Winning Space Race with Data Science

Jatinkumar Patel

06/17/2023

Jatinpatel_77@hotmail.com

Outline

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

Executive Summary

- Summary of methodologies
  - Following methodologies were used to perform predictive analysis of successful rocket launches.
    - **Data Collection**: REST API, Web Scrapping
    - **Data Processing:** Transform data to have usable information for data analysis and visualization using pandas and SQL
    - **Data Analysis and Visualization:** Exploratory data analysis using Pandas, SQL, Scikit Learn, Folium, Plotly
    - **Build Model:** The predictive model was built based on a training data set using logistic regression, support vector machine, decision tree, and K-nearest neighbor and then analyzed using a test dataset.
- Summary of all results
  - There were 4 orbits that had a 100% success rate (SSO, HEO, GEO, ES-L1)
  - The landing site with the highest success rate was KSC LC 39A
  - Success rate was increased over time except for a year when it was reduced.
  - Decision tree model had the best performance out of all models.

Introduction

- SpaceX is a revolutionary company that has disrupted the space industry by offering rocket launches specifically Falcon 9 as low as 62 million dollars; while other providers cost upward of 165 million dollars each. Most of these savings are due to SpaceX's astounding idea to reuse the launch's first stage by re-landing the rocket to be used on the next mission. Repeating this process will make the price down even further. As a data scientist of a startup rivaling SpaceX, this project aims to create a machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

  In order to accomplish this, we will work on the following:

  - How payload mass, launch site, number of flights, and orbits affect first-stage landing success
  - Rate of successful landings over time
  - Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Data collected using REST API and web scrapping.
- Perform data wrangling: Extracted useful data in a new data frame and applied on hot encoding to convert data in a useful form.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

    Models were built using a training dataset and then it was evaluated for different parameters to find the best parameters using grid search. Once we had the best parameters available, models were further evaluated using a test dataset to see how it performs.

Data Collection

- Data Collection
  - SpaceX API
  - Web Scrapping

Data Collection – SpaceX API

- response = requests.get(spacex_url)
- data = response.json()
- data = pd.json_normalize(data)

Request SPACEX data using get request → decode content with JSON() → Convert data to the data frame using normalize → Filter data frame to keep useful content

- jupyter-labs-spacex-data-collection-api

df=pd.DataFrame(launch_dict)

Data Collection - Scraping

- response=requests.get(static_url).text
- soup = BeautifulSoup(response,"html.parser")
- html_tables = soup.find_all('table')
- first_launch_table = html_tables[2]

- jupyter-labs-webscraping

Request data using HTTP get → Create beautiful soup object → Extract HTML Table → Create data frame from HTML Table

Data Wrangling

- Replace missing values
- Use hot encoding to convert categorical dataset into numeric data for analyzing

- Data Wrangling

EDA with Data Visualization

- Scattered plots were used to understand various relationships.
- Bar graphs were used to compare various columns,

- Exploring and Preparing Data

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the 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 date when the first successful landing outcome in the ground pad was achieved.
- List the names of the boosters which have success in drone ships and 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 records which will display the month names, failure landing_outcomes in drone ship , booster versions, and launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- SQL Notebook

Build an Interactive Map with Folium

- ## Map with Folium
- **Markers Indicating Launch Sites**
  - Added a blue circle at NASA Johnson Space Center's coordinate with a popup label
  - Added red circles at all launch sites coordinates with a popup label
- **Colored Markers of Launch Outcomes**
  - Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- **Distances Between a Launch Site to Proximities**
  - Added colored lines to show distance between launch site CCAFS SLC 40 and its proximity to the nearest coastline, railway, highway, and city
- Launch Sites Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- **Dropdown List with Launch Site Names**
  - Allow user to select all launch sites or a certain launch site
- **Slider of Payload Mass**
  - Allow user to select payload mass range
- **Pie Chart Showing Successful Launches**
  - Allow users to see successful and unsuccessful launches as a percent of the total
- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**
  - Allow user to see the correlation between Payload and Launch Success
- Plotly

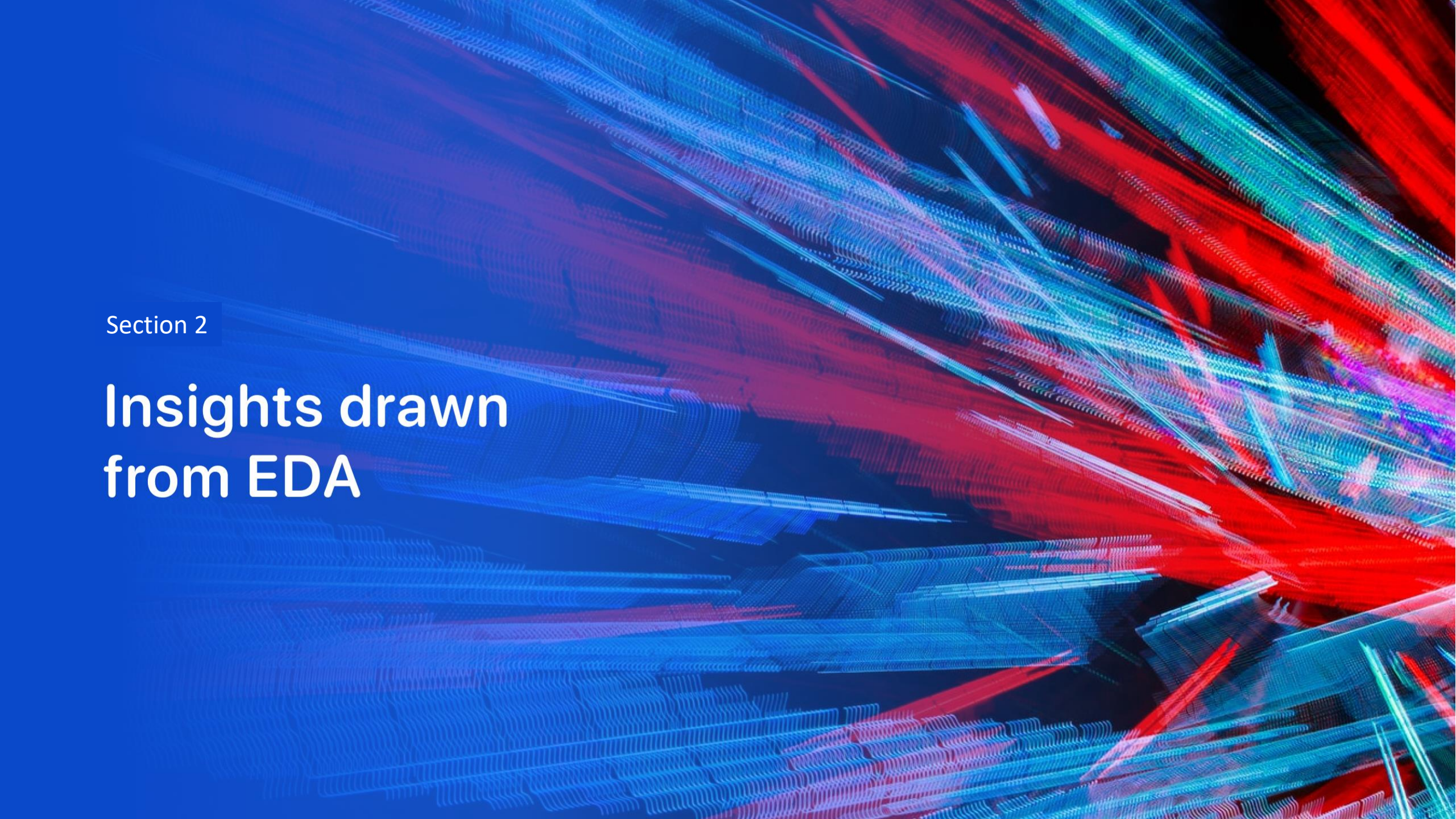
Predictive Analysis (Classification)

- Create a NumPy array
- Standardize the data using a standard scaler
- Use the function train_test_split to split the data into training and test data
- Create various machine learning objects and GridSearchCV object
- Fit the ML objects to find the best parameters from the dictionary parameters
- Calculate the accuracy of the test data using the method score for each ML object.
- Find the best ML object that performed best on test data.

- Machine Learning Prediction

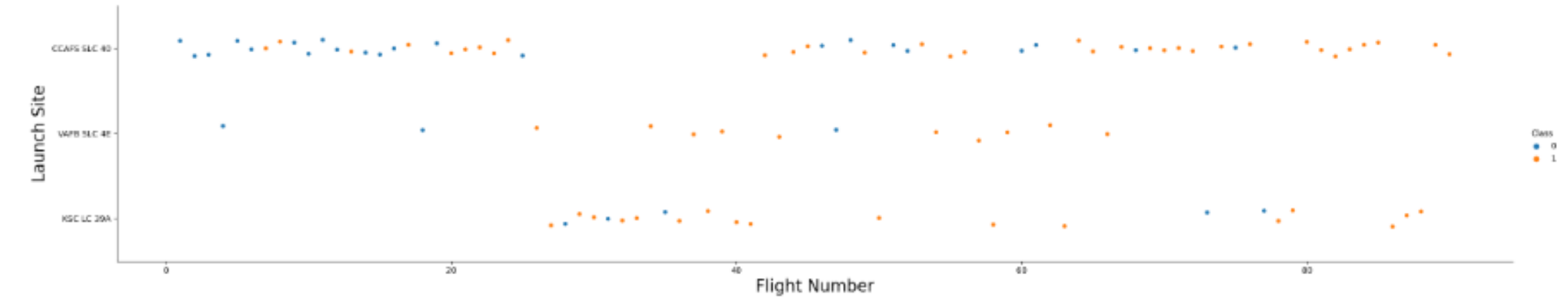
Results

- Exploratory data analysis results
  - 4 Orbits called Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.
  - The landing site with the highest success rate is KSC LC-39A.
  - Over time launch success has improved.
- Interactive analytics demo in screenshots
  - Most launch sites are near the equator, and all are close to the coast
- Predictive analysis results
  - Decision Tree performs best for this dataset


Section 2
Insights drawn from EDA

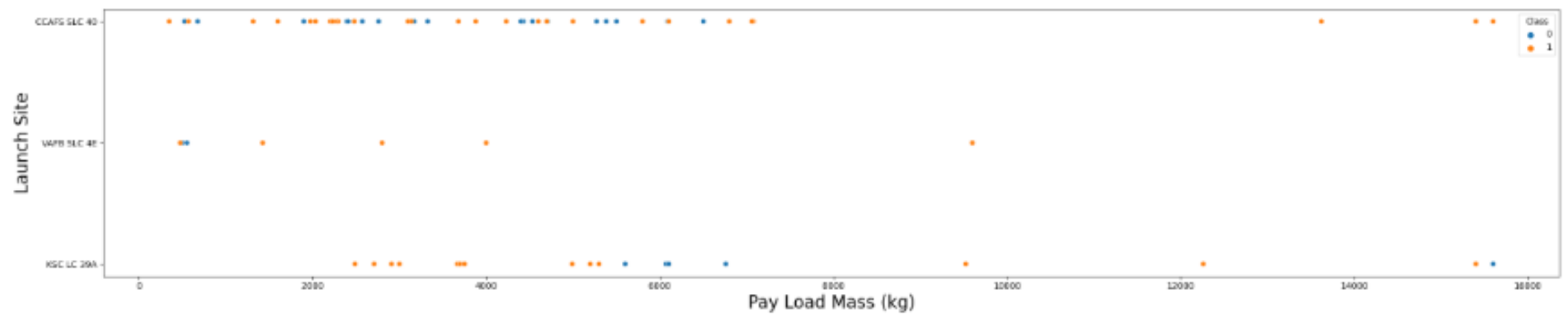
Flight Number vs. Launch Site

- It’s possible to verify that the best launch site nowadays is CCAF5 SLC 40.
- Around half of the launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- New launches have a higher success rate



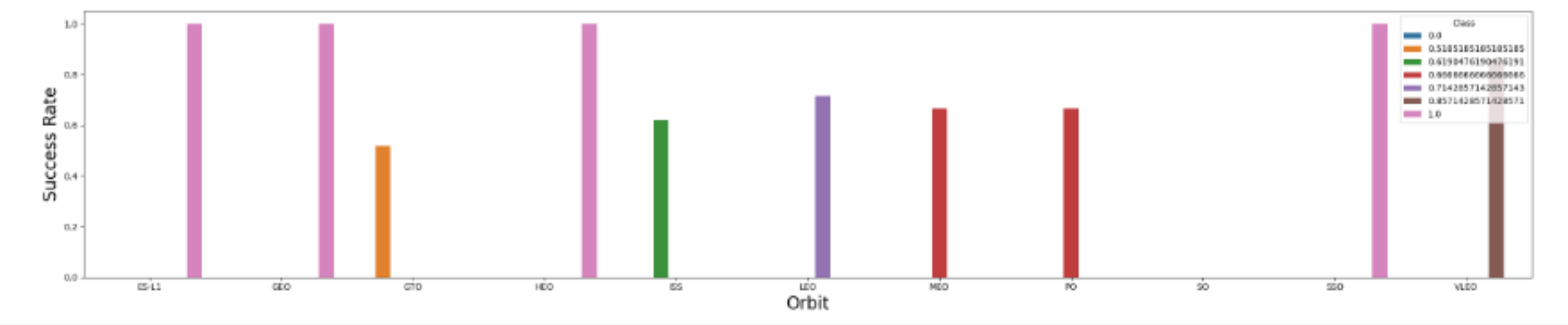
Payload vs. Launch Site

- Higher the payload mass, the higher the success rate
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than 10,000 kg



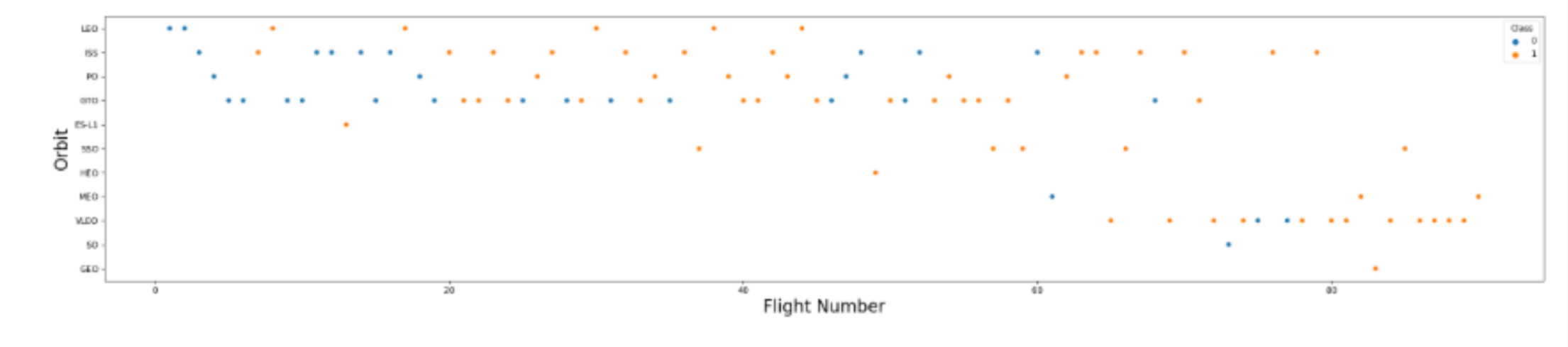
Success Rate vs. Orbit Type

- **Exploratory Data Analysis**
- ES-L1, GEO, HEO, and SSO had a 100% success rate
- GTO, ISS, LEO, MEO, PO had a success rate in the range of 50% to 100%
- SO has not achieved any success yet.



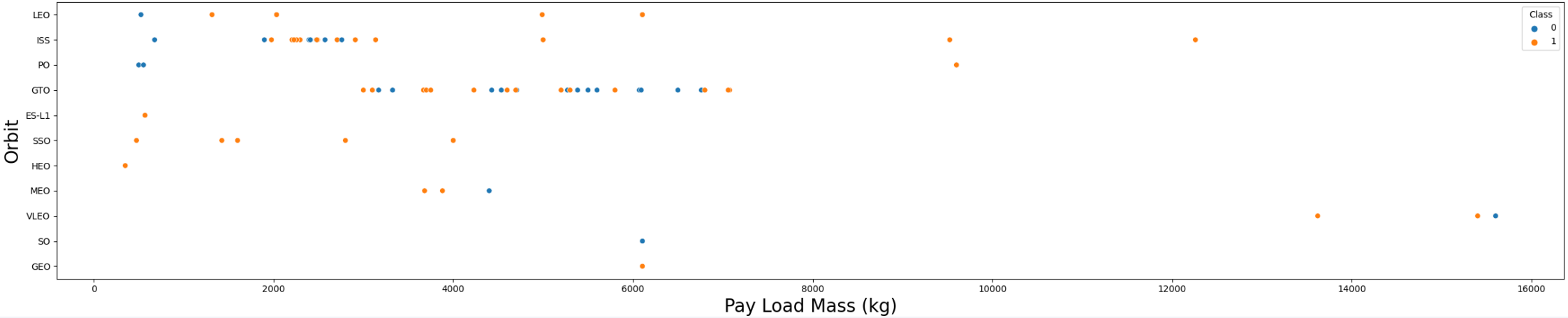
Flight Number vs. Orbit Type

- LEO orbits 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.



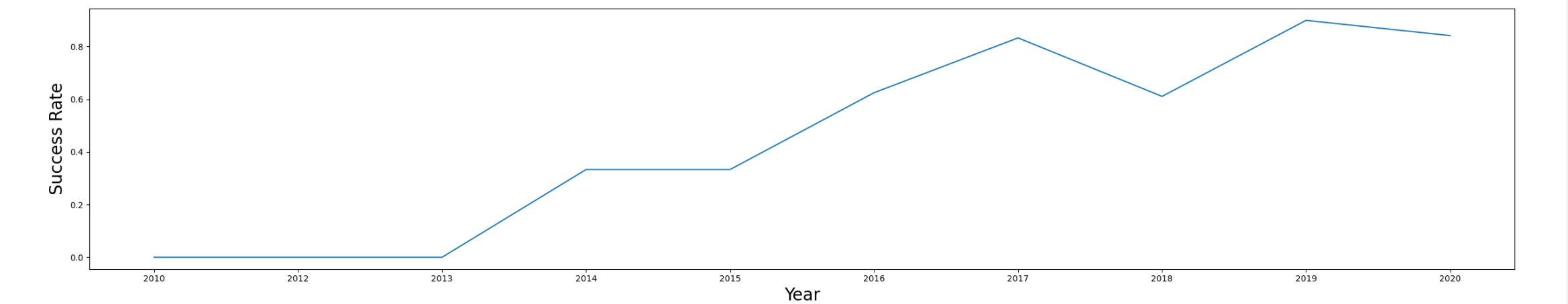
Payload vs. Orbit Type

- With heavy payloads the successful landing rate is 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 there here.



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

In [9]:
```
%sql select LAUNCH_SITE from SPACEXTBL group by LAUNCH_SITE
```

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

Out[9]:

Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

In [13]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where "Customer"="NASA (CRS)"
```

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

Out[13]:

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

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

In [14]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" like "F9 v1.1"
```

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

Out[14]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [15]:
```
%sql select min("Date") from SPACEXTBL where "Landing_Outcome" like "Success"
```

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

Out[15]:

min("Date")
01/07/2020

Successful Drone Ship Landing with Payload between 4000 and 6000

In [16]:
```
%sql select "Booster_Version" from SPACEXTBL where "Landing_Outcome" like "Success (drone ship)" and 4000<PAYLOAD_MASS__KG_
```

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

Out[16]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [17]:
```
%sql select count("Mission_Outcome") from SPACEXTBL
```

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

Out[17]:

count("Mission_Outcome")
101

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [19]:
```
%sql select "Booster_Version" from SPACEXTBL where PAYLOAD_MASS__KG_ = (Select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[19]:

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

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql select substr(Date,4,2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTBL where "Landing_Ou
```

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

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

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[15]:
```
%sql SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCOME) AS "Total Count" FROM SPACEXTBL \
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' \
GROUP BY LANDING_OUTCOME \
ORDER BY COUNT(LANDING_OUTCOME) DESC ;
```

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

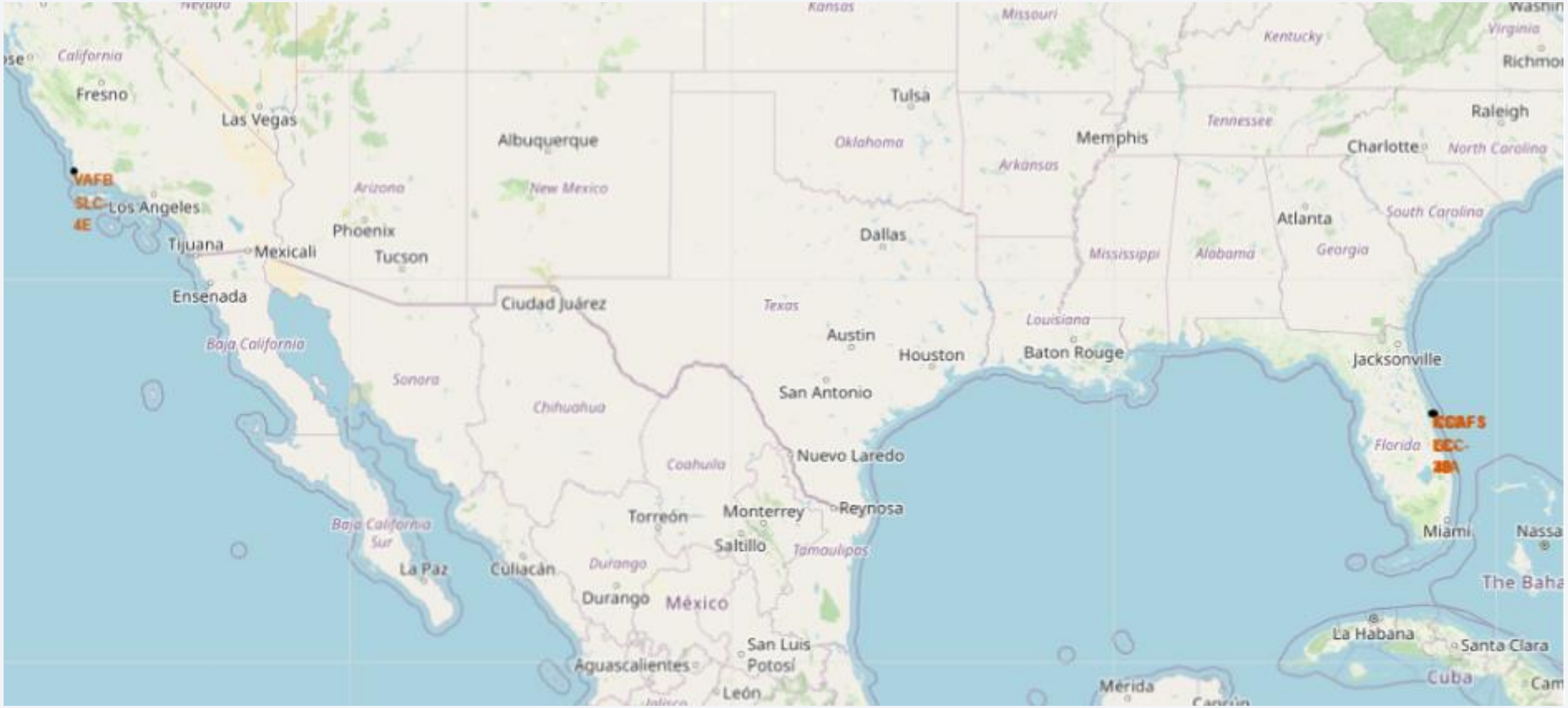
[15]:

Landing Outcome	Total Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

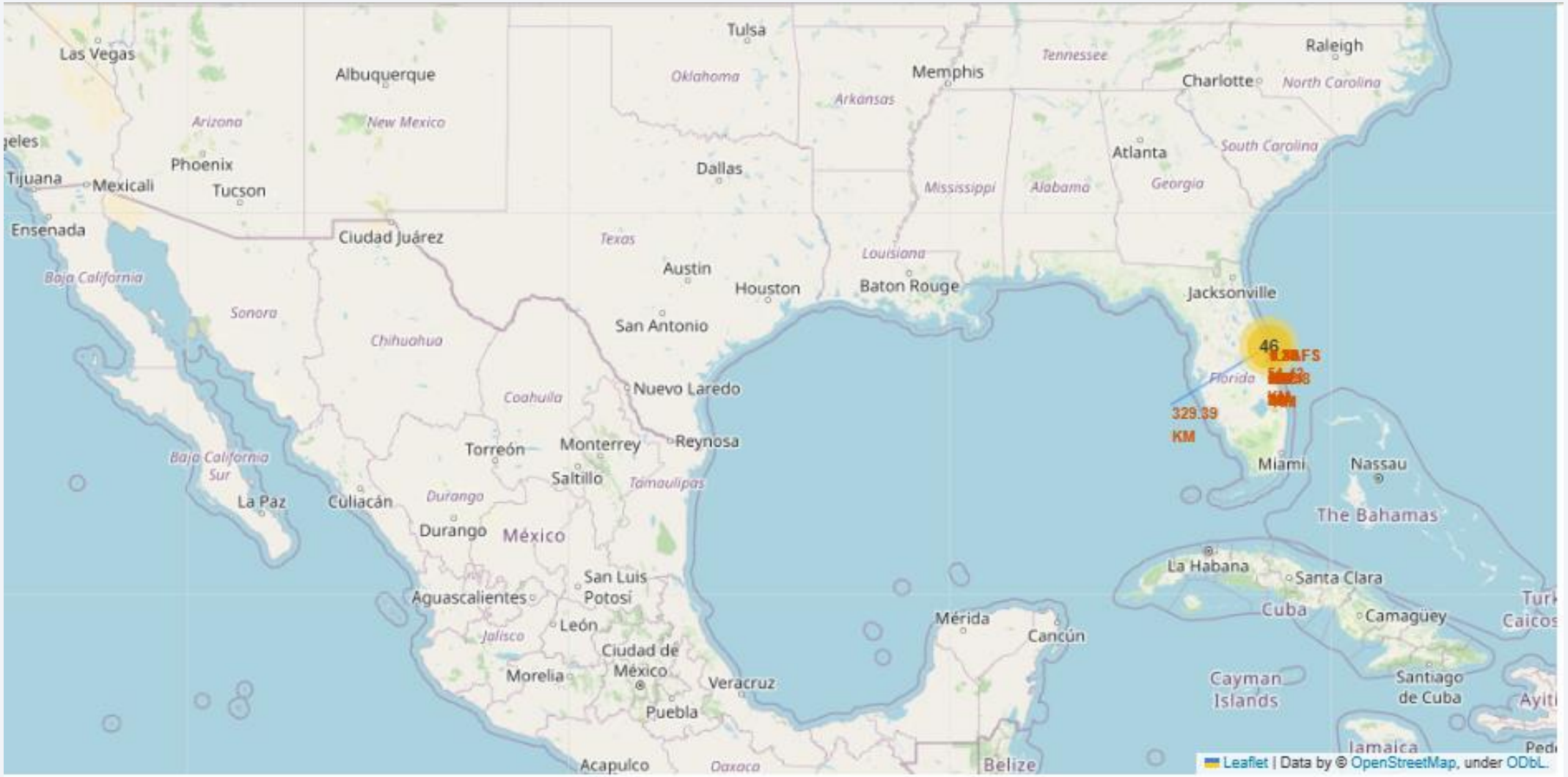
Launch Sites Proximities Analysis

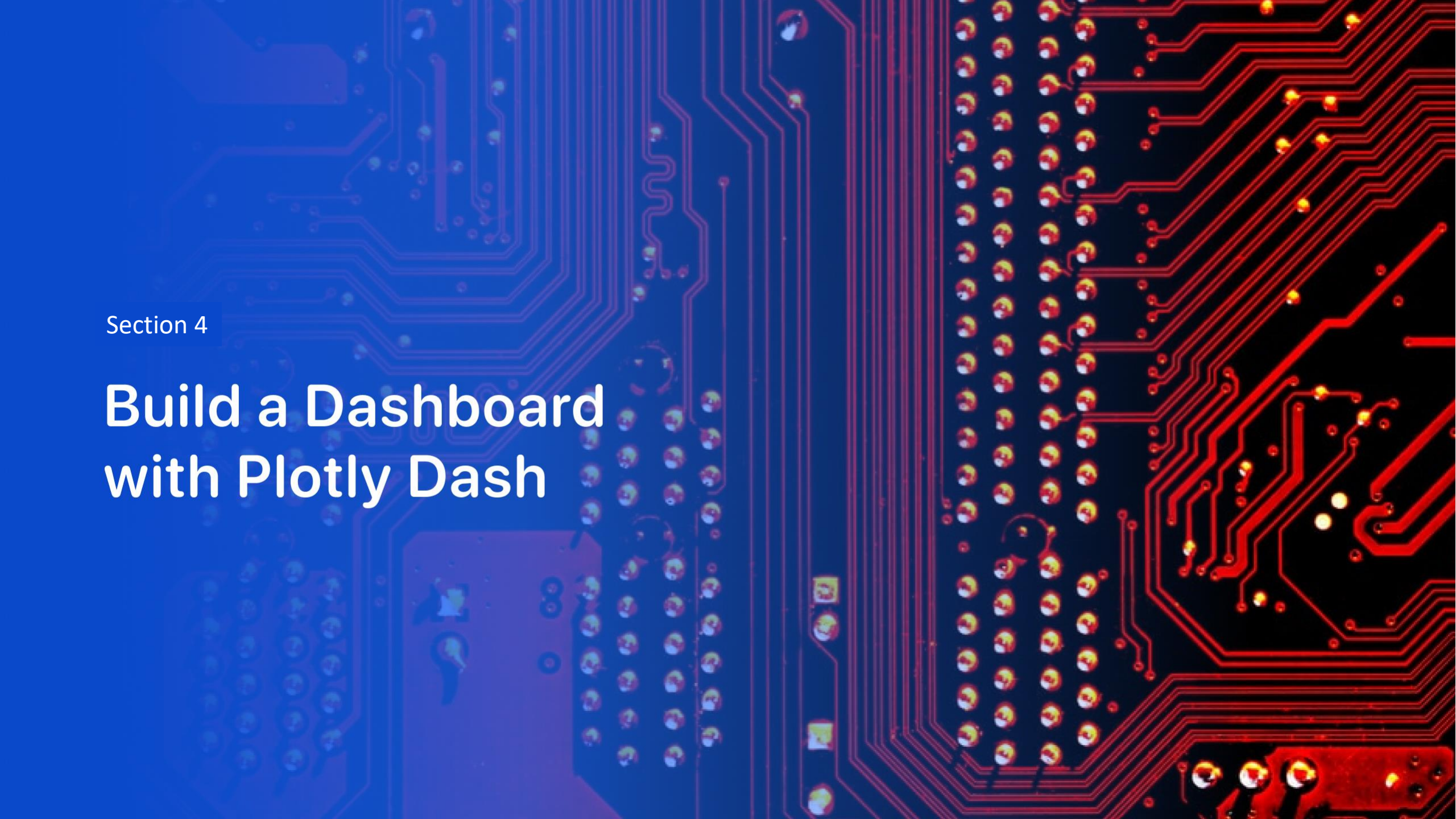
SpaceX Launch Records – All Sites



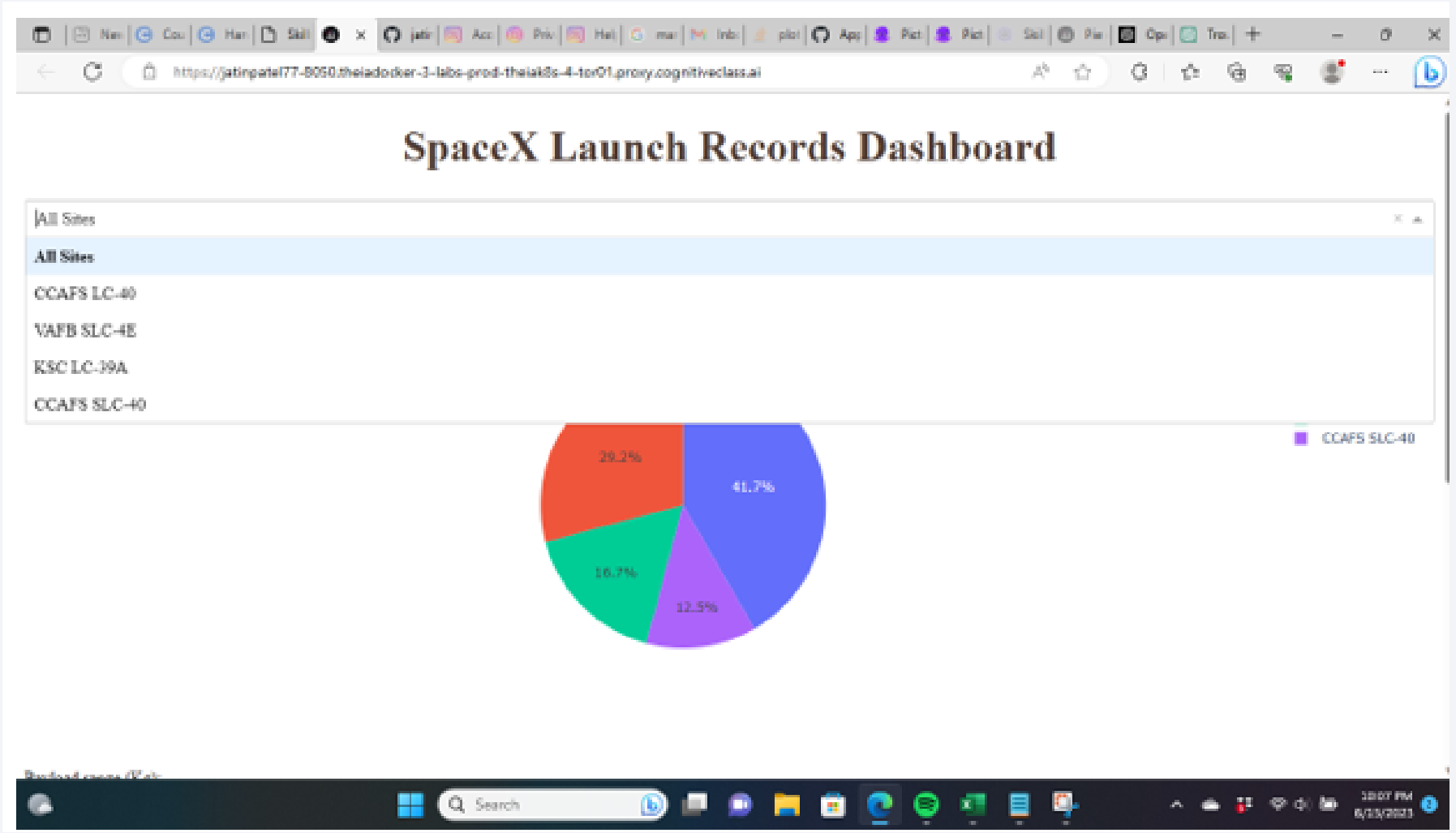
Successful outcomes

Distance from Launch Site

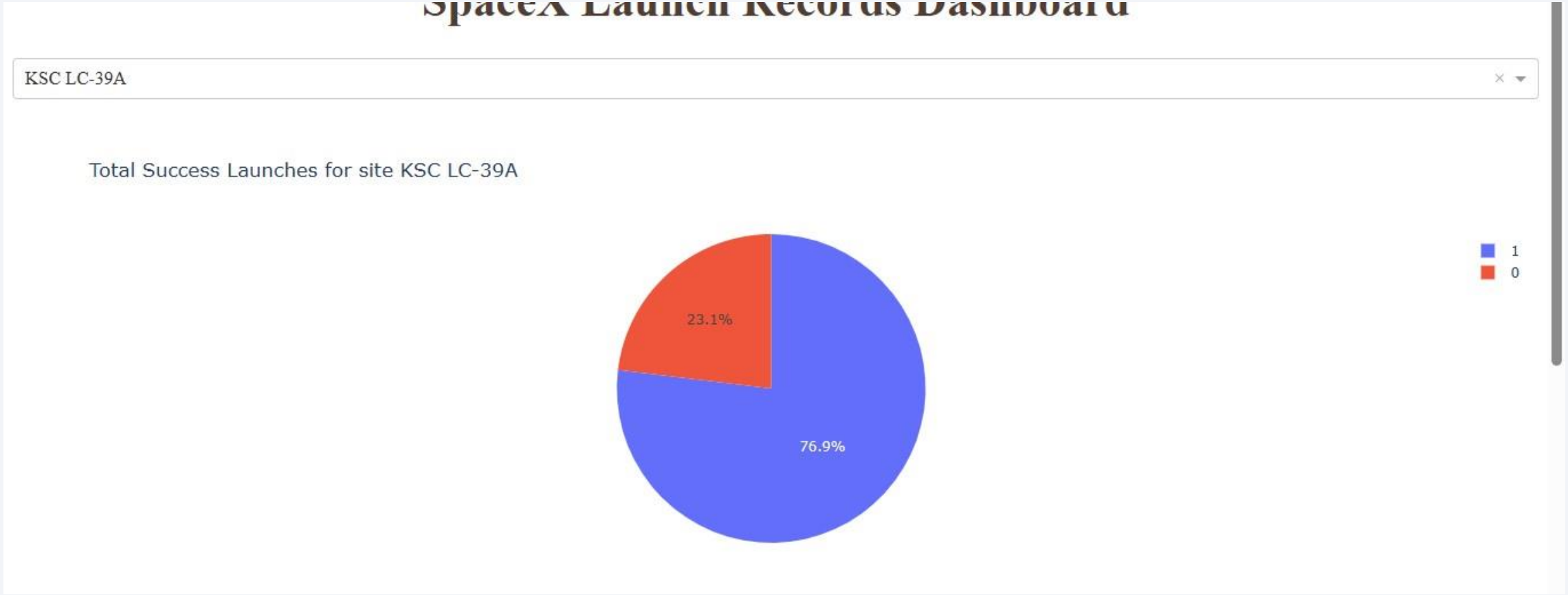



Section 4
Build a Dashboard with Plotly Dash

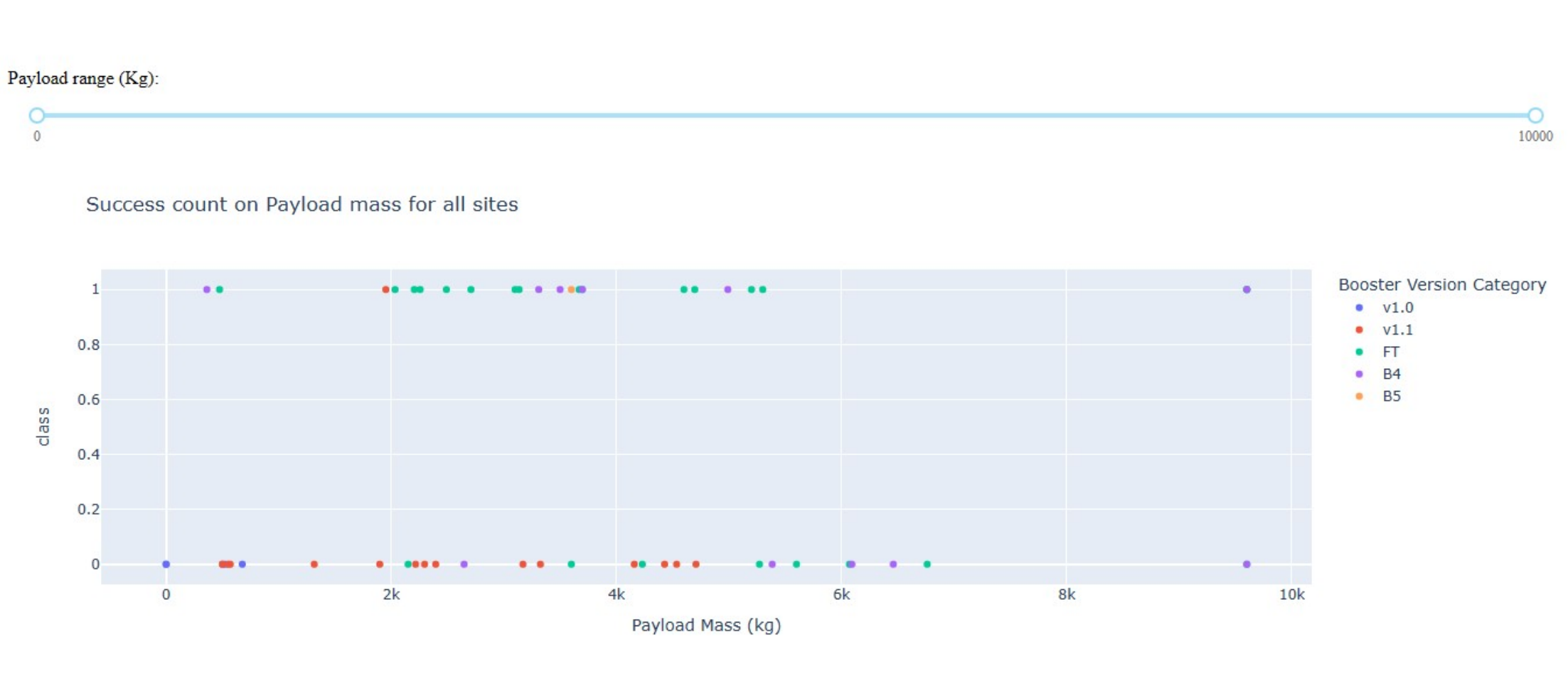
Launch success count for all sites



Launch site with highest success rate



Payload vs. Launch Outcome scatter plot for all sites

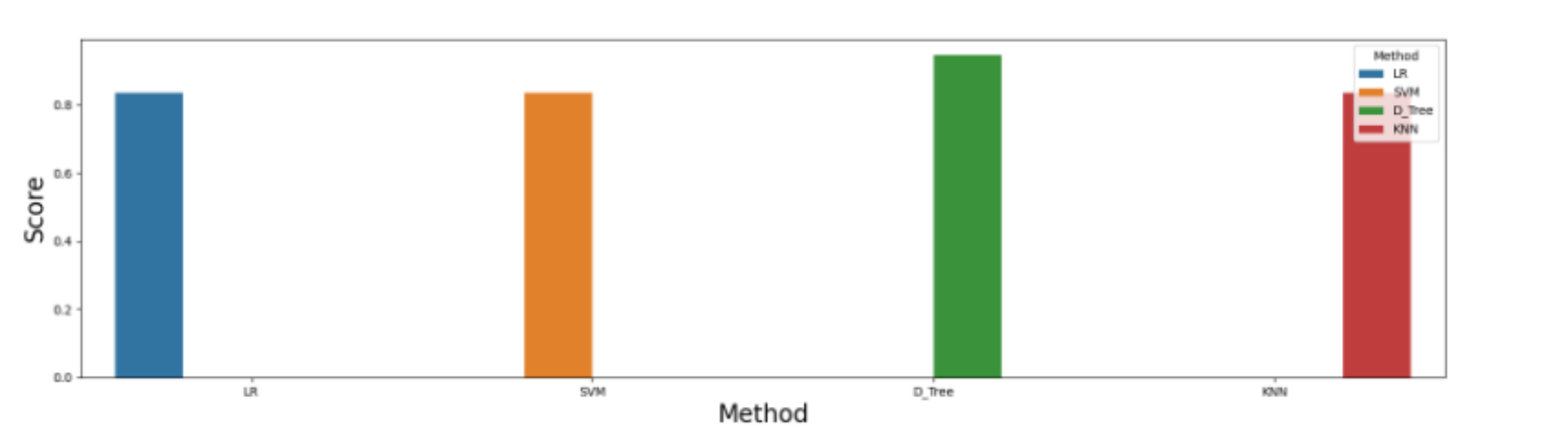


Section 5

Predictive Analysis (Classification)

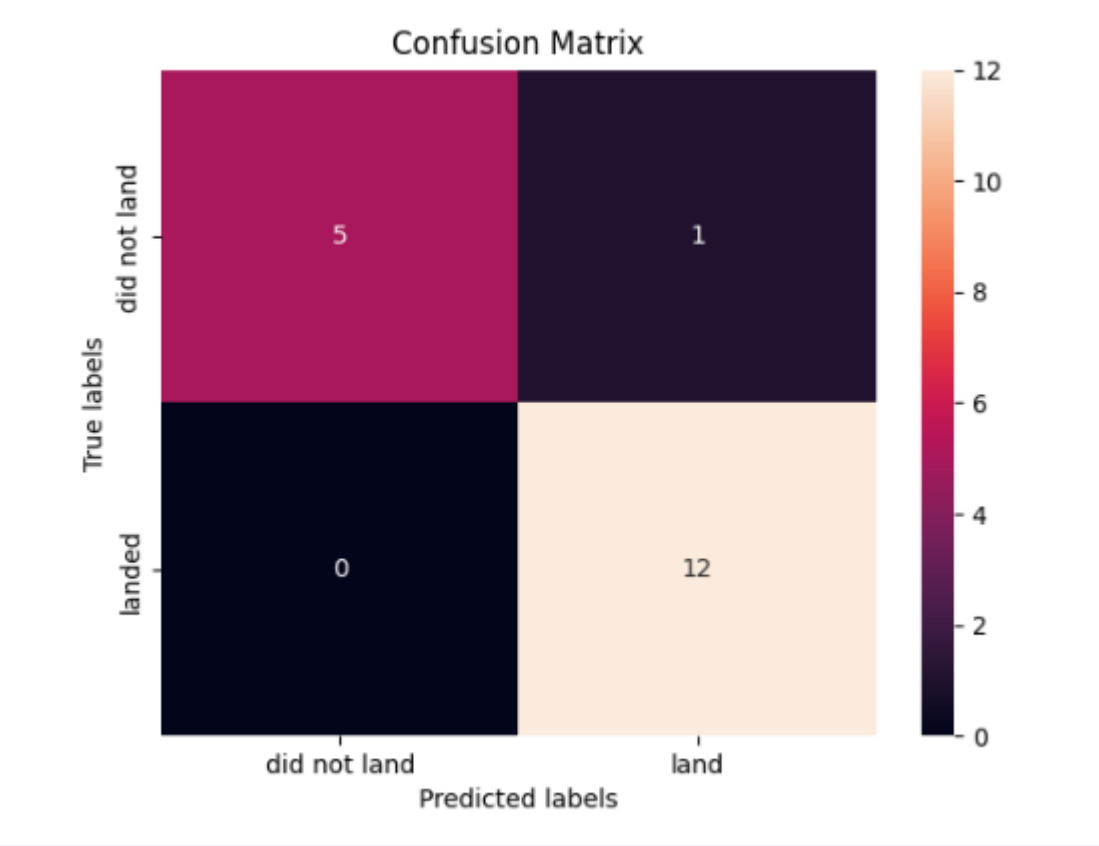
Classification Accuracy

- Decision tree has the highest accuracy



Confusion Matrix

- Decision Tree



Conclusions

- Decision Tree provided the best accuracy on test data, it should be considered to use on real-time data.
- With heavy payloads the successful landing rate is more for Polar, LEO, and ISS.
- In general, the Higher the payload mass, the higher the success rate
- Launch sites VAFB SLC 4E and KSC LC 39A have higher success rates, a study is recommended to understand the reasons for higher success rates to see if those differences can be applied at other sites to improve chances of success at other sites as well.

Appendix

- All Noebooks

Thank you!