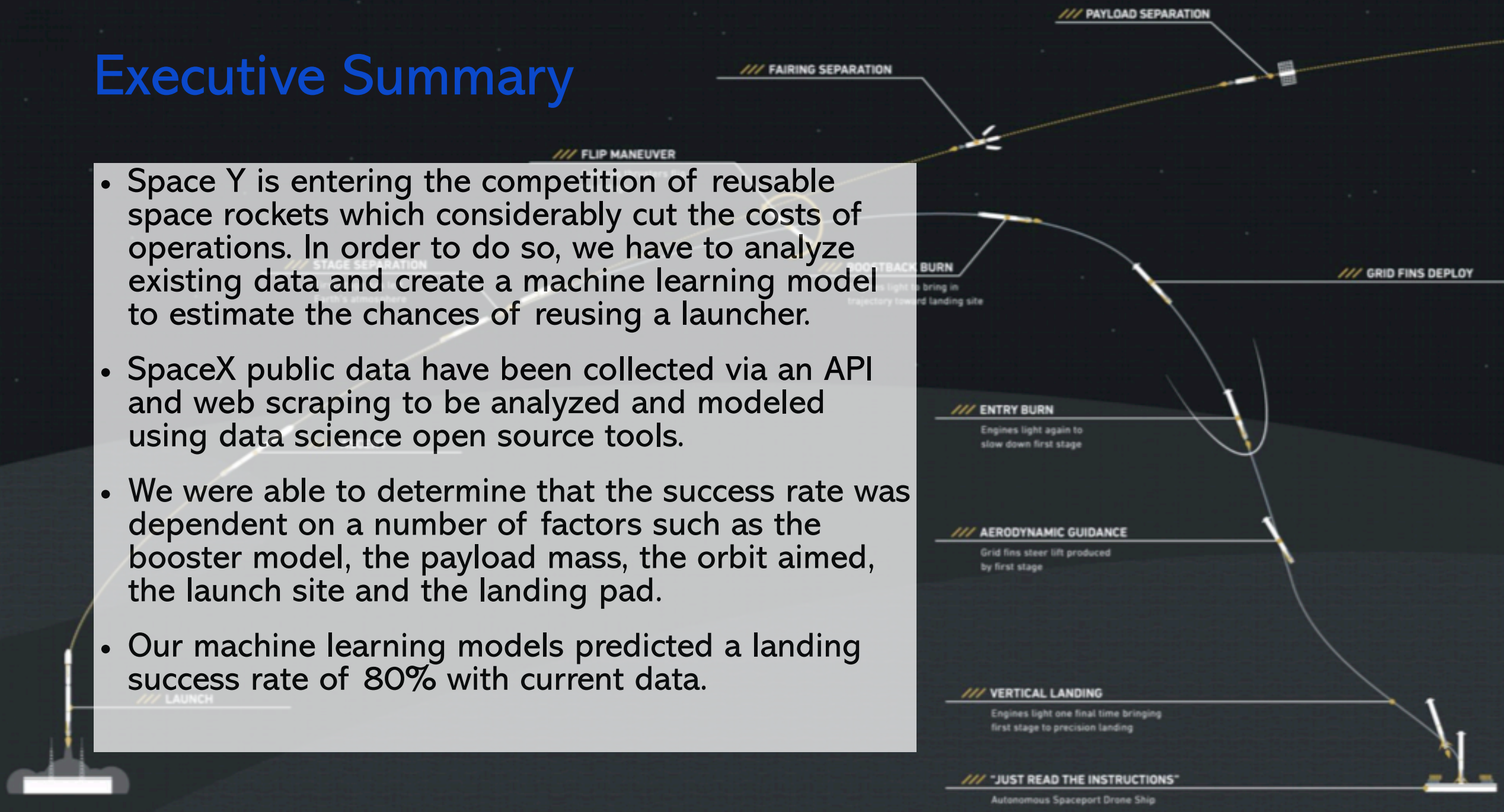
Winning Space Race with Data Science

Thierry Coton
16 Feb 2022

Table of content

Executive Summary

- Space Y is entering the competition of reusable space rockets which considerably cut the costs of operations. In order to do so, we have to analyze existing data and create a machine learning model to estimate the chances of reusing a launcher.
- SpaceX public data have been collected via an API and web scraping to be analyzed and modeled using data science open source tools.
- We were able to determine that the success rate was dependent on a number of factors such as the booster model, the payload mass, the orbit aimed, the launch site and the landing pad.
- Our machine learning models predicted a landing success rate of 80% with current data.



Introduction

This report has been prepared as part of the IBM Professional Certificate in Data Science provided by Coursera.

This document will present the results about a study of the success rate of SpaceX rocket launches using publicly available documentation and open source data science tools.

FALCON 9

OVERVIEW

HEIGHT	70 m / 229.6 ft
DIAMETER	3.7 m / 12 ft
MASS	549,054 kg / 1,207,920 lb
PAYLOAD TO LEO	22,800 kg / 50,265 lb
PAYLOAD TO GTO	8,300 kg / 18,300 lb
PAYLOAD TO MARS	4,020 kg / 8,860 lb

SpaceX advertises its Falcon 9 rockets starting at $62 M up to 5.5mT to GTO orbit.

They can only advertise this price because they can reuse most of the costly launcher components contained in the first stage.

Can we predict the rate of successful landing of this first stage?

Official SpaceX numbers as of February 2022

139	99	79
TOTAL LAUNCHES	TOTAL LANDINGS	REFLOWN ROCKETS

Section 1

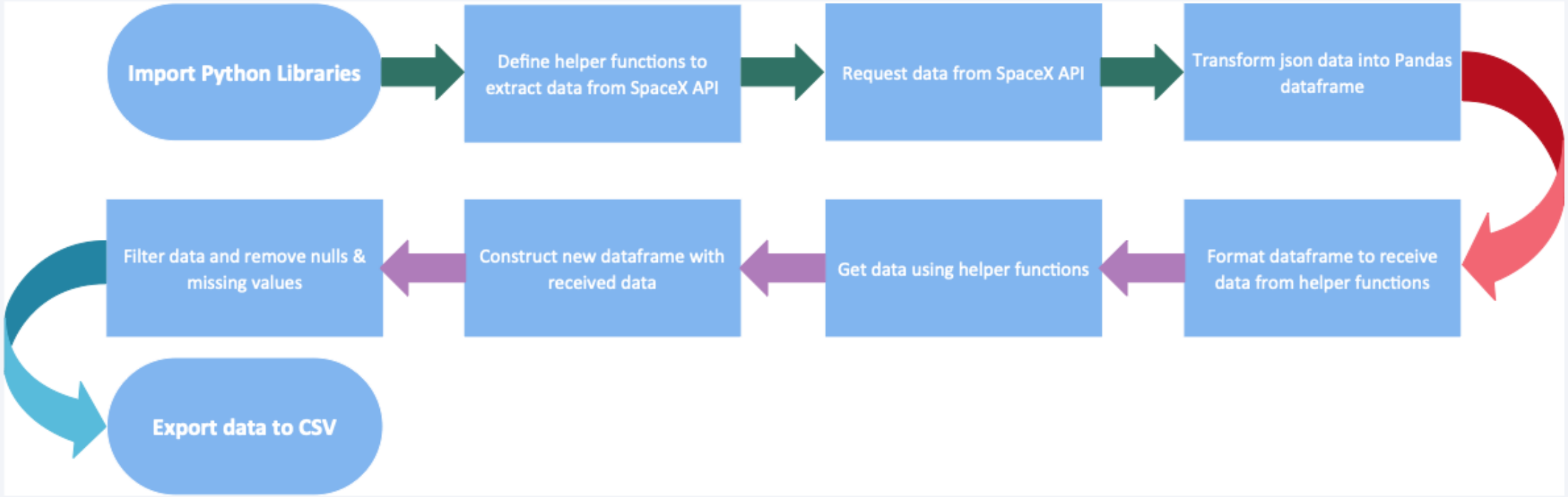
Methodology

Methodology

Executive Summary

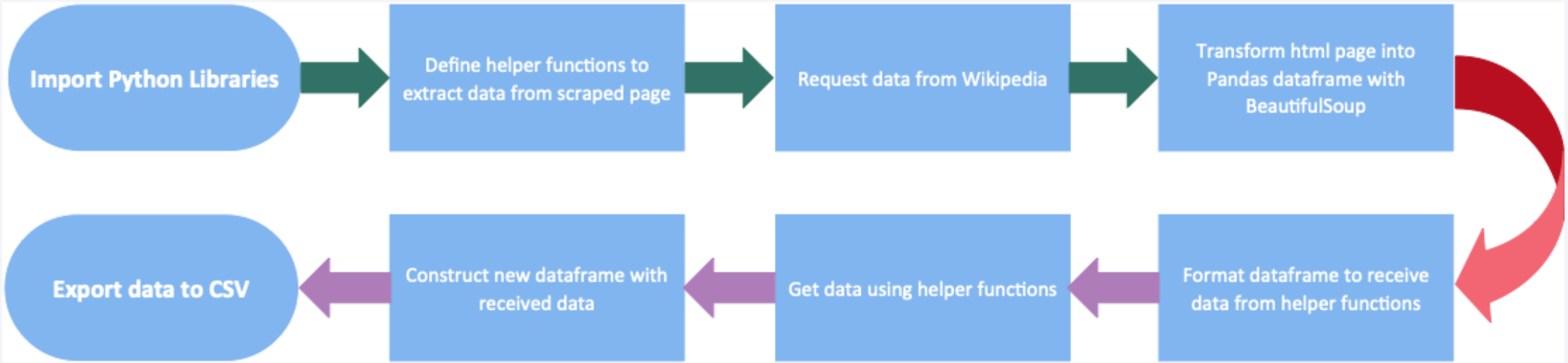
- Data collection methodology:
  - Data were collected on the Internet using SpaceX API and web scraping SpaceX's Wikipedia webpage. Some datasets were provided « as is » for the SQL and Folium labs.
- Perform data wrangling
  - Data were processed in Jupiter Notebook using Python Pandas and Numpy.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We split the data in train and test sets to use with the different classification models. Each classification model has been evaluated by searching the best algorithm parameters along its accuracy score and a printed confusion matrix. All scores were finally compared to determine which model was the most accurate.

Data Collection – SpaceX API



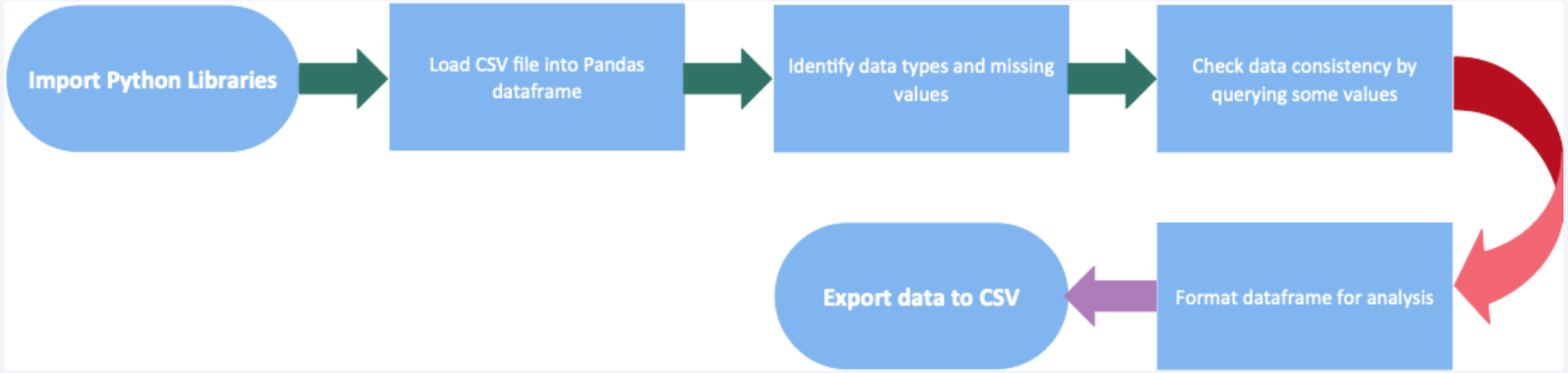
https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/1-Data%20Collection%20API%20Lab%20-%20Notebook.ipynb

Data Collection - Scraping



https://github.com/Tcoton/Applied-Data-Science-labs/x/master/2-Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

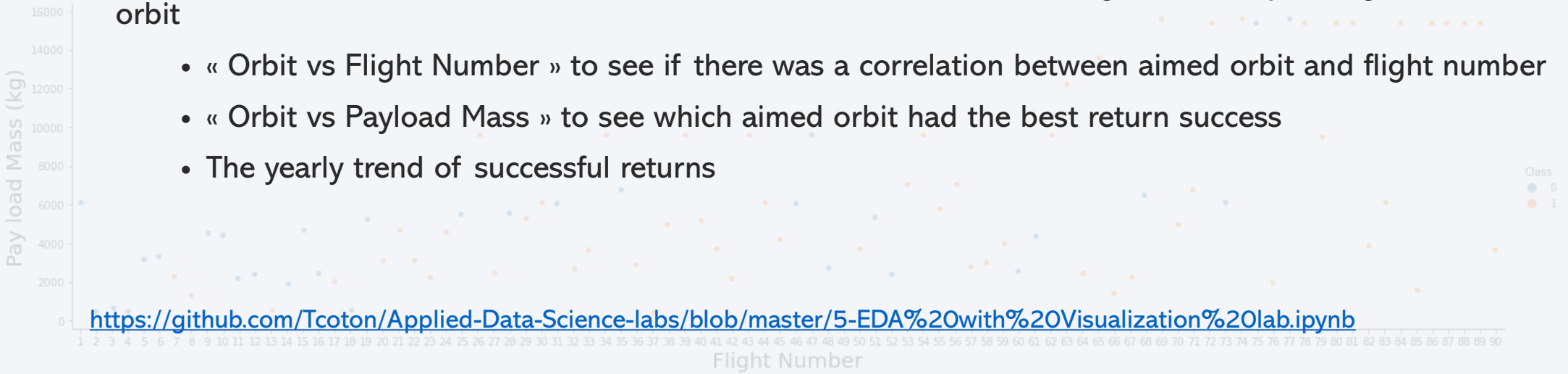
Data Wrangling



https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/3-Data%20Wrangling%20-%20EDA%20lab.ipynb

EDA with Data Visualization

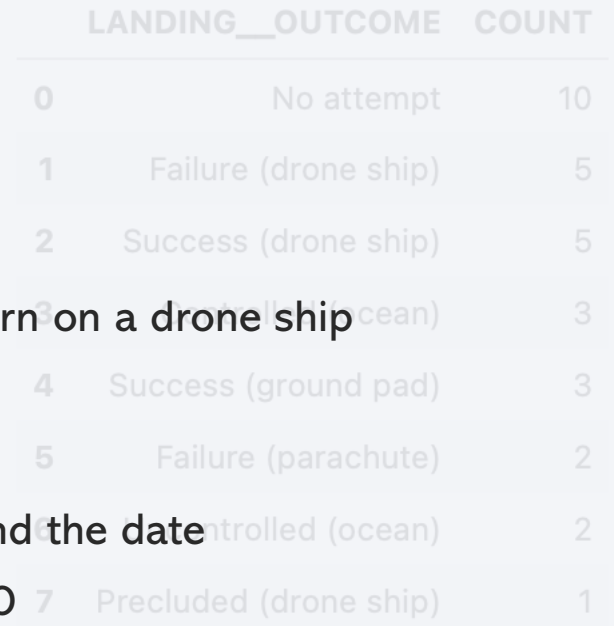
- We plotted the following charts:
  - « Payload vs Flight Number » to see the ratio of payload mass vs first stage return
  - « Launch site vs Flight Number » to see which launch site had the best outcomes
  - « Payload Mass vs Launch Site » to see which site launched the heaviest payload
  - « Orbit vs Success rate » to determine the success rate of first stage return depending on aimed orbit
  - « Orbit vs Flight Number » to see if there was a correlation between aimed orbit and flight number
  - « Orbit vs Payload Mass » to see which aimed orbit had the best return success
  - The yearly trend of successful returns



https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/5-EDA%20with%20Visualization%20lab.ipynb

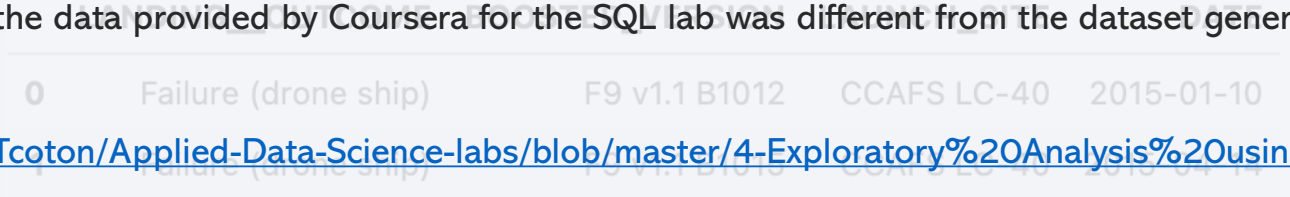
EDA with SQL

- Using DB2 SQL we retrieved the following information from a provided dataset:
  - The names of each launch site
  - The total and average payload mass carried by booster or customer
  - The date of the first return success on a ground pad
  - The booster name which carried a certain payload and succeeded to return on a drone ship
  - The count of mission outcomes
  - A list of booster having carried the maximum payload mass
  - A list of booster with failed outcomes along with the launch site name and the date
  - A rank count of return outcomes between 2010/06/04 and 2017/03/20



	LANDING__OUTCOME	COUNT
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	[illegible]ocean)	3
4	Success (ground pad)	3
5	Failure (parachute)	2
6	[illegible]ntrolled (ocean)	2
7	Precluded (drone ship)	1

- Please note that the data provided by Coursera for the SQL lab was different from the dataset generated by our data collection.



	LANDING_OUTCOME	BOOSTER_VERSION	LAUNCH_SITE	DATE
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
1	Failure (drone ship)	F9 v1.1 B1013	CCAFS LC-40	2015-04-14

https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/4-Exploratory%20Analysis%20using%20SQL.ipynb

Build an Interactive Map with Folium

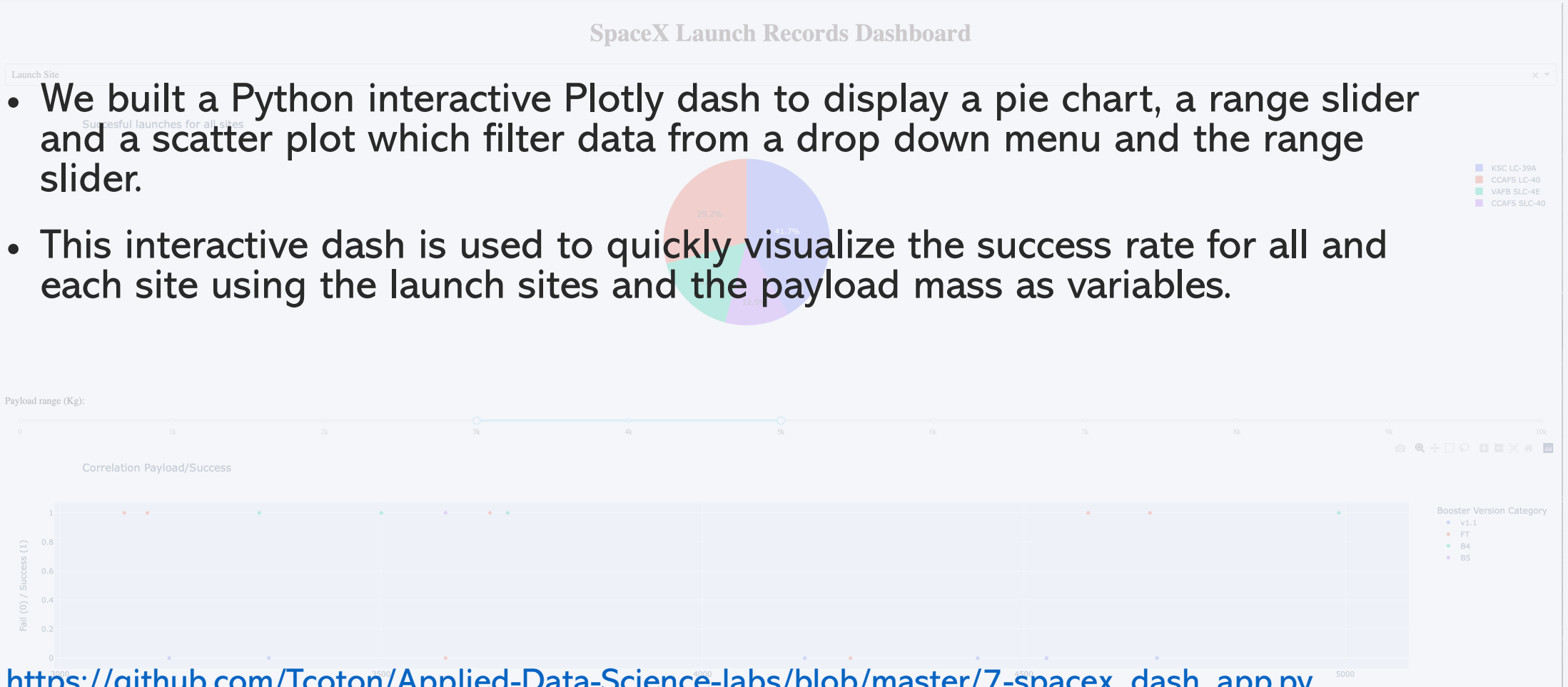
- We marked on maps the location of each launch site using circle markers
- On top of those location, we added some marker clusters to visualize the success and failure for each site
- We then added the distances from different point on the map using mouse position, distance markers and distance lines

- Please note that the data provided by Coursera for this lab was different from the data collection.

https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/6-Interactive%20Visual%20Analytics%20with%20Folium.ipynb

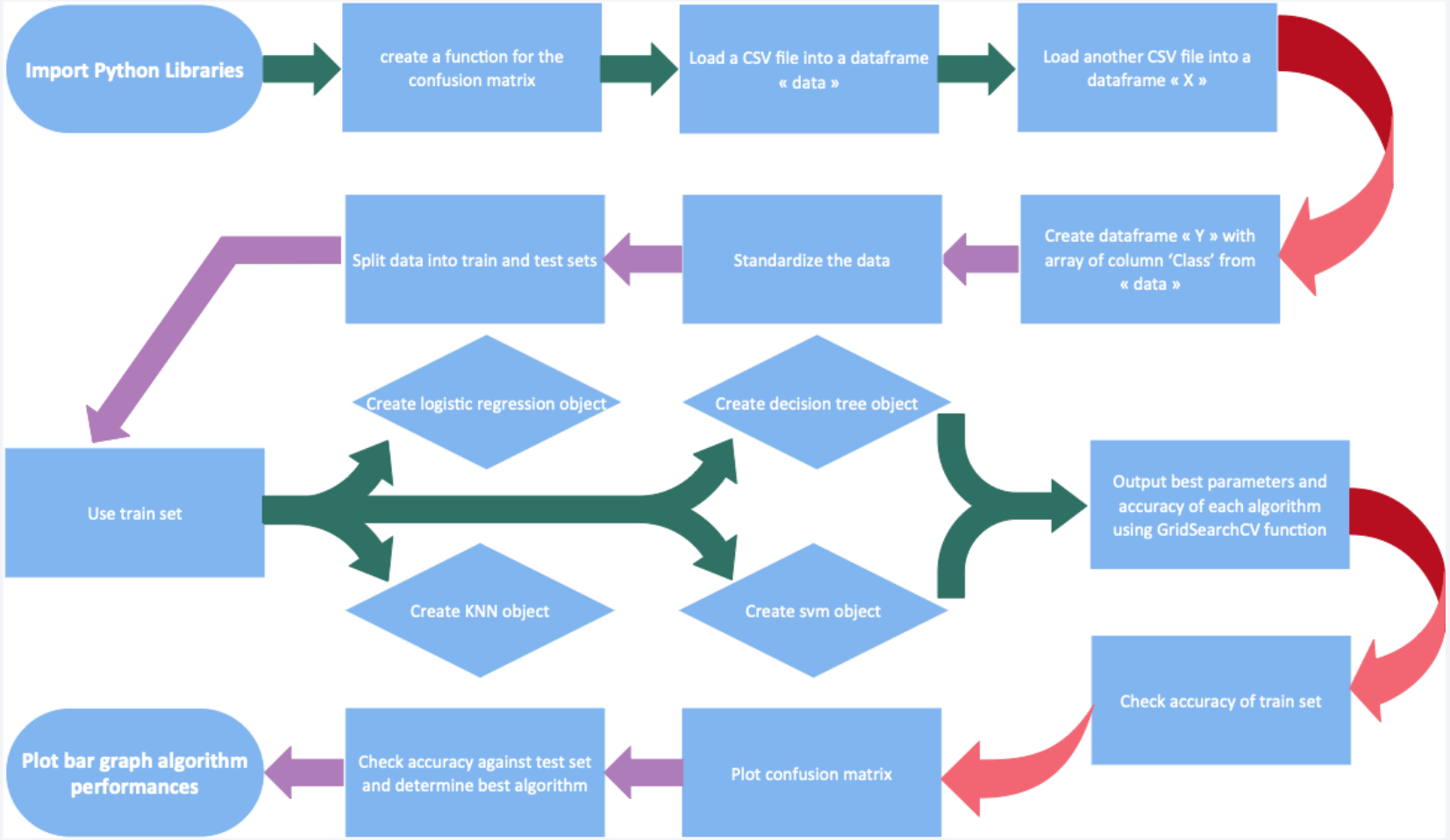
Build a Dashboard with Plotly Dash

- We built a Python interactive Plotly dash to display a pie chart, a range slider and a scatter plot which filter data from a drop down menu and the range slider.
- This interactive dash is used to quickly visualize the success rate for all and each site using the launch sites and the payload mass as variables.



https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/7-spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/Tcoton/Applied-Data-Science-labs/blob/master/8-Machine%20Learning%20Prediction%20.ipynb

Results

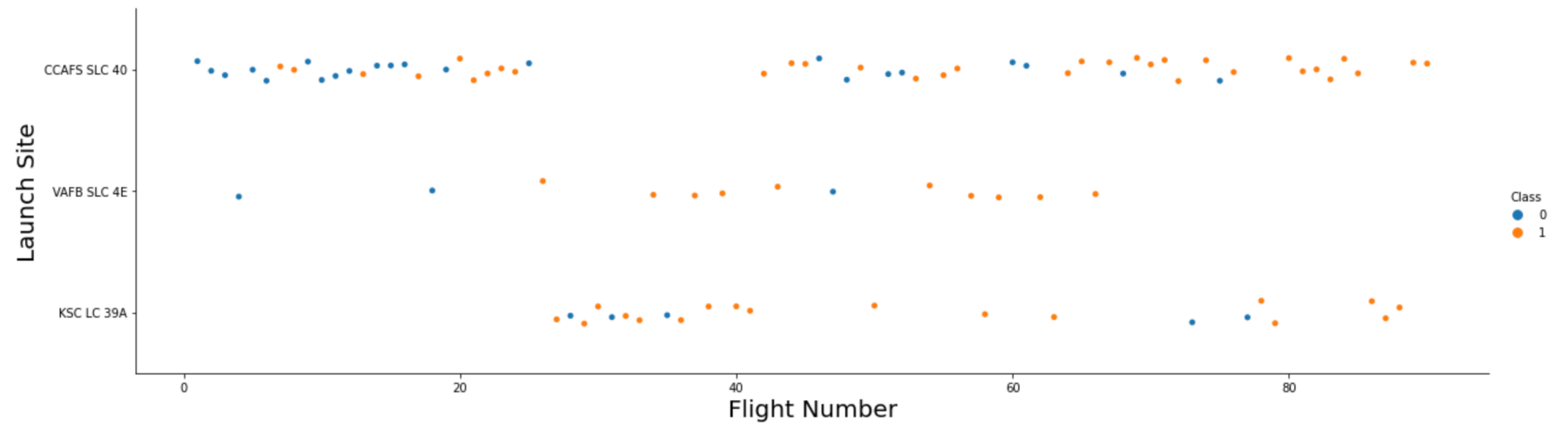
- The exploratory analysis using SQL showed us 4 launch sites with a total of 100 successful outcomes vs 1 failure, one of the successes has an unclear payload status. The first successful return on ground occurred on December 22nd, 2015 and we could see that over a period of seven years, there was 10 occurrences over 31 where there was no attempts to return the first stage on planet Earth.
- The dataset provided for the SQL analysis was different from the other dataset used for the other analysis.

- The predictive analysis showed us that all models had the same accuracy of around 80%, the most accurate being the decision tree due to the low data cardinality but the most performing was the K nearest neighbors algorithm.
- Let's explore the results in details

Section 2

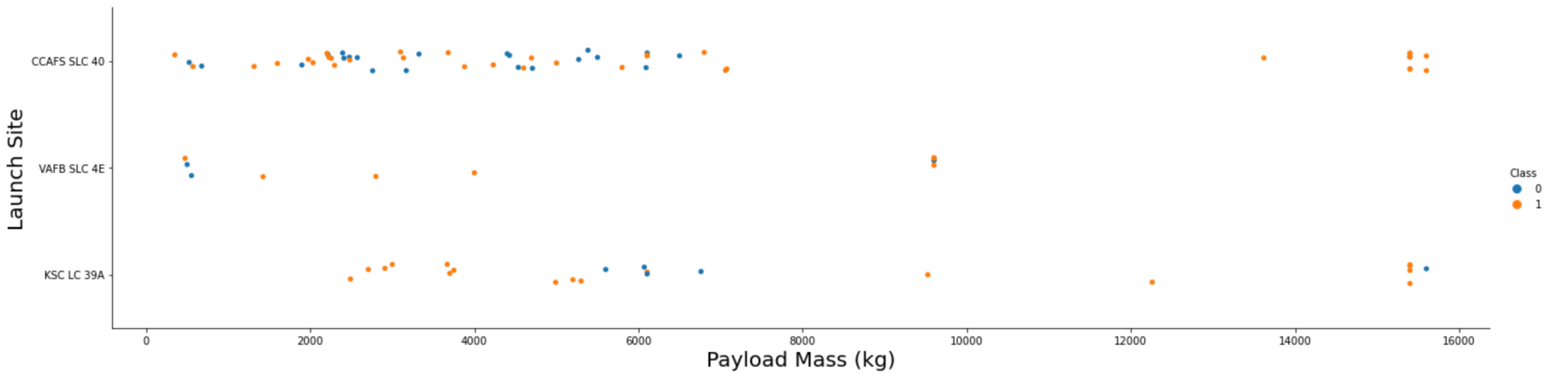
Insights drawn from EDA

Flight Number vs. Launch Site



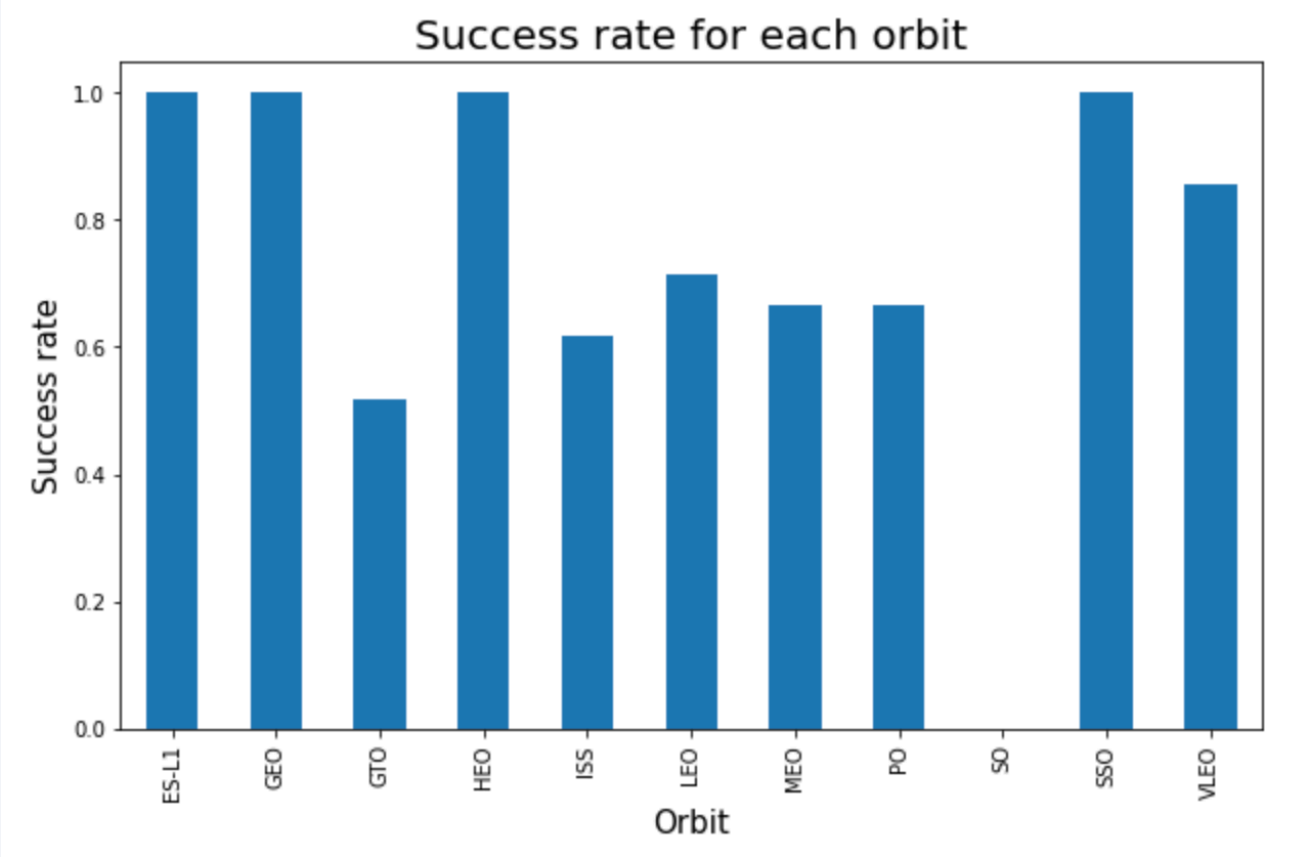
In this scatter plot, we can see the ratio of success (1) / failure (0) for each launch site over time represented by the Flight number. We also see that CCAFS SLC 40 has launched more rockets than the other 2.

Payload vs. Launch Site



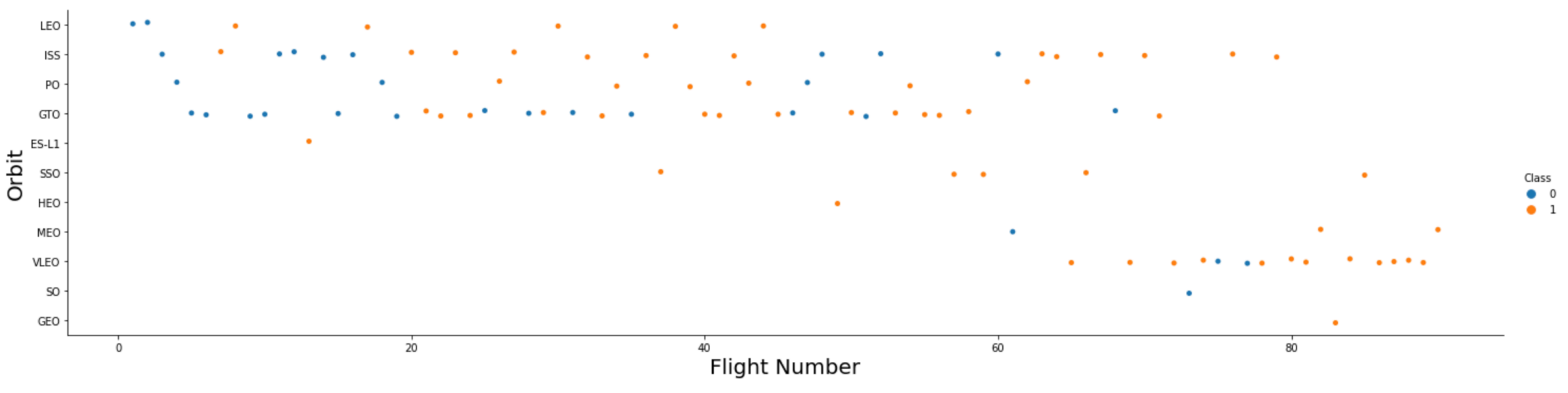
In this scatter plot, we can see only 2 sites have successfully launched payloads over 14000 KG with CCAFS SLC 40 being the most successful.

Success Rate vs. Orbit Type



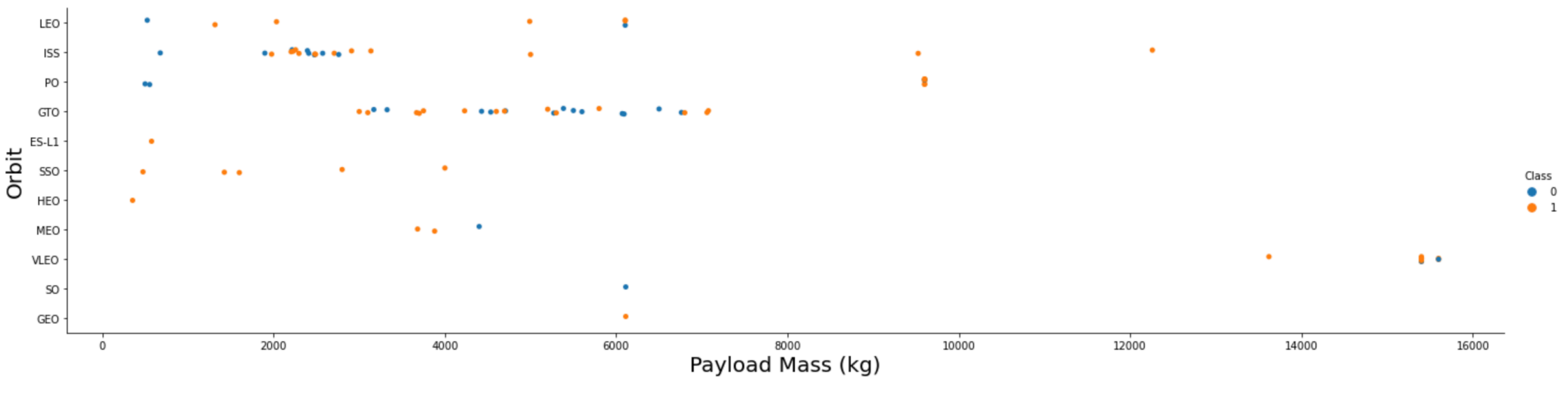
This bar charts displays the success rate for each orbit aimed, the most successful being ES-L1, GEO, HEO and SSO

Flight Number vs. Orbit Type



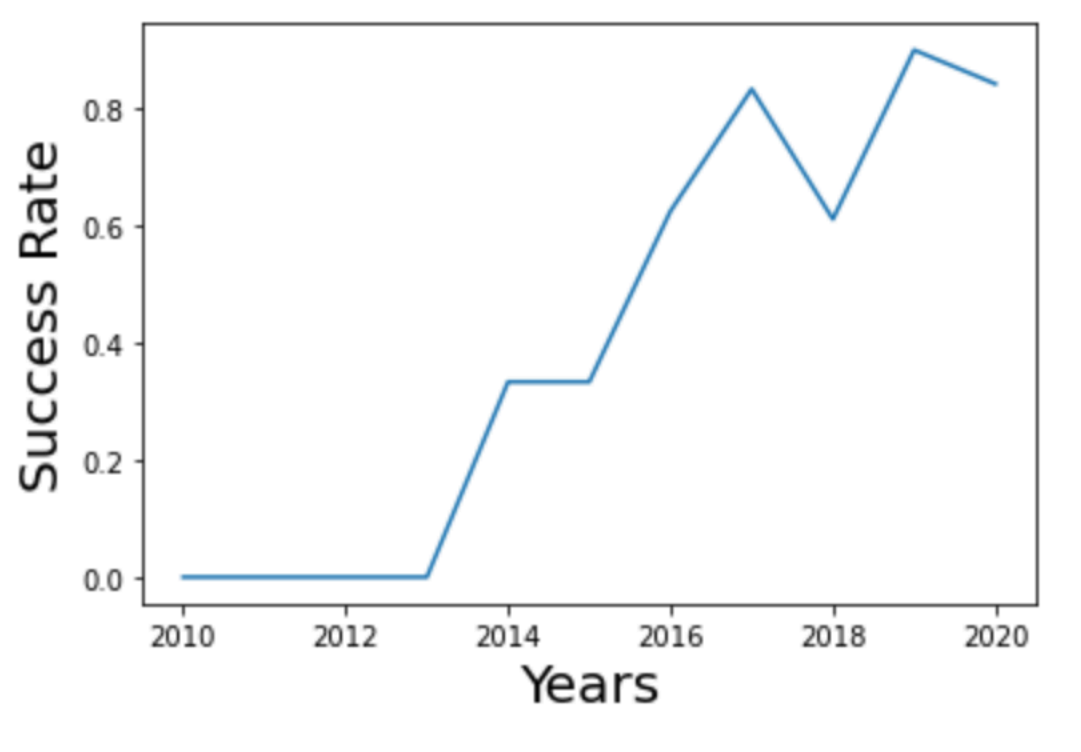
We could not find any correlation between the Flight numbers and the orbits to determine a success rate even though LEO seems to have a higher success rate on the long term.

Payload vs. Orbit Type



In this plot, we can see that SSO has a high success rate with loads under 5000Kg while ISS and PO orbits have a higher success with loads above 8000Kg than the other orbits which have mix results below 8000Kg of payload.

Launch Success Yearly Trend



In this graph, we see the landing success rate improving sharply after 2013 with a plateau between 2014 and 2015 followed by a steep increase until 2017 and a succession of decline and rises from 2017 to 2020 while remaining steadily above 60%.

All Launch Site Names

	Launch Site
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

A list of all launch sites using a SQL query

Launch Site Names Begin with 'CCA'

	Launch Site
0	CCAFS LC-40
1	CCAFS LC-40
2	CCAFS LC-40
3	CCAFS LC-40
4	CCAFS LC-40

A limited list of all occurrences of the launch sites name beginning with « CCA » in the database.

Total Payload Mass

	Total Payload Mass Launched by NASA (CRS)
0	45596

The total payload mass in Kg launched by NASA (CRS) customer

Average Payload Mass by F9 v1.1

	Average payload mass carried by F9 v1.1
0	2928

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

First Successful Ground Landing Date

	Date 1st Success on ground pad
0	2015-12-22

The dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

	Booster Version	Payload mass KG	Landing Outcome
0	F9 FT B1022	4696	Success (drone ship)
1	F9 FT B1026	4600	Success (drone ship)
2	F9 FT B1021.2	5300	Success (drone ship)
3	F9 FT B1031.2	5200	Success (drone ship)

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

	Mission Outcome	COUNT
0	Failure (in flight)	1
1	Success	99
2	Success (payload status unclear)	1

The total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

	Booster Version	Payload mass KG
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

The names of the boosters having carried the maximum payload mass

2015 Launch Records

	Landing Outcome	Booster Version	Launch Site	DATE
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

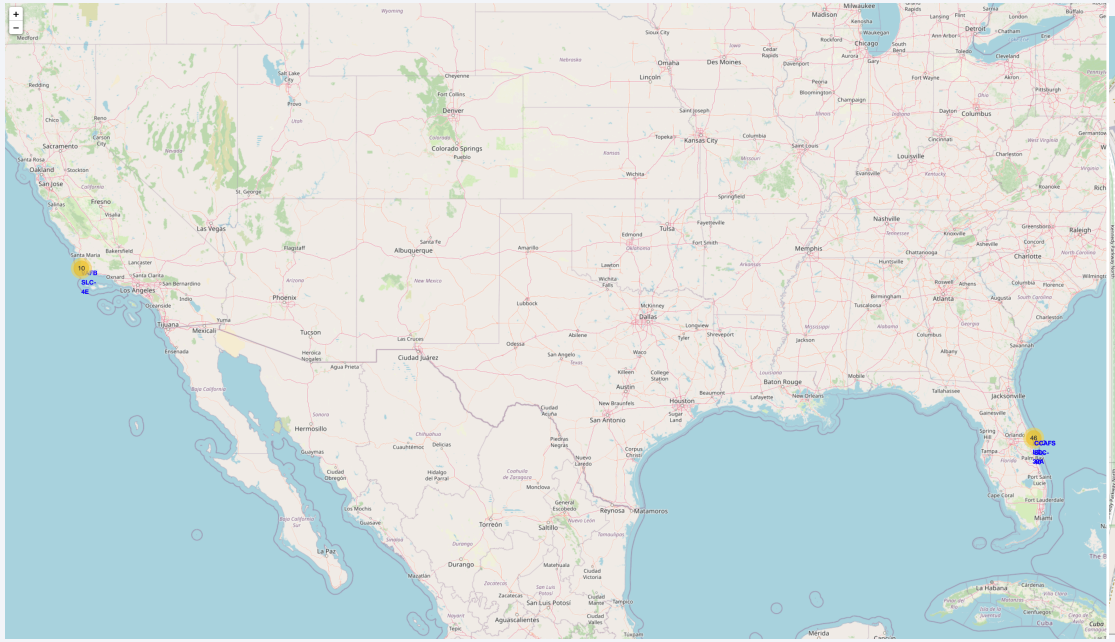
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

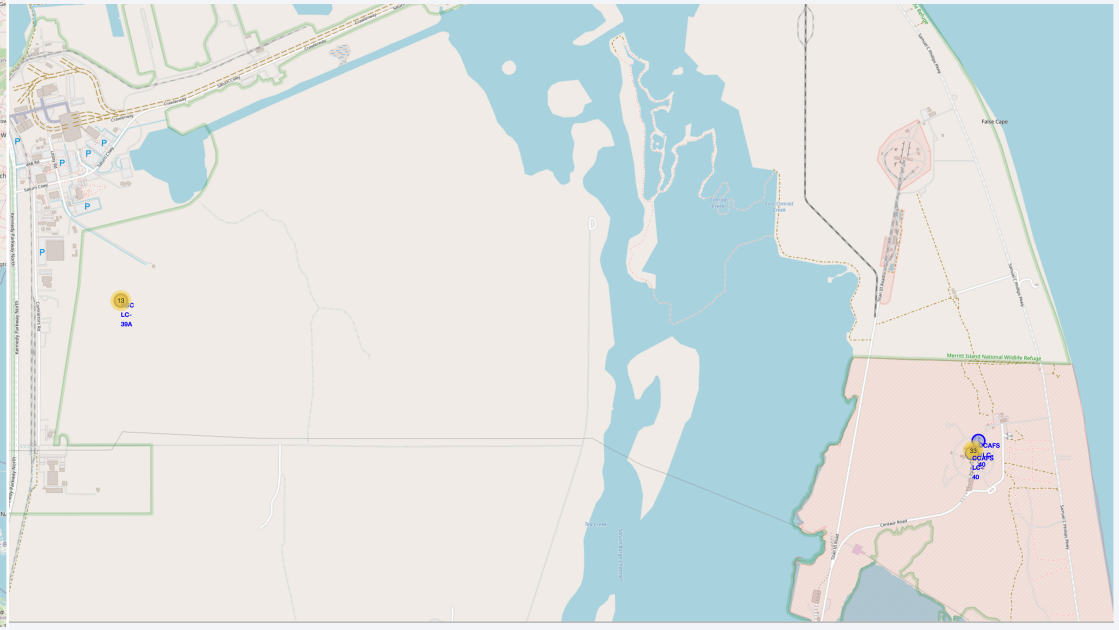
	Landing Outcome	COUNT
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Failure (parachute)	2
6	Uncontrolled (ocean)	2
7	Precluded (drone ship)	1

The count of landing outcomes between the date 2010-06-04 and 2017-03-20


Section 3
Launch Sites
Proximities Analysis

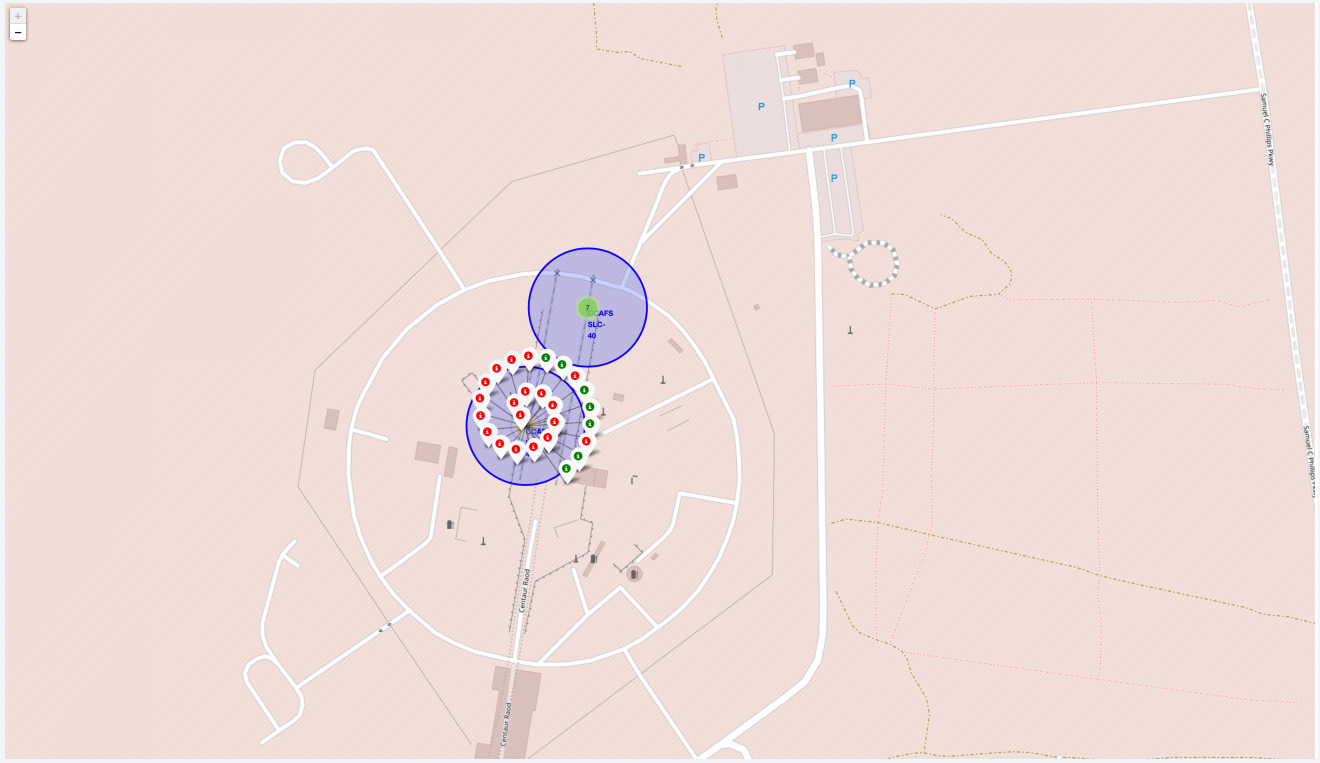
Location of all launch site on a map





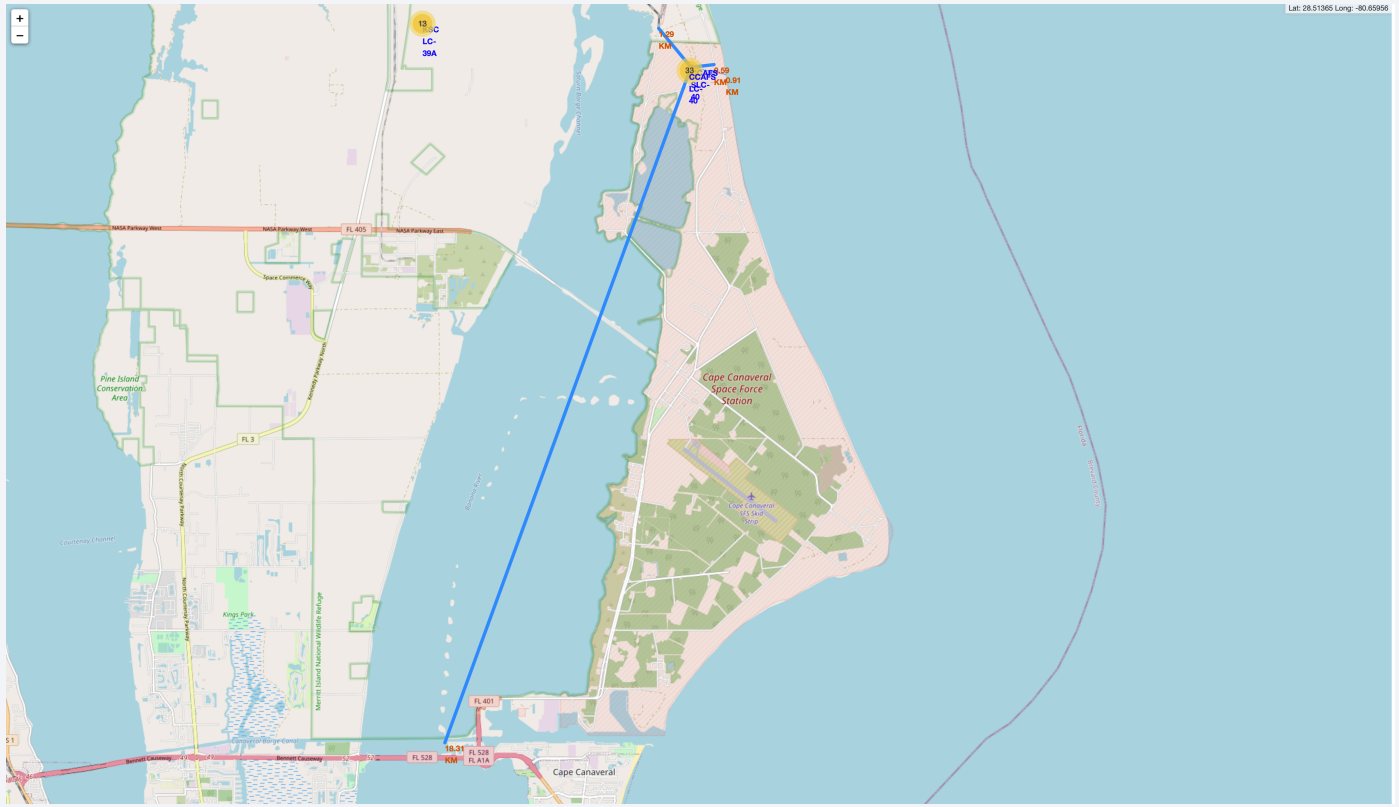
- On this map we see the location of the site in California with 10 launches and the location of the sites in Florida totaling 46 launches. I we zoom on Florida, 3 sites can be distinguished.

Outcomes mapping

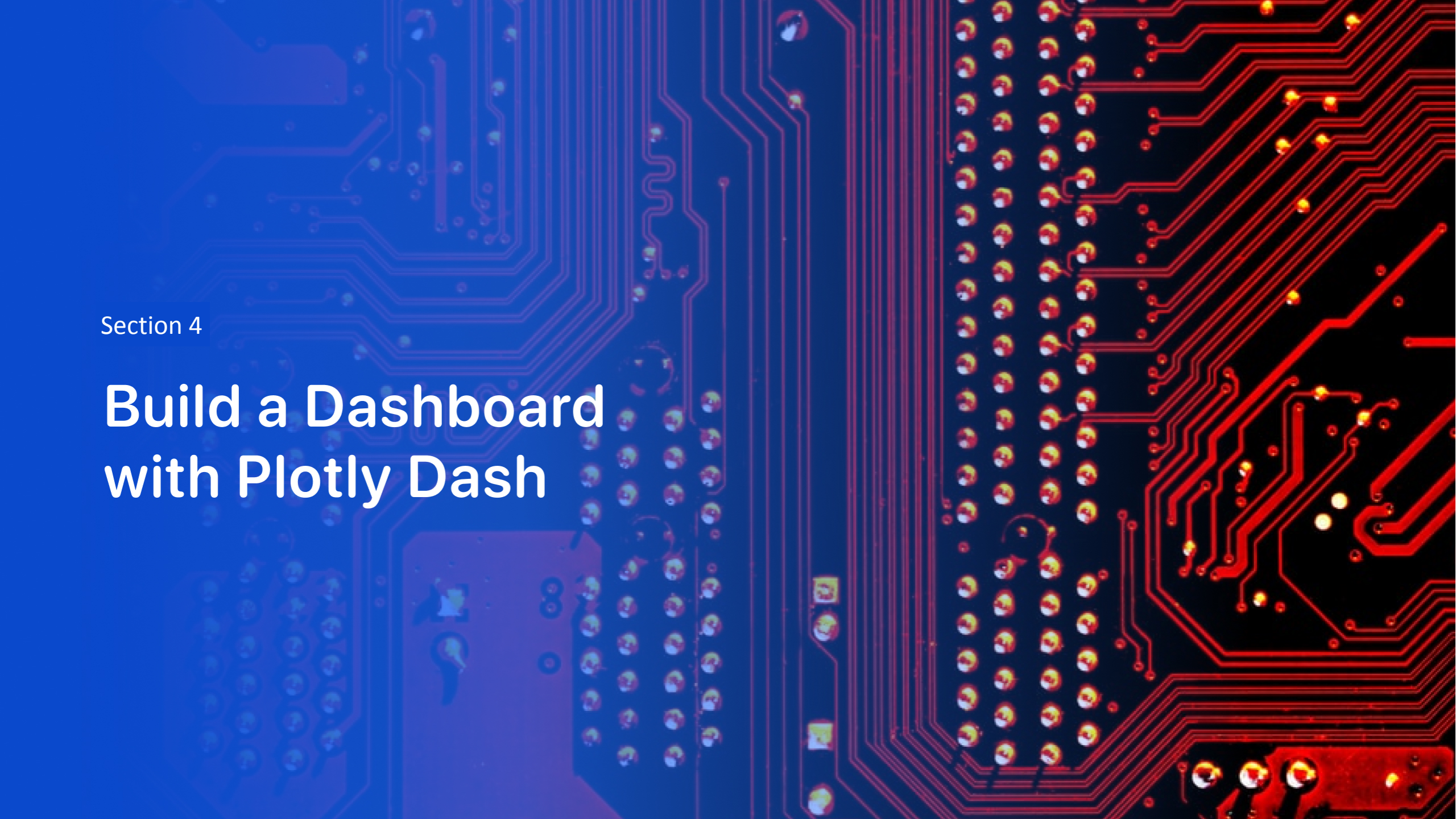


If we zoom even more, we can see the location of each launch outcome, green being success and red, failure.

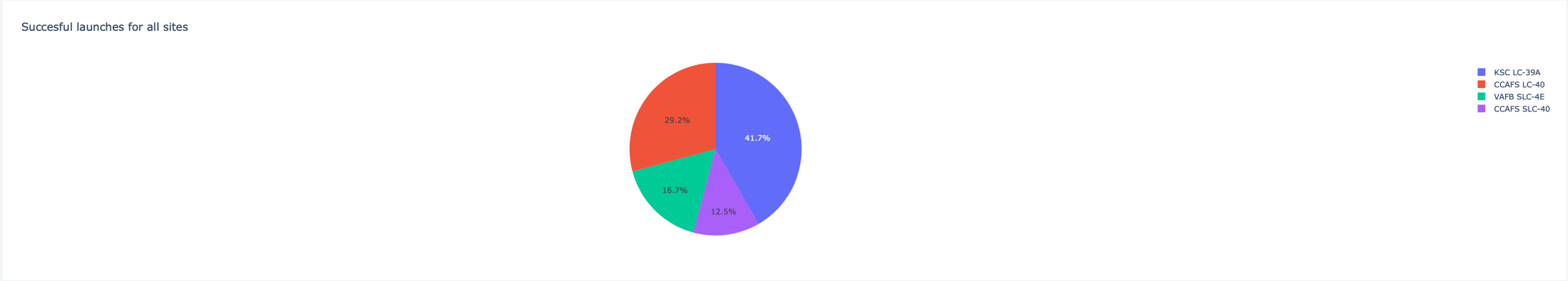
Folium distance markers



We have marked on the map the distance of the closest City, railway, highway and shore from the launch site CCAFS SLC-40 in Florida

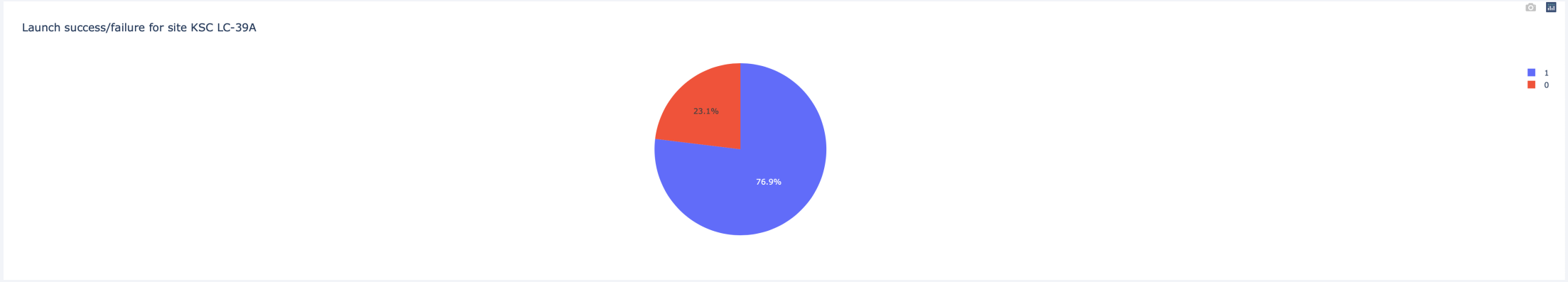

Section 4
Build a Dashboard with Plotly Dash

Success rate for all sites



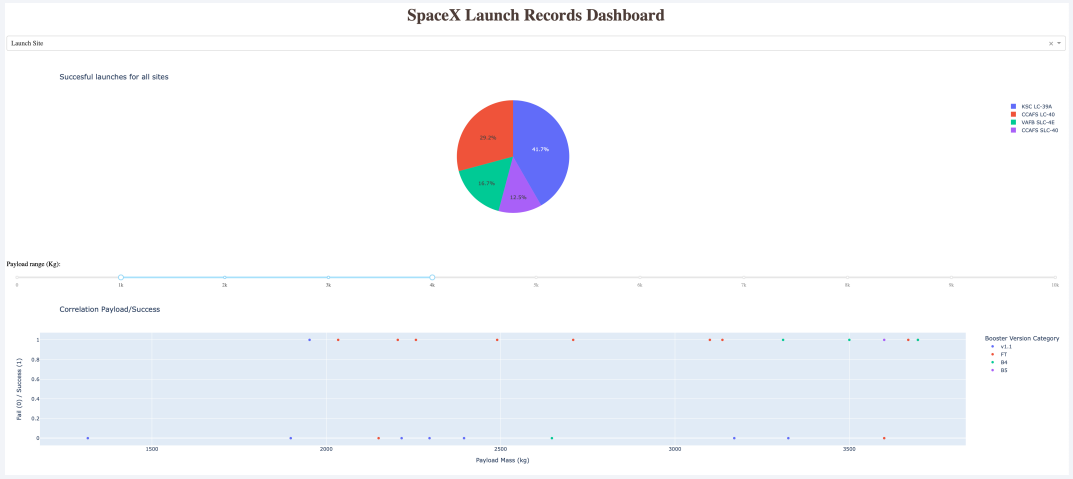
This pie chart displays to total success rate for all launch site

Highest success rate

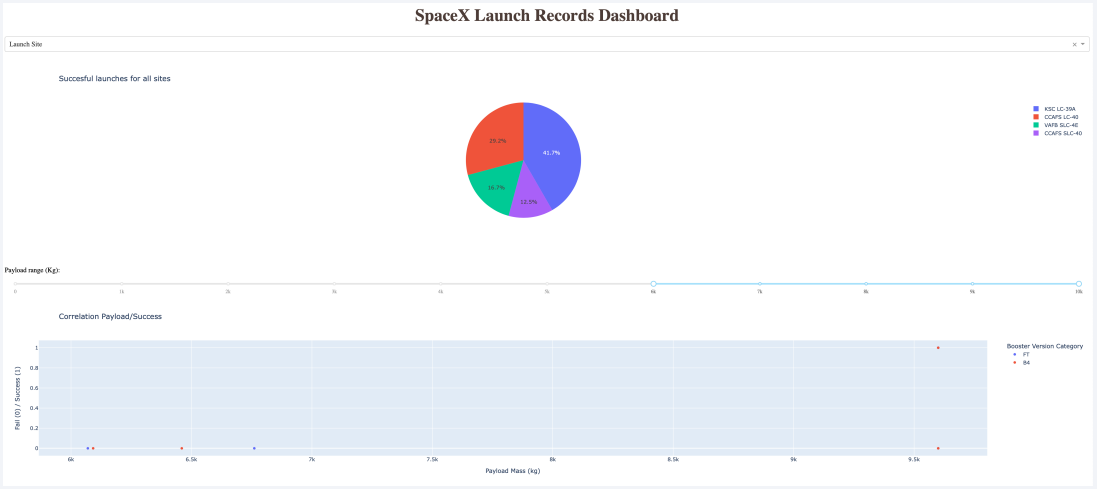


This pie chart displays the site having the highest success rate

Checking payload differences



Payloads between 1K and 5K Kg for all sites



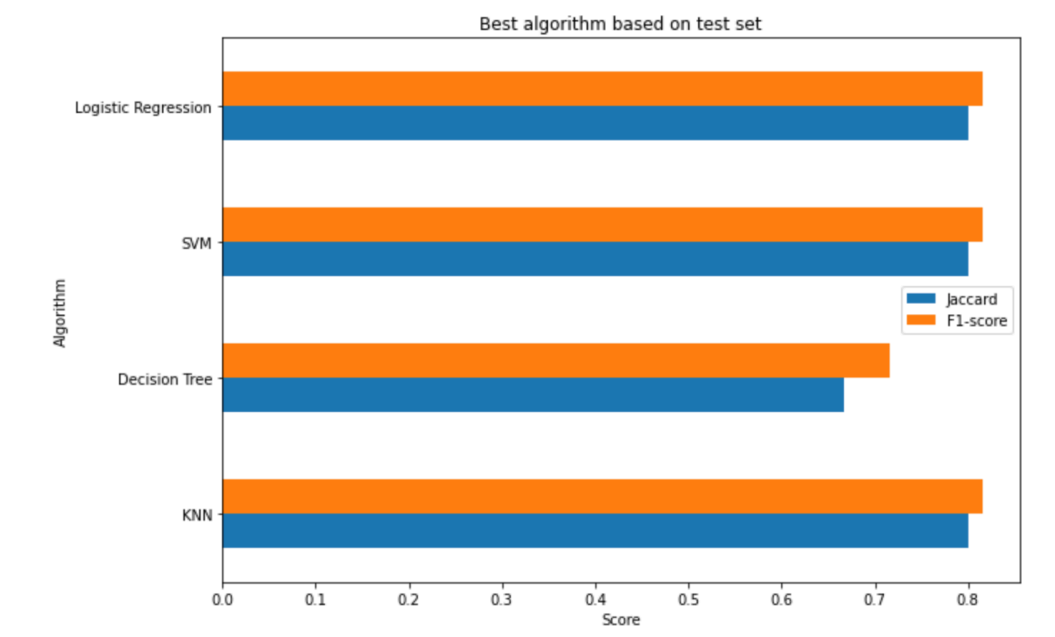
Payloads between 6K and 10K Kg for all sites

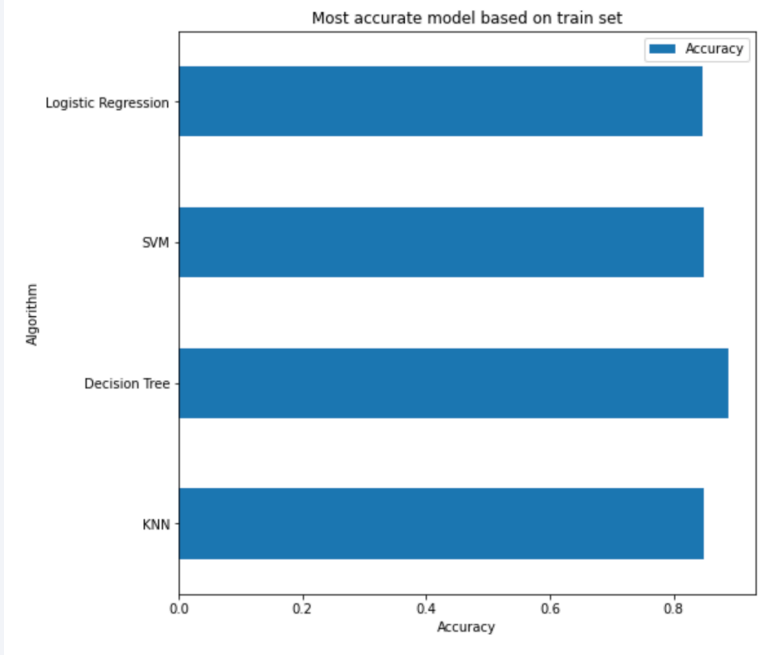
Using a slider, we can fine tune the payload filter to modify the scatter plot in real time

Section 5

Predictive Analysis (Classification)

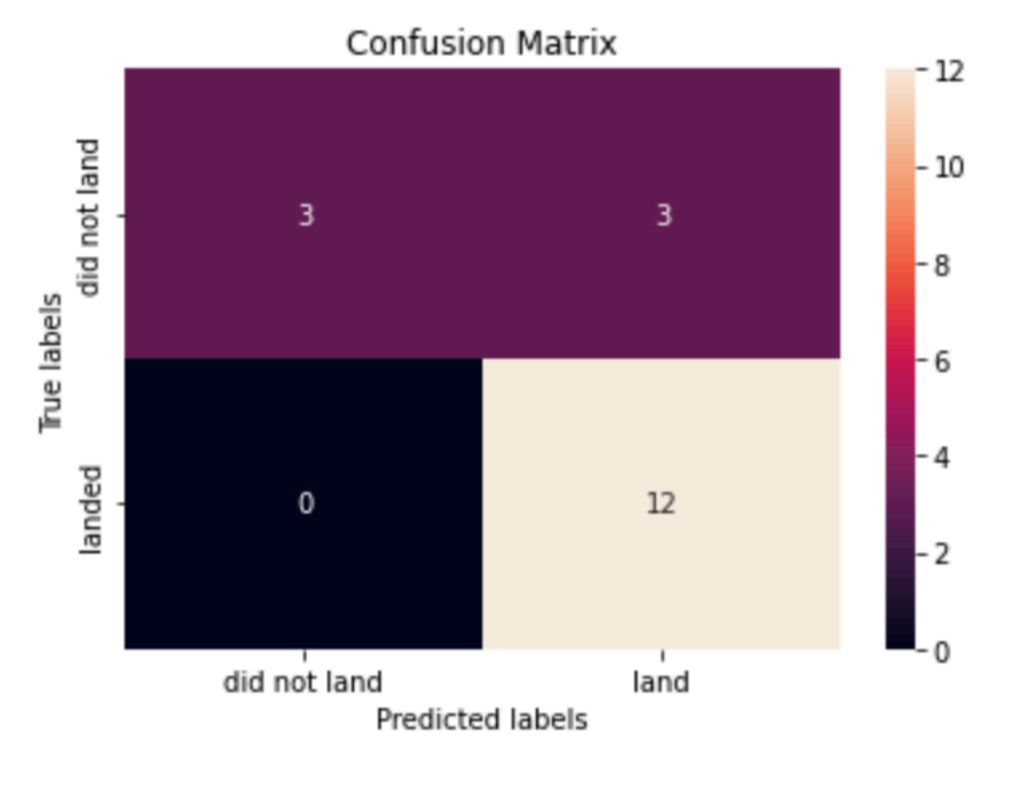
Classification Accuracy





The most accurate model with a train set is the decision tree but it is outperformed by the all other models when using a test set. Our analysis revealed the KNN algorithm to be the best model overall.

Confusion Matrix



This confusion matrix shows us that given the limited range of data, there will be a high rate of successful landing prediction with very little false positives.

Conclusions

- Using the models from these analysis, Space Y can predict a landing outcome of the first stage depending on the payload, the launch site and the orbit aimed with a confidence of 80%.
- To be more competitive, Space Y would need to analyze how many time the first stage can be reused; the more reuse, the more cost effective.
- The prediction model could be more efficient with more data available.

Appendix

- Falcon 9 information are available on SpaceX webpage :https://www.spacex.com/vehicles/falcon-9/
- The full set of Jupyter Notebooks is available on my Github: https://github.com/Tcoton/Applied-Data-Science-labs
- SpaceX Falcon 9 Wikipedia : https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
- DB2 extension for Jupyter Notebooks: https://github.com/DB2-Samples/db2jupyter

Thank you!