IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

IBM

Applied Data Science Specialization

Chamika Dilshan Hapuarachchi
2025/03/18

# Outline

# Executive Summary



- This project aims to predict the successful landing of the SpaceX Falcon 9 first stage using various machine learning classification algorithms. The project follows these key steps:

1. Data collection, cleaning, and preparation
2. Exploratory data analysis
3. Interactive data visualization
4. Machine learning prediction

The analysis reveals that certain features of the rocket launches show a correlation with the landing outcome, whether successful or not. Based on the findings, it appears that the decision tree algorithm is the most effective for predicting the success of the Falcon 9 first stage landing.

# Introduction



- This capstone project focuses on predicting whether the Falcon 9 first stage will land successfully. SpaceX markets Falcon 9 rocket launches on its website at a cost of 62 million dollars, significantly lower than other providers, whose costs start at over 165 million dollars. The primary savings come from SpaceX's ability to reuse the first stage of the rocket. Therefore, accurately predicting whether the first stage will land successfully can provide insights into the cost of a launch. This information could be valuable to competing companies looking to bid against SpaceX for rocket launches.
- Most failed landings are intentional, as SpaceX sometimes conducts controlled landings in the ocean. The central question in this project is: given a set of features such as payload mass, orbit type, launch site, and others, can we predict whether the Falcon 9 first stage will successfully land?

Section 1

# Methodology

# Methodology



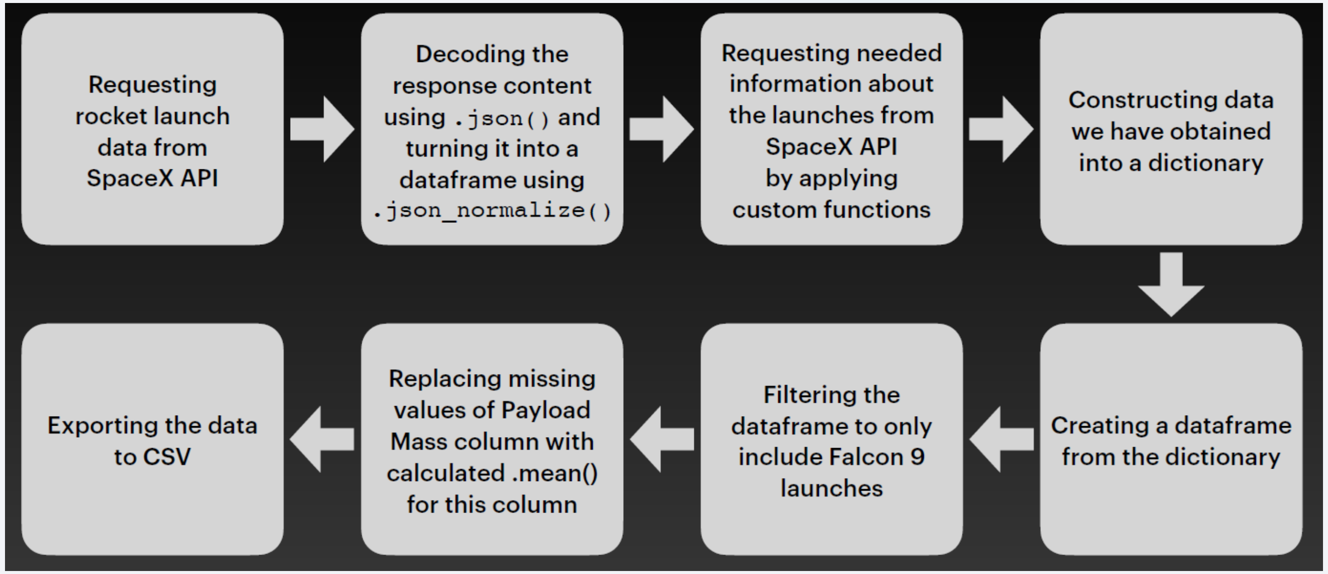
The overall methodology includes:

- **Data collection, wrangling, and formatting, using:** SpaceX API & Web scraping Wikipedia pages
- **Exploratory data analysis (EDA), using:** Pandas and NumPy SQL
- **Data visualization, using**: Matplotlib and Seaborn Folium Dash
- **Machine learning models used** : Logistic regression, Support vector machine (SVM), Decision tree, K-nearest neighbors (KNN)

# Data Collection – SpaceX API

1. SpaceX API:

- API used: https://api.spacexdata.com/v4/rockets/
- Provides data on various SpaceX rocket launches.
- respective column.
- The final dataset consists of:90 rows (instances) 17 columns (features)

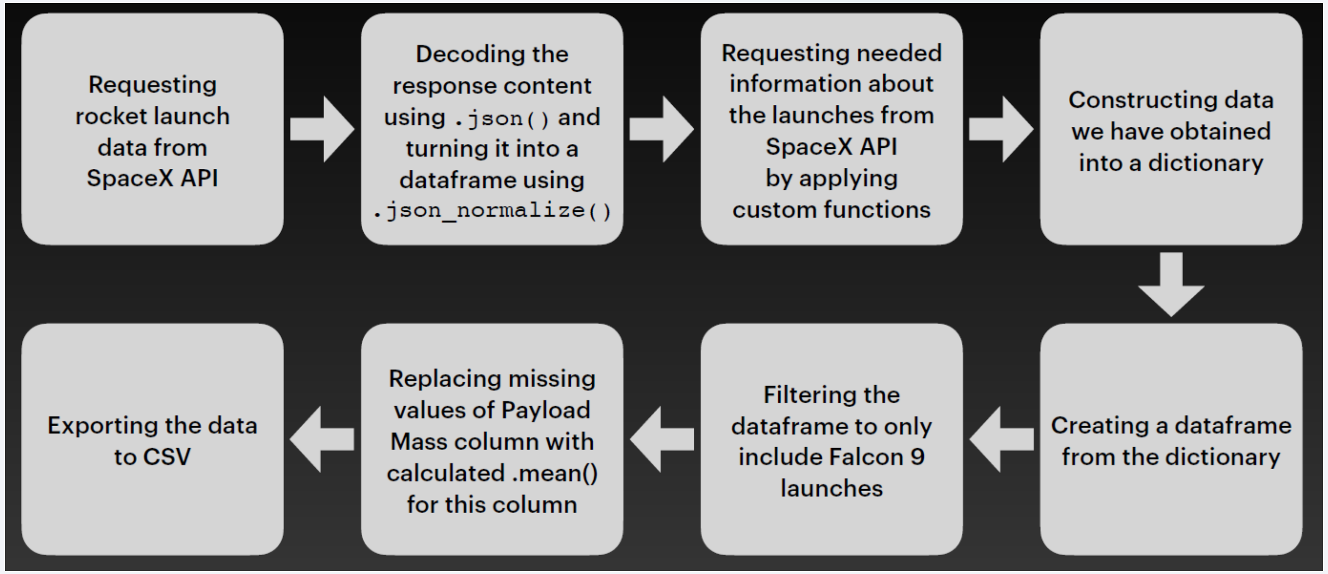



GitHub URL: Data Collection API

# Data Collection – SpaceX API

2. Web scraping:

- API used: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The final dataset consists of:121 rows (instances)11 columns (features)




GitHub URL: Web Scrapping

# Data Wrangling



- The data is processed to eliminate any missing entries, and categorical features are encoded using one-hot encoding. A new column, 'Class', is added to the dataset, where a value of 0 indicates a failed launch and 1 indicates a successful one. This results in a dataset with 90 rows (instances) and 83 columns (features). The dataset also includes different cases of unsuccessful booster landings. For instance, "True Ocean" refers to a successful landing in the ocean, while "False Ocean" means the landing was unsuccessful. Similarly, "True RTLS" indicates a successful landing on a ground pad, and "False RTLS" represents a failed landing. "True ASDS" means a successful landing on a drone ship, whereas "False ASDS" means an unsuccessful landing. These outcomes are converted into training labels, where "1" denotes a successful landing and "0" denotes an unsuccessful one.

GitHub URL: Data Wrangling

# EDA with Data Visualization

Various charts were plotted to analyze the data, including:

- 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
- Success Rate Yearly Trend

Scatter plots were used to explore relationships between variables, which could be useful for building machine learning models if patterns exist. Bar charts were utilized to compare discrete categories, highlighting relationships between different categories and measured values. Line charts were employed to analyze trends over time, providing insights into time series patterns.

GitHub URL: EDA

# EDA with SQL

The following SQL queries were executed to analyze the dataset:

- Retrieving unique launch site names
- Fetching 5 records where launch sites start with ‘CCA’
- Calculating the total payload mass carried by NASA (CRS) boosters
- Finding the average payload mass carried by booster version F9 v1.1
- Identifying the date of the first successful landing on a ground pad
- Listing boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000
- Counting the total number of successful and failed mission outcomes
- Finding booster versions that carried the maximum payload mass
- Listing failed drone ship landings in 2015, along with booster versions and launch site names

Ranking landing outcomes (e.g., Failure on a drone ship, Success on a ground pad) between 2010-06-04 and 2017-03-20 in descending order

GitHub URL: EDA with SQL

# Build an Interactive Map with Folium

- **Markers for All Launch Sites:**
  - Placed a circle marker with a popup label and text label at NASA Johnson Space Center using its latitude and longitude as the starting location.
  - Added markers for all launch sites, showing their geographical positions and proximity to the equator and coastlines.
- **Colored Markers for Launch Outcomes:**
  - Used green markers for successful launches and red markers for failed launches.
  - Applied Marker Cluster to visualize success rates at different launch sites.
- **Measuring Distances from a Launch Site:**
  - Drew colored lines to represent distances between Launch Site KSC LC-39A and nearby locations such as railways, highways, coastlines, and the nearest city.

GitHub URL: interactive Map Visualization with Folium

# Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
  - Implemented a dropdown menu to allow users to select a specific launch site.

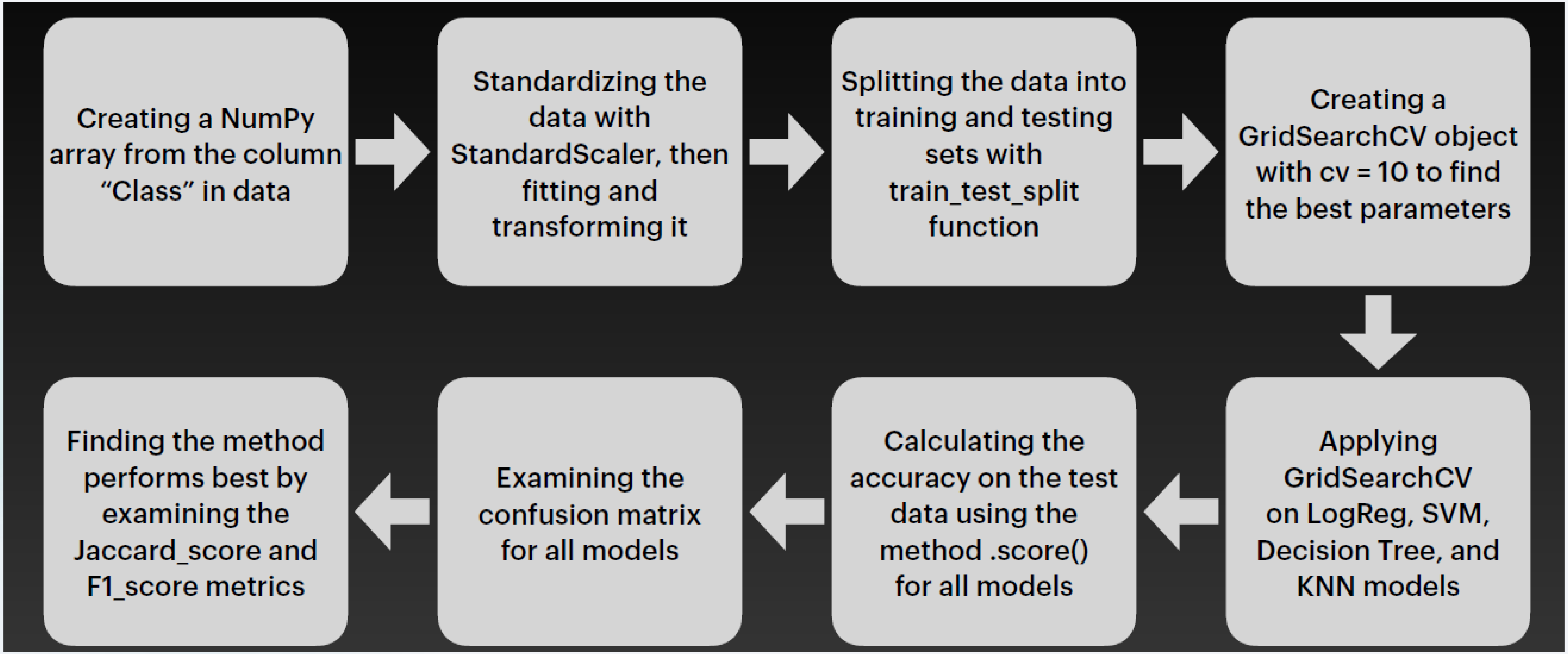
- **Pie Chart for Success Rates:**
  - Displayed a pie chart showing the total successful launches across all sites.
  - If a specific site is selected, the chart updates to show the Success vs. Failure count for that site.

- **Payload Mass Range Slider:**
  - Added a slider to enable users to filter data based on the payload mass range.

- **Scatter Chart: Payload Mass vs. Success Rate:**
  - Plotted a scatter chart to analyze the correlation between payload mass and launch success, categorized by different booster versions.

GitHub URL: Plotly Dashboard

# Predictive Analysis (Classification)




[GitHub URL: Machine Learning Prediction]

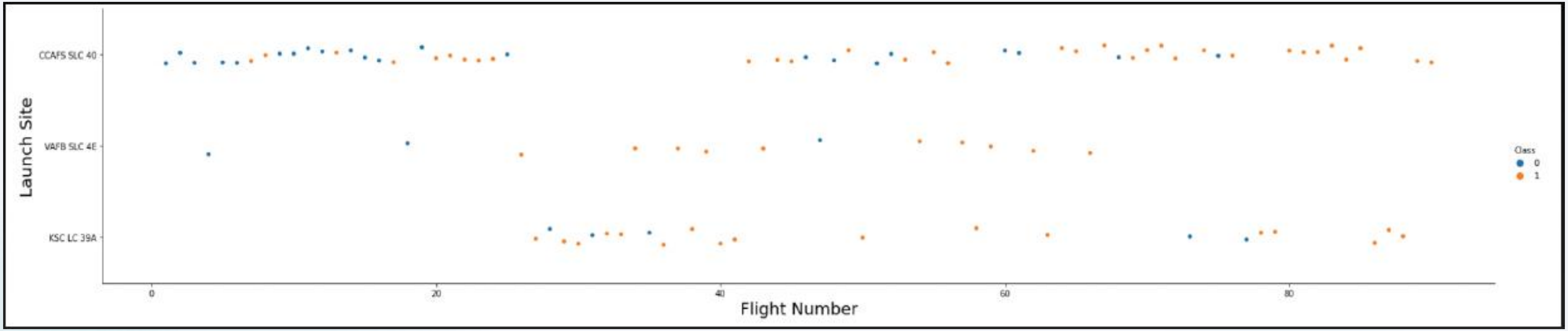
# Results

- The results are divided into five key sections:
  - SQL (EDA with SQL):Performed exploratory data analysis using SQL queries to extract insights from the dataset.
  - Matplotlib and Seaborn (EDA with Visualization):Created various visualizations to identify trends and relationships in the data.
  - Folium:Mapped launch sites and outcomes using interactive geographical visualizations.
  - Dash:Developed an interactive dashboard to explore launch data dynamically.
  - Predictive Analysis:Implemented machine learning models to predict whether a Falcon 9 first-stage landing would be successful.
- In all graphs, Class 0 represents a failed launch, while Class 1 represents a successful launch.

Section 2

# Insights drawn from EDA

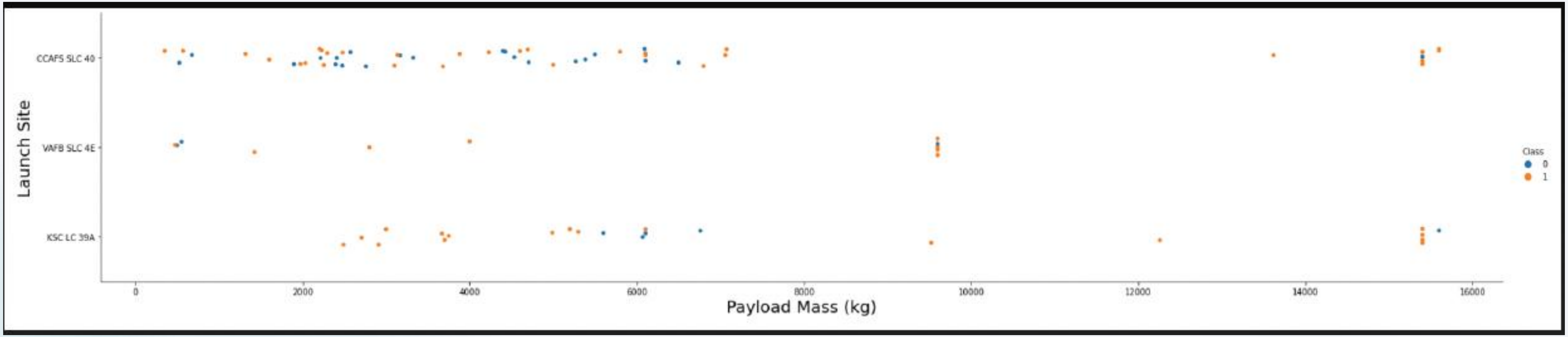
# Flight Number vs. Launch Site




- 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. Launch Site

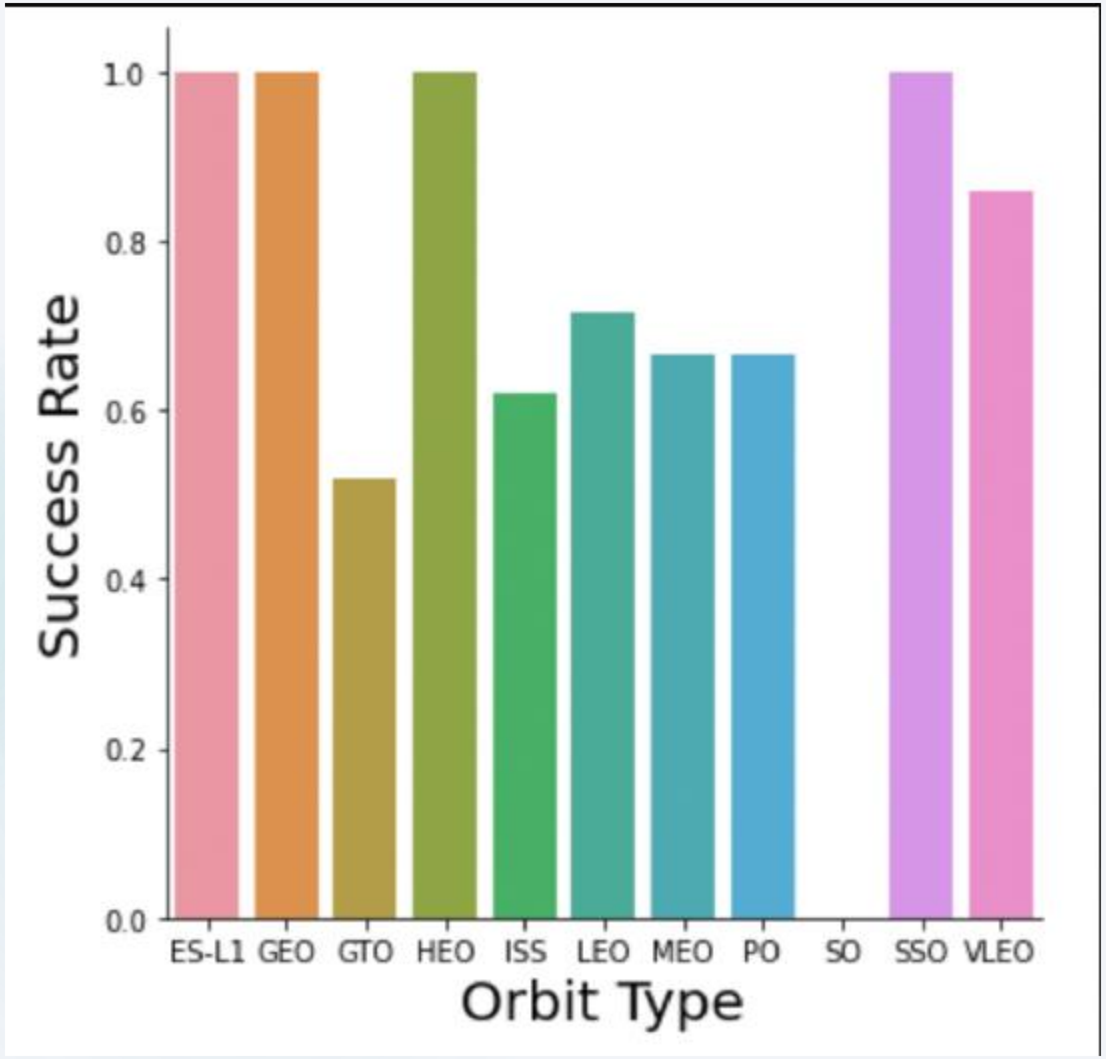





- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

# Success Rate vs. Orbit Type

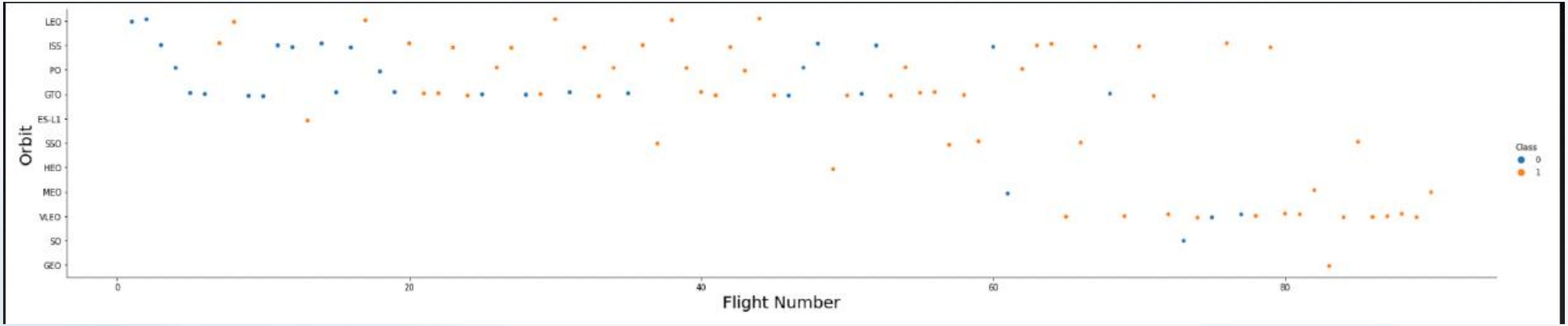





- Explanation:
- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

# Flight Number vs. Orbit Type

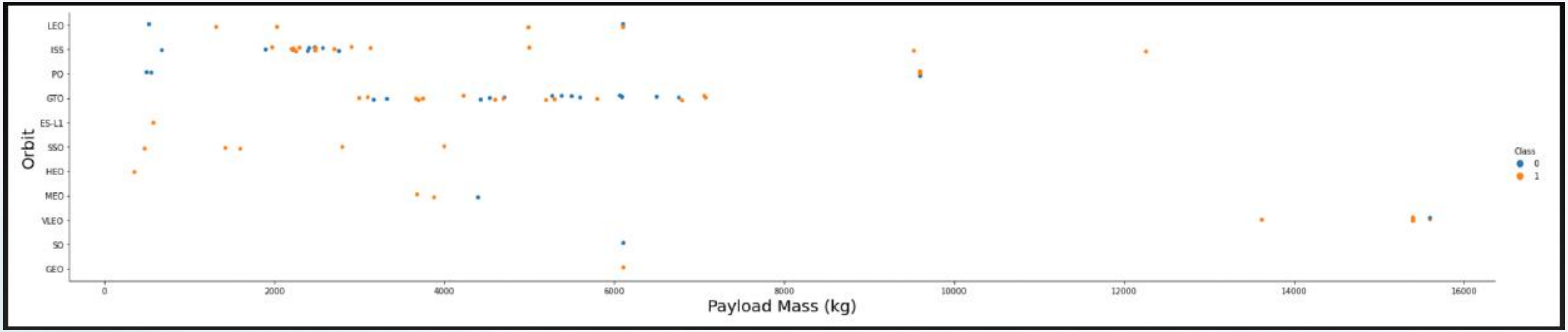





- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

# Payload vs. Orbit Type

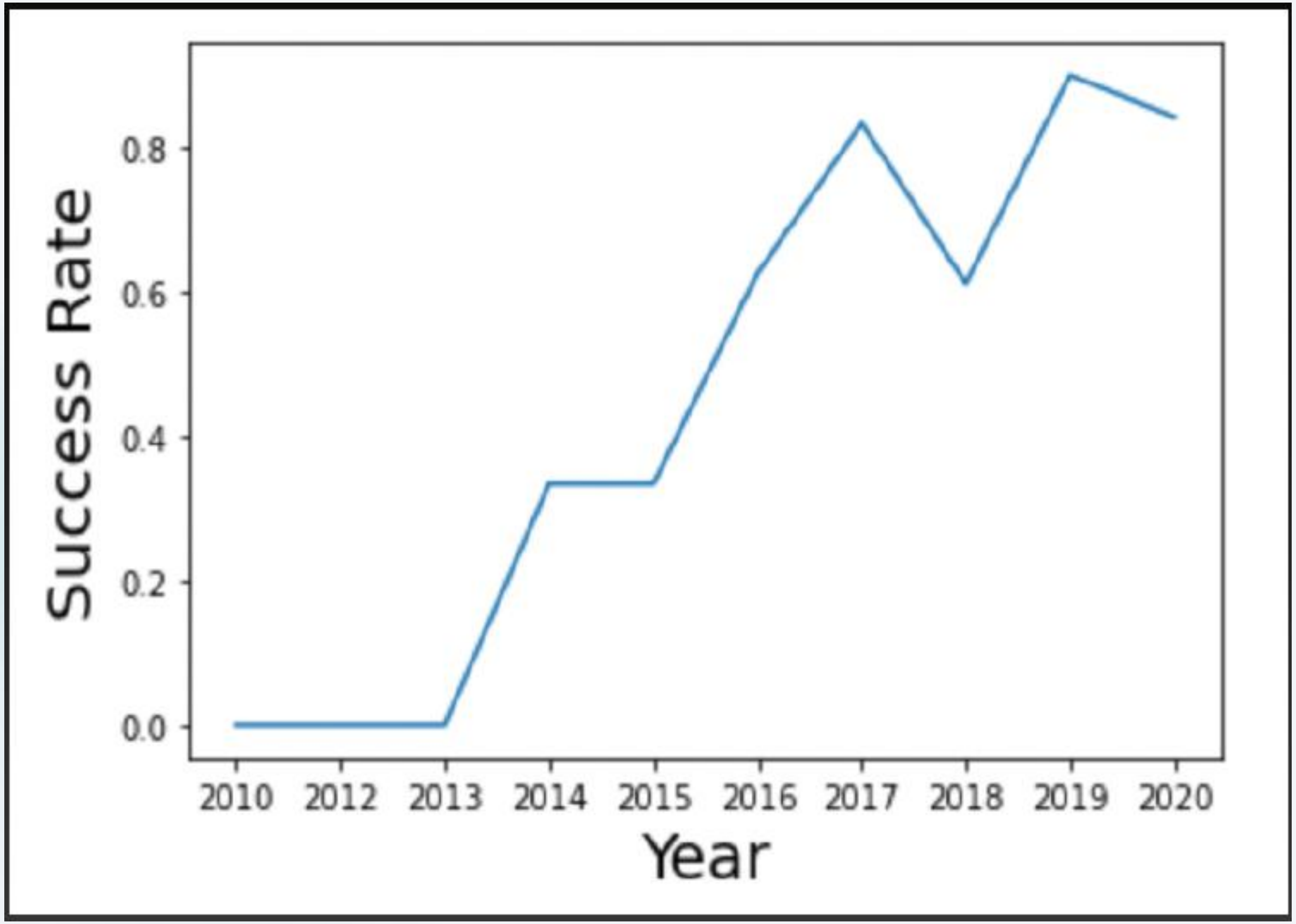





- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

# Launch Success Yearly Trend






- The success rate since 2013 kept increasing till 2020.

# All Launch Site Names



```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

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

# Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

| 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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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 |

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

# Total Payload Mass



In [6]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

| total_payload_mass |
|---|
| 45596 |

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

# Average Payload Mass by F9 v1.1



```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

| average_payload_mass |
|---|
| 2534 |

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

# First Successful Ground Landing Date



```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[8]:
```

| first_successful_landing |
| --- |
| 2015-12-22 |

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

# Successful Drone Ship Landing with Payload between 4000 and 6000



```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

- 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



```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[10]:
```

| mission_outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

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

# Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[11]:
```

| 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 |

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

# 2015 Launch Records



```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[12]:
```

| MONTH | DATE | booster_version | launch_site | landing__outcome |
|---|---|---|---|---|
| January | 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| April | 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

- 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

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
               where date between '2010-06-04' and '2017-03-20'
               group by landing__outcome
               order by count_outcomes desc;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[13]:
```

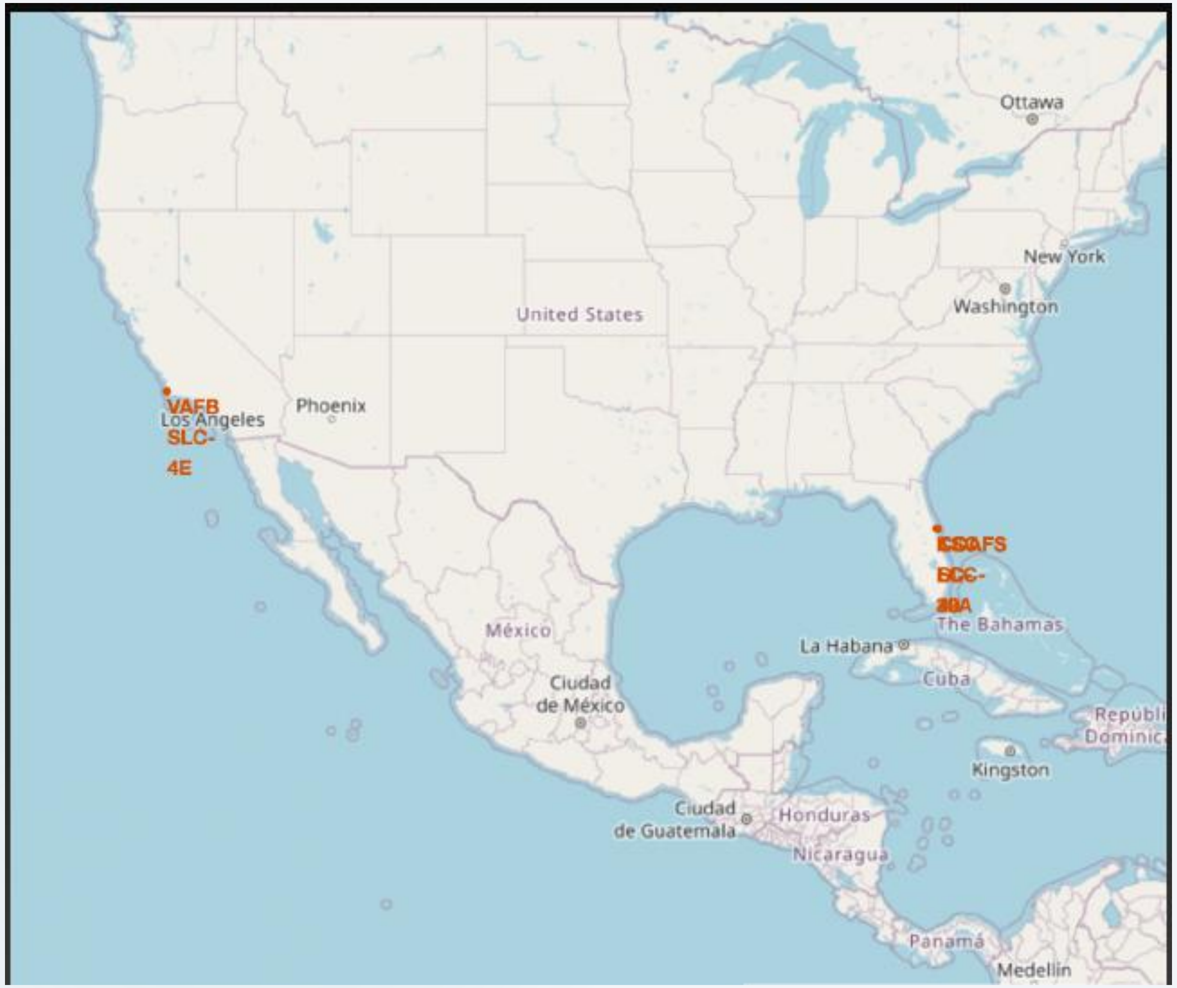
| landing__outcome | count_outcomes |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

- 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

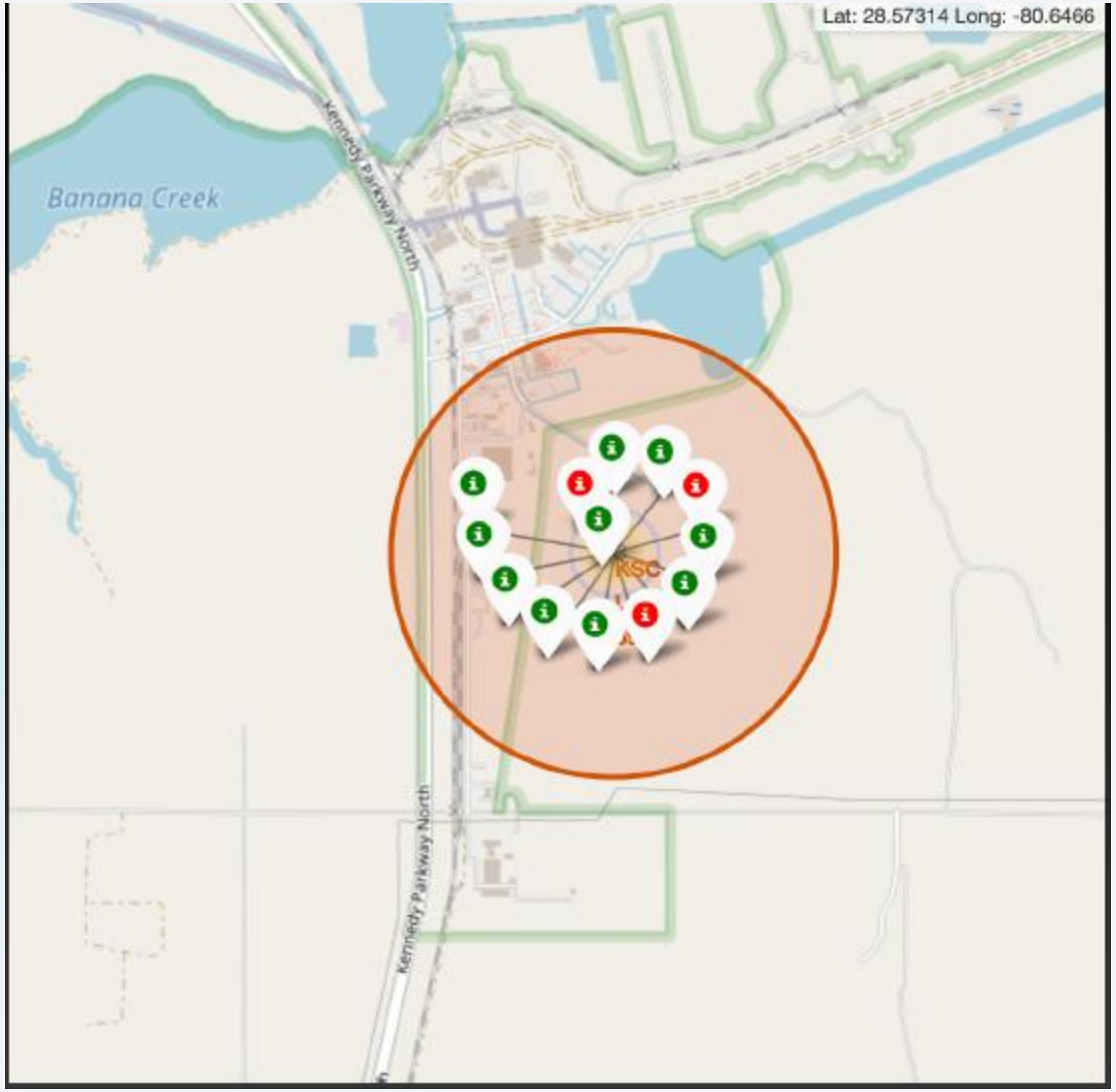
# Launch Sites Proximities Analysis

# All launch sites' location markers on a global map



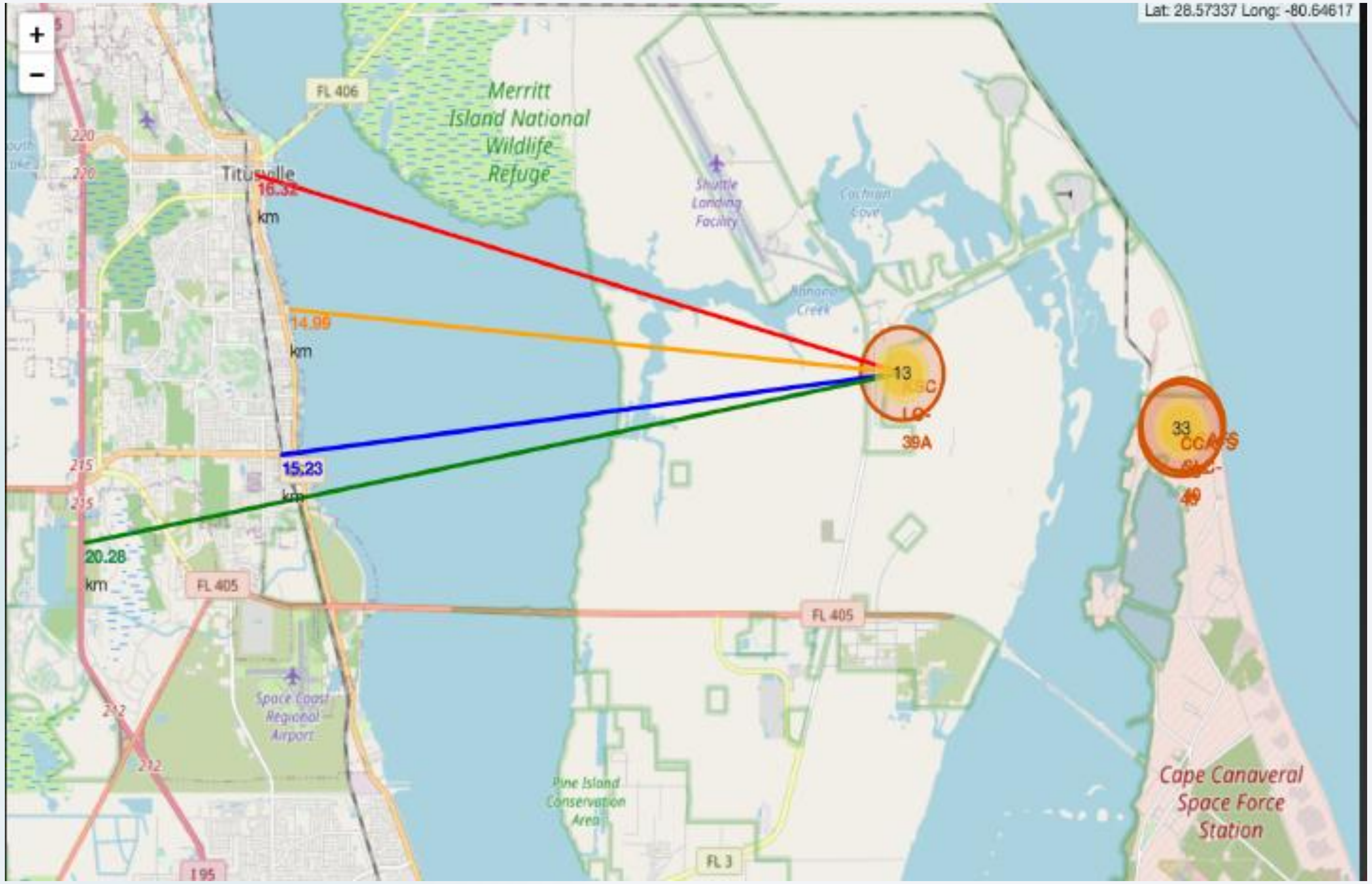

- Proximity to the Equator:
  - The Earth's rotation is fastest at the equator (1670 km/h).
  - Launching from the equator provides an additional speed boost due to inertia, helping spacecraft stay in orbit.
- Proximity to the Coast:
  - Launching over the ocean reduces risks by ensuring debris or failed launches do not endanger populated areas.

# Color-labeled cluster launch records on the map




- Color-coded markers for easy identification:
  - Green = Successful launch
  - Red = Failed launch
- Key Observation:
  - Launch Site KSC LC-39A has a very high success rate.

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

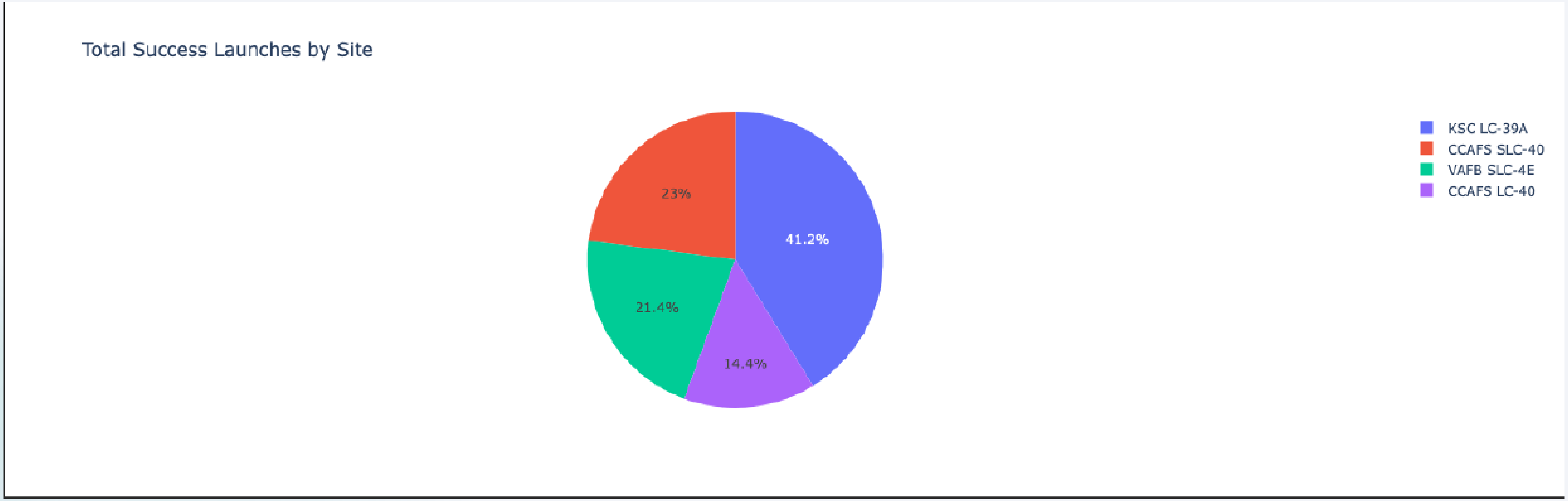



- Close to key infrastructures:
  - 🚆 Railway: 15.23 km
  - 🛣️ Highway: 20.28 km
  - 🌊 Coastline: 14.99 km
  - 🏙️ Closest City (Titusville): 16.32 km
- Safety Concern:
  - A failed rocket traveling at high speeds can cover 15-20 km in seconds, posing risks to populated areas.

Section 4

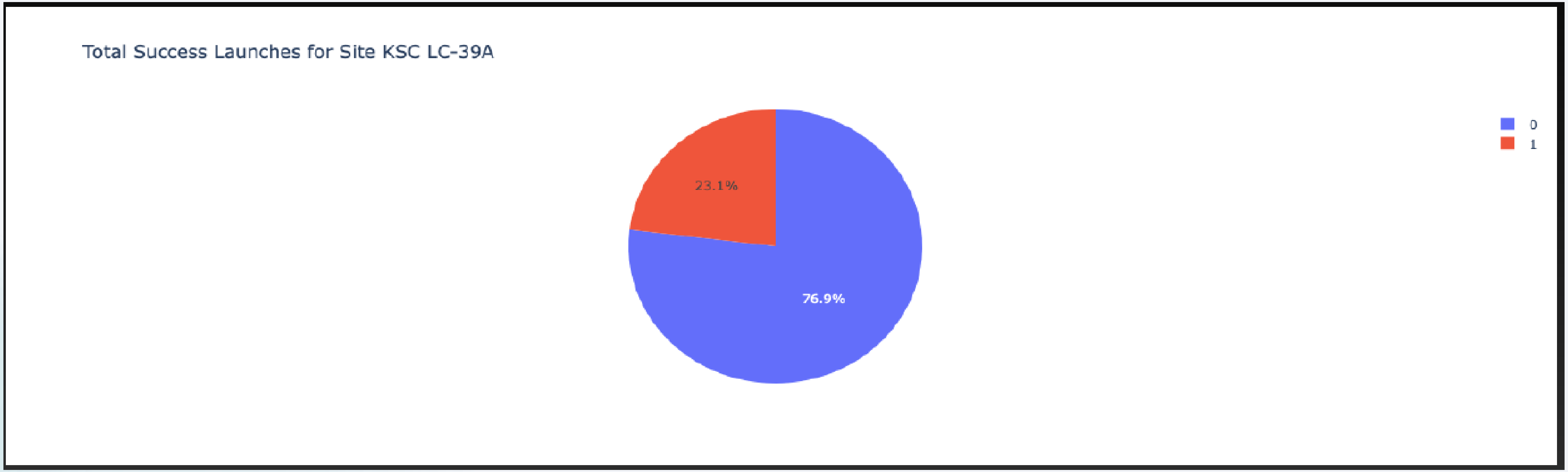
# Build a Dashboard with Plotly Dash

# Launch success percentages for all sites




- KSC LC-39A has the highest number of successful launches among all launch sites.

# Launch site with highest launch success ratio




- 🚀 KSC LC-39A Achieves the Highest Success Rate!
- ✅ 76.9% Success Rate (10 Successful, 3 Failed Landings)

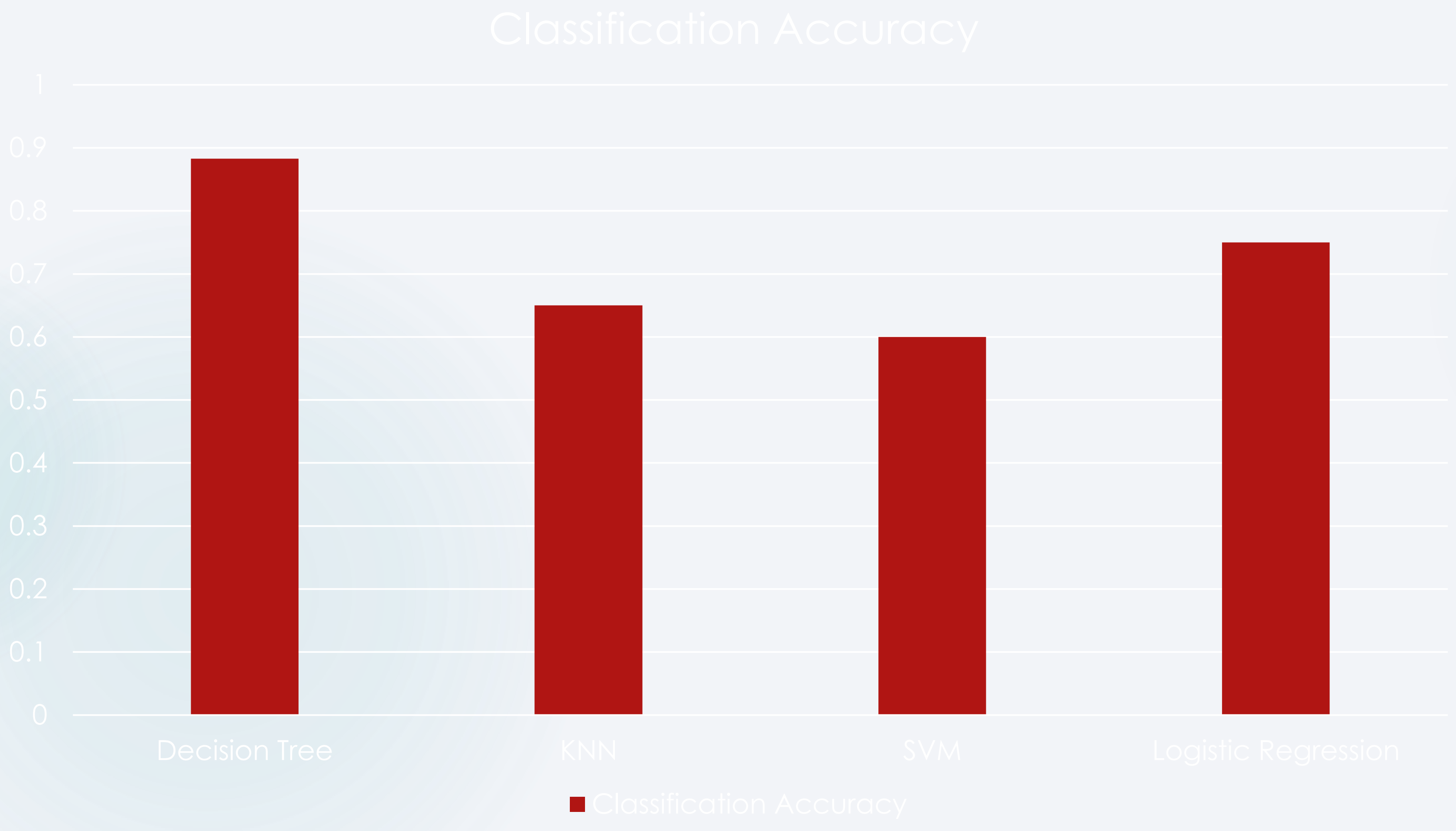
# Payload Mass vs. Launch Outcome for all sites


- 📊 Optimal Payload Range for Success
- ✅ Payloads between 2000 - 5500 kg have the highest success rate.

Section 5

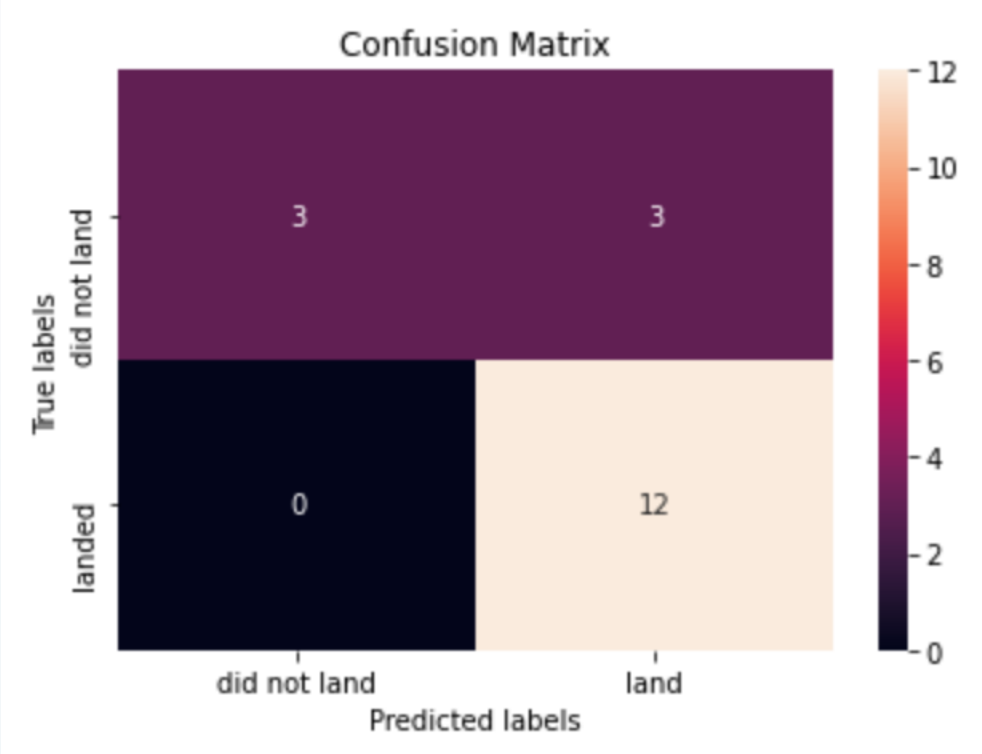
# Predictive Analysis (Classification)

# Classification Accuracy




- Examining the accuracies , we see that Decision Tree Model has the highest classification accuracy.

# Confusion Matrix – Decision Tree Model




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

# Conclusions



- Best Algorithm: The Decision Tree Model proves to be the most effective for predicting launch success.
- Payload Impact: Lower payload masses generally result in higher success rates.
- Proximity Advantage: Most launch sites are near the Equator and coastlines, which enhances safety.
- Improving Success: The overall success rate of launches has increased over the years.
- Top Performing Site: KSC LC-39A stands out with the highest success rate.
- Successful Orbits: Orbits like ES-L1, GEO, HEO, and SSO boast a 100% success rate.

# Thank you!