IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mateusz Bejger
6th May 2023

Outline

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

Executive Summary

- Part 1: Scrape / acquire data – utilize API / web-scraping to collect data in preparation to cleanse it. This section highlights the first part of data preparation.
- Part 2: Data wrangling – clean the data. During this stage, the data is wrangled and any errors or unwanted values are treated accordingly. This is to be done before next stage is initialized.
- Part 3: Visualize – showcase the data in a visual format to help make sense of results. Visual usage is both for the benefit of data exploration/understand.
- Part 4: Machine Learning – build a machine learning model to predict the findings of the provided question. This model will use computing to predict outcomes based of existing data. This model will be trained on a training set and performance checked.

Introduction

- Space X is creating a more cost-efficient alternative to space travel through utilizing, self-landing, reusable thrusters. Through predicting if the first stage will land successfully, the information can be used for companies to bid for the launches.
- In order to determine if the launch will include a successful landing from the first stage, a calculated estimation will be performed to predict the outcome.

Section 1

Methodology

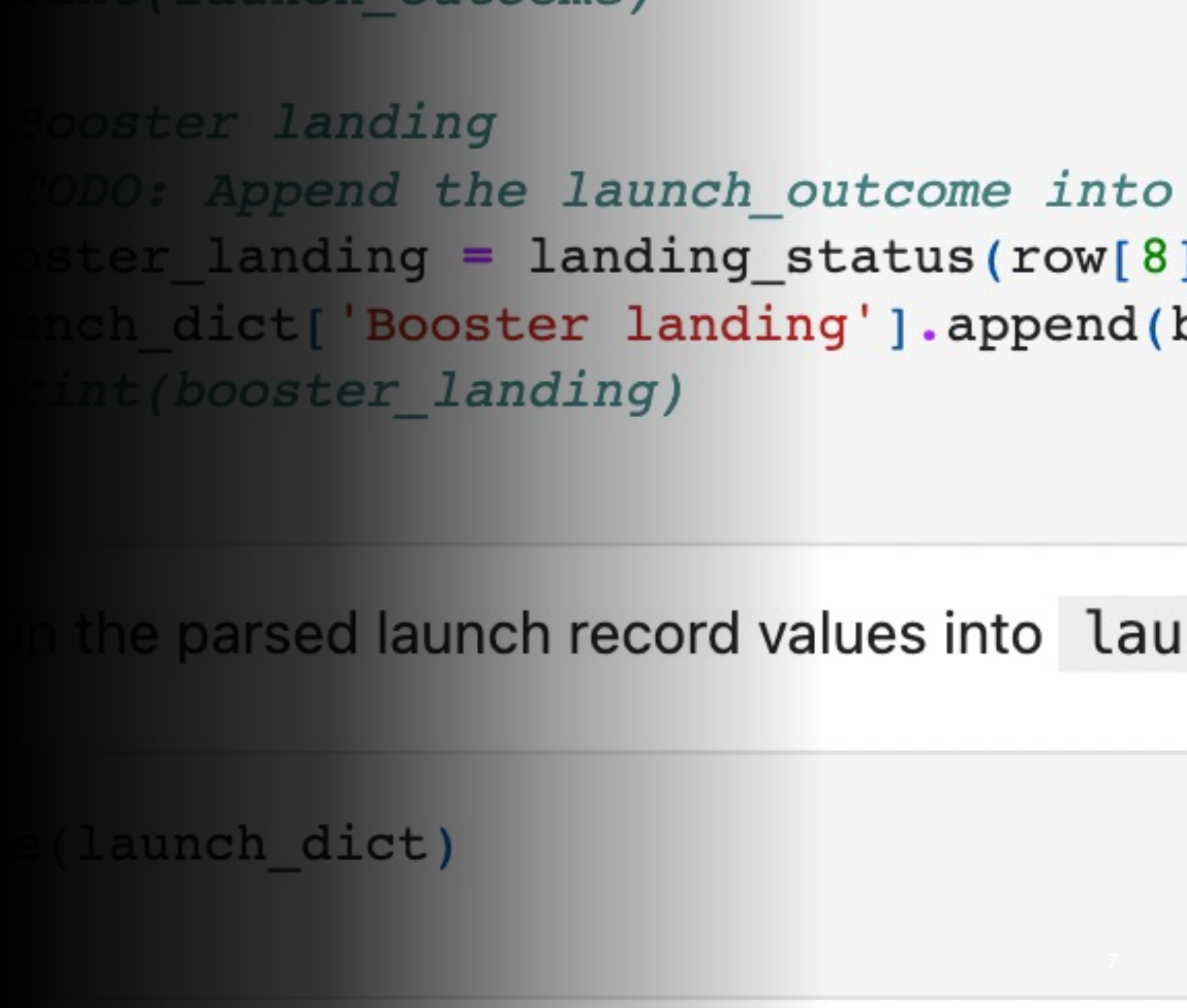
Methodology

Executive Summary

- Data collection methodology:
  - Collecting data through an API and web-scraping official tables
- Perform data wrangling
  - Data was cleansed with python library assisted functions / SQL querying
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build a machine learning model and validate accuracy

Data Collection

- API data collection through a get request.
- Web-scraping official table information for launches with BeautifulSoup, Python library.
- JSON normalization to proceed in converting the data to a pandas Dataframe.



Data Collection – SpaceX API

- Create a GET request to normalize JSON package, in order to transform the data to a pandas Dataframe
- https://github.com/codebymateus/Capstone_Project/blob/main/1_API.ipynb

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json result into a dataframe
response_1 = response.json()
data = pd.json_normalize(response_1)
```

```
# Lets take a subset of our dataframe keeping only the features we want and the
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_

# We will remove rows with multiple cores because those are falcon rockets with
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single v
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then extractin
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

Data Collection - Scraping

- Create a static URL variable and a response using a GET request. Create a soup object and find all tables. Apply column names and create a loop to iterate through the scraped data.
- https://github.com/codebymateus/Capstone_Project/blob/main/2_WebScraping.ipynb

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Fal
```

```
response = requests.get(static_url)
response.status_code
```

```
# Use BeautifulSoup() to create a BeautifulSoup object from a respor
soup = BeautifulSoup(response.text, "html.parser")
```

```
html_tables = soup.find_all('table')
html_tables
```

```
column_names = []
# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_fron
# Append the Non-empty column name (`if name is not None and len(nar
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if (name != None and len(name) > 0):
        column_names.append(name)
```

```
df=pd.DataFrame(launch_dict)
df
```

Data Wrangling

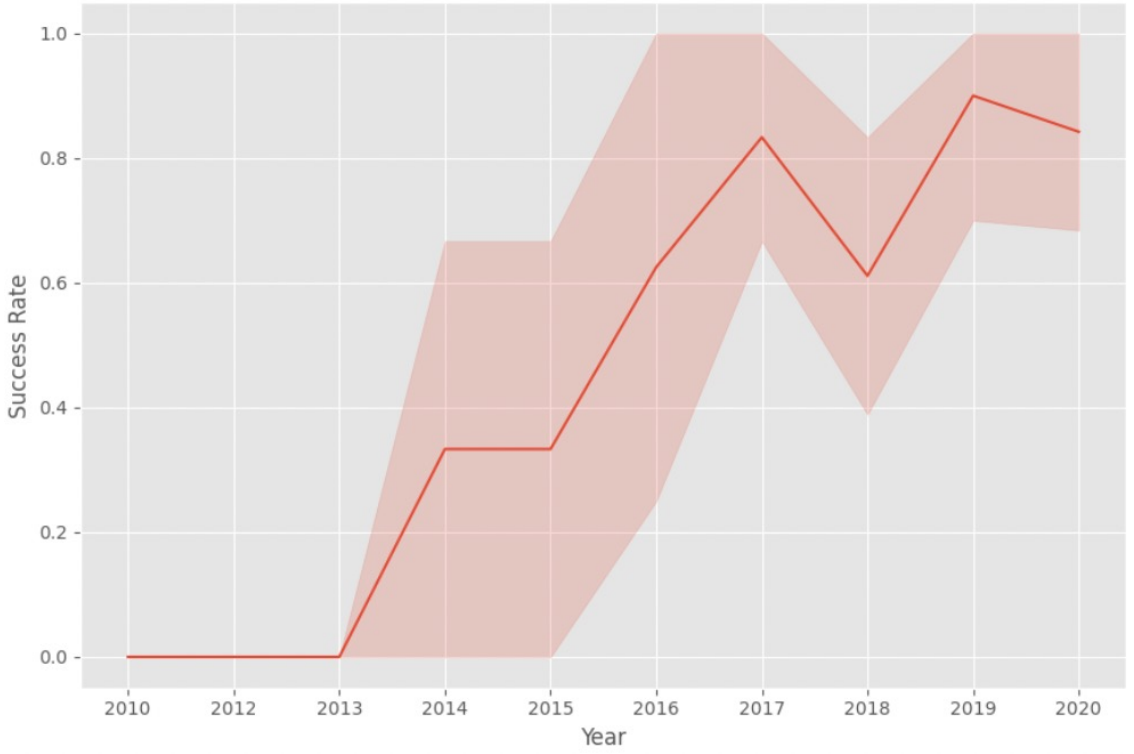
- Cleaning data through pandas filtering and new Dataframe processing within new set variables. Search for null values and replace applicable with the mean for data acquired within the column. Ensure data types for each column are sensible and can be computed accordingly.
- https://github.com/codebymateus/Capstone_Project/blob/main/3_DataWrangling.ipynb

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

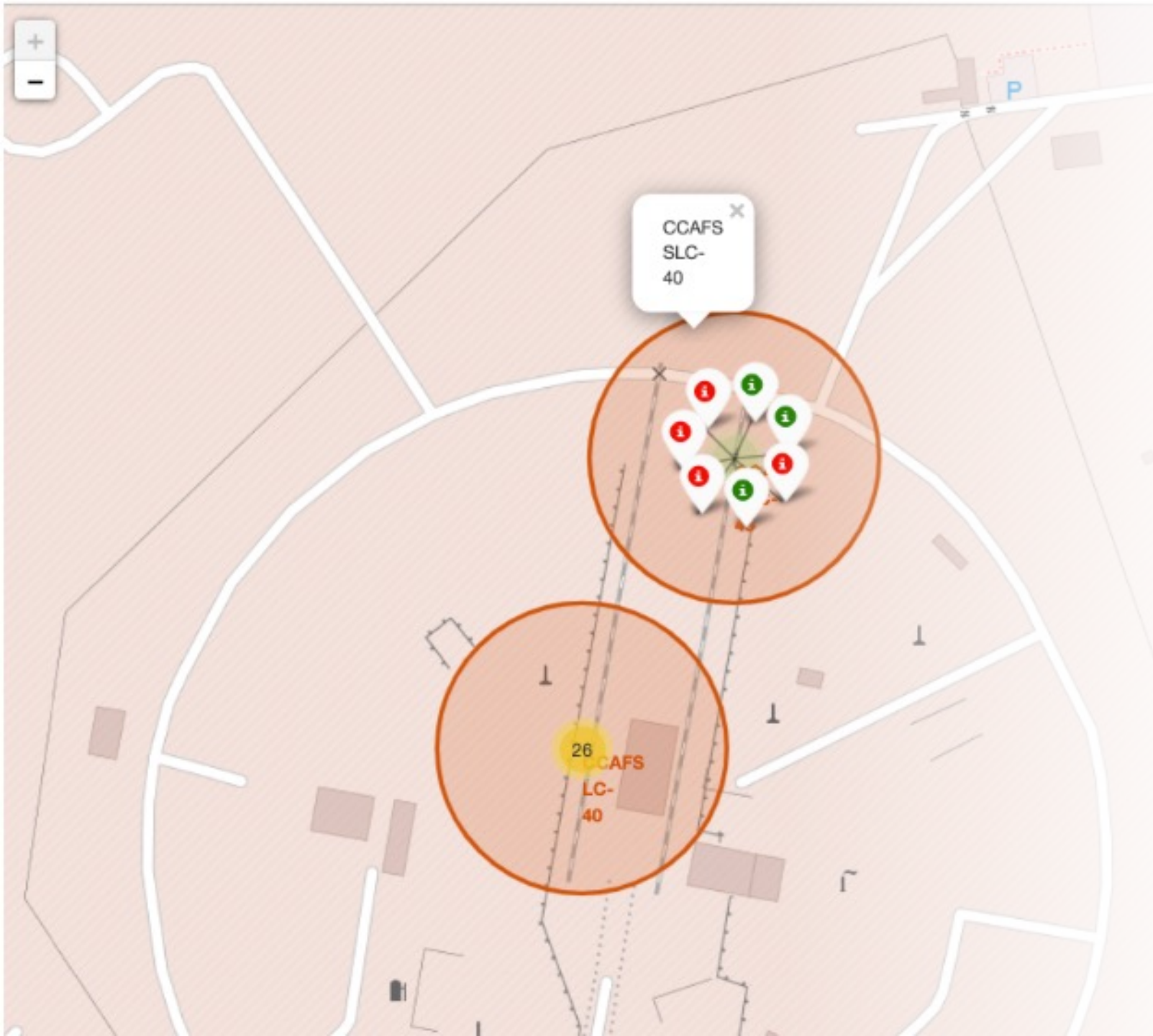
EDA with Data Visualization

- Scatter plots to show relationship between variables
- Bar charts to show the relationship between variables
- Line chart was used to show the trends in success rate over time
- https://github.com/codebymateus/Capstone_Project/blob/main/5_Visualization.ipynb



EDA with SQL

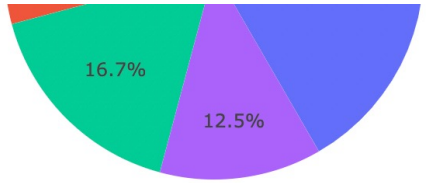
- Find unique values for ‘Launch Sites’
- Find total payload mass for boosters launched by NASA (CRS)
- Find average payload mass for booster version F9 v1.1
- Find date where first successful landing outcome was achieved
- List booster versions that carried the maximum payload mass
- Rank count of successful landing outcomes between specified date range
- https://github.com/codebymateus/Capstone_Project/blob/main/4_SQL.ipynb



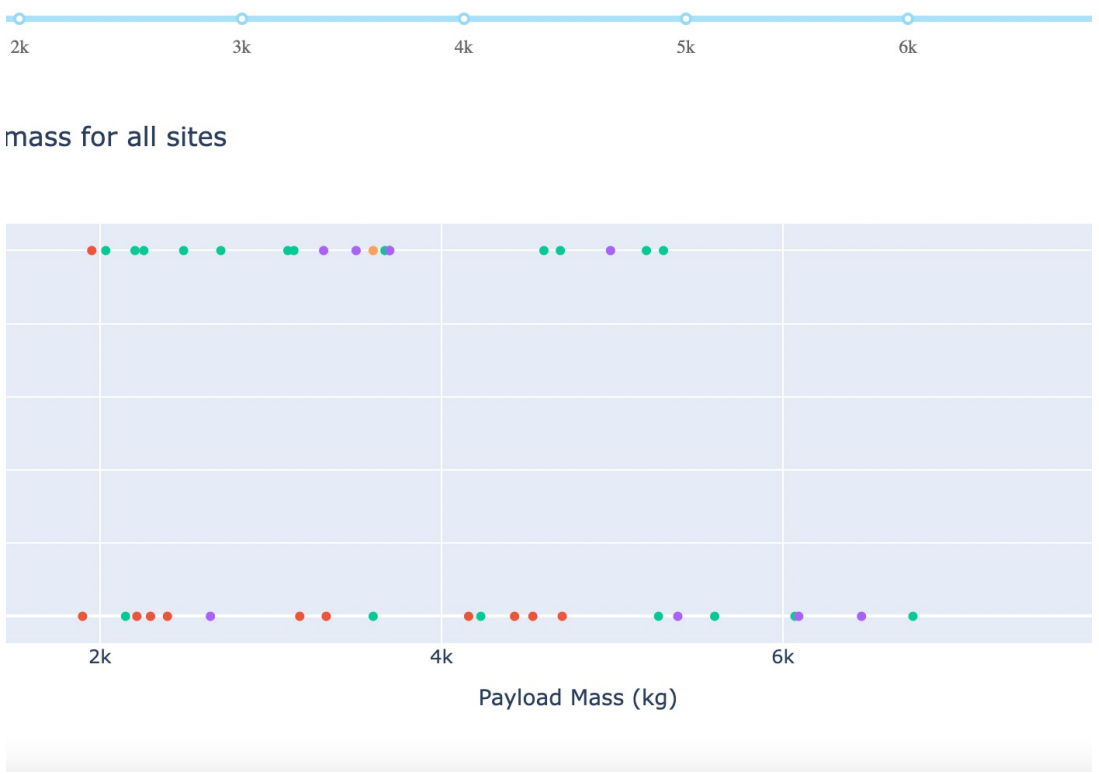
Build an Interactive Map with Folium

- Created folium map markers for launch sites, color coded based off success outcome
- Highlight potential correlation between successful outcomes to location launch
- https://github.com/codebymateus/Capstone_Project/blob/main/6_%20Folium.ipynb

Build a Dashboard with Plotly Dash



- Interactive drop downs for parameter selection. Range slider for date filtering. Scatter plot for class classification. Bar chart for visual insight in variance.
- https://github.com/codebymateus/Capstone_Project/blob/main/7_Dash.py



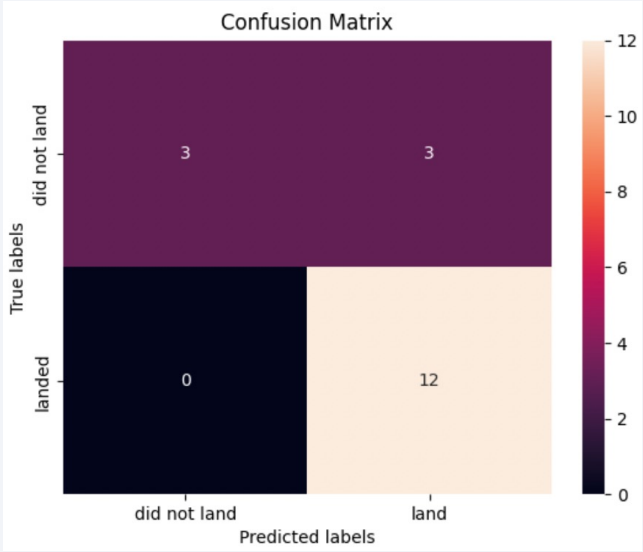
Predictive Analysis (Classification)

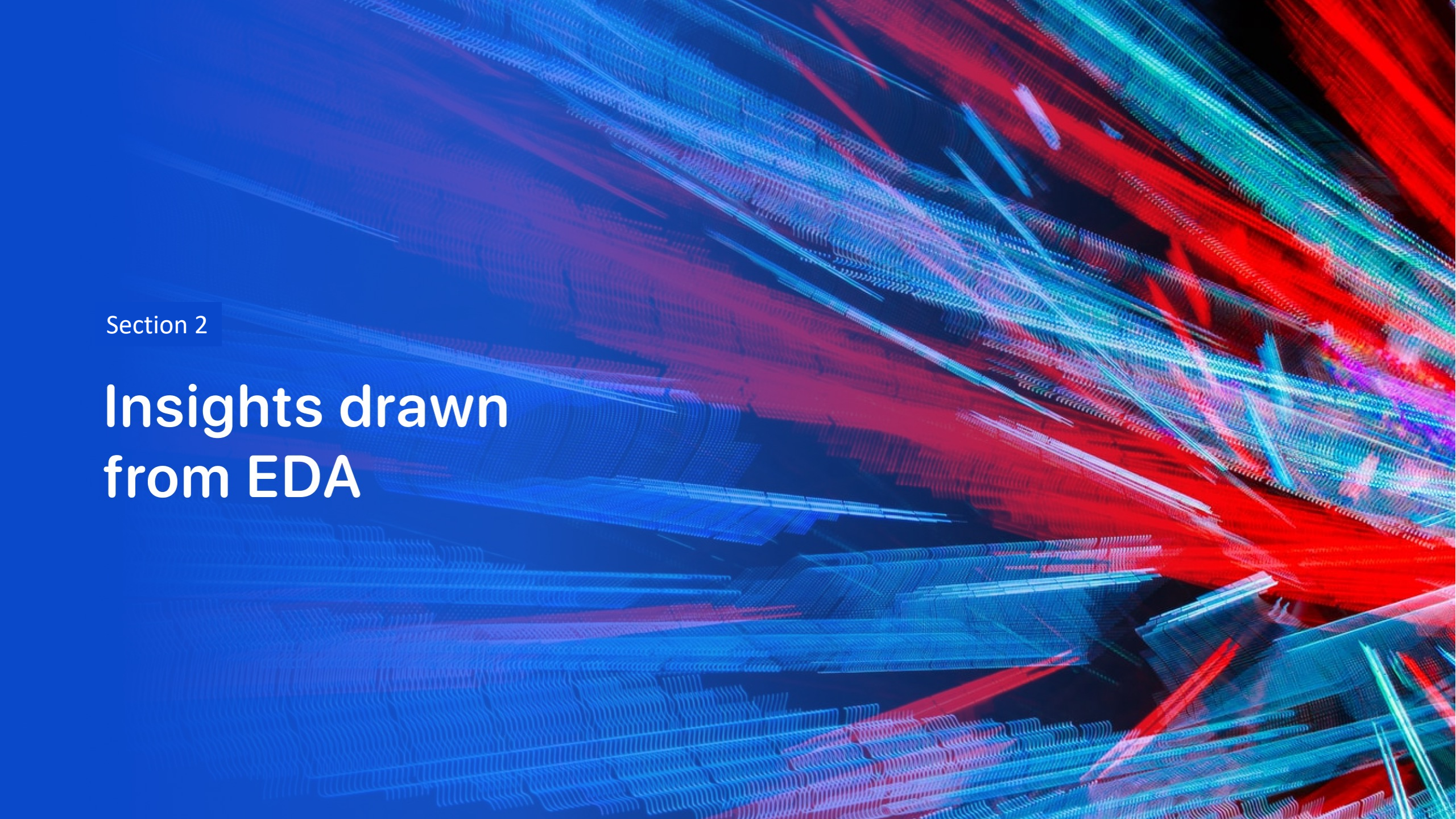
- Import relevant libraries, sklearn, pandas, numpy
- Determine training labels, find outcome variables using numpy
- With the Scaler function, standardize the dataset to split into training/test data sets
- Choose between Support Vector Machine, k nearest neighbor, class trees and logistic regression with the advantage of a confusion matrix and .score function calculation for model accuracy
- https://github.com/codebymateus/Capstone_Project/blob/main/8_ML_Prediction.ipynb

Results

Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

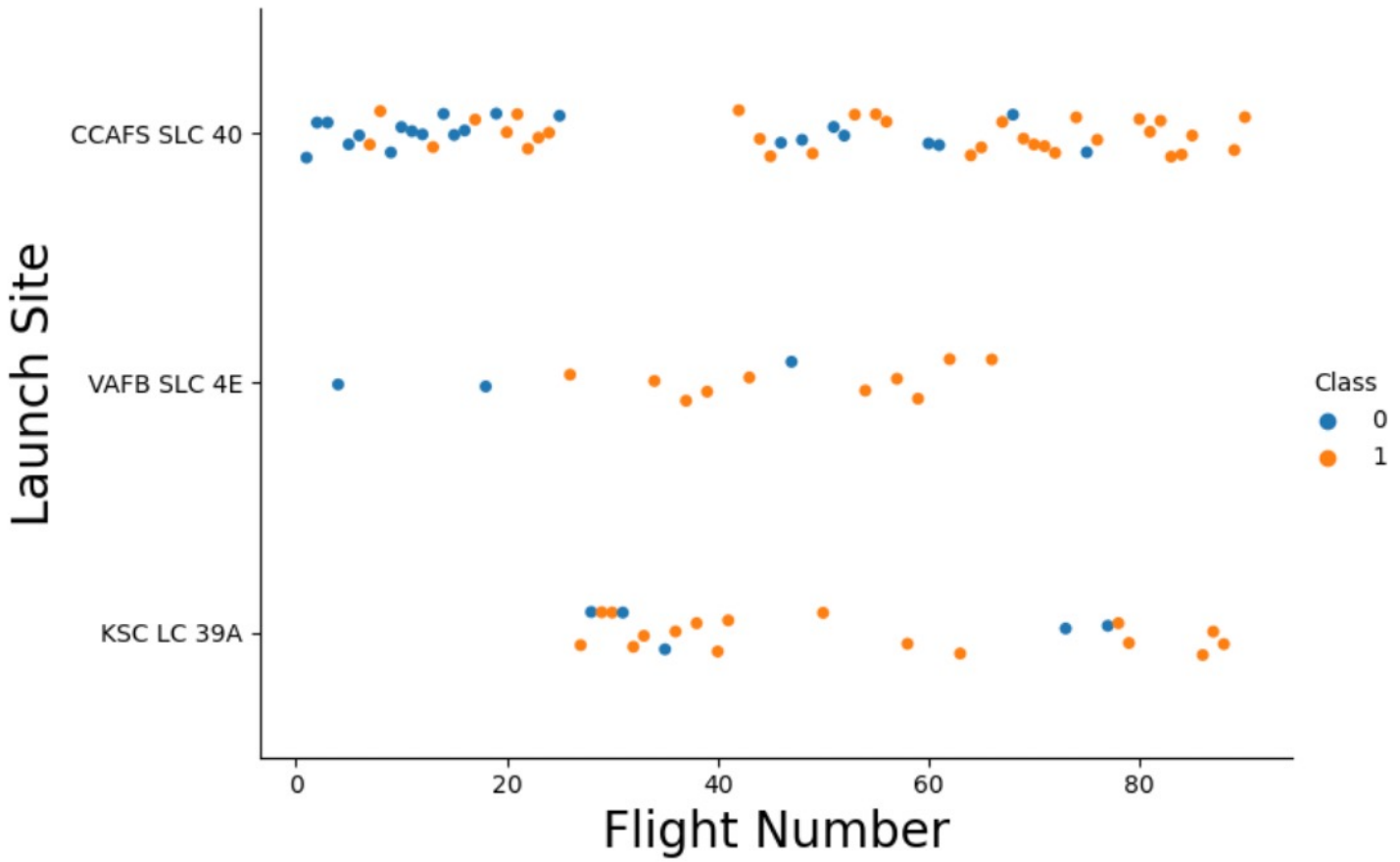
All models performed equally and shall be chosen for use upon convenience / availability




Section 2
Insights drawn from EDA

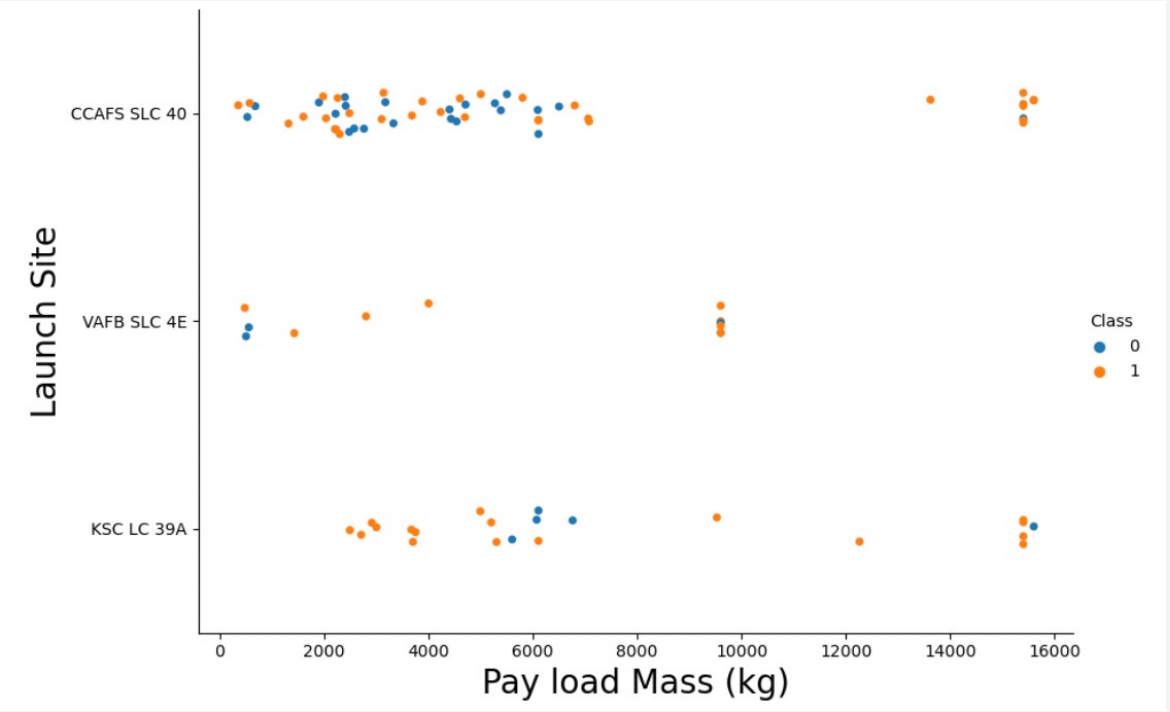
Flight Number vs. Launch Site

- We can deduce that, as the flight number increases in each of the 3 launcg sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight.



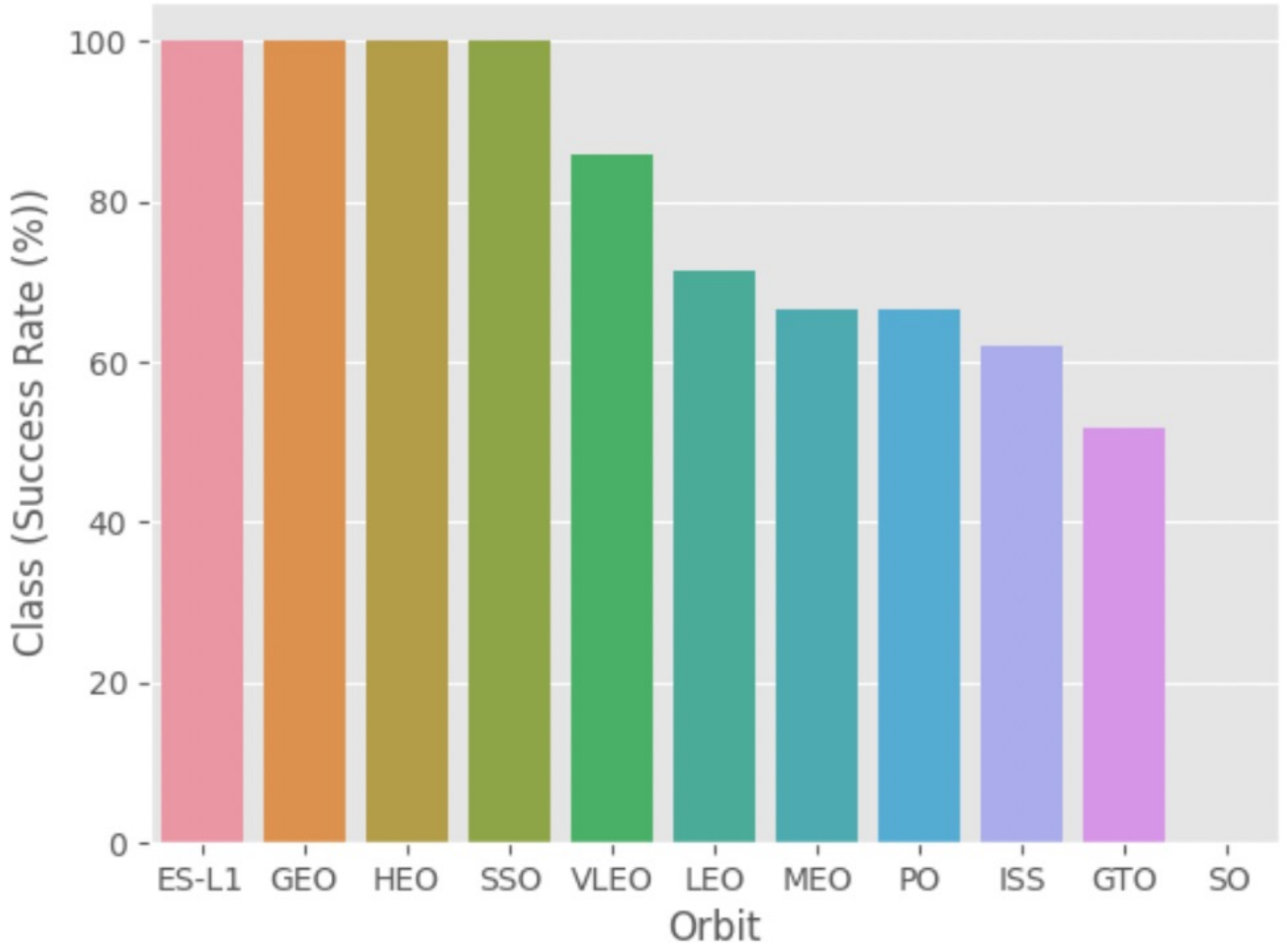
Payload vs. Launch Site

- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



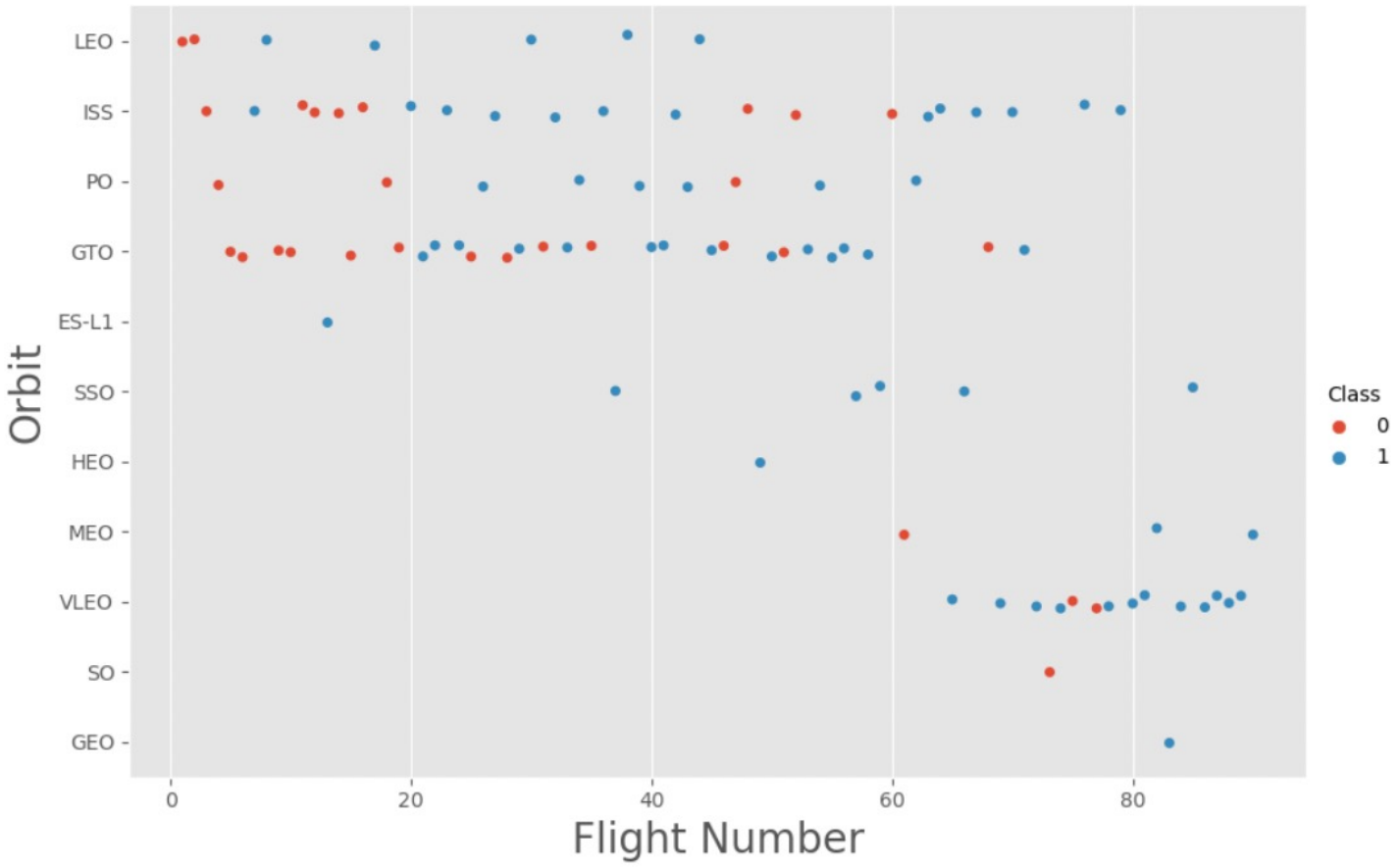
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.



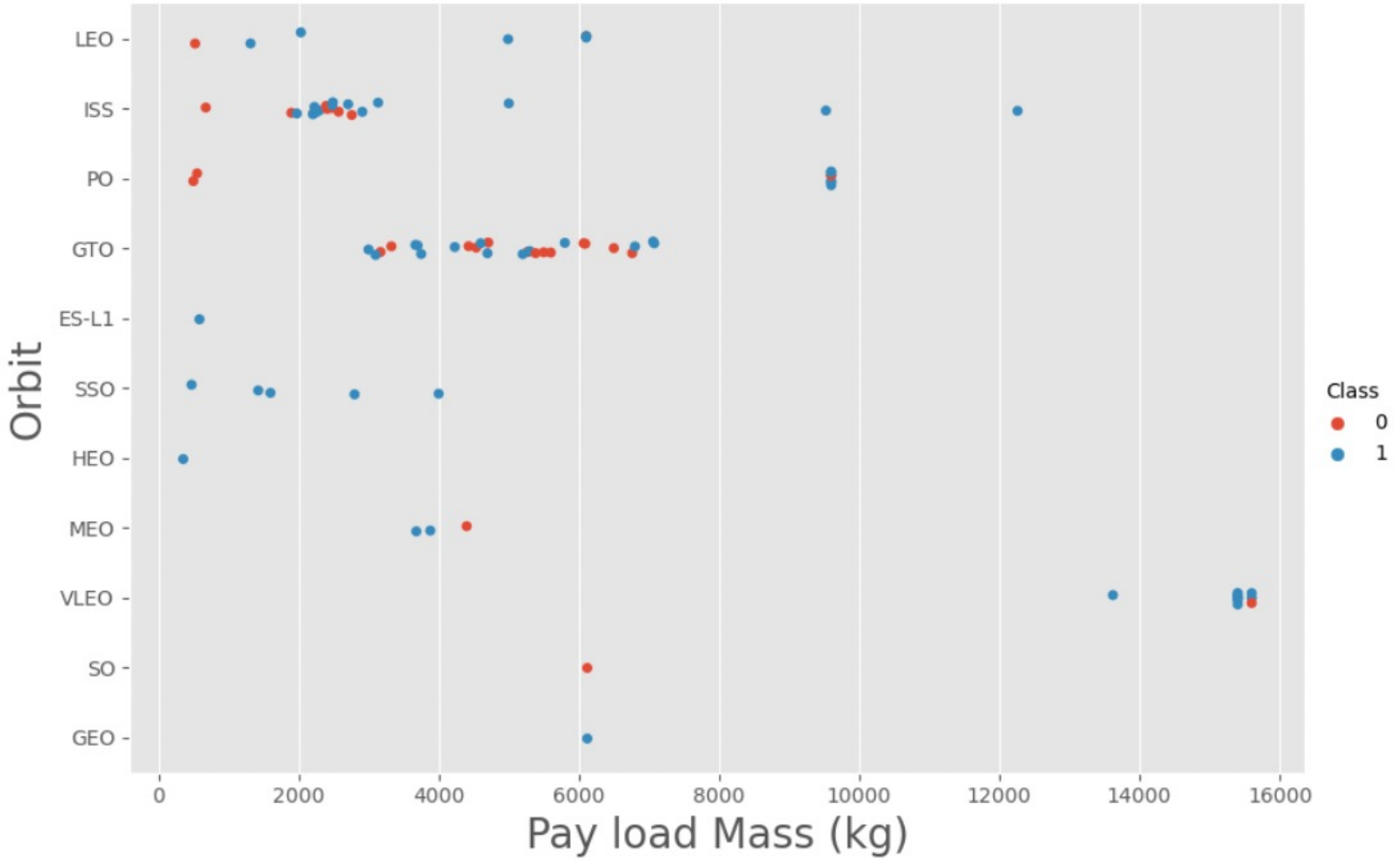
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



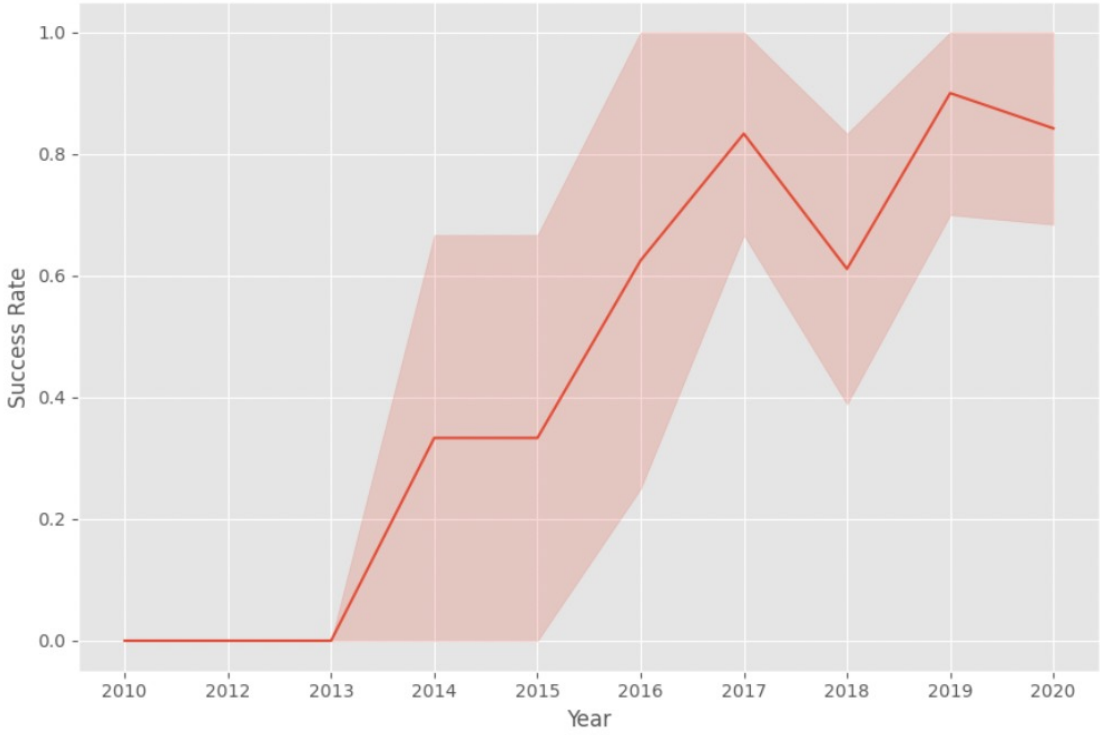
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- Unique launch sites were found to be:

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

Launch Site Names Begin with 'CCA'

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

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

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

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

MIN(DATE)
01-05-2017

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

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

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6

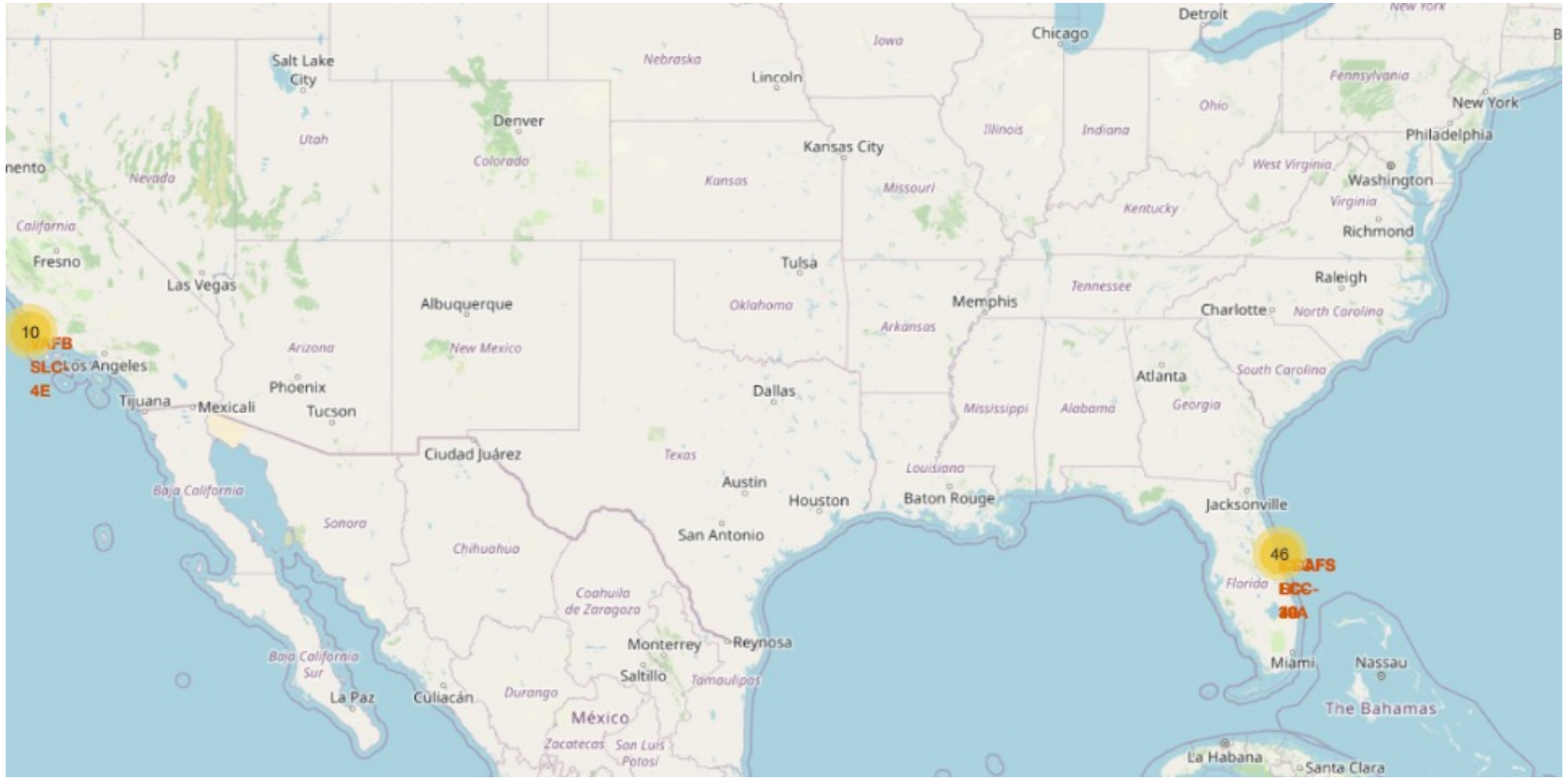
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0
18-08-2020	14:31:00	F9 B5 B1049.6	CCAFS SLC-40	Starlink 10 v1.0, SkySat-19, -20, -21, SAOCOM 1B
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9
18-04-2018	22:51:00	F9 B4 B1045.1	CCAFS SLC-40	Transiting Exoplanet Survey Satellite (TESS)
17-12-2019	00:10:00	F9 B5 B1056.3	CCAFS SLC-40	JCSat-18 / Kacific 1, Starlink 2 v1.0
16-11-2020	00:27:00	F9 B5B1061.1	KSC LC-39A	Crew-1, Sentinel-6 Michael Freilich
15-12-2017	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13
15-11-2018	20:46:00	F9 B5 B1047.2	KSC LC-39A	Es hail 2
14-08-2017	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12

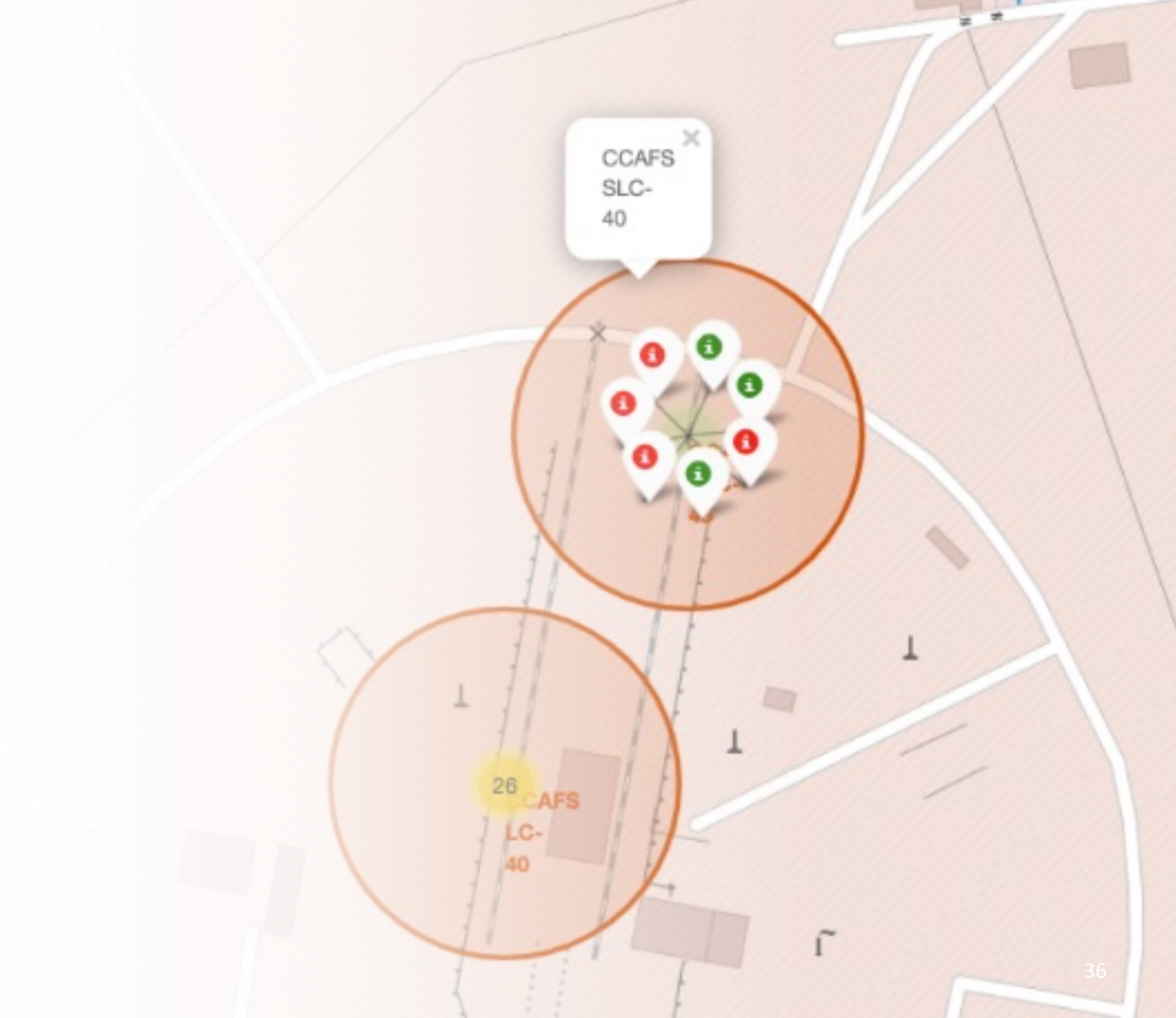

Section 3
Launch Sites
Proximities Analysis

Folium Map for all Locations

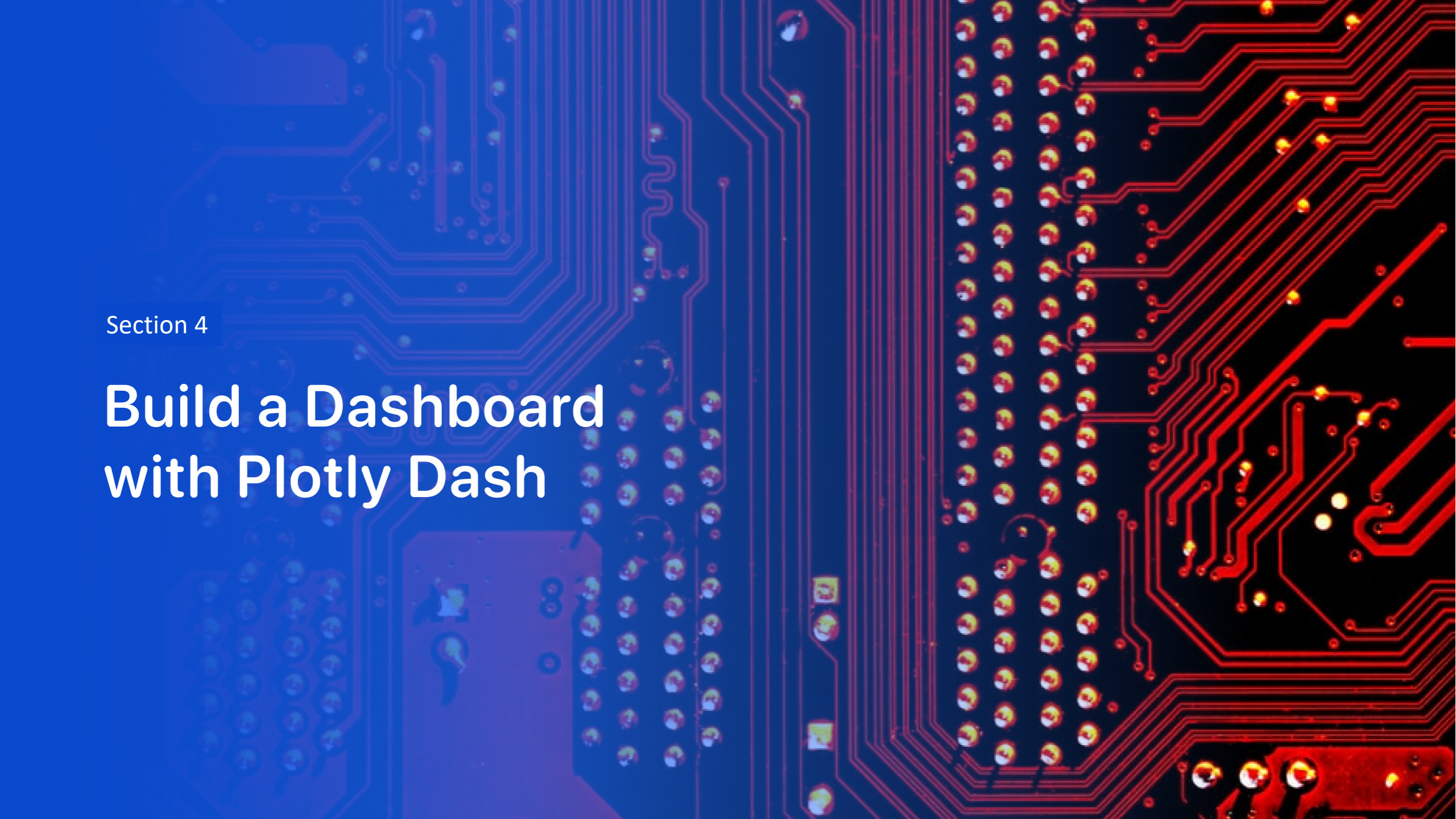


Folium Map for Launch Sites

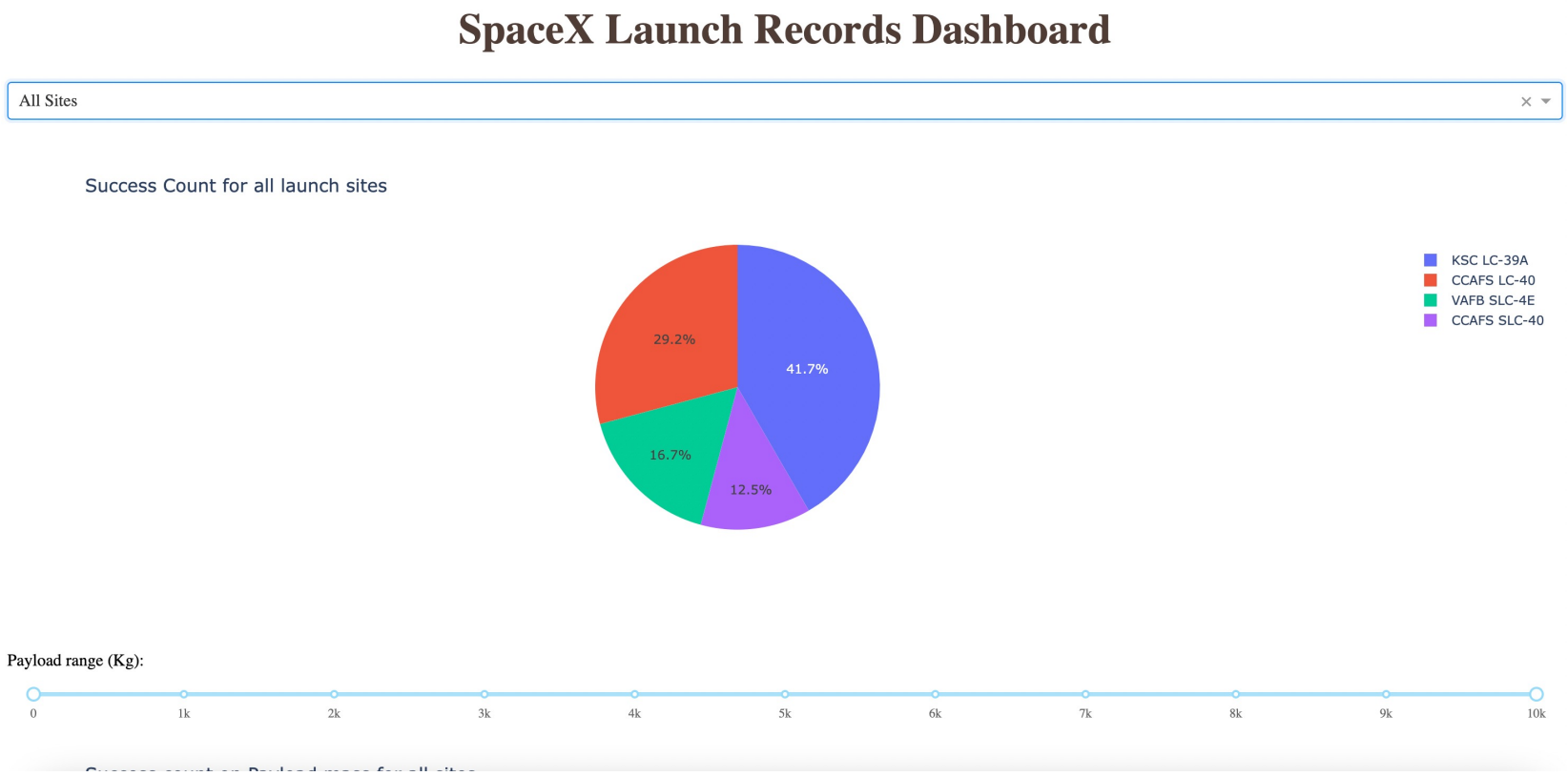
- Markers colored based on on success outcome



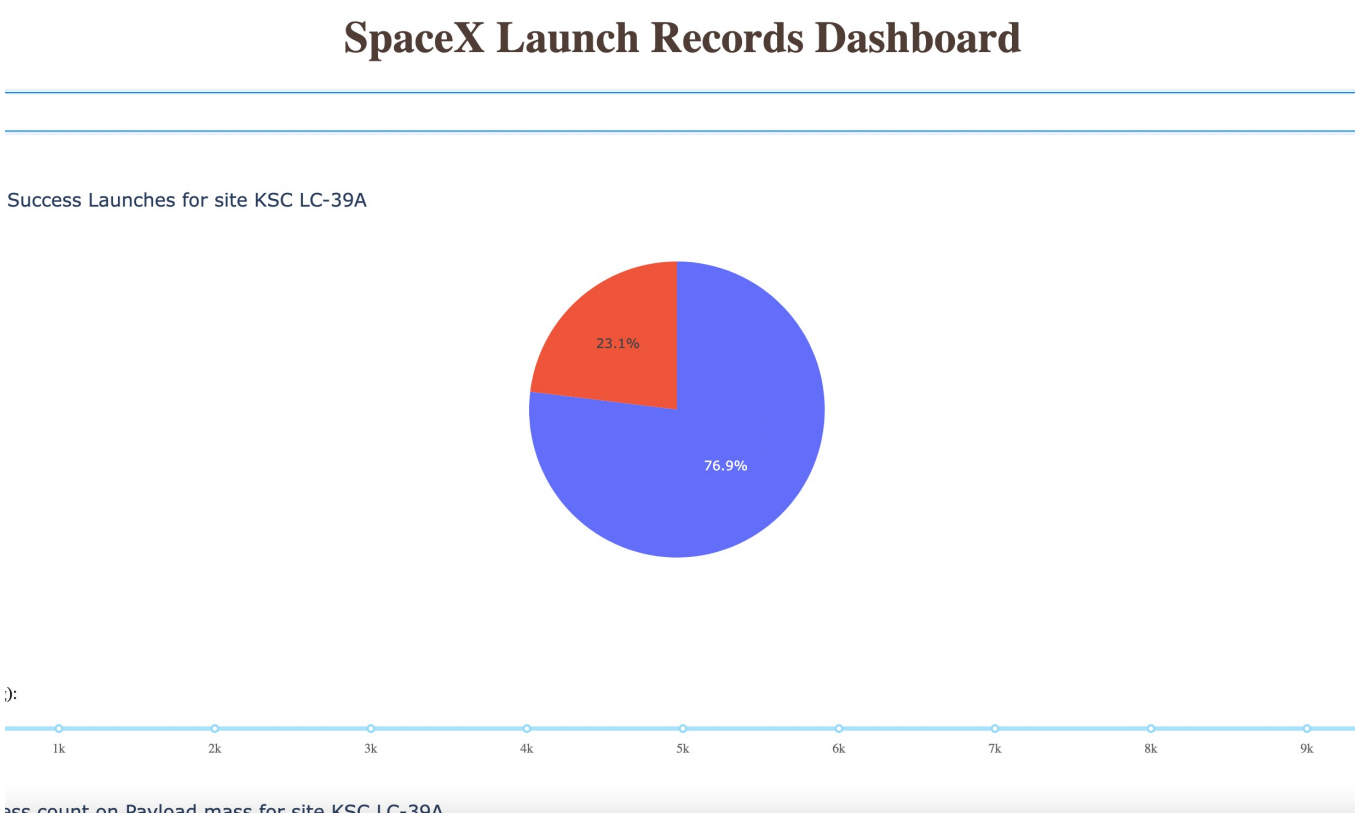
Folium Map for Proximity – Utilizing distance spheres


Section 4
Build a Dashboard with Plotly Dash

Dash: Success count for all sites



Dash: Highest Success Count



Dash: Payload vs Success Count

- Version B4 has the only success count for a payload above 8k (kg).

Section 5

Predictive Analysis (Classification)

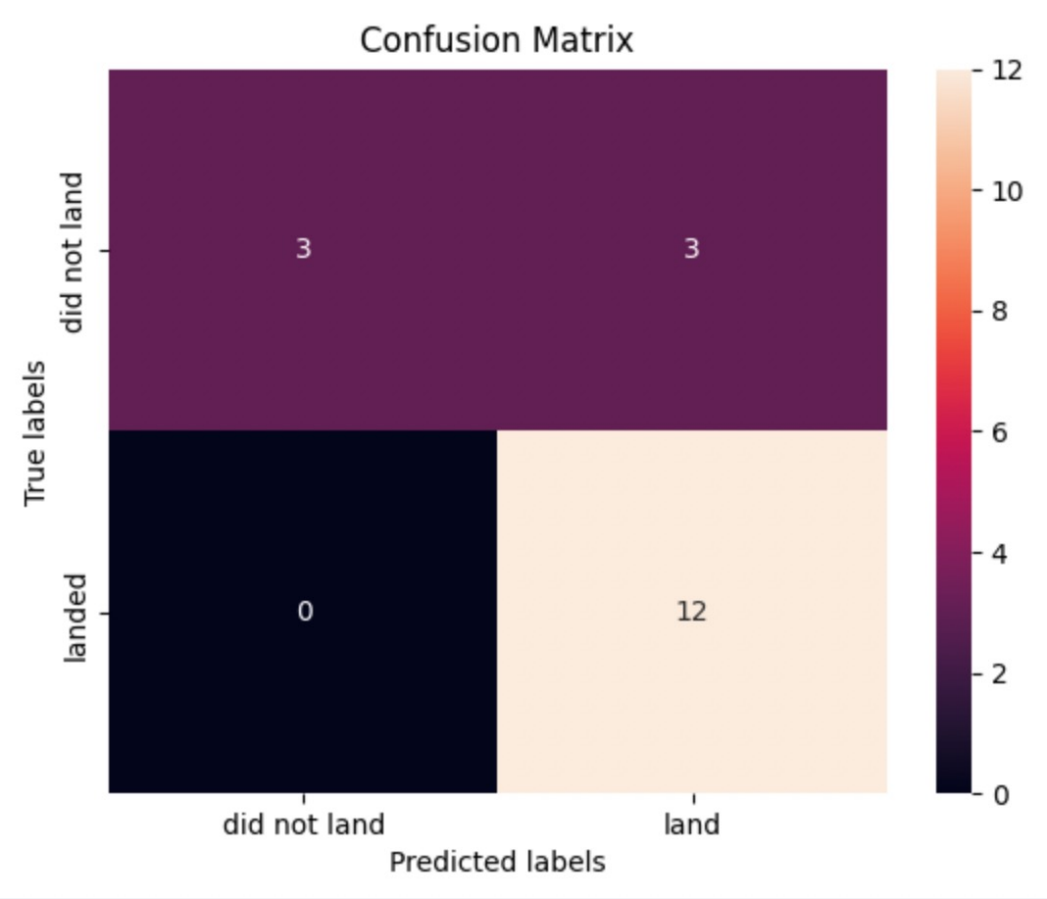
Classification Accuracy

- Equal accuracy for the four selected models

Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

Confusion Matrix

- The confusion matrix for validation purposes displayed for KNN model



Conclusions

- All models performed equally
- More data would be classed as beneficial in improving the reliability of the dataset / calculations
- The launch success has been increasing since 2013
- B4 version showed the highest success count

Appendix

- https://github.com/codebymateus/Capstone_Project

Thank you!