IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Janvi Patel

30th dec, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodology
Data collection was done by using a SpaceX rest API and web web scraping
Data wrangling was done to filter data and removing unnecessary column from the data set
Train test split data was used to train a predict values
Various machine learning models such as logistic regression decision tree svm and K nearest neighbor we are used to predict

Summary of results
After finding certain insights of data it can be concluded that most of the launch sites were near coastline which is safest.
Decision three model is chosen for the best fit model

Introduction

SpaceX American Aerospace Company founded in 2002 that helped usher in the era of commercial spaceflight SpaceX advertises Falcon nine rocket launches on its website with a cost of $62 million other providers cost upward of $165,000,000 each must of saving is because SpaceX 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 SpaceX for a rocket launch by using public data applying data analytics methodologies and building machine learning model it can be determined that SpaceX could whether reuse its first state or not.

How payload mass orbit launch site correlate after first stage landing success rate?

what building model performs best for prediction.

change in success of lending over the timeline.

Section 1

Methodology

Methodology

Executive summary

Data collection:

Data is downloaded from two sources : SpaceX API and Web Scrapping using BeautifulSoup.

Data wrangling:

Data is cleaned by finding and filling missing values, removing unnecessary data and resetting index.

Exploratory data analysis using visualization and SQL:

Ploting charts, finding correlations and performing queries to filter data.

Visual analytics using folium and Plotly dash:

Mapping and creating creative dashboards.

Predicting analysis using classification model:

Various classification model such as logistic regression, decision tree, KNN and support vector system to build and predict outcomes. Model with most accuracy score fits the best.

Data Collection

SpaceX API:

https://api.spacexdata.com/v4/launches/past

Web Scrapping:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

In [5]:
```
# Takes the dataset and uses the launchpad column to call the API and append the data to the list
def getLaunchSite(data):
    for x in data['launchpad']:
       if x:
         response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
         Longitude.append(response['longitude'])
         Latitude.append(response['latitude'])
         LaunchSite.append(response['name'])
```

From the `payload` we would like to learn the mass of the payload and the orbit that it is going to.

In [6]:
```
# Takes the dataset and uses the payloads column to call the API and append the data to the lists
def getPayloadData(data):
    for load in data['payloads']:
       if load:
        response = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
        PayloadMass.append(response['mass_kg'])
        Orbit.append(response['orbit'])
```

From `cores` we would like to learn the outcome of the landing, the type of the landing, number of flights with that core, wh
were used, wheter the core is reused, wheter legs were used, the landing pad used, the block of the core which is a number u
version of cores, the number of times this specific core has been reused, and the serial of the core.

In [7]:
```
# Takes the dataset and uses the cores column to call the API and append the data to the lists
def getCoreData(data):
    for core in data['cores']:
            if core['core'] != None:
                response = requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
```

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b859130005fae363/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP res

```
# use requests.get() method with the provided static_url
requests.get(static_url)
# assign the response to a object
response = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response)
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/jupyter-labs-webscraping.ipynb

Data Wrangling

Initially, we filtered falcon 9 launch data from the dataset. Next, various functions were performed to get statistical summary of data such as occurrence of each orbit type, number of launches on each site and occurrence of mission outcome of orbit. Finally, Created landing outcome label from outcome label.

Source:

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

To get more insights of data effectively, scatter plots, bar charts and line charts were plotted to know relation between variables.

payload mass X flight number X launch site X orbit

Insights of success rate can determine from the various plotted charts.

Source:

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed on dataset to find:

- Names of the unique sites in the space mission.
- 5 records of missions starts with “CCA”.
- Total payload mass carried by a booster launched by NASA(CRS).
- Average payload mass carried by booster “F9 v1.1”.
- When the first landing outcome was on ground pad was achieved.
- Boosters with success in drone ship with given payload mass.
- Count of successful and failure outcomes of mission.
- A booster that have carried maximum payload mass .
- records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- 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.

Source:
https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers, circles, lines and clusters were used in the folium map.

- ❖Markers indicates launch sites.
- ❖Circles indicates specific coordinates.
- ❖Marker circles used for grouping launching sites.
- ❖Lines specifies distance between two points.

Source:

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Dashboard was created to show following graphs:

- Pie chart
- Scatter plot

Pie chart is to reflect percentage by launching sites.

Scatterplot is to depicts payload range over the years with launch orbit.

This allows audience to get clear perception of best launching site according to payload.

Source:

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/SpaceX_Interactive_Visual_Analytics_Plotly.ipynb.py

Predictive Analysis (Classification)

Firstly data was split into train and test set. Train set is used to build different models. Test set to predict values then actual value and predicted value get compared. Model with the highest accuracy is best fit model.

Classification models are logistic regression, KNN, SVM, decision tree.

Source:

https://github.com/janvi-patel04/applied-data-science-capestone/blob/3e5e7a07da1e0dbf4a5824cb1b85913005fae363/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results

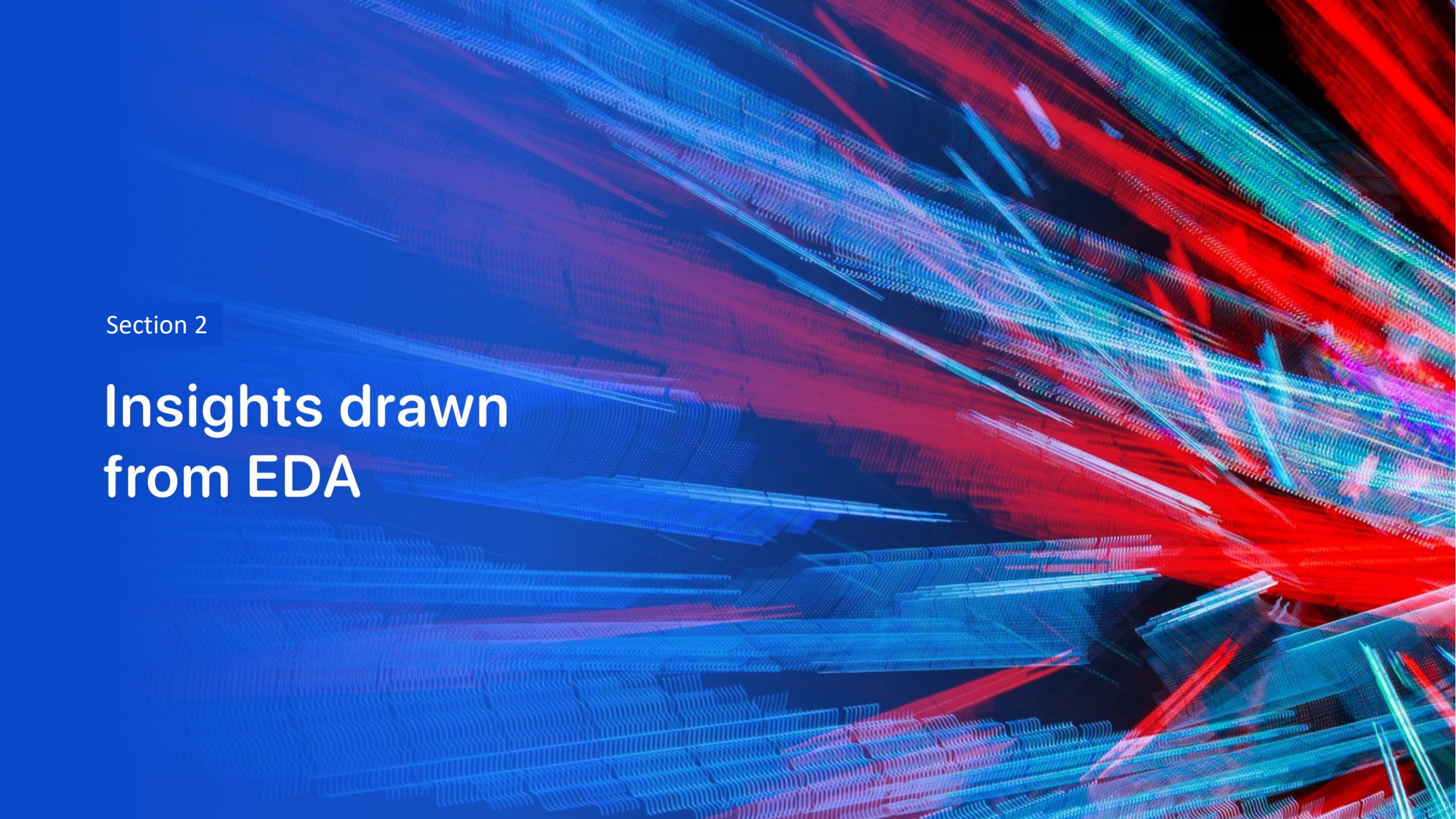
Space uses 4 launch sites.
Average payload mass of F9 v1.1 is 2928 kg
First landing outcome was done in 2015.
Almost 100% of mission outcome were successful.
There were two booster landing failure in 2015.
Count on successful landing increases by year.

- Exploratory data analysis results

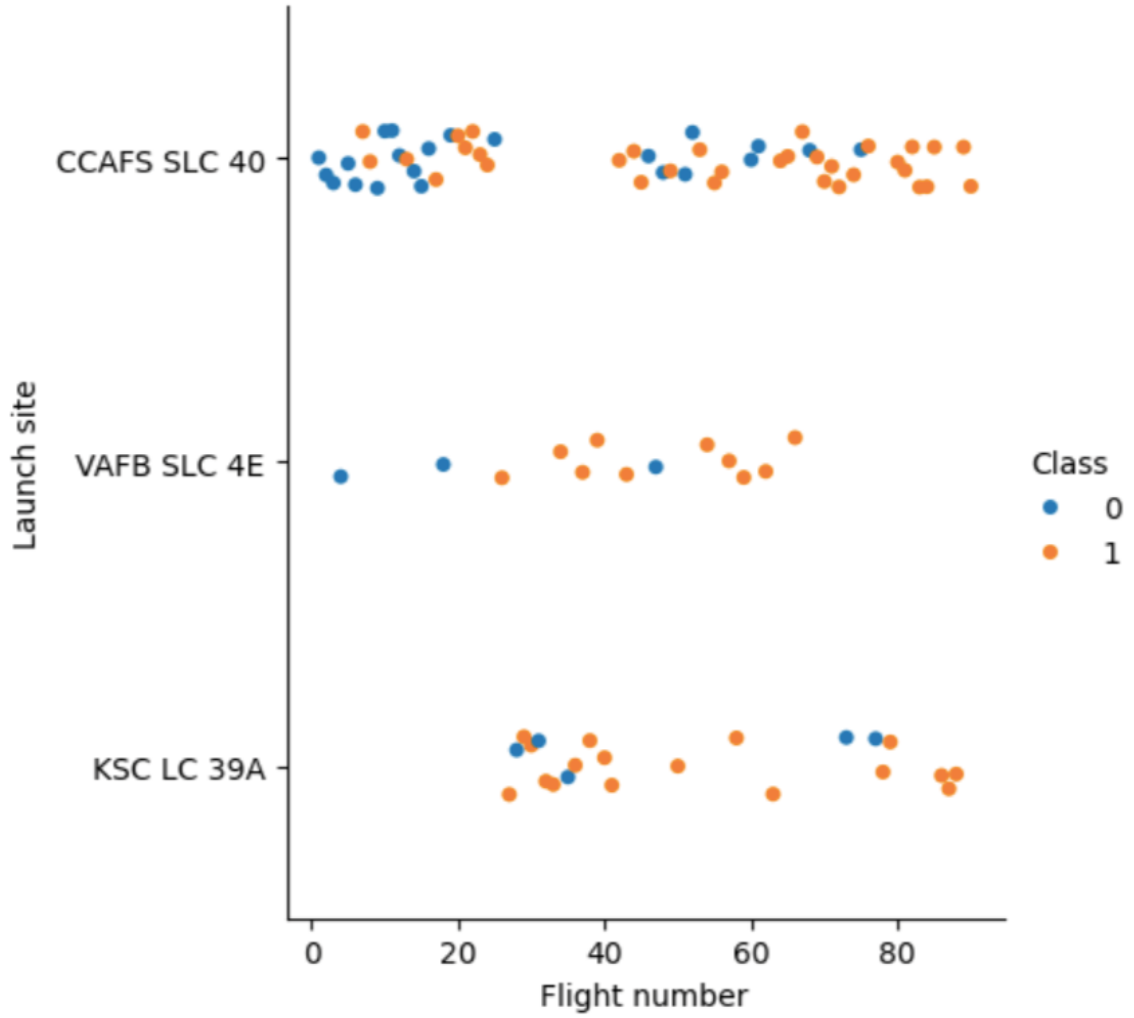
Predictive analysis shows that decision tree is the best fit model for predicting landing outcome with the highest accuracy.

- Interactive analysis results

Visualizing map helps to identify nearby Infrastructure that launch site is safe.
Most launch site is already near coastline.


Section 2
Insights drawn from EDA

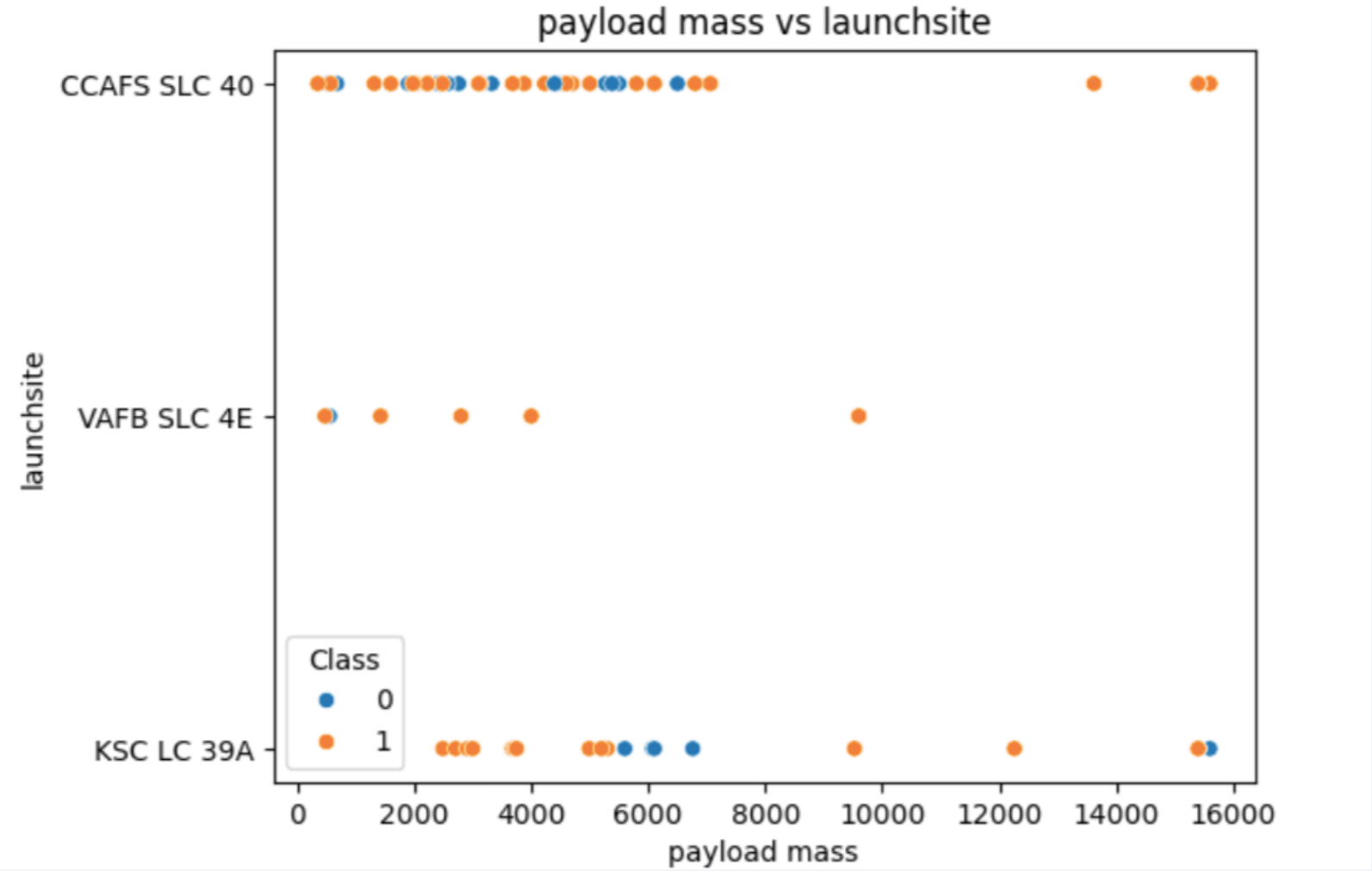
Flight Number vs. Launch Site



It can be seen that launch site CCAFS SLC 40 is with the most successful Launches over the years.

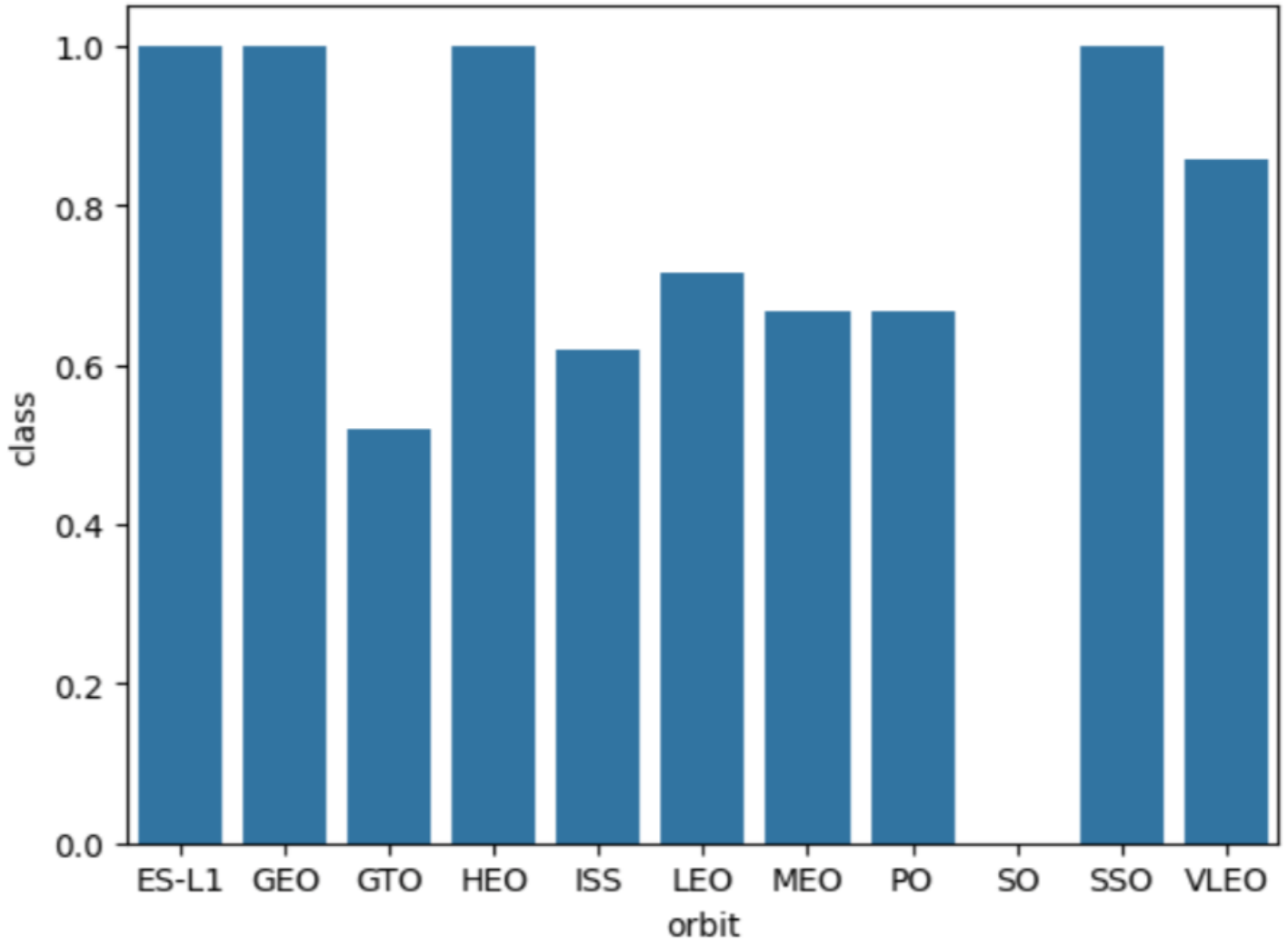
Least used launching site is VAFB SLC 4E But it cannot be denied that It has the most successful outcome with counts.

Payload vs. Launch Site



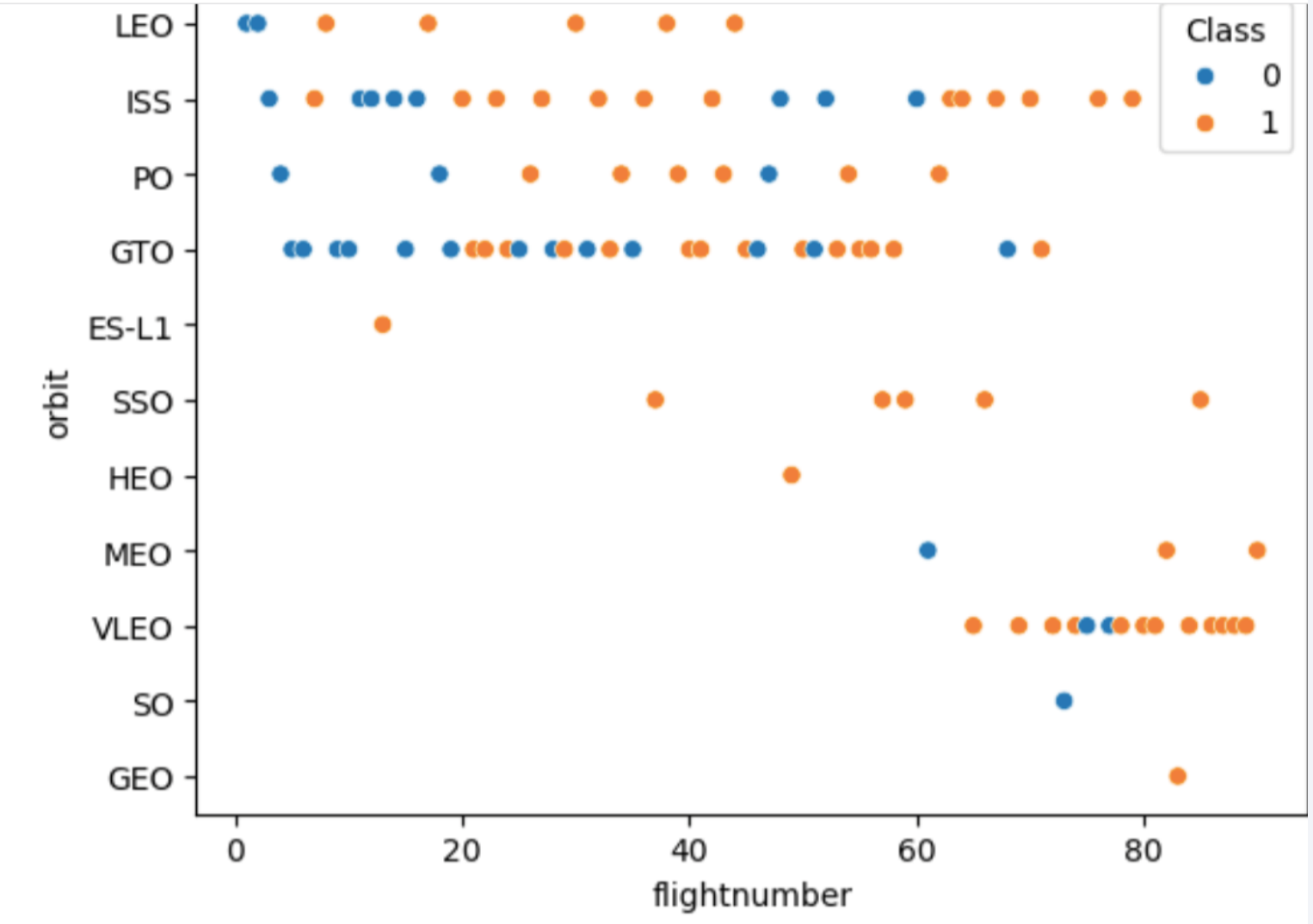
KSC LC 39A, launch site launces with lesser payload mass compared to other two.
CCAFS SLC 40 launch site used the maximum payload mass so far with successful launching.

Success Rate vs. Orbit Type



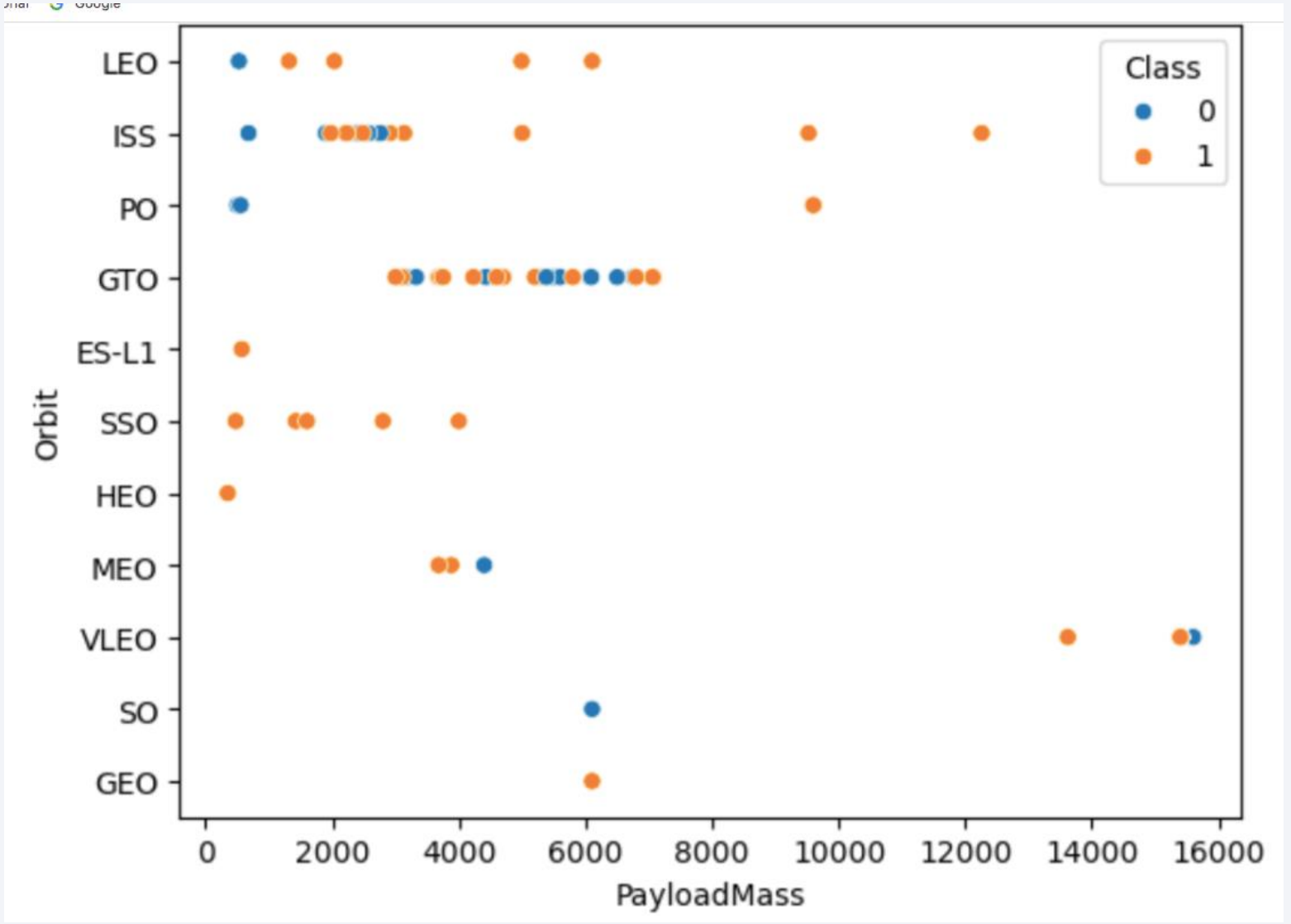
This bar chart depicts success rate of each orbit. It is clear that Sites ES-L1, GEO, HEO, SSO having same and highest rate, where GTO is having least rate and SO orbit with No attempts.

Flight Number vs. Orbit Type



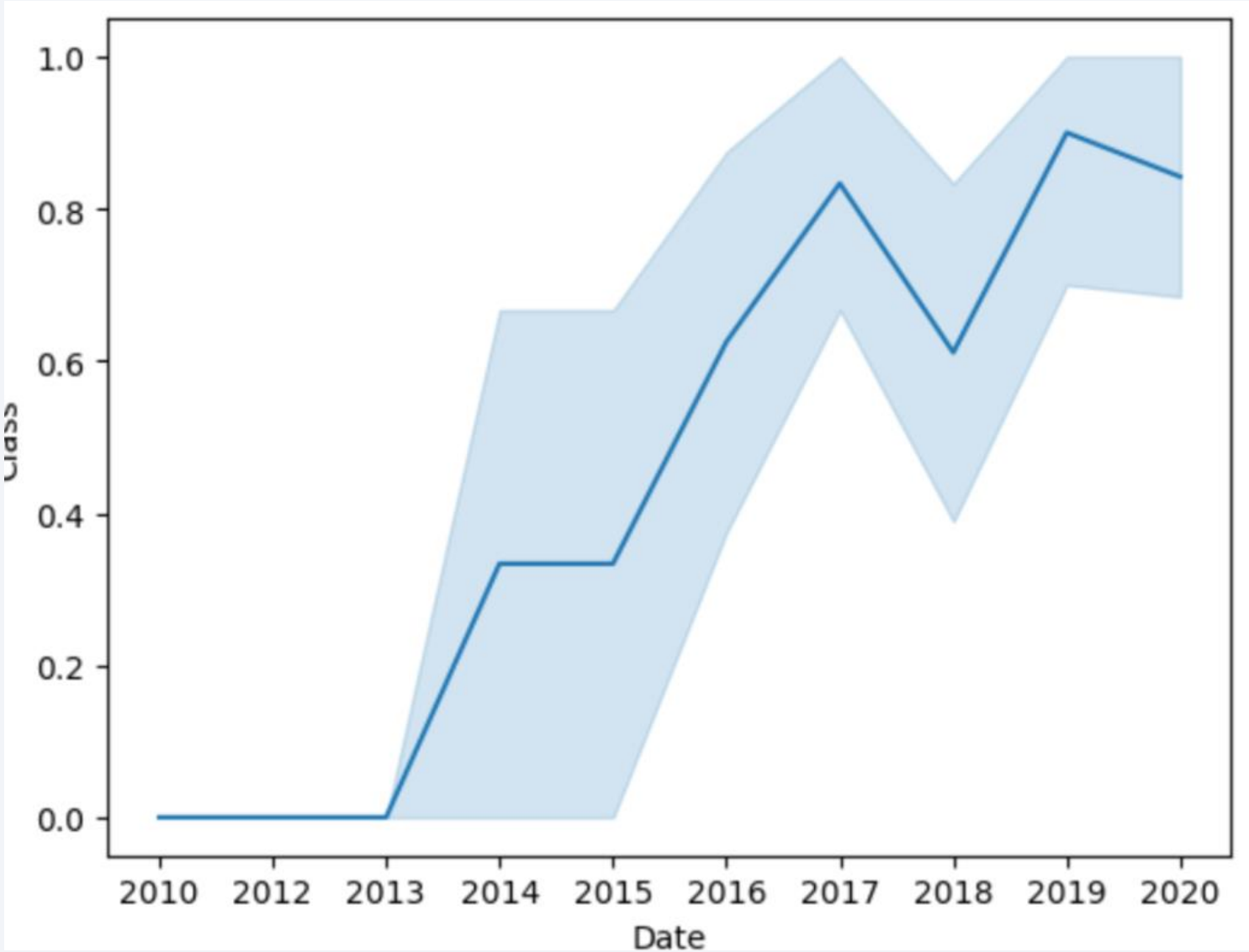
This scatterplot shows distribution of flight number to each orbit with launching outcomes. SSO orbit is one with most success outcome. ISS is with most attempts VLEO is with higher flight number also with higher Success.

Payload vs. Orbit Type



Scatter plot shows payload mass of each orbit type with success/failure on SSO orbit launching goes safe with lesser payload mass with successful outcomes. Whereas VLEO goes with higher payload mass with major successful outcome. On ISS orbit missions were launched with various range of payload mass and succussed in most mission.

Launch Success Yearly Trend



From the line plot shown in figure it can be said that over the year graph line increases over time. Number of mission and its successful outcomes Increases simultaneously.

All Launch Site Names

Now write and execute SQL queries to solve the assignment tasks.

Note: If the column names are in mixed case enclose it in double quote

Task 1

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

[15]:
```
%sql select distinct "Launch_Site" from spacextable
```

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

[15]:

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

Task 2 ¶

Launch Site Names Begin with 'CCA'

```
features = df[['FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', '
features.head()
```

	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

Here the code filtered the launch site that starts with CCA to get an overview of certain insights.

Total Payload Mass

Display the total payload mass carried by boosters launched by NAS

In [24]:
```
%sql select sum(PAYLOAD_MASS__KG_) as total, Customer fro
```

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

Out[24]:

total	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

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

In [30]:
```
%sql select avg(PAYLOAD_MASS__KG_) as average_mass, Booster_Version from spacex
```

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

Out[30]:

average_mass	Booster_Version
2928.4	F9 v1.1

A SQL query that retrieves average pay load mass for the booster version F9 v1.1

First Successful Ground Landing Date

```
In [34]: %sql select min(Date), Landing_Outcome from spacext
 * sqlite:///my_data1.db
Done.
```

Out[34]:

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

A SQL query to find the what was the day when first success ground pad landing occurred.

Successful Drone Ship Landing with Payload between 4000 and 6000

t[52]:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

A SQL query to show all successful drone ship landing outcomes with its payload mass and booster version.

Total Number of Successful and Failure Mission Outcomes

57]:

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

A SQL query to find count of mission outcomes. Here we can see that most outcomes are successful.

Boosters Carried Maximum Payload

```
In [60]: %sql select Booster_Version, max(PAYLOAD_MASS__K
 * sqlite:///my_data1.db
Done.
```

Out[60]:

Booster_Version	max(PAYLOAD_MASS__KG_)
F9 B5 B1048.4	15600

A SQL query that shows a booster version with maximum payload mass.

2015 Launch Records

In [18]:
```
%sql select substr(Date, 6,2) as Month, Booster_Version, Date
from spacextable \
where (Landing_Outcome) = 'Failure (drone ship)' and substr(D
```

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

Out[18]:

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

In first and fourth month failure drone ship landing outcome was on same launch site in year 2015.

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

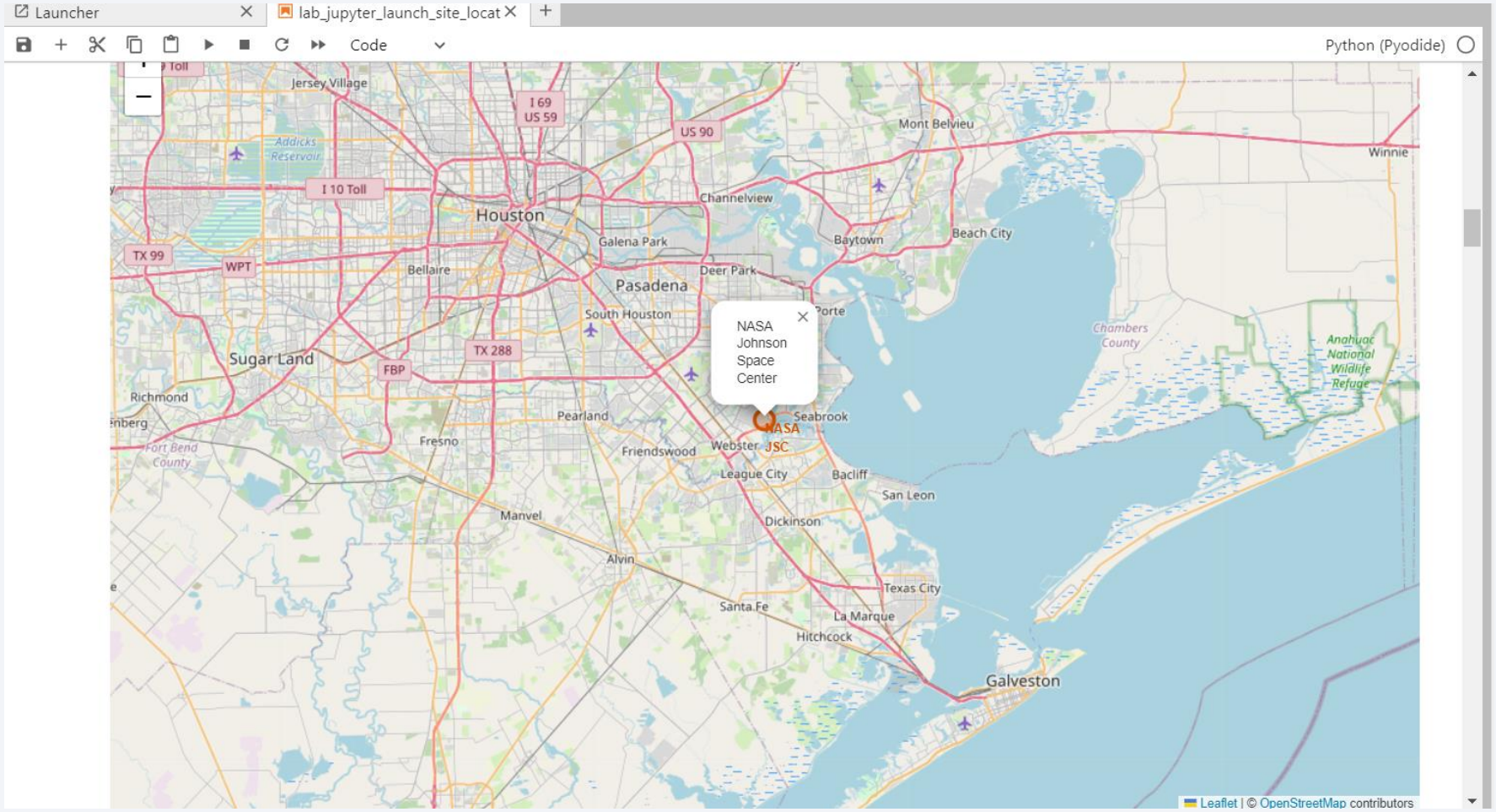
Out[21]:

Date	Landing_Outcome	counts
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

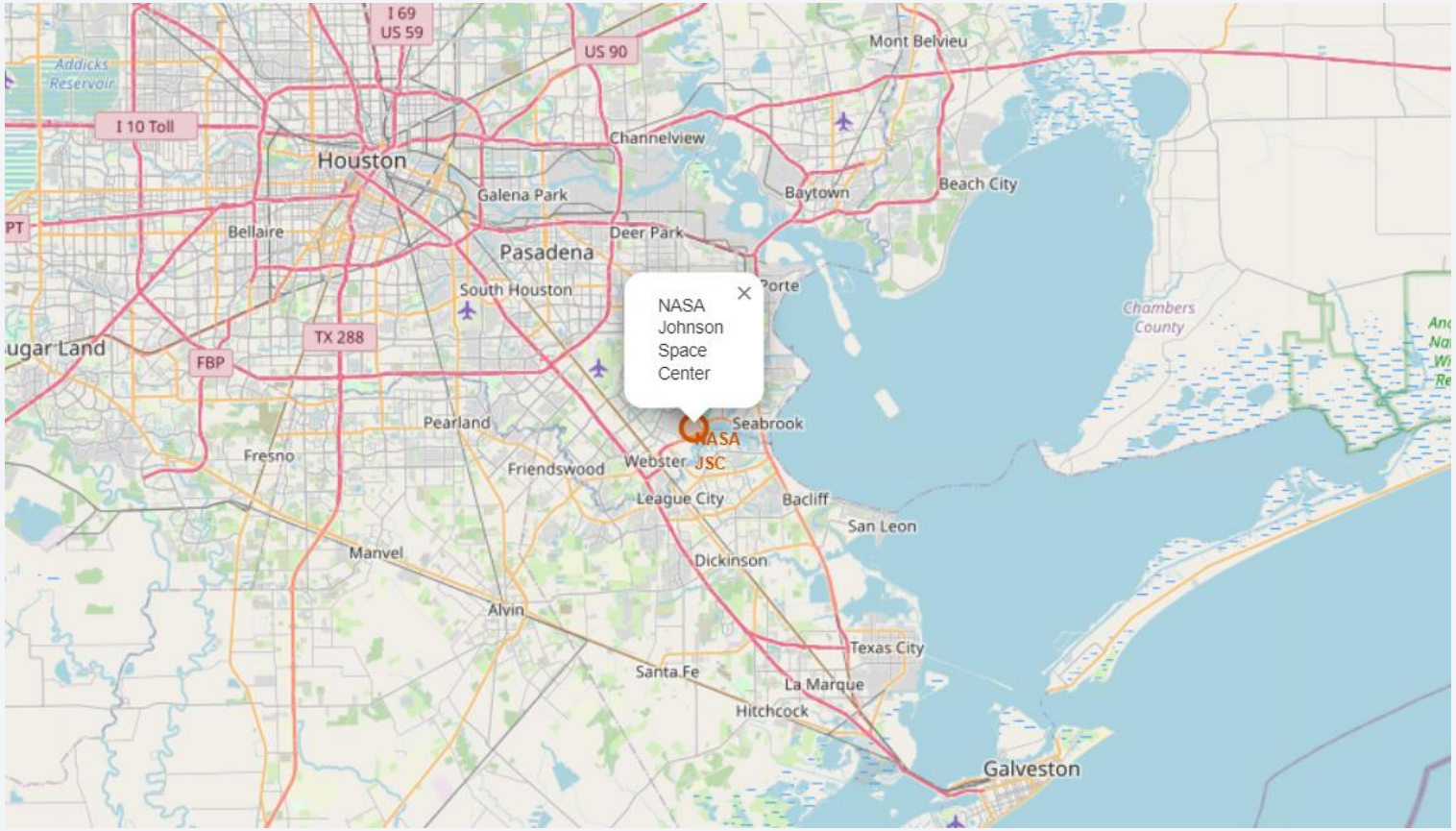
Section 3

Launch Sites Proximities Analysis

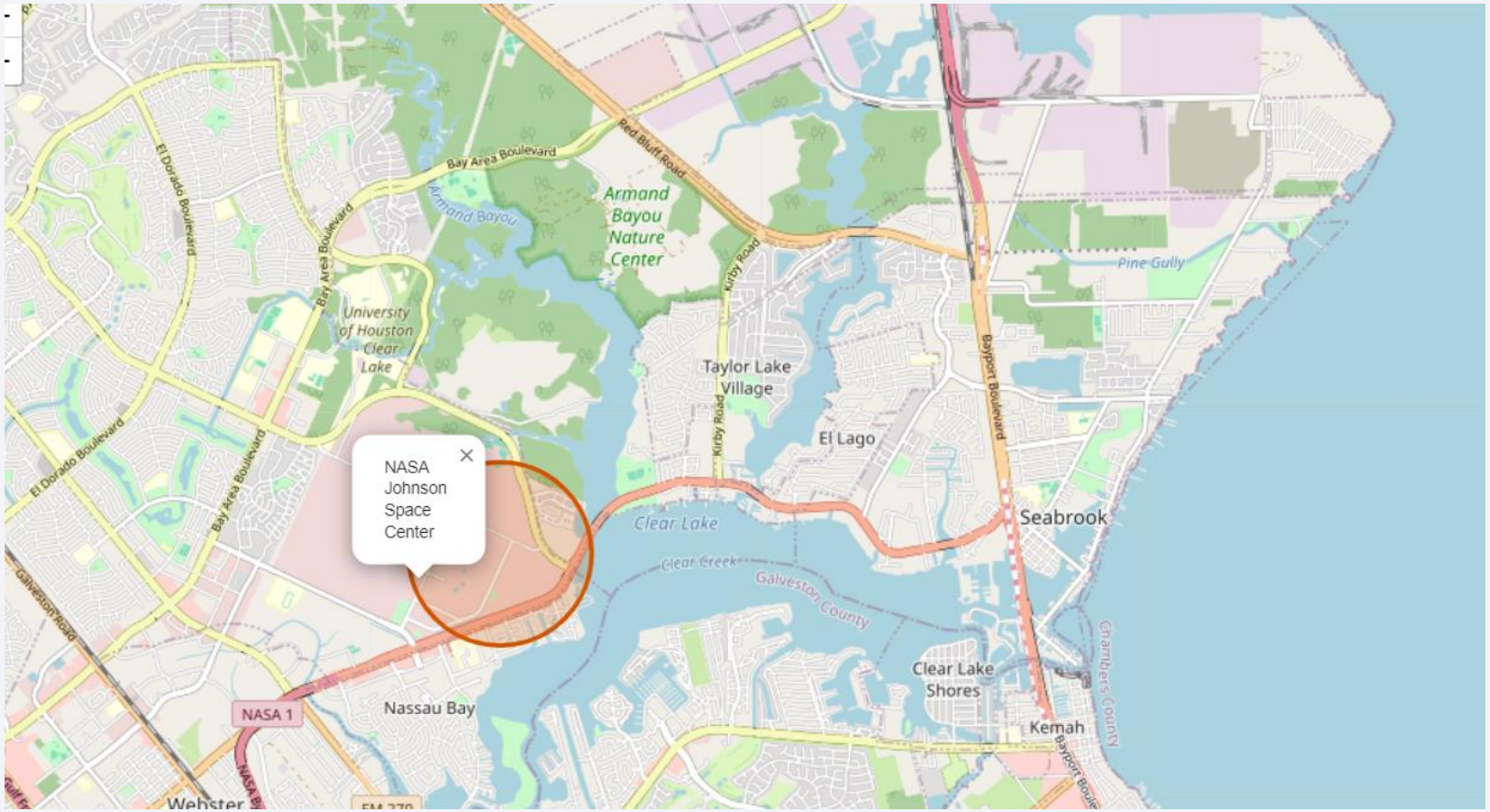
<Folium Map Screenshot 1>

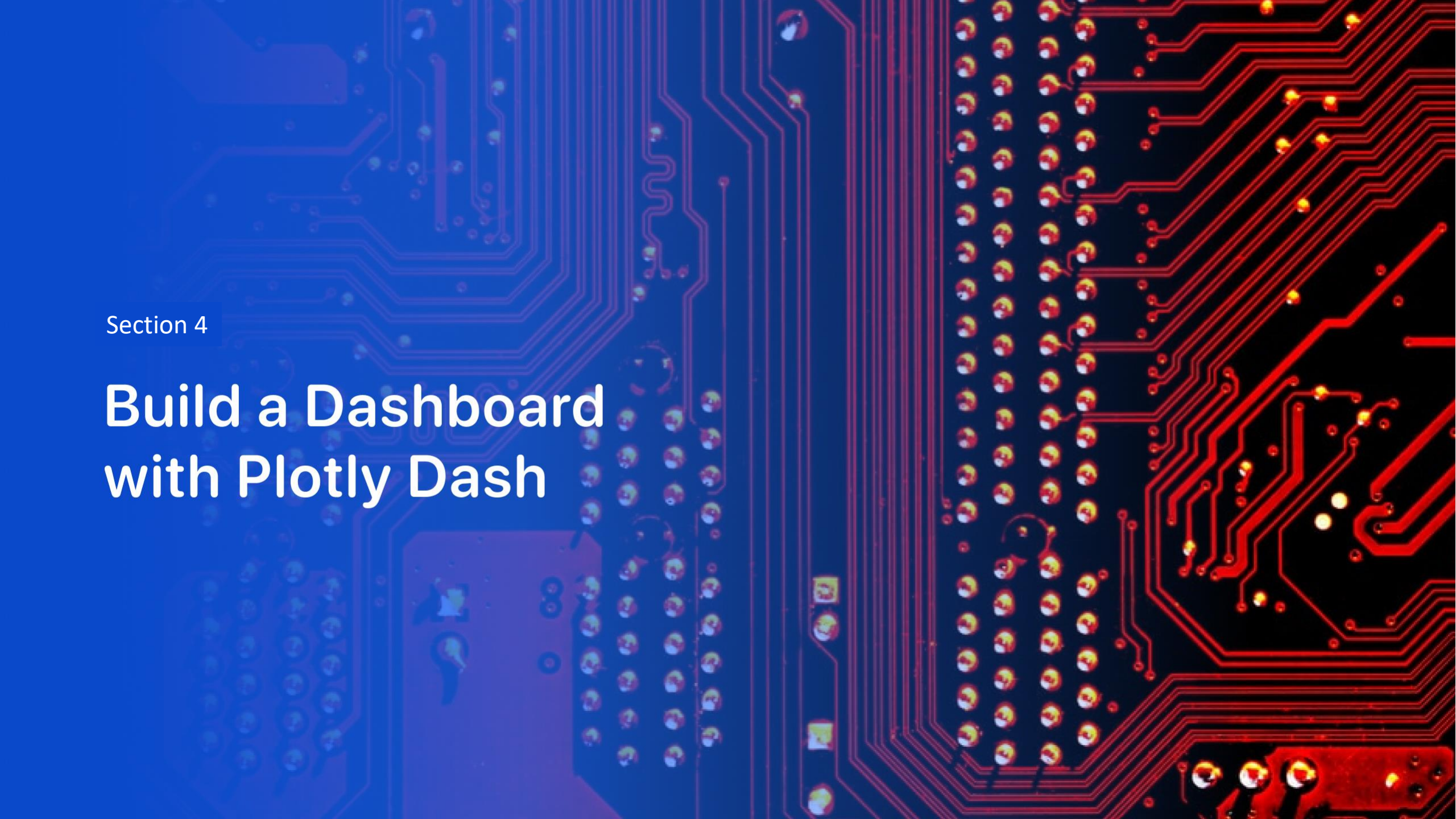


<Folium Map Screenshot 2>



<Folium Map Screenshot 3>

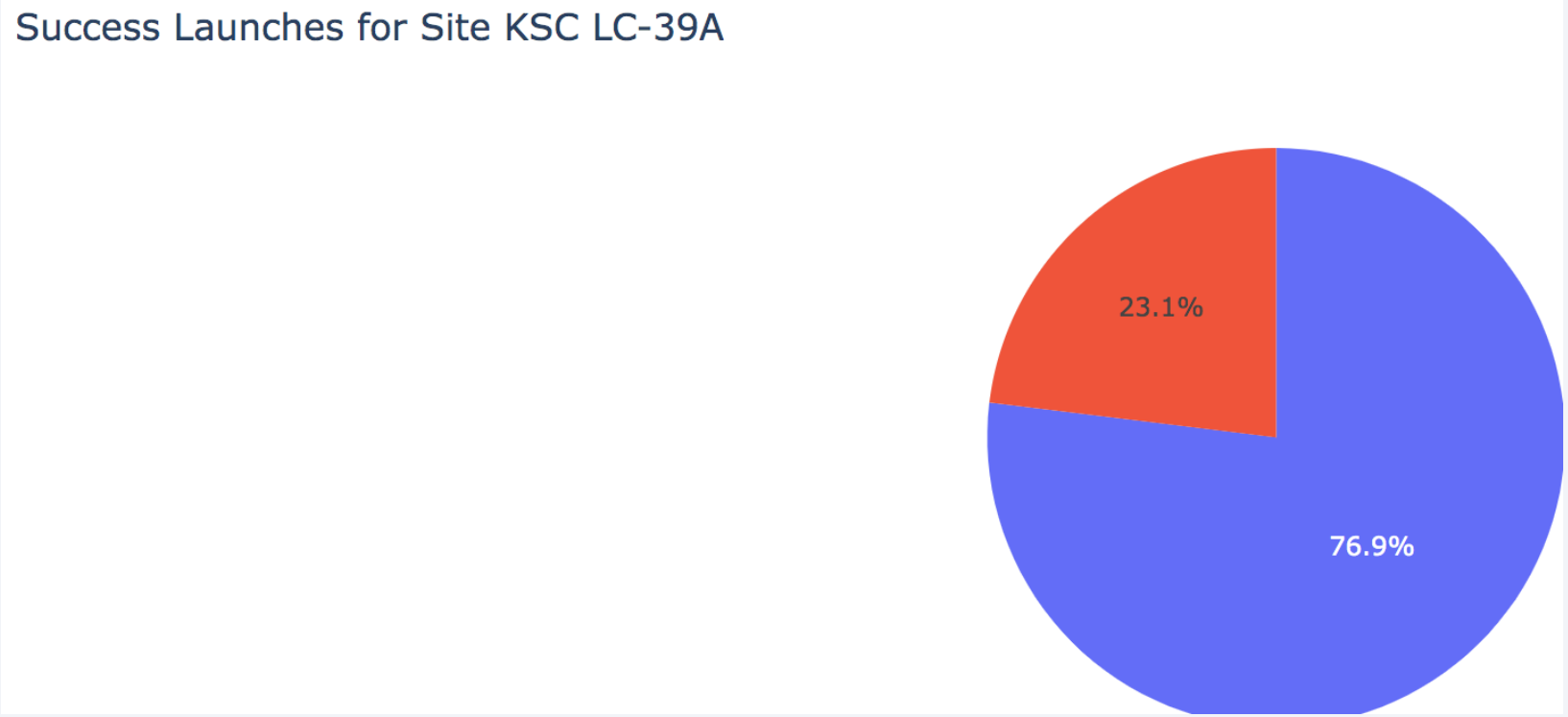



Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

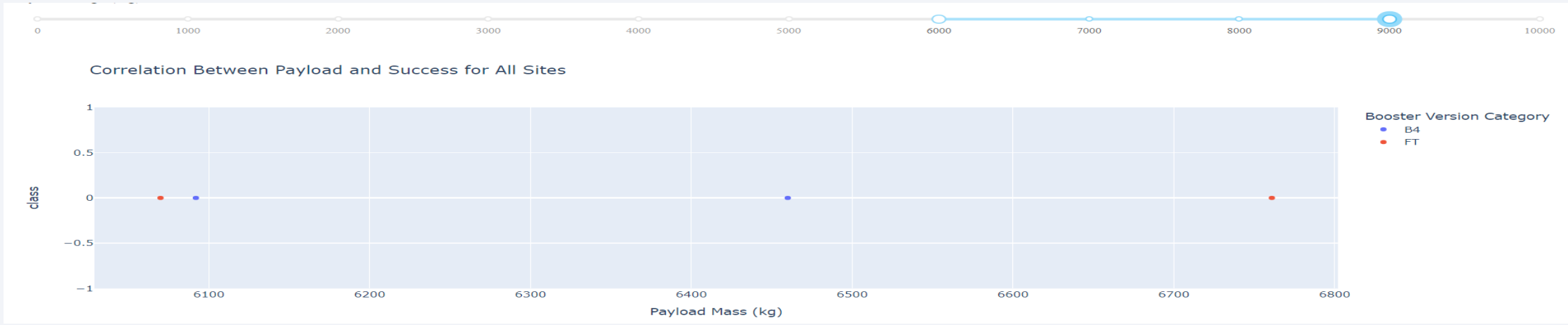
KSC LC -39A as a launch site with maximum use of 41.2 %. Where the least use site CCAFS LC-40

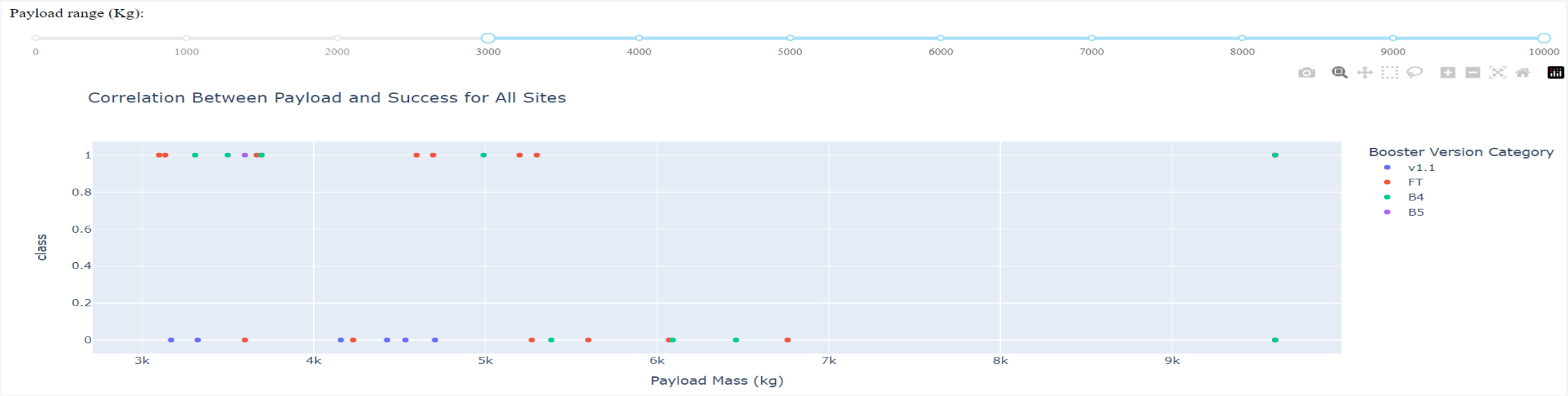
<Dashboard Screenshot 2>



Success launches for this site is highest among all other sites with 76.9% rate

<Dashboard Screenshot 3>





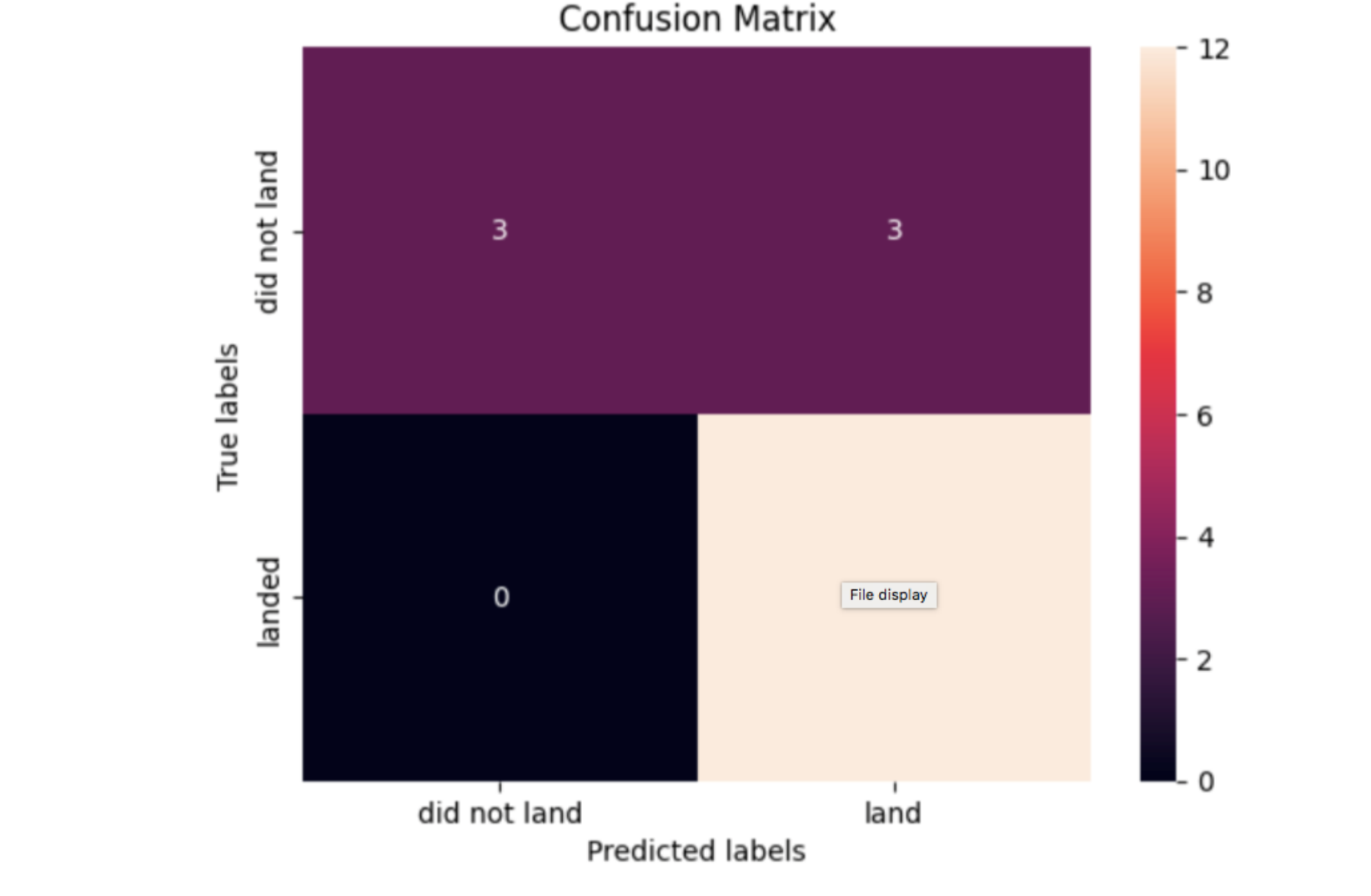
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- First a logistic regression model was build with accuracy score of 83.33
- Secondly a K-Nearest Neighbor model was build with different K values and a good accuracy score.
- A decision tree model build with highest accuracy score which was chosen as a best fit model for predicting landing outcomes.
- Support victor machine model was build with reliable jaccard index score and r-squared score

Confusion Matrix



Conclusions

- In conclusion, SpaceX can use its first launching to save cost.
- Machine learning classification algorithm called decision tree is bet fit model.
- Payload mass relates with launching outcomes in mission.
- All launching sites are safe to conduct mission.

Thank you!