IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Jose Yanez
2024-02-24

# Outline

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

# Executive Summary

- **Summary of methodologies:** I worked on this capstone first making a Data Collection using web scraping and SpaceX API after it make Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics and finally I make Machine Learning Predictions.

- **Summary of all results:** Valuable data was gathered from public sources, enabling an insightful exploration through exploratory data analysis (EDA) to pinpoint the most predictive features for launch success. Machine learning prediction then determined the optimal model for effectively leveraging these characteristics, harnessing the entirety of the collected data to drive opportunities with precision.

# Introduction

- The main objective of this project is to evaluate the viability of the new company Space Y to compete with Space X performing data analysis of Space X information.
- Specific objectives:

  *Know the best way to estimate the total cost for make the launches predicting the successful landings of the first stage of rockets.

  *Know from the analysis where is the best place to make launches.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:

Data from SpaceX was sourced from two distinct channels: the SpaceX API (https://api.spacexdata.com/v4/rockets/) and through web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), ensuring a comprehensive dataset for analysis.

- Perform data wrangling:

The data collection process was enhanced by generating a landing outcome classification, which was derived from outcome data subsequent to summarizing and scrutinizing the features.

- Perform exploratory data analysis (EDA) using visualization and SQL:
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

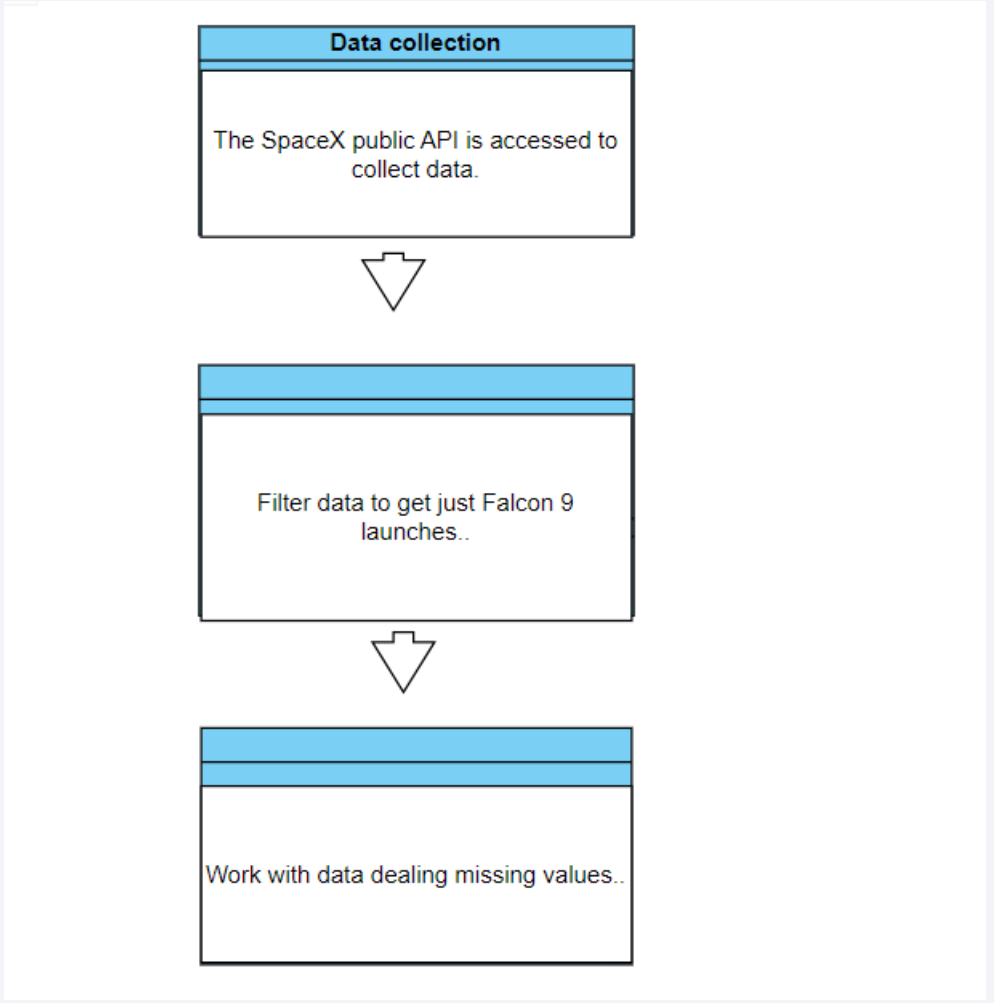
The data gathered up to this stage underwent normalization, partitioning into training and testing sets, and assessment using four distinct classification models. Each model's accuracy was assessed utilizing various parameter combinations to ensure thorough evaluation.

# Data Collection

- Data sets were acquired through two sources: the SpaceX API (https://api.spacexdata.com/v4/rockets/) and Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), employing web scraping techniques.

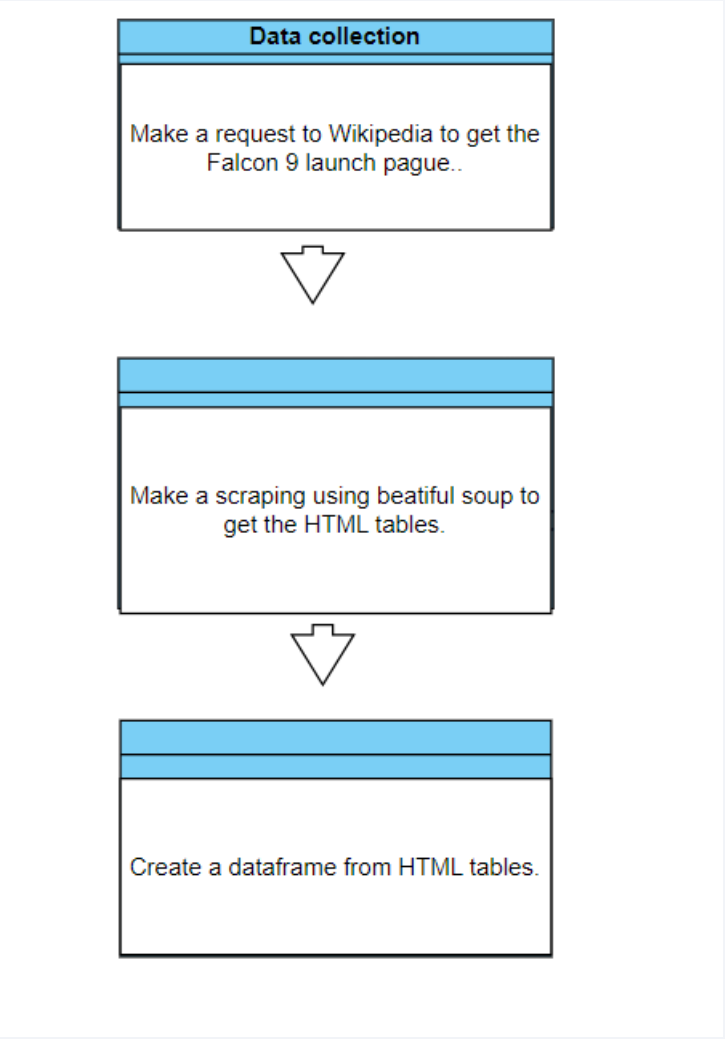
# Data Collection – SpaceX API

- SpaceX provides a publicly accessible API for obtaining data, which is subsequently utilized according to the accompanying flowchart, and the retrieved data is then persisted for further use.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/92680a34e1a0287a8c27bbf9e175d20c80f298b0/jupyter-labs-spacex-data-collection-api.ipynb



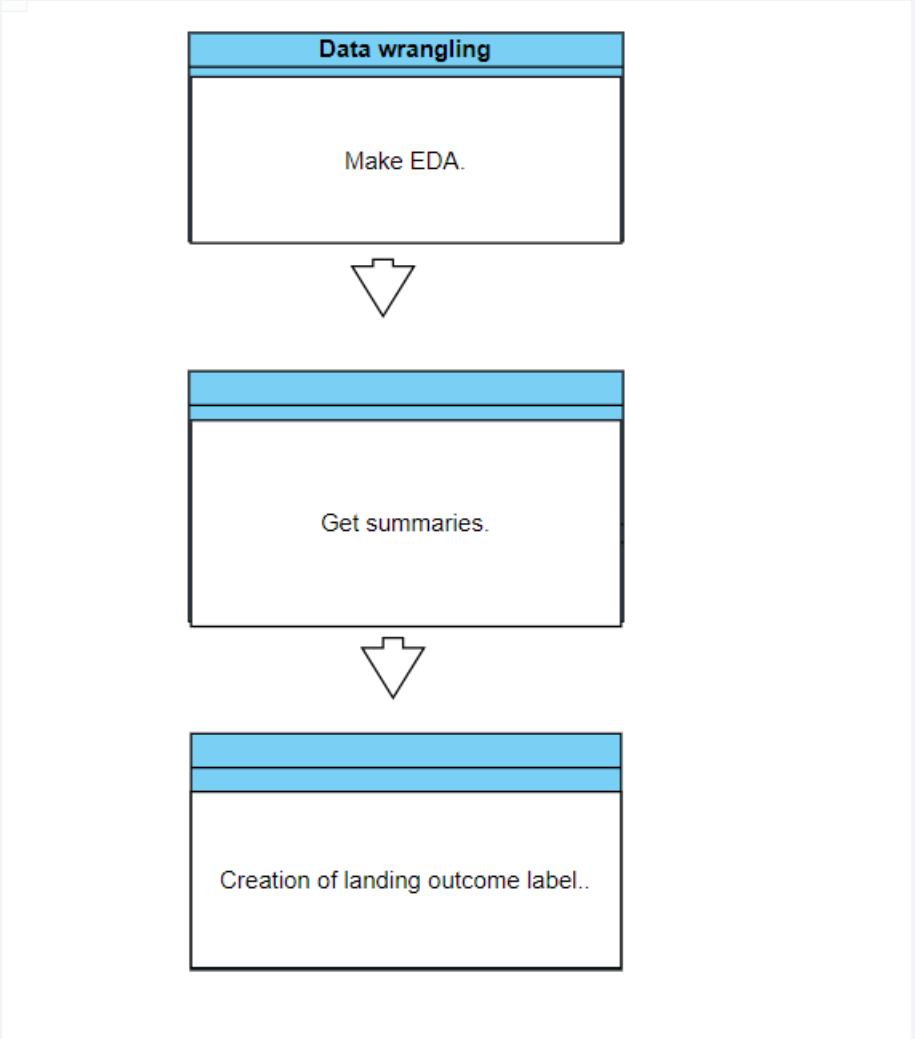
# Data Collection - Scraping

- Data regarding SpaceX launches can also be sourced from Wikipedia. The data is downloaded from Wikipedia following the established flowchart and then stored for future use.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/92680a34e1a0287a8c27bbf9e175d20c80f298b0/jupyter-labs-webscraping.ipynb



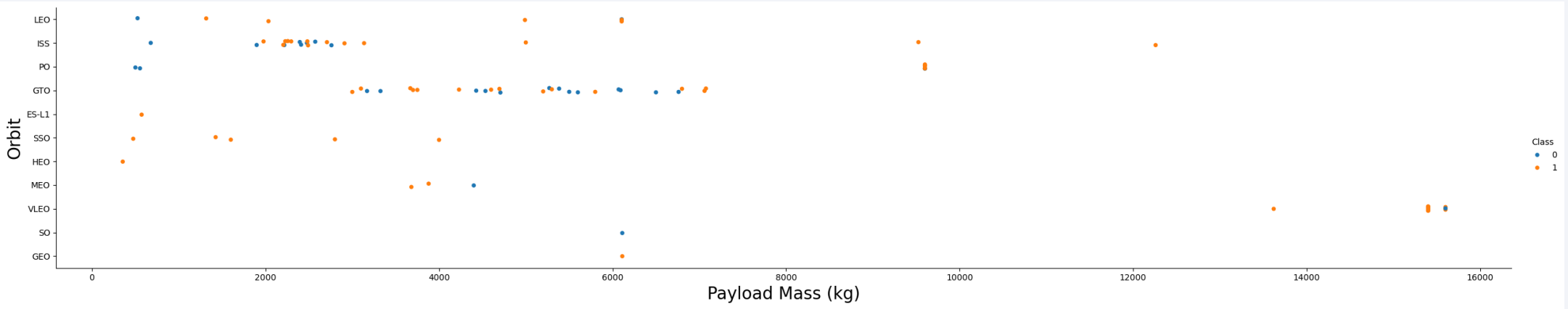
# Data Wrangling

- Initially, exploratory data analysis (EDA) was conducted on the dataset, involving the examination and interpretation of various data points. Subsequently, analyses were conducted to determine the frequency of launches per site, occurrences of each orbit, and the distribution of mission outcomes across different orbit types. Finally, a landing outcome label was generated based on the information in the Outcome column of the dataset.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/92680a34e1a0287a8c27bbf9e175d20c80f298b0/labs-jupyter-spacex-Data%20wrangling.ipynb



# EDA with Data Visualization

- Scatterplots and bar plots were employed to delve into the dataset, enabling the visualization of associations between pairs of features. These included Payload Mass versus Flight Number, Launch Site versus Flight Number, Launch Site versus Payload Mass, Orbit versus Flight Number, as well as Payload versus Orbit.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/92680a34e1a0287a8c27bbf9e175d20c80f298b0/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



# EDA with SQL

- SQL queries performed:
  - Identification of unique launch sites involved in space missions.
  - Determination of the top 5 launch sites with names starting with 'CCA'.
  - Calculation of the total payload mass transported by boosters launched by NASA (CRS).
  - Assessment of the average payload mass carried by booster version F9 v1.1.
  - Retrieval of the date when the first successful landing outcome occurred on a ground pad.
  - Compilation of the names of boosters achieving success on a drone ship with payload masses ranging between 4000 and 6000 kg.
  - Tabulation of the total number of successful and failed mission outcomes.
  - Listing of booster versions that have carried the maximum payload mass.
  - Identification of failed landing outcomes on a drone ship, along with their corresponding booster versions and launch site names for the year 2015.
  - Ranking of the count of landing outcomes (e.g., Failure on a drone ship or Success on a ground pad) between June 4, 2010, and March 20, 2017.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/92680a34e1a0287a8c27bbf9e175d20c80f298b0/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

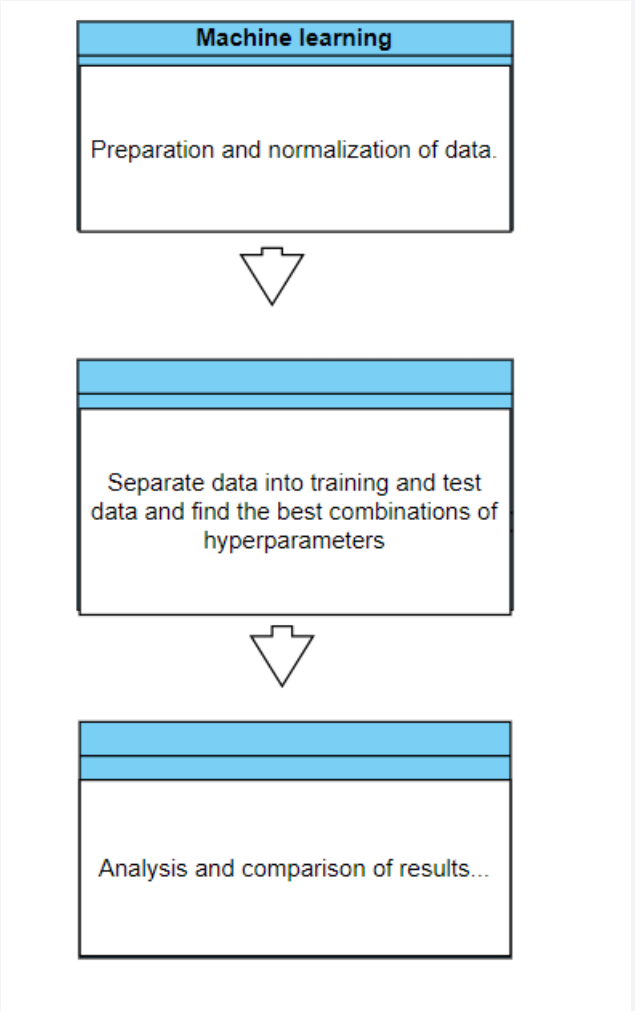
- Folium Maps utilized various elements such as markers, circles, lines, and marker clusters to enhance visualization:
- Markers were employed to pinpoint specific locations such as launch sites on the map.
- Circles were utilized to highlight designated areas around particular coordinates, for instance, the NASA Johnson Space Center.
- Marker clusters were employed to group related events within each coordinate, such as launches at a particular launch site.
- Lines were drawn on the map to represent distances between pairs of coordinates, facilitating spatial understanding.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/f6a8262c6d6fe888d621452488605c62afe1969e/lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

# Build a Dashboard with Plotly Dash

- A variety of graphs and plots were utilized to visually represent the data, including the percentage distribution of launches by site and the payload range. This amalgamation facilitated a swift examination of the correlation between payloads and launch sites, aiding in the identification of optimal launch locations based on payload considerations.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/bf006bc2643bf3eb7afc15cc64bec758363b2b2c/dash_spaceX.py

# Predictive Analysis (Classification)

- Four classification algorithms underwent comparison: logistic regression, support vector machine, decision tree, and k-nearest neighbors.
- Source code: https://github.com/josedyanez/capstone_data.science/blob/bf006bc2643bf3eb7afc15cc64bec758363b2b2c/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



# Results

- SpaceX operates from 4 distinct launch sites.
- Initial launches were conducted for SpaceX itself and NASA.
- The average payload for the F9 v1.1 booster stands at 2,928 kg.
- The first successful landing occurred in 2015, five years subsequent to the inaugural launch.
- Several versions of the Falcon 9 booster achieved successful landings on drone ships, with payloads surpassing the average.
- Nearly all mission outcomes attained success.
- In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landing attempts on drone ships.
- Over time, the rate of successful landing outcomes exhibited improvement.

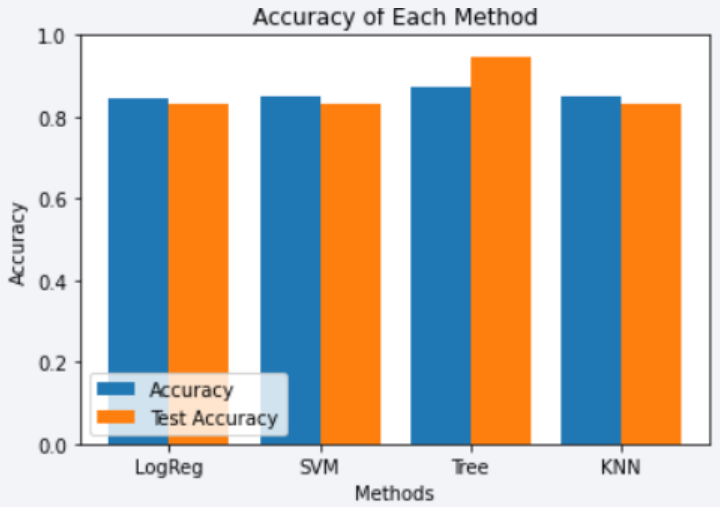
# Results

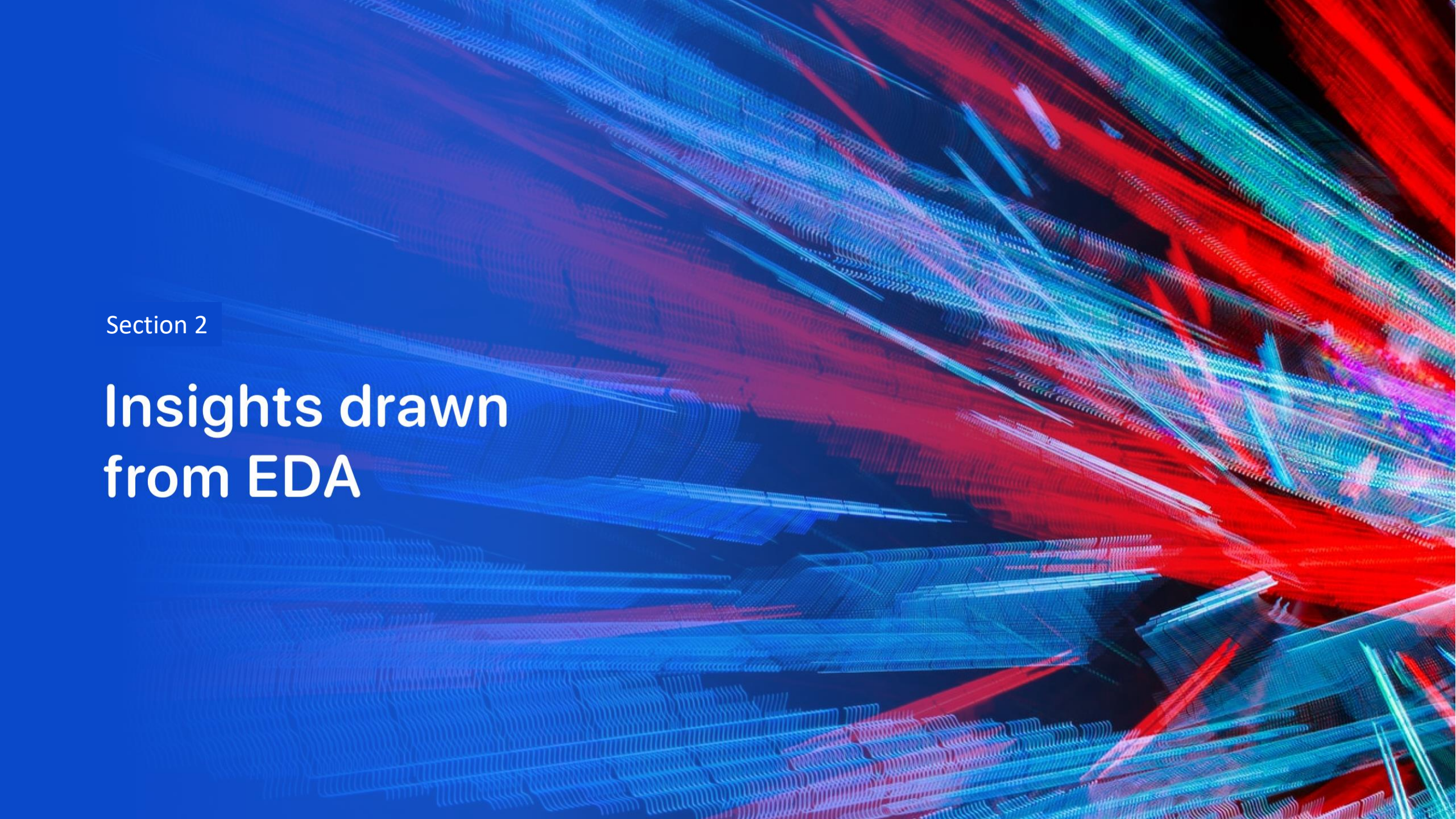
- SpaceX operates from 4 distinct launch sites.

- Through interactive analytics, it was feasible to discern that launch sites are typically situated in secure locations, often in proximity to the sea, and boasting robust logistical infrastructure. Additionally, the majority of launches occur at launch sites located along the east coast.


# Results

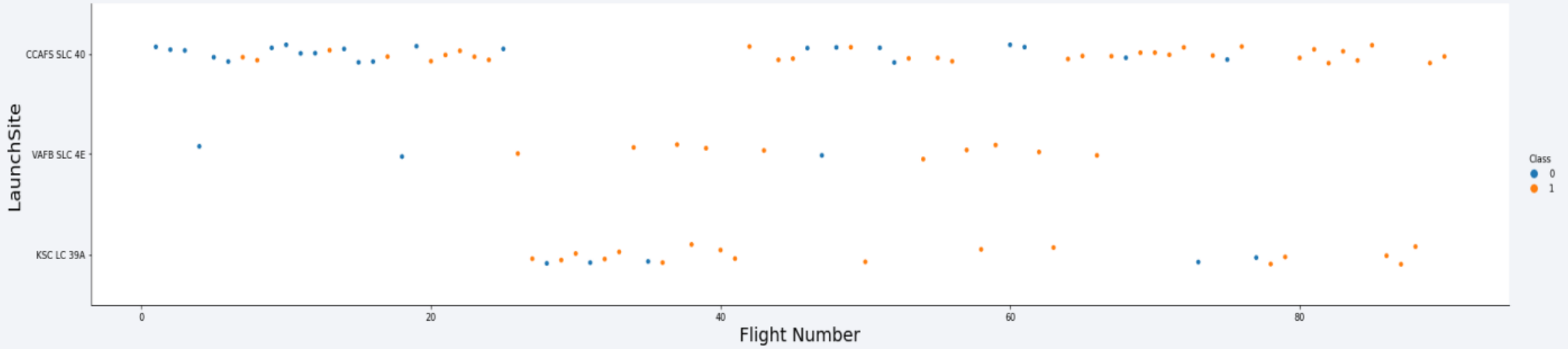
- SpaceX operates from 4 distinct launch sites.
- According to predictive analysis, the Decision Tree Classifier emerged as the optimal model for forecasting successful landings, boasting an accuracy rate exceeding 87%, with a test data accuracy surpassing 94%.




Section 2
Insights drawn from EDA

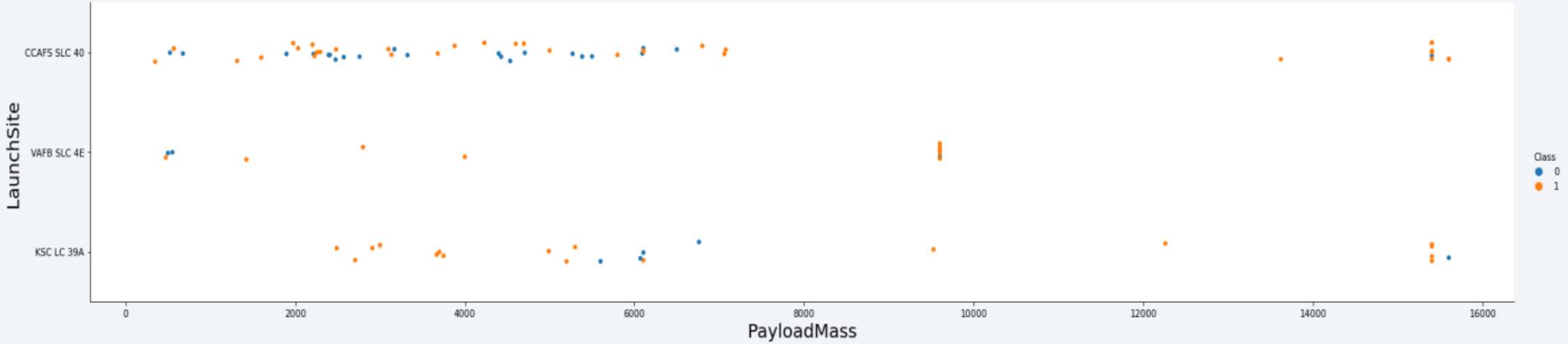
# Flight Number vs. Launch Site

- As indicated by the aforementioned plot, it is evident that the current optimal launch site is CCAF5 SLC 40, with the majority of recent launches culminating in success. Following closely in second place is VAFB SLC 4E, succeeded by KSC LC 39A in third place. Moreover, a discernible trend reveals an overall enhancement in the success rate over time.LC 39A in third place. Moreover, a discernible trend reveals an overall enhancement in the success rate over time.



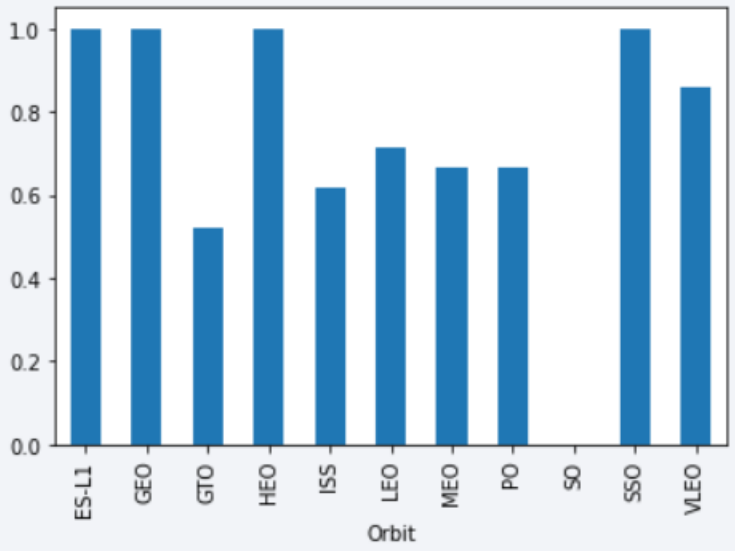
# Payload vs. Launch Site

- Payloads exceeding 9,000kg, roughly equivalent to the weight of a school bus, exhibit a notably high success rate. Additionally, payloads surpassing 12,000kg appear to be feasible exclusively at the CCAFS SLC 40 and KSC LC 39A launch sites.



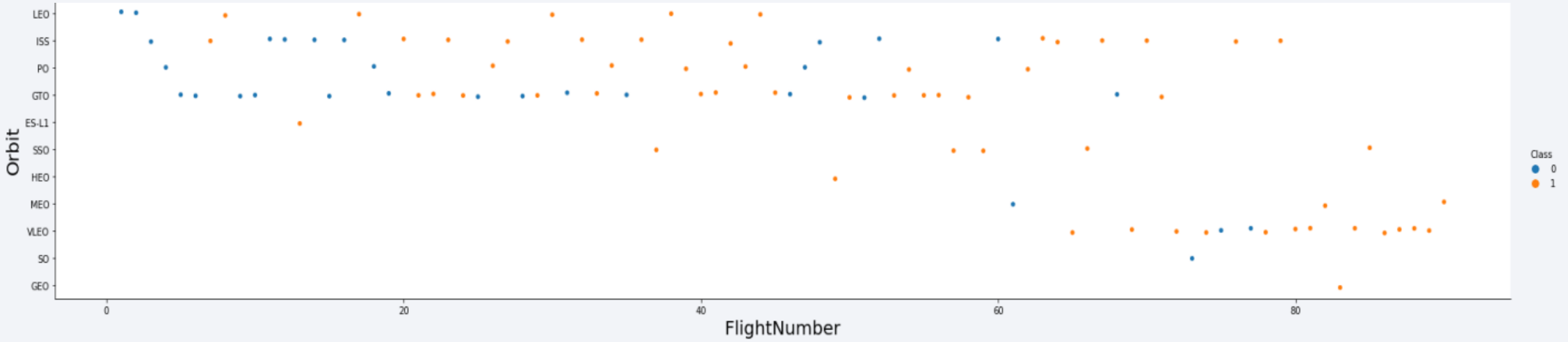
# Success Rate vs. Orbit Type

- The highest success rates occur in orbits such as ES-L1, GEO, HEO, and SSO, closely followed by VLEO, with a success rate above 80%, and LFO, with a success rate above 70%.



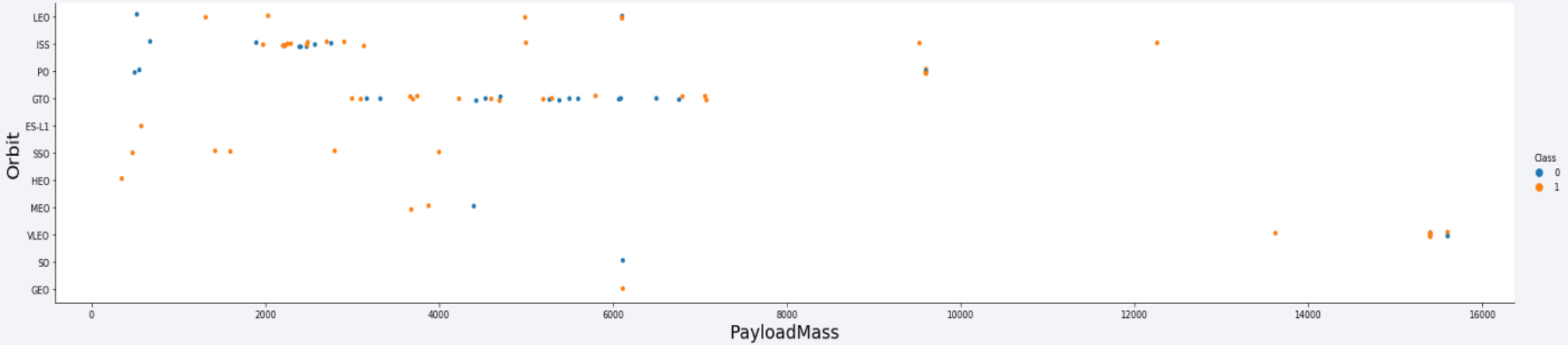
# Flight Number vs. Orbit Type

- Evidently, there has been an improvement in the success rate across all orbits as time has progressed. Additionally, the recent surge in the frequency of launches to the VLEO orbit suggests a burgeoning business opportunity in this sector.



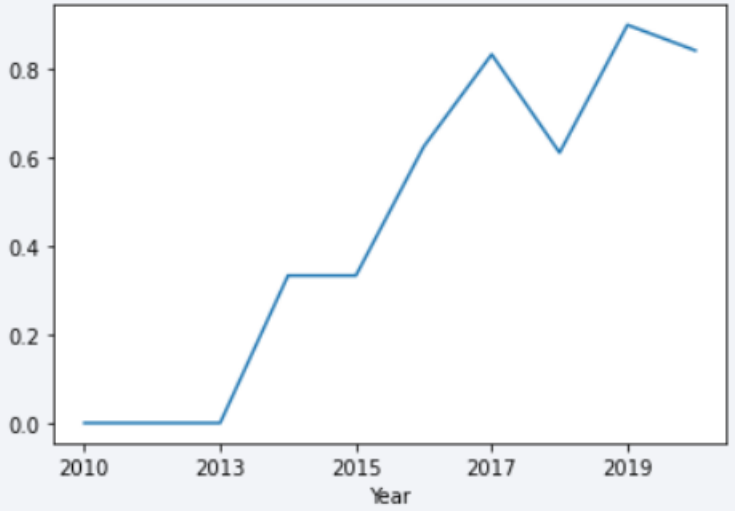
# Payload vs. Orbit Type

- It seems that there is no discernible correlation between payload size and the success rate for the GTO orbit. The ISS orbit boasts the broadest range of payload capacities, coupled with a commendable success rate. Conversely, there are relatively few launches to the SO and GEO orbits.



# Launch Success Yearly Trend

- The success rate began its ascent in 2013 and continued to rise steadily until 2020. It appears that the initial three years marked a phase of adjustments and technological advancements, leading to subsequent improvements.



# All Launch Site Names

- Launch sites obtained from the data.

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

# Launch Site Names Begin with 'CCA'

| 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 | **CCA**FS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | **CCA**FS 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 | **CCA**FS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | F9 v1.0 B0006 | **CCA**FS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | **CCA**FS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attemp |

# Total Payload Mass

- The total payload calculated earlier was derived by aggregating the payloads associated with codes containing 'CRS', representing payloads transported by NASA.

```
[ ]  %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

| total_payload_mass |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

- By filtering the data based on the booster version mentioned and computing the average payload mass, we determined a value of 2,534.66 kg.

```
[ ]  %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';

     * sqlite:///my_data1.db
    Done.
     average_payload_mass
     2534.6666666666665
```

# First Successful Ground Landing Date

- Filtering the data to isolate instances of successful landings on ground pads and extracting the earliest date allows us to pinpoint the first occurrence, which took place on December 22, 2015.

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
first_successful_landing
2015-12-22
```

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters that achieved successful landings on drone ships and had a payload mass exceeding 4000 but less than 6000.

```
 * sqlite:///my_data1.db
Done.
 Booster_Version
 F9 FT B1022
 F9 FT B1026
 F9 FT B1021.2
 F9 FT B1031.2
```

# Total Number of Successful and Failure Mission Outcomes

- The process of grouping mission outcomes and tallying the records within each group resulted in the summary provided below.

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

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

# Boosters Carried Maximum Payload

- These are the boosters that have transported the highest payload masses documented in the dataset.

```
 * 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
```

# 2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

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

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

- This perspective on the data highlights the importance of considering instances labeled as "No attempt.“

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

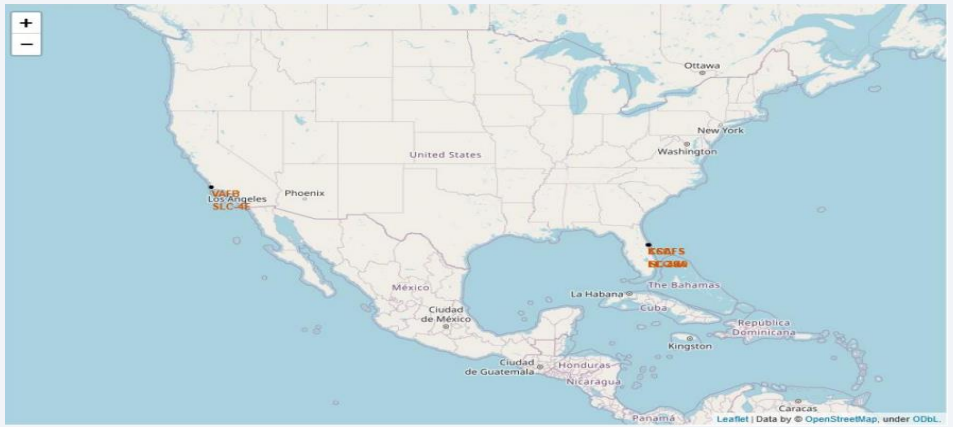
| Landing_Outcome | count_outcomes |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

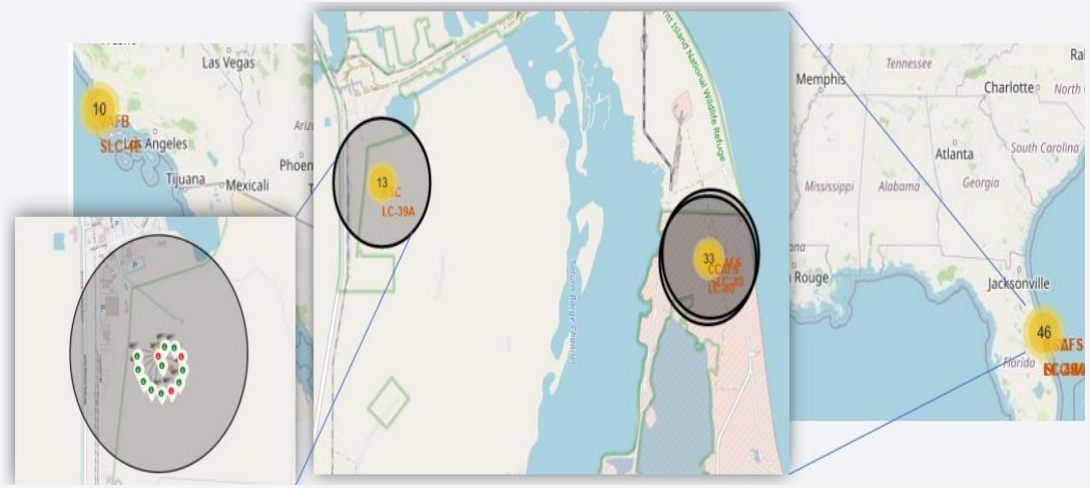
# All launch sites

- Launch sites are situated in close proximity to the sea, likely for safety reasons, yet remain conveniently accessible to roads and railroads.



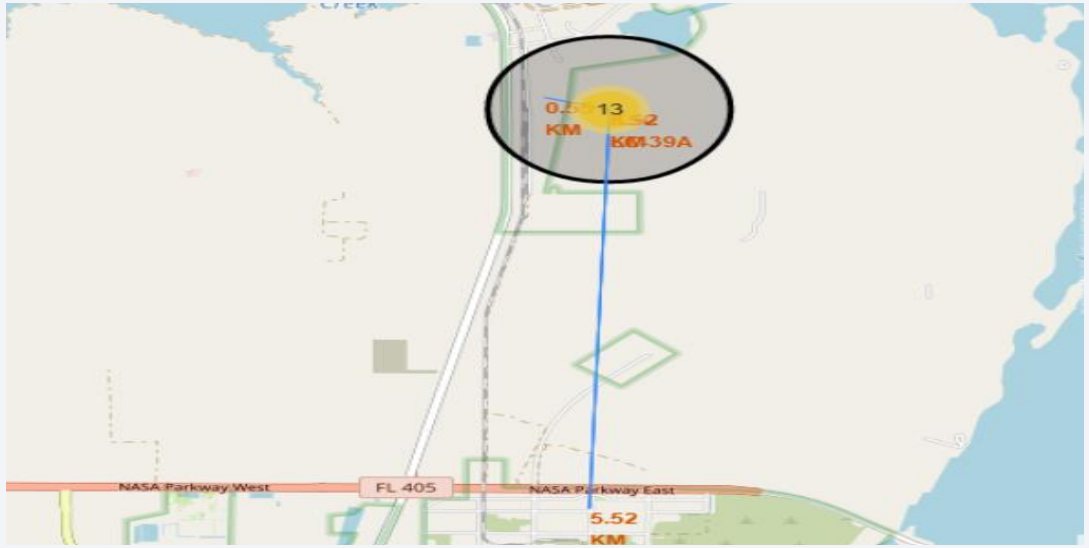
# Launch Outcomes by site

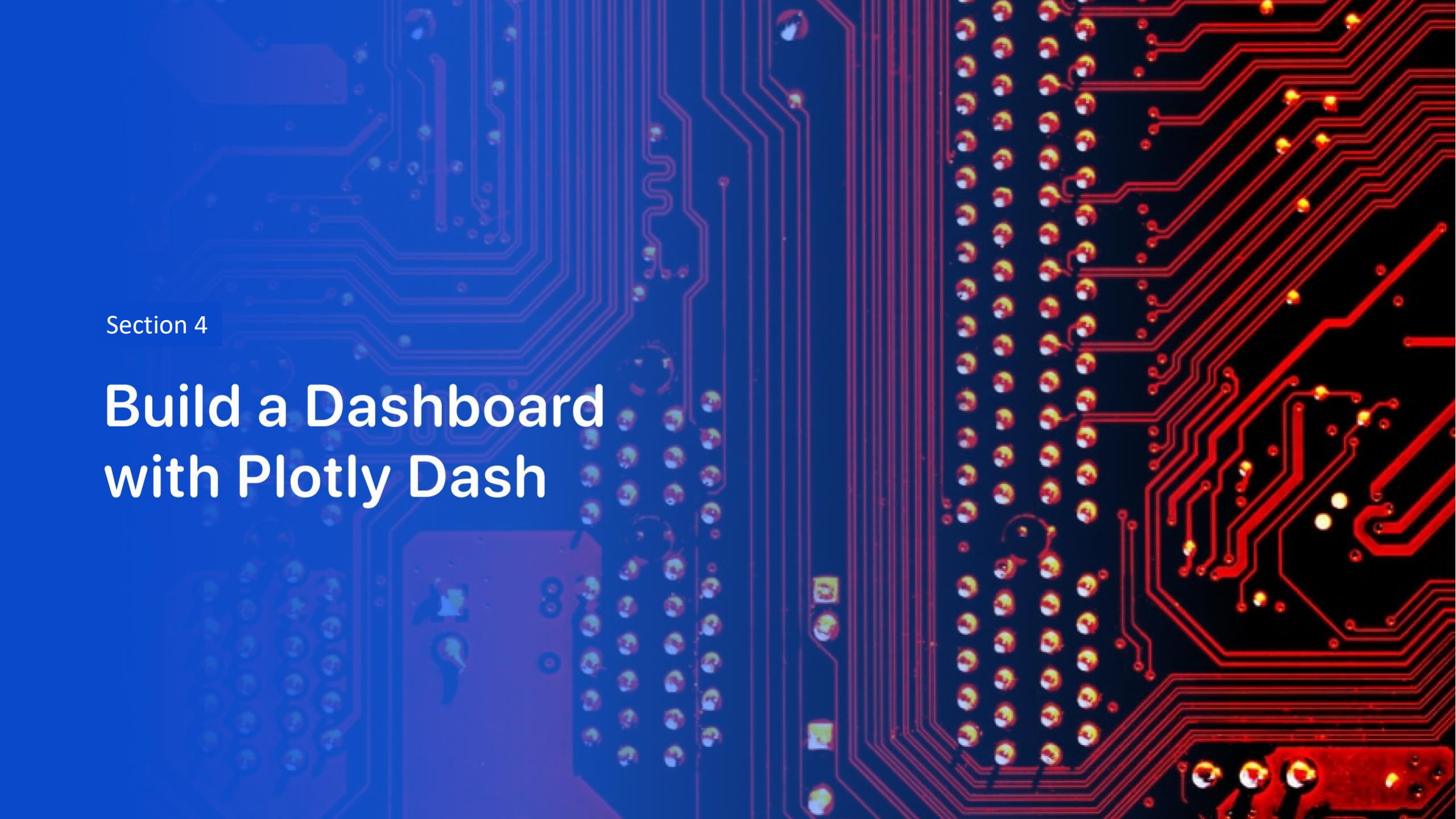
TheGreen markers denote successful outcomes, while red ones signify failures.



# Logistics and Safety

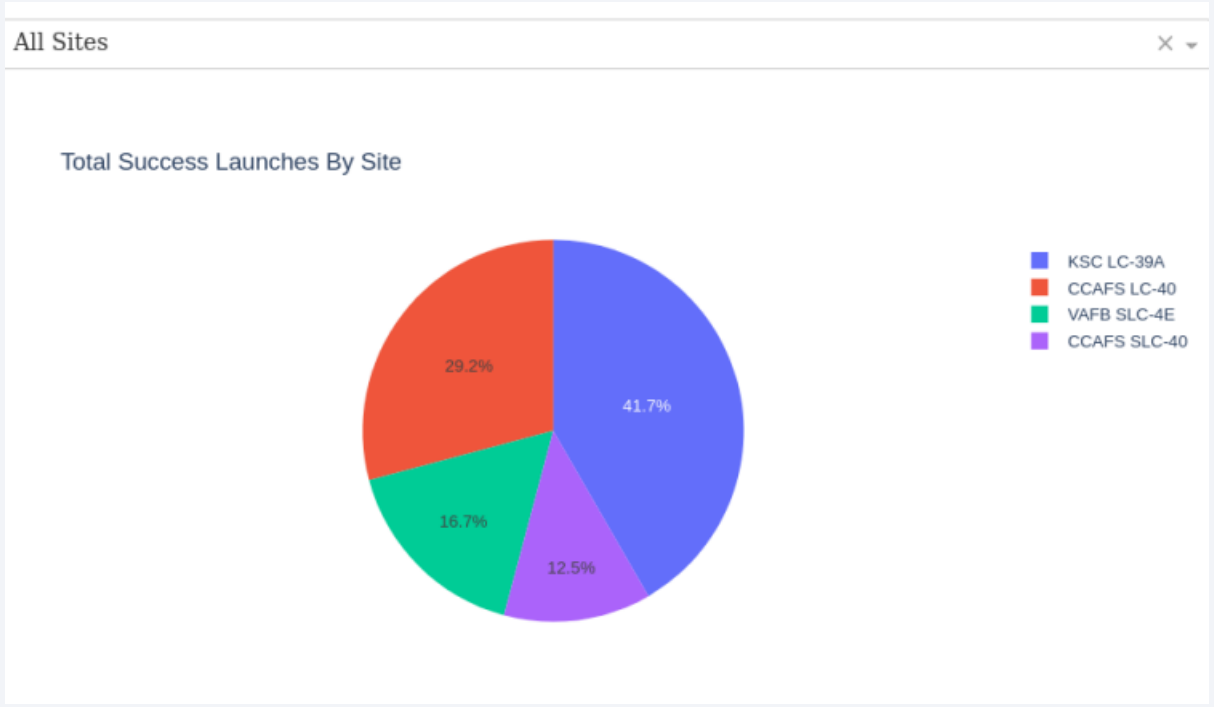
The KSC LC-39A launch site boasts favorable logistical features, with its proximity to both railroads and roads, while also being situated at a considerable distance from populated areas.




Section 4
Build a Dashboard with Plotly Dash

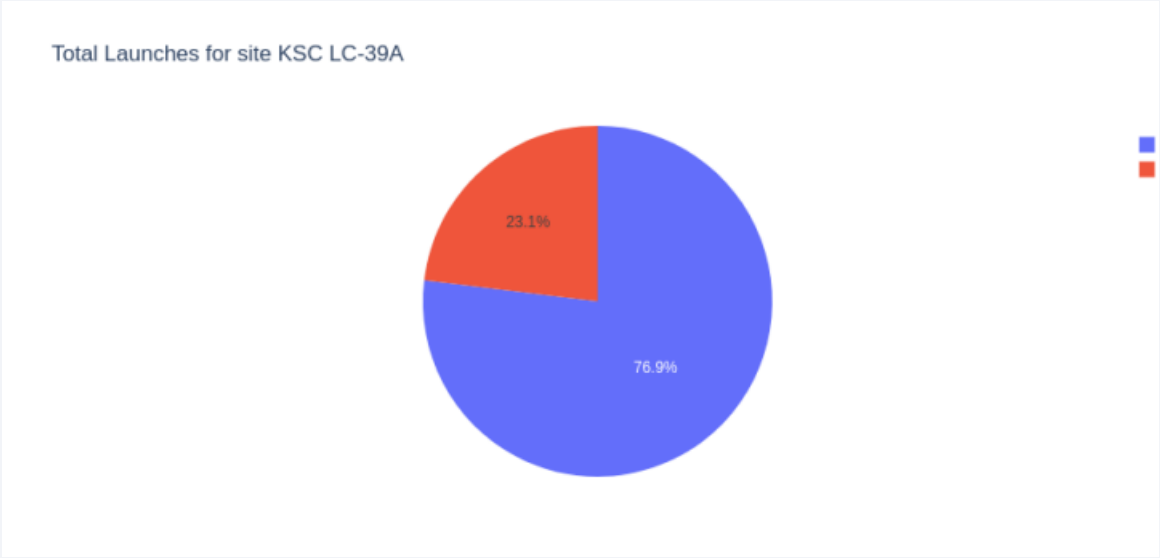
# Successful Launches per Site

- The location where launches take place appears to be a crucial determinant of mission success.



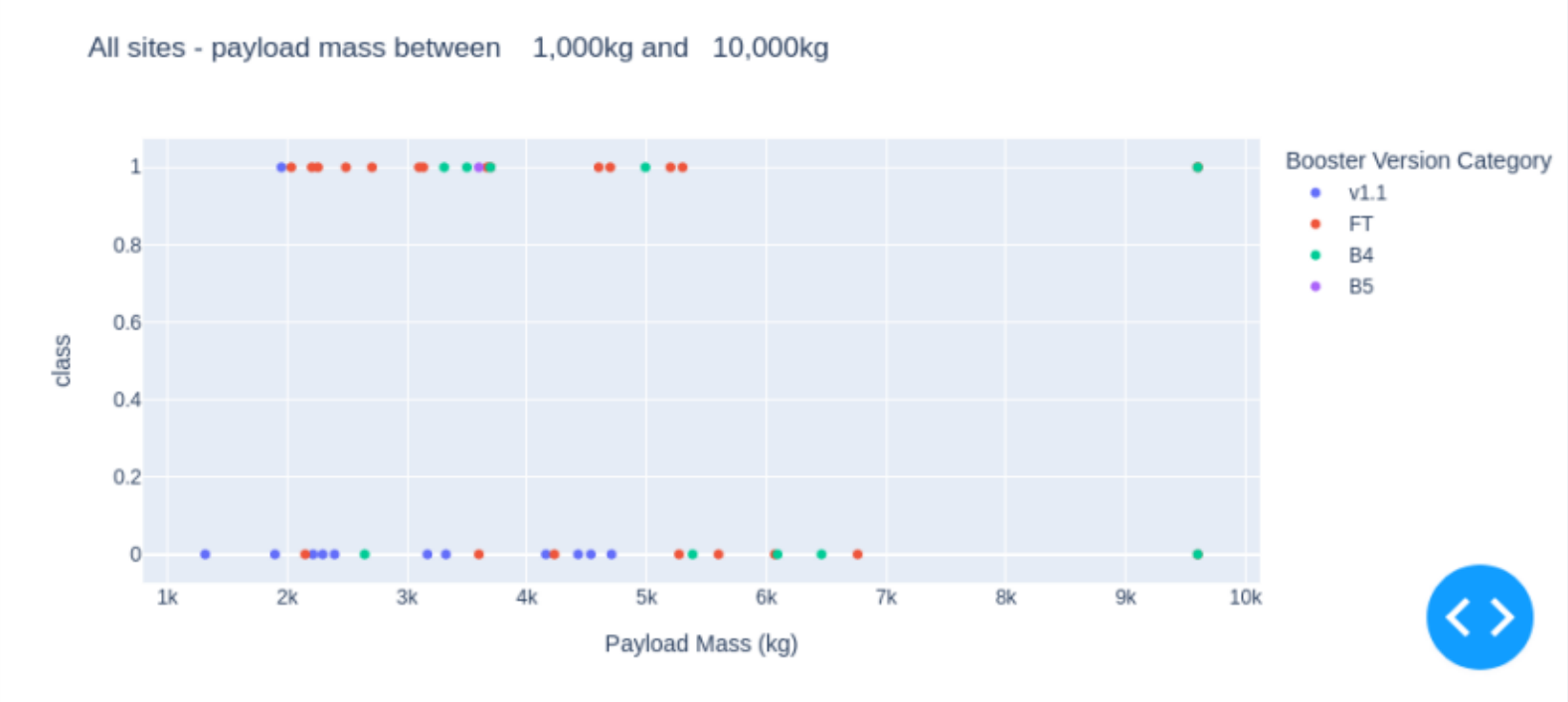
# Launch Success Ratio for KSC LC-39A

- We can see that 76.9% of launches are successful on this site.



# Payload vs Launch Outcome

The combination of payloads weighing less than 6,000kg paired with FT boosters yields the highest success rates.

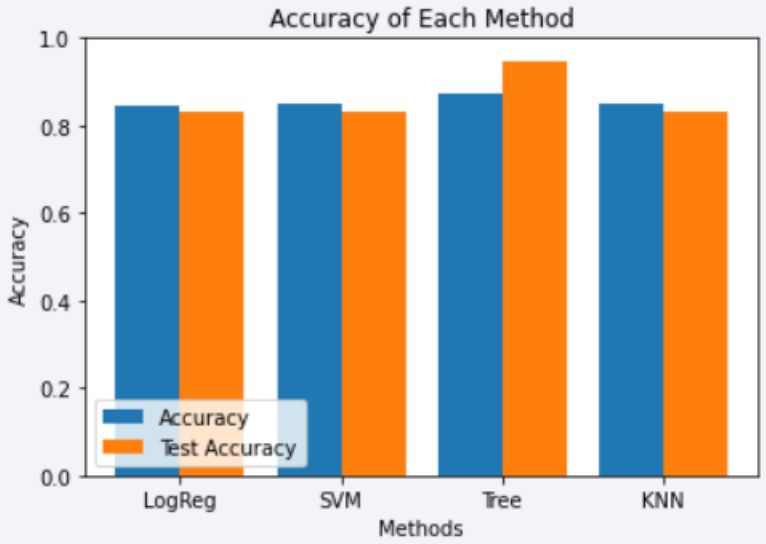


Section 5

# Predictive Analysis (Classification)

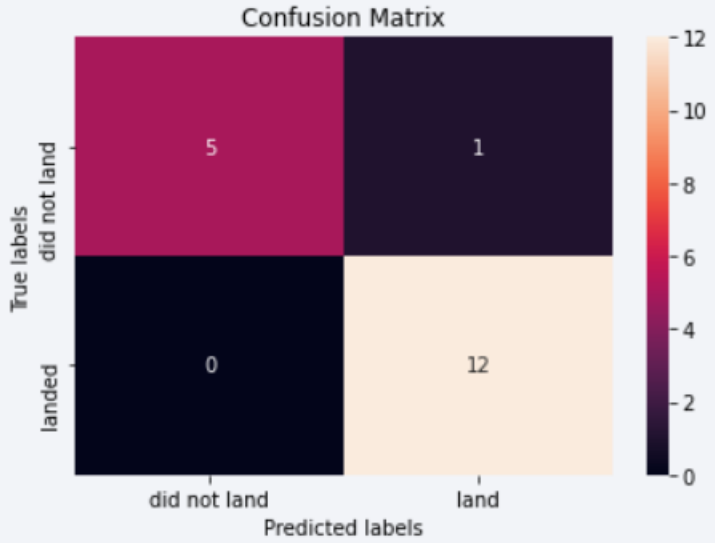
# Classification Accuracy

- Four classification models underwent evaluation, and their respective accuracies are depicted in the accompanying plot.
- The Decision Tree Classifier emerged as the model with the highest classification accuracy, surpassing 87%.



# Confusion Matrix

The accuracy of the Decision Tree Classifier is substantiated by the confusion matrix, which reveals a notable prevalence of true positive and true negative values in comparison to the false ones.



# Conclusions

- The Decision Tree Classifier emerged as a valuable tool for predicting successful landings, thus holding potential to enhance profitability.
- Despite the predominantly successful mission outcomes, there was observed improvement in successful landing outcomes over time, attributed to the evolution of processes and rocket technology.
- The analysis identified KSC LC-39A as the premier launch site.
- Launches with payloads exceeding 7,000kg were determined to carry lower levels of risk.
- Different data sources underwent analysis, contributing to the iterative refinement of conclusions throughout the process.

# Thank you!