IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luong Hoai Nam
11/09/2023

Outline

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

Executive Summary

- In this project, we will predict if SpaceX's Falcon 9 rocket will successfully launch its first stage, to determine the cost of a launch, using their historical public information
- We will collect, wrangle, analyze that data and train machine learning model in order to accurately predict the probability

Introduction

- SpaceX, is an American spacecraft manufacturer, launch service provider and satellite communications company headquartered in Hawthorne, California. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Collect data using the API from https://api.spacexdata.com/v4/launches/past
  - Collect data using Web Scraping from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Create a Data Frame from relevant data extracted from those 2 source
- Perform data wrangling
  - Replace missing values in Payload Mass with the average value and create a new column to indicate if the first stage land successfully or not.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build 4 machine learning model and take turn evaluate each one with a range of parameters to determine which model with which parameters performs the best

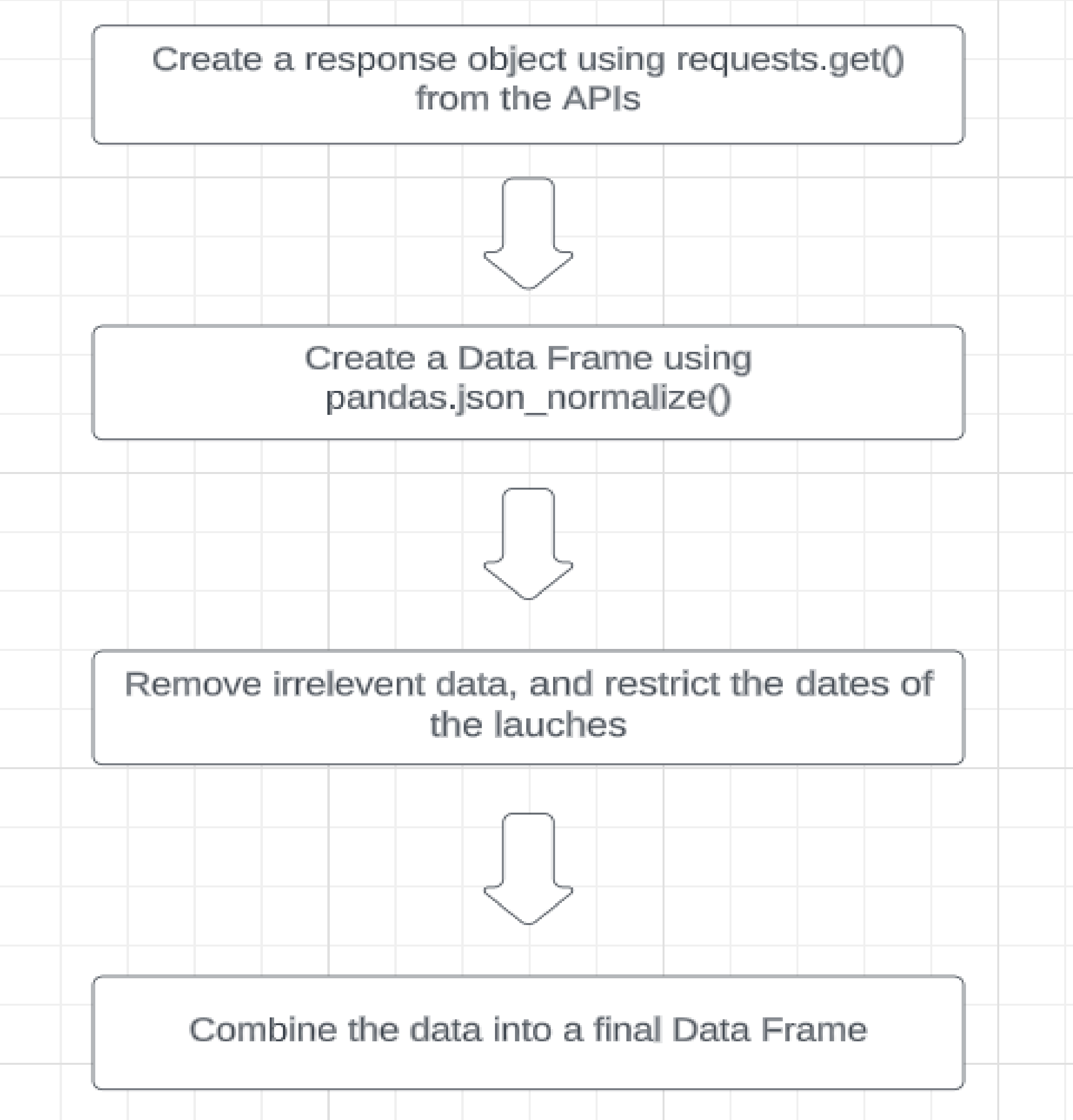
Data Collection

In general, we will try to collect public historical data of SpaceX's rocket launches, then create a data frame using only relevant data from those sources

Data Collection – SpaceX API

- GitHub URL for the notebook:

https://github.com/namL03/Data-Science-Capstone/blob/89b4f2e170ffa32e40c4c6a3683d0ddbc7206472/jupyter-labs-spacex-data-collection-api.ipynb



Create a response object using requests.get() from the APIs

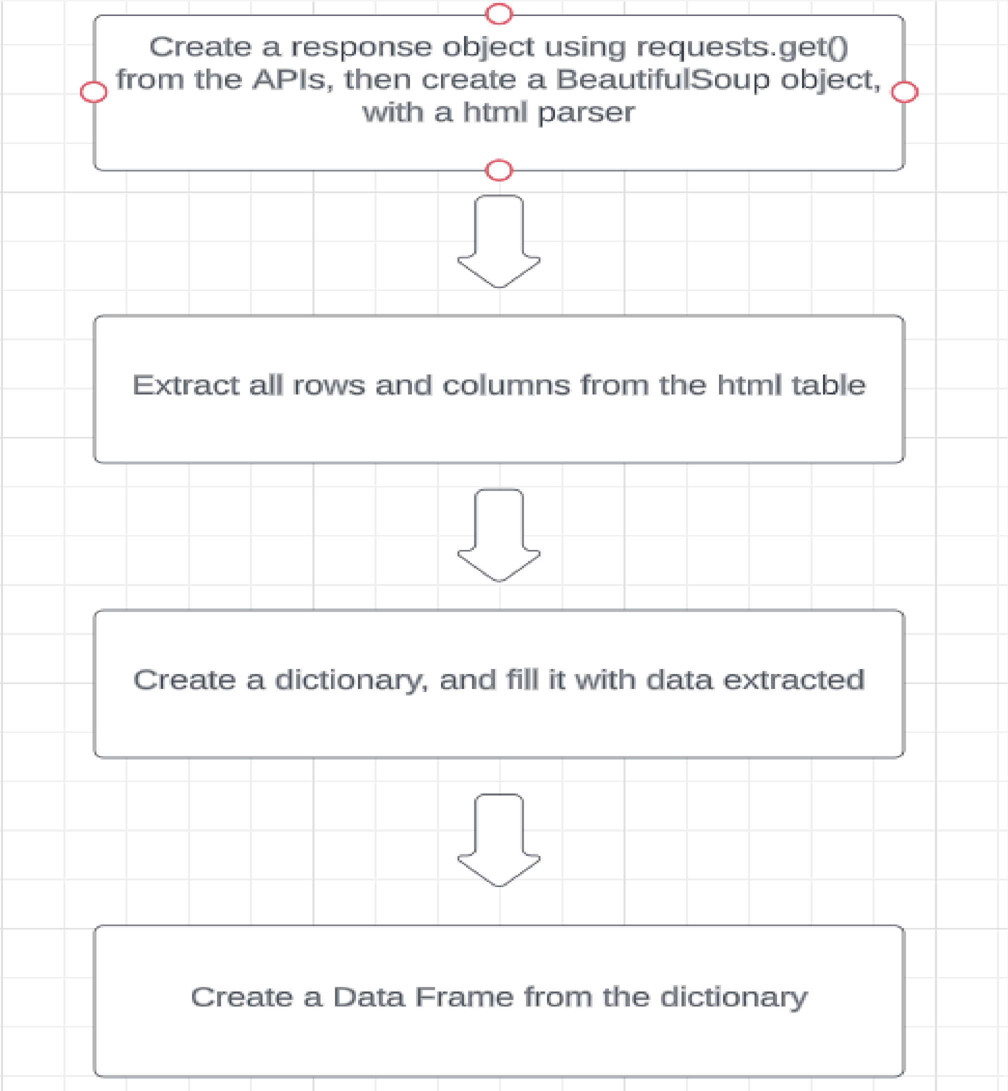
Create a Data Frame using pandas.json_normalize()

Remove irrelevent data, and restrict the dates of the lauches

Combine the data into a final Data Frame

Data Collection - Scraping

- GitHub URL for the notebook: https://github.com/namL03/Data-Science-Capstone/blob/89b4f2e170ffa32e40c4c6a3683d0ddbc7206472/jupyter-labs-webscraping.ipynb



Create a response object using requests.get() from the APIs, then create a BeautifulSoup object, with a html parser

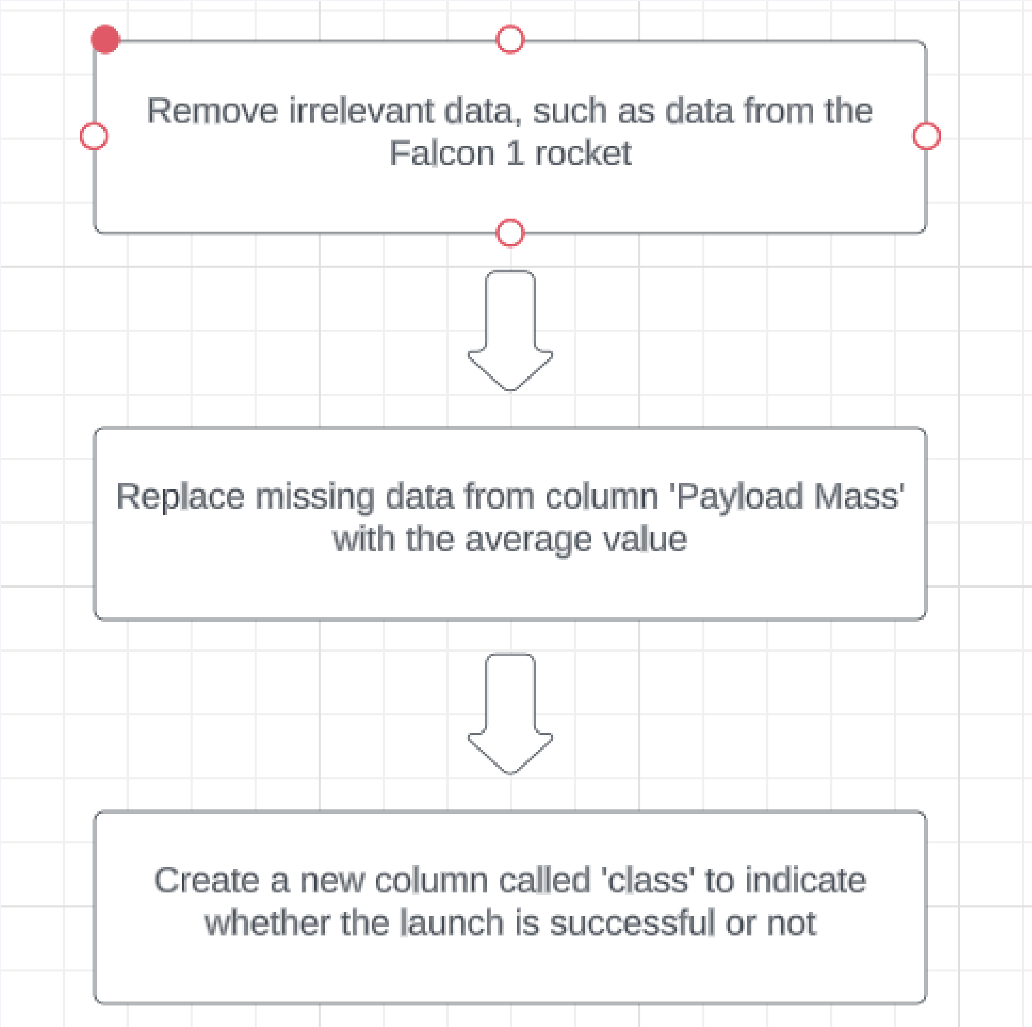
Extract all rows and columns from the html table

Create a dictionary, and fill it with data extracted

Create a Data Frame from the dictionary

Data Wrangling

- GitHub URL for the notebook: https://github.com/namL03/Data-Science-Capstone/blob/89b4f2e170ffa32e40c4c6a3683d0ddbc7206472/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



Remove irrelevant data, such as data from the Falcon 1 rocket

Replace missing data from column 'Payload Mass' with the average value

Create a new column called 'class' to indicate whether the launch is successful or not

EDA with Data Visualization

- Use a scatter plot to see how Flight number and Payload affect the launch outcome
- Use a scatter plot to see how Flight number and Launch site affect the launch outcome
- Use a scatter plot to see how Payload and Launch site affect the launch outcome
- Use a scatter plot to see how Flight number and Orbit type affect the launch outcome
- Use a scatter plot to see how Payload and Orbit type affect the launch outcome
- Use a line chart to see the success rate over the year
- Use a bar chart to compare the success rate of each Orbit
- GitHub URL: https://github.com/namL03/Data-Science-Capstone/blob/89b4f2e170ffa32e40c4c6a3683d0ddbc7206472/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Display the names of unique launch sites
- Display 5 records where launch sites begin with ‘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 successful landing outcome in ground pad was achieved
- 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
- 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: https://github.com/namL03/Data-Science-Capstone/blob/89b4f2e170ffa32e40c4c6a3683d0ddbc7206472/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Add a circle and a marker for each launch site to mark its location on the map
- Add a marker cluster with markers referring to each launch being successful (colored in green) or not (colored in red)
- Add a mouse position to find the coordinates on the map
- Add markers and Polylines to show the distance between a particular launch site to the nearest coastline, highway, railway or city
- GitHub URL: https://github.com/namL03/Data-Science-Capstone/blob/dc9ad3867ab0587b7761597a64269549544f7bc4/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Add pie charts to show the proportions of Success launch by Site and the number of Success launches for each Site
- Add scatter plots to show the correlation between Payload and Success for all Sites and for each Site
- GitHub URL: https://github.com/namL03/Data-Science-Capstone/blob/dc9ad3867ab0587b7761597a64269549544f7bc4/spacex_dash_app.py

Predictive Analysis (Classification)

- Firstly, reformat categorical features as numerical ones using one-hot-encoding
- Standardize the data using StandardScaler and divide the dataset into train and test set
- In turn, examine the 4 ML algorithm: Logistic Regression, SVM, Decision Tree and KNN. Use GridSearchCV to find the best parameters for each model. Test each model with the test set and examine the accuracy score, as well as the confusion matrix
- GitHub URL: https://github.com/namL03/Data-Science-Capstone/blob/dc9ad3867ab0587b7761597a64269549544f7bc4/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

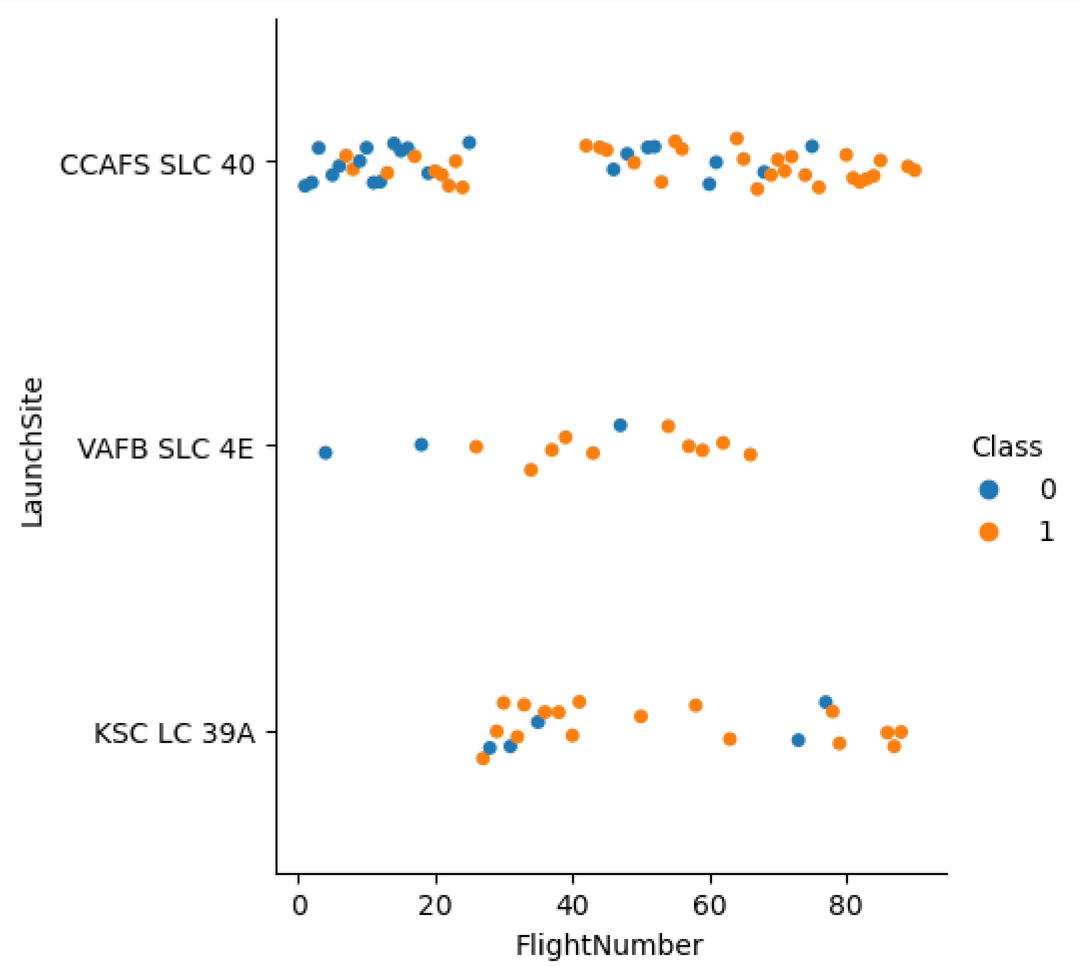
Results

- The massive the payload and the large the fight number, the first stage is more likely to land successfully
- ES-L1, GEO, HEO and SSO have the highest success rate of 100% among all Orbit types
- The general success rate kept increasing from 2010 until 2020
- KNN, Logistic Regression and SVM produce the same accuracy of 83.3% and the same confusion matrix, while that amount of Decision Tree is 94.4% with best parameters
- The major problem of those models is false positives

Section 2

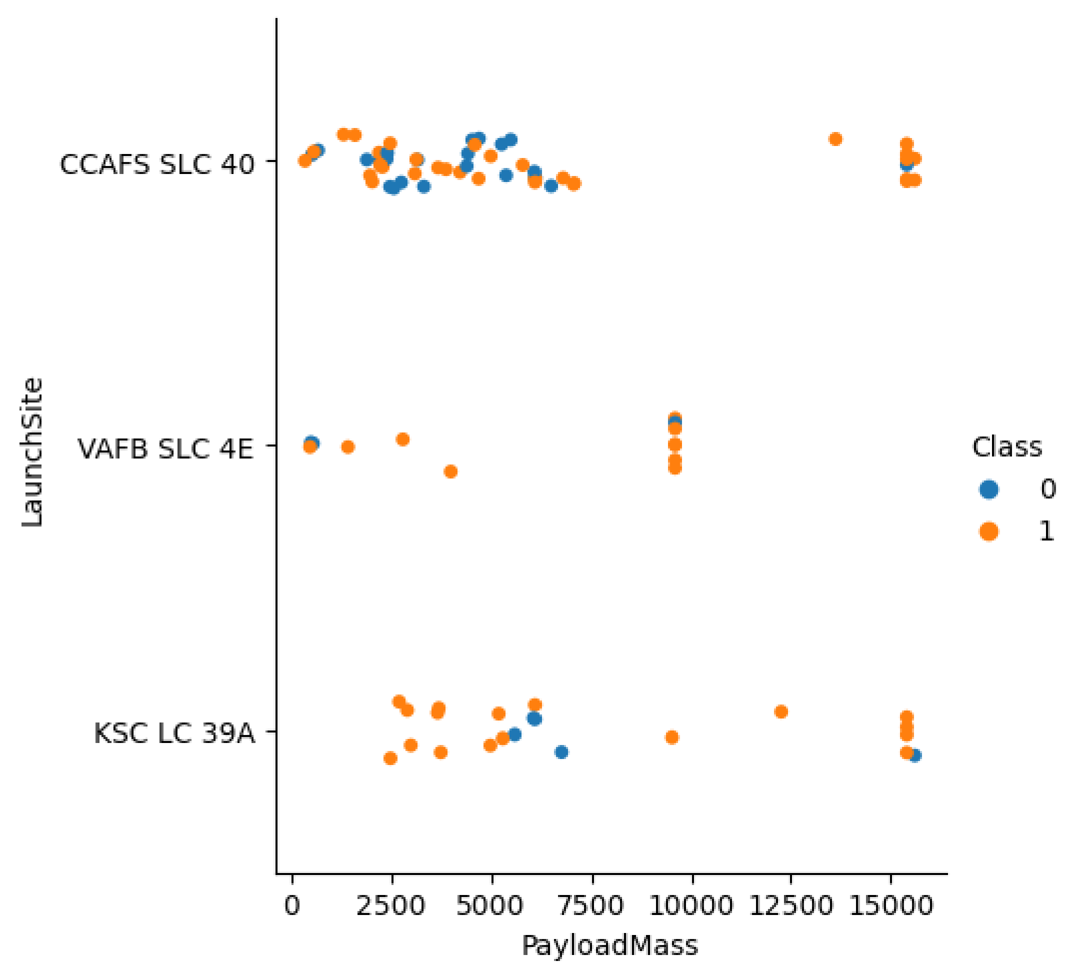
Insights drawn from EDA

Flight Number vs. Launch Site



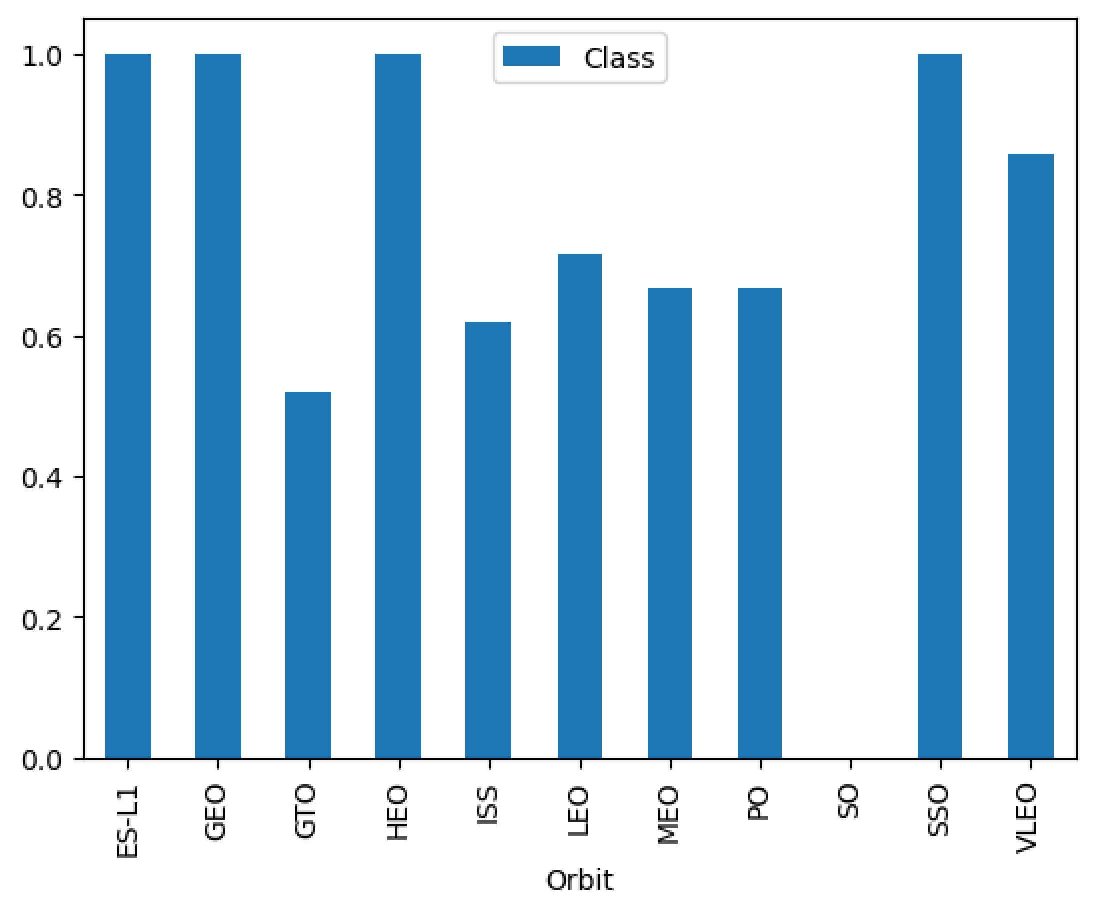
- The higher the Flight Number, the more successful the launch would be
- KSC LC 39A has the highest successful chance of 77.2%

Payload vs. Launch Site



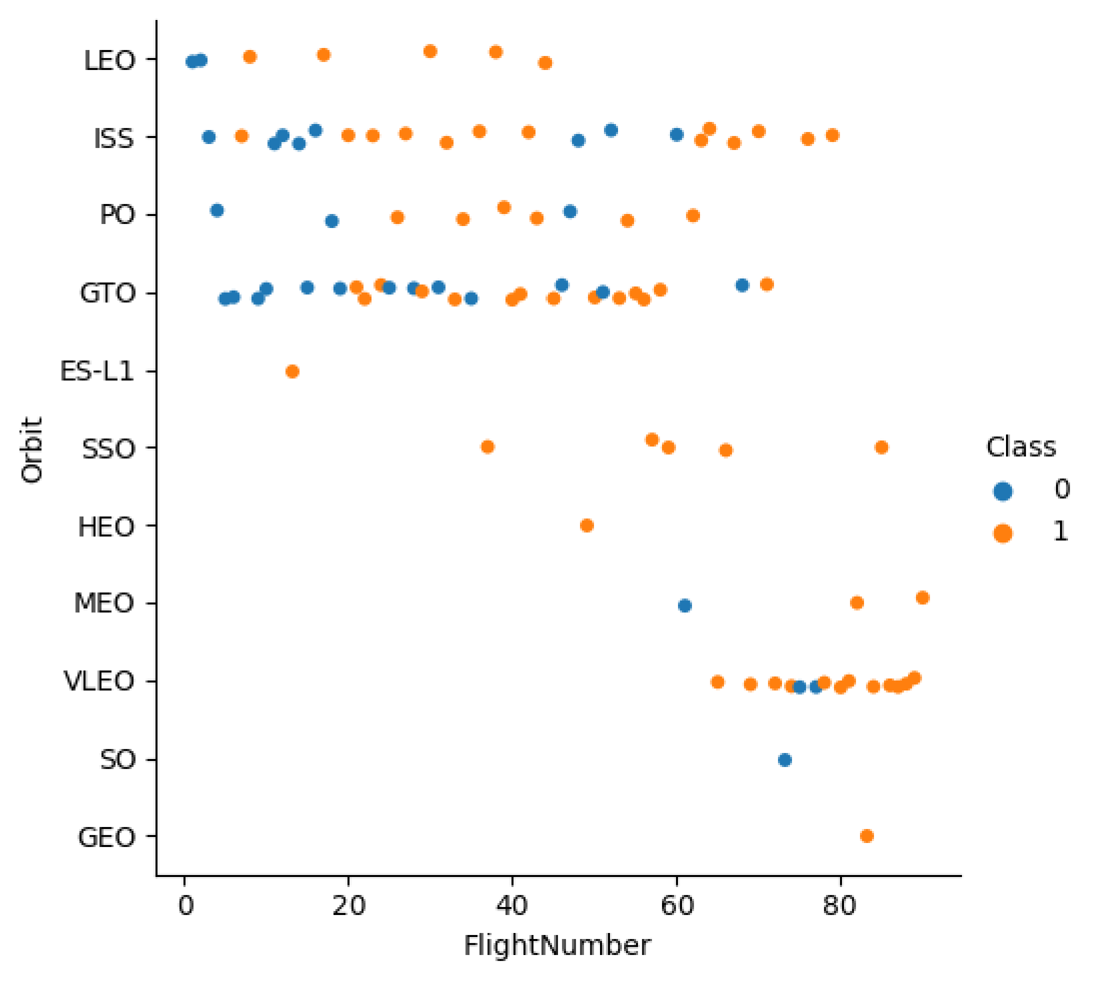
- It seems that the larger the Payload Mass, the higher chance that the launch at CCAFS SLC 40 would be successful
- However, at the other two Launch Sites, the Payload seems to have little effect on the probability
- For some reasons, launches at VAFB SLC 4E don’t exceed 10000 kg Payload

Success Rate vs. Orbit Type



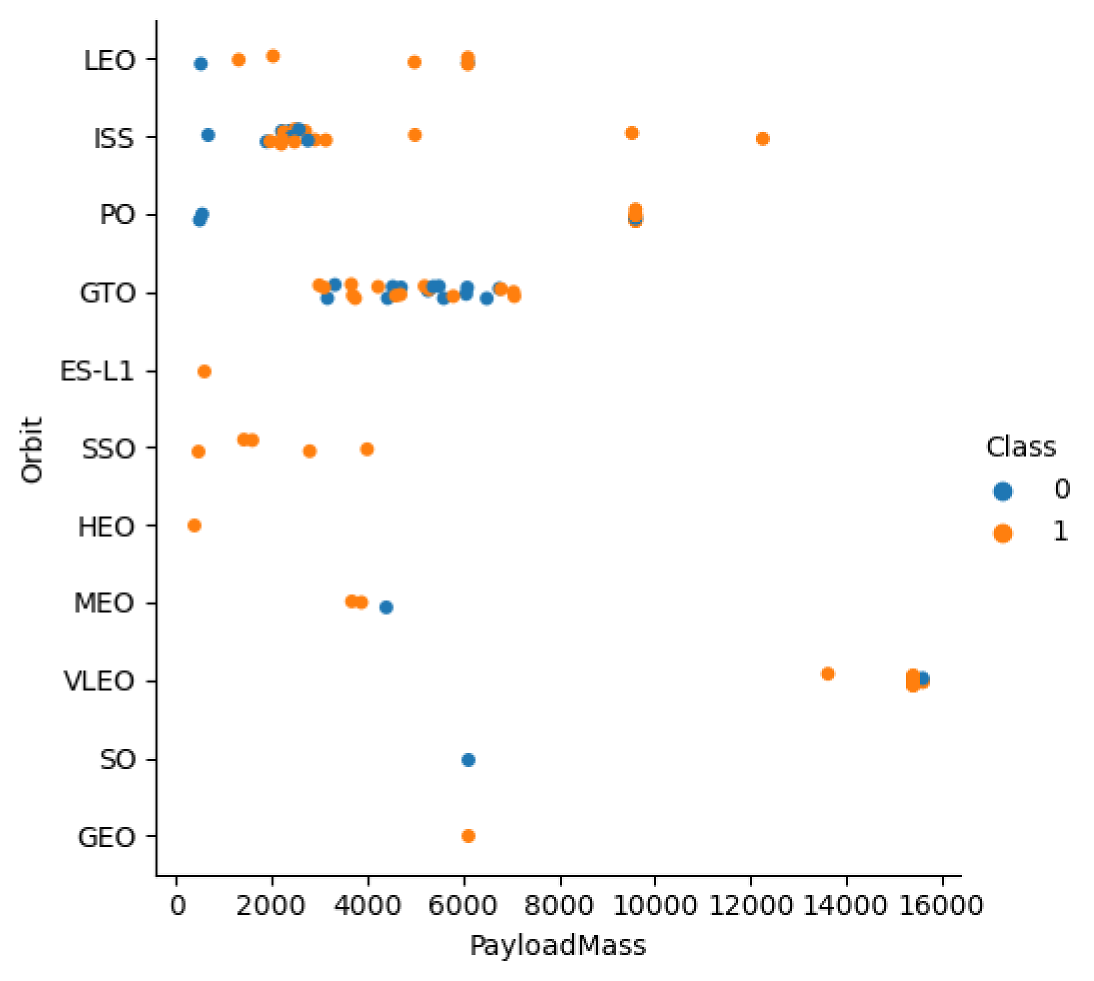
- ES-L1, GEO, HEO and SSO have the highest successful chance of 100%, while SO has the lowest of 0%
- However, as we could not see the exact number of each type’s launches, we cannot really tell which orbit type is better than one another

Flight Number vs. Orbit Type



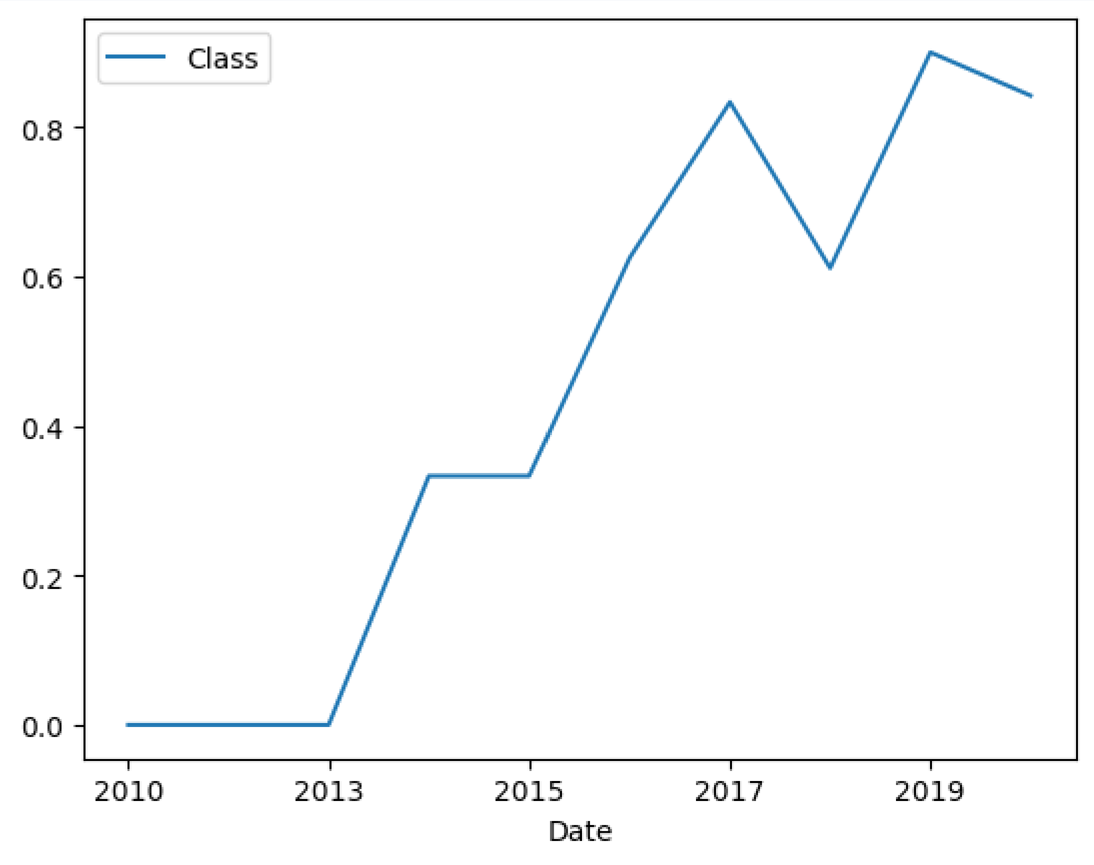
- With VLEO, LEO and ISS, it seems like the higher the Flight Number, the more successful the launch would be
- On the other hand, Flight Number has little effect on GTO's launches
- There is too few launches for the other orbit types, so not much information could be extracted from them

Payload vs. Orbit Type



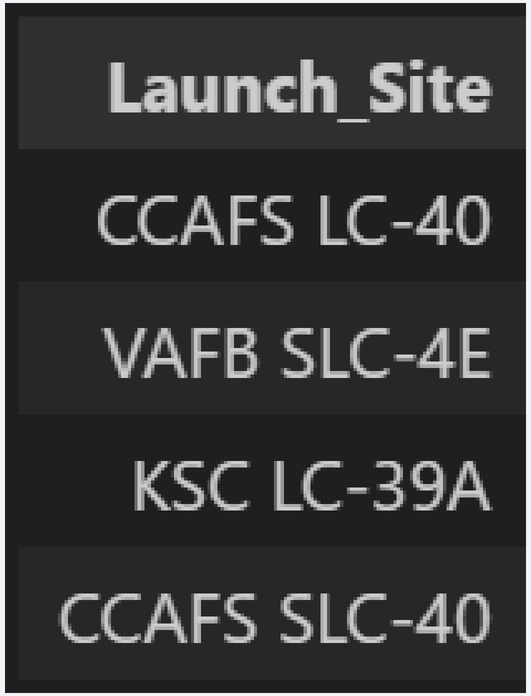
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
- However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing are both there here

Launch Success Yearly Trend



The successful rate has an increasing trends over the years from 2013, except for 2 period: 2017 – 2018 and 2019 - 2020

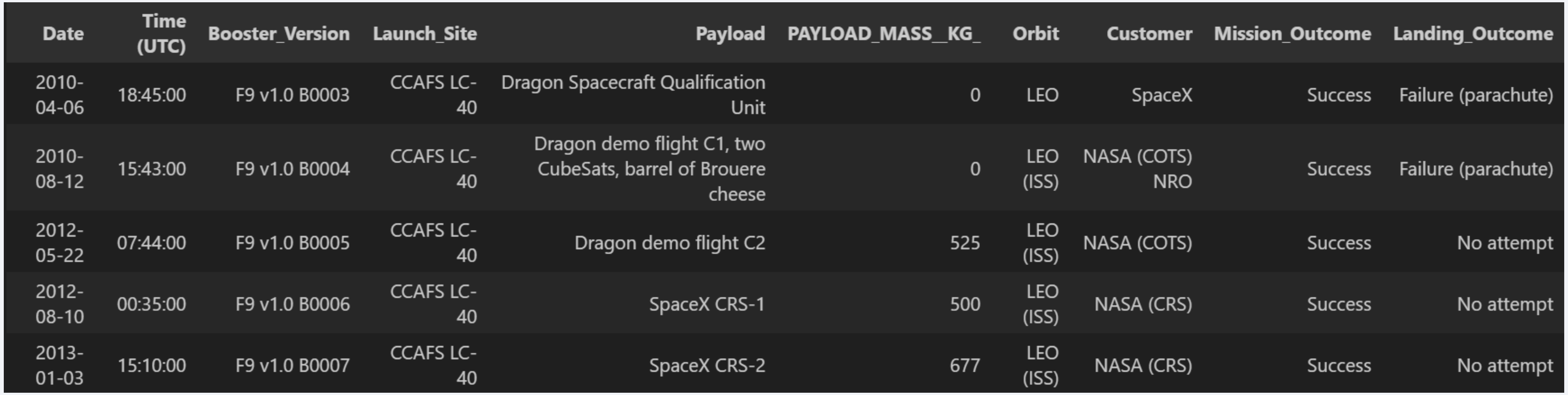
All Launch Site Names



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

All the launches are executed at 1 of 4 launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFAS SLC-40

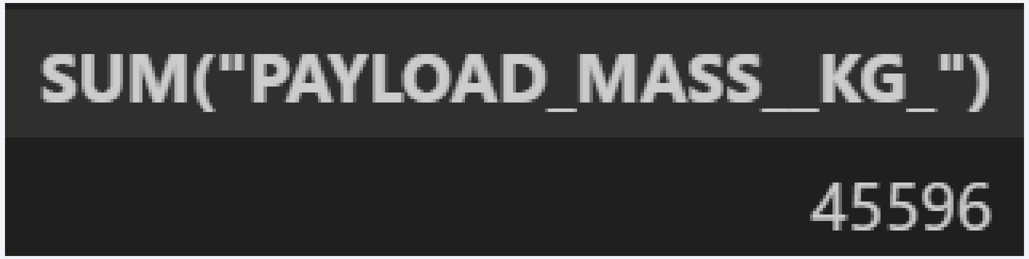
Launch Site Names Begin with 'CCA'



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

Information about 5 launches at a launch site begin with ‘CCA’

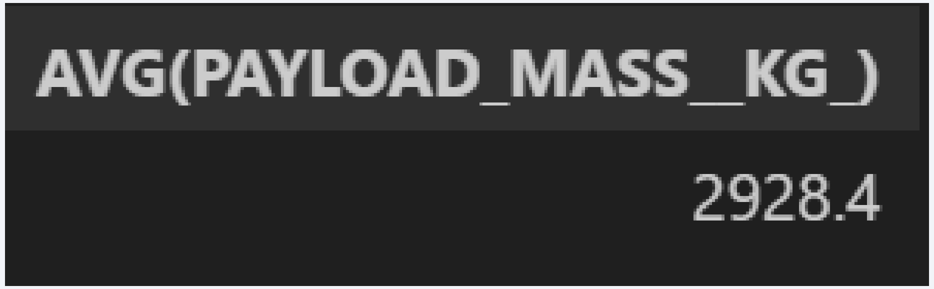
Total Payload Mass



SUM("PAYLOAD_MASS__KG_")
45596

Total Payload of all launches is 45596 kg

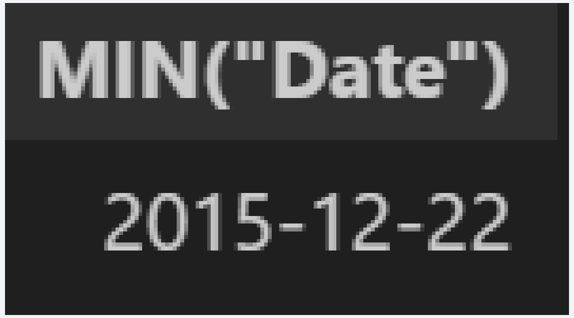
Average Payload Mass by F9 v1.1



AVG(PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass by F9 v1.1 is 2928.4 kg

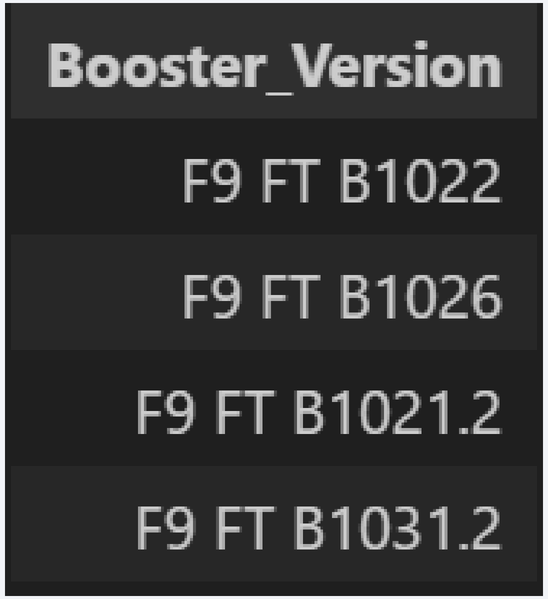
First Successful Ground Landing Date



MIN("Date")
2015-12-22

The first successful ground landing is on December 22nd, 2015

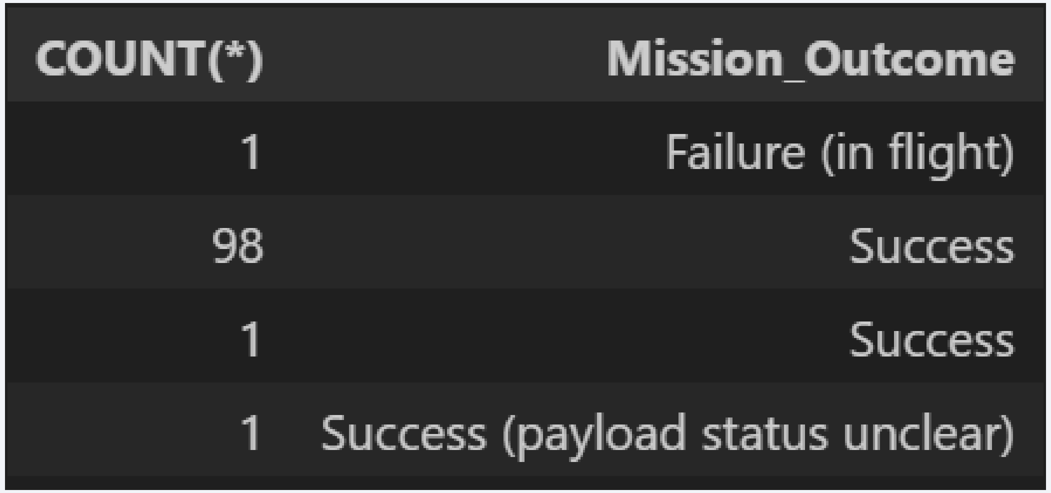
Successful Drone Ship Landing with Payload between 4000 and 6000



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

These are the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

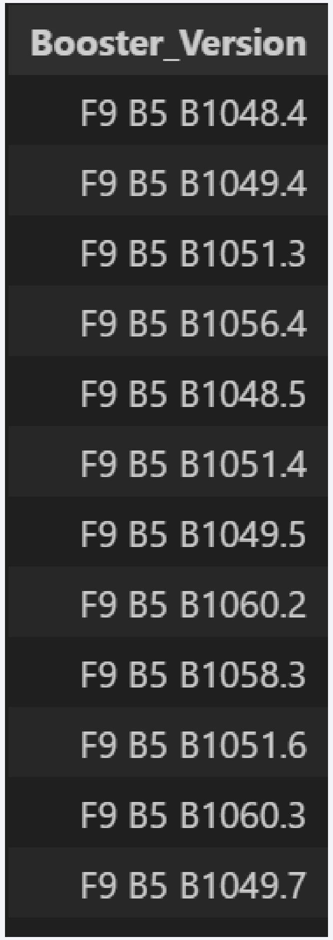
Total Number of Successful and Failure Mission Outcomes



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

Only 1 out of 101 launches failed in flight, while the other 100 succeeded, with 1 has an unclear payload status

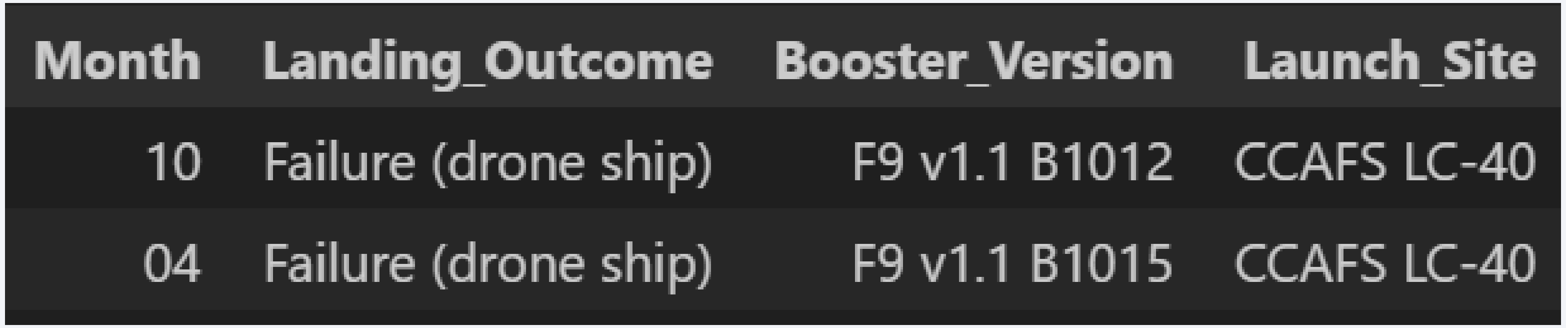
Boosters Carried Maximum Payload



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

These are the names of the boosters which have carried the maximum payload mass

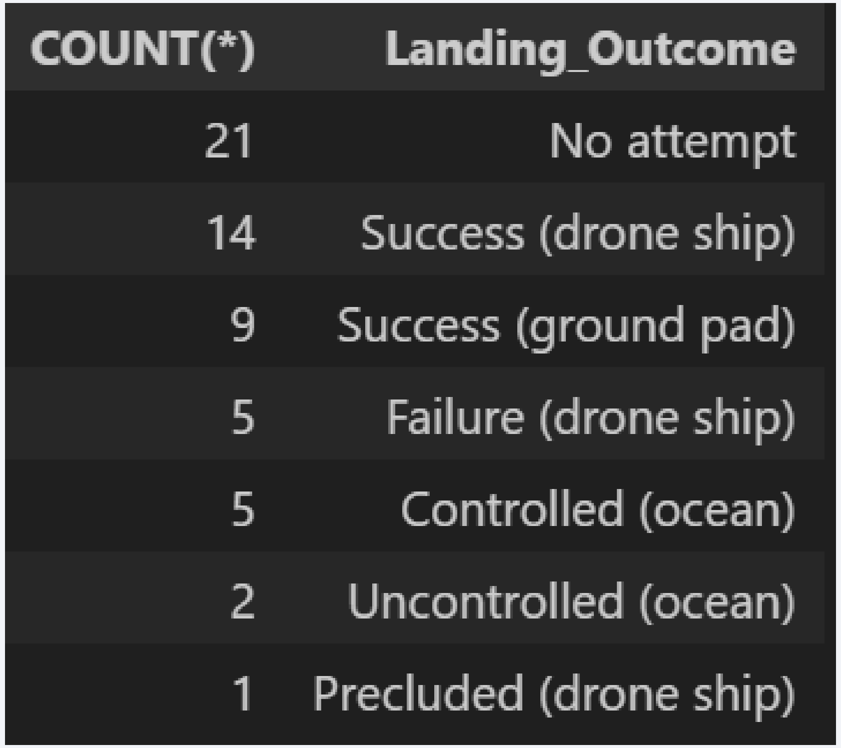
2015 Launch Records



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

Launches occurred in 2015

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



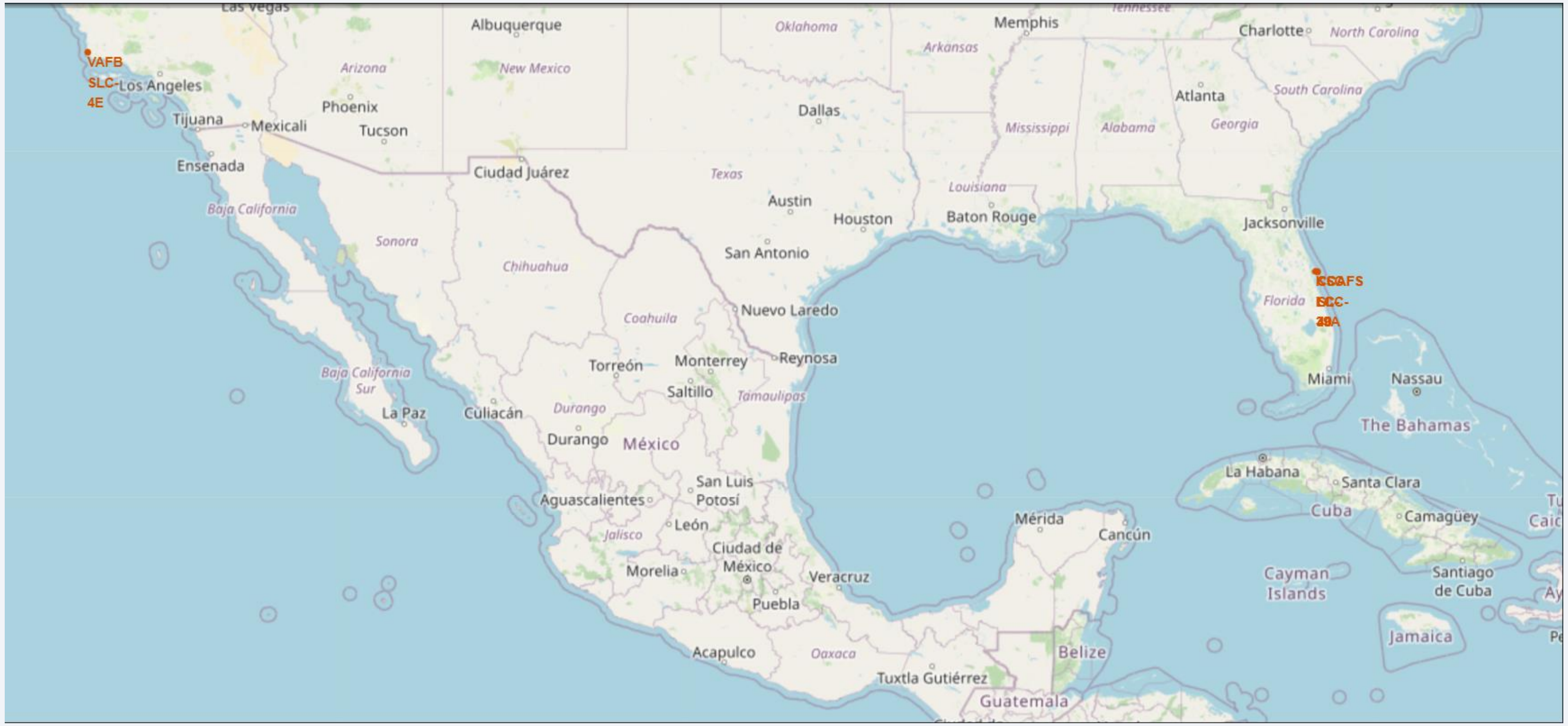
COUNT(*)	Landing_Outcome
21	No attempt
14	Success (drone ship)
9	Success (ground pad)
5	Failure (drone ship)
5	Controlled (ocean)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Types of Landing Outcome and the number of launches for each type, from June 4th, 2010 to March 20th, 2017

Section 3

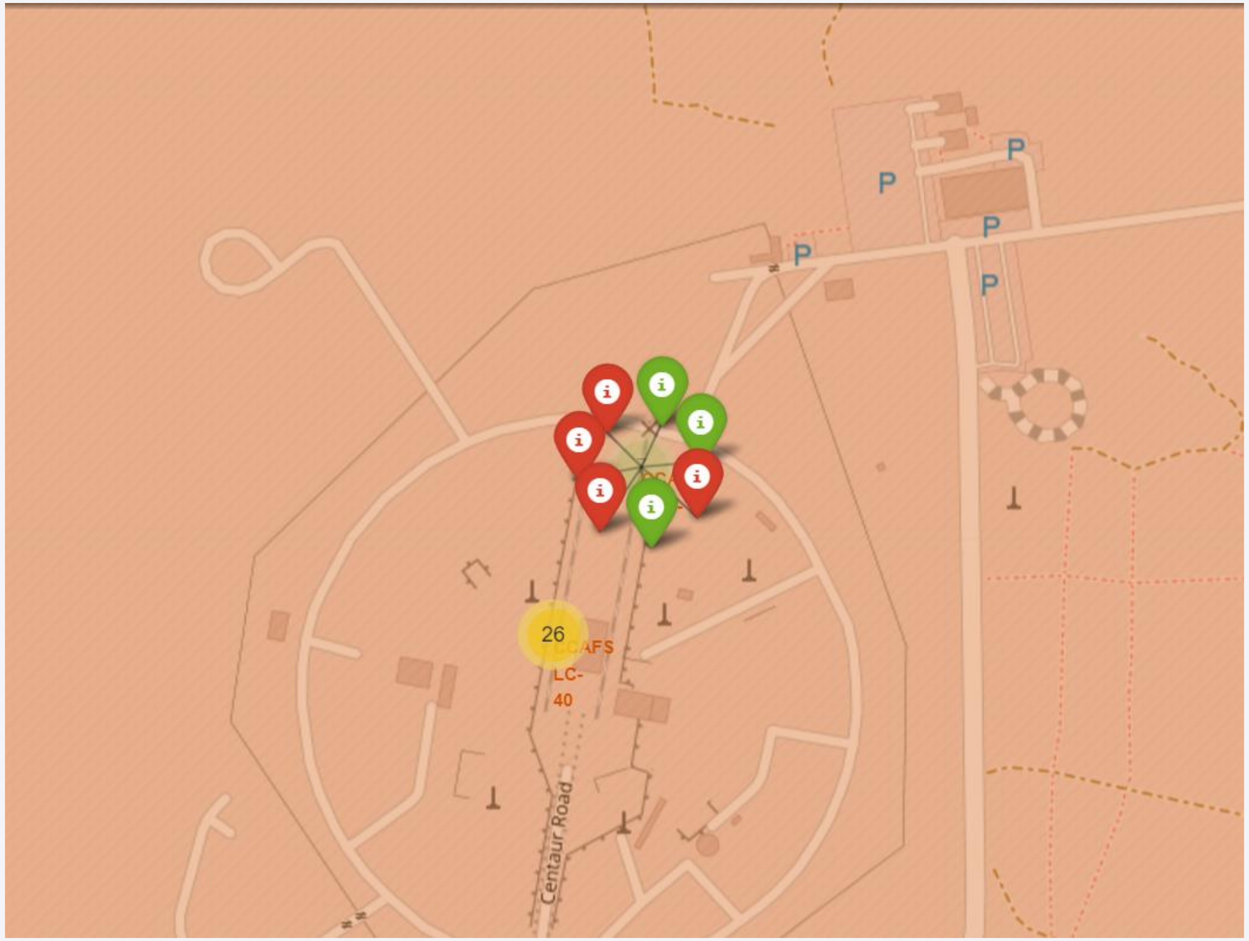
Launch Sites Proximities Analysis

Location of Launch Sites on the World Map



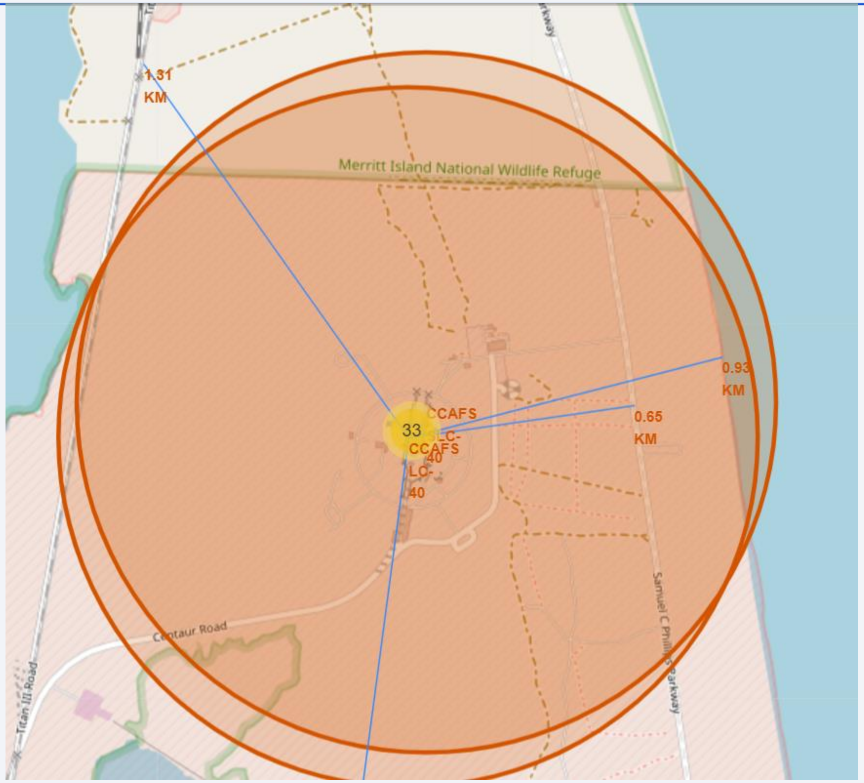
We can see that VAFB SLC-4E located quite far from the other three launch sites

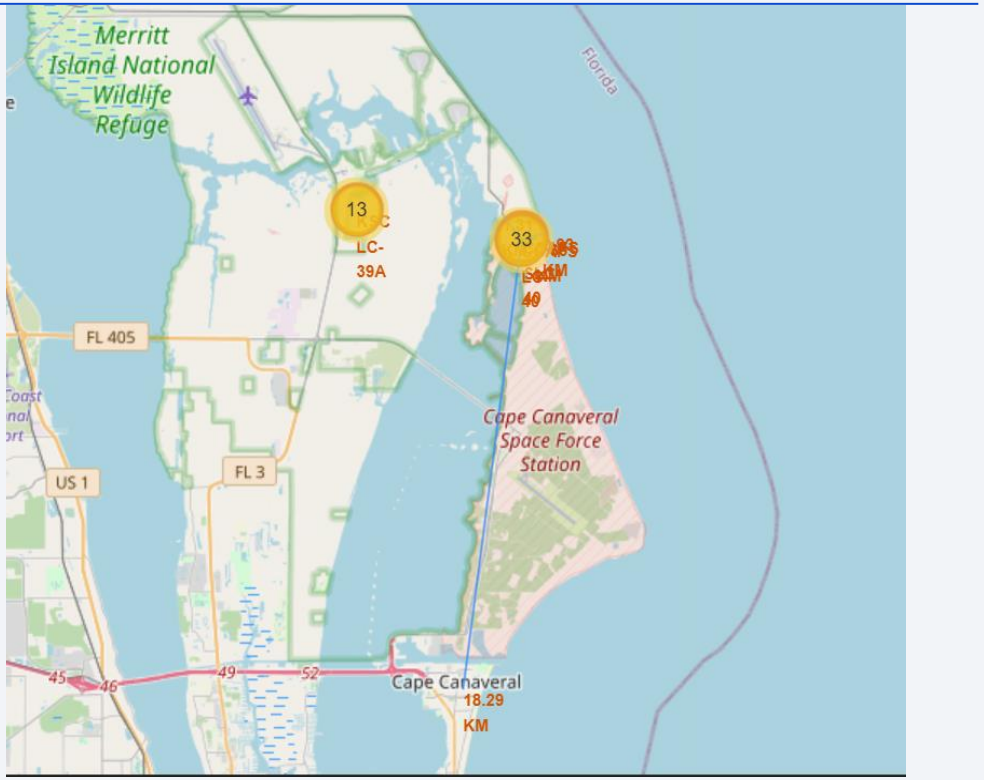
Successful and failed launches distinguished by color



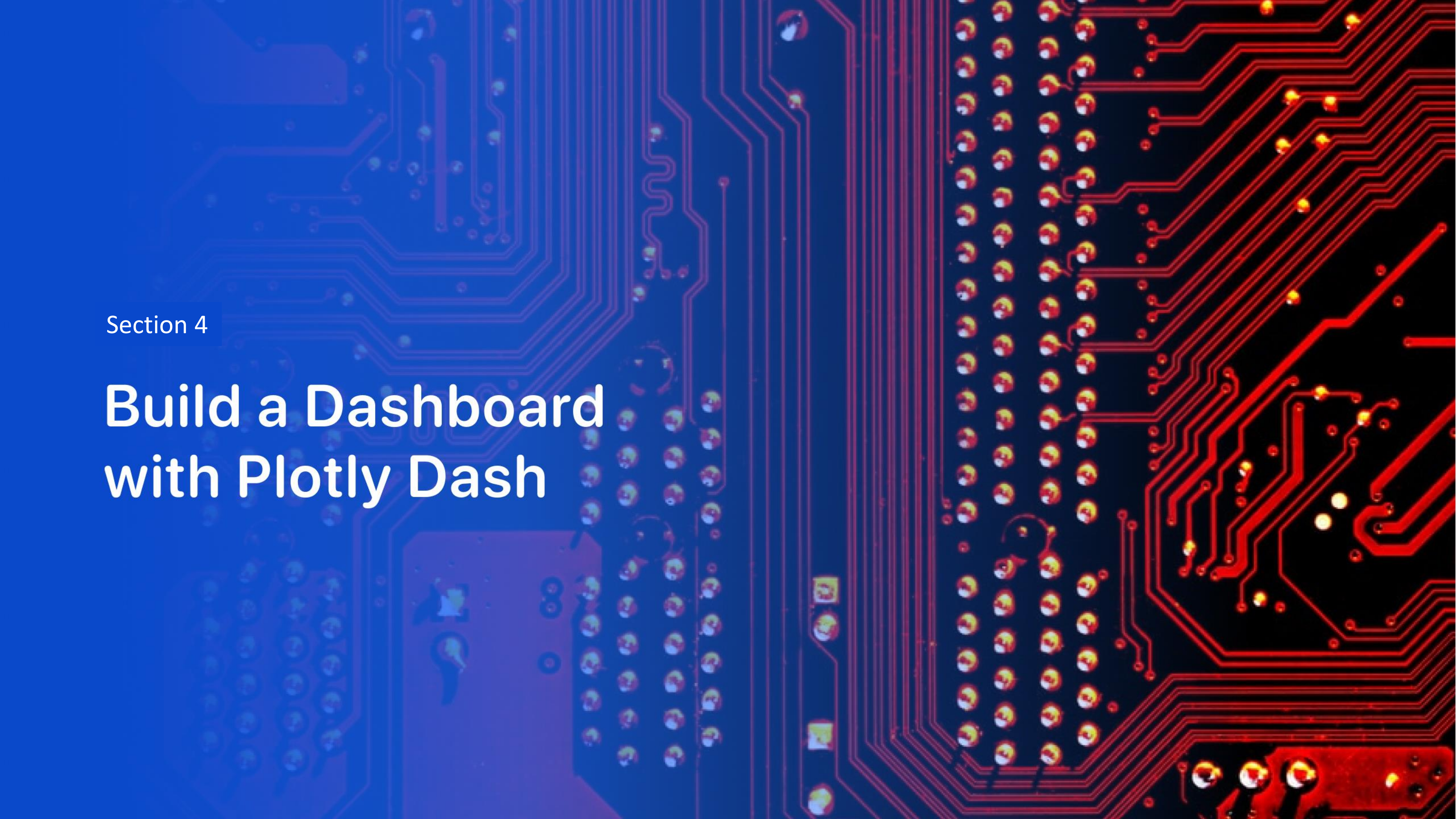
We can see that there are 3 successful and 4 failed launches at CCAFS SLC-40

Distance between CCAFS LC-40 and its nearest coastline, railway highway and city





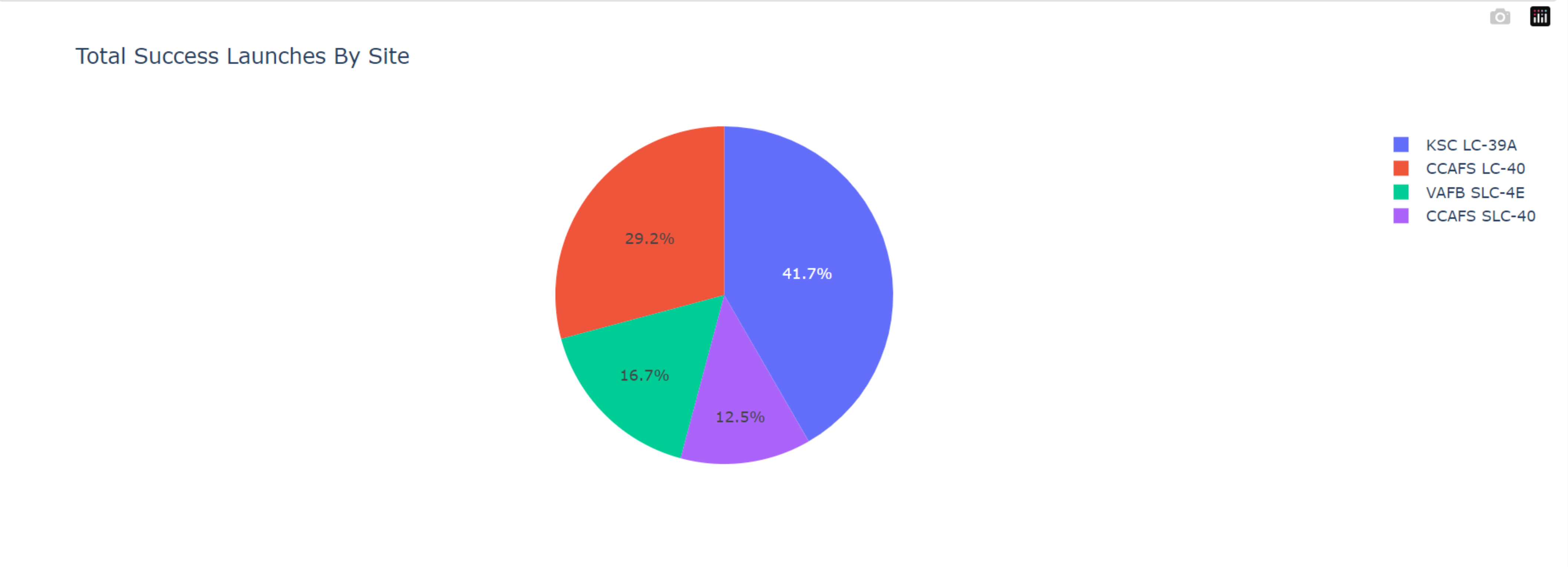
CCAFS LC-40 is located pretty near railway, highway and coastline but far from city



Section 4

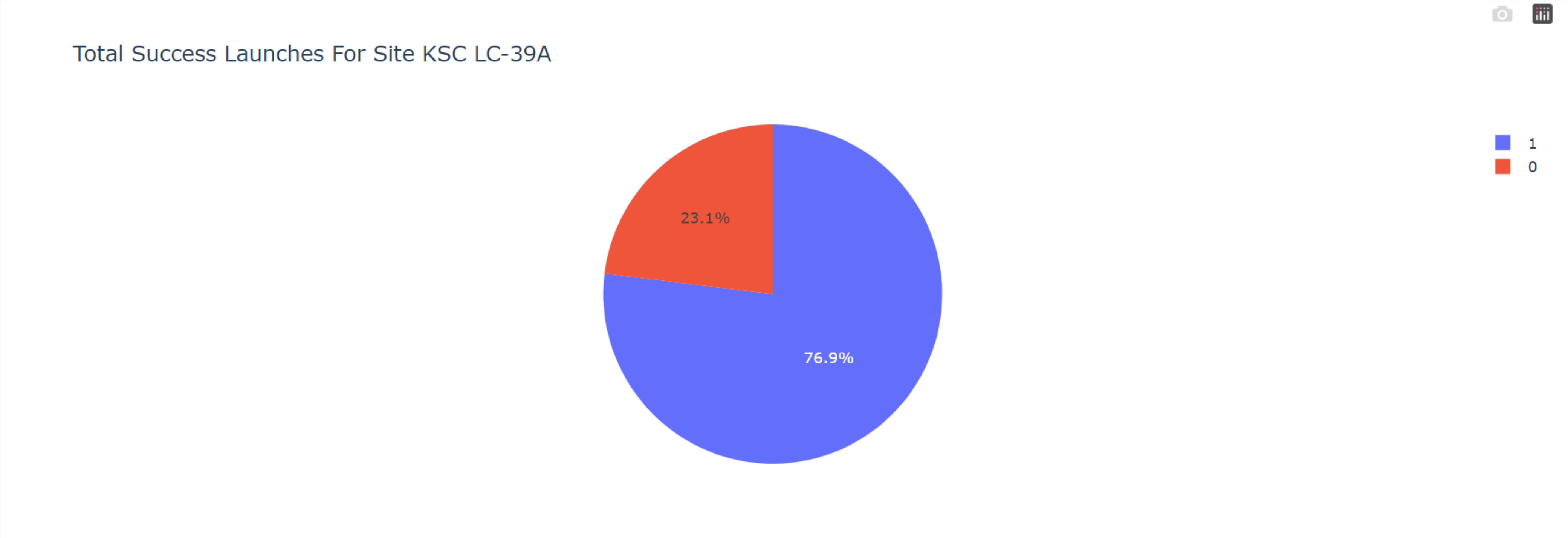
Build a Dashboard with Plotly Dash

Percentages of successful launches by Site



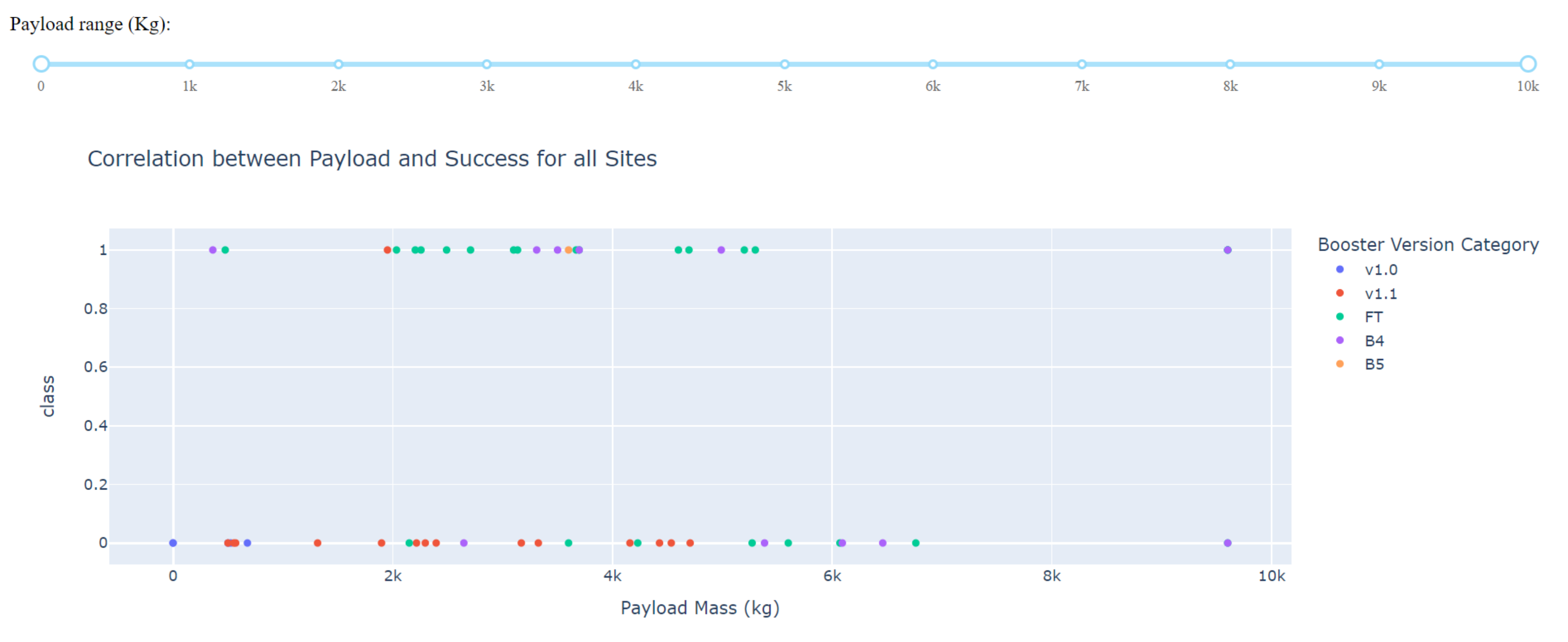
KSC LC-39A has the highest proportions (41.7%), while CCAFS SLC-40 has the lowest (12.5%)

Percentages of successful launches at KSC LC-39A



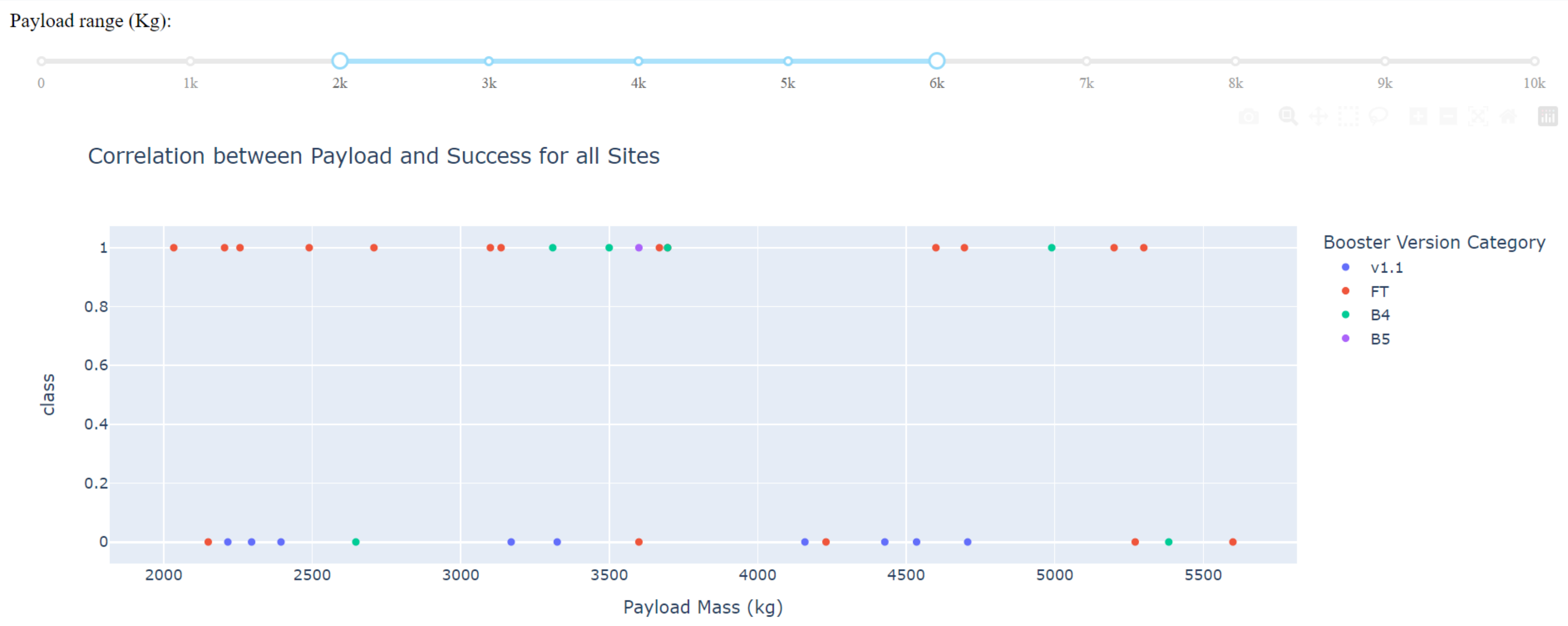
76.9% of KSC LC-39A’s launches are successful, which is the highest among 4 sites

Payload vs. Booster Version (at 0-10000 kg range)


Payload range (Kg):

- B5 has the highest successful chance at 100%; however, there is only 1 launch using this version
- 65% of launches (13 out of 20) using FT are successful, which is the second highest among 5 booster version

Payload vs. Booster Version (at 2000-6000 kg range)



Payload range (Kg):

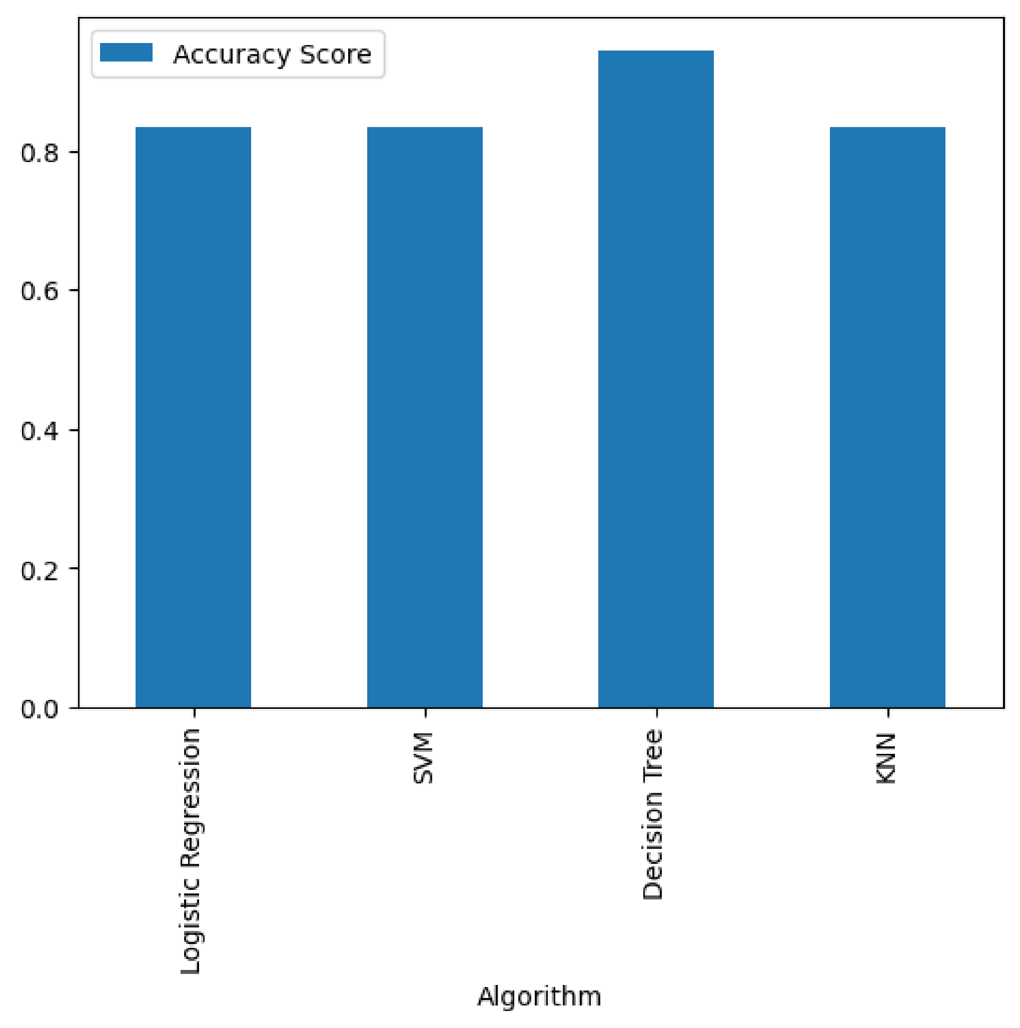
Correlation between Payload and Success for all Sites

70% of launches that has the Payload Mass between 3000 and 4000 kg are successful

Section 5

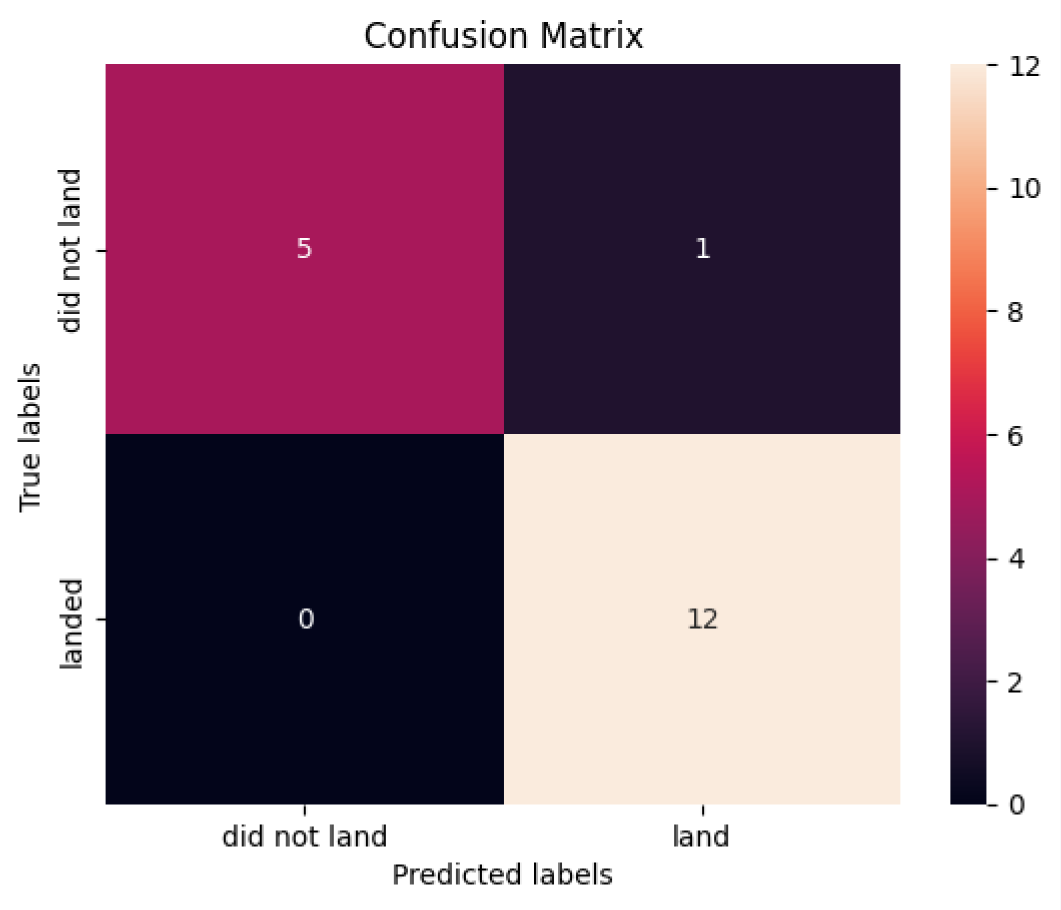
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree Model produces the highest accuracy score (at 94.4%), while the other 3 models all share the same accuracy at 83.3%

Confusion Matrix



Only 1 failed launch is wrongly predicted as successful, while all the other launches in the test set are accurately predicted

Conclusions

- The massive the payload and the large the fight number, the first stage is more likely to land successfully
- The successful rate has an increasing trends over the years from 2013, except for 2 period: 2017 – 2018 and 2019 – 2020
- VAFB SLC-4E located quite far from the other three launch sites
- Launch Site is located near railway, highway and coastline but far from city
- Decision Tree Model performs the best with an accuracy score of 94.4% with the exact parameters
- The major problem with all models is false positives

Appendix

Here is the parameters with the highest accuracy for the Decision Tree Model:

{'criterion': 'entropy', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}

accuracy : 0.9017857142857142

We can see that the splitter is random, so the model would not be consistent after each rebuild. As a result, when you run the notebook provided, you might not get the desired accuracy and the confusion matrix would not be the same as the one I provided.

You need to rebuild the model multiple times until you get the accuracy score high enough for the test set.

Thank you!