Winning Space Race with Data Science

Alexander Cruz
7/21/22

Outline

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

Executive Summary

In this project we follow a methodology of:

- Collecting our data on SpaceX Falcon 9 launches
- Process, transform, and clean the data
- Perform exploratory data analysis using descriptive statistics and advanced visualizations
- Generate and evaluate machine learning models to help us predict successful first stage rocket landings

Summarily, we find key insights regarding launches and are able to generate predictive models that are able to determine whether a SpaceX Falcon 9 rocket launch will have a successful first stage landing given information on certain features/variables of a launch.

Introduction

SpaceX 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.

For this project we want to analyze SpaceX Falcon 9 data and see if we can determine if the first stage of a rocket launch will land.

If we are able to determine if the first stage will land, we can determine the cost of a launch. This information can then be used by an alternate company that wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We collect our data using a combination for REST API callouts to api.spacexdata and web scraping of web page tables of Falcon 9 records
- Perform data wrangling
  - We use one-hot encoding of categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected using REST API callouts to SpaceX endpoints :

- GET requests were made to the appropriate SpaceX API endpoints.
- HTTP response JSONs was parsed out and then normalized into a Pandas dataframe.
- The data was then sampled and null values were replaced with mean column data where applicable

We also performed web scraping of a Wikipedia page listing Falcon 9 launches and then parsed the scraped records using BeautifulSoup.

Data Collection - SpaceX API

We retrieve data through API callouts to the following SpaceX endpoints:

- launches/past
- /rockets
- /launchpads
- /payloads
- /cores

See the Data Collection - API notebook for further details on our data collection flow.

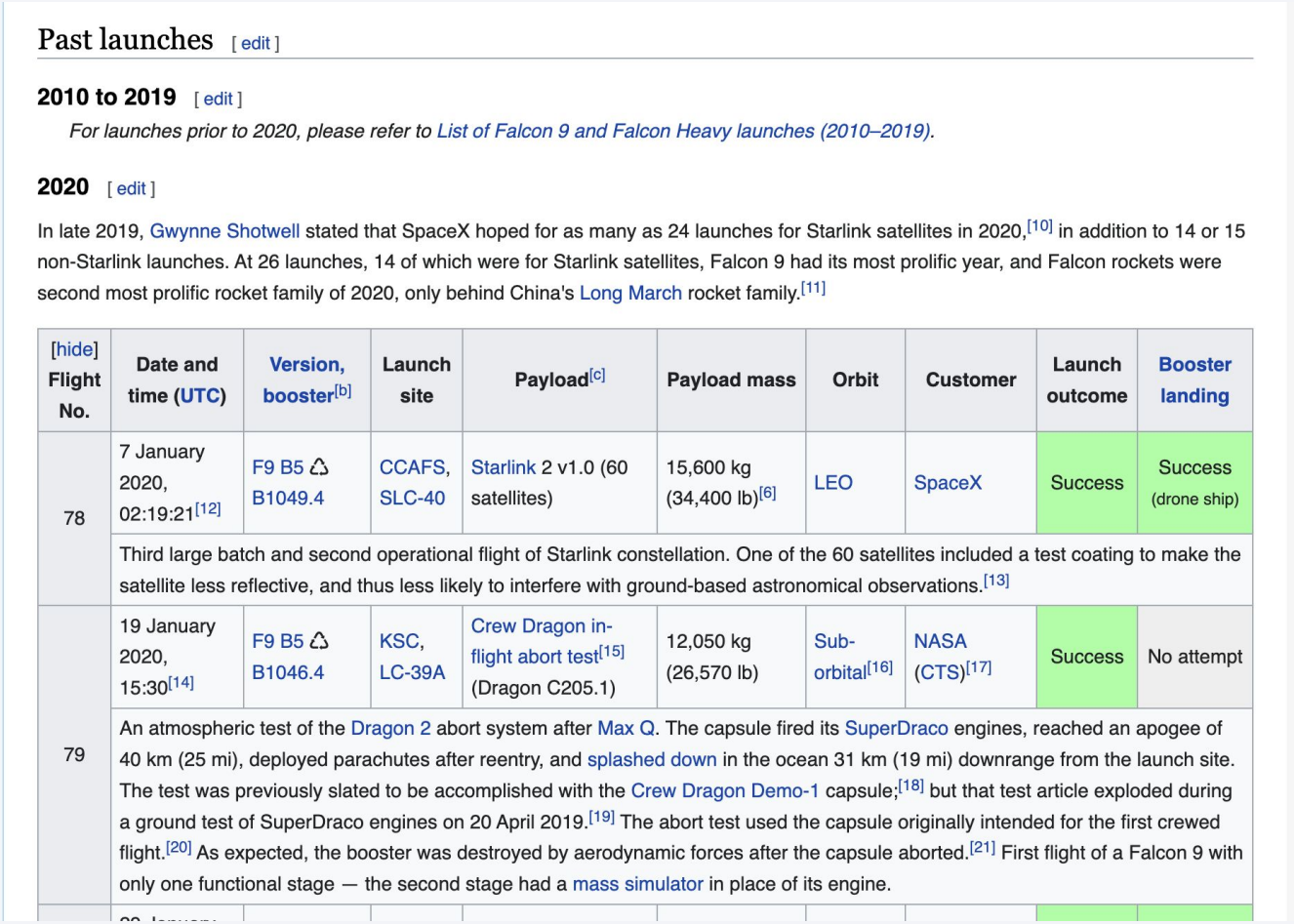
```
In [6]: 1 spacex_url="https://api.spacexdata.com/v4/launches/past"

In [7]: 1 response = requests.get(spacex_url)

        Check the content of the response

In [8]: 1 print(response.content)
```


```
b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"link
s":{"patch":{"small":"https://images2.imgbox.com/3c/0e/T8iJcSN3_o.png","large":"https://image
s2.imgbox.com/40/e3/GypSkayF_o.png"},"reddit":{"campaign":null,"launch":null,"media":null,"re
covery":null},"flickr":{"small":[],"original":[]},"presskit":null,"webcast":"https://www.yout
ube.com/watch?v=0a_00nJ_Y88","youtube_id":"0a_00nJ_Y88","article":"https://www.space.com/2196
-spacex-inaugural-falcon-1-rocket-lost-launch.html","wikipedia":"https://en.wikipedia.org/wik
i/DemoSat"},"static_fire_date_utc":"2006-03-17T00:00:00.000Z","static_fire_date_unix":1142553
600,"net":false,"window":0,"rocket":"5e9d0d95eda69955f709d1eb","success":false,"failures":
[{"time":33,"altitude":null,"reason":"merlin engine failure"}],"details":"Engine failure at 3
3 seconds and loss of vehicle","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b5b6c3bb
0006eeb1e1"],"launchpad":"5e9e4502f5090995de566f86","flight_number":1,"name":"FalconSat","dat
e_utc":"2006-03-24T22:30:00.000Z","date_unix":1143239400,"date_local":"2006-03-25T10:30:00+1
2:00","date_precision":"hour","upcoming":false,"cores":[{"core":"5e9e289df35918033d3b2623","f
light":1,"gridfins":false,"legs":false,"reused":false,"landing_attempt":false,"landing_succes
s":null,"landing_type":null,"landpad":null}],"auto_update":true,"tbd":false,"launch_library_i
d":null,"id":"5eb87cd9ffd86e000604b32a"},{"fairings":{"reused":false,"recovery_attempt":fals
e,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/4f/e3/I
0lkuJ2e_o.png","large":"https://images2.imgbox.com/be/e7/iNqsqVYM_o.png"},"reddit":{"campaig
n":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":[],"original":[]},"pres
```

Data Collection - Scraping

We web scraped a Wikipedia page listing Falcon 9 launches and then parsed the scraped records from an HTML table using BeautifulSoup.

See the Data Collection - Web Scraping notebook for further details on our data collection flow.



Past launches [edit]

2010 to 2019 [edit]

For launches prior to 2020, please refer to List of Falcon 9 and Falcon Heavy launches (2010–2019).

2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[10] in addition to 14 or 15 non-Starlink launches. At 26 launches, 14 of which were for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[11]

[hide] Flight No.	Date and time (UTC)	Version, booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[12]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[6]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[13]								
79	19 January 2020, 15:30[14]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[15] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[16]	NASA (CTS)[17]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[18] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[19] The abort test used the capsule originally intended for the first crewed flight.[20] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[21] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								

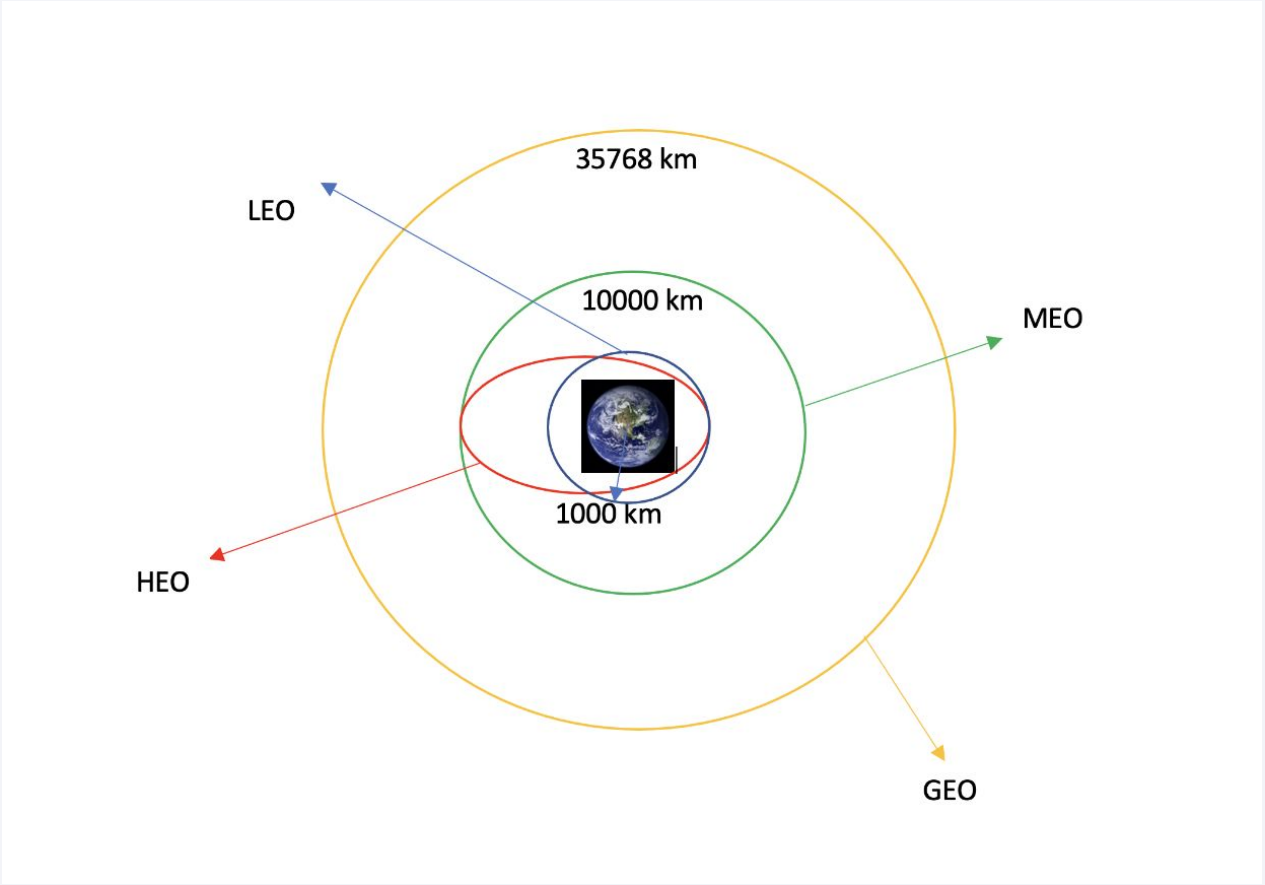
Data Wrangling

We did an exploratory data analysis to determine the correct training labels.

We then calculated the number of launches per site as well as the number of occurrences for each type of orbit. See image.

We then created a landing outcome label from the raw outcome column and exported to a .csv file for further usage.

See the Data Wrangling notebook for further details.



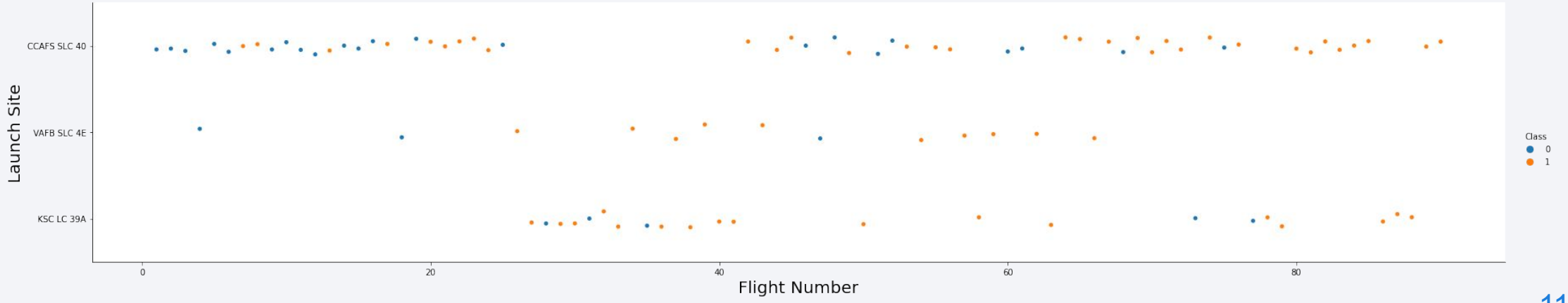
EDA with Data Visualization

We use scatter plots to see the relationship, if any, between our attributes

We then use bar graphs and line graphs to further explore any relations that seem to have a pattern.

We use feature engineering to to create dummy categorical variables for further analysis.

See the EDA - Data Visualization notebook for further details.



EDA with SQL

To further explore the data, we used the following SQL queries on the data:

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string ‘CCA’.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the 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 sites names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

See the EDA - SQL notebook for details.

Build an Interactive Map with Folium

- We add launch markers to an interactive map to visualize the geolocations of the launches.
- We took the latitude and longitude coordinates at each launch site and added a circle marker around each launch site with a label of the name of the launch site.
- We then classified the launches as either failure or success with Red and Green coloring respectively. We markers of these launches in the map to a MarkerCluster().
- We then use Haversine's formula to calculate the distance of the launch sites to various landmark to understand how close the launch sites were to railways, highways, coastlines, and cities.

See the EDA - Folium notebook for more details.

Build a Dashboard with Plotly Dash

We built an interactive dashboard with Plotly that allows a user to dynamically view the data across different visualizations.

We plotted pie charts showing the total launches by a certain sites.

We plotted a scatter graph showing the relationship of Outcome vs. Payload Mass (Kg) for the different booster version.

See the source file, spacex_dash.py, for details.

Predictive Analysis (Classification)

We load the data using a combination of numpy and pandas, transform the data, and then split our data into training and testing sets.

We built a selection of machine learning models and optimize and tune to different hyperparameters using GridSearchCV.

We evaluate the models by accuracy and improve the models.

At the end, we find the best performing classification model given our parameters.

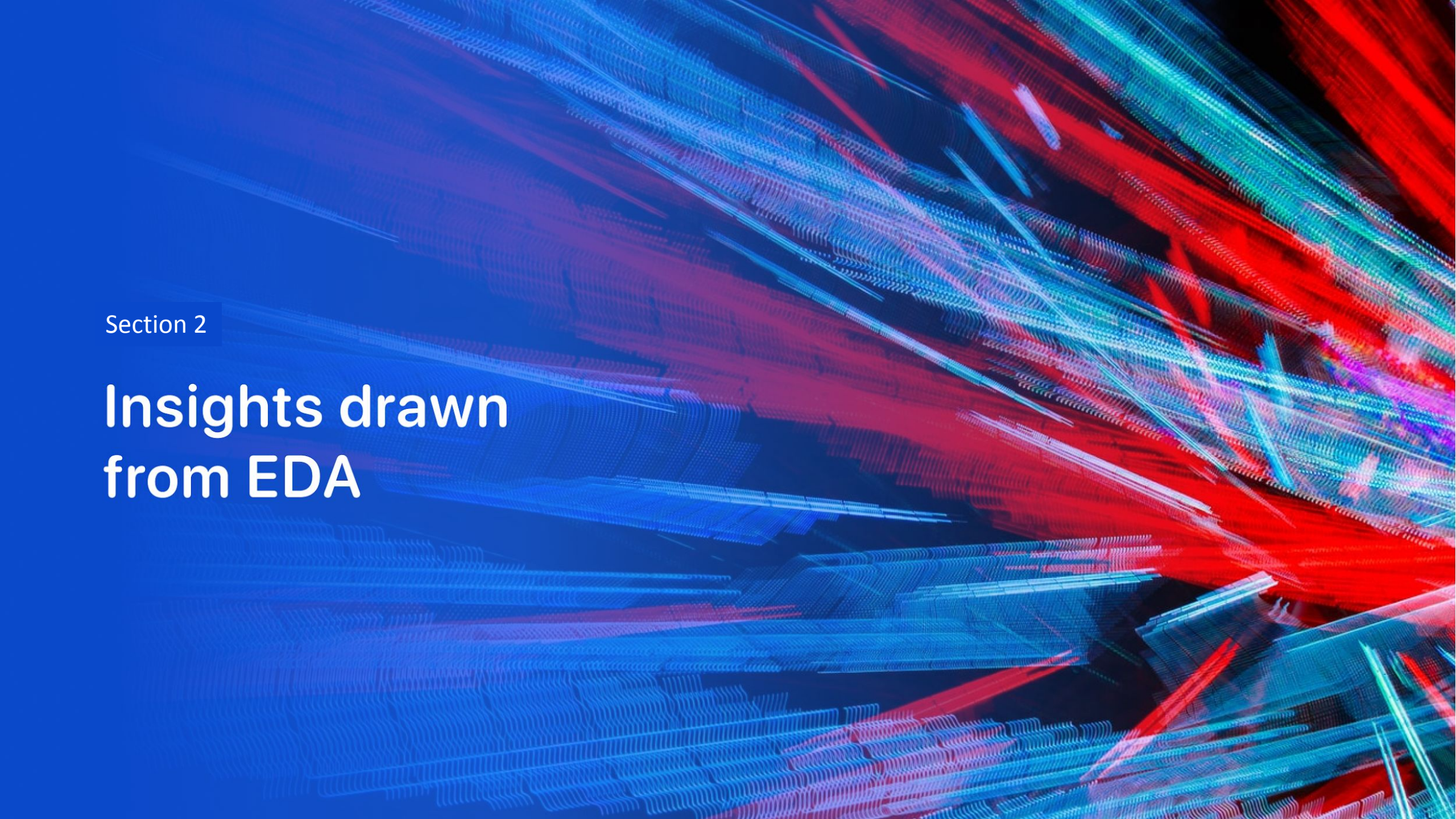
See the Machine Learning notebook for details.

Results

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

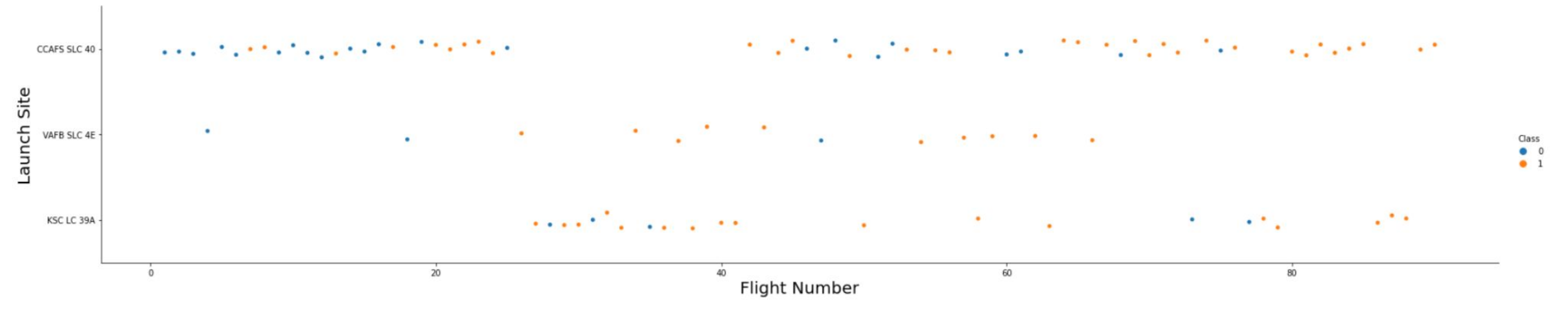
Section 2

Insights drawn from EDA



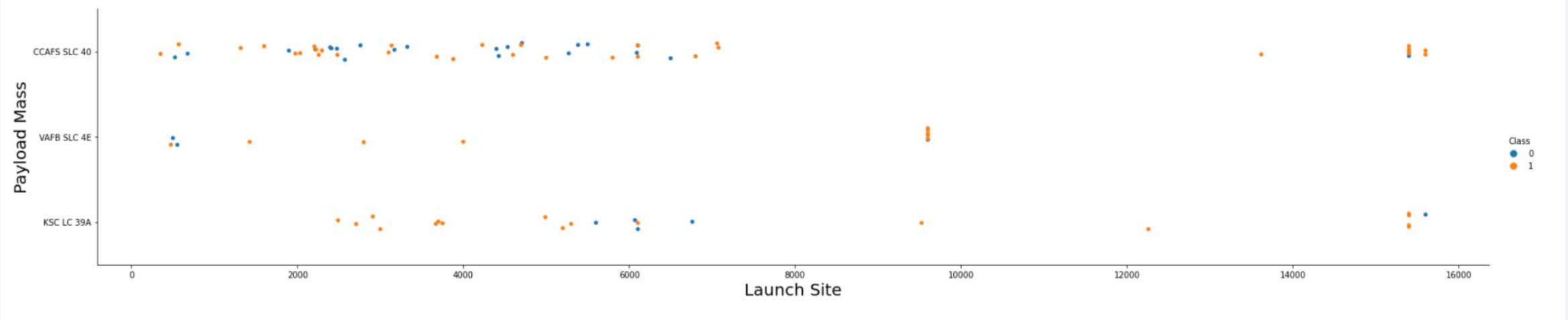
Flight Number vs. Launch Site

From the scatter plot, we can see the that the greater the flight amount at a launch site, the higher the success rate for the site. However, we that the CCAFS SLC 40 site tend to deviate from this pattern.



Payload vs. Launch Site

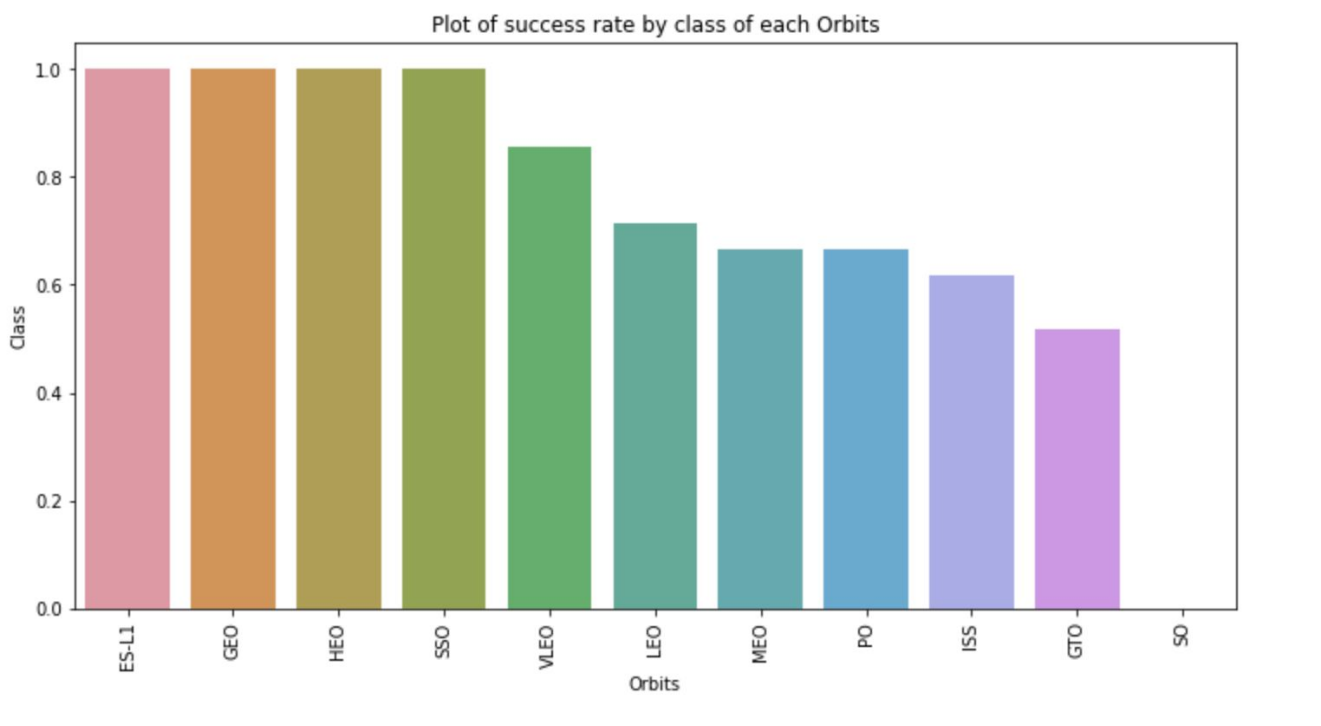
From the scatter plot we can see in general the higher the payload mass the higher the success rate of launches for a launch site.



Success Rate vs. Orbit Type

We can see from the bar chart the differing success rates in descending for each orbital.

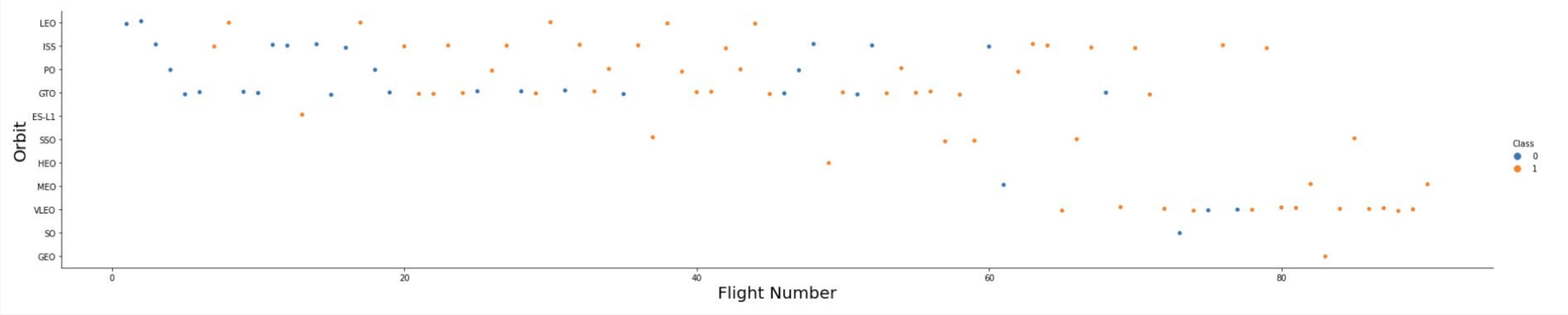
Of note though is that some of the high success rates Orbits, like GEO, SO, HEO and ES-L , only had one launch so further data may be required before these success rate can be more generalized.



Flight Number vs. Orbit Type

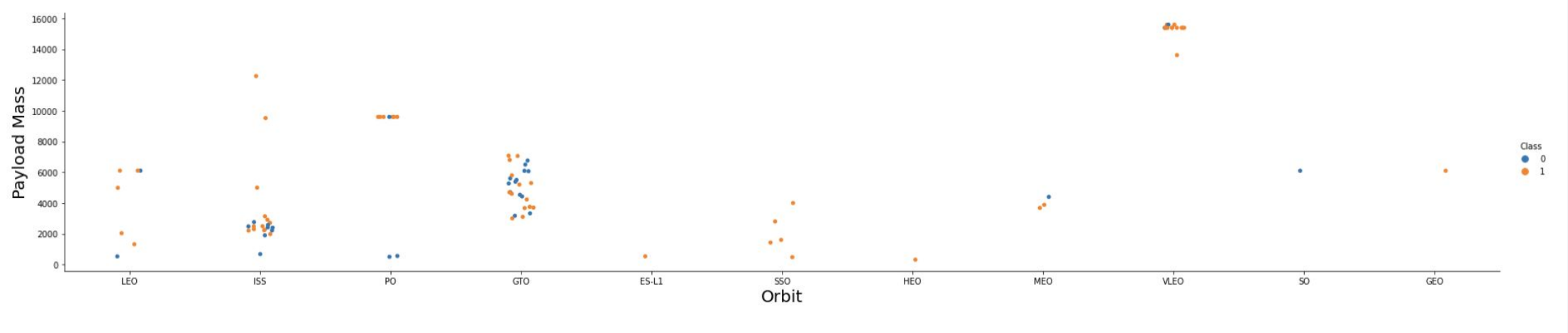
We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

This scatter plot shows that generally, larger flight numbers for an orbits also show greater success rates, except for GTO orbit which doesn't seem to have a pattern.



Payload vs. Orbit Type

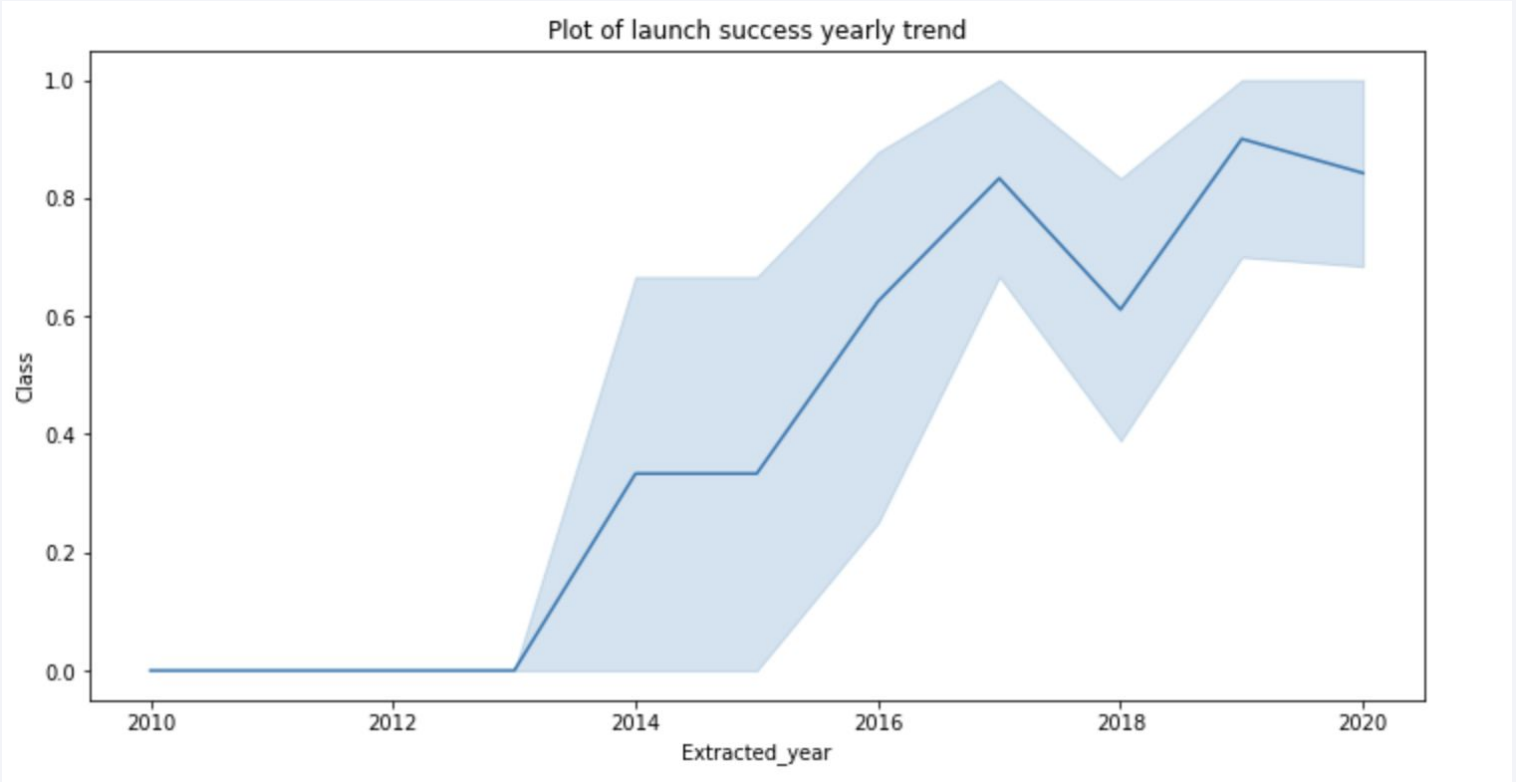
We see that larger payloads seem to coincide with higher success rates for orbits.



Launch Success Yearly Trend

We see that there is an increasing trend of success rate from 2013 to 2020.

There is a noticeable dip around 2017 - 2018, but the trend corrects back up in 2019.



All Launch Site Names

We use DISTINCT in a query to find the specific launch sites. We see a total of 4 unique launch sites.

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

In [10]:

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

We query to find 5 records where launch sites begin with `CCA`.

In [11]:

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

We use SUM in a query to find the total payload mass of carriers from NASA.
We find that the total payload mass as 45596 kg.

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

We use AVG in a query to find the total payload mass for boosters version F9 v1.1. We find that the average payload mass as 2928.4 kg.

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

We find the date of the first successful landing outcome on ground pad using MIN.

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We find the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 using query conditionals.

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

We run queries to find the total number of successful and failure mission outcomes using COUNT.

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

Boosters Carried Maximum Payload

We list the names of the booster which have carried the maximum payload mass using a subquery.

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

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

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

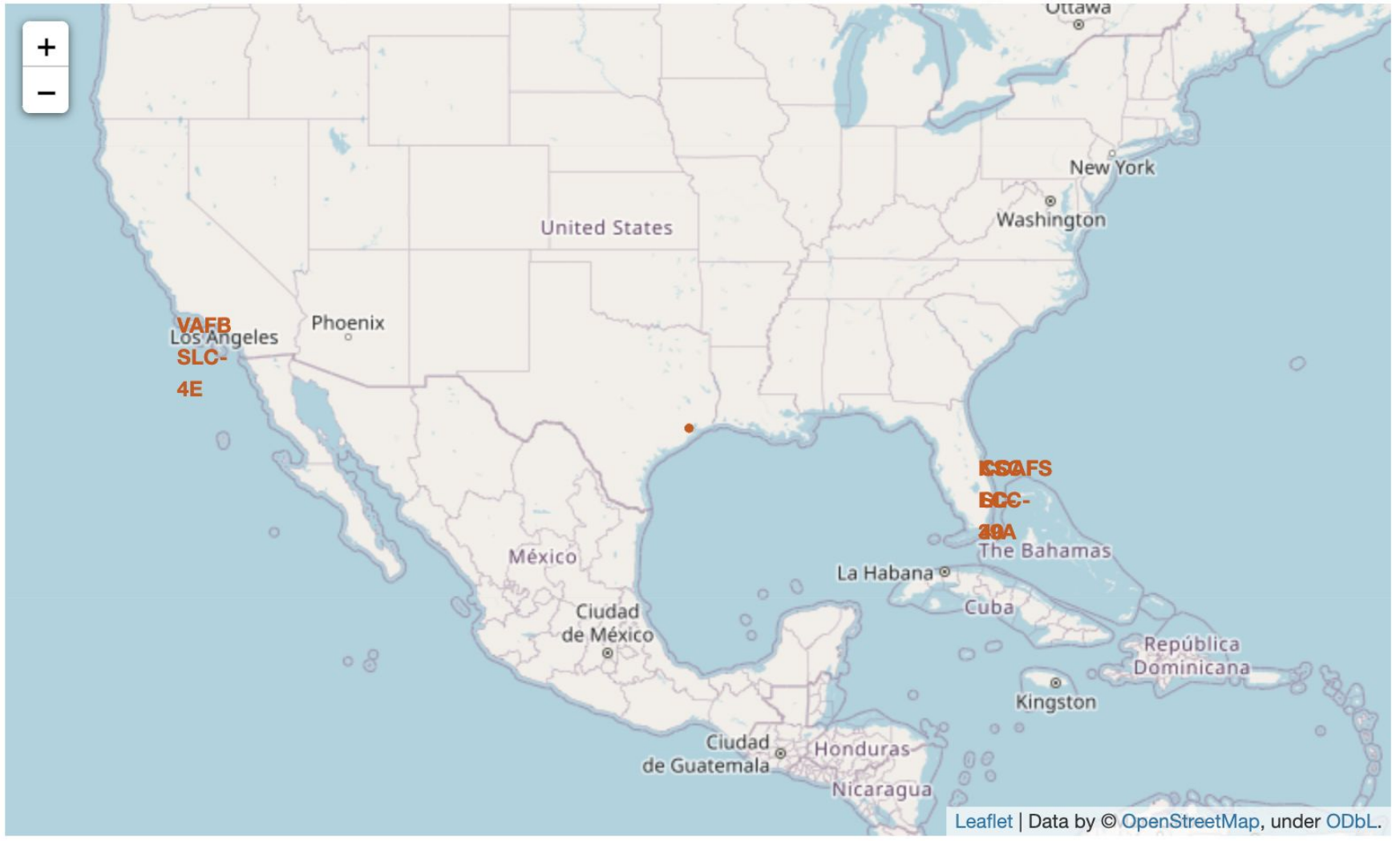
We rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

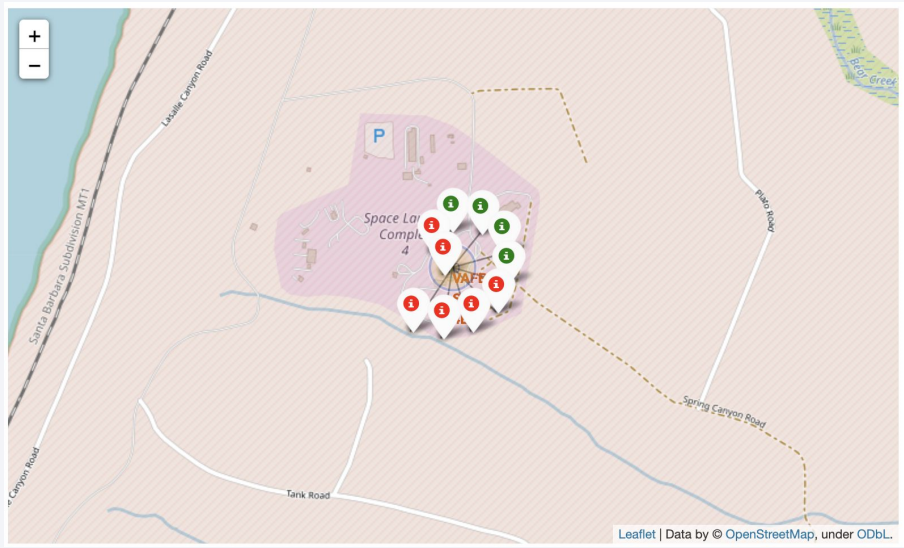

Section 3
Launch Sites
Proximities Analysis

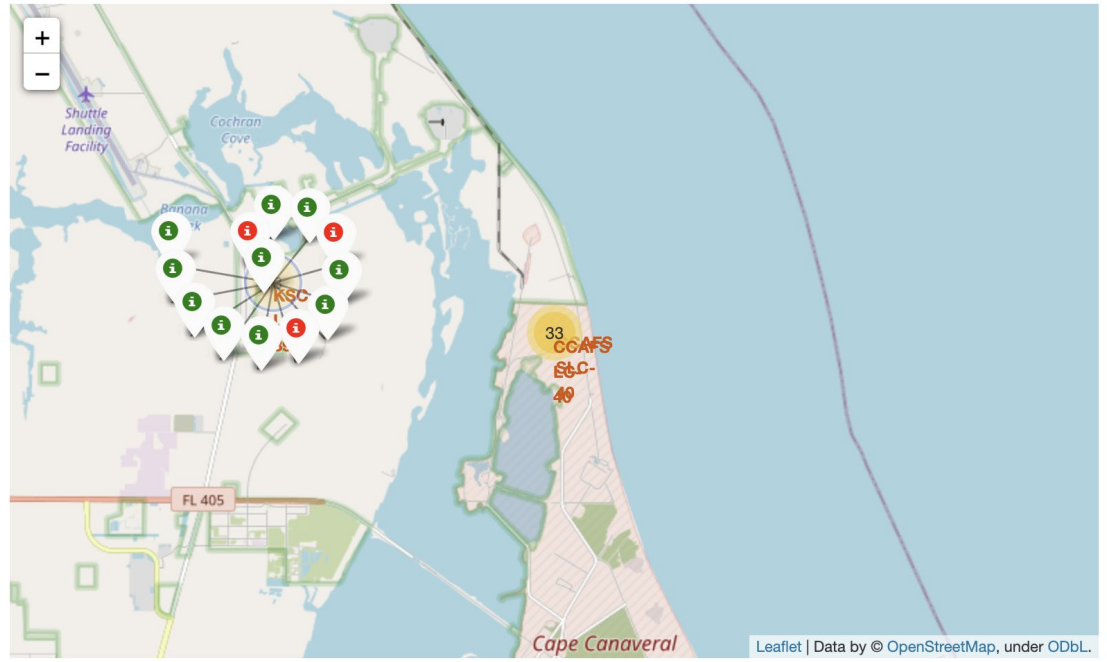
Launch Site Locations



The launch sites are primarily located in the U.S. along coastlines in Florida and Southern California

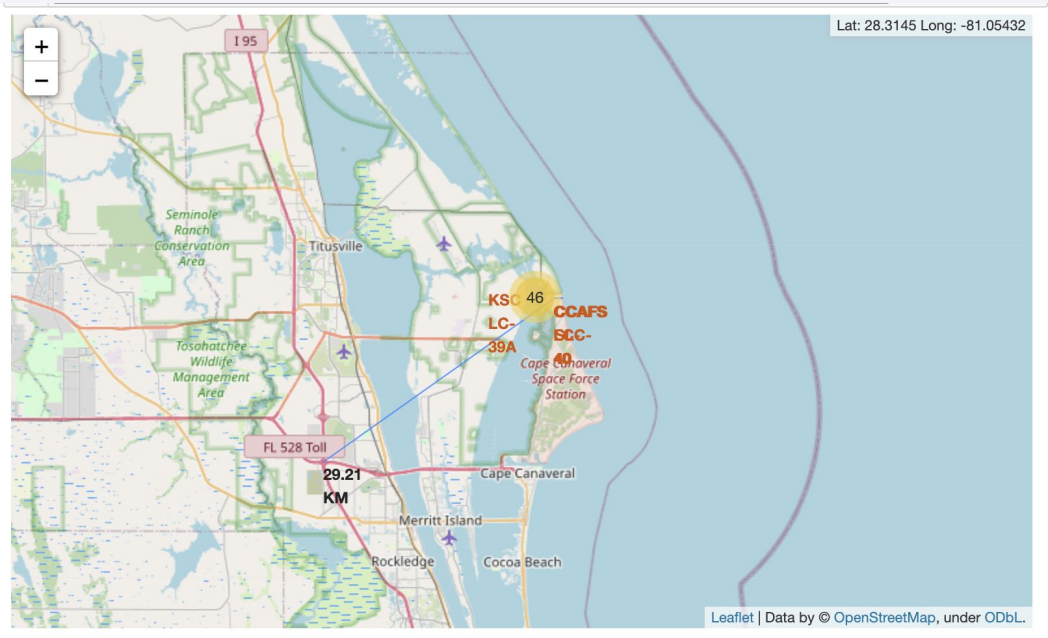
Launch Sites with Success, Fail Markers

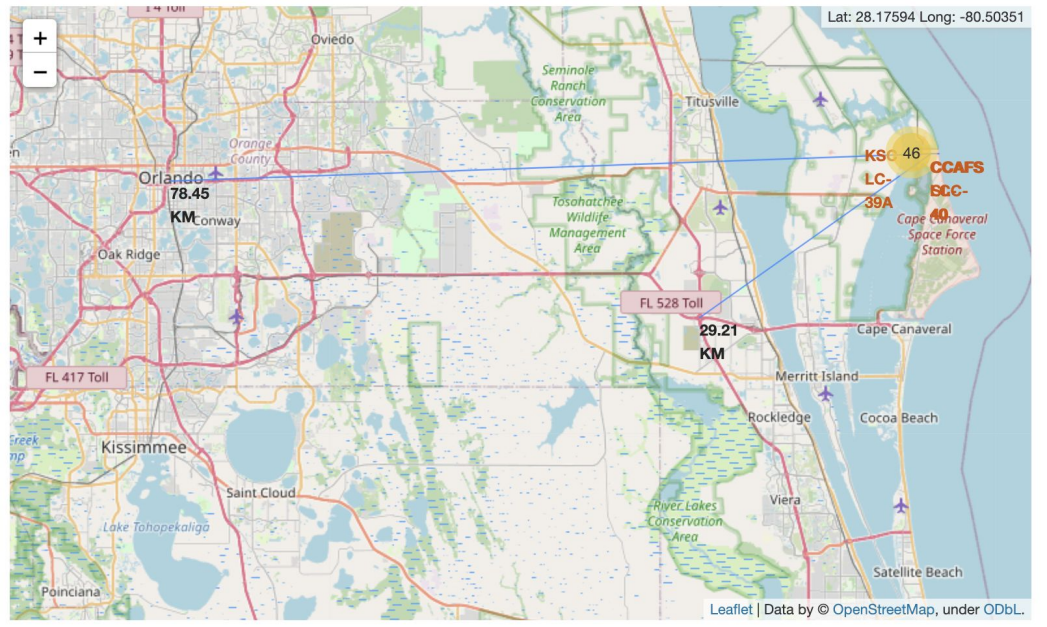




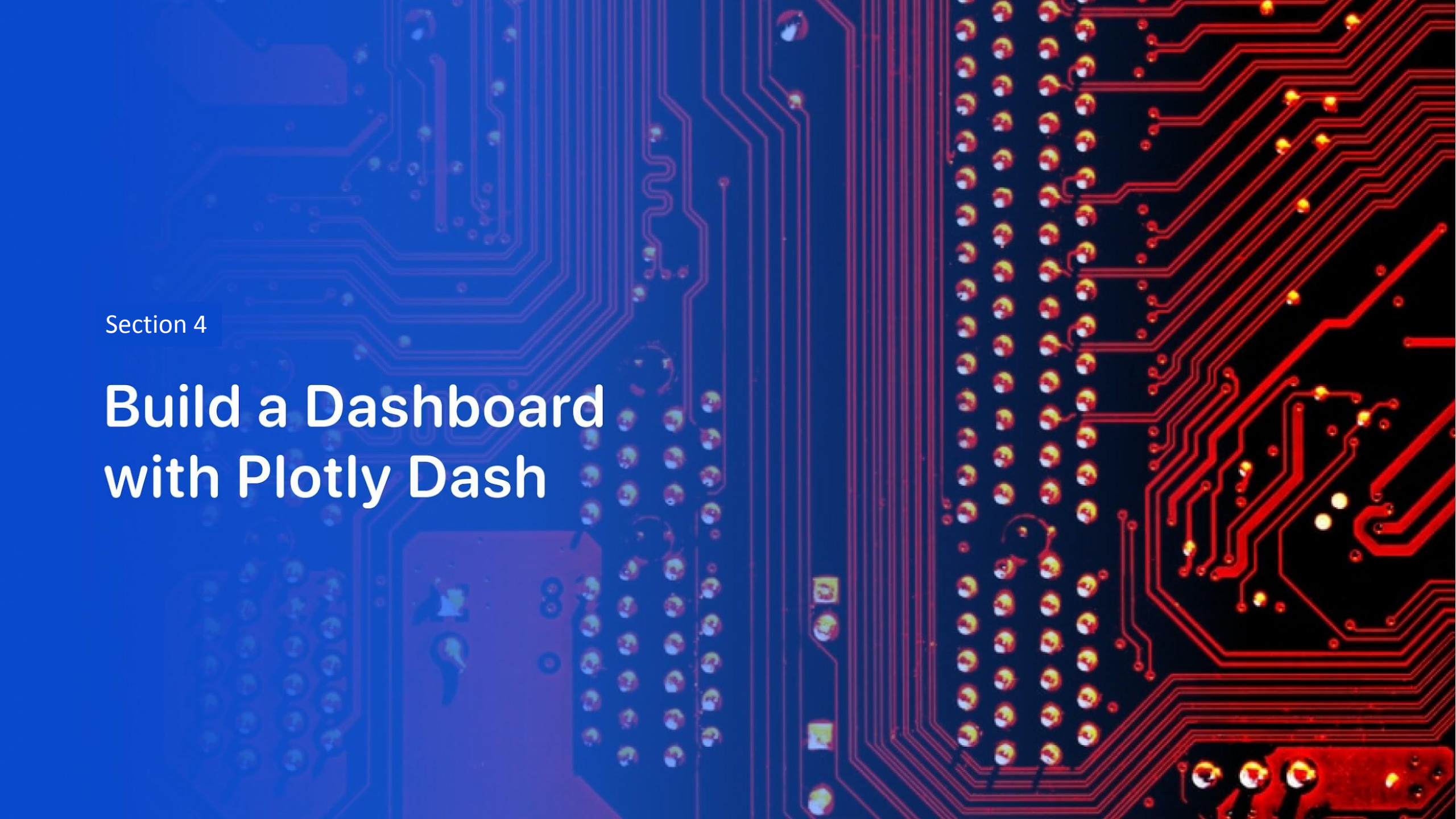
We see the launch sites with markers plotted for successful (green) and failed (red) launches.

Distance to Proximities

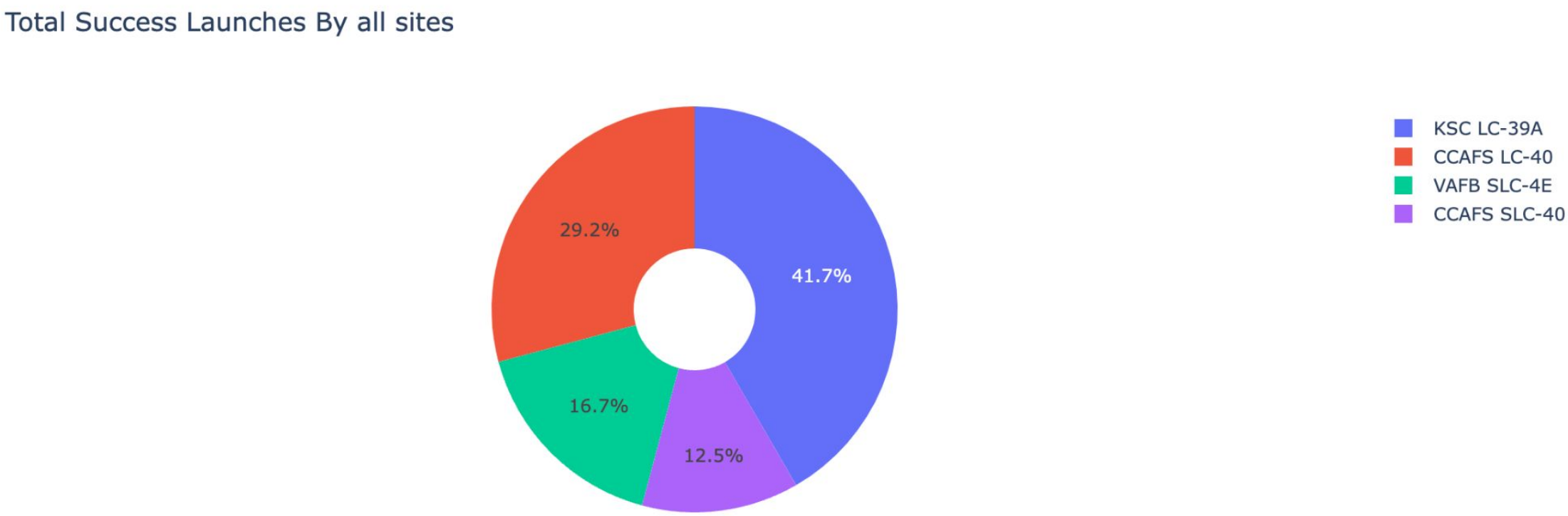




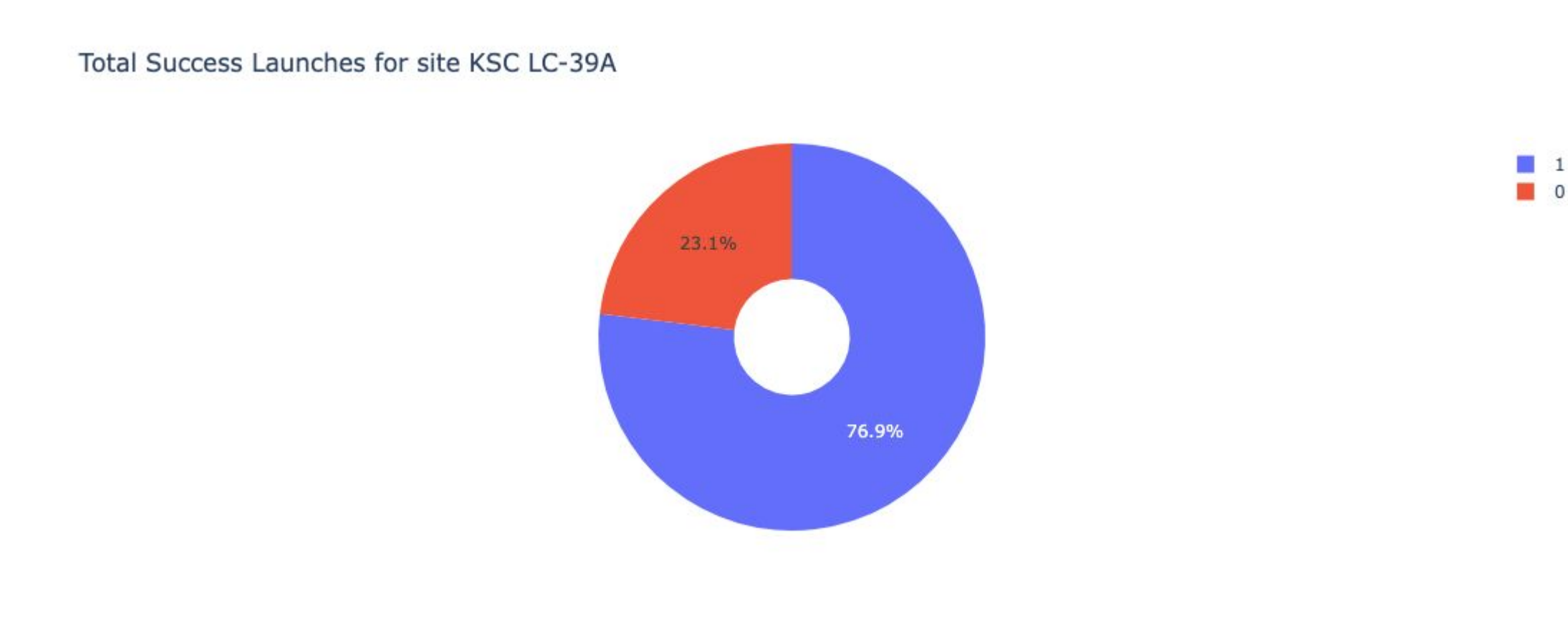
We show the distance between different proximity locations like a nearby highway and a Florida city.


Section 4
Build a Dashboard with Plotly Dash

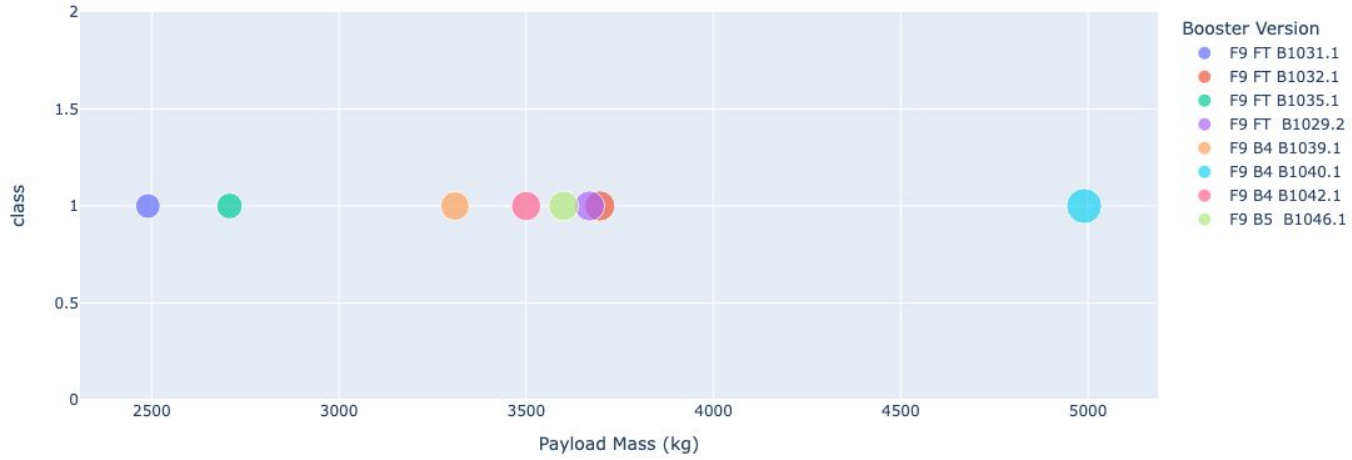
Success Rates by Launch Site

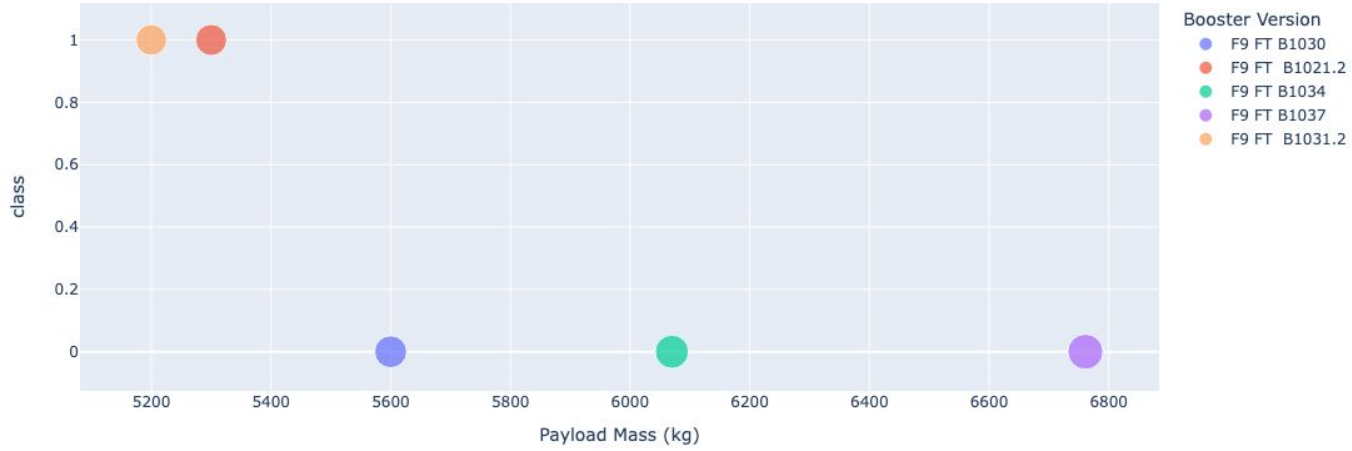


Launch Success Ratio for KSC LC-39A



Success Rate for Low Weight vs High Weight Payload





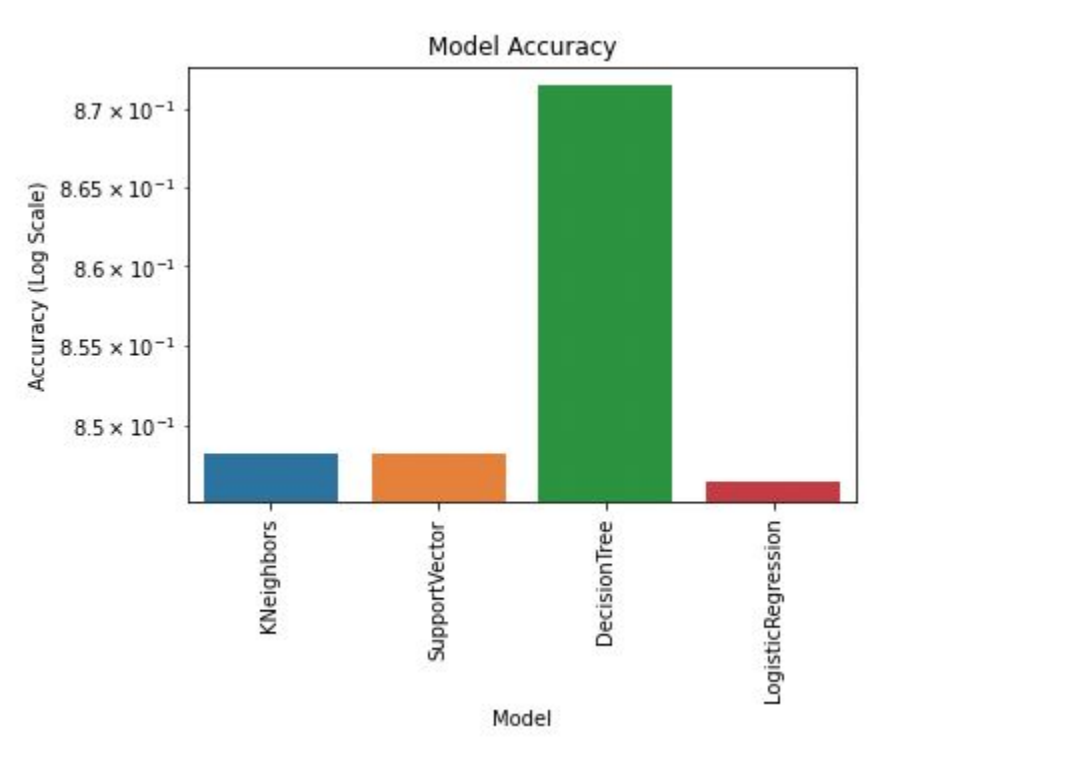
On the graph we see payload mass vs success rate. The top graph shows success rate for lower weight payloads of 5000kg or less. The bottom shows success rate for higher weight payloads of greater than 5000kg.

We see that lower payloads have a higher success rate compared against heavier payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



We graph the accuracy of our models against each in bar chart with accuracy on a log scale to better visualize the differences.

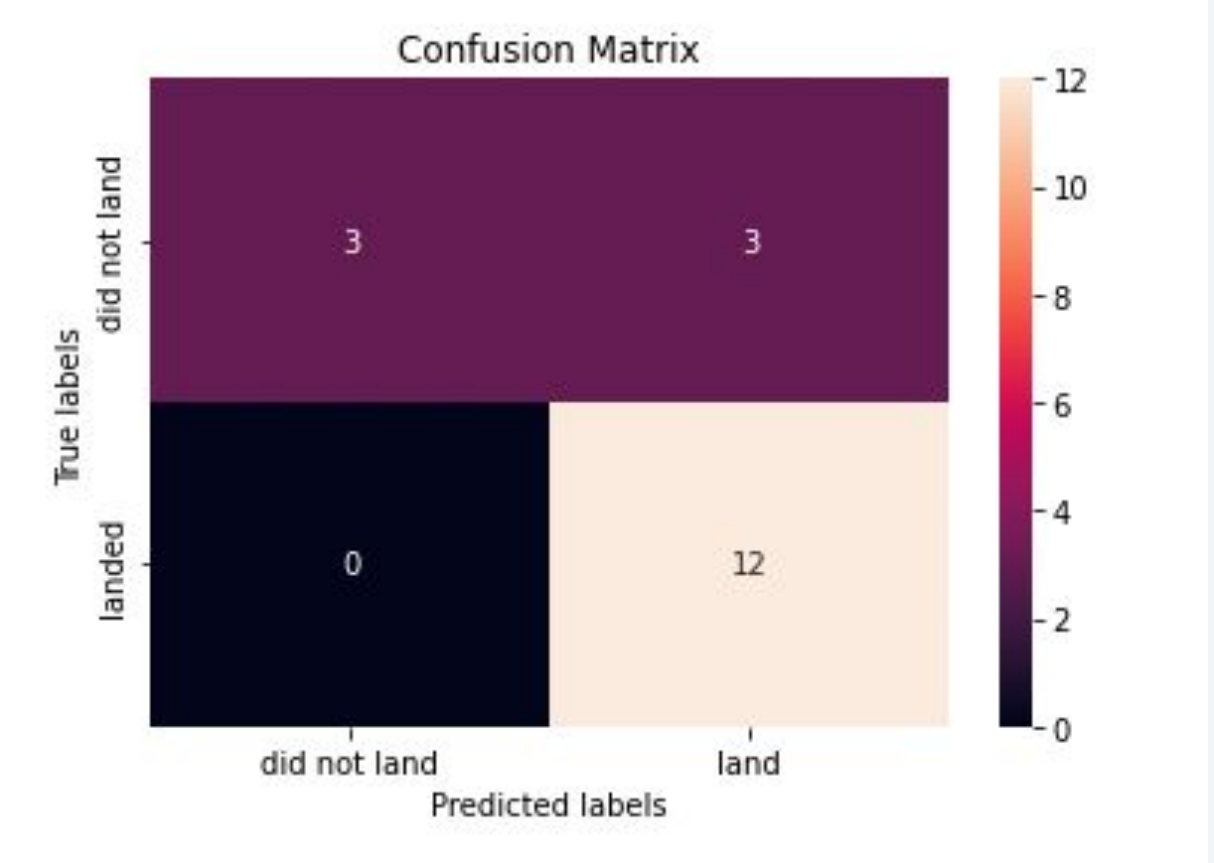
As can be seen, the Decision Tree model had the best accuracy with a score of 0.8714 and params: {'criterion': 'entropy', 'max_depth': 10, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}.

Confusion Matrix

We can display the confusion matrix for the decision tree.

The matrix shows that the DecisionTree classifier can distinguish between the different landing classes.

However, there is also a problem with false positives (unsuccessful landing marked as successful landing).



Conclusions

From our analysis of the data we have found a few key takeaways:

- Lower weighted payloads typically had higher launcher success rates than higher weighted payloads
- In general, the larger the amount of launches at site, the greater the success rate.
- Orbits ES-I1, GEO, VLEO, HEO, and SSO generally had higher success rates than other orbits.
- KSC LC-39A has the most successful launches of the other sites.
- Based on the data, the Decision tree classifier had the best accuracy and is the best machine learning algorithm for this scenario compared against our other models.

Appendix

All source code for the linked notebooks and dashboard can be found in the following GitHub repo:

https://github.com/axcruz/data-capstone

Thank you!