Data Science Capstone Project

Edgar N. Tafaleng
20241004

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Acknowledgements

Executive Summary

1. Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA) using SQL
- EDA using Data Visualization
- Creating Interactive Visuals with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

2. Summary of Results

- EDA using SQL
- EDA using Data Visualization
- Creating an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Introduction

Background

SpaceX has emerged as the leading player in the commercial space age, revolutionizing space travel by making it more affordable. The company promotes its Falcon 9 rocket launches on its website, offering them at a competitive price of $62 million. In contrast, other providers charge upwards of $165 million per launch. This significant cost reduction is largely due to SpaceX's ability to reuse the first stage of its rockets.

To further enhance our understanding of launch costs, we aim to predict whether the first stage will successfully land and be reused. By leveraging public data and advanced machine learning models, we will analyze the factors that influence the likelihood of reusability, ultimately providing insights into the overall cost of a SpaceX launch.

Introduction

Questions

- How do variables such as payload mass, launch site, number of flights, and orbit types influence the success rate of first stage landings?
- Has the rate of successful first stage landings improved over the years, and if so, what factors contribute to this trend?
- Which algorithm is most effective for binary classification in predicting first stage landing success in this context?

Methodology

1. Data Collection
   - Gathered data from the SpaceX REST Application Programming Interface (API)
   - Conducted web scraping from Wikipedia for additional information
2. Data Wrangling
   - Filtered and cleaned the dataset
   - Addressed missing values appropriately
   - Applied one-hot encoding to prepare the data for binary classification
3. Exploratory Data Analysis (EDA)
   - Conducted EDA using SQL queries and visualization techniques to uncover insights

Methodology

4. Interactive Visual Analytics
    - Developed interactive visualizations using Folium and Plotly Dash for enhanced data exploration
5. Predictive Analysis
    - Built, tuned, and evaluated various classification models to optimize performance and ensure the best predictive accuracy

Data Collection

The data collection process was conducted using a combination of API requests from the SpaceX REST API and web scraping from a table in SpaceX's Wikipedia entry. Both methods were employed to gather comprehensive information about the launches, enabling a more detailed analysis.

The following data columns were obtained from the SpaceX REST API: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, and Latitude.

From the Wikipedia web scraping, the following data columns were extracted: Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, and Time

Data Collection – SpaceX API

1. Requested rocket launch data from the SpaceX API.
2. Decoded the response content using .json() and converted it into a DataFrame using pd.json_normalize().
3. Extracted necessary information about the launches by applying custom functions to the DataFrame.
4. Constructed a dictionary with the processed data.
5. Created a DataFrame from the constructed dictionary.
6. Filtered the DataFrame to include only Falcon 9 launches.
7. Replaced missing values in the Payload Mass column with the calculated mean for that column.
8. Exported the DataFrame to a CSV file.

Data Collection from API

Data Collection – Webscraping

1. Requested Falcon 9 launch data from Wikipedia.
2. Created a BeautifulSoup object from the HTML response.
3. Extracted all column names from the HTML table header.
4. Parsed the HTML tables to collect the data.
5. Constructed a dictionary from the collected data.
6. Created a DataFrame from the dictionary.
7. Exported the DataFrame to a CSV file.

Data Collection using Webscraping

Data Wrangling

In the dataset, several instances are noted where the booster did not land successfully. Some landings were attempted but failed due to various reasons. For example, "True Ocean" is used to indicate that the mission was successfully landed in a specific region of the ocean, while "False Ocean" is used to denote an unsuccessful landing in the ocean. Similarly, "True RTLS" is signified for a successful landing on a ground pad, whereas "False RTLS" is indicated for an unsuccessful landing on the ground pad. In the case of drone ship landings, "True ASDS" denotes a successful landing on a drone ship, while "False ASDS" signifies an unsuccessful landing on a drone ship.

To simplify the analysis, these outcomes are converted into training labels: "1" is assigned to represent a successful landing of the booster, while "0" is assigned to represent an unsuccessful landing.

Data Wrangling

1. Performed exploratory data analysis and determined training labels.
2. Calculated the number of launches at each site.
3. Calculated the number and occurrence of each orbit type.
4. Calculated the number and occurrence of mission outcomes for each orbit type.
5. Created a landing outcome label from the Outcome column.
6. Exported the data to a CSV file

Data Wrangling

EDA with SQL

Performed SQL queries:

1. Displayed the names of the unique launch sites in the space mission.
2. Displayed 5 records where launch sites began with the string ‘CCA’.
3. Displayed the total payload mass carried by boosters launched by NASA (CRS).
4. Displayed the average payload mass carried by booster version F9 v1.1.
5. Listed the date when the first successful landing outcome on the ground pad was achieved.
6. Listed the names of the boosters that succeeded in landing on a drone ship and had a payload mass greater than 4000 but less than 6000.
7. Listed the total number of successful and failed mission outcomes.

EDA with SQL

8. Listed the names of the booster versions that carried the maximum payload mass.
9. Listed the failed landing outcomes on drone ships, along with their booster versions and launch site names for the months in the year 2015.
10. Ranked the count of landing outcomes (such as Failure on drone ship or Success on ground pad) between the dates 2010-06-04 and 2017-03-20 in descending order.

EDA using Data Visualization

1. Charts were plotted for the following comparisons: 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.
2. Scatter plots were used to illustrate the relationships between variables. If a relationship existed, these plots could be utilized in a machine learning model.
3. Bar charts were created to show comparisons among discrete categories, with the goal of highlighting the relationships between the specific categories being compared and a measured value.
4. Line charts were employed to depict trends in data over time (time series).

Creating Interactive Visuals using Folium

1. Added markers for all launch sites
   - marker with a circle, popup label, and text label for NASA Johnson Space Center using its latitude and longitude coordinates as a starting location.
   - markers with circles, popup labels, and text labels for all launch sites using their latitude and longitude coordinates to show their geographical locations and proximity to the equator and coasts.
2. Added colored markers for the launch outcomes at each launch site
   - colored markers for successful launches (green) and failed launches (red) using marker clusters to identify which launch sites had relatively high success rates.
3. Included distances between a launch site and its proximities
   - colored lines to show distances from the launch site CCA FS SLC-40 (as an example) to the coastline, railway, highway, and closest city.

Interactive Visuals using Folium

Building a Dashboard with Plotly Dash

1. Generated Launch Site dropdown list
   - list to enable launch site selection
2. Generated a pie chart showing successful launches (all sites/certain site)
   - pie chart to display the total count of successful launches for all sites and the success vs. failed counts for a specific launch site, if selected
3. Generated a slider of payload mass range
   - slider to select the payload mass range
4. Generated scatter chart of payload mass vs. success rate for the different booster versions
   - scatter chart to show the correlation between payload mass and launch success rate

Predictive Analysis (Classification)

1. Created a NumPy array from the column "Class" in the data.
2. Standardized the data using StandardScaler, then fitted and transformed it.
3. Split the data into training and testing sets using the train_test_split function.
4. Created a GridSearchCV object with cv = 10 to find the best parameters.
5. Applied GridSearchCV to the LogReg, SVM, Decision Tree, and KNN models.
6. Calculated the accuracy on the test data using the .score() method for all models.
7. Examined the confusion matrix for all models.
8. Determined the method that performed best by examining the Jaccard score and F1 score metrics.

Results

- EDA using SQL
- EDA using Data Visualization
- Creating an Interactive Visuals using Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

EDA using SQL

Names of all launch sites from the dataset

Task 1

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

First 5 records where launch site starts with “CCA”

Task 2

Display 5 records where launch sites begin with the string '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

Total payload mass carried by boosters launched by NASA (CRS)

Task 3

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

```
%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

Average payload mass by booster F9 v1.1

Task 4

Display average payload mass carried by booster version 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

Date of first successful landing in ground pad

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

Boosters that were successful in drone ship landing with payload between 4000 and 6000

Task 6

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

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
```

```
 * 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 failed missions

Task 7

List the total number of successful and failure mission outcomes

```
%sql select distinct 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	99
Success (payload status unclear)	1

Booster versions that carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%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

Details regarding launches in 2015 that failed landing in drone ship

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql select substr(date, 6, 2) as month, date, booster_version, launch_site, landing_outcome
from SPACEXTABLE
where landing_outcome = 'Failure (drone ship)'
 and substr(date, 1, 4) = '2015';
```

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

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)

Landing outcomes between 2010-06-04 and 2017-03-20 ranked in descending order by count

Task 10

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.

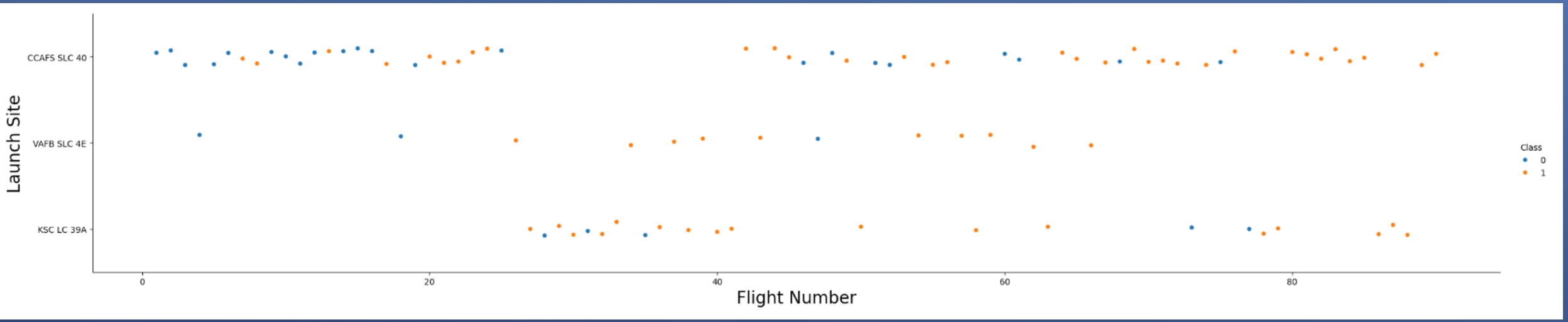
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

```
 * 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

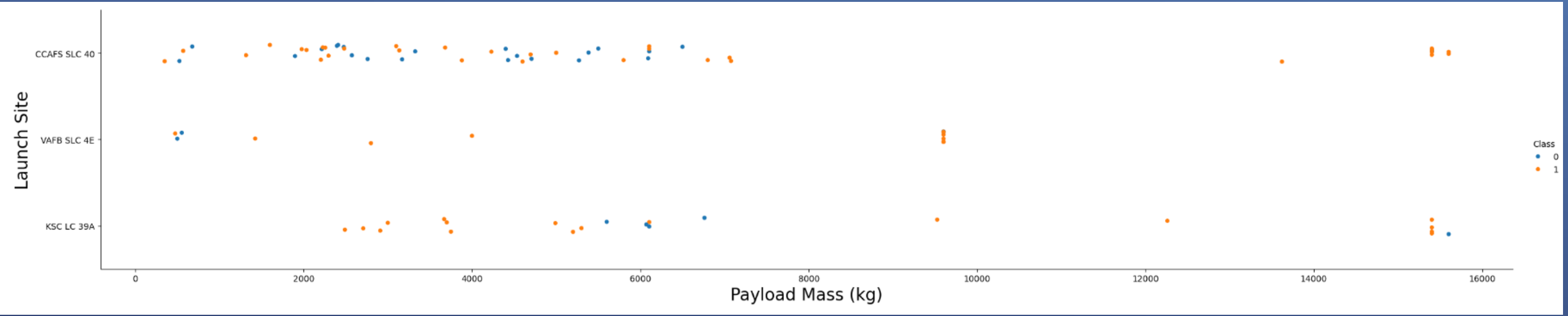
EDA using Visualization

Flight Number vs. Launch Site



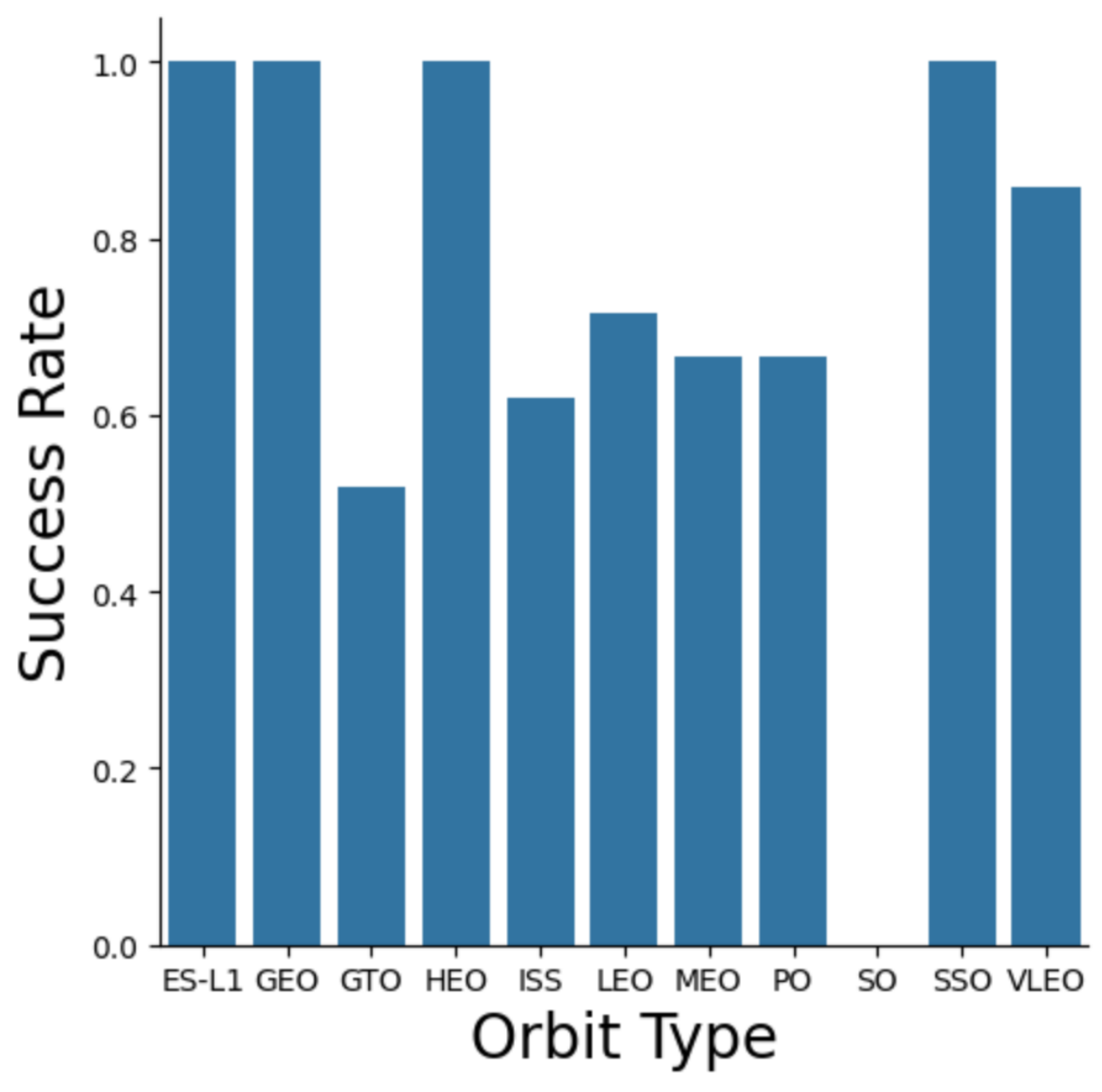
- Most of the earlier flights failed while most of the later flights succeeded.
- The CCAFS SLC 40 site is where majority of all flights were launched.
- VAFB SLC 4E and KSC LC 39A have higher success rates compared to CCAFS SLC 40.

Payload Mass vs. Launch Site



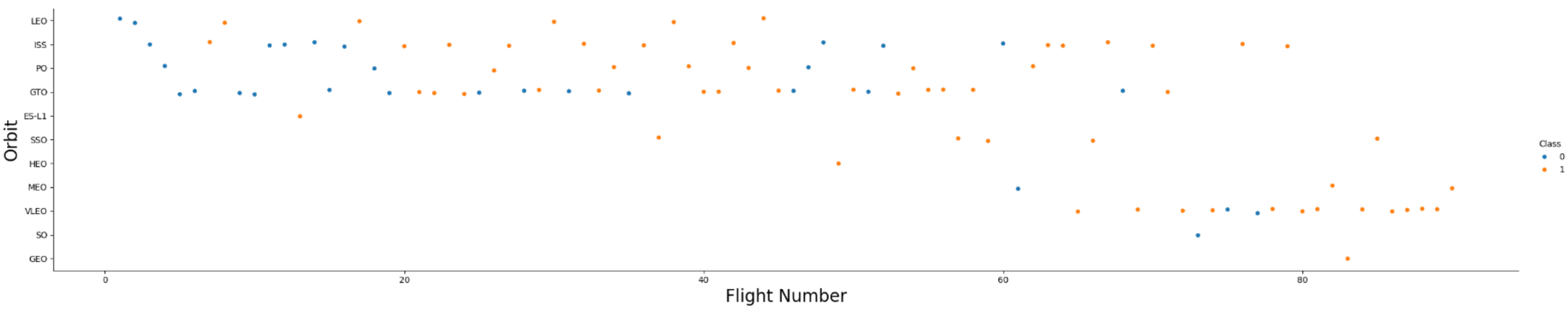
- In general, a higher payload mass leads to better chances of success.
- Most of the flights ccarrying a payload mass above 7000 kg were successfull.
- KSC LC 39A has a 100% success rate for payload mass below 5500 kg.

Success rates of different orbit types



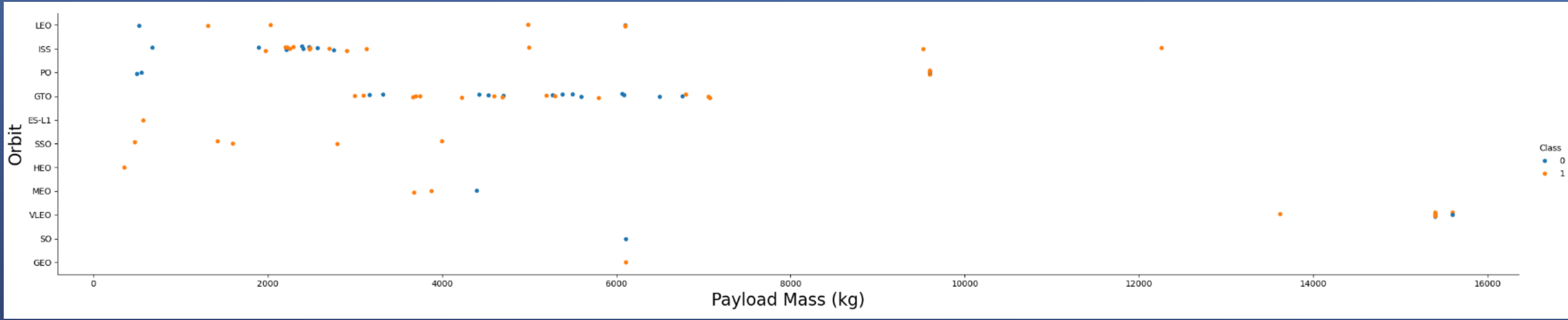
- ES-L1, GEO, HEO, and SSO orbit types had 100% success rates
- GTO, ISS, LEO, MEO, and PO orbit types had a success rate between 50% and 85%.
- SO orbit type had a 0% success rate.

Flight Number vs. Orbit Type



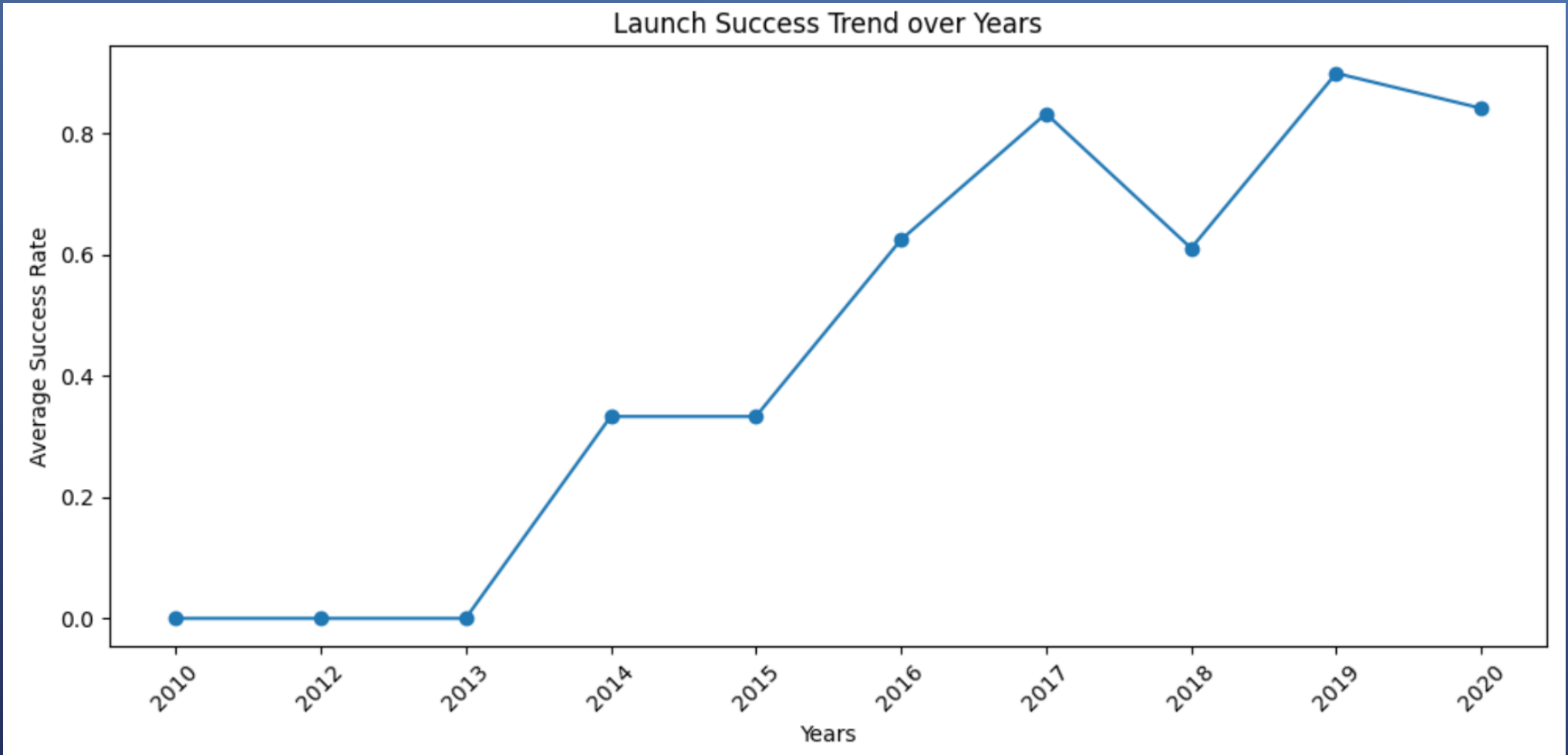
- In the LEO orbit type, success appears to be related to the number of flights
- There seems to be no relationship between the number of flights and success in the GTO orbit.

Flight Number vs. Orbit Type



- For PO, LEO and ISS orbit types, heavier payloads seems to increase the rates of successful landing.
- For GTO, it is difficult to find a trend because successful and unsuccessful landings randomly occur regardless of payload mass

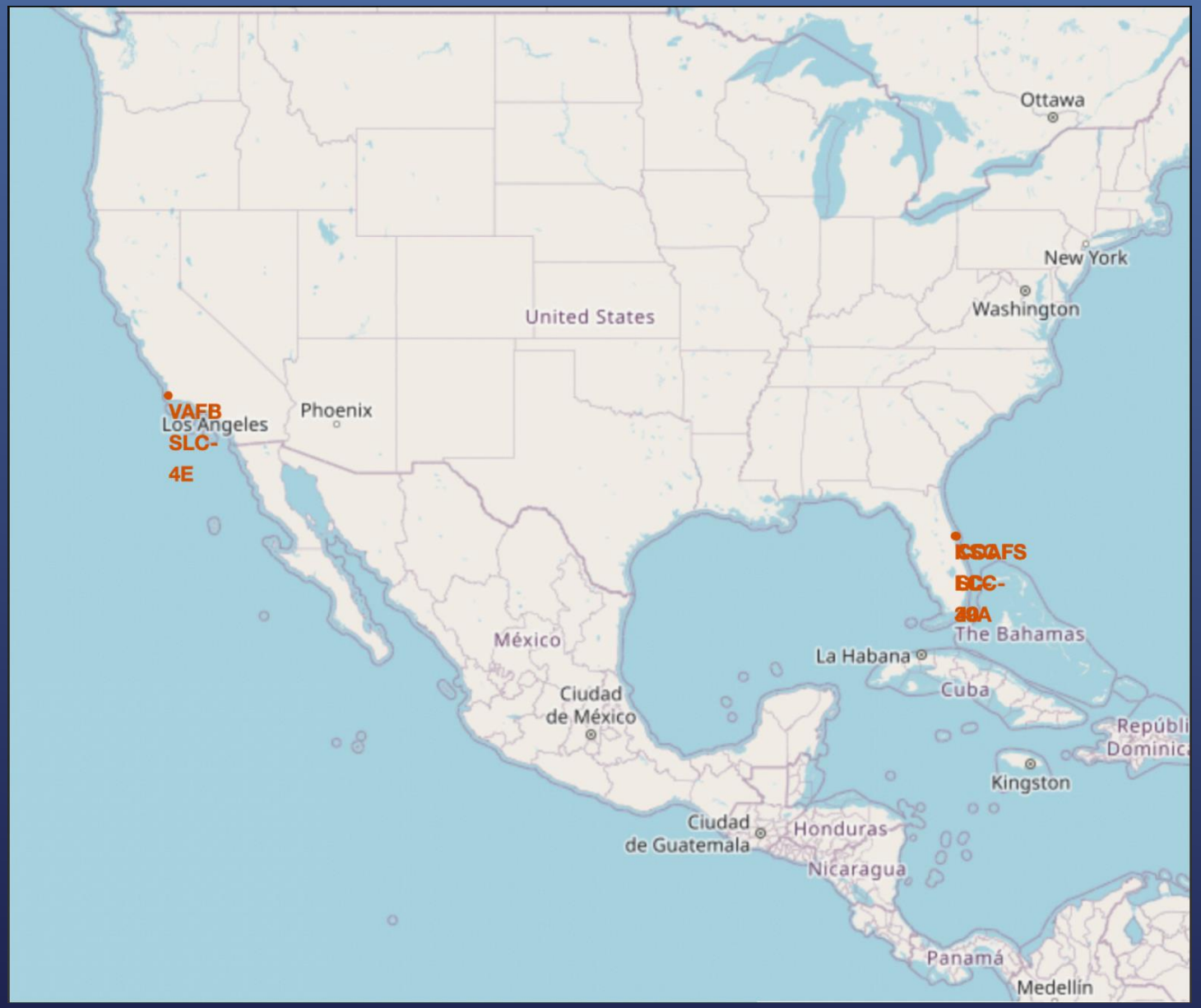
Launch success trend over years



- Launch success rates have generally been improving from 2013 to 2020

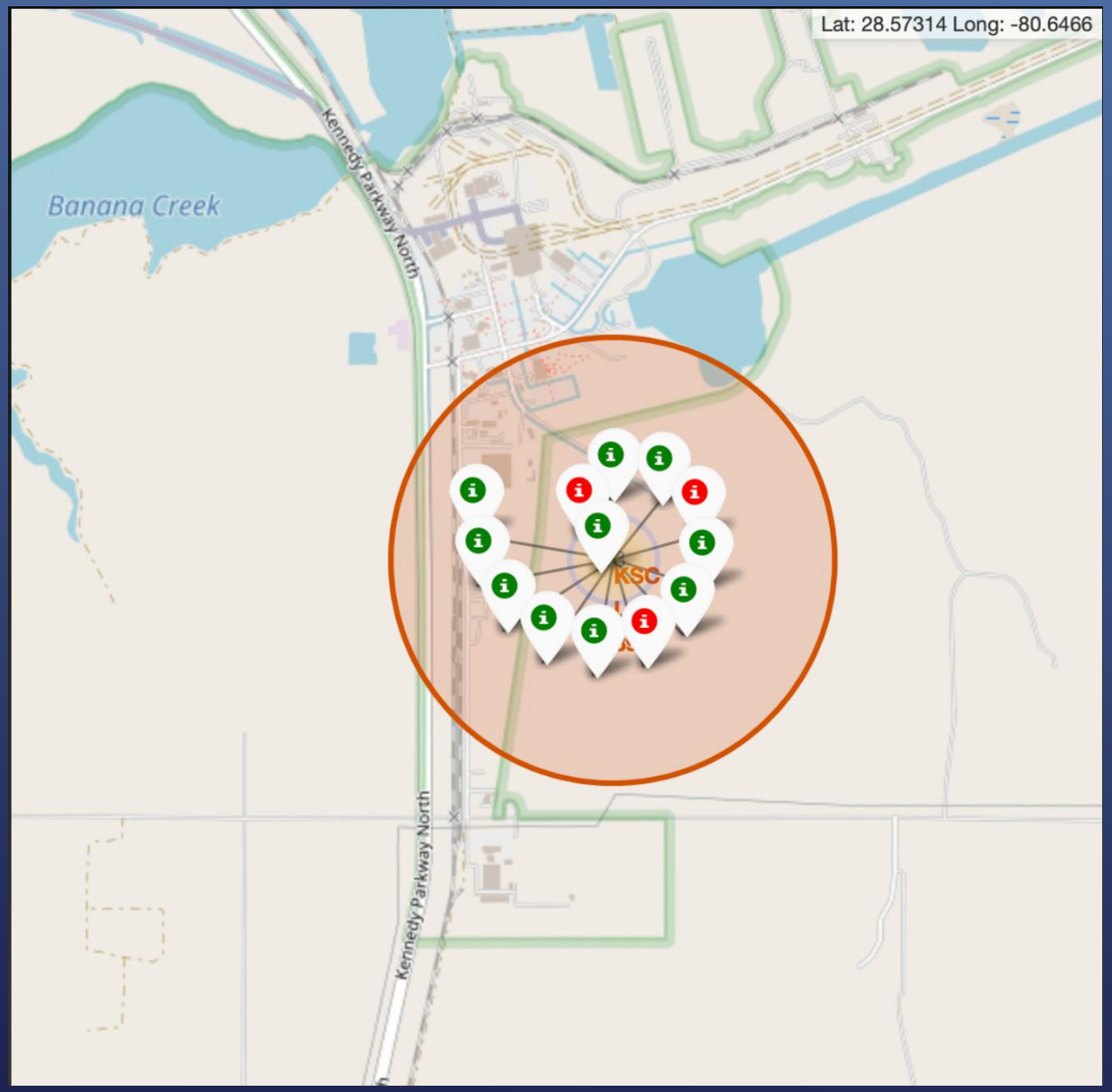
Creating an Interactive Visuals using Folium

Launch site locations in the US



- Most launch sites are located near the equator. The Earth's surface at the equator moves faster than at any other point, at a speed of approximately 1670 km/hour. When a spacecraft is launched from the equator, it carries this initial velocity into space due to inertia, helping it achieve the necessary speed to stay in orbit
- Additionally, all launch sites are positioned close to the coast. Launching rockets over the ocean minimizes the risk of debris falling or explosions occurring near populated areas.

Successful and failed launches in KSC LC-39A



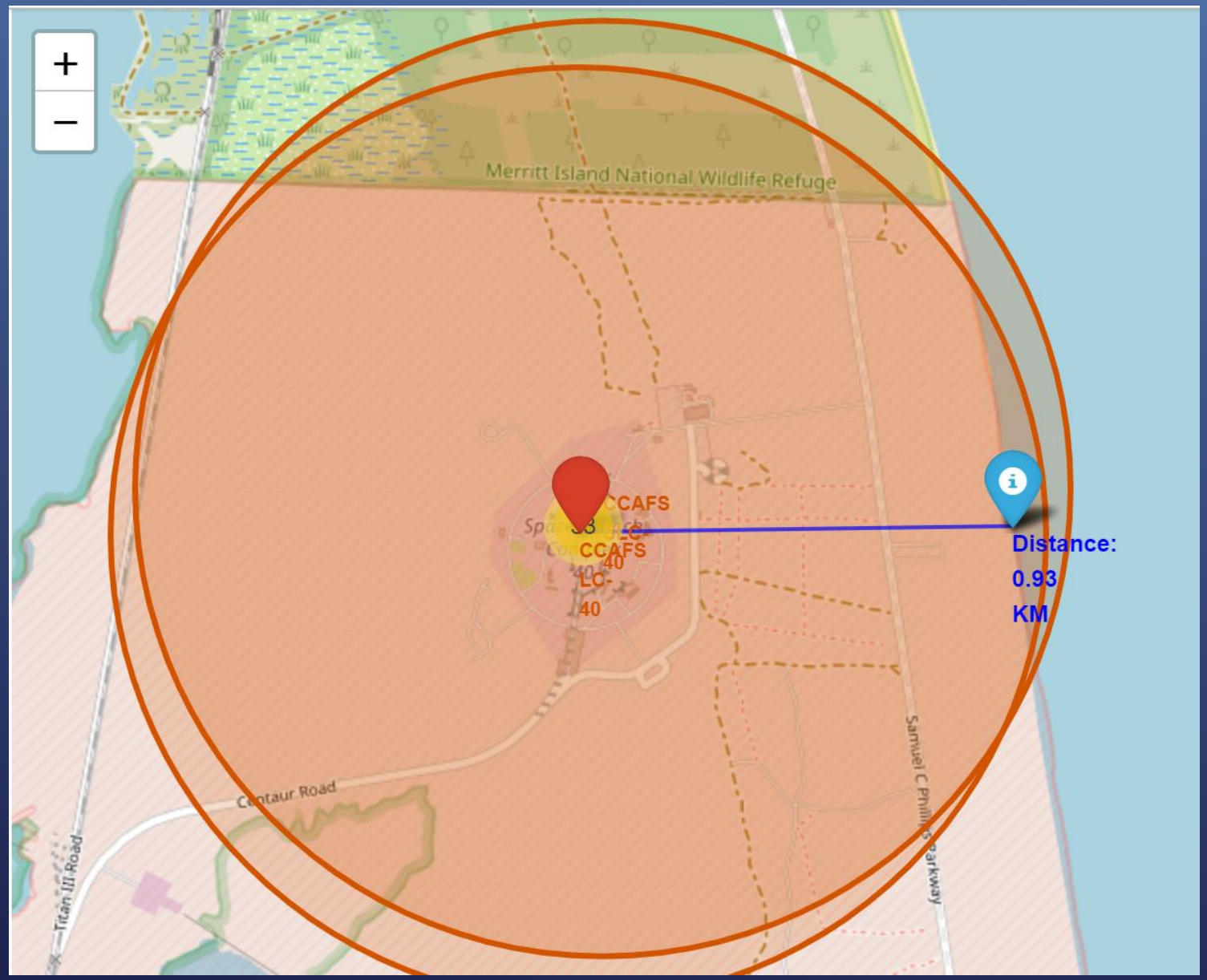
- The color-coded markers make it easy to identify which launch sites have relatively high success rates
- The KSC LC-39A launch site has a very high success rate for launches

LEGEND

Green marker: Successful launch

Red marker: Failed launch

Distance between CCAFS LC-40 and the coastline



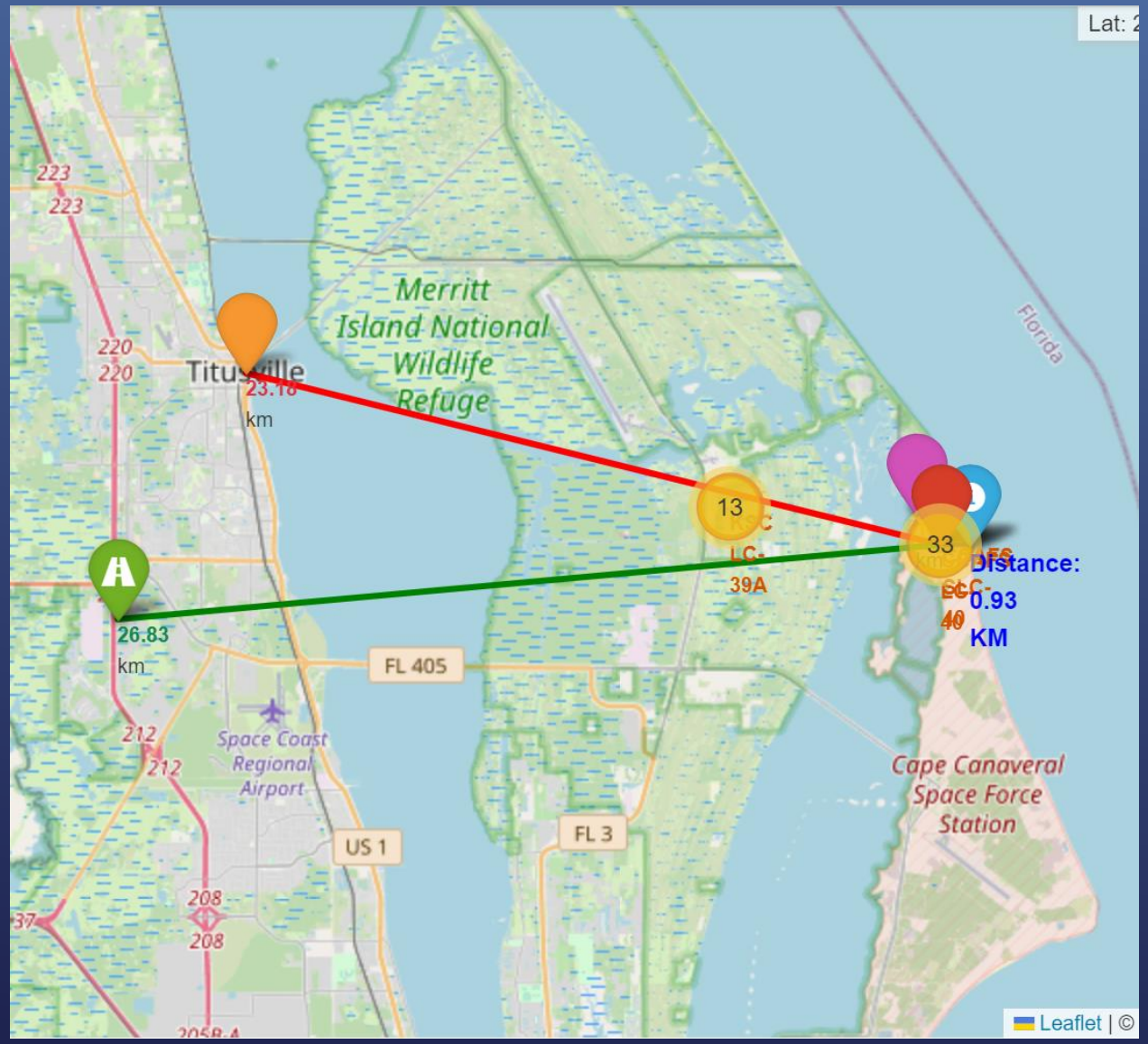
LEGEND

Red marker: CCAFS SLC-40

Blue marker: Coastline

Blue line: Distance (0.93 km)

Distance between CCAFS LC-40 and the nearest coastline, highway, city, and railroad



LEGEND

Red marker: CCAFS SLC-40

Blue marker: Coastline

Green marker: Highway

Orange marker: City

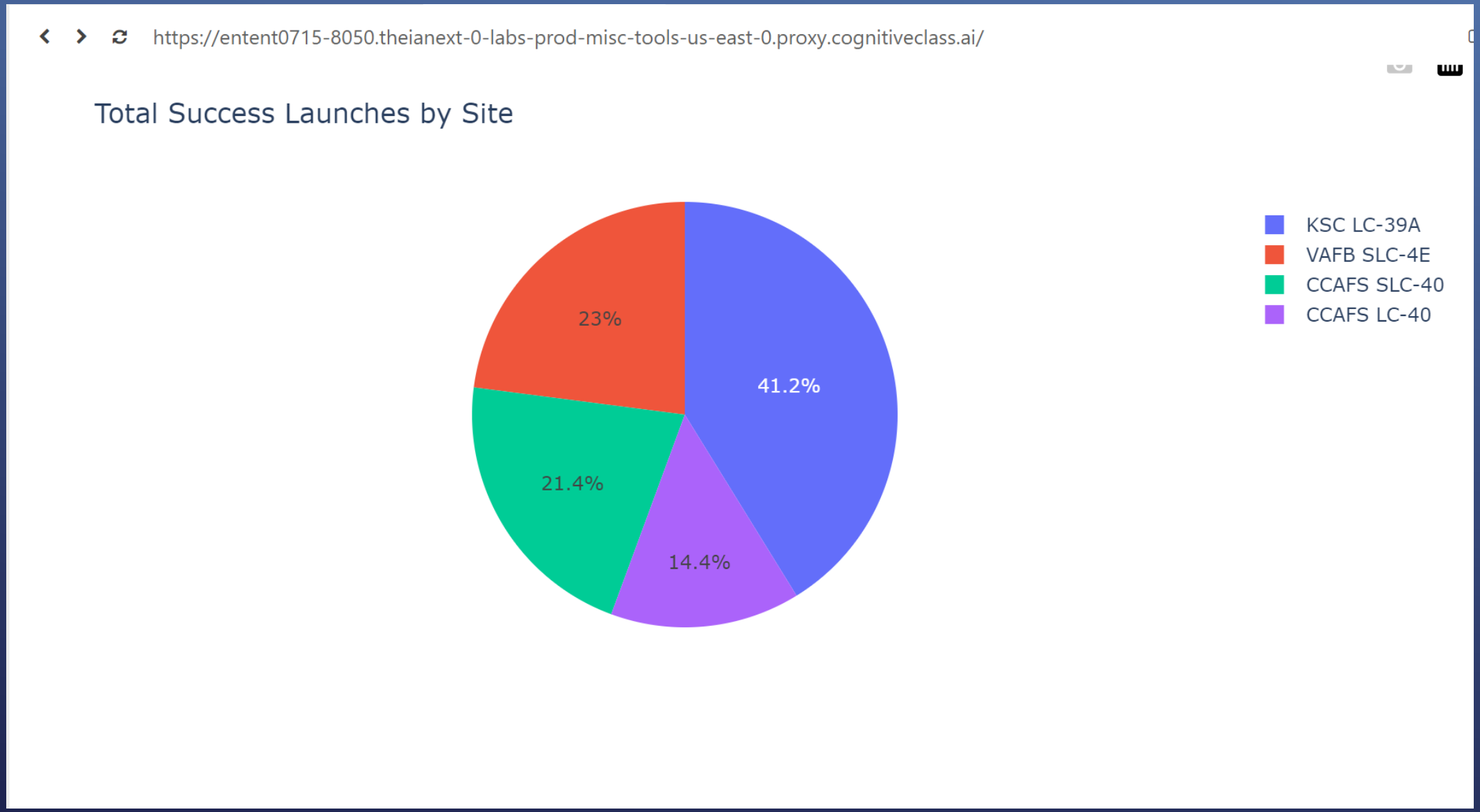
Blue line: 0.93 km

Green line: 26.83 km

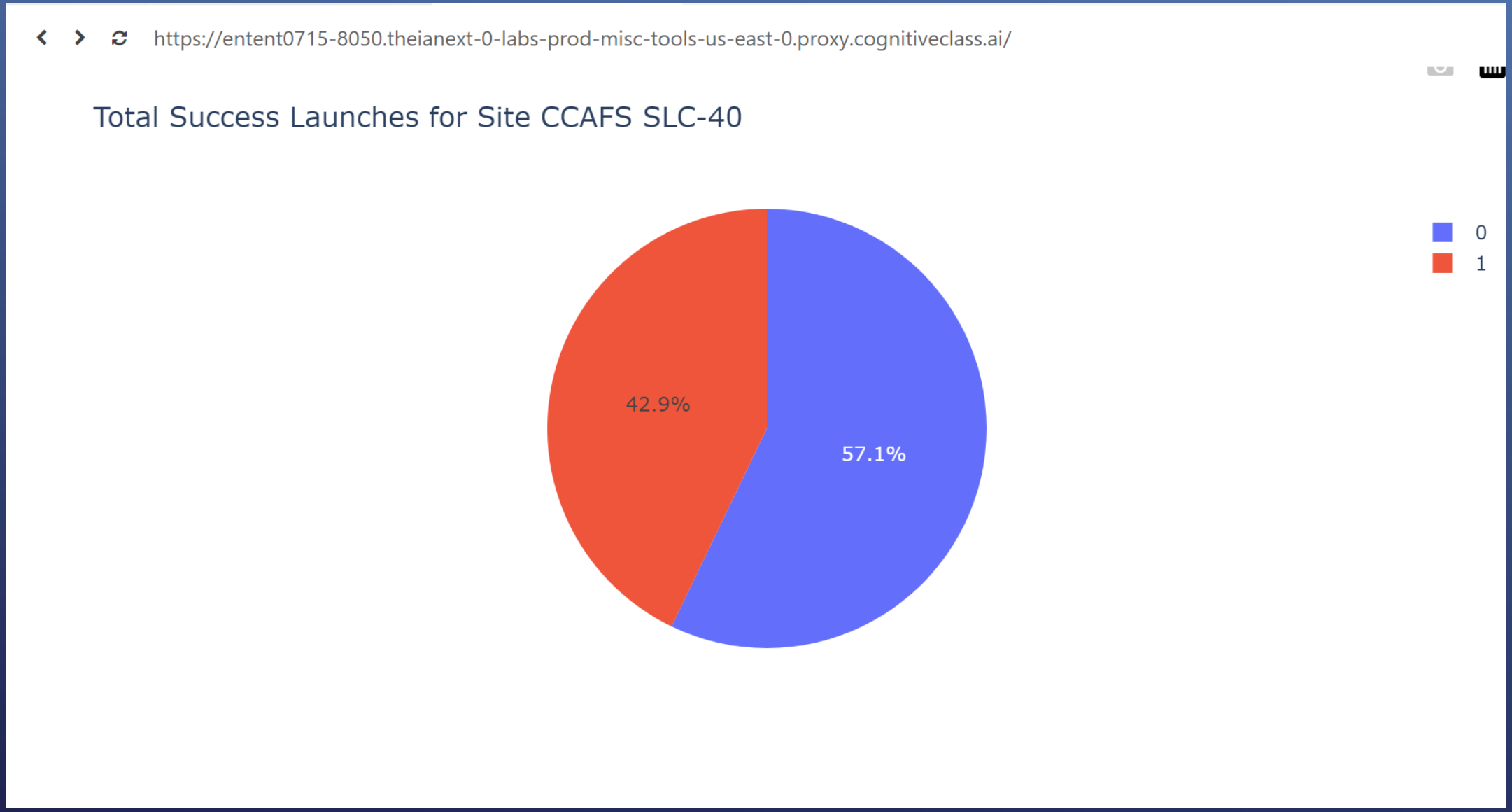
Red line: 23.18 km

Building a Dashboard with Plotly Dash

KSC LC-39A is the Site with largest successful launches



CCAFS SLC-40 has the highest launch success rate



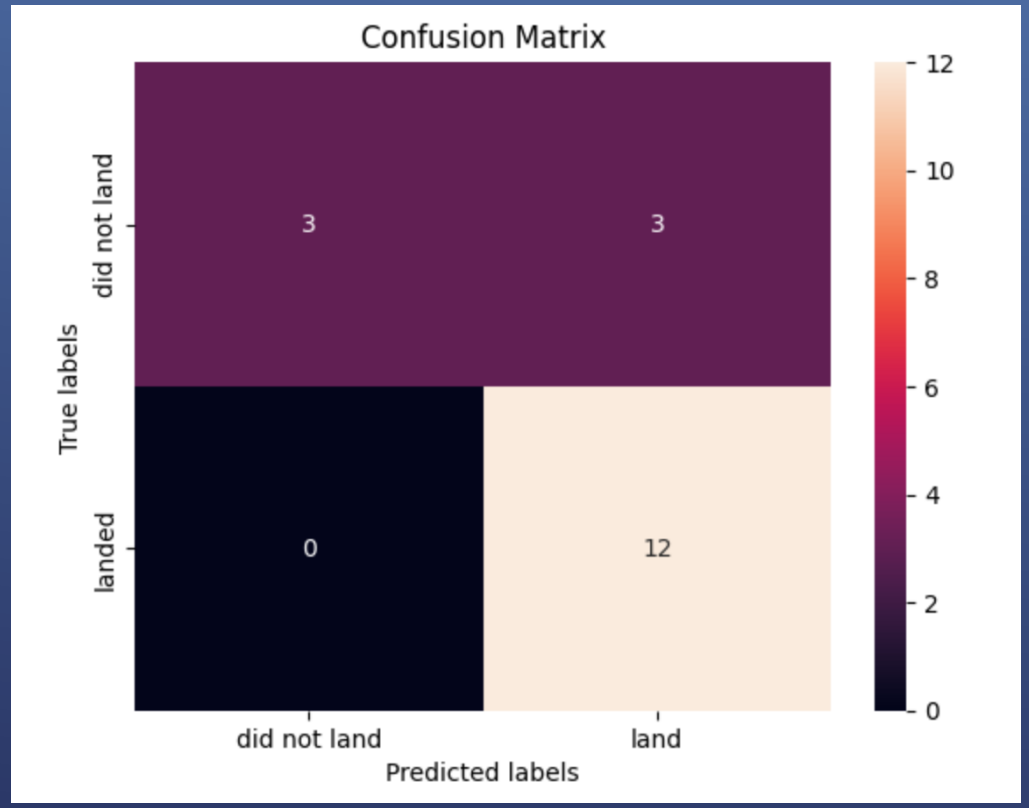
The payload range from 3000-4000 has the highest launch success rate (70%)

The payload range from 6000 to 7000 has the lowest launch success rate (0%)

B5 booster has the highest launch success rate (100%)

Predictive Analysis (Classification)

Confusion Matrix (same for all models tested)



- True Negatives: 3
- False Positives: 3
- False Negatives: 0
- True Positives: 12

- The model shows a moderate performance in predicting the "landed" class, with 3 false positives.
- It has a strong performance in predicting the “did not land” class, with 0 false negatives.
- Overall, the matrix suggests that the model is generally reliable but could benefit from improved identification of successful landings.

Jaccard Index, F1-score, and Accuracy scores

Test Set Only

	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	0.833333	0.833333

All Data

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

- Based on the scores of the test set, it could not be determined which method performed best. The identical test set scores and accuracy may have been due to the small sample size (18 samples).
- All methods were therefore tested on the entire dataset. The scores of the full dataset identified the SVM model as the best, as it not only had highest scores but also the highest accuracy.

Conclusion

- Most launch sites were located in proximity to the Equator line, and all the sites were in very close proximity to the coast.
- The success rate of launches increased over the years, with KSC LC-39A having the highest success rate among all launch sites.
- Launches with a low payload mass showed better results than launches with a larger payload mass.
- Orbits ES-L1, GEO, HEO, and SSO had a 100% success rate.
- The SVM model was identified as the best algorithm for predicting first stage landing success

Acknowledgement

I extend my sincere gratitude to Coursera and IBM for their resources, to the instructors for their guidance, and to my peers for their support. Your contributions greatly enriched my learning experience. Thank you!