IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diego Allendes Rojas
18/02/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX REST API and web Scraping
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics using Folium and Dashboard.
  - Machine Learning Prediction of Outcome of Flight.
- Summary of all results
  - Its possible collect data from open to public sources, like Wikipedia and government and private database.
  - EDA allowed to identify the best variables to predict the success of the landings.
  - Machine Learning Prediction showed the best model to predict which characteristics are important, using all collected data.

Introduction

- The objective of this research is to assess the feasibility of a new company called SpaceY to compete with SpaceX.
- Desirable answers:
  - The best way to estimate the total cost of launches, predicting successful landings of the first stage rocket.
  - How do variables affect the success of the first stage landing.
  - Does the rate of successful landings increase over the years.
  - It is possible to predict the success of the rocket's first stage landing using the variables that are available prior to launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data form SpaceX launches was obtained from 2 sources:
    - SpaceX REST API
    - Web Scraping from Wikipedia Falcon 9 page
- Perform data wrangling
  - Filtering the data, deleting old rocket launches
  - Dealing whit 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
  - After filtering, dealing and binary encoding, the data were divided into training and test data sets, evaluated using four different classification models, and the accuracy of each model was obtained.

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 page.

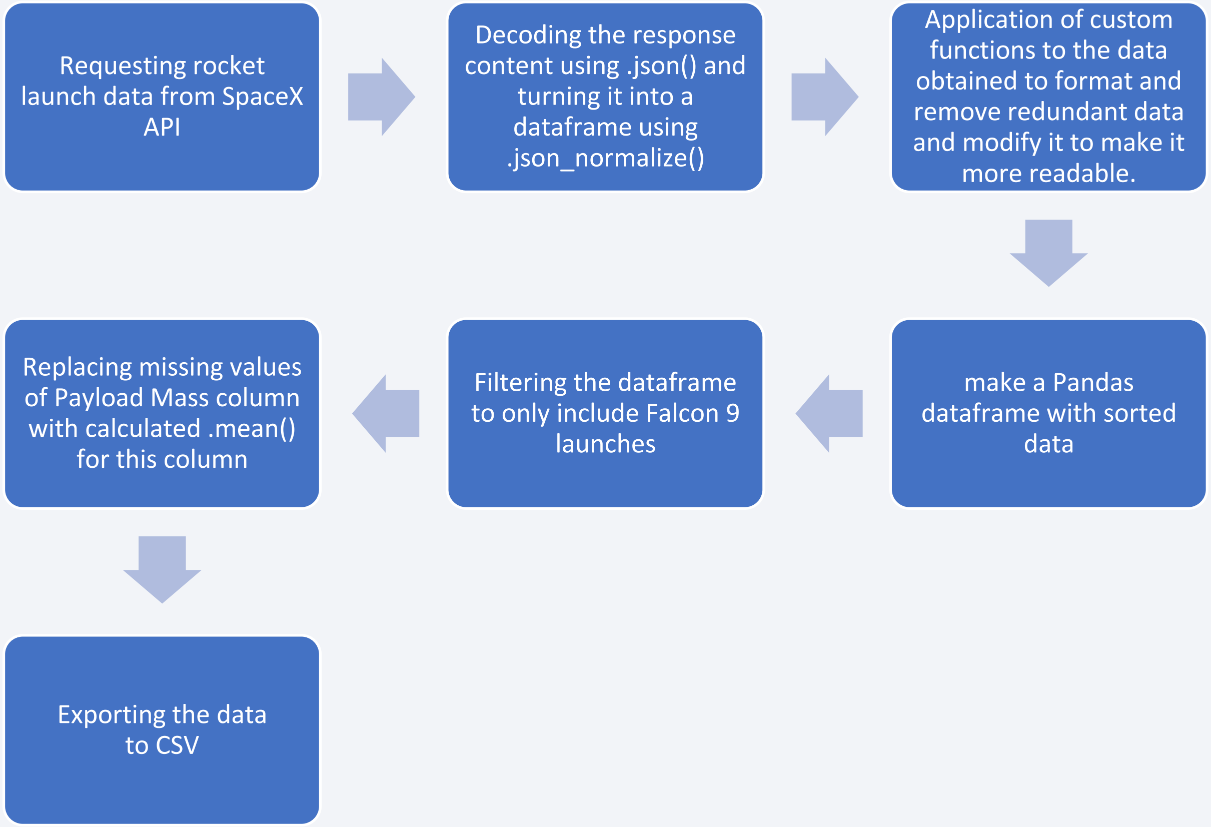
data had to be obtained from these two sources to generate a table with as much information as possible about Falcon 9 rocket launches and their derivatives.

- 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

- SpaceX offers a public REST API from which data and where data can be obtained.
- This API was used according to the flowchart

Data Collection Notebook



Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Application of custom functions to the data obtained to format and remove redundant data and modify it to make it more readable.

make a Pandas dataframe with sorted data

Filtering the dataframe to only include Falcon 9 launches

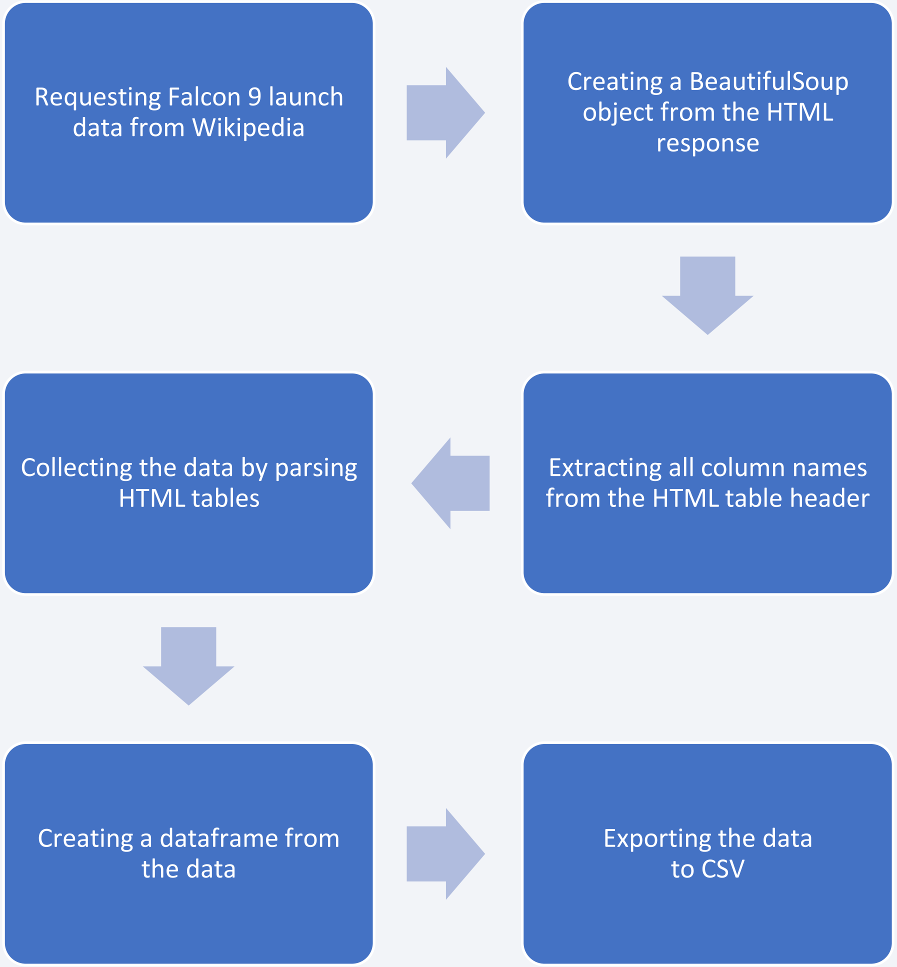
Replacing missing values of Payload Mass column with calculated .mean() for this column

Exporting the data to CSV

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia
- The data scraping was executed according to the flowchart

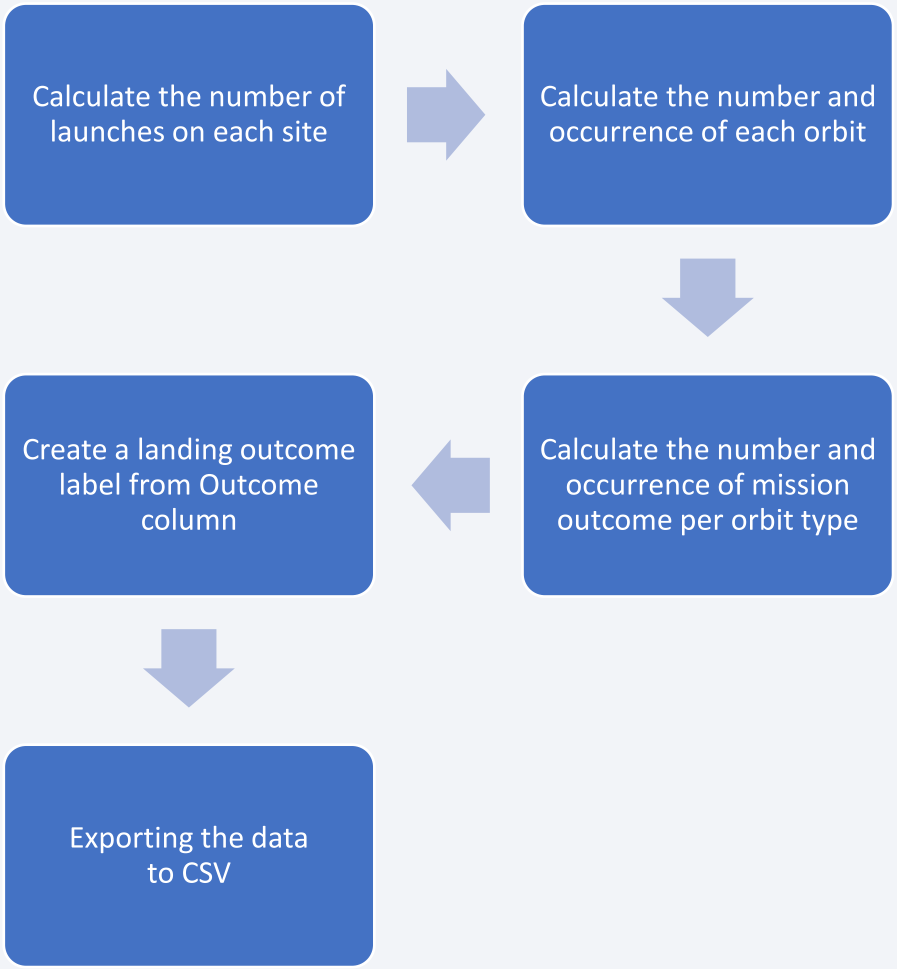
[Webscraping Notebook]


Requesting Falcon 9 launch data from Wikipedia
Creating a BeautifulSoup object from the HTML response
Extracting all column names from the HTML table header
Collecting the data by parsing HTML tables
Creating a dataframe from the data
Exporting the data to CSV

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- The summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Sometimes a landing was attempted but fail due to an accident, in the dataset the column called "Outcome" describe if the rocket was successfuly landing or not, and it was landing in the ocean, ground pad or ship, and this column is used to make a new column called "Class" was describe if the landing is a bad outcome or otherwise, tis column is used in the present work.

Data Wrangling Notebook



Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

EDA with Data Visualization

- To explore the data, , scatterplots and barplots were used to visualize the relationship between variables, looking for relationships and trends.
- 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).

Data Visualization Notebook

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;
- 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

SQL Notebook

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.

Interactive Map Notebook

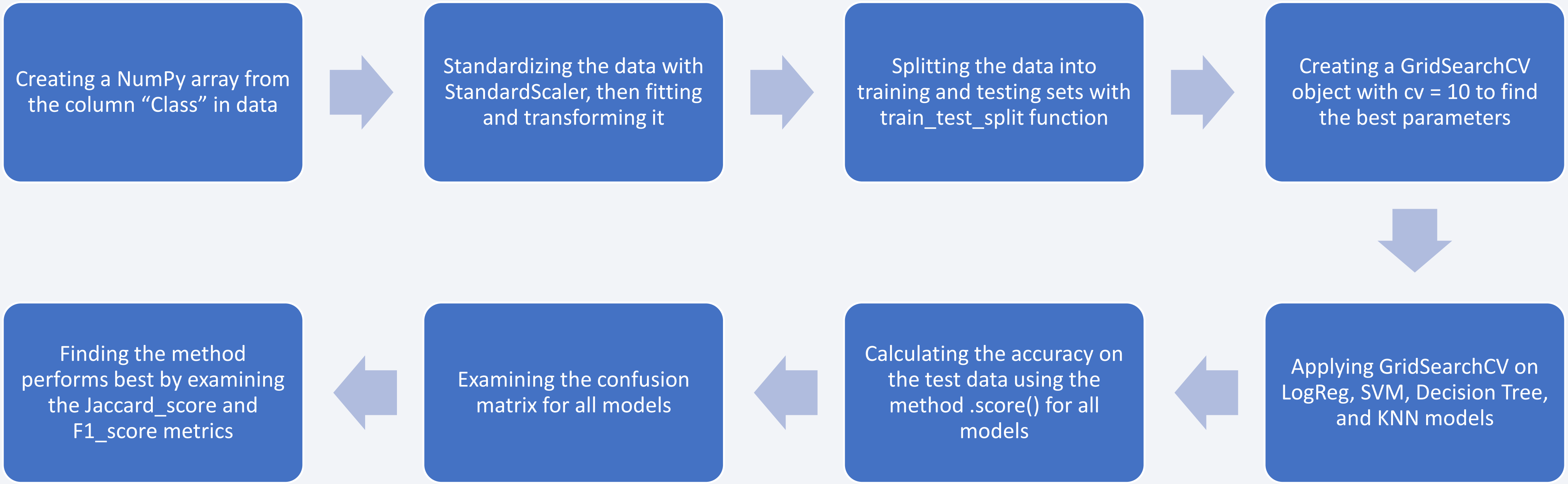
Build a Dashboard with Plotly Dash

- 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.

Interactive Dashboard Notebook

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

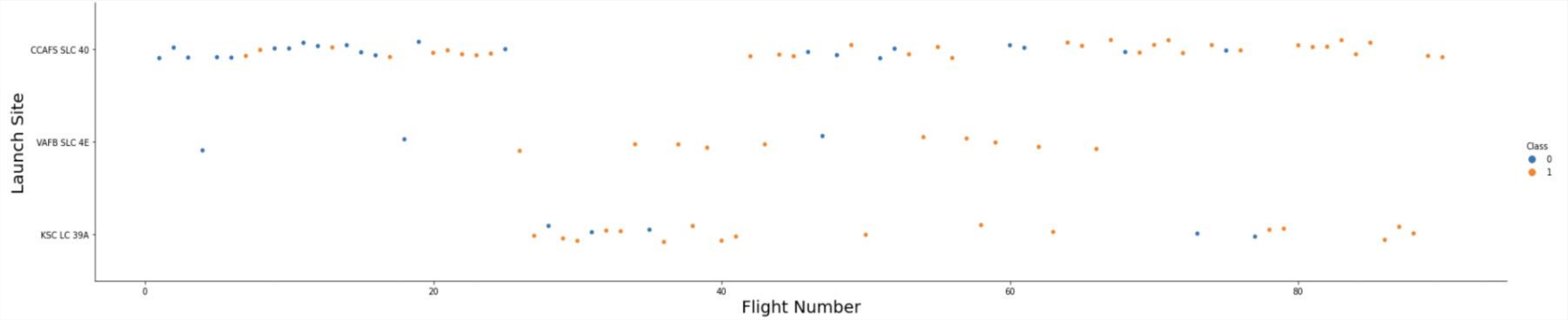
Results

- Exploratory data analysis 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;
  - The number of landing outcomes became as better as years passe;
- 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.
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%

Section 2

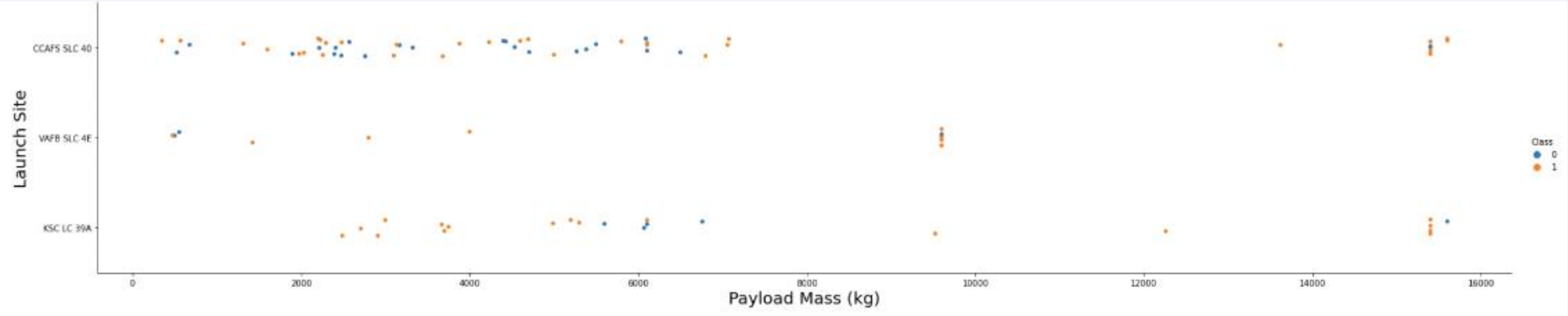
Insights drawn from EDA

Flight Number vs. Launch Site



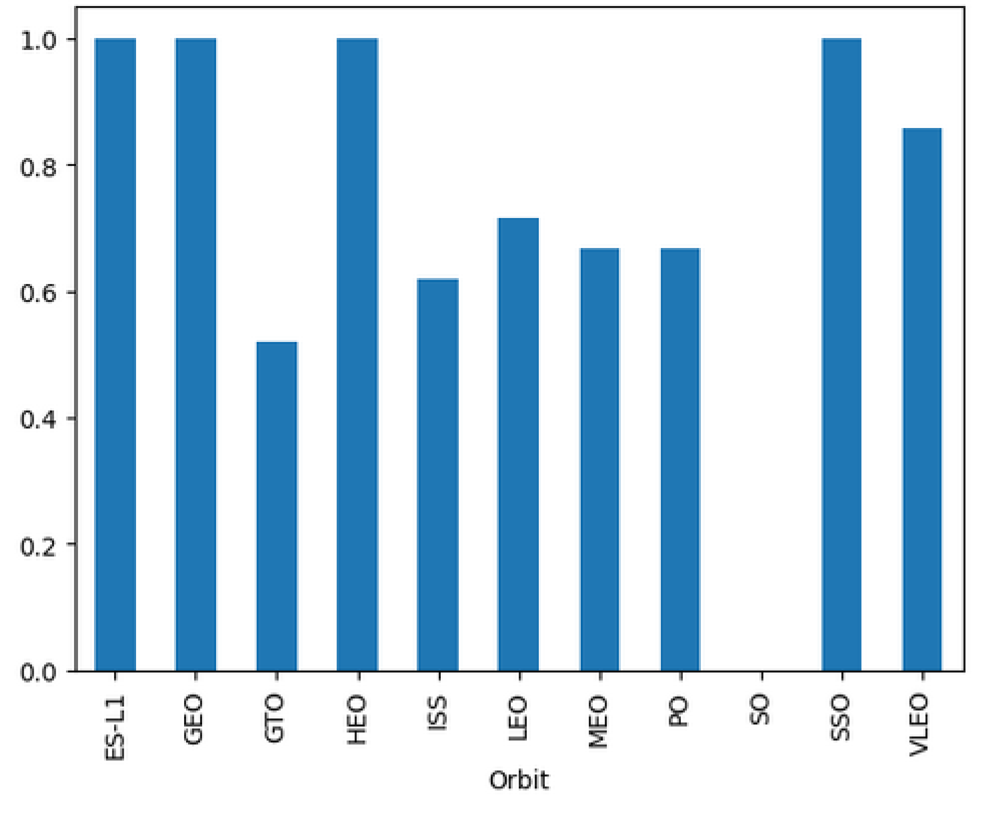
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



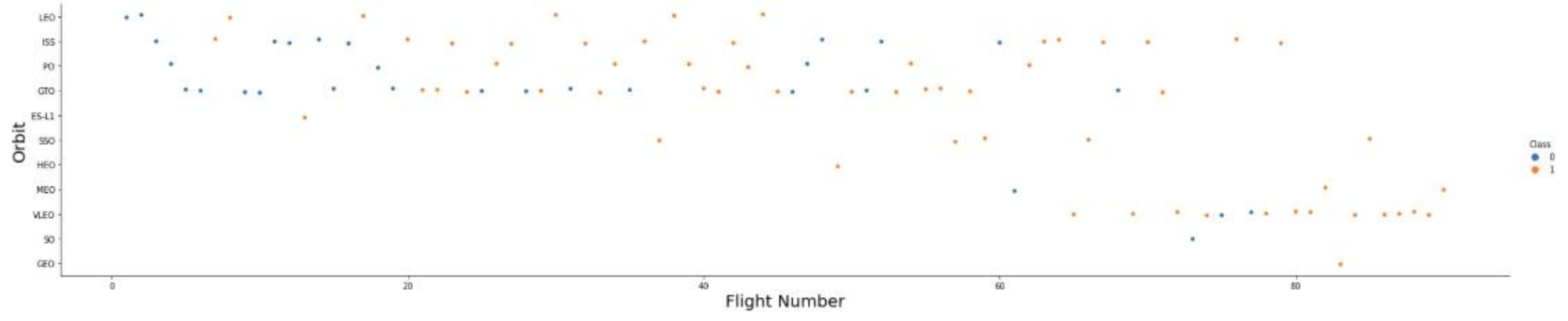
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



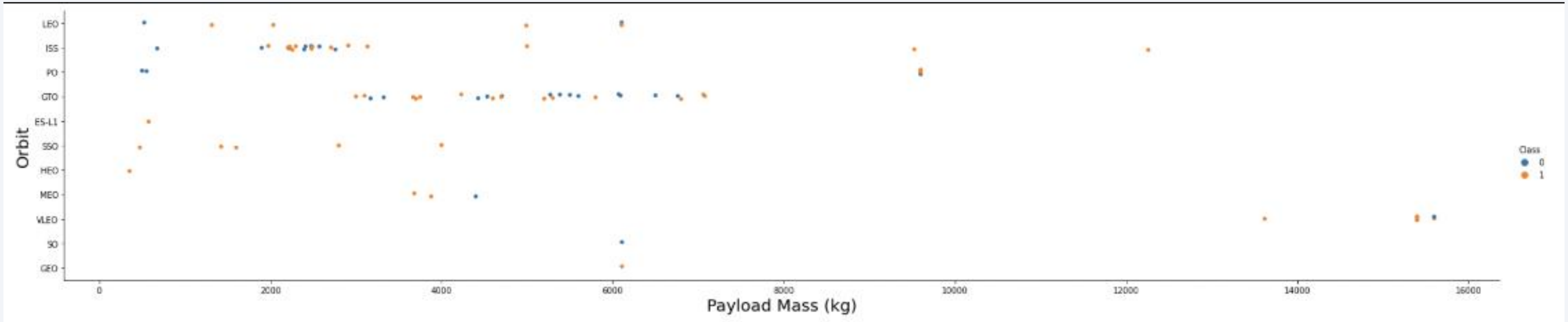
- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO, VLEO

Flight Number vs. Orbit Type



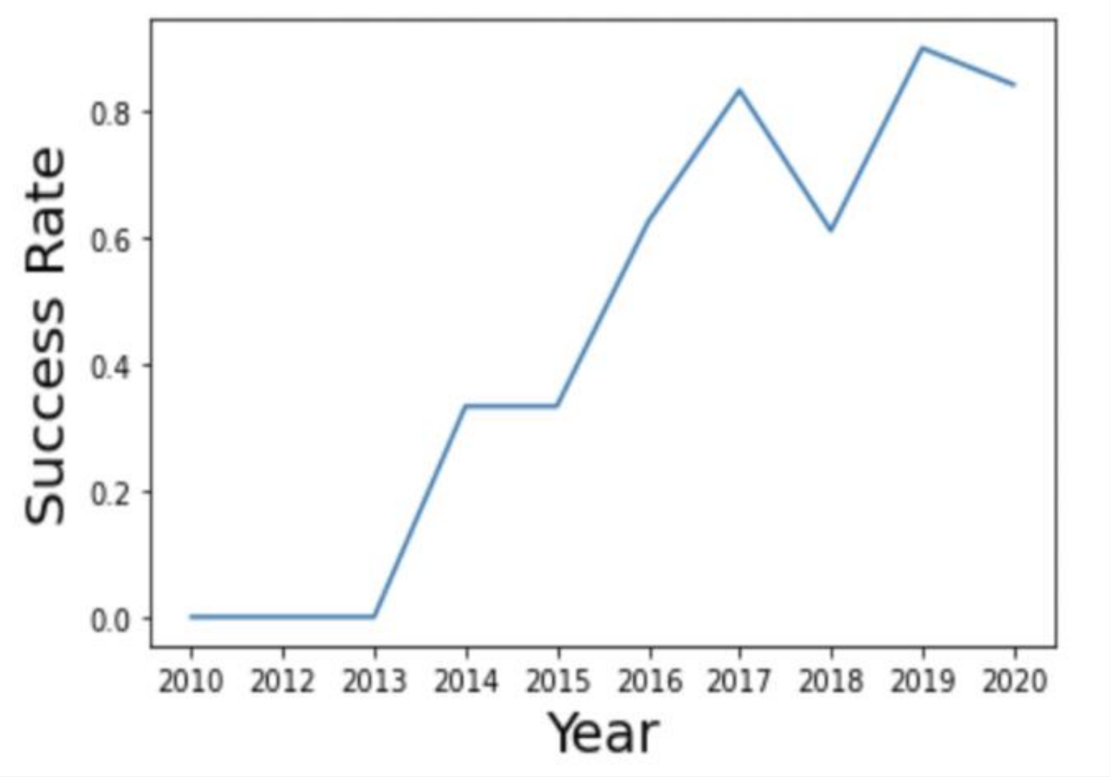
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

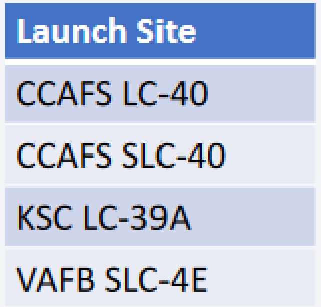
Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020.

All Launch Site Names

- According to data, there are four launch sites:

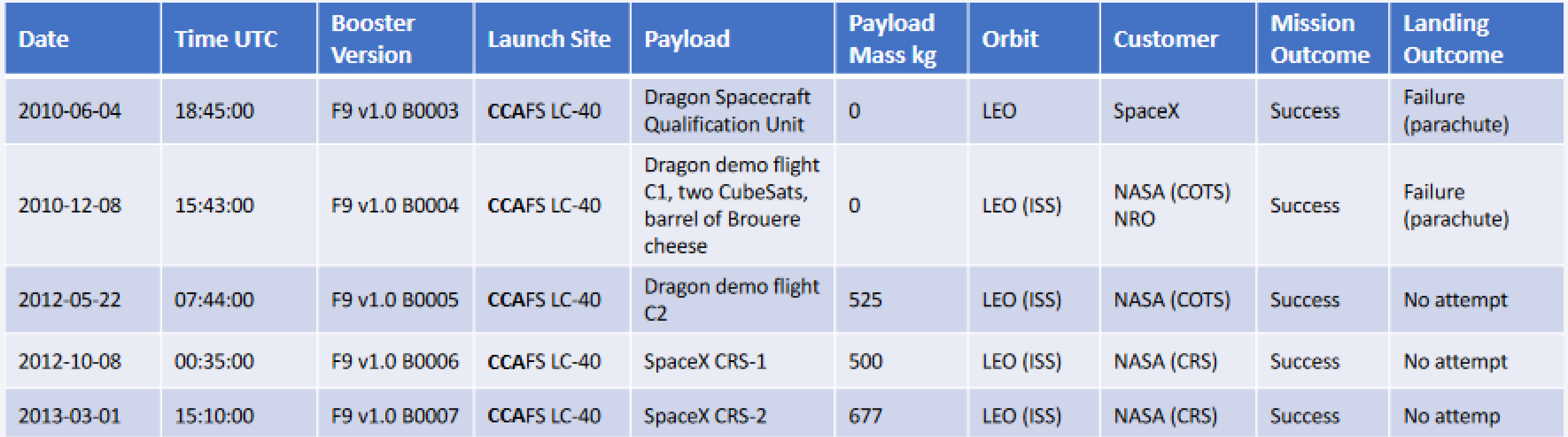


Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

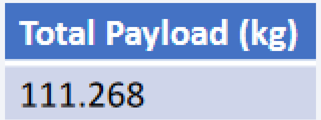
- 5 records where launch sites begin with `CCA` (Cape Canaveral Air):



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 attemp

Total Payload Mass

- The total payload carried by boosters from NASA:

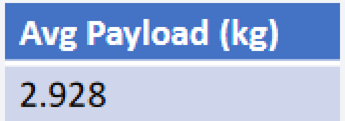


Total Payload (kg)
111.268

Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA

Average Payload Mass by F9 v1.1

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

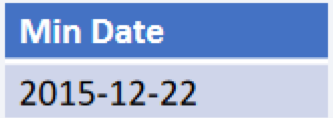


Avg Payload (kg)
2.928

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:

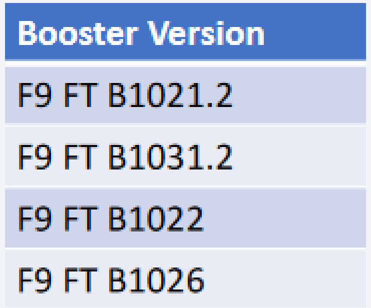


Min Date
2015-12-22

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

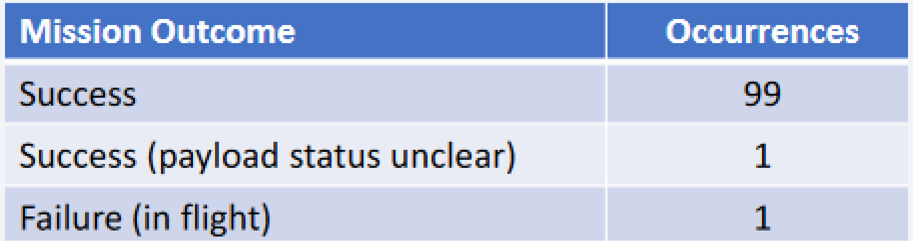


Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

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

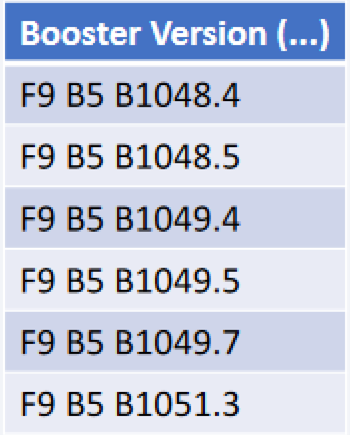


Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

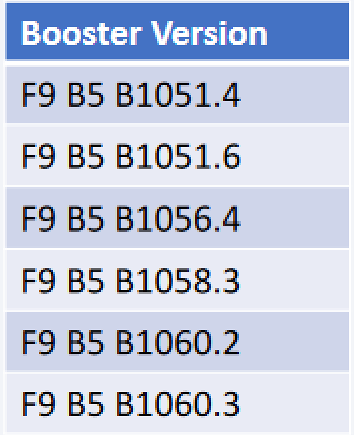
Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Booster 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

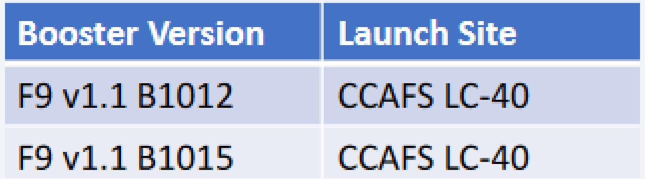


Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

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

2015 Launch Records

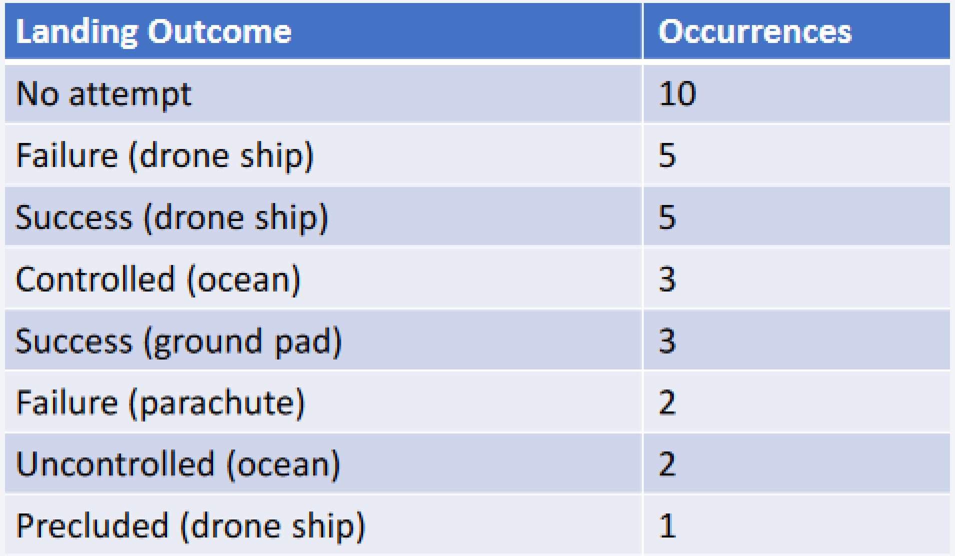
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

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

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



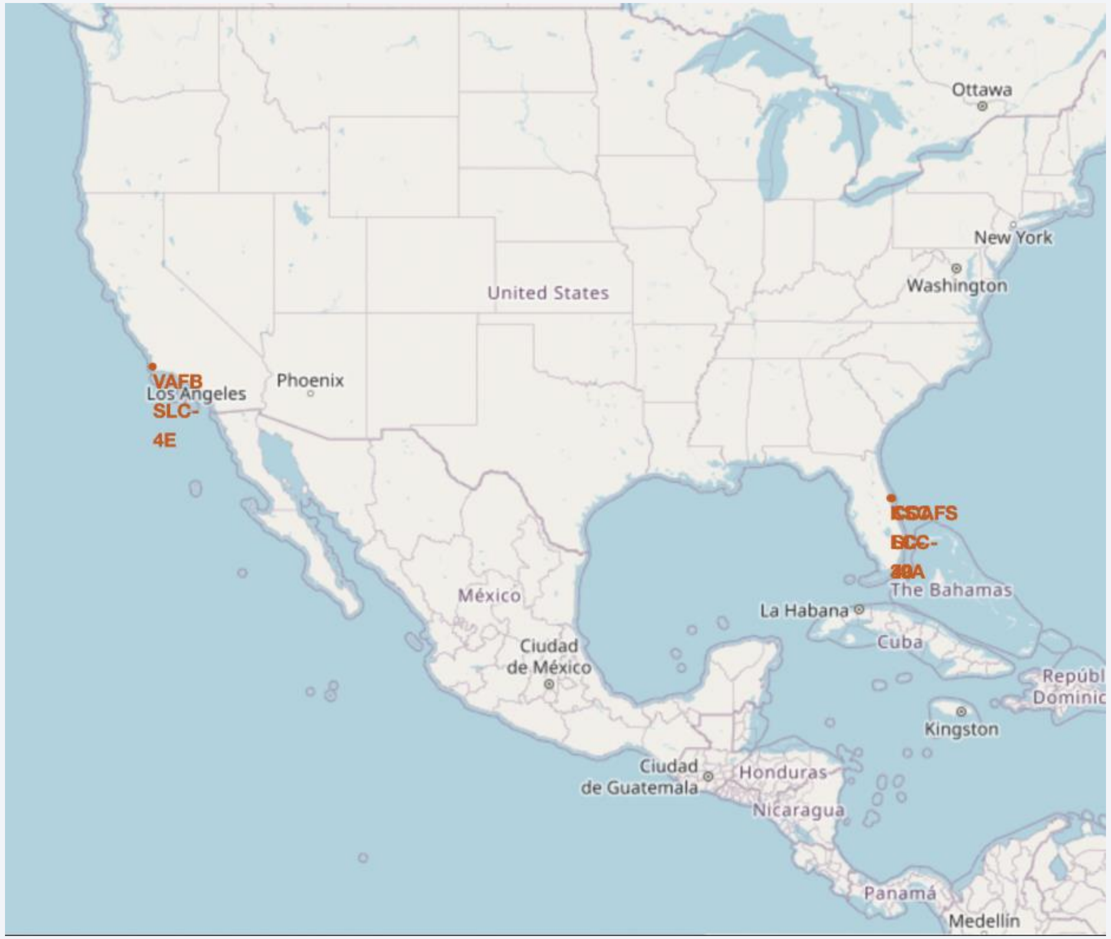
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

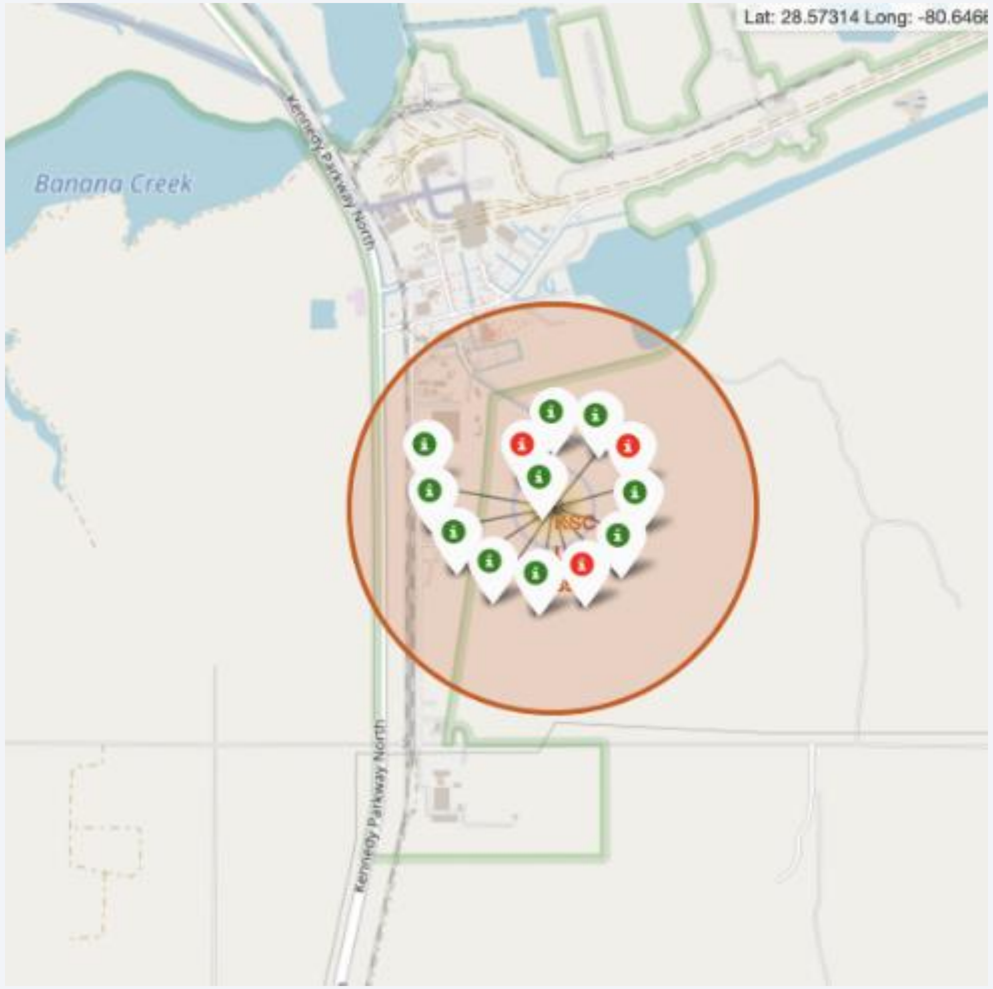
All Launch Sites

- 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.



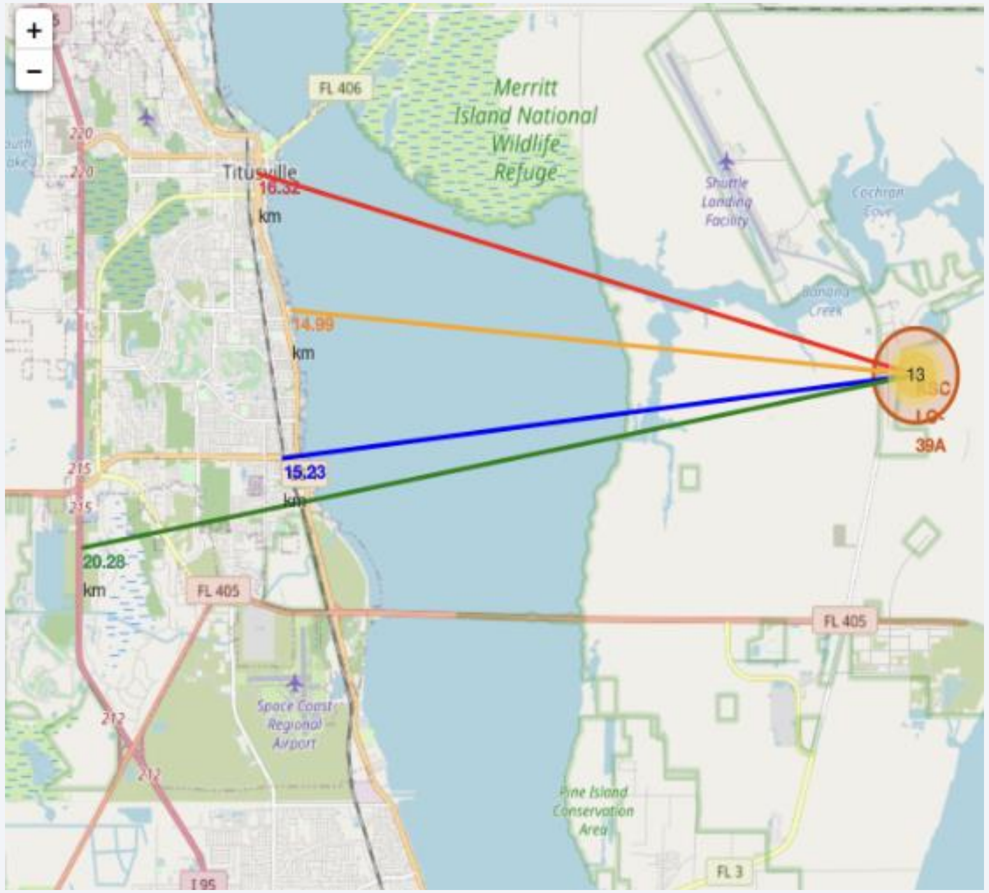
Launch Outcomes by Site

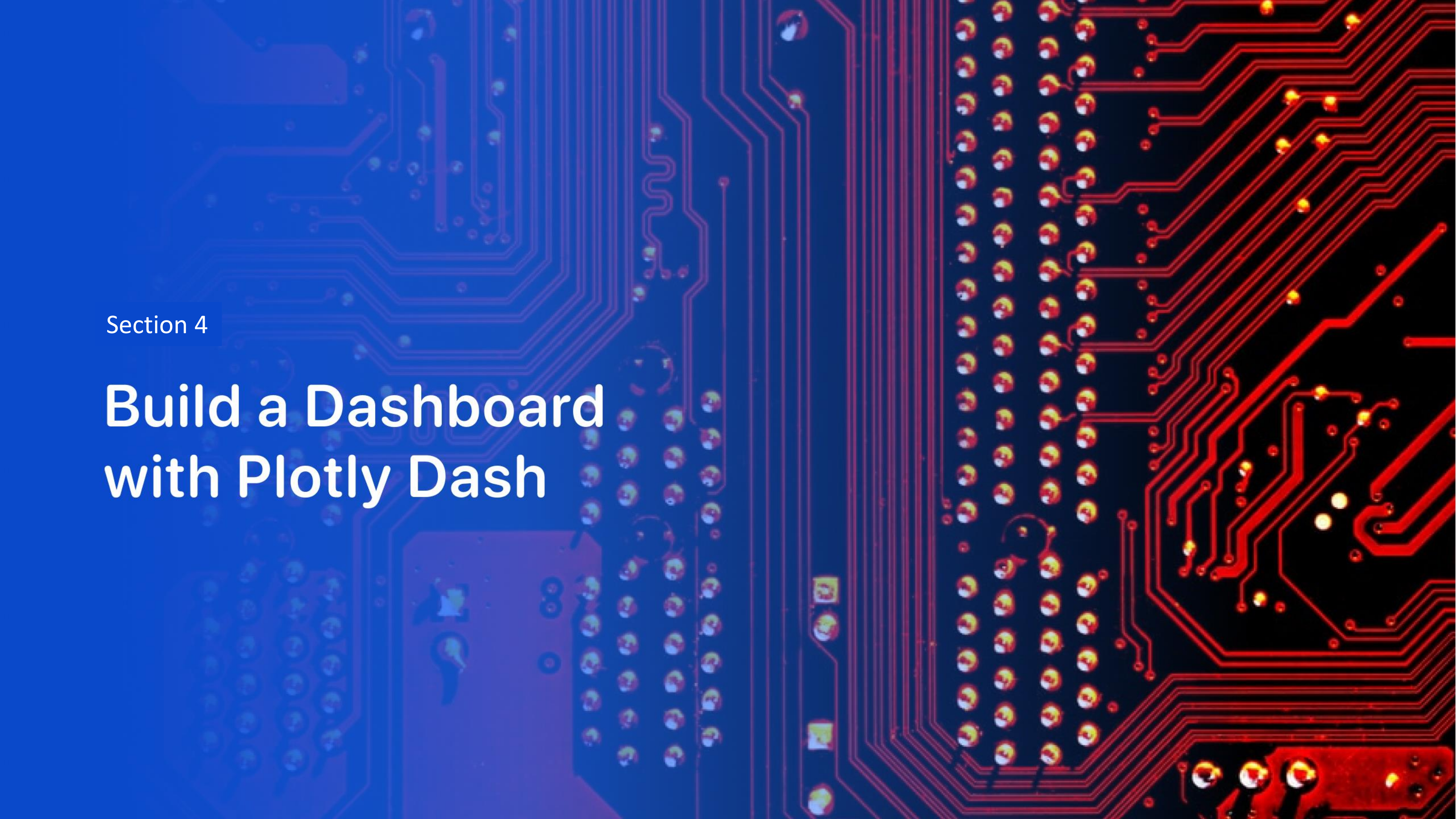
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- •Launch Site KSC LC-39A has a very high Success Rate.



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

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 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.

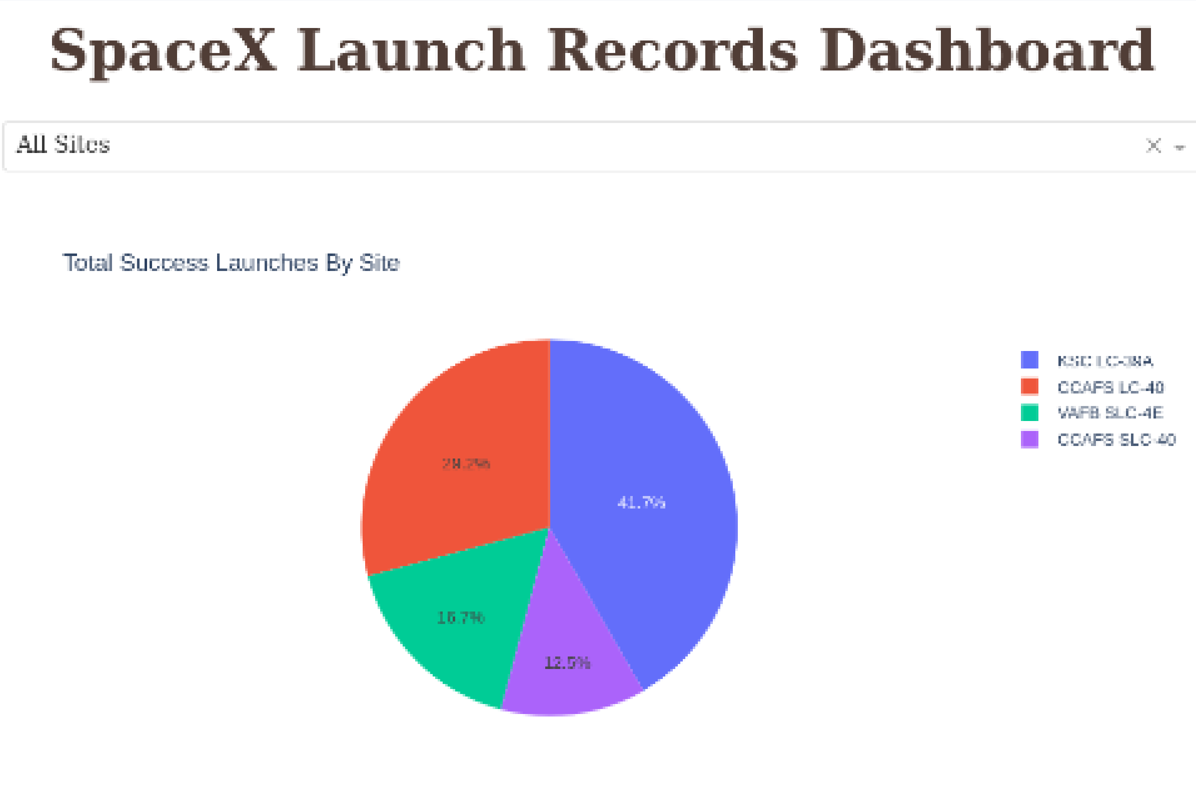




Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

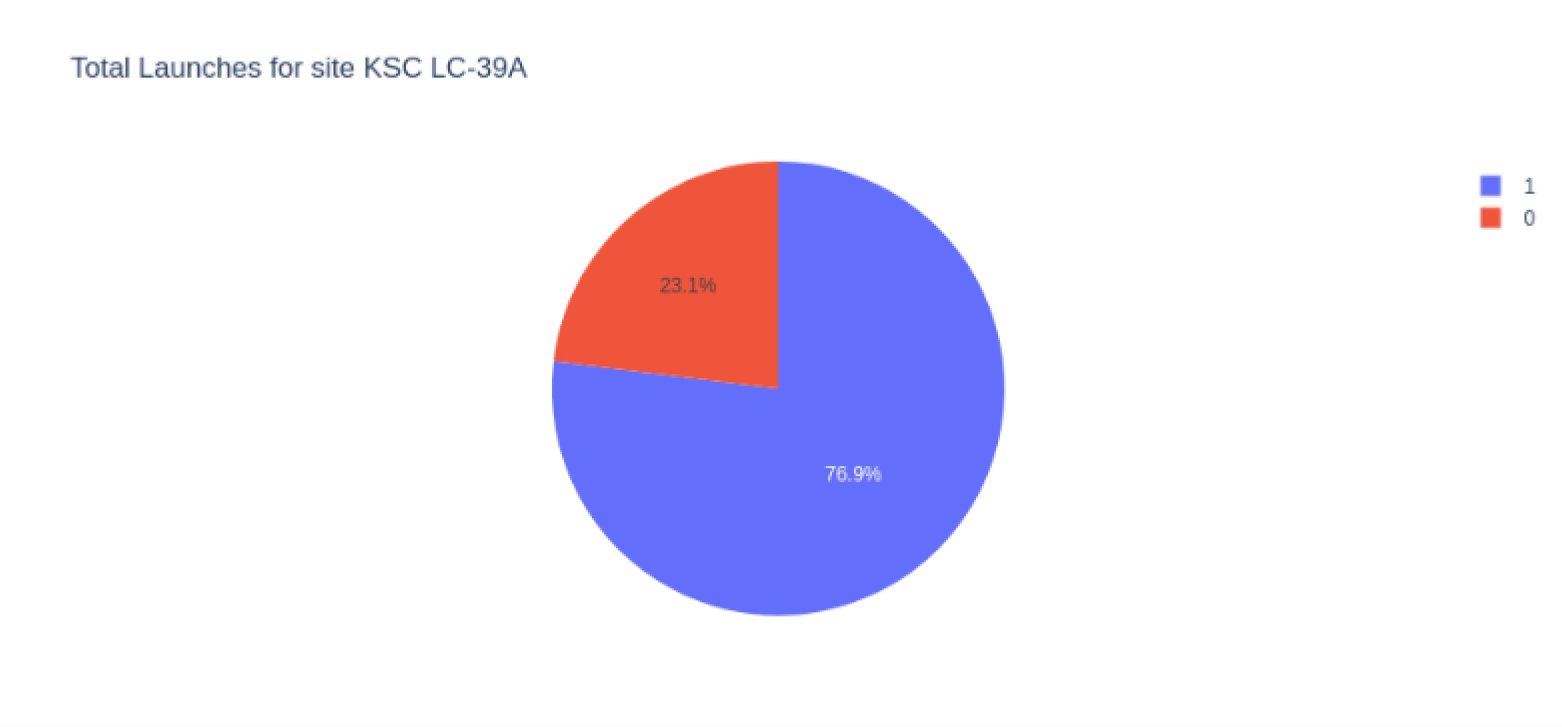


SpaceX Launch Records Dashboard

All Sites

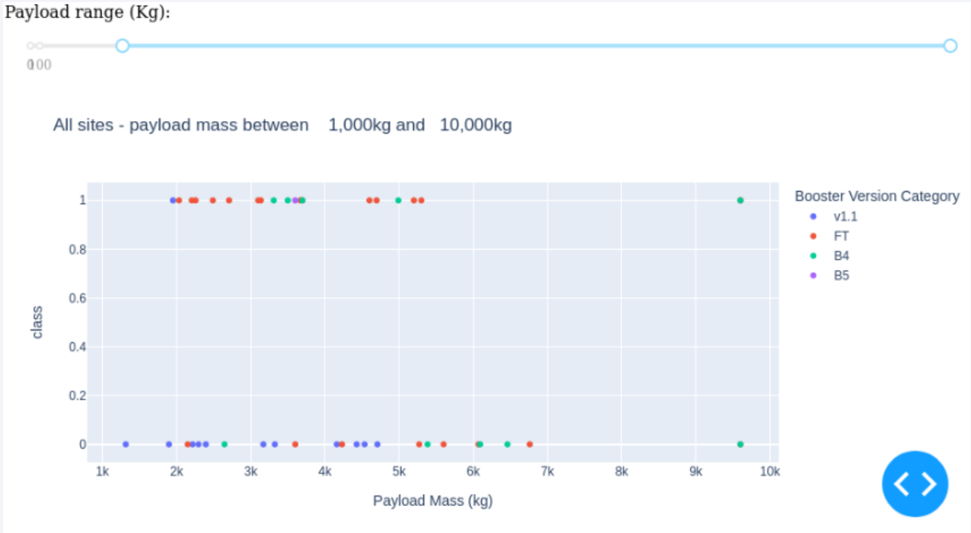
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



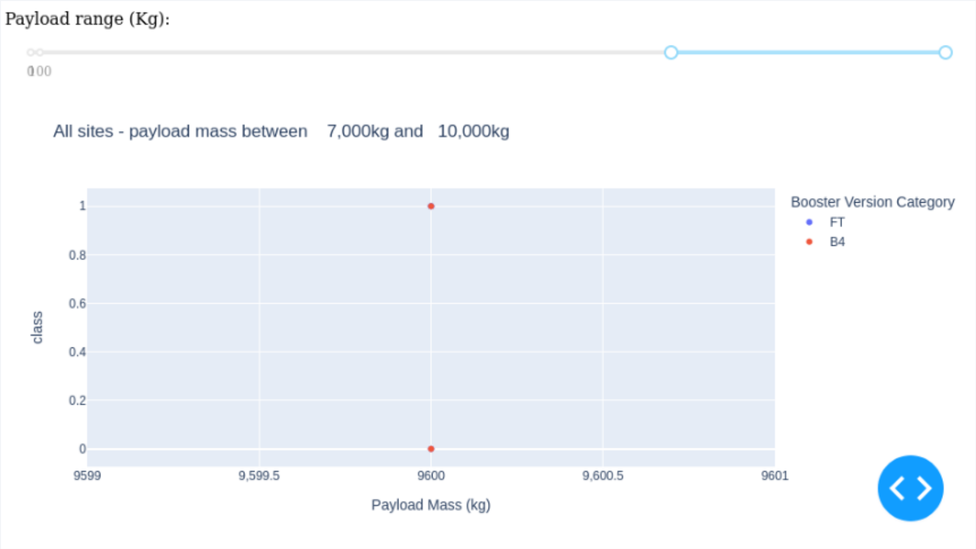
- 76.9% of launches are successful in this site

Payload vs. Launch Outcome



Payload range (Kg):

All sites - payload mass between 1,000kg and 10,000kg



Payload range (Kg):

All sites - payload mass between 7,000kg and 10,000kg

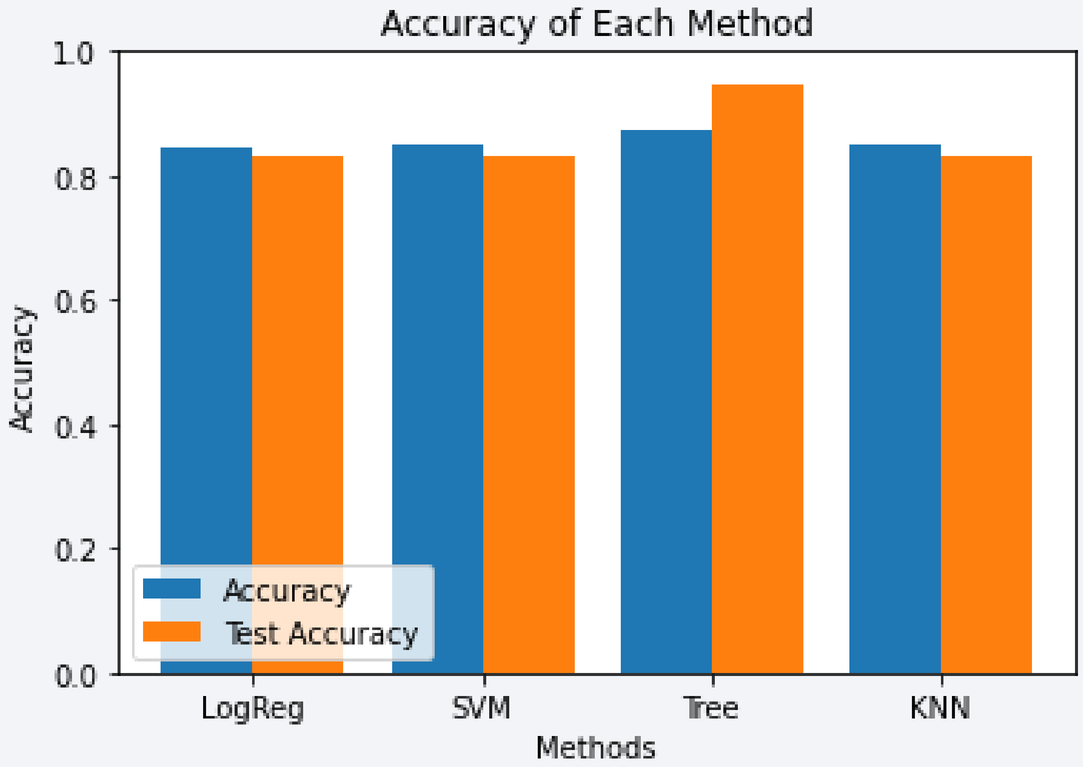
- Payloads under 6,000kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg

Section 5

Predictive Analysis (Classification)

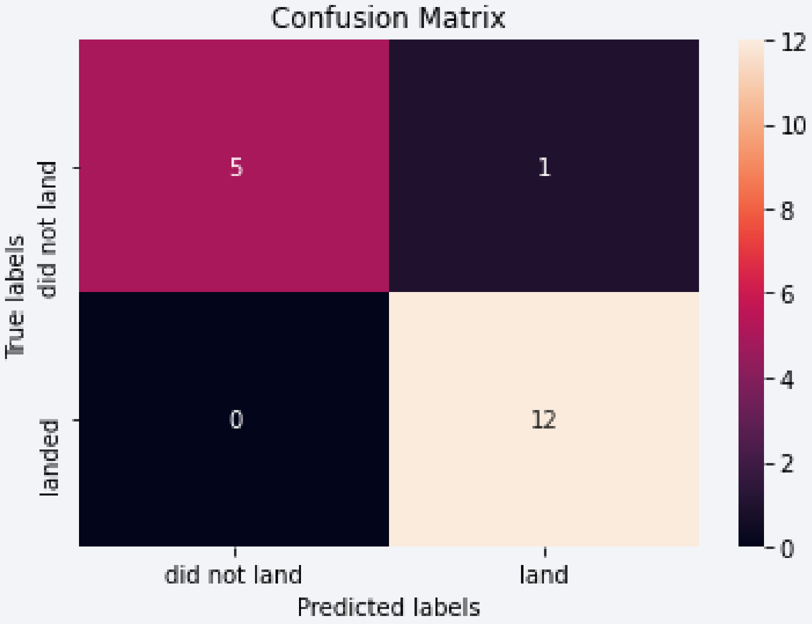
Classification Accuracy

- 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 87%.



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.



Conclusions

- Different data sources were analyzed, refining conclusions along the process.
- 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.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- The success rate of launches increases over the years.

Thank you!