IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gijo Joseph George
18 July 2022

Outline

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

Executive Summary

The following methodologies were used to analyze data:

- Data Collection using web scraping and SpaceX API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
- Machine Learning Prediction.

Summary of all results:

- Success of launches improved as years passed.
- A Decision tree classifier model predicted launch outcome with an accuracy of over 87%

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.

- Problem Statement :
  - Predict whether next falcon 9 launches would succeed or not.
  - What criterions or parameters influence a successful launch ?
  - What insights can you draw from a successful launch ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- 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

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

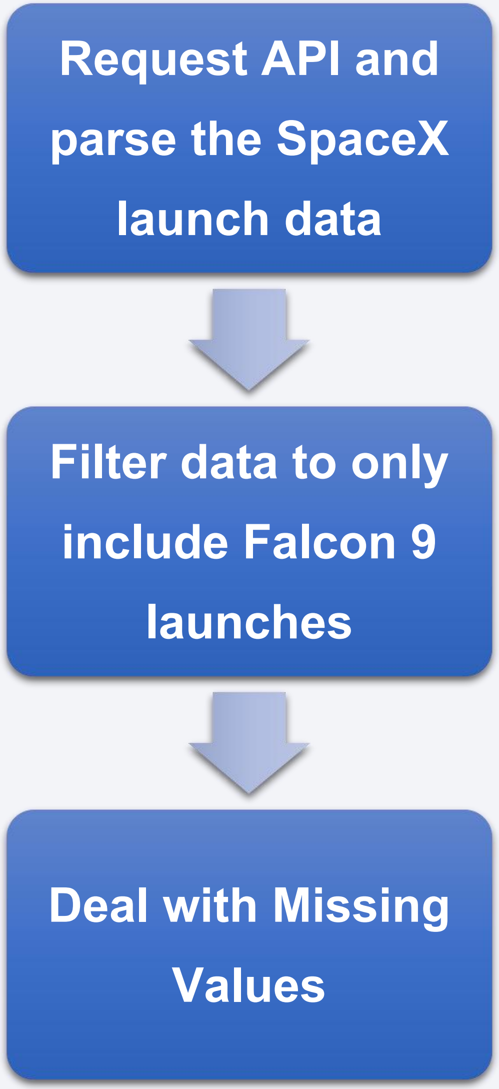
Data Collection

Data sets were collected from :

- Space X API (https://api.spacexdata.com/v4/rockets/),and
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
  - Source Code : https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/1.%20jupyter-labs-spacex-data-collection-api.ipynb



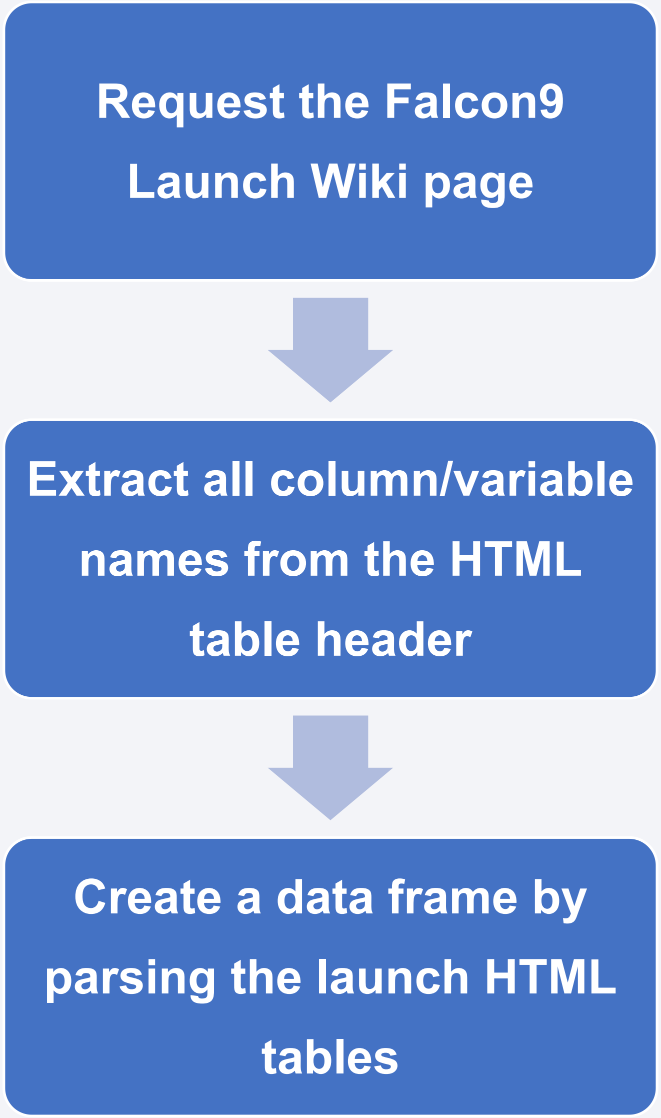
Request API and parse the SpaceX launch data

Filter data to only include Falcon 9 launches

Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
    - Source code: https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/2.%20jupyter-labs-webscraping.ipynb



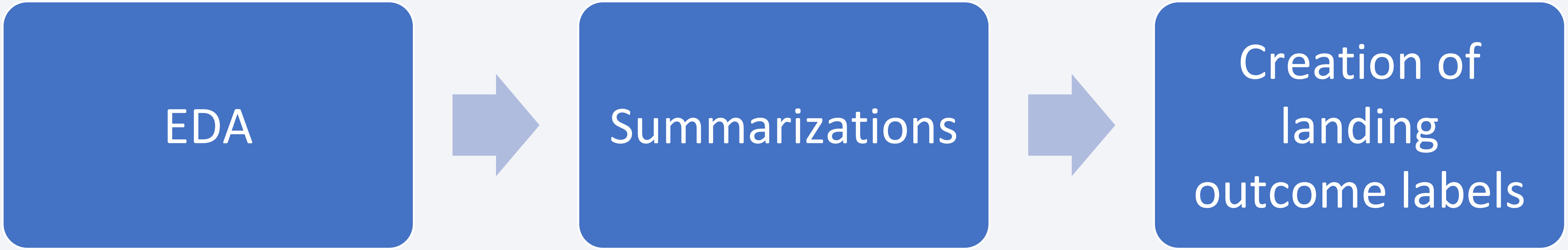
Request the Falcon9 Launch Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



EDA

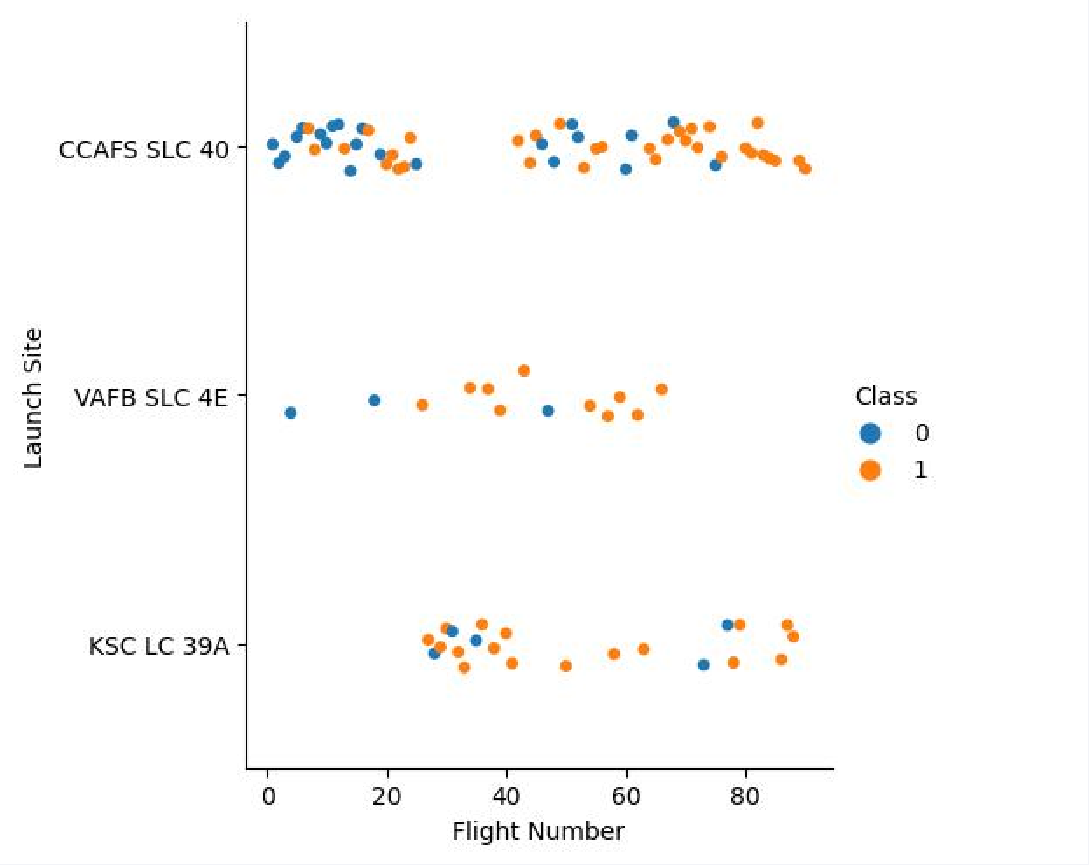
Summarizations

Creation of landing outcome labels

- Source Code : https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
    - Payload Mass & Flight Number, Launch Site & Flight Number, Launch Site & Payload Mass, Orbit & Flight Number, Payload & Orbit
- Source code : https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/5.%20jupyter-labs-eda-dataviz.ipynb



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.

Source Code : https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium MapsExplain why you added those objects

- 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; and
- Lines are used to indicate distances between two coordinates.

Source code : https://github.com/gijojozf/Project-IBM-Final/blob/80299270cc146ab4d90c2b9c432b47bf2a9f8201/6.%20lab_jupyter_launch_site_location.ipynb

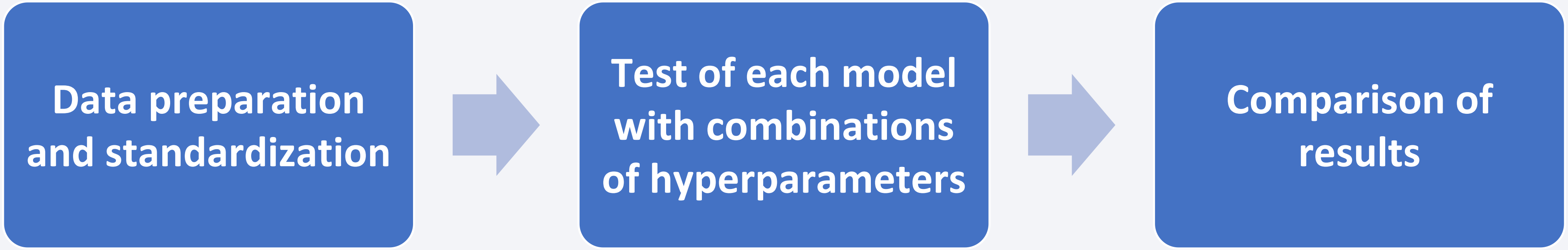
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source Code : https://github.com/gijojozf/Project-IBM-Final/blob/37abb1253af3a85bf5df320f2872b1556b91ff8a/7.%20spacex_dash_app.py

Predictive Analysis (Classification)

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



Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

Source code : https://github.com/gijojozf/Project-IBM-Final/blob/37abb1253af3a85bf5df320f2872b1556b91ff8a/8.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

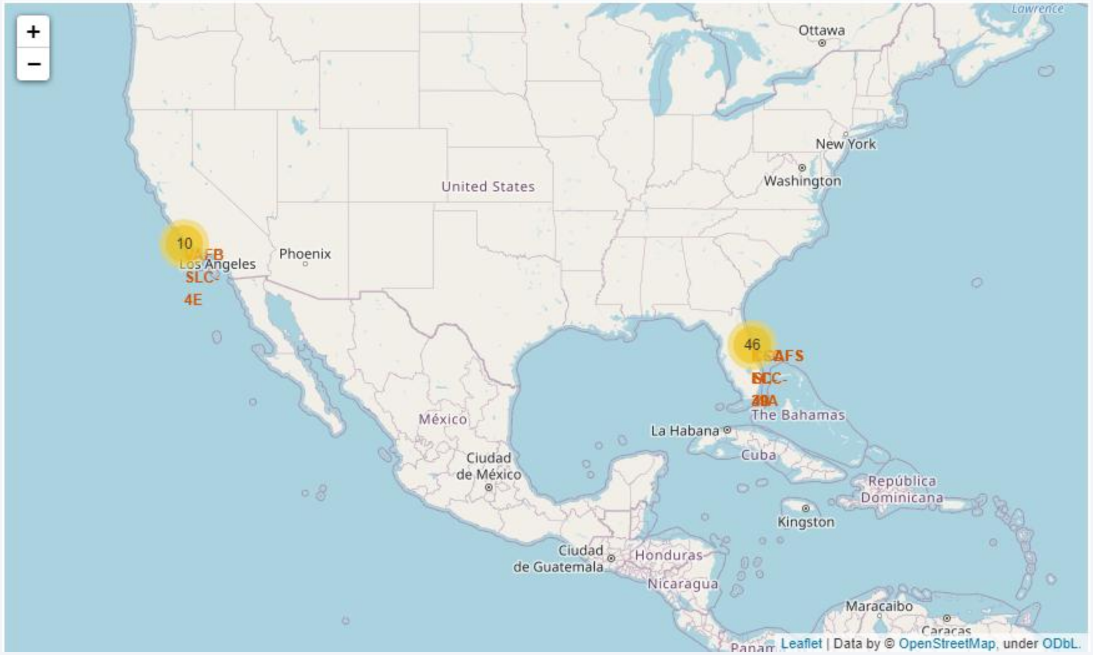
Results

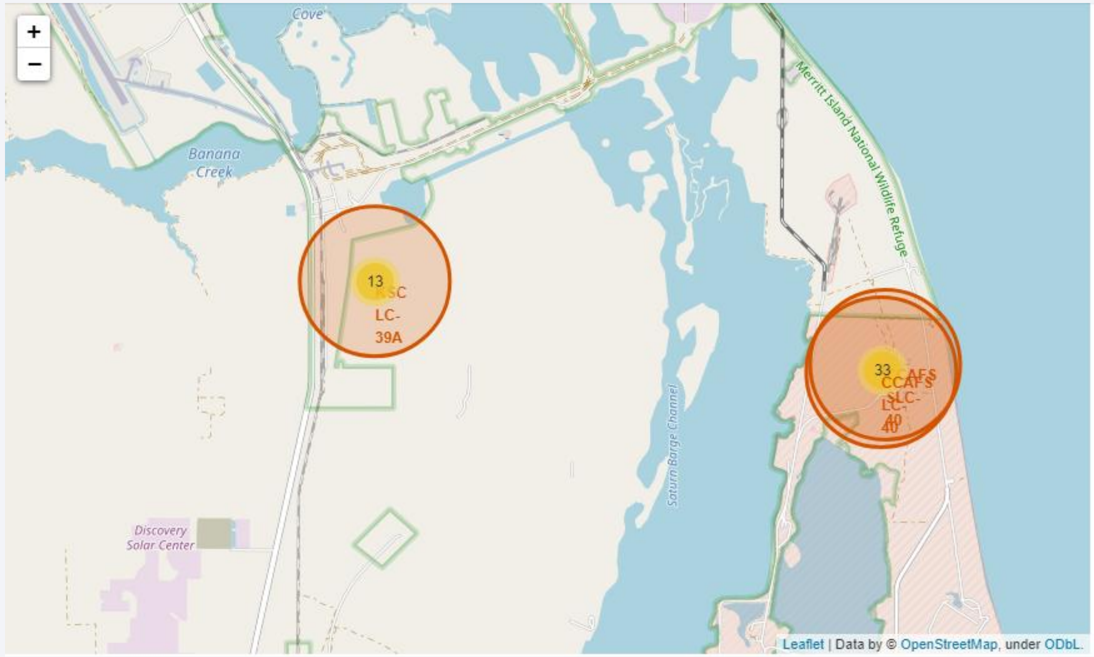
Exploratory data analysis results:

- 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 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

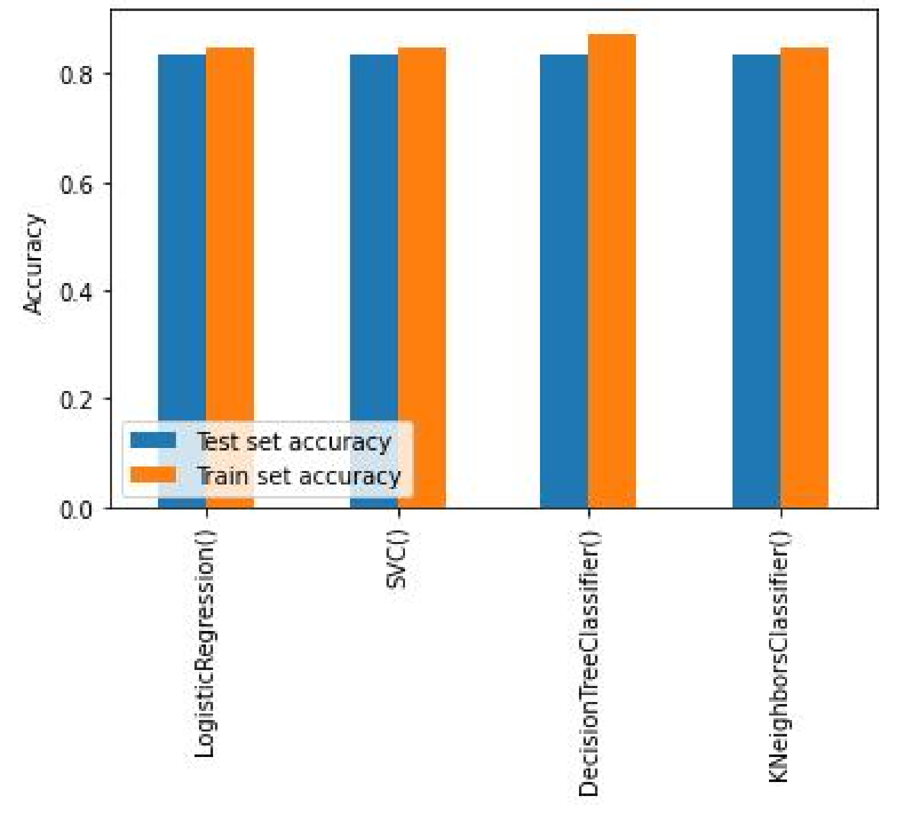
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.




+
−
Cove
Banana
Creek
Merritt Island National Wildlife Refuge
13
KSC
LC-
39A
33
CCAFS
SLC-
40
Saturn Barge Channel
Discovery
Solar Center
Leaflet | Data by © OpenStreetMap, under ODbL.

Results

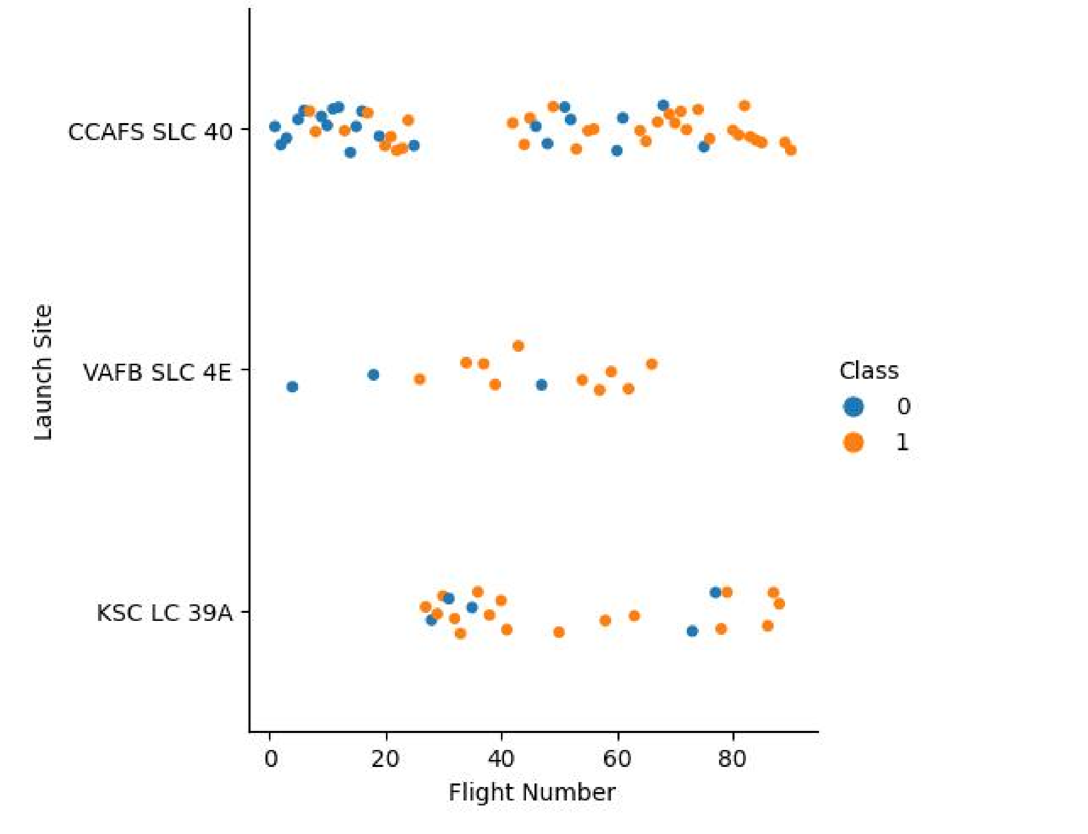
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 83%.



Section 2

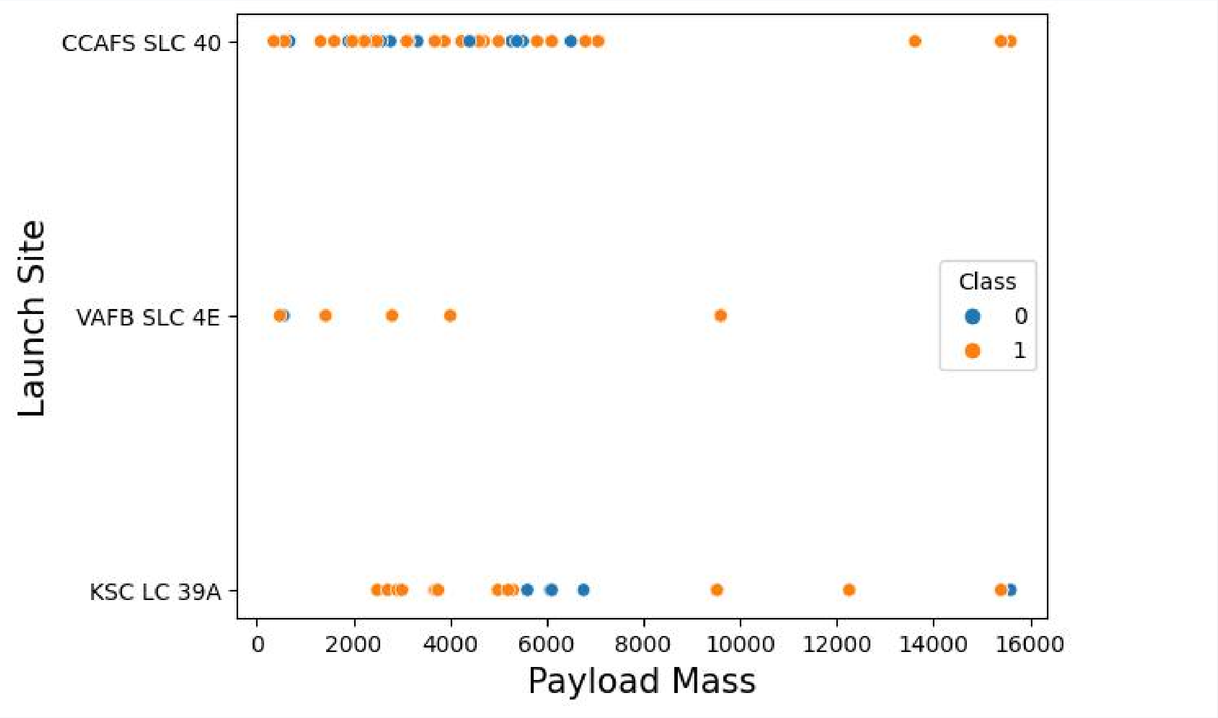
Insights drawn from EDA

Flight Number vs. Launch Site



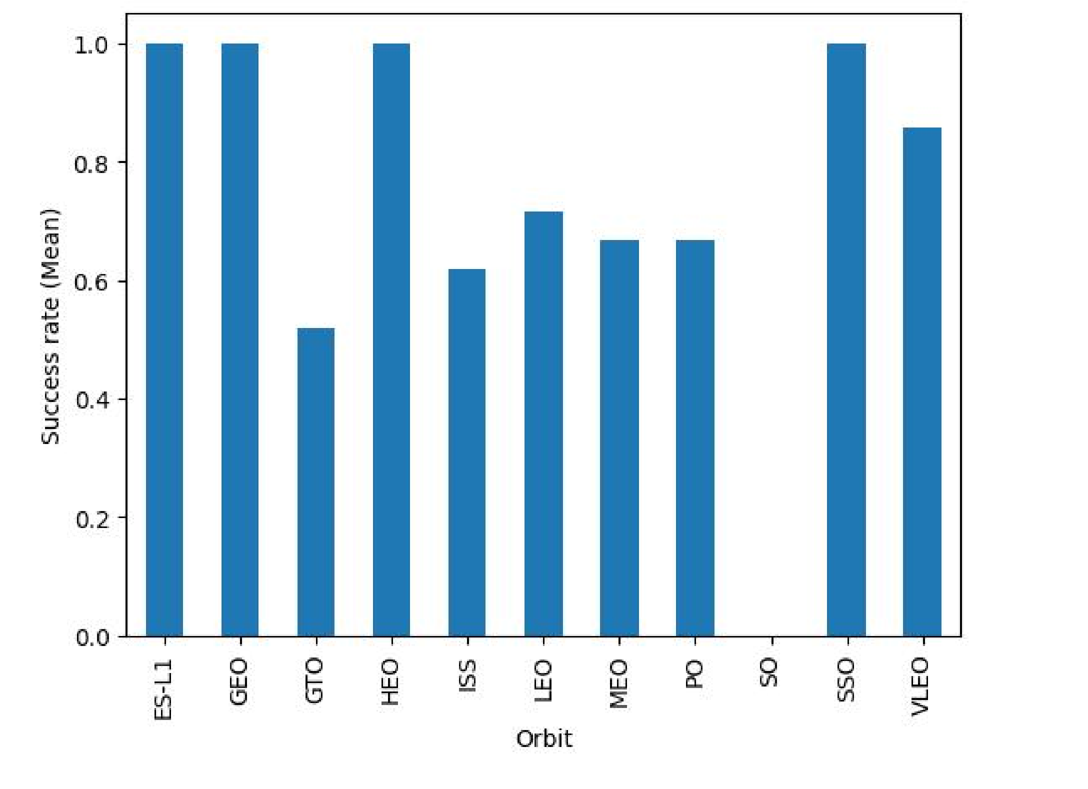
- According to the plot, it’s possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It’s also possible to see that the general success rate improved over time.

Payload vs. Launch Site



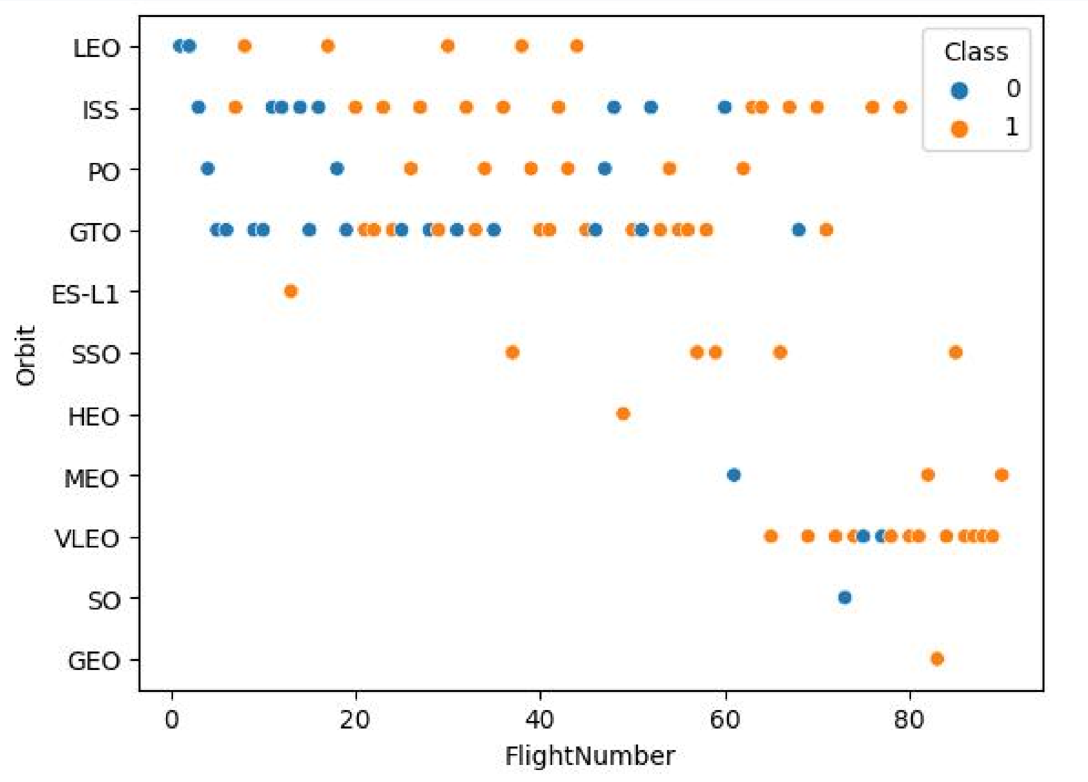
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Success Rate vs. Orbit Type



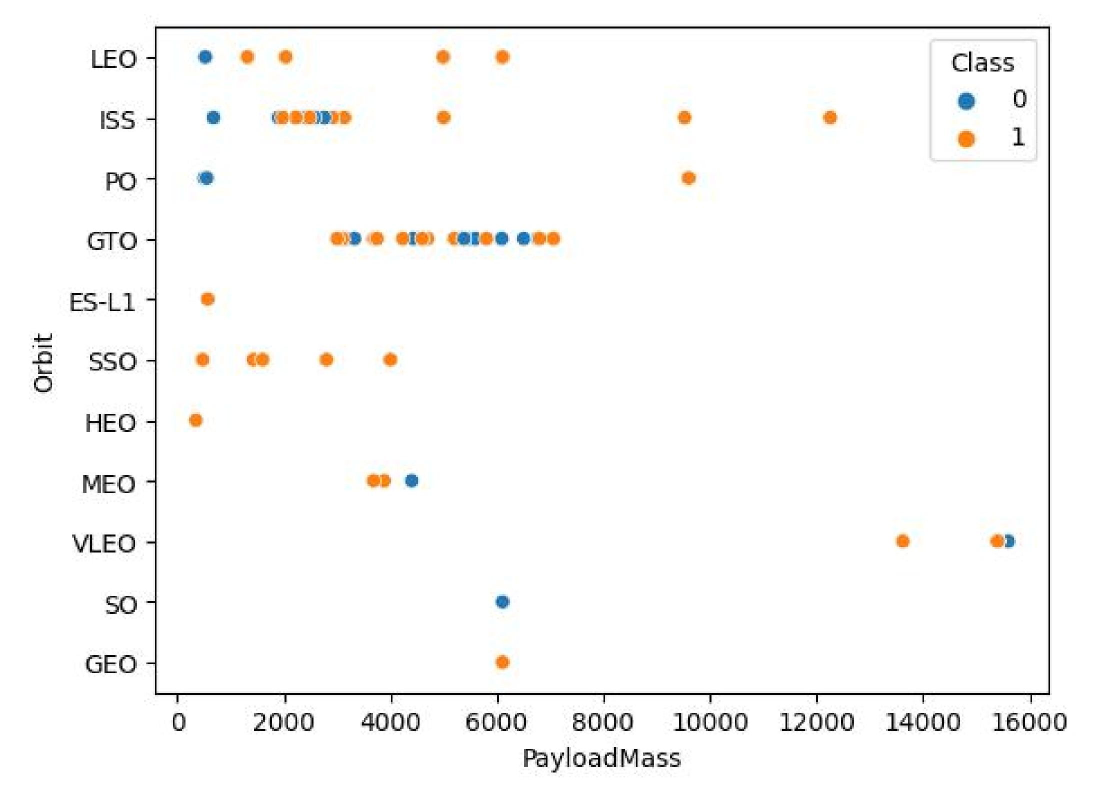
- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LEO (above 70%).

Flight Number vs. Orbit Type



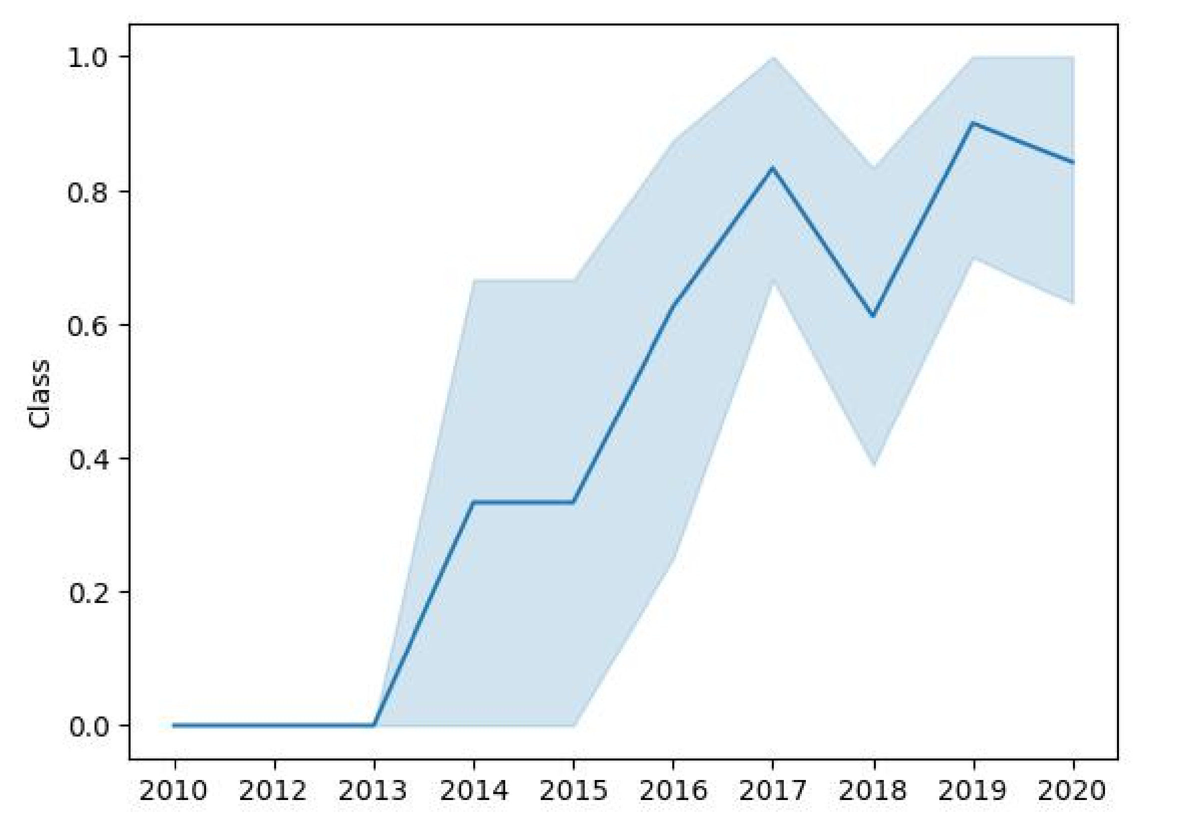
- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

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'

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

Total Payload Mass

- Total payload carried by boosters from NASA:

customer	payload_mass
NASA (CRS)	45596

- 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)
2534

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,534 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 B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- 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

- Boosters which have carried the maximum payload mass :

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

Booster Version
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- 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

MONTH	booster_version	launch_site	landing__outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- 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:

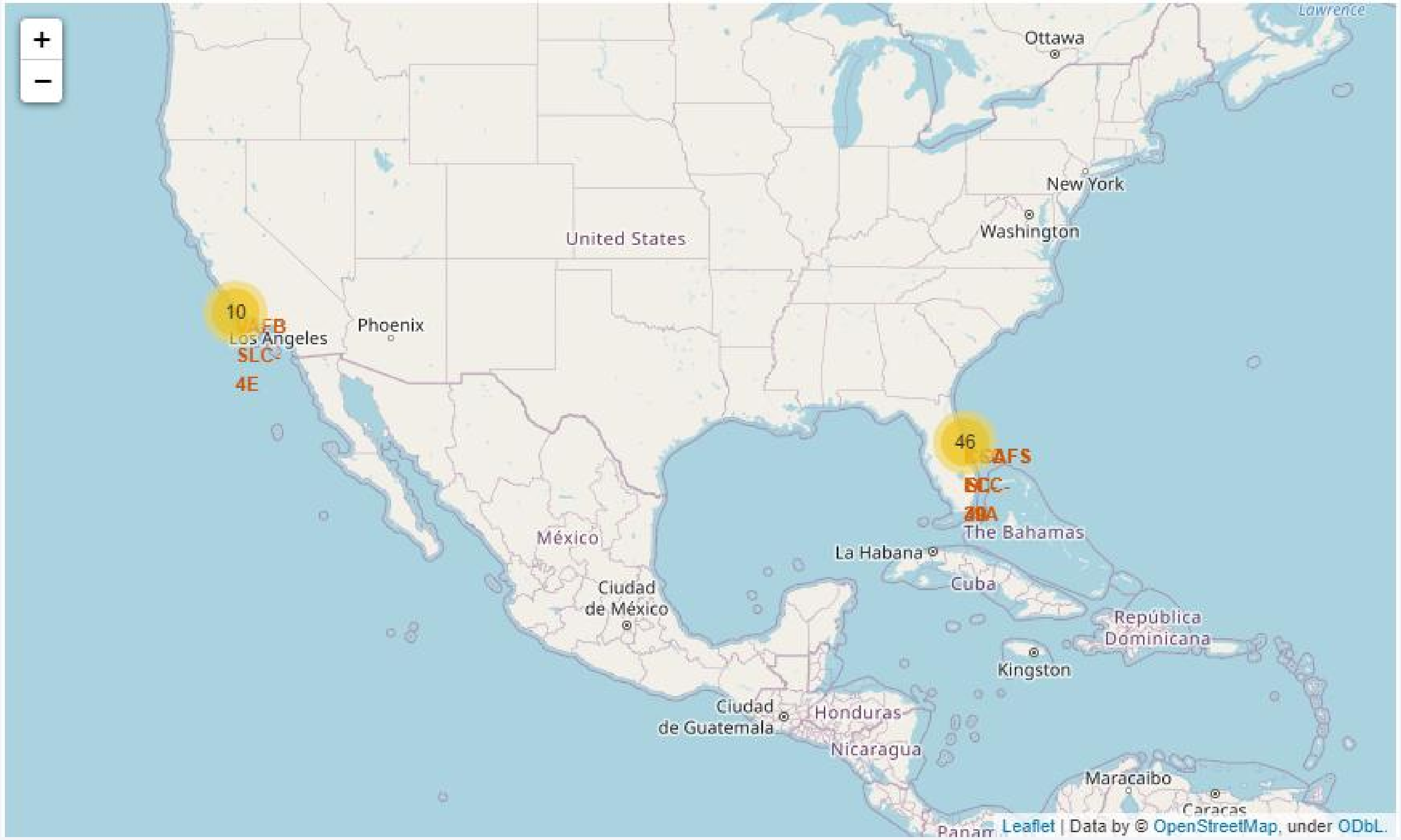
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 3

Launch Sites Proximities Analysis

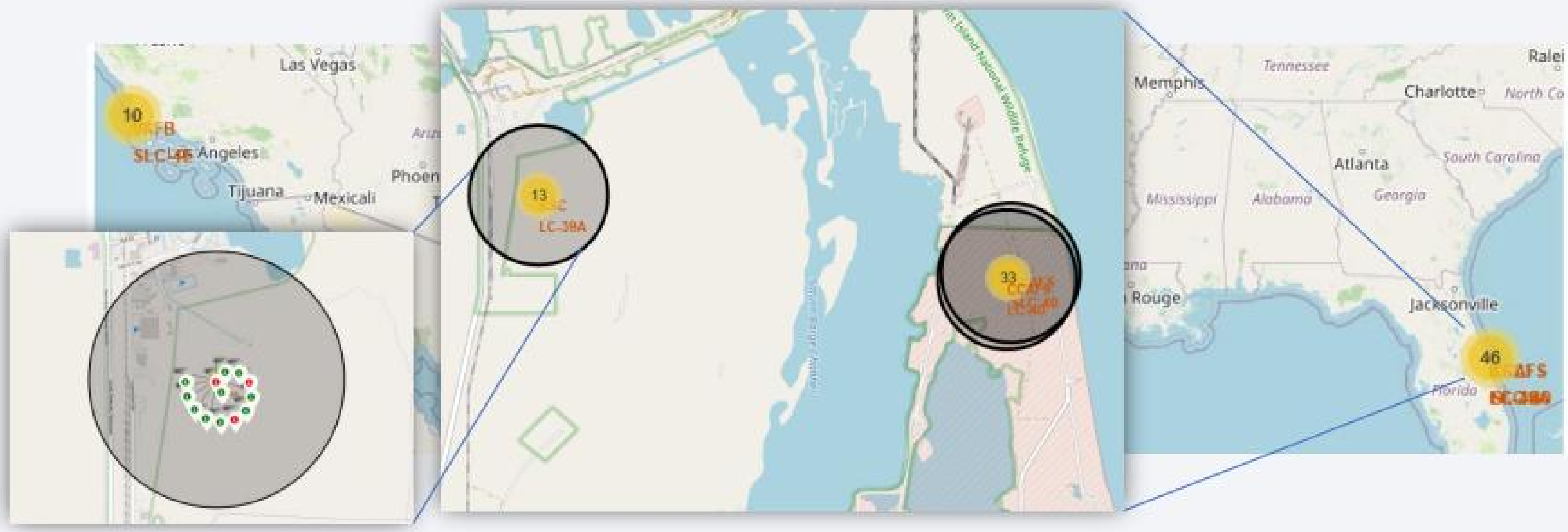
All launch sites



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

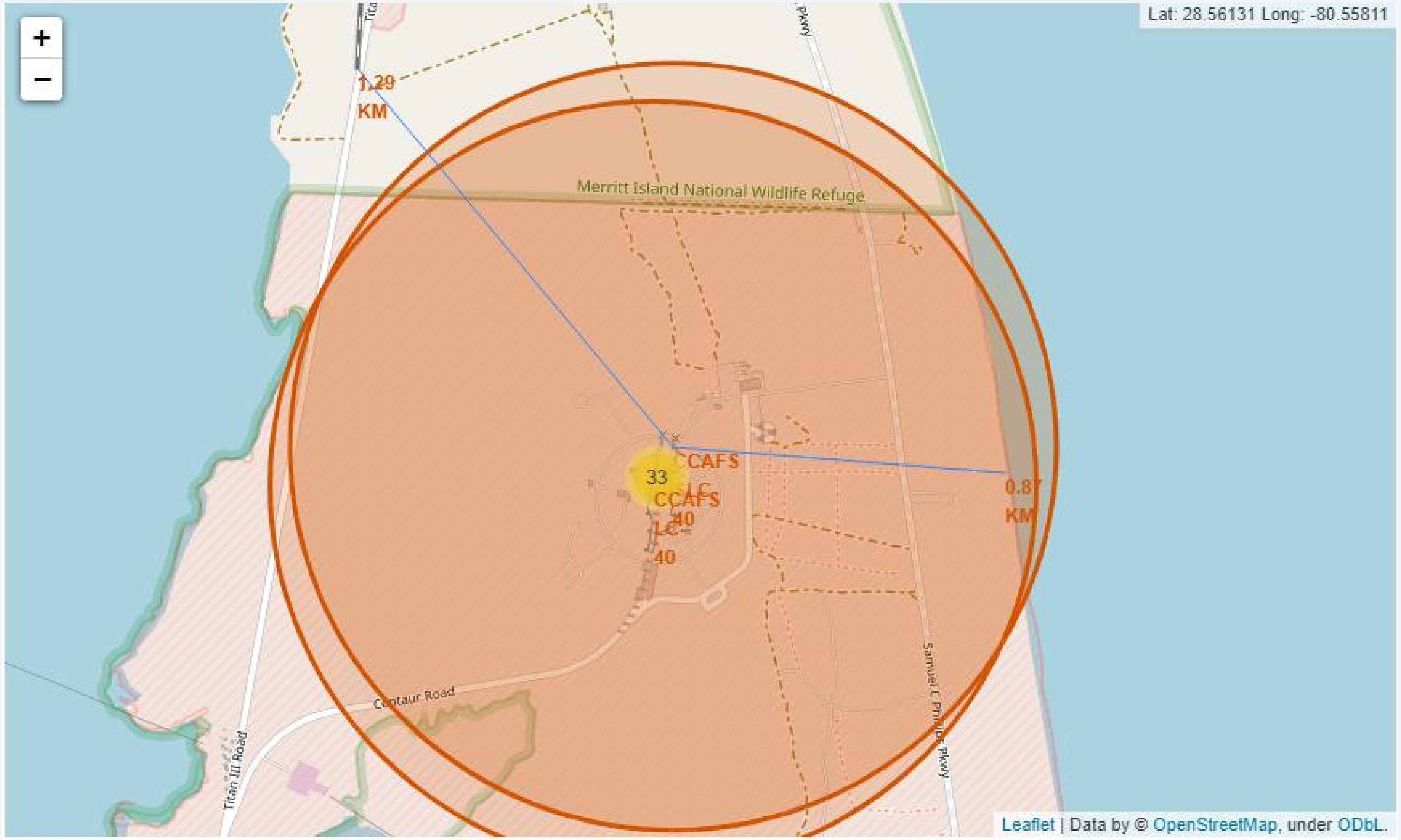
Launch Outcomes by Site

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



Green markers indicate successful and red ones indicate failure.

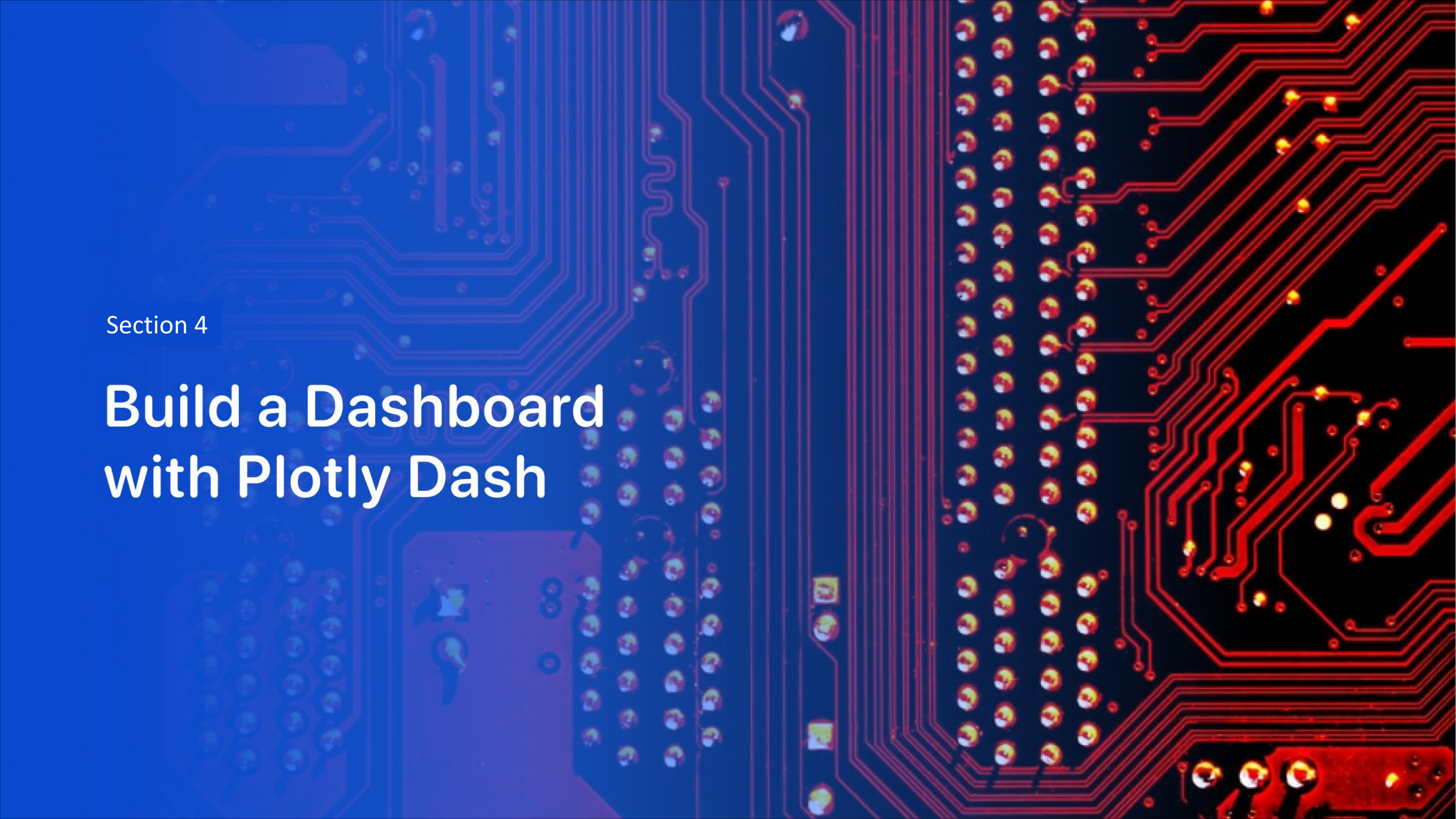
Logistics and Safety


+
−
Lat: 28.56131 Long: -80.55811
1.29 KM
Merritt Island National Wildlife Refuge
33
CCAFS LC-40
CCAFS LC-40
0.87 KM
Centaur Road
Titan III Road
Samuel C Phillips Pkwy
Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

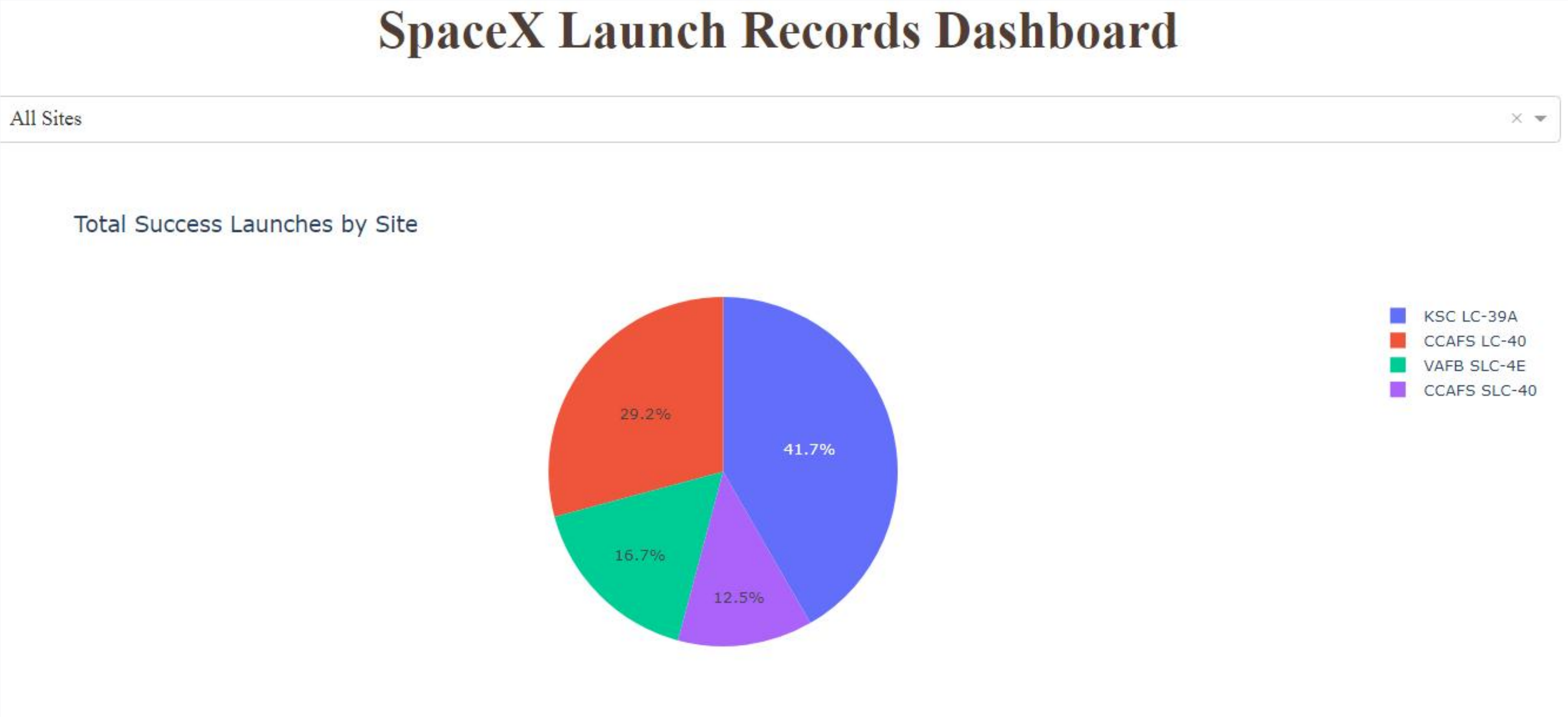
Launch site CCAFS SLC -40 has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 4

Build a Dashboard with Plotly Dash



Successful Launches by Site



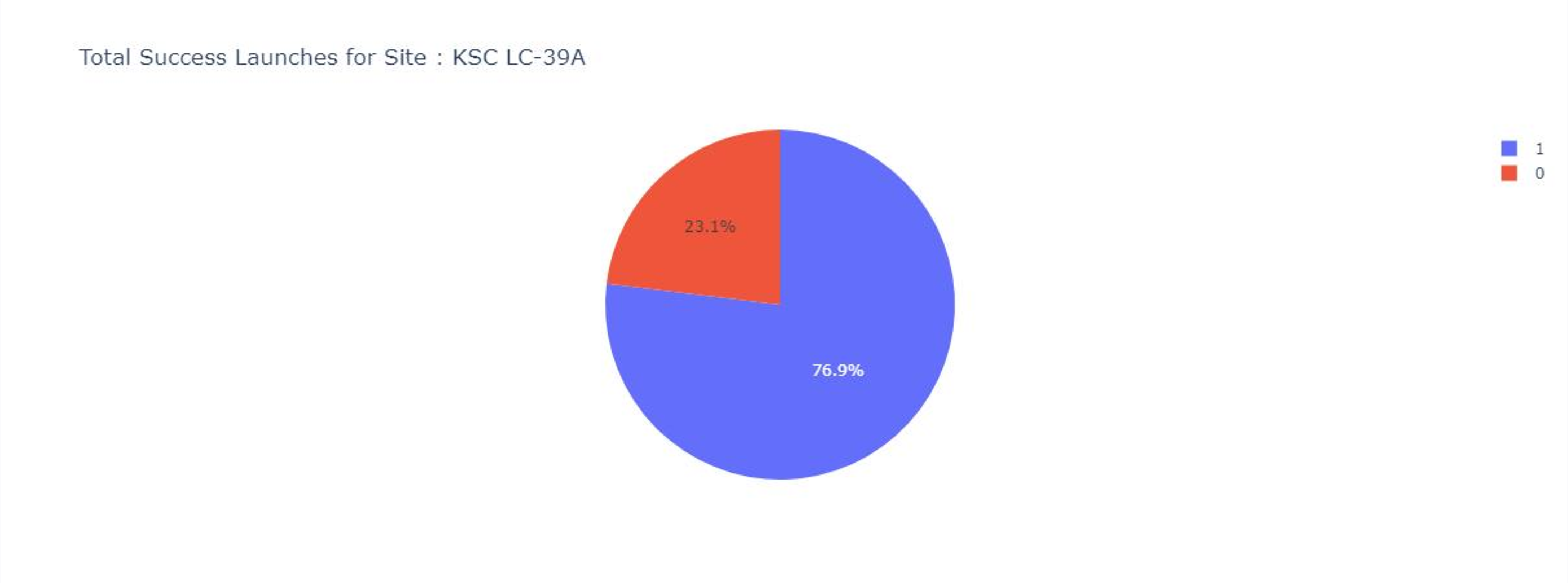
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

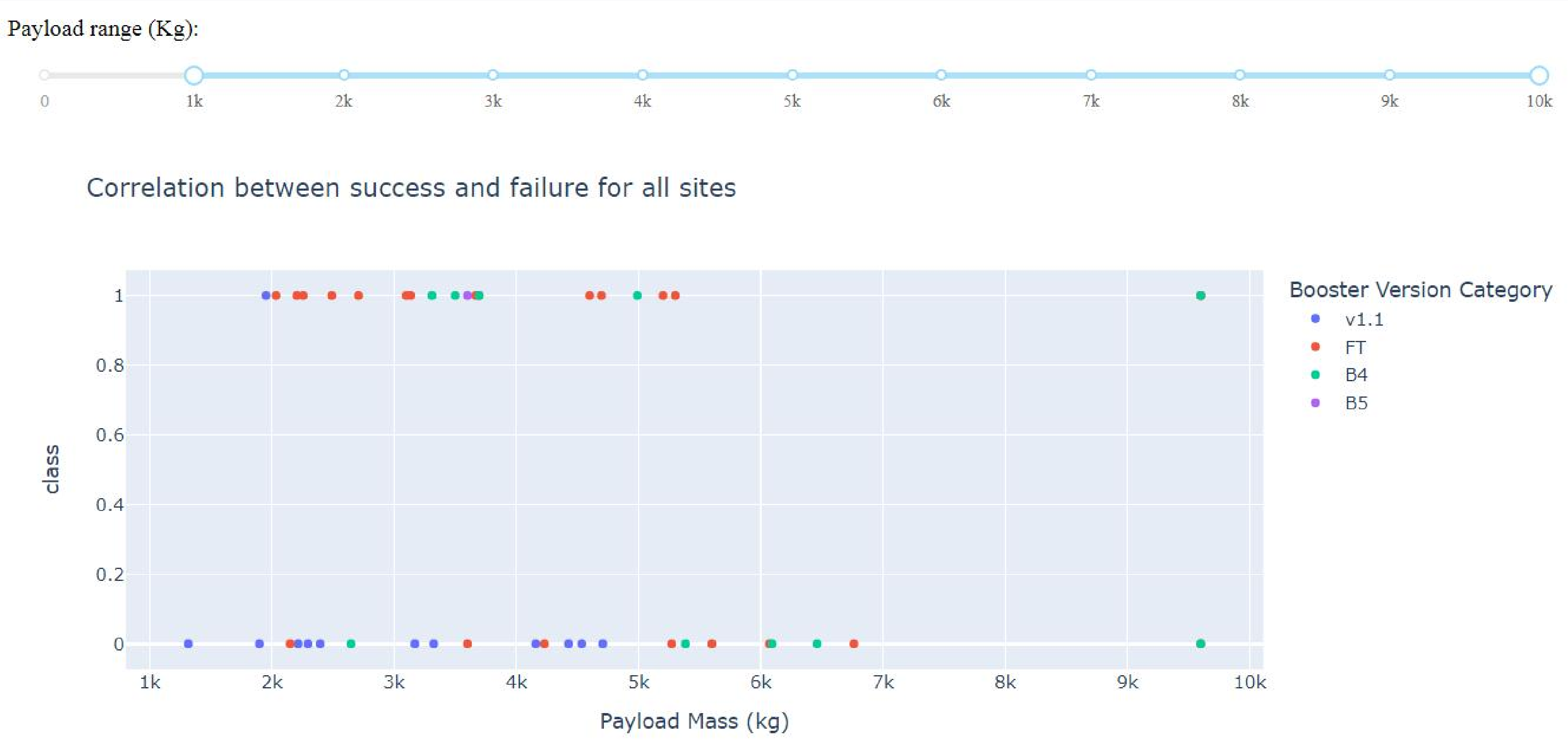
The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



76.9% of launches are successful in this site.

Payload vs. Launch Outcome



Payload range (Kg):

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

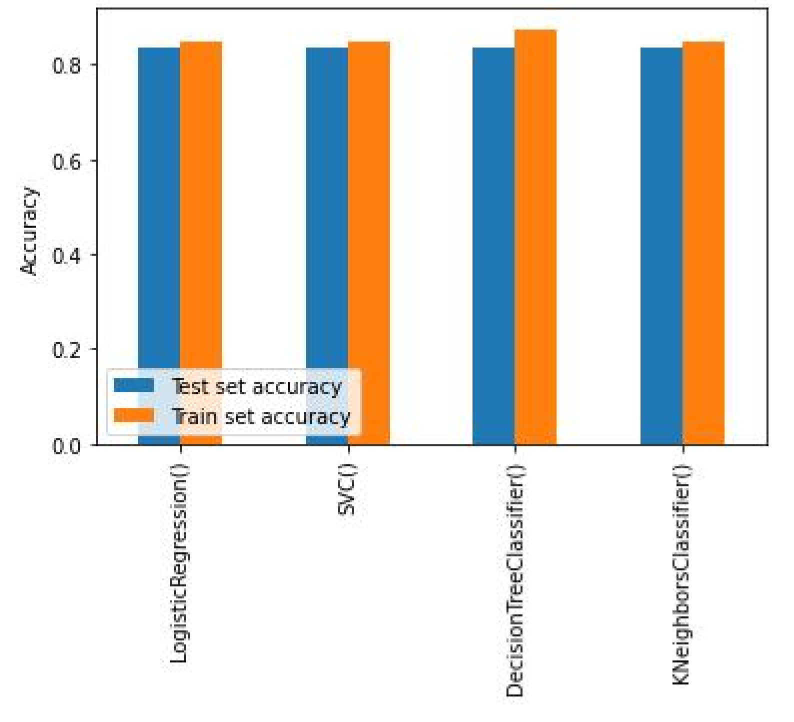
Section 5

Predictive Analysis (Classification)

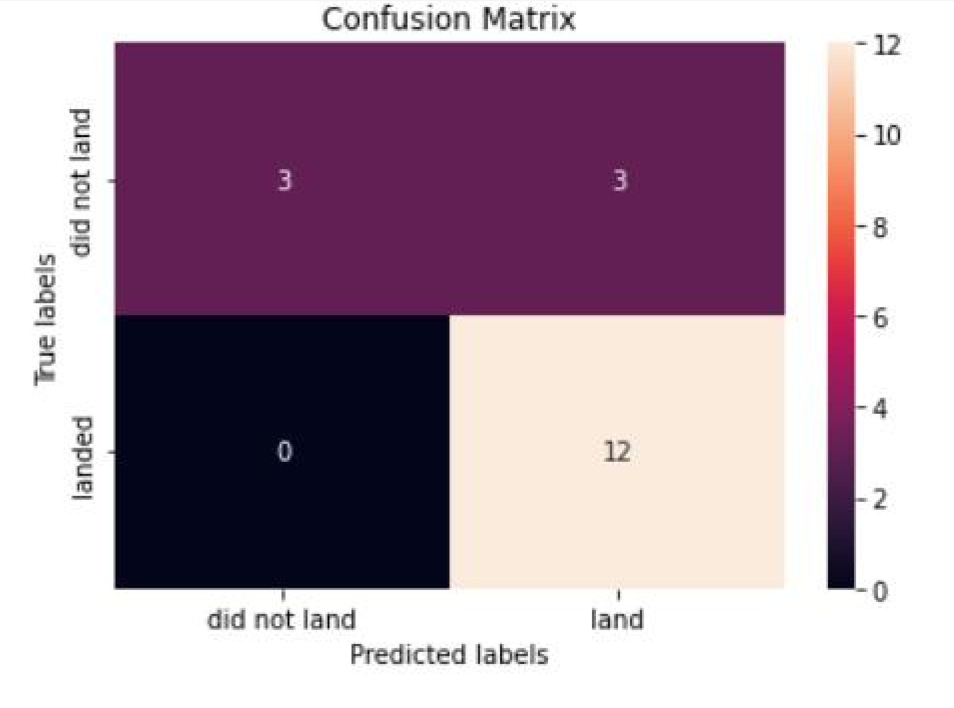
Classification Accuracy

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



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the bignumbers 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.

Appendix

- To standardize the result set, data of launch from **June 2010 to December 2020** was only used for analysis.
- For EDA with sql queries, as I was unable to load data of sql table using skills network lab, I created a connection to my db2 account of IBM.
- Reference to Github repository : https://github.com/gijojozf/Project-IBM-Final.git

Thank you!