IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nguyễn Huy Hoàng
12/16/2021

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with SpaceX API
  - Data collection with Web Scraping
  - Data wrangling
  - EDA with Data visualization
  - EDA with SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- In the era of commercial space, companies are aiming for space travel. Among them, SpaceX are the most prominent one since that their rocket launching costs are relatively inexpensive. It is known that the launching cost will significantly reduce if we can predict whether the first stage will land.

- Therefore, in order to be able to compete with SpaceX by reducing launching cost, we need to build a machine learning model to predict the outcome of the first stage (success/failed) based on SpaceX information.

Section 1

Methodology

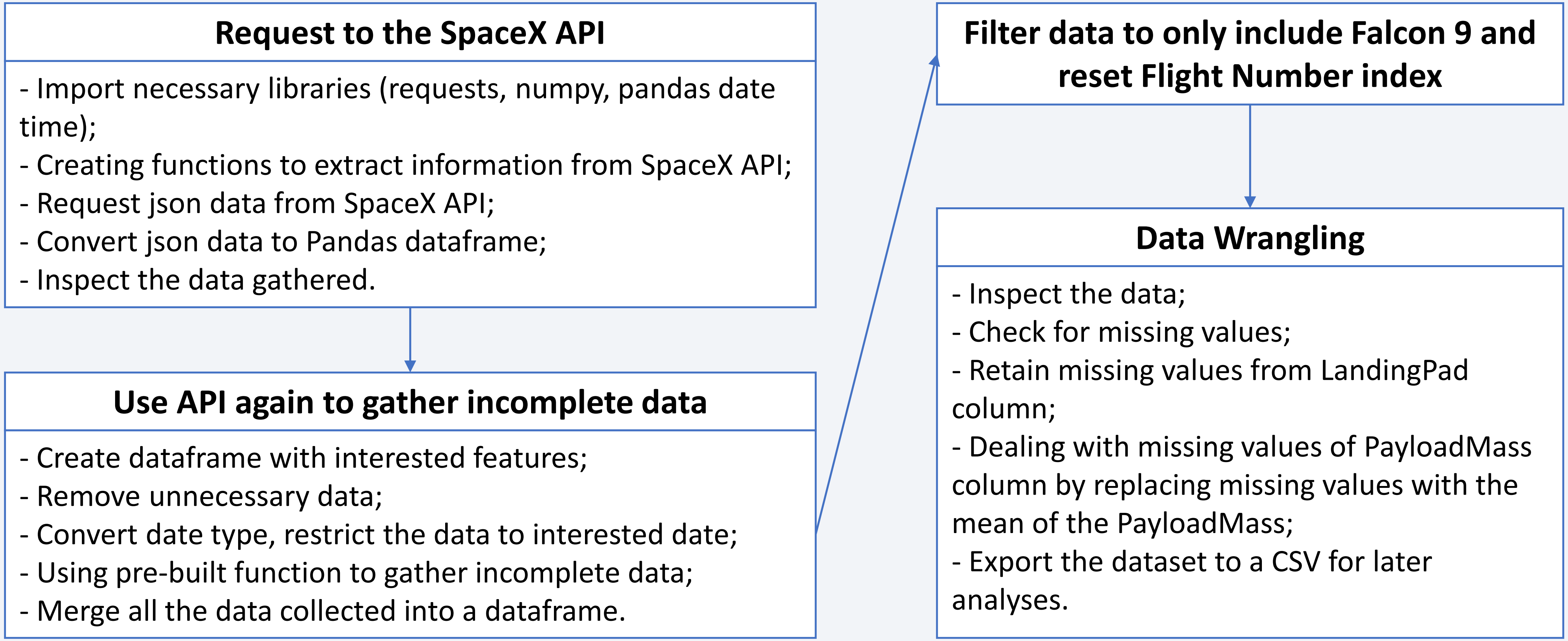
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX Rest API and scraped from the Wikipedia page List of Falcon 9 and Falcon Heavy launches.
- Perform data wrangling
  - Transform launch outcome variable to a binary variable (success/failure landing)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The set was splited into training and testing set. Classification models using different methods was built using the training set. The model was then evaluated using the testing set.

Data Collection – SpaceX API

Flowchart of SpaceX API



Request to the SpaceX API

- Import necessary libraries (requests, numpy, pandas date time);
- Creating functions to extract information from SpaceX API;
- Request json data from SpaceX API;
- Convert json data to Pandas dataframe;
- Inspect the data gathered.

Use API again to gather incomplete data

- Create dataframe with interested features;
- Remove unnecessary data;
- Convert date type, restrict the data to interested date;
- Using pre-built function to gather incomplete data;
- Merge all the data collected into a dataframe.

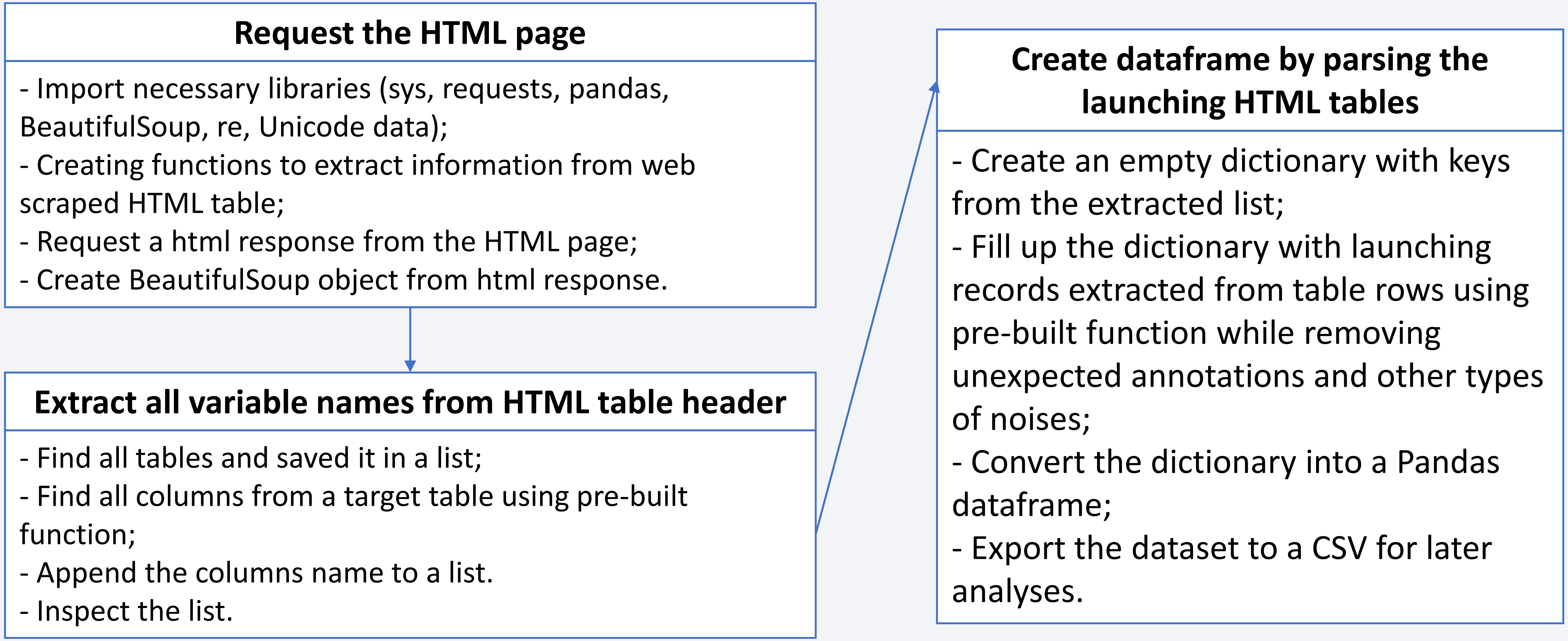
Filter data to only include Falcon 9 and reset Flight Number index

Data Wrangling

- Inspect the data;
- Check for missing values;
- Retain missing values from LandingPad column;
- Dealing with missing values of PayloadMass column by replacing missing values with the mean of the PayloadMass;
- Export the dataset to a CSV for later analyses.

Data Collection - Scraping

Flowchart of Web Scraping



Request the HTML page

- Import necessary libraries (sys, requests, pandas, BeautifulSoup, re, Unicode data);
- Creating functions to extract information from web scraped HTML table;
- Request a html response from the HTML page;
- Create BeautifulSoup object from html response.

Extract all variable names from HTML table header

- Find all tables and saved it in a list;
- Find all columns from a target table using pre-built function;
- Append the columns name to a list.
- Inspect the list.

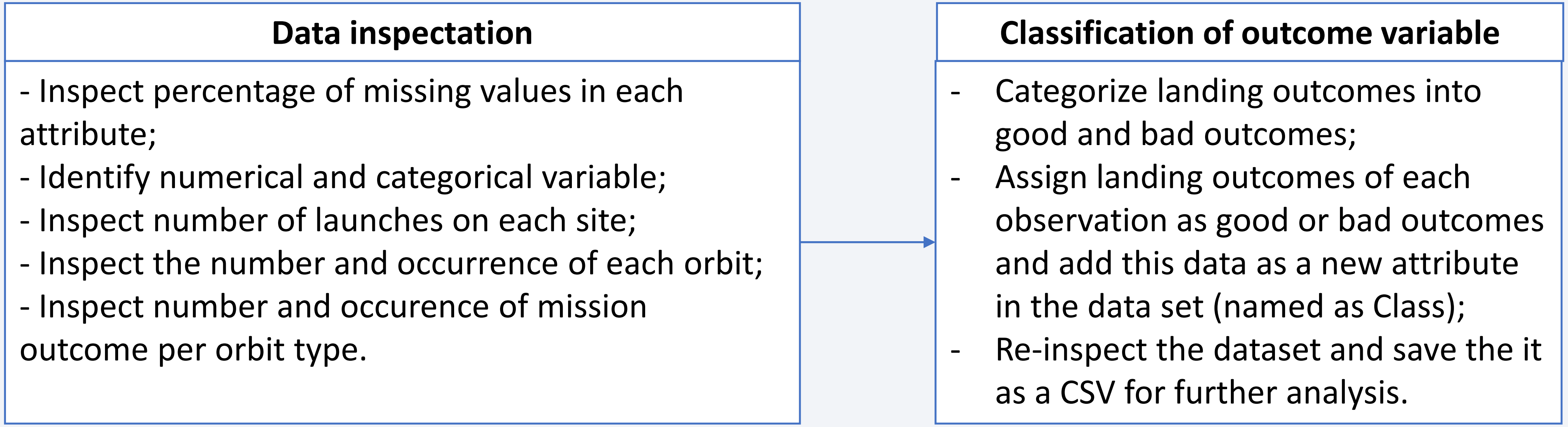
Create dataframe by parsing the launching HTML tables

- Create an empty dictionary with keys from the extracted list;
- Fill up the dictionary with launching records extracted from table rows using pre-built function while removing unexpected annotations and other types of noises;
- Convert the dictionary into a Pandas dataframe;
- Export the dataset to a CSV for later analyses.

Data Wrangling

- In this process, we aim to convert the launch outcomes to binary variable (0 as bad outcome/1 as good outcome) for model classification.

Flowchart of Data Wrangling



Data inspectation

- Inspect percentage of missing values in each attribute;
- Identify numerical and categorical variable;
- Inspect number of launches on each site;
- Inspect the number and occurrence of each orbit;
- Inspect number and occurence of mission outcome per orbit type.

Classification of outcome variable

- Categorize landing outcomes into good and bad outcomes;
- Assign landing outcomes of each observation as good or bad outcomes and add this data as a new attribute in the data set (named as Class);
- Re-inspect the dataset and save the it as a CSV for further analysis.

EDA with Data Visualization

- To visualize the relationship between two variables and launch outcomes, these scatter charts were used:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit type
  - Payload vs. Orbit type
- We used bar chart to visualize the relationship between Orbit type and it’s success rate.
- We used line chart to visualize launch success yearly trend.

EDA with SQL

- SQL queries used in the present report:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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.

Build an Interactive Map with Folium

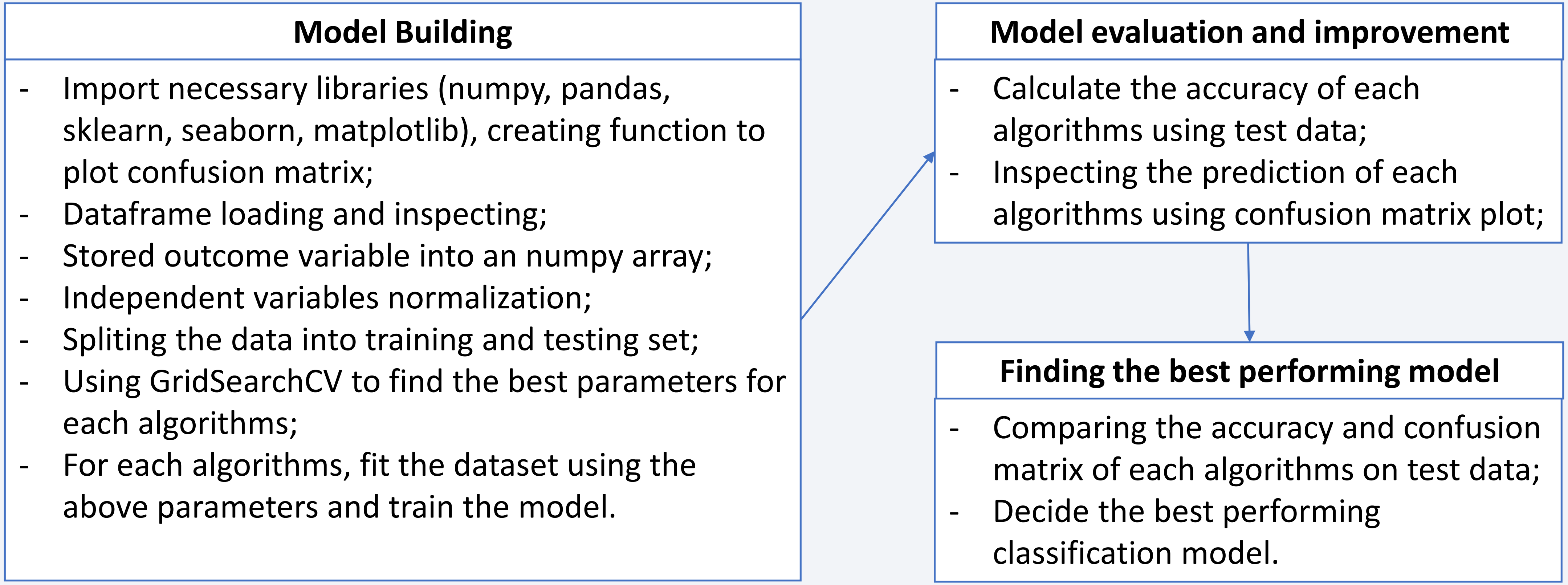
- We analyzed the existing launch site locations to find the optimal location for building a launch site.
- We used red circles and text to highlight all launch sites on the map.
- We used cluster objects to add markers of success and failure outcome on each site with Green markers represent success outcomes and Red markers represent failure outcomes.
- We used line objects to show the distance between a launch site (CCAFS SLC-40) and its proximities (closest railway, highway, coastline, city).

Build a Dashboard with Plotly Dash

- A Plotly Dash application was built for users to perform interactive visual analytics on SpaceX launch data in real-time.
- Users could select the graph for each launch site and the range of payload mass.
- We used bar chart to illustrate the total success launches for all site and the success rate of each site.
- We used scatter chart to illustrate the correlation between Payload and Success outcomes, varying across Booster version.

Predictive Analysis (Classification)

Flowchart of Classification Process



Model Building

- Import necessary libraries (numpy, pandas, sklearn, seaborn, matplotlib), creating function to plot confusion matrix;
- Dataframe loading and inspecting;
- Stored outcome variable into an numpy array;
- Independent variables normalization;
- Spliting the data into training and testing set;
- Using GridSearchCV to find the best parameters for each algorithms;
- For each algorithms, fit the dataset using the above parameters and train the model.

Model evaluation and improvement

- Calculate the accuracy of each algorithms using test data;
- Inspecting the prediction of each algorithms using confusion matrix plot;

Finding the best performing model

- Comparing the accuracy and confusion matrix of each algorithms on test data;
- Decide the best performing classification model.

Results

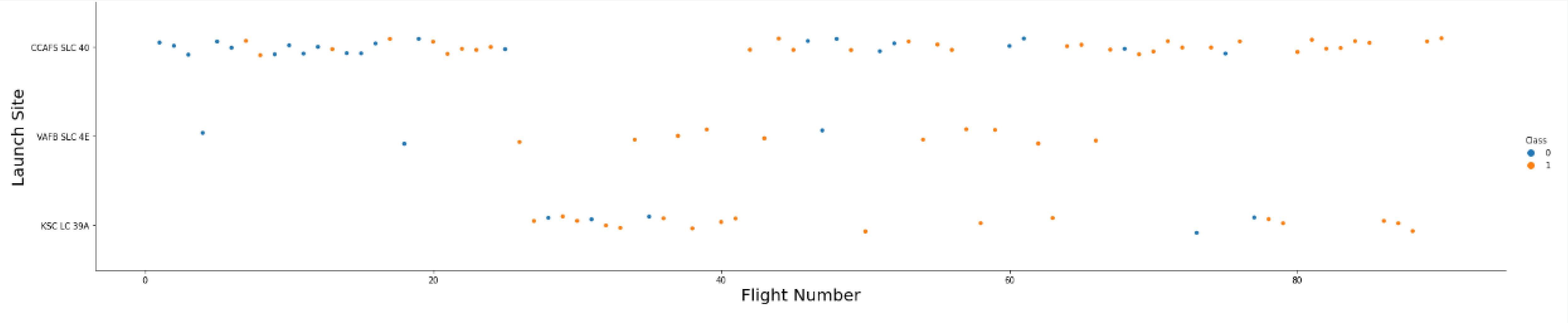
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

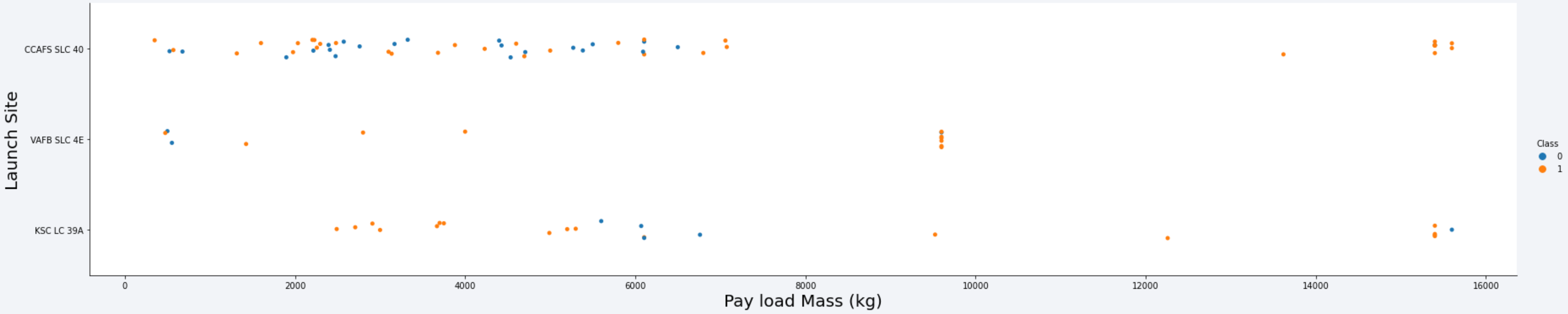
Flight Number vs. Launch Site

- In all sites, higher flight number are more likely to to be success.
- CCAFS SLC 40 has the most flight numbers and a large proportion of the first 25 flights are failed missions.



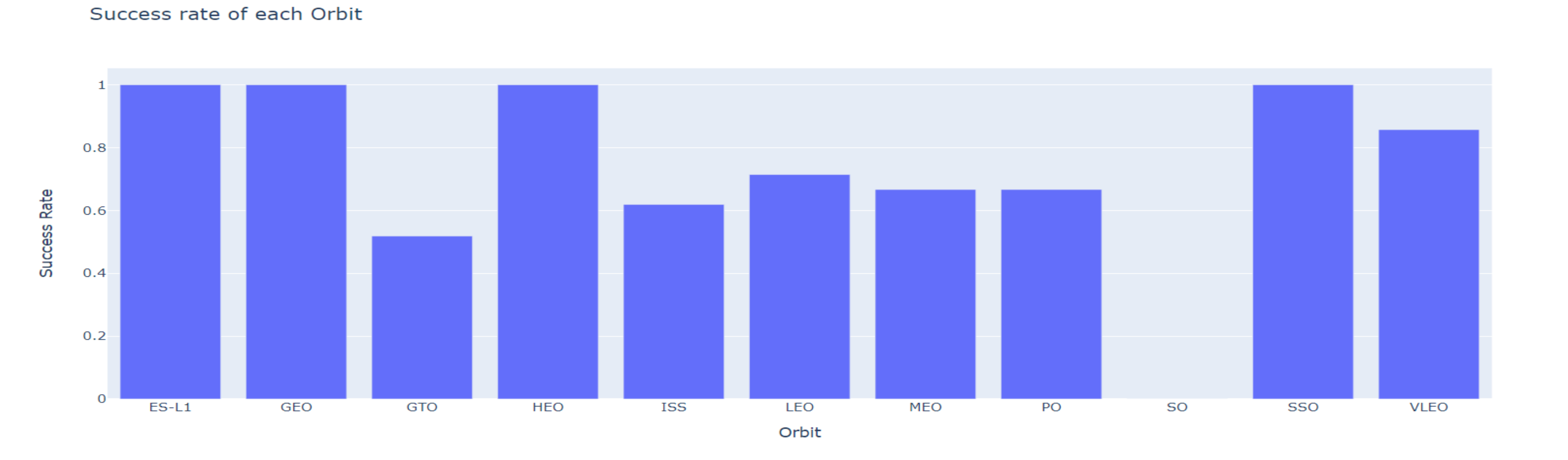
Payload vs. Launch Site

- In all sites, medium-to-high payload mass (8000 to 16000 kg) have high success outcomes.
- In all sites, the outcome is more ambiguous in the low payload mass flights (less than 8000 kg).



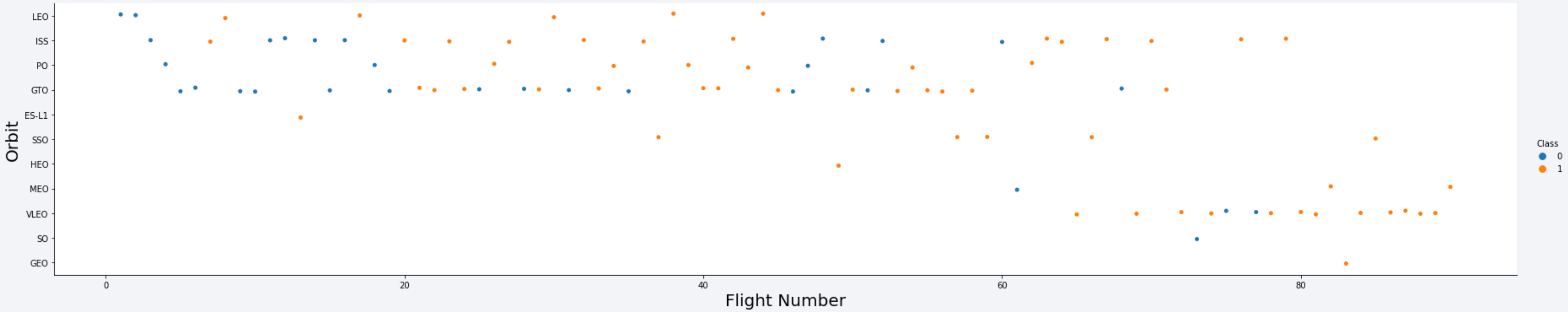
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbit types have highest success rate.
- No success observation was found for SO orbit type.



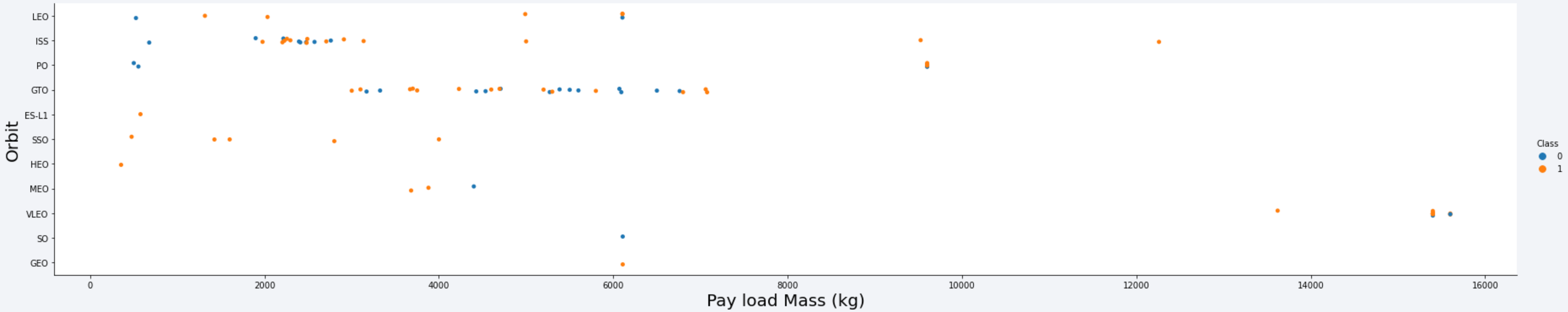
Flight Number vs. Orbit Type

- In LEO orbit type, high flight number is more likely to have success outcomes.



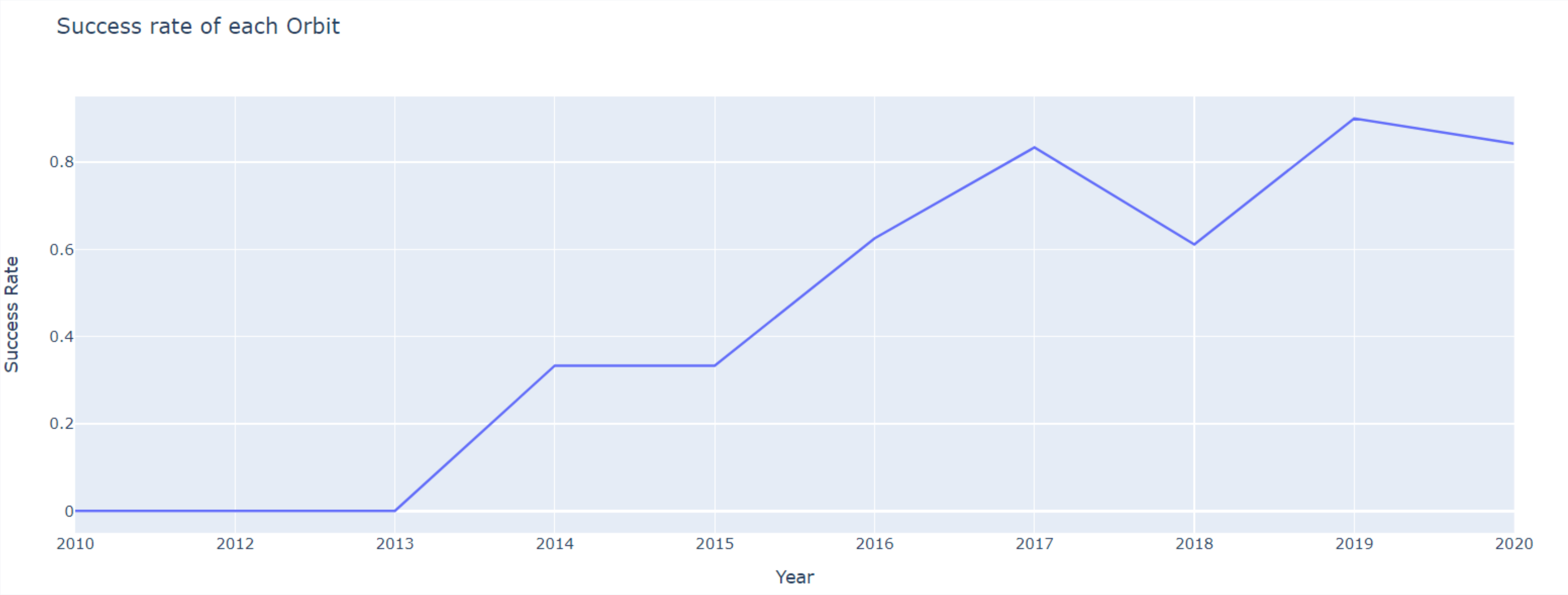
Payload vs. Orbit Type

- In LEO and ISS orbit types, higher playload mass is linked to success outcomes.



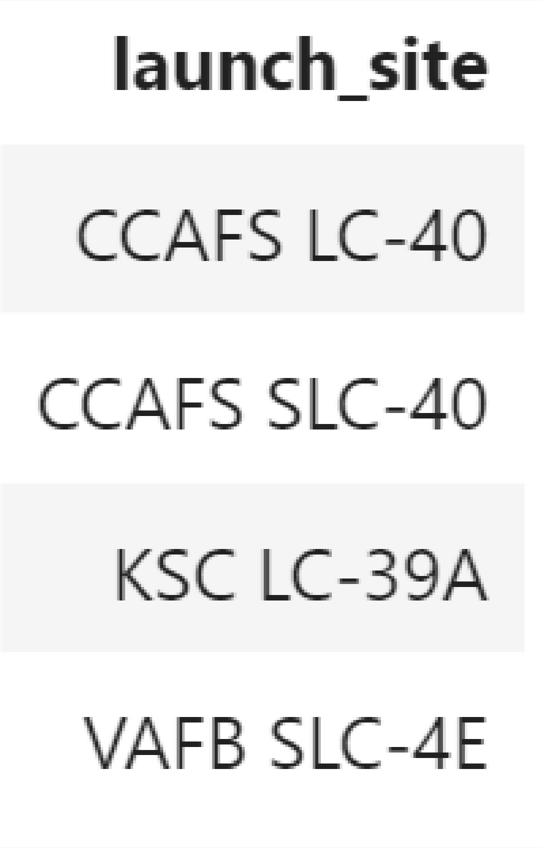
Launch Success Yearly Trend

- The chart indicates that the success rate increased over years starting from 2014.



All Launch Site Names

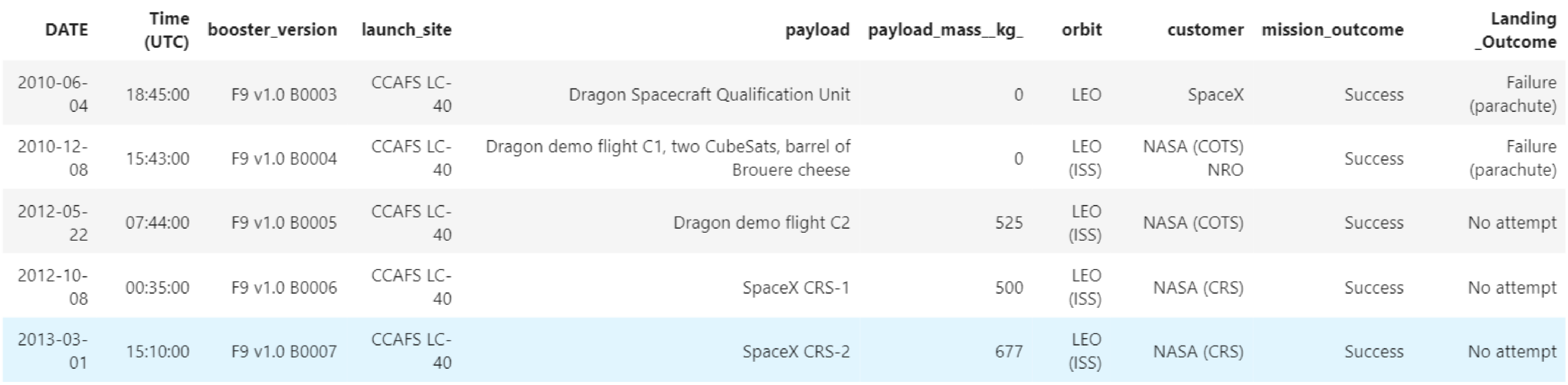
- SQL query: **select** DISTINCT(launch_site) **from** SPACEXDATASET
- There are 4 launch sites in total. Namely, CCAFS LC-40, CCAFS SLC-40, KSC LC 39A, and VAFB SLC-4E.



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

Launch Site Names Begin with 'CCA'

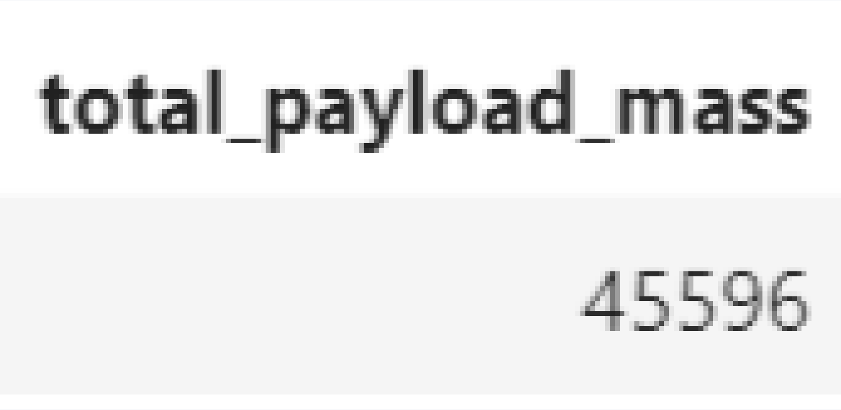
- SQL query: **select** * **from** SPACEXDATASET **where** launch_site **LIKE** 'CCA%' **LIMIT** 5
- The first 5 records where launch sites begin with the string 'CCA‘ were all belong to CCAFS LC-40 launch site.



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

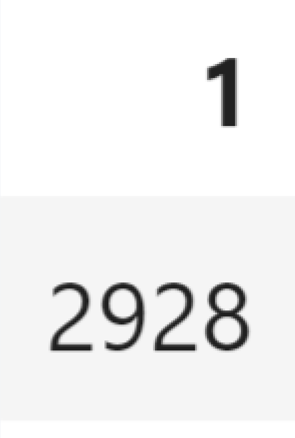
- SQL query:
  - **select** SUM(payload_mass__kg_) AS total_payload_mass **from** SPACEXDATASET **where** customer = 'NASA (CRS)'
- The total payload mass carried by boosters launched by NASA (CRS) is 45596 kg.



total_payload_mass
45596

Average Payload Mass by F9 v1.1

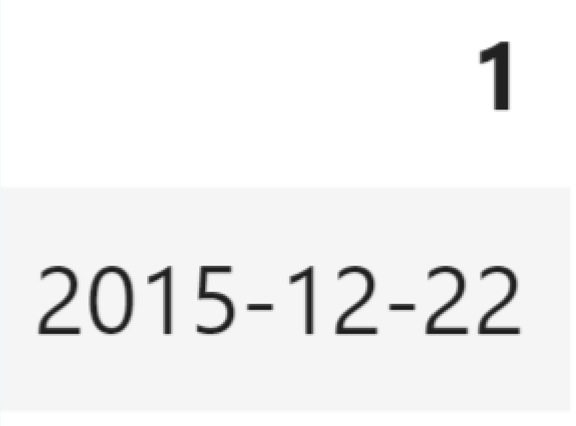
- SQL query:
  - **SELECT** AVG(payload_mass__kg_) **FROM** SPACEXDATASET **WHERE** booster_version = 'F9 v1.1'
- The average payload mass carried by booster version F9 v1.1 is 2928 kg.



1
2928

First Successful Ground Landing Date

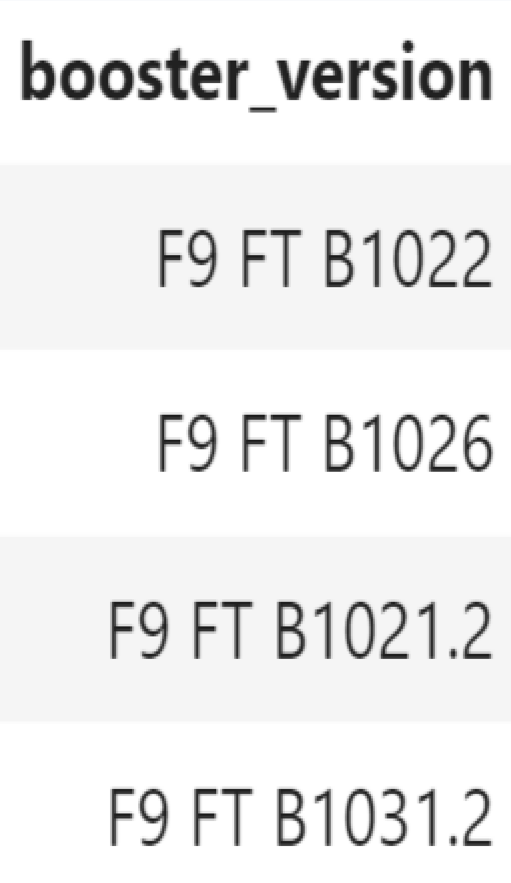
- SQL query: **SELECT** MIN(DATE) **from** SPACEXDATASET **WHERE** LANDING_OUTCOME = 'Success (ground pad)'
- The first successful landing outcome on ground pad was on 2015-12-22



1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

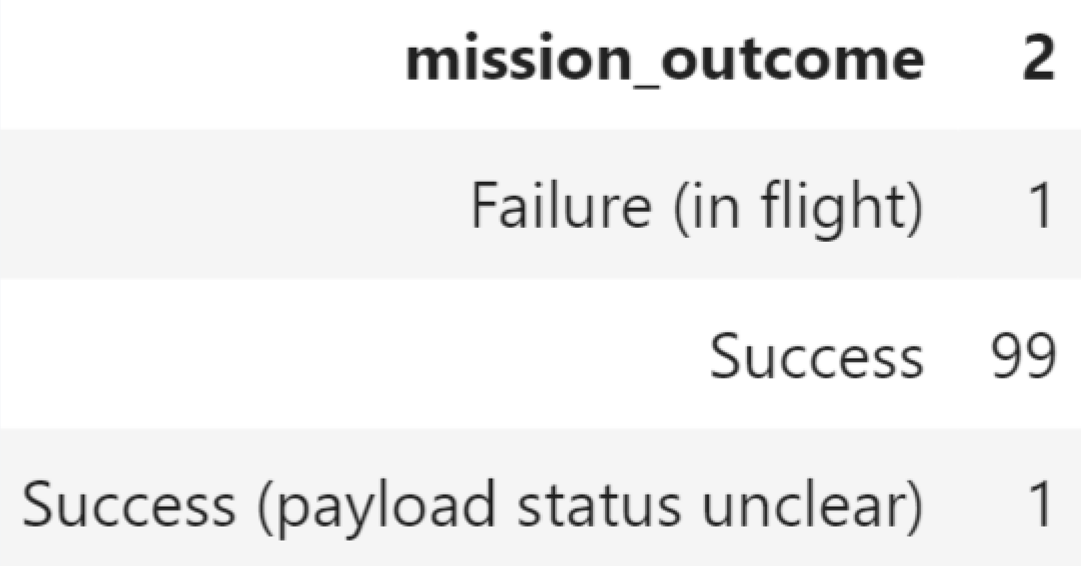
- SQL query: **SELECT** booster_version **from** SPACEXDATASET **where** (LANDING_OUTCOME = 'Success (drone ship)') **and** (payload_mass__kg_ Between 4000 and 6000)
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2.
- Note: The column "Landing _Outcome" was rename as "Landing_Outcome" when loading table in SQL (DB2)



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

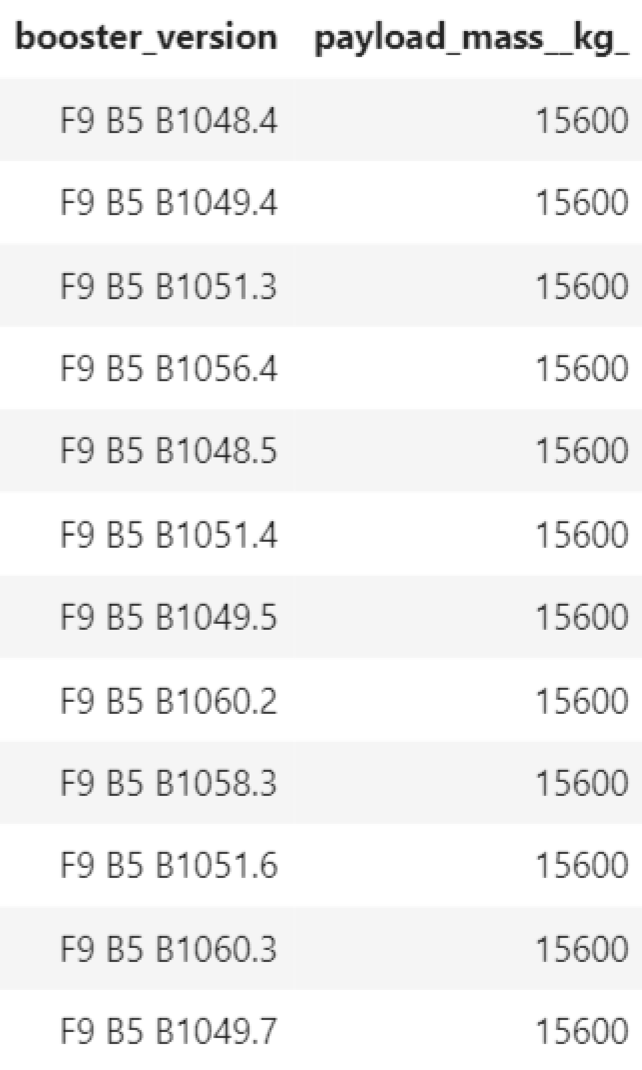
Total Number of Successful and Failure Mission Outcomes

- SQL query: **select** mission_outcome, count(mission_outcome) **from** SPACEXDATASET **GROUP BY** mission_outcome
- The total number of successful mission outcomes is 100.
- There is only 1 failure mission.



mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

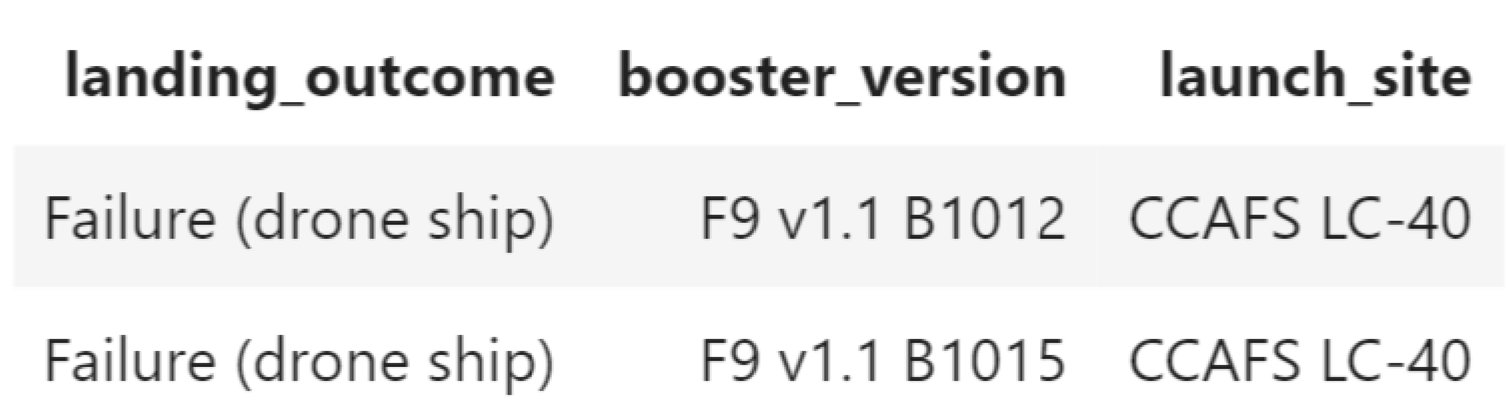


booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- SQL query: **select** booster_version, payload_mass__kg_ **from** SPACEXDATASET **where** payload_mass__kg_ = (**select** max(payload_mass__kg_) **from** SPACEXDATASET)
- There are 12 boosters which have carried the maximum payload mass. Namely, F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1051.3, F9 B5 B1056.4, F9 B5 B1048.5, F9 B5B1051.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3, F9 B5 B1051.6, F9 B5 B1060.3, F9 B5 B1049.7.

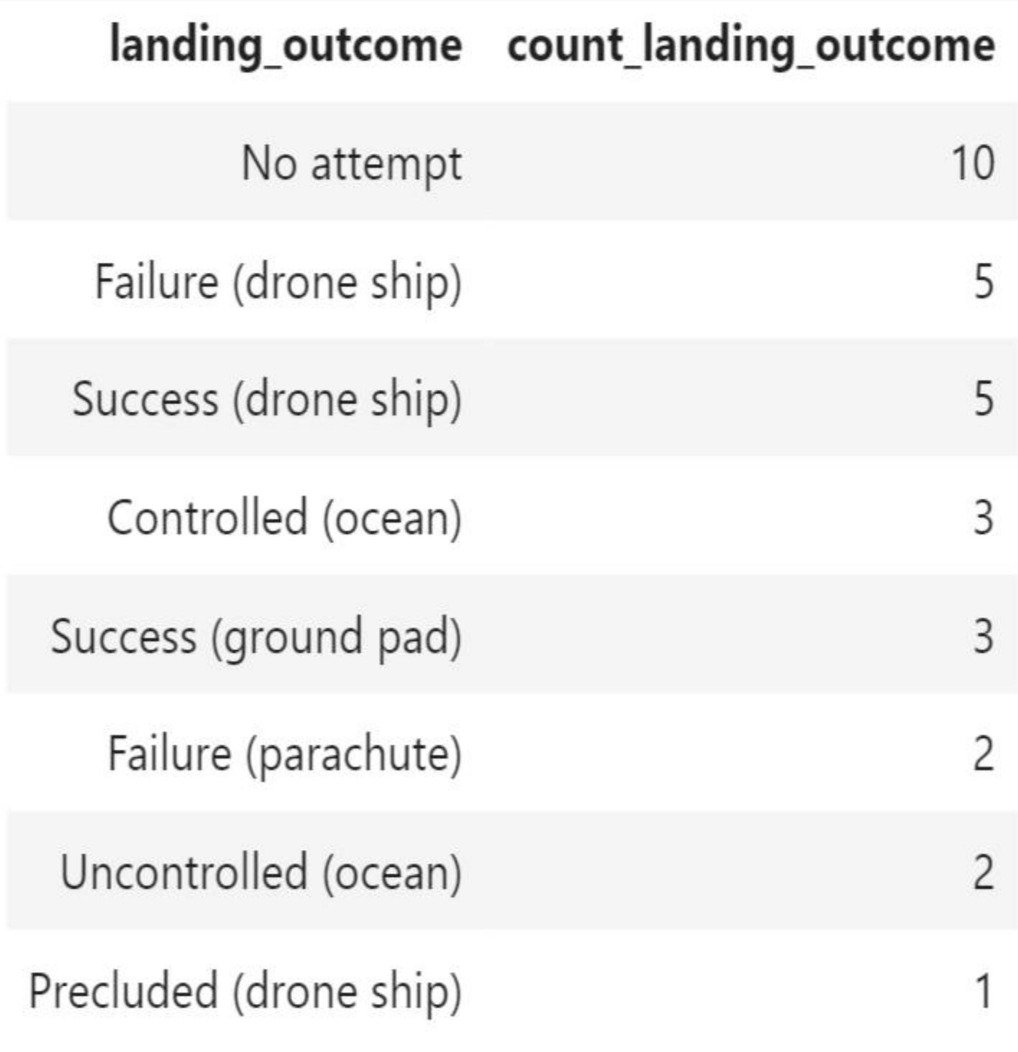
2015 Launch Records

- SQL query: **select** landing_outcome, booster_version, launch_site **from** SPACEXDATASET **where** (LANDING_OUTCOME = 'Failure (drone ship)') **and** YEAR(DATE) = 2015)
- There are 2 failed landing outcomes in drone ship in 2015 in CCAFS LC-40 launch site. The booster version of these two missions are F9 v1.1 B1012 and F9 v1.1 B1015.



landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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



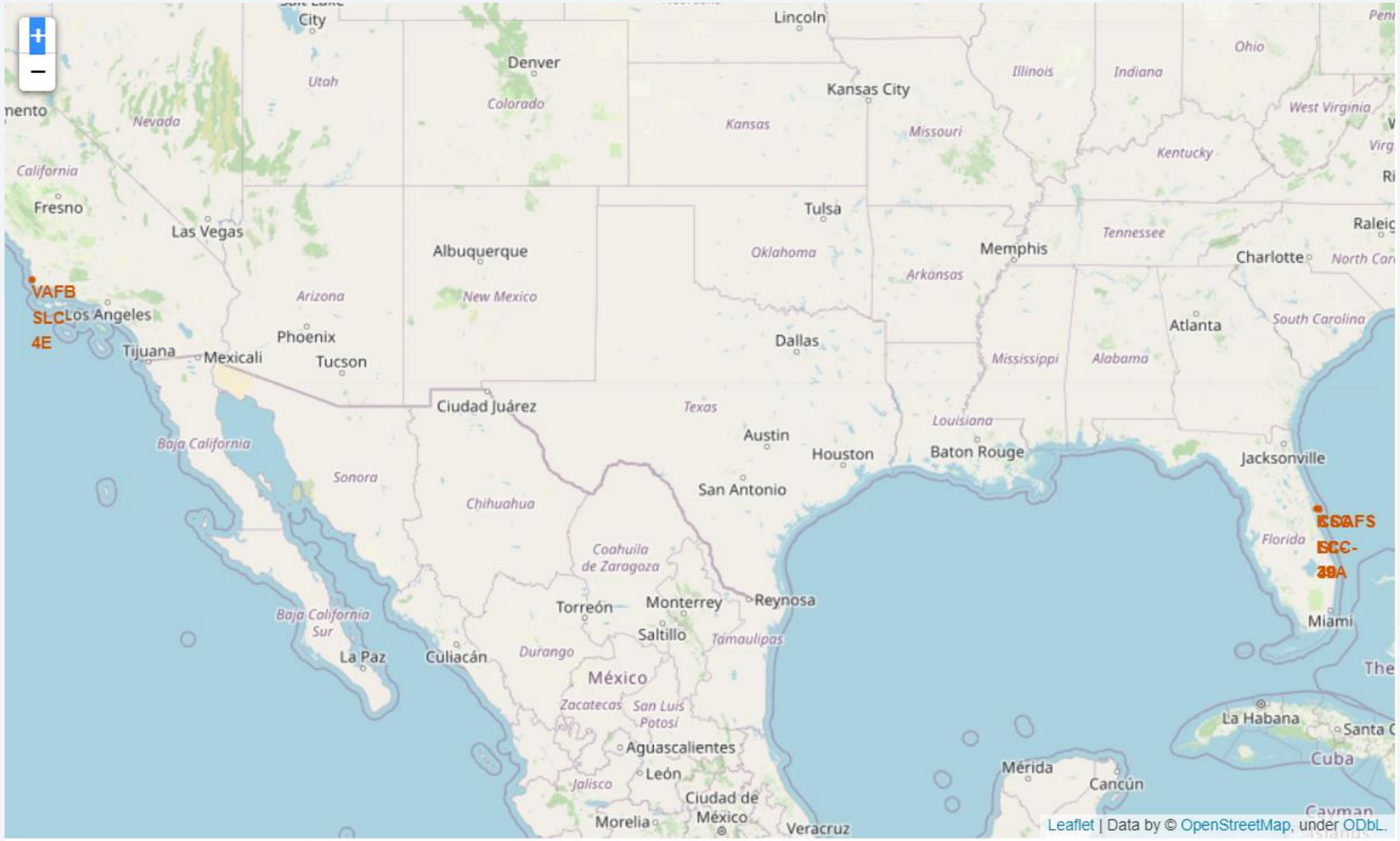
landing_outcome	count_landing_outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- SQL query: **select** landing_outcome,count(landing_outcome) AS count_landing_outcome **from** SPACEXDATASET **WHERE** DATE **BETWEEN** '2010-06-04' AND '2017-03-20' **GROUP BY** landing_outcome **ORDER BY** count(landing_outcome) DESC
- Between 2010-06-04 and 2017-03-20:
  - There are 10 missions with no attempt on landing outcome.
  - There are 5 successes and 5 failure outcome of landing on drone ship.

Section 4

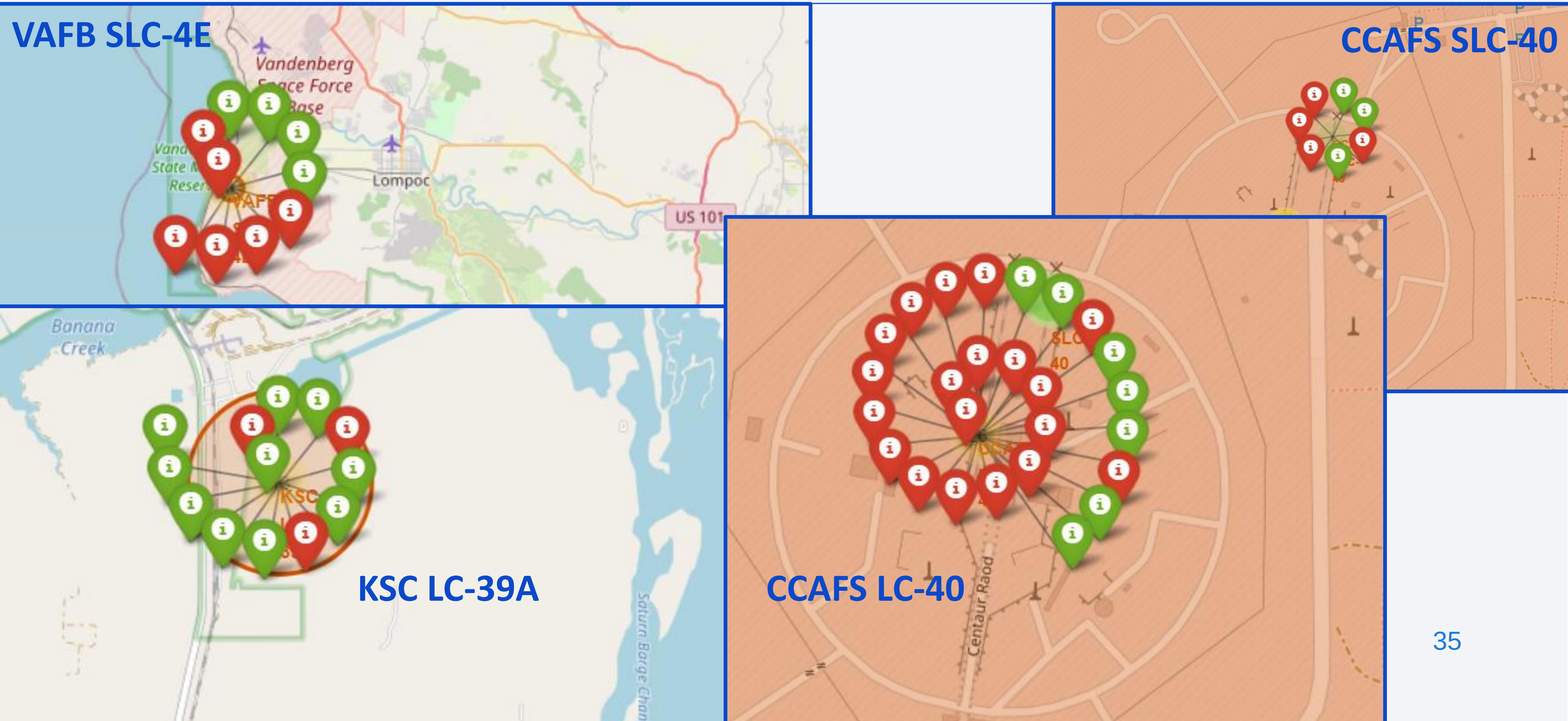
Launch Sites Proximities Analysis

Launch sites’ location on the map



- All launching sites are located in the U.S.
- 3 out of 4 launch sites were closely located together in Florida.

Launch outcomes on each site



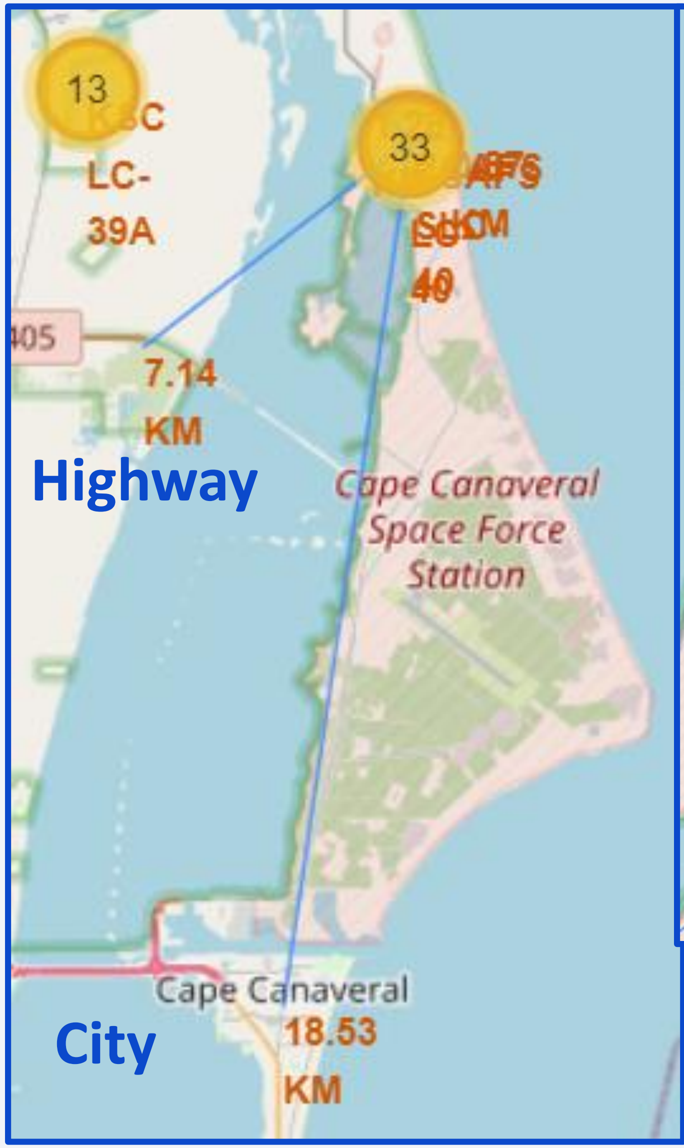
VAFB SLC-4E

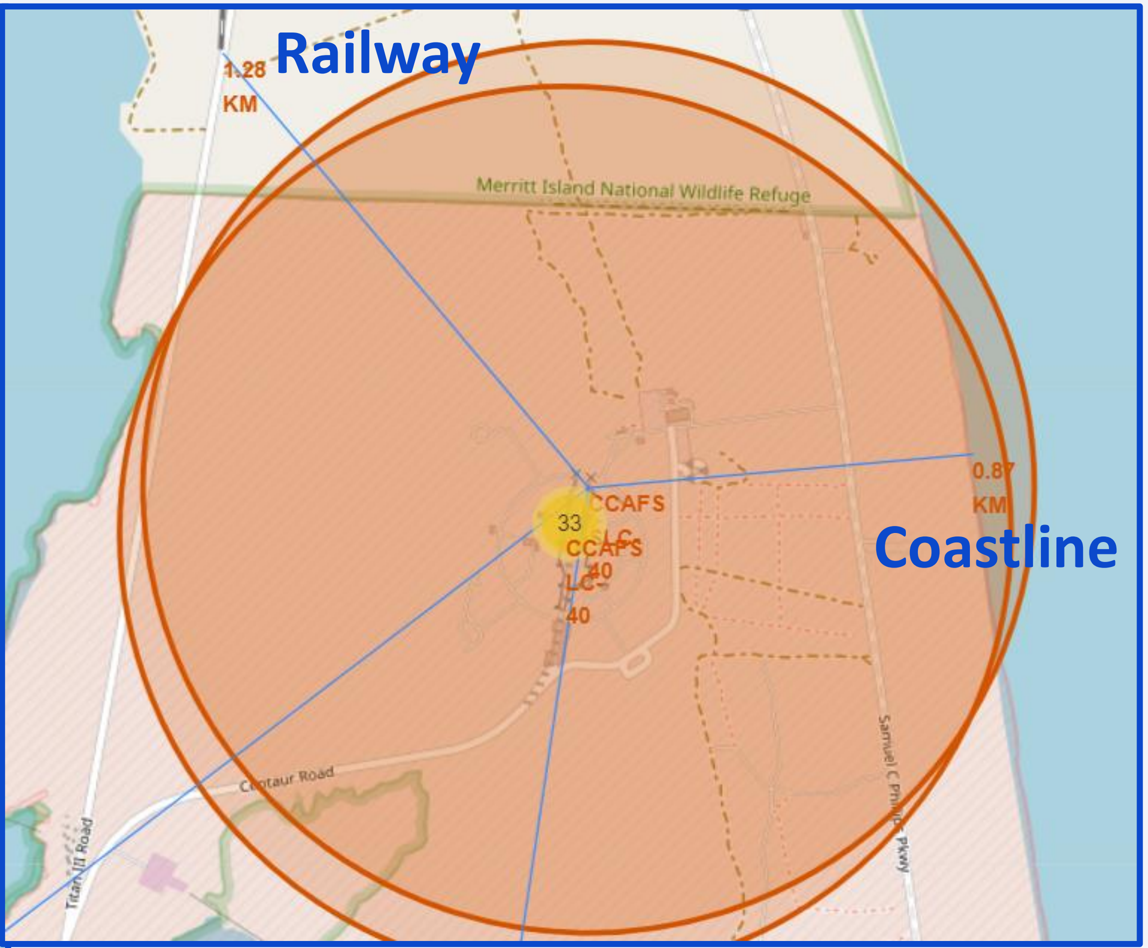
CCAFS SLC-40

KSC LC-39A

CCAFS LC-40

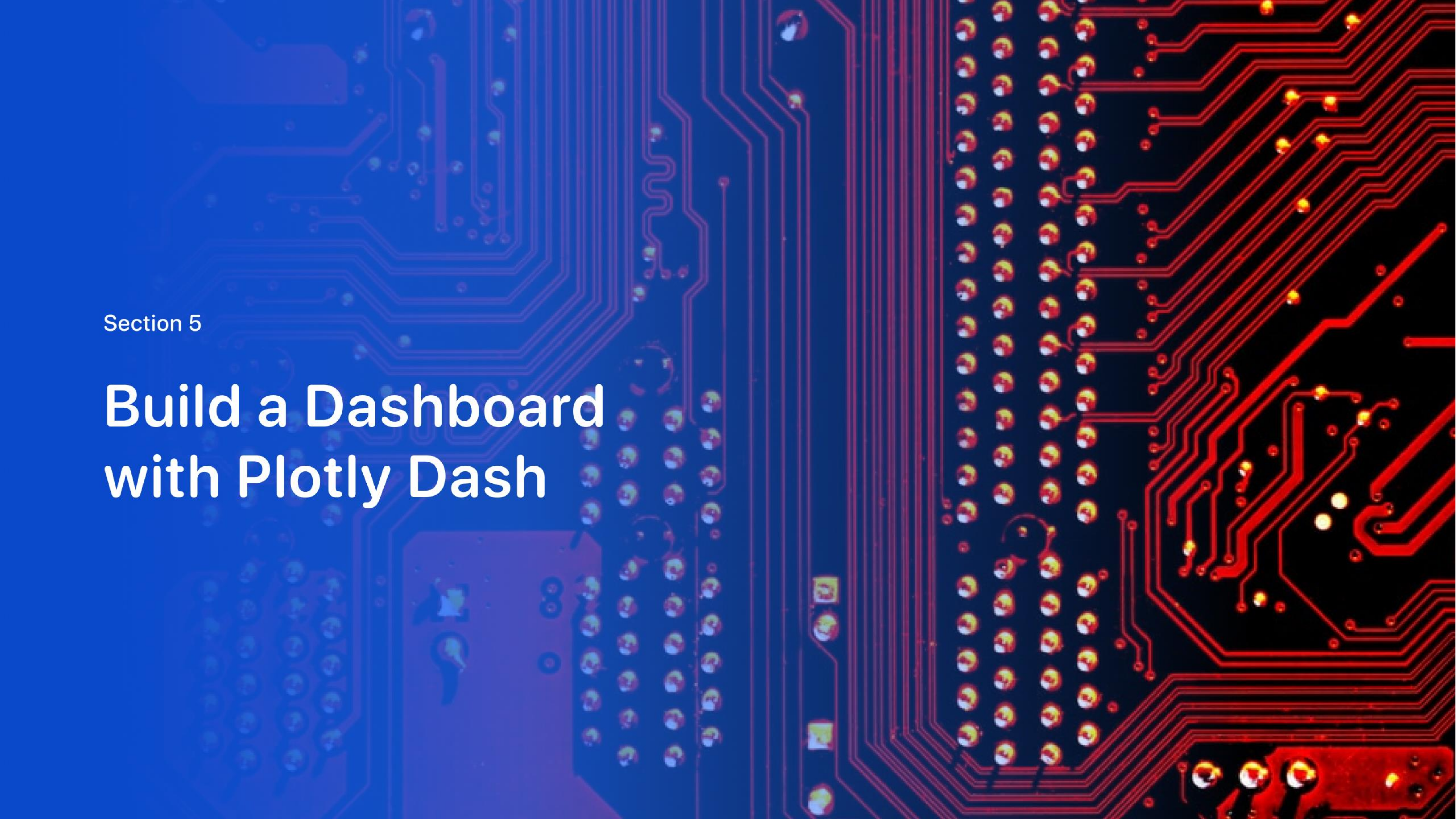
Distance from CCAFS SLC-40 to its proximities





- CCAFS SLC-40 is in close proximity to the nearest coastline.

- CCAFS SLC-40 keeps certain distance away from the closest railway, highway, and city.

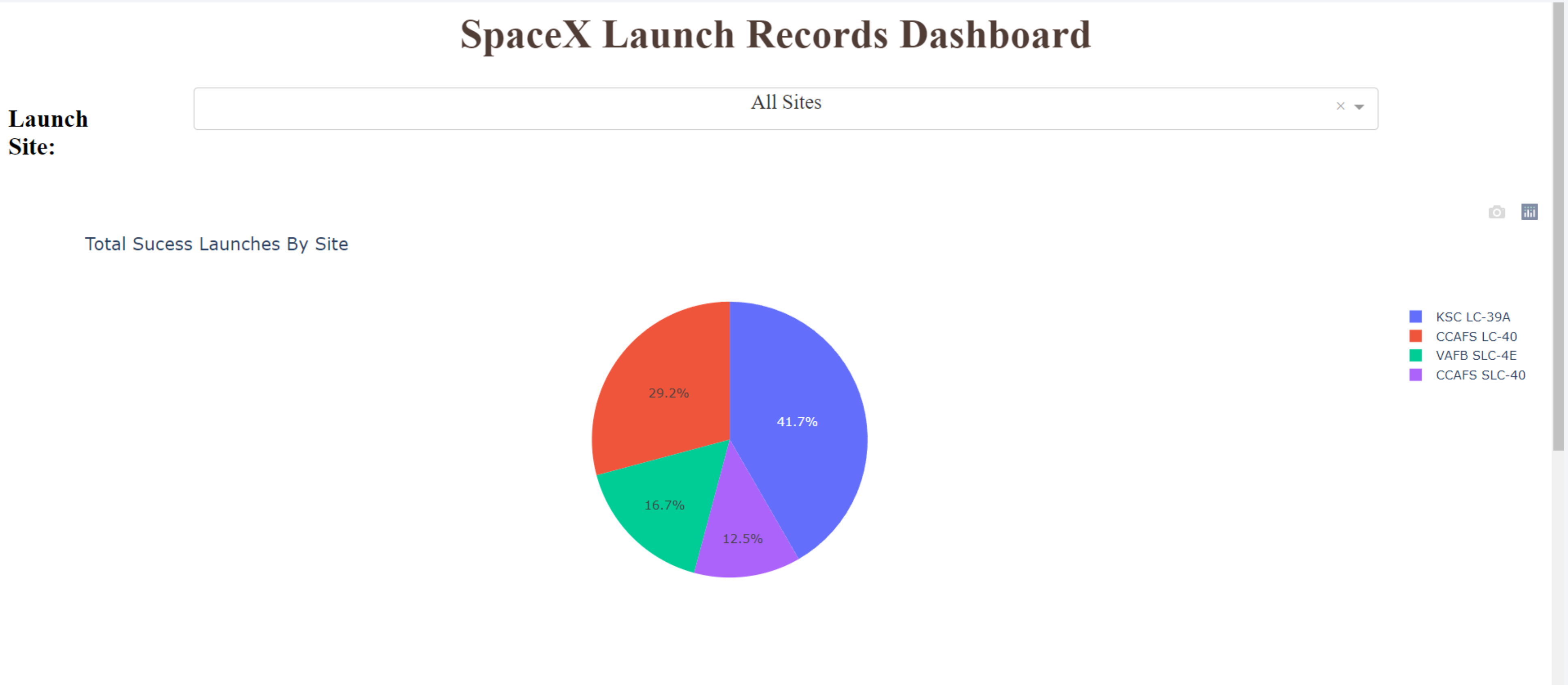


Section 5

Build a Dashboard with Plotly Dash

Total Success Launches By Site

- The pie chart indicates that KSC LC-39A launch site has the highest success flights. CCAFS SLC-40 has the lowest success flights.



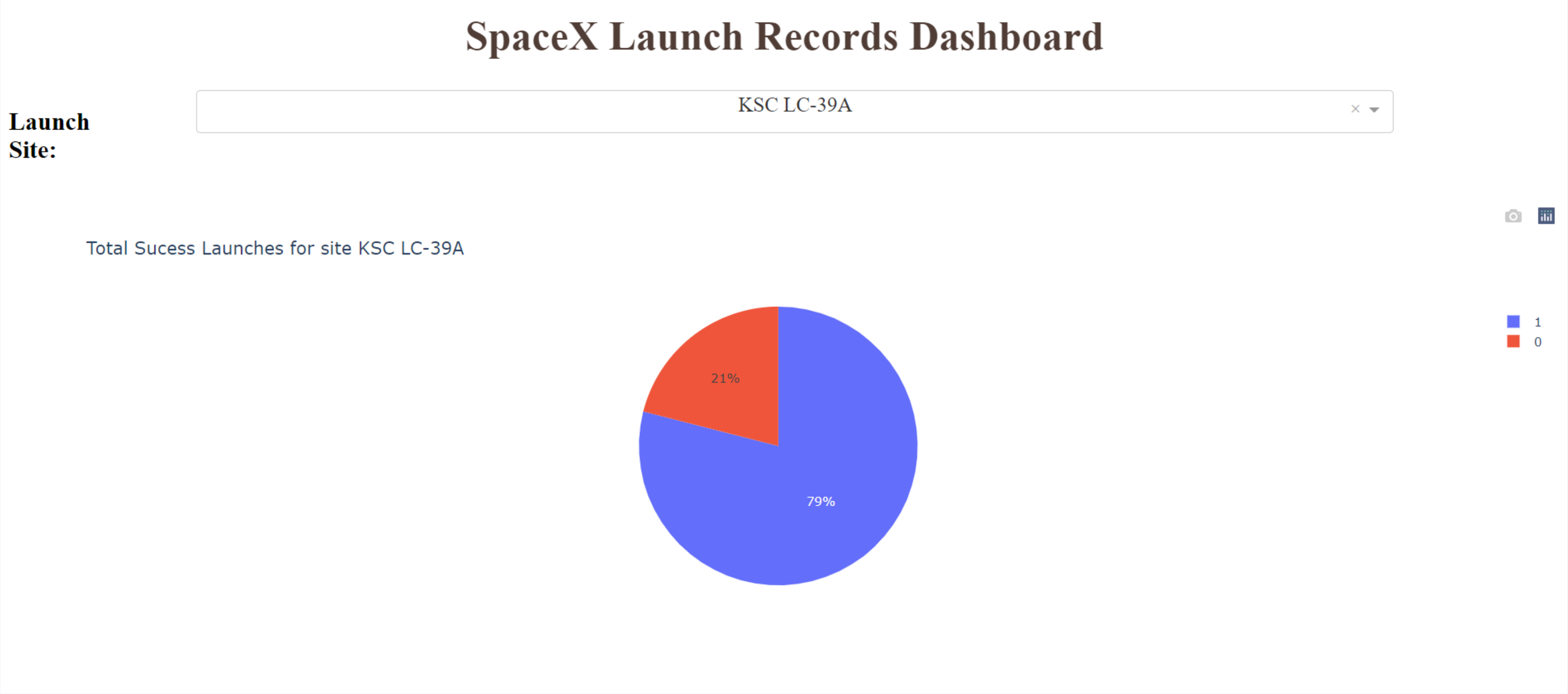
SpaceX Launch Records Dashboard

Launch Site: All Sites

Total Sucess Launches By Site

Total Success Launches for site KSC LC-39A

- KSC LC-39A—the launch site with highest launch success ratio, has a success rate of 79% and a failure rate of 21%.



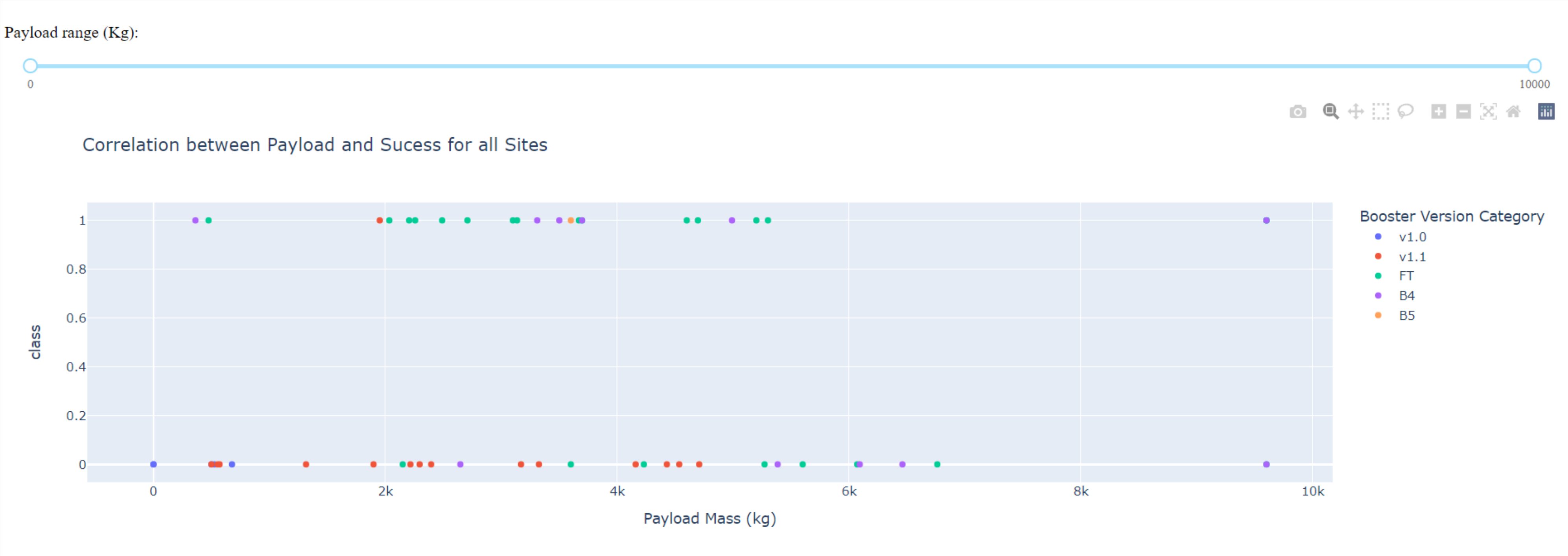
SpaceX Launch Records Dashboard

Launch Site: KSC LC-39A

Total Sucess Launches for site KSC LC-39A

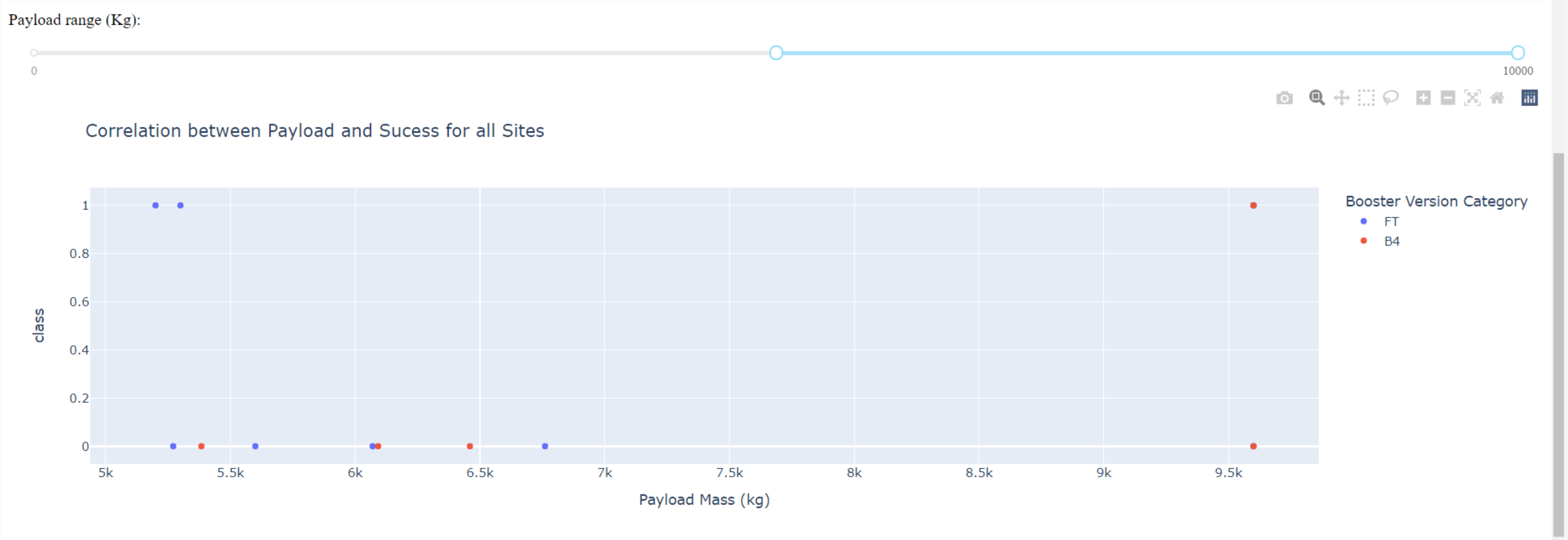
Payload (full range) vs. Launch Outcome

- High payload in all sites are rarely success.
- Flights using v1.1 booster are rarely success regardless of payload.
- Low-to-medium payload flights using FT booster have more success outcomes.



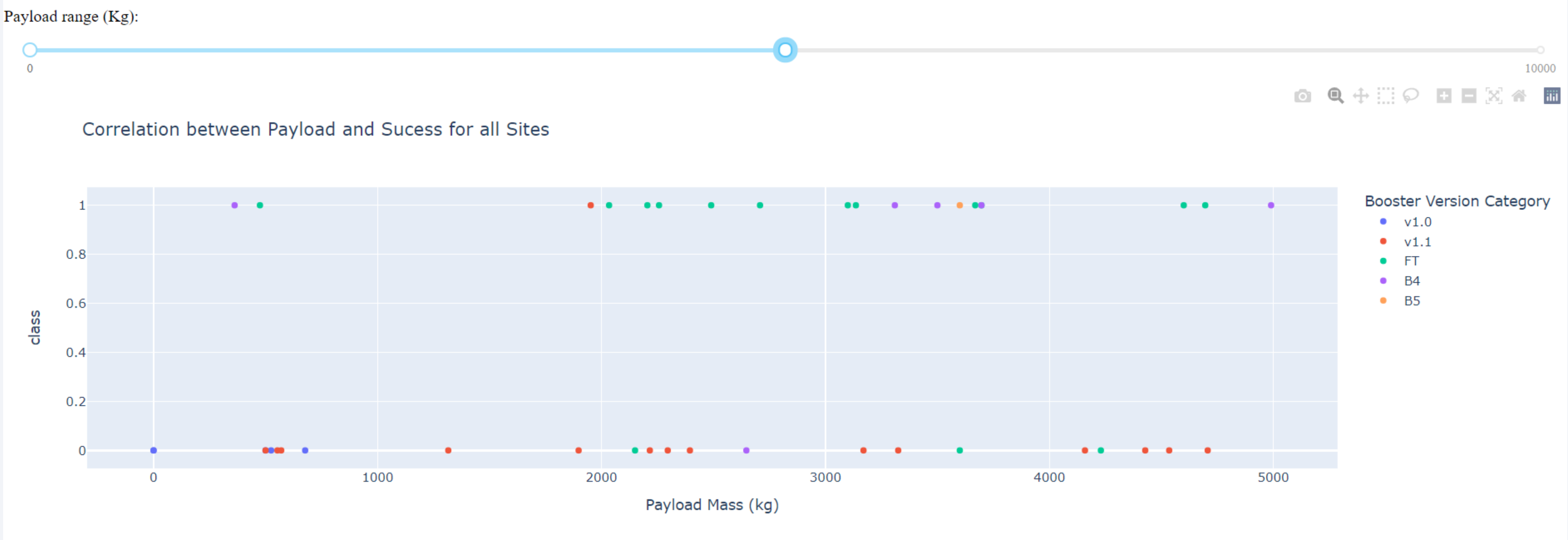
Medium-to-high Payload vs. Launch Outcome

- There are only a few flights with payload above 5000 kg, and only FT and B4 booster were used for these flights.
- The success flight with highest payload used B4 booster.



Low-to-medium Payload vs. Launch Outcome

- Flights with low-to-medium payload using FT and B4 booster are more likely to be success.
- Flights using v1.0 and v1.1 are more likely to be failure.

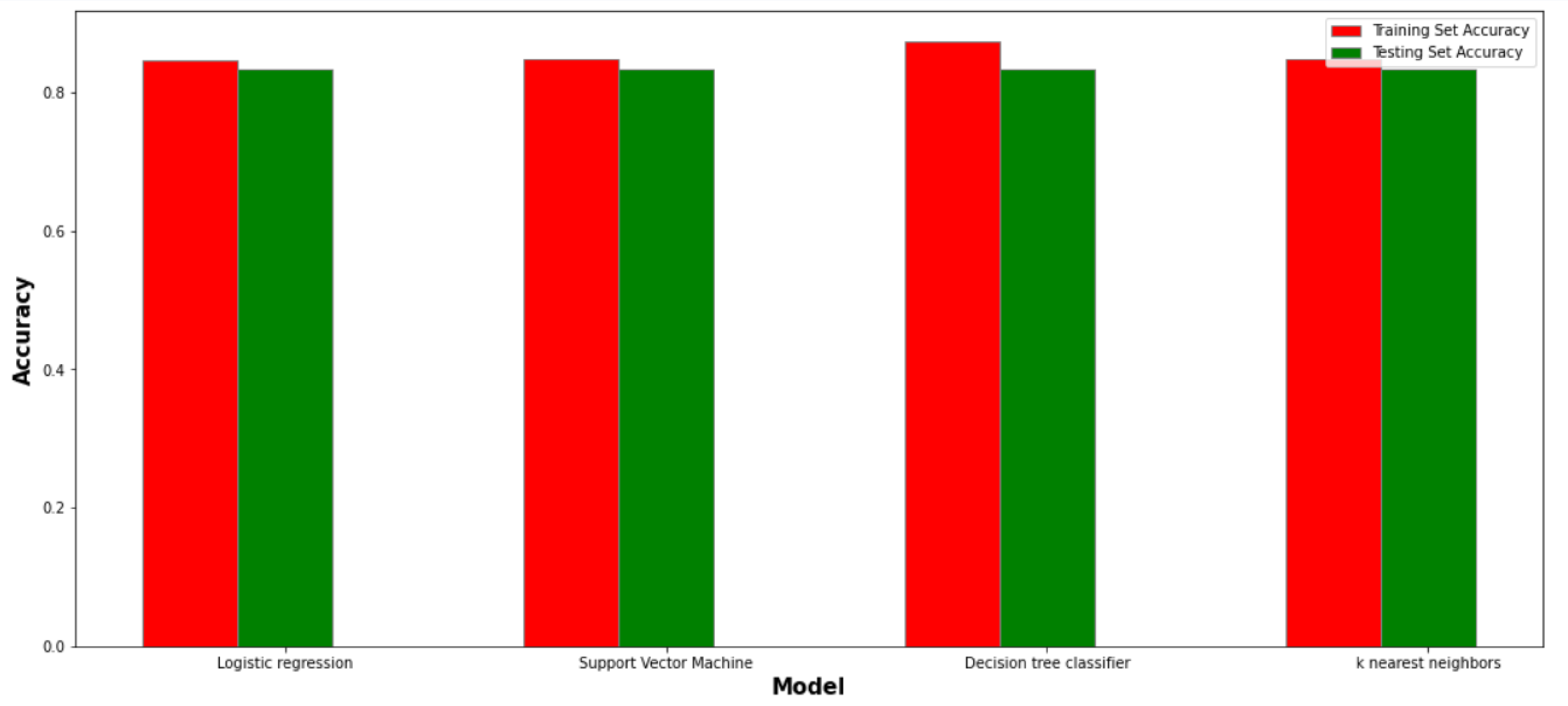


Section 6

Predictive Analysis (Classification)

Classification Accuracy

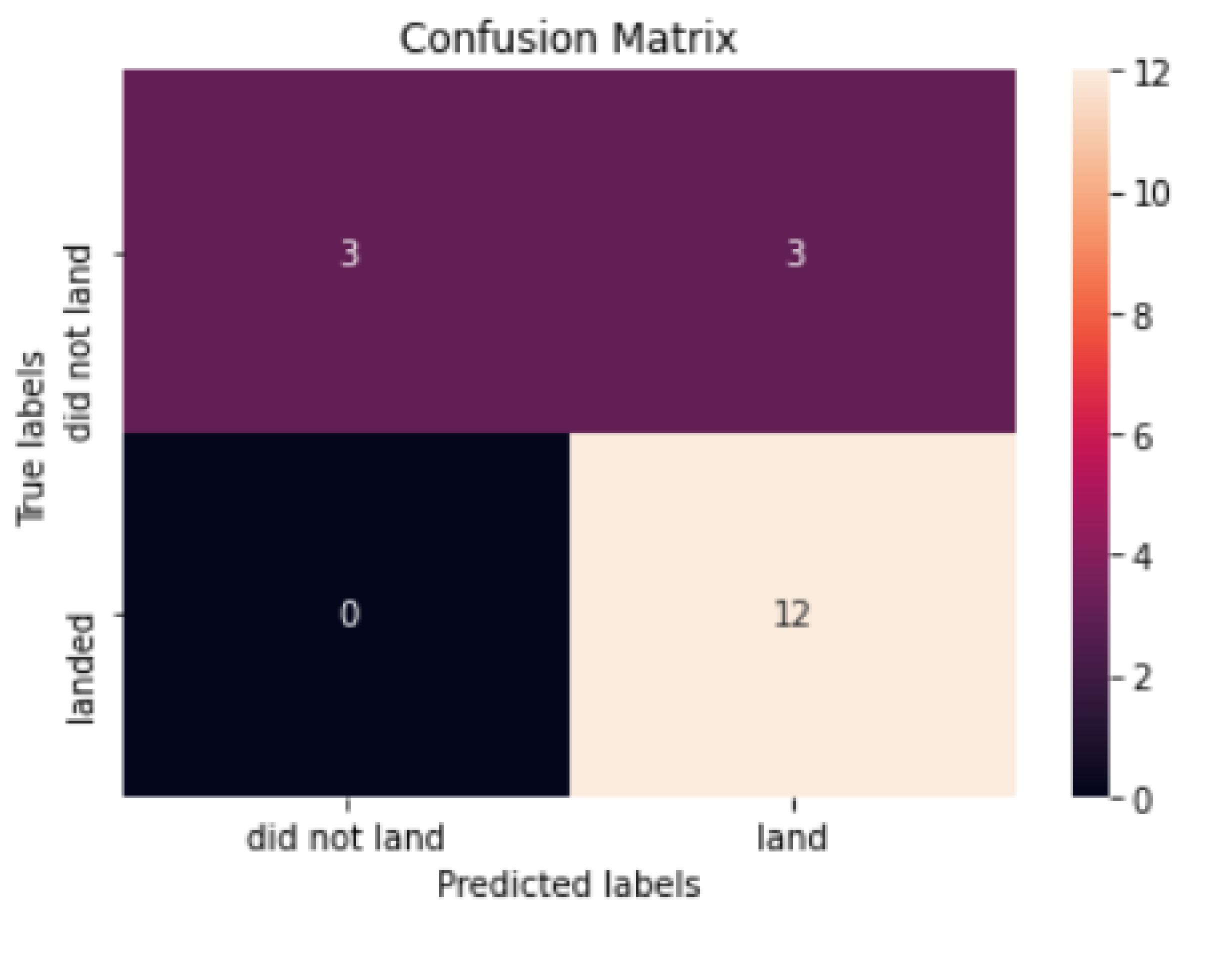
- The bar chart shows training and testing set accuracy for each model.
- All model have the same testing set accuracy (83.33%)



- However, Decision tree classifier have the highest training set accuracy (87.50%) among all algorithms.

Confusion Matrix

- Decision tree classifier’s confusion matrix on testing set shows that the model could correctly predict 15 out of 18 cases (83.33%).
- The 3 incorrect predictions indicate false positives issue of the model.



Conclusions

- Starting from 2014, the launching success rate increase over years.
- ES-L1, GEO, HEO, SSO orbit types have highest success rate.
- KSC LC-39A launch site has highest success flights with 79% success rate.
- The usage of FT booster is associated with success outcomes, especially in low-to-medium payload flights.
- Decision tree classifier algorithms is shown as the best model to predict launch outcomes.

Appendix

- Link to GitHub repository: Here
    - Link to Data collection API: Here
    - Link to Data collection Web Scrapping: Here
    - Link to Data Wrangling: Here
    - Link to EDA with Data visualization: Here
    - Link to EDA with SQL: Here
    - Link to Interactive Visual Analytics with Folium: Here
    - Link to Interactive Dashboard with Ploty Dash: Here
    - Link to Machine Learning Prediction: Here

Thank you!