IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bruno Gratival
March, 2023

Outline

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

Executive Summary

- This is a data science project developed for an aerospace systems manufacturer (SpaceX) that seeks to determine the possibility of this company to reuse the first stage (one of the parts of a space launcher) of its rockets.
- Rather than using rocket science to determine whether the first stage will land successfully, we train a machine learning model (classification algorithm) based on public data to predict whether SpaceX will reuse the first stage. Several data science techniques were used, from data collection and wrangling to exploratory and predictive analysis.
- All trained and tested models presented accuracy levels close to 0.83 when classifying the classes, in this case, in predicting the result of the first stage landing. These are good results that reinforce the feasibility of the data sets used and the application of the adopted methodology.

Introduction

The commercial space age is here and companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX.

SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space.

The launch of the first Falcon 9 v1.1

Introduction

One reason SpaceX can do this is the rocket launches are relatively inexpensive. 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.

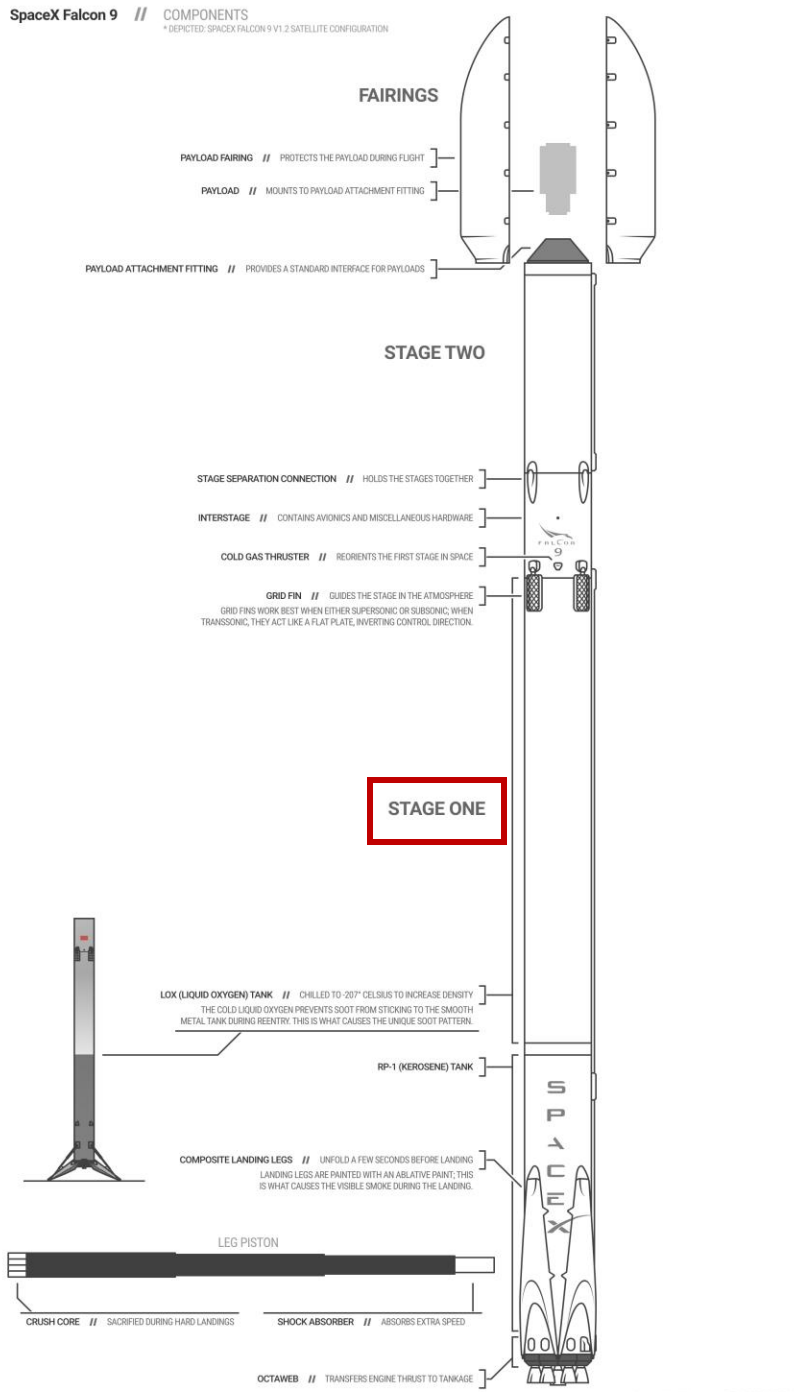
Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Falcon 9's first stage successfully landing for the first time.

Introduction

Spaces X's Falcon 9 launch like regular rockets. The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. This stage does most of the work and is much larger than the second stage. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

This project gathered data about SpaceX and the Falcon 9 rocket to apply data science techniques around this problem: **is it possible to predict if SpaceX will reuse the first stage?**



Section 1

Methodology

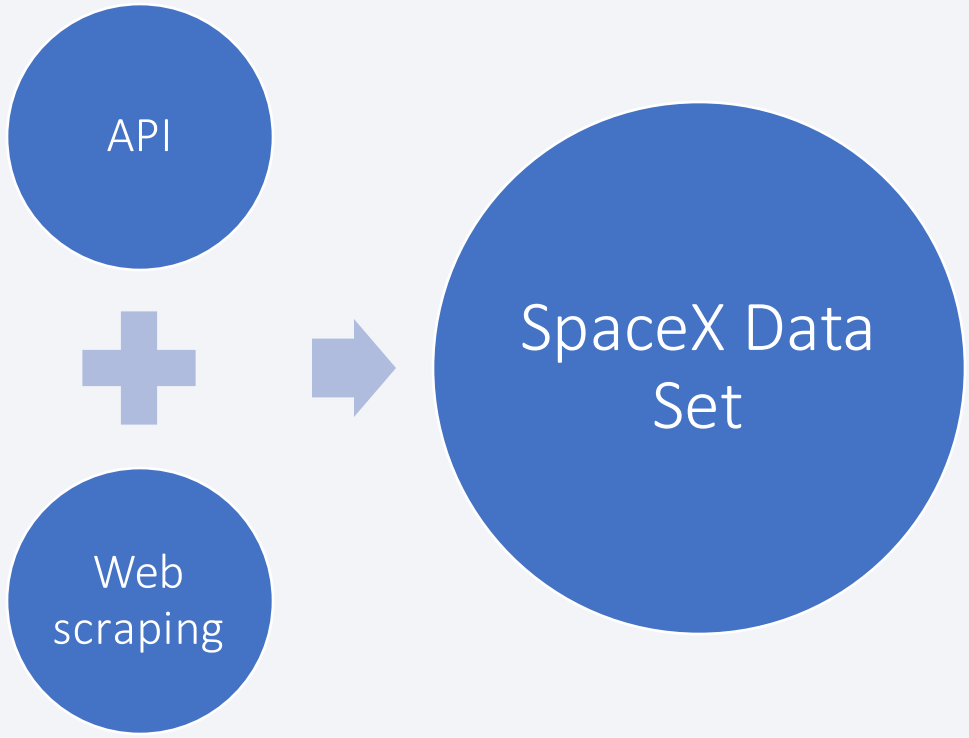
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

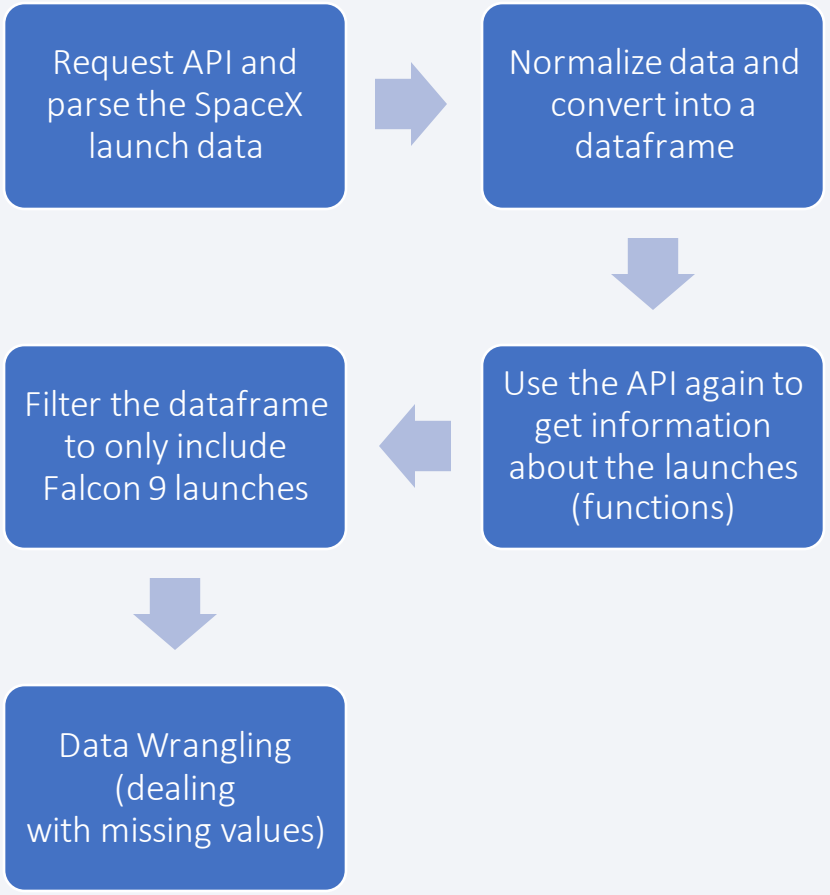
Data Collection

- The main data set was collected from an API (SpaceX REST API) and additional data were extracted from HTML tables (Wikipedia) using web scraping technics.



Data Collection – SpaceX API

- The SpaceX launch data was gathered from an specifically API: SPACEX REST API.
- This API provides data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- This is data that can be used to predict whether SpaceX will attempt to land a rocket or not.



Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Dataframe screenshot

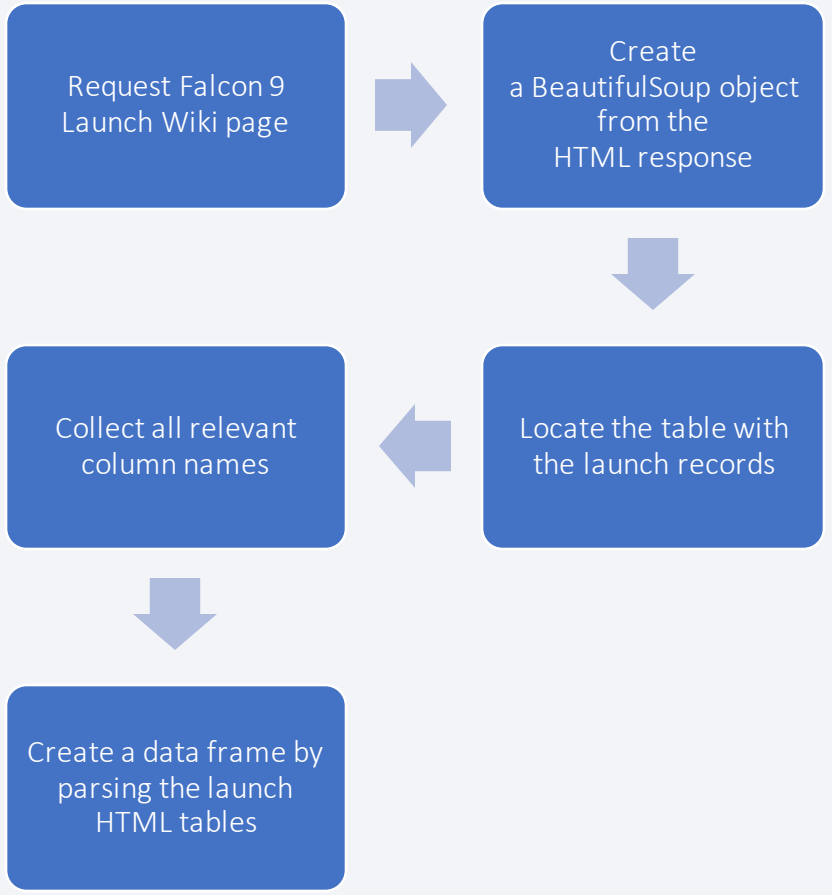
Notebook link

Data Collection - Scraping

- A Wikipedia page (HTML tables) that contain valuable Falcon 9 launch records was also used in data collection.
- This source contains data such as booster version, payload mass and orbit.

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								

Table from Wiki page screenshot



Notebook link

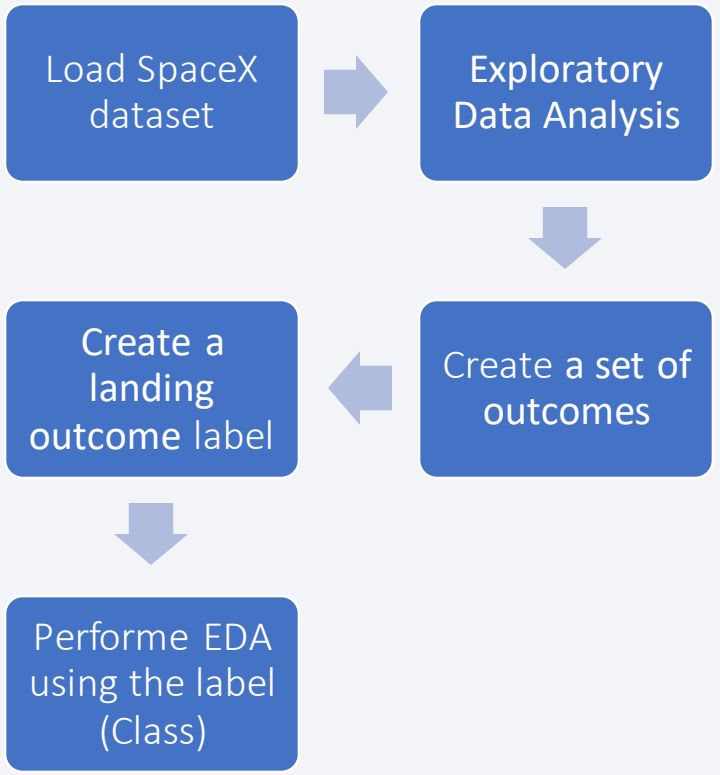
Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. The main objective in this process was to convert those outcomes into Training Labels (Class*):

[1] means the booster successfully landed;

[0] means it was unsuccessful.

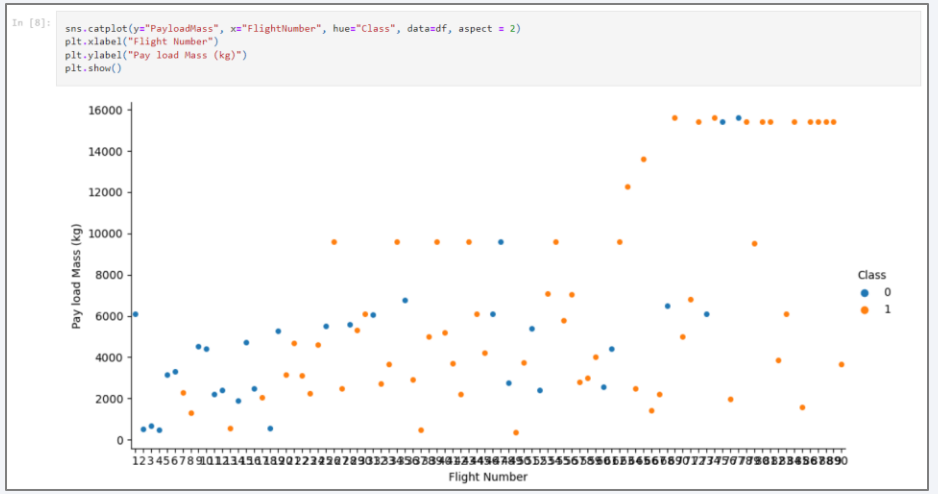
Outcomes	Class*	Definition
True ASDS	1	The mission outcome was successfully landed on a drone ship
None None	0	Failure to land
True RTLS	1	The mission outcome was successfully landed to a ground pad
False ASDS	0	The mission outcome was unsuccessfully landed on a drone ship
True Ocean	1	The mission outcome was successfully landed to a specific region of the ocean
False Ocean	0	The mission outcome was unsuccessfully landed to a specific region of the ocean
None ASDS	0	Failure to land
False RTLS	0	The mission outcome was unsuccessfully landed to a ground pad



Notebook link

EDA with Data Visualization

- Visual methods were used to verify that the data can be used to automatically determine whether Falcon 9's second stage will land.
- Scatter plots made it possible to understand different attributes (such as number of flights or payload mass) and their relationship with the landing outcome.
- Bar graphs were used to compare values by category (such as orbit type) and also line graphs showing the evolution of results over time.
- Also, dummy variables were created with 0/1 values to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.



Scatter plot (notebook screenshot)

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Serial_B1050	Serial_B1051	Serial_B10
0	1.0	6104.959412	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	2.0	525.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
2	3.0	677.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
3	4.0	500.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	5.0	3170.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
85	86.0	15400.000000	2.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
86	87.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
87	88.0	15400.000000	6.0	1.0	1.0	1.0	5.0	5.0	0.0	0.0	...	0.0	0.0	0.0	1.0	
88	89.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
89	90.0	3681.000000	1.0	1.0	0.0	1.0	5.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	

Dummy variables (dataframe) screenshot

Notebook link

EDA with SQL

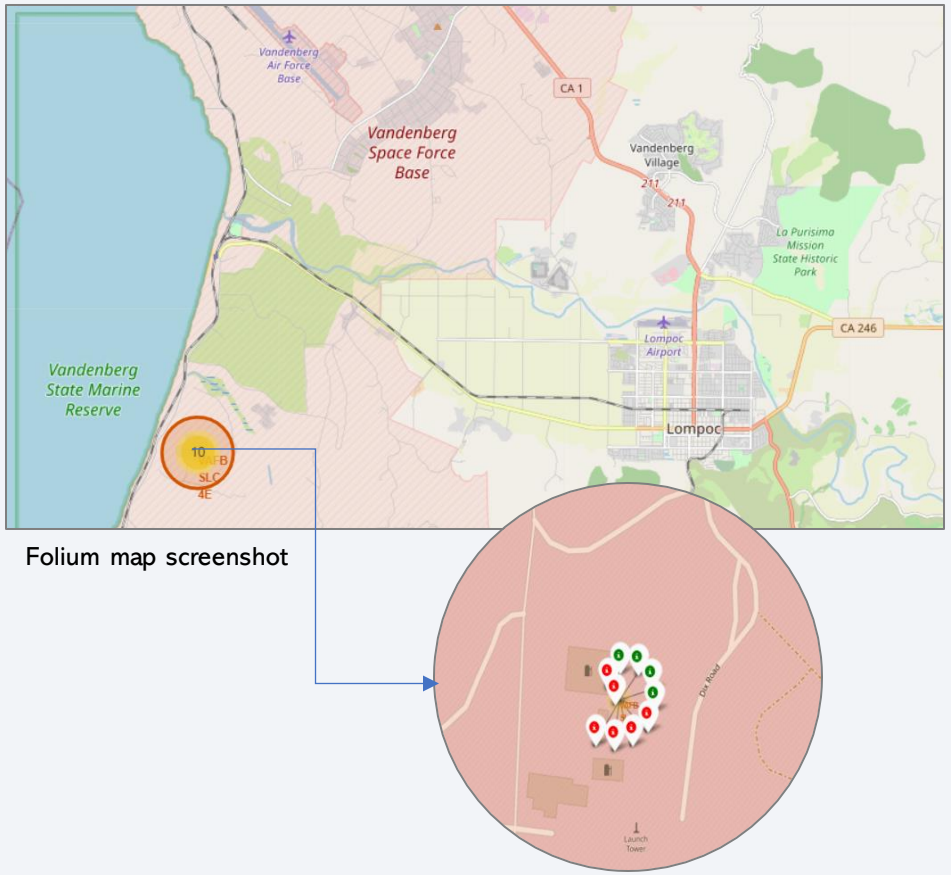
Some SQL queries were applied in the search for answers to specific questions about the launch data.

- Names of the unique launch sites in the space mission
- Records where launch sites names begin with specific strings
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Booster version, launch site and date (2015) for failure landing outcome in drone ship
- Count of successful landing outcomes between the date 04-06-2010 and 20-03-2017.

Notebook link

Build an Interactive Map with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Interactive visual analytics using Folium were performed to clarify this kind of question.
- Launch site geo and proximities were marked and explored on an interactive map to discover patterns. The idea is to be able to choose an optimal launch site location.

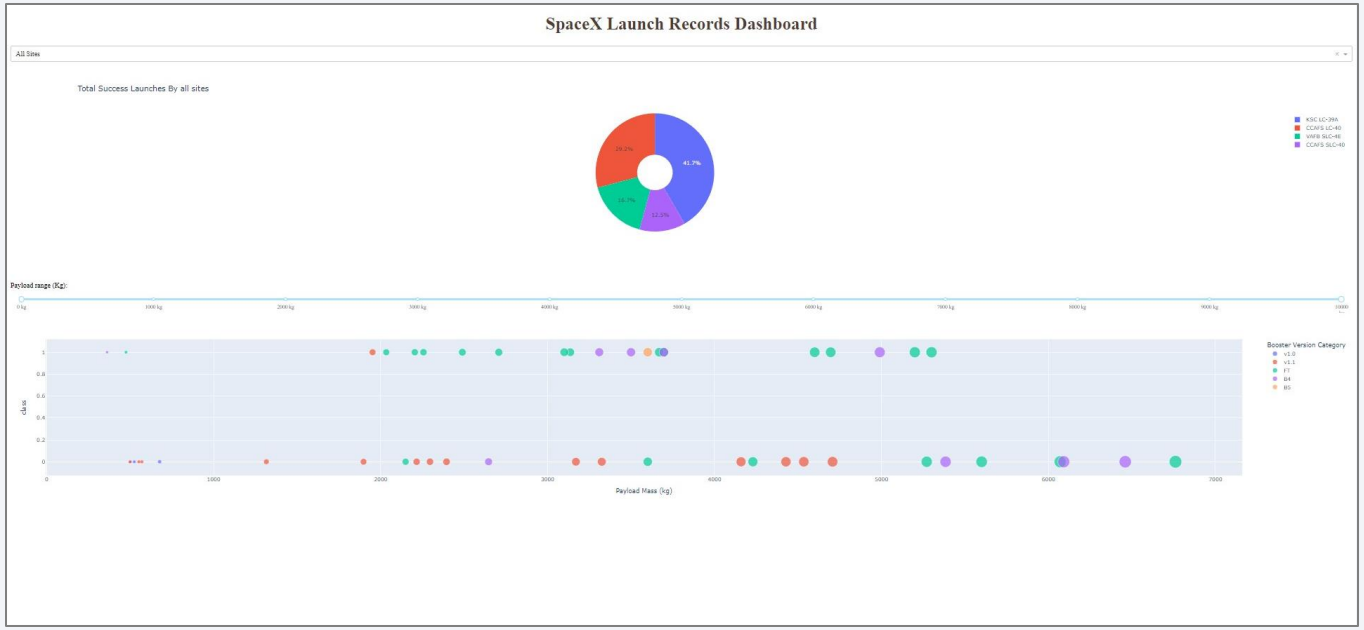


Folium map screenshot

Notebook link

Build a Dashboard with Plotly Dash

- A dashboard was built to find visual patterns faster and more effectively.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

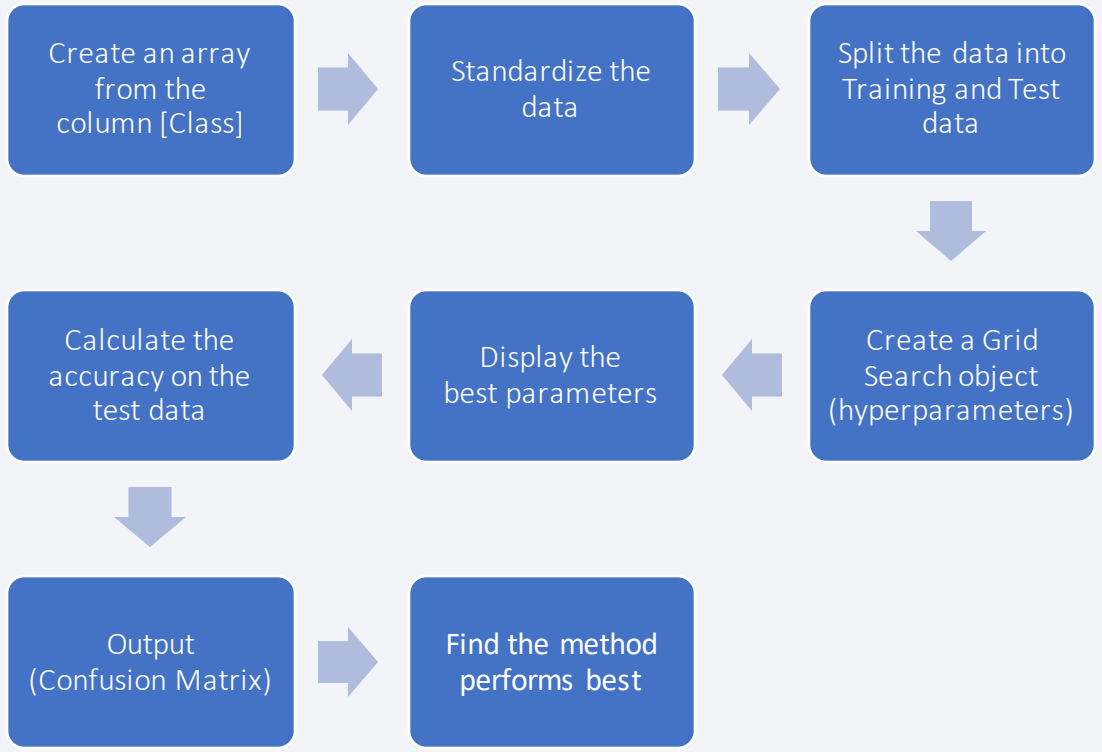


Dashboard screenshot

Notebook link

Predictive Analysis (Classification)

- A machine learning pipeline was created to predict if the first stage will land given all the data from the preceding steps. This stage begins with preprocessing and data training/ test splitting.
- For better algorithms performance, grid search technique was applied, allowing to determine the model with the best accuracy using the training data.



Notebook link

Predictive Analysis (Classification)

- Four classification algorithms were tested:
  - ✓ Logistic Regression
  - ✓ Support Vector Machine
  - ✓ Decision Tree
  - ✓ K-Nearest Neighbors
- Finally, the models' performance was evaluated using a confusion matrix.

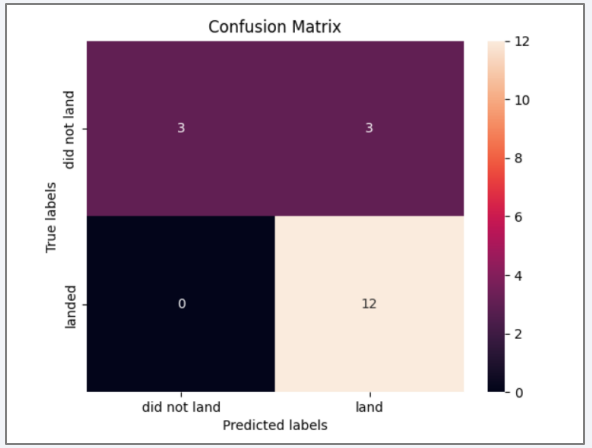
```
svm_cv = GridSearchCV(svm, parameters, cv = 10)
svm_cv.fit(X_train, Y_train)
```

```
GridSearchCV(cv=10, estimator=SVC(),
             param_grid={'C': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
       1.00000000e+03]),
                         'gamma': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
       1.00000000e+03]),
                         'kernel': ('linear', 'rbf', 'poly', 'rbf', 'sigmoid')})
```

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

Grid Search/ Hyperparameters code screenshot (SVM algorithm)



Confusion matrix screenshot

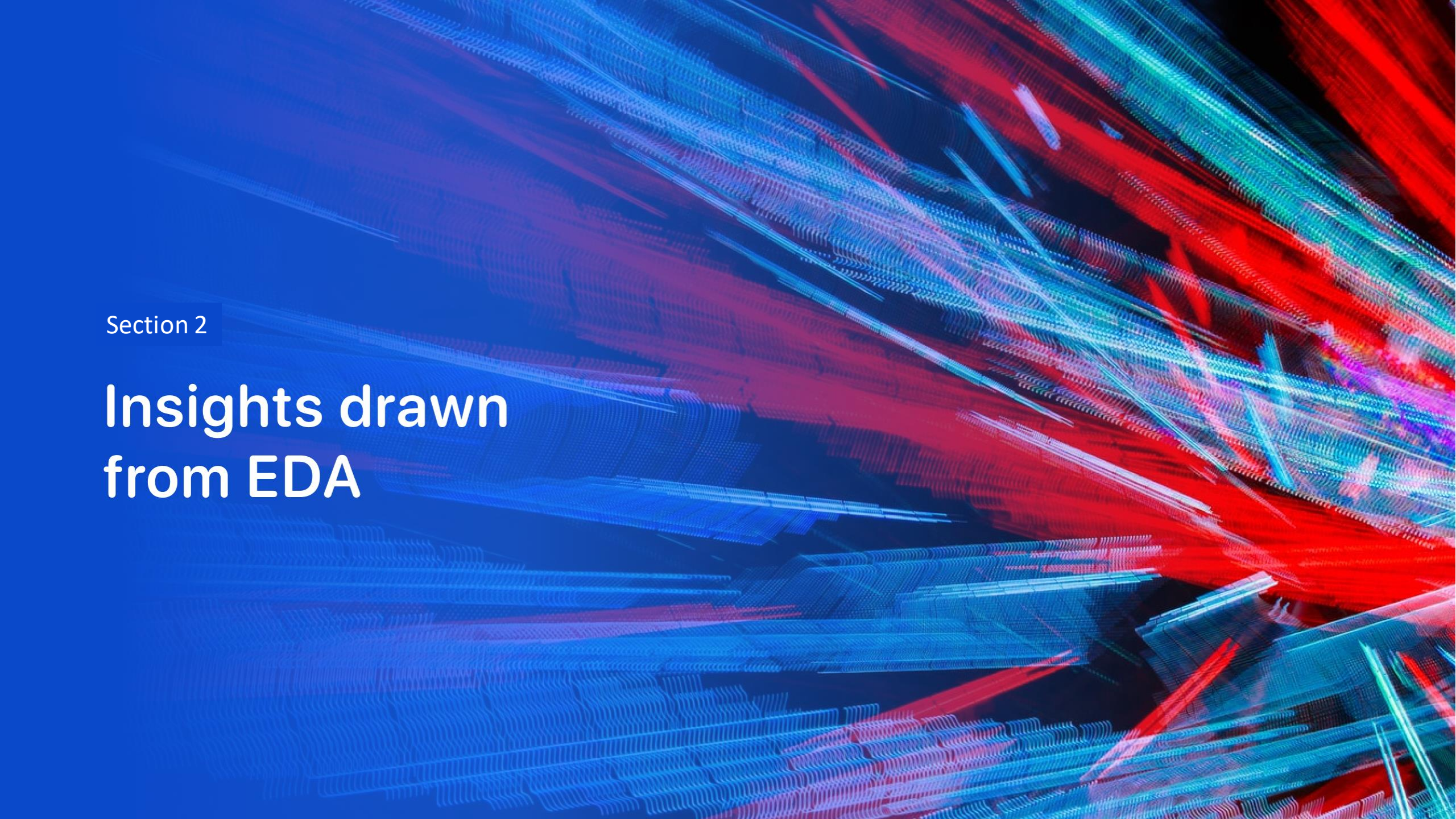
Notebook link

Results

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

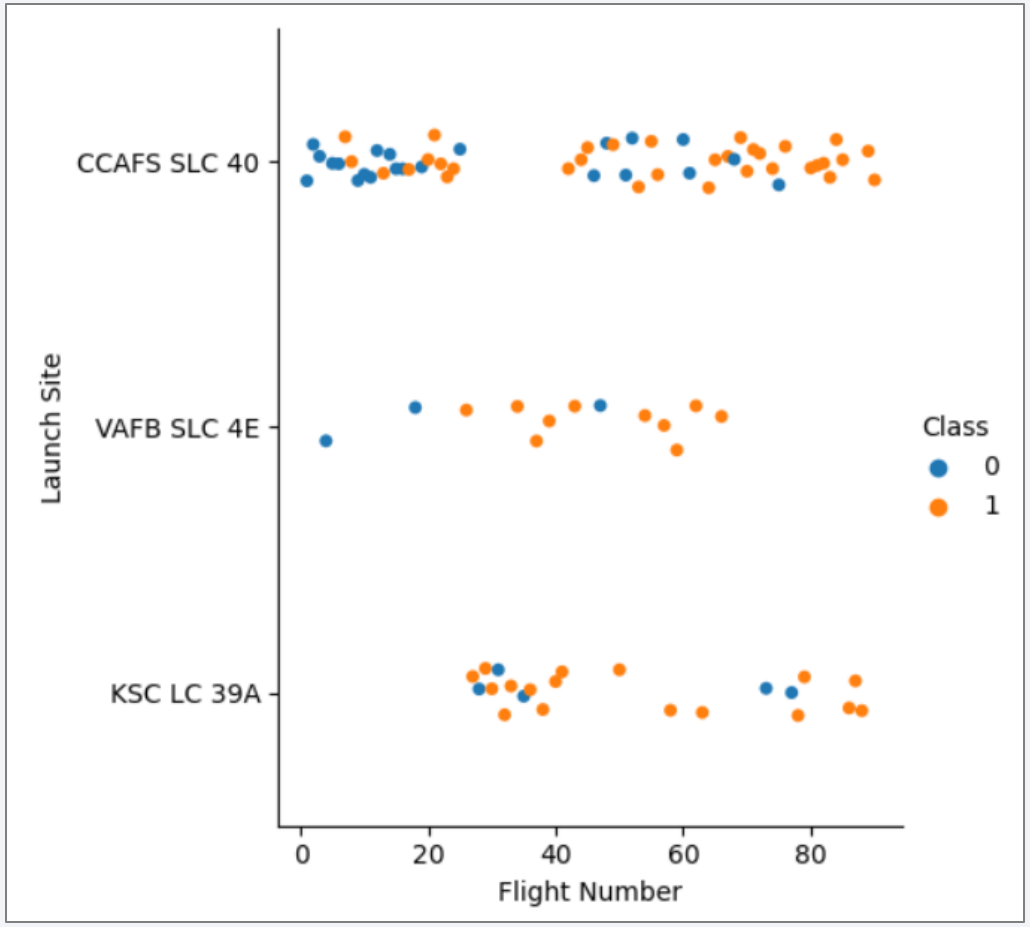
Section 2

Insights drawn from EDA



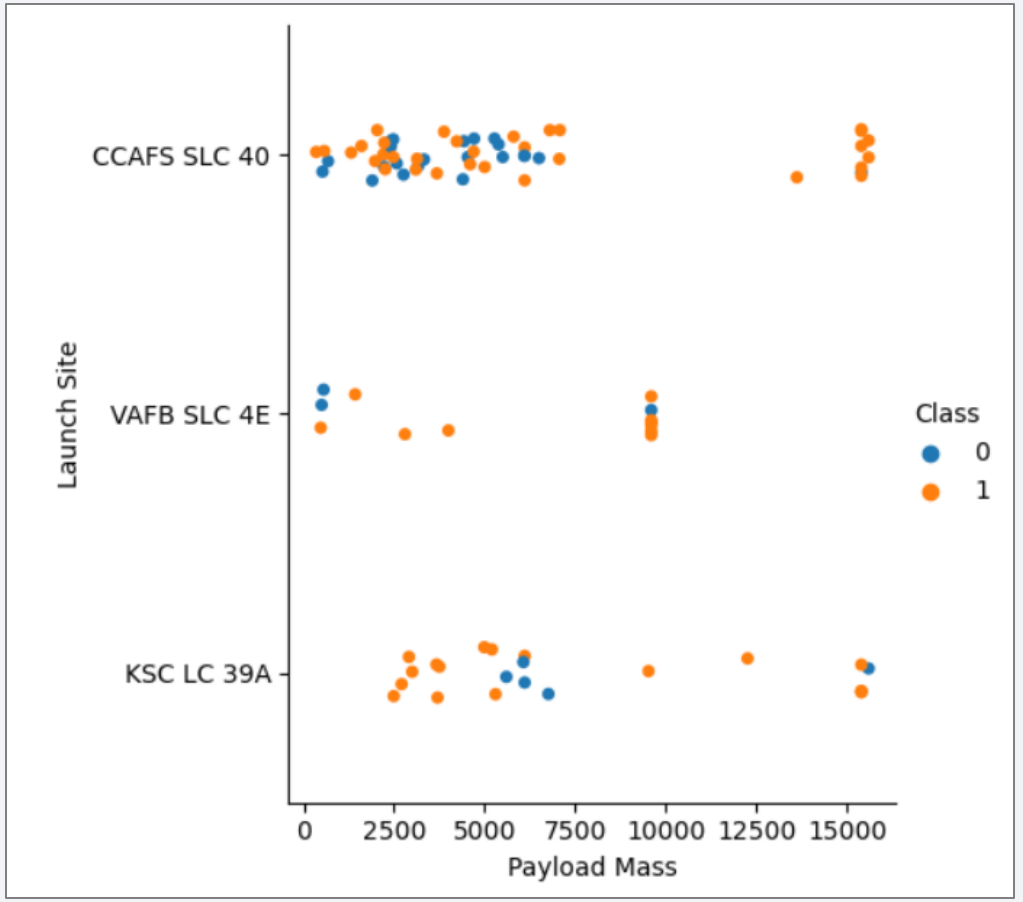
Flight Number vs. Launch Site

- At all launch sites, data shows that the first stage is more likely to land successfully as the number of flights increases.



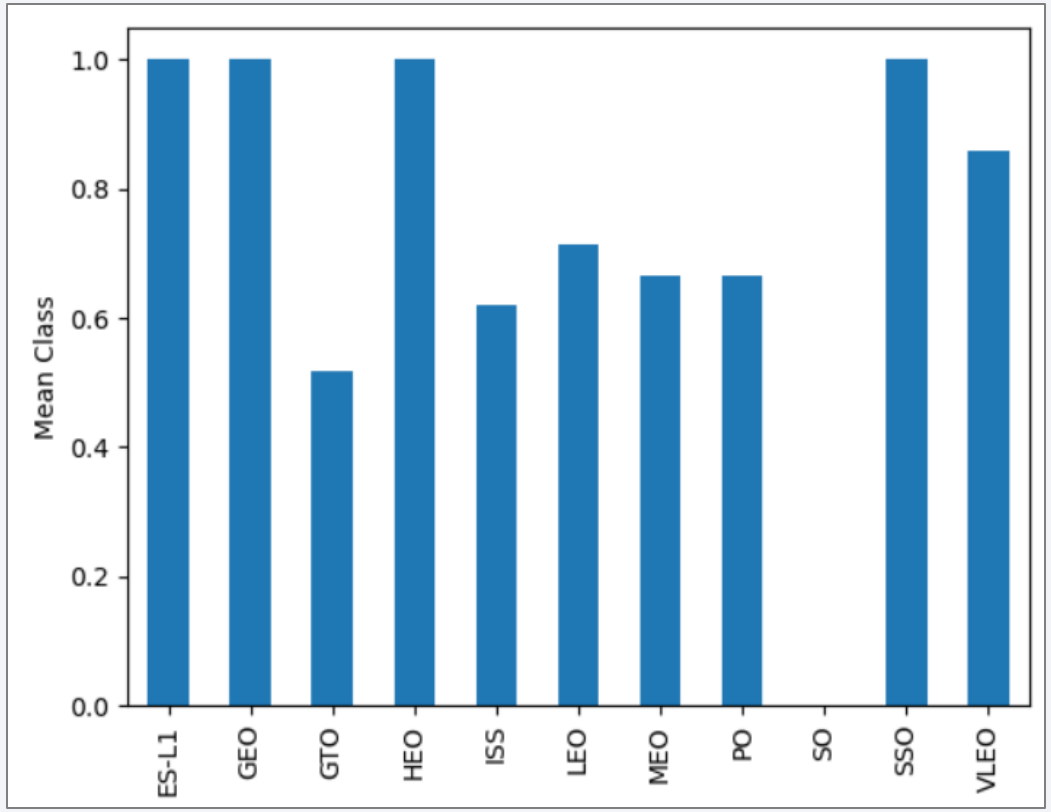
Payload vs. Launch Site

- Launch site CCAFS SLC 40 shows good landing results with heavier payloads (above 12,500 kg). In fact, considering payloads above 10,000 kg, we have only one bad landing, despite the low number of samples.
- At KSC LC 39A site, landings with lower payload mass (between 2,500 – 6,000 kg) were successful.
- For the VAFB SLC 4E launch site there are no rockets launched for heavy payload mass (greater than 10,000 kg).



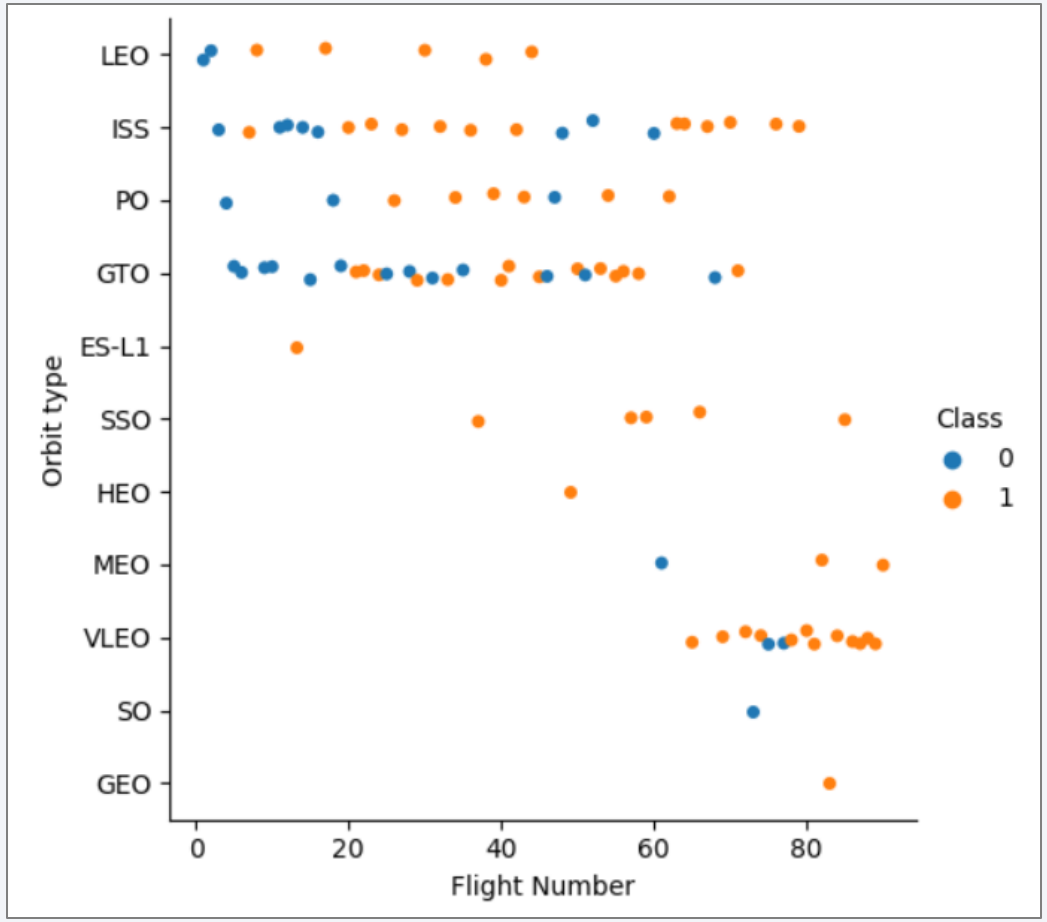
Success Rate vs. Orbit Type

- GTO type represents the orbit with the worst landing success rate.
- ES-L1, GEO, HEO and SSO have 100% successful landings.



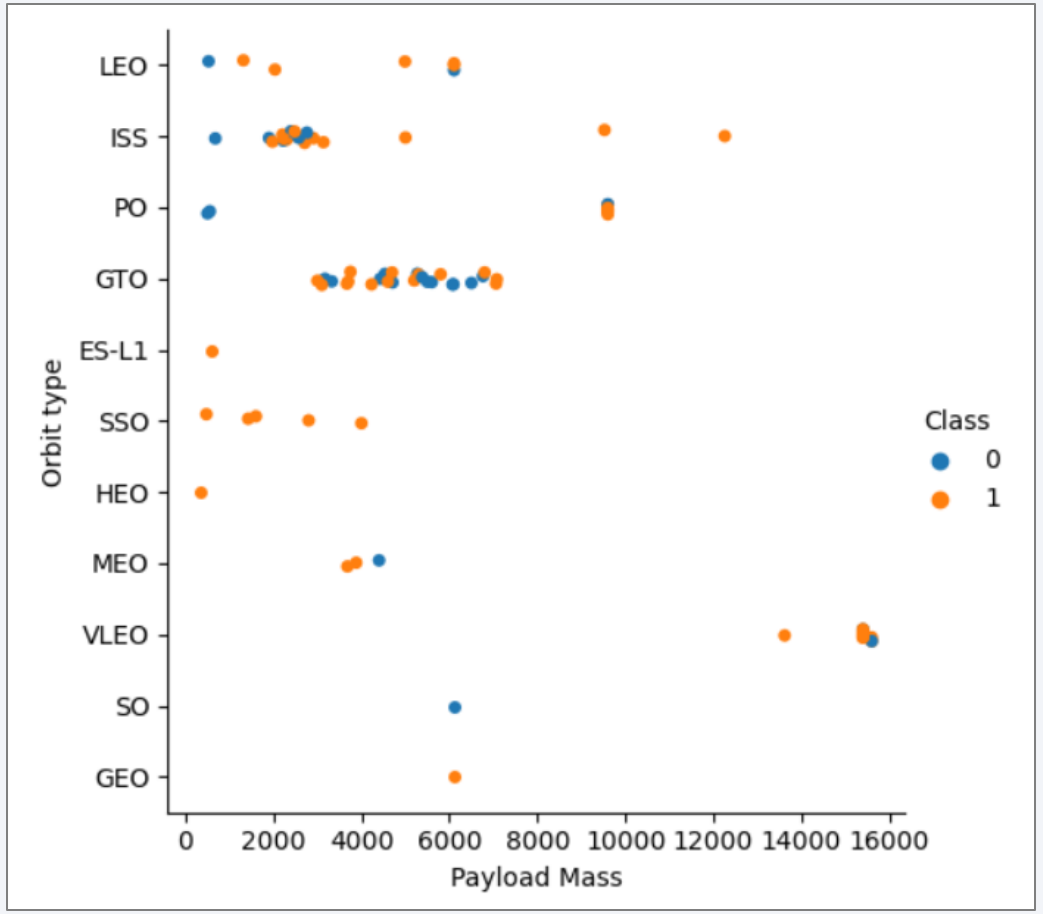
Flight Number vs. Orbit Type

- Again some data show that the first stage is more likely to return as the number of flights increases, for example in the LEO orbit.
- The GTO type is a notable exception showing no pattern in the behavior of its data.



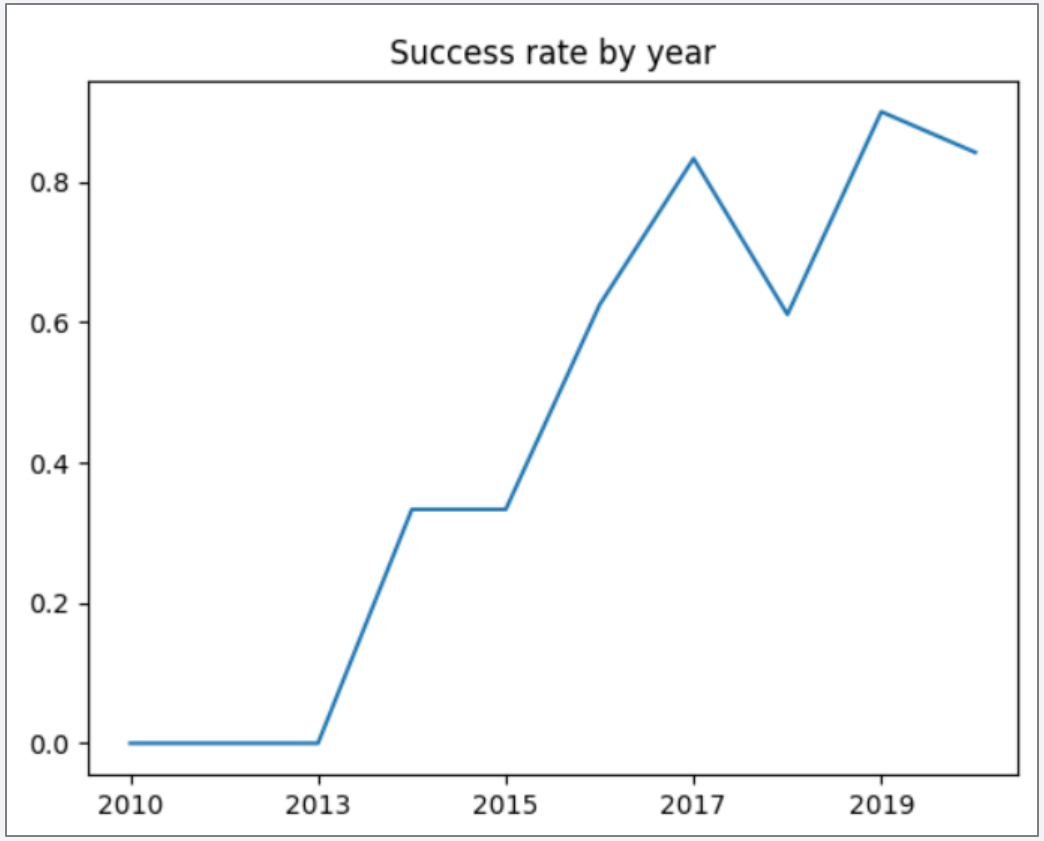
Payload vs. Orbit Type

- Despite not having a large data sample, the SSO orbit stands out for not presenting any unsuccessful landing outcome.
- ISS orbit seems to have achieved better results with heavier payloads.
- LEO orbit performed best with payload mass in the range of 1,500 – 6,000 kg.



Launch Success Yearly Trend

- Increased success rate since 2013, with the exception of a drop in 2018 that was recovered the following year, returning to levels close to 90%.



All Launch Site Names

- The table below shows the names of all launch sites registered in the database.

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

Launch Site Names Begin with 'CCA'

- In the table below, a demonstration of five database records that contain the characters 'CCA' in the name of the launch site.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- If we consider only payload carried by boosters launched by NASA (CRS), we have a total mass of 45,596 kg.

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- If we consider only payload carried by boosters launched by F9 v1.1, we have an average mass of 2,534 kg.

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- January 5th, 2017 is the date when the first successful landing outcome in ground pad was achieved.

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the names of the boosters which have successfully landed on drone ship and had payload mass greater than 4,000 but less than 6,000 kg.

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 numbers show only one mission with an unsuccessful outcome.

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

- These are the names of the twelve booster versions which have carried the maximum payload mass.

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

- In January and April 2015, there were two failed drone ship landings at CCAFS LC-40 sites.

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- Between the dates 04-06-2010 and 20-03-2017, there are the following numbers of successful landings.

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

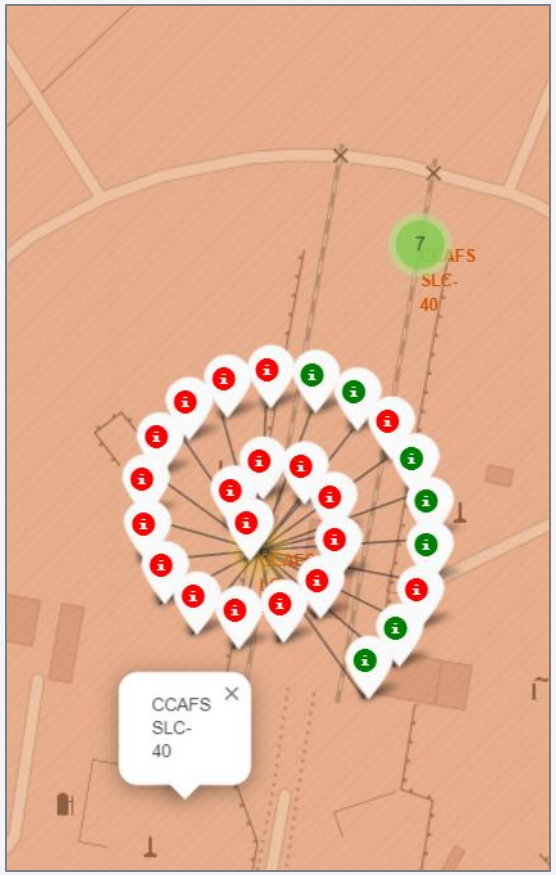
Launch Sites Location

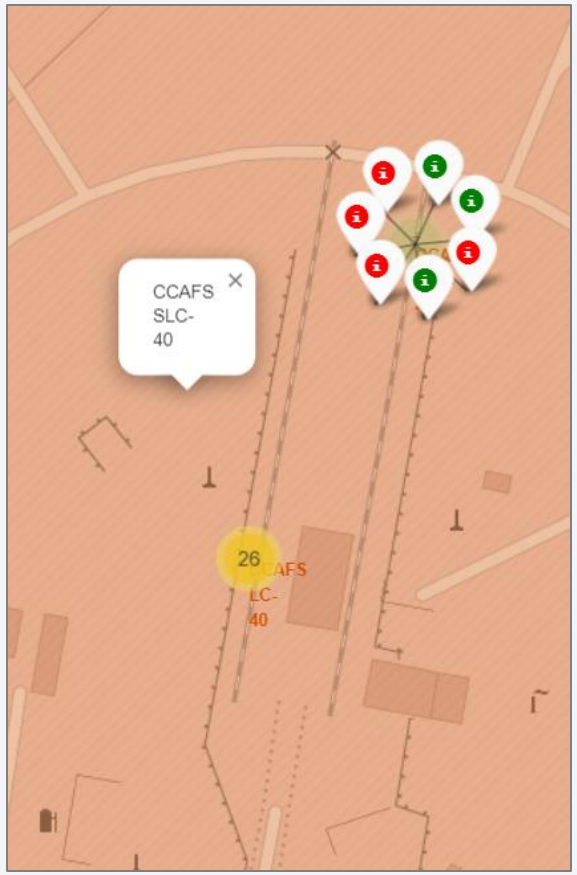
This map shows the launch sites location:

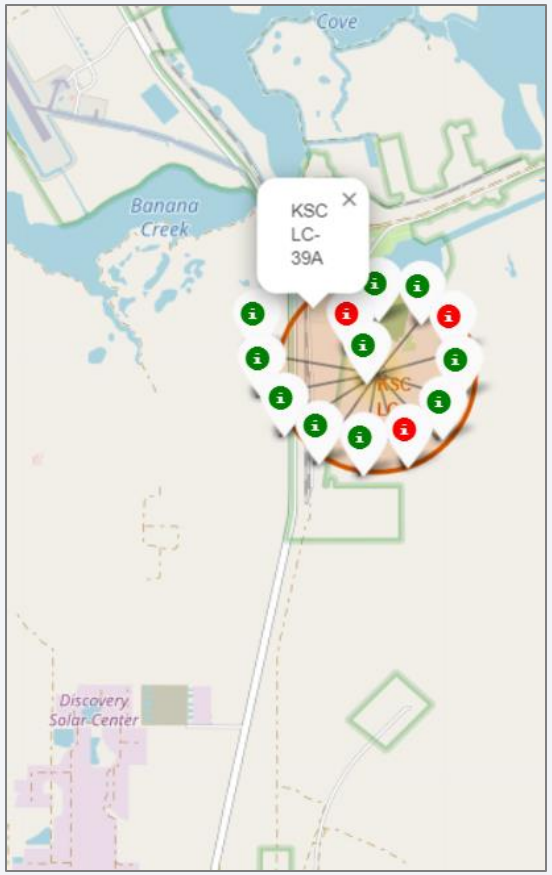
- Vandenberg Space Launch Complex, in California.
- Space Launch Complex (SLC-40 and LC-40) in Cape Canaveral Space Force Station, Florida.
- Kennedy Space Center Launch Complex 39 in Merritt Island , Florida.

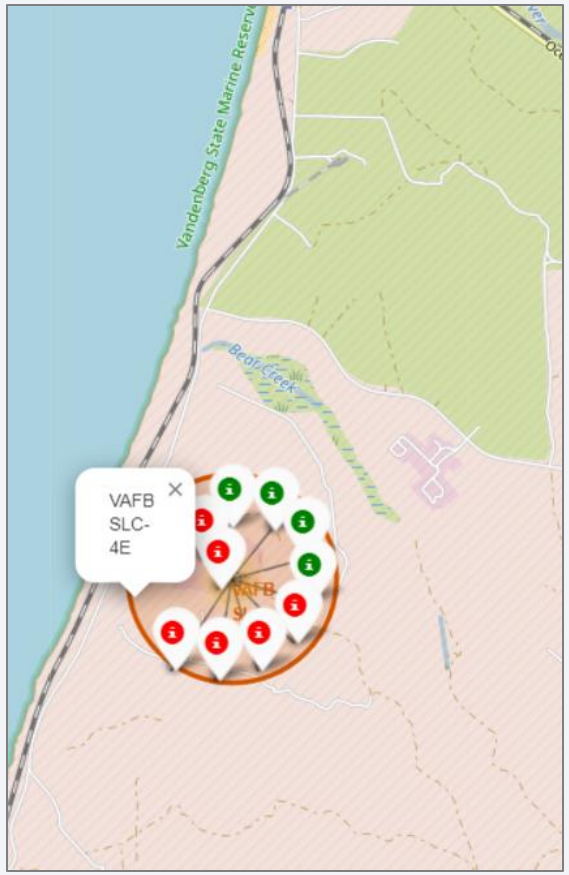
Launch outcomes

- The map shows all launches at each site with green markers for successful outcome and red markers for failed launches.





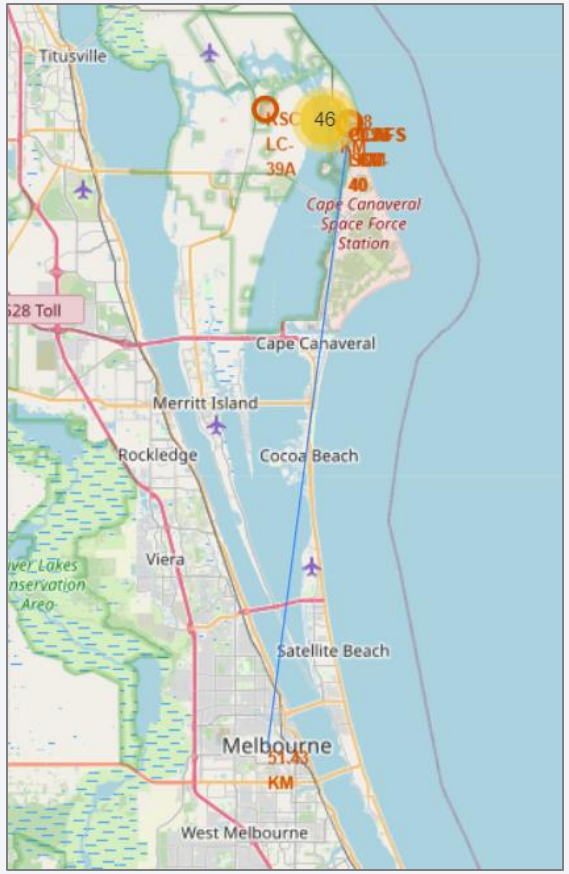




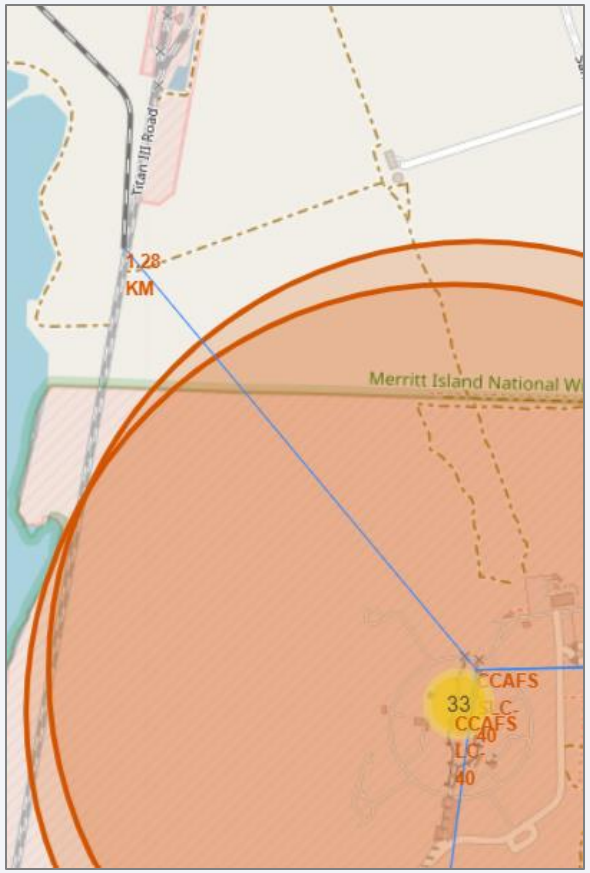
CCAFS SLC-40 launch site and its proximities

- Exploring the proximity of Cape Canaveral launch site, the distances and accesses to the installation are visualized.

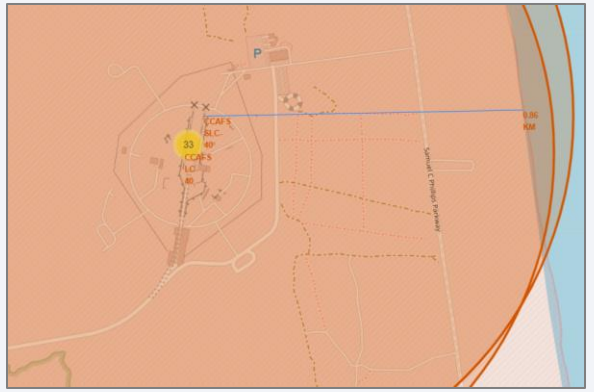
CCAFS - Melbourne: 51.43 km



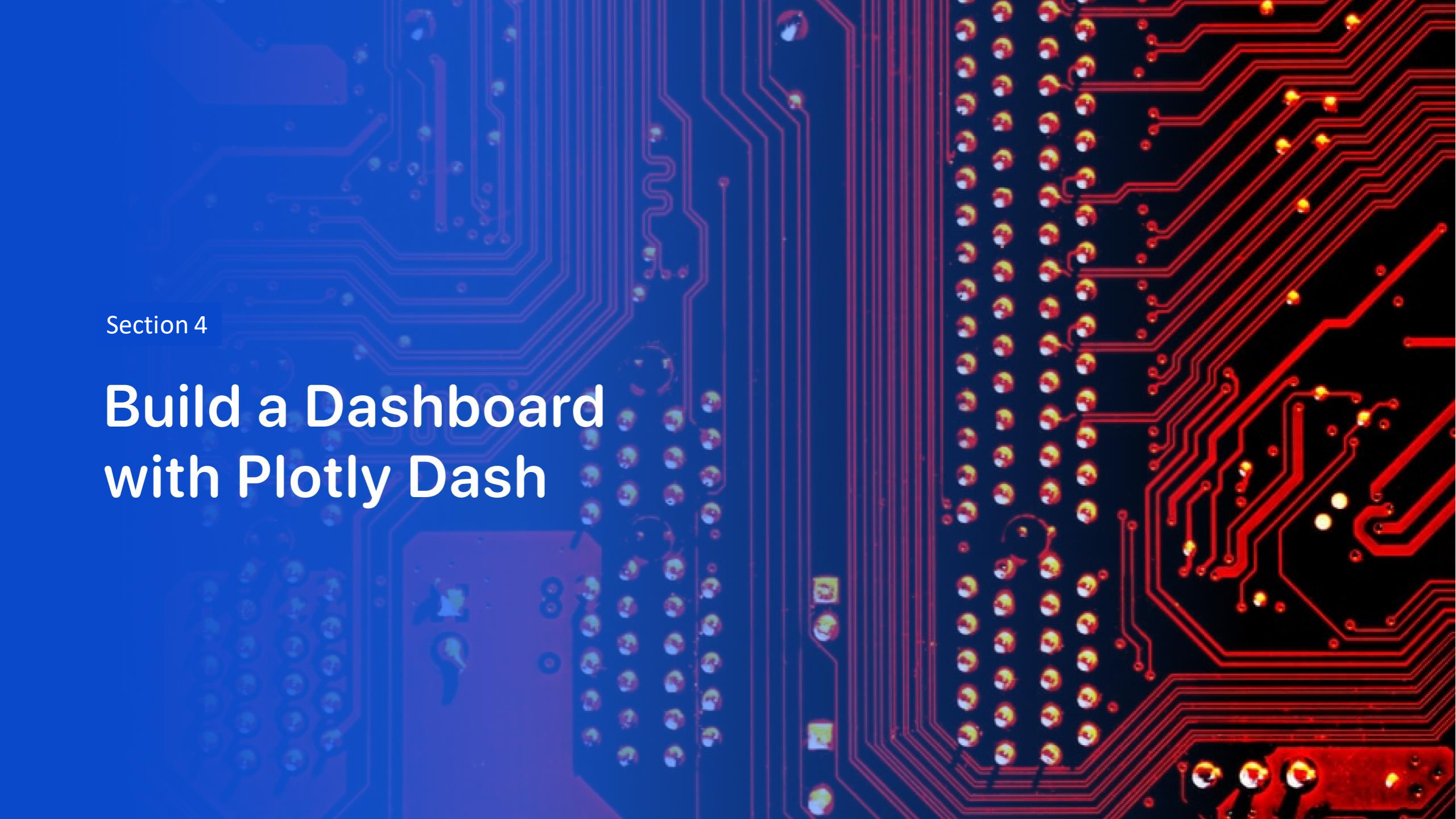
CCAFS - NASA Railroad: 1.28 km



CCAFS - coastline: 0.86 km



CCAFS - Samuel C Phillips Parkway: 0.58 km

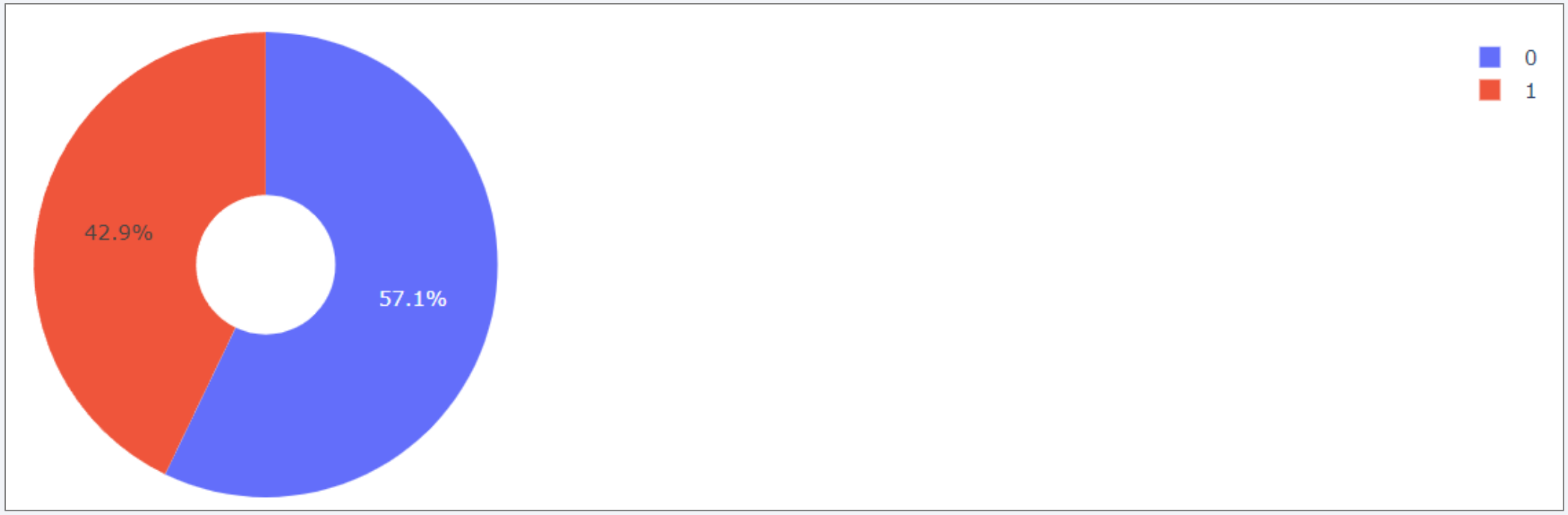

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches

- Analyzing landings from all sites, Kennedy Space Center stands out with the highest rate of successful launches (41.7%), followed by CCAFS LC-40 (29,2%).

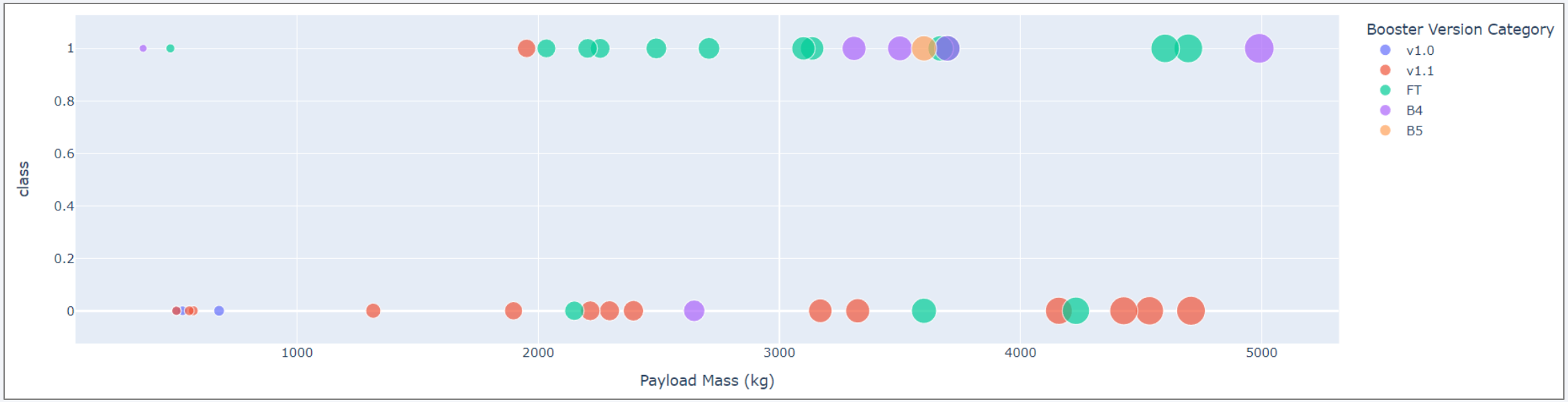
Total Success Launches for site CCAFS SLC-40

- Despite the low representation in the total of successful launches, Cape Canaveral Space Force Station SLC-40 has the best outcome with 42,9% success rate in the analysis by launch site.



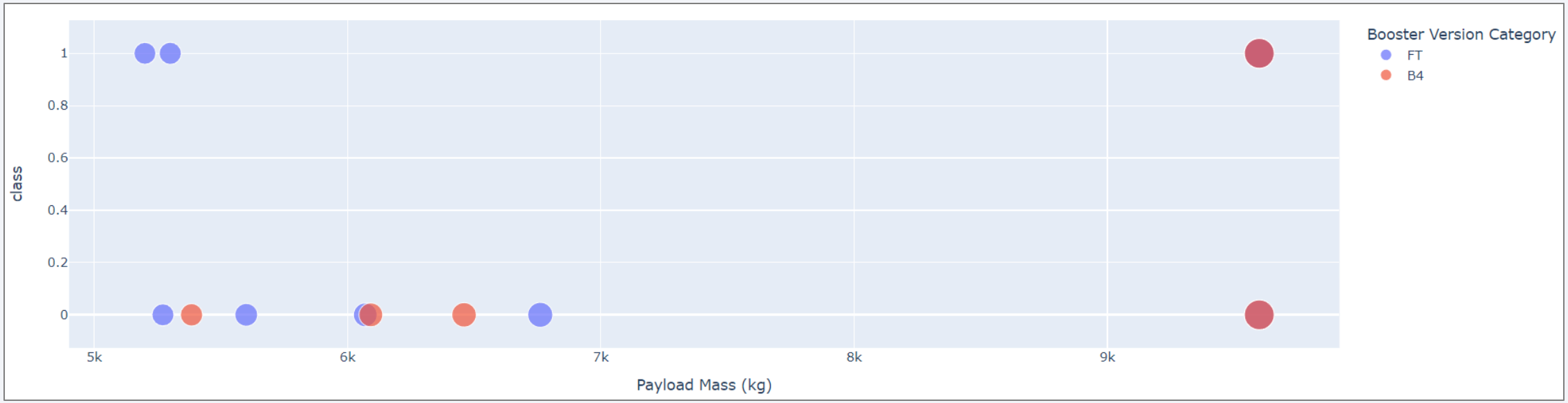
Payload Mass (≤ 5,000 kg) vs Launch Outcome

- Analyzing the outcome in a lower payload mass range, FT booster shows a higher number of successful launches, compared to the other version categories. Version 1.1 definitely achieved the worst results.



Payload Mass (> 5,000 kg) vs Launch Outcome

- In the upper payload mass range, only two versions of boosters are used (FT and B4), there are few launches, and their success rate is low.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

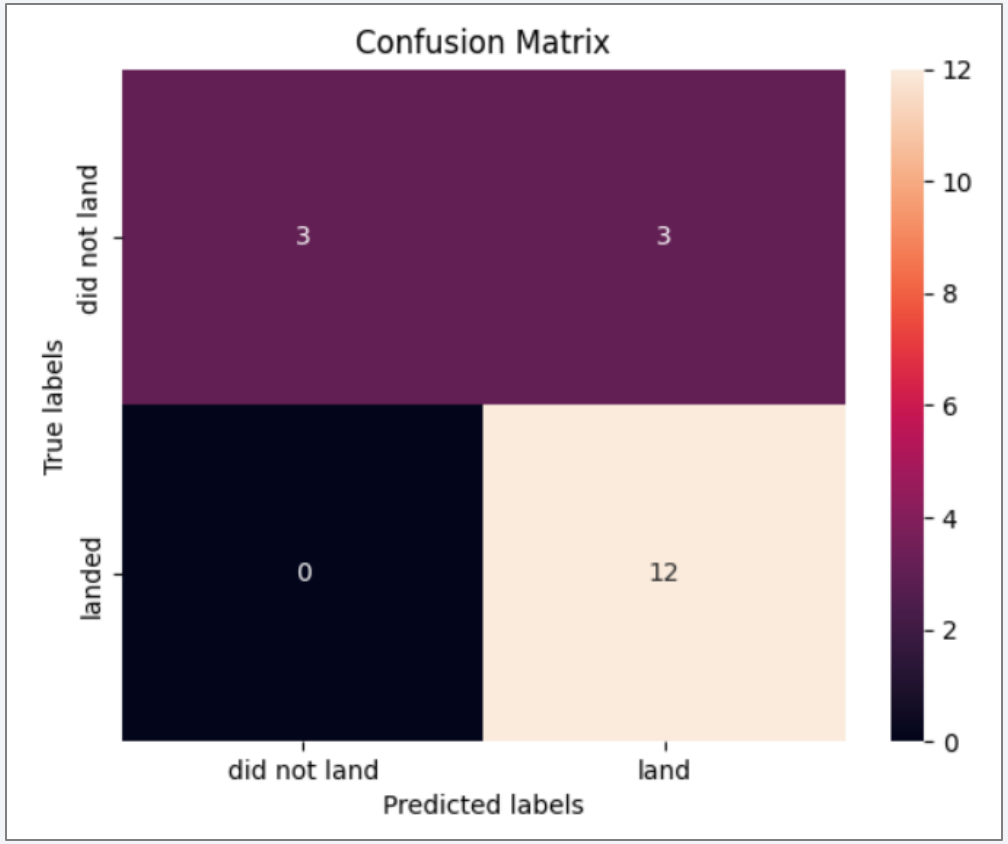
- The built classification models reached very similar results, mainly considering the score in the test data.
- As for the training/validation data, the highest classification accuracy was obtained with the Decision Tree classifier.

```
tree_cv = GridSearchCV(tree, parameters, cv = 10)
tree_cv.fit(X_train, Y_train)

GridSearchCV(cv=10, estimator=DecisionTreeClassifier(),
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [2, 4, 6, 8, 10, 12, 14, 16, 18],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10],
                         'splitter': ['best', 'random']})

print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples
_split': 10, 'splitter': 'random'}
accuracy : 0.8767857142857143
```

Decision tree hyperparameters

Confusion Matrix – Decision Tree Model

- Examining the confusion matrix, we see that the decision tree model can distinguish between different classes with good prediction results.
- The main point of improvement in the model would be the false positives (three predicted unsuccessful landings at actual values that were predicted as successful landings).



Conclusions

- It is concluded that the model can indeed be used to predict the outcome of the first stage landings considering the relatively high level of accuracy achieved with the machine learning process.
- Also noteworthy is the effectiveness of the entire methodology adopted, with a wide variety of data science techniques, largely contributing to accurate analysis and application of machine learning models.
- An interesting possibility of continuity is also identified, adding new data samples and considering other SpaceX rockets in the dataset, allowing a more generalist and less focused analysis from the point of view of aerospace activity.
- A new development could also start from the addition of new variables (attributes), making the whole exploratory data analysis process more complex, however, richer in nuances regarding the issue of reuse of the rocket's first stage.

Appendix

- GitHub repository:

https://github.com/brunogratival/Applied-Data-Science-Capstone

Thank you!