IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

D. Jungheim
19.12.2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results

Executive Summary

In this capstone project we found a model for predicting if the Falcon 9 first stage rocket of SpaceX will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. We also got insights in the interdependencies between certain parameters (such as launch site) and the success rate of a landing.

For that purpose we:

- Collected and cleared data

- Performed Exploratory Data Analysis (EDA) and visualized the data in a Dashboard

- Created a best fitting predictive model with simple machine learning techniques

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, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

We wanted to answer the following questions:

- How big is the success rate for a first stage landing?

- Is the success rate dependent on certain parameters (such as launch site etc.)?

- Can we - based on the impact of certain parameters - predict the outcome of a new rocket launch first stage landing?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

We collected the data from an API. We also performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page.

Perform data wrangling:

We converted rocket launch outcomes into Training Labels.

Perform Exploratory Data Analysis (EDA) using visualization and SQL:

We executed SQL queries and Feature Engineering on a dataset with a record for each payload carried during a SpaceX mission into outer space.

Perform interactive visual analytics using Folium and Plotly Dash

We visualized the dataset using matplotlib and seaborn to discovered some preliminary correlations. Then we performed an interactive visual analytics using Folium.

Perform predictive analysis using classification models

We split the data into training data and test data to find the best Hyperparameter for SVM, Classification Trees, KNN and Logistic Regression.

Data Collection – SpaceX API

We performed a GET-Request to the SpaceX API and cleaned the requested data:

1. Requested and parsed the SpaceX launch data using the GET request

2. Filtered the dataframe to only include Falcon 9 launches

3. Dealt with Missing Values

GitHub URL (SpaceX API Calls notebook):
https://github.com/DJungheim/ibm_capstone/blob/dd2057e5fe5aecd5e75b7af8d0aeff99bc71aa9d/01-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Web scraped Falcon 9 launch records with BeautifulSoup: Extracted a Falcon 9 launch records HTML table from Wikipedia, parse the table and converted it into a Pandas data frame:

1. Requested the Falcon9 Launch Wiki page from its URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

2. Extracted all variable names from the HTML table header

3. Created a data frame by parsing the launch HTML tables

GitHub URL (Web scraping notebook): https://github.com/DJungheim/ibm_capstone/blob/e18c0bcc99d20abd5bd9d57cf926a883bd52d505/02-jupyter-labs-webscraping.ipynb

Data Wrangling

In the data set, there are several different cases where the booster did or did not land successfully. We converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful:

1. Calculated the number of launches on each site
2. Calculated the number and occurrence of each orbit
3. Calculated the number and occurrence of mission outcome of the orbits
4. Created a landing outcome label from Outcome column

GitHub URL (Data wrangling notebook):
https://github.com/DJungheim/ibm_capstone/blob/b14fc1f727caccf26749705e20d598fb1f108f93/03-labs-jupyter-spacex-data%20wrangling_jupyterlite.ipynb

EDA with Data Visualization

We performed Exploratory Data Analysis (EDA) with Data Visualization and Feature Engineering using Pandas and Matplotlib:

1. Created Scatterplots: Relationship between flight number/payload mass and launch site

2. Created Barplot: Relationship between success rate of each orbit type

3. Created Scatterplots: Relationship between flight number/payload mass and orbit type

4. Created Lineplot: Launch success yearly trend

GitHub URL (EDA with data visualization notebook): https://github.com/DJungheim/ibm_capstone/blob/2fafa7c0ce4030ef693b7e2c95016848fd2269e8/04-jupyter-labs-datavisualization.ipynb

EDA with SQL

We loaded the dataset into a corresponding table in a Db2 database and executed SQL queries:

1. Displayed the names of the unique launch sites in the space mission

2. Displayed 5 records where launch sites begin with the string 'KSC'

3. Displayed the total payload mass carried by boosters launched by NASA (CRS)

4. Displayed average payload mass carried by booster version F9 v1.1

5. Listed the date where the successful landing outcome in drone ship was achieved

6. Listed the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

7. Listed the total number of successful and failure mission outcomes

8. Listed all the booster versions that have carried the maximum payload mass by using a subquery

9. Listed the records which will display the month names, succesful landing outcomes in ground pad, booster versions and launch site for the year 2017

10. Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20

GitHub URL (EDA with SQL):
https://github.com/DJungheim/ibm_capstone/blob/d15369c4fe71f182409383602ae74f6dfb44159c/05%20jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

The launch success may also depend on the location and proximities of a launch site, i.e. the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors. We discovered some of the factors by analyzing the existing launch site locations and performing interactive visual analytics using Folium. We...

1. marked all launch sites on a map,

2. marked the success/failed launches for each site on the map

3. and calculate the distances between a launch site to its proximities.

GitHub URL (Map with Folium):
https://github.com/DJungheim/ibm_capstone/blob/9fbdb82968c2587abb0a5ec07c96b879878f7305/06%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

We built a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart. We...

1. added a Launch Site Drop-down Input Component,

2. added a callback function to render success-pie-chart based on selected site dropdown,

3. added a Range Slider to Select Payload and

4. added a callback function to render the success-payload-scatter-chart scatter plot.

GitHub URL (Dashboard with Plotly Dash/Python script): https://github.com/DJungheim/ibm_capstone/blob/393fdbd42605bc8d33c5db689501e7f9c7d13993/07-spacex-dash-app.py

Predictive Analysis (Classification)

We built predictive Classification models using different machine learning techniques:

1. We performed Exploratory Data Analysis and determined Training Labels,

2. created a column for the class,

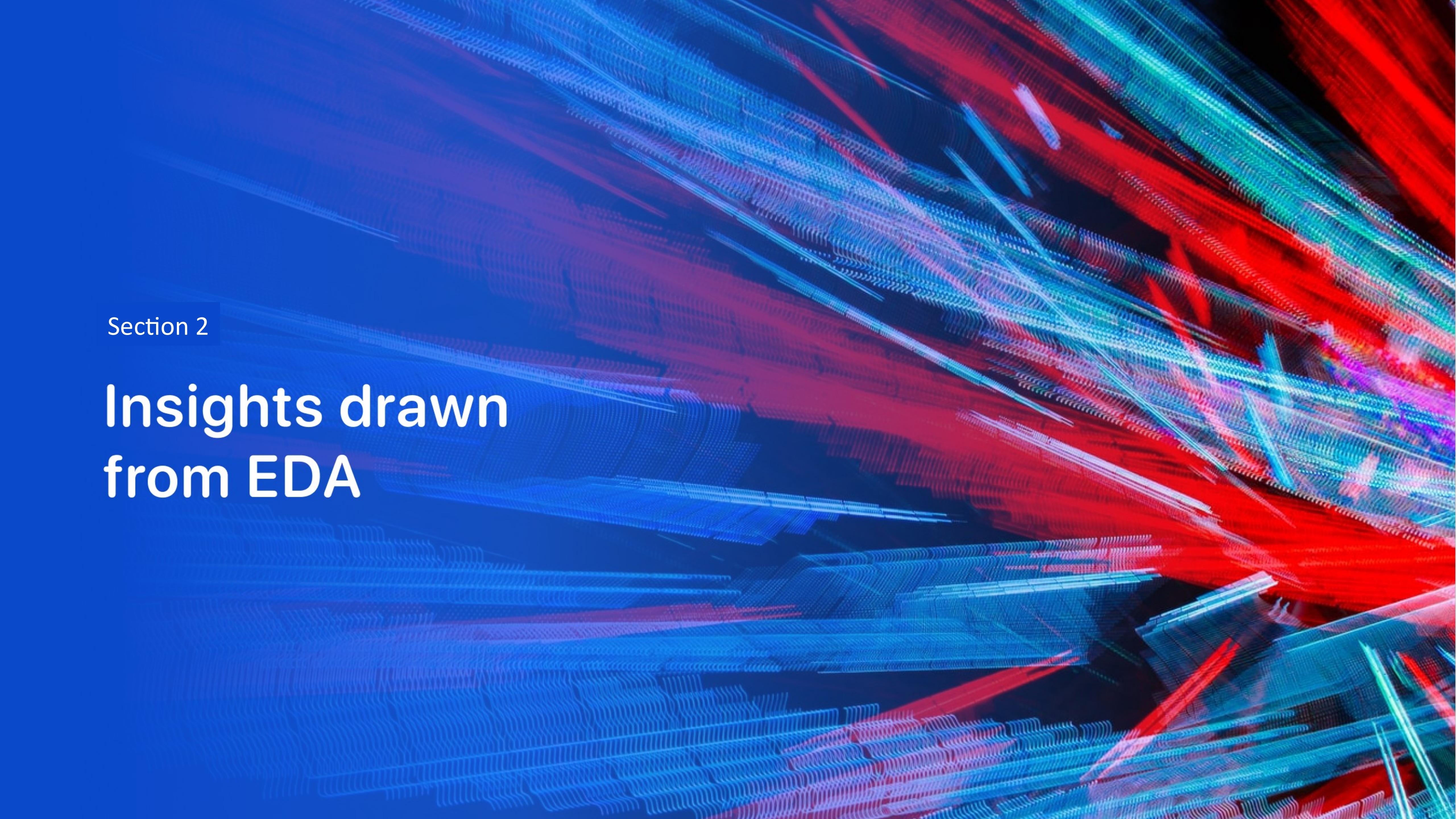
3. standardized the data,

4. split the data set into training data and test data,

5. found the best Hyperparameters for SVM, Classification Trees, KNN and Logistic Regression and

6. found the method that performs best using test data.

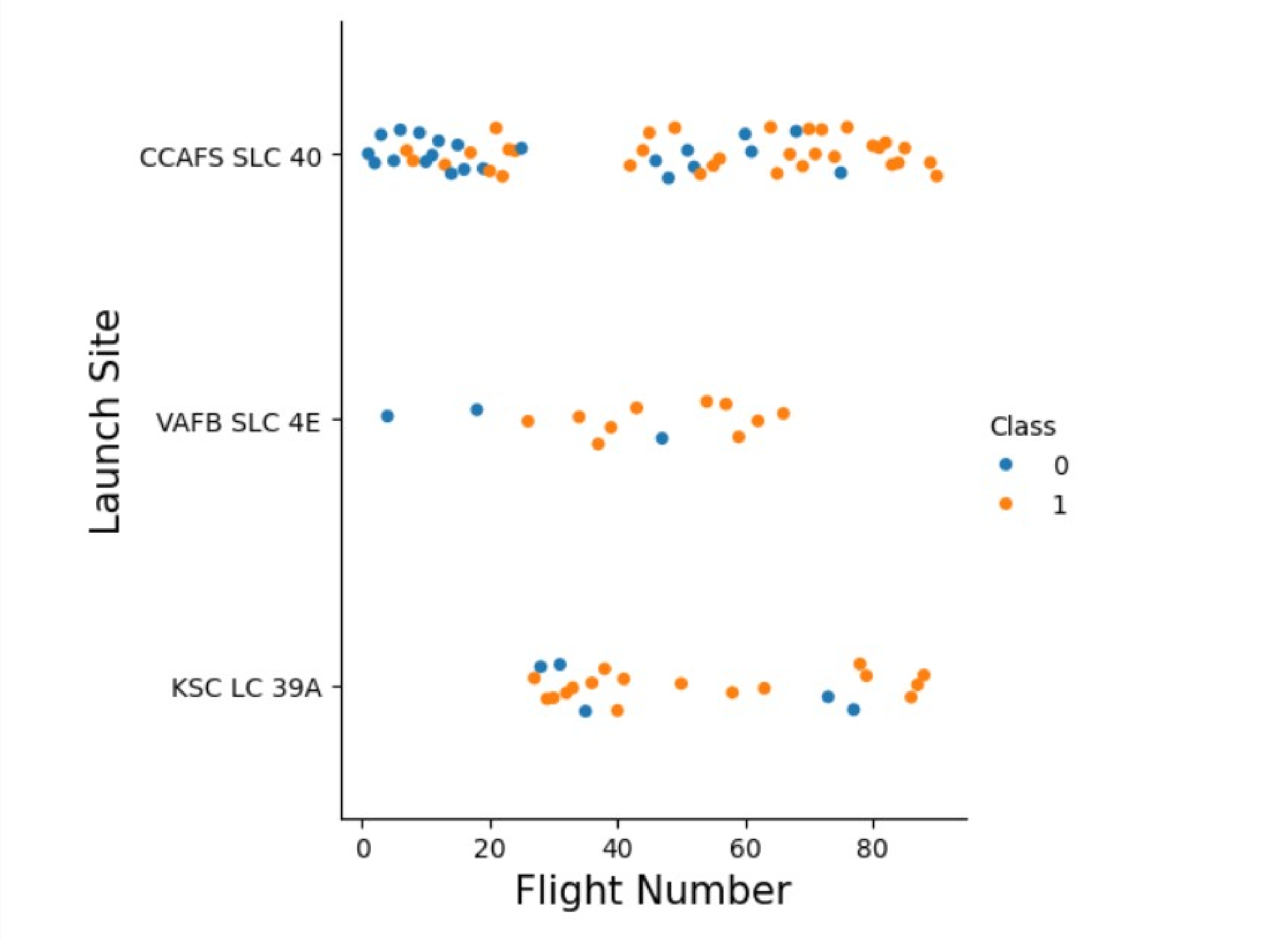
GitHub URL (Predictive analysis):
https://github.com/DJungheim/ibm_capstone/blob/4f522c5eb923467bda62852493d8aa591cb151f0/08-jupyter_labs_predictivemodels.ipynb



Section 2

Insights drawn from EDA

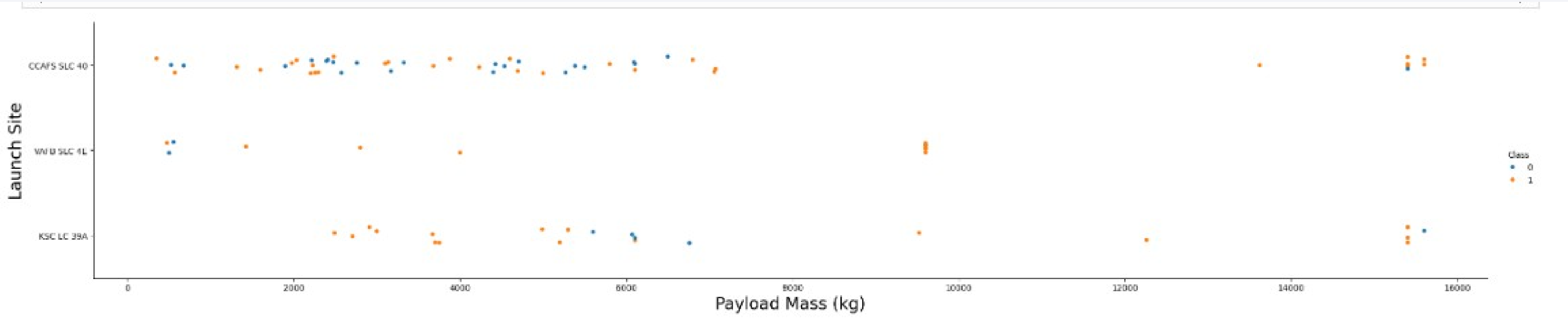
Flight Number vs. Launch Site



Class 0: unsuccessful landing
Class 1: successful landing

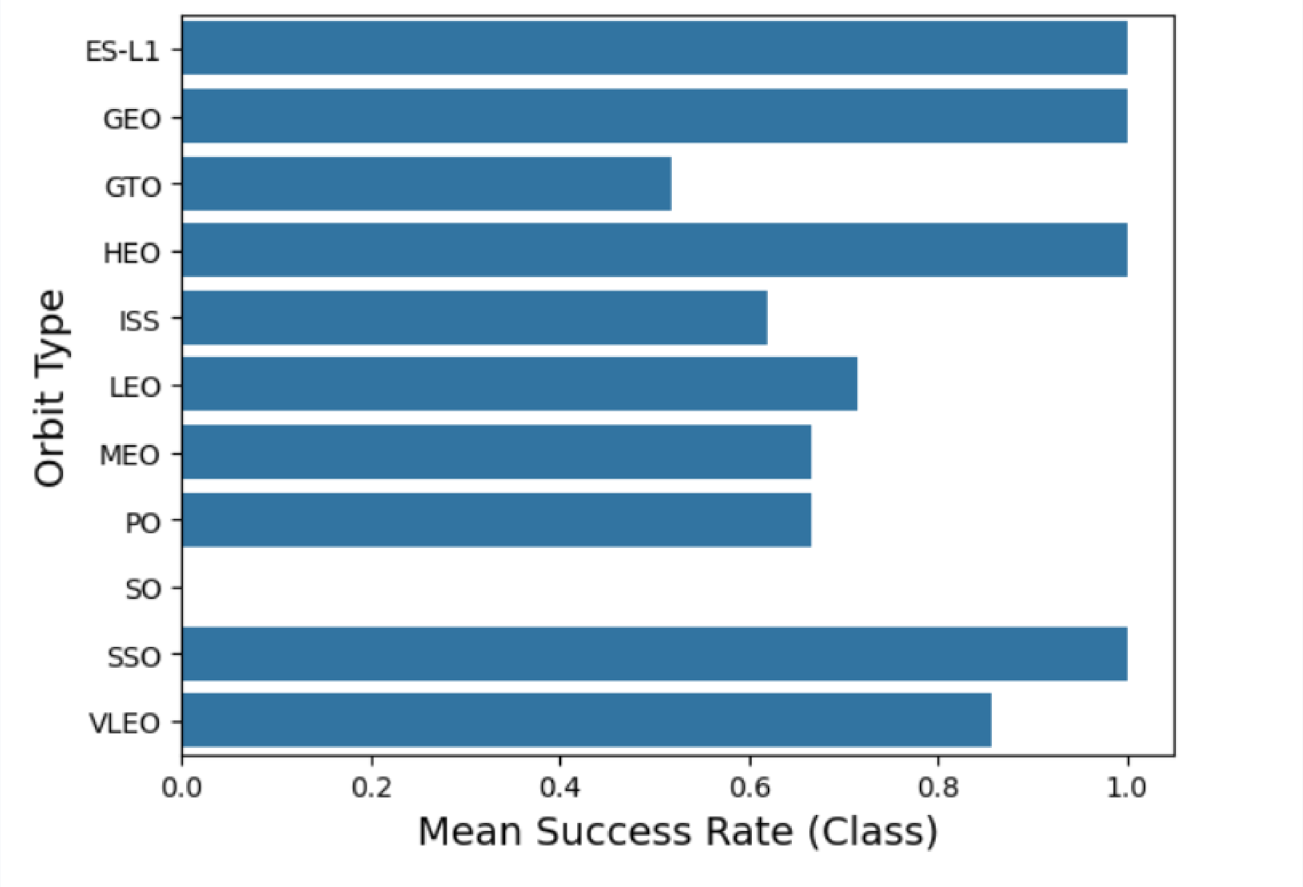
Most unsuccessful landings in early phase, especially at Cape Canaveral (CCAFS)

Payload vs. Launch Site



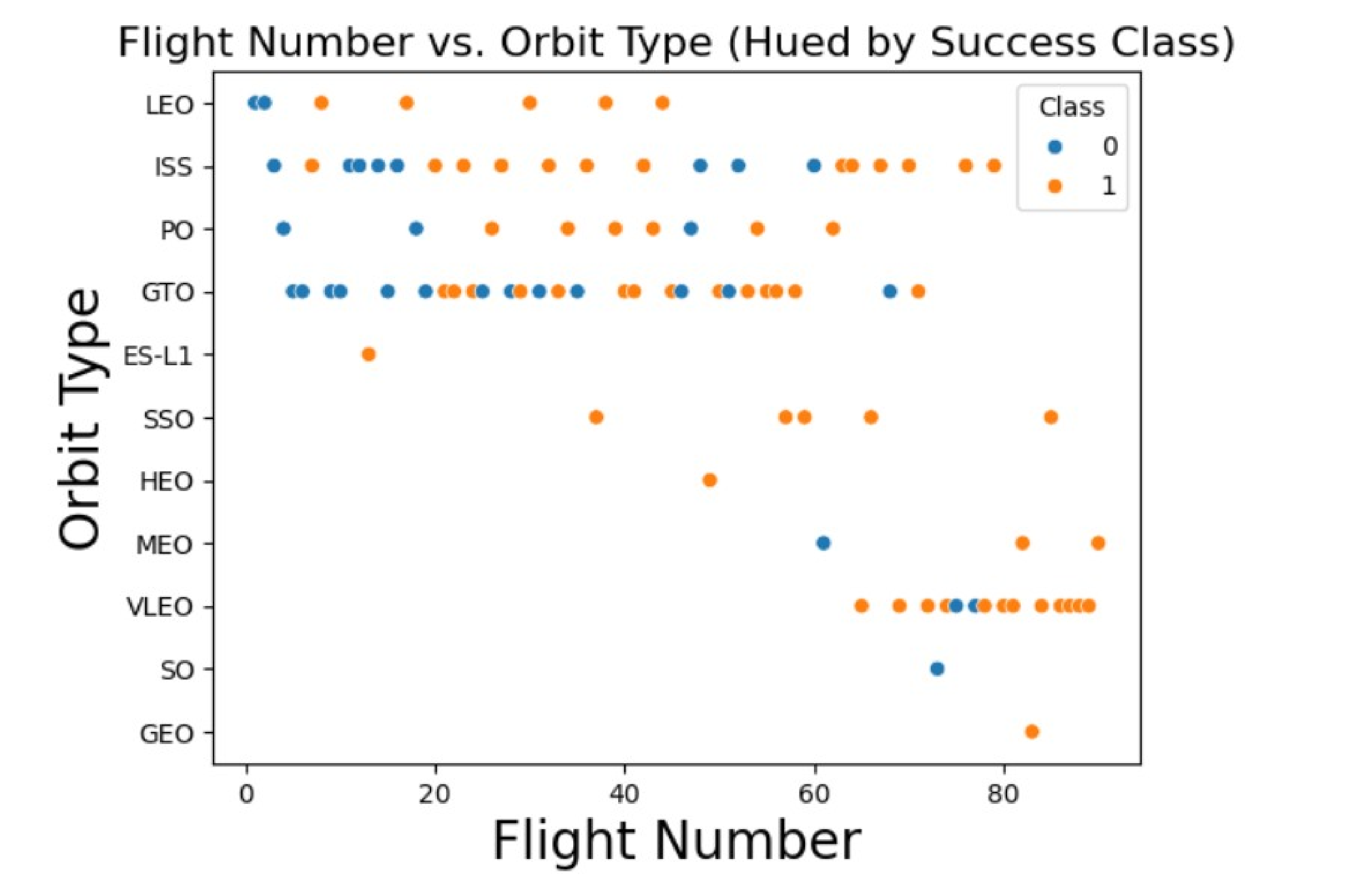
No strong correlation between payload mass and success rate, but a weak trend: higher payload mass --> slightly higher success rate

Success Rate vs. Orbit Type



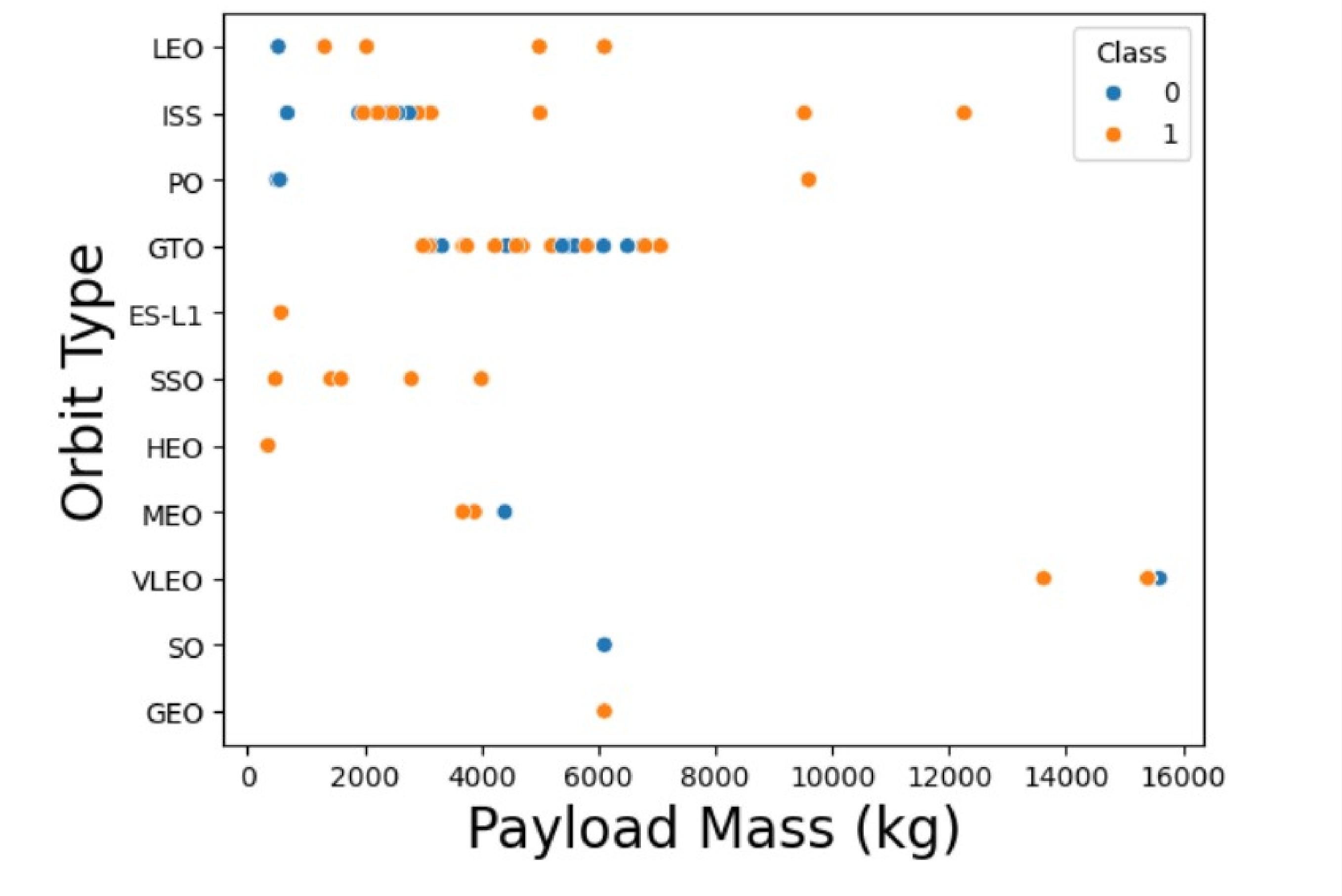
Mean Success Rate the highest (100%) at Orbit Types ES-L1, GEO, HEO and SSO, the lowest at GTO (around 50%), the rest in the middle field (60-70%).

Flight Number vs. Orbit Type

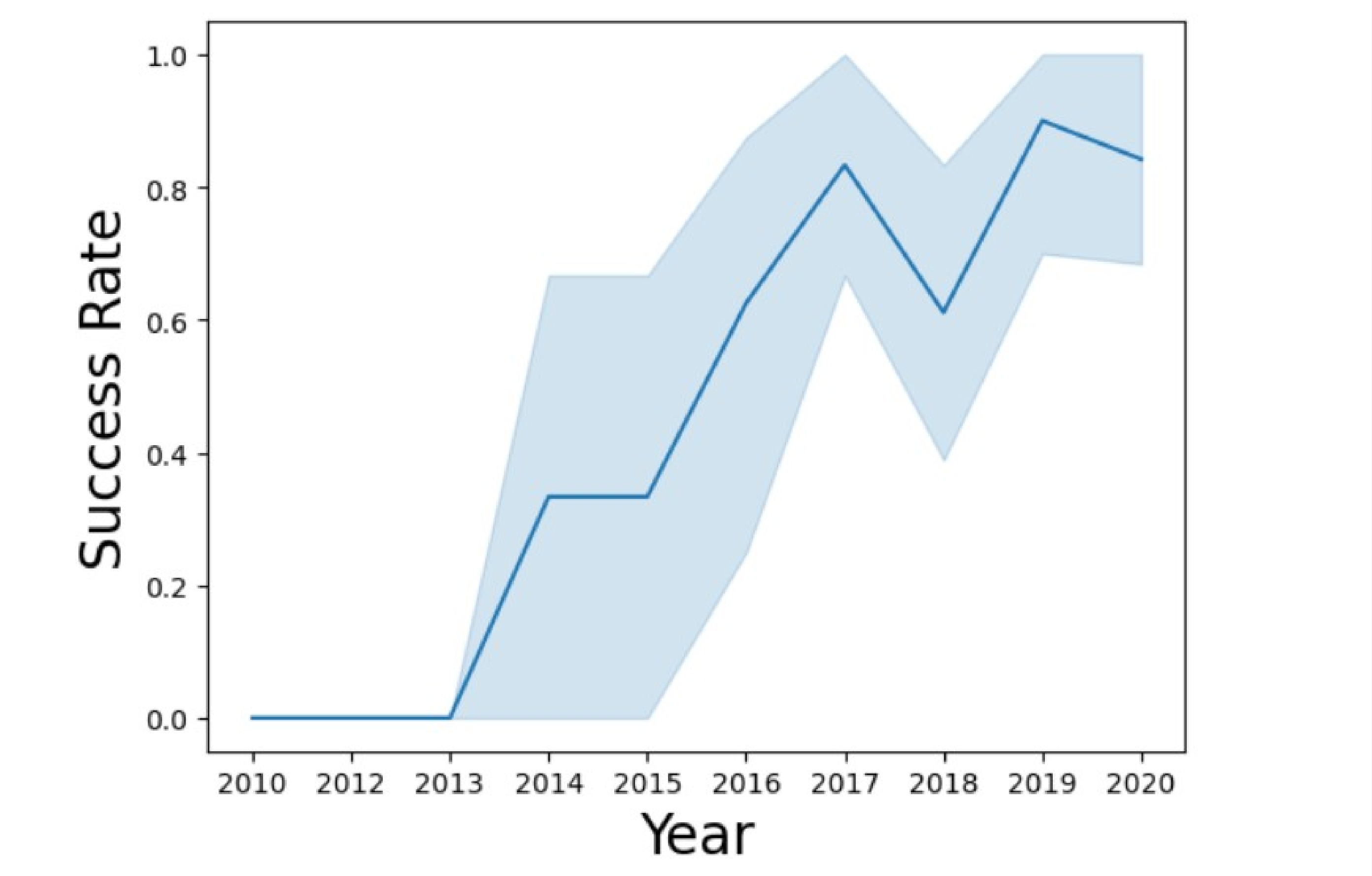


100% success rate at ES-L1, HEO and GEO is explainable by the small number of flights to that orbit types!

Payload vs. Orbit Type

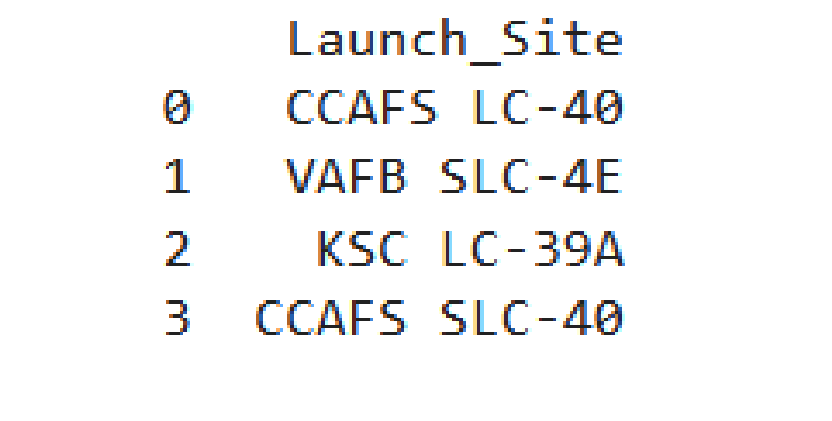


Launch Success Yearly Trend



Success rate rising after 2013, small collapse 2018.

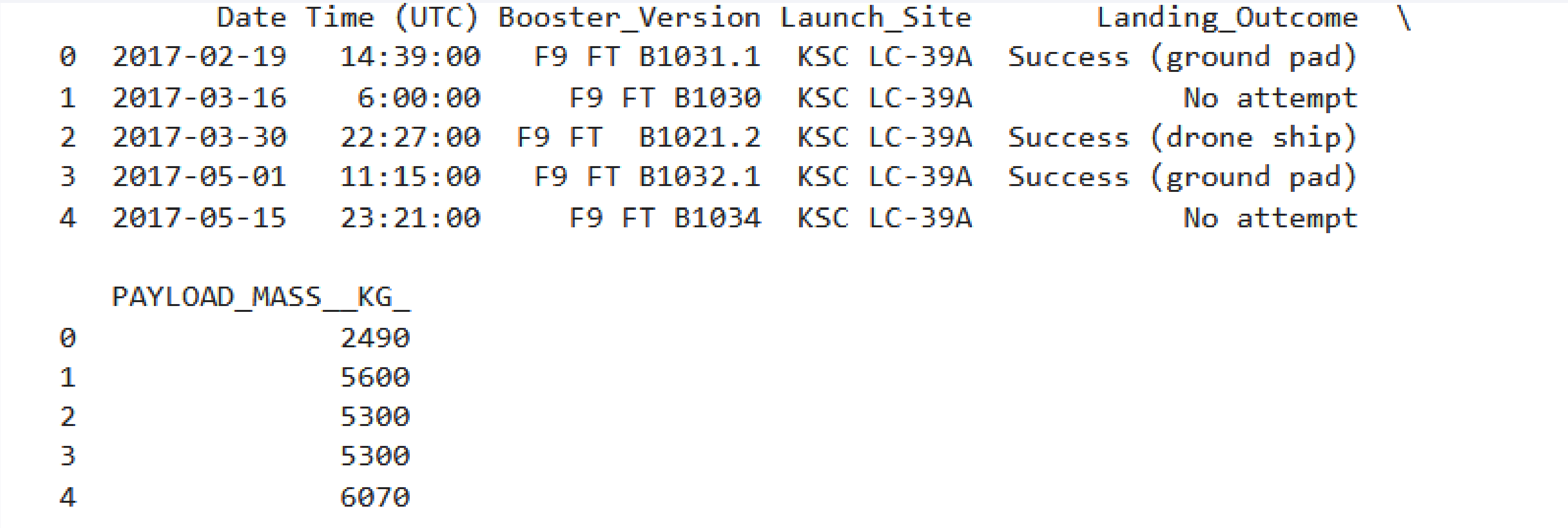
All Launch Site Names



```
     Launch_Site
0    CCAFS LC-40
1    VAFB SLC-4E
2     KSC LC-39A
3   CCAFS SLC-40
```

CCAFS LC-40: Cape Canaveral Air Force Station, Launch Complex 40 (Florida, USA)
VAFB SLC-4E: Vandenberg Air Force Base, Space Launch Complex 4 East (California, USA)
KSC LC-39A: Kennedy Space Center, Launch Complex 39A (Florida, USA)
CCAFS SLC-40: Cape Canaveral Air Force Station, Space Launch Complex 40 (Florida, USA)

Launch Site Names Begin with 'KSC'

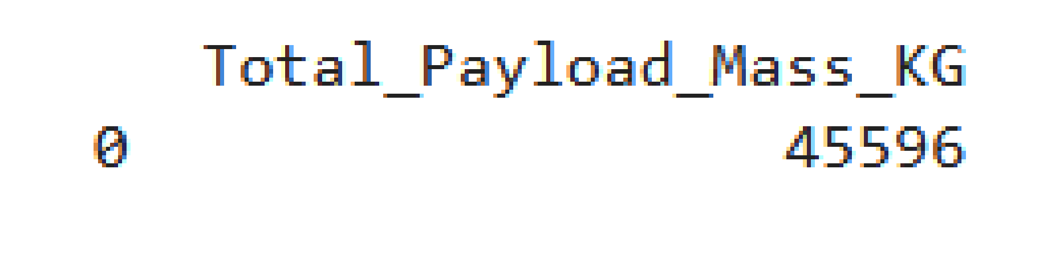


```
         Date Time (UTC) Booster_Version Launch_Site       Landing_Outcome  \
0  2017-02-19   14:39:00   F9 FT B1031.1  KSC LC-39A  Success (ground pad)   
1  2017-03-16    6:00:00     F9 FT B1030  KSC LC-39A            No attempt   
2  2017-03-30   22:27:00  F9 FT  B1021.2  KSC LC-39A  Success (drone ship)   
3  2017-05-01   11:15:00   F9 FT B1032.1  KSC LC-39A  Success (ground pad)   
4  2017-05-15   23:21:00     F9 FT B1034  KSC LC-39A            No attempt   

   PAYLOAD_MASS__KG_  
0               2490  
1               5600  
2               5300  
3               5300  
4               6070  
```

5 records where launch sites begin with the string 'KSC'

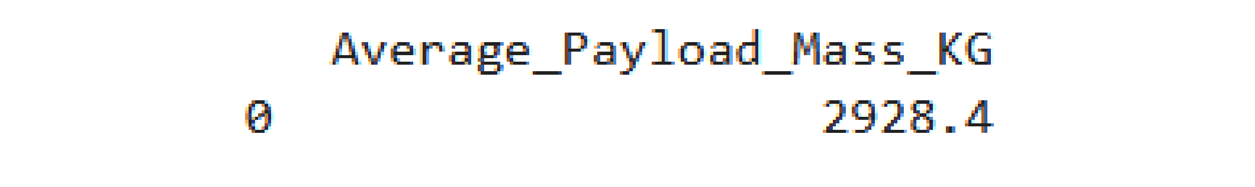
Total Payload Mass



```
   Total_Payload_Mass_KG
0                  45596
```

Total payload mass carried by boosters launched by NASA (CRS): 45,596 kg

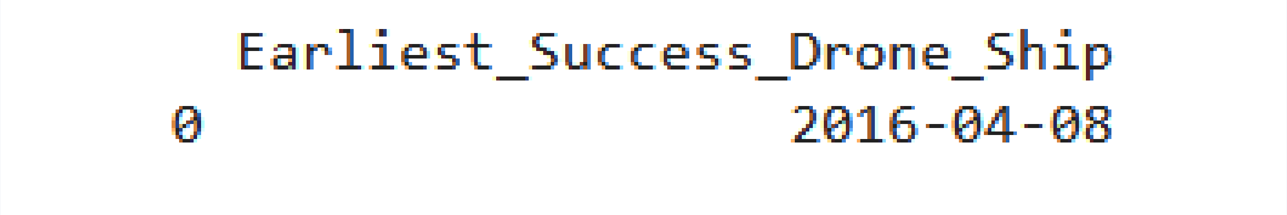
Average Payload Mass by F9 v1.1



```
   Average_Payload_Mass_KG
0                   2928.4
```

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

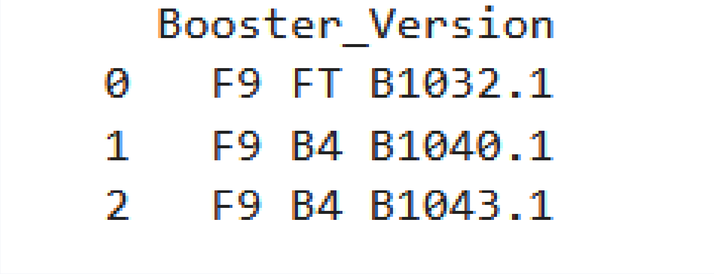
First Successful Ground Landing Date



```
   Earliest_Success_Drone_Ship
0                   2016-04-08
```

Date where the succesful landing outcome in drone ship was achieved: 8th April 2016

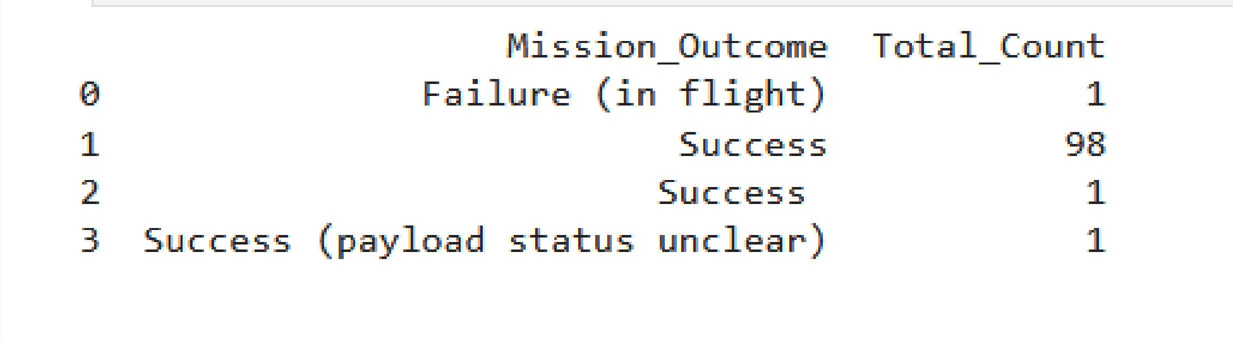
Successful Drone Ship Landing with Payload between 4000 and 6000



```
  Booster_Version
0   F9 FT B1032.1
1   F9 B4 B1040.1
2   F9 B4 B1043.1
```

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

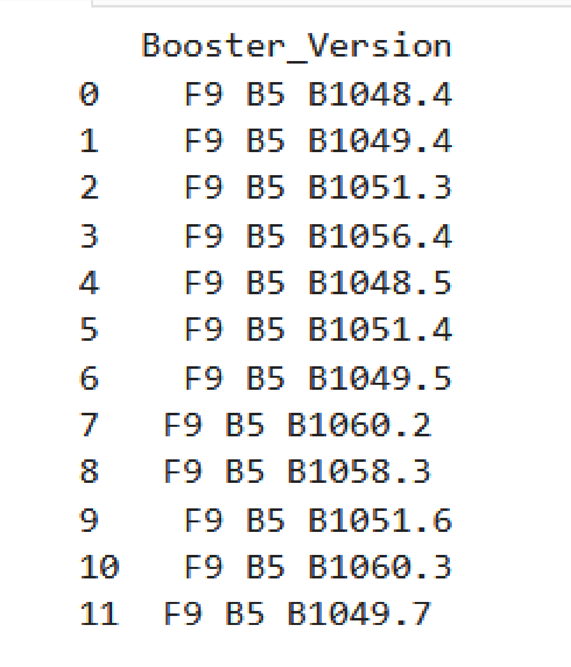
Total Number of Successful and Failure Mission Outcomes



```
                    Mission_Outcome  Total_Count
0               Failure (in flight)            1
1                           Success           98
2                          Success             1
3  Success (payload status unclear)            1
```

Total number of successful and failure mission outcomes

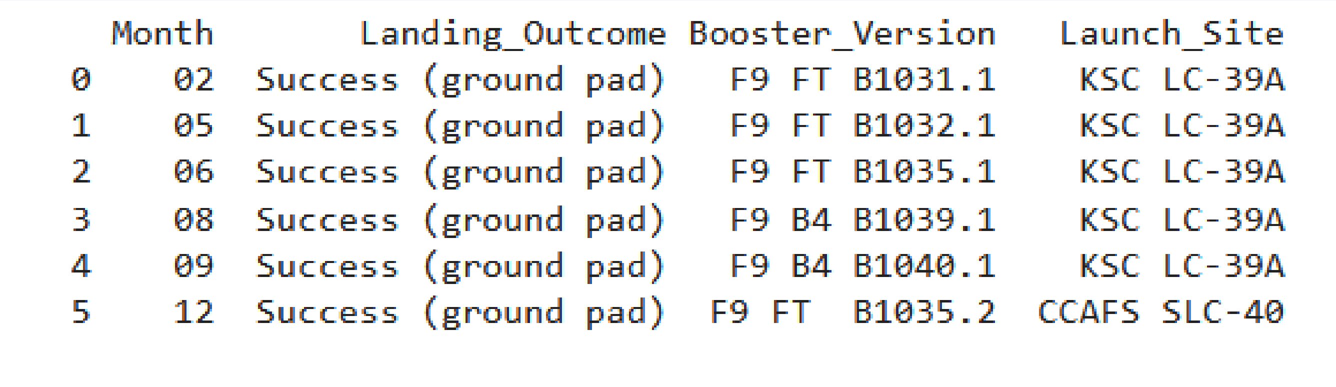
Boosters Carried Maximum Payload



```
   Booster_Version
0    F9 B5 B1048.4
1    F9 B5 B1049.4
2    F9 B5 B1051.3
3    F9 B5 B1056.4
4    F9 B5 B1048.5
5    F9 B5 B1051.4
6    F9 B5 B1049.5
7   F9 B5 B1060.2
8   F9 B5 B1058.3
9    F9 B5 B1051.6
10   F9 B5 B1060.3
11  F9 B5 B1049.7
```

Names of the boosters which have carried the maximum payload mass

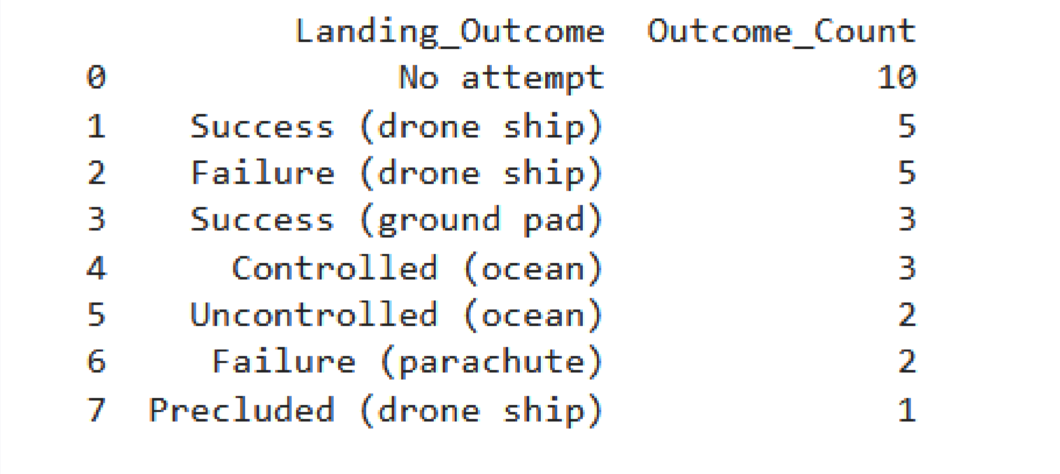
2017 Launch Records



```
  Month       Landing_Outcome Booster_Version   Launch_Site
0    02  Success (ground pad)   F9 FT B1031.1    KSC LC-39A
1    05  Success (ground pad)   F9 FT B1032.1    KSC LC-39A
2    06  Success (ground pad)   F9 FT B1035.1    KSC LC-39A
3    08  Success (ground pad)   F9 B4 B1039.1    KSC LC-39A
4    09  Success (ground pad)   F9 B4 B1040.1    KSC LC-39A
5    12  Success (ground pad)  F9 FT  B1035.2  CCAFS SLC-40
```

Month names, successful landing outcomes in ground pad, booster versions and launch site for the months in year 2017

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



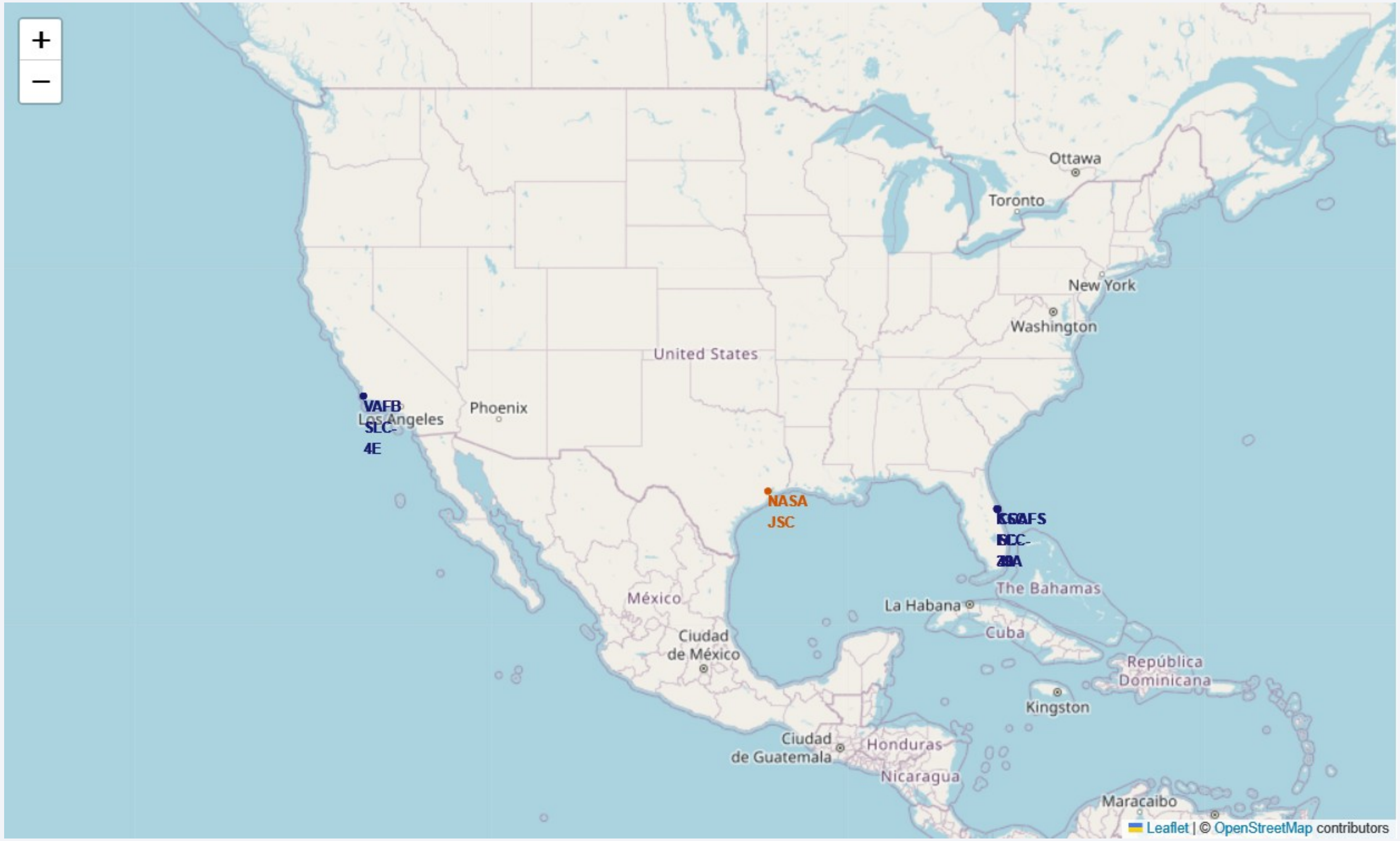
```
          Landing_Outcome  Outcome_Count
0              No attempt             10
1    Success (drone ship)              5
2    Failure (drone ship)              5
3   Success (ground pad)               3
4      Controlled (ocean)              3
5    Uncontrolled (ocean)              2
6     Failure (parachute)              2
7  Precluded (drone ship)              1
```

Count of landing outcomes between dates 2010-06-04 and 2017-03-20 in descending order

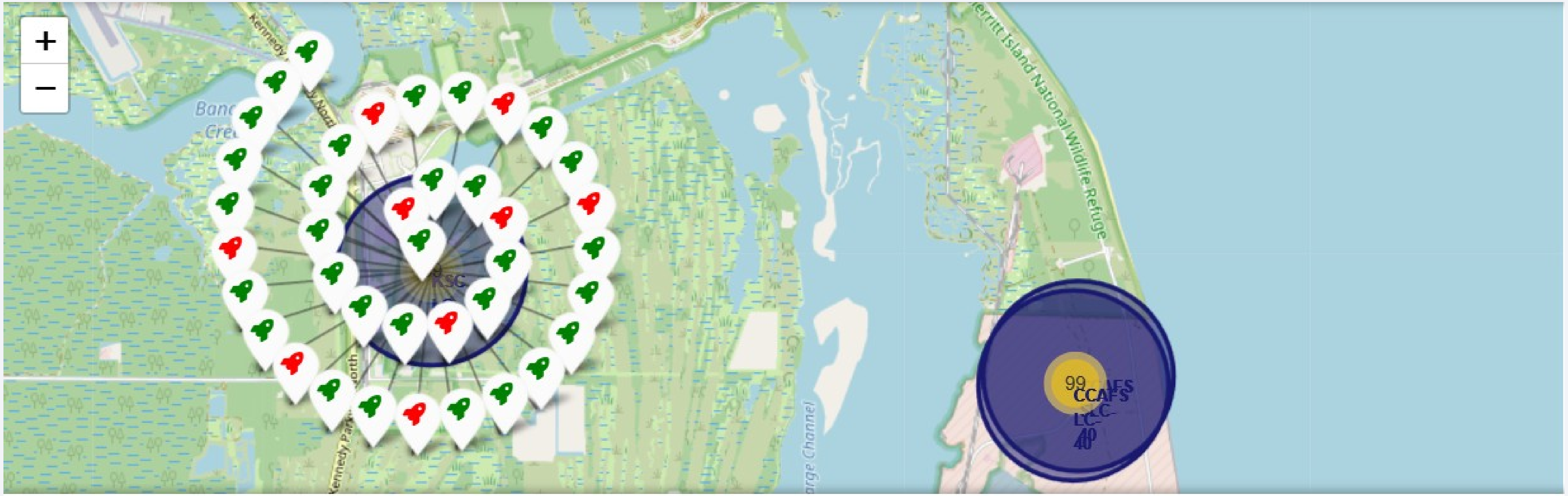
Section 3

Launch Sites Proximities Analysis

Launch Site Locations (Folium)

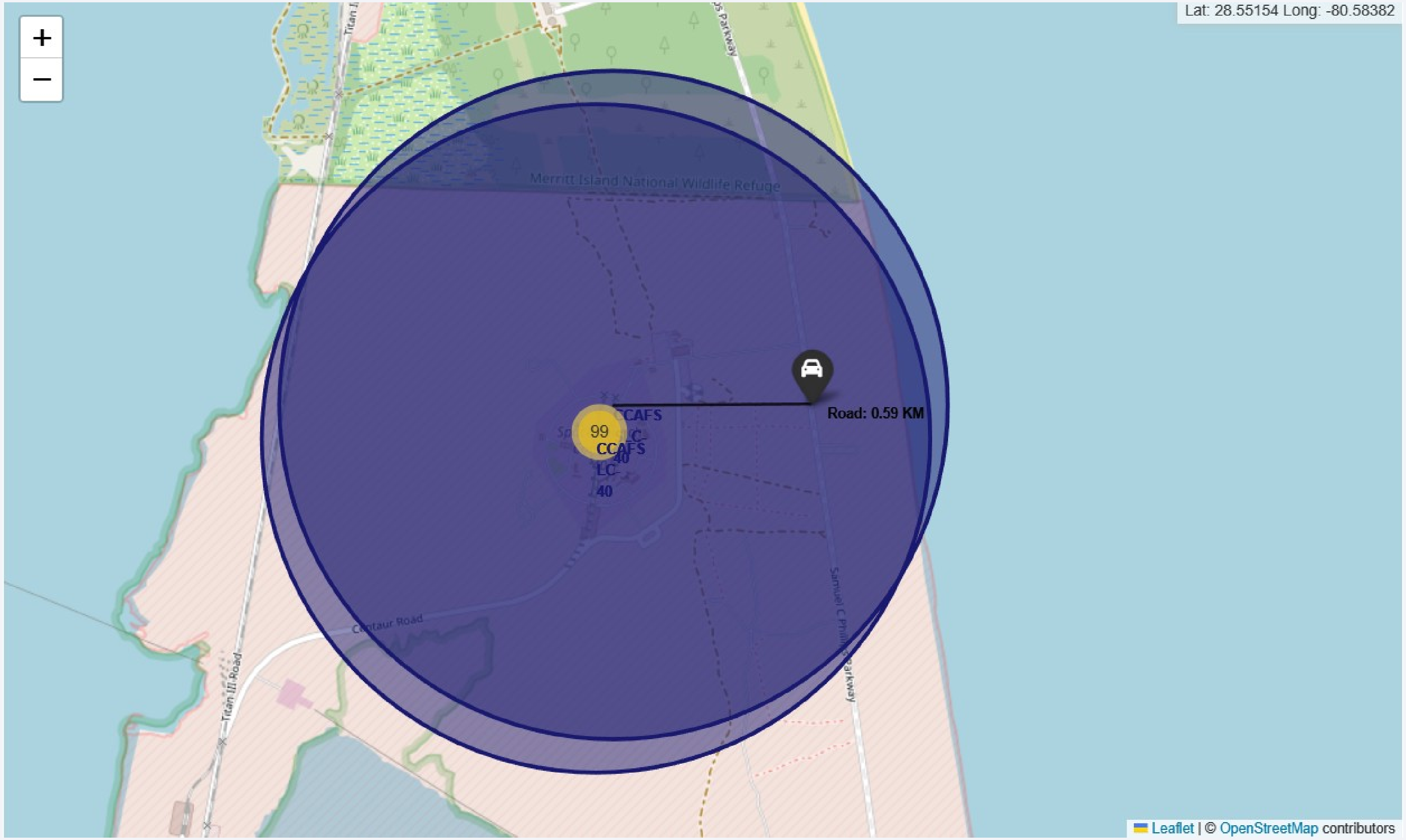


All Launch Records (Folium)

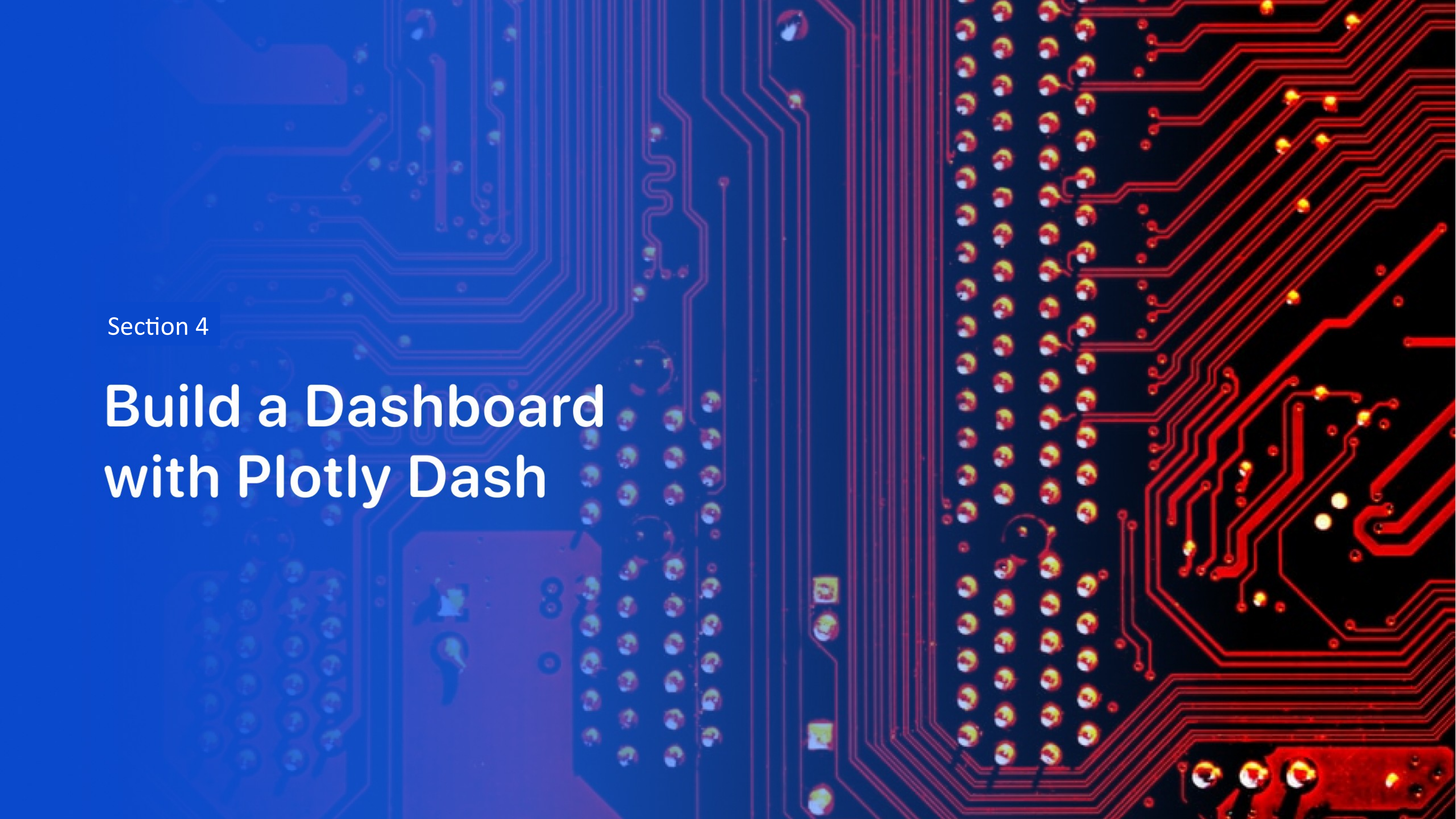


green: successful landing, red: failed landing

Distances: Launch Site – Proximity/Road (Folium)



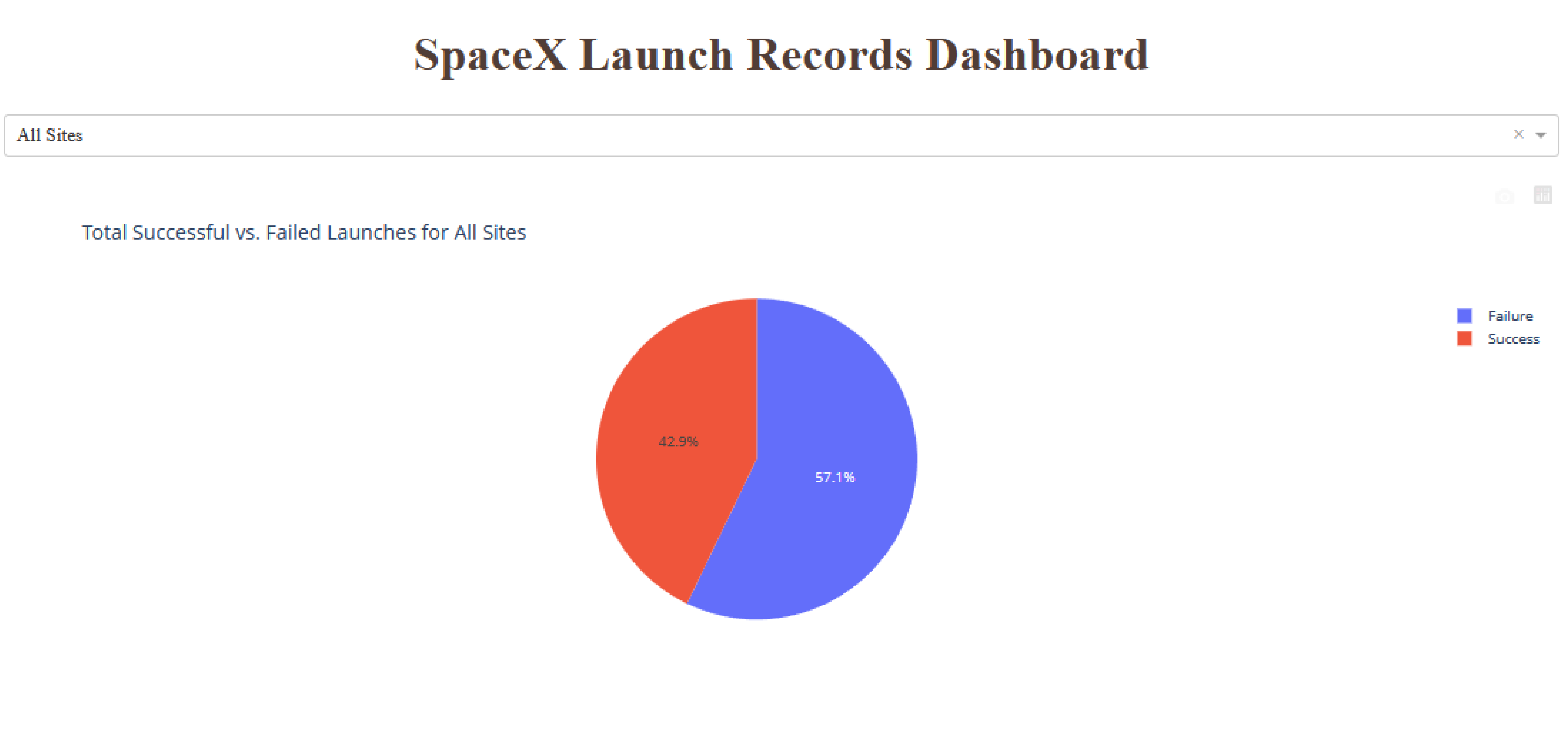
+
−
Lat: 28.55154 Long: -80.58382
Merritt Island National Wildlife Refuge
99
CCAFS LC-40
CCAFS LC-40
Road: 0.59 KM
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors



Section 4

Build a Dashboard with Plotly Dash

Piechart: Launch Success Count (Dash)

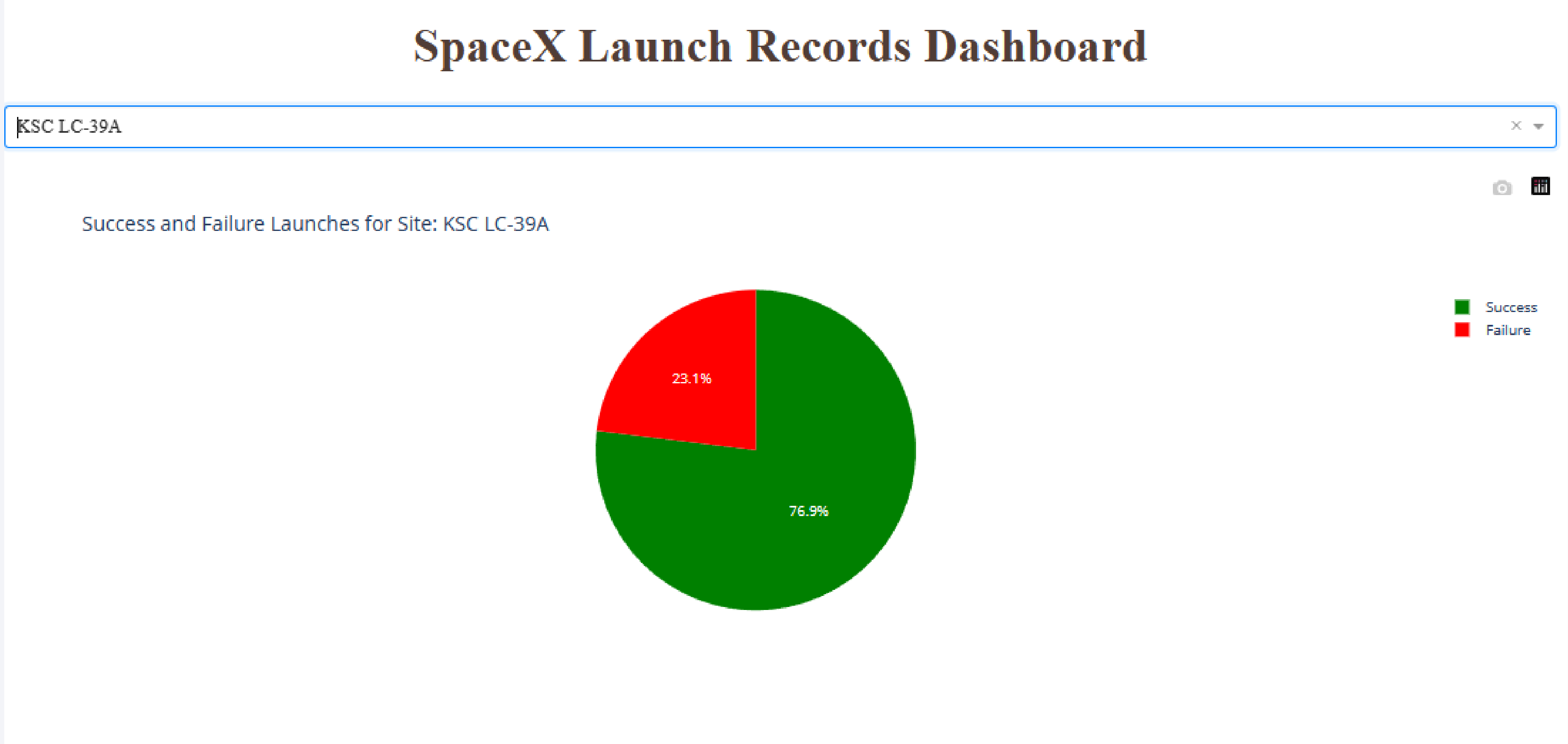


SpaceX Launch Records Dashboard

All Sites

Total Successful vs. Failed Launches for All Sites

Piechart: Launch Success Count for KSC (Dash)



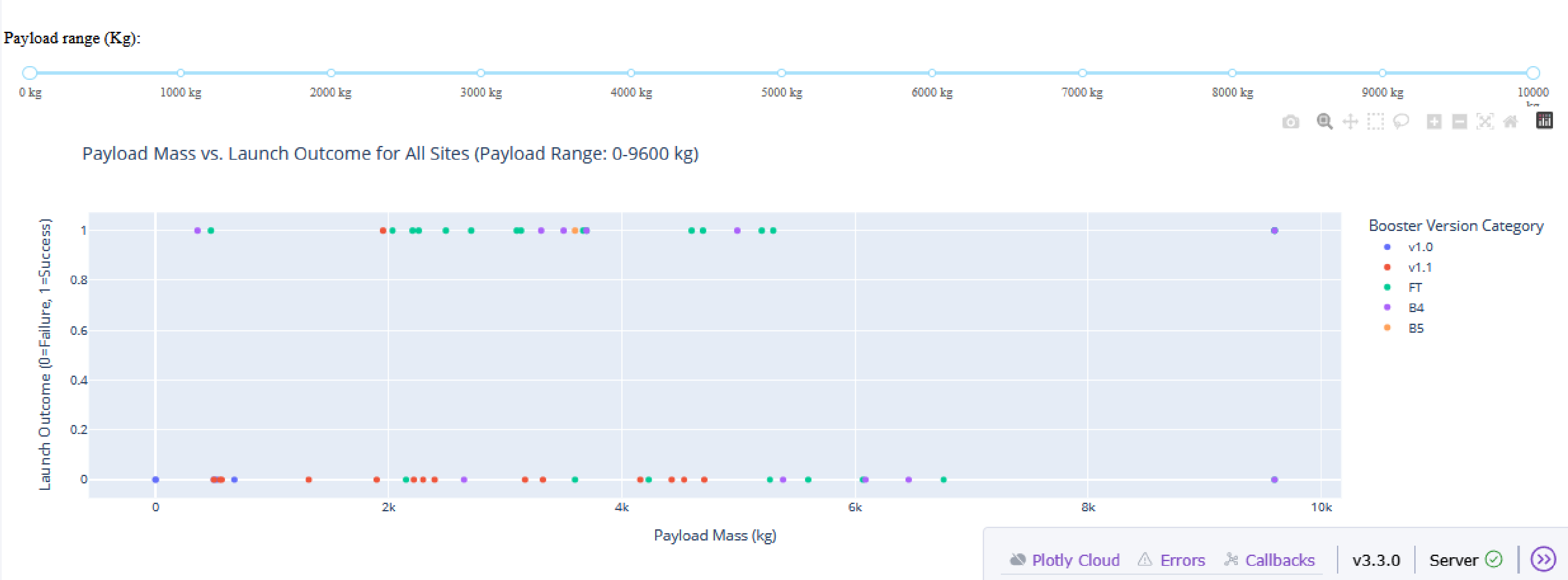
SpaceX Launch Records Dashboard

KSC LC-39A

Success and Failure Launches for Site: KSC LC-39A

Highest success rate of all Locations: Kennedy Space Center: 76,9%

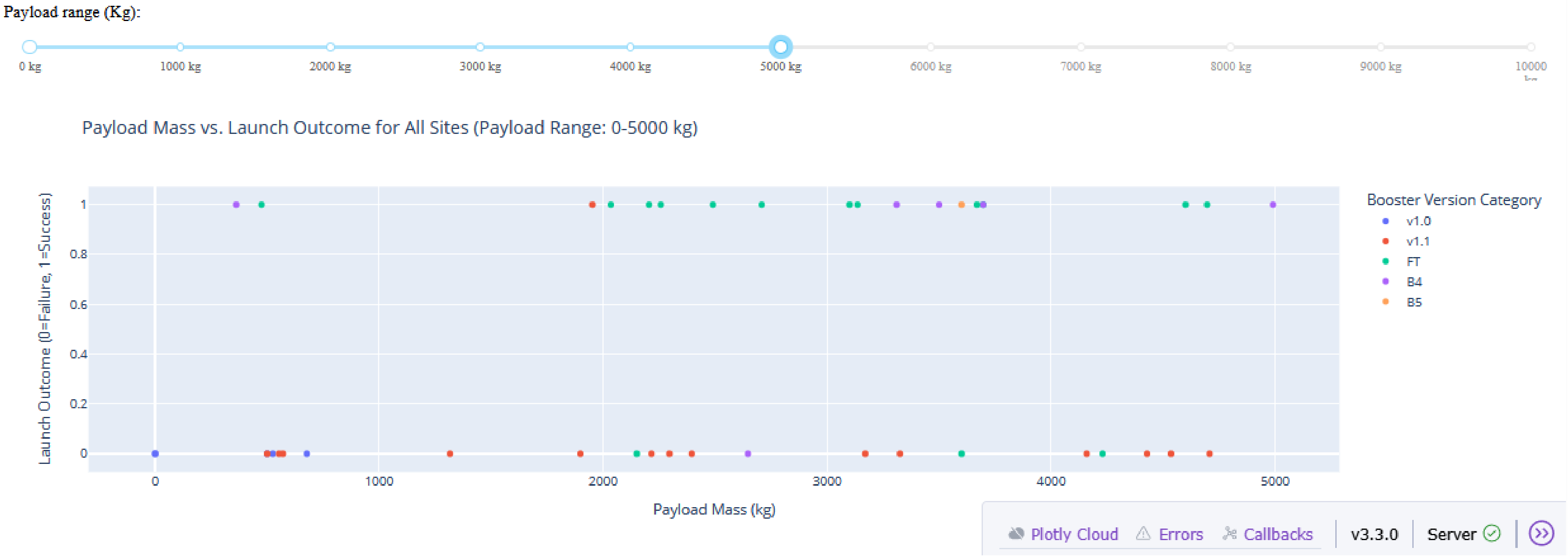
Scatterplots: Launch Outcome vs. Payload (Dash)



Payload range (Kg):

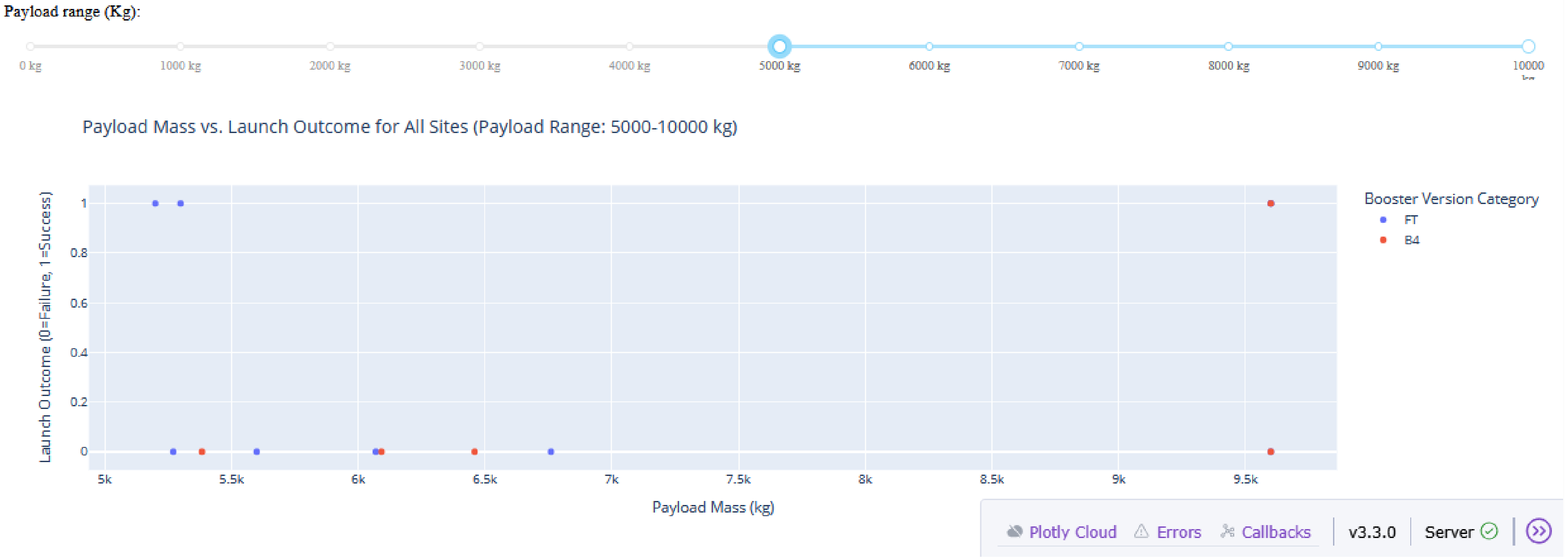
Payload Range 0-10000 kg

Payload range:
0-5000 kg



Payload Mass vs. Launch Outcome for All Sites (Payload Range: 0-5000 kg)

Payload range:
5000-10000 kg

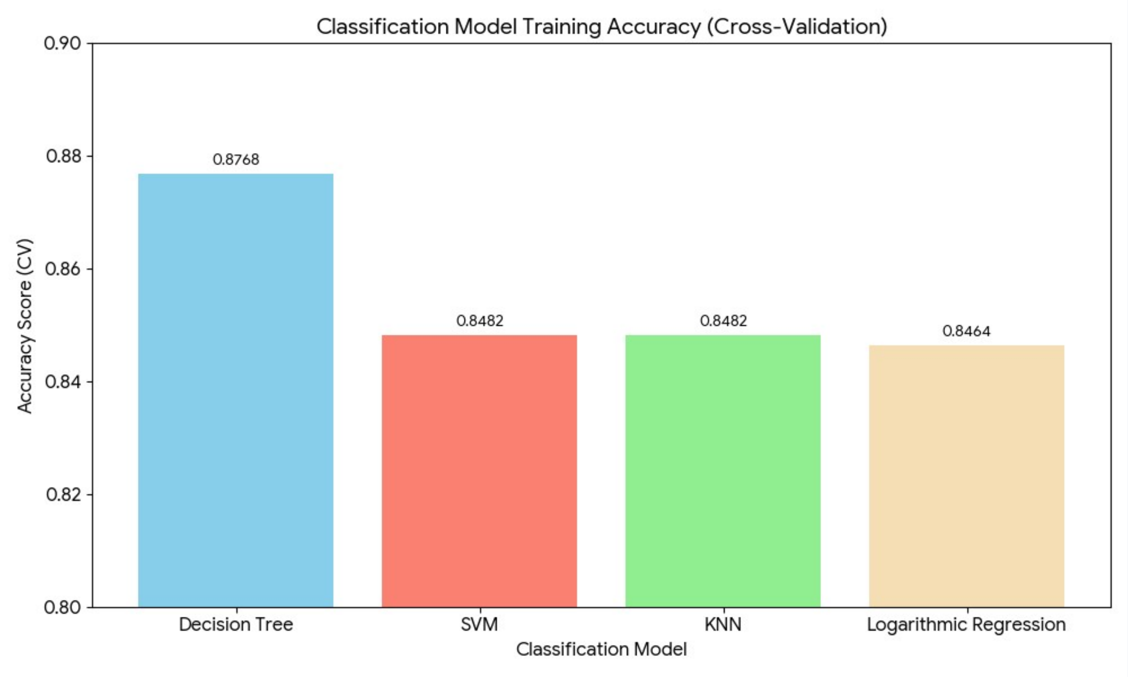


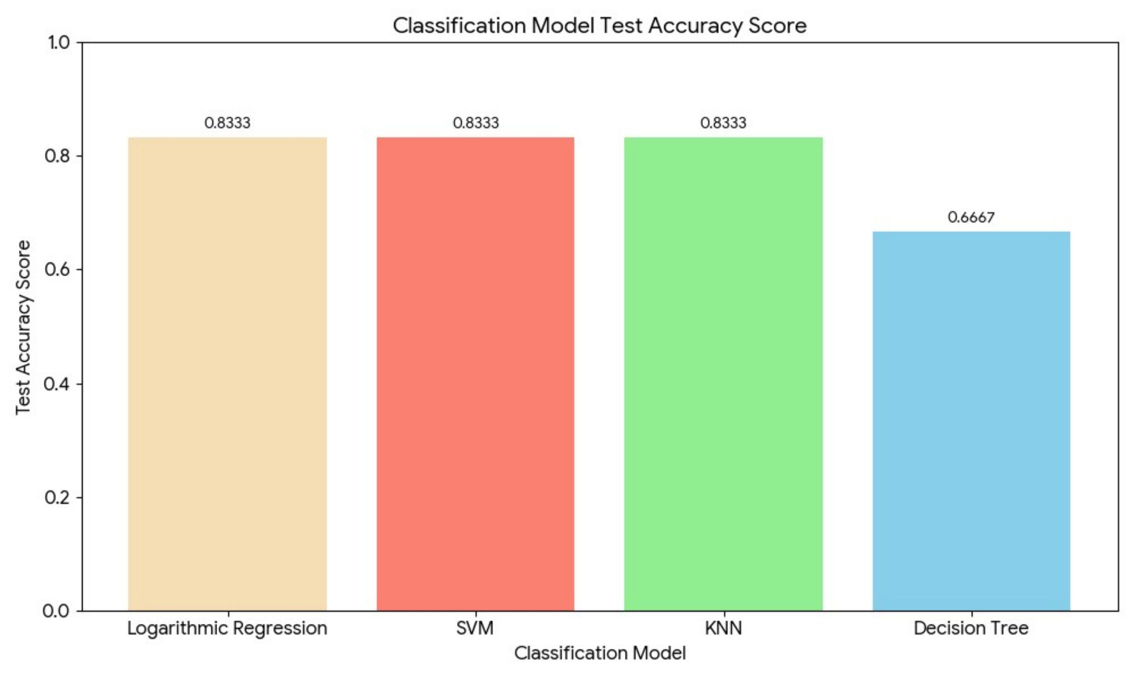
Payload Mass vs. Launch Outcome for All Sites (Payload Range: 5000-10000 kg)

Section 5

Predictive Analysis (Classification)

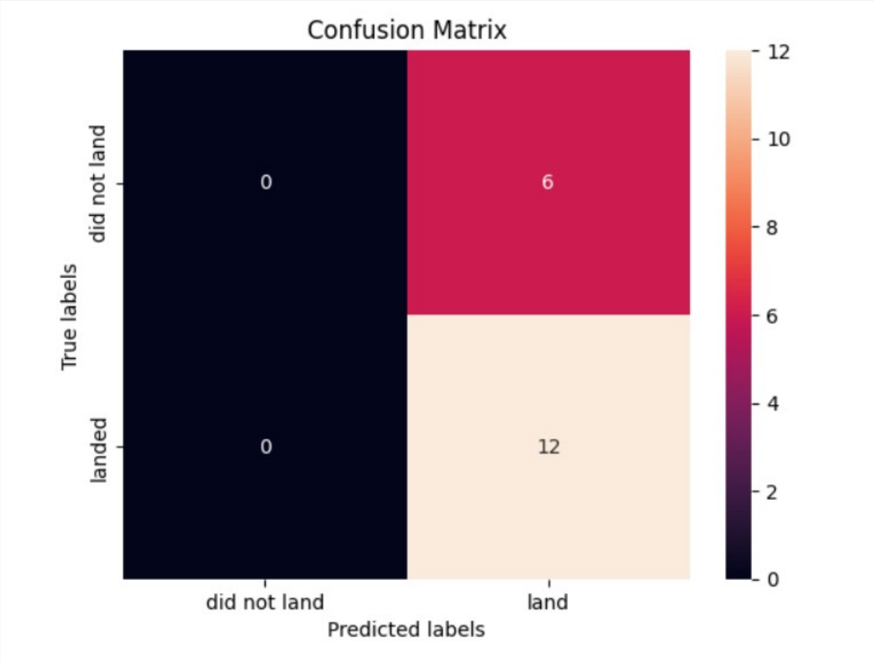
Classification Models (Accuracy Score, Test Score)

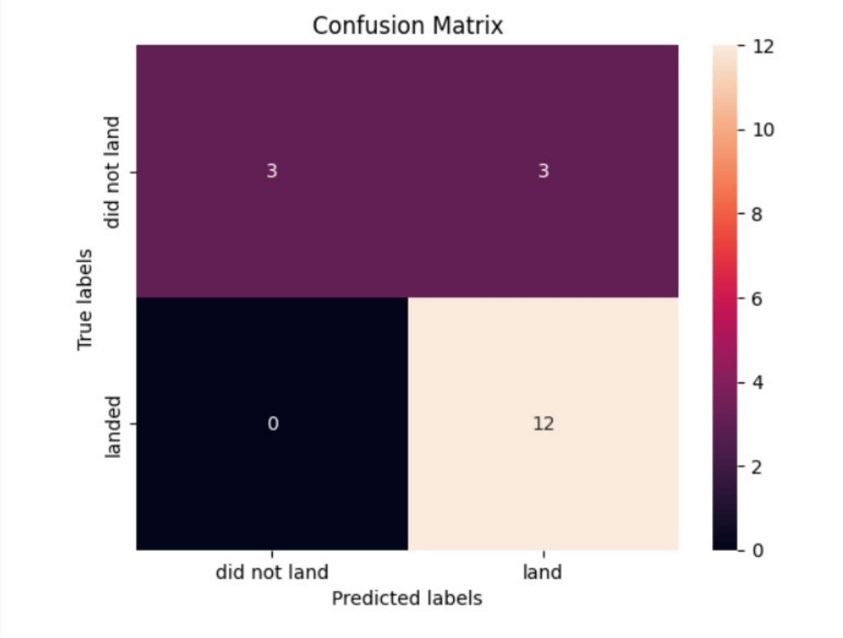




Although Decision Tree seems to be the best model based on the Accuracy score of the training data, the Test Accuracy Score indicates, that it is overfitting. SVM or KNN seems to be more balanced!

Confusion Matrix (Decision Tree, SVM)





Decision Tree:
True positive: 12
True negative: 6
False positive: 0
False negative: 0

The model is overfitting!

SVM:
True positive: 12
True negative: 3
False positive: 0
False negative: 3

No overfitting!

General Problem:
Dataset is to small!

Thank you!