IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aline Tanja Polak
8th of February 2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection (API and Web Scraping)
- Data wrangling
- Exploratory data analysis with SQL
- Exploratory data analysis with Data Visualization
- Interactive visual analytics with Folium
- Interactive visual analytics with Plotly Dash
- Machine Learning predictive analysis

Summary of results

- Exploratory data analysis results
- Interactive analytics results
- Predictive analysis results

Introduction

Project background and context

Space X is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage. Therefore, by determining if the first stage will land, we can estimate the cost of a launch. This information can be used for an alternate company, Space Y, that wants to bid against Space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Questions to be answered

- What variables affect the success of the first stage landing? How do their interactions influence the success rate?
- What is the tendency of the rate of successful landings throughout the years?
- What is the best algorithm that can be used for binary classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Applying One-hot encoding to prepare data 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
  - Building, tuning and evaluation of classification models to ensure the best results.

Data Collection

The data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in Space X’s Wikipedia entry.

Obtained data columns through SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Obtained data colums through Wikipedia Web Scraping:

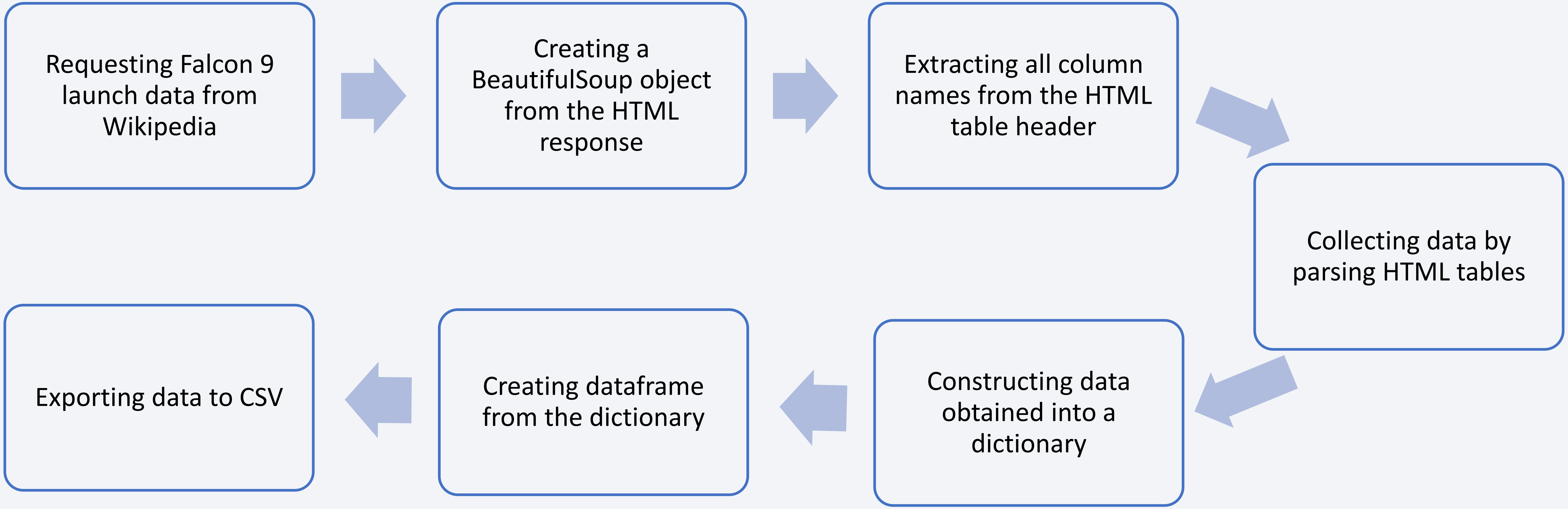
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection - SpaceX API

- Requesting rocket launch data from Space X API
- Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
- Requesting needed information about launches with custom functions
- Constructing obtained data into a dictionary
- Creating data frame from the dictionary
- Filtering dataframe to only include Falcon 9 launches
- Replacing missing values with calculated .mean() (Payload Mass column)
- Exporting data to CSV

[GitHub URL: Data Collection API]

Data Collection – Web Scraping



[GitHub URL: Data Collection with Web Scraping]

Data Wrangling

- Calculate number of launches on each site
- Calculate number of occurrence of each orbit
- Calculate number and occurrence of mission outcome per orbit type
- Create landing outcome label from Outcome column
- Exporting data to CSV

GitHub URL: Data Wrangling

EDA with Data Visualization

Charts were plotted:

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 and Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

[GitHub URL: EDA Data Visualization]

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 which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which 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 year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

[GitHub URL: EDA with SQL]()

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured lines to show distances between the Launch Site KSC LC-39A.

GitHub URL: Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL: Plotly Dash App

Predictive Analysis (Classification)

1. Creating a Numpy array from the column “Class”
2. Standardizing the data with StandardScaler, fitting and transforming
3. Splitting data into training and testing sets with train_test_split function
4. Creating a GridSearchCV object with cv=10 to find best parameters
5. Applying GridSearchCV on LogReg, SVM, Decision Tree and KNN models
6. Calculating accuracy on test data using method.score() for all models
7. Examining confusión matrix for all models
8. Finding best performing method by examining Jaccard_score and F1_score metrics

[GitHub URL: ML Prediction]

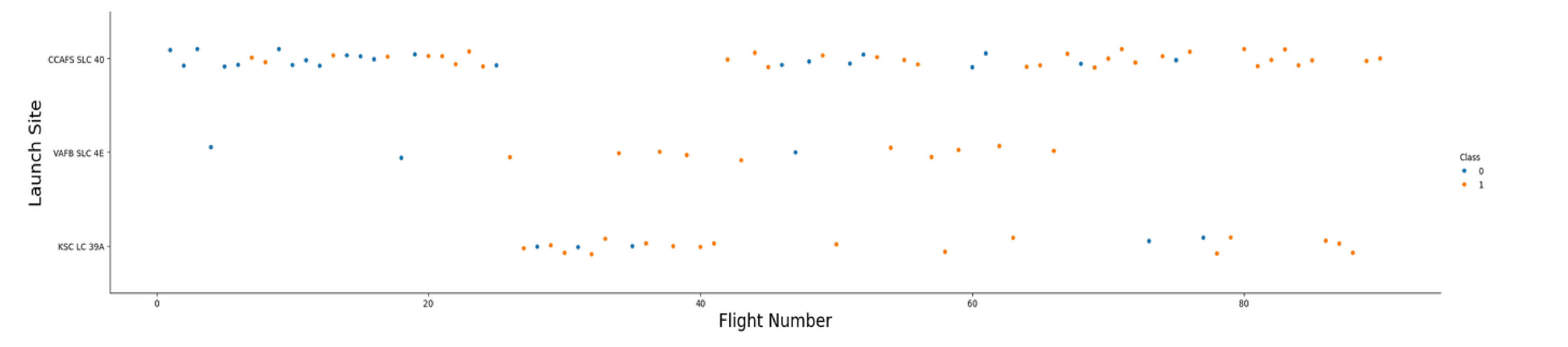
Results

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

Section 2

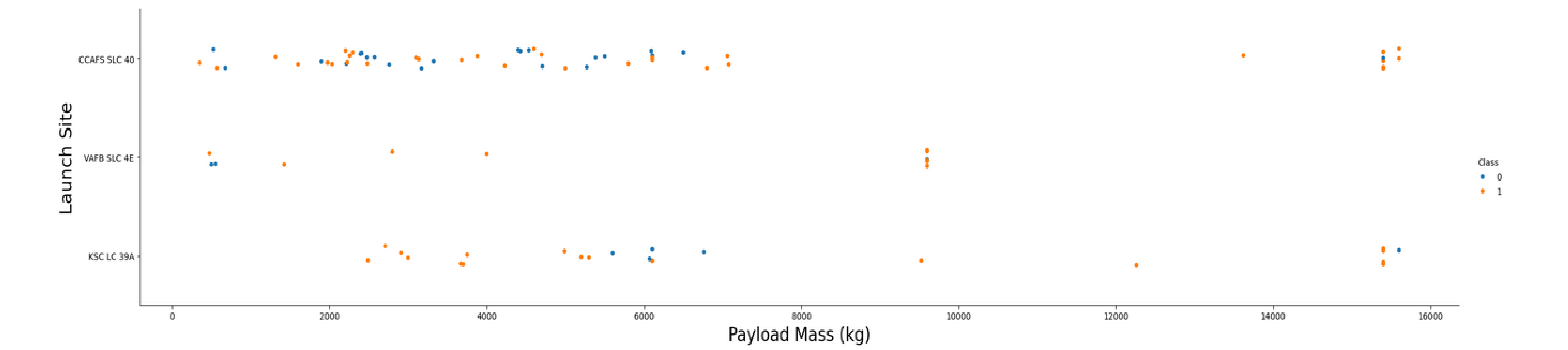
Insights drawn from EDA

Flight Number vs. Launch Site



- Earlier flights accumulated most failures, whereas later flights accumulated more successful ones.
- Therefore, the higher the flight number at a launch site, the greater the success rate.

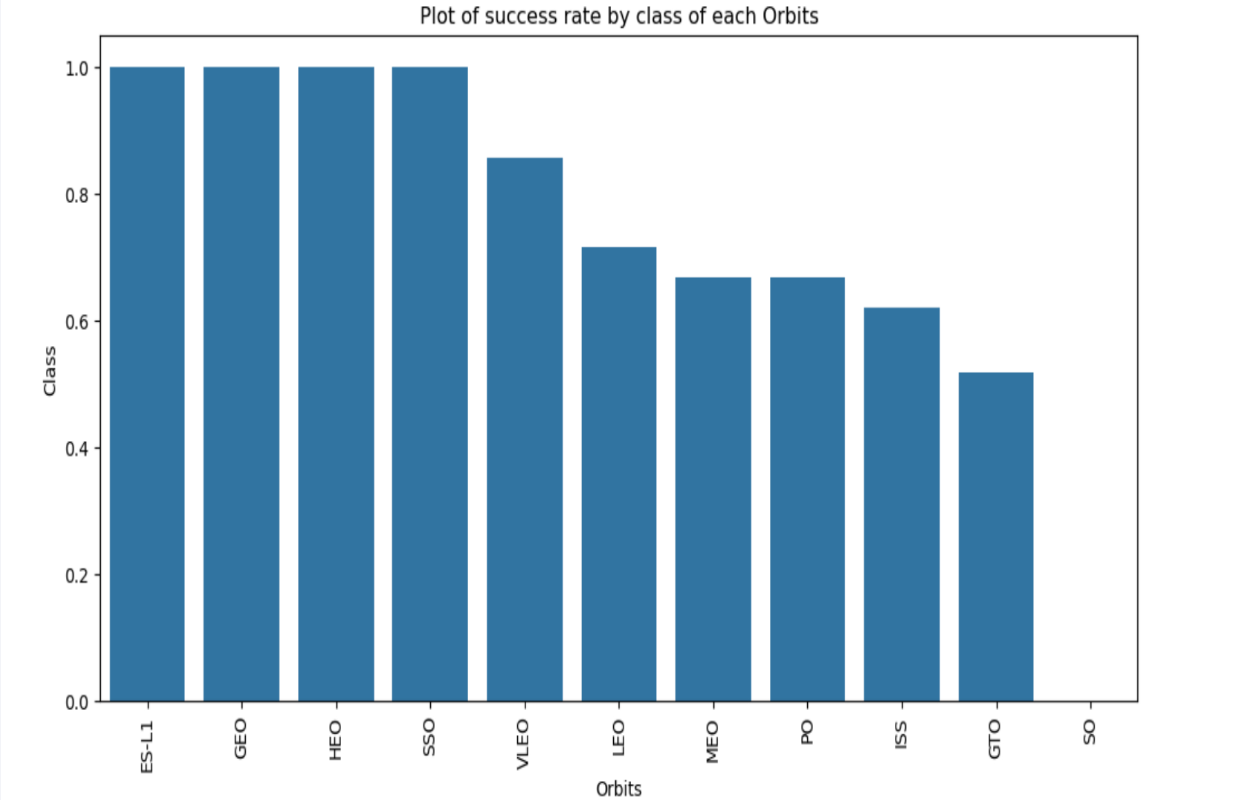
Payload vs. Launch Site



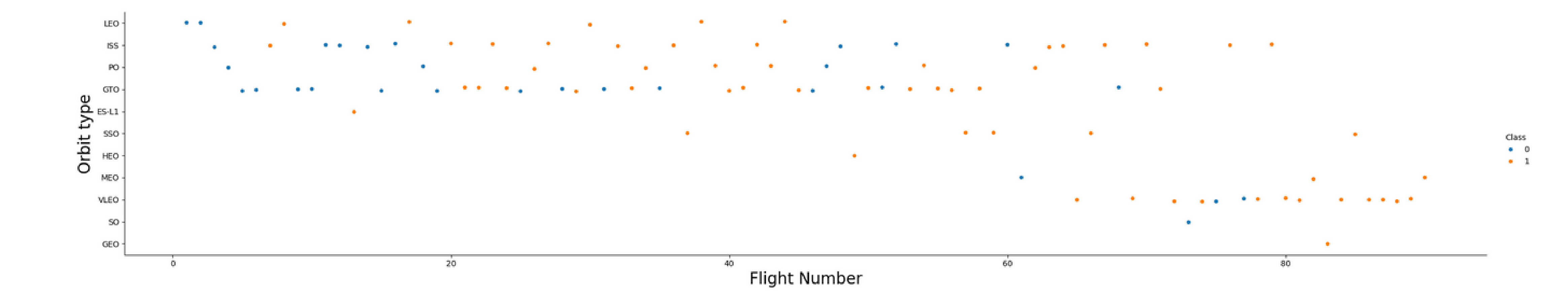
- Most of the launches with payload mass over 7000 kg were successful for every launch site.
- KSC LC 39A accumulated most successful flights under 5000 kg of payload mass.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO had the most successful flights (100% success rate)
- Orbit SO had the least successful flights (0% success rate)

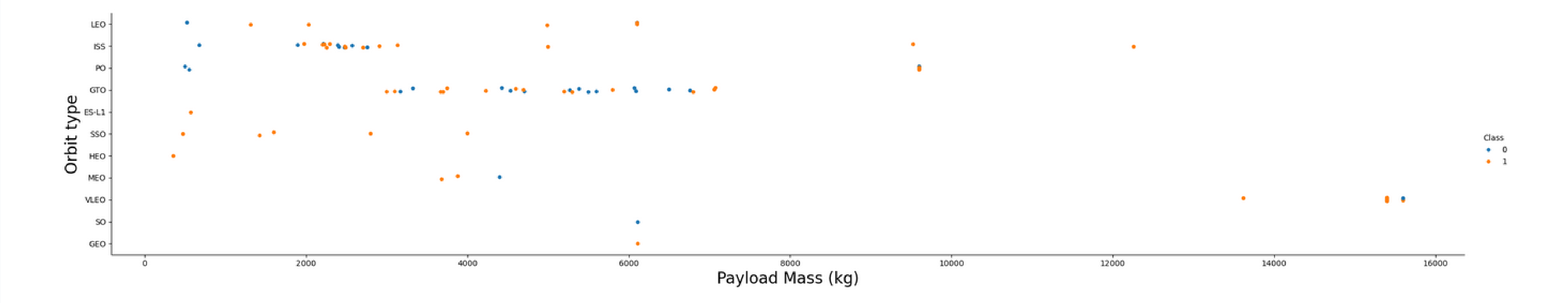


Flight Number vs. Orbit Type



In the LEO orbit, the success rate is associated with the number of flights whereas the plot indicates no relationship with flight number in the GTO orbit.

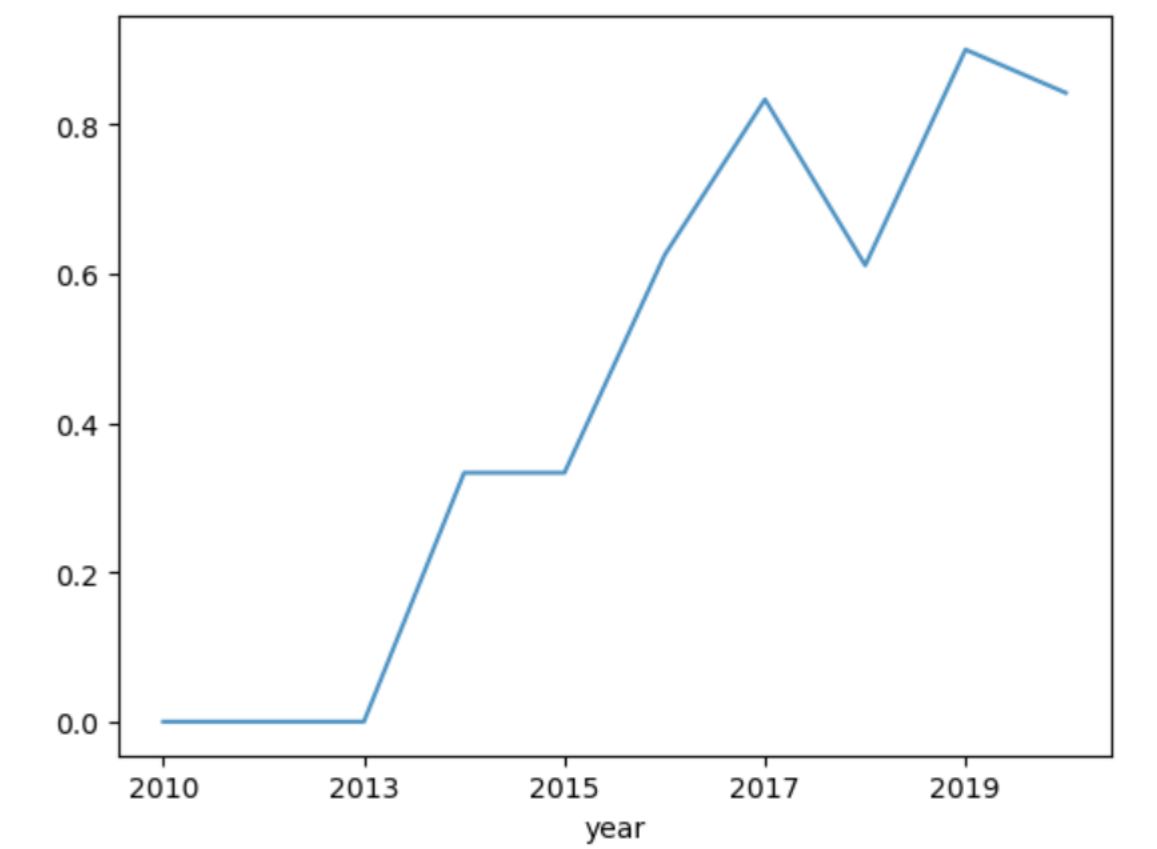
Payload vs. Orbit Type



The orbits PO, LEO and ISS accumulate more successful landings with heavy payloads.

Launch Success Yearly Trend

The success rate since 2013 started increasing, stabilizing in 2014, until 2017 and in 2018 after a slight decline started increasing again.



All Launch Site Names

Displaying the different names of all launch sites.

Display the names of the unique launch sites in the space mission

In [27]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

Out[27]:

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

Launch Site Names Begin with 'CCA'

Displaying five records of launch sites beginning with the string ‘CCA’.

In [31]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Out[31]:

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

Total Payload Mass

```
In [67]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'

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

Out[67]:

SUM(PAYLOAD_MASS__KG_)
45596

Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

In [75]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

Out[75]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Displaying the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
[39]: %sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

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

[39]:

MIN(DATE)
2015-12-22

Displaying the date the first successful landing outcome was achieved in ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[41]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

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

[41]:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Displaying the booster names with a successful drone ship landing and a payload mass between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

```
[95]: %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE

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

[95]:

COUNT(Mission_Outcome)
101

Displaying the total number of successful and failed mission outcomes.

Boosters Carried Maximum Payload

Displaying boster names that carried máximum payload mass.

[117]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

[117]:

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

```
[47]: %%sql SELECT SUBSTR(Date, 6, 2) AS Month, SUBSTR(Date, 0, 5) AS Year, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure%' AND Year = '2015'

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

[47]:

Month	Year	Booster_Version	Launch_Site
01	2015	F9 v1.1 B1012	CCAFS LC-40
04	2015	F9 v1.1 B1015	CCAFS LC-40

Displaying failed landing outcomes in drone ships filtering by months in the year 2015, booster versions, and launch sites.

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

```
[51]: %%sql SELECT landing_outcome, COUNT(Landing_Outcome) AS Total_Landing FROM SPACEXTABLE
      WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Total_Landing DESC

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

[51]:

Landing_Outcome	Total_Landing
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

Displaying a rank of the number of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

Section 3

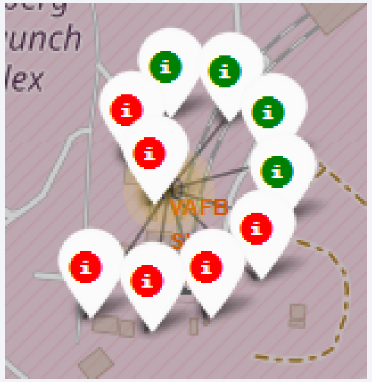
Launch Sites Proximities Analysis

Launch sites' locations on a global map

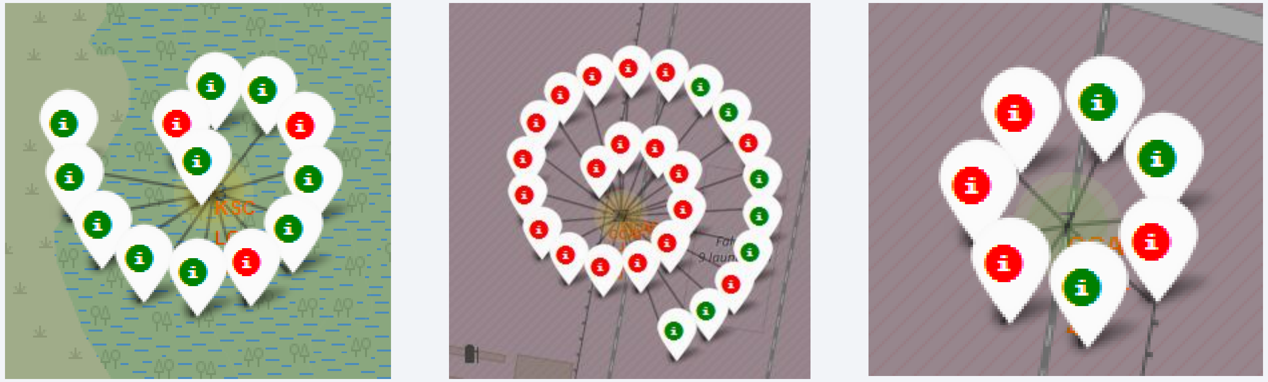
SpaceX launch sites are located on the United States of America coastlines, specifically in Florida and California.

Launch sites with outcomes (color-labeled)

California launch site



Florida launch sites

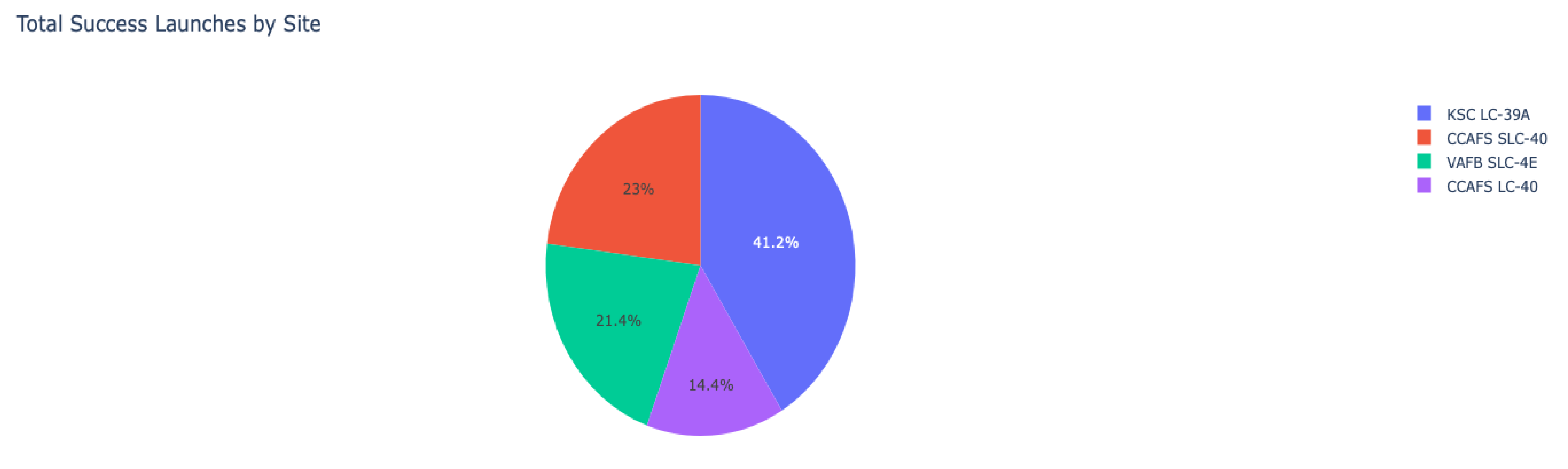


Green markers show successful launch outcomes, whereas red markers show failed outcomes.

Section 4

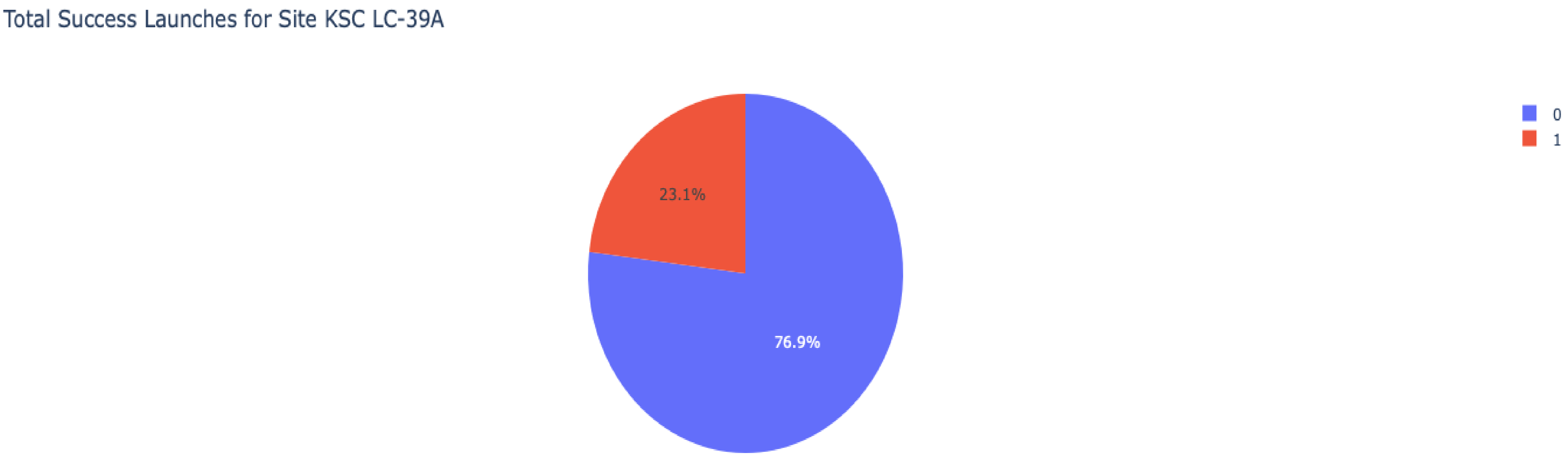
Build a Dashboard with Plotly Dash

Launch success percentage of each launch site



KSC LC-39A had the most successful launches of all launch sites.

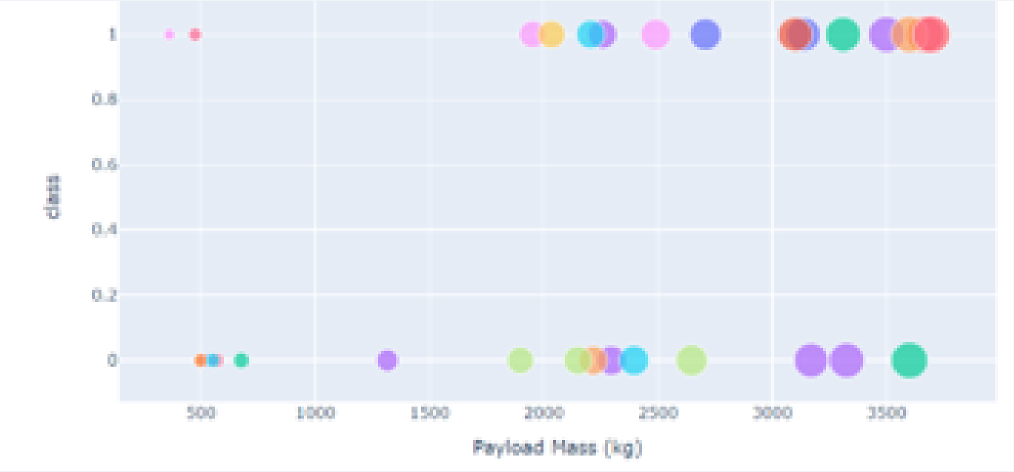
Launch site with the highest launch success rate



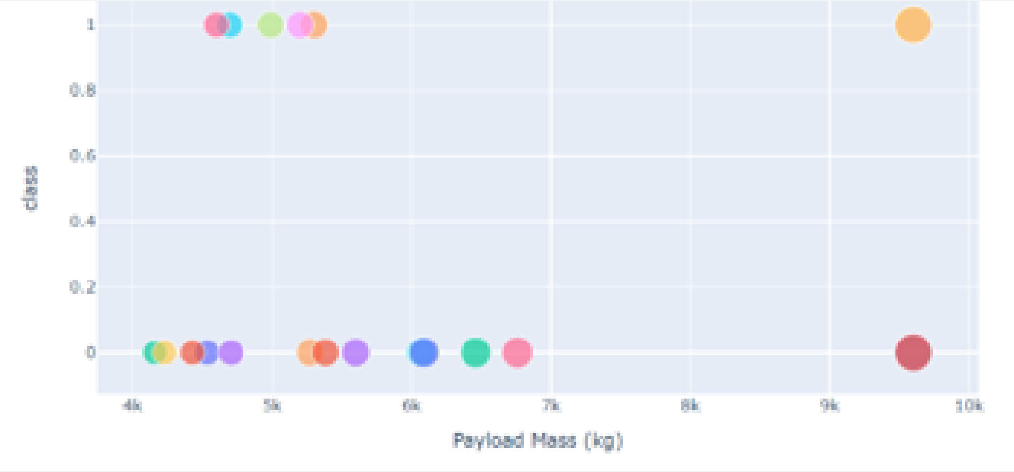
KSC LC-39A had the highest success rate (76,9%) with 10 successful launches and only 3 failed ones.

Payload mass vs. Launch outcome for all sites

Light-weighted payload mass (0-4000kg)



Heavy-weighted payload mass (4000-10.000kg)



In general, light-weighted payload mass have higher success outcomes than heavy-weighted ones. Specifically, payloads between 2000 and 5500 have the highest success rate.

Section 5

Predictive Analysis (Classification)

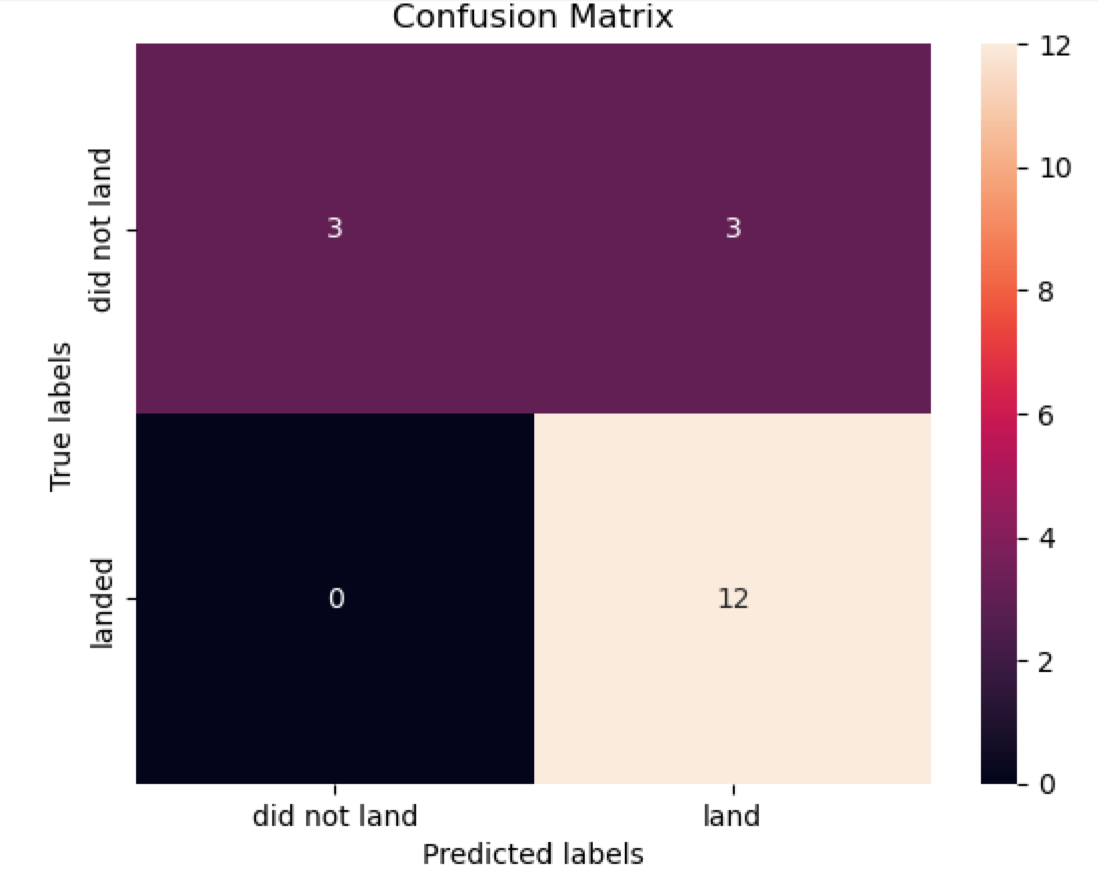
Classification Accuracy

```
[60]: models = {'KNeighbors':knn_cv.best_score_,
                'DecisionTree':tree_cv.best_score_,
                'LogisticRegression':logreg_cv.best_score_,
                'SupportVector': svm_cv.best_score_}

      bestalgorithm = max(models, key=models.get)
      print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
      if bestalgorithm == 'DecisionTree':
          print('Best params is :', tree_cv.best_params_)
      if bestalgorithm == 'KNeighbors':
          print('Best params is :', knn_cv.best_params_)
      if bestalgorithm == 'LogisticRegression':
          print('Best params is :', logreg_cv.best_params_)
      if bestalgorithm == 'SupportVector':
          print('Best params is :', svm_cv.best_params_)

      Best model is DecisionTree with a score of 0.8732142857142857
      Best params is : {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}
```

When taking into consideration the scores of the whole data set, the decision tree model performed best, with higher scores and the highest accuracy.

Confusion Matrix

The decisión tree classifier was able to distinguish between different classes, however, false positives seem to be a problem, which means unsuccessful landings are marked as successful landings.



Conclusions

- The larger the number of flights at a launch site, the greater its success rate.
- Launch success rate overall increased from 2013 until 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO have the highest success rates (100%).
- KSC LC-39A had the most successful launches of any sites.
- All launch sites are near the coastline.
- Light-weighted payload mass were associated with better success rates than heavier-weighted ones.
- The Decision tree model is the best machine learning algorithm for this dataset.

Thank you!