IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Yuan Du
Jun 22, 2022

# Outline

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

# Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection by using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA) by data wrangling, data visualization and interactive visual analytics;
  - Predictive Analysis
- Summary of all results
  - 11 orbit types ES L1, GEO, HEO, SSO were 100 successful with less than 6000 kg payload;
  - SpaceX has 4 launch sites, one is near California, the other three is near Florida and South Texas. All the sites are in near proximity to ocean and all the sites are bit far away from the city;
  - Machine Learning Prediction shows that all models performed similarly when trained with the data at hand

# Introduction

- Project background and context:
  - SpaceX is a company that aims to make commercial space travel more affordable for everyone.
  - This company can launch rockets for a cost of around 60 million dollars. In contrast, other providers require165 million dollars for one launch. This is due to the fact that SpaceX can reuse the first stage of the rocket Falcon9.
  - The primary cost saving agent is the high success rate of stage 1 landing and thus its reusability in future launches.
  - The challenge here is to set a right costing fore cast of the rocket launches through predicting its potential to land stage 1 successfully

- Problems we want to find answers:
- What are the features impact SpaceX to achieve the best results?
- What will be the accuracy of the prediction models?

Section 1

# Methodology

# Methodology

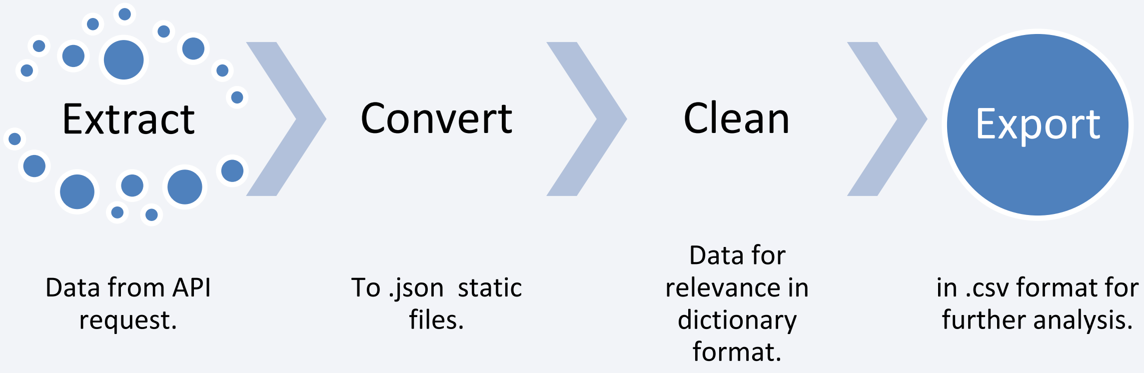
- Data collection methodology:
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics
- Perform predictive analysis using classification models
  - Build, tune, and evaluate classification models (KNN, SVM, Regression, Tree maps)

# Data Collection - SpaceX API

Data collection is the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques.

Data sets were collected from SpaceX API:

- Get response from API
- Convert Response to a .json file
- Apply custom functions to clean data
- Assign list to dictionary then create dataframe
- Export dataframe and export to .CSV

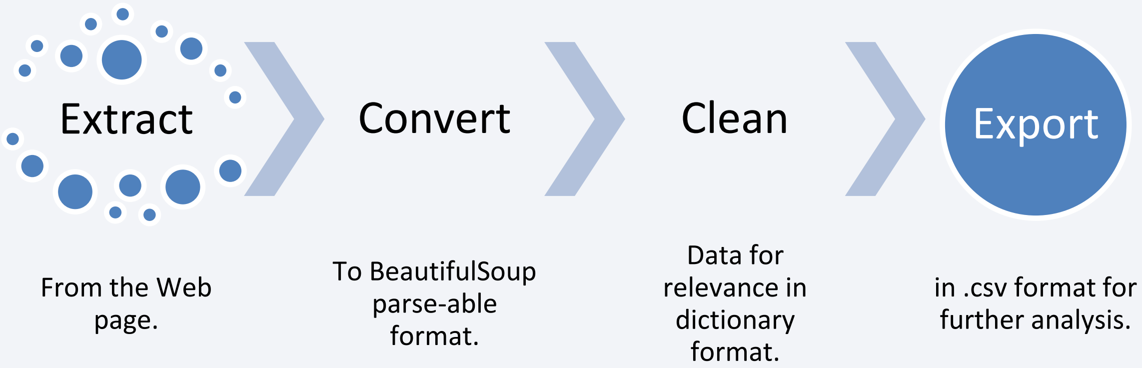



[GIT API]

# Data Collection - Web Scraping

Using web scraping technics to collect data:

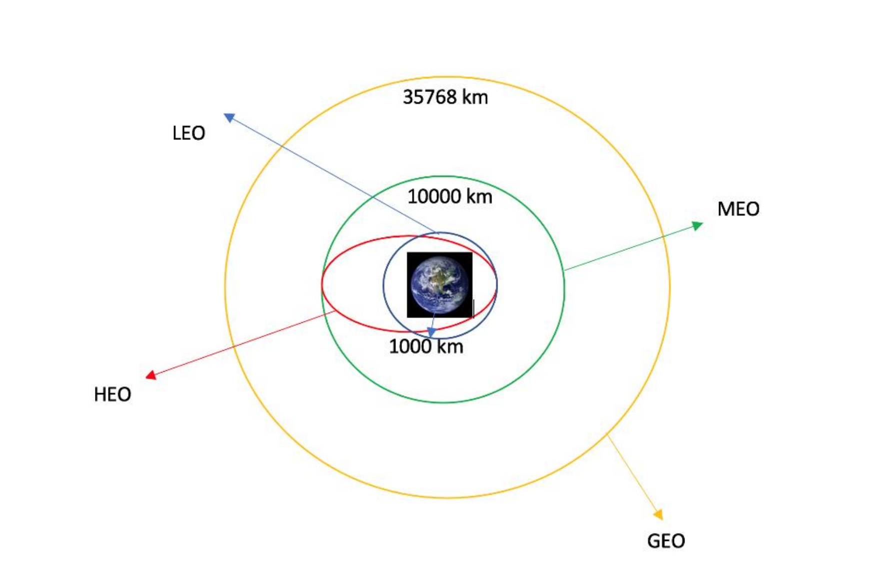
- Get response from HTML
- Create BeautifulSoup Object
- Find HTML Table
- Get column names
- Create dictionary and append data to keys
- Convert dictionary to dataframe
- Export dataframe to .CSV




[GIT Web Scraping]

# Data Wrangling

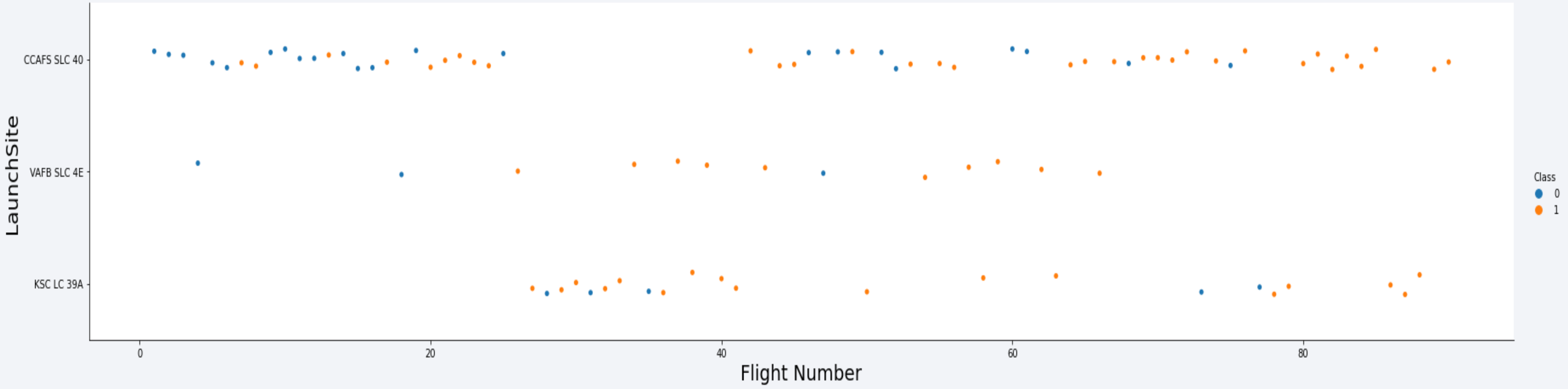
1. Calculate the number of launches per site.
2. Number of occurrence of each orbit.
3. Number of occurrences of outcome per orbit.
4. Create landing outcome label.




[GIT Data Wrangling]

# EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit




[GIT EDA with Data Visualization]()

# EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of 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.

[GIT EDA with SQL]

# Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers: Added to mark a specific area with a text label on a specific coordinate.
  - Circles: Added to highlight circle areas with a text label on a specific coordinate.
  - Marker Cluster: Marker clusters were used to simplify the containing many markers having the same coordinates.
  - Mouse Position: Used to get coordinate for a mouse over a point on the map (proximities). It helps to find the coordinates easily of any points of interests while exploring the map.
  - Polyline: It draws polyline overlays on a map. It was used to denote the distance between a launch site and its proximities(such as Railway station, city, etc.).

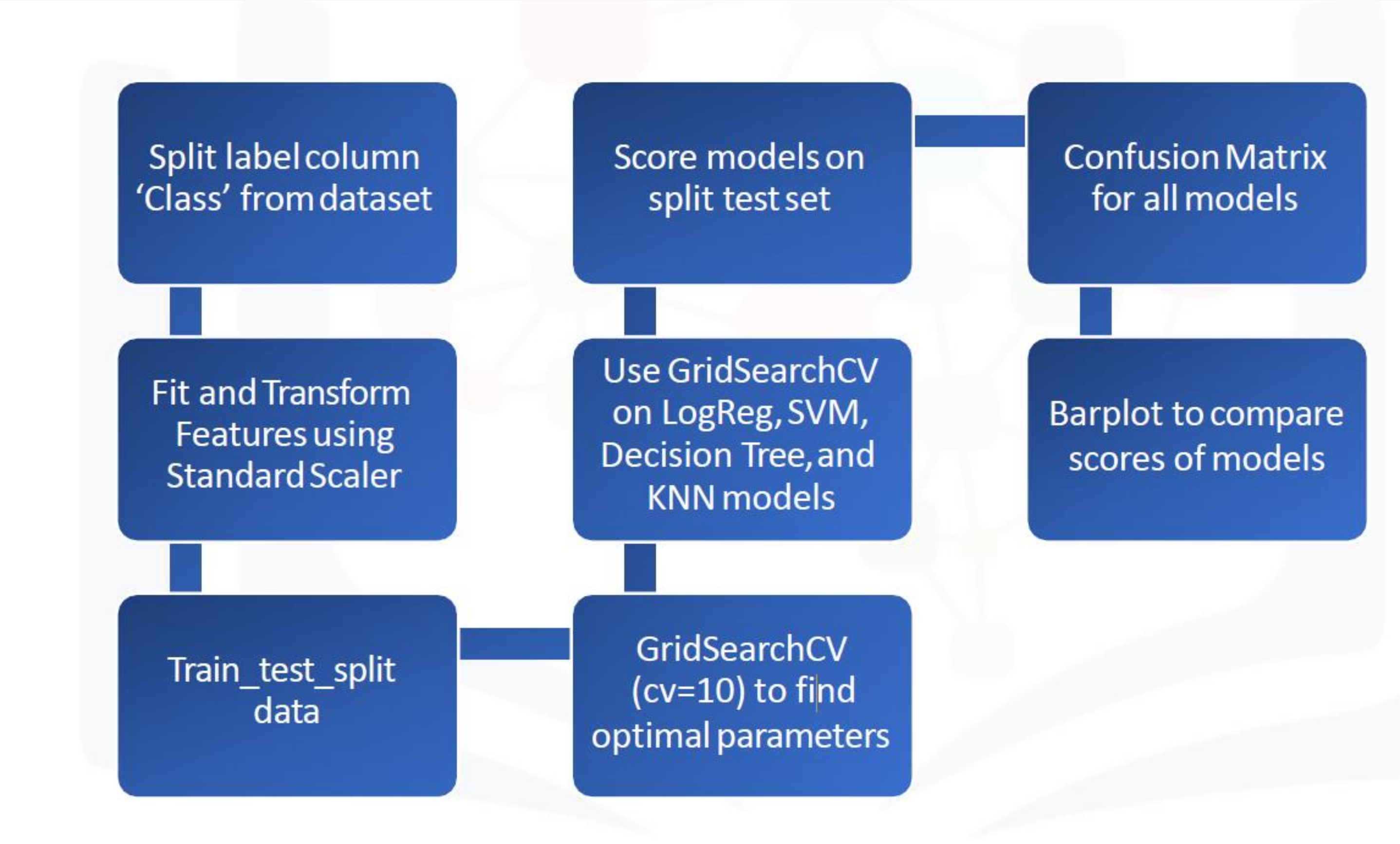
GIT Interactive Visual Analytics

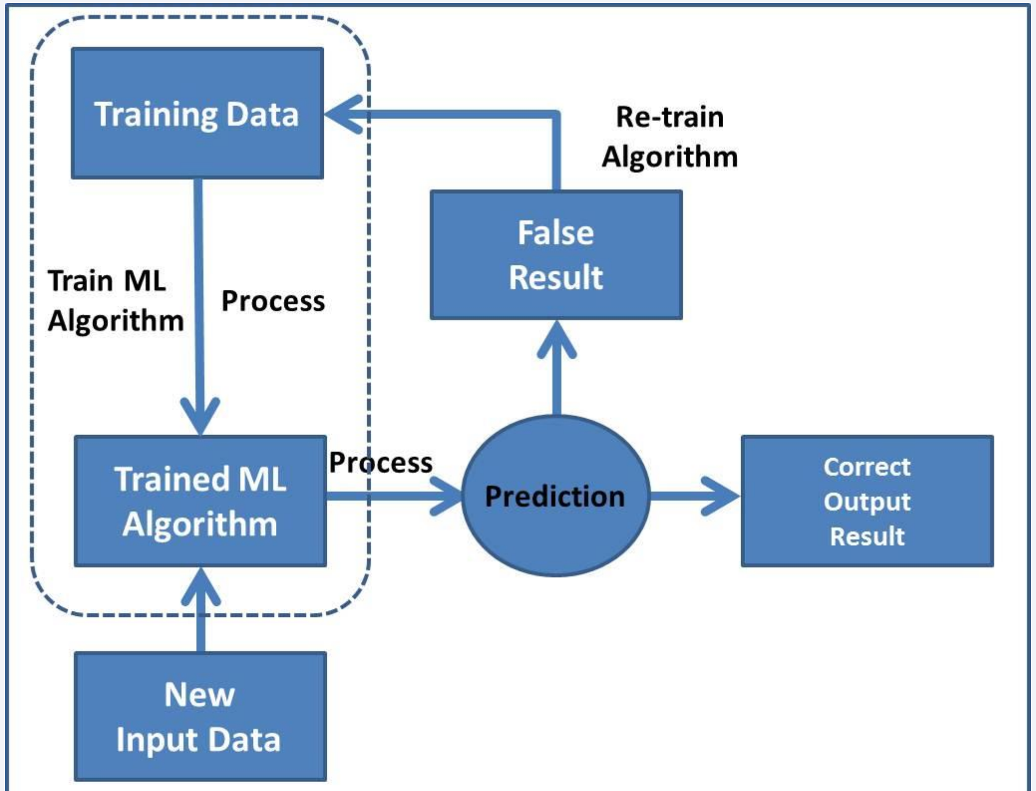
# Build a Dashboard with Plotly Dash

- Pie chart (total launches for a selected site or the total sites collection)
  - Percentage of success vs. failure for a given site
  - Relative proportions of different sites successful landing distribution
- Scatter Plot
  - correlation between Outcome and Payload Mass(Kg) for different Booster Versions with freedom of selection of the range of payload mass of Interest.

[GIT SpaceX Dash App]

# Predictive Analysis (Classification)



Machine Learning Algorithm Processing Flow Chart

[GIT Machine Learning Prediction]

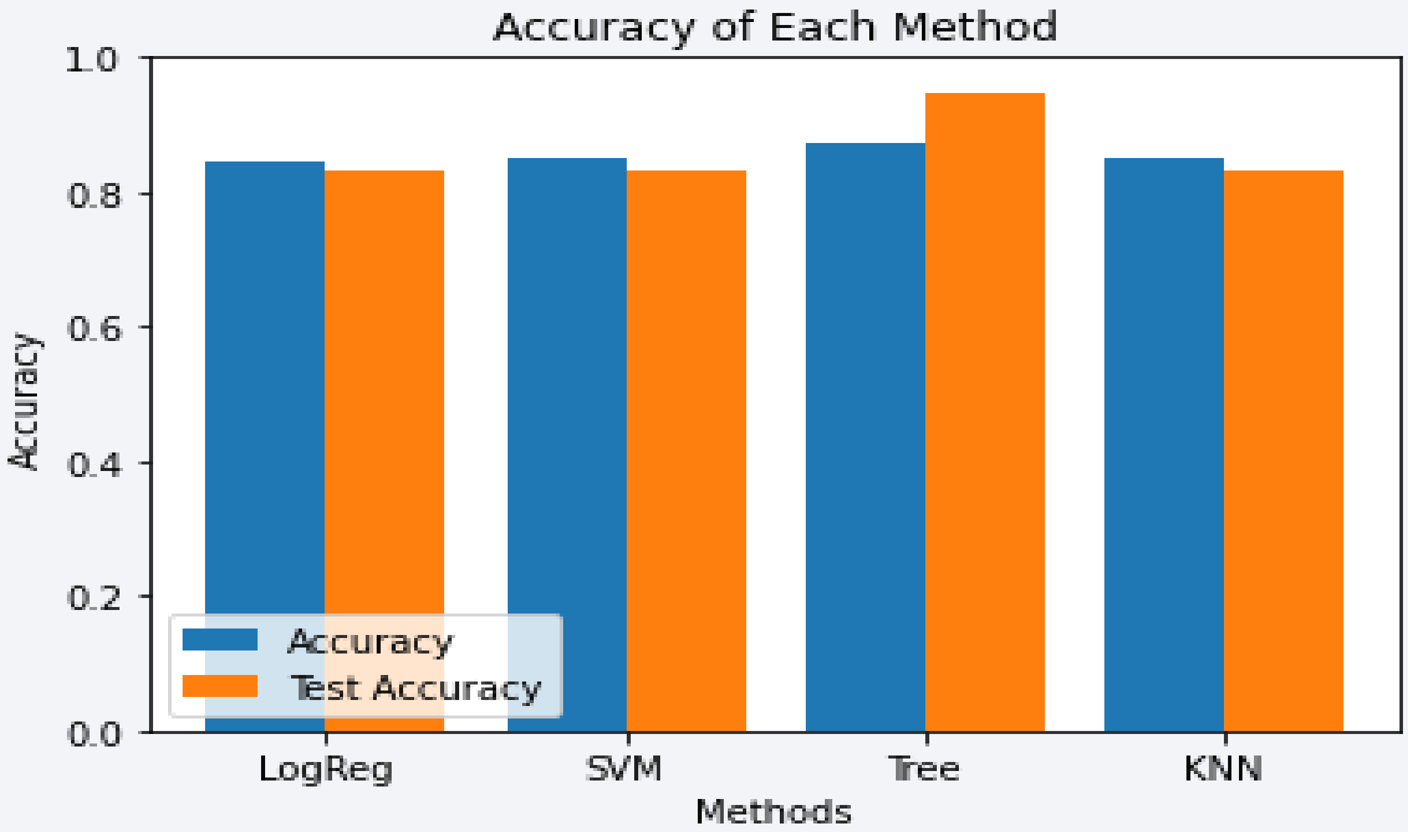
# Results

- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

# Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.


# Results

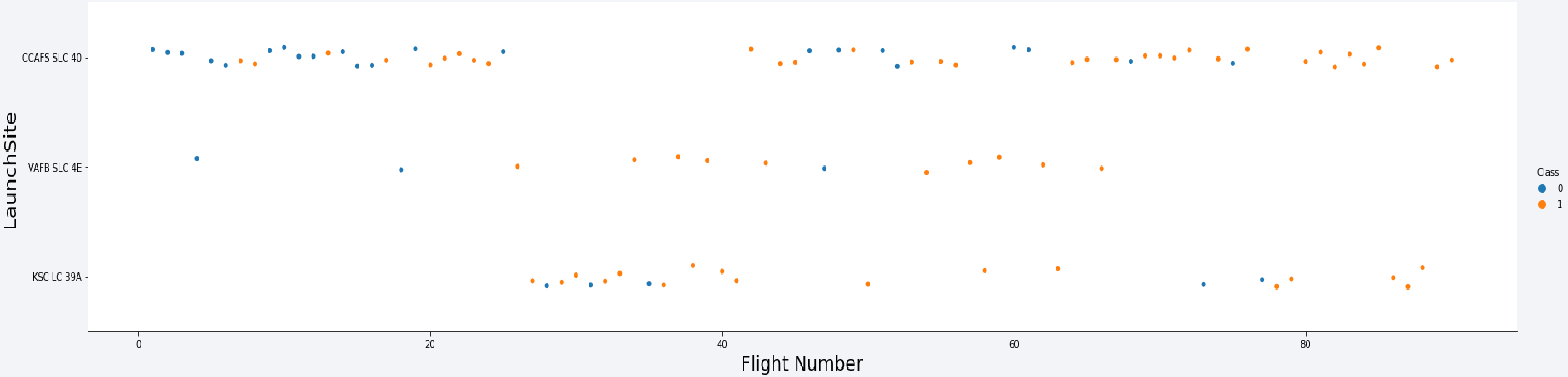



- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 0.884 and accuracy for test data over 0.946

Section 2

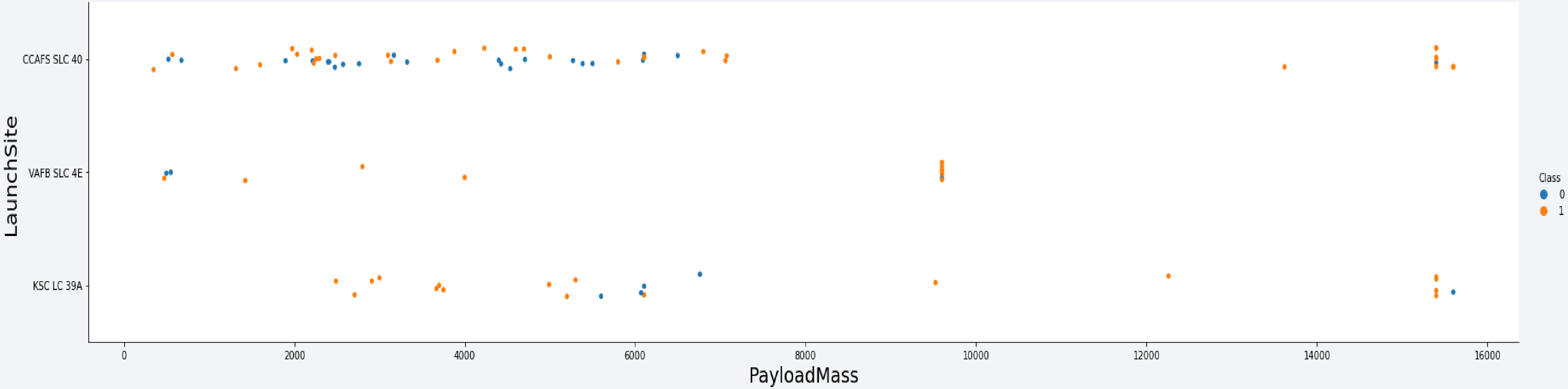
# Insights drawn from EDA

# Flight Number vs. Launch Site




- According to the plot 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 KSCLC 39A;
- It’s also possible to see that the general success rate improved over time.

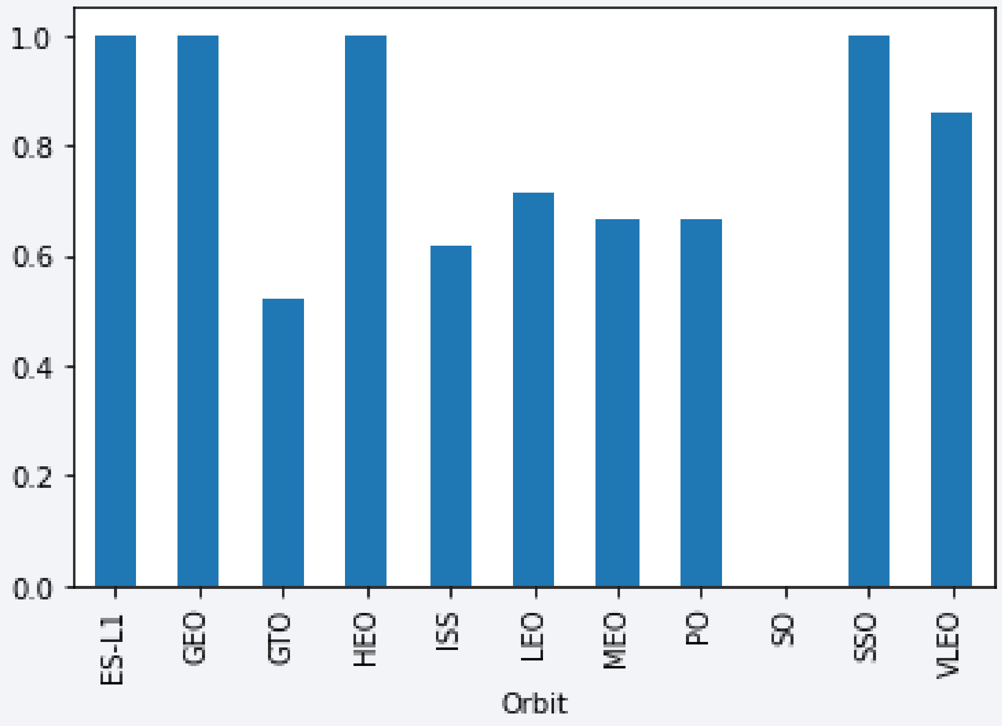
# Payload vs. Launch Site



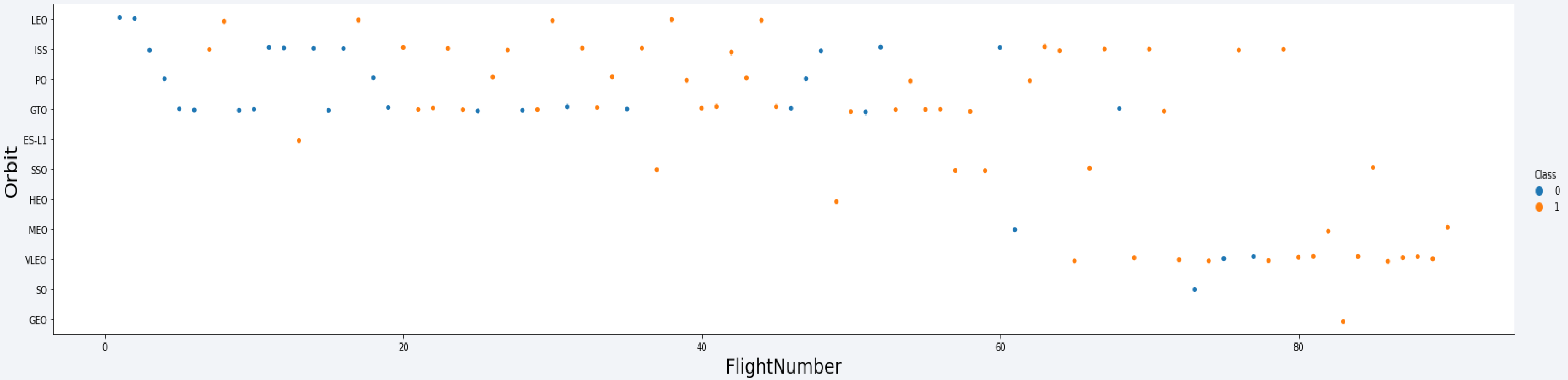

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

# Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

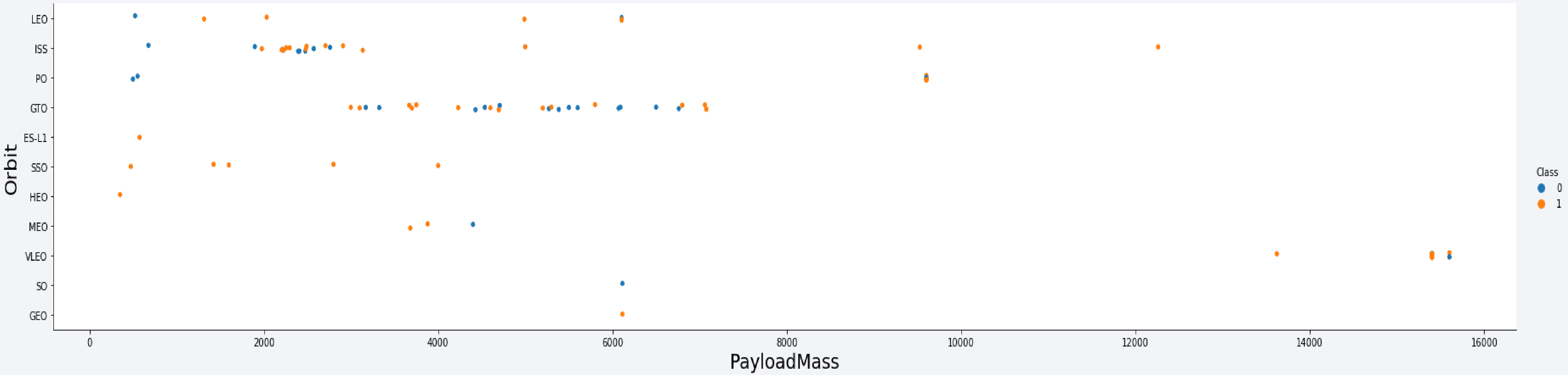


# Flight Number vs. Orbit Type




- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

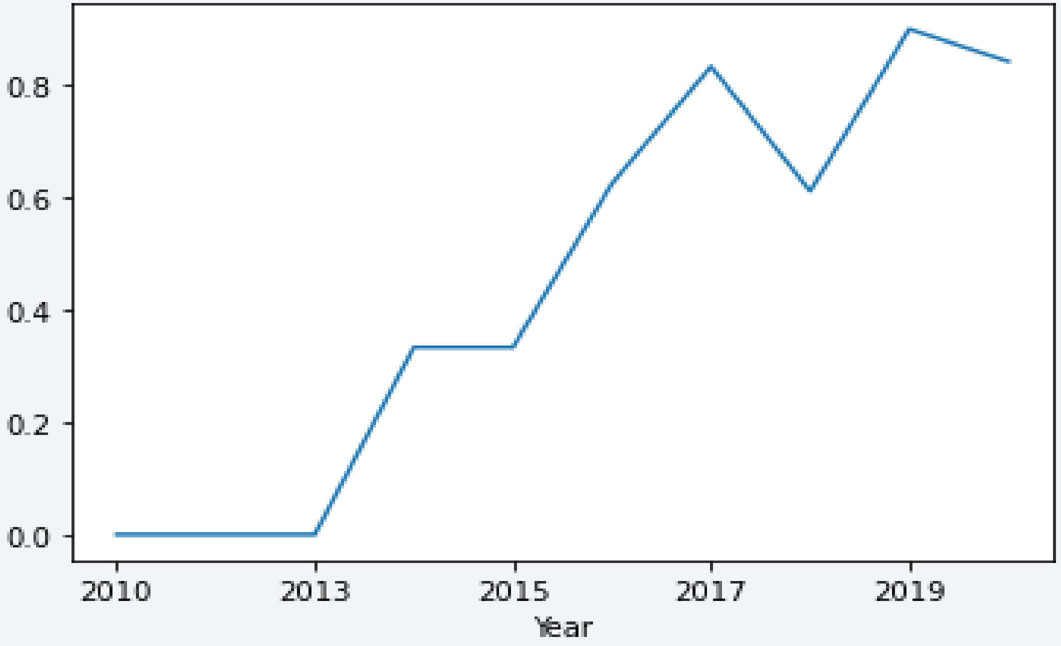
# Payload vs. Orbit Type




- Apparently, 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.

# Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



# All Launch Site Names

- According to data, there are four launch sites:

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

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

# Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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 |

- Here we can see five samples of Cape Canaveral launches.

# Total Payload Mass

- Total payload carried by boosters from NASA:

| Total Payload (kg) |
| --- |
| 111.268 |

- Total payload calculated above, by summing all payloads whose codes contain 'CRS', which corresponds to NASA.

# Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

| Avg Payload (kg) |
|---|
| 2.928 |

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

# First Successful Ground Landing Date

- First successful landing outcome on ground pad:

| Min Date |
|---|
| 2015-12-22 |

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

| Booster Version |
| --- |
| F9 FT B1021.2 |
| F9 FT B1031.2 |
| F9 FT B1022 |
| F9 FT B1026 |

- Selecting distinct booster versions according to the filters above, these 4 are the result.

# Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

| mission_outcome | 2 |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- Grouping mission outcomes and counting records for each group led us to the summary above.

# Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

| Booster Version (...) |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1048.5 |
| F9 B5 B1049.4 |
| F9 B5 B1049.5 |
| F9 B5 B1049.7 |
| F9 B5 B1051.3 |

| Booster Version |
|---|
| F9 B5 B1051.4 |
| F9 B5 B1051.6 |
| F9 B5 B1056.4 |
| F9 B5 B1058.3 |
| F9 B5 B1060.2 |
| F9 B5 B1060.3 |

- These are the boosters which have carried the maximum payload mass registered in the dataset.

# 2015 Launch Records

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

| Booster Version | Launch Site |
| --- | --- |
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

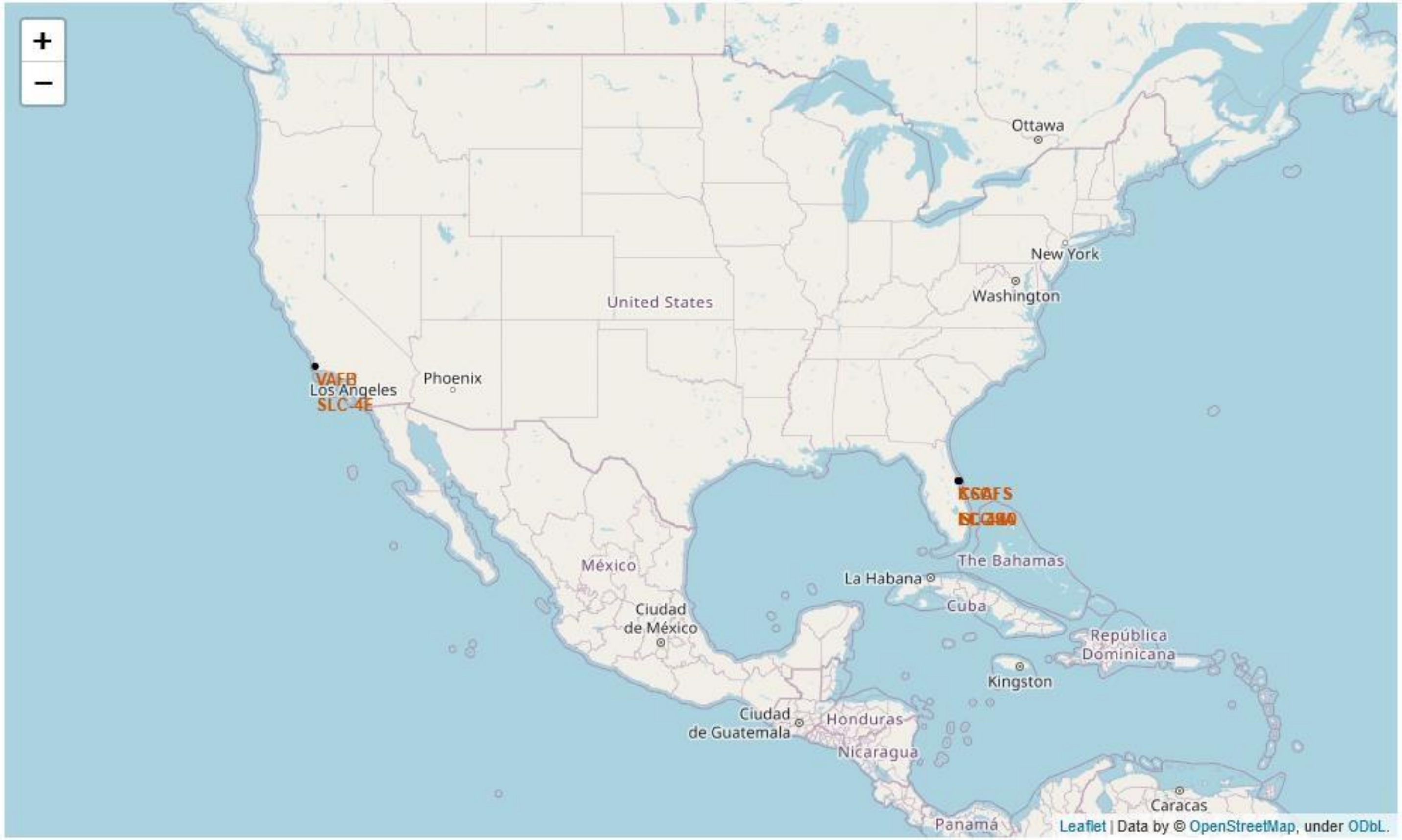
| Landing Outcome | Occurrences |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

- This view of data alerts us that “No attempt” must be taken in account.

Section 4

# Launch Sites Proximities Analysis

# All launch sites



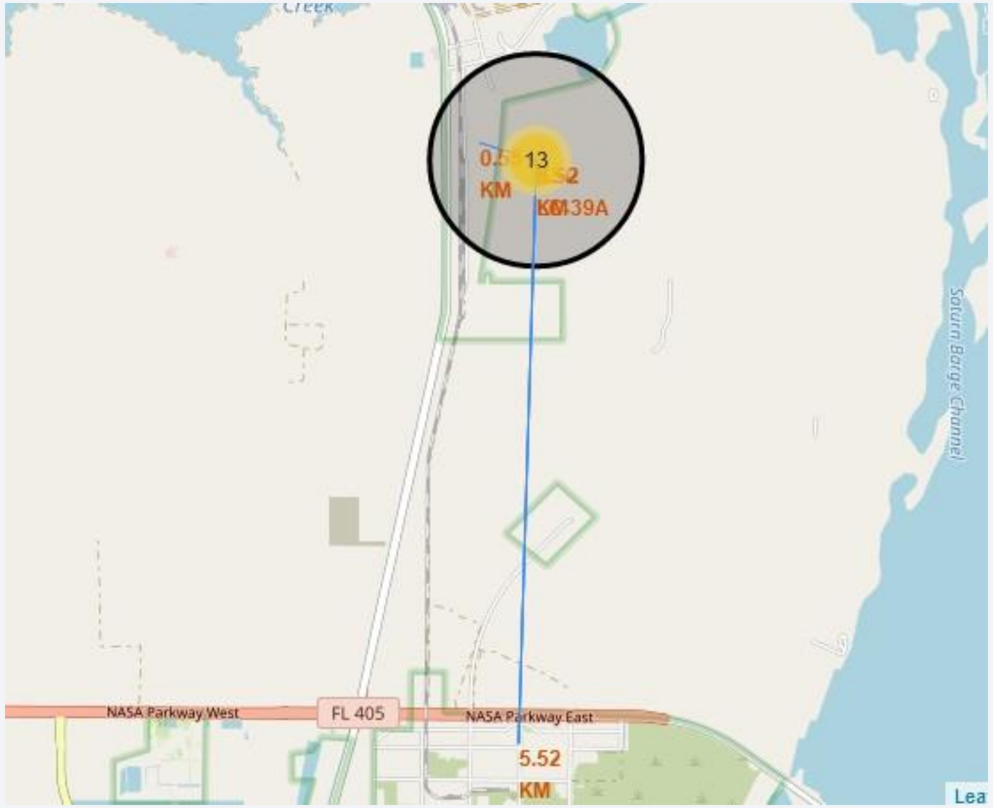

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

# Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes


- Green markers indicate successful and red ones indicate failure.

# Logistics and Safety

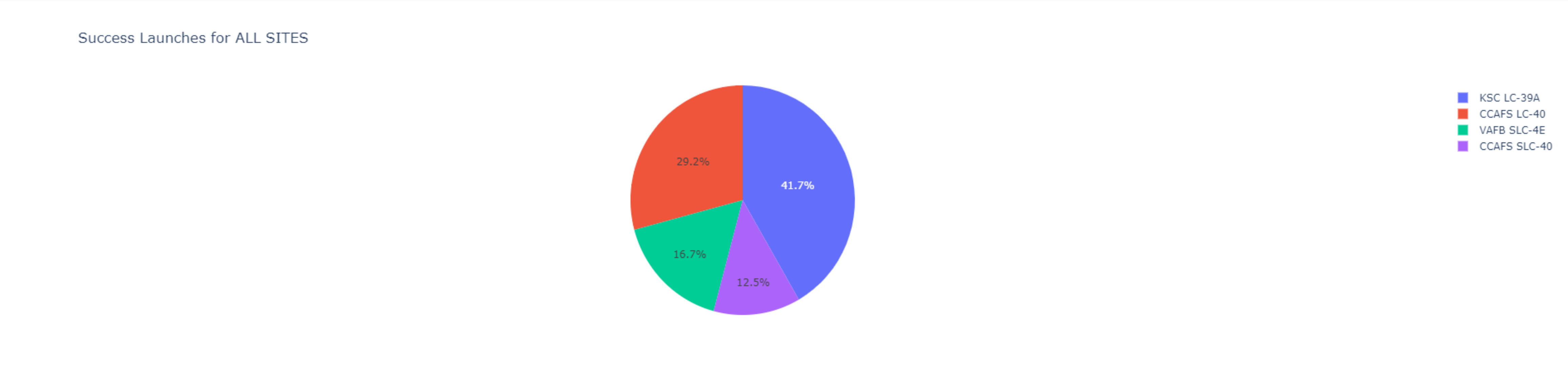



- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 5

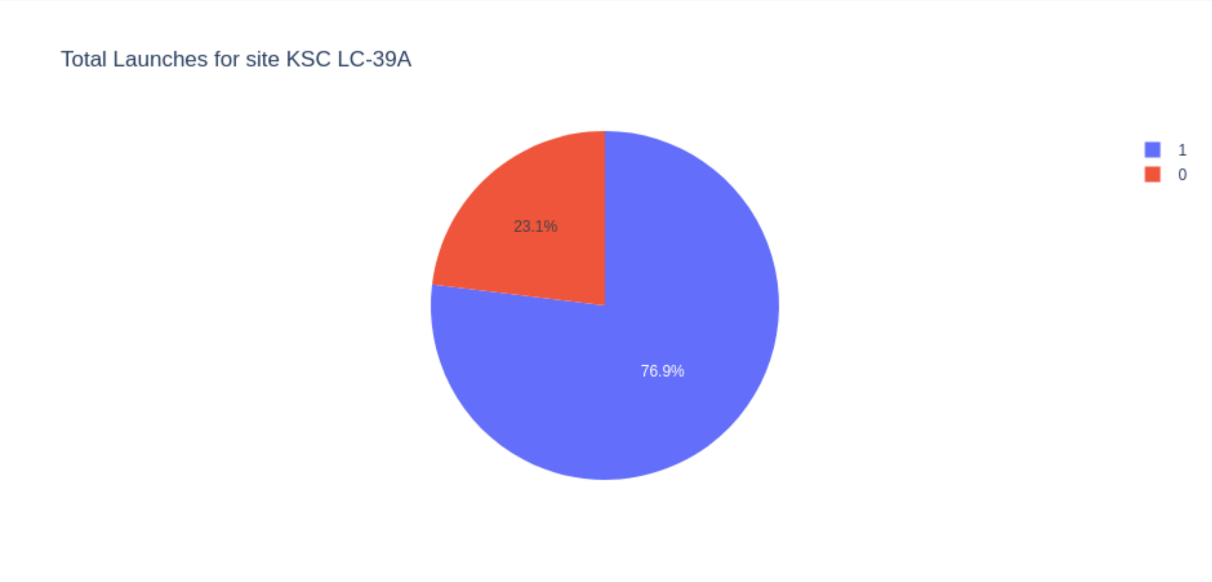
# Build a Dashboard with Plotly Dash

# Successful Launches by Site



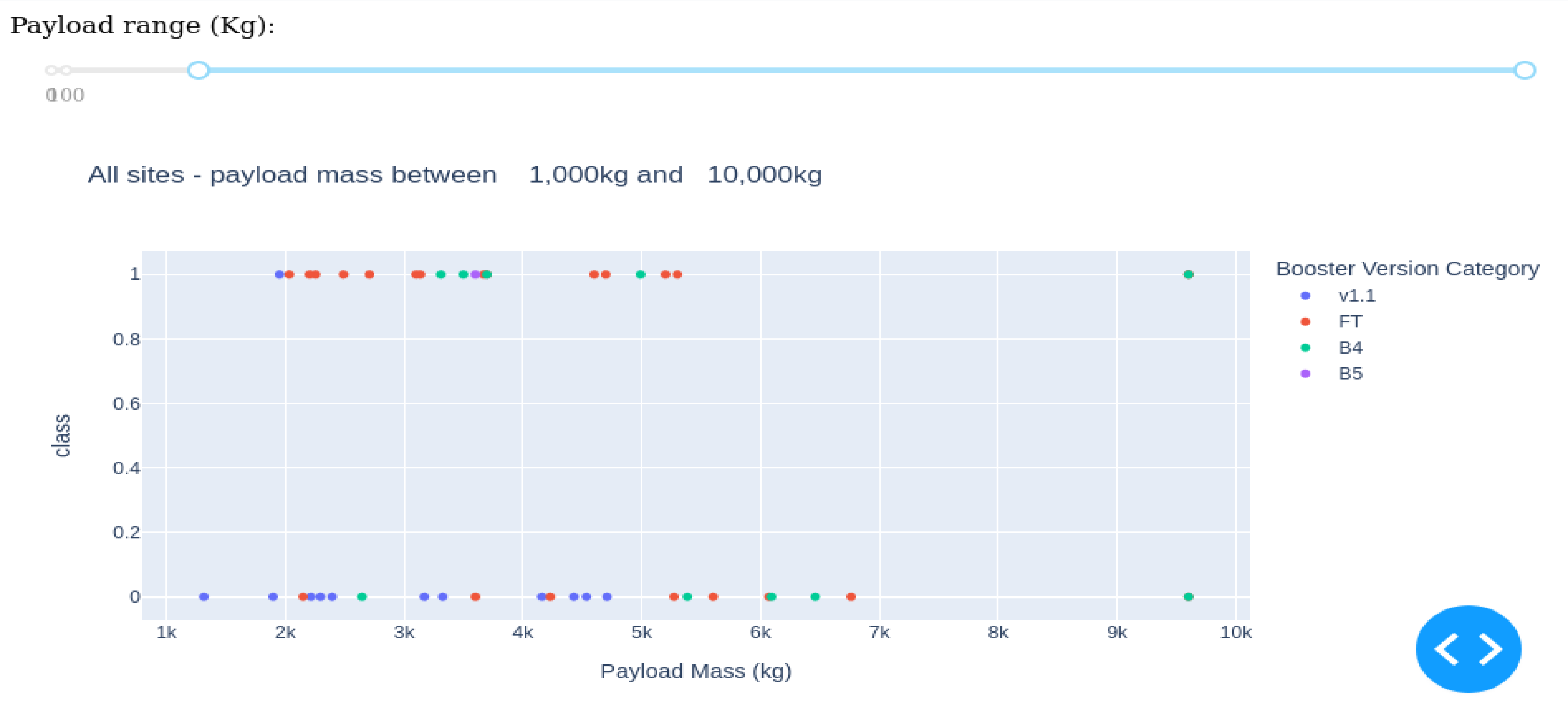

- The place from where launches are done seems to be a very important factor of success of missions.

# Launch Success Ratio for KSC LC-39A



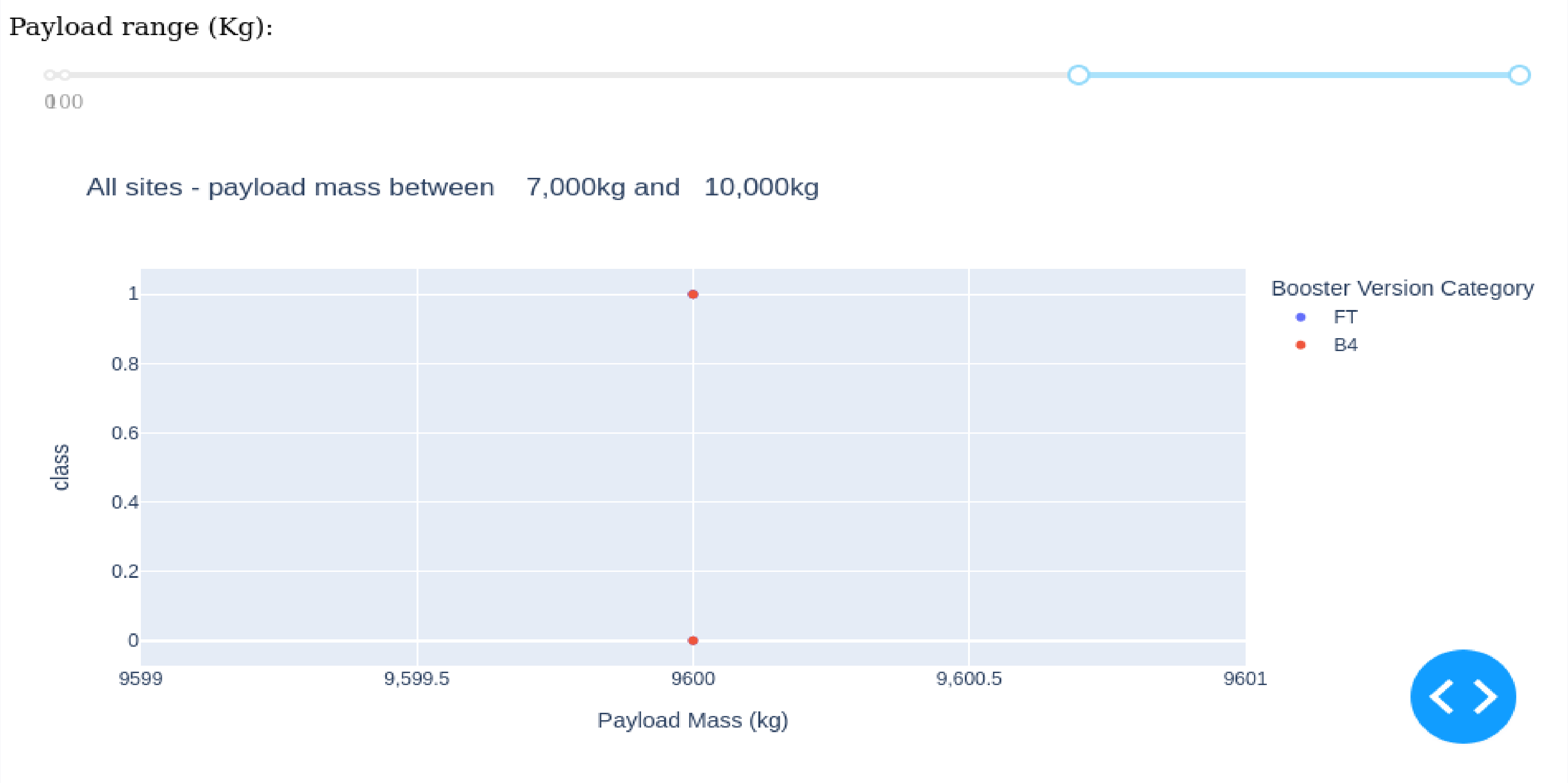

- 76.9% of launches are successful in this site.

# Payload vs. Launch Outcome




- Payloads under 6,000kg and FT boosters are the most successful combination.

# Payload vs. Launch Outcome




- There's not enough data to estimate risk of launches over 7,000kg

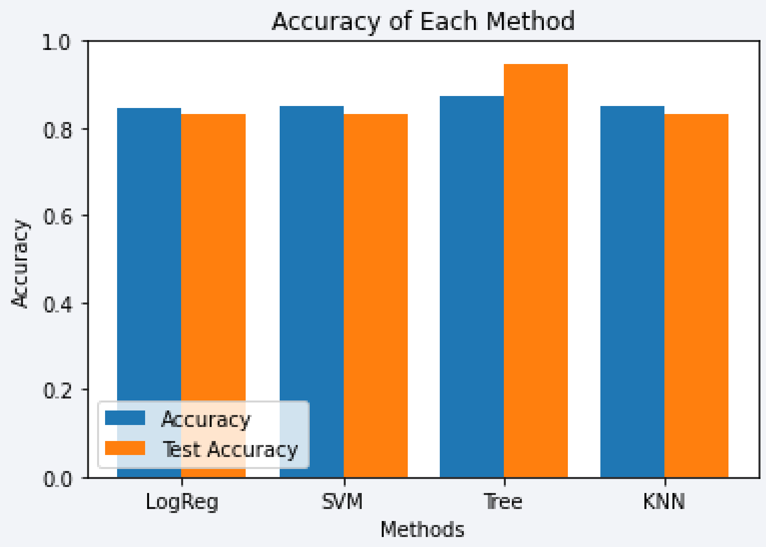
Section 6

# Predictive Analysis (Classification)

# Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 0.884



# Confusion Matrix

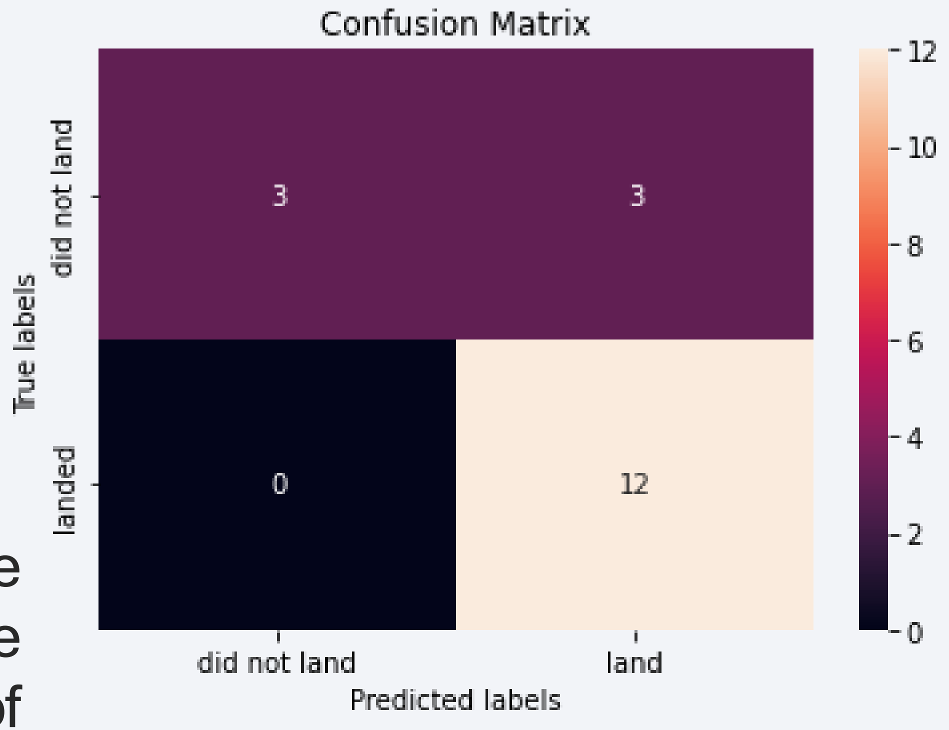
The Confusion Matrix of the Decision Tree Classifier

True Positive : 12

False Negative : 0

True Negative : 3

False Positive : 3




The model is quite interesting as it predicts a lot of times the good labels, however 3 times it predicted the success of the mission and the mission failed. Reducing the amount of False Positive would be a good idea to avoid spending Millions and years of work.

It could be done using Boosting or maybe look at a model with a lower accuracy but a better precision.

# Conclusions

- There are many parameters when considering launching rockets in space:
  - The Booster version is definitely one of this essential parameter.
  - The Orbit, Payload Mass are also important.
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- More data would be useful to have better, I have no doubt that engineer and scientists use these data in their predictions.

# Thank you!