Winning Space Race with Data Science

D. F.
2022.05.12

Outline

Executive Summary

- Using several machine learning algorithms, we will predict whether the SpaceX Falcon 9 first stage will land successfully. To make the prediction, we first collected, wrangled, and formatted the data. Then, exploratory data analysis was performed, which was then used to create interactive data visualization, to aid in understanding the data and what it meant. Multiple machine learning algorithms were implemented and compared to find which model would produce the best results for our data and purpose.
- Overall, the rate of rocket launch success is increasing for SpaceX. There is a relatively high rate of launches occurring at the Kennedy Space Center but relatively low success rate for launches at Cape Canaveral Launch Complex 40. We can predict with about 83.33% accuracy whether or not the next launch that SpaceX does will be a success or failure.

Introduction

Space travel is becoming more affordable, making commercial space flight possible. SpaceX does this effectively by reusing the first stage of their rockets, enabling them to reduce the average cost of 165 million dollars per launch for other providers to 62 million dollars as advertised on their website. However, based on mission parameters, SpaceX does not always preserve the first stage. We aim to predict whether or not SpaceX will reuse the first stage of the Falcon 9 rocket based on a variety of its features. This will aid in predicting the price of a launch and can be used in a bid for a rocket launch against SpaceX.

Section 1

Methodology

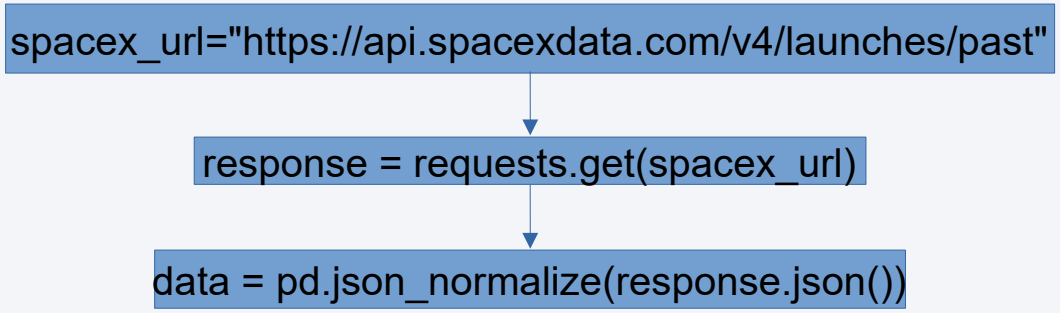
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping
- Perform data wrangling
  - Pandas
  - NumPy
- Perform exploratory data analysis (EDA)
  - Matplotlib
  - Seaborn
  - SQL
- Perform interactive visual analytics
  - Folium
  - Plotly Dash
- Perform predictive analysis using classification models – Machine Learning
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection – SpaceX API

- SpaceX API [GitHub Notebook]
  - SpaceX REST API has information on the launches such as launch specifications, landing specifications, landing outcome, etc.
  - We are using endpoint https://api.spacexdata.com/v4/launches/past
  - Using a get request, launch data was obtained
  - The launch data was viewed in the form of a JSON and then normalized into a table
  - Only the Falcon 9 data was kept – there was Falcon 1 included in the original data that was discarded
  - The final table included 90 rows and 17 feature columns



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- Web scraping [GitHub URL]
  - The HTML records were scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 with BeautifulSoup
  - This only contained Falcon 9 data
  - Data was parsed from HTML tables and converted to Pandas data frame
  - The final table had 121 rows and 11 column features

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
        ↓
contents = urllib.request.urlopen(static_url).read()
        ↓
html_tables = soup.find_all('table')
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	[SpaceX, \n]	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	[\n, NASA, (, COTS,), \n, NRO, \n, \n]	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	[NASA, (, COTS,)\n]	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	[NASA, (, CRS,)\n]	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[NASA, (, CRS,)\n]	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- [GitHub URL]
- The data was further processed (wrangled) to address the missing value entries.
- For some of the data, such as the “PayloadMass” feature, the mean value for all entries replaced the empty entries.
- Categorical features were encoded with one-hot encoding
- An additional feature “class” was added to the data frame. This contained either a 1 for successful launch or a 0 for a failed launch

EDA with Data Visualization

[GitHub]
Scatter plots were used to visualize if variables are correlated to one another.

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Flight Number vs Orbit Type
- Payload vs Orbit Type

Bar charts were used to compare categories with discrete values

- Orbit Type Success Rate

Line Charts were used to show trends in the data over time

- Launch Success Yearly Trend

EDA with SQL

[GitHub]

- Display the unique names of the launch sites
- Display 5 records where launch sites begin with string ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the names of the boosters which have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for 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

Build an Interactive Map with Folium

[GitHub]

Folium makes it easy to visualize data in interactive maps. Each launch site was plotted with its latitude and longitudinal coordinates and labeled. Each launch sites had its corresponding launches surrounding it marked as success or failure.

- Map Markers (folium.Marker()) is the object to to make marks on the map
- Circle Markers (folium.Circle()) creates a circle at the marker coordinates
- Icon Markers (folium.Icon()) creates icons at marker coordinates
- Polyline (folium.Polyline()) draws a line between points
- Marker Cluster (MarkerCluster()) clusters together multiple markers at the same coordinates

Build a Dashboard with Plotly Dash

[GitHub]

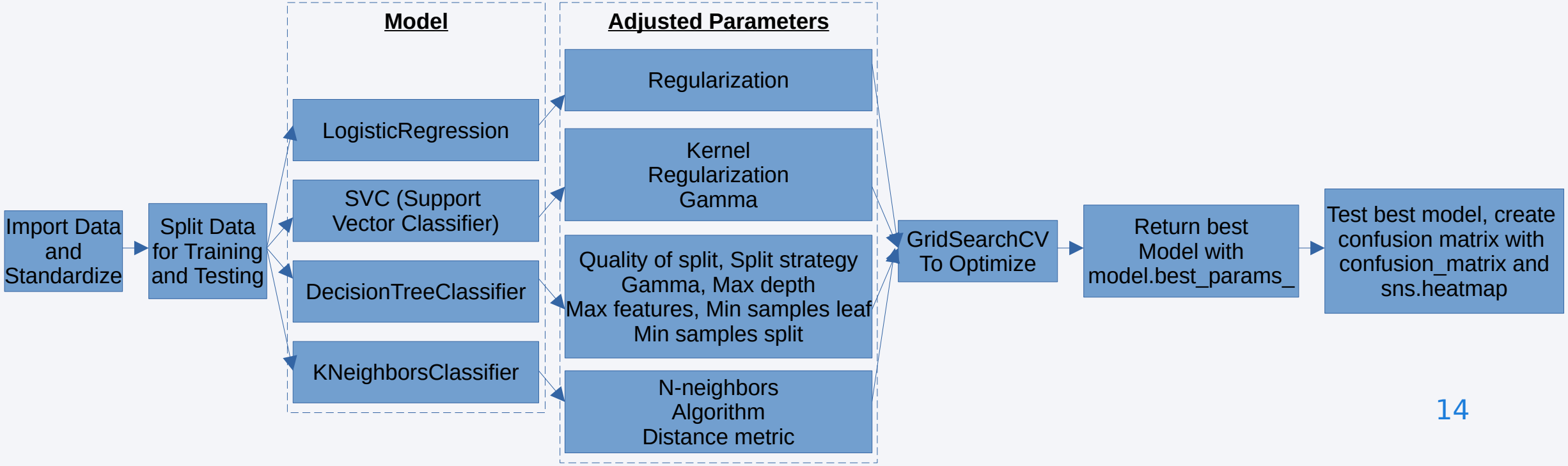
Insights can be made from an interactive dashboard that may be more difficult to make from static plots. Here, the dashboard includes:

- Dropdown list (dcc.Dropdown()) which creates a dropdown list of launch site options to select from
- Range slider (dcc.RangeSlider()) which creates a bar with sliders to select the range of payload mass we want to observe data from
- Pie chart (px.pie()) which displayed the launch success rate of whatever site(s) was selected from the dropdown
- Scatter plot (px.scatter()) which created a scatter plot to find correlations based on the launch site(s) and payload range selected

Predictive Analysis (Classification)

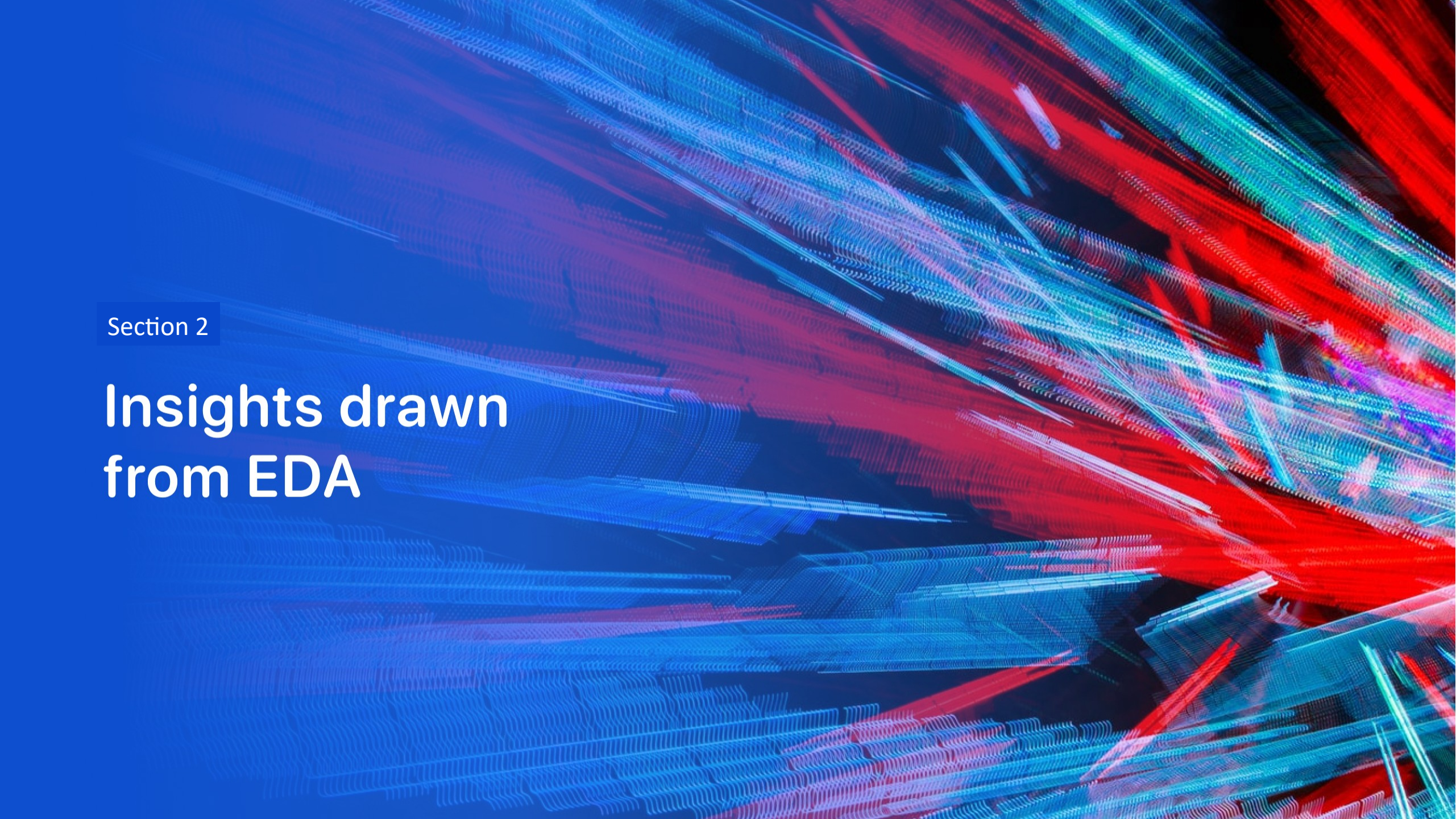
[GitHub]

Four models were built, evaluated, improved, selected for best performance, and compared from the sklearn library:



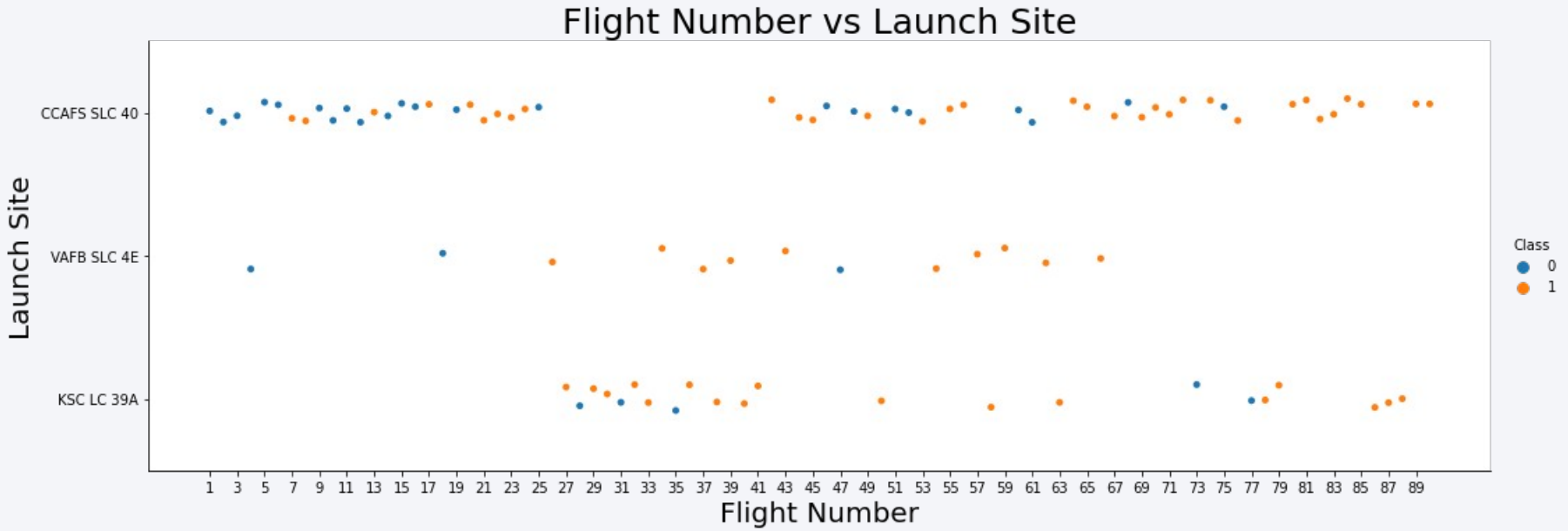
Results

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


Section 2
Insights drawn from EDA

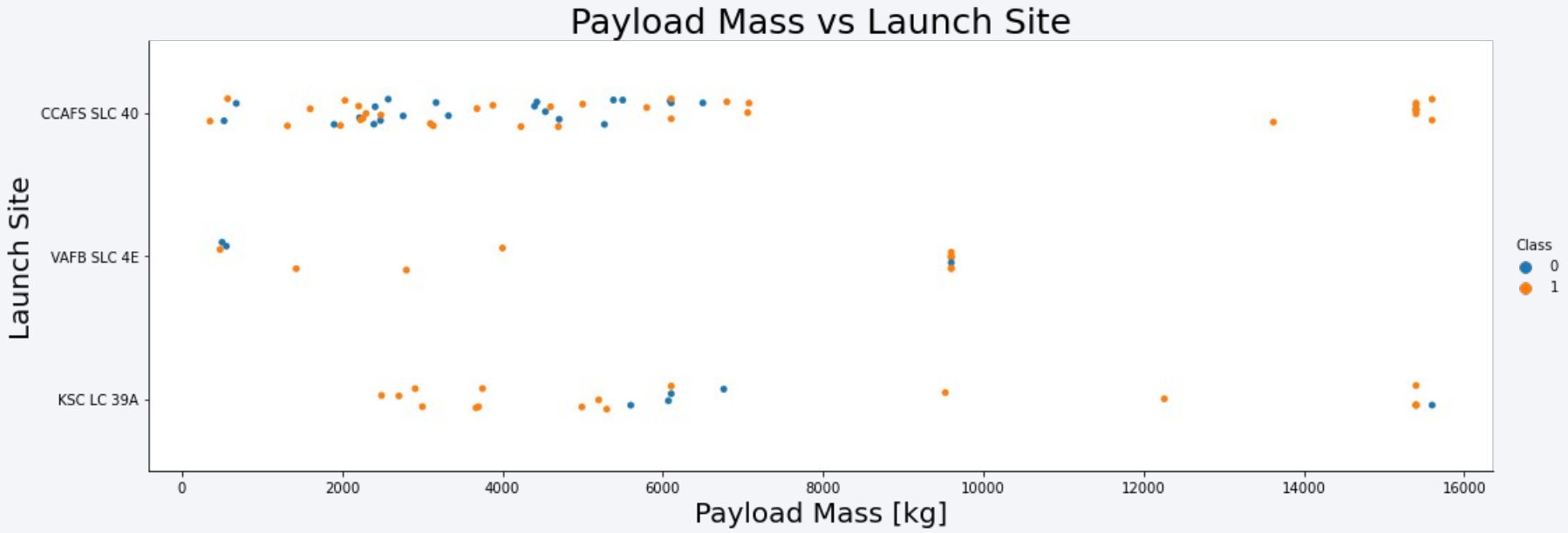
Flight Number vs. Launch Site

- Flight Number vs. Launch Site
  - As flight number increases, the rate of success increases
  - The largest number of flights occurred at CCFAS



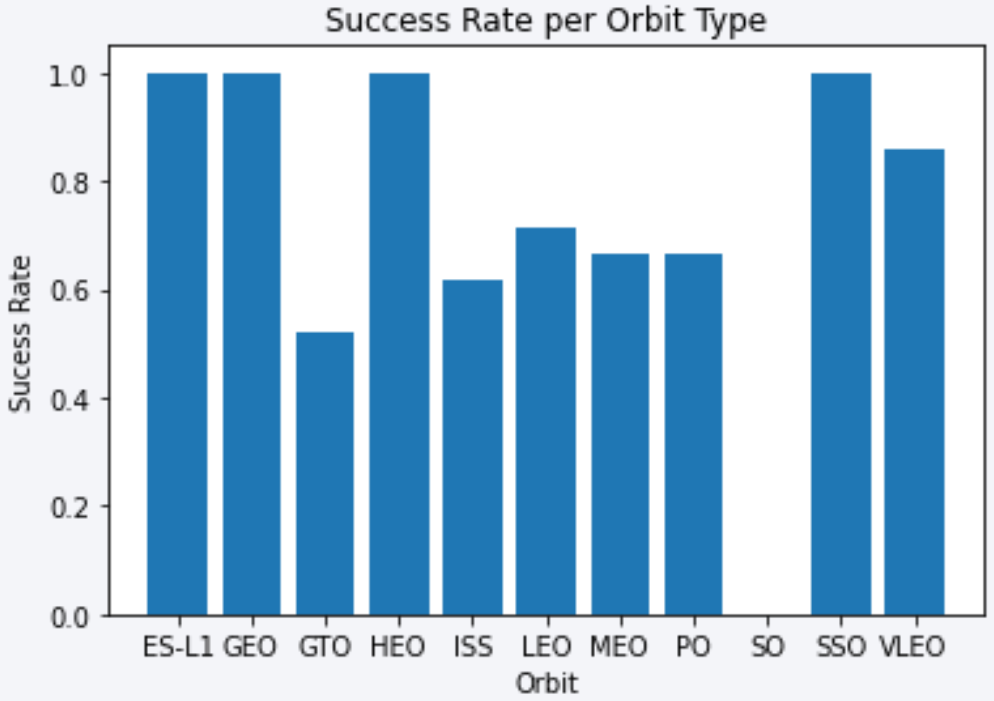
Payload vs. Launch Site

- Payload Mass vs. Launch Site
  - As the payload mass increases, the rate of success increases
  - The largest number of missions with a relatively low payload mass occurred at CCAFS



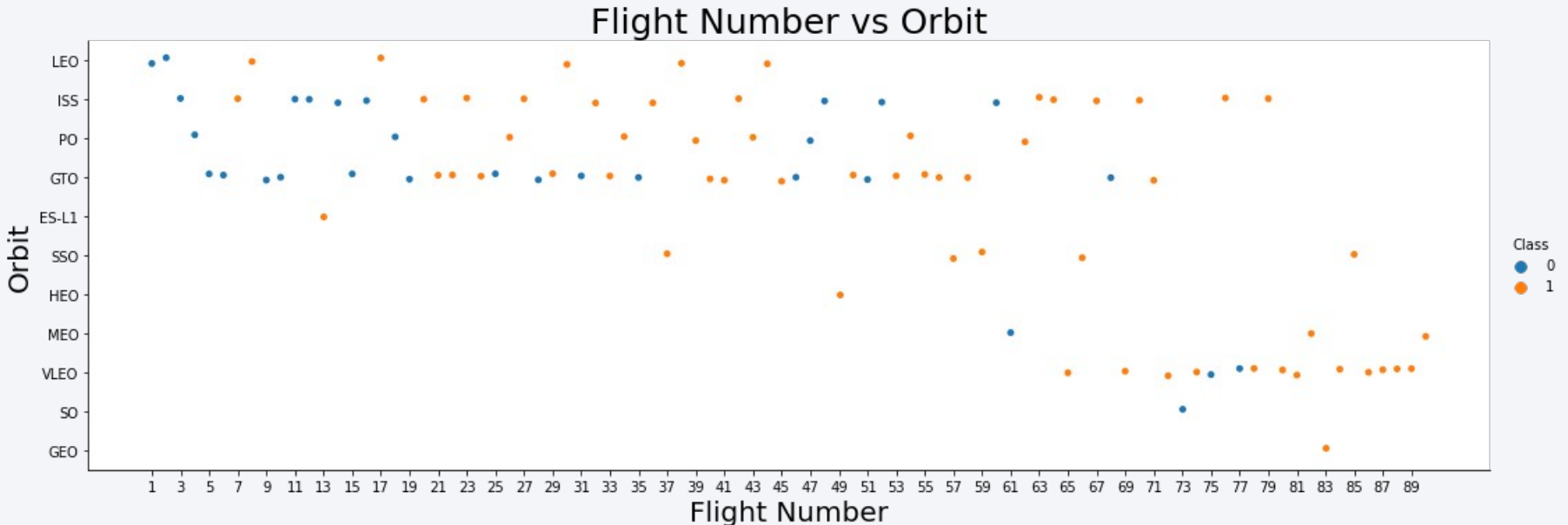
Success Rate vs. Orbit Type

- Success rate of each orbit type
  - The most successful launch percentages occur at the ES-L1, GEO, HEO, and SSO with 100% success.
  - The least successful launch percentage occurs at SO



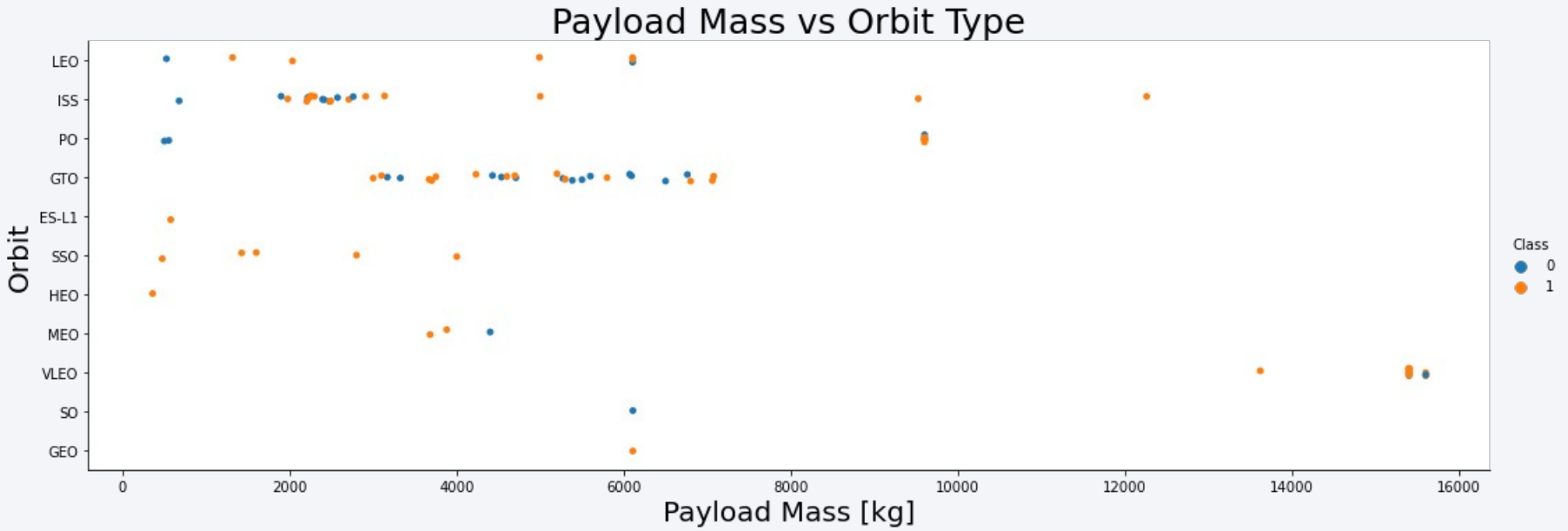
Flight Number vs. Orbit Type

- Flight number vs. Orbit type
  - For LEO, success seems to be dependent on flight number
  - For ES-L1, SSO, HEO, and GEO, all missions were successful
  - The number of flights occurring at the orbits of further distance increase as flight number increases, as well as the rate of success



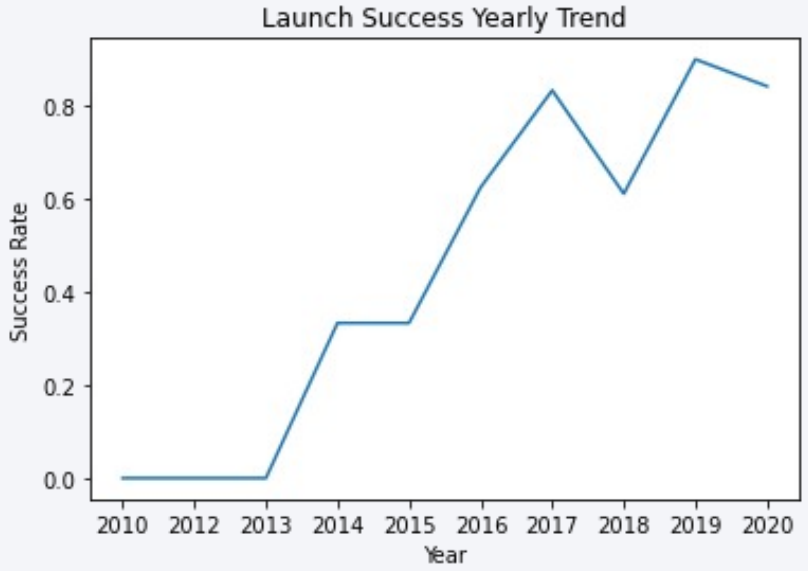
Payload vs. Orbit Type

- Payload vs. orbit type
  - There does not seem to be a strong correlation between payload mass, orbit type, and success rate



Launch Success Yearly Trend

- Yearly average success rate
  - Launches are having a higher rate of success over time



All Launch Site Names

To find the names of the unique launch sites, we SELECT the DISTINCT names from the column LAUNCH_SITE within the table SPACEXTBL

```
%%sql

SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

To find 5 records where launch sites begin with `CCA`, we SELECT all the columns (*) from the table SPACEXTBL where the column LAUNCH_SITE is has ‘CCA’ at the beginning (LIKE ‘CCA%’) and only LIMIT the output to 5 records

```
%%sql

SELECT *
FROM SPACEXTBL
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

To calculate the total payload carried by boosters from NASA, we SELECT the calculated SUM in the column PAYLOAD_MASS_KG from the table SPACEXTBL where the column CUSTOMER is ‘NASA (CRS)’

```
%%sql

SELECT SUM(PAYLOAD_MASS_KG)
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

To calculate the average payload mass carried by booster version F9 v1.1, we SELECT the calculated AVERAGE in the column PAYLOAD_MASS_KG from the table SPACEXTBL where the column BOOSTER_VERSION is ‘F9 v1.1’

```
%%sql

SELECT AVG(PAYLOAD_MASS_KG)
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

1
2928

First Successful Ground Landing Date

To find the date of the first successful landing outcome on the ground pad, we SELECT the minimum (MIN) DATE from the table SPACEXTBL where the column LANDING_OUTCOME is ‘Success (ground pad)’

```
%%sql

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

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

To find the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, we SELECT the columns BOOSTER_VERSION and PAYLOAD_MASS_KG from the table SPACEXTBL where the column LANDING_OUTCOME is ‘Success (drone ship)’ and PAYLOAD_MASS_KG is BETWEEN 4000 and 6000

```
%%sql

SELECT BOOSTER_VERSION, PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE (LANDING_OUTCOME='Success (drone ship)' )
AND PAYLOAD_MASS_KG BETWEEN 4000 AND 6000;
```

booster_version	payload_mass_kg
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

To calculate the total number of successful mission outcomes, we SELECT the COUNT in the column MISSION_OUTCOME from the table SPACEXTBL where the column MISSION_OUTCOME has ‘Success’ at the beginning (LIKE ‘Success%’)

To calculate the total number of failure mission outcomes, repeat the above except MISSION_OUTCOME has ‘Failure’ at the beginning (LIKE ‘Failure%’)

```
%%sql

SELECT COUNT(MISSION_OUTCOME)
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Success%';
```

1
100

```
%%sql

SELECT COUNT(MISSION_OUTCOME)
FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Failure%';
```

1
1

Boosters Carried Maximum Payload

To find the list the names of the booster which have carried the maximum payload mass, we SELECT the columns BOOSTER_VERSION and PAYLOAD_MASS_KG from the table SPACEXTBL where the PAYLOAD_MASS_KG is equal to the value produced when we SELECT the calculated maximum (MAX) of the column PAYLOAD_MASS_KG from the table SPACEXTBL

```
%%sql

SELECT BOOSTER_VERSION, PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE PAYLOAD_MASS_KG = (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTBL);
```

booster_version	payload_mass_kg
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

To find the list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, we SELECT the columns LANDING_OUTCOME, DATE, BOOSTER_VERSION, and LAUNCH_SITE from the table SPACEXTBL where the column LANDING_OUTCOME is ‘Failure (drone ship)’ and the YEAR of the column DATE is equal to 2015

```
%%sql

SELECT LANDING_OUTCOME, DATE, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME='Failure (drone ship)' AND YEAR(DATE)=2015;
```

landing_outcome	DATE	booster_version	launch_site
Failure (drone ship)	2015-01-10	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

To 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, we SELECT the LANDING_OUTCOME and the COUNT all columns (*), creating column totalCount, from the table SPACEXTBL where the DATE is BETWEEN '2010-06-04' and '2017-03-20'. We GROUP BY the column LANDING_OUTCOME (so the categories in this column are what we are trying to count) and then ORDER BY the column TOTALCOUNT in decending (DESC) order

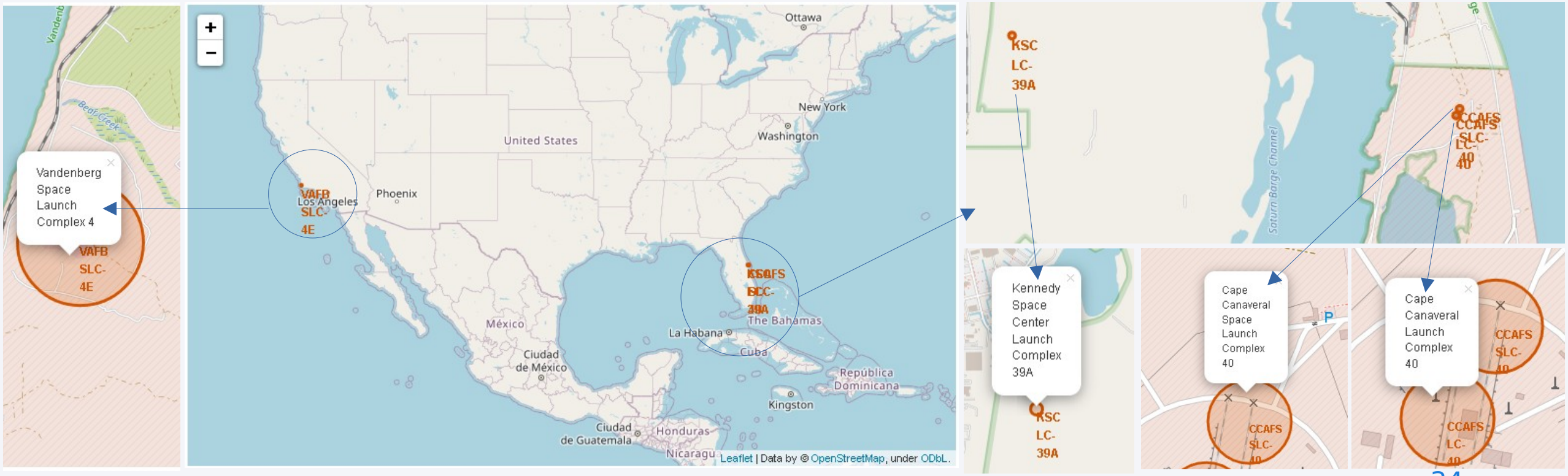
```
%%sql

SELECT LANDING_OUTCOME, COUNT(*) totalCount
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTALCOUNT DESC
```

landing_outcome	totalcount
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

Folium Map – Launch Site Locations

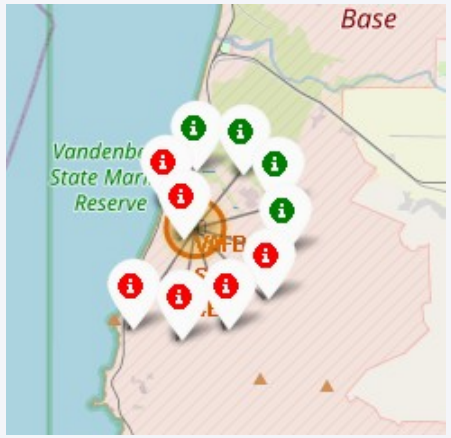
Four locations are displayed on the map: one in California and three in Florida. All locations are along the coastlines and are close to the equator.

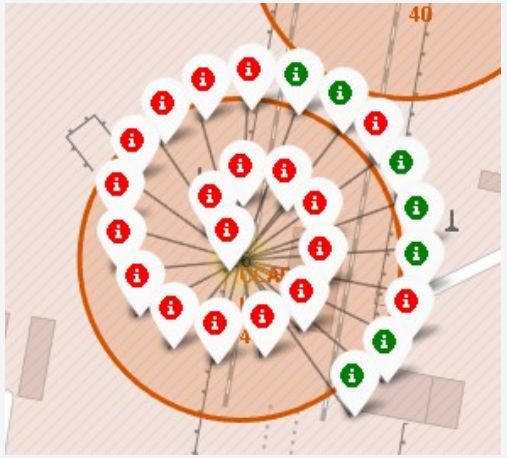


Folium Map – Launch Outcomes

From the colored markers, it can be easily identify which launch sires have relatively high success (green) rate, such as launches as Kennedy Space center, and relatively high failure (red) rate, such as launches at Cape Canaveral Launch Complex 40

Vandenberg Space Launch Complex 4	Cape Canaveral Launch Complex 40	Cape Canaveral Space Launch Complex 40	Kennedy Space Center Launch Complex 39A



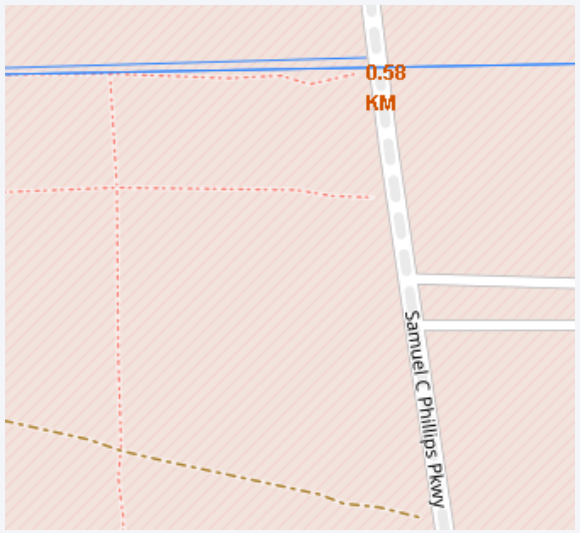


Folium Map – Distance from Launch Site to Proximities

From the lines dawn on the map, we can see that launch sites are relatively close to coast lines, highways, and railroads, but relatively far away from cities

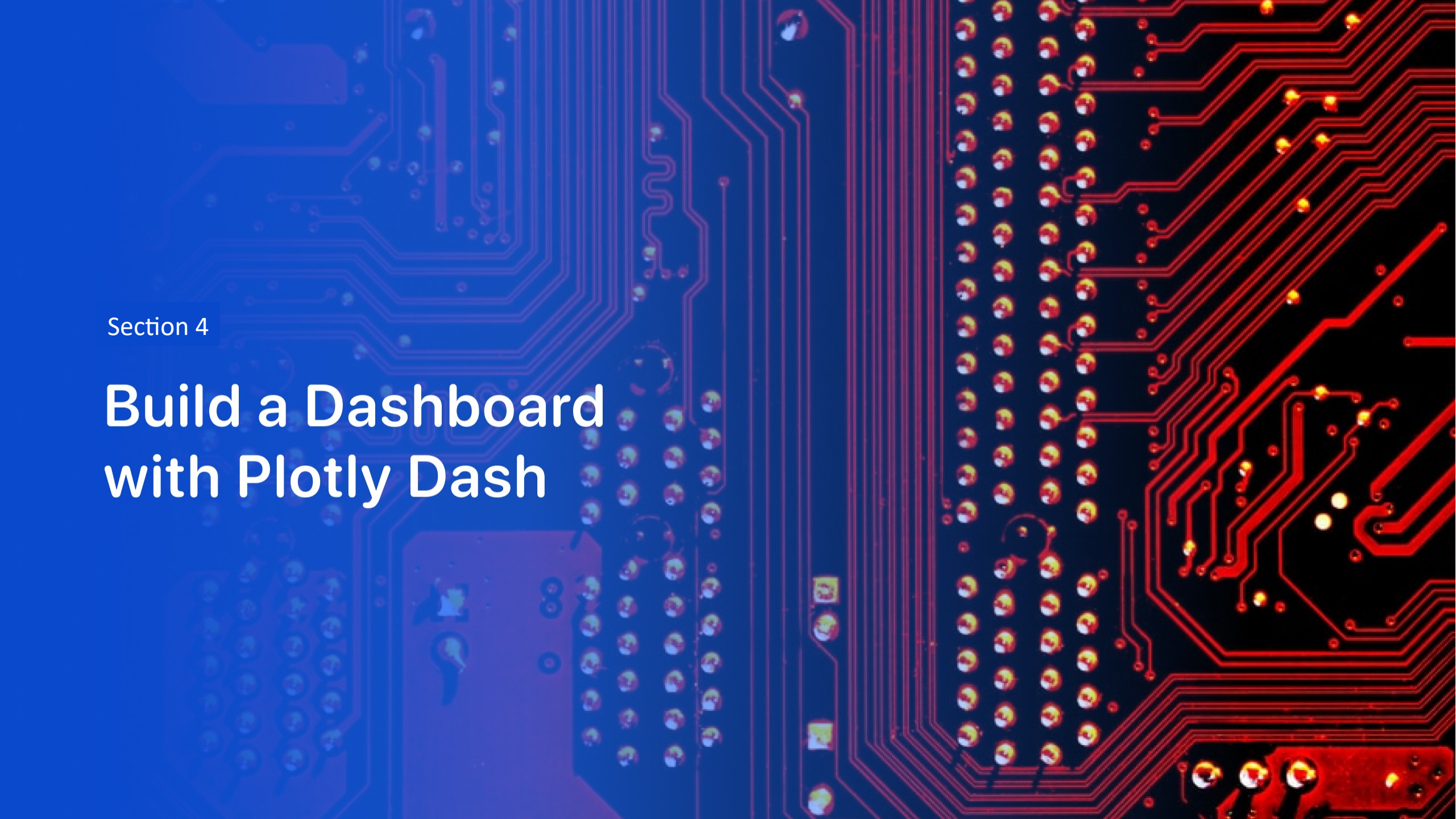
Coastline

Highway



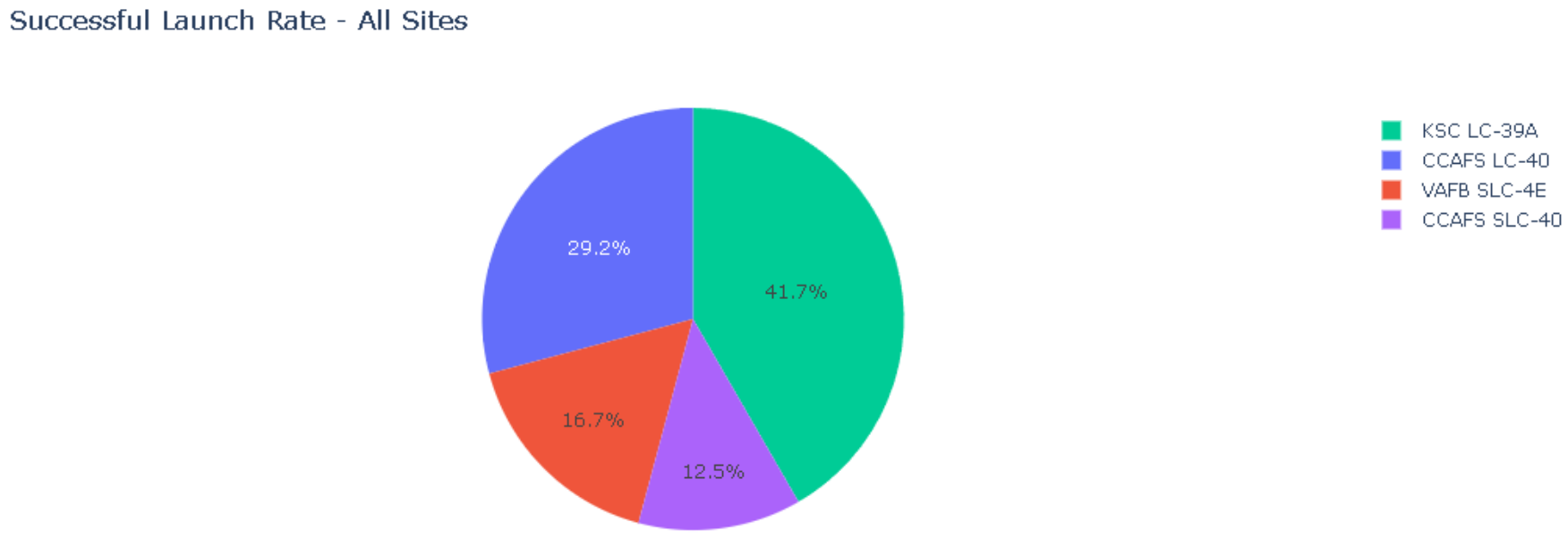
Coastline

Coastline


Section 4
Build a Dashboard with Plotly Dash

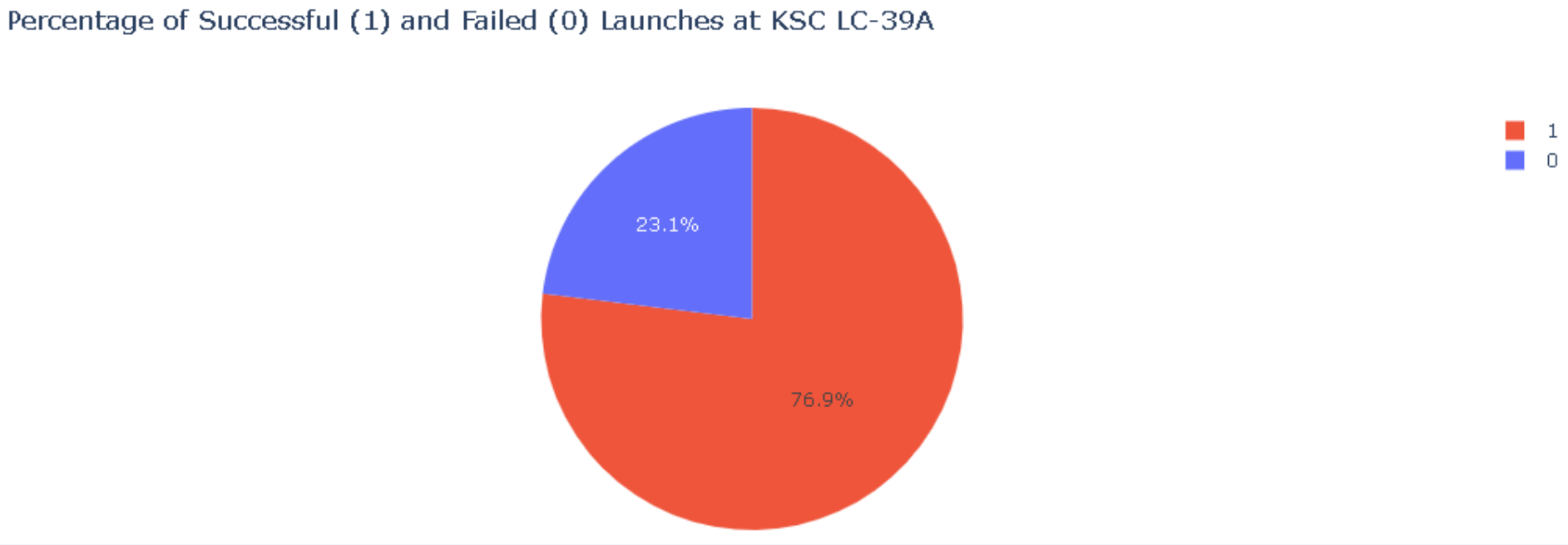
Dashboard – Launch Success Rate

From the pie chart, we can see that, the Kennedy Space Center had the highest number of successful launches, Cape Canaveral Launch Complex 40 had the second highest, Vandenberg Space Launch Complex 4 had the third highest, and Cape Canaveral Space Launch Complex 40 had the fewest.



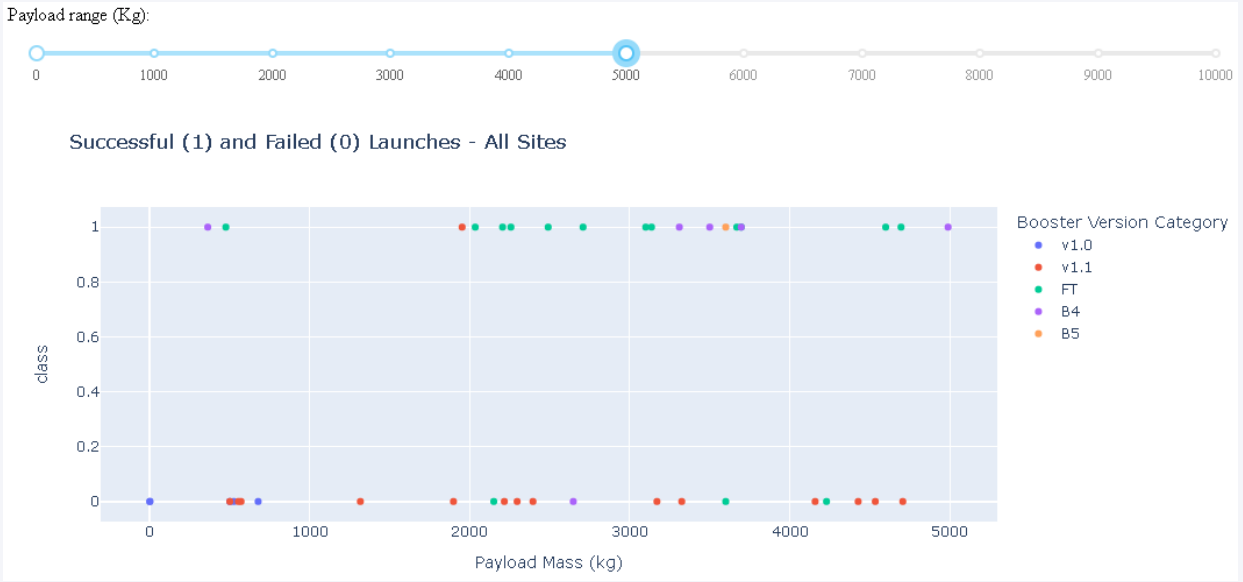
<Dashboard – Launch Success Rate KSC

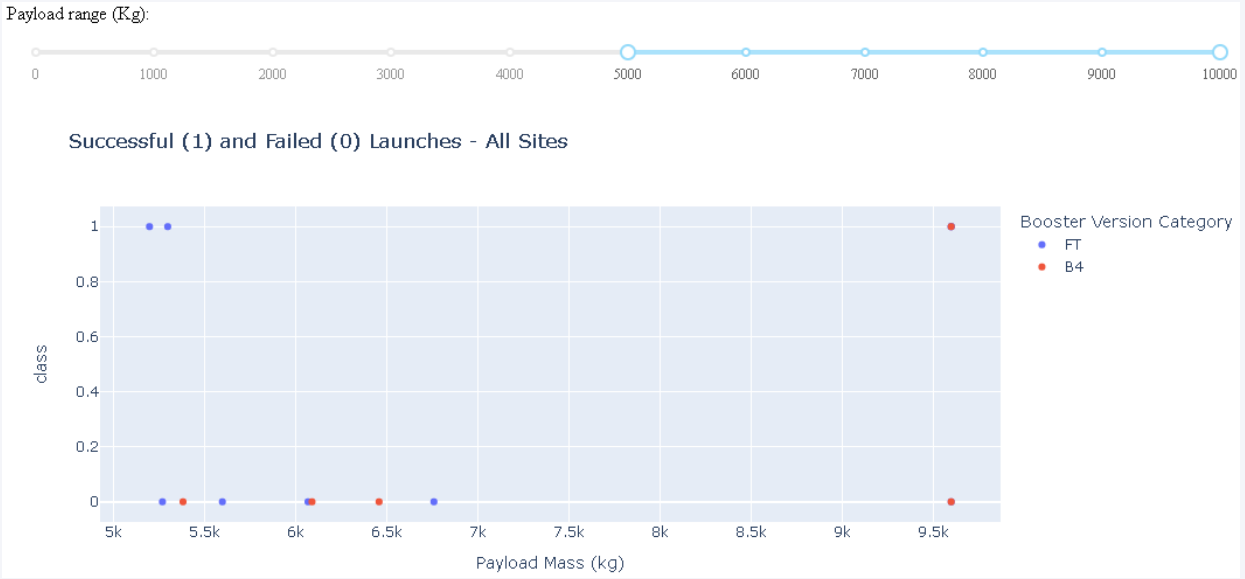
The launch site with highest launch success ratio is the Kennedy Space Center with 76.9% success 23.1% failure. This means that if SpaceX decides to have their next launch at KCS, it is likely to be a success.



Dashboard - Payload vs. Launch Outcome

From the plots, we can see that the success rate when the payload range is between 0 kg and 5000 kg is higher than when the payload range is between 5000 kg and 10000 kg. We can also see that the FT booster has a relatively high success rate and the v1.1 has a relatively low success rate



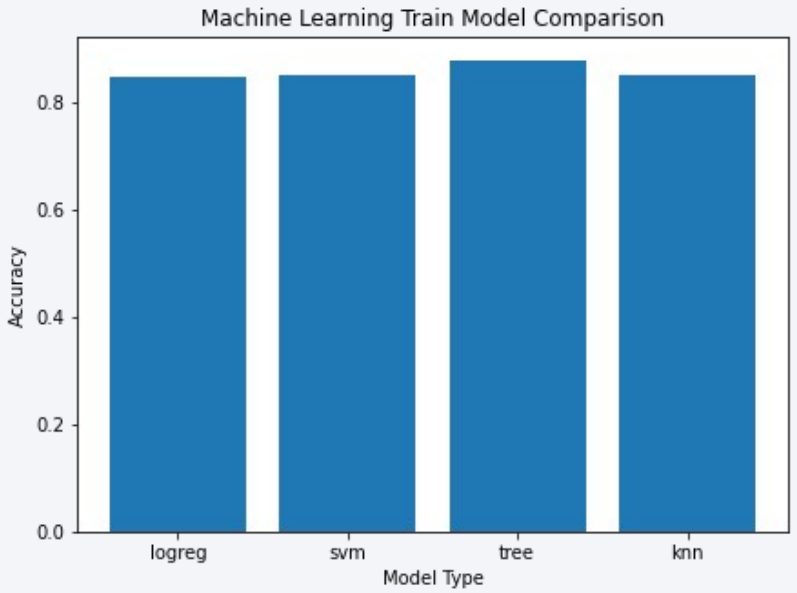


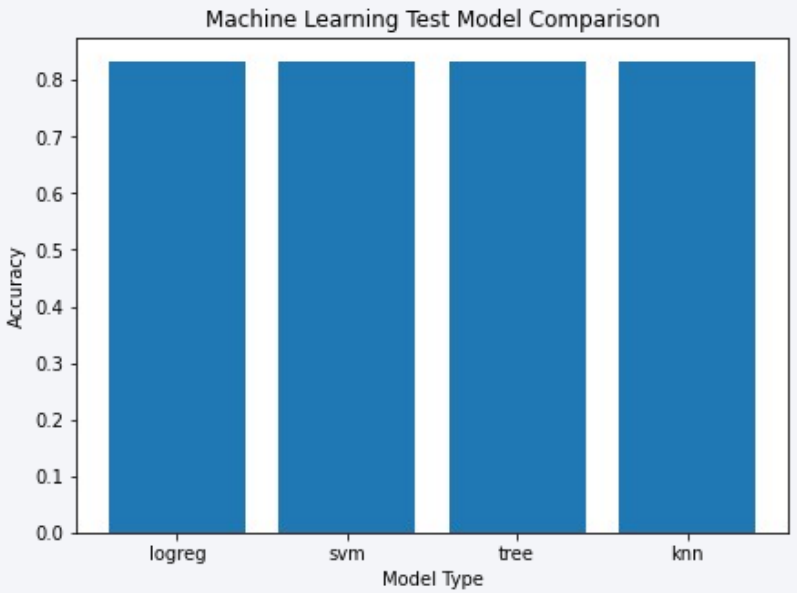
Section 5

Predictive Analysis (Classification)

Classification Accuracy

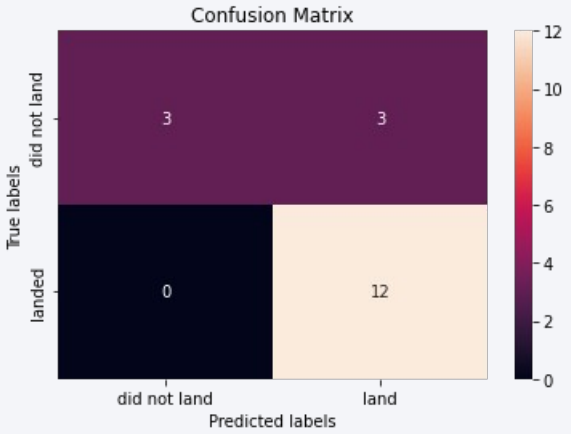
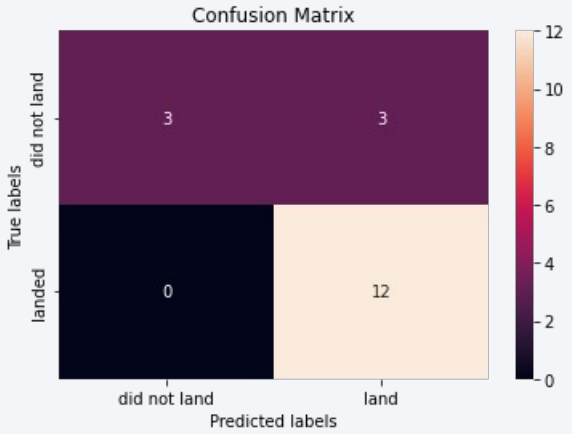
- During training, the decision tree had the highest accuracy. During testing, all four models had the same accuracy.





Confusion Matrix

Based on the confusion matrices of the models, all models perform relatively the same. If we were to continue to fine-tune the parameters, we may find one model becomes slightly better than the rest, but it could be considered “better” due to random error. All models are provide sufficient predictive capabilities.

Linear Regression	SVM	Decision Tree	KNN
			
Accuracy: 83.33%	Accuracy: 83.33%	Accuracy: 83.33%	Accuracy: 83.33%

Conclusions

Over time, the overall rate of rocket launch success increases.

Percentage wise, the most successful launches occur at the ES-L1, GEO, HEO, and SSO with perfect success. However, there are few launches that have occurred in these orbits, so it can not be determined with certainty that, if the next launch occurs at one of these sites, it will be successful because there is not enough data.

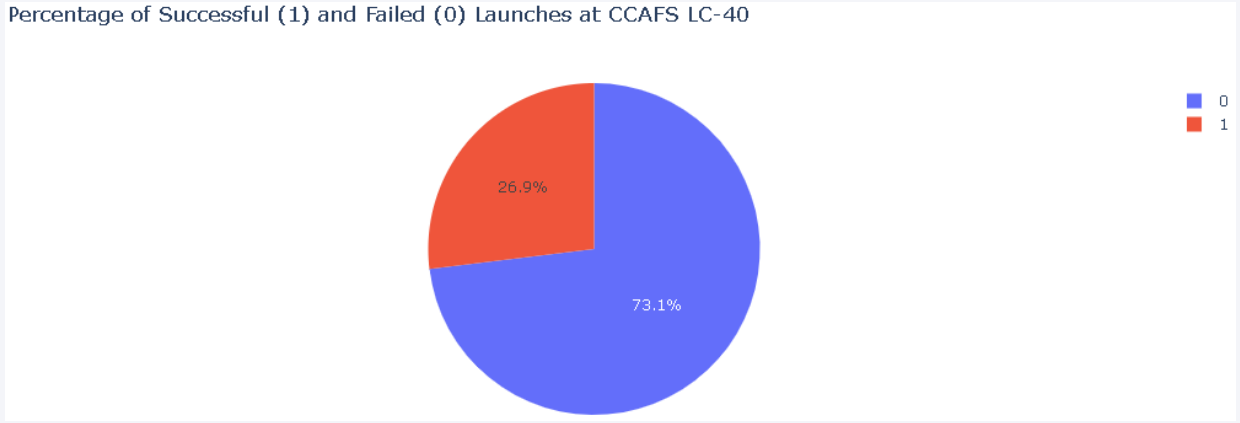
There is a high rate of success for launches occurring at the Kennedy Space Center where 76.9% of the launches are successful and consists of 41.7% of the total successful launches that have occurred It is likely that, if the next launch were to happen from the Kennedy Space Center, that it will be successful.

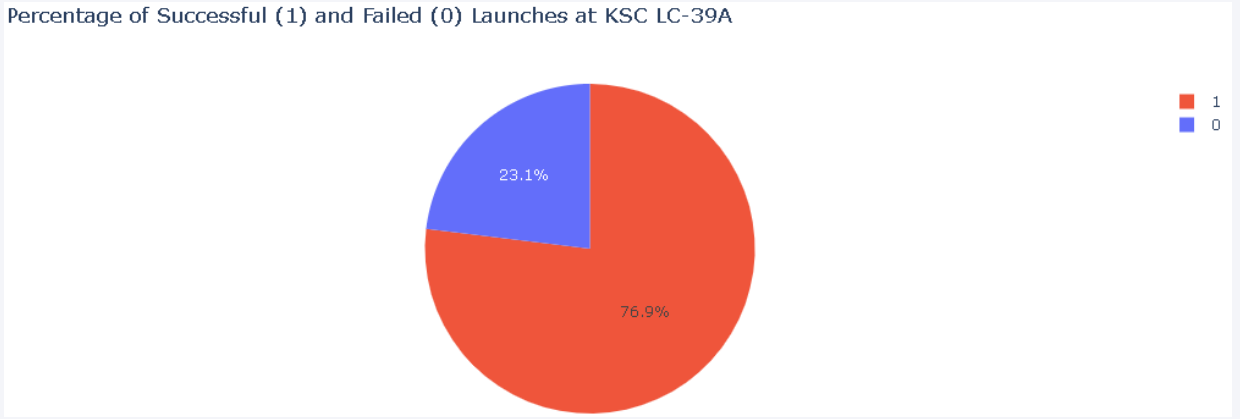
There is a relatively low success rate for launches at Cape Canaveral Launch Complex 40 - only 26.9% of launches succeed from that location. Is is likely that, if the next launch were to occur at this location, that it will not be successful.

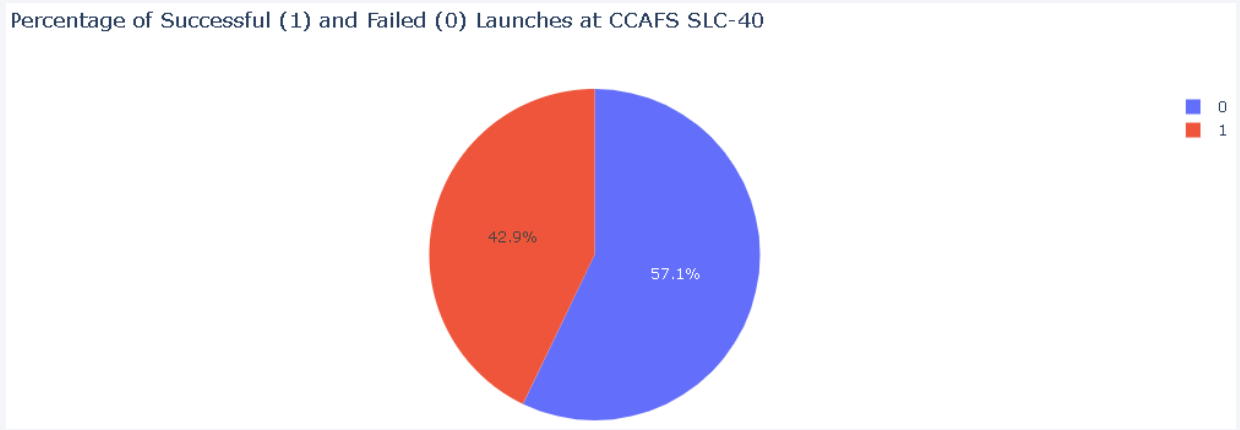
There is also a dependency on payload mass – as payload mass increases, the rate of success decreases.

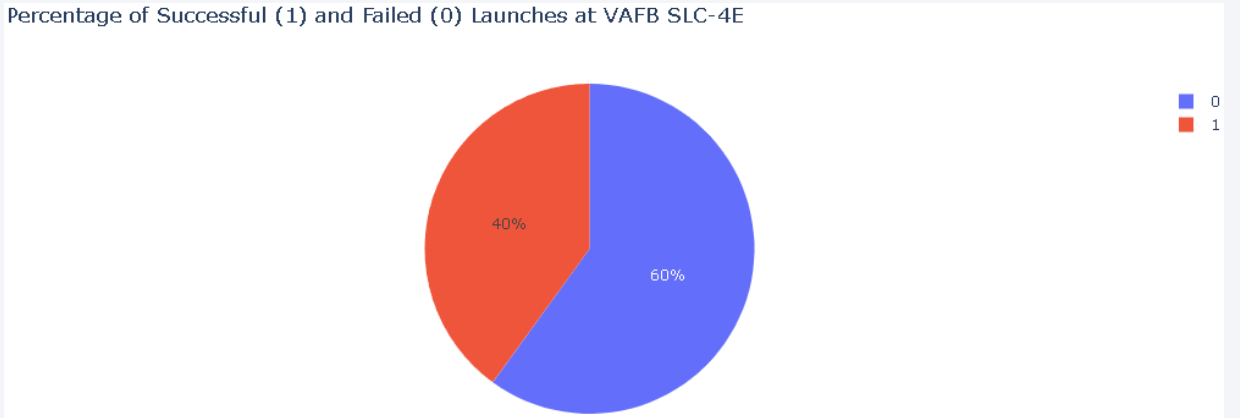
Four models were created, tuned, and tested to predict whether the next SpaceX launch of the Falcon 9 will be successful. All four models, based on the given parameters, are able to predict this outcome with 83.33% accuracy.

Appendix


Percentage of Successful (1) and Failed (0) Launches at CCAFS LC-40
26.9%
73.1%
0
1


Percentage of Successful (1) and Failed (0) Launches at KSC LC-39A
23.1%
76.9%
1
0


Percentage of Successful (1) and Failed (0) Launches at CCAFS SLC-40
42.9%
57.1%
0
1


Percentage of Successful (1) and Failed (0) Launches at VAFB SLC-4E
40%
60%
0
1

Thank you!