IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nuttapat Pianarnupap
14 June 2024

Outline

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

Executive Summary

Summary of methodologies

- **Data was collected** from the SpaceX public API

- **Utilized SQL queries** and various data visualizations to uncover insights within the dataset

- Use **Grid Search method** to find the best Machine Learning Model to predict the classification of next landing

Summary of all results

- **Exploratory Data Analysis** result

- **Predictive Analytics** result

Introduction

- **Project background and context**

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings 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.

- **Problems to Research**

  - Can the historical launch data be used to predict the success of a new launch's first stage landing?

  - What operational conditions are necessary to guarantee a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - Encoded using one-hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV to find best fit model.

Data Collection

- To analyze the Falcon 9 rocket launches, data was gathered from multiple sources and processed accordingly:

    1. **SpaceX API**
    2. **Web Scraping from Wikipedia**

Data Collection – SpaceX API

- **SpaceX API**:
- The primary data source is the SpaceX API at https://api.spacexdata.com/v4/rockets/, filtered specifically for Falcon 9 launches.
- Missing values in the dataset were replaced with the mean of their respective columns.
- https://github.com/llouislouis/iBM_Data Science_Capstone_SPACE_X/blob/main/ Lab1_Data%20Collection%20API.ipynb

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

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

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

	static_fire_date_utc	static_fire_date_unix	tbd	net	window	rocket	success	details	ships	ca
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	False	0.0	5e9d0d95eda69955f709d1eb	False	Engine failure at 33 seconds and loss of vehicle	[]	
1	None	NaN	False	False	0.0	5e9d0d95eda69955f709d1eb	False	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	

Data Collection - Scraping

- **Web Scraping from Wikipedia**:
- Additional data was scraped from Wikipedia using the URL https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab2_Data%20Collection%20with%20Web%20Scraping.ipynb

In [6]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(html_data.text, 'html5lib')
```

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

In [7]:
```
# Use soup.title attribute
soup.title
```

Out[7]:
```
<title>List of Falcon 9 and Falcon Heavy launches – Wikipedia</title>
```

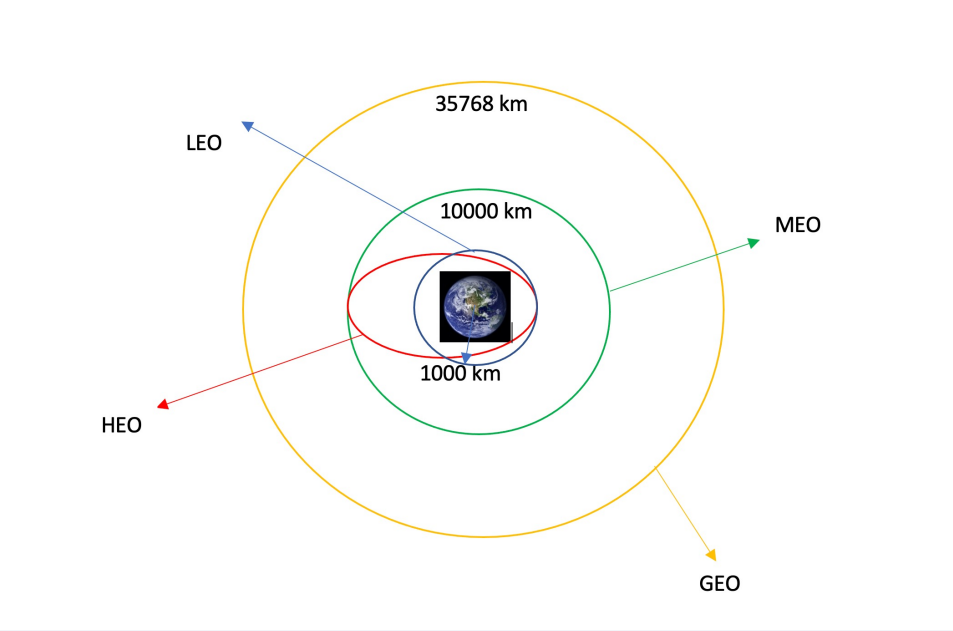
In [8]:
```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

Starting from the third table is our target table contains the actual launch records.

In [9]:
```
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

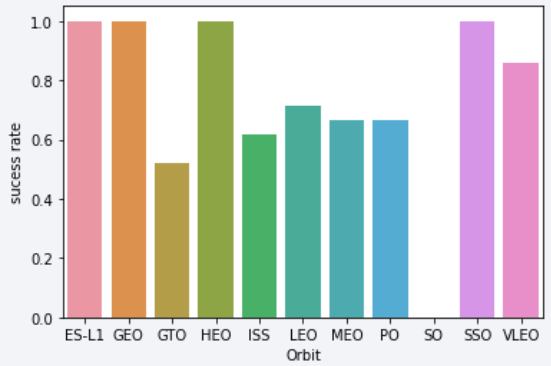
Data Wrangling

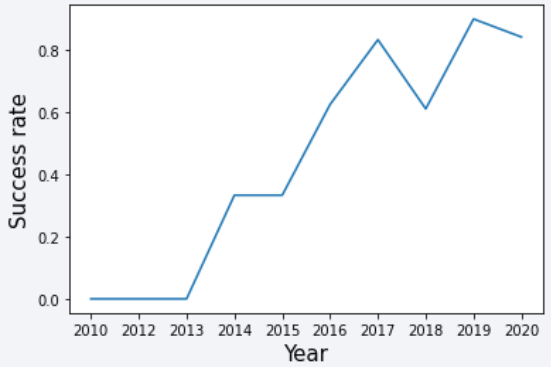
- conducted exploratory data analysis to identify the training labels.
- analyzed the number of launches at each site and examined the frequency and types of orbits
- Derived the landing outcome labels from the outcome column and exported the results to a CSV file
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab3_Data%20Wrangling.ipynb

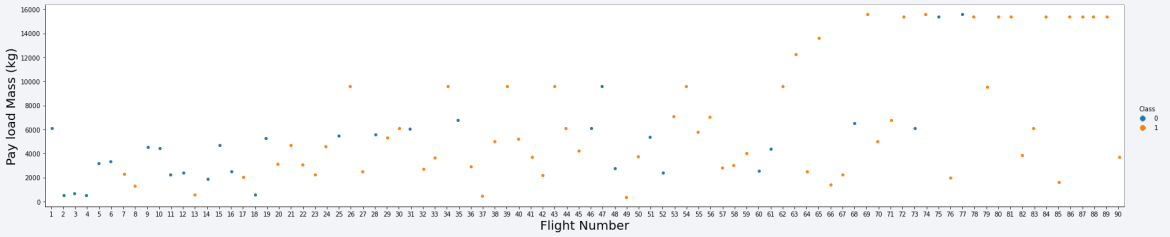


some common orbit types

EDA with Data Visualization







- The total success launches from each launch site
- The correlation between payload mass and mission outcome (success or failure) for each launch site
- Trend by year
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab5_Data_visual.ipynb

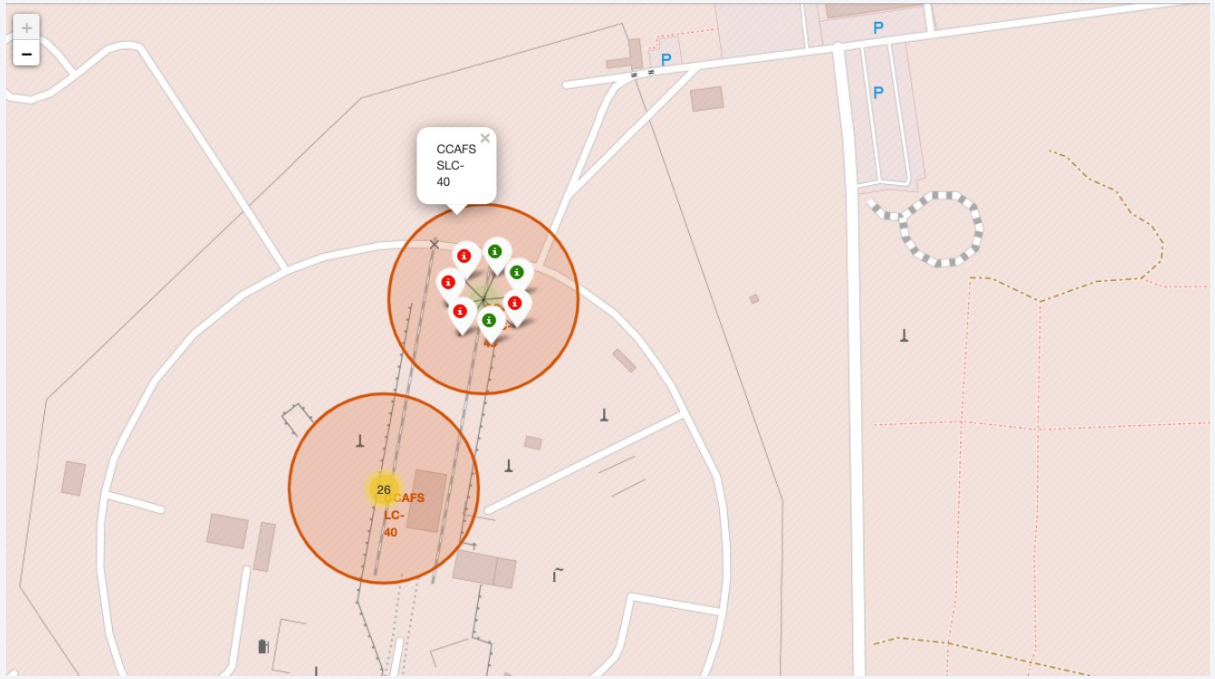
EDA with SQL

- Retrieve the names of the unique launch sites used in the space missions.
- Identify the booster versions that have carried the maximum payload mass.
- Count the total number of successful and failed mission outcomes.
- List the names of the boosters that have successfully landed on a drone ship and have a payload mass within a specific range.
- Rank the successful landing outcomes within a given date range in descending order.
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab4_EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

- Markers were added for launch sites and for the NASA Johnson Space Center
- Circles were added for the launch sites.
- Lines were added to show the distance
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab6_site_location_folium.ipynb



Build a Dashboard with Plotly Dash

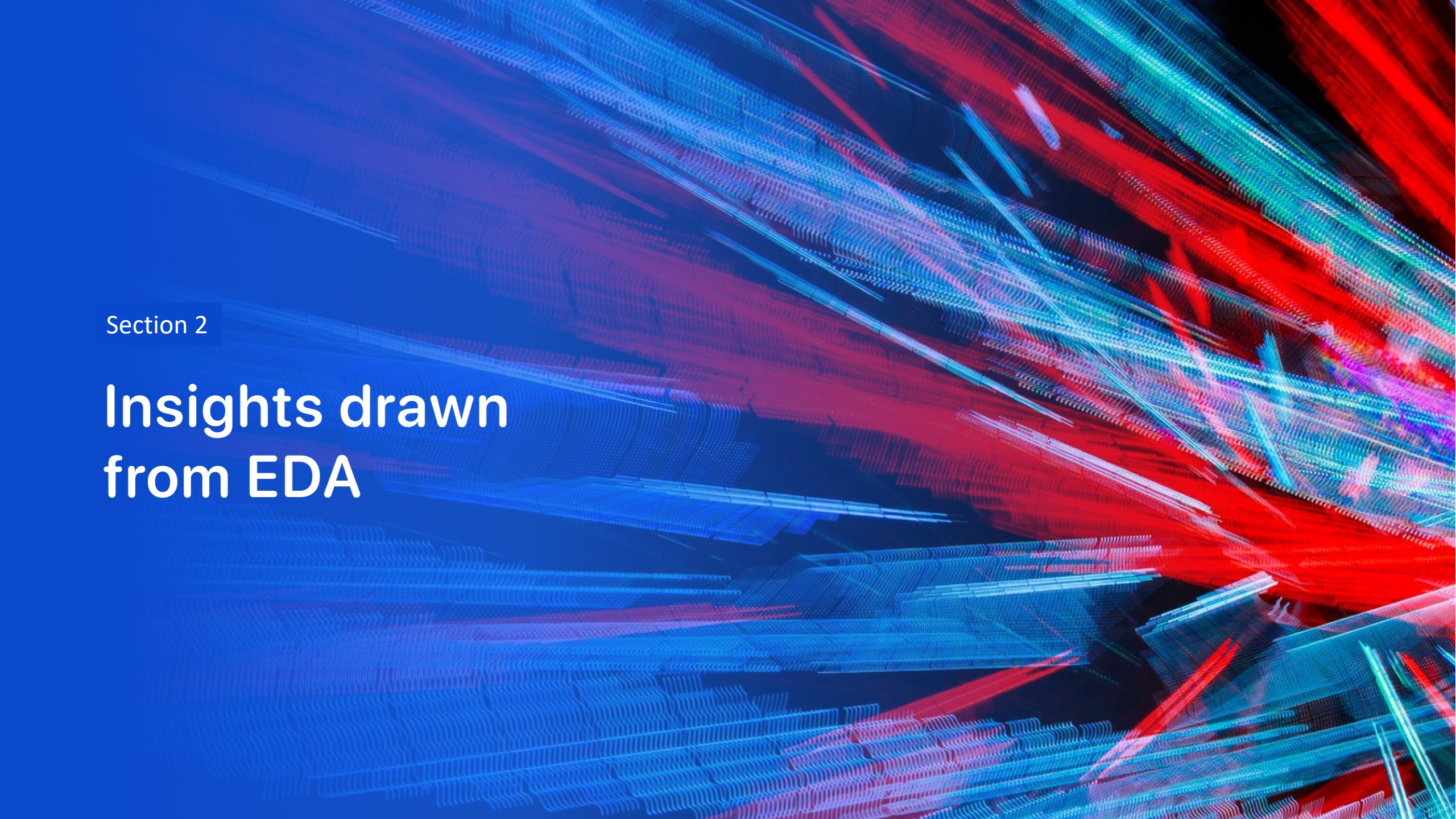
- developed an interactive dashboard using Plotly Dash.
- Created pie charts displaying the total number of launches by specific sites.
- generated scatter plots illustrating the relationship between Outcome and Payload Mass (kg) for different booster versions.
- https://github.com/llouislouis/iBM_DataScience_Capstone_SPACE_X/blob/main/dash_app.py

Predictive Analysis (Classification)

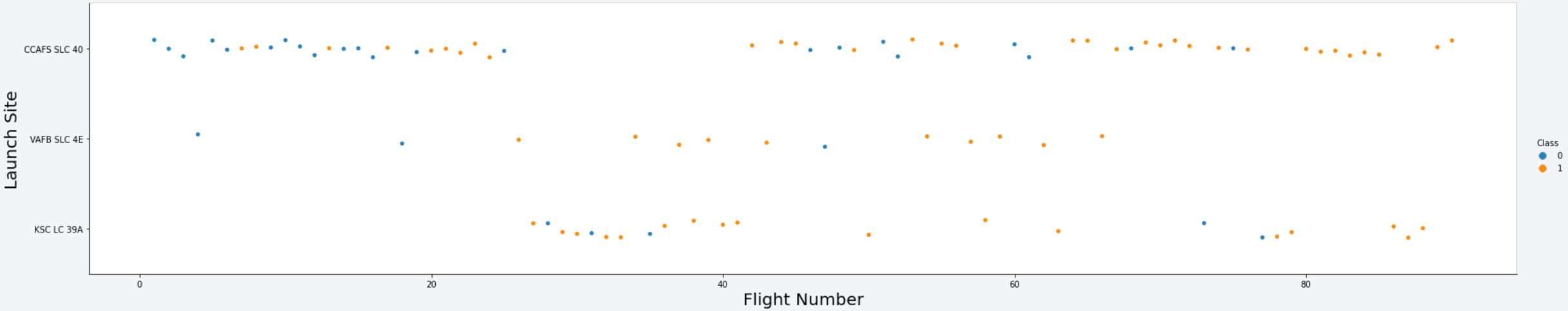
- Loaded and transformed data using numpy and pandas, then split into training and testing sets.
- Built various machine learning models and tuned hyperparameters with GridSearchCV.
- Identified the best-performing classification model.
- https://github.com/llouisIouis/iBM_DataScience_Capstone_SPACE_X/blob/main/Lab7_ML.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

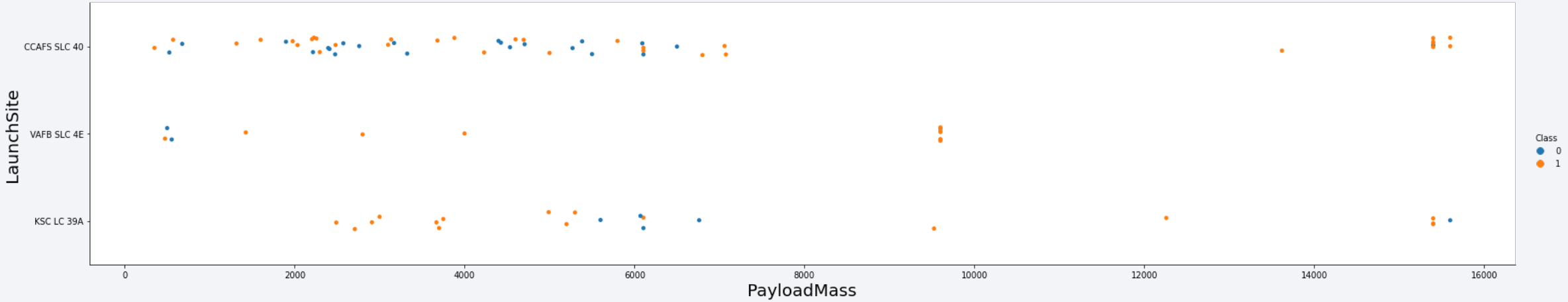

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- **Explanation**: We can see from the scatter plot that as flight number increases, there are more successful first stage landing. With small flight numbers, launches happens more in the site CCAFS SLC 40 and with much lower success rate. Although there are less launches in VAFB SLC 4E and KSC LC 39A, higher success rate can be seen in these two sites.

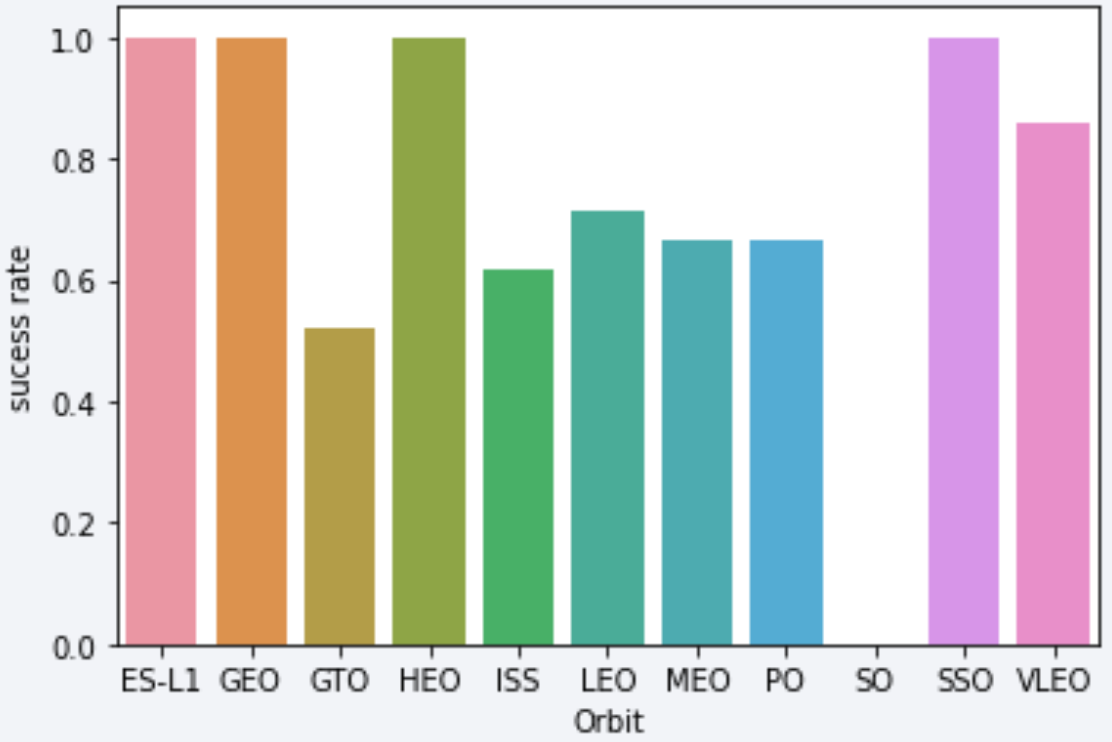
Payload vs. Launch Site



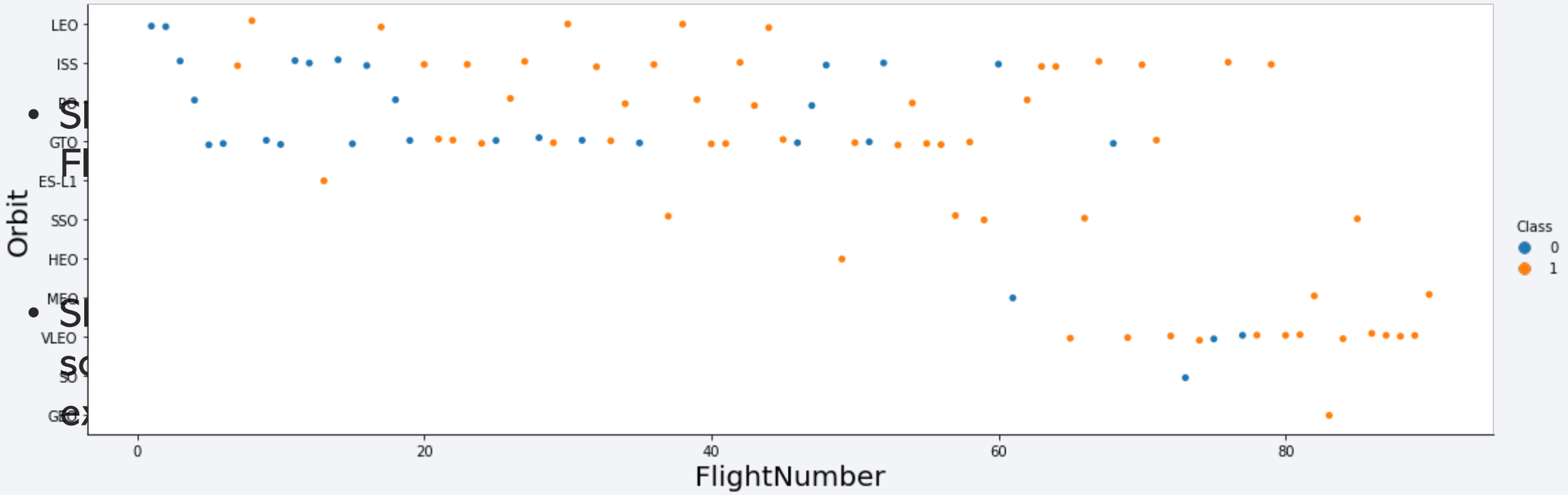
- **Explanation:** With higher Payload the success rate is much higher. And in KSC LC39A launchsite we can see much higher success rate with low Payload whereas this rate is mucher lower in CCAFS SLC 40 launchsite. Besides, there no rockets launched in VAFB-SLC for Payload greater than 10000. Furthermore, with Payload more than 9500, we can see very high success rate overall.

Success Rate vs. Orbit Type

- From the Bar Plot we can see for Orbit type ES-L1, GEO, HEO, and SSO have the highest success rate, which is 100%. And we also find in SO orbit, the rate is zero.

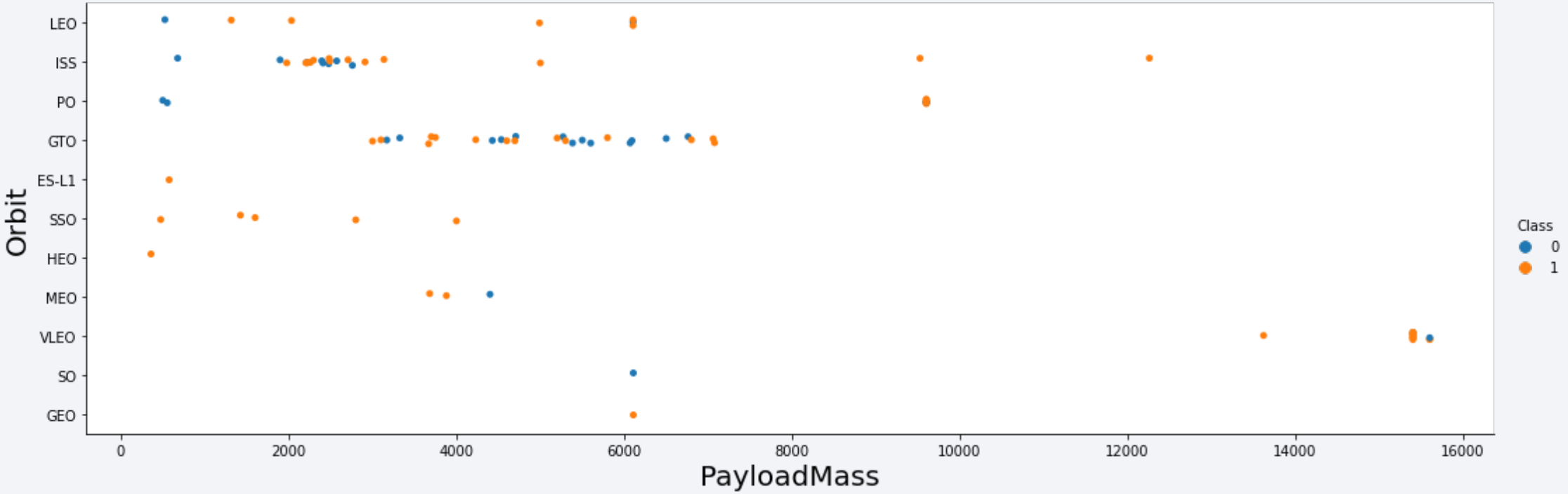


Flight Number vs. Orbit Type



- **Explanation:** In ES-L1, GEO, HEO, and SSO orbits, all launches are successful. There is clear relationship between flight number and success rate in LEO orbit since as flightnumber increases, the success rate increases. In contrast, there is no such obvious relationship in GTO orbit.

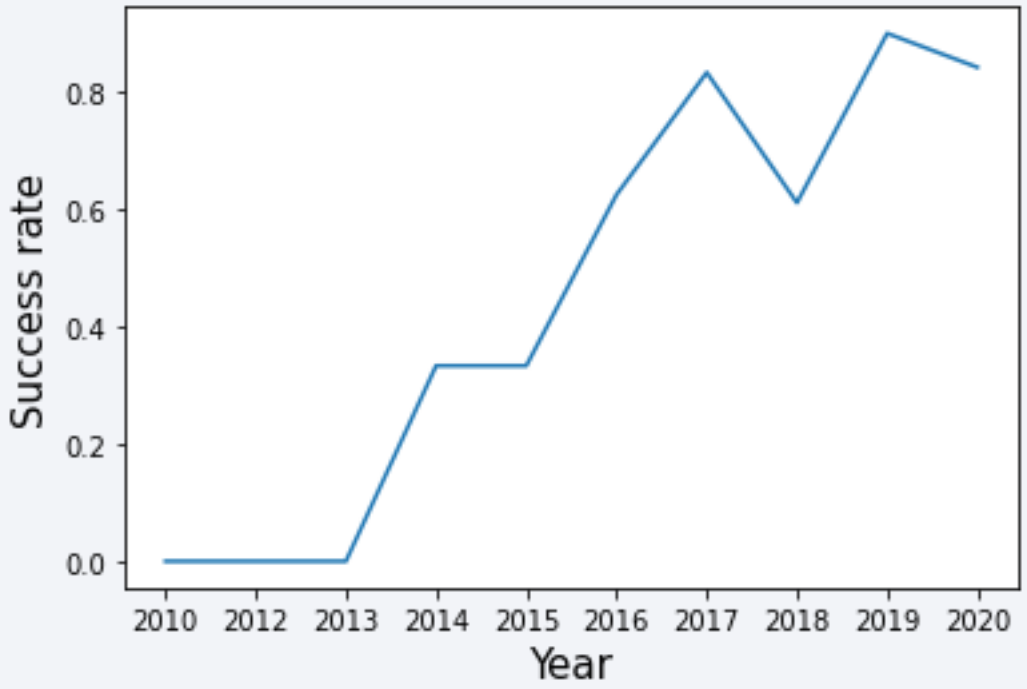
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- **Explanation**: The observe show that the success rate since 2013 kept increasing till 2020



All Launch Site Names

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

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

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

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

Utilized the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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 (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: these 5 launches happened in LEO orbit, and four of them were from customer NASA.

Total Payload Mass

```
[9]  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA%'
```

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

sum(PAYLOAD_MASS__KG_)
99980

Explanation: The total payload carried by boosters from NASA is **99980**.

Average Payload Mass by F9 v1.1

```
[ ]  %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

Explanation: the average payload mass carried by booster version F9 v1.1 is **2534.67**.

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)"
```

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

min(Date)
01-05-2017

Explanation: the first successful landing outcome on ground pad is **01-05-2017**.

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_>4000
    and PAYLOAD_MASS__KG_ < 6000
```

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

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

Explanation: names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select distinct "Mission_Outcome" from SPACEXTBL
```

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

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

```
%%sql

select count(*) from SPACEXTBL
where "Mission_Outcome" like "Success%"
```

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

count(*)
100

```
%%sql

select count(*) from SPACEXTBL
where "Mission_Outcome" like "Failure%"
```

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

count(*)
1

Explanation:

- the total number of **successful** mission outcomes is **100**
- the total number of **failure** mission outcomes is **1**

Boosters Carried Maximum Payload

```
%%sql

select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

Names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql

select substr(Date, 4, 2) as Month, Booster_Version, Launch_Site from SPACEXTBL
where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)"
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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.

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

```
%%sql

select "Landing _Outcome",
    count("Landing _Outcome") as landings
from SPACEXTBL
where Date >= "04-06-2010" and Date <= "20-03-2017"
group by "Landing _Outcome"
order by landings desc
```

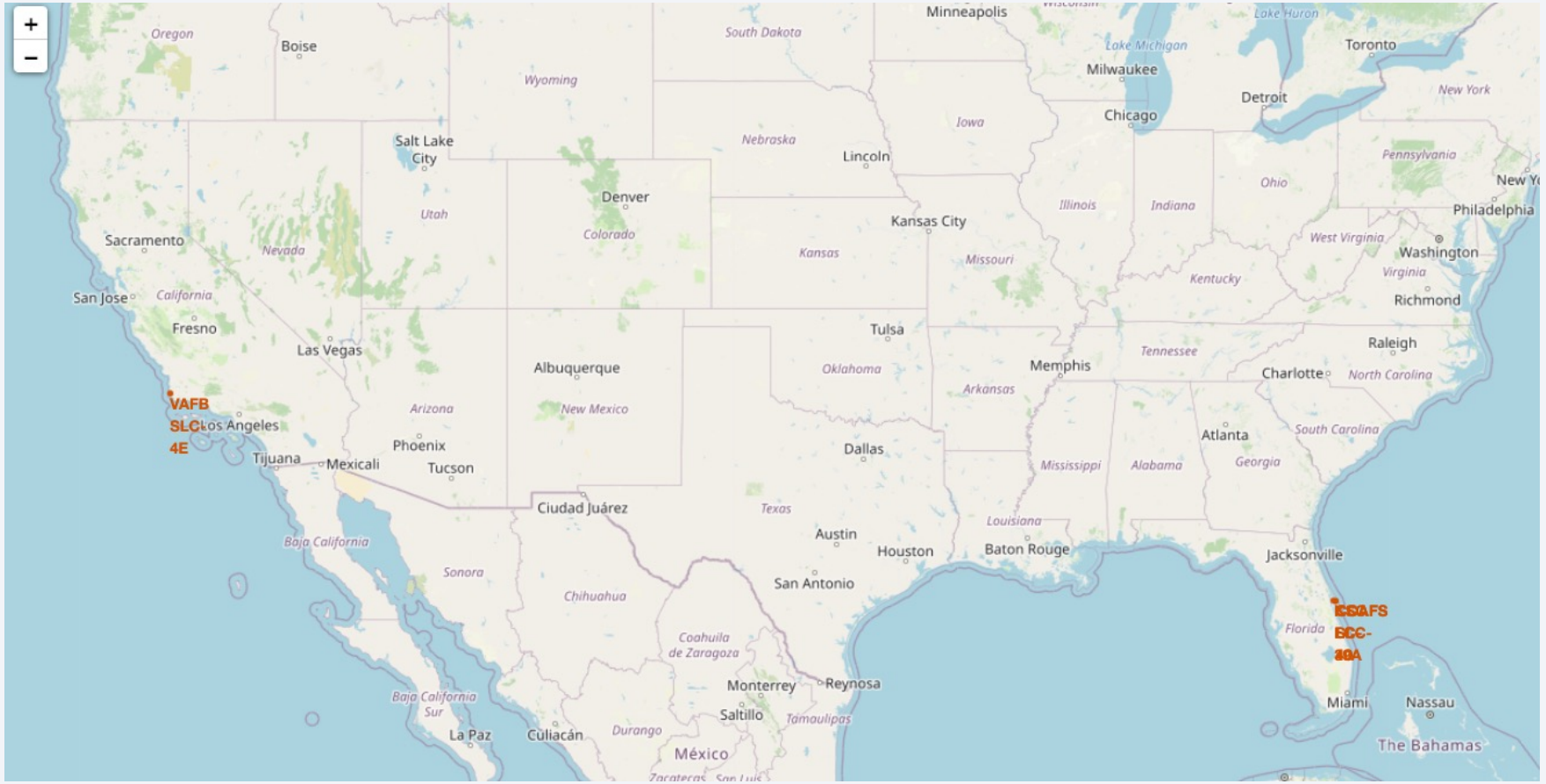
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

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

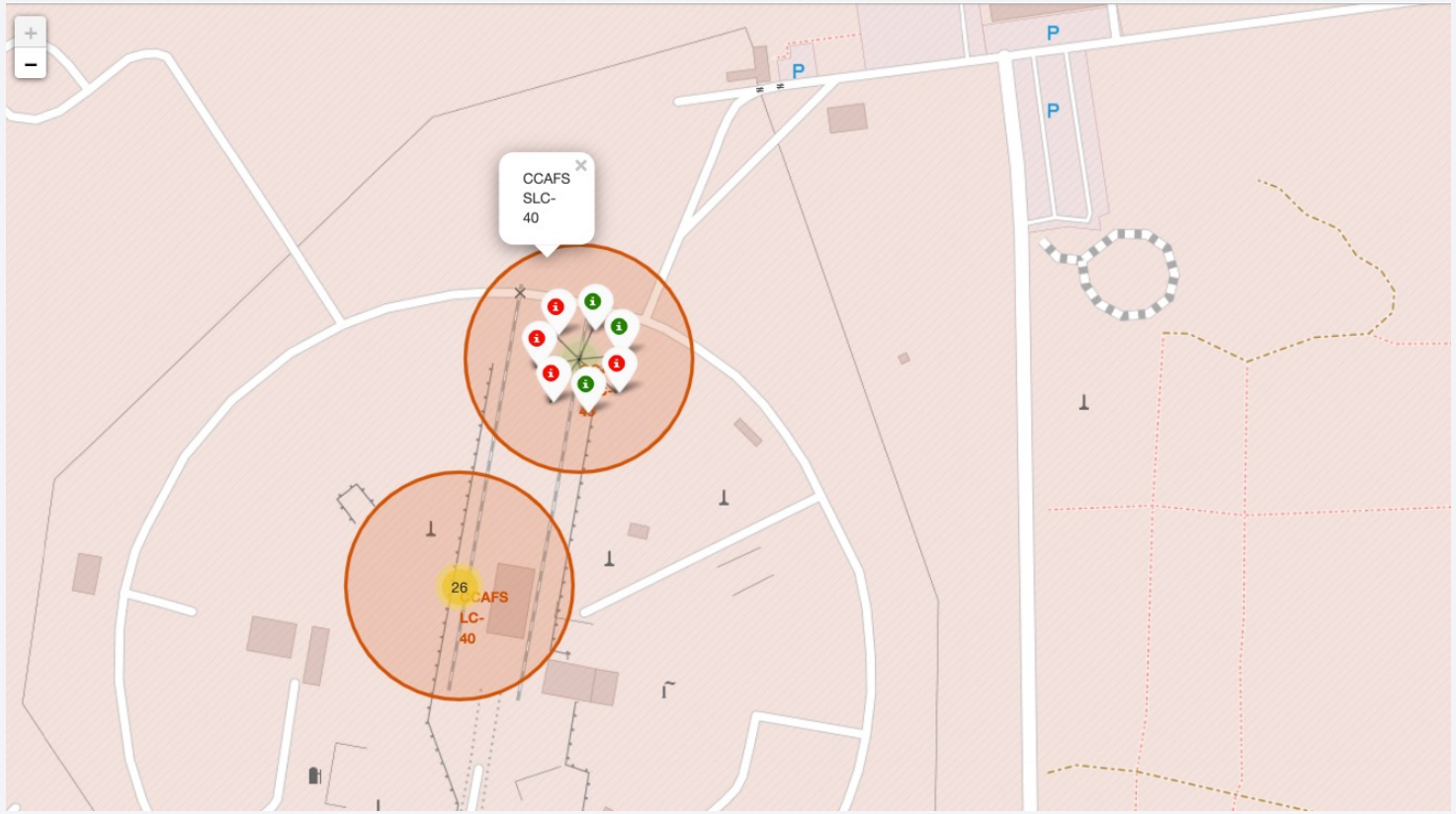

Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>



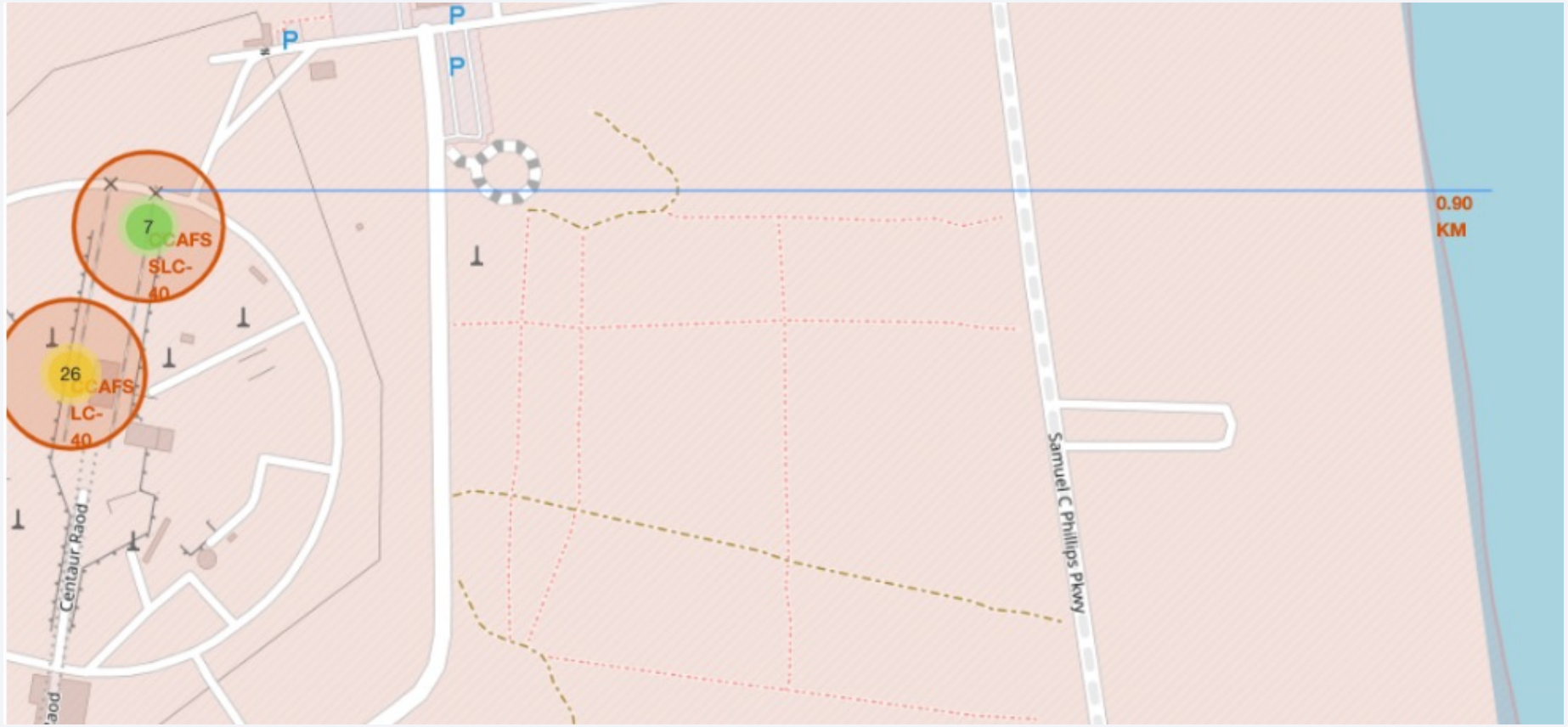
The SpaceX launch sites are show in the USA coasts.

<Folium Map Screenshot 2>

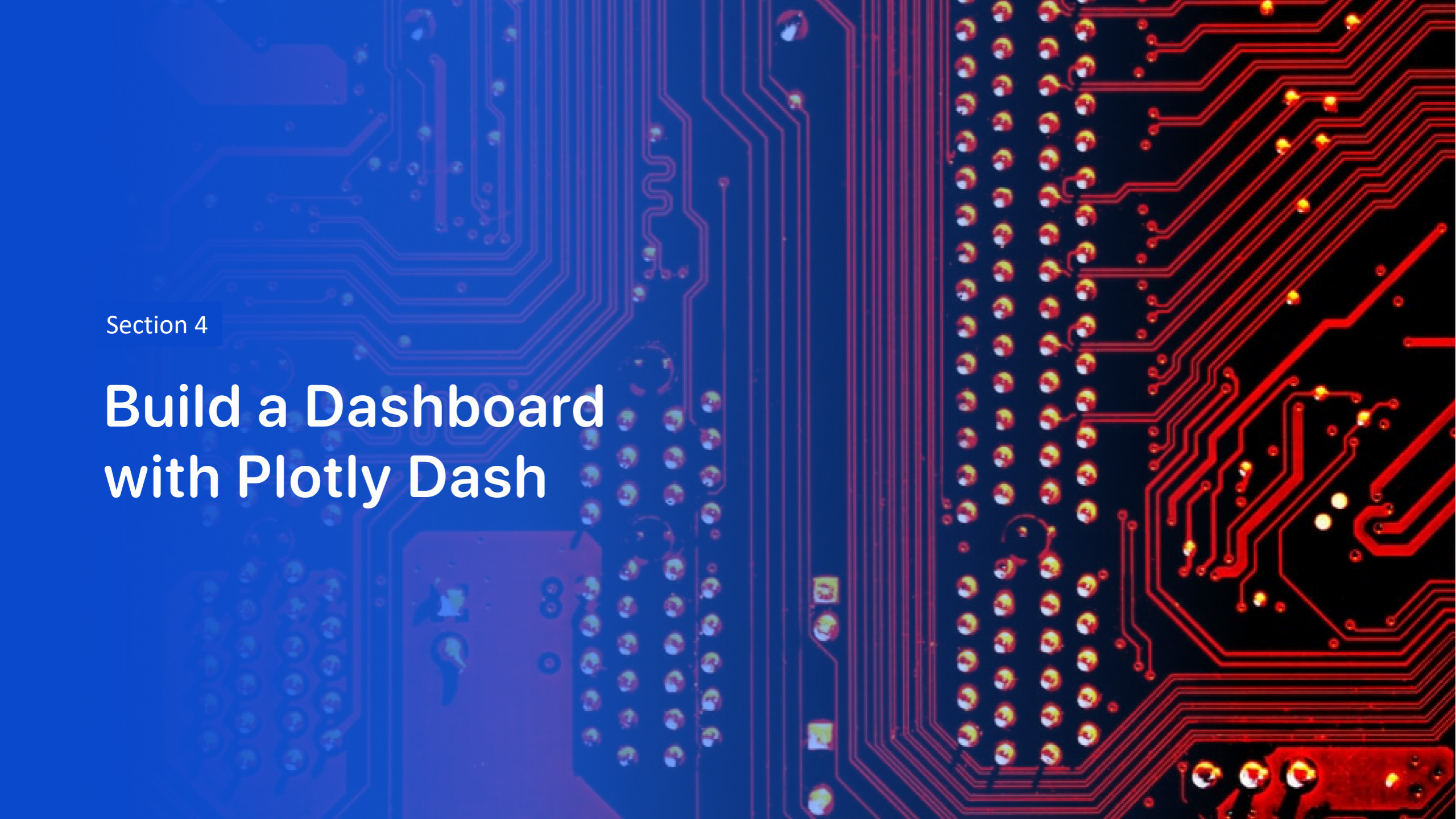


The succeeded launches and failed launches for each site on map

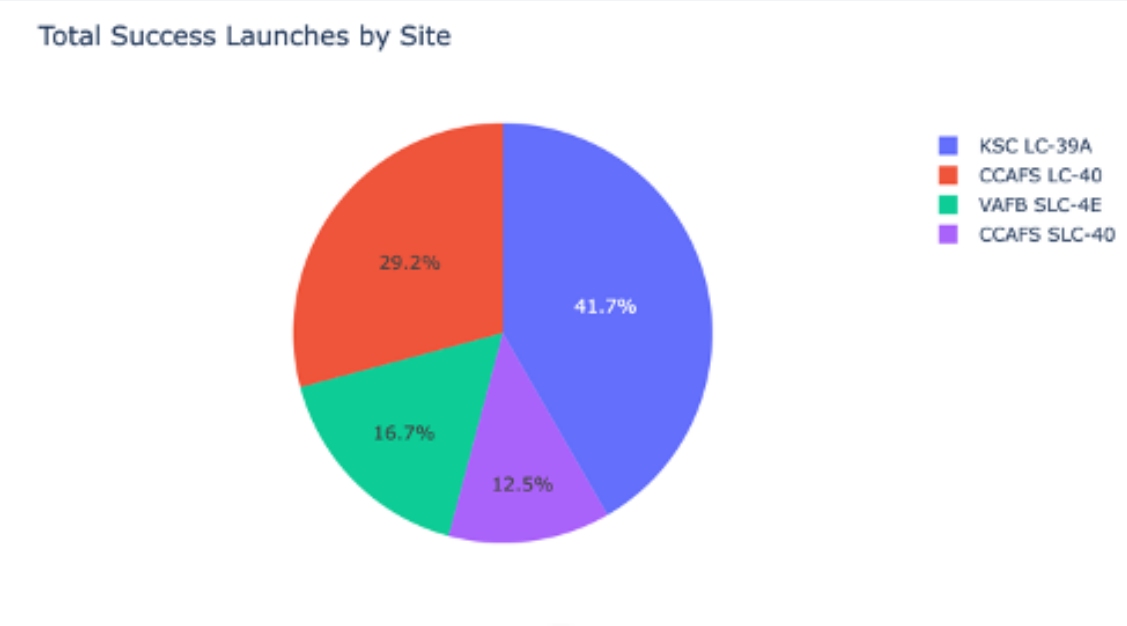
<Folium Map Screenshot 3>



Able to proximities a distance by line between a launch site to its closest city, railway, highway, etc.


Section 4
Build a Dashboard with Plotly Dash

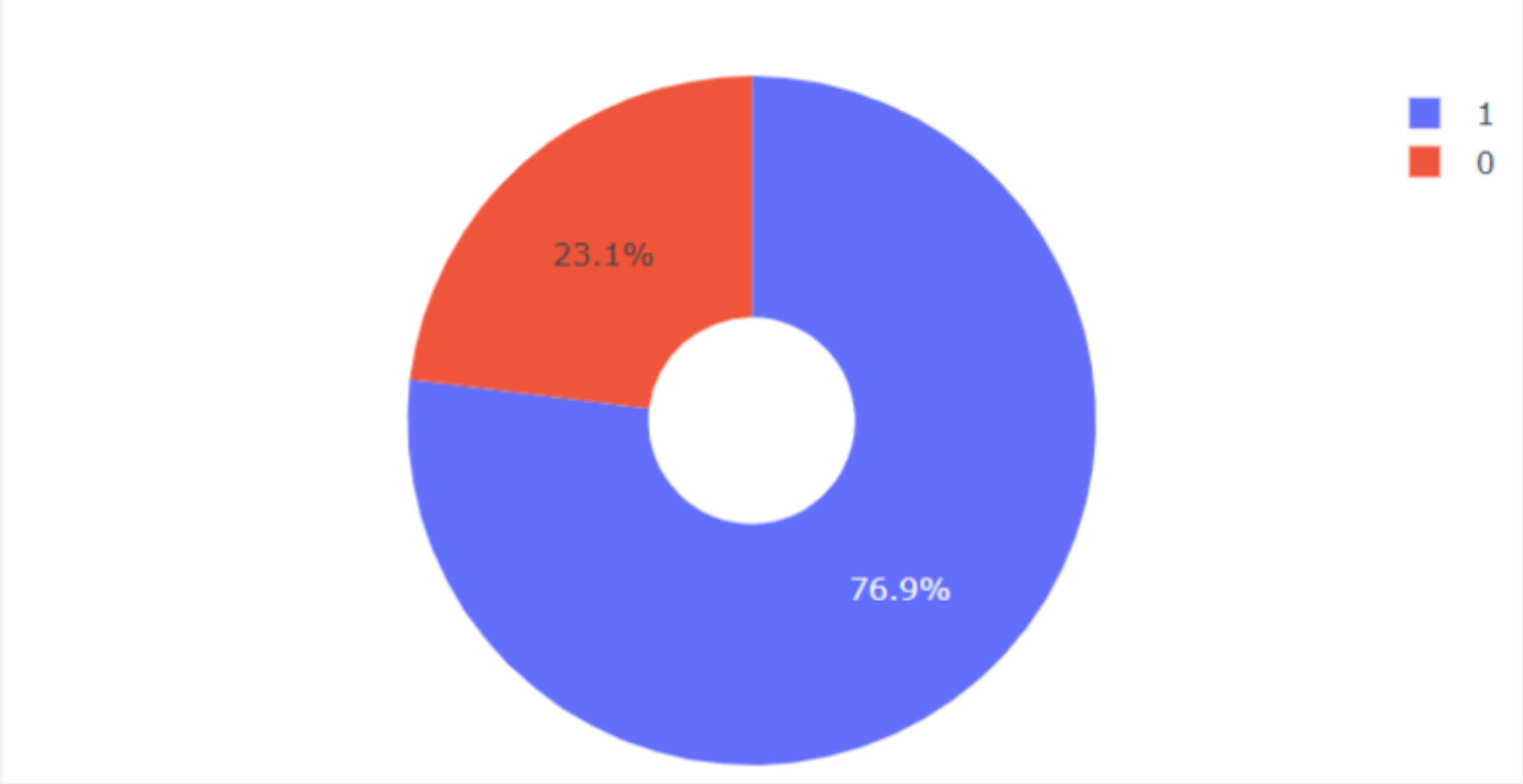
Pie chart showing the success percentage achieved by each launch site



Total Success Launches for All Sites is

- CCAFS LC-40: 29.2%
- VAFB SLC-4E: 16.7%
- KSC LC-39A: 41.7%
- CCAFS SLC-40: 12.5%

Pie chart showing the Launch site with the highest launch success ratio

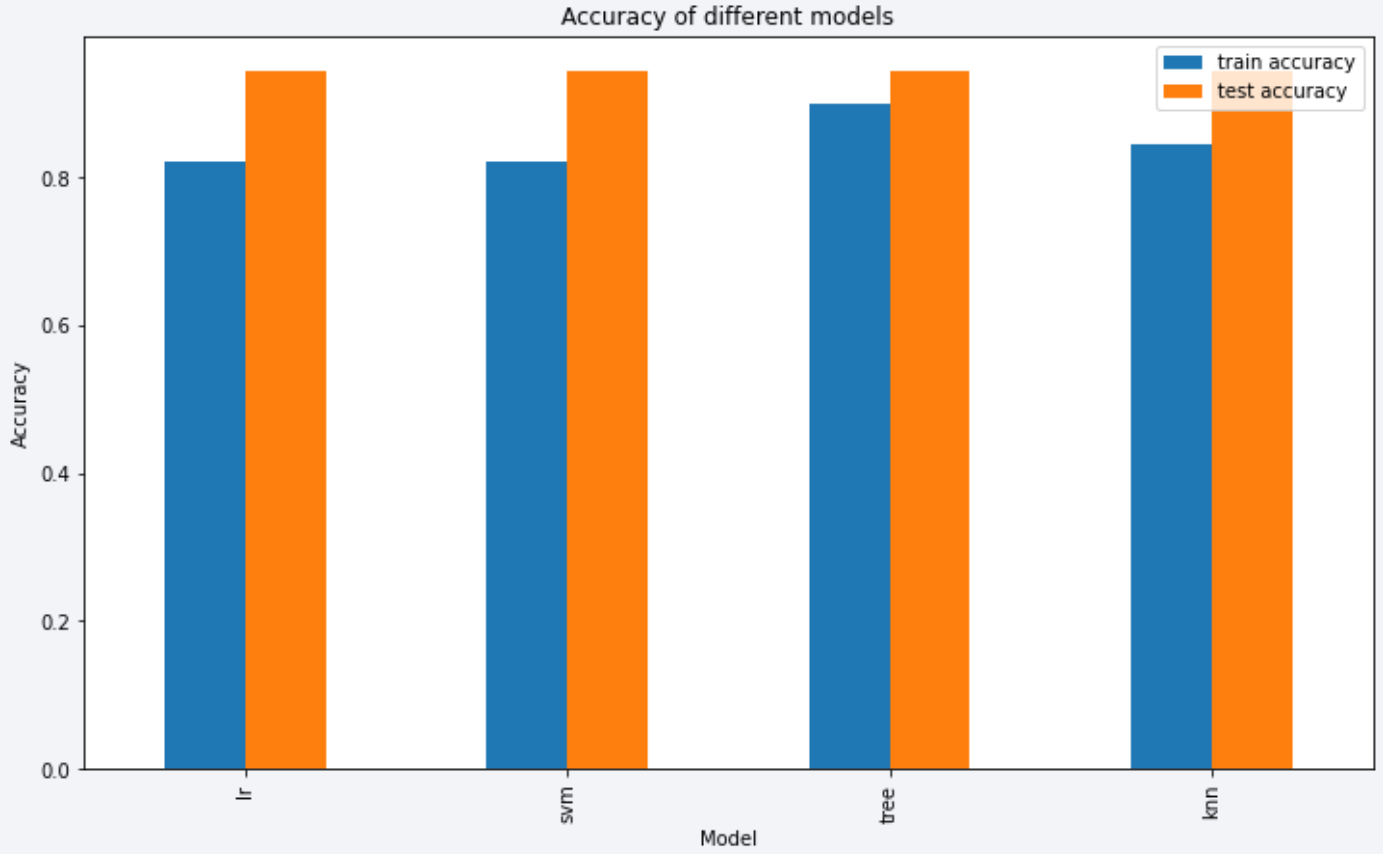


Payload vs. Launch Outcome

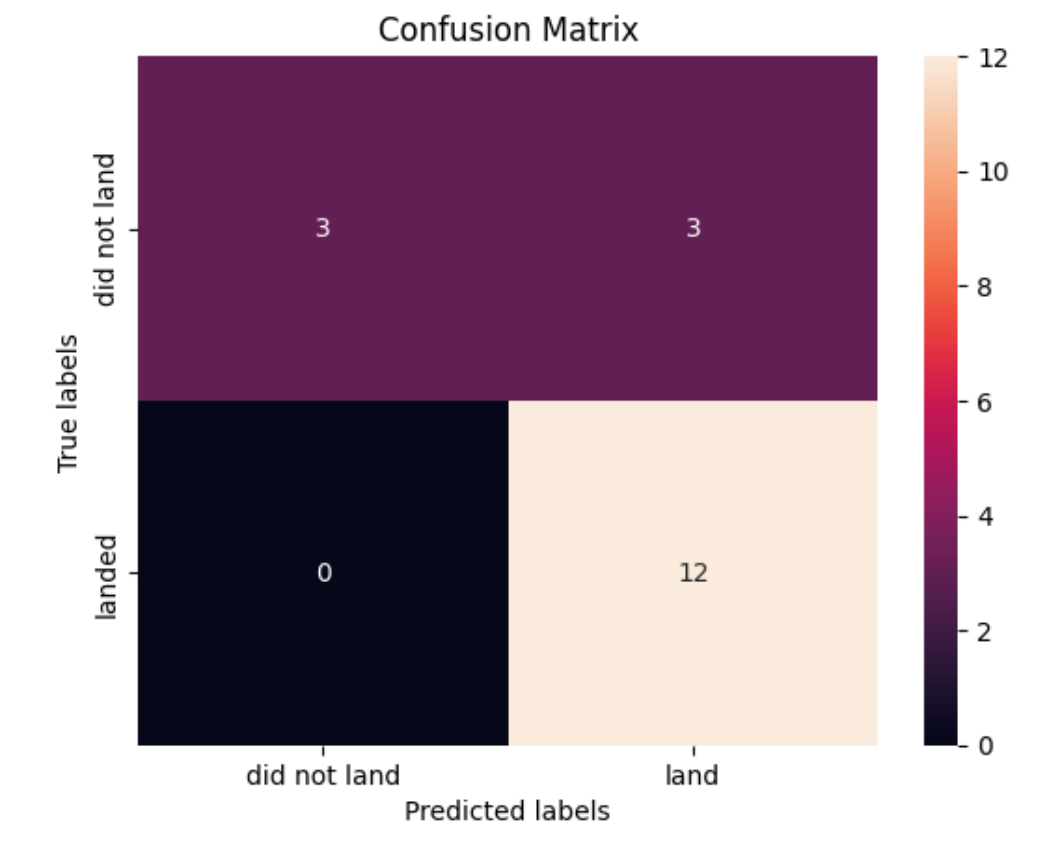
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- We trained four models using GridSearchCV, with the Decision Tree model performing best on the test dataset, though it may have issues with false positives affecting bid estimations for rocket launches.

- The dataset comprises 90 rows and 83 columns, split 80/20 into 72 training rows and 18 testing rows.

- Key findings includes higher flight amounts at a launch site correlate with greater success rates; launch success rates increased from 2013 to 2020; orbits ES-L1, GEO, HEO, SSO, and VLEO had the highest success rates; and KSC LC-39A had the most successful launches.

- The project aimed to predict Falcon 9 first stage landing outcomes to determine launch costs, with features like payload mass and orbit type influencing mission outcomes; the Decision Tree classifier was the best predictive model among those tested.

Thank you!