IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stacey Adams
2021-09-05

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Data was collected via REST API and web scraping.
- Exploratory Data Analysis was performed using data visualization tools and SQL queries.
- A dashboard was created to interactively examine aspects of the data.
- A machine learning algorithm was selected to predict future launch states.
- A decision tree classification model can be used to predict whether a launch will be successful or not.
- As the payload mass goes up, the likelyhood of a successful launch decreases.

INTRODUCTION

Rocket launches are expensive, and by reusing components, we can bring the price of a launch down

We want to predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

METHODOLOGY

Executive Summary

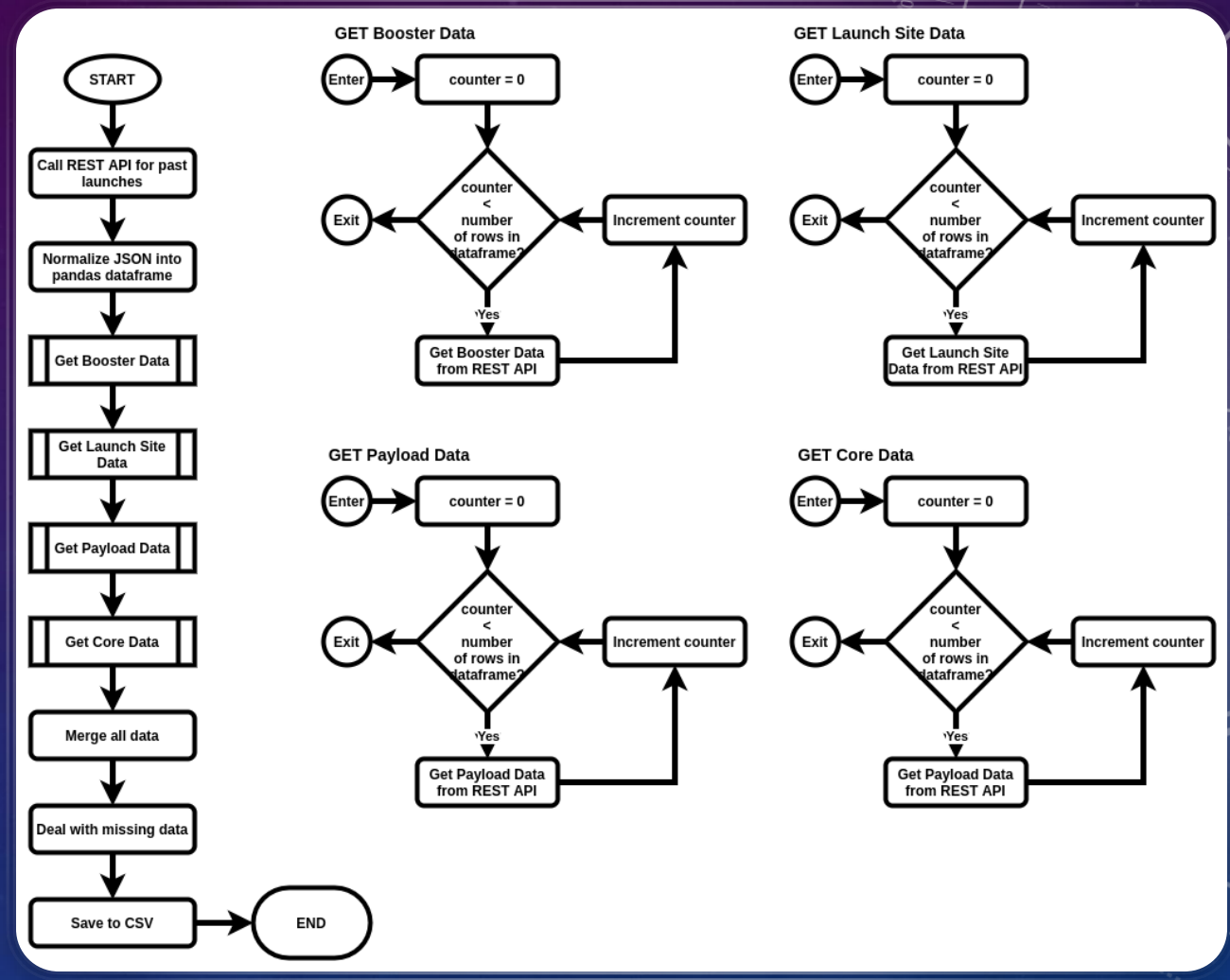
- Data collection methodology:
  - Data was collected via REST API calls and web scraping into CSV files
- Perform data wrangling
  - Pandas dataframes were used to wrangle the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was split into train/test sets
  - Various classification models were optimized using a cross-validated grid search over their parameter grids
  - Each model was evaluated with optimized parameters

DATA COLLECTION

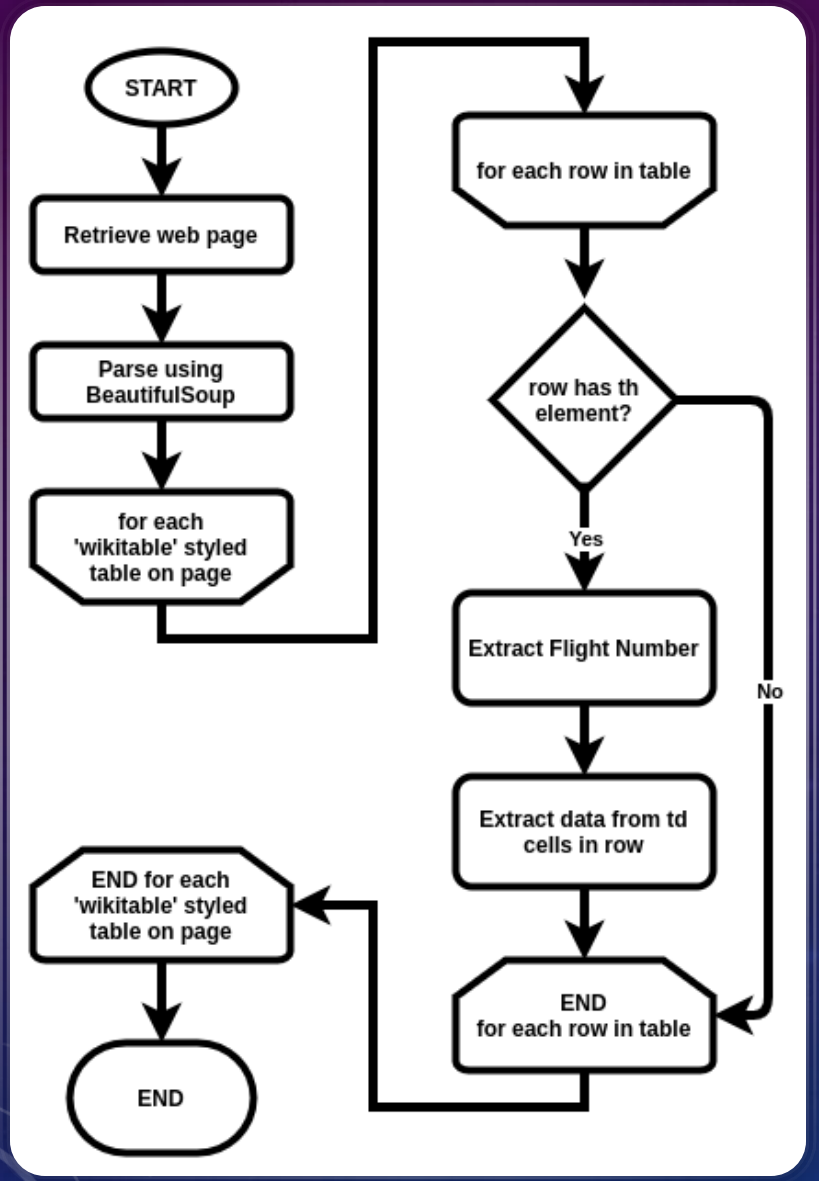
- Data sets were collected via REST API and web scraping
- REST API endpoints used were
  - https://api.spacexdata.com/v4/rockets/{rocket_id}
  - https://api.spacexdata.com/v4/launchpads/{launchpad_id}
  - https://api.spacexdata.com/v4/payloads/{payload_id}
  - https://api.spacexdata.com/v4/cores/{core_id}
  - https://api.spacexdata.com/v4/launches/past
- Web Scraping was performed on the Wikipedia page:
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

DATA COLLECTION – SPACEX API

Data was collected via REST API, and normalized using pandas dataframes



Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/SpaceX%20Data%20Collection.ipynb



DATA COLLECTION - SCRAPING

Data was scraped from wikipedia's page of launches using BeautifulSoup library, extracting table cells into a pandas dataframe. The dataframe was then saved to a CSV file for further use.

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Webscraping%20Data%20Collection.ipynb

DATA WRANGLING

Pandas dataframes were used to wrangle the data. Landing outcomes were converted to a 'class' field, with class=1 indicating a good outcome, and class=0 indicating a bad outcome.

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Data%20Wrangling.ipynb

EDA WITH DATA VISUALIZATION

Data examined

- Payload Mass vs Flight Number
- Launch Site vs Flight Number
- Launch Site vs Payload Mass
- Success Rate vs Orbit Type
- Orbit Type vs Flight Number
- Orbit Type vs Payload Mass
- Success Rate vs Year of Launch

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Exploratory%20Analysis%20with%20Data%20Visualization.ipynb

EDA WITH SQL

SQL Queries performed

- select unique LAUNCH_SITE from SPACEXTBL
- select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
- select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
- select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
- select min(date) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
- select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
- select mission_outcome, count(*) from SPACEXTBL group by mission_outcome
- select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
- select monthname(date), landing__outcome, booster_version, launch_site from SPACEXTBL where year(date) = 2015
- select landing__outcome, count(landing__outcome) as landing_outcome_count from SPACEXTBL where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome) desc

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Exploratory%20Analysis%20with%20SQL.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- On the launch site map, circles and markers were added to indicate locations of launch sites.
- On the map showing success/failed launches, marker clusters were added to each launch location. Each marker cluster contained a set of pins for each launch at that site. The pins were colored green for successful launches, and red for failed launches.
- On the proximities map, lines were added between launch sites and coastlines, highways, railroads, and cities, and were each tagged with distances from the launch site.

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

BUILD A DASHBOARD WITH PLOTLY DASH

- On the dashboard, a dropdown to select a specific launch site, or "all sites", was added to allow the user to examine data per launch site, or cumulative data.
- A pie chart was added to examine the percent of successful launches per launch site. If "all sites" was chosen, the pie chart displayed counts of successful launches per site.
- A scatterplot was added to examine the relationship between payload mass and success/failure of the launch. A slider allowed the user to change the payload mass range displayed in the scatterplot.

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/spacex_dash_app.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

- All standard classification models were built and evaluated, using a 20% split between test and train data in the data set.
- A cross-value grid search was performed on each model's parameter space to determine the best parameters for the model.
- The best parameter sets were used to generate the models.
- Models were evaluated based on their mean accuracy score.
- All models had the same mean accuracy score, so the model with the best confusion matrix (fewest false positives + false negatives) was chosen.

Reference: https://github.com/staceybellerose/data_science_capstone/blob/master/Machine%20Learning%20Prediction.ipynb

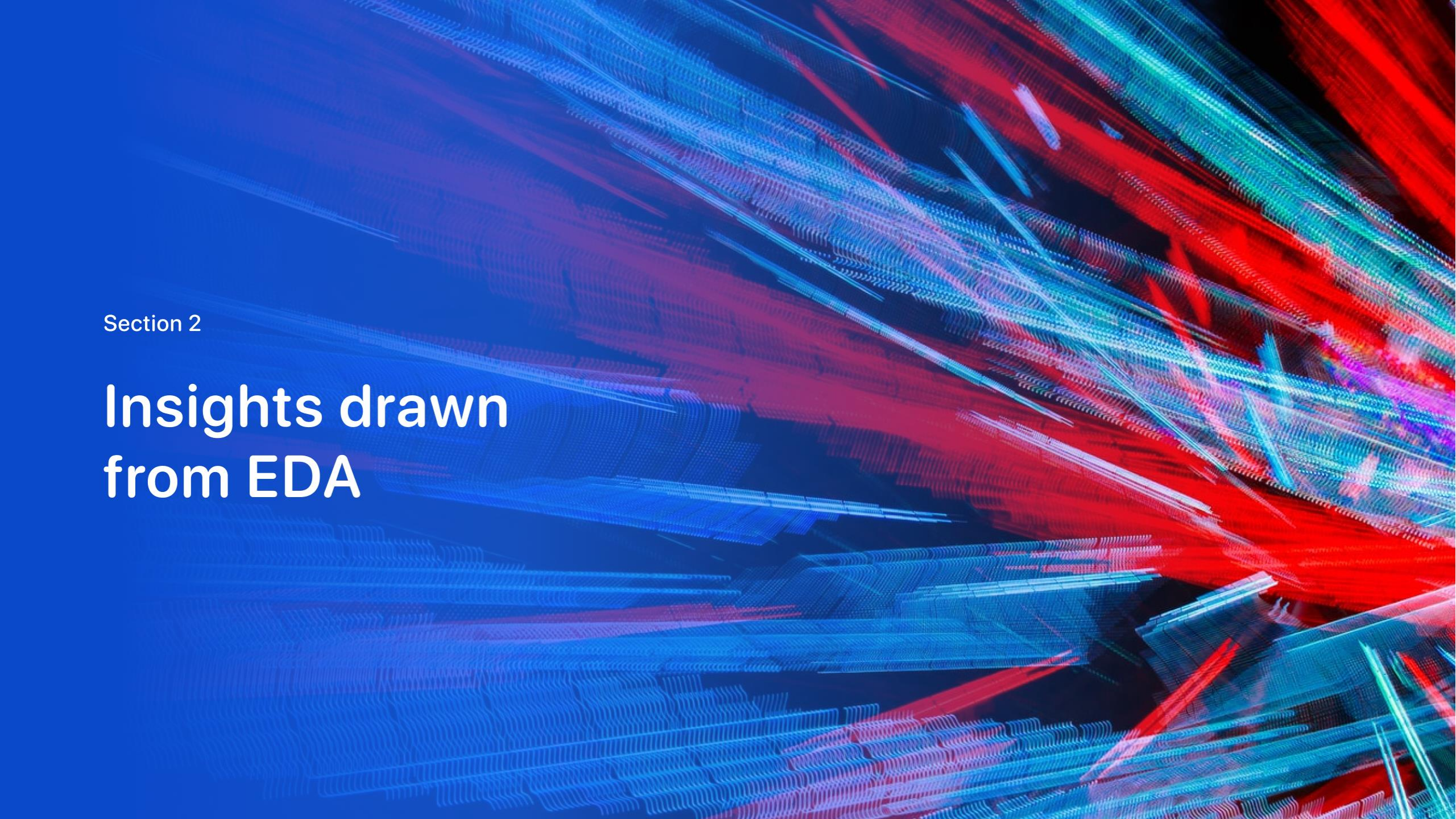
RESULTS

Successful landings of boosters increases over time, as the company gains more experience.

Launches should be done near the coast, with sites having easy access to highways and railroads, but at some distance from major cities.

As the payload mass goes up, the likelyhood of a successful landing decreases.

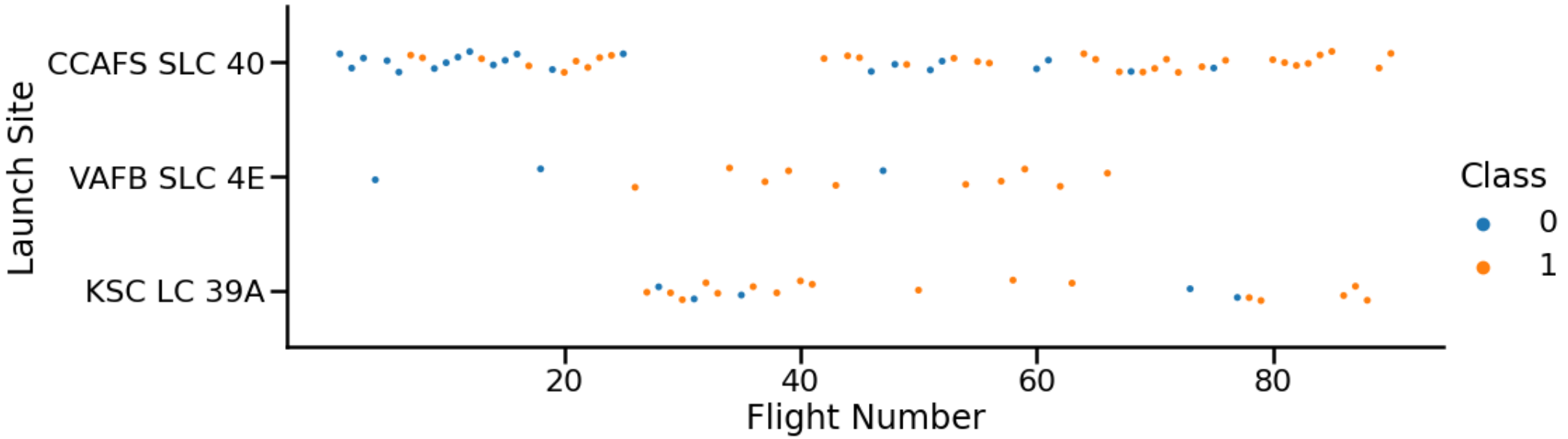
A decision tree classification model can be used to predict whether a launch will be successful or not.



Section 2

Insights drawn from EDA

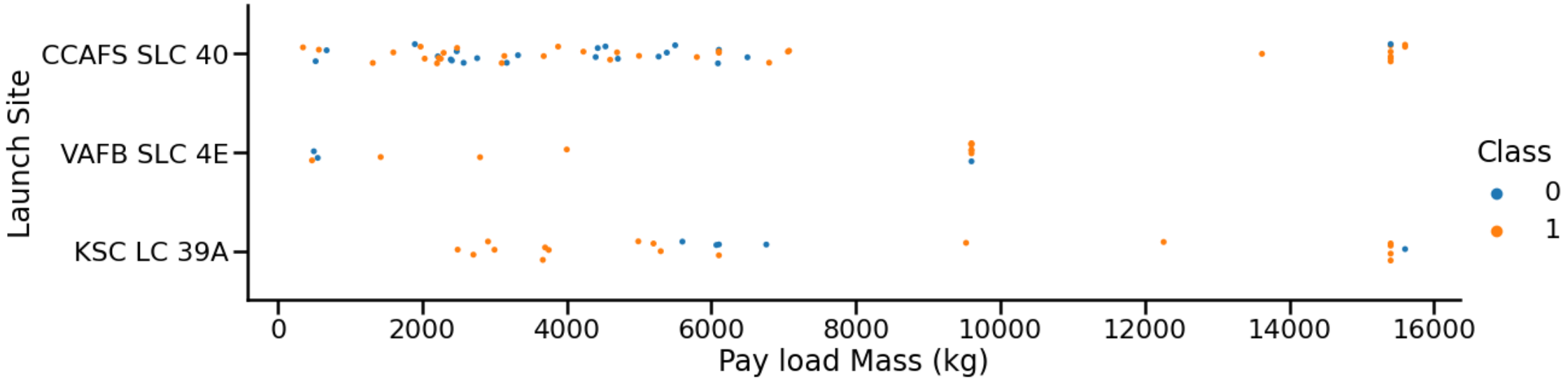
FLIGHT NUMBER VS. LAUNCH SITE



We see that CCAFS SLC 40 launch site was used almost exclusively in early flights, and that VAFB SLC 4E was only used sporadically.

KSC LC 39A was only used after flight 25, and use of this launch site tapered off after flight 40.

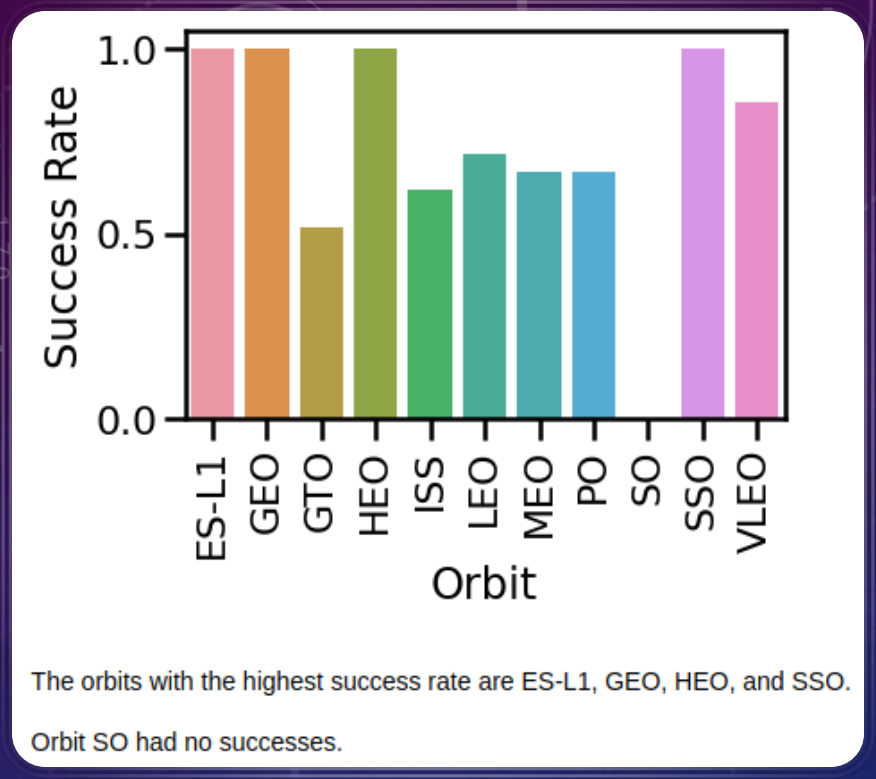
PAYLOAD VS. LAUNCH SITE



CCAFS SLC 40 was use most often, and had no apparent correlation between launch success and payload mass.

VAFB SLC 4E was used the least often, with the smallest payloads.

KSC LC 39A was had the most success with its heaviest payloads.

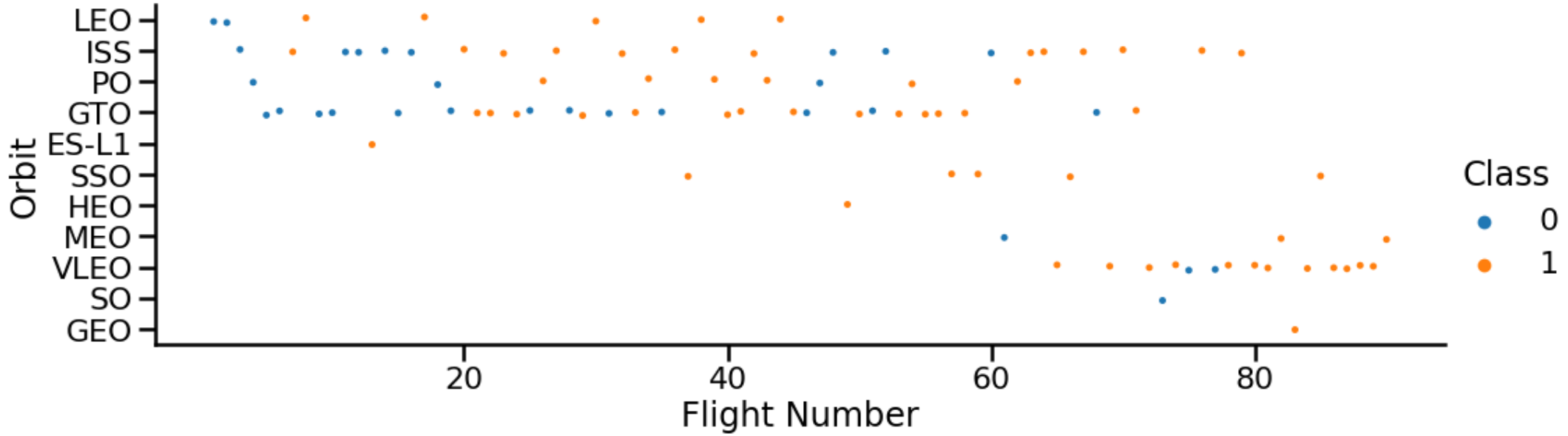


The orbits with the highest success rate are ES-L1, GEO, HEO, and SSO.

Orbit SO had no successes.

SUCCESS RATE VS. ORBIT TYPE

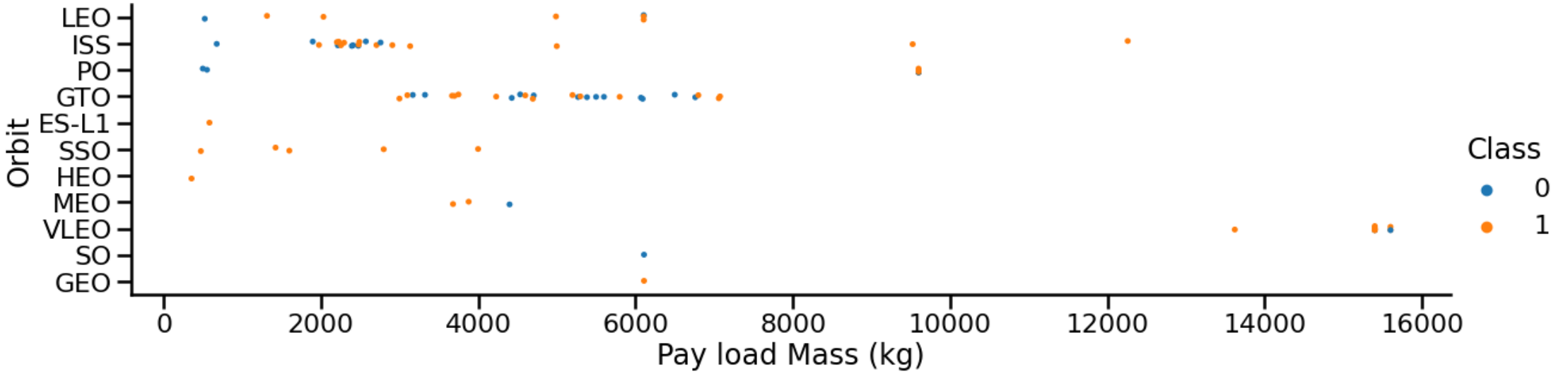
FLIGHT NUMBER VS. ORBIT TYPE



You should see that in the LEO orbit the Success appears related to the number of flights.

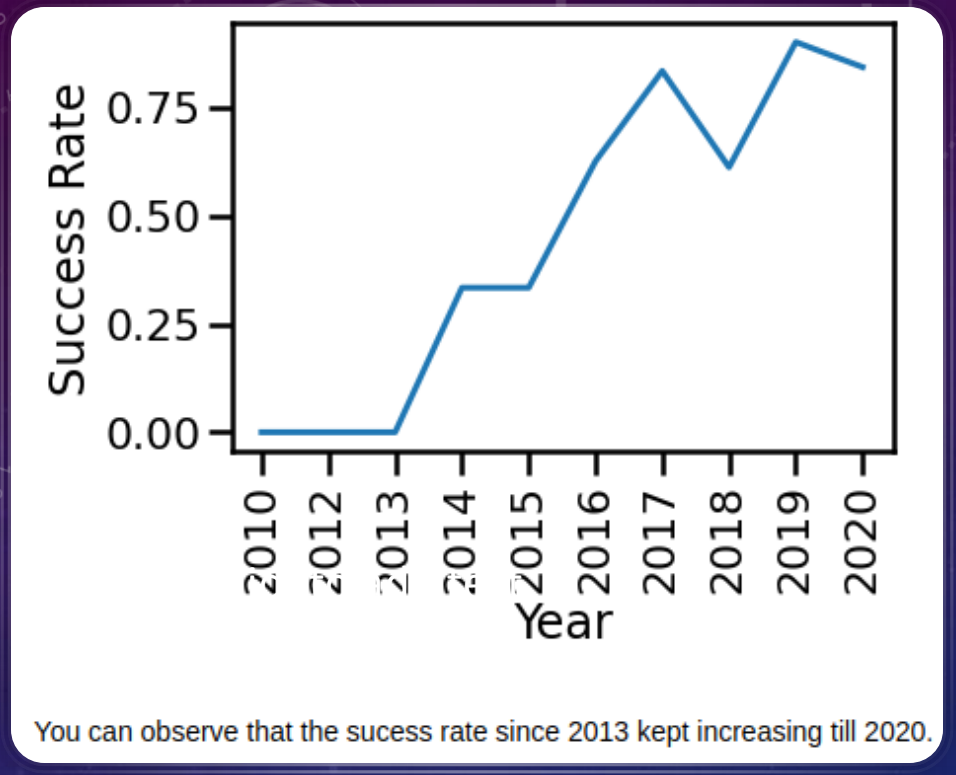
On the other hand, there seems to be no relationship between flight number when in GTO orbit.

PAYLOAD VS. ORBIT TYPE



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND



You can observe that the sucess rate since 2013 kept increasing till 2020.

ALL LAUNCH SITE NAMES

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

After importing flight data into SQL table, query was run to select the unique launch site names.

Date	Time	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

A SQL query was run to retrieve the first 5 records where the launch site name started with "CCA".

LAUNCH SITE NAMES BEGIN WITH 'CCA'

TOTAL PAYLOAD MASS FOR NASA

Total payload carried by boosters from NASA is **45,596 kg**

A SQL query was run to sum the payload mass for all launches for the customer "NASA (CRS)".

AVERAGE PAYLOAD MASS BY F9 V1.1

Average payload mass carried by booster version F9 v1.1 is **2928.4 kg**

A SQL Query was run to average the payload mass for all launches with a booster version of "F9 v1.1".

FIRST SUCCESSFUL GROUND LANDING DATE

First successful landing outcome on ground pad was on **2015-12-22**

A SQL query was run to select the minimum date for a launch where the landing outcome was "Success (ground pad)".

SUCCESSFUL DRONE SHIP LANDING BOOSTERS WITH PAYLOAD BETWEEN 4000 AND 6000

- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

A query was run to select the booster version from launches where the landing outcome was "Success (drone ship)" and the payload mass was between 4000 and 6000 kg.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

A query was run to count the number of mission outcomes.

Mission Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

BOOSTERS CARRYING MAXIMUM PAYLOAD

Booster Version
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

A query was run to select the booster version from all launches where the launch payload was the maximum use payload.

2015 LAUNCH RECORDS FOR DRONE SHIP LANDINGS

A query was run to show the month, landing outcome, booster version, and launch site for all launches in 2015 where the landing outcome mentioned "drone ship".

Month	Landing Outcome	Booster Version	Launch Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

RANKED LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

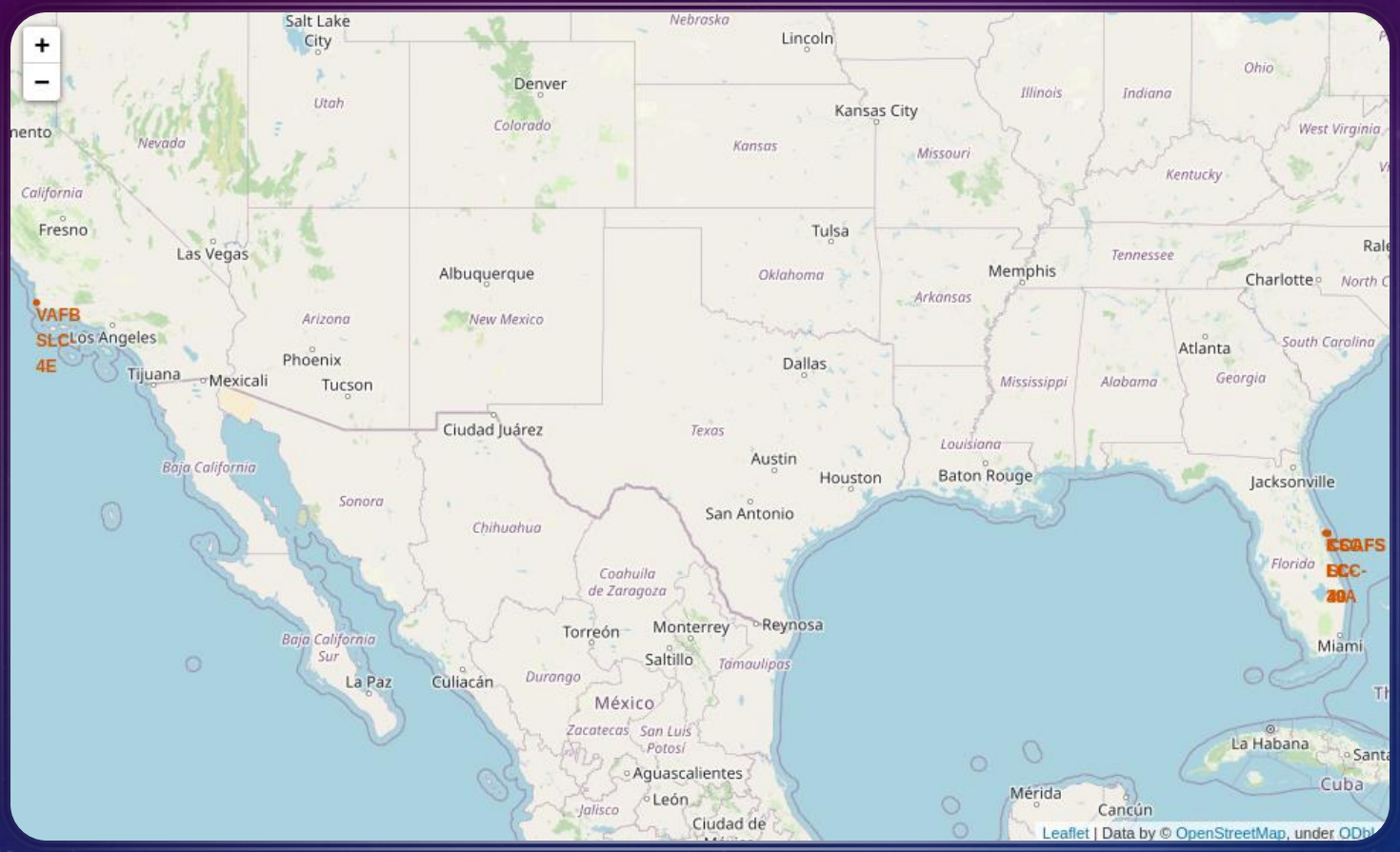
Landing Outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

A query was run to count all the landing outcome values in the requested date range, sorted by descending count.

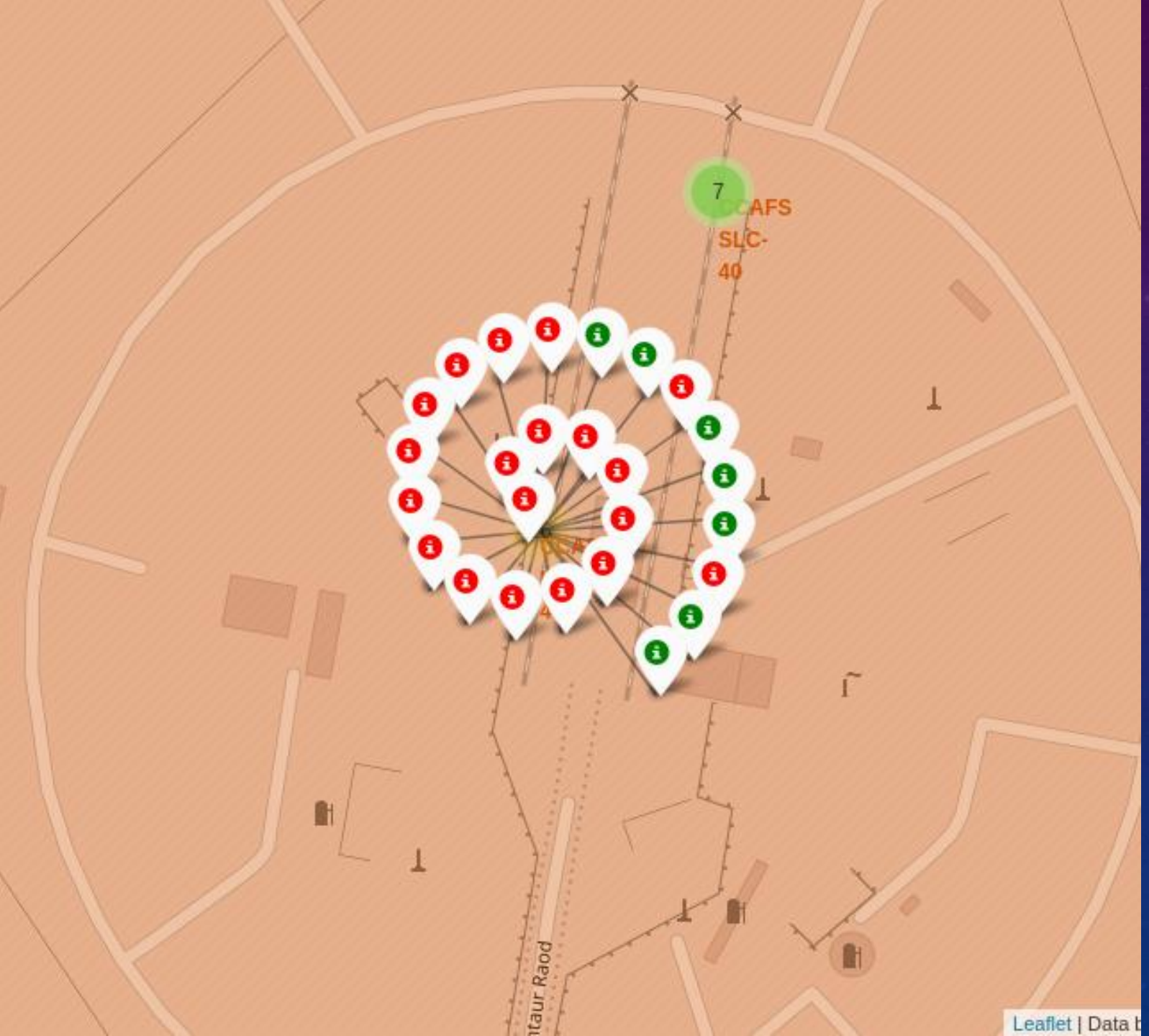
Section 4

Launch Sites Proximities Analysis

LAUNCH SITE LOCATIONS



Launch sites areshown in orange. They are located near the coast, in California and Florida. This is for safety reasons, to allow a failing rocket to be ditched into the ocean and not affect populations on land.

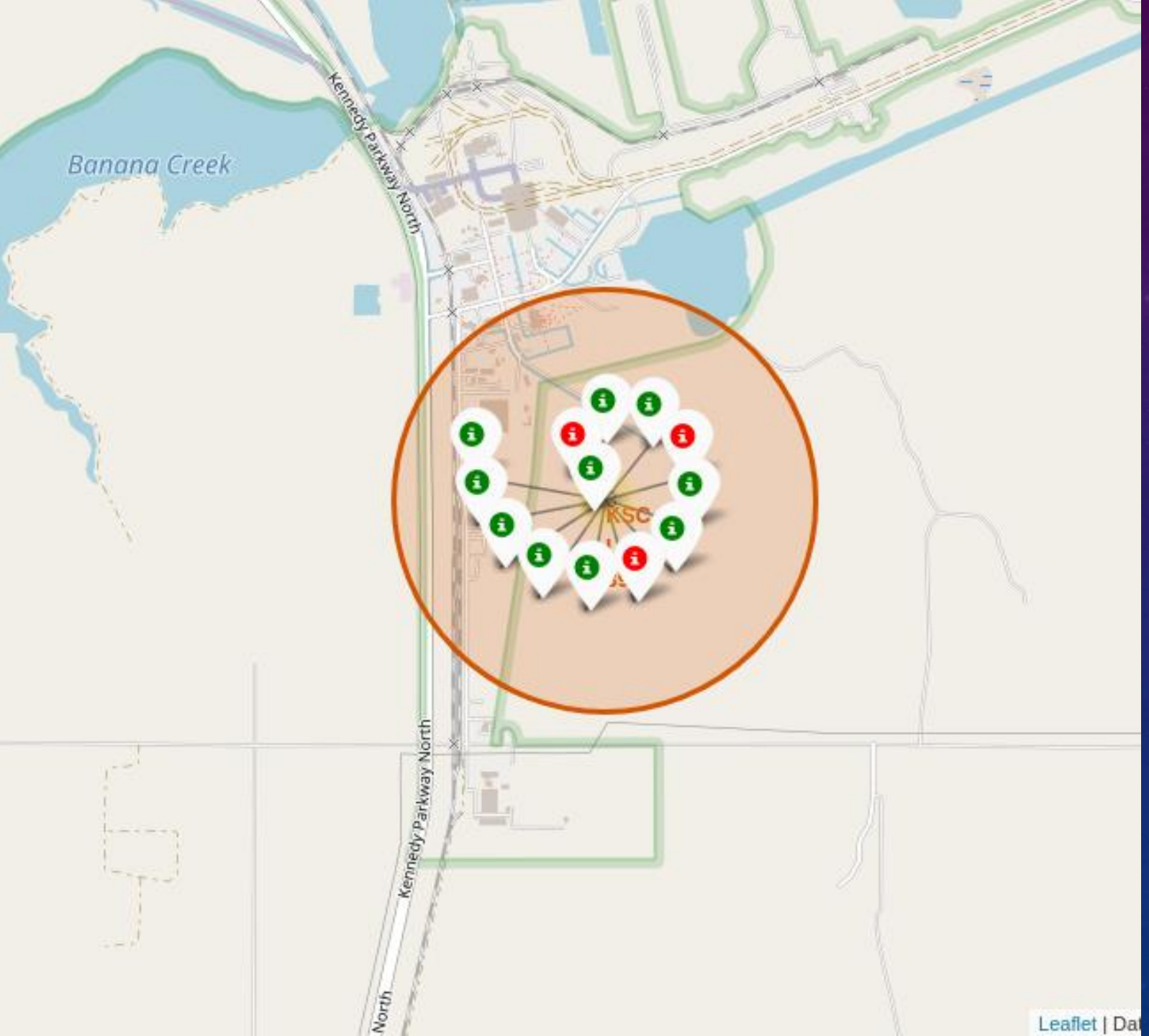


LAUNCH OUTCOMES FOR CCAFS LC-40

When looking at the sequence of launches from CCAFS LC-40, one can see that the first launches failed to land the booster. Later launches succeeded.

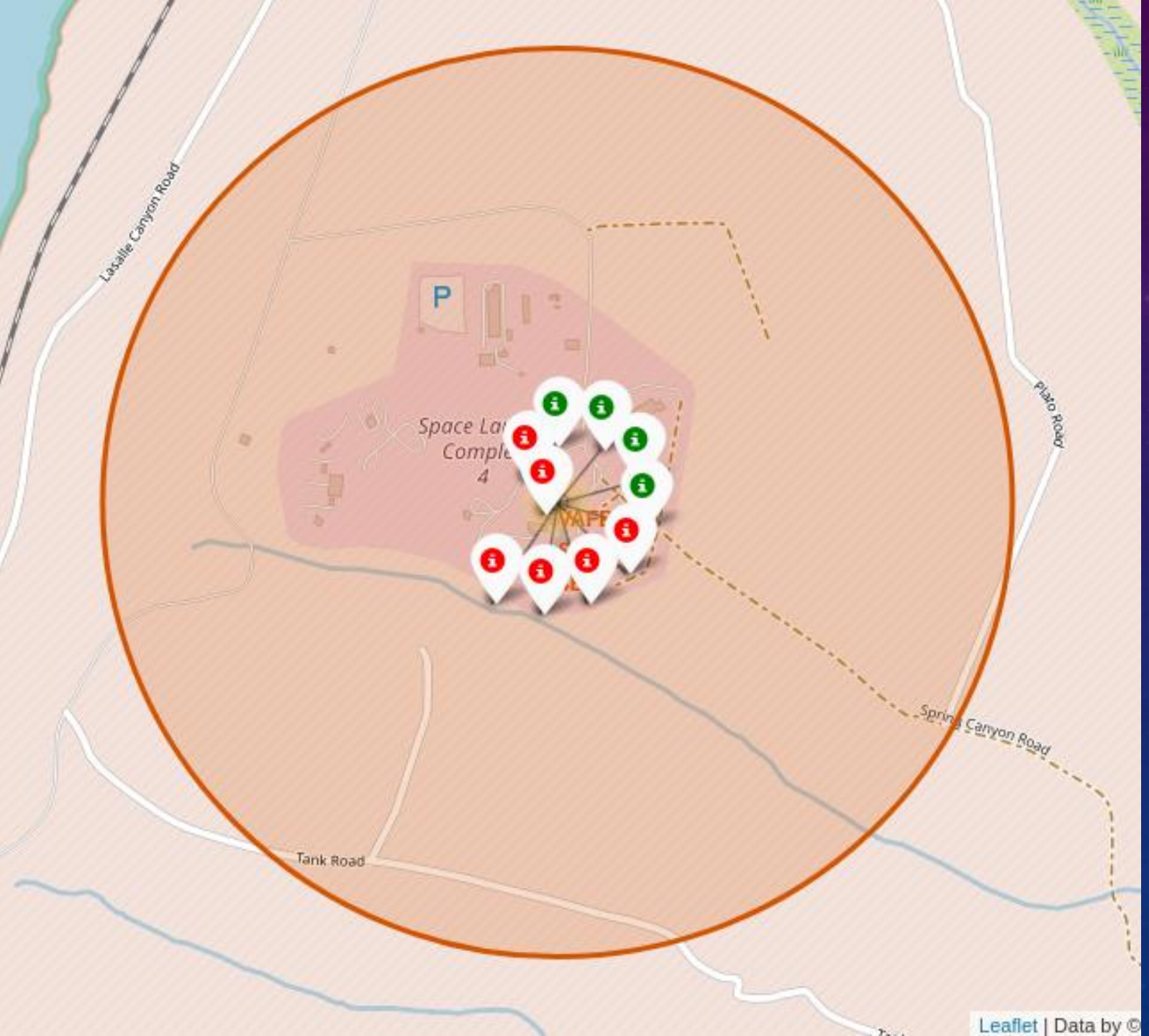
LAUNCH OUTCOMES FOR CCAFS SLC-40

When looking at the sequence of launches from CCAFS SLC-40, one can see that the first launches failed to land the booster. Later launches succeeded.



LAUNCH OUTCOMES FOR KSC LC-39A

When looking at the sequence of launches from KSC LC-39A, one can see that the nearly all of the launches succeeded.



LAUNCH OUTCOMES FOR VAFB SLC-4E

When looking at the sequence of launches from VAFB SLC-4E, one can see that the first launches failed to land the booster. Launches later in the sequence succeeded. Finally, the last launches failed.

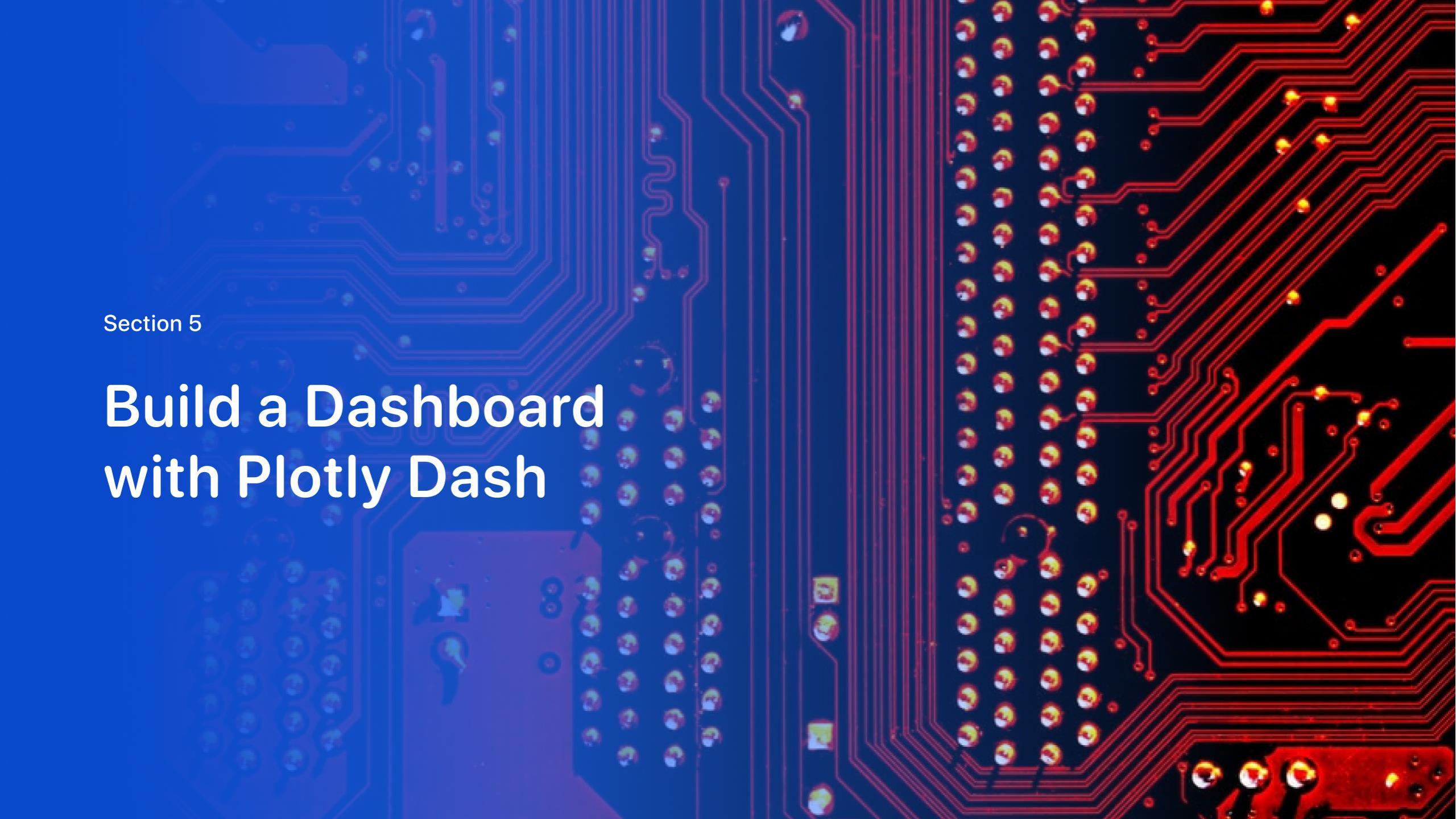
DISTANCE BETWEEN CCAFS SLC-40 AND COASTLINE

The launch site is 0.90 km from the coastline, and is near to highways and railroads. It is distant from major population centers.

Coastline: nearness allows launches to be diverted to the ocean in case of malfunction.

Highways & railroads: nearness allows launch vehicles to be assembled elsewhere and easily transported to the launch site.

Population Centers: distance allows protection to civilians in case of malfunction.

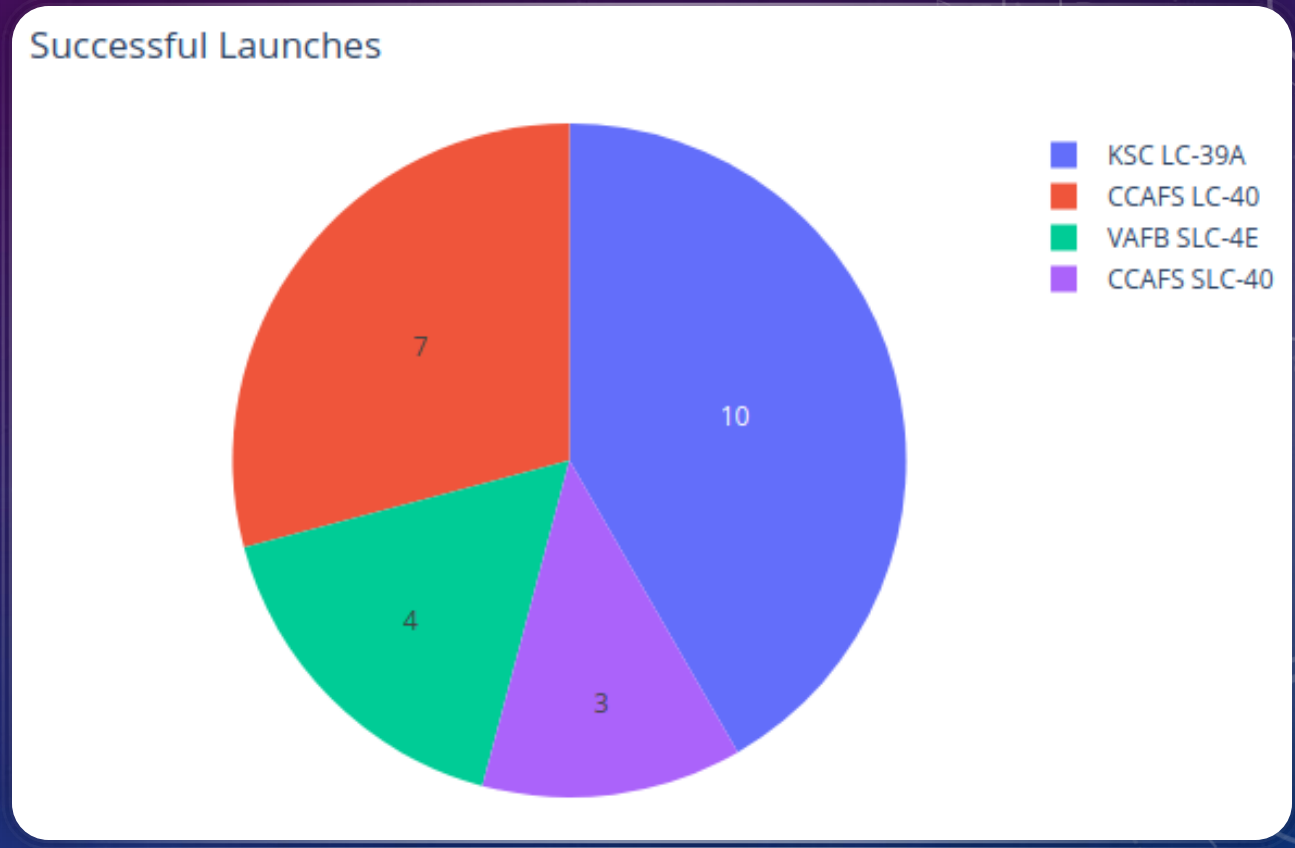


Section 5

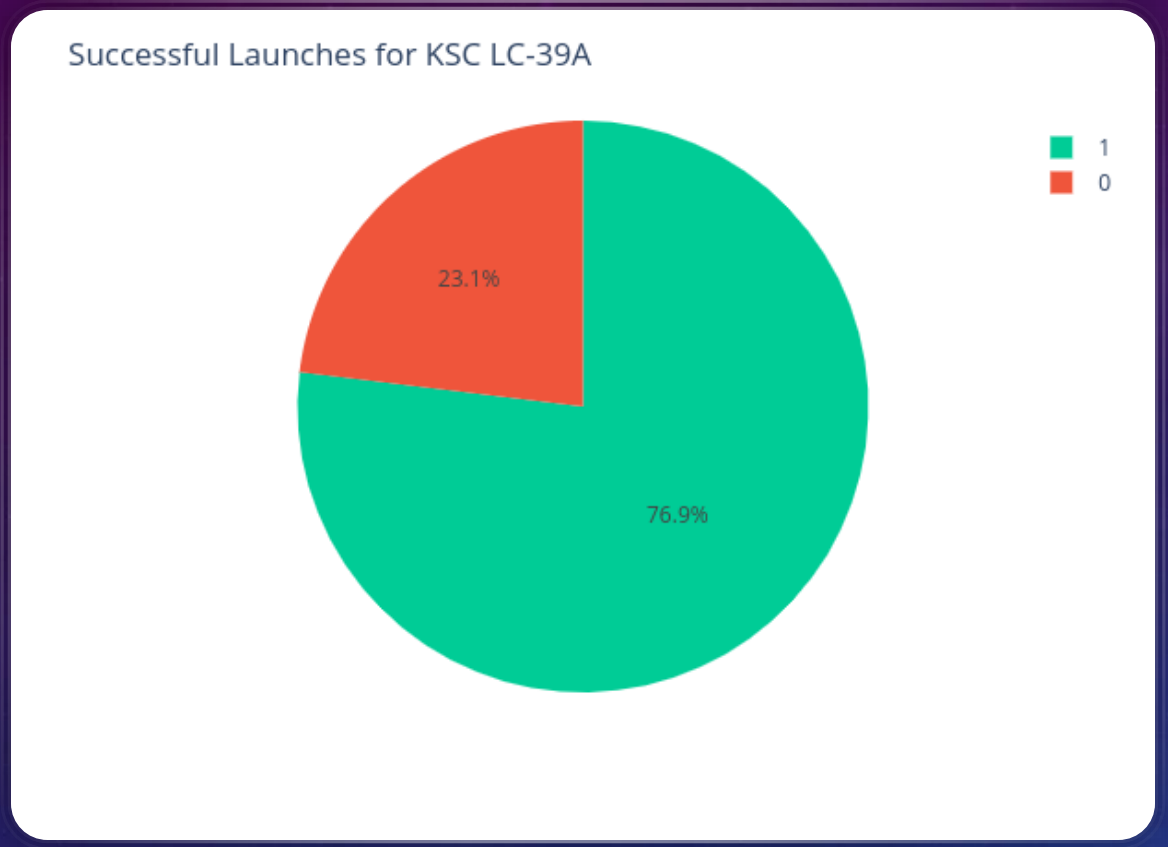
Build a Dashboard with Plotly Dash

SUCCESSFUL LAUNCHES BY LAUNCH SITE

The launch site KSC LC-39A shows the most successful launches, while CCAFS SLC-40 shows the least number of successful launches.

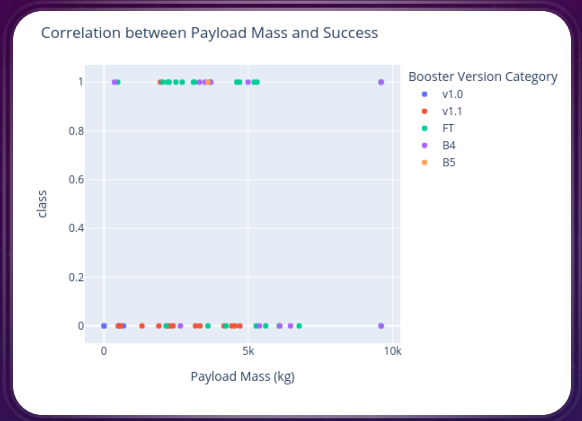


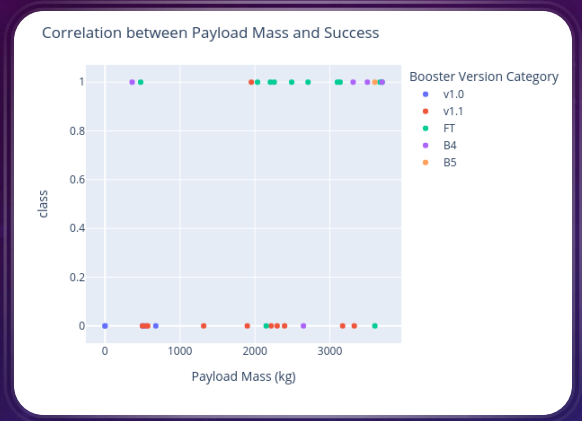
KSC LC-39A: HIGHEST LAUNCH SUCCESS RATIO

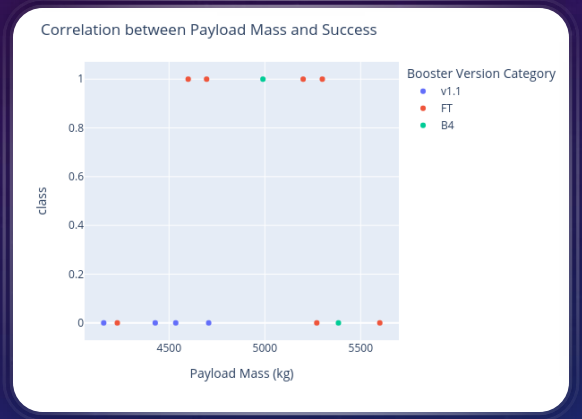


Launch site KSC LC-39A has the highest ratio of launch successes at **76.9%**.

PAYLOAD MASS VS LAUNCH OUTCOME







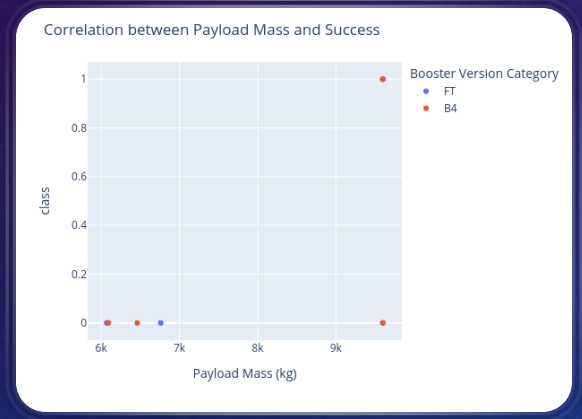


Image 1: Full Mass Range (0-10000 kg)

Image 2: Low Mass Range (0-4000 kg)

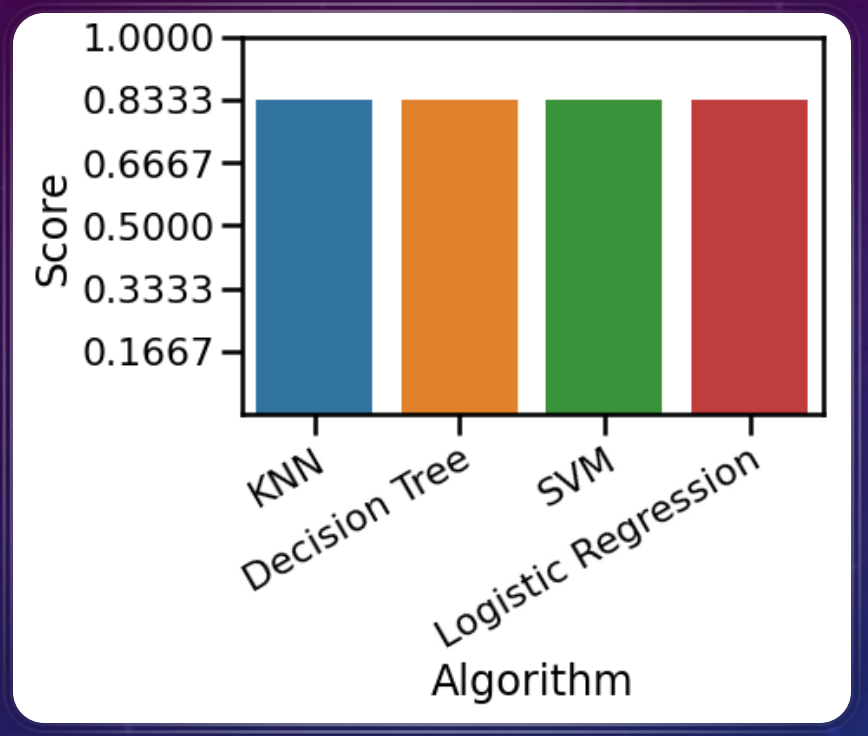
Image 3: Mid Mass Range (4000-6000 kg)

Image 4: High Mass Range (6000-10000 kg)

Section 6

Predictive Analysis (Classification)

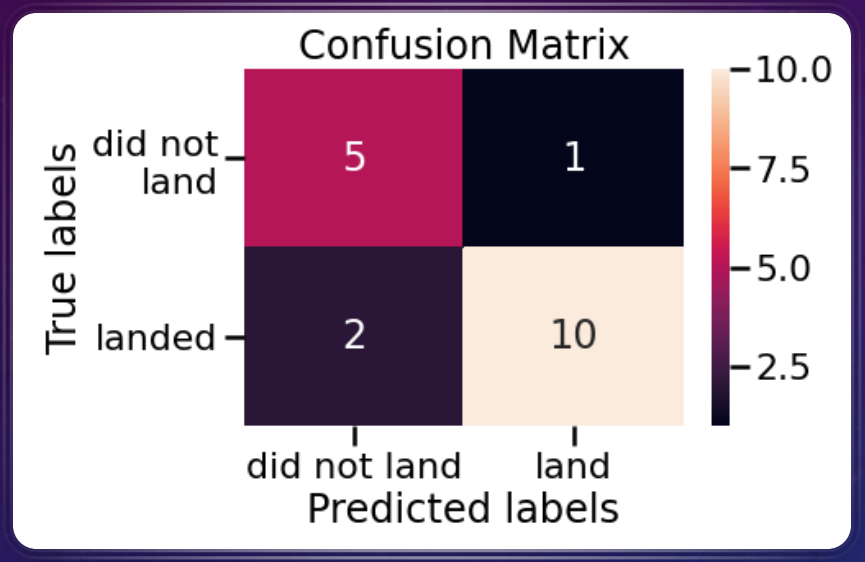
CLASSIFICATION ACCURACY



All models show the same accuracy. This is most likely due to a small test sample (only 18 samples to test). Better differentiation between algorithms can be made using a larger data set.

Decision Tree algorithm was chosen due do a better confusion matrix (see next slide).

CONFUSION MATRIX



Examining the confusion matrix, we see that a decision tree classifier can distinguish between the different classes.

CONCLUSIONS

- Successful landings of boosters increases over time, as the company gains more experience.
- Launches should be done near the coast, with sites having easy access to highways and railroads, but at some distance from major cities.
- As the payload mass goes up, the likelyhood of a successful landing decreases.
- A decision tree classification model can be used to predict whether a launch will be successful or not.

Thank you!