Winning Space Race with Data Science

José Carro Yanes

2022/05/06

Outline

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

Executive Summary

- Summary of methodologies:

On this course we use Data Collection using web scraping and SpaceX API, Exploratory Data Analysis (EDA) including data wrangling, data visualization and interactive visual analytics; Machine Learning Prediction to create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.

- Summary of all results:

It was possible to collected valuable data from public sources; EDA allowed to identify which features are the best to predict success of launchings; Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective of the project is to evaluate the possibility of a company called Space Y competing with Space X
- guiding questions:
  - we can estimate the total cost of launches by predicting successful landings of the first stage rockets?
  - Where is the best place to make launches?
  - Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - it was used a SpaceX API and WebScraping
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data that was collected was normalized, partitioned into training and test data sets and evaluated using four different classification models, with the accuracy of each model being evaluated using different combinations of parameters.

Data Collection

Data sets were collected from:

i. Space X API: https://api.spacexdata.com/v4/rockets/

ii. Wikipedia using web scraping technics: (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights,

GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude,Latitude

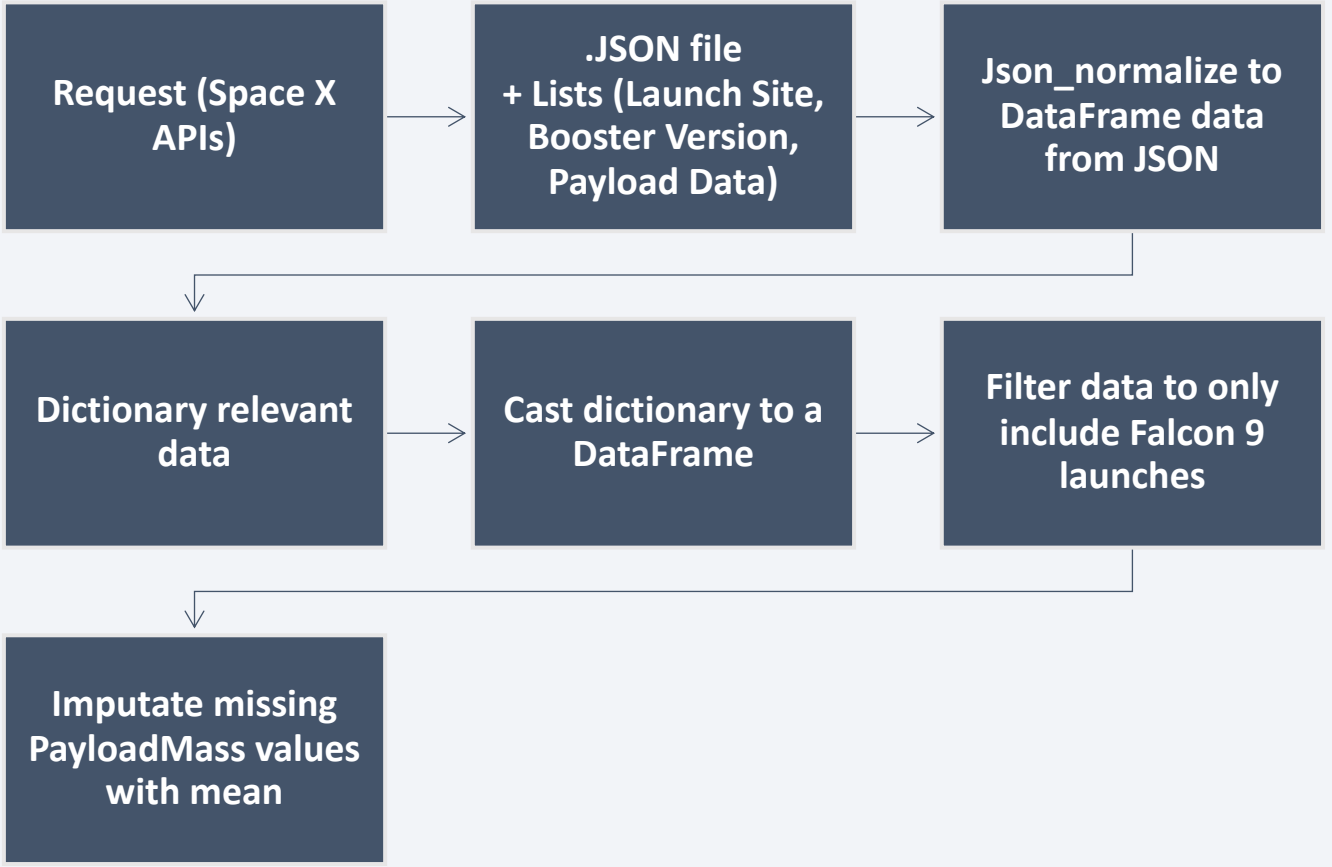
Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version

Booster, Booster landing, Date, Time

Data Collection – SpaceX API

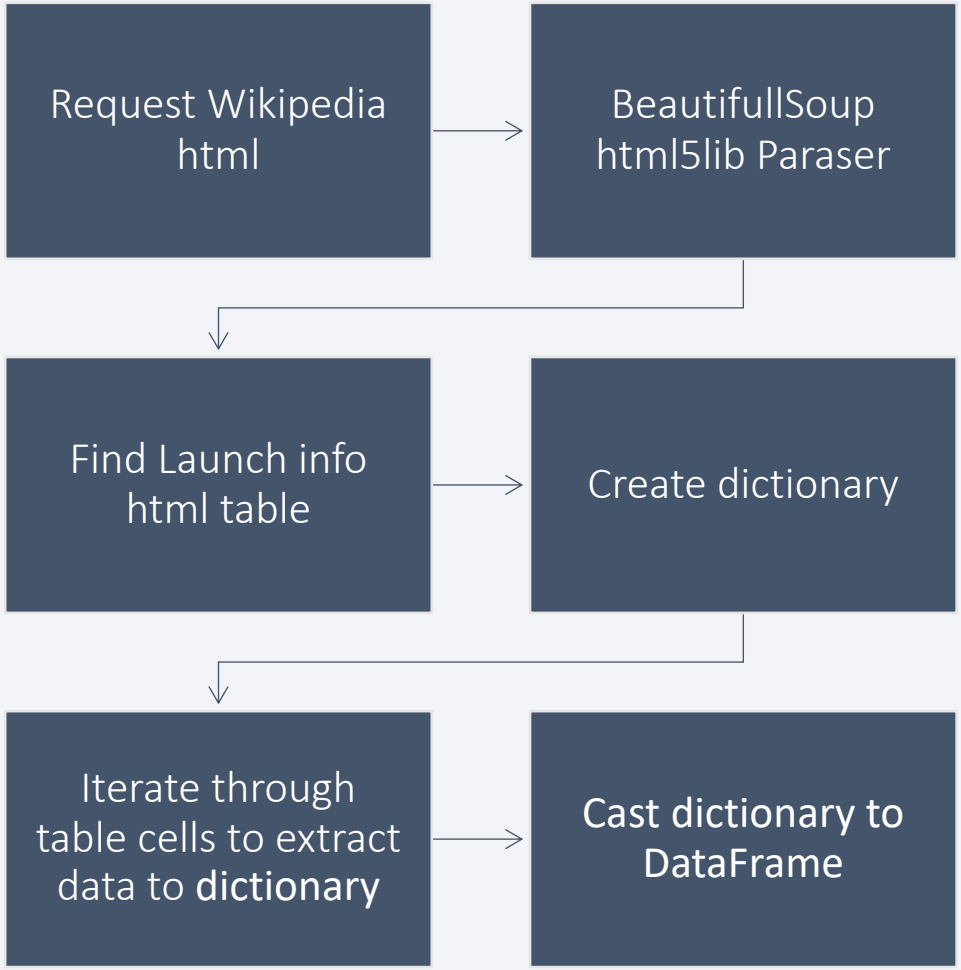
- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.



Github URL:
https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone.git

Data Collection - Scraping

Data Collection – Web Scraping



Github URL:

https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone.git

Data Wrangling

Create a training label with landing outcomes where successful = 1 & failure = 0.

Outcome column has two components: ‘Mission Outcome’ ‘LandingLocation’

New training label column ‘class’ with a value of 1 if ‘MissionOutcome’ is True and 0 otherwise.

<u>Value Mapping:</u>

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.

You can view the graphs in the following link

Github: https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone/blob/cebe9af170dcd02e7435cf0daf735ba6b88171b3/5-%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - 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 inground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - 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.

https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone/blob/dc7b4d0a9d54798cb78fd6d6c419404ebfd8508d/4-%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

GitHub: https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone/blob/b34206c3fdc5188dadbb17e1fcf803b4c6848026/6-%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

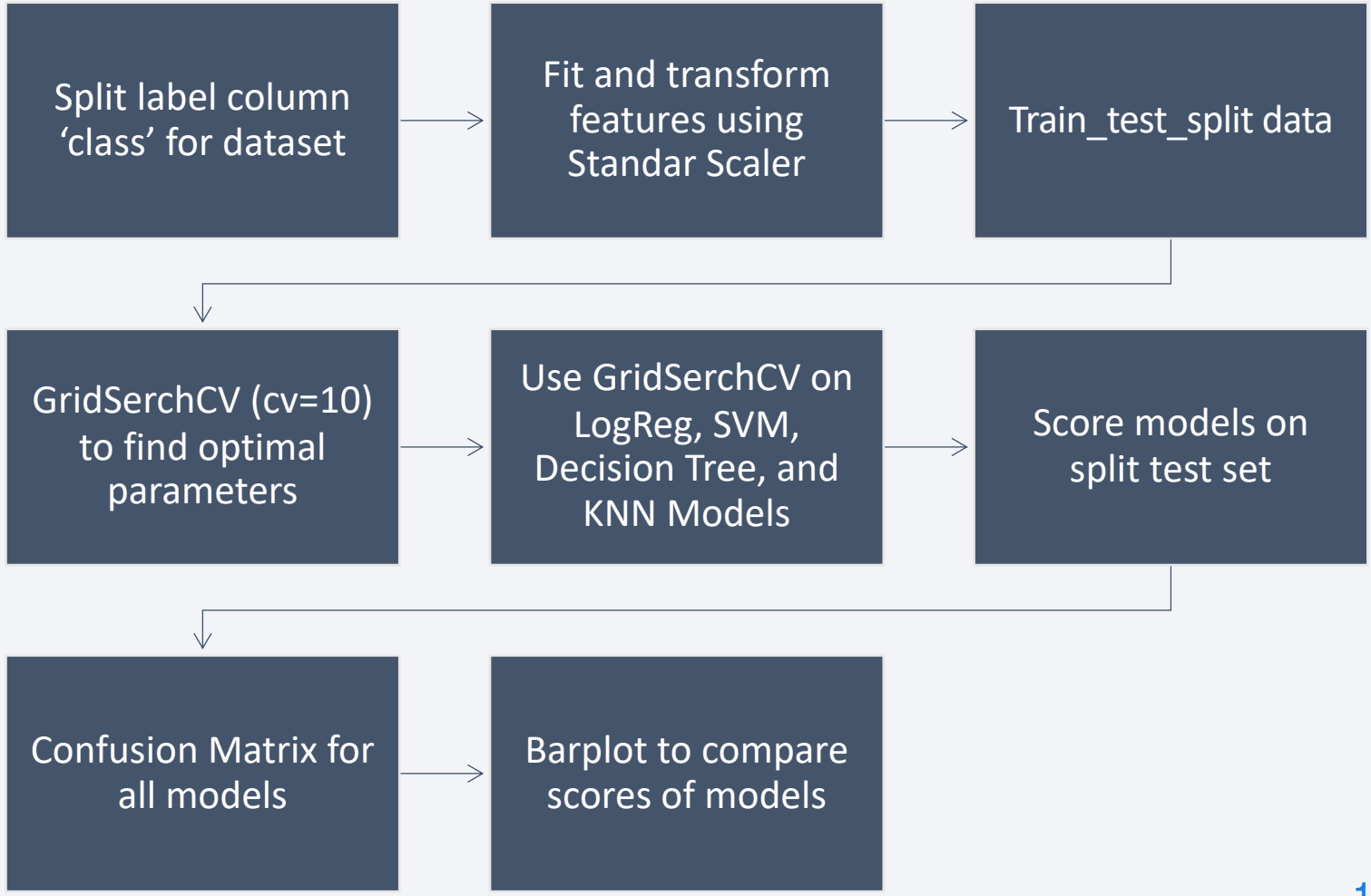
- Percentage of launches by site
- Payload range

This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone/blob/main/spacex%20dash%20app.py

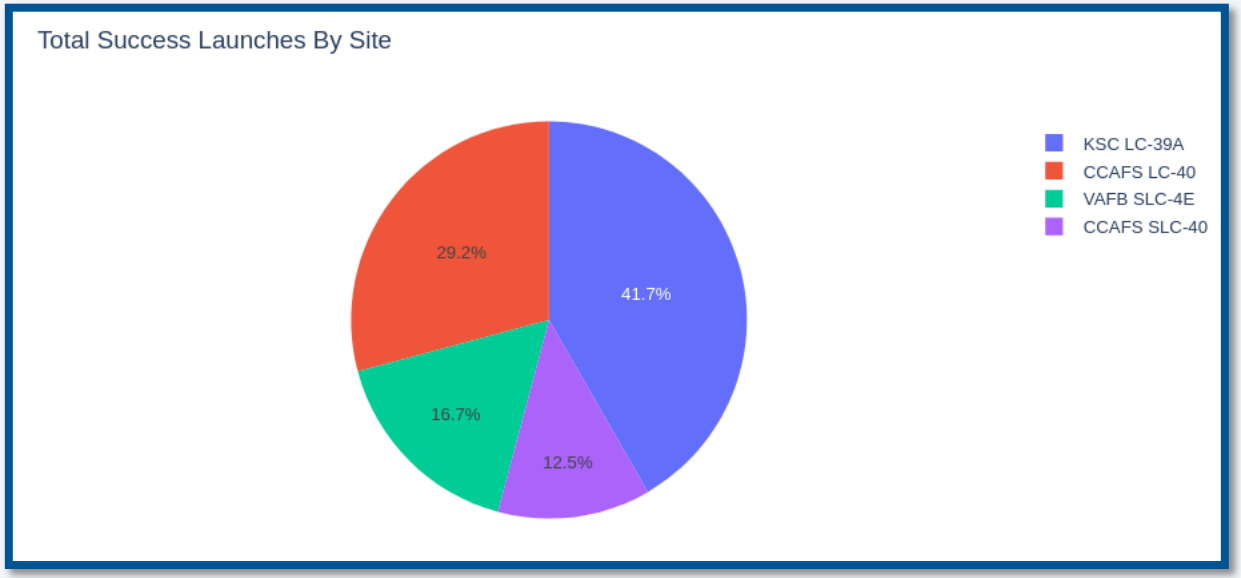
Predictive Analysis (Classification)

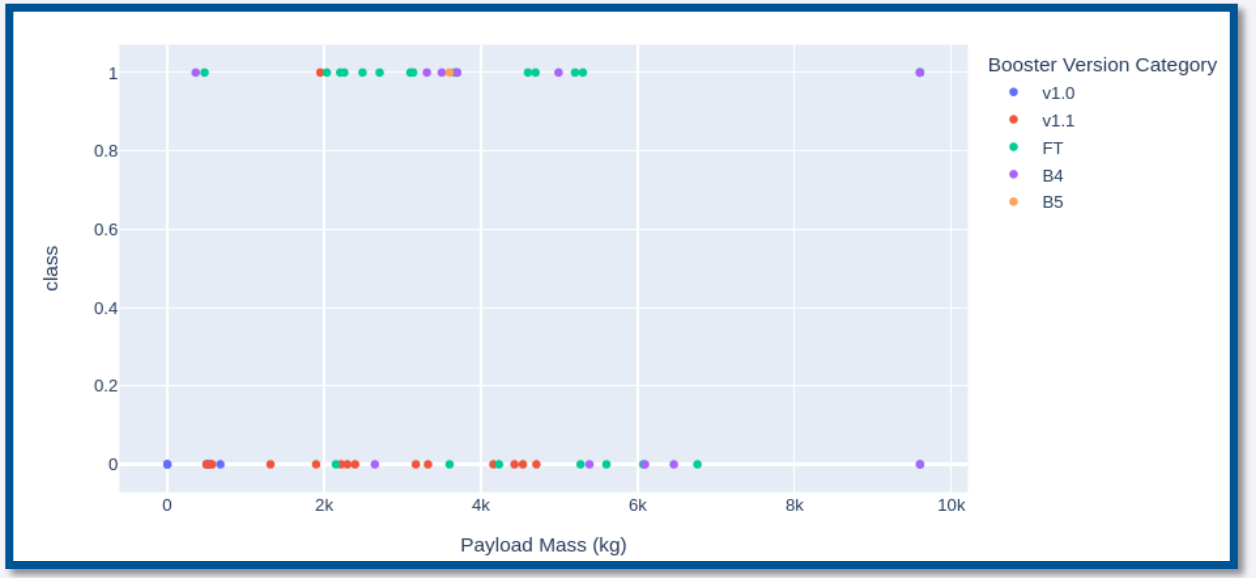
Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



Results

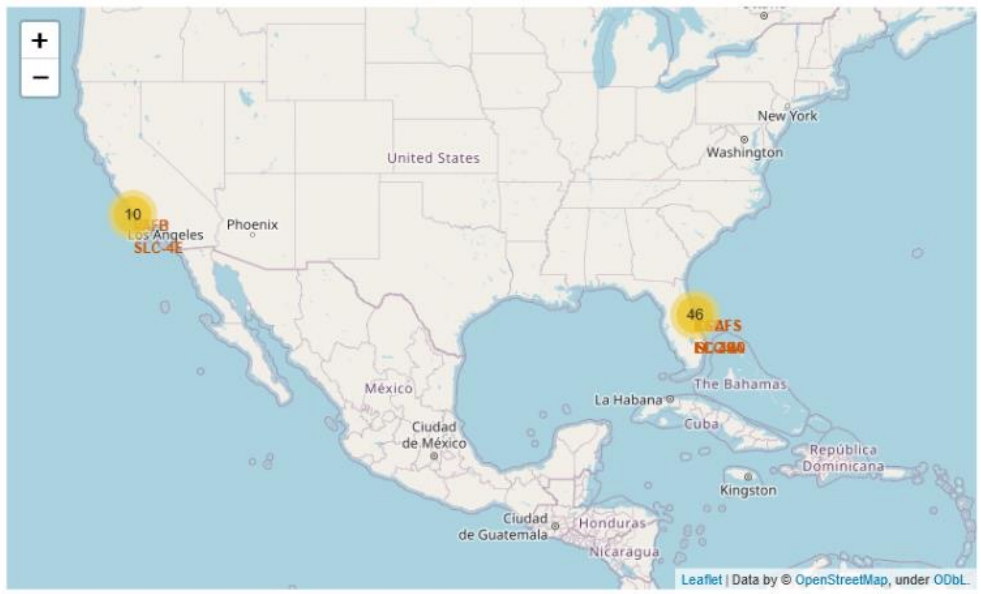
This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.





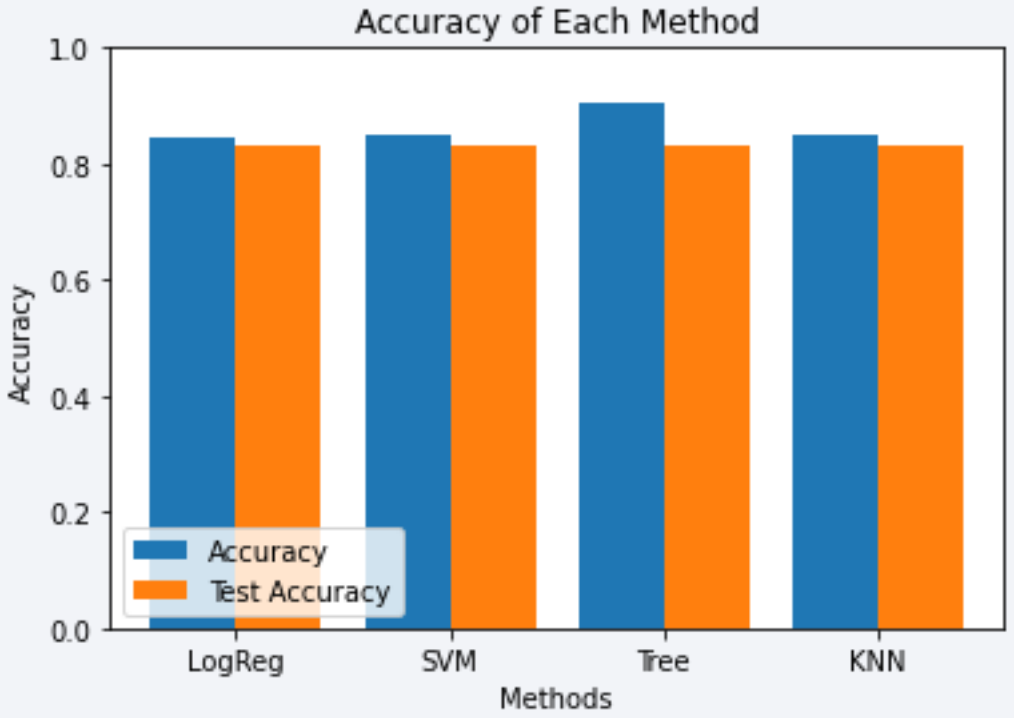
Results

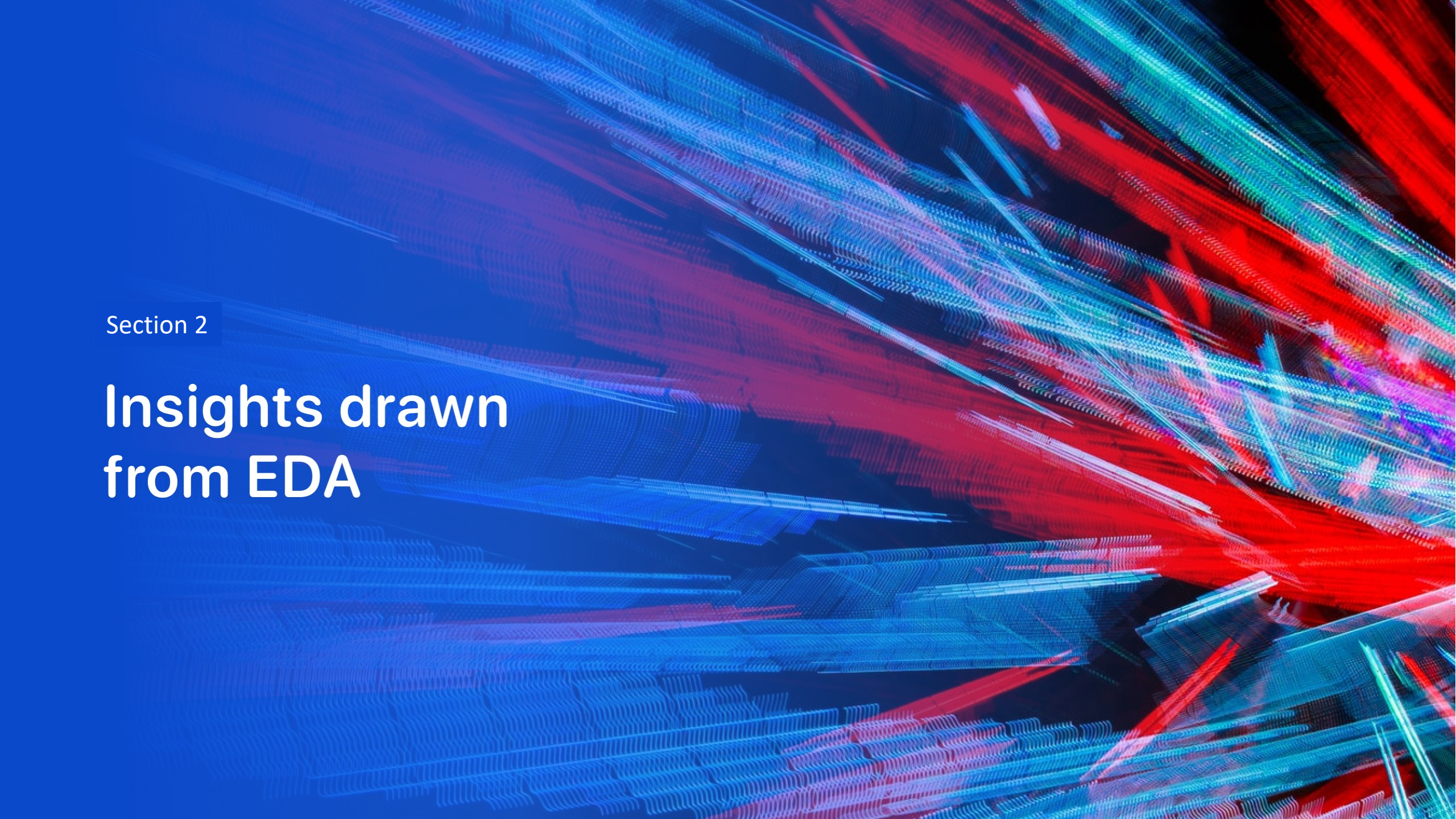
Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
Most launches happens at east cost launch sites.



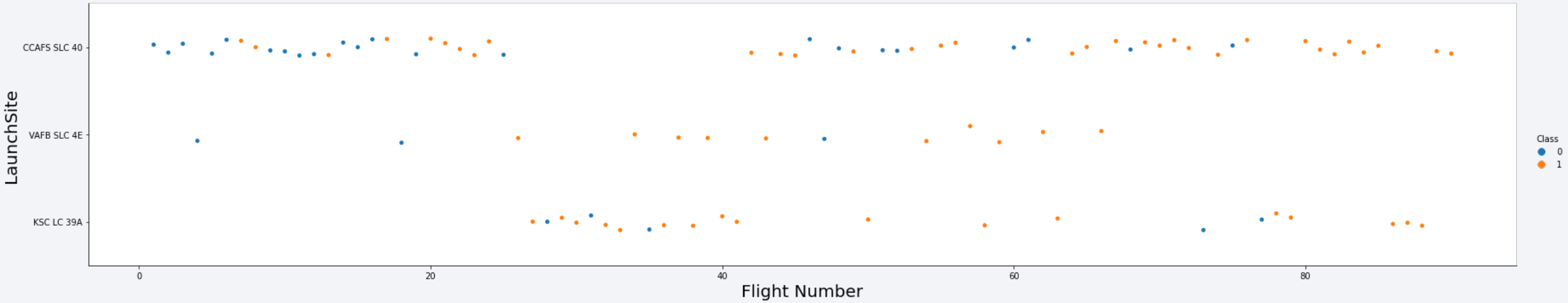
Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.




Section 2
Insights drawn from EDA

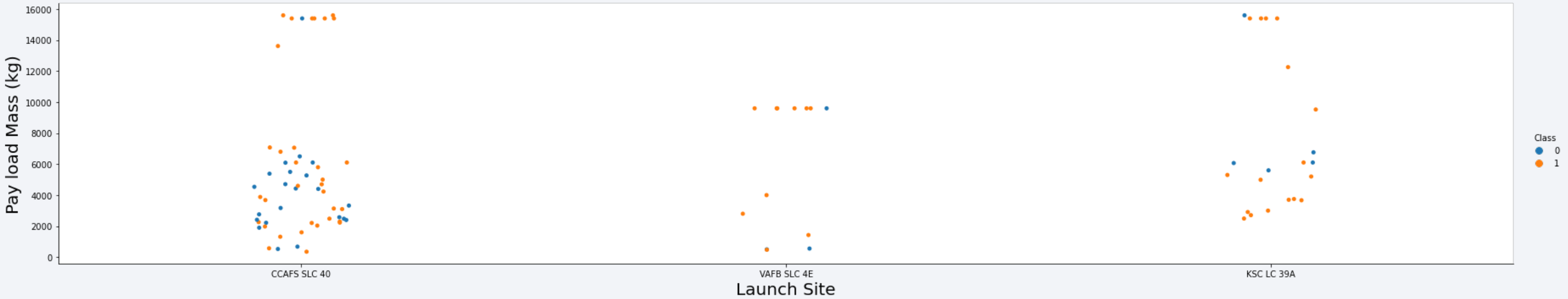
Flight Number vs. Launch Site



Blue indicates successful launch; Orange indicates unsuccessful launch.

Graphic suggests that CCAF5 SLC 40 seems to be the main launch site, as it has the highest volume. In second place VAFB SLC 4E and third place KSC LC 39A

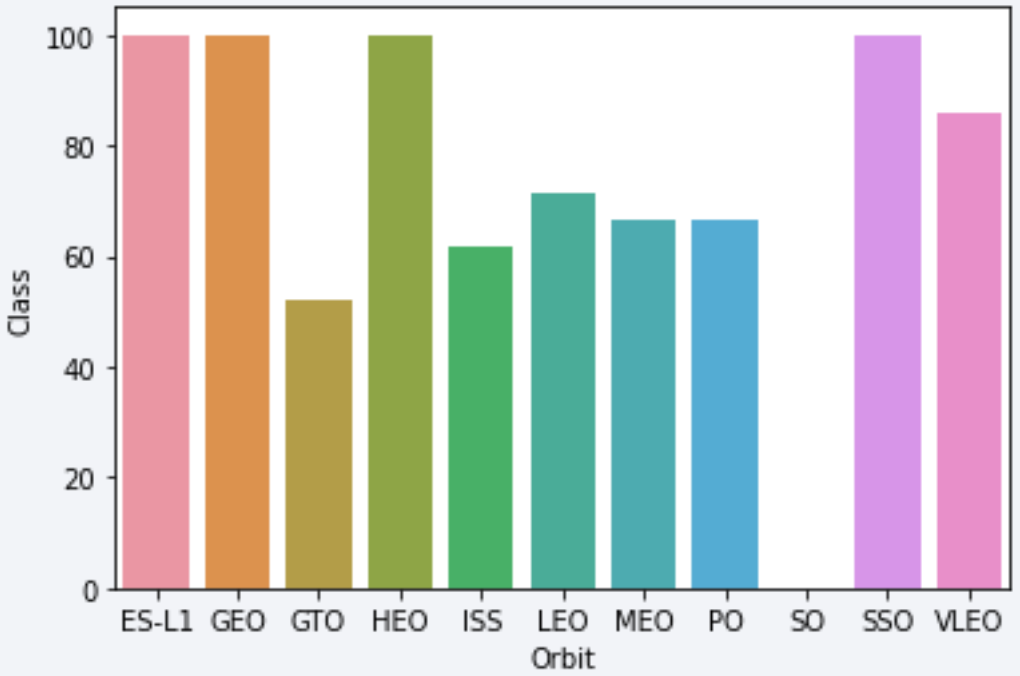
Payload vs. Launch Site



Blue indicates successful launch; Orange indicates unsuccessful launch.

Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass. Payloads over 9,000kg (about the weight of a school bus) have excellent success rate. Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



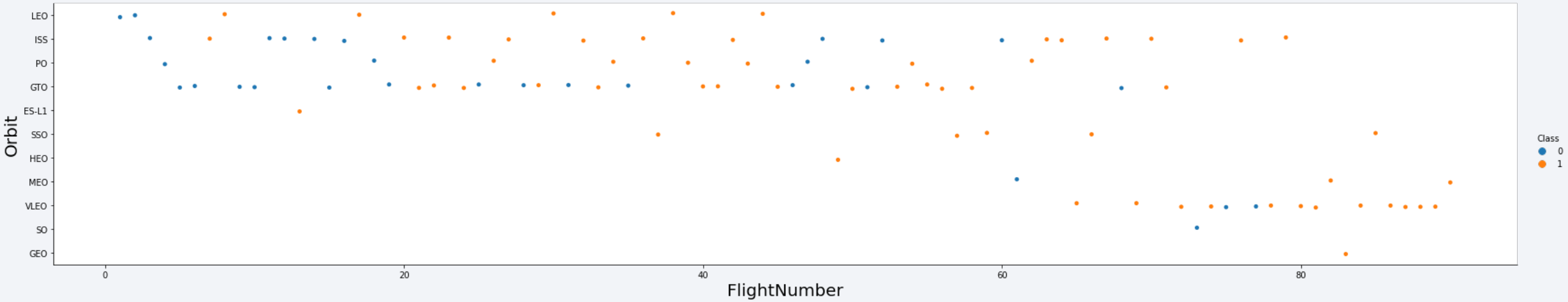
ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

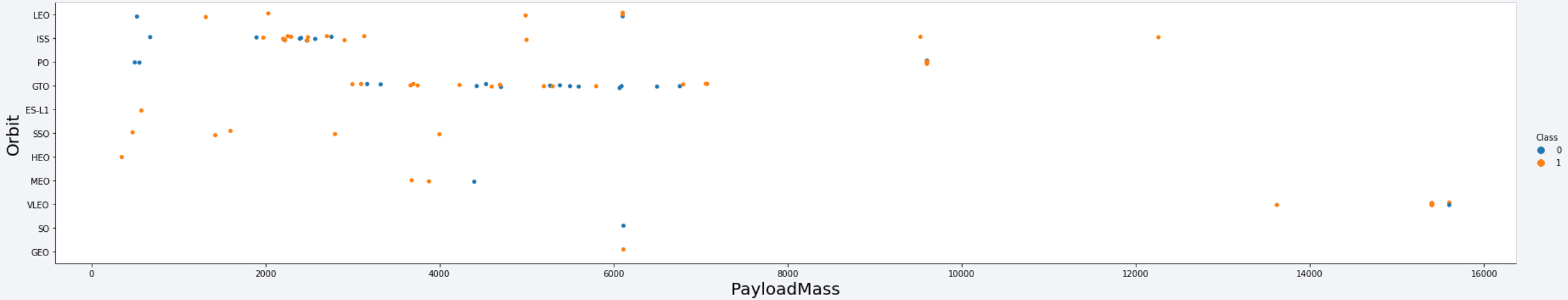
Flight Number vs. Orbit Type



Blue indicates successful launch; Orange indicates unsuccessful launch.

Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference. SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches. SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



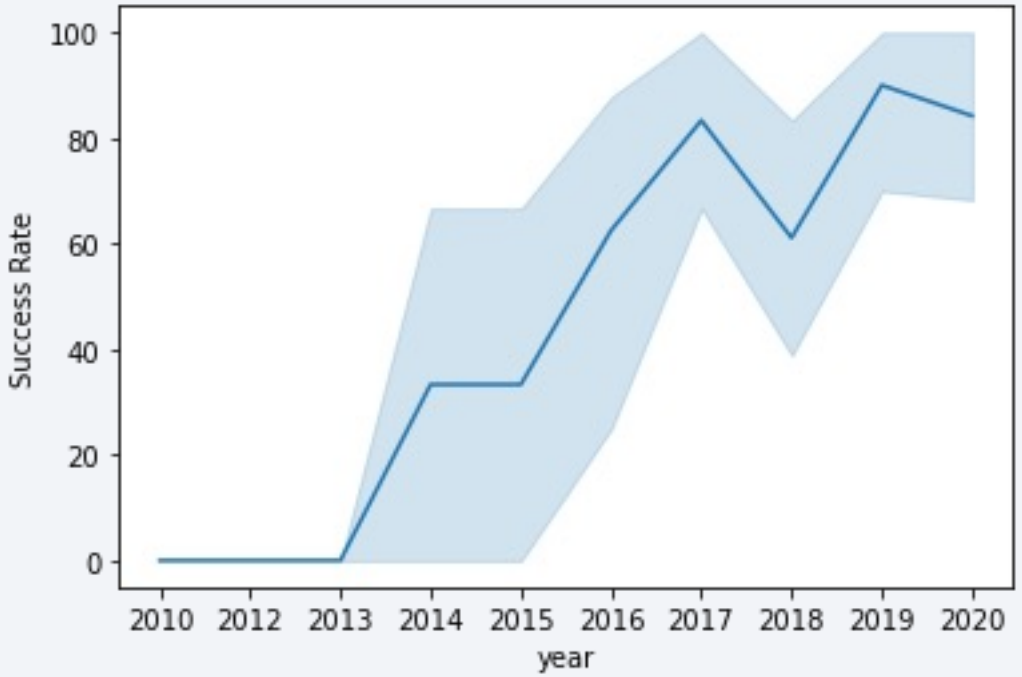
Blue indicates successful launch; Orange indicates unsuccessful launch.

Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



95% confidence interval (light blue shading)

Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

All Launch Site Names

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Query unique launch site names from database.

They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

SUM
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

In [8]:
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

avg_payload
2928

First Successful Ground Landing Date

By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

In [9]:
```
sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE '%uccess%';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

1
2015-12-22

This query returns the first successful ground pad landing date. First ground pad landing wasn't until the end of 2015. Successful landings in general appear starting 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

In [17]:
```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[17]:

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Successful and Failure Mission Outcomes

Number of successful and failure mission outcomes:

In [50]:
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[50]:

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

Grouping mission outcomes and counting records for each group led us to the summary above.
This query returns a count of each mission outcome.
SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing failures are intended.

One launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload

Boosters which have carried the maximum payload mass

```
maxm = %sql select max(payload_mass__kg_) from SPACEXDATASET
maxv = maxm[0][0]
%sql select booster_version from SPACEXDATASET where
payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
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

These are the boosters which have carried the maximum payload mass registered in the dataset (15600 kg.).

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [49]:
```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEXDATASET where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[49]:

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

This query returns the Month, Landing Outcome, Booster Version, and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

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

```
%sql select landing__outcome, count(*) as count from SPACEXDATASET
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

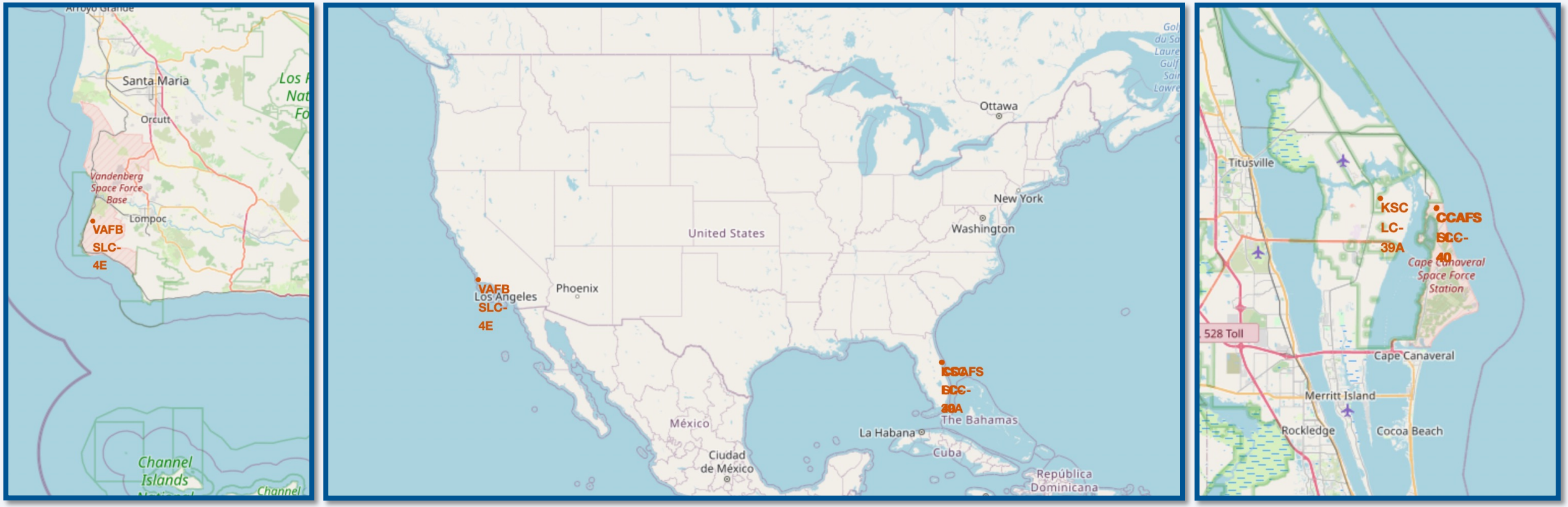
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

There are two types of successful landing outcomes: drone ship and ground pad landings.

There were 8 successful landings in total during this time period


Section 3
Launch Sites
Proximities Analysis

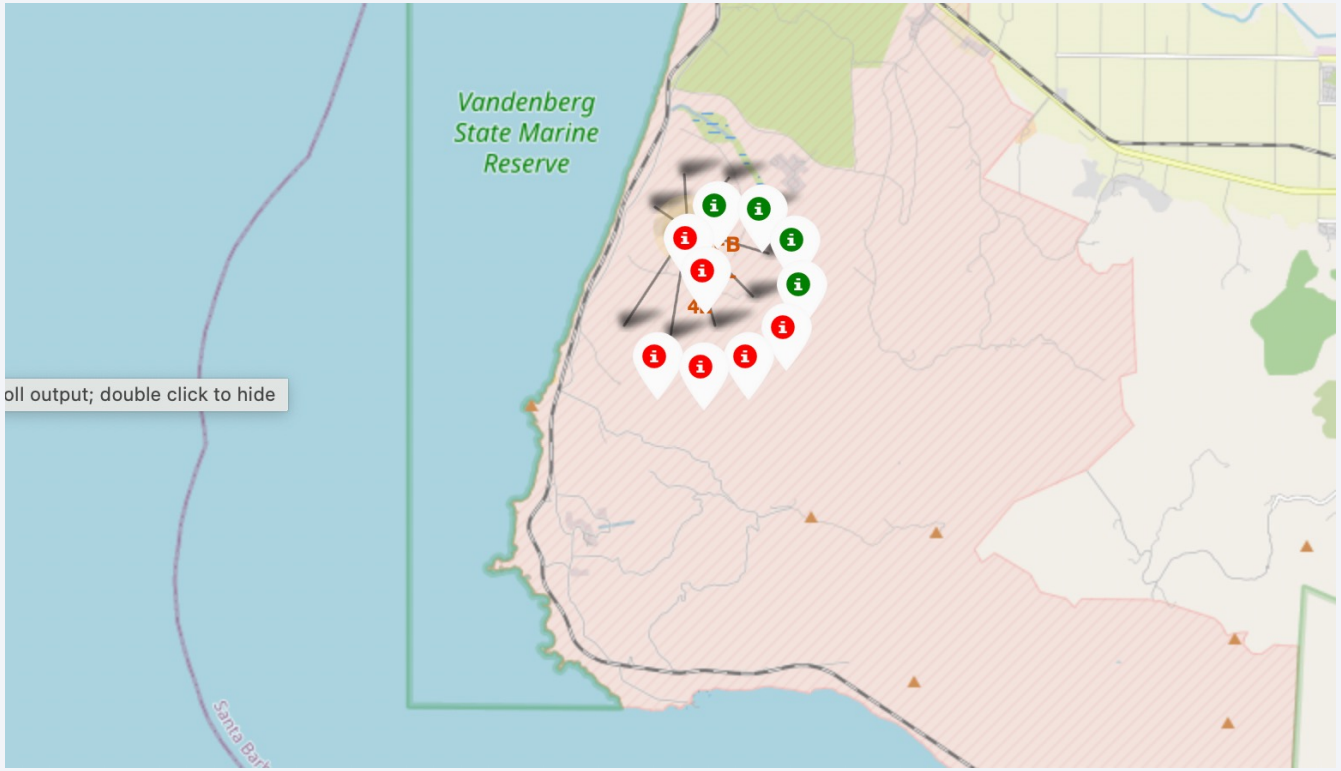
Launch site locations



Launch sites are near sea, probably by safety, but not too far from roads and railroads.

The left map shows the Los Angeles launch site, the center map shows all the launch sites relative to the US map, and the right map shows the two Florida launch sites as they are in close proximity.

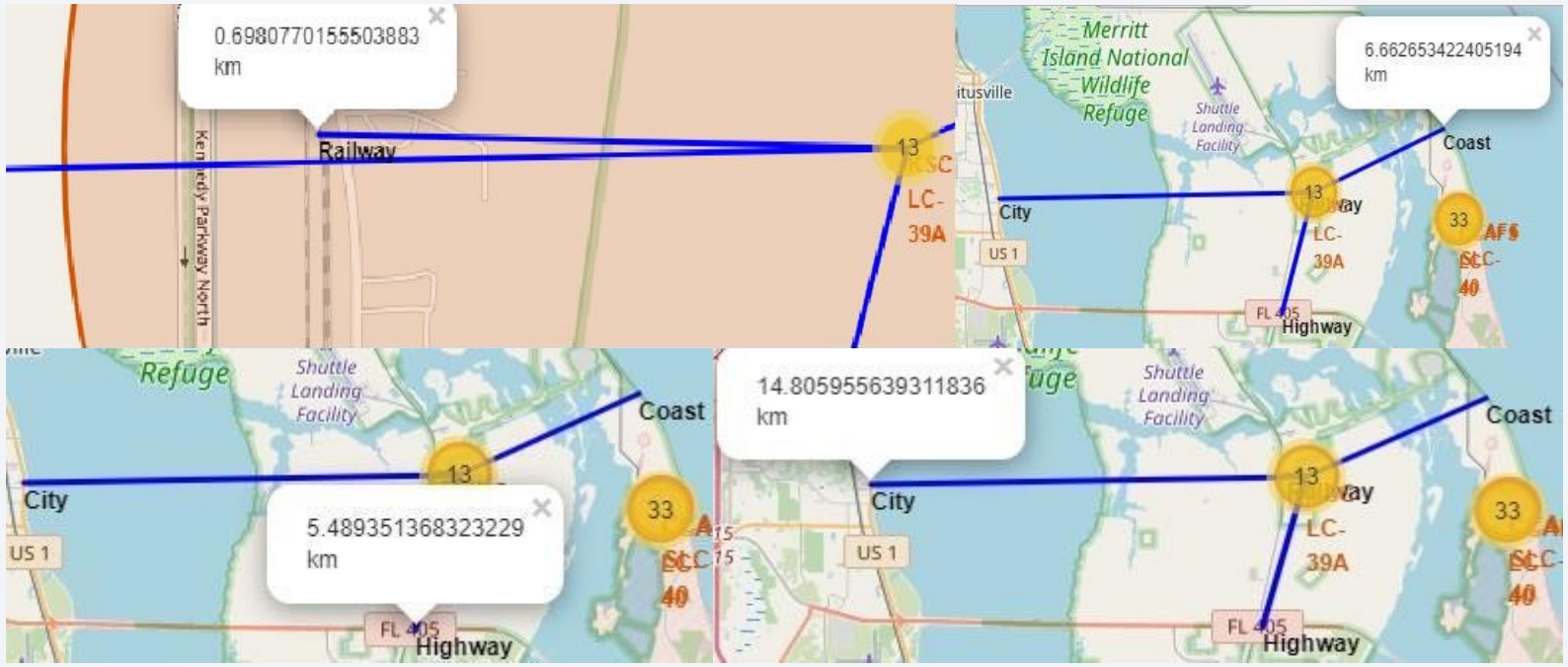
Site Color-coded launch markers



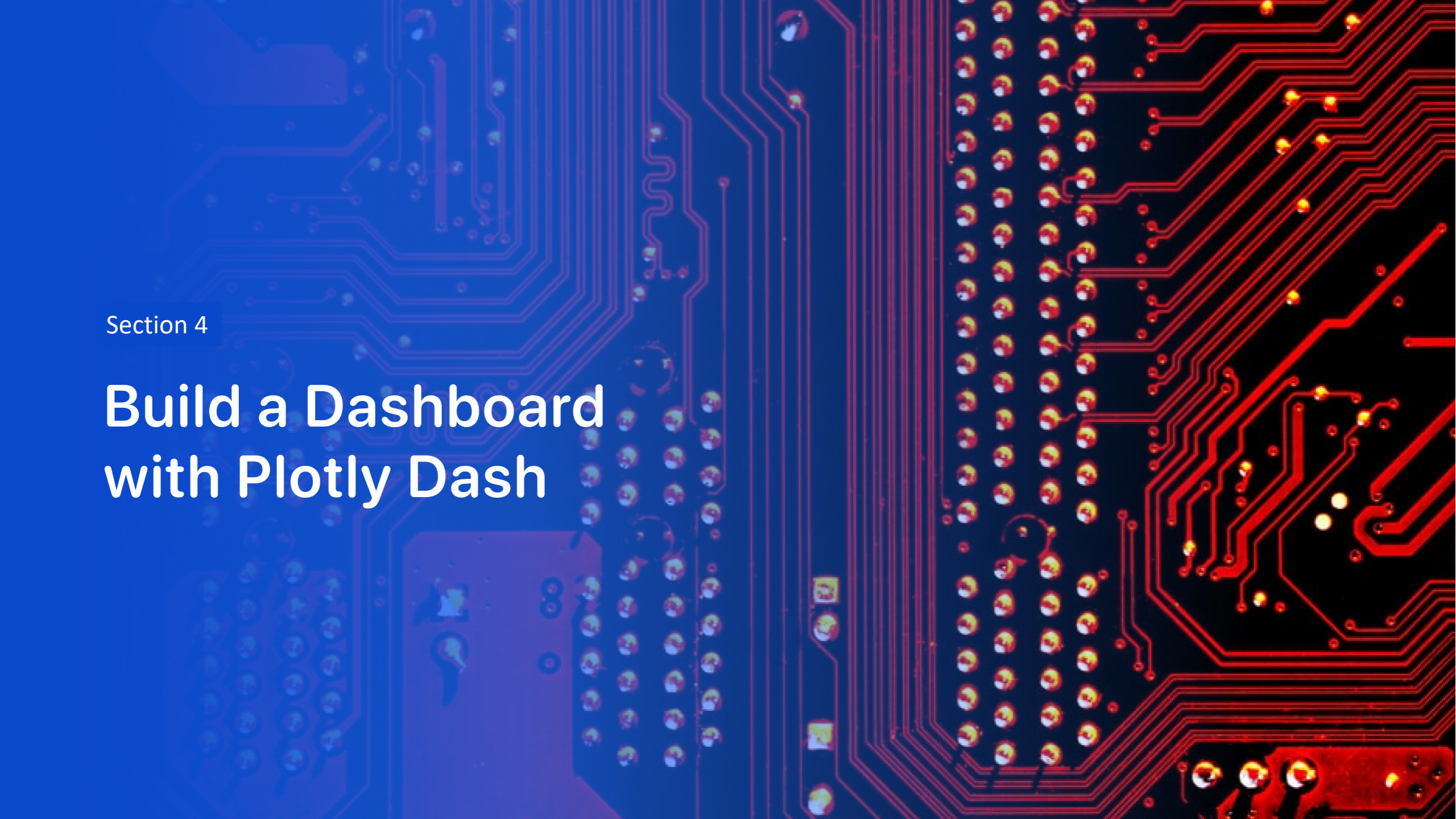
Green markers indicate successful and red ones indicate failure.

Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC- 4E shows 4 successful landings and 6 failed landings.

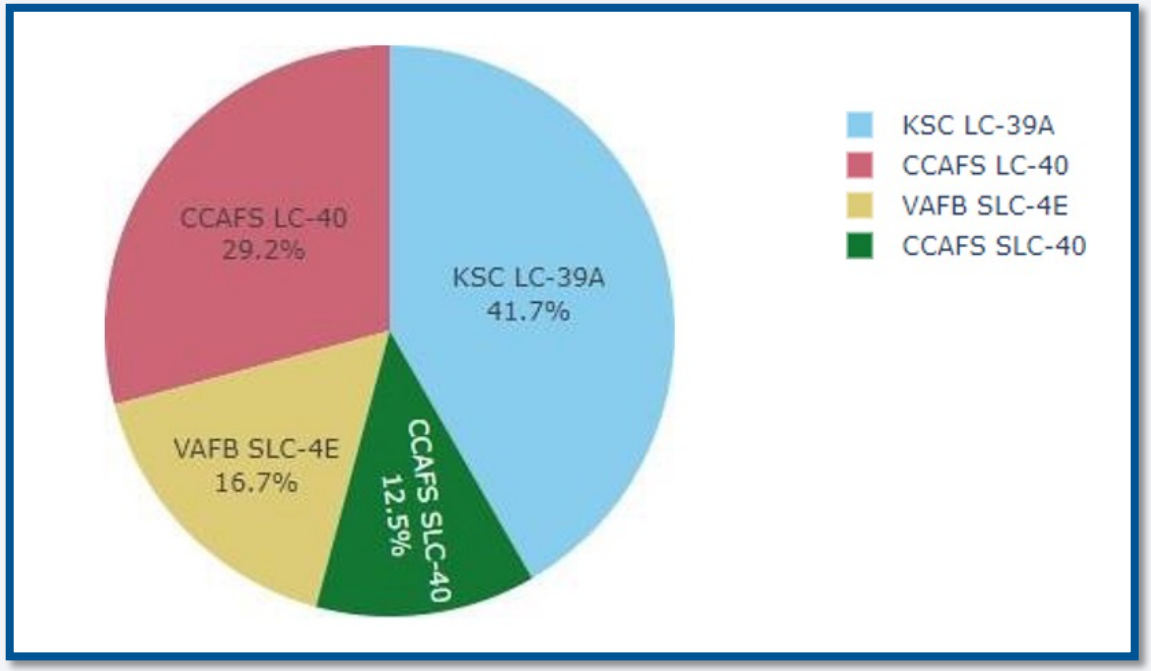
Logistics and Safety Locations



Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.


Section 4
Build a Dashboard with Plotly Dash

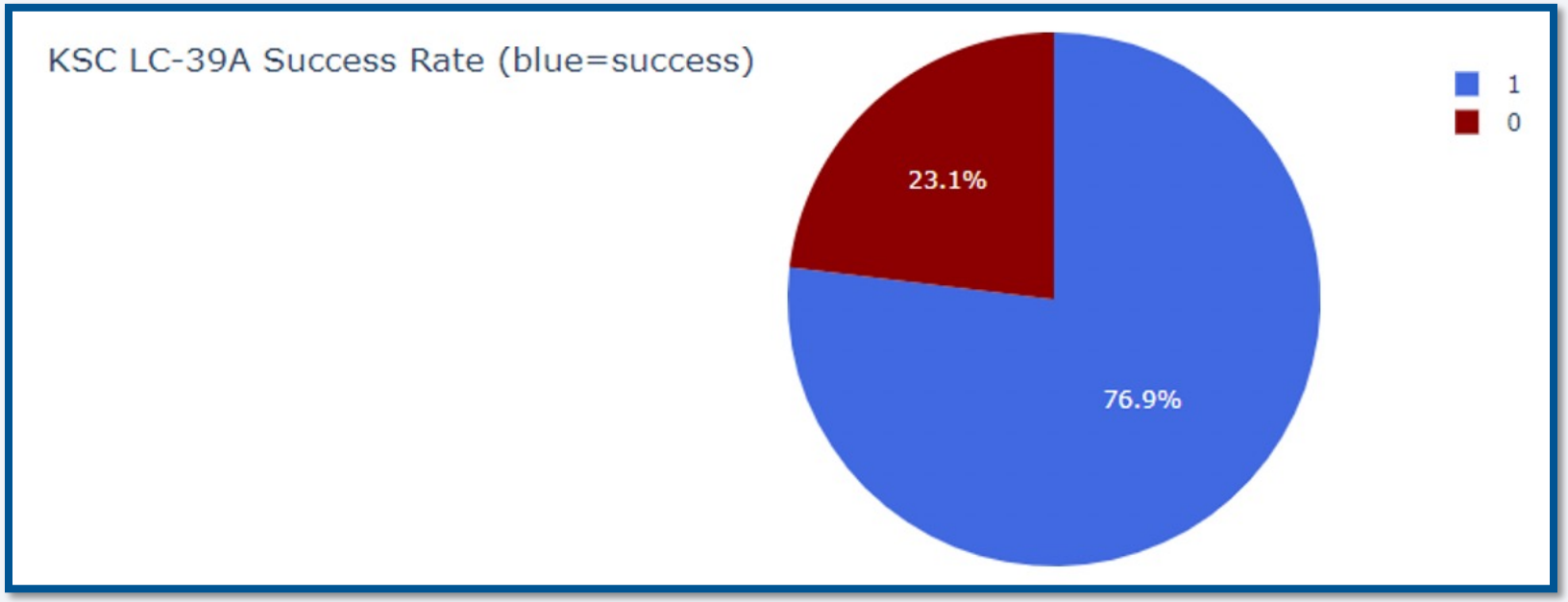
Successful Launches by site



The place from where launches are done seems to be a very important factor of success of missions.

This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

<Dashboard Screenshot 2>



76.9% of launches are successful in this site.
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

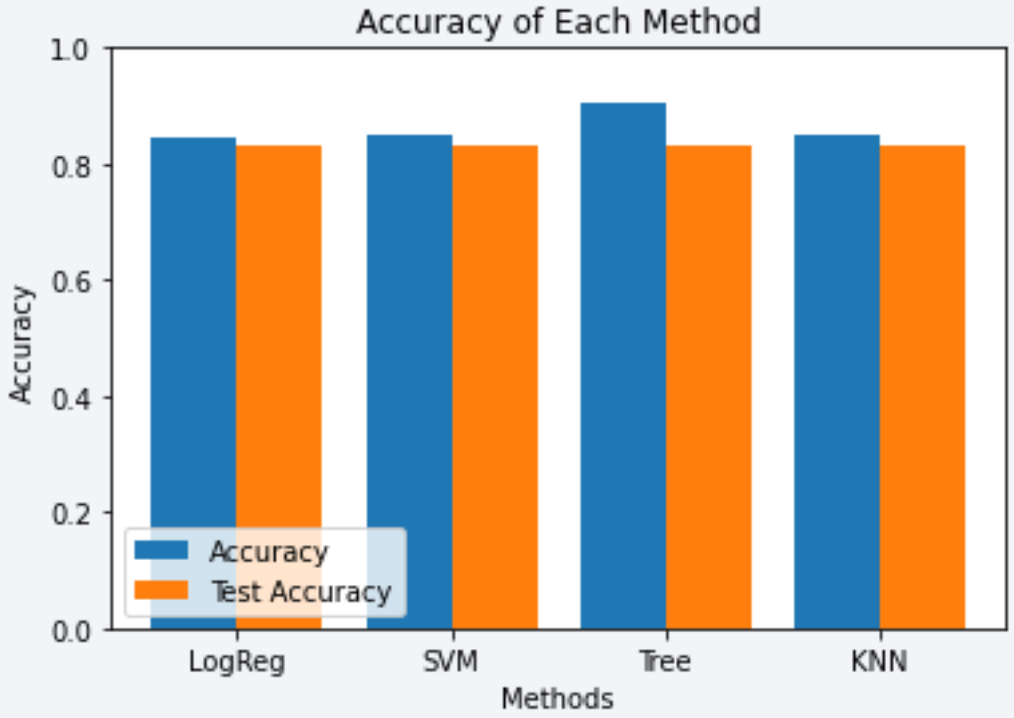
Payload Mass vs Success vs Booster Version Category

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

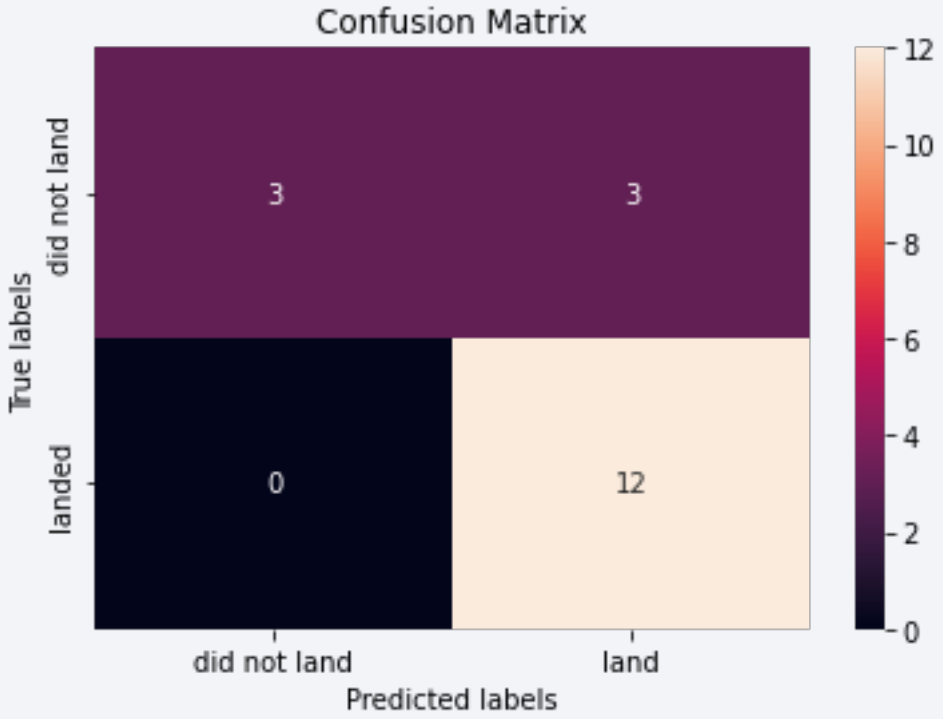


Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.90357	0.83333
KNN	0.84821	0.83333

Four classification models were tested, and their accuracies are plotted beside;

The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies is 90,36%

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.

The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.

The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

•The objective of this project was to develop a machine learning model for Space Y that seeks to compete with SpaceX. The goal of model is to predict when Stage 1 will successfully land to save 100 million USD
•Different data sources were analysed, refining conclusions along the process. Used data from a public SpaceX API and web scraping SpaceX Wikipedia page. Created data labels and stored data into a DB2 SQL database
•a machine learning model was developed with an accuracy of 83%
•SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing prior to launch, to determine whether or not the launch should proceed.
•To estimate a better machine learning model, more data must be collected to obtain greater accuracy.

According to the model:
•The best launch site is KSC LC-39A;
•Launches above 7,000kg are less risky;
•Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;

Appendix

- As an improvement for model tests, it's important to set a value to np.random.seed variable;
- Folium didn't show maps on Github, so I took screenshots.

GitHub Repository:

https://github.com/josecarro96/Final-Project---osecarro96-Final-Project---Applied-Data-Science-Capstone.git

Thank you!