Data Science Capstone project

Arnaud VIGNERON

31/08/2021

Table of contents

Executive Summary

- Summary of methodologies :
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis
  - Interactive visual analytics
  - Predictive analysis

- Summary of all results :
  - Success rate increase since 2013 and is around 85% in 2020
  - Success rate is impacted by the Launch Site, the number of flights, the payload and the orbit
  - The prediction of the Falcon 9 first stage landing successfully or not has a 83% accuracy

Introduction

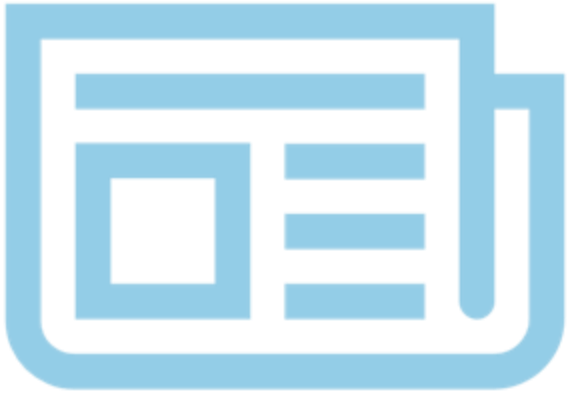
- Project background and context :

The new rocket company Space Y want to compete with Space X. And the reason for which Space X has great results is that they can reuse the first stage of the Falcon 9.

- Problems :

We want to know the cost of each launch and for that we need to know if the first stage will land successfully or not.

Methodology



- Data collection methodology:
  - Get request to the SpaceX API
- Perform data wrangling
  - Dealing with missing values, creating new columns
- 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

Methodology

Data collection

- To get the data we used a Get request to the Space X API

  - We used the api.spacexdata.com/v4/launches/past endpoint to get past launch data

- Webscraping from Wikipedia :

  - We used the page :

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

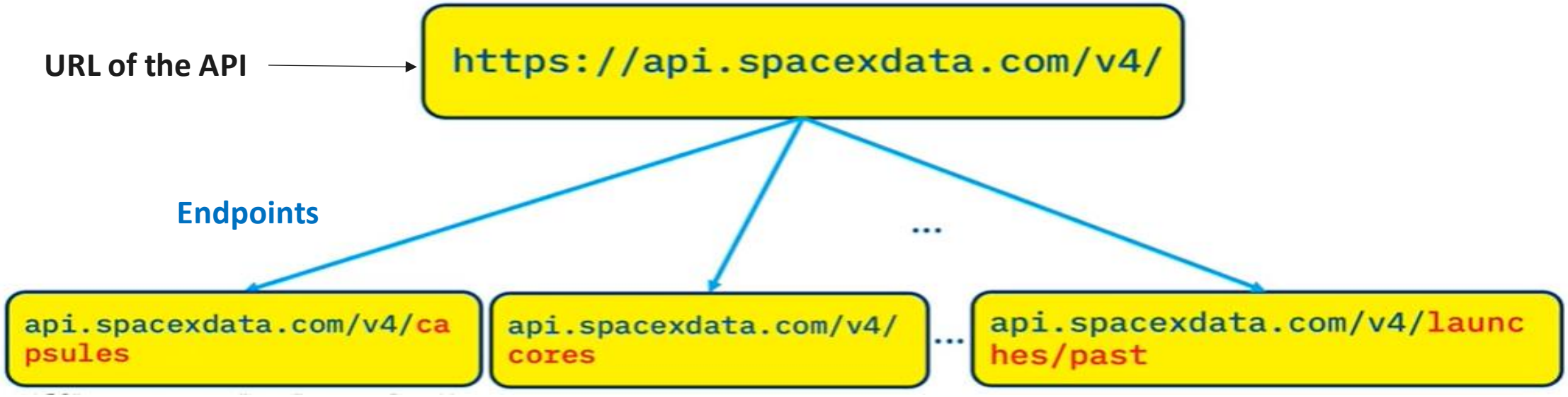
To get the historical launch records

Data collection – SpaceX API

To get the data we used a Get request to the Space X API

We used the api.spacexdata.com/v4/launches/past endpoint to get past launch data

Flowchart of the API



GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Data%20collection%20-%20SpaceX%20API.ipynb

Data collection – Web scraping

- Webscraping from Wikipedia :
- We used the page :

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

to collect the Falcon 9 historical launch records.

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Webscraping%20-%20SpaceX.ipynb

Data wrangling

1. Checking the missing values of the data
2. We identify the data types of each variable to see if they are correct
3. We create a variable class that represents the landing outcome :

Class = 1 if the landing is successful

Class = 0 if the landing is a failure

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Data%20wrangling%20-%20SpaceX.ipynb

EDA with data visualization

- We used Scatter chart to see the relationships between different variables and the landing outcome:

Flight Number and Launch Site

Payload and Launch Site

Flight Number and Orbit type

Payload and Orbit Type

We used a bar chart to see the success rate of each orbit and determine if this has an impact on the result

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Data%20Visualization%20-%20SpaceX.ipynb

EDA with SQL

Queries :

- Get the names of unique launch sites
- Getting all the launch sites that starts with 'CCA'
- Display total or average payload mass from different customers and for different booster version
- Date of the first successful outcome
- Names of the boosters which have success in drone ship for a certain payload
- Number of successful and failure mission outcome
- List the names of the booster versions which have carried the maximum payload mass
- List the the booster versions and launch sites with failed landing outcomes in drone ship
- Count the the different landing outcomes between 04/06/2010 and 20/03/2017

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Exploratory%20Data%20Analysis%20-%20SpaceX.ipynb

Build an interactive map with Folium

- Creation of a map with the different launch sites, their names, their success and failed launches, and the distance from railways, highways, coastline and cities with a line.

Names, success/failed launches were added to see easily the best success rate site

The distances were added to see if all the sites share the same characteristics, far from cities, close to coastline.

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Dashboard%20-%20SpaceX.ipynb

Build a Dashboard with Plotly Dash

- Creation of a dropdown for the sites and a range slider for the payload.

- A pie chart which returns the number of successful outcomes for each site if no sites are selected in the dropdown or return the success and failure landing outcome for a particular site if a site is chosen.

It helps us see the success rate of each site and also which site has the most successful outcomes.

- A Scatter chart between the Payload and the landing outcome for the sites selected in the dropdown site for each booster version category.

It helps us the relationship between Payload Mass, Bosster Version Category and the landing outcome.

Predictive analysis (Classification)

Steps for the classification :

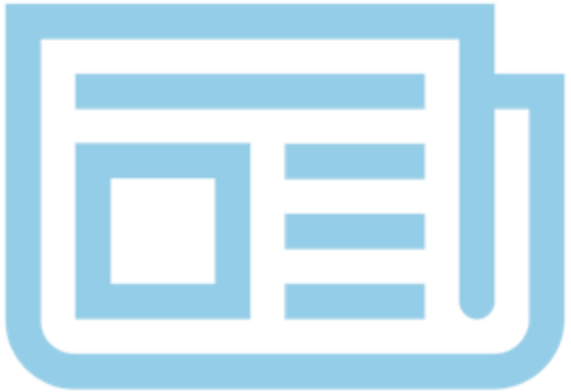
1. Standardize the data
2. Split the data between train and test data
3. To find the best parameters for each model we use the GridSearchCV method
4. We fit the model with the best parameters
5. We test the accuracy of the model on the data test and we look at the confusion matrix

We do the step 3 to 5 for the four different models :

- Logistic Regression
- SVM
- Decision tree
- KNN

GitHub URL : https://github.com/ArnaudVIGNERON/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction%20-%20SpaceX.ipynb

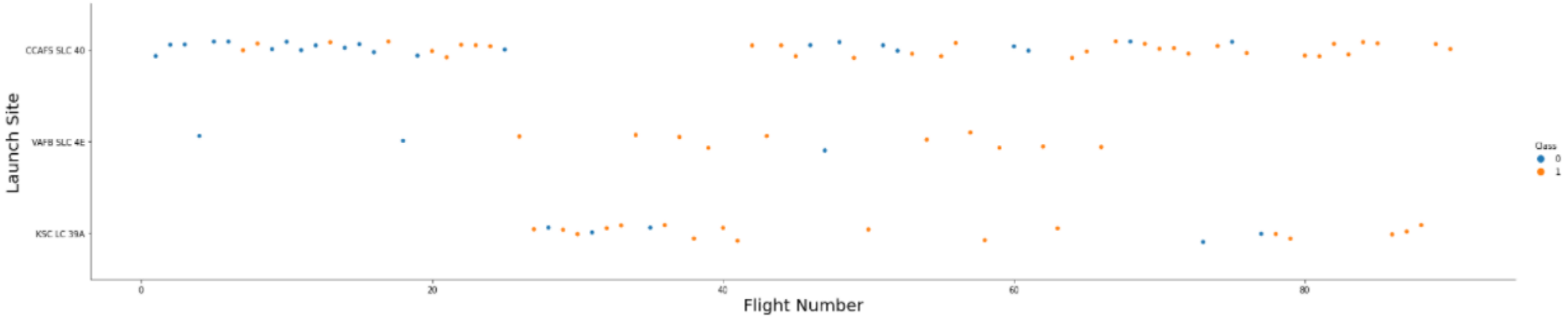
Results



- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

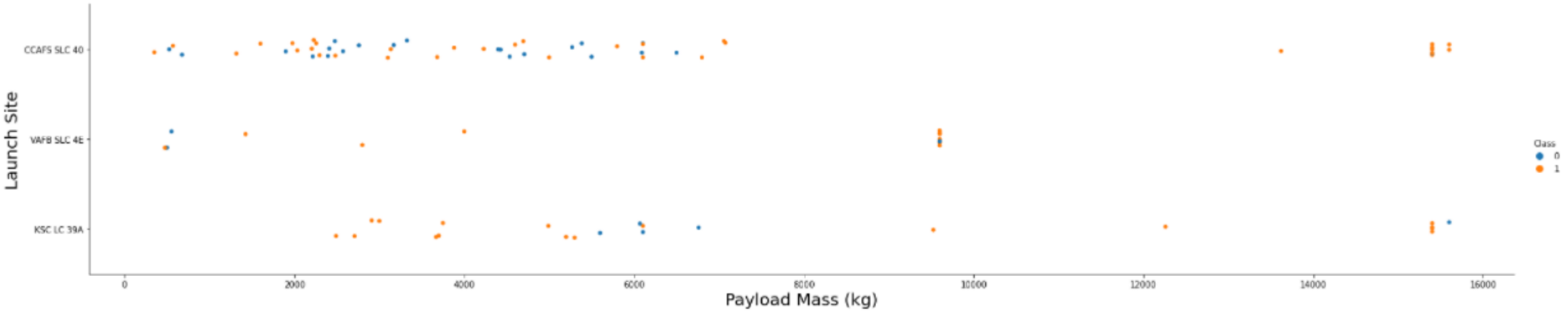
Flight Number vs. Launch Site



For the Launch Site CCAFS SLC 40 as the flight number increase, the success rate increase aswell.

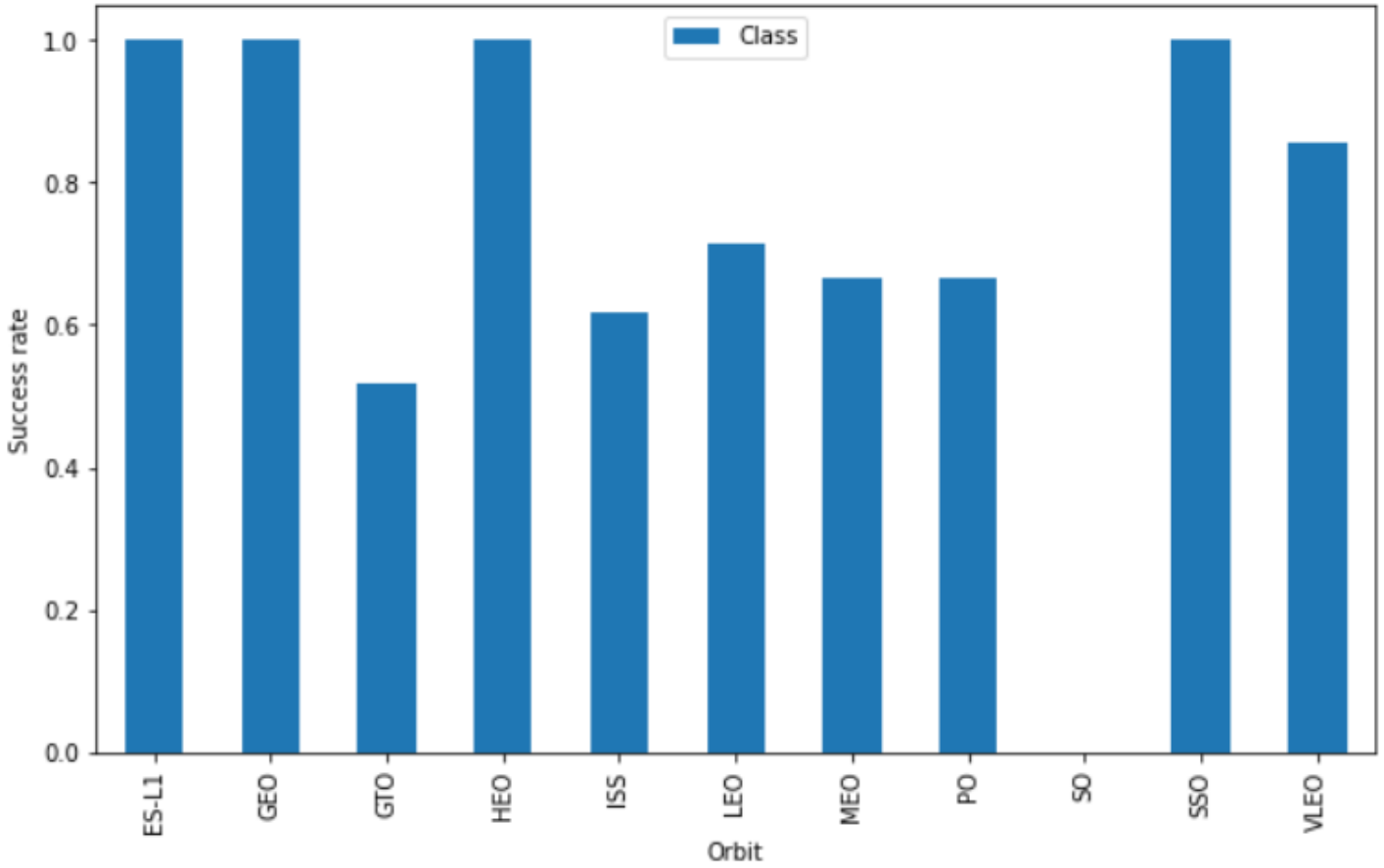
It seems to be the same for VAFB SLC 4E but we don't have enough data And it's not as clear for KSC LC 39A.

Payload vs. Launch Site



When the payload mass is over 10 000 we have a better success rate.

Success rate vs. Orbit type



We have 4 orbits with a perfect success rate :

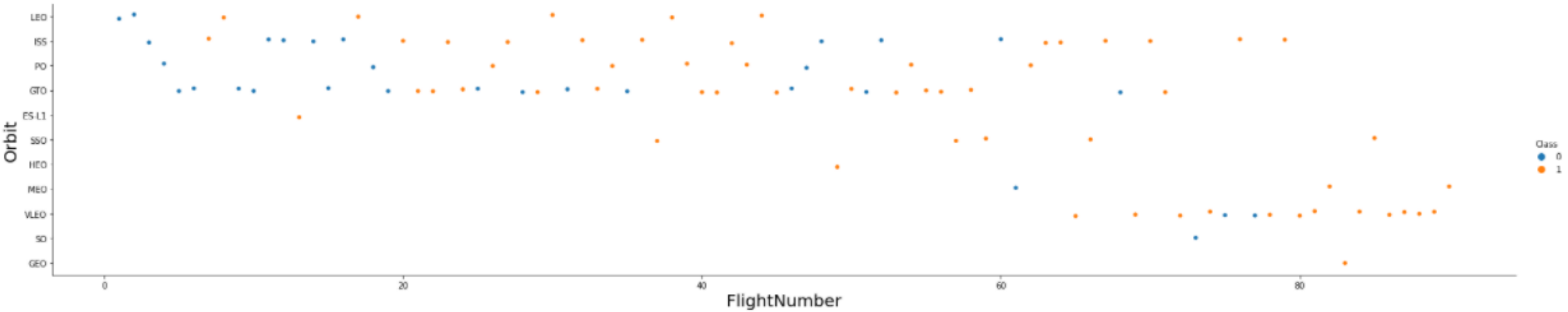
ES-L1, GEO, HEO, SSO

1 orbit with a 0% succes rate :

SO

We can assume that the orbit has an impact on the succes rate of the landing

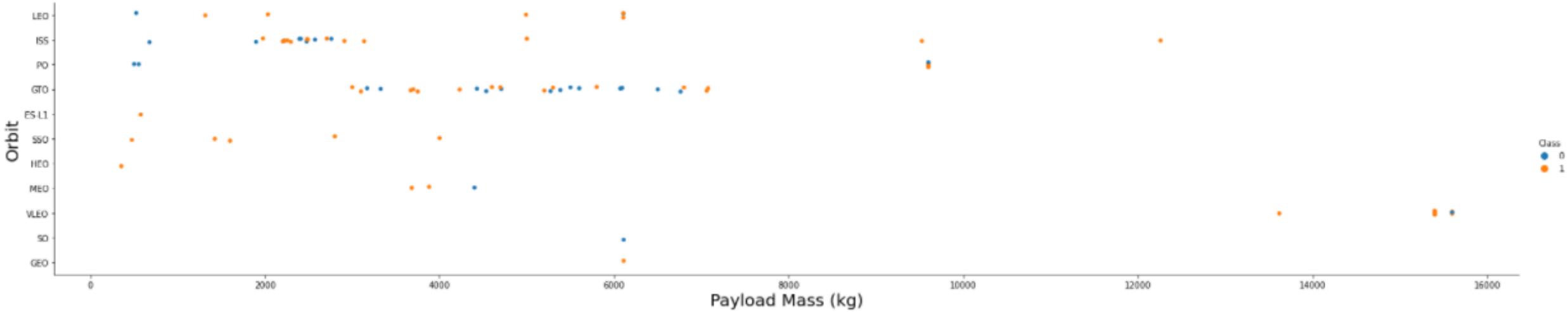
Flight Number vs. Orbit type



The success in the LEO orbit appears to be related to the number of flights

Ther seems to be no relationship between the other orbits and the number of flights

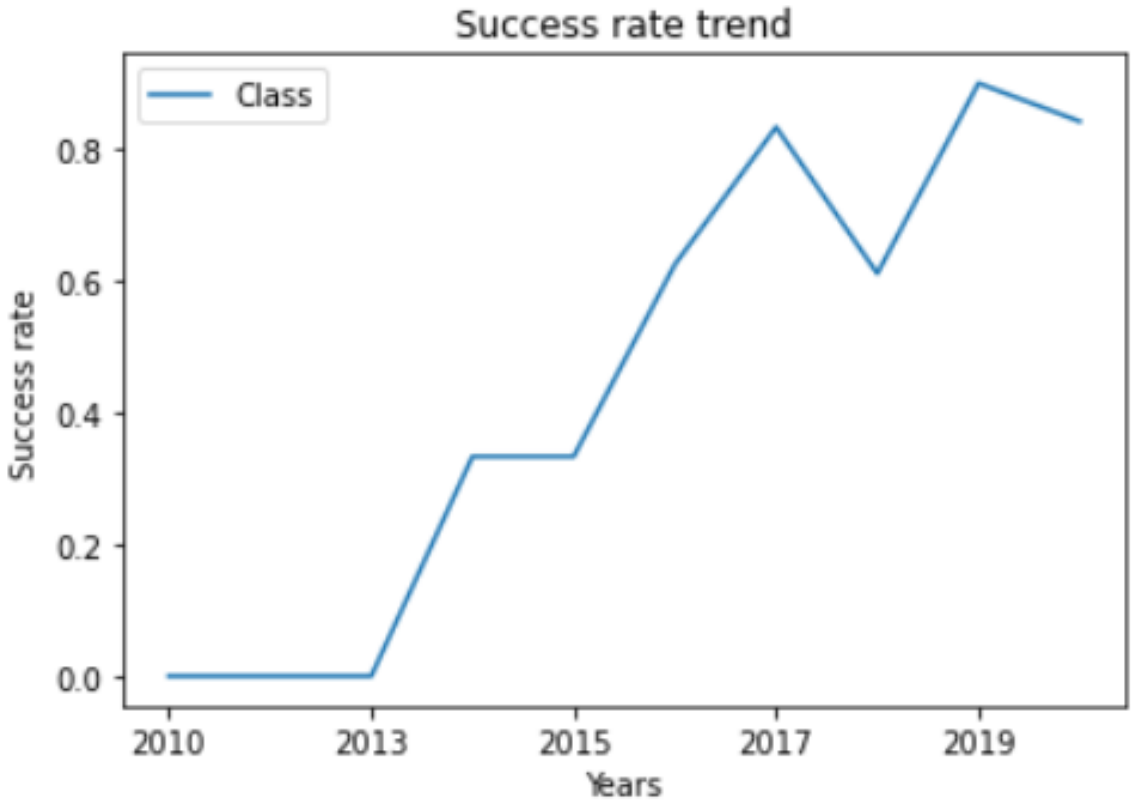
Payload vs. Orbit type



Heavy payloads have negative influence on GTO orbit.

Heavy payloads have a positive influence on the LEO and ISS orbits.

Launch success yearly trend



The success rate increase since 2013 till 2020

EDA with SQL

All launch site names

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

Those are the unique launch sites in the space mission

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

I display 5 records where launch site begin with 'CCA'.
Since the launch site begin with 'CCA' it can only be either 'CCAFS LC-40' or 'CCAFS SLC-40'.

Total payload mass

pmass
45596

The total payload mass carried by boosters launched by NASA (CRS) is 45 596 kg

Average payload mass by F9 v1.1

booster_version	pmass
F9 v1.1	2928
F9 v1.1 B1003	500
F9 v1.1 B1010	2216
F9 v1.1 B1011	4428
F9 v1.1 B1012	2395
F9 v1.1 B1013	570
F9 v1.1 B1014	4159
F9 v1.1 B1015	1898
F9 v1.1 B1016	4707
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

The booster versions F9 v1.1 B1003, F9 v1.1 B1013, F9 v1.1 B1017 have very light average payload mass

The booster versions F9 v1.1 B1011, F9 v1.1 B1014, F9 v1.1 B1016 have very heavy average payload mass

First successful ground landing date

DATE
2010-06-04

In 04/06/2010 there was the first successful ground landing

Successful drone ship landing with payload between 4000 and 6000

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

Those 4 booster versions have success in drone ship and have payload mass included between 4000 kg and 6000 kg

Total number of successful and failure mission outcomes

mission_outcome	nb
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Warning : Mission outcome and landing outcome aren't the same !

There was only 1 Failure in flight and 1 Success with payload status unclear.

All the other missions outcome were successful

Boosters carried maximum payload

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

All those booster versions have carried at least once the maximum payload

2015 launch records

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Those 2 booster versions had at least one failed landing outcome in drone ship in 2015

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

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

The count of landing outcomes sorted by nb in descending order between 04/06/2010 and 20/03/2017

There is the same number of failure and success in drone ship

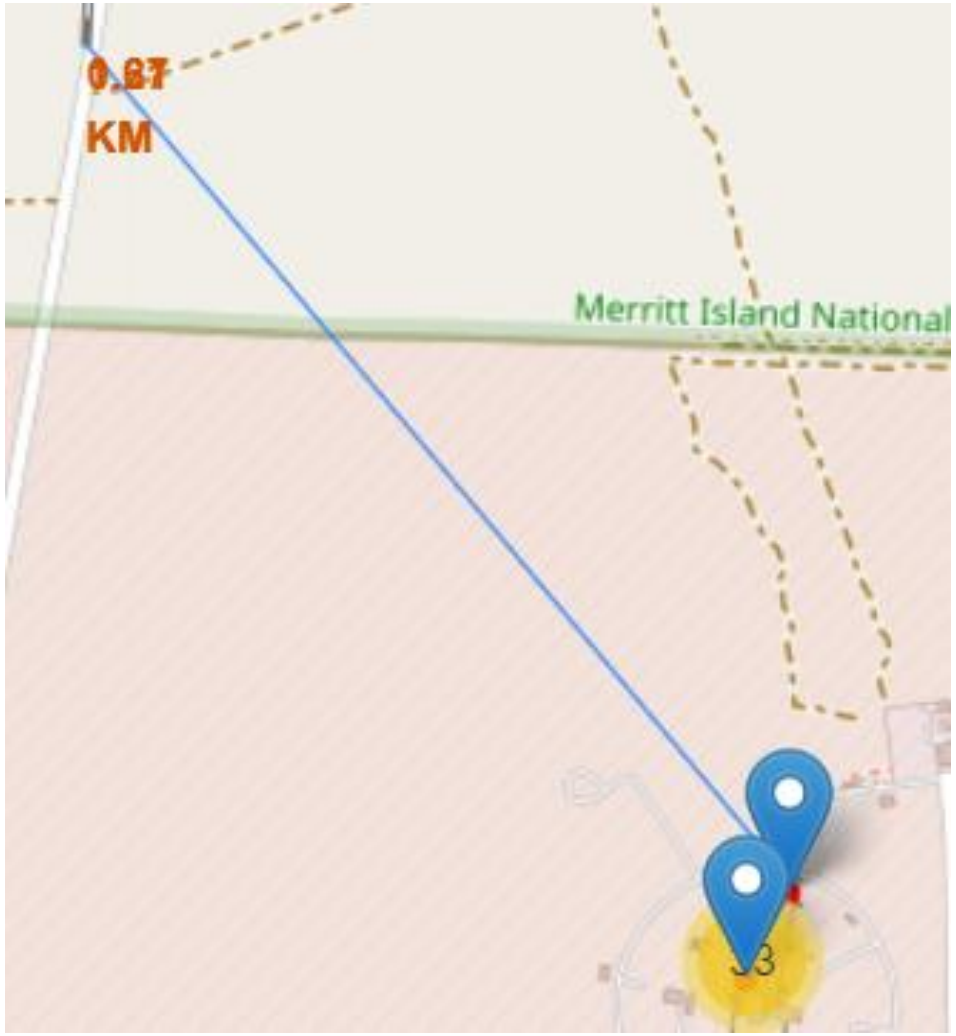
Interactive map with Folium

All the launch sites on a map

The launch sites are really close to the coast

Three of them are almost in the same place on the east coast of Florida

Distance between a launch site and a railway

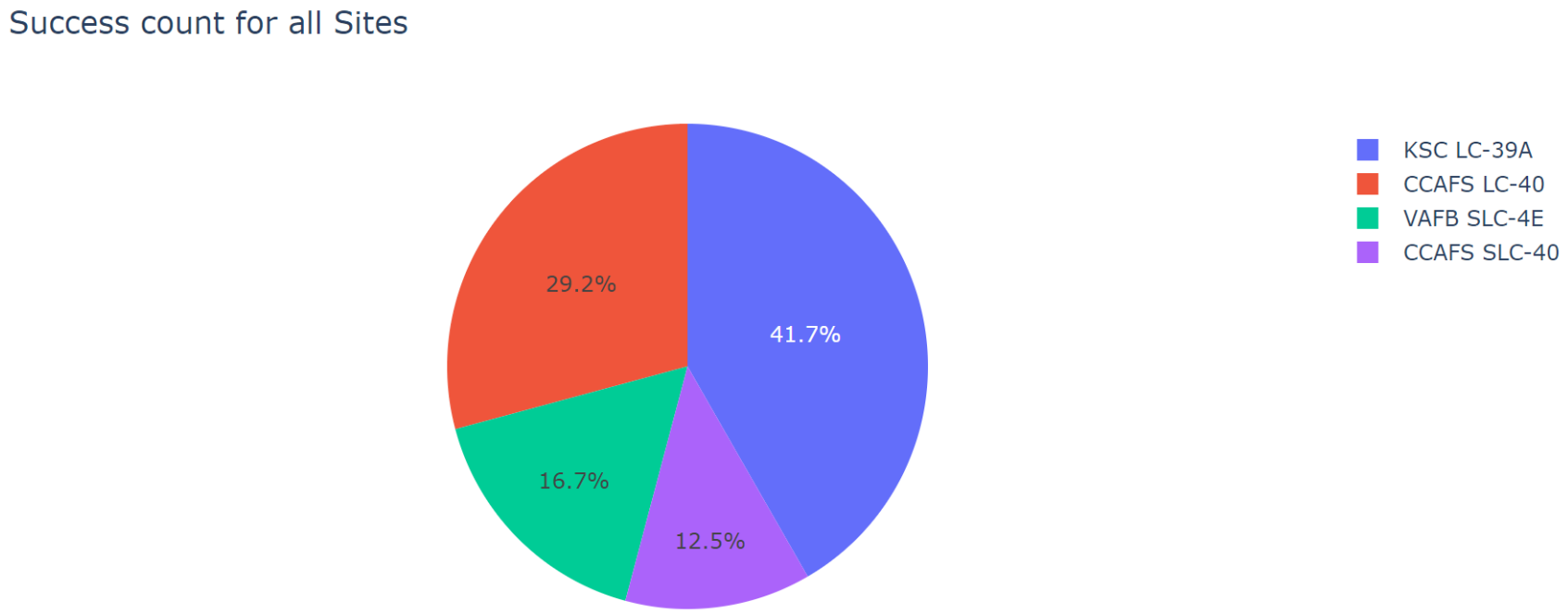


The launch sites in Florida are close to a railway and the coast

They are also far from cities and we can imagine that those characteristics are needed for a launch site

Build a Dashboard with Plotly Dash

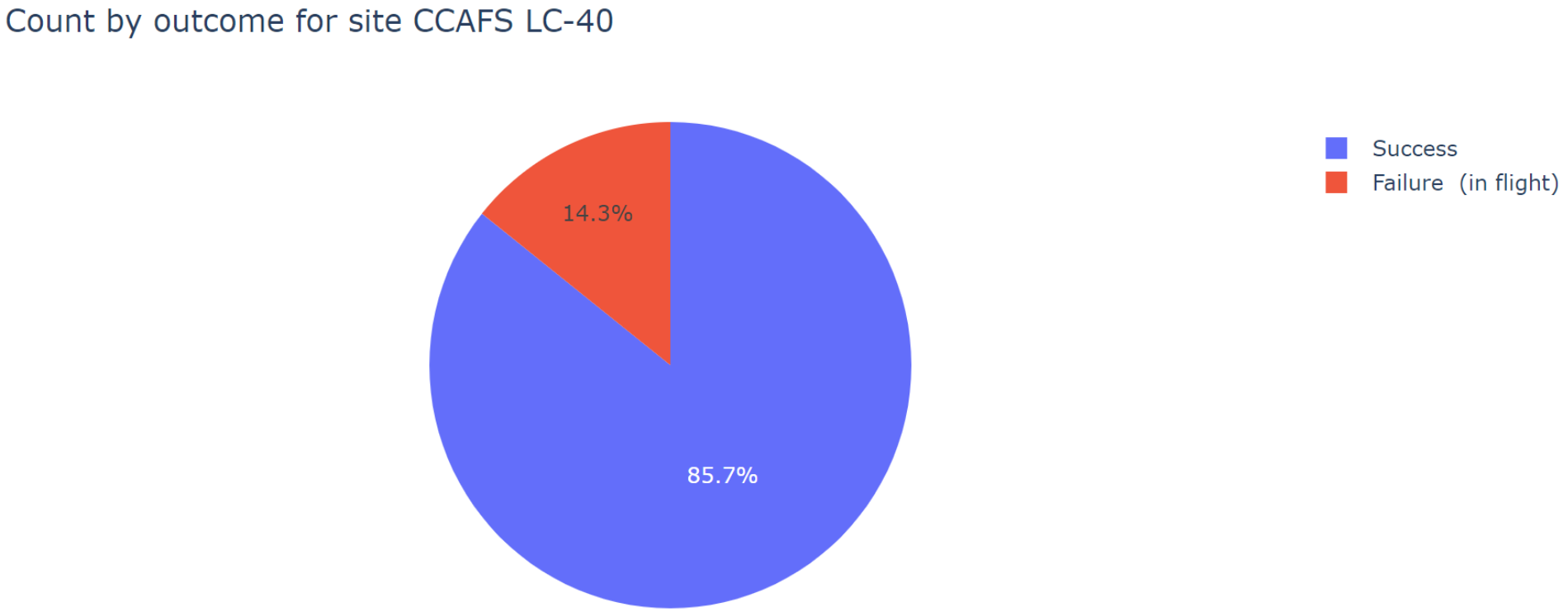
Proportion of successful landing outcomes by site



The KSC LC-39A has the most successful landing outcomes

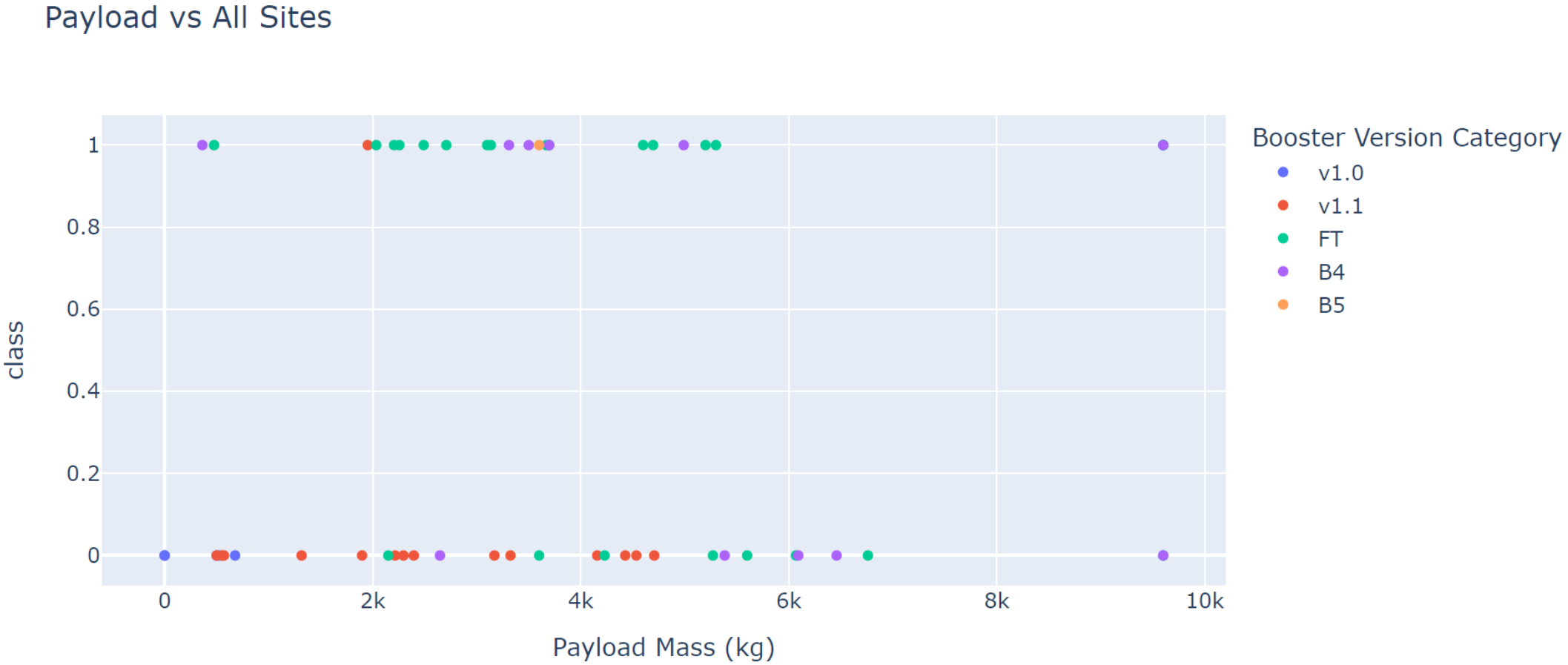
The CCAFS SLC-40 has the least successful landing outcomes

Count by mission outcome for site CCAFS LC-40



The proportion of successful mission outcome for CCAFS LC-40 is 85.7%

Relationship between Payload mass, Booster Version and outcome



There is no relationships between Payload Mass, Booster Version Category and the landing outcome for all sites

Predictive analysis (Classification)

Classification Accuracy and performance

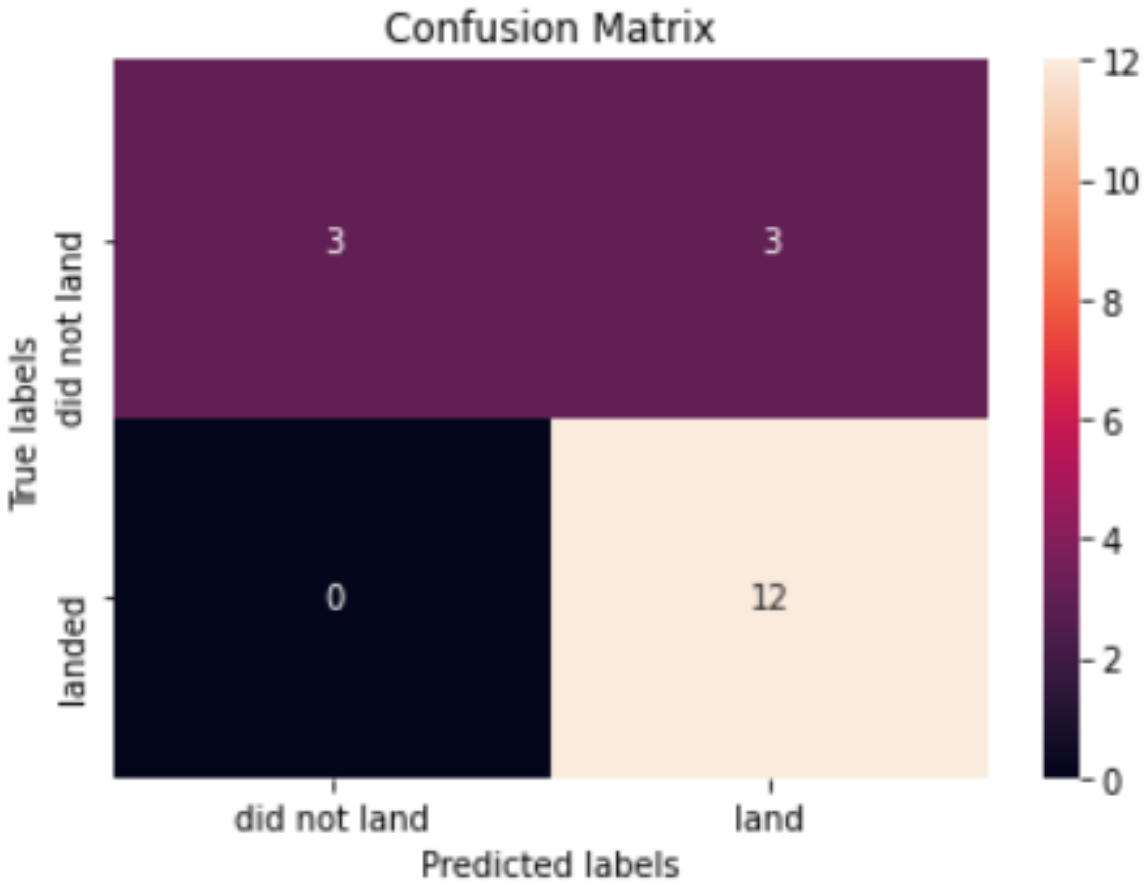
Performance of each model

Algorithm	Jaccard	F1-score	LogLoss
LogisticRegression	0.80	0.89	0.48
SVM	0.80	0.89	NA
Decision Tree	0.80	0.89	NA
KNN	0.80	0.89	NA

All the models have the same accuracy which is 83.3%

The models also performs the same way according to the Jaccard Index and F1 Score

Confusion Matrix



There is 3 False positives which means that 3 times he did not land but we predict that it lands.

But everytime it lands, we always predict that it will land.

The issue is that we are going to surevaluate the number of successful landing and underevaluate the number of failure landing.

CONCLUSION

- Success rate of the landing outcome increase since 2013 and is around 85%
- Success rate of the landing outcome have a correlation with the Launch Site, the number of flights, the payload and the orbit
- The prediction of the Falcon 9 first stage landing successfully or not has a 83% accuracy and will surevaluate the number of successful landing

- From the trend we can guess that the success rate will get even higher
- To compete with Space X, we are going to need to lower the cost to the same cost as the Falcon 9 when the landing of the first stage is a success

THANK YOU