IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marco Ramos
2024-05-27

Outline

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

Executive Summary

- Methodologies used in this project
  - Data collection, using web scrapping and SpaceX API;
  - Exploratory data analysis, using static and dynamic/interactive data visualizations;
  - Feature engineering, on selected features to process categorical variables;
  - Predictive analysis, building and testing different classification machine learning models;
- Summary of all results
  - From all the features collected, the best were selected to be used for machine learning, which include Flight Number, Payload Mass, Flights, Block, Reused Count, Orbit, Launch Site, Landing Pad, Serial, Grid Fins, Reused, Reused, Legs;
  - After testing and tunning a decision tree classifier model was selected due to presenting the higher accuracy of the tested models;
  - Its possible to predict the success of a booster landing with the features available;

Introduction

- A new rocket Company Space Y, founded by Billionaire industrialist Allon Musk would like to compete with SpaceX.
- The main problem is predicting the cost of each launch and so, some questions were defined:
  - Can we determine if the first stage/booster will perform a successful landing?
  - Where is the best place to make launches?

Section 1

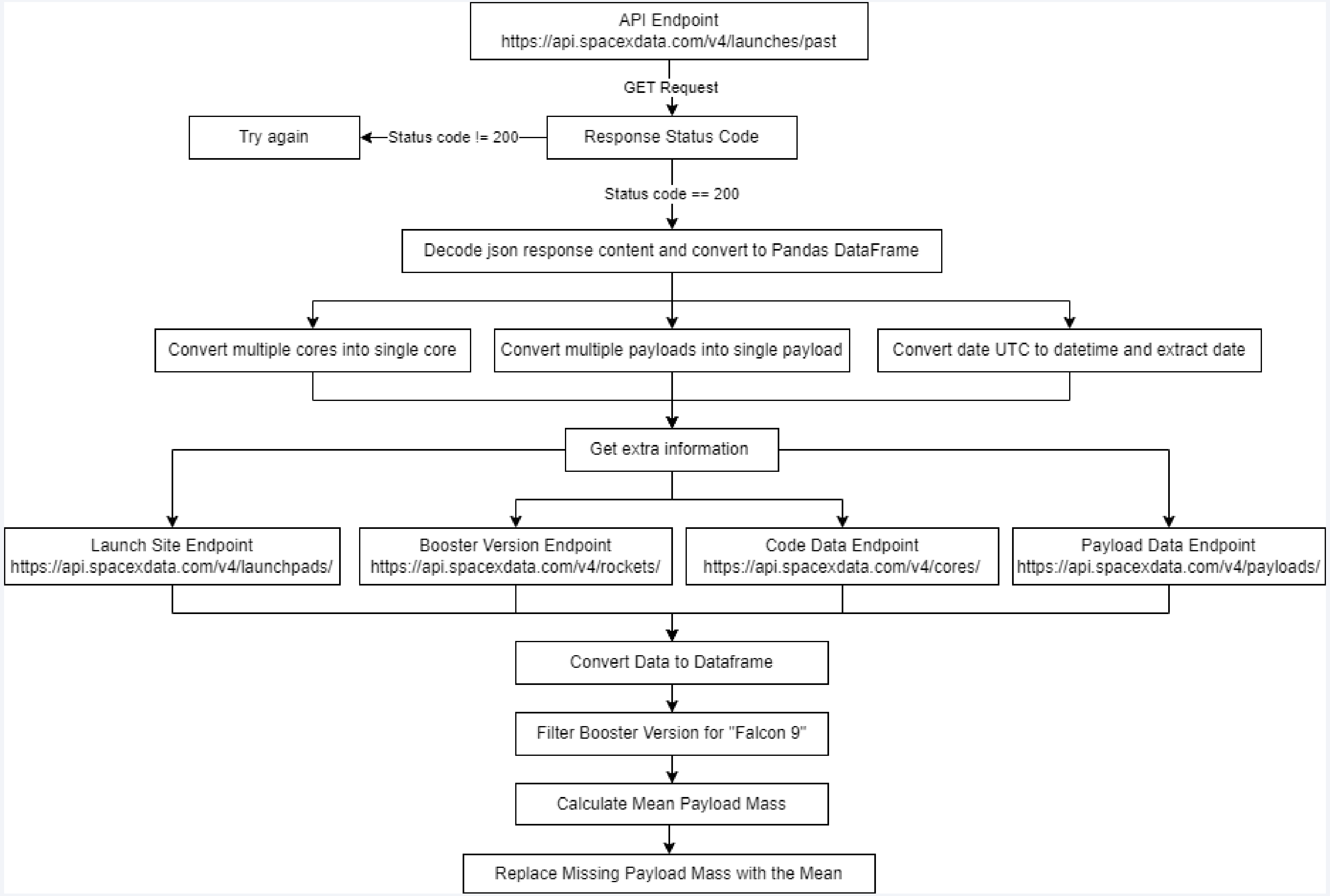
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    - SpaceX public API;
    - Web scrapping Wikipedia page about Falcon Rocket launches:
- Perform data wrangling
  - Data after collection as analyzed and cleaned/formated by filtering unwanted launches related to old booster versions.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API



API Endpoint
https://api.spacexdata.com/v4/launches/past

GET Request

Response Status Code

Status code != 200

Try again

Status code == 200

Decode json response content and convert to Pandas DataFrame

Convert multiple cores into single core

Convert multiple payloads into single payload

Convert date UTC to datetime and extract date

Get extra information

Launch Site Endpoint
https://api.spacexdata.com/v4/launchpads/

Booster Version Endpoint
https://api.spacexdata.com/v4/rockets/

Code Data Endpoint
https://api.spacexdata.com/v4/cores/

Payload Data Endpoint
https://api.spacexdata.com/v4/payloads/

Convert Data to Dataframe

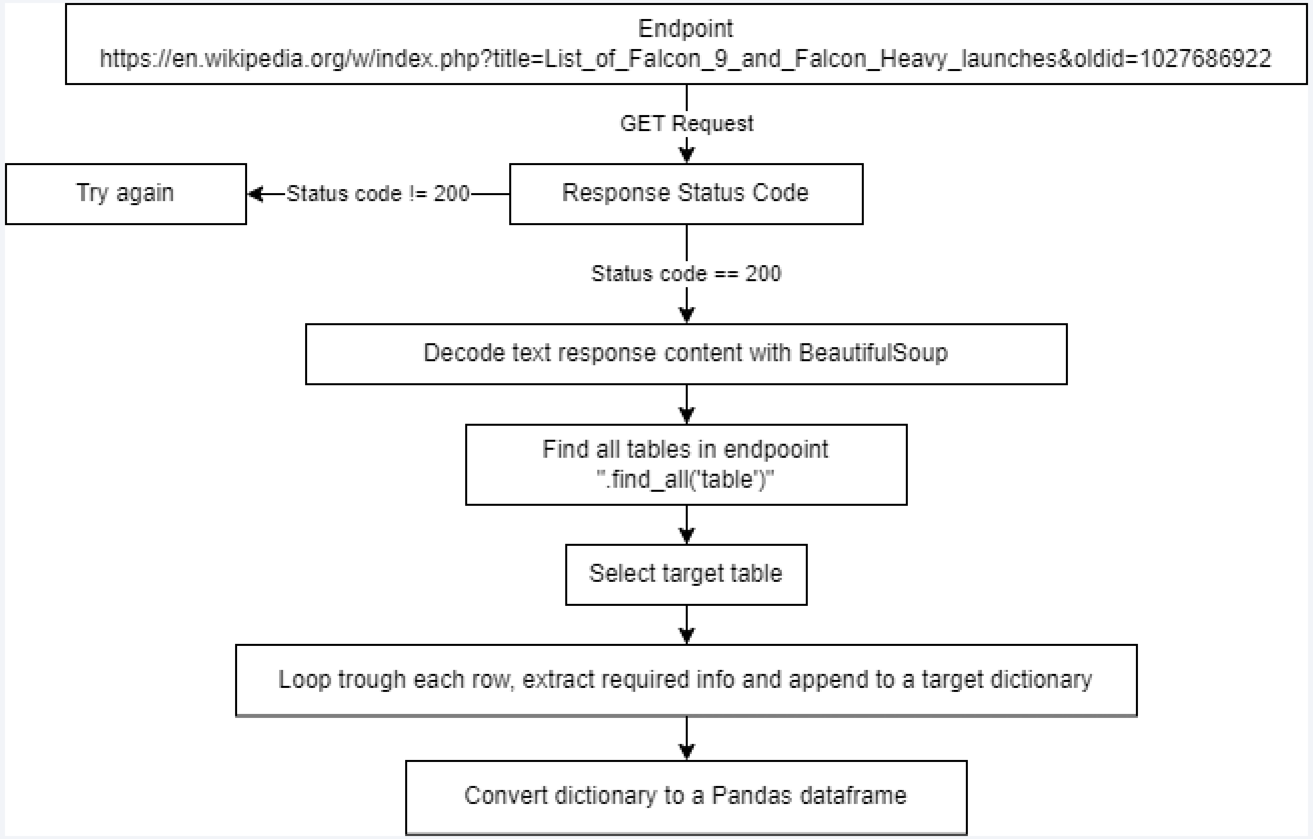
Filter Booster Version for "Falcon 9"

Calculate Mean Payload Mass

Replace Missing Payload Mass with the Mean

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Endpoint
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

GET Request

Response Status Code

Status code != 200

Try again

Status code == 200

Decode text response content with BeautifulSoup

Find all tables in endpooint
".find_all('table')"

Select target table

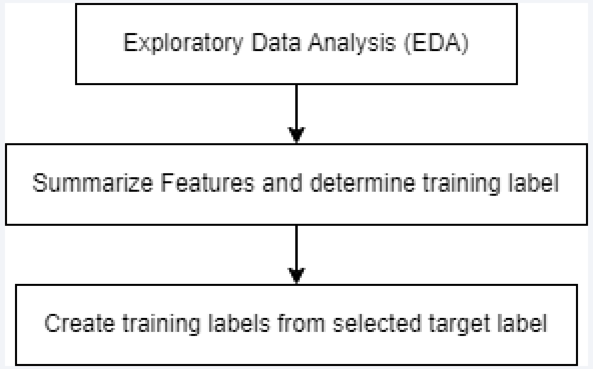
Loop trough each row, extract required info and append to a target dictionary

Convert dictionary to a Pandas dataframe

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/jupyter-labs-webscraping.ipynb

Data Wrangling

- First Exploratory Data Analysis was performed on the dataset analyzing the launch sites, target orbits and launch outcomes.
- Summarize some of the features and determine the training labels.
- Create the training label (landing outcome) for the Outcome column.



Exploratory Data Analysis (EDA)

Summarize Features and determine training label

Create training labels from selected target label

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To perform EDA trough visualizations, the plot types were used:
  - Scatter plots;
  - Bar plots;
  - Line plots;

- Features used for plotting:
  - Payload Mass
  - Flight Number
  - Launch Site
  - Orbit

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The names of the unique launch sites in the space mission;
- 5 records where launch sites 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 successfully landed in drone ship and have payload mass greater than 4000 but less than 6000;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site by month 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;

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- An interactive map was created using Folium with markers, circles and lines used to represent the launch sites and other features around them
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates, like between a launch site and nearest city, train line, etc;

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/lab_jupyter_launch_site_location.ipynb

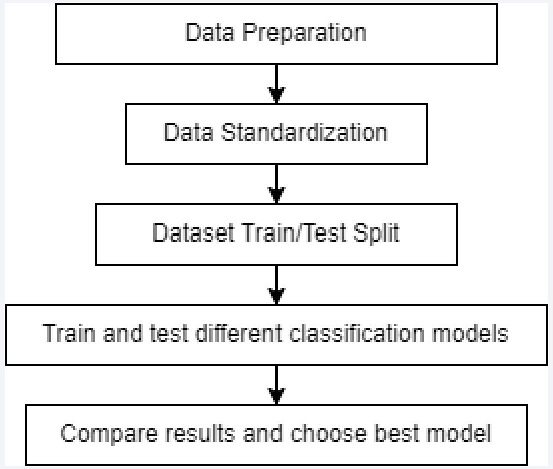
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of launches by site;
  - Percentage success launches by site;
  - Payload mass vs success launch vs booster version;
- These interactive visualizations allowed to quickly analyze the success ratio by launch site and the relation between payloads mass and launch sites, helping to identify where is best place to launch.

Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/spacex_dash_app.py

Predictive Analysis (Classification)

- Dataset was prepared by:
  - Performing feature engineering by creating dummy variables from the categorical features
  - Applying standardization to the features;
- The dataset was split in train and test dataset with choosing a test size of 20%;
- 4 different classification models were tested:
  - Logistic Regression;
  - Support Vector Machine;
  - Decision Tree Classifier;
  - K Nearest Neighbors;
- The results were compared using 3 different metrics: accuracy, jaccard score and F1 score;



Data Preparation

Data Standardization

Dataset Train/Test Split

Train and test different classification models

Compare results and choose best model

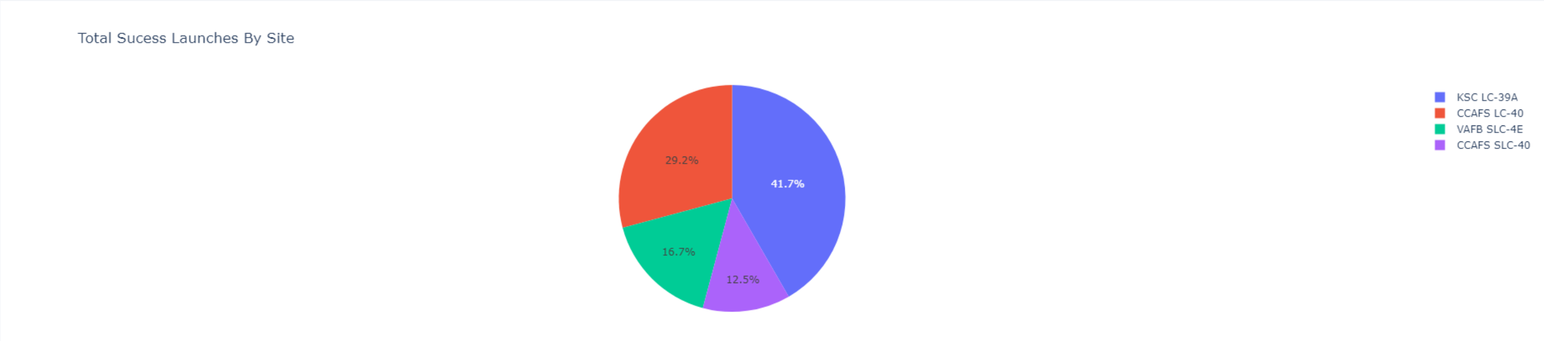
Source Code: https://github.com/Majramos/coursera-applied-data-science-capstone/blob/main/assignments/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

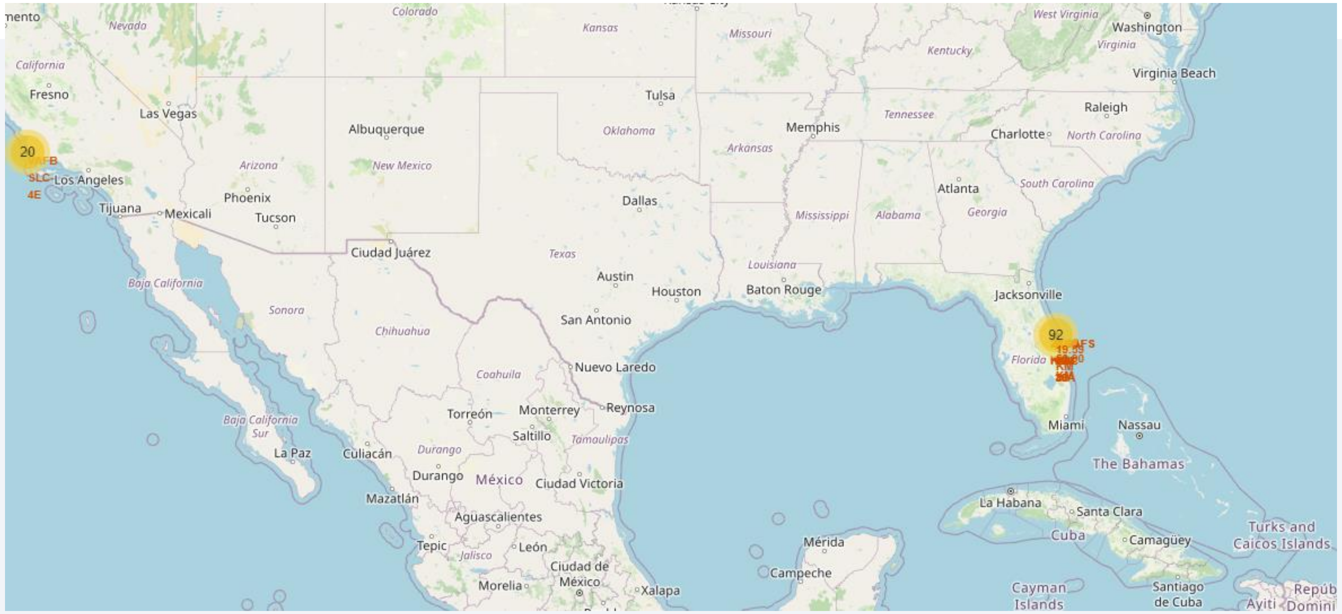
Results

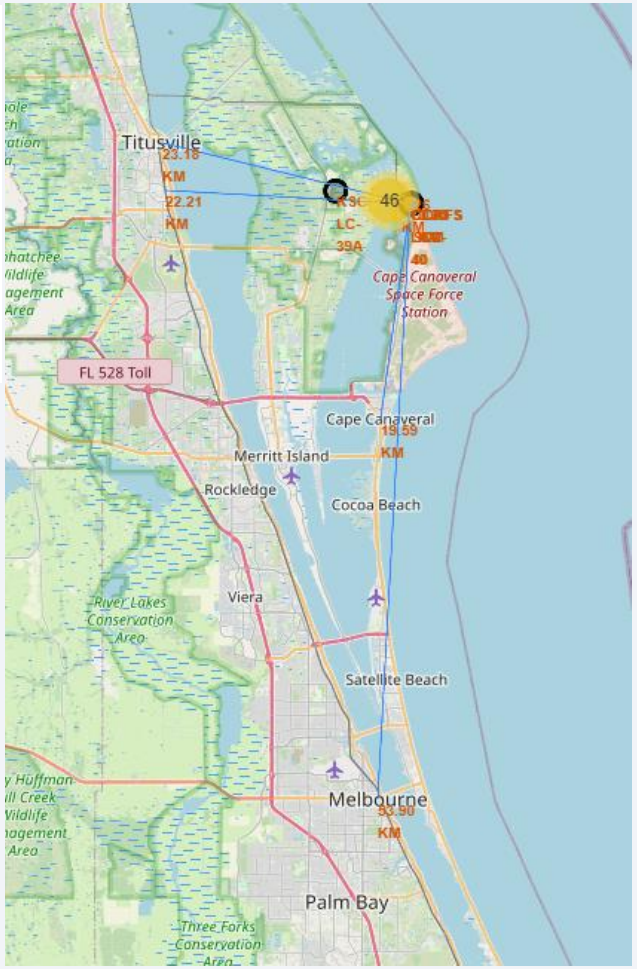
- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome in a ground pad happened in 2015;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Only one mission outcome in 102 failed;
  - The booster version to carry the max payload until date was the F9 B5;
  - Payload mass and success ratio has been increasing with the number of flights/years;
  - VAFB-SLC launch site has no rockets launched for heavy payload mass (greater than 10000);
  - There are 4 target orbits with 100% success ratio (ES-L1, GEO, HEO, SSO);
  - The heaviest payloads were targeting VLEO orbit;

Results

- Using interactive analytics, it was possible to identify the best launch site based on historical success ratio and proximity to required logistic infrastructure;

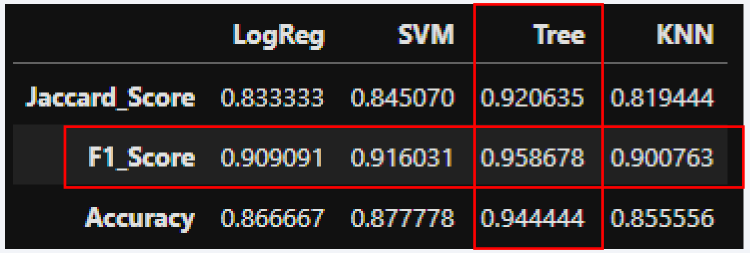






Results

- Predictive analysis results showed that Decision Tree Classifier has the best results to predict successful landings, based on the F1 score of the model

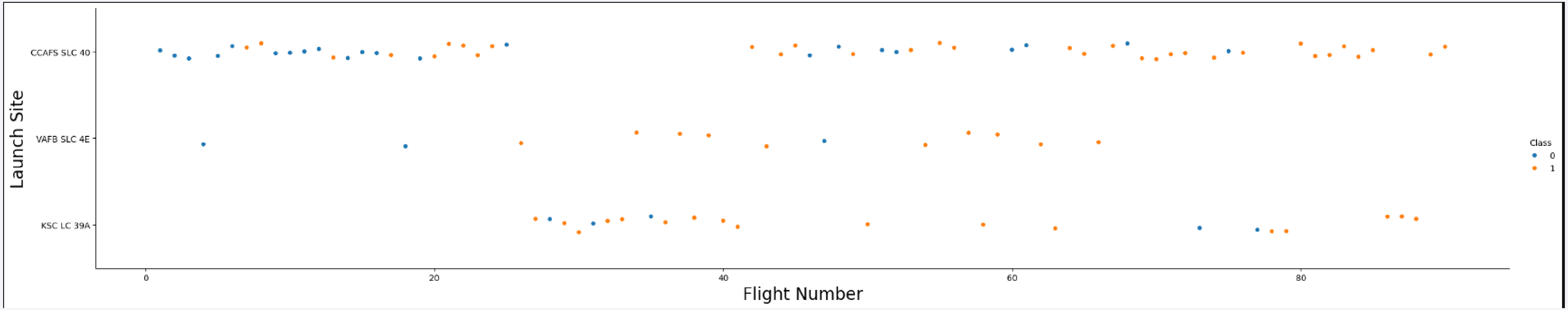


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.920635	0.819444
F1_Score	0.909091	0.916031	0.958678	0.900763
Accuracy	0.866667	0.877778	0.944444	0.855556

Section 2

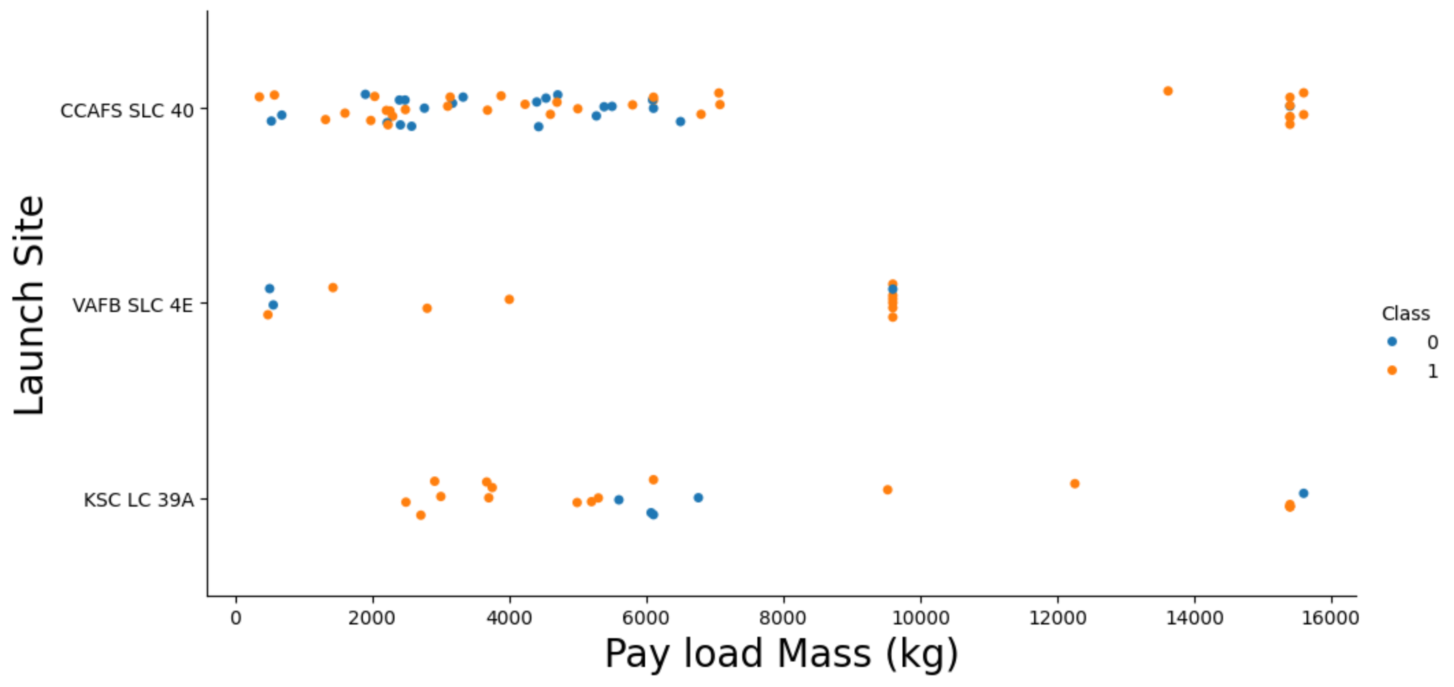
Insights drawn from EDA

Flight Number vs. Launch Site



- The general success rate increased with flight number time;
- The site CCAFS SLC 40 has had the most launches overall;
- All 3 sites had success booster landings in their last 5 launches;

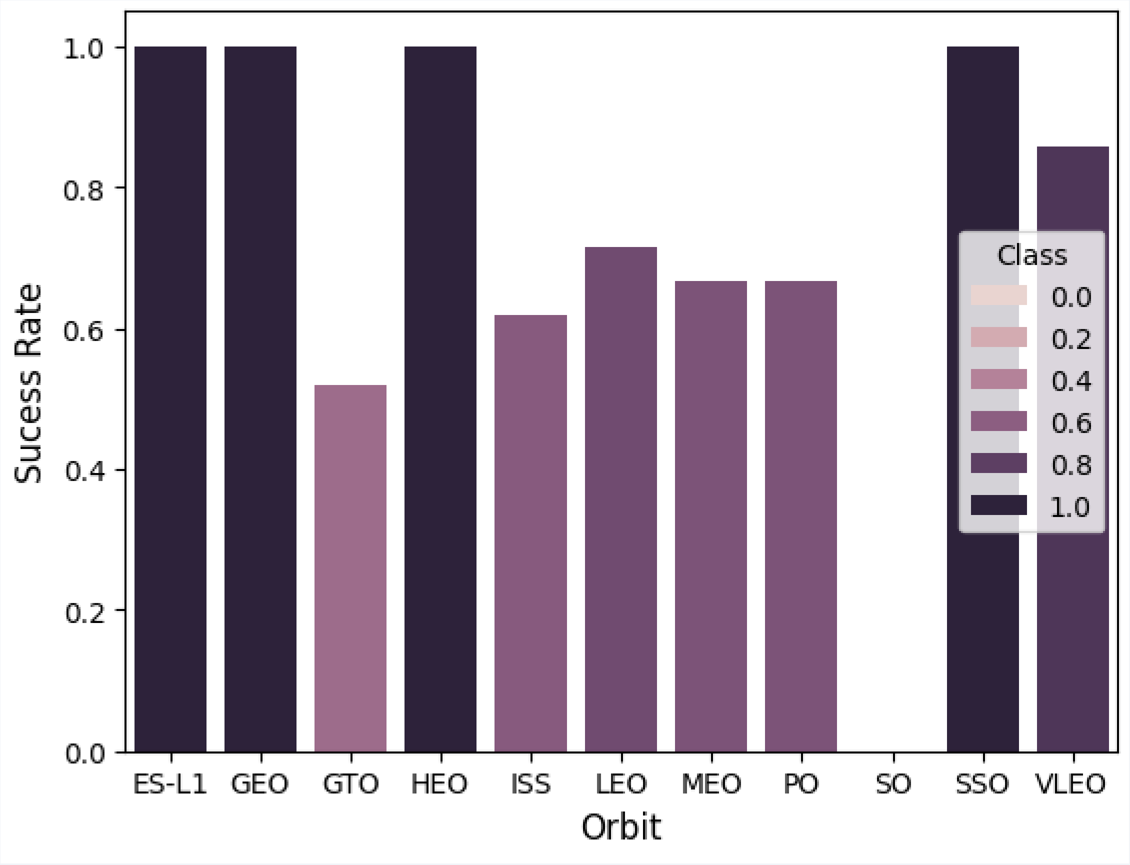
Payload vs. Launch Site



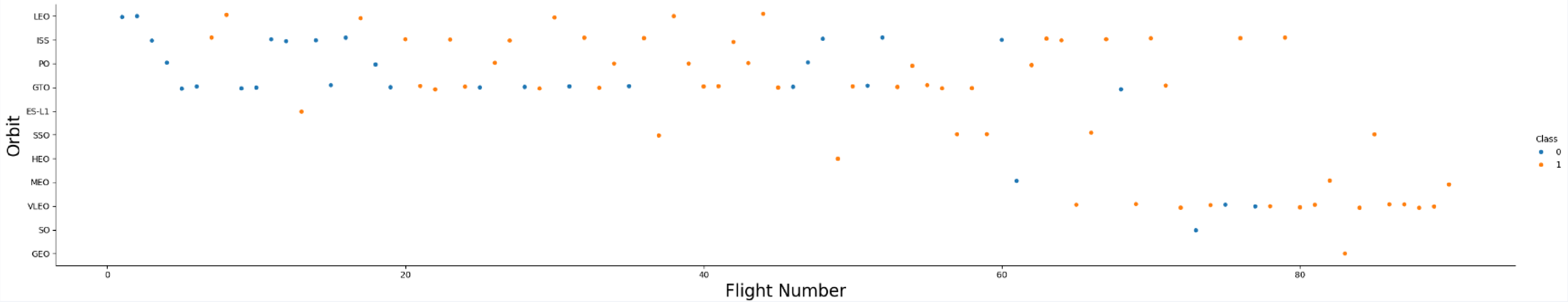
- High payloads (over 9000kg) have high success ratio;
- Launch site VAFB SLC 4E hasn't launched payloads above 10000kg. Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites;

Success Rate vs. Orbit Type

- Four orbits have 100% success rate: ES-L1, GEO, HEO, SSO;
- One orbit, VLEO, with high success rate (above 80%,;
- Orbits GTO, ISS, LEO, MEO, PO have moderate success rate (between 50% and 80%);
- One orbit, SO, never had a successful booster recover;



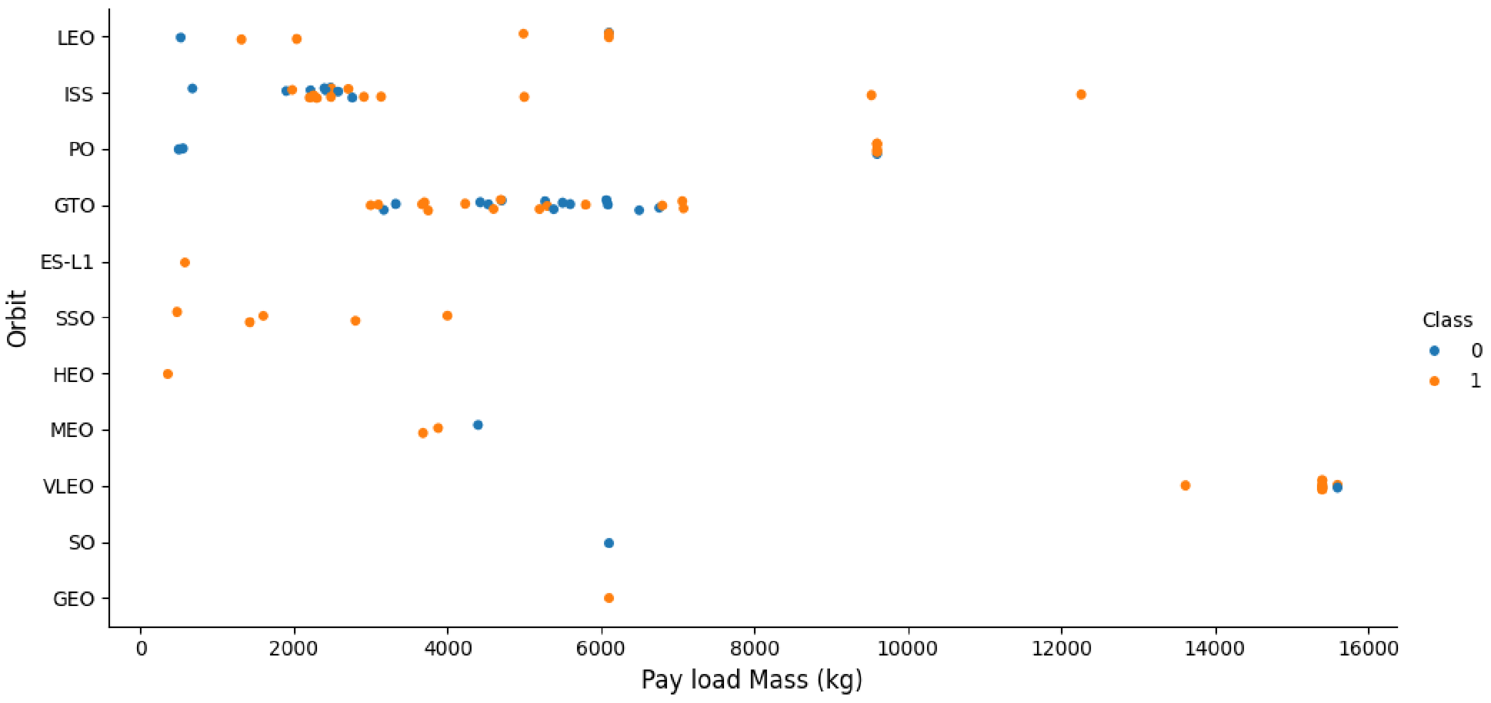
Flight Number vs. Orbit Type



- GEO orbit only with one recent and successful launch;
- MEO, VLEO, SO, GEO are all recently targeted orbits with apparently successfully launches;

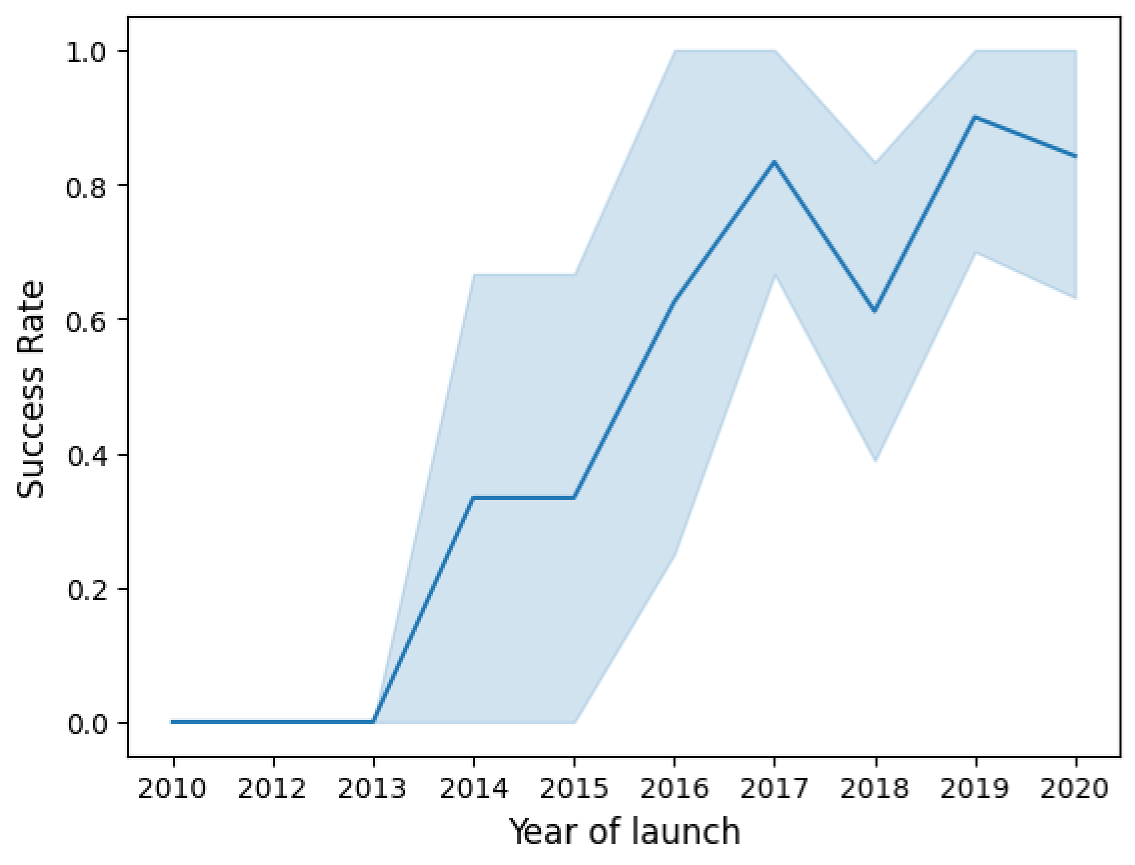
Payload vs. Orbit Type

- Only the VLEO orbit has received high payload masses;
- Orbits ES-L1, SSO, HEO have 100% success rate with low payload;
- No apparent relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success masses;
- There are few launches to the orbits SO and GEO;

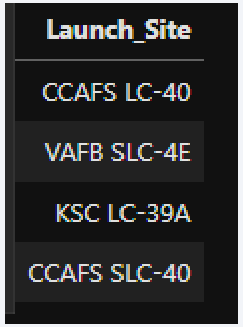


Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2017
- It stayed stable in 2014;
- After 2015 it started increasing until 2017;
- In 2017 there was a decrease in the success ratio, but by the end of 2019 it had recover;
- 2019 seems to be the year with the highest observed success ratio;



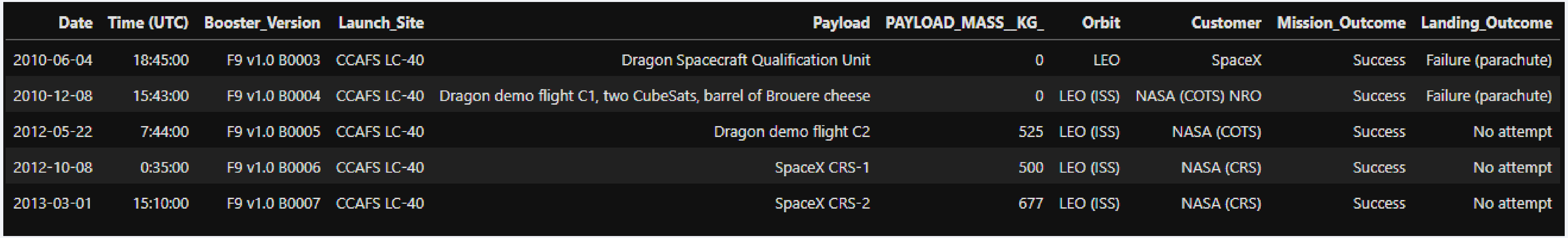
All Launch Site Names



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

- All launch site names were obtained by selecting the distinct values of “Launch_Site” from the dataset;
- Only 4 launch sites available in the dataset;

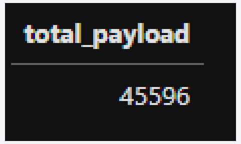
Launch Site Names Begin with 'CCA'



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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The 5 records were selected by filtering the field “Launch_site” starting by the string ‘CCA’ using a wildcard ‘CCA%’;
- Launch site beginning with CCA refer to the CCAFS LC-40 launch site;
- The first 5 records for launch site CCAFS LC-40 were all using booster versions F9 v1.0, all with successful mission outcome and no recovery of booster;

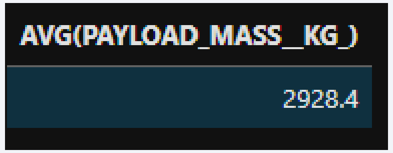
Total Payload Mass



total_payload
45596

- The total payload mass carried by boosters launched by NASA (CRS) was obtained trough the sum of the payload mass (PAYLOAD_MASS__KG_, renamed to total_payload) where the “Customer” is equal to ‘NASA CRS’;

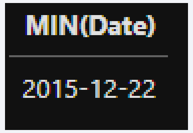
Average Payload Mass by F9 v1.1



AVG(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass carried by boosters of version F9 v1.1 was obtained trough the avg of the payload mass (PAYLOAD_MASS__KG_) where the “Booster_Version” is equal to ‘F9 v1.1’;

First Successful Ground Landing Date

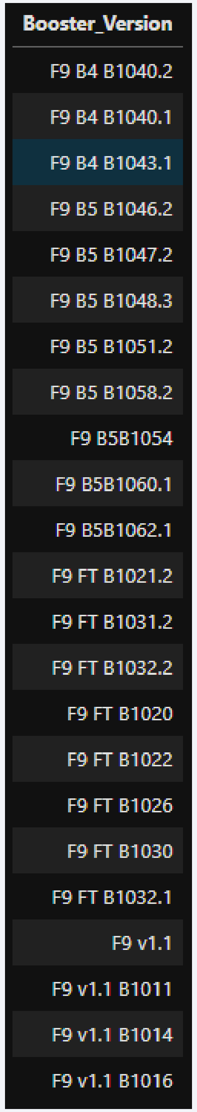


MIN(Date)
2015-12-22

- The first date of a success landing in a ground pad was obtained trough the minimum value of “Date” were the “Landing_Outcome” is equal ‘Success (ground pad)’;

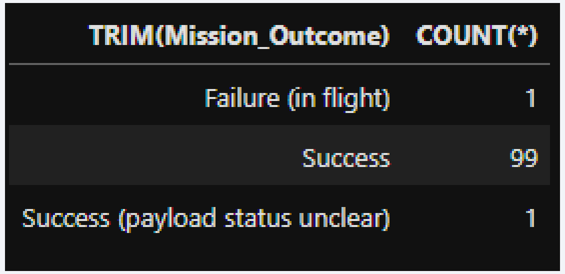
Successful Drone Ship Landing with Payload between 4000 and 6000

- The distinct values of “Booster_Version” were selected by selecting where “PAYLOAD_MASS__KG_” larger than 4000 and less than 6000;



Booster_Version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

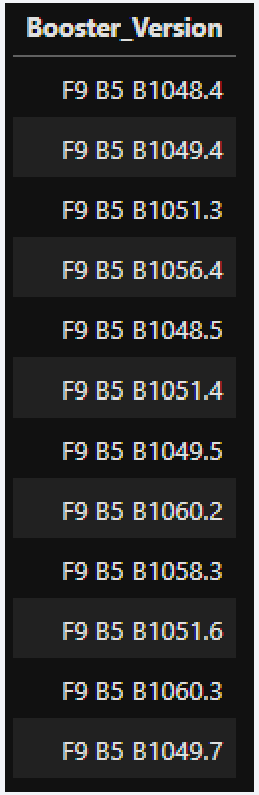


TRIM(Mission_Outcome)	COUNT(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The total number of successful and failure mission outcomes was obtained by grouping mission outcomes and counting records/rows for each group;

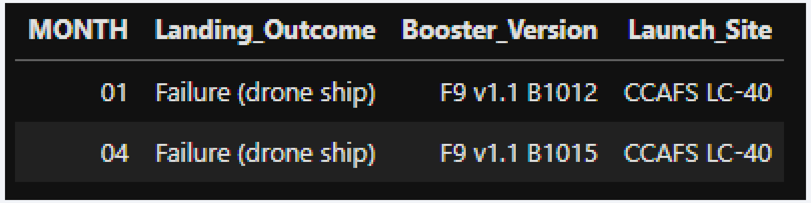
Boosters Carried Maximum Payload

- The list of booster names which carried the maximum payload were obtained by selecting the distinct values of “Booster_Version” where the payload mass given by the field “PAYLOAD_MASS__KG_” is equal to the maximum values of “PAYLOAD_MASS__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

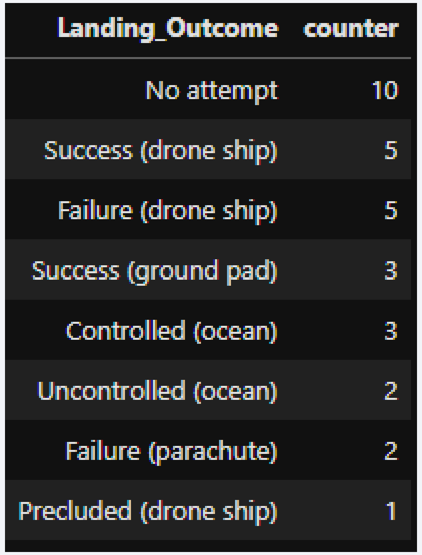


MONTH	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 were obtained by selecting the features “Landing_Outcome”, “Booster_Version”, “Launch_Site” filtering where the first 4 characters of “Date” are equal to ‘’2015’ and “Landing_Outcome” is equal to ‘Failure (drone ship)’;

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

- To get the 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, one grouped landing outcomes and counted records/rows for each group while filtering “Date” between ‘2010-06-04’ and ‘2017-03-20’ ordering the results by the count of rows of each group;



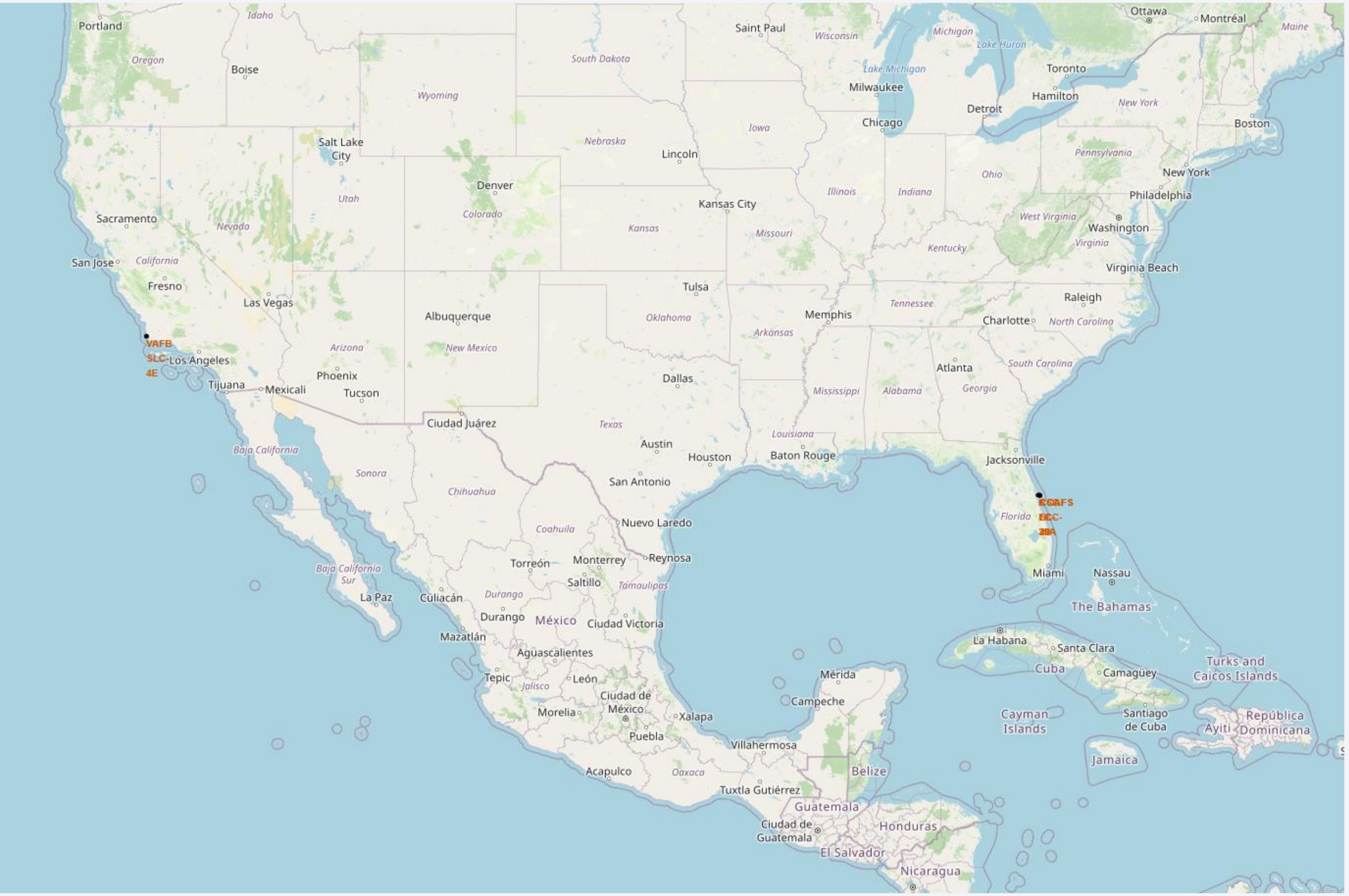
Landing_Outcome	counter
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

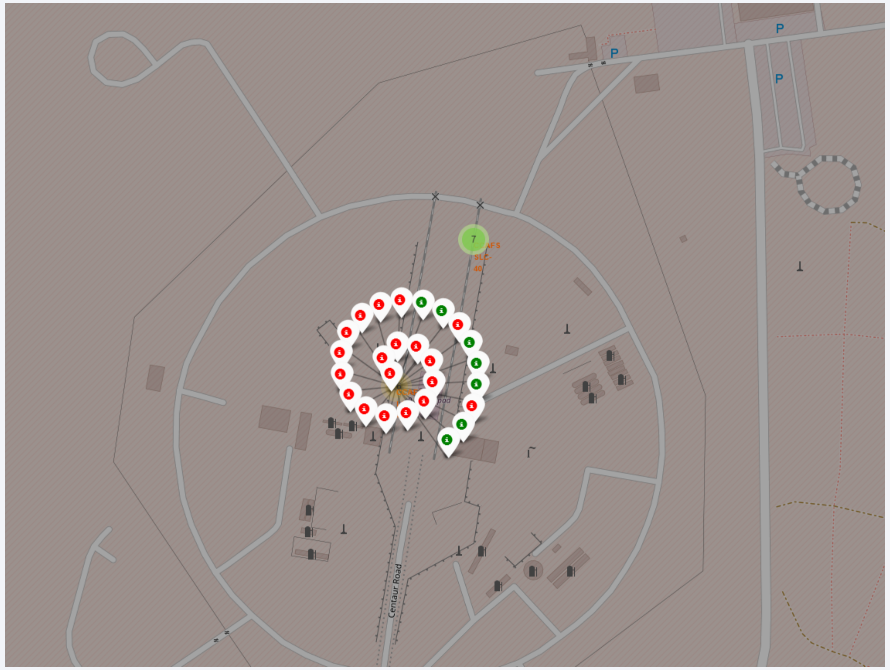
Launch Sites Proximities Analysis

All launch sites

- All the launch sites are close to the ocean, far from cities and as south as possible in the United States America;



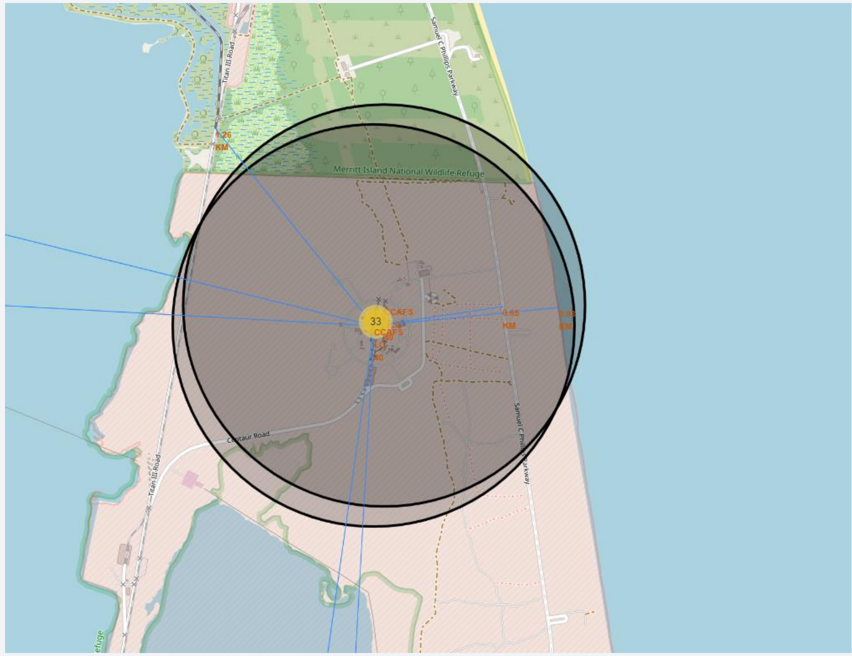
Launch Outcomes by Site

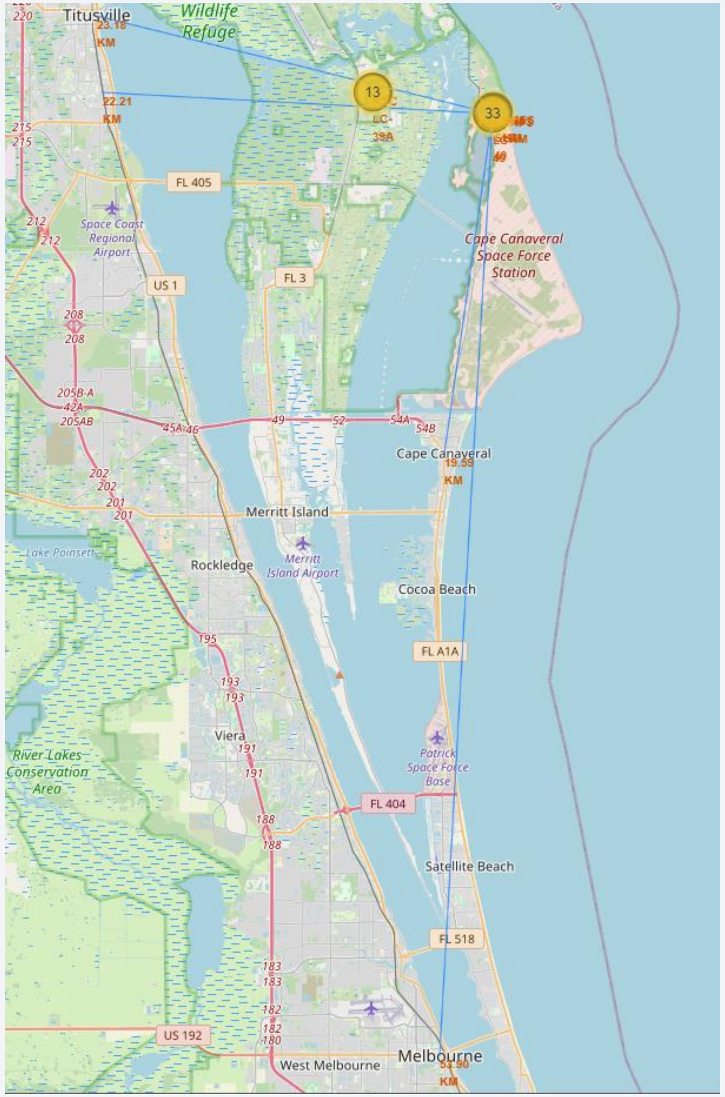


- Green markers indicate successful and red ones indicate failure;

Logistics and Safety

- Launch sites at the Cape Canaveral Space Force Station present a good distance to utilities like roads and train and enough distance to cities;

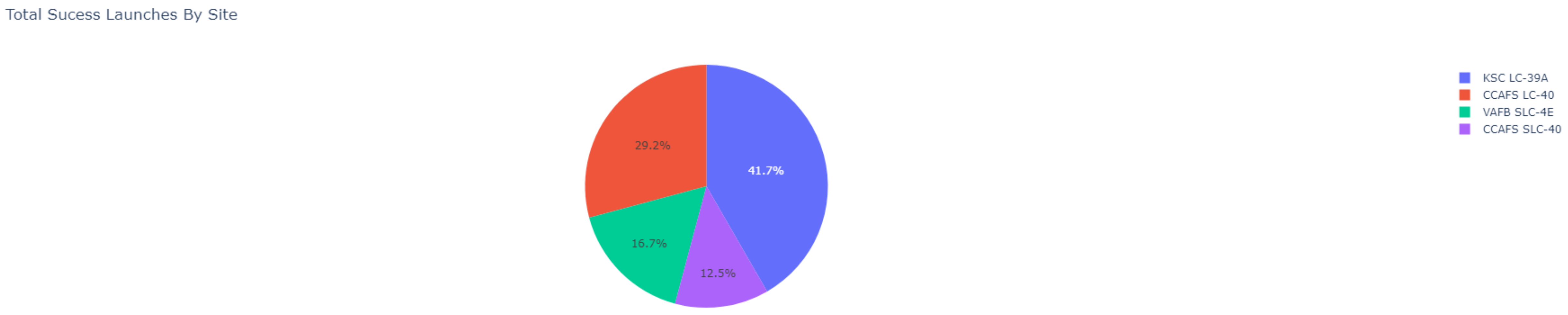

Merritt Island National Wildlife Refuge
33
Centaur Road
Samuel C Phillips Parkway



Section 4

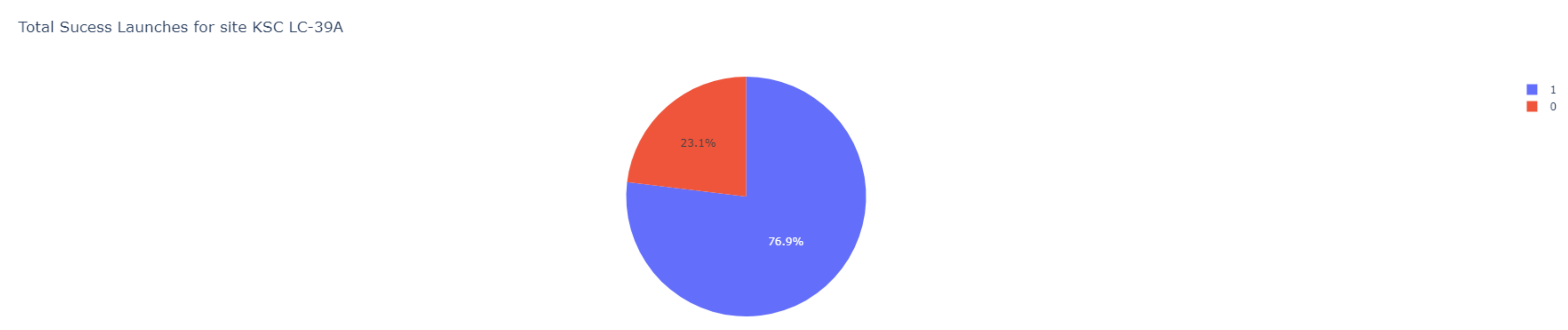
Build a Dashboard with Plotly Dash

Successful Launches by Site



- The launch site with the most success launches is KSC LC-39A with 41.7% of all the successful launches.

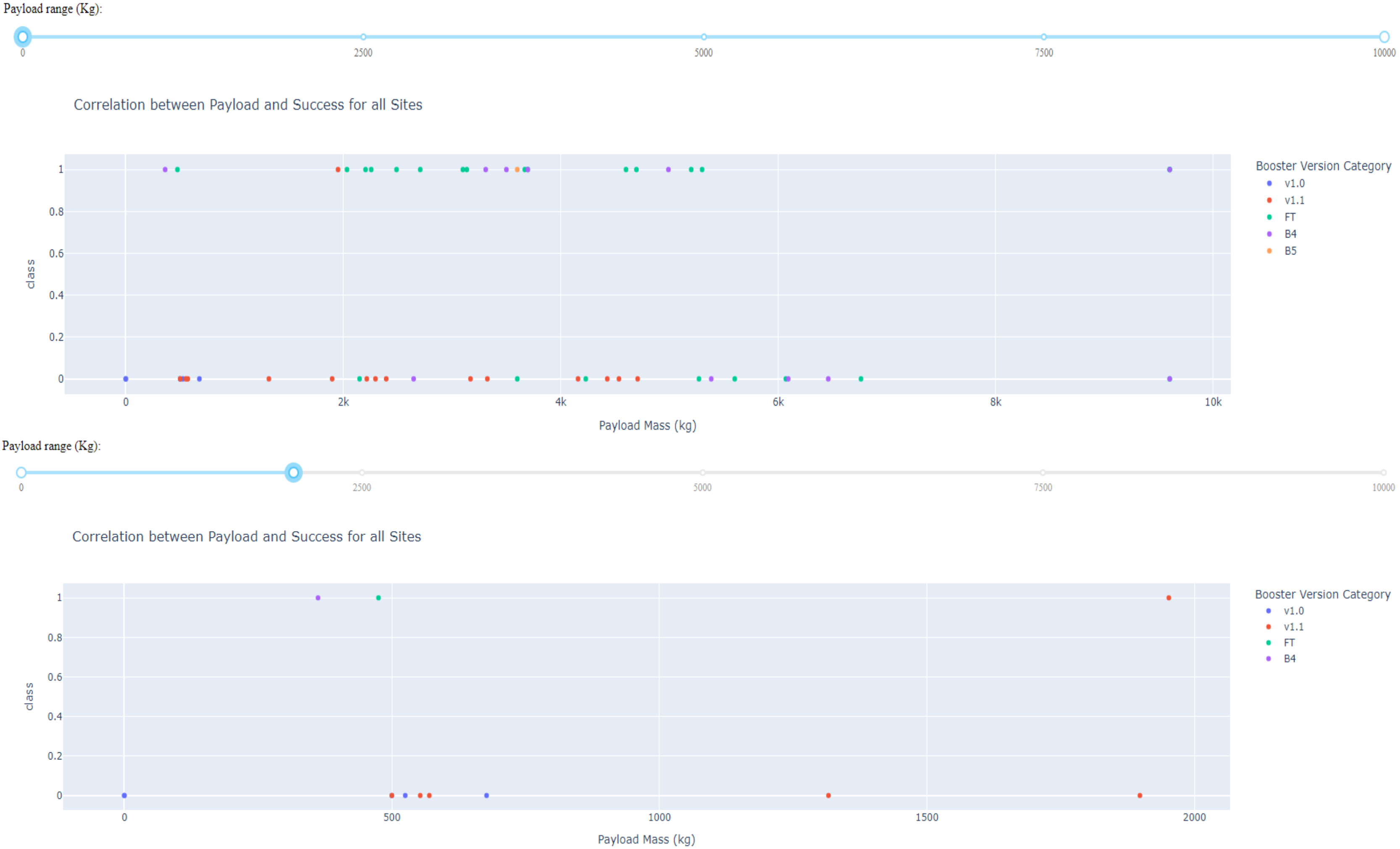
Launch Success Ratio for KSC-39A



Total Sucess Launches for site KSC LC-39A

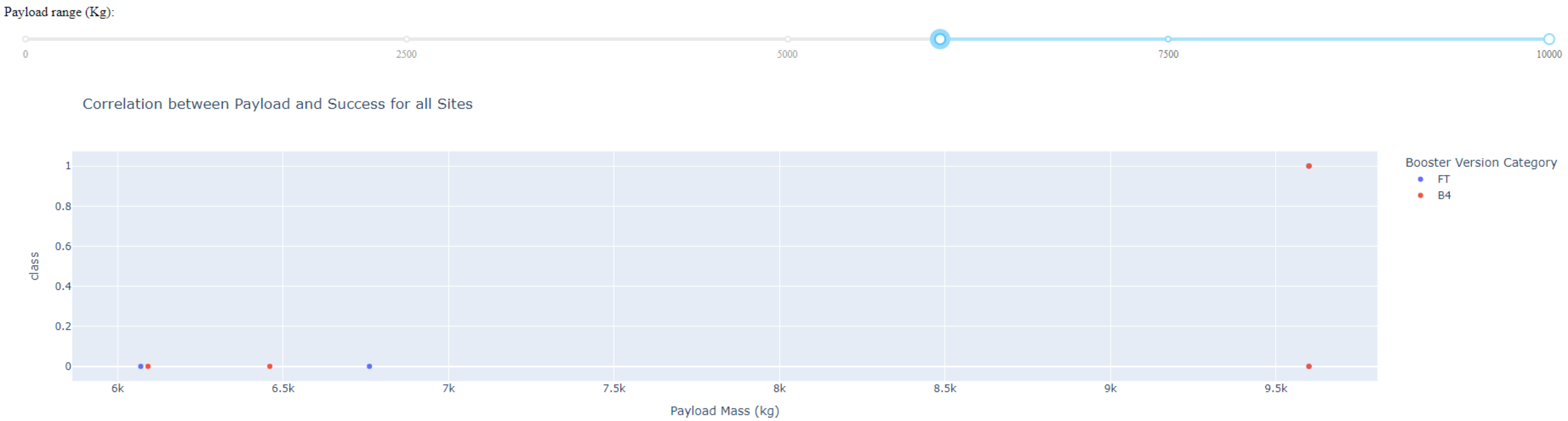
- The site KSC LC-39A has a launch success ratio of 76.9%;

Payload vs. Launch Outcome vs. Booster Version



- Versions v1.0 and v1.1 have low payload launches and success ratio;
- For booster version B5, only have one point with a successful launch;
- With lower payload mass, more recent booster versions and a high success ratio;

Payload vs. Launch Outcome vs. Booster Version



Payload range (Kg):

Correlation between Payload and Success for all Sites

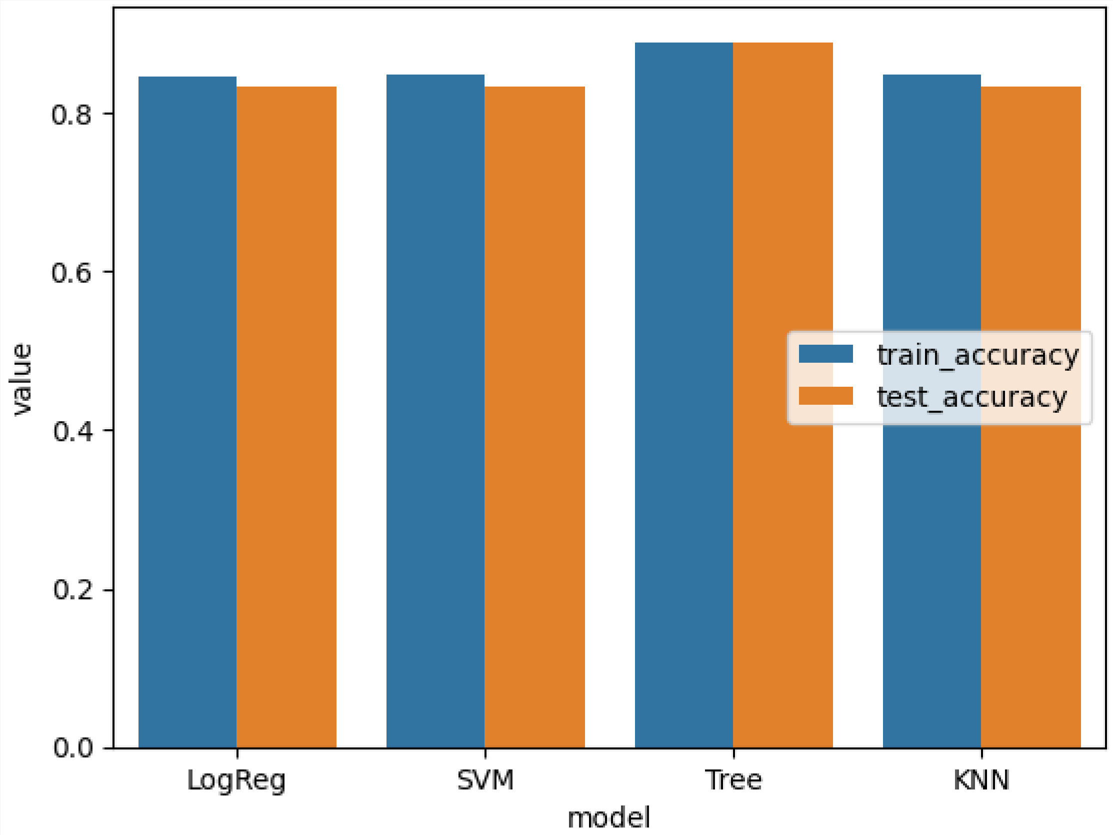
- Both FT and B4 had a success and fail launch with payload higher than 7000kg making impossible with only these two variables make proper predictions to high mass payloads;

Section 5

Predictive Analysis (Classification)

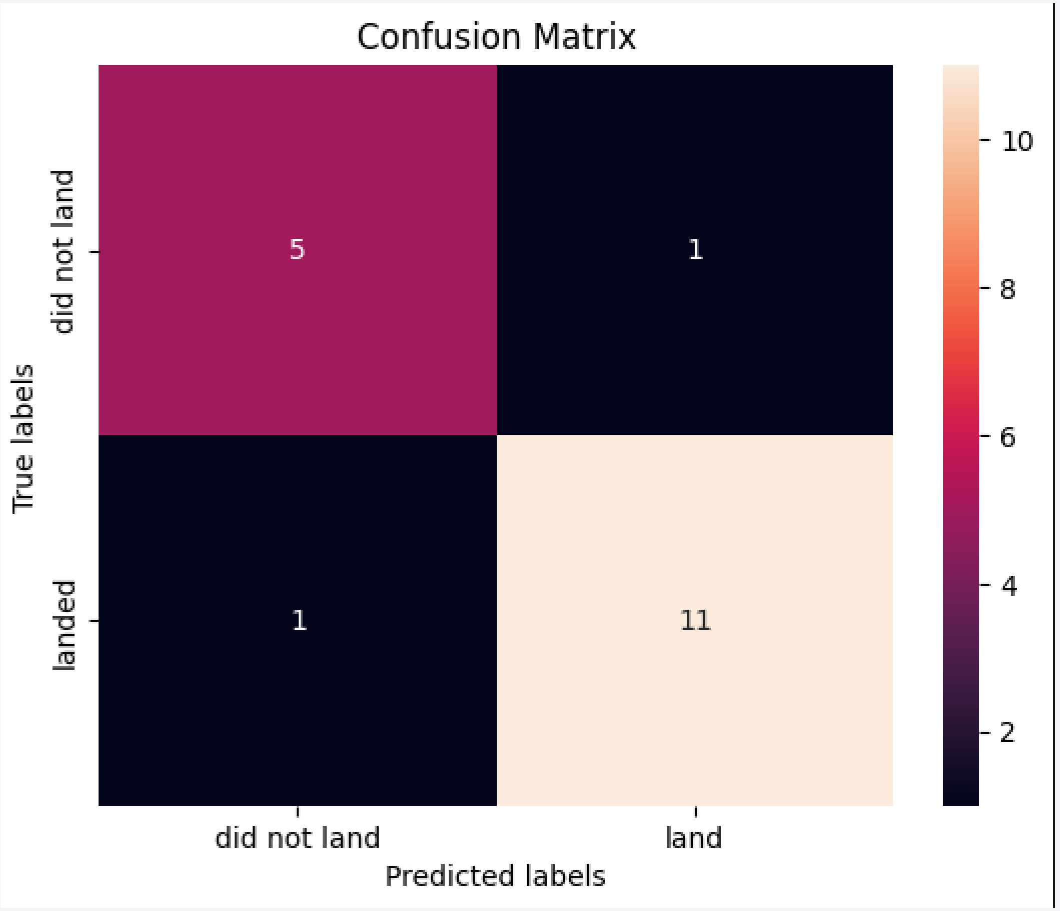
Classification Accuracy

- The best model is the decision tree classifier due to obtaining highest accuracy no only in the train dataset but also in the test dataset;



Confusion Matrix

- The confusion matrix for the decision tree classifier showed only one false negative and one false positive;



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The site KSC LC-39A based on historic successes and proximity to utilities, presents itself as the best location for launches;
- Launches with high payloads masses present an apparent higher success ratio, but it could be biased due to high payloads only happened in more recent flights;
- 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 booster landings success and thus increase profits;

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!