IBM Professional Data Science Capstone Project

SpaceX landing

Pasha Bayramli

25.4.2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
  - Data collection and data wrangling
  - EDA and Interactive visual analytics methodology
  - Predictive analysis methodology
- Results
  - Visualization – Charts
  - Dashboard
  - EDA with SQL
  - Predictive analysis (Classification)
- Discussion
  - Findings & Implications
- Conclusion

Executive Summary

- Main object: obtain data, analyze, visualize and predict the possibility of case study upon successful first launch landing of SpaceX rockets on different launch locations, boosters, payload mass, and etc. data types.
- The data used in this analysis is obtained from SpaceX API and filtered/cleaned upon necessary data types.
- Data manipulation is done via Pandas data-frames and SQL
- After manipulation data was explained by visual means through matplotlib, plotly dashboard tools
- Normalization of data and creation of dummy columns were made to further predictive analysis mainly through sklearn ML packages. Models were trained and tested in order to obtain accuracy and further evaluation
- Eventually, successful landing possibilities were found via three ML models such as logistic regression, support vector machine and decision tree.
- Successful landing possibility, practically, is the same for all models : 0.833 (1.0 is the highest confidence of successful landing)

Introduction

- In the capstone project, the prediction of first stage landing of the Falcon 9 space rocket will be done. Main purpose of the whole project is to reduce the company costs by predicting successful first landing which allows the company to reuse the rocket itself for further launches leading to a enormous cost savings for the company. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Methodology

Data collection and data wrangling

Create functions to retrieve data with required keywords

- Create get request
- Retrieve requested content and filter by already created functions
- Normalize data
- Create corresponding columns and append data into the one dataframe

Data filtering and manipulation

- Choose Falcon 9 booster
- Replace missing values with mean value with payload mass
- Save data-frame

Data wrangling

- Create outcome column as "Class"
- Save the dataset

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	M
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	M
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	M
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	M
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	S
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B

rsion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
con 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
con 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
con 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
con 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
con 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Methodology

EDA and Interactive visual analytics methodology

Create charts to understand relationships of success rates

- Payload Mass vs Flight number
- Launch site vs Flight number
- Launch site vs Payload mass
- Bar chart: success rate vs Orbit type (group by)
- Orbit type vs Flight number
- Orbit type vs Payload

Visualize launch success yearly trend

- Extract years out of date column
- Group success rate by launch year

Define features for further analysis

- Depending on the data visualization done earlier determine features for prediction models
- Create dummies Orbit, Launch sites, Landing Pads, and Serial numbers
- Turn each column into float type

Create dashboards

- Create pie chart showing Launch sites and success rates
- Create Payload vs Success rate each launch sites with booster version information

Methodology

Predictive analysis methodology

Introduce features and normalize data → Train – test data → Prediction models → Choose best model

Introduce features and normalize data

- X and Y values from the data
- Normalization of each data feature column

Train – test data

- Split data into training and testing data
- Visualize the shape of each data

Prediction models

- Use Grid search Cv for tuning each model
- ML model - Logistic regression
- ML model – Support Vector Machine
- ML model – Decision Tree
- ML model – K nearest neighbors
- Evaluate each model accuracy and test the prediction accuracy
- Create confusion matrix for better understanding the accuracy of each model

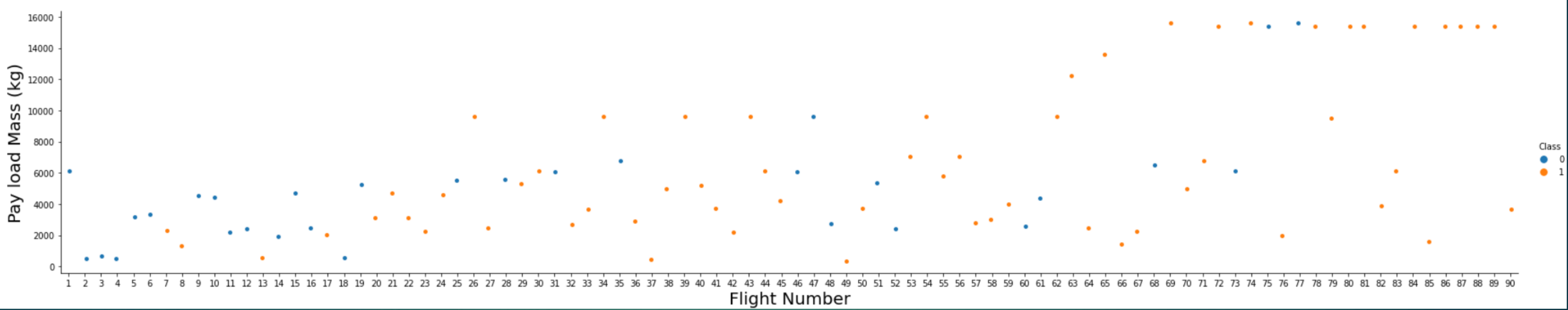
Choose best model

- Choosing best accurate model

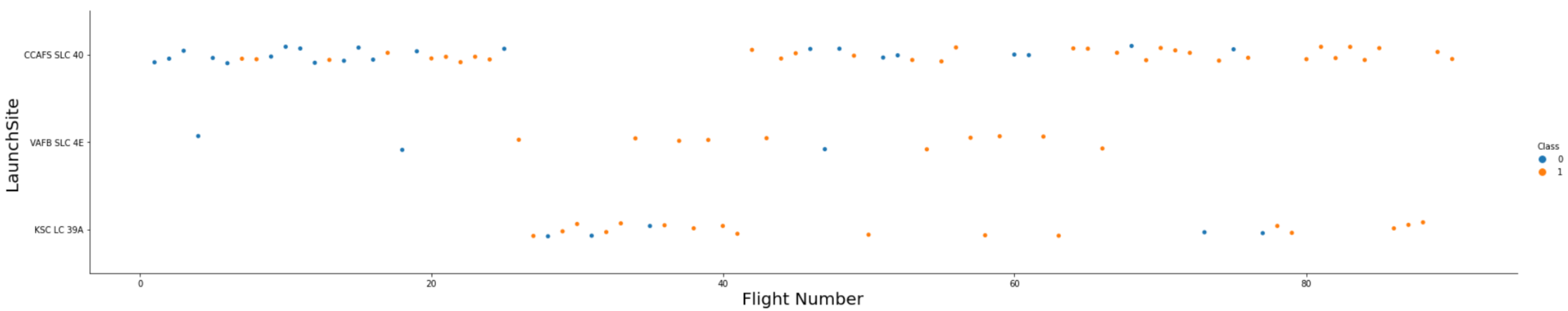
Results

EDA with visualization

Payload mass, kg vs Flight number



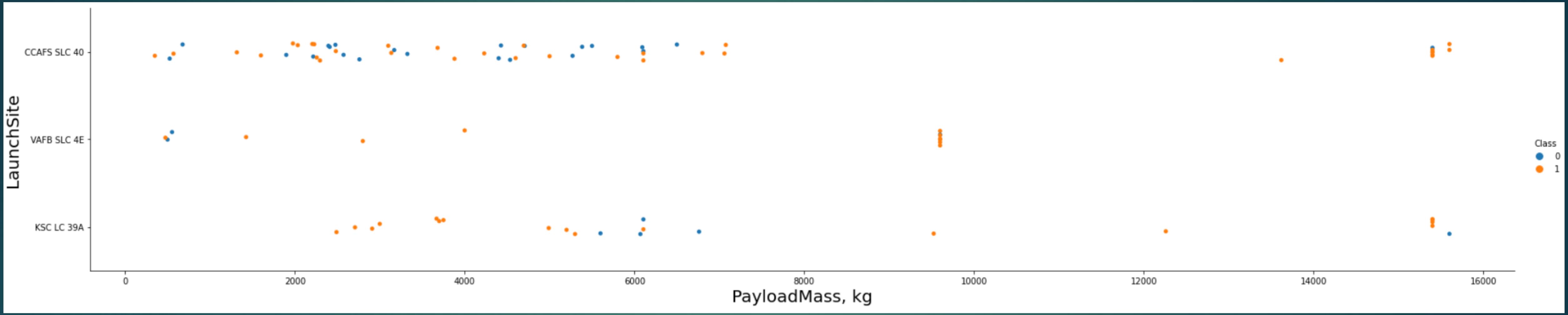
Launch site vs Flight number



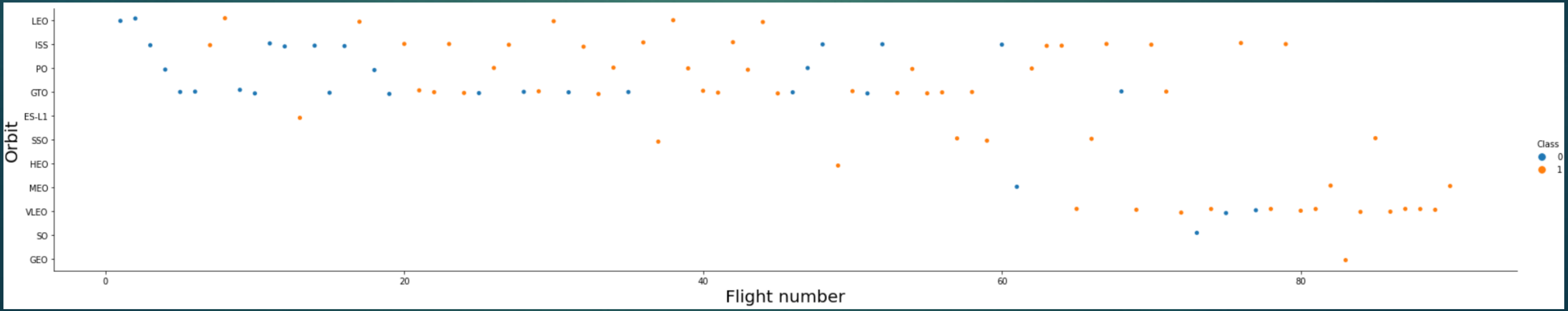
Results

EDA with visualization

Launch site vs Paylaod mass, kg



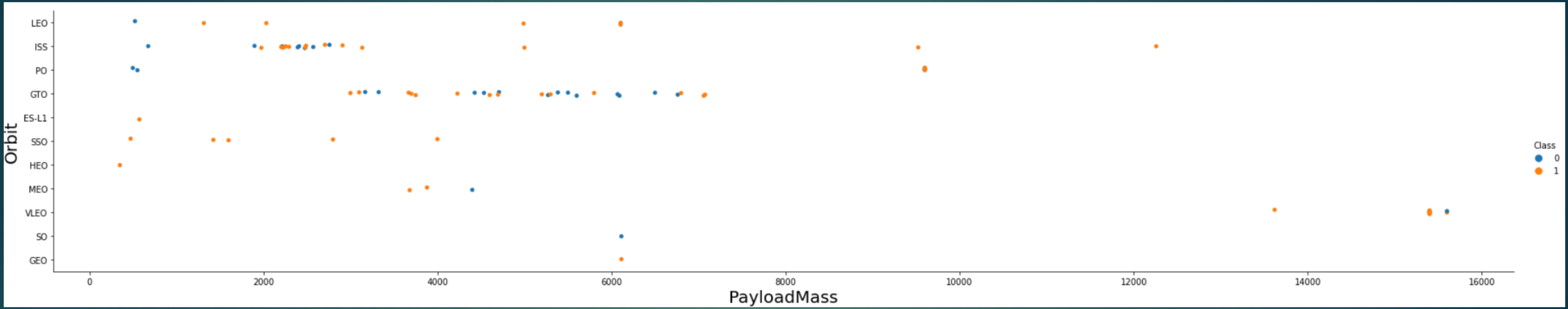
Orbit vs Flight number



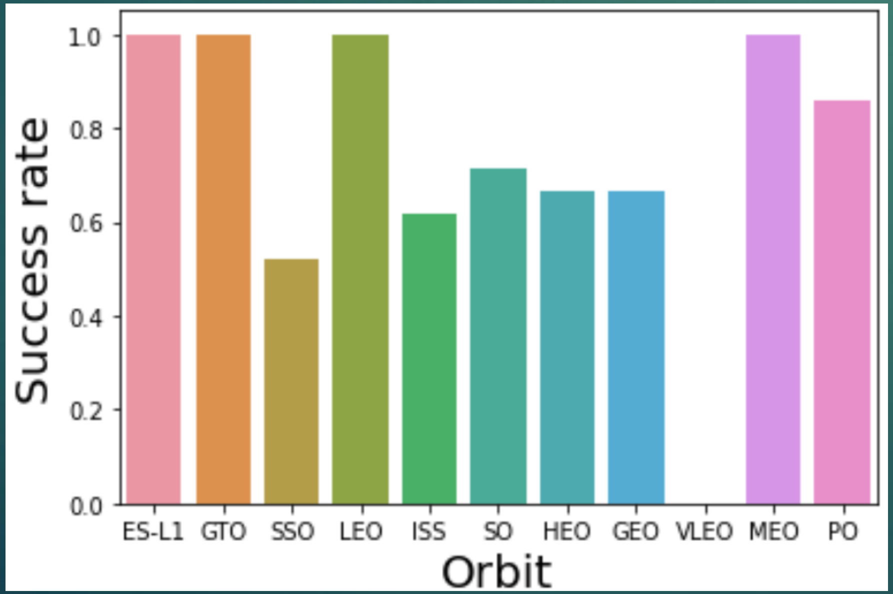
Results

EDA with visualization

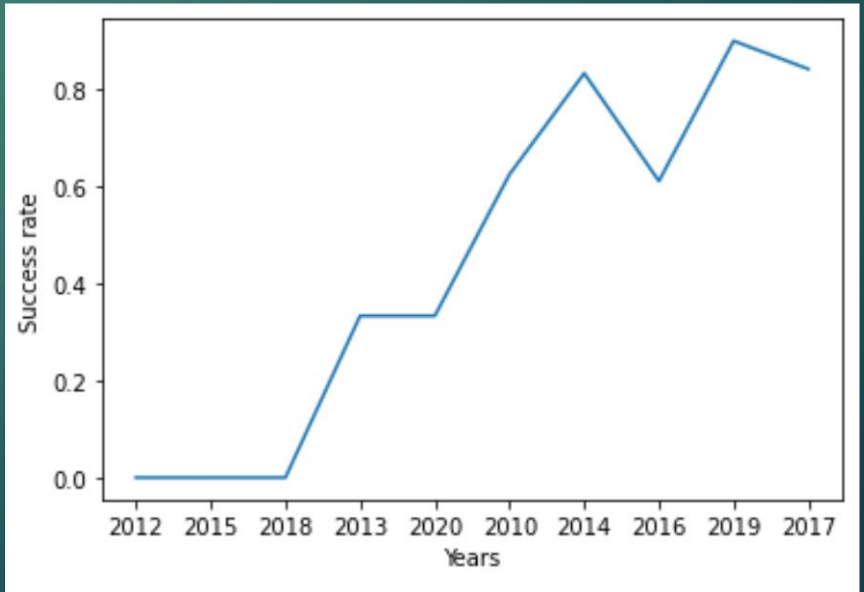
Orbit vs Payload mass, kg



Success rate vs Orbit



Success rate vs Years



Results

EDA with SQL

- **Names of unique launch sites**

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

- **Launch sites beginning 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

- **Total payload mass of NASA (CRS)'s boosters**

total_mass_nasa_crs_kg
45596

- **Average payload mass carried by booster version F9 v 1.1**

average_mass_f9_v1_1_kg
2928

- **First successful landing date with ground pad outcome**

first_successfull_landing
2015-12-22

- **Booster versions with success in drone ship and payload mass 4000-6000 kg**

booster_version	landing__outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Results

EDA with SQL

- **Total number of successful and failed missions**

success_count	failure_count
100	1

- **Boosters with maximum payload mass (15600 kg)**

max_pm_boosters
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

- **Failed cases with drone ship outcome in 2015**

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

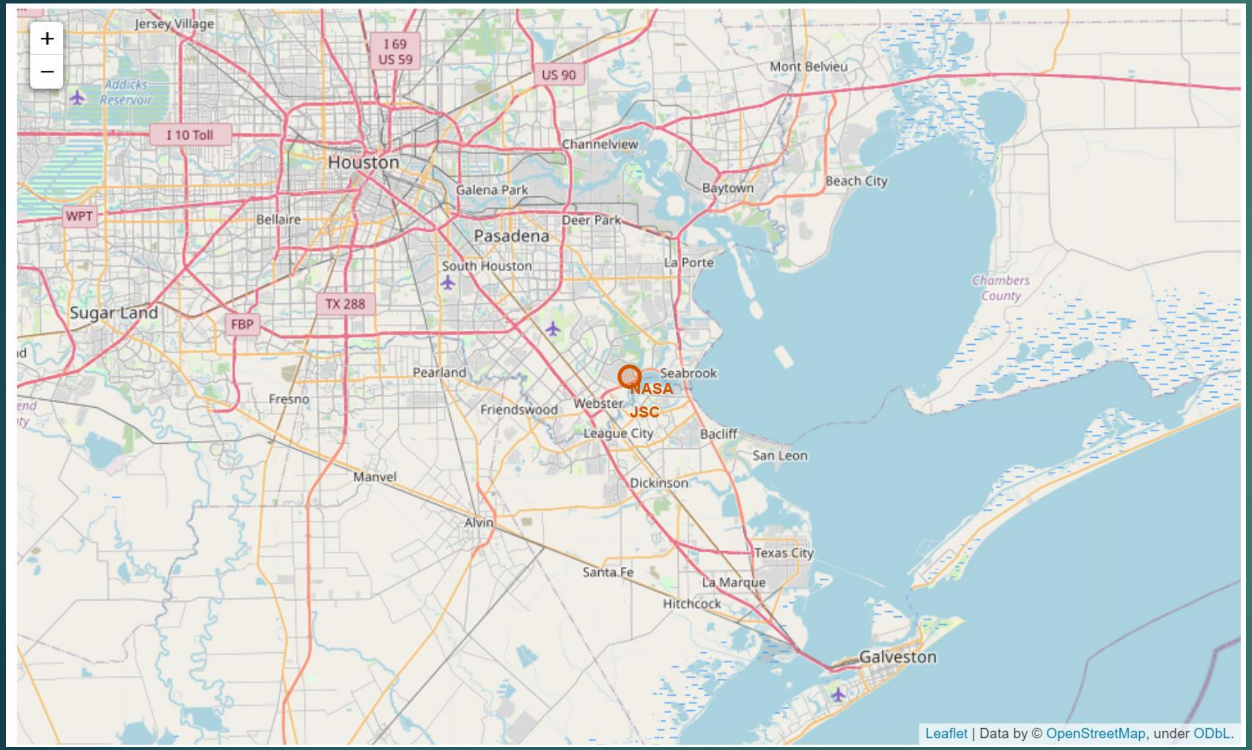
- **Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20**

landing__outcome	COUNT
Success (ground pad)	9
Failure (drone ship)	5

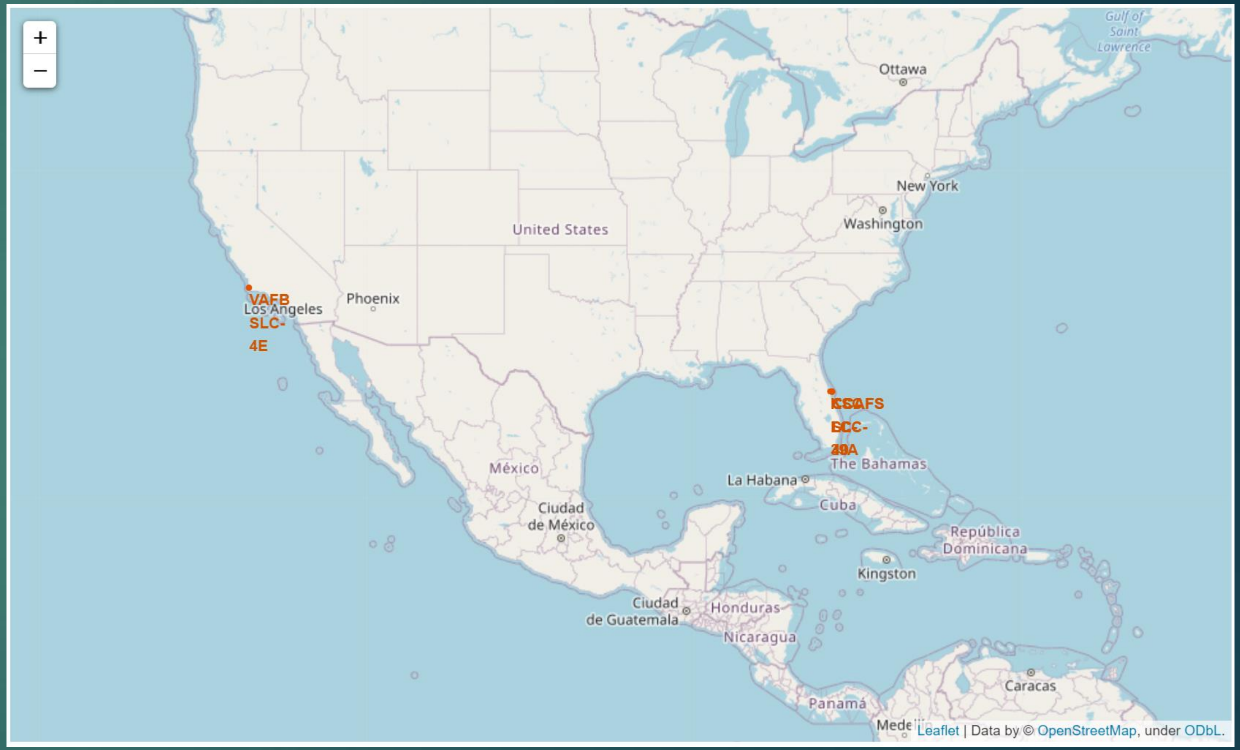
Results

Interactive map with Folium

Nasa Johnson Space center



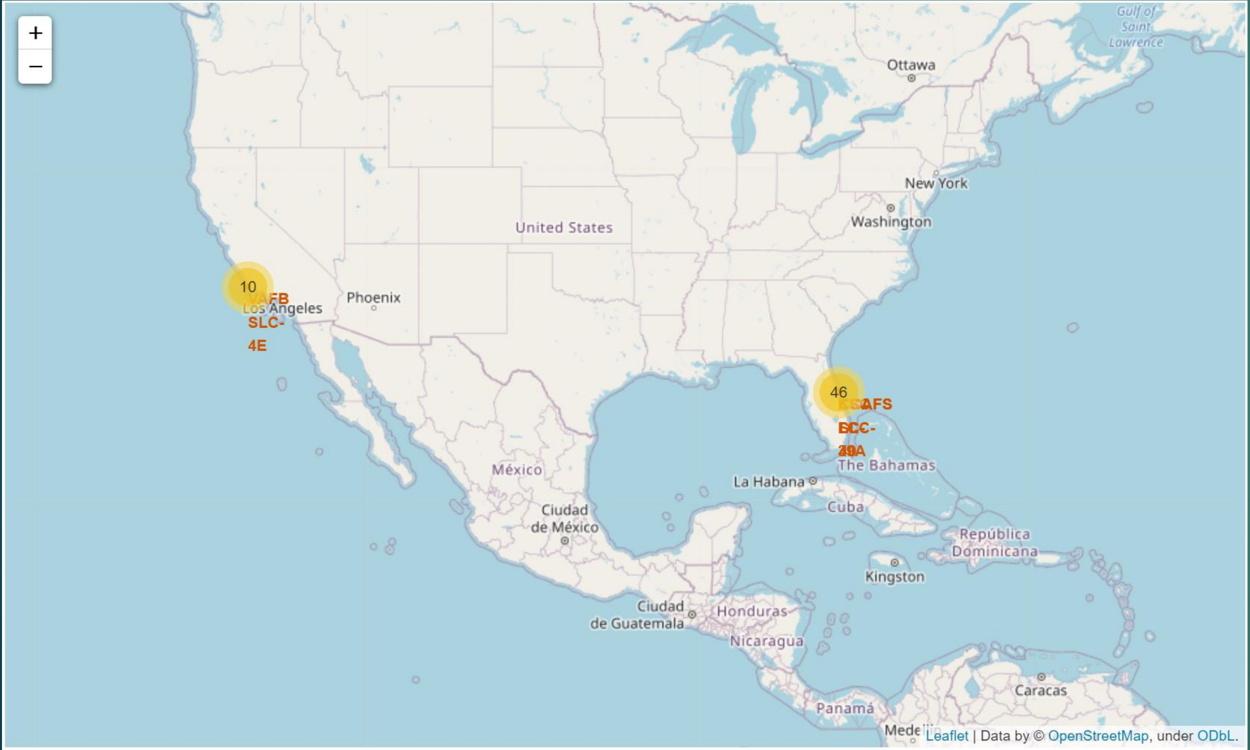
Launch sites with names shown



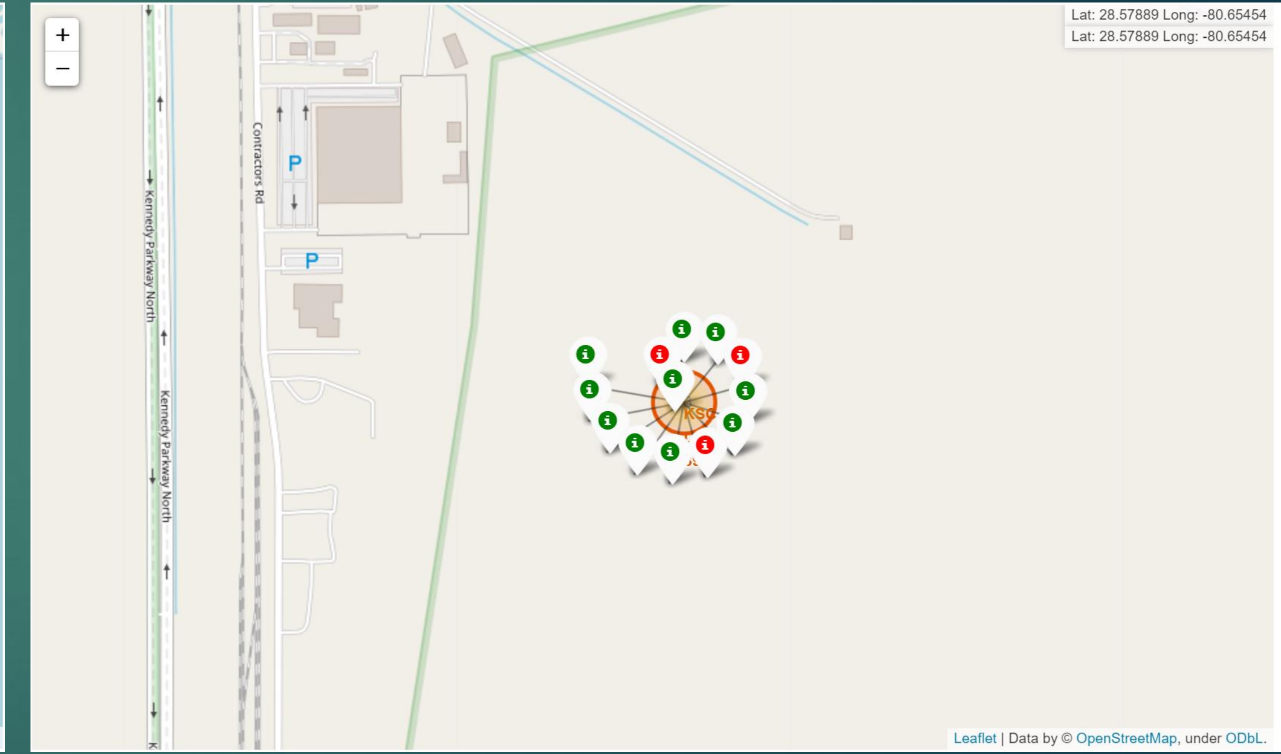
Results

Interactive map with Folium

Launch site – each success and failure cases



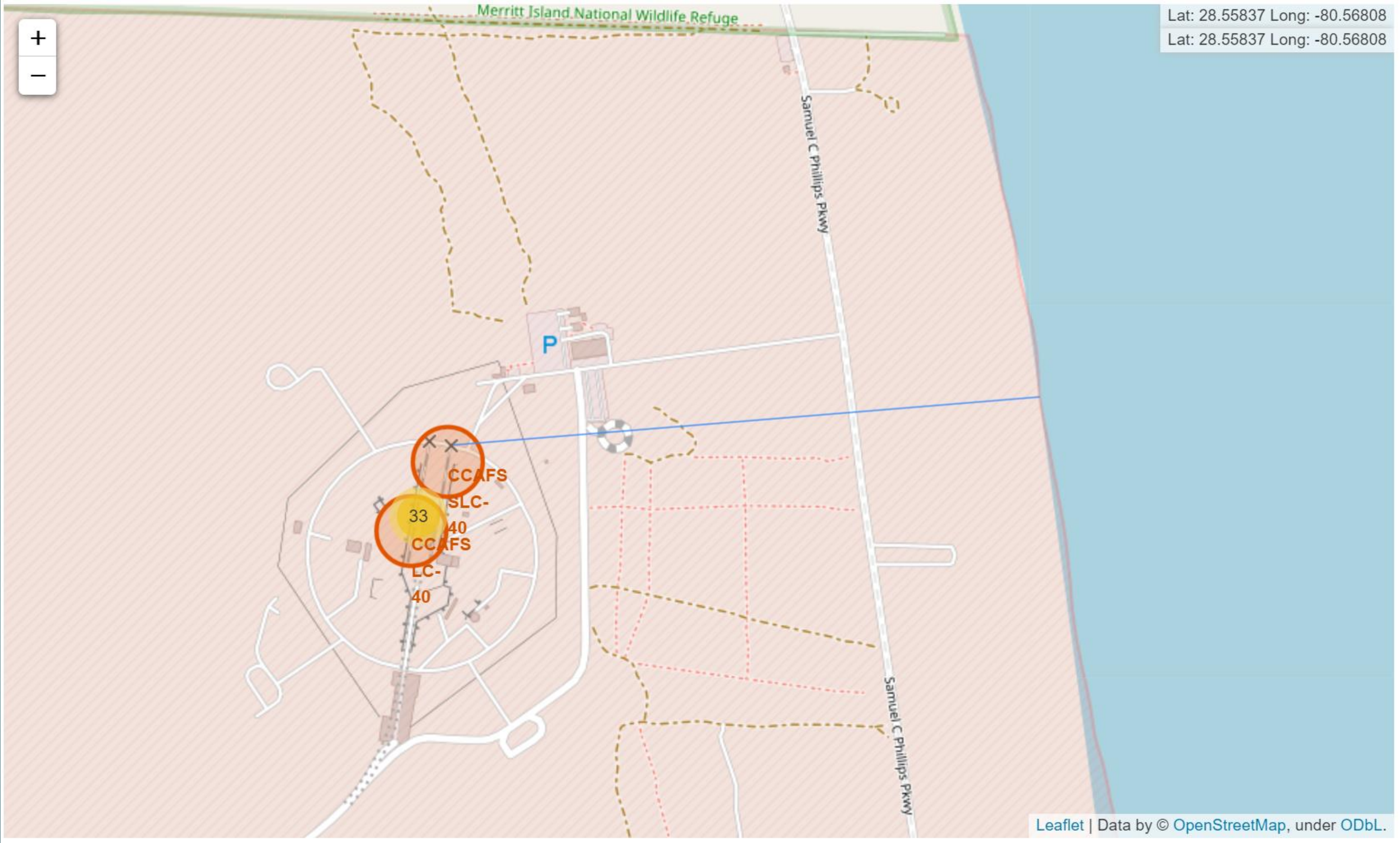
Marking successfull (green) and failure (red) launches



Results

Interactive map with Folium

Distance to the coastline from the closest launch site

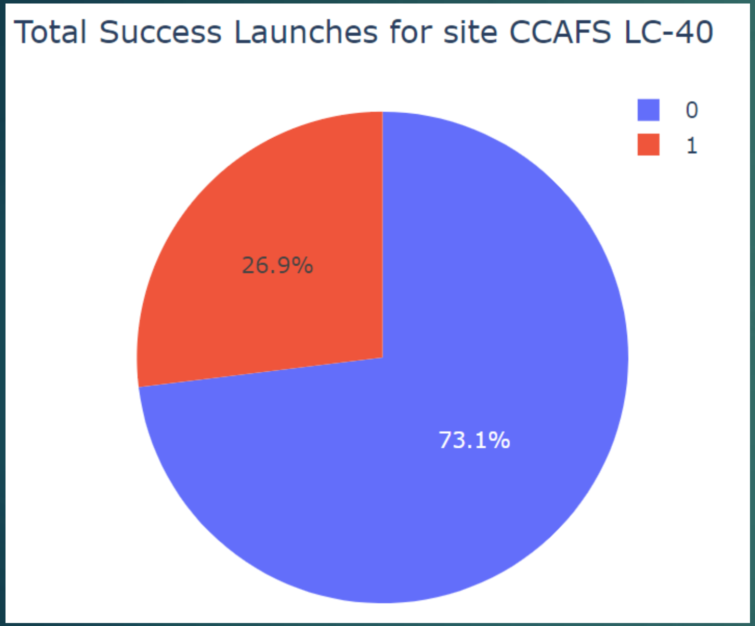


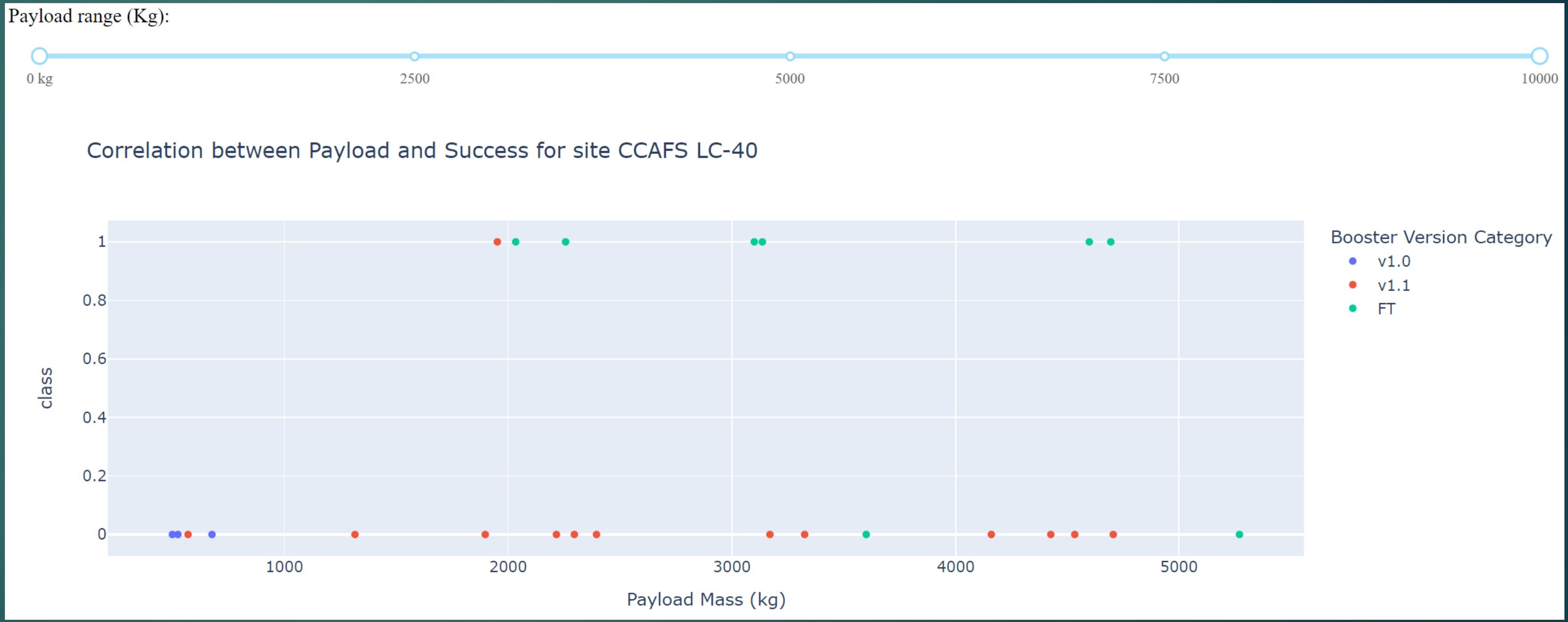
Results

Plotly Dash dashboard

Launch site details: CCAFS LC-40

- **Success rate**
- **Payload mass distribution vs success rates along booster version categories**



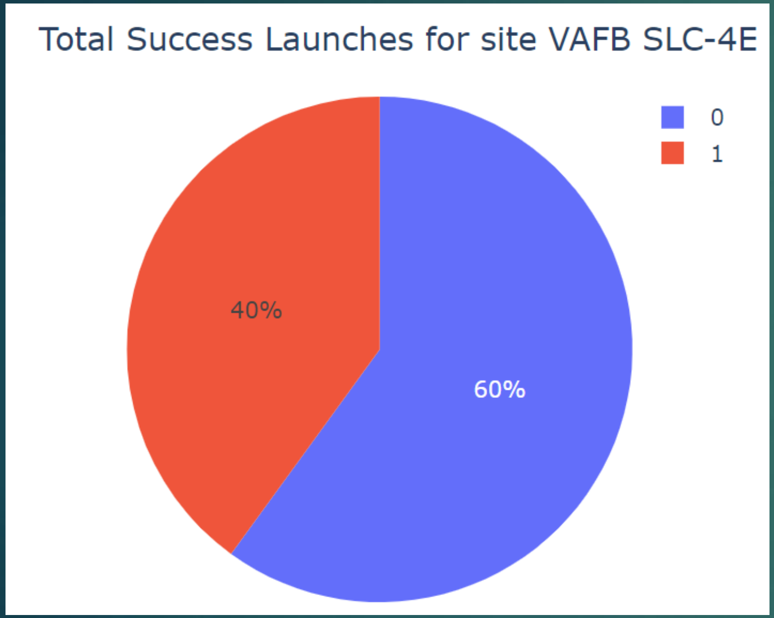


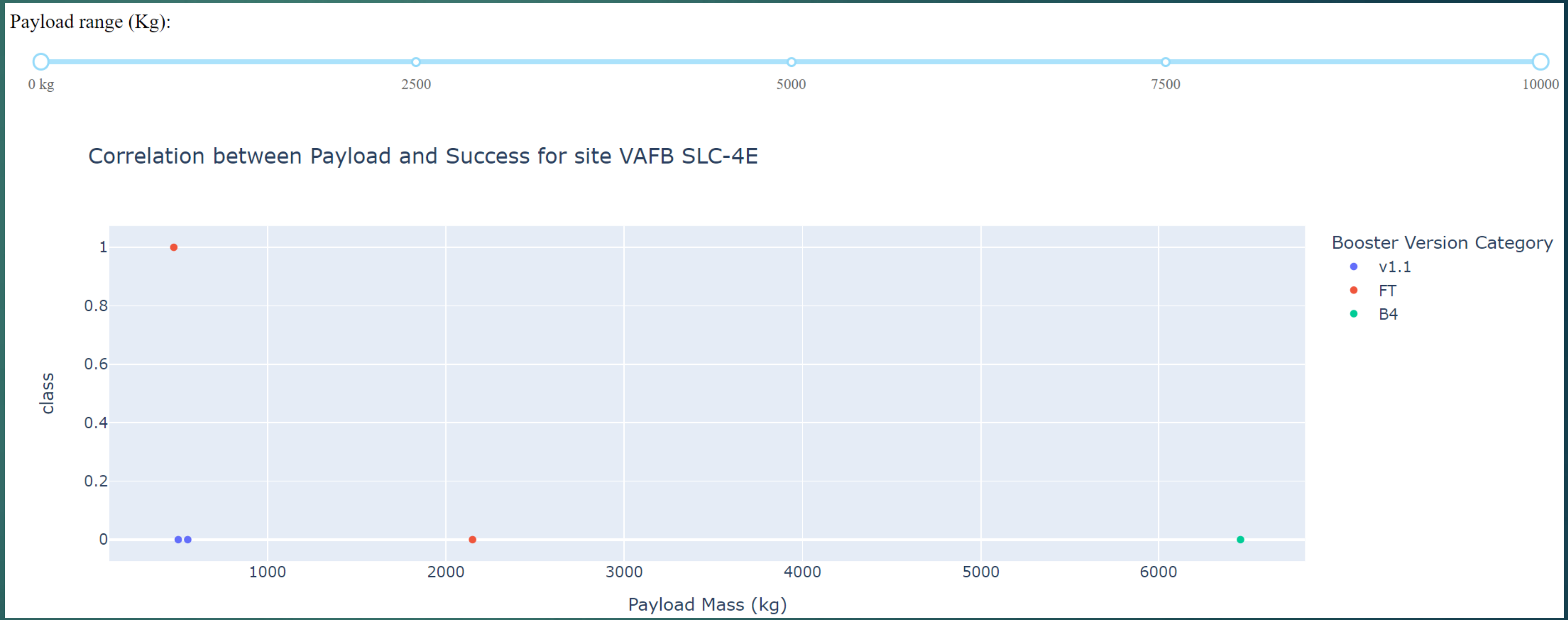
Results

Plotly Dash dashboard

Launch site details: VAFB SLC-4E

- **Success rate**
- **Payload mass distribution vs success rates along booster version categories**



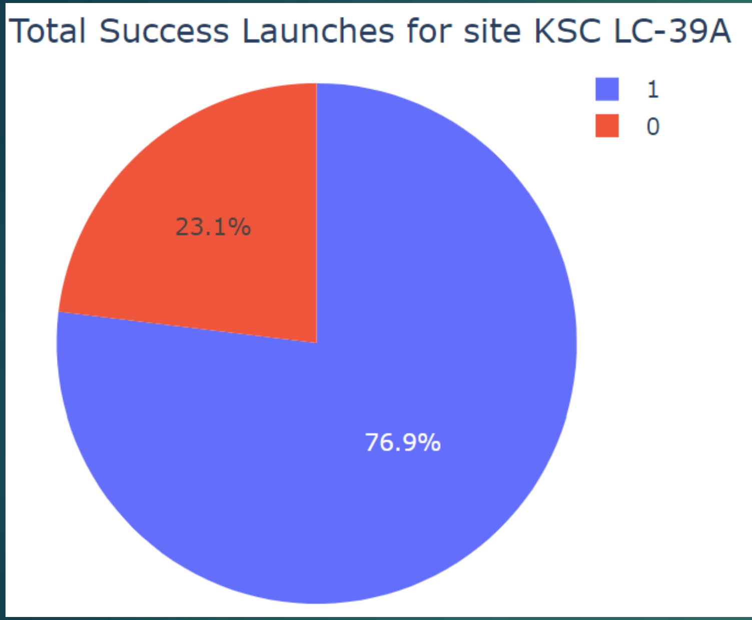


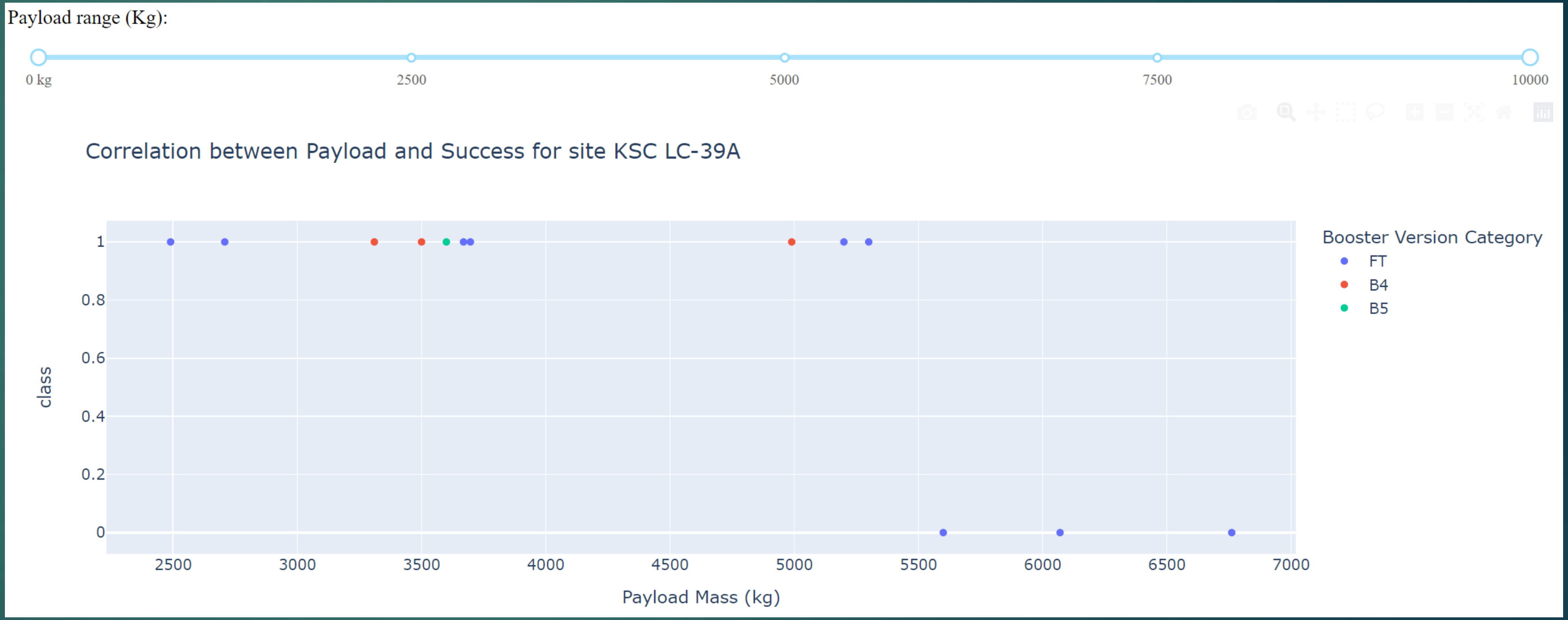
Results

Plotly Dash dashboard

Launch site details: KSC LC-39A

- **Success rate**
- **Payload mass distribution vs success rates along booster version categories**



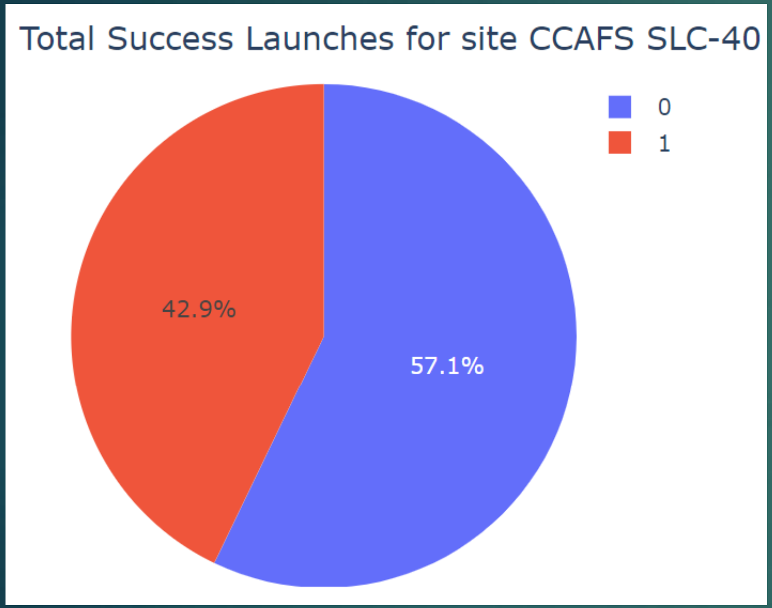


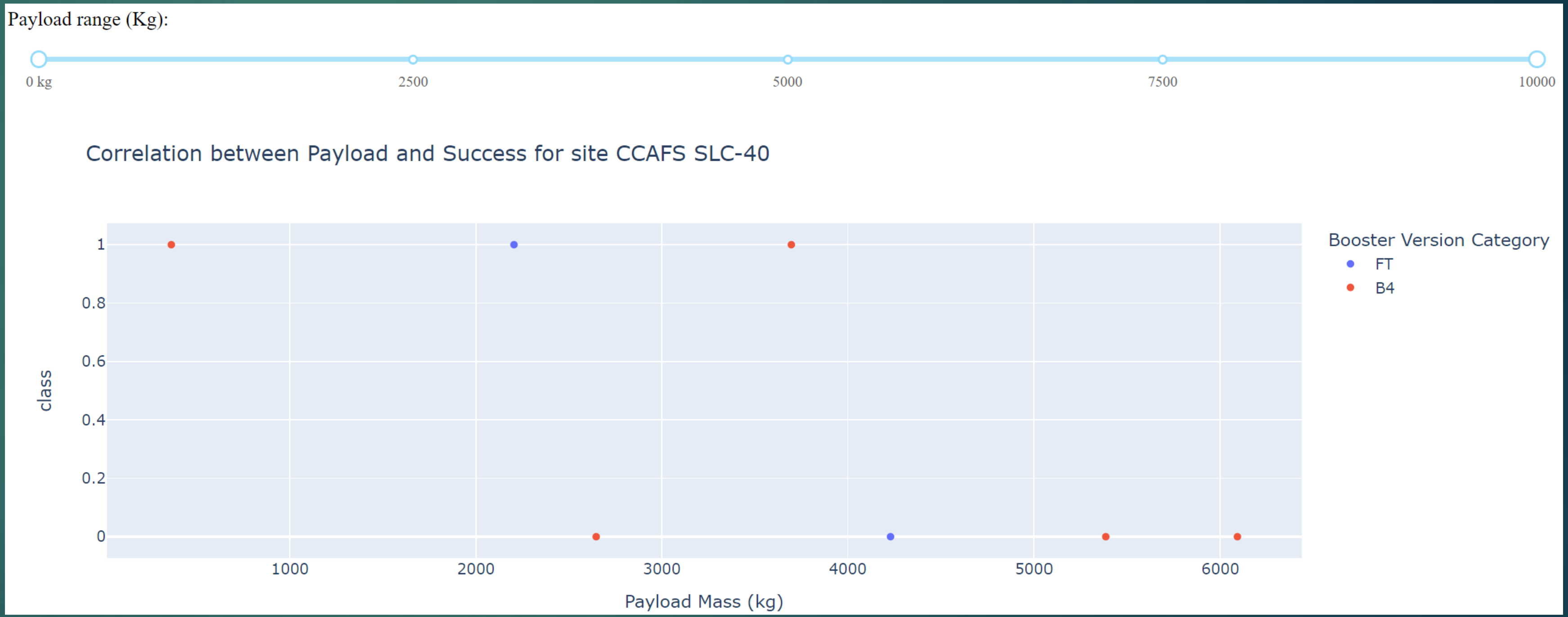
Results

Plotly Dash dashboard

Launch site details: CCAFS SLC-40

- **Success rate**
- **Payload mass distribution vs success rates along booster version categories**





Results

Predictive analysis (classification)

Training and testing data shapes

```
Training and testing data shapes
Train set: (72, 83) (72, 1)
Test set: (18, 83) (18, 1)
```

Normalized data with dummies (categorical data)

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058	Serial_B1059	Seri
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	

5 rows × 83 columns

Results

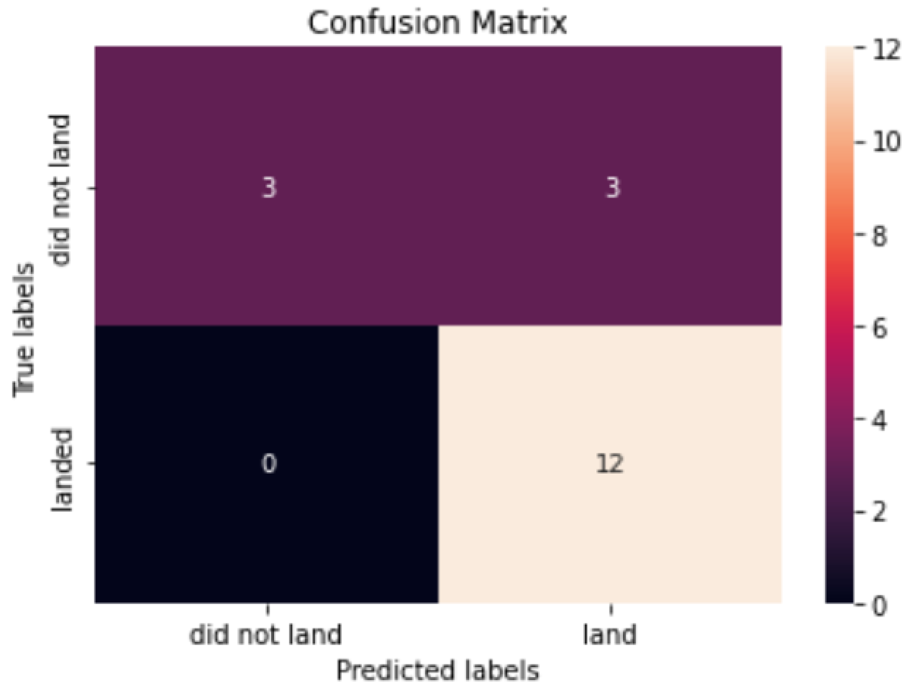
Predictive analysis (classification)

- **Logistic regression model:**

- **tuning parameters and accuracy**
- **Testing data accuracy**

```
Logistic regression model details with grid search parameters
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
best score, i.e. accuracy : 0.8464285714285713

Logistic regression test data accuracy: 0.8333333333333334
```

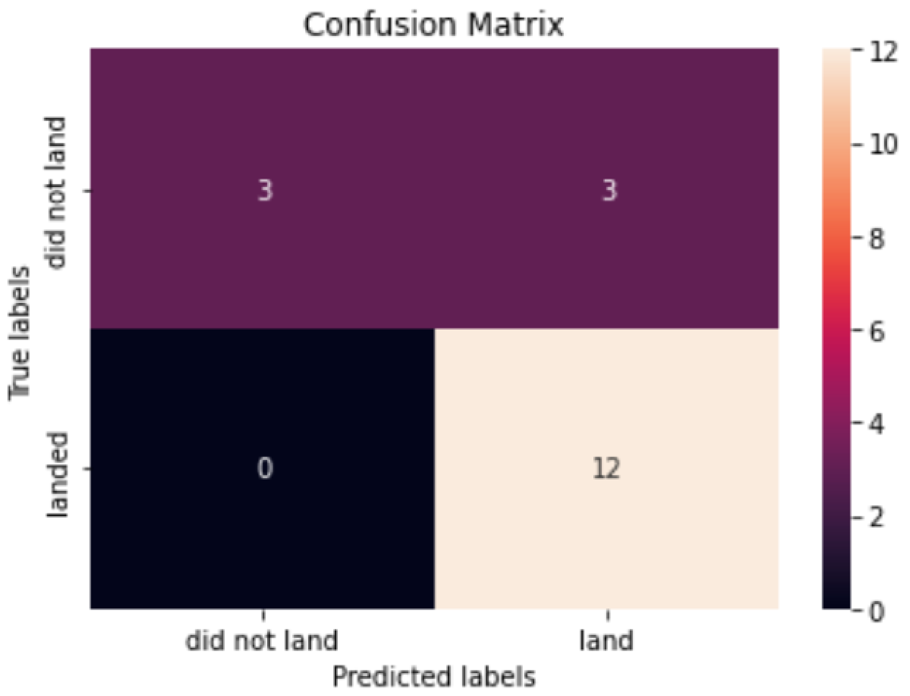


Results

Predictive analysis (classification)

- **Support vector machine model**
- **tuning parameters and accuracy**
- **Testing data accuracy**

```
Support vector machine model details with grid search parameters
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
Support vector machine model test data accuracy: 0.8333333333333334
```

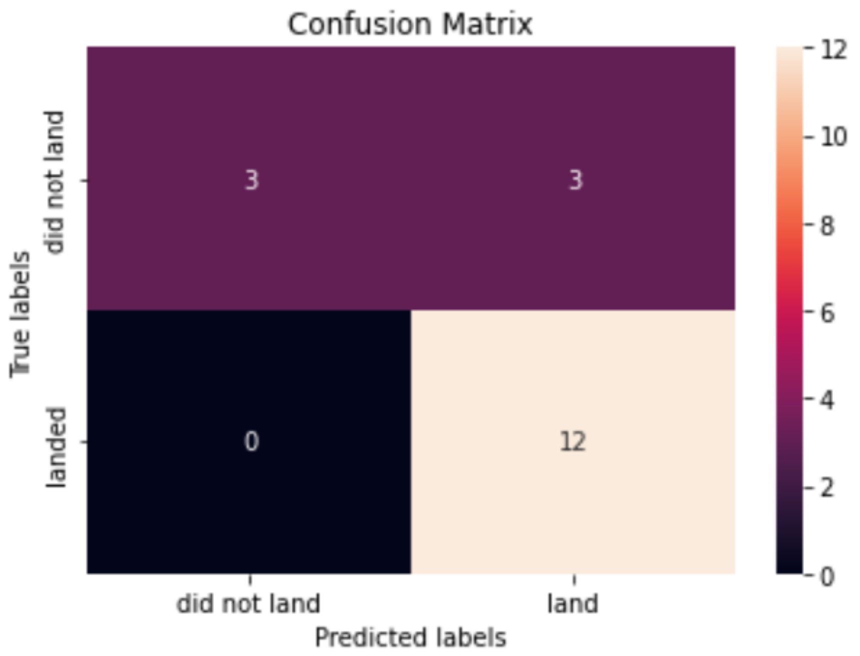


Results

Predictive analysis (classification)

- **Decision tree model**
- **tuning parameters and accuracy**
- **Testing data accuracy**

```
Decision tree model details with grid search parameters
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
accuracy : 0.8767857142857143
Decision tree model test data accuracy: 0.8333333333333334
```

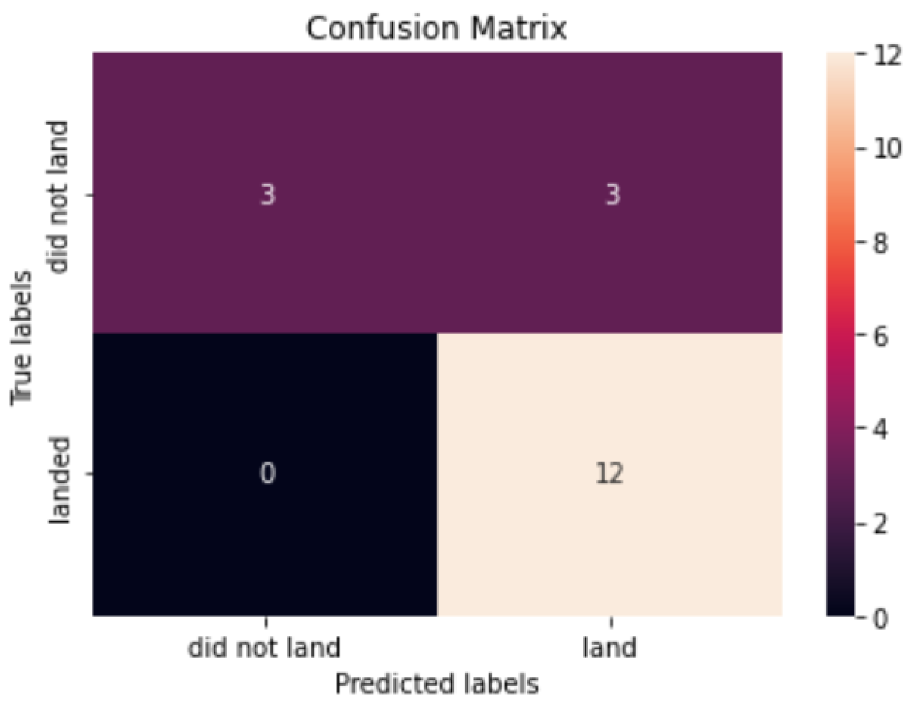


Results

Predictive analysis (classification)

- **K nearest neighbours model**
- **tuning parameters and accuracy**
- **Testing data accuracy**

```
K nearest neighbors model details with grid search parameters
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
K nearest neighbors model test data accuracy: 0.8333333333333334
```



Discussion

Visualization tools:

- payload mass vs flight number: We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- as flight number increases there is high probability that landing would be succesfull, especially after FN = 35.
- CCAFS SLC 40 booster has high probability of successfull landing after flight number around 75
- VAFB SLC 4E has high probability of successfull landing overall, specifically after 20
- overall all launch sites have successfull landing with payload masses higher than 8000 kg. Especially VAFB SLC 4E site has exact payload mass around 9500 kg that always landed successfully
- success rate vs orbit: orbits ES-L1, GTO, LEO, MEO has highest success rate, slight decrease in PO orbit.
- orbit vs Flight number: overall doesn't have significant correlation
- orbit vs payload mass: this correlation doesn't have a big story line except regardless of orbit type, successful landing is observed with payload masses over 8000 kg
- success rate vs years: mainly success rate is in increasing tendency after 2013 till 2020

Dashboard:

- according to the launch site success rates: highest rate is observed in KSC LC-39A (76.9%) followed by CCAFS LC-40 (73.1%)

Discussion

Site locations with Folium mapping:

- launches already conducted are distributed into two regions: 10 for the locations in California state, 46 for state Florida
- according to the visual data - Launch site KSC LC-39a has more successfull launches than any other ones.

Prediction models:

- practically all models have same test data accuracy (0.833) and same confusion matrix distribution. However best grid search cv tuning best scores are held by Decision tree model (0.876).
- The other two models : svm and logistic regression have almost tuning accuracy around 0.84.

Conclusion

- Successful landing possibilities were found via three ML models such as logistic regression, support vector machine and decision tree.
- Successful landing possibility, practically, is the same for all models : 0.833 (1.0 is the highest confidence of successful landing).

ML model name	Tuned best hyper parameters for Grid Search CV	Tuning accuracy's best score	Test data accuracy
Logistic Regression	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	0.8464	0.8333
Support Vector Machine	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}	0.8482	0.8333
Decision Tree	{'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}	0.8767	0.8333