IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AKASH SHANKAR JADHAV
05/03/2025

Outline

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

Executive Summary

- **Project Objective:** Predict whether SpaceX's Falcon 9 rockets will successfully land in the first stage.
- **Cost Advantage:** SpaceX offers rockets at $62 million, significantly lower than other providers at $165 million, due to its ability to reuse the first stage (IBM).
- **Techniques Applied:**
  - **Data Collection:** Used APIs and web scraping.
  - **Data Processing:** Performed data wrangling to restructure the data.
  - **Exploratory Data Analysis (EDA):** Visualized data to uncover insights, patterns, and trends.
  - **Feature Engineering:** Preprocessed the data for machine learning models.
- **Machine Learning Models:** Applied four different algorithms for predictive modeling.
- **Best Performing Model:** DecisionTree achieved the highest score of **87.32**%

Introduction

- **Role & Objective:**
  - I am a **Data Scientist** at **SpaceY**, a rocket company founded by billionaire **Allon Mask**.
  - My job is to analyze publicly available information about **SpaceX** to predict whether the first-stage landing of their **Falcon 9** rockets will be successful.
  - This research helps SpaceY make informed bids against SpaceX, as Falcon 9's reusability gives them a competitive cost advantage.
- **Data Collection & Processing:**
  - Gathered public data from **SpaceX's API** using **HTTP requests** and **BeautifulSoup**.
  - Flattened the collected data into a **pandas DataFrame** for further analysis.

Introduction

- **Data Wrangling & Analysis:**
  - Processed and restructured data using **pandas**.
  - Conducted in-depth analysis using **SQLite** for queries.
  - Used **Matplotlib** and **Seaborn** for visualizing patterns and relationships between key features.
- **Visualization & Insights:**
  - **Folium** was used to map and analyze launch sites.
  - Built an **interactive dashboard** using **Dash** to explore:
    - How **payload** size affects landing success.
    - The impact of **orbit type** and **flight number** on success rates.

Section 1

Methodology

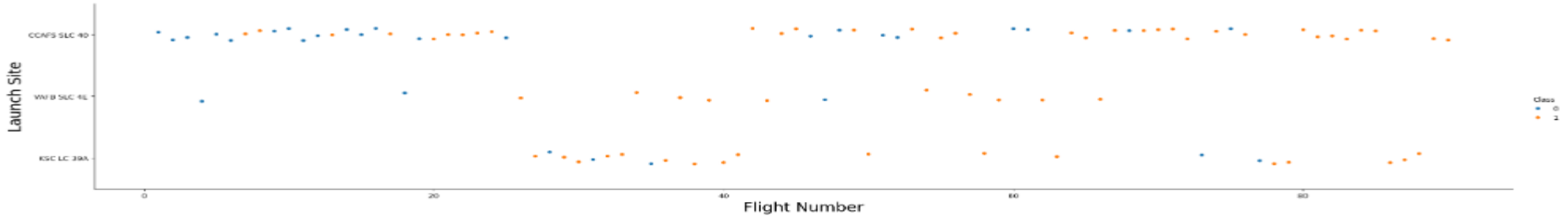
Methodology

- **Data Collection:**
  - SpaceX API
  - Web Scraping
- **Data Processing:**
  - Data Wrangling
- **Data Analysis & Visualization:**
  - SQL
  - Pandas
  - Folium (Launch Site Visualization)
  - Matplotlib & Seaborn (Data Exploration & Patterns)
- **Machine Learning:**
  - Predicting the success of the first-stage landing.

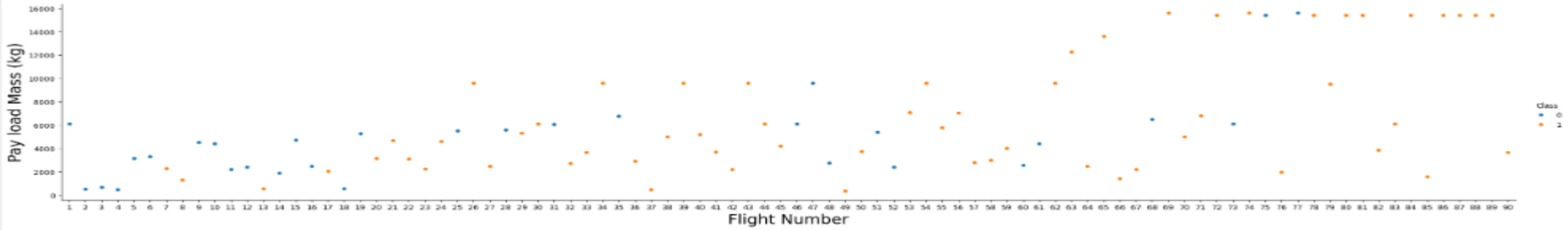
Methodology

METHODOLOGY: EDA & Interactive Visual Analytics (1/3)

[6]:
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

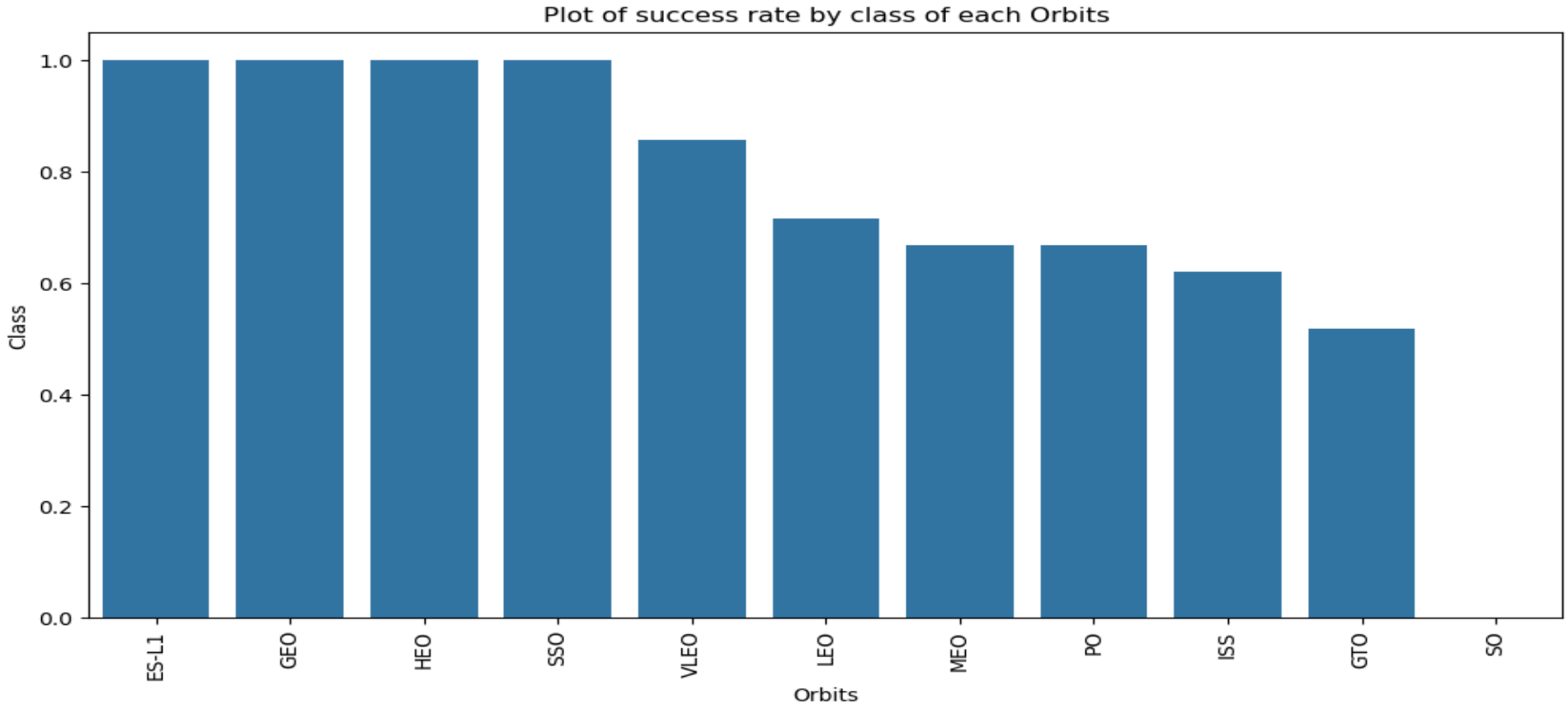


[4]:
```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Pay load Mass (kg)", fontsize=20)
plt.show()
```



Methodology

METHODOLOGY: EDA & Interactive Visual Analytics (2/3)



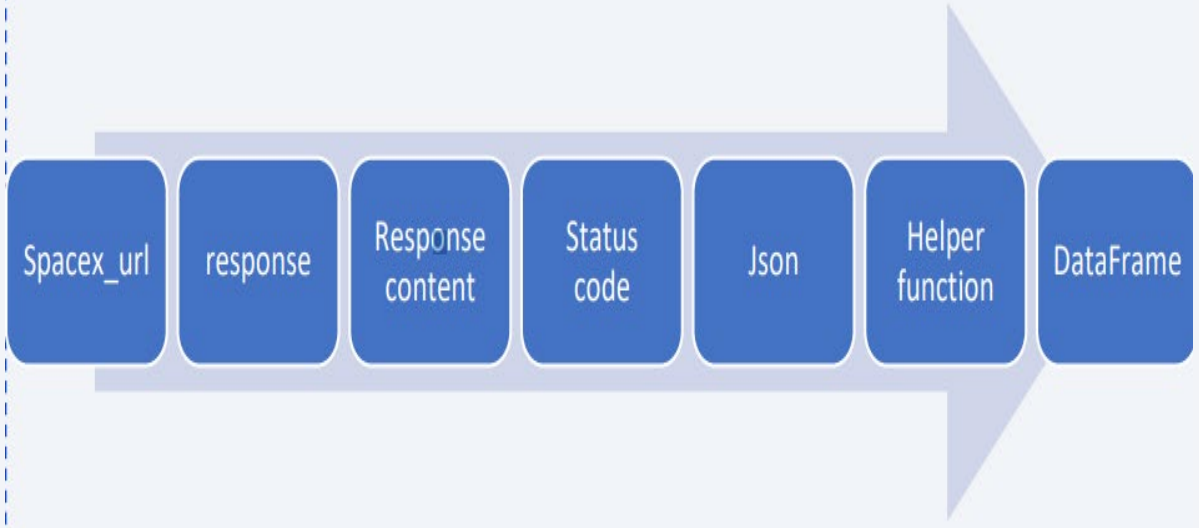
Methodology

METHODOLOGY: EDA & Interactive Visual Analytics

- Larger the flight amount at a launch site, the greater the success rate at a launch site.
- ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

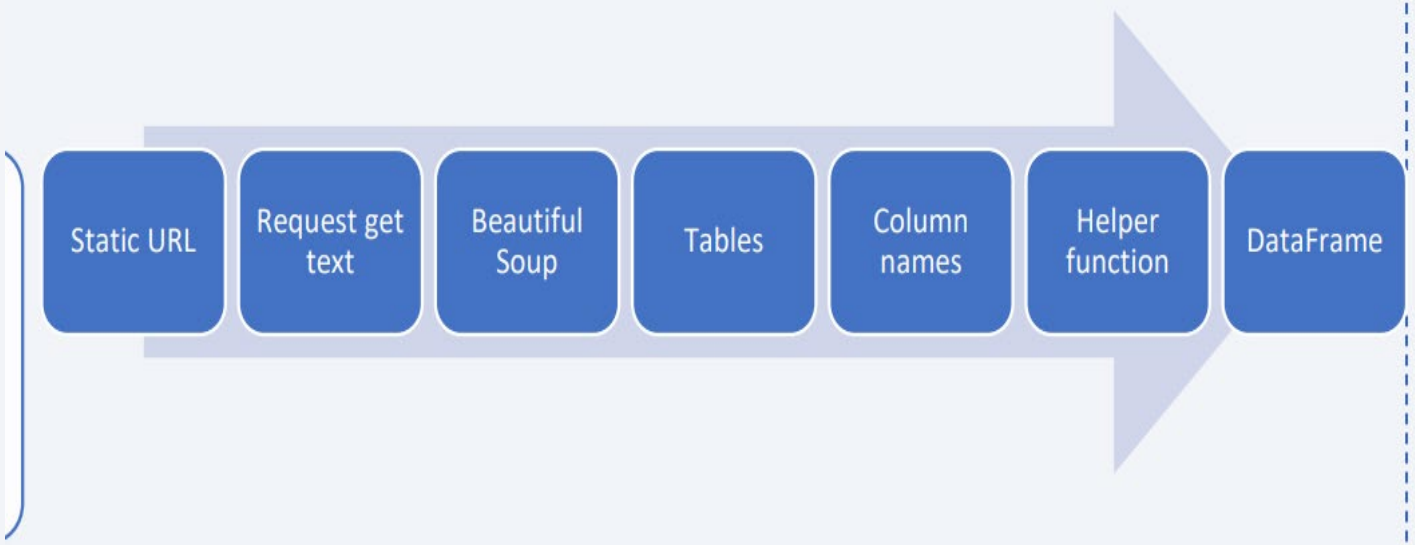
Data Collection – SpaceX API

- SPACEX URL-> RESPONSE-> JSON-> DATAFRAME
- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/1_spacex_datacollection_api.ipynb



Data Collection - Scraping

- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/2-data-collection-webscraping.ipynb



Data Wrangling

- The cleaned data was imported, and I first analyzed the percentage of missing values in the **LaunchingPad** column, as it was the only column still containing missing values, indicating instances where "no LaunchingPad" was used.
- Next, I examined the data types of each column, identifying four different types: **int64, object, float64, and bool**.
- Further analysis included evaluating the value counts of **LaunchSite**, revealing that **Cape Canaveral Space Launch Complex 40** and **VAFB SLC 4E** had the highest count of **55**.
- Additionally, I created a new feature called **"class"** from the **outcome** column. Any outcome containing **"False"** or **"None"** was categorized as **bad** and assigned a value of **0**, while all other outcomes were considered **good** and assigned a value of **1**.
- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/3_data_wrangling.ipynb

EDA with Data Visualization

OPEN FOR REFERENCE:

https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/5_EDA_with_Data_Visualization.ipynb

EDA with SQL

- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/4_EDA_WITH_SQL.ipynb

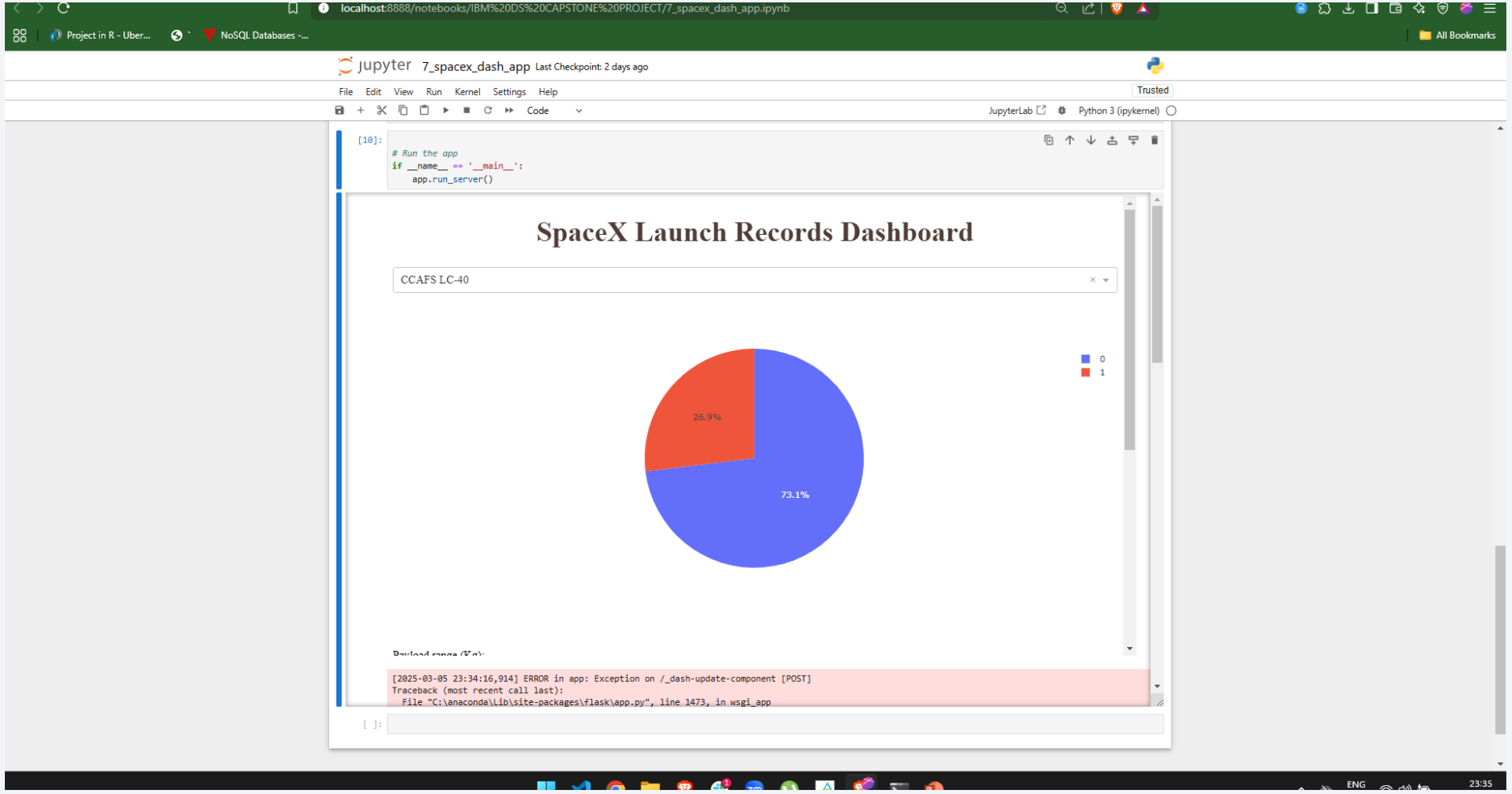
Build an Interactive Map with Folium

- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/6_launch_site_location_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- https://github.com/akash-jae/IBM_DATA-SCIENCE_FALCON9_SPACEX/blob/main/7_spacex_dash_app.ipynb

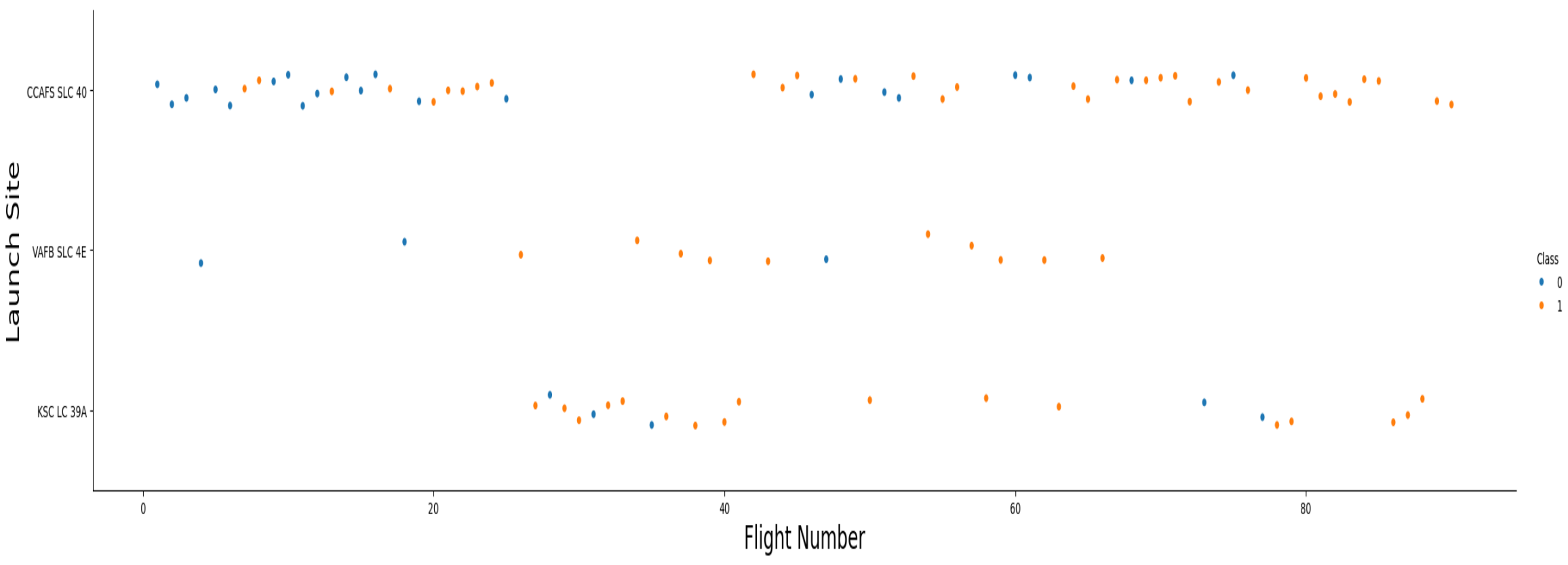
Build a Dashboard with Plotly Dash



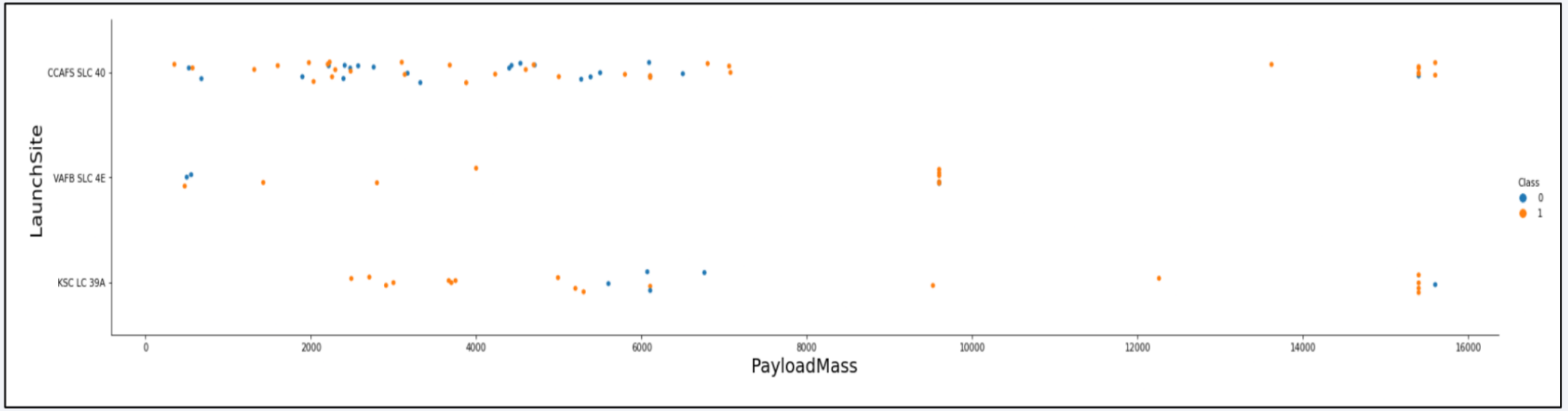
Section 2

Insights drawn from EDA

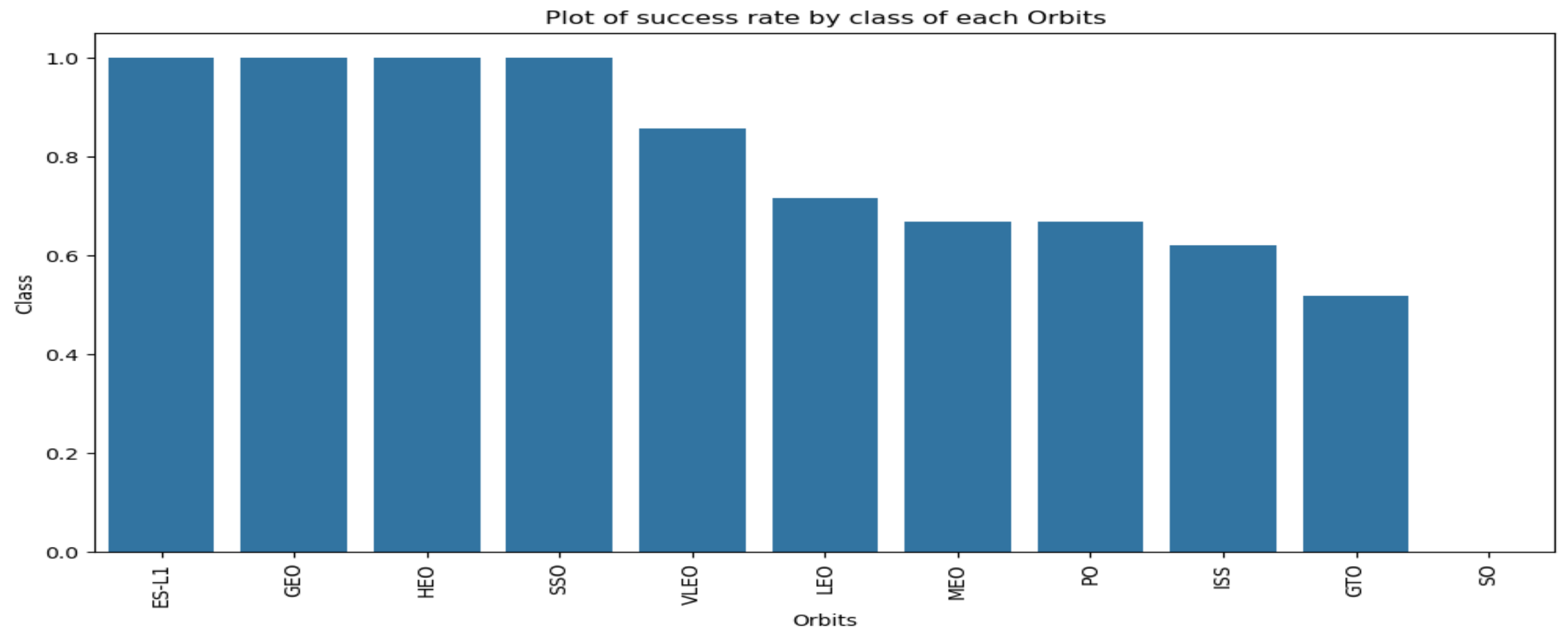
Flight Number vs. Launch Site



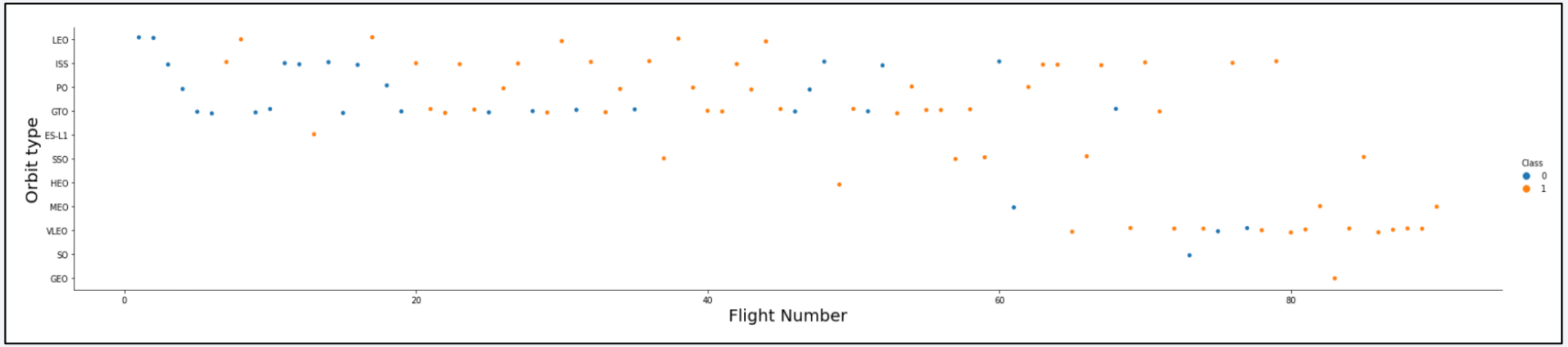
Payload vs. Launch Site



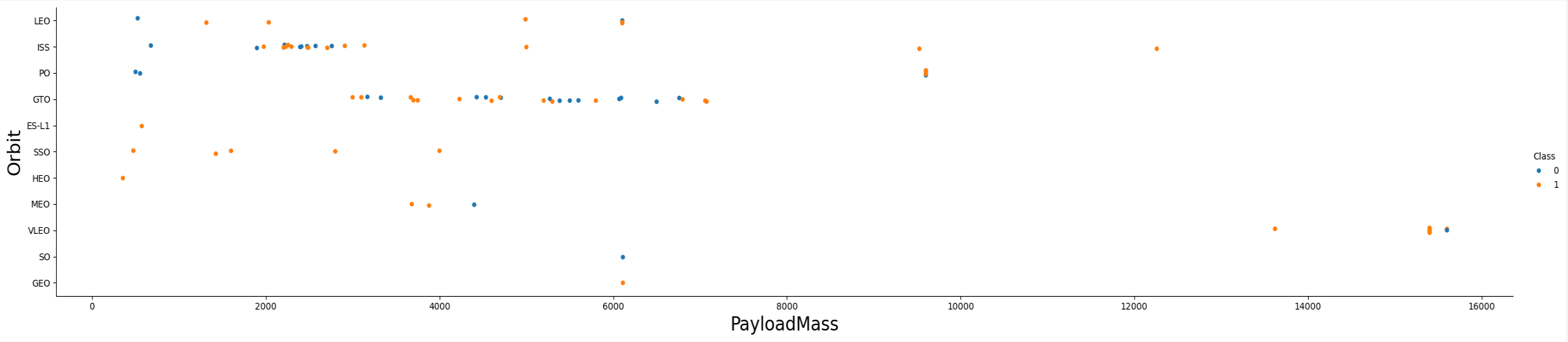
Success Rate vs. Orbit Type



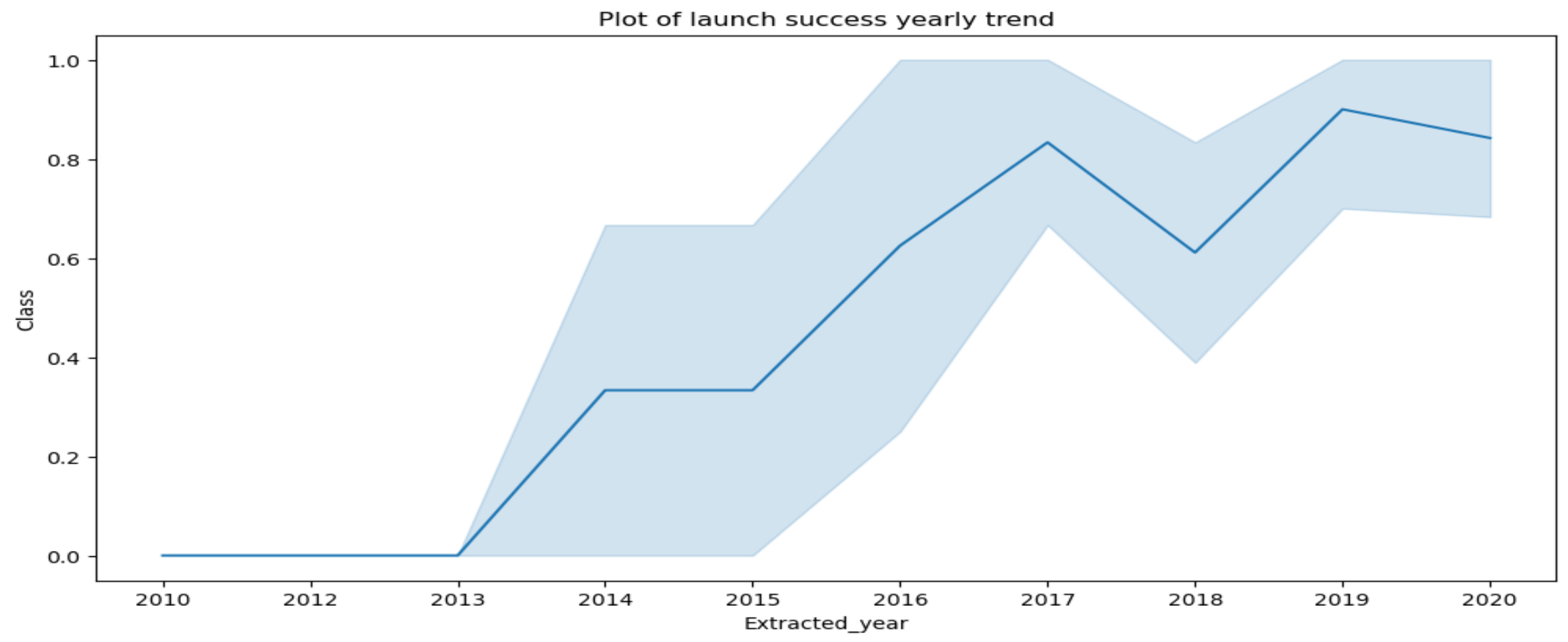
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- CCAFS LC-40: Space Launch Complex 40 formerly Launch Complex 40 (LC-40) is an orbital launch pad located in northern Cape Canaveral , Florida (Wikipedia).

- VAFB SLC-4E: Vandenberg AFB Space Launch Complex 4 is a launch and landing site at Vandenberg Space Force Base, California, U.S. It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 for SpaceX landings(Wikipedia).

- KSC LC-39A: Kennedy Space Center Launch Complex 39A Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida(Wikipedia).

Launch Site Names Begin with 'CCA'

Task 2

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

In [13]:

```
%%sql
select *
from SPACEXTBL
where LAUNCH_SITE like "CCA%"
limit 5;
```

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

Out[13]:

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

Total Payload Mass

Task 3

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

In [16]:
```
%%sql
select Customer, sum(PAYLOAD_MASS__KG_) as Total_NASA_CRS_mass
from SPACEXTBL
where Customer = "NASA (CRS)";
```

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

Out[16]:

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

Task 4

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

In [20]:
```
%%sql
select Booster_Version, avg(PAYLOAD_MASS__KG_) as avg_Booster_versionF9_v1_1
from SPACEXTBL
where Booster_Version = "F9 v1.1";
```

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

Out[20]:

Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

First Successful Ground Landing Date

F9 v1.1 2928.4

Task 5

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

Hint:Use min function

In [23]:
```
%%sql
select  Mission_Outcome, min(Date) as Date_First_Succ_Land
from SPACEXTBL
where Landing_Outcome ='Success (ground pad)';
```

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

Out[23]:

Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Success 2015-12-22

Task 6

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

In [24]:
```
%%sql
select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_
from SPACEXTBL
where (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
      and Landing_Outcome = 'Success (drone ship)';
```

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

Out[24]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [27]:
```
%%sql
select Mission_Outcome, count(Mission_Outcome) as "Total (Success or failure)"
from SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Out[27]:

Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task 8

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

In [28]:
```
%%sql
select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_)
                            from SPACEXTBL);
```

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

Out[28]:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Success	15600
F9 B5 B1049.4	Success	15600
F9 B5 B1051.3	Success	15600
F9 B5 B1056.4	Failure	15600
F9 B5 B1048.5	Failure	15600
F9 B5 B1051.4	Success	15600
F9 B5 B1049.5	Success	15600
F9 B5 B1060.2	Success	15600
F9 B5 B1058.3	Success	15600
F9 B5 B1051.6	Success	15600
F9 B5 B1060.3	Success	15600
F9 B5 B1049.7	Success	15600

2015 Launch Records

Task 9

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.

In [29]:
```
%%sql
SELECT  Date, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTBL
where Landing_Outcome= 'Failure (drone ship)' and Date <= "2015-12-31";
```

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

Out[29]:

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

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

Task 10

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.

In [30]:

```
%%sql
select Landing_Outcome, count(Landing_Outcome) as "Total Count"
from SPACEXTBL
where Landing_Outcome = "Failure (drone ship)" or Landing_Outcome = "Success (ground pad)" and
Date between "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome
order by Landing_Outcome desc;
```

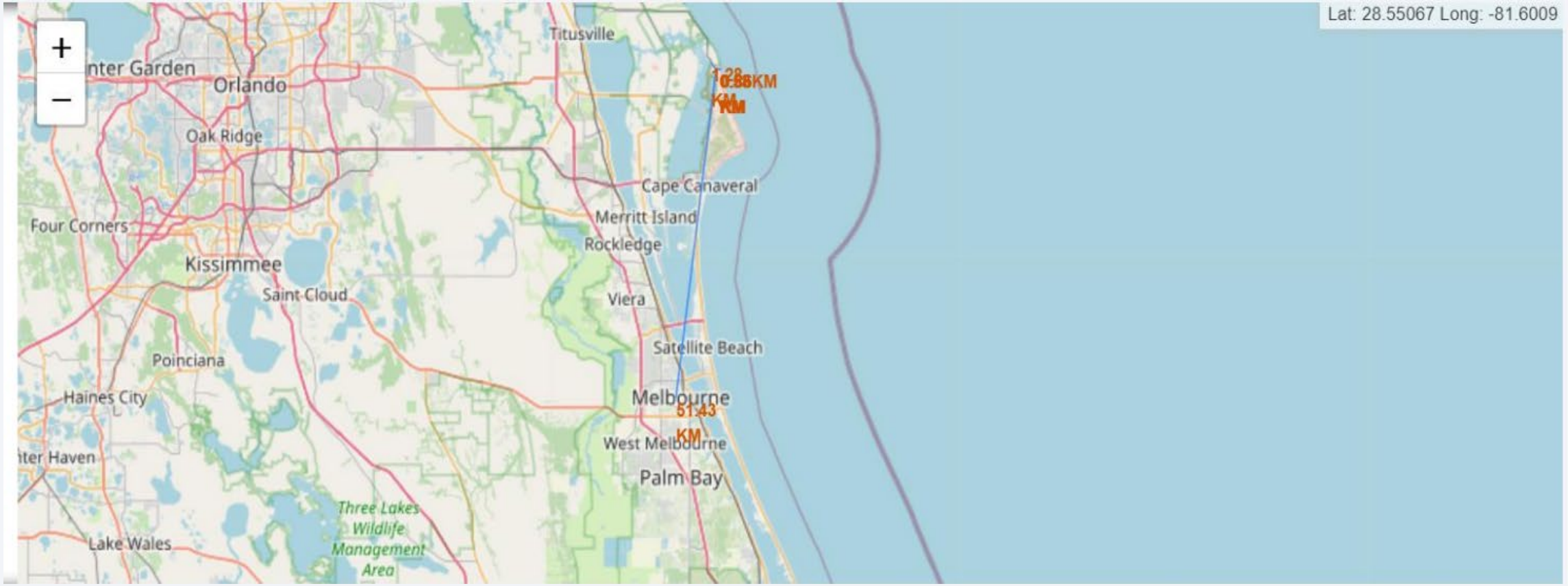
```
 * sqlite:///my_data1.db
Done.
```

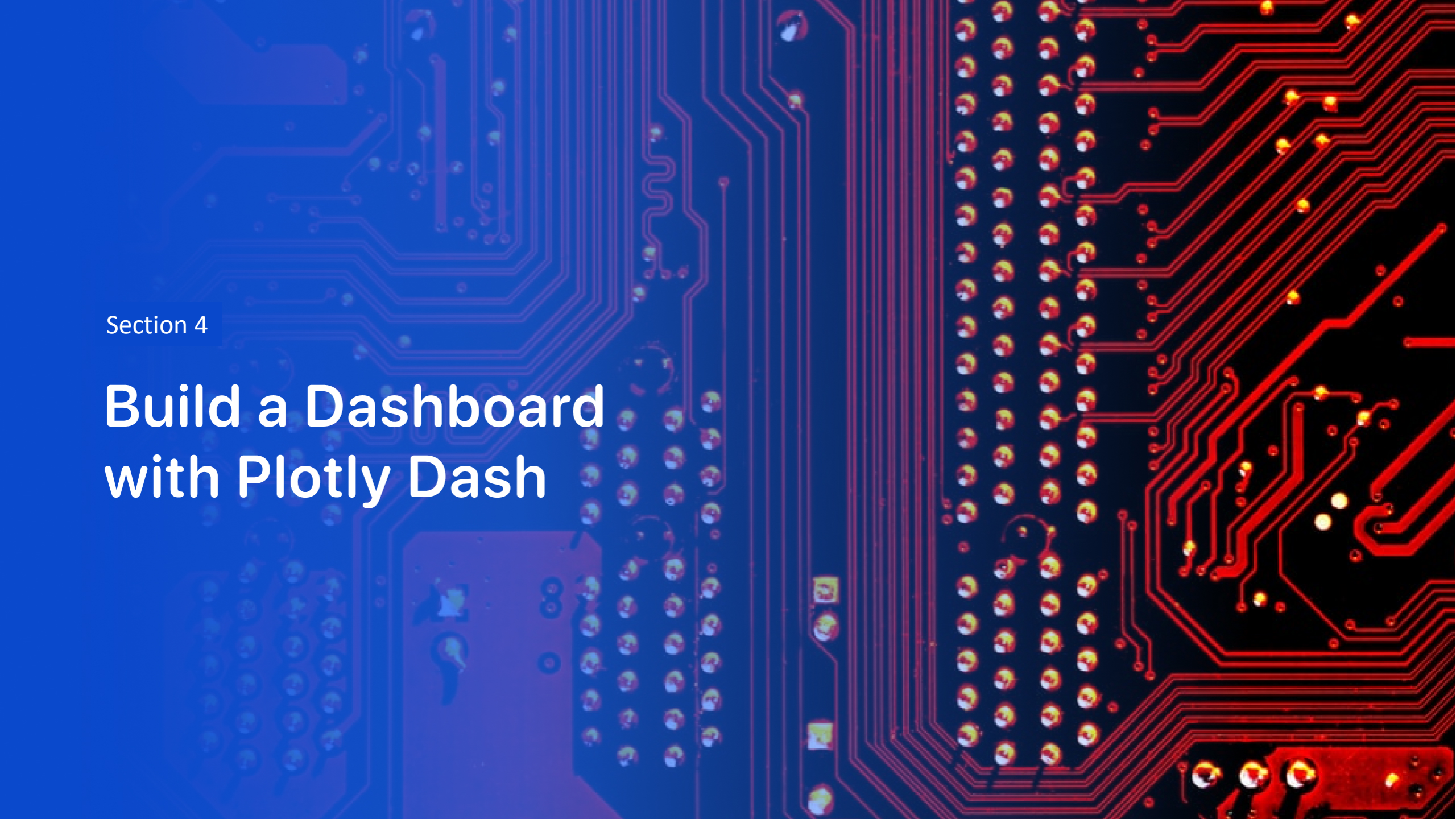
Out[30]:

Landing_Outcome	Total Count
Success (ground pad)	5
Failure (drone ship)	5

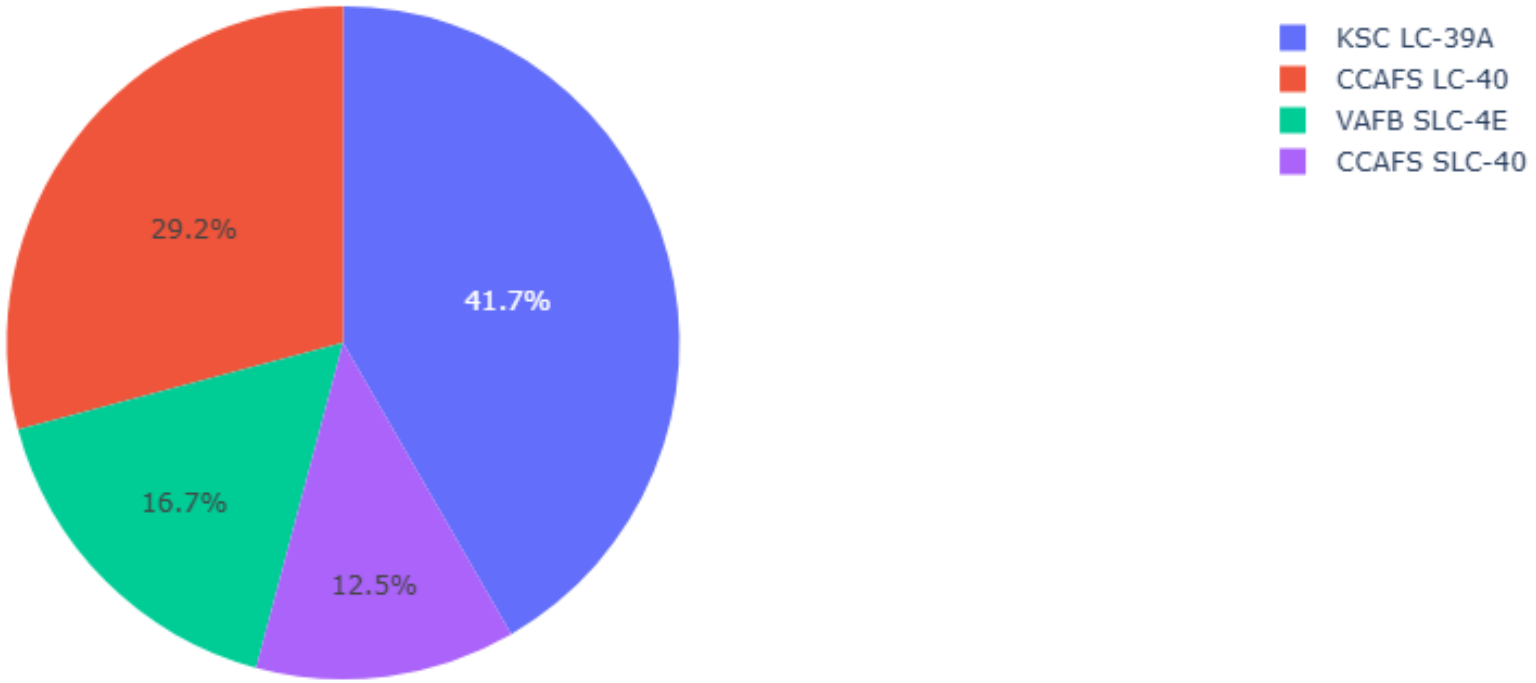
Section 3

Launch Sites Proximities Analysis

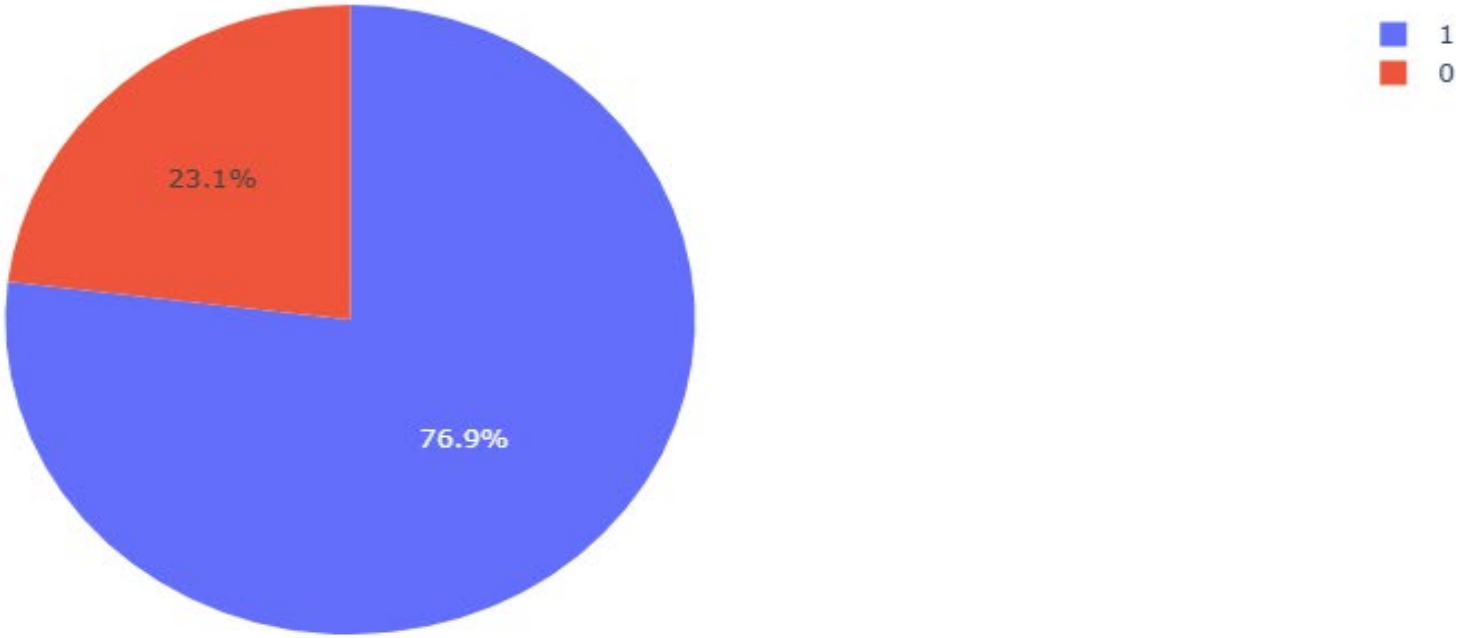

Lat: 28.55067 Long: -81.6009
Titusville
Orlando
Oak Ridge
Cape Canaveral
Merritt Island
Rockledge
Four Corners
Kissimmee
Saint Cloud
Viera
Satellite Beach
Poinciana
Haines City
Melbourne
51.43
KM
West Melbourne
Palm Bay
Three Lakes
Wildlife
Management
Area
Lake Wales


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Dashboard



KSC LC-39A

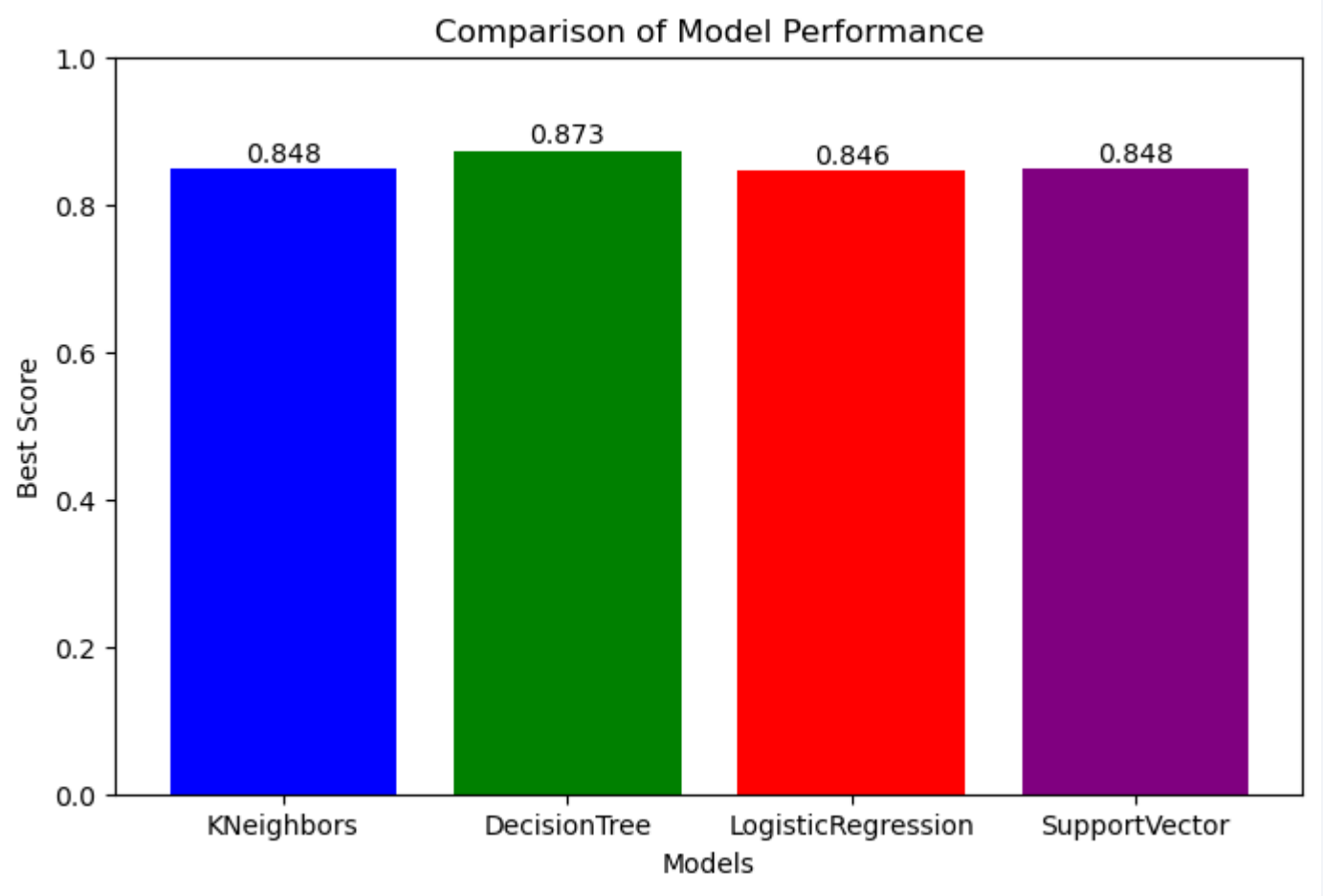


Site KSC LC 39A has a higher count with a 76.9% success rate of Falcon 9 landing.

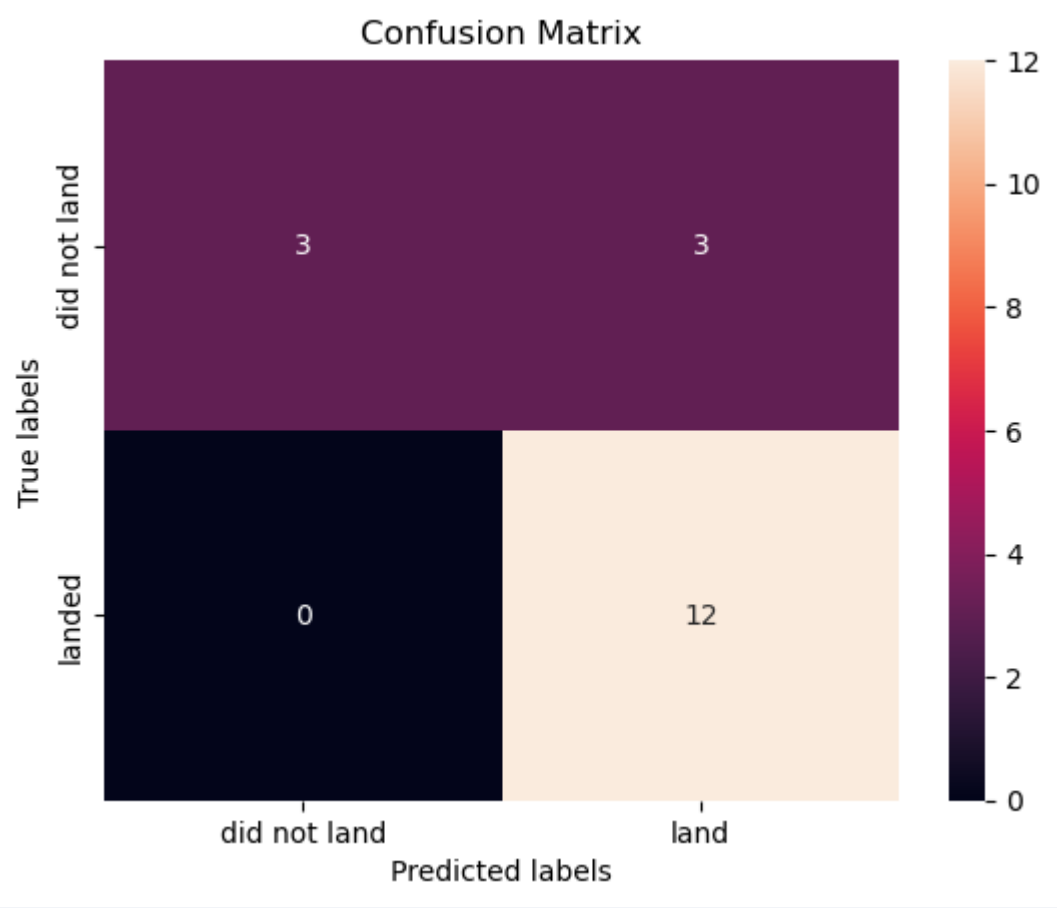
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

The best-performing model is the Decision Tree, achieving a score of 0.8732. The optimal hyperparameters are:

- **Criterion:** Gini
- **Max Depth:** 6
- **Max Features:** Square root
- **Min Samples Leaf:** 2
- **Min Samples Split:** 5
- **Splitter:** Random

Thank you!