IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael M
2023-04-23

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection SpaceX API and web scraping
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive analytics
  - Prediction using Machine Learning
- Summary of all results
  - Best launch site: KSC LC-39A
  - Best ML model to predict the success of a landing: Decision Tree

Introduction

- Objective: evaluate if Space Y can compete with Space X
- Desirable findings:
  - Estimate the total cost for launches by predicting successful landings of the first stage rockets
  - Which launch sites have the highest success rates

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API
  - Webscraping from Wikipedia
- Perform data wrangling
  - Raw data was extended by a variable for the landing outcome
  - One hot encoding for categorial features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
    - Features were normalized
    - Four different classification models were applied:
        - Logistic Regression
        - Support Vector Machine
        - Decision Tree
        - K-nearest neighbors
    - Best model: Decision Tree Classifier

Data Collection

- Source 1: SpaceX API (https://api.spacexdata.com/v4/rockets/)
- Source 2: Wikipedia (webscraping) (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX REST API was used as primary datacource
- Webscraping was applied to include SpaceX data from Wikipedia
- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Hands-on%20Lab_%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- Webscraping was applied to include SpaceX data from Wikipedia

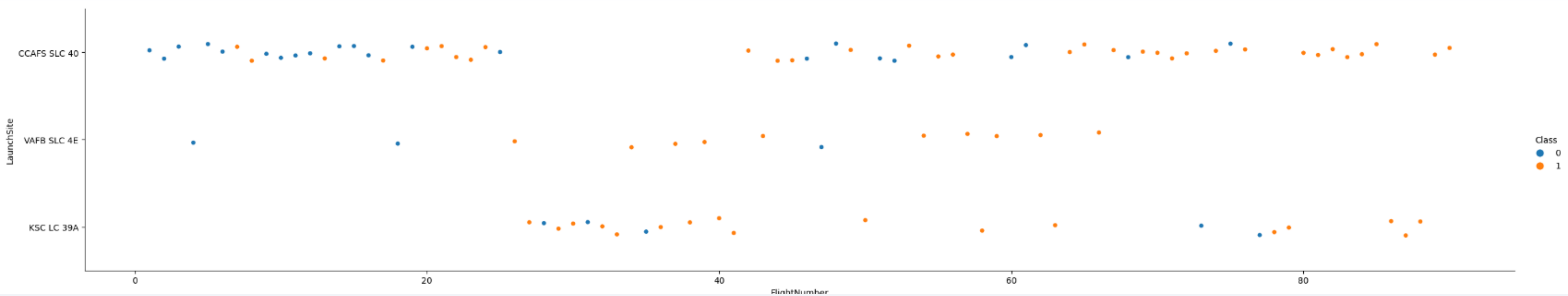
- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- First, Exploratory Data Analysis (EDA) was applied to gain first insights
- Second, the following parameters were analyzed:
  - Launches per site
  - Occurrences of each orbit
  - Occurrences of mission outcome per orbit type
- Finally, the landing outcome was generated based on the Outcome column and reduced from multiple categories to a true or false value
- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Hands-on%20Lab_%20Data%20Wrangling.ipynb

EDA with Data Visualization

- The following pairs were analyzed on their correlations between each other:
  - Payload Mass vs. Flight Number, Launch Site vs. Flight Number, Launch Site vs. Payload Mass, Orbit vs, Flight Number, Payload vs. Orbit



Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- SQL queries performed:
  - unique launch sites in the space mission
  - Top 5 launch sites beginning with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass for F9 v1.1 booster version
  - First date of successful landing
  - boosters which have success in drone ship and payload mass between 4000 and 6000 kg
  - Total number of successful and failure outcomes
  - booster versions with maximum payload mass
  - Failed landing outcomes in drone ship incl. booster versions, and launch site for year 2015
  - Ranks of the count of landing outcomes between the date 2010-06-04 and 2017-03-20.
- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Hands-on%20Lab_%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Added markers, circles, lines and marker clusters to the Folium map:
  - Markers indicate specific points on the map, e.g. launch sites
  - Circles indicate areas around specific coordinates
  - Markers clusters indicates a group of events, e.g. launches for launch sites
  - Lines indicate the distance between two coordinates

- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Complete%20the%20Data%20Visualization%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

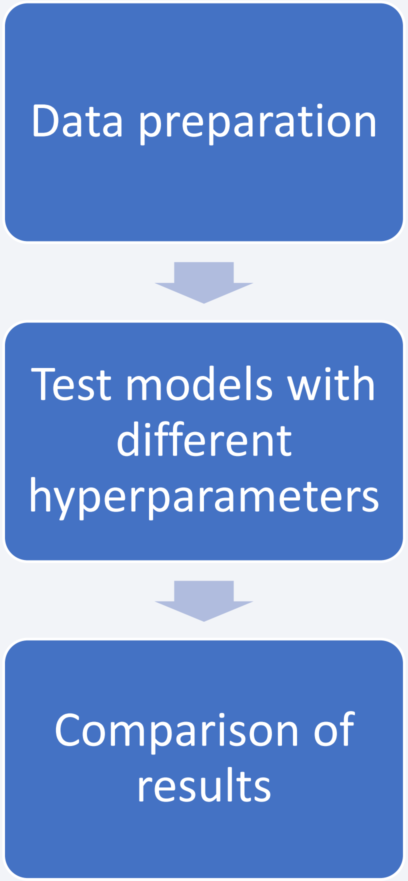
- The dashboard includes the following graphs:
  - Pie chart for percentages of successful launches per launch site
  - Scatter chart on Payload mass
- The charts can be used to analyze the relation between payloads and launch sites in order to determine the best place to launch

- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were applied:
  - Logistic regression,
  - Support vector machine
  - Decision tree
  - K-nearest neighbors

- Source code: https://github.com/michaelmaetzig/ibm-data-science-course/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

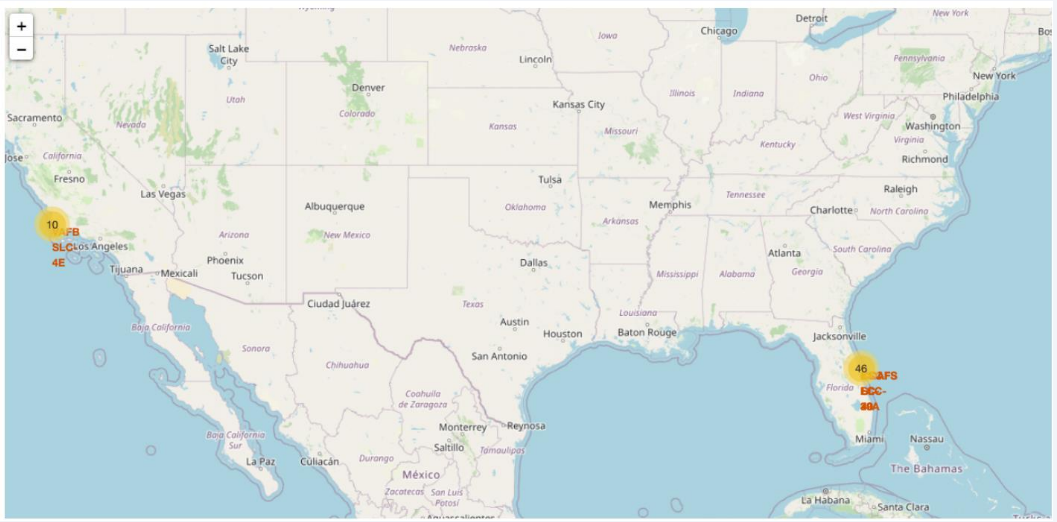


Results

- Exploratory data analysis results:
  - Four different launch sites were used by Space X
  - 2010: first launch
  - 2015: first successful landing
  - Number of landing outcomes became better with more launches
  - Average payload of F9 v1.1 booster: 2,928 kg
  - Many F9 booster version landings were successful having above average payloads

Results

- Interactive analytics – identify launch sites in safety places
  - Safety places = e.g. near sea, good logistic infrastructure around
  - Most launches happen at launch sites at the East coast



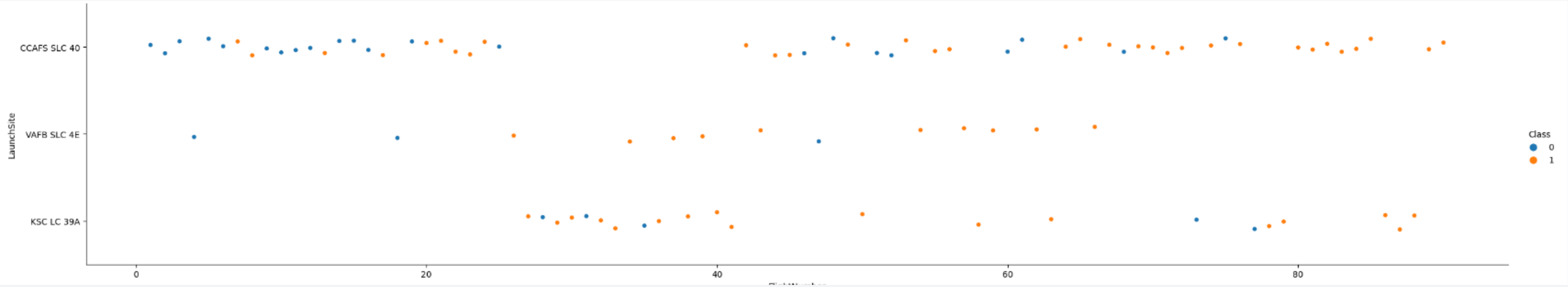
Results

- Predictive analytics – predict successful landings:
  - Best model: Decision Tree Classifier
  - Accuracy: 87%
  - Test accuracy: 94%

Section 2

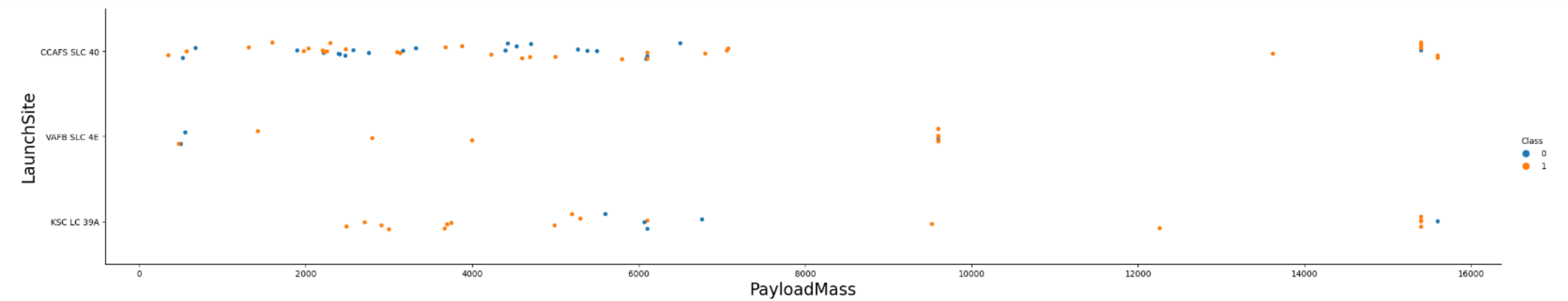
Insights drawn from EDA

Flight Number vs. Launch Site



- Best launch site with recently most successful landings: CCAF5 SLC 40
- In general: success rate improved over time

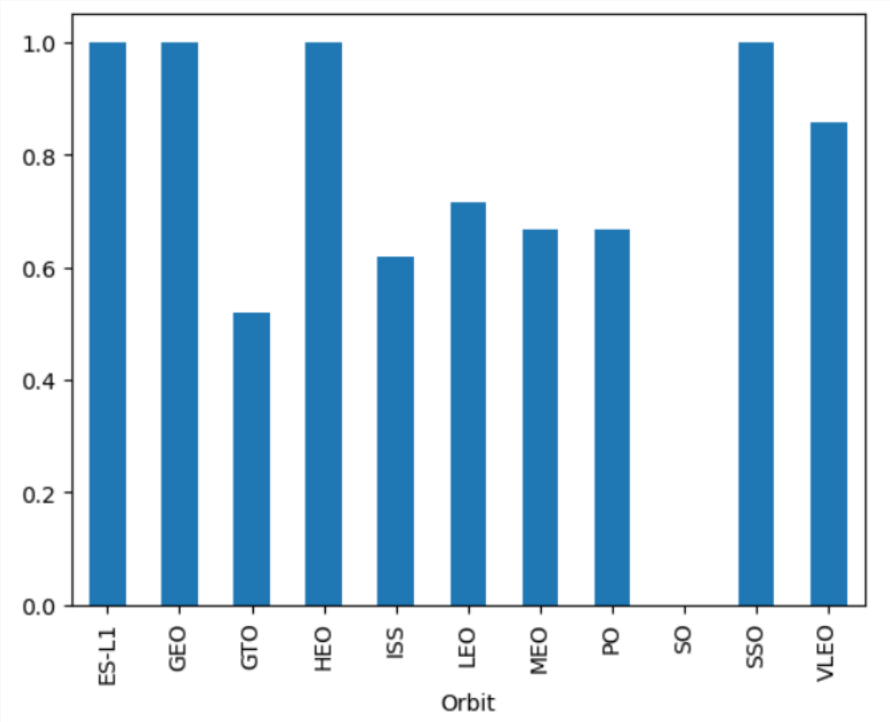
Payload vs. Launch Site



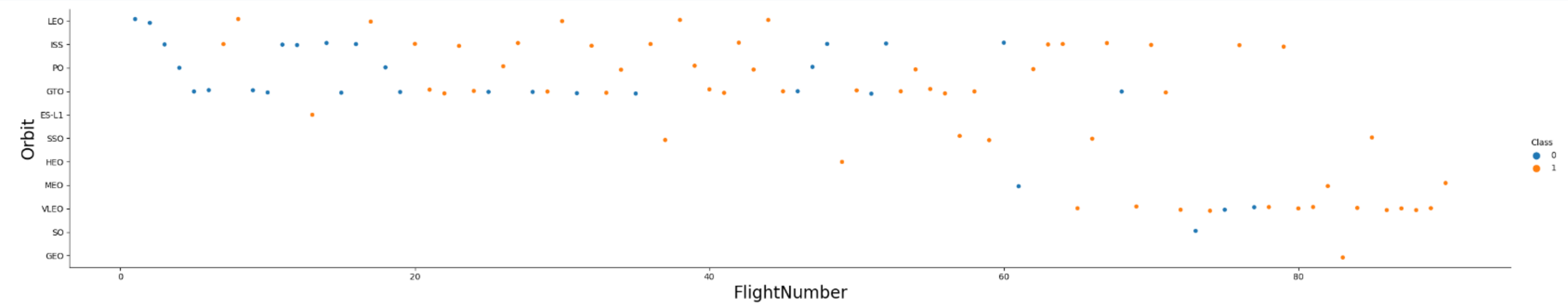
- Payloads over 9,000kg: high success rate
- Payloads over 12,000kg: only possible for CCAFS SLC 40 and KSC LC 39A

Success Rate vs. Orbit Type

- Orbits with the highest success rates:
  - ES-L1
  - GEO
  - HEO
  - SSO

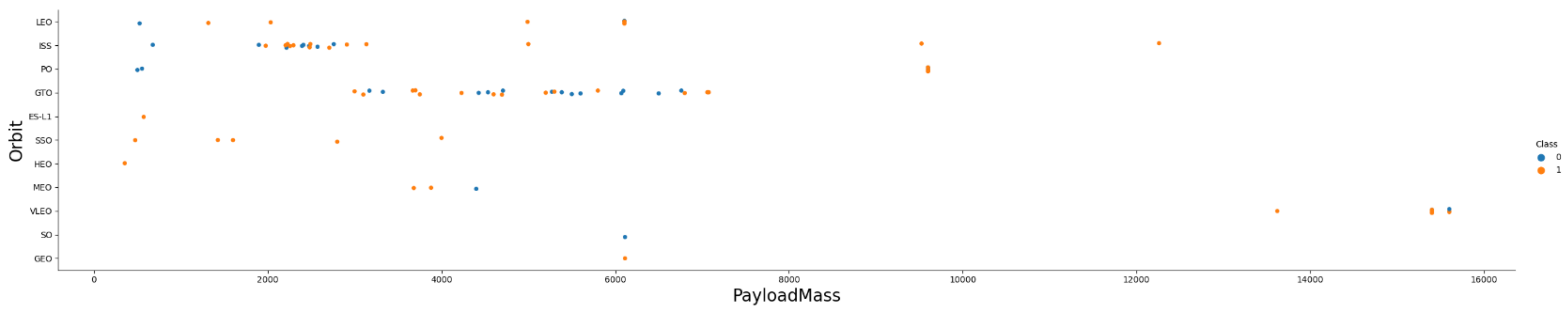


Flight Number vs. Orbit Type



- In general: success rate improved over time for all orbits
- Recent increase of frequency: VLEO orbit

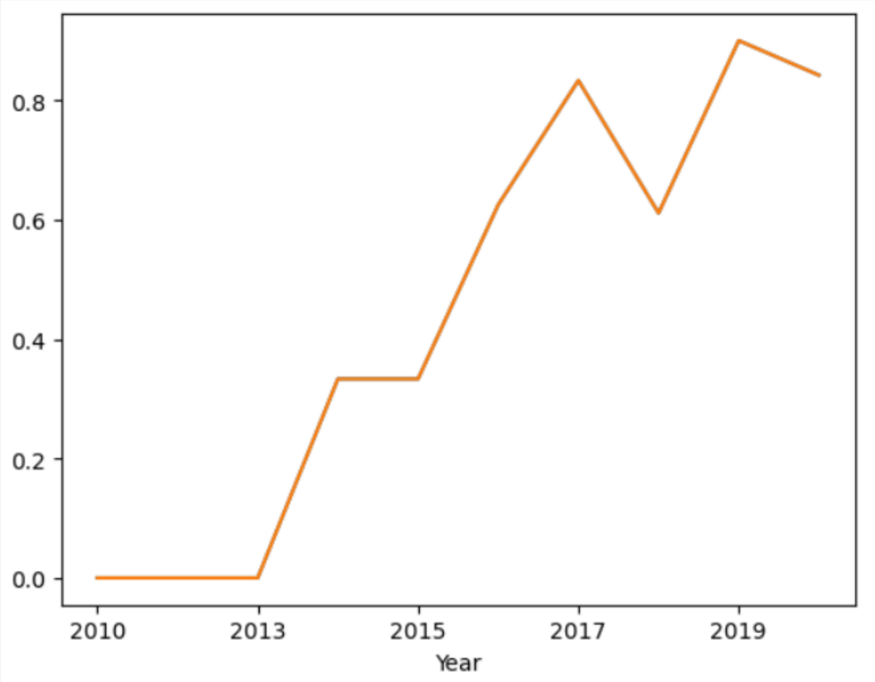
Payload vs. Orbit Type



- In general: no relation between payload and success rate to orbit GTO
- Widest range of payload and high success rate: ISS orbit

Launch Success Yearly Trend

- Success rate increased from 2013 until 2020
- First three years seem to have been a period of pure testing and development



All Launch Site Names

- Four launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Select unique occurrences for launch_site in the dataset

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

Launch Site Names Begin with 'CCA'

- Five samples of Cape Canaveral launches

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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

- Total payload carried by boosters from NASA: 45,596kg
- Calculated by summing all payloads with code “CRS” (NASA)

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- Average payload mass for F9 v1.1 booster version: 2,928kg
- Filter data on the booster version and calculating the average payload mass of the filtered dataset

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- First successful landing: 2015-12-22
- Filter the data on successful landing outcomes and getting the minimum value for date

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Successful landed booster versions on drone ship with had payload mass greater than 4000 but less than 6000:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- Select distinct booster versions after applying the filter statement

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

- Group mission outcomes, counting records

Boosters Carried Maximum Payload

- booster which have carried the maximum payload mass:

- Get the maximum payload mass and the corresponding booster versions

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

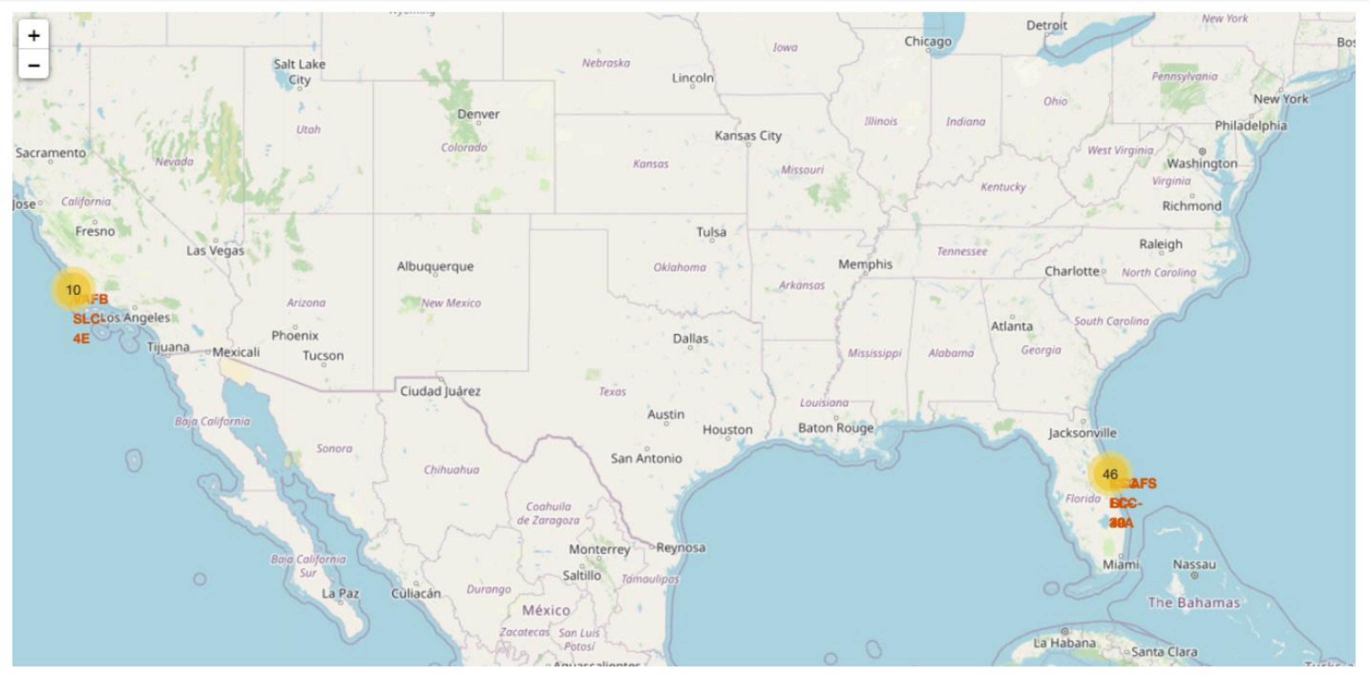
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- No attempt also has to be considered.

Section 3

Launch Sites Proximities Analysis

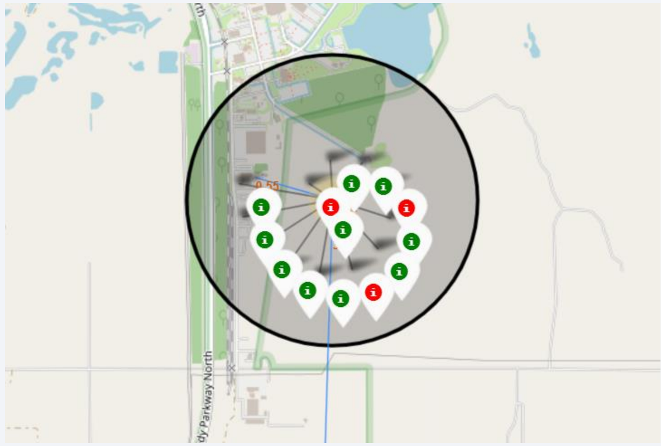
Launch sites



- Launch sites main characteristics:
  - Near sea
  - Close to roads and railroads

Launch outcomes

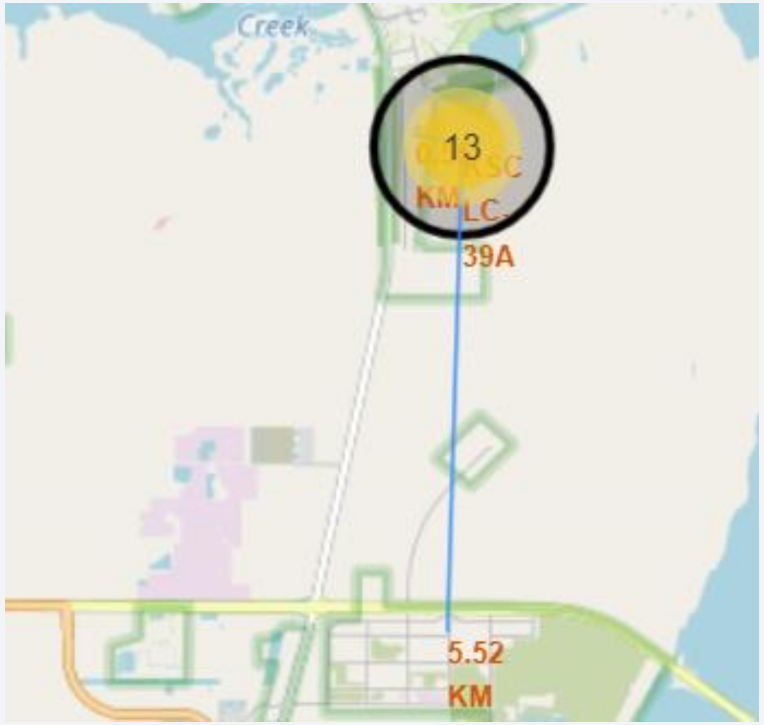
- KSC LC-39A launch site outcomes:



- Green marker: successful
- Red marker: failure

Logistics

- Logistic characteristics of KSC LC-39A:
  - Near road and railroad
  - Far from inhabited areas

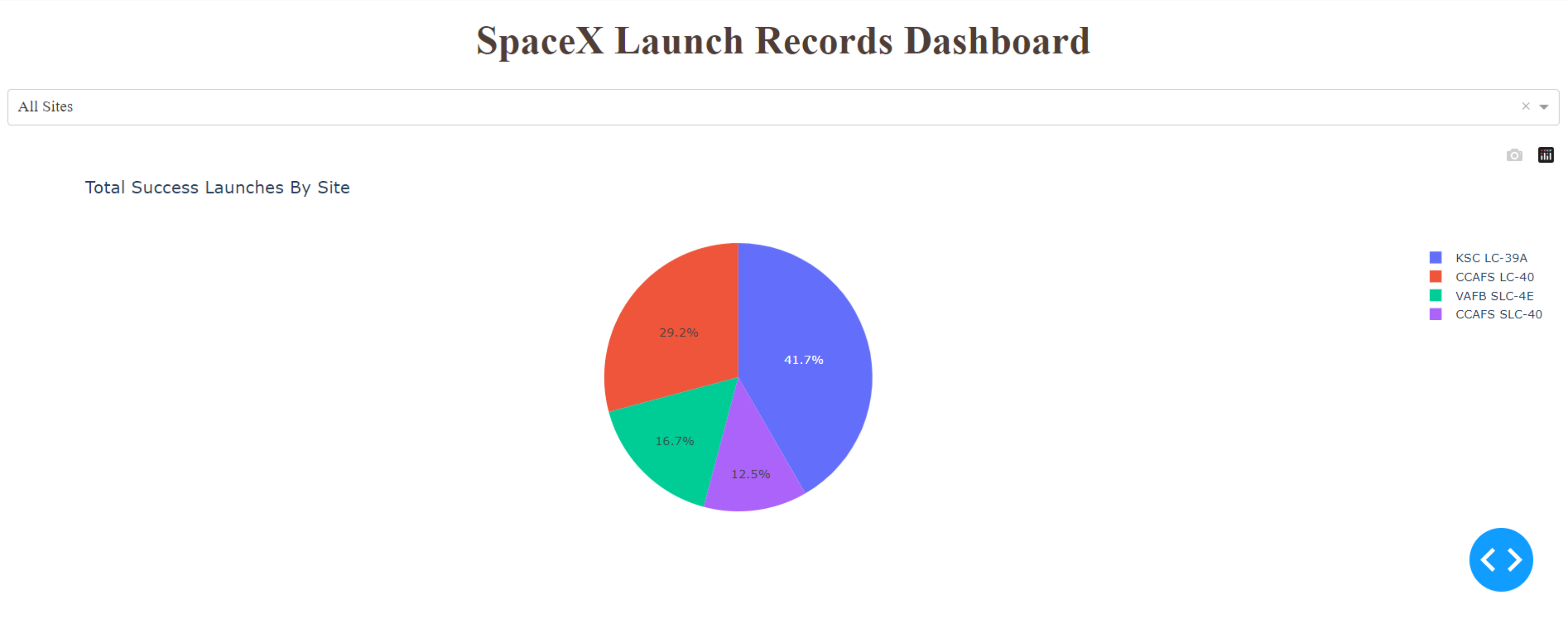


Section 4

Build a Dashboard with Plotly Dash

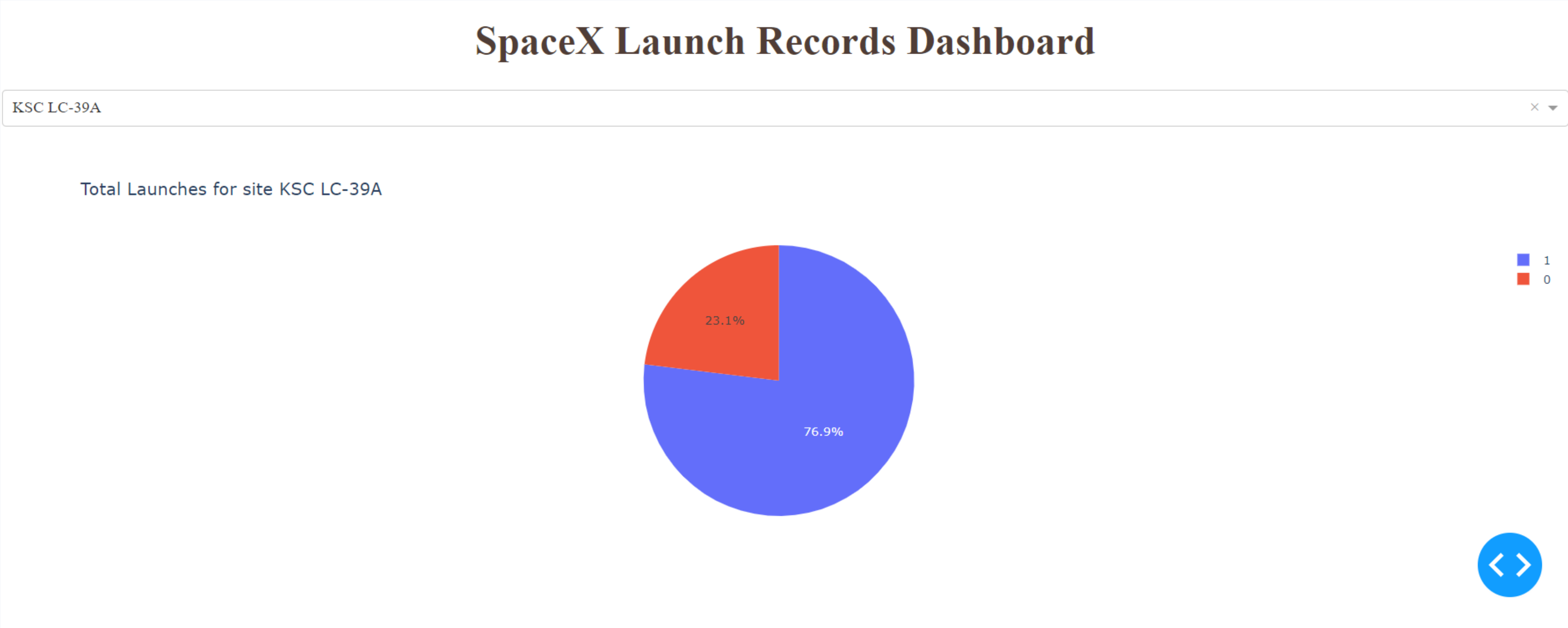
Dashboard

- Launch sites can be selected via dropdown
- The launch sites is an important factor on the success rate

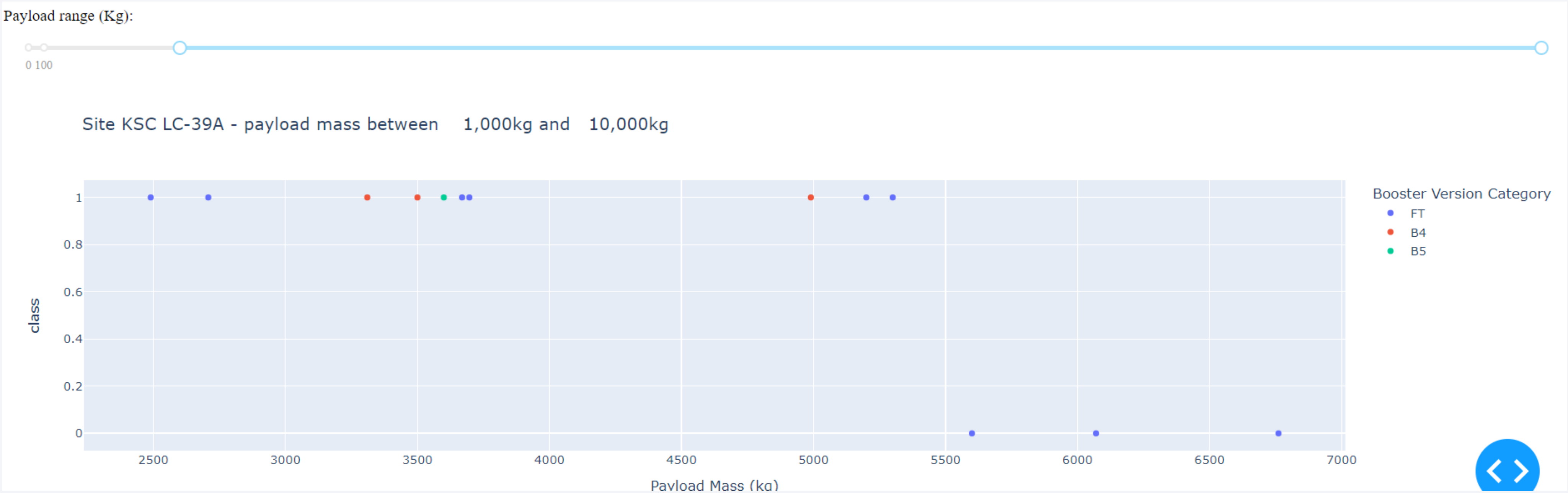


Dashboard

- Launch site with highest success rate: KSC LC-39A (76.9%)



Dashboard



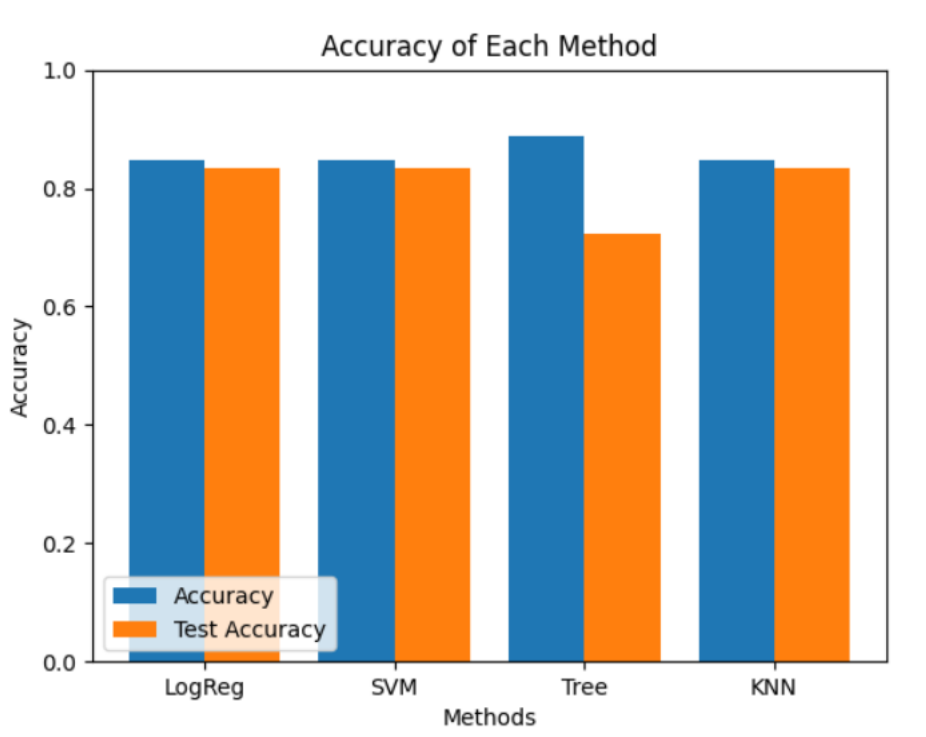
- Highest success rate: payload mass < 6,000kg and FT boosters

Section 5

Predictive Analysis (Classification)

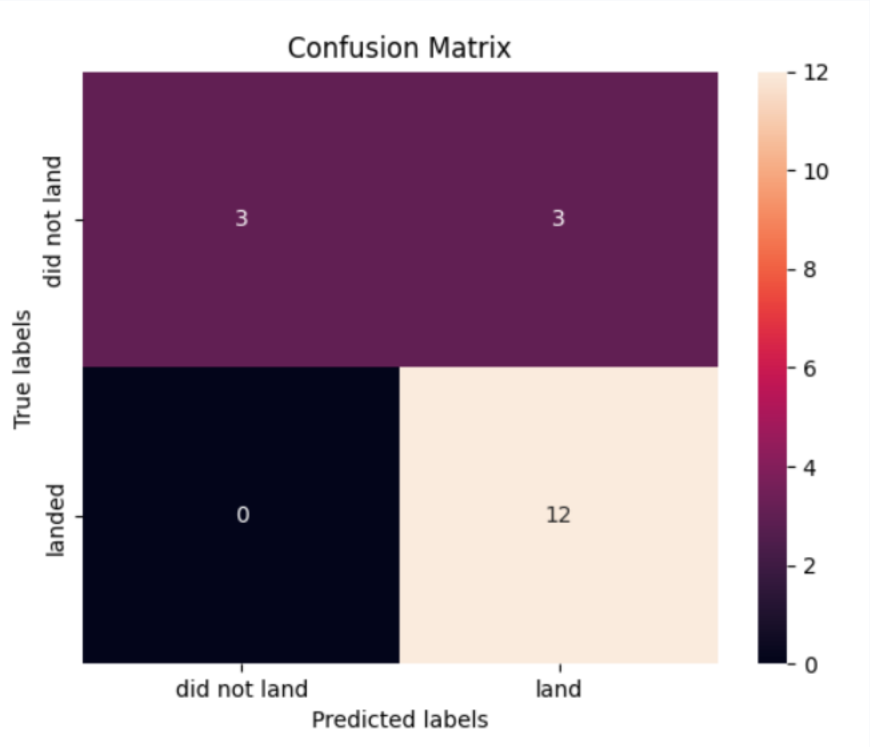
Classification Accuracy

- Four models were applied:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree
  - K-nearest neighbors
- Highest accuracy: Decision Tree
- Highest test accuracy: all models except Decision Tree



Confusion Matrix

- Confusion matrix of Decision Tree Classifier (left) and other models (right):



Conclusions

- Most successful launch site: KSC LC-39A
- Launches with payload mass > 7,000kg are less risky
- Successful landing outcomes seem to improve over time
- Decision Tree Classifier suits best to predict successful landings

Appendix

- Set np.random.see variable when testing models

Thank you!