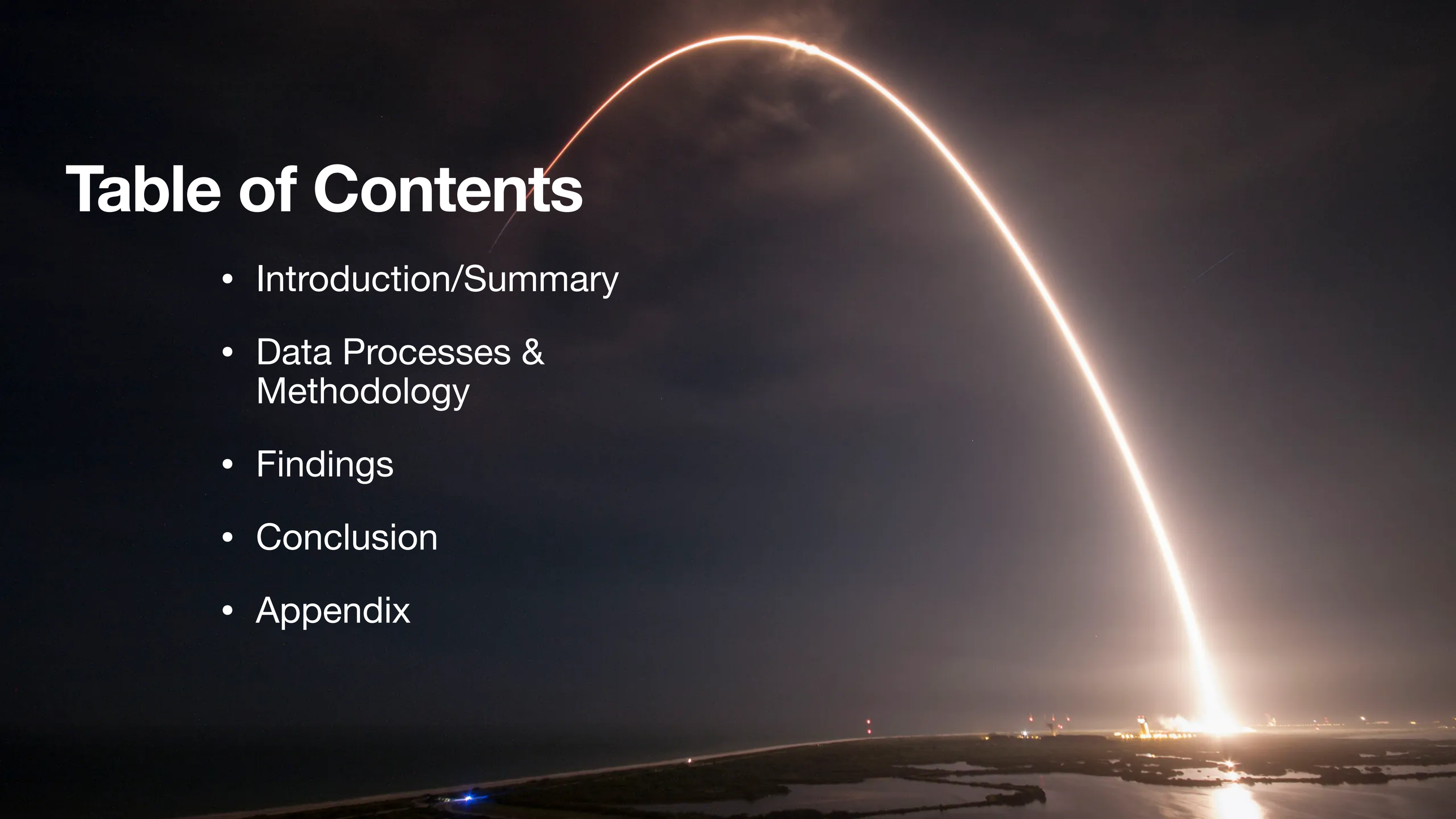
Applied Data Science Capstone

11 July 2024

G. Oistad

Table of Contents



Project Overview & Goals

The objective is to evaluate the viability of the new company Space Y to compete with Space X.

Desirable answers:

- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
- Where is the best place to make launches.

Executive Summary

The following methodologies were used to analyze data:

- Data Collection using web scraping and SpaceX API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics
- Machine Learning Prediction.

- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Data Methodologies

Data Collection

- **Sources of Data:**
  - **SpaceX API:** https://api.spacexdata.com/v4/rockets/
  - **Web Scraping:** Wikipedia - List of Falcon 9 and Falcon Heavy launches

Data Wrangling

Data Enrichment:

- Created a landing outcome label based on outcome data
- Summarized and analyzed features

Exploratory Data Analysis (EDA)

- **Visualization Techniques:**
  - Utilized various visualizations (Including SQL) to perform EDA

SPACEX

Interactive Visual Data Analytics

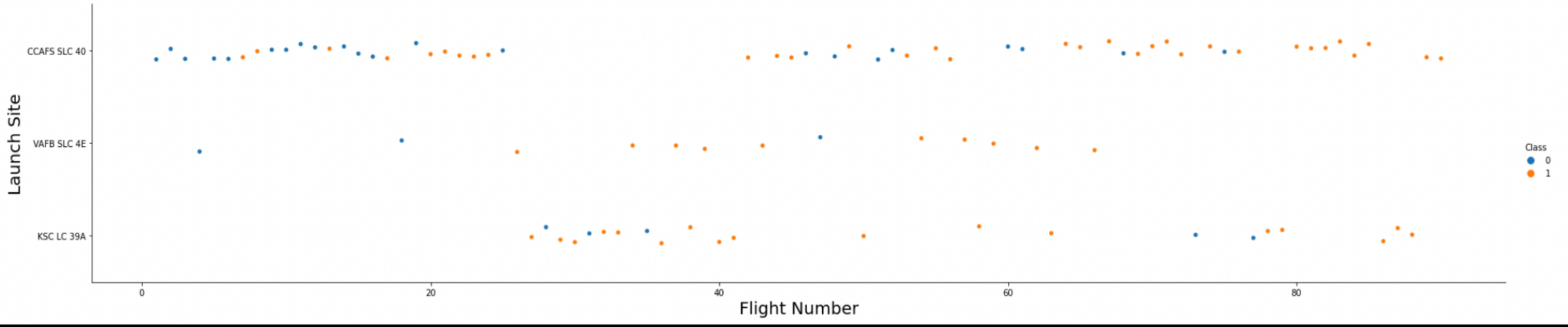
- Utilized SQL, Plotly Dash, & Folium within Jupyter Notebook/Cloud IDE

Predictive Data Analysis

- Conducted using four separate classification models with test training data. The accuracy of each model was determined using multiple, unique parameters.

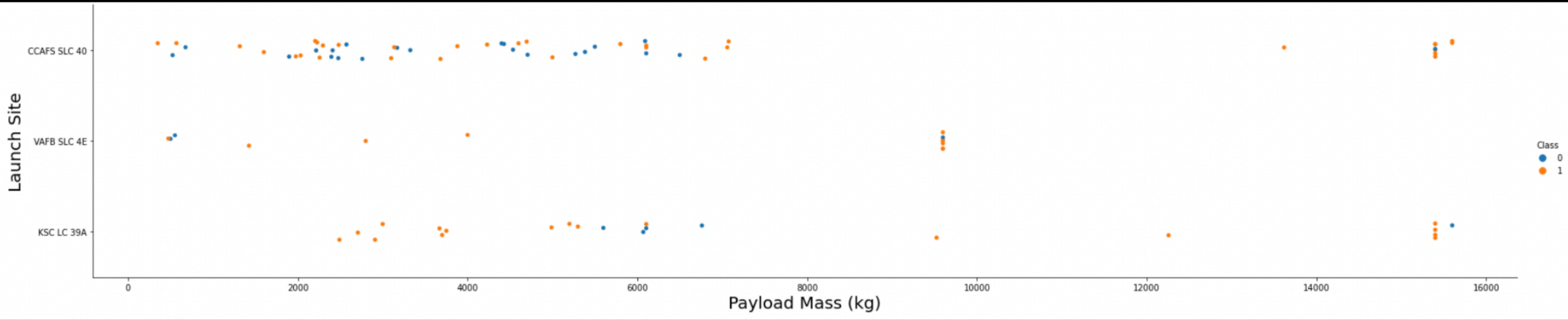
Data Driven Results

EDA Results - Launch Site Successes



- The earliest flights all failed while the latest flights were all successful
- The CCAFS SLC 40 launch site has the most number of launches
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success due to continuous process improvement and regular R&D.

EDA Results - Launch Site Payload

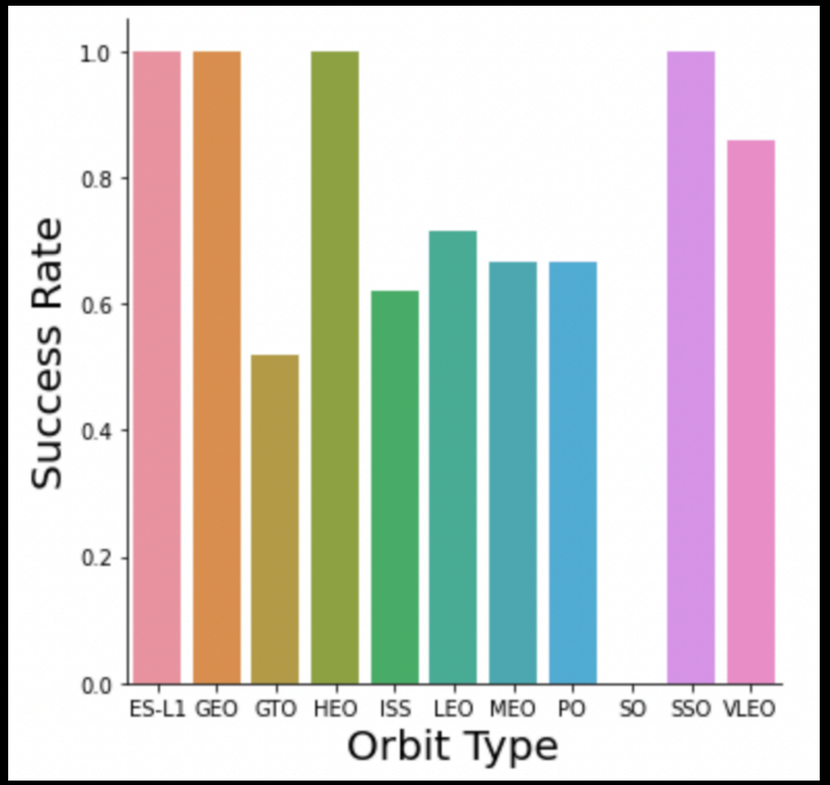


- The higher the payload mass=the higher the success rate
- A majority of launches with payload mass >=7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg

SPACEX

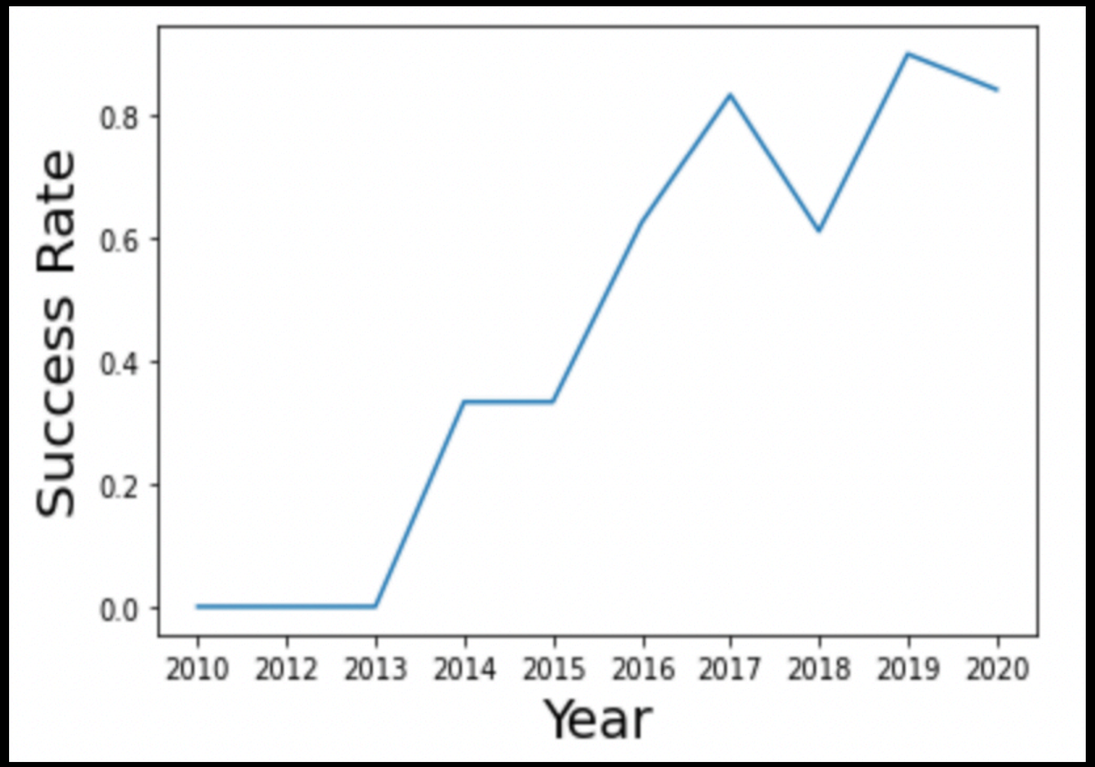
EDA Results - Successes Per Orbit

- 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 Results - Successes Per Orbit

Successful Launch rates have been increasing expeditiously since 2010 with 10+ years of research and development



EDA w/ SQL Results - Overview

Generating list of all launch sites.

```
In [15]: %sql select distinct launch_site from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

```
In [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Date	Time_(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	L
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Generating list of all launch sites including “CCA”.

EDA w/ SQL Results - Overview

```
In [17]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[17]: total_payload_mass
         45596
```

Generated Total Payload moved: 45596KG

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
 * sqlite:///my_data1.db
Done.
 average_payload_mass
 2534.6666666666665
```

Average Payload moved: 2545KG

SPACEX

EDA w/ SQL Results - Overview

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
2015-12-22

First Sucessful Launch: 22 December 2015

```
%sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ b
```

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

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

Boosters that have successfully landed with payload between 4k & 6k KG

SPACEX

EDA w/ SQL Results - Overview

Successful/Failed Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

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

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTB
```

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

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

Boosters that have carried maximum payload

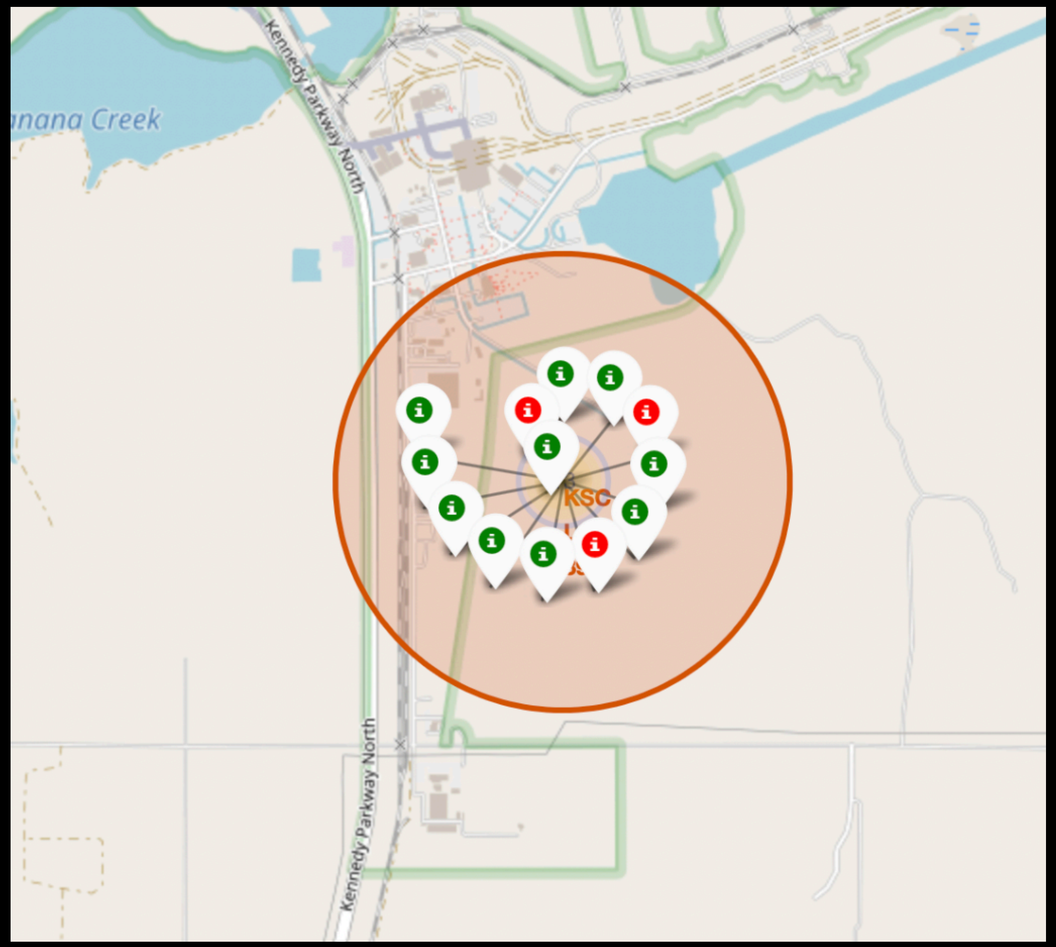
Interactive Map w/ Folium - Launch Locations

• Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay 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 w/ Folium - Launch Locations

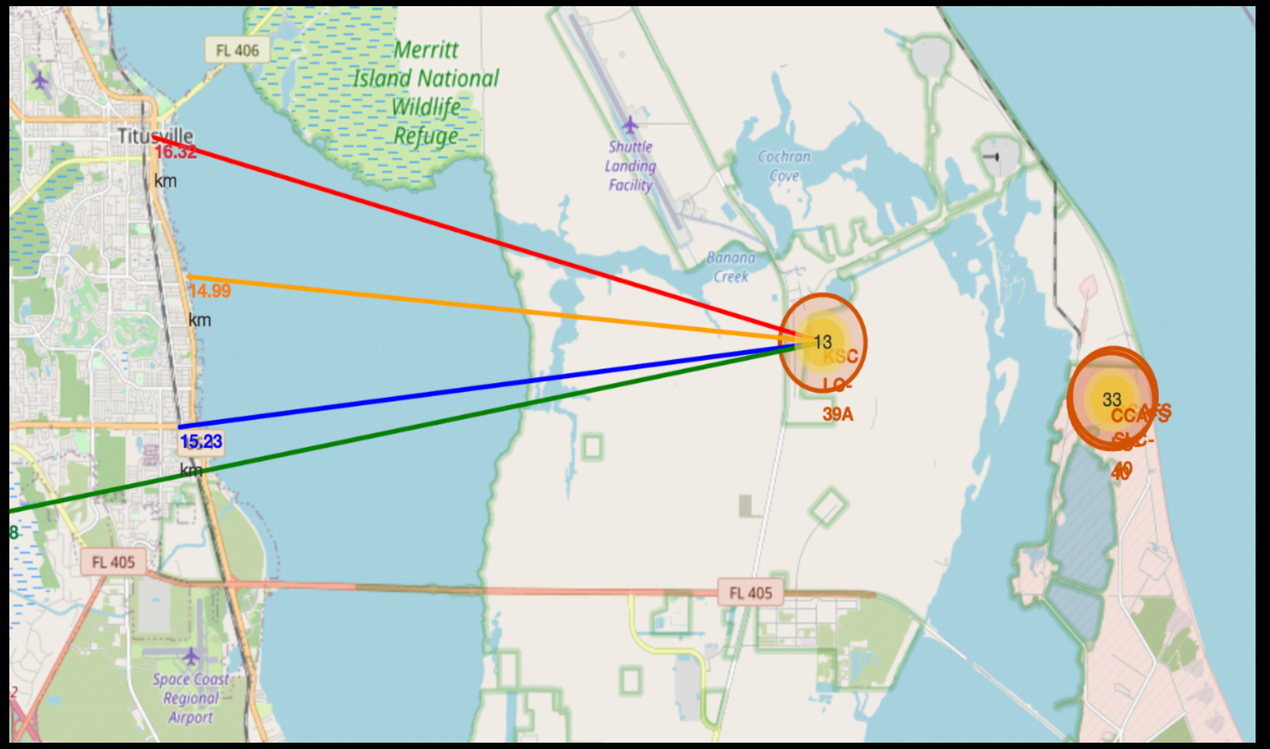
- Green Marker = Successful Launch
- Red Marker = Failed Launch

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

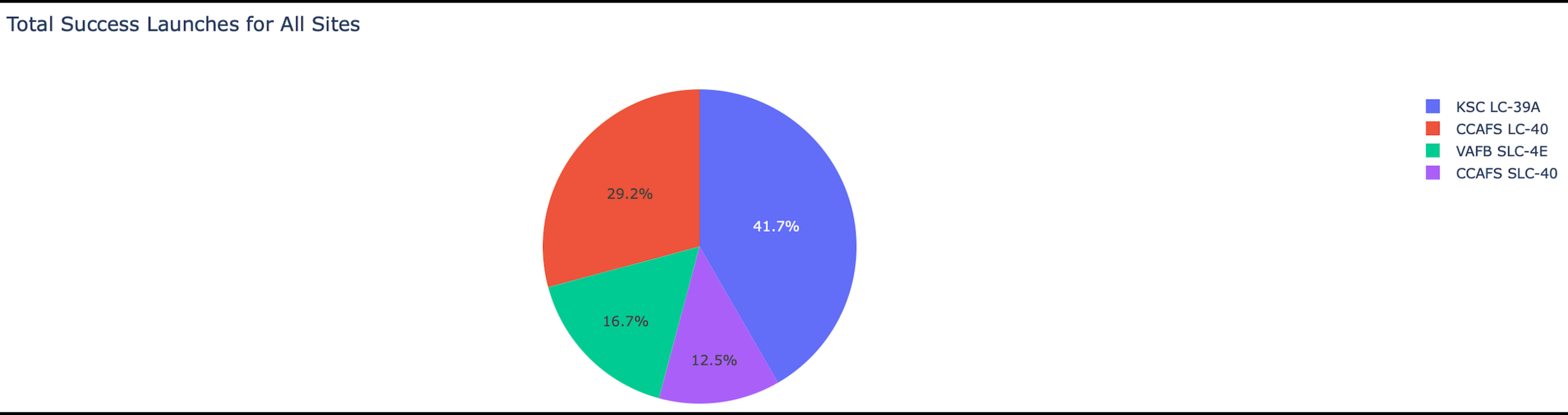


Interactive Map w/ Folium - Launch Locations

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.



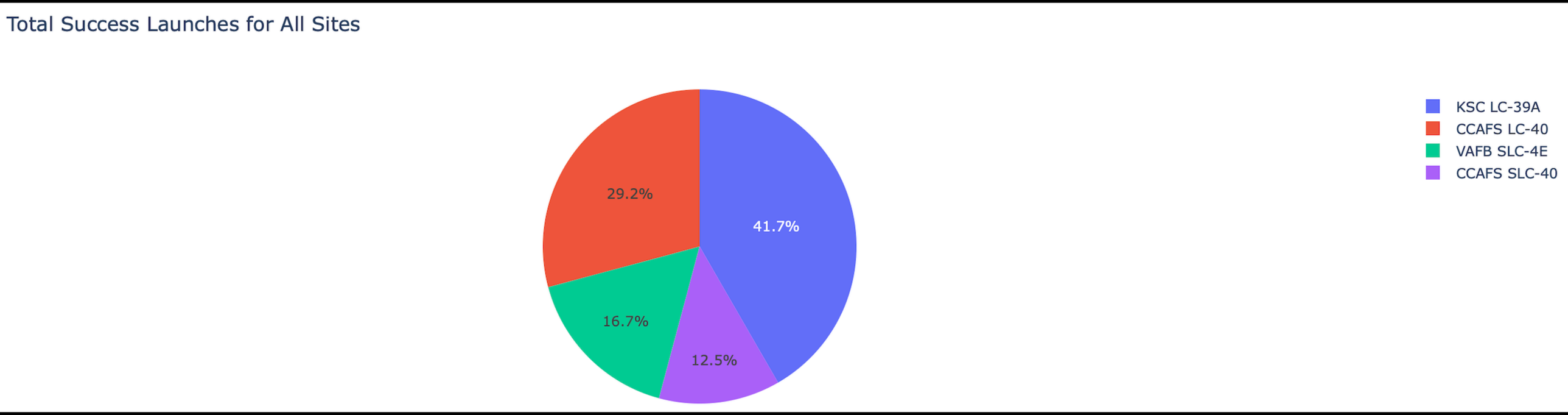
Interactive Application with Plotly Dash



Interactive Pie Chart showing the percentage of successful launches per launch locations

SPACEX

Interactive Application with Plotly Dash



Interactive Pie Chart showing the percentage of successful launches per launch locations

Drop Down Menu to sort by launch site

Interactive Application with Plotly Dash

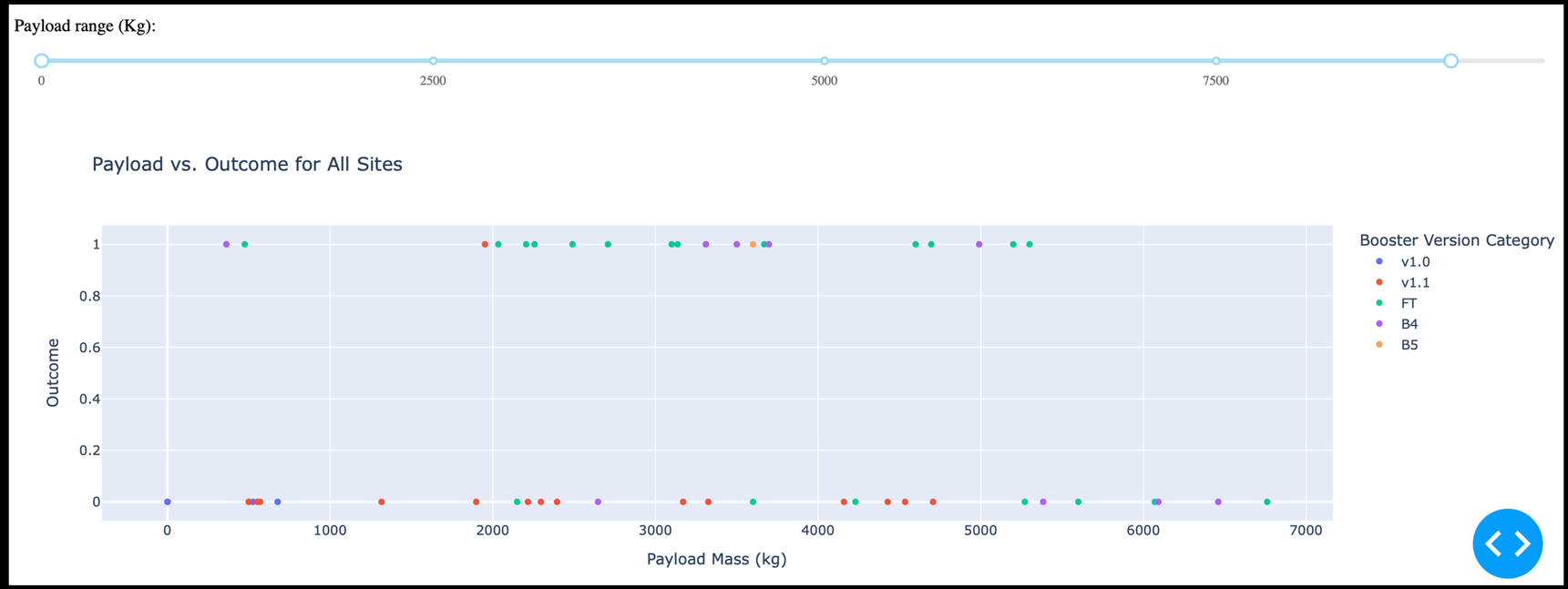


Chart showing payload & success outcomes per booster type. Also is an added slider bar so that you can sort the chart by total payload if required.

Predictive Analysis/Machine Learning

Based on the scores of the Test Set, we can not confirm which method performs best…

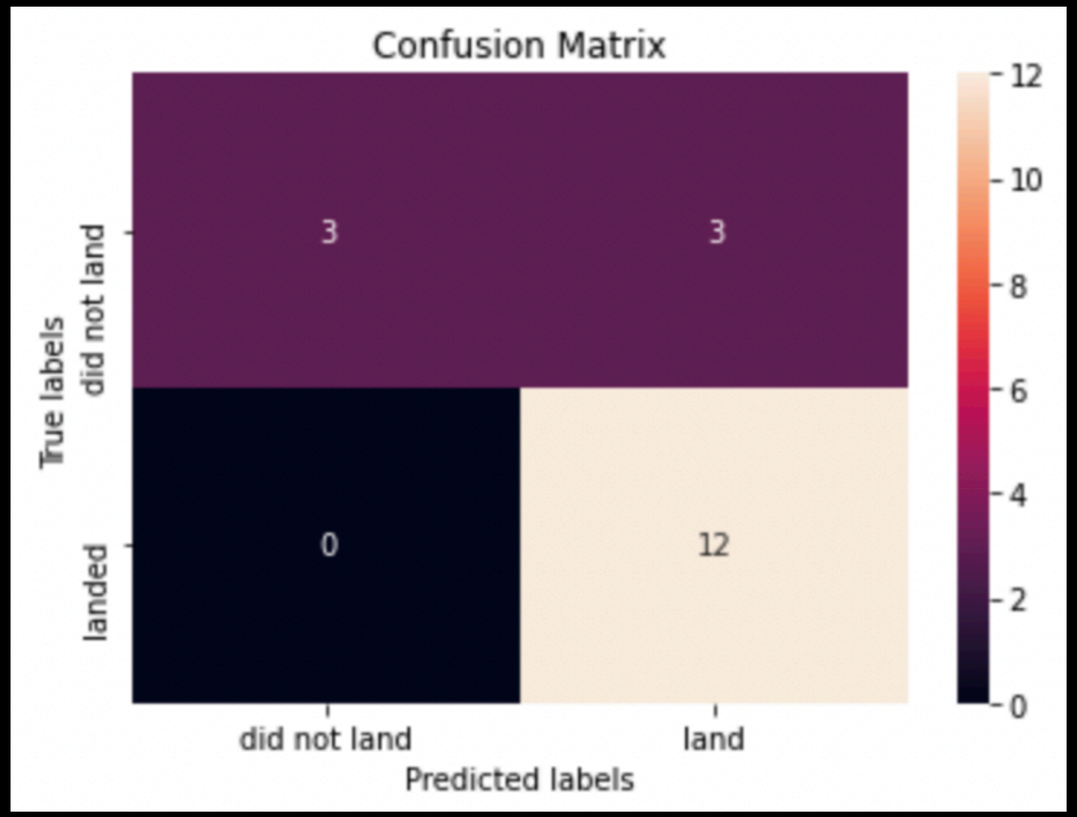
• Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
• The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

	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

	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

Predictive Analysis/Machine Learning

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives



Conclusion

- Decision Tree Model is the best algorithm for this dataset.

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

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

- The success rate of launches increases over the years.

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

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