IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Edwin Reyes
April 14th, 2025

Outline

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

Executive Summary

- *The following methodologies were used to collect and analyze the data:*
  - Data collection using SpaceX API and web scraping.
  - Exploratory Data Analysis (EDA); data wrangling, data visualization, and interactive visual analytics.
  - Machine Learning (ML) predictions.
- *Summary of all results:*
  - EDA allow to identify the site, payload range, and booster version of larges successful launches.
  - ML predictions showed the best model to determine if the first stage will land.

Introduction

- *Project background and context:*
  - Space Y is a company founded by Allon Musk focused on aerospace exploration.
  - The objective is to determine the competitive position of this company in the market.

- *Questions to solve:*
  - Is possible to predict a successful land of the first stage of rockets?
  - Which site has the largest successful launches?

Section 1

Methodology

Methodology

Executive Summary

- *Data collection methodology:*
  - The necessary data was obtained given the following two procedures:
    - API Python request process, and
    - Tabular information search using WebScraping.

- *Perform data wrangling:*
  - The data from API was normalized, a booster version was obtained, relevant data set was construct, and, finally, missing data were handled.
  - The data from WebScraping was extracted; the data frame was created and indexed.

Methodology

Executive Summary

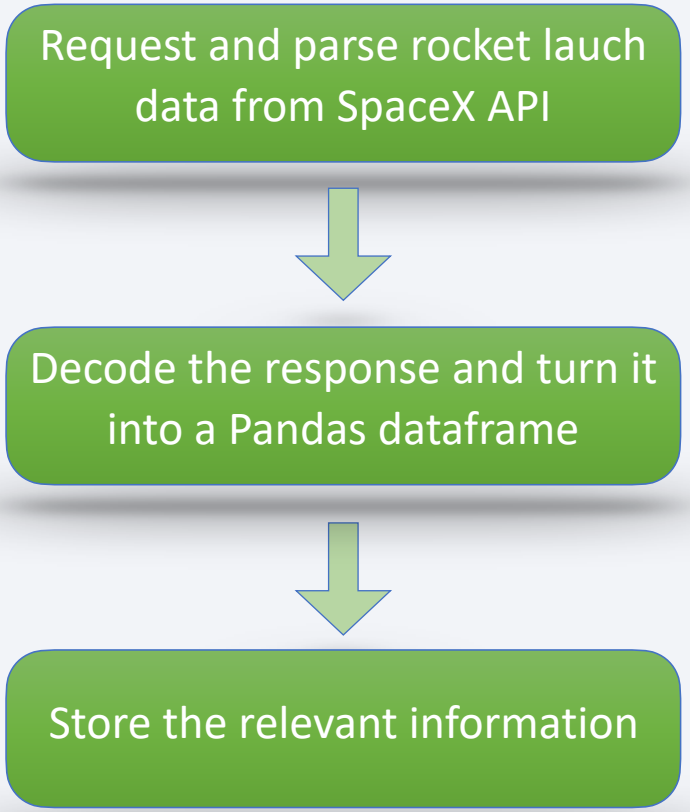
- *Perform interactive visual analytics using Folium and Plotly Dash*
- *Perform predictive analysis using classification models*
  - Using the information obtained, the numerical data of interest were standardized, and the training and test sets were split.
  - The optimal set of hyperparameters for SVM, Classification Trees, KNN, and Logistic Regression were determined to find the model that best performs the desired prediction.

Data Collection

- Sources:
  - The necessary data was obtained from:
    - Space X API ("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=102768692")

Data Collection – SpaceX API

- SpaceX API is a public source where data was obtained.
- Given the next flowchart data collection procedure can be achieved.

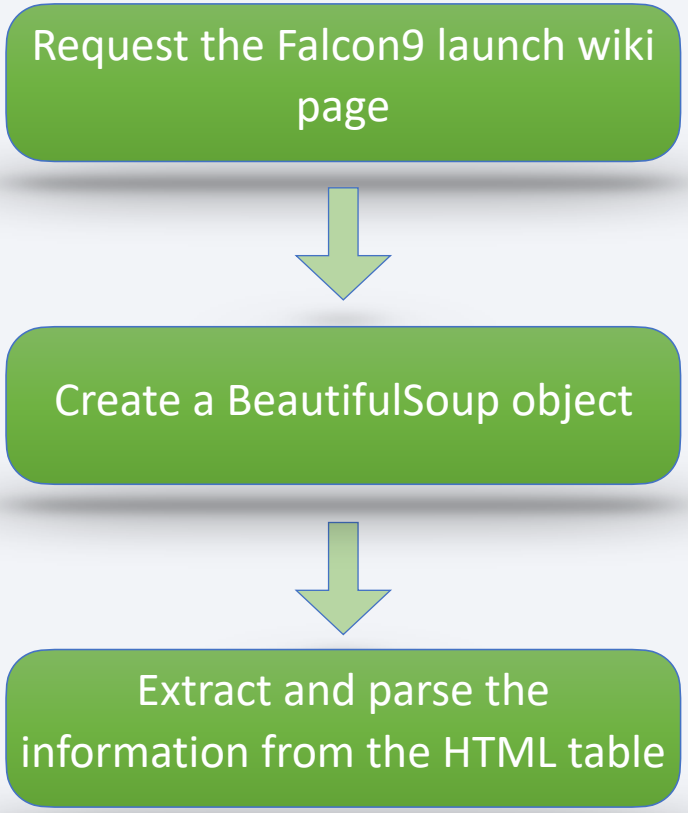


Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/Data%20collection%20API.ipynb

Data Collection - Scraping

- Relevant information can also be obtained from Wikipedia.

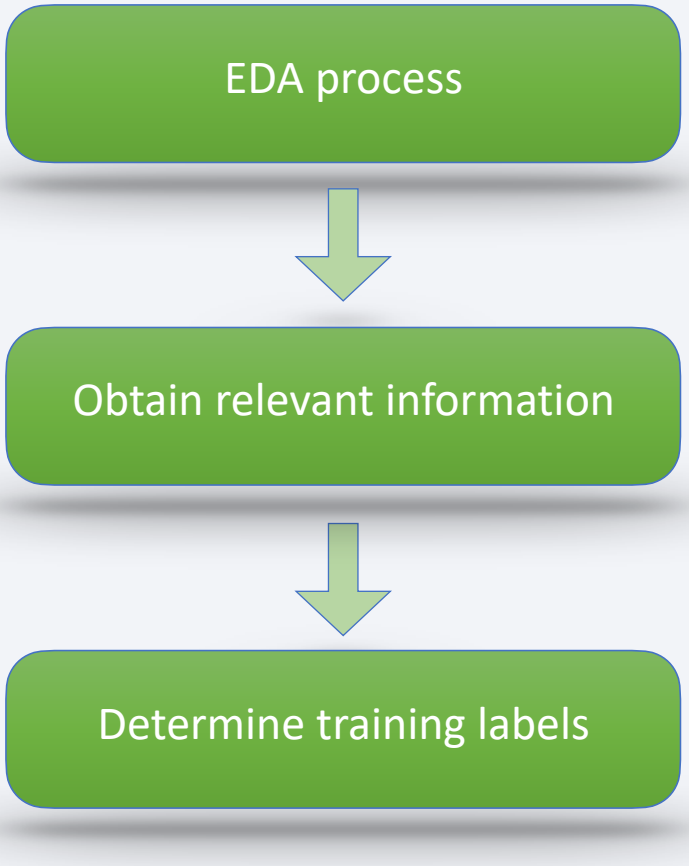
- Data is requested, extracted and parsed according to the next flowchart:



Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/Data%20collection%20WebScraping.ipynb

Data Wrangling

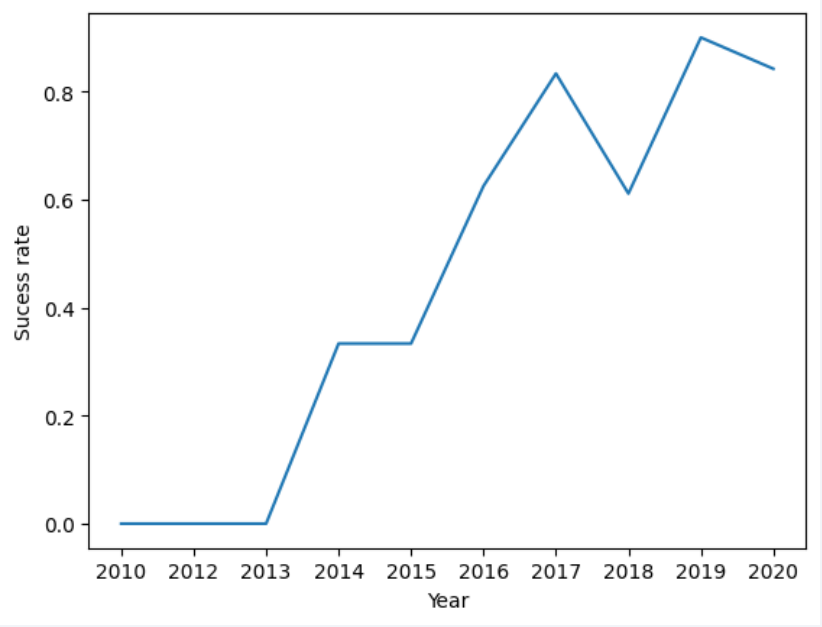
- First, an exploratory Data Analysis was performed.
- Relevant number of events were obtained.
- Finally, a landing outcome label was created for future training.



Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/Data%20wrangling.ipynb

EDA with Data Visualization

- To explore relationships in data scatter, bar and line charts were used as visual tools.
- For example:



Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/EDA%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Display of the unique launch sites and records of with specific content.
  - Total mass carried by booster launched by NASA (CRS) and average payload mass carried by booster version F9 v1.1.
  - Date of the first successful landing outcome in ground pad was achieved and information about month, booster version, launch site and landing outcome in 2015..
  - List of the names of the boosters which have success in drone ship with payload mass in the range (4000 kg, 6000 kg).
  - Total number of successful and failure mission outcomes.
  - Booster versions that have carried the maximum payload mass.
  - Rank of landing outcomes between 2010 and 2017.

Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/EDA%20using%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and clusters were created and added to a folium map.
  - Markers and circles were used to identify launch sites.
  - Clusters represent groups of events.
  - Lines indicate distances between relevant positions.

Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/Interactive%20Visual%20Analytics%20Folium.ipynb

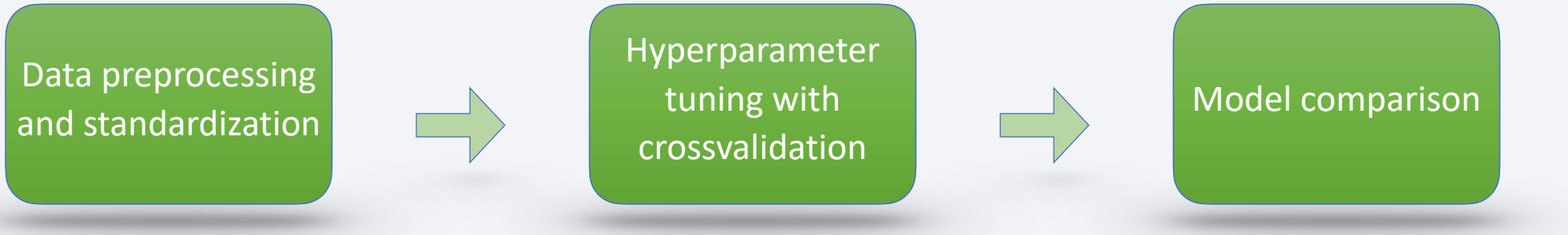
Build a Dashboard with Plotly Dash

- *A dashboard was designed with the following graphics:*
  - Total and partial (by launch site) percentage of outcomes shown in a pie chart.
  - Outcome (class) vs payload mass given a scatter plot with an adjustable interval in the domain.
- *These graphics allow you to quickly obtain an understanding of the information and gain insight from data.*

Code: https://github.com/MartinReyesS/Applied-Data-Science-Capstone/blob/2767bc31a549aae188fb01c6b252e0b5bd8b6755/Interactive%20Visual%20Analytics%20Dash.py

Predictive Analysis (Classification)

- Four models were trained, performed and compared: logistic regression, support vector machine, decision tree and K nearest neighbors.
- GridSearchCV were implemented for hyperparameter tuning with crossvalidation.



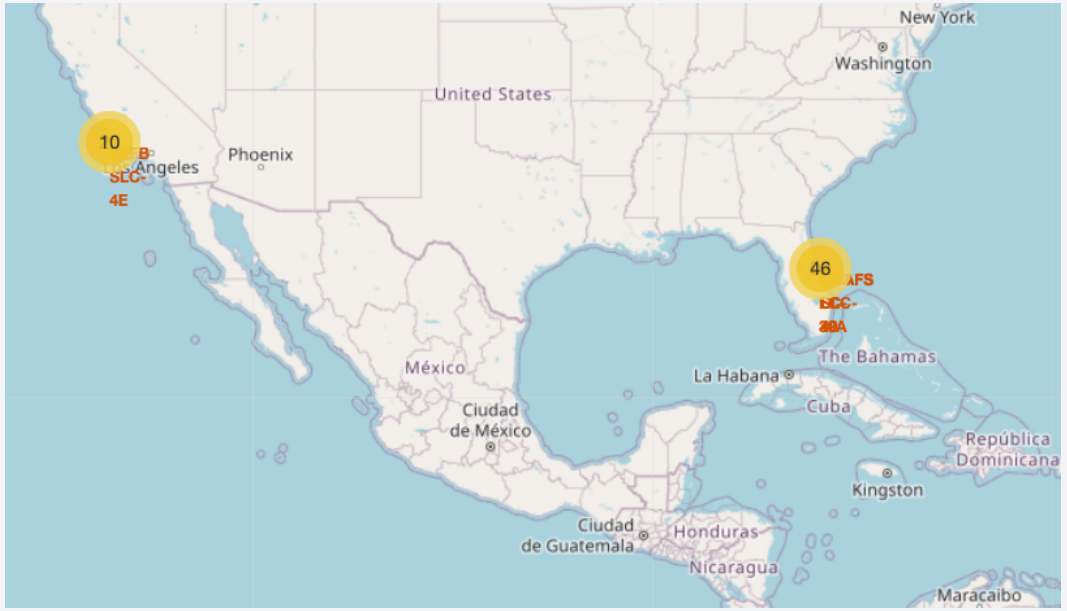
Results

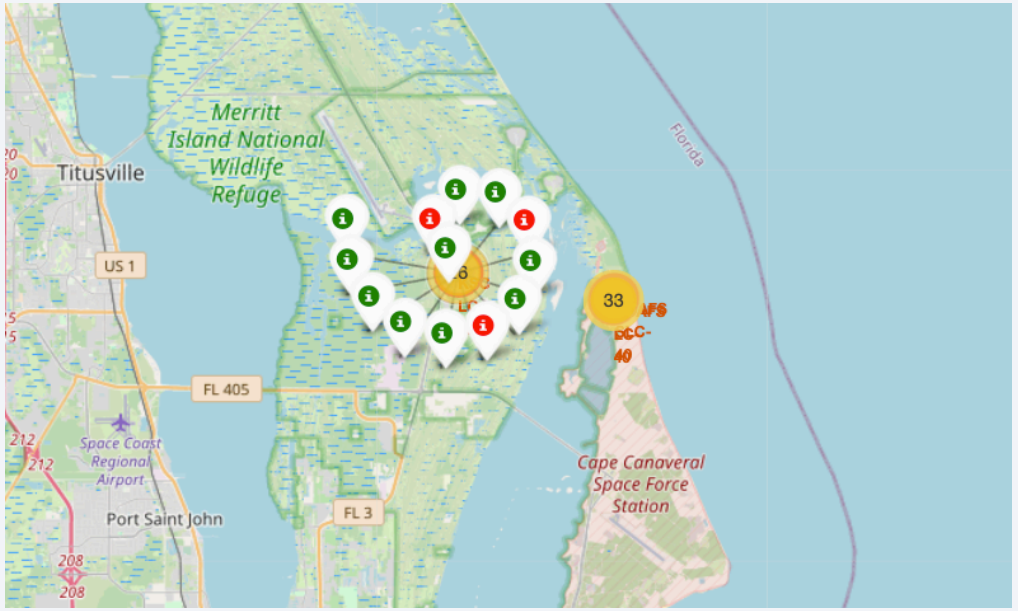
- *Exploratory data analysis results:*

  - For VAFB-SLC launchsite there are no rockets launched for heavy payload mass.
  - For LEO orbit the success rate is related with the number of flights.
  - In general, success rate is increasing with time.

Results

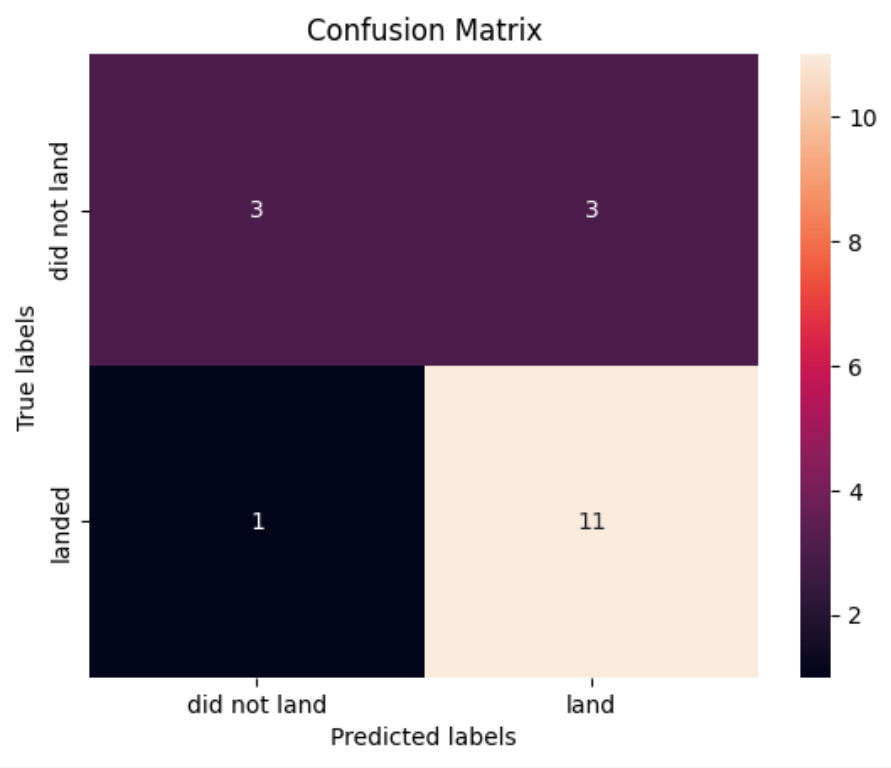
- *Interactive analytics: it is possible to identify the launch sites with outcome information.*





Results

- *Predictive analysis results:*
  - The decision tree classifier was the most accurate model to study the present dataset.

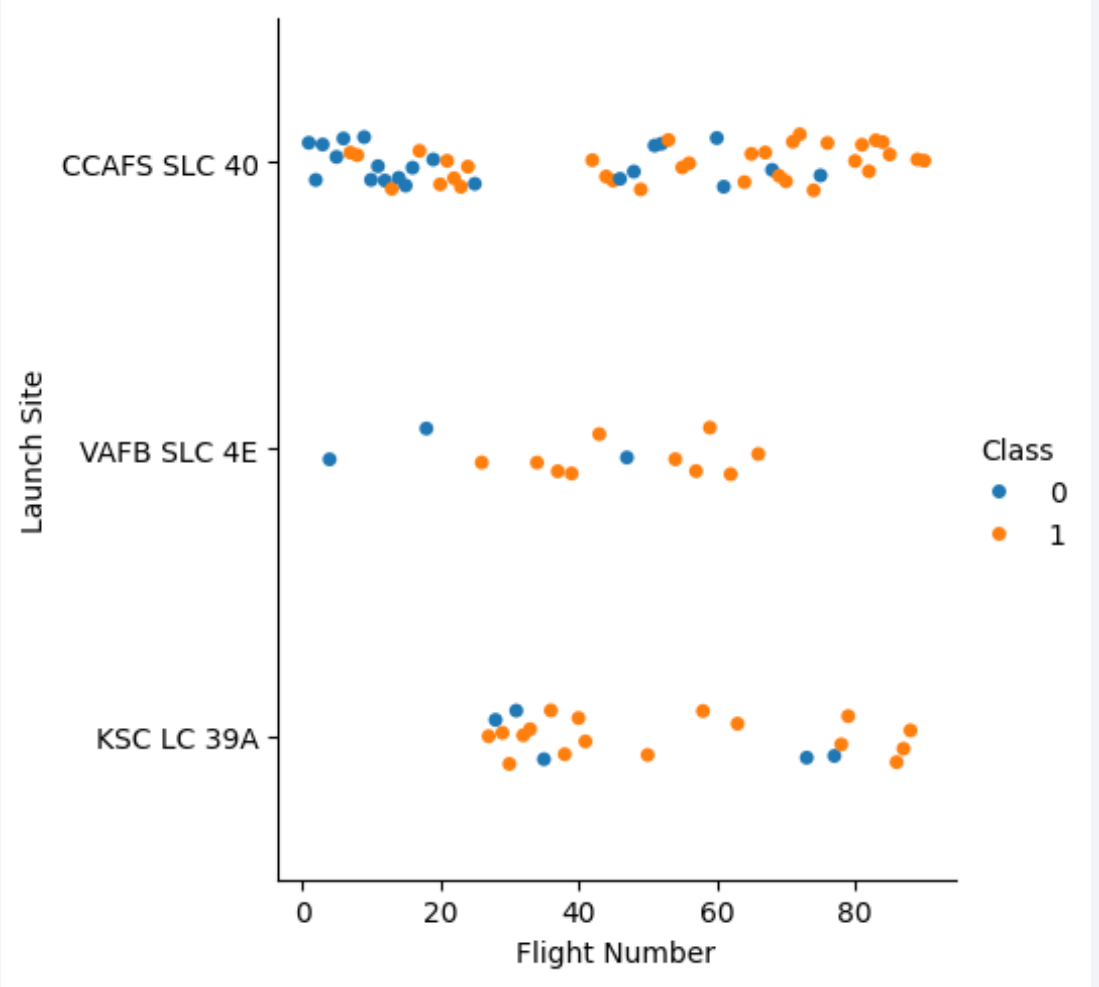


Section 2

Insights drawn from EDA

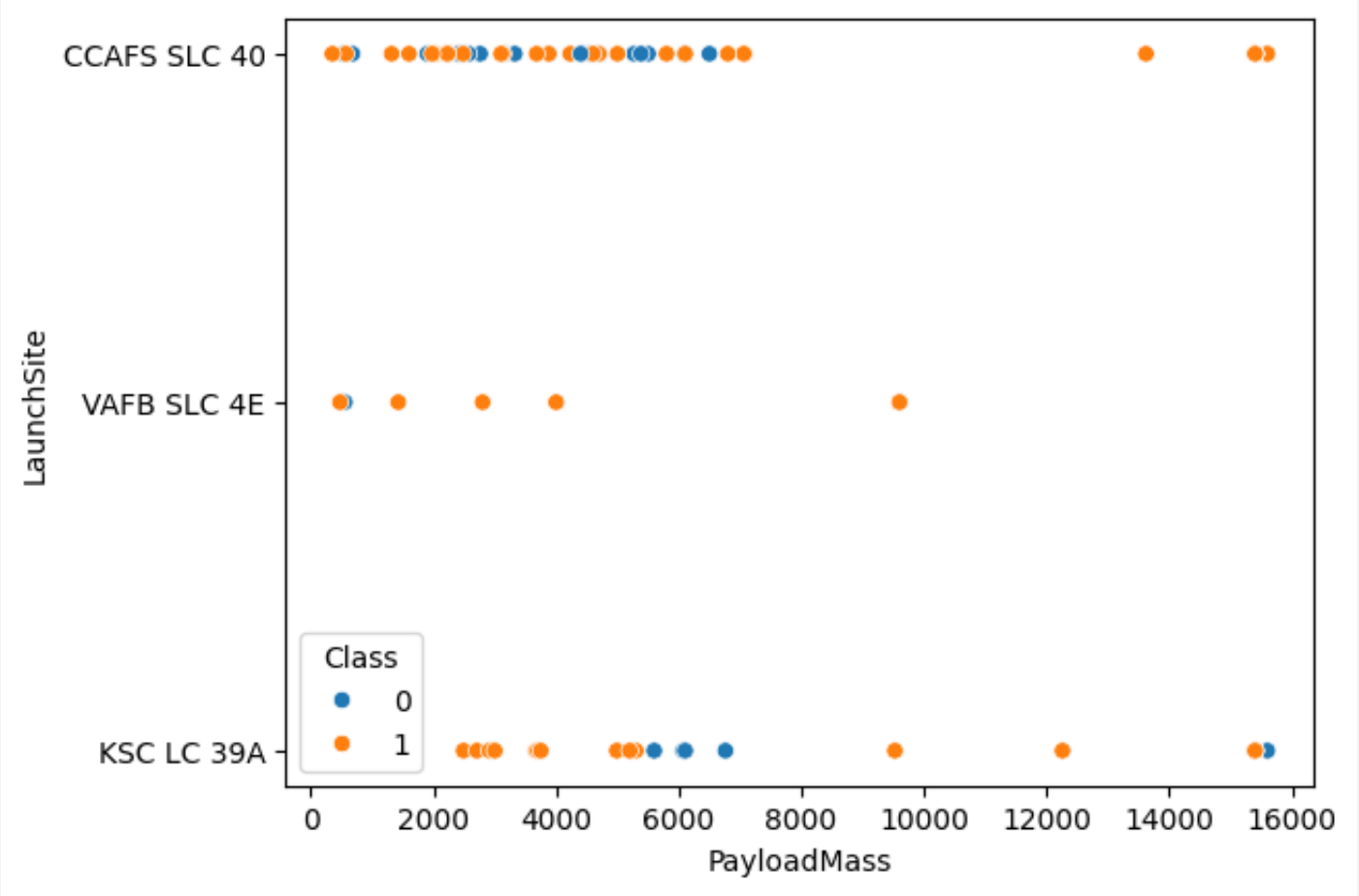
Flight Number vs. Launch Site

- In general, we observe a success rate improvement over time (Flight Number)
- The latest releases are concentrated in CCAFS SLC 40 and KSC LC 39A with a favorable balance.



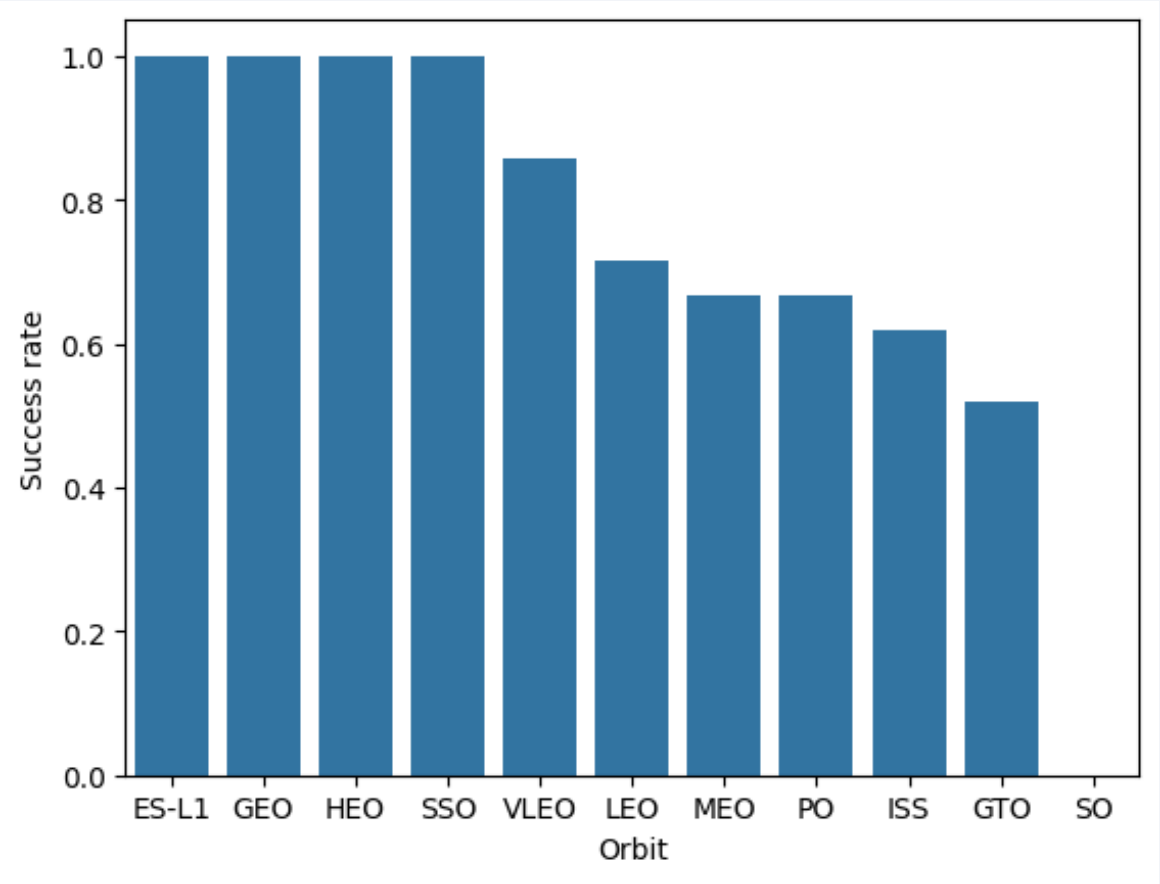
Payload vs. Launch Site

- Throws over 8000 kg have a very high success rate.
- Launches above 10,000kg appear to be impossible at VAFB SLC 4E.



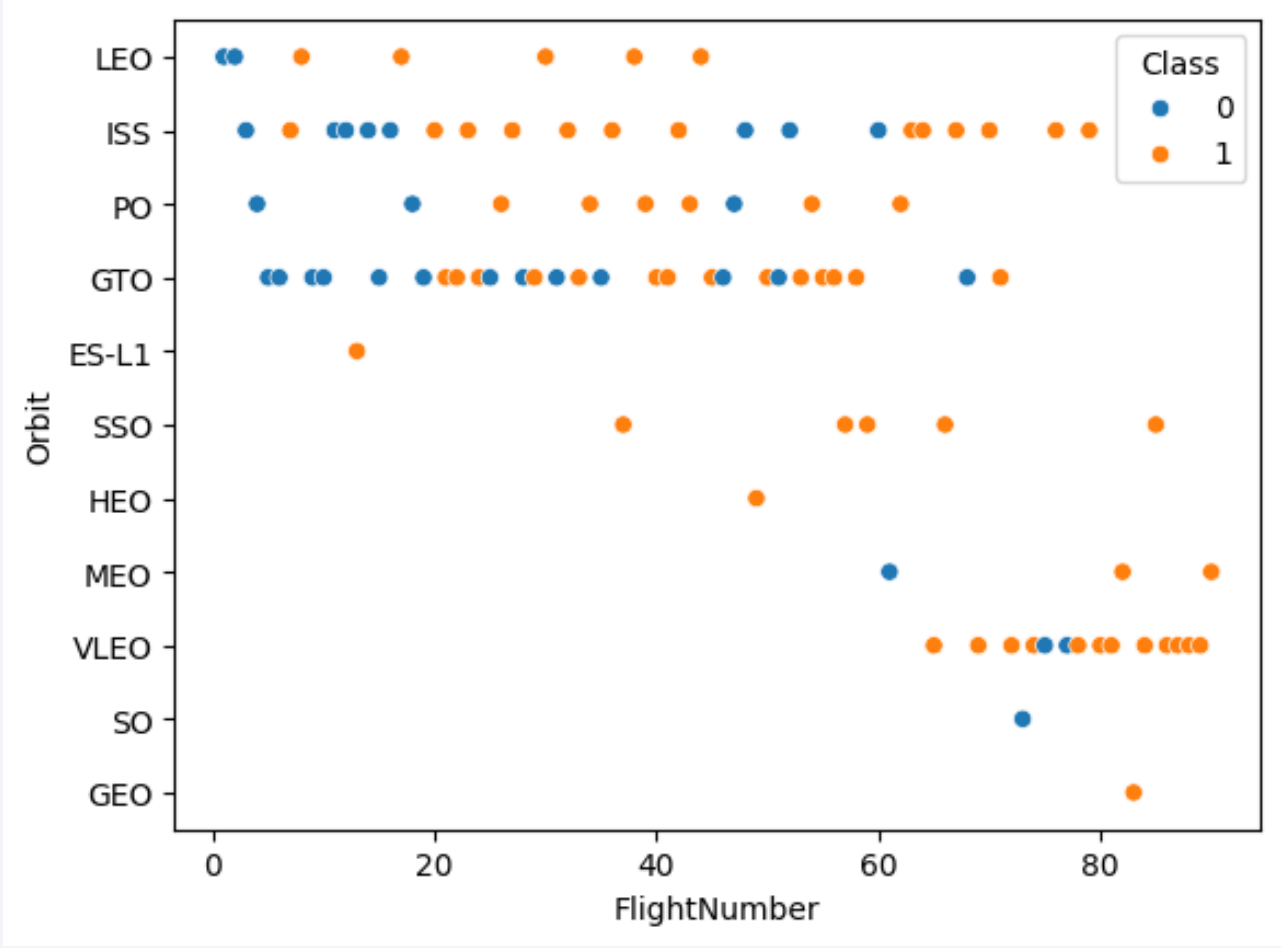
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have a perfect success ratio.
- The SO orbit has no successful launches.



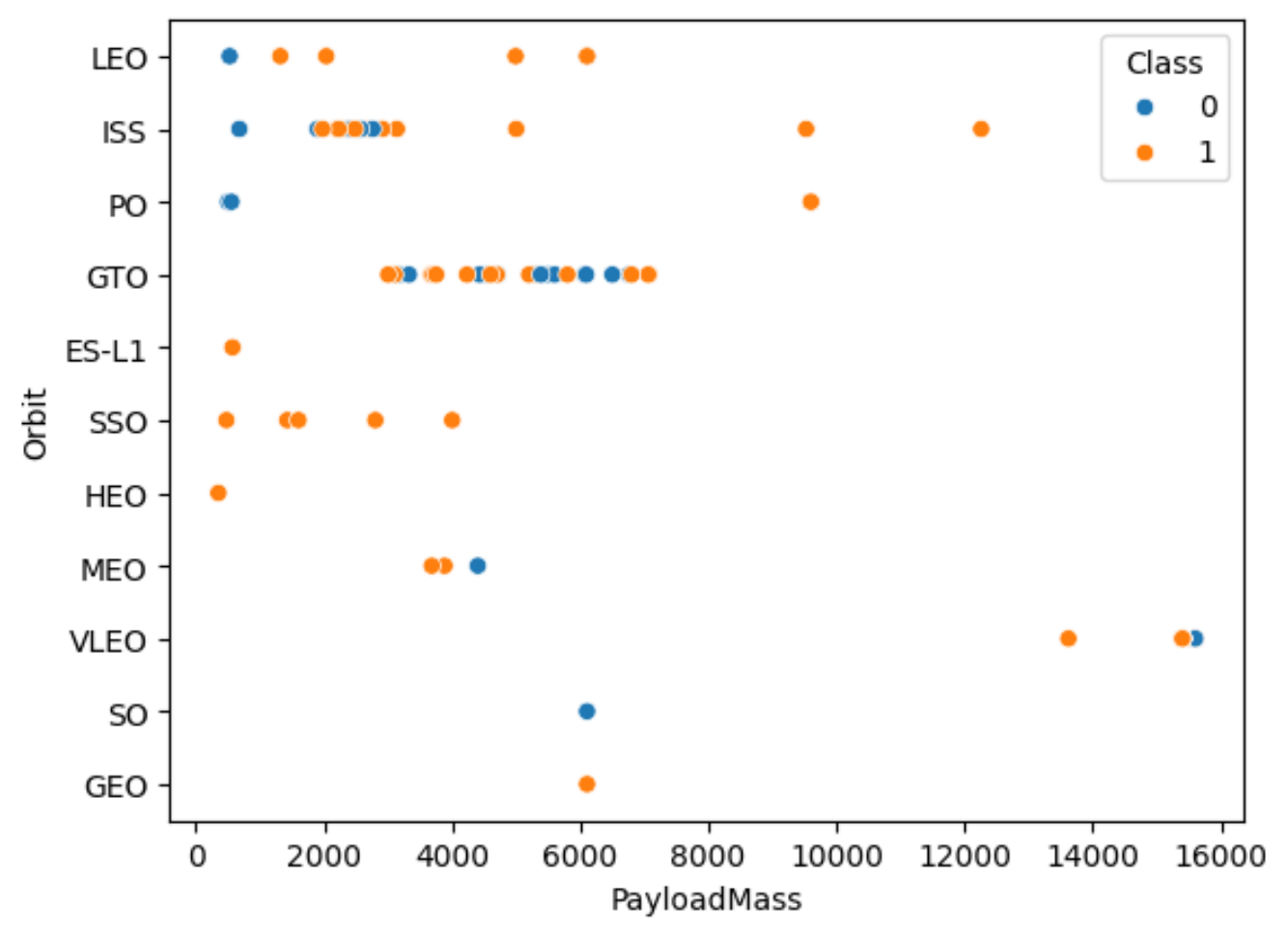
Flight Number vs. Orbit Type

- Except for the SO orbit, in all orbits an increase in successful launches is observed as a function of the Launch Number.
- ES-L1, SSO and HEO are seen as orbits with very safe launches.



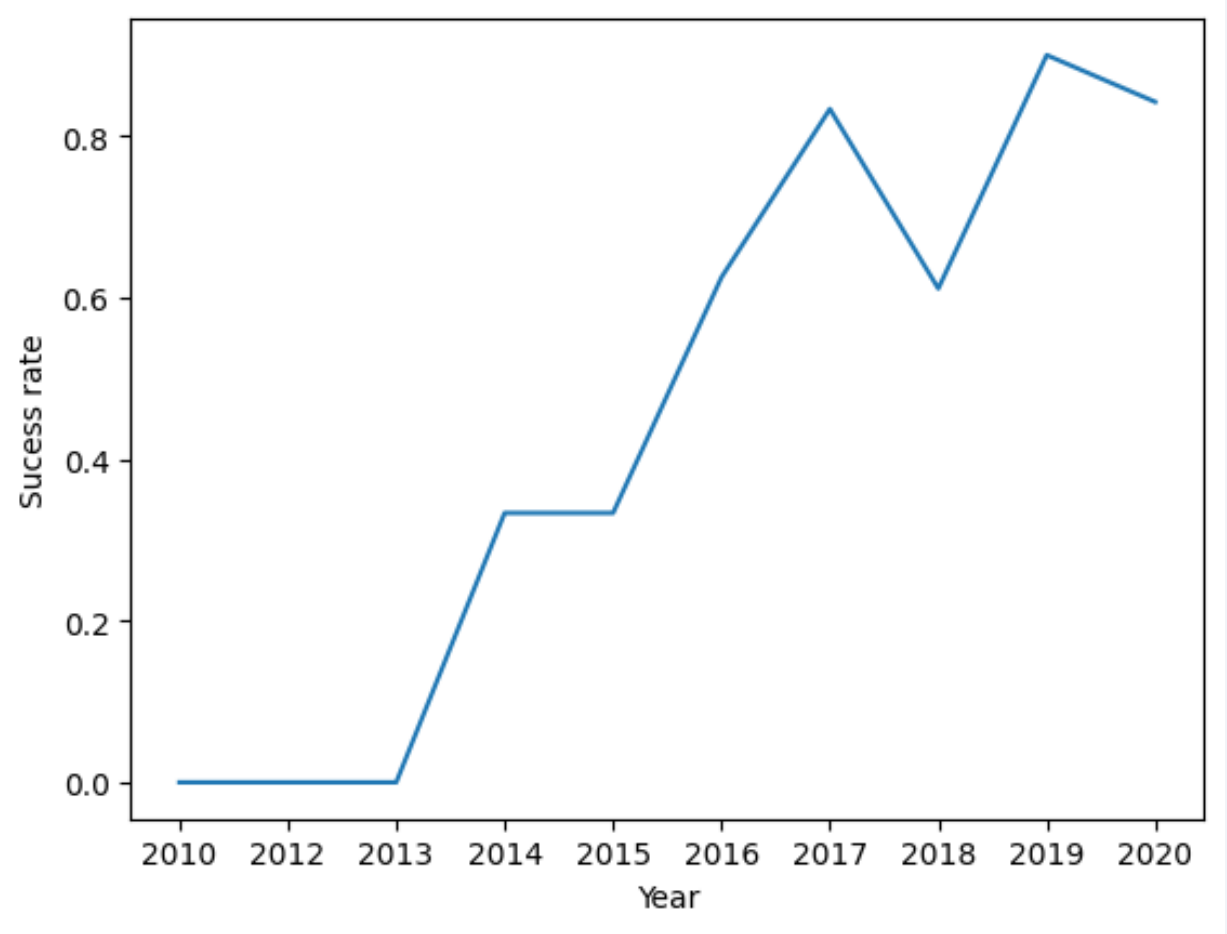
Payload vs. Orbit Type

- With heavy Payload Mass the successful landing rate is concentrated in LEO, ISS and PO.
- For GTO is not possible to distinguish a behavior related with the Payload Mass.



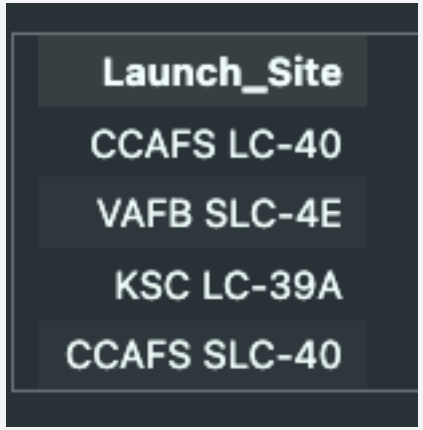
Launch Success Yearly Trend

- As we commented before, the success rate is highly related with time.
- Although, around 2018 there exist a negative slope (negative trend).



All Launch Site Names

- According to the data, there are four Launch Sites:



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

- The result is obtained given the Distinct values for “Launch_Site”.

Launch Site Names Begin with 'CCA'

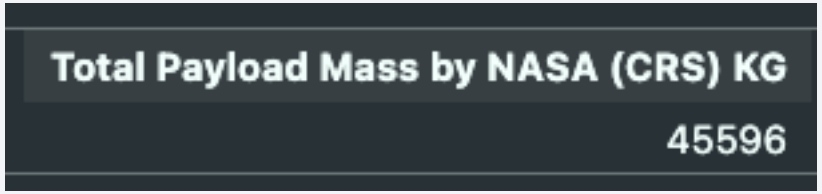
- These are five records associated with “CCA” Launch Site.
- The result is obtained given the “like” information in the corresponding table.

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

Total Payload Mass

- Here the total Payload Mass carried by boosters launched by NASA (CRS).

- The result is obtained given the “sum” of records associated with “Customer” == “NASA (CRS)”.



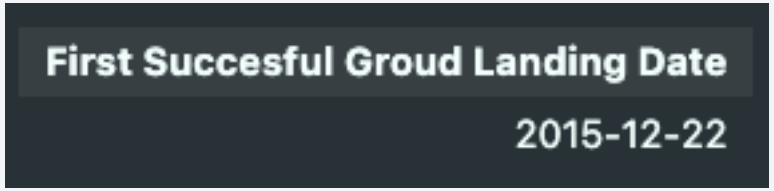
Average Payload Mass by F9 v1.1

- Here we present the average payload mass carried by booster version F9 v1.1
- The result is obtained using “AVG” function over the relevant records in relation with “Booster_Version”==“F9 v1.1”.

AVG Payload Mass for F9v1.1
2928.4

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad is given.
- The result is observed a “Min” function from the “Success (ground pad)” landing outcome.



Successful Drone Ship Landing with Payload between 4000 and 6000

- List of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kg.
- The list is achieved as a double request over the “Landing_Outcome” and the “Payload_Mass”.

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The count of de different outcome results where “Success” or “Failure” are involved.
- The query result is given as a “Count” for the different “Group by” outcomes.

Outcome	COUNT
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

Boosters Carried Maximum Payload

- List of all booster versions that have carried the maximum payload mass.
- Using a subquery routine it is possible to know the “Max” payload mass and relate it to the Booster Version.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List of the records with month name, failure landing outcome in drone ship, booster version and launch site in the year 2015.
- With a “substr” function it is possible to obtain the month and year, which allow us to recover the desired outcome.

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

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

- Rank of the landing outcomes between 2010-06-04 and 2017-03-20.

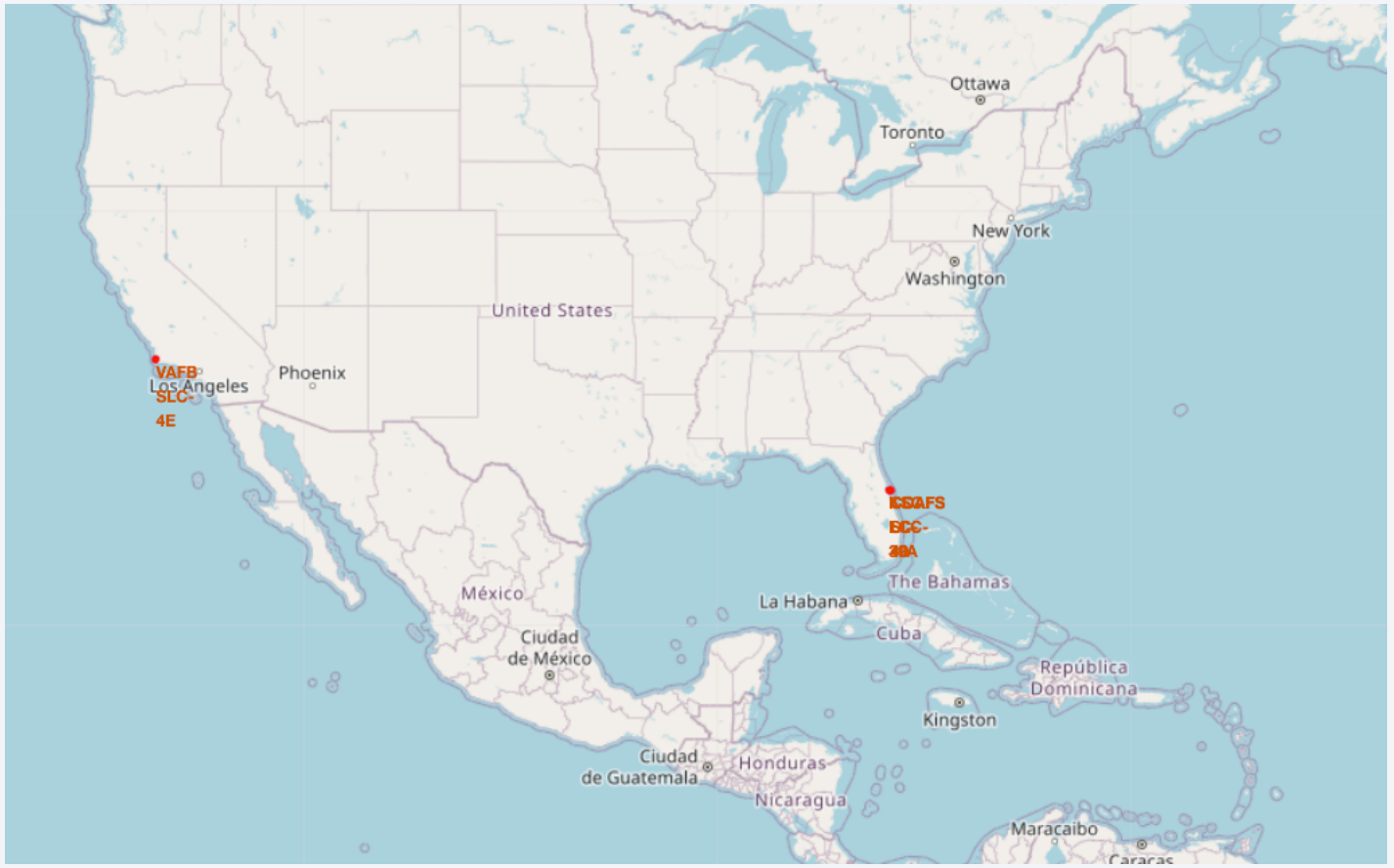
- With a subquery routine it is possible to find the time interval of interest and then “Count” the different sets obtained by a “Group by” action.

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

Section 3

Launch Sites Proximities Analysis

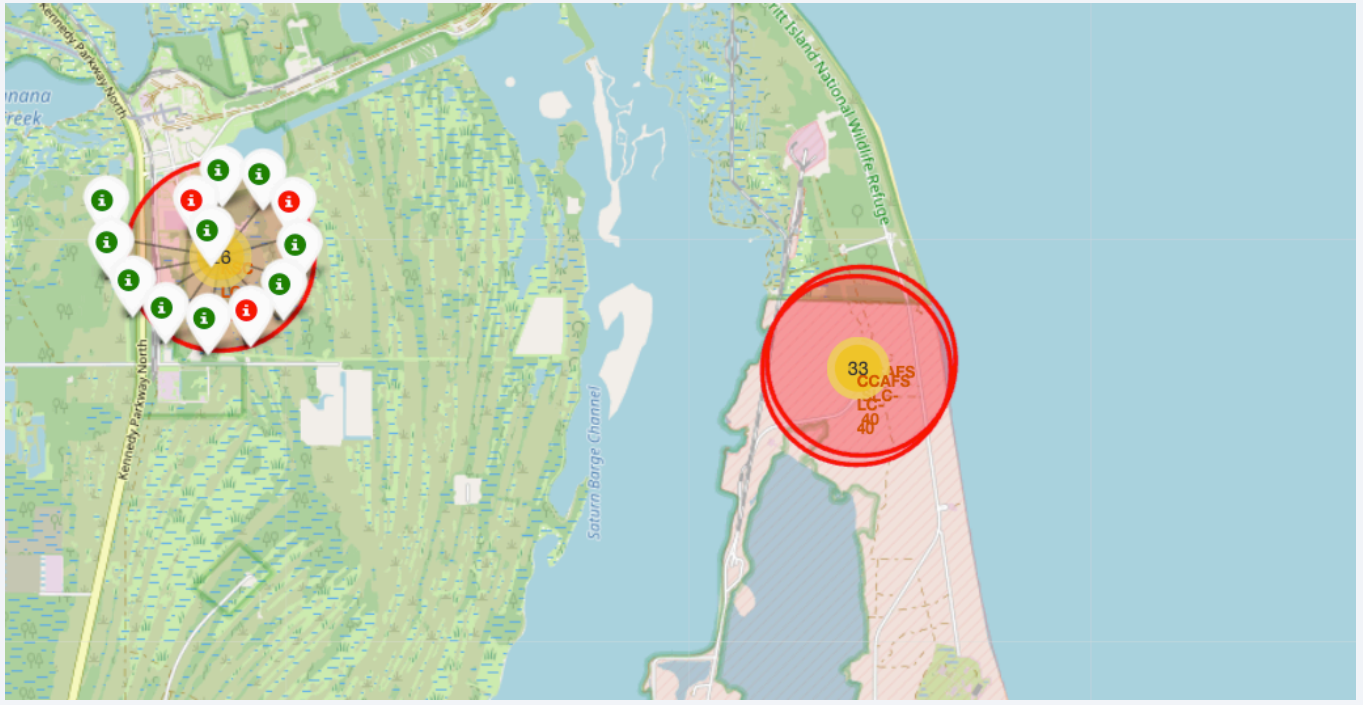
All Launch Sites



- There are launch sites near both oceans with which the US has contact.

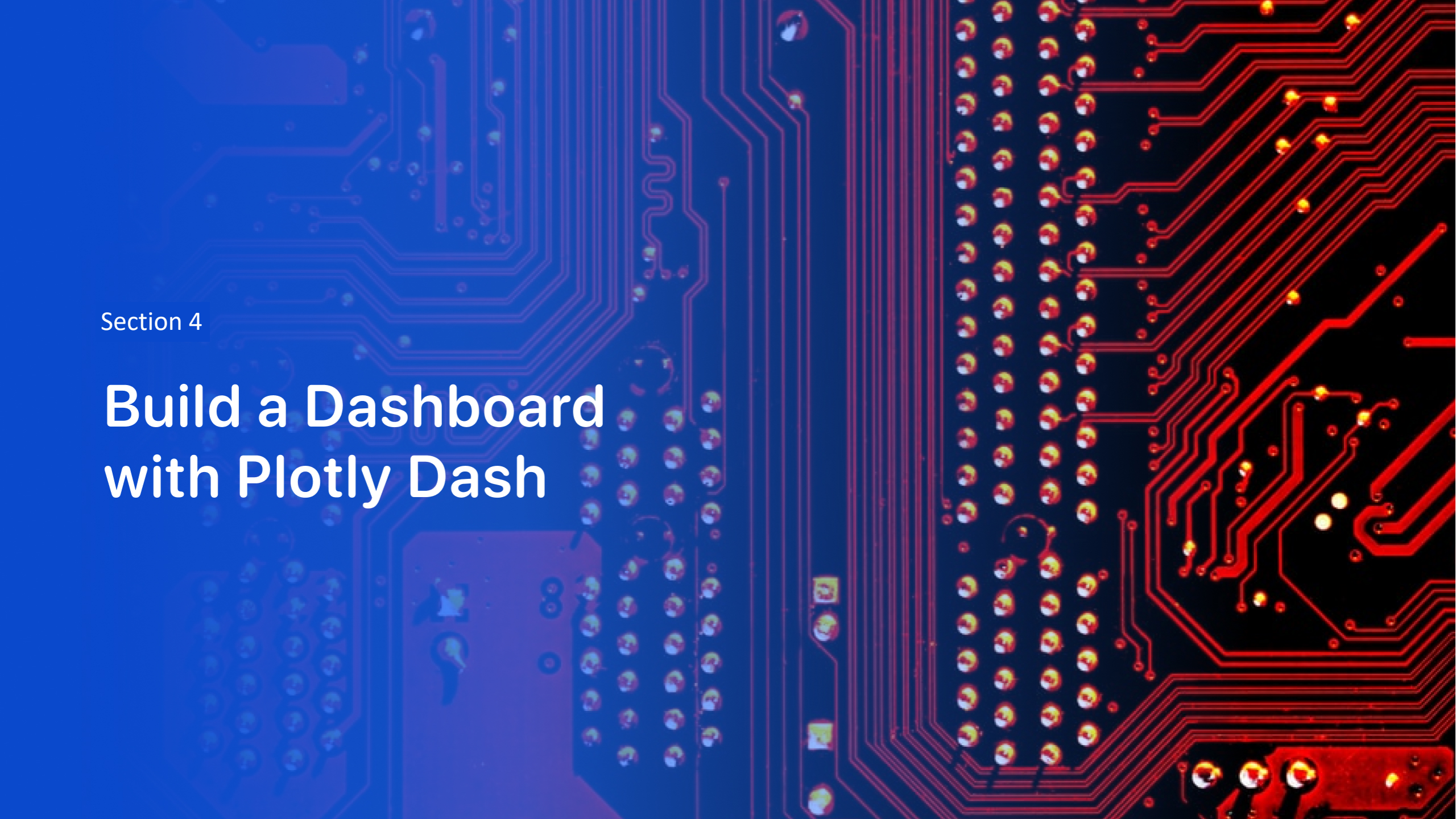
Launch Outcomes by Site

- In a Folium map it was possible to color-label the latch outcomes; in green the positive and in red the negative outcomes.
- E.g. the results for KSC LC-39A is shown.



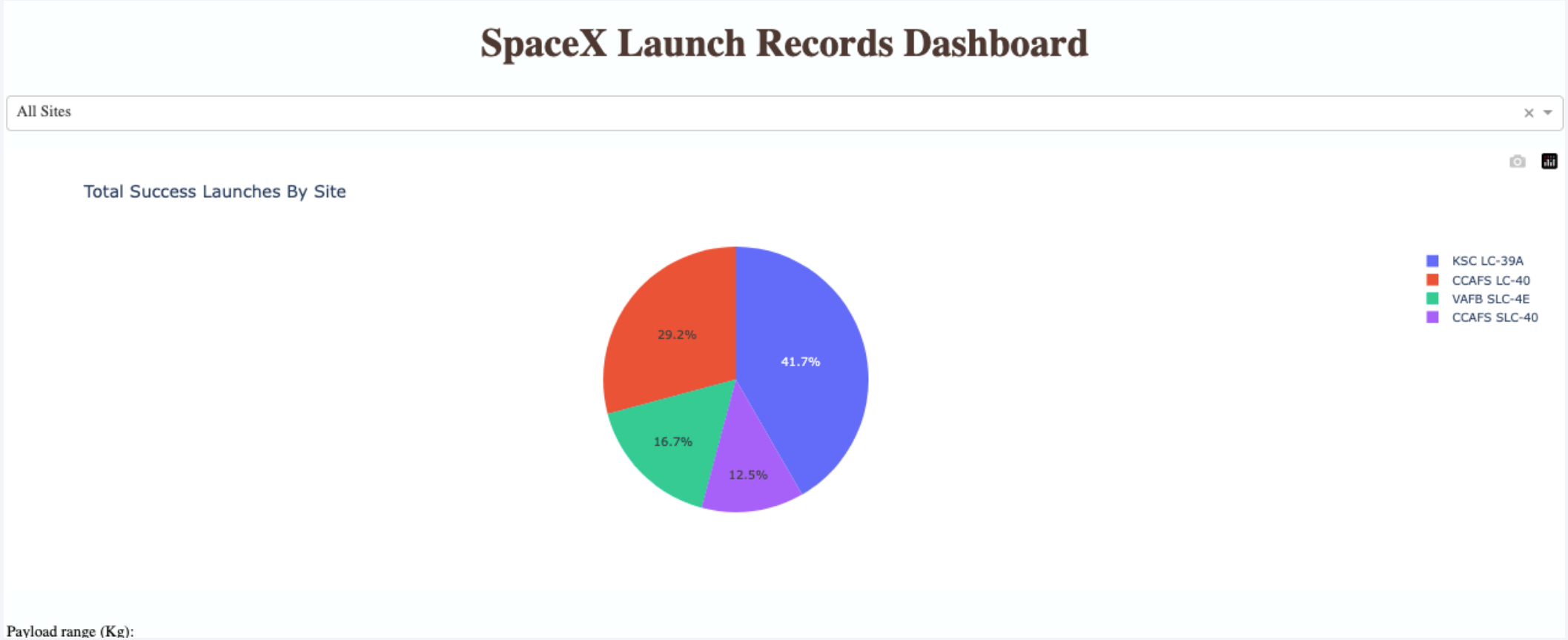
Launch Sites and Proximities

- It was also possible to explore the proximities of the launch sites such as railway, highway coastline, with distance calculated and displayed
- E.g. the nearest coastline to CCAFS SLC-40 is shown.


Section 4
Build a Dashboard with Plotly Dash

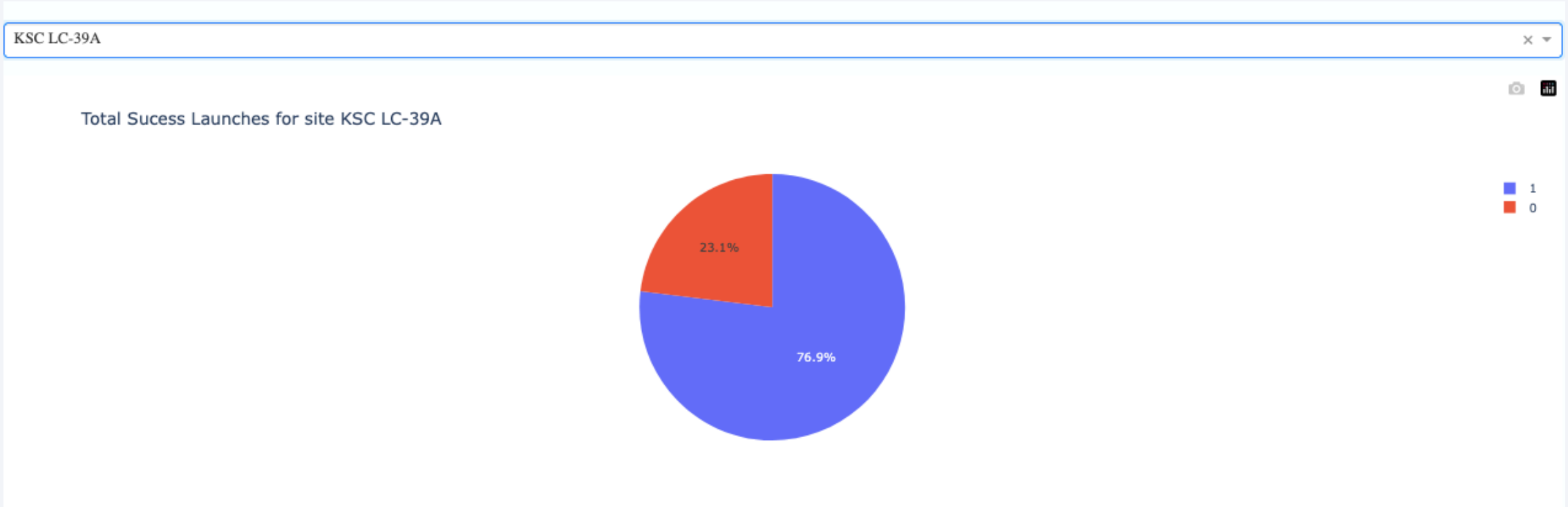
SpaceX Launch Records Dashboard

- A pie chart of Success Launches by site is shown. By total amount, KSC LC-39A is the most successful location.



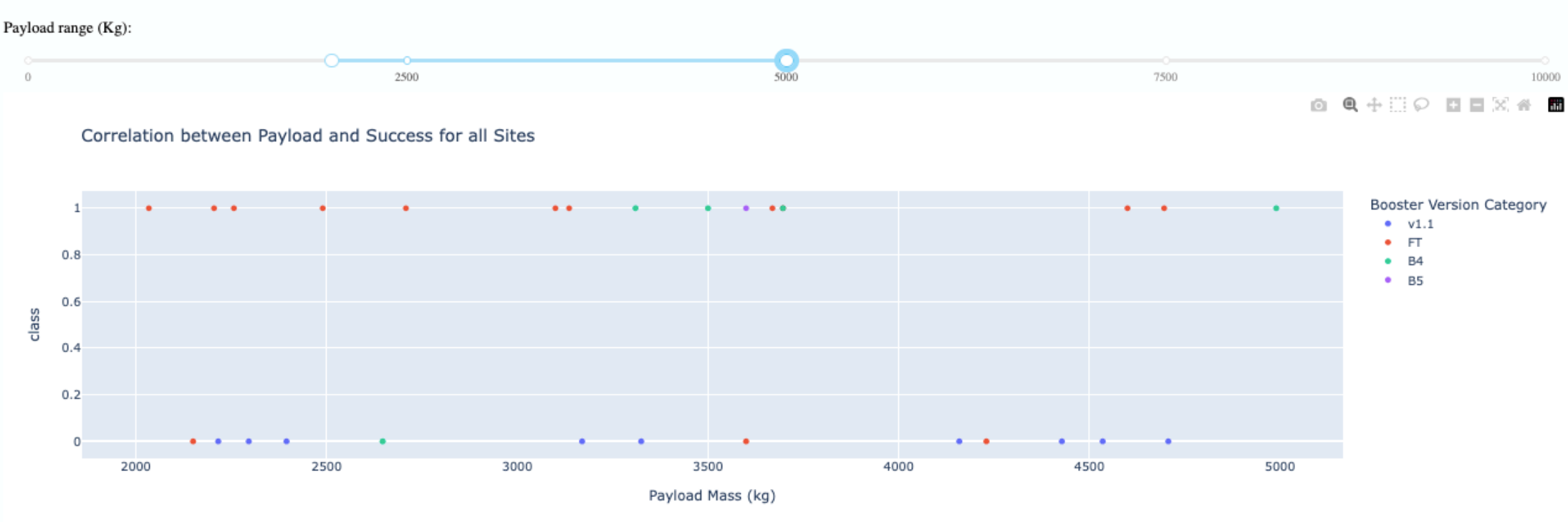
Launch site with highest launch success ratio

- The most successful site is KSC LC-39A with a success ratio of the 76.9%



Payload vs Launch Outcome

- A Payload Mass range around (2000kg,6000kg) seems to collect the highest number of successful launches.

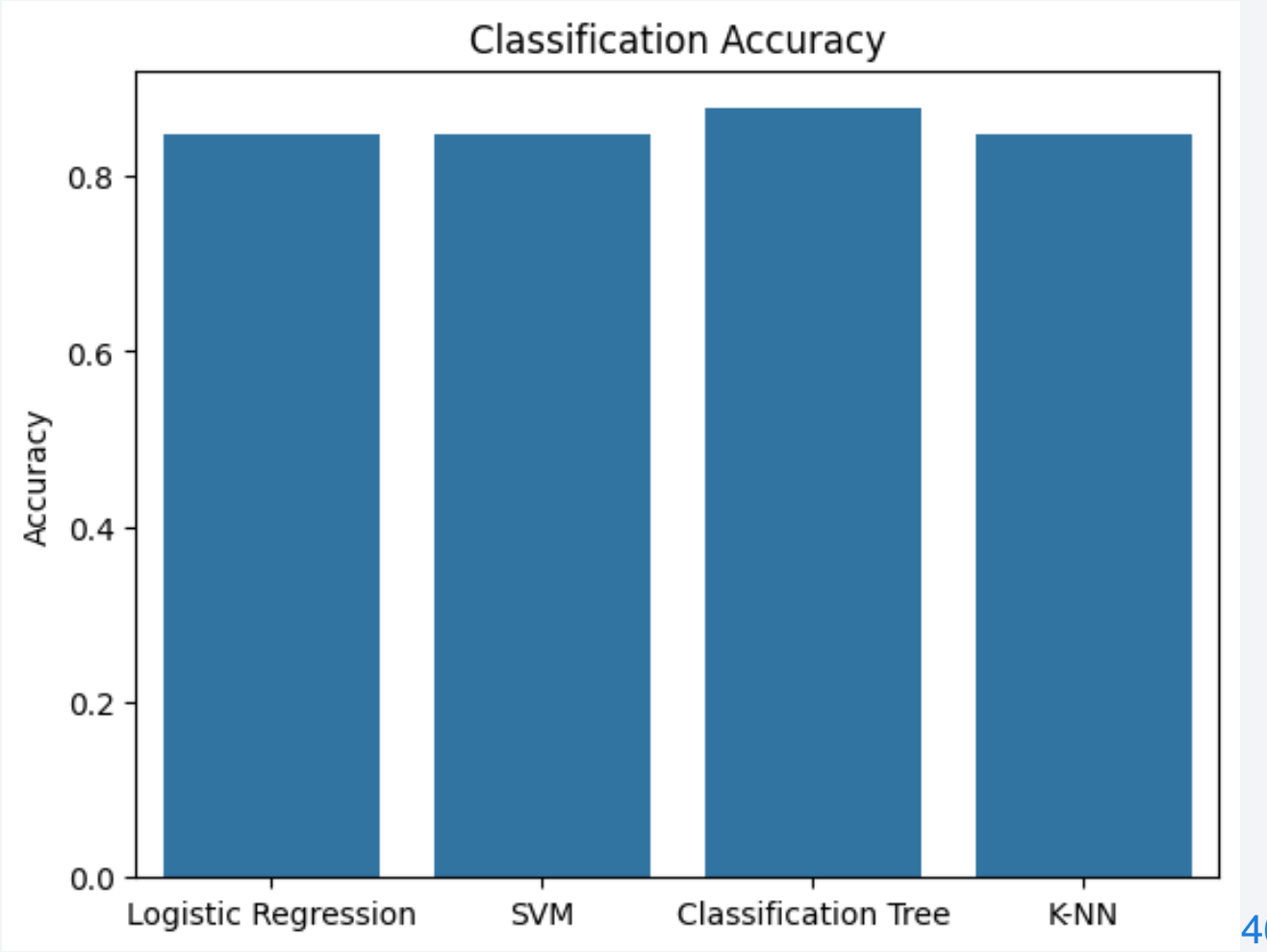


Section 5

Predictive Analysis (Classification)

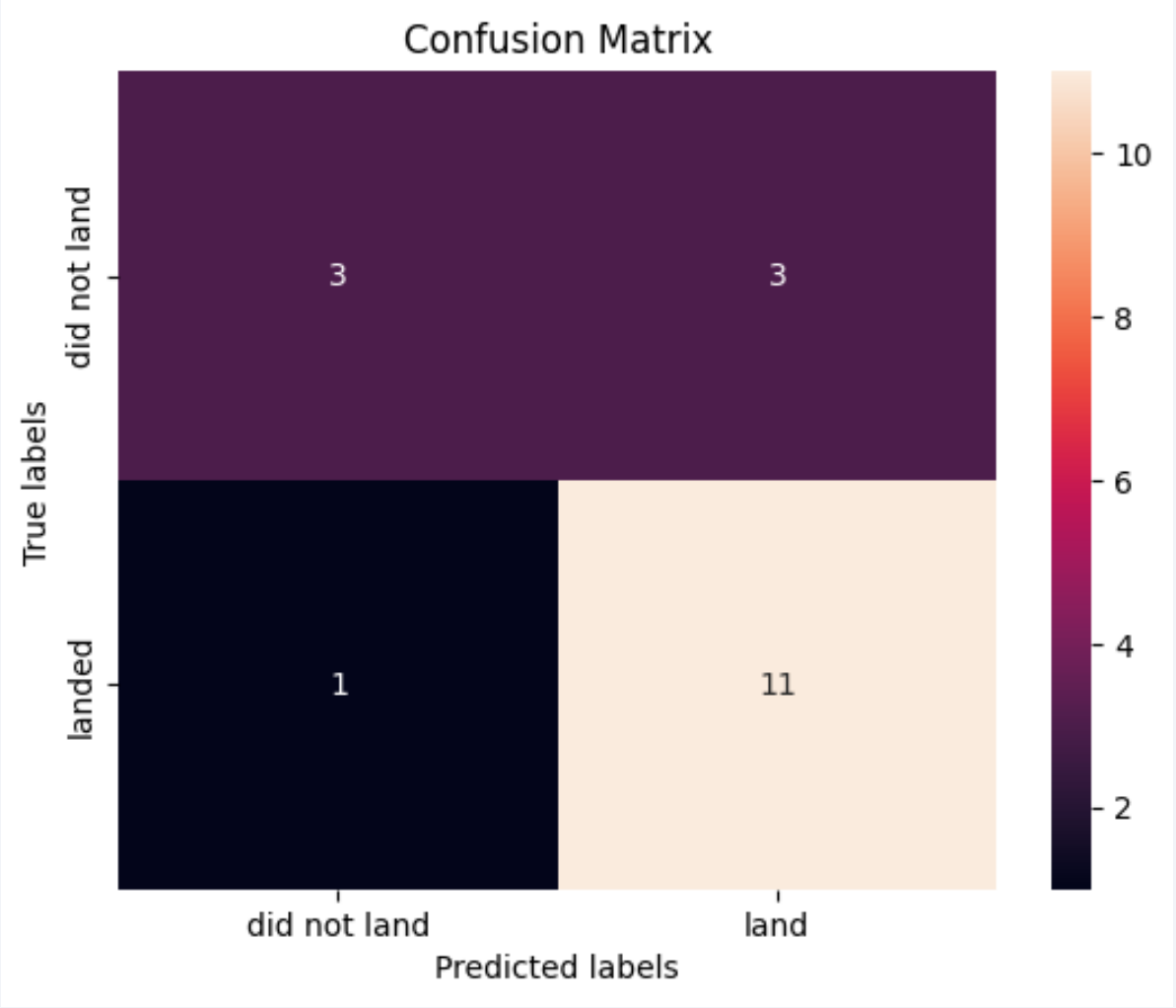
Classification Accuracy

- Classification Tree with an accuracy of 0.877 is the model that best performs.



Confusion Matrix

- The confusion matrix of the Decision Tree Classifier is shown.
- It has 14 over 18 correct classifications given the test set.



Conclusions

- Different data sources were explored in order to obtain the necessary information.
- It is possible to recognize the most successful launch site: KSC LC-39A.
- A Payload Mass around 4500kg appears to be the best bet for a test launch.
- For future attempts by SpaceY, the Tree Classification model will be refined as it has proven to be the best classifier.
- The launch process is showing more promise over time.

Appendix

- All resources used in this exhibition can be consulted in the corresponding repository.

Thank you!