# IBM – Data Science Capstone Project

Diego de Mattos

May 2023

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

# 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 result

# Introduction

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company has the Falcon 9 rocket that can reuse its first stage if they have a successful launch. Based on public information and machine learning models, we are going to try to predict if the next launch will succeed.

## 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?

# Methodology

Data collection

- Using SpaceX Rest API.
- Using Web Scrapping from Wikipedia.

Performed data wrangling

- Filtering the data.
- Dealing with missing values.
- Using One Hot Encoding to prepare the data to a binary classification.

Performed exploratory data analysis (EDA) using visualization and SQL.

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models.

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

# Data Collection

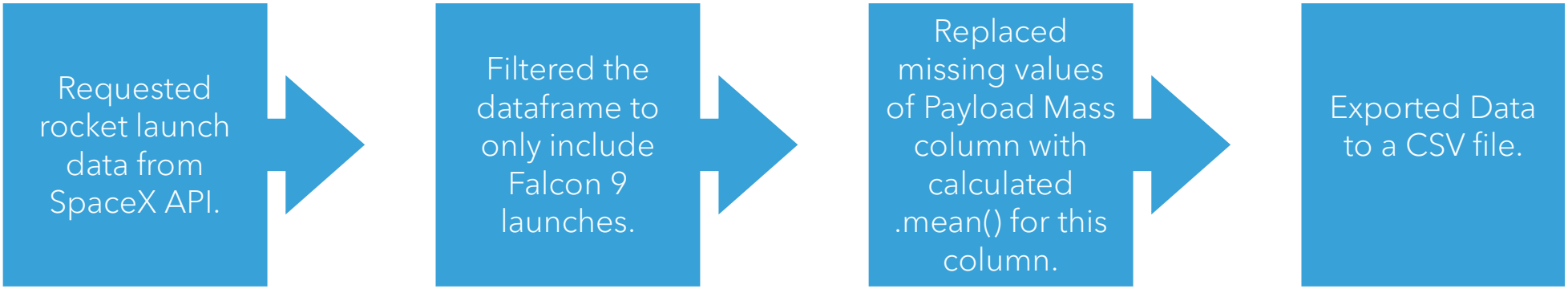
Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both 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.

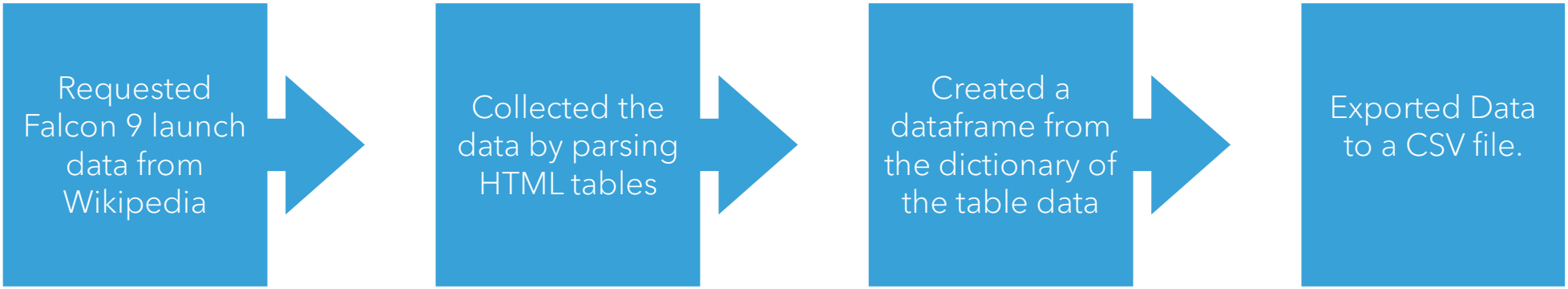
Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

# Data Collection – SpaceX API




Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20Space%20X%20Data%20Collection%20API.ipynb

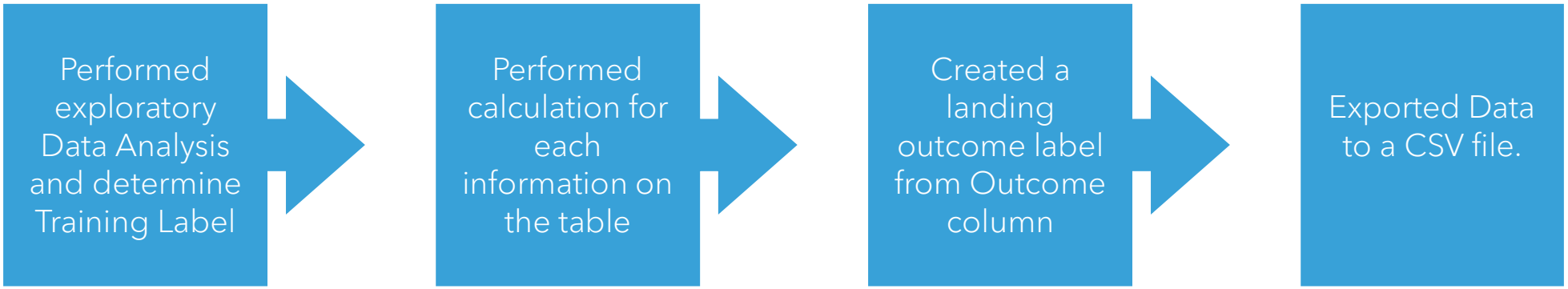
# Data Collection – Web Scraping




Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20Webscraping.ipynb

# 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. We mainly convert the outcomes from the table into Training Labels with "1" means the booster successfully landed, "0" means it was unsuccessful.




Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20Data%20Wrangling.ipynb

## EDA with data visualization

### Charts 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.

Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20EDA%20DataViz.ipynb

# 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.

Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20EDA%20SQL.ipynb

# 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.

Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20Maps.ipynb

# 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:

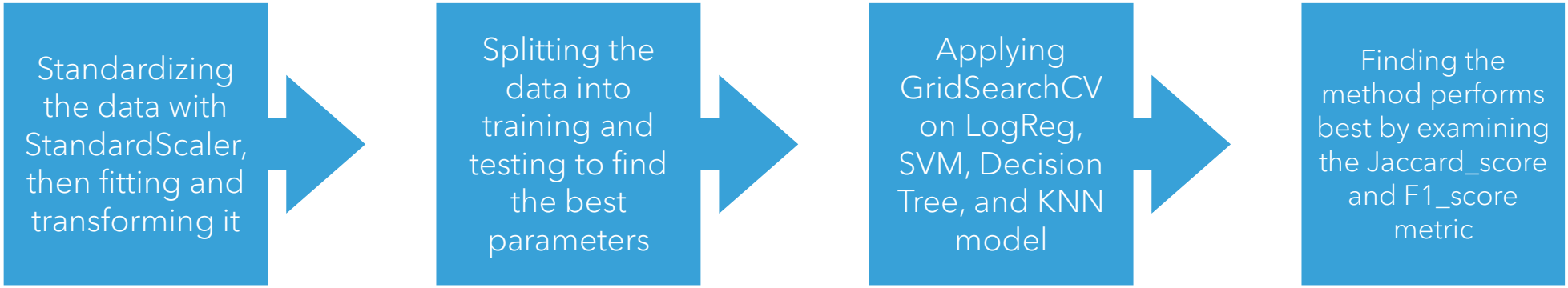
- Added a slider to select Payload range.

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.

Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20DashBoard.py

# Predictive analysis (Classification)




Source: https://github.com/diguitarrista/IBM-Data-Science-Capstone-Project/blob/main/IBM%20SpaceX%20Machine%20Learning%20Prediction.ipynb

# Results

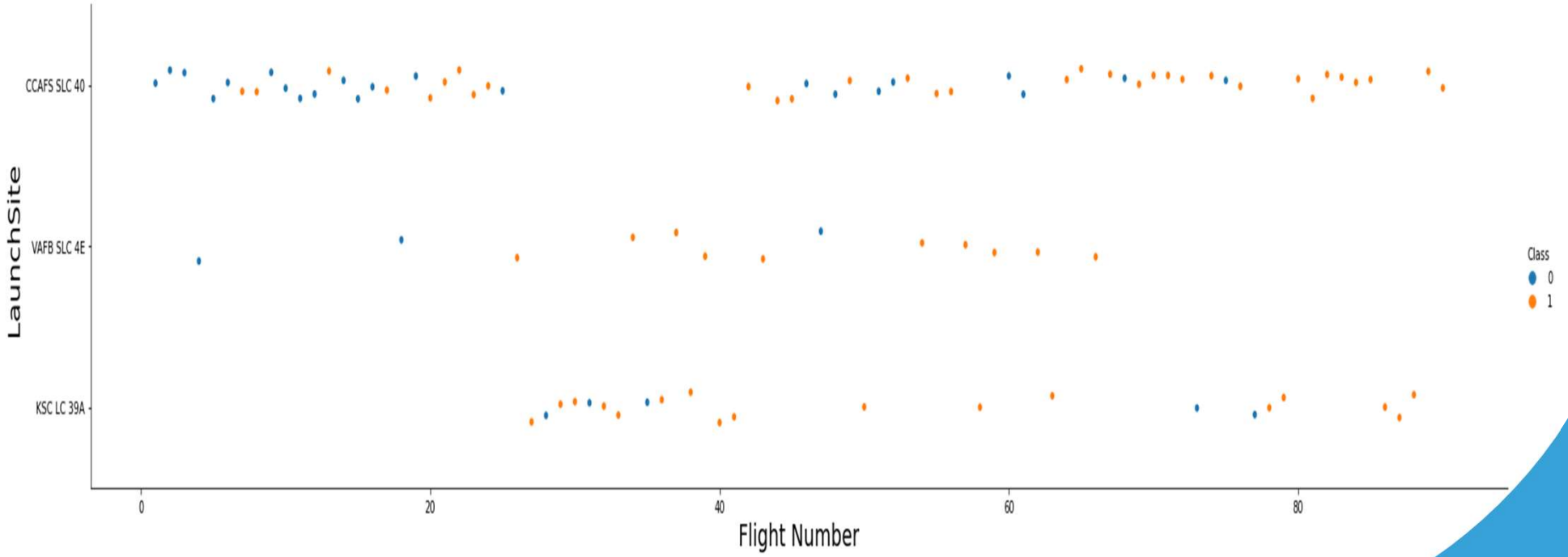
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

# EDA with Visualization

Flight Number vs. Launch Site



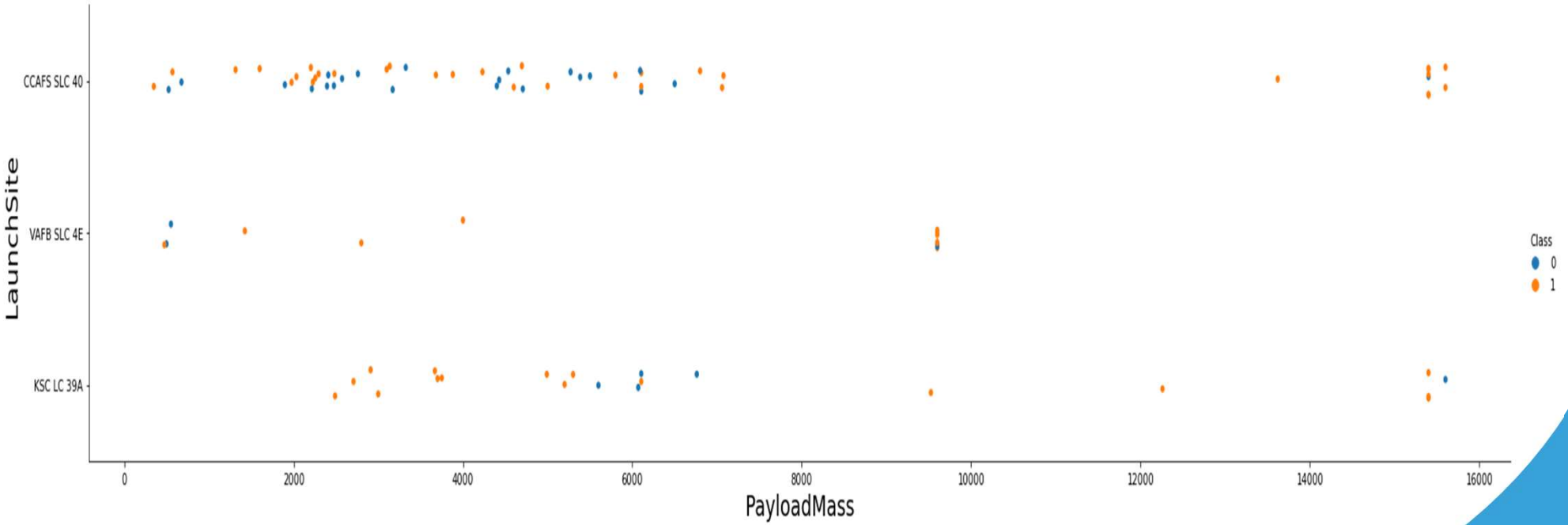
# EDA with Visualization

Flight Number vs. Launch Site

- According to the chart above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A.

# EDA with Visualization

Payload vs. Launch Site



# EDA with Visualization

Payload vs. Launch Site

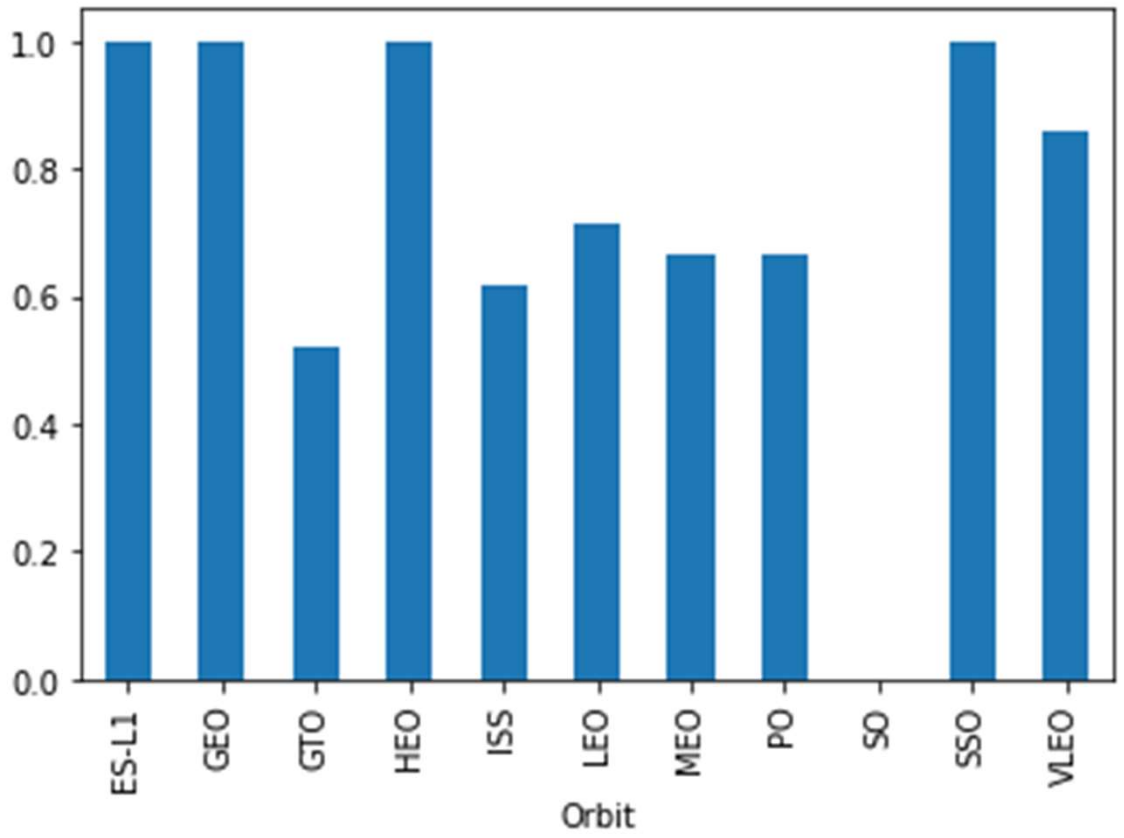
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rates;
- Payloads over 12,000kg are only possible on CCAFS SLC 40 and KSC LC 39A launch sites.

## EDA with Visualization

Success rate vs. Orbit type

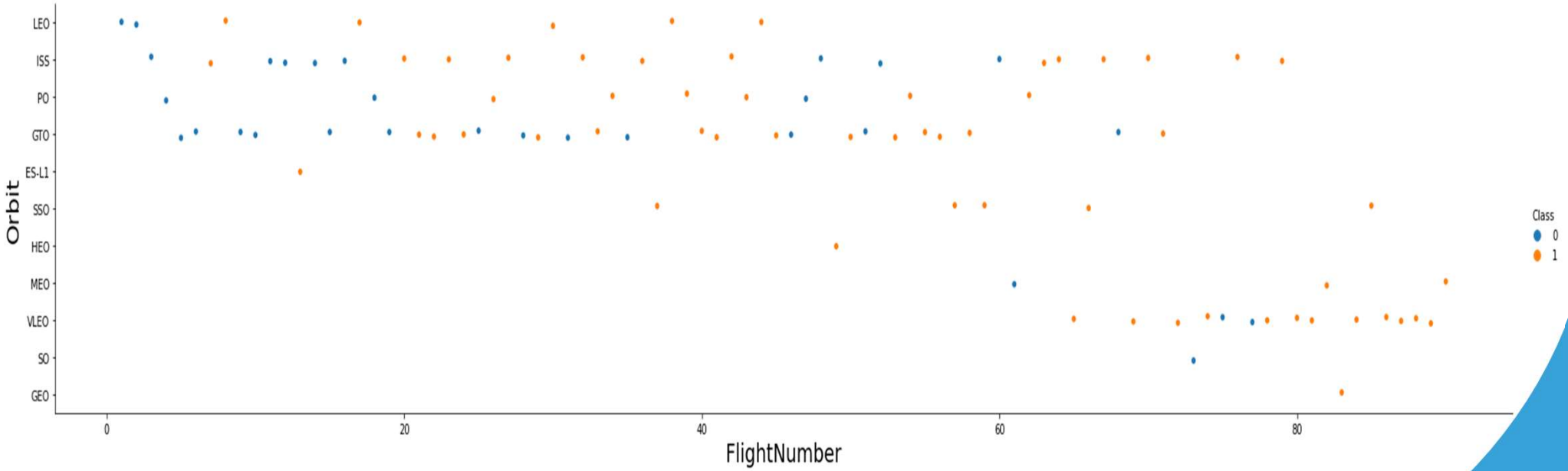
The biggest success rates happens to orbits:

- ES-L1;
- GEO;
- HEO;
- SSO.
- VLEO
- LFO



# EDA with Visualization

Flight Number vs. Orbit type



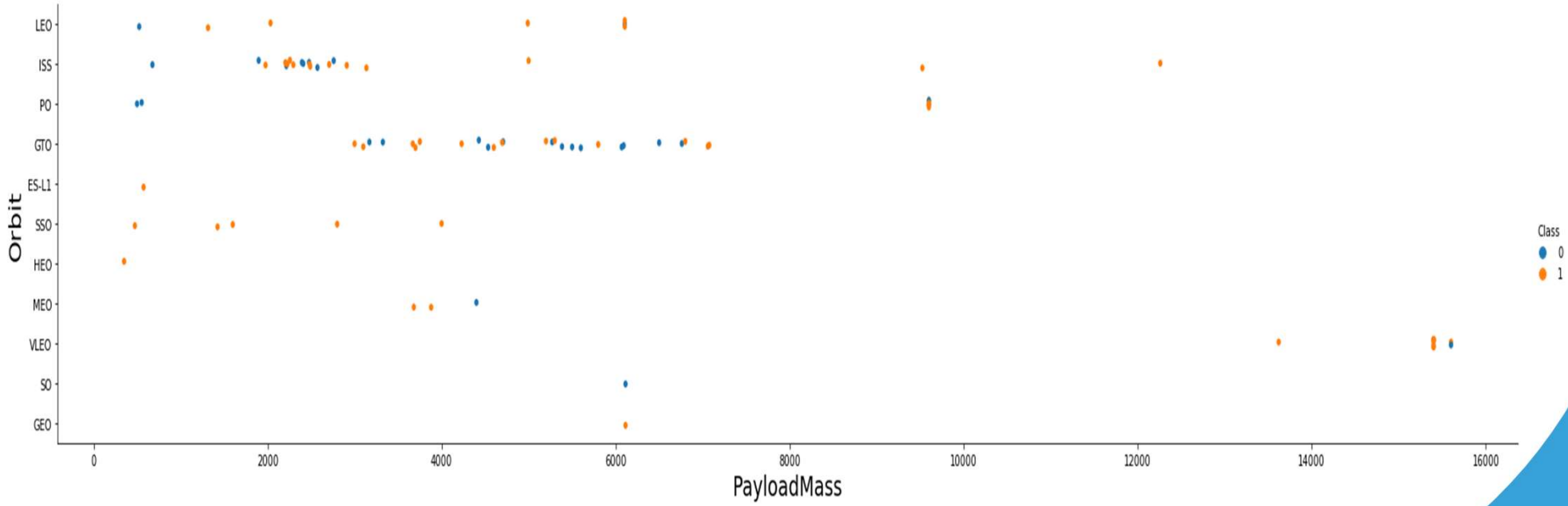
# EDA with Visualization

Flight Number vs. Orbit type

- Success rate improved over time to all orbits.

# EDA with Visualization

Payload vs. Orbit Type



# EDA with Visualization

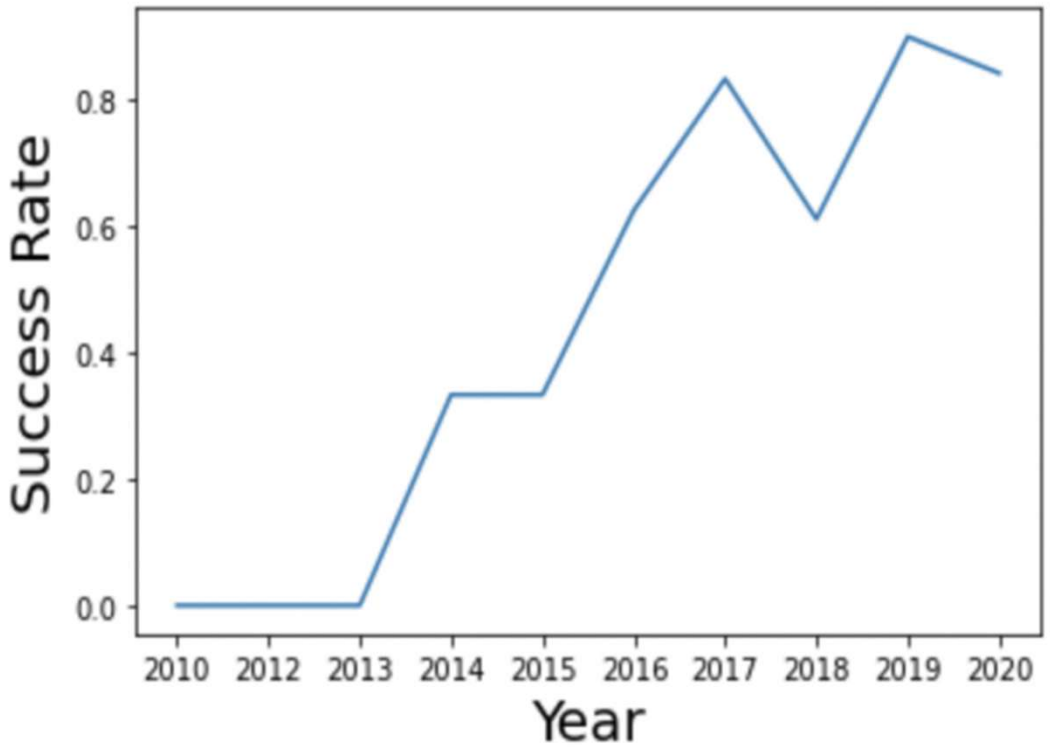
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

## EDA with Visualization

Launch success yearly trend

Success rate started increasing in 2013 and kept up until 2020.



# EDA with SQL

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 | 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 |

# EDA with SQL

- All launch site names: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- Total Payload Mass: 45596 kg
- Avarage Payload Mass: 2564 kg
- First successful ground landing date: 2015-12-22
- Successful drone ship landing with payload between 4000 and 6000:

    Booster version: F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, F9 FT B1026

# EDA with SQL

Total number of successful and failure mission outcomes:

Failure (in flight): 1

Success: 99

Success (payload status unclear): 1

Boosters carried maximum payload: F9 B5 B1048.4, F9 B5 B1048.5, F9 B5 B1049.4, F9 B5 B1049.5, F9 B5 B1049.7, F9 B5 B1051.3, F9 B5 B1051.4, F9 B5 B1051.6, F9 B5 B1056.4, F9 B5 B1058.3, F9 B5 B1060.2, F9 B5 B1060.3.

# EDA with SQL

2015 launch records:

| 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) |

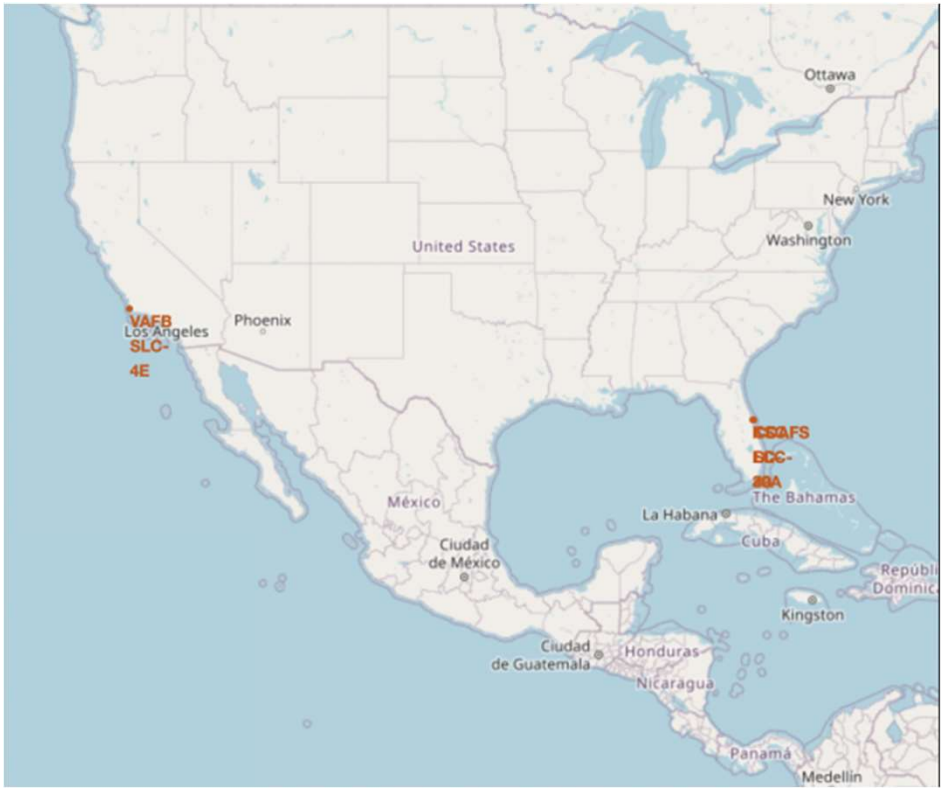
## EDA with SQL

Rank success count between 2010-06-04 and 2017-03-20

| 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 |

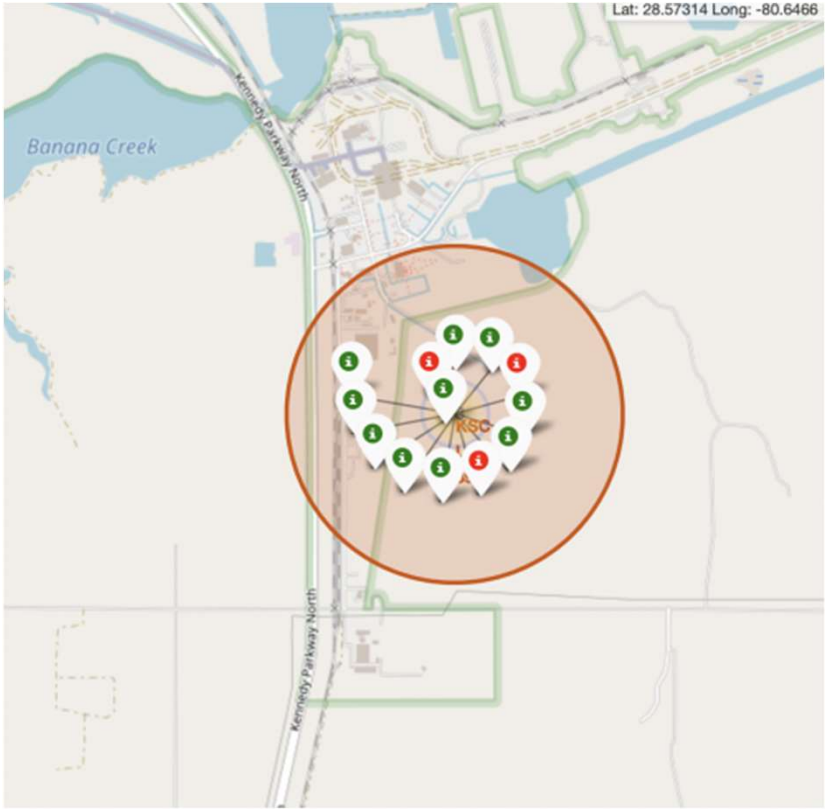
# Interactive map with Folium

All launch sites' location markers on a global map



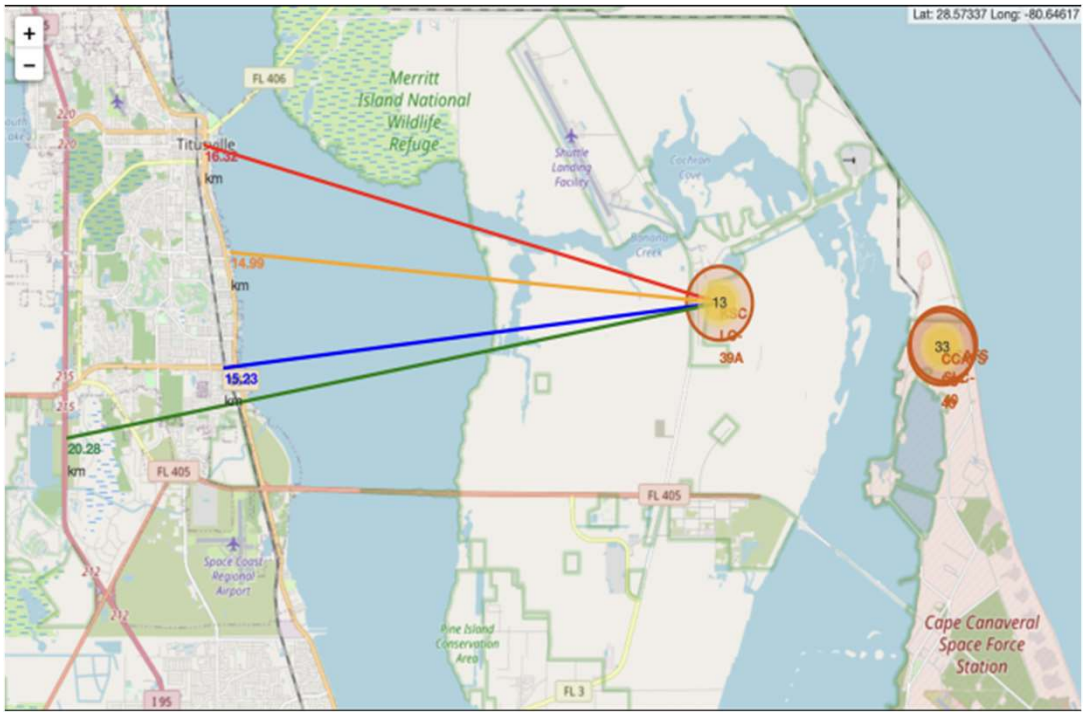
# Interactive map with Folium

Colour-labeled launch records on the map



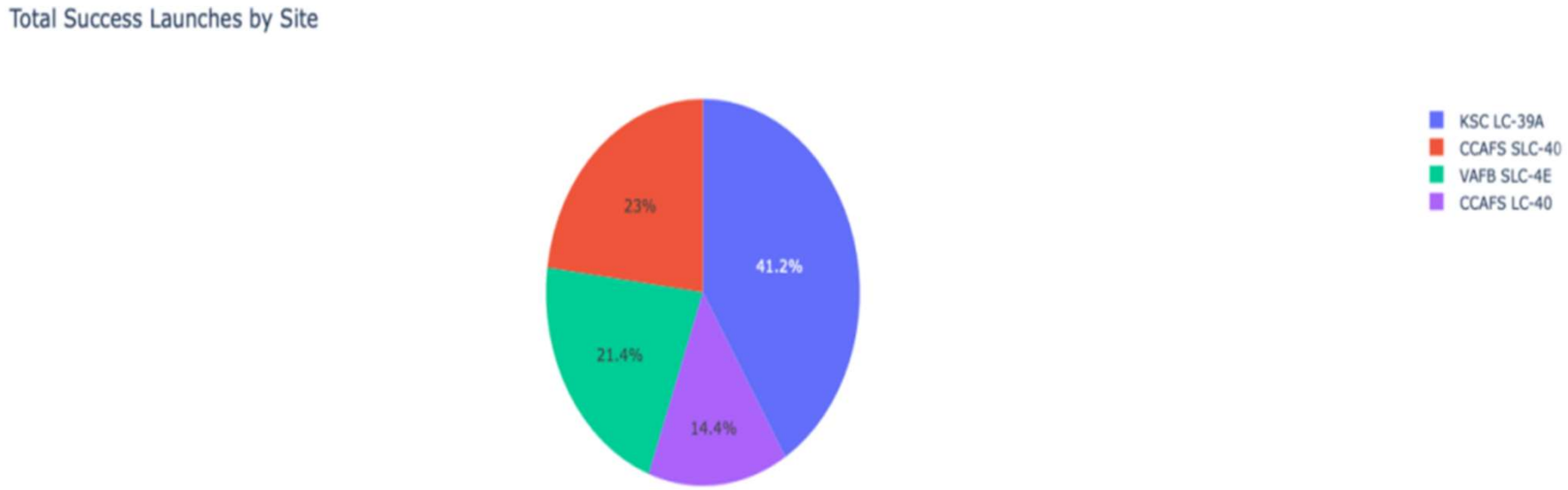
# Interactive map with Folium

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



# Build a Dashboard with Plotly Dash

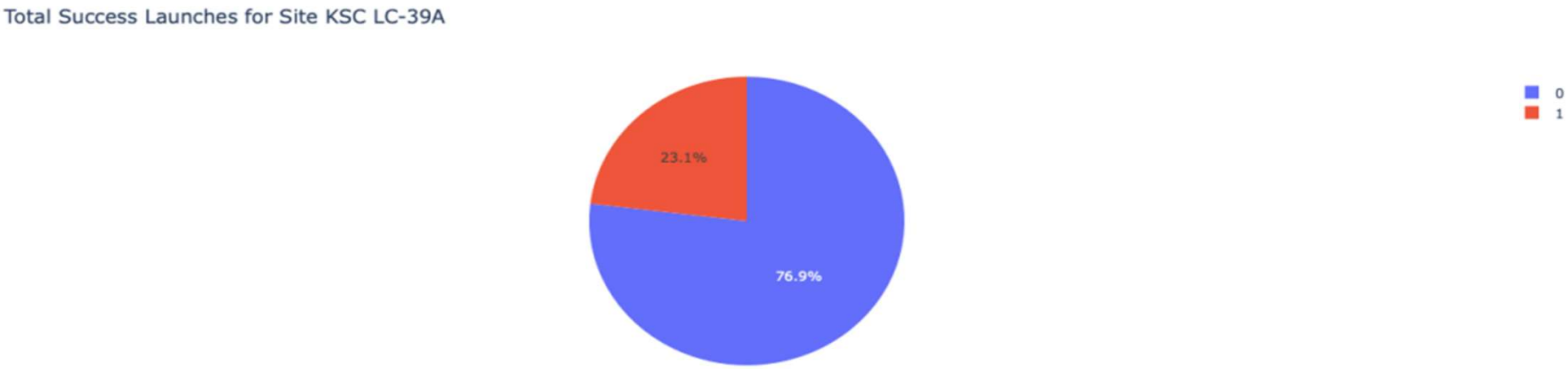
Launch success count for all sites




The place from where launches are done shows to be a very important factor of success of missions.

# Build a Dashboard with Plotly Dash

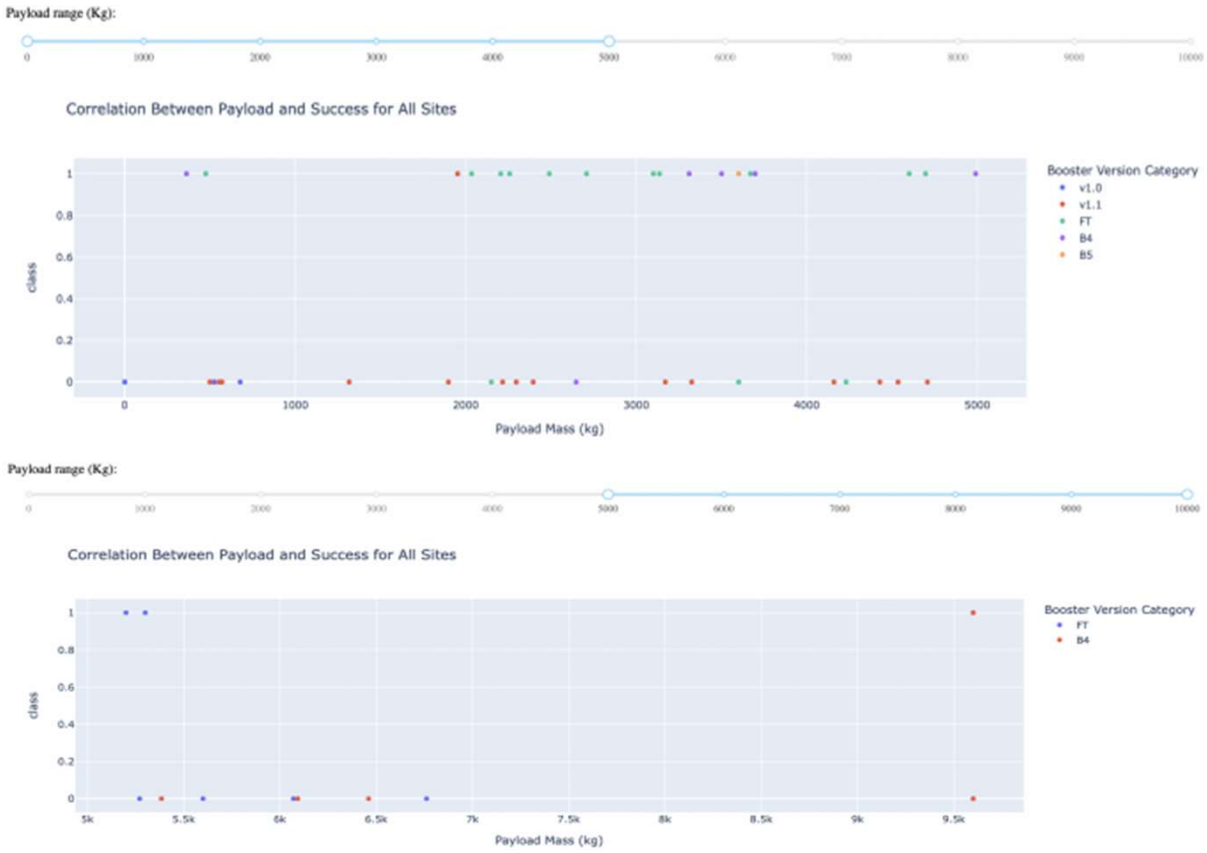
Launch site with highest launch success ratio




76.9% of launches are successful in this site.

# Build a Dashboard with Plotly Dash

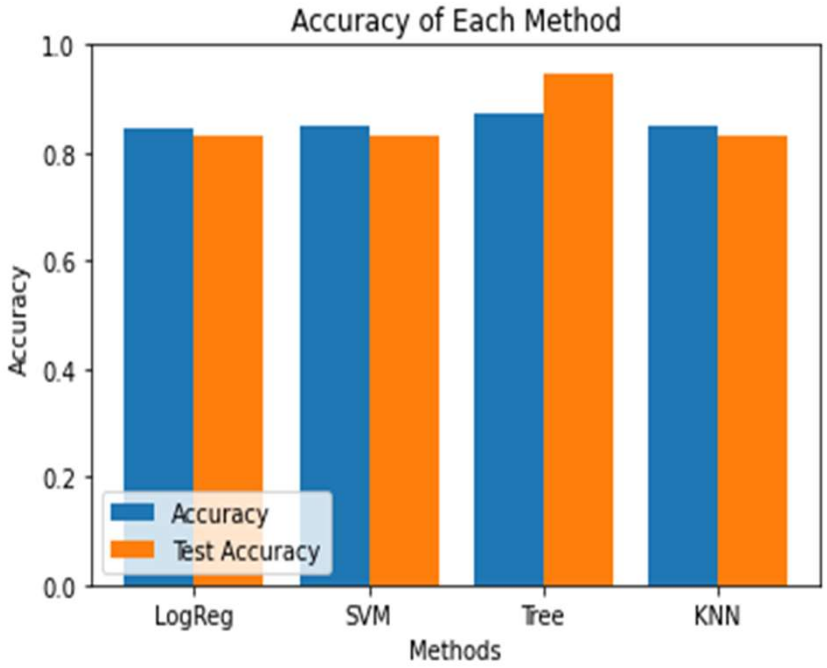
Payload Mass vs. Launch Outcome for all sites




- Payloads under 6,000kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg

# Predictive analysis (Classification)

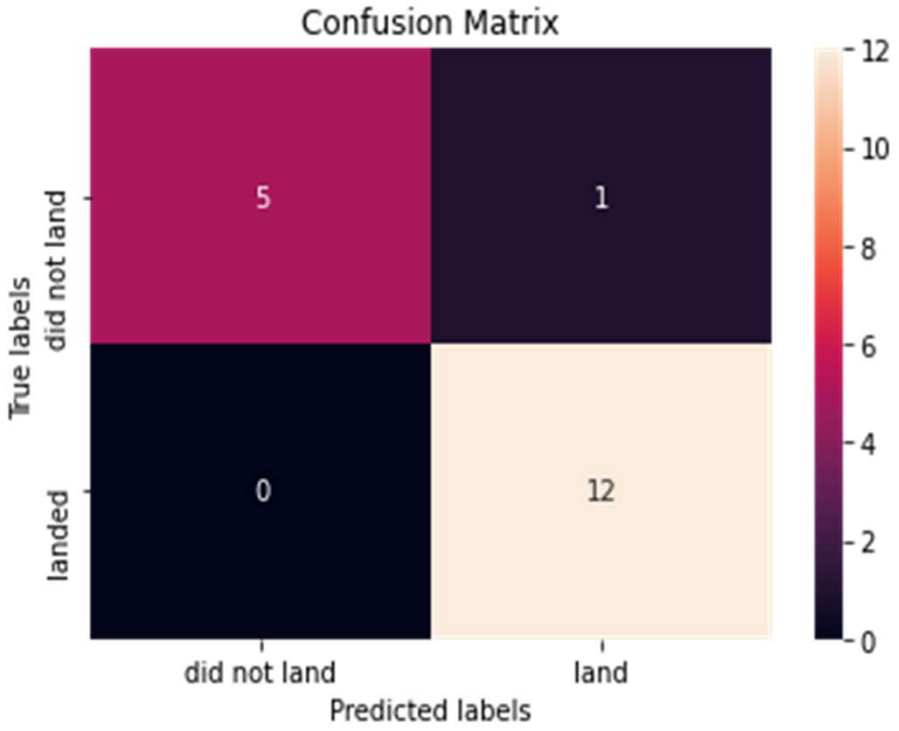
The scores of the whole Dataset confirm that the best model is the Decision Tree Model. Not only does this model have higher scores, but also the highest accuracy.




| Model | Accuracy | TestAccuracy |
|---|---|---|
| LogReg | 0.84643 | 0.83333 |
| SVM | 0.84821 | 0.83333 |
| Tree | 0.875 | 0.94444 |
| KNN | 0.84821 | 0.83333 |

# Confusion Matrix of Decision Tree Classifier

Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



# Conclusion

- According to the data gathered throughout this project and showed in this presentation, the Decision Tree Model is the best algorithm for this dataset.
- As for the characteristics of the successful launches, the data shows that a lower payload mass has better results than a larger payload mass.
- The places chosen for the launches are both closer to the Equator line and to the coast. The KSC LC-39A site has the highest success rate of the launches.
- And lastly, the success rate has been increasing over the years.