Winning Space Race with Data Science

Daniel Bluff
06/02/2022

Outline

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

Executive Summary

Using data from SpaceX API and webscraping, we're able to understand which factors play a role in success rate of a launch. This was achieved by data wrangling, interactive plots and predictive analytics.

From this, we're able to predict an 83% success rate of any launch going forward.

Introduction

SpaceY is on a mission to compete with SpaceX.

We will determine:

- The price of each launch by gathering and understanding data from SpaceX
- If SpaceX will reuse the first stage allowing us to understand costs for future launches

Section 1

Methodology

Data Collection

- Data was obtained in several ways; through the SpaceX API and webscraping.
- Over the next few slides, we will cover each in detail to understand what was obtained and how.

Data Collection – SpaceX API

- Lots of data was obtained through the SpaceX API.
- This was achieved through a requests.get method within Python.
- The data obtained was in json format so simple cleaning operation took place to convert this to a dataframe.
- Further data wrangling was used on the data set to ensure optimal cleaning

Flowchart:

Obtain json from requests.get(spacex_url)

Use functions such as getLaunchSite(data) to extract data about specific areas

Convert to a dataframe from a dictionary

Perform data wrangling (discussed further on)

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%201%20-%20Data%20Collection%20-%20API.ipynb

Data Collection - Scraping

- Falcon 9 launch records were obtained through Wikipedia by webscraping.
- This was done using the BeautifulSoup library in Python.
- A number of attributes were extracted from the tables in the provided URL.

Flowchart:

Extract data using requests.get(static_url)

Parse the data into a readable format using BeautifulSoup: BeautifulSoup(response.text, 'html.parser')

Extract key elements from the data and save as a dataframe for later use.

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%202%20-%20Data%20Collection%20-%20Webscraping.ipynb

Data Wrangling

- Data wrangling took place to ensure the data was in the cleanest possible state before having to perform any analysis on it. This is to avoid any issues with modelling or overall outcomes.
- Imputation (the process of replacing missing values with actual data) was achieved on the payload mass.
- This was done by finding the average payload mass from the other launches.
- The only other field with missing data was the landing pad but unfortunately this couldn't be avoided.

Github (appeared in step 1 within the API section and data wrangling section):

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%201%20-%20Data%20Collection%20-%20API.ipynb

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%203%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

Several visualizations were created to understand the data more:

- A scatter plot of payload mass vs flight number to understand the likelihood of the first stage completing. In fact, as the flight number increases, the more likely the stage is to complete.
- A scatterplot of launch site vs flight number to understand if launch site has any impact on successful launches. It shows some sites have a higher success rate than others but CCAFS SLC 40 is the preferred site for launches.
- A barplot to understand the relationship between orbit type and the success rate.
- A lineplot to show the success rate over time so we can understand the likely impact on SpaceY.
- Plus several other graphs which will be available.

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%204%20-%20EDA%20Data%20Visualisations.ipynb

EDA with SQL

- I used magic SQL to query within a python notebook.
- The queries used a combination of common select statements, joins, where clauses and mathematical operations like min and sum.

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%205%20-%20EDA%20SQL.ipynb

Build an Interactive Map with Folium

- Using Folium, I mapped the locations of the different launch sites.
- Each of these sites were a marker on the map
- Circles were added to show the success/failures of each launch
- Line were also added to show the proximity to other locations.

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%206%20-%20Folium.ipynb

Build a Dashboard with Plotly Dash

- In the interactive dashboard you're able to filter by launch site and payload mass.
- You can see the success rate for each site and also look at the correlation between payload and success rate.

Github:

https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%207%20-%20Dash.py

Predictive Analysis (Classification)

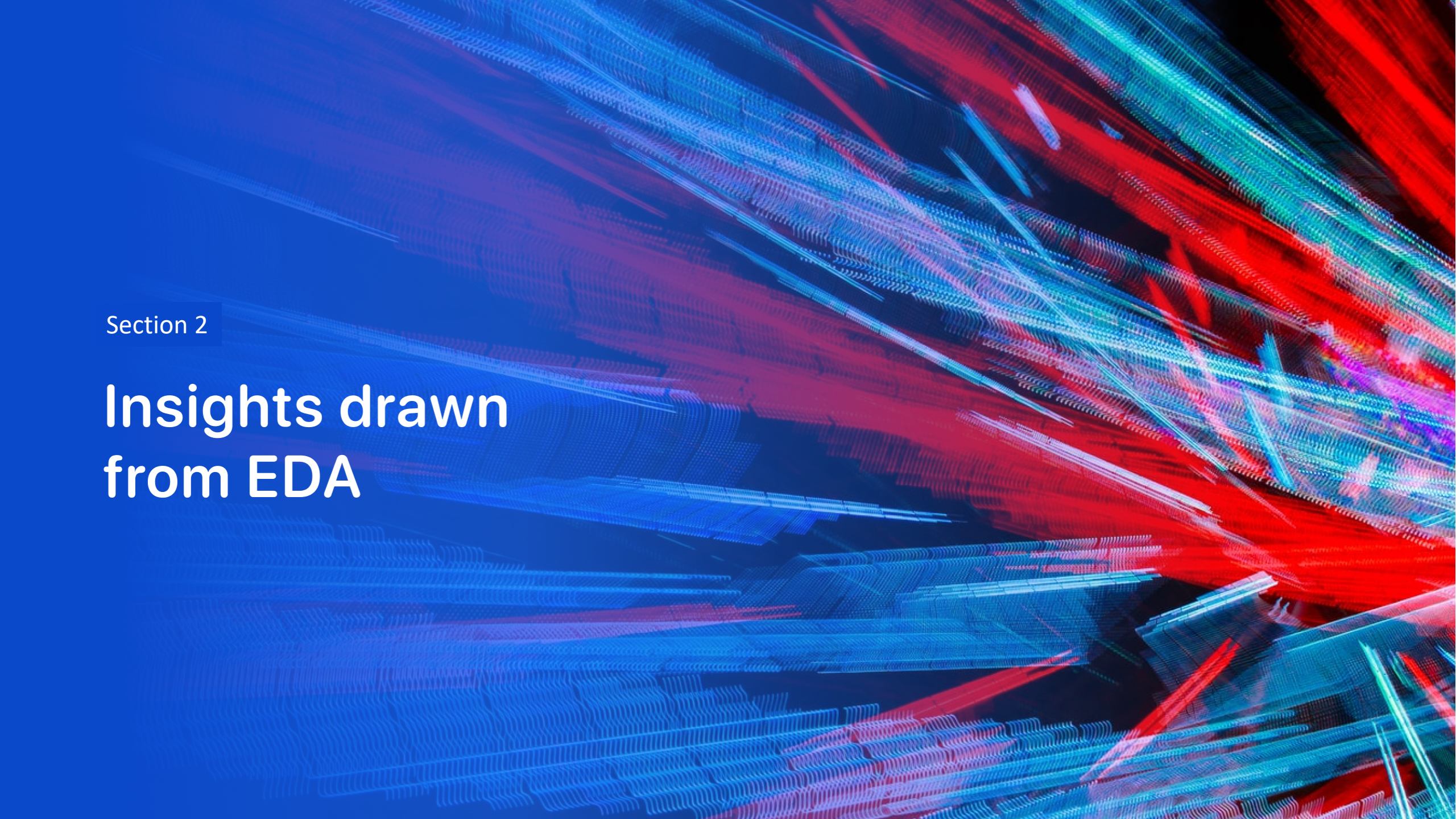
- First, I standardized the data to get all data into the same unit.
- Then, I split the data into a training set (80%) and testing set (20%).
- For each model in logistic regression, support vector machine, decision tree and K-nearest neighbour I used a grid search to understand the best parameters for each model.
- Then using the score, I could see which model performed the best in predicting whether a launch was to land or not.

Github:

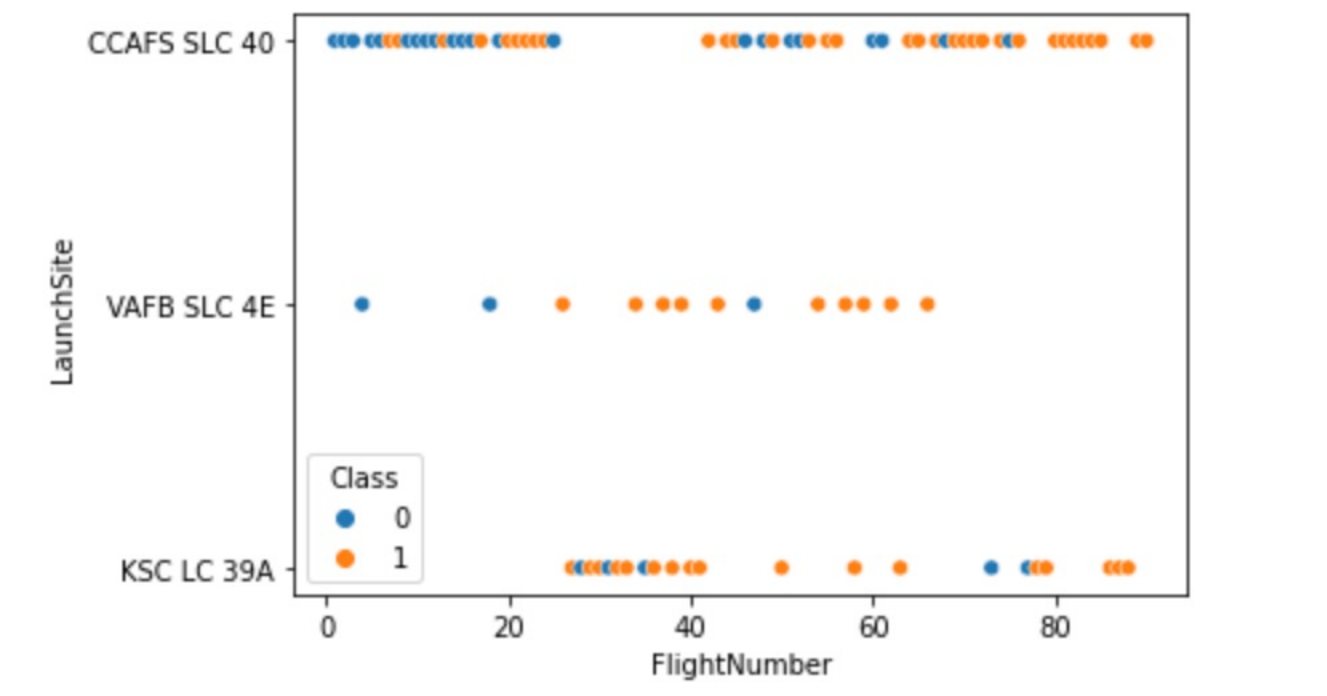
https://github.com/DannyBluff/IBM-Data-Science---Coursera/blob/master/Step%208%20-%20Predictive%20Analysis.ipynb

Results

- The KSC LC-39A launch site has the best success rate of 77%.
- With an 83% accuracy, we're able to determine whether a first stage launch was successful

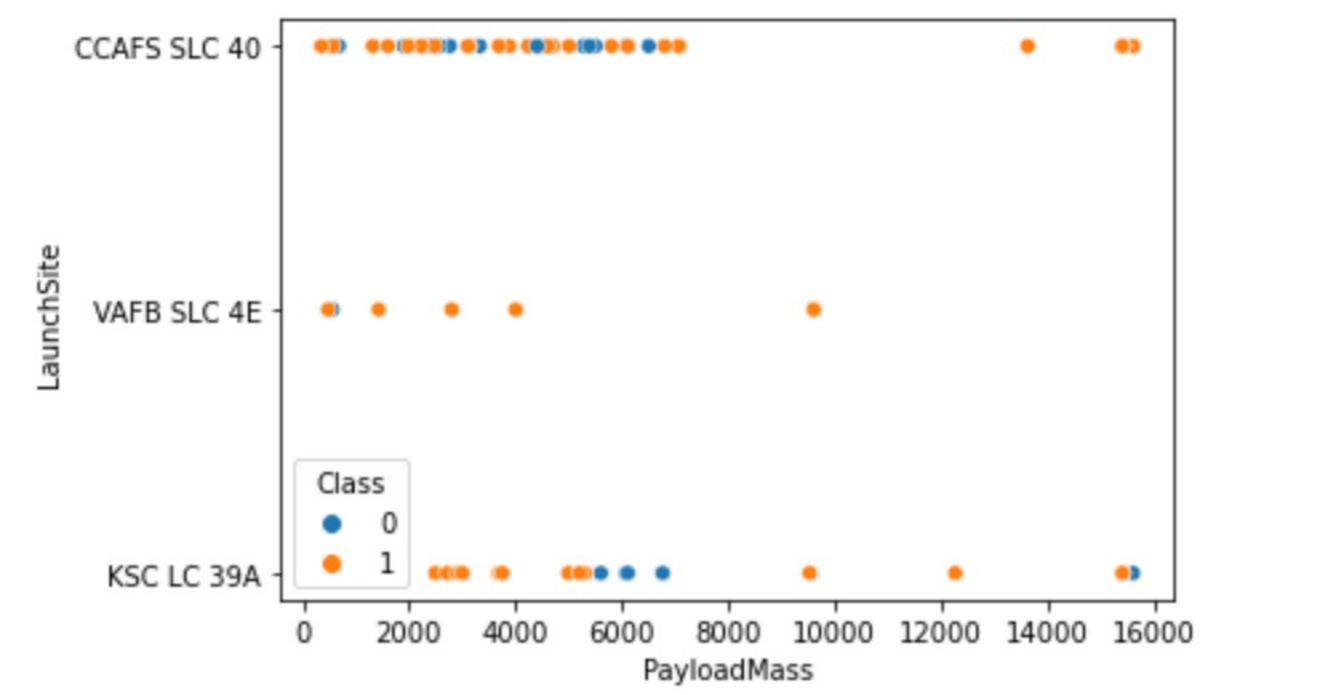

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



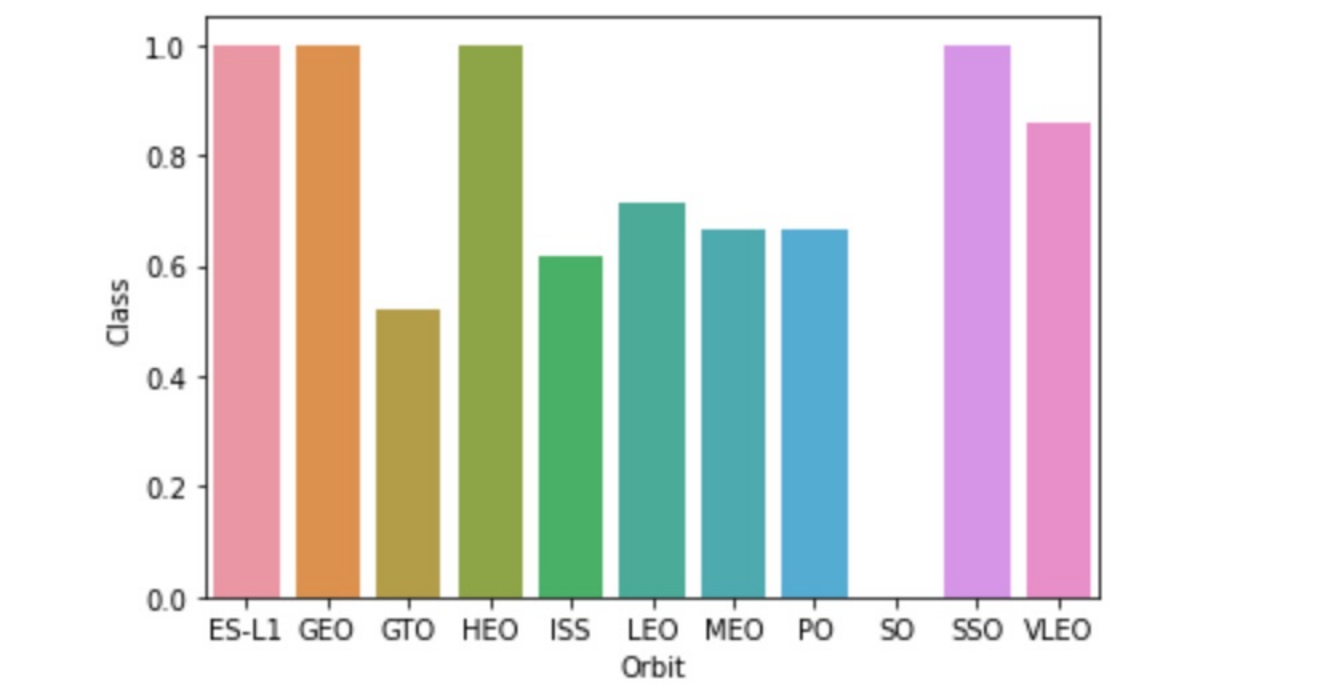
CCAFS SLC 40 was the most popular launch site which had a number of failures to begin with but improved over time. The other sites were less popular.

Payload vs. Launch Site



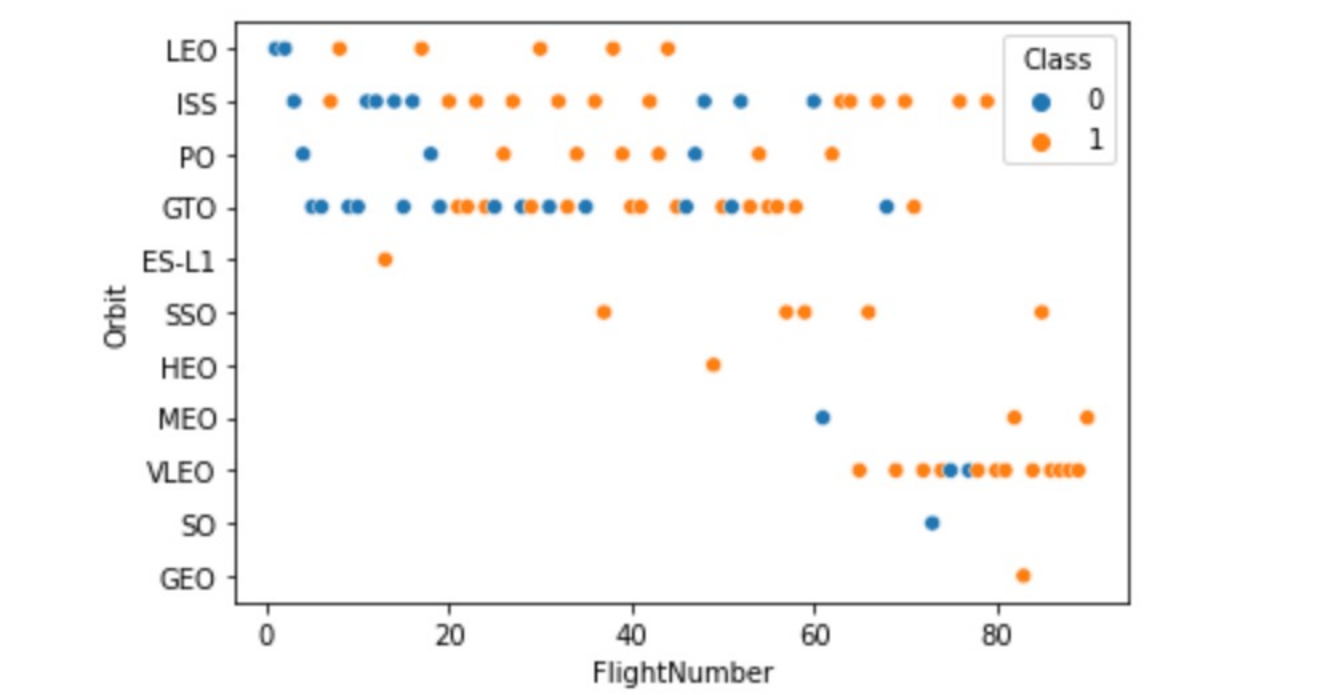
Whilst most launches were between 2,000-8,000kg, in VAFB SLC 4E there were no launches with a payload mass greater than 10,000 kg.

Success Rate vs. Orbit Type



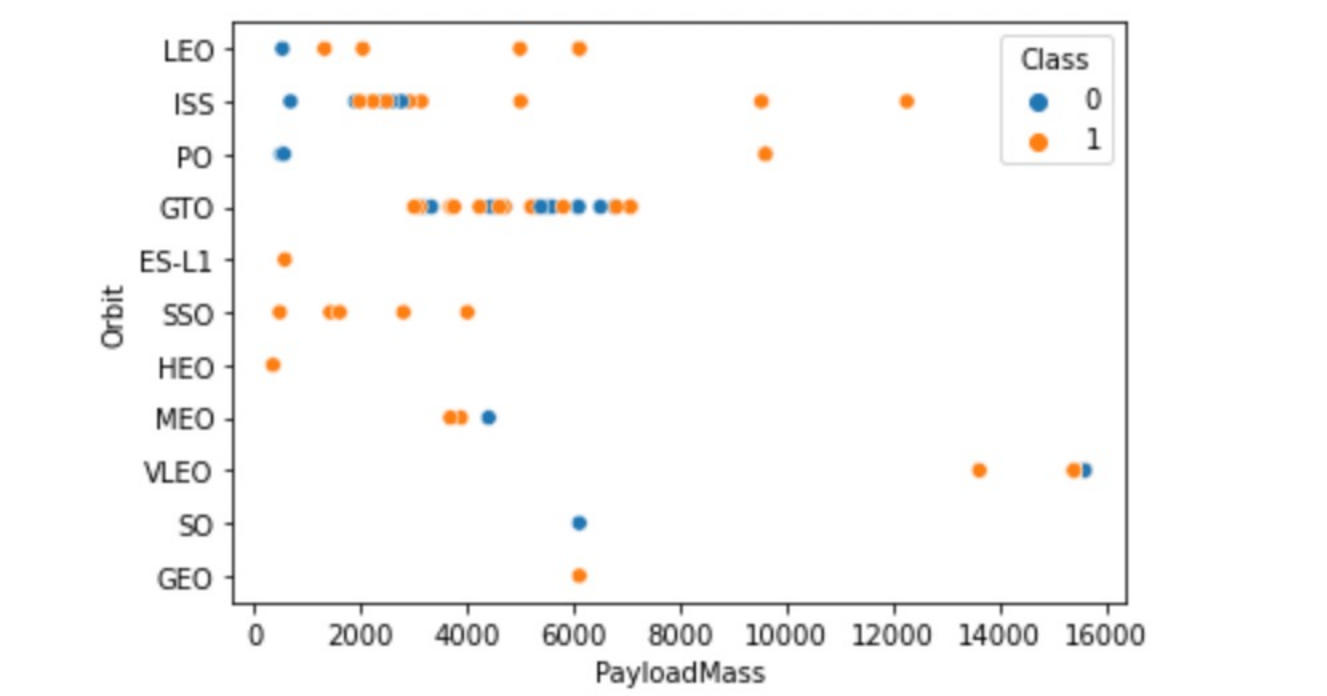
ES-L1, GEO, HEO, SSO all have best success rate.

Flight Number vs. Orbit Type



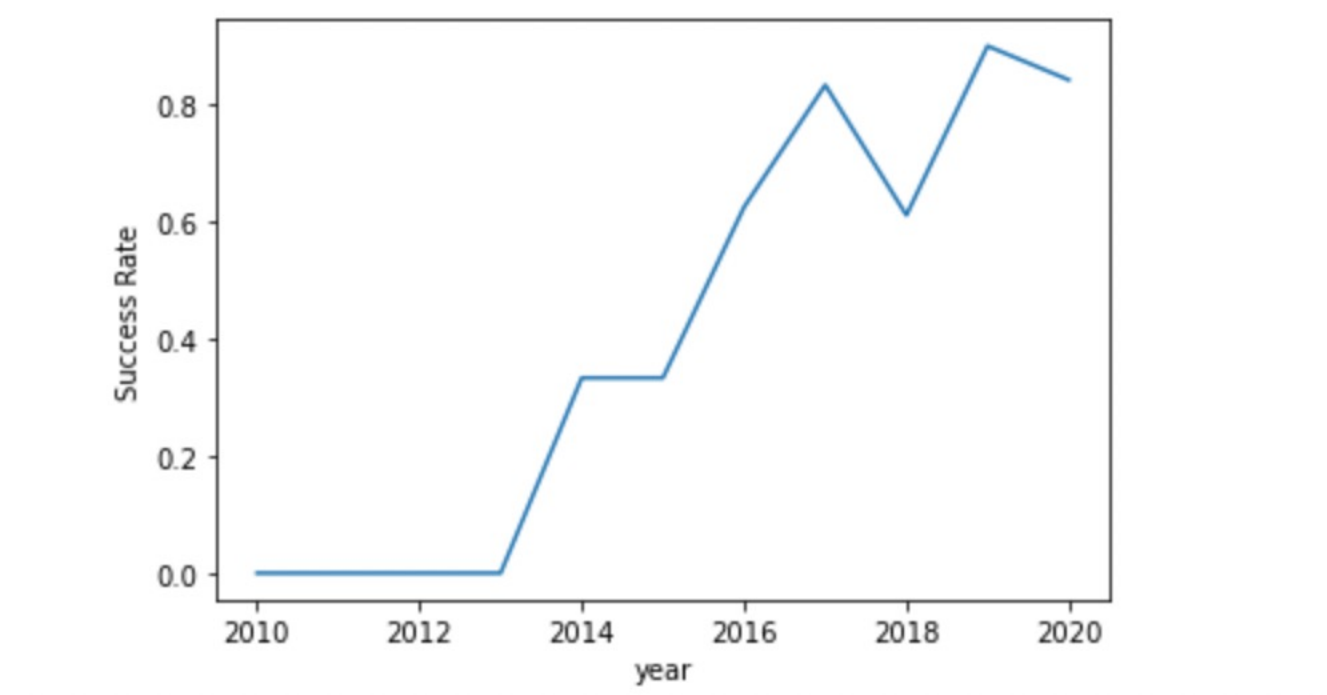
The LEO orbit shows that there is a relationship between success and number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads the successful landing rate are more for LEO, PO and ISS.
However, for orbits like GTO we can’t determine a success rate based on payload mass.

Launch Success Yearly Trend



Success rate has improved over time, albeit with a drop in 2018, but does appear to be plateauing.

All Launch Site Names

All launch site names are provided here using the following code in python:

%%sql

select distinct launch_site from SPACEXDATASET

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

%%sql

select * from SPACEXDATASET where launch_site like 'CCA%' limit 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

- Calculate the total payload carried by boosters from NASA

%%sql

select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)'

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

%%sql

select cast(cast(sum(payload_mass__kg_) as float) / cast(count(*) as float) as float) as average_payload_mass from SPACEXDATASET where booster_version = 'F9 v1.1'

average_payload_mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%%sql

select min(DATE) as first_date from SPACEXDATASET where
landing__outcome = 'Success (ground pad)'
```

first_date
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 SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql

select sum(case when mission_outcome like '%Success%' then 1 else 0 end)
as successes, sum(case when mission_outcome like '%Failure%' then 1 else 0
end) as failures from SPACEXDATASET
```

successes	failures
100	1

Boosters Carried Maximum Payload

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

```
%%sql
select distinct booster_version
from SPACEXDATASET s
right join (
        select max(payload_mass__kg_) as max_payload_mass
        from SPACEXDATASET
) as tbl1
on s.payload_mass__kg_ = tbl1.max_payload_mass
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

%%sql

select DATE, booster_version, launch_site from SPACEXDATASET where landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	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

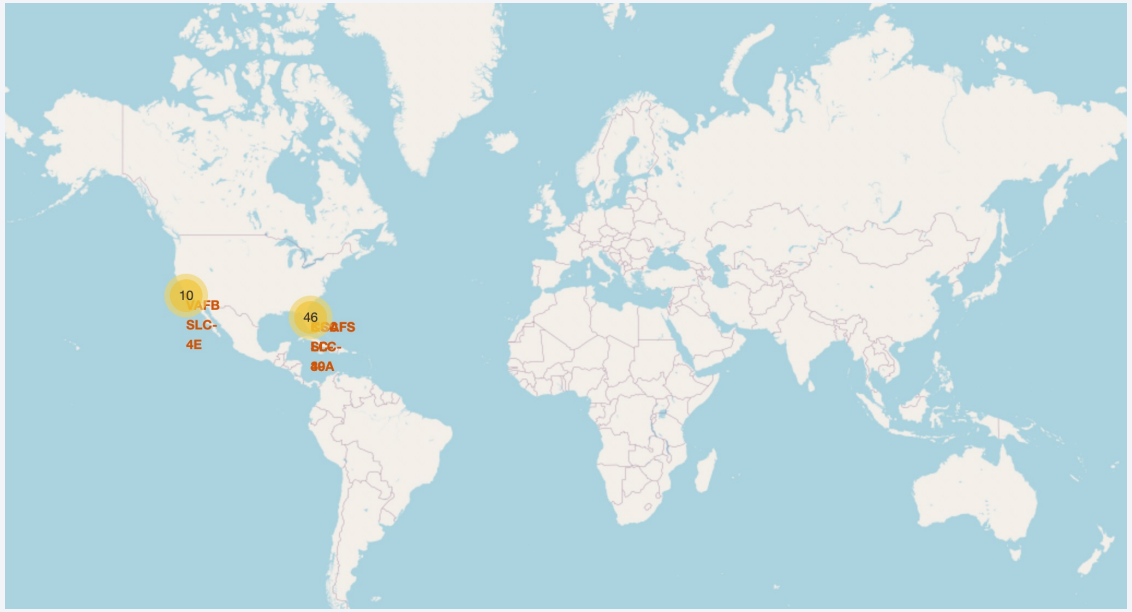
```
%%sql

select landing__outcome, count(*) as occurrences from SPACEXDATASET where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by occurrences desc
```

landing__outcome	occurrences
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


Section 3
Launch Sites
Proximities Analysis

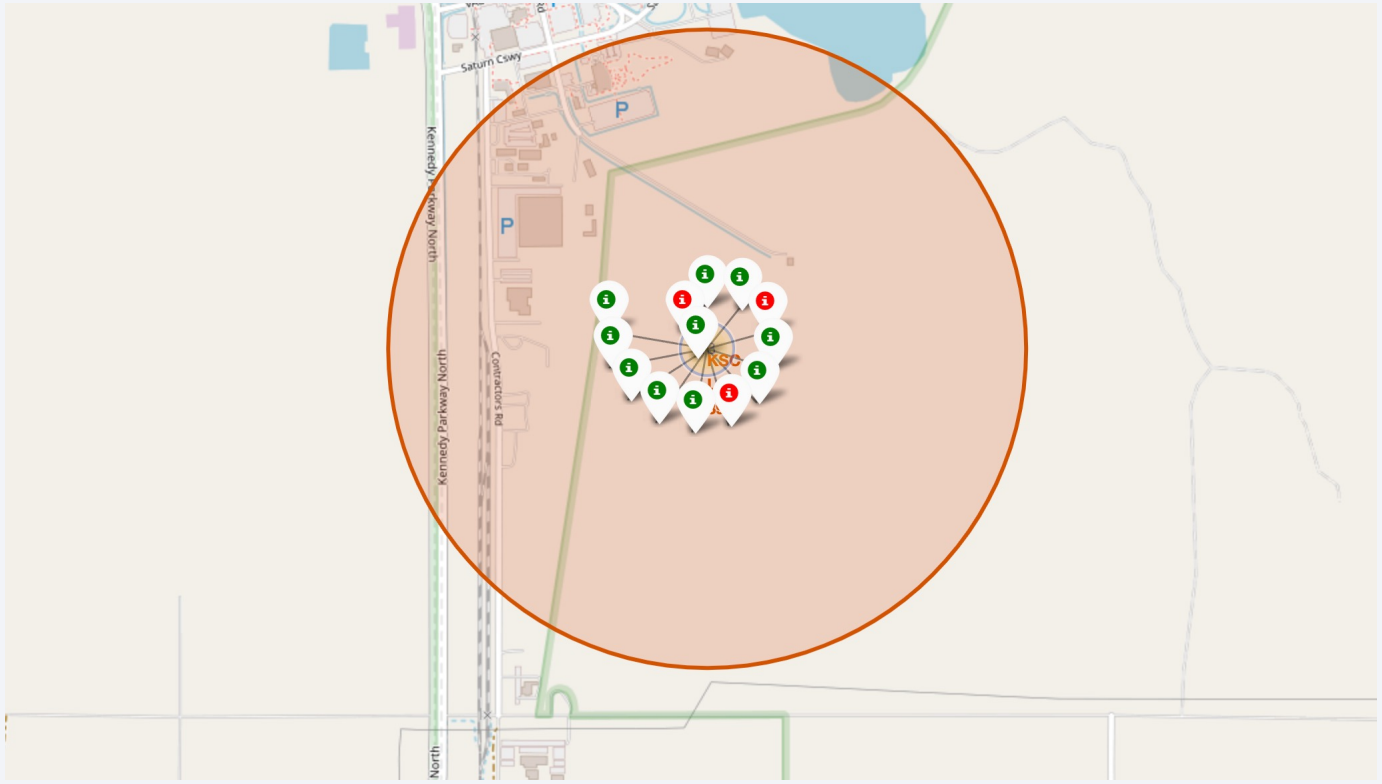
Launch Site Map

- All launch sites were located in the USA.
- The site were situated on the east coast in Florida or on the west coast in California.



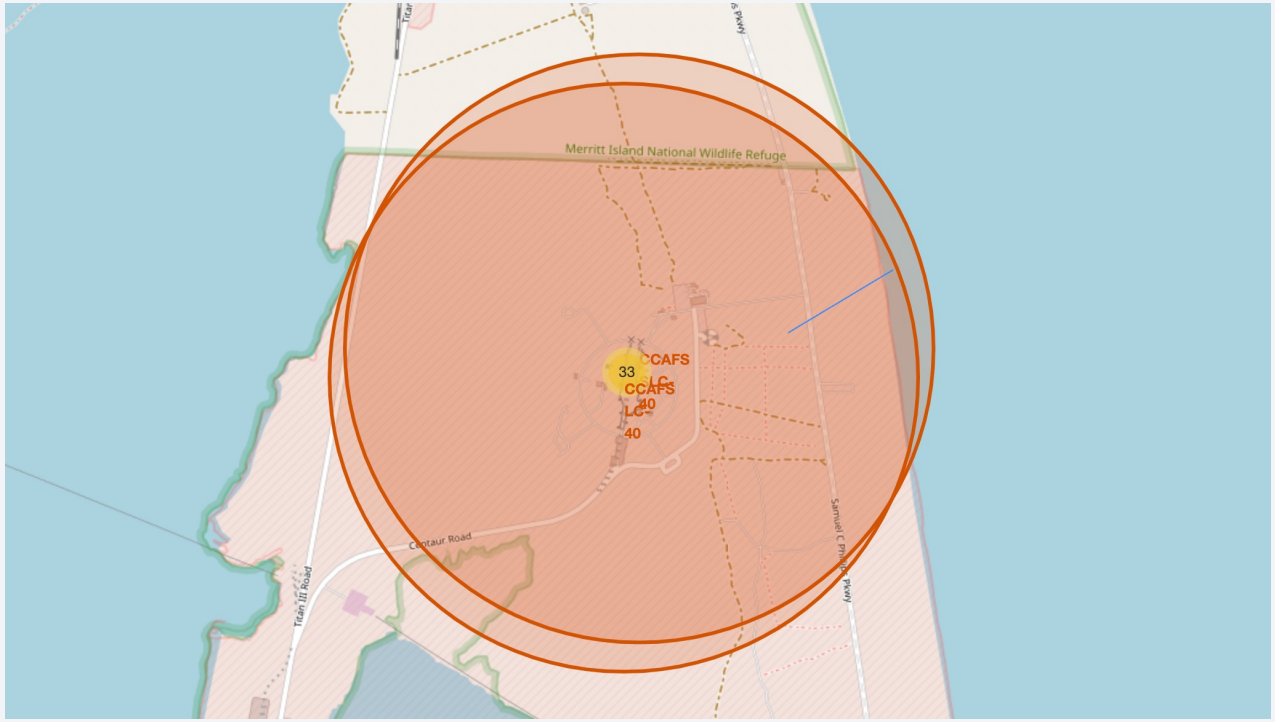
Success and Failures

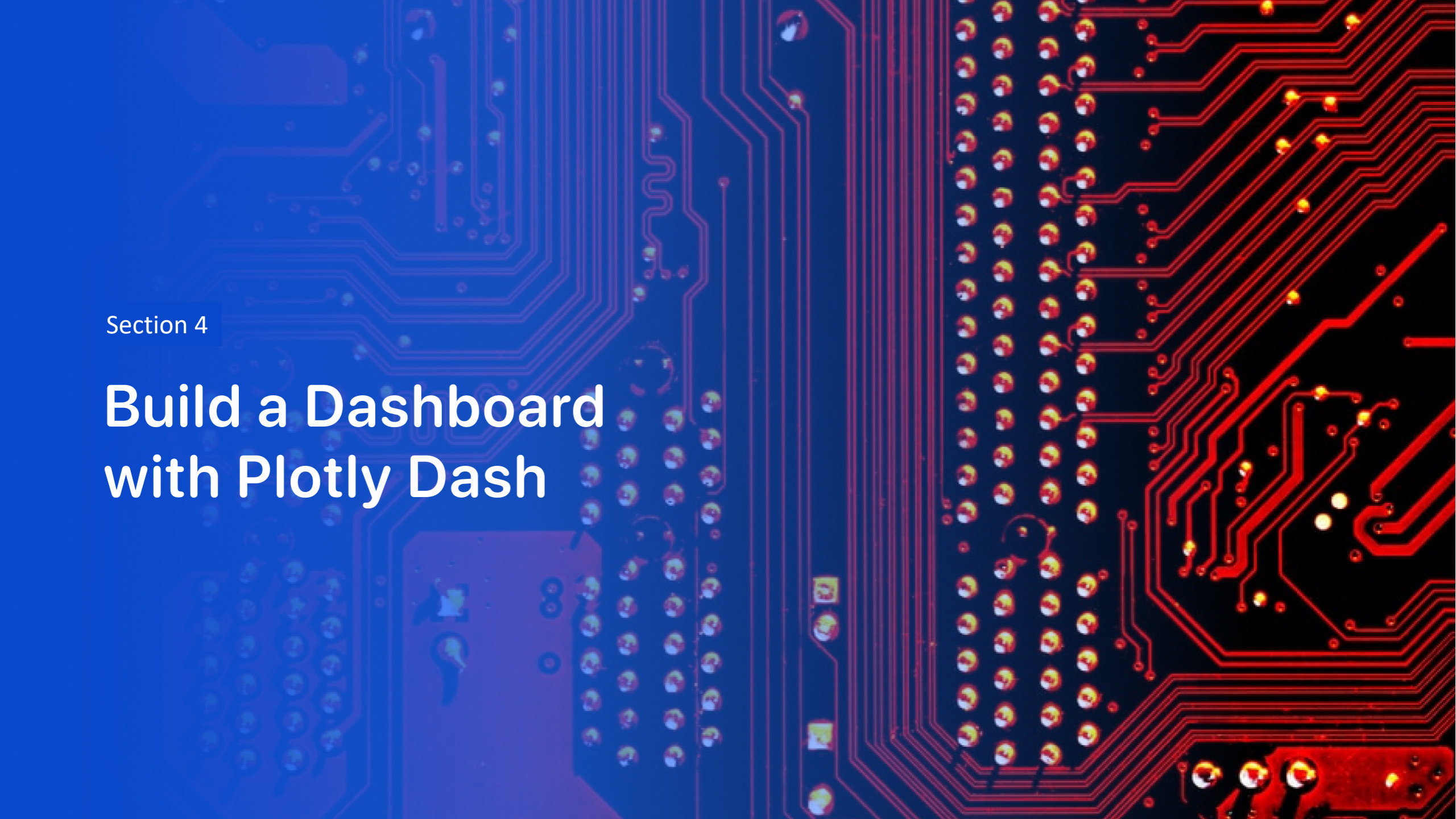
- It's possible to zoom in and view the success and failures at each of the launch sites.



Proximity Locations

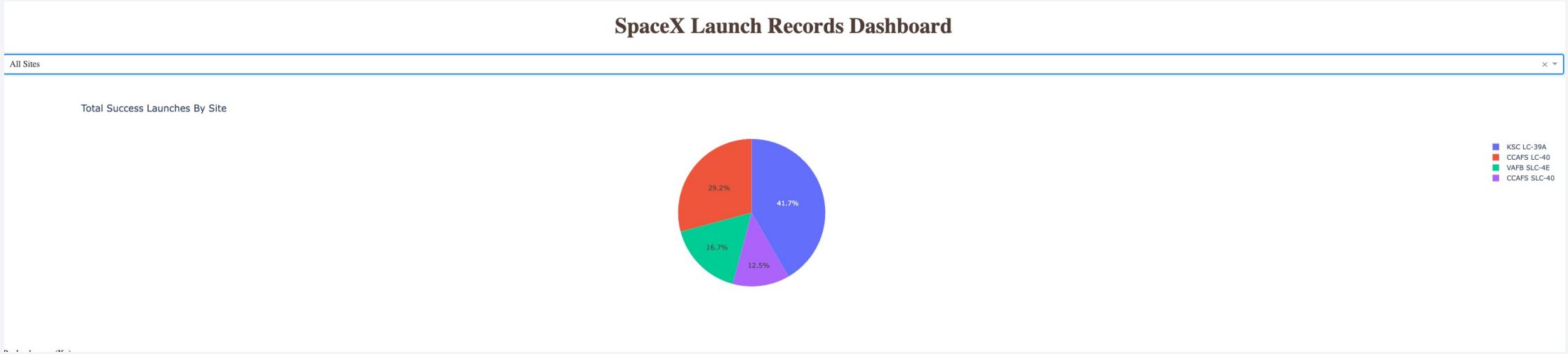
- We’re able to add lines onto the maps to determine the proximity to other locations.
- For example, the railway is location 0.68km away from KSC-LC39A launch site.




Section 4
Build a Dashboard with Plotly Dash

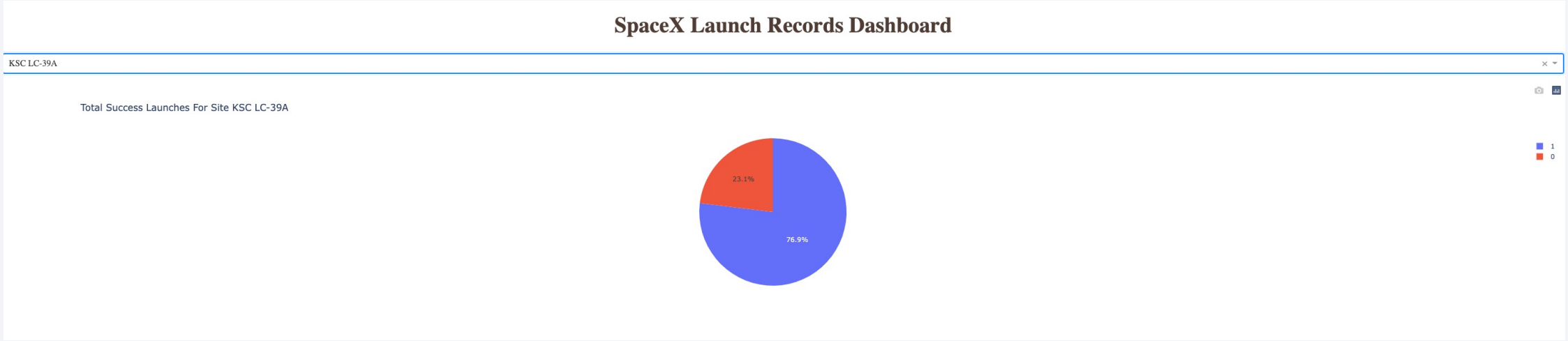
Success Launches

- Success launches across all sites:



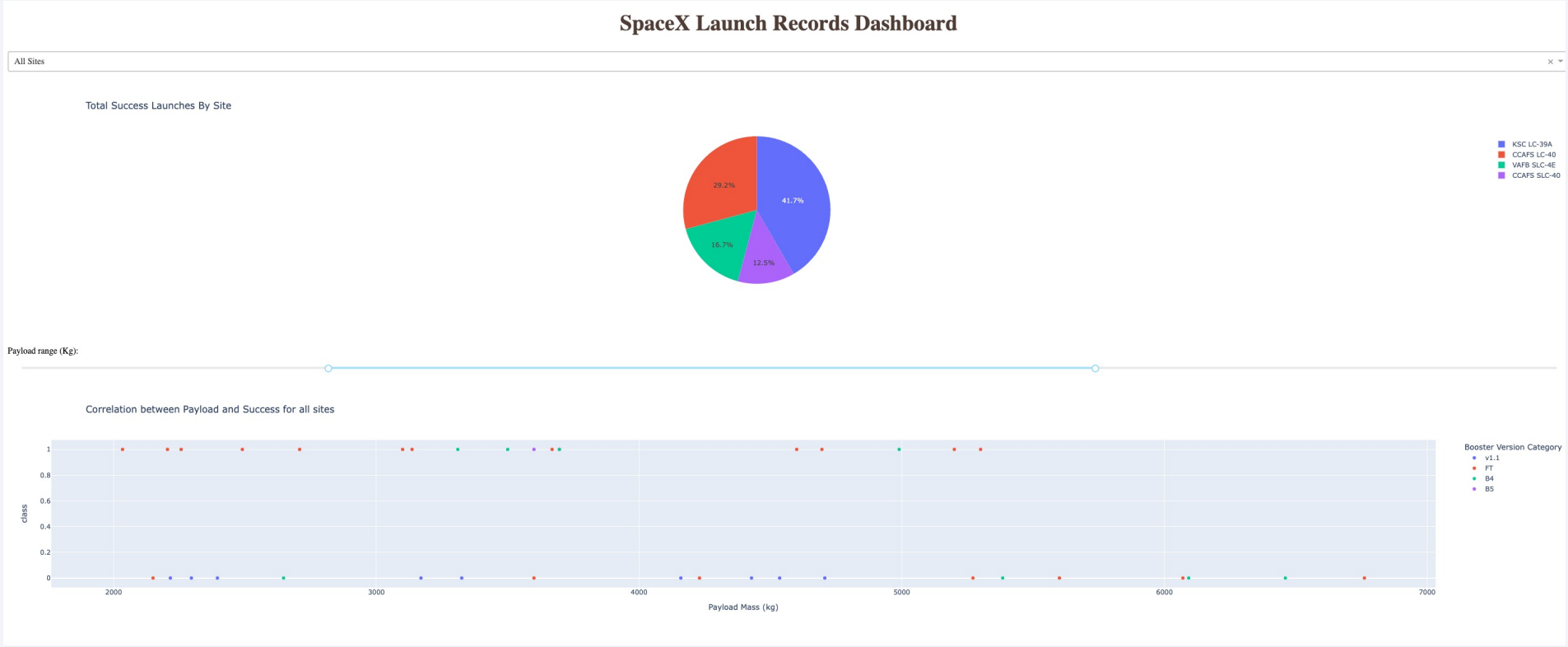
Highest Launch Site Ratio

- Highest launch site ratio showing 77% success rate for KSC LC-39A



Payload vs Launch Outcome

- Looking at all sites but filtering on payload mass range between 2000 and 7000:



Section 5

Predictive Analysis (Classification)

Classification Accuracy

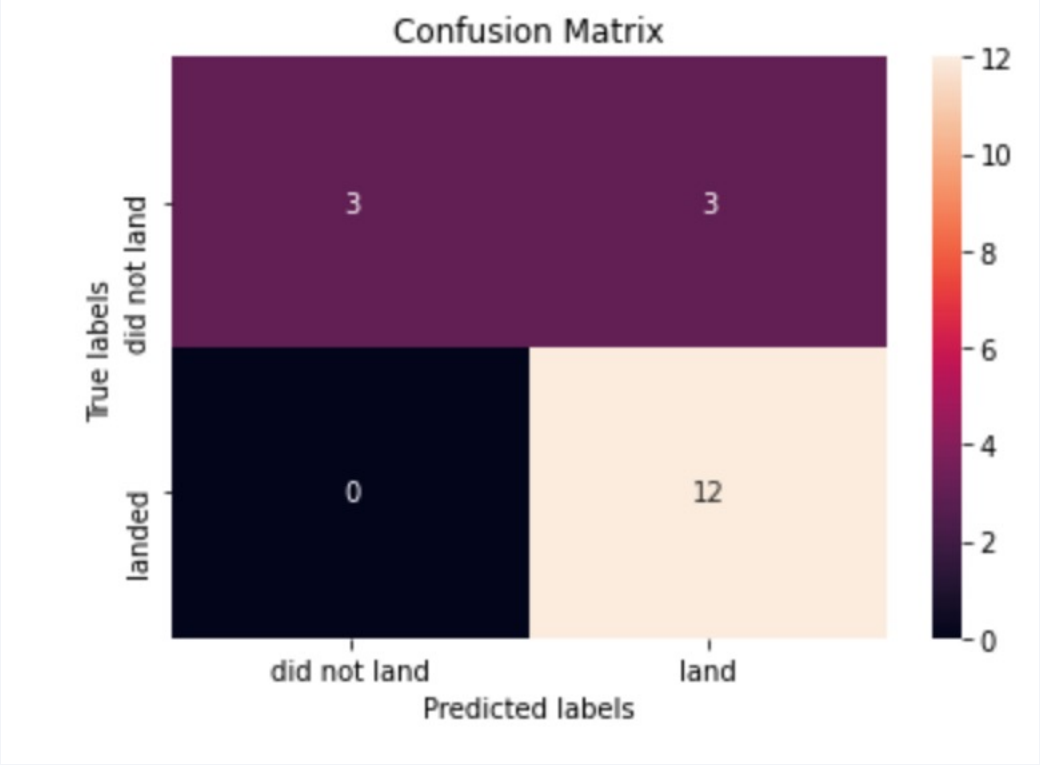
- All models showed the same accuracy when looking at the score for each model.

```
print('LR score:',lr.score(X_test, Y_test))
print('SVM score:',svm.score(X_test, Y_test))
print('DT score:',tree.score(X_test, Y_test))
print('KNN score:',KNN.score(X_test, Y_test))
```

```
LR score: 0.8333333333333334
SVM score: 0.8333333333333334
DT score: 0.8333333333333334
KNN score: 0.8333333333333334
```

Confusion Matrix

- The confusion matrix showed promising results but was impacted by the false positives.



Conclusions

- After a reiterative process, the CCAFS SLC 40 is the best launch site. The success rate appears to be lower than the others but this was due to very bad start with numerous failures. This site now has a proven record recently and can accommodate large payload masses in excess of 10,000 kg.
- SpaceY can be successful since we're able to correctly determine launch success 83% of the time.

Thank you!