IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sahil Gobade
June 24, 2023

Outline

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

Executive Summary

- To analyze the data, the following methodologies were used:
  - Data Collection methods: Web scraping and API Methods
  - Exploratory Data Analysis: Data Visualization and Interactive Visual Analytics
  - Machine Learning Methods: Logistic Regression, KNN, SVM, Decision Tree
- Summary of all results
  - Public sources are a great way to collect usable data for competitor analyses.
  - EDA helped to identify the features that would best predict a successful launching.
  - Machine Learning showed that the Decision Tree Model is the best classification model to predict important driving characteristics.

Introduction

- We are Space Y. We are looking at bidding against Space X for a future rocket launch. To do so, we are looking determining the cost of a launch when the first stage lands and is reusable.
- The problems we would like to tackle in this analysis is how much would a successful total launch cost and where would be the best launch site.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was scraped from the SpaceX API and then cleaned to better fit this project.
- Perform data wrangling
  - Doing exploratory data analysis (EDA) and then determining training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four Models used: Logistic Regression, SVM, Decision Tree, KNN

Data Collection

- Data was collected using two methods
  - SpaceX API
  - Webscraping from Wikipedia

- Each Data Collection had three major steps, as seen with the flow charts on the next two slides.

Data Collection – SpaceX API

Request API and parse the SpaceX launch data → Filter data to only include Falcon 9 launches → Deal with any missing values that may present

- Data was grabbed from SpaceX's public API. Data from the public API was put through the process above before moving to exploratory analysis.

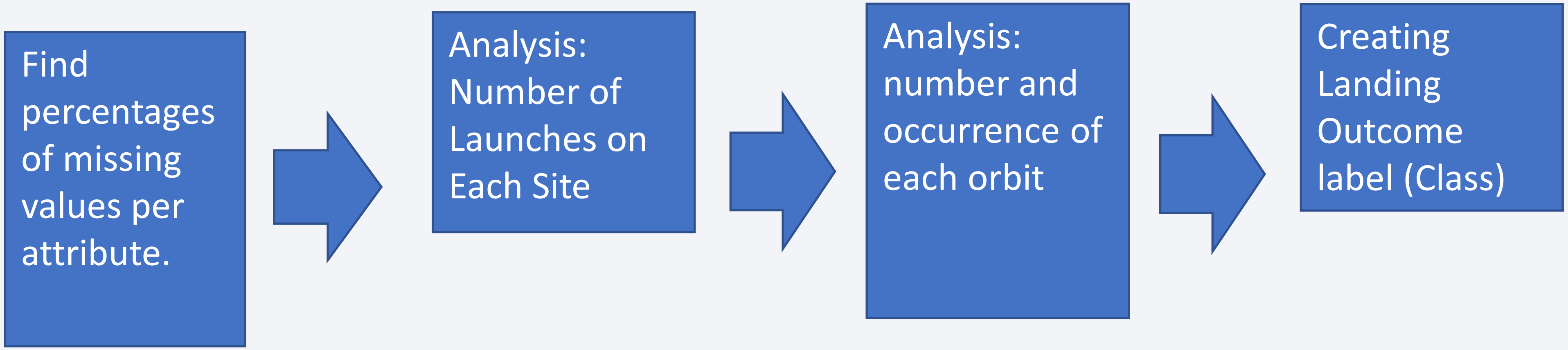
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping

Request Falcon 9 Launch page from Wikipedia → Extract all Column/Variable names from the HTML table header → Create data frame by parsing the launch HTML tables

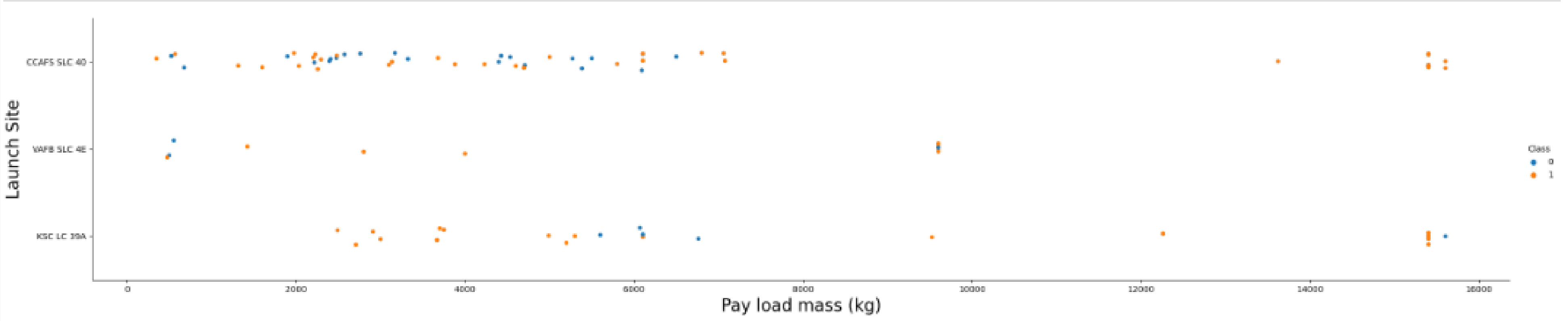
- SpaceX launch data was obtained through Wikipedia. Data scraped was put through the process above before moving to exploratory analysis.
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



- Data Wrangling was completed by doing exploratory data analysis (EDA) and then determine training labels.
- The EDA done is included in the flow chart above.
- Source Data: https://github.com/bubblypanda/skillsnetworklab/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization



- Scatterplot charts were created for every relationship that was pertinent including the following:
    - Flight Number and Payload Mass
    - Flight Number and Launch Site
    - Payload Mass and Launch Site (included above)
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The following was found utilizing SQL queries:

- Unique Launch Sites: CCAFS LC-40, VAFB SLC-4E, KSCLC-39A, CCAFS SLC-40
- 5 Records where launch sites begin with the string ("CCA%")
- Total Payload Mass Carried by Boosters launched by NASA (CRS): 45,596
- Average Payload Mass Carried by Booster Version F9 v1.1: 2,928.4
- Date when the first successful landing outcome in ground pad was achieved: 01-05-2017
- Boosters that have success in Drone Ship and have Payload Mass between 4,000 and 6,000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2
- Total Successful Missions: 100, Total failure Missions: 1
- The 12 Boosters that have carried the max payload mass.
- The months where two drone ship failures occurred at Launch Site CCAFS LC-40: January and April
- The rank of successful landing_outcomes between 04-06-2010 and 20-03-2017 descending: success, success (drone ship) and success (ground pad)
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In Folium, we did the following:
  - Created Circle markers for each launch site
  - Created Marker Cluster for all launch records (green markers for successful, red for unsuccessful)
  - Added a line for closest city to a particular launch site.
  - Added a line for closest coast to a particular launch site.
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash was used to create a Dashboard with a breakout of percentages of successful launches that occur at each site, the percentages of total launches performed at each site in a pie chart, and a scatterplot of the different launches, their respective payload mass in kilograms, and whether they were successful (class = 1) or unsuccessful (class = 0).
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/Build%20a%20Dashboard%20Application%20with%20Plotly%20Dash%20skill%20lab%20network

Predictive Analysis (Classification)

Perform each of the four classification models. → Create Confusion Matrix based on Each model. → Compare Accuracy of each Model on the Test data set → Select Best Model Method for this Analysis

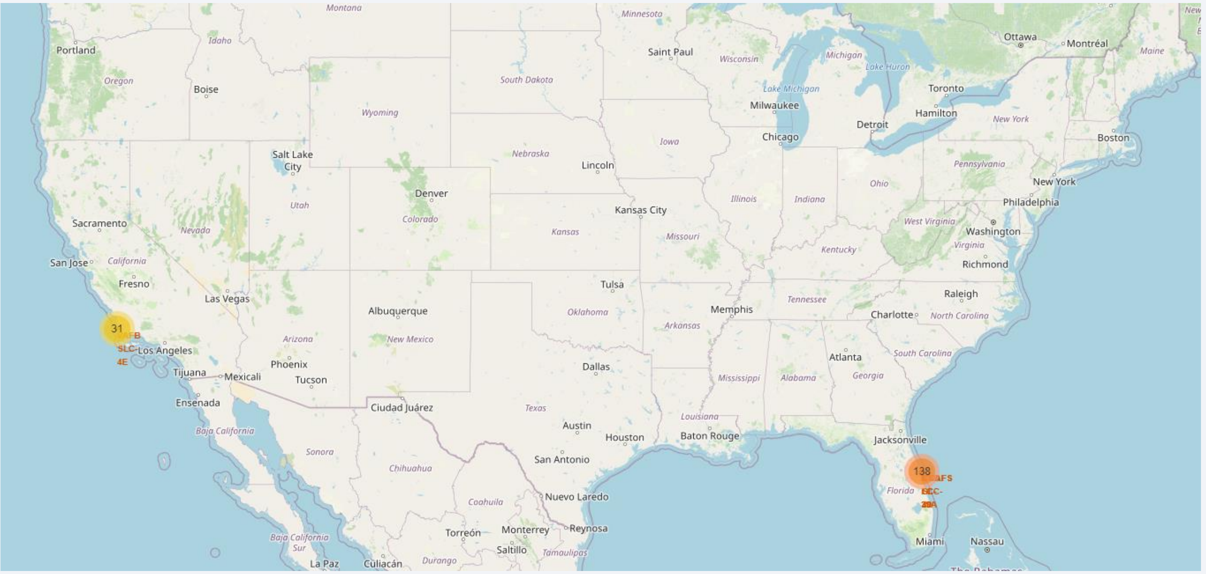
- For this predictive analysis, we focused on four classification models to find the best performing classification model. The models are Logistic Regression, Support Vector Machine (SVM), Decision Tree (Tree), and K Nearest Neighbors.
- Source Code: https://github.com/bubblypanda/skillsnetworklab/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

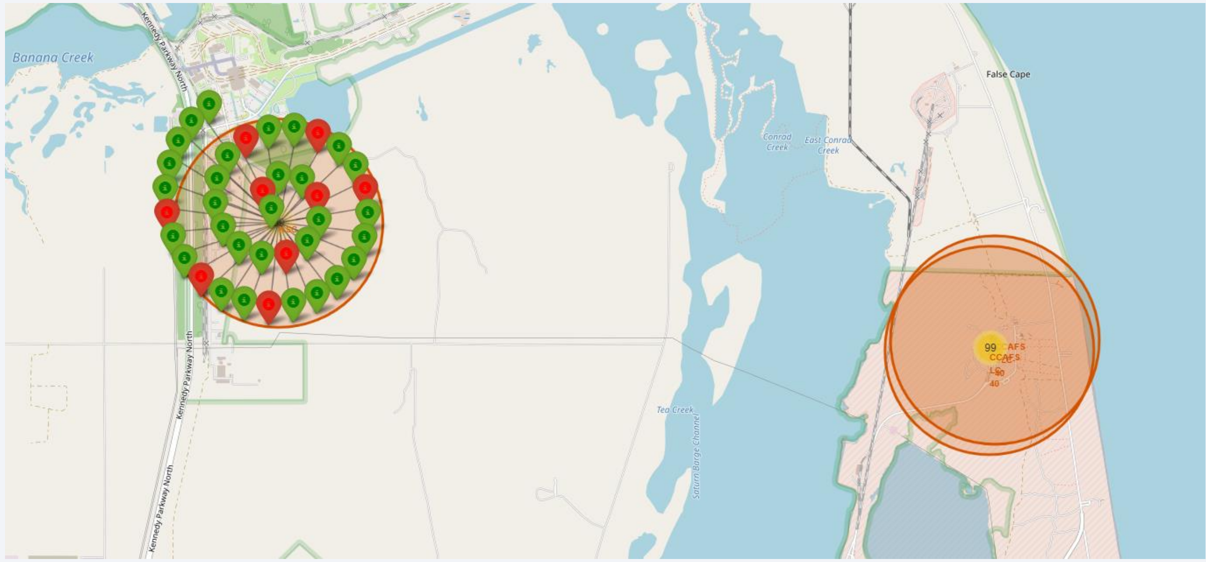
Results

- Looking at four unique launch sites, the earliest successful landing outcome in ground pad occurred on 01-05-2017.
- Landings were significantly more often successful than failures.
- The success rate has increased since these launches began in 2010, with 2019 having the greatest percentage of successful launches across all sites.
- Four orbits had 100% success rates: ES-L1, GEO, HEO, and SSO.

Results

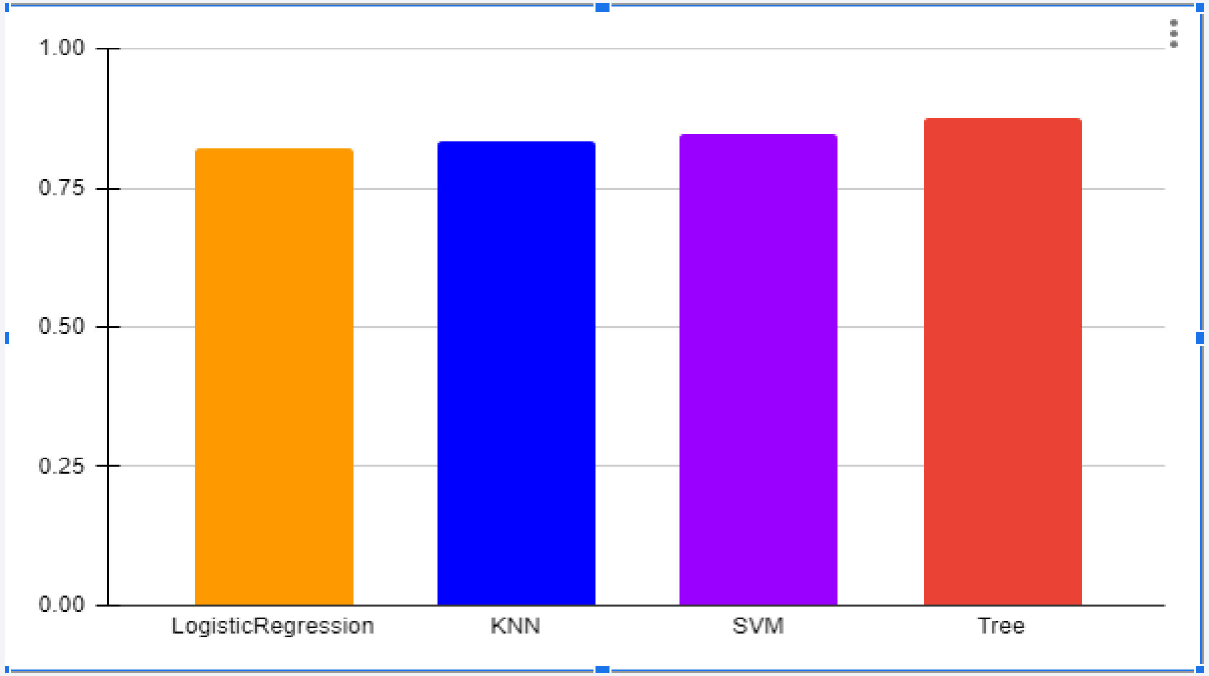
- All launches occurred in the United States, with most occurring in the state of Florida. This can be seen in the US map.
- Launch Site KSC LC-39A had the greatest percentage of successful launches, as seen by the many green markers in the second screenshot.





Results

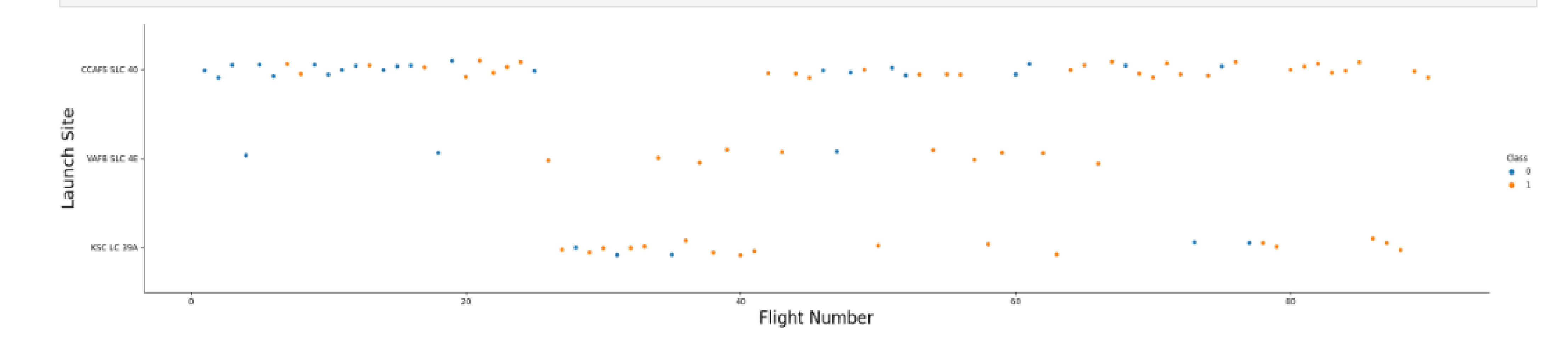
- The Decision Tree model was found to be the best Classification model to predict a successful launch in this project.



Section 2

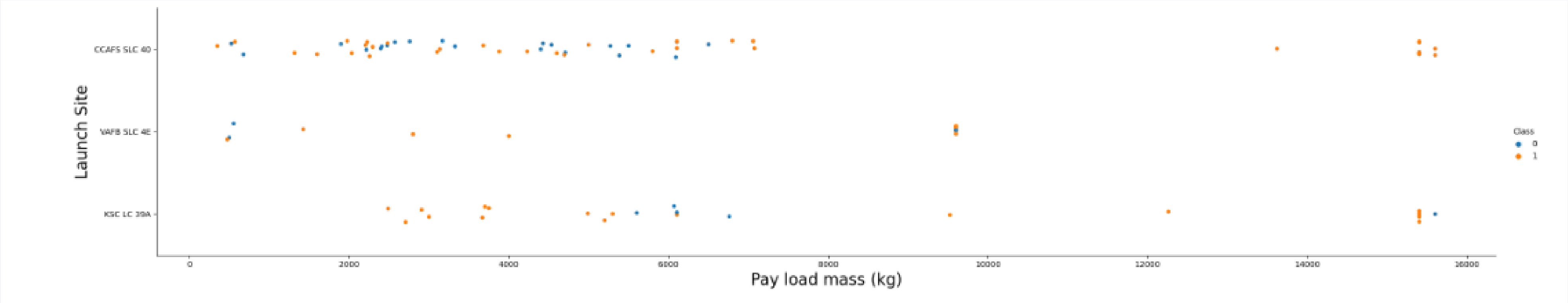
Insights drawn from EDA

Flight Number vs. Launch Site



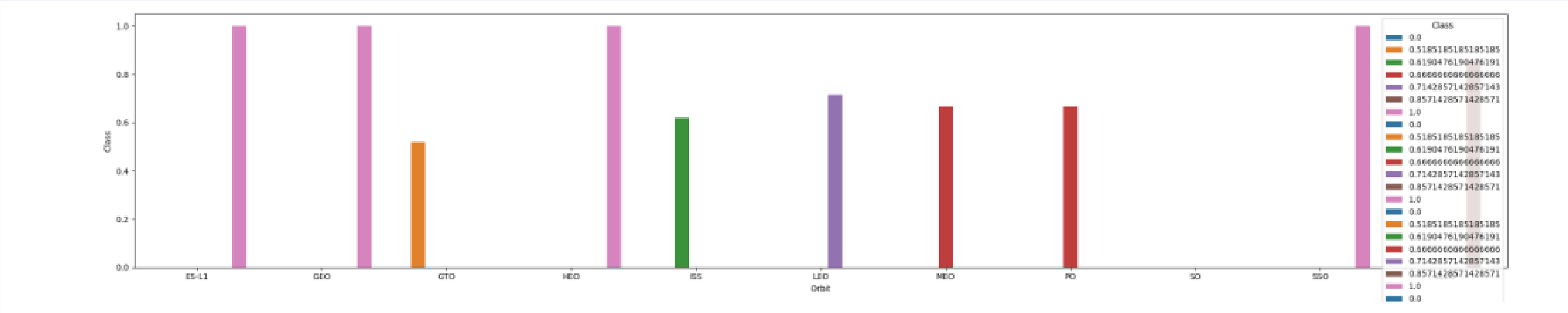
- The launch site with the least amount of flight numbers is VAFB SLC-4E.
- Launch site CCAFS SLC-40 had more flights with higher flight numbers than either of the other launch sites.

Payload vs. Launch Site



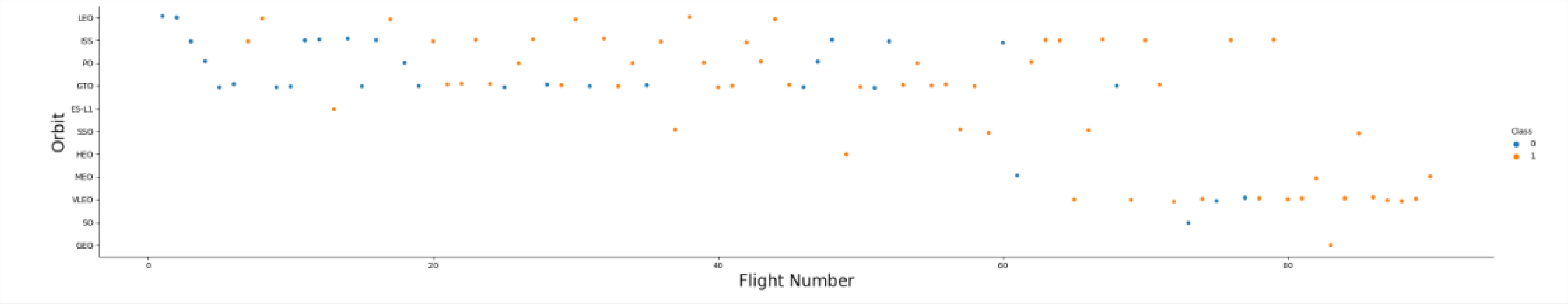
- The VAFB-SLC launch site has no rockets launched with heavier payload mass (or a heavier payload mass of over 10,000 kg).
- The CCAFS SLC-40 launch site also had more launches with smaller payloads than both the other launch sites.

Success Rate vs. Orbit Type



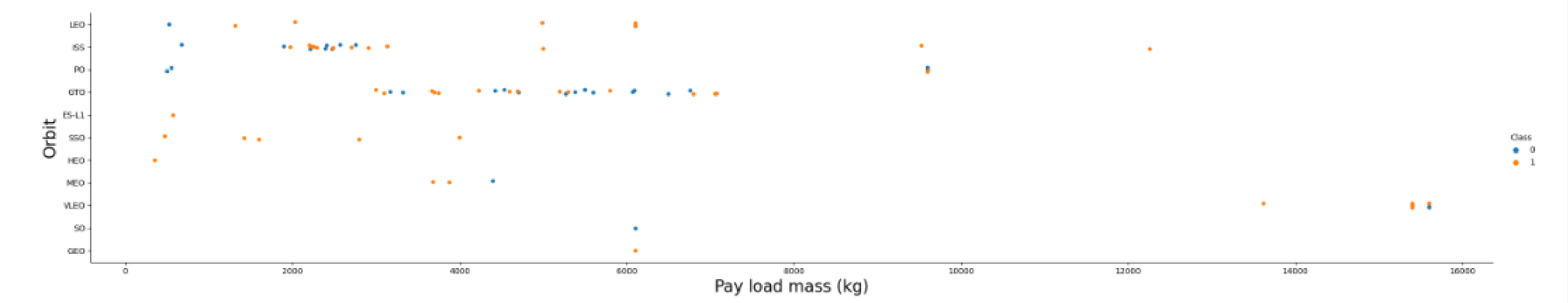
- There were four orbit types that were 100% successful and one orbit type that was 100% unsuccessful.
- The 100% successful orbit types are featured in pink in the bar chart above.

Flight Number vs. Orbit Type



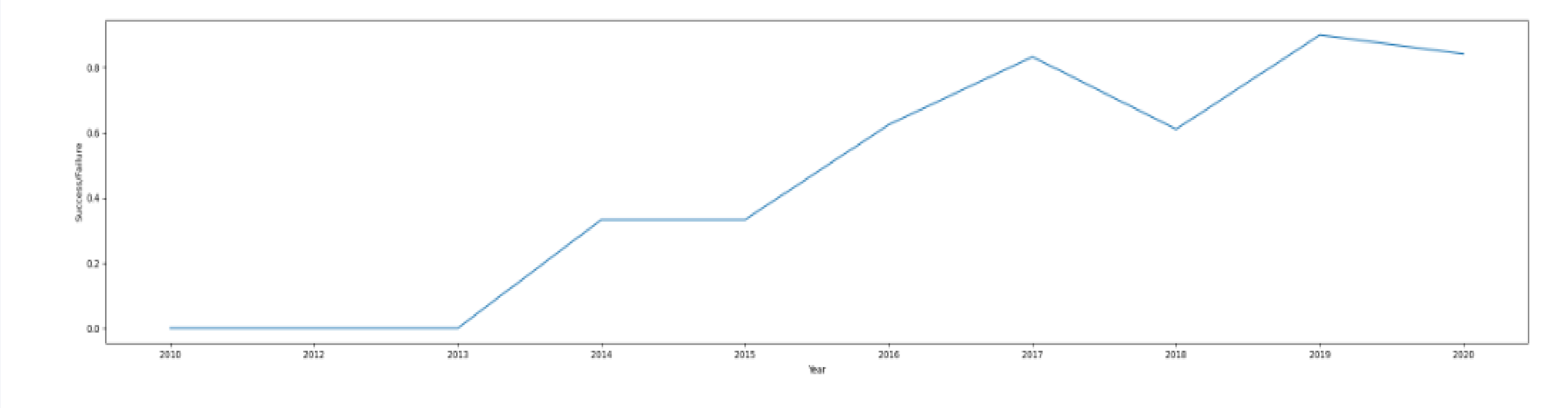
- There are more flight numbers with several types of orbit including: ISS, GTO, and VLEO.

Payload vs. Orbit Type



- Lower total payload mass (in kilograms) is associated with more variety in types of orbit.
- Only the VLEO orbit type has the highest payload mass.

Launch Success Yearly Trend



- There was an increase in the success rate as the years passed from 2010-2020. The most successful year was 2019.

All Launch Site Names

- These are the distinct launch_site names found in the dataset.

Unique Launch Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- These were the first 5 records in the dataset where launch sites begin with `CCA`. These records all occurred at the same launch site: CCAFS LC-40.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass in kilograms carried by boosters launched by NASA (CRS) was 45,598 kg. This was the sum of all records that had a customer name of NASA (CRS).

45,598 kg

Average Payload Mass by F9 v1.1

- The average payload mass in kilograms carried by booster version F9 v1.1 was 2,928.4 kg. This was the average payload mass per record where the booster_version was F9 v1.1.

2,928.4 kg

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 01-05-2017. This was the date of the first record where the landing outcome was success (ground pad).

01-05-2017

Successful Drone Ship Landing with Payload 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 displayed in the table here. Only 4 boosters fit these qualifications.

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes was 100. This included 2 Mission outcomes: Success and Success (Payload Status Unclear). There was only 1 failure (in flight) mission outcome.

Mission Outcome	Total Number
Failure (in flight)	1
Success	99
Success (Payload Status Unclear)	1

Boosters Carried Maximum Payload

- The table to the right includes the names of the booster which have carried the maximum payload mass. There are 12 boosters that have carried this 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

- Below are the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015. They occurred in January and April at the same Launch Site: CCAFS LC-40.

month	month_name	Landing _Outcome	Booster_Version	Launch_Site
01	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Below is the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked in descending order. Success ranked highest in descending order.

Landing _Outcome	rank
Success	1
Success (drone ship)	2
Success (ground pad)	3

Section 3

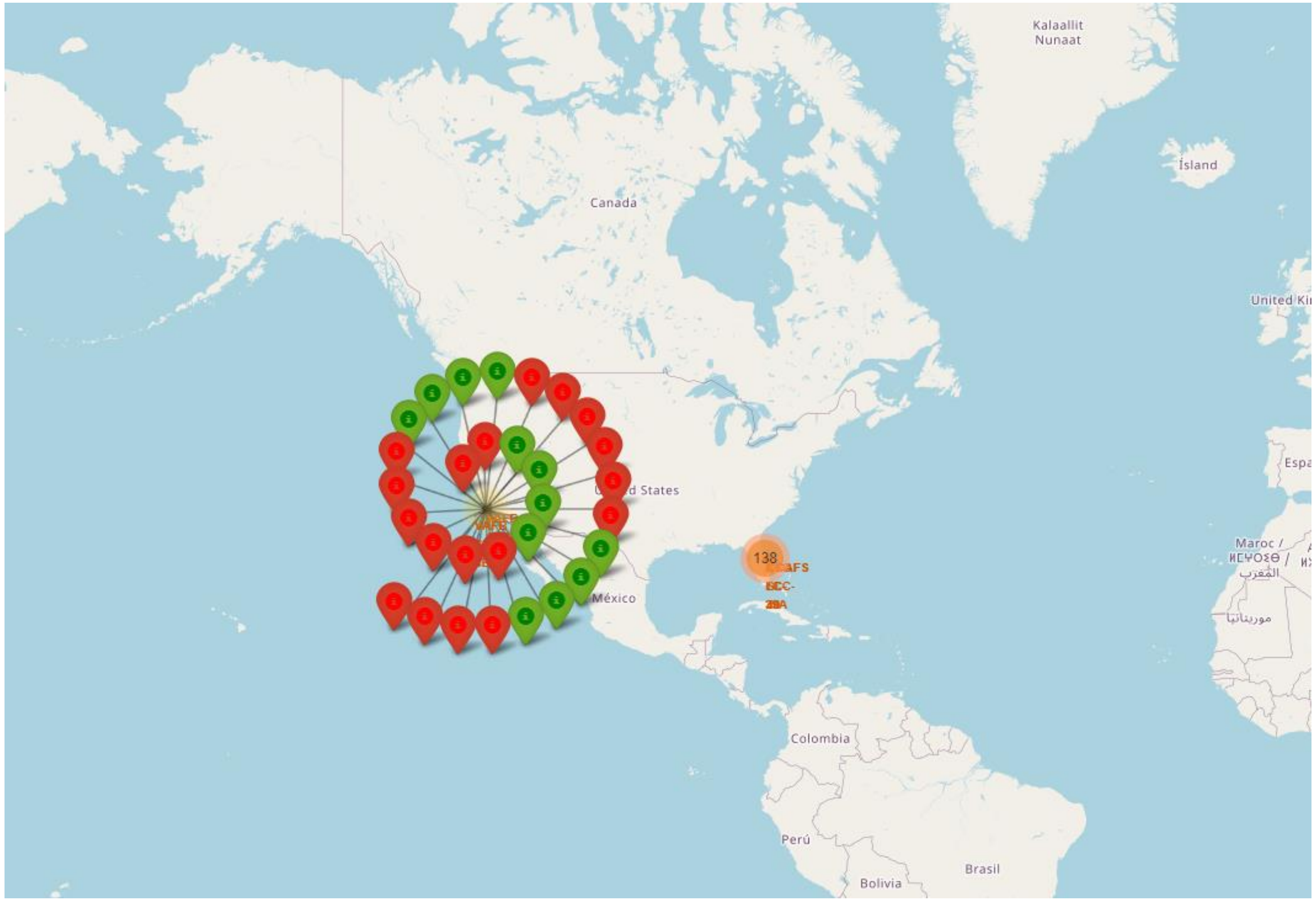
Launch Sites Proximities Analysis

Global Launch Sites

- This screenshot includes all launch sites’ location markers on a global map.
- There are multiple launch sites in Florida, which explains the overlap seen in the labels of the sites covering Florida in the United States.

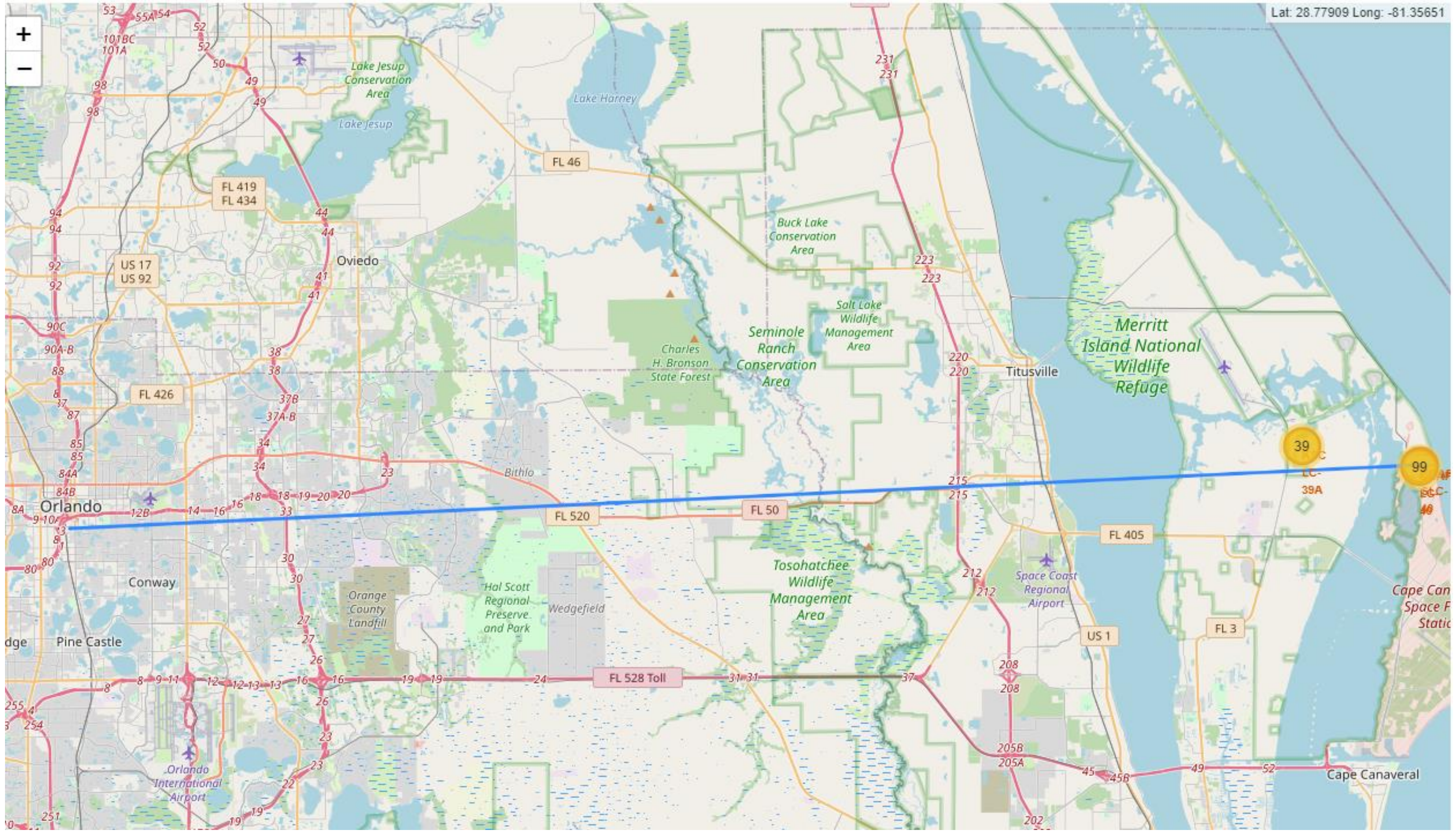
Successful and Unsuccessful Launches

- This screenshot shows the color-labeled launch outcomes on the map.
- All successful launches are shown in green, while all unsuccessful launches are shown in red.
- This map breaks down each launch site into similar successful/unsuccessful launch breakouts.



CCAFS SLC-40 Breakout

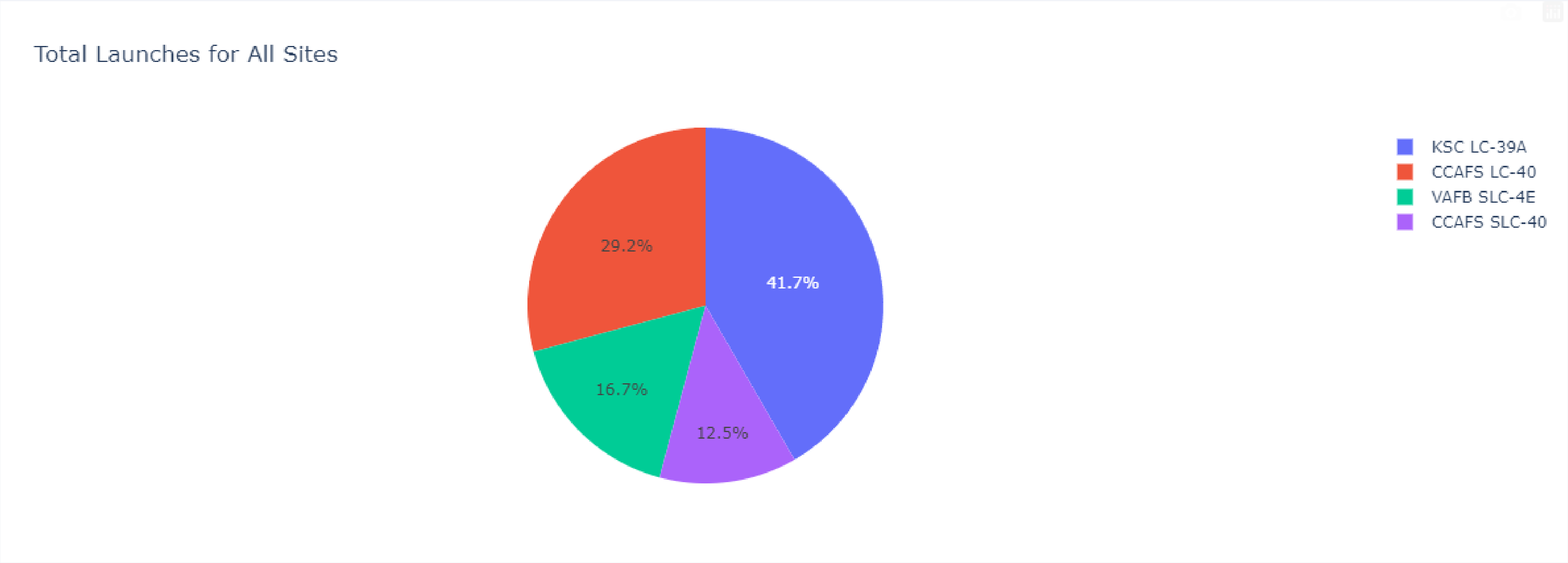
- The 99 includes two launch site locations: CCAFS SLC-40 and CCAFS LC-40, but these lines are specifically originating from the CCAFS SLC-40 launch site.
- This screenshot of CCAFS SLC-40 shows its proximity to the nearest town (Orlando, Florida) and to the nearest coast.



Section 4

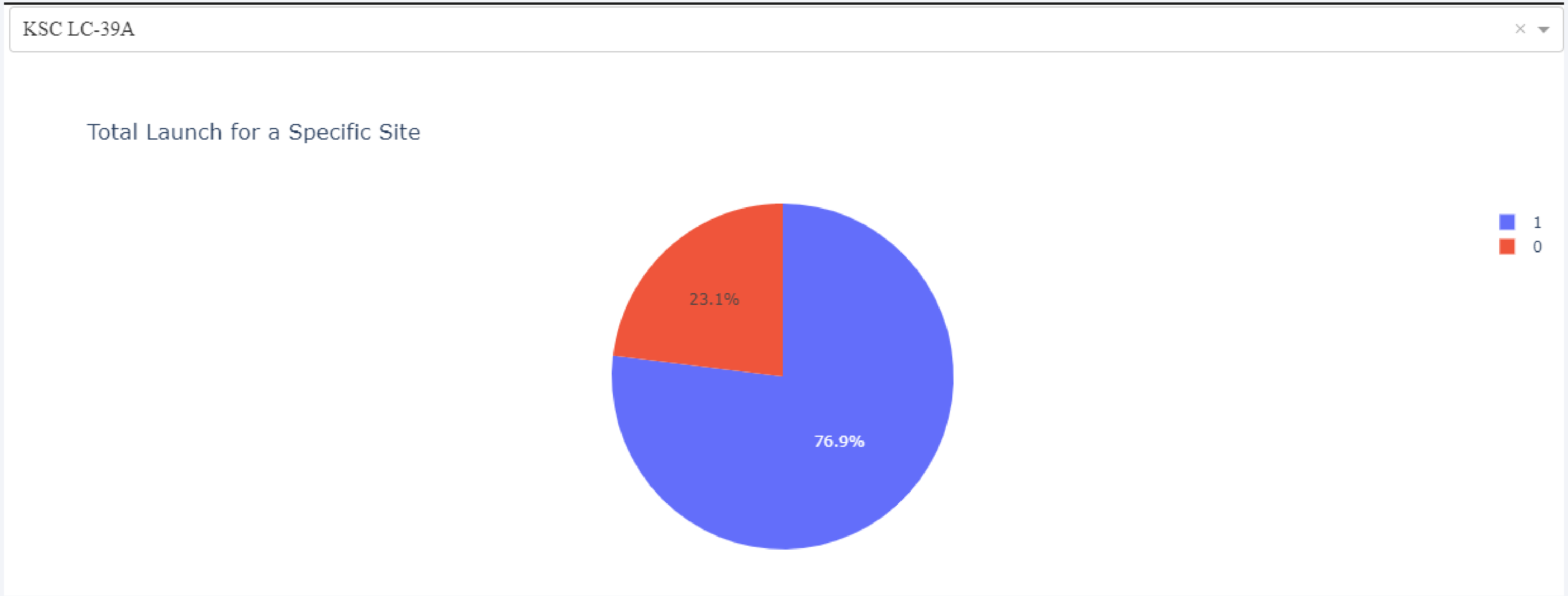
Build a Dashboard with Plotly Dash

Total Successful Launches for All Sites



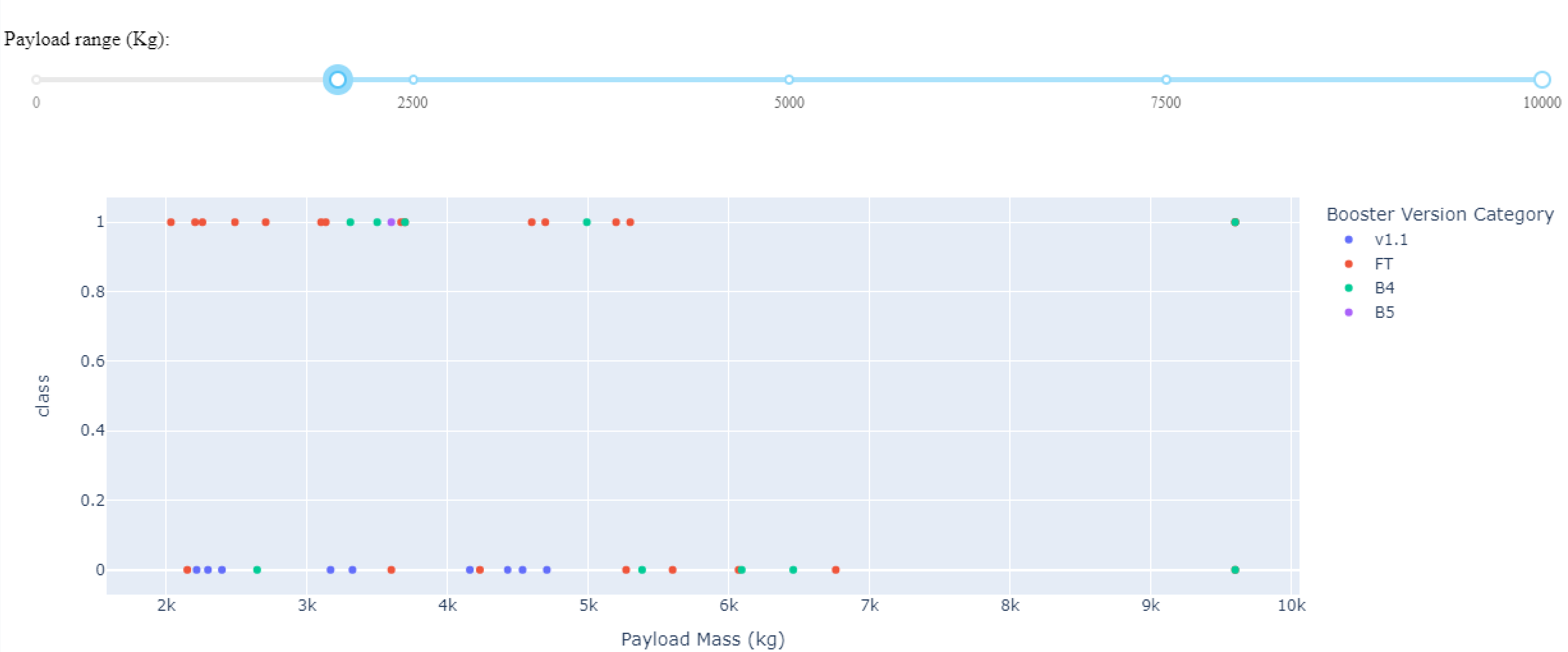
- This pie chart shows the percentage of launches for all sites. KSC LC-39A has the greatest percentage of launches and CCAFS SLC-40 has the lowest percentage of total launches.

Total Launches for KSC LC-39A



- KSC LC-39A, the launch site with highest launch success ratio, has a successful launch rate of 76.9%.

Payload vs Launch Outcomes for All Launch Sites

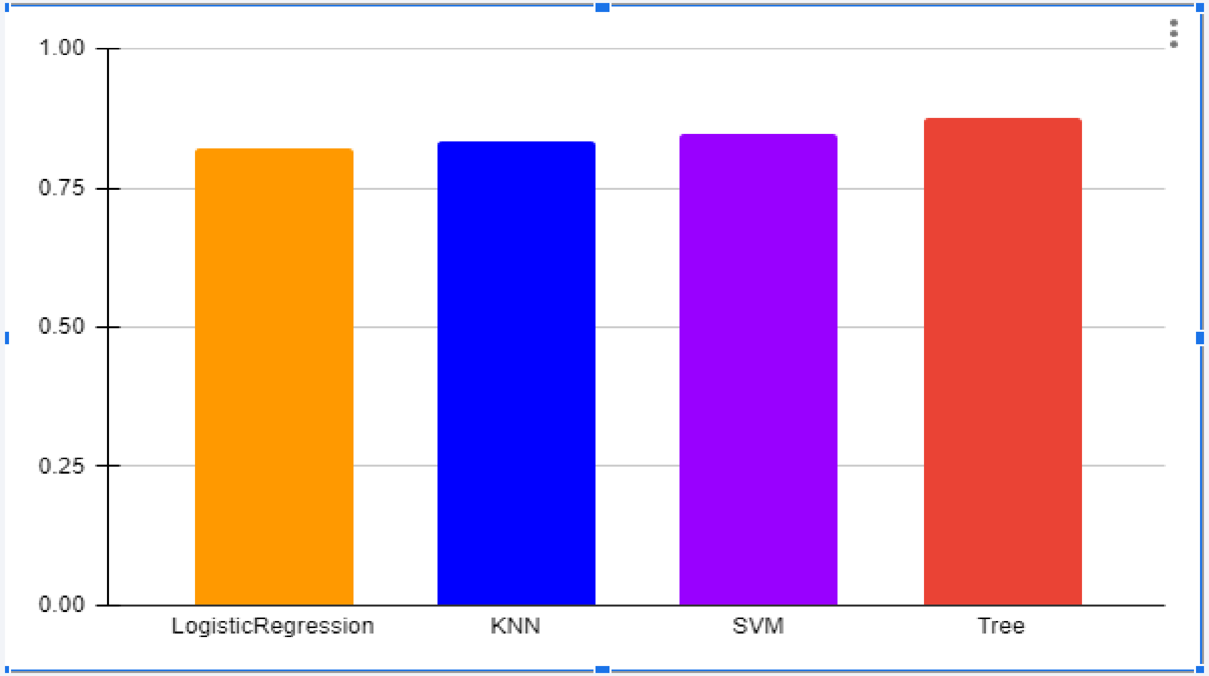


- This is the Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider. Booster B4 had both successful and unsuccessful launches with the highest payload mass in kilograms.

Section 5

Predictive Analysis (Classification)

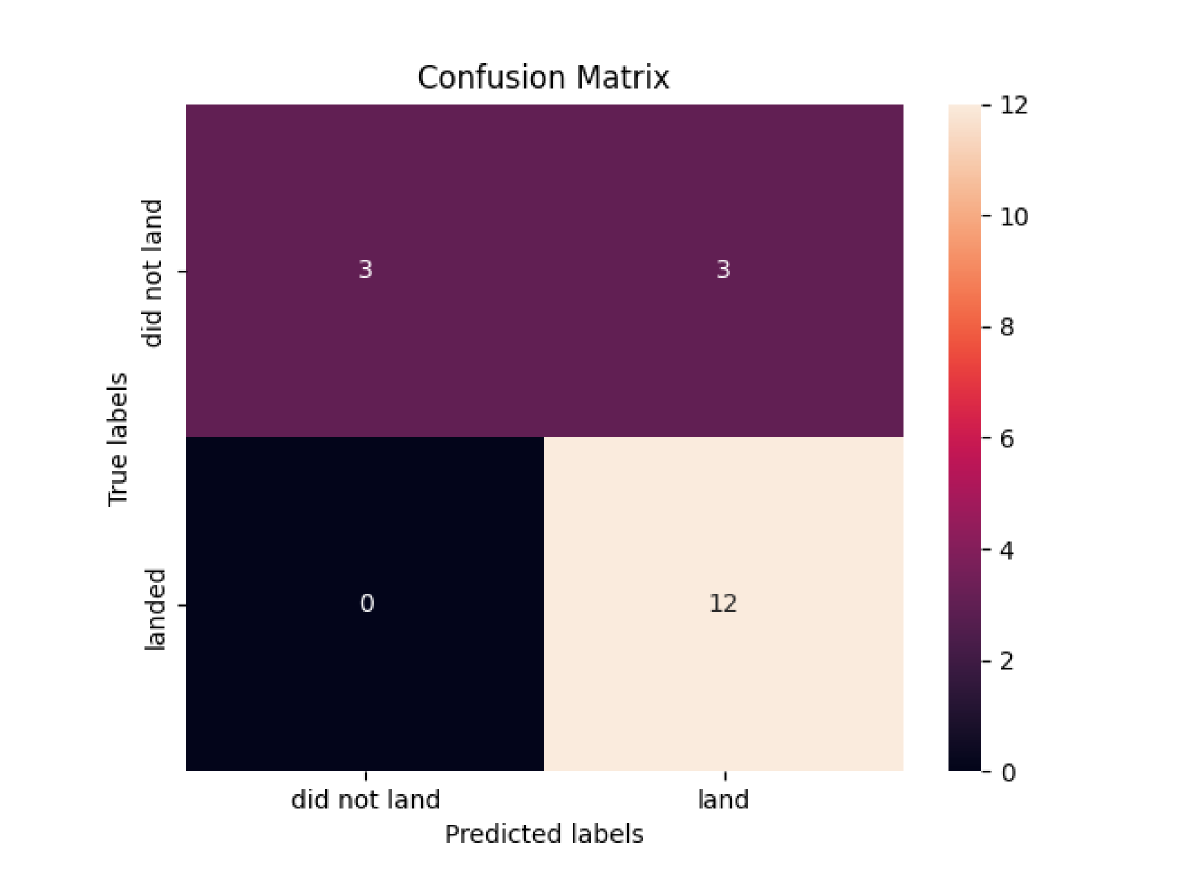
Classification Accuracy



	Algorithm	Train Data Accuracy
0	LogisticRegression	0.821429
1	KNN	0.833929
2	SVM	0.848214
3	Tree	0.876786

- While the accuracy for these four models was close, the Decision Tree Model has the highest classification accuracy.

Confusion Matrix



- This is the confusion matrix for the Decision Tree Model. It had the highest accuracy amongst the three other models used, correctly predicting 12 land labels.

Conclusions

- With an almost 100% Mission success rate, it is safe to assume that the Space X launch price will be closer to the lower estimate, a cost of 62 million dollars.
- To compete with Space X, Space Y will have to compose a closer bid to this cost and be able to reuse the first stage.
- Space Y also should use Landing Site KSC LC-39A, as it had the highest percentage of successful landings of the four unique sites used by Space X.
- Our machine learning work suggests that the Decision Tree model would be the most accurate classification model for predicting whether a landing will be successful.

Thank you!