Winning Space Race with Data Science

Eduardo Schiavo
3/11/2021

Outline

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

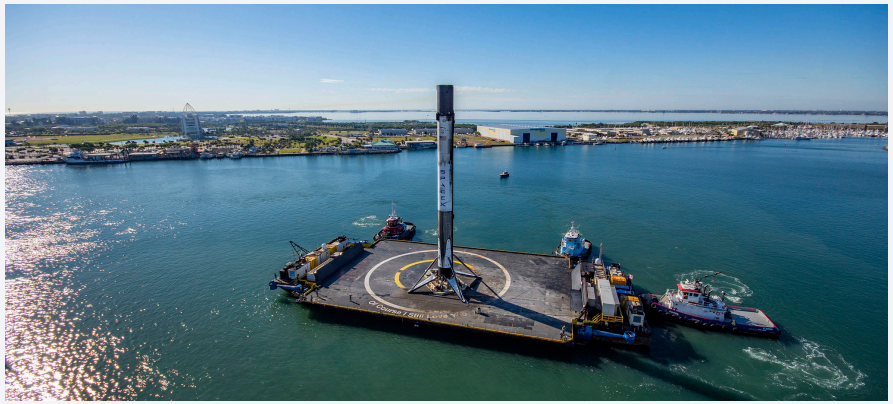
Executive Summary

- Exploratory Data Analysis is carried out on a Space X dataset (via webscraping, interactive visualization, dashboard building, geographical data visualization etc.)
- Different Data Visualization techniques allow us to rationalize what are the factors affecting the successful landing of the first stage.
- This has a direct economical impact, as the possible reutilization of the first stage is the main factor that makes Falcon 9 launches much cheaper that all the other space programs

Introduction

A Falcon 9 Launch from Space X costs **62$ million**

Launches from other companies cost > **165$ million**



The main difference stems from the fact that Falcon 9 can reuse the first stage…

…if it lands without blowing up

We can investigate which factors affect the correct landing of the First Stage in order to increase the number of successful landings!

Section 1

Methodology

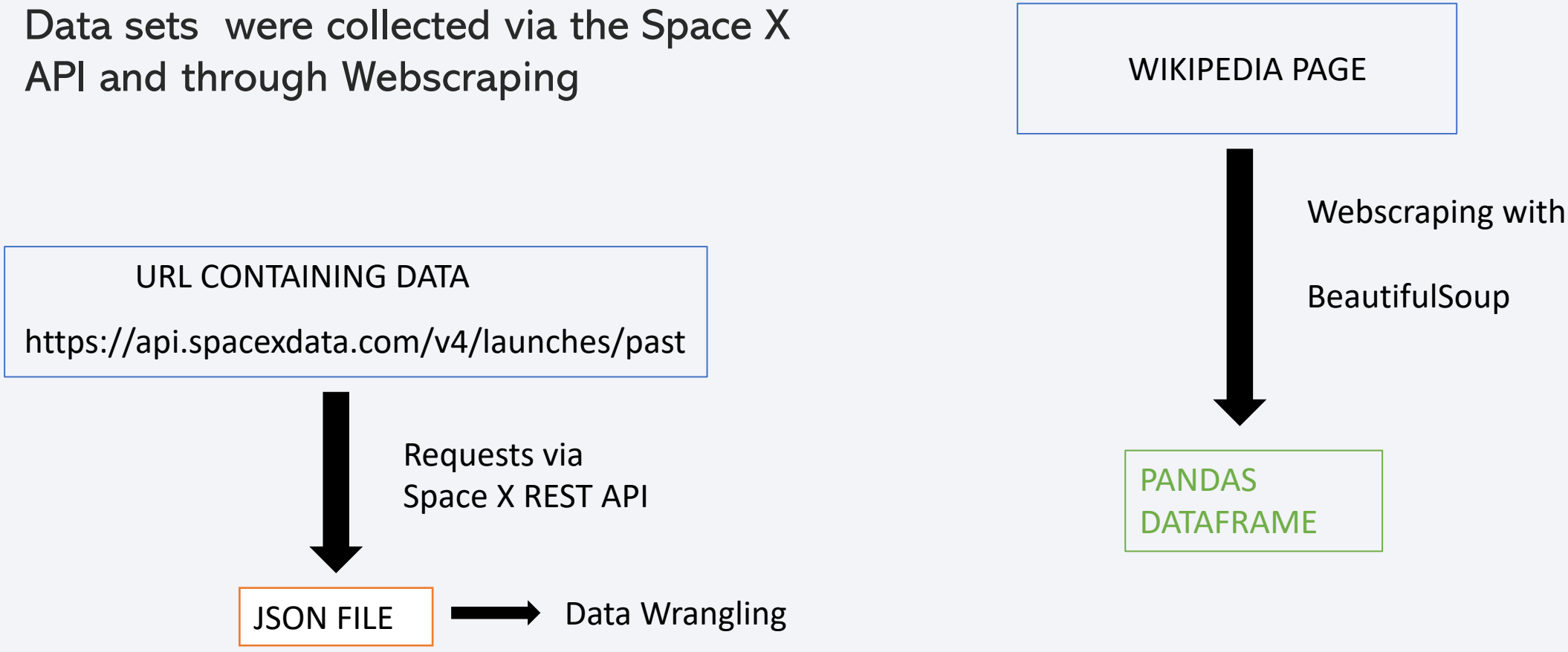
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected mainly through requests to the Space X API
- Perform data wrangling
  - A landing Class with values 0/1 for unsuccessful/successful landing was created
  - Am Orbit column was added, containing the target orbit for the Payload.
- 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 sets were collected via the Space X API and through Webscraping

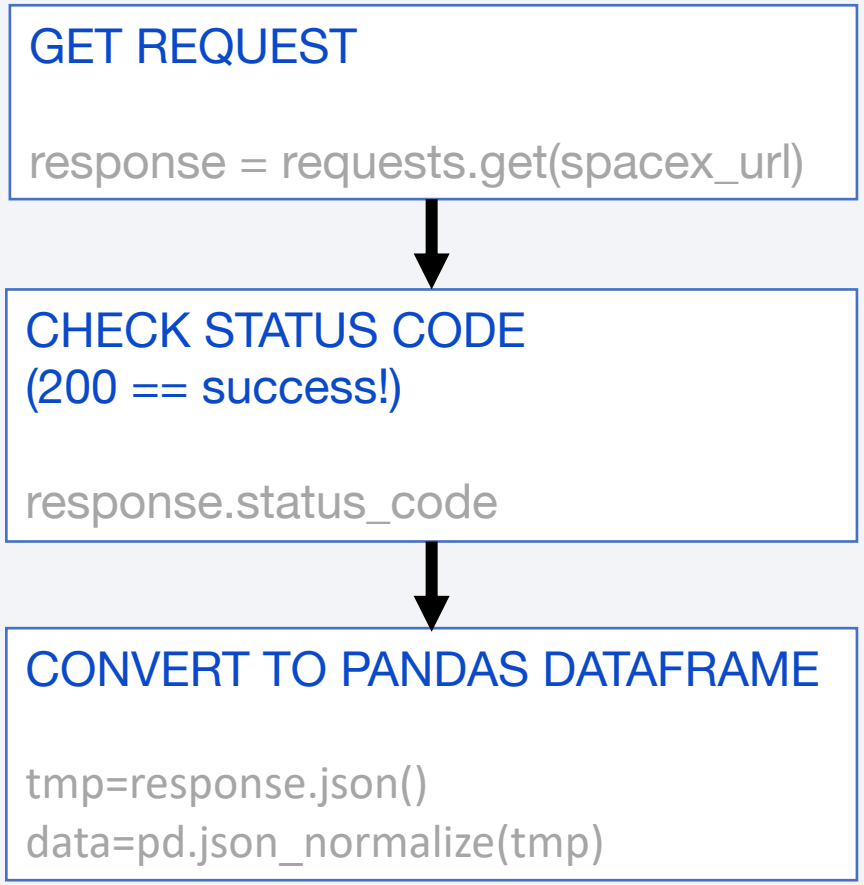


Data Collection – SpaceX API

Scheme of Space X REST API Data collection

Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/DataCollactionAPI.ipynb

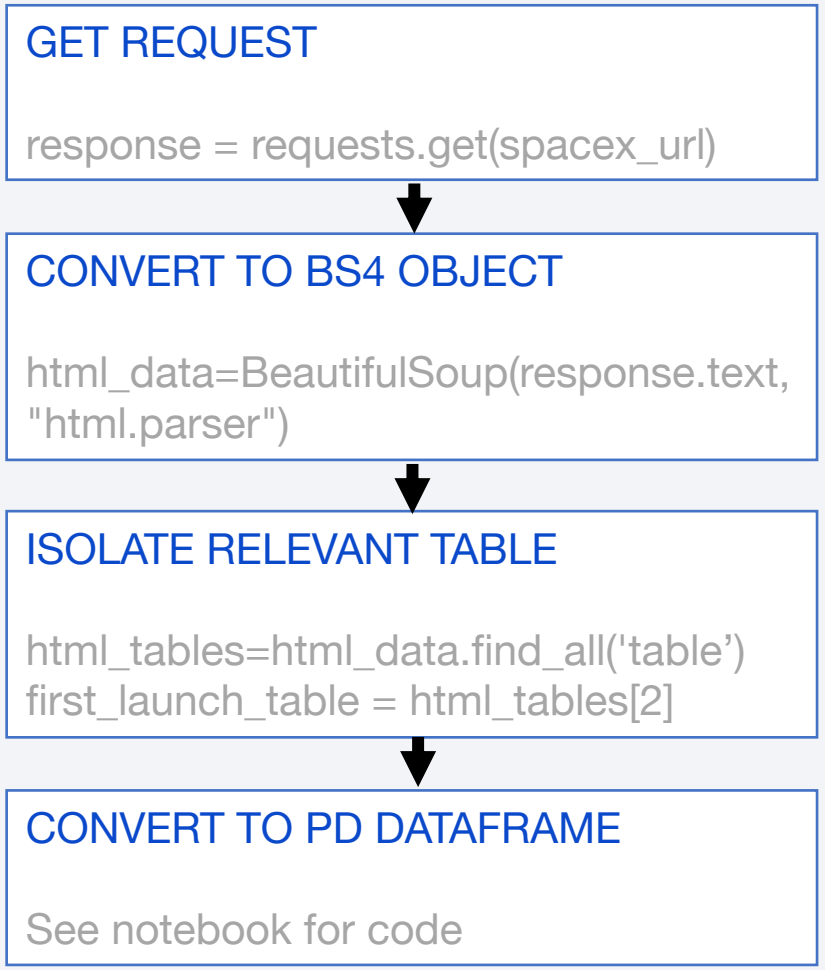


Data Collection - Scraping

Webscraping scheme

Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/Webscraping.ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) is performed in order to find patterns in the data

(i.e. the number of launches per orbit and per site are counted)

- The dataset is processed in order to have suitable labels for the subsequent training of supervised machine learning (ML) models

(i.e. the mission outcomes are converted in a Class entry)

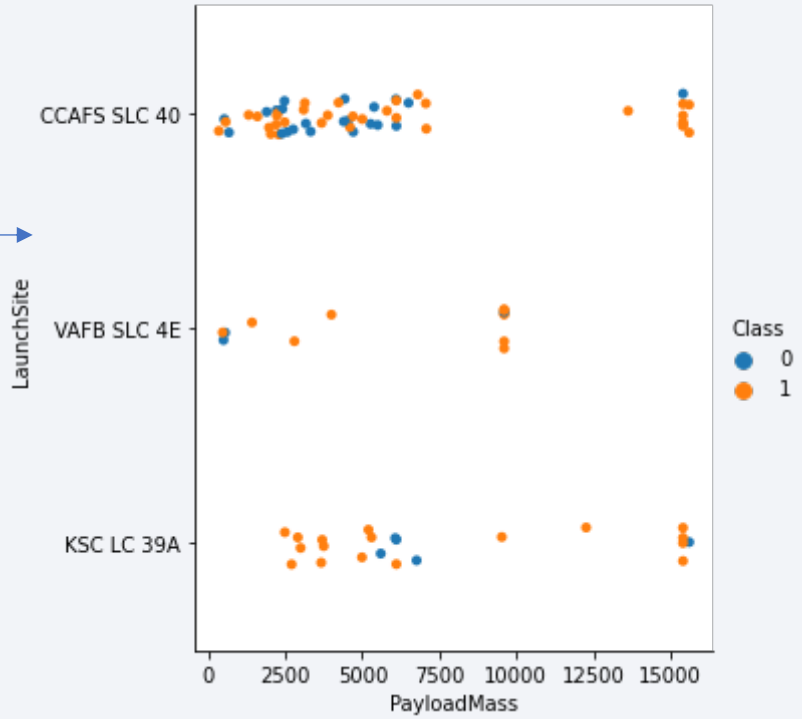
Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/EDA.ipynb

EDA with Data Visualization

The following relationships are visually explored:

- Launch Sites vs. Flight Number
- Launch Sites vs. Payload Mass
- Orbit Type vs. Success Rate
- Orbit Type vs. Flight Number
- Orbit Type vs. Payload Mass
- Class vs. Year



Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/EDA%20with%20DATA%20Visualization.ipynb

EDA with SQL

EDA was performed through SQL queries, in particular:

- Launch site names
- Launches from sites starting with CCA
- Payload mass carried by NASA CRS boosters
- Average payload mass carried by F9 v1.1 boosters
- Date of first successful landing outcome
- Name of boosters with succesfuls outcome and payload mass in a certain range

EDA with SQL

- Total number of successes and fails
- Booster versions that have carried out the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranked 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

• Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

On a map we marked:

-launch sites

-success and fails for launch sites

-distances between a launch site to its proximities

This allows us to visualize the sites with the highest rate of success and to assess the risk due the proximities of human settlements and or infrastructures. See notebook for further details

Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/EDA%20with%20SQL.ipynb

Build a Dashboard with Plotly Dash

A simple dashboard was created, visualizing the outcomes for launch site and the outcome against payload mass, color coded with respect to the site. A dropdown menu allows for selection of the site and a slider helps us select the payload mass range.

GitHub Link to the python code:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/83b78656bf8eeede3a8f6cfa84b4c472d650a17a/spacex_dash.py

Predictive Analysis (Classification)

A Class label was created, data was standardized and split into a training and test set.

Through GridSearchCV we found the best parameters set for:

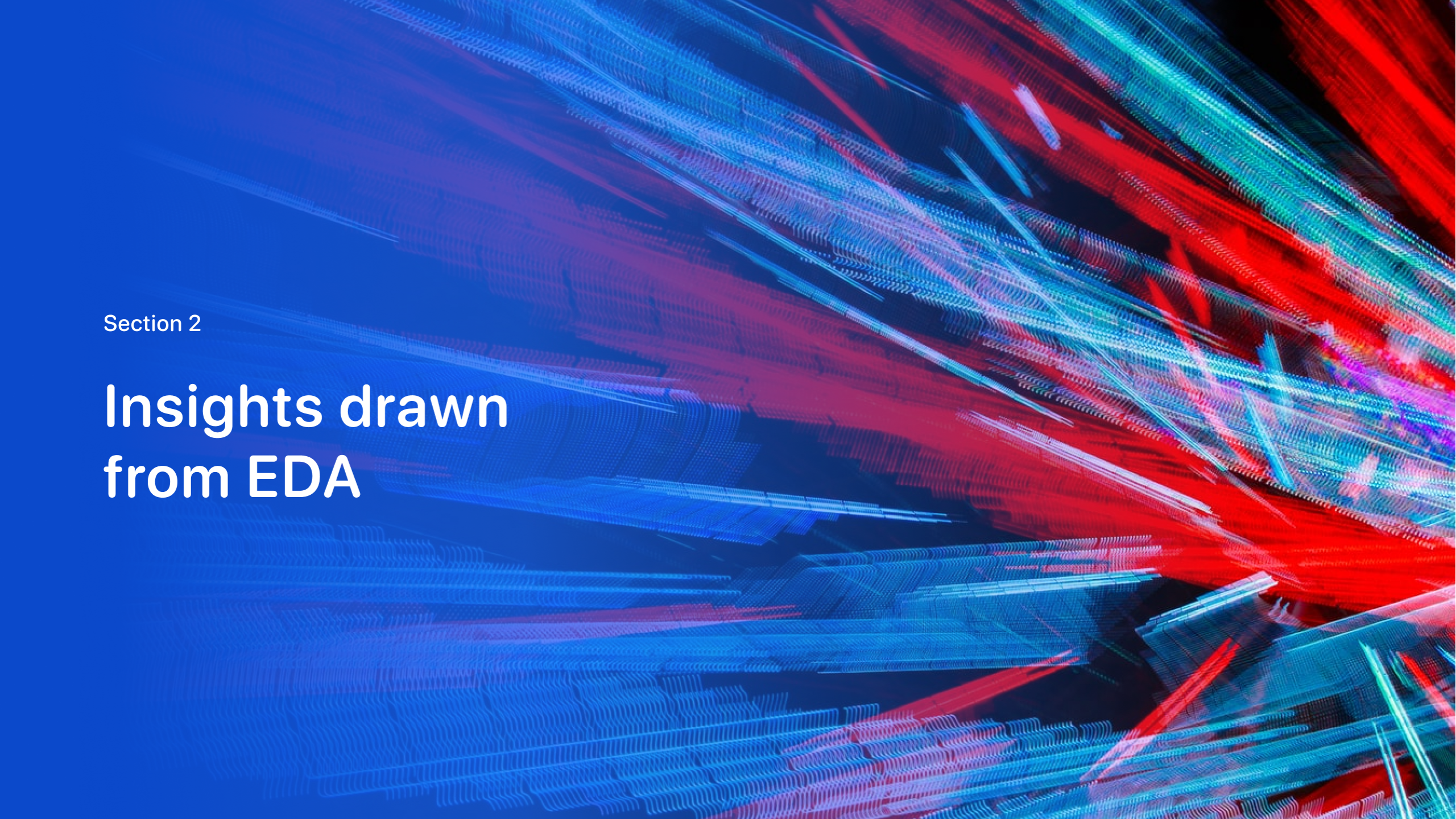
Logistic Regression, Decision Tree, SVM and KNN

Notebook link:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone/blob/596da586cda3137638b1ab75b20b009296b9e271/Machine%20Learning%20Prediction.ipynb

Results

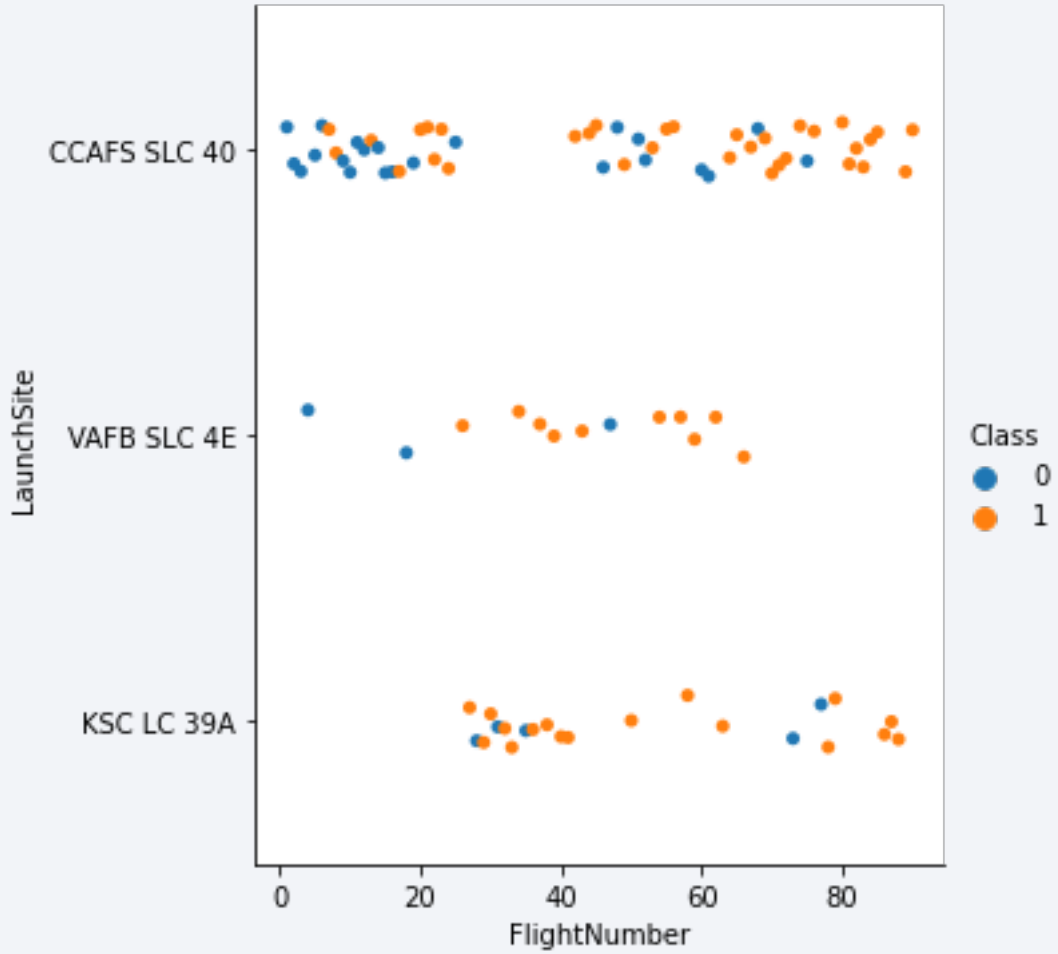
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

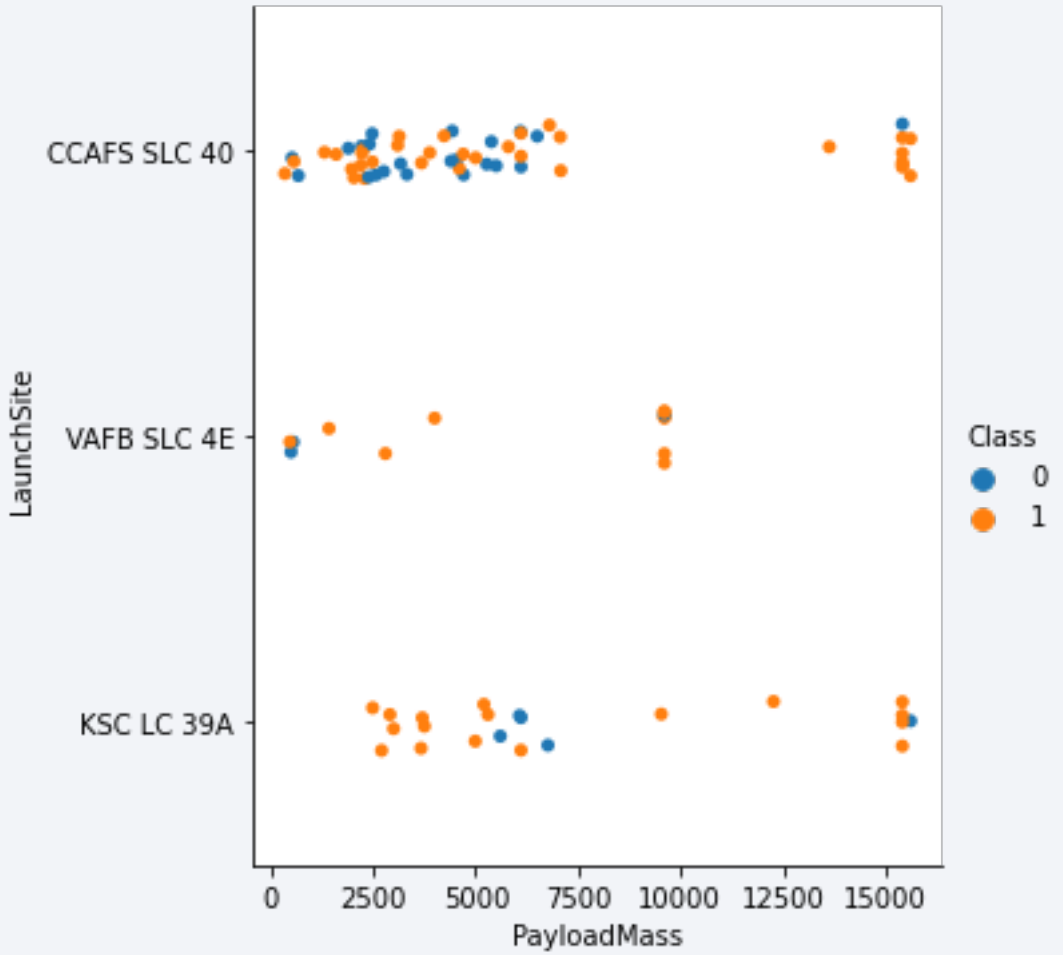


For CCAFS SLC 40 there seems to be a correlation between Flight Number and outcome

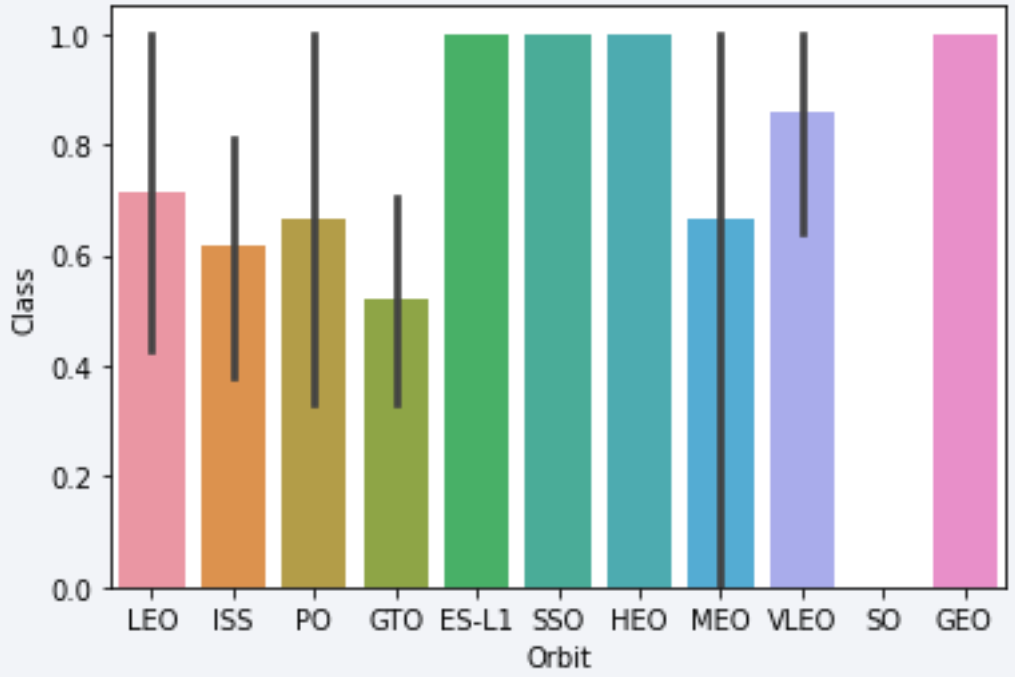
Payload vs. Launch Site

For VAFB SLC 4E there are no launches with payload mass greater than 10000

In general, there seems to be a higher success rate for larger payloads across all the Lauch Sites



Success Rate vs. Orbit Type



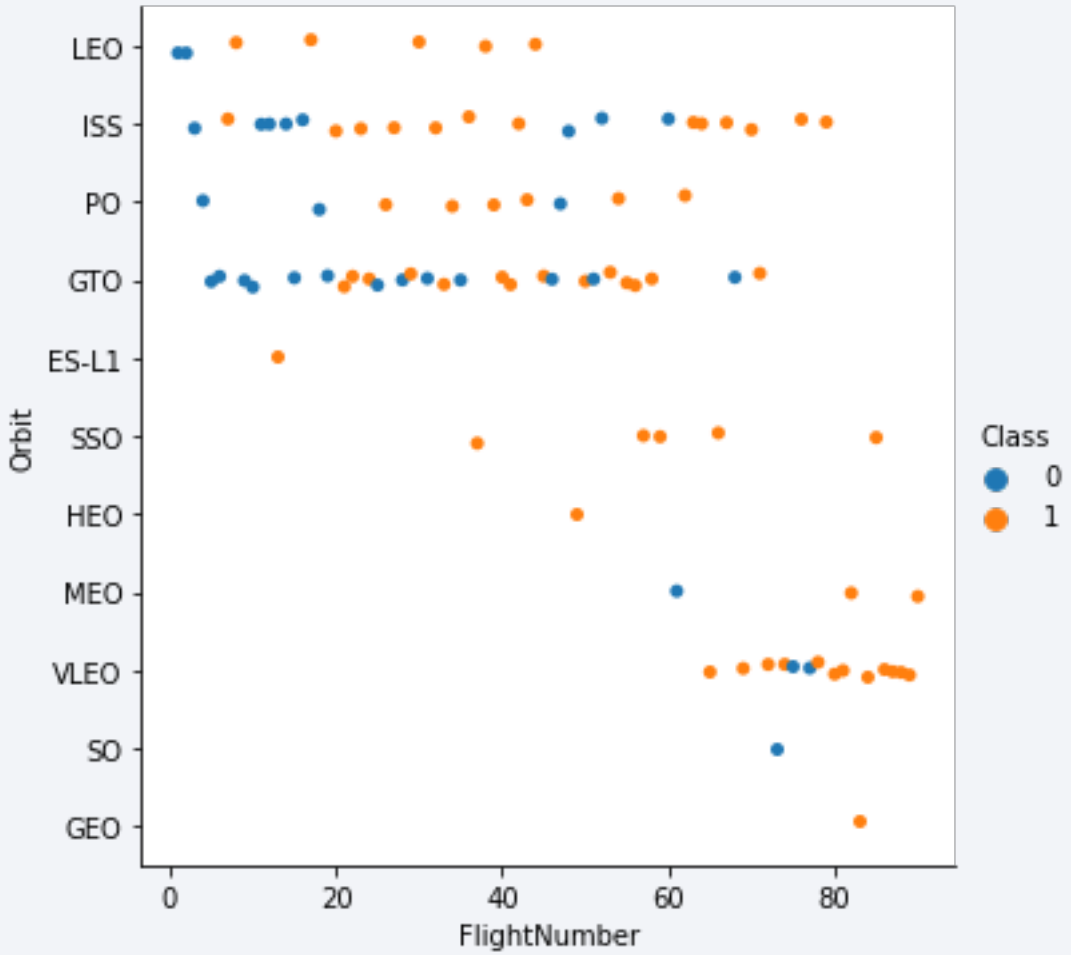
ES-L1 SSO HEO and GEO have the highest success rate

Flight Number vs. Orbit Type

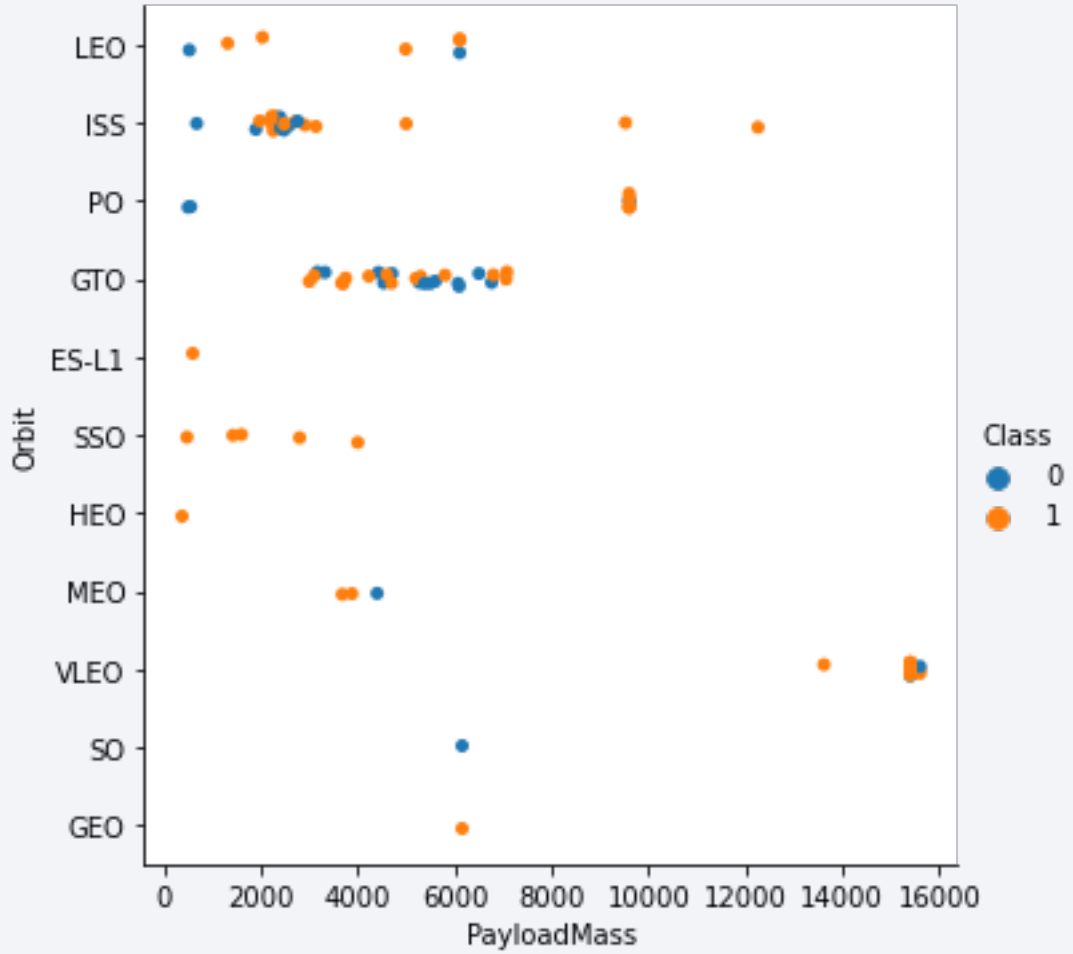
There is a relationship between number of flights and success rate for the LEO orbit

For other orbits (e.g. GTO) this is not the case.

For certain orbits we did not measure unsuccessful flights



Payload vs. Orbit Type



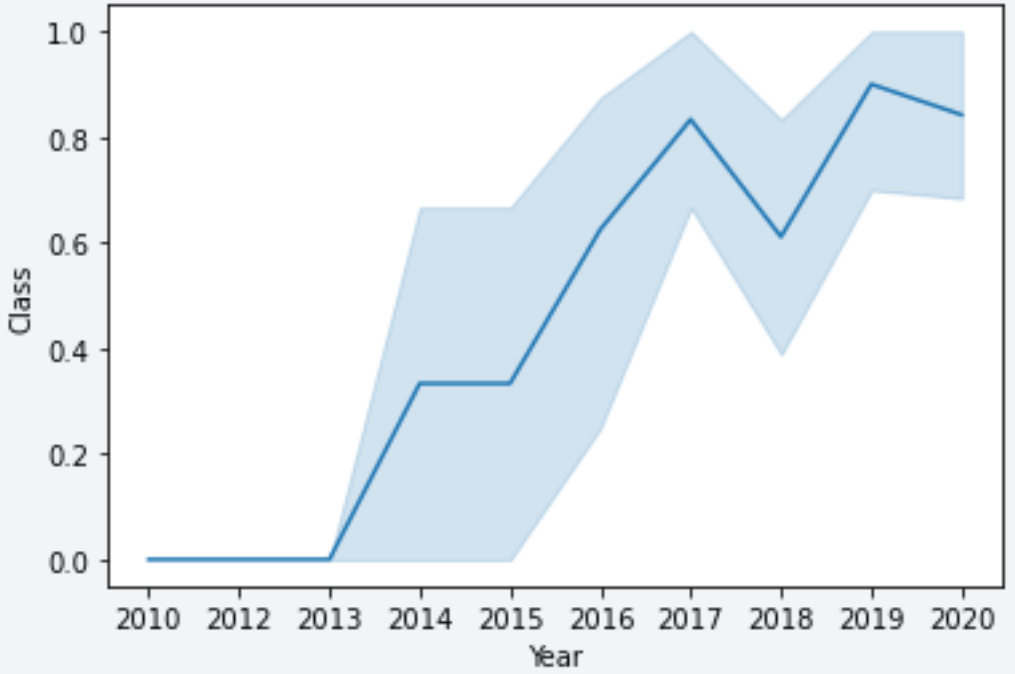
Again, for GTO we don't observe any trend

Heavier payloads are more frequently launched to the ISS, PO and VLEO orbits

On ISS and PO we also notice that heavier payloads correlate with a higher rate of success

Launch Success Yearly Trend

Launch success generally increases over time



All Launch Site Names

SQL query retrieving the launch sites from a table containing all the data relative to the launches

Display the names of the unique launch sites in the space mission

```
In [19]: %sql select unique(launch_site) from SPACEXTBL ;
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c
Done.
```

Out[19]:

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

Launch Site Names Begin with 'CCA'

SQL queries all the entries whose launch_site begins with ‘CCA’, then limits output to 5 entries

Display 5 records where launch sites begin with the string 'CCA'

In [22]:
```
%%sql select * from SPACEXTBL
where launch_site like 'CCA%'
limit 5;
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[22]:

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

Selecting the sum of the payload mass values of all the entries where customer is equal to 'NASA (CRS)'

Display the total payload mass carried by boosters launched by NASA (CRS)

In [29]:
```
%%sql
select SUM(payload_mass__kg_) as tot_payload from SPACEXTBL
where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a0
Done.
```

Out[29]:

tot_payload
45596

Average Payload Mass by F9 v1.1

Selecting the average of the payload mass values of all the entries where booster version starts with ‘F9 v1.1’

Display average payload mass carried by booster version F9 v1.1

In [30]:
```
%%sql
select avg(payload_mass__kg_) as avg_payload from SPACEXTBL
where booster_version like 'F9 v1.1%';
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c
Done.
```

Out[30]:

avg_payload
2534

First Successful Ground Landing Date

Querying the minimum of the Data column. The first successful landing occurred on the 04/06/2010

```
In [33]: %%sql select min(DATE) from SPACEXTBL
         where mission_outcome ='Success';

          * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-48
         Done.
```

Out[33]:

1
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

File display

```
%%sql
select booster_version from SPACEXTBL
where payload_mass__kg_>4000 and payload_mass__kg_<6000 and mission_outcome='Success';
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.dat
Done.
```

Out[37]:

booster_version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1

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

The screenshot is cropped to fit into the slides. See the Notebook for the full output.

Total Number of Successful and Failure Mission Outcomes

Missions are grouped my mission outcomes and outcomes are counted.

Note that the second line is commented out

List the total number of successful and failure mission outcomes

In [61]:
```
%%sql
select mission_outcome, count(mission_outcome) from SPACEXTBL
--where mission_outcome like 'Success%' or  mission_outcome like 'Failure%'
group by mission_outcome
;
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu
Done.
```

Out[61]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [63]:
```
%%sql
select booster_version, payload_mass__kg_        from SPACEXTBL
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
;
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.datab
Done.
```

Out[63]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600

Boosters carrying the maximum payload are queried. The maximum payload is obtained through a subquery to the SPACEXTBL

Not that the output is cropped to fit into the slides. See notebook for the full output

2015 Launch Records

SQL query to all failed llaunches in 2015

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

In [67]:
```
%%sql
select  DATE, booster_version, launch_site, landing__outcome from SPACEXTBL
where landing__outcome like 'Failure%' and DATE like '2015%'
;
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databa
Done.
```

Out[67]:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The landing outcome and its counts are queried and grouped by outcome then ordered in descending order

Rank 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

In [71]:
```
%%sql
select landing__outcome, count(landing__outcome) from SPACEXTBL
group by landing__outcome
order by count(landing__outcome) desc
;
```

```
 * ibm_db_sa://kzj02380:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

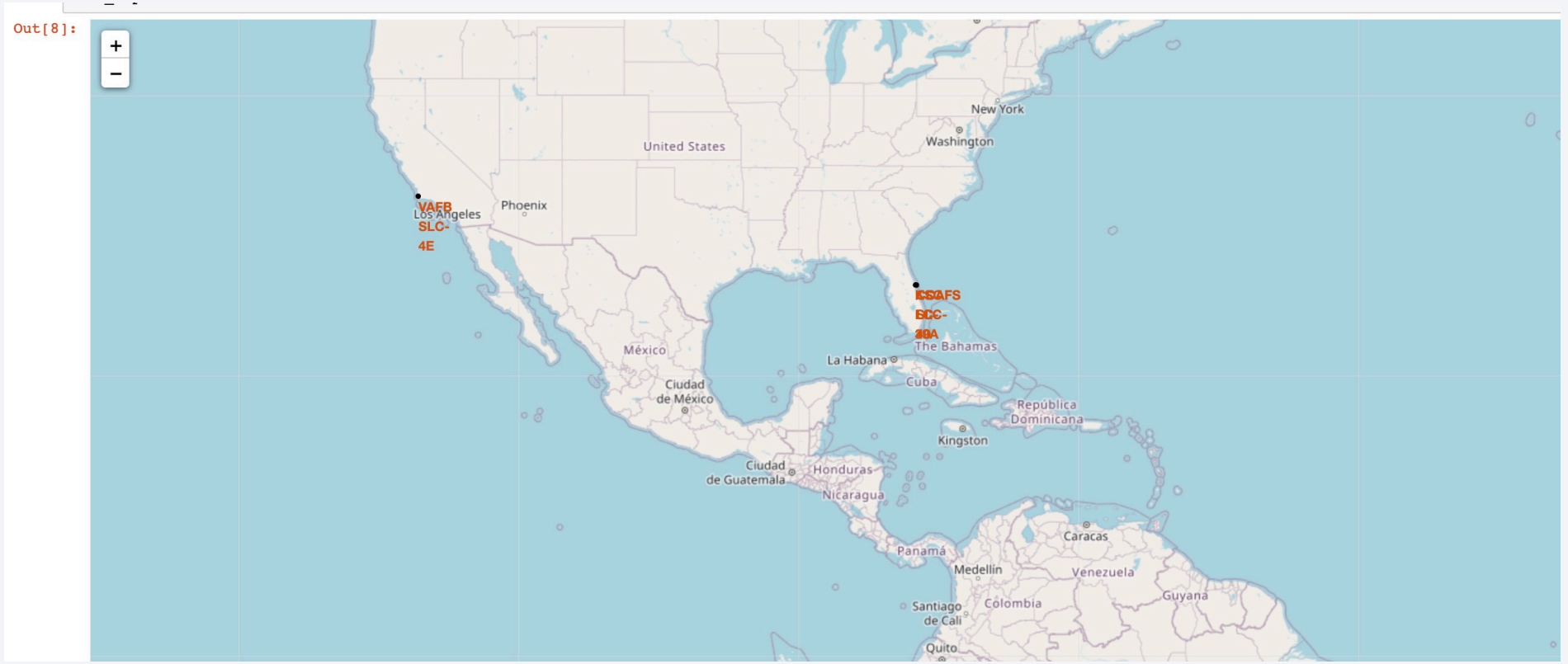
Out[71]:

landing__outcome	2
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 4

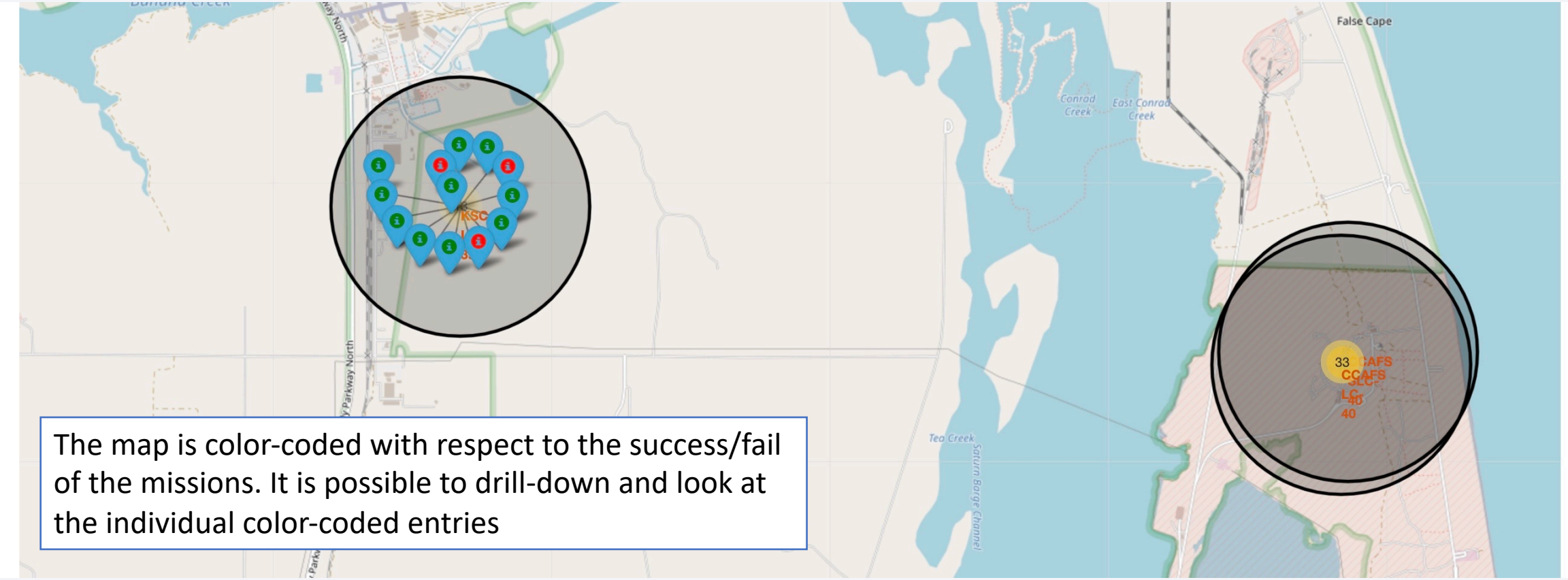
Launch Sites Proximities Analysis

Launch Sites' Locations



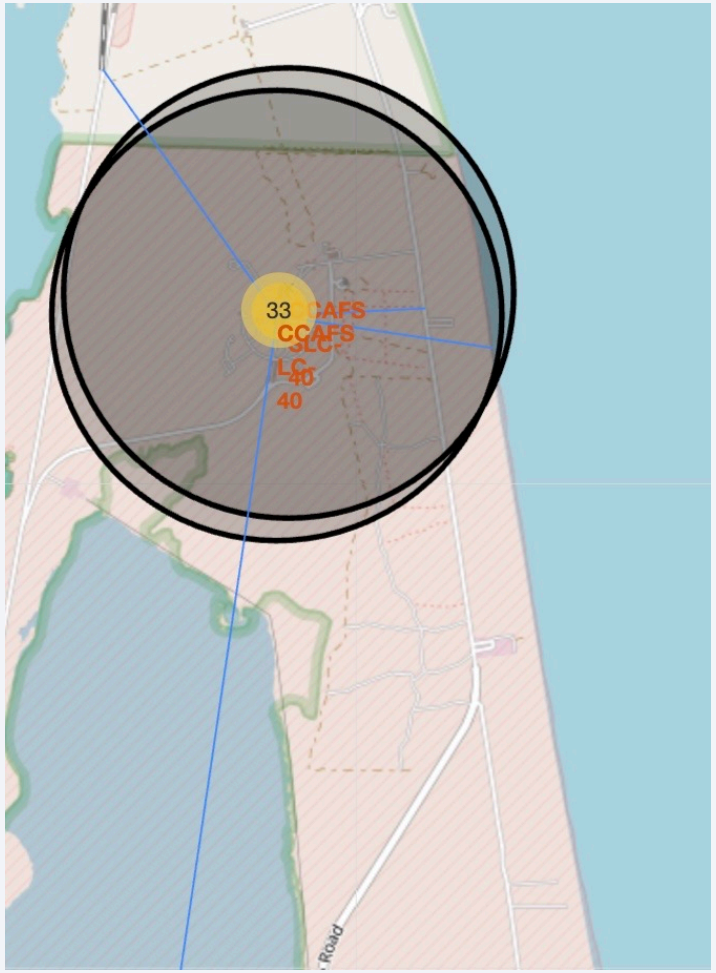
All the launch sites are located in coastal areas in proximity of the equator

Success rate of Launch sites



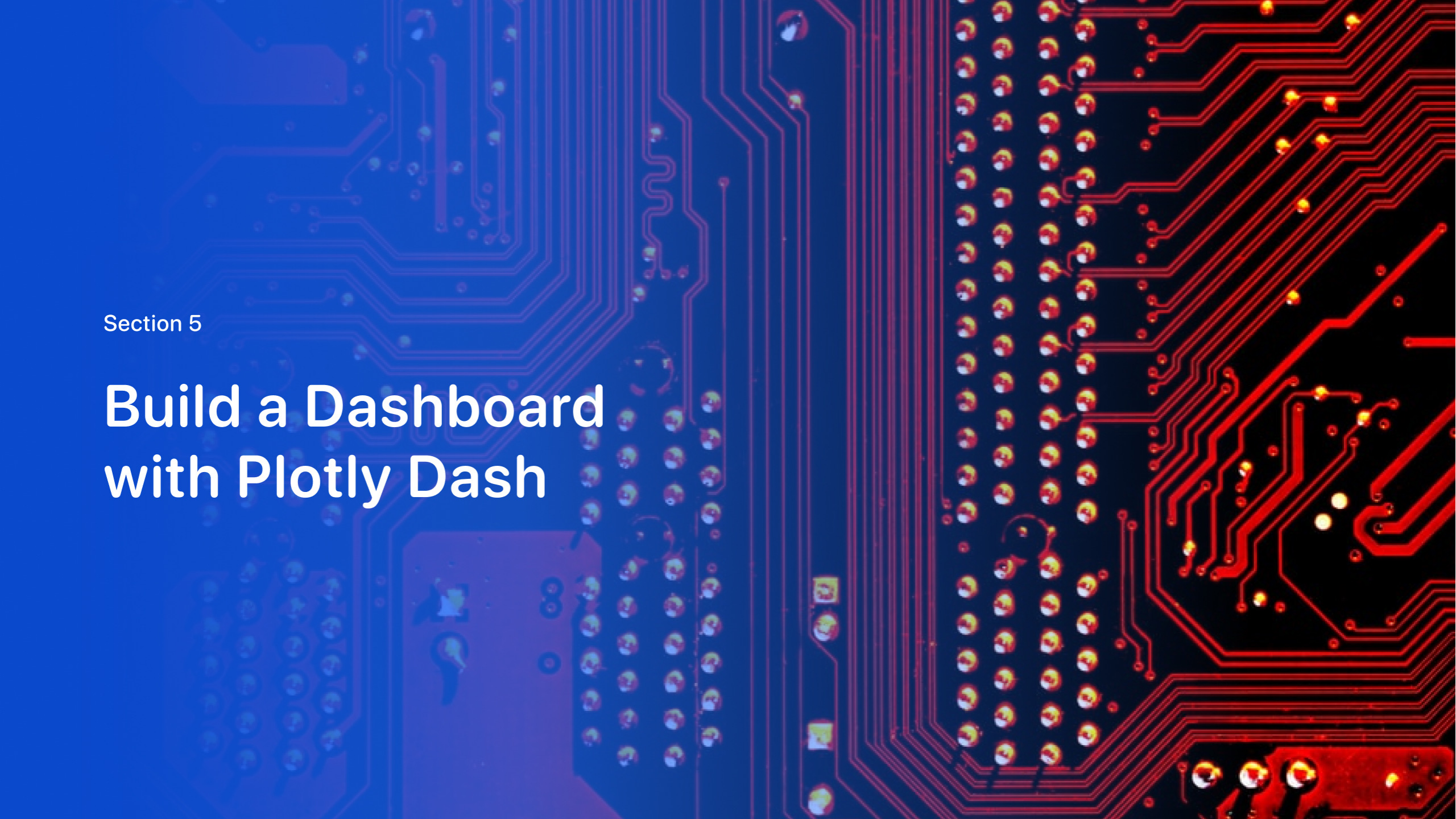
The map is color-coded with respect to the success/fail of the missions. It is possible to drill-down and look at the individual color-coded entries

Proximities of Launch Site



The map shows lines connecting launch site CCAFS-LC40 to its proximities

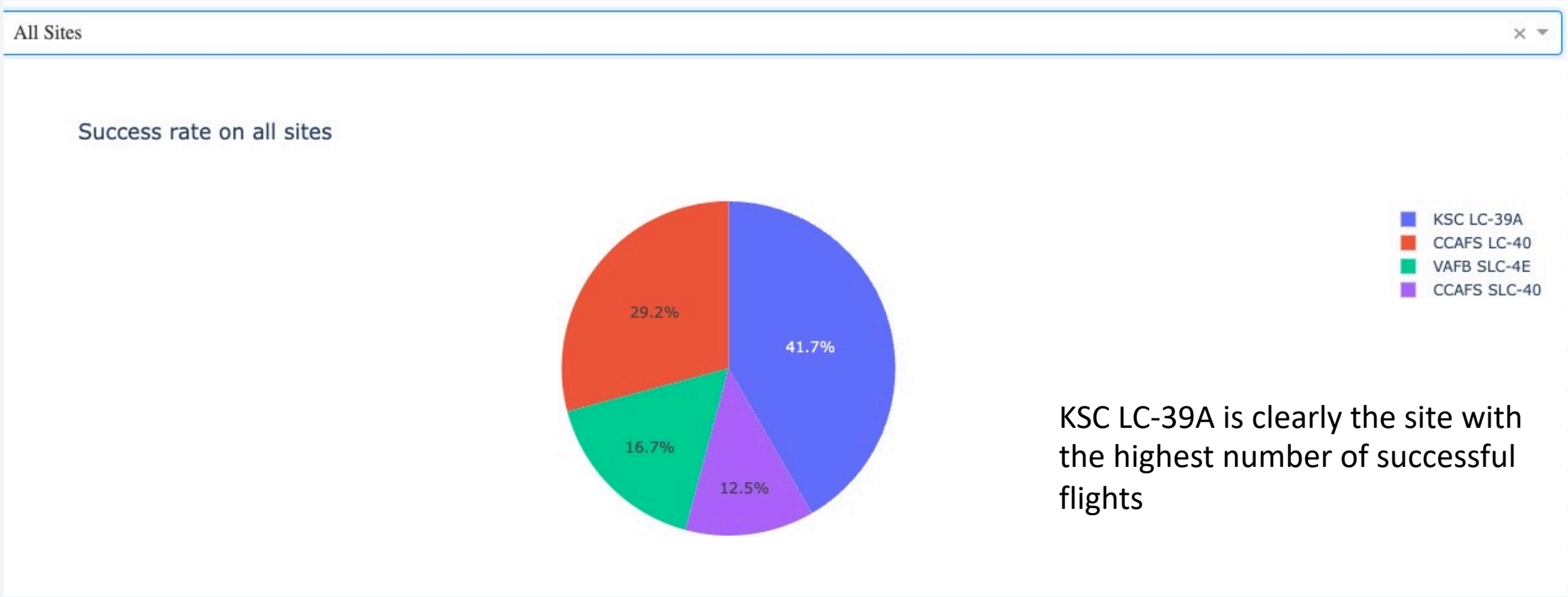
The distances to places like human settlements, railorads, streets and coastline are calculated and highlighted in the map.



Section 5

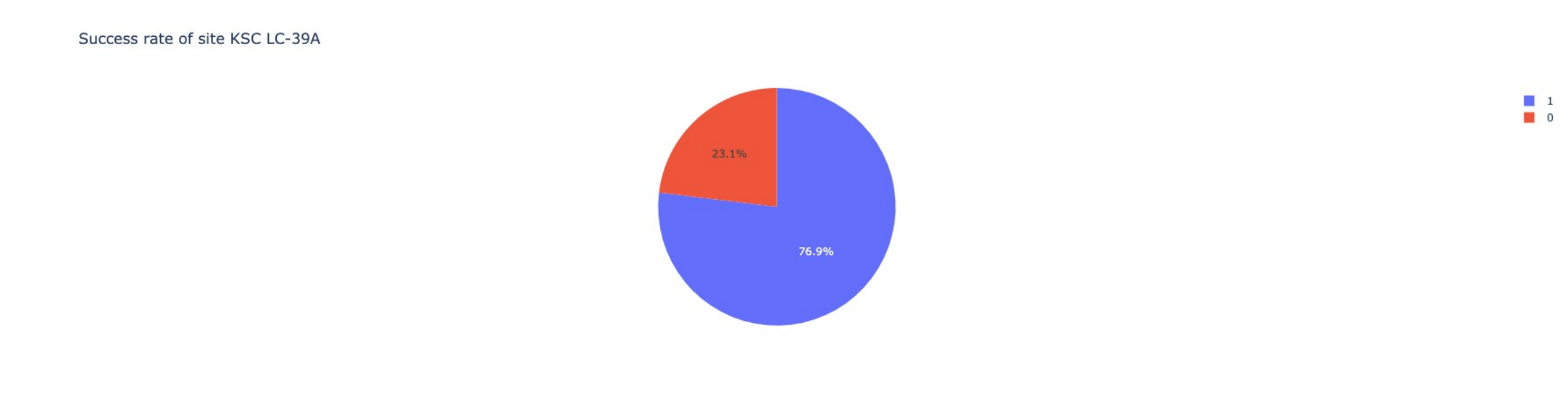
Build a Dashboard with Plotly Dash

Success Rate on all Sites



KSC LC-39A is clearly the site with the highest number of successful flights

Success Ratio of site KSC LC-39A



40 % of the total successes come from this site

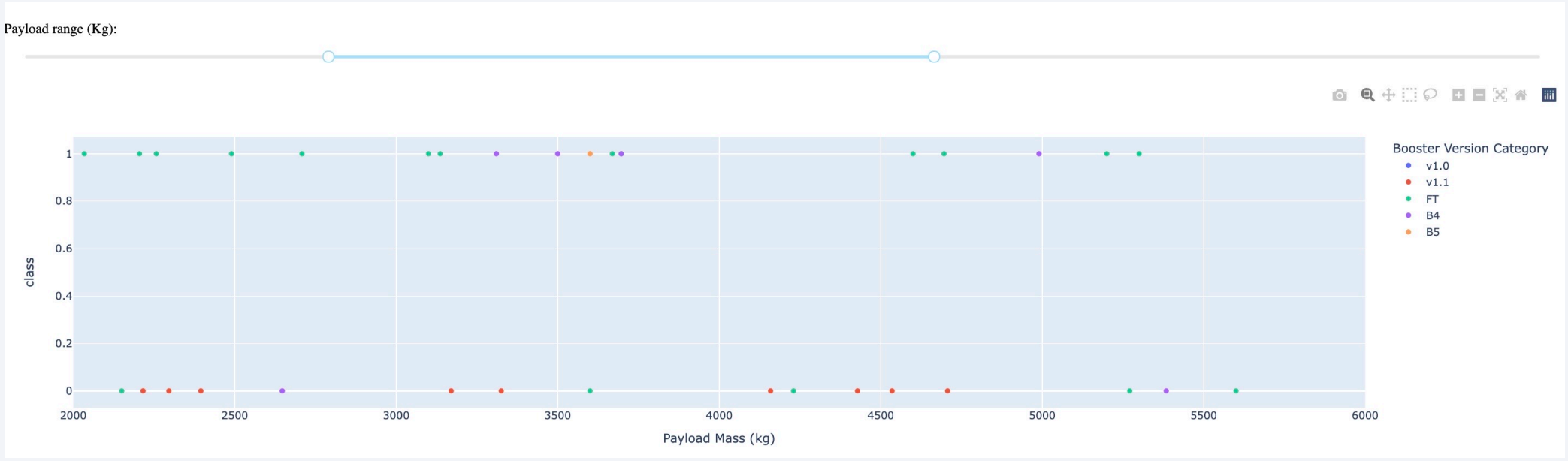
Within the site, 76.9% of the launches have been successful

Success vs Payload

We can see how different booster types have different success rate depending on the mass payload

Success vs Payload

In particular, most successful launches are obtained in the 2000 – 5500 kg range, with the FT booster version

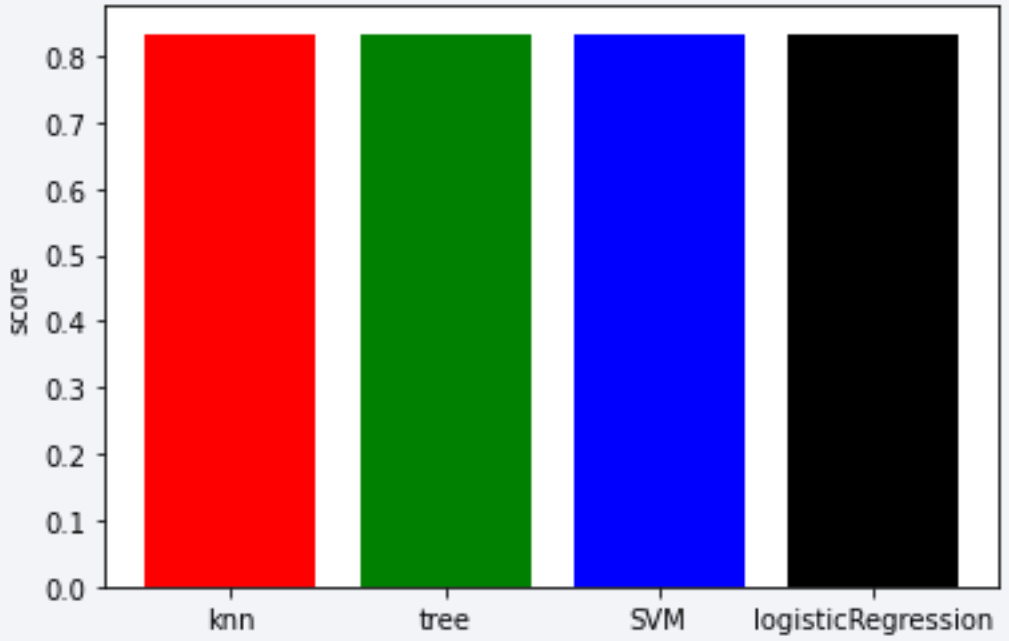


Section 6

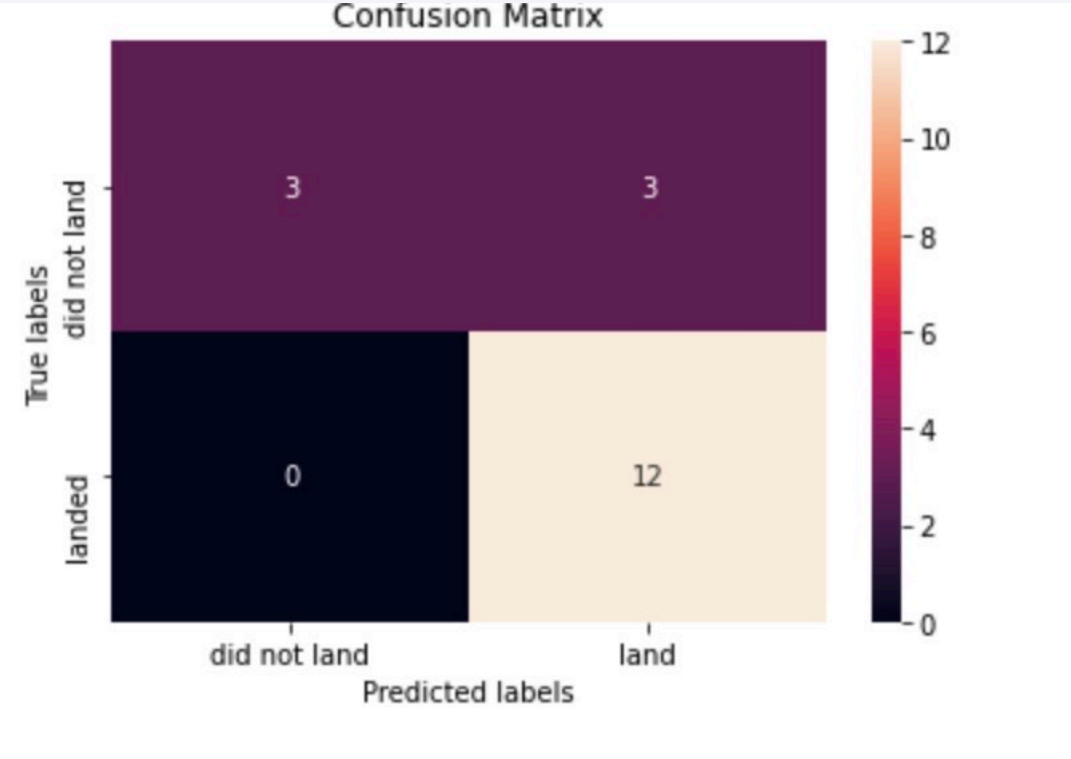
Predictive Analysis (Classification)

Classification Accuracy

All the models have the same score of about 0.833



Confusion Matrix



Here is the confusion matrix of one of the methods

The major problem here are the 12 false positive results

Conclusions

- EDA allows us to have a first impression of which factors affect a successful landing of the first stage. The latter being the major reason for the relative low cost of space X missions
- Launch site, Payload mass and target orbit are among the main factors
- KSC-LC39 A is the site with the highest success rate
- Low Payload launches are most successful when launched to the SSO, while High Payloads have a better chance of success when launched to PO and ISS
- A general increase of success rate over time is observed
- All the classification models tested performed in the same way, with some issues due to the high number of false positives

Appendix

- The following github page contains all the code used for the analysis presented so far:

https://github.com/EduardoSchiavo/AppliedDataScienceCapstone

Thank you!