Winning Space Race With Data Science

Oksana Moskalyuk

31.03.2025

Outline

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

Executive Summary

This report provides an in-depth analysis of data collected from SpaceX launch data. The primary goal is to determine the successful landing rate of the first stage of a rocket launch.

To work with the lauch data, we collected the data using the SpaceX REST API and Web Scraping. After the Data Collection, we applied Data Wrangling to get a clean dataset. Our next step was to do an Exploratory Data Analysis, where we used SQL and Data Visualization for gaining insights on the important features of the dataset. We proceeded by using Interactive Visual Analytics to discover patterns. The final step was Predictive Analysis, where we built a machine learning pipeline to predict the success of landing.

SUMMARY OF ALL RESULTS

Introduction

SpaceX 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 SpaceX can reuse the first stage.
Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

HOW

do payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

DOES

the rate of successful landings increase over the years?

WHAT

is the best algorithm that can be used for binary classification in this case?

Methodology

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning prediction, using:
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

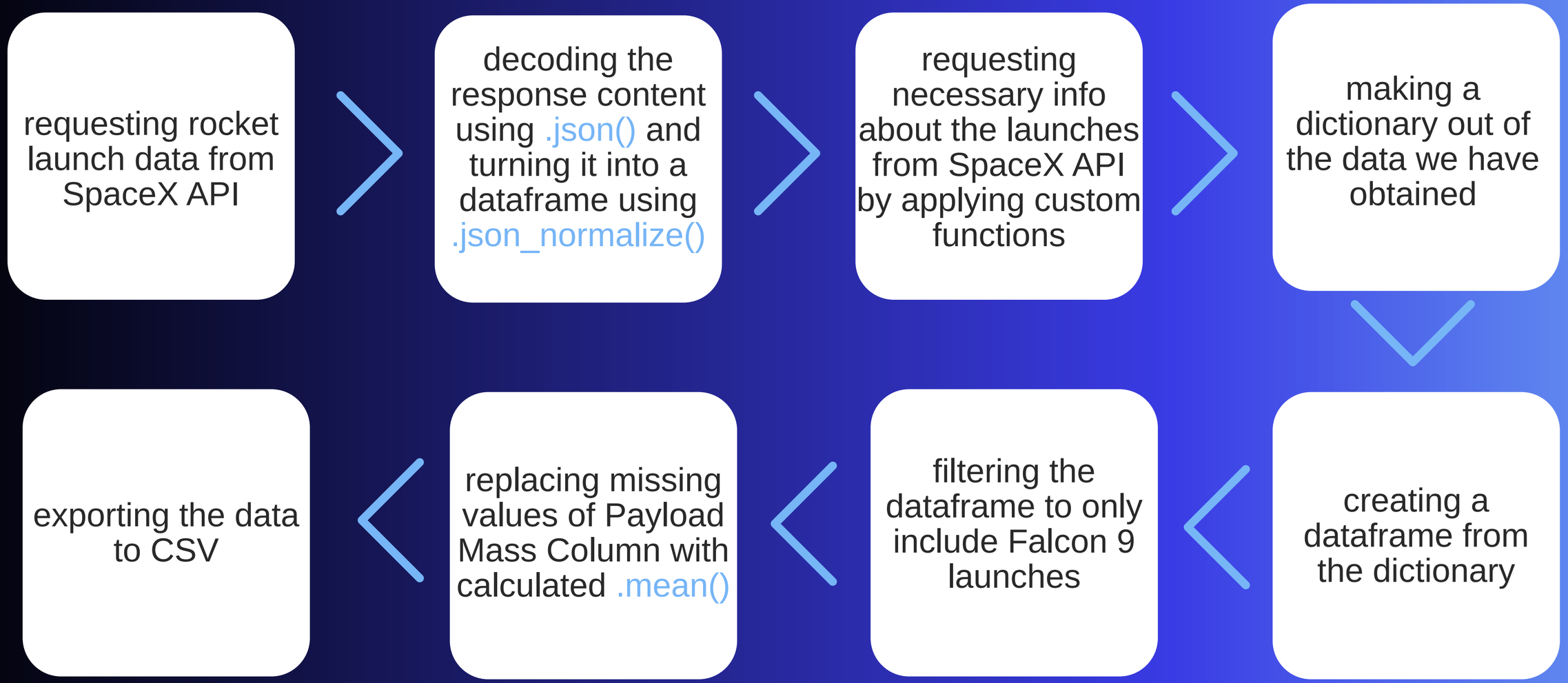
USING SPACEX REST API WE OBTAINED:

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

USING WIKIPEDIA WEB SCRAPING WE OBTAINED:

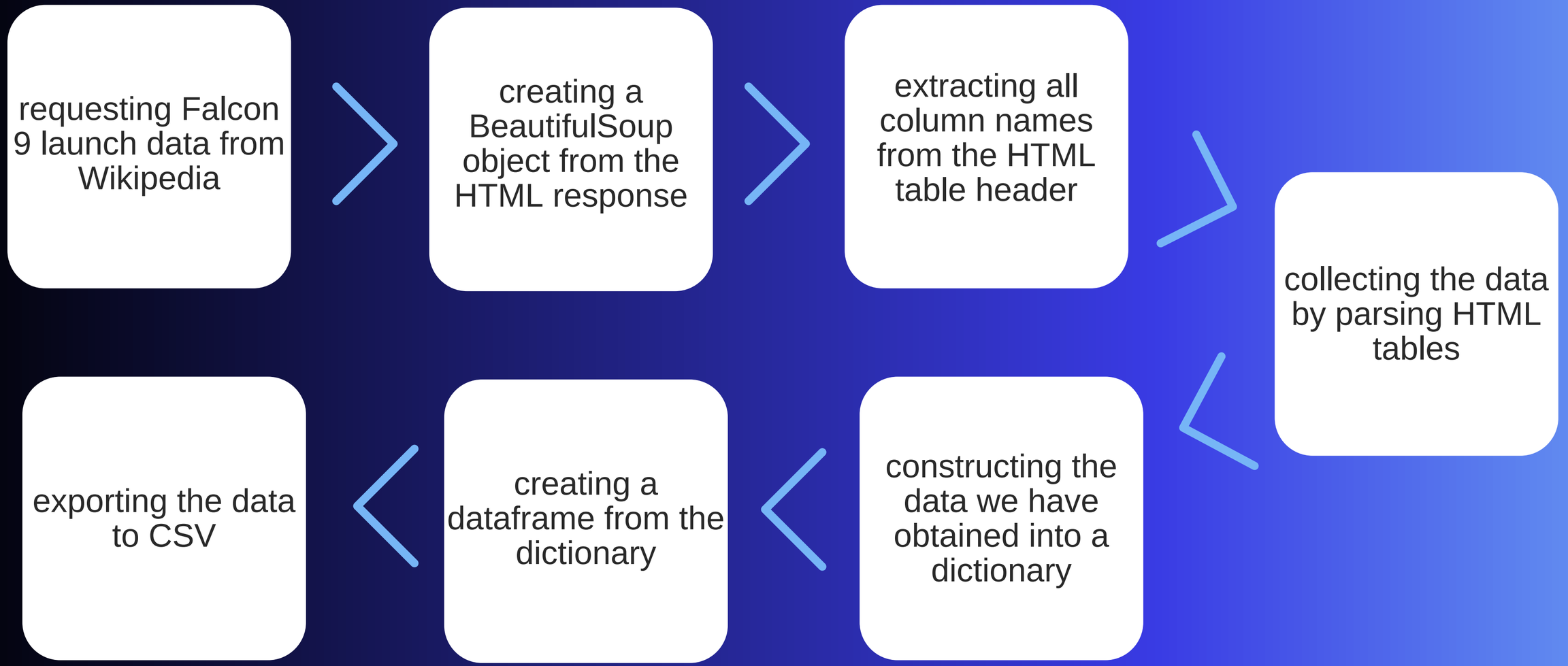
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection - SpaceX API



GitHub: Data Collection API

Data Collection - Web Scraping



GitHub: Data Collection with Web Scraping

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

performing Exploratory Data Analysis and determine Training Labels

calculating the number of lanches on each site

calculating the number and occurrence of each orbit

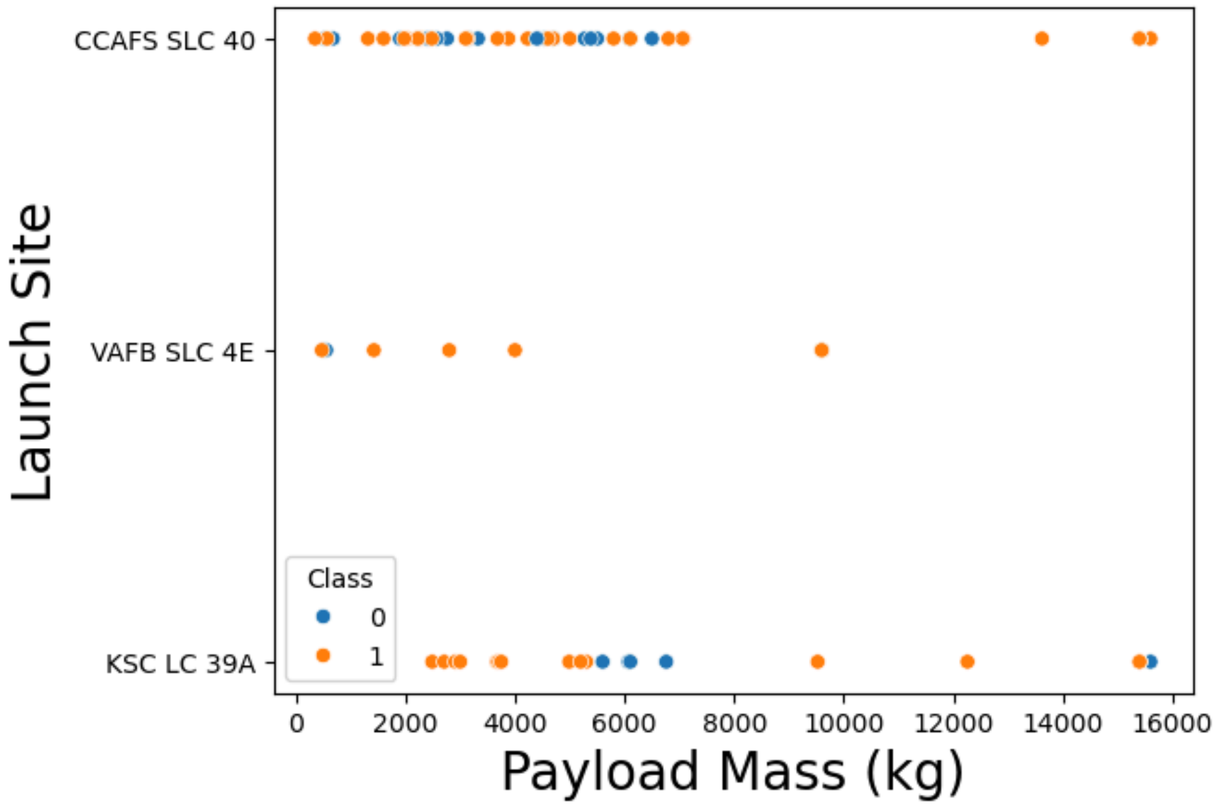
calculating the number and occurrence of mission outcome per orbit type

creating a landing outcome label from Outcome column

exporting the data to CSV

GitHub: Data Wrangling

EDA with Data Visialization



Example: Payload Mass Vs. Launch Site scatter plot

[GitHub: EDA with Data Visualization]

During Exploratory Data Analysis we plotted the following:

- scatter plots - to show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- bar charts - to show the relationship between the specific categories being compared and a measured value.
- line charts - show trends in data over time (time series).

Charts plotted: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mas vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Perfromed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the ‘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 between the date 2010-06-04 and 2017-03-20 in descending order

GitHub: 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

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 (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub: Interactive Visuals with Folium

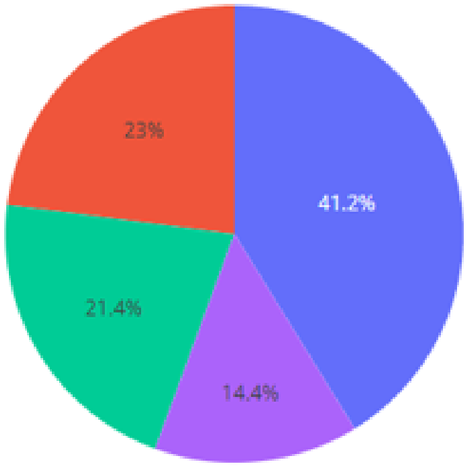
Build a Dashboard with Plotly Dash

1. Added a dropdown list to enable Launch Site selection

2. 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

3. Added a slider to select Payload range

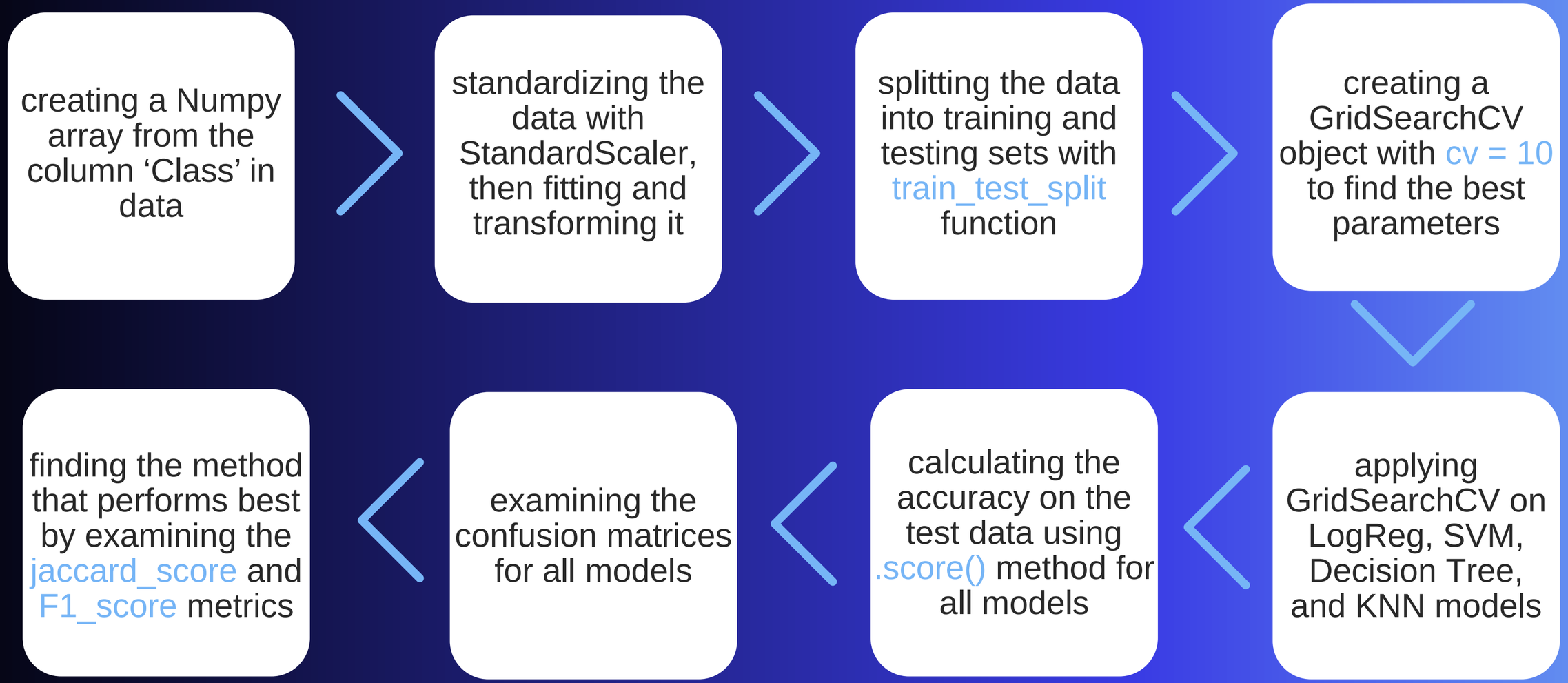
4. Added a scatter chart to show the correlation between Payload and Launch Success



Example: Total Success Launches by Site

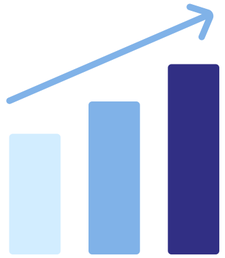
GitHub: SpaceX Dash App

Predictive Analysis (Classification)



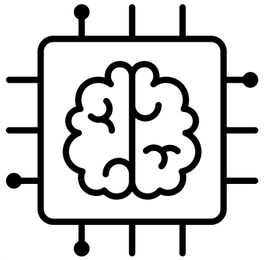
GitHub: Machine Learning Prediction

Results



EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS



PREDICTIVE ANALYSIS RESULTS

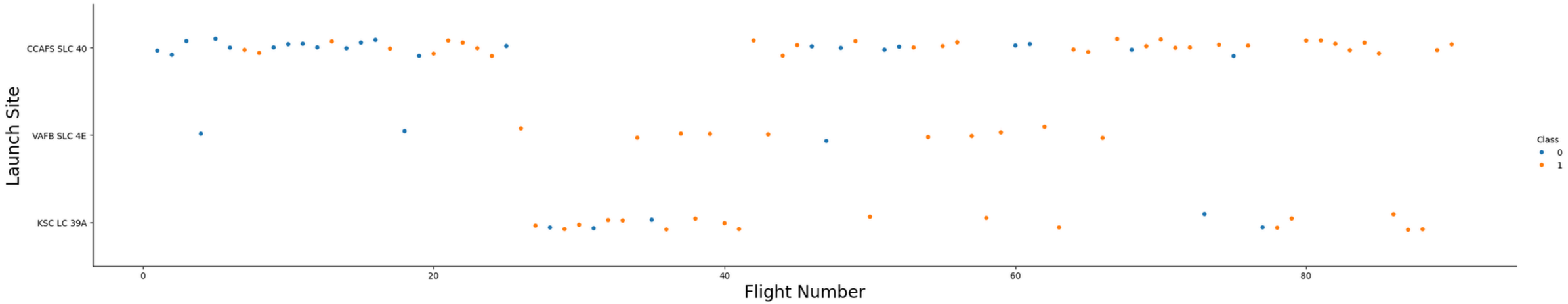
EDA with Visualization

Flight Number vs. Launch Site

The earliest flights all failed while the latest flights all succeeded. It can be assumed that each new launch has a higher rate of success.

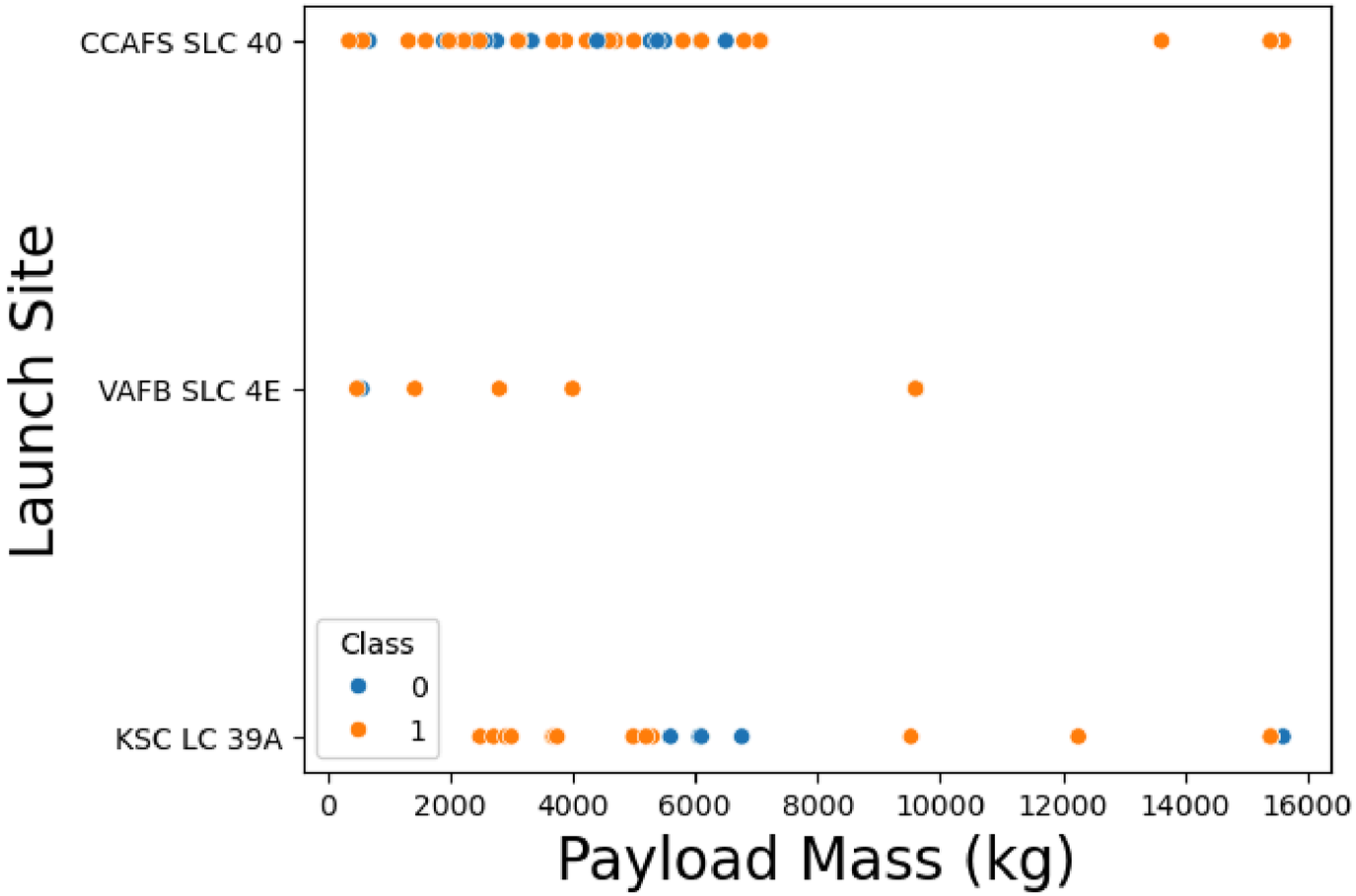
The CCAFS SLC 40 launch site has about a half of all launches.

VAFB SLC 4E and KSC LC 39A have higher success rates.



EDA with visualization

Payload vs. Launch Site



Most of the launches with payload mass over 7000 kg were successful.
For every launch site the higher the payload mass, the higher the success rate.

KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

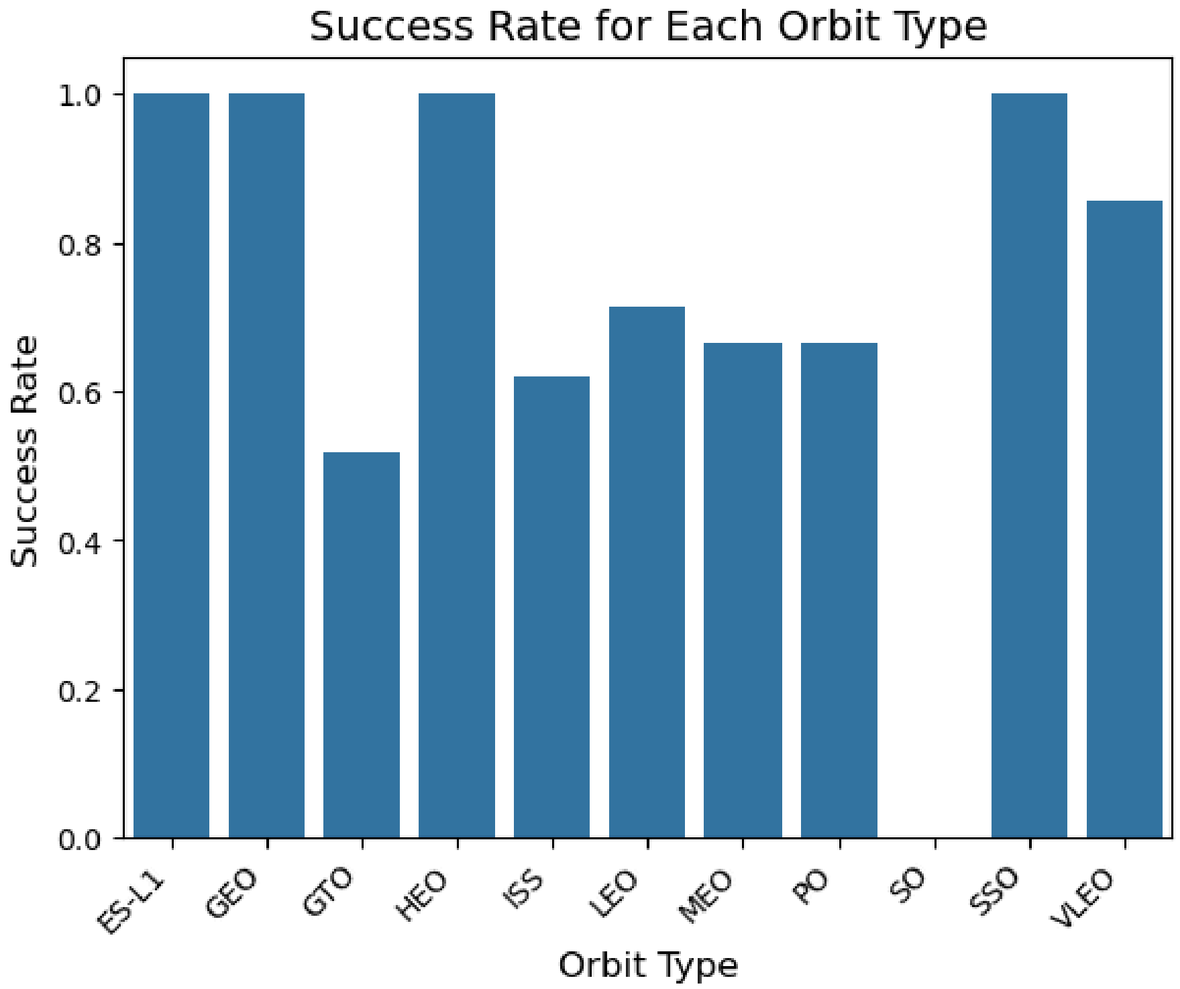
EDA with Visualization

Success Rate vs. Orbit Type

Orbits with 100% success rate are ES-L1, GEO, HEO, SSO.

Orbit with 0% success rate is SO.

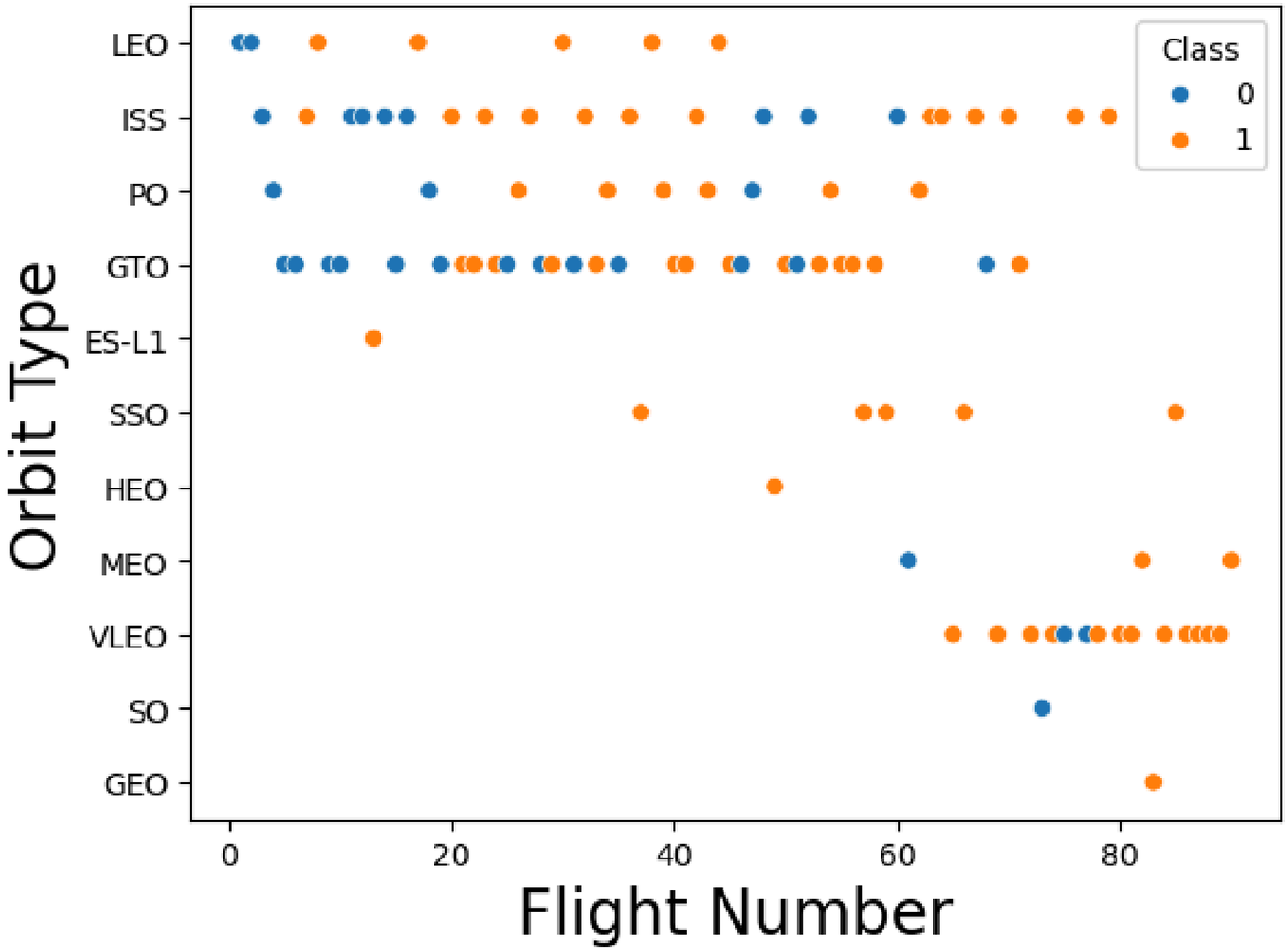
Orbits with success rate between 50% and 85% are GTO, ISS, LEO, MEO, PO.



EDA with Visualization

Flight Number vs. Orbit Type

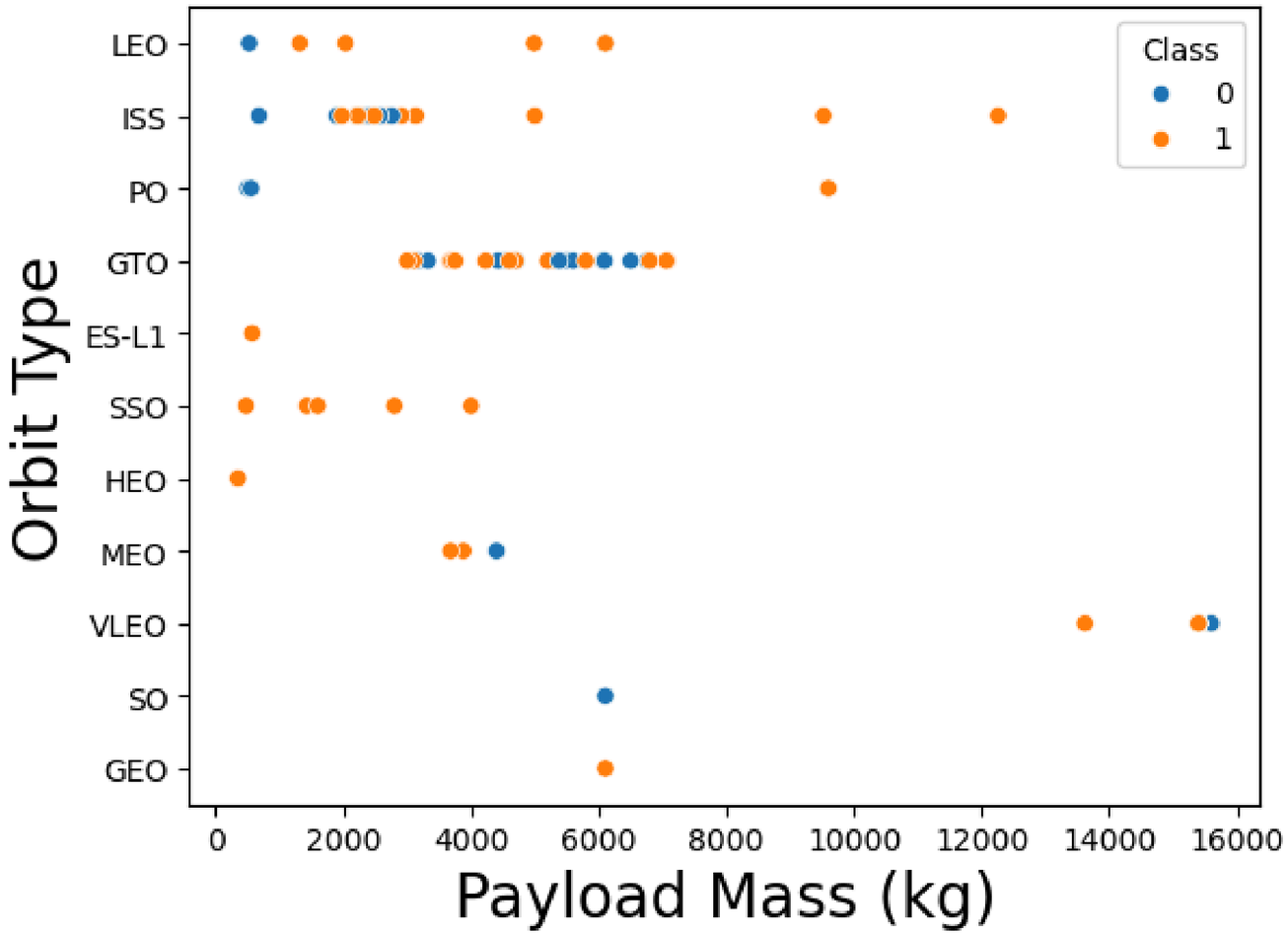
In the LEO orbit the Success appears related to the number of flights;
on the other hand, there seems to be no relationship between flight number when in GTO orbit.



EDA with Visualization

Payload vs. Orbit Type

Heavy payloads have a negative influence on GTO orbits and positive on LEO, ISS, and PO orbits.

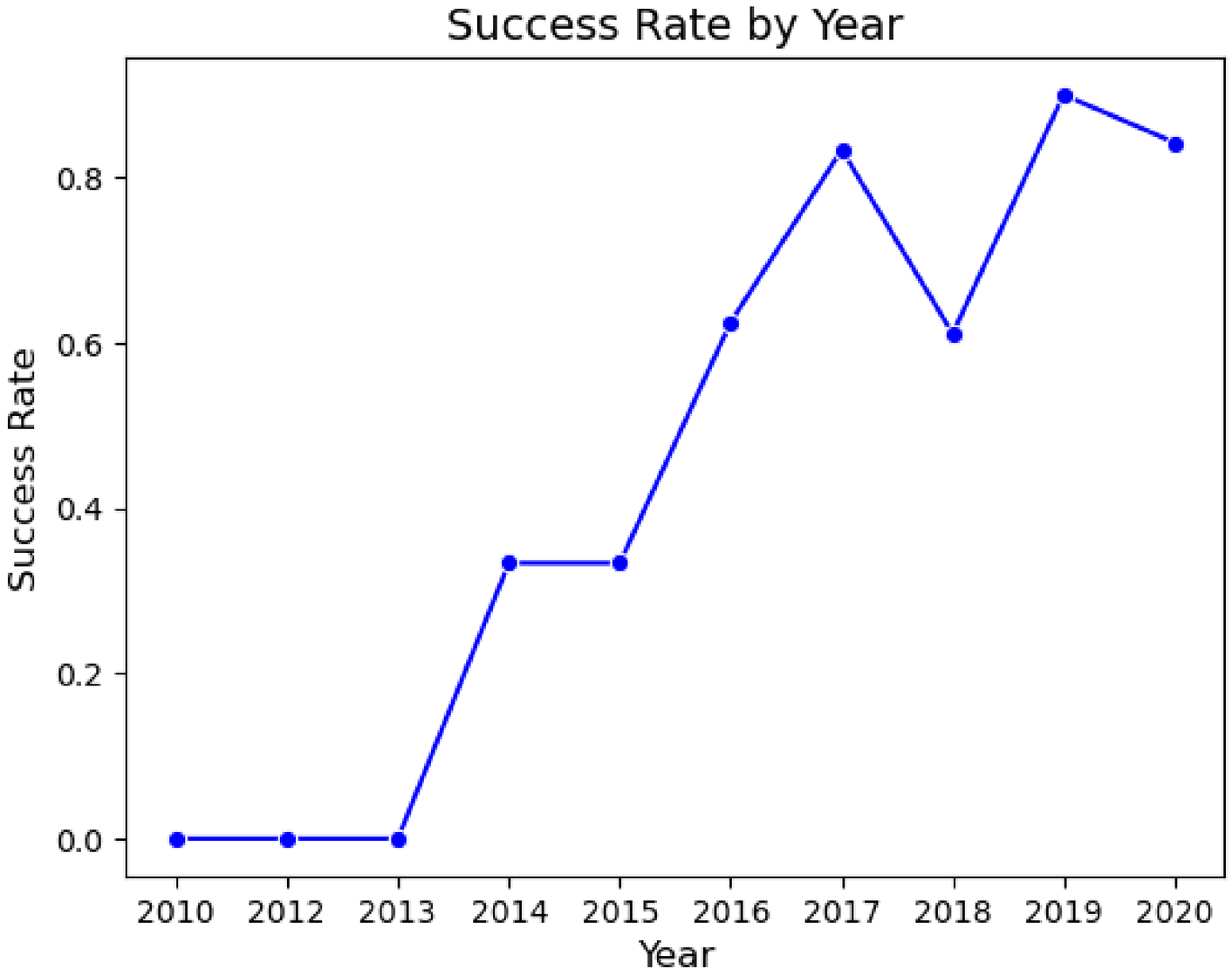


EDA with Visualization

Launch Success yearly trend

The success rate since 2013 kept increasing till 2020.

We can observe a sudden drop in 2018, which presumably happened due to adding more experimental elements into their launches.



EDA with SQL

All Launch Site names

```
%sql select distinct Launch_Site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
Out[15]:
```

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

Displaying the names of the unique launch sites in the space mission.

EDA with SQL

Launch Site names begin with ‘CCA’

```
%sql select * from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5
```

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

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

Displaying 5 records where launch sites begin with the string 'CCA'.

EDA with SQL

Total Payload Mass

```
%sql select SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

Total_Payload_Mass
45596

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

EDA with SQL

Average Payload Mass by F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

Average_Payload_Mass
2534.6666666666665

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

EDA with SQL

First Successful Ground Landing Date

```
%sql select MIN(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

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

EDA with SQL

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version     from SPACEXTABLE where (Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 AND 6000)
```

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

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

Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

EDA with SQL

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(*) as total_number from SPACEXTABLE group by Mission_Outcome;
```

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

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

Listing the total number of successful and failure mission outcomes.

EDA with SQL

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

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

Listing the names of the booster versions which have carried the maximum payload mass.

EDA with SQL

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

```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
      where Date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

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

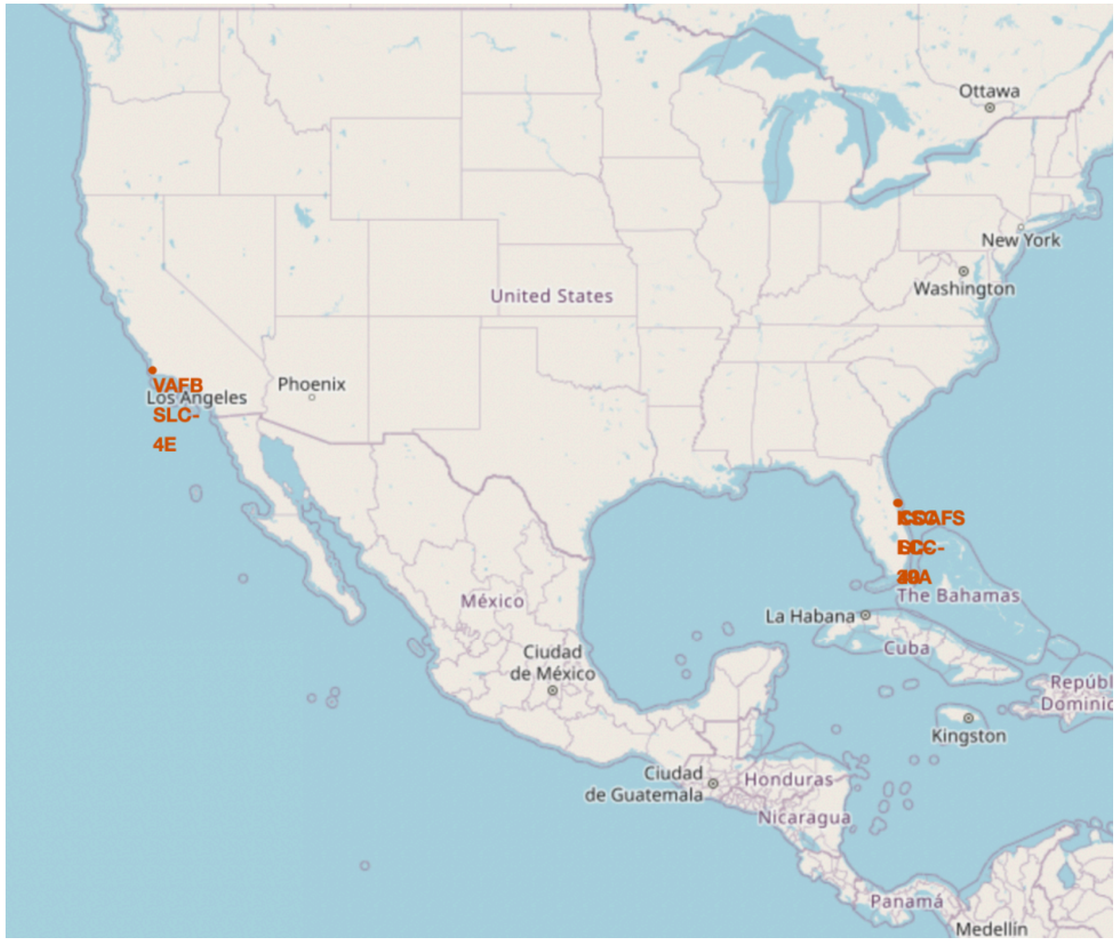
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Interactive map with Folium

All Launch Sites’ Location Markers on a Global Map

Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

All launch sites are in very close proximity to the coast. Launching rockets towards the ocean minimises the risk of having any debris dropping or exploding near people.

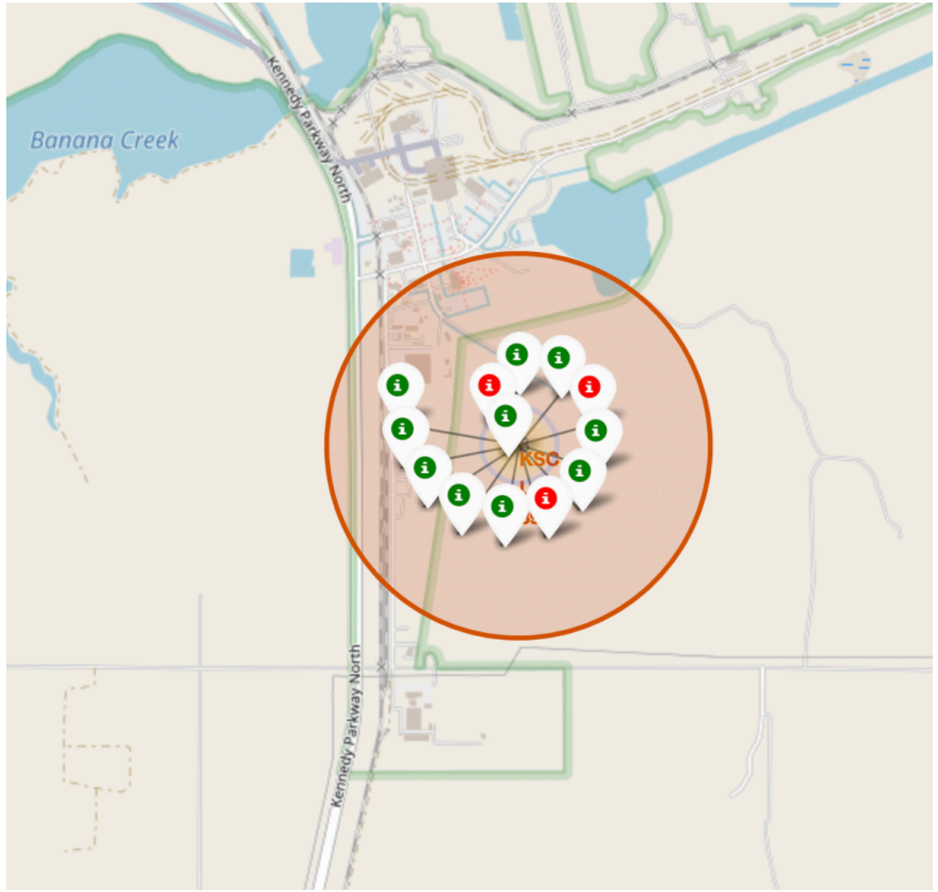


Interactive map with Folium

Colour-labeled Launch Records on the Map

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
Green Marker = Successful Launch
Red Marker = Failed Launch

Launch Site KSC LC-39A has a very high Success Rate.



Interactive map with Folium

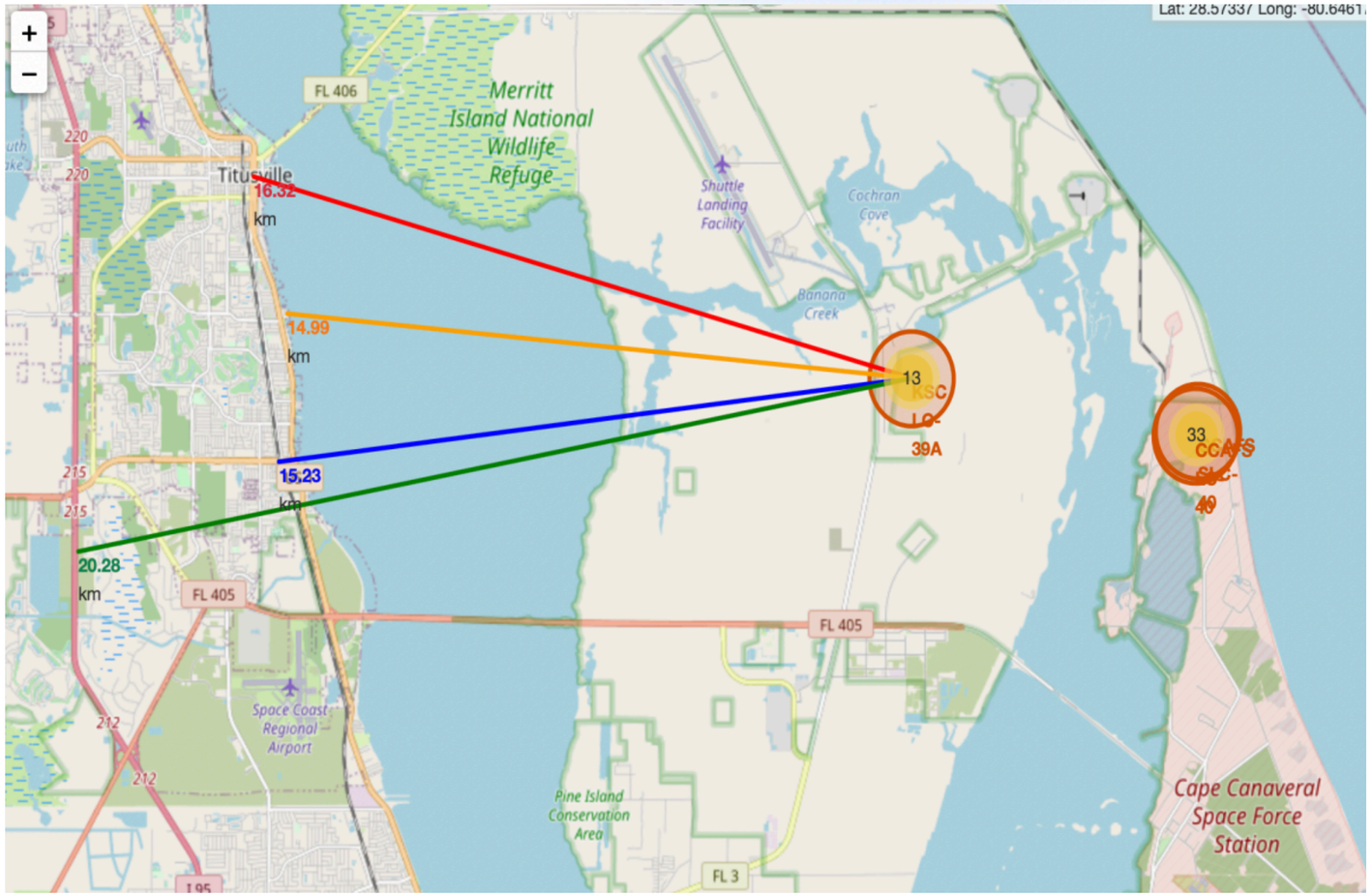
Distance from the Launch Site KSC LC-39A to its Proximities

From the visual analysis of KSC LC-39A we can clearly see that it is:

- relatively close to railway (15.23 km)
- relatively close to highway (20.28 km)
- relatively close to coastline (14.99 km)

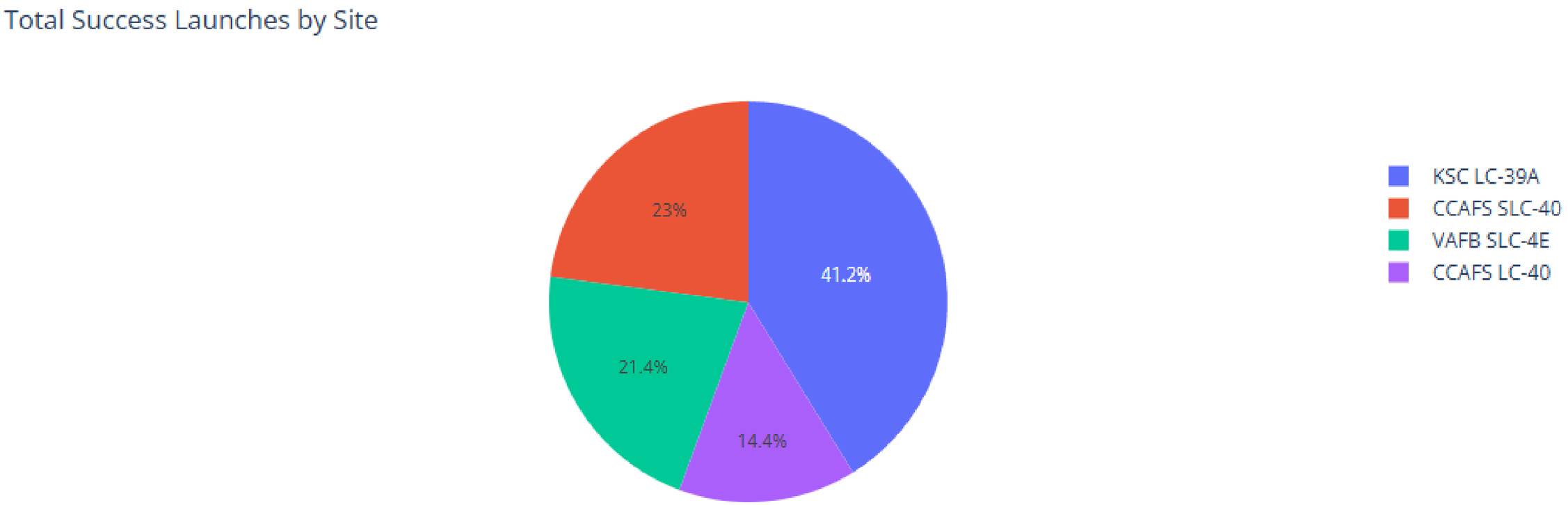
Also the launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).

This relative closeness is, on the one hand, giving the site accesssibility and logistical support, but on the other hand, arises potential safety concerns.



Build a Dashboard with Plotly Dash

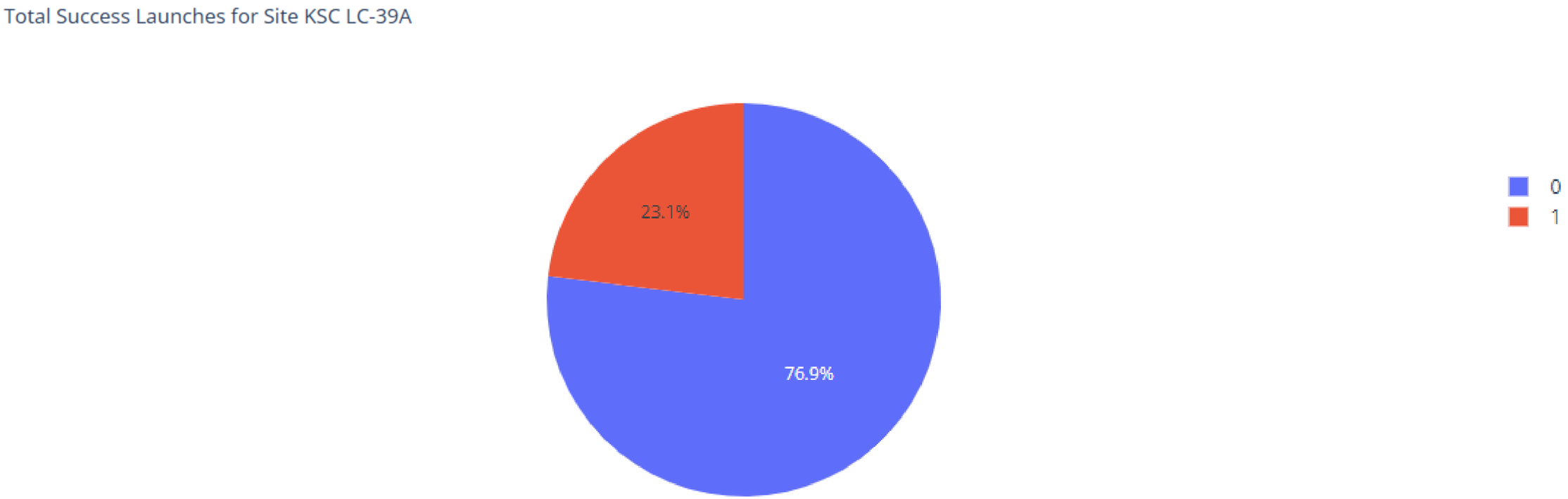
Total Success Launches by Site



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Build a Dashboard with Plotly Dash

Launch Site with Highest Launch Success Ratio



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Build a Dashboard with Plotly Dash

Payload Mass vs. Launch Outcome for all sites

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Predictive Analysis

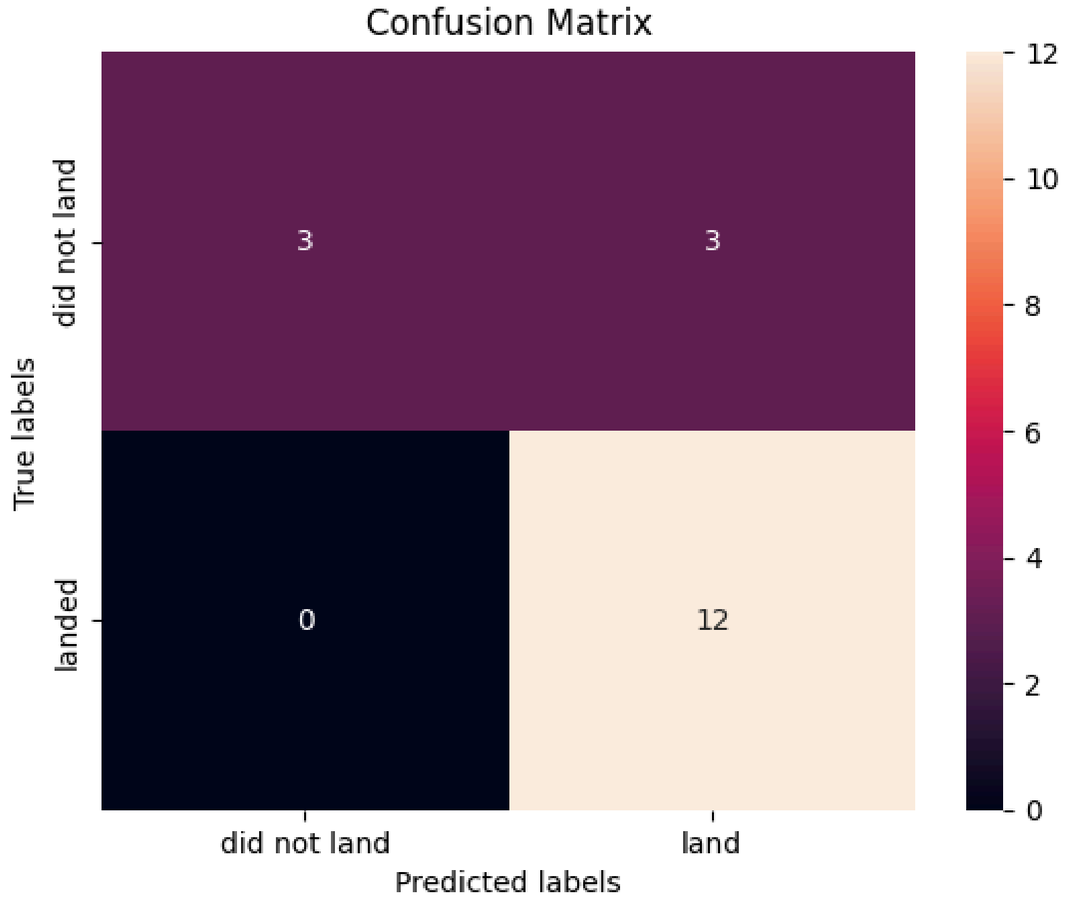
Classification Accuracy

	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	0.846429	0.848214	0.875000	0.848214

Based on the scores of the Test Set, we can assume that the Decision Tree Model performs best. Although, the size of the test set is relatively small, and there may be a need for further testing.

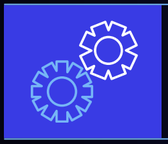
Predictive Analysis

Confusion Matrix

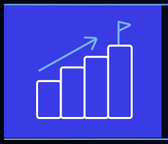


Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusion



Decision Tree Model is the best algorithm for this dataset.



Launches with a low payload mass show better results than launches with a larger payload mass.

Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.

The success rate of launches increases over the years.

KSC LC-39A has the highest success rate of the launches from all the sites.

Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Special Thanks to:

Coursera

IBM