IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

<JAVID KAMAL>
<28-07-2023>

# Outline

# Executive Summary

- **Summary of methodologies:**
- **Data collection**: Readable data deriving from the SpaceX API web server data and converting it into a Pandas dataframe and also from Wikipedia by the process of web scraping and converting it into CSV file using beautiful soup.
- **Data wrangling:** Taking care of NAN values and outliers and substituting it with the mean values of the data using the statistical approach.
- **Exploratory data analysis:** EDA using SQL queries and graphical visualization of the data using Matplotlib.
- **Interactive data analysis:** Alleviating and leveraging interactive data analysis responding to varying inputs from the user or the stake holders using Folium and Plotly Dash.
- **Predictive data analysis**: Availing various classification models and arriving at the best model to predict the target, given the input data using Machine learning algorithms.

# Executive Summary

- **Summary of all results:**
- **Recovery of the SpaceX rocket first stage and its maximum success rate:** It is deduced from the data analysis that SpaceX – Falcon9 is very successful in the recovery when it has the following characteristics:
- Minimum payload i.e. between 2000 kg and 4000 kg accounts for its maximum success.
- The success story manifests itself from 2013 AC onwards and is exponentially increasing and has the maximum success between 2017 AC and 2020 AC.
- Out of all the four launch sites including CCA SLC and CCA LC or say VAFB or KSC LC, the highest success rate is for KSC LC 39A.
- Out of various landing sites available like Ocean landing, ground pad landing or drone ship landing, drone ship landing wins the race of the maximum success.

# Introduction

- **Project background and context:**
- SpaceX, a rocket launching firm attracting many customers towards it based on a competitive strategy of cutting down the space tour to $62 million with its falcon9 when compared to its rivals which costs around $165 million.
- The objective of the project is to find out the probability of successful rocket's first stage recovery, given that it is the most expensive part of the rocket launch and is reusable.
- **Problems you want to find answers:**
- Best launch site, optimum payload and the landing scenario that results in the maximum success rate of the first stage recovery of the rocket so that the company banks and put emphasis on these features to maximize the success rate.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Availing SpaceX REST API with the web server, reading the JSON file which is much like a dictionary and get the data and converting the data into a readable Pandas data frame enabling us to perform manipulation operations on it.
  - Availing Wiki article data which is the html type and using web scraping tool called Beautiful soup which makes use of requests, get and response methods and the data is converted in the Pandas data frame and eventually to CSV file.
- Perform data wrangling
  - Raw data was analyzed for null values and unusual and suspicious values which the data frame contains and replaced by the mean values derived from statistical libraries.

# Methodology

- **Perform exploratory data analysis (EDA) using visualization and SQL:**
- Matplotlib tool was used to plot line charts, scatter plots for EDA visualization giving an appealing graphical representation of the data features an enhanced effect of data reality to the stake holders.
- SQL queries were used to establish and isolate the desired results as per the requirements of the project and further exploring the data and deriving the insights from the data.

# Methodology

- **Perform interactive visual analytics using Folium and Plotly Dash:**
- Folium library was used for mapping terrain, especially when it comes to finding out the distances from one point to the other, given the varying latitude and longitude values. Instant outputs based on the then entered inputs, however fluctuating the parameters are, Folium with the help of markers yields the map and accurate results, for example, minimum distance of the launch site from the coastline or even the successful launches versus the unsuccessful ones.
- Plotly dash signifies and projects the most stunning pie charts, along with range sliders, given the selection in the dropdown list boxes, for example. Different launch sites and varying payloads were considered in the project.

# Methodology

- Perform predictive analysis using classification models:
  - Out of the striking machine learning algorithms, the most relevant classification models such as logistic regression, KNN, decision tree model and SVM were considered for predictive data analysis. It is deciphered that most of the models yield 83% results accuracy on the test data. The best parameters were found out for each model respectively and the score was tallied against each other to uncover the maximum score accuracy to find out the best classification model for the project. The data was split 80% for training and 20% for test data and the score was estimated for the test data after having gone through various phases of GridSearchCV object creation, normalizing, transforming and fitting the training data on individual classification models.

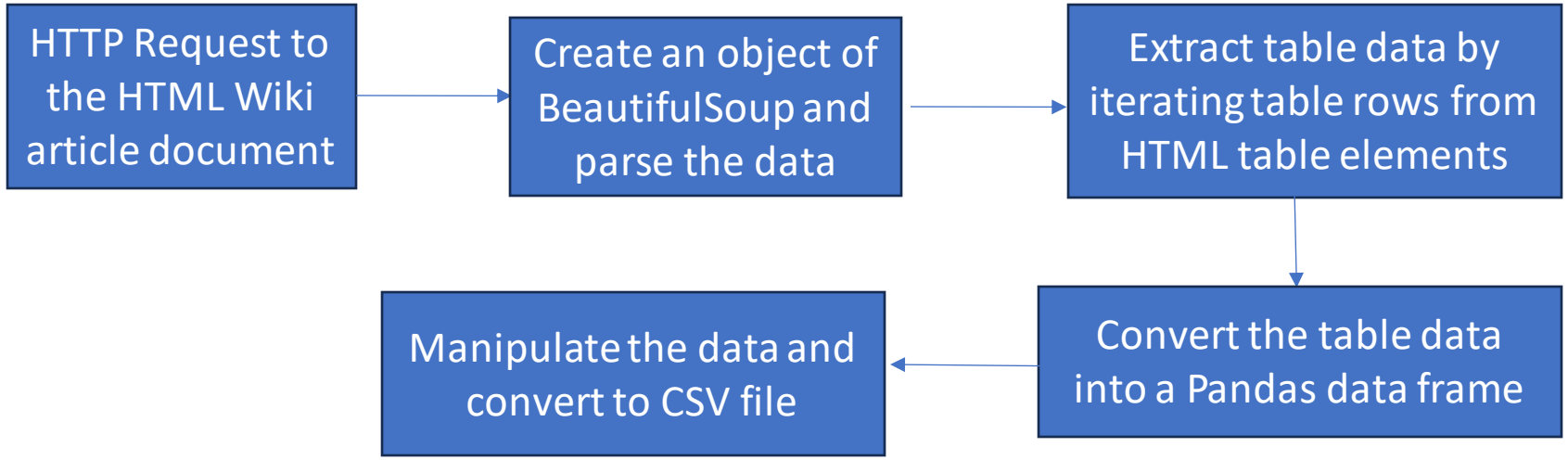
# Data Collection

- There are two ways in which the data can be collected:
- SpaceX web server through REST API, in which the client sends a HTTP request and the server responds with a JSON data format, a dictionary with the status code of 200 but only when it is a successful connection between the client and the server. The data is then stored in an individually appended list and finally it is converted into pandas data frame, a readable structured data format ready to be manipulated, eventually to a CSV file.
- Web scraping with Beautiful soup library of bs4 package is another way of extracting the raw data from wiki article on the internet dealing with SpaceX. It is a HTML document from which table data is extracted by iterating through the table rows from the table elements, which is converted to Pandas data frame to which several manipulating operations are done and then converted to CSV file.

# Data Collection – SpaceX API


SpaceX API

# Data Collection – Web Scraping



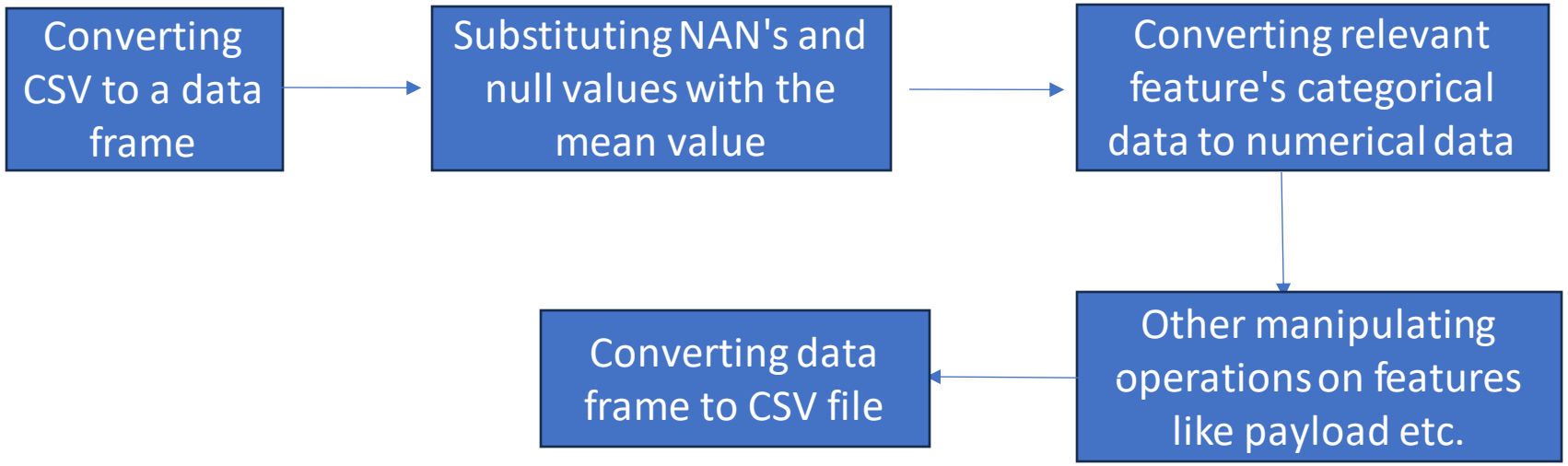



Web Scraping

# Data Wrangling

- Data wrangling is a process of making the raw data originating from the data frame in a more understandable form so that insights can be drawn from the data of relevance. In this project, we have pondered on the values of all features of the Falcon9 and care was taken to replace the NAN's and null values by the mean value of that particular feature which happens to be the column of the data frame. There will be a scenario in which some of the values will be unusually large or small making it to be indigestible for that particular feature and it will also be replaced by the mean value. Sometimes, mode or median can also be used instead of mean values.
- Converting some of the relevant features of the data frame from categorical variable to numerical variable will become handy prior to applying the machine learning algorithms, especially the classification models which will not run on, say, Boolean variables but can implement their methods on the numerical data which splits the columns retaining the data intact..

# Data Wrangling






Data Wrangling

# EDA with Data Visualization

- Matplotlib tool was used to depict various charts including but not restricted to scatter plots to arrive at the conclusion if the mission outcome is successful or not. The following features of Falcon9 were considered for EDA visualization to determine the scenario, specific set of features of relevance in which the first stage recovery was successful:
- Launch site including CCAC SLC, VAFB, KSC SLC etc. versus Flight number.
- Payload range to maximize the success rate versus Flight number.
- Orbit type such as LEO etc. , an ideal orbit type versus payload.
- Yearly trend which has kept increasing since 2013 AC and has maximized since 2017 AC until 2020 AC versus mission or landing outcome.
- Landing type such as drone ship, ground pad etc. versus flight number.

EDA with Data Visualization

# EDA with SQL

- **SQL Queries performed:**
- First successful drone ship landing.
- Booster versions with successful drone ship landing having payload mass between 4000 kg and 6000 kg
- Total number of both successful and unsuccessful landing or mission outcomes grouped by Launch sites.
- Launch sites and booster versions carrying maximum payload and were having successful landing outcome.

EDA with SQL

# Build an Interactive Map with Folium

- The map objects such as markers, circles, lines, etc. were created and added to a interactive folium map for the following reasons:
- Circles: The circles indicate the launch site location.
- Markers: There were two types of markers, one green marker which signifies the successful outcome for a given launch site and a red marker indicating that it is a failure.
- Lines: There is a blue colored line which defines the distance between amenities such as coastline, highway, railway or a closest city.
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose



Interactive Map with Folium

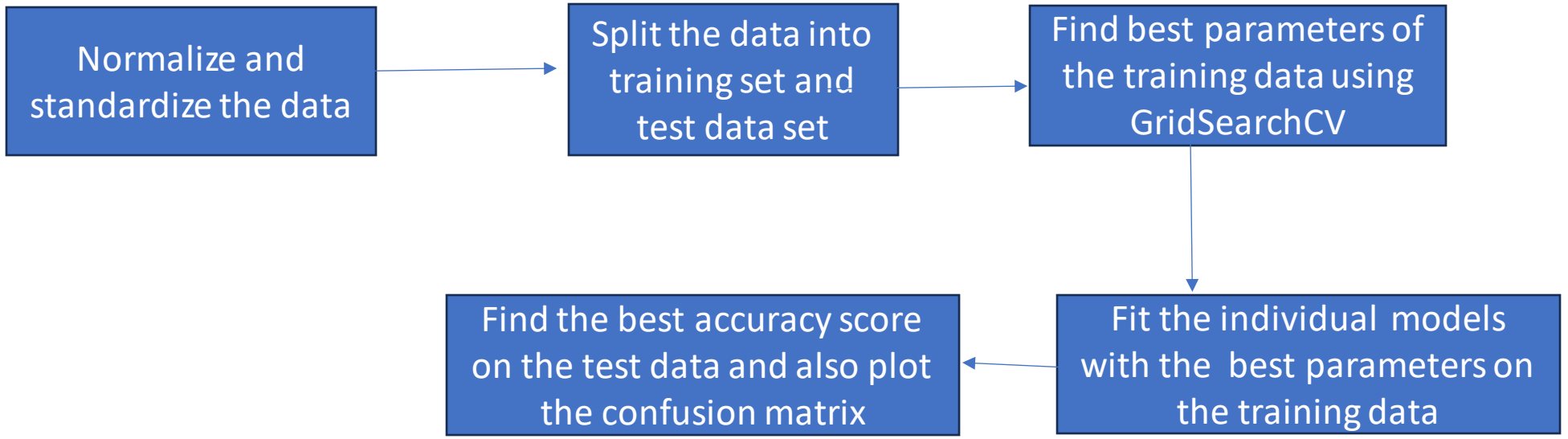
# Build a Dashboard with Plotly Dash

- The plots/graphs and interactions which have been added to a dashboard are mainly Pie charts which designate the instantaneous variation in the launch site and reflecting the change in the appearance of the pie chart. The success rate is hence determined visually.
- There was also a range slider which reflects the payload variation and which affects the look of pie chart resulting in varying success rate.
- There is a dropdown list box which consists of all launch sites individually and any of the launch sites can be selected from it or ALL sites can be selected from the list.
- The payload range slider has a varying payload mass between 0 kg and 10, 000 kg, which alters the pie chart according to the selected payload mass.

# Predictive Analysis (Classification)

- The four relevant and applicable classification models were built, evaluated, improved, and the best performing classification model was found.
- The classification models used are logistic regression, support vector machine, k nearest neighbor and decision tree model.
- The data set was evaluated first by normalizing and standardizing it. The data was then then split into training and test data with train_test_split method 80% training data and 20% test data.
- The various classification models were trained for the trained data by finding the best parameters derived from GridSearchCV algorithm of Sci-kit learn library and applying it to train the training data by fitting on it.
- The best accuracy score amongst the various classification models was determined thus when applying them on the test data and the most efficient model is applied to the test data.

# Predictive Analysis (Classification)

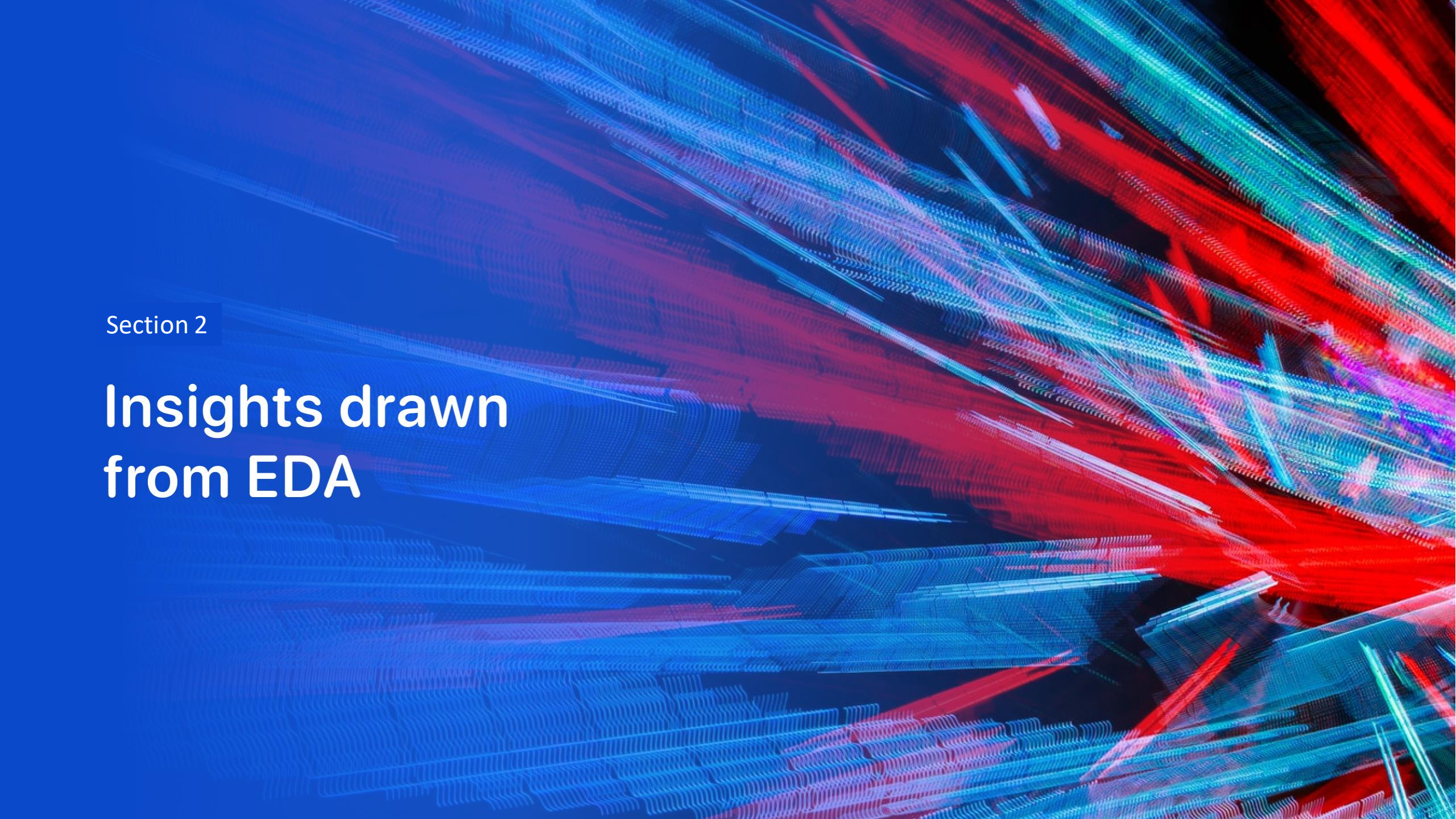


# Results

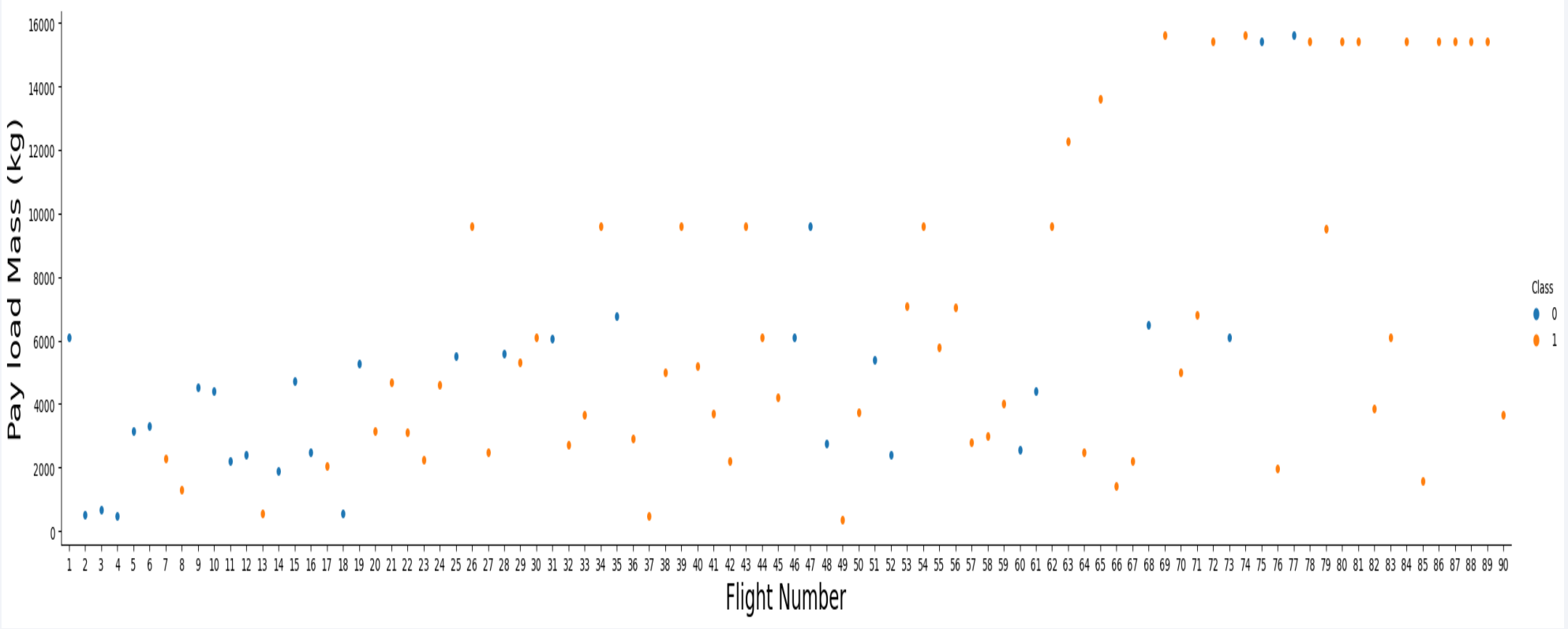
- **Exploratory data analysis results:**
- It is observed that there is a maximum rate of success i.e. first stage recovery of the Falcon9 when the following characteristics are implemented:
- Payload mass between 2000 kg and 4000 kg.
- Launch site being KSC SLC 39A.
- Landing scenario is drone ship.
- Minimum date of launch is 2017 AC as the yearly trend keeps on increasing since the 2013 AC mark and is maximum success frate after 2017 AC
- **Interactive analytics demo in screenshots:** Folim Map and Plotly Dash is used for the data analytics and the results of pie charts and maps follows this section.
- **Predictive analysis results:** It is estimated that the maximum score accuracy score amongst all classification models is 83.33 % and any of them can be used here.

Section 2

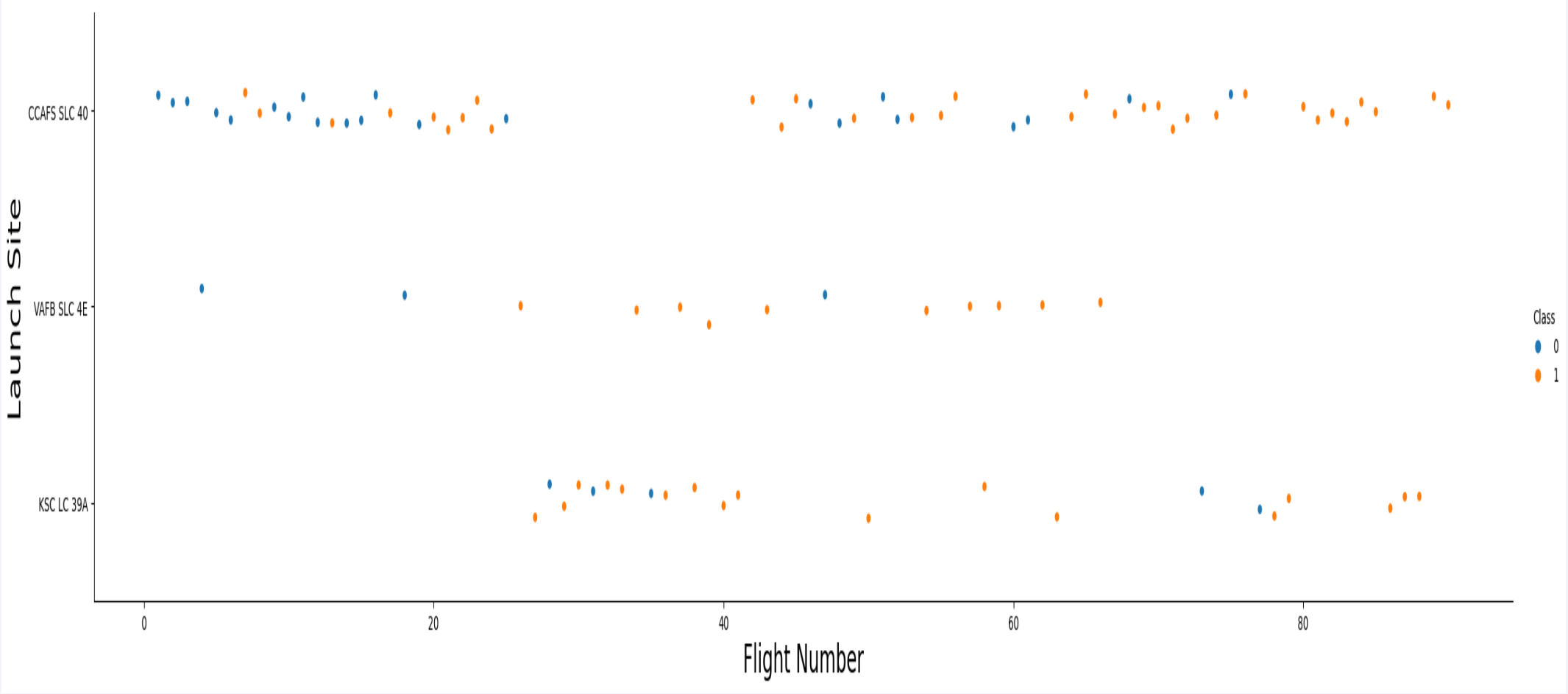
# Insights drawn from EDA



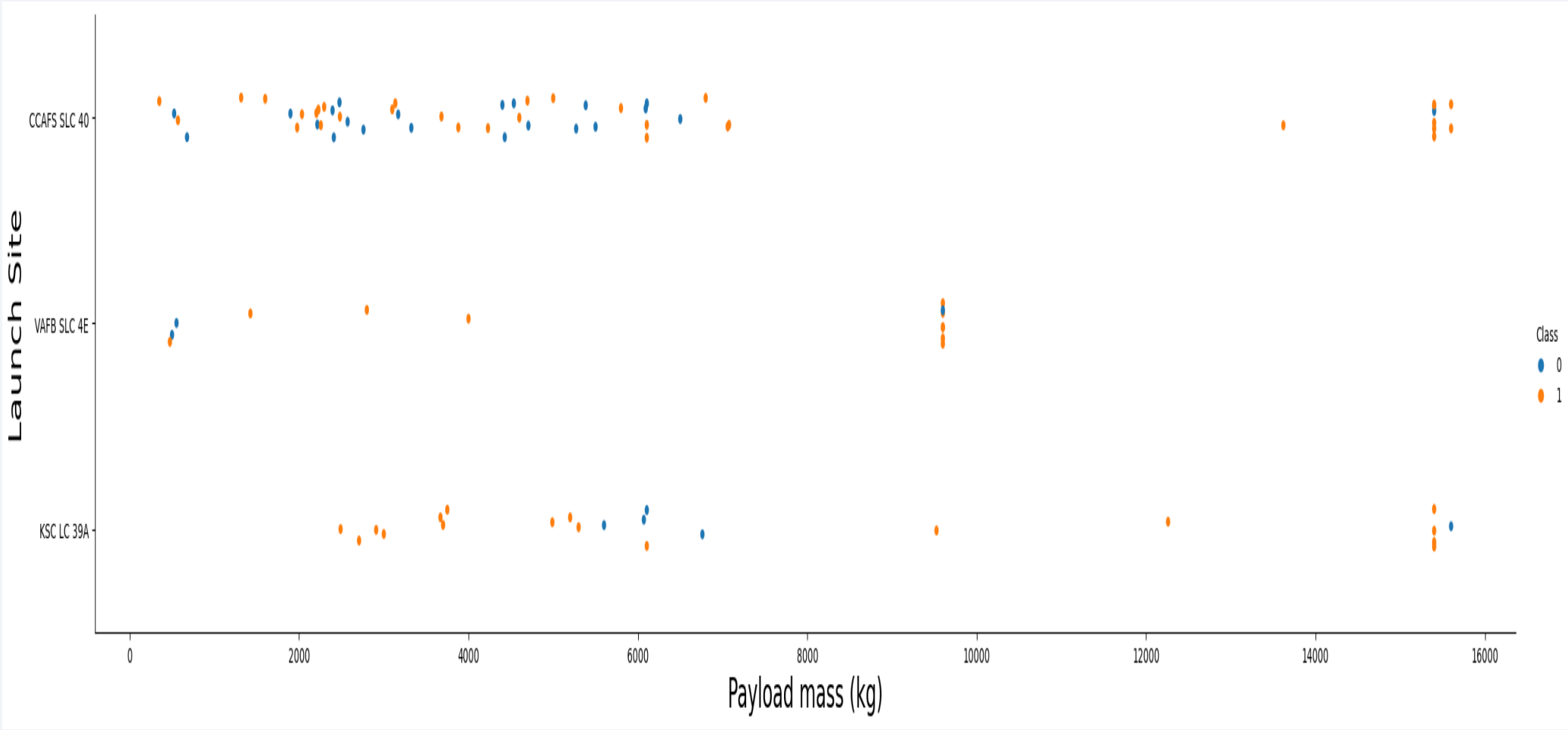
# Flight Number vs. Payload Mass



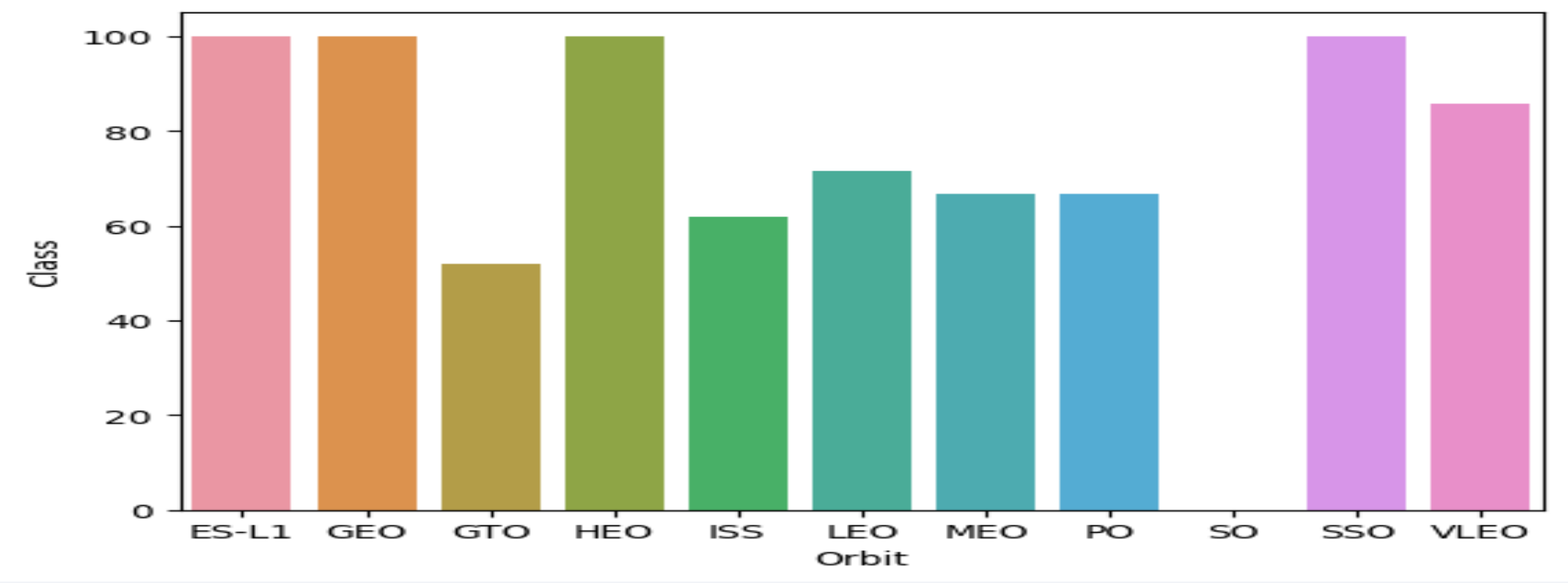
# Flight Number vs. Launch Site



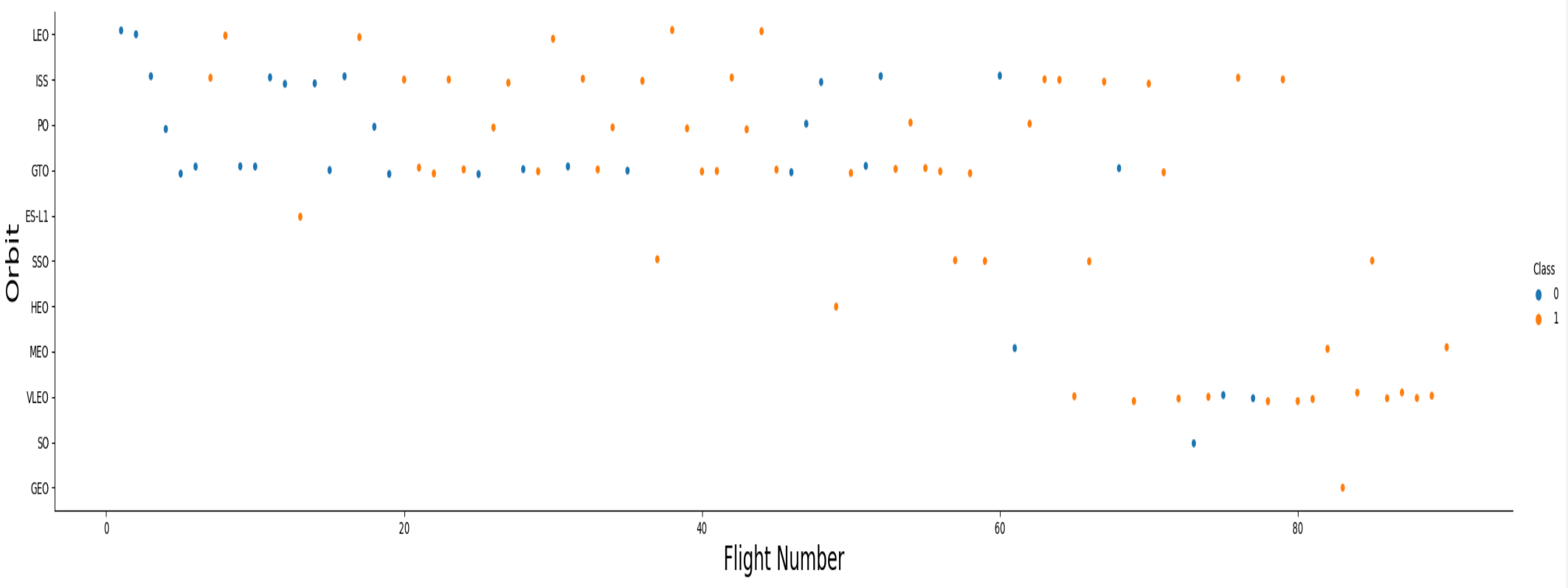
# Payload Mass vs. Launch Site



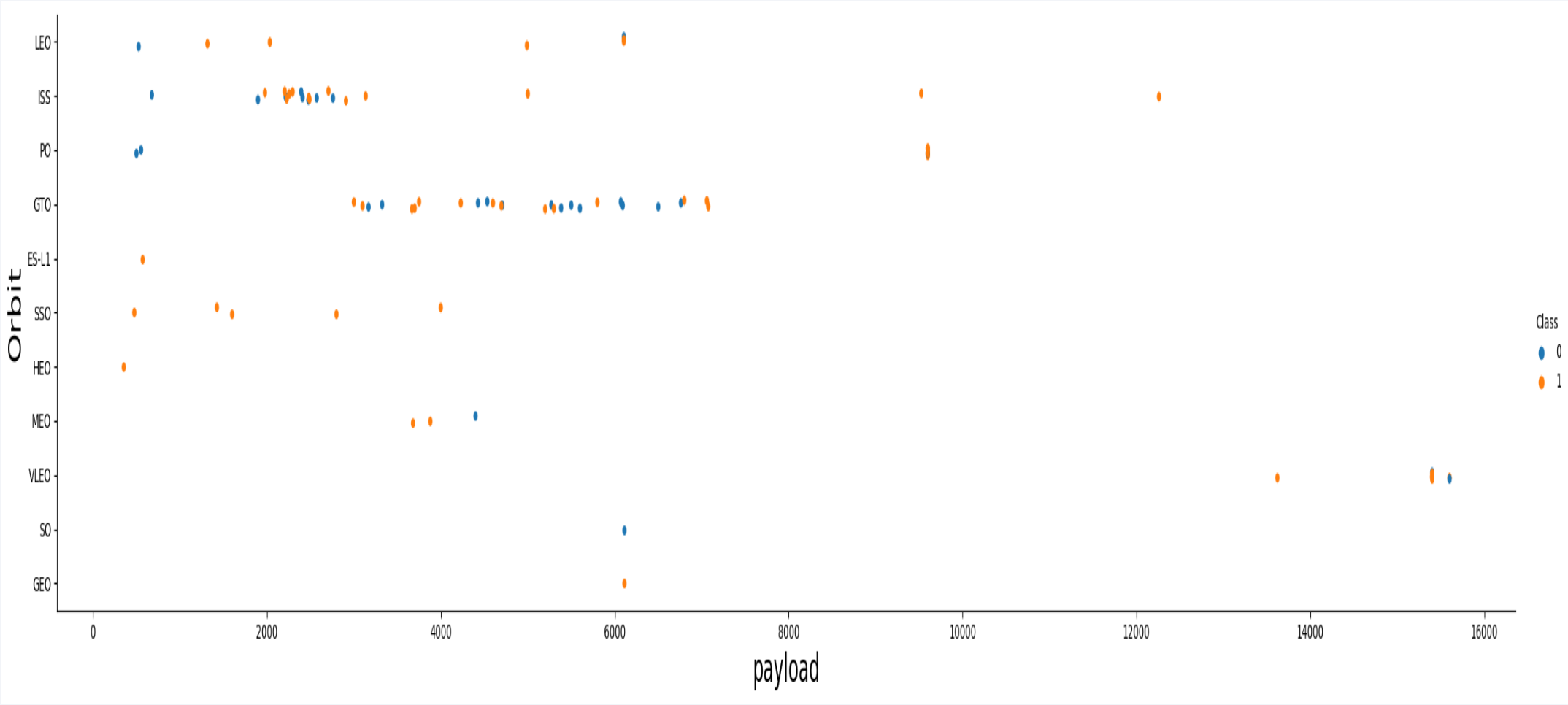
# Class vs. Orbit Type



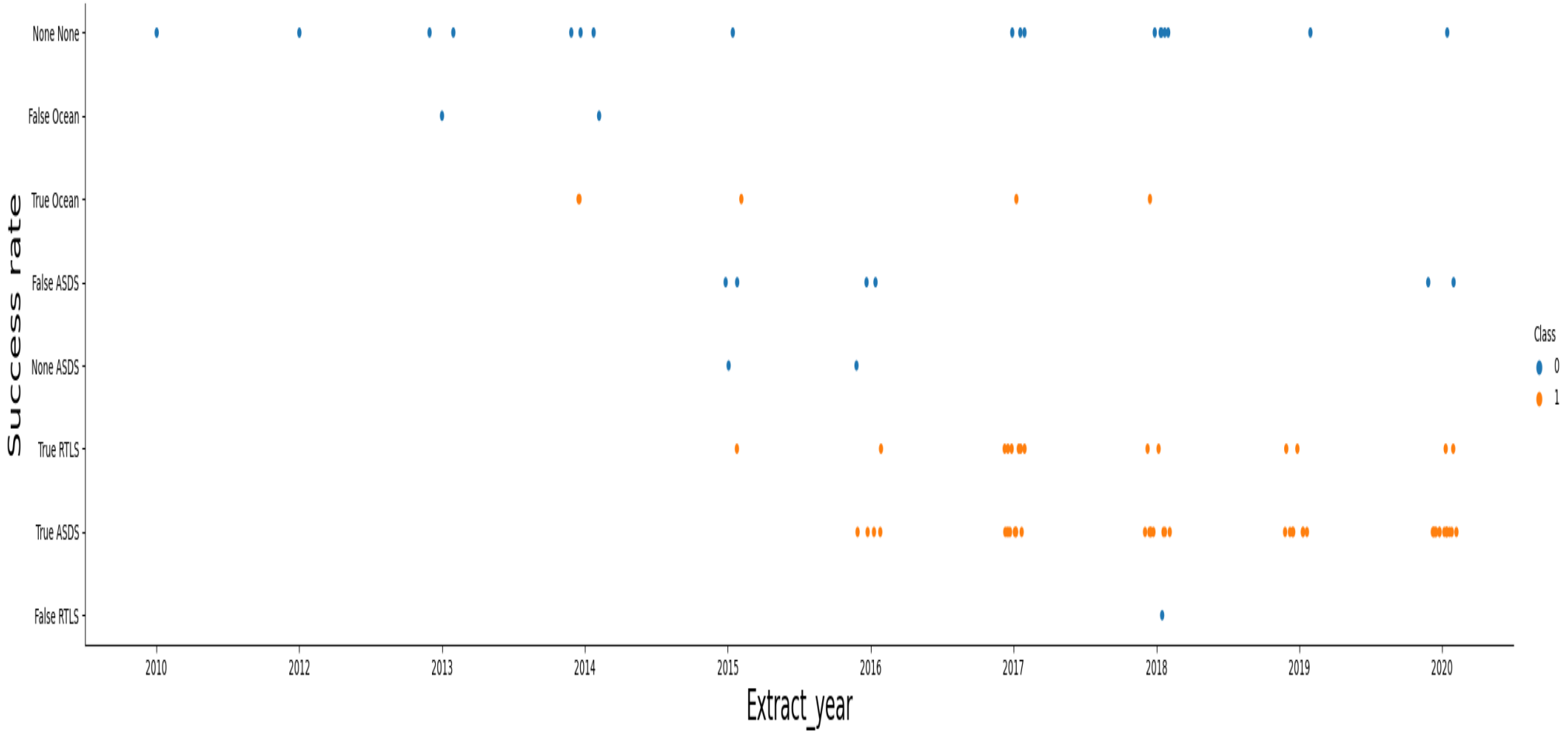
# Flight Number vs. Orbit



# Payload vs Orbit



# Yearly Trend vs Success Rate



# All Launch Site Names

**select distinct**(Launch_Site) **from** SPACEXTBL;

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

# Launch Site Names Begin with 'CCA'

**select * from** spacextbl **where** Launch_Site **like** '%CCA%' **limit** 5

| Date | Time (UTC) | Booster_Version | Launch_Site | 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

```
select sum(PAYLOAD_MASS__KG_) from
  spacextbl where customer like '%nasa (crs)%'
```

| sum(PAYLOAD_MASS__KG_) |
| --- |
| 48213.0 |

# Average Payload Mass by F9 v1.1

```
select avg(PAYLOAD_MASS__KG_) from spacextbl where booster_version like '%F9 v1.1%'
```

| avg(PAYLOAD_MASS__KG_) |
| --- |
| 2534.6666666666665 |

# First Successful Ground Landing Date

```
select min(date) from spacextbl where landing_outcome like
  'success (ground pad)'
```

| min(date) |
| --- |
| 01/08/2018 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
select booster_version from spacextbl where landing_outcome
like'Success (drone ship)'

AND PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

# Total Number of Successful and Failure Mission Outcomes

**select** mission_outcome, **count**(mission_outcome) **from** spacextbl **group by** mission_outcome

| Mission_Outcome | count(mission_outcome) |
|---|---|
| None | 0 |
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

```
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 | 15600.0 |
| F9 B5 B1049.4 | 15600.0 |
| F9 B5 B1051.3 | 15600.0 |
| F9 B5 B1056.4 | 15600.0 |
| F9 B5 B1048.5 | 15600.0 |
| F9 B5 B1051.4 | 15600.0 |
| F9 B5 B1049.5 | 15600.0 |
| F9 B5 B1060.2 | 15600.0 |
| F9 B5 B1058.3 | 15600.0 |
| F9 B5 B1051.6 | 15600.0 |
| F9 B5 B1060.3 | 15600.0 |
| F9 B5 B1049.7 | 15600.0 |

# 2015 Launch Records

```
select SUBSTR(Date, 4, 2) as month,substr(date, 7, 4) as year,
  landing_outcome, booster_version, launch_site FROM
  SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone
  ship)' AND "DATE" LIKE '%2015%'
```

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

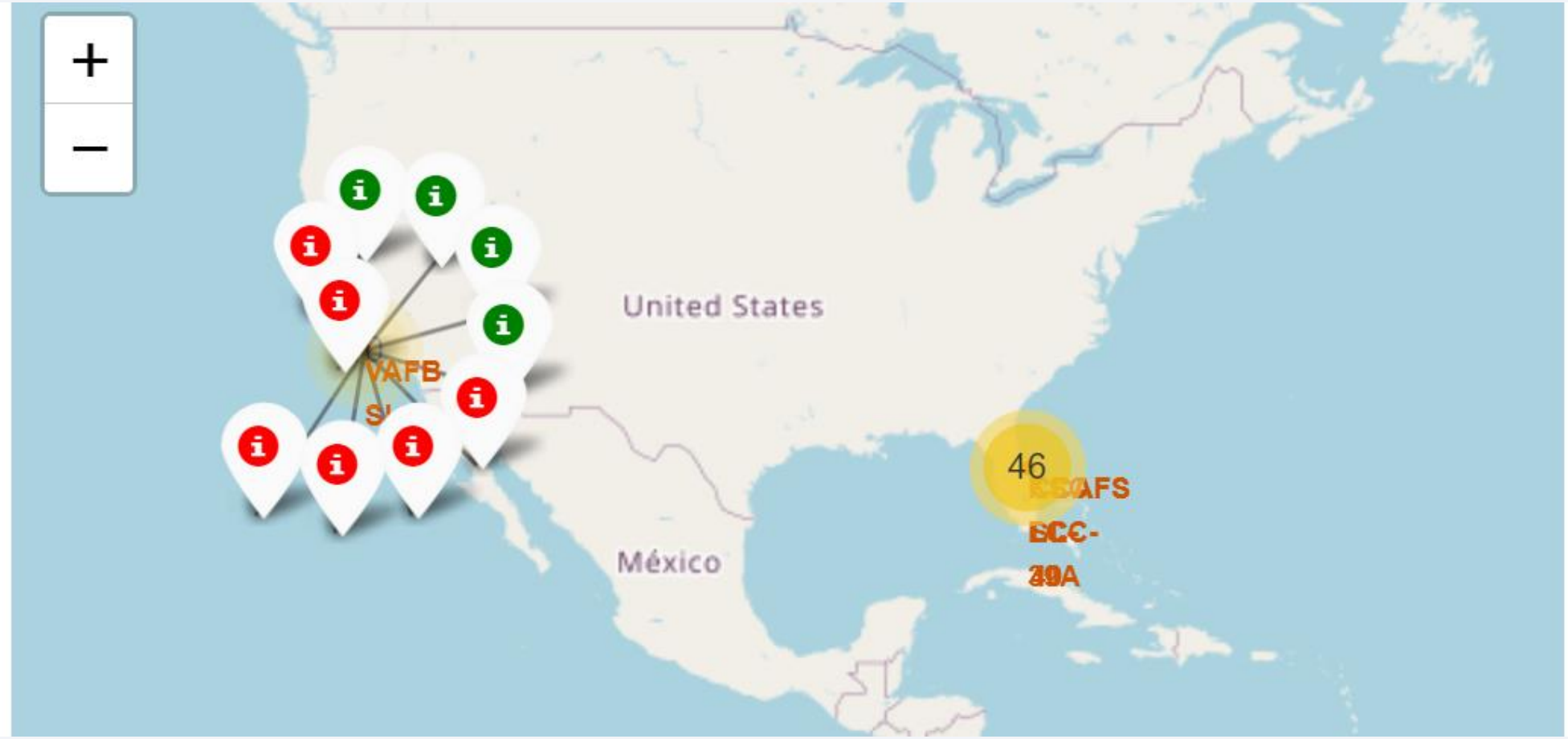
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
select landing_outcome, count(landing_outcome) from
  spacextbl where '2010-06-04' < date < '2017-03-20' group
  by landing_outcome order by count(landing_outcome) desc
```

| Landing_Outcome | count(landing_outcome) |
|---|---|
| Success | 38 |
| No attempt | 21 |
| Success (drone ship) | 14 |
| Success (ground pad) | 9 |
| Failure (drone ship) | 5 |
| Controlled (ocean) | 5 |
| Failure | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |
| No attempt | 1 |

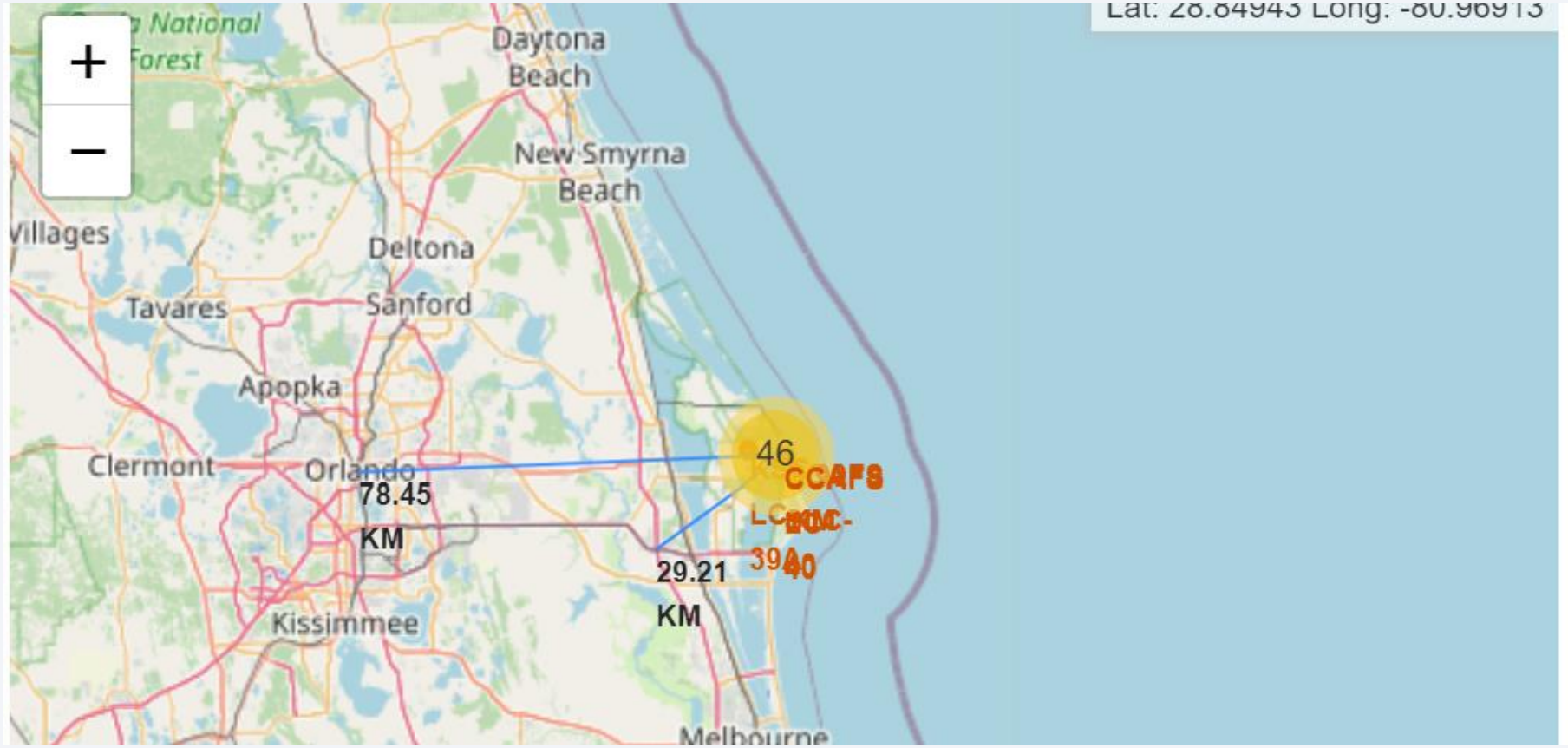

Section 3
Launch Sites
Proximities Analysis

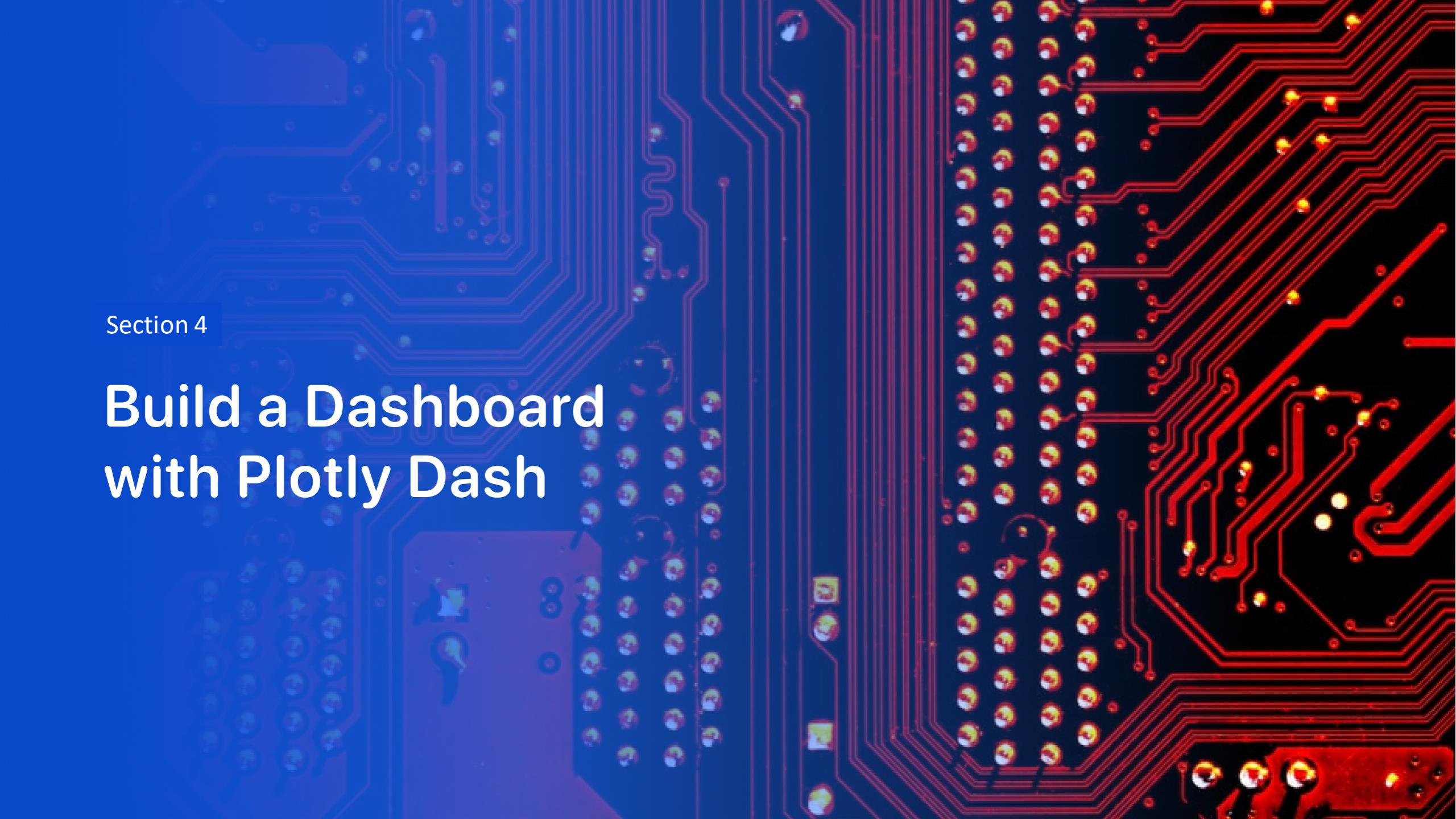
# Folium Map - Success and Failure Analysis of Launch Sites Using Colored Markers



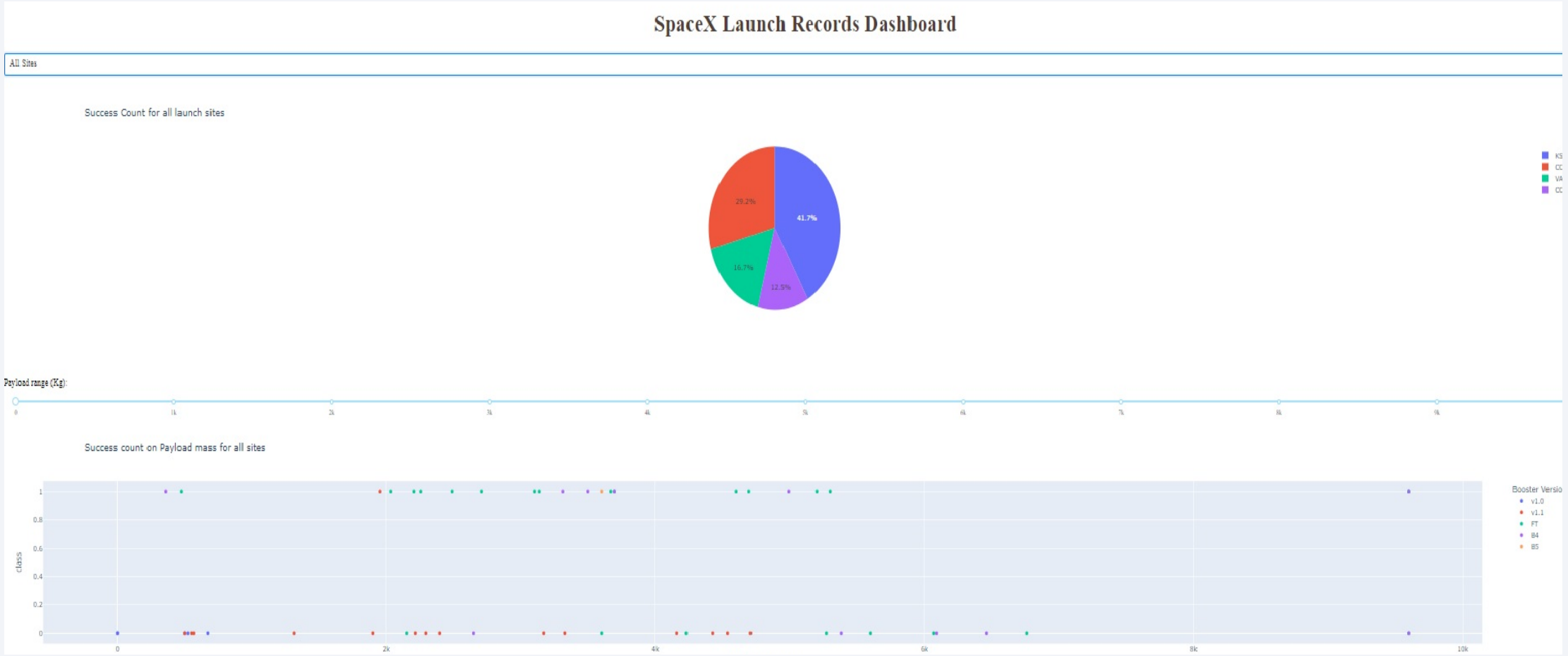
# Proximities Analysis of Launch Sites

# Amenities Distances Determination like Nearest Coastline, Closest City, Railway or a Highway

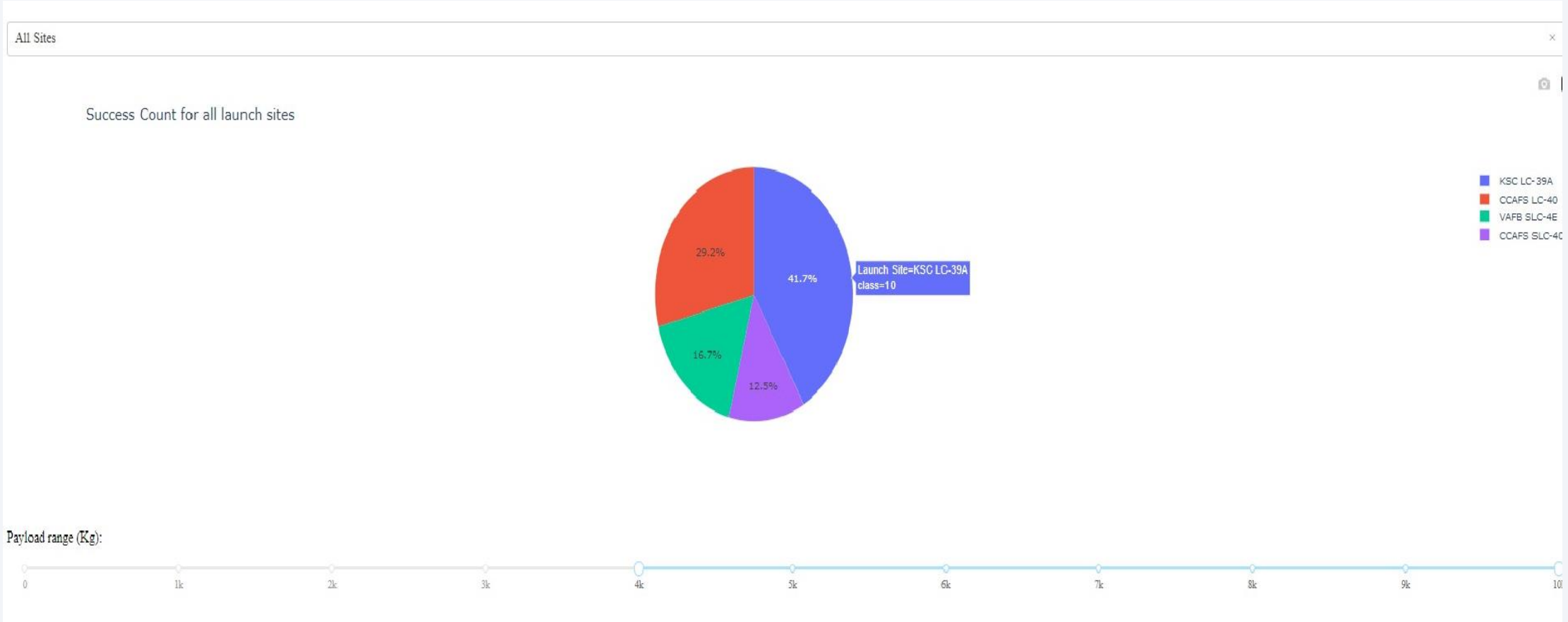



Section 4
Build a Dashboard with Plotly Dash

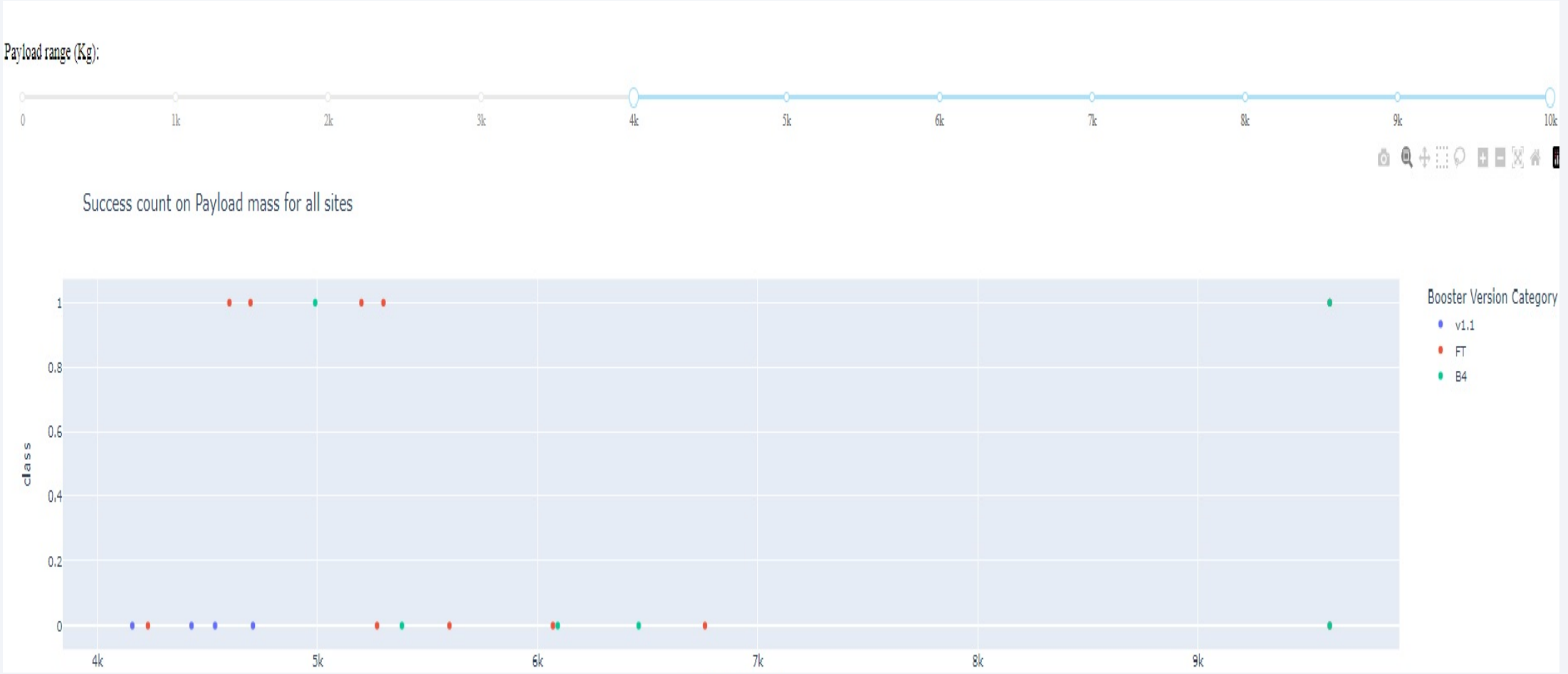
# ALL SITES SUCCESS RATE (SpaceX Dashboard)



# PIE CHART SUCCESS RATE OF PAYLOAD WITH 4, 000 KG (ALL SITES)



# SCATTER PLOT SUCCESS RATE OF PAYLOAD RANGE SLIDER WITH PAYLOAD AS 4, 000 KG



Section 5

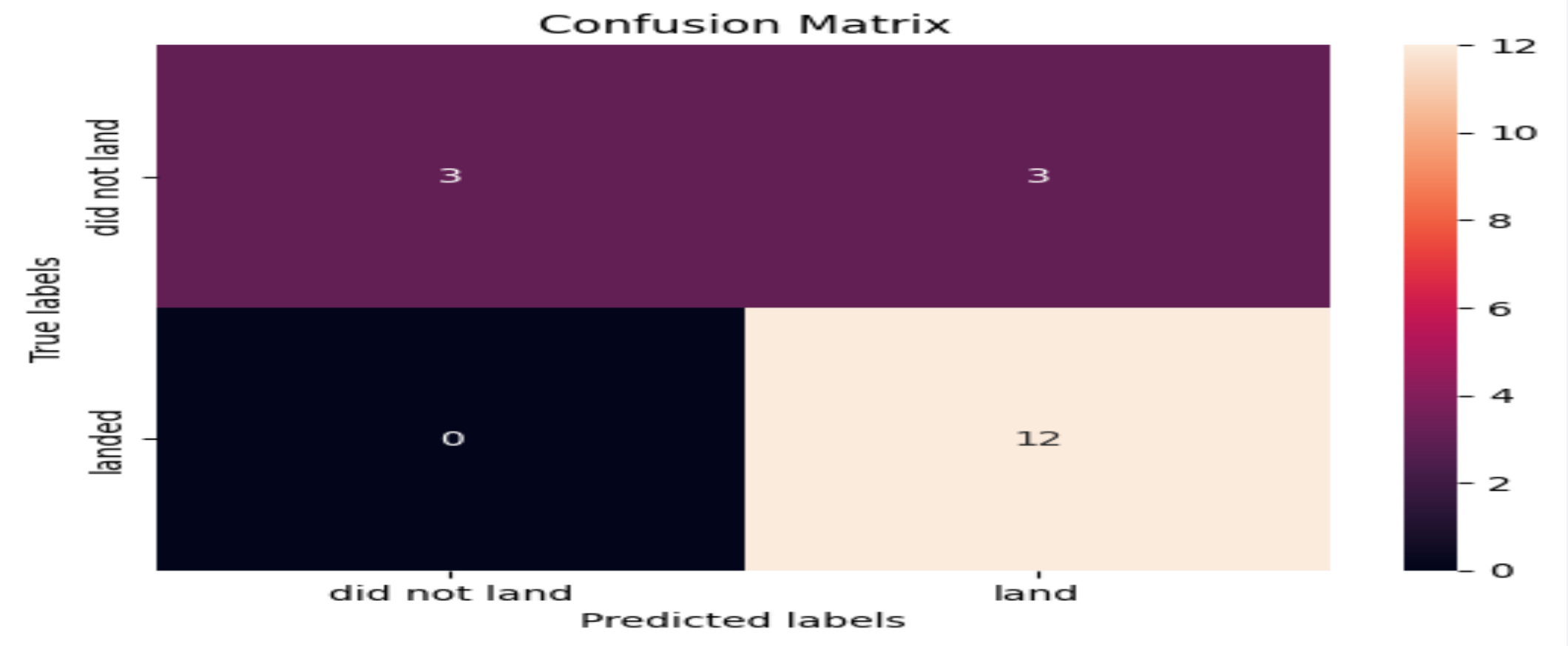
# Predictive Analysis (Classification)

# Classification Accuracy

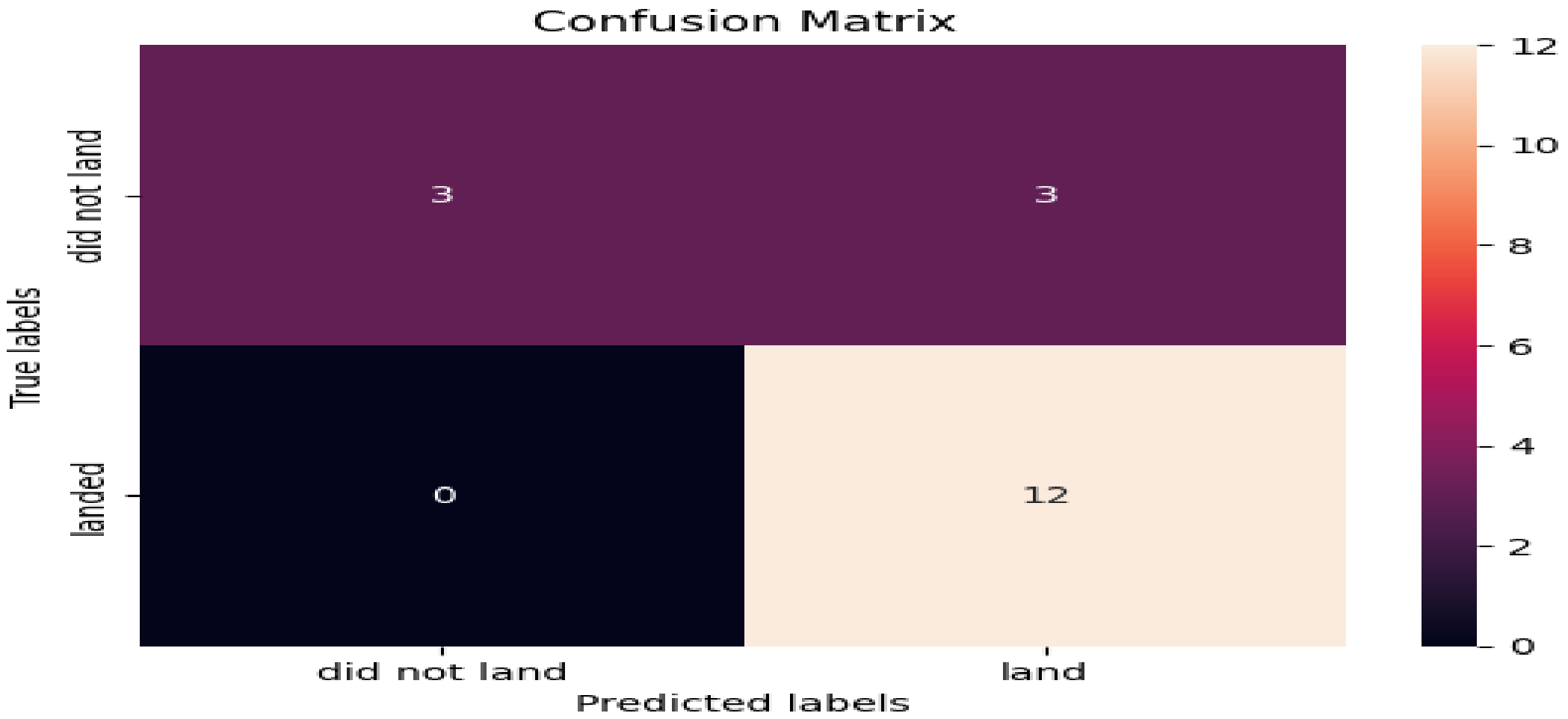
- Comparing all the classification models, it has been deduced that all of the classification models including logistic regression, decision tree, KNN and SVM has an equal score accuracy of 0.833 which is 83%.
- The inference which the comparison gives is that any of the classification models can be used for the purpose of prediction on the test data.
- For the comparison of the classification models as far as the maximum accuracy score is concerned, please refer to the code attached named Machine Learning, a jupyter notebook file in GitHub.
- The confusion matrices for all the classification models follows.

# Confusion Matrix

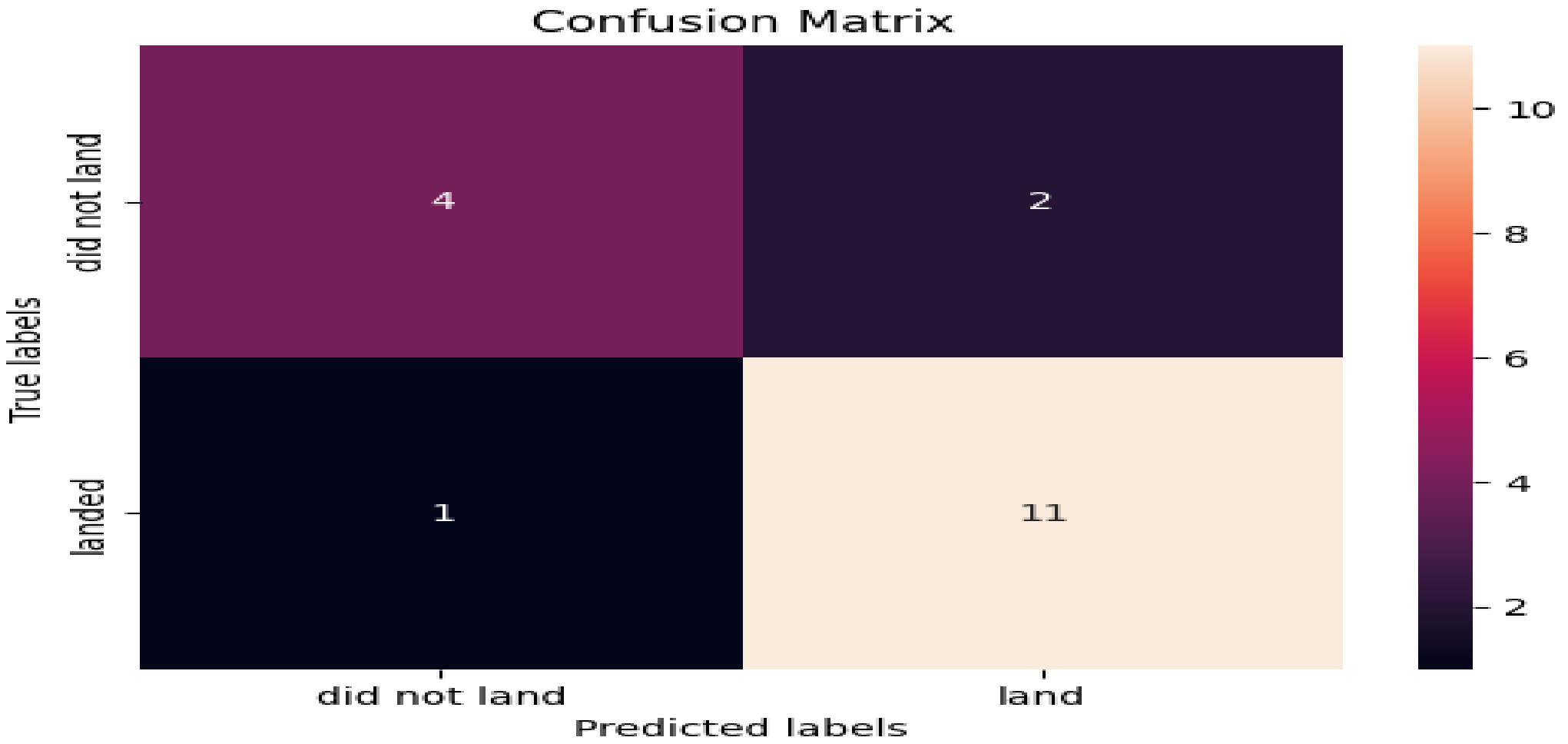
## Logistic Regression Classification Model



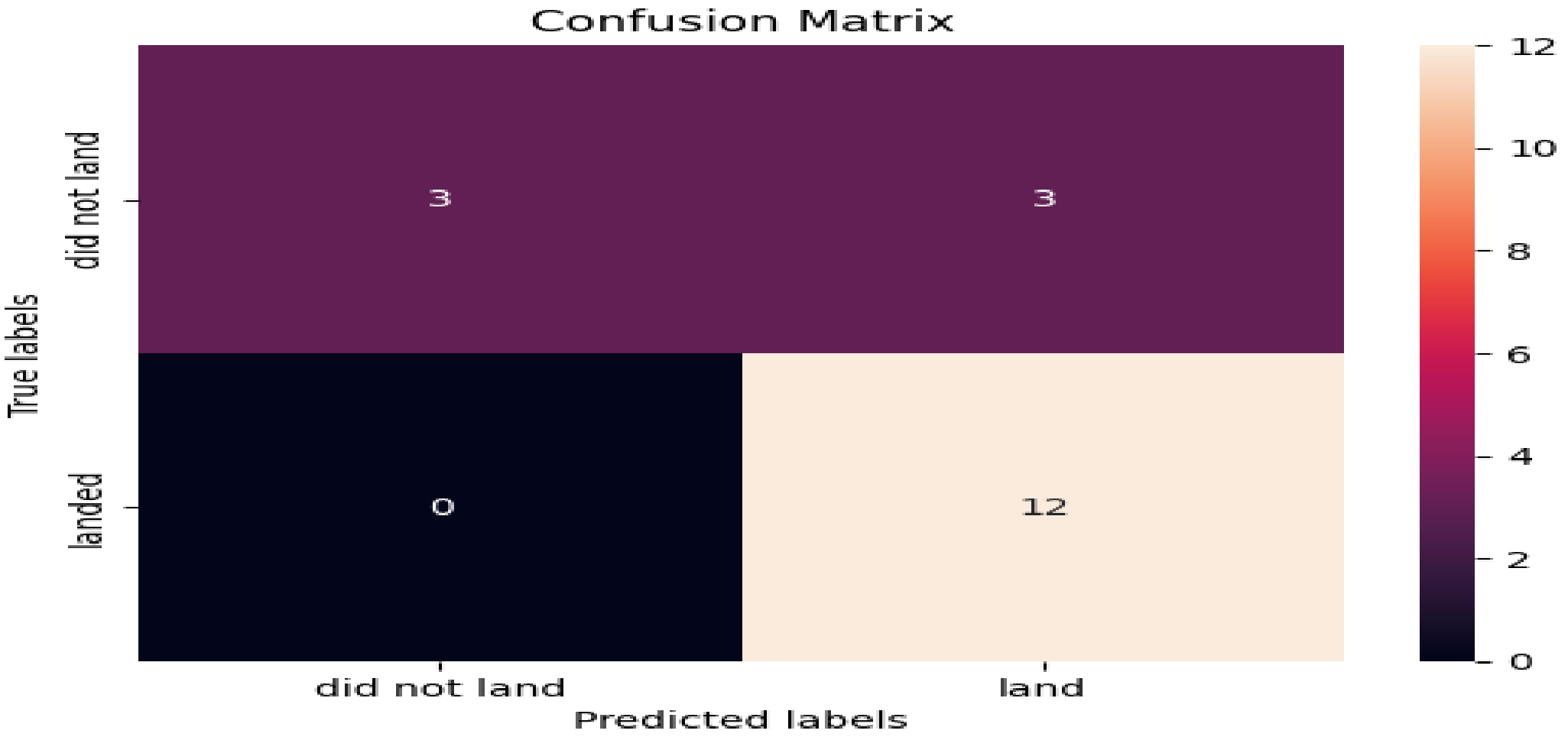
# Support Vector Machine Classification Model



# Decision Tree Classification Model



# K Nearest Neighbors Classification Model

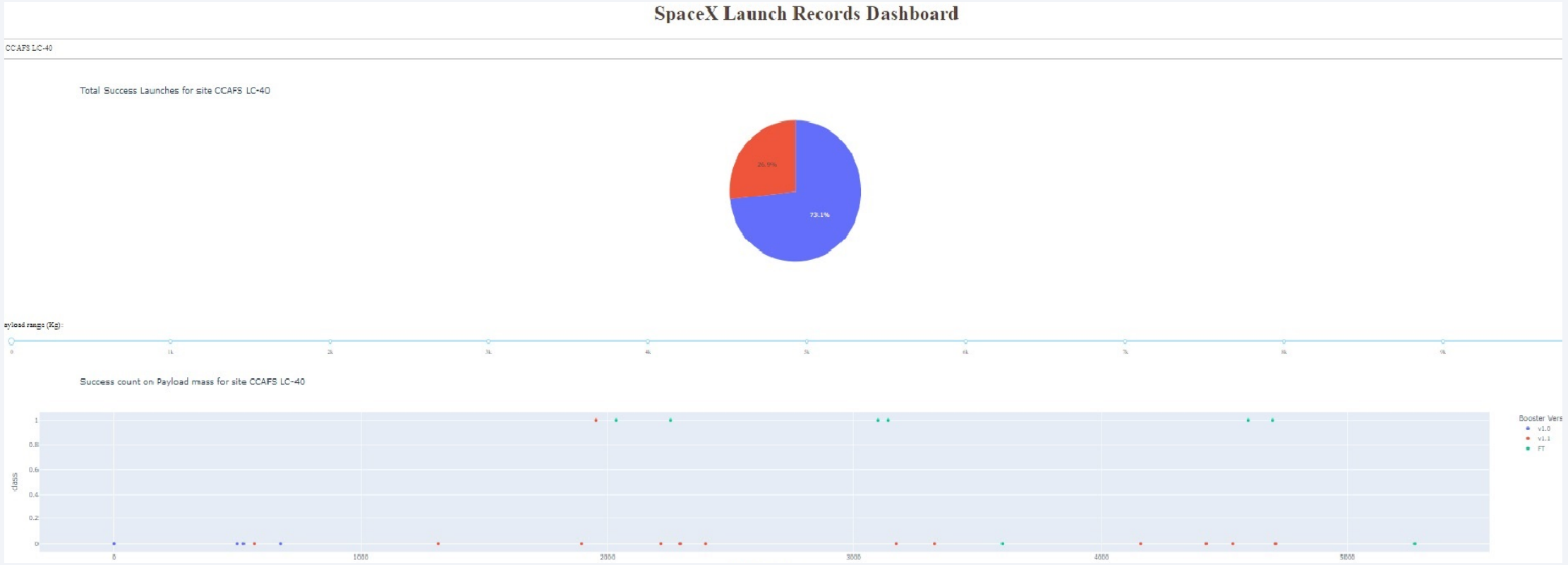


# Conclusions

- It is concluded that there is a **maximum rate of success** i.e. first stage recovery of the SpaceX Falcon9 when the following characteristics are implemented:
- **Payload mass** between 2000 kg and 4000 kg.
- **Launch site** being KSC LC 39A.
- Landing scenario is drone ship.
- **Minimum date** of launch is 2017 AC as the yearly trend keeps on increasing since the 2013 AC mark and is maximum success frate after 2017 AC
- It is estimated that the maximum **score accuracy** amongst all classification models is **83.33 %** and any of them can be used here.

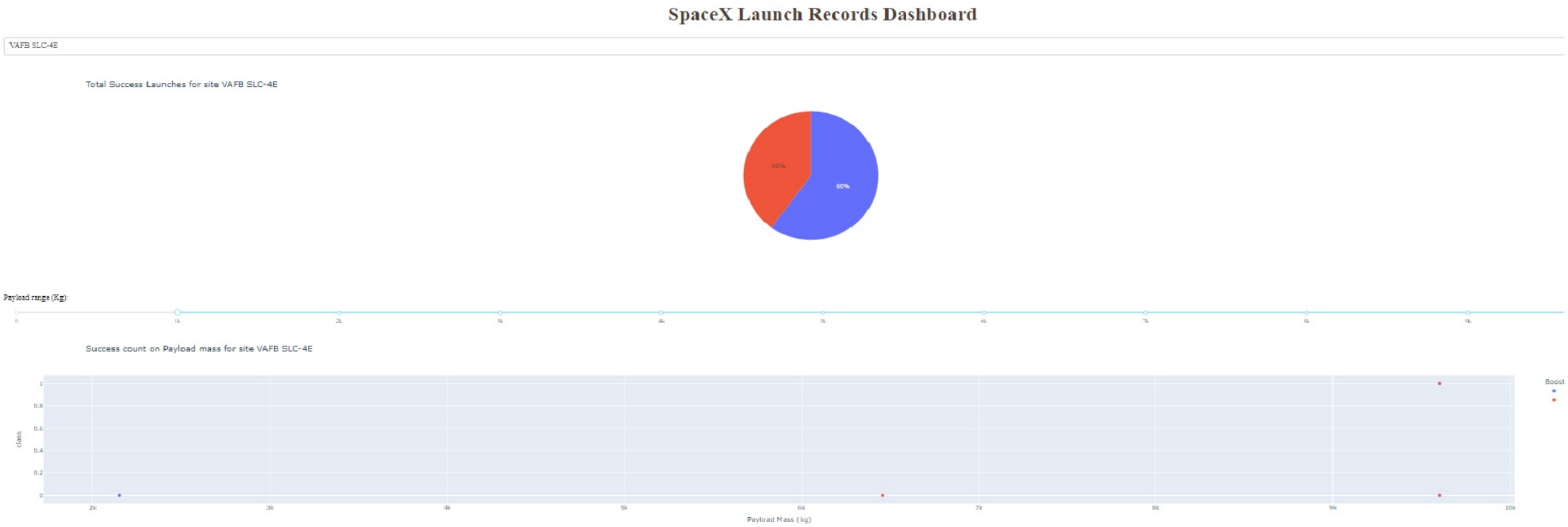
# Appendix

CCAF SLC –40 Success Rate



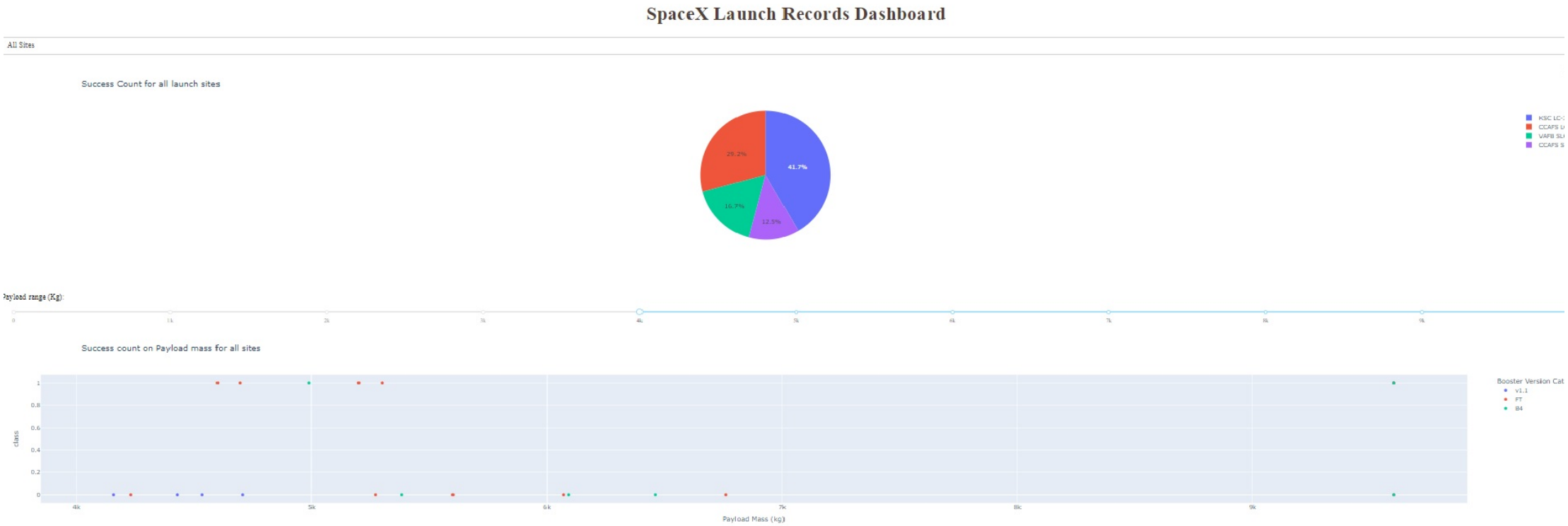
# Appendix

VAFB SLC-4E Success Rate with Payload as 1, 000 KG



# Appendix

ALL Sites Success Rate With Payload as 4, 000 KG



# Thank you!