IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammed Fayan Ahmed

06/09/2023

Outline

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory data analysis with results
  - Creating graphs, Dash's and interactive Folium Maps.

Introduction

- Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. We want to use Data analysis and Machine Learning to predict if the first stage will land successfully.
- We want to find out how various conditions will affect the landing of stage one and which conditions are optimal for landing.

Section 1

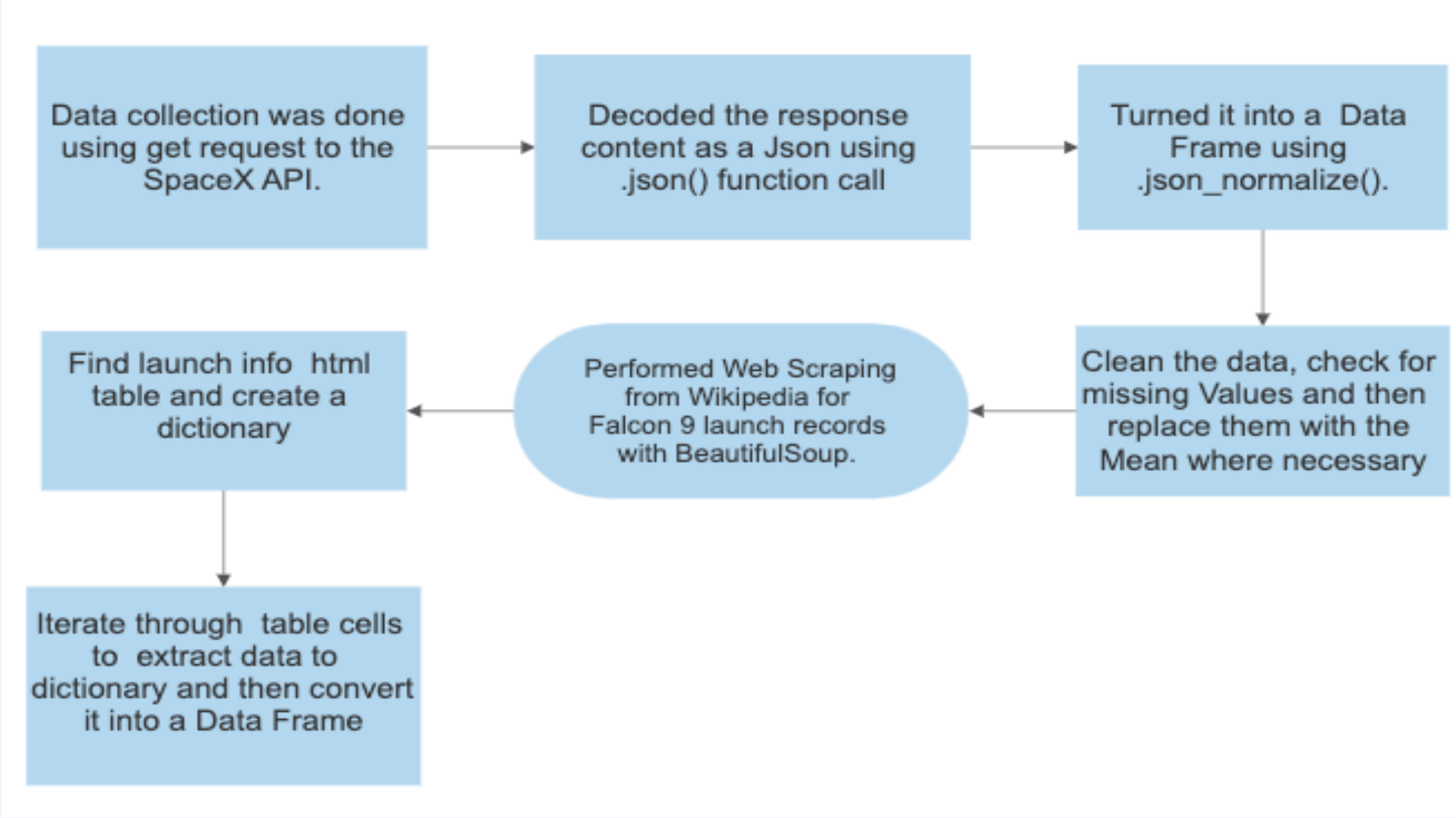
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry
- Perform data wrangling
  - One-hot encoding was applied to 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 : Flow chart for SpaceXAPI and Web Scraping



Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/W1L1-Collecting%20the%20data.ipynb

Data Collection - Scraping

- We applied Web Wcrapping to scrape Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a Pandas Data Frame.

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/Web%20scrapping.ipynb

Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and how many times each orbit was entered.
- We created landing outcome labels from the outcome column and exported the results to a csv.

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/data%20wrangling.ipynb

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

The plots that were used are:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend, Scatter plots, line charts, and bar plots.

These were used to compare relationships between variables to decide if relationships exists so that they could be used in training the machine learning model

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/eda%20Data%20Viz.ipynb

EDA with SQL

We applied EDA with SQL to get insight from the data. We wrote queries to find:

- ❑ The names of unique launch sites in the space mission.
- ❑ The total payload mass carried by boosters launched by NASA (CRS)
- ❑ The average payload mass carried by booster version F9 v1.1
- ❑ The total number of successful and failure mission outcomes
- ❑ The failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/EDA%20SQL.ipynb

Build an Interactive Map with Folium

We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map. Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate. This allows us to understand why launch sites may be located where they are. It also helps to visualizes successful landings relative to their location and surroundings.

https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/folium.ipynb

Build a Dashboard with Plotly Dash

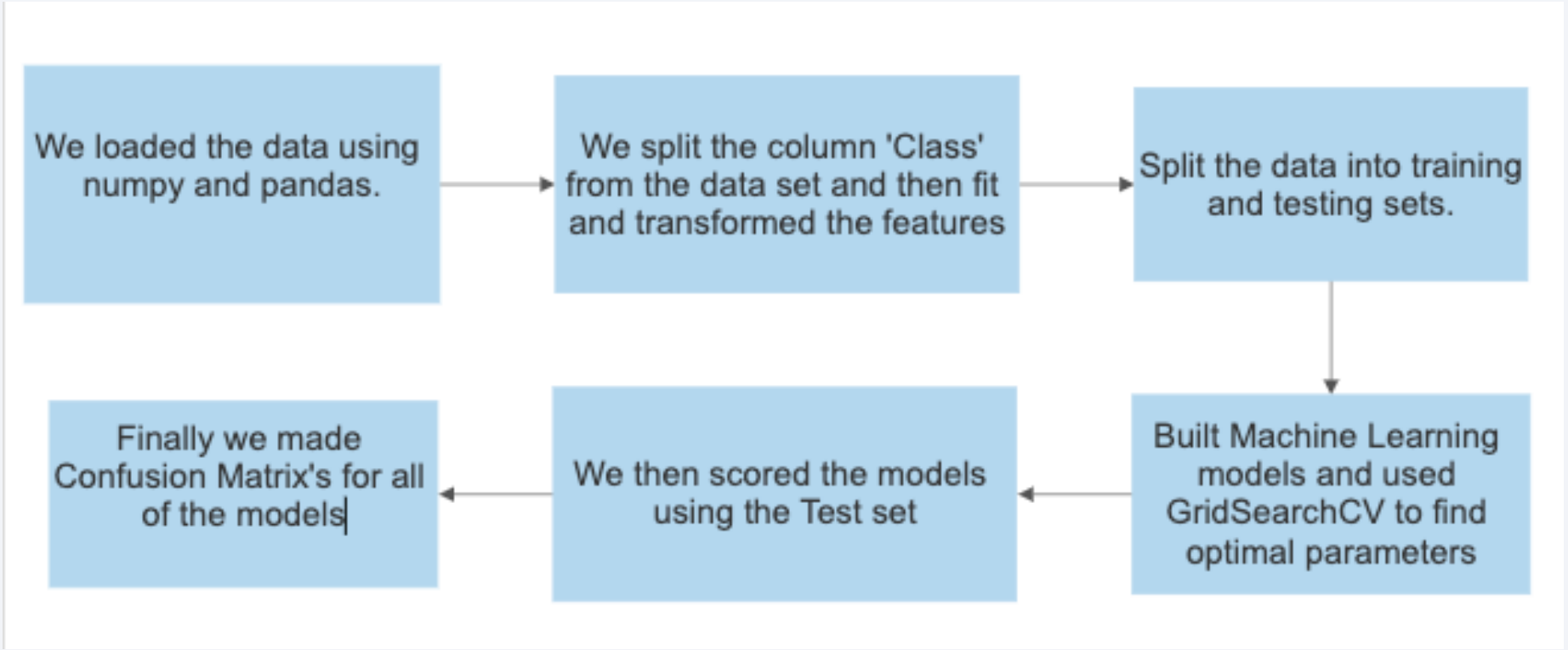
We built an interactive dashboard with Plotly dash.

We plotted a pie chart to show the successful and unsuccessful launches by a certain sites.

We plotted a scatter graph to help us see how success varies across launch sites, payload mass, and booster version category.

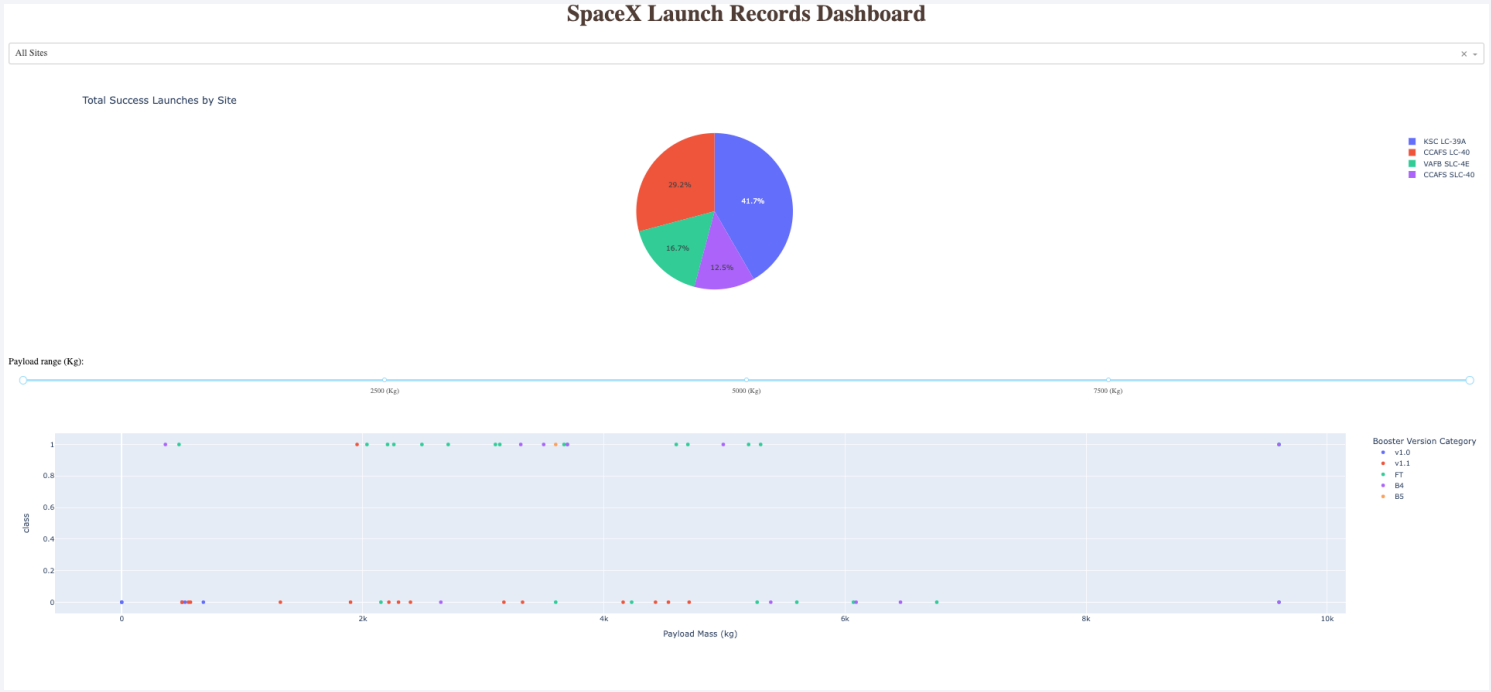
https://github.com/rak1-7/IBM-Data-Science-Final-Project-/blob/b846eb92c20dd678b349e28f7bd2d29e12b929cd/spacex_dash_app%20Main.py

Predictive Analysis (Classification)

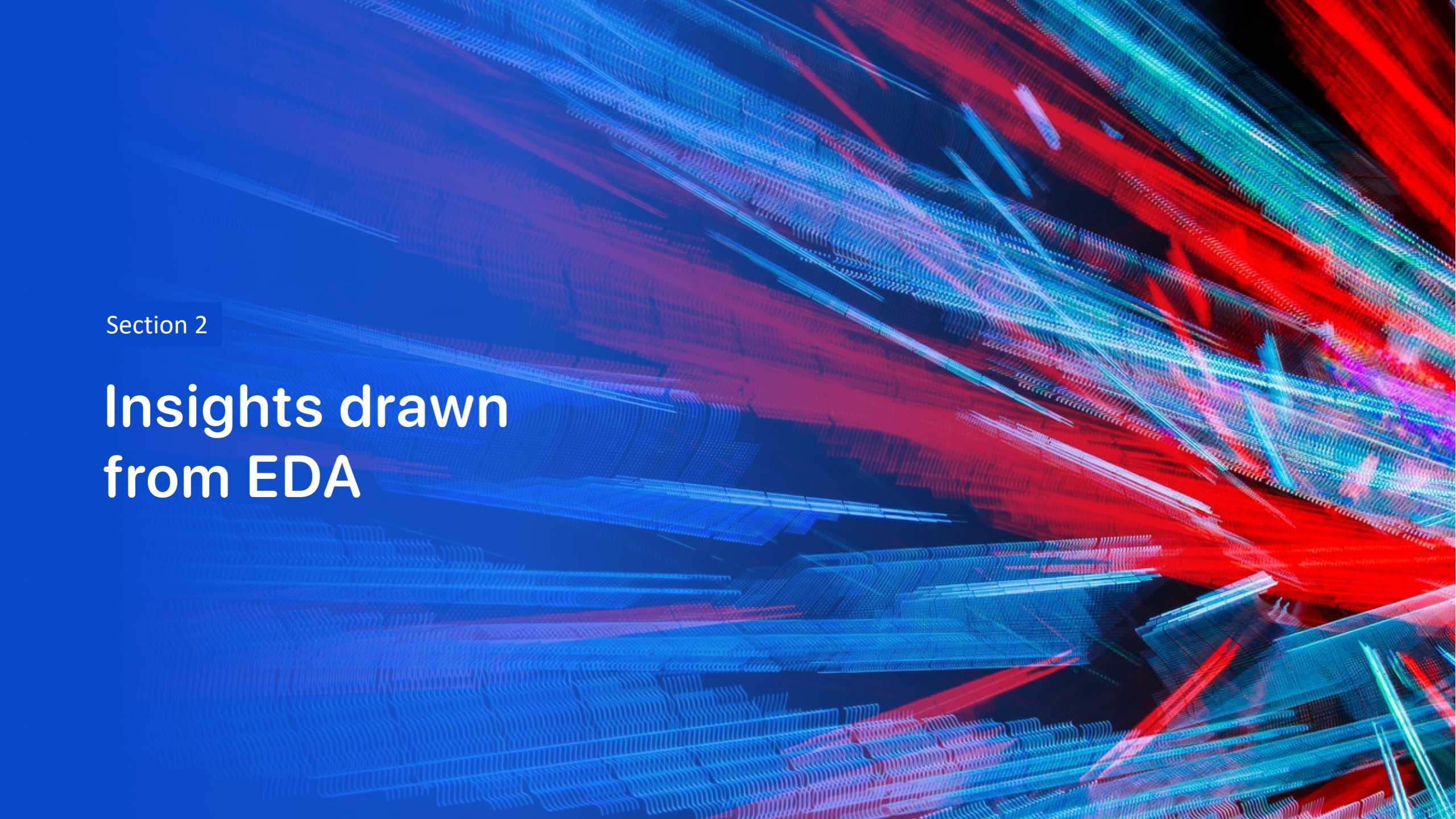


- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

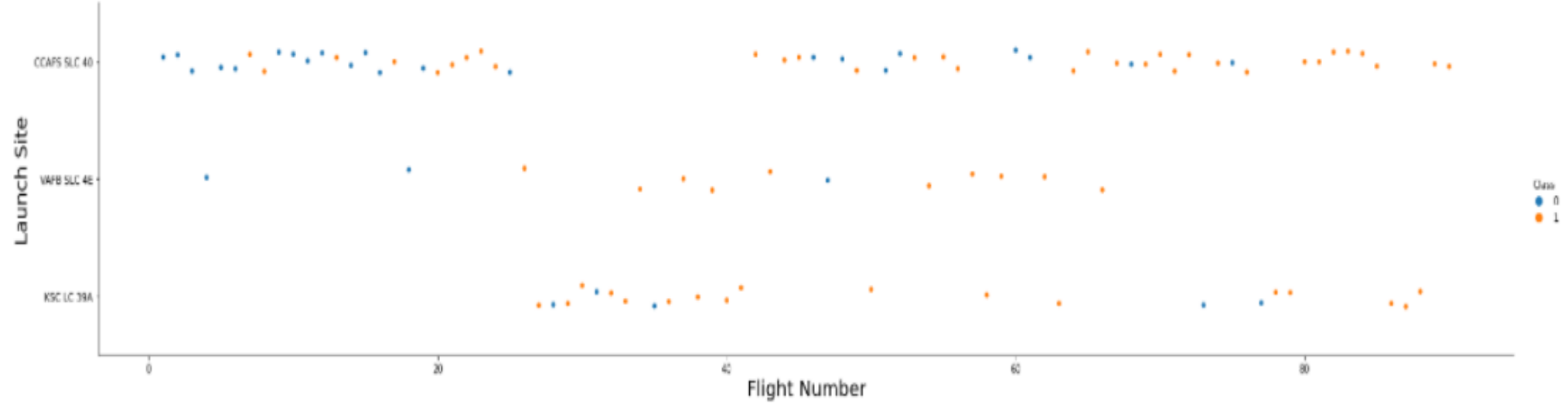
Results



Above is the Dash we made. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

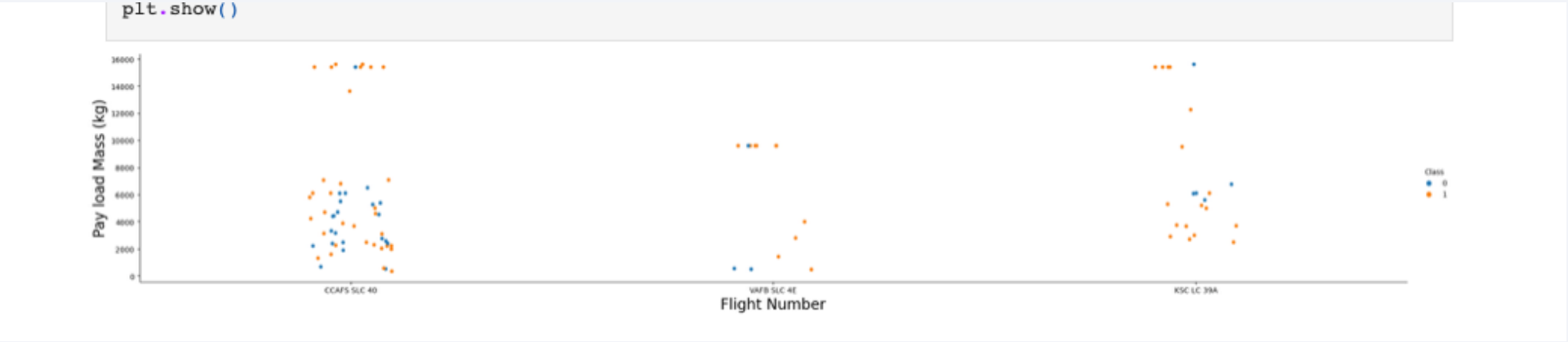

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



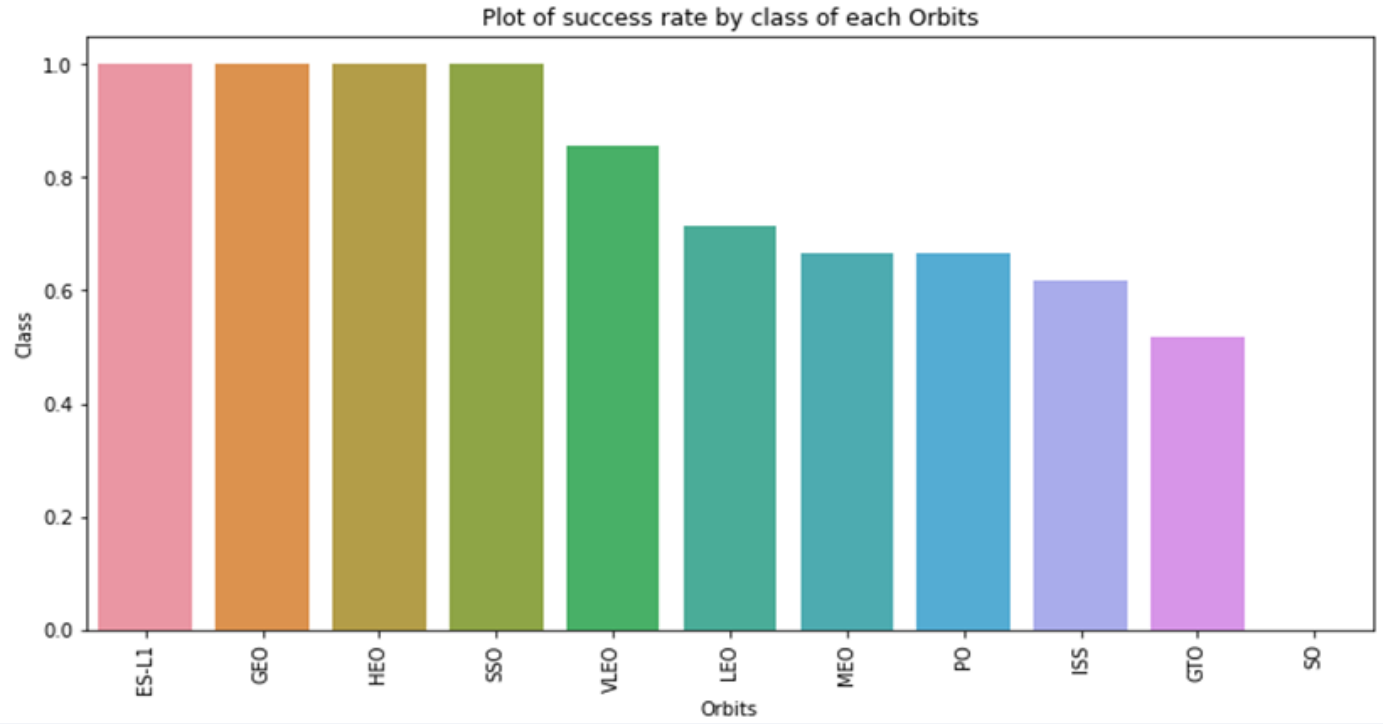
From the plot, we found flight number and launch site do not have a correlation.

Payload vs. Launch Site



Payload mass appears to fall mostly between 0-6000 kg.
Different launch sites also seem to use different payload mass

Success Rate vs. Orbit Type



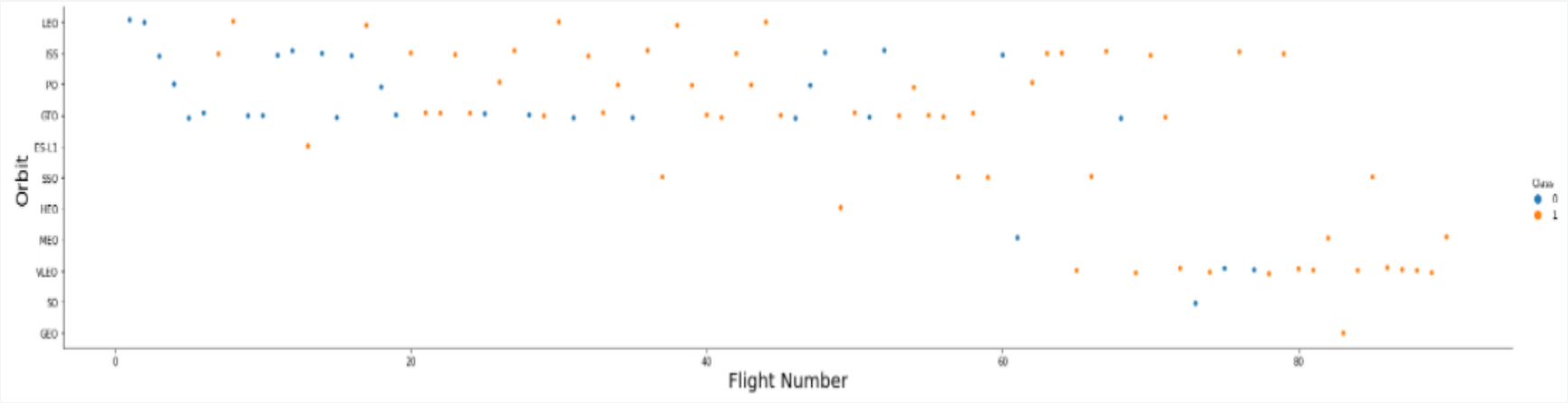
ES-L1 (1), GEO (1), HEO (1) , SSO (5) have 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type



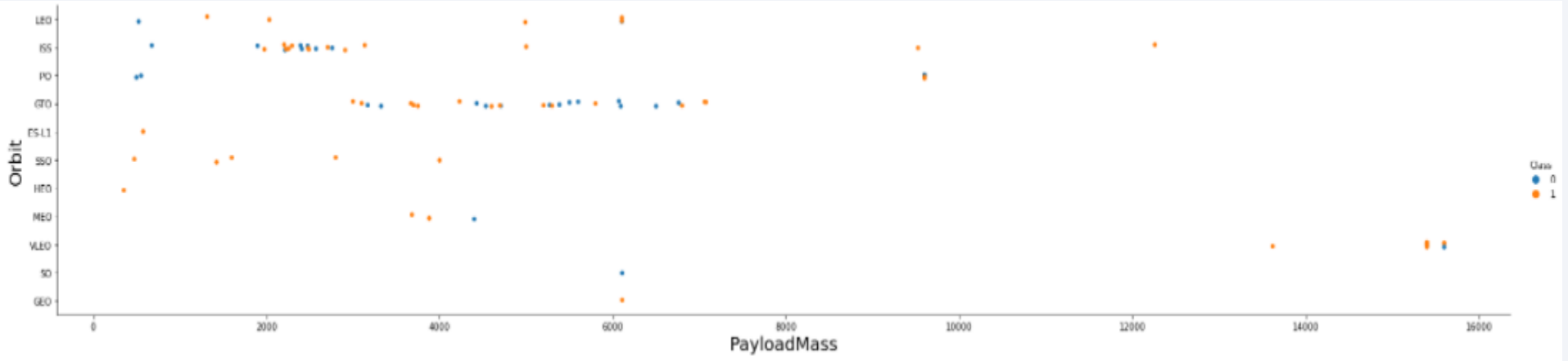
Launch Orbit preferences changed over Flight Number. Launch

Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

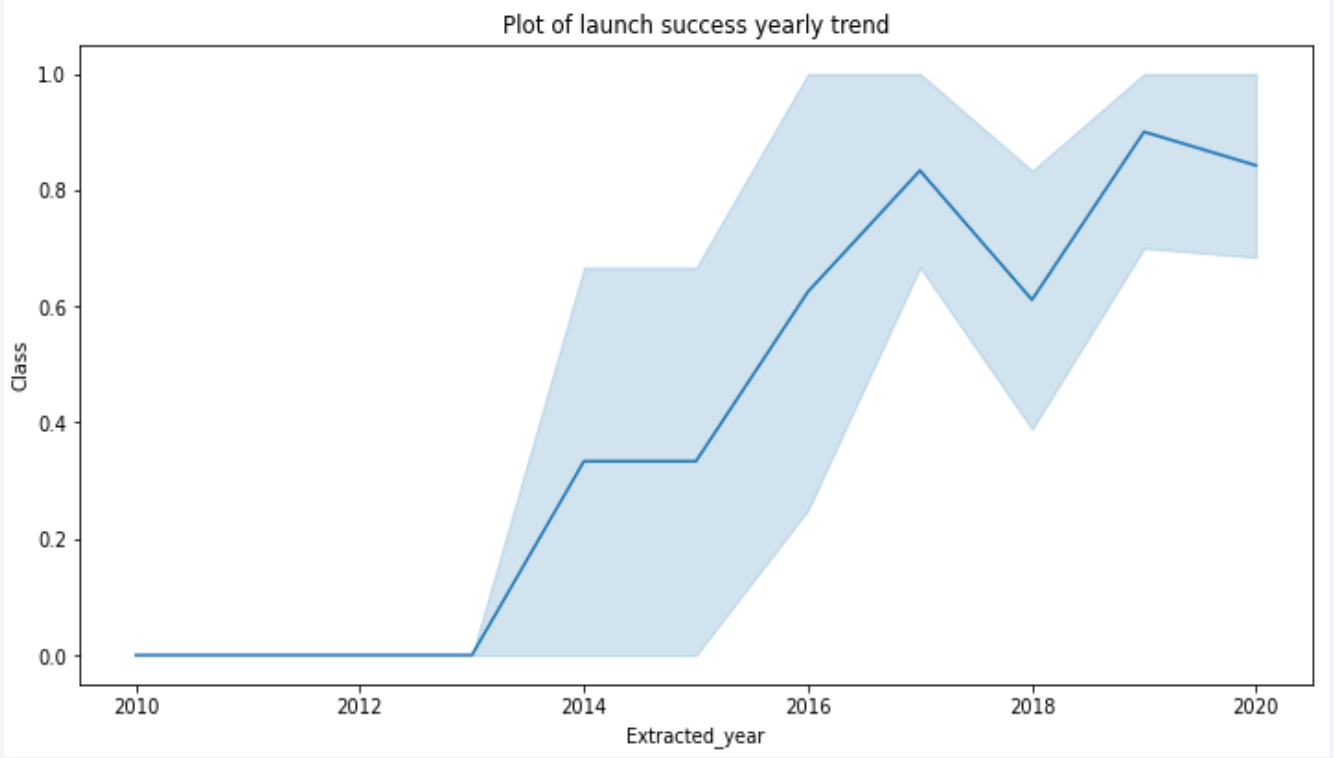
Payload vs. Orbit Type



We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%



All Launch Site Names

We used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

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

In [9]:
```
%sql SELECT distinct LAUNCH_SITE from SPACEXTABLE
```

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

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

In [10]:
```
%sql select * from SPACEXTABLE where LAUNCH_SITE LIKE 'CCA%' limit 5
```

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

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (
2010-08-12	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 (
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

We used the query above to display 5 records where launch sites begin with `CCA`

Total Payload Mass

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

In [13]:
```
%sql select sum(PAYLOAD_MASS__KG_) AS TOTAL from SPACEXTABLE where customer like 'NASA (CRS)'
```

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

Out[13]:

TOTAL
45596

Average Payload Mass by F9 v1.1

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

```
In [17]: %sql select AVG(PAYLOAD_MASS__KG_) AS AVG from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[17]:
```

AVG
2534.6666666666665

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

First Successful Ground Landing Date

In [14]:

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

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

In [15]:

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

Out[15]:

	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

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

In [41]:
```
%sql select DATE AS Month, Landing_Outcome, booster_version, launch_site from SPACEXTABLE where DATE like '2015%'
```

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

Out[41]:

Month	Landing_Outcome	Booster_Version	Launch_Site
2015-10-01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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.

In [44]:
```
%sql select Landing_Outcome, count(*) as total from SPACEXTABLE where date <= '2017-03-20' and date>=' 2010-06-04
```

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

Out[44]:

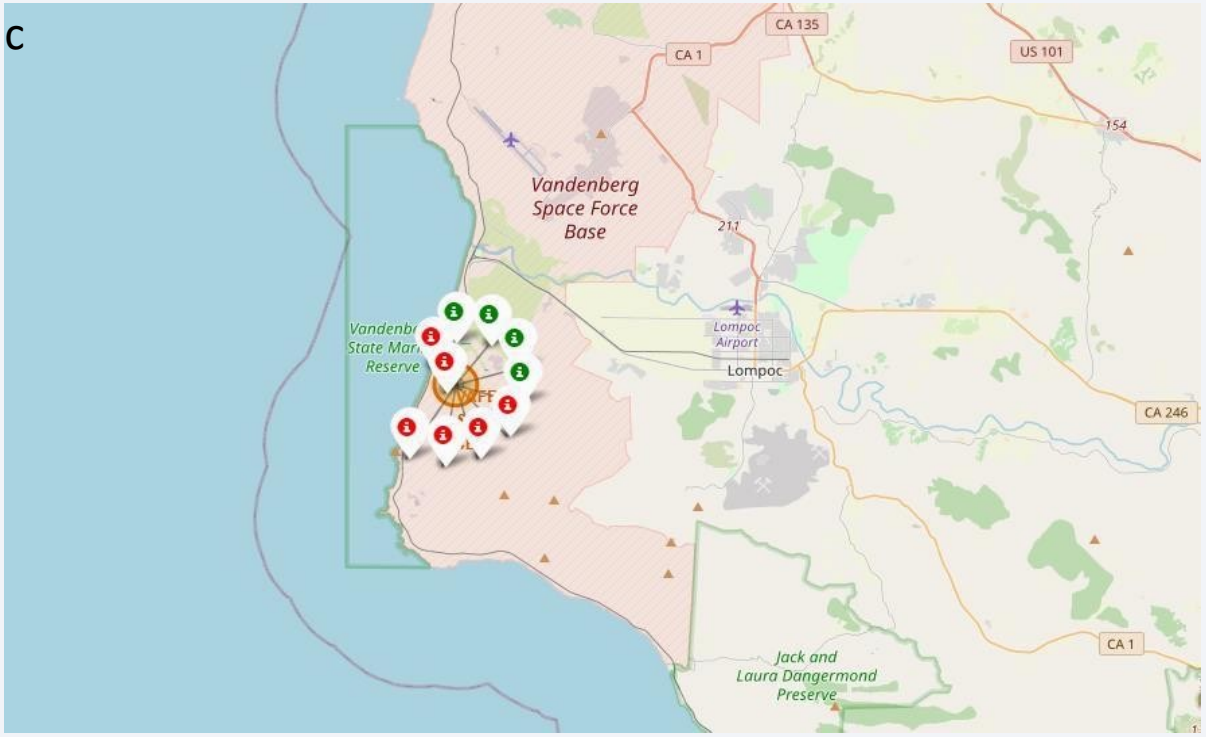
Landing_Outcome	total
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

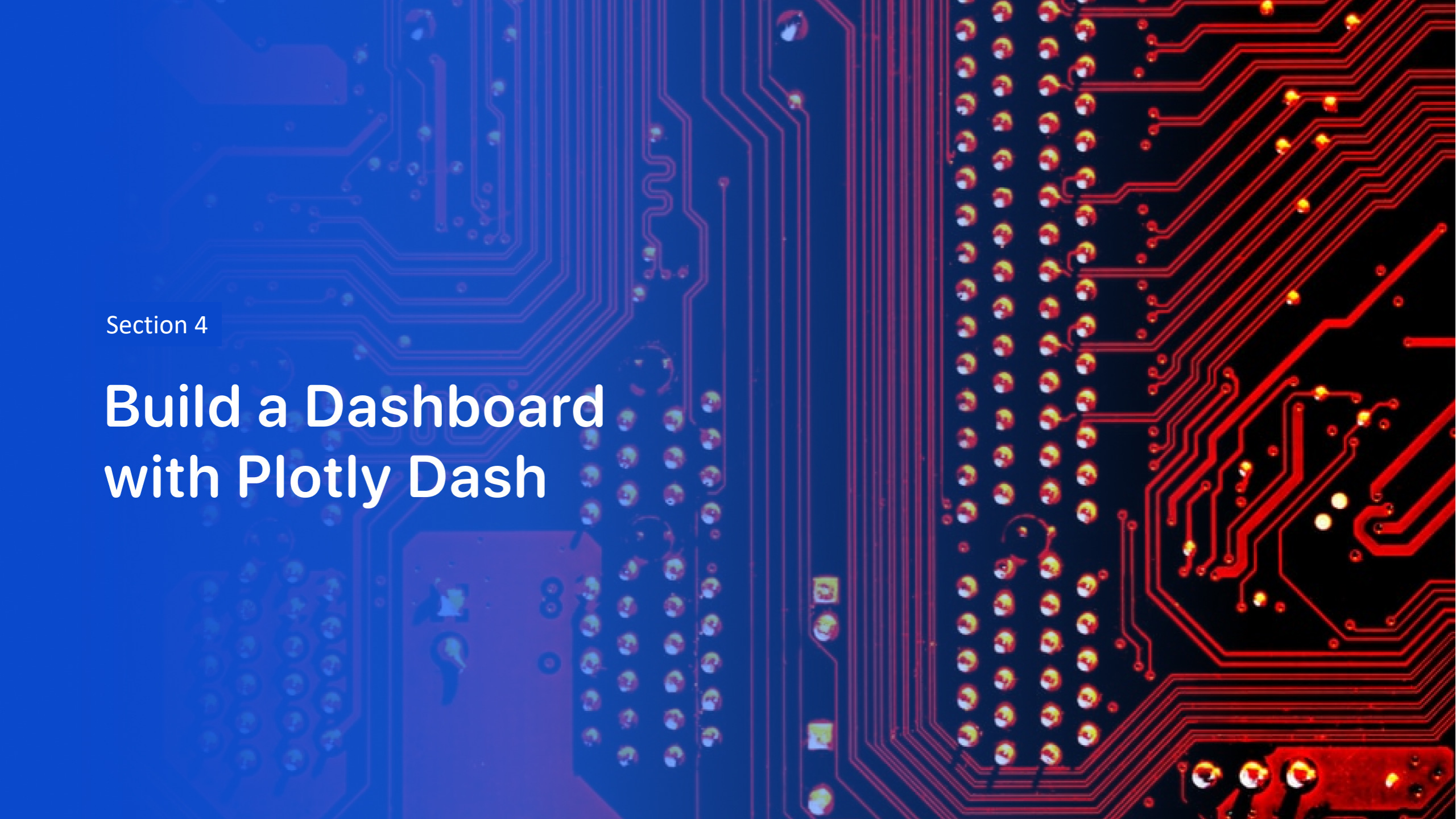
We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Color-Coded Launch Markers

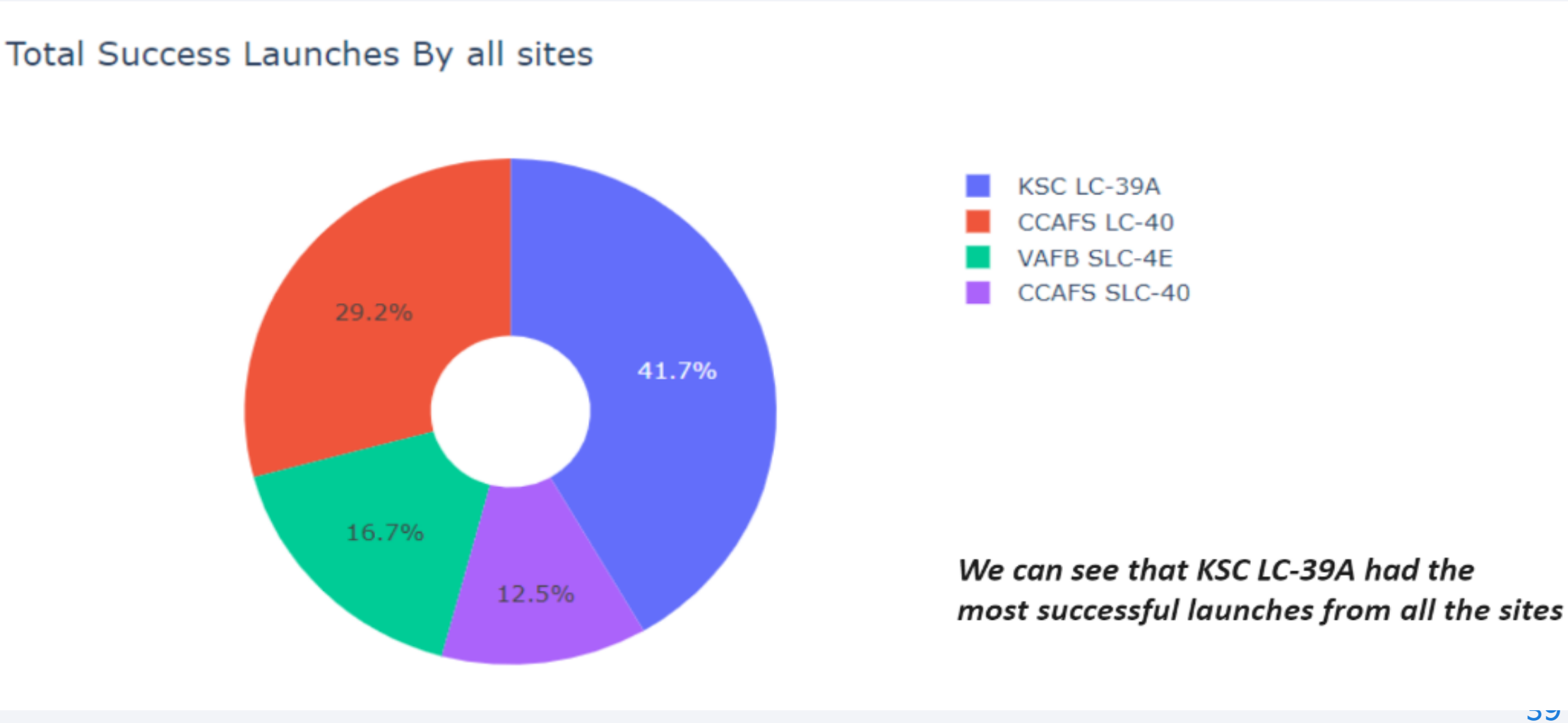


Green markers show successful launches and red markers show failures.

<Folium Map Screenshot 3>

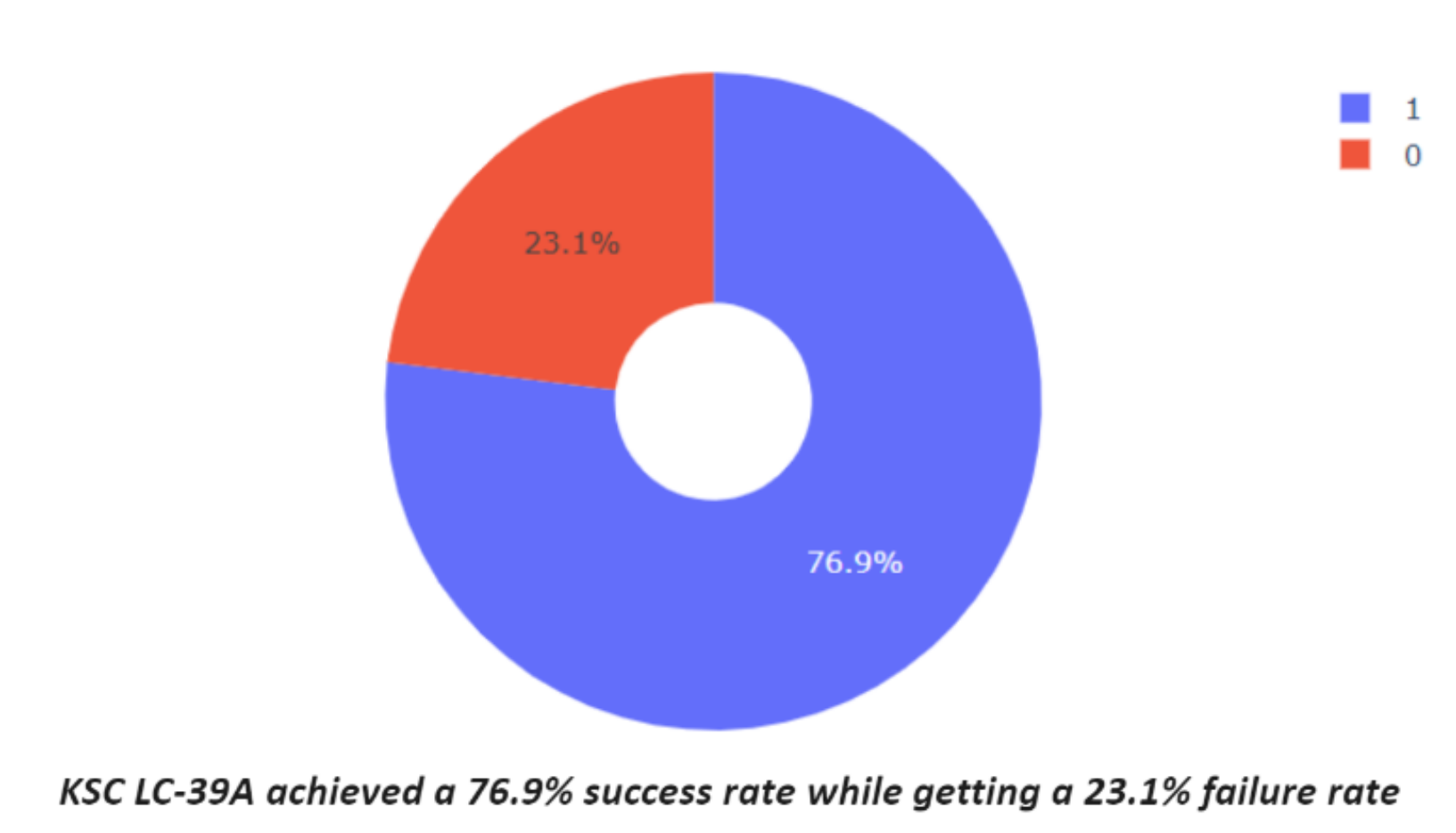

Section 4
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



We can see that KSC LC-39A had the most successful launches from all the sites

Highest Success Rate Launch Site

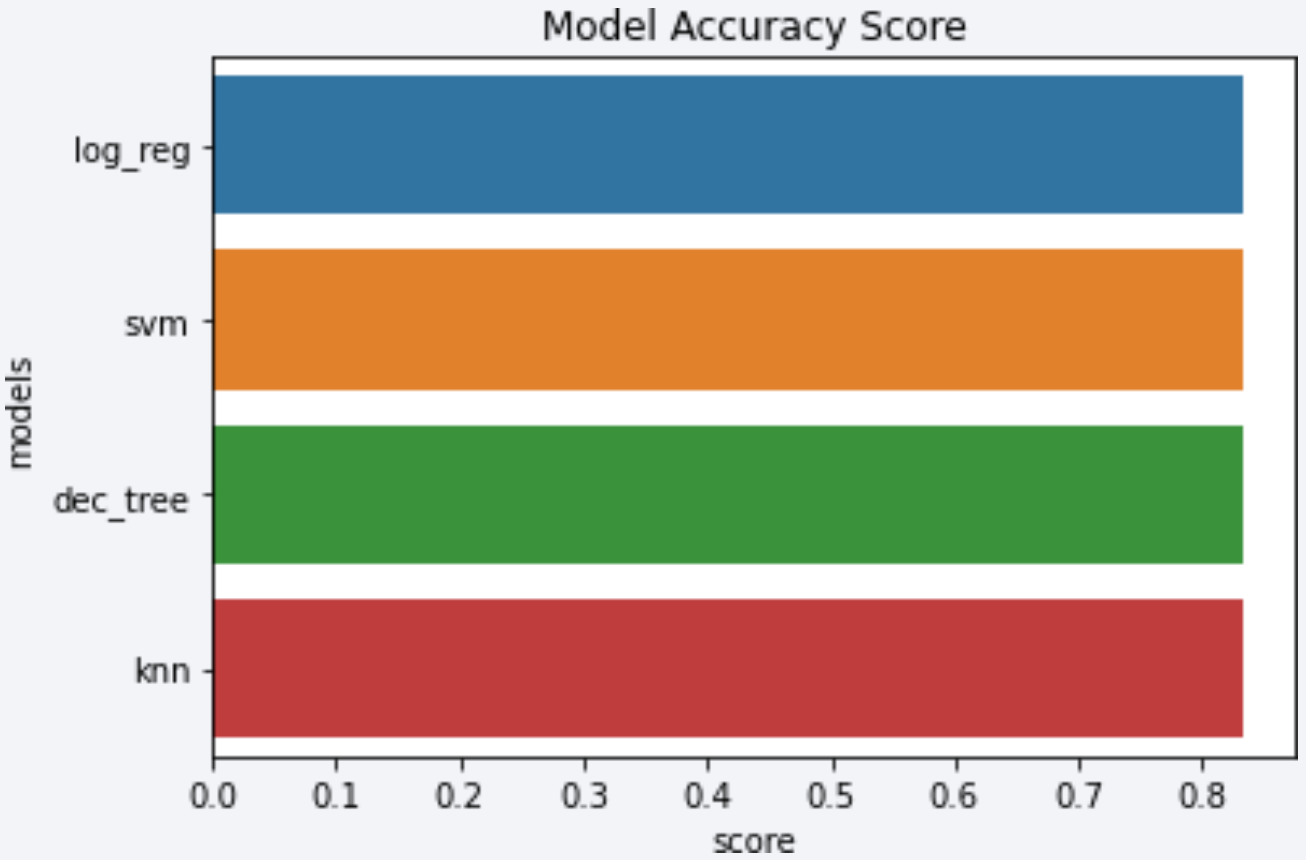


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Section 5

Predictive Analysis (Classification)

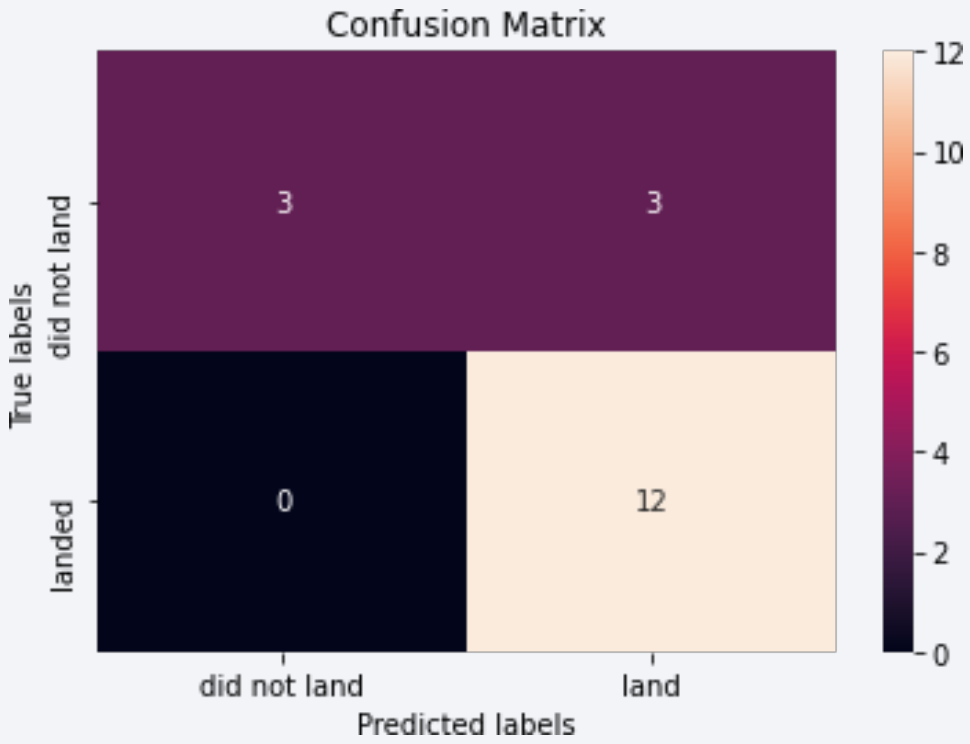
Classification Accuracy



All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We conclude:

- Launch success rate started to increase in 2013 until 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!