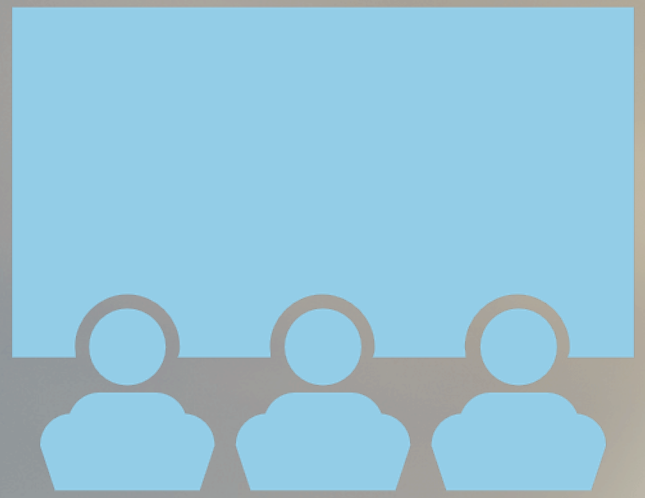
Analysis and Prediction of SpaceX First Stage Reentry performance

Data Science Capstone project

Alessandro Guidotti, Ph. D.

August 14, 2021

Outline

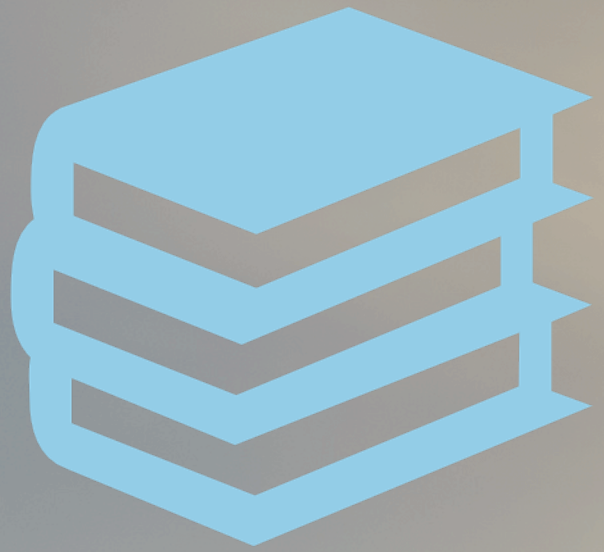


- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Through web-scraping (BeautifulSoup) and REST API (provided by SpaceX) we obtained a dataframe reporting the first stage reentry outcome based on various attributes
  - Orbit, payload mass, booster version, launch site, …
- Focusing on Falcon 9, the data was pre-processed
  - OneHot encoding for categorical variables, dataset filtering, standardization of the features
- Exploratory Data Analysis was applied to gain insight on the launch sites and correlation among features
  - Valuable insight has been gained in terms of the impact of the payload mass/orbit/experience on the success rate
- A Support Vector Machine classification model was optimized, providing an accuracy of 88% in the prediction
- **Thus, we are indeed able to predict, with a good accuracy, the outcome of the stage one reentry, i.e., of the launch cost**

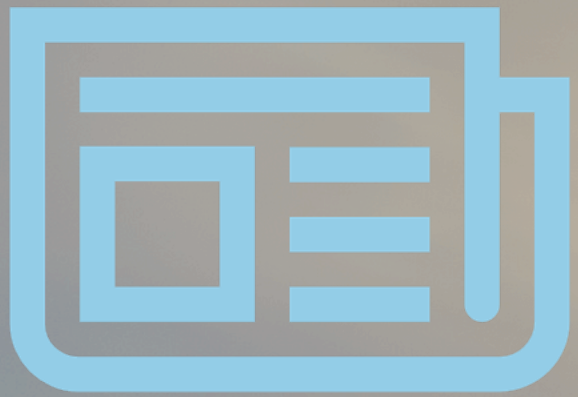
Introduction



- SpaceX advertises Falcon 9 launches at 62 M$
  - That's 100 M$ less compared to other launch providers!
- A key difference between SpaceX and other providers is that the first stage is reused for multiple launches
  - Fundamental impact on reducing the costs
- If we predict whether a first stage rocket will safely return to ground or not, we predict the overall cost of a launch

Can we actually predict the first stage return…?

Methodology

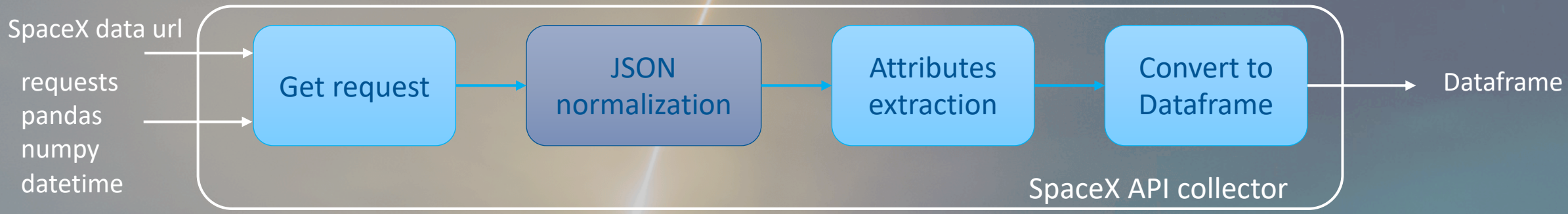


- Data collection methodology
  - Request to SpaceX API, with a dataframe obtained from the normalized JSON response
  - Extraction from the Wikipedia page on *List of Falcon 9 and Falcon Heavy launches*, with the dataframe obtained scraping the html through BeautifulSoup
- Data wrangling
  - Generation of the successful landing labels from the scraped outcomes
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - KNN, SVM, Decision Tree, Logistic Regression
  - Optimization of the hyperparameters with cross-validation
  - Performance evaluation: Jaccard index, F1-score, Log-Loss (LR only)

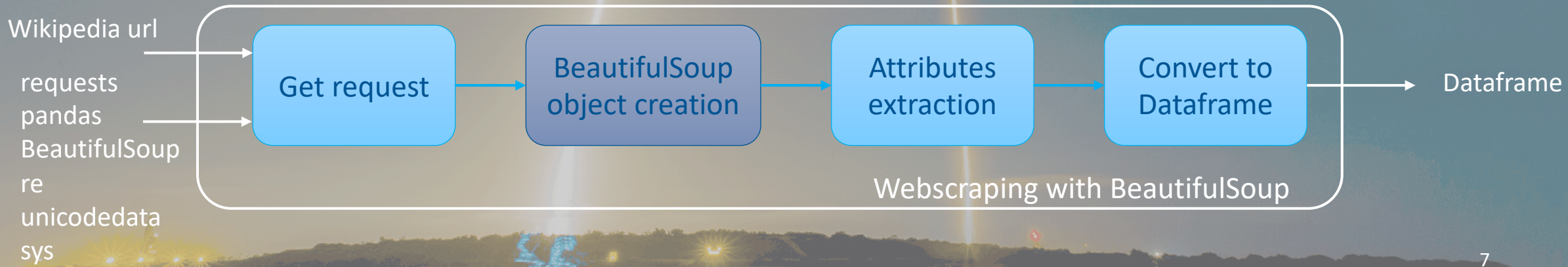
Methodology

Data collection

- Request to SpaceX API



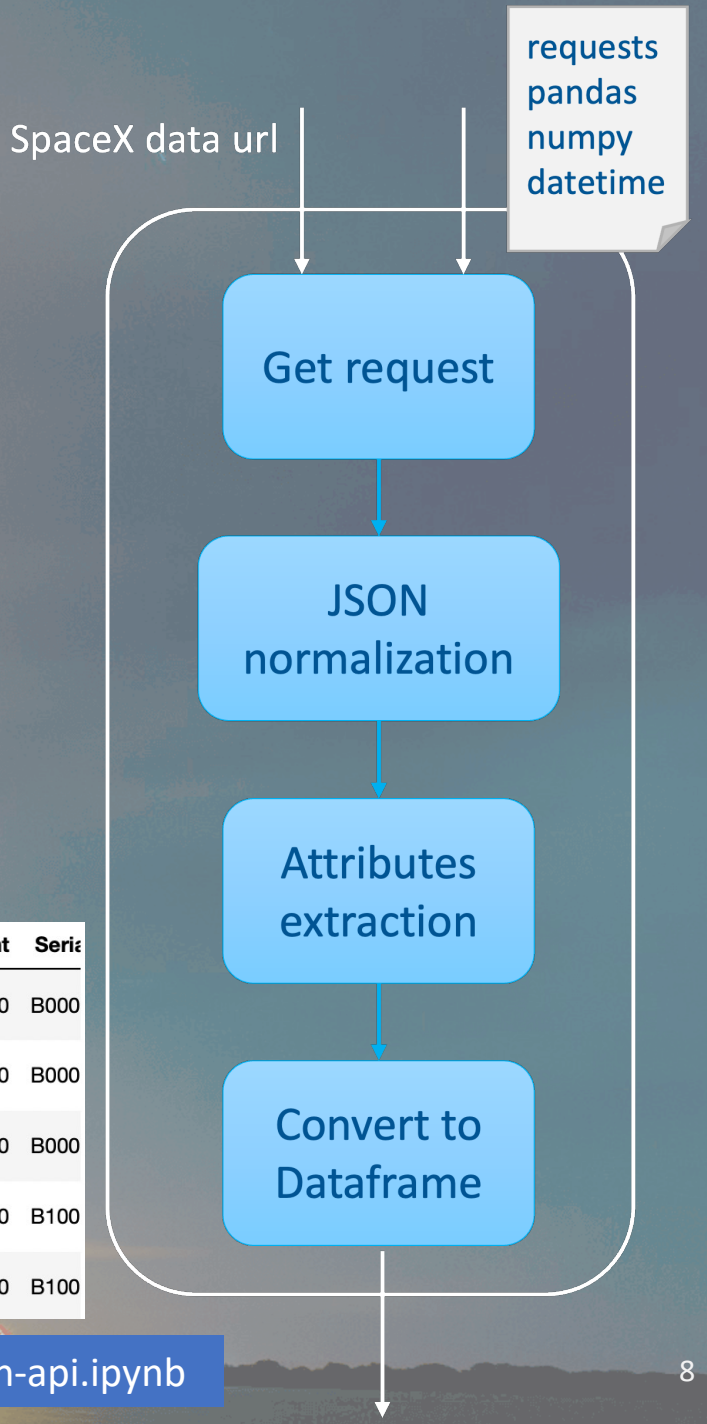
- Web scraping from Wikipedia (link)



Data collection: SpaceX API

- Get request: rocket launch data requested with the SpaceX REST API
  - https://api.spacexdata.com/v4/launches/past
- The JSON response was normalized into a Pandas Dataframe
  - Note: a static JSON response object was used as provided by IBM
- Attributes extraction
  - A list per attribute, to be used to generate the dataframe
  - Data wrangling: filter to keep only Falcon 9 and replace missing values

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Seria
1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B000
2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B000
3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B000
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B100
5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B100

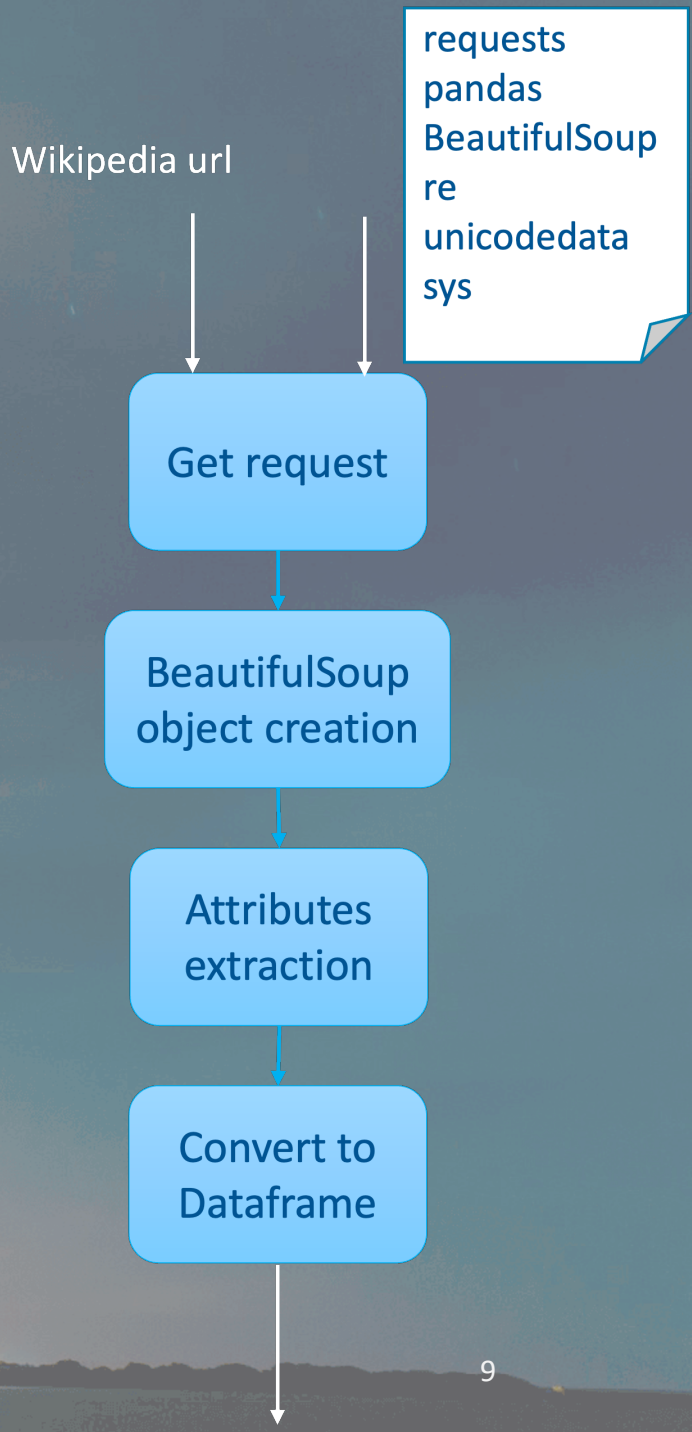


https://github.com/aleg83/applied_data_science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data collection: Web scraping

- Get request: html requested from Wikipedia “List of Falcon 9 and Falcon Heavy launches”
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The BeautifulSoup object is then analyzed to extract the relevant html tags
  - In this case, the html table rows with attribute “tr”
- Attributes extraction from the html table rows
  - A list per attribute, to be used to generate the dataframe

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success	F9 v1.0B0007.1	No attempt	1 March 2013	15:10



https://github.com/aleg83/applied_data_science/blob/main/jupyter-labs-webscraping.ipynb

Data wrangling

- Step 1 (SpaceX API only)
  - A filter was applied to only select Falcon 9 launchers
  - The missing values in the payload mass were replaced with the average
- Step 2 (both collection methods)
  - The target labels were obtained from the landing outcome and appended to the dataframe

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 None ASDS
6 False Ocean
7 False RTLS
```

0 (failure)
1 (success)

https://github.com/aleg83/applied_data_science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with data visualization

- Scatter plot to identify the combined impact of different features on the outcome
  - number of flights vs payload mass
  - number of flights vs launch site
  - payload mass vs launch site
  - number of flights vs orbit
  - payload mass vs launch site
- Bar plot to identify the impact of the orbit type on the success rate
- Line plot to show the average success rate as a function of the year

https://github.com/aleg83/applied_data_science/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Names of unique launch sites
- List of records with launch site starting with “CCA”
- Computation of the total payload mass carried by boosters launched by NASA (CRS)
- Computation of the average payload mass carried by F9 v1.1 boosters
- Obtain the date of the first successful landing outcome on ground pad
- List of the boosters names with a success in landing on a drone ship and payload mass between 4000 kg and 6000 kg
- List of the total number of successful and failure mission outcomes
- Subquery to list the names of the booster versions that have carried the maximum payload mass
- List of the records displaying the month names, failure landing outcomes on a drone ship, booster versions, and launch site in 2015
- Rank of the successful landing outcomes between June 4, 2010 and March 20, 2017 in descending order

https://github.com/aleg83/applied_data_science/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an interactive map with Folium

- Circles with marker
  - (example from IBM: NASA Johnson Space Center)
  - Launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-44E
    - To easily locate the launch sites on the map
  - Marker cluster reporting the successful/failed launches per site
    - Visual understanding of the success rate per launch site
  - Mouse position
    - To mark down the coordinates of locations of interest, see next point
  - Line with marker reporting the distance from the nearest railway, city, coastline, highway
    - Understand whether relevant locations are typically close or far away from the launch site

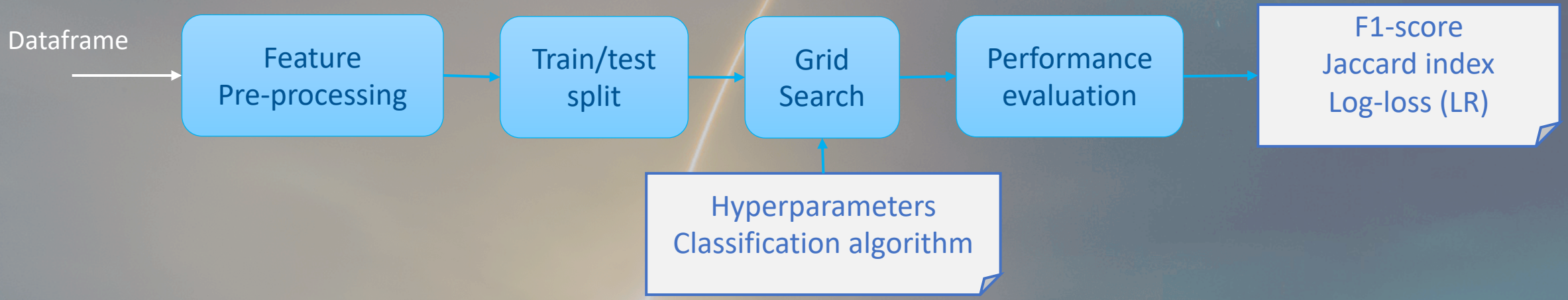
https://github.com/aleg83/applied_data_science/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart
  - Success count over all launch sites
  - Success percentage over all launches of all of the launch site
    - To visually understand which launch site has the best success rate
  - Success percentage per launch site
    - To visually understand the success rate of a single launch site
- Scatter plot
  - payload mass vs outcome based on the booster version category
  - The payload mass is tunable with a range slider
  - This plot helped in understanding the correlation between the mass of the launched payload and the outcome, as a function of the booster version

https://github.com/aleg83/applied_data_science/blob/main/spacex_dash_app.py

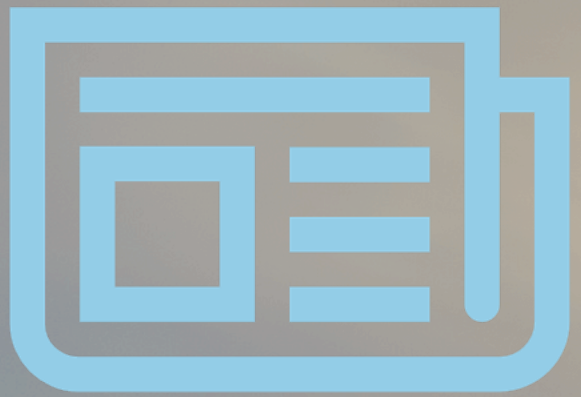
Predictive analysis (Classification)



- Feature pre-processing: standardization → $\hat{x} = (x - \mu)/\sigma$
- Train/test split: 20% of the observations retained for testing purposes
- Grid search
  - Performed for all classification algorithms: KNN, LR, SVM, DT
  - Hyperparameters defined depending on the specific model
- Each optimized model has been evaluated based on the F1-score and Jaccard index on the test set
  - Log-loss has also been used for the Logistic Regression model

https://github.com/aleg83/applied_data_science/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

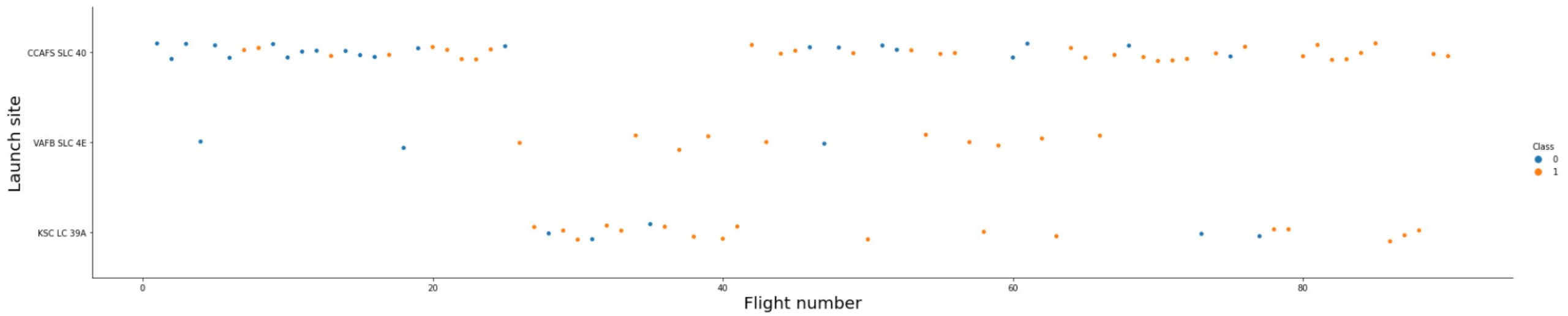


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


EDA with Visualization

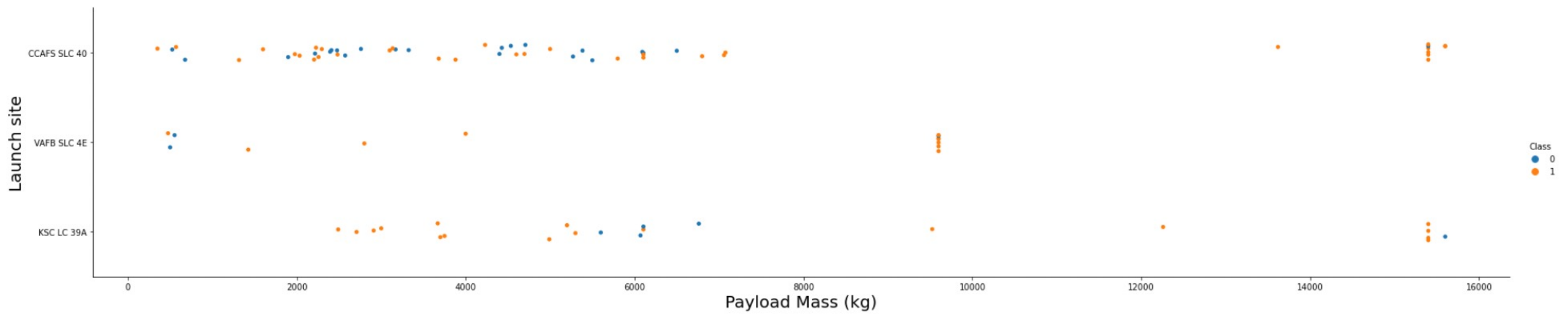
Flight Number vs. Launch Site

- Many of the Falcon 9 launches were performed from CCAFS SLC-40
- In general, for an increasing number of flight attempts the successful return of the stage 1 is more likely
  - CCAFS SLC-40: the first launches were particularly difficult, but learning from the experience obtained from other sites, the probability of a successful return of the stage 1 improved
  - from VAFB SLC-4E: a similar trend can be observed, even though with a much more limited number of launches
  - the same applies also to KSC LC-39A, but the failure of two returns in late launches shall be monitored thanks to future data to be gathered



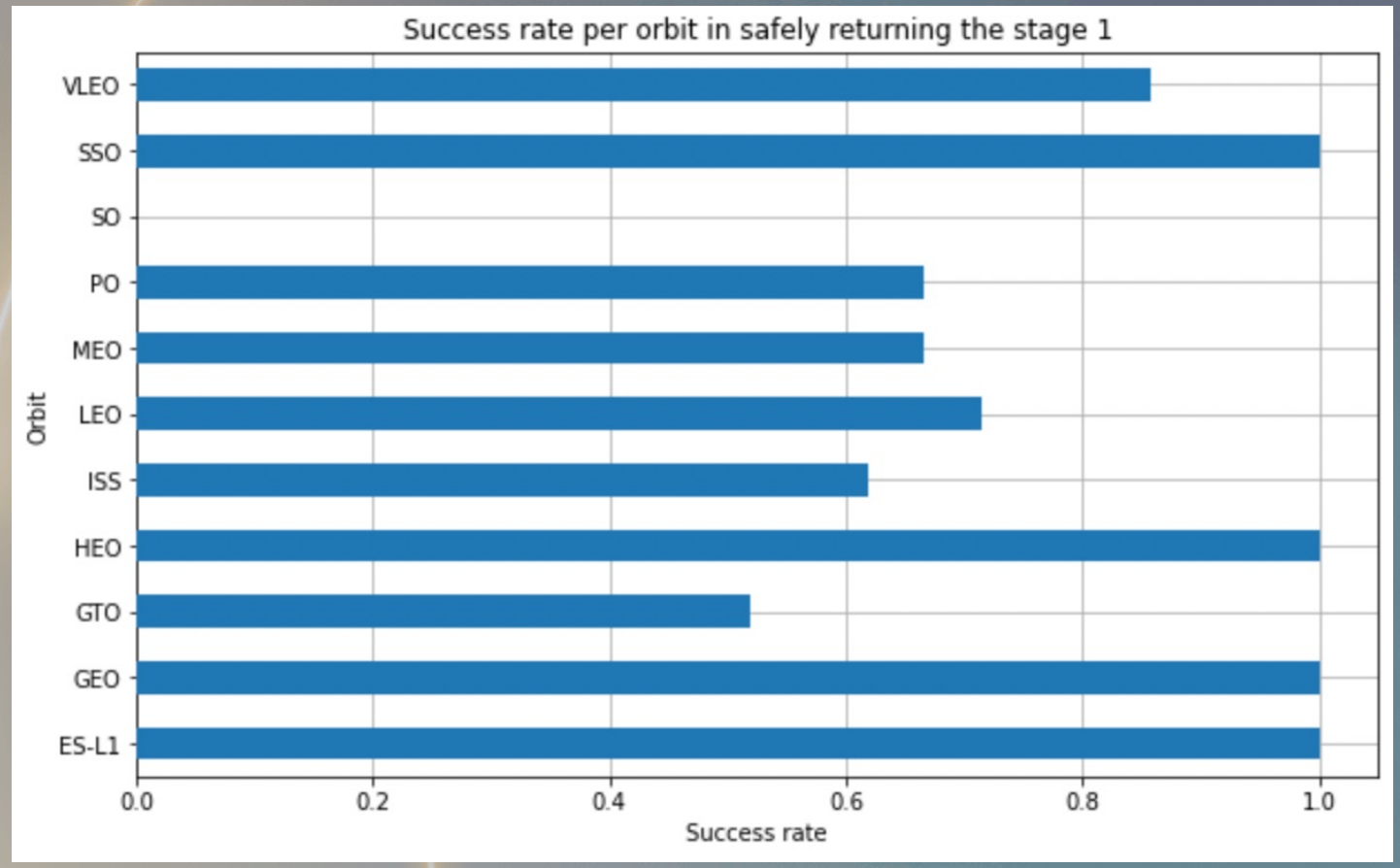
Payload vs. Launch Site

- In general, the larger the payload mass, the more likely is the stage 1 to safely return to ground
  - CCAFS SLC 40: only one return failed with a payload of approximately 16000 kg, while the remaining failures all happened for a payload with a mass lower than 8000 kg
    - In the latter case, the payload mass of failed returns seems to be equally distributed between 1000 and 7000 kg
  - KSC LC 39A: a mass between 5500 and 7000 kg seems critical, resulting in 4 failures, with only one failure with a payload mass is in the order of 16000 kg
  - WAFB SLC 4E: no launches were performed for payloads heavier than 10000 kg (1 failure occurred), while 2 failures were experienced with a very limited mass (approx. 500 kg)



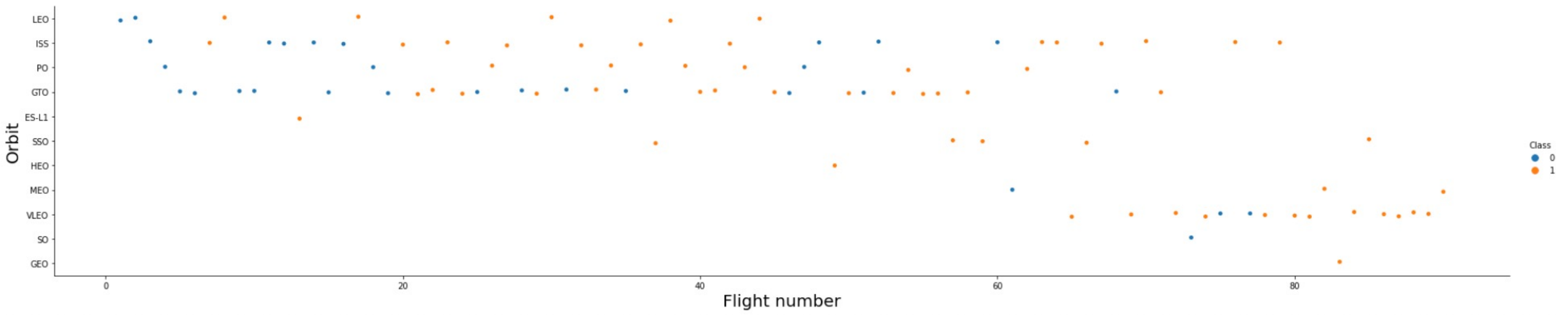
Success rate vs. Orbit type

- 4 orbits at a 100% success:
  - Sun-Synchronous Orbit (SSO)
  - Highly Elliptical Orbit (HEO)
  - Geostationary Transfer Orbit (GTO)
  - Earth-Sun L1 (ES-L1) orbit
  - HEO, GTO, and ES-L1 only had one launch, while for SSO 5 launches were attempted
- Many launches were attempted to Very Low Earth Orbit (VLEO)
  - 12 successes over 14 launches: excellent 85.7% of success
- Between 60% and 70%, we have the success for Polar Orbit (PO), Medium Earth Orbit (MEO), Low Earth Orbit (LEO), and ISS launches
  - achieved after 9, 3, 7, and 21 launches respectively
- Worst orbit: Geostationary Transfer Orbit (GTO)
  - 51.8% of success after 27 launches



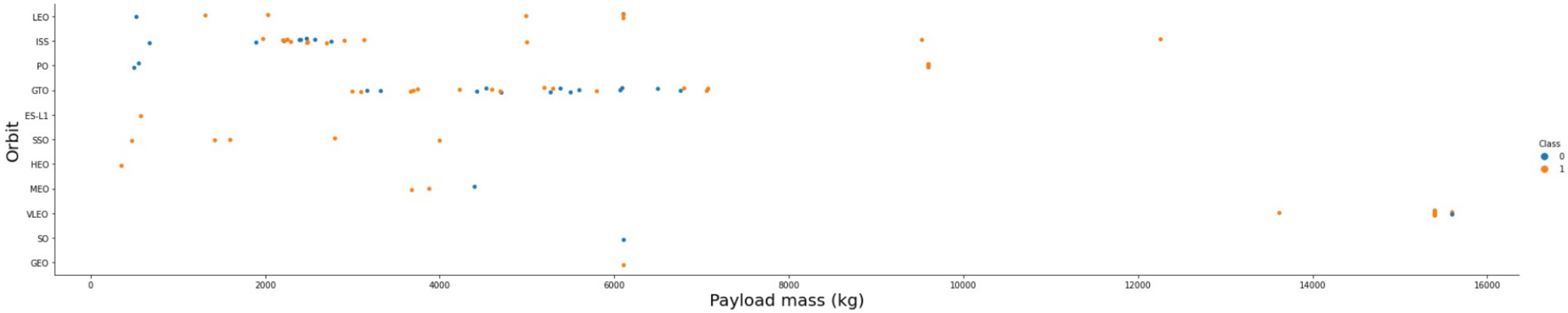
Flight Number vs. Orbit type

- LEO, VLEO, ISS
  - The probability of success appears to be related to the number of flights, i.e., to the experience
  - This is particularly evident for LEO
- GTO: no apparent relation arises between the probability of success and the number of flights
- The other orbits have a too limited number of launches to outline a clear trend



Payload vs. Orbit type

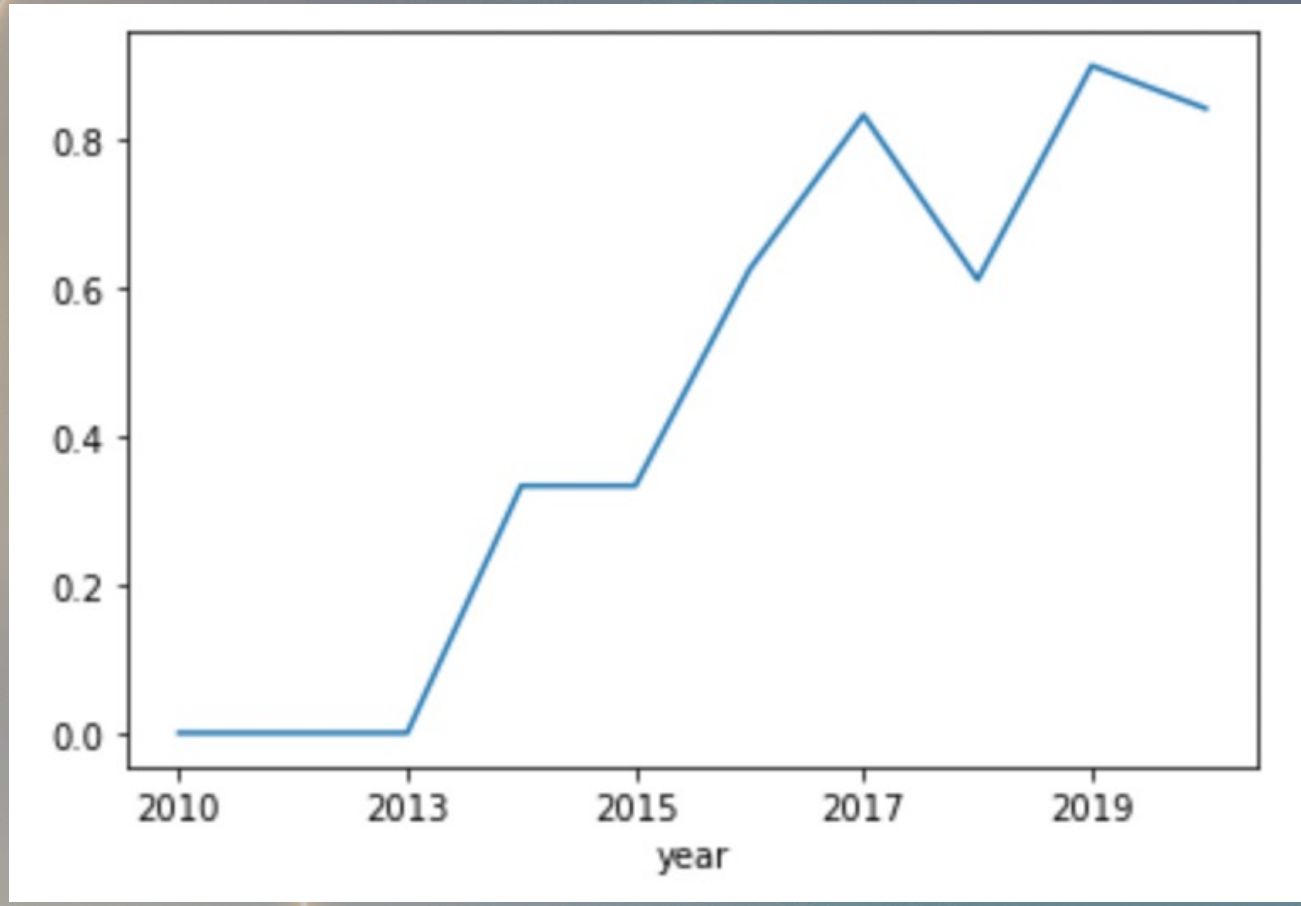
- Heavy payloads (4000+ kg) have a bad impact on the success at GTO
- LEO, ISS, PO: the trend seems to be the opposite → larger payloads are related to an increased probability of success
- The other orbits have a too limited number of launches to outline a clear trend



Launch success yearly trend

- Since 2013, the yearly success rate almost constantly increased
  - The only exception is in 2018

- This clearly indicates that an increased experience in launches brought to a valuable advantage in the mission success




EDA with SQL

All launch site names

- Query to find the names of the unique launch sites:

```
select distinct(LAUNCH_SITE) from SPACEXDATASET
```

- We can notice that 5 launch sites are present in the database
- However, one is related to a typo
  - CCAFSSLC-40 is CCAFS SLC-40

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

Launch site names begin with `CCA`

- Database records in which the launch sites beginning with `CCA`

```
select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

- Here, we limited the search result to 5 entries
  - Clearly, these results are not impacted by the typo in the previous query

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

- Calculate the total payload carried by boosters from NASA (CRS)

```
select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER like 'NASA (CRS)'
```

- Here we have a single output providing the total payload mass
  - 45596 kg

1
45596

Average payload mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION like 'F9 v1.1'
```

- Again, we have a single output providing the result
  - 2928 kg per launch with F9 v1.1

First successful ground landing date

- Date when the first successful landing outcome in ground pad

```
select min(DATE) from SPACEXDATASET where MISSION_OUTCOME like 'Success'
```

- Here, we have a single output reporting the date: June 4, 2010

1
2010-06-04

Successful drone ship landing with payload between 4000 and 6000

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

```
select BOOSTER_VERSION from SPACEXDATASET where
lower(LANDING__OUTCOME) like '%drone%' and LOWER(MISSION_OUTCOME) like
'%success%' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

- We have 5 F9 booster versions that satisfy the above requirements

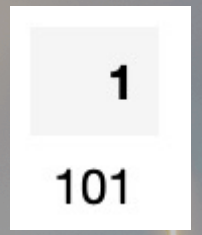
booster_version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total number of successful and failure mission outcomes

- Calculate the total number of successful and failure mission outcomes

```
select count(*) from SPACEXDATASET where lower(MISSION_OUTCOME) like
'%success%' or lower(MISSION_OUTCOME) like '%failure%’
```

- We have 101 mission outcomes



Boosters carried maximum payload

- List the names of the booster which have carried the maximum payload mass (using a sub-query)

```
select BOOSTER_VERSION from SPACEXDATASET where PAYLOAD_MASS__KG_ =
        (select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET)
```

- We have 12 boosters that 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

- List the records which will display the month names, failure landing outcomes on a drone ship, booster versions, launch site for 2015

```
select monthname(DATE) as month_name, LANDING__OUTCOME, BOOSTER_VERSION,
LAUNCH_SITE from SPACEXDATASET where lower(LANDING__OUTCOME) like '%drone%' and
LOWER(MISSION_OUTCOME) like '%failure%' and year(DATE)=2015
```

- A single record satisfies the above requirements:

month_name	landing__outcome	booster_version	launch_site
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

```
select count(LANDING__OUTCOME) as outcome_count from SPACEXDATASET
where DATE between '2010-06-04' and '2017-03-20' group by
LANDING__OUTCOME order by outcome_count DESC
```

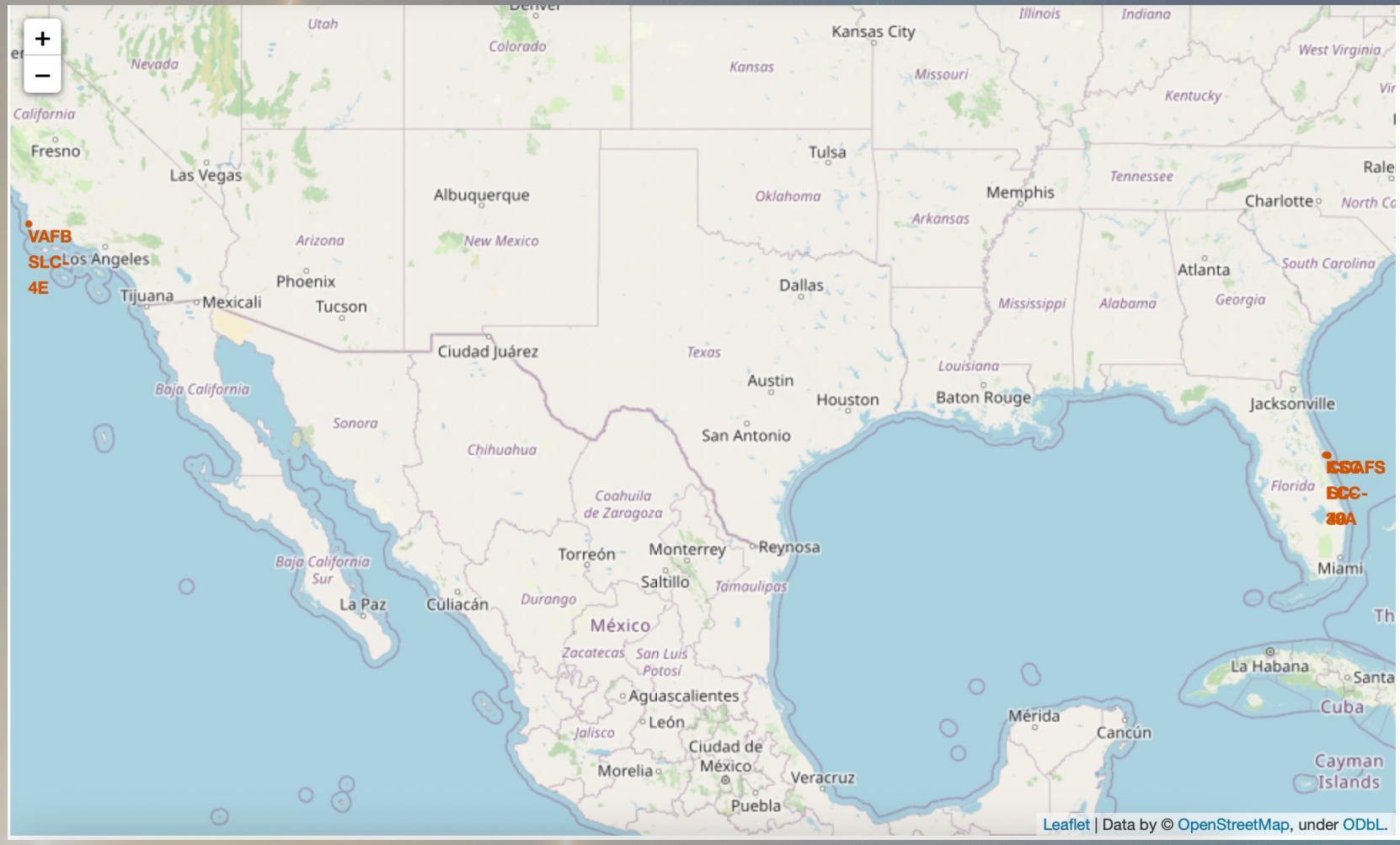
- On the right, we have the result for this query:

outcome_count
10
5
5
3
3
2
2
1

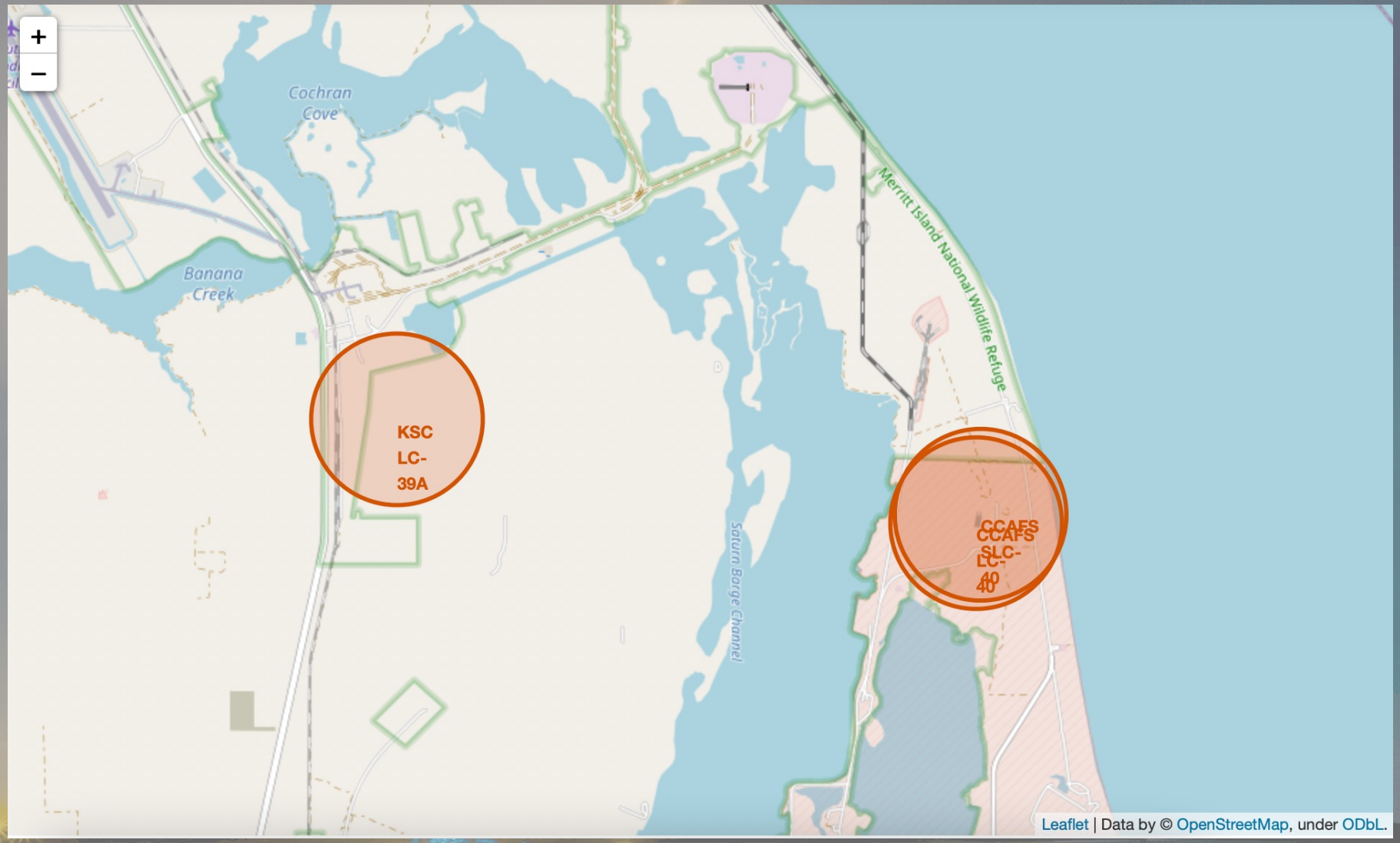

Interactive map with Folium

Launch sites locations

- 4 launch sites
- 1 on the West Coast
  - VAFB SLC-4E
- 3 on the East Coast
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - These sites are basically co-located (see next slide)

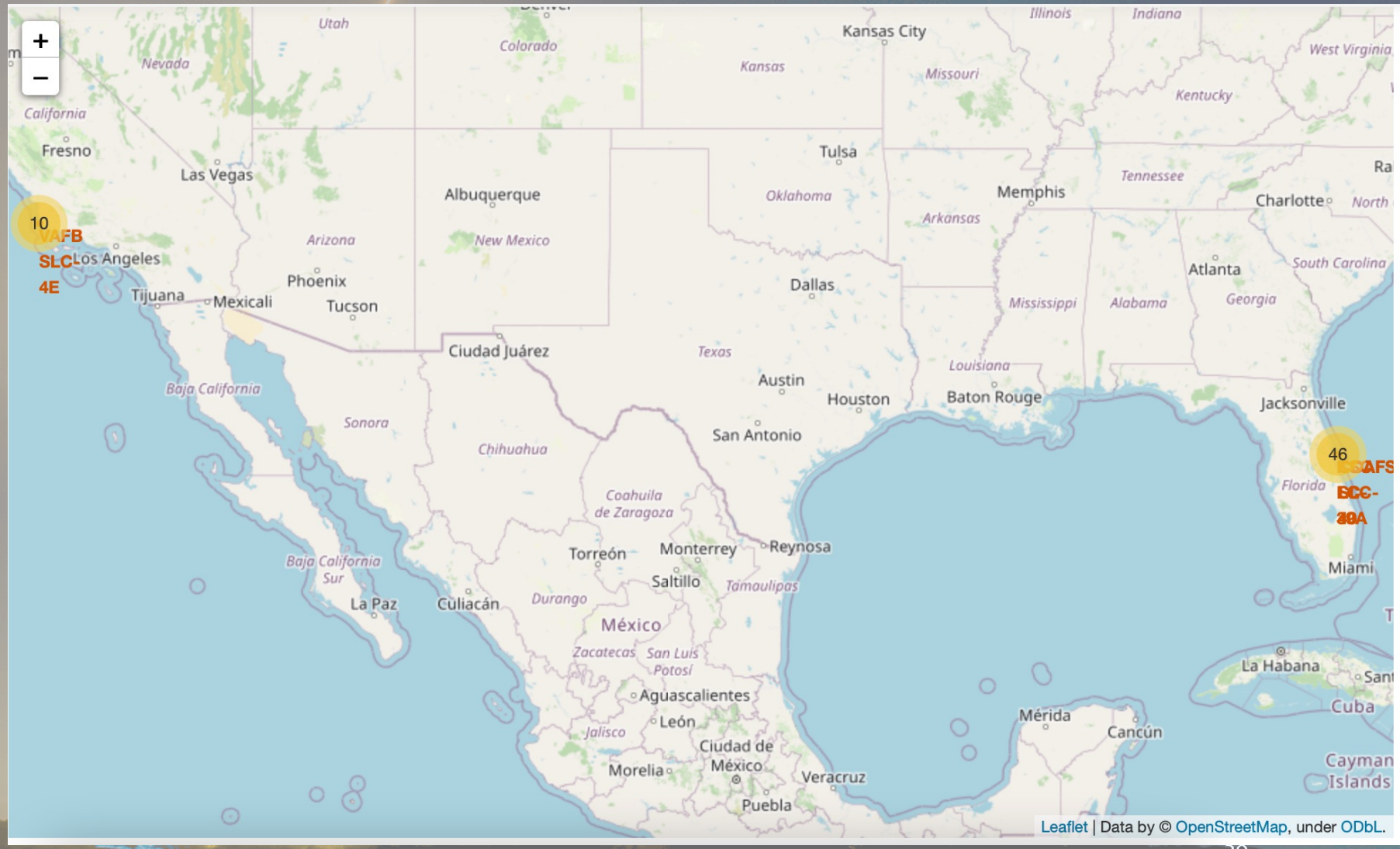


Launch sites locations: East Coast



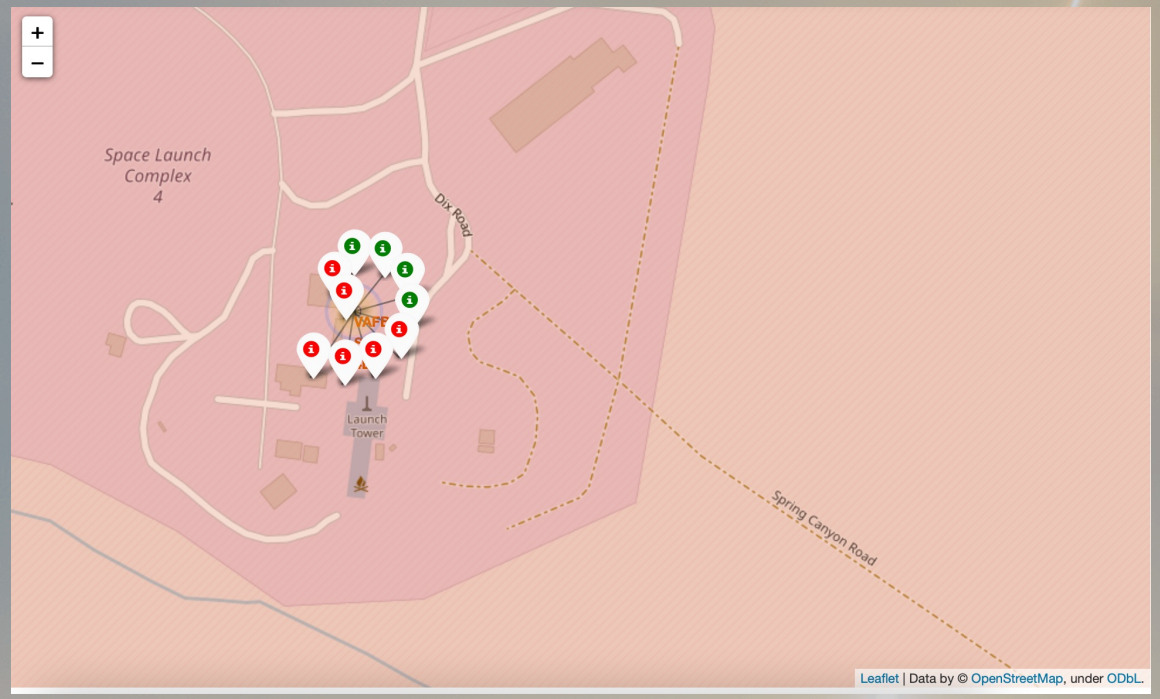
Launch records per site

- 10 launches from the West Coast site
- 46 launches from the 3 sites on the Atlantic

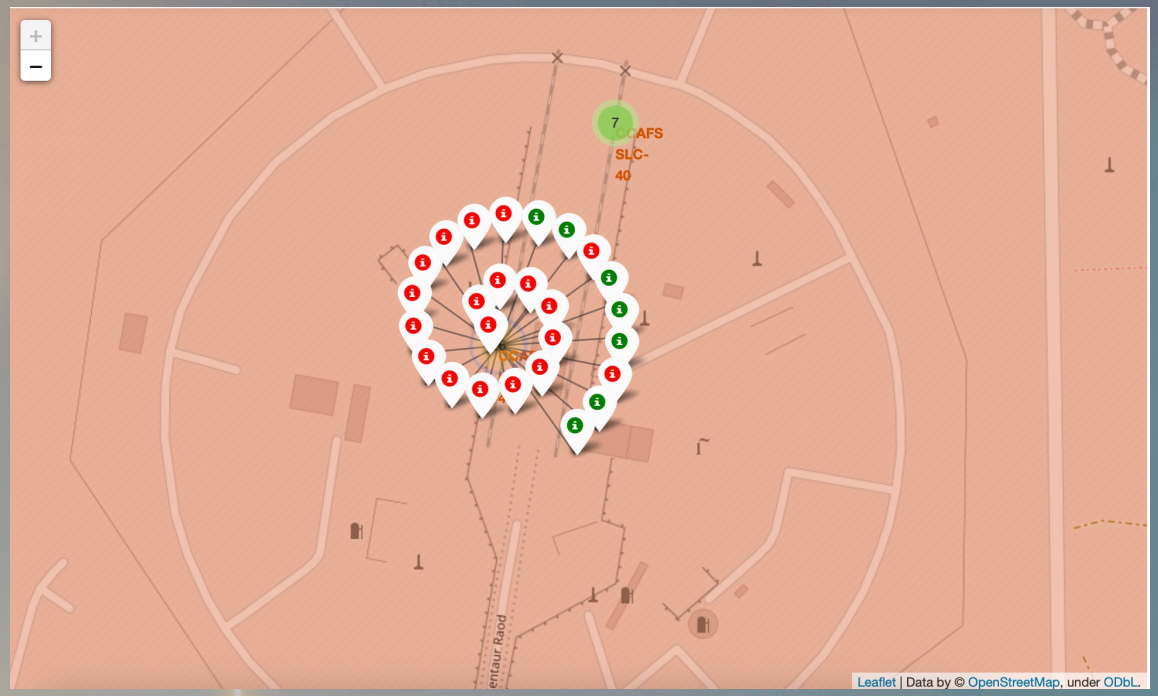


Launch records per site: details

- VAFB SLC-4E

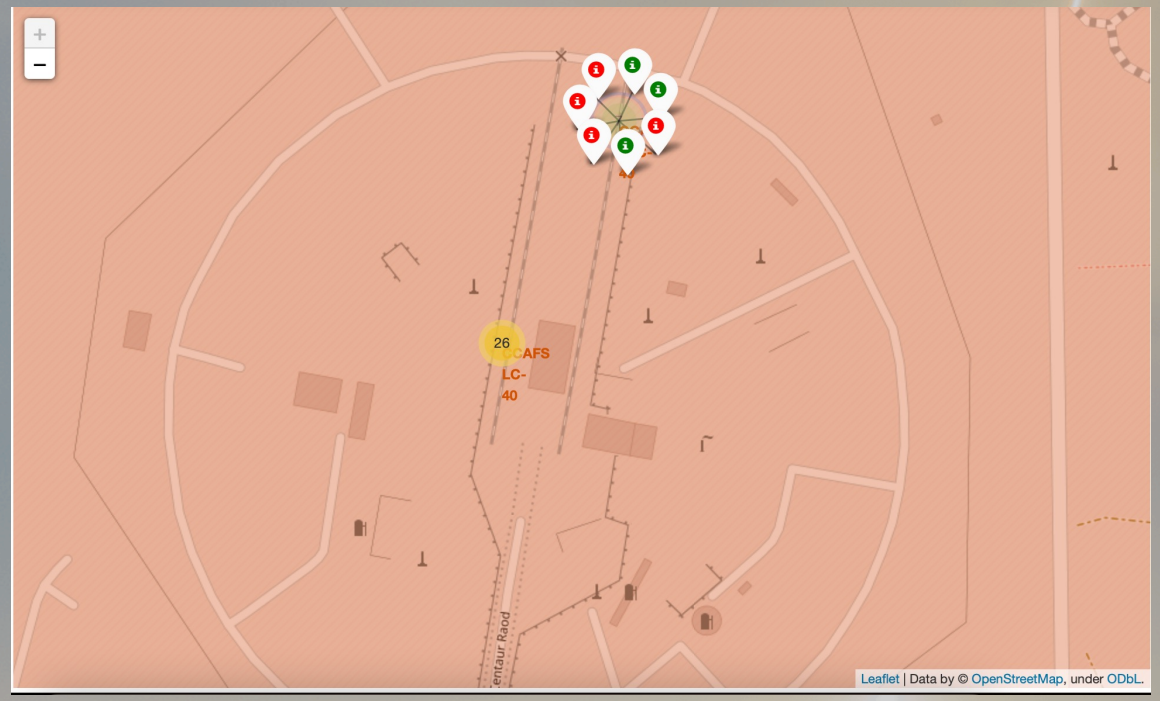


- CCAFS LC-40

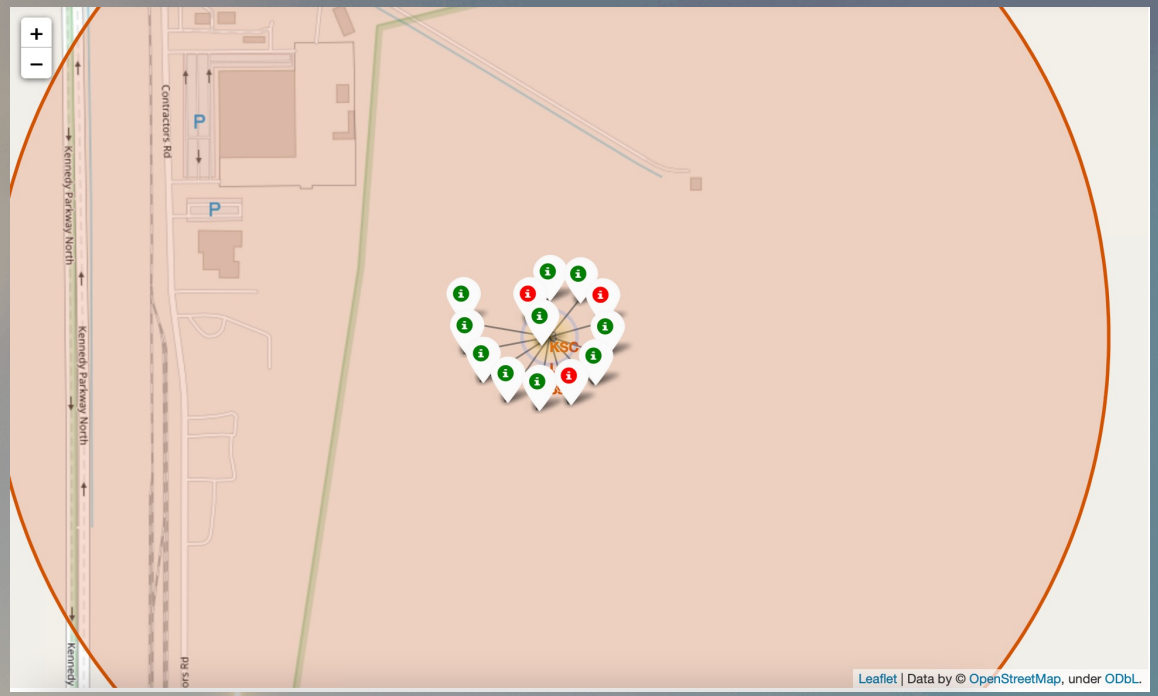


Launch records per site: details

- CCAFS SLC-40

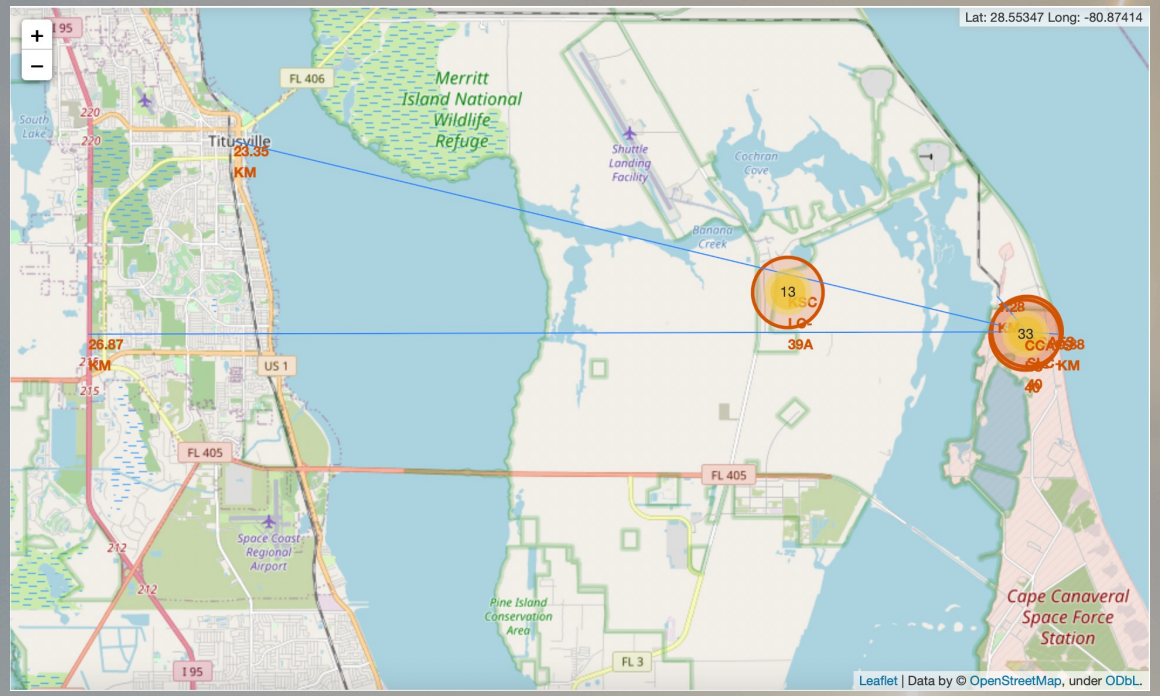


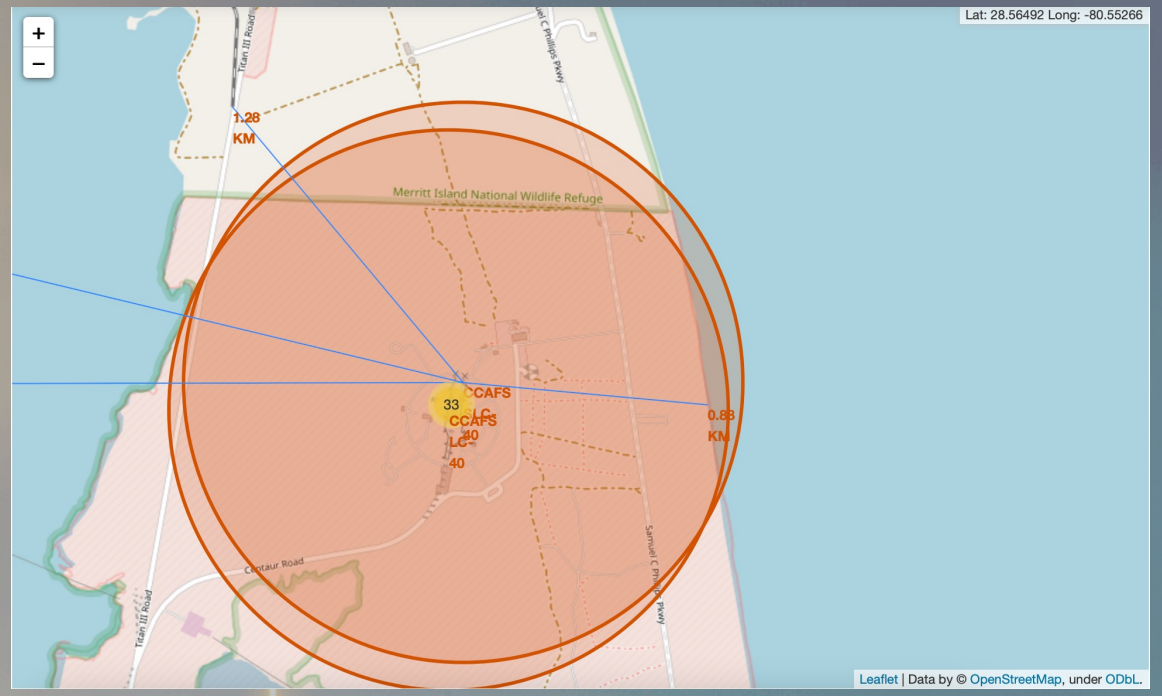
- KSC LC-39A



Proximities of CCAFS SLC-40

- nearest city (Titusville): approx. 23.35 km
- nearest highway: approx. 26.87 km
- nearest railway: approx. 1.28 km
- nearest coastline: approx. 0.88 km



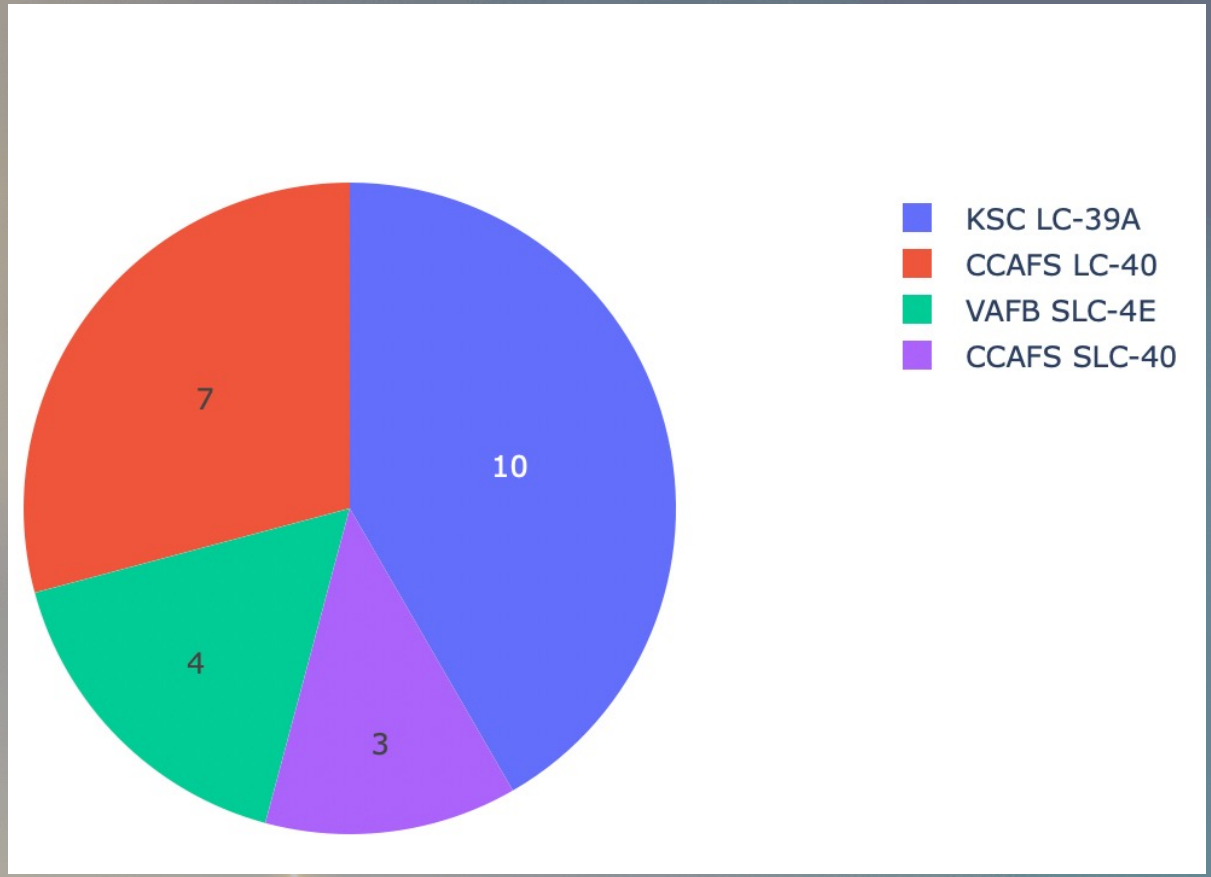


- launch sites away from cities and critical infrastructures (e.g., a highway) is fundamental to avoid harming people or infrastructures in case of disruptive failures
- a nearby source of water (the Ocean here) is an advantage to deal with fire and explosions


Build a Dashboard with Plotly Dash

Successful launcher per site

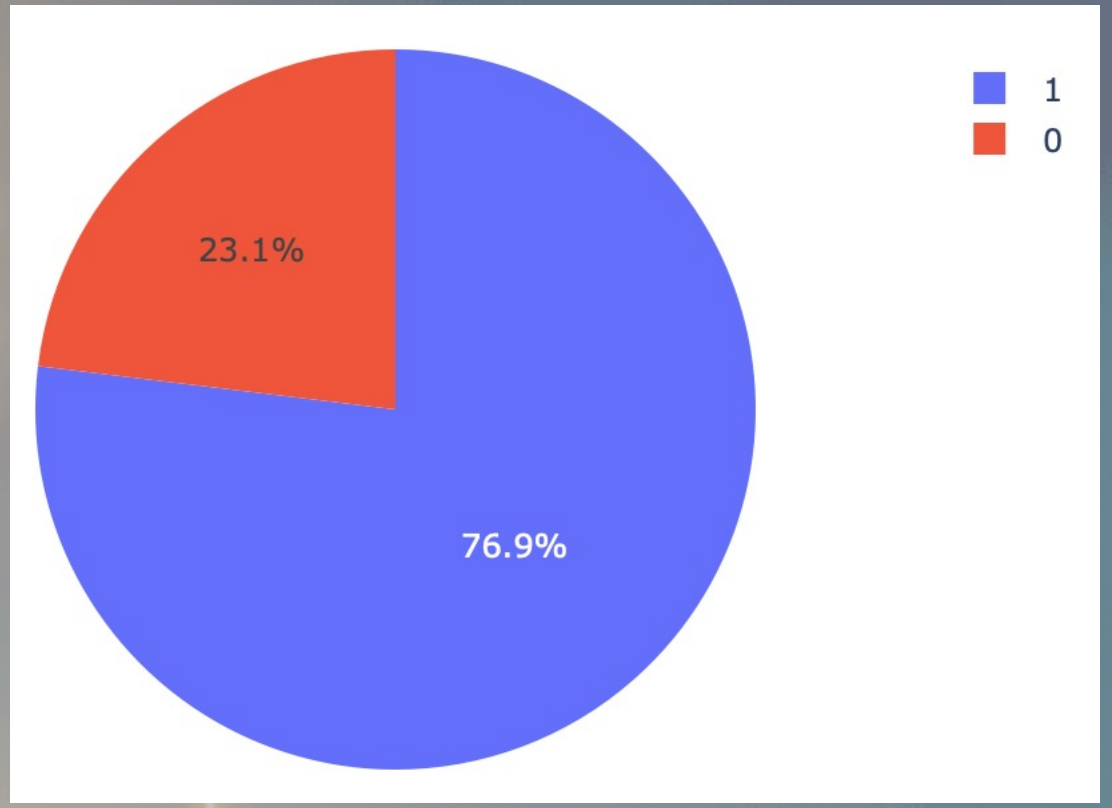
- There have been 24 launches that saw a successful outcome
- The best performing site is KSC LC39-A with 10 successes
- The worst site is CCAFS SLC-40, with only 3 successes



Success ratio at KSC LC-39A

- The KSC LC-39A is also the site with the best success ratio

- From 13 overall launches:
  - 10 successes → 76.9%
  - 3 failures → 23.1%



Payload mass vs mission outcome

- In general, the FT booster version seems to be the best one
- Payloads between 2000 kg and 5500 kg have the best outcomes

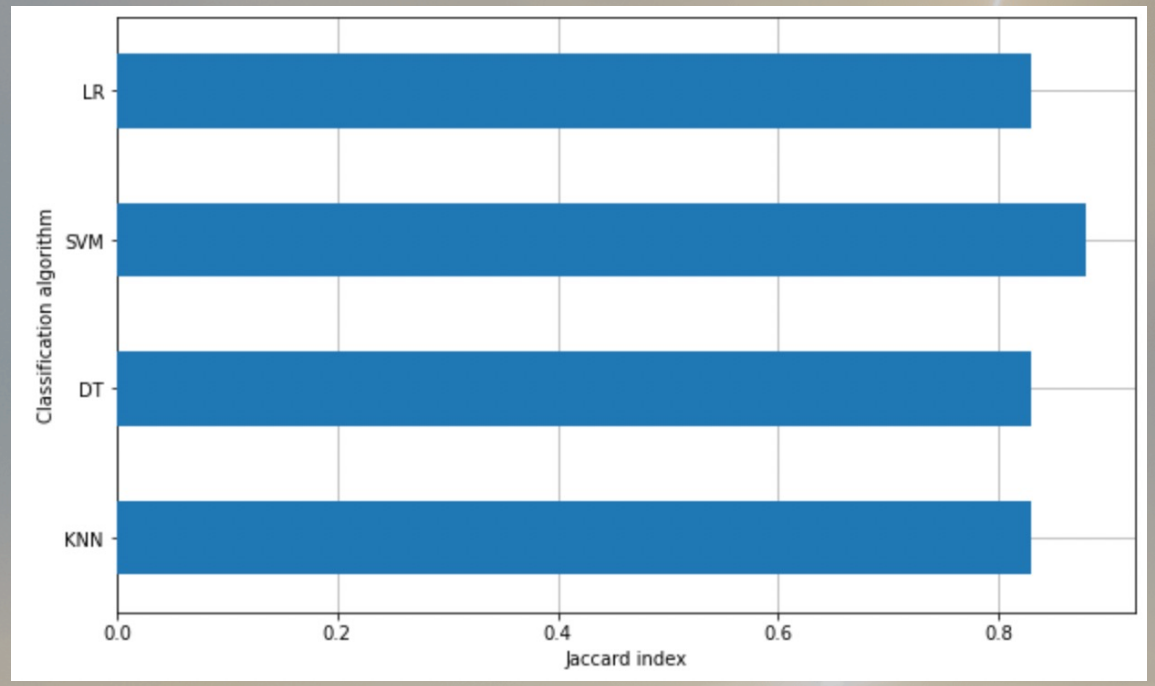
Payload mass vs mission outcome: detail

- Zooming in, indeed the FT booster is much better compared to others
- A payload mass between 2000 kg and 4000 kg does provide a good portion of successes, but it also has come failures (most of them with the v1.1 booster)

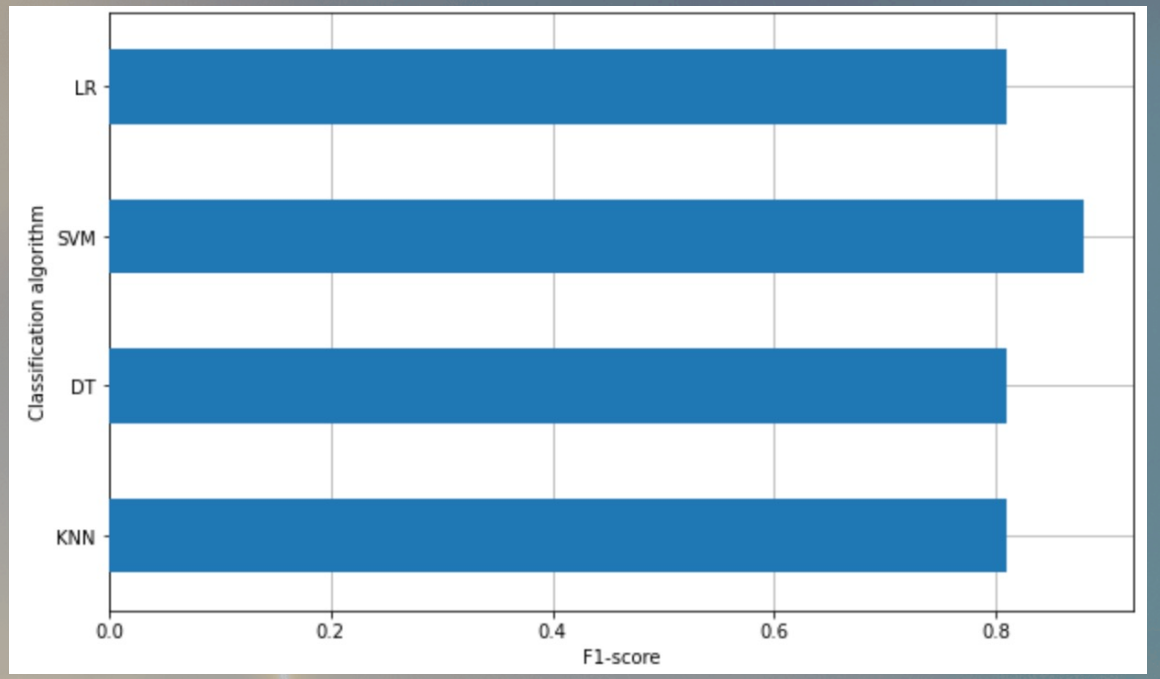
Predictive analysis (Classification)

Classification Accuracy

- Jaccard index: SVM is the best classification algorithm

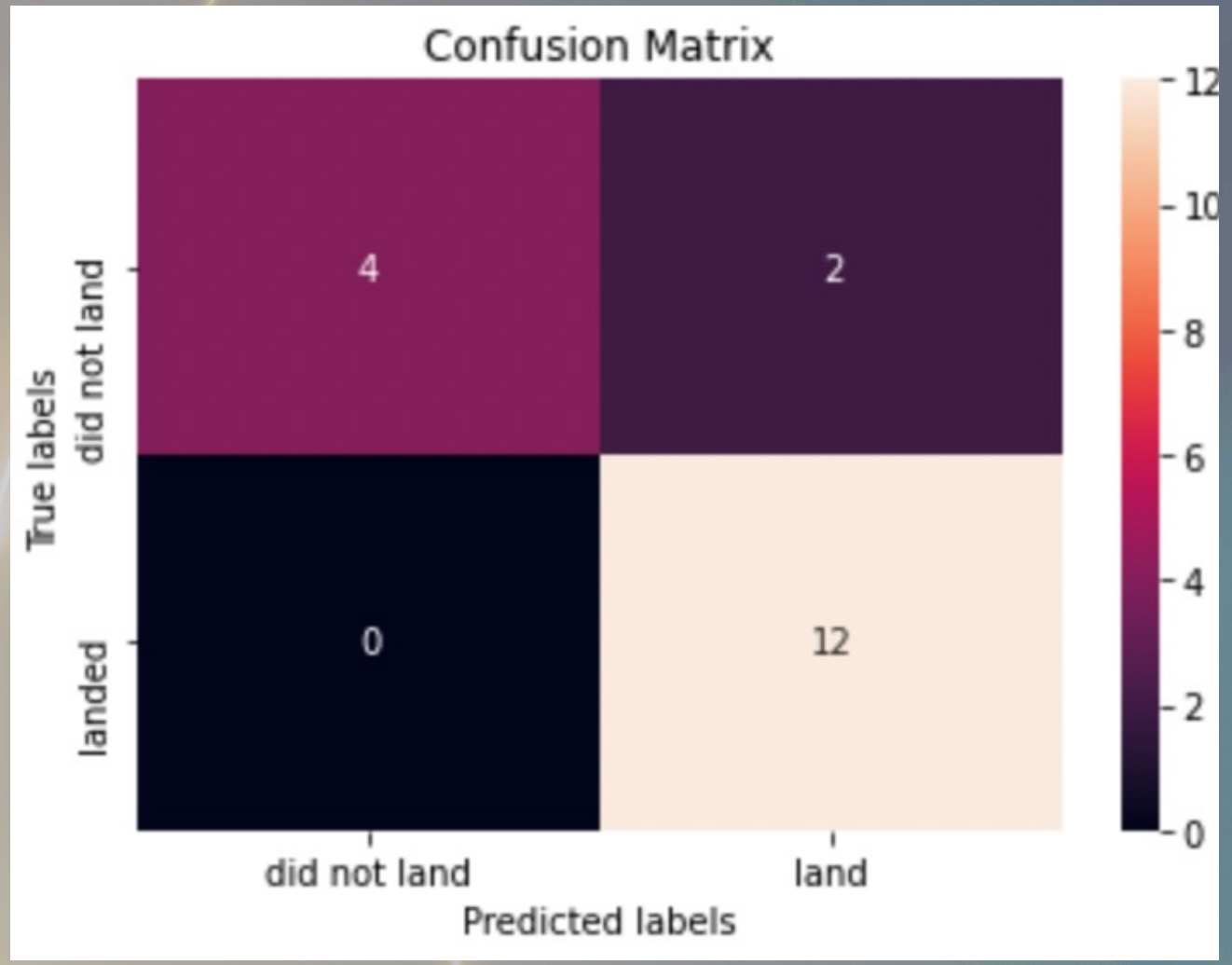


- F1-score: SVM is the best classification algorithm



Confusion Matrix

- The optimized SVM classification provides an excellent performance
- No False Negatives are present, with all true negatives (12) correctly classified
- There are 2 false positives, over a total of 6 negative classes, which might need some further exploration



Cost estimation

- With the SVM model, the accuracy is 88%
  - This means that in the 88% of cases, we will correctly predict the first stage reentry outcome
  - In the remaining 12% of the cases, it appears (see the confusion matrix in the previous slide) that the trend is towards False Positives
    - More landing predicted compared to the actually successful outcomes
- Thus, the classification model might provide lower budget estimations compared to those actually needed
- However, this needs to be accurately verified with much more data
  - Here, we only had 18 samples in the test set

Conclusions: factors impacting the success

- With Exploratory Data Analysis, many interesting correlations were observed
- The success probability increases when
  - The flight number increases, thanks to a better experience gained from past launches, as also shown in the yearly trend of the success rate
  - A larger payload mass brings benefit to the success rate at LEO, ISS, and PO, but not at GTO
  - In terms of the orbit, LEO/ISS and VLEO (85.7%) have an excellent success probability, considering the large number of launches

Conclusions: launch site analysis

- Analyzing the surroundings of the launch sites some considerations were relevant
- The 4 launch sites are all close to the coast
  - Having a large water pool is an advantage, even though not mandatory, in case of explosive outcomes
  - All sites are away from cities and critical infrastructures (highways, public railways, …)
    - This ensures that no harm is provided to people or infrastructures in case of incidents
  - However, a private/reserved railway close to the launch site can be useful to move components or even entire rockets

Conclusions: classification model

- We assessed, with hyperparameter optimization, the performance of
  - Support Vector Machine
  - K-Nearest Neighbour
  - Decision Tree
  - Logistic Regression
- SVM is the best performing algorithm, with 88% of accuracy on the test set
- It shall be noticed that further data might be needed to better assess the performance in the real world

We can effectively predict the outcome of the first stage reentry, which means correctly estimating the launch cost