Data Science Capstone Project

CALVIN MLANGENI

01.03.2024

Outline

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 results
- Interactive analytics demo in screenshots
- Predictive analysis results

Executive Summary

PROJECT BACKGROUND AND CONTEXT

SpaceX stands at the forefront of the commercial space era, revolutionizing space exploration by making it more cost-effective. The company promotes its Falcon 9 rocket launches on its website for 62 million dollars, significantly lower than the 165 million dollars charged by other providers. This cost-efficiency largely stems from SpaceX's ability to recycle the first stage of the rocket. By predicting the likelihood of the first stage's successful landing, we can estimate the overall cost of a launch. Utilizing publicly available data and machine learning techniques, we aim to forecast SpaceX's reusability of the first stage.

Key inquiries include:
- How do factors such as the mass of the payload, launch site, flight frequency, and orbital paths influence the successful landing of the first stage?
- Has there been an improvement in the rate of successful first stage landings over time?
- Which algorithm is most suitable for binary classification in predicting the first stage's reusability?

Introduction

Methodology

Data Gathering Approach:

Accessed data through the SpaceX REST API.

Employed web scraping techniques to extract information from Wikipedia.

Executed data cleansing operations, including:
- Refining the dataset.
- Addressing missing data entries.

Prepared the dataset for binary classification by applying One Hot Encoding.

Conducted exploratory data analysis (EDA) through both visualization techniques and SQL queries.

Utilized interactive visualization tools such as Folium and Plotly Dash for visual analytics.

Implemented classification models for predictive analysis
- Building, tuning and evaluation of classification models to ensure the best results

Methodology

Data collection

Data Columns are obtained by using SpaceX REST API:

Data Columns are obtained by using Wikipedia Web Scraping:

Data collection – SpaceX API

1. Requesting rocket launch data from SpaceX API
2. Decoding the response content using `.json()` and turning it into a dataframe using `.json_normalize()`
3. Requesting needed information about the launches from SpaceX API by applying custom functions
4. Constructing data we have obtained into a dictionary
5. Creating a dataframe from the dictionary
6. Filtering the dataframe to only include Falcon 9 launches
7. Replacing missing values of Payload Mass column with calculated .mean() for this column
8. Exporting the data to CSV

[GitHub URL: Data Collection API]()

Data collection – Web scraping

Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV

[GitHub URL: Data Collection with Web Scraping]

Data wrangling

For our analysis, we transform these outcomes into training labels, where a "1" signifies a successful booster landing, and a "0" indicates an unsuccessful attempt.

GitHub URL for Data Preparation:

Within the dataset, various instances depict unsuccessful booster landings. In some cases, a landing attempt was made but resulted in failure due to mishaps. For instance, "True Ocean" indicates that the mission successfully achieved a landing in a designated ocean area, whereas "False Ocean" signifies that the mission failed to land in the intended ocean region. Similarly, "True RTLS" (Return to Launch Site) denotes a successful return and landing at the launch site, while "False RTLS" indicates a failure to land back at the launch site. Furthermore, "True ASDS" (Autonomous Spaceport Drone Ship) represents a successful landing on a drone ship, in contrast to "False ASDS," which implies a failed attempt to land on a drone ship.

Perform exploratory Data Analysis and determine Training Labels

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 from Outcome column

Exporting the data to CSV

EDA with data visualization

CHARTS WERE PLOTTED:

GitHub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

[GitHub URL: EDA with SQL]

Build an interactive map with Folium

MARKERS OF ALL LAUNCH SITES:

Coloured Markers of the launch outcomes for each Launch Site:

Distances between a Launch Site to its proximities:

GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

LAUNCH SITES DROPDOWN LIST:

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

Slider of Payload Mass Range:

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

[GitHub URL: SpaceX Dash App]

Predictive analysis (Classification)

Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

[GitHub URL: Machine Learning Prediction]

Results

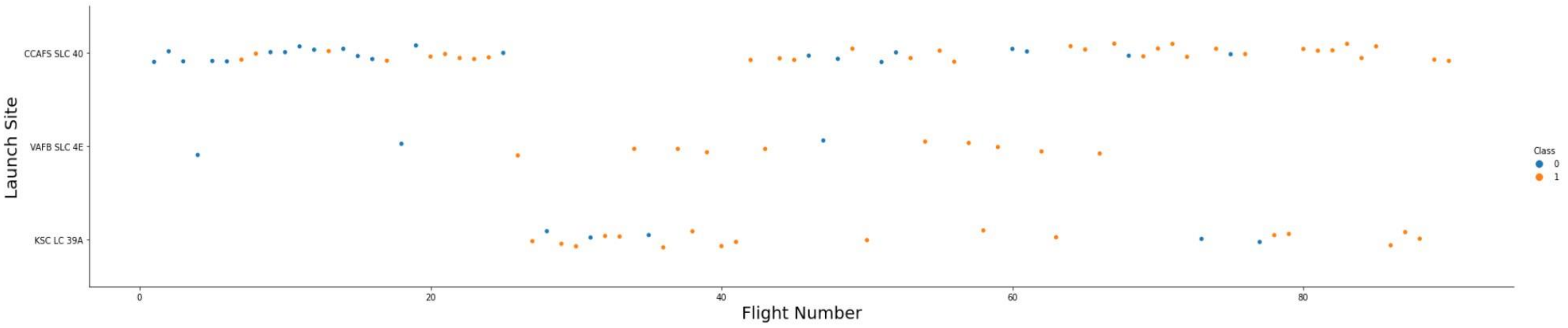
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

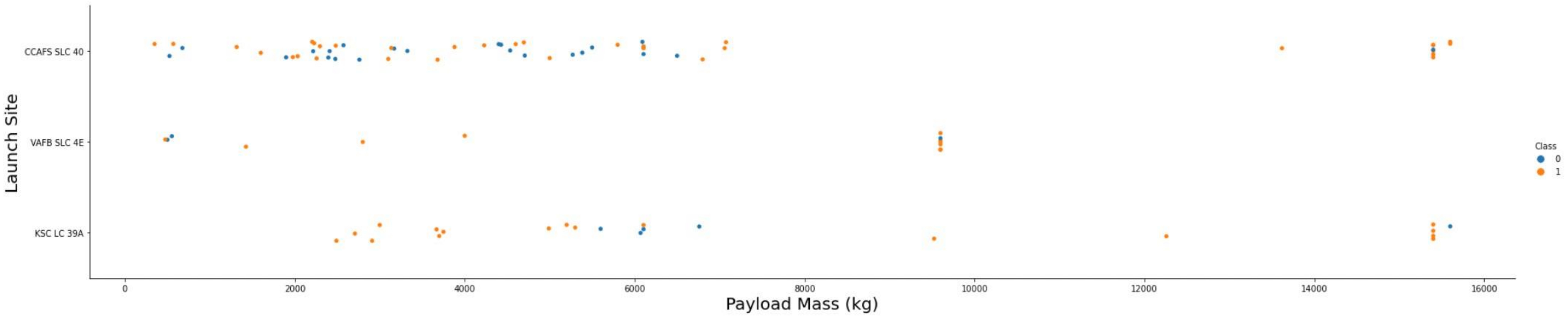
EDA with Visualization

Flight Number vs. Launch Site



Explanation: CCAPS SLC 40 at has the most flight attempts

Payload vs. Launch Site



Explanation: Site CCAPS SLC 40 has launched a variety of payloads with a certain payload being preferred towards the end of the launch data.

Success rate vs. Orbit type

EXPLANATION:

Orbits with 100% success rate:

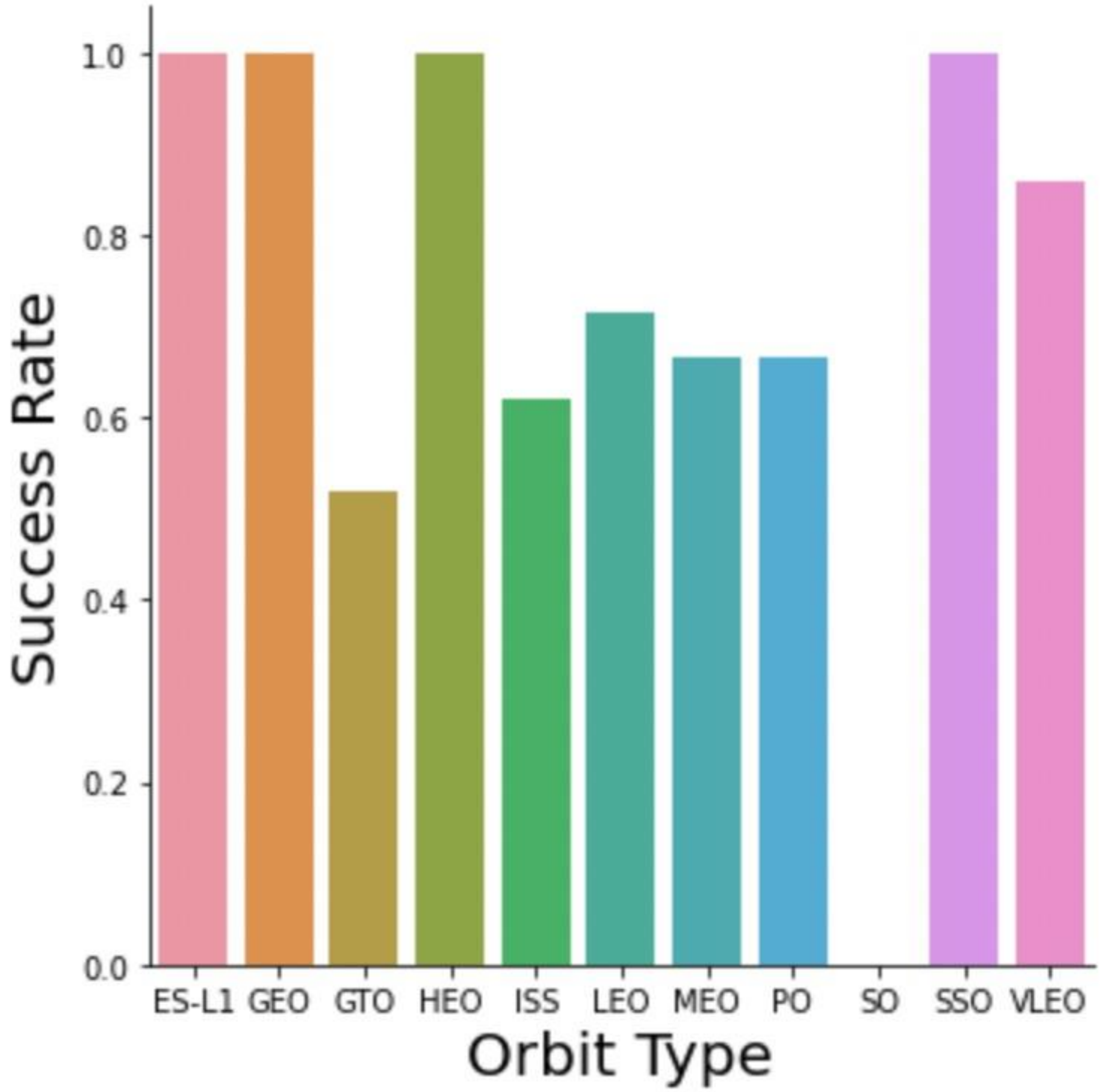
- ES-L1, GEO, HEO, SSO

Orbits with success rate between 50% and 85%:

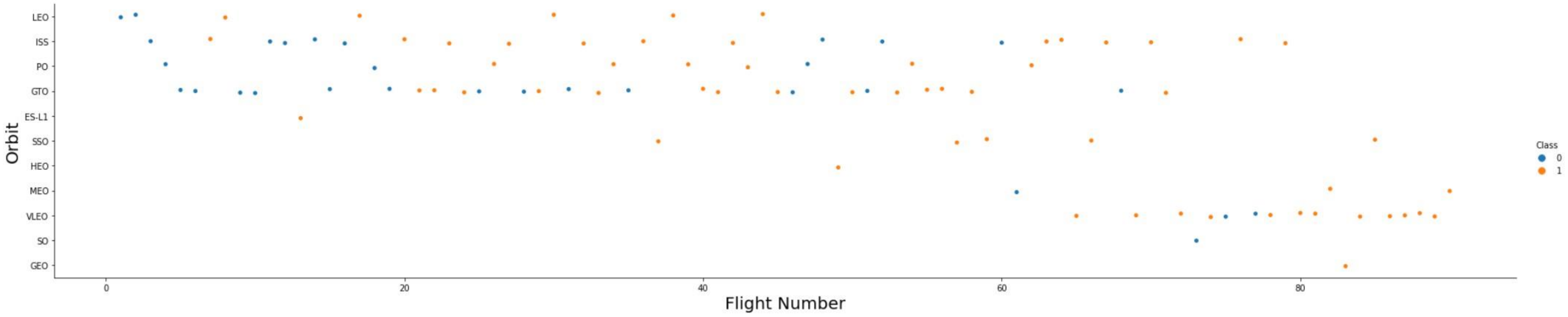
- GTO, ISS, LEO, MEO, PO

Orbits with 0% success rate:

- SO

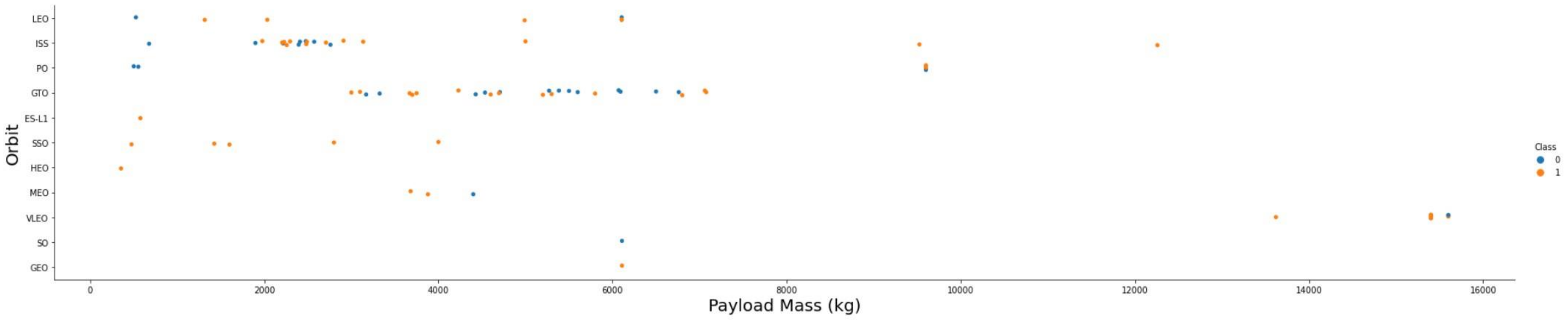


Flight Number vs. Orbit type



Explanation: VLEO orbit has the most number of success in launches with GEO having the least.

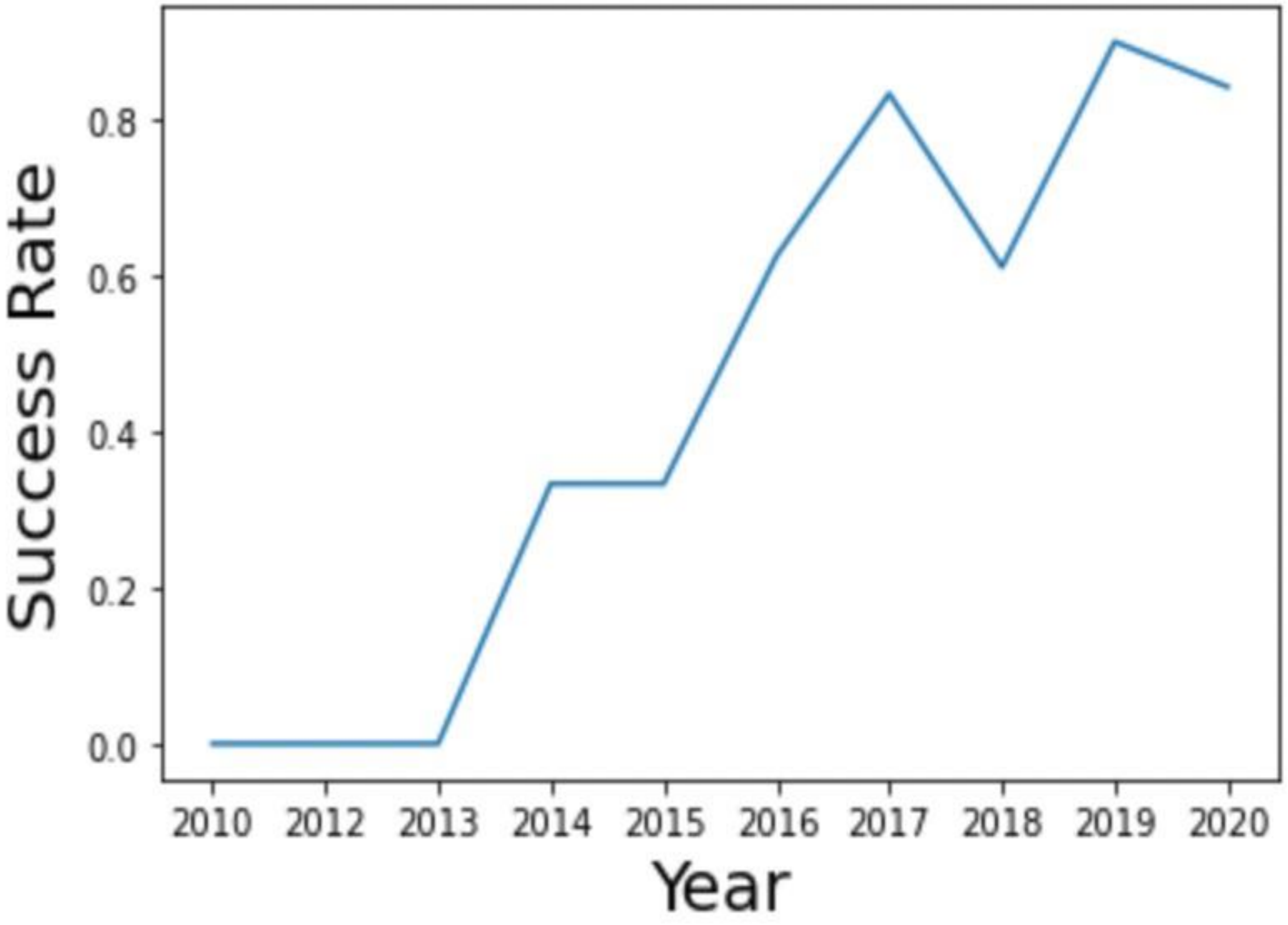
Payload Mass vs. Orbit type



Explanation: Above average payloads tend to travel to further orbits.

Launch success yearly trend

EXPLANATION: LAUNCHES ARE GETTING BETTER WITH TIME!!



EDA with SQL

All launch site names

In [4]:
```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Explanation: Only four unique launch sites

Launch site names begin with `CCA`

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

In [6]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Explanation: The total mass is 45596kg

Average payload mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Explanation: The average payload 2534kg

First successful ground landing date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[8]:
```

first_successful_landing
2015-12-22

Explanation: The first successful landing was on the 2015-12-22

Successful drone ship landing with payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Explanation: There are numerous booster versions that can achieve success

Total number of successful and failure mission outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation: There is more success than failure

Boosters carried maximum payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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

Explanation: A variety of boosters carried payload.

2015 launch records

In [12]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank success count between 2010-06-04 and 2017-03-20

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

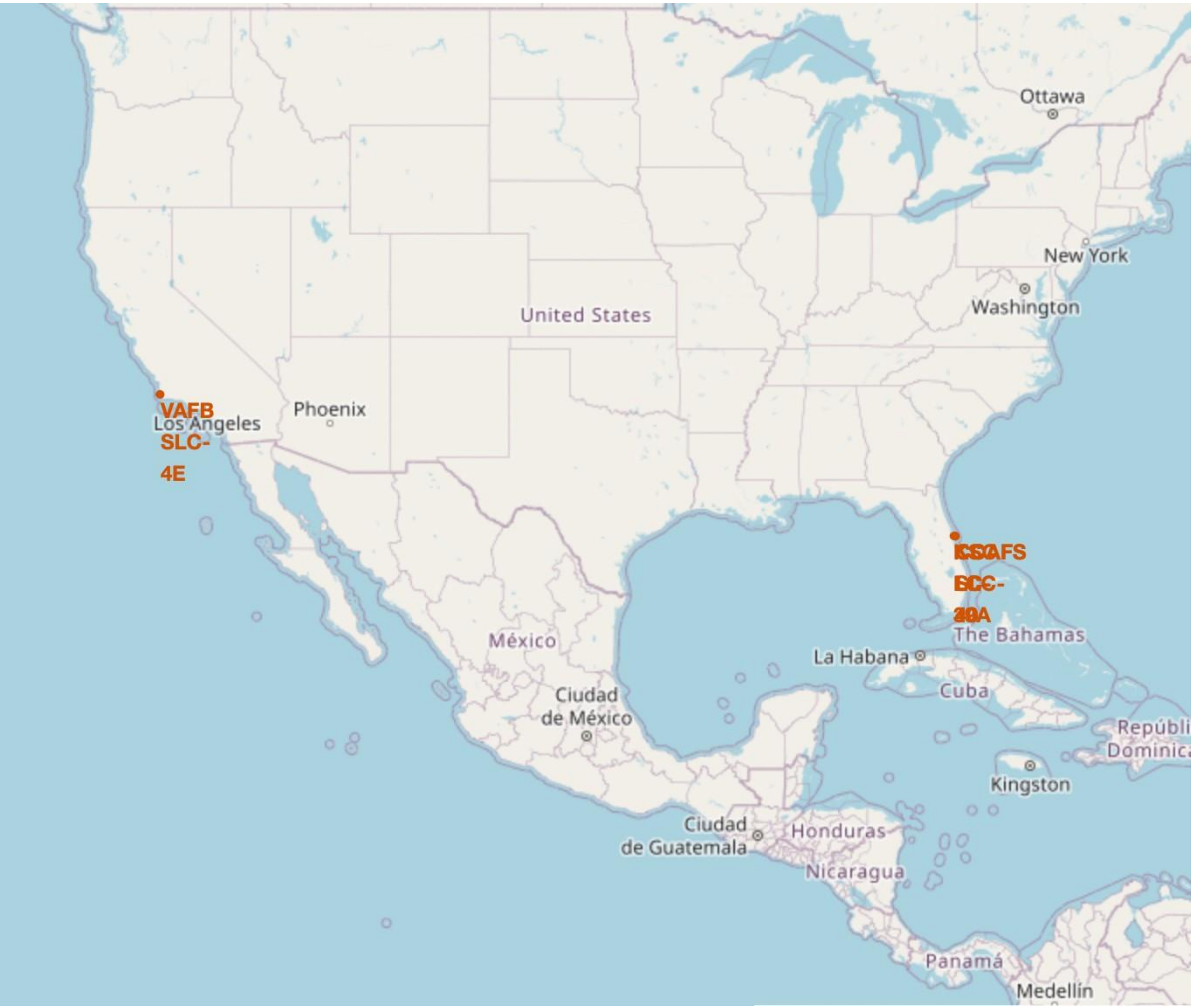
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:

Interactive map with Folium

All launch sites' location markers on a global map

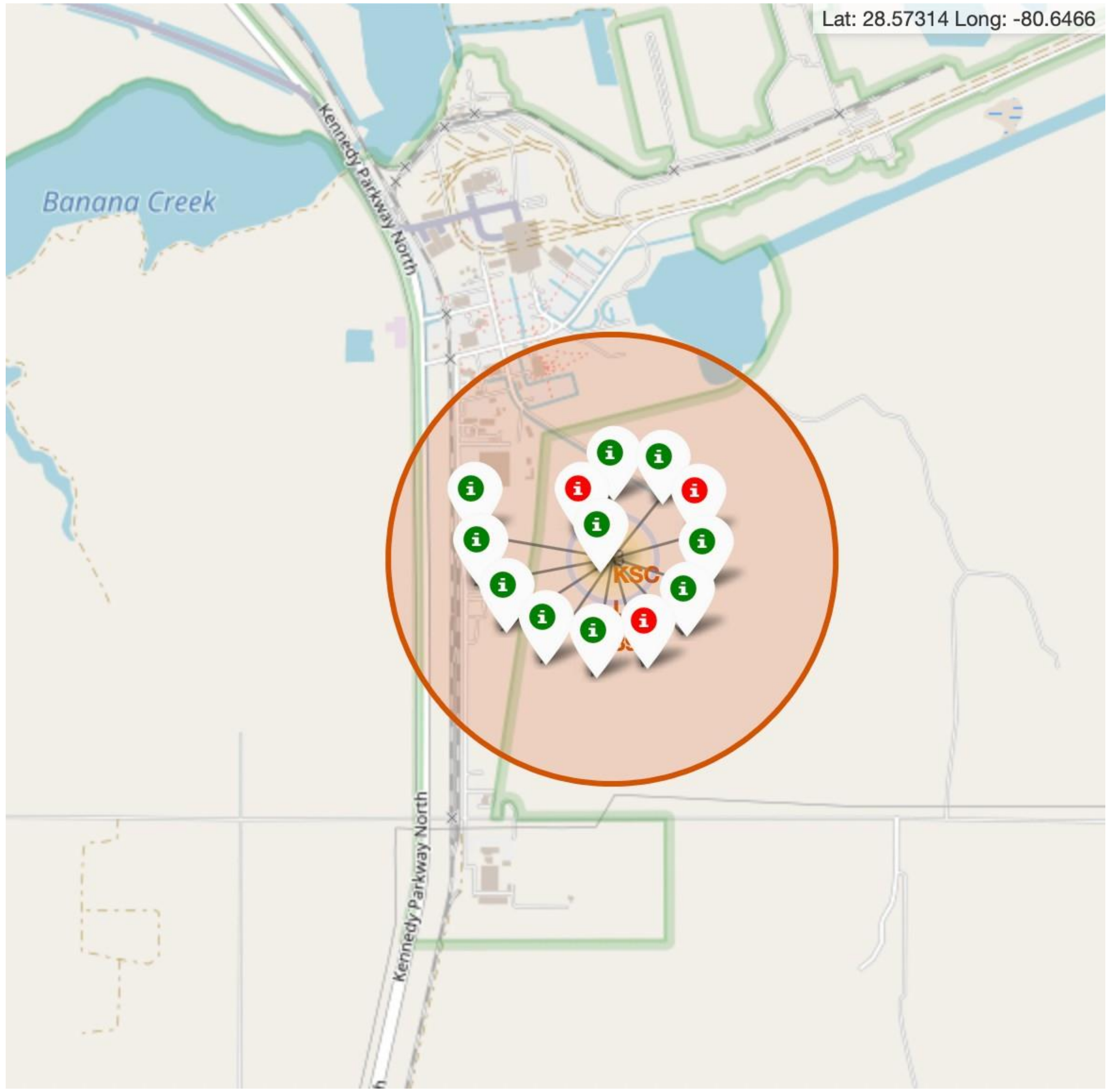
Explanation: Various locations of launches conducted.



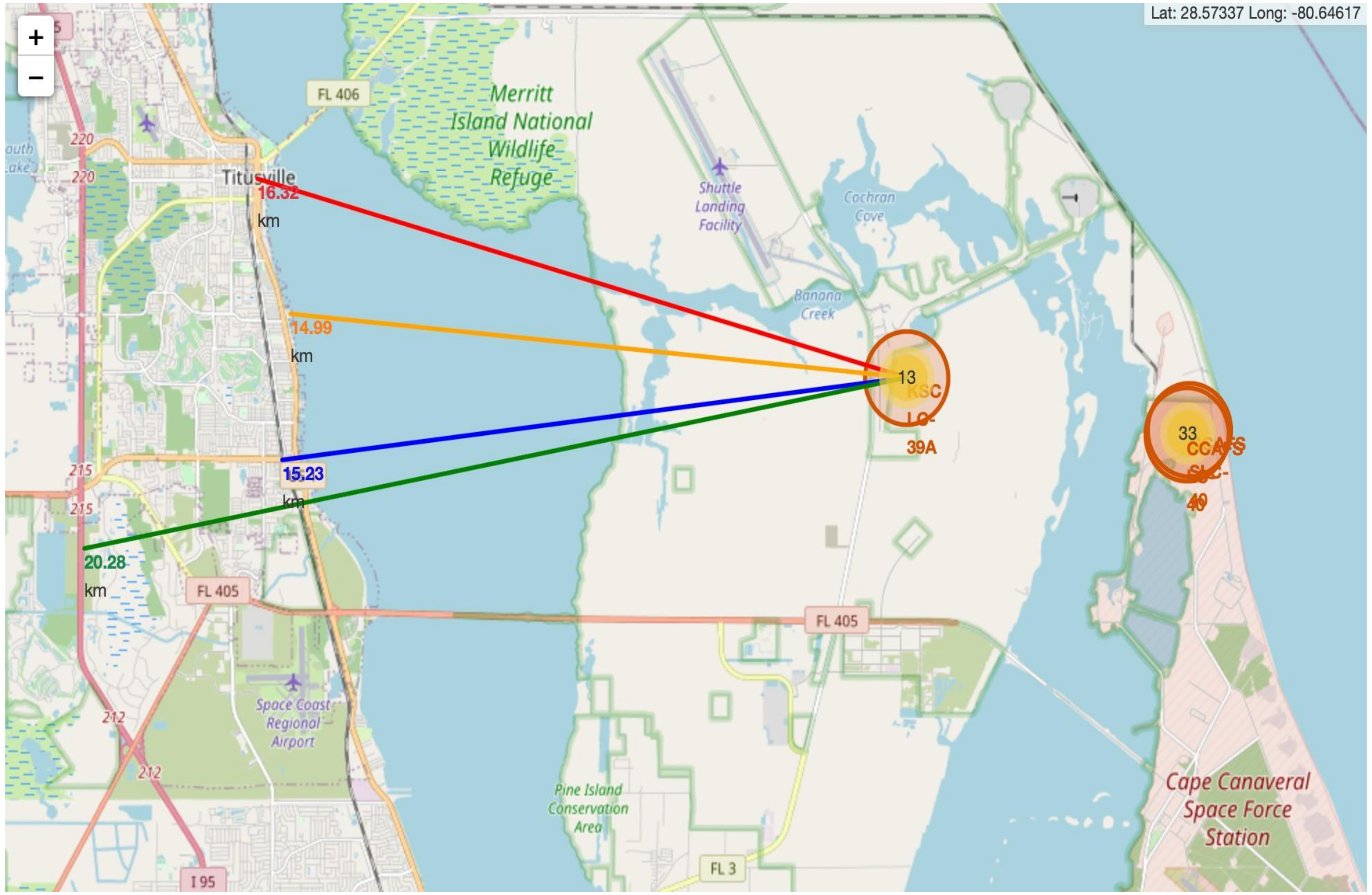
Colour-labeled launch records on the map

Green Marker

Red Marker

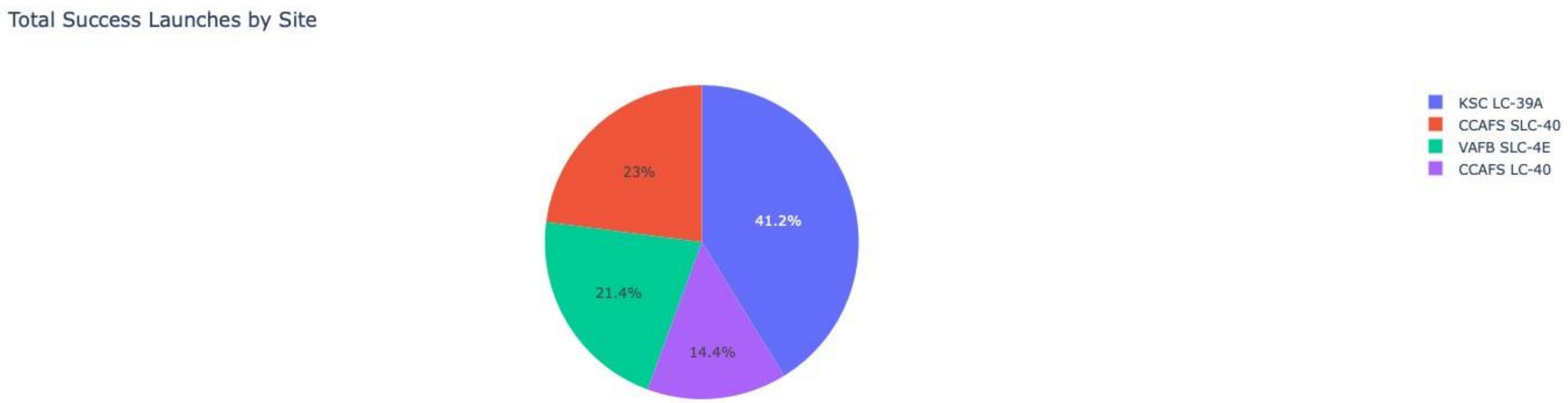


Distance from the launch site KSC LC-39A to its proximities



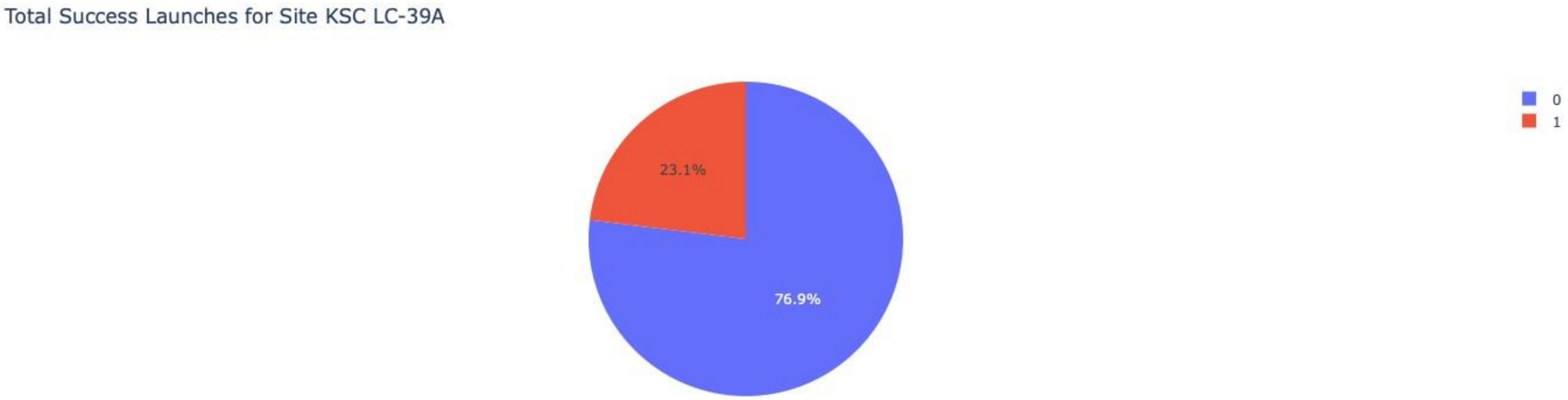
Build a Dashboard with Plotly Dash

Launch success count for all sites



Explanation: KSC LC-39-A has the most successful launches

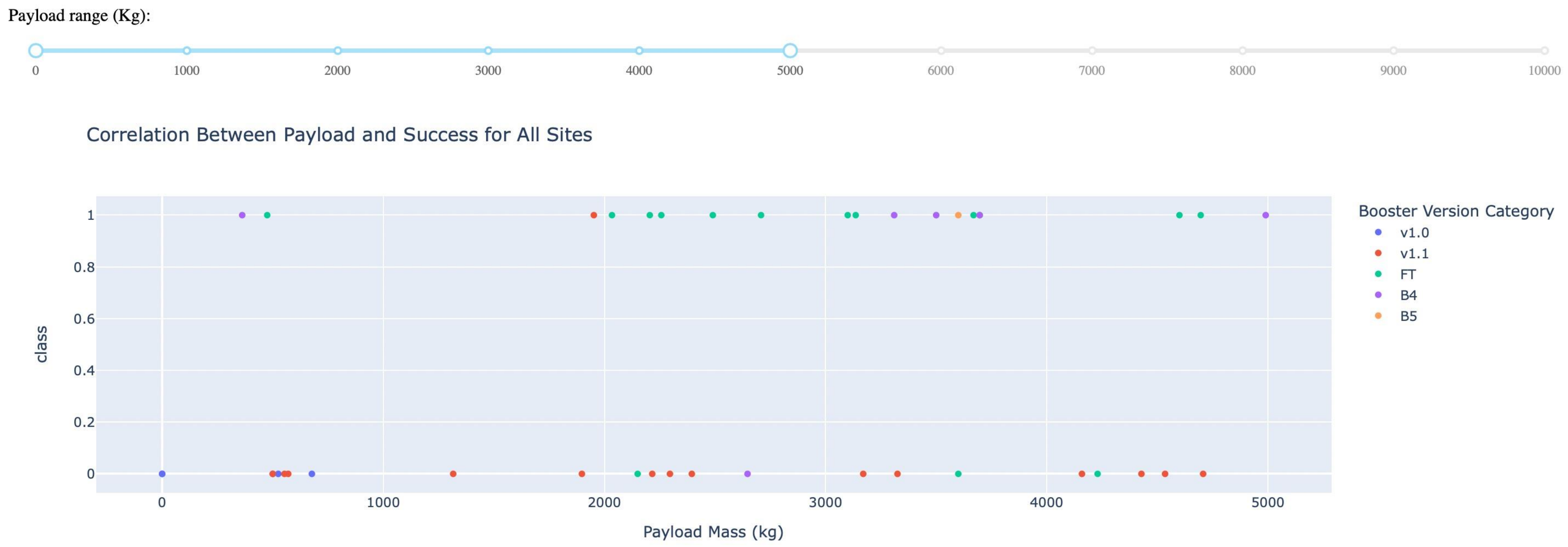
Launch site with highest launch success ratio



Explanation:Only 23.1% of launches where successful.

Payload Mass vs. Launch Outcome for all sites

EXPLANATION:



Predictive analysis (Classification)

Classification Accuracy

Explanation:

SCORES AND ACCURACY OF THE TEST SET

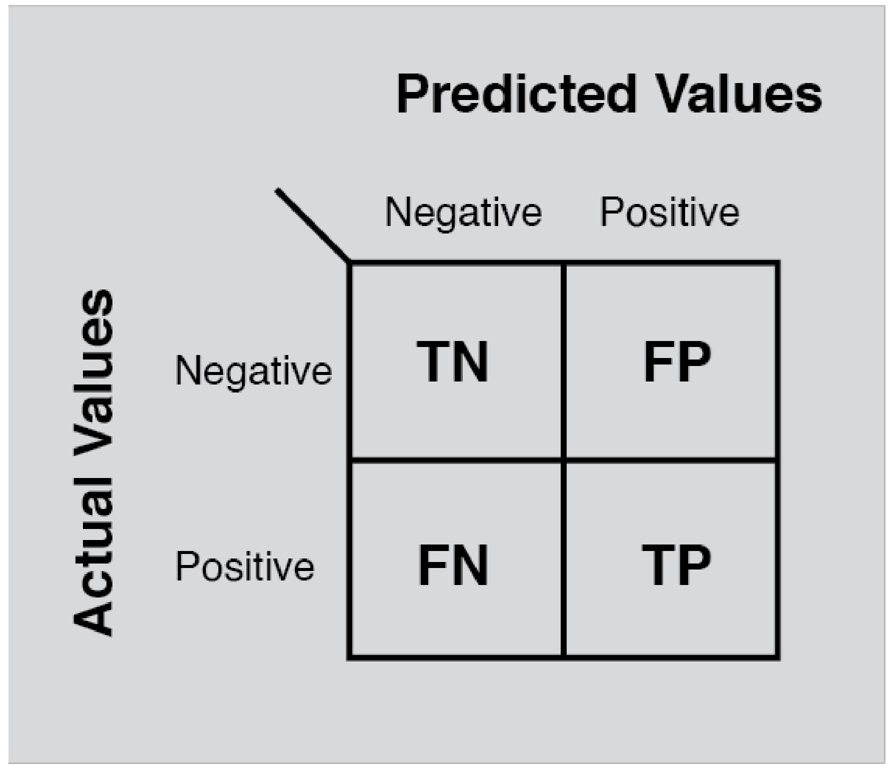
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

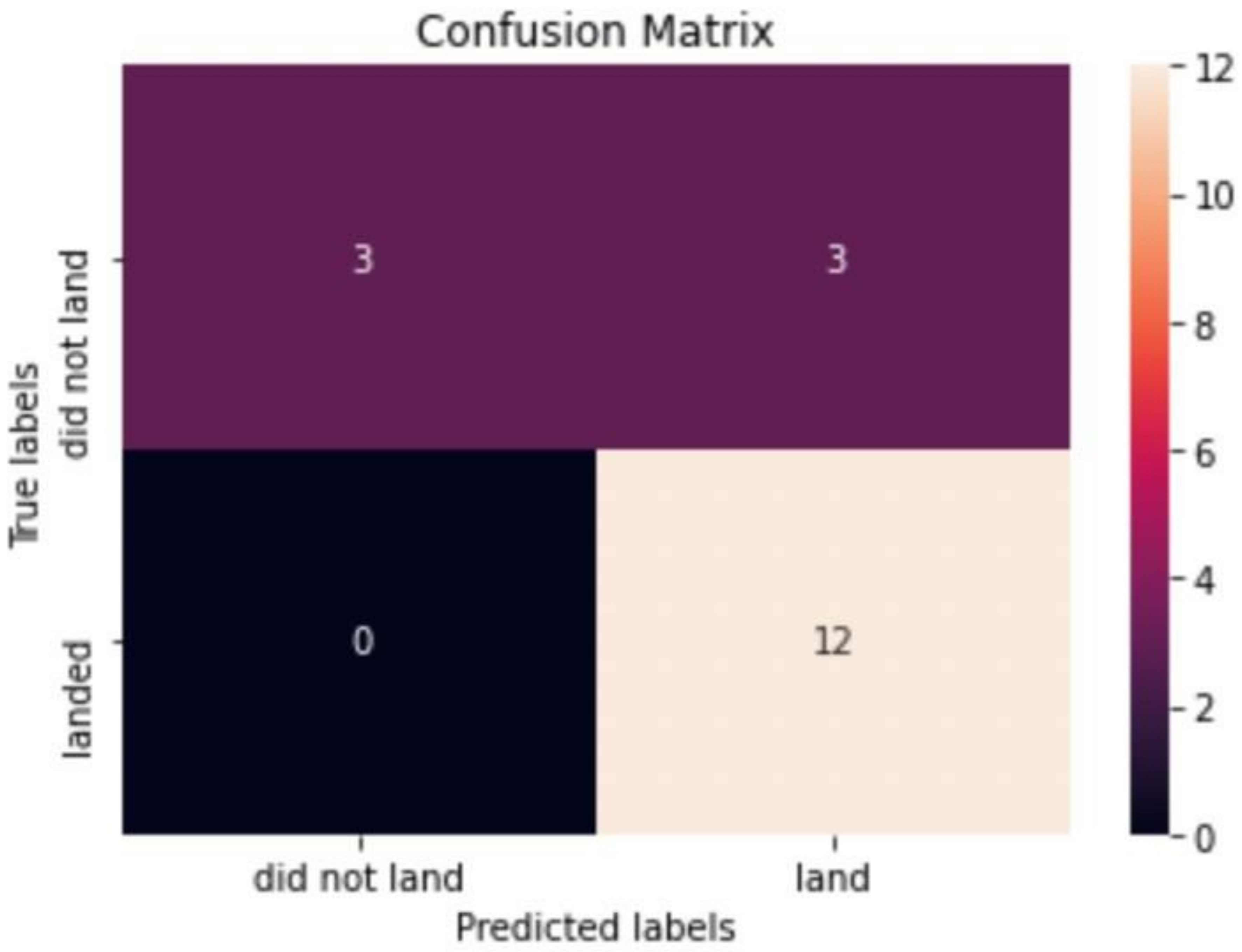
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

EXPLANATION:





Conclusion

- The Decision Tree Model emerges as the most effective algorithm for analyzing this dataset.
- Launch attempts carrying lighter payloads are more likely to succeed compared to those with heavier payloads.
- The launch sites are predominantly situated near the Equator, with all sites being a short distance from coastal areas.
- There has been a noticeable improvement in the success rate of launches over time.
- Among all launch sites, KSC LC-39A boasts the highest rate of successful launches.
- The orbits designated as ES-L1, GEO, HEO, and SSO achieve a 100% success rate in launches

Appendix

A REALLY

GREAT COURSE

THAT OFFERED

A LOT