SpaceY
First Stage Launch Analysis

Andrei Boicu
2024-12-16

Content

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

Executive Summary

Summary of methodologies

The purpose of this analysis is to identify and predict those variables that count for a successful rocket landing. In my attempt to make this determination, I have used open-sourced SpaceX data together with the following methodologies:

- **Collect** data using SpaceX REST API and Wikipedia web scraping
- **Wrangle** data to create success / fail outcome variables
- **Explore** data with visualization techniques considering payload mass, launch site, flight number and orbits.
- **Analyze** the data with SQL calculating relevant statistics
- **Compute** launch site success rates and proximity to geographical targets (such as railways, cities, coastline)
- **Visualize** those launch sites with the highest landing success rate counting for payload mass as well
- **Build ML models** to predict landing outcomes using several models: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbor.

Results

- Launch success rate has improved over time (starting 2013)
- KSC LC-39A has the highest success rage among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success landing rate
- Most launch sites are near the equator, and all are close to the coast
- All models performed similarly on the test set. The Decision Tree model slightly outperformed.

Introduction

Background

SpaceX, a leader in the commercial space industry, is revolutionizing space travel by making it more affordable and accessible. Its achievements include sending spacecraft to the International Space Station, launching a satellite constellation to provide internet access, and conducting manned space missions. Central to SpaceX's success is its innovative reuse of the first stage of its Falcon 9 rocket, which reduces launch costs to $62 million, compared to upwards of $165 million charged by other providers who cannot reuse this component.

This cost advantage underscores the importance of determining whether the first stage of a rocket can be reused, as it directly impacts the affordability of a launch. Using public data and machine learning models, we aim to predict the likelihood of the first stage successfully landing and being reused, which can inform the pricing and feasibility of launches for SpaceX and its competitors.

Explore

- How variables such as payload mass, launch site, number of flights, and orbits affect the success of first-stage Falcon 9 landing
- The rate of successful landings over the years
- Best predictive model for successful landings (binary classification)

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - I have used SpaceX REST APIs and Web Scraping techniques to collect data
- Perform data wrangling
  - Cleaning the data by handling missing values, filtering data, and applying one-hot-encoding to prepare the categorical variables for later use.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models that were used: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbor.

Data Collection – API

- First request data from SpaceX API
- Decode response using .json() and convert it to a DataFrame by .json_normalize()
- Request information about the launches form SpaceX API using custom functions

resource: github.com

Data Collection – Web Scraping

- First request data from Wikipedia (Falcon 9 data)
- Use BeautifulSoup Python library to extract HTML response
- Create DataFrame and store the data that was extracted
- Save all of it to a .csv file

resource: github.com

Data Wrangling

- Filtered only for Falcon 9 data
- Replaced missing values on Payload mass with the average value for the same
- Created a landing outcome label for successful / failed launches

resource: github.com

EDA with Data Visualization

- Flight Number vs. Payload mass
- Flight Number vs. Launch Site
- Payload mass (kg) vs. Launch Site
- Payload mass (kg) vs. Orbit

Scatter plots: to better understand the relationship between these variables

Bar charts: to show comparison between categorical variables.

resource: github.com

EDA with SQL

- Display name of 4 unique launch sites
- First 5 rows starting with CCA
- Total amount of payload mass (kg) carried by booster launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass between 4,000 and 6,000
- Total number of successful and failed launches
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 descending.

resource: github.com

Interactive Map with Folium

- Added circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added circles at all launch sites coordinates with a popup label showing its name using its latitude and longitude coordinates
- Added colored markers of successful and unsuccessful launches at each launch site to show which launch sites have high success rates
- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

resource: github.com

Dashboard with Plotly Dash

- Allow user to select one or all launch sites
- Pie chart showing successful and failed launches as a percent of the total
- Range tool to select payload mass (kg)
- Scatter plot displaying correlation between Payload mass and Launch Success

resource: github.com

Predictive Analysis

- Create NumPy array from Class column
- Standardize the data with StandardScaler. Fit and train the data
- Split the data for train and test using train_test_split.
- For parameter optimization, create a GridSearchCV object with cv=10
- On different algorithms (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor) apply GridSearchCV
- Using .score() function, calculate accuracy on test data for each of these models
- Generate confusion matrix for each model
- By a set of measure (Accuracy, Jaccard Score, and F1 Score) asses the best model.

resource: github.com

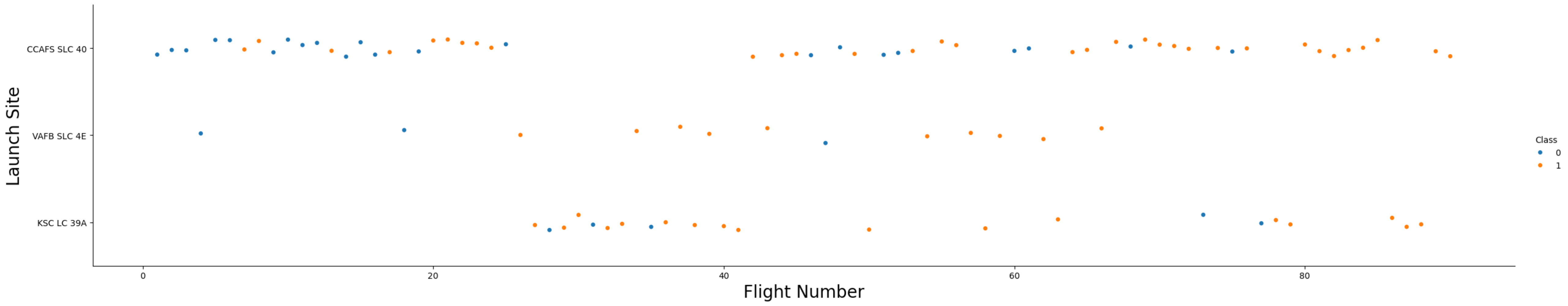
Results

- Launch success has improved over time (starting 2013)
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rate
- Most launched sites are neat the equator and are close to the coast
- Lunch sites are far from cities, railways so that in case of a failure not to cause any harm
- Decision Tree model is slightly better than the other models for the give dataset.

Insights drawn from EDA

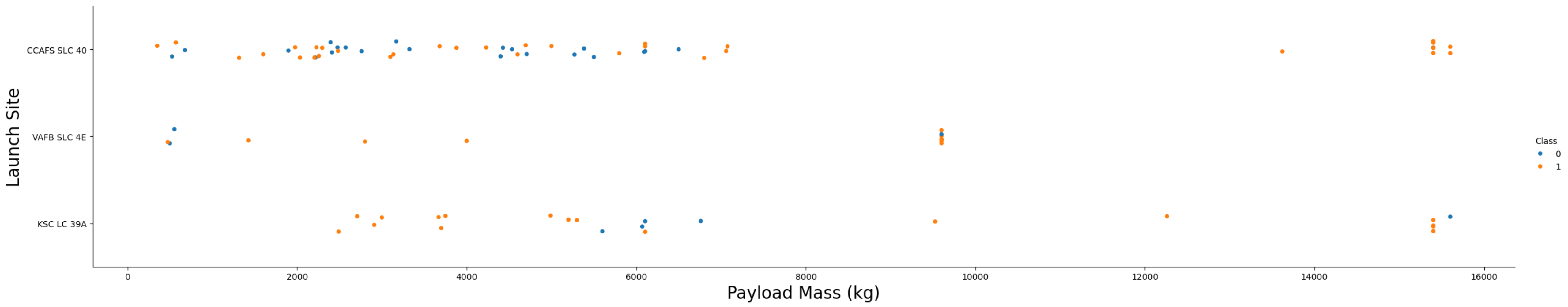
Flight Number vs. Launch Site

- After 77th flight, we can see only succeeded landings
- Launch site VAFB SLC 4E has no launch after 65th flight.
- Most flights were performed from CCAFS SLC 40 site



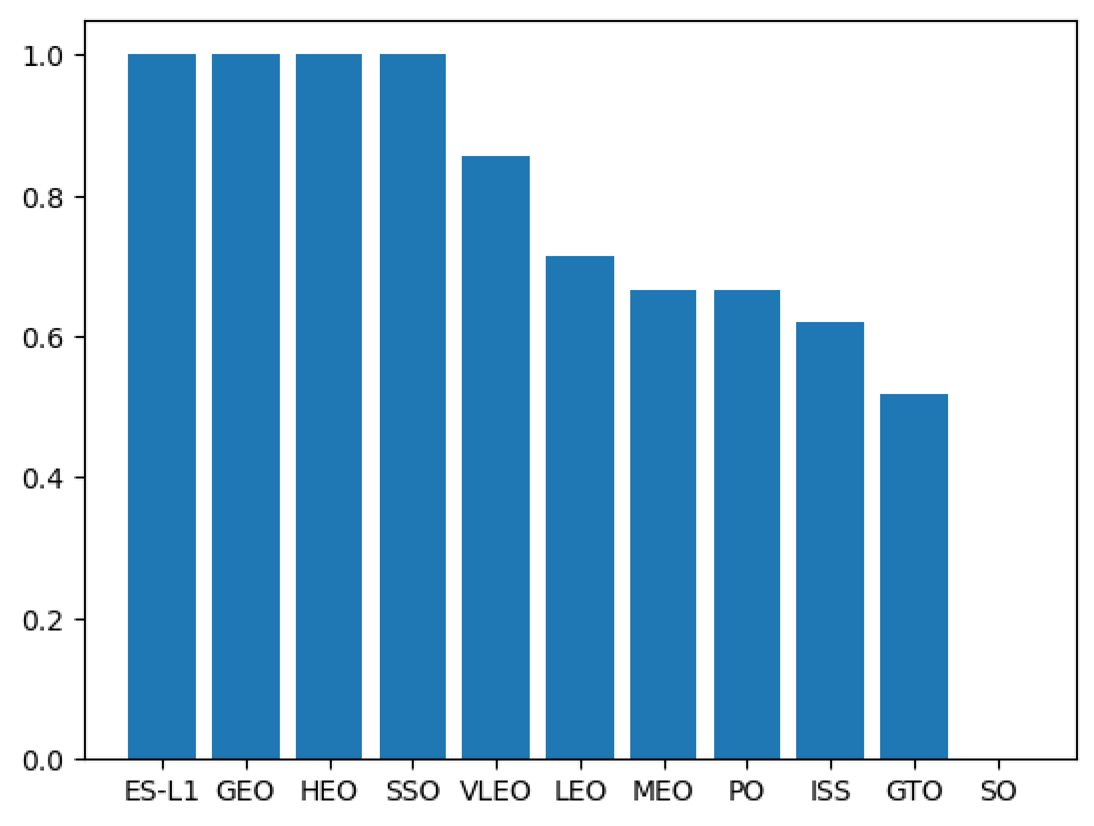
Payload vs. Launch Site

- The higher the payload mas, the greater the success rate
- VAFB SLC 4E has not launched anything greater than 10,000 kg
- Most launches with a payload mass greater than 7,000 kg were successful



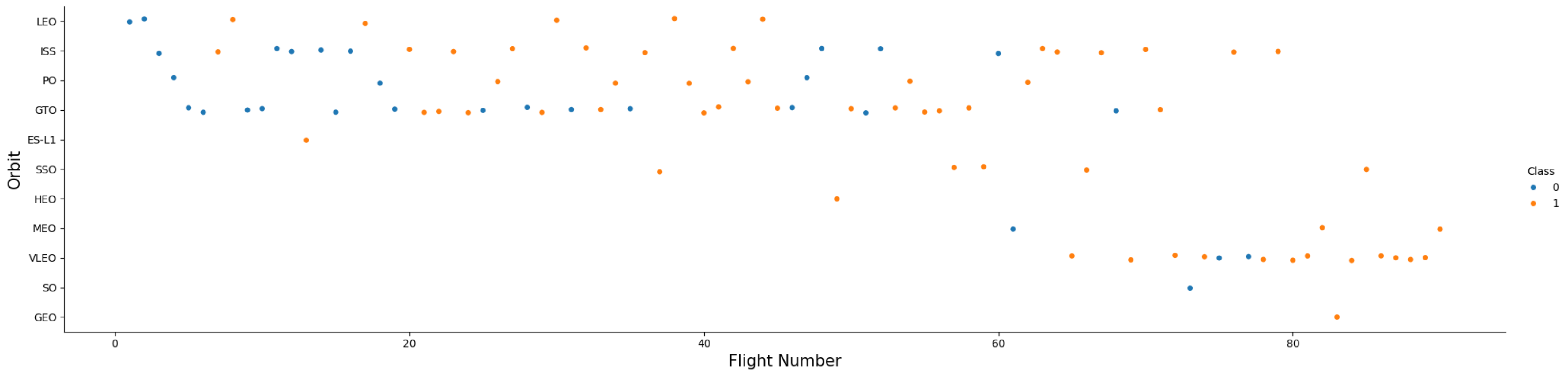
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have all 100% success rate
- SO has 0% success rate



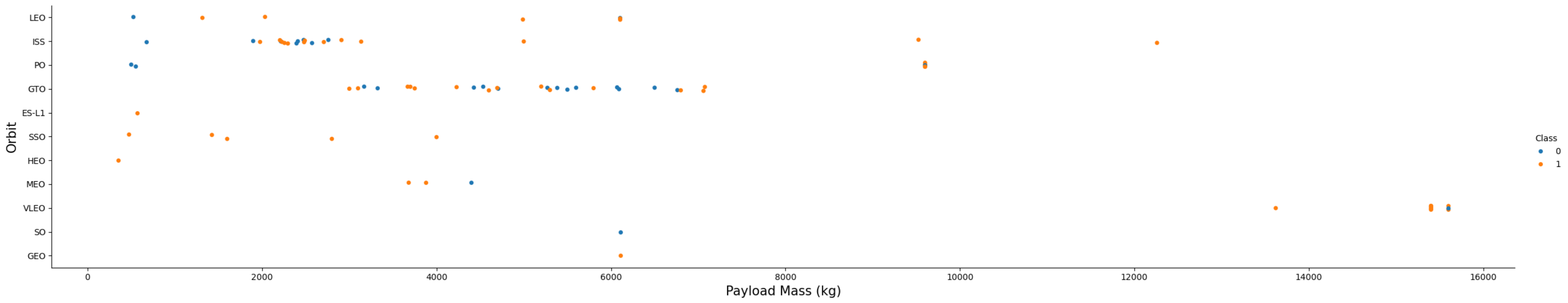
Flight Number vs. Orbit Type

- With each flight, we see that the success rate increases
- Everything beyond 80^{th} flight is a success



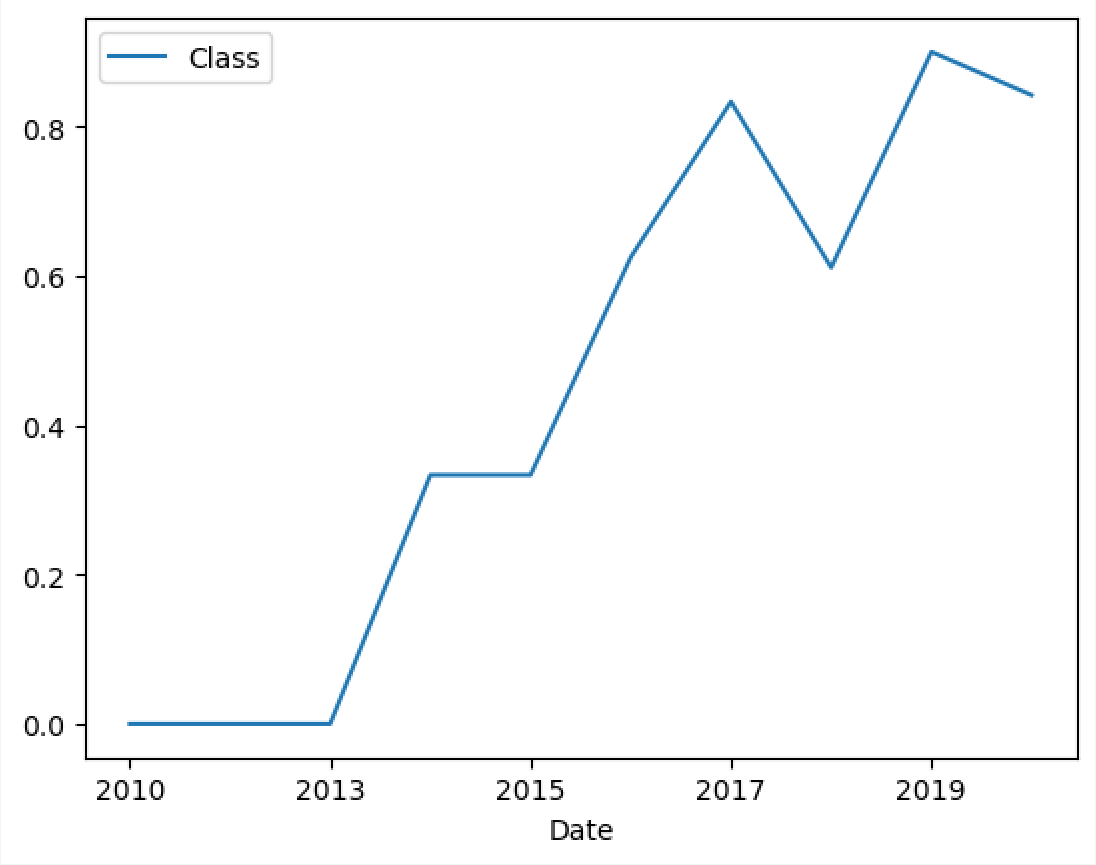
Payload vs. Orbit Type

- Only VLEO launched flights with a payload greater than 15,000 kg
- Almost all launches with a payload greater than 10,000 kg are successful



Launch Success Yearly Trend

- Success rate improved considerably from 2013 reaching to more than 80% in 2019
- There is a dip in success rate in 2018, but it recovered fast in 2019



Launch Sites

- All unique Site Names: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A VAFB SLC-4E
- First five records where Launch Site starts with 'CCA'

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Payload Mass

- Total Payload Mass carried by NASA (CRS) is 45,596 kg

```
%sql select Customer, sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Customer	total_payload_mass
NASA (CRS)	45596

- Average Payload Mass carried by booster version F9 v1.1 is 2,928 kg

```
[25]  %sql select avg(PAYLOAD_MASS__KG_) avg_payload_mass from SPACEXTABLE where Booster_Version like 'F9 v1.1'
```

```
...  * sqlite:///my_data1.db
Done.
```

...

avg_payload_mass
2928.4

Landing Data

- First successful Landing on Ground Pad occurred on 2015-12-22
- Booster versions where payload mass was between 4,000 and 6,000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- There were 99 successful missions, 1 successful mission, but with payload status unclear, and 1 failure in flight

Boosters Data

- Boosters that carried the maximum payload are listed below

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

2015 Launch Records

- There were two failures on drone ship during the year of 2015, one in January, and the other in April

monthName	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The outcomes and the number of landing outcomes can be seen below

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

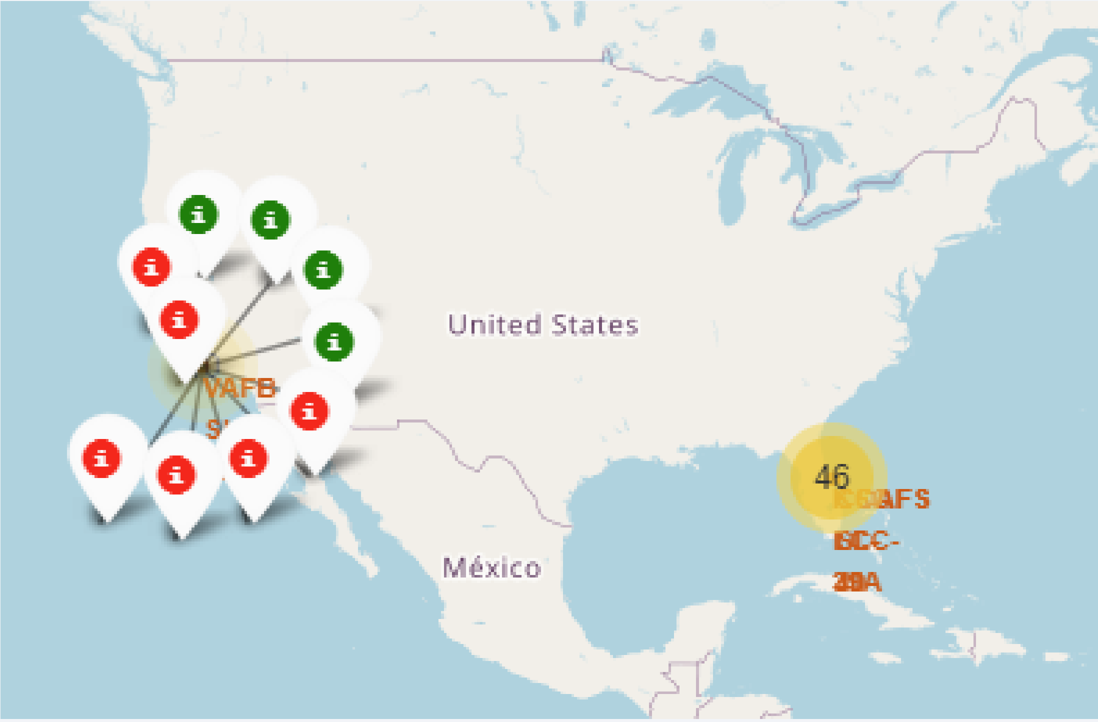
Launch Sites Analysis

Launch Sites

- The rockets launched from sites near the Equator get an additional natural boost due to the rotational speed of Earth, that helps save the cost of putting in extra fuel and boosters.

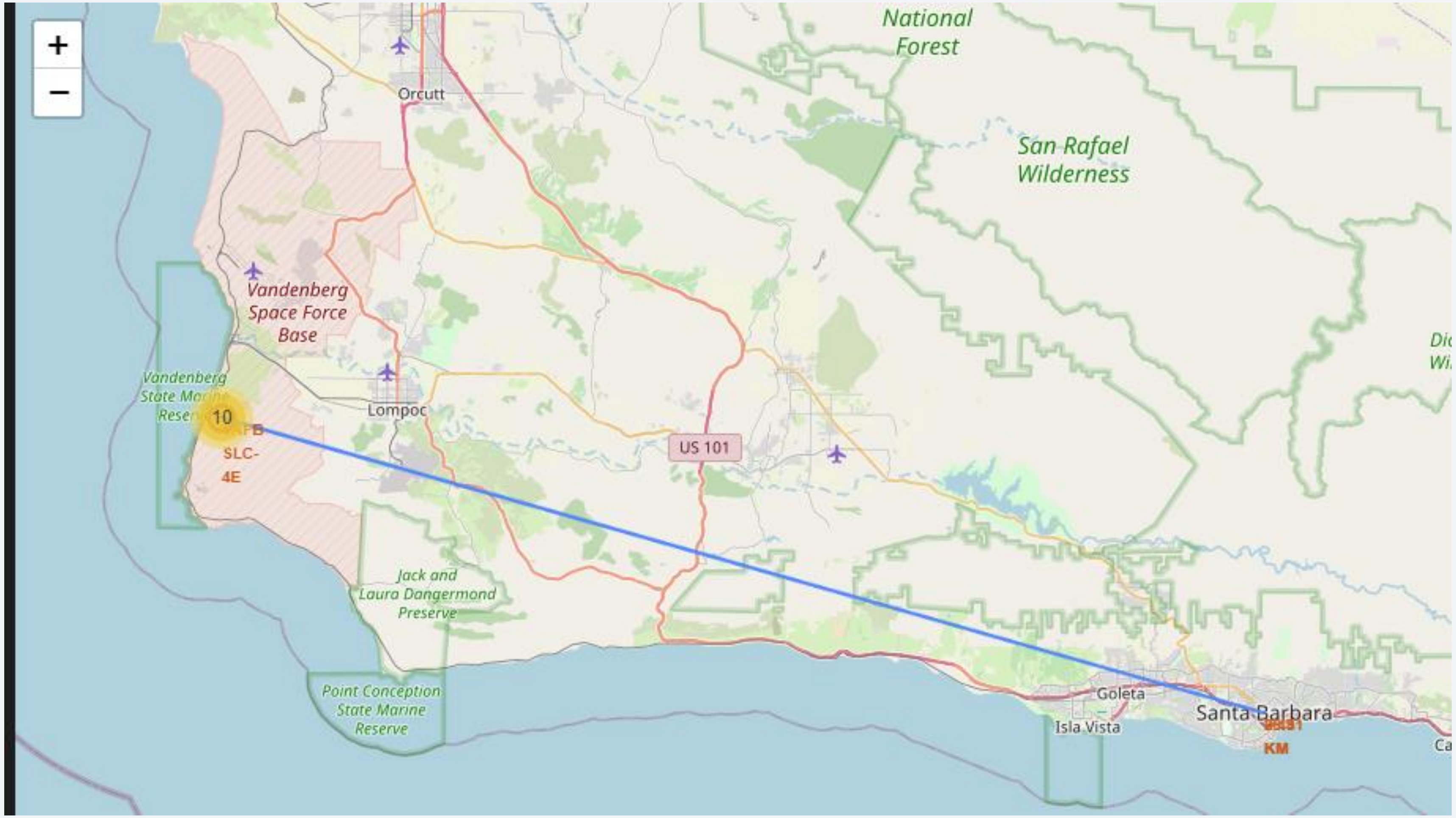
Launch Outcomes

- Green markers are for successful launches, while red markers are for failed ones
- VAFB SLC-4E, the site from the west coast had a 40% success rate (4/10)



Distance to Proximities

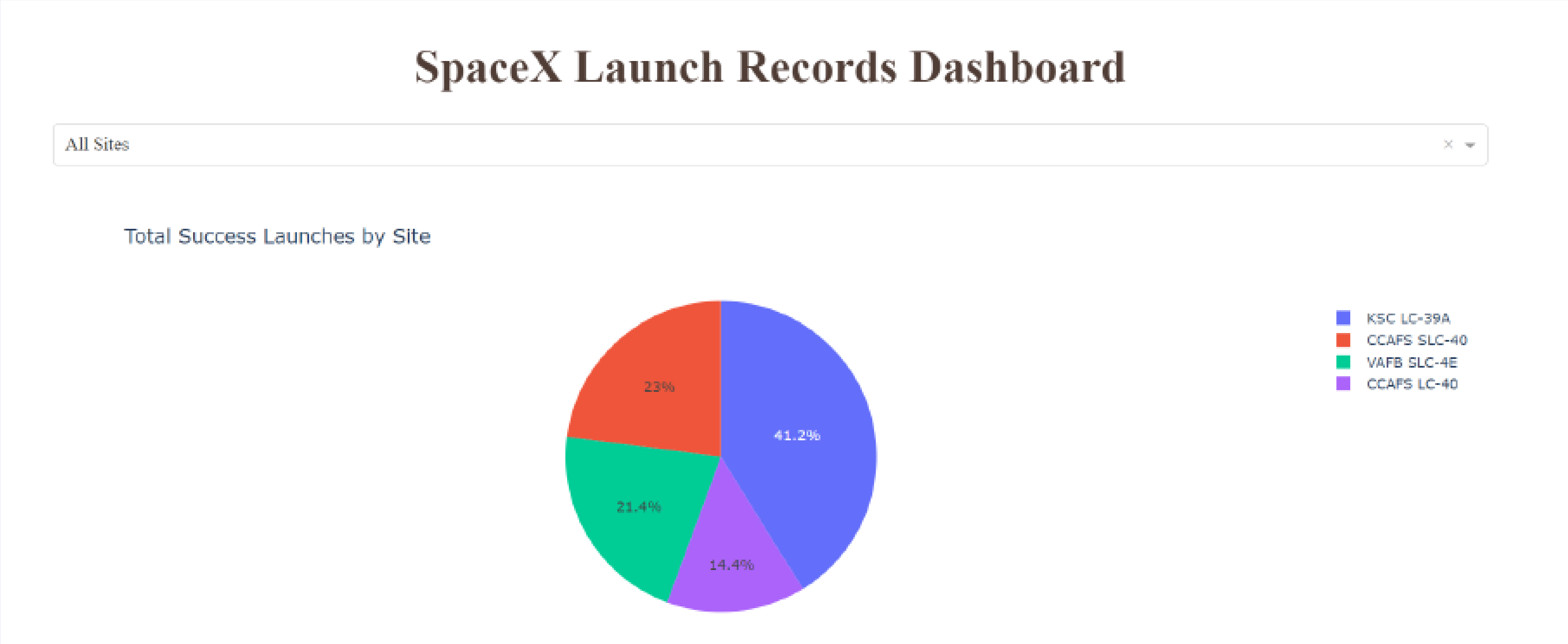
- Distance from VAFB SLC-4E to Santa Barbara



Dashboard with Plotly Dash

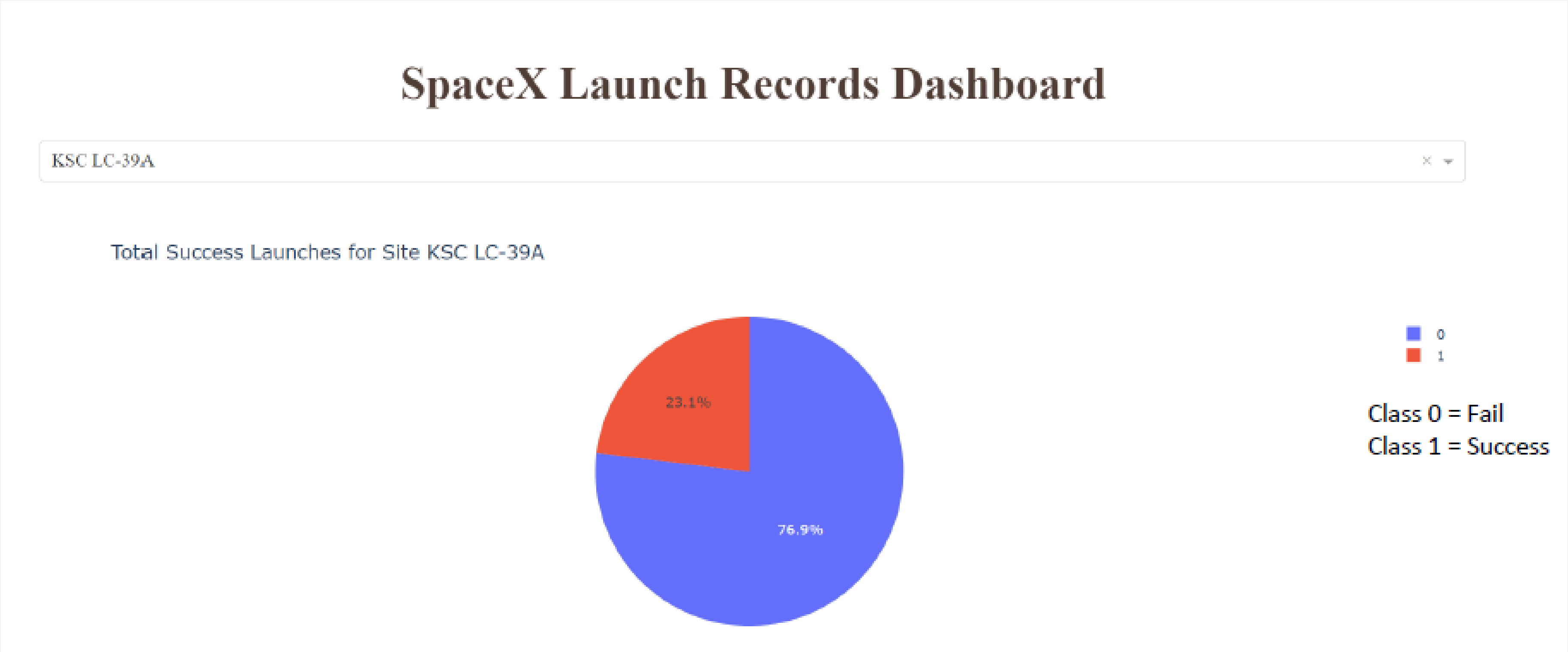
Launch Success by Site

- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

- KSC LC-29A has the highest success rate among the sites (42.1%) and a success rate of 76.9%
- only 3 failed launces out of 13 total launches



Payload Mass and Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate

Predictive Analysis

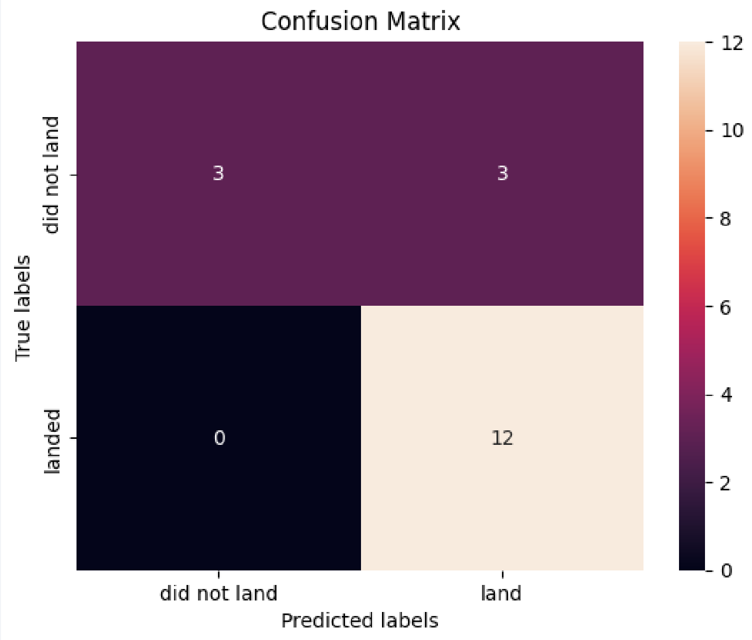
Classification Accuracy

- All the models performed at about the same level. The one that performed slightly better is the Decision Tree model

	ML Method	Accuracy Score (%)
0	Logistic Regression	83.333333
1	Support Vector Machine	83.333333
2	Decision Tree	88.888889
3	K Nearest Neighbour	83.333333

Confusion Matrix

- The confusion matrix summarizes the performance of all four classification models
- Confusion matrix was the same for all four models
    - 12 True Positive (TP)
    - 3 True Negative (TN)
    - 3 False Positive (FP)
    - 0 False Negative (FN)



- Precision = TP / (TP + FP) = 12 / 15 = 0.8
- Recall = TP / (TP + FN) = 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.8 * 1) / (0.8 + 1) = 0.89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = 15 / 18 = 0.833

Conclusions

- The model performed similarly on all models. Decision Tree performed slightly better
- Most of the launch sites are near Equator which gives a natural advantage due to the Earth’s rotation
- All sites are close to coastline, but slightly far from cities, railways, airports, etc.
- Launch Success rate increased over time, passing 80% mark in 2019
- KSC LC-39A has the highest success rate among all sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate, while SO has 0%
- The trend looks like the higher the payload mass, the higher the success rate

Thank you

Andrei Boicu
2024-12-16