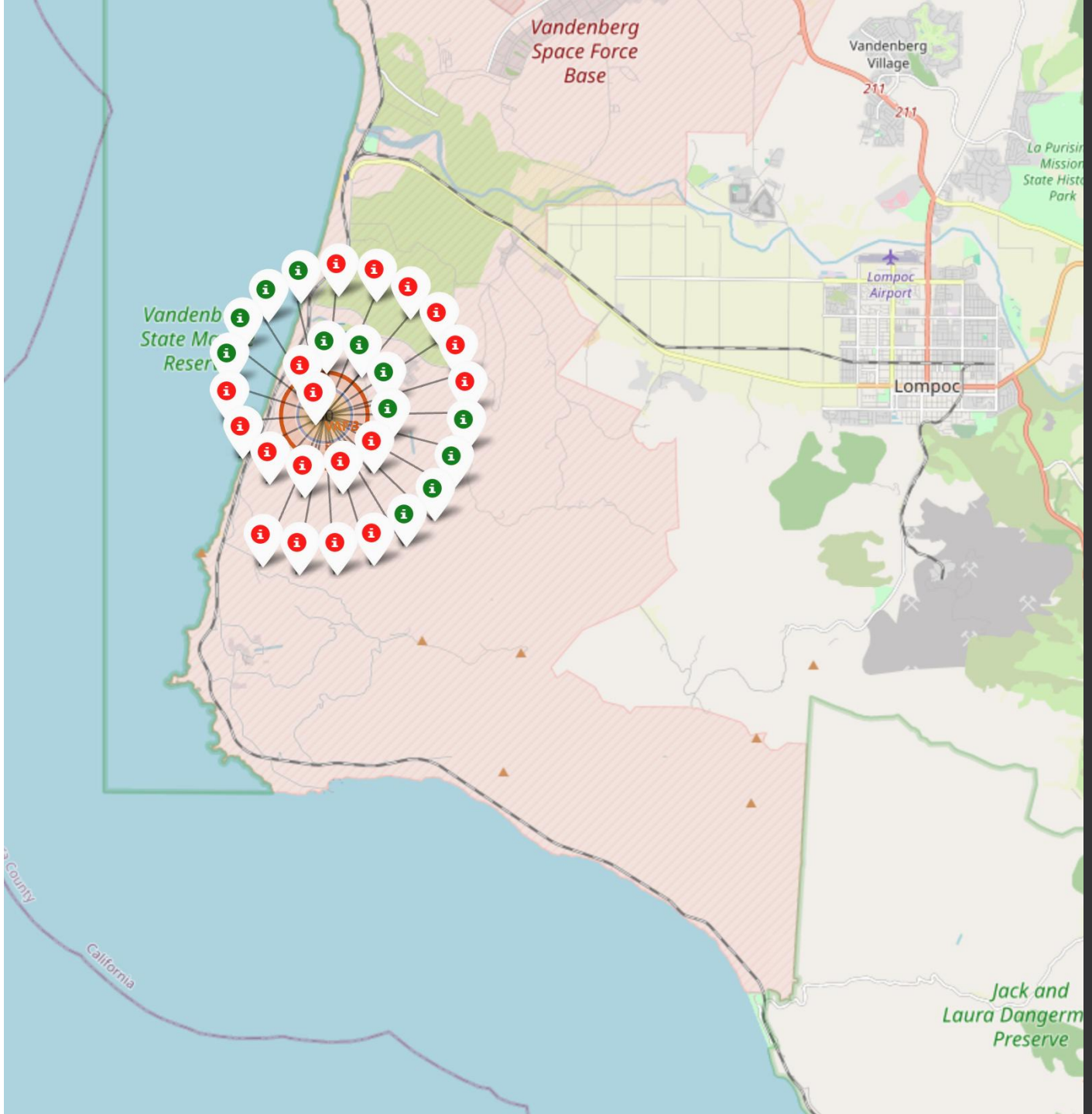
Section 3

All launch sites on the Folium map

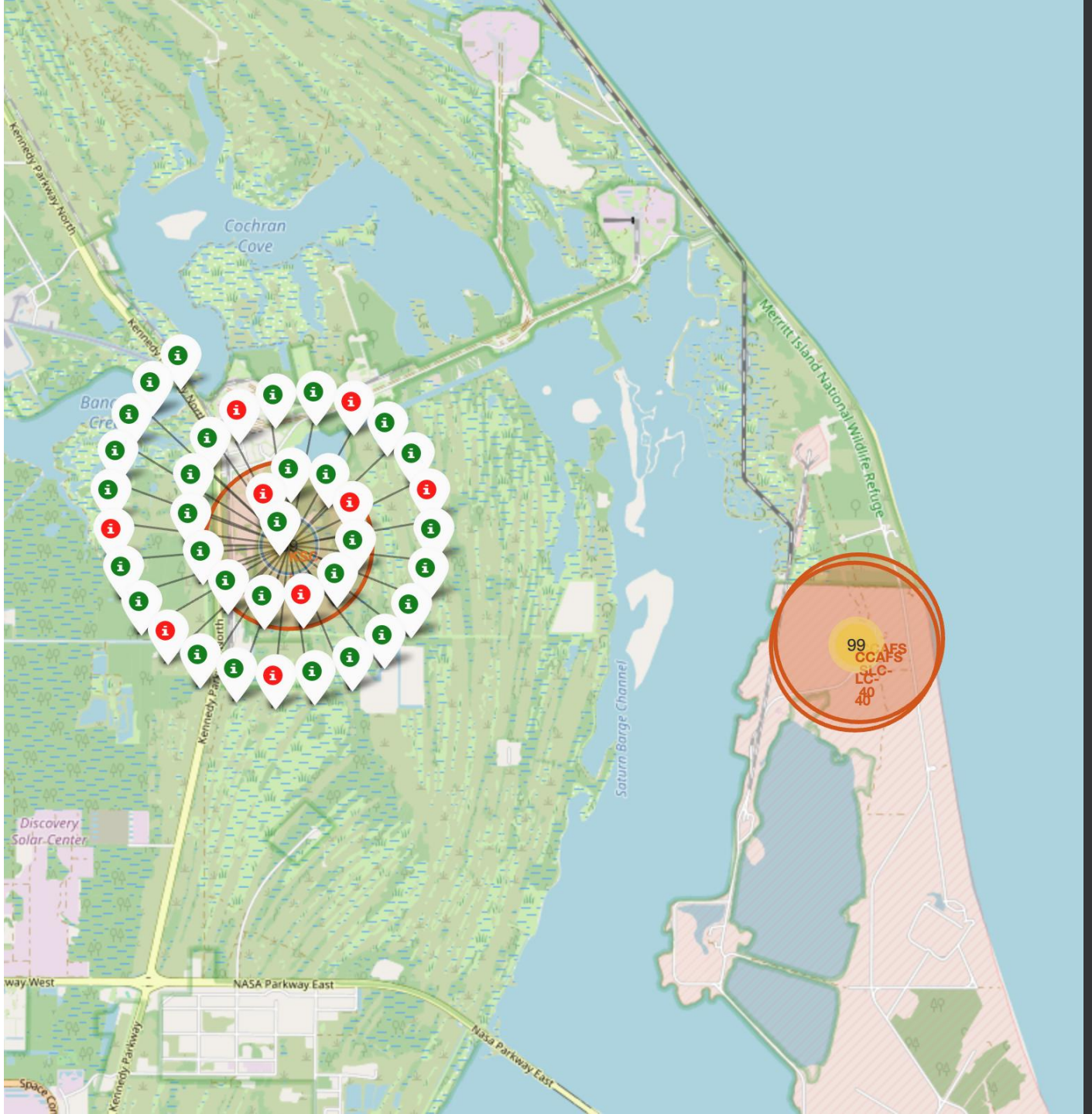
Florida launch sites from closer up



California launch site successes (green) and failures (red)

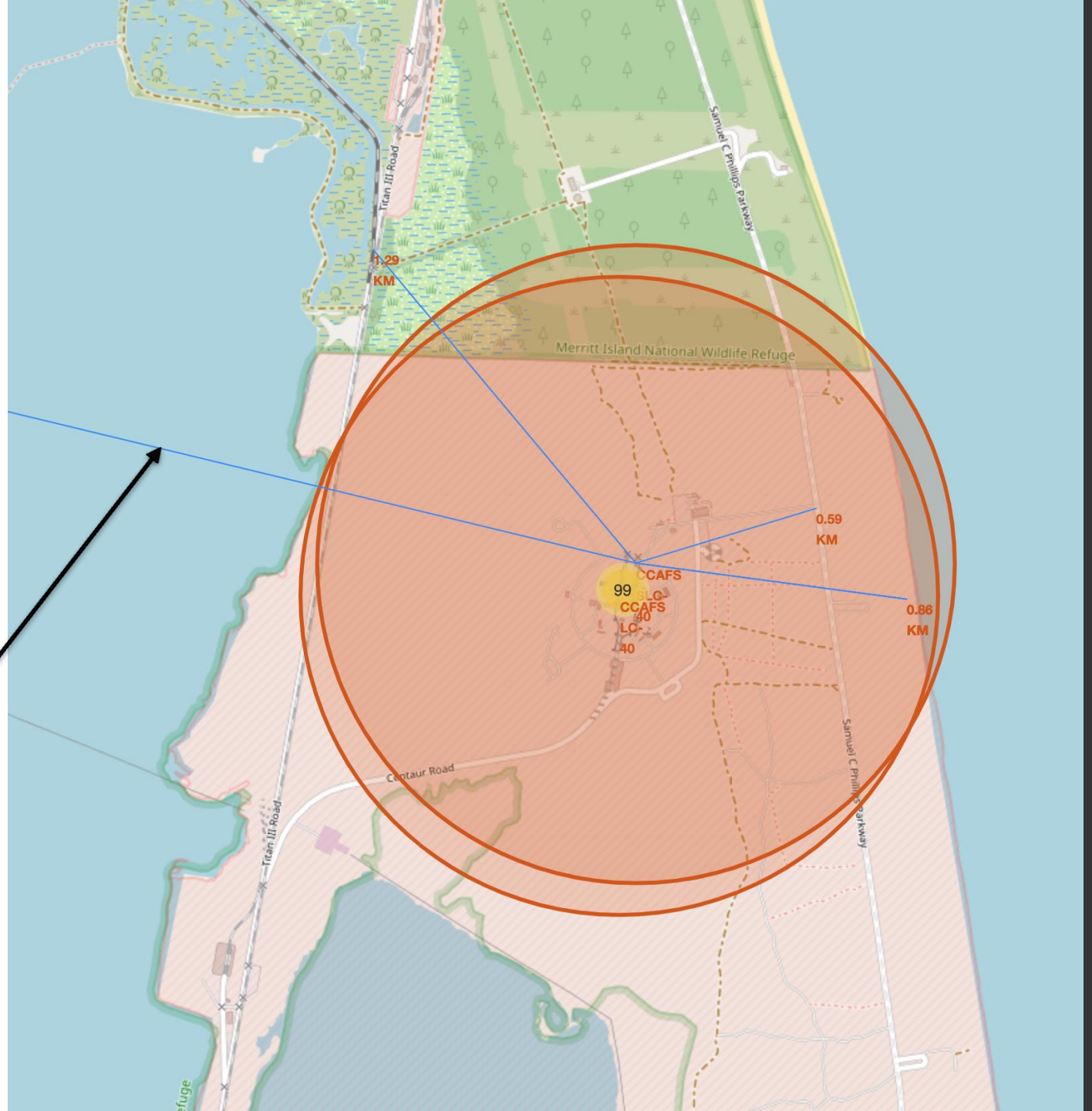


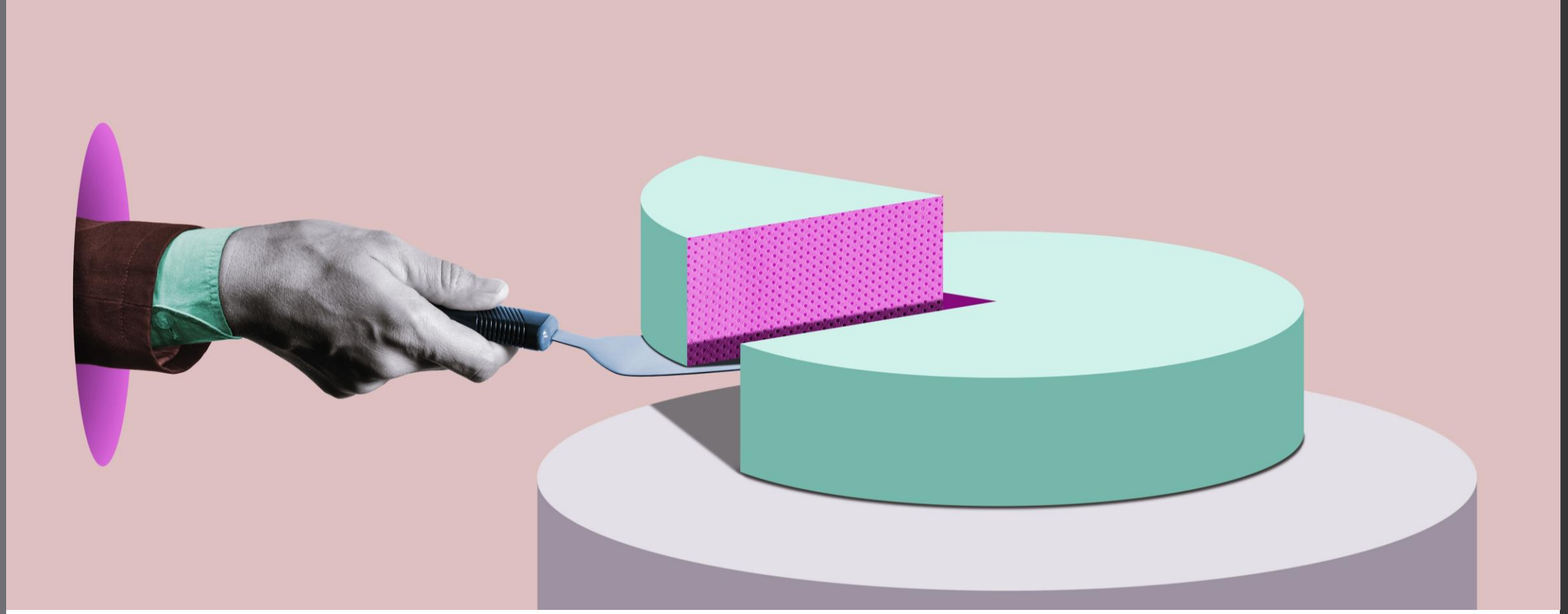
Same way in Florida, each site has labels for successes and failures



Proximity of a launch site to other objects: railways, highways, coast, cities

Line to the nearest city Titusville





Dashboard with Plotly Dash

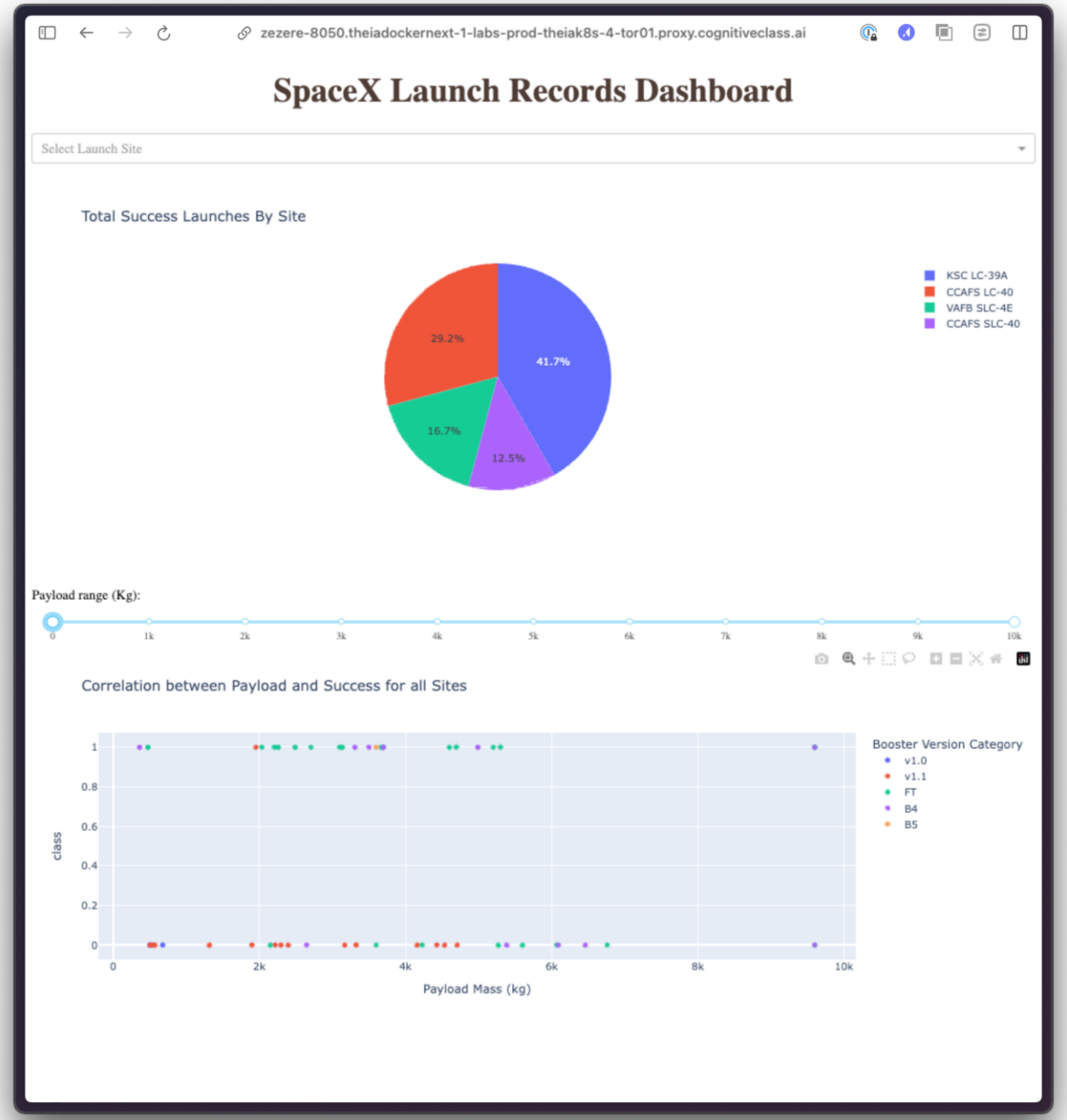
Section 4

Interactive dashboard

Interactive dashboard consists of two parts:

1. Pie chart that shows successful launches by site. We can choose all or a particular site from the dropdown.

2. Scatterplot that shows correlation between payload and Success rate. We can “zoom in” by using a slider to limit the range of payload mass we are interested in.



Interactive dashboard

This is how it looks if we choose site KSC LC-39A:

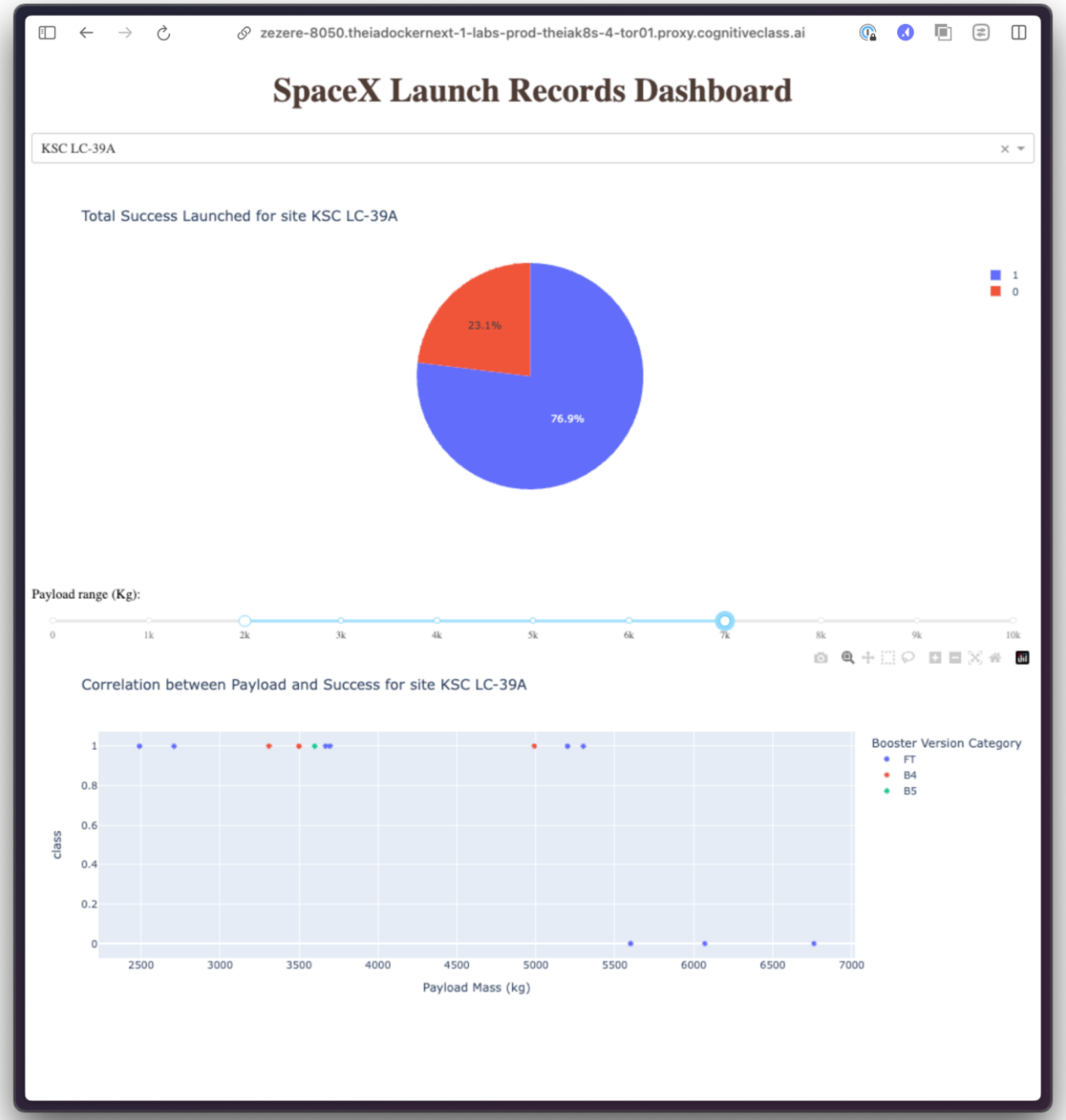
- 76.9 % successful landings

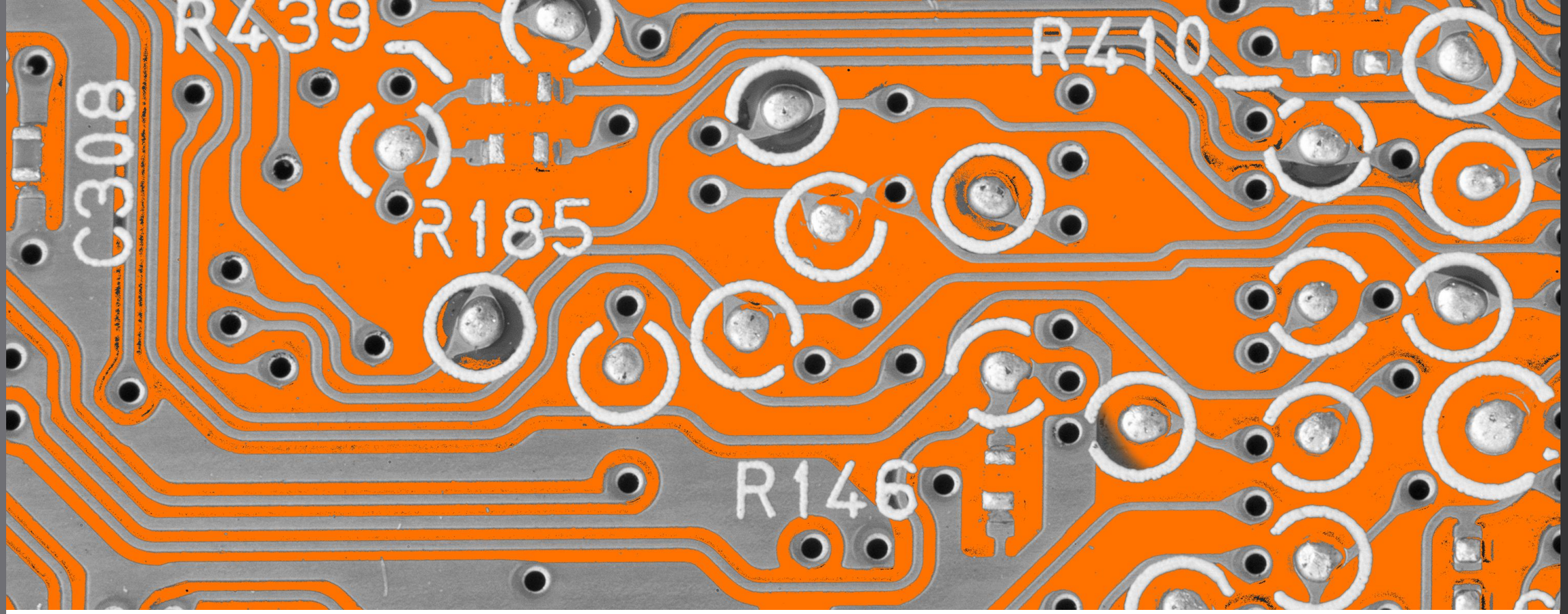
- 23.1 % unsuccessful

Payload range between 2000 and 7000 kg:

- 3 failed landings

- 10 successful landings





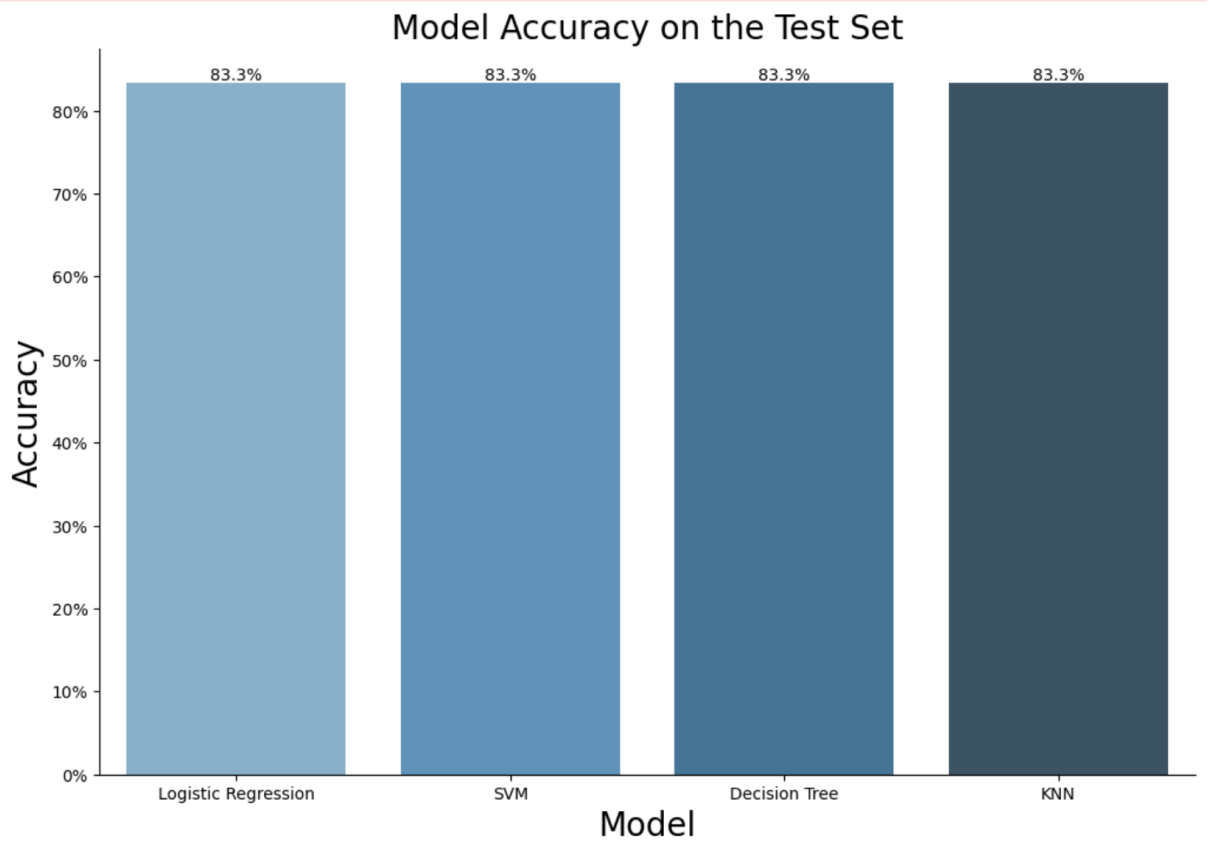
Predictive Analysis (Classification)

Section 5

Classification Accuracy

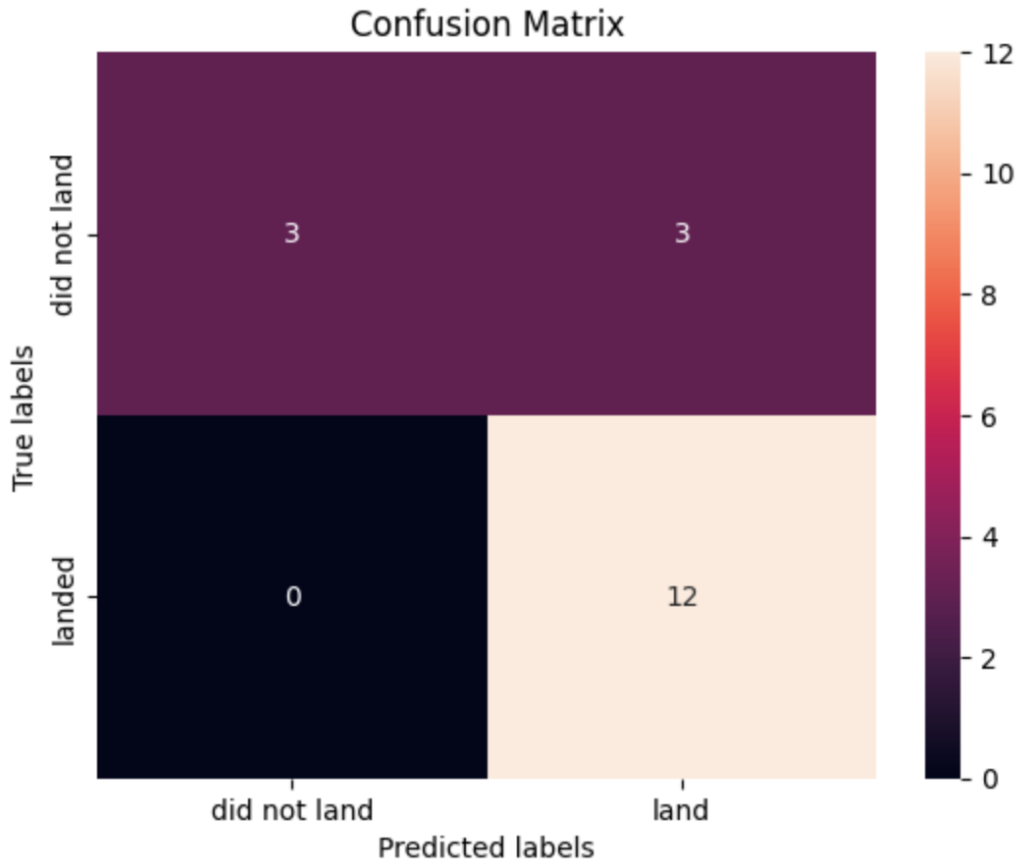
All models showed the same accuracy when tested on the test dataset.

However, Decision tree model broke out in errors due to faulty data, so I would assume that with good data it would perform differently.



Confusion Matrix

Confusion matrix for all models was the same.



Conclusion

- Here are discovered ways to increase the success of a launch:
  - Less payload
  - Use launch site KSC LC-39A
  - Launch to orbits GEO, HEO, SOO and ES-L1
- Time is on our side: there is a steady increase in the success rate of launches of Falcon 9.
- Our predictive capability is at a good level of 83.3 % with any model.

Thank You for Your Time!