Winning Space Race with Data Science

Wille Faler

2023-07-02

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API and with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL & with Data Visualization
  - Interactive Visual Analytics with Folium
  - Predictive Analysis with Machine Learning
- Summary of all results
  - Public sources allowed collection of valuable data for analysis
  - Exploratory Data Analysis allowed identification of most predictive features
  - Classification testing discovered best model to predict success or failure of launch with collected data

Introduction

- Project background and context
  - Explore viability of Space Y to compete with Space X. Space X uses reusable first stage rockets for cost savings, analysis of success of this data from public sources allows us to figure out if the thesis of Space X is correct.
- Problems you want to find answers
  - What factors are predictive of whether a rocket will land successfully?
  - What interaction among those factors/features determine the overall success rate.
  - What conditions need to be in place to maximise chances of successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web-scraping of Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features, to give them numerical values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was normalized, divided into training and test datasets, evaluated against for different classification models, with the accuracy of each being evaluated with different combinations of parameters.

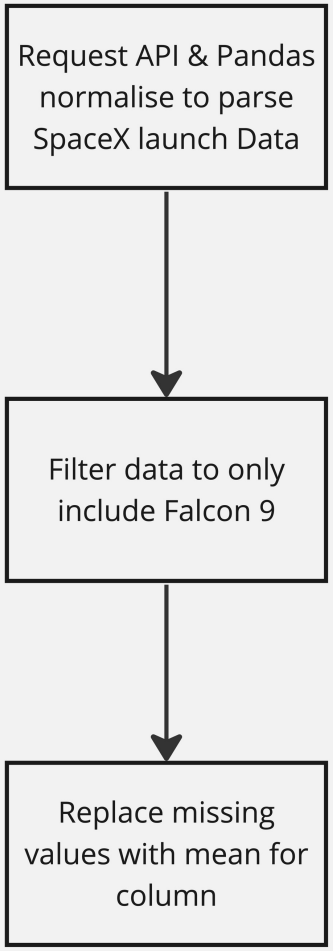
Data Collection

- Datasets were collected from SpaceX API – https://api.spacexdata.com/v4/rockets,
- and from Wikipedia - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches using web scraping

Data Collection – SpaceX API

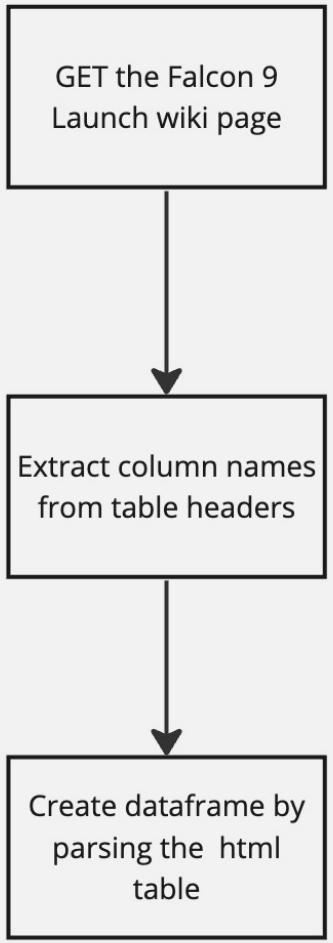
- SpaceX has a Public API, from which data was retrieved.
- API was used as described in the chart to the left.

- Source Jupyter Playbook: https://github.com/wfaler/data-science-capstone/blob/main/module-1-part-1-jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

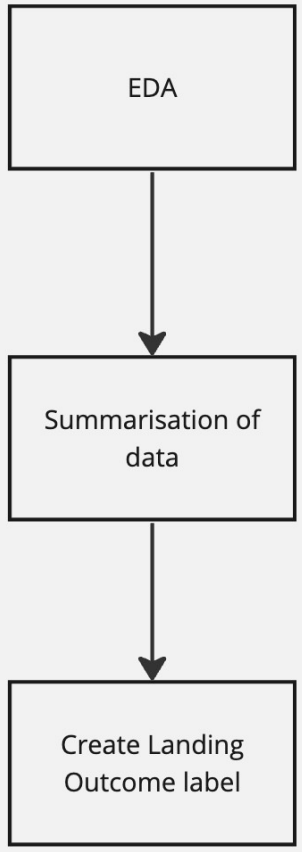
- We used web-scraping from Wikipedia to scrape Falcon 9 launches, parsing data with BeautifulSoup
- Parsed the table and converted to a Pandas dataframe
- Github link to playbook: https://github.com/wfaler/data-science-capstone/blob/main/module-1-part-2-jupyter-labs-webscraping.ipynb



Data Wrangling

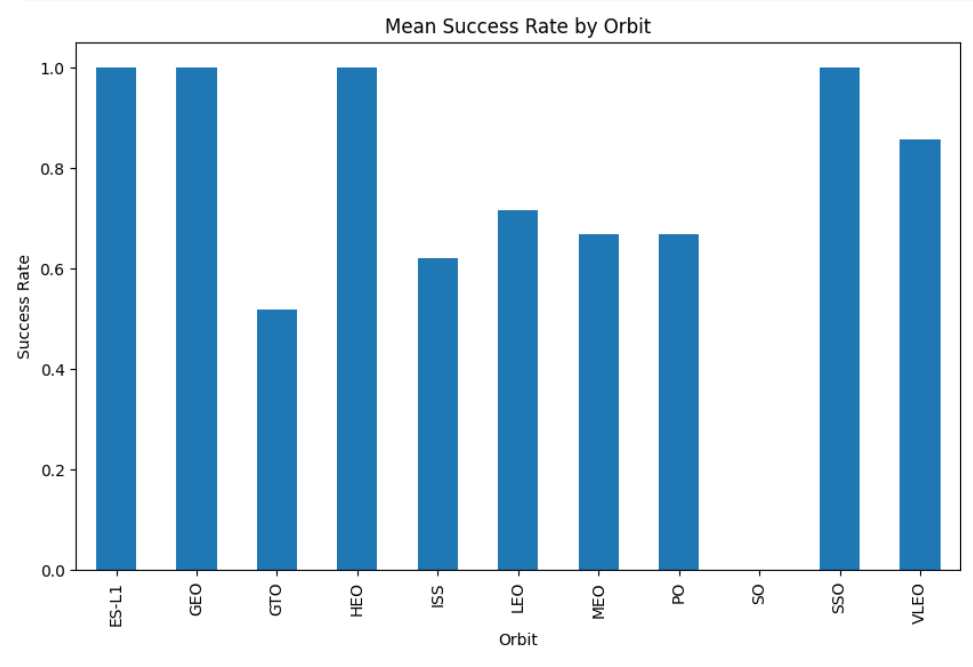
- First, we performed some exploratory analysis, looking at contents, types of data in the data set.
- We calculated the number of launches from each site
- Finally, we created a new Landing Outcome label and exported to CSV

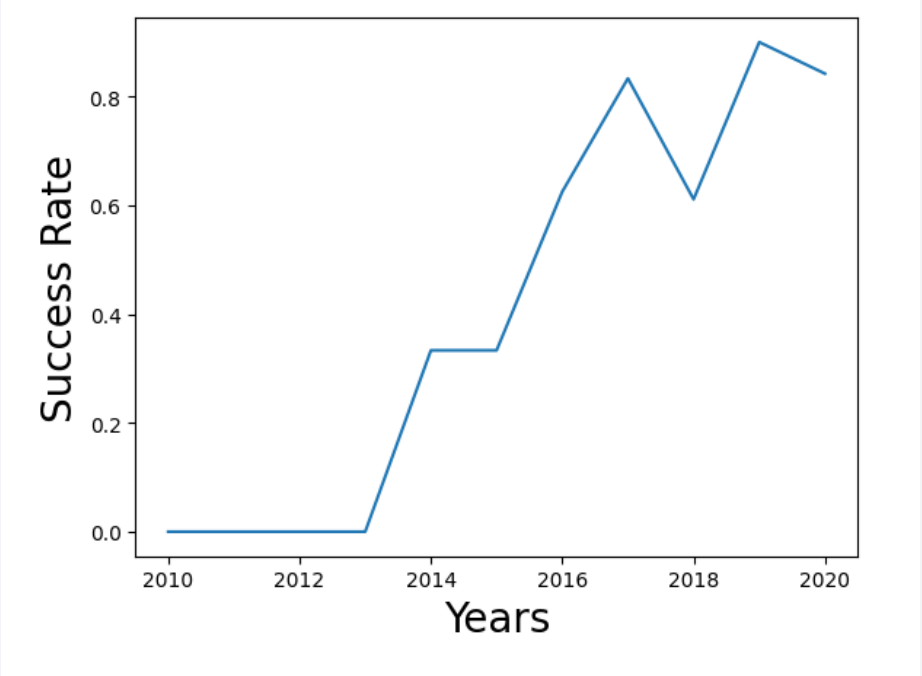
- Github: https://github.com/wfaler/data-science-capstone/blob/main/module-1-part-3-labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- We explored the data by visualizing various relationships in the data, such as success rate per orbit type, success rate over time, payload and orbit.
- We found that the success-rate has increased over time
- Github URL for EDA: https://github.com/wfaler/data-science-capstone/blob/main/module-2-part-2-jupyter-labs-eda-dataviz.ipynb





EDA with SQL

- We did queries to find out the following:
  - Find distinct launch sites (select distinct Launch_Site from SPACEXTBL)
  - Display 5 records for launch sites beginning with 'CCA'
  - Display total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 1.1
  - List names of boosters which have success in drone ship and have payload between 4000 and 6000 KG
  - 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.
  - 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.

Github URL for results: https://github.com/wfaler/data-science-capstone/blob/main/module-2-part-1-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

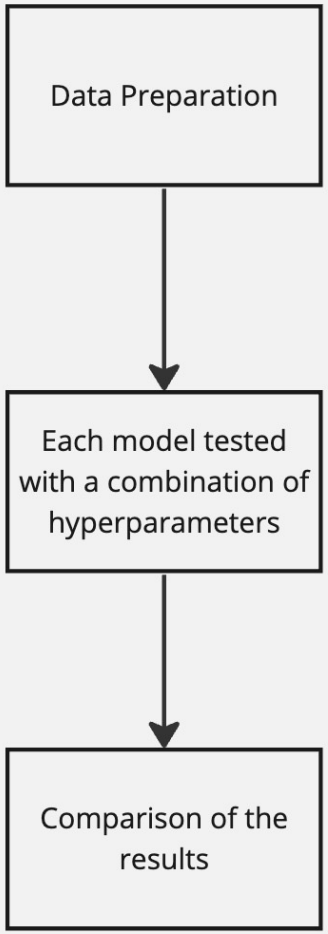
- Marked all launch cites, added map objects such as markers, lines and circles to indicate success or failure of launches.
- We calculated the distances between launch sites and proximities
- Github URL: https://github.com/wfaler/data-science-capstone/blob/main/module-3-part-1-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive Plotly dashboard
- We plotted pie charts showing total launches by site
- We plotted a scatter graph showing the relationship with outcome and payload for different sites.
- Code: https://github.com/wfaler/data-science-capstone/blob/main/spacex_dash_app.py - due to changes in dash, I also had to change some of the imports (outside of scope of this task)

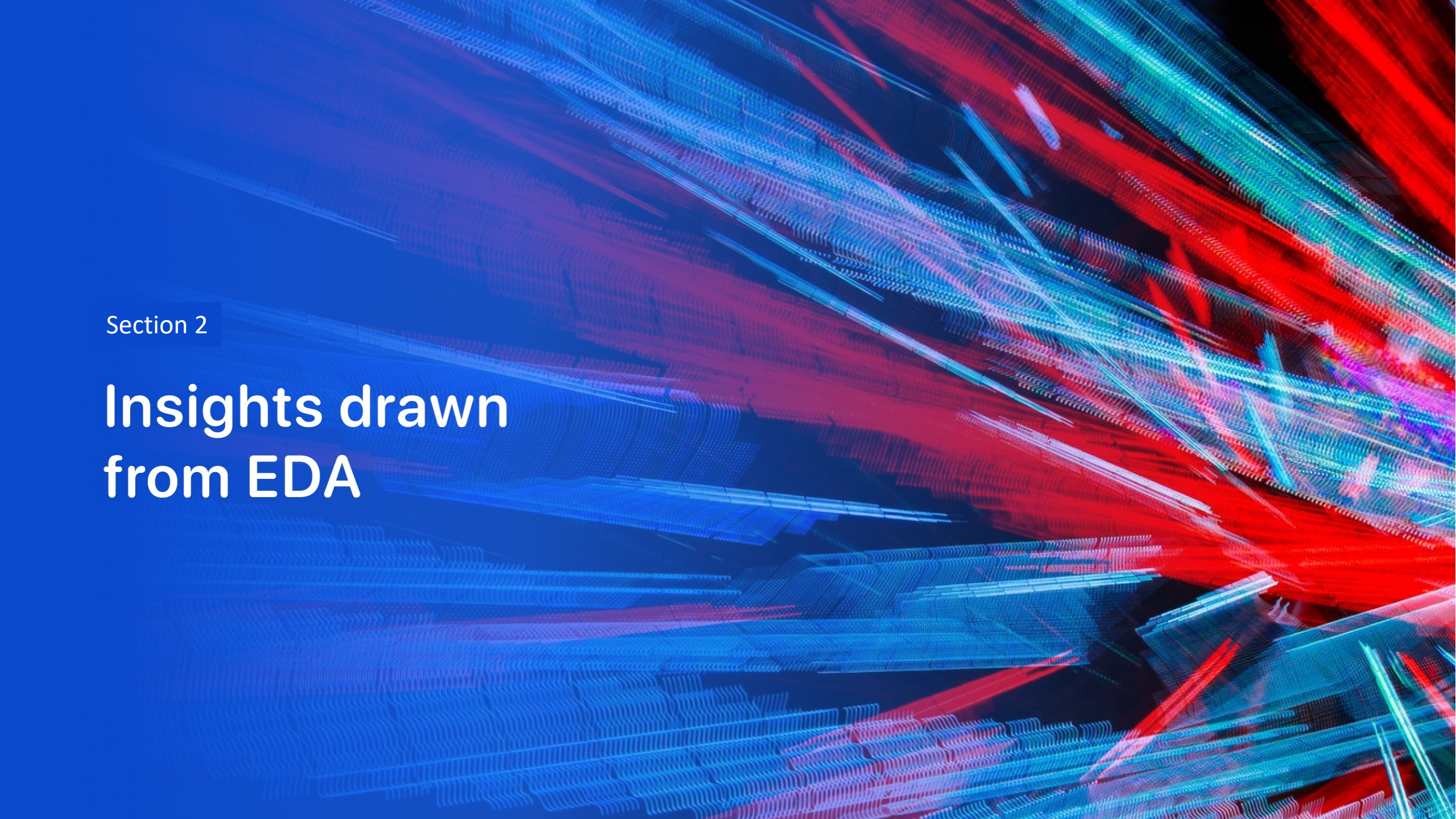
Predictive Analysis (Classification)

- Four classification models were compared in their predictive ability of predicting a successful launch:
  - Logistic Regression
  - Decision Tree
  - Support Vector Machines
  - K Nearest Neighbour
- Github Source: https://github.com/wfaler/data-science-capstone/blob/main/module-4-part-1-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



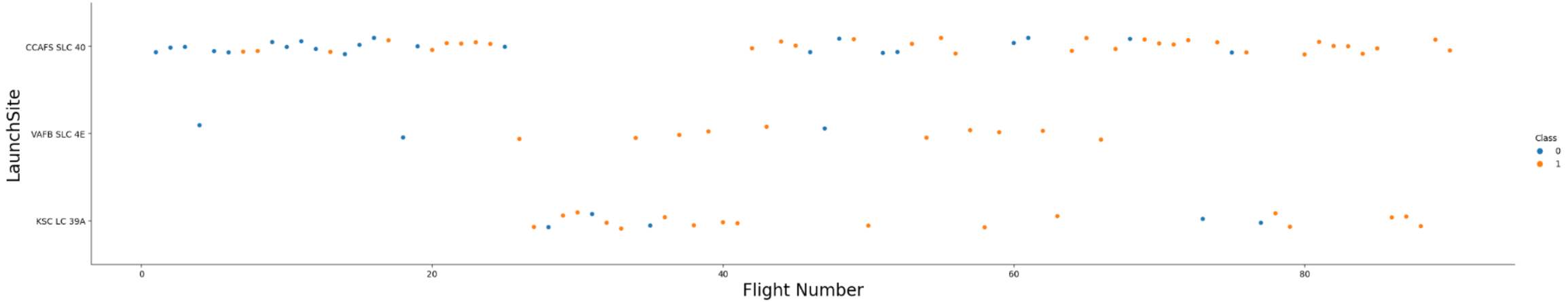
Results

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


Section 2
Insights drawn from EDA

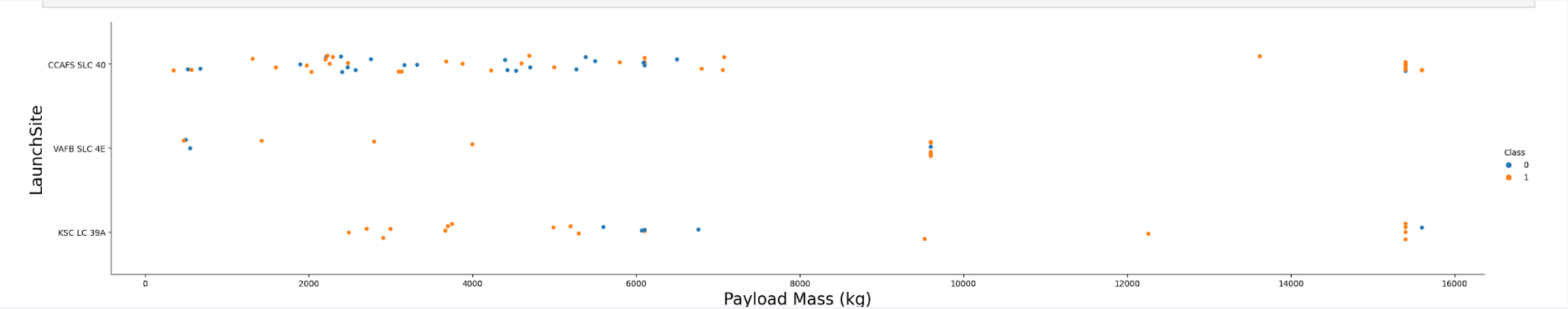
Flight Number vs. Launch Site

- We can see a trend towards higher success rates over time
- CCAFS SLC 40 has recently had a high number of successful launches, indicating it is a good place to launch



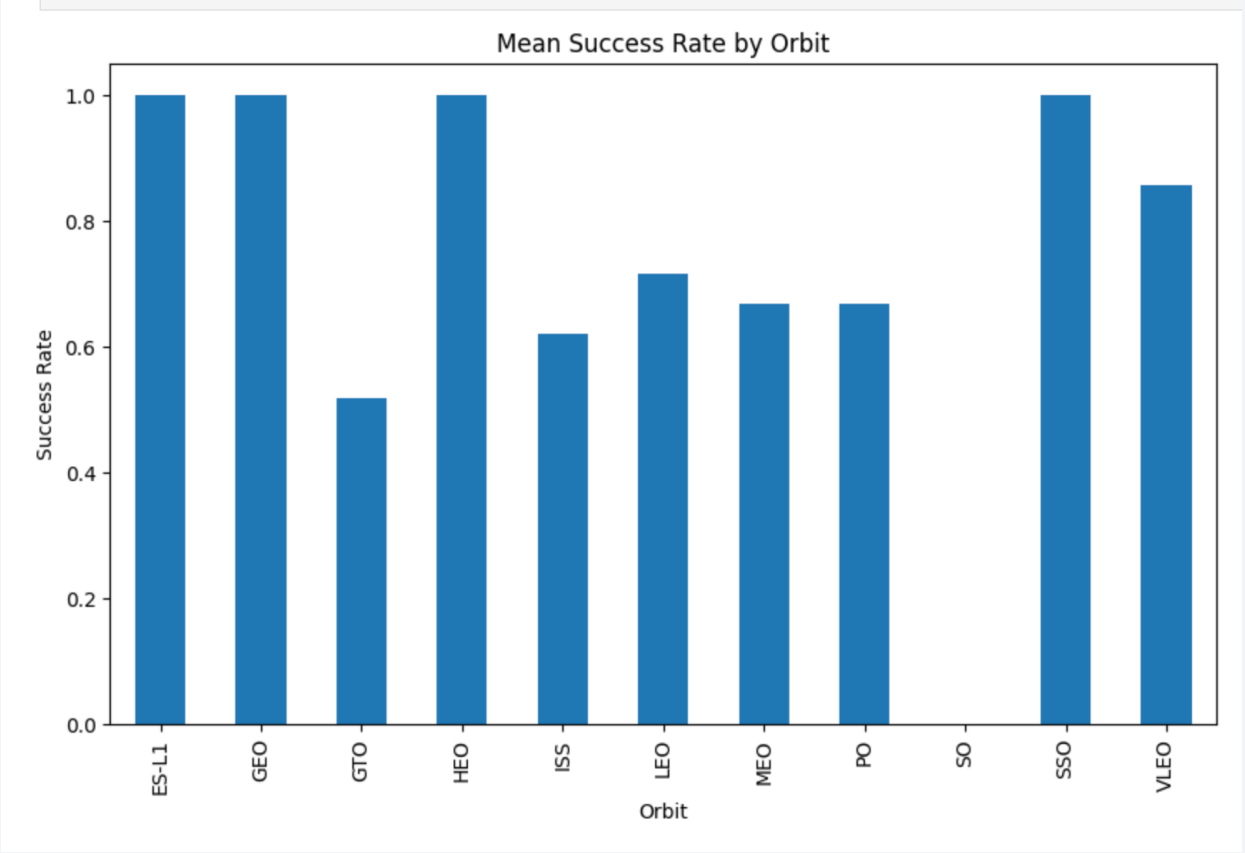
Payload vs. Launch Site

- We can see the higher the payload, the higher the success rate.
- However, payload may be determining success, but rather that payload has increased over time, as confidence in success of a launch has risen (correlation, not causation)



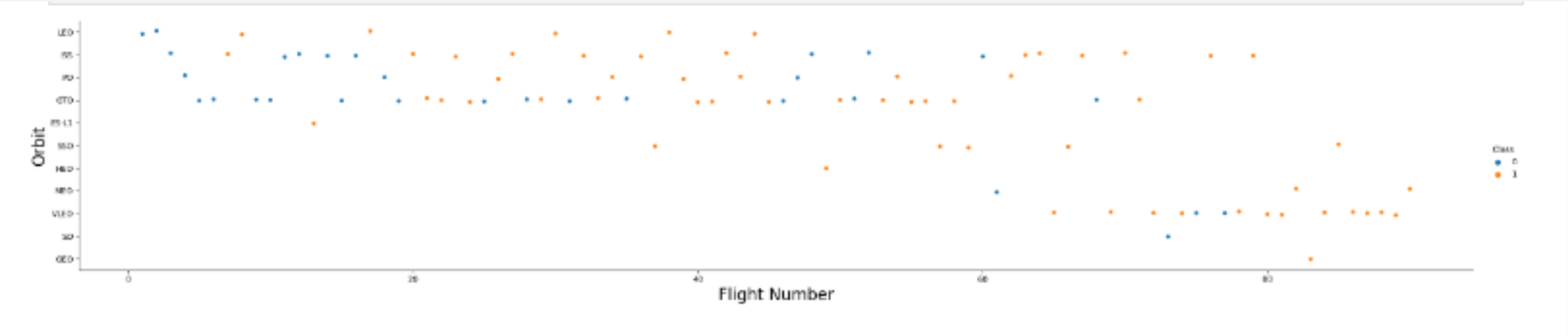
Success Rate vs. Orbit Type

- Highest success-rate by orbit type are:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by: VLEO and LEO



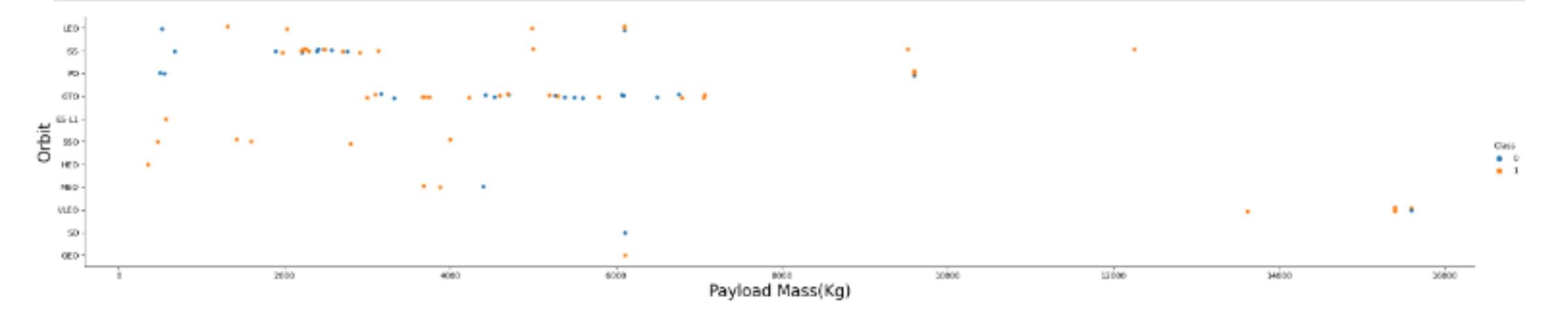
Flight Number vs. Orbit Type

- Success rate improved over time
- VLEO has become increasingly popular later on, likely representing increased revenue opportunities



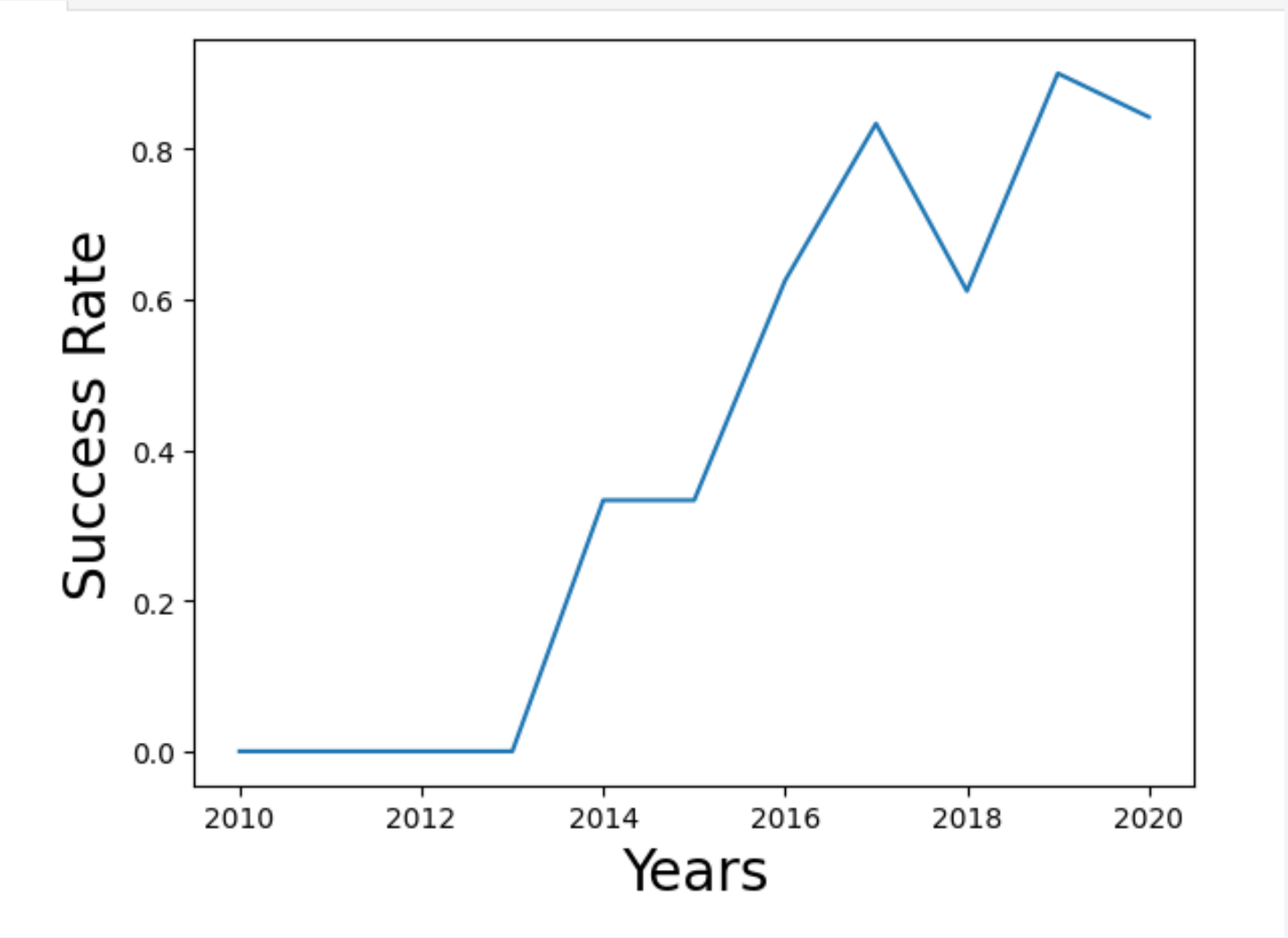
Payload vs. Orbit Type

- VLEO orbits have higher payloads, seemingly confirming the theory from the previous slide, that VLEO likely has a higher financial reward.
- Higher payloads have higher success-rates, indicating payloads have been increased over time, as success-rates in general increased



Launch Success Yearly Trend

- Success-rates have increased dramatically over time, with a slight, temporary dip in 2018.
- Success-rates were zero for the first 3 years, indicating technology was not yet mature enough back then.



All Launch Site Names

- According to the data, there are for launch sites:

```
In [12]: %sql select distinct Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

Launch Site Names Begin with 'CCA'

- Querying for the five records starting with ‘CCA’ imply this launch site was used early in Space X history (2010-2013)

Display 5 records where launch sites begin with the string 'CCA'

In [23]:
```
%%sql
select * from SPACEXTBL
where Launch_Site like 'CCA%'
limit 5
```

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

Out[23]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA:
  - Total payload is 45596kg

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

In [15]:
```
%%sql
select sum(PAYLOAD_MASS__KG_) from SPACEXTBL
where Customer = 'NASA (CRS)'
```

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

Out[15]:

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Average payload carried by the F9 V1.1 is 2928.4kg

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

In [16]:
```
%%sql
select avg(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version = 'F9 v1.1'
```

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

Out[16]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- First success is 22/12/2015.
- Please pay attention to note about ”min” use in the screenshot.

```
[5]:
%%sql
select date from SPACEXTBL
where Landing_Outcome = 'Success (ground pad)'
order by substr(Date,7,4)
limit 1
```

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

[5]:

Date
22/12/2015

Note: Could not use min function, as dates in SQLite are stored as strings, not date objects, so whether `min()` works depends on the date format used. In this case, it did not.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here is the query, and the names of the boosters that have had success with payloads between 4000-6000kg

In [18]:
```
%%sql
select Booster_Version from SPACEXTBL
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[18]:

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

Total Number of Successful and Failure Mission Outcomes

- 1 outright failure, 99 successes and 1 success with unknown status of payload.
- I am assuming “success” does not account for “landing outcome”.

In [19]:

```
%%sql
select trim(Mission_Outcome) as Outcome,count(Mission_Outcome) from SPACEXTBL
where (Mission_Outcome like 'Success%' or Mission_Outcome like 'Failure%')
group by Outcome
```

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

Out[19]:

Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Here is the query, and the names of the boosters that have carried the maximum payload.
- All are versions of F9 B5

In [20]:
```
%%sql
select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[20]:

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

2015 Launch Records

- Failed landing outcomes in 2015. Result: in April and October.

In [21]:

```
%%sql
select substr(Date,4,2) as Month,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTBL
where substr(Date,7,4) = '2015' and
Landing_Outcome like 'Failure%'
```

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

Out[21]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Landing outcomes for the selected date range, in order of frequency

In [22]:

```
%%sql
SELECT LANDING_OUTCOME,COUNT(LANDING_OUTCOME) AS COUNT FROM SPACEXTBL
WHERE (substr(DATE,7)||substr(DATE,4,2)||substr(DATE,1,2)
      BETWEEN '20100604' AND '20170320')
GROUP BY LANDING_OUTCOME
ORDER BY COUNT DESC
```

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

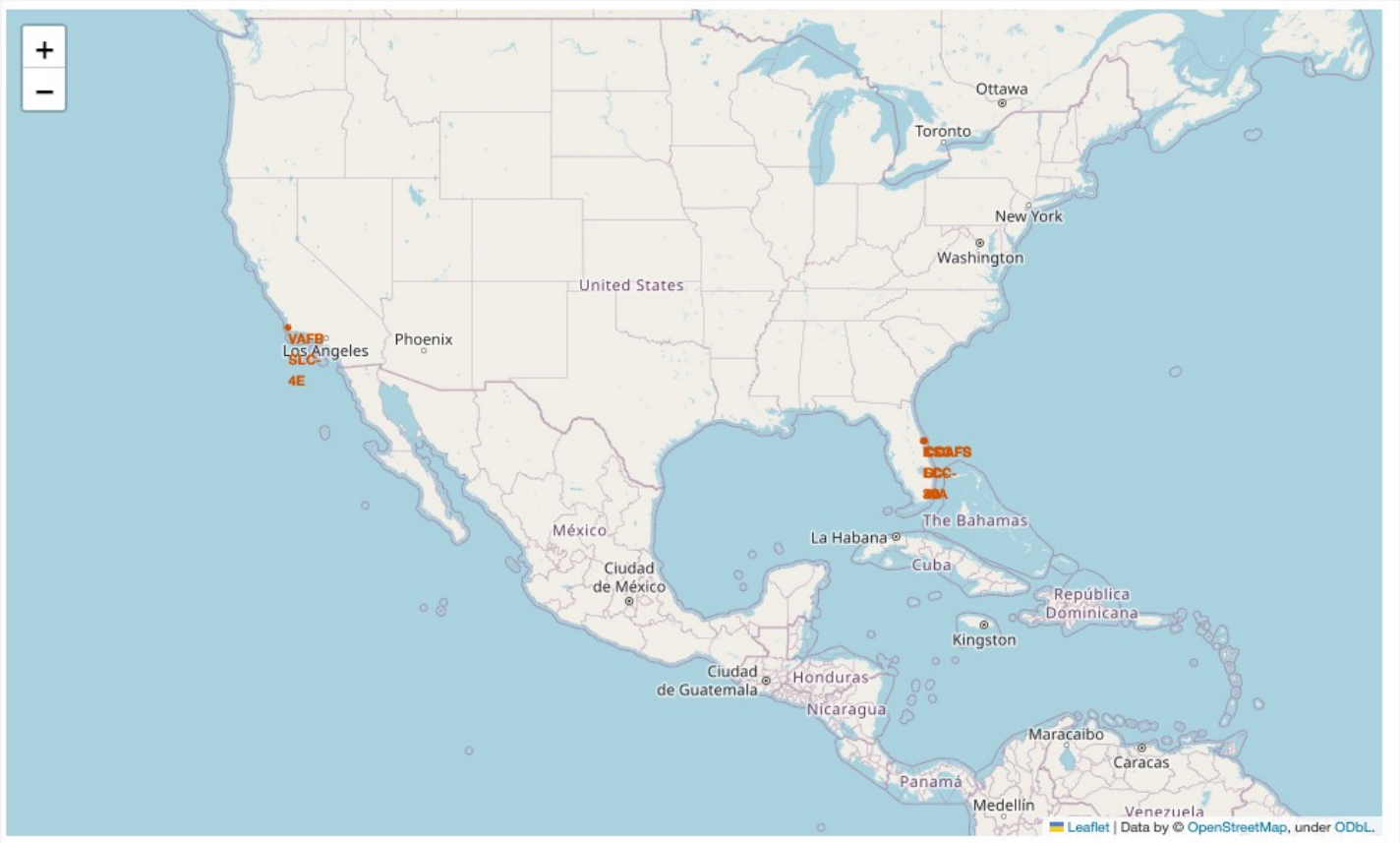
Out[22]:

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


Section 3
Launch Sites
Proximities Analysis

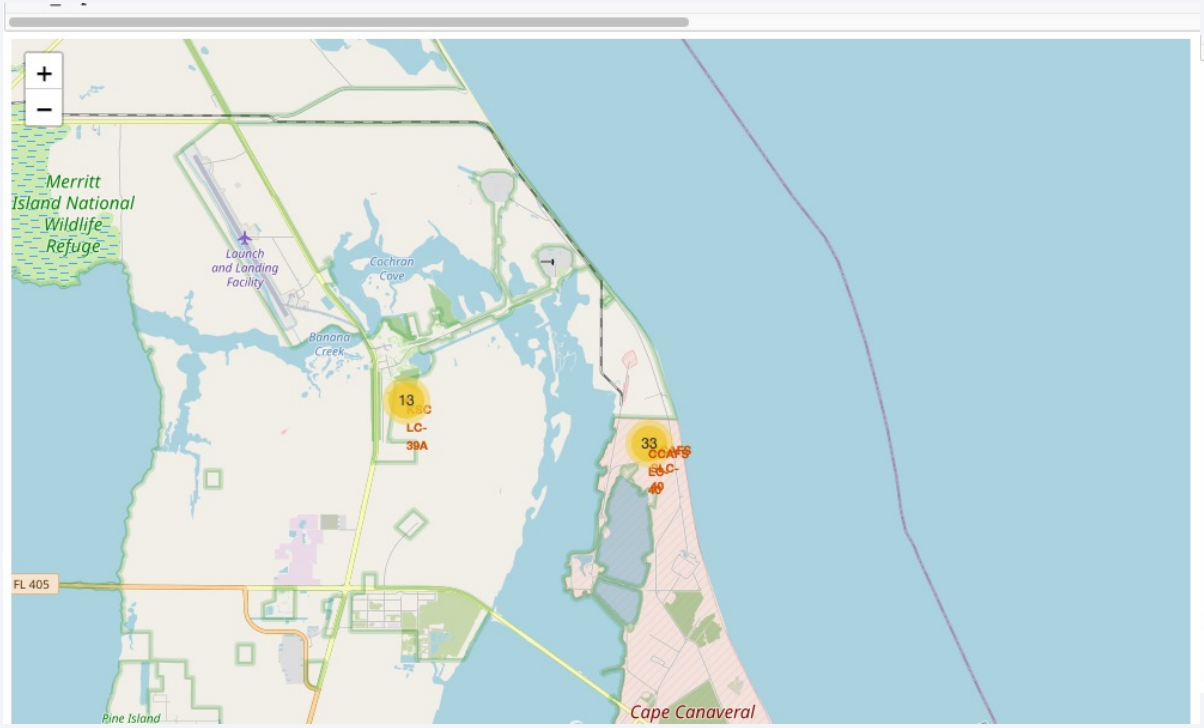
Map of all launch sites

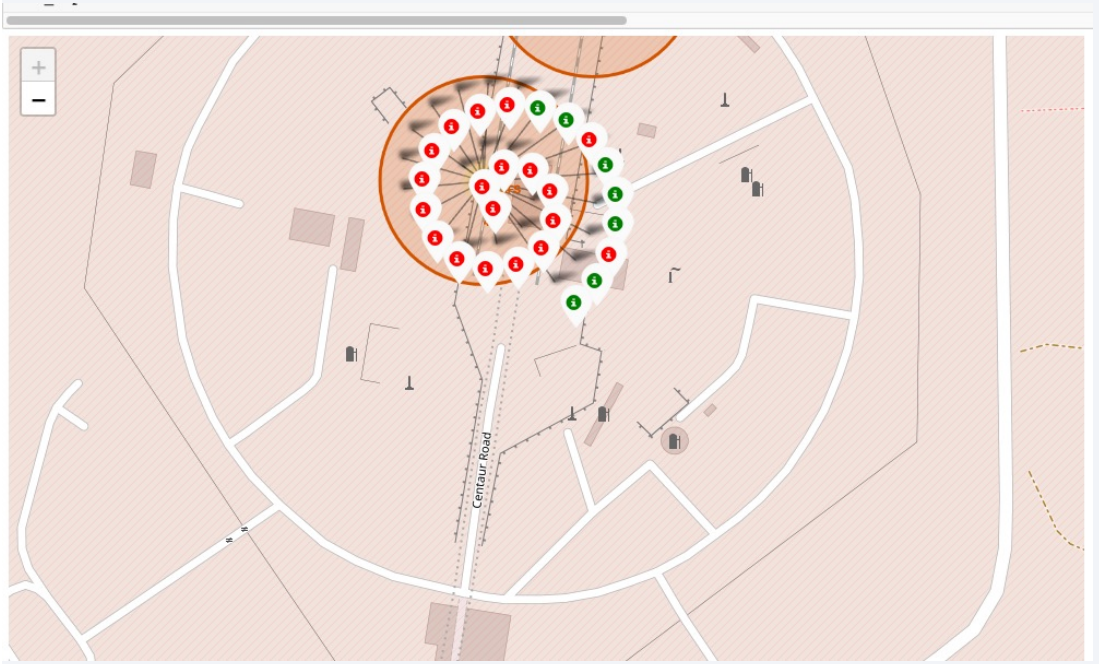
- Launch-sites are all in proximity of water, allowing for the safe destruction of rockets, in case of failure



Launch outcomes by site

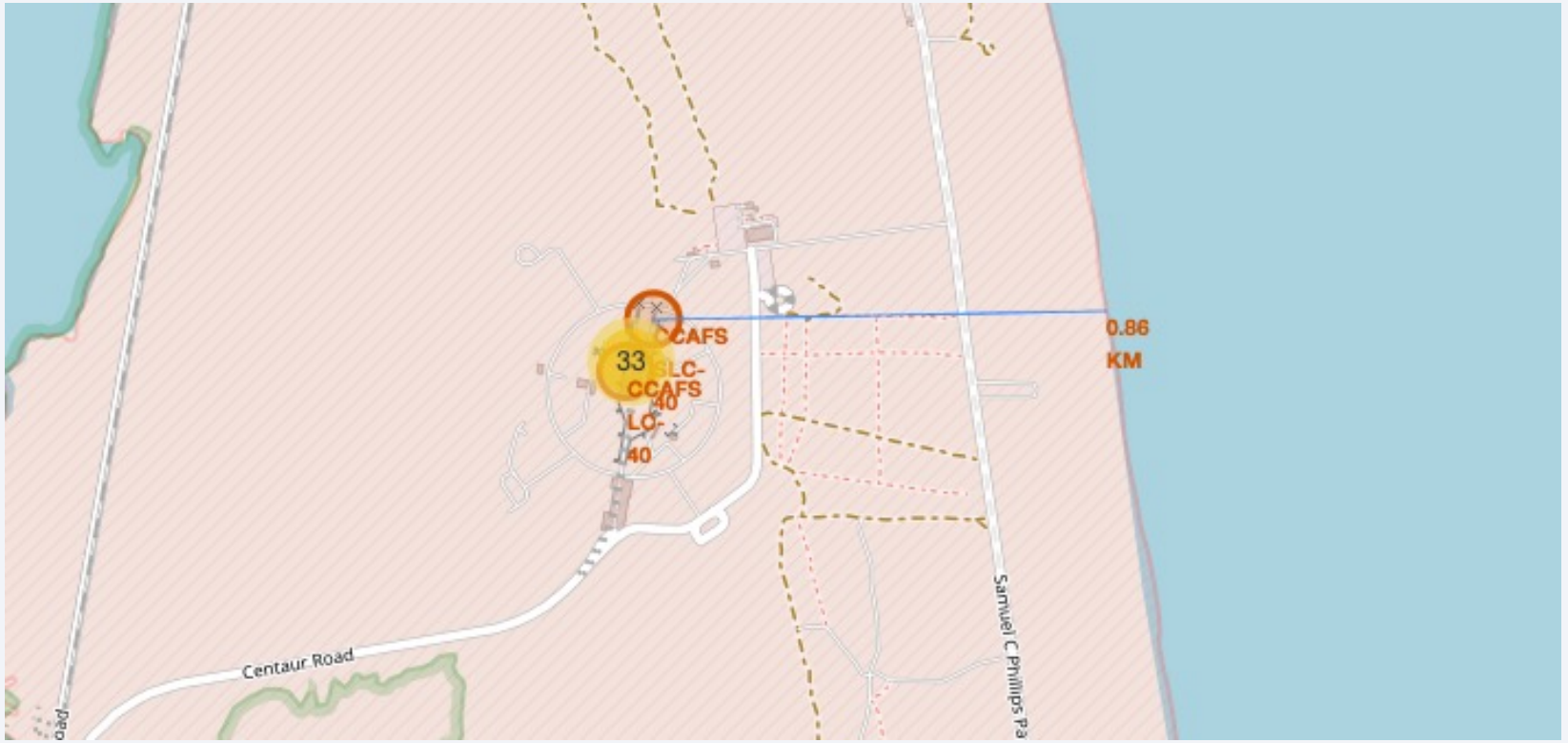
- We created markings on the Folium map, denoting the launch outcome in red or green, for failure or success, which can be inspected by zooming in/out

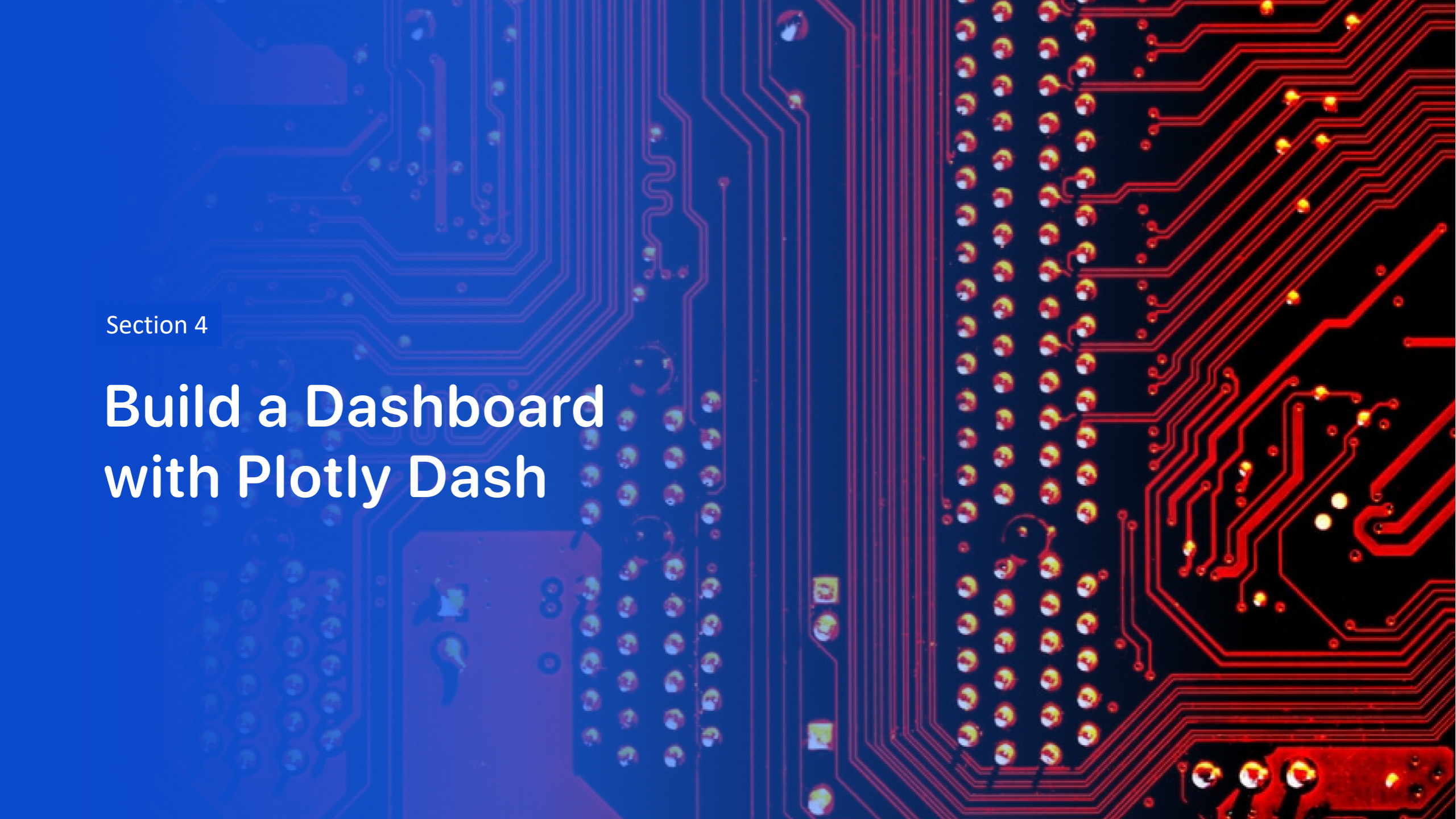




Distances

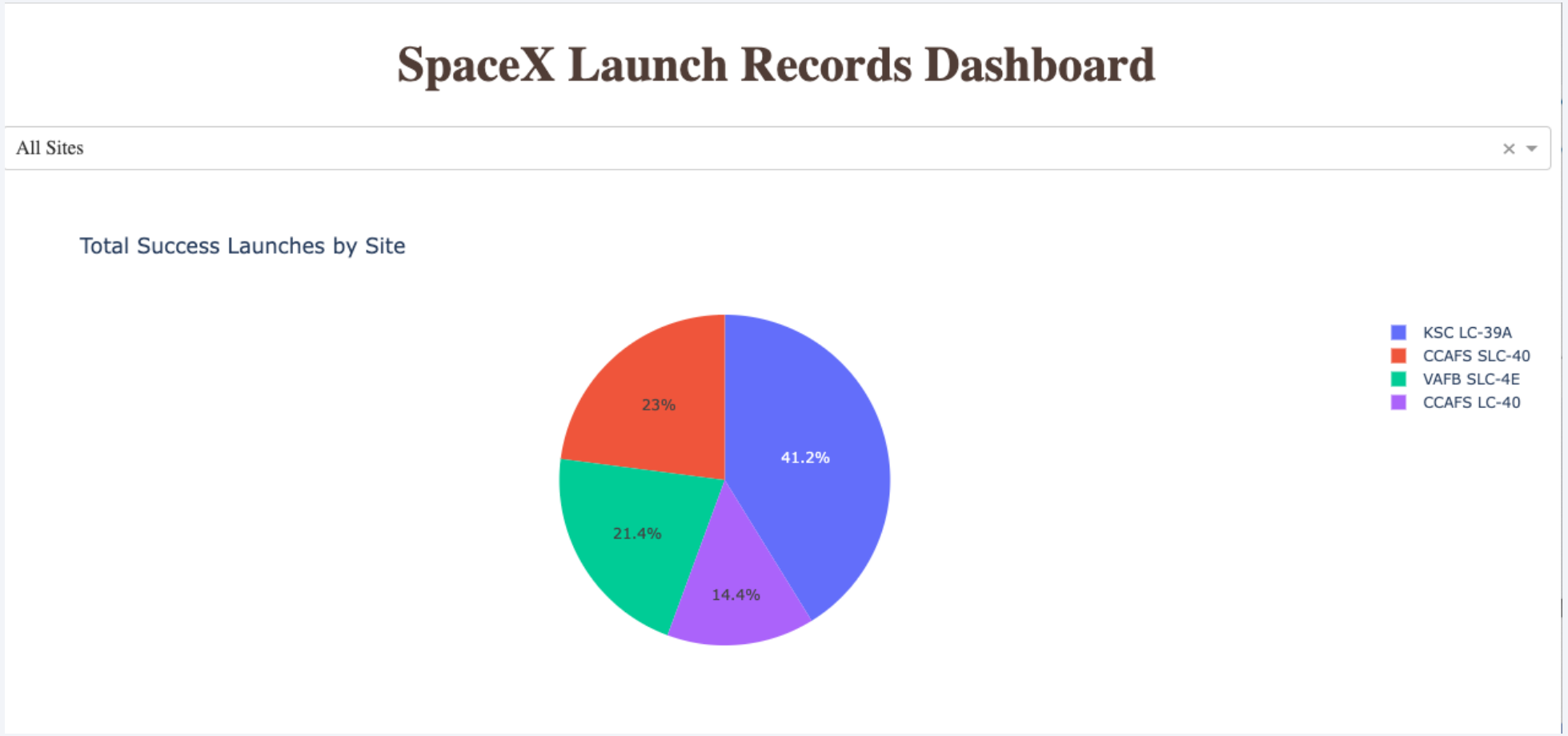
- Finally, we mapped distances to logistics and safety. As seen below, we have distance to the sea mapped out




Section 4
Build a Dashboard with Plotly Dash

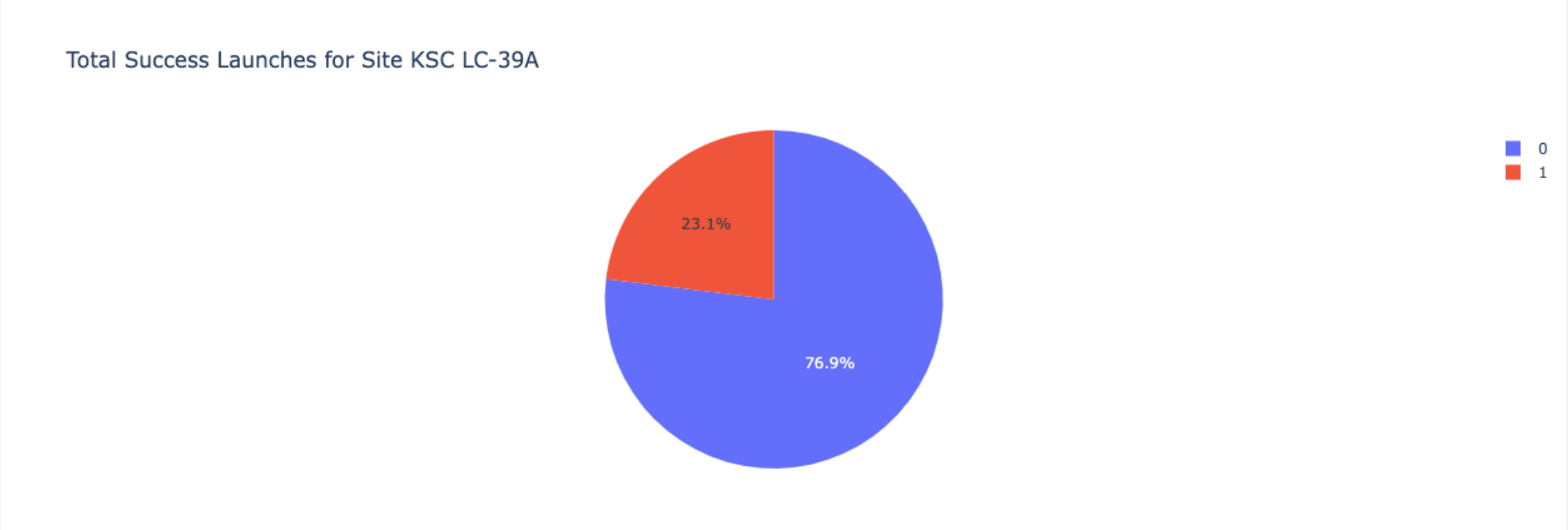
Locations of successful launches

KSC LC-39A has the highest amount of successful launches of all sites



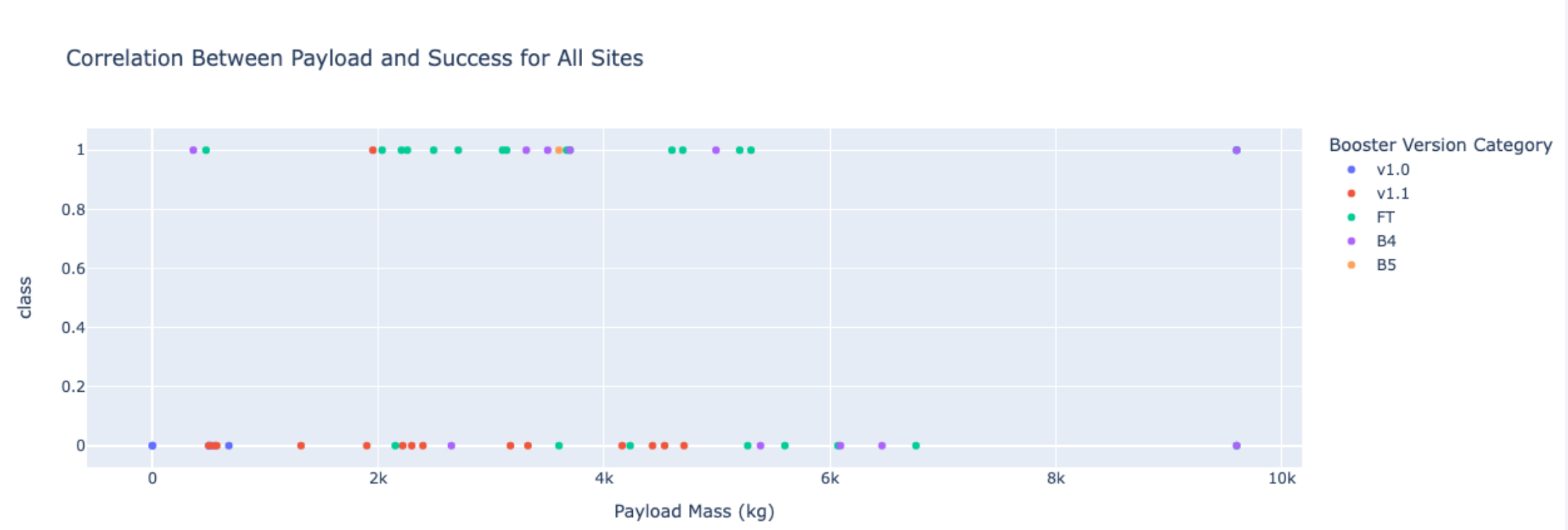
Successful launches from KCS LC-39A

- 76.9% of launches from this locations have been successful



Correlation between payload and success for all sites by booster type

- FT boosters and payloads below 6000kg have the highest success rate

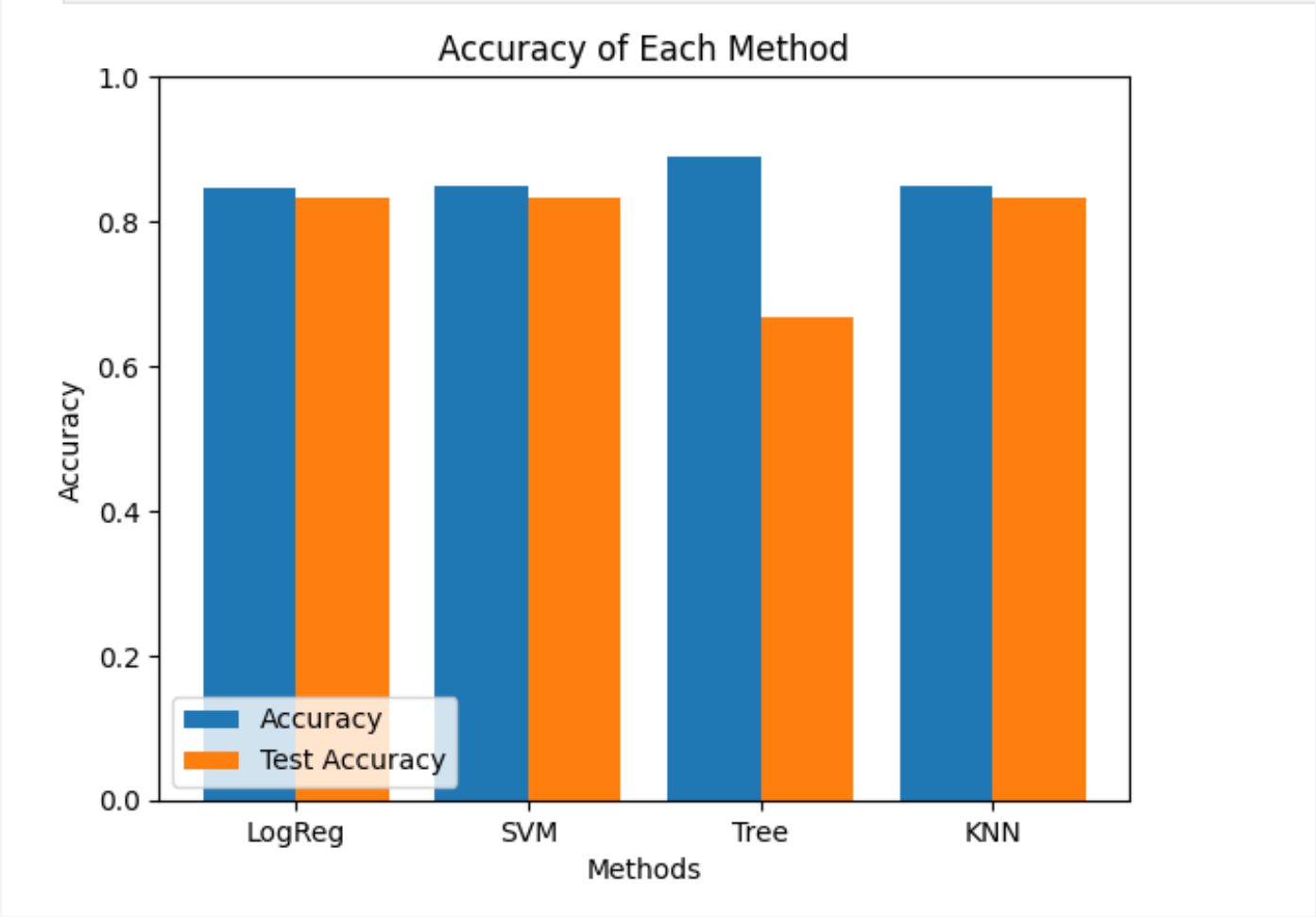


Section 5

Predictive Analysis (Classification)

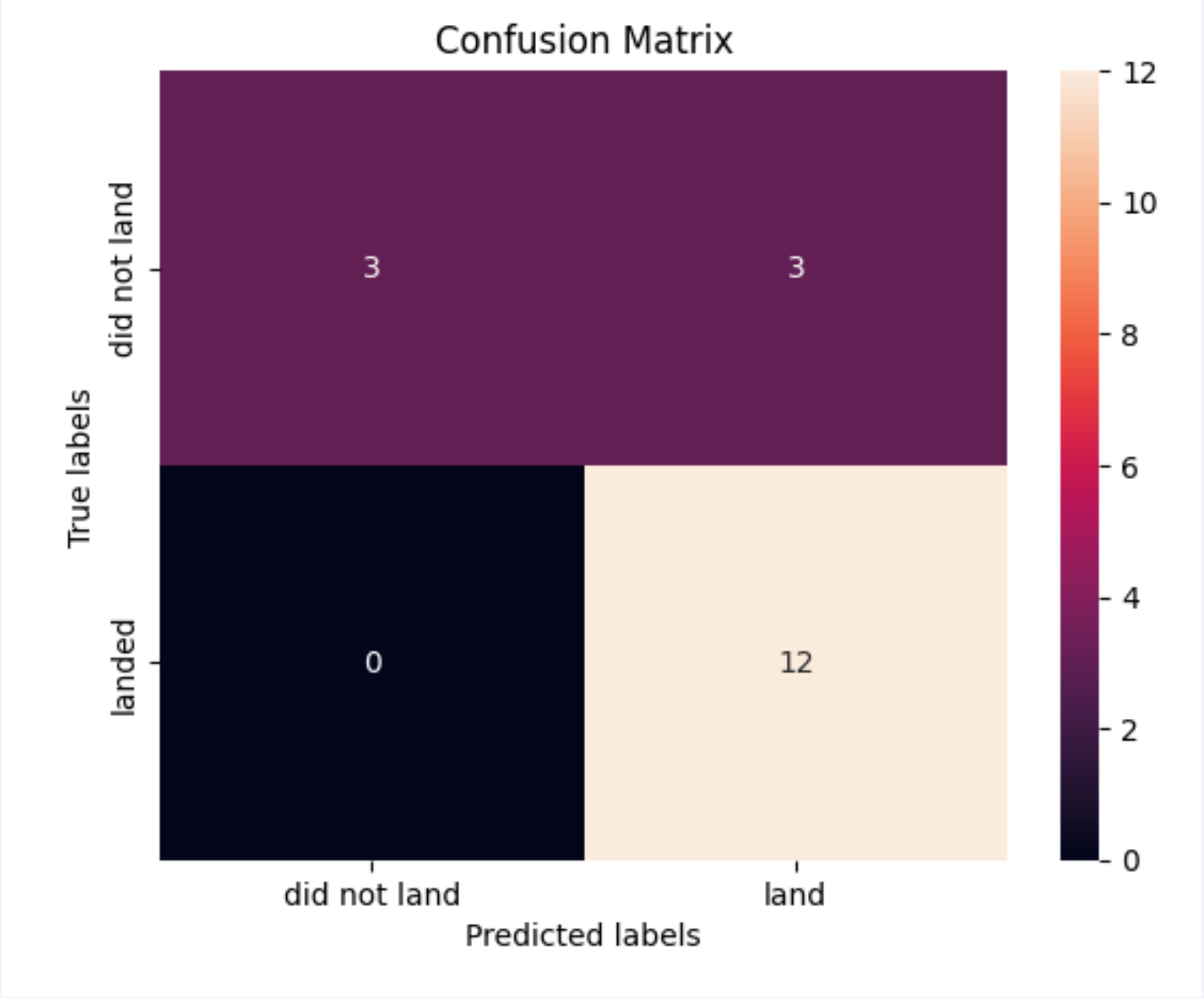
Classification Accuracy

- Decision Tree had the highest accuracy, but lowest test accuracy (88.9% vs 66.7%)
- SVM and KNN had high accuracy both in- and out of test (84.8%/83.3% for both).



Confusion Matrix

- The Decision Tree confusion matrix showed the lowest number of wrong classifications, proving it to be the most accurate accounting for false results.



Conclusions

- We can conclude that:
- Success rate has increased over time, starting with first successes in 2013
- The best launch-site is KSC LC-39A
- Higher payloads correlate with higher success-rates, but may not indicate causation, as it is likely payloads at risk increased as success-rates increased.
- Overall mission success has been consistently high, but landing outcomes have dramatically improved over time.
- Decision Tree Classification is the best choice for predicting successful landings

Appendix

- All used project resources are available in the github repository at: https://github.com/wfaler/data-science-capstone/
- In certain cases, such as the SQL tasks, as well as Dashboard tasks, adjustments/deviations from instructions were done, due to timezone formatting and incompatible dependencies due to updated libraries.

Thank you!