IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Atif Hussain
August, 2023

# Executive Summary

## Methodologies

### Data Collection

- REST API
- Web Scraping

### Data Wrangling

- Creation of binary Outcome Variable

### Visualization

- Basic Graphs
- Maps using Folium
- Dashboard

### Analysis

- Basic Analysis
- Predictive Models

## Results

- Improved Success rate over the time
- KSC LC-39A is the most successful site
- ES-L1, GEO, HEO, and SSO have the highest success rate
- All sites are close to coastline
- Decision Trees model has the highest accuracy

# Introduction

## Project background and context

- SpaceX is a leader in the space industry. It has launched several missions in some of which first stage has been successful and unsuccessful in others. We need to analyze publicly available data for a competitor of SpaceX.

## Specific questions

- How various factors (launch site, payload, orbit etc) affect success/failure of first stage landing?
- Is there any trend in success rate of first stage landings over time?
- What ML model best predicts success?

Section 1

# Methodology

# Methodology

**Data collection:**

Collected through REST API and Web Scraping

**Perform data wrangling**

Filtering, missing values handling, one hot endcode

**Visualization and EDA with SQL**

**Visual analytics using Folium and Plotly Dash**

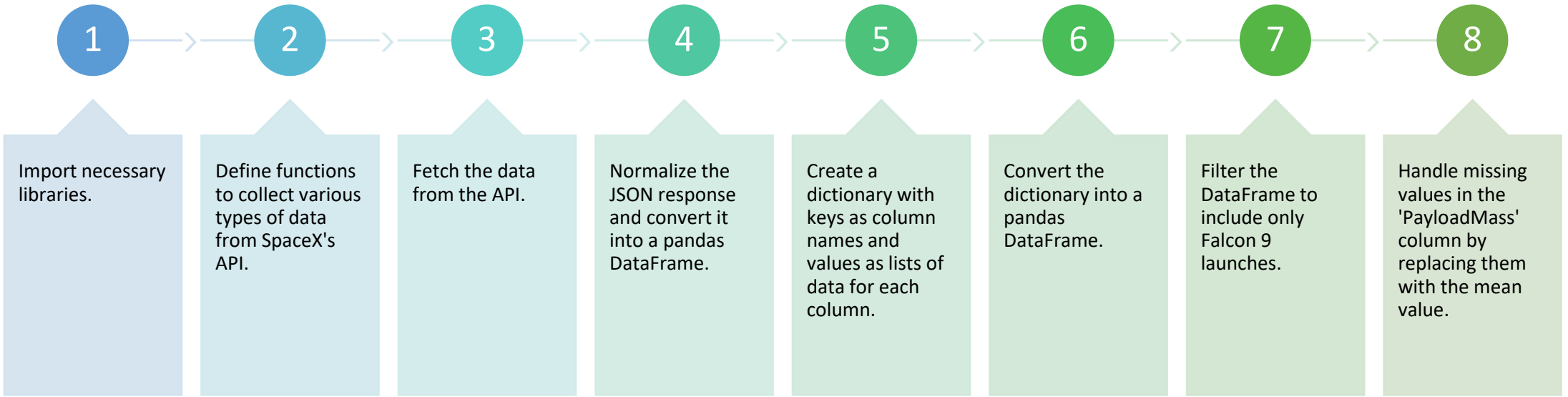
**Predictive Analytics**

Various classification models built and tested

# Data Collection

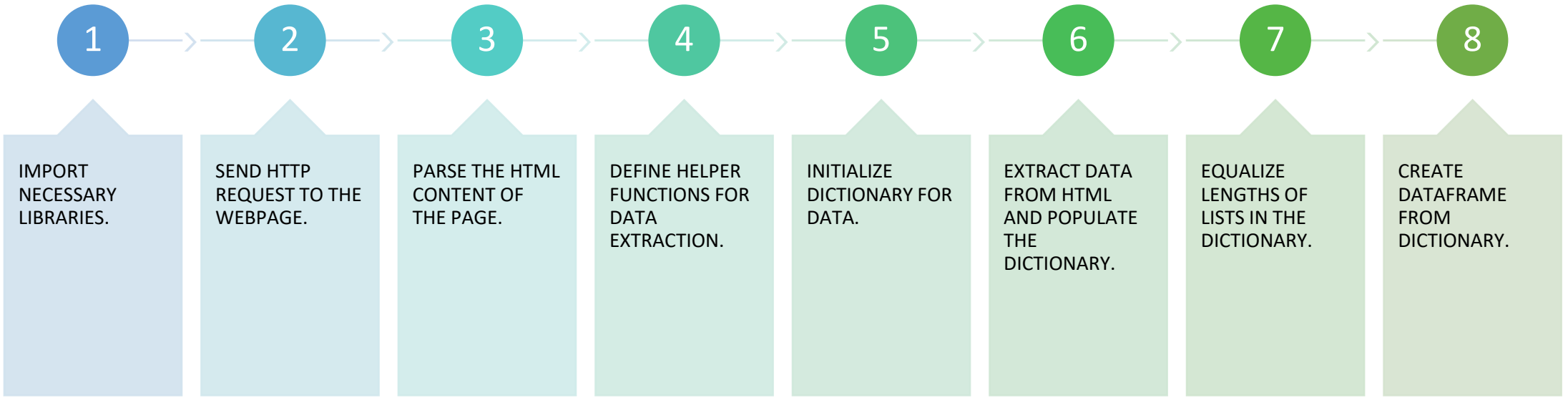
1. Request/Fetch Data
2. Convert to Data Frame
3. Filter Data
4. Replace missing values
5. Export to csv

# Data Collection – SpaceX API




Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/api%20collection.ipynb

# Data Collection - Scraping




Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/web%20scraping.ipynb

# Data Wrangling

Import necessary libraries (**numpy**, **pandas**)

Load the dataset

**D**ata analysis:

Missing values percentage

Column types

**Calculations**:

Number of launches on each site

Number and occurrence of each orbit

Number and occurrence of mission outcome per orbit type

New variable ‘Class’

Export the dataframe to a CSV file

Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/data%20wrangling.ipynb

# EDA with Data Visualization

| Plot | Purpose |
|---|---|
| FlightNumber vs. PayloadMass | Investigate the impact of flight number and payload mass on landing success |
| FlightNumber vs. LaunchSite | Examine the influence of flight number and launch site on success rates |
| Payload vs. LaunchSite | Determine the relationship between payload mass and choice of launch site |
| Success Rate vs. Orbit Type | Identify which orbit types have higher success rates |
| FlightNumber vs. Orbit Type | Explore the correlation between flight number and orbit type |
| Payload vs. Orbit Type | Analyze how payload mass and orbit type affect landing success |
| Yearly Success Trend | Assess the trend of launch success rate over time |

Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/eda.ipynb

# EDA with SQL 1/2

## Select

Select distinct launch site names.

- Query: SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

## Display

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

- Query: %sql SELECT SUM(PAYLOAD_MASS_KG_) as Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';.

## Display

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

- Query: SELECT SUM("PAYLOAD_MASS__KG_") as Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';

## Display

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

- Query: SELECT AVG("PAYLOAD_MASS__KG_") as Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%';

## List

List the date when the first successful landing outcome in ground pad was achieved.

- Query: SELECT MIN("Date") as First_Successful_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success%';

Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/sql%20analysis.ipynb

# EDA with SQL 2/2

## List

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

- Query: SELECT DISTINCT "Payload" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

## List

List the total number of successful and failure mission outcomes.

- Query: SELECT SUM(CASE WHEN "Mission_Outcome" LIKE 'Success%' THEN 1 ELSE 0 END) AS Successful_Missions, SUM(CASE WHEN "Mission_Outcome" LIKE 'Failure%' THEN 1 ELSE 0 END) AS Failed_Missions FROM SPACEXTABLE;

## List

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

- Query: SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

## List

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

- Query: SELECT SUBSTR("Date", 4, 2) as Month, "Booster_Version", "Launch_Site", "Landing_Outcome" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Failure (drone ship)%' AND SUBSTR("Date",7,4)='2015';

## Rank

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

- Query: SELECT "Landing_Outcome", COUNT(*) as Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC;

Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/sql%20analysis.ipynb

# Build an Interactive Map with Folium

| Object Added | Purpose |
|---|---|
| folium.map.Marker | Mark the exact location of each launch site with a text label marker. The marker has the launch site name. |
| folium.PolyLine | Draw a line from the launch site to the coastline. |
| folium.PolyLine | Draw a line from a nearby city to the launch site and display the distance. |
| folium.PolyLine | Draw a line from a nearby railway to the launch site and display the distance. |
| folium.PolyLine | Draw a line from a nearby highway to the launch site and display the distance. |

Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/Folium.ipynb

# Build a Dashboard with Plotly Dash

| Chart | Purpose |
| --- | --- |
| Chart- Successful Launches by Site | To see which site has the highest share in the successful launches |
| Chart – KSC LC 39A Success Failure comparison | To see the percentage of successful launches from the site |
| Scatter Chart | To see correlation between payload and launch success |

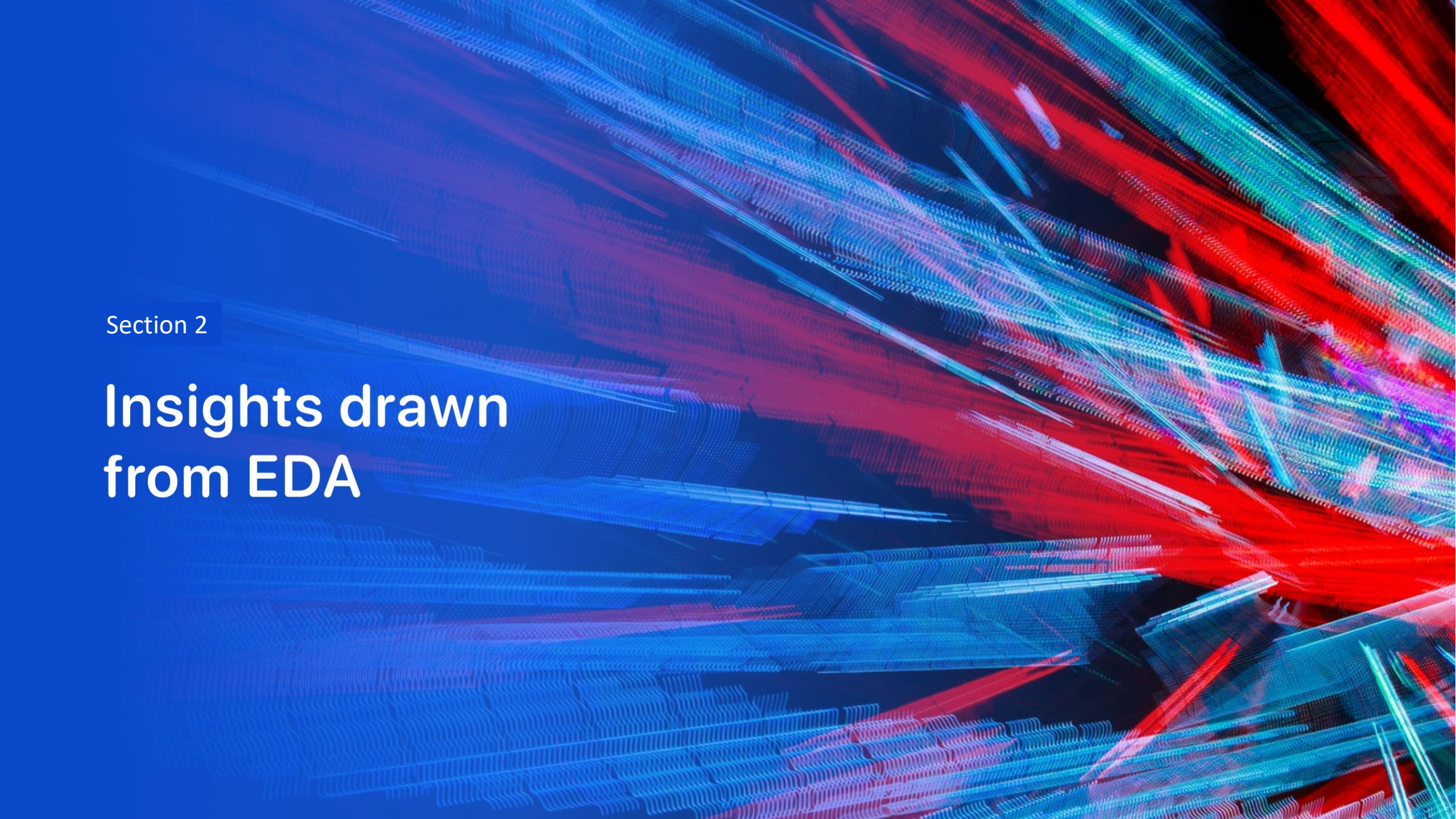
# Predictive Analysis (Classification)

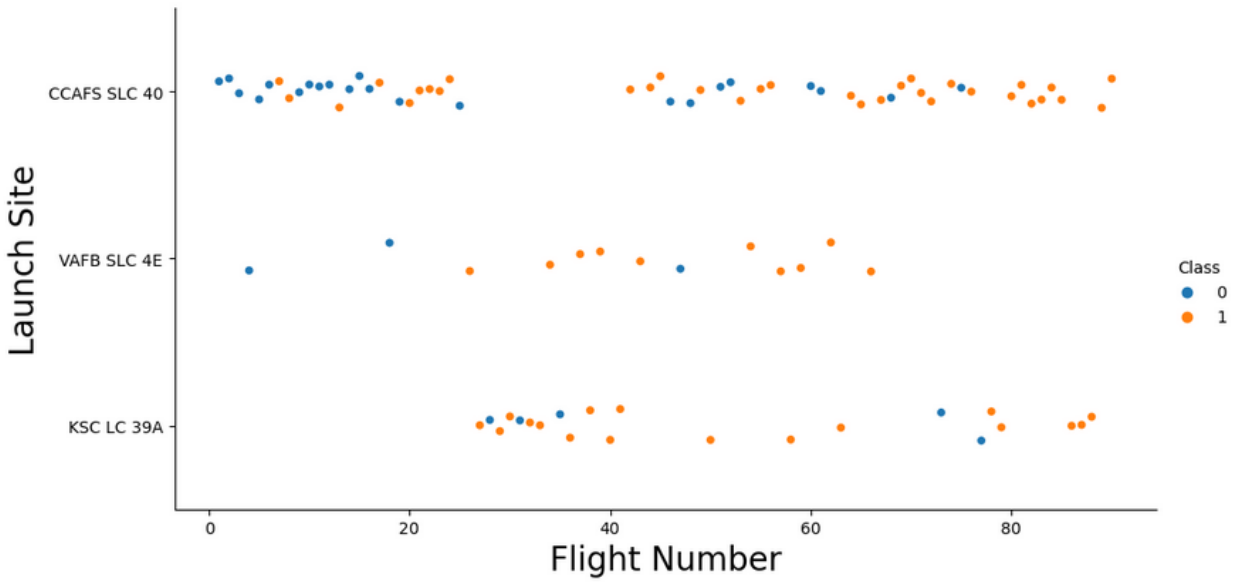
- Data Loaded
- Models Built
- Models tuned using GridSearchCV
- Models compared using accuracy score
- Best model selected

- Github Link: https://github.com/Atifhussain40/data_science_capstone/blob/main/Predictive.ipynb

# Main Results

- Success rate has improved over the time
- KSC LC 39A has the best success rate
- Decision Trees based model performs best in predicting outcome

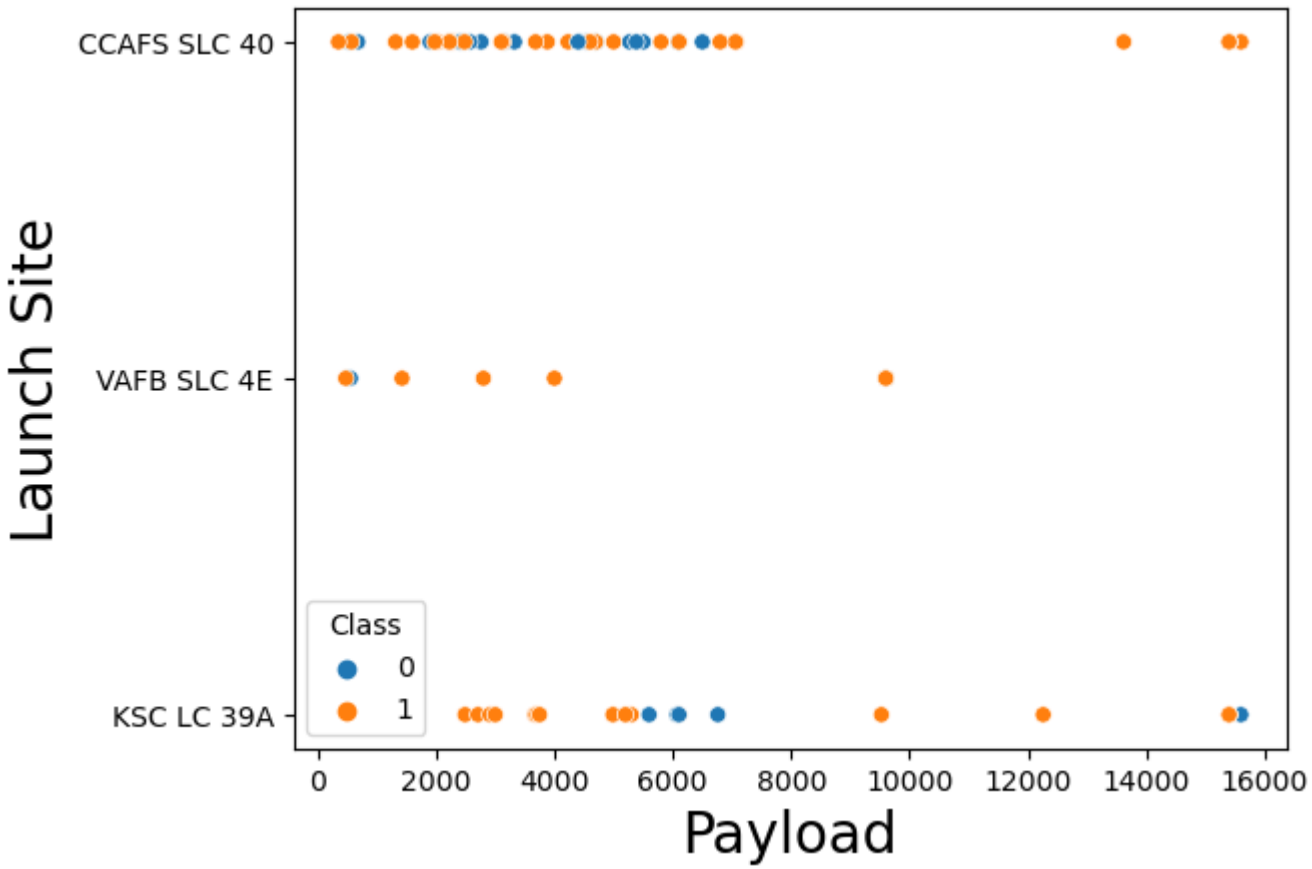

Section 2
Insights drawn from EDA




# Flight Number vs. Launch Site

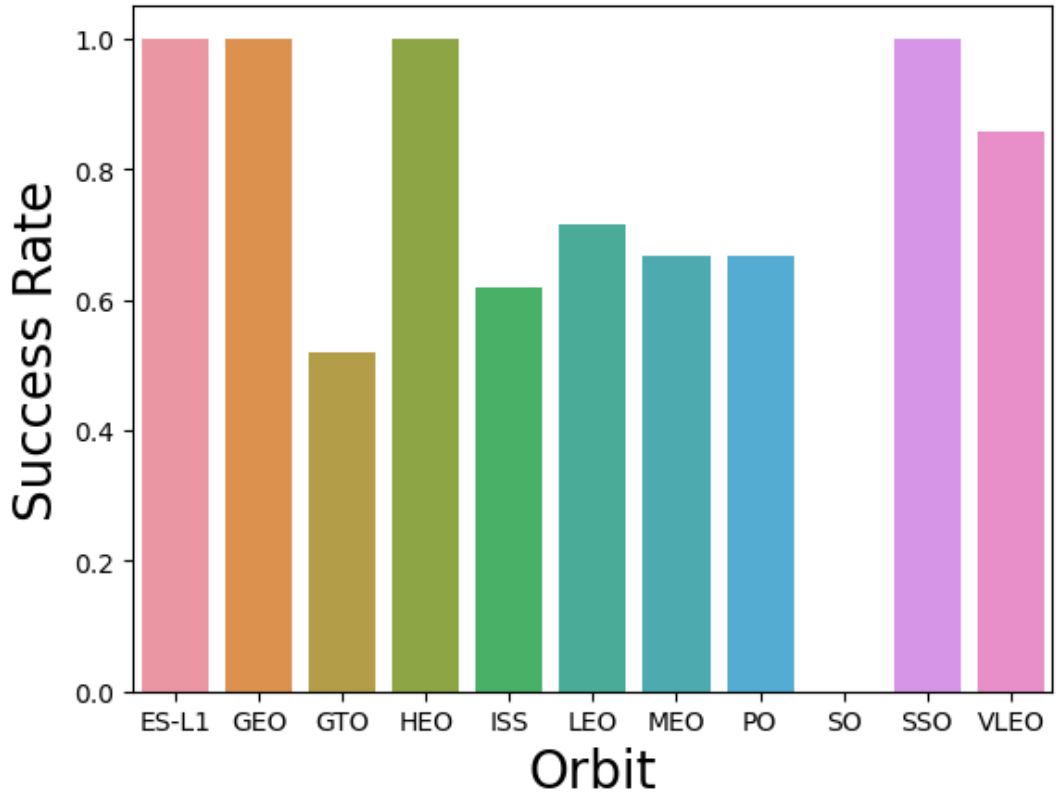
- CCAFS SLC 40 has the highest number of flights, however, many of those have been unsuccessful
- KSC LC 39A has very few unsuccessful flights

# Payload vs. Launch Site




- CCAFS SLC 40 has more success with greater mass loads
- KSC LC 39A also has greater success with small mass loads

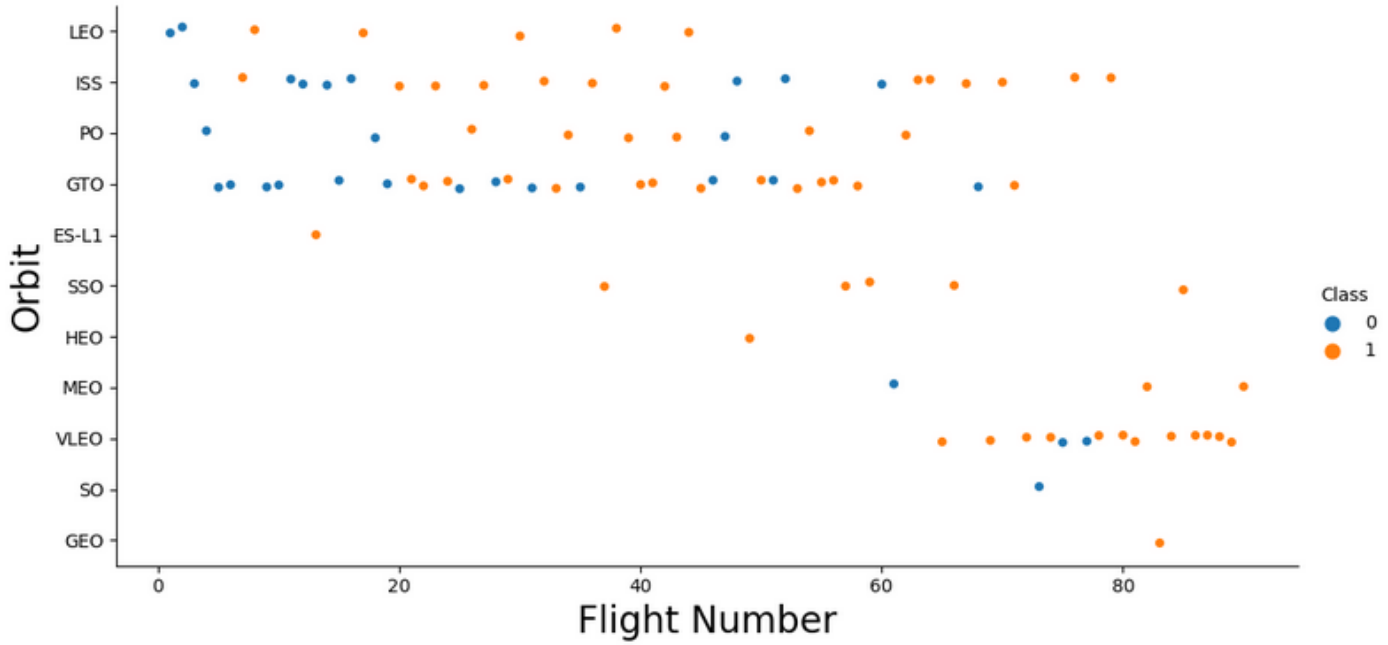
# Success Rate vs. Orbit Type




- ES-L1, GEO, HEO and SSO have the highest success rates
- GTO orbit has the lowest success rate

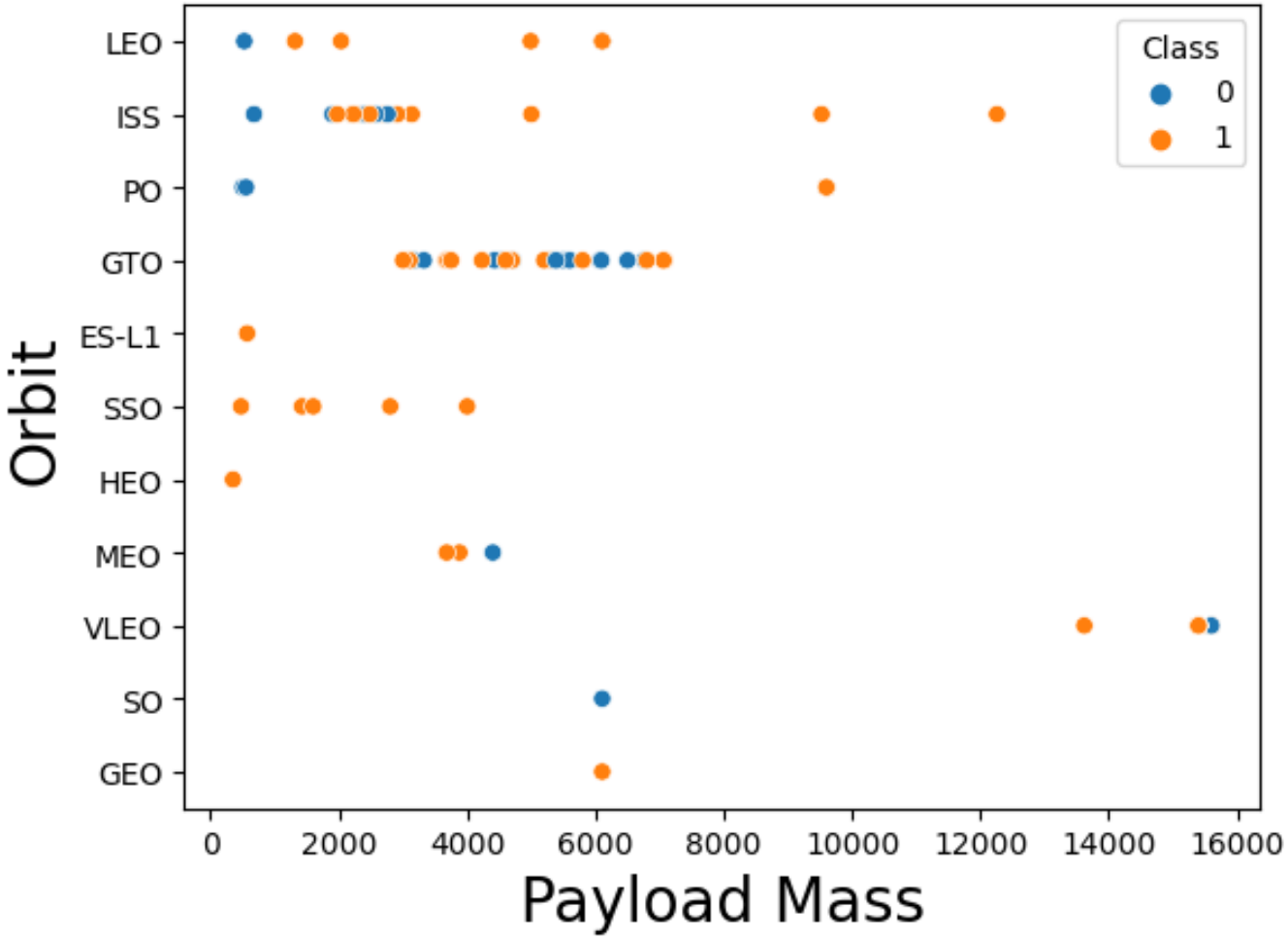
# Flight Number vs. Orbit Type

- LEO orbit, success is related to the flight number whereas
- In the GTO orbit, there is no relationship between flight number and the success



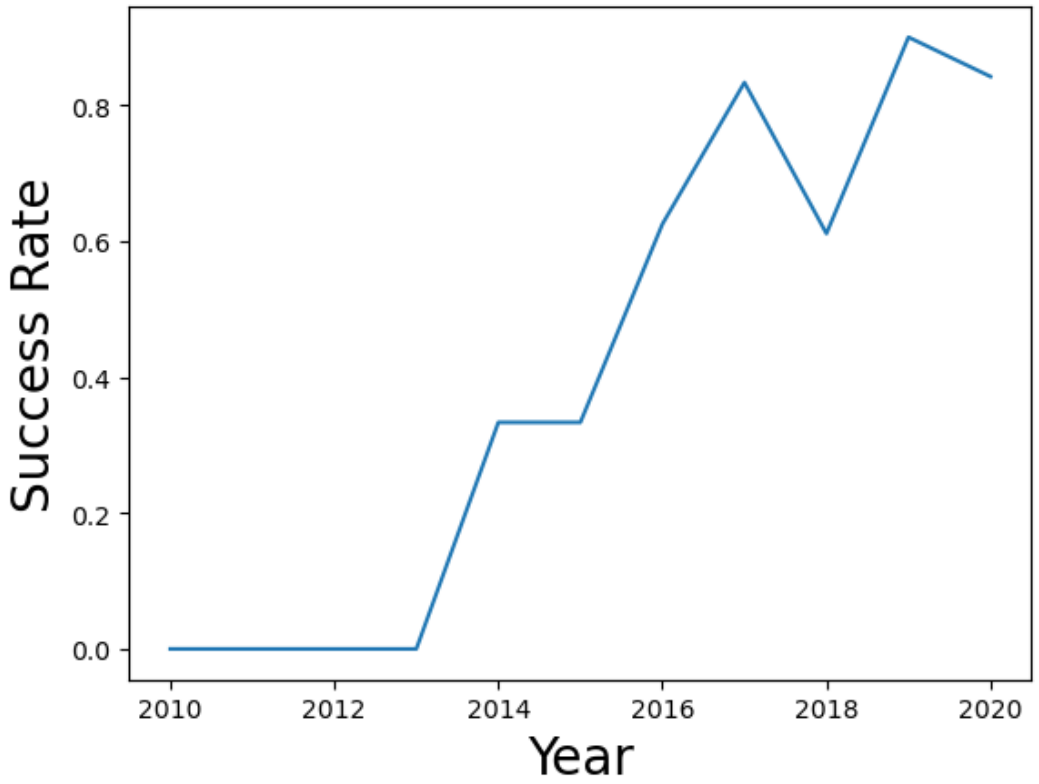
# Payload vs. Orbit Type

In ISS orbit, success is associated with higher payloads



# Launch Success Yearly Trend

Success rate has generally increased over the years.



# All Launch Site Names

Query has selected distinct/unique launch site names from SPACEXTABLE

```
In [16]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

# Launch Site Names Begin with 'CCA'

Selected records where launch sites began with `CCA` using LIKE

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission |
|---|---|---|---|---|---|---|---|---|
| 06/04/2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0.0 | LEO | SpaceX | |
| 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 | |
| 22/05/2012 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525.0 | LEO (ISS) | NASA (COTS) | |

# Total Payload Mass

```
In [22]: %sql SELECT SUM("PAYLOAD_MASS__KG_") as Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[22]:
```

| Total_Payload_Mass |
|---|
| 45596.0 |

Total payload carried by boosters from NASA is 45,596 Kgs

# Average Payload Mass by F9 v1.1

In [12]:
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

Out[12]:

| AVG("PAYLOAD_MASS__KG_") |
|---|
| 2928.4 |

**Average Payload Mass carried by F9 v1.1 is 2534.67 Kgs**

# First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

| MIN("Date") |
| --- |
| 01/08/2018 |

**The first successful ground landing was on 01/07/2020**

# Successful Drone Ship Landing with Payload between 4000 and 6000

In [27]:
```
%sql SELECT DISTINCT "Payload" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAY
```

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

Out[27]:

| Payload |
| --- |
| JCSAT-14 |
| JCSAT-16 |
| SES-10 |
| SES-11 / EchoStar 105 |

Names of the boosters with payloads between 4000 and 6000 and successful landing are JCSAT-14, JCSAT-16, SES-10, SES-11/EchoStar 105

# Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [28]:
```
%sql SELECT SUM(CASE WHEN "Mission_Outcome" LIKE 'Success%' THEN 1 ELSE 0 END) AS Successful_Missions, SUM(CA
```

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

Out[28]:

| Successful_Missions | Failed_Missions |
|---|---|
| 100 | 1 |

100 missions have been successful and 1 unsuccessful

# Boosters Carried Maximum Payload

In [15]:
```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("
```

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

Out[15]:

| 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 |

**Boosters carrying the maximum payload**

# 2015 Launch Records

In [36]:
```
%sql SELECT SUBSTR("Date", 4, 2) as Month, "Booster_Version", "Launch_Site", "Landing_Outcome" FROM SPACEXTABL
```

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

Out[36]:

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

**Two Boosters in 2015 failing in drone ship**

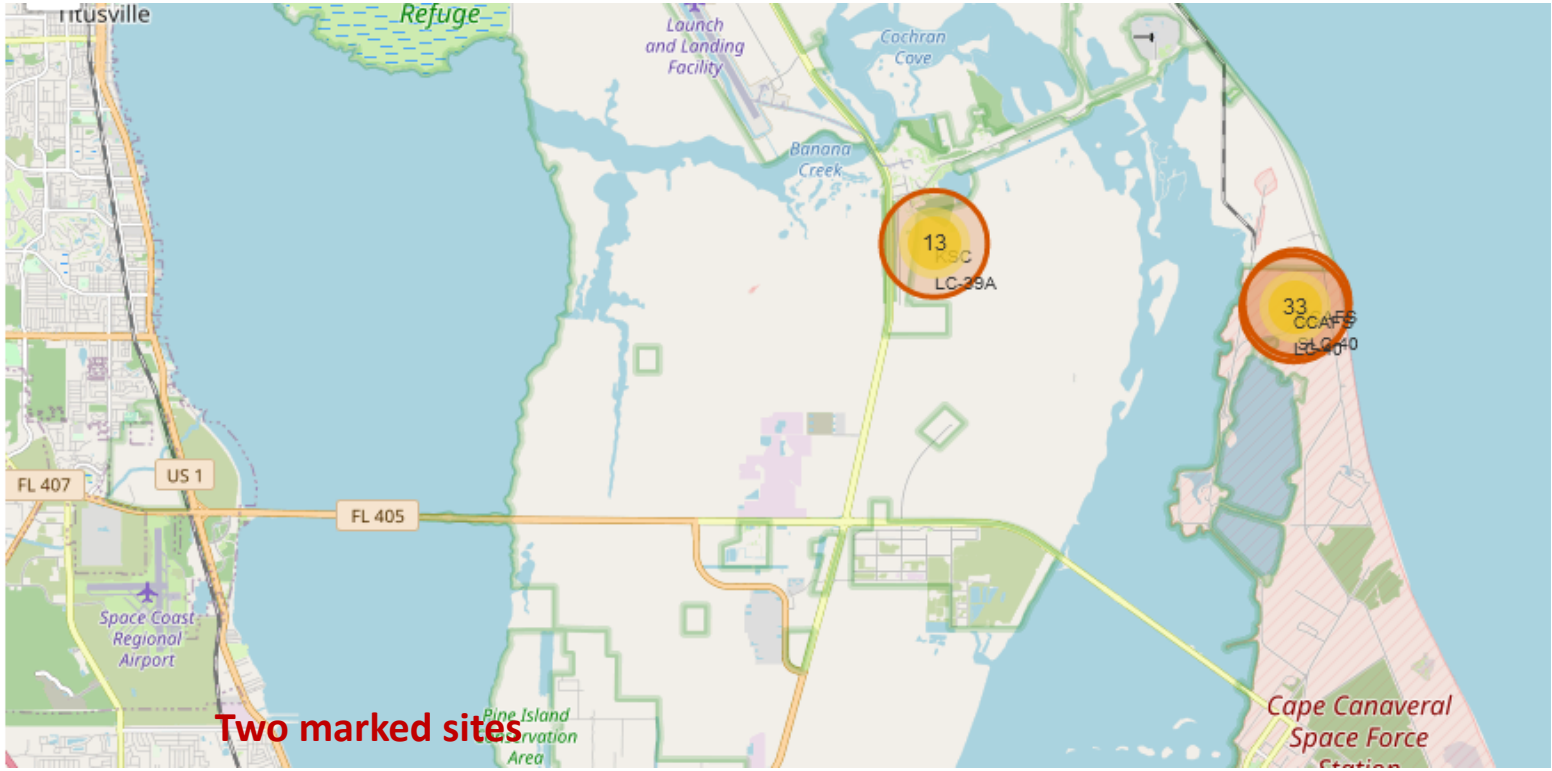
Section 3

# Launch Sites Proximities Analysis

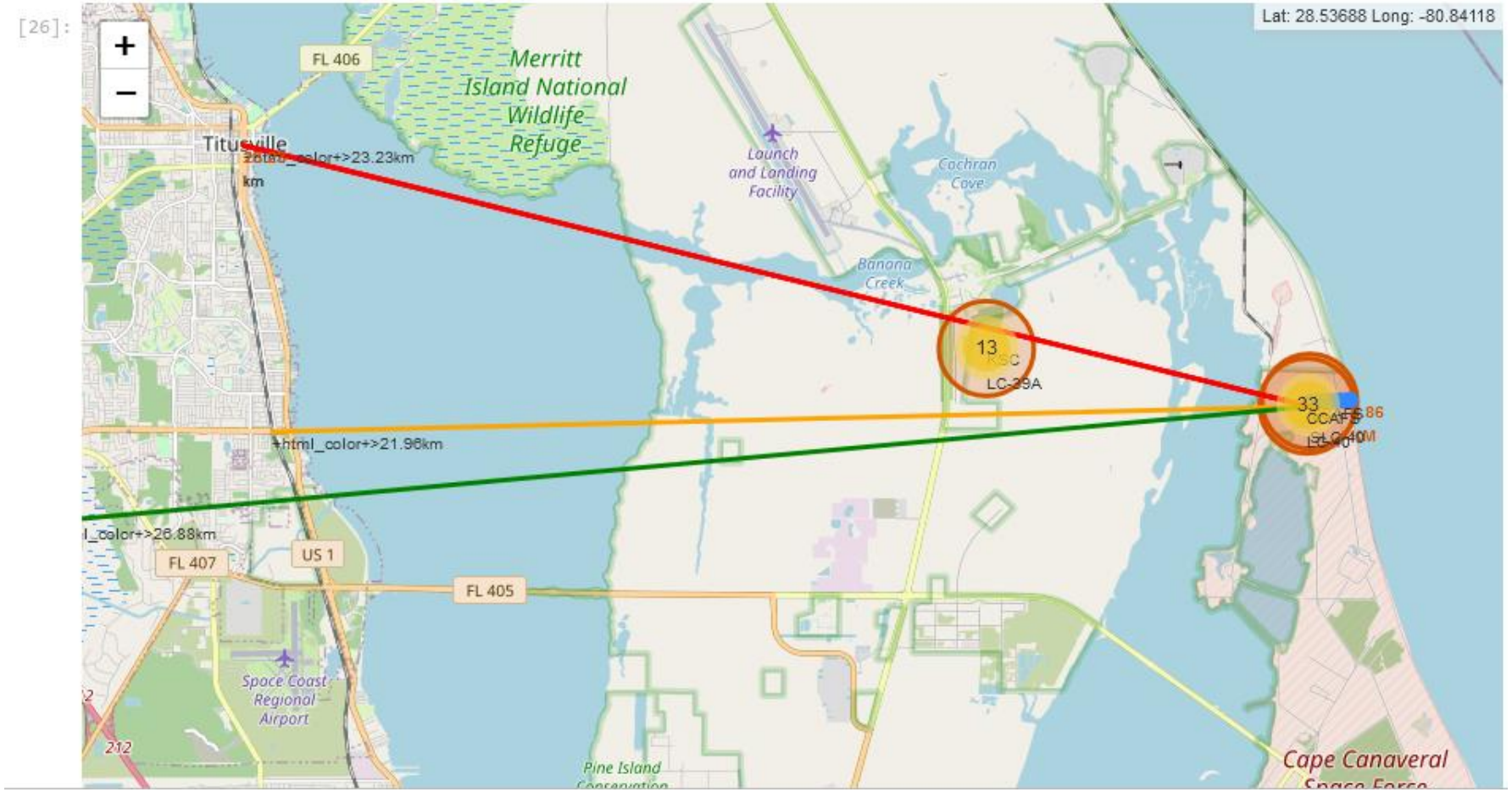
# Launch Sites in USA

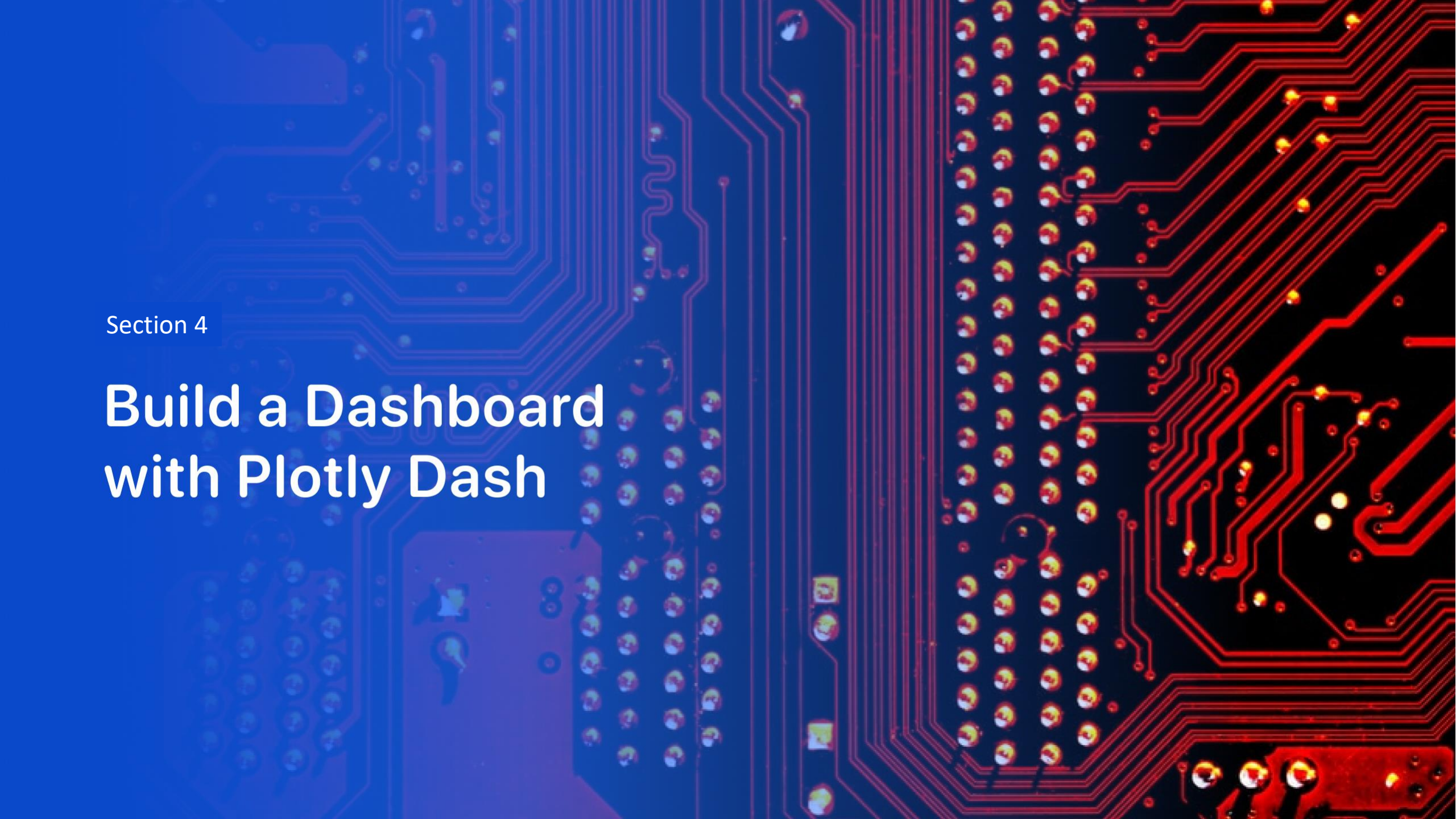

Launch sites in USA

# Marked Sites

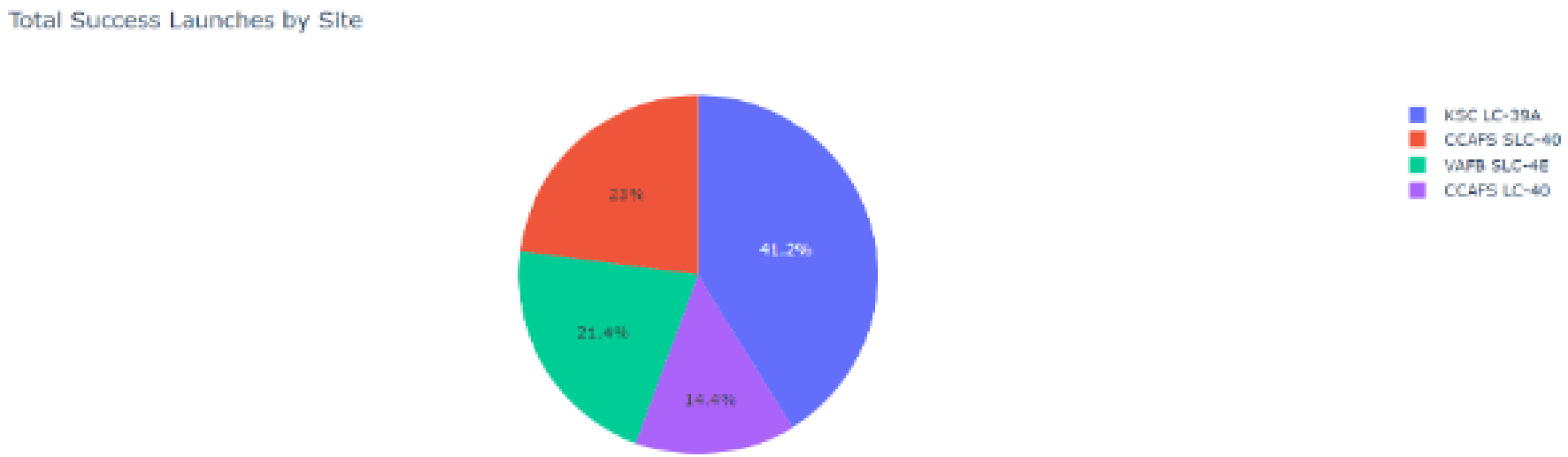


# Distances from city, railway, and highway



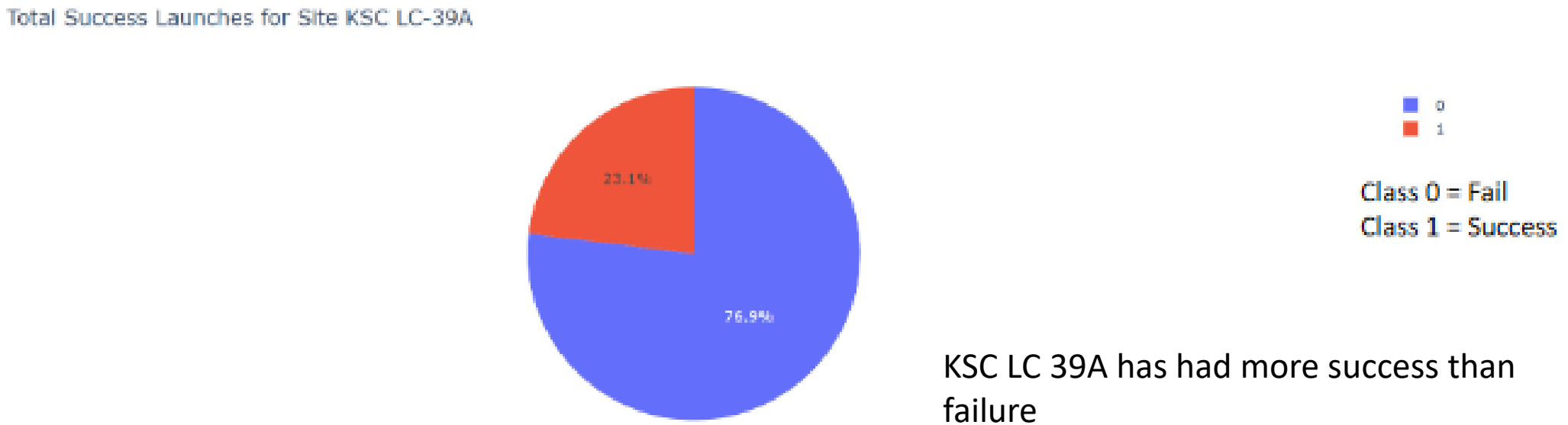

Section 4
Build a Dashboard with Plotly Dash

# Successful Launches by site




KSC LC 39A has the highest successful launches

# KSC LC 39A Success vs Failure

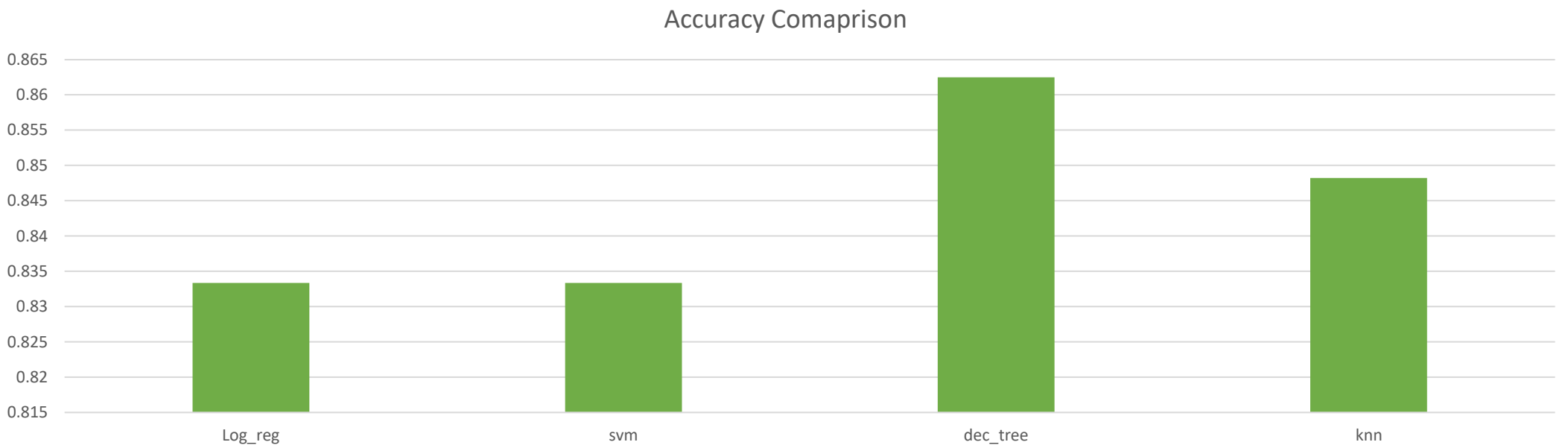



KSC LC 39A has had more success than failure

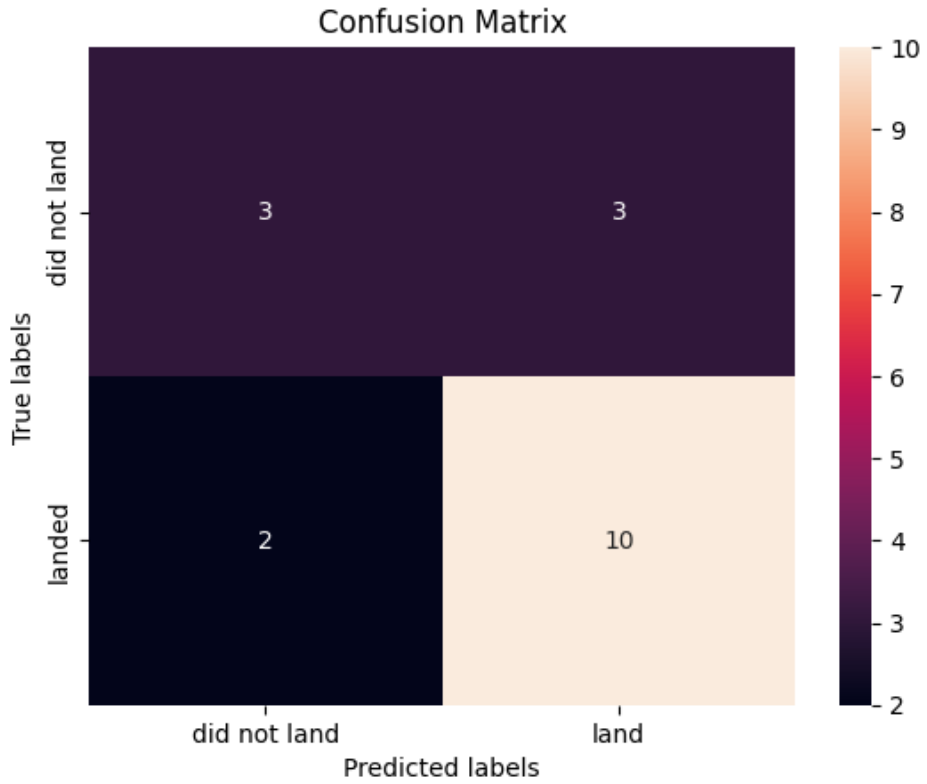
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy



# Confusion Matrix of Decision Tree




The matrix shows that the classifier misclassified 3 failed landings as successful

# Conclusions

Launch success rate has generally increased over the time.

Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success

KSC LC-39A has enjoyed the most success.

The Decision tree classifier is the best classifier in the current context

# Conclusions

- Launch success rate has generally increased over the time.
- Higher payloads are associated with greater success
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success
- KSC LC-39A has enjoyed the most success.
- Sites are near the coastline
- The Decision tree classifier is the best classifier in the current context

# Thank you!