IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

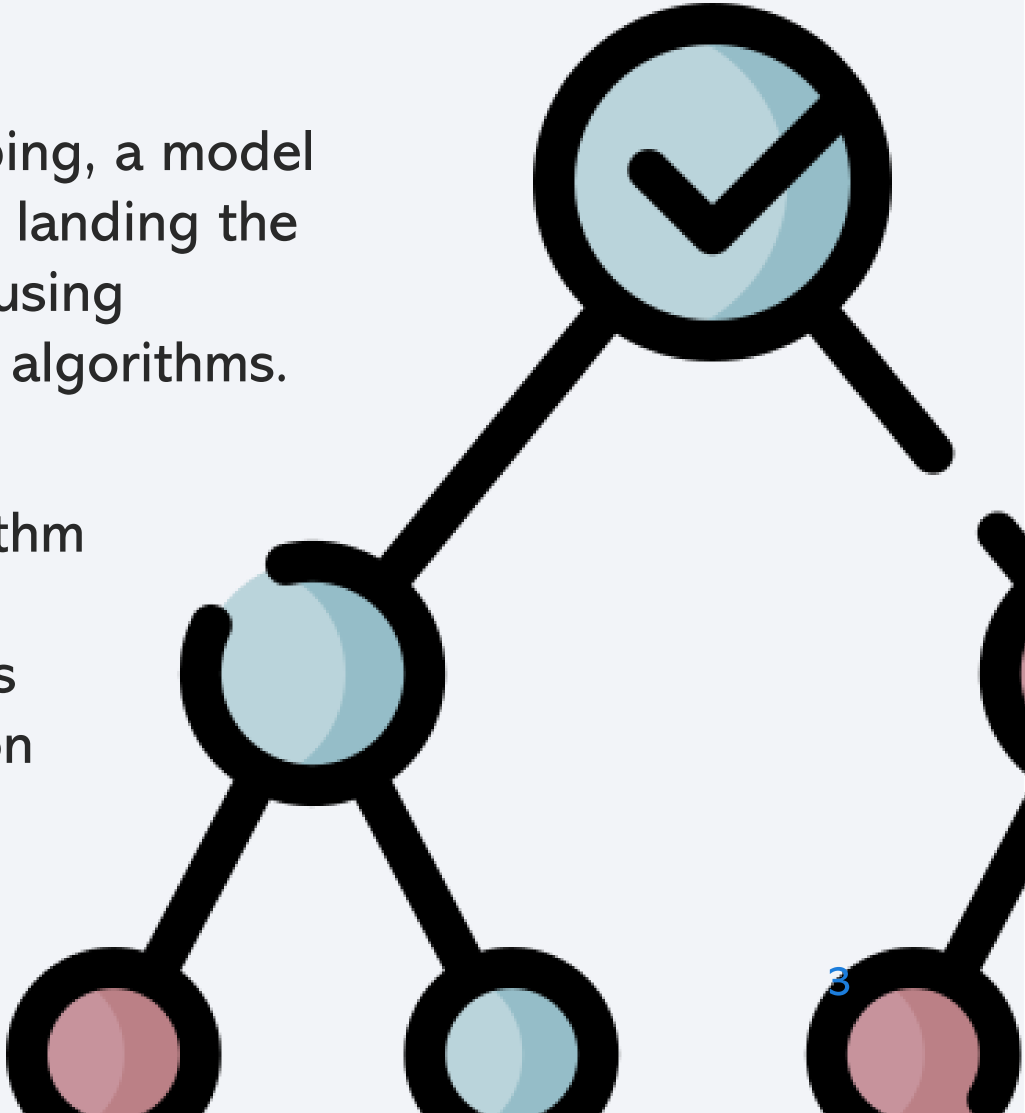
David Francisco Ortiz Gutiérrez
2022/08/25

Outline

Executive Summary

- Obtaining data from SpaceX API and by web scraping, a model was trained for predicting the success or failure of landing the first stage of a Falcon9 rocket. This was achieved using technologies as SQL Python and machine learning algorithms.

- The best model obtained is a Decision Tree algorithm with an accuracy of 0.83 on unseen data. Also, an interactive Dash Board for visualizing the data was created, where is possible to extract information on how the different variables involved in a lunching affect the outcome of a landing



Introduction

- SpaceX is a spacecraft manufacturer who builds rockets with the capability of landing some of its stages after launching. Of course, landing a rocket is not an easy task and it sometimes fails.
- As part of an attempt to study what variables can affect the landing and how to improve the success rate of it, a predictive model is trained. This model receive the variables for a future launching and estimates if will have a successful landing or not.

Section 1

Methodology

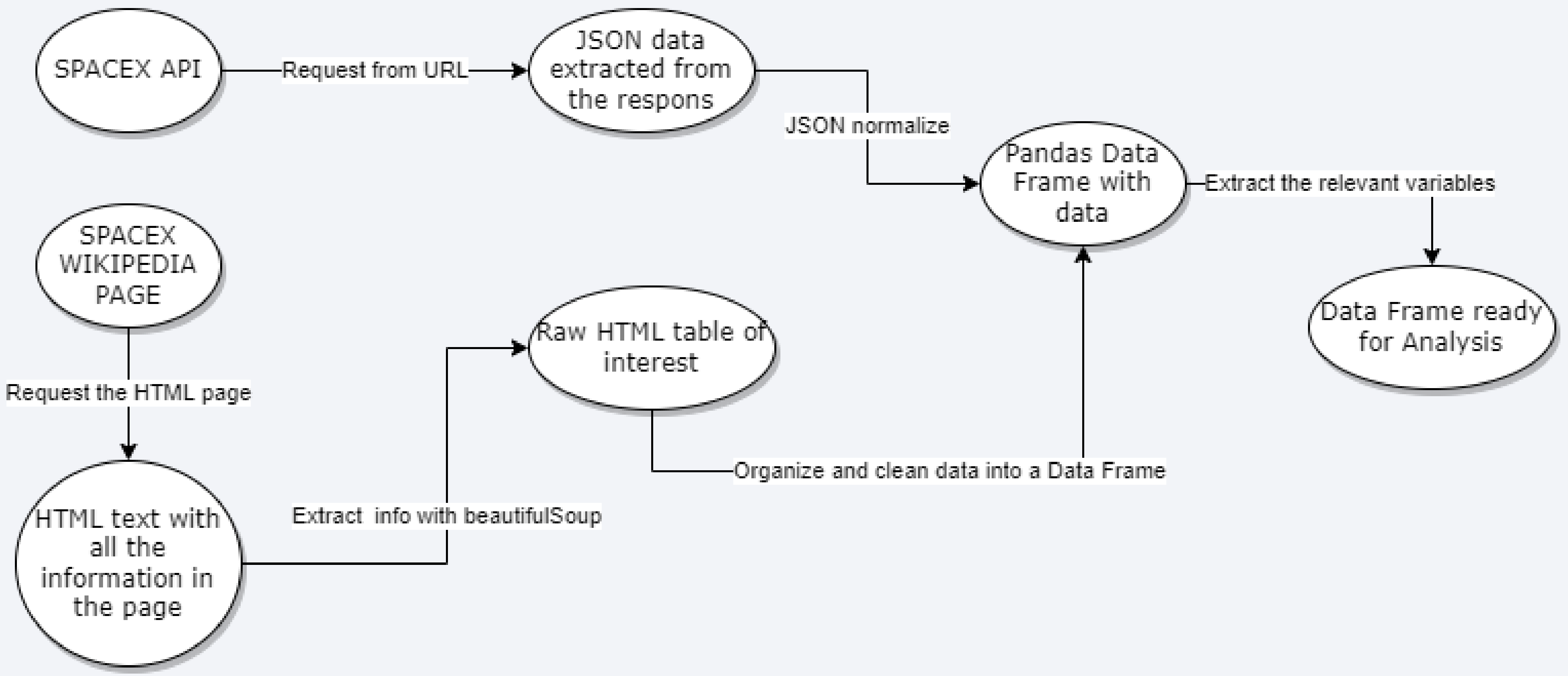
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected in the SpaceX API and via web scraping
- Perform data wrangling
  - Data was cleaned using Python and Pandas
- 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

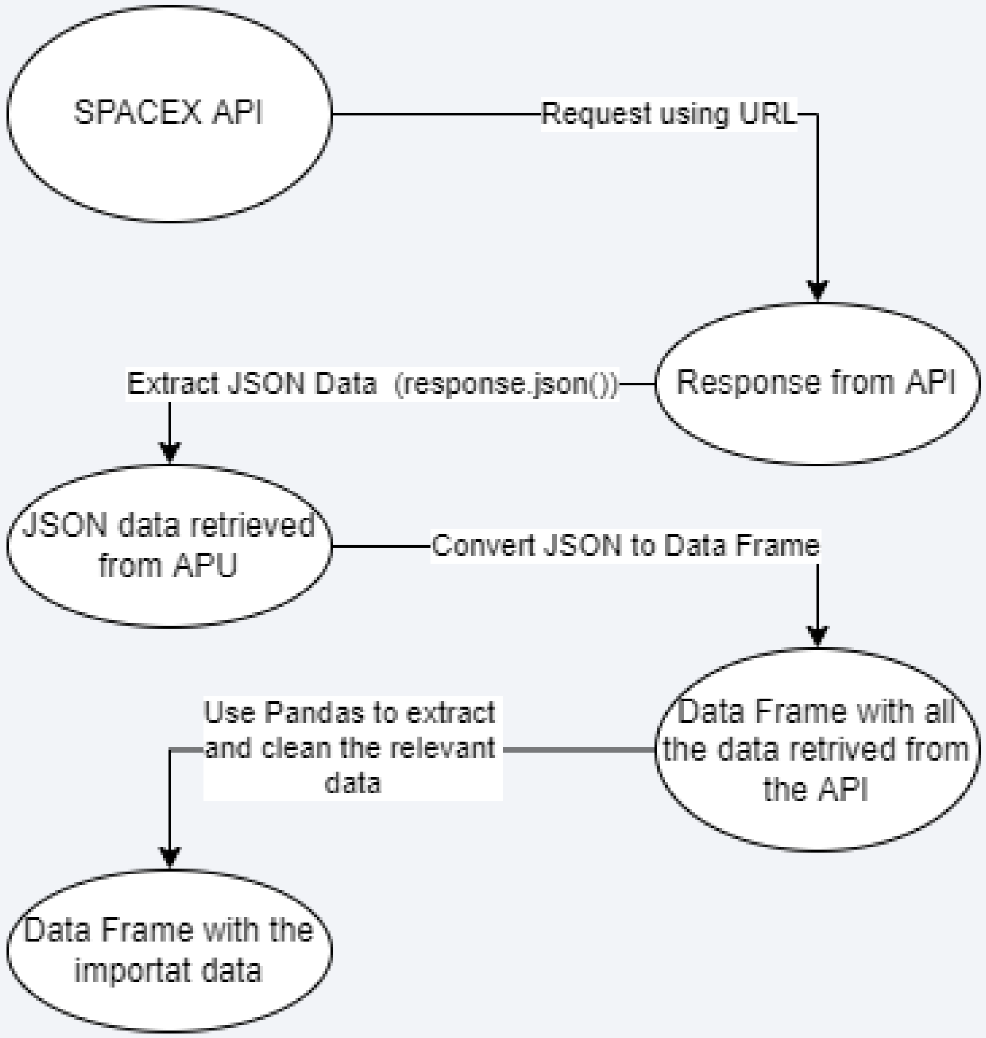
Data Collection

- Data was extracted using the SpaceX API and from the Wikipedia page for the SpaceX launches



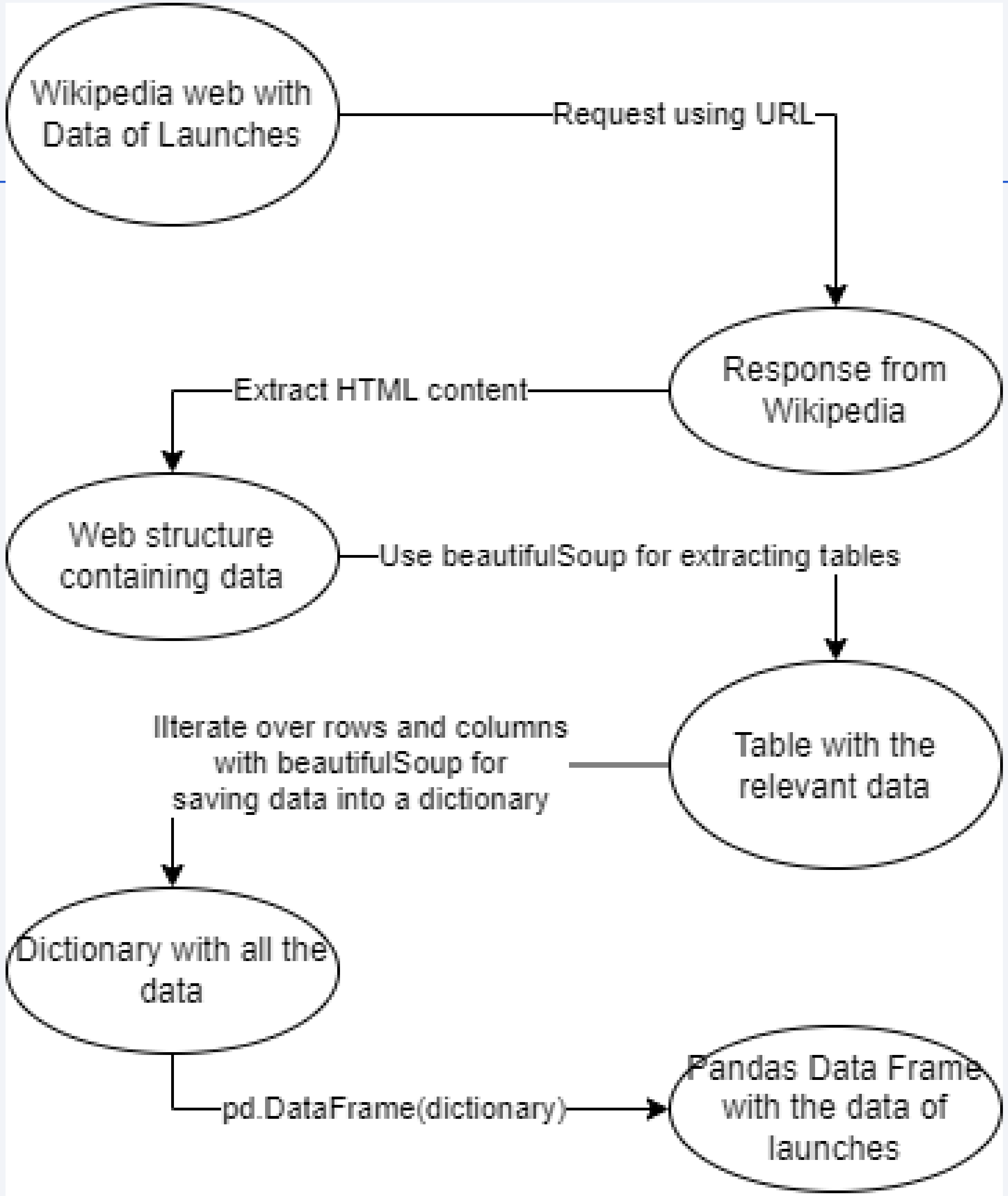
Data Collection – SpaceX API

- Data Collection via SpaceX API shown below ([Link to Notebook]())
- Request module was used for retrieving data
- Pandas is used for cleaning and filtering data



Data Collection -Web Scraping

- Data collection via Web Scraping (Link to Notebook)
- Web scraping of HTML tables using beautifulSoup
- Convert data to Pandas Data Frame

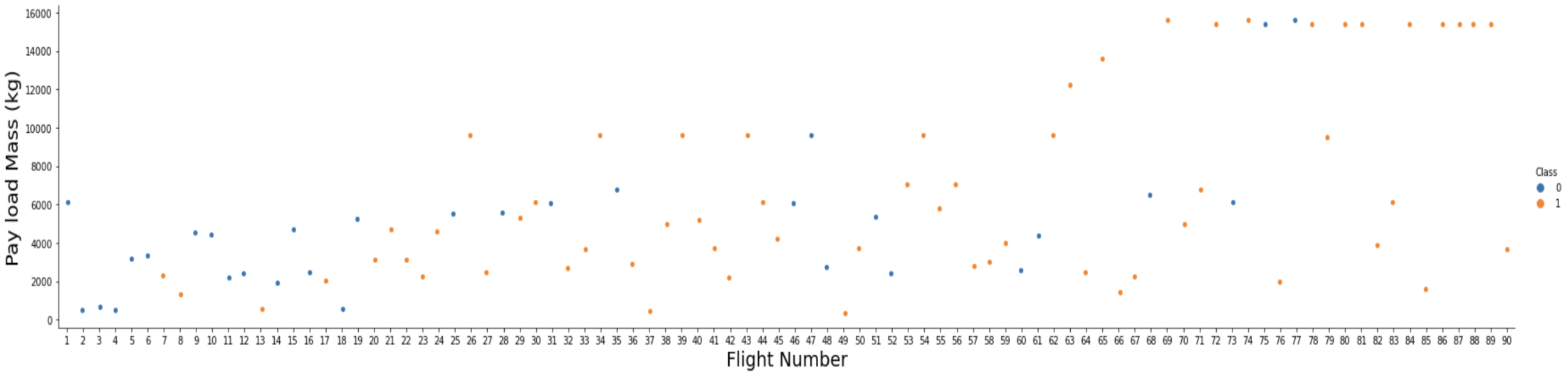


Data Wrangling

- For each of the launches the data extracted contains variables like Launch Site, kind of orbit, booster version and Outcome (How the landing ended)
- Using Pandas, 3 different Launch sites were found, as well as the amount of launches in each site.
- The same was done with the Orbit
- The kind of Outcome is analyzed and categorized as successful or failure. The new category is called class and is added to the Data Frame
- (Link to Notebook)

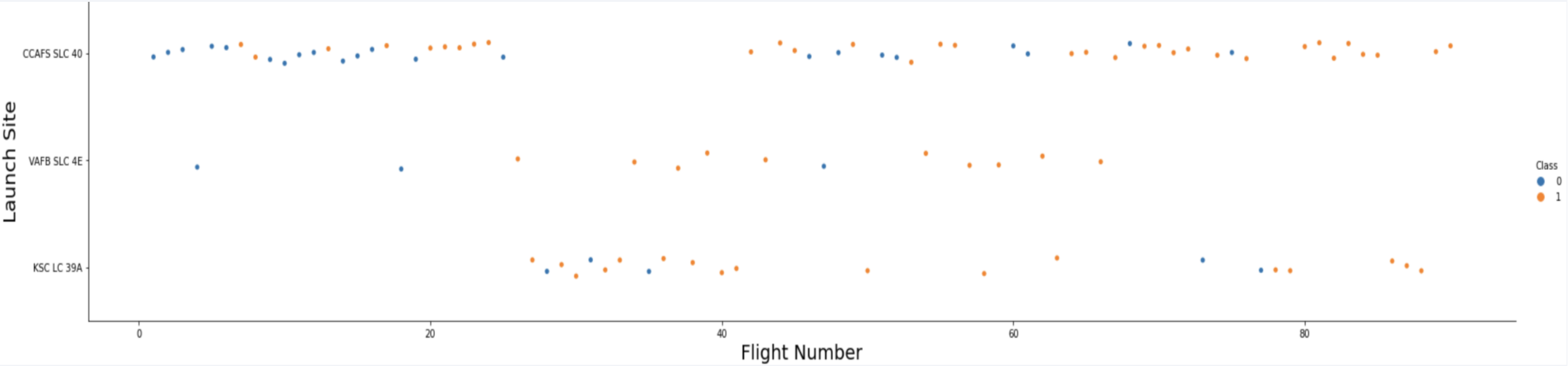
EDA with Data Visualization

- (Link to Notebook)
- Relation between Payload Mass and the success (class 1) or failure (class 0)



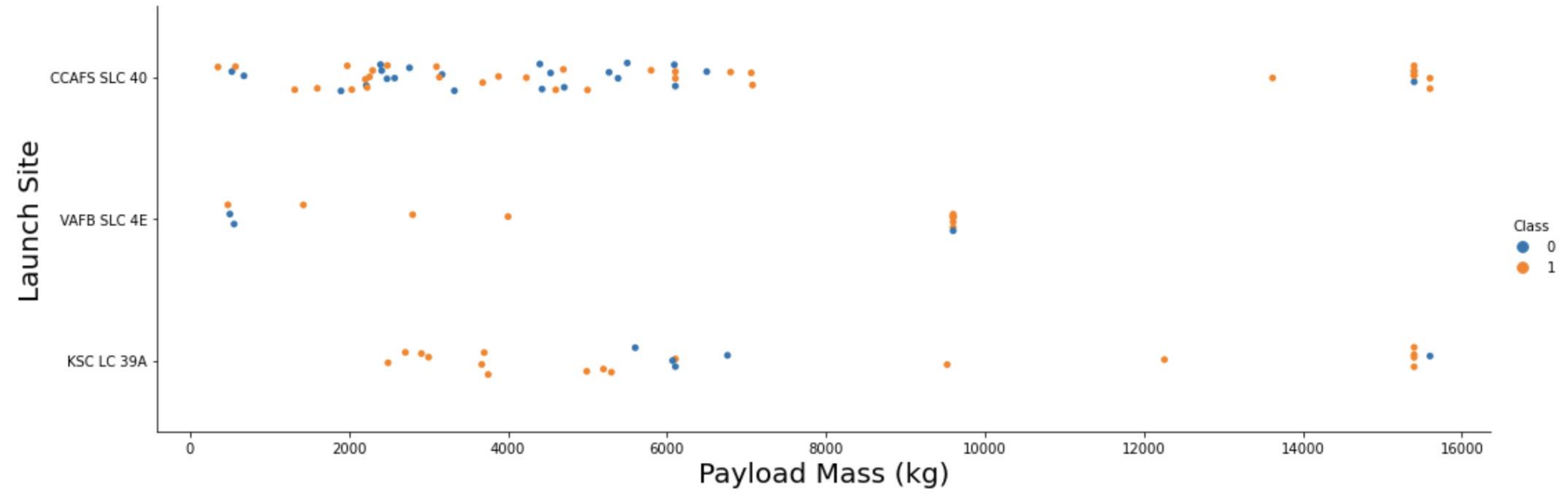
EDA with Data Visualization

- Relation between launch site and class



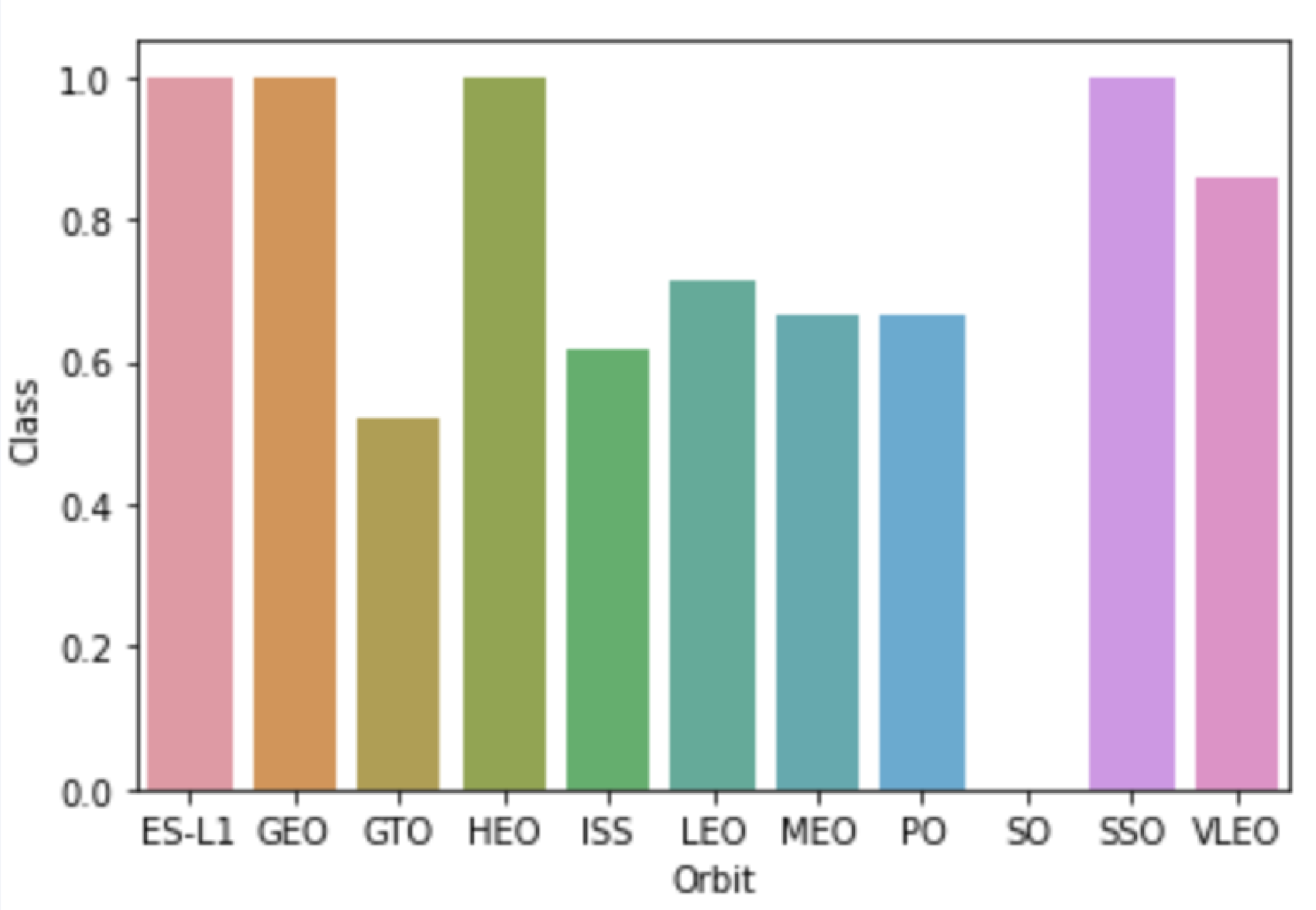
EDA with Data Visualization

- Relation between Payload mass, Launch Site and Class



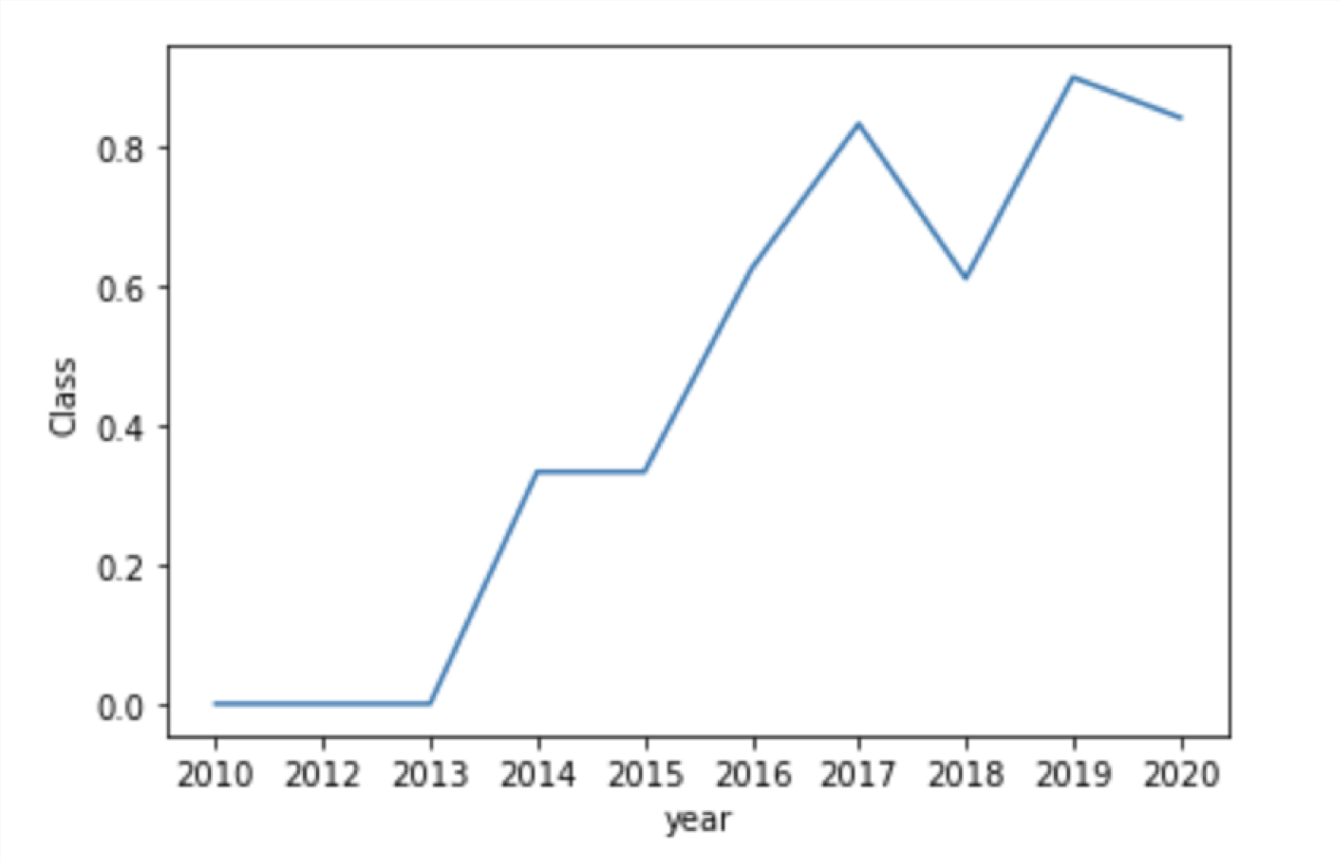
EDA with Data Visualization

- Rate of success for each kind of orbit



EDA with Data Visualization

- Success rate over time (yearly)



EDA with SQL

- Find of unique categories in variables using DISTINCT
- Summarize data using GROUP BY for finding averages, totals and count
- Find specific values in a variable using WHERE
- ([Link to Notebook])

Build an Interactive Map with Folium

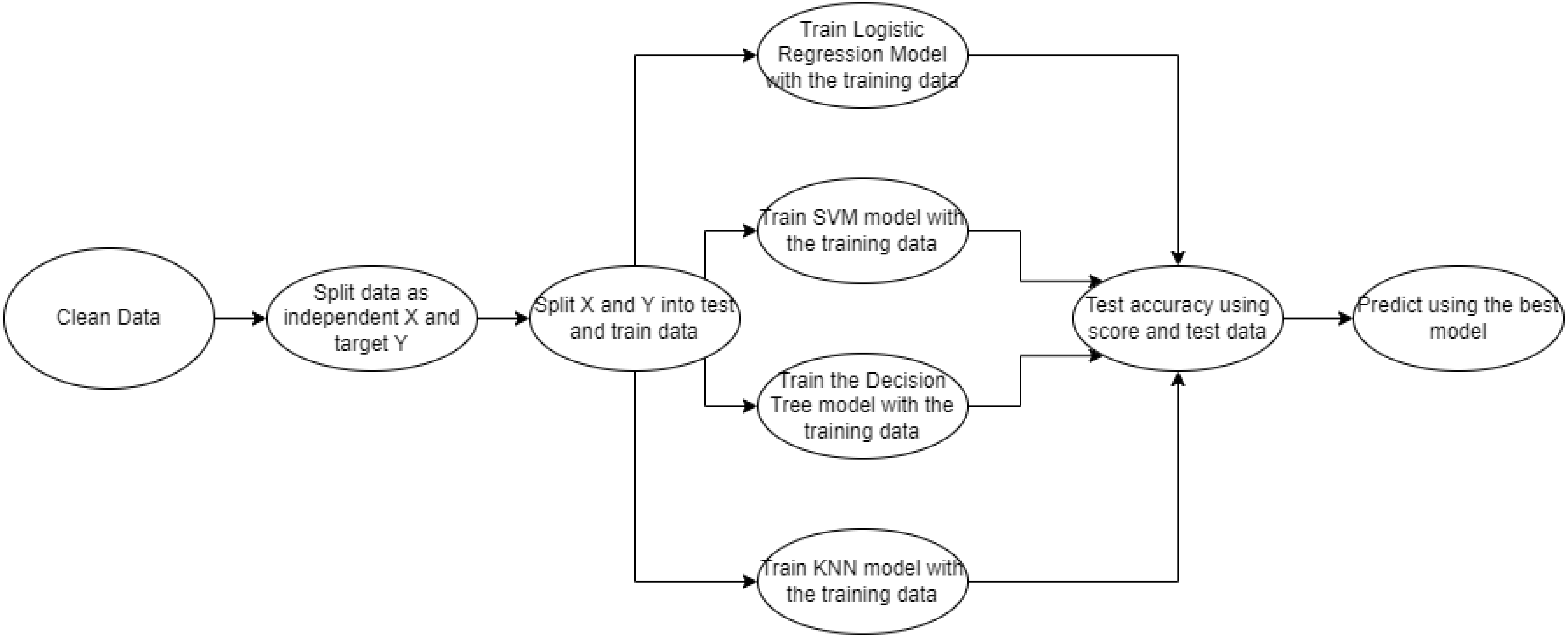
- On Folium where added some items like:
  - Circles to show the location of something and show its size
  - Lines to show the distance between points
  - Markers and groups of markers to show the success of a landing
- This objects allow to show how some of the different variables relate geographicaly
- (Link to Notebook)

Build a Dashboard with Plotly Dash

- A pie chart to show the percentage of launches for each launching site. This can be changed dynamically to show the success rate for each site. To change this a dropdown list is used
- Also, a scatter plot showing the success or failure of each launch in relation to is Payload mass. With the dropdown list the launch site to show can be selected and with a slider the payload mass range to show in tha plt can be changed
- (Link to script)

Predictive Analysis (Classification)

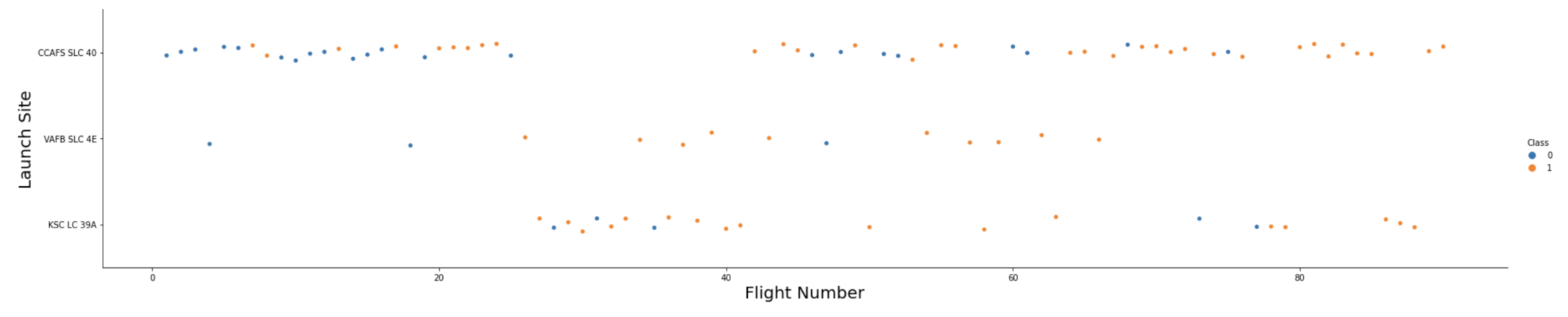
- The data was splited into test and train data
- Multiple models where trained using the train data
- Using the test data each of the trained models was tested. It was made using the score method in each of the models
- The models trained were: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors
- (Link to Notebook)


Clean Data
Split data as independent X and target Y
Split X and Y into test and train data
Train Logistic Regression Model with the training data
Train SVM model with the training data
Train the Decision Tree model with the training data
Train KNN model with the training data
Test accuracy using score and test data
Predict using the best model

Section 2

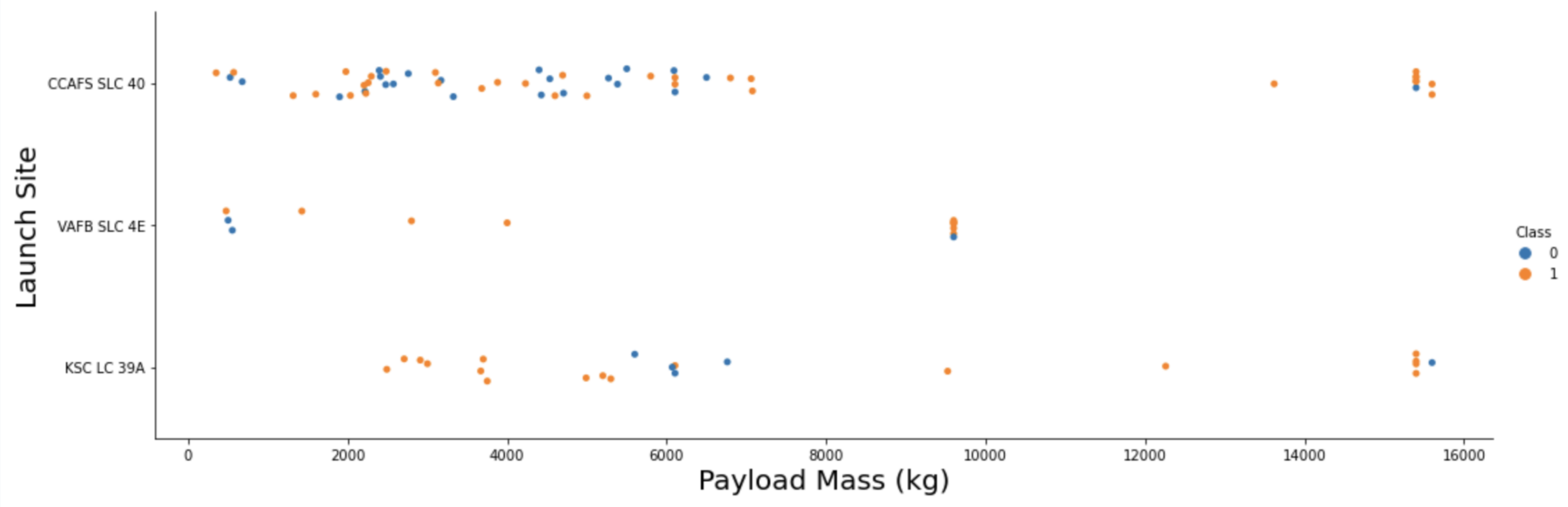
Insights drawn from EDA

Flight Number vs. Launch Site



- The main Launch Site is CCAFS SLC 40
- The number of successful landings increased with the Flight Number (With time)

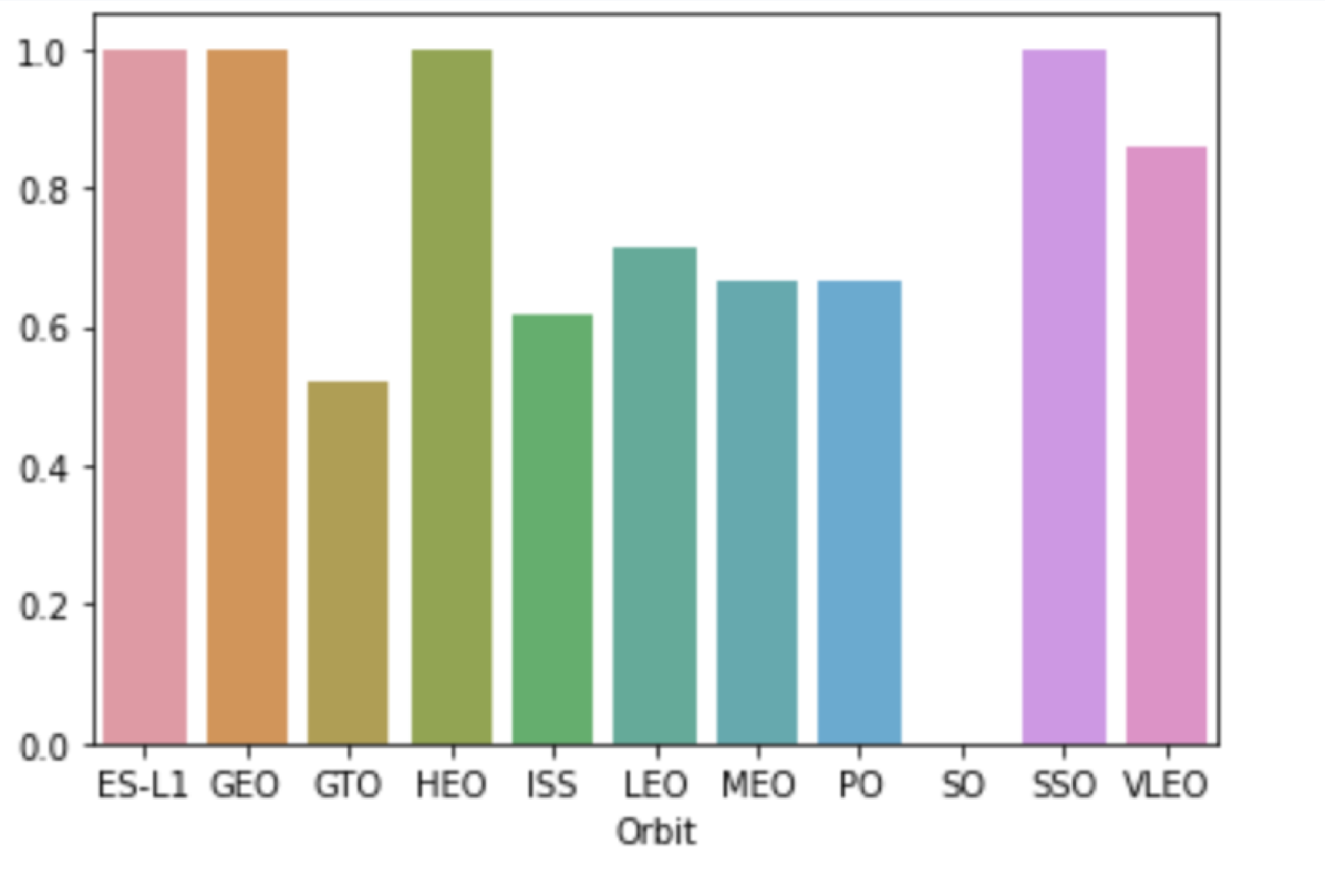
Payload vs. Launch Site



- For CCAFS SLC 40 the heavier the payload, the higher change of successful landing
- For KSC LC 39A is a region around 6000 Kg were failure is max
- The best launching site is VAFB SLC 4E

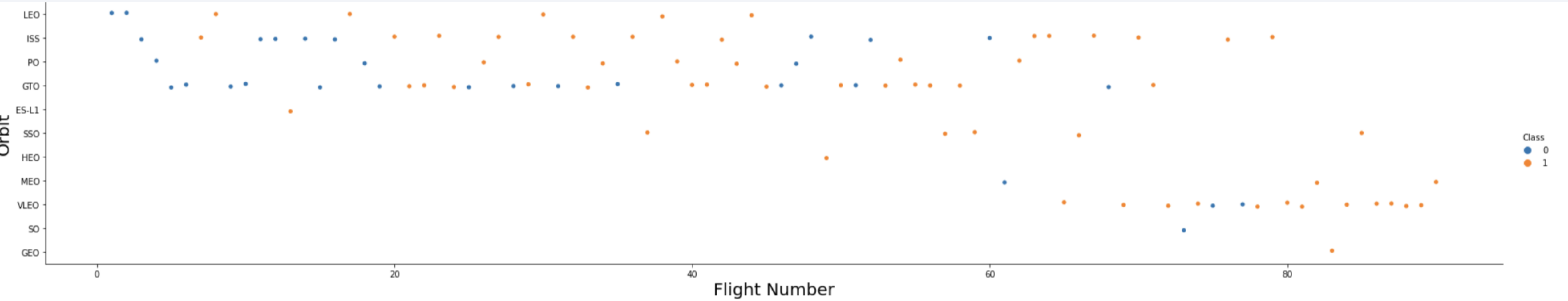
Success Rate vs. Orbit Type

- The lowest success rate for landing the first stage is for GTO
- Seems like the safest orbits are ES-L1, GEO, HEO and SSO but…

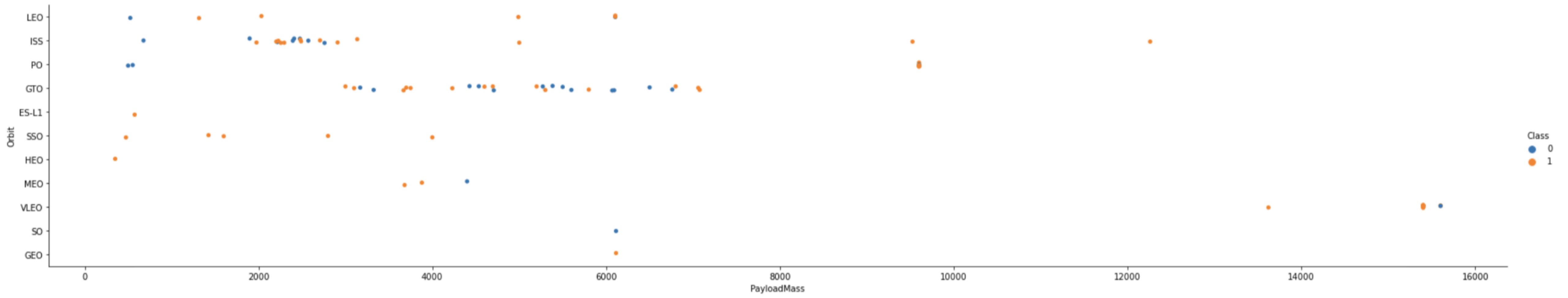


Flight Number vs. Orbit Type

- The orbits that seem to be the safest in the previous plot, are the ones with less launches, This means that those look like the safest launching site for landing but the amount of data is not significative



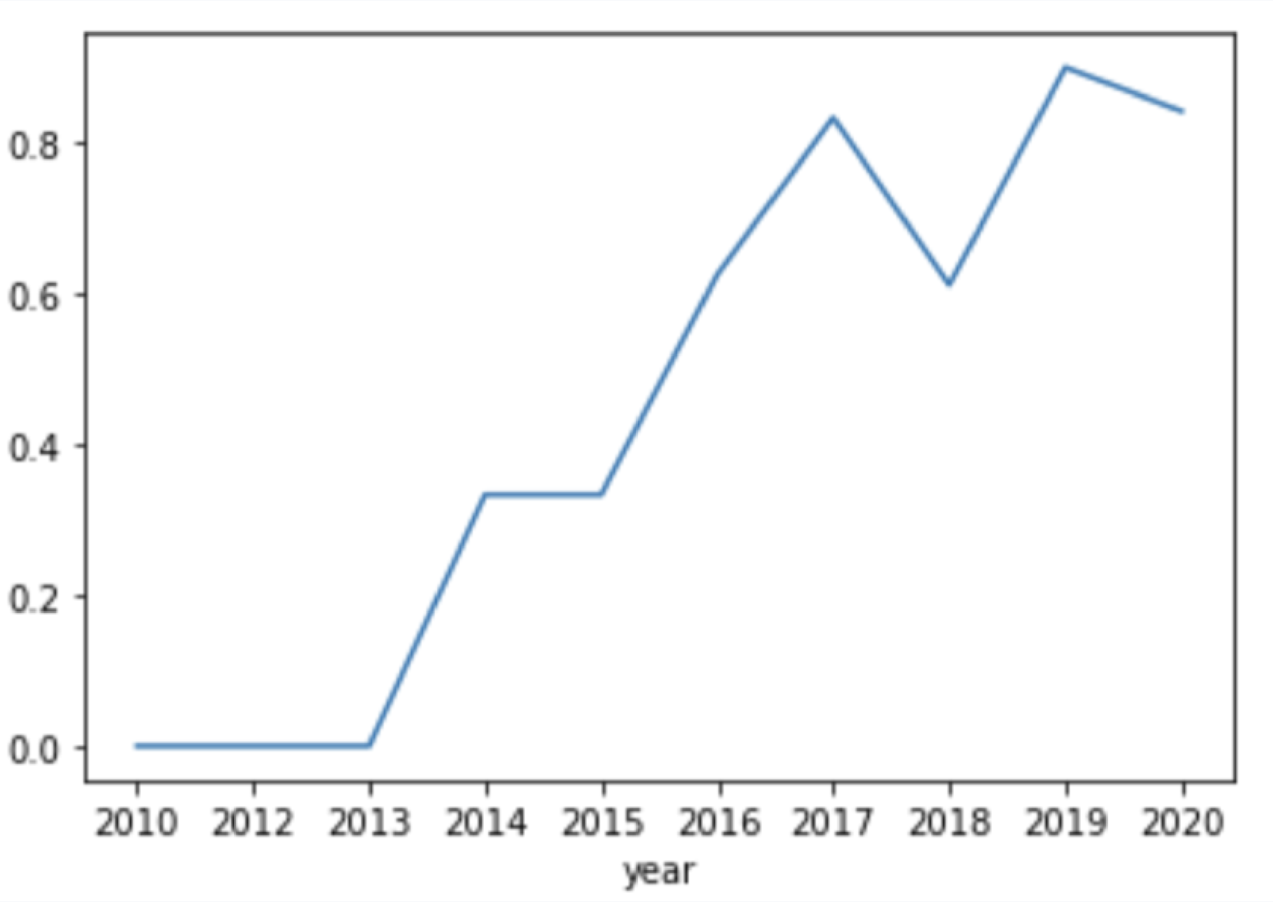
Payload vs. Orbit Type



- Starting from 8000 kg of payload, the success increase. This could be because the technology was good enough for launching higher payloads and therefor the landing was more successful

Launch Success Yearly Trend

- As time foes one the success of landing keeps increasing, as mentioned in the previous slide.



All Launch Site Names

- A list of all launch sites can be obtained using DISTINCT.
- 4 sites are returned. For the previous plots the CCAFS LC-40 and CCAFS SLC-40 were taken as a single one

```
%%sql
    SELECT DISTINCT LAUNCH_SITE
    FROM SPACE_X
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%%sql
    SELECT *
    FROM SPACE_X
    WHERE LAUNCH_SITE LIKE 'CCA%'
    LIMIT 5
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	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

- Total payload carried by boosters from NASA (CRS)

```
%%sql

    SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
    FROM SPACE_X
    WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb618
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

```
%%sql

    SELECT AVG(PAYLOAD_MASS__KG_)
    FROM SPACE_X
    WHERE BOOSTER_VERSION LIKE 'F9 v1.1%%'
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46e
Done.
```

1
2534

First Successful Ground Landing Date

- The first successful la landing took place on 2015/10/22

```
%%sql

SELECT MIN(DATE)
FROM SPACE_X
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1f
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql

SELECT DISTINCT BOOSTER_VERSION
FROM SPACE_X
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculation of the total number of successful and failure mission outcome

```
%%sql

SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME)
FROM SPACE_X
GROUP BY LANDING__OUTCOME
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb61
Done.
```

landing__outcome	2
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

```
%%sql

SELECT BOOSTER_VERSION
FROM SPACE_X
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACE_X)
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nc
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

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

```
%%sql

SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACE_X
WHERE LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = 2015
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41c
Done.
```

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

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

- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

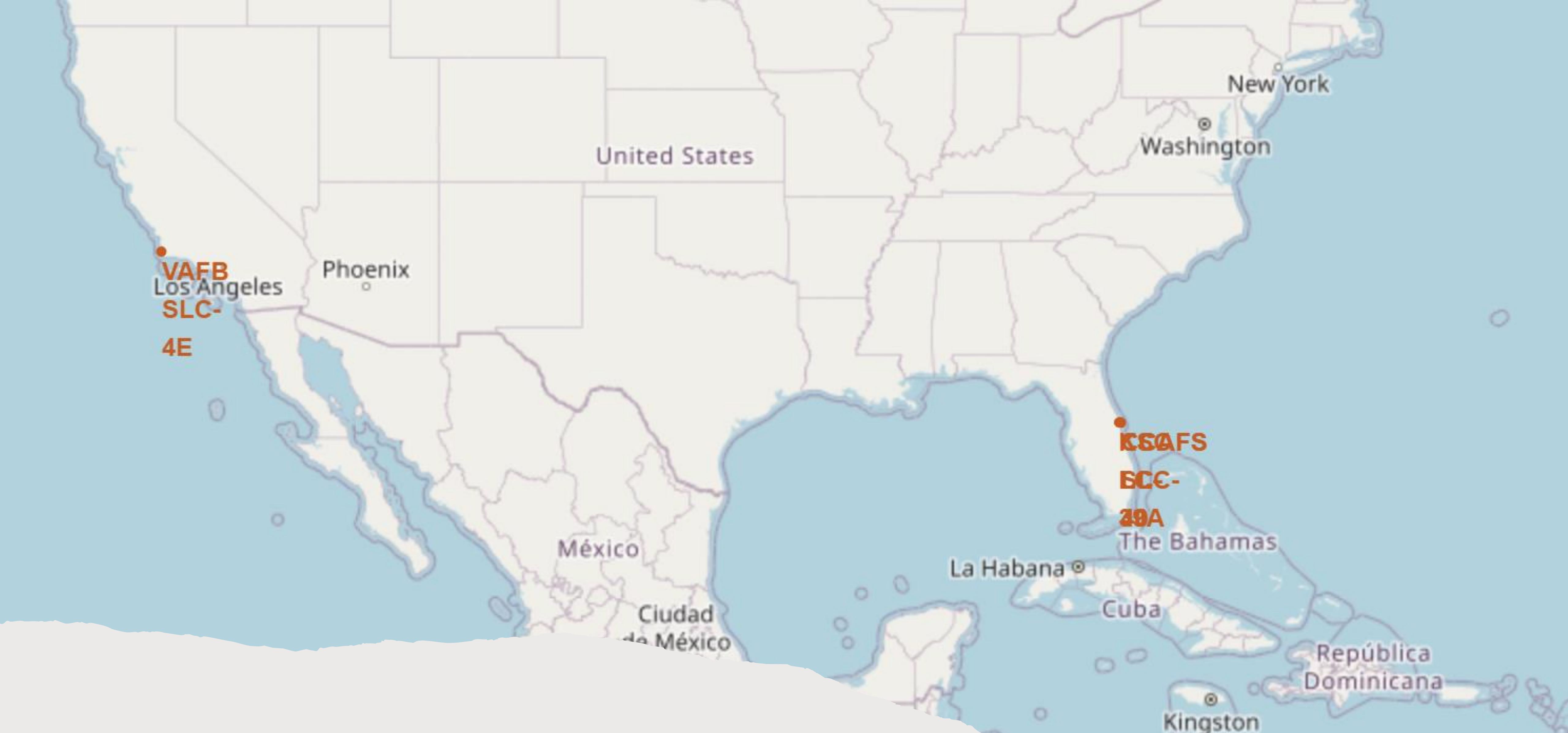
```
%%sql

SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT_LANDING_OUTCOME
FROM SPACE_X
GROUP BY LANDING__OUTCOME
ORDER BY COUNT_LANDING_OUTCOME DESC
```

```
 * ibm_db_sa://dcv80908:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnr
Done.
```

landing__outcome	count_landing_outcome
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

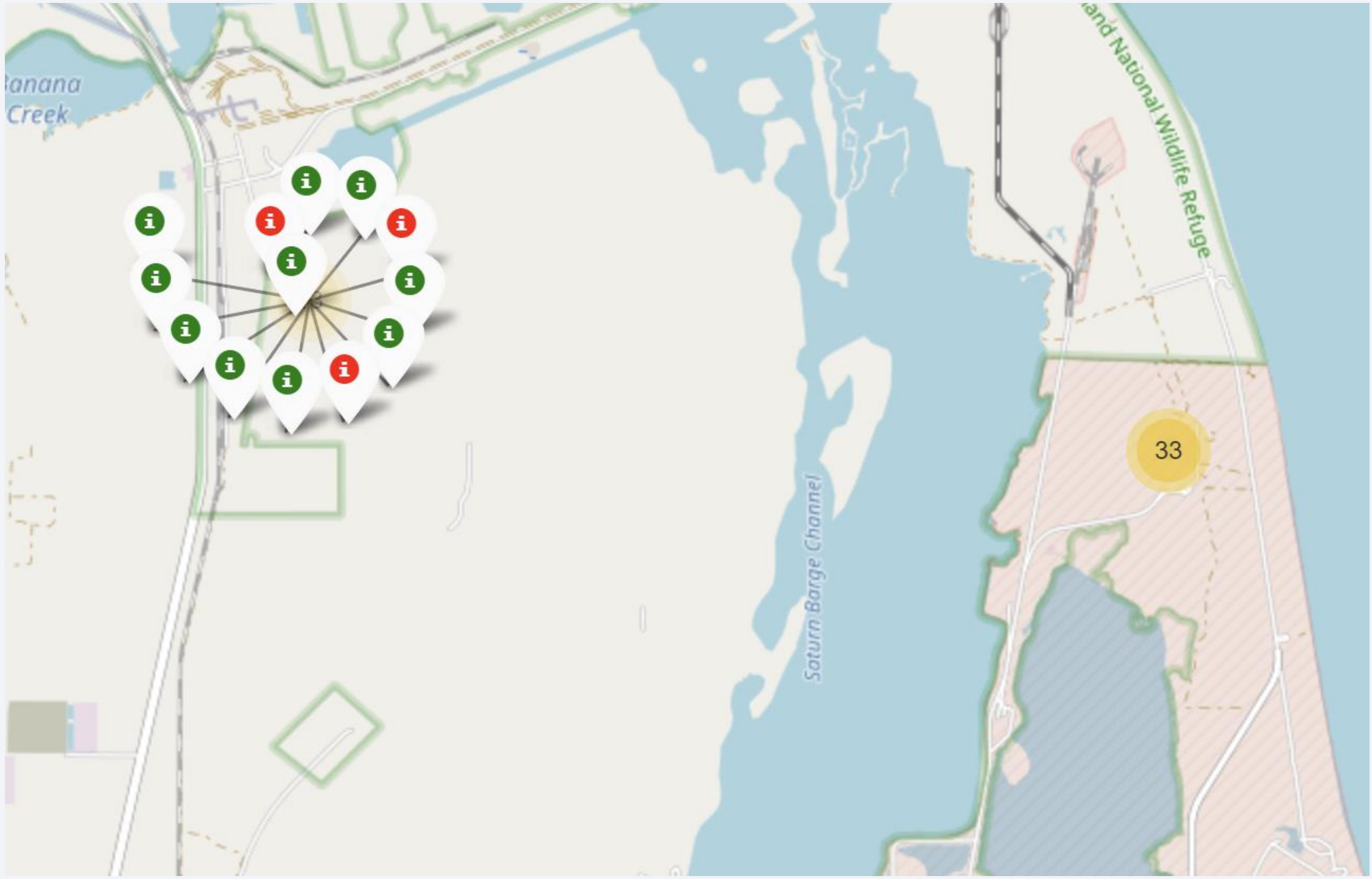
Launch Sites Proximities Analysis



Location of Launching sites

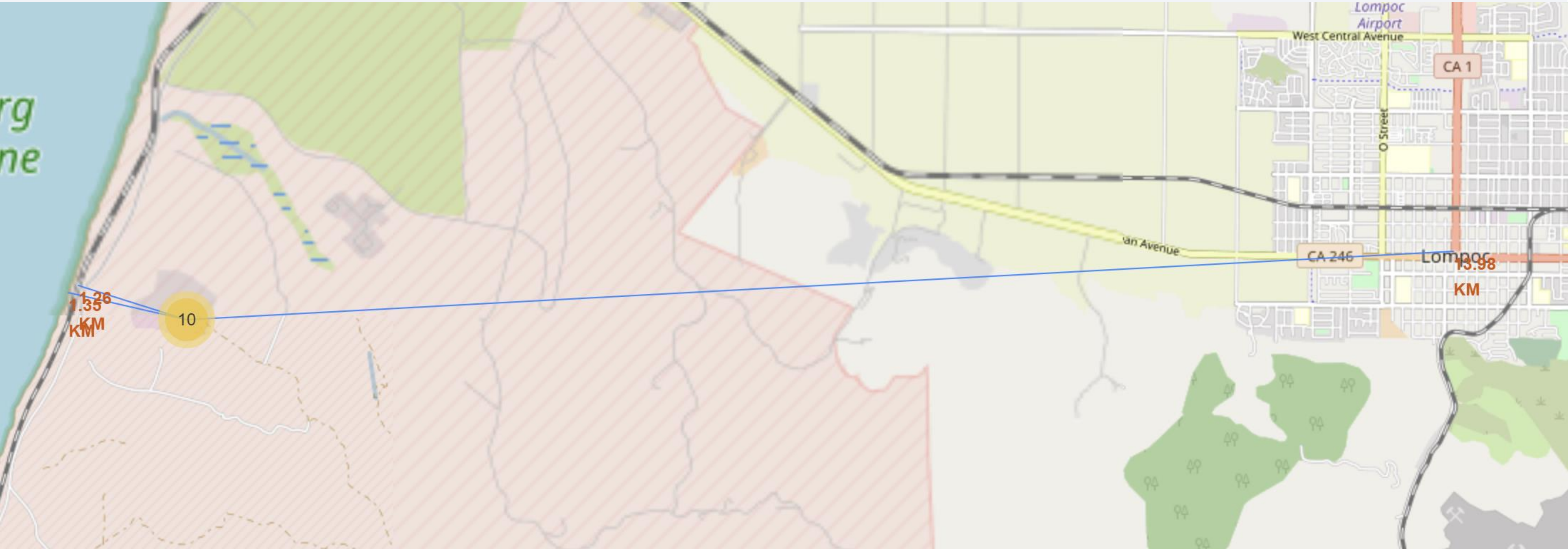
Visualize the success

- Zoom into the east launching sites and the group of markers showing in green the successful landing, and in red the unsuccessful ones.



Relevant close locations

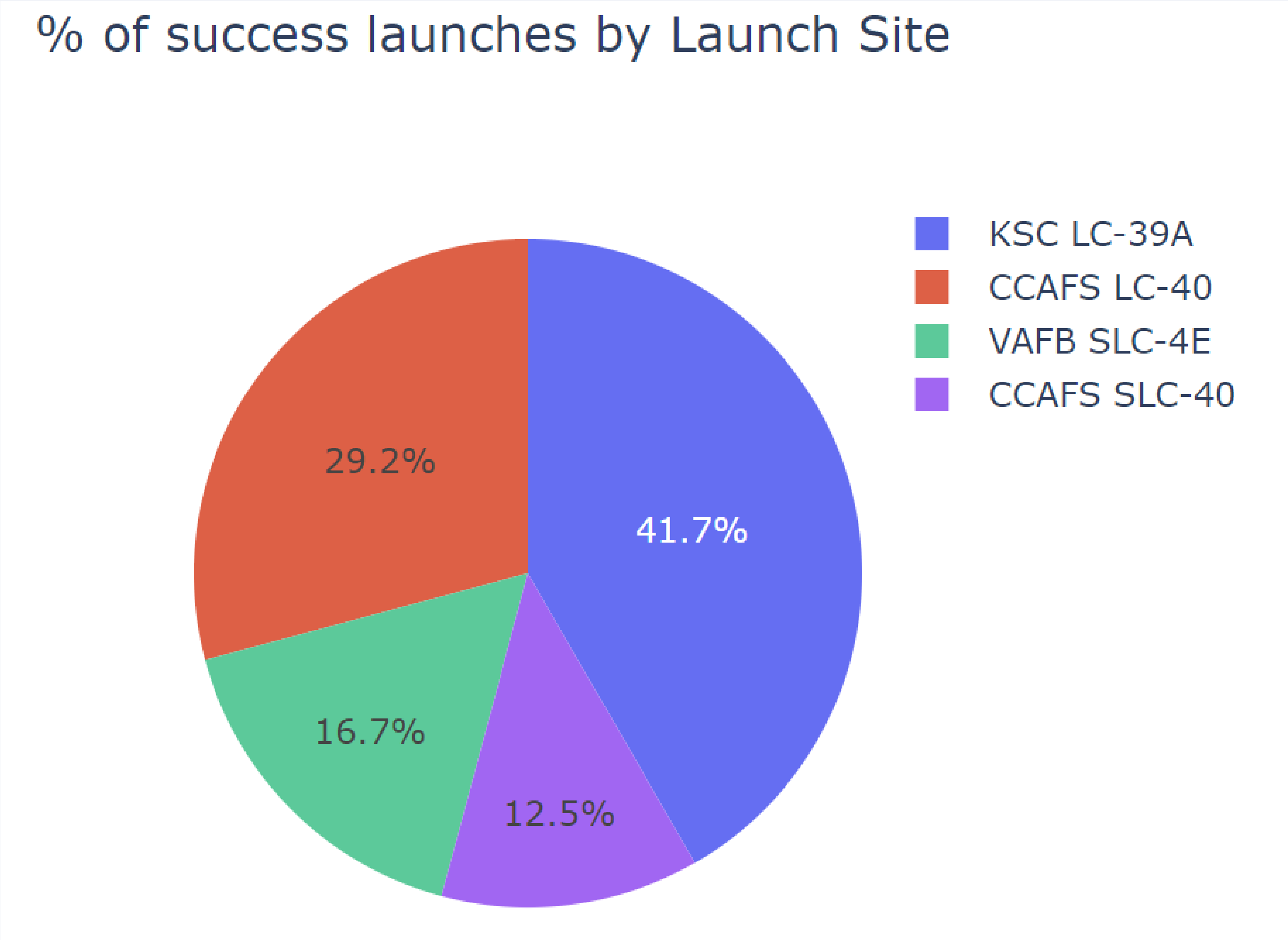
- Closest coastline, railway and city



Section 4

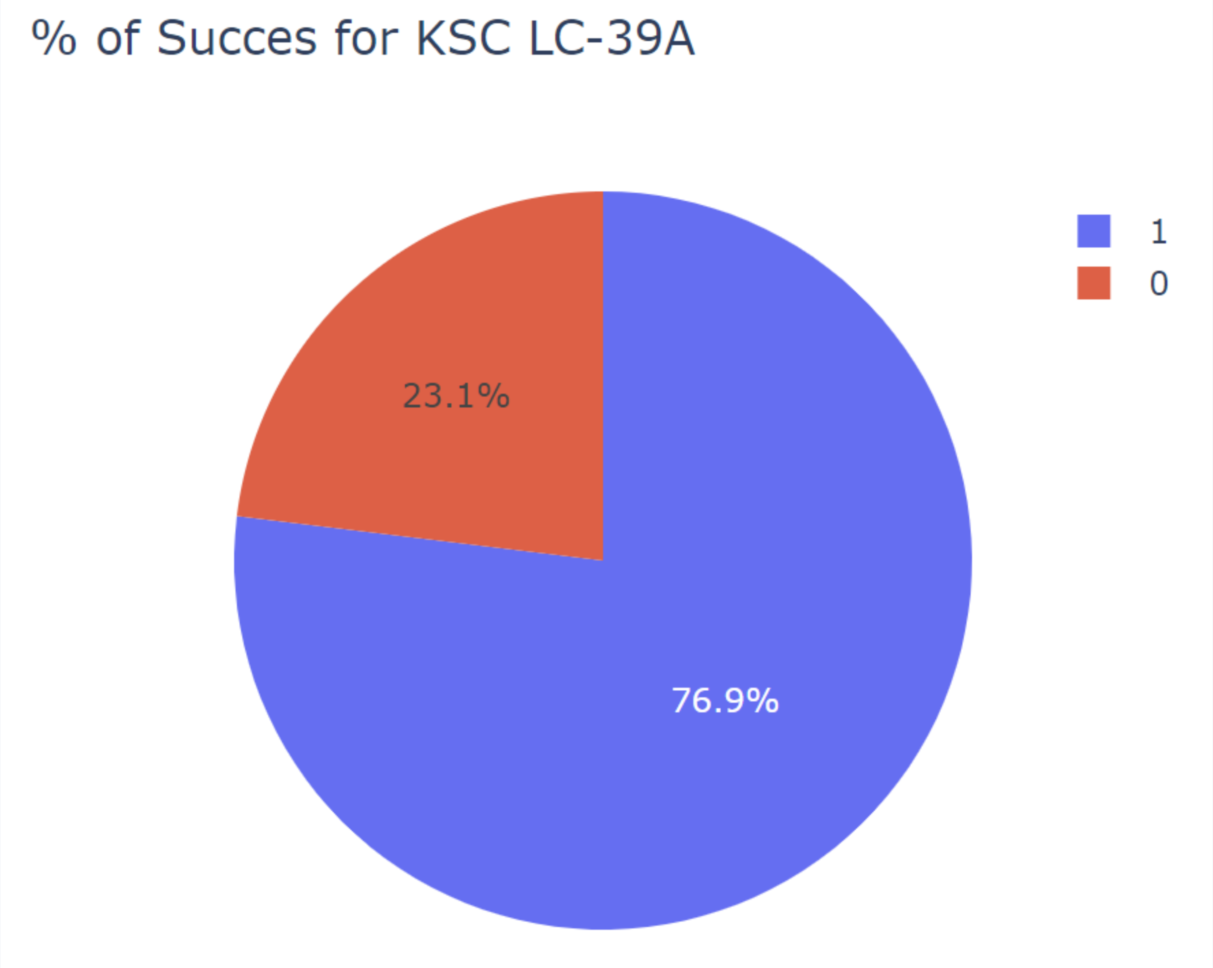
Build a Dashboard with Plotly Dash

Total Launches for each site

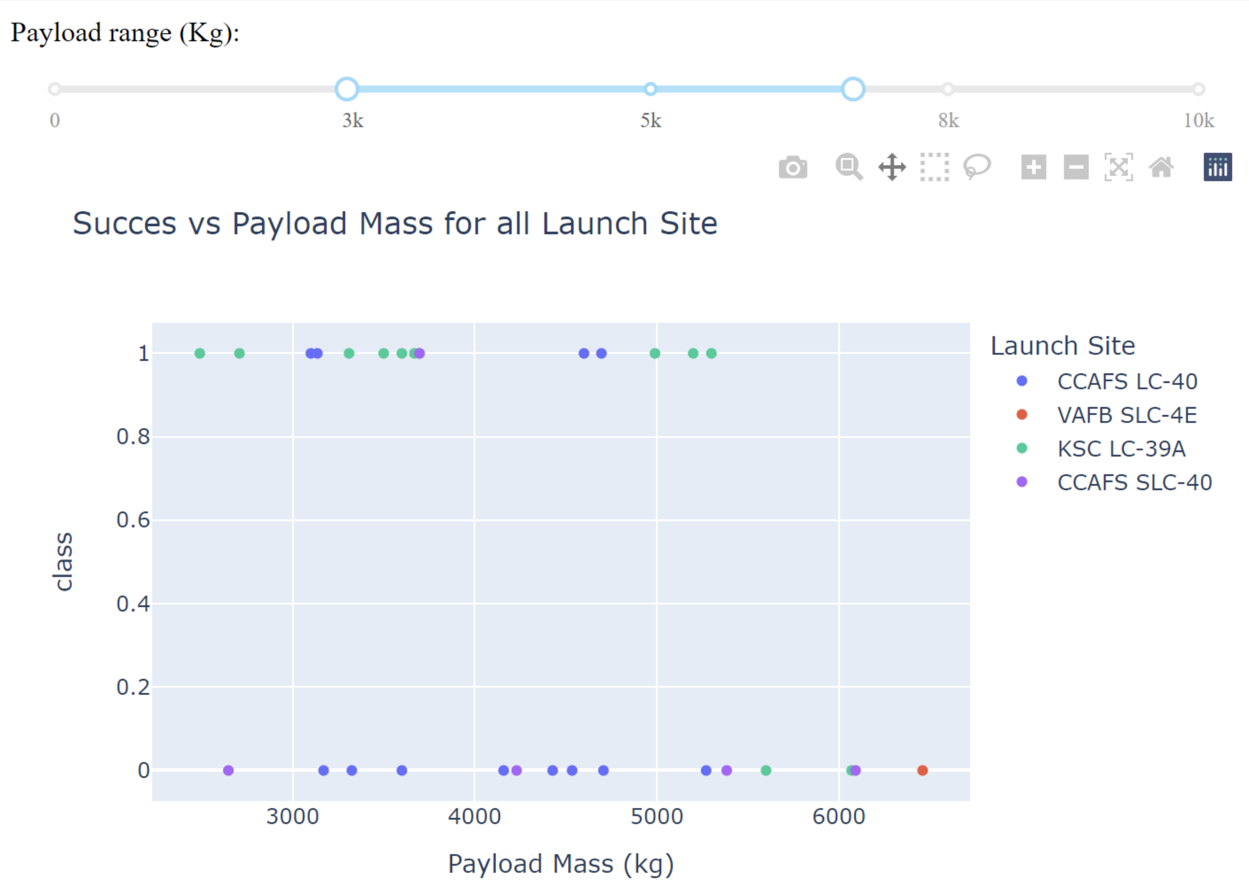


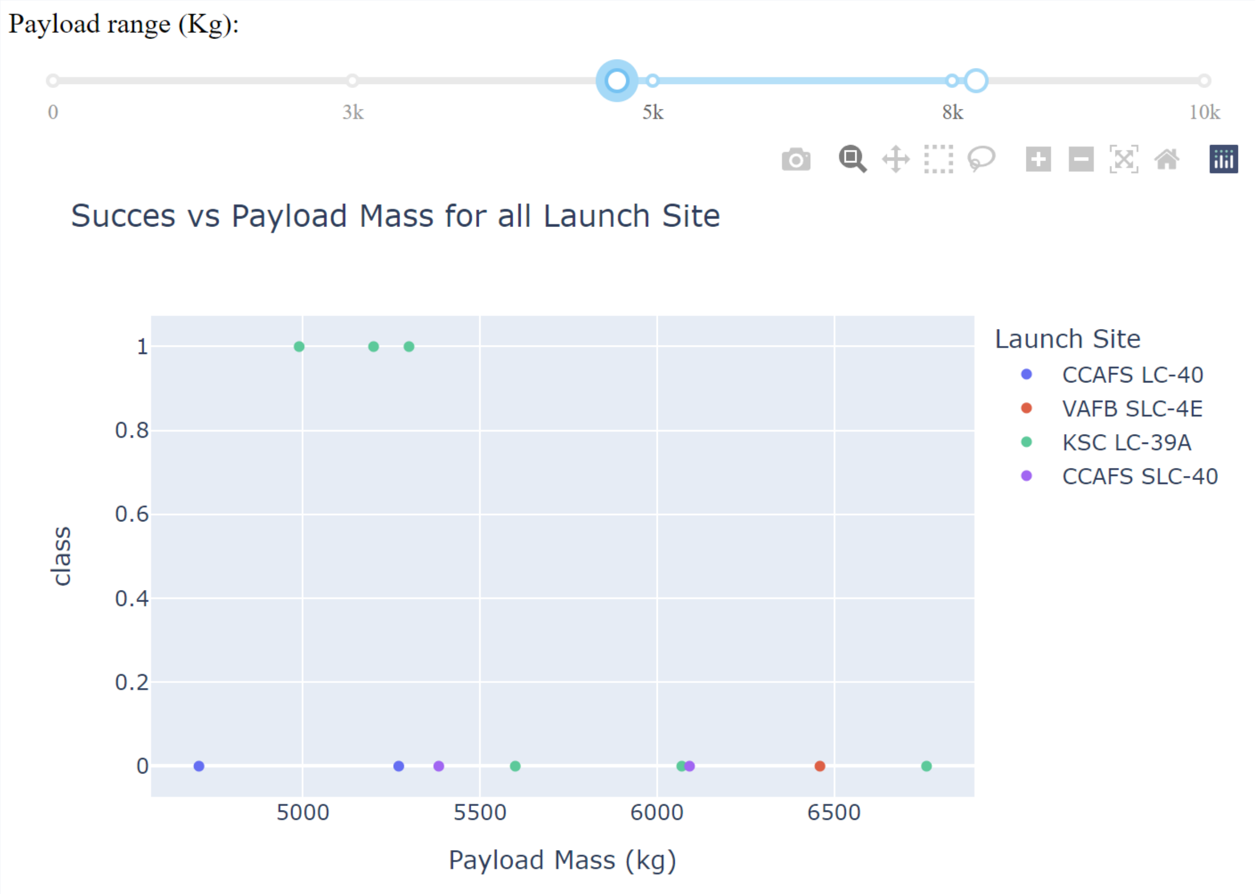
Most Successful launch site by rate

The best launch site by success rate is KSC LC 39A



Class vs payload mass slider



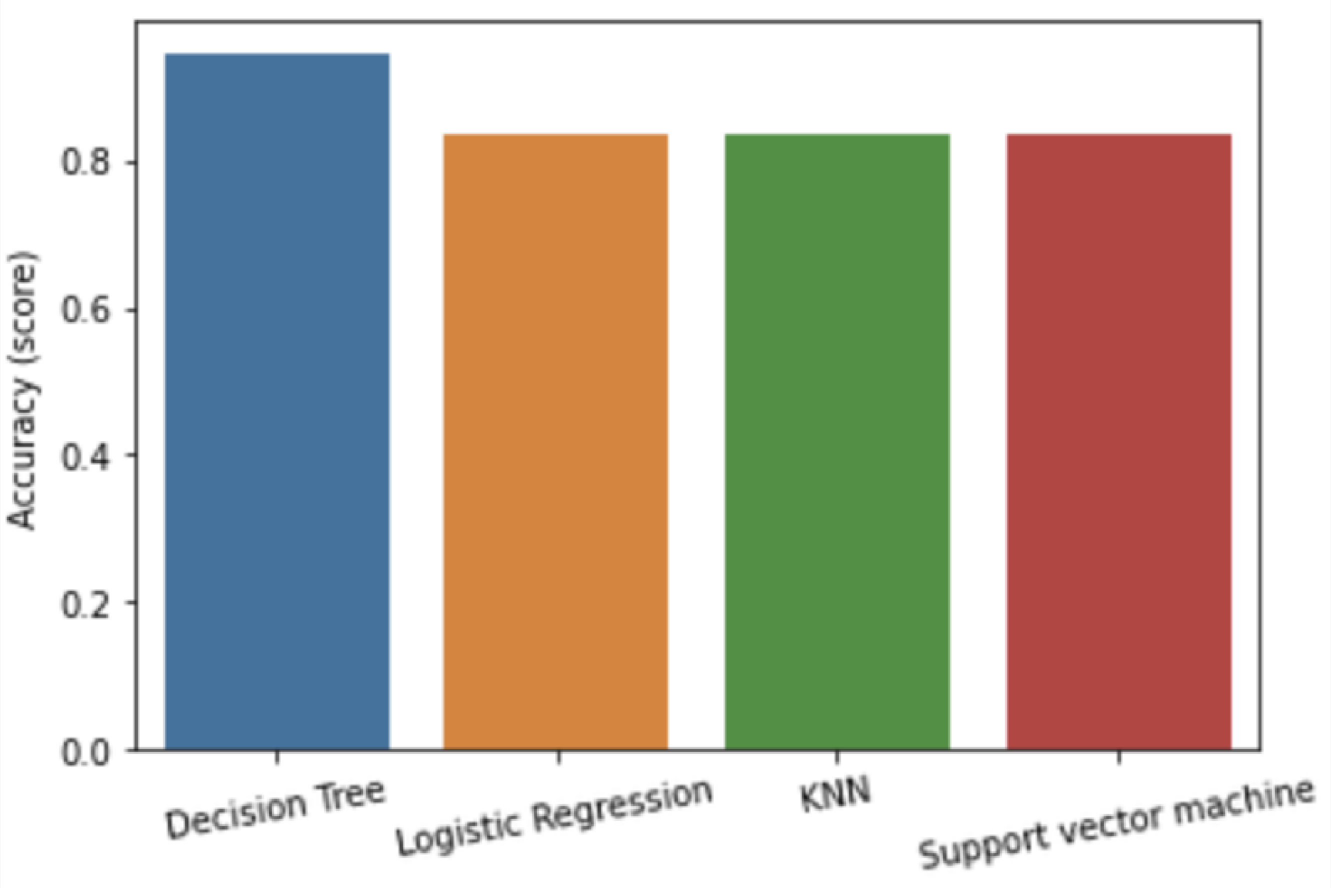


Section 5

Predictive Analysis (Classification)

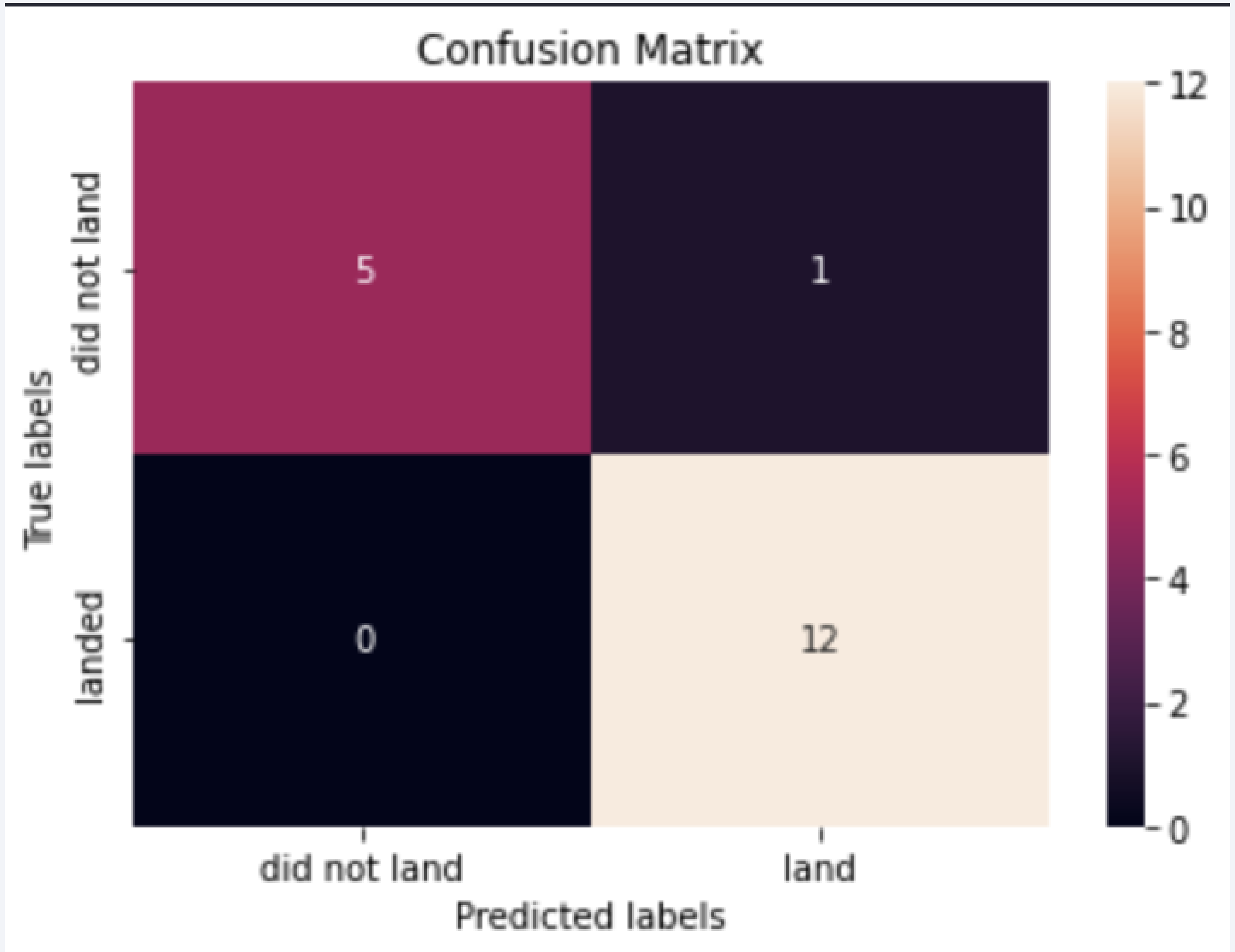
Classification Accuracy

- By calculating the score for each model was found that the best model is the Decision Tree



Confusion Matrix for Decision Tree

- The confusion matrix shows a really good performance for classification, the only wrongly classified label is a did not land label classified as land.
- The Decision Tree improved respect to the other models since the other models had a higher false positive rate



Conclusions

- The most significant variables in the success or failure of a landing are the Launching site and the year it was made. Clearly as technology advances is expected to have a better performance
- A model for less time could improve the accuracy of the model, since the advance in technology change the over all behavior of the data
- The best model of the ones tested is the Decision Tree Classifier, with an accuracy of 0.94 over 1

Appendix

- https://github.com/dfortizgu/space-Y-project

Thank you!