IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Harisai Marisa
13th Dec 2024
Discord: iamharisai
LinkedIn: harisai-m

GitHub - Link

# Outline

# Executive Summary

## Summary of methodologies

- Data collection
- - Data wrangling
- - Exploratory Data Analysis with Data Visualization
- - Exploratory Data Analysis with SQL
- - Building an interactive map with Folium
- - Building a Dashboard with Plotly Dash
- - Predictive analysis (Classification)

## Summary of all results

- - Exploratory Data Analysis results
- - Interactive analytics demo in screenshots
- - Predictive analysis results

# Introduction

- **Project background and context**

- SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon. 9 rocket launches on its website, costing 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we will predict if SpaceX will reuse the first stage.

**Questions to be answered**

- - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- - Does the rate of successful landings increase over the years?
- - What is the best algorithm that can be used for binary classification in this case?

Section 1

# Methodology

# Methodology

- Executive Summary
- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Comparing them using jacard score, F1 score and accuracy measures

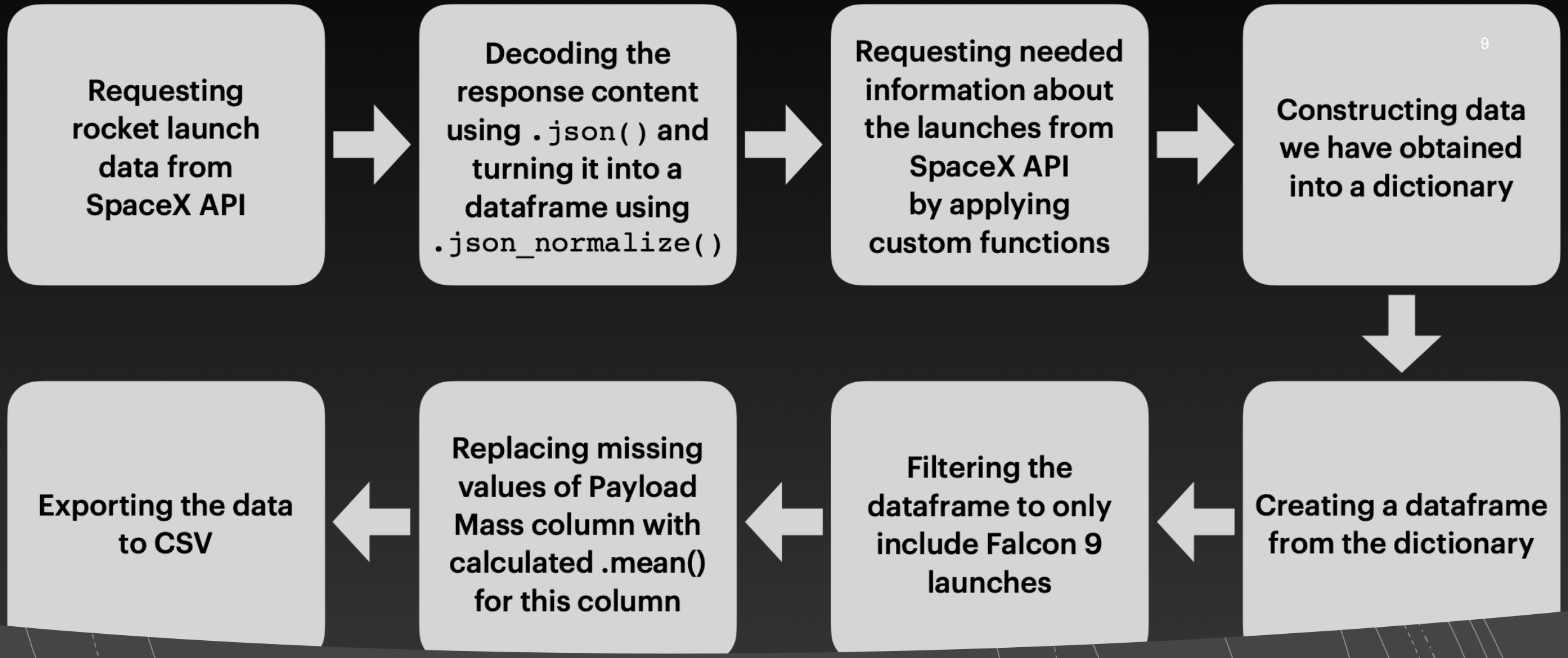
# Data Collection

- **Data Collection Process**
- **Data Sources**:
  - **SpaceX REST API**: Used for retrieving structured launch-related data.
  - **Wikipedia Web Scraping**: Extracted additional data from SpaceX's Wikipedia entry.
- **Purpose**:
  - Combined both methods to gather comprehensive information for detailed analysis of SpaceX launches.

# Data Columns Collected

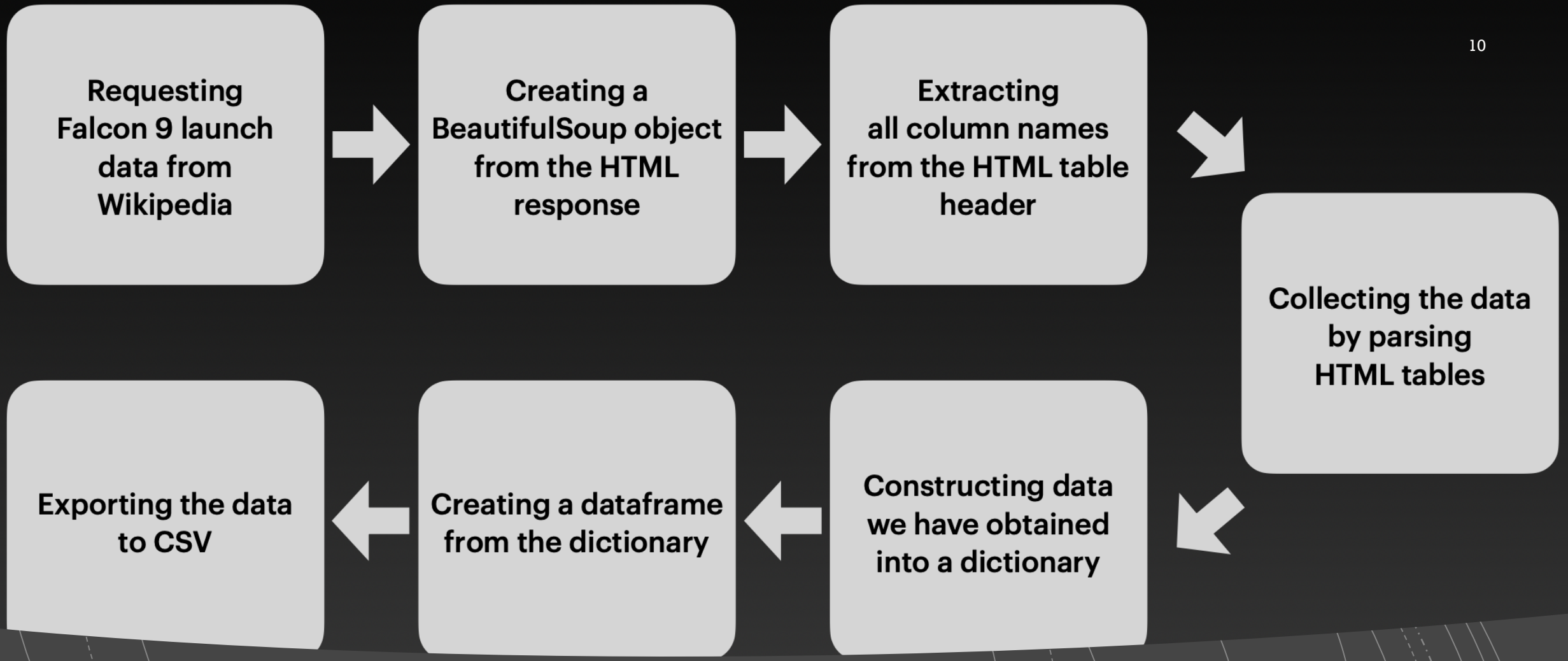
From the **SpaceX REST API**, we collected the following columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

From **Wikipedia Web Scraping**, we extracted the following columns: Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time.




# Data Collection – SpaceX API

GitHub - Link




# Data Collection - Scraping

GitHub - Link

# Data Wrangling

- Perform exploratory Data Analysis and determine Training Labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the datato CSV

GitHub - Link

# EDA with Data Visualization

Charts were plotted:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass
- vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type,
- Payload Mass vs Orbit Type and Success Rate Yearly Trend

---

Scatter plots show the relationship between variables. If a relationship exists, they could be used in a machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series).

GitHub - Link

# EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in the ground pad was achieved
- Listing the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failed mission outcomes
- Listing the names of the booster versions that have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in the year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order

GitHub - Link

# Build an Interactive Map with Folium

- Markers of all Launch Sites:
- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to the Equator and coasts.

- Coloured Markers of the launch outcomes for each Launch Site:
- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- Distances between a Launch Site to its proximities:
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

GitHub - Link

# Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

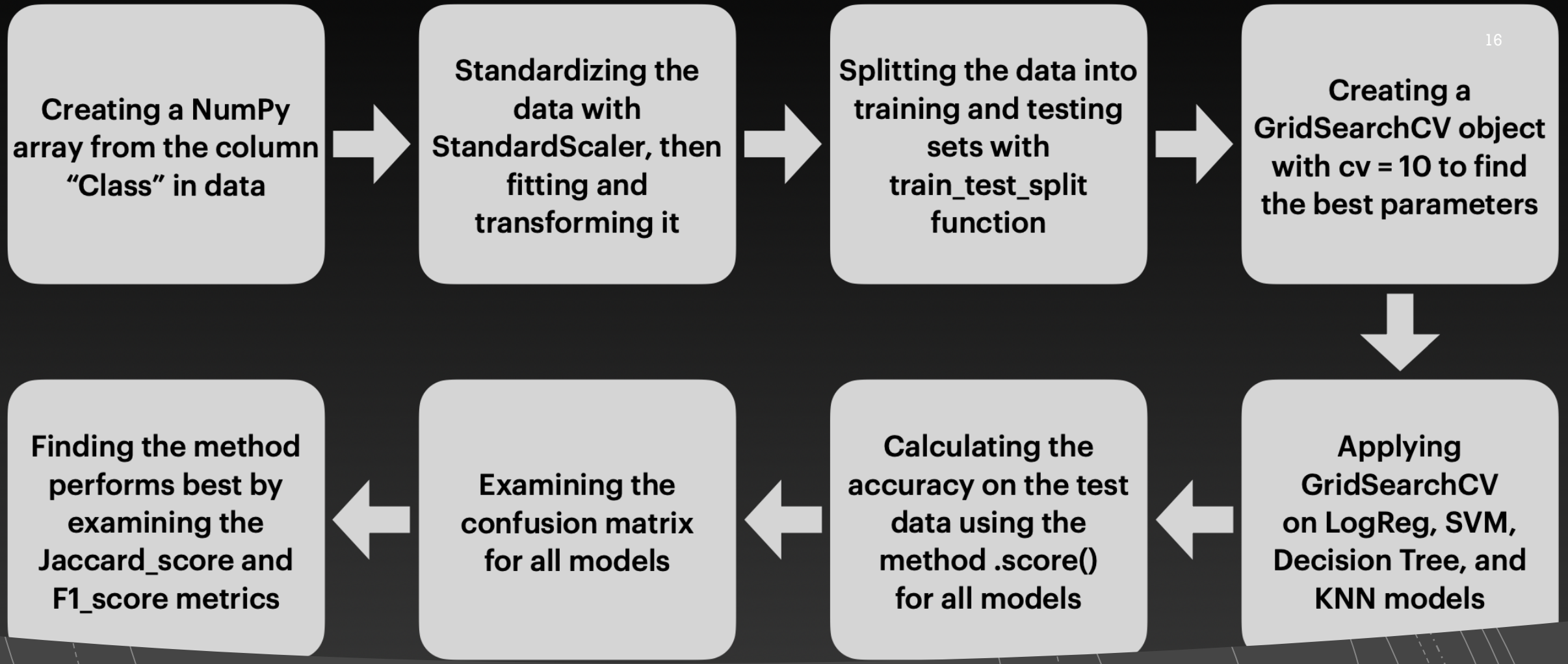
Slider of Payload Mass Range:

Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub - Link




# Predictive Analysis (Classification)

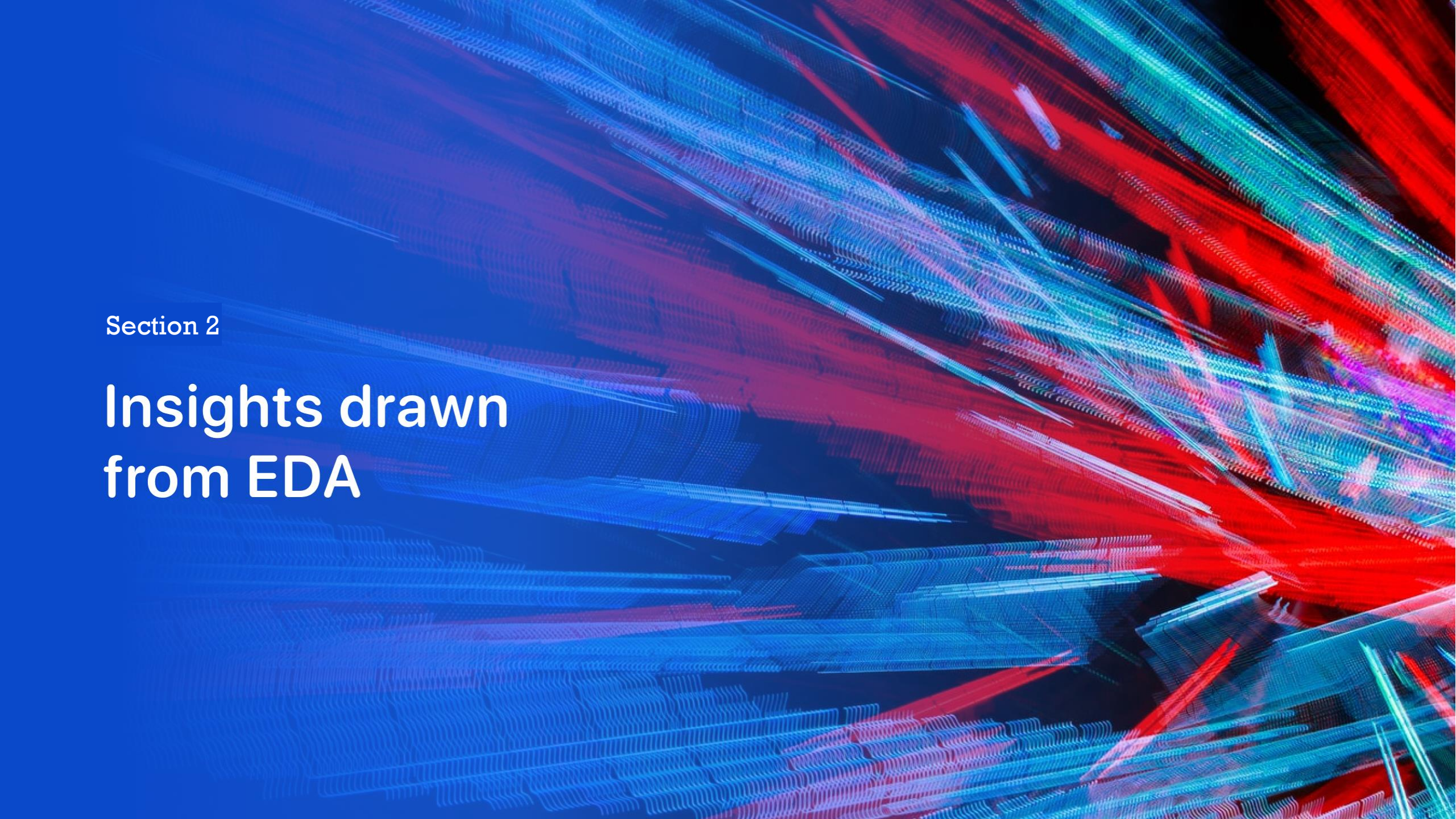
GitHub - Link

# Results

Exploratory data analysis results

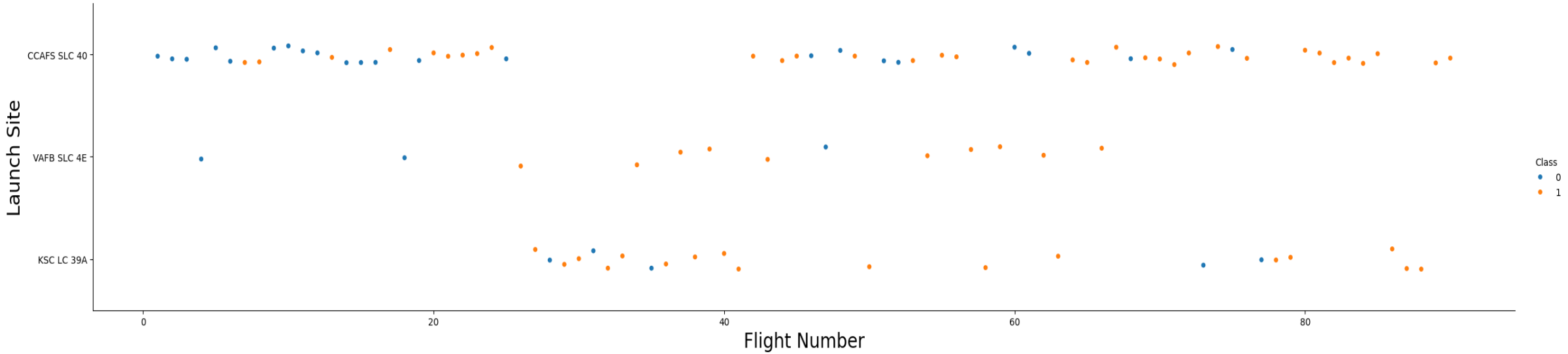
Interactive analytics demo in screenshots

Predictive analysis results



Section 2

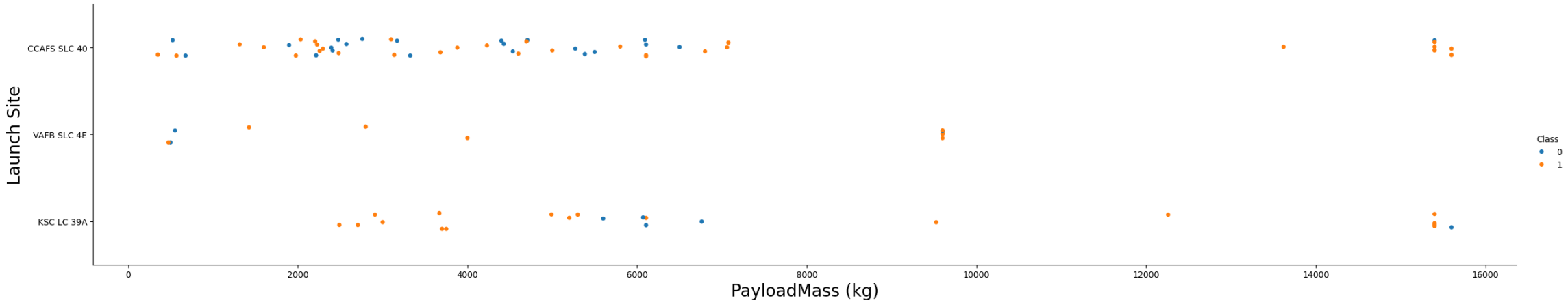
# Insights drawn from EDA




# Flight Number vs. Launch Site

Explanation:

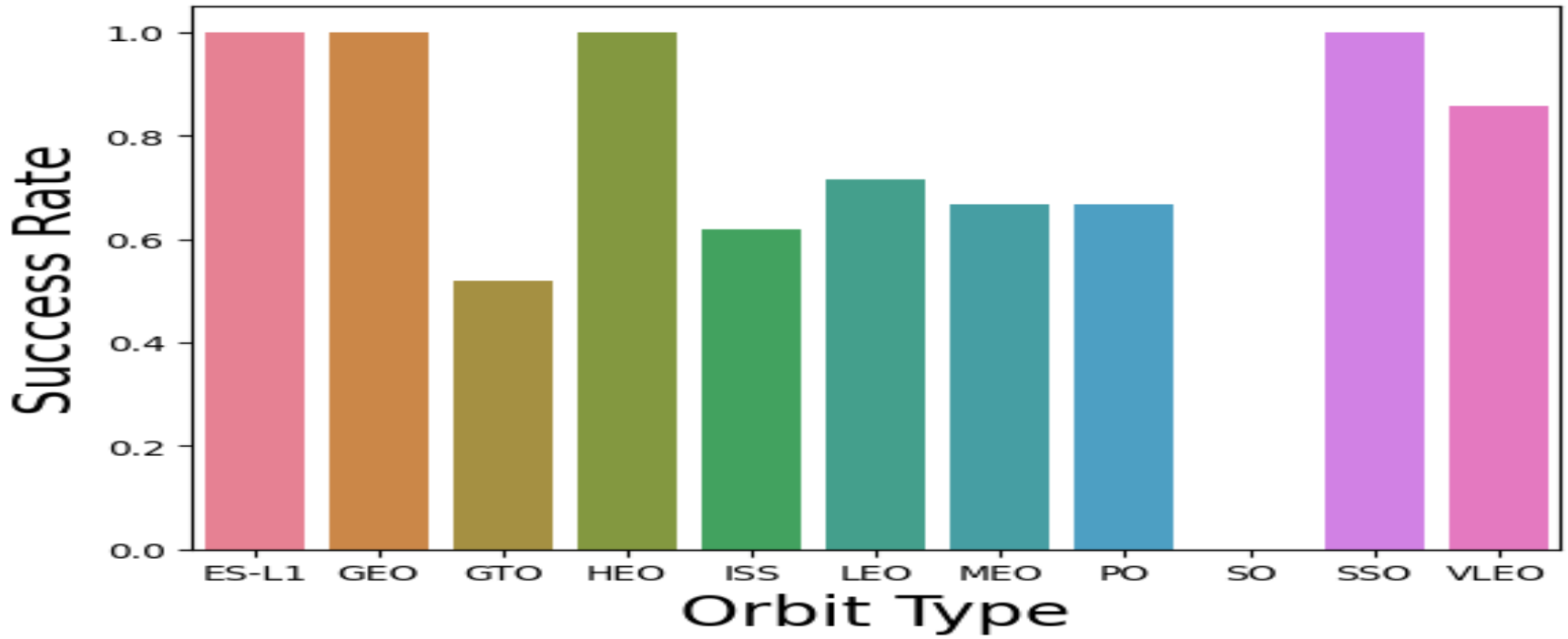
- • The earliest flights all failed while the latest flights all succeeded.
- • The CCAFS SLC 40 launch site has about half of all launches.
- • VAFB SLC 4E and KSC LC 39A have higher success rates.
- • It can be assumed that each new launch has a higher rate of success.




# Payload vs. Launch Site

Explanation:

- • For every launch site the higher the payload mass, the higher the success
- rate.
- • Most of the launches with payload mass over 7000 kg were successful.
- • KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



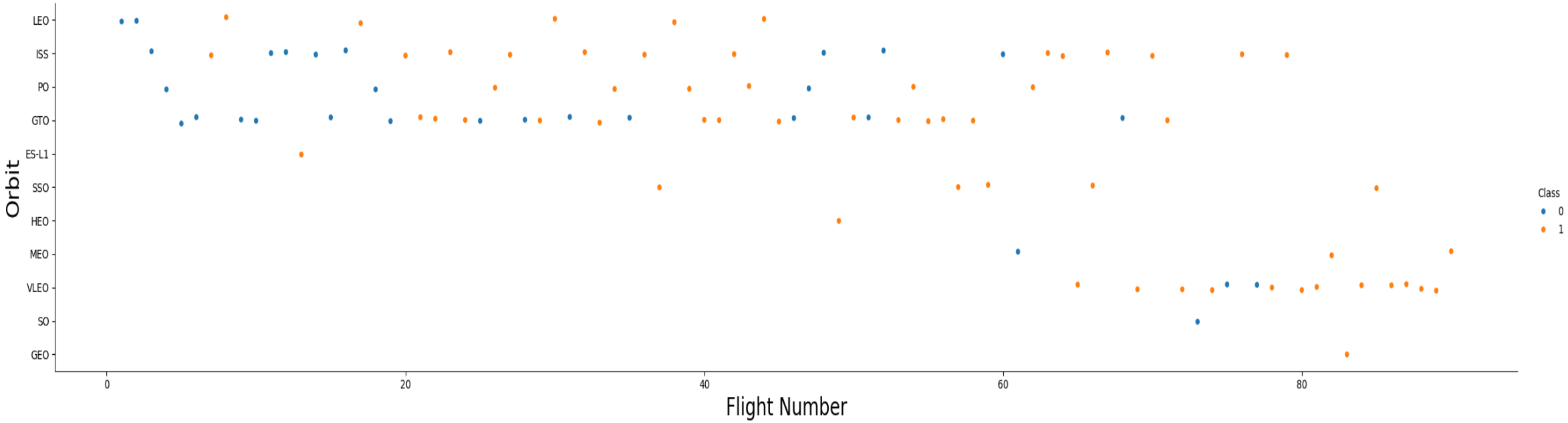

# Success Rate vs. Orbit Type

Explanation:

Orbits with 100% success rate: ES-L1, GEO, HEO, SSO

Orbits with 0% success rate: SO

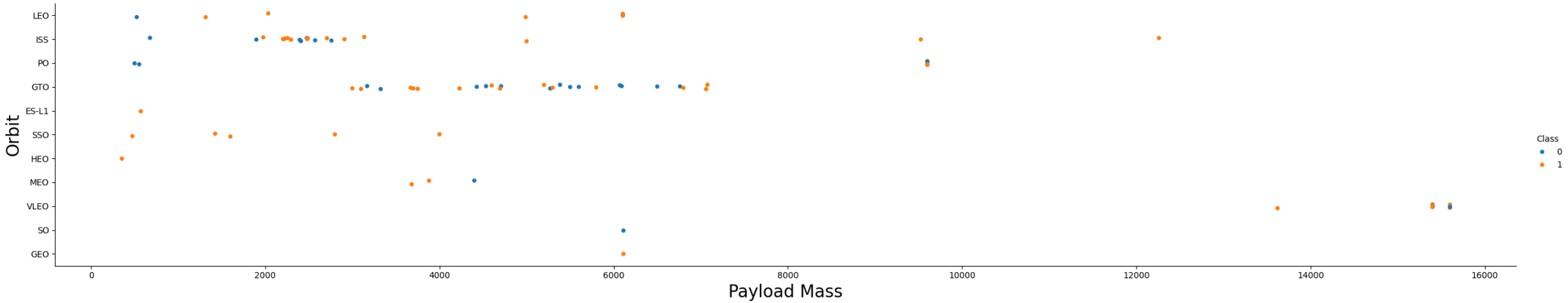
Orbits with a success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO




# Flight Number vs. Orbit Type

Explanation:

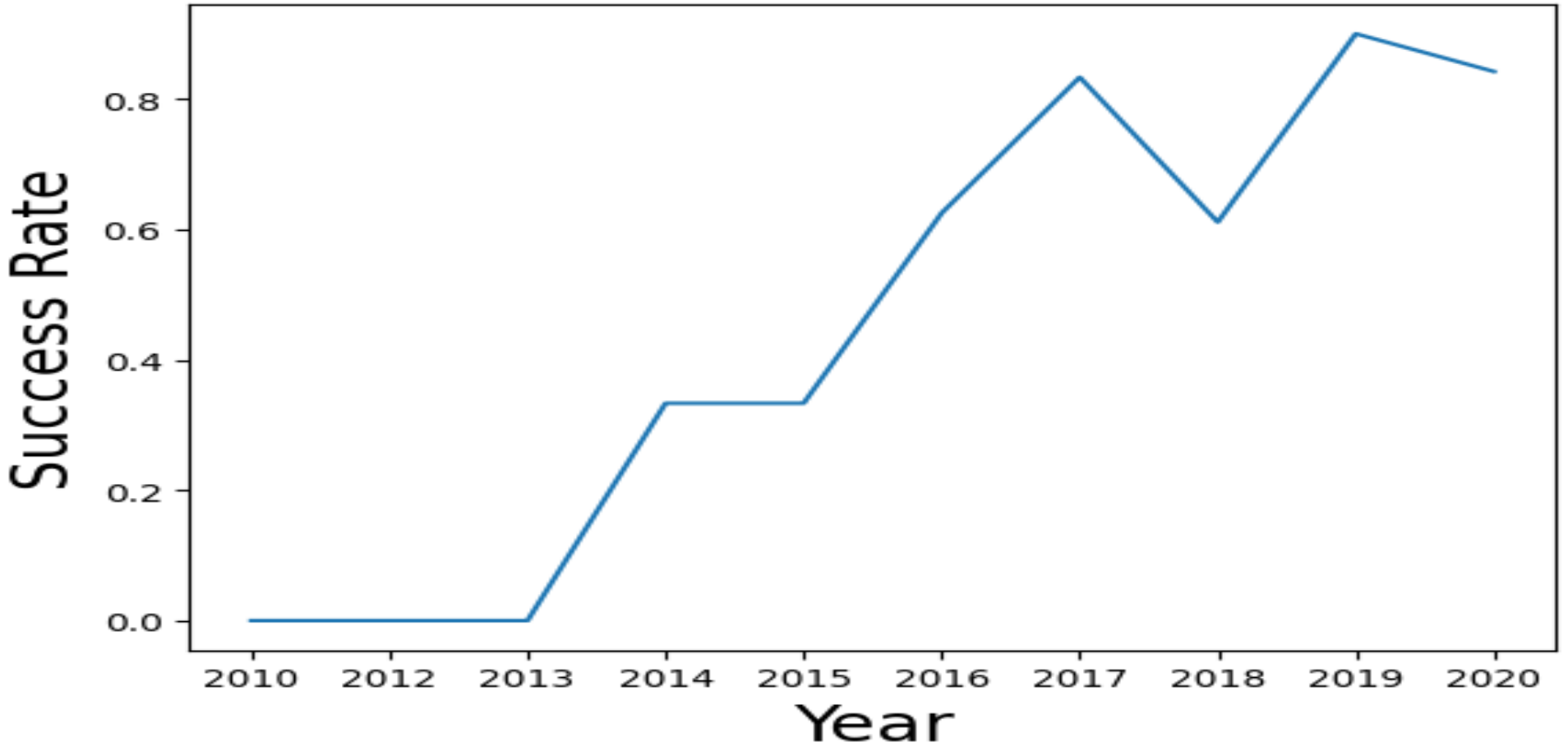
In the LEO orbit the Success appears related to the number of flights on the other hand, there seems to be no relationship between flight number when in GTO orbit.




# Payload vs. Orbit Type

**Explanation:**

Heavy payloads have a negative influence on GTO orbits and a positive on GTO and Polar LEO (ISS) orbits.




# Launch Success Yearly Trend

Explanation: The success rate since 2013 kept increasing till 2020.

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

```
[11]  %sql select distinct Launch_site from SPACEXTABLE
```

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

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

# All Launch Site Names

Explanation: I have used distinct keyword to get the unique names of launch sites

```
%sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5
```

[13] Python

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

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 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 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Launch Site Names Begin with 'CCA'

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

```
%sql select customer,sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)' group by Customer
```

[25] Python

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

| Customer | sum(PAYLOAD_MASS__KG_) |
|---|---|
| NASA (CRS) | 45596 |

# Total Payload Mass

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

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

```
[28] %sql select Booster_Version,avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1' group by Booster_Version
```

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

| Booster_Version | avg(PAYLOAD_MASS__KG_) |
|---|---|
| F9 v1.1 | 2928.4 |

# Average Payload Mass by F9 v1.1

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

```
[30] %sql select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

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

| min(Date) |
| --- |
| 2015-12-22 |

# First Successful Ground Landing Date

Explanation: Displaying average payload mass carried by booster version F9 v1.1.

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

```
[32] %sql select Booster_Version from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

+ Code + Markdown

# Successful Drone Ship Landing with Payload between 4000 and 6000

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

List the total number of successful and failure mission outcomes

```
[33]  %sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

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

| Mission_Outcome | count(*) |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Total Number of Successful and Failure Mission Outcomes

Explanation: Listing the total number of successful and failed mission outcomes.

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

[34]

```
 * 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 |

# Boosters Carried Maximum Payload

**List the names of the booster versions that have carried the maximum payload mass**

```
%sql select substr(Date, 6,2) as Month, Landing_Outcome,Booster_Version, Launch_Site from SPACEXTABLE
    where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015'
```

[38]

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

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

# 2015 Launch Records

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in the year 2015.

```
%sql SELECT Landing_Outcome, COUNT(*) AS Count, RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC;
```

[43]

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

| Landing_Outcome | Count | Rank |
|---|---|---|
| No attempt | 10 | 1 |
| Success (drone ship) | 5 | 2 |
| Failure (drone ship) | 5 | 2 |
| Success (ground pad) | 3 | 4 |
| Controlled (ocean) | 3 | 4 |
| Uncontrolled (ocean) | 2 | 6 |
| Failure (parachute) | 2 | 6 |
| Precluded (drone ship) | 1 | 8 |

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

Ranking 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.

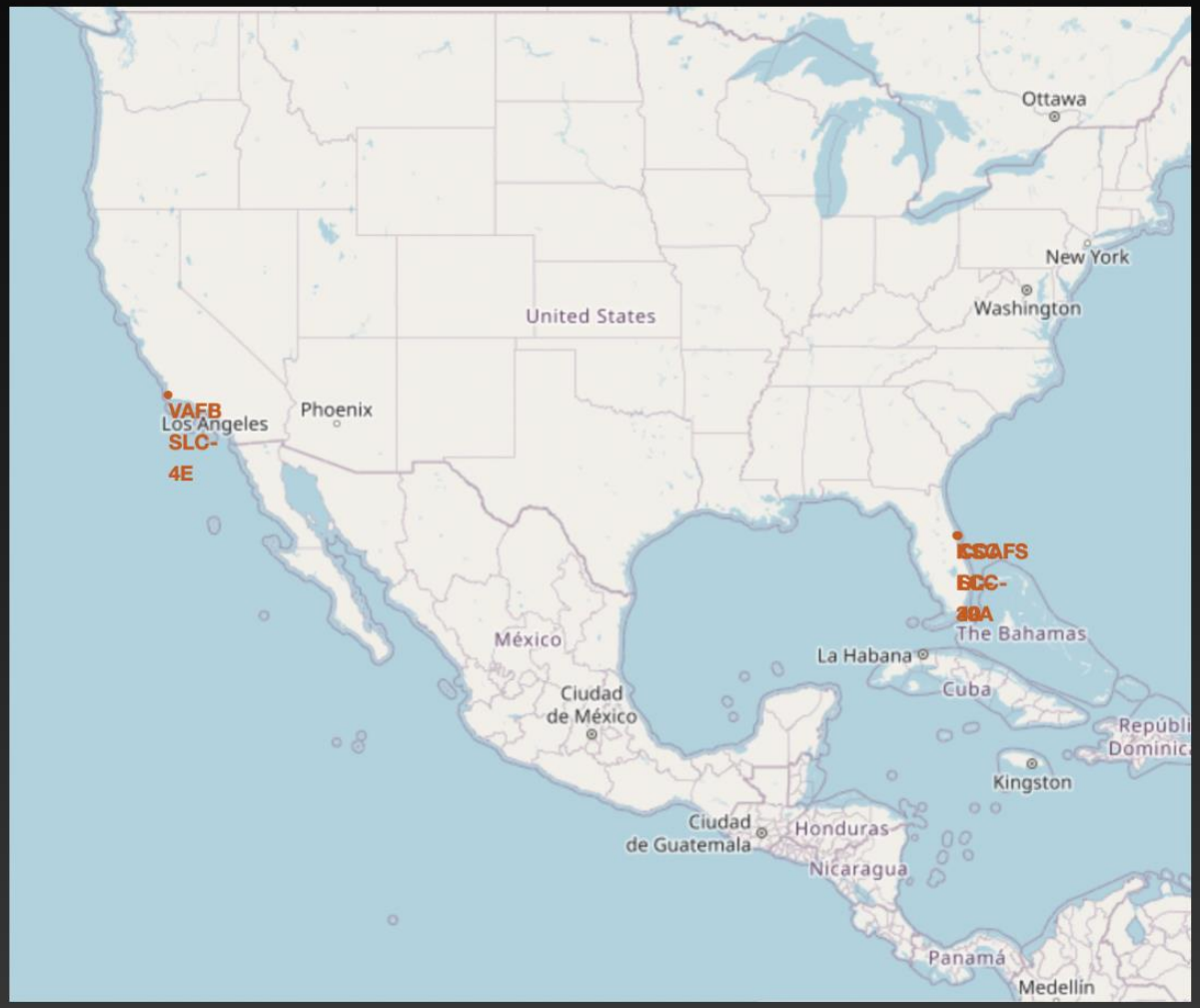
Section 3

# Launch Sites Proximities Analysis

# All launch sites' location markers on a global map

**Explanation:**

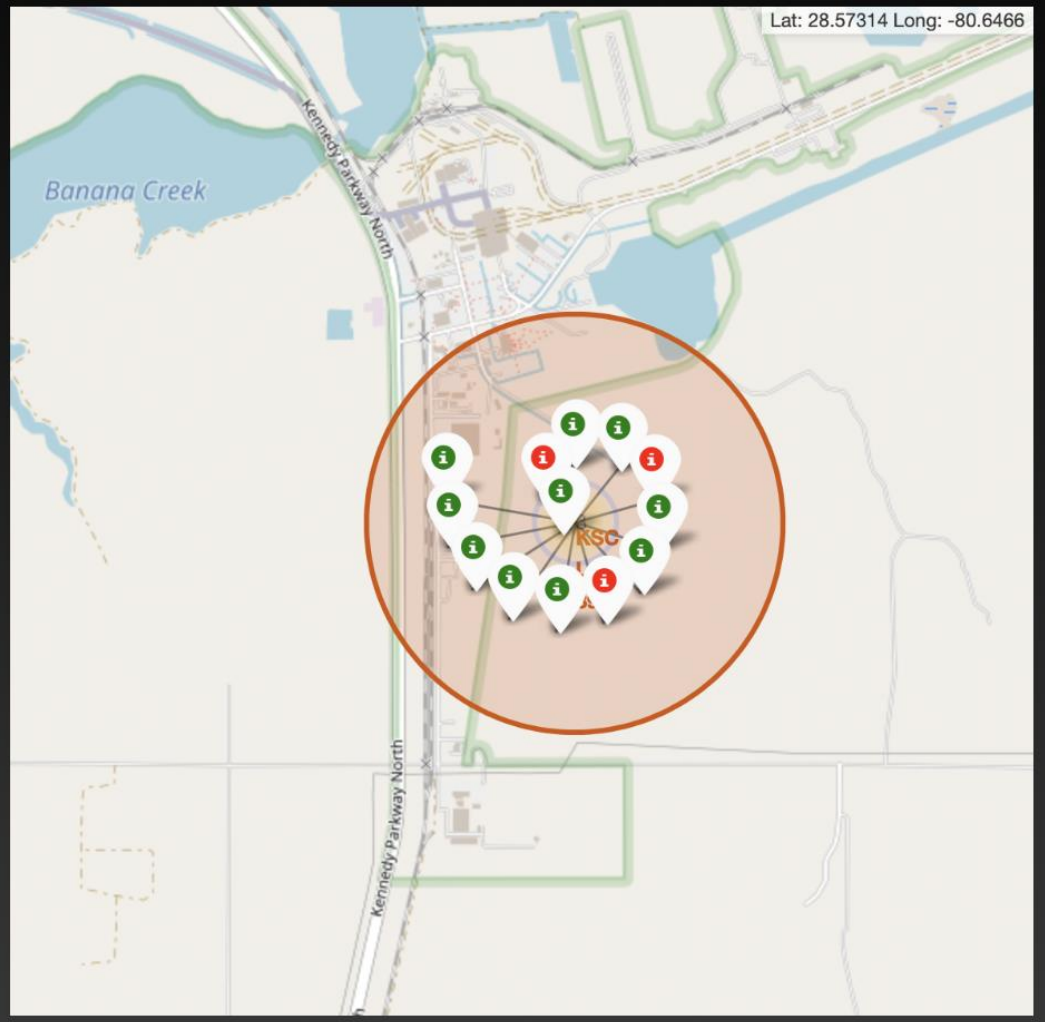
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



# Colour-labeled launch records on the map

## Explanation:

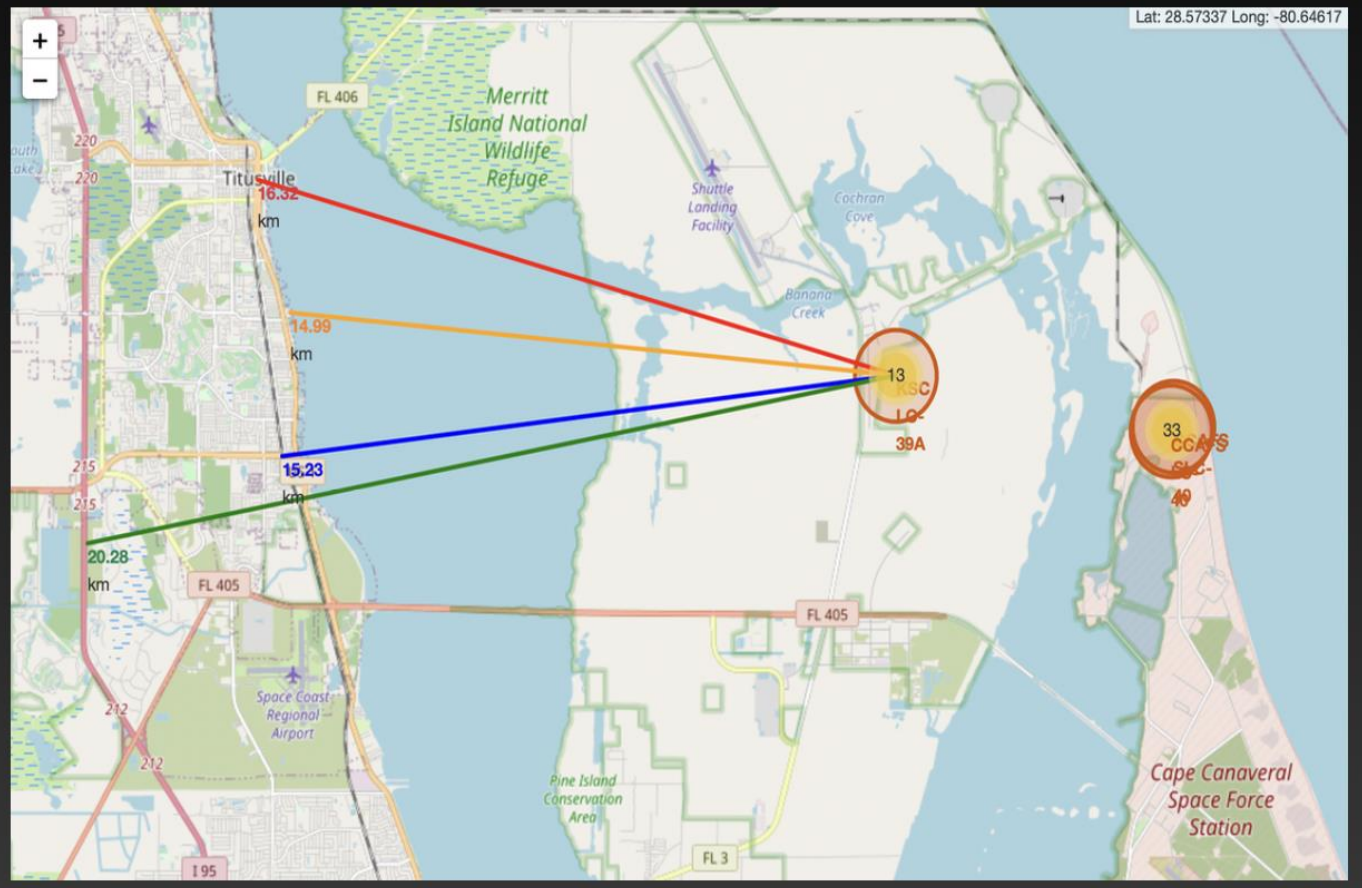
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

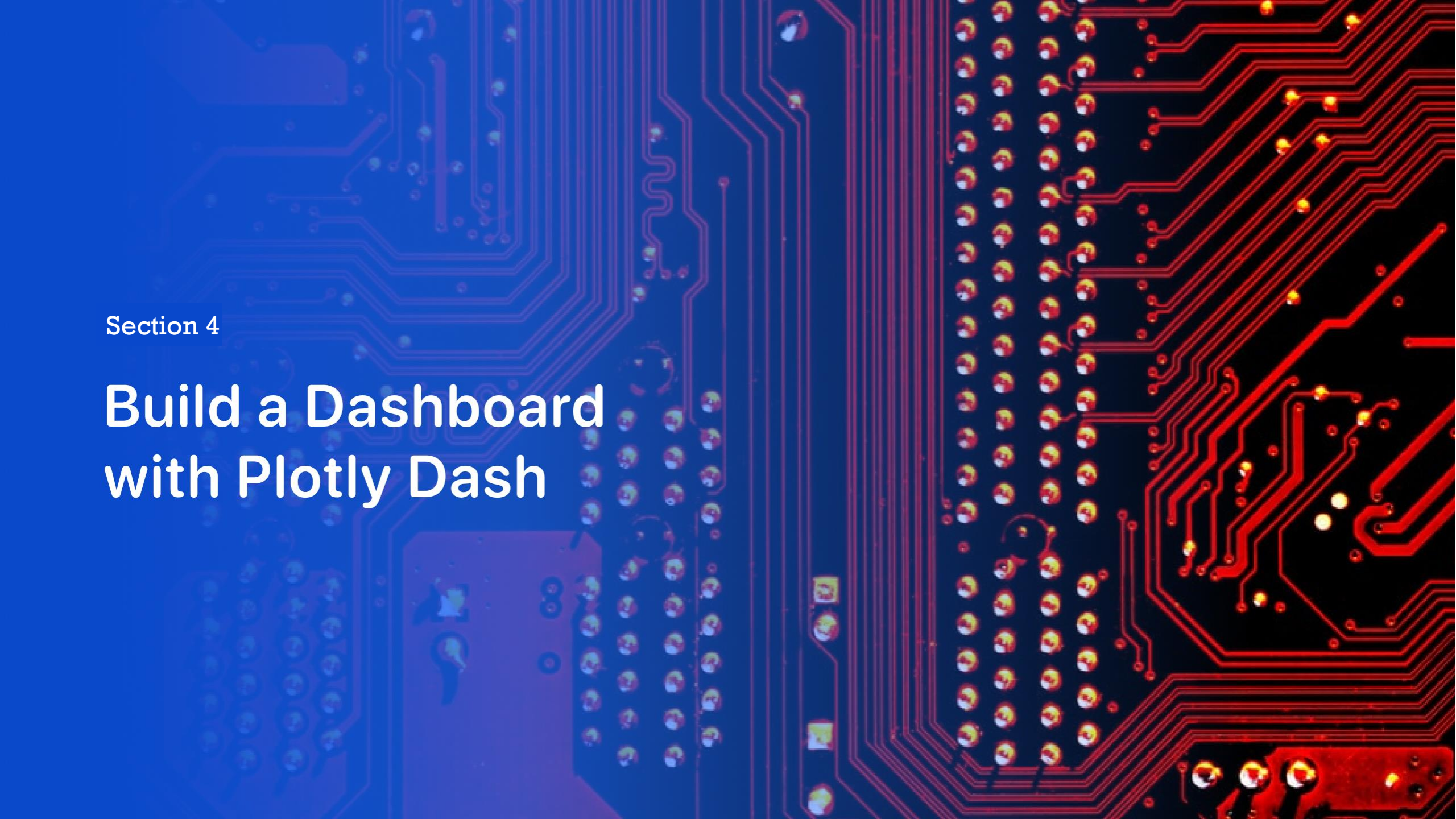


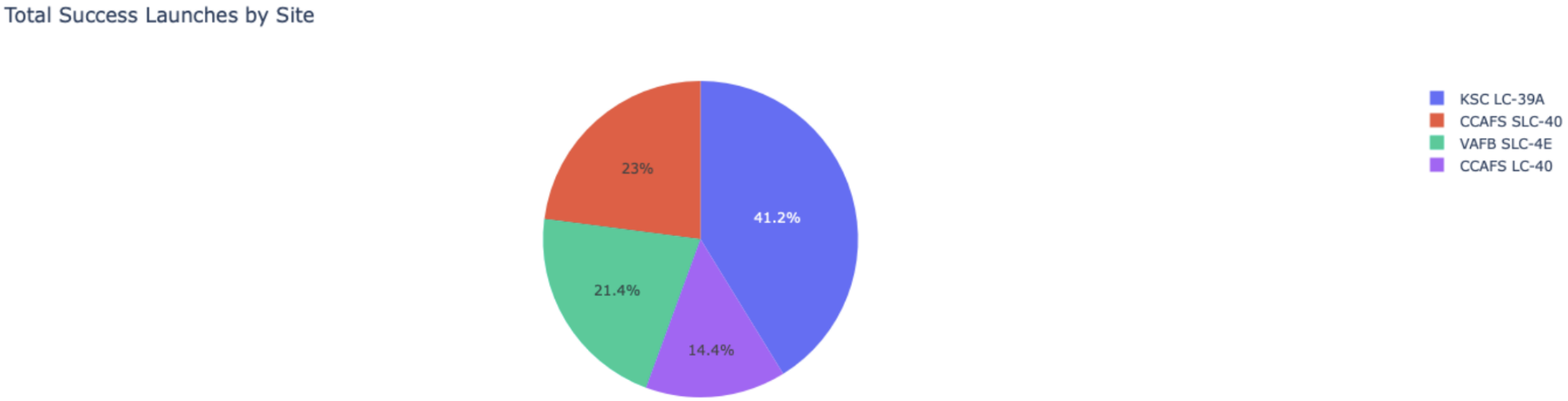
# Distance from the launch site KSC LC-39A to its proximities

**Explanation:**

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

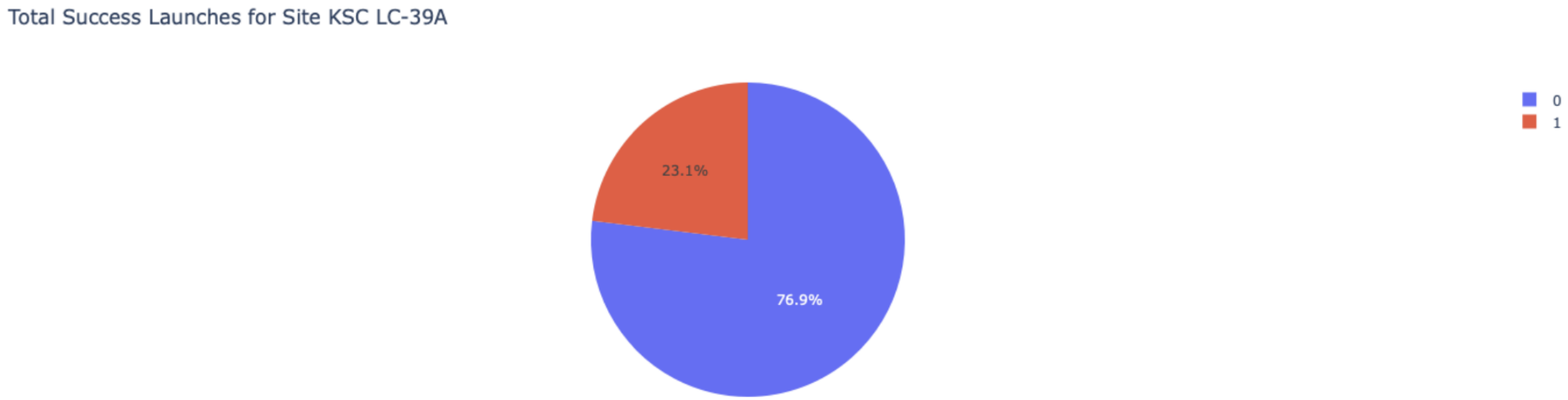



Section 4
Build a Dashboard with Plotly Dash




# Launch success count for all sites

The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches




# Launch site with highest launch success ratio

KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings


# Payload Mass vs. Launch Outcome for all sites

Explanation: The charts show that payloads between 2000 and 5500 kg have the highest success rate

Section 5

# Predictive Analysis (Classification)

| | LogReg | SVM | Tree |
|---|---|---|---|
| Jaccard_Score | 0.833333 | 0.845070 | 0.833333 |
| F1_Score | 0.909091 | 0.916031 | 0.909091 |
| Accuracy | 0.866667 | 0.877778 | 0.866667 |

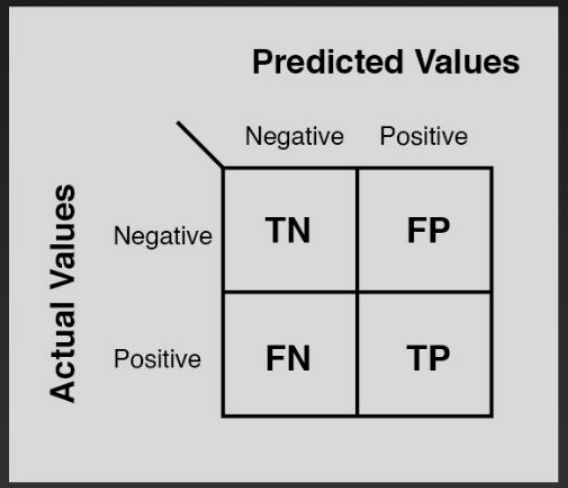
# Classification accuracy

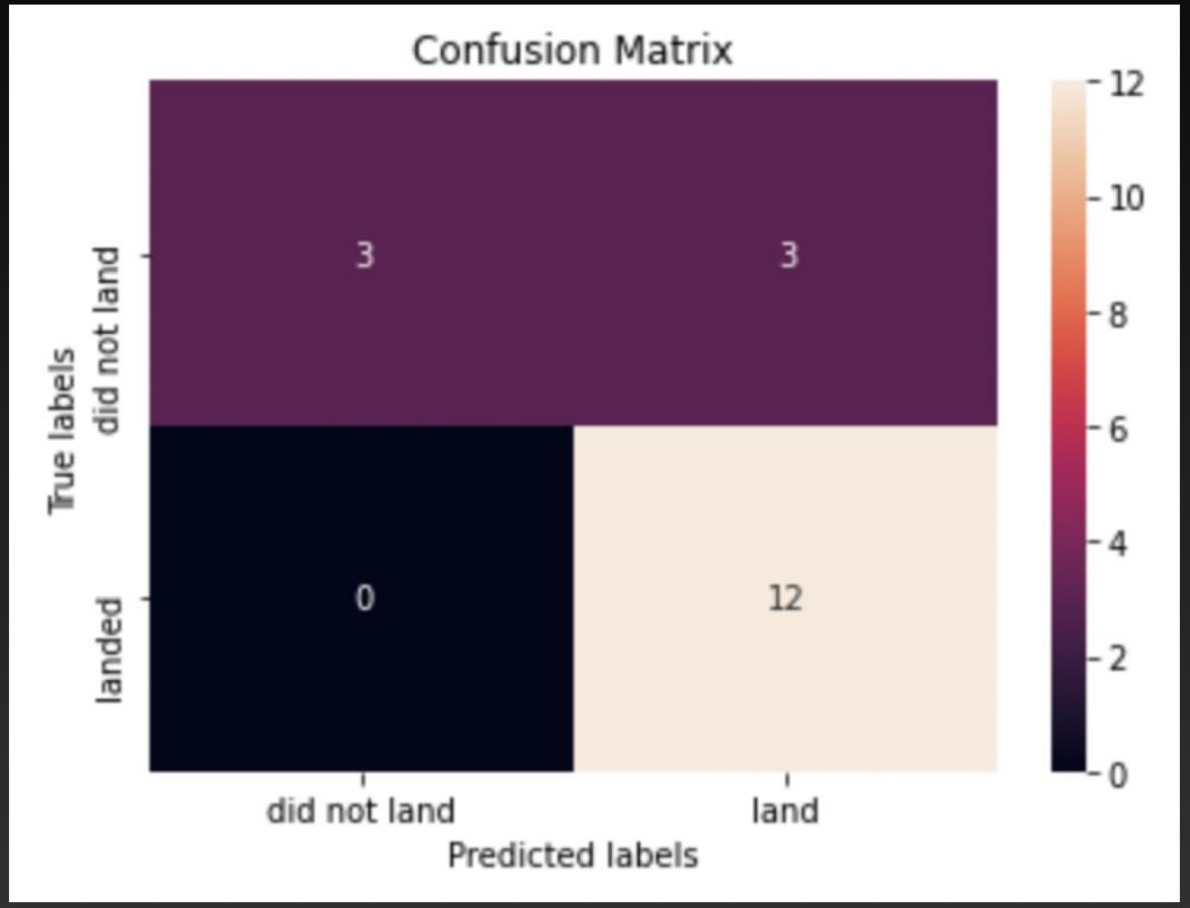
- The scores of each model on the whole Dataset confirm that the best model is the SVM Model. This model has not only higher scores but also the highest accuracy.

# Confusion Matrix

**Explanation:**

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



# Conclusions

- The SVM Model is the best algorithm for this dataset.
- • Launches with a low payload mass show better results than launches with a larger payload mass.
- • Most of the launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- • The success rate of launches increases over the years.
- • KSC LC-39A has the highest success rate of the launches from all the sites.
- • Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

# Appendix

- I have created a few extra charts and SQL queries during this project, you can find them in my GitHub repository. Please use the below link to access my git repo.

IBM-Data-Science-Capstone

# Thank you!