IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jean-Philippe Savard

August 14th, 2023

Outline

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

Executive Summary

- Summary of the methodologies that were used for data analysis:
  - Data Collection using SpaceX API, WebScraping and SQL Queries
  - Exploratory Data Analysis(EDA); data wrangling and visualization, interactive visual analytics
  - Experimentations with multiple Machine Learning Algorithms for predictions
- Summary of all results
  - Data were collected from valuable and official sources
  - EDA allowed to identify which data features are a best fit to predict the launches success
  - Machine Learning Algorithms experimentations revealed which model would be the best in predicting launches success based on the collected data

Introduction

- The objective of the analysis is to evaluate if the company Space Y could compete with Space X.
- We try to find answers:
  - To estimate the cost of launches, predict successful landings of the first stage rockets and the best places to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API, multiple endpoints were used to collect data from https://api.spacexdata.com/v4/
    - /rockets/, /launchpads/, /payloads/, /cores/, /launches/past
  - WebScraping from:
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Data was analyzed to find patterns and determine the label for successful and unsuccessful landings.

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected data were used to evaluate the accuracy of different models. The data was split into training and testing sets and Machine Learning Algorithms were run with different parameters, trying to find the best hyperparameters for every models.

Data Collection

- SpaceX API:
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
  - https://api.spacexdata.com/v4/launches/past"
- WebScraping:
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- SpaceX exposes public API that can be used to obtain data. The API Documentation can be found at: https://docs.spacexdata.com/
- GitHub URL of the SpaceX API calls notebook: [SpaceX Data Collection API Notebook]()

SpaceX API
- Request data from endpoints

Response
- Data is returned in a JSON format as array of objects

Prepare data
- Filter necessary data to keep only the one related to Falcon 9 launches
- Deal with missing values

Data Collection - Scraping

- SpaceX launches data were obtained from wikipedia: [SpaceX Falcon 9 Launches]
- GitHub URL of the SpaceX WebScraping notebook: [SpaceX WebScraping Notebook]

SpaceX WikiPedia
- Request data from WikiPedia using WebScraping techniques

Response
- Extract relevant tables and columns/variables names and values

Prepare data
- Prepare data, filter necessary values and create DataFrame

Data Wrangling

- Some EDA was done to find pattern in the data and determine what would be the best features to uses for training and testing purposes.
- Preliminary evaluation of data and ; null values, object types, launch sites, become familiar with the different kind of orbit types.

- GitHub URL of the Data Wrangling notebook: [Data Wrangling Notebook]()

EDA

- Exploratory Data Analysis of SpaceX data and become familiar with the available information

Summary

- Launches success rate, orbit types

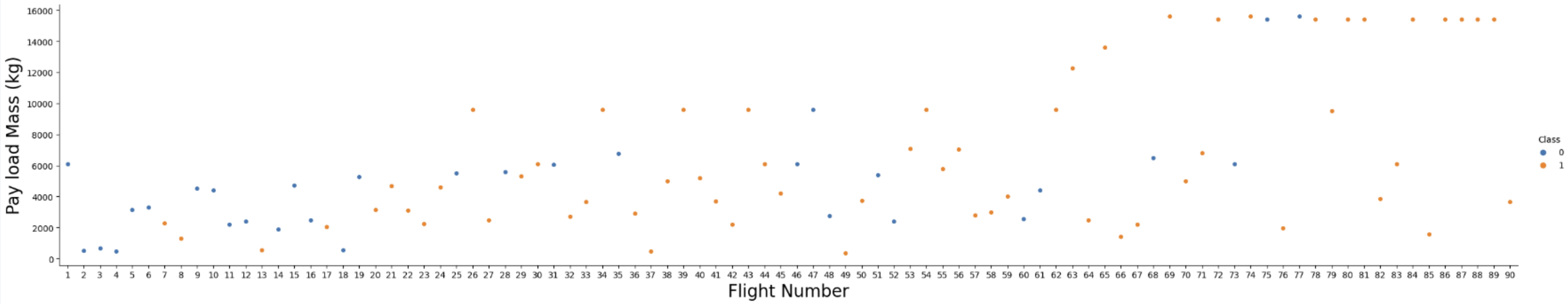
Prepare data

- Creation of ‘Class’ outcome label based on the available outcome information

EDA with Data Visualization

- Different types of charts were used to find relationship/correlation between the different variables; Flight Number VS Pay load Mass(kg), Flight Number VS Launch Site, Launch Site VS Payload Mass, Orbit VS Success Rate, Flight Number VS Orbit, Payload Mass VS Orbit,
- GitHub URL of the EDA Data Visualisation notebook: [EDA Data Visualisation Notebook]()

Exemple:



EDA with SQL

After creating a locale SQLite DB for EDA purpose, these queries were executed to find insight:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank 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.

GitHub URL of the EDA with SQL notebook: [EDA with SQL]

Build an Interactive Map with Folium

- A map was build, using Folium, to give a visual geospatial representation of the data. Different types of objects were used:
  - Markers: represent a point on a map with specific latitude/longitude. In our case, point represents launch sites.
  - Marker Clusters: Group of point on a map
  - Circles: an “circled” area that is highlighted around specific geo-coordinates.
  - Lines: define a line between two coordinates, in our case it was used to show the distance between two points.

GitHub URL of the Interactive Map with Folium notebook: [Launch Site Location]

Build a Dashboard with Plotly Dash

Two types of interactive graphs were used to visualize the data.

- Pie: success launches by site
- Scatter with range slider: to see the successful launches by site based on payload mass and booster version.

The combination of these two chart and slider allowed us to quickly find the relation between payload, launch site and booster version. That can also gives hint on the best place to launch based on these characteristics.

GitHub URL of the Dashboard with Plotly Dash Python file: SpaceX Dash

Predictive Analysis (Classification)

Four(4) Classification Models were used to compare their performances; Logistic Regression, Support Vector Machine(SVM), Decision Tree and K-Nearest Neighbors(KNN).

- The four models were trained and tested on the same set of training and testing data.

GitHub URL of the Predictive Analysis notebook: [ML Prediction]

Data

- Preparation of the data, standardization
- Creation of the training/testing sets

Testing

- Testing of the different Machine Learning Algorithms

Results

- Comparaison of the results obtained by the different models being evaluated

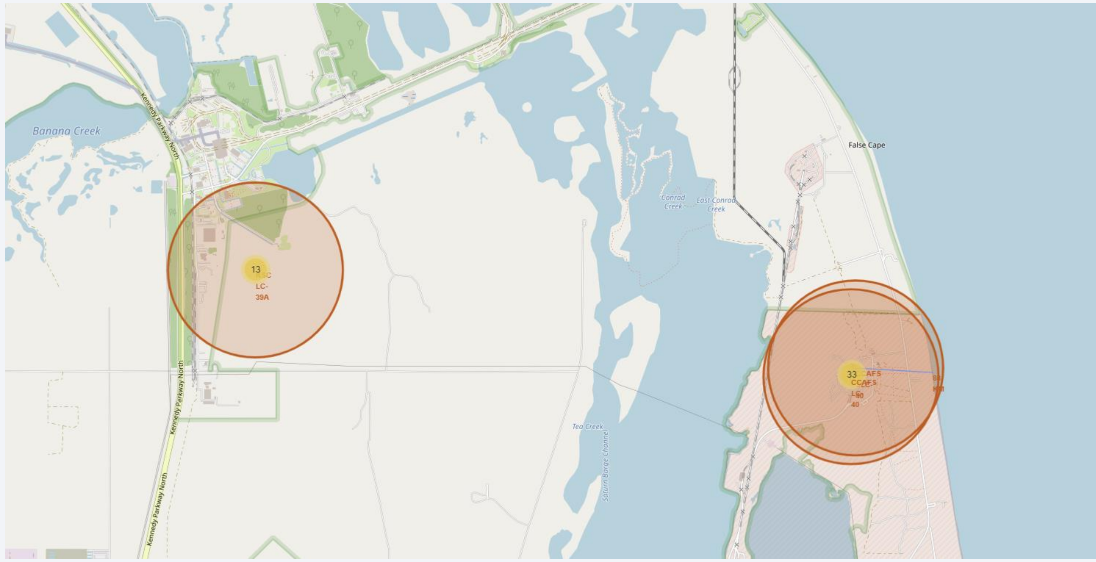
Results

- EDA results
  - There is 4 x different launch sites used by SpaceX.
  - The payload mass sum of boosters launched by NASA (CRS) is 45596 kg.
  - Average payload mass of F9 v1.1 booster is 2928.4 kg.
  - The first successful ground pad launch happened in 2015-12-22.
  - The boosters which have success in drone ship and have a payload mass between 4000 and 6000 kg are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2.
  - ~99% of mission outcomes were successful, with 100 x success and 1 x failure.
  - Booster version that carried the maximum payload mass of 15600 kg are: 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.
  - There were apparently 2 x failures in the year 2015 with booster versions F9 v1.1 B1012 and F9 v1.1 B1015.
  - Outcomes are more successful over time.

Results

Interactive analytics

Most of the launches were done on the east coast and from a distance of approximately 90 km from the sea. It could be considered as a safe place.



Results

Predictive analysis results

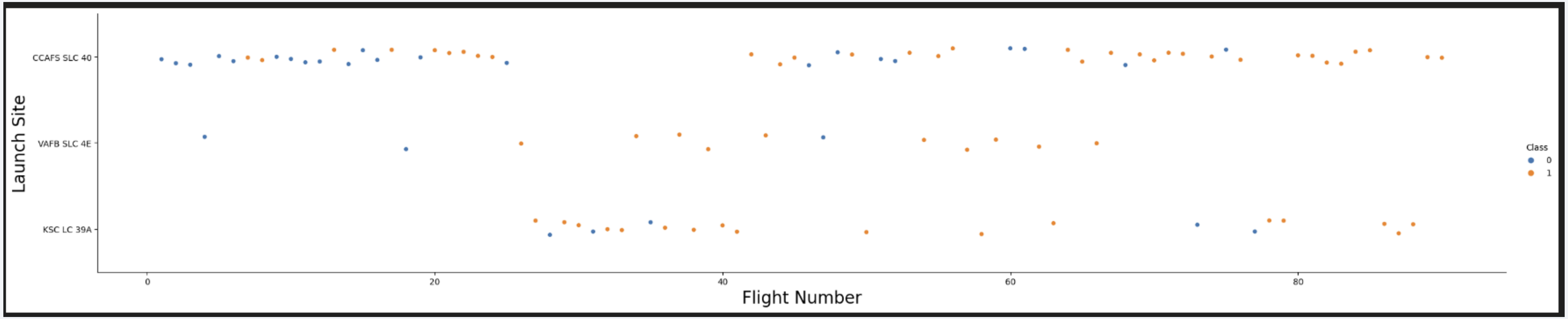
Experimentations with different Machine Learning Algorithms showed that Decision Tree Classification Model is the best in predicting successful landings with accuracy over 88%.

	method	score
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.888889
3	KNN	0.833333

Section 2

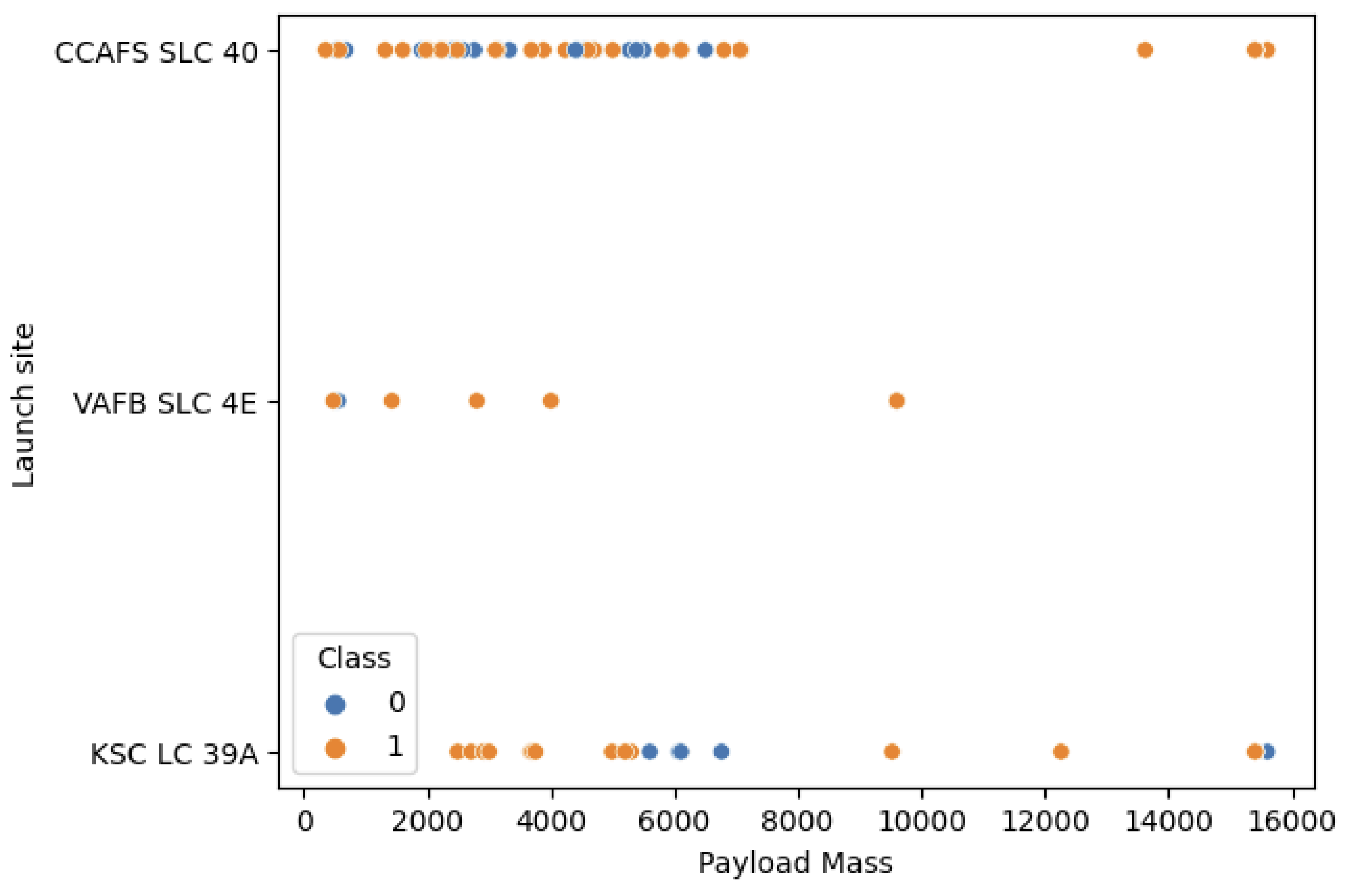
Insights drawn from EDA

Flight Number vs. Launch Site



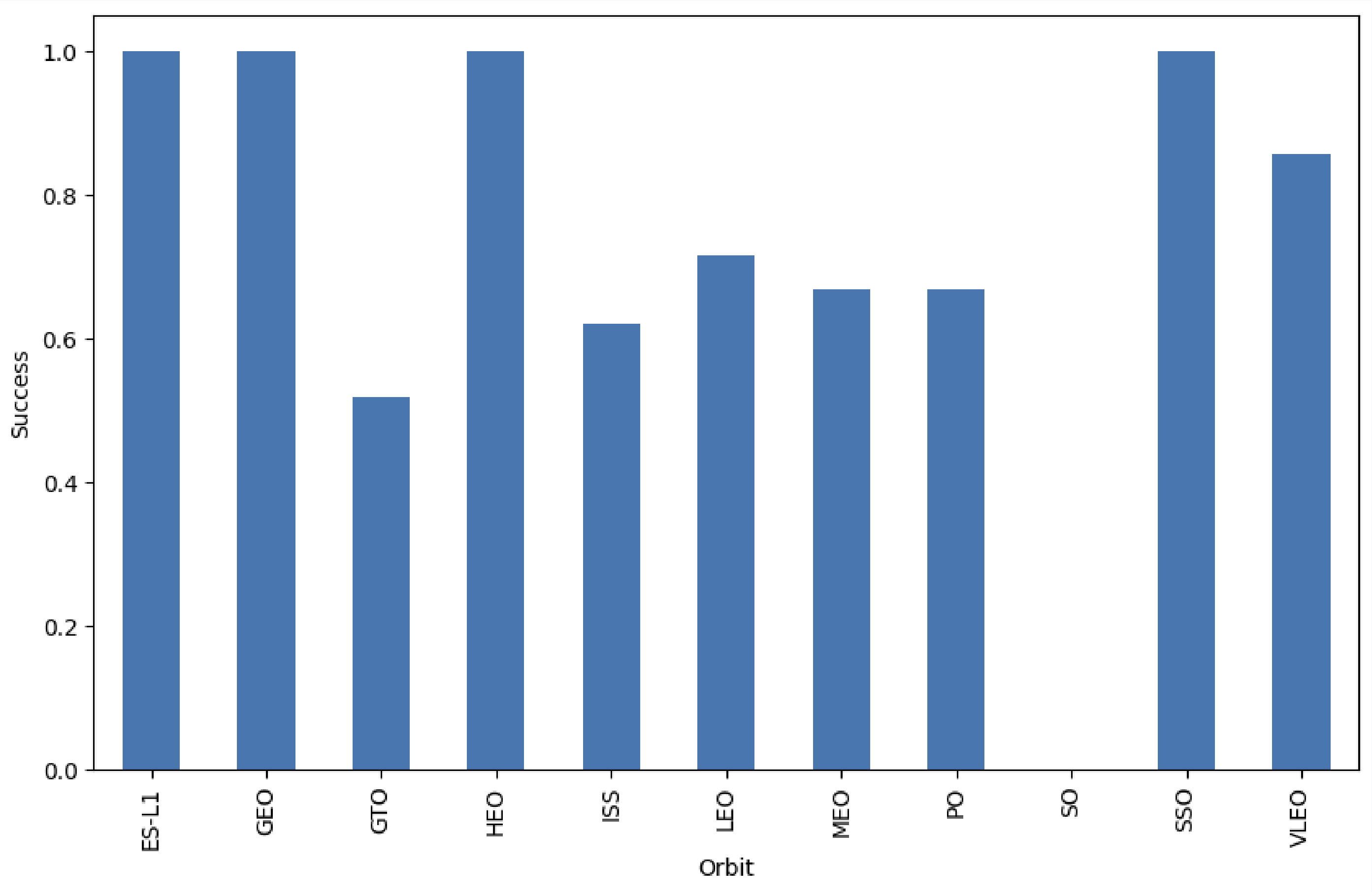
- Based on the graphic above, it looks like the site CCAFS SLC 40 is the best locations, since most of the recent launches were successful.
- Also, the plot seems to confirm that, when comparing the recent launches of all sites, the global success was getting better over time.

Payload vs. Launch Site



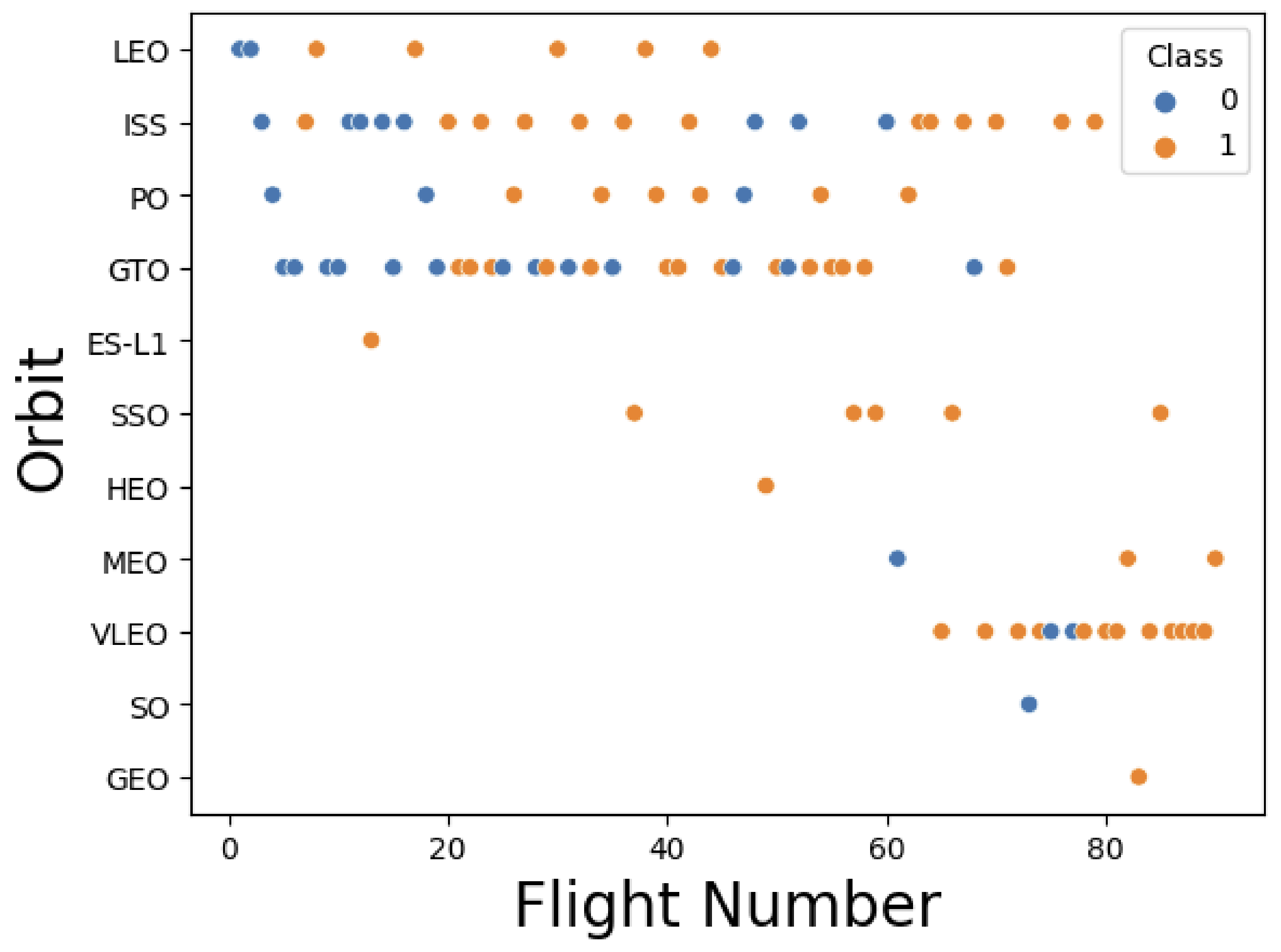
- The graphic show that launches with heavier payload have better success rate; over 9000 kg ar mostly successful.
- Most launches under 8000 kg were made on the location CCAFS SLC 40.
- CCAFS SLC 40 and KSC LC39A seems to be the only sites with launches over 10000 kg.

Success Rate vs. Orbit Type



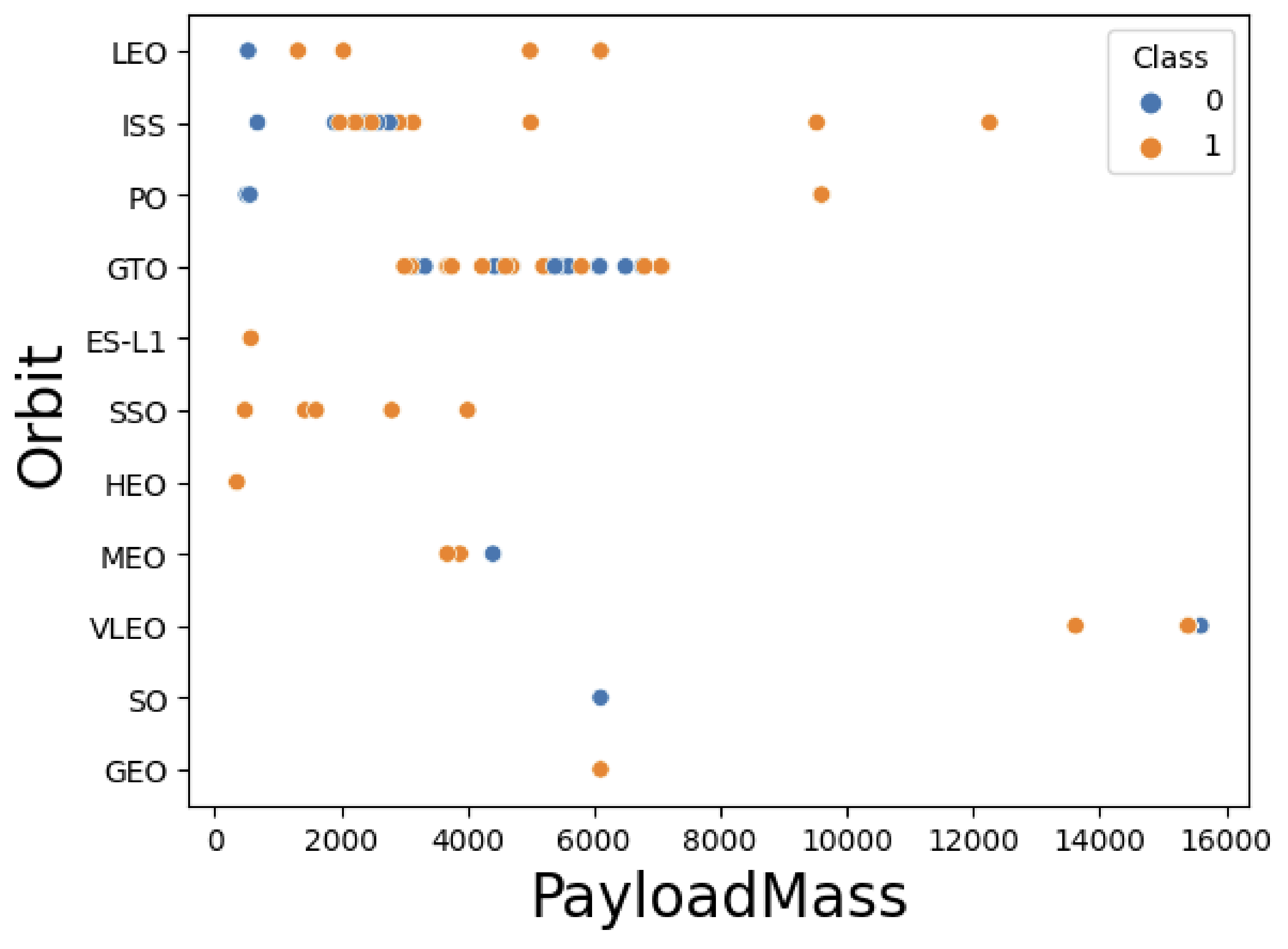
- Orbits with the best success rate, of 100%, are: EL-L1, GEO, HEO, SSO.
- As for the other orbits:
    - VLEO: ~82%
    - LEO: ~70%
    - MEO, PO: ~65%
    - ISS: ~60%
    - GTO: ~50%

Flight Number vs. Orbit Type



- Orbits SSO, HEO, MEO SO and GEO are not as frequent as the others, where ISS and GTO are most used and the use of VLEO seems to increase over time.

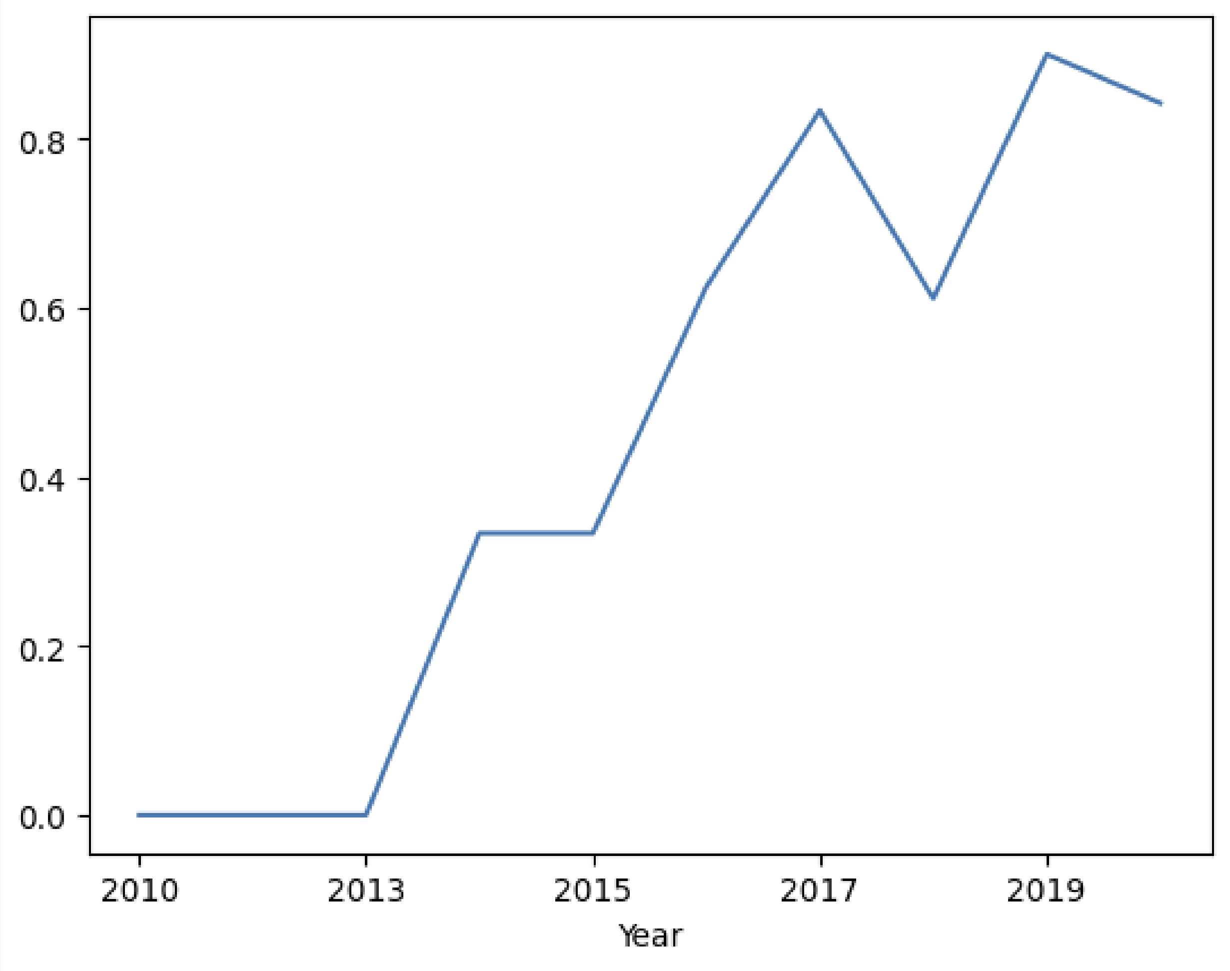
Payload vs. Orbit Type



- There do not seem to be a correlation between Payload Mass and success rate to Orbits.

Launch Success Yearly Trend

- Success rate keep getting better over time, since 2013, with a small drop in 2018.
- The better outcomes could be due to adjustments in launches calculation algorithms.



All Launch Site Names

- Based on the extracted data, there is 4 x launch sites.

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Sample of 5 x launch sites where the name begins with ‘CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Sum of the total payload carried by boosters from NASA, calculated using the SQL SUM Method.

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload carried by booster version F9 v1.1, calculated using the SQL AVG Method.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad. The results were filtered out to find the minimum date value.

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. Filtered out by finding the unique values.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes, calculated by grouping the mission outcomes.

success	failure
100	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass. Was filtered out on the value of 15600 kg.

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.

substr(Date, 6, 2)	substr(Date, 0, 5)	Booster_Version	Landing_Outcome	Launch_Site
10	2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
11	2015	F9 v1.1 B1013	Controlled (ocean)	CCAFS LC-40
02	2015	F9 v1.1 B1014	No attempt	CCAFS LC-40
04	2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40
04	2015	F9 v1.1 B1016	No attempt	CCAFS LC-40
06	2015	F9 v1.1 B1018	Precluded (drone ship)	CCAFS LC-40
12	2015	F9 FT B1019	Success (ground pad)	CCAFS LC-40

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

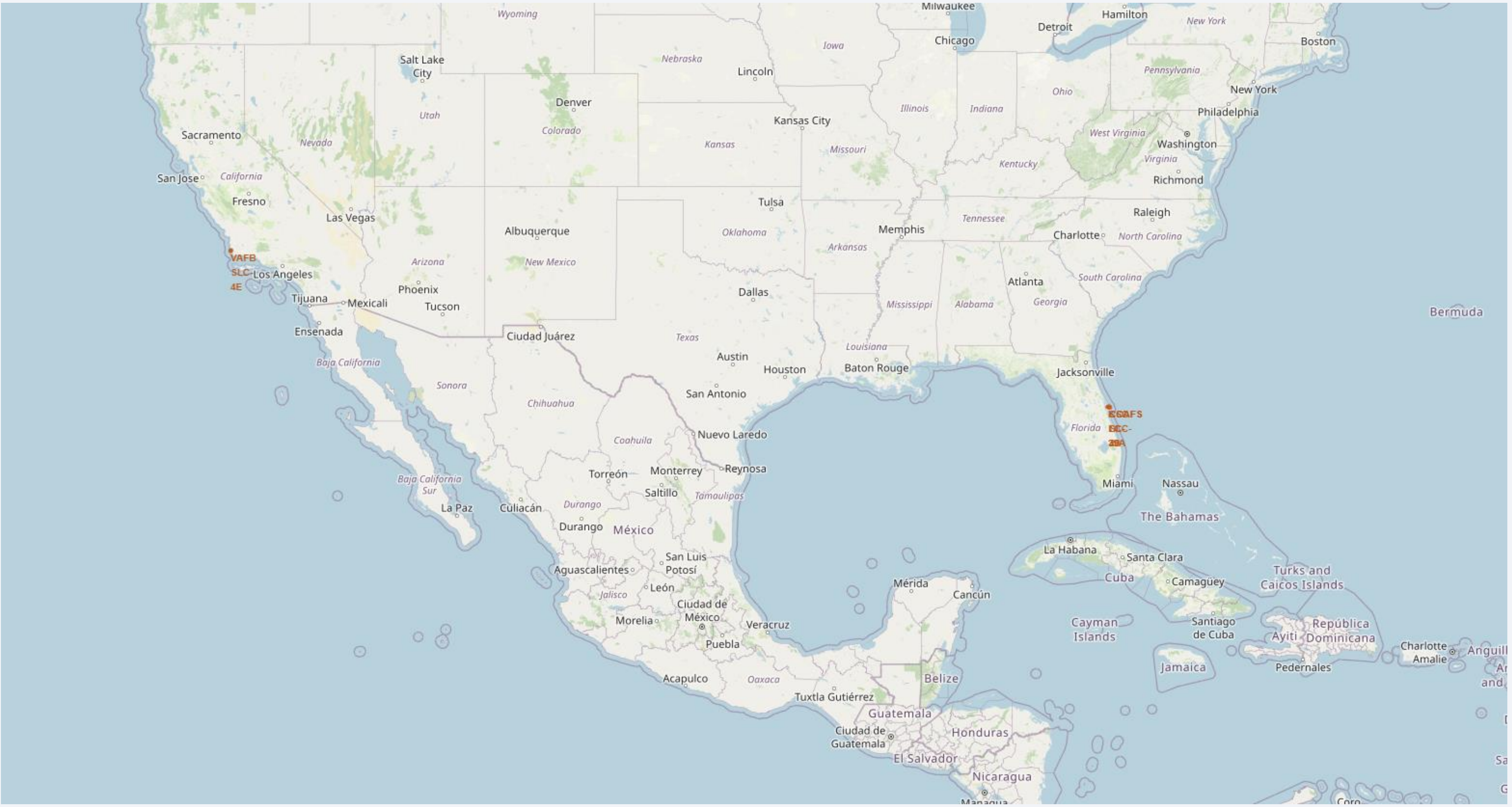
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

success	failure
5	1

Section 3

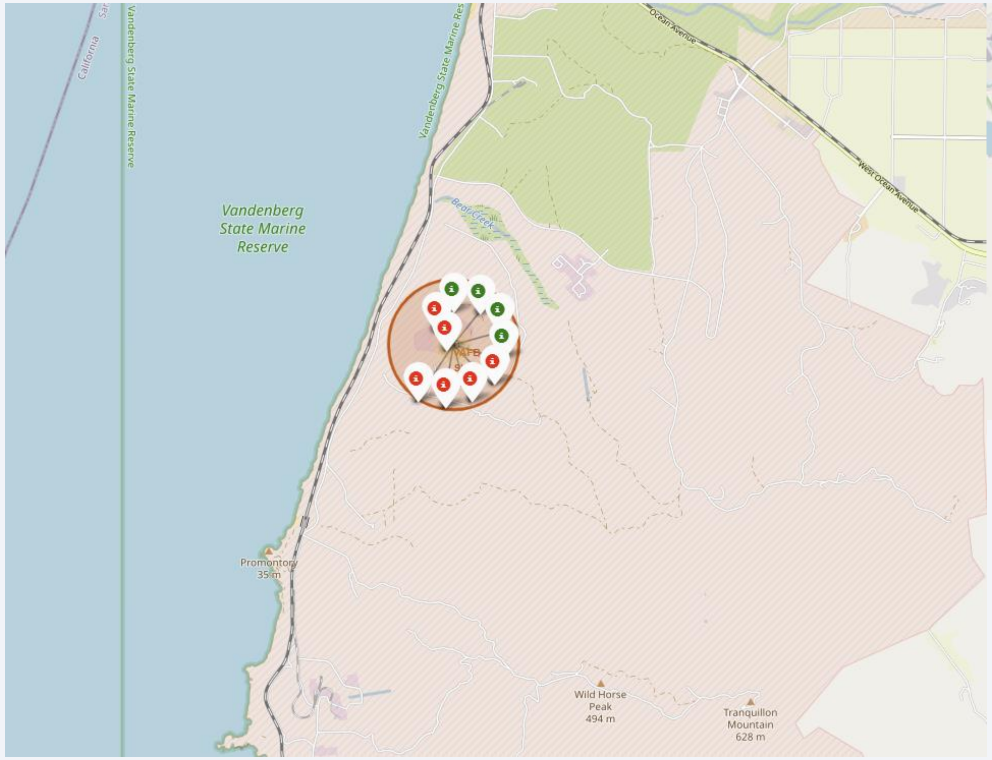
Launch Sites Proximities Analysis

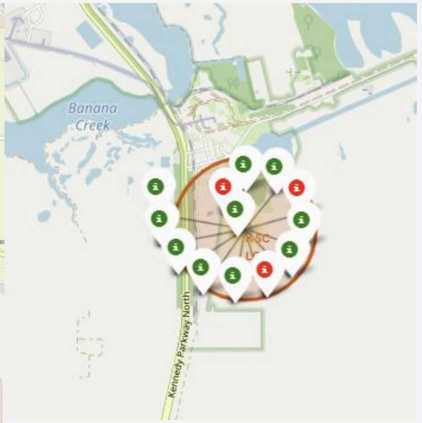
All launch locations

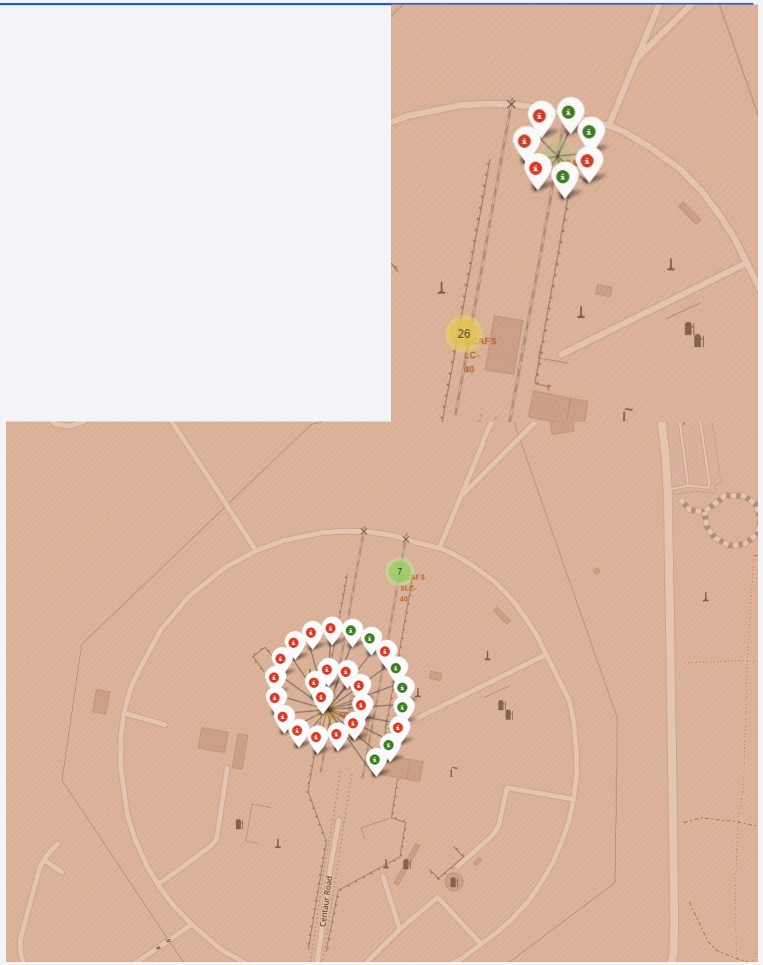


All launch sites looks to have been chosen near the coastline, potentially for safety reasons.

Launch outcomes by locations







All launches outcomes by sites, where green markers indicate success and red indicate failure.

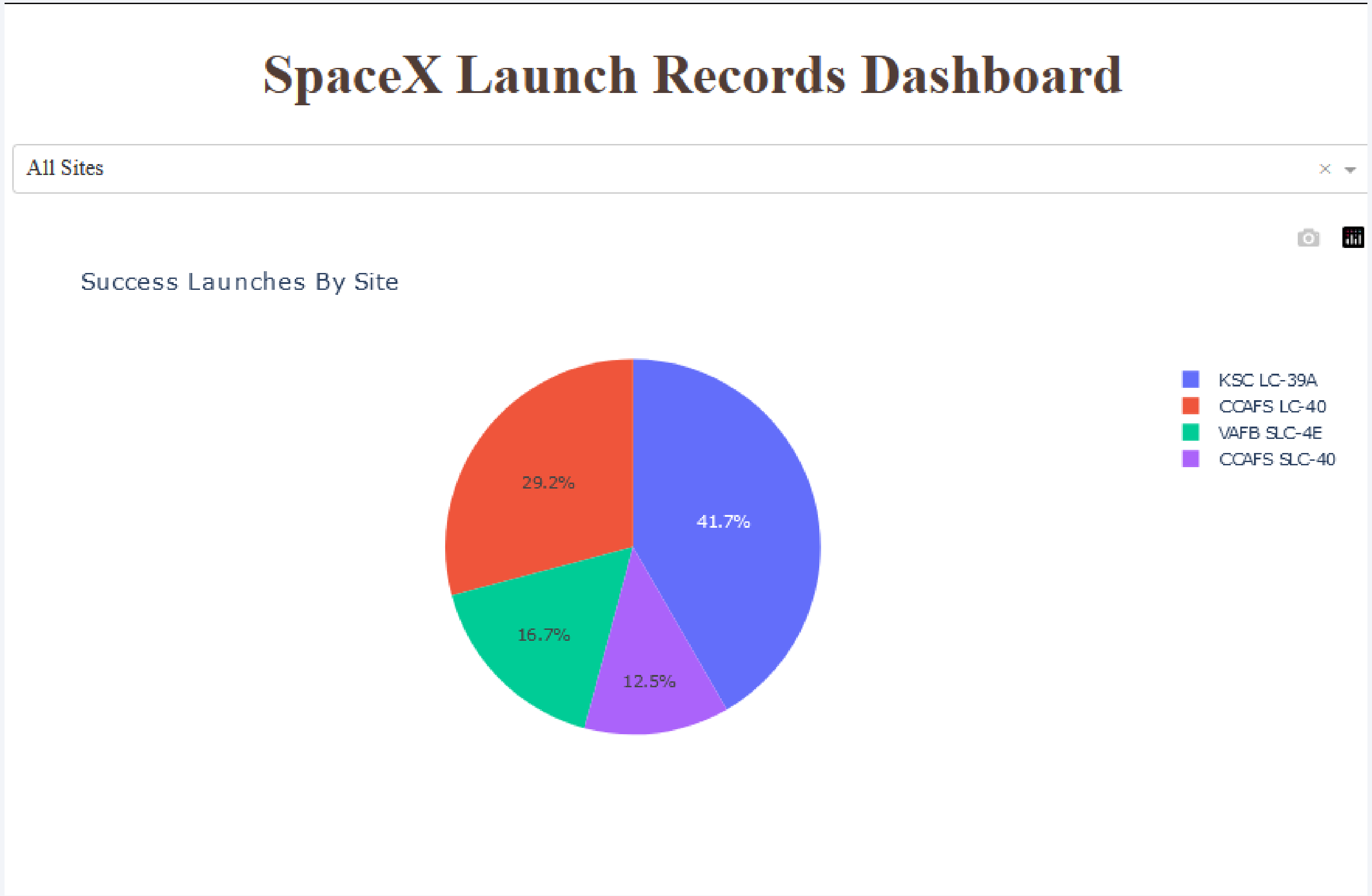
Launch locations Safety

Most sites look to be within 90 km of the coast line and far from inhabited areas.

Section 4

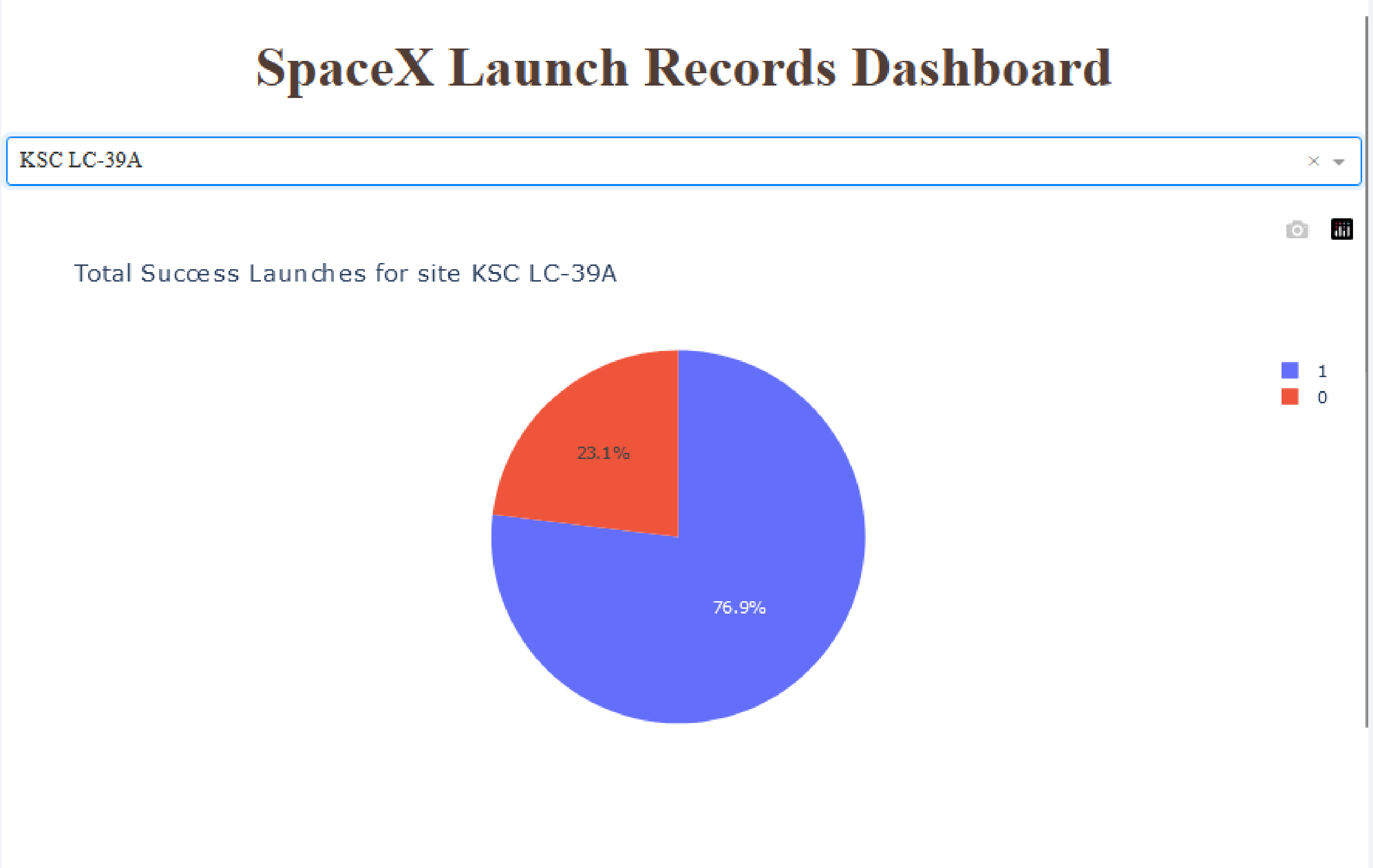
Build a Dashboard with Plotly Dash

Successful launches by Sites



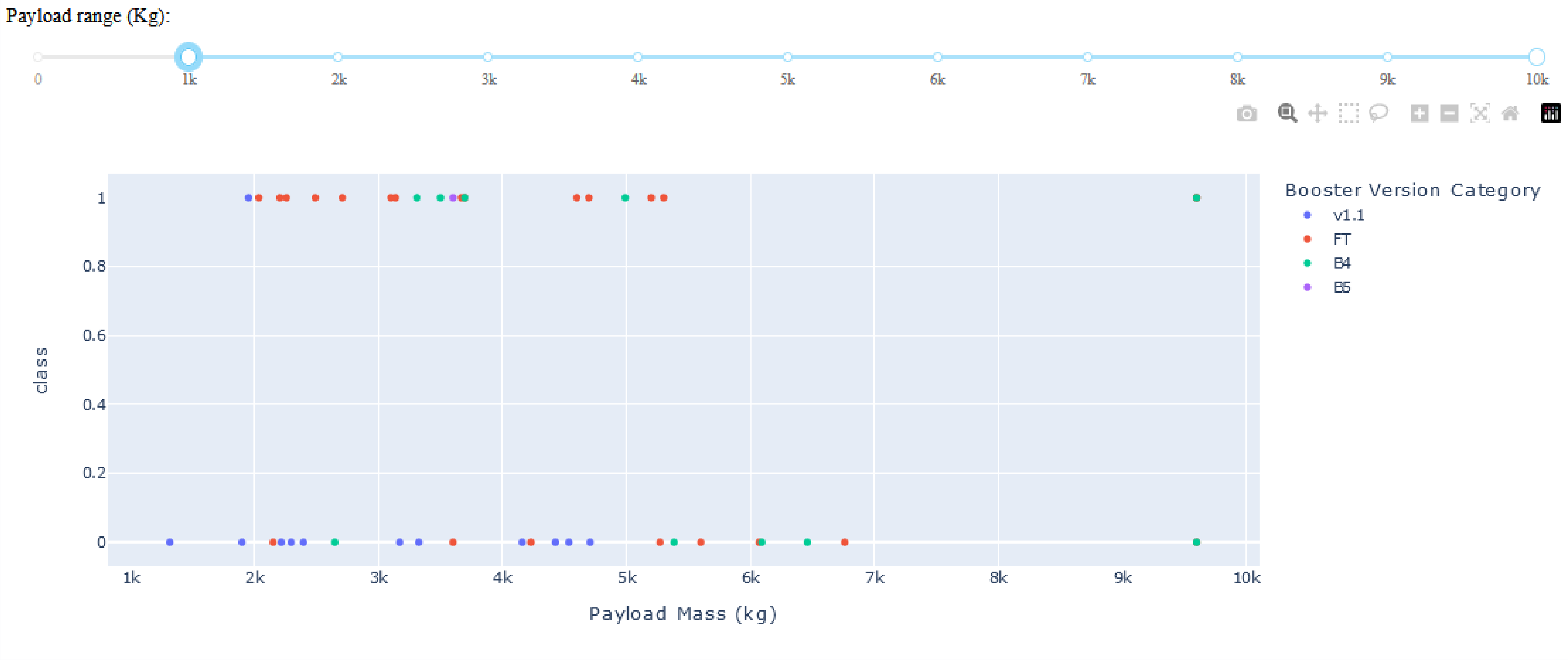
The choice of launch location seems to have an impact on the success rate.

Success ratio for site KSC KC-39A



Site KSC KC-39A has 76.9% success rate.

Payload Mass (kg) VS Launch Outcomes(class)

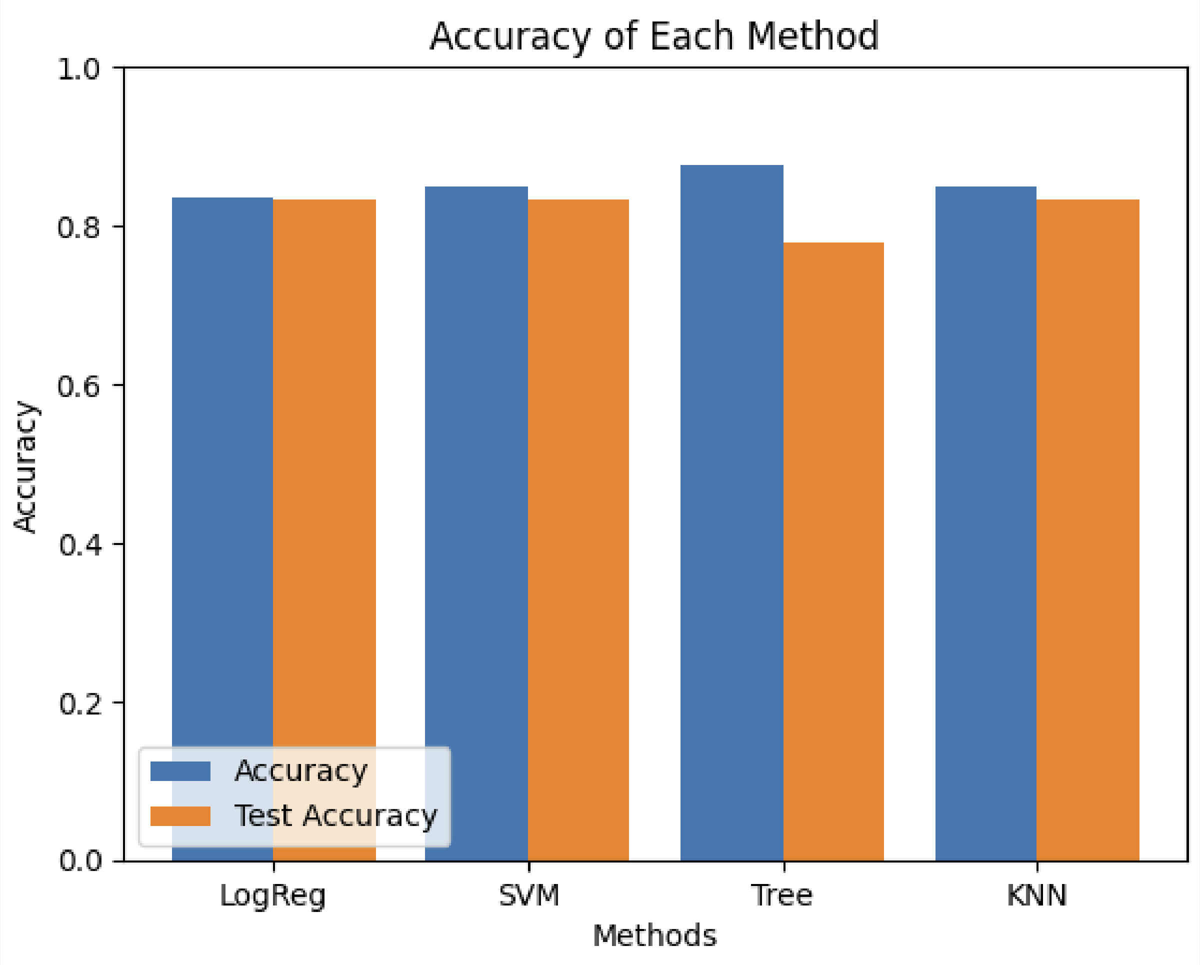


Payload Mass under 6000 kg, combined with FT booster version, is the most successful combination.
Also, there is not a lot of data to make a an estimate on Payload Mass over 6000 kg.

Section 5

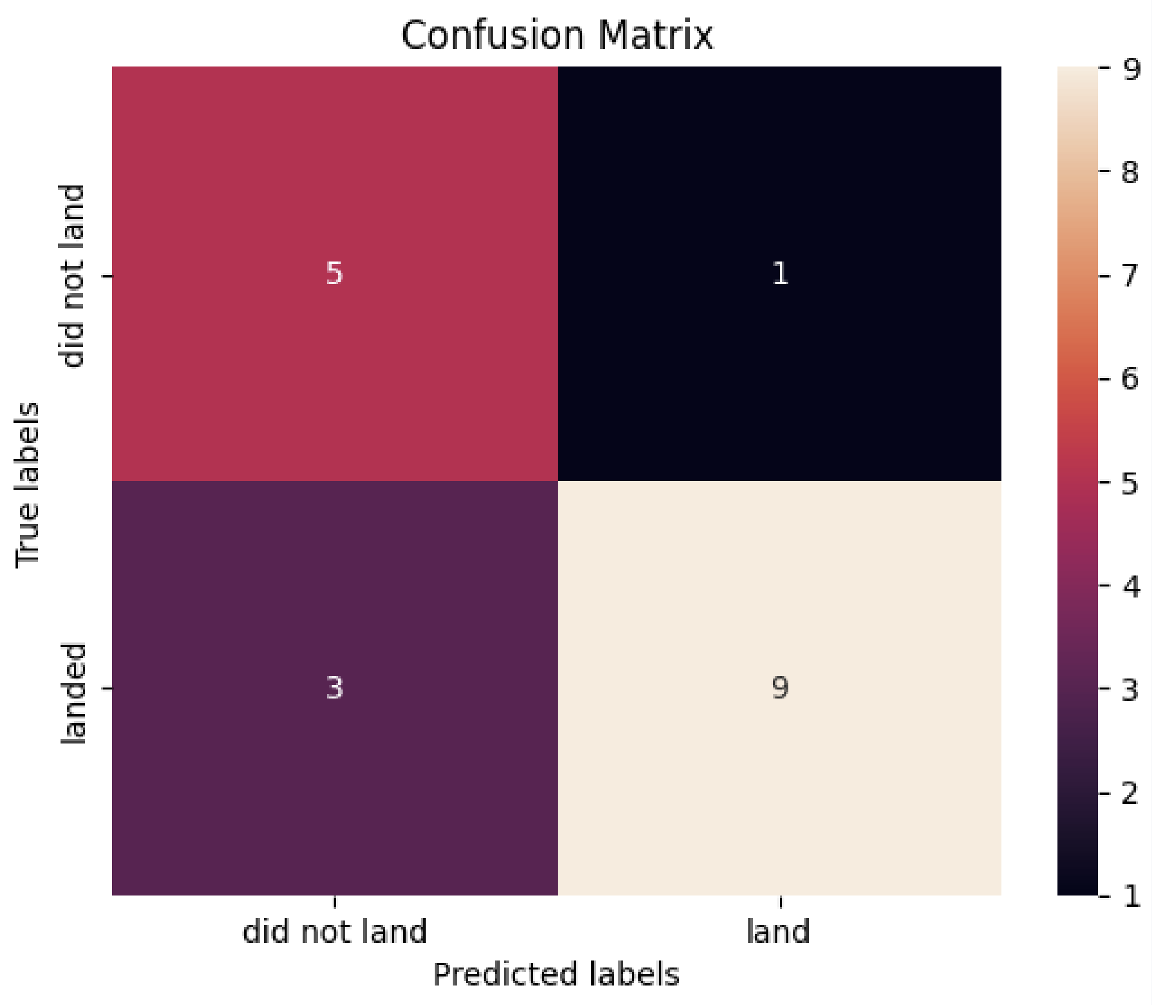
Predictive Analysis (Classification)

Classification Accuracy



- Four classification models were evaluated.
- The most successful model in predicting outcome is the Decision Tree Algorithm with an accuracy over 87%.

Confusion Matrix



The confusion matrix of the Decision Tree Classification Model, show True/False Positive and True/False Negative.

Conclusions

- Different sources of data were evaluated, from API extracted information to WebScraping, revealing insights
- Decision Tree Classification Model proved to be an effective choice for prediction
- The site with the most successful launches is KSC LC-39A
- While the mission outcomes improved over time, there is still room for improvement and that could depend on the evolution of rockets and other factors

Appendix

- Folium maps are not visible on GitHub

Thank you!