IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<MagB>
<May 2023>

Outline

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

Executive Summary

- Summary of methodologies
    - Data collection using API
    - Data collection with web scraping
    - Data wrangling
    - Exploratory data Analysis with SQL
    - Exploratory data Analysis with Data Visualization
    - Interactive Visual Analytics with Folium
    - Machine learning prediction
- Summary of all results
    - EDA results
    - Interactive and predictive analysis results

Introduction

- Project background and context
  - A new rocket company is trying to determine what are the success factors for SpaceX launch and if SpaceX will reuse the first stage.
  - In order to study these points, we will use and train a machine learning model, with public informations.
  - The outcome of this study can be further use to determine the cost of a launch for alternate company.

- Problems
  - Which features/conditions contribute to the success of the launching (payload, orbit, launch site…)
  - What is the trend of the success rate over time?
  - What is the most reliable machine learning model to determine the success of a launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection
  - 2 sources for data :
    - SpaceX API : (https://api.spacexdata.com/v4/rockets/)
    - Webscraping from Wikipedia : (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Data wrangling
  - Data cleaning before analysis (replace null value by the mean)
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Finding out the best machine learning model with the best accuracy

Data Collection – Space X API

Stage 1

1. Input the url of the data to analyse
2. Use the GET request to get and parse the data for SpaceX launch
3. Convert the response into a JSON file, then in Panda dataframe

Stage 2

1. Export information about the launches with API
2. Filter only on Falcon 9 input
3. Excecute Data cleaning and replace null value

Git hub link :
https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20Jupyter%201%20Data%20collection.ipynb

Data Collection – Web Scraping

Request the Facon 9 wiki page

Create a Beautiful Soup object

Extract all columns names from the table HTML

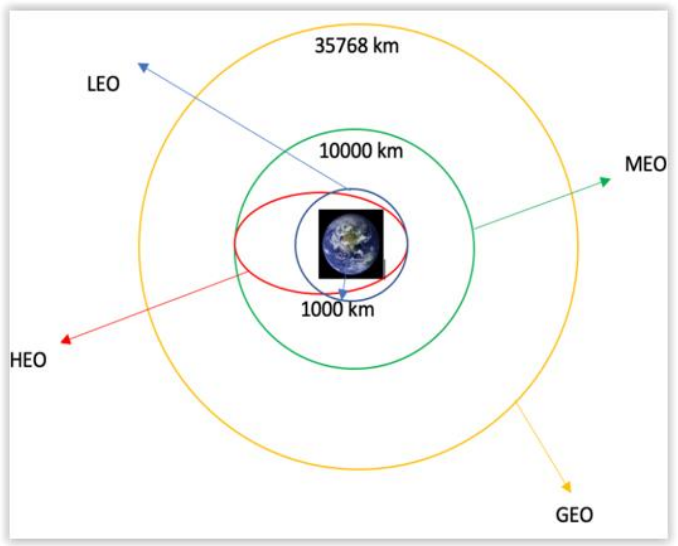
Create a empty dictionnay with the columns names

Fill the dictionnary with launch input

Turn the dictionnary into csv dataset

Git hub link: https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%202%20Web%20scraping.ipynb

Data Wrangling



Calculate de number of launches for each site → Calculate the number of occurrences for each orbit → Calculate the number of occurrences for mission outcome/orbit type → Create a landing outcome label → Calculate the workout succès rate for every launching in dataset

Git hub link:
https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%203.%20Data%20wrangling.ipynb

EDA with Data Visualization

- The charts used are :

  - **Bar chart**: Flight number vs Payload mass/Launch sites/Orbit type
    Payload mass vs Orbit type/Launch sites
  - **Scatter plot** : success rate per orbit
  - **Line plot** : success rate and year

Git hub link :
https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%205%20EDA%20Visualisation.ipynb

EDA with SQL

Summary of SQL query used :

- Names of unique launch sites
- Display 5 records where launch site begins with “CCA
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster F9 v1.1
- Dates for first successful landing in ground pad
- Names of the boosters with success in drone ship and payload mass between 4 and 6 Kg
- Total number of successful and failure mission outcomes
- Name of the booster version which have carried the maximum payload mass (with subquery)
- List of failed landing outcomes (drone ship type) or success (group pad key) between 2010-06-04 and 2017-03-20 in descending order

Git hub link : https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%204%20SQL.ipynb

Build an Interactive Map with Folium

- Markers (for launch sites), lines (for distance) and markers clusters (for events such as launch) were used with Folium map to show the nearest important landmarks such as railways, highways, cities and coastlines.
- Draw red line to show rocket launch failures and green lines to show the successes.
- Then, using the Haversines formula, calculate the distance from the launch site to the various landmarks

Git hub link :
https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%205%20Visual%20Analytics%20with%20Folium.ipynb

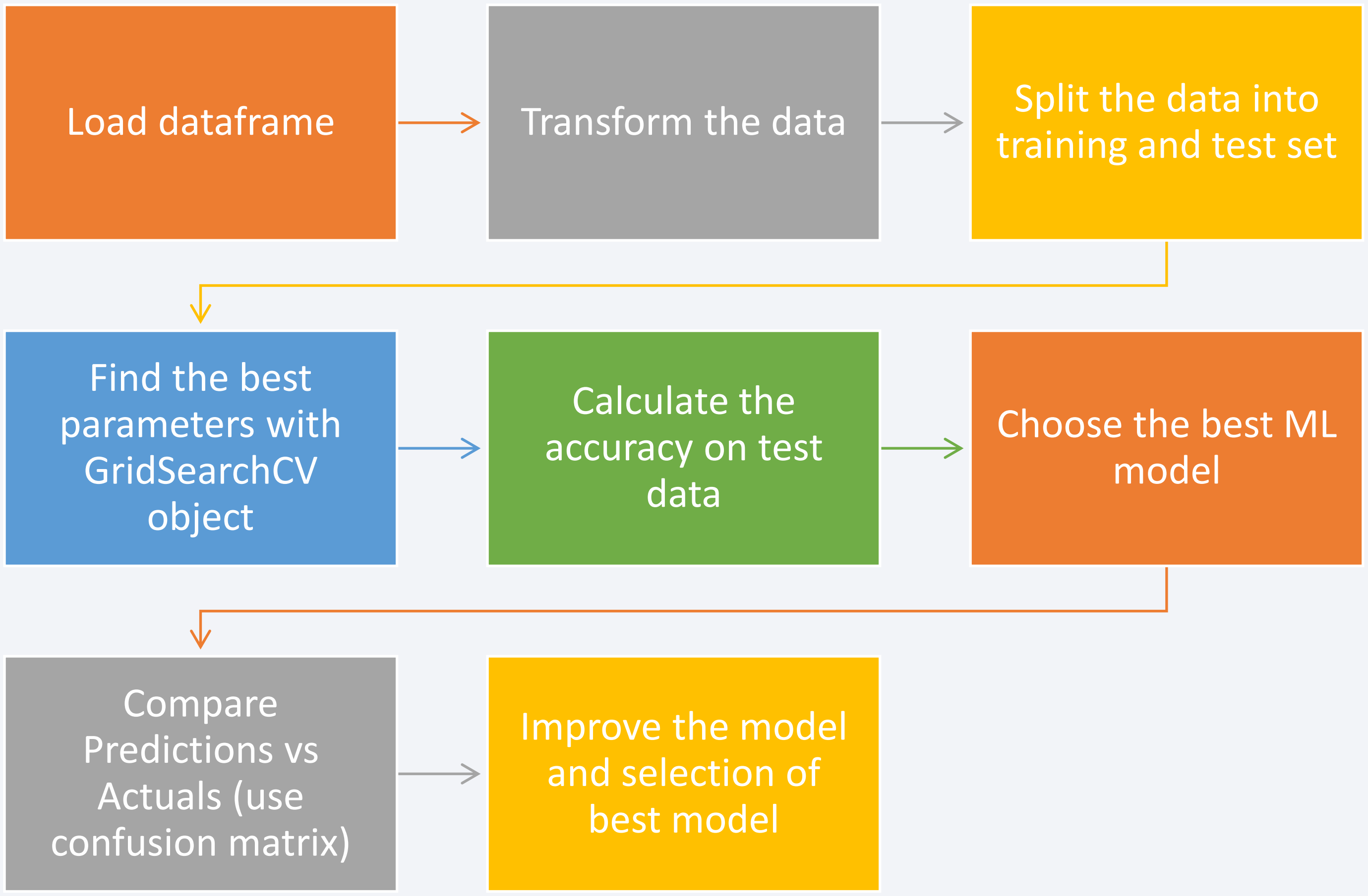
Build a Dashboard with Plotly Dash

- Launch records of SpaceX were visualized with Pie Charts and scatters Charts.
- We display the rocket launch success rate per site to understand which factors may influence the success rate on each site.
- We try to understand the correlation between factors like payload mass and launch site.
- Successful launches are represented by a “1” and failures by a “0”.

Predictive Analysis (Classification)

We use Scikit-learn for predictive analysis (evaluate the success chance of landing at first stage)

4 classifications models were compared (logistic regression, SVM, Decision tree, K nearest neighbours)



Git hub link: https://github.com/coockiem/Capstone_project_SpaceX_MB/blob/main/Notebook%20jupyter%206%20Machine%20learning.ipynb

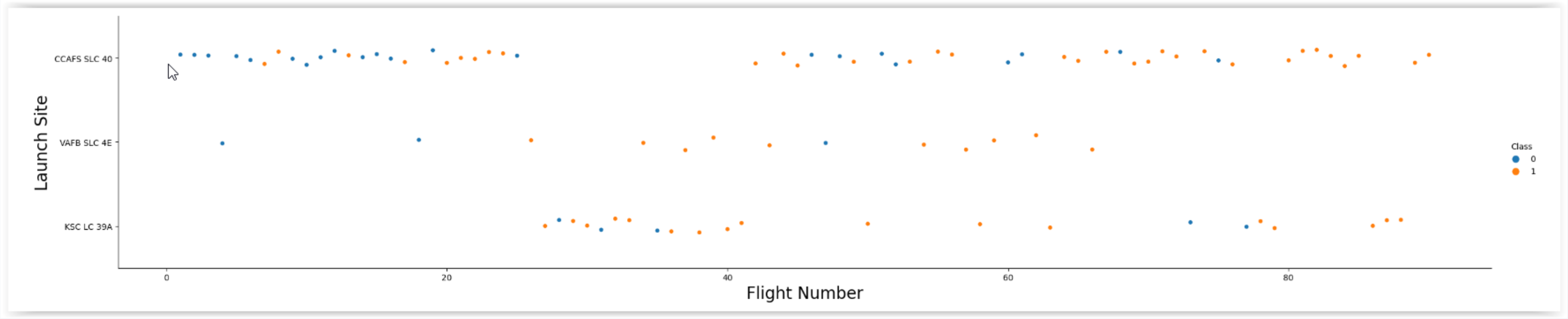
Results

- Exploratory data analysis results
  - There are 4 different launch sites with different success rate. The best sites are located near the coastline (probably for safety reason), and not too far from transportations infrastructures such as highway and railways (make easier transportations of people and materials)
  - First success appears to be in 2015
  - The average payload is 2928 kg
  - 2 failures occurred with the booster version FXXXXX in 2015
  - The rate of success increase as years passed.
- Predictive analysis results
  - The machine learning model was able to predict the landing success of rockets with an accuracy of 88%

Section 2

Insights drawn from EDA

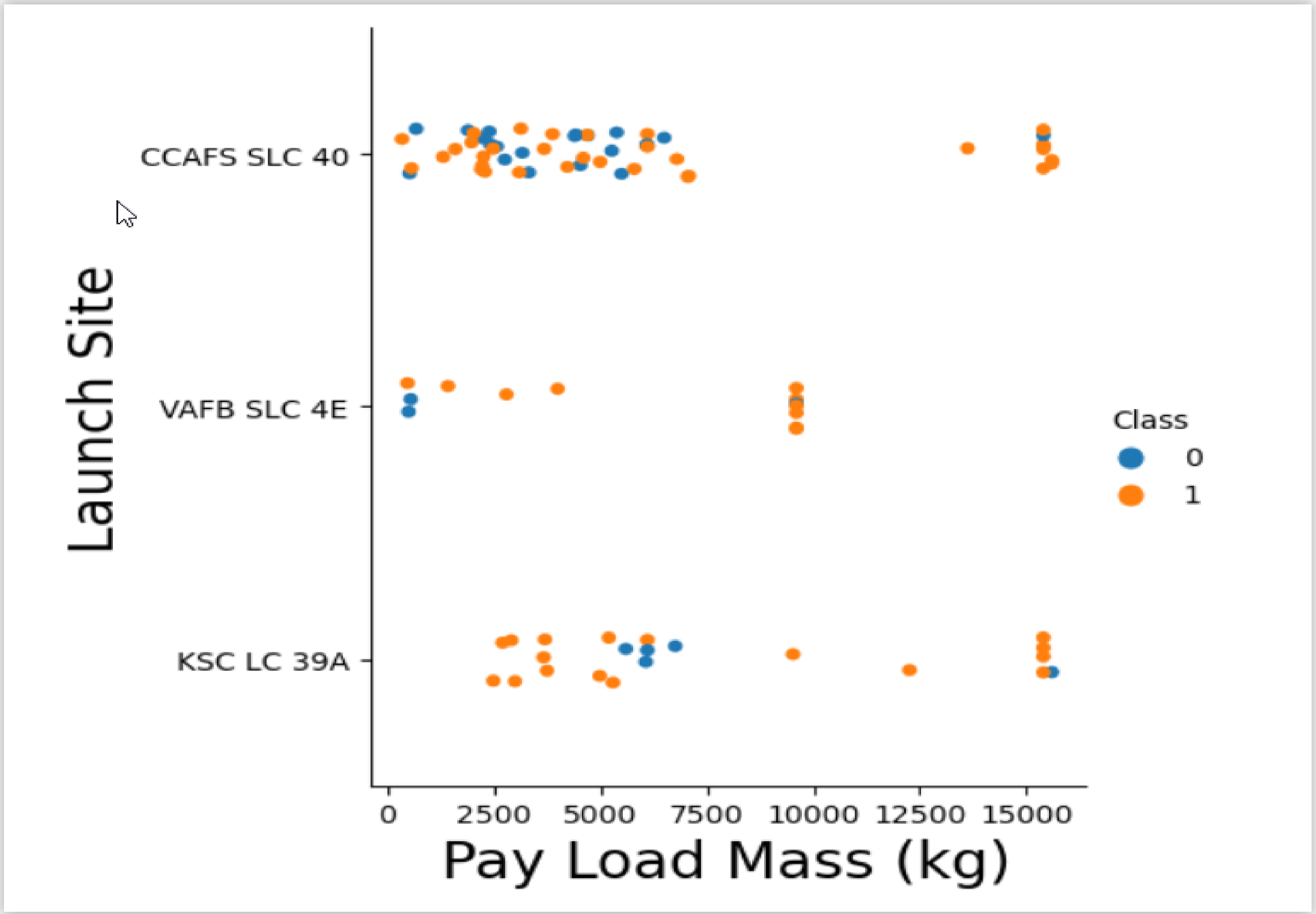
Flight Number vs. Launch Site



- It appears clearly that most successful landing took off from CCAFS sites, but also from KSC site, although the first flights were less successful.
- The success increases with the number of flights.

Payload vs. Launch Site

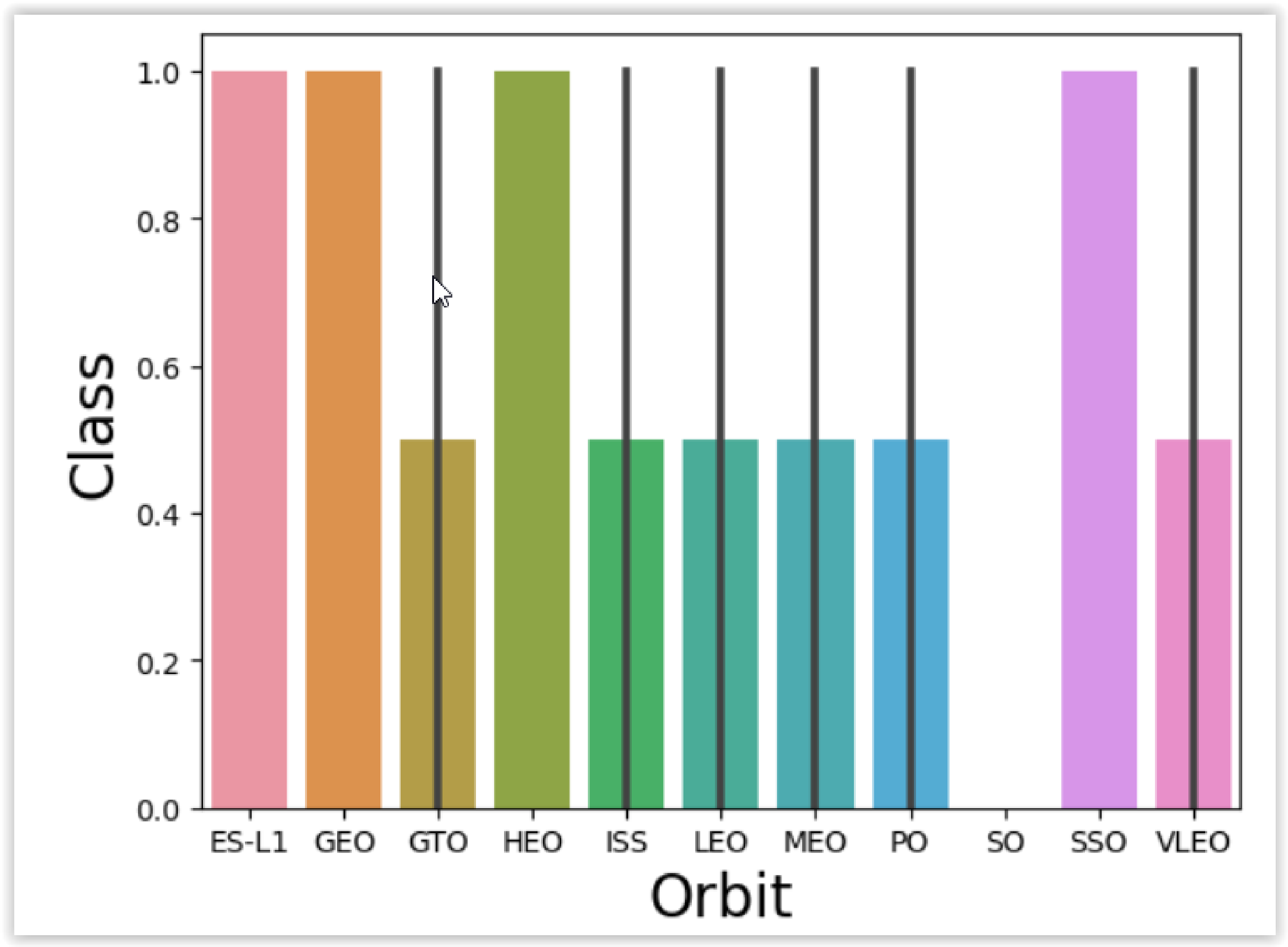
- Only 2 sites allows the comparison for heavy payload (> 10K Kg).
- It seems that success launches are higher when payload is > 10K Kg.
- However, it would be surprinsing to find that there is a clear connection between launch sites and payload.



Success Rate vs. Orbit Type

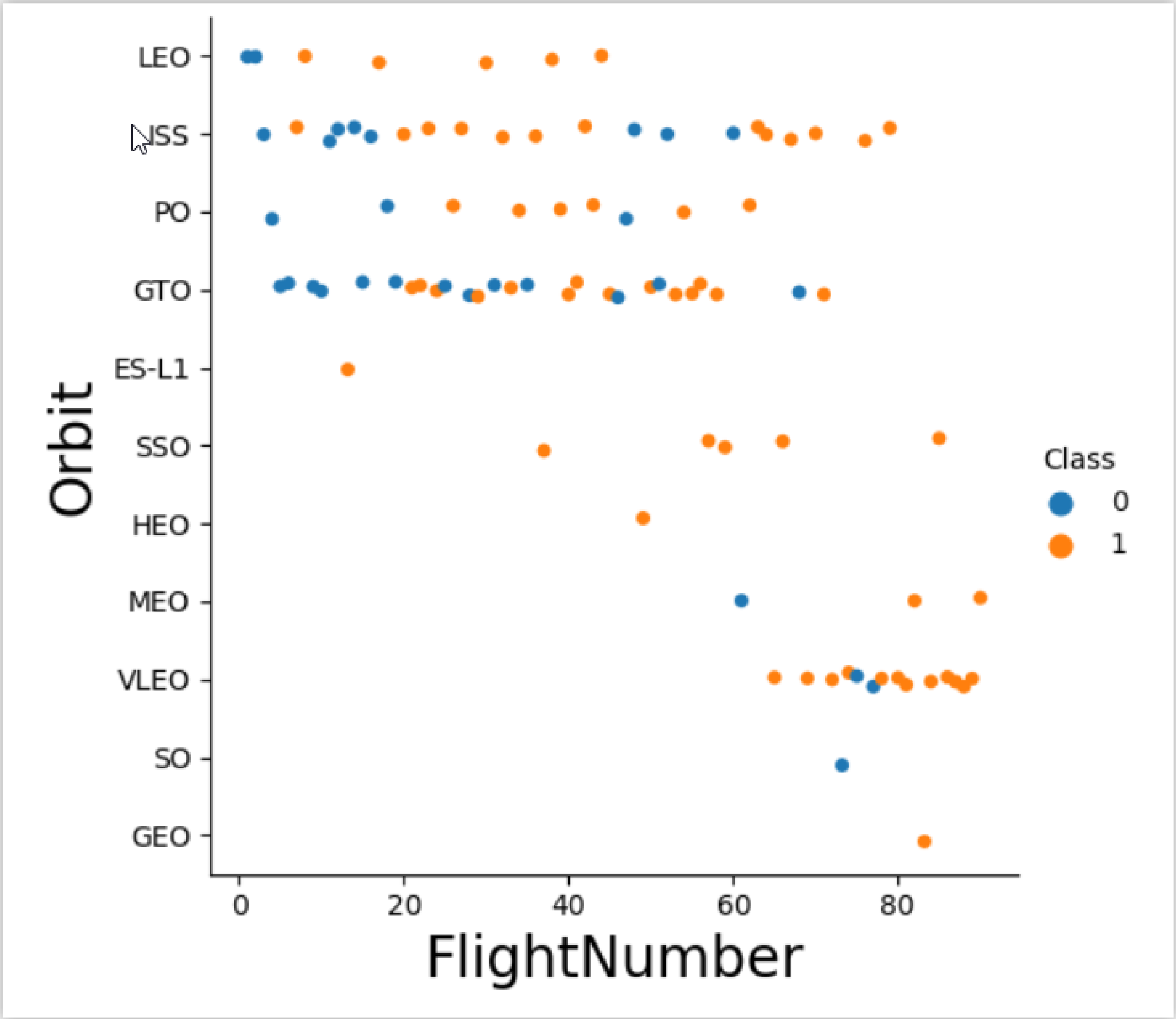
- There is a 100% success rate for ORBITS below :
- ES-L1
- GEO
- SSO
- HEO

- We can notice that SO orbit has a success rate of 0% (but with only 1 launch).



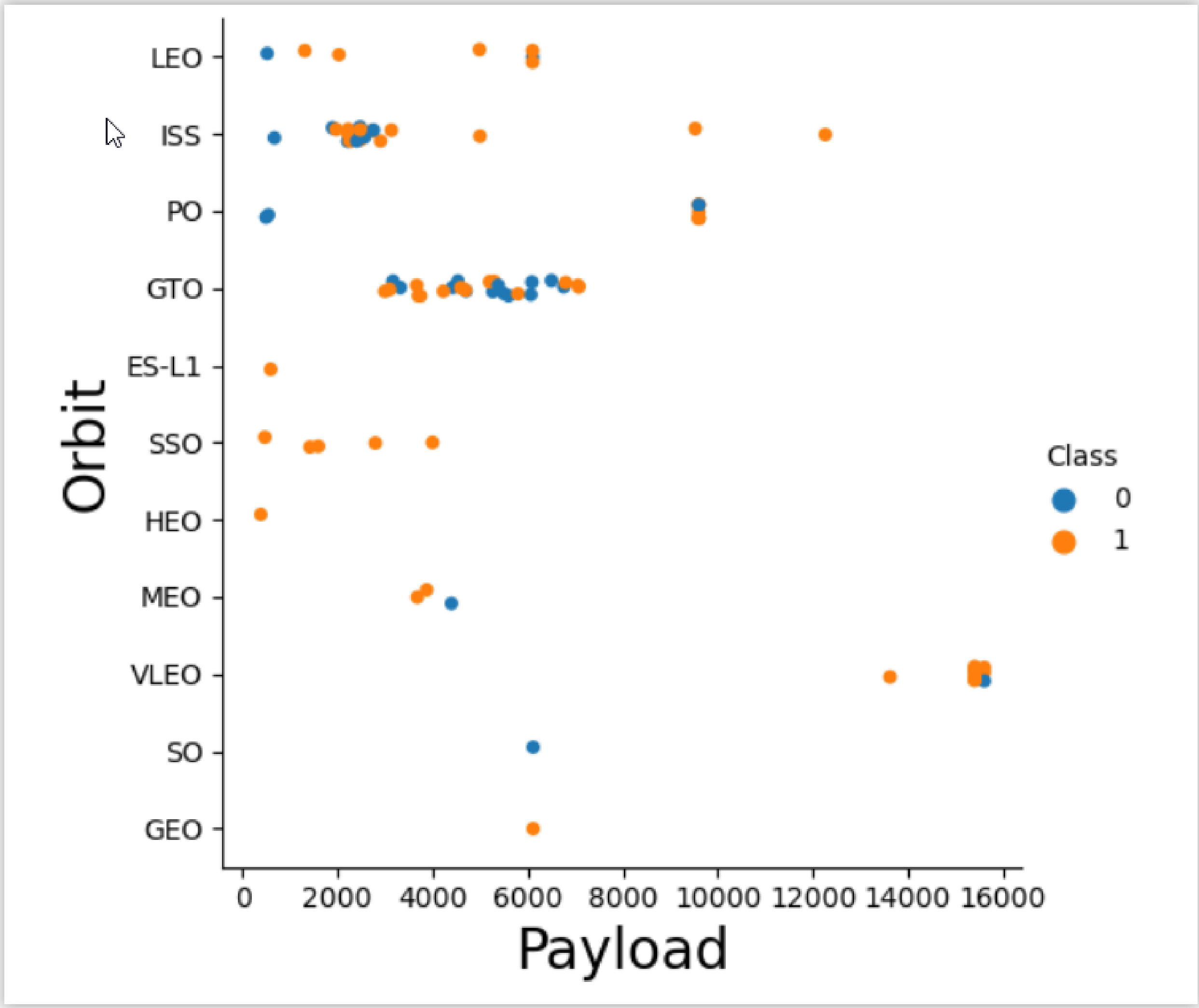
Flight Number vs. Orbit Type

- It seems that LEO and PO orbits have an increasing success rate with the number of flights.
- Some orbits have very few flights (Ex: VLEO) so it is difficult to compare with others.



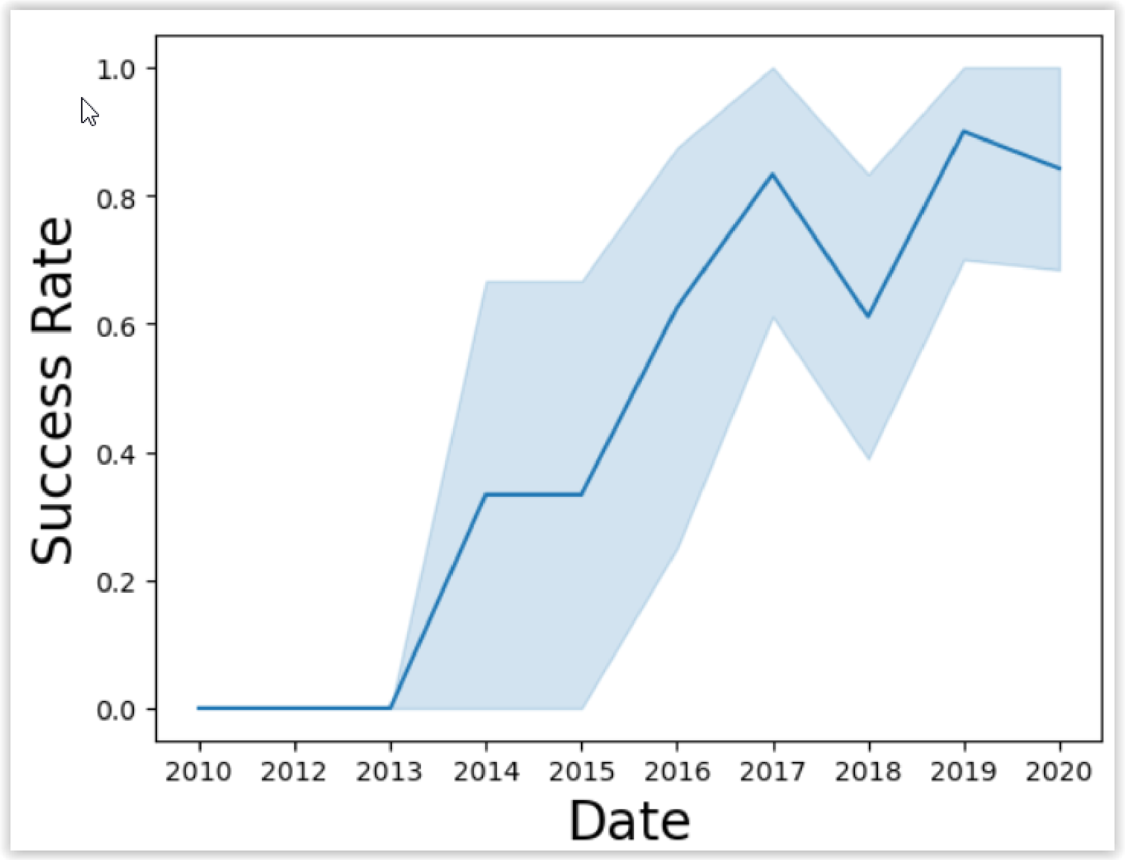
Payload vs. Orbit Type

- SSO orbit is 100% successful.
- Despite the great number of flight, it is not obvious to state if GTO orbit is successful or not.
- We can also see a clear distinction between low and high payload (> 8K Kg) : there are not a lot of flights to compare above this payload mass.



Launch Success Yearly Trend

- It is clear that the success rate increases with the year (maybe due to more experiences that make adjustments possible and evolution of technology)



All Launch Site Names

- The unique launch sites are listed below :
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- We use the SELECT DISTINCT command to retrieve the unique launch site names

Display the names of the unique launch sites in the space mission

```
%sql  SELECT DISTINCT LAUNCH_SITE FROM NDL78246.SPACEX;
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- List of 5 records where launch sites begin with `CCA` where we can see other organization than NASA.
- We use the LIMIT command to retrieve 5 records only.

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM NDL78246.SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

- The total payload mass carried by boosters from NASA as customer was 45596 kg
- We use the SUM function to calculate the total payload mass and the filter customer = ‘NASA (CRS)’ to calculate NASA only results.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass_kg_) FROM NDL78246.SPACEX WHERE customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928,4 Kg
- We use the AVG() function to retrieve the average payload mass and the filter Booster version = 'F9 V1.1' to calculate result for this booster version only.

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

```
%sql SELECT AVG(payload_mass_kg_) FROM NDL78246.SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

1
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 22 December 2015.
- We use the MIN() function to retrieve the smallest value date, with a filter on landing outcome ='Success (ground pad)''.

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(DATE) from NDL78246.SPACEX where Landing_Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are only 4 boosters with a payload mass between 4K and 6K kg.
- We use 2 conditions here to filter the records by landing outcome and payload mass.

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

```
%sql SELECT BOOSTER_VERSION from NDL78246.SPACEX WHERE landing_outcome = 'Success (drone ship)' and  payload_mass_kg_  >4000 and payload_mass_kg_ <6000;
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is 100.
- We use COUNT() function to retrieve the number of records for mission outcome = ‘Success’ or ‘Failure (in flight)’

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) from NDL78246.SPACEX where mission_outcome = 'Success' or mission_outcome = 'Failure (in flight)'
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

1
100

Boosters Carried Maximum Payload

- There are 12 boosters which have carried a maximum payload mass of 15600 Kg
- We use a sub query to retrieve the maximum payload mass associated to the MAX() function.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT BOOSTER_VERSION FROM NDL78246.SPACEX WHERE PAYLOAD_MASS_KG_ = (SELECT max(PAYLOAD_MASS_KG_) FROM NDL78246.SPACEX);
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

2015 Launch Records

- There were 2 boosters (drone ship) that failed to land in 2015, both situated in the launch site CCAFS LC-40 :
  - F9 v1.B1012
  - F9 v1.B1015
- We use the YEAR() function to retrieve the year part of the date.

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM NDL78246.SPACEX WHERE year(DATE) = '2015' AND \
LANDING_OUTCOME = 'Failure (drone ship)'
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

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

- We are ranking 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.
- We use order by date desc to sort by1st column in descending order

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

```
%sql select * from NDL78246.SPACEX where LANDING_OUTCOME = 'Success (ground pad)' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * ibm_db_sa://ndl78246:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

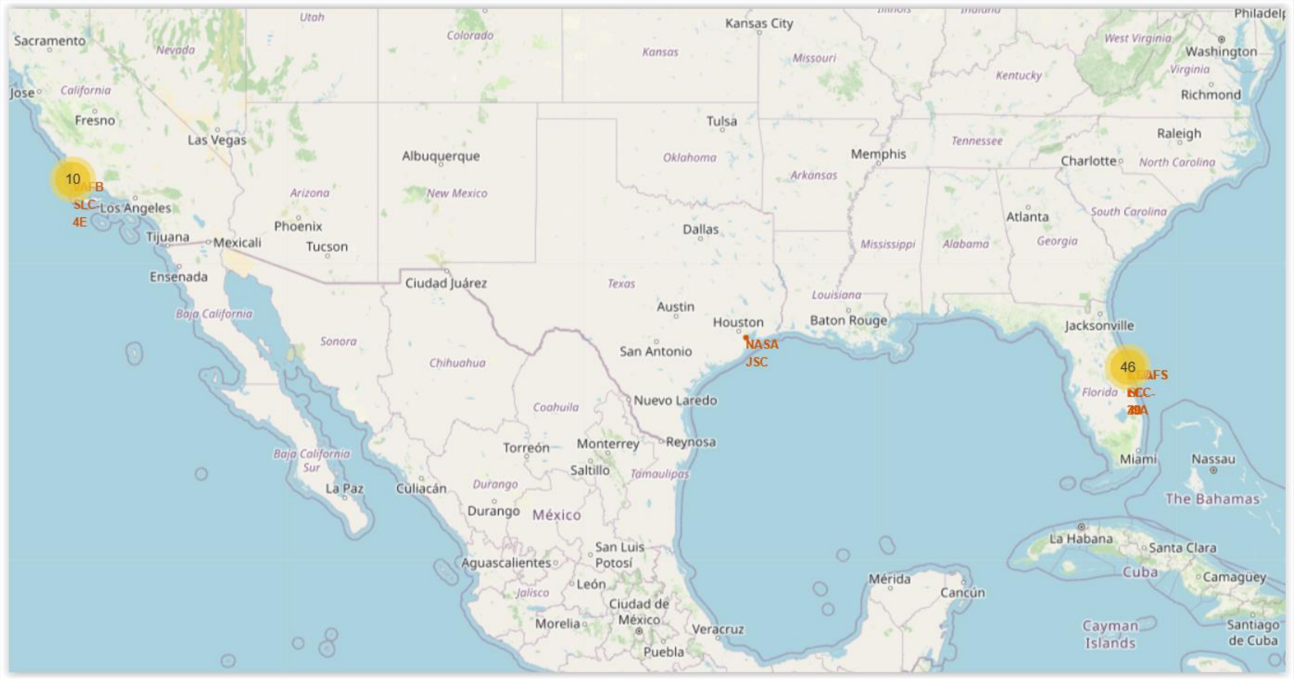
DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

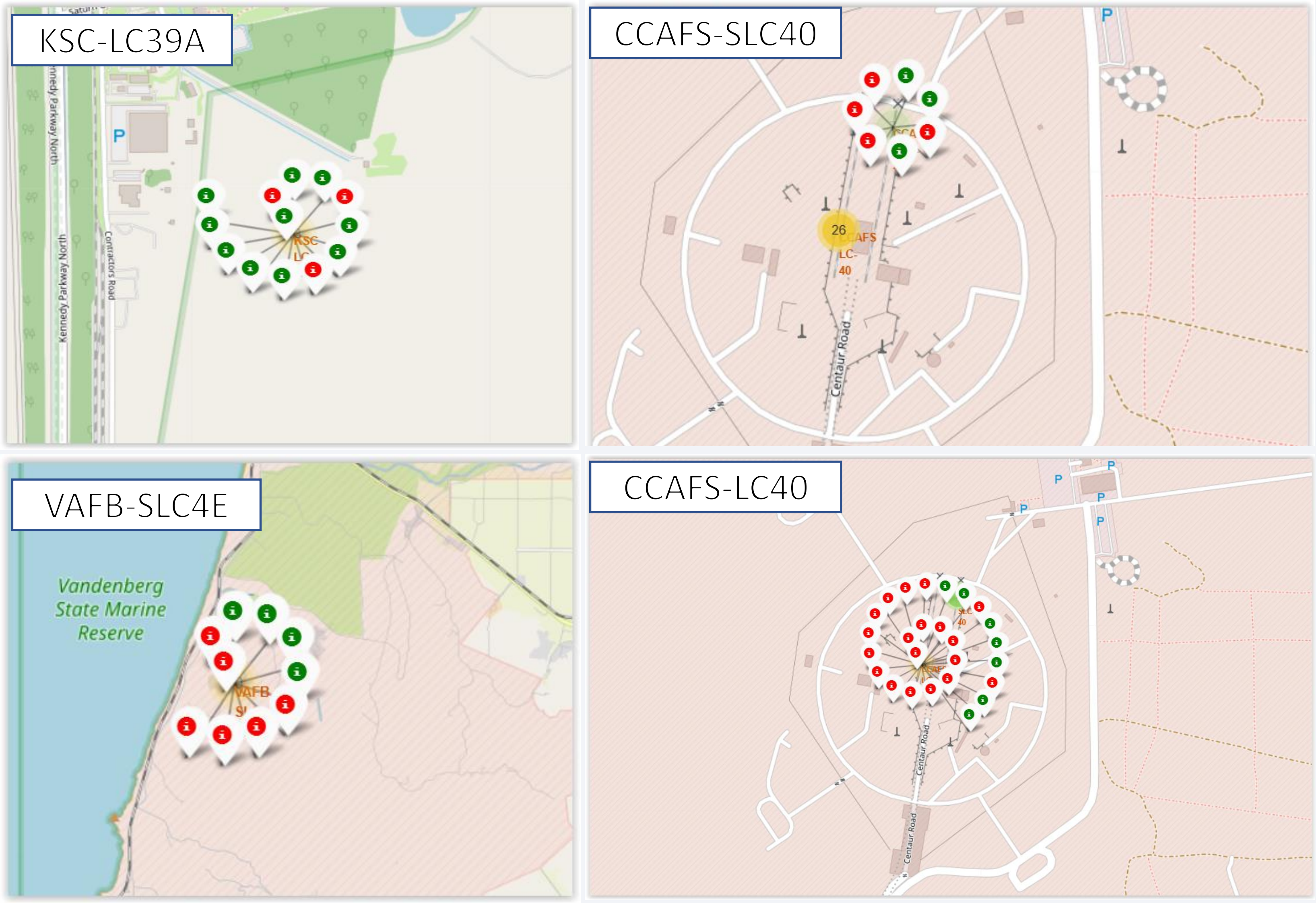
Launch sites

- The main lauchsites are opposite to each other, one in Florida (east coast) and the other in California (west coast) to be able to choose the best regarding different factors.
- The choice of launch site is very important in order to avoid crash as the one of Starphip in April 2023 and the huge amount of damage caused.



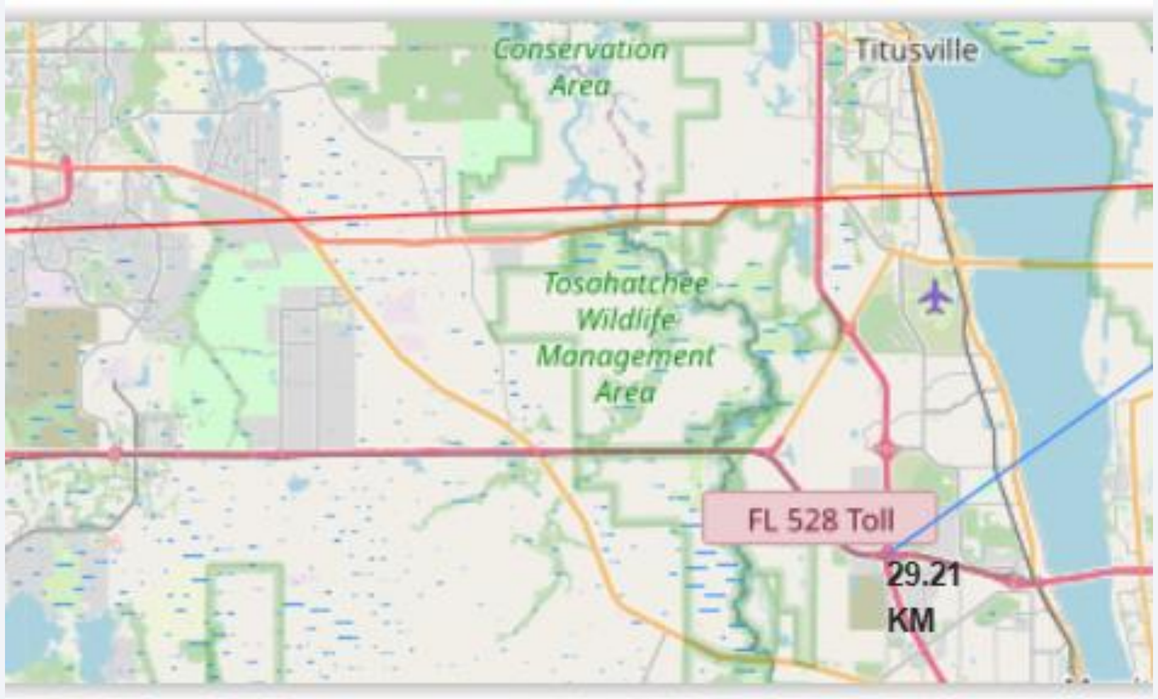
Success rate of launches per site

- Example for sites CCAFS-SLC40, CCAFS_LC40, VAFB-SLC4E & KSC-LC39A
- Green markers indicate success and red ones failures.
- We can already noticed that KSC-LC39A seems to have a good ratio (green/red) of successful launches, compared to the others.



Distance to proximities

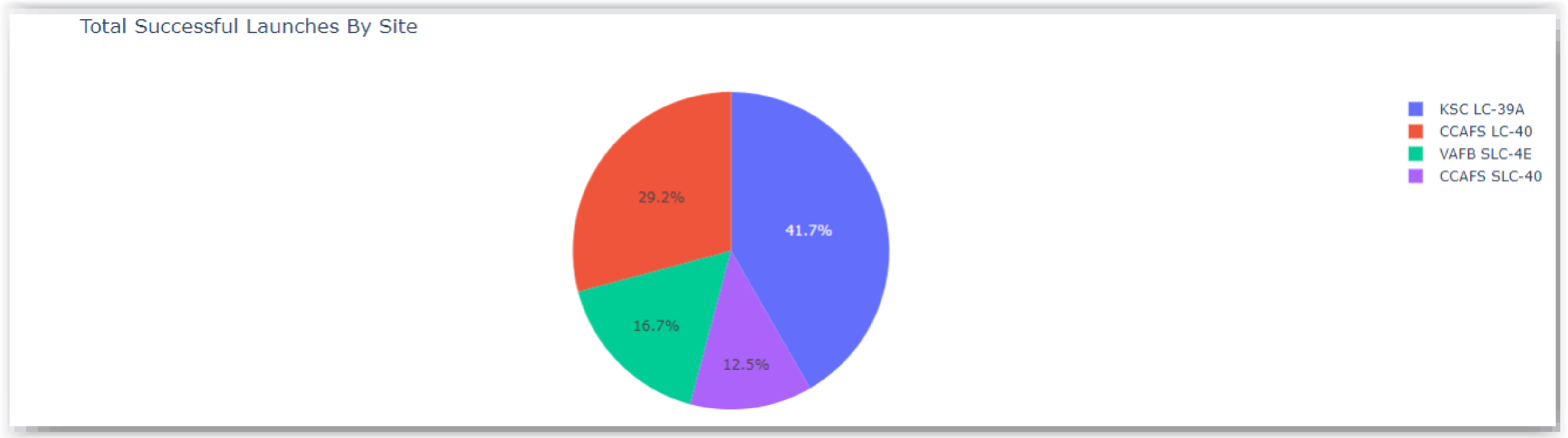
- Launch sites are located (example for launch site CCAFS-SLC40)
  - Far from cities for safety (almost 79 Km from Orlando for CCAFS site)
  - Not too far from highway (29 km) for easy transportation of people/evacuation in case of a incident
  - Close to railways (0.86 km) for easy logistic
  - Close to coastline (0.59 km) to avoid damages in case of launch failures



Section 4

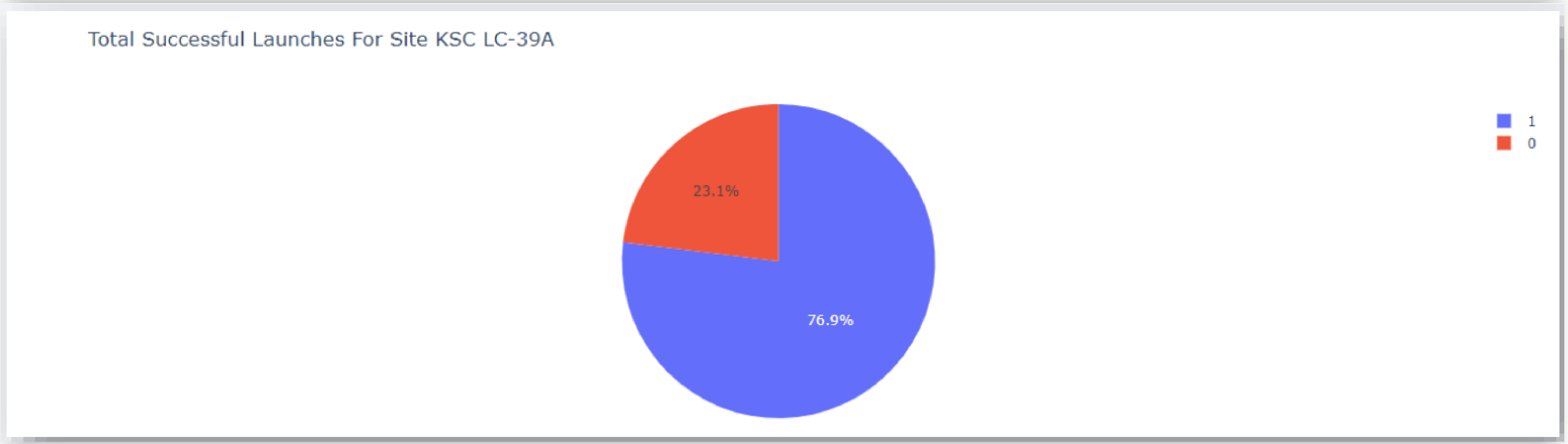
Build a Dashboard with Plotly Dash

Successful launches by site



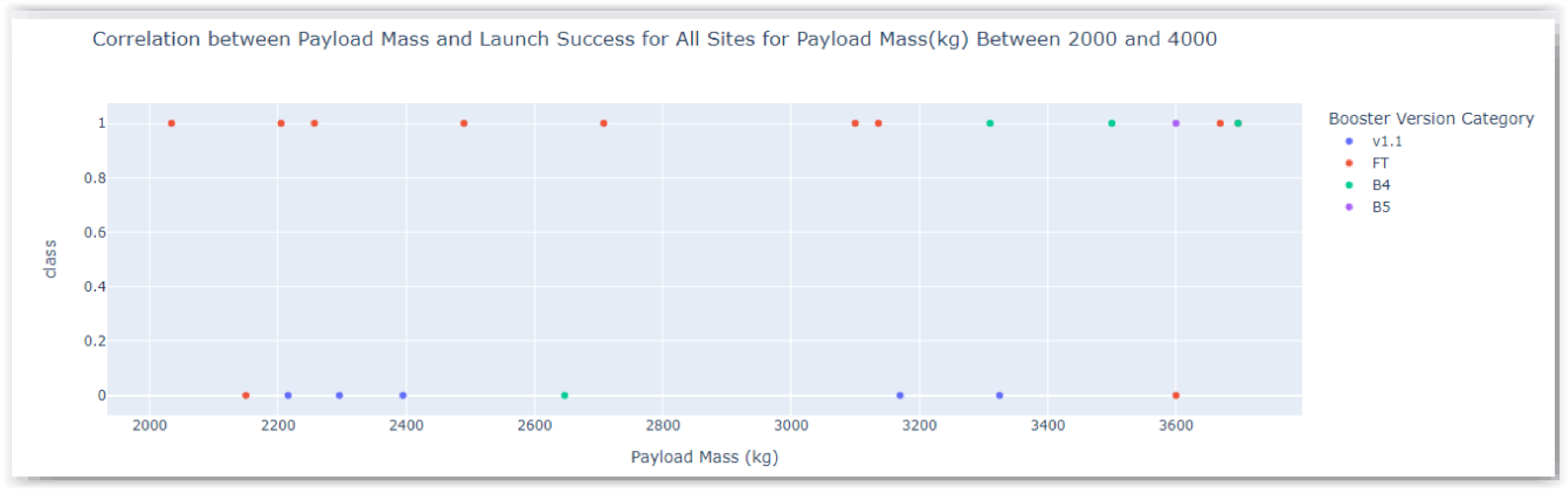
- The place of launching is a very important factor for success of mission. The majority of launches took place in KSC LC-39A site (41,7%).

Successful launches by site



- The site KSC LC-39A seems to be the most successful with a rate of 76,9% of success, not surprisingly at it was the site with most launches.

Payload mass vs launch outcome for all site



- It appears that lower payload under 4K Kg has the highest success rate if combined with FT boosters.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 4 classifications models were tested and their accuracy calculated (score method).
- The model with the highest accuracy is the decision tree with a rate of 88%.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	88.888889

Confusion Matrix

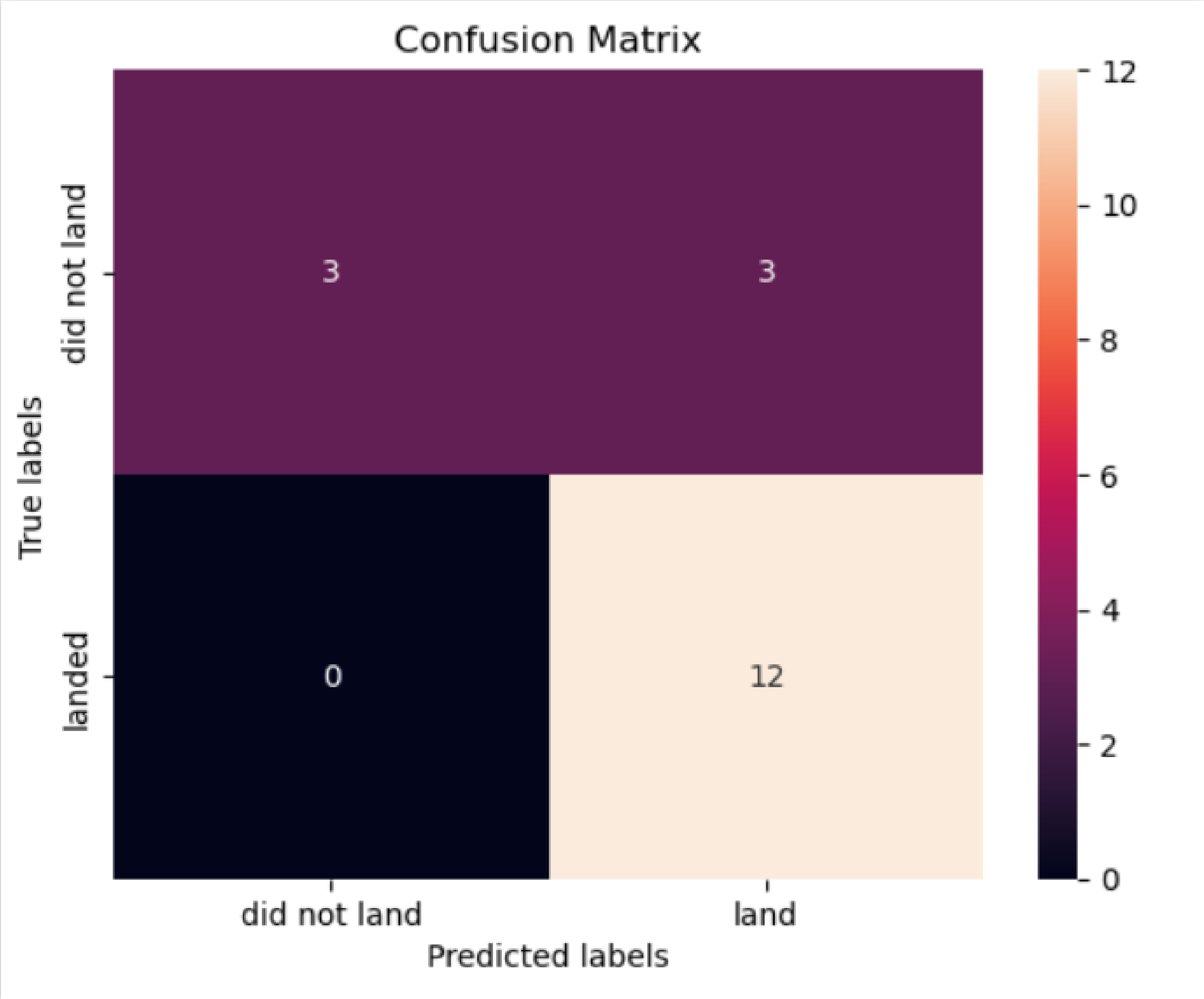
The confusion matrix for decision tree has been plotted for test data. Unfortunately, all the matrix confusion are the same for all models, so not much can be said, except that :

- The number of false positive (3) is quite high for such activity as launching rocket, which is a major problem in terms of safety and cost.
- Accuracy (correctly classified data) is the same for the 18 tests data for all models (TP+TN/Total = 0.83)

we can see we only have 18 test samples.

```
In [9]: Y_test.shape

Out[9]: (18,)
```



Conclusions

- Main findings about factors of success for launching rockets :
  - The tree classifier has the best accuracy rate (with this dataset) (88%).
  - The location of launch site is a very important factor and need to be at a fair distance of infrastructures.
  - Launch site with the highest rate of success (77%) has been found to be KSC LC-39A.
  - SSO orbit has the highest success rate with GEO and HEO (100%)
  - Lower payload (<4kg) seems to be more successful than higher payload.
  - There is a growing trend of success since 2015, as the more launches done, the more improvements are made.

Thank you!