Applied Data Science Capstone

SAMUEL SCHWARZKOPF

1/2/2024

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Visual Analytics with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis using Machine Learning

Summary of Results

- EDA Results
- Interactive Analytics with Screenshots
- Predictive Analytics Results

Introduction

Project Background and Context

- SpaceX is the leading company in the space travel market. They advertise their Falcon 9 rocket launches on their website at a cost much lower than their competitors. The cost savings can be attributed to their ability to reuse the first stage. By determining if the first stage will land, we can find out the cost of a launch. This project will predict if SpaceX will reuse the first stage using public information and machine learning models.

Questions to be Answered

- How do variables such as launch site, orbits, number of flights, and payload mass affect the probability of success of the first-stage landing?
- As time goes on, does the rate of successful first-stage landings increase?
- What is the best predictive analytics method to determine if the first stage will be reused?

Section 1: Methodology

Methodology

Executive Summary

Data collection Methodology

- Use of SpaceX Rest API
- Web Scraping from Wikipedia

Perform Data Wrangling

- Filter the data
- Replace missing values
- Convert outcomes to 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

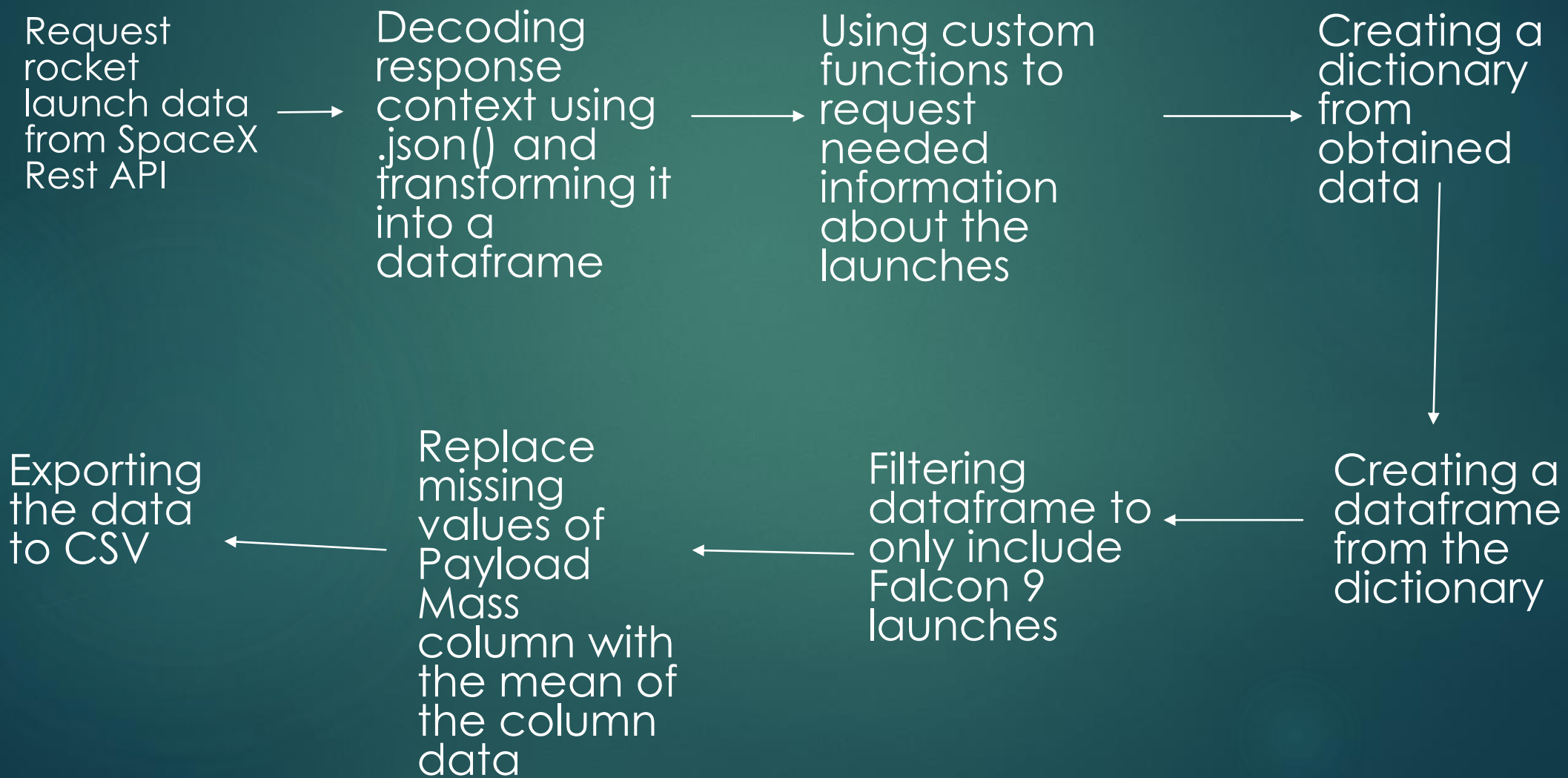
- Building, refining, and testing analytical models to determine the most accurate model

Data Collection

- Data collection was performed using API requests from SpaceX Rest API and Web Scraping SpaceX's Wikipedia page

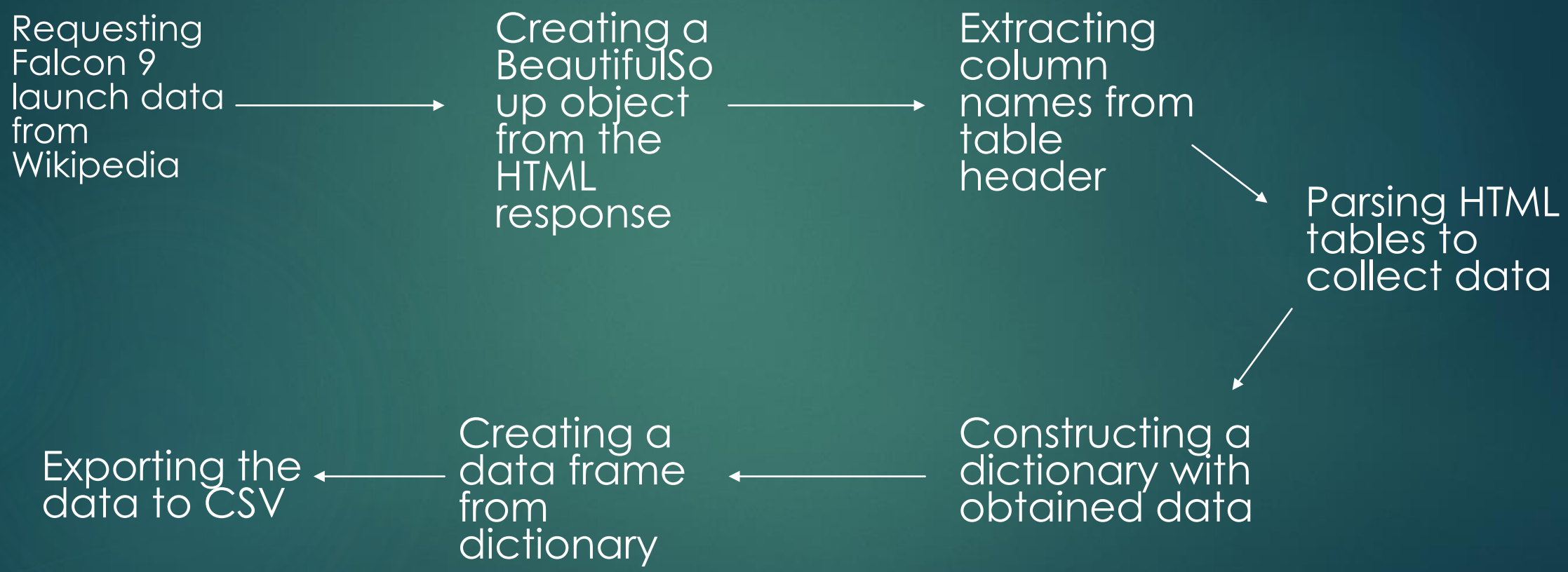
Data Collection: SpaceX Rest API

GitHub Data Collection API



Data Collection: Web Scraping

Git Hub: Data Collection Web Scraping



Data Wrangling

- Purpose was to convert different outcomes into binary classification: 1 = successful landing , 0 = unsuccessful landing

Git Hub: Data Wrangling

Steps

1. Exploratory Data Analysis
2. Calculate number of launches on each site
3. Calculate number and occurrence of each orbit
4. Calculate number and occurrence of mission outcome of orbits
5. Create landing outcome label from Outcome column
6. Export to CSV

EDA with Data Visualization

- Charts produced: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, Success Rate Yearly Trend
- Scatter plots were used to show the relationship between variables
- Bar plots were used to compare discrete categorical variables with a measured value
- Line charts were used to show data trends over time

Git Hub: Data Visualization

EDA with SQL

Git Hub: EDA with SQL

Performed SQL queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display 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 have success in drone ship and have a payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Build an Interactive Map with Folium

Git Hub: Mapping with Folium

Adding Markers of all Launch Sites:

- Added marker with circle, popup label and text label of Nasa Johnson Space Center and all launch sites using longitude and latitude coordinates

Colored Markers of Launch Outcomes:

- Added colored markers of launches with green(success) and red(failure) using marker cluster to identify which launch sites have higher success rates

Distance Markers

- Added lines to show distances between a particular launch site and its proximities to railway, coastline, closest city, and highway

Build a Dashboard with Plotly Dash

- Added a dropdown list containing different launch sites
- Added pie charts showing total successful launch counts and the success vs. failure counts for selected launch site
- Added a slider to select payload range
- Built scatter chart of payload mass vs success rate for each booster version to show the correlation

Git Hub: Dashboard with Dash Plotly

Predictive Analysis (Classification)

Git Hub: Predictive Analysis

Using to_numpy() to create a NumPy array from the column Class

Standardize the data and reassign it to variable x using transform

Split data X and Y into training and test data

Create logistic regression object then create a GridSearchCV object and find best parameters

Calculate accuracy on the test data using the method score

Inspect the confusion matrix of each model

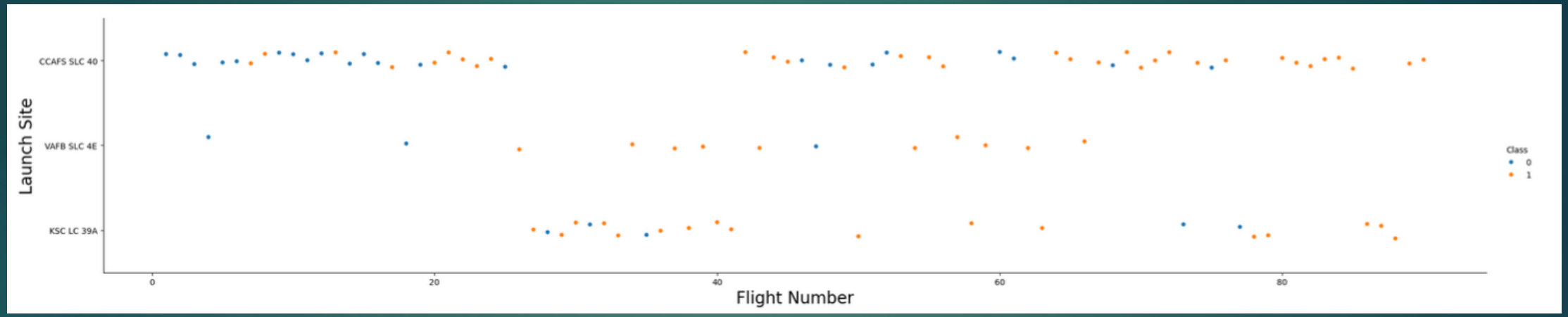
Use Jaccard_score and F1_score metrics to find the best performing model

Results

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

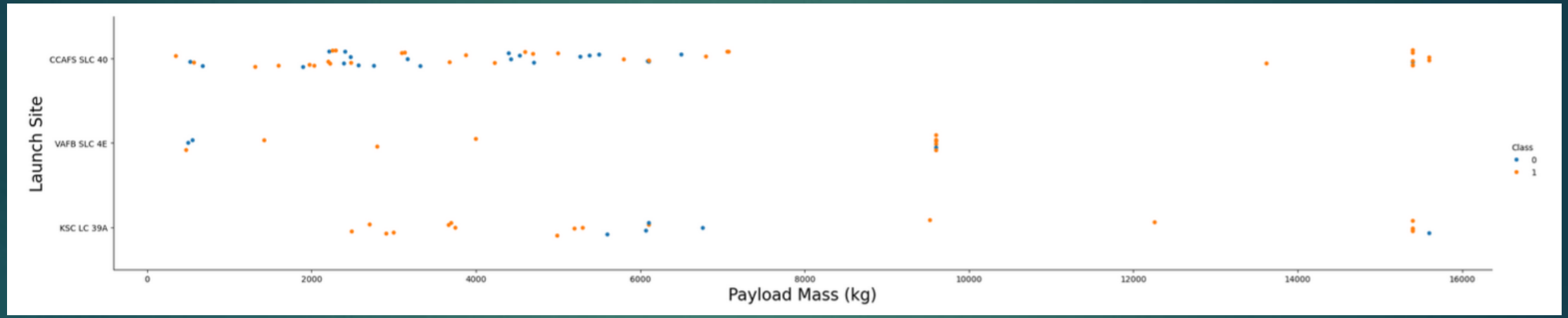
Section 2: Insights Drawn from EDA

Flight Number vs. Launch Site



- As the number of flights increases from each launch site, the rate of successful launches increases
- Most launches came from the CCAFS SLC 40 launch site
- The rate of success is higher than the rate of failure for all launch sites

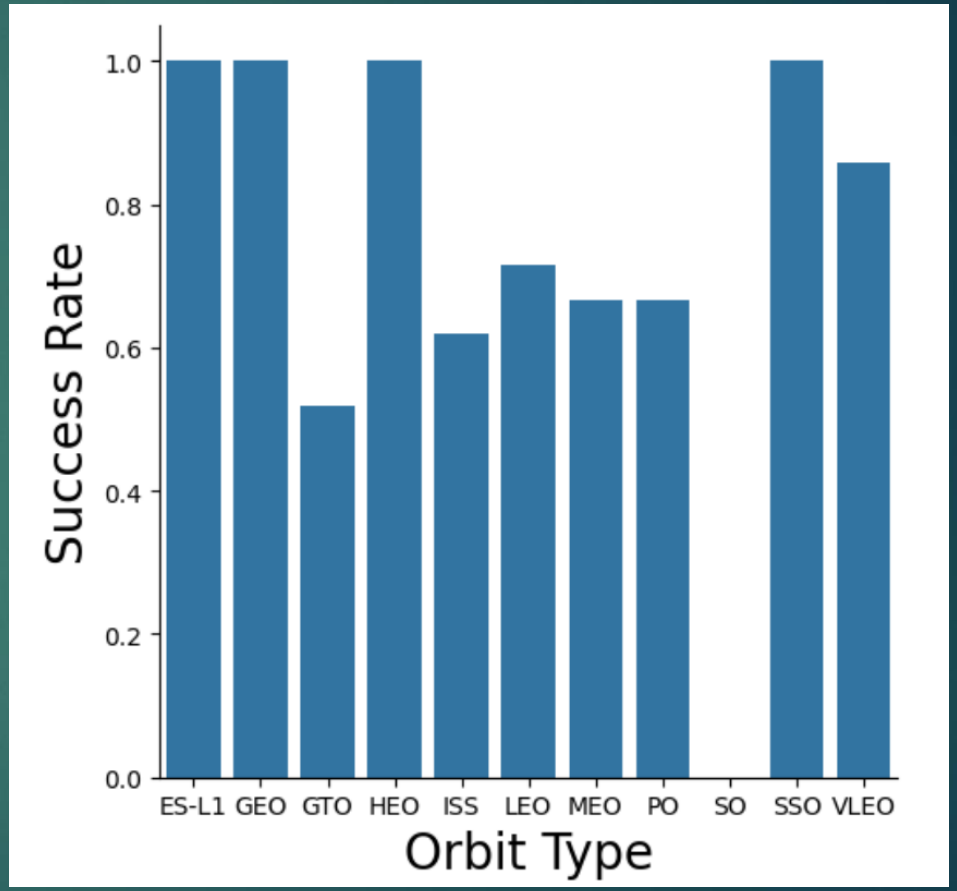
Payload vs. Launch Site



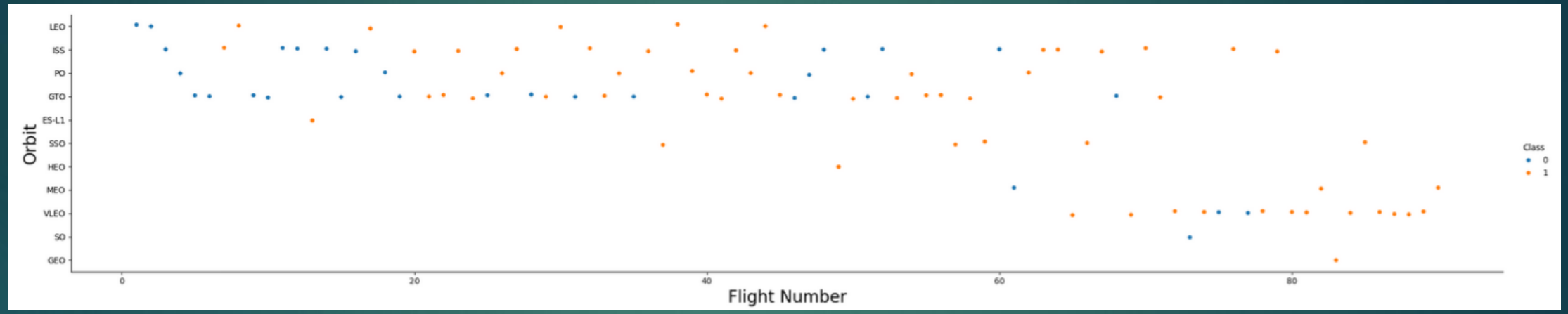
- Launch sites CCAFS SLC 40 and KSC LC 39A had the highest payloads with success
- Most payloads were below 7000 kg
- The rate of success increases as the payload mass increases

Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO and SSO have 100% success rates
- SO has no success
- GTO, ISS, LEO, MEO and PO had average success rates between 50% and 70%

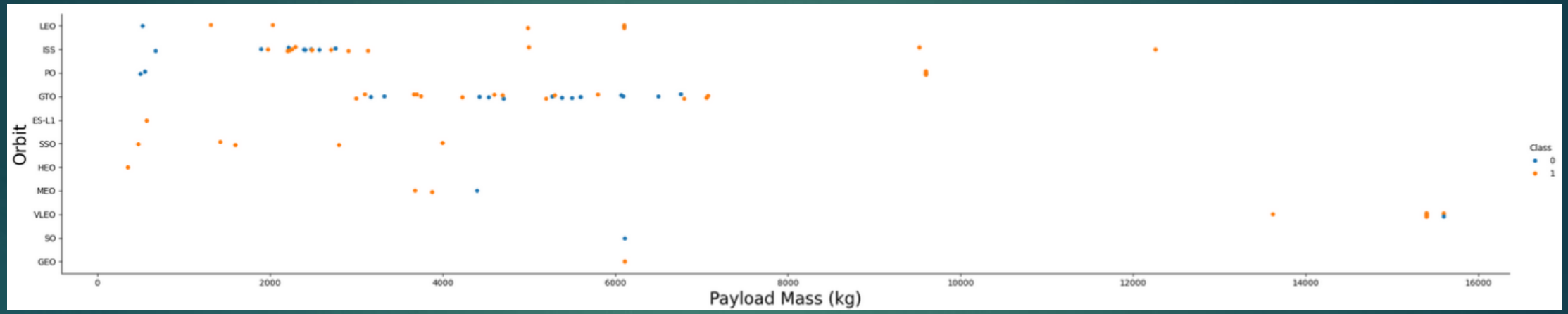


Flight Number vs. Orbit Type



- The first flights with orbit types LEO, ISS, PO, GTO, MEO and SO resulted in failure
- Orbit types ISS and GTO had the highest number of flights
- Orbit types SSO, HEO and GEO had 100% success

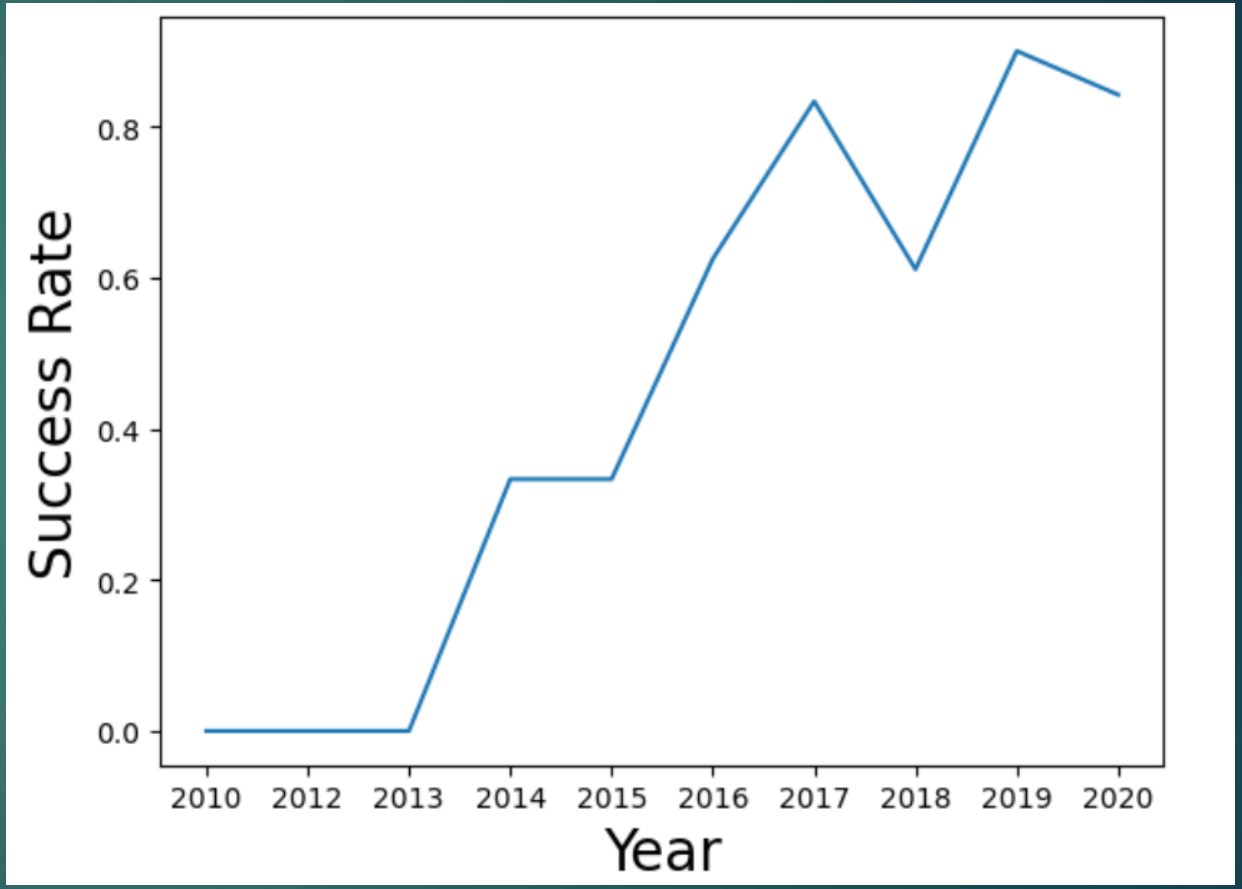
Payload vs. Orbit Type



- PO, ISS and VLEO had success with payloads greater than 8000kg
- SSO had success with all payloads attempted

Launch Success Yearly Trend

- There is a positive correlation between success rate and year
- 2018 experienced a dip in success rate



All Launch Site Names

Displaying all launch site names

Out[20]:

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

Launch Site Names Begin with ‘CCA’

Displaying 5 launch site names beginning with the string ‘CCA’

Out[21]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parac
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 (parac
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

Query displaying total payload mass in kg from boosters launched by NASA

Out[22]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Query displaying the average payload mass by booster version F9 v1.1

Out[23]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

Query displaying the date of the first successful landing in ground pad

Out[25]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Query displaying booster versions with successful drone ship landing holding payloads between 4000 and 6000

Out[26]:

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

Total Number of Successful and Failure Mission Outcomes

Query displaying total number of successful and failure mission outcomes

Out[28]:

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

Boosters Carried Maximum Payload

Displaying a list of boosters carrying the maximum payload mass

Out[29]:

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

Query displaying launches with landing outcomes resulting in failure during the year 2015

[34]:

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)

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

Query displaying landing outcomes between 2010-06-04 and 2017-03-20

[36]:

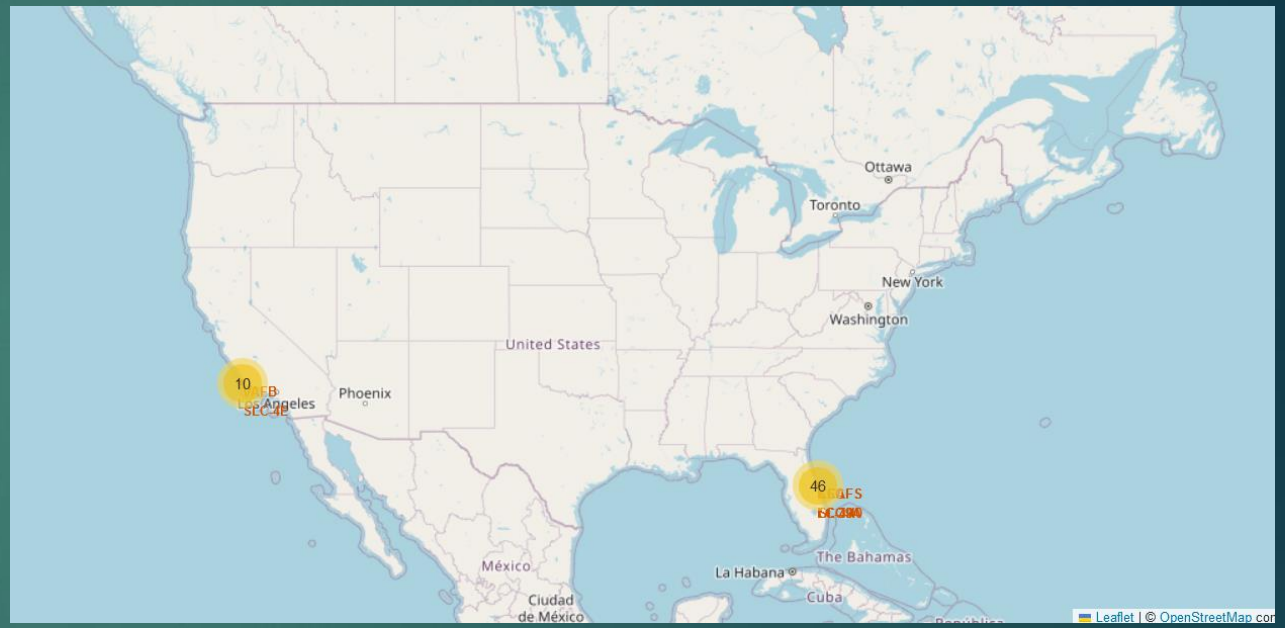
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3: Launch Sites Proximity Analysis

All Launch Site Location Markers

All launch sites are in the southern half of the United States, close to the equator. Launches near the equator travel at a faster speed than further from the equator. Launches are launched close to the equator to better match the speed of Earth and keep in orbit

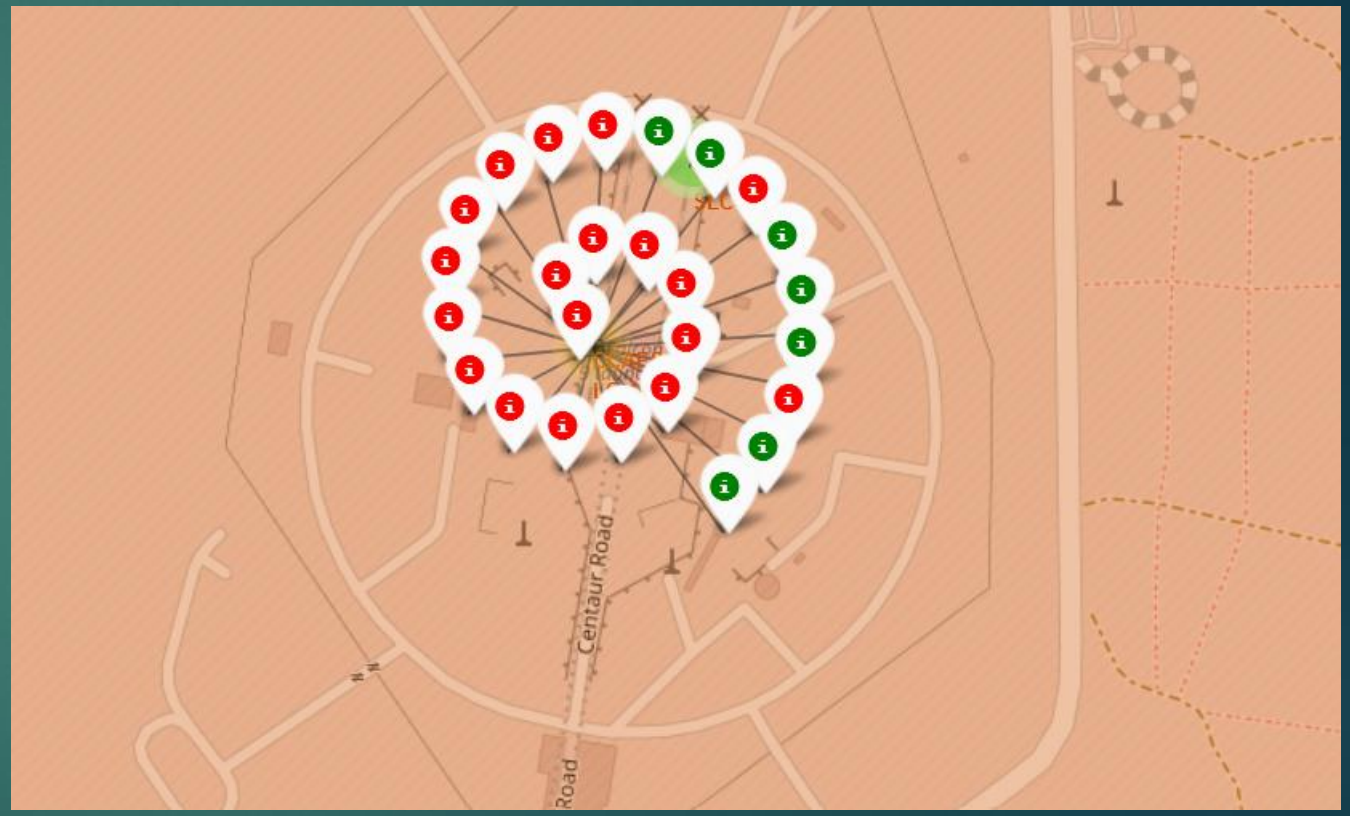
Launch sites are close to the coast to minimize damage to infrastructure in case of failure



Color Labeled Launch Records

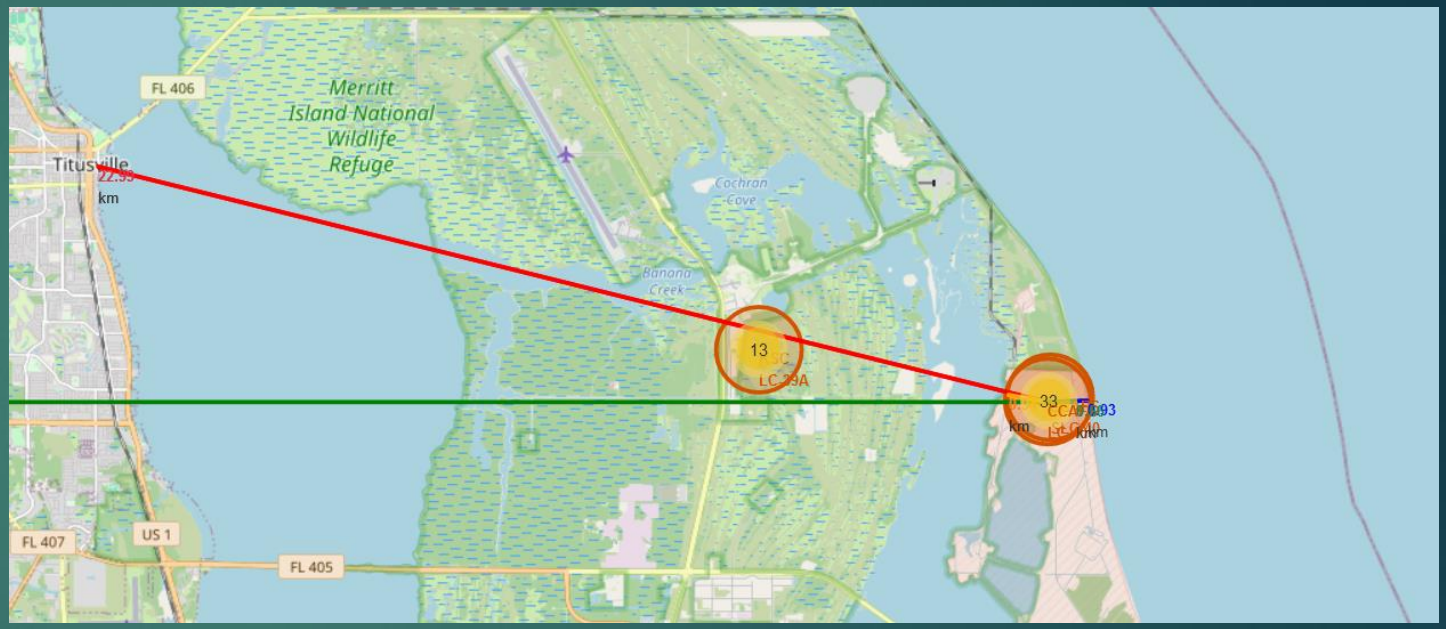
The color labels tell us if a launch was successful (green) or a failure (red).

This allows us to observe how successful a launch site is.



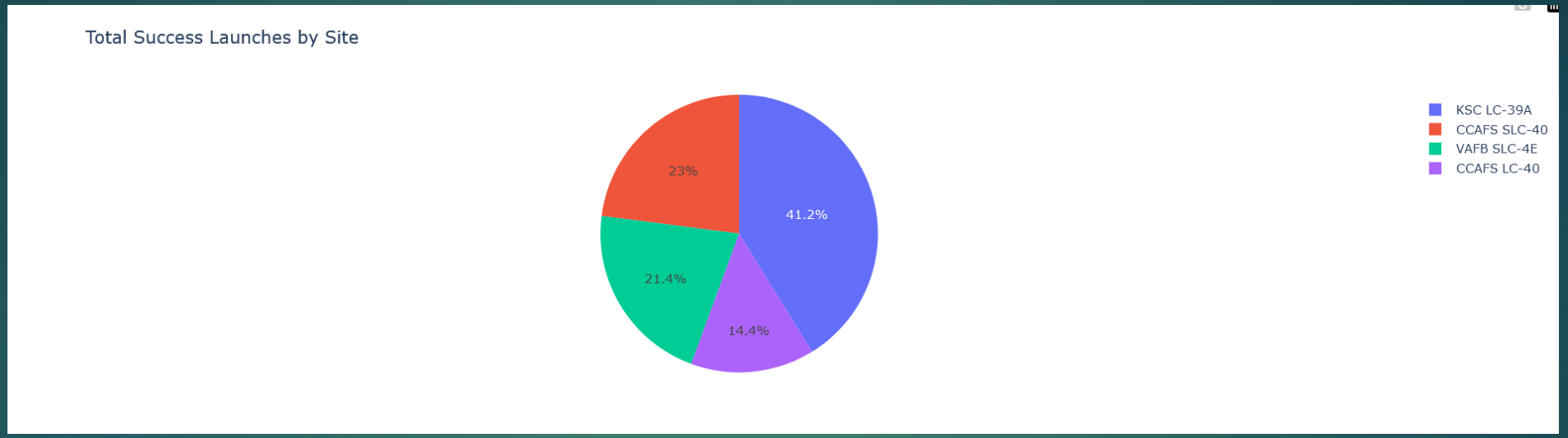
Distance from Launch Site CCAFS LC 40 to Proximities

Launch site CCAFS LC 40 is within one km from the coastline, 23km from the nearest city and less than 1km away from the nearest railroad



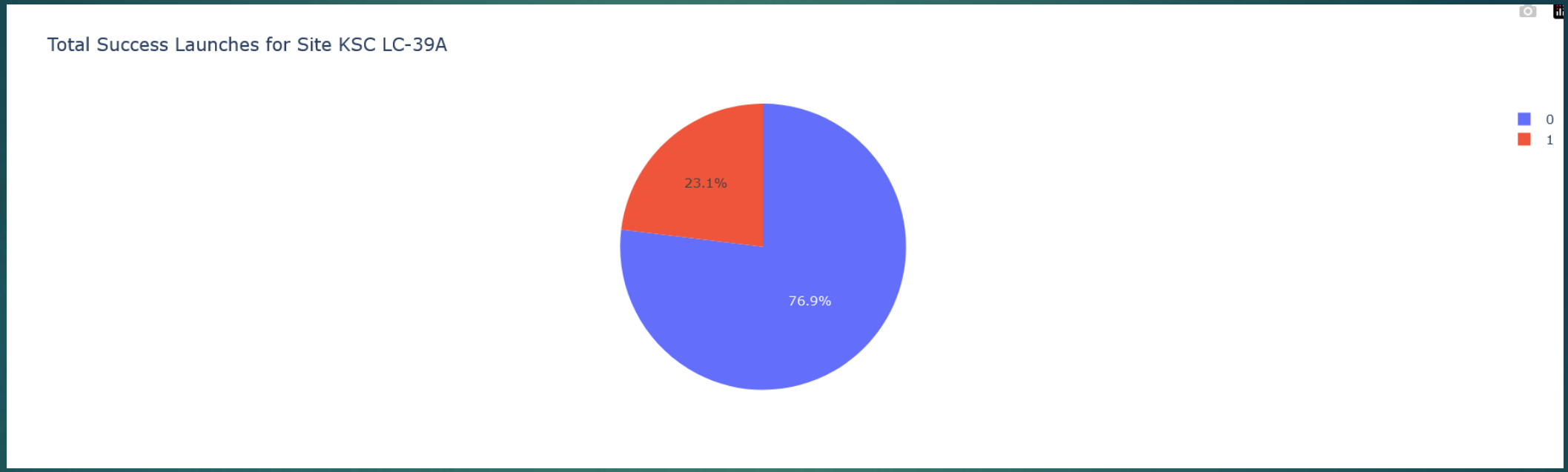
Section 4: Build a Dashboard with Plotly Dash

Total Successful Launches by Site



The pie chart shows that launch site KSC LC-39A has the most successful launches with 41.2%

Launch Site with the Highest Launch Success Rate



Launch site KSC LC-39A had the highest rate of success, with a 76.92% success rate

Payload vs. Launch Outcome Scatter Plot

The payload range with the highest success rate is between 2000 and 5000 with 15 successes and 13 failures

Booster version FT has the highest number of successful outcomes

Section 5: Predictive Analysis (Classification)

Classification Accuracy

Each method has the same Jaccard Score and F1 Score, but the Tree method has the lowest accuracy score. The other models have an accuracy score of 83.33%.

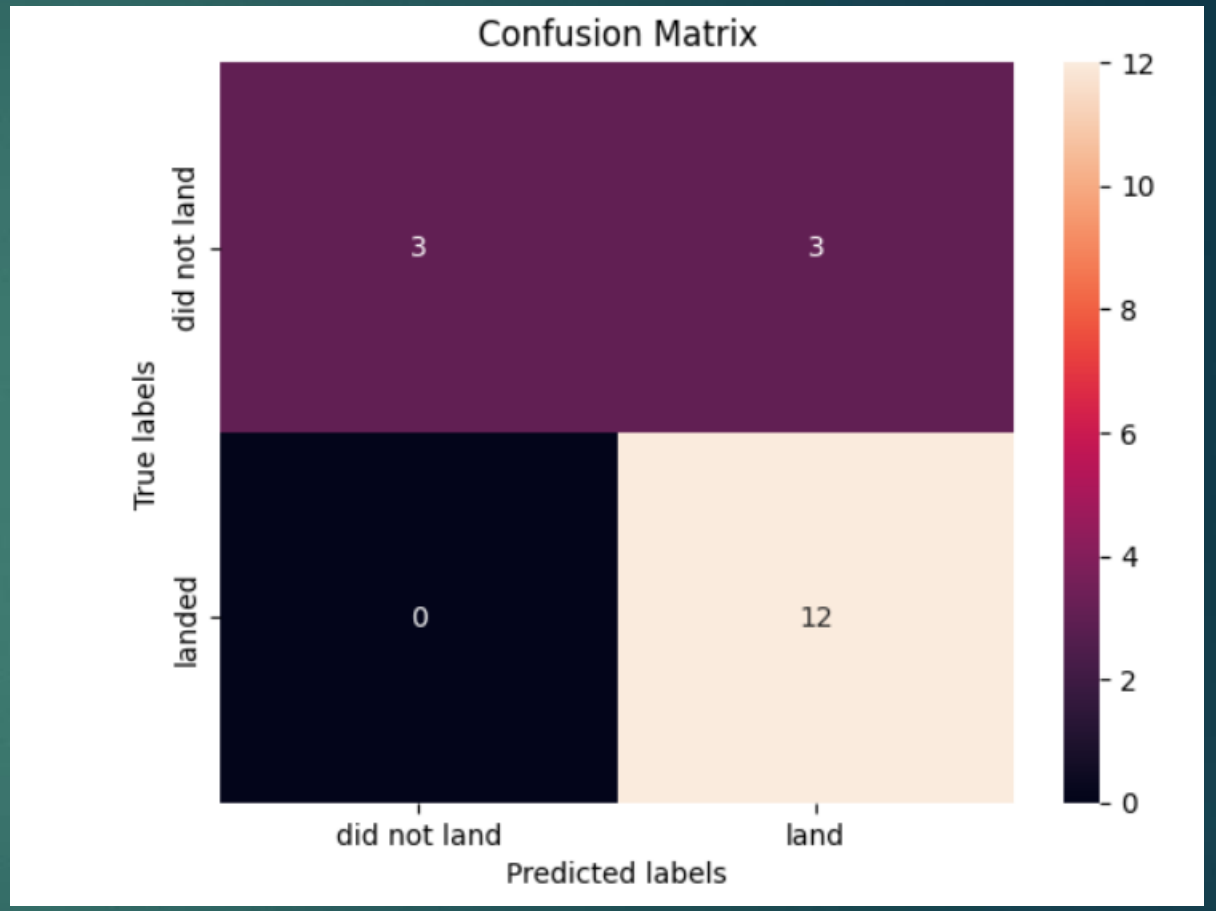
It is unclear which method has the best accuracy.

[50]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.611111	0.833333

Confusion Matrix

Judging from the confusion matrix, it seems that false positives give us the most trouble, with 3 instances.



Conclusion

- As years increase, launch success rates increase
- Most launches occur closer to the equator
- Orbit types ES-L1, GEO, HEO and SSO have 100% success rates
- The rate of success is higher than the rate of failure for all launch sites
- Launch site KSC LC-39A had the highest rate of success, with a 76.92% success rate