IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ansh Srivastava
Oct,4,2022

Outline

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

Executive Summary

- Principal methods
  - Data collection through web scraping
  - Data wrangling with Pandas
  - Exploratory analysis with SQL
  - Exploratory analysis with data visualization
  - Interactive mapping with Folium
  - Machine learning predictions
- Important results
  - SQL revealed that successful landing outcomes are increasing over time.
  - Interactive analytics enables user-input exploration of launch sites
  - Machine learning with Decision Tree models revealed the highest training accuracies

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX’s accomplishments include sending spacecraft to the International Space Station; Starlink, a satellite internet constellation providing satellite Internet access; And sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Spaces X’s Falcon 9 launches like regular rockets. The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. The first stage does most of the work and is much larger than the second stage. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and the customer’s needs.

For fictional company Space Y, our job was to determine the price of each launch. We do this by gathering information about Space X and creating dashboards. We also determined when SpaceX would reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, we trained a machine learning model and used public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected using SpaceX API and Wikipedia
- Data was wrangled using Matplotlib and Pandas
- Exploratory data analysis (EDA) was conducted using visualization and SQL
- Application was developed for interactive visual analytics using Folium and Plotly Dash
- Predictive analysis was performed using various classification models
    - Logistic regression
    - SVM
    - Classification Tree
    - KNN

Data Collection

Data was collected using two methods:

- Get requests on the SpaceX API
- Web scraping using the BeautifulSoup package on the Falcon 9 launch records presented on Wikipedia
- Ultimately launch records were extracted as an html table and converted to a pandas dataframe

Data Collection – SpaceX API

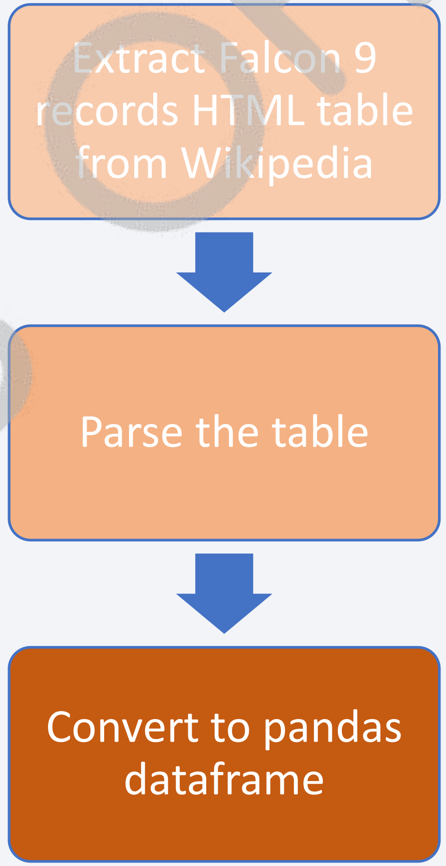
- Collected data from the SpaceX API using the python package *Requests*

- The github code can be found here.

Data Collection - Scraping

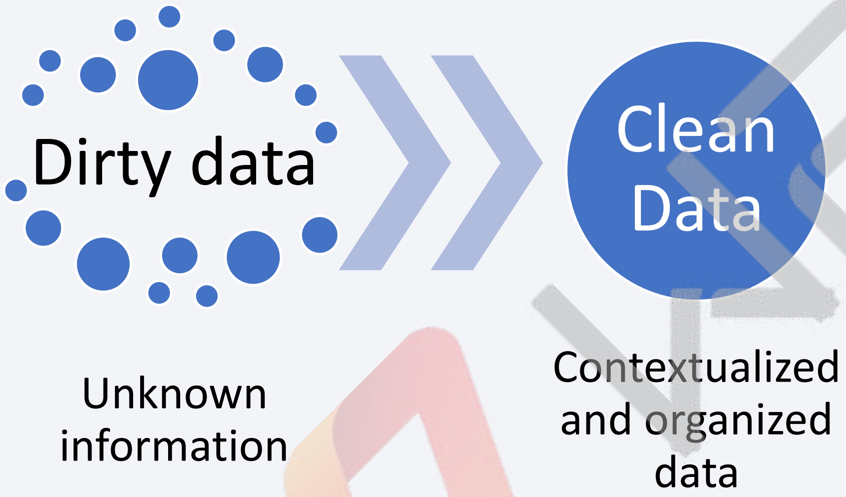
- Collected data via web scraping the Falcon 9 Wikipedia page using the python package *BeautifulSoup*

- The code to the scraping exercise can be found here.



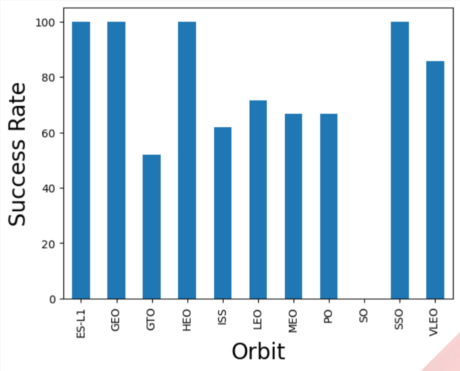
Data Wrangling

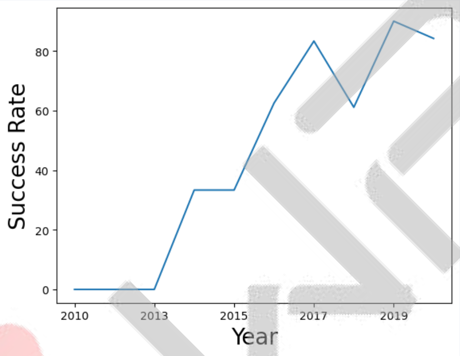
- Dataframe from the web scraping + API request was used
- Data was initially explored using value_counts(). Additionally, columns for launch outcomes were generated for visual analysis.
- The code can be found [here](#).



EDA with Data Visualization

- Visual analysis was performed using scatter and bar plots.
  - These plots enable relational dynamics to be easily identified. See example below.
- The code for this process can be found here.



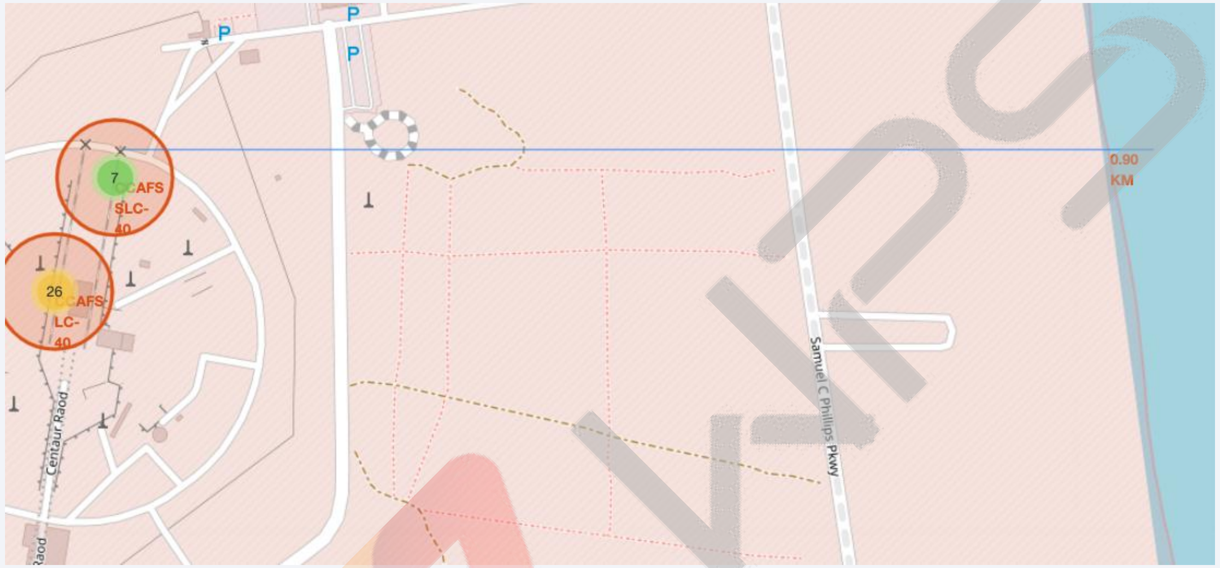


EDA with SQL

- 10 SQL queries were performed to display the following:
  - Names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string '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 successful landing outcome in ground pad was achieved
  - Names of the 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
  - Names of the booster versions which have carried the maximum payload mass
  - Month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in year 2015
  - Count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order
- The code for these queries and their results can be found [here](#).

Build an Interactive Map with Folium

- Markers, circles, lines, etc. were used to indicate launch sites.
- The code for these interactive maps can be found here.



Build a Dashboard with Plotly Dash

- Pie charts and scatter plots were used for visualization of:
  - Launches by site (pie)
  - Payload range by class (slider, scatter)
- The pie chart will allow you to visualize percentage of launches for a given site.
- The scatter plot will show payload ranges for the given site and given range.
- The code for this interactive app can be found here.

Predictive Analysis (Classification)

- Predictive models were used to assess accuracy of predicted outcomes given the SpaceX dataset.
- The four models used: logistic regression (LR), support vector machine (SVM), decision tree (DT), and k nearest neighbors (KNN).
- The code for this model assessment can be found here.

Standardize the data

- This prevents a wide range of values and outliers from skewing the data

Split the data

- Splitting the data allows us to create a training set to train the model and a test set to test the accuracy

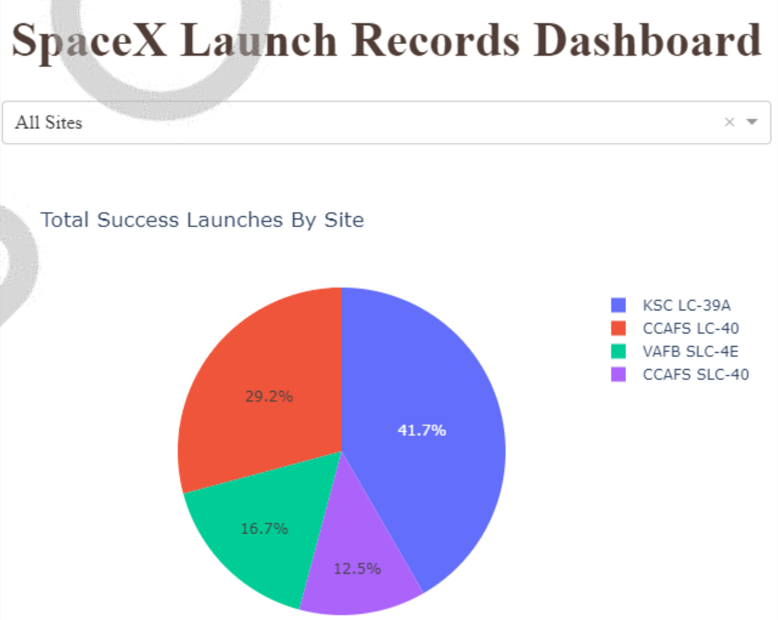
Evaluate the data using each model

- Comparing the accuracy of prediction for each model will allow us to choose a best fit option

Results

- There are 4 SpaceX launch sites
  - LC-40, SLC-40, LC-39A, and SLC-4E
- SLC-40 or Cape Canaveral, has the most successful launches and largest payloads
- The average payload mass carried by F9 v1.1 is 2,928 kg
- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- Nearly 100% of launches have been successful

- Successful landing outcomes have increased over time



- Interactive folium maps revealed most launches occur at east coast sites near the coastline.
- Prediction models were all considerably accurate on the test data (>83%) but Decision Tree boasted the highest (89%) accuracy in training.

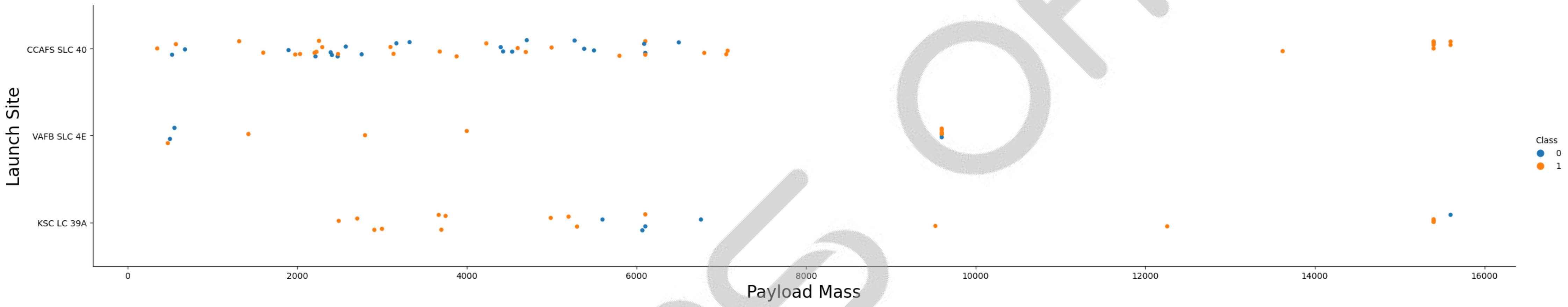
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- The Cape Canaveral SLC 40 has the highest number of successful landing outcomes, especially as time goes on.

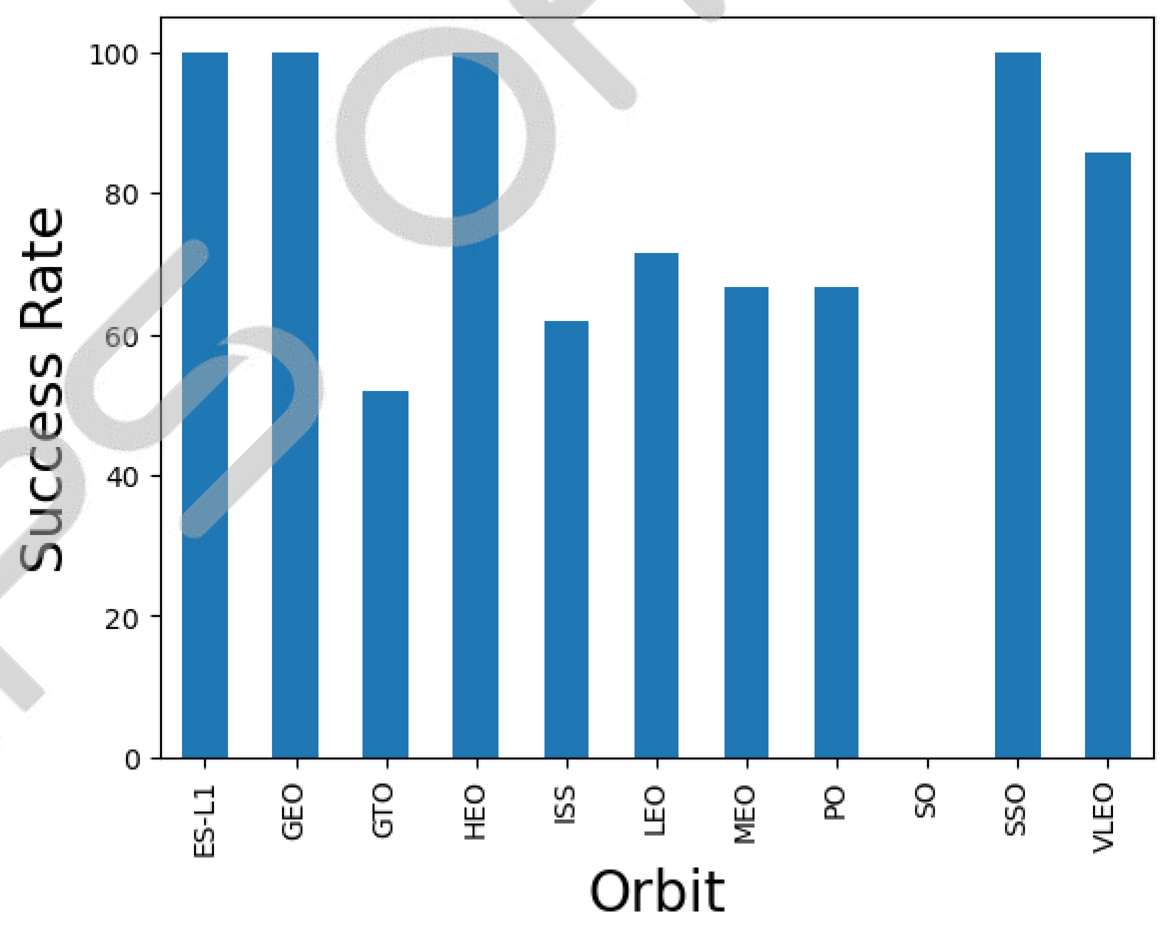
Payload vs. Launch Site



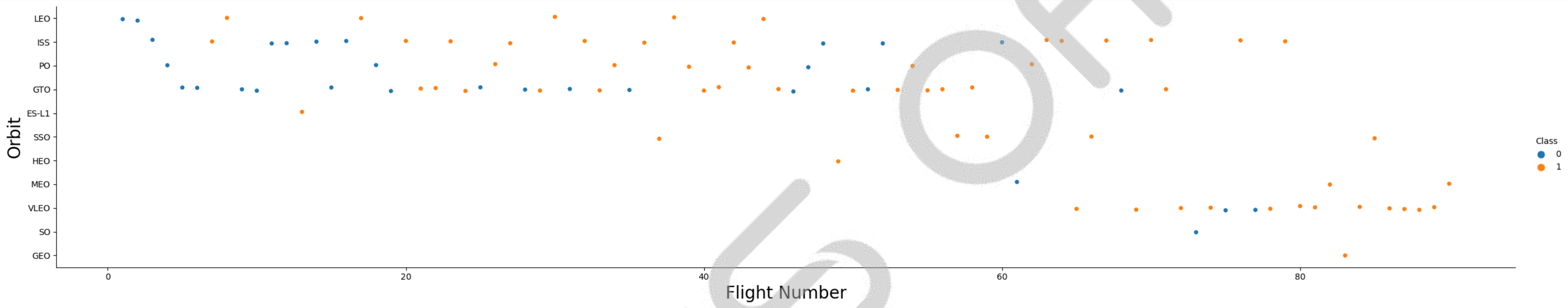
- Payloads over 9,000 kg have the highest success rate. Payloads over 10,000 kg are launched from either SLC 40 or LC39A.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO are the most successful orbits to launch into.
- GTO has the lowest success.

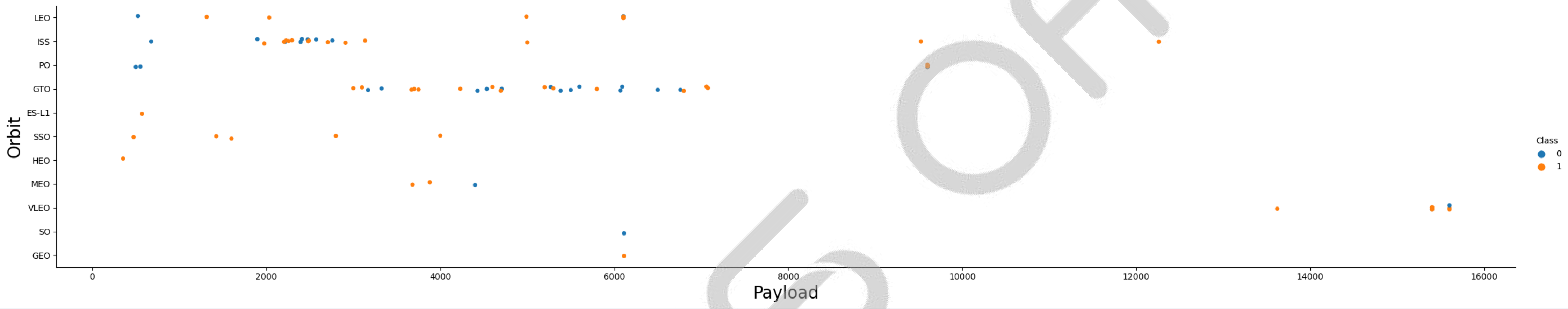


Flight Number vs. Orbit Type



- Even though ES-L1, GEO, HEO, and SSO are the most successful, here we can see that they represent a low number of flights.
- Alternatively, LEO has had only successful flights since being the first orbit SpaceX explored. They also represent a fair number of flights.
- LEO could be a good target for early flights.

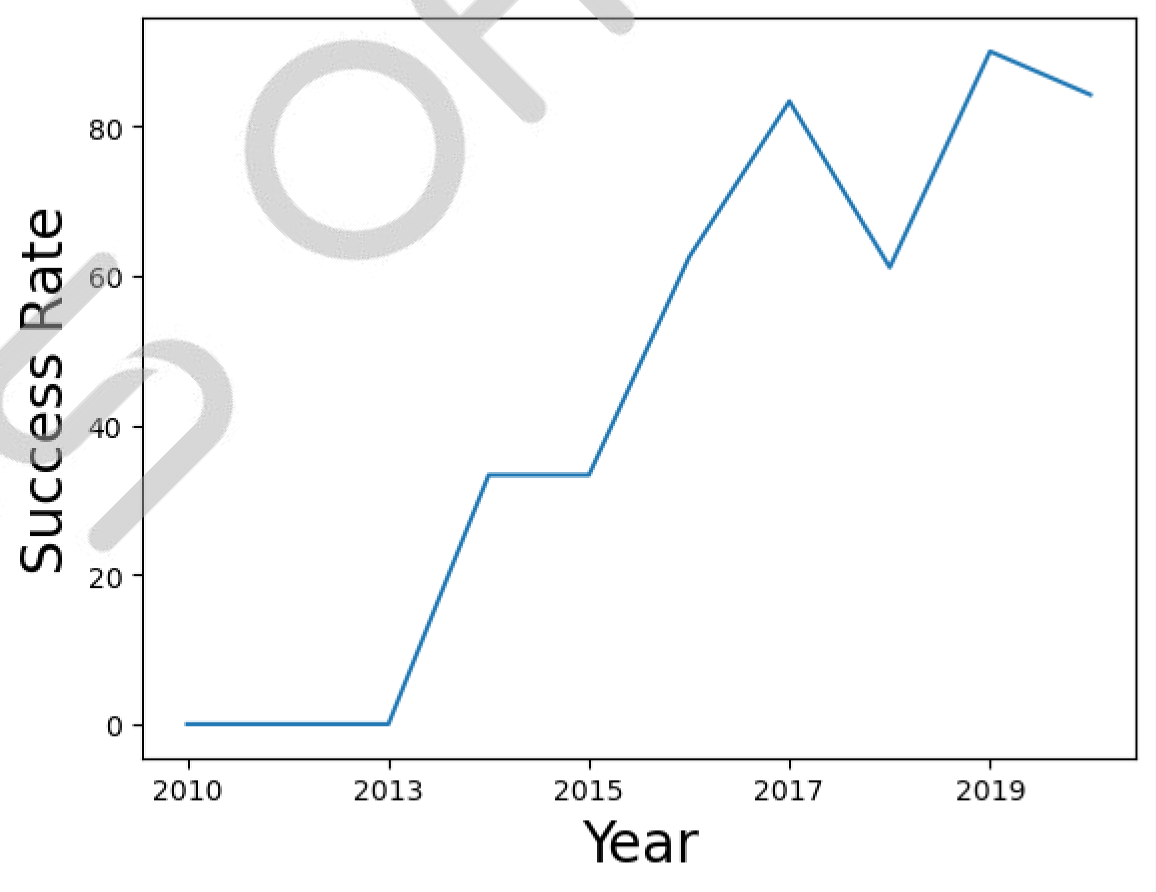
Payload vs. Orbit Type



- Larger payloads (>13,000 kg) fly to VLEO only.
- Payloads between 8,000 and 13,000 kg typically launch into ISS or PO.
- SSO is highly successful for payloads under 5,000 kg.

Launch Success Yearly Trend

- Launch success improved dramatically from 2013 to 2017 before taking a dip in 2018.
- 2019 represents the greatest success yet.



All Launch Site Names

- Launch sites
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
- Majority customer is NASA (4), with one customer self-representing SpaceX.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
- 111,268 kg

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
- 2928.4 kg

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';
- 01-05-2017

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

- sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "Landing _Outcome" = 'Success (drone ship)';
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

Mission Outcome	QTY
Failure (in flight)	1
Success	100

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date, 4, 2) AND substr(Date,7,4) = '2015';

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

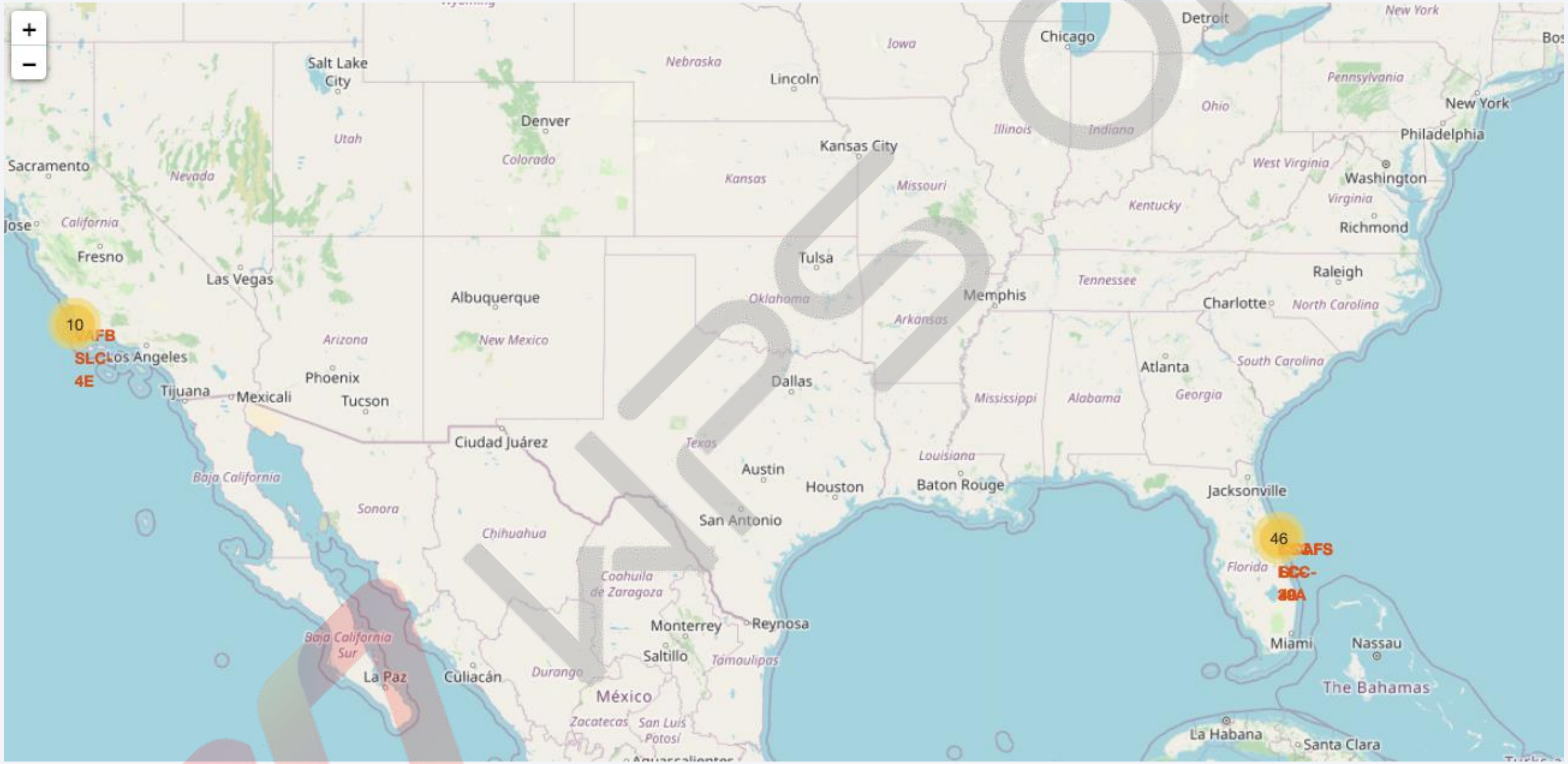
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- sql SELECT "Landing _Outcome", COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN substr(Date,7,4) = '2010' AND substr(Date, 4, 2) AND substr(Date,7,4) = '2017' GROUP BY "Landing _Outcome" ORDER BY QTY DESC;

Landing Outcome	QTY
Success (ground pad)	3
Success (drone ship)	1
No attempt	1

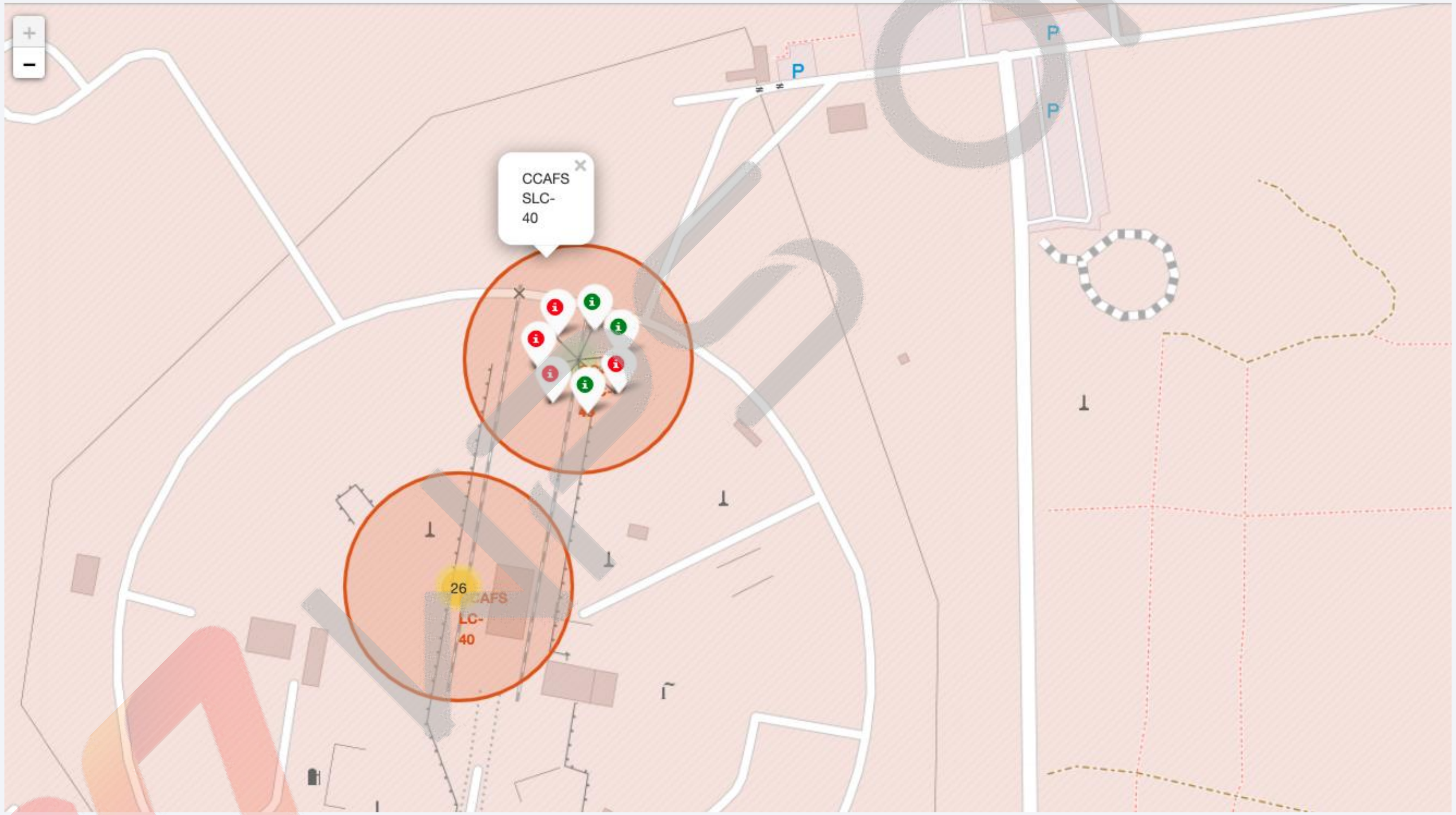
Section 3

Launch Sites Proximities Analysis

Map of Launch Sites



Launch Site Success

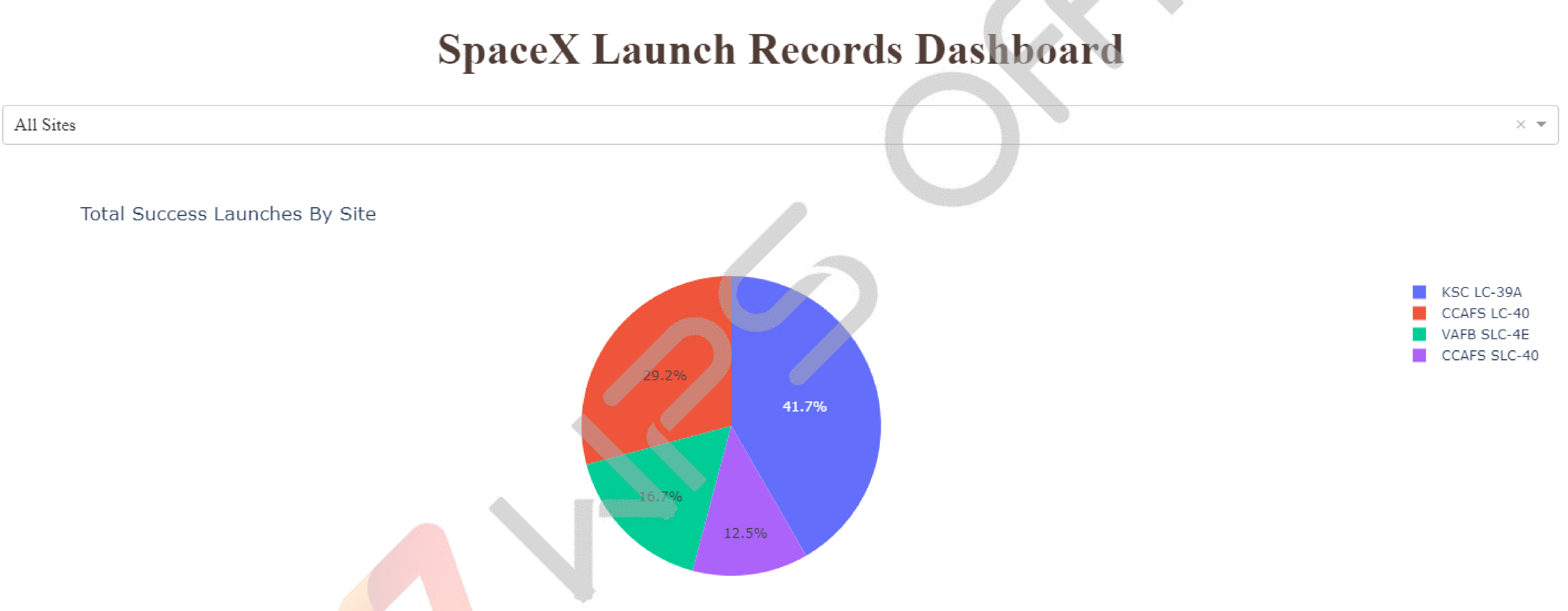


Map of Coastline Proximity

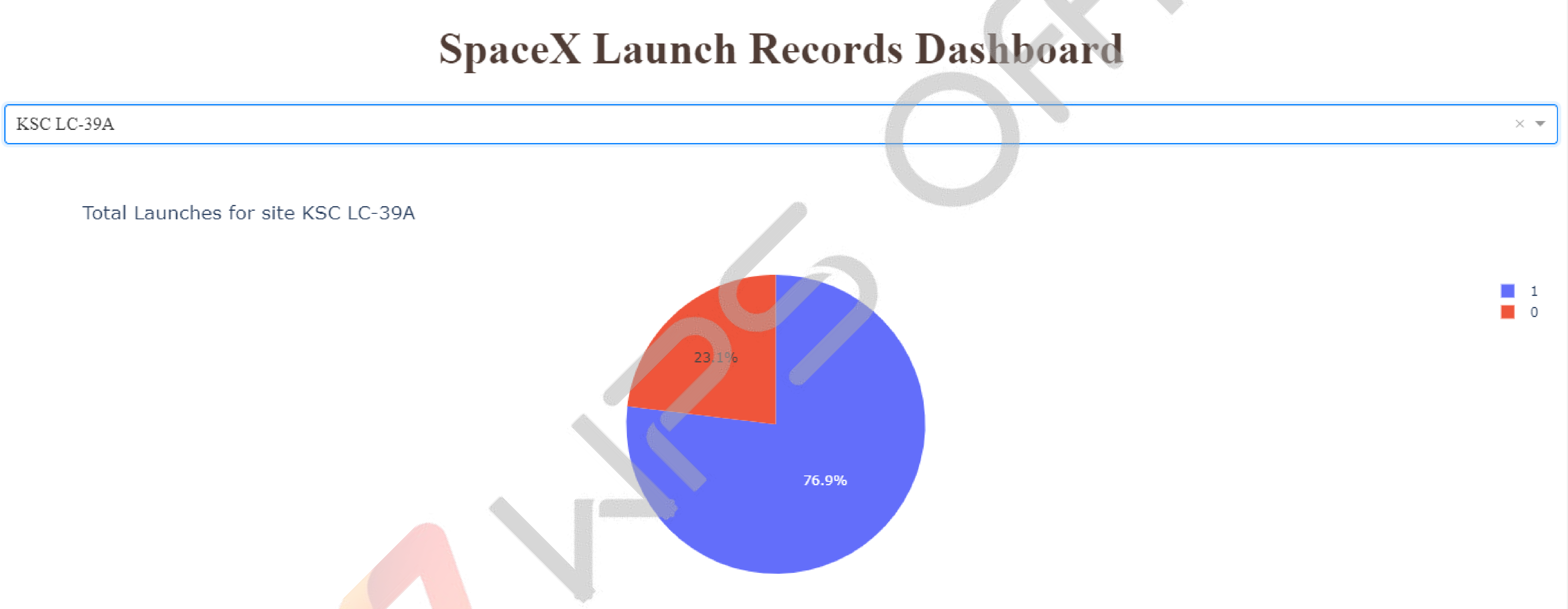
Section 4

Build a Dashboard with Plotly Dash

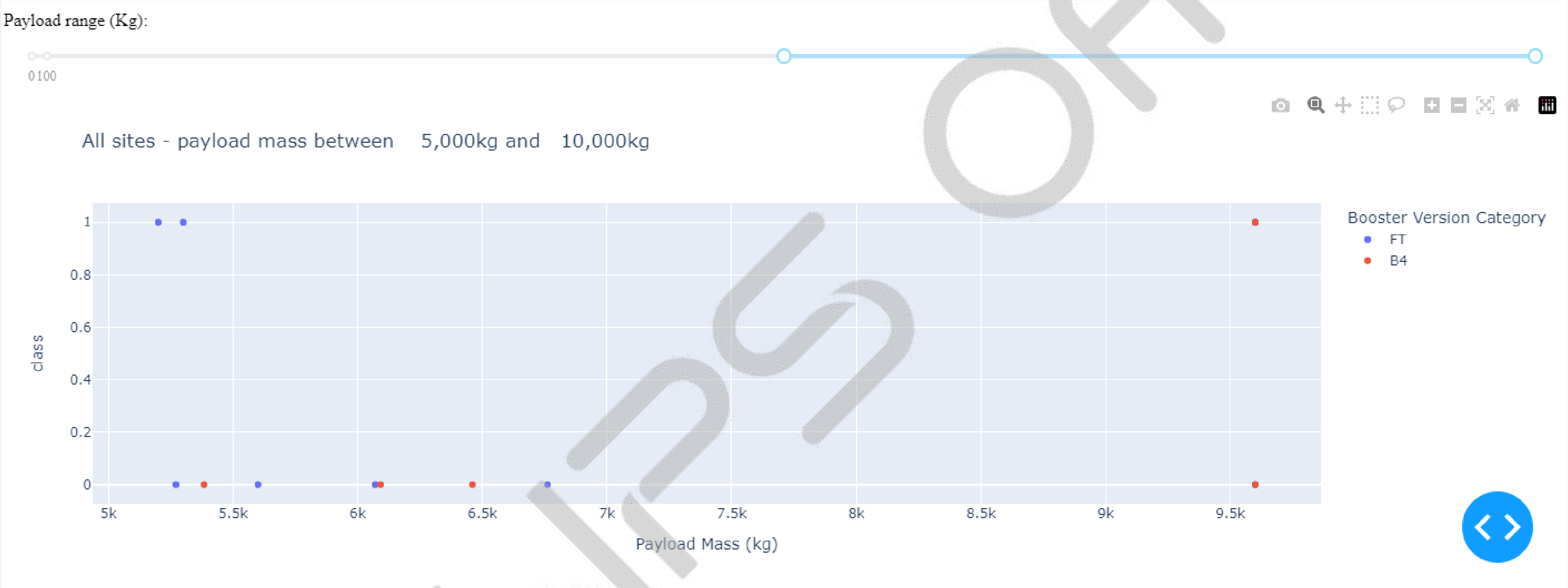
Interactive Dashboard App for Launch Site Exploration



Launch Site with Highest Success Rate



Payload VS Launch Outcome



- From 5,000 to 9,000 kg, FT boosters had the only successful launches (2).

Section 5

Predictive Analysis (Classification)

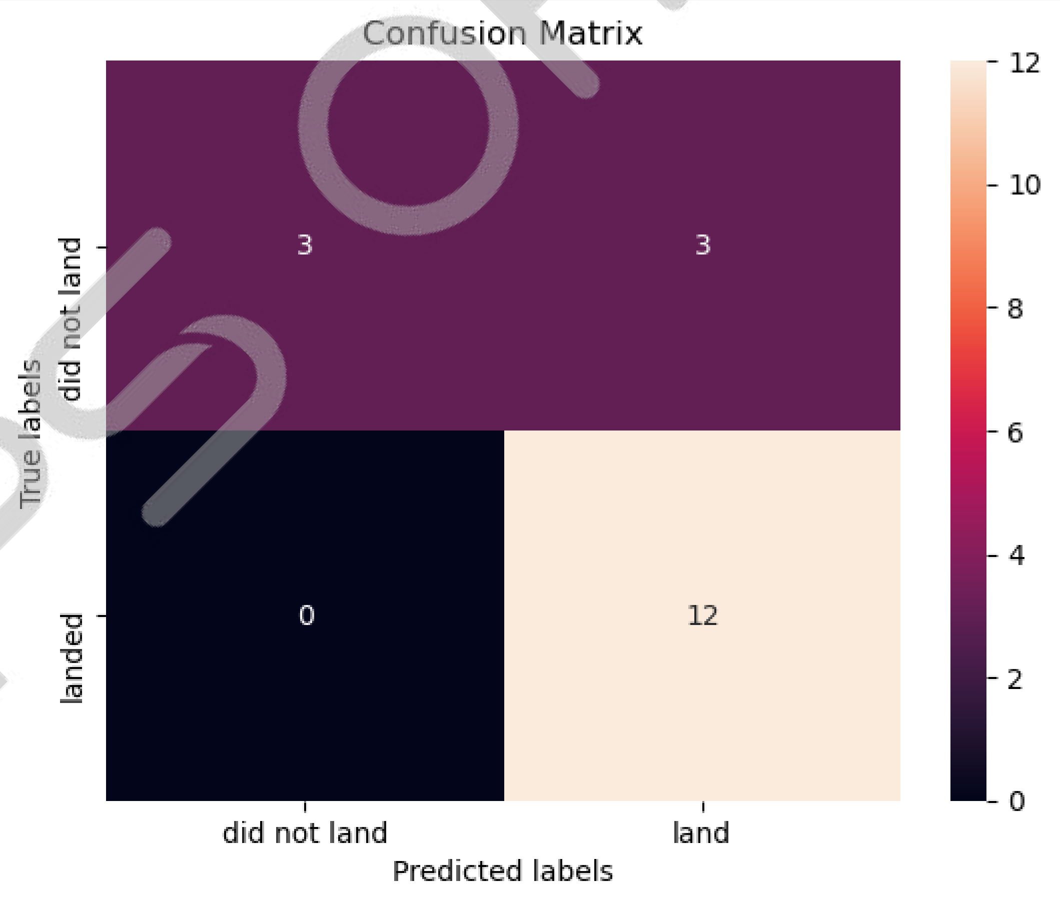
Classification Accuracy

We can see here that Decision Tree has the highest training accuracy, with all methods sharing the same test accuracy.

Model	Accuracy	Test Accuracy
Log Reg	0.84722	0.83333
SVM	0.84722	0.83333
Tree	0.88889	0.83333
KNN	0.84722	0.83333

Confusion Matrix

- All models had the same confusion matrix results, with 12 out of 15 landing.



Conclusions

- Cape Canaveral SLC-40 Launch Site proves the best launching location for successful launches.
- LEO orbit should be a target for test flights.
- The heavier the payload, the better the result.
- Decision Tree classification should be used for predictive modeling.

Thank you!