IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

**Hashim Shaikh**
**28 September 2025**

# Outline

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

# Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with Data 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 analysis 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, 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Problems you want to find answers**

  - 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 Scraping 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**

  Building, tuning and evaluation of classification models to ensure the best results

# Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

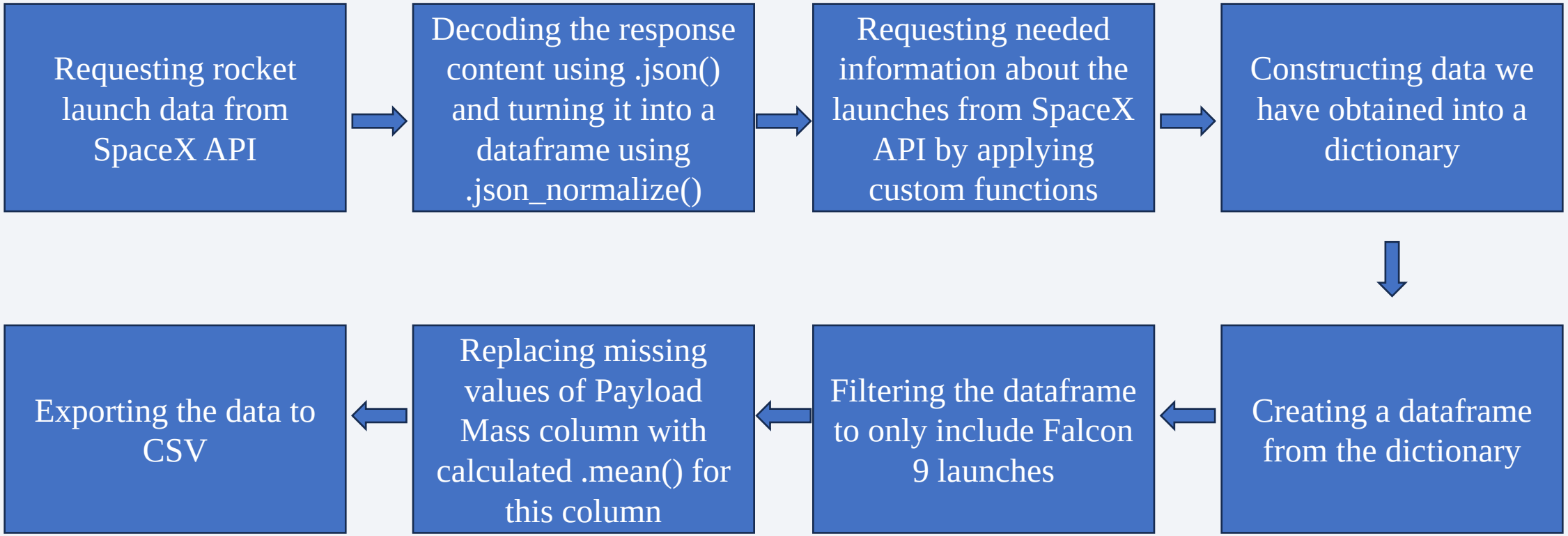
## Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

## Data Columns are obtained by using Wikipedia Web Scraping:

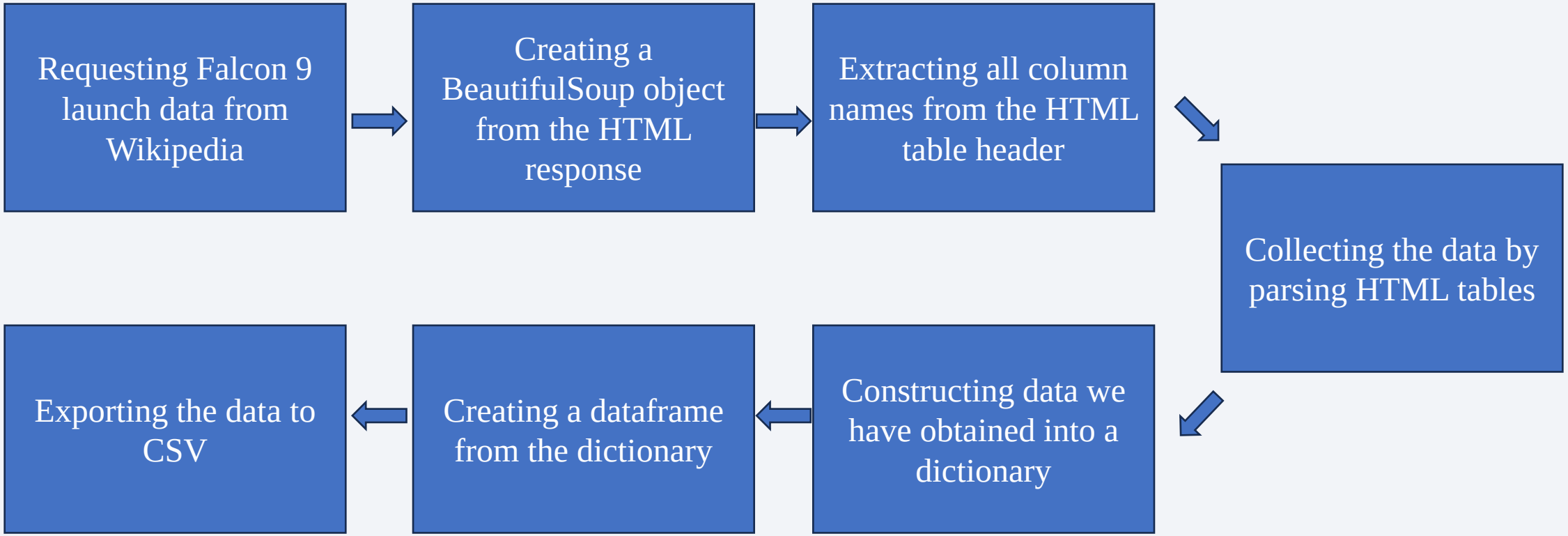
Flight Number, Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

# Data Collection – SpaceX API




GitHub URL: Data Collection API

# Data Collection – Web Scraping

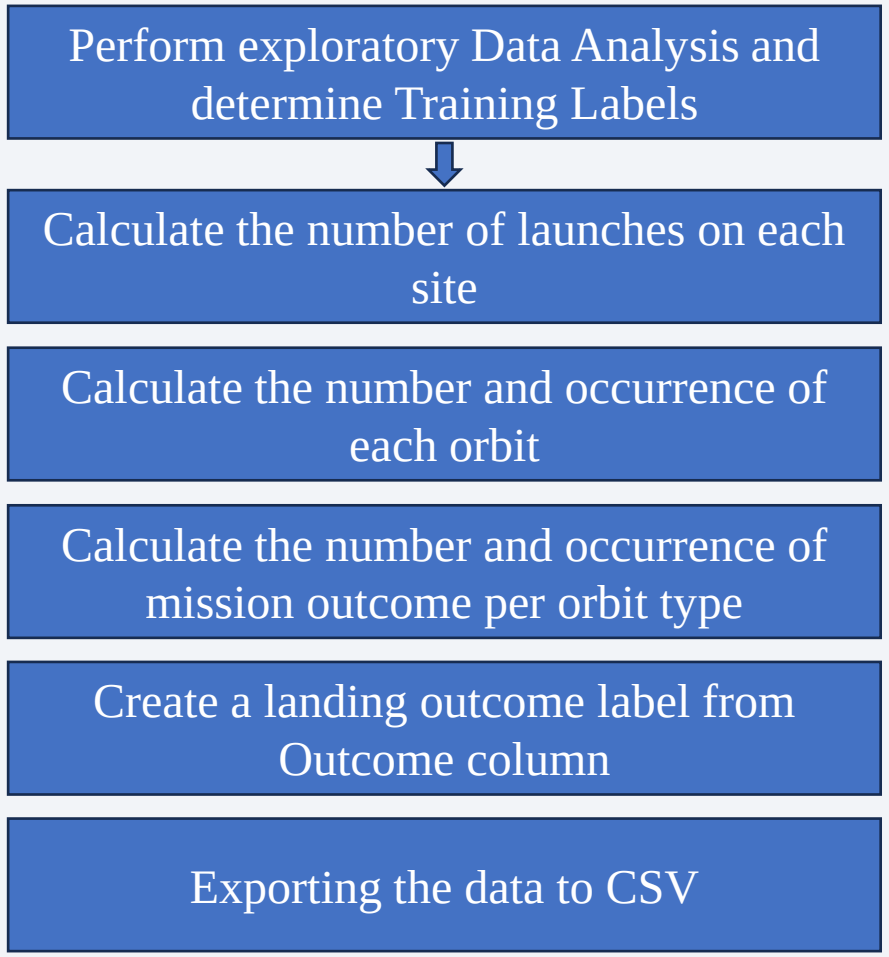



GitHub URL: Data Collection Web Scraping

# Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.




GitHub URL: Data Wrangling

# 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 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 URL: EDA with Data Visualization

# 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 ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which 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 year 2015
  - 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

GitHub URL: EDA with SQL

# 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 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 URL: Interactive Visual Analytics with Folium

# 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:**

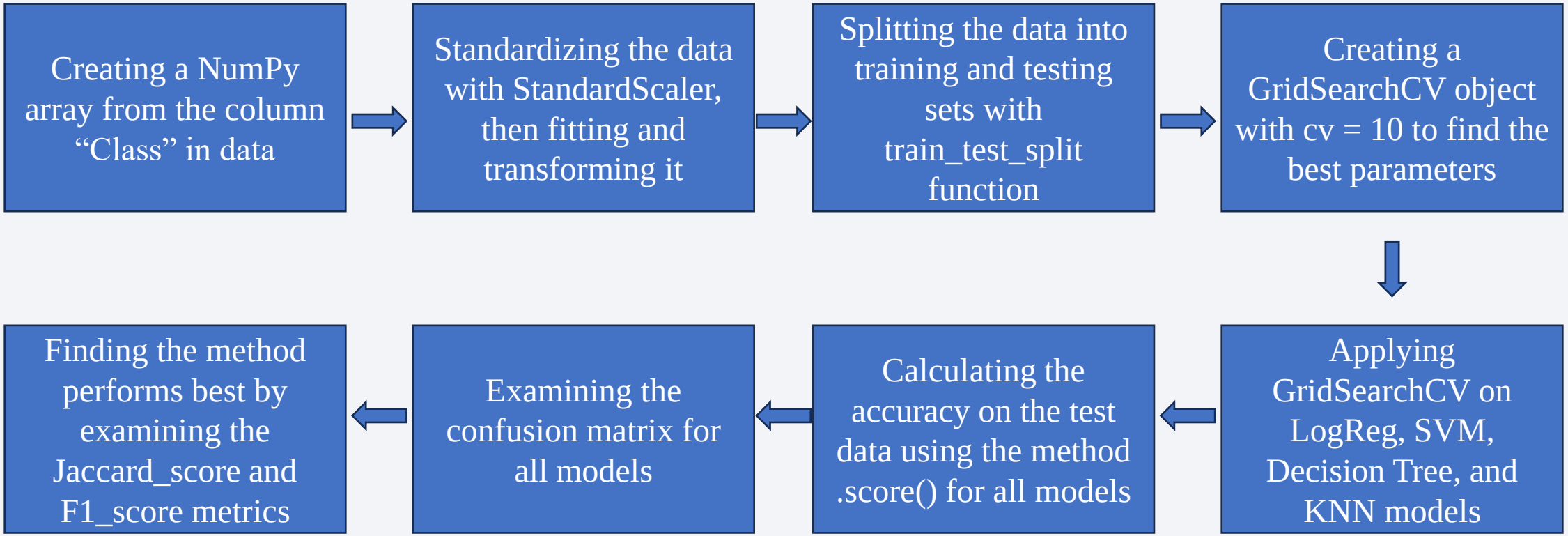
  - Added a slider to select Payload range.

- **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 URL: Plotly Dash App

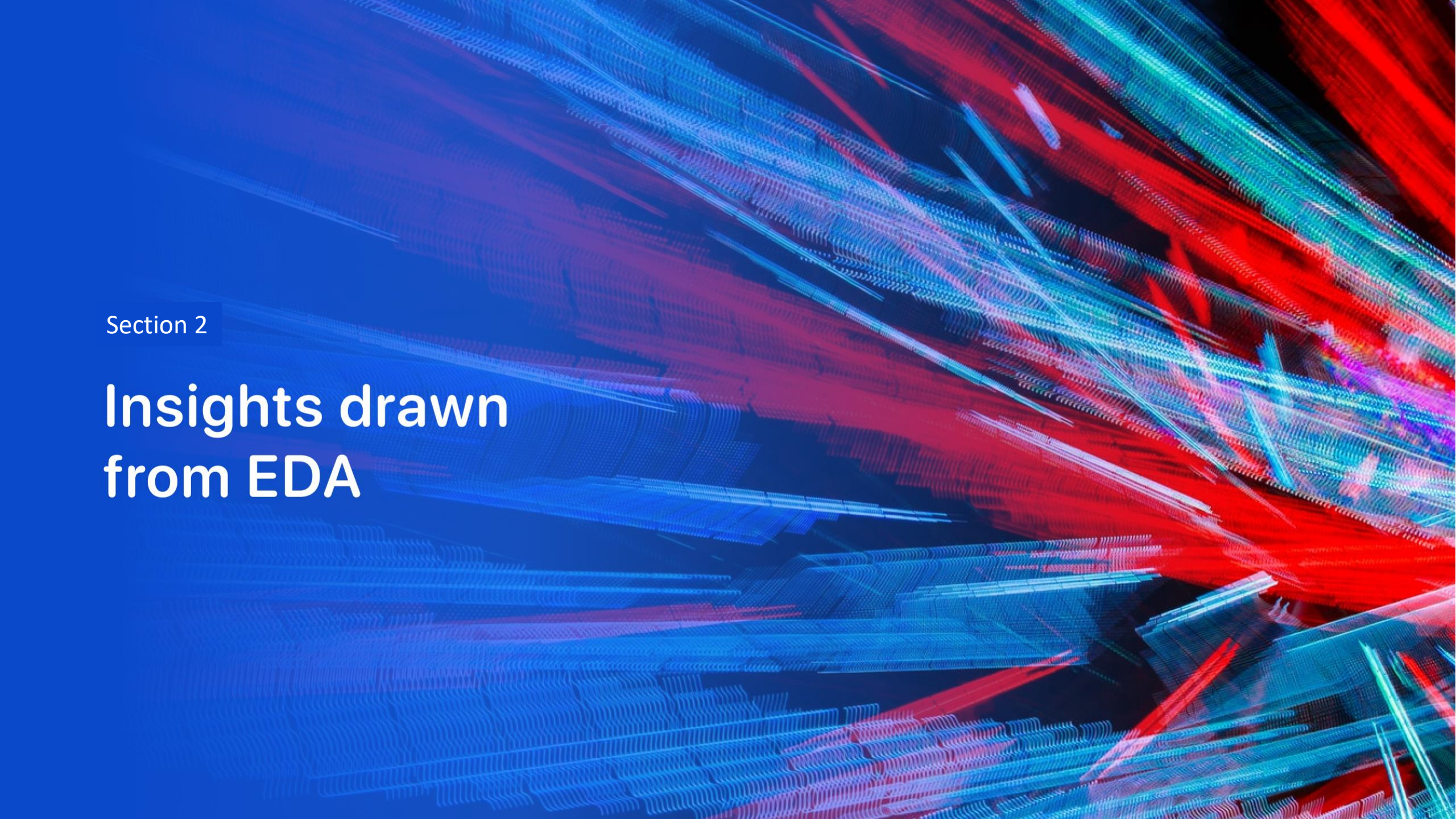
# Predictive Analysis (Classification)



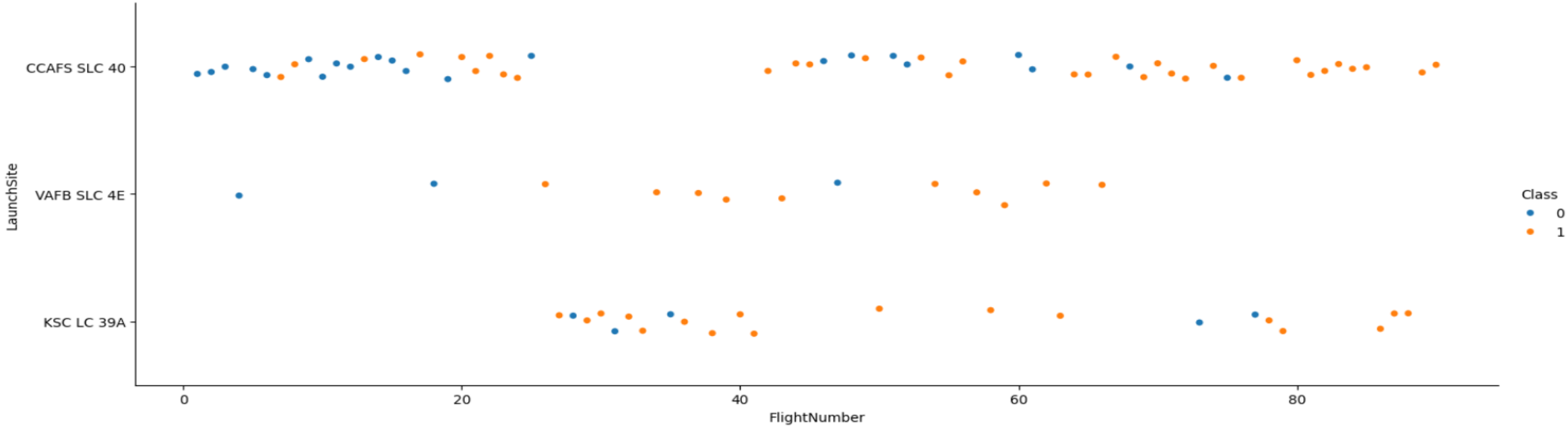

GitHub URL: Machine Learning Prediction

# Results

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


Section 2
Insights drawn from EDA

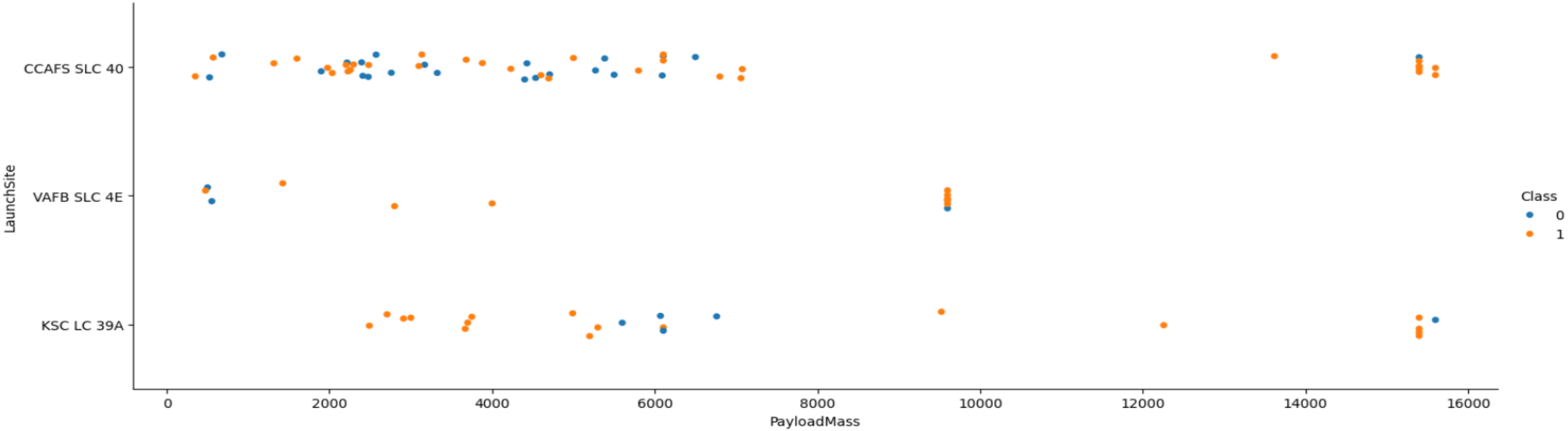
# Flight Number vs. Launch Site




**Explanation:**

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a 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



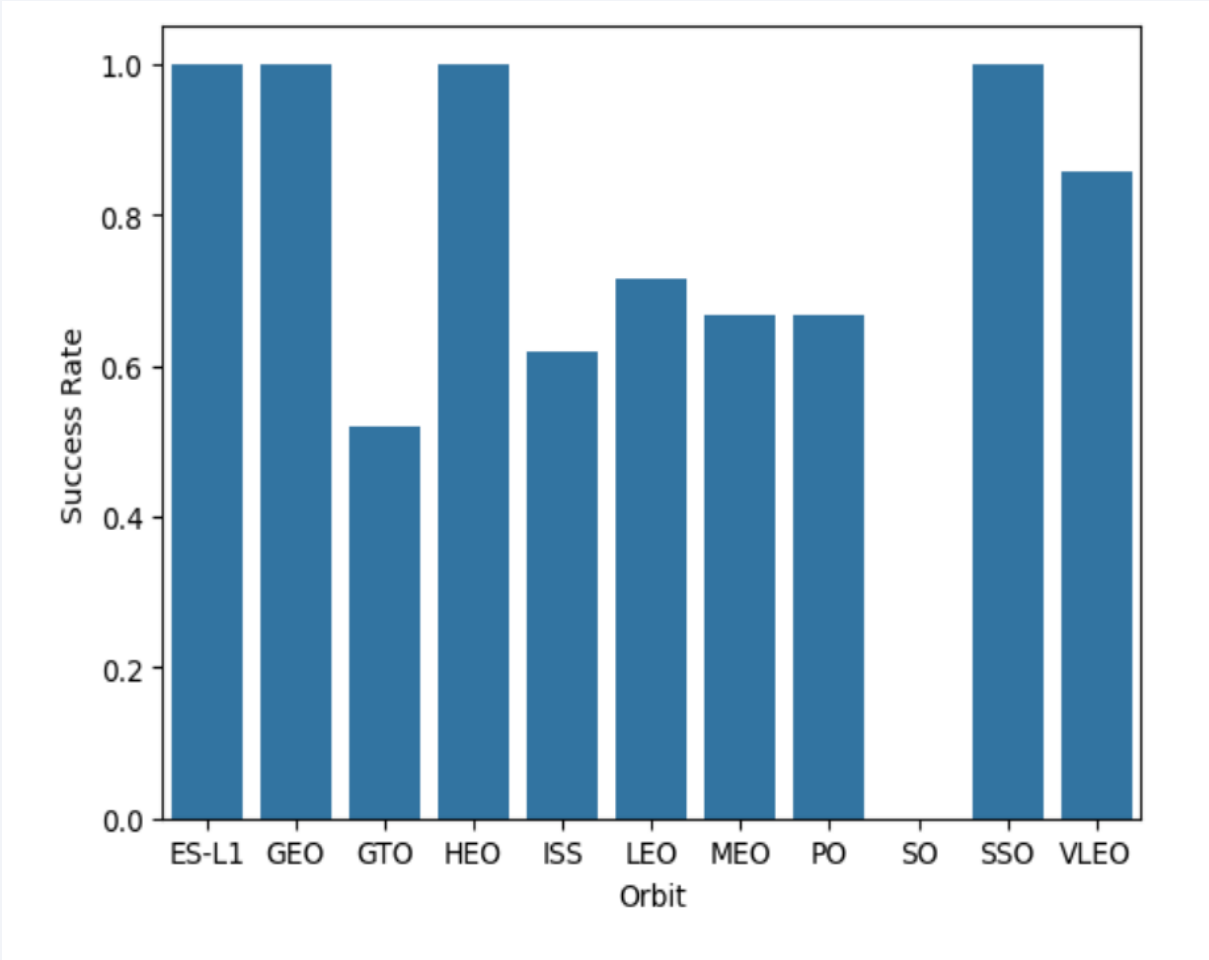

## Explanation:

- 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.

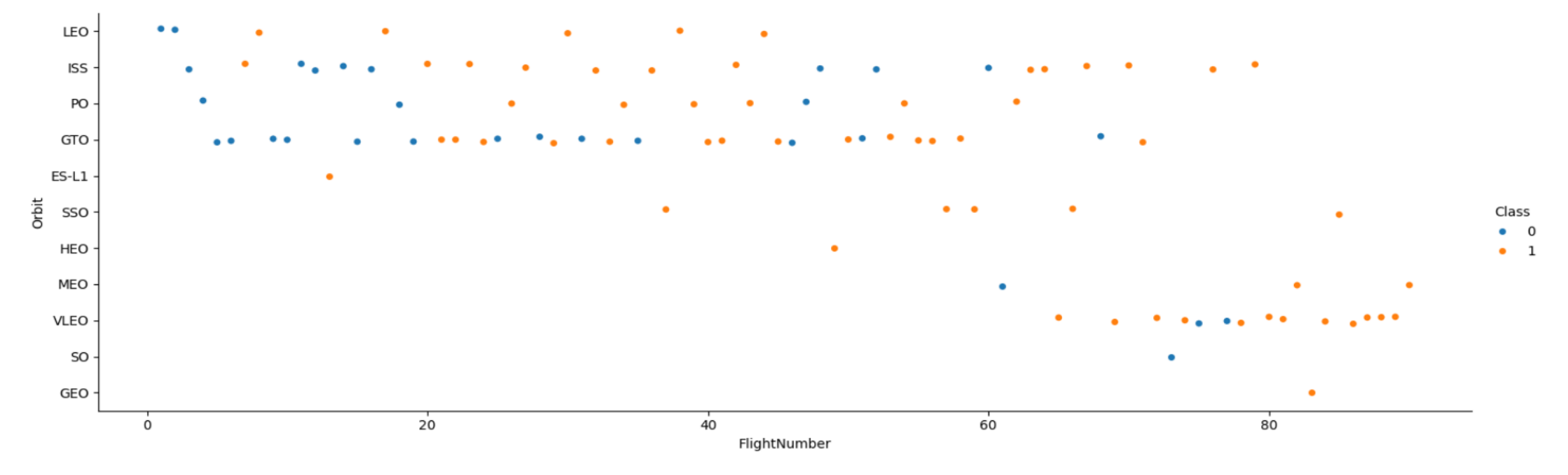
# Success Rate vs. Orbit Type

**Explanation:**

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO



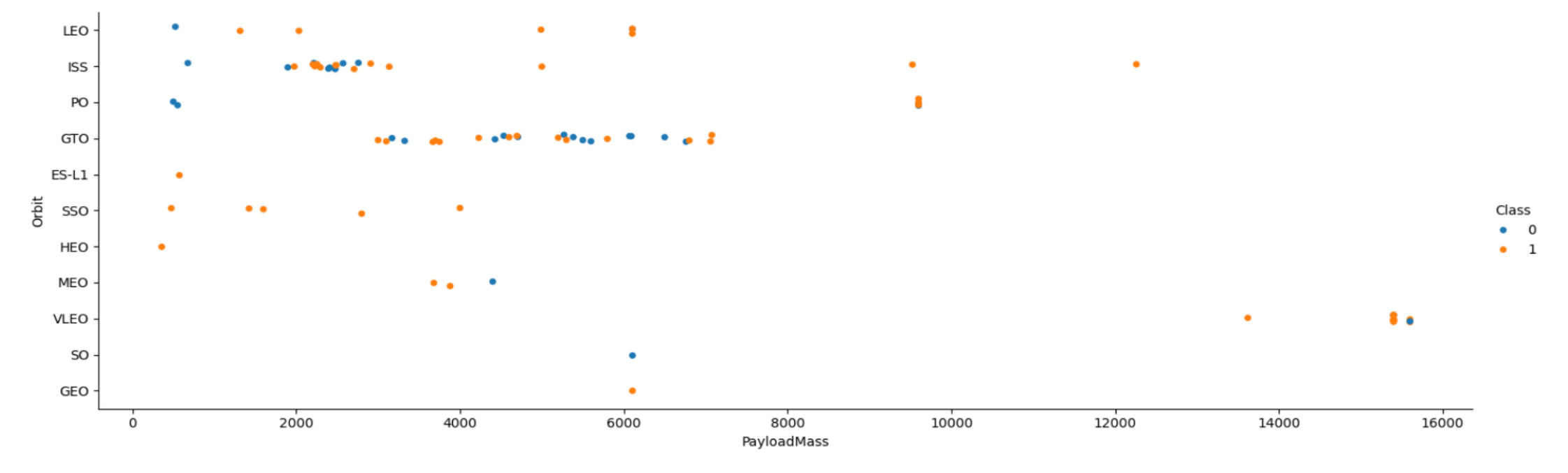
# Flight Number vs. Orbit Type




**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



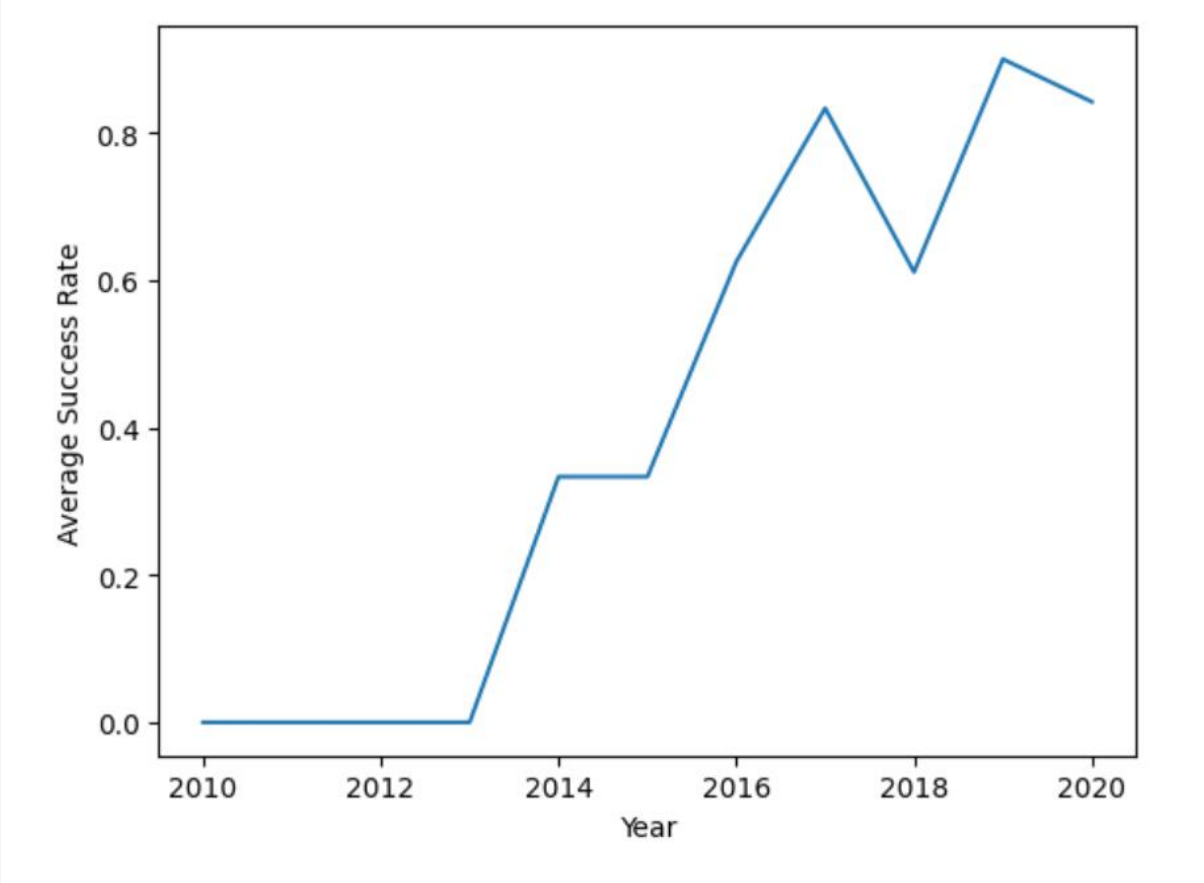

**Explanation:**

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

# Launch Success Yearly Trend

**Explanation:**

- The success rate since 2013 kept increasing till 2020.



# All Launch Site Names

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

[11]:

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

**Explanation:**

- Displaying the names of the unique launch sites in the space mission.

# Launch Site Names Begin with 'CCA'

```
[14]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[14]:

| 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 |

**Explanation:**

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

# Total Payload Mass

```
[16]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
[16]: Total_Payload
45596
```

**Explanation:**

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

# Average Payload Mass by F9 v1.1

```
[17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

[17]:

| Avg_Payload |
|---|
| 2928.4 |

**Explanation:**

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

# First Successful Ground Landing Date

```
[18]: %sql SELECT MIN(Date) AS First_Success_GroundPad FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[18]:

| First_Success_GroundPad |
|---|
| 2015-12-22 |

**Explanation:**

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

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
[20]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
 * sqlite:///my_data1.db
Done.
```

[20]:

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

## Explanation:

- Listing the 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

```
[21]: %sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

[21]:

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

**Explanation:**

- Listing the total number of successful and failure mission outcomes.

# Boosters Carried Maximum Payload

```
[22]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[22]:

| 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 |

**Explanation:**

- Listing the names of the booster versions which have carried the maximum payload mass.

# 2015 Launch Records

```
[24]: %sql SELECT substr(Date,6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Failure (drone ship)%'
```

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

[24]:

| 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 |

**Explanation:**

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

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

```
[25]: %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER
 * sqlite:///my_data1.db
Done.
```

[25]:

| Landing_Outcome | Outcome_Count |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

## Explanation:

- 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 on Global Map

**Explanation:**

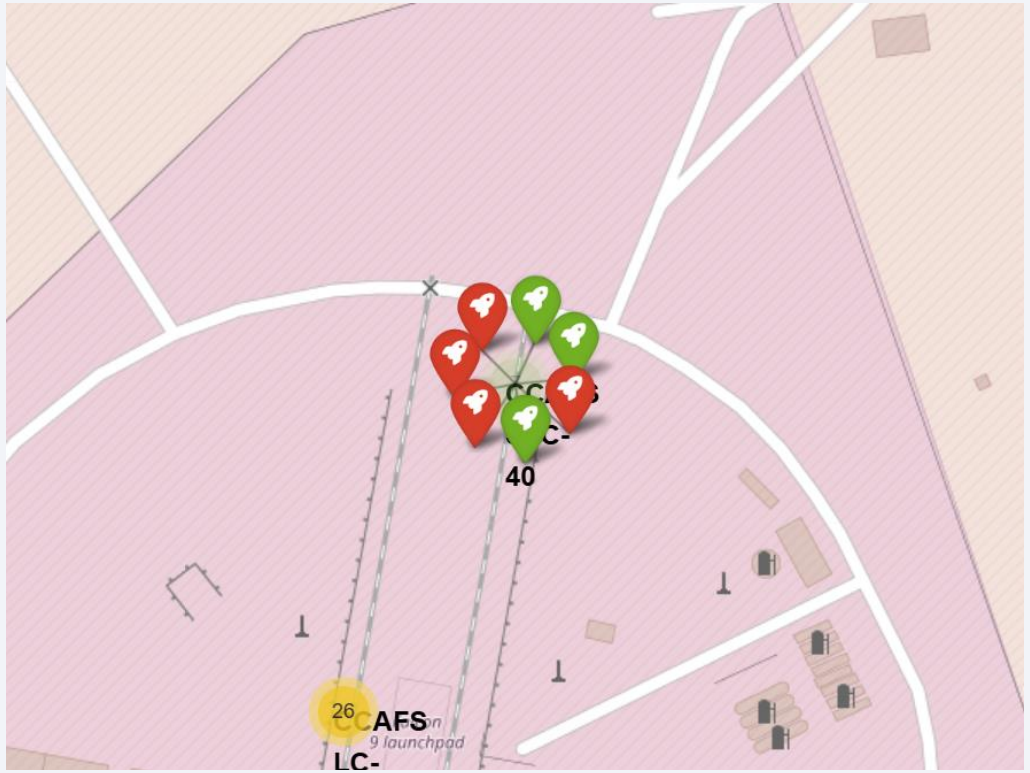
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, 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.

# Color-Labeled Launch outcomes on the Map

**Explanation:**

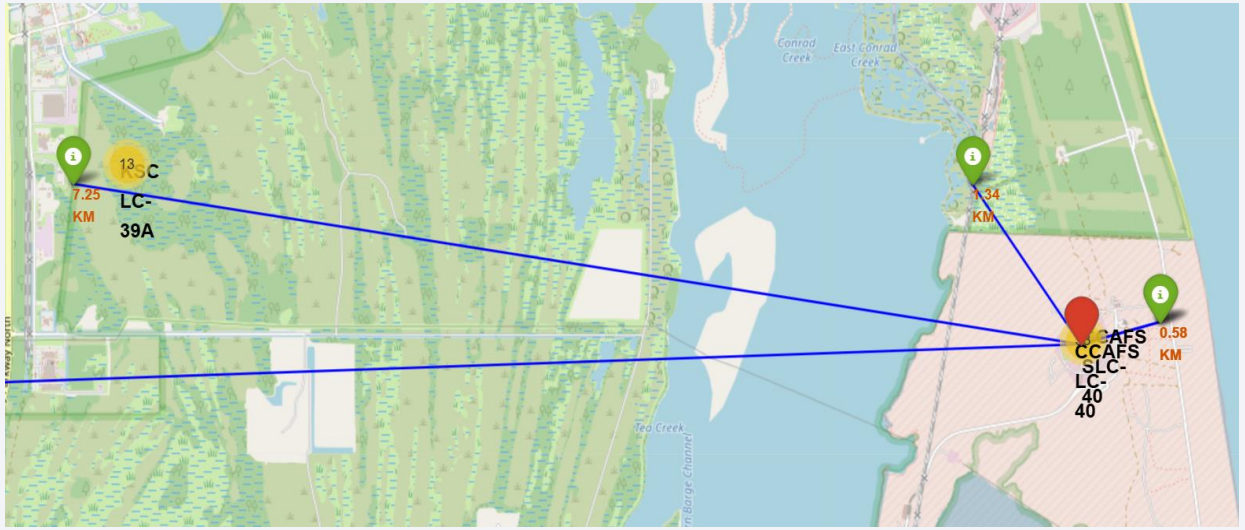
- Using the color-labeled markers, we can easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site CCAFS LC-40 shows a high success rate based on marker analysis, making it a significant site in the project.

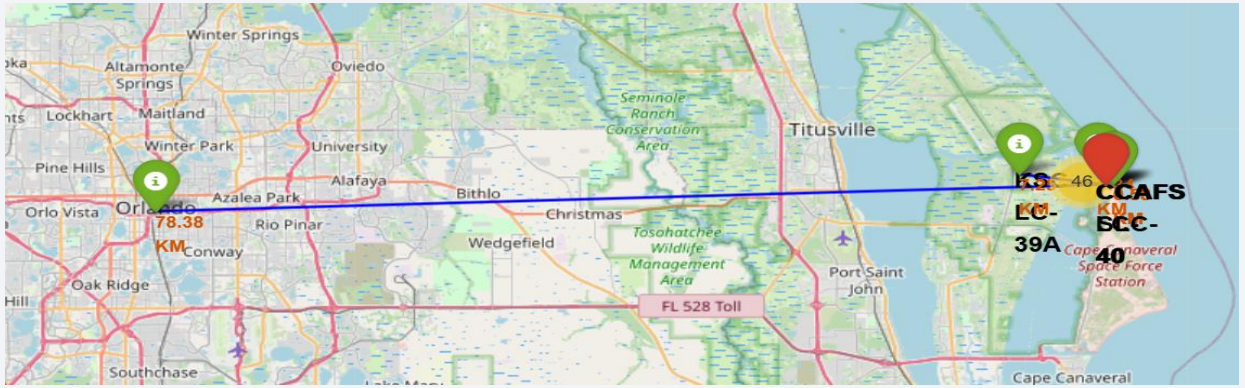


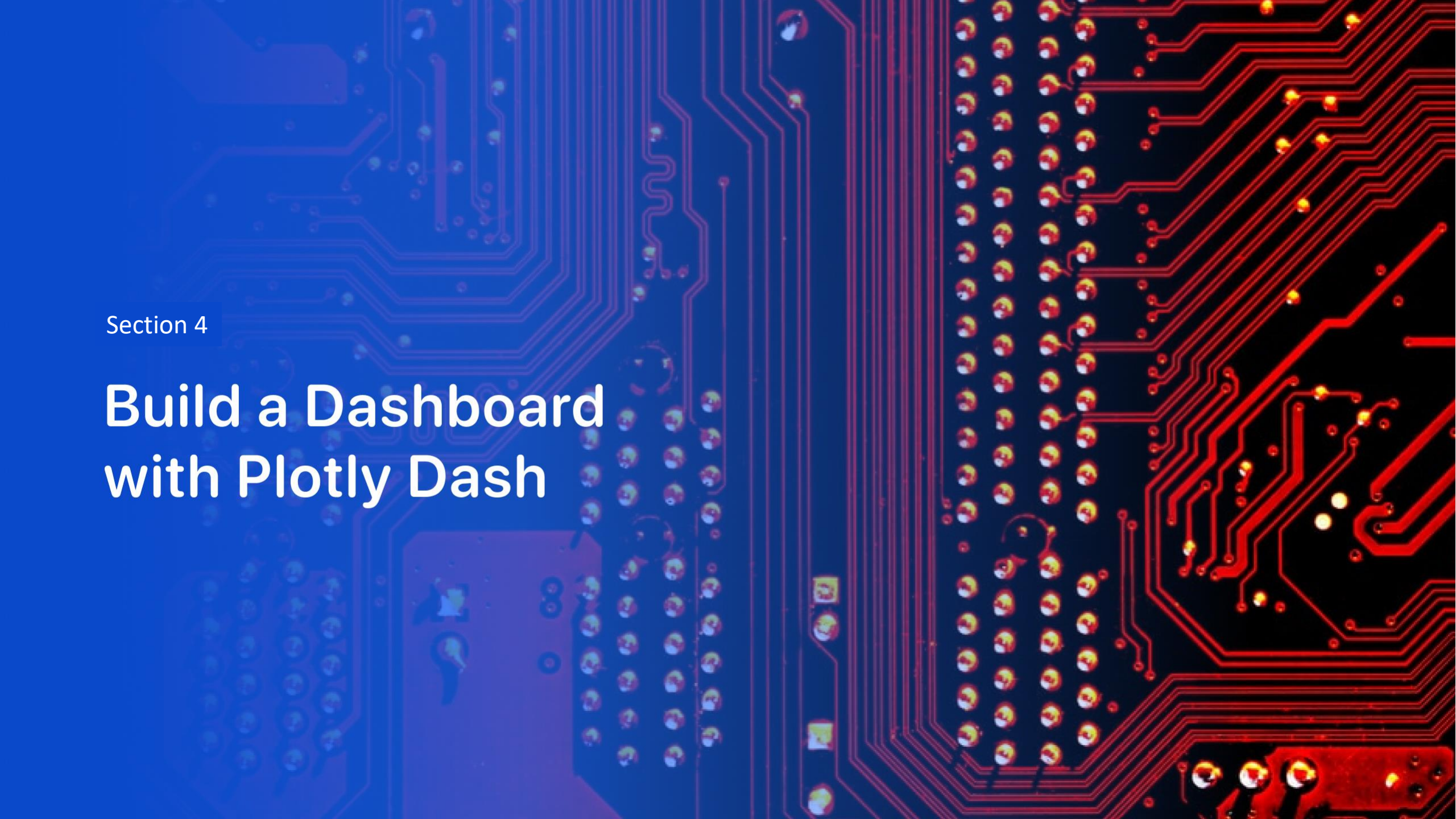
# Distance from the Launch Site CCAFS LC-40 to its proximities

**Explanation:**

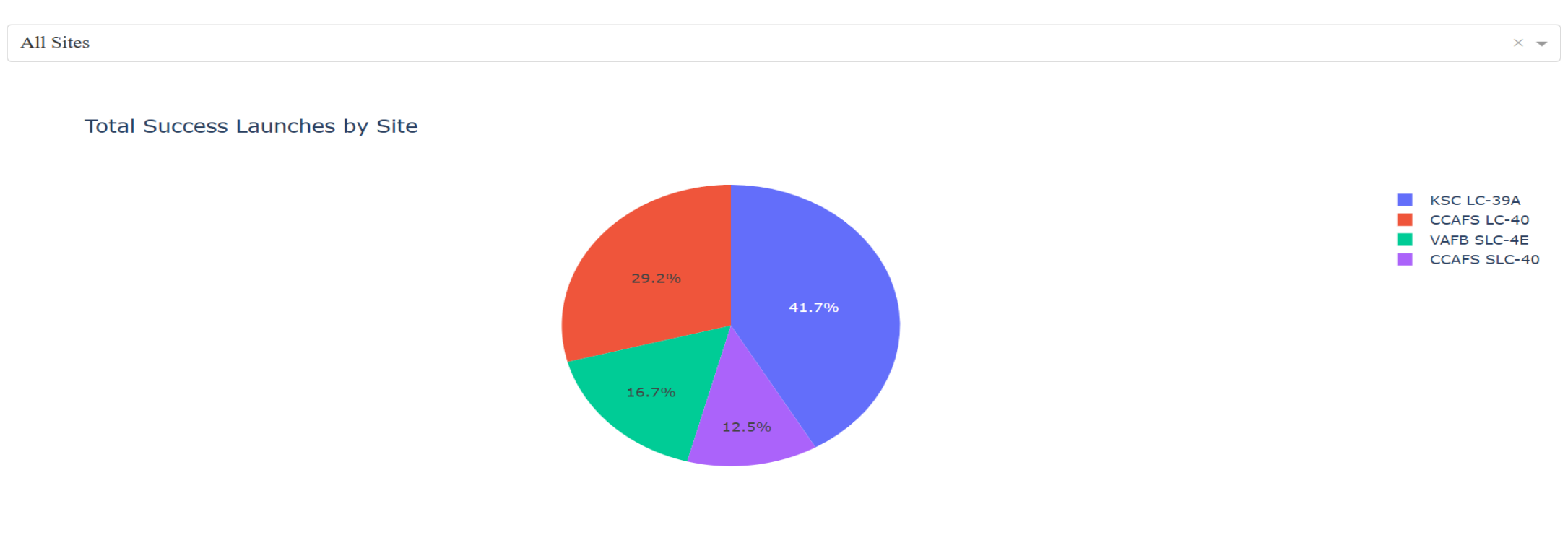
- From the visual analysis of the launch site CCAFS LC-40 we can clearly see that it is:
  very close to railway (1.34 km)
  extremely close to coastline (0.58 km)
  relative close to highway (7.25 km)
- However, the launch site CCAFS LC-40 is not close to its closest city Orlando (78.38 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas and nearby infrastructure.




Section 4
Build a Dashboard with Plotly Dash

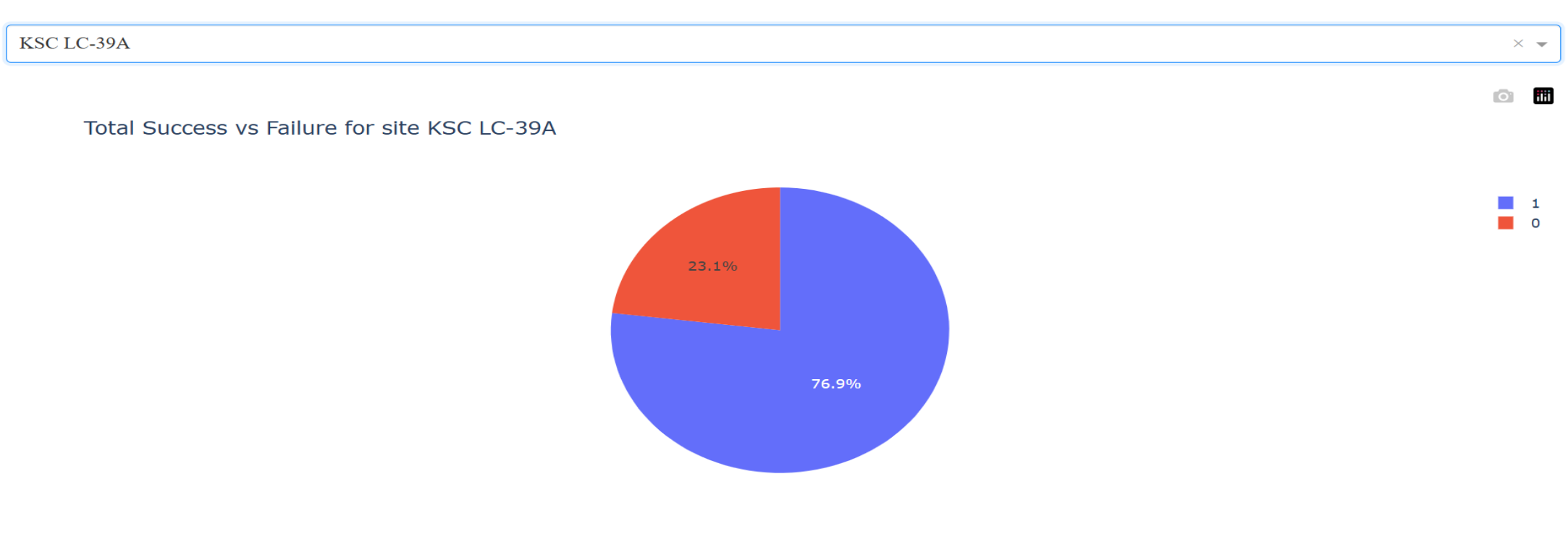
# Launch success count for all Sites




**Explanation:**

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

# Launch Site with highest launch success ratio



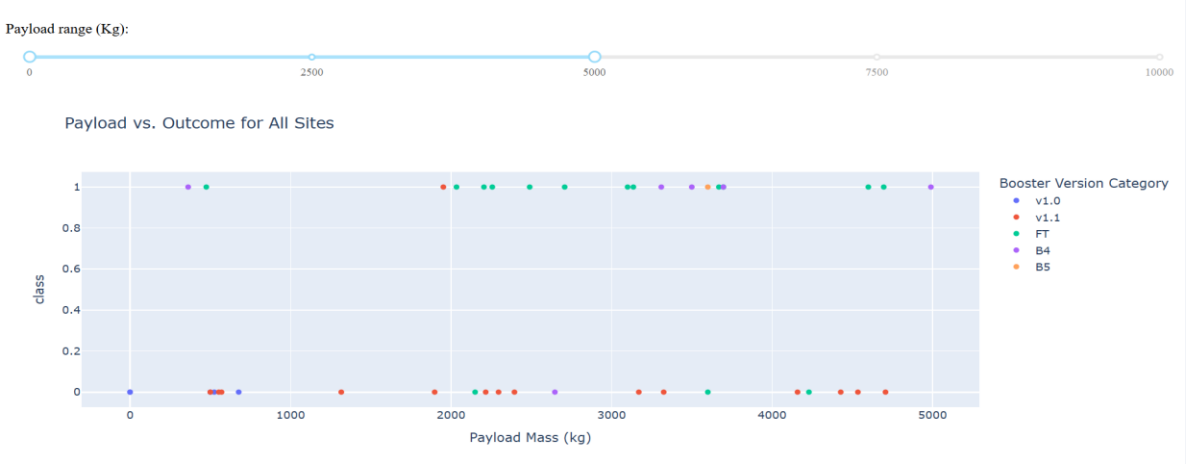

**Explanation:**

- 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)

# Classification Accuracy

```
[31]: print("Final Test Accuracies:")
      print("Logistic Regression:", logreg_cv.score(X_test, Y_test))
      print("SVM:", svm_cv.score(X_test, Y_test))
      print("Decision Tree:", tree_cv.score(X_test, Y_test))
      print("KNN:", knn_cv.score(X_test, Y_test))

      Final Test Accuracies:
      Logistic Regression: 0.8333333333333334
      SVM: 0.8333333333333334
      Decision Tree: 0.8333333333333334
      KNN: 0.8333333333333334
```

**Explanation:**

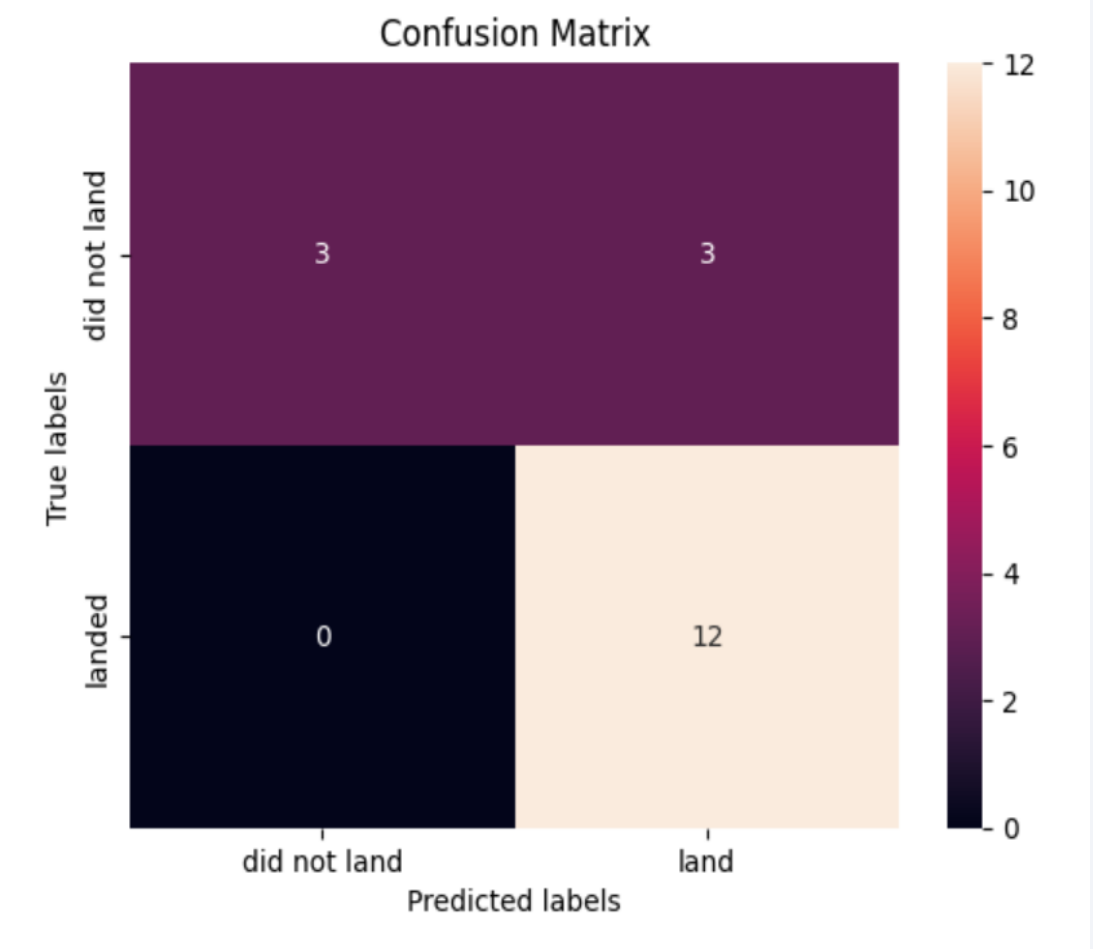
- All four models — Logistic Regression, SVM, Decision Tree, and KNN — achieved the same Test Set accuracy of 0.8333.
- Since the performance is identical, we cannot determine which method performs best based solely on the Test Set results.
- This outcome may also be influenced by the small Test Set size (18 samples), which limits the ability to distinguish model performance effectively.
- Therefore, the results indicate that all models are comparable on the Test Set, and no single model can be confirmed as the best performer in this case.

# Confusion Matrix

**Explanation:**

- The Confusion Matrix shows the prediction performance of the model on the Test Set (18 samples).
- Out of the 18 samples:
  - The model correctly classified 12 landings (True Positives).
  - The model correctly identified 3 non-landings (True Negatives).
  - The model misclassified 3 non-landings as landings (False Positives).
  - There were no landings misclassified as non-landings (False Negatives = 0).

- This gives an overall accuracy of 15/18 = 0.8333 (83.3%), which matches the Test Set accuracy results.
- The results indicate that the model performs very well at detecting landings, but it has some difficulty in correctly predicting when a launch does not land.



# Conclusions

- Decision Tree 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 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

- This project was completed as part of the **Applied Data Science Capstone - IBM**.
- I would like to sincerely thank the IBM instructors for their valuable guidance and support throughout this project.

Instructors

IBM

# Thank you!