Winning Space Race with Data Science

Fardowsa Abdulle
Feb 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: from 1) SpaceX API 2) Web Scrapping from Wikipedia
  - Data Wrangling : the collected data was refined by creating a landing outcome label
  - Exploratory Data Analysis: with 1) SQL 2) Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Predictive Analysis using Classification
- Summary of all results
  - Data collection results
  - Exploratory Data Analysis determination
  - Predictive analysis results on which classification models

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers can cost upwards of $165 million dollars each. The difference in price is mostly due to SpaceX's practice of re-using the first stage.

If we can determine the success of the first stage landing, we can determine the cost of a launch. This determination can then be used in case an alternate company wants to bid against SpaceX for a rocket launch

Project Goal

Create a machine learning pipeline to predict the success of first stage landing

Questions

1) What factors determine if the rocket will land successfully?

2) What interactions contribute to the success rate of a successful landing?

3) What operating conditions need to be in place for success in landing?

Section 1

Methodology

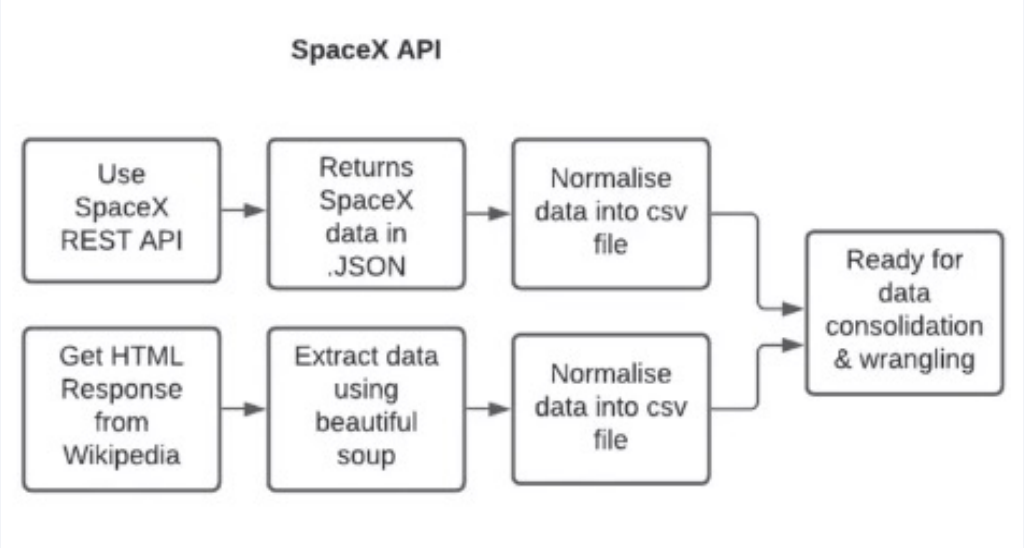
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features and null values were cleaned
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models (Logistic Regression, Support Vector Model, Decision Tree, and K Nearest Neighbors model) were evaluated to determine the most accurate for the project

Data Collection

- Datasets were collected
  - SpaceX launch data gathered from the SpaceX API
  - API: gives us data about launches, information about the rockets used, payload, as well as specifications on launching and landing. The API also includes landing outcomes.
  - Falcon 9 launch records were obtained using web scraping from Wikipedia
  - BeautifulSoup was used to extract launch records as a HTML table, parse information in the table, and convert into understandable data and stored in dataframes using pandas



Data Collection – SpaceX API

- Data collection with SpaceX REST calls
- Github link

1 – Getting response from API

```
In [6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

In [7]: response = requests.get(spacex_url)
```

2- Converting json result to dataframe

```
In [11]: # Use json_normalize meethod to convert the json result into a dataframe
         data = pd.json_normalize(response.json())
```

3- Data is cleaned and list is assigned to dictionary

```
In [18]: # Call getLaunchSite
         getLaunchSite(data)

In [19]: # Call getPayloadData
         getPayloadData(data)

In [20]: # Call getCoreData
         getCoreData(data)
```

```
In [21]: launch_dict = {'FlightNumber': list(data['flight_number']),
         'Date': list(data['date']),
         'BoosterVersion':BoosterVersion,
         'PayloadMass':PayloadMass,
         'Orbit':Orbit,
         'LaunchSite':LaunchSite,
         'Outcome':Outcome,
         'Flights':Flights,
         'GridFins':GridFins,
         'Reused':Reused,
         'Legs':Legs,
         'LandingPad':LandingPad,
         'Block':Block,
         'ReusedCount':ReusedCount,
         'Serial':Serial,
         'Longitude': Longitude,
         'Latitude': Latitude}
```

4- Data was filtered to only include

```
In [25]: # Hint data['BoosterVersion']!='Falcon 1'
         data_falcon9=df[df['BoosterVersion']!='Falcon 1']
```

Data Collection - Scraping

- Data collection with with web scrapping
- Github link

1- Getting HTTP response from HTML

```
r = requests.get(static_url)
data = r.text
```

2- Creating BeautifulSoup object

```
soup = BeautifulSoup(data,"html.parser")
```

3- Finding tables

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

5- Appending data to keys

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wik
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corr
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
```

6- Converting dictionary to dataframe and saving to CSV

```
df = pd.DataFrame.from_dict(launch_dict)
df.to_csv('spacex_web_scraped.csv', index=False)
```

4 – Getting Column names and creating dictionary

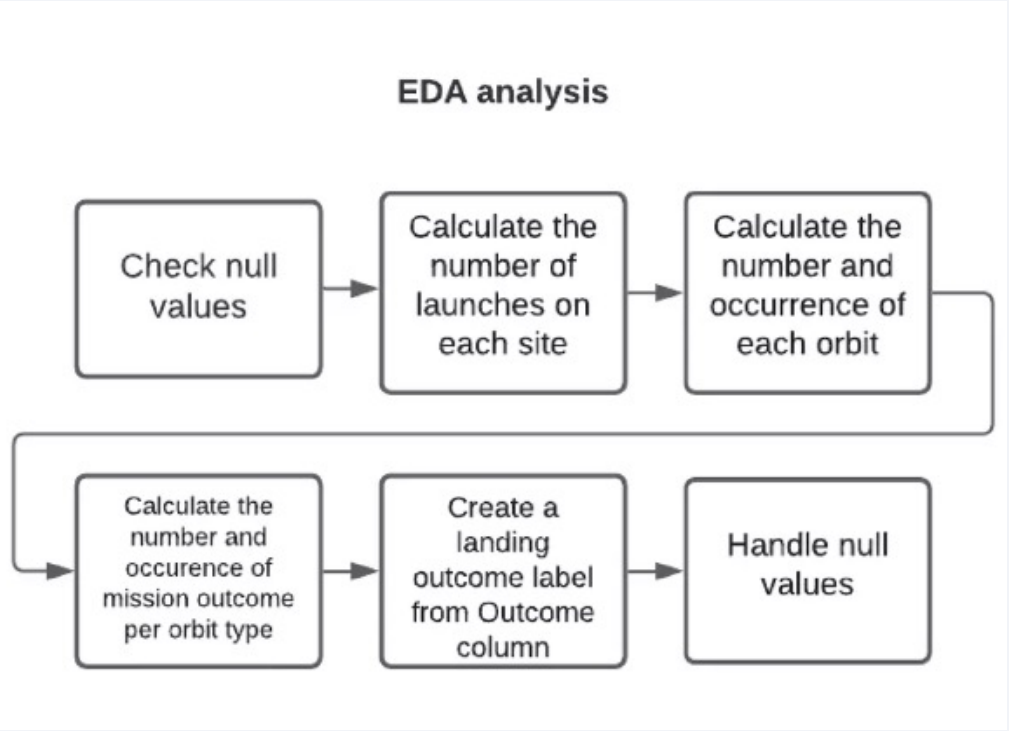
```
column_names = []
table_headers = first_launch_table.find_all('th')

for j, table_header in enumerate(table_headers):
    name = extract_column_from_header(table_header)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

```
In [34]:

launch_dict= dict.fromkeys(column_name

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with e
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

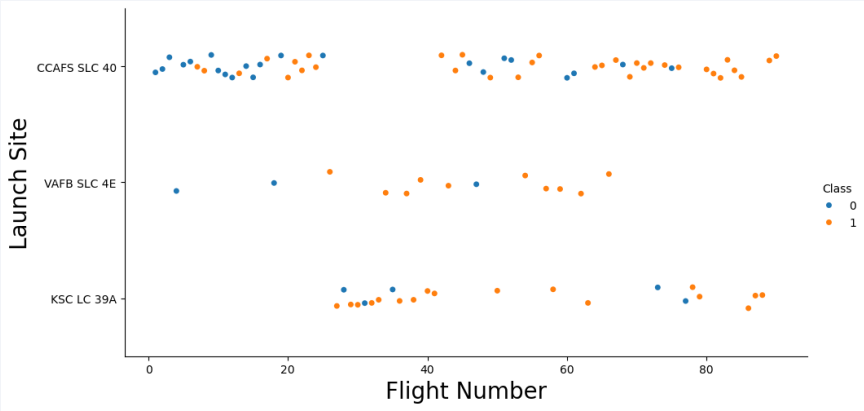
Data Wrangling

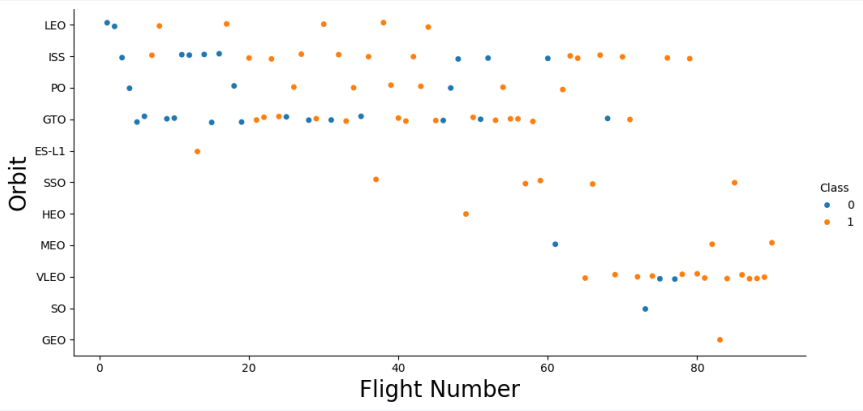
- Exploratory data analysis was performed to determine training labels
- We calculated:
  - The number of launches at each site
  - The number and occurrence of each orbits
- The label “Landing Outcome” was created from the outcome column
- The results were exported to CSV

Github link



EDA with Data Visualization

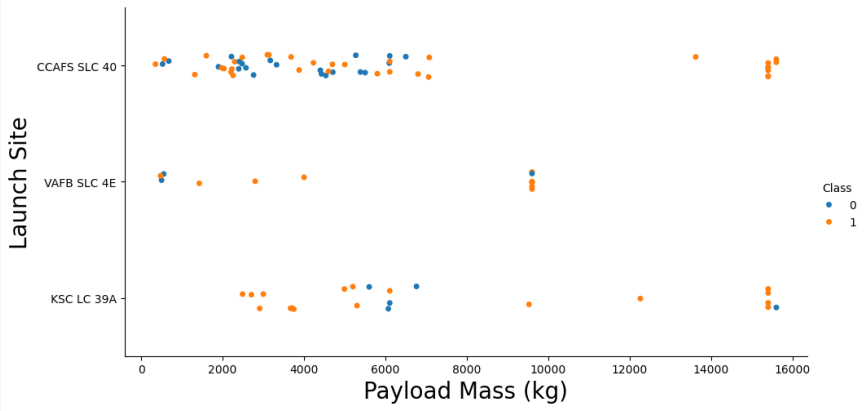


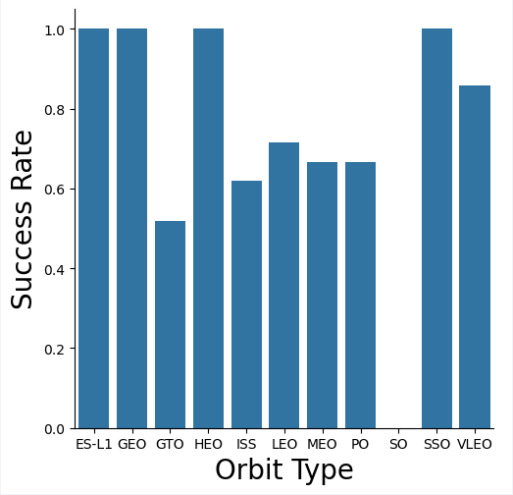


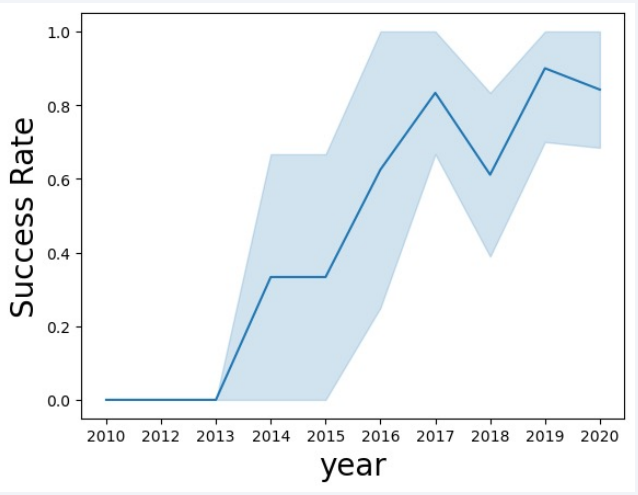
From top to bottom, left to right

1) Successful flights by launch site
2) Success by orbit
3) Successful flights by payload mass
4) Success rate per orbit type
5) Success rate per year

[Github Link]



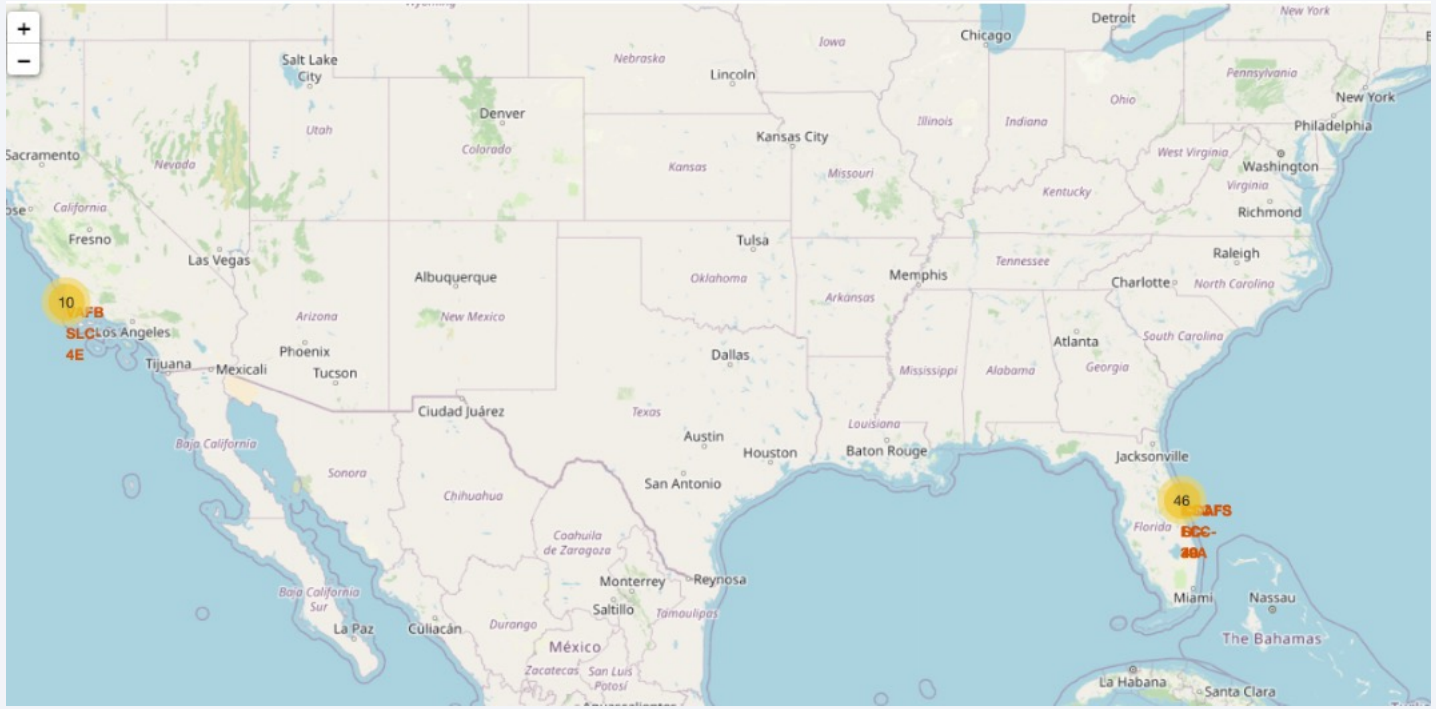




EDA with SQL

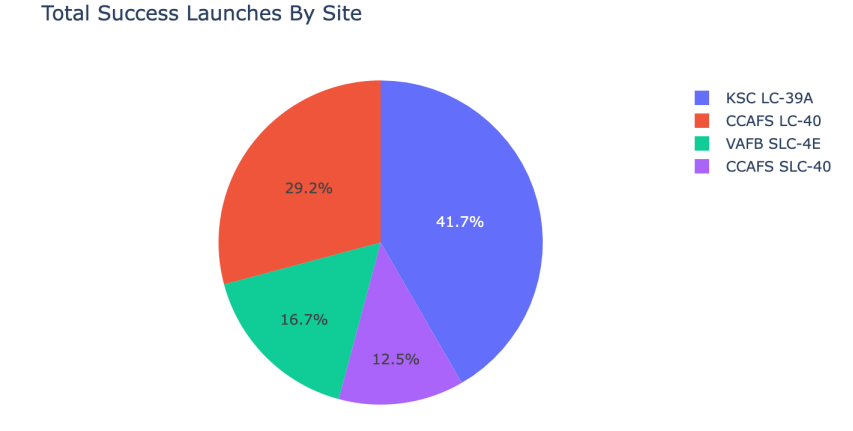
- The SpaceX dataset was loaded into PostegreSQL database in the jupyter notebook
- SQL queries performed included
  - Displaying the names of the unique launch sites
  - Displaying the total payload mass carried by different sites
  - Listing dates where successful landing outcomes occurred
  - Listing the total number of successful and unsuccessful missions
- Github Link

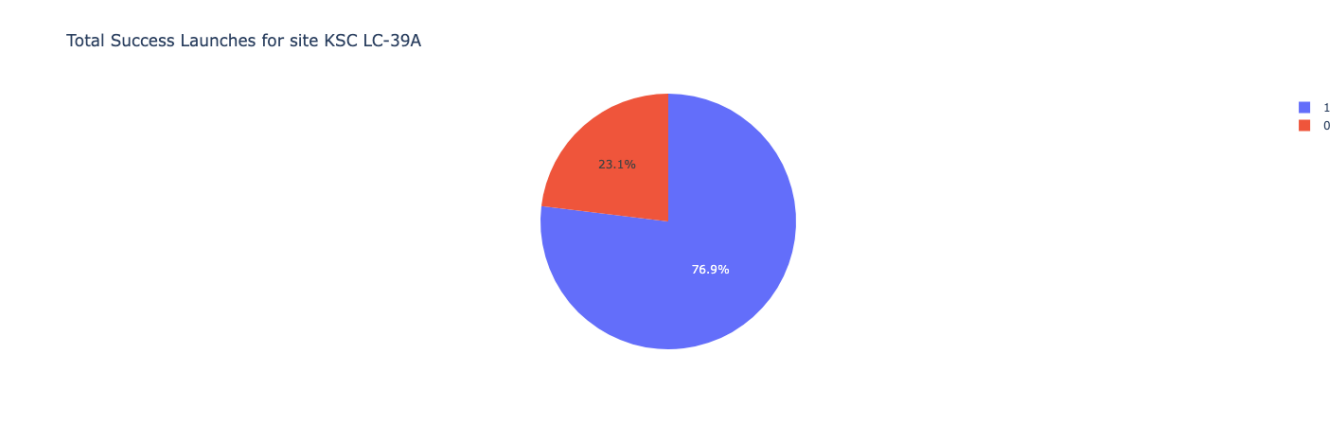
Build an Interactive Map with Folium



- Map markers were added to the map with the aim of comparing launch site locations and finding the optimal site
- [Github Link]

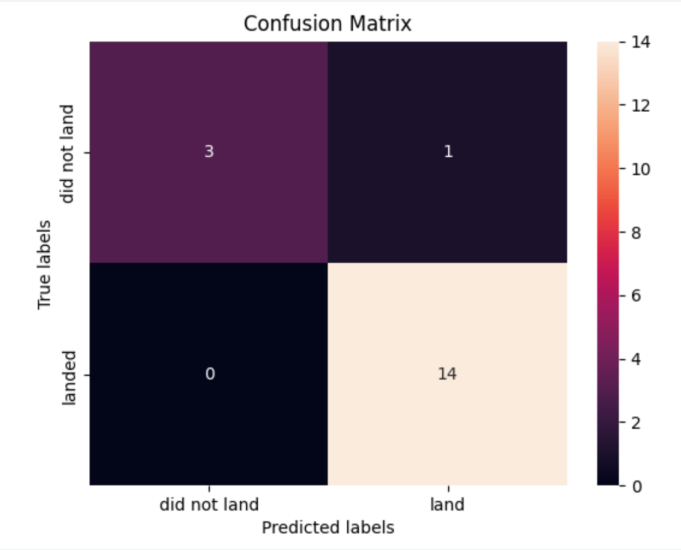
Build a Dashboard with Plotly Dash

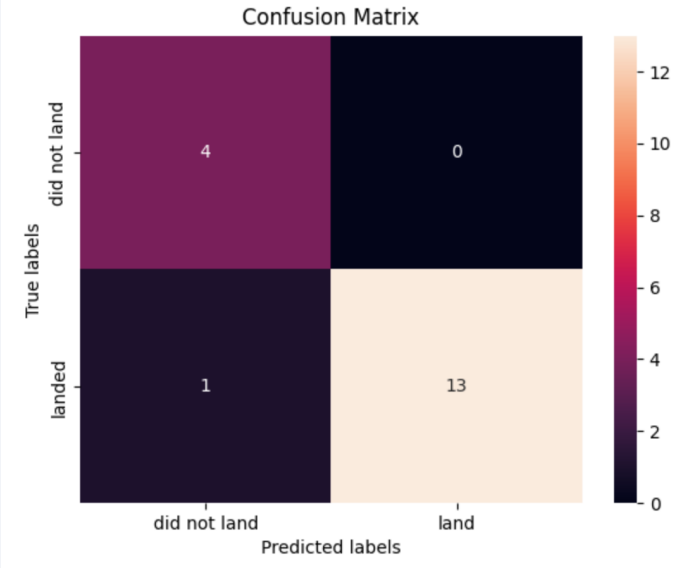


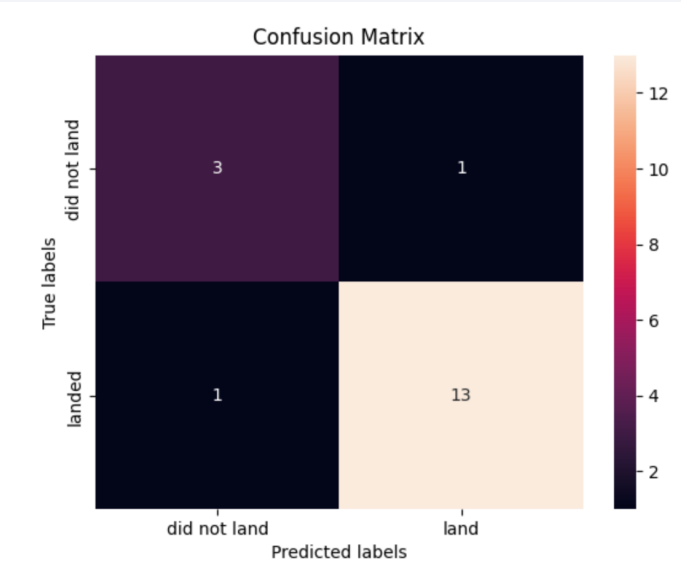


- Launch site KSC LC-39A had the most successful launches, with a success rate of 76.9%
- The success rates of low weighted payloads are higher than high weighted payloads
- Github link

Predictive Analysis (Classification)



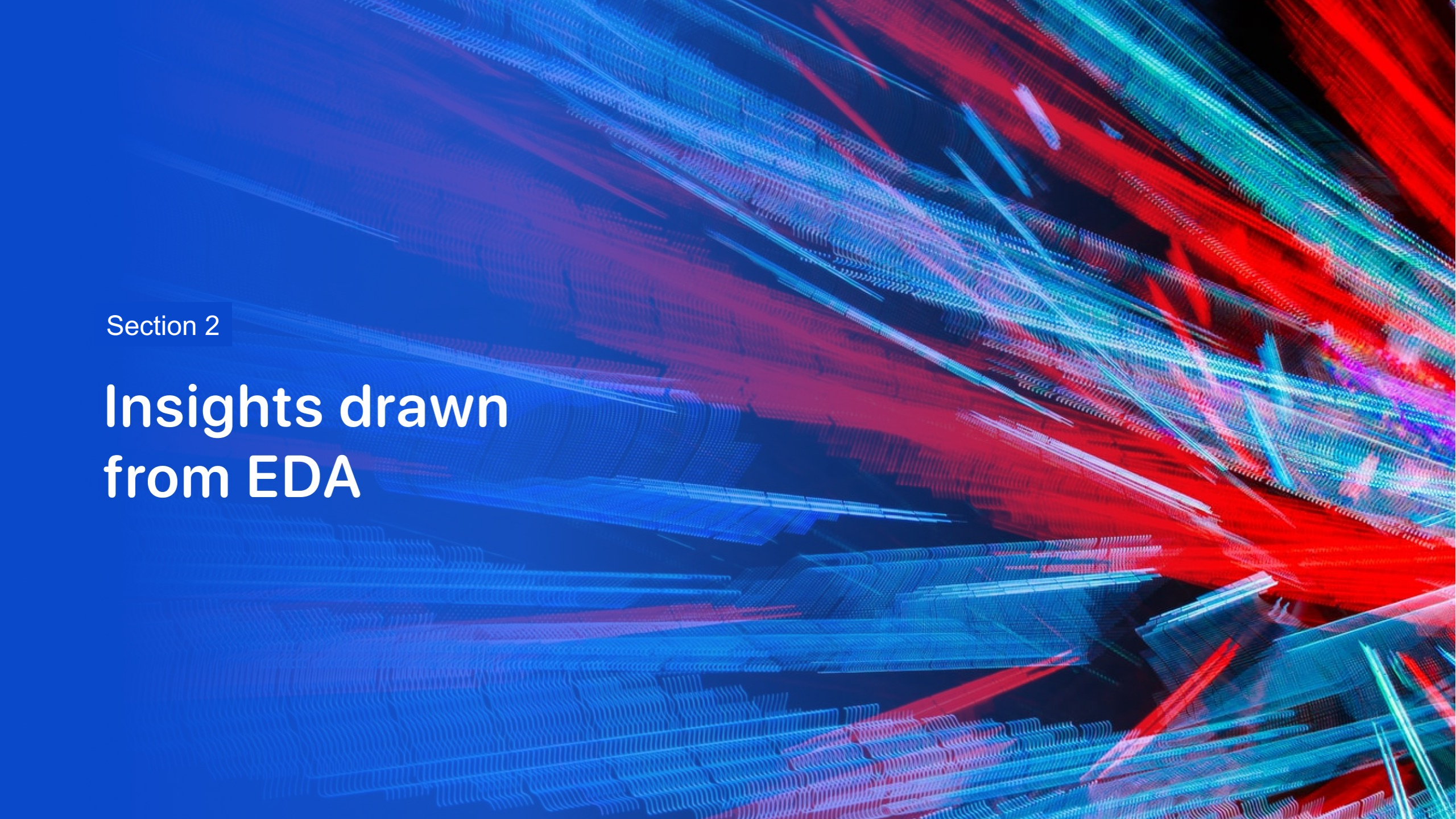




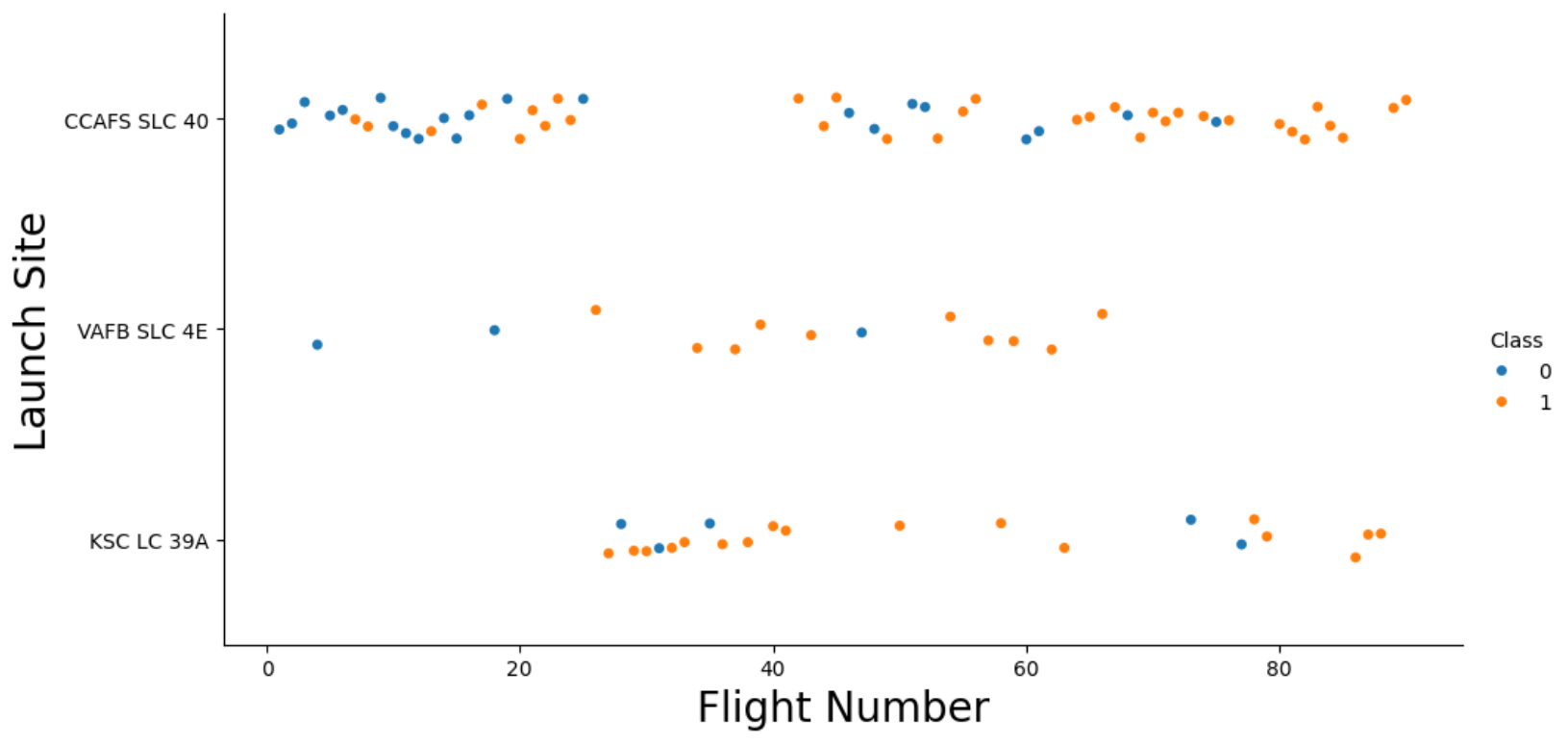
- The KNN, Decision Tree, and logistic regression models achieved the highest accuracy at 94%. The SVM also had the best area under the curve calculation of 0.958.
- Github link

Results

- From the various labs we found
  - Launch site KSC LC 39A has the most successful launches out of all sites
  - Low weighted payloads have better success rates than higher weighted payloads
  - The KNN, SVM, and logistic regression models are the best in terms of prediction accuracy testing of the dataset

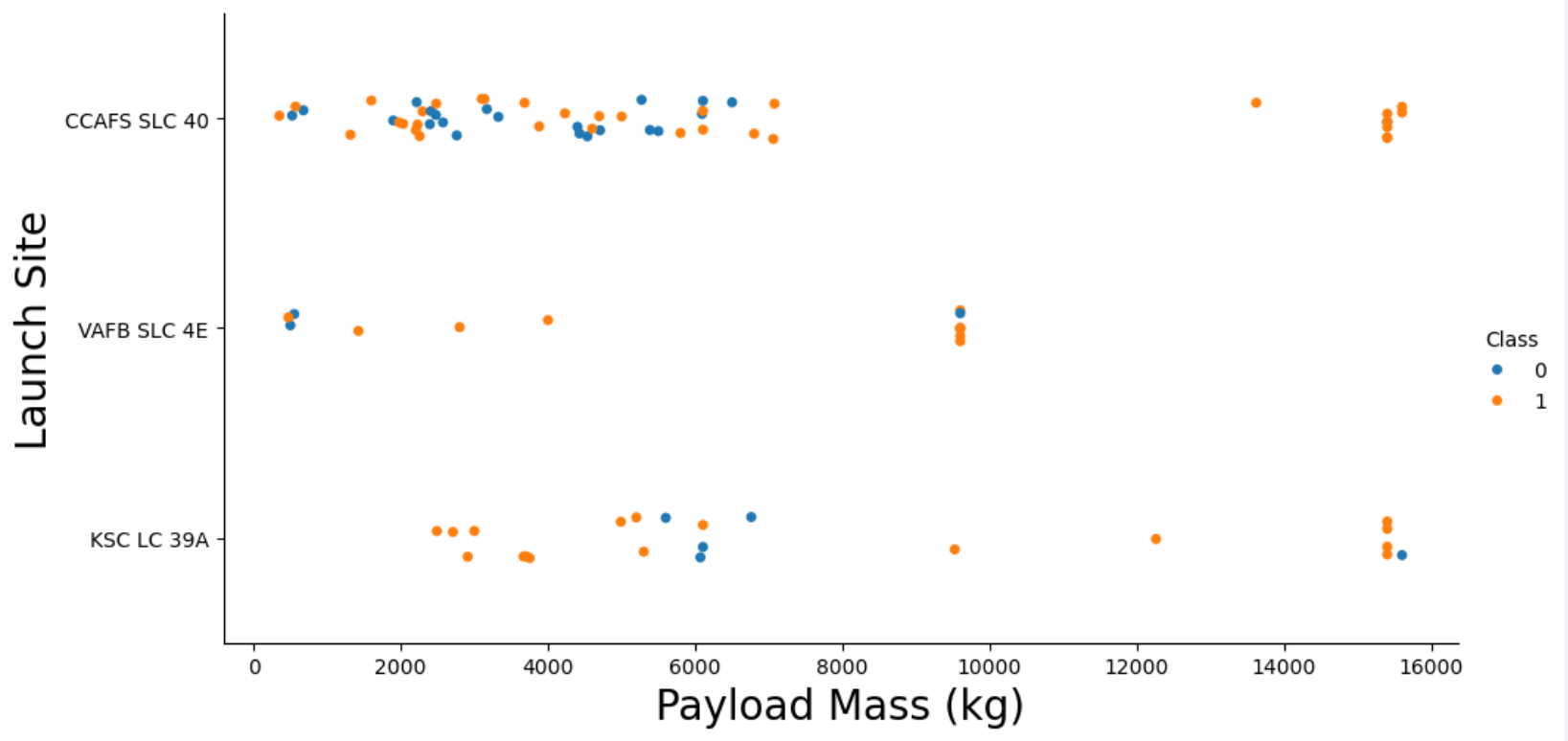

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



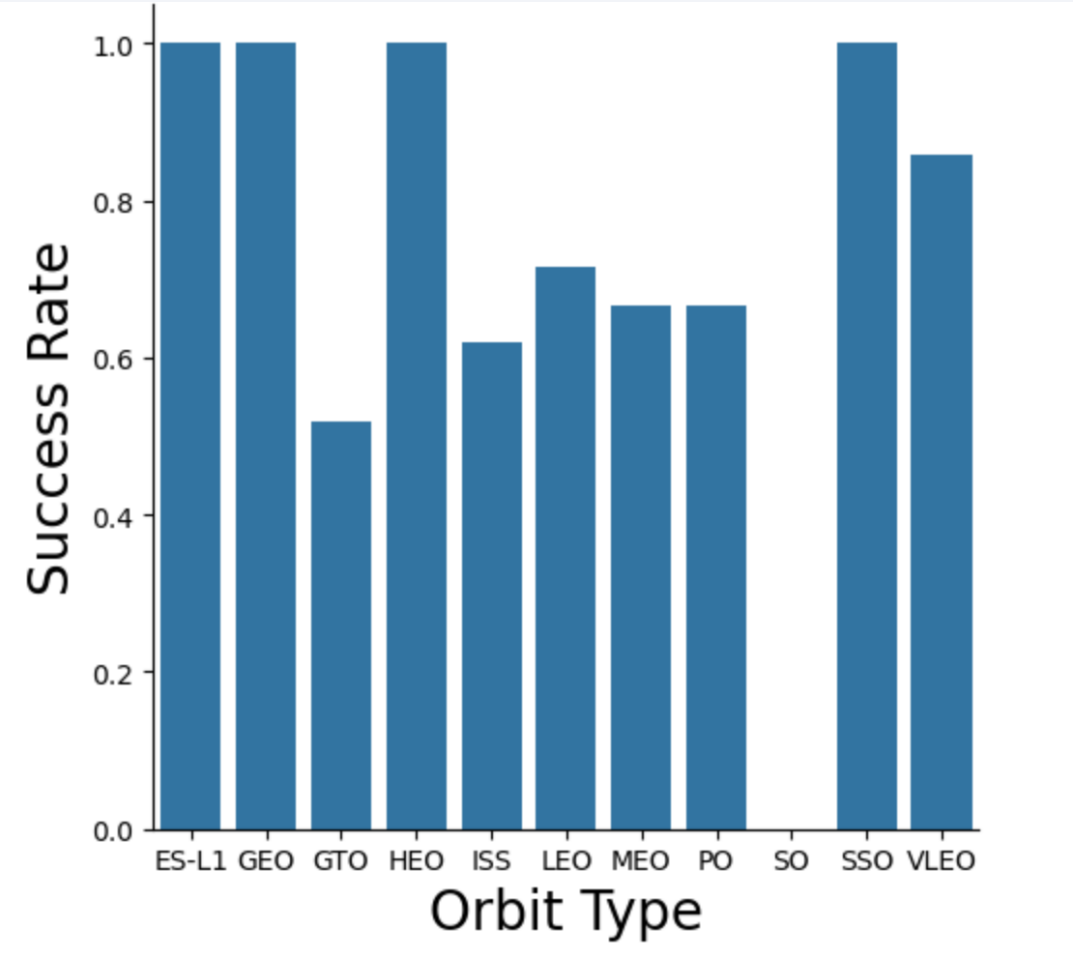
Launches from the site CCAFS SLC 40 are higher than launches from other sites

Payload vs. Launch Site



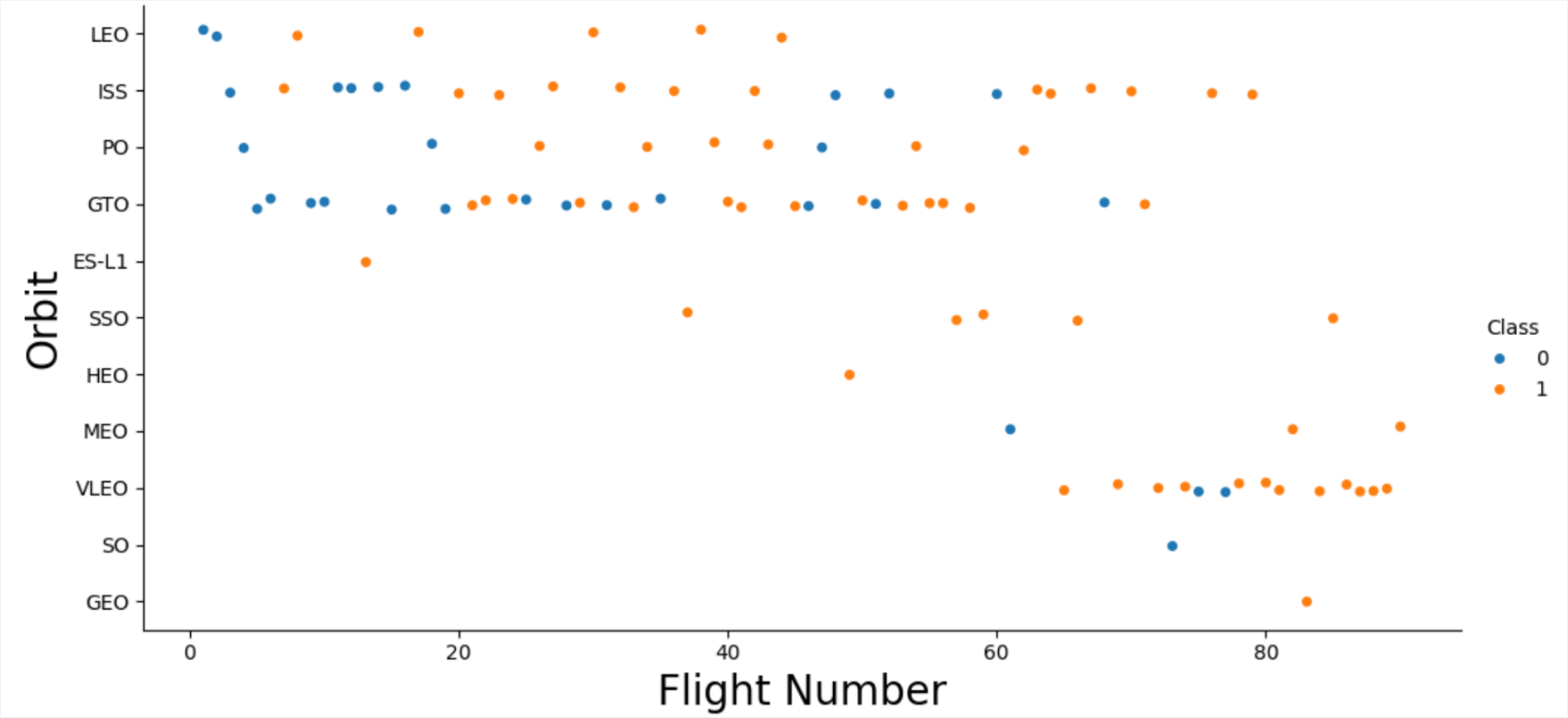
The launch site CCAFS SLC 40, which has the most launches, also launches the lowest payload masses

Success Rate vs. Orbit Type



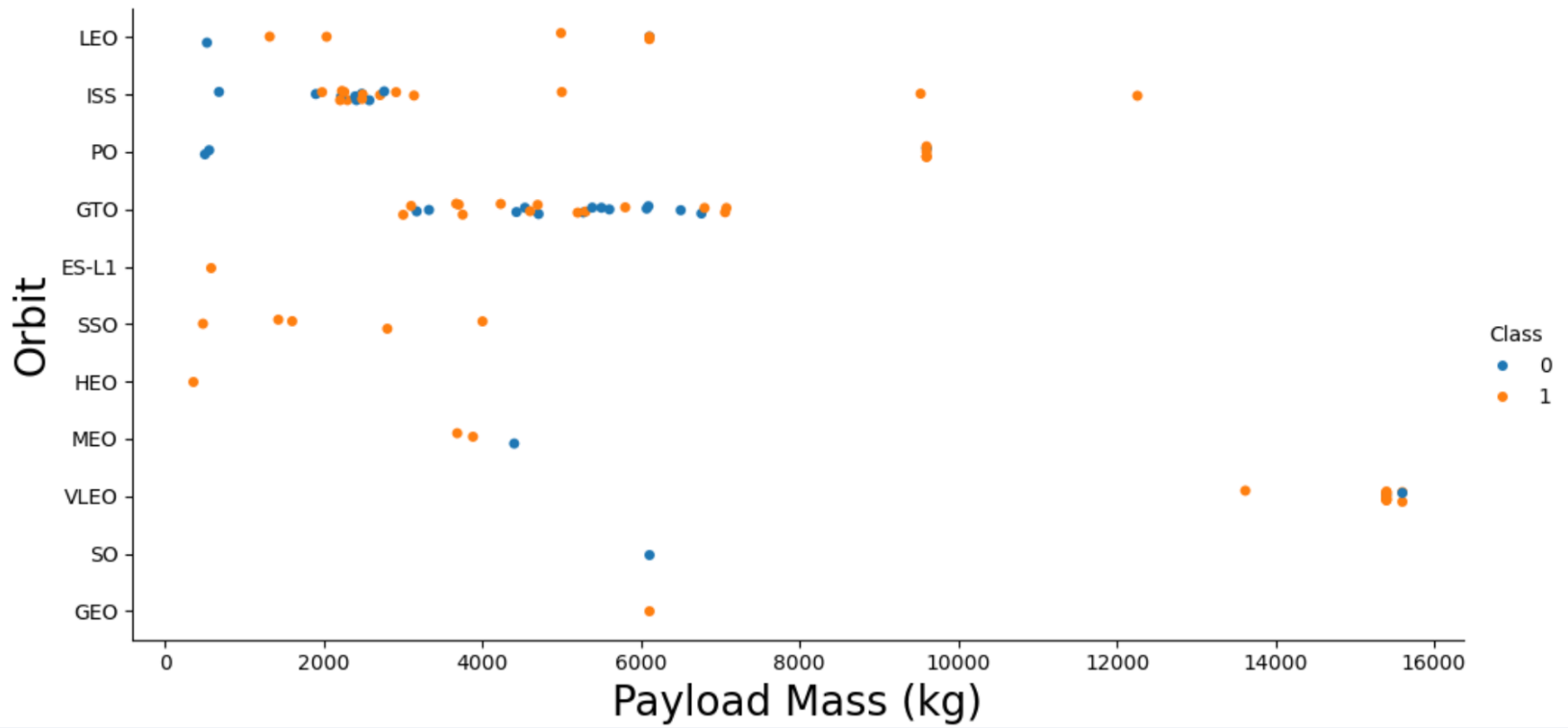
The orbit types of ES-L 1, GEO, HEO, SSO have the highest success rate

Flight Number vs. Orbit Type



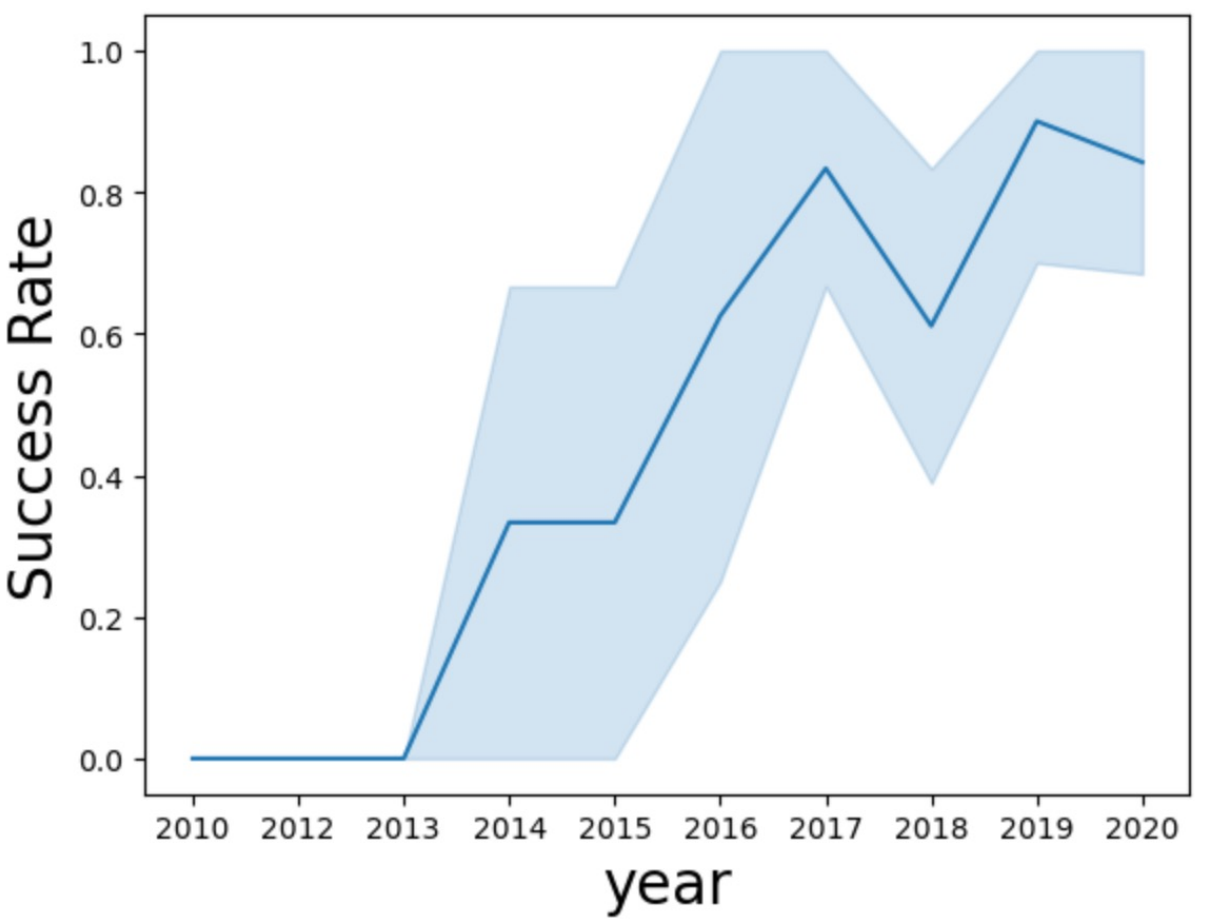
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend



The success rate improved greatly during 2013 and has been steadily increasing since. There was a slight dip in 2018 but the rate improved in 2019 again and seems to have stabilized.

All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

```
%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_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
```

sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

```
%sql SELECT min(date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BET
```

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

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

Boosters Carried Maximum Payload

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

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

2015 Launch Records

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

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

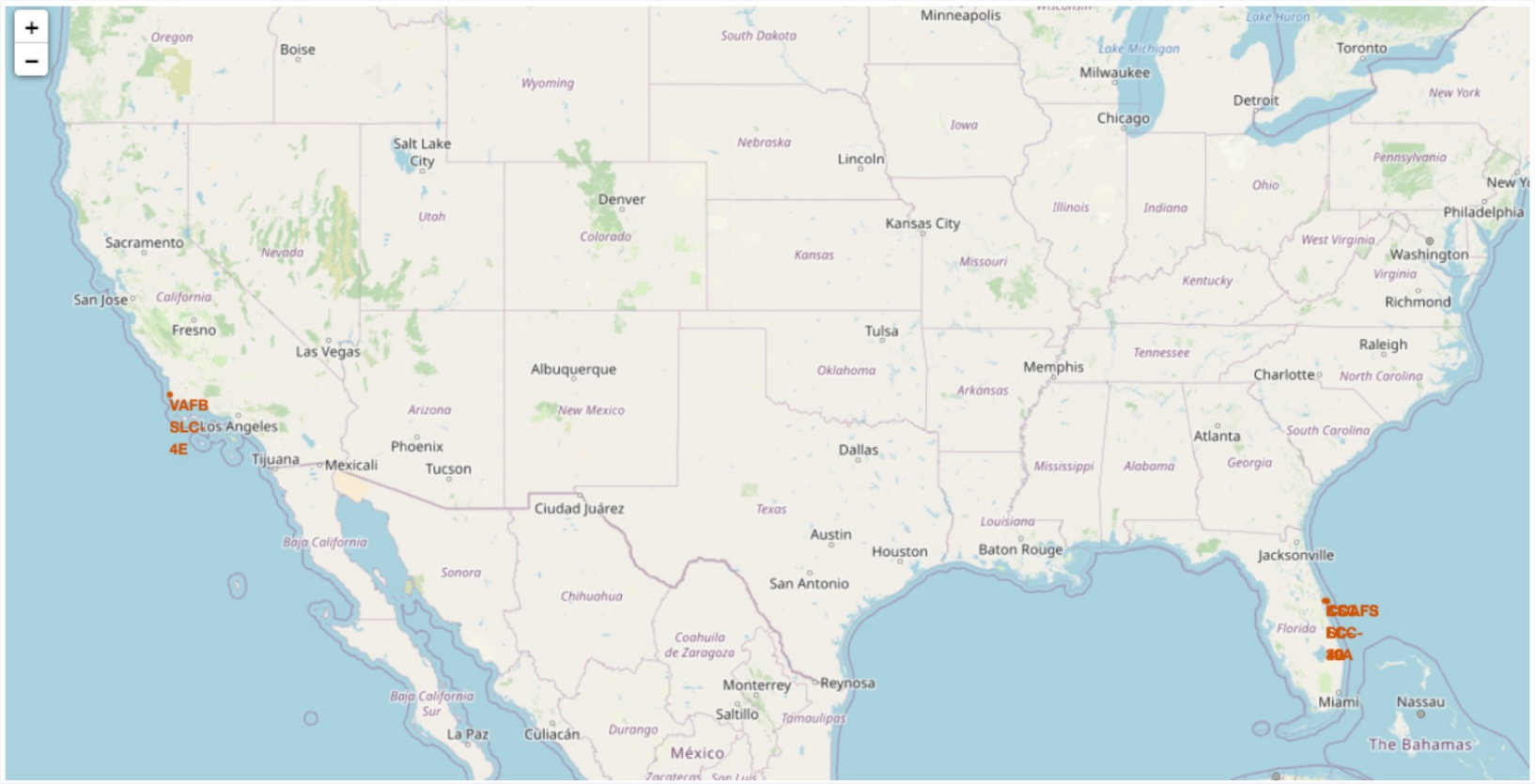
```
%sql select count(Landing_Outcome), Landing_Outcome from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' group by Landing_Outcome\
order by count(Landing_Outcome) desc
```

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

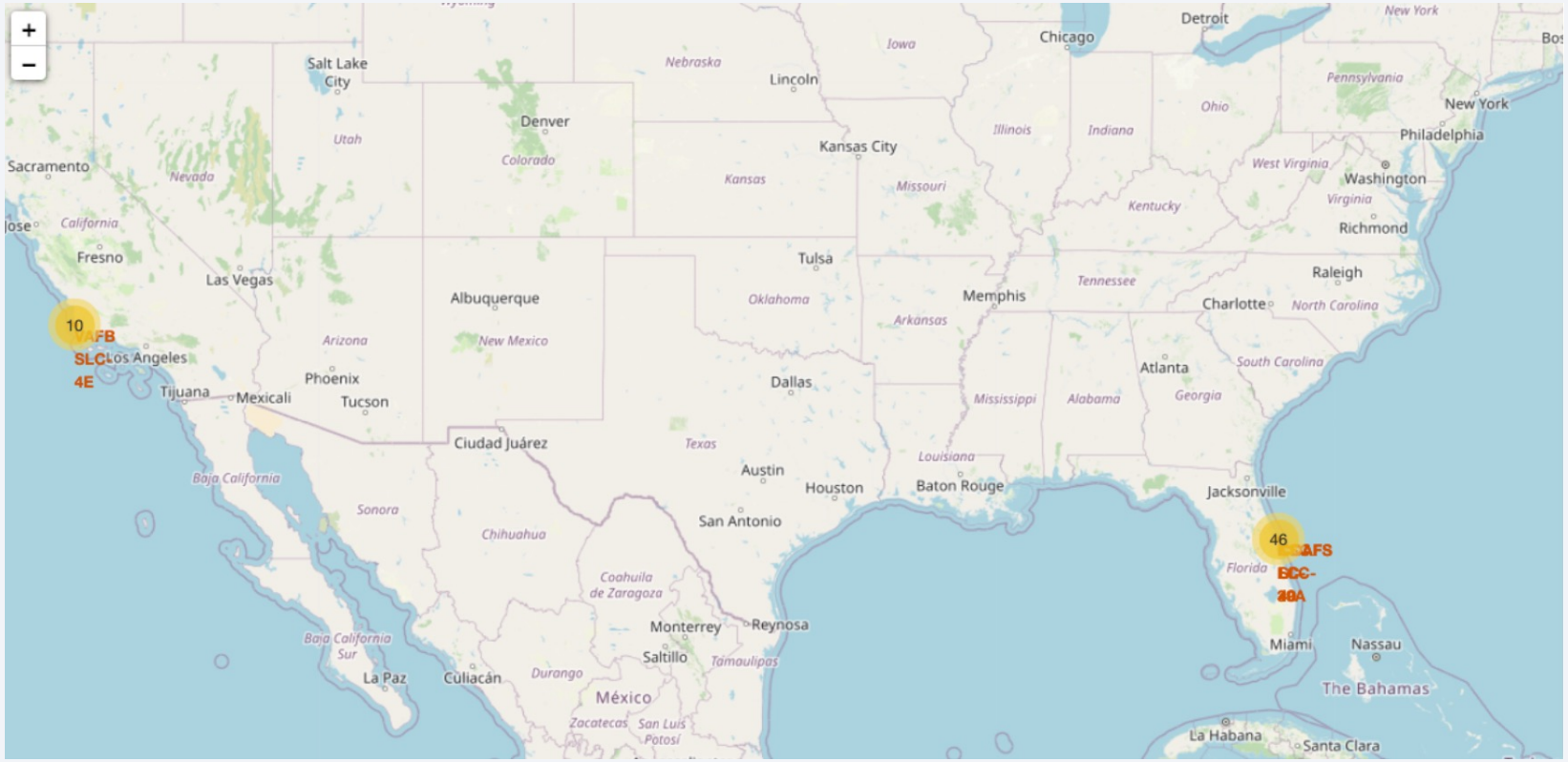
Section 3

Launch Sites Proximities Analysis

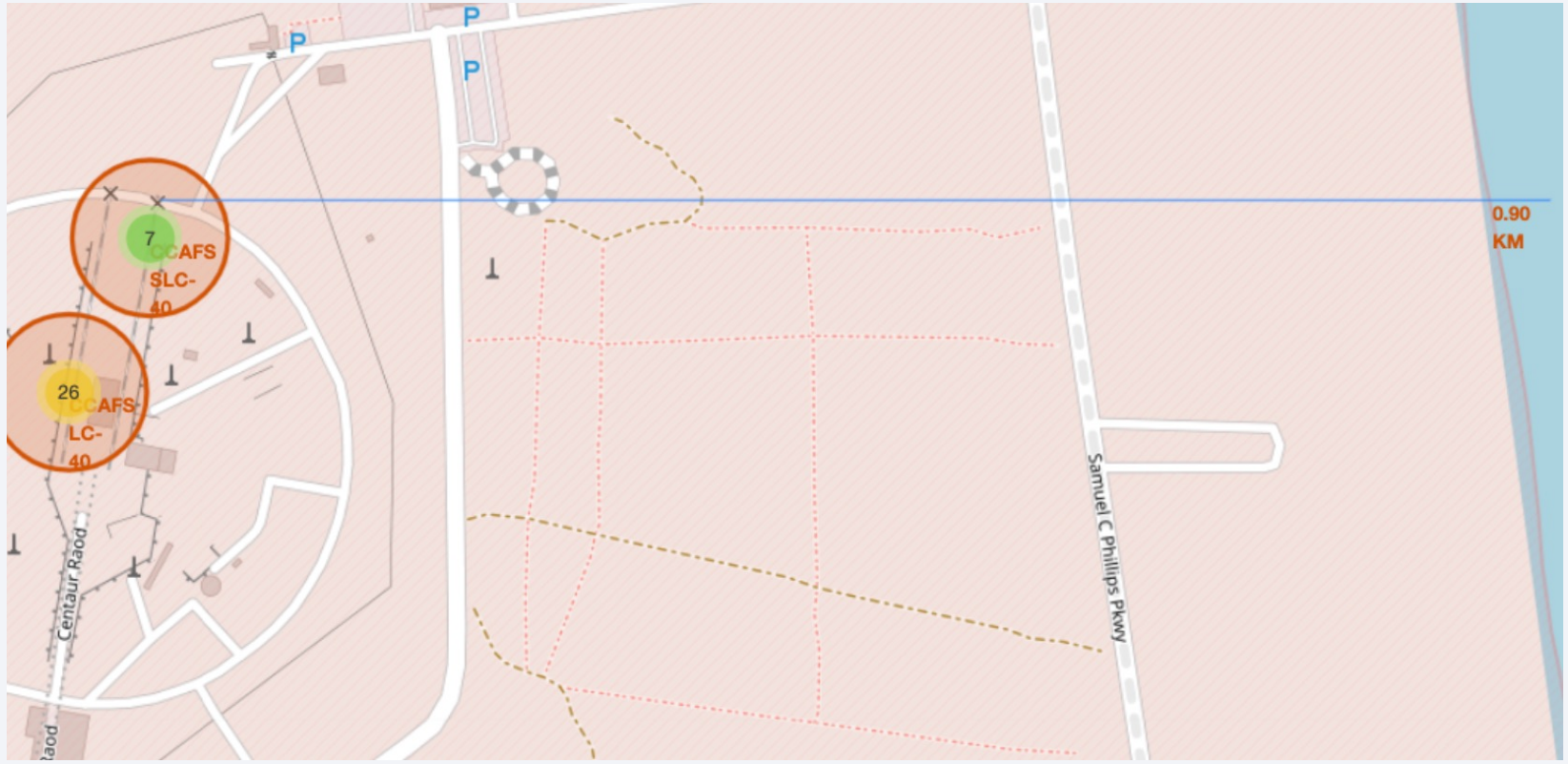
All launch sites marked on Map

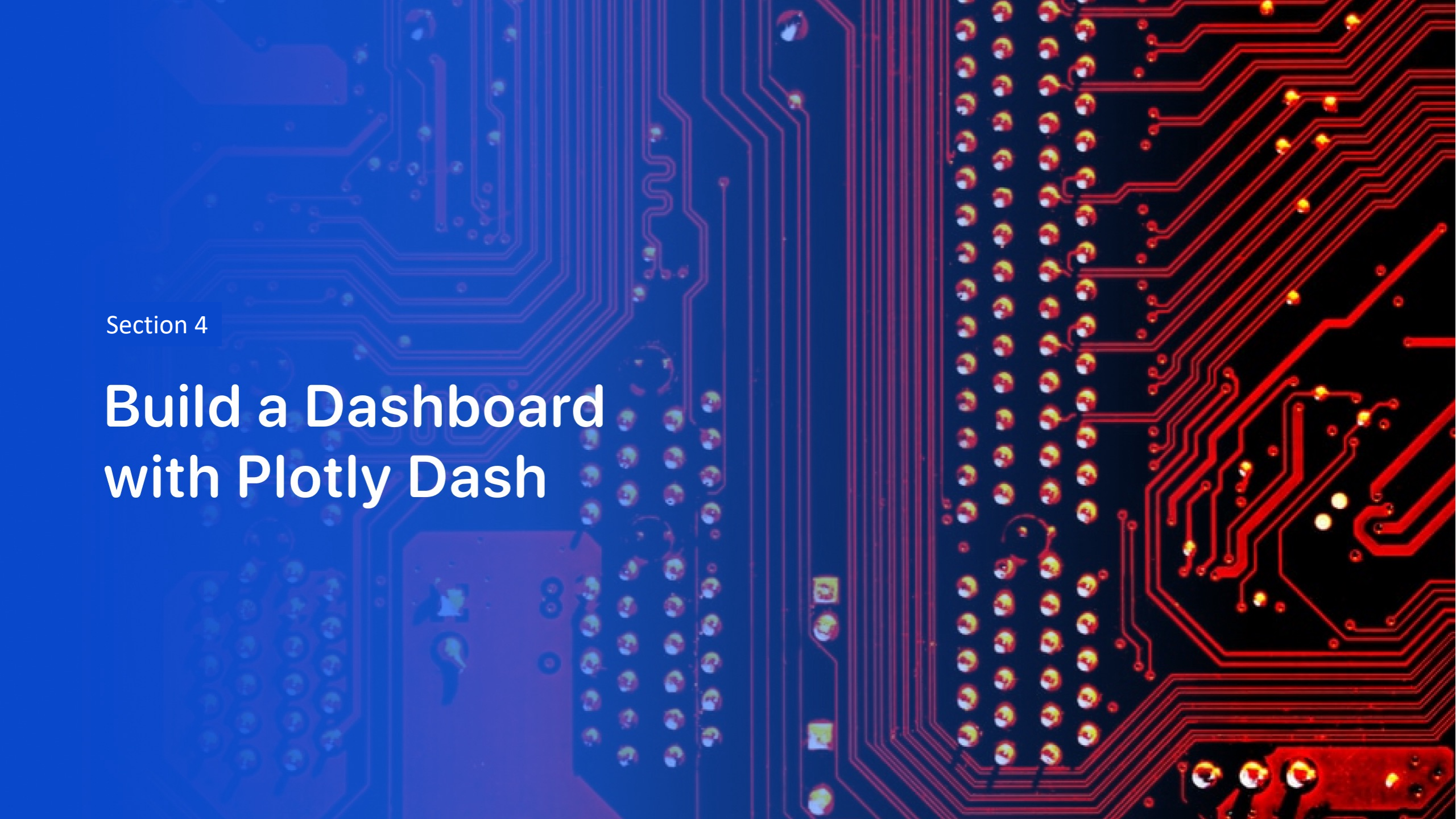


Successful/failed launches marked on Map

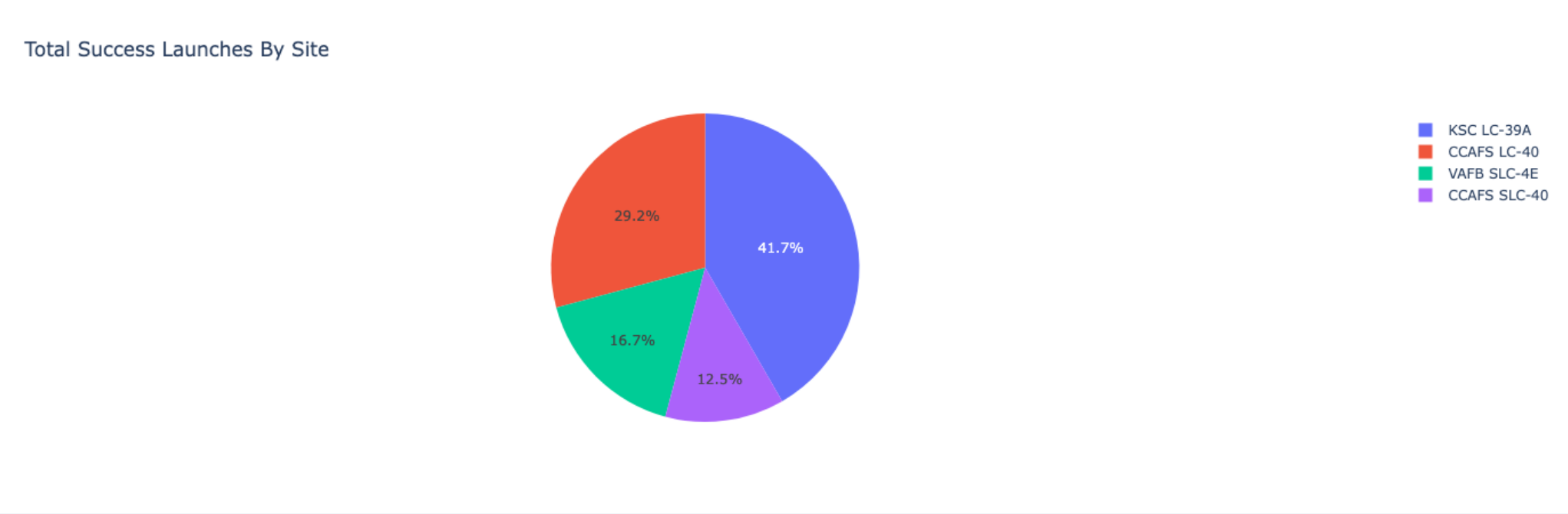


Distances between launch site

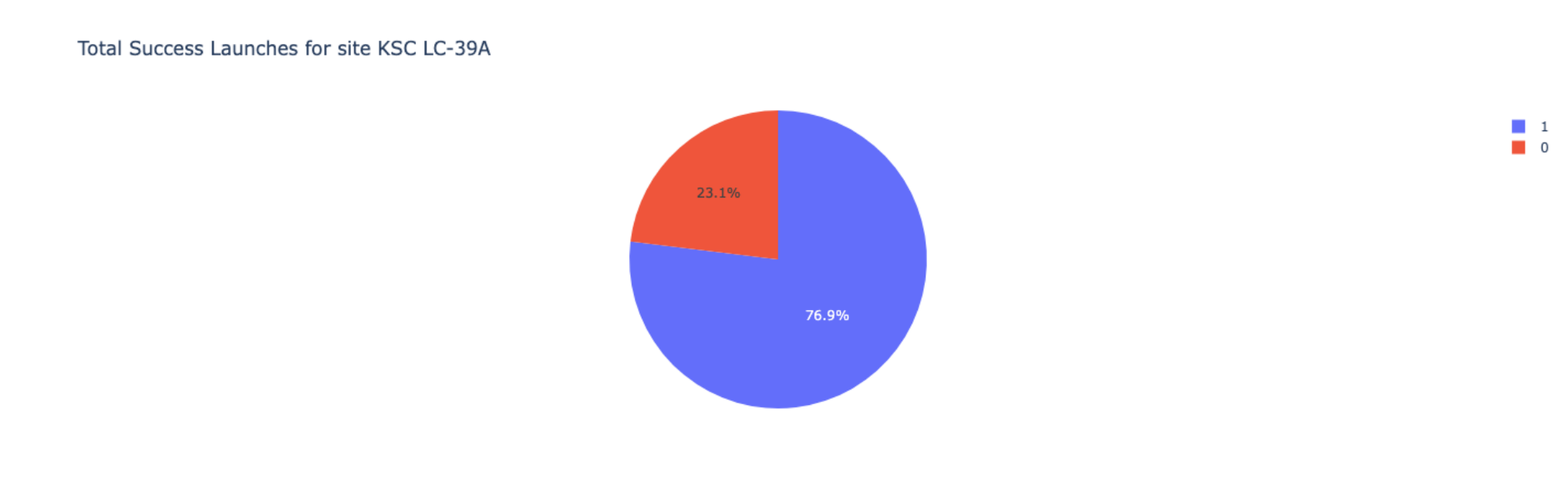



Section 4
Build a Dashboard with Plotly Dash

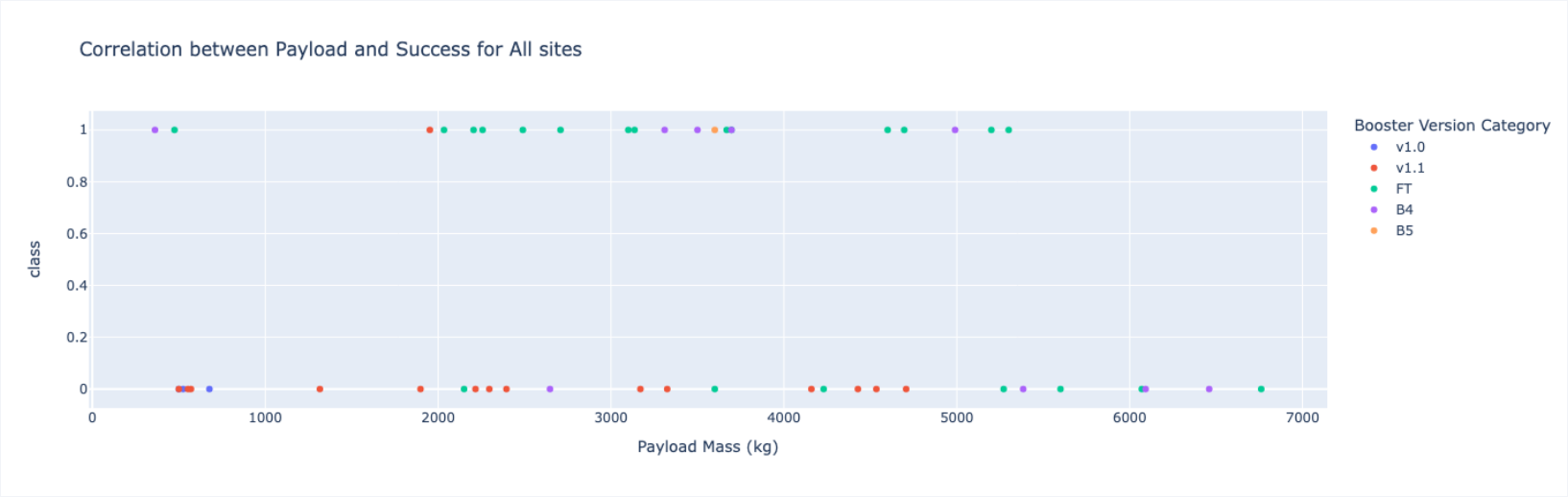
Total Success Launches by Site



Launch site: KSC LC-39A



Payload vs Launch Outcome



Lower payload

Higher payload

Section 5

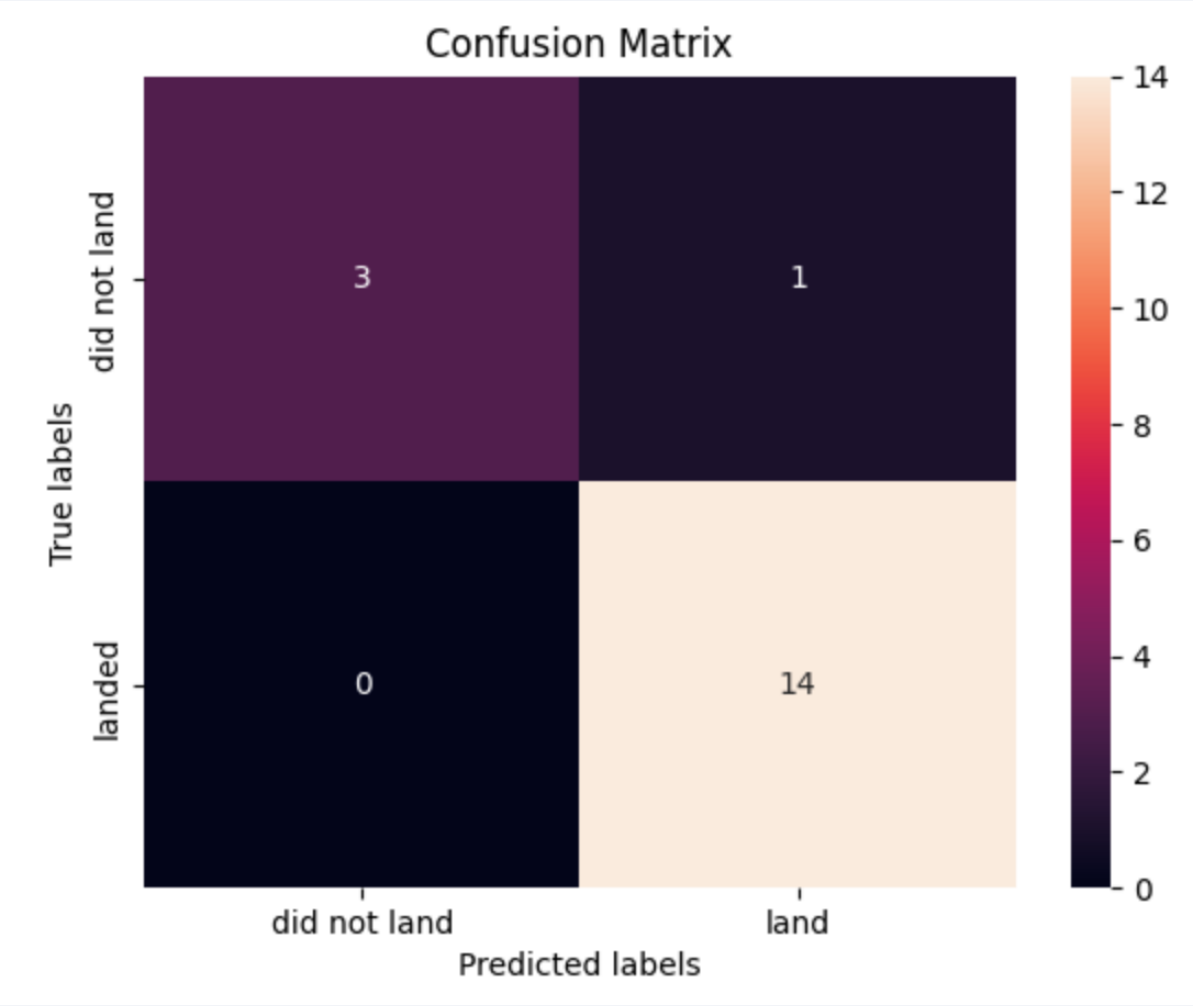
Predictive Analysis (Classification)

Classification Accuracy

```
Logistic Reg.'s Accuracy:  0.9444444444444444
SVM's Accuracy:  0.8888888888888888
DecisionTrees's Accuracy:  0.9444444444444444
KNN's Accuracy:  0.9444444444444444
```

The KNN, Decision Tree, and logistic regression models achieved the highest accuracy at 94%.

Confusion Matrix



This is the confusion matrix for the KNN model.

Conclusions

- Launch site KSC LC 39A has the most successful launches out of all sites
- Low weighted payloads have better success rates than higher weighted payloads
- The KNN, decision tree, and logistic regression models are the best in terms of prediction accuracy testing of the dataset

Thank you!