Best parameters for successful landing of first stage

SpaceY

Vusi Mtsweni

03 July 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results - Key Findings
    - Exploratory Data Analysis (SQL)
    - Exploratory Data Analysis (Matplotlib)
    - Interactive Visualisation - Folium
    - Interactive visual analysis - Plotly Dash
    - Predictive Models
    - Model evaluation
- Conclusion

EXECUTIVE SUMMARY

- The purpose of the report is to determine the best parameters that optimise first stage rocket landing. Recycling of the first stage during rocket launches greatly reduces the cost of launching into space.
- SpaceX is a leader in the field. Without getting into the engineering details of launching, the report focusses on the parameters, such as:
  - Location of Launches
  - Model types and payload mass of the rockets
  - Target orbits of the rockets;

  To identify most efficient launches
- Company SpaceY is positioning itself to rival SpaceX dominance in the field. Therefore, analysing this data is important for outbidding SpaceX in future.

INTRODUCTION

- SpaceX currently charges $62m to launch into space, which is nearly three times less than most competitors who charge upwards of $165m.
- Analysing SpaceX launches will help SpaceY compete by identifying most efficient launch operations.
- The analysis is based on Location, Model types , Pay load mass of rockets and target orbits parameters. Varying success rates are analysed based on these parameters.
- The data is sourced from SpaceX. Exploratory data analysis, visualisation and machine learning is applied to gain insights.

METHODOLOGY

- **Webscraping**: Through the SpaceX REST API V4, data is extracted and converted into pandas dataframes
- **Preprocessing**: Data is parsed through and data wrangling is applied to filter and organise.
- **Exploratory Data Analysis**
  - SQL query through SQL Magic functions
  - Matplotlib used for visualising plots, with data sourced from Pandas Dataframe
- **Machine learning**
  - Data is seperated into training and testing sets and LogReg, Decision trees, SVM and KNN models applied to assess predictability of landing success.
  - Confusion matrices and Accuracy scores used to evaluate accuracy of models.

METHODOLOGY (Webscraping)

- **Define functions** to extract data through SpaceX API
- **Request.get()** function to retrieve data from SpaceX url
- **Decode** the response content as a Json using .json() and **convert to dataframe** using .json_normalize()
- To convert data from **ID`s to meaningful names** , SpaceX API used to **assign new values** and stored on lists to be converted to dataframe
- **Create dictionary** and use DataFrame.from_dict() convert dictionary to **refined dataframe**
- **Data Wrangling:** Find rows with **missing values, replace** with mean for PayloadMass and leave as none value for launching pads
- **Create landing class** column to classify outcome to binary results **0(unsuccesfull) and 1(successful)**
- Export as CSV

METHODOLOGY (Exploratory Data Analysis)

- **SQL**
  - **Connect to SQL** using SQL Magic function
  - **Explore data** by **grouping, filtering** independent variables to detect relationships with target variable (Landing class)
- **Pandas and Matplotlib**
  - **Visualise data** by creating **Catergory plots, Scatterplots, Barplots, Lineplots** to compare relationship between independent variables and target variable (Landing class)
  - Use the **function get_dummies()** and features dataframe to apply **OneHotEncoder** on Catergorial columns (i.e. Orbits, LaunchSite, LandingPad) . Dataframe will be **used to create predictive models**.

METHODOLOGY (Interactive Visual Analytics and Dashboards)

- **Folium**
  - **Mark location of launch sites** on map and calculate distance to proximities (coastline, railways, highways)
  - **Add landing Class markers** to visualise locations associated with Landing class success and failures
- **Interactive Dashboard with Plotly Dash**
  - **Dropdown menu** for all sites and individual sites
  - **Piechart for percentage of success** for all sites and piecharts of percentage of success for individual sites
  - **Slider for range** of PayLoadMasses
  - **Scatter chart PayLoadMass vs Landing class** for each Launch site and Booster version

METHODOLOGY (Predictive Models)

Developing Predictive Models

- **Create a NumPy array** from Landing Class column set as Y variable
- **Transform X variables** by applying transform.fit_transform(X) function to standardise data
- Use the **function train_test_split to split the data X and Y** . Set the parameter test_size to 0.2 and random_state to 2
- **Create a GridSearchCV** object with cv = 10. Find the best dictionary parameters
    - Perform GridSearchCV for Logistic regreesion, Decision trees, SVM and KNN models

Evaluating Predictive Models

- **Calculate the accuracy** on the test data using the method .score() for all the models
- **Create confusion matrices** for each model using:

yhat = cv.predict(X_test)

plot_confusion_matrix(Y_test,yhat)

RESULTS – Key Findings

- **Exploratory Data Analysis**
  - **Mission outcomes** success is **not related** to success of **first stage landing**
  - First stage landing **success rate** has **increased over time**
  - All sites have a **high increase** in landing success **beyond 7500 kg** PayLoadMass
  - **ES-L1, GEO, HEO , SSO and VLEO** Orbits have high success rate
    - **High success linked to limited launches** only SSO has more than 1 launch (5 launches)

- **Interactive Visualisation (Folium)**
  - **Marked launchsites** are near coastlines, railroads and far enough from major highways and cities. All launch sites near the equator.
  - **Proximities** are related to operational logistics (railroads), landing sites (offshore), infrastucture and human preservation (Cities & Highways) and physics (equator)

- **Interactive Visual analysis (Plotly Dash)**
  - **KSC LC-39A** most successful site at **41.7%**.
  - **Highest success** rate at **2500 – 5500 kg** PayloadMass. **Influenced by KSC LC-39A** high success
  - Booster Version FT the most successful booster
- **Predictive Models**
  - 3 Models have similar accuracy score at 83.33%. **Decision trees Model best at 88.88%**

Exploratory Data Analysis(SQL)

- **Four unique sites**

```
%sql Select Distinct Launch_Site from SPACEXTABLE
```

* sqlite:///my_data1.db
Done.

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

- **First Successful landing**

```
%sql Select MIN(Date) from SPACEXTABLE where  Landing_Outcome Like 'Success%';
```

* sqlite:///my_data1.db
Done.

MIN(Date)
2015-12-22

Exploratory Data Analysis(SQL)

- Successful **mission outcome (99% successful)**, cannot be related to first stage **landing success (variable)**

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS TOTAL_NUMBER
```

* sqlite:///my_data1.db
Done.

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

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

* sqlite:///my_data1.db
Done.

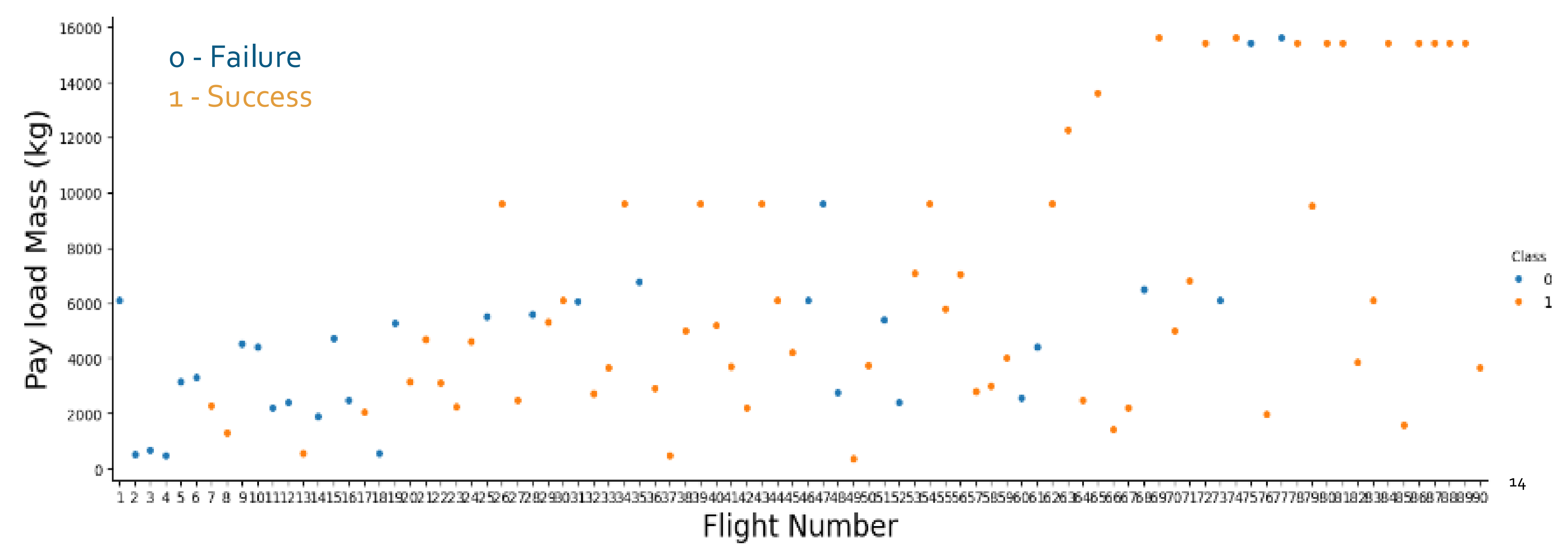
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Exploratory Data Analysis

Matplotlib

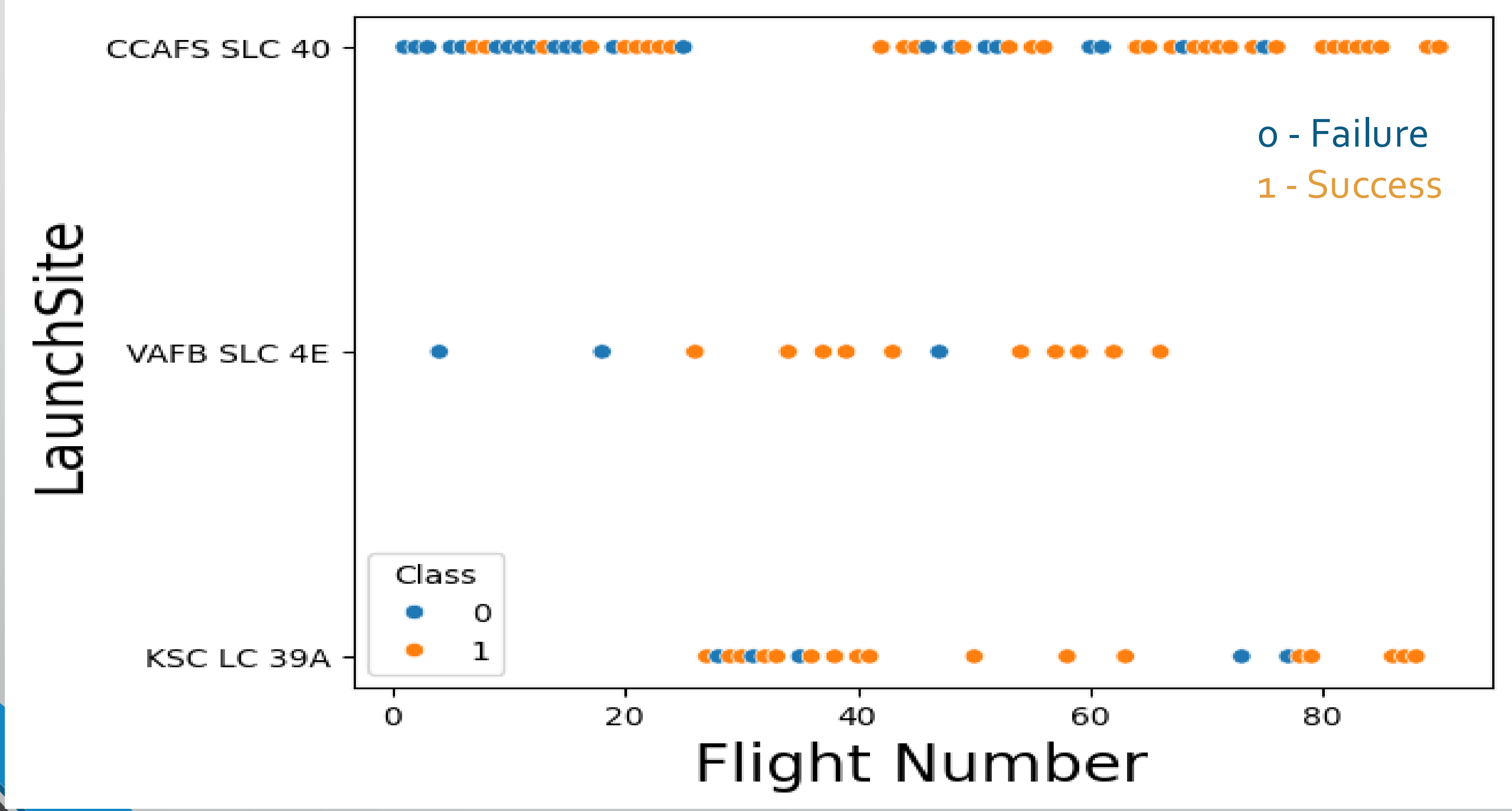
Flight Number vs PayLoadMass

- The is a **general increase in landing success** with increase in number of flights and PayLoadMass



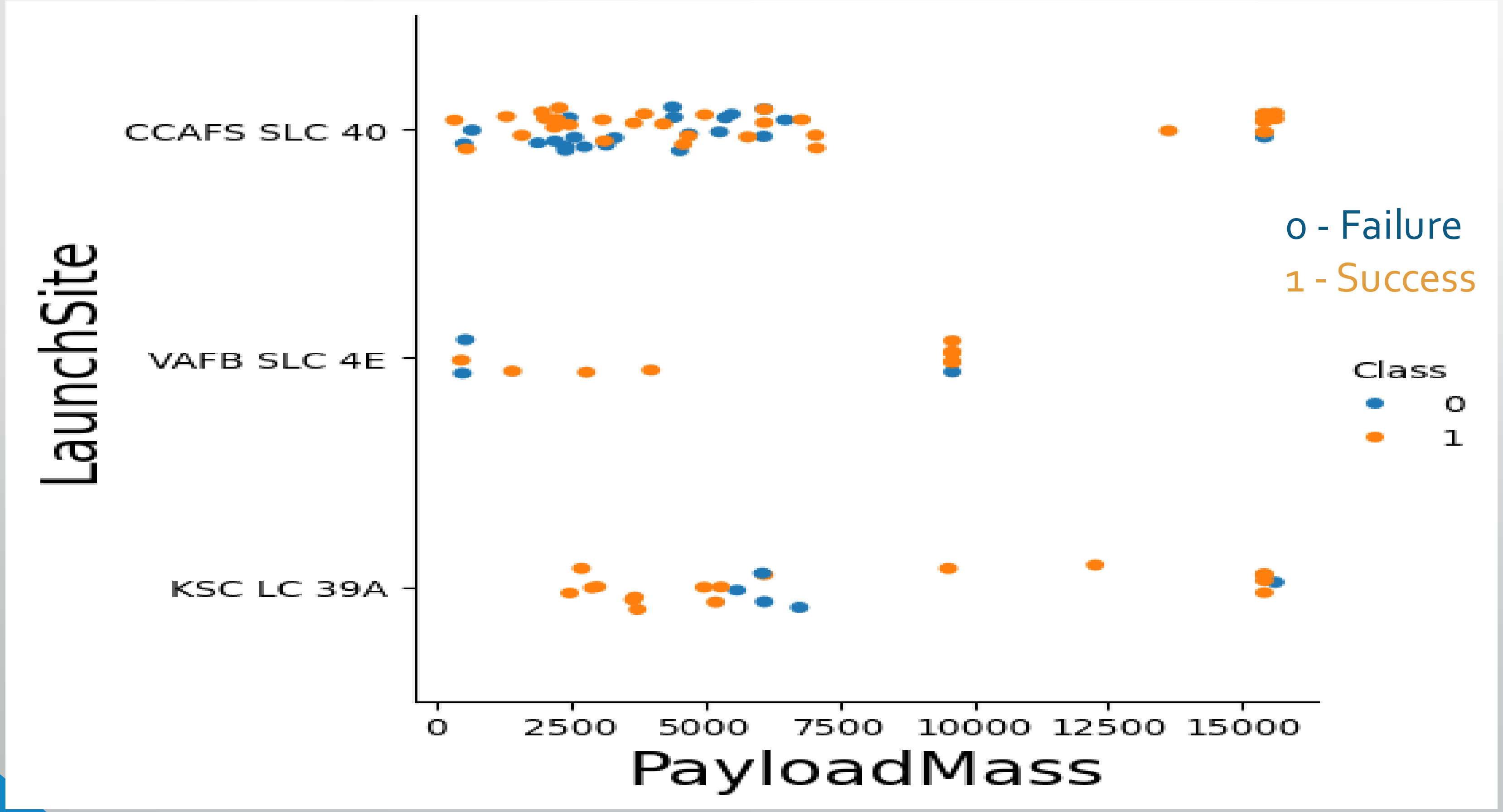
Flightnumber vs Launchsite

- The is a **general increase** in successful landing with increase in number of flights across **all Launch sites**



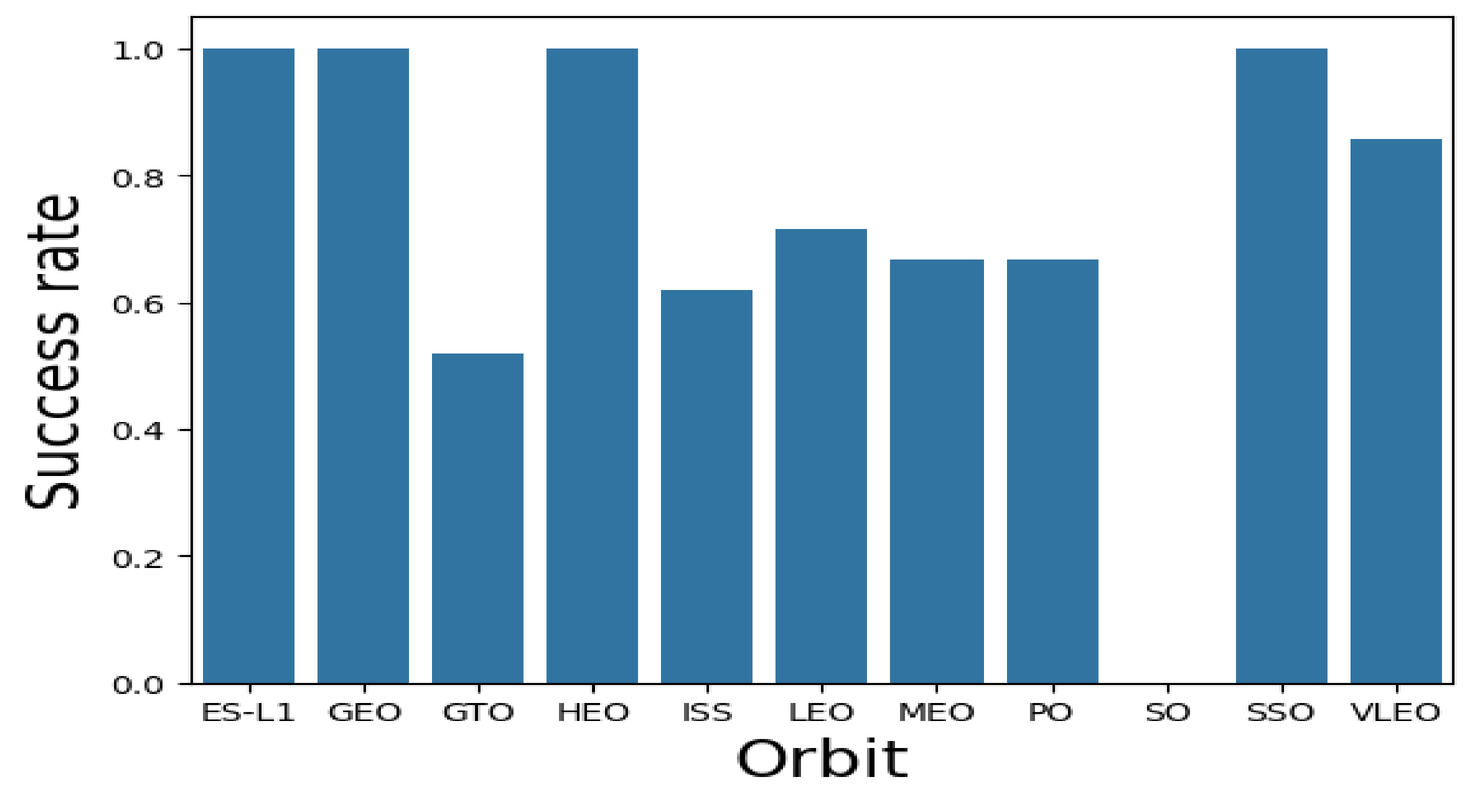
PayLoadMass vs Launchsite

- **Mixed correlation** between PayLoadMass and Launchsite
  - Amount of **successful landings increase** compared to failed landings for All the sites at **payloads bigger than 7500kg**
  - **KSC LC 39A** has **high successful landings** when payloads **smaller than 5000kg**



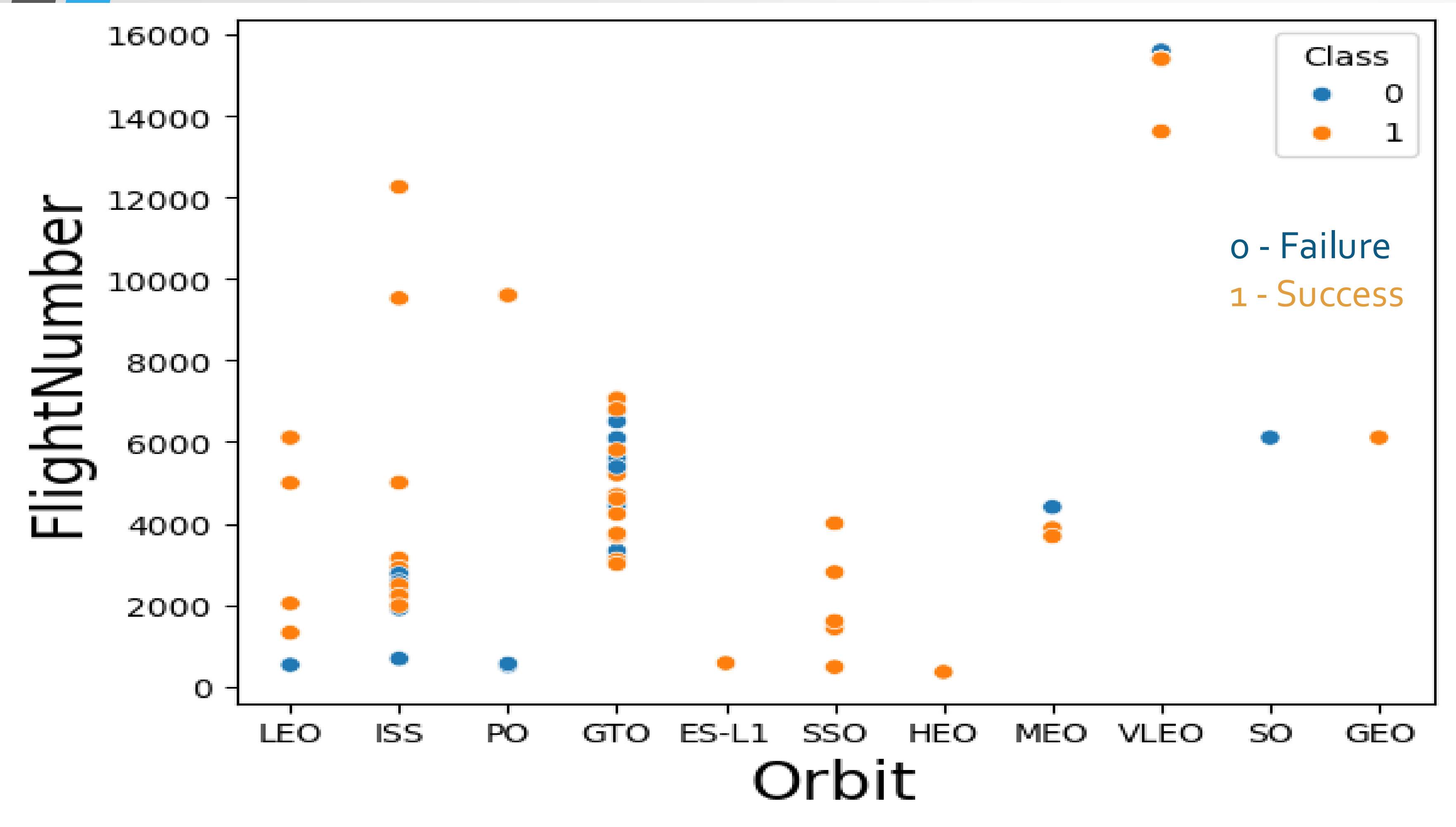
Orbit Success Rate

	Orbit Name	Count
High Success Rate [80-100%]:	ES-L1, GEO, HEO and SSO	4
Moderate Success Rate [50-80%]	GTO, ISS, LEO, MEO, PO, VLEO	6
No Success [0%]	SO	1
	Total	11



Orbit vs FlightNumber

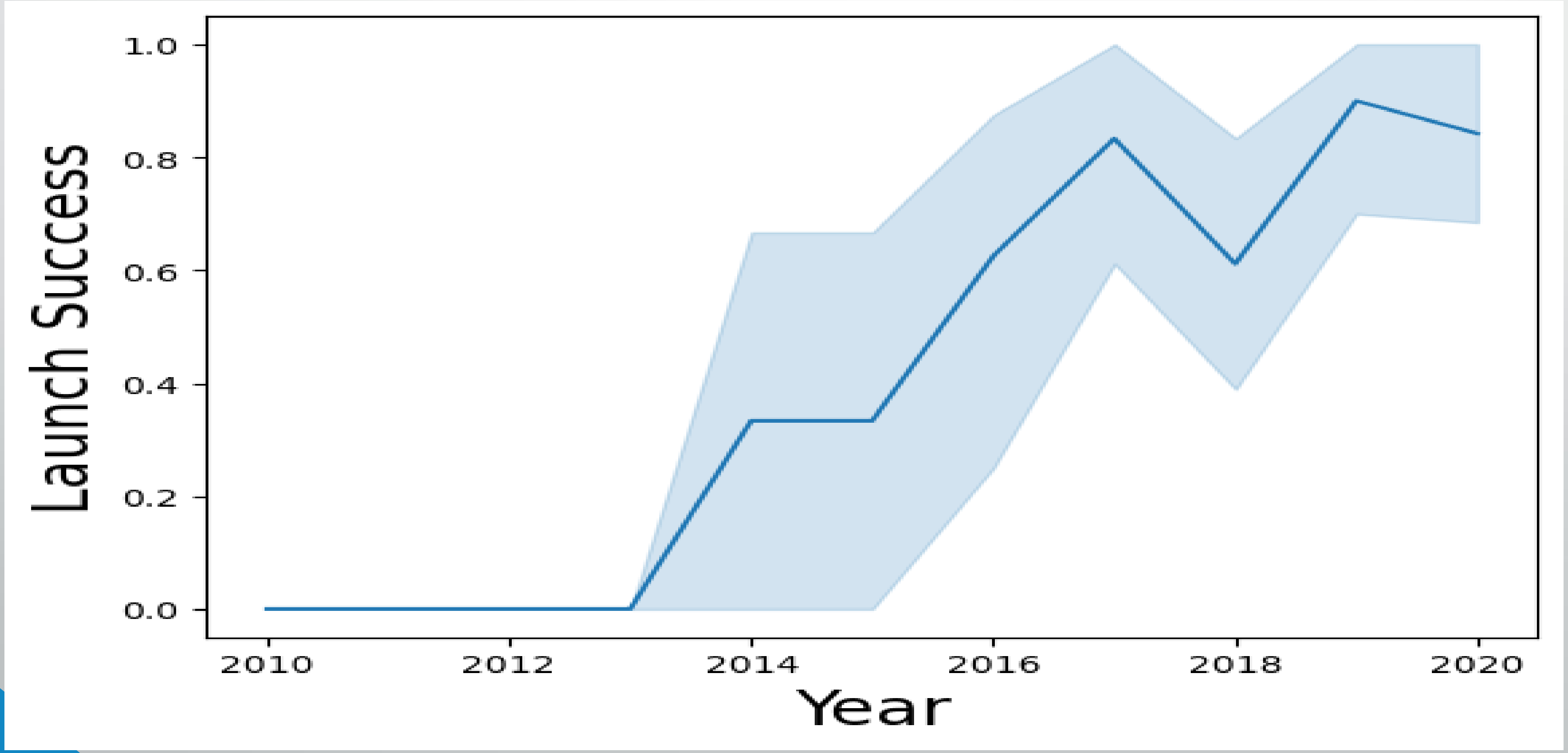
- Orbits with **highest success rates** associated with **fewer flight numbers**



- Sun-Synchronous Orbit (SSO) has more flight numbers and success rate. High success rate based on synchronicity with source of light making it easier to land payload.

Time series of successful launches

- Since intial succcessful landing on 2015-12-22, **Launch success has increased** only major dip was in 2018



Interactive Visualisation

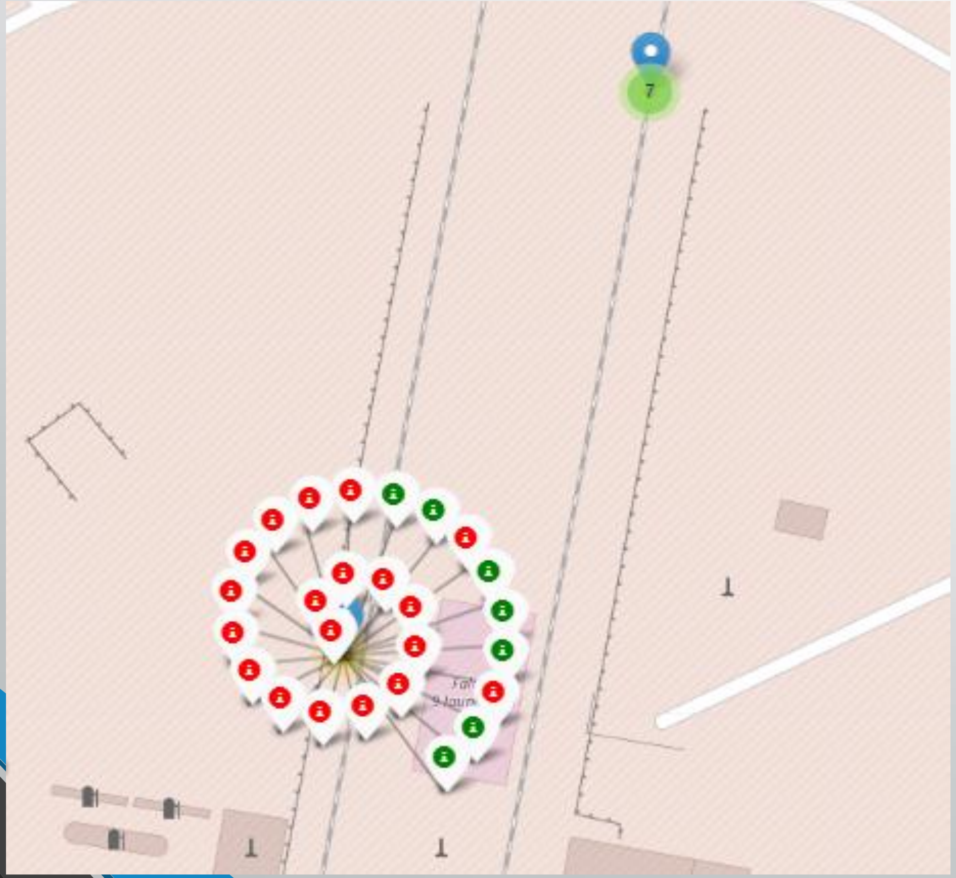
Folium

Initial Map with Location Markers

- East and west coast location markers showing number of launches


KSC
LC-
39A
13
33
Kennedy Parkway North
Saturn Barge Channel
National Wildlife Refuge

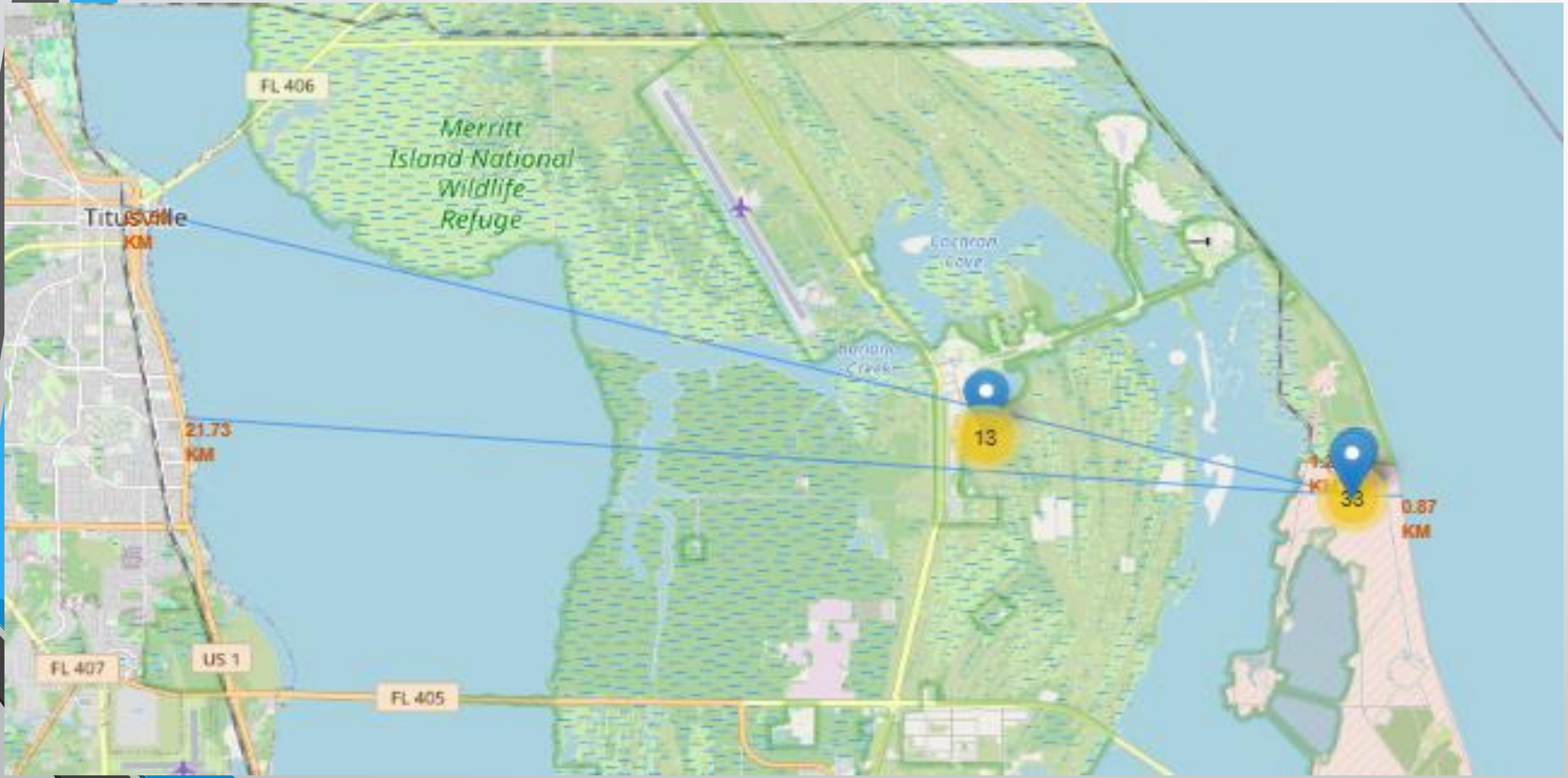
- **CCAFS LC-40 Launch site** with associated tags for successful landings green and unsuccessful launches red



- **VAFB SLC-4E Launch site** with associated tags for successful landings green and unsuccessful launches red

Launchsite Proximities

- `Distance_highway = 21.732582300517876 km`
- `Distance_railroad = 1.216752521066563 km`
- `Distance_city = 23.501031770852844 km`
  `Distance_coastline = 0.87 km`



- Launches are **near the equator** in an easterly direction, as this **maximizes** use of the Earth's **rotational speed** (465 m/s at the equator)

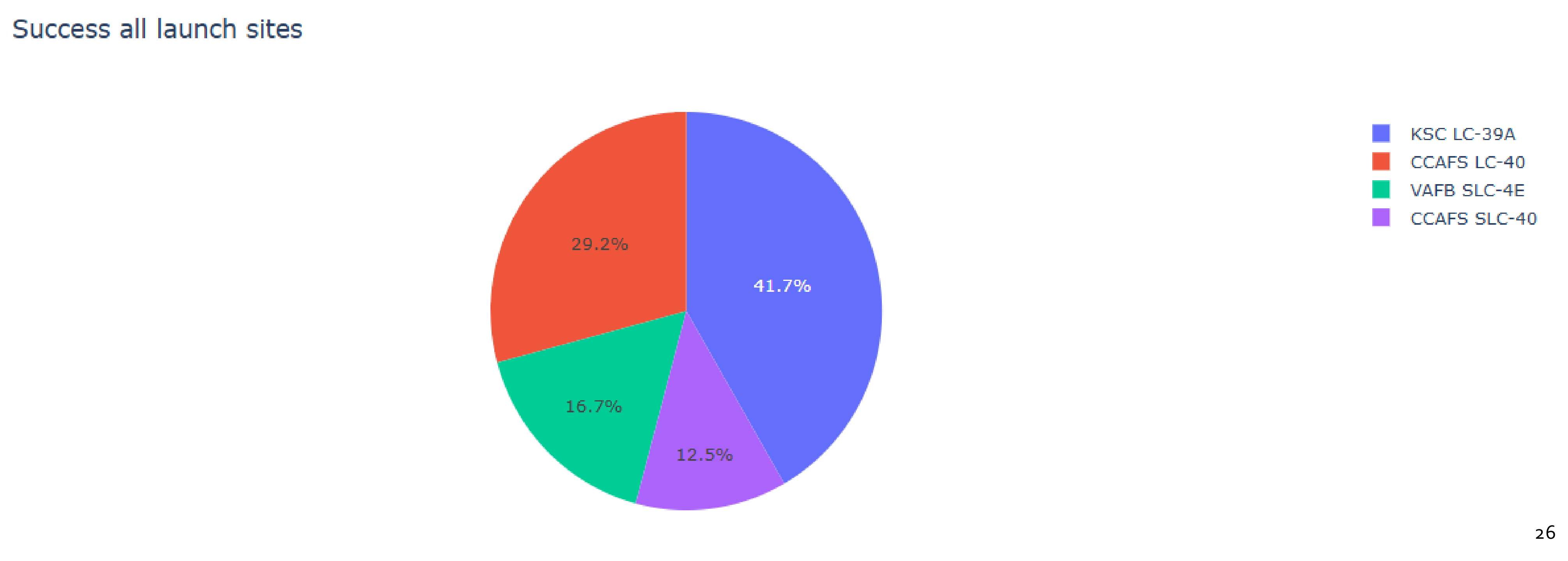
- Launchsites are **close to coast** proximate to landing site
- **Far enough** from city, highway & railroad to **avoid civilian and infrastructure damage** incase of crash
- **Close enough for logistics**

Interactive Visual analysis

Plotly Dash

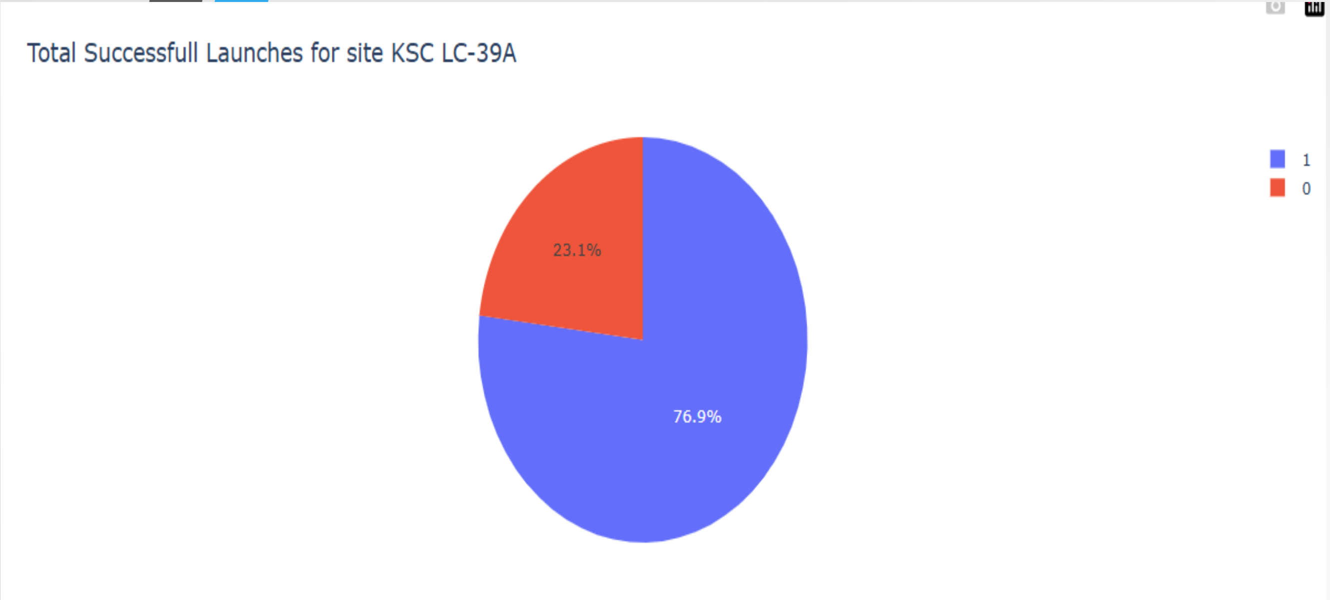
Percentage of Success for all Launch sites

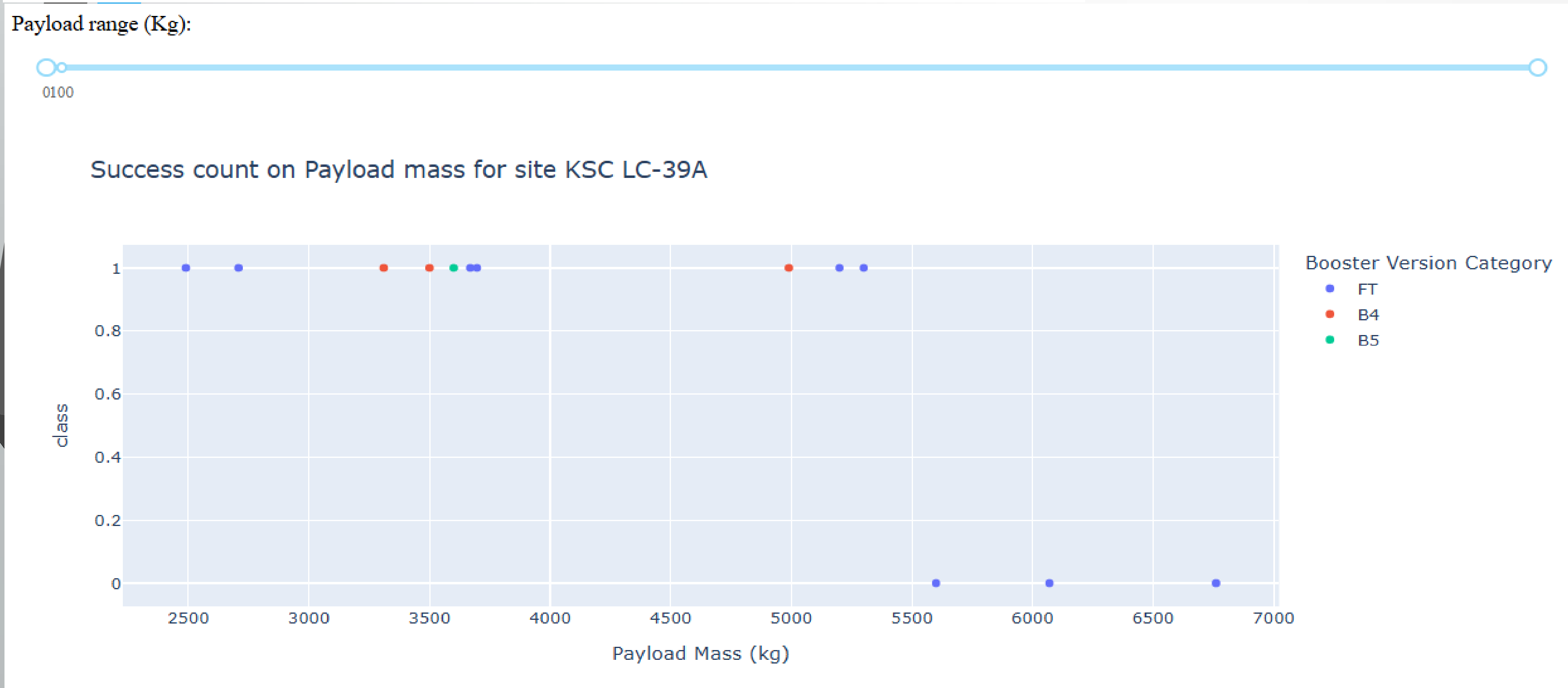
- **KSC LC-39A** most successful with a share of **41.7%**



Most Successful site

KSC LC-39A

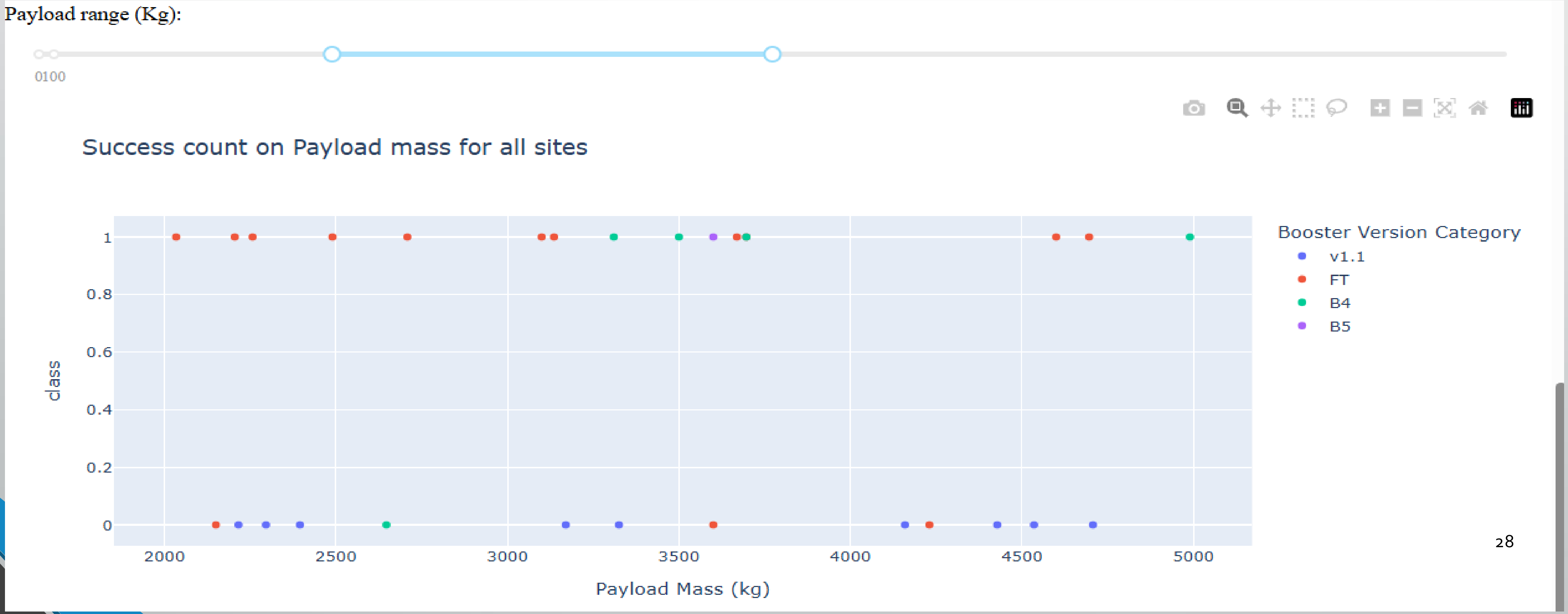




- **77%** of KSC LC -39A Landings are **successful**
- All **Booster version Payload** masses **successful** between **2500-5500 kg**
- **Failure** associated with **Booster version FT** between **5500-7000 kg**

Most Successful PayLoadMass

- **PayLoadMass** between **2000-5000 kg** represents **most successful landings**

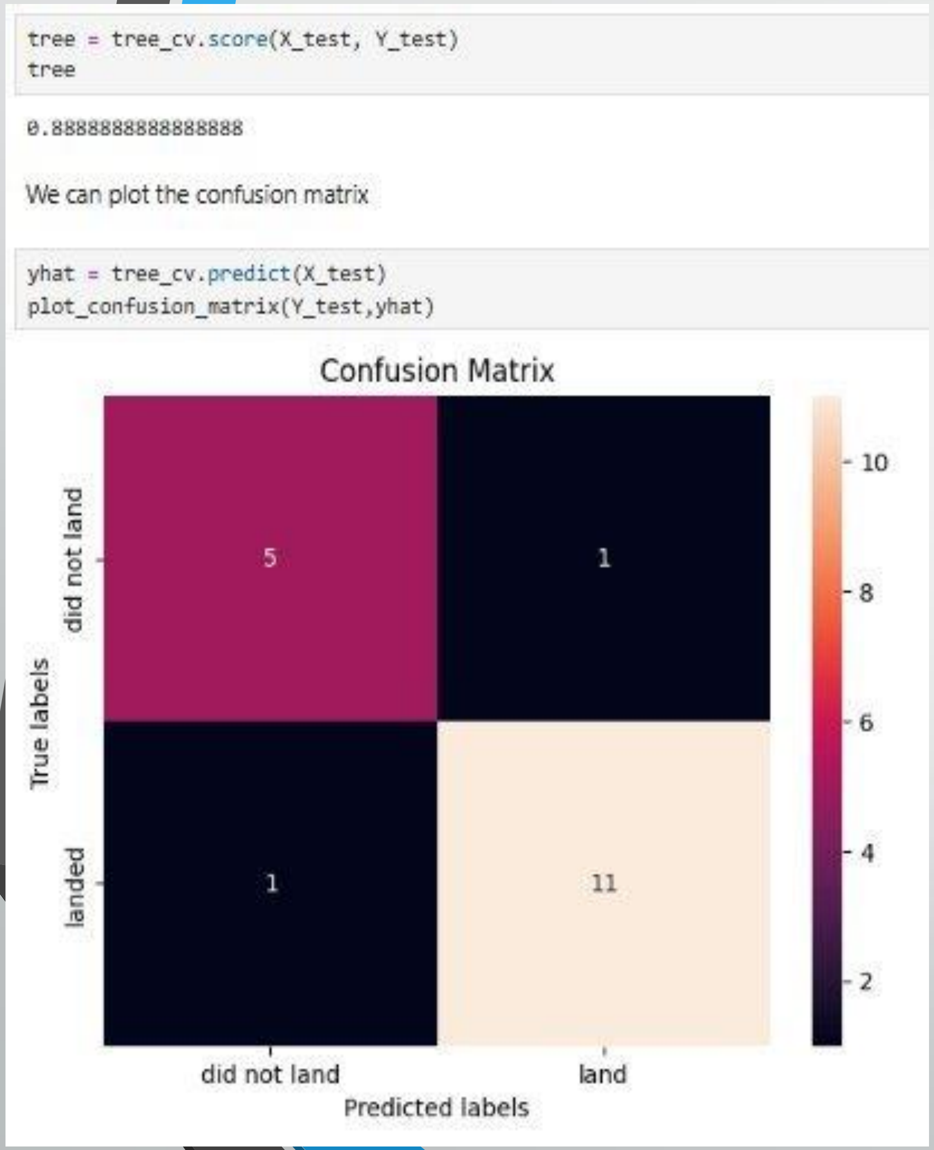


Booster with highest successful rate

- Booster Version FT has 65.3% successful landings
- The success rate increases to 77% when its confined to **2000-5000 kg PayloadMass**

Predictive Models

- Having a **small data set** and **training** a large portion of the data set **(80%)** yielded **high accuracy** in the test sets for the four models built:
  - Logistical Regression
  - Decision Trees
  - SVM
  - KNN
- **Lunchsite, PayLoadMass, Orbits, Booster version** independent variables are sufficient to build the models because they show **distinct relationships** with target variable **landing success**



Model Evaluation

- **LR, SVM, KNN** yielded similar **accuracy ~ 83.33%**
- The **Decision trees** model yielded the best accuracy score of all the models **88.88%**
- The **confusion matrix** algorithm predicted
  - 11 true positives vs 1 false positive
  - 1 false negative vs 5 True negatives

CONCLUSION

- Success rate increases with time
- **Choosing launch** site is important for maximising landing success for SpaceY
  - Launch site proximities should be considered (i.e. coastline, railways, cities)
  - **KSC LC-39A** site should be pursued because it **shows high success at low (2500-5500 kg) and high(<7500 kg) PayLoadMasses**.
  - The is a **general increase** in success **with increase in PayLoadMass** for all sites
- SpaceY should also **choose missions based on targeted Orbit** to maximise landing success
  - Out of the 4 Orbits with 100% success, SSO is the only Orbit with considerable number of launches to pursue.
- **Parameters chosen** for predictive modeling show **strong correlation with target variable** (Landing success)
  - The 4 LR, Decision trees, SVM, KNN Models chosen yielded high accuracy scores. **Decision trees** edges the other 3 with **88.88% accuracy**
  - A **larger data** set should be considered for future studies