2024

Data Science Capstone Project

Roberto Aguirre

2024

OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

Executive Sumary

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 results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes Falcon 9 rocket flights, which start at 62 million dollars; in comparison, other suppliers charge up to 165 million dollars per launch; a large portion of the cost savings are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the cost of a launch if we can ascertain if the first stage will land. We are going to make a prediction about whether SpaceX will reuse the first stage based on publicly available data and learning models. Currently, the need has arisen to provide affordable space traveling. In this scenario, the world known company SpaceX has the business problem of estimating the cost of a Falcon 9 launch, which is determined if the first stage of Falcon 9 will land.

Questions to be answered

- What impact do factors like orbits, payload mass, launch site, and number of flights have on the first stage landing's success?
- Does the percentage of successful landings rise with time?
- What is the best algorithm that can be used for binary classification

Methodology

Methodology

Data collection methodology:

- Primary data source: SpaceX API.
- Secondary data source: data available on Wikipedia.

Perform data wrangling

- SpaceX API data has a relatively easy data wrangling, only filtering Falcon 9 data and missing data quick handling. Data available on Wikipedia required HTML syntaxis cleaning, and a HTML parse to Python data frame.

- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models.
- A standard mythology for fitting and testing the classification models were carried out.

2024

Data collection

A combination of web scraping data from a table in SpaceX's Wikipedia entry and API queries from the company's REST API were used in the data collection procedure.

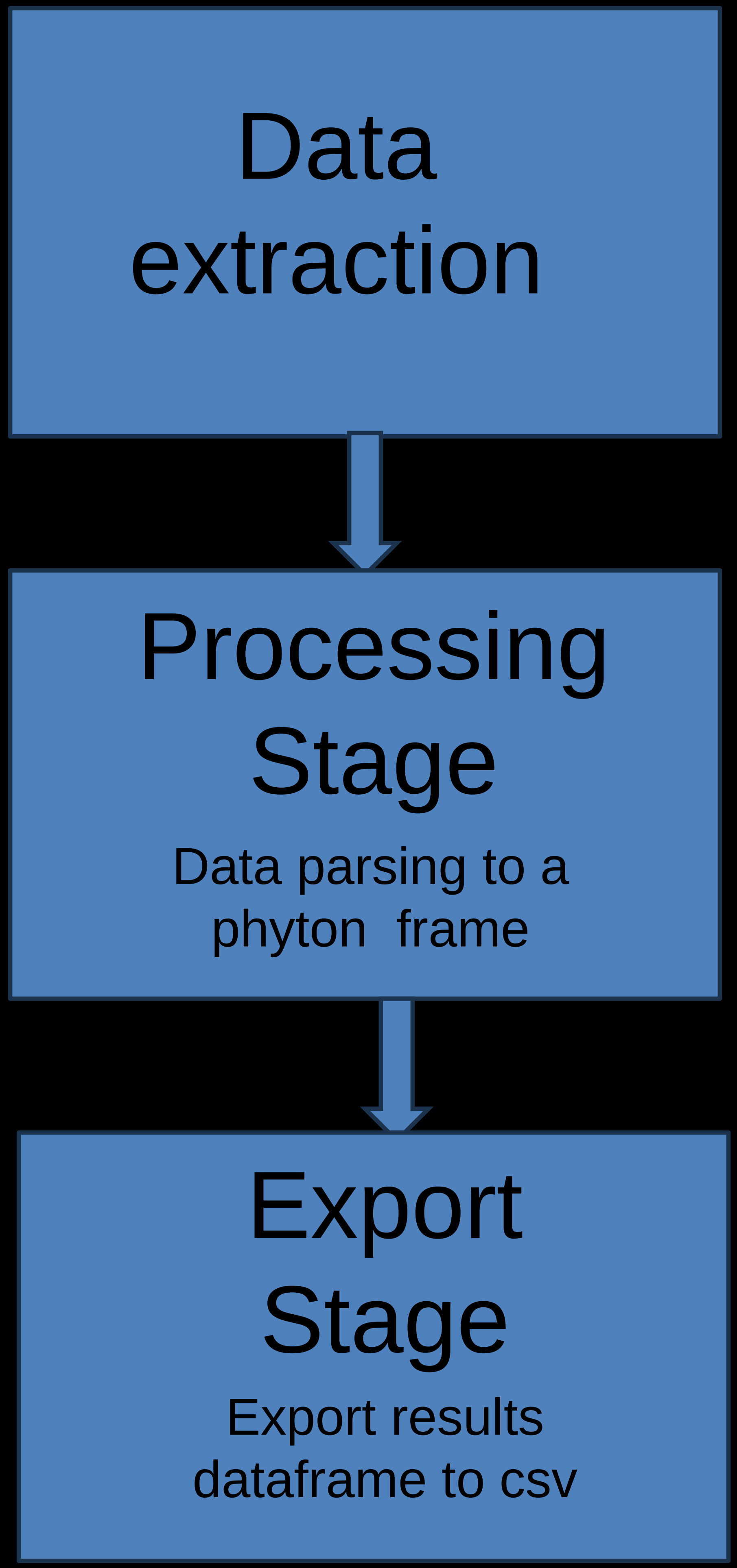
In order to obtain comprehensive information about the launches for a more in-depth analysis, we had to employ both of these data collection techniques.

Data Columns are obtained by using SpaceX REST API:

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

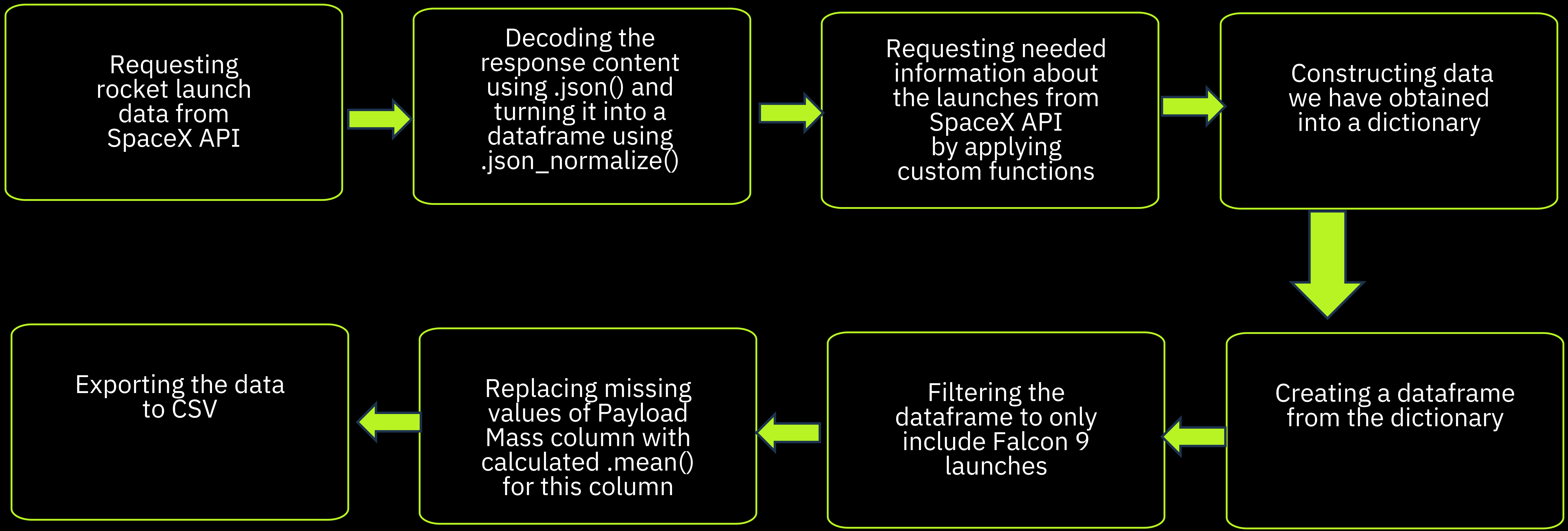
Data Columns are obtained by using Wikipedia Web Scraping:

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



2024

Data collection – SpaceX 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

Data wrangling

There are other instances in the data set where the booster failed to land properly. There are instances where a landing attempt was made but was unsuccessful due to an accident; for instance, True Ocean denotes a successful landing in a particular area of the ocean, whereas False Ocean denotes an unsuccessful landing in a particular area of the ocean.

When a mission is declared successful and lands on a ground pad, it is said to have true RTLS. A mission outcome that was unsuccessfully landed on a ground pad is indicated by a false RTLS.A successful landing of the mission's outcome on a drone ship is referred to as true ASDS. An unsuccessful landing on a drone ship is indicated by a false ASDS.
Mostly, we translate those results into Training Labels, where "0" denotes an unsuccessful landing and "1" indicates a successful booster landing.

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:

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.

Disk plots display the correlation between variables. These could be included in a machine learning model if a relationship is found.

Bar charts display discrete category comparisons. The intention is to illustrate the correlation between a measured value and the particular categories under comparison.

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

• Generally speaking, the graph type selected for each feature relationship analysis was directly related to the necessity of understanding the connection among three or two features.
• The bar plot lets us analyze the relationship between success rate and orbit type. Lastly, a line chart shows the relationship between year and average success rate.

EDA with SQL

Performed SQL queries:

• Listing the names of the space mission's distinct launch locations.
• Showing the average payload mass carried by rocket version F9 v1.1; • Showing the total payload mass carried by boosters launched by NASA (CRS); • Showing five records where launch sites start with the term "CCA."
• A list of the names of the boosters that have successfully landed on a drone ship with a payload mass of more than 4,000 but less than 6,000, along with the date of the first successful landing outcome in a ground pad
• Counting the overall number of mission outcomes that were successful and unsuccessful. • 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

Build an interactive map with Folium

Markers of all Launch Sites:

- Added a NASA Johnson Space Center marker with a circle, pop-up label, and text label, utilizing the center's latitude and longitude as a starting point.

Coloured Markers of the launch outcomes for each Launch Site:

In order to determine which launch sites have comparatively high success rates, colored markers representing successful (Green) and unsuccessful (Red) launches were added using Marker Cluster.

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.

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:
- Added a dropdown list to enable Launch Site selection.

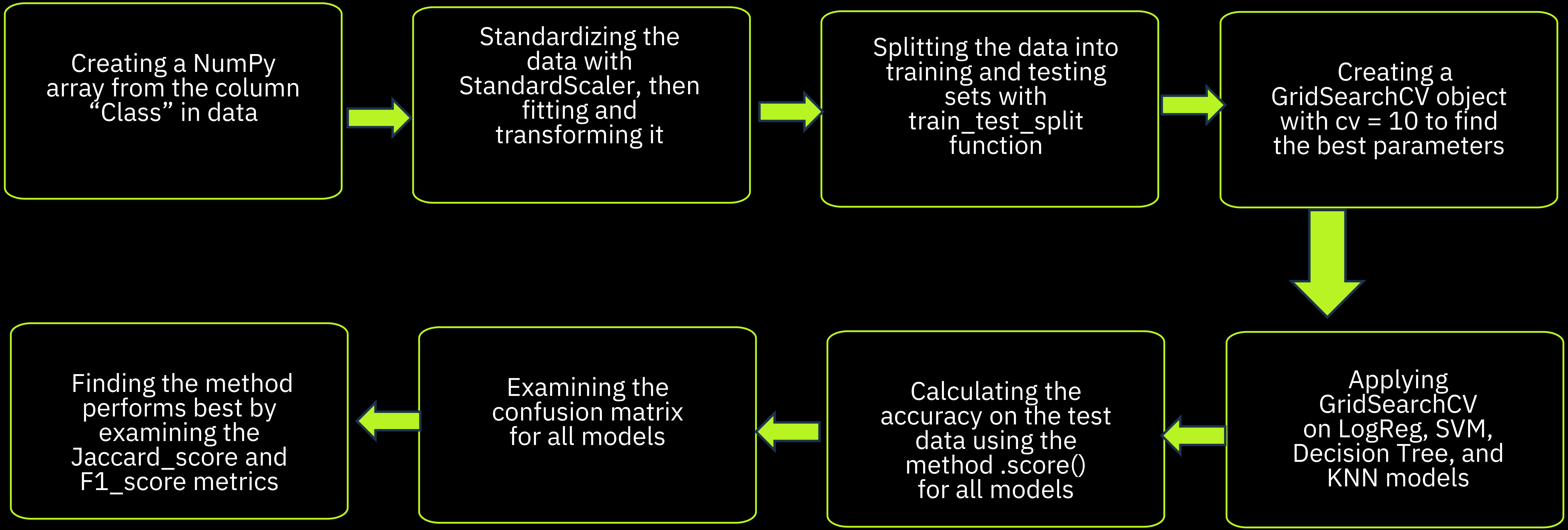
Pie Chart showing Success Launches (All Sites/Certain Site):
Included a pie chart that displays the total number of successful launches across all sites as well as the success versus failure counts for each site that was chosen as a launch location.

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 Succes

Predictive analysis (Classification)

2024

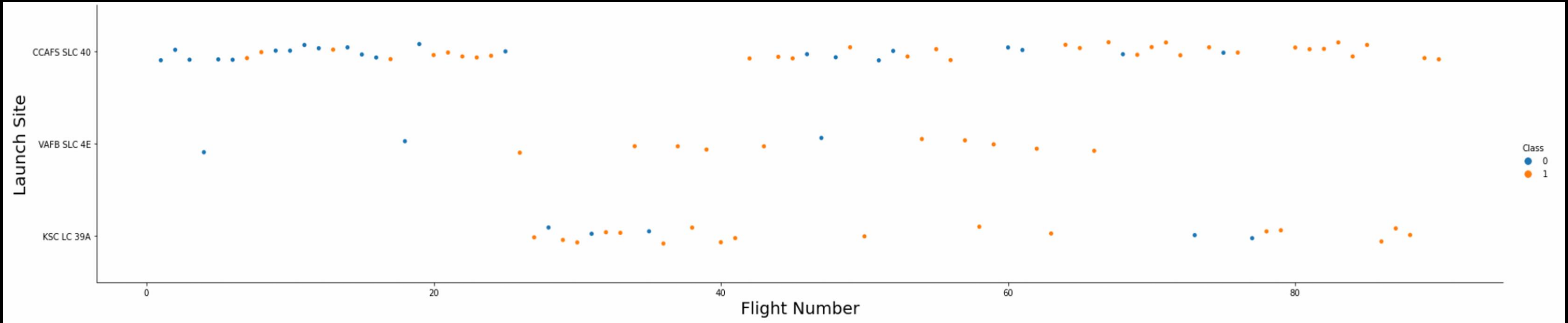


RESULTS

EDA with Visualization

2024

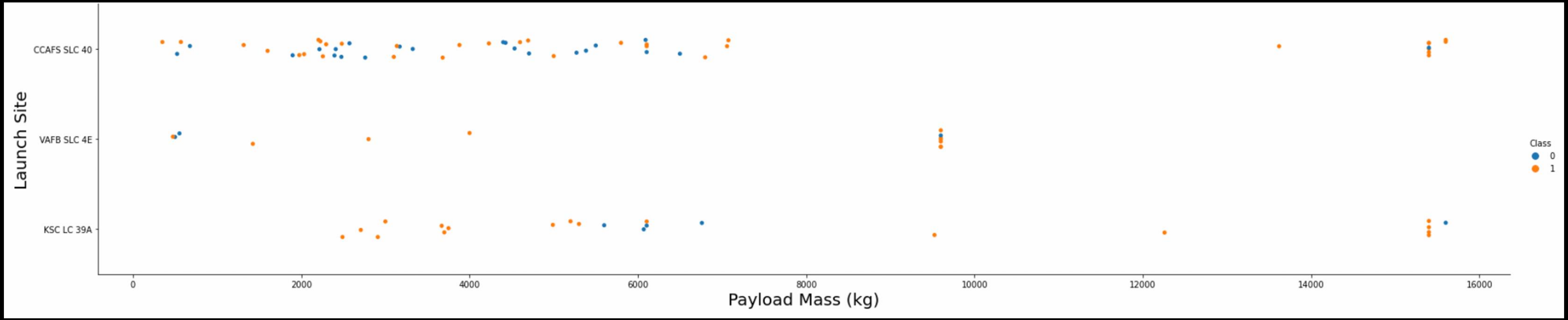
Flight Number vs. Launch Site



All of the most recent flights were successful, whereas none of the earlier ones were.
Approximately 50% of all launches occur at the CCAFS SLC 40 launch location.
Success rates are greater at KSC LC 39A and VAFB SLC 4E.
One can expect that the success rate increases with each fresh launch.

EDA with Visualization

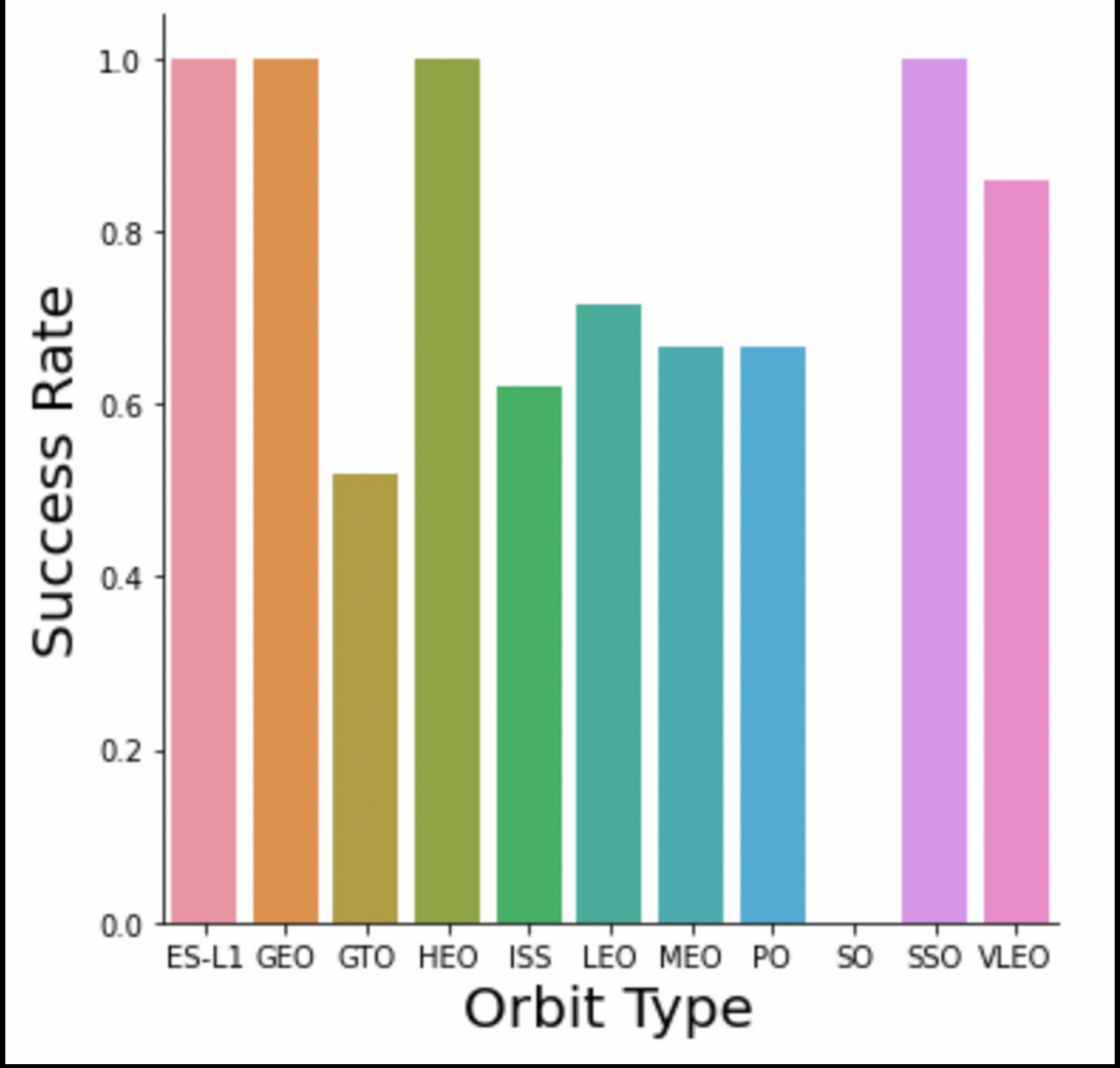
Payload vs. Launch Site



The success rate increases with payload mass at each launch site.
With payload masses exceeding 7000 kg, the majority of launches were successful.
Under 5500 kg, KSC LC 39A also has a 100% success rate for payload mass.

EDA with Visualization

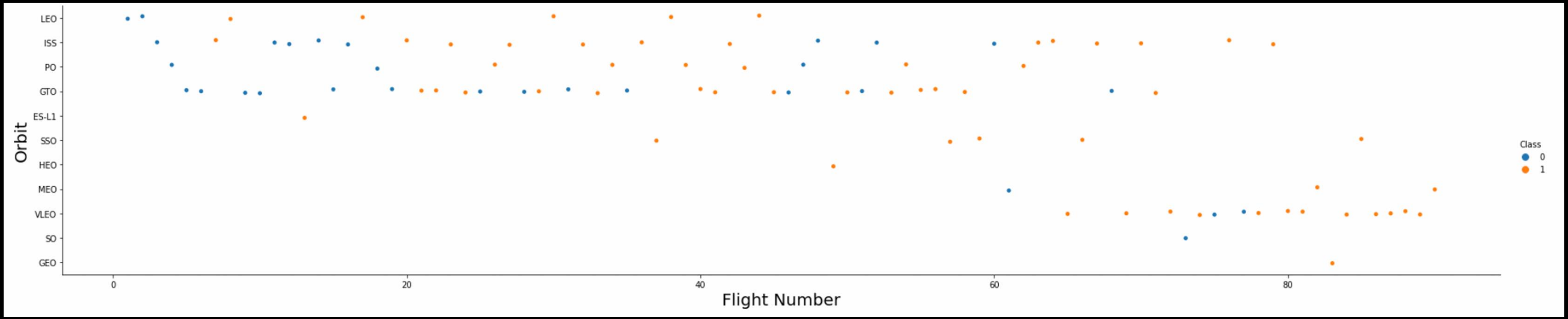
Success rate vs. Orbit type



Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO
Orbits with 0% success rate:
- SO Orbits with success rate
between 50% and 85%:
- GTO, ISS, LEO, MEO, PO

EDA with Visualization

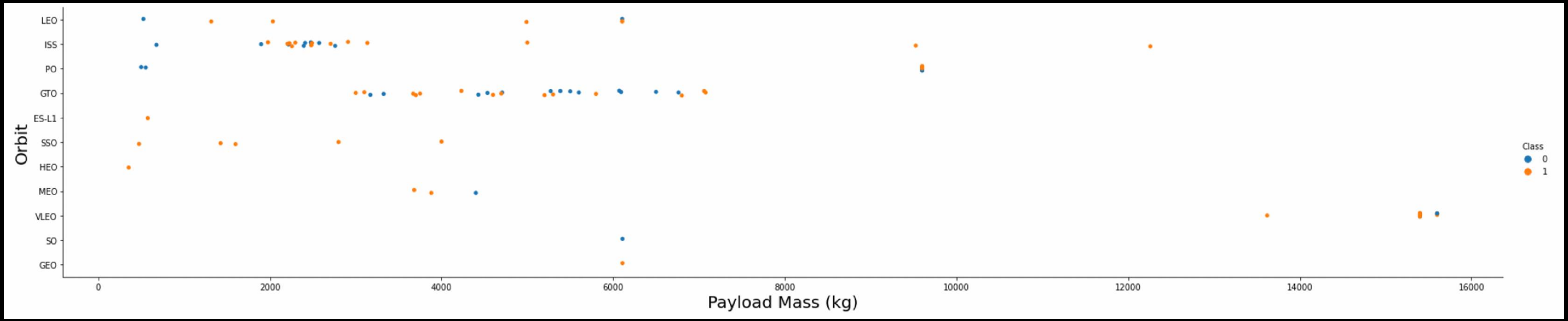
Flight Number vs. Orbit type



While in GTO orbit, there doesn't seem to be any correlation between the number of flights and success; however, in LEO orbit, it does appear to be related.

EDA with Visualization

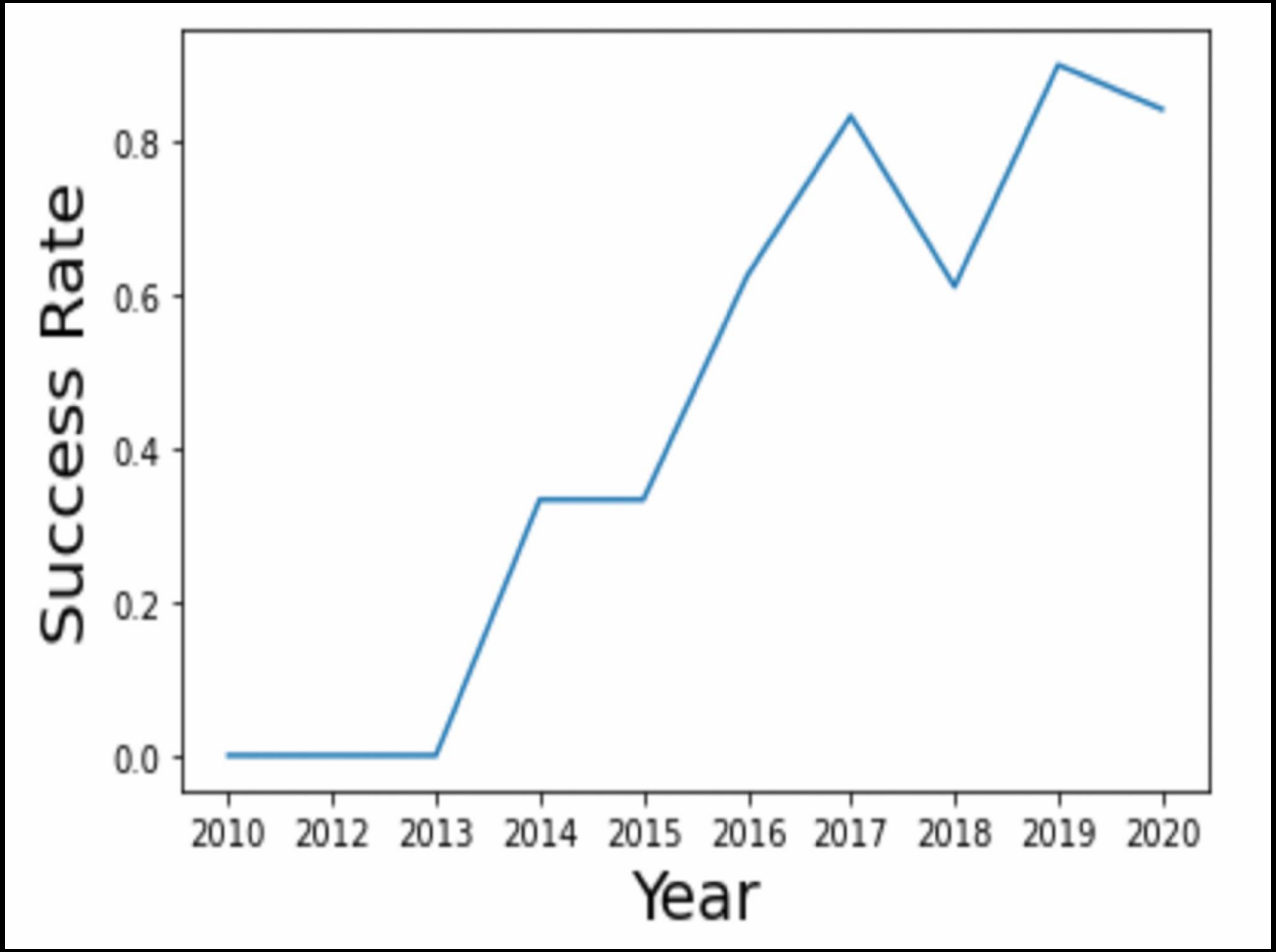
Payload Mass vs. Orbit type



Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

EDA with Visualization

Launch success yearly trend



The success rate since 2013 kept increasing till 2020.

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

putting the names of the space mission's distinct launch sites on display.

EDA with SQL

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

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

EDA with SQL

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

Displaying the total tonnage of payload carried by NASA-launched rockets (CRS)

EDA with SQL

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

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

EDA with SQL

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

Stating the date on which the ground pad had its first successful landing.

EDA with SQL

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

A list of the boosters' names that have been successful in drone ships with payload masses over 4,000 but under 6,000

EDA with SQL

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

Listing the total number of successful and failure mission outcomes.

EDA with SQL

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

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

EDA with SQL

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)

A list of the drone ship's unsuccessful landings along with the names of the launch sites and the versions of their boosters for each month in 2015.

EDA with SQL

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.
```

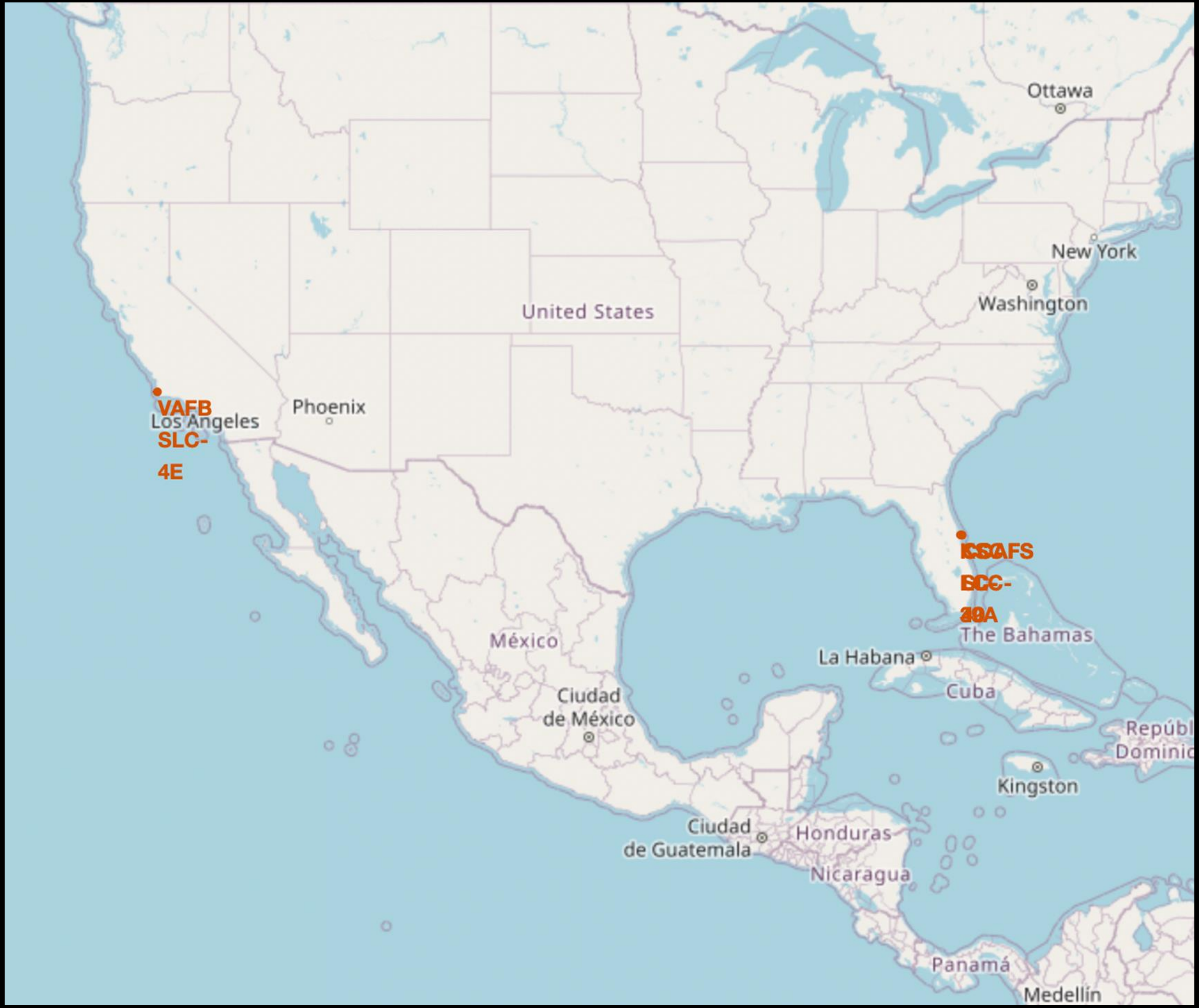
Out[13]:

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

Sorting the number of landing results (e.g., ground pad success or drone ship failure) between 2010-06-04 and 2017-03-20 in descending order.

Interactive map with Folium

All launch sites' location markers on a global map



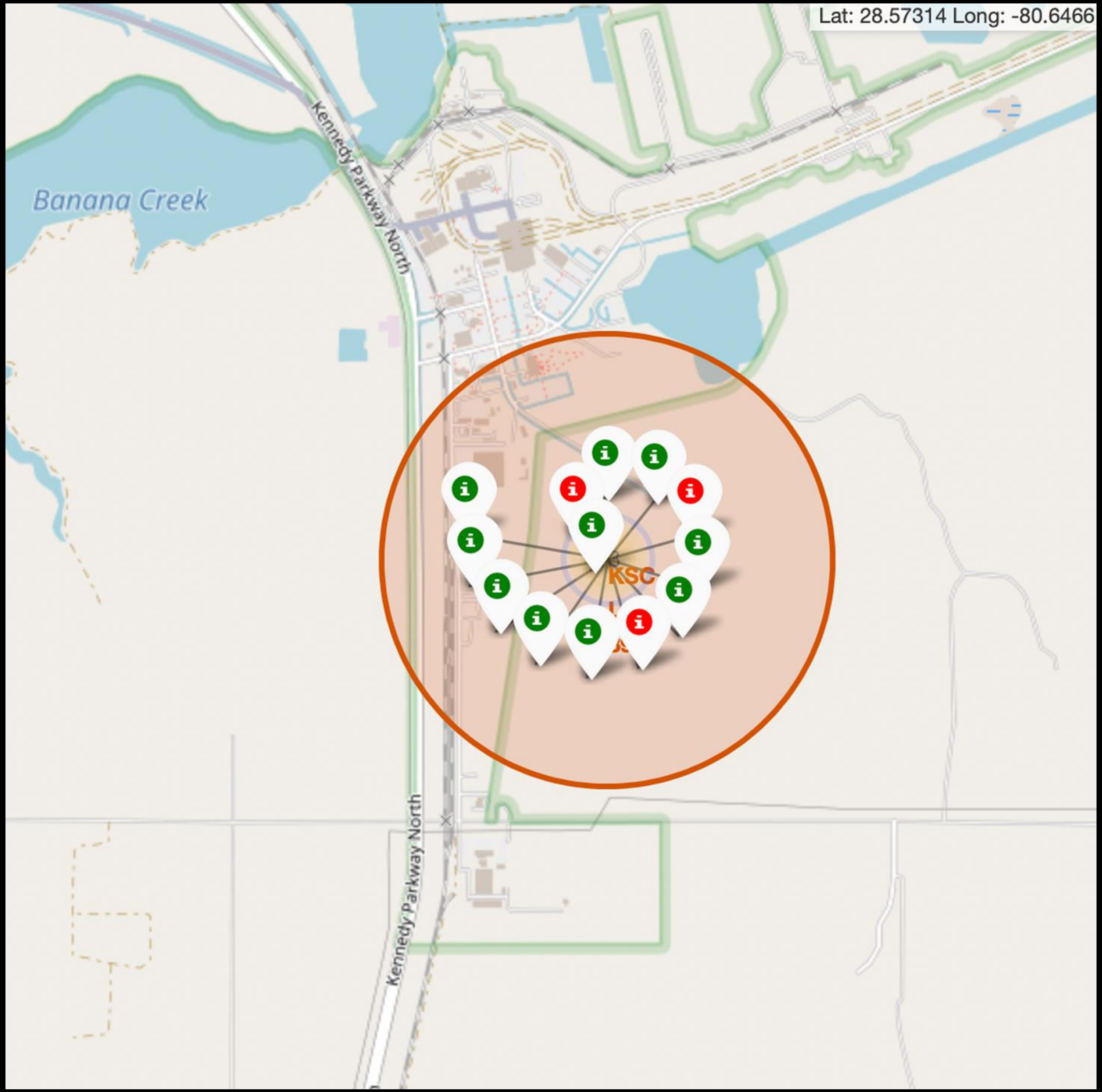
The majority of launch sites are close to the Equator. At the equator, land moves more quickly than it does anywhere else on Earth's surface. At the equator, everything on Earth is already travelling at a speed of 1670 km/h.

A ship launched from the equator travels through space at the same speed as it did prior to launch, as well as around the planet. Inertia is the cause of this. This velocity will assist the spacecraft in maintaining a sufficient speed to remain in orbit.

All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it 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.

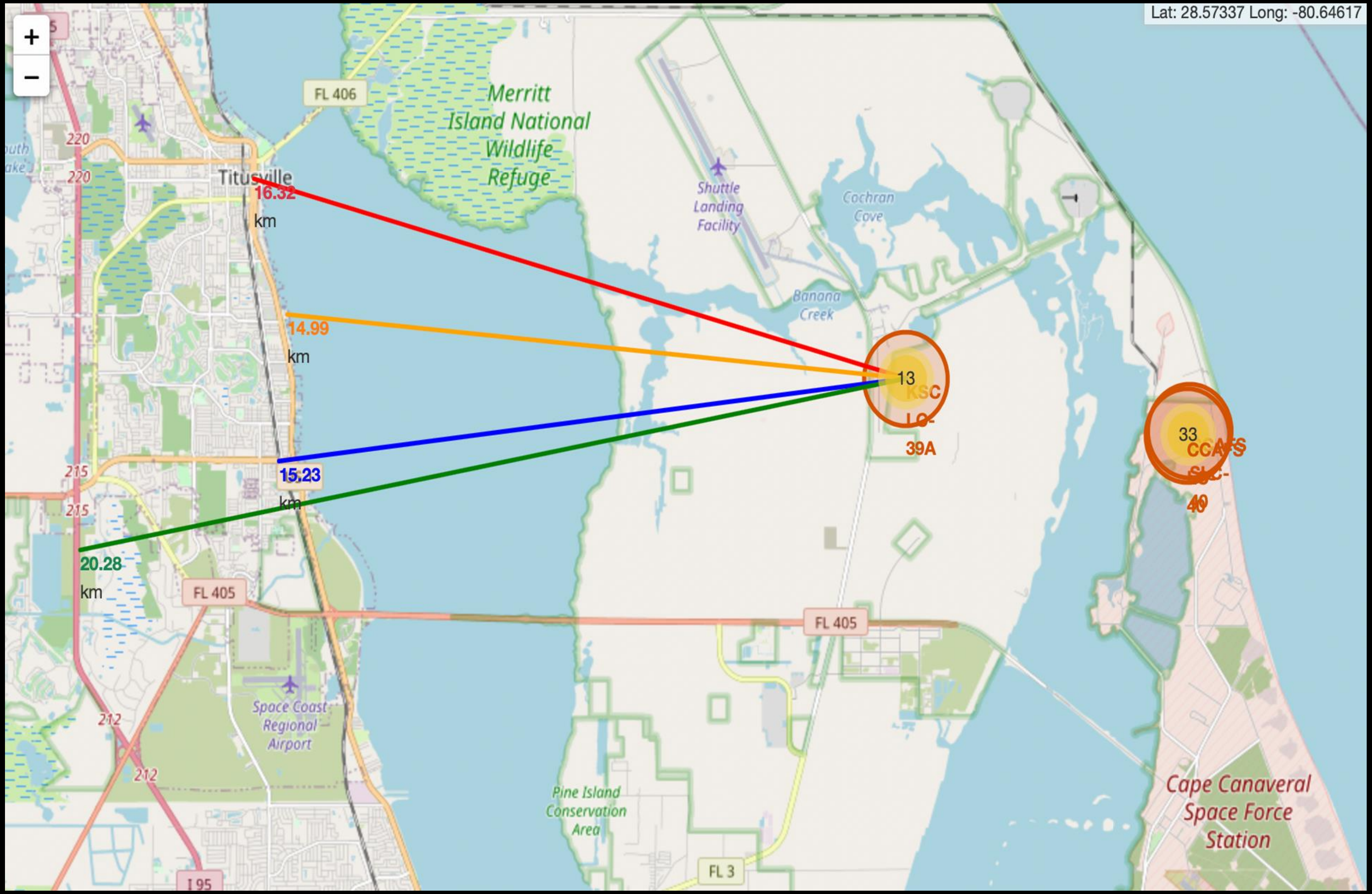
- Green Marker = Successful Launch
- Red Marker = Failed Launch

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

Interactive map with Folium

2024

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



From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

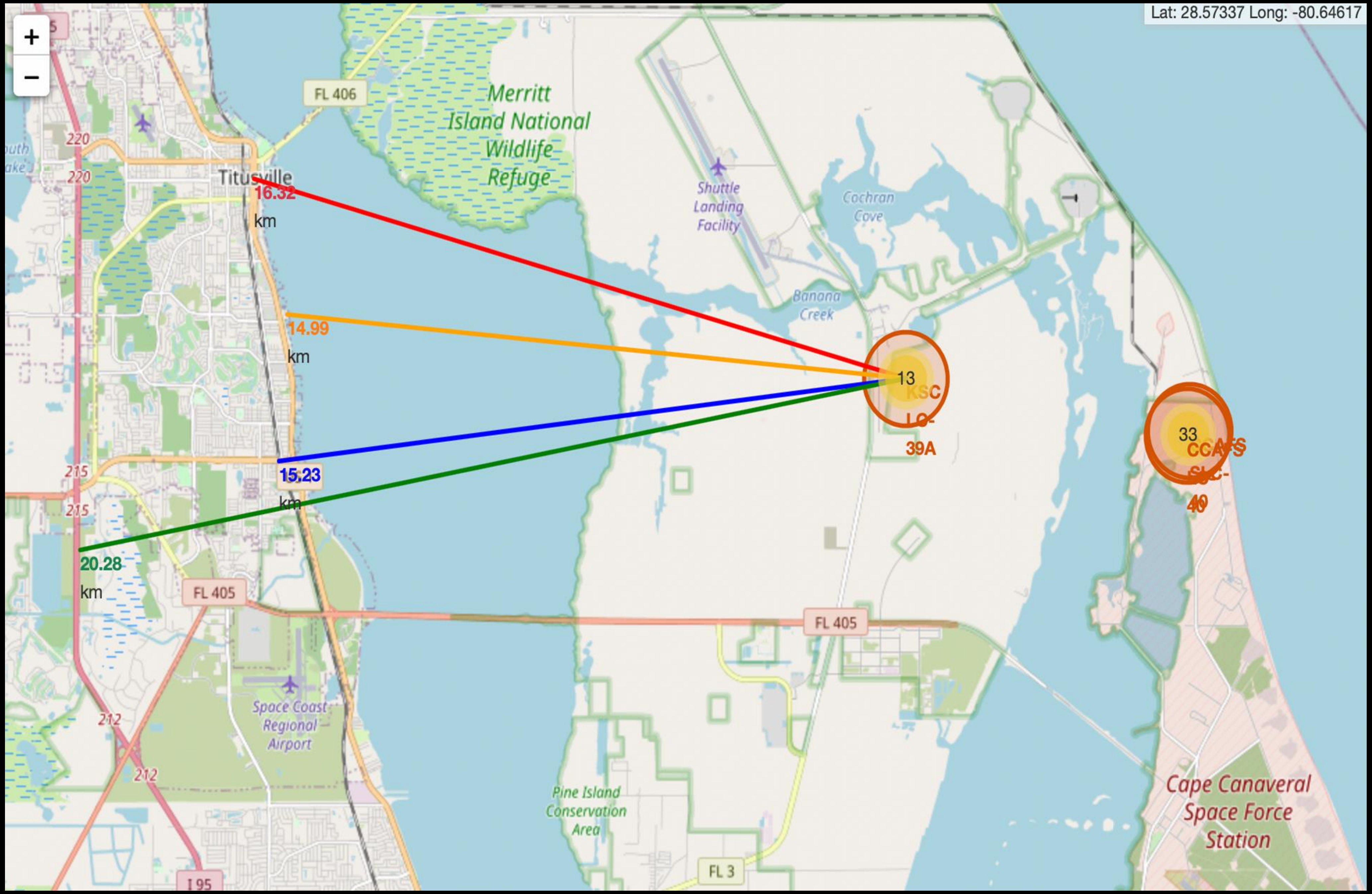
Additionally, Titusville, the closest city (16.32 km) is relatively close to the launch site KSC LC-39A.

A failed rocket can travel up to 15-20 km in a matter of seconds due to its high speed. It might pose a threat to densely inhabited areas.

Interactive map with Folium

2024

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



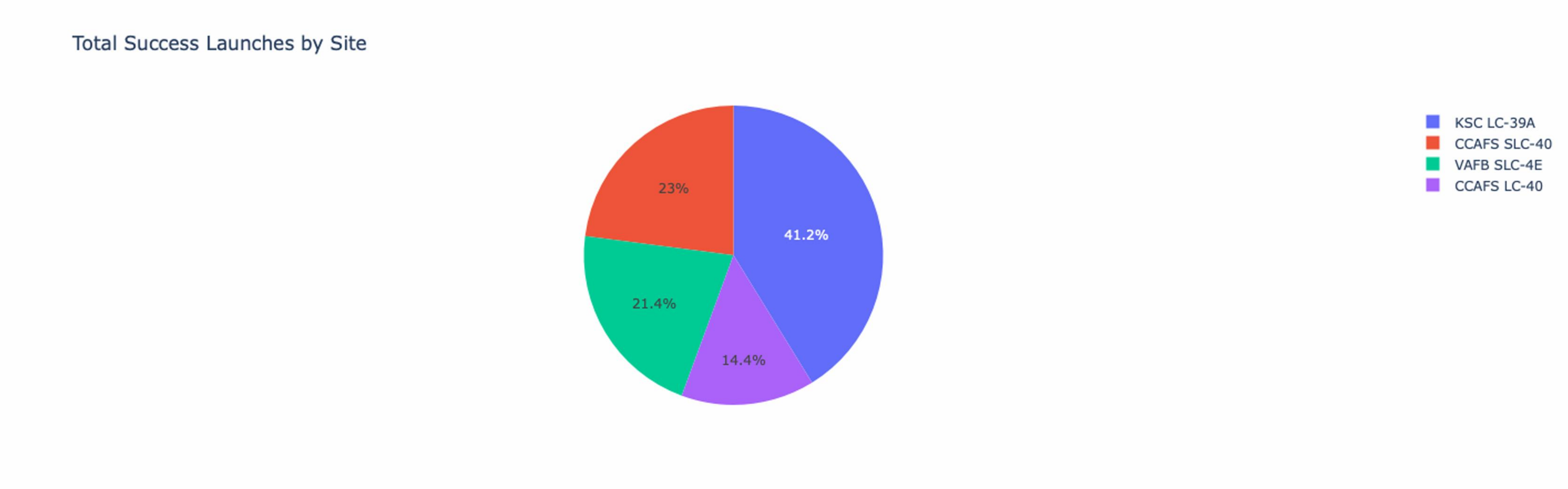
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

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

Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

Build a Dashboard with Plotly Dash

Launch success count for all sites

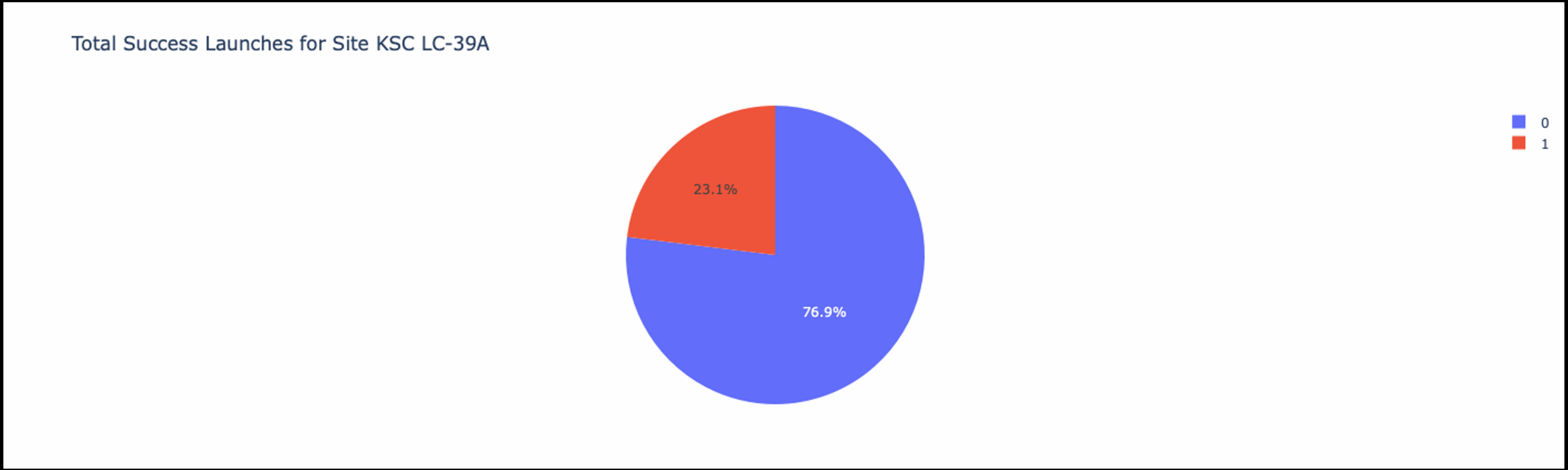


The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Build a Dashboard with Plotly Dash

2024

Launch site with highest launch success ratio

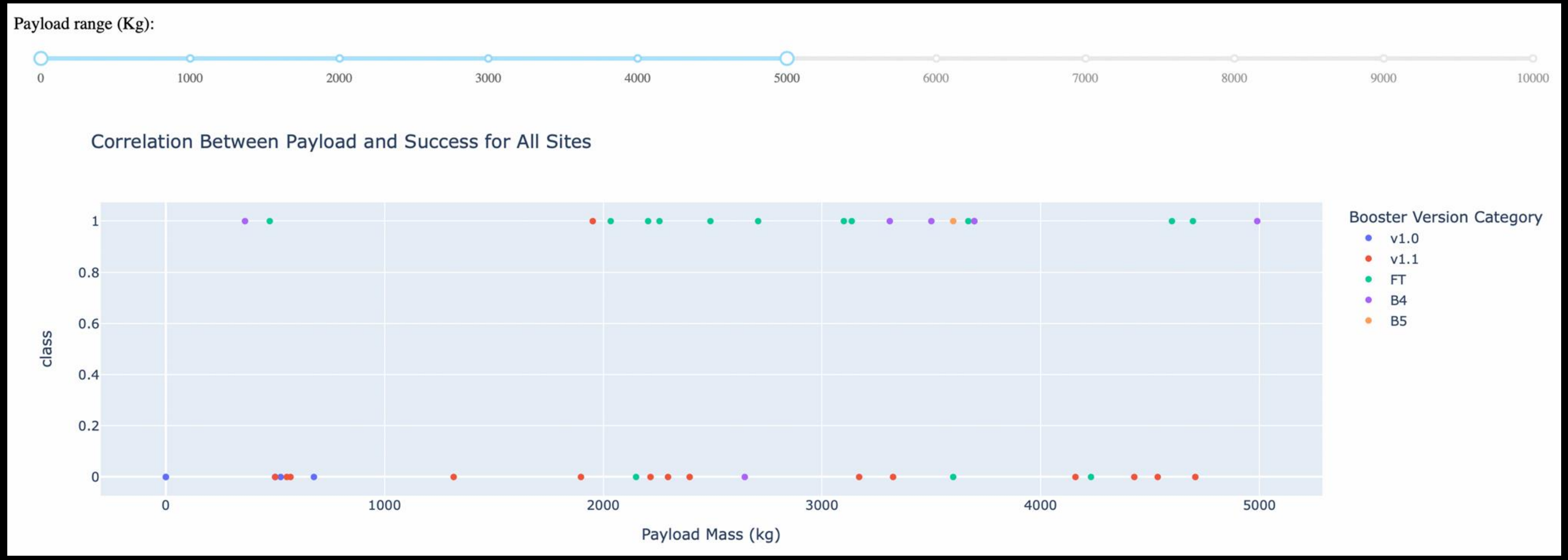


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Build a Dashboard with Plotly Dash

2024

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)

Classification Accuracy

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

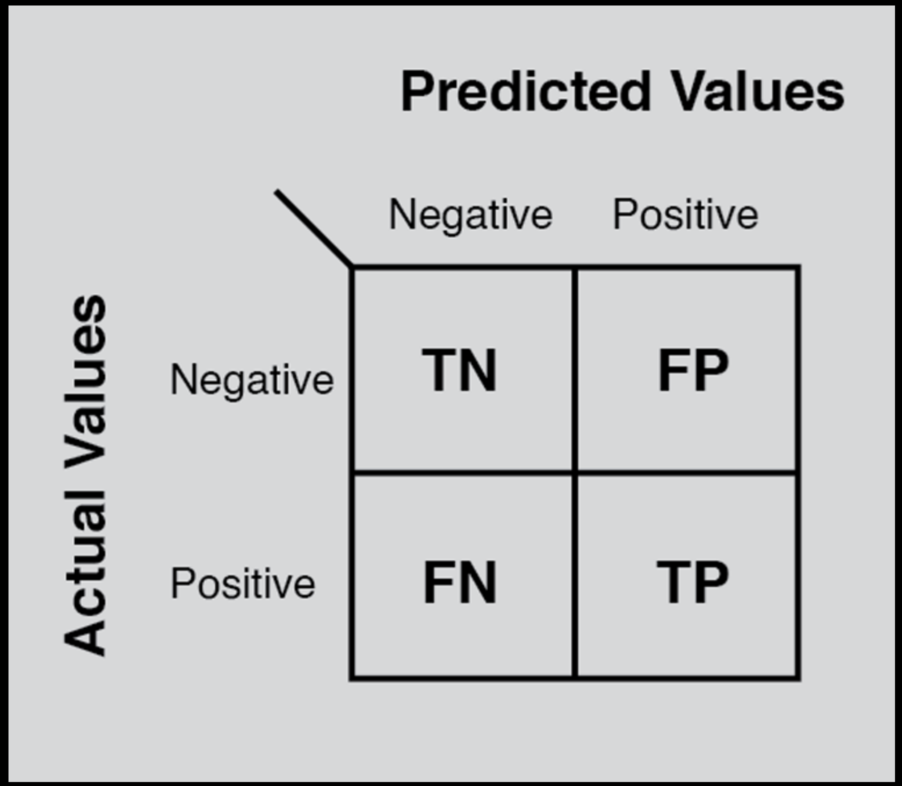
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

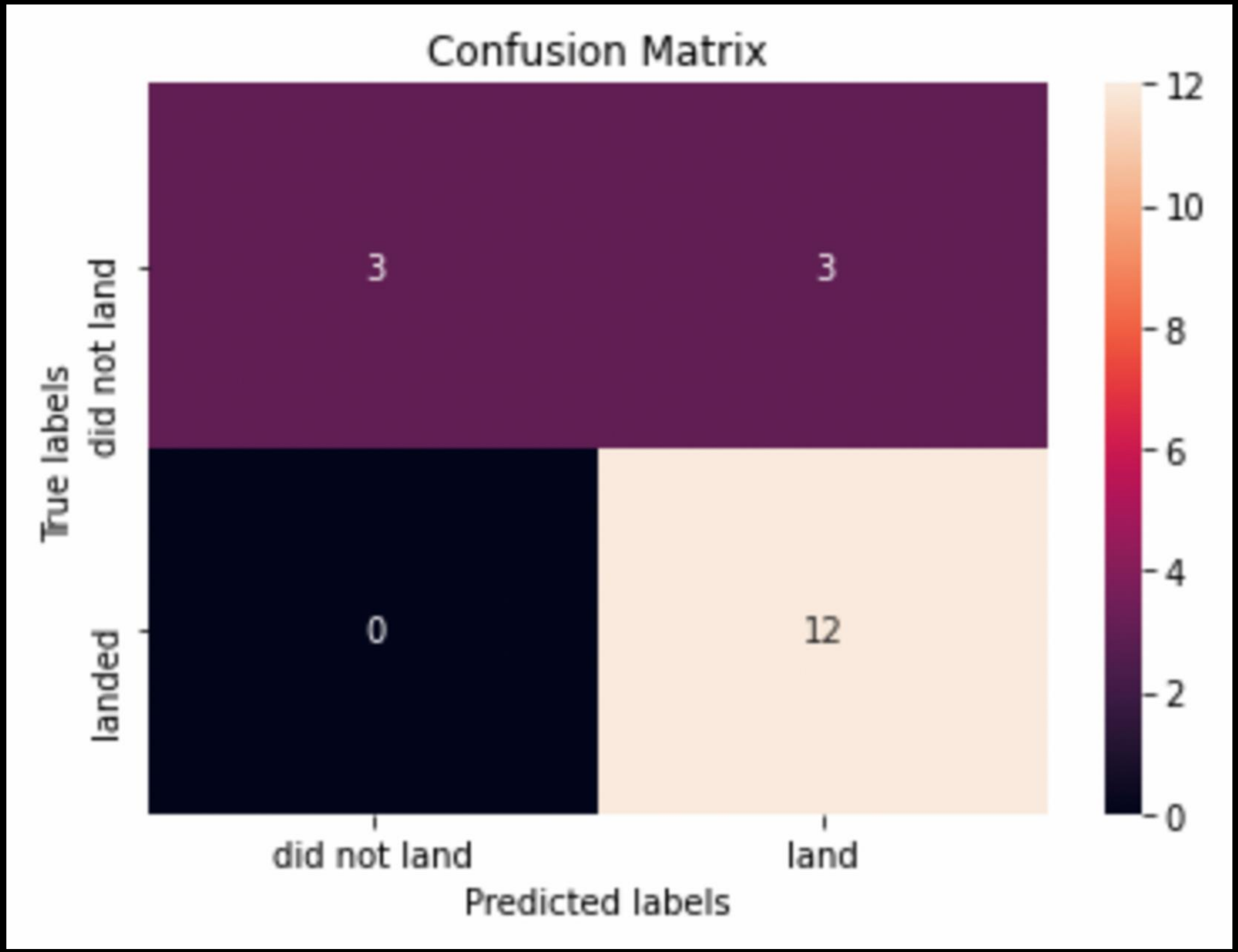
- We are unable to verify the optimal strategy based on the Test Set scores.
- The short test sample size (18 samples) may be the cause of the same test set scores. As a result, we used the entire dataset to test every approach.
- The Decision Tree Model is the best model, as confirmed by the scores of the entire dataset. This model offers the highest accuracy as well as higher scores.

Predictive analysis (Classification)

Confusion Matrix

It is evident by looking at the confusion matrix that logistic regression is capable of differentiating between the various classes. It is evident that false positives are the main issue.





CONCLUSIONS

Decision Tree Model is the best algorithm for this dataset.

The success rate of launches increases over the years.

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

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

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

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