IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ștefan Epure

August 11, 2022

Table of Contents

Executive Summary

Methodology summary

- data collection
- data wrangling
- exploratory data analysis
- plotting maps with folium
- building a dashboard with slicers
- classification

Executive Summary

Results summary

- web scraping final dataset
- SQL findings
- visualization insights
- selecting the best hyperparameter

Introduction

Background and context

- Falcon 9 rockets launches cost 62 million USD
- competitors' upper cost is 165 million USD
- savings are a result of first stage reusability
- our company is looking to predict first stage successful landings
- this will be useful for bidding against SpaceX rocket launches

Introduction

Problems to answer

- How do we collect the necessary data?
- What are the variables which impact successful landings?
- What is the best classifier for predicting successful landings?

Methodology

Methodology

Executive Summary

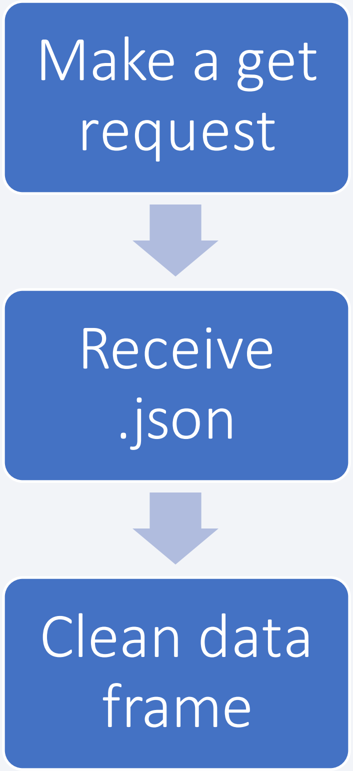
- Data collection methodology
  - SpaceX REST API, web scraping with BeautifulSoup
- Perform data wrangling
  - replace missing values with the mean
  - create an outcome label, One-Hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build, tune and evaluate classification models

Data Collection

- make a get request to the SpaceX API
- clean the requested data
- extract a Falcon 9 launch records HTML table from Wikipedia with BeautifulSoup
- parse the table and convert it to a Pandas data frame

Data Collection – SpaceX API

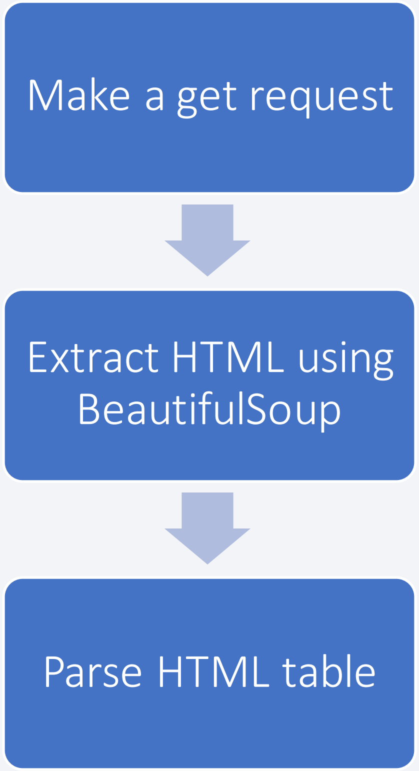
```python
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)

# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())

launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```



More on Github

Data Collection – Scraping

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url).text

# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, 'html5lib')

# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables=soup.find_all("table")

launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```



More on [Github]()

Data Wrangling

- Perform EDA on the final dataset
  - Handle missing values
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label
  - Determine the success rate

More on [Github]()

EDA with Data Visualization

- Plot graphs for the following
  - the relationship between Flight Number and Launch Site
  - the relationship between Payload and Launch Site
  - the relationship between success rate of each orbit type
  - the relationship between Flight Number and Orbit type
  - the relationship between Payload and Orbit type
  - the launch success yearly trend

More on Github

EDA with Data Visualization

- Required graphs for this task
  - scatter plots to see the relationship between variables
  - bar charts to compare changes between groups
  - line charts to see trends clearly

More on Github

EDA with SQL

- Perform queries using SQL to
  - 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

More on [Github]()

EDA with SQL

- Perform more queries with SQL to
  - List the names of the boosters which have success in drone ship and have 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 with a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in 2015
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

More on [Github]()

Build an Interactive Map with Folium

- Include the following
  - markers for all launch sites
  - markers for all launches with different colors for the outcome
  - distances lines between a launch site to its proximities

More on Github

Build a Dashboard with Plotly Dash

- Use pie charts to
  - display the total launches segmented by sites
- and scatter plots to
  - see the relationship between the outcome and payload mass for any booster version

More on Github

Predictive Analysis (Classification)

- Steps for model building
  - load dataset, transform data, determine training labels
  - perform a test – train split, standardize the data
  - find the best hyperparameter for logistic regression, SVM, decision trees and KNN
  - select the classification model with the best accuracy
  - fit the train data, make predictions with the test data

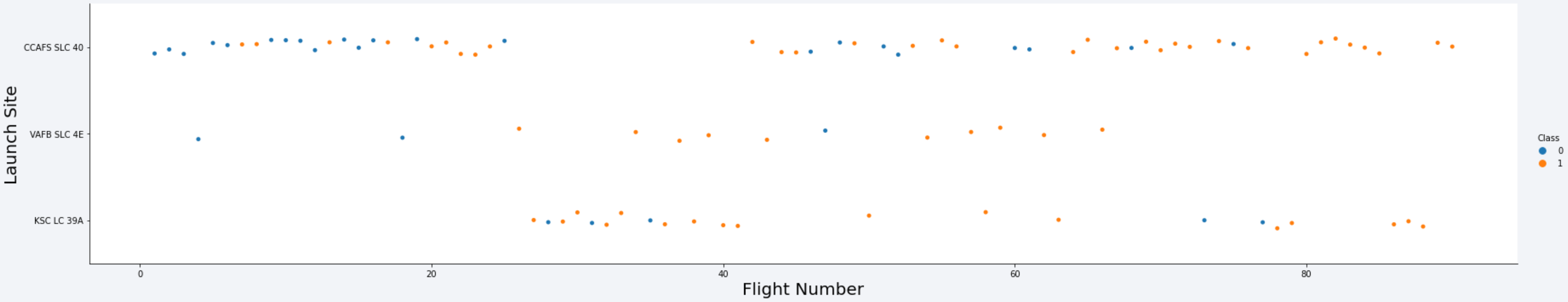
More on Github

Results

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

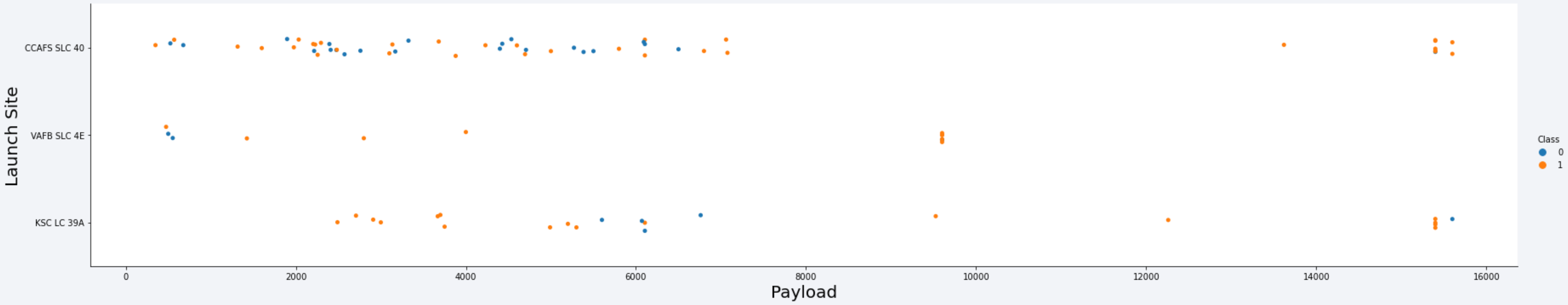
Insights drawn from EDA

Flight Number vs. Launch Site



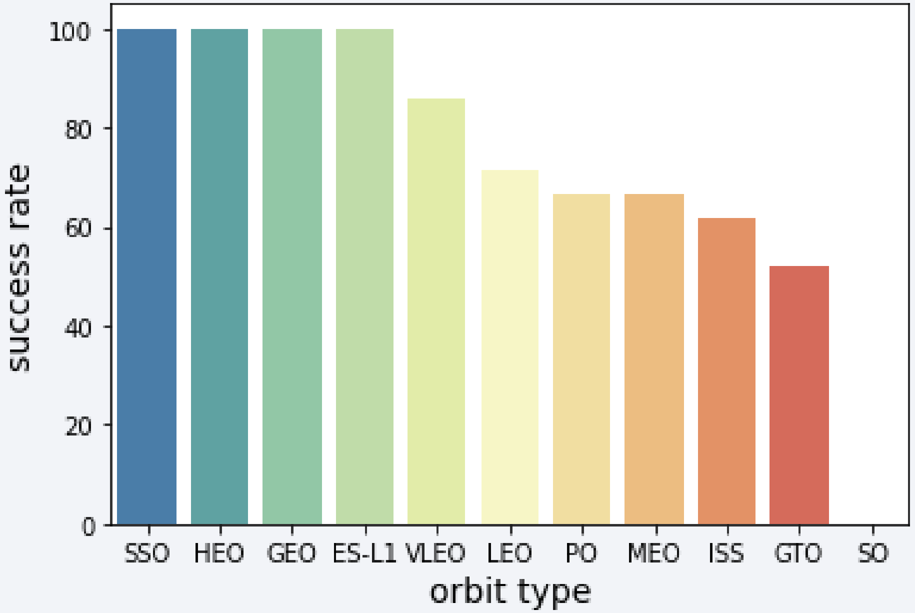
At a quick glance, it seems that more launches result in a higher success rate

Payload vs. Launch Site



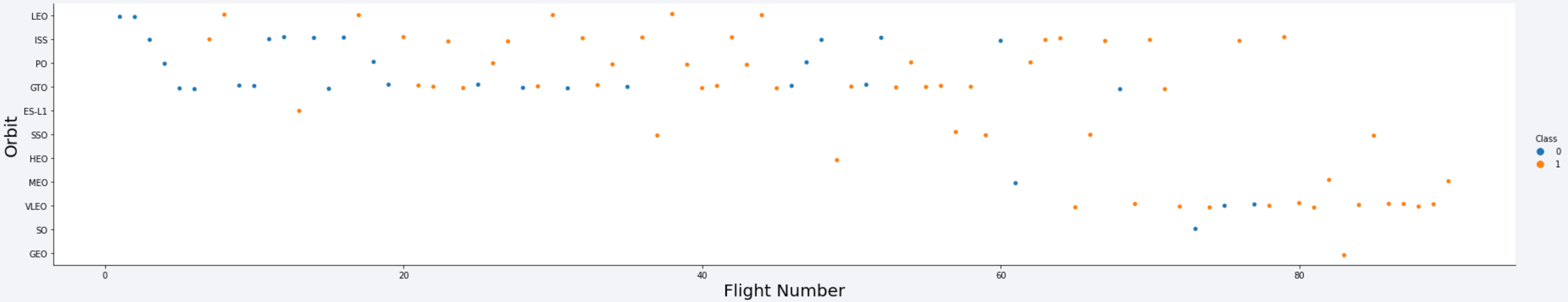
Here we can see that for the VAFB-SLC launch site there are no launches for heavy payload mass, greater than10 000 kg.

Success Rate vs. Orbit Type



The first four orbits from the left have the highest success rate.

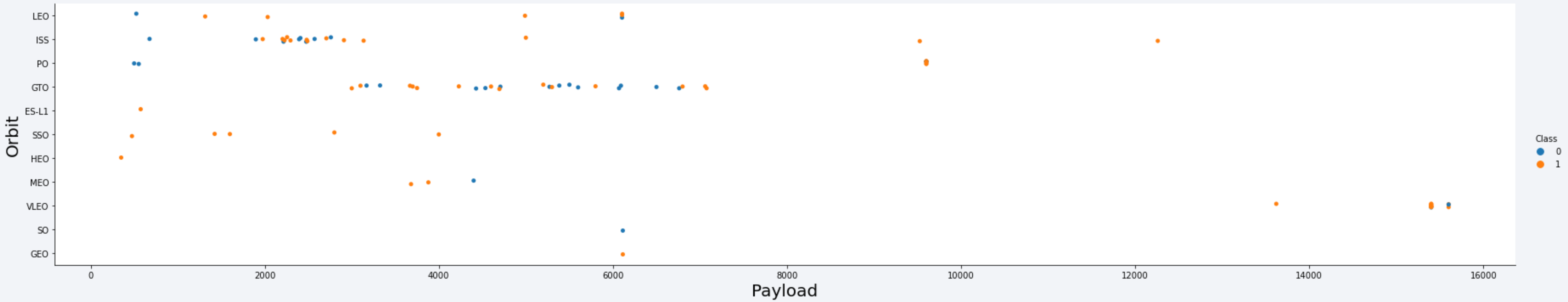
Flight Number vs. Orbit Type



In the LEO orbit, the success seems to be related to the number of flights.

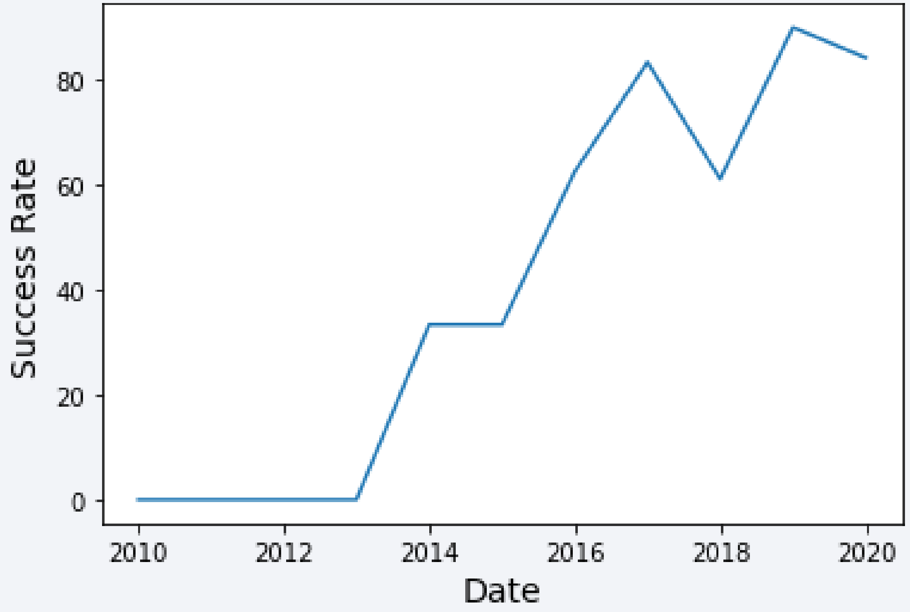
On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads, Polar, LEO and ISS have the higher successful landing rates.

Launch Success Yearly Trend



We can observe that the sucess rate, since 2013, kept increasing until 2020.

It also seems like the rate's growth has stagnated recently.

All Launch Site Names

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

We use the DISTINCT keyword to get the unique values.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Custom
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of	0	LEO (ISS)	NASA (COTS) NRO

We make use of the Where clause and limit the results to just 5.

Total Payload Mass

total_payload_mass
48213

We make use of the SUM() function.

Average Payload Mass by F9 v1.1

average_payload_mass
340.4

Here we use the AVG() function.

First Successful Ground Landing Date

Date	Time (UTC)	Payload
22-12-2015	01:29:00	OG2 Mission 2 11 Orbcomm-OG2 satellites

The dates are already ordered, so we use the Where clause to filter out unsuccessful launches and limit the results to the first row only.

Successful Drone Ship Landing with Payload between 4 000 and 6 000 kg

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Here we use two conditions in the Where clause.

Total Number of Successful and Failure Mission Outcomes

Success	Failure
100	1

We include two subqueries in the Select clause.

Boosters Carried Maximum Payload

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

We use a subquery, in the Where clause, which finds the max value.

2015 Launch Records

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

We specify the year and outcome in the Where clause.

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

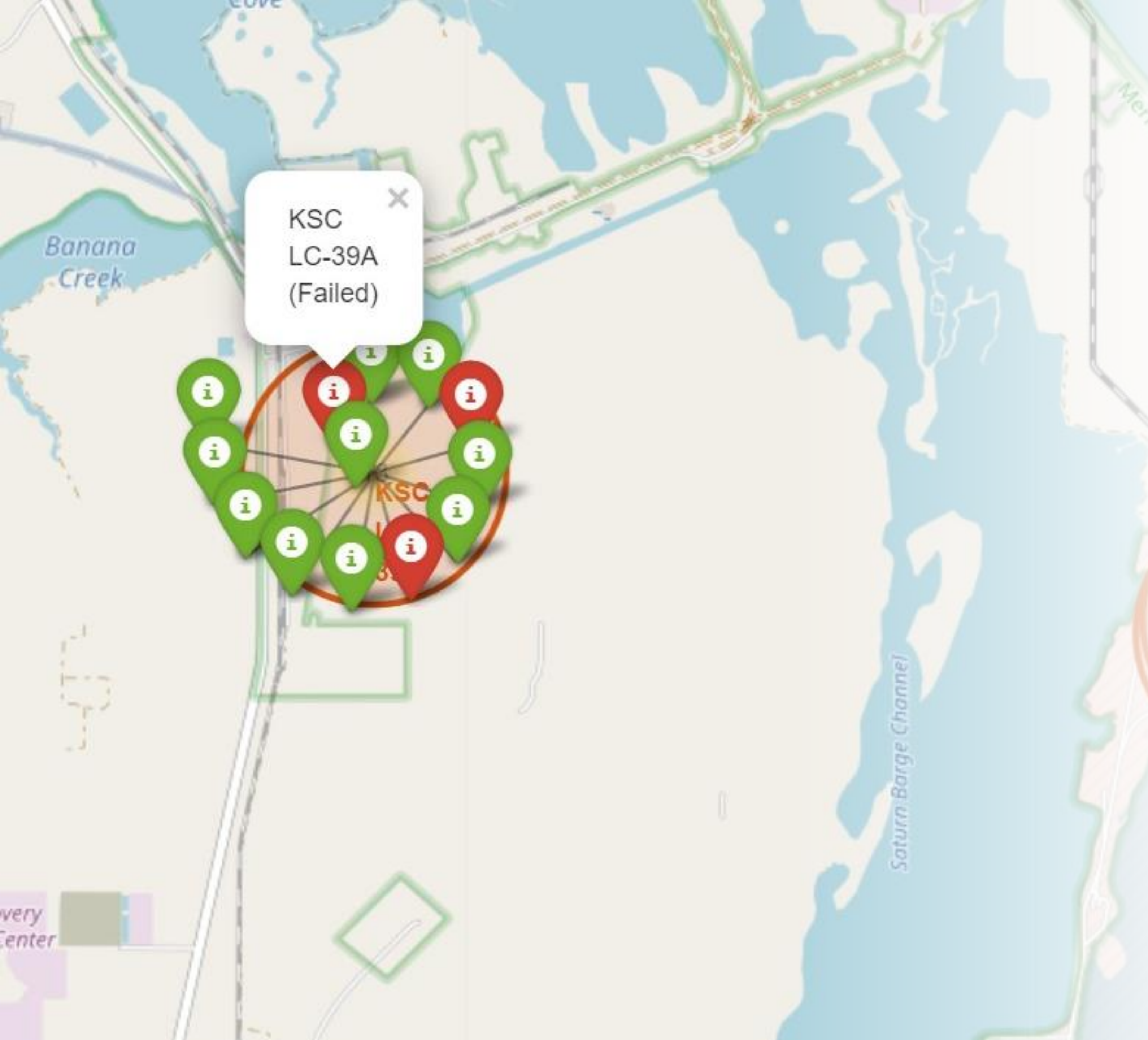
Landing _Outcome	successful_landing_outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

We make use of the COUNT() function and then we rank the values by specific types of success.


Launch Sites
Proximities Analysis

Launch Site Locations

The launch sites used by SpaceX are located in the USA, more specifically on the coasts of Florida and California.



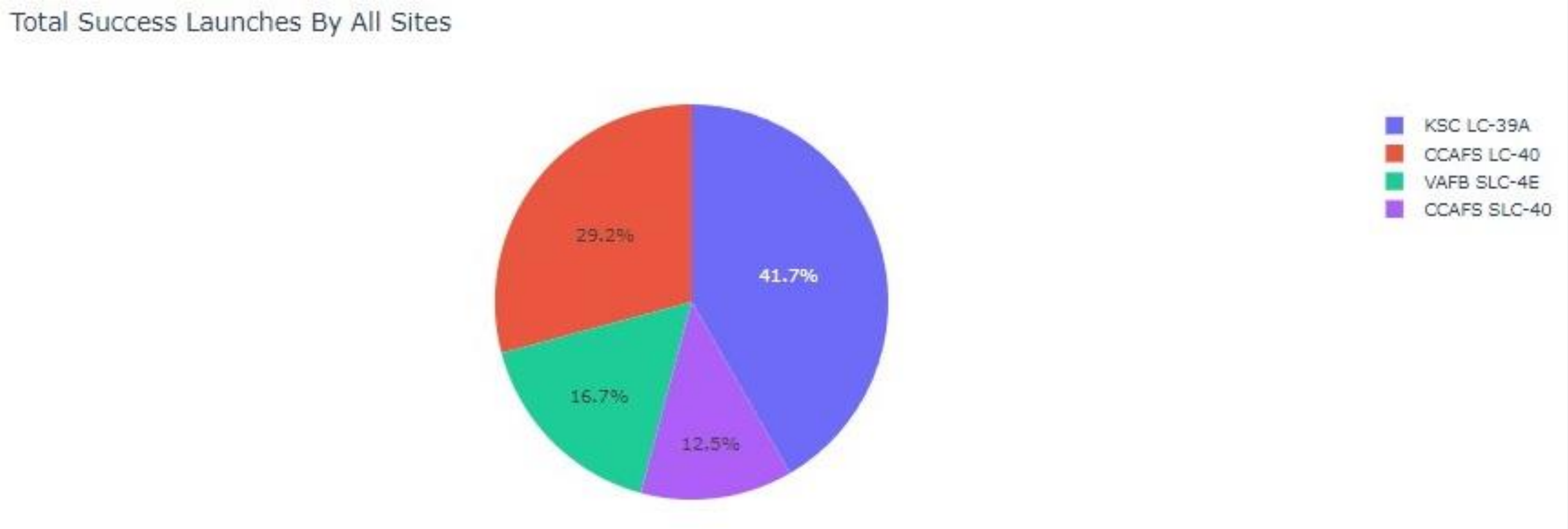
Launch Locations with Colors

- The green markers on this map indicate successful launches, the contrary is true for red markers.

Launch Site Distances

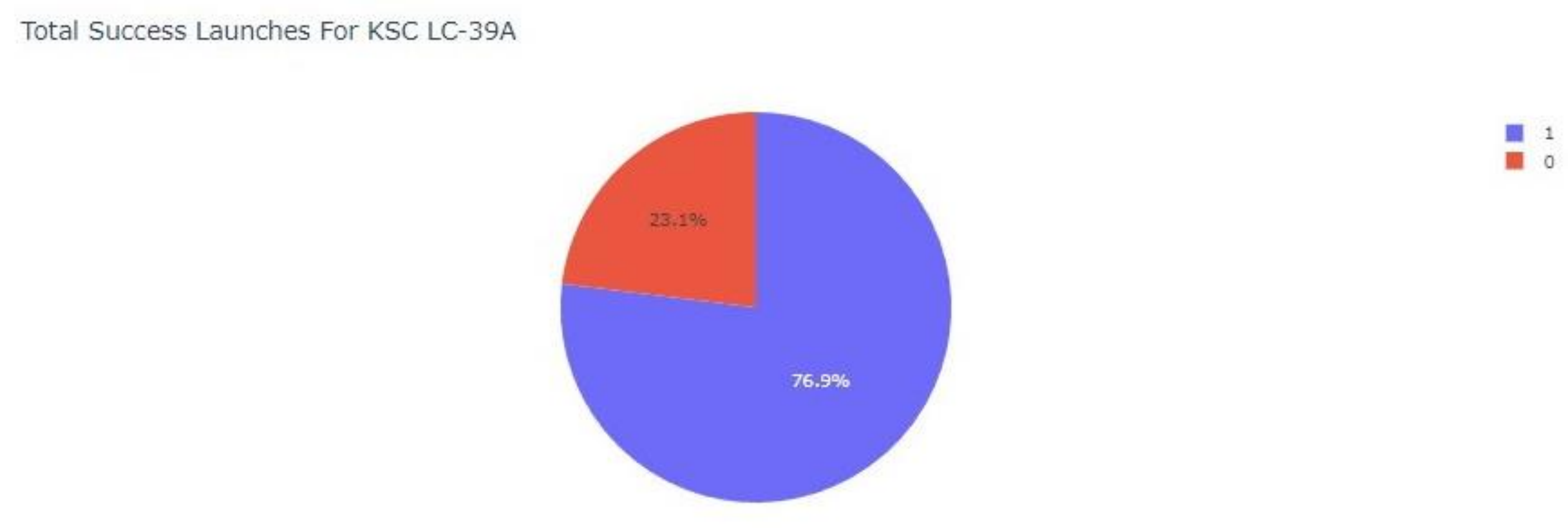
Build a Dashboard with Plotly Dash

Successful Launches per Site



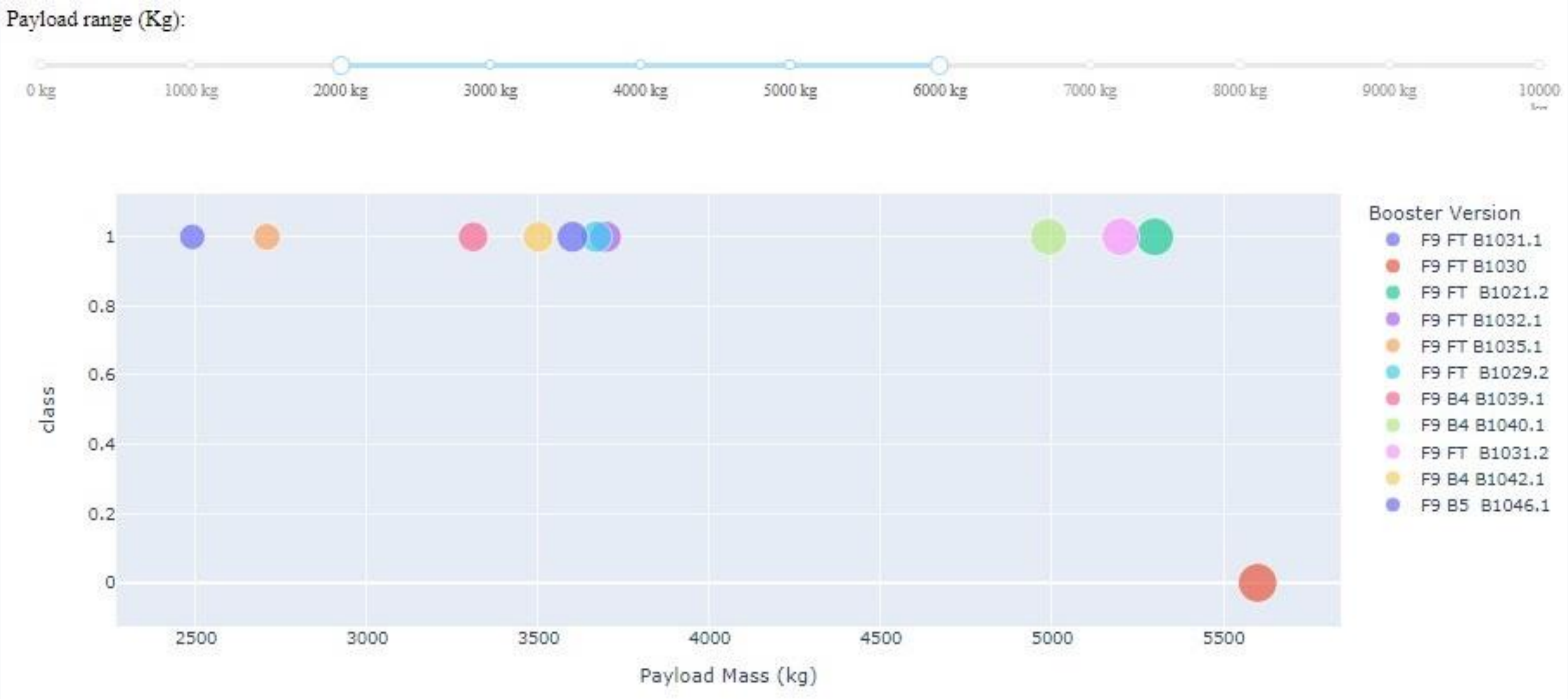
KSC LC-39A has the most successful rates out of all sites.

The Best Performing Site



KSC LC-39A has a 76.9% success rate.

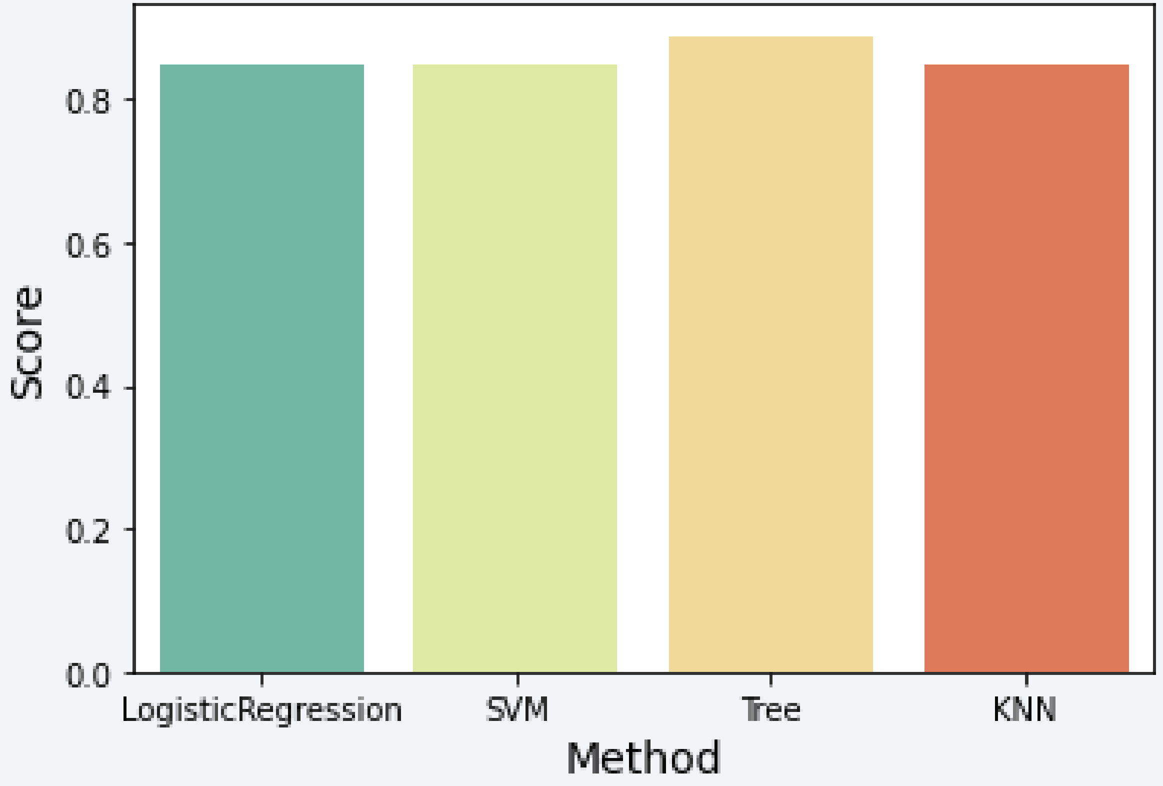
Plotting Outcomes by Payload Mass



We notice that successful launches tend to group up at the lower end of the payload range.

Predictive Analysis (Classification)

Classification Accuracy



Using the best hyperparameter for each method, we can see that they are all very accurate, but the decision tree classifier is the best choice.

Confusion Matrix

Examining the confusion matrix, we see that the decision tree can distinguish between the different classes. However, we also notice that the major problem is false positives.

Conclusions

- The following orbits have the best success rate: SSO, HEO, GEO, ES-L1
- Launch site KSC LC-39A has the most successful launches
- Lower payload weights tend to be more successful
- As time goes by, SpaceX is constantly perfecting the launch process, therefore the success rate will keep growing
- All classification methods are similar in terms of accuracy and, aside from the false positive issue, the decision tree classifier is the best choice for making further predictions

Thank you!