IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

P.ATHISPAT
June 19, 2023

Outline

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

Executive Summary

The methodologies begin with collecting the Falcon 9 from 2 sources: SpaceX API data and Wiki page by scraping. Data wrangling was applied to the data by dealing with missing values. Perform exploratory data using many tools including SQL, visualize with graphs, made interactive visual analytics in map using Folium and create dashboard by Dash. Finally, training various machine learning models to predict Falcon 9 rocket landing success or not

According to the results of 4 models including Logistic regression, Support vector machine, Decision tree, and K-nearest neighbor. Decision tree gave the best performance on accuracy 90.35%

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward 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 successfully, we can determine the cost of launch.

Section 1

Methodology

Methodology

Executive Summary

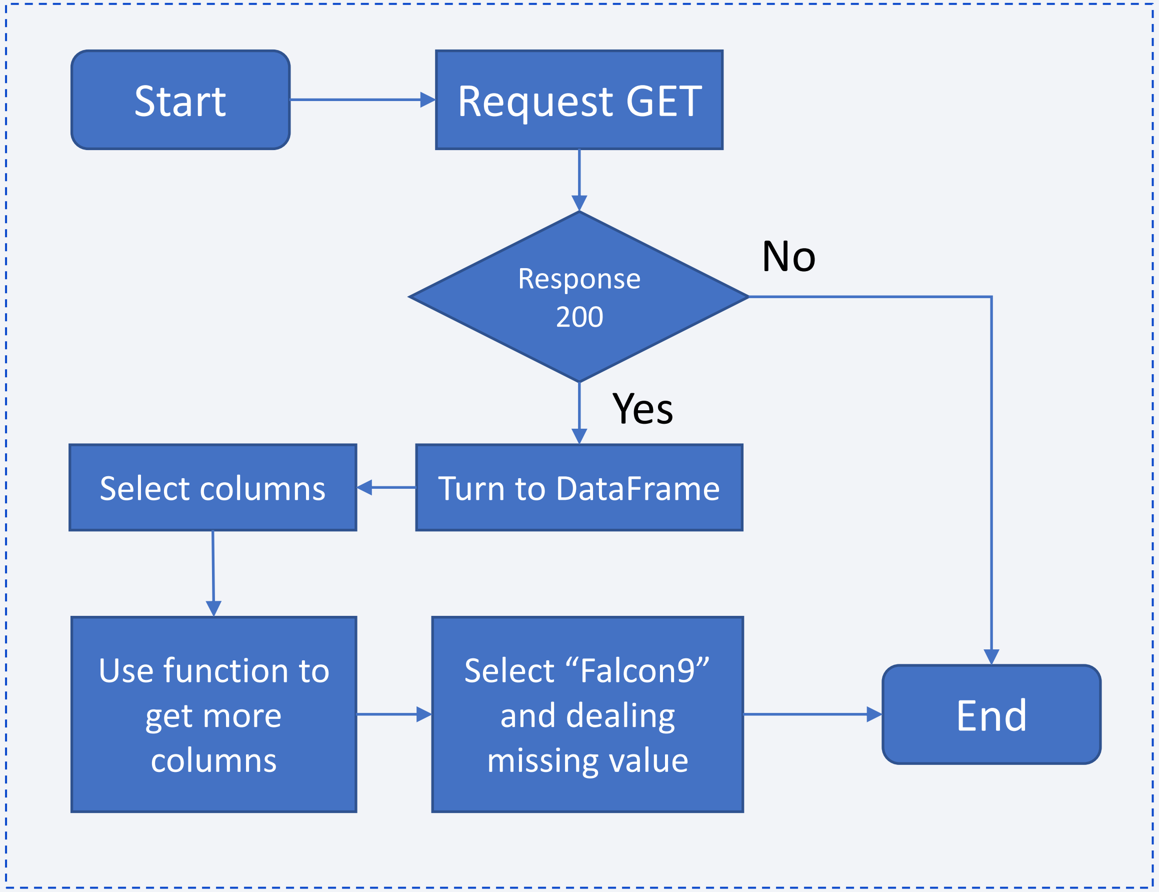
- Data collection methodology:
  - Collect data from SpaceX data API and scraping from Wiki page
- Perform data wrangling
  - Dealing with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize, train test split in preprocess data step
  - GridSearch for tuning in various models and evaluate by confusion matrix

Data Collection

- SpaceX - API
  - Collect data from Space X API including: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LangingPad, lock, ReusedCount, Serial, Longitude, ad Latitude
- Space X Falcon9 – Scraping
  - Collect data from Falcon9 Launch Wiki page from its URL, and then transform to Data frame including: Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time

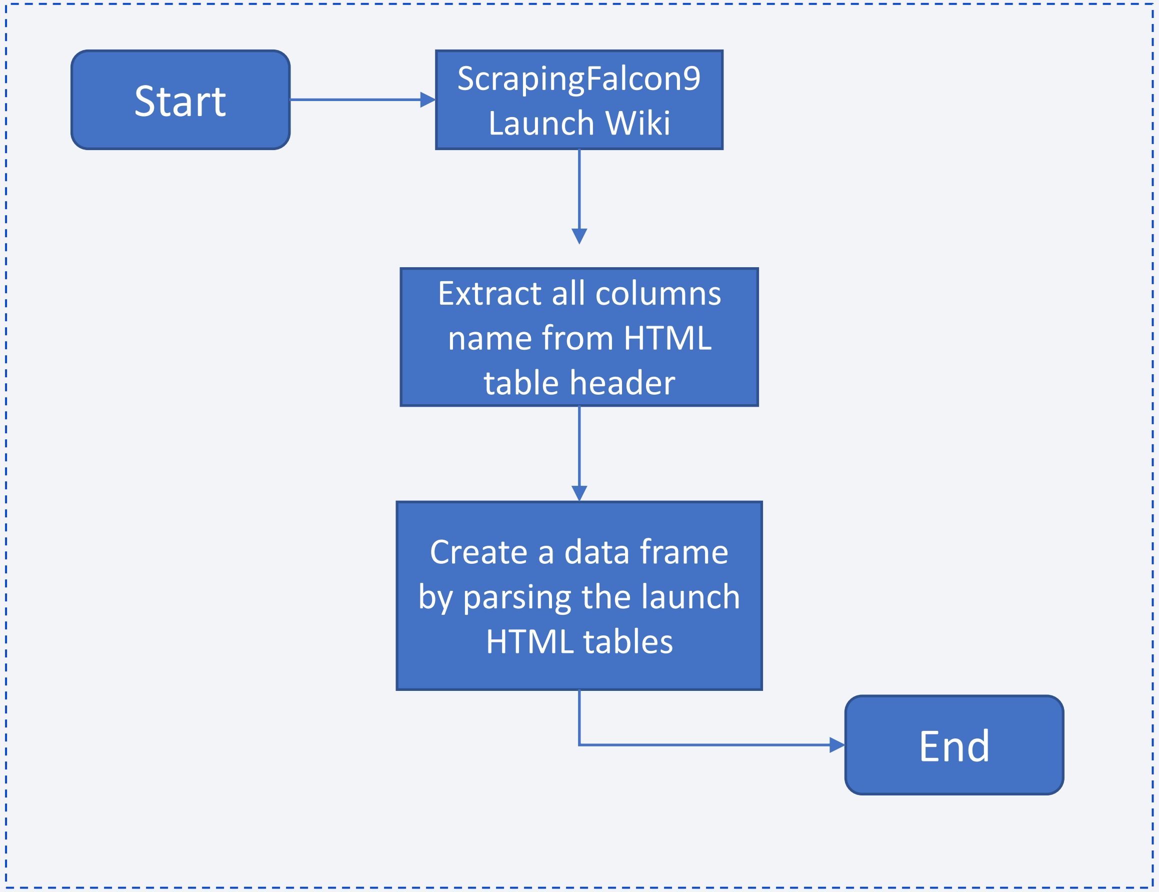
Data Collection – SpaceX API

- Request and parse the SpaceX data using GET request
    - Turn JSON to DataFrame
    - Using columns: rocket, payloads, launchpad, and cores
    - Get BooterVersion, LaunchSite, PayloadData and CoreData and others
- Filter the dataframe to only include “Falcon 9” in BoosterVersion
- Replace missing value with Mean
- GitHub: [spacex-data-collection-api.ipynb]()



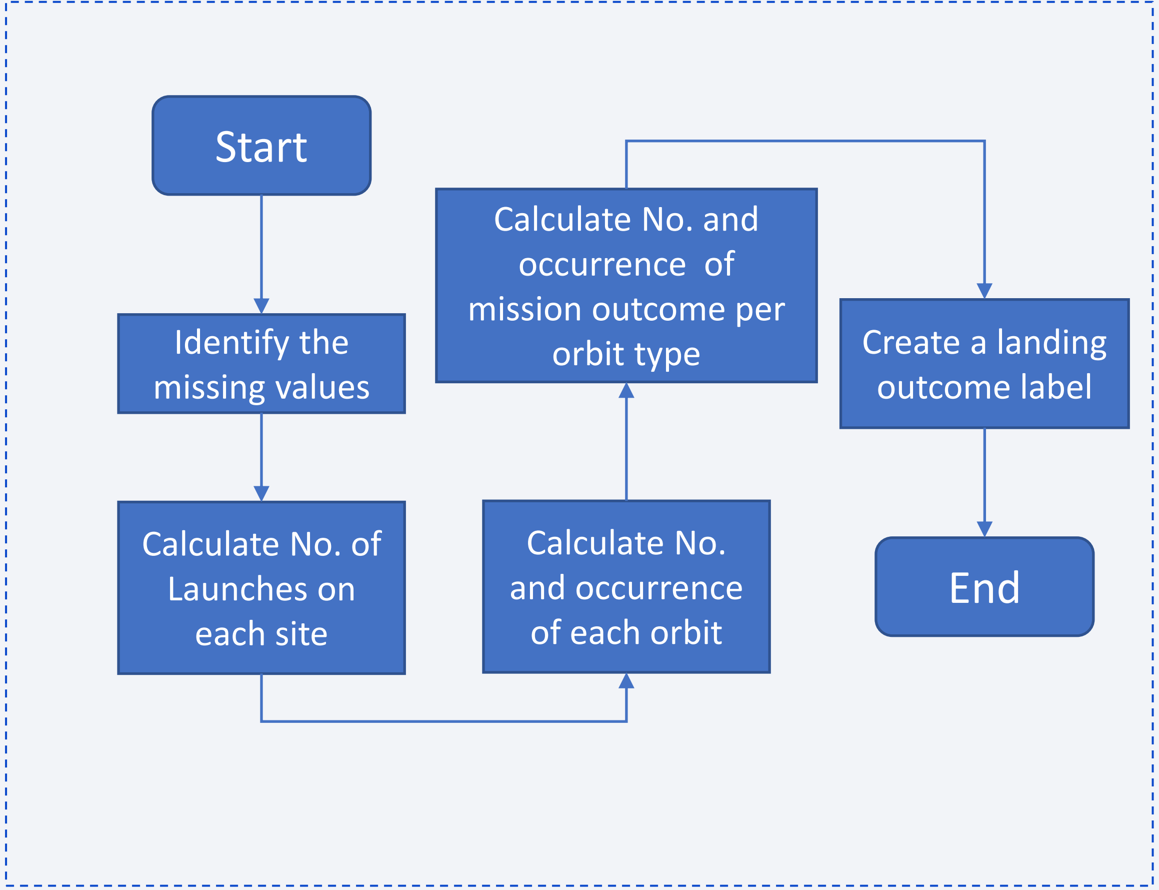
Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from URL
  - Using BeautifulSoup for scraping
  - Extract all columns name from HTML table header
  - Create a data frame by parsing the launch HTML tables
- GitHub: [webscraping.ipynb]



Data Wrangling

- Analyze the data
  - Identify the missing values
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label (class)
- GitHub: [spacex-data_wrangling.ipynb]



EDA with Data Visualization

- Scatter and Line chart
  - Using with numerical to numerical data
  - To observe the trend and relationship between 2 numerical data
- Bar chart
  - Using with categorical to numerical data
  - To compare values among the categorical data
- GitHub: [eda-dataviz.ipynb]

EDA with SQL

- SELECT DISTINCT: display the unique records
- WHERE Launch_Site LIKE 'CCA%': display records where Launch_Site with pattern 'CCA%'
- SUM(PAYLOAD_MASS_KG_): display total payload mass
- AVG (PAYLOAD_MASS_KG_): display average payload mass
- MIN(Date): display minimum date
- PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000: condition mass greater than 4000 but less than 6000
- GROUP BY Mission_Outcome: group the data by Mission_Outcome
- GitHub: [eda-sql-coursera_sqllite.ipynb]

Build an Interactive Map with Folium

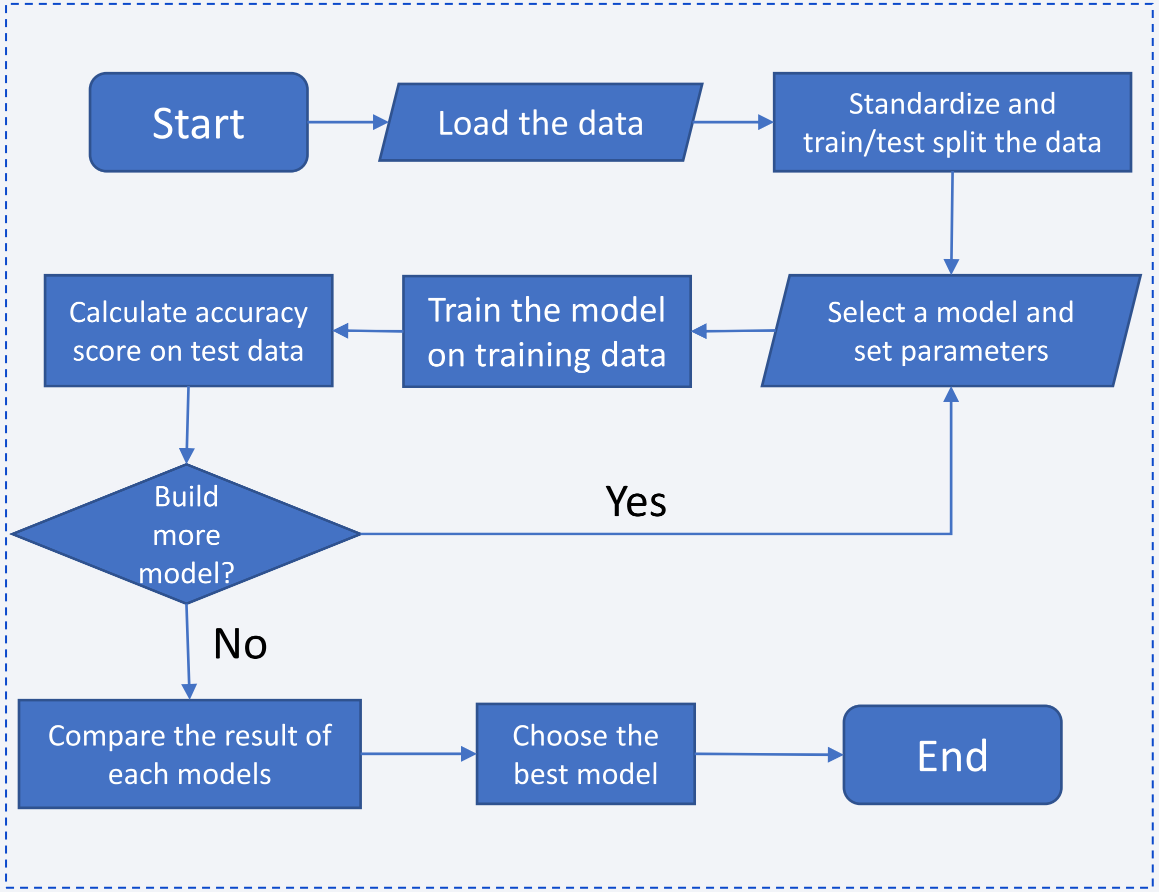
- Map: initial center location of the map
- Marker: Create marker of location on the map
- Circle: Create circle around the desire location
- Lines: Create the line between points on the map
- MarkerCluster: A map containing many markers having the same coordination
- GitHub: launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown: use for getting the launch site type as an input, user can select options provided for displaying various graph style according to the option
- Pie chart: Display the proportions of success of all launch site for option "All Sites" and display the proportions between success/fail of each launch site selected
- Scatter chart: Display compare between Payload Mass (kg) and class
- GitHub: dashboard

Predictive Analysis (Classification)

- Prepare the data by standardize and split the data into train/test set
- Select a model and set parameter of its
- Start train the model on training data
- Calculate the accuracy score on test data
- Decide whether need to train more models or not
- Compare result of all models then choose the best model
- GitHub: SpaceX_Machine_Learning_Prediction.ipynb



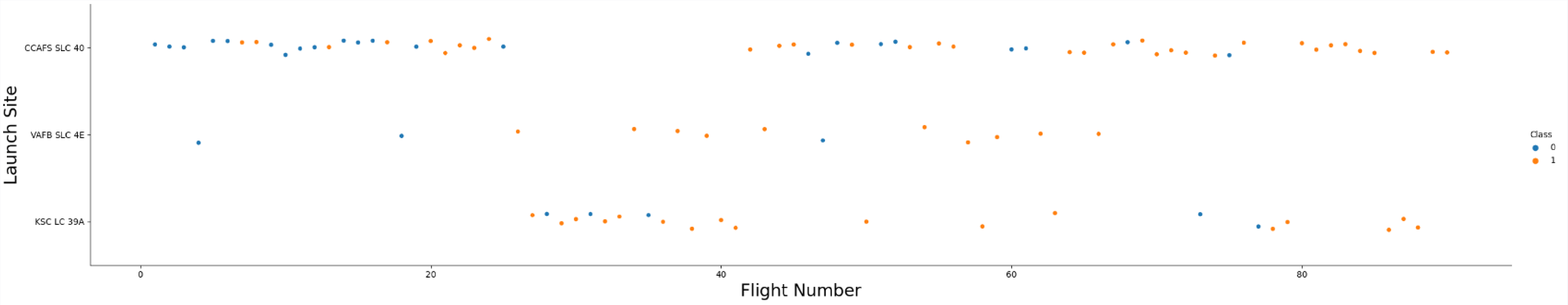
Results

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

Section 2

Insights drawn from EDA

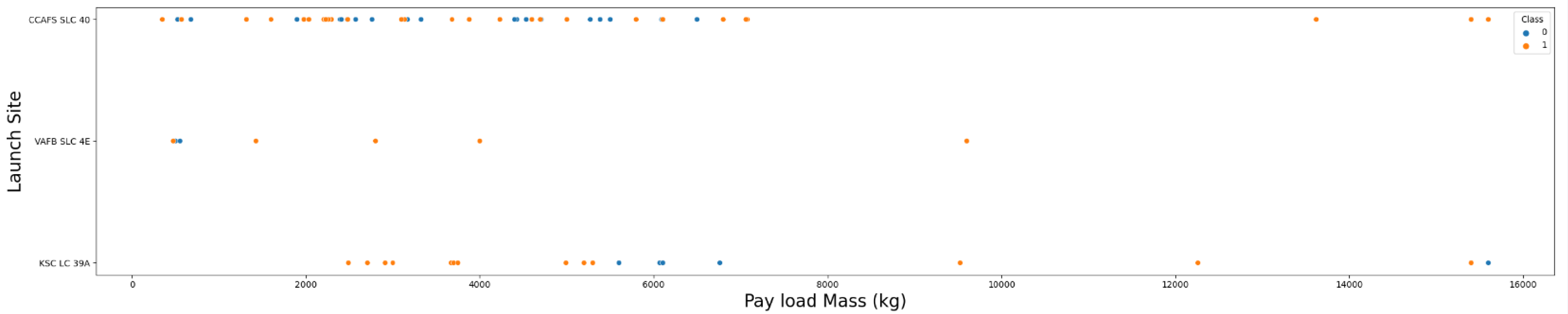
Flight Number vs. Launch Site



The CCAFS SLC 40 has no Flight number from around 25 to 40.

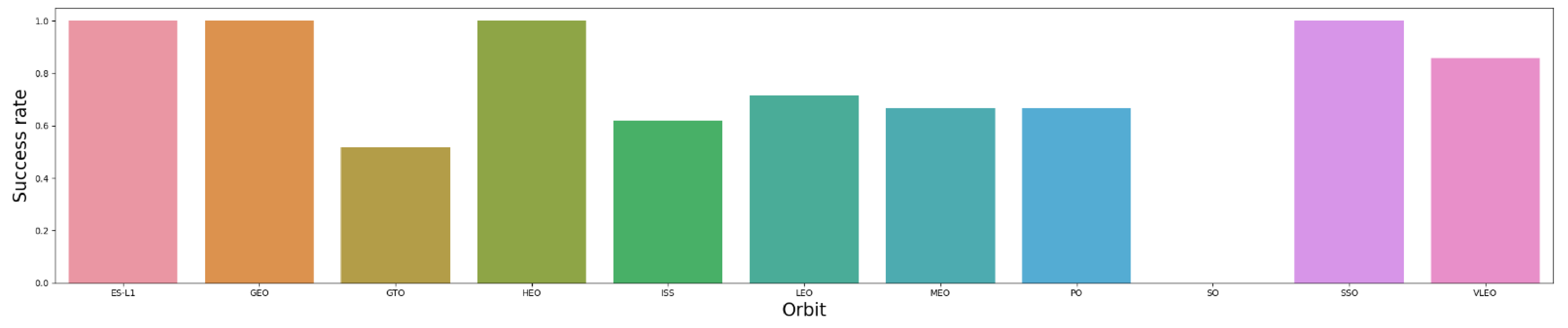
The KSC-LC 39A has no Flight number from 0 to around 20.

Payload vs. Launch Site



The VAFB-SLC is no rockets launched for heavy payload mass(greater than 10000).

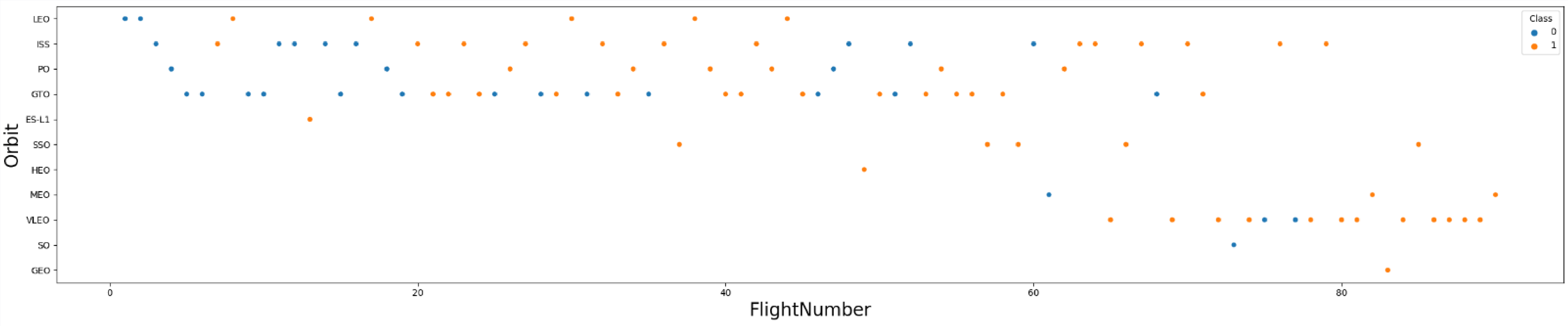
Success Rate vs. Orbit Type



The SO Orbit type has 0 success rate.

The ES-L1, GEO, HEO, and SSO have highest success rate.

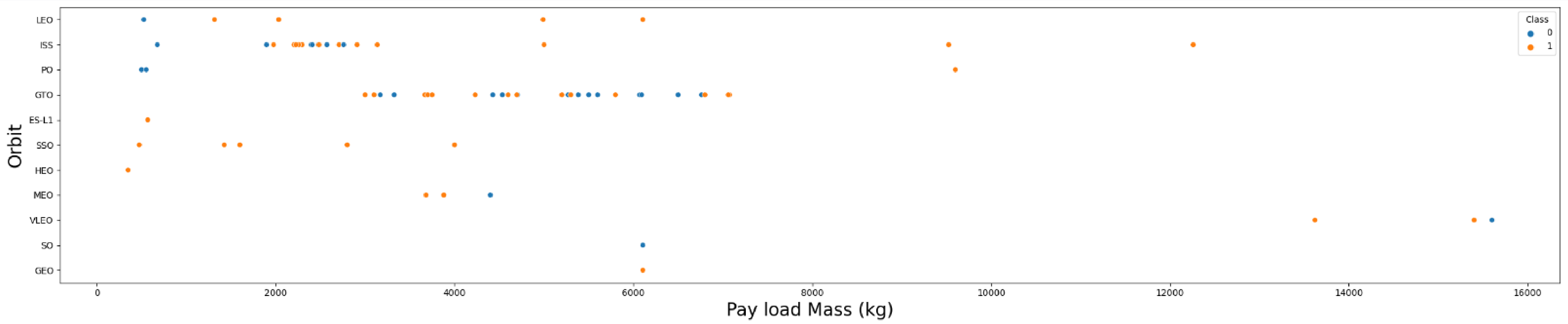
Flight Number vs. Orbit Type



The LEO orbit the Success appears related to the number of flights.

There seems to be no relationship between flight number when in GTO orbit.

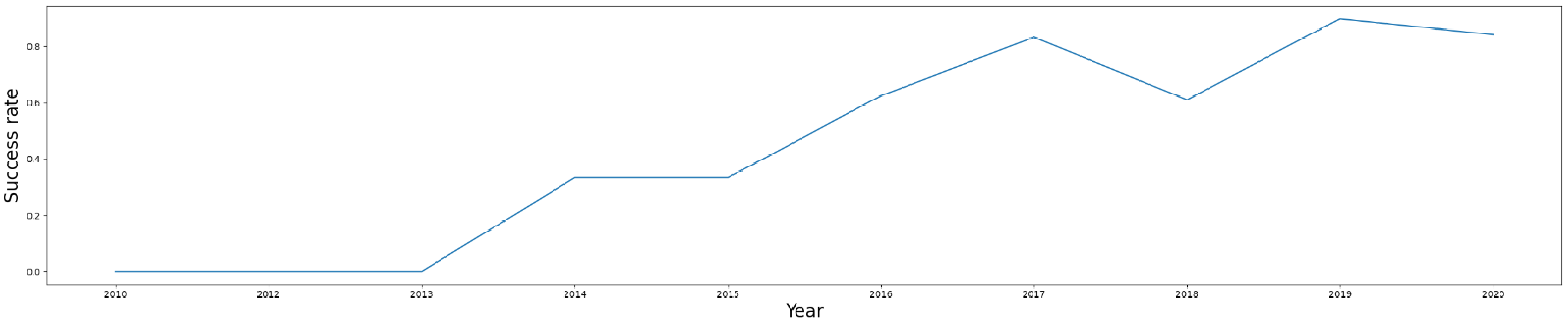
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for LEO and ISS.

GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission).

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

All Launch Site Names

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

There are 4 distinct Launch Site name including:

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site name which begin with ‘CCA’ is CCAFS LC-40

Total Payload Mass

Total payload mass carried by boosters launched by NASA (CRS) is **45596 KG**

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is **August 1, 2018**

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

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

There are 4 names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	COUNT
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful is 100 and failure mission outcomes is 1

Boosters Carried 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

There are list the names of the booster which have carried the maximum payload mass

2015 Launch Records

There are list the failed Landing Outcomes in drone ship, their booster versions, and launch site names for in year 2015

SUBSTR(Date,4,2)	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	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)
Success (ground pad)	7
Success (drone ship)	8
Success	20

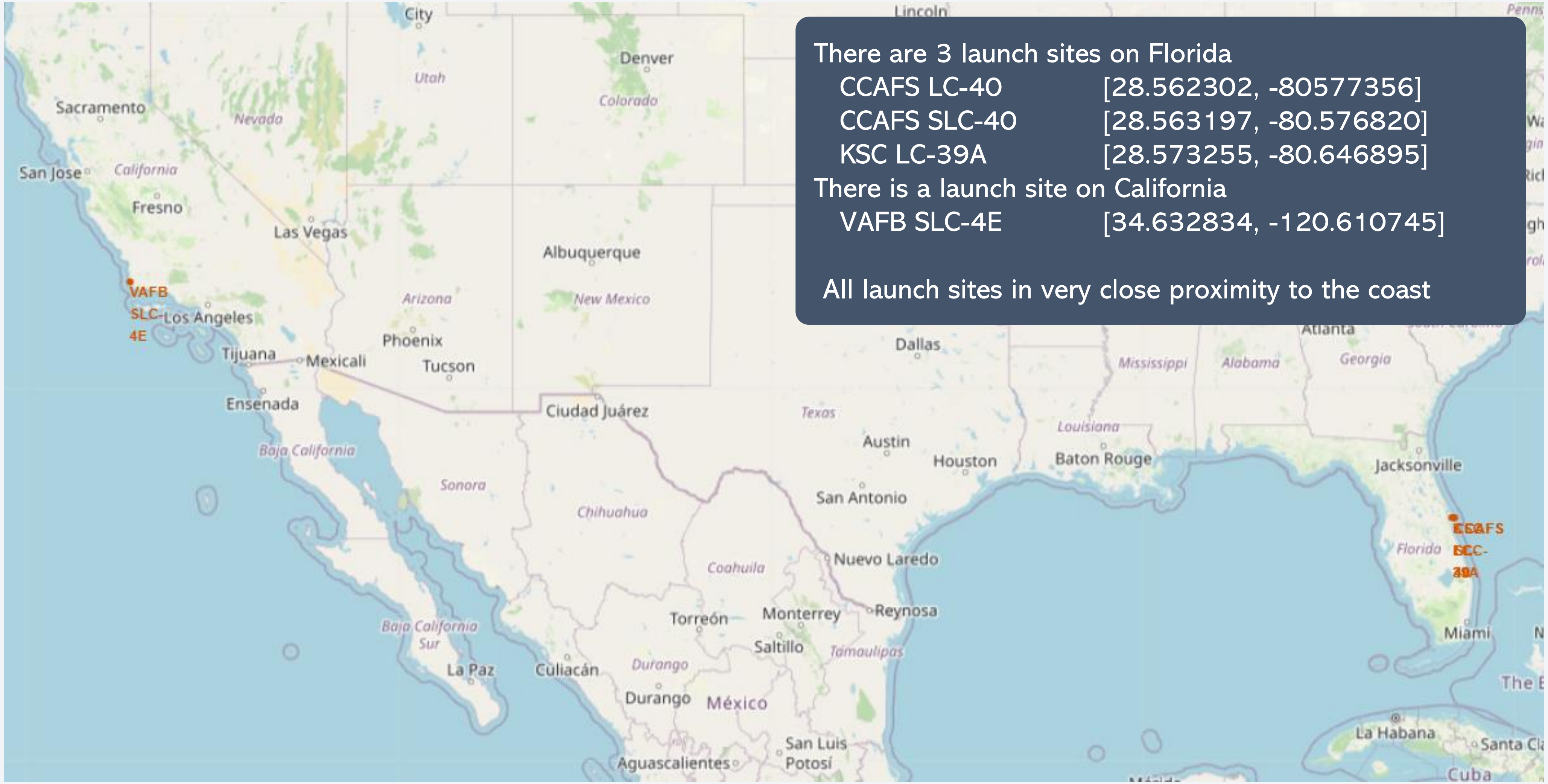
There are the Success Landing Outcomes that ranked by count in descending between the date 2010-06-04 and 2017-03-20

- Success (ground pad) is 7
- Success (drone ship) is 8
- Success is 20

Section 3

Launch Sites Proximities Analysis

All launch sites



There are 3 launch sites on Florida

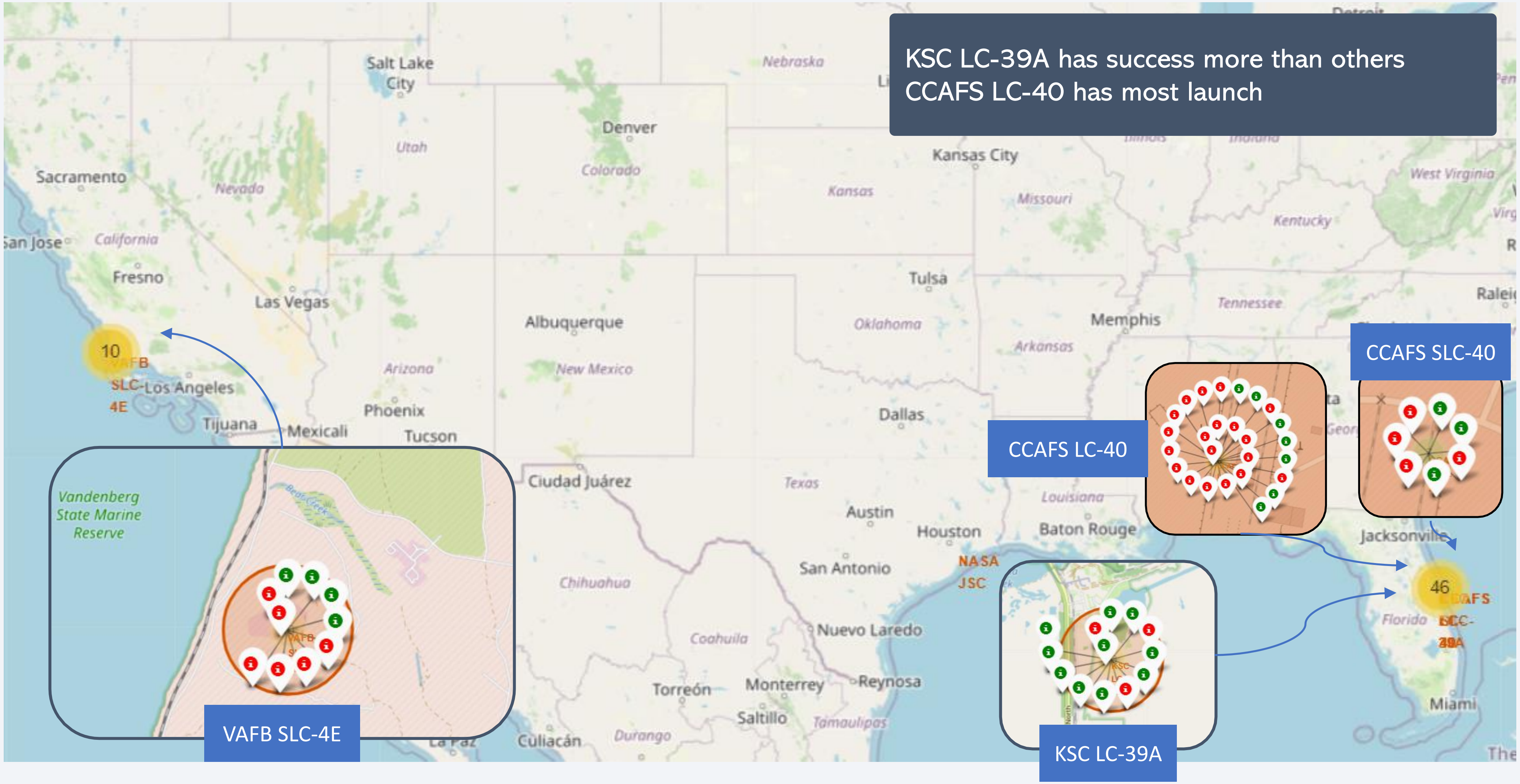
CCAFS LC-40	[28.562302, -80577356]
CCAFS SLC-40	[28.563197, -80.576820]
KSC LC-39A	[28.573255, -80.646895]

There is a launch site on California

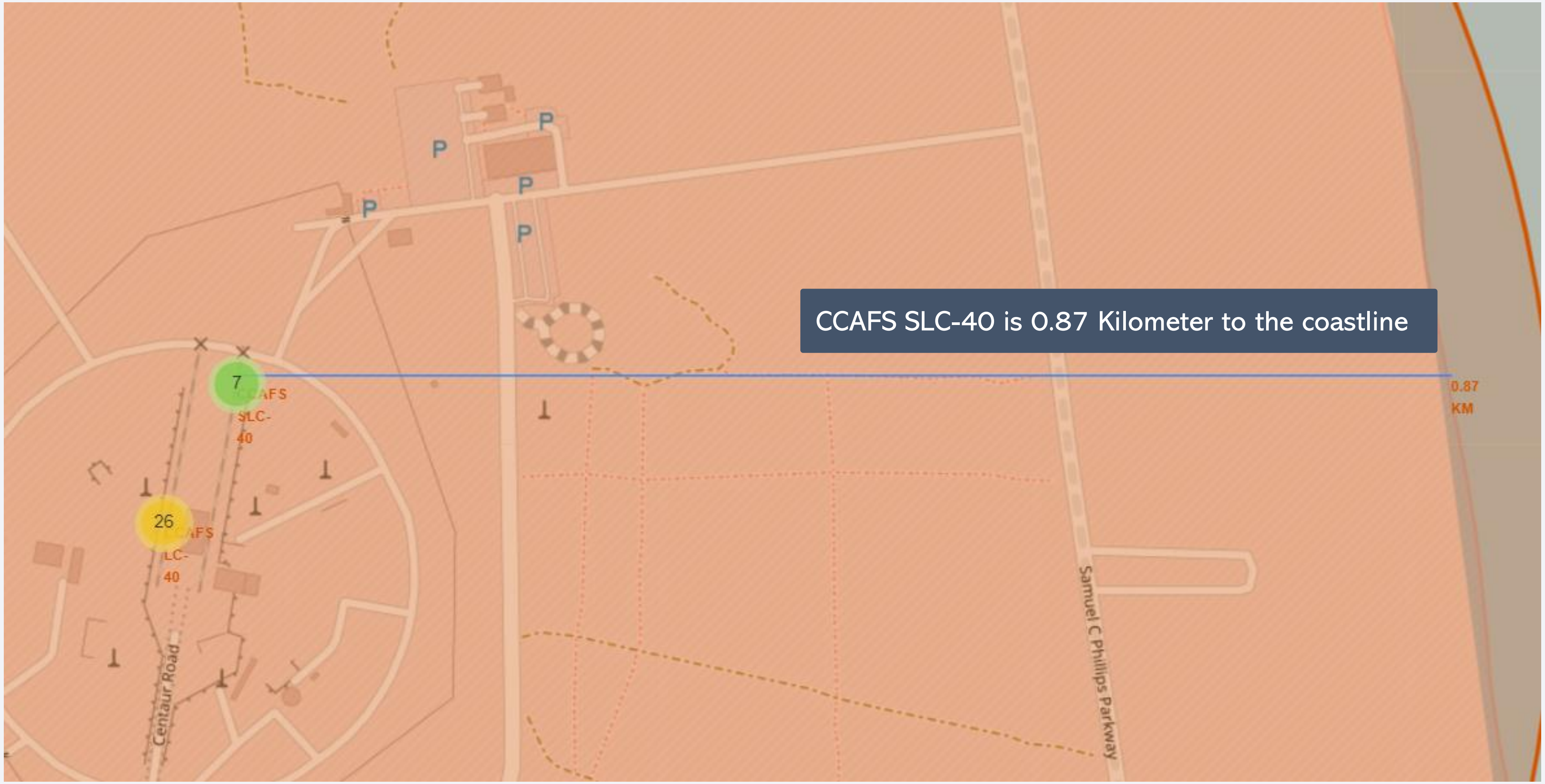
VAFB SLC-4E	[34.632834, -120.610745]

All launch sites in very close proximity to the coast

The success/failed launches for each site



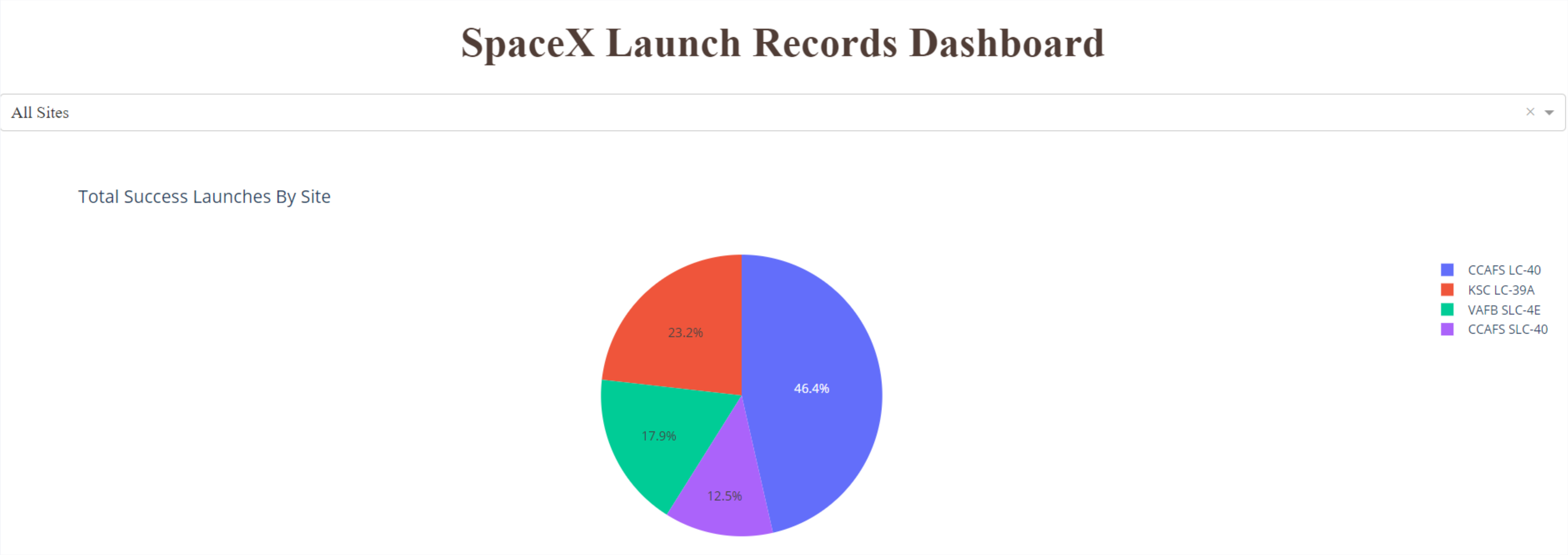
Distances between a launch site to Coast



Section 4

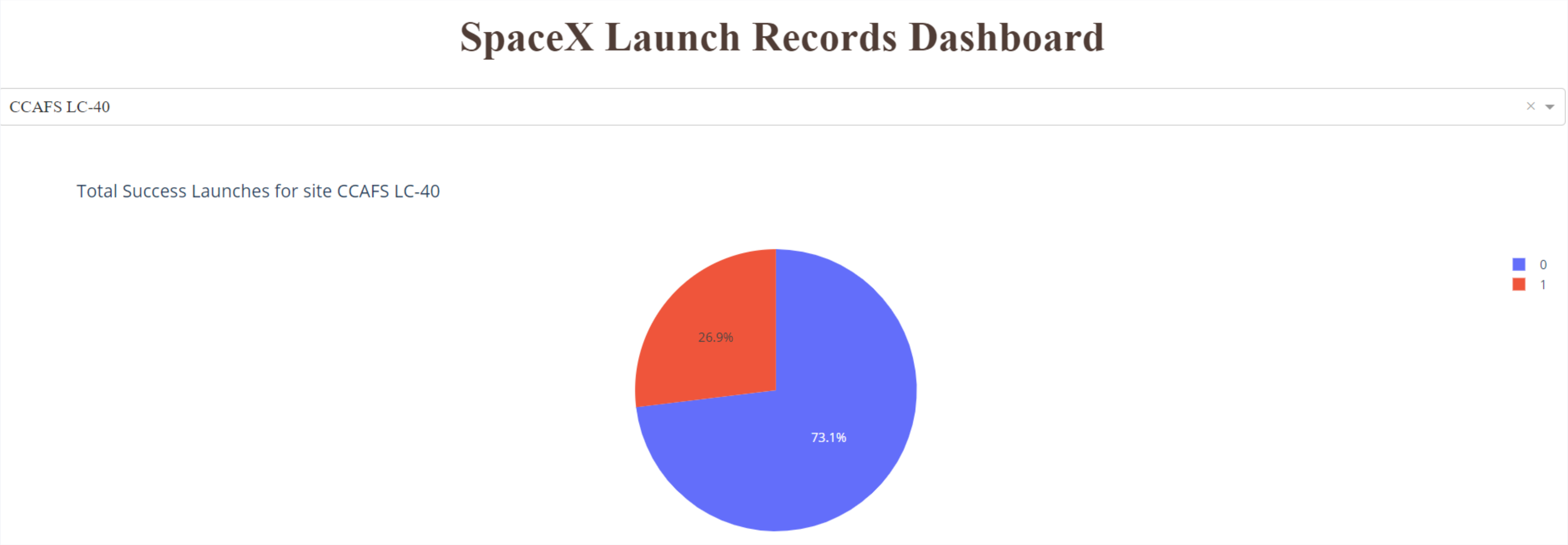
Build a Dashboard with Plotly Dash

Launch success count for all sites



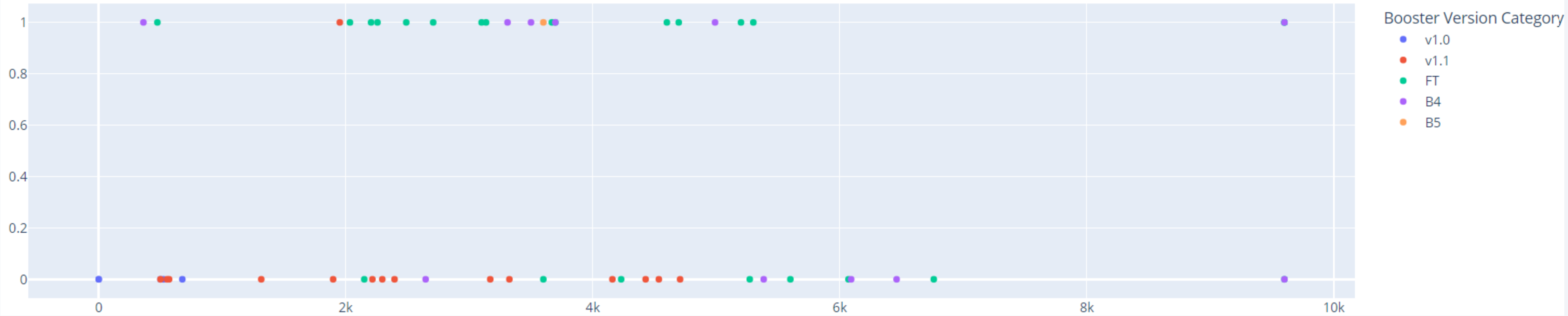
The CCAFS LC-40 has most number of success launch (46.4%)

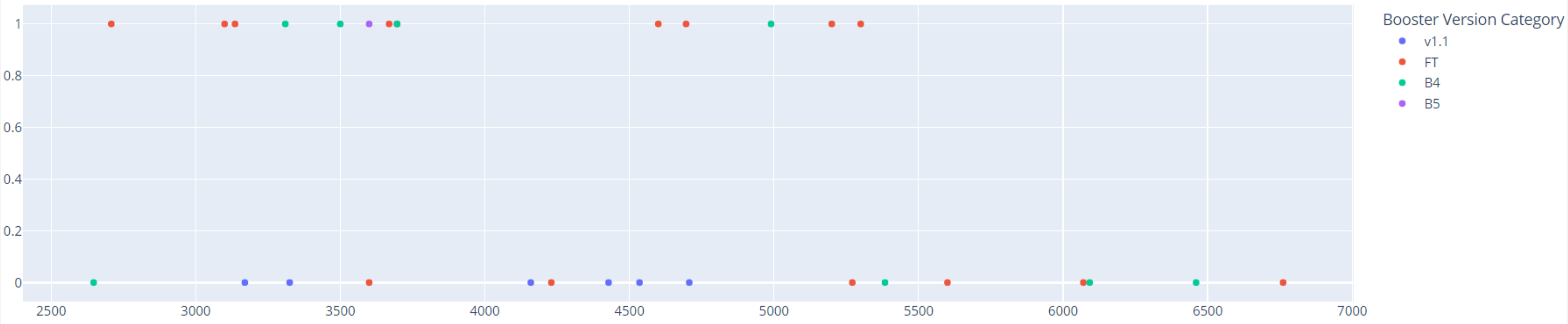
The launch site with highest success ratio



Although the CCAFS LC-40 has high number of success launch but if we look in to the ration, the success rate is quite low 26.9%

Payload vs. Launch Outcome





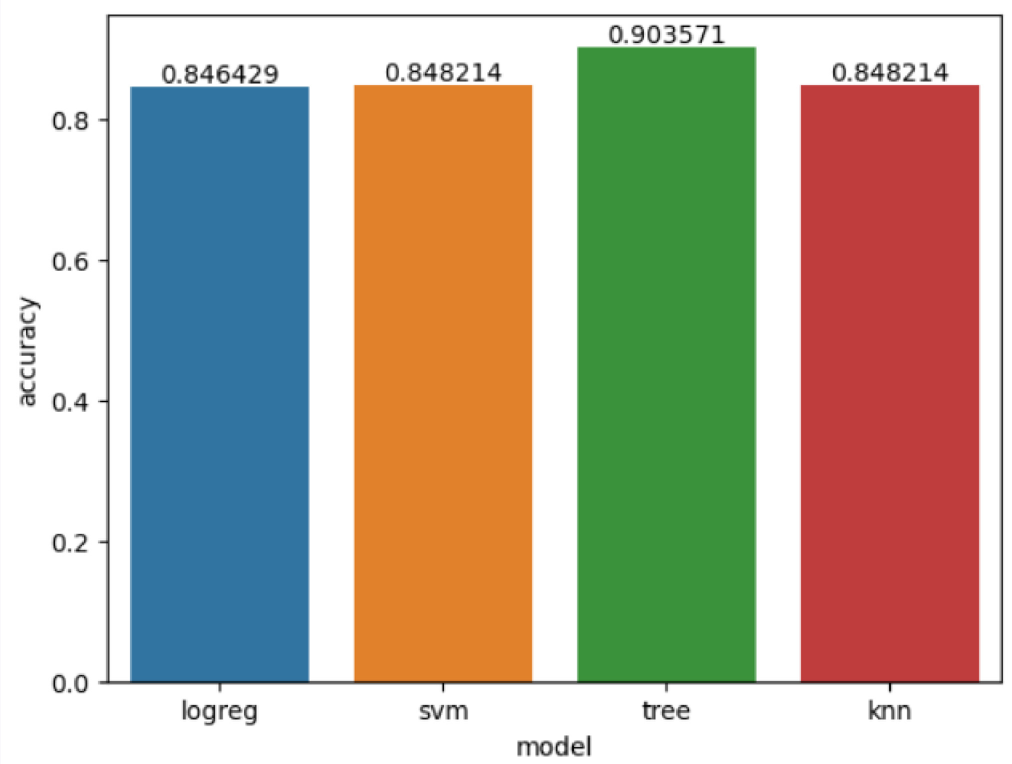
The B4 has success launches more than fail.

The v.1.0 is the only Booter Version has Payload more than around 7K.

Section 5

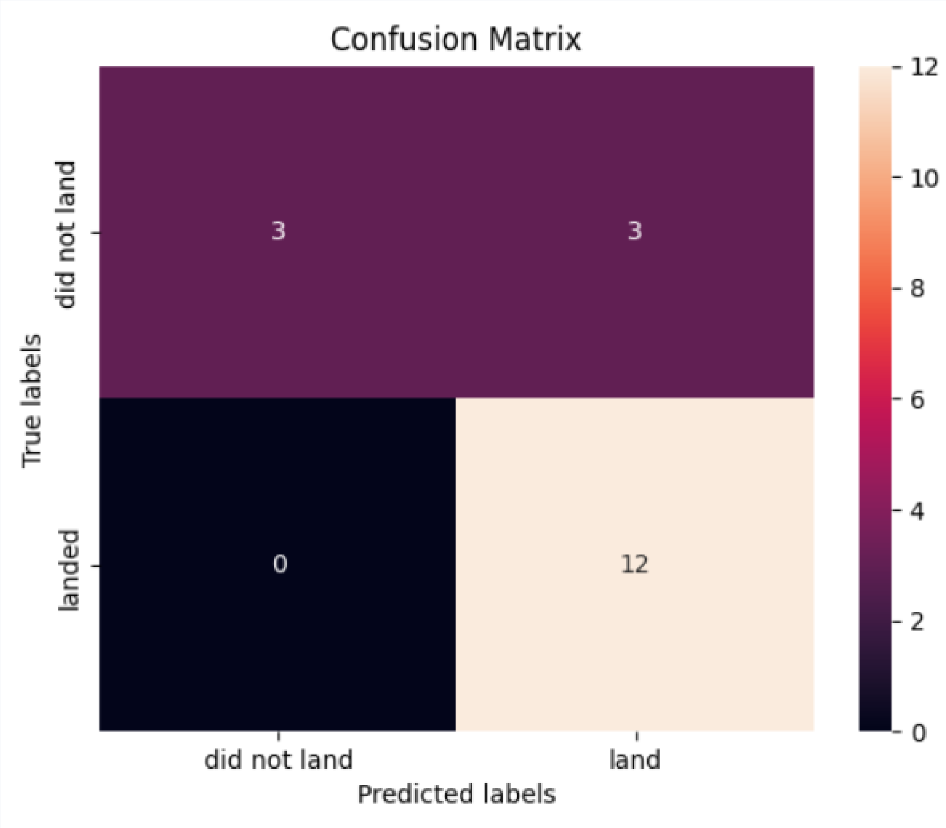
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree classification has highest accuracy (90.35%)

Confusion Matrix



The model can predict the "land" class (positive class) well but "did not land" (negative class) quite not good.

TN: 3 FP: 3

FN: 0 TP: 12

Conclusions

- The orbit type of ES-L1, GEO, HEO, and SSO have highest success rate
- The LEO orbit success appears related to number of flights but GTO is not
- Success rate since 2013 kept increasing till 2020
- All launch sites location is near to coast line
- The CCAFS LC-40 has high number of success launch buy success rate is quite low
- The Decision Tree classification model is the best performance on accuracy 90.35%

Appendix

The relevant assets was stored on GitHub, it's include all Notebook and Python code that I have created during this project

GitHub: https://github.com/atpluem/dscapstone.git

Thank you!