IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexander Boren

02/13/2025

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of All Results

- Exploratory Data Analysis Outcomes
- Interactive Analysis Demos Screenshots
- Predictive Analysis Outcomes

Introduction

- Project Background and Context
  - SpaceX is the leading provider for cost effective space flights. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.
- Questions to answer
  - How does the variables launch site, payload, payload mass, orbit type affect the outcome of successful landing of the 1st stage?
  - Where are the launch sites located and does location matter?
  - What is the best classification model to produce the most accurate result?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Using SpaceX REST API
    - Using Web Scraping from Wikipedia
- Perform data wrangling
    - Filter the data
    - Replace missing or null values
    - Preform One Hot Encoding to simplify categorical variables into binary format
- 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 each classification models to ensure the most optimal result

Data Collection

Data Collection involved two types of collection to get the most amount of data

1. Using API request from SpaceX REST API
2. Using Web Scraping the data from a table in SpaceX's Wikipedia site

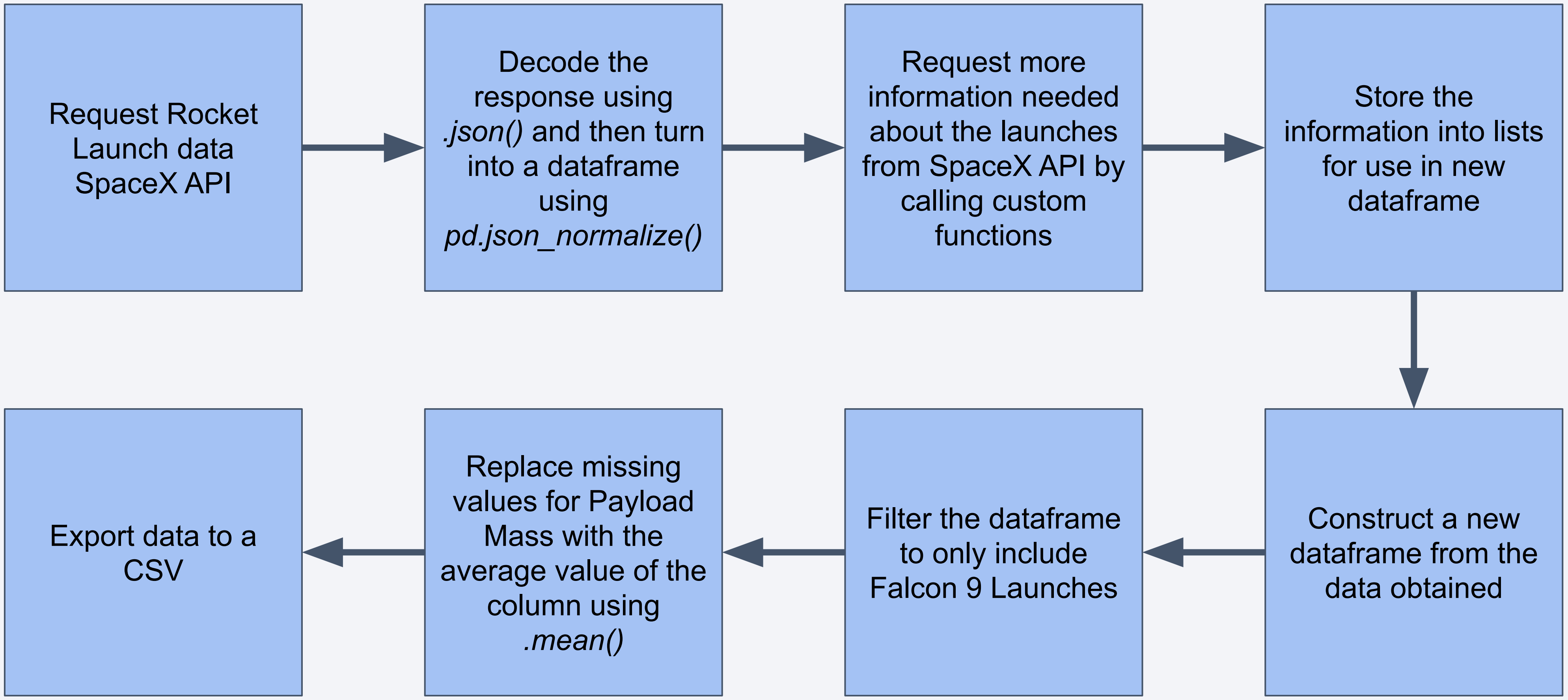
Data Columns obtained from SpaceX Rest API:

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

Data Columns obtained from Wikipedia Web Scraping:

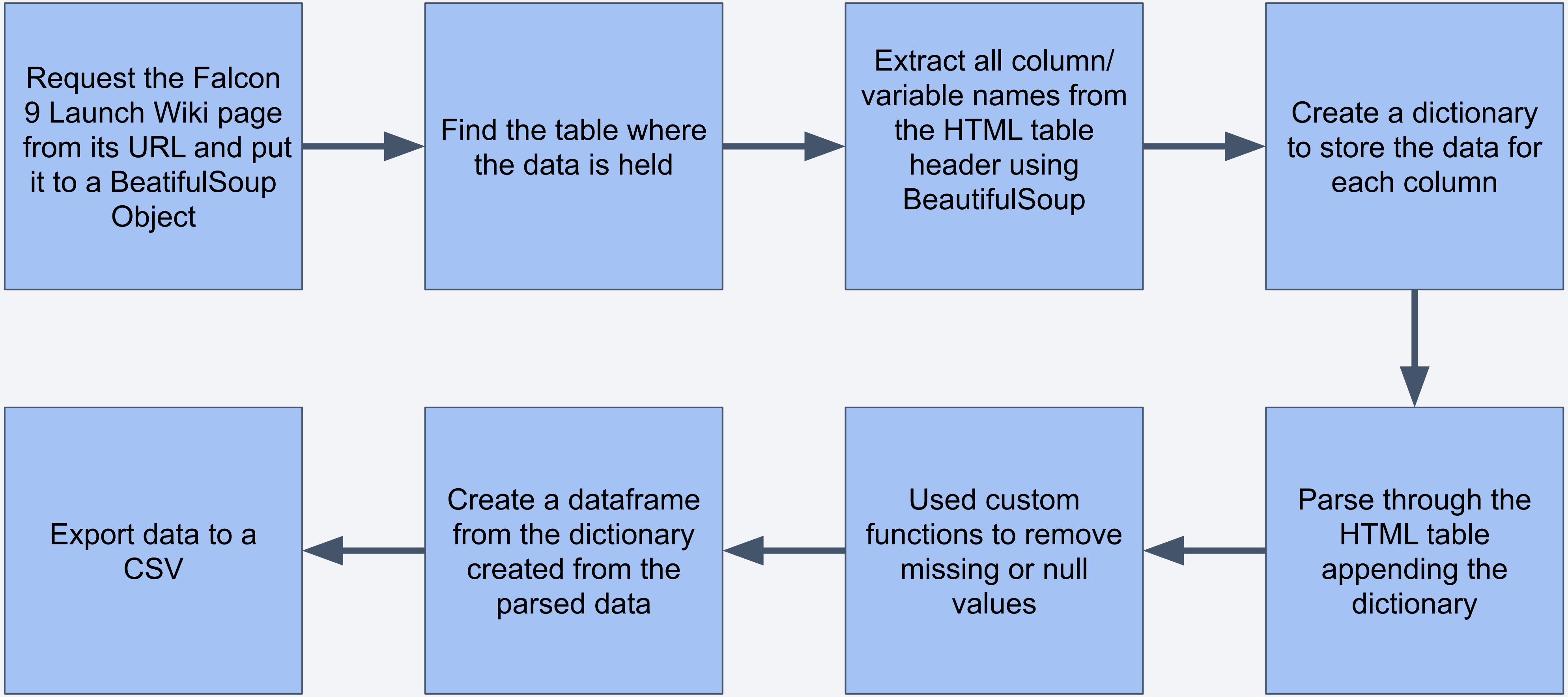
Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Date, and Time

Data Collection – SpaceX API



[GitHub URL: Data Collection API]

Data Collection - Scraping

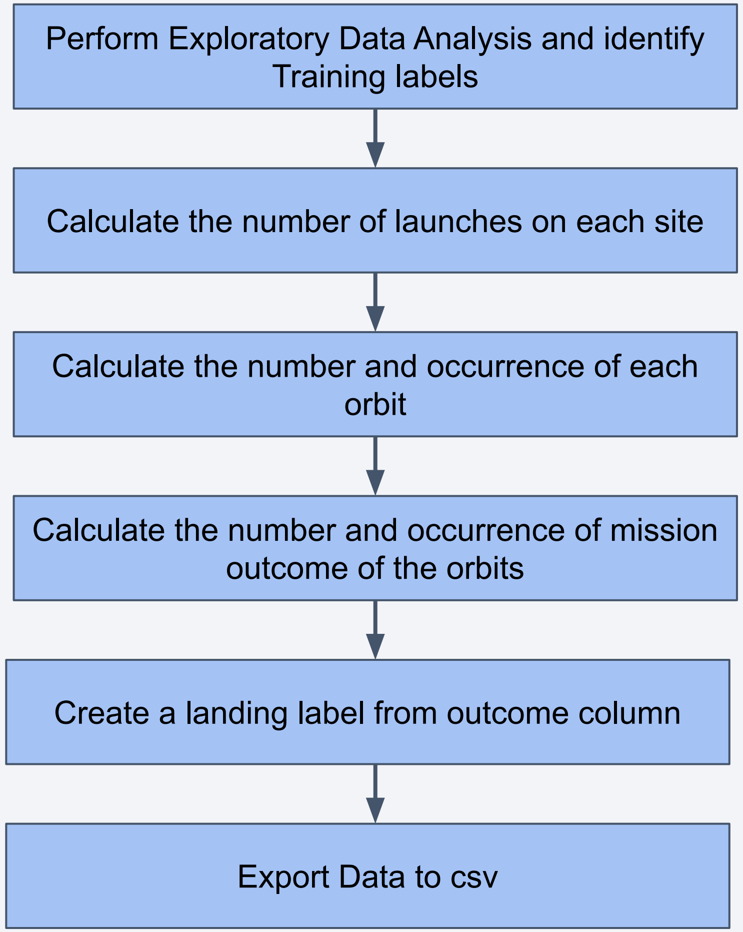


[GitHub URL: Data Collection Web Scraping]()

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. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We will mainly convert those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.



[GitHub URL: Data Wrangling]

EDA with Data Visualization

Charts plotted:

Flight Number and Payload Mass,Flight Number and Launch Site, Payload Mass and Launch Site, Success rate per Orbit type, Flight Number and Orbit Type, Payload Mass and Orbit Type, and Success Rate Yearly Trend.

Scatter Plot was used to show correlations between variables. If there was a correlation, the variables could be used for future machine learning application.

Bar Chart was used to show comparisons between categorical data. It helps identify outliers in a category when comparing against numerical values.

Line Charts was used to show change of data over time.

[GitHub URL: EDA with Data Visualization]

EDA with SQL

Preformed 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. Using 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 year 2015.
- Rank 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

Mark all launch sites on a map

- Added Circle and Marker objects to highlight the location for each launch site. To help differentiate each launch site by adding a Popup and Text Label.

Color-Labeled launch outcomes on the Map

- Created marker cluster for each launch site.
- Each cluster holds an icon showing red or green. Which signifies a successful launch or failure.
- Each cluster also counts and able to show how many launches happened at each site.

Launch Site CCAFS SLC-40 distance to Proximities

- Added colored lines to between Launch Site CCAFS SLC-40 and closest significant locations:Coastline, Highway, Railway, and City. To help show the distance.

[GitHub URL: Interactive Visual Analytics with Folium]

Build a Dashboard with Plotly Dash

Launch Site Drop-Down List

- To be able to choose all or individual Launch Sites

Pie Chart showing Launch Outcomes

- To show all launch sites total count of successful outcomes.
- When selected individual sites are chosen. The pie chart shows the launch outcome ratio between successful and failure.

Slider for Payload Mass Range

- To be able to limit the weight of the rockets. This help show launch outcomes per weight class.

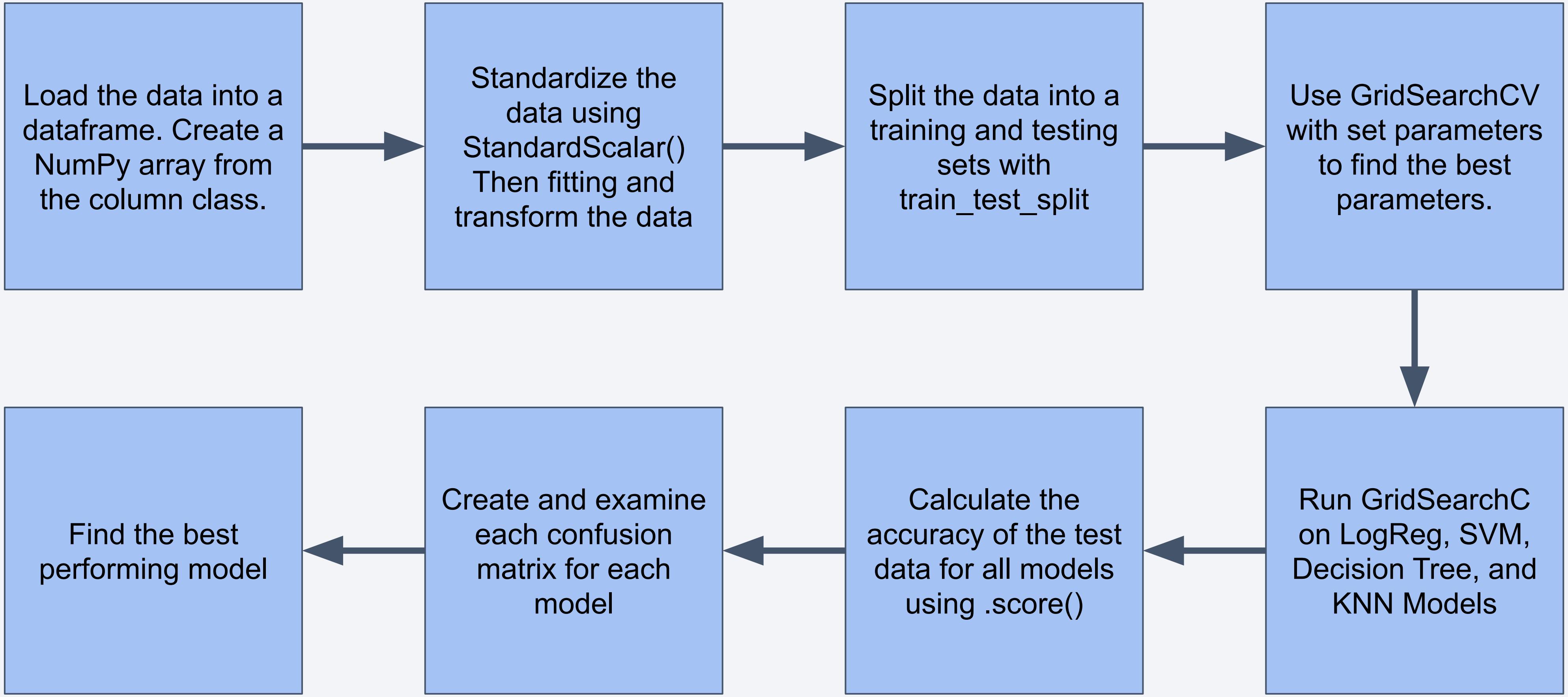
Scatter Chart of Payload vs. Launch Outcome

- To be able to show the correlation between Payload mass and the Launch Outcome.

GitHub URL: Interactive Visual Analytics with Dashboard

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart



[GitHub URL: Predictive Analysis]()

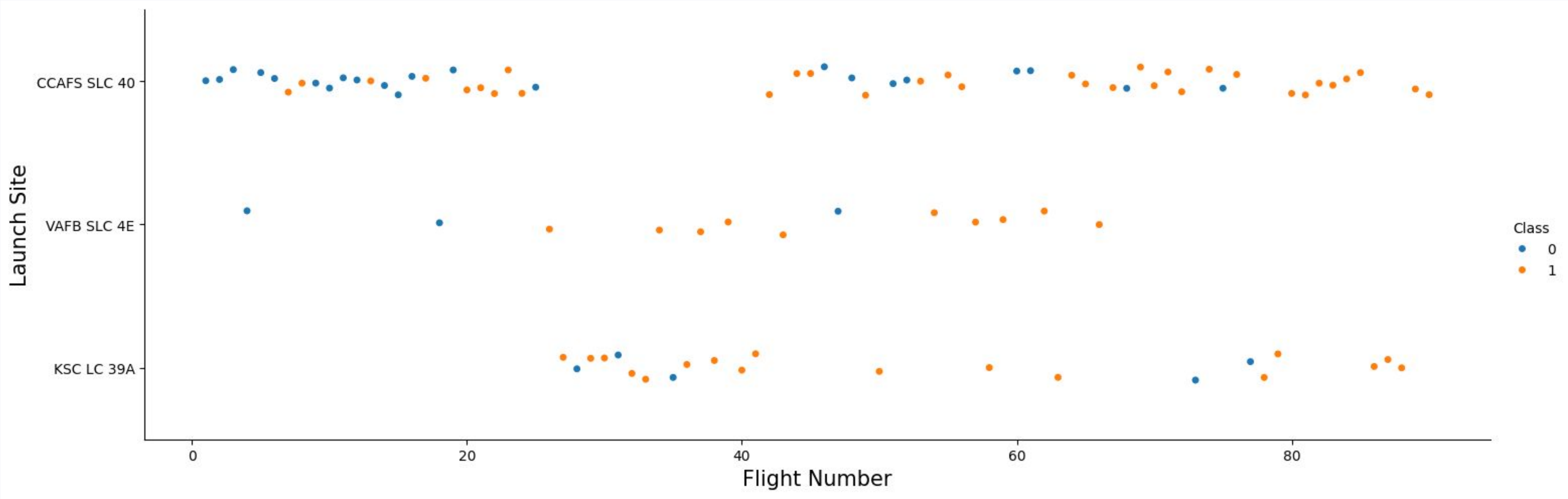
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



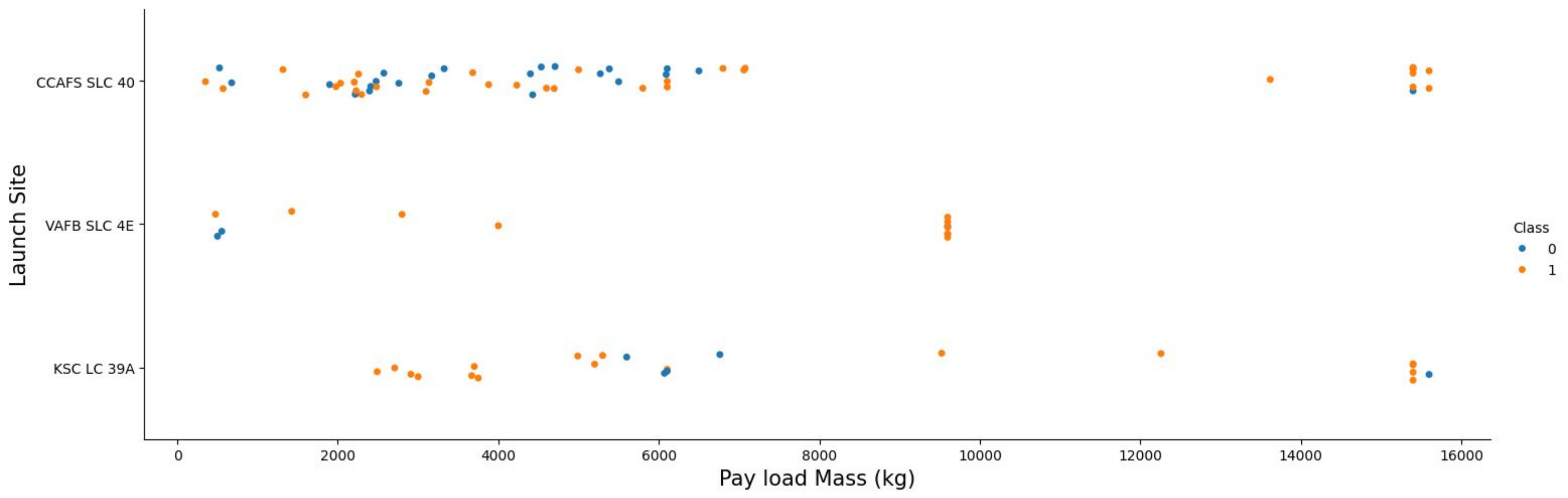
The earliest flights failed the most and the latest flights succeeded.

CCAFS SLC 40 has the most launches

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

The more info they gathered from the launches. The higher the rate of success.

Payload vs. Launch Site

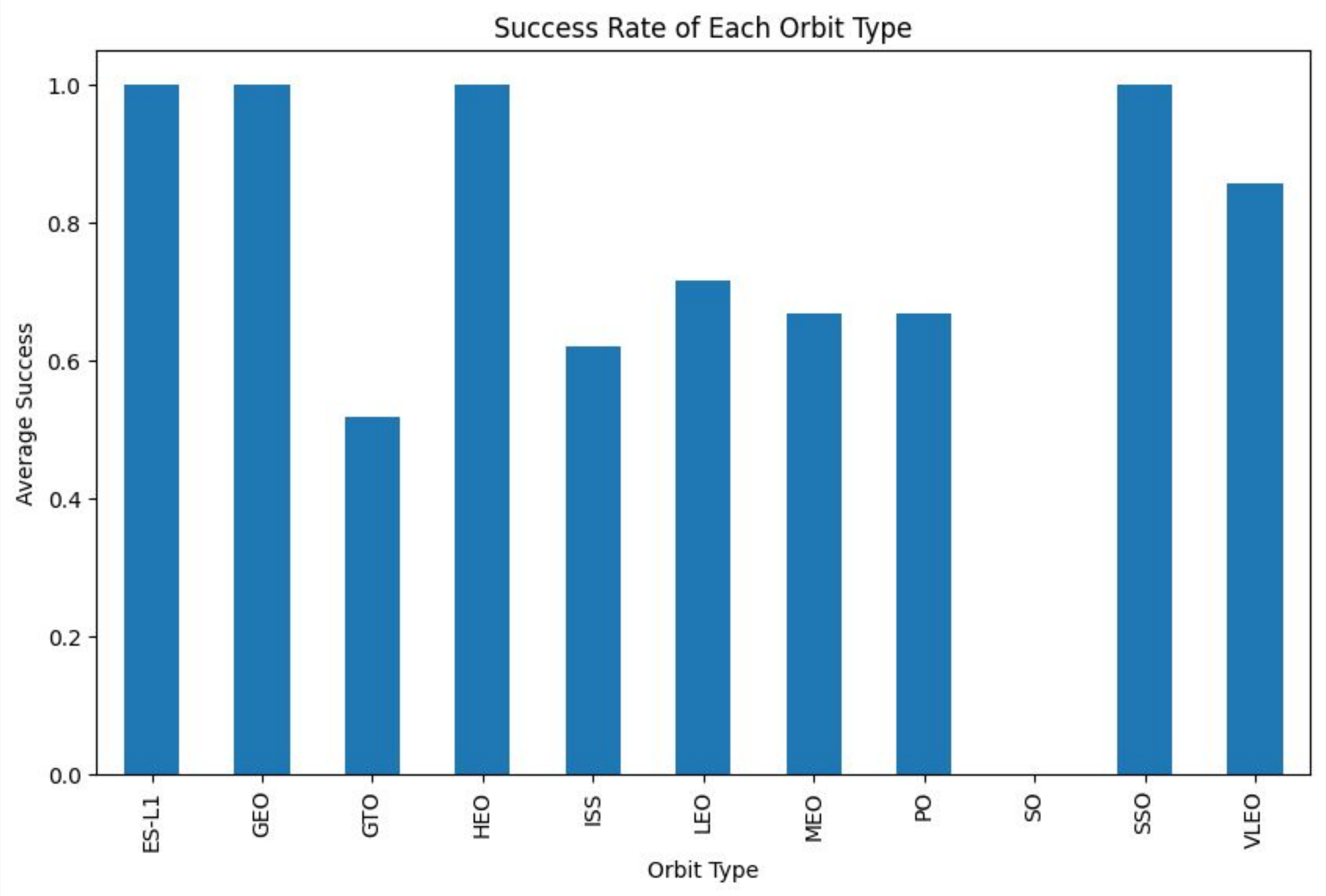


There are more launches less than 8000 kg.

Though the launches over 8000 kg are mostly successful.

There are no rockets launched at VAFB SLC 4E over 10000 kg.

Success Rate vs. Orbit Type



Orbits with 100% success rate

- ES-L1, GEO, HEO, and SSO

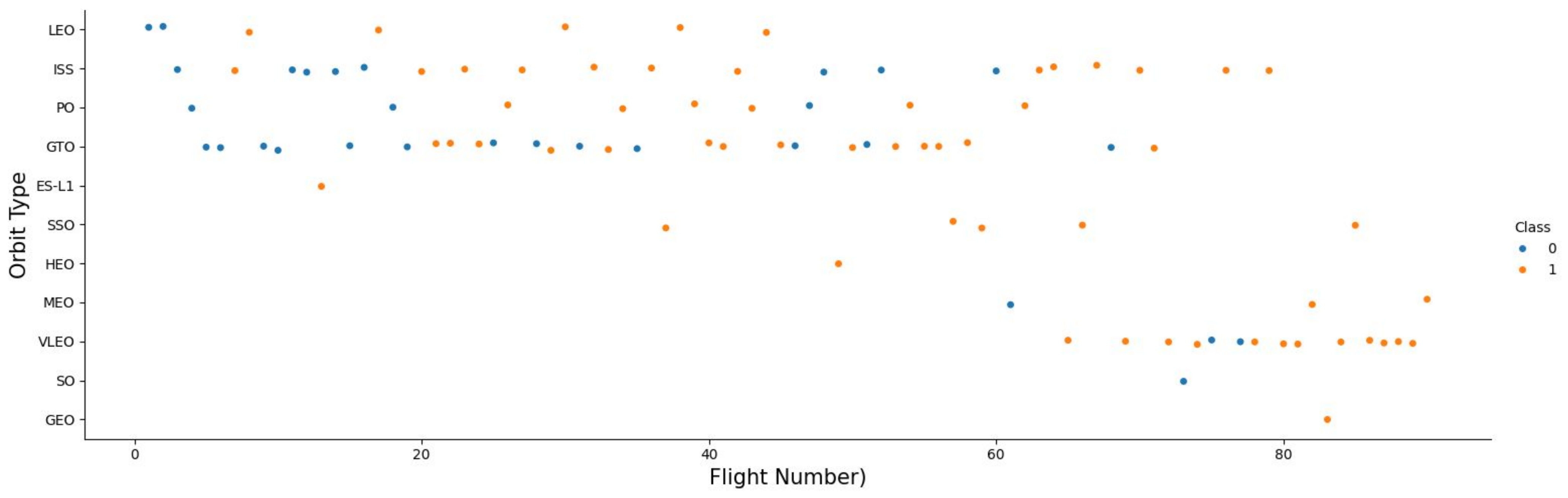
Orbits with success rates between 50% and 85%

- GTO, ISS, LEO, MEO, PO, and VLEO

Orbits with 0% success rate

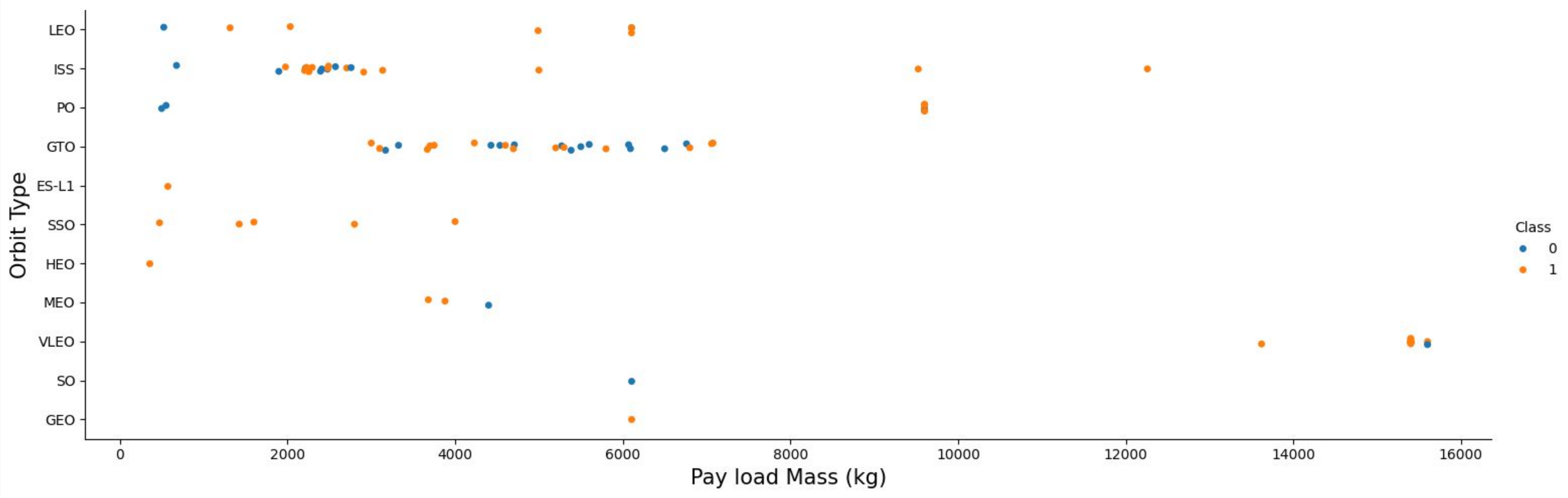
- SO

Flight Number vs. Orbit Type



In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

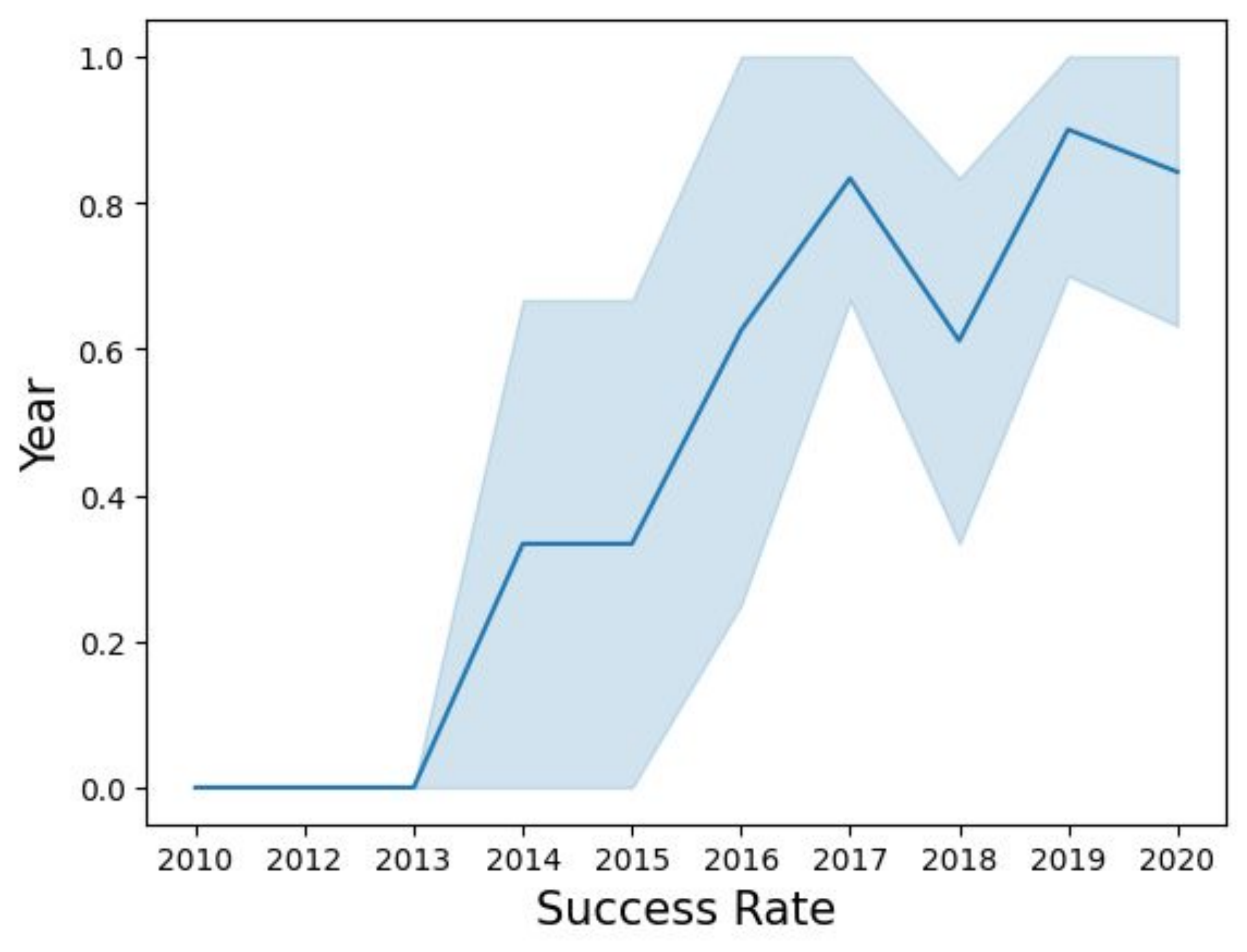
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO, it’s difficult to distinguish between successful and unsuccessful landings as both outcomes are present

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020. Though there is a dip in success during 2018.

All Launch Site Names

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

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

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

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string ‘CCA’

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

Python

* 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

Total Payload Mass

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

```
%sql select SUM(PAYLOAD_MASS__KG_) as 'Total Payload Mass carried by NASA (CRS)' from SPACEXTABLE where Customer = 'NASA (CRS)';
```

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

Total Payload Mass carried by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select AVG(PAYLOAD_MASS__KG_) as 'Average Payload Mass Carried by F9 v1.1 Booster' from SPACEXTABLE where Booster_Version = 'F9 v1.1' ;
```

Python

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

Average Payload Mass Carried by F9 v1.1 Booster
2928.4

First Successful Ground Landing Date

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

```
%sql select MIN(Date) as 'First Successful Landing on Ground Pad' from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

Python

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

First Successful Landing on Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Python

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

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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
%sql select  Mission_Outcome, COUNT(Mission_Outcome) as 'Total' from SPACEXTABLE GROUP BY Mission_Outcome
```

Python

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ >= (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

Python

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

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
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(Date, 6,2) as 'Month', substr(Date,0,5) as 'Year', Landing_Outcome, Booster_Version, Launch_Site  from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015'
```

Python

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

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank 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

```
%sql select Landing_Outcome, COUNT(Landing_Outcome) from SPACEXTABLE where Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome order by count(Landing_Outcome) DESC;
```

Python

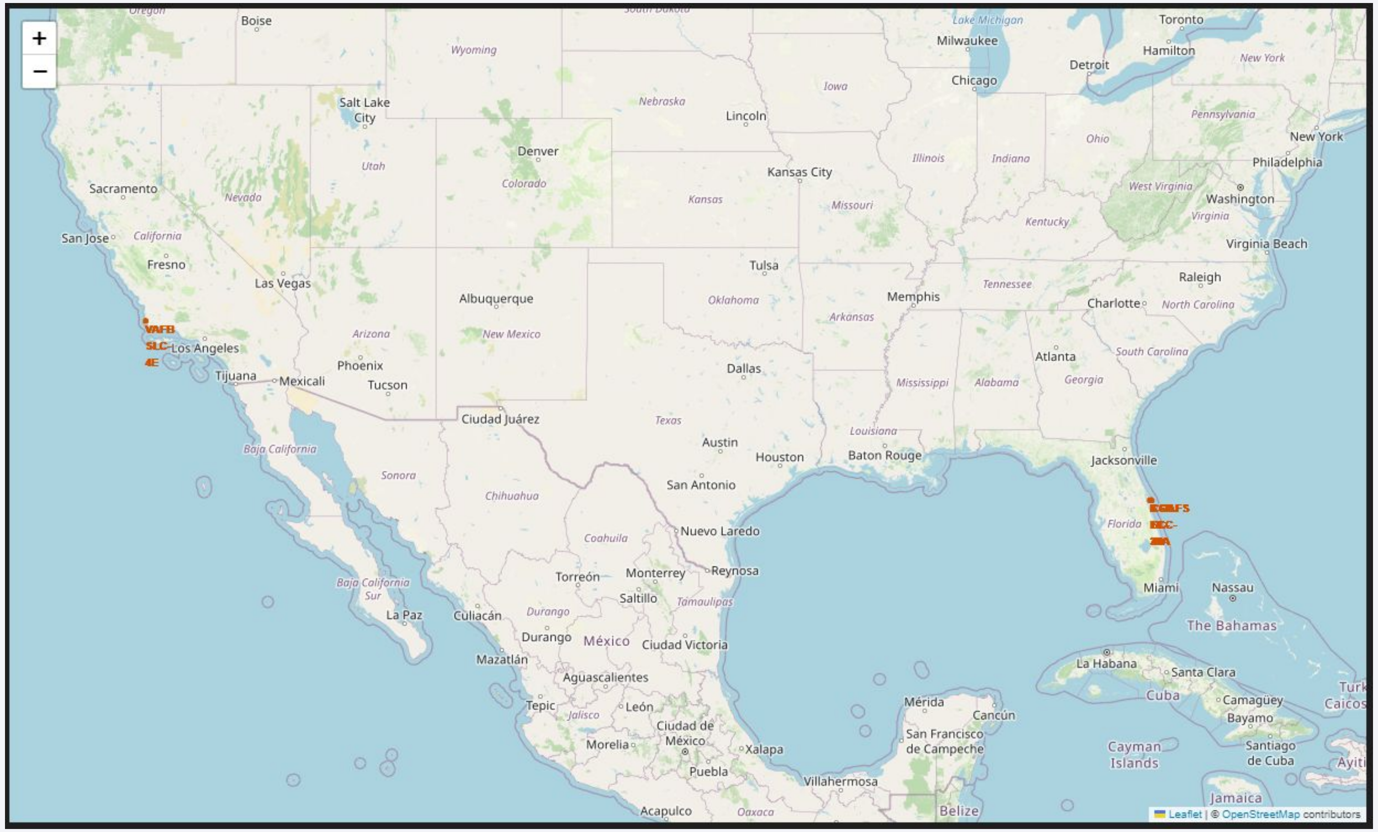
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	COUNT(Landing_Outcome)
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 Proximities Analysis

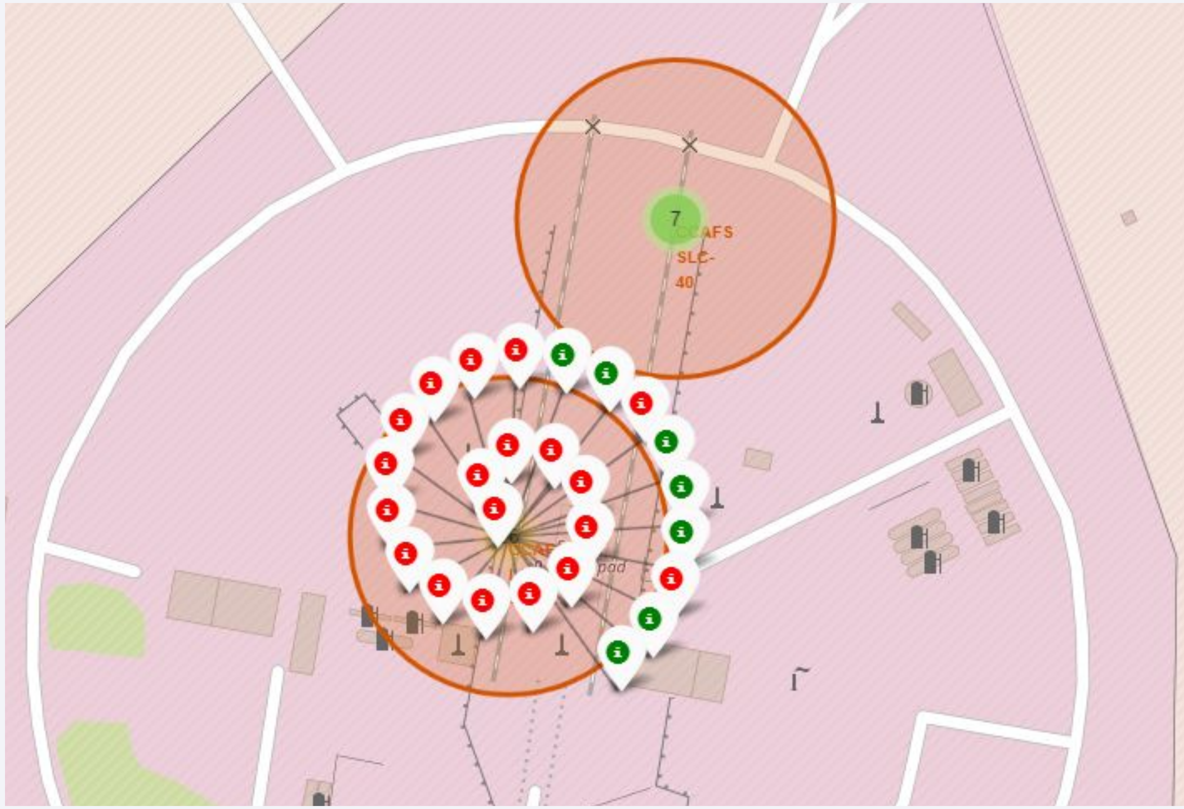
All Launch Sites marked on a global map



The launch sites are as close as they can be to the equator for the US. Since it is easier to launch due to rotation speed of the Earth.

Launch sites are very close to the coastlines. So that if the rocket fails. It would have a higher chance to not cause damage.

Color-Labeled launch outcomes on the Map



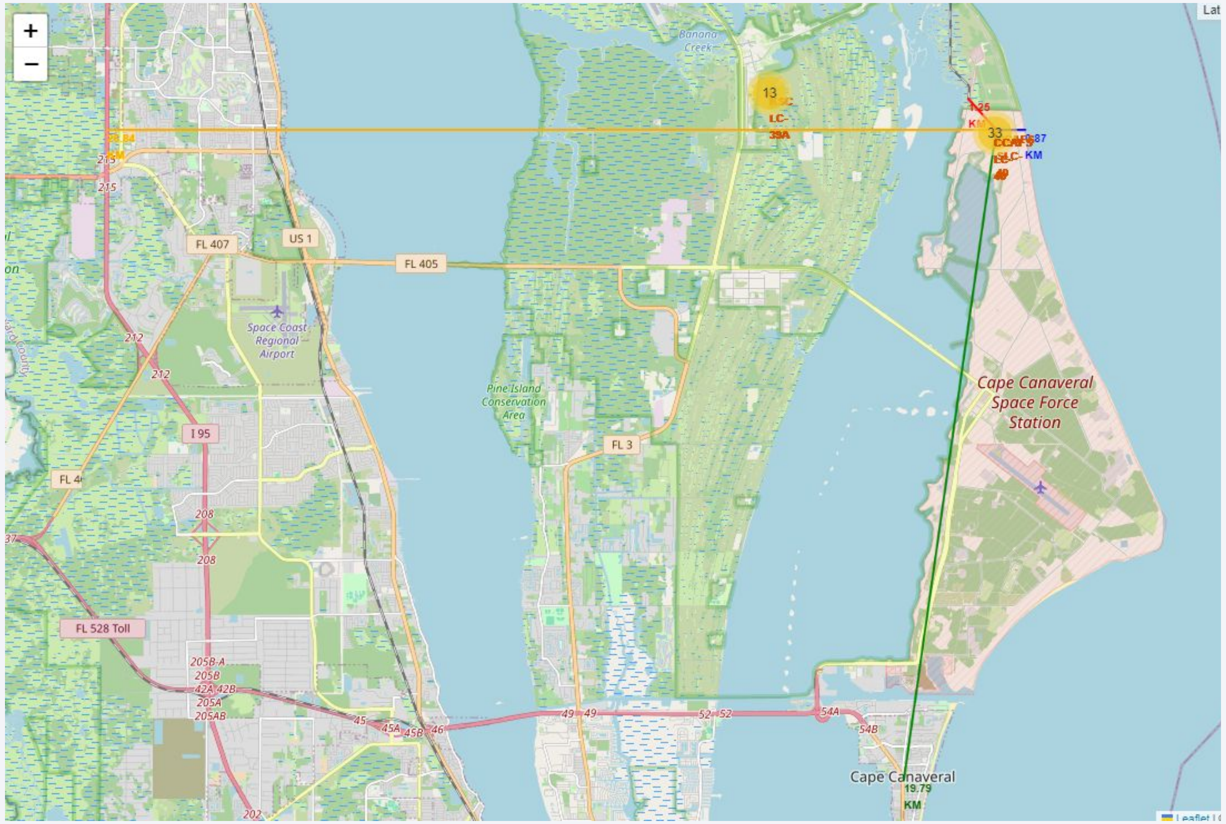
Since the markers are color labeled. We can make a quick glance to identify the success rate of the launch site.

Green Marker is Success

Red Marker is Failure

CCAFS LC-40 has a high amount of failed launches

Launch Site CCAFS SLC-40 distance to Proximities



Looking at Launch site CCAFS SLC-40.
We can see that:

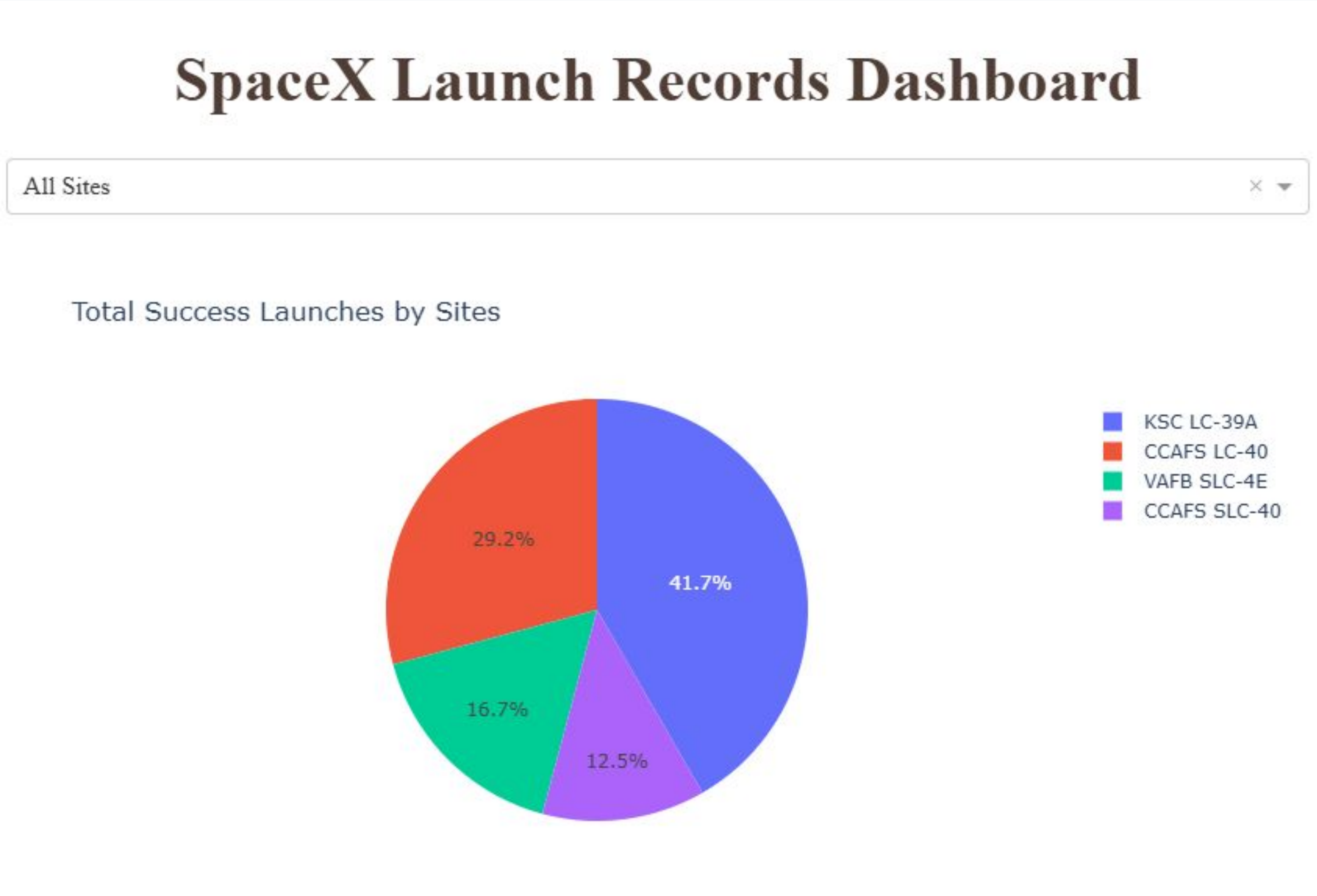
- Extremely close to the coastline (0.87 KM)
- Extremely close to a railroad (1.25 KM)
- Far away from a major city (19.79 KM)
- Far away from major highway (26.84 KM)

The launch site is far enough away from major cities and close enough to a coastline to not cause damage. While being close enough to a railroad to help with shipments.

Section 4

Build a Dashboard with Plotly Dash

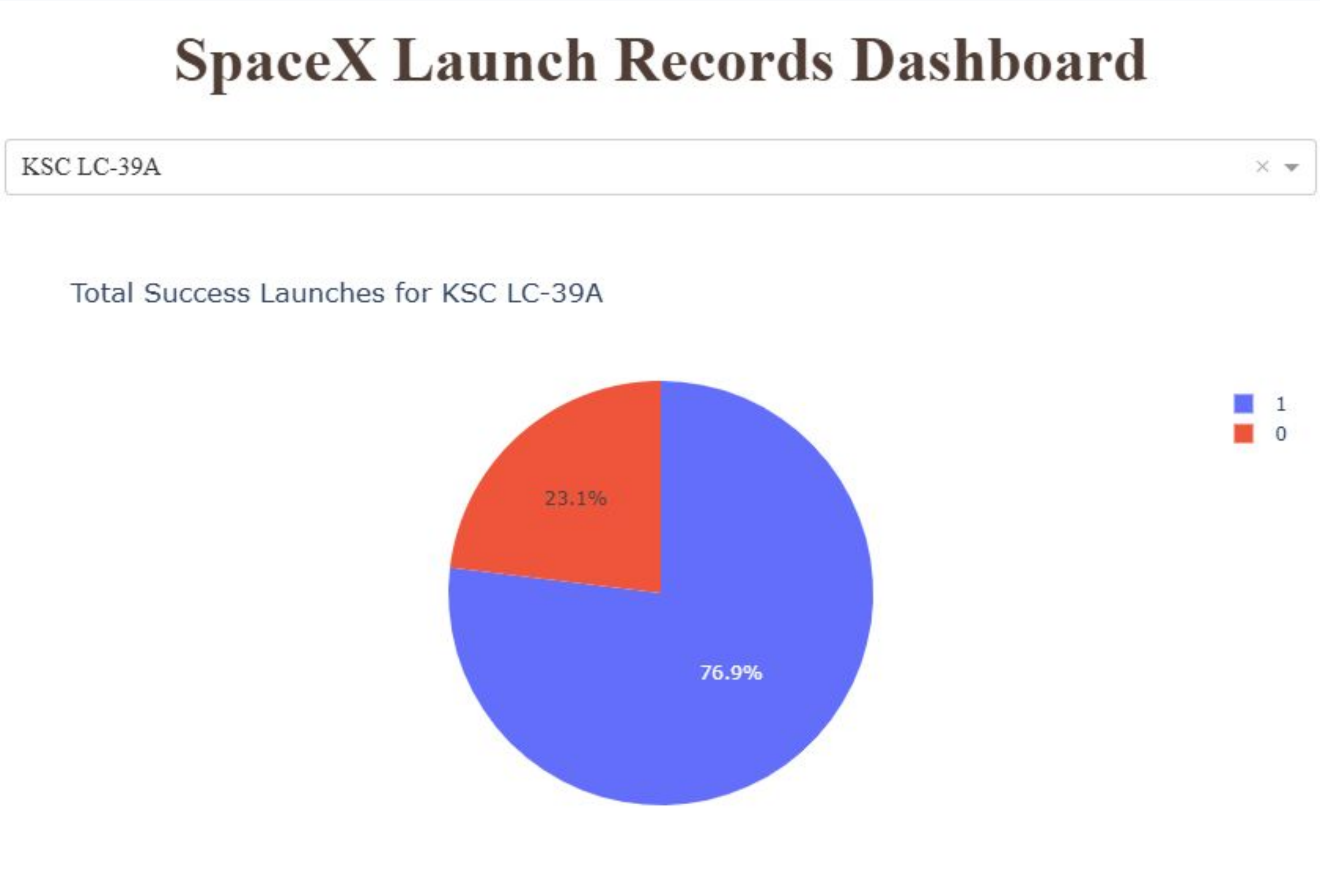
Launch Success Count for All Sites



The chart shows that KSC LC-39A is the most successful launches.

While CCAFS SLC-40 has had the least successful launches.

Launch Site with Highest Launch Success Ratio



KSC LC-39A has the highest success ratio with:
10 successful launches(76.9%)
3 failed launches(23.1%).

Payload vs Launch Outcome for all sites

Between the two charts:

The payload between 2000 and 5000 has more successes.

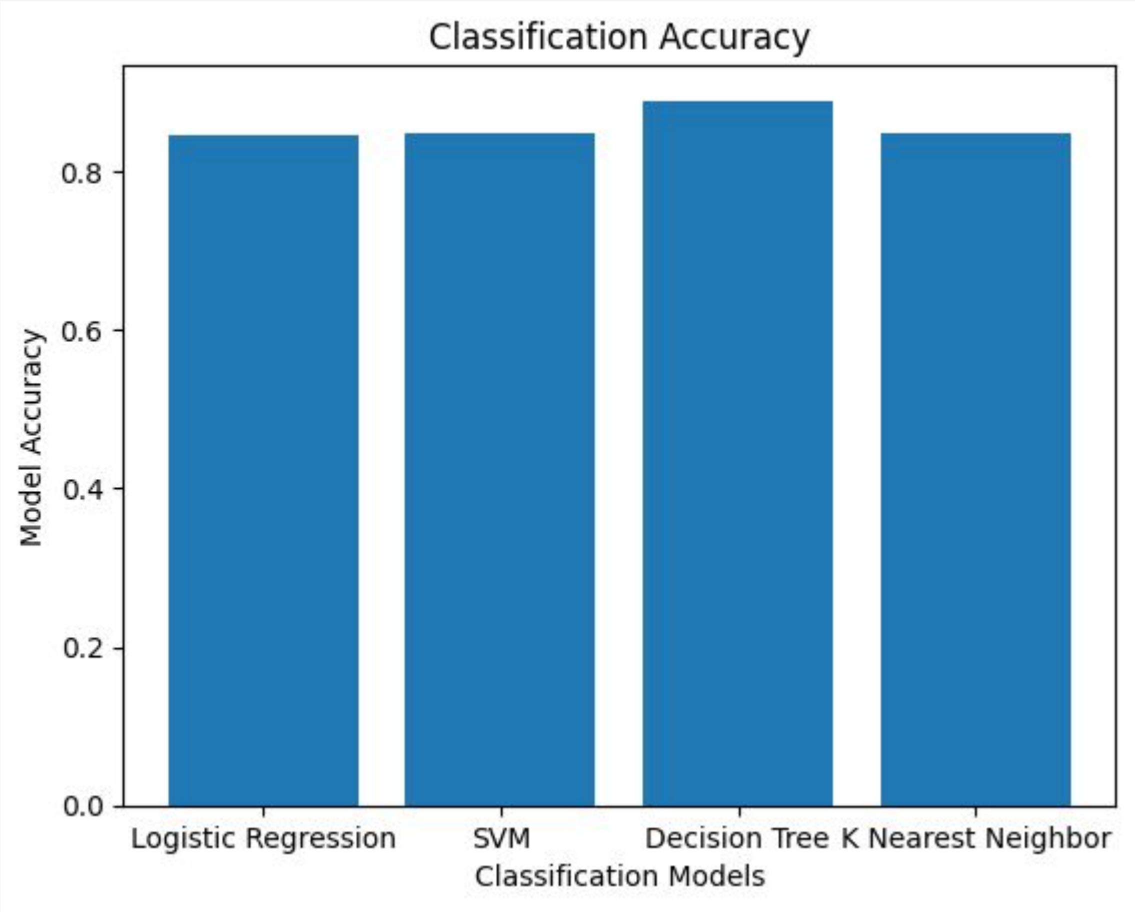
The FT Booster Version has had the most success launch outcomes.

Section 5

Predictive Analysis (Classification)

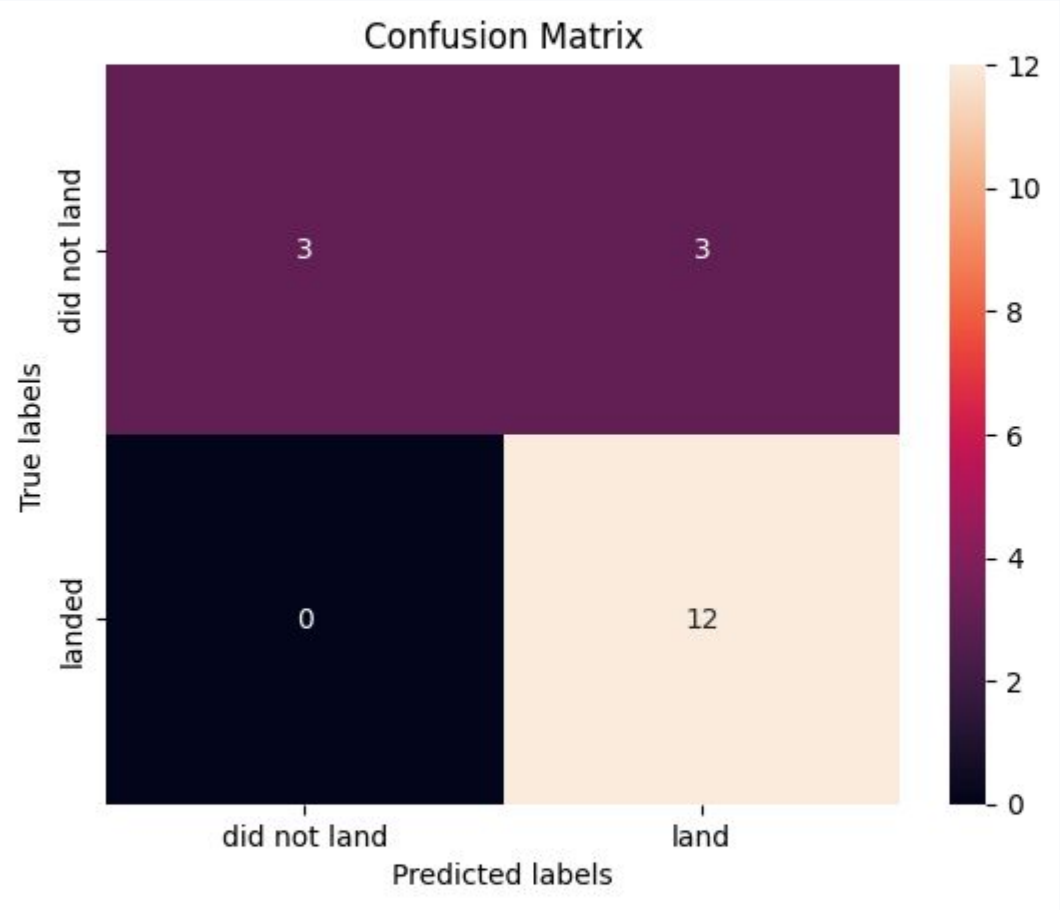
Classification Accuracy

The model with the highest classification accuracy is the Decision Tree model.



Confusion Matrix

Even though the Decision Tree model is is the most accurate. Running the test data through produces the same result for each model. This could be due to the low testing data.



Conclusions

- The Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Lower Payload masses have better outcomes of success.
- The success rate of launch increases over the years. This could be from more data from the missions
- The launch sites should be close to the Equator line and be close to the coastline in case of failure.
- The KSC LC-39A launch site has the highest chances of success.
- The Decision Tree Model is the best algorithm for the dataset.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!