IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Haydee M. Dominguez
October 10th, 2023

Outline

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

Executive Summary

- Our company SpaceY aims to provide affordable space travel. To be successful we must offer competitive prices and understand which attributes increase the chances of a successful first stage landing. Data from our top competitor, SpaceX, was collected from two sources, cleaned and prepared for a variety of analyses including: EDA with SQL and visualizations, interactive visual analytics with Folium and Plotly Dash, and machine learning classification models.
- In general we found, as expected, that landing success rate of the first stage increased with time, presumably due to improvements in experience and technology. Considering all launch sites, the payload mass between approximately 2000 to 5500 kg had the largest success rate. Considering all launch sites and all payload masses, booster version FT showed the highest number of successful landings. Our three best predictive models K Nearest Neighbor, Logistic Regression, and Support Vector Machine, returned an accuracy score of 0.83 on the test data and an acceptable confusion matrix.

Introduction

- Our company SpaceY aims to provide affordable space travel
  - Most providers are currently offering launches that cost over USD$165 million
  - SpaceX is currently advertising Falcon 9 rocket launches at a cost of USD$62 million
  - By reusing the first stage of Falcon 9 rockets, SpaceX can significantly reduce costs
  - For our enterprise to be successful, we must be able to offer competitive prices
- As a first step, our goals are:
  - To gather information about SpaceX launches
  - To explore the relationship between predictor variables and landing outcome
  - To determine if SpaceX will reuse the first stage of their launches
  - To determine the price of SpaceX launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We used the SpaceX REST API and web scraped a Wiki page for past SpaceX launch data
- Perform data wrangling
  - We inspected the data and created a target variable Y by converting the Outcome attribute into successful (1) and unsuccessful (0) landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using four classification models
  - For each model we split the data into training and testing sets, we built and obtained the best model parameters based on the train set, and evaluated the model using the test set

Data Collection

- Python code was used for data collection and analysis
- Two main sources of information were used to collect data:

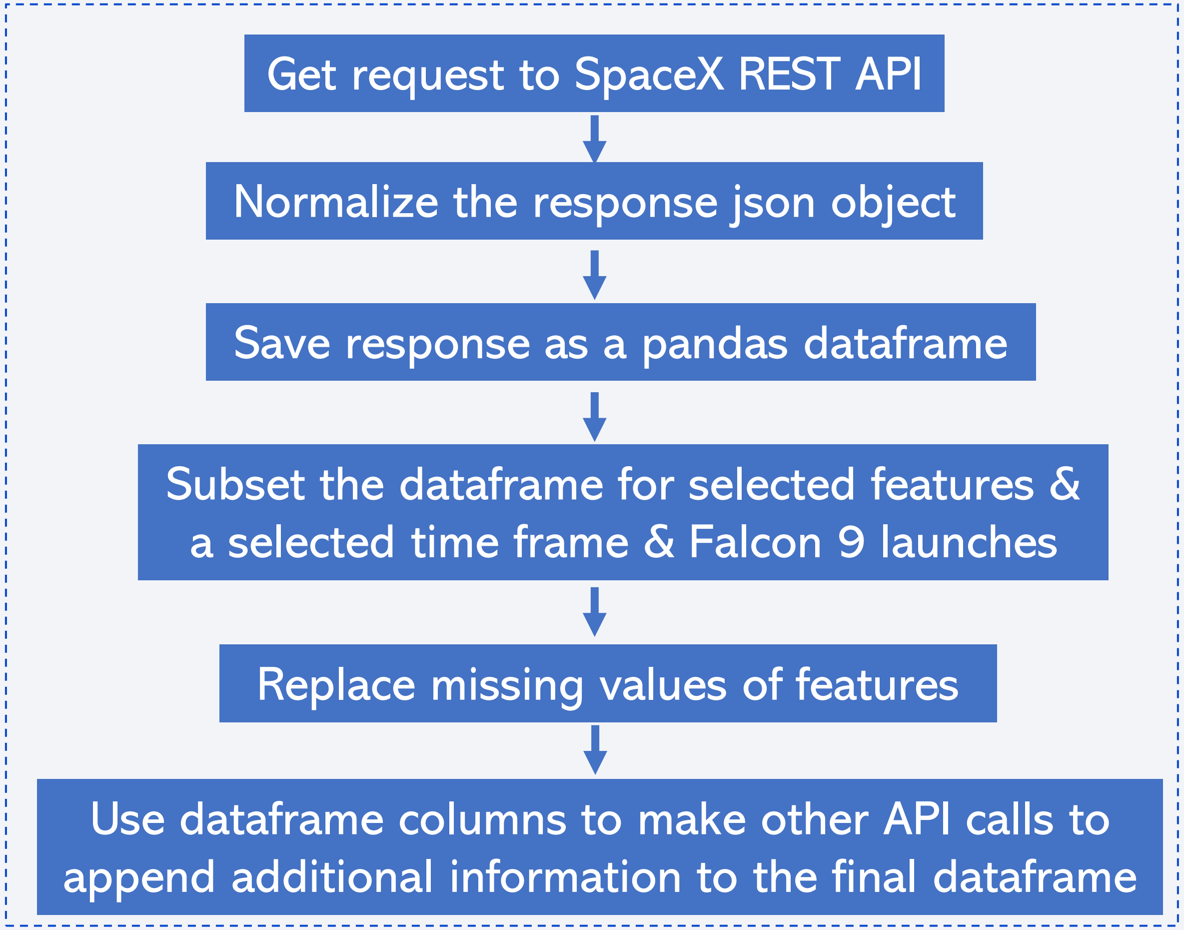
1) Get requests were made to several endpoints of the SpaceX REST API to collect data on launches including information about:
    - The rocket used, payload delivered, launch specifications, landing specifications, and landing outcome

2) Web scraping was used on a selected Wiki page:
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Falcon 9 launch records were extracted from an HTML table using the BeautifulSoup package
    - The rows of the relevant HTML table were scraped to save launch records into a python dictionary that was converted to a dataframe

Data Collection – SpaceX API

- Get requests were performed on different endpoints of the SpaceX REST API to collect information on past launches and save them into a pandas dataframe

- GitHub URL for the SpaceX API calls notebook: https://github.com/mtejo/ibmcapstone/blob/main/jupyter-labs-spacex-data-collection-api-HMDTCompleted.ipynb

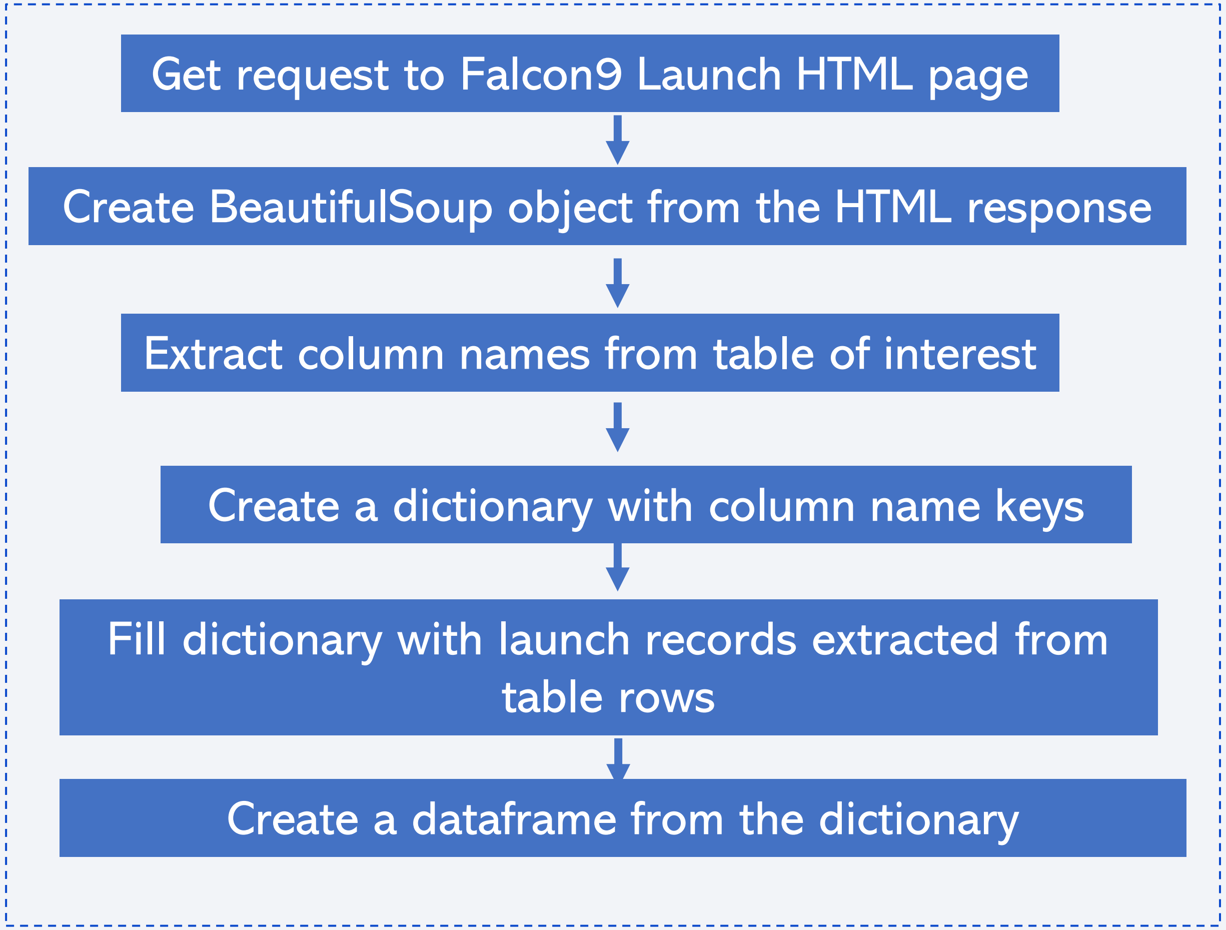


Create a BeautifulSoup object from the HTML response

Data Collection - Scraping

- A get request was performed on a selected Wiki page to extract launch records from an HTML table into a python dictionary and convert it to a dataframe

- GitHub URL for the web scraping notebook: https://github.com/mtejo/ibmcapstone/blob/main/jupyter-labs-webscraping-HMDTCompleted.ipynb



Data Wrangling

- Basic EDA was used to explore launches per site, per orbit, and launch outcome per orbit type
- The outcome feature was further processed to create our target variable ‘Class’, which classified successful and failed launches
- GitHub URL for the data wrangling notebooks: https://github.com/mtejo/ibmcapstone/blob/main/SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite-HMDTComplete.jupyterlite.ipynb

Calculate number of launches per launch site

↓

Calculate number of launches per orbit type

↓

Calculate mission outcome per orbit type

↓

Explore the different landing outcomes

↓

Create a landing class variable to separate successful (1) vs unsuccessful (0) landings

↓

Append the variable as the target feature ‘Class’ to the launch dataframe

EDA with Data Visualization

- Three types of charts were used to explore the data:
  - Scatter plots to observe relationships between the following variables and the color-coded landing outcome: Flight Number vs. Launch Site, Payload vs. Launch Site, Flight number vs. Orbit type, and Payload vs. Orbit type.
  - A bar chart to visually inspect the relationship between success rate and orbit type
  - A line chart to observe trends in average launch success rate per year
- Feature engineering was achieved by one-hot encoding (creating dummies) of the categorical columns followed by their conversion to float type
- GitHub URL for EDA with data visualization notebook: https://github.com/mtejo/ibmcapstone/blob/main/SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb-HMDTComplete.jupyterlite.ipynb

EDA with SQL

- SQL queries were performed to:
    - Obtain the names of the unique launch sites
    - Display 5 records where the launch site began with the string ‘CCA’
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display the average payload mass carried by booster version F9 v1.1
    - List the date when the first successful landing outcome in ground pad was achieved
    - List the names of the boosters which had success in drone ship and had payload mass greater than 4000 but less than 6000
    - List the total number of successful and failed mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass
    - List records with month, failure landing outcomes in drone ship, booster versions and launch_site for the months of 2015
    - Rank the count of landing between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL for EDA with SQL notebook: https://github.com/mtejo/ibmcapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite-HMDTComplete.ipynb

Build an Interactive Map with Folium

- Using Folium, a map was created centered around the NASA Johnson Space Center in Houston, Texas, and a circle object was created around the Space Center
- Circle objects with pop up labels were added to the map to locate the different launch sites based on their latitude and longitude coordinates
- Markers were created with a MarkerCluster object to show the result (success or failure) for all launch records, and allow the user to zoom in and out of the results
- Line objects were created to showcase the distance from launch sites to different points of interest such as the nearest coastline point, railway, highway or city
- GitHub URL for interactive map with Folium: https://github.com/mtejo/ibmcapstone/blob/main/SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location-HMDTCompleted.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A SpaceX Launch Records Dashboard was created with Plotly Dash to allow for an interactive exploration of the data
- A dropdown menu was added to allow the user to see results for all launch sites combined, or for a particular launch site
- A pie chart was added to show the total successful launches per site when all sites are selected, or the percent of successful and unsuccessful launches per site when one site is selected
- A scatter plot of Payload Mass and Class was included, additionally showing the booster versions in different colors, to explore the relationship between these variables
- A Payload Range slicer was added to allow the user to filter the results of the scatter plot within selected limits of Payload Mass
- GitHub URL for Plotly Dash lab: https://github.com/mtejo/ibmcapstone/blob/main/spacex_dash_app-HMDTCompleted.py

Predictive Analysis (Classification)

- The dataframe was prepared for predictive analysis
- Four modelling algorithms were selected: logistic regression, support vector machine, decision trees, and k nearest neighbors
- For each model, the training set was used to train the model and obtain the best model parameters. The model was then evaluated using the test data.
- Models were compared based on their accuracy score and confusion matrix results
- GitHub URL for predictive analysis lab: https://github.com/mtejo/ibmcapstone/blob/main/SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5-HMDTComplete.jupyterlite.ipynb

Create Target Variable (Y) and feature (X) dataframes

↓

Standardize the features in the X dataframe

↓

Split the dataframes X and Y into training and test data

↓

Create a model object and a dictionary of parameters

↓

Use Gridsearch to fit the model on the training set and save the model with the best parameters

↓

Calculate the accuracy of the model on the test data, plot the confusion matrix, and compare results

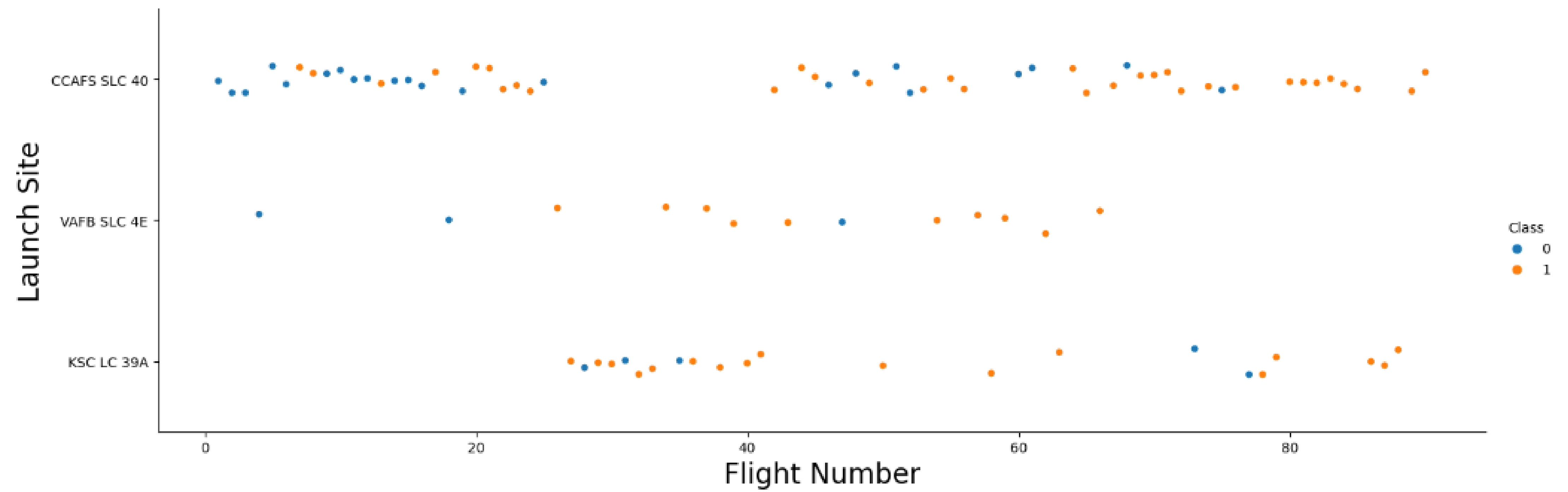
Results

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

Section 2

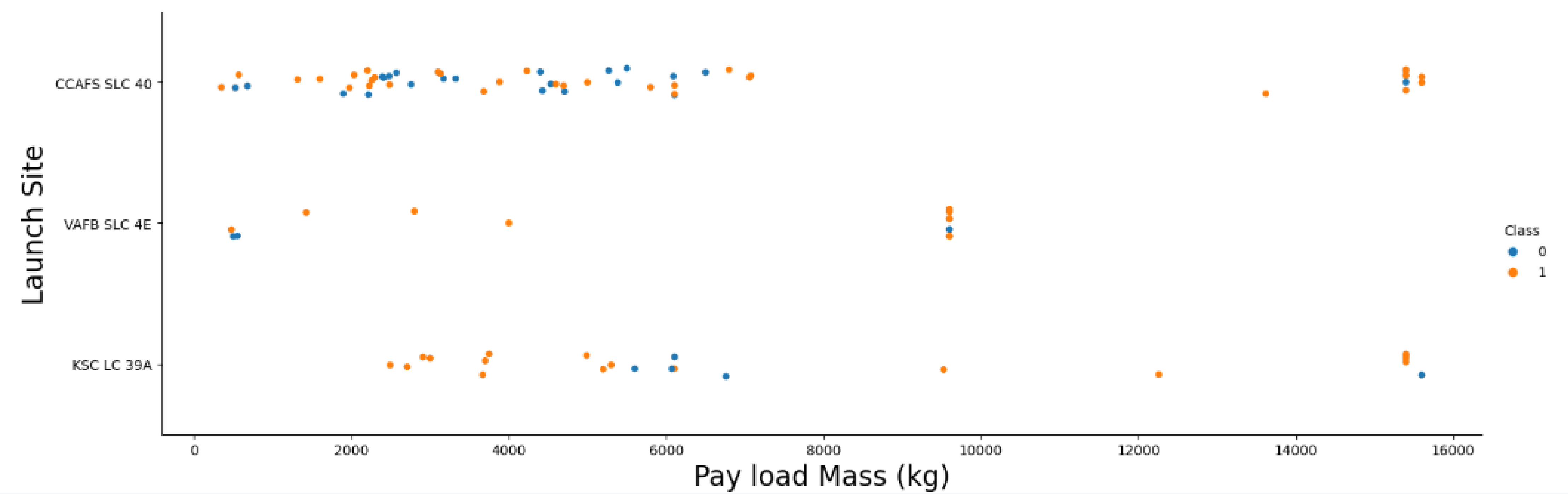
Insights drawn from EDA

Flight Number vs. Launch Site



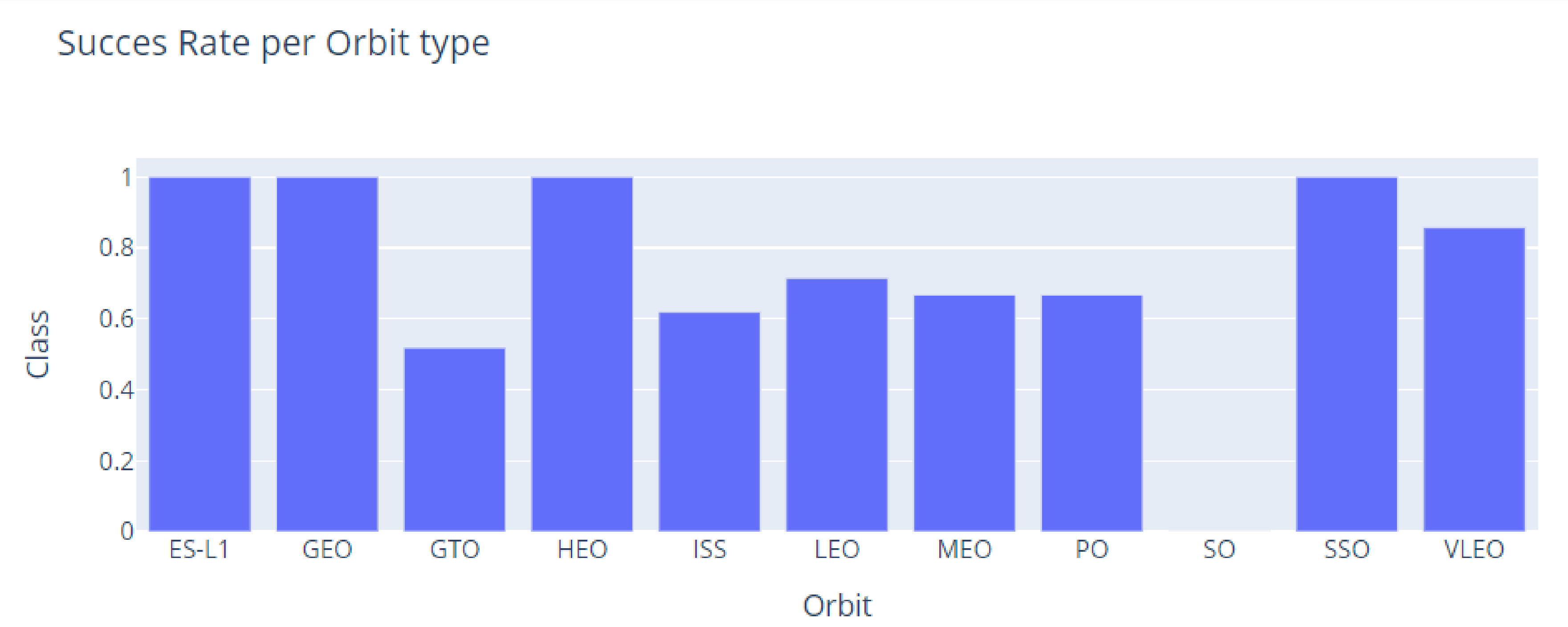
- A similar patter was observed for the three launch sites: as flight numbers increased, there tended to be more successful landings
- The CCAFS SLC-40 launch site had most of the earliest flights (flight numbers 1-25), the majority of which had an unsuccessful landing. At the KSC LC-39A site, flights started the latest

Payload vs. Launch Site



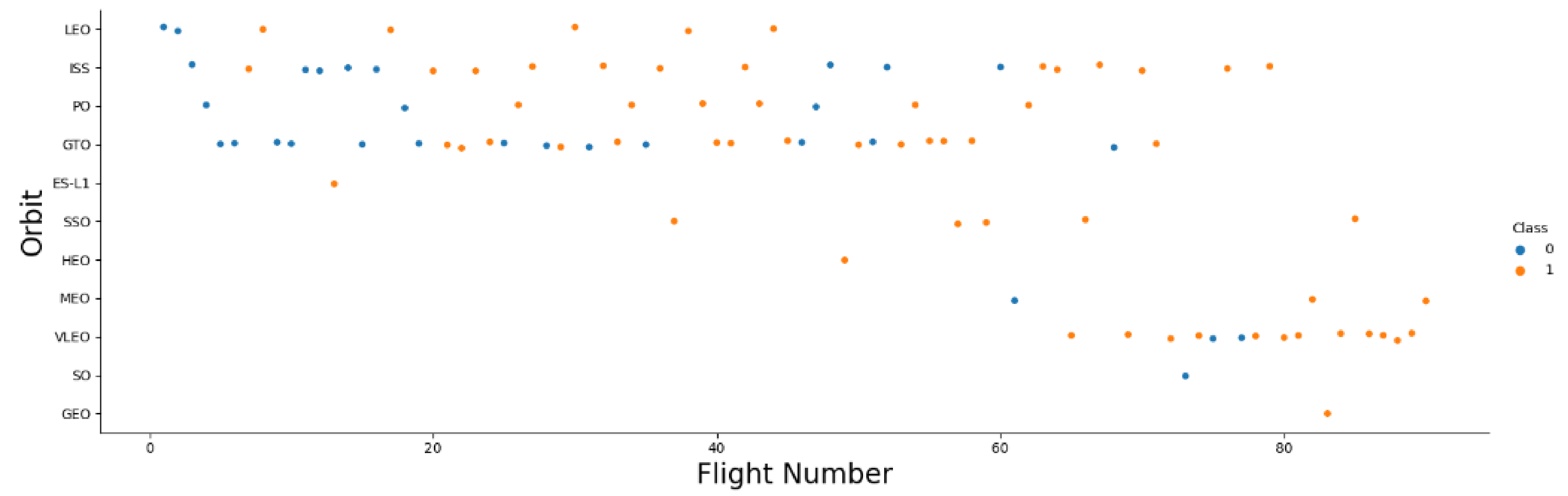
- A similar patter was observed for the three launch sites: as payload mass increased, there tended to be more successful landings
- The VAFB-SLC launch site did not have launches with payload mass heavier than 10000 kg

Success Rate vs. Orbit Type



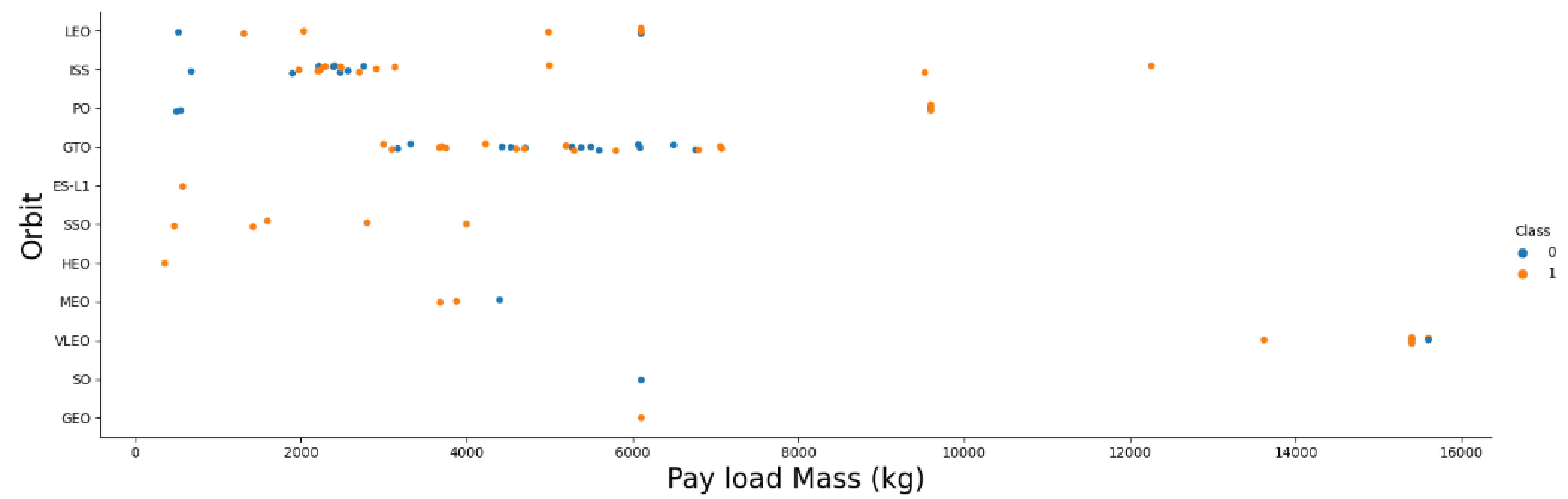
- There was no apparent relationship between the altitude of the orbit and landing success
- The orbit types with the greatest success rates were ES-L1, GEO, HEO and SSO

Flight Number vs. Orbit Type



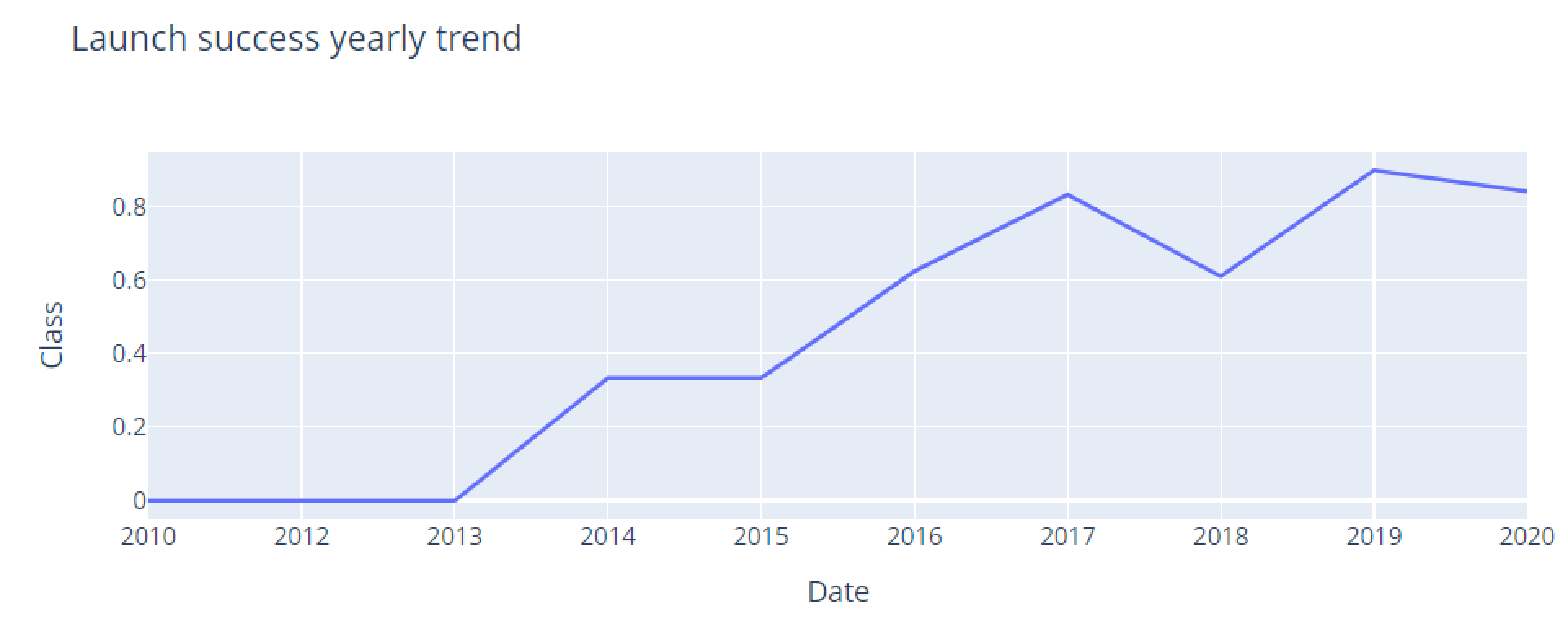
- For some orbits like LEO and PO success rate tended to increase with flight number, but for other orbit types like ISS and GTO no relationship was evident
- The orbit types with the greatest success rates ES-L1, GEO, HEO and SSO had only 5 or fewer flights and the flights were fairly recent (flight number greater than 40)

Payload vs. Orbit Type



- Success rate for orbit types LEO, ISS and PO increased with payload mass, but for the GTO orbit no relationship was evident
- The orbit types with the greatest success rates ES-L1, GEO, HEO and SSO had payload masses of around 6000 kg or less

Launch Success Yearly Trend



- Since 2013 there has been an overall increase in average success rate per year

All Launch Site Names

- Find the names of the unique launch sites
- Falcon 9 rockets are launched by SpaceX from the four following launch sites:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- The following five records belong to the launch site CCAFS LC-40:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- The total payload mass that has been carried by SpaceX rockets for all launches for their NASA (CRS) customer is over 45000 kg:

Customer	Total Payload Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass that has been carried by the booster version F9 v1.1 is just over 2900 kg:

Booster_Version	Average Payload Mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- Find the date of the first successful landing outcome on ground pad
- The date when the first successful landing outcome on ground pad was achieved was Dec 22nd, 2015:

MIN(Date)
2015-12-22

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
- The following four booster versions carrying a payload mass larger than 4000 kg but smaller than 6000 kg have successfully landed on drone ships:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failed mission outcomes
- The Mission Outcome of SpaceX launches has been largely successful except for one case of a Failure (in flight):

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The 12 booster versions listed on the right-side table have all carried the maximum payload mass:

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

2015 Launch Records

- List the records displaying the month, failure landing_outcomes in drone ship, booster version and launch_site for the months of 2015

- In April and October 2015 two booster versions launched from the CCAFS LC-40 launch site failed to land in a drone ship:

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Between 2010-06-04 and 2017-03-20, there were ten cases where landing was not attempted. There were five cases each of successful landings in ground pads and drone ships, and five cases of failed landings in drone ships.
- Other landing outcomes had 3 or fewer cases.

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

Section 3

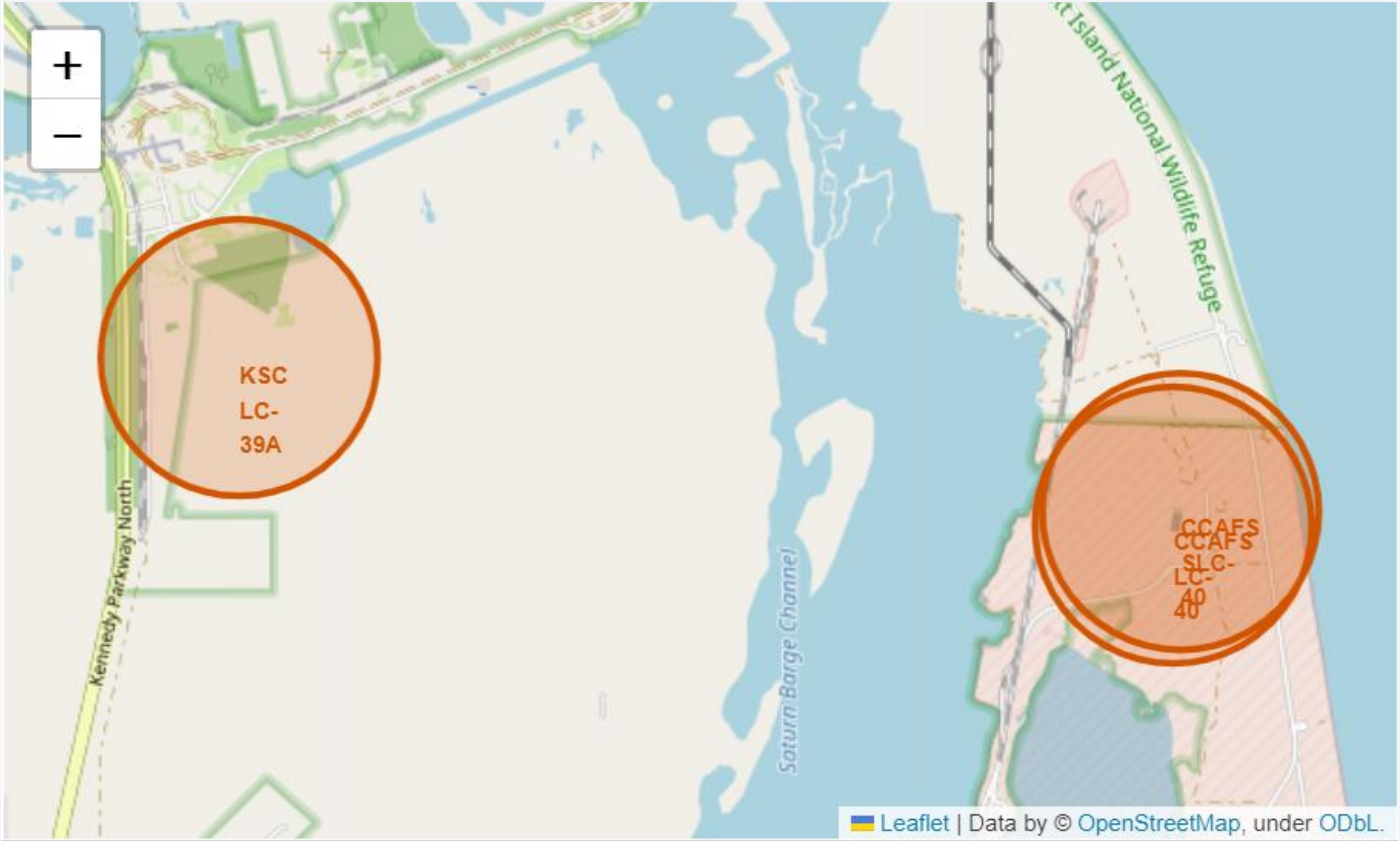
Launch Sites Proximities Analysis

Launch Site Locations

- All launch sites are in close proximity to the coast
- Launch site VAFB SLC-4E is located in the Pacific coast of the USA
- The remaining three launch sites are located in the Atlantic coast of the USA

Launch Sites in the Atlantic coast of the USA

- Launch sites CCAFS LC-40 and CCAFS SLC-40 are in close proximity of each other



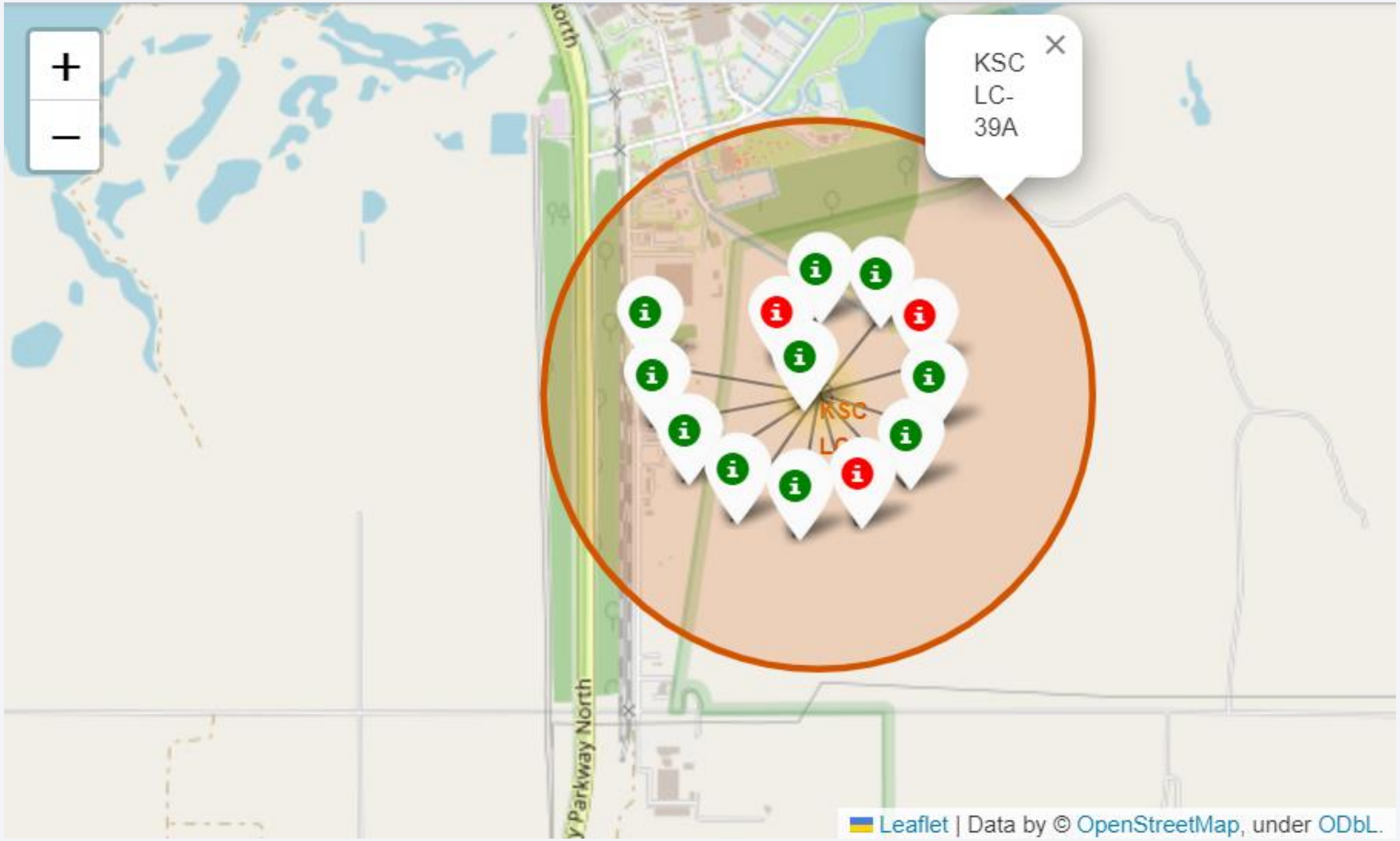
Location and Number of launches

- Most launches have taken place at launch sites in the Atlantic coast of the USA

- Only 10 launches have been recorded in the Pacific coast launch site

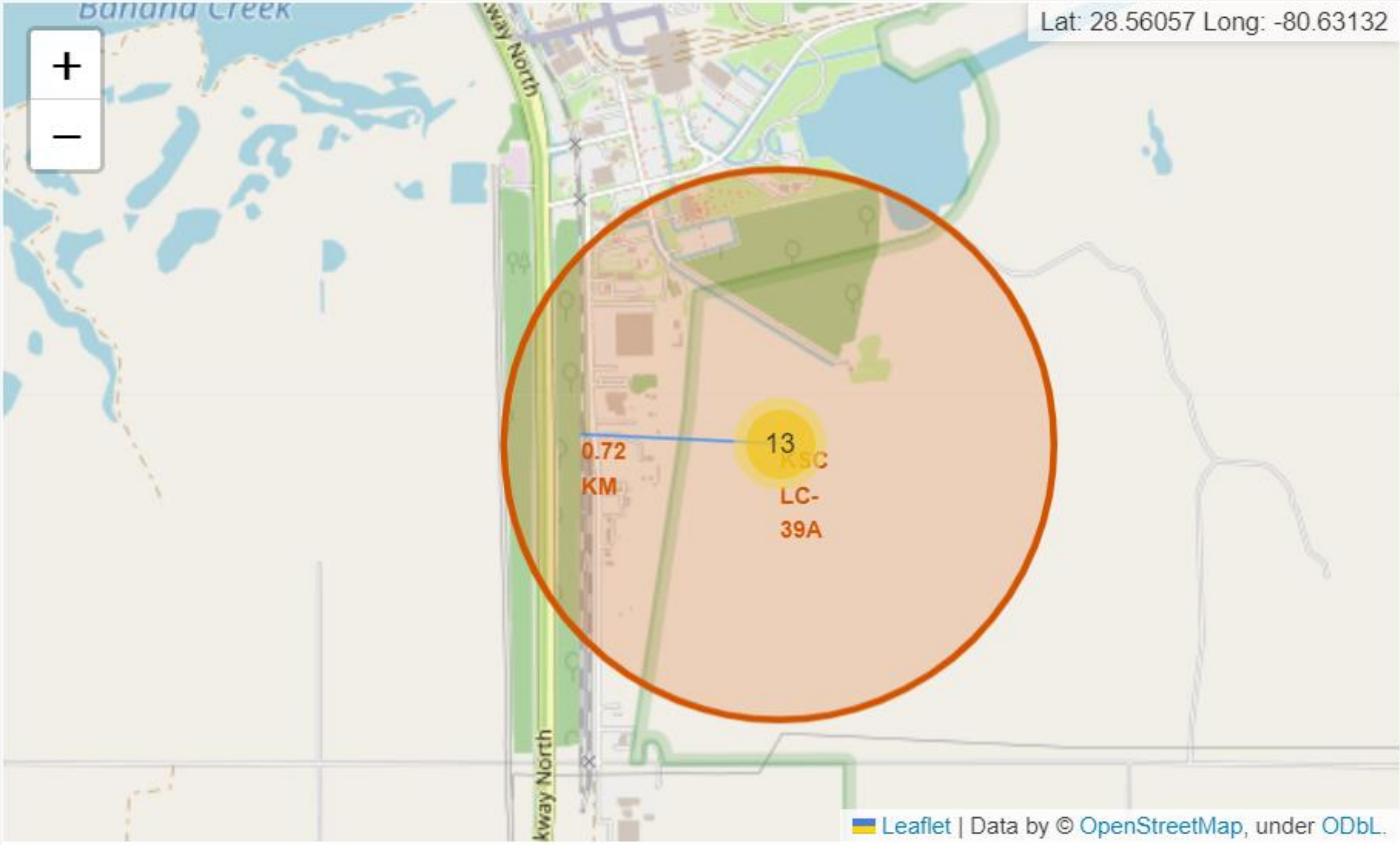
Launch Outcomes at Site KSC LC-39A

- 10 out of the 13 launch records at the KSC LC-39A launch site had a successful landing



Proximities of Launch Site KSC LC-39A

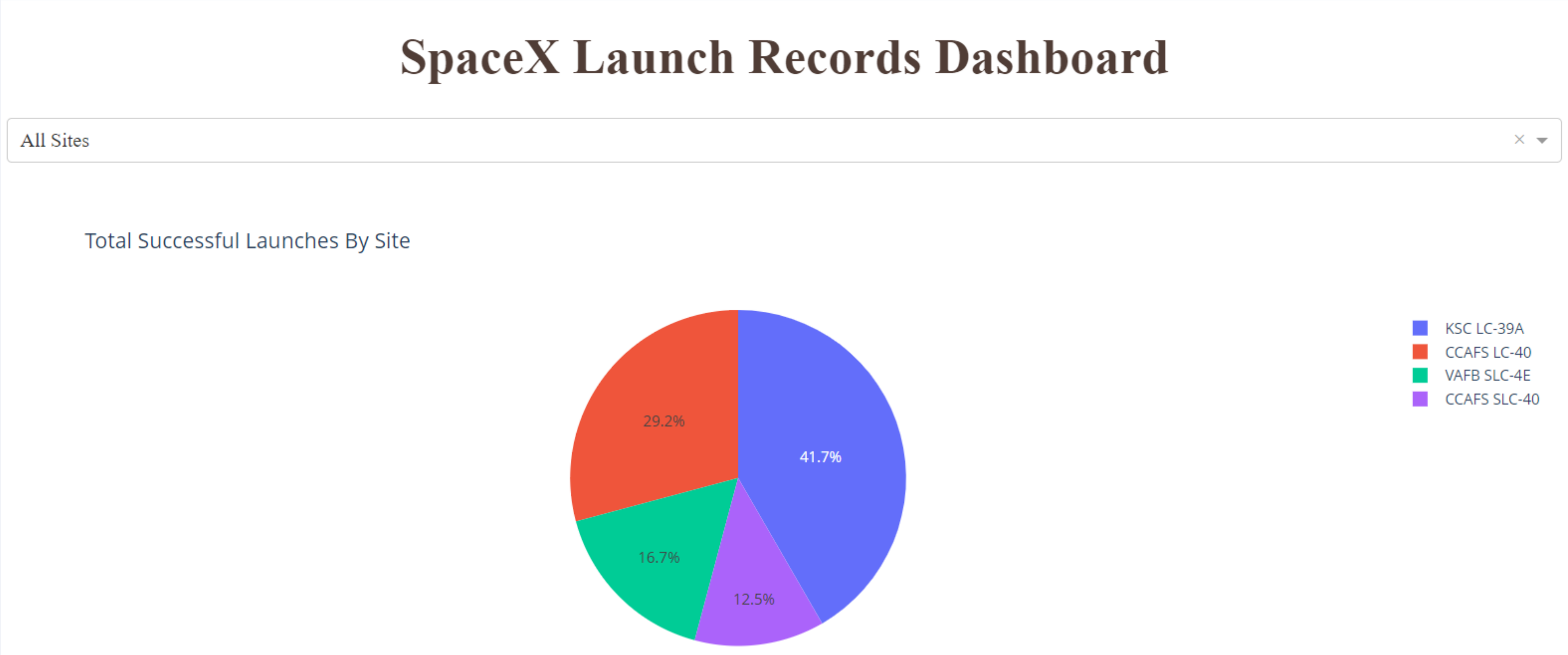
- All launch sites are close to the coast and far away from city centers
- The KSC LC-39A launch site is 0.72 km away from the nearest railway



Section 4

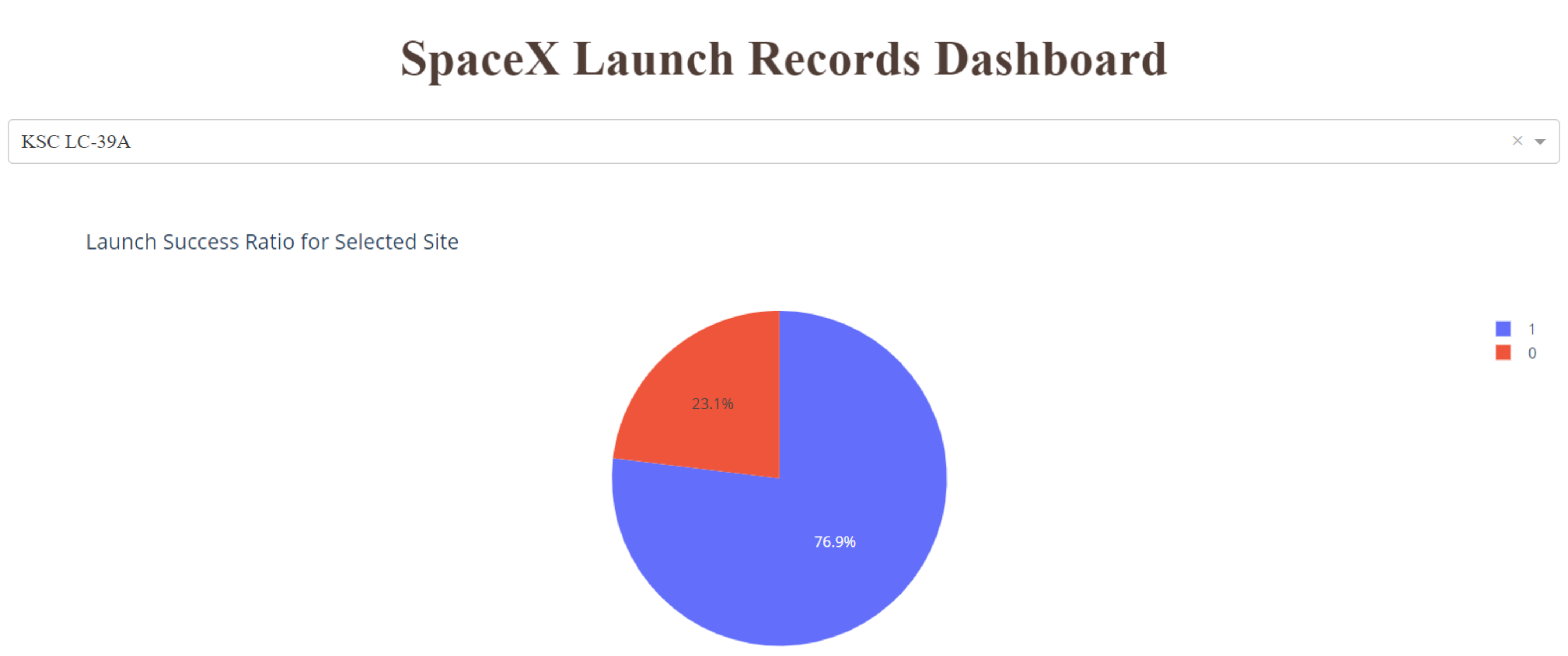
Build a Dashboard with Plotly Dash

Proportion of Launch Success for All Sites



- The launch site with the greatest proportion of successful launches is the KSC LC-39A site with 41.7%, followed by the CCAFS LC-40 site with 29.2%

Launch Site with Highest Success Rate



- Launch site KSC LC-39A had the highest launch success ratio of 76.9% (10 successful launches out of 13 in total)

Correlation between payload and launch outcome (1)

- Considering all launch sites, the payload mass between approximately 2000 to 5500 kg had the largest success rate

Correlation between payload and launch outcome (2)

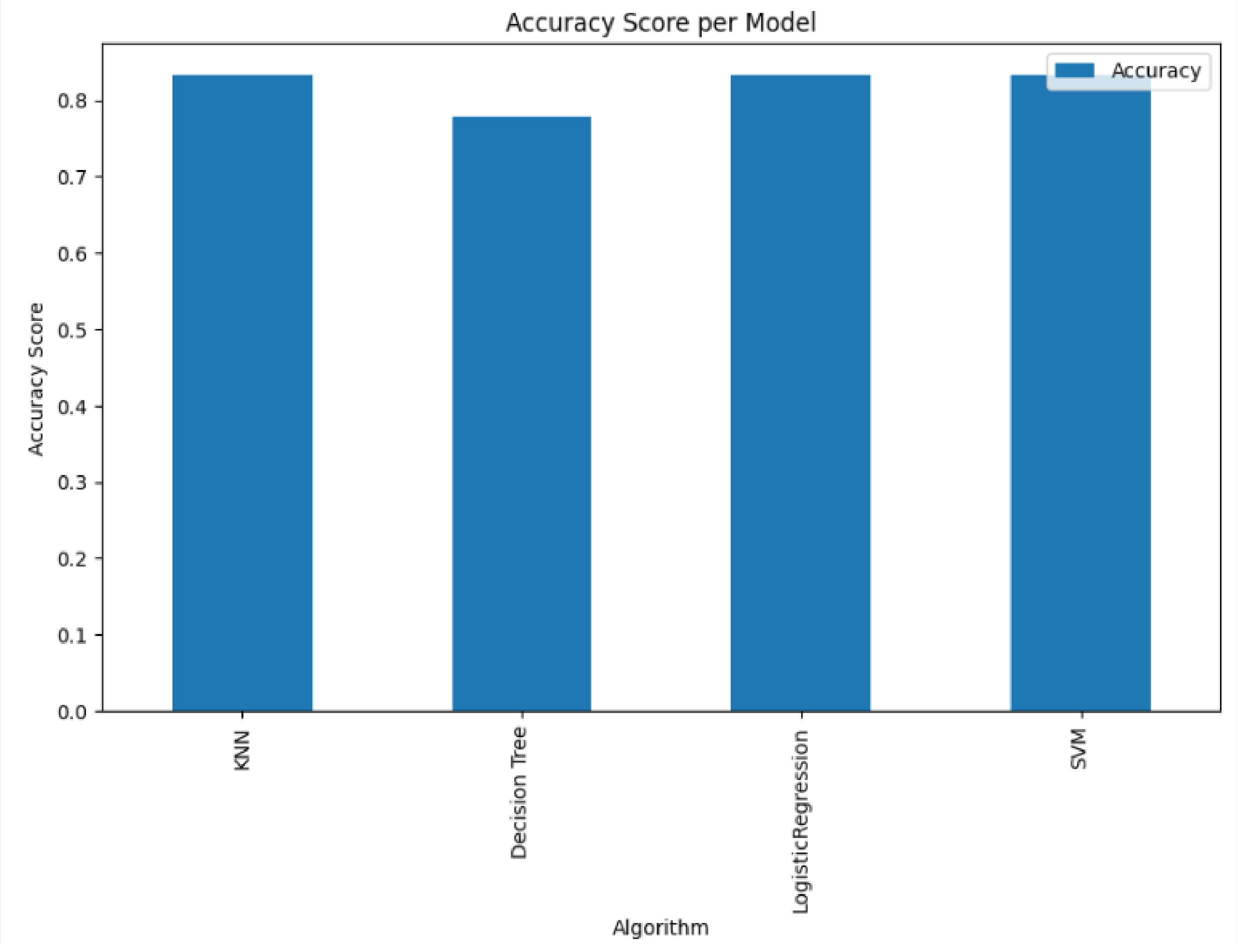
- Considering all launch sites and all payload masses, booster version FT showed the highest number of successful landings, while booster version v1.1 showed the highest number of failed landings

Section 5

Predictive Analysis (Classification)

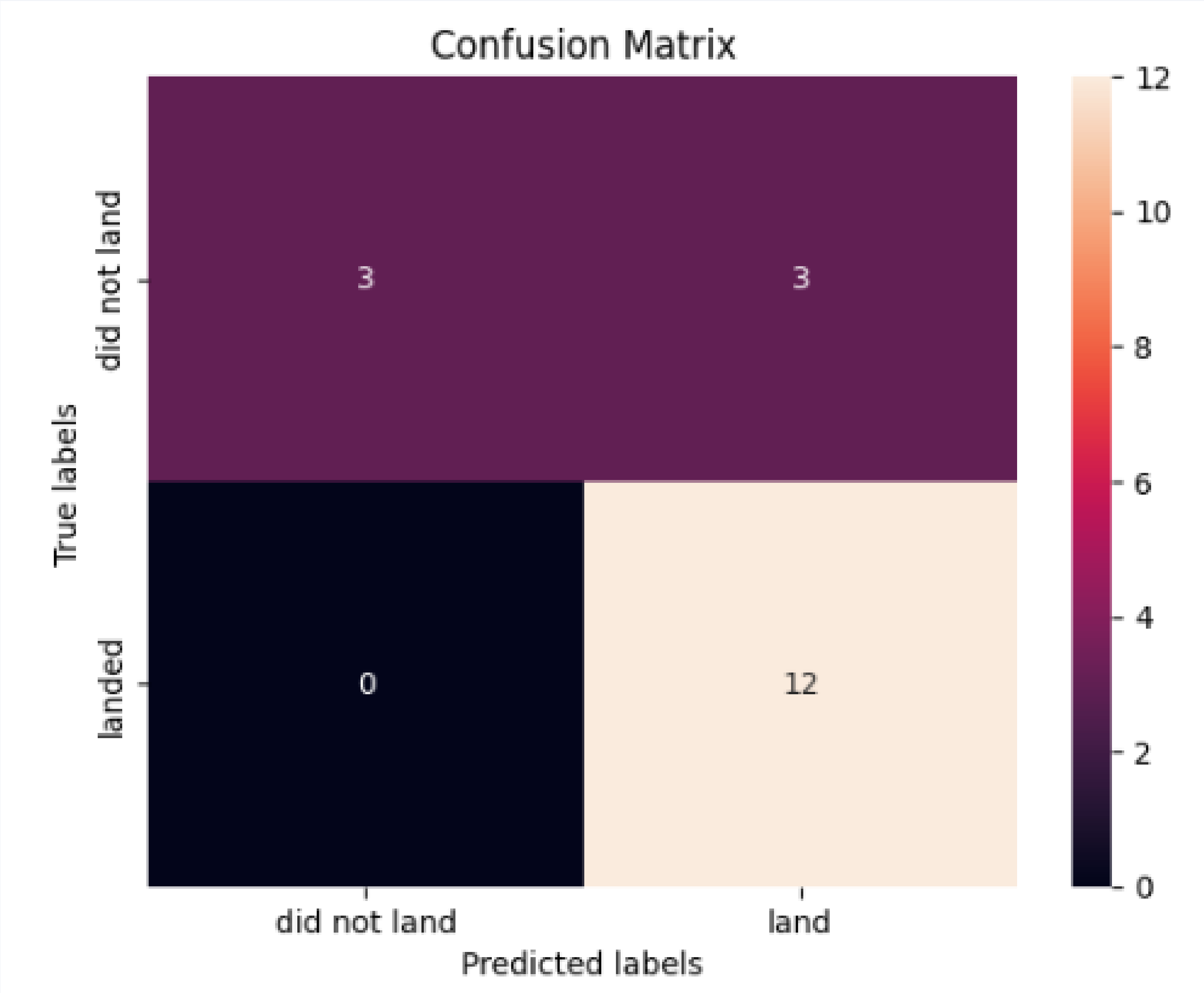
Classification Accuracy

- The K Nearest Neighbor, Logistic Regression and Support Vector Machine models resulted in the same accuracy score of 0.8333 on the test data
- The Decision Tree model had the lowest performance with an accuracy score of 0.7777 on the test data



Confusion Matrix

- The K Nearest Neighbor, Logistic Regression and Support Vector Machine models showed the same results for the confusion matrix
- All successful landings were correctly classified, but three of the failed landings were incorrectly classified as successful



Conclusions

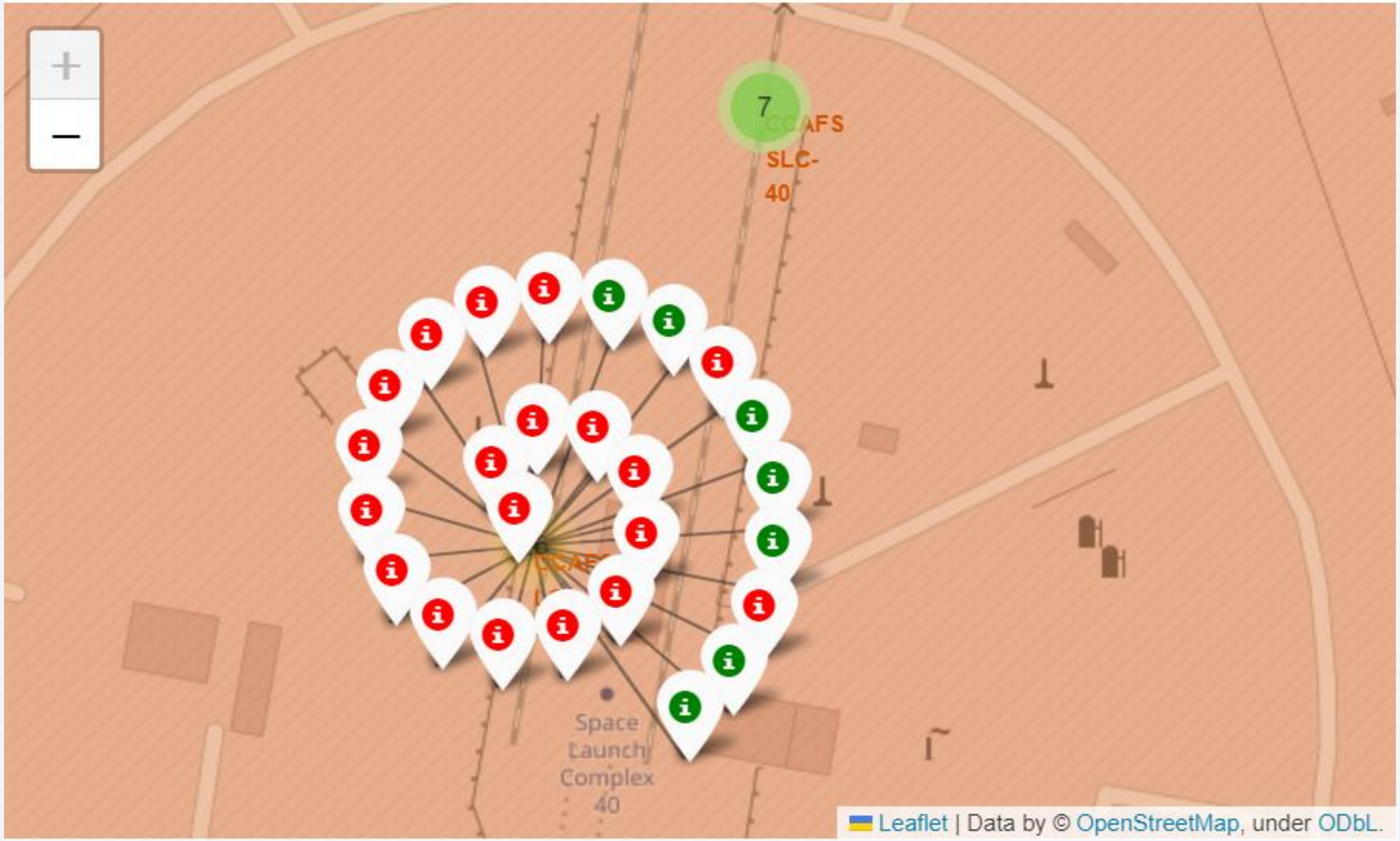
- As flight numbers increased there were more successful landings, due likely to the experience gained and improvements made with time
- In fact, the average success rate per year clearly increased between 2013 -2020
- Most of the earliest flights (flight numbers 1-25), were conducted at the CCAFS SLC-40 launch site and the majority of them had unsuccessful landings
- Launch site KSC LC-39A had the highest launch success ratio of 76.9%, and flights in this site started later than in other launch sites
- There was no clear relationship between orbit altitude and success rate; the orbit types with higher success rates had few and fairly recent flights with payload masses of around 6000 kg or less
- Considering all launch sites, the payload mass between approximately 2000 to 5500 kg had the largest success rate
- Considering all launch sites and all payload masses, booster version FT showed the highest number of successful landings, while booster version v1.1 showed the highest number of failed landings
- The K Nearest Neighbor, Logistic Regression and Support Vector Machine models resulted in the same accuracy score of 0.8333 on the test data, and the same confusion matrix results

Appendix

- Successful and failed launches at the CCAFS SLC-40 site

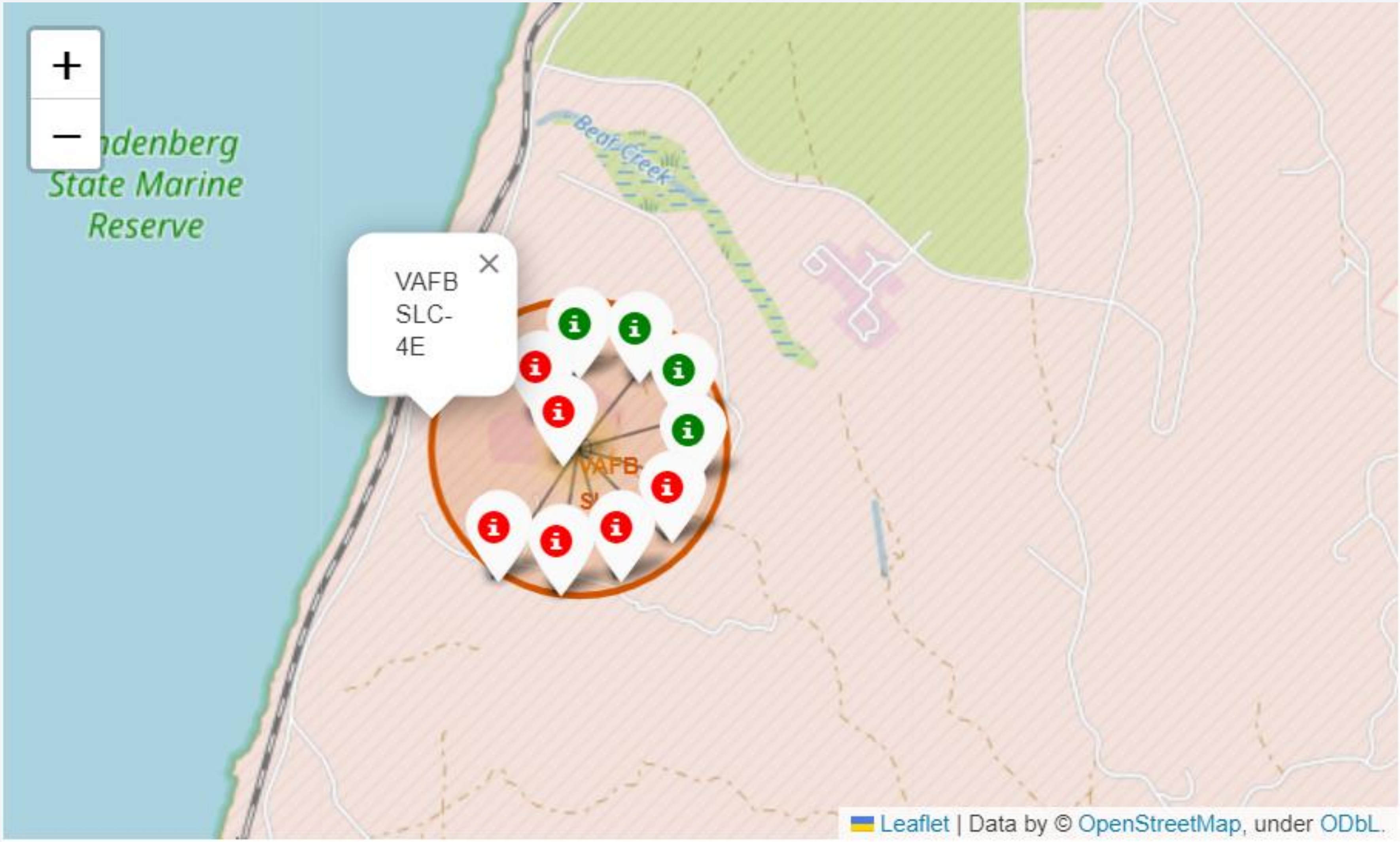
Appendix

- Successful and failed launches at the CCAFS LC-40 site

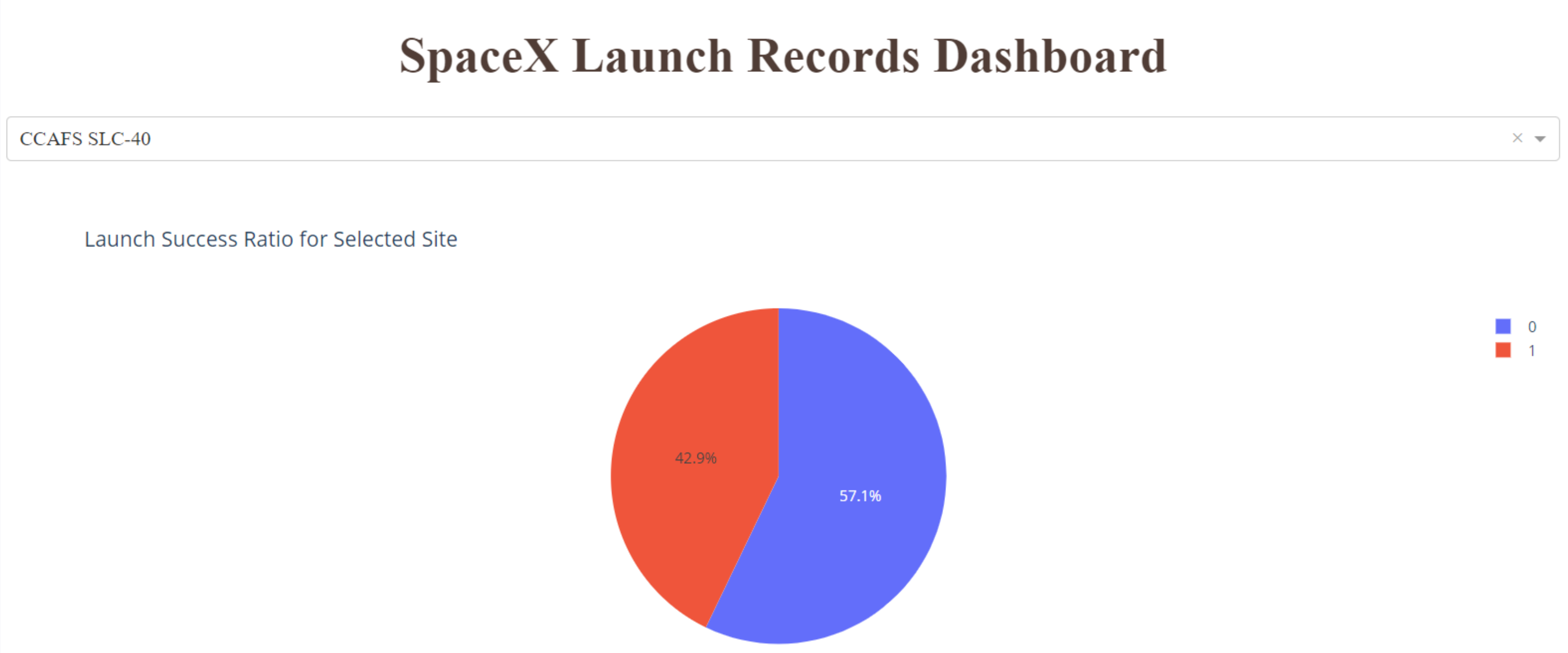


Appendix

- Successful and failed launches at the VAFB SLC-4E site

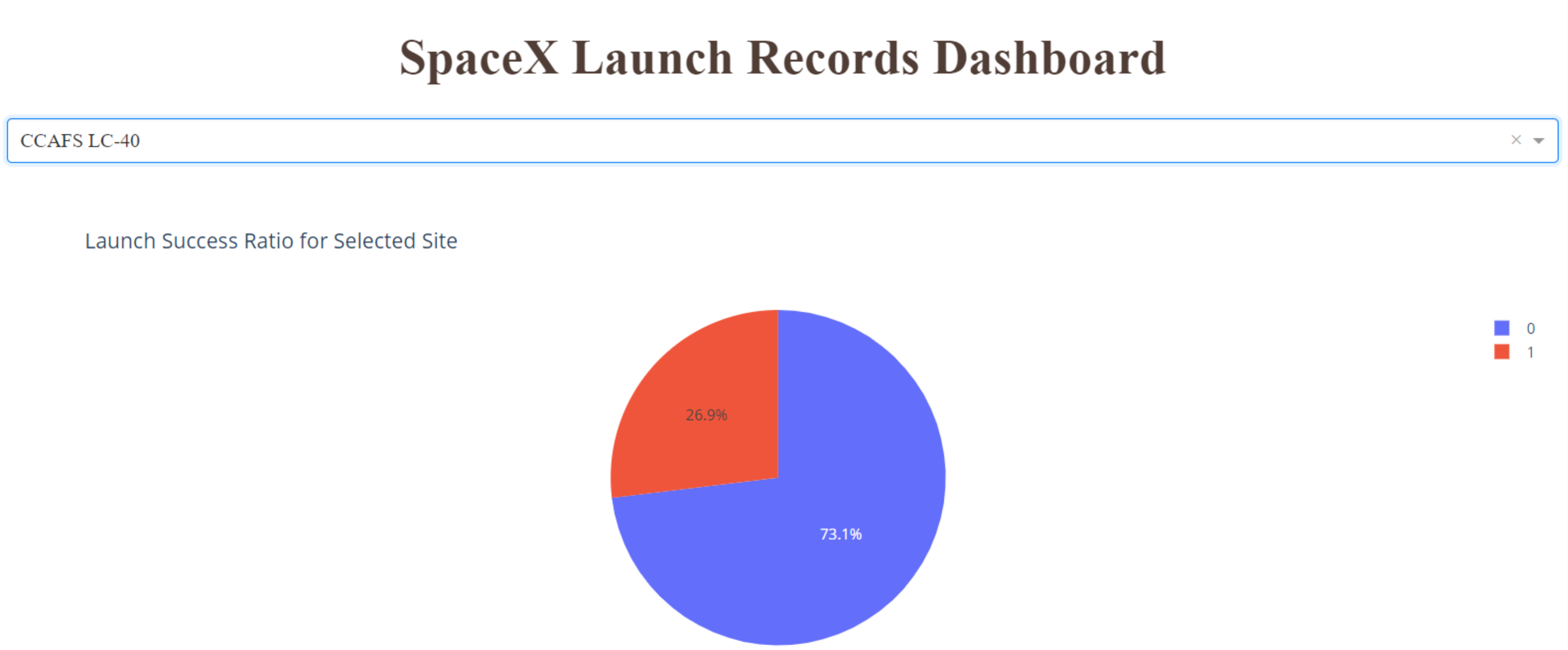


Appendix



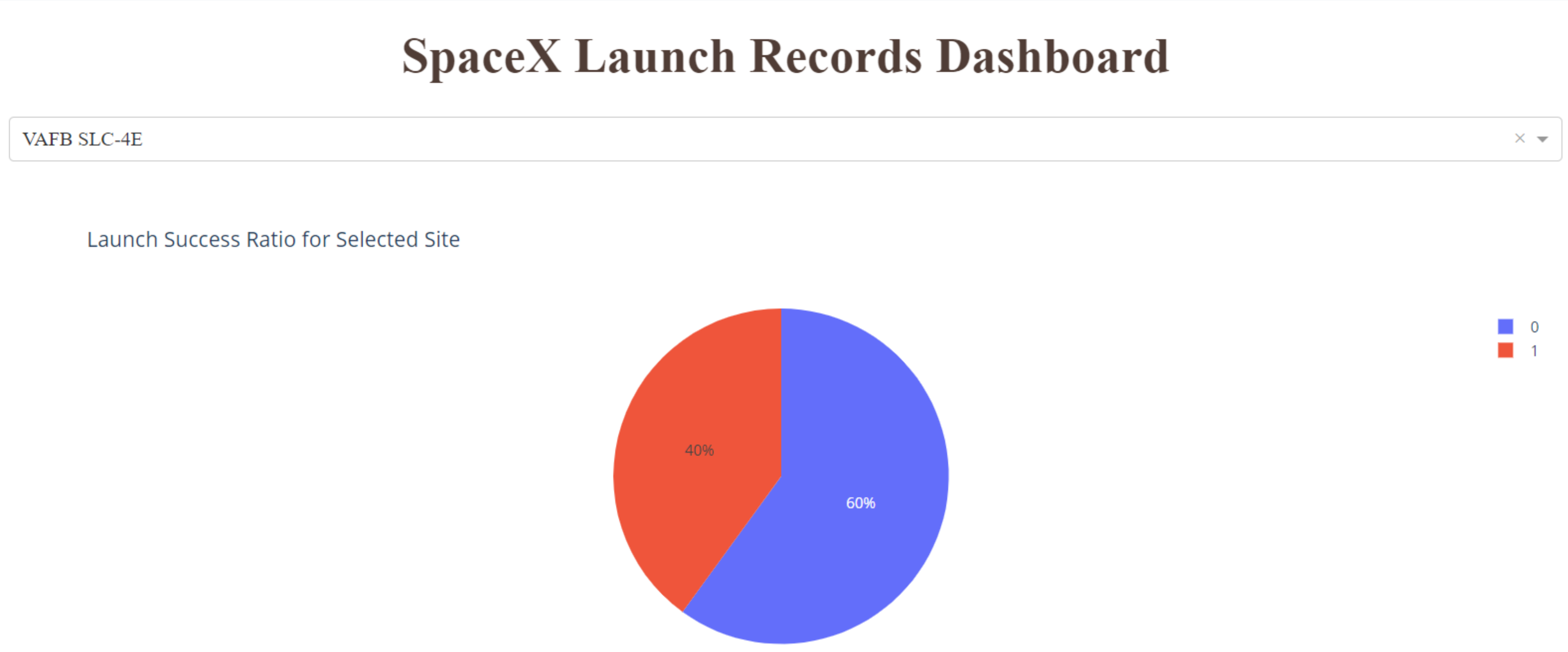
- The launch success ratio for site CCAFS SLC-40 was 42.9%

Appendix



- The launch success ratio for site CCAFS LC-40 was 26.9%

Appendix



- The launch success ratio for site VAFB SLC-4E was 40%

Appendix

- The accuracy scores on the test data for the four classification models were the following:

	Accuracy
Algorithm	
KNN	0.833333
Decision Tree	0.777778
LogisticRegression	0.833333
SVM	0.833333

Thank you!