IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Benjamin Leppard
6/7/25

Outline

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

Executive Summary

- Existing SpaceX data was used to build models that aimed to predict whether first-stage boosters would land successfully. The data was first processed, and exploratory data analysis was used to determine which features might be of importance. The data was split into training and testing datasets to assess how well the model performed with unseen data. Logistic Regression(LR), Support Vector Machine(SVM), and K-Nearest-Neighbors(KNN) models were trained and tuned using GridSearchCV to determine the optimal parameters.
- It was found that LR, SVM, and KNN models all performed equally well on the testing data at predicting successful landings, with 83.3% accuracy. A decision tree model most closely fits the training data at 89.1% accuracy. However, it performed poorly on the testing data, with an accuracy of 72.2%.

Introduction

- SpaceX is currently the most successful company involved in space travel. Their success is due in part to their relatively inexpensive launch costs, as compared to their competition. A major reason for these lower prices is the ability to successfully land first-stage boosters and reuse them.

- To predict the price of a launch, we must predict whether the first booster will land successfully. Using existing data, we can determine how factors such as payload mass, launch site, and orbit affect landing success. We will also determine the best algorithm for predicting landing success.

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX REST API
  - Web scraping of Wikipedia pages
- Perform data wrangling:
  - Filtering data
  - Dealing with missing values
  - Using one-hot encoding for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logarithmic Regression, Decision Tree, SVM, and KNN models were trained and tuned using GridSearchCV and then evaluated using unseen data.

Data Collection

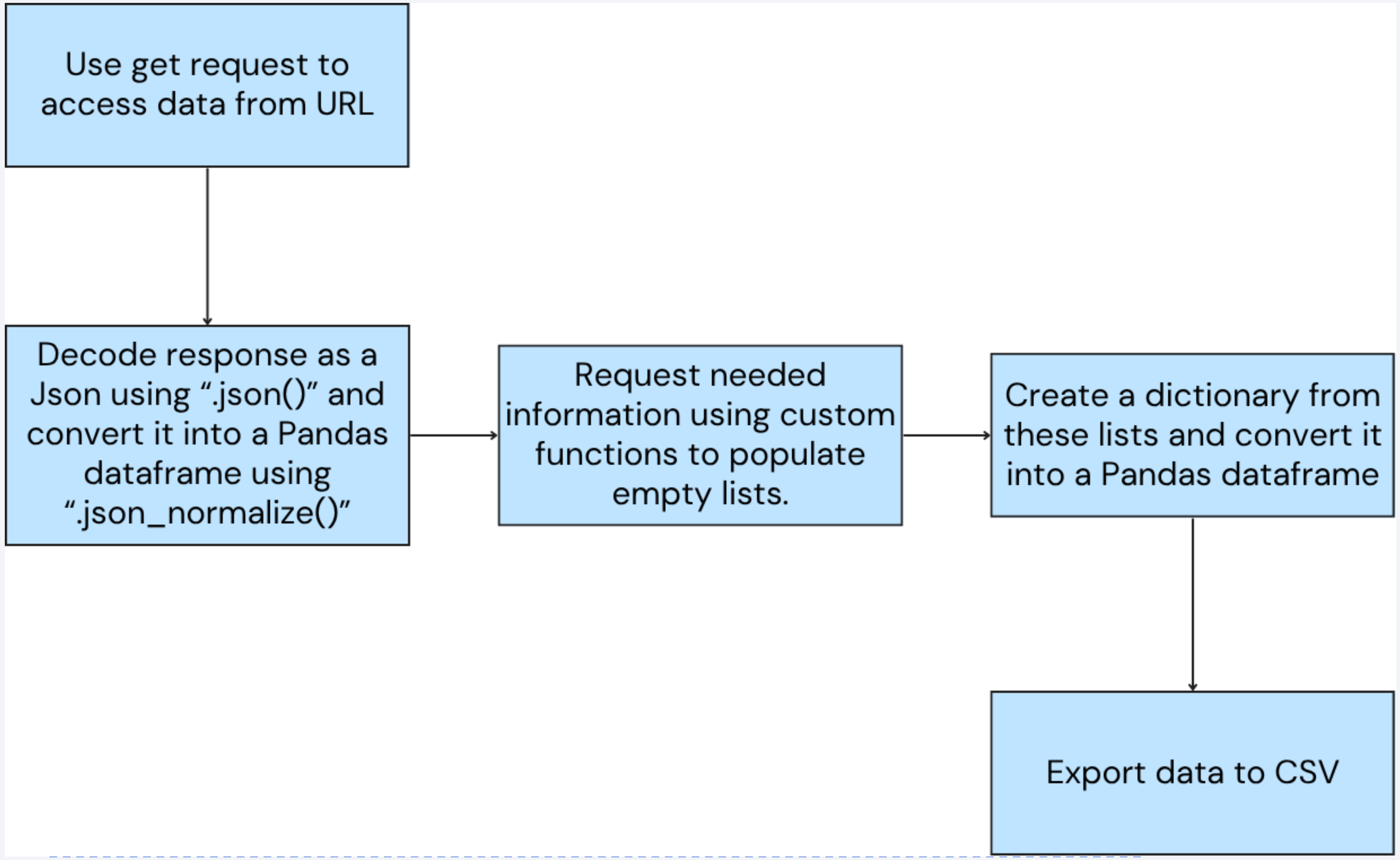
- Data was first collected using the SpaceX REST API, specifically the “/launches/past” endpoint. (api.spacexdata.com/v4/launches/past)
  - Columns collected:

    Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Sites, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

- Data was also collected by web scraping the Wikipedia page titled “List of Falcon 9 Heavy Launches.” (https://en.Wikipedia.org/wiki/List_of_Falcon _9_and_Falcon_Heavy_launches)
  - Columns collected:

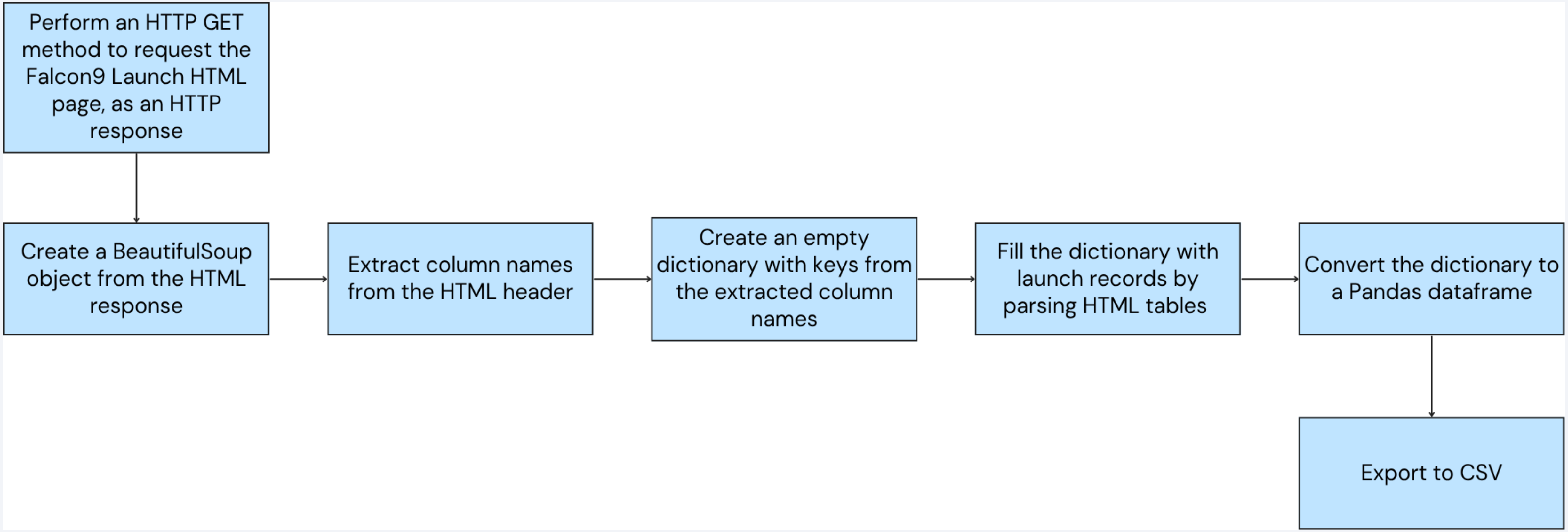
    Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



https://github.com/leppardben/Capstone_Project/blob/main/1_Data_Collection_API.ipynb

Data Collection - Scraping



https://github.com/leppardben/Capstone_Project/blob/main/2_Data_Collection_Webscraping.ipynb

Data Wrangling

- Data was filtered to only include the following features of interest:
  - ‘rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc’
- Rows with multiple cores and multiple payloads were removed, as well as all rows that included Falcon 1 launches.
- Rows where the landing outcome is missing were removed.
- Missing values for Payload Mass were replaced with the mean Payload Mass.
- Assigned numeric values to each Landing Outcome and mapped those values to 0 for unsuccessful outcomes and 1 for successful landings.

https://github.com/leppardben/Capstone_Project/blob/main/3_Data_Wrangling.ipynb

EDA with Data Visualization

- Data visualization was used to better understand the relationship between variables and investigate trends.
- Charts/plots used:
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Orbit Type vs Landing Outcome
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type

https://github.com/leppardben/Capstone_Project/blob/main/5_EDA_Viz.ipynb

EDA with SQL

- Performed SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1, listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters that have been successful on drone ships and have a payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failed mission outcomes
  - Listing the names of the booster versions that have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for the months in the year 2015
  - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 6/4/2010 and 3/20/2017 in descending order

https://github.com/leppardben/Capstone_Project/blob/main/4_EDA_SQL.ipynb

Build an Interactive Map with Folium

- Launch Site Markers:
  - Added Markers with Circles, Popup Labels, and Text Labels to NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circles, Popup Labels, and Text Labels to all Launch Sites using their latitude and longitude coordinates.
- Colored Markers displaying launch outcomes at each Launch Site:
  - Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distance between a Launch Site and its proximities:
  - Added colored Lines to show distances between the Launch Sites and their proximities like railways, highways, and Coastlines.

https://github.com/leppardben/Capstone_Project/blob/main/6_Folium_map.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection.
- Pie chart displaying total successful launches by site.
- Scatter chart of Payload Mass vs. Success Rate for each Booster Version with a slider to select Payload Mass range.

https://github.com/leppardben/Capstone_Project/blob/main/7_spacex_dash_app.py

Predictive Analysis (Classification)

- Built machine learning models using the following steps:
  - Created a Numpy array from the column “Class”
  - Standardized data using the StandardScaler and fit_transform functions
  - Split the data into training and testing sets using the test_train_split function
  - Use GridSearchCV to create models using the following algorithms:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K Nearest Neighbors (KNN)
  - Fit models to the training dataset
  - Find the best hyperparameters for each model
  - Evaluate models using the testing dataset, based on the accuracy scores and confusion matrices

https://github.com/leppardben/Capstone_Project/blob/main/8_Predictive_Analysis.ipynb

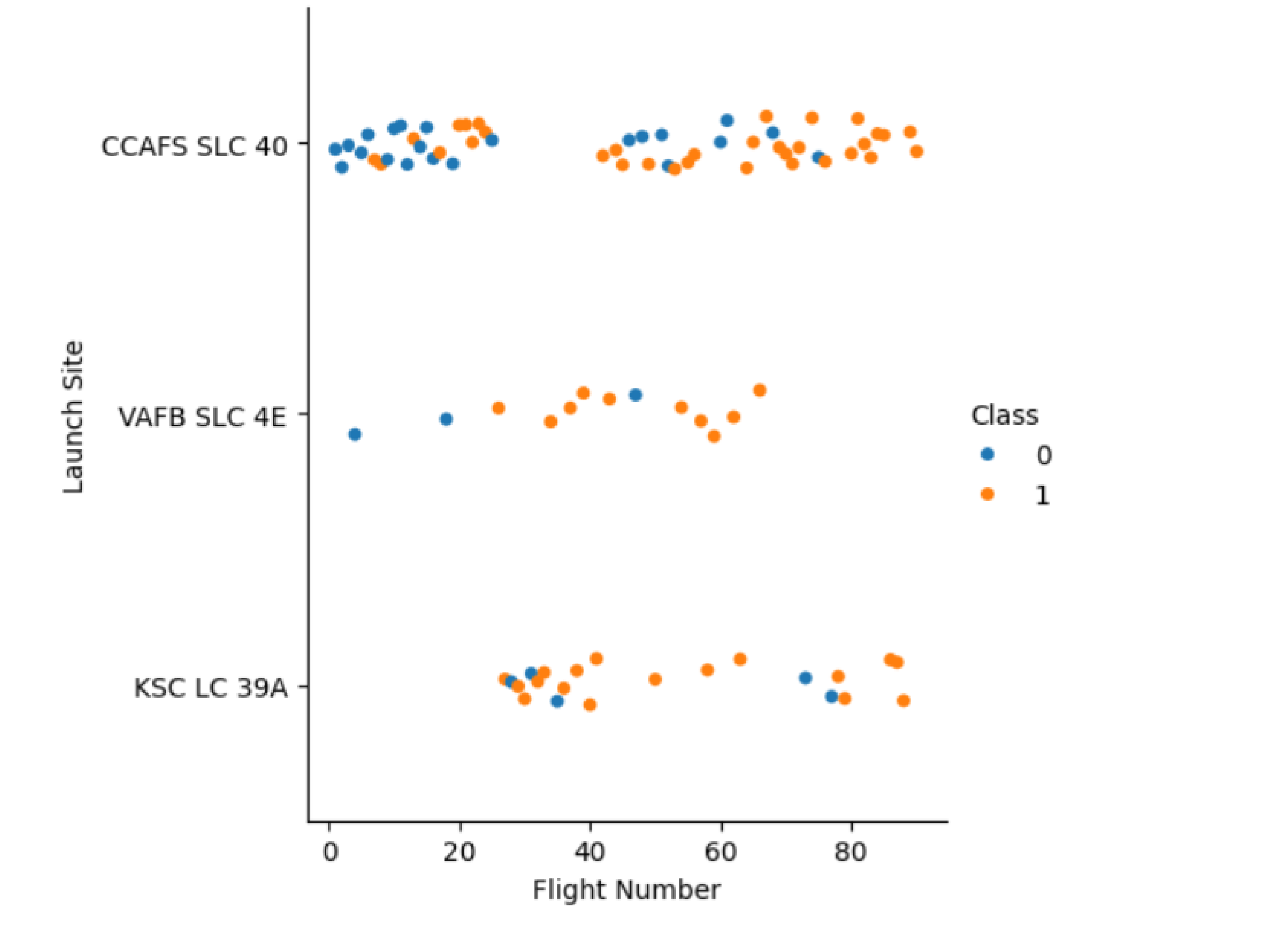
Results

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

Section 2

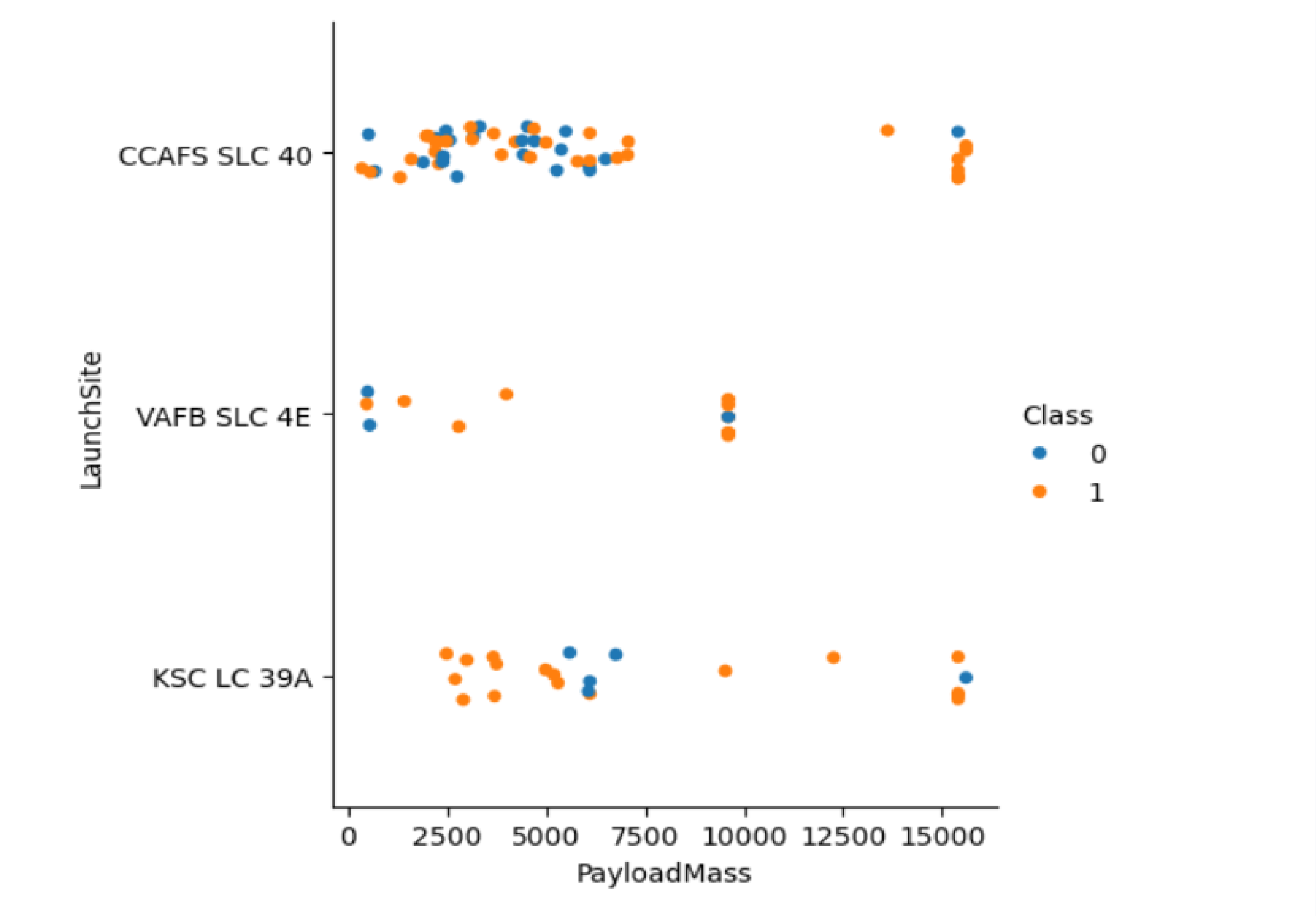
Insights drawn from EDA

Flight Number vs. Launch Site



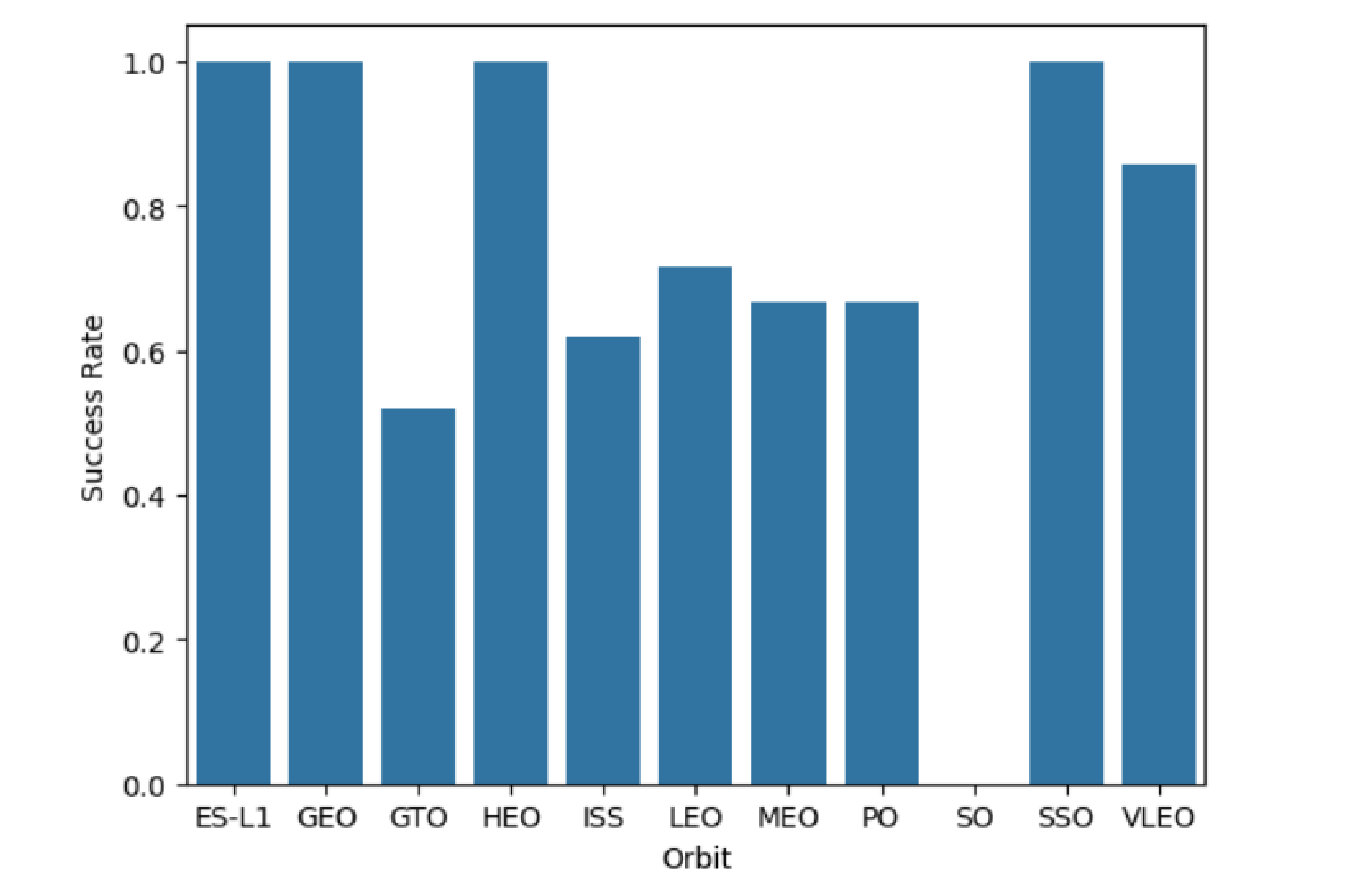
- For all launch sites, the first stage is more likely to land successfully as Flight Number increases.
- CCAFS SLC 40 is the most frequent launch site and has the worst success rate.

Payload vs. Launch Site



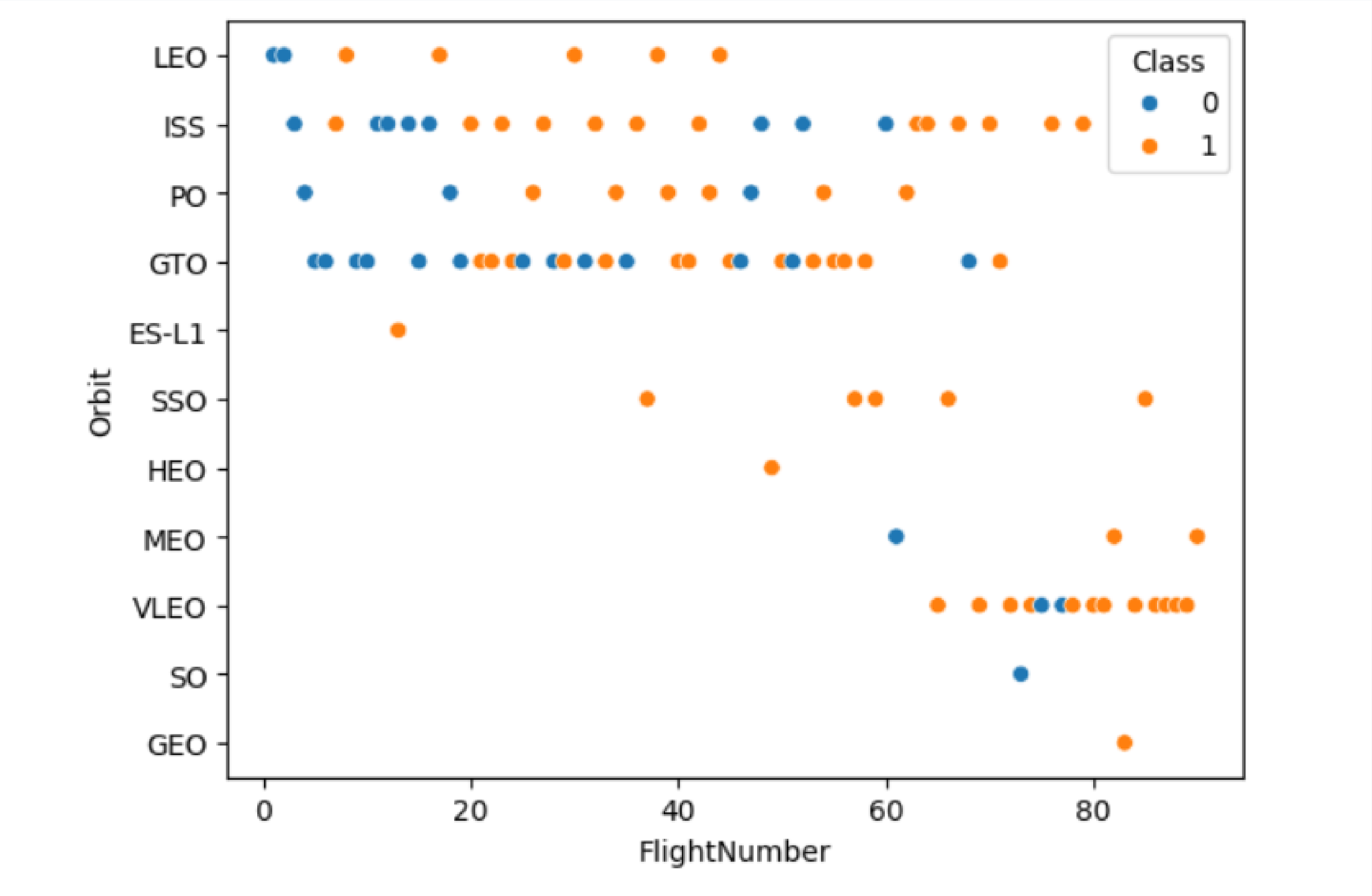
- For “CCAFS SLC 40”, the success rate significantly increases for heavy payloads.
- “KSC LC 39A” saw its highest success rate for payloads below 5000kg and above 7500kg

Success Rate vs. Orbit Type



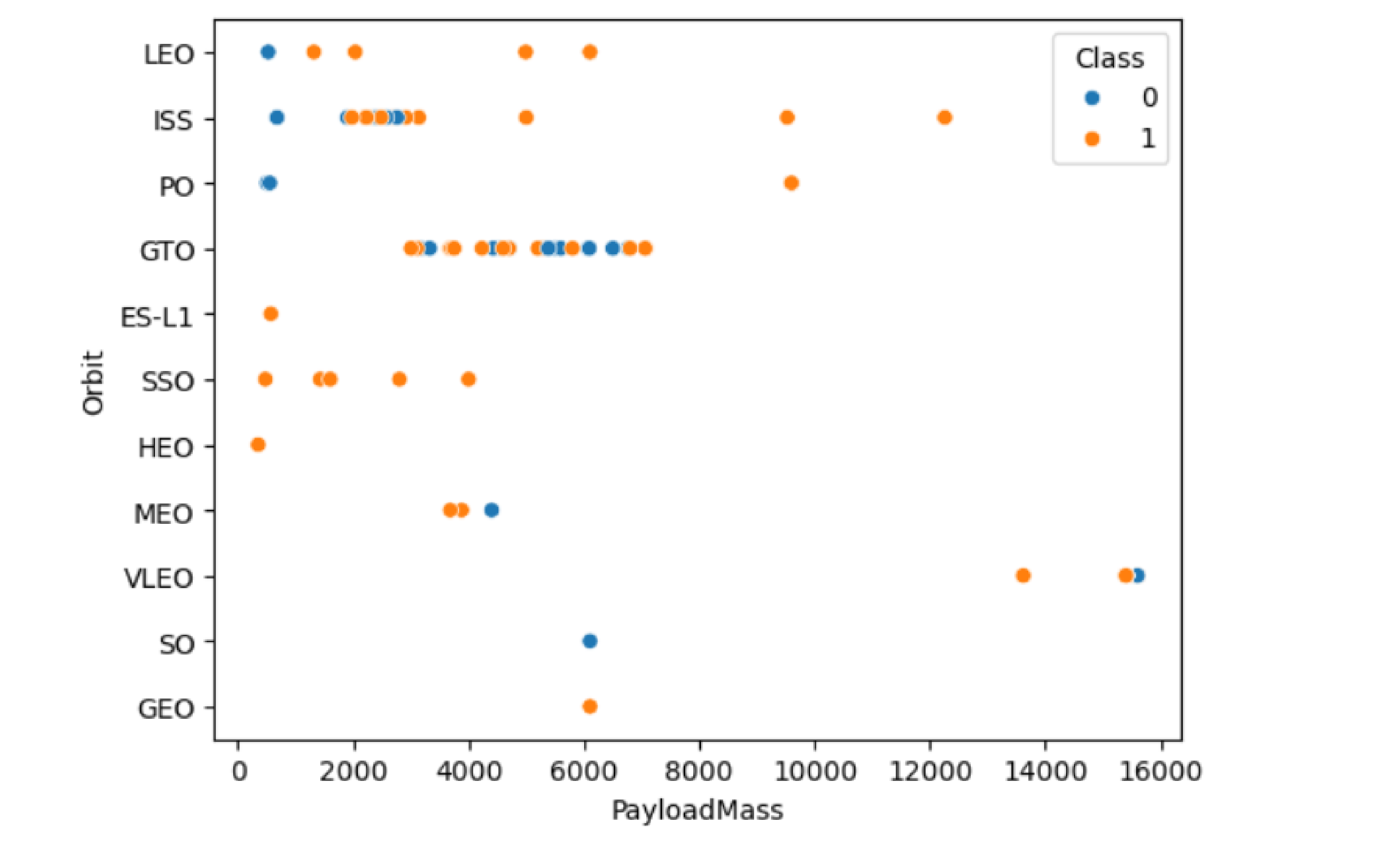
- Orbit types: “ES-L1”, “GEO”, “HEO”, and “SSO” all have a 100% success rate.
- The orbit type “SO” has yet to record a successful landing.

Flight Number vs. Orbit Type



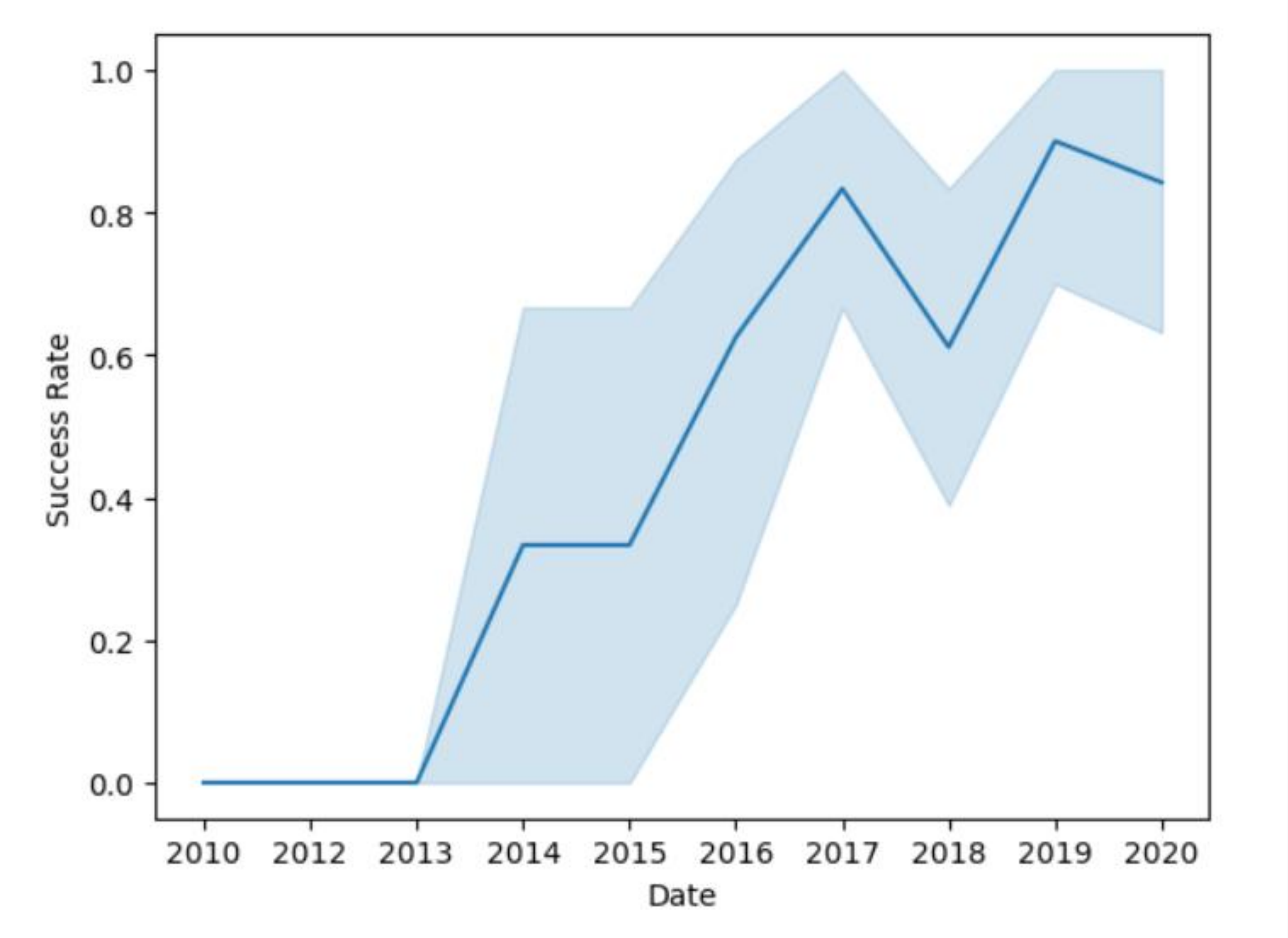
- For all orbit types, the success rate tends to go up with flight number.

Payload vs. Orbit Type



- For orbit types: LEO, ISS, and PO, the unsuccessful landings tended to occur with small payloads.

Launch Success Yearly Trend



- Overall, the success rate increases over time, starting in 2013 with the first successful landing.
- After 2017, this upward trend seems to level out, with some ups and downs.

All Launch Site Names

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

- Query result displaying the names of all launch sites.

Launch Site Names Begin 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

- Query results showing records for launch sites beginning with ‘CCA’.

Total Payload Mass

total_payload_mass
45596

- Query result showing the total payload mass carried by NASA boosters.

Average Payload Mass by F9 v1.1

average_payload_mass
2534.6666666666665

- Query result showing the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

first_successful_landing
2015-12-22

- Query results showing the date of the first successful landing outcome on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Query result showing the 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_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Query result showing the total number of successful and failed mission outcomes.

Boosters Carried Maximum Payload

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

- Query result showing the names of the boosters that have carried the maximum payload mass

2015 Launch Records

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Query result showing the failed landing outcomes on drone ships, their booster versions, and launch site names for the year 2015

Rank Landing Outcomes Between 6/4/2010 and 3/20/2017

Landing_Outcome	count_outcomes
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

- Query result displaying the counts of landing outcomes between the dates 6/4/2010 and 3/20/2017, in descending order.

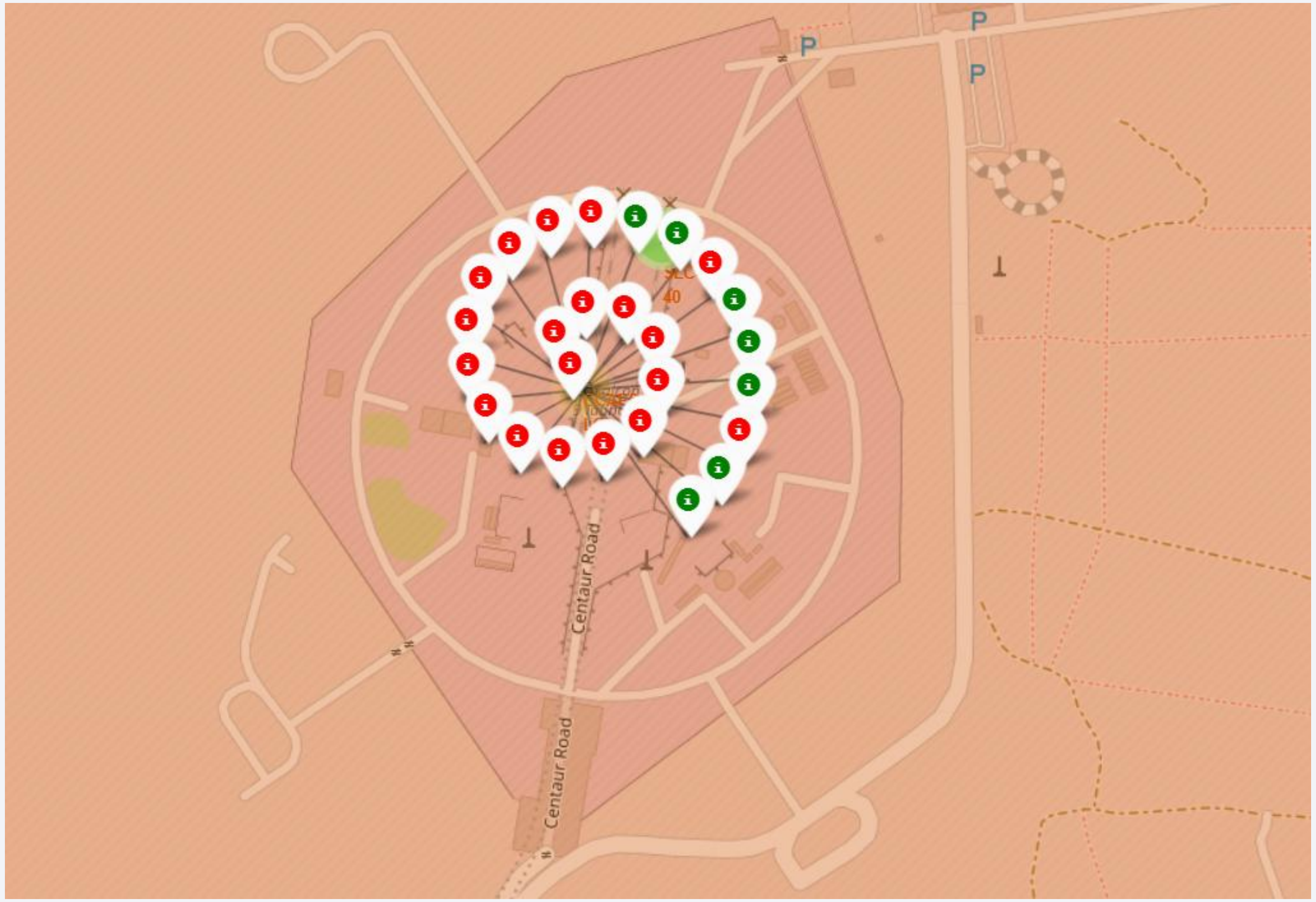
Section 3

Launch Sites Proximities Analysis

Folium Map Showing All Launch Sites

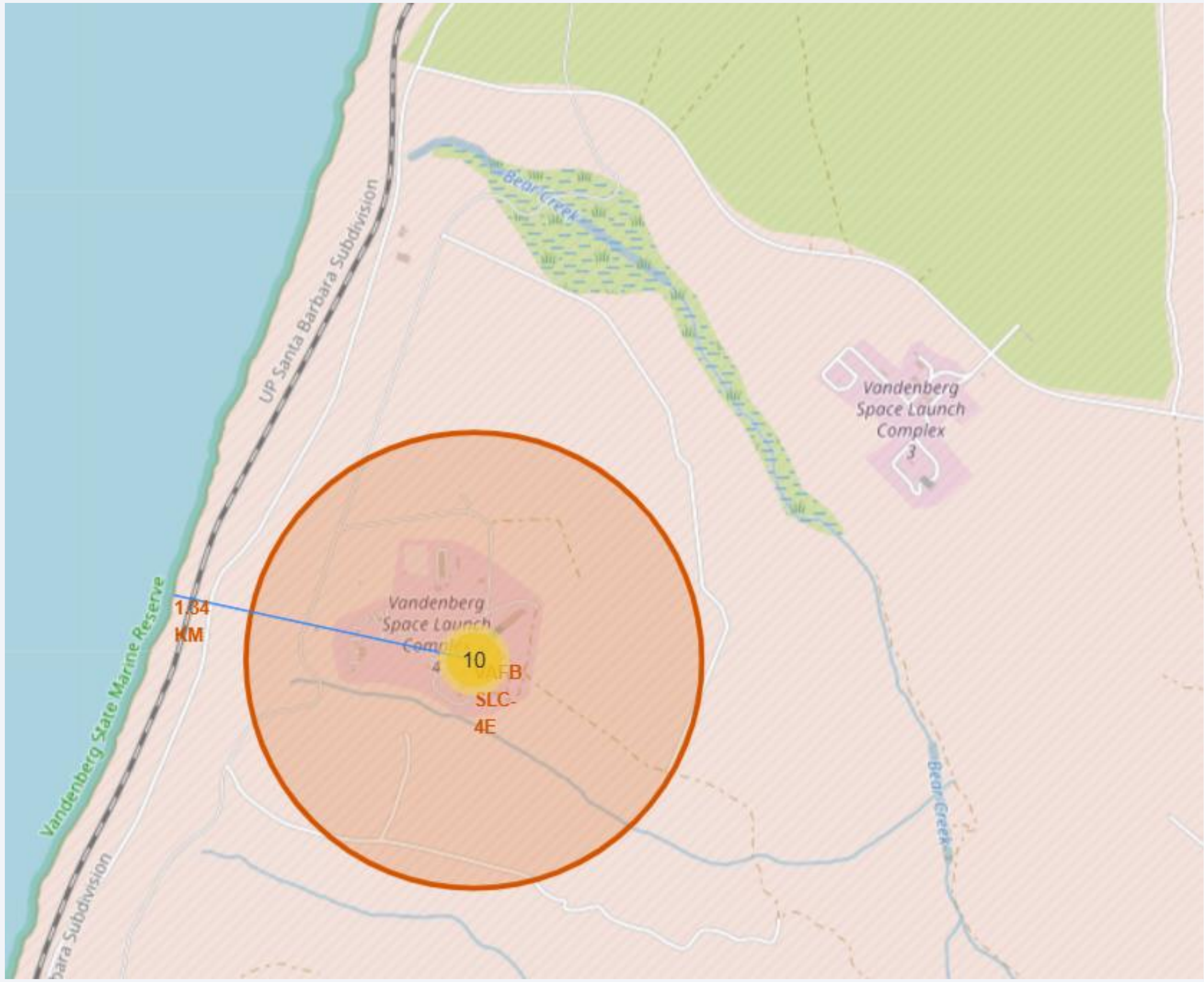
- All launch sites are located towards the equator.
- Launch sites are also located in close proximity to the coastline

Folium Map Showing Launch Site Success Rate



- Green dots represent a successful landing
- Red dots represent failed landings

Folium Map Showing Distance from Launch Site to Coastline

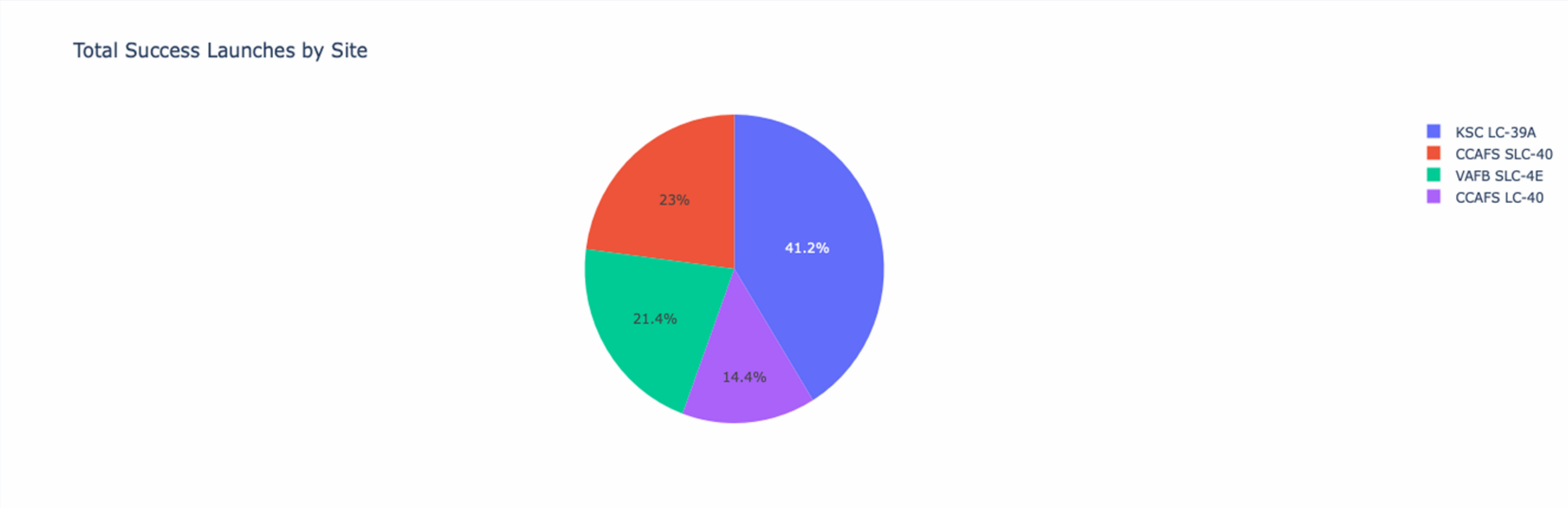


- The distance line highlights the proximity to the coastline (1.54 km)

Section 4

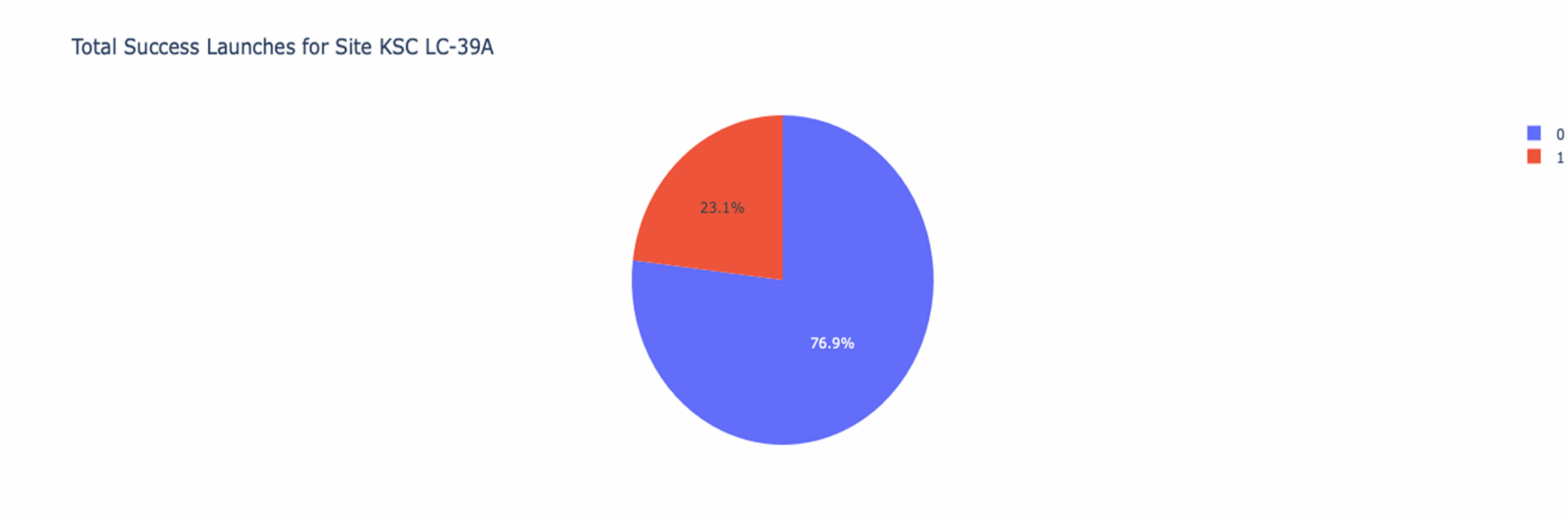
Build a Dashboard with Plotly Dash

Total Successful Landings by Site



- Site “KSC LC-39A” has the most successful landings

Landing Success Rate for “KSC LC-39A”

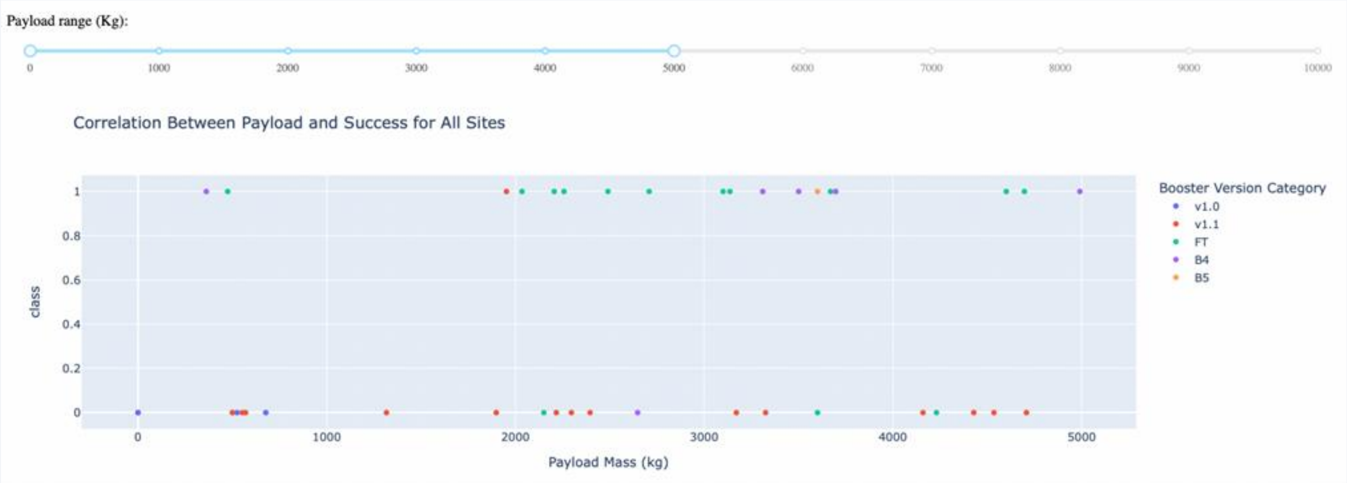


- Site “KSC LC-39A” has a success rate of 76.9%

Section 5

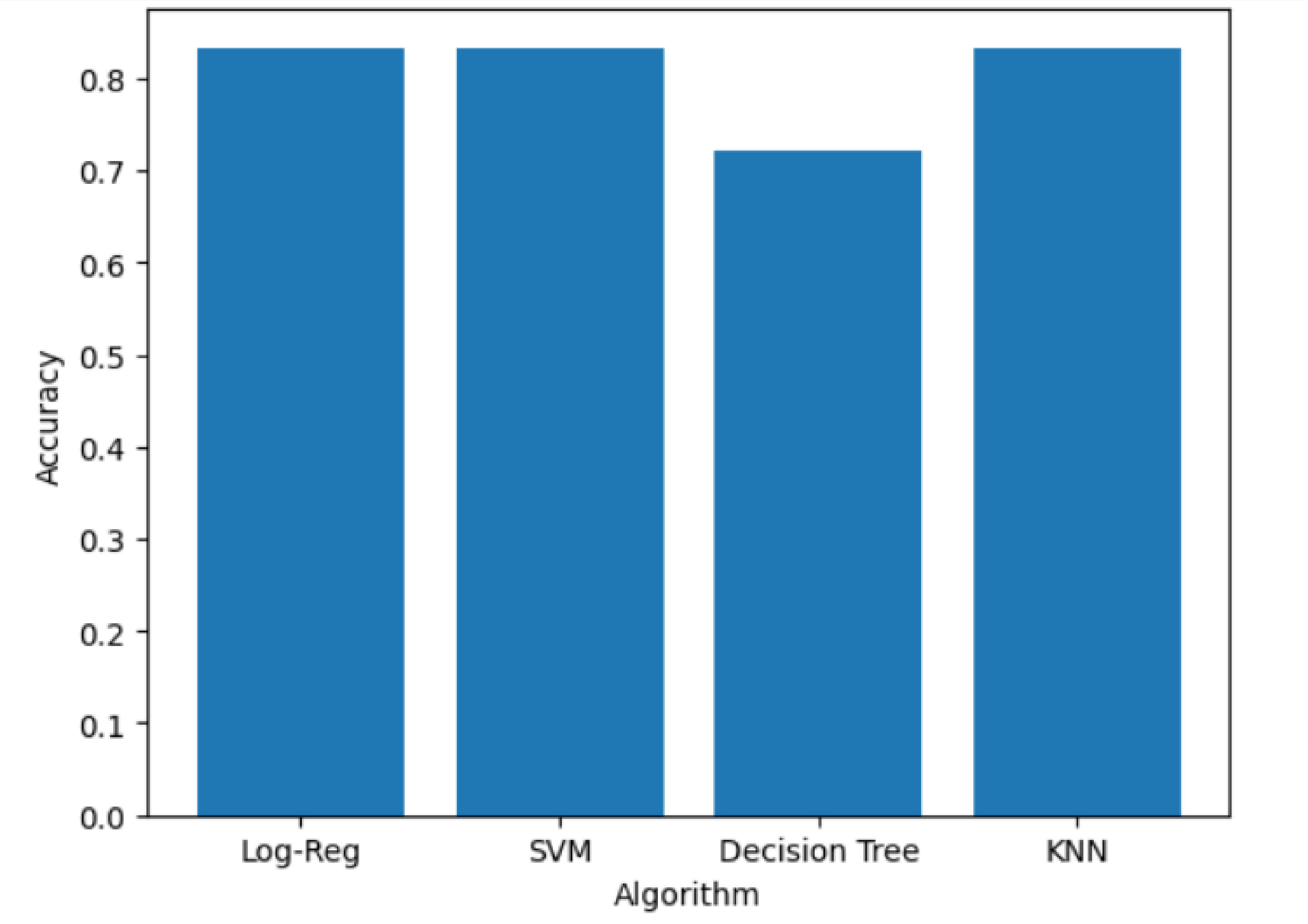
Predictive Analysis (Classification)

Payload Mass vs Landing Success



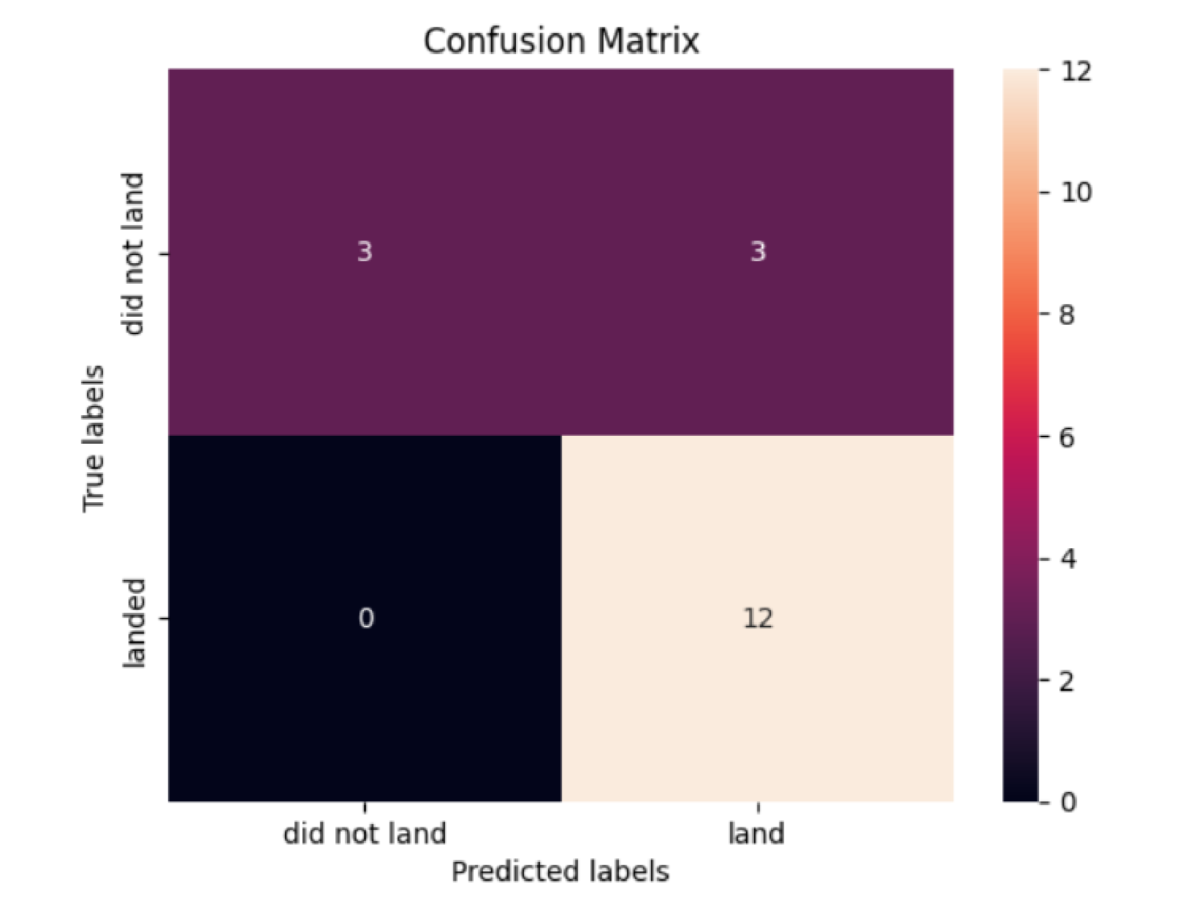
- Higher payloads resulted in a higher rate of success

Classification Accuracy



- Logistic Regression, SVM, KNN models all displayed approximately the same accuracy in predicting test data.

Confusion Matrix



- The confusion matrix shows that the mislabeled classes were false positives.

Conclusions

- The data can be most effectively modeled using KNN, SVM, or Logistic Regression
- Landing success rates have increased over time
- Payload mass has a positive correlation with landing success.
- KSC LC-39A has the highest success rate of all launch sites.
- Most launch sites are located near the equator and are in close proximity to the coast.
- Orbits ES-L1, GEO, HEO, and SSO have a100% success rate.

Thank you!