IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Malick Diene
06th December 2022

Outline

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

Executive Summary

- In order to respond to our problematic, we needed first to **COLLECT DATA** via an *API* and using *web scraping* methodology, then we proceed to the **DATA WRANGLING** stage to transform and map our collected data. After the data collected and prepared to be used, we performed **EDA** -exploratory data analysis- (for initial analysis and findings) using *Data Visualization and SQL.* For advance analysis we used **Interactive Visual Analytics** and **Dashboard** to explore and manipulate data in an interactive and real-time way. We concluded our work with **Predictive Analysis** by building a machine learning pipeline to predict our target value.

- Since 2013 the success rate of Falcon 9's landing successfully has improved a lot. Some launch sites have better success rate than other : CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E have a success rate of around 77%. There are also good correlation between success rate and others attributes, like payload mass and specially with F9 Booster version (FT booster have the best success rate (78%) when payload mass is under 5500kg)

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful space compagnie is SpaceX. One reason for SpaceX's success is that their rocket launches are relatively inexpensive compare to others compagnies. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars other providers cost upwards of 165 million dollars each, much of the savings is because unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage.

- Can we predict if SpaceX will reuse the first stage ? Therefore, if we can determine if the first stage will land, we can determine the cost of a launch and we will be able to minimize it.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting data from SpaceX REST API and web scraping from a Wikipedia page
- Perform data wrangling
  - Understanding key attributes in the data (launch sites, orbit, and outcome), and creating a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing our data, split it into Train and Test data, finding hyperparameters with Grid Search for the following models : Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors

Data Collection

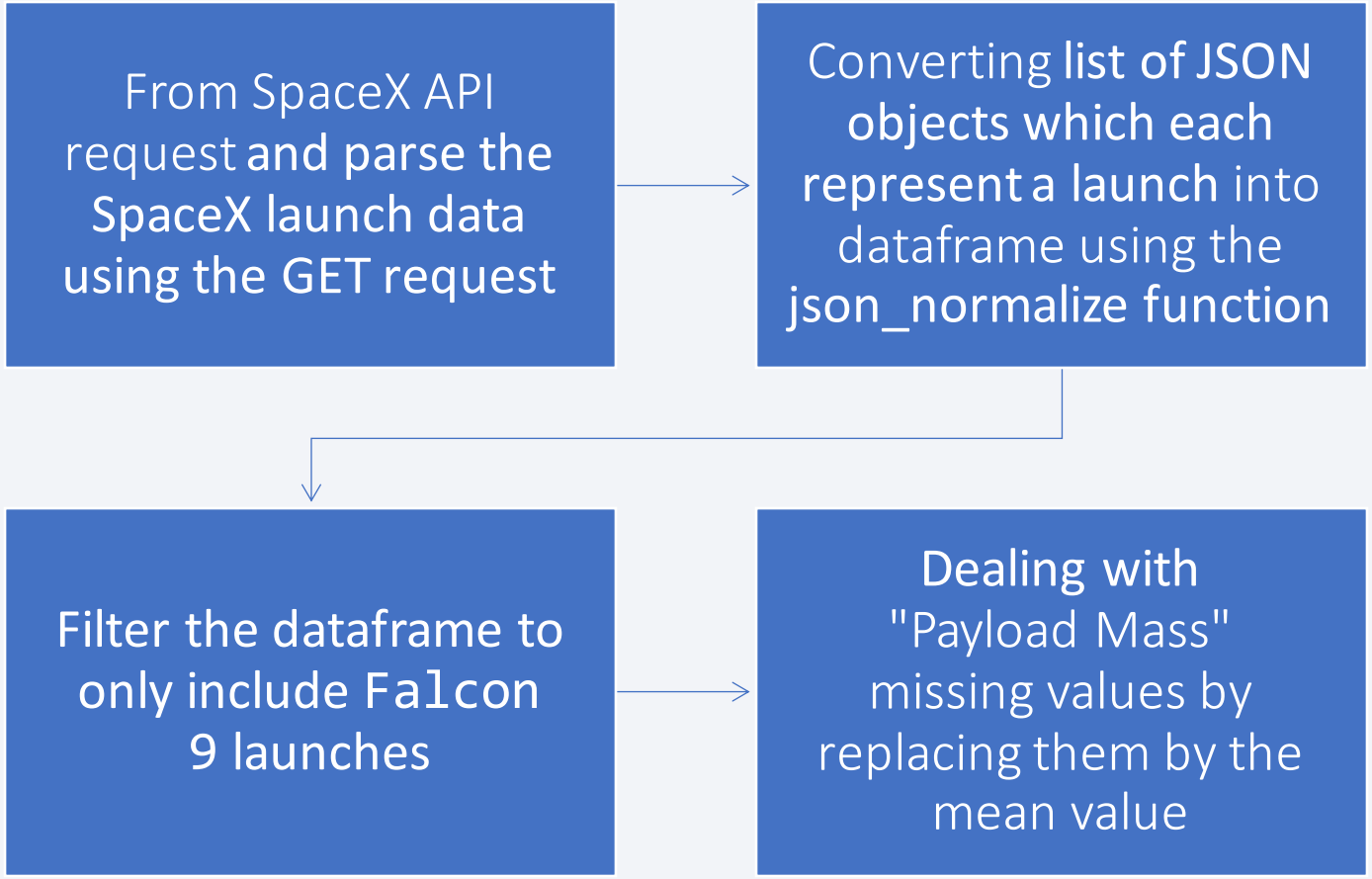
SpaceX REST API : using requests**. Data were cleaned after collected.**

- The API give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome

Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches" : using web scraping.

- Python BeautifulSoup package were used to web scrape some HTML tables that contain valuable Falcon 9 launch records such as landing outcome, date, booster version, and customer.

Data Collection – SpaceX API

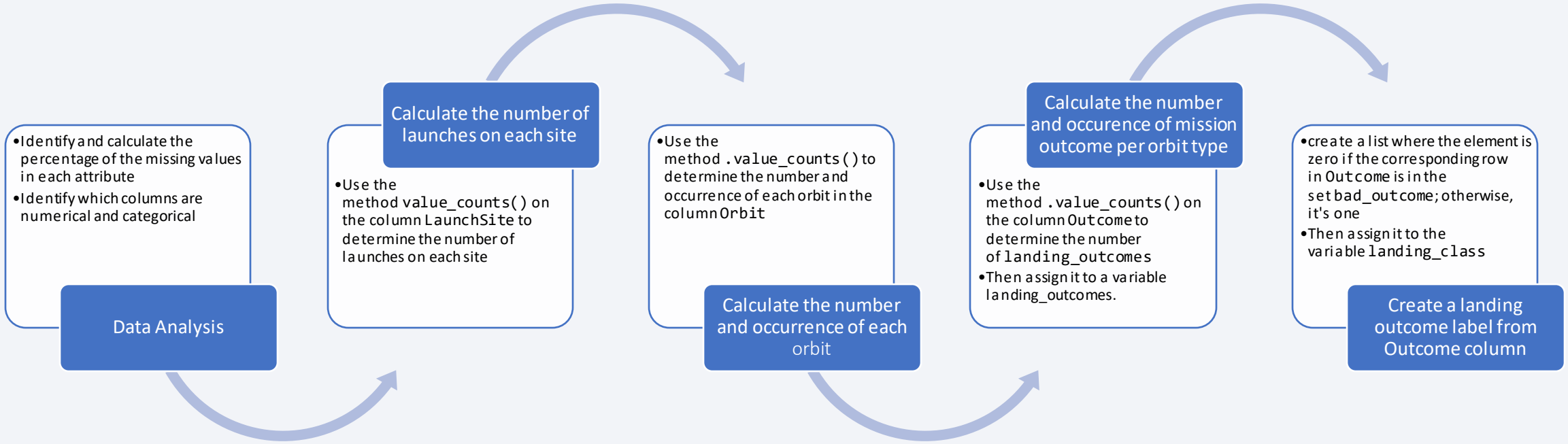


Data Collection via SpaceX API Notebook link

Data Collection - Scraping

Data Collection via Web Scraping Notebook link

Data Wrangling



The "Landing_class" variable created at the end represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully.

[Data Wrangling Notebook link]

EDA with Data Visualization

- Several charts were plotted to find insights in the data :
  - scatter plot of Flight Number vs. Launch Site : to find impact of site launch on success landing
  - scatter plot of Payload vs. Launch Site : to find relationship between launch sites and their payload mass
  - bar chart for the success rate of each Orbit type : to visually correlation between success rate and orbit type
  - scatter point of Flight number vs. Orbit type : to find relationship between flight number and orbit type
  - scatter point of payload vs. Orbit type : to reveal the relationship between Payload and Orbit type
  - line chart of yearly average success rate : to get the average launch success trend

EDA with Data Visualization Notebook link

EDA with SQL

- The following SQL queries were performed to retrieve information :
  - SELECT **DISTINCT** : to display the names of the unique launch sites in the space mission
  - SELECT * FROM x WHERE y **LIKE** z **LIMIT** : to display 5 records where launch sites begin with the string 'CCA'
  - SELECT **SUM** : to display the total payload mass carried by boosters launched
  - SELECT **AVG** : to display the average payload mass carried by a particular booster
  - SELECT **MIN** : to list the date when the first successful landing outcome was achieved
  - SELECT … **BETWEEN** a **AND** b : to display booster names which have payload mass greater than 4000 but less than 6000
  - SELECT... COUNT … GROUP BY : to list the total number of successful and failure mission outcomes
  - SELECT DISTINCT … WHERE x=(**SELECT MAX...**) : to list the boost which have carried the maximum payload mass
  - SELECT **CASE** … THEN... WHERE **SUBSTR()** : list the records which will display the month names [...] for the months in year 2015
  - SELECT … **ORDER BY** … **DESC** : to Rank the count of successful landing_outcomes in descending order

EDA with SQL Notebook link

Build an Interactive Map with Folium

- The following map objects were created and added to a folium map :
  - Circle and marker : to generate a map with all launch sites marked
  - A marker to a Marker Cluster : to populate the map with markers for all launch records (If a launch was successful, we use a green marker and if a launch was failed, we use a red marker). Many launch records will have the exact same coordinate. Marker clusters is way to simplify a map containing many markers having the same coordinate.
  - Mouse position : to get coordinate for a mouse over a point on the map
  - Polyline : To draw line between coordinates on the map

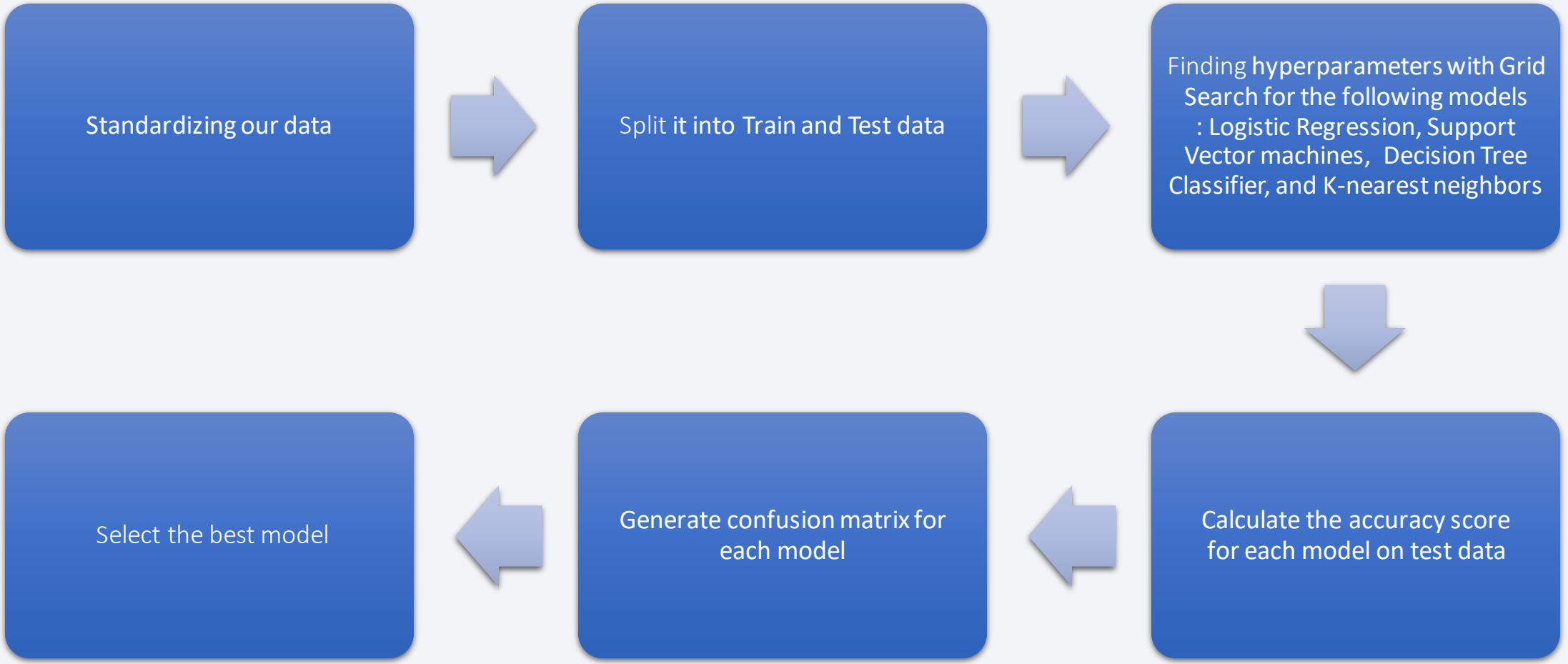
Interactive Map with Folium

Build a Dashboard with Plotly Dash

- A dropdown list of launch sites and a range slider of payload mass were added to a dashboard to interact with the following plot :
  - Pie Chart : to display launch success rate based on all sites or select launch site
  - Scatter Point Chart : to observe how payload may be correlated with mission outcomes for selected site(s)

Dashboard with Ploty Dash Notebook link

Predictive Analysis (Classification)



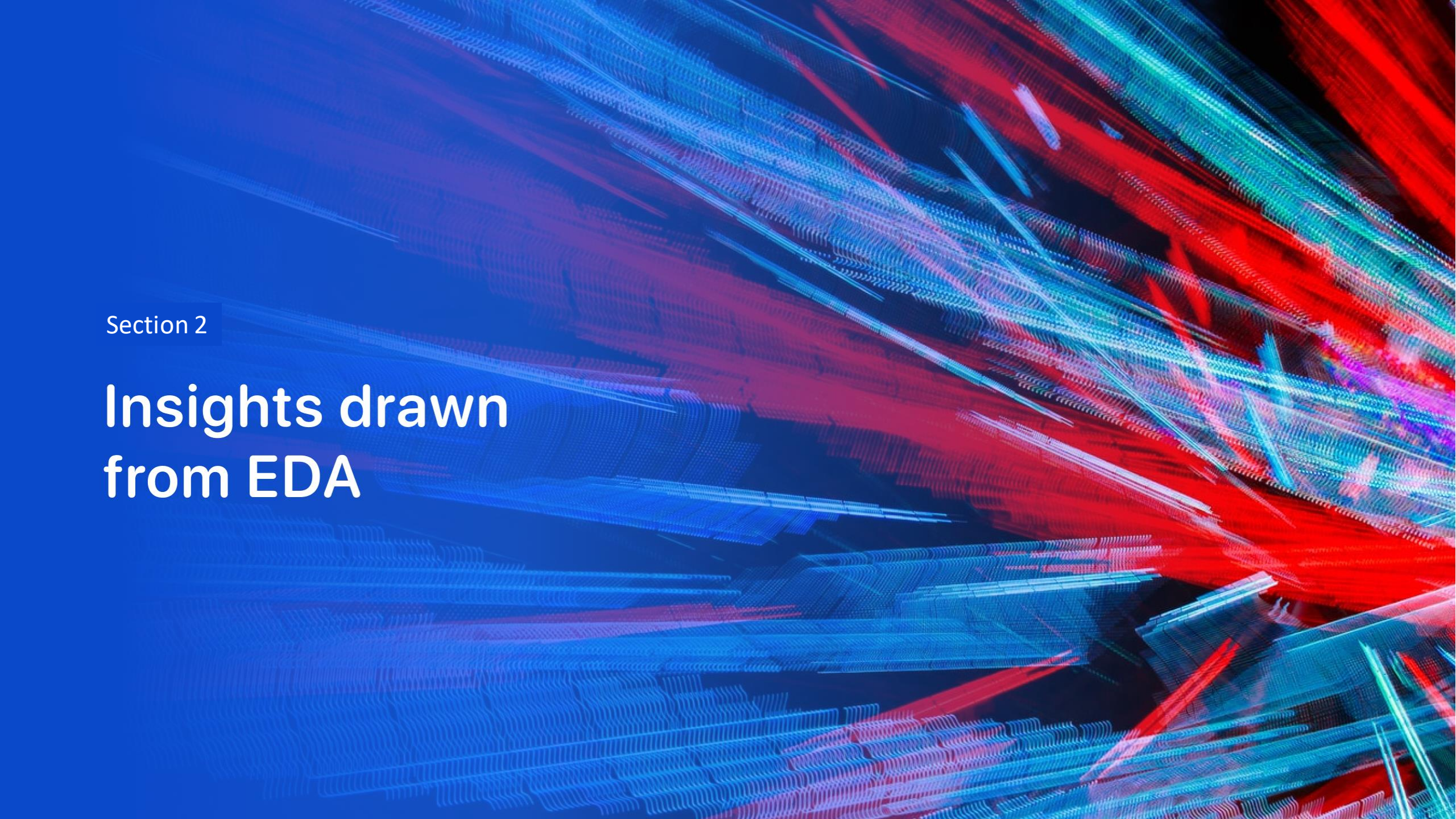
[Predictive Analysis (Classification) Notebook link]

Results

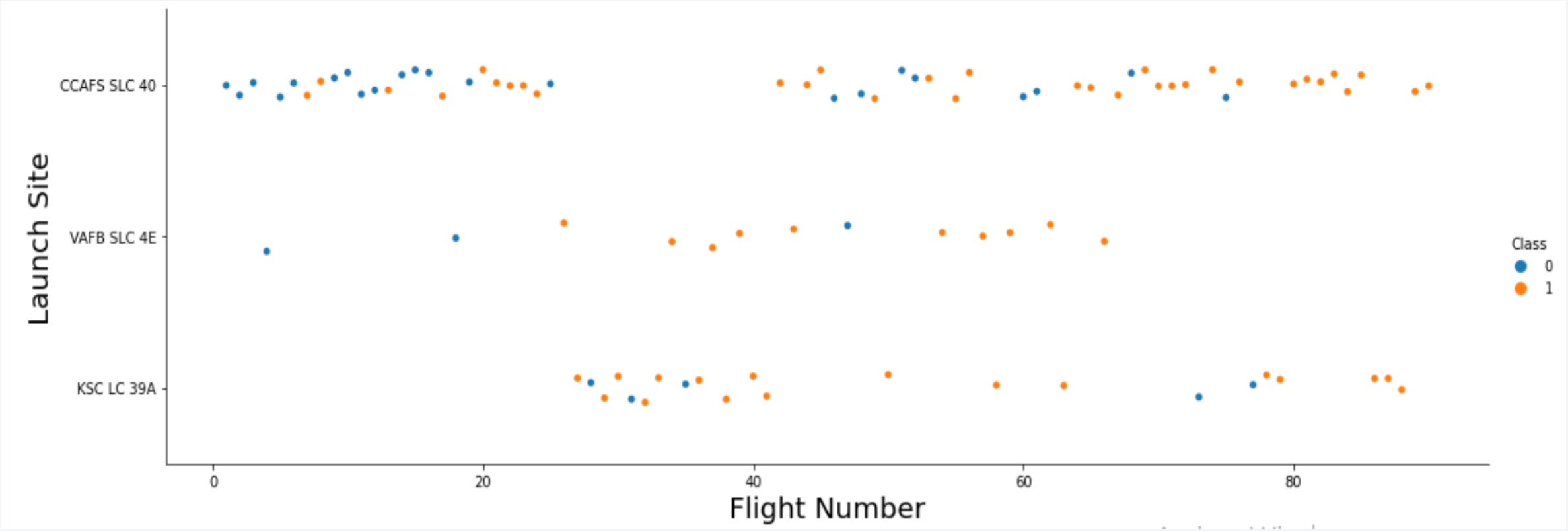
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



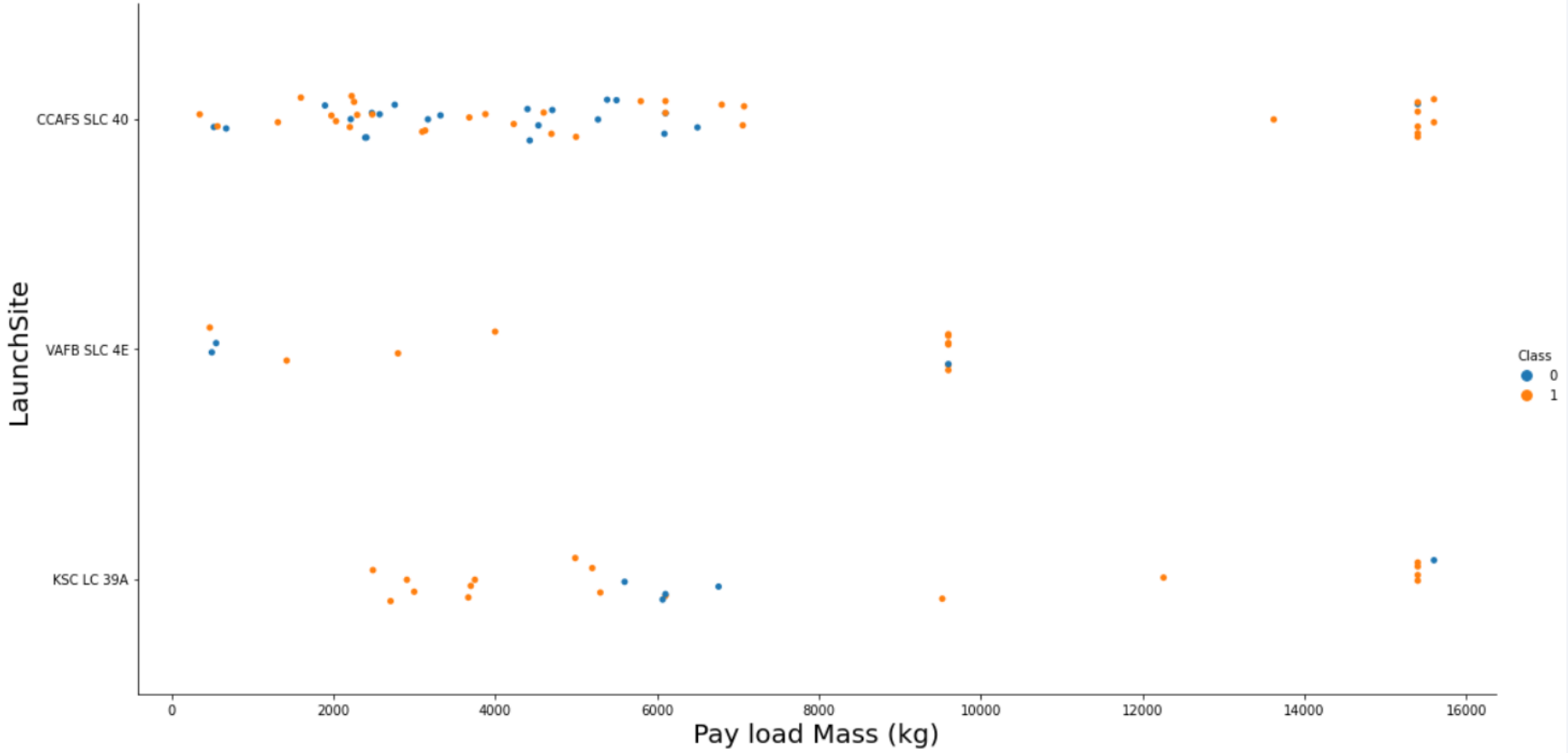
Flight Number vs. Launch Site



We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

We can also notice that for each launch site, the success rate improving with the flight number (more bad outcomes with first tries)

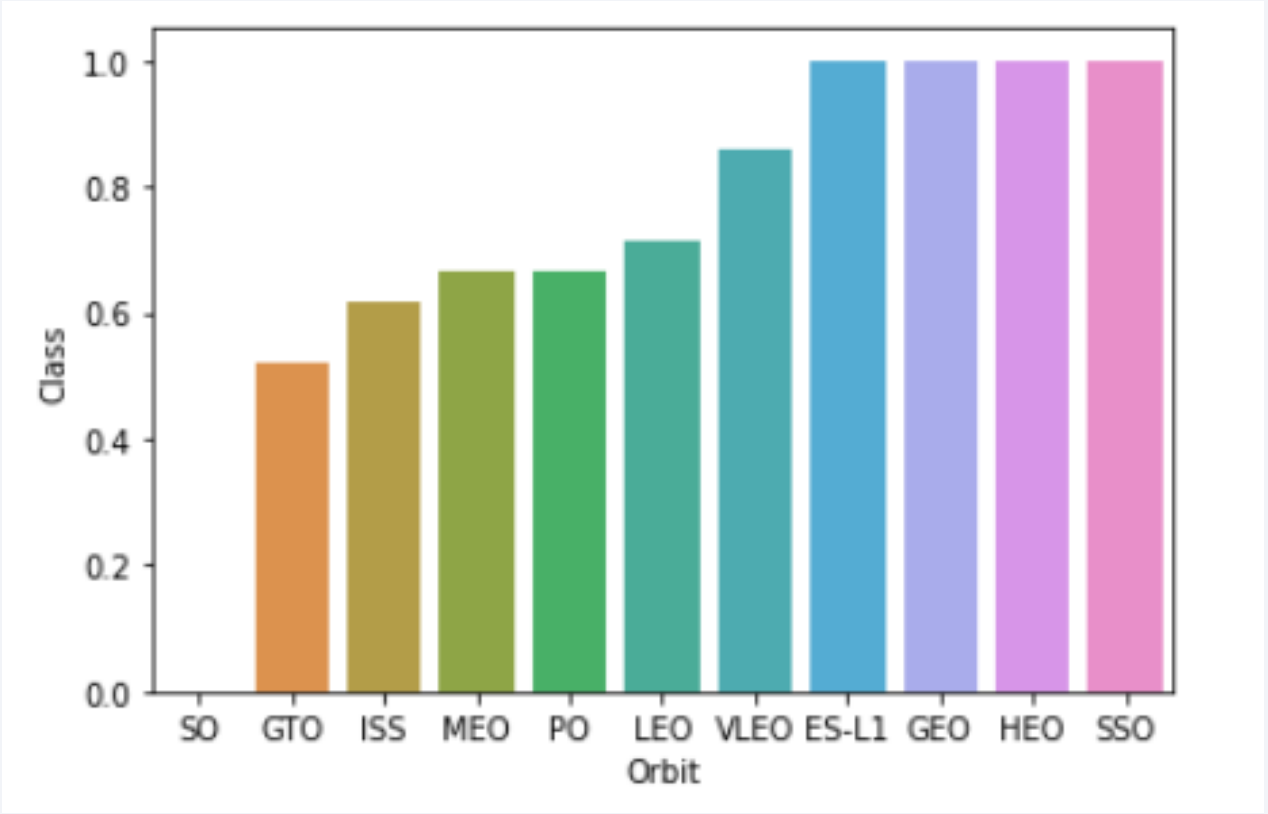
Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

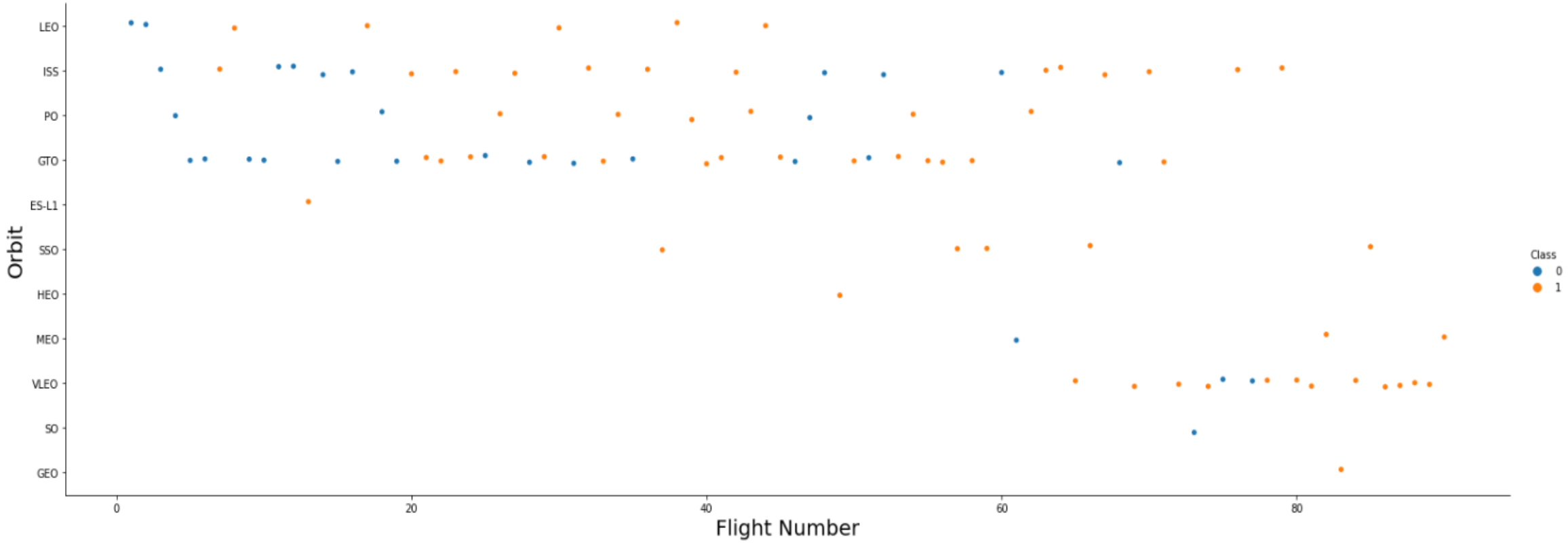
We can also observe that heavypayload (greater than 10000KG) have a success rate greater than 90%

Success Rate vs. Orbit Type



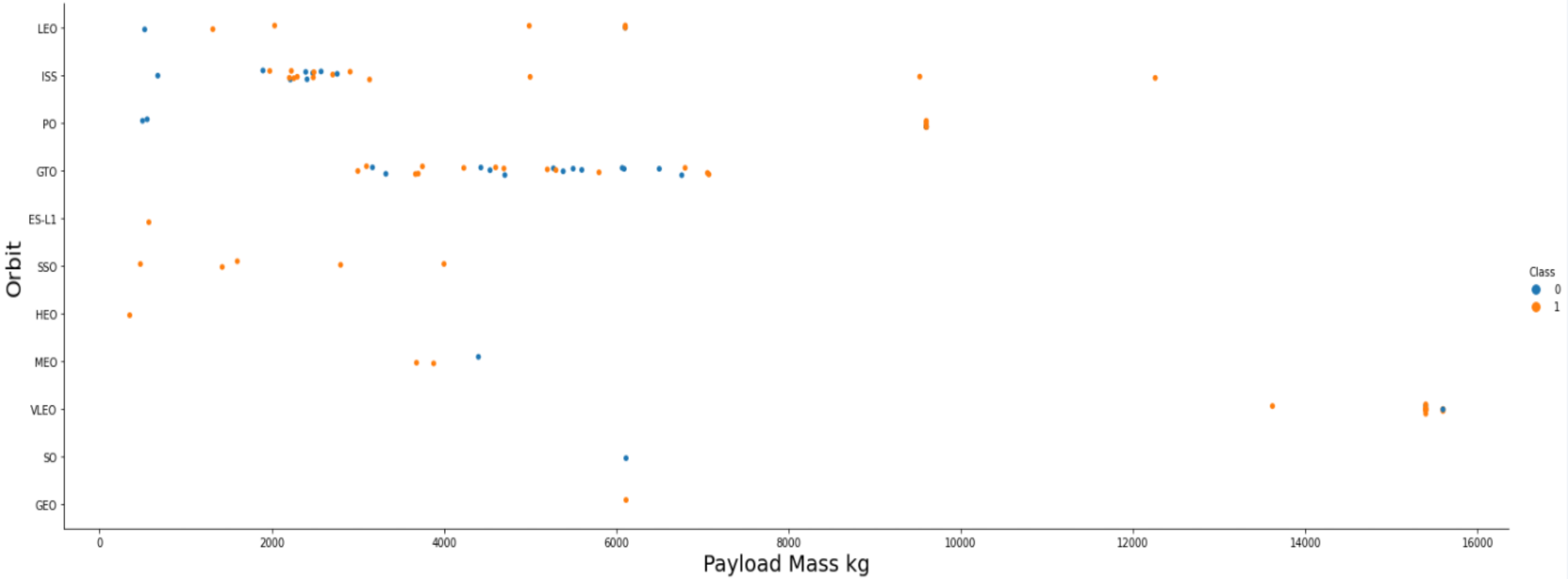
We can see that ES-L1, GEO, HEO and SSO have a success rate of 100% BUT the first 3 of them have only 1 landing attempt. and SO have a success rate of 0%.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

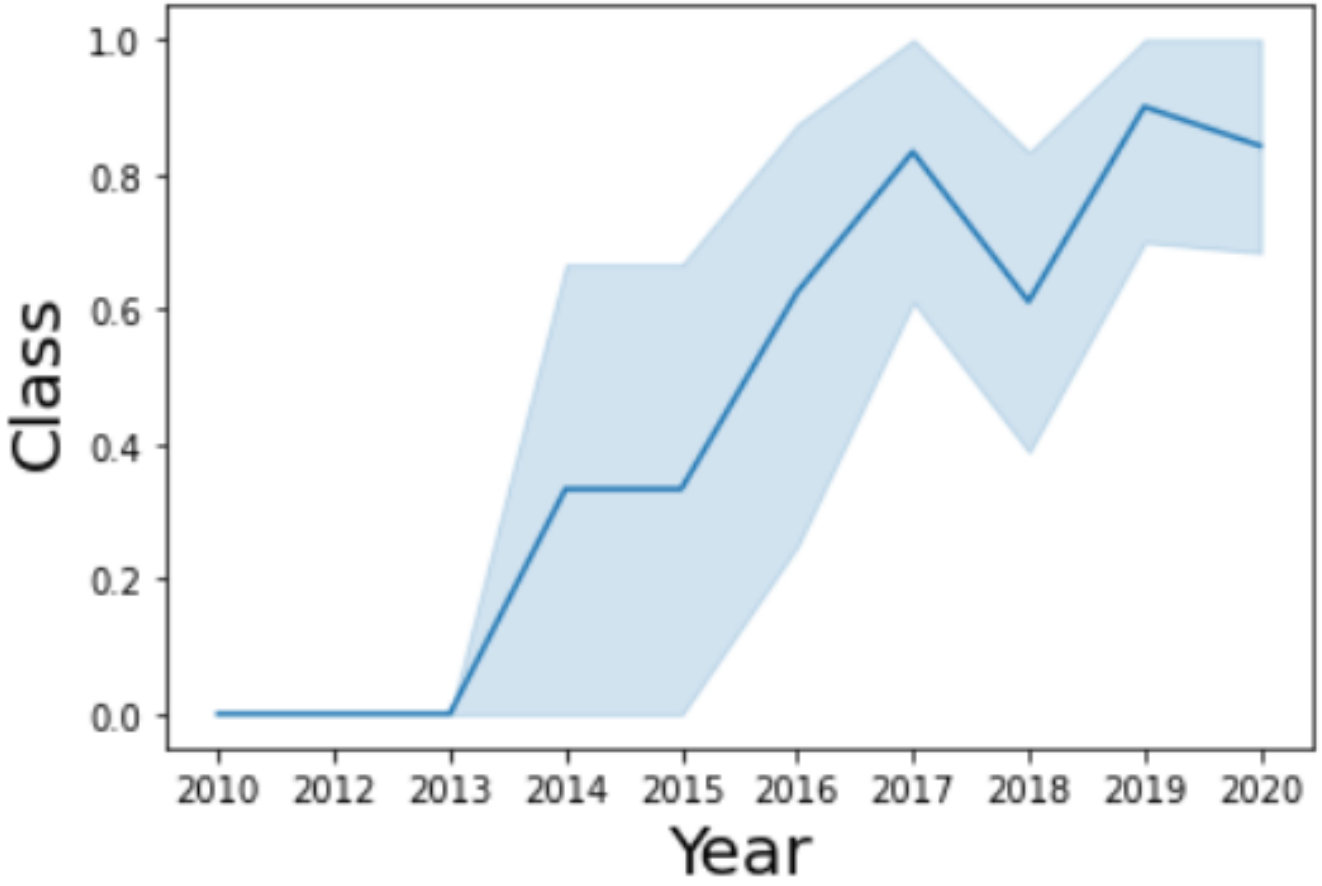
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

- There is only 4 Launch Sites in the space mission

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

We can observe launch sites begin with the string 'CCA' limited to the 5 first records

Total Payload Mass

SUM("PAYLOAD_MASS__KG_")
45596

The total payload mass carried by boosters launched by NASA (CRS) is equal to 45596 KG

Average Payload Mass by F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

The average payload mass carried by booster version F9 v1.1 is equal to 2928.4 KG

First Successful Ground Landing Date

Date
22-12-2015

The first succesful landing outcome in ground pad was achieved on December 22th 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Only four boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	COUNT("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful and failure mission outcomes. Only 1 of 101 mission outcome was a failure.

Boosters Carried Maximum Payload

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

12 different booster version have carried the maximum payload mass of 15600KG

2015 Launch Records

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

In year 2015, we had 2 failed landing in drone ship. One in January and the other in April, both where launch from CCAFS LC-40

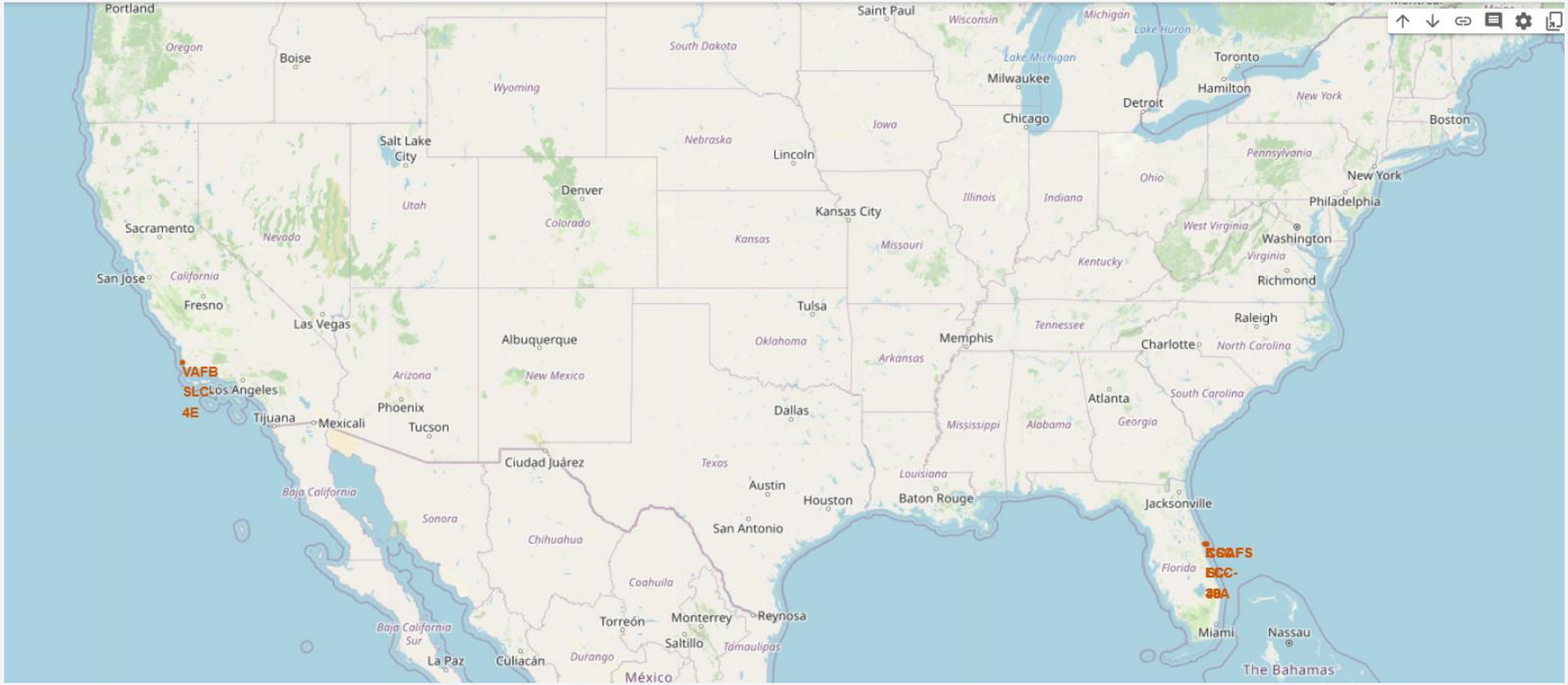
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
14-01-2017	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
27-05-2016	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

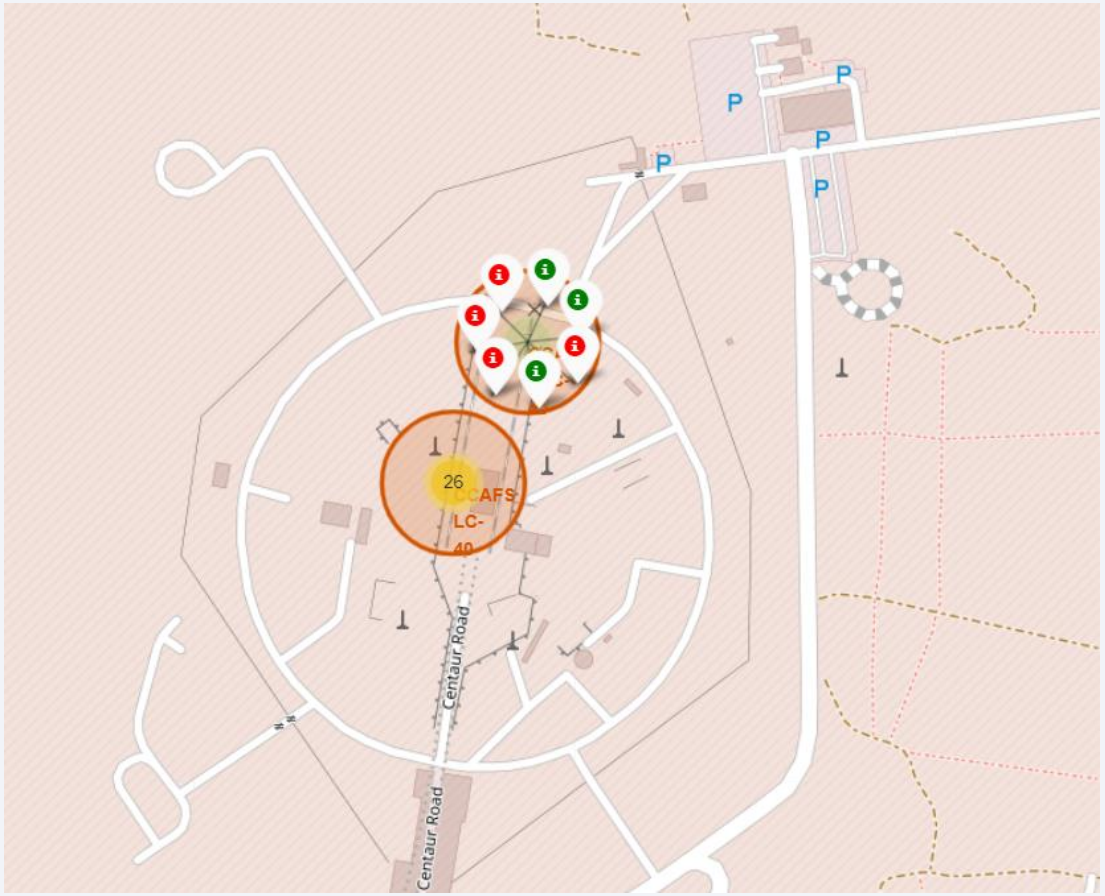

Section 3
Launch Sites
Proximities Analysis

Launch Sites location on Folium map



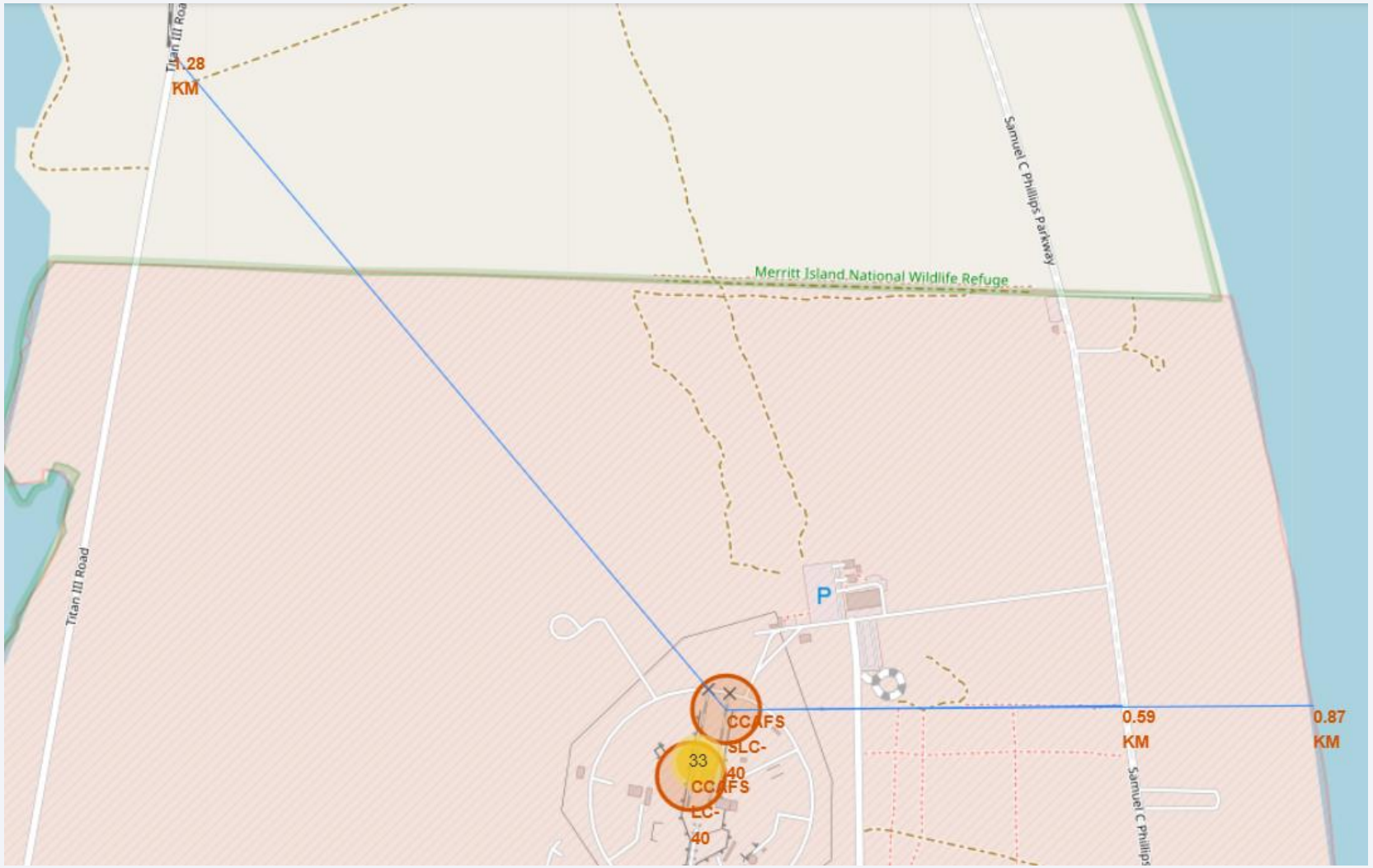
We can confirm that all launch sites are close proximity to the coast and in proximity to the Equator line

Succes/Failed Launches for each site on Folium map



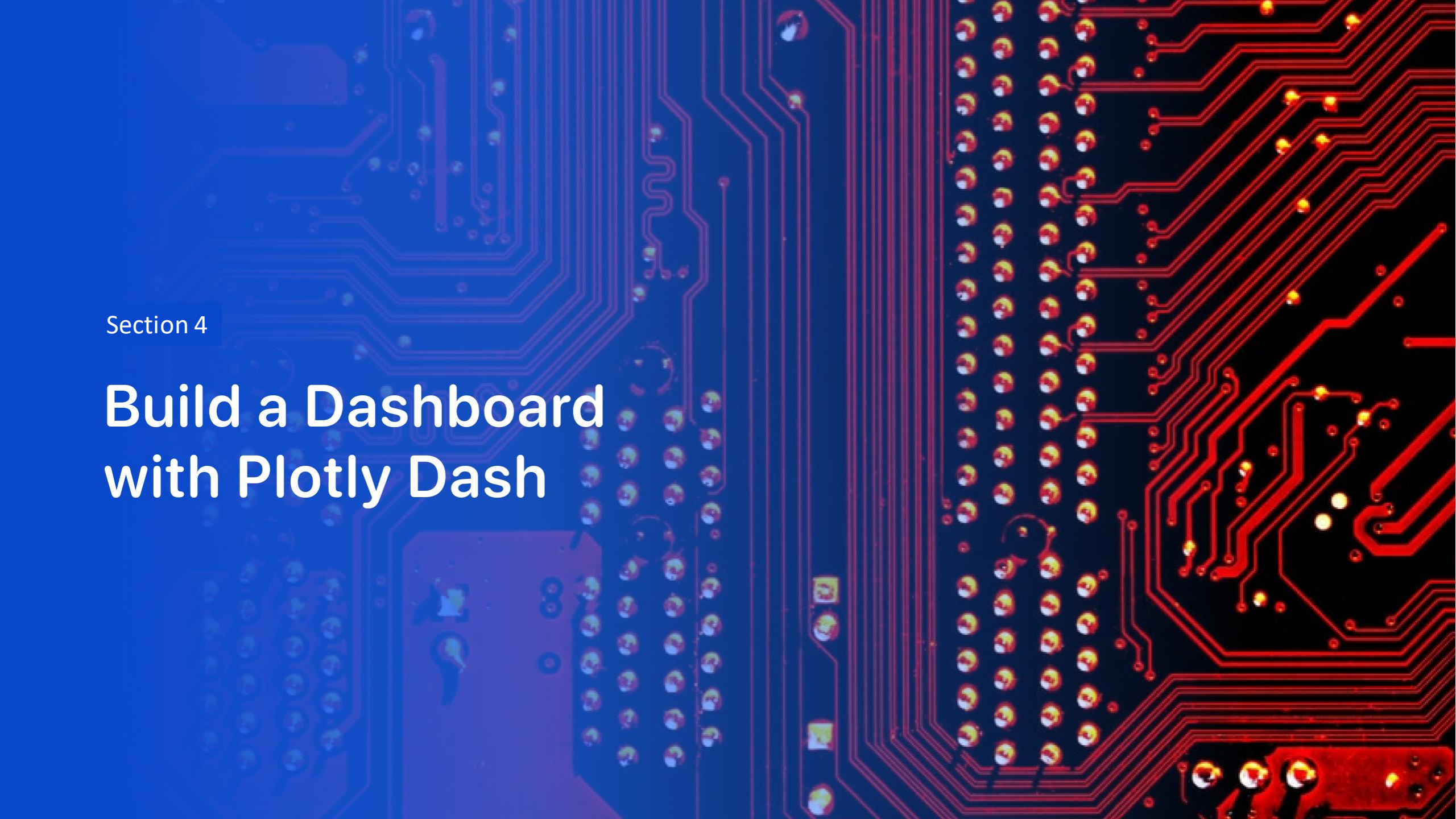
With the color-labeled markers in marker clusters, green for successful launch and red otherwise, we are able to easily identify sites with relativly high success rate. KSC LC-39A come first.

Launch site and its proximities on Folium Map

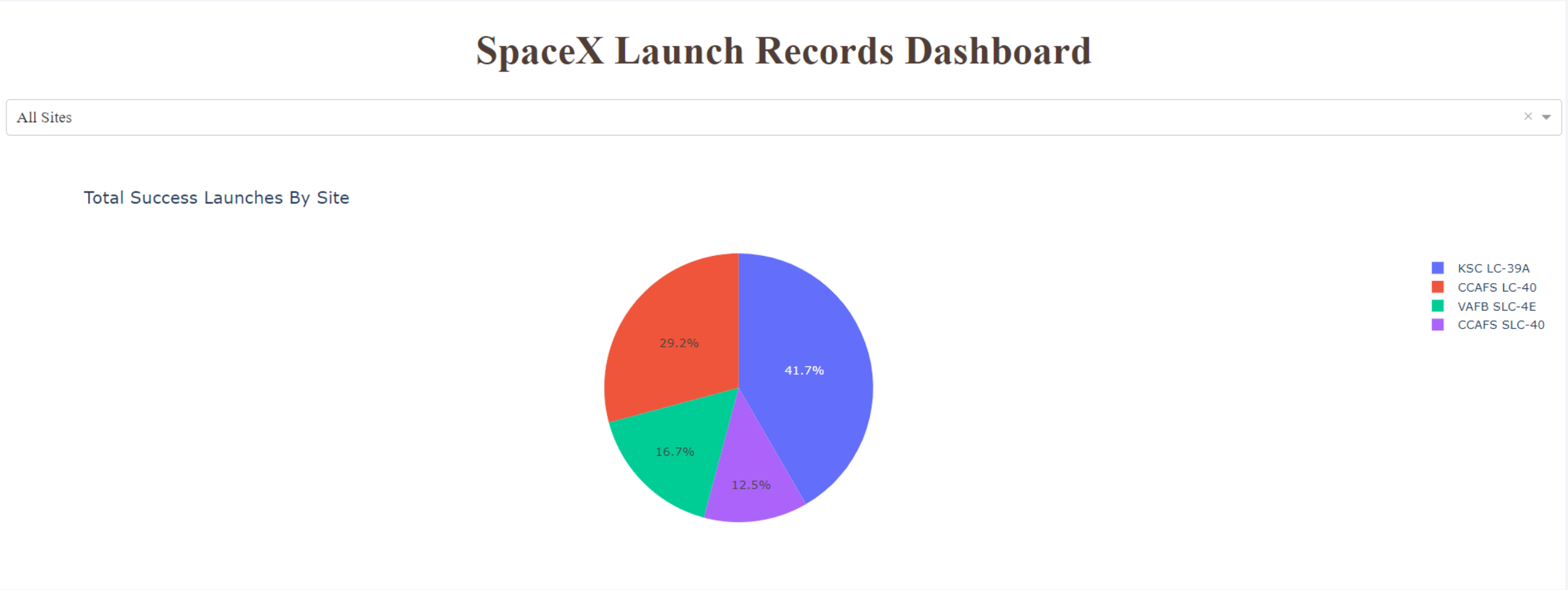


When we select CCAFS SLC-40 we can observe that this launch site is kept at a certain distance away from cities and at close proximity to coastline (0.87KM). When we analyze distance from road (highway or railway) we can see that the closest high way is at 0.59 KM when the closest railway is at 1.28 KM.

Proximity from Launch Site have to be taking account for security measure.

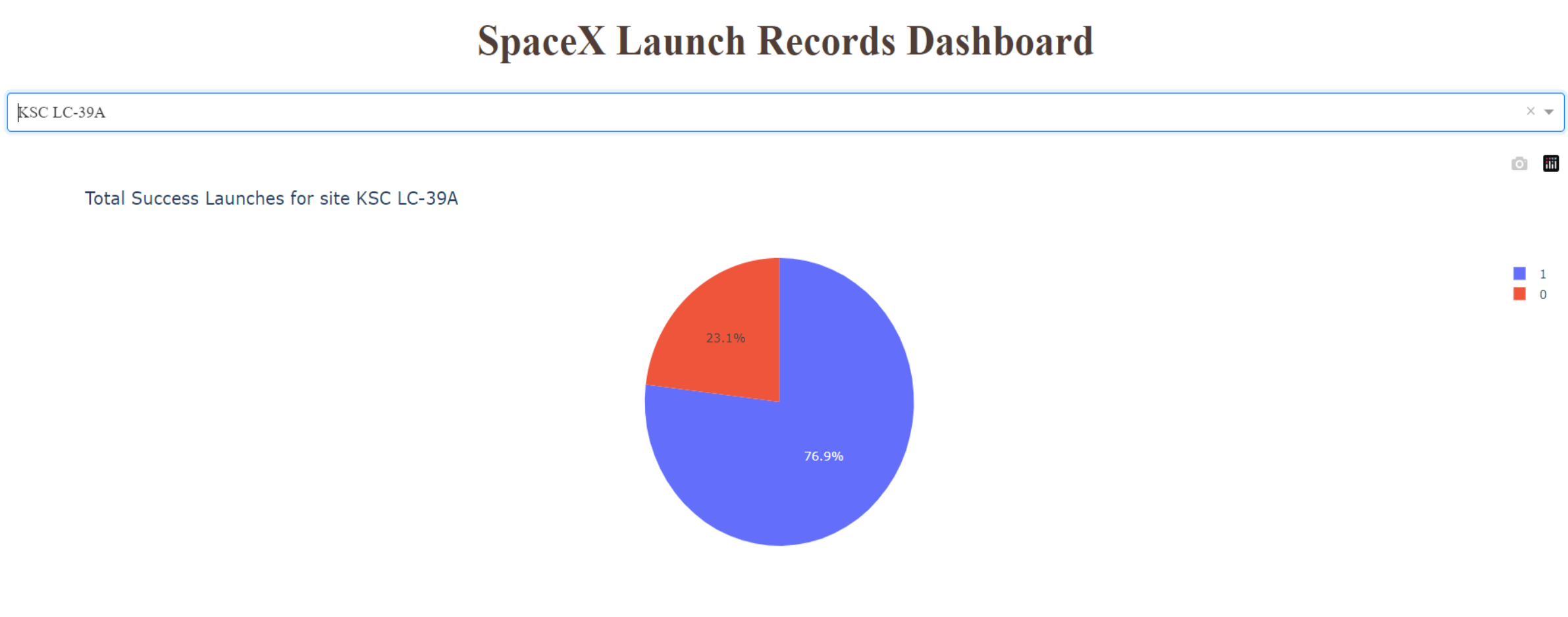

Section 4
Build a Dashboard with Plotly Dash

Success Launches counts by Site on Dashboard



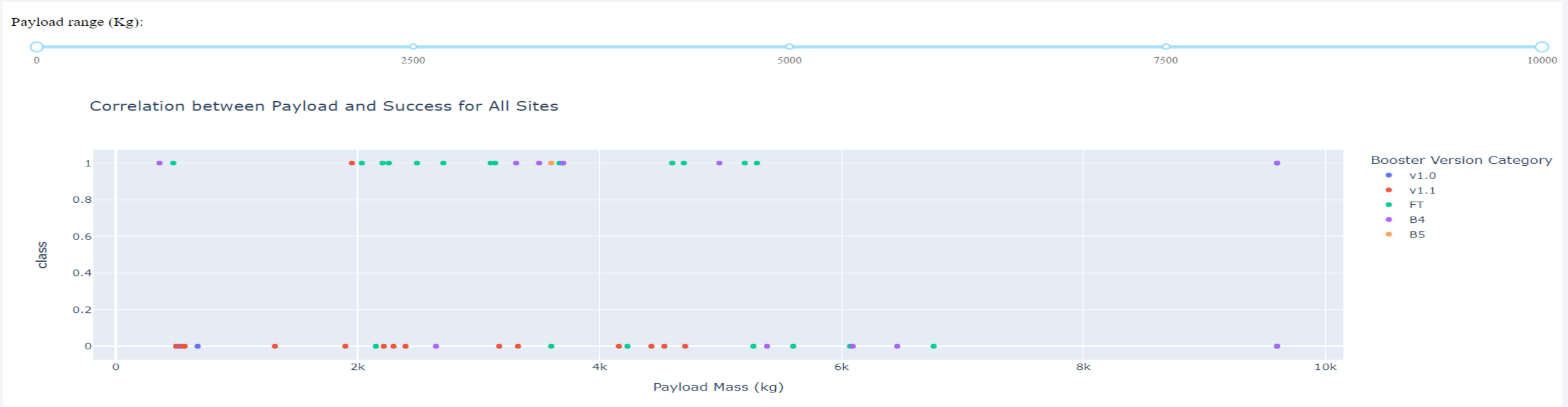
KSC LC-39A have the most successful launch counts with 42% of all success launches, and the site with the less successful counts with 12.5% is CCAFS SLC-40

Launch Site with highest launch success on Dashboard



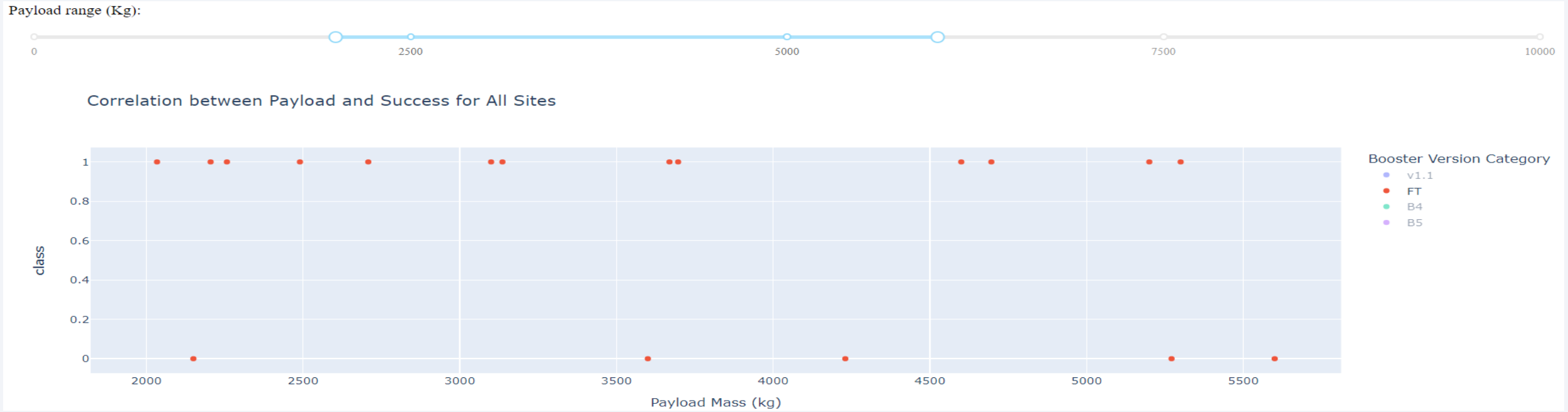
In addition of being the site with most successful launch counts, KSC LC-39A have the highest success ratio with 77% of successful launches.

Payload vs. Launch Outcome on Dashboard



We can observe that there is a correlation between Payload mass/Booster Version with launch outcome.
Most of successful launches have a payload mass between 2000 and 6000 KG.

Payload vs. Launch Outcome on Dashboard (next)



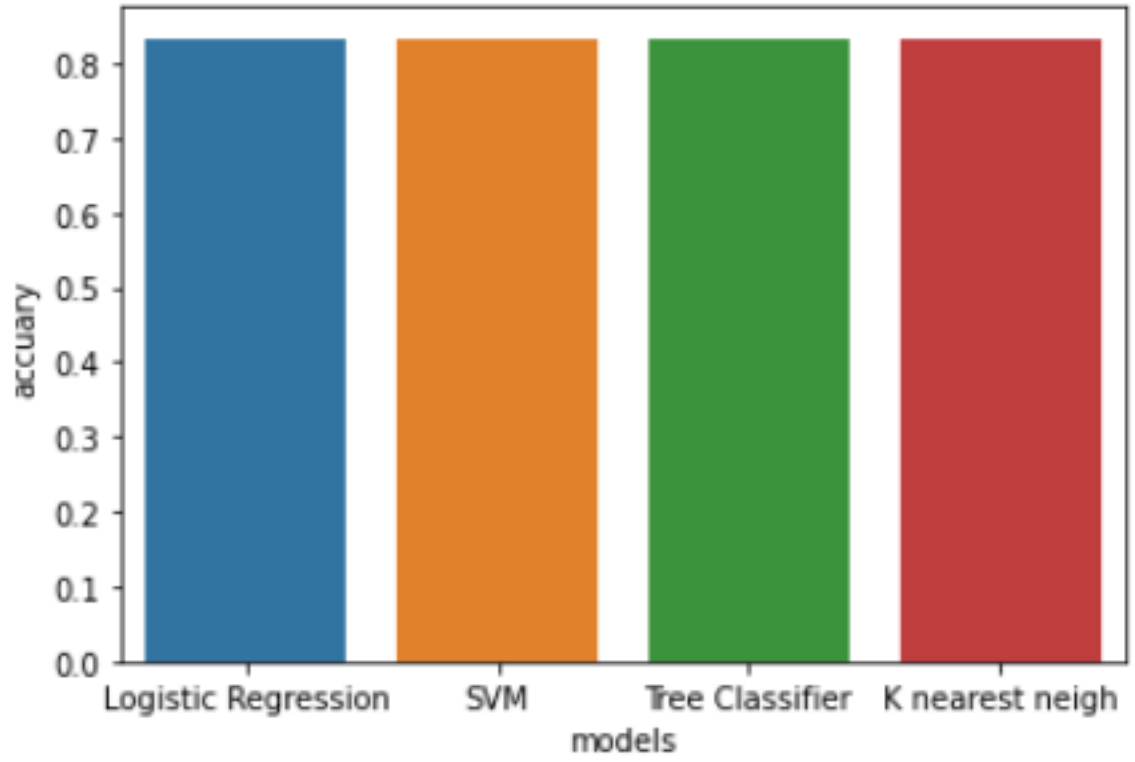
The most successful Booster Version Category is "FT" with the highest success rate specially when the payload mass is under 6000KG.

Section 5

Predictive Analysis (Classification)

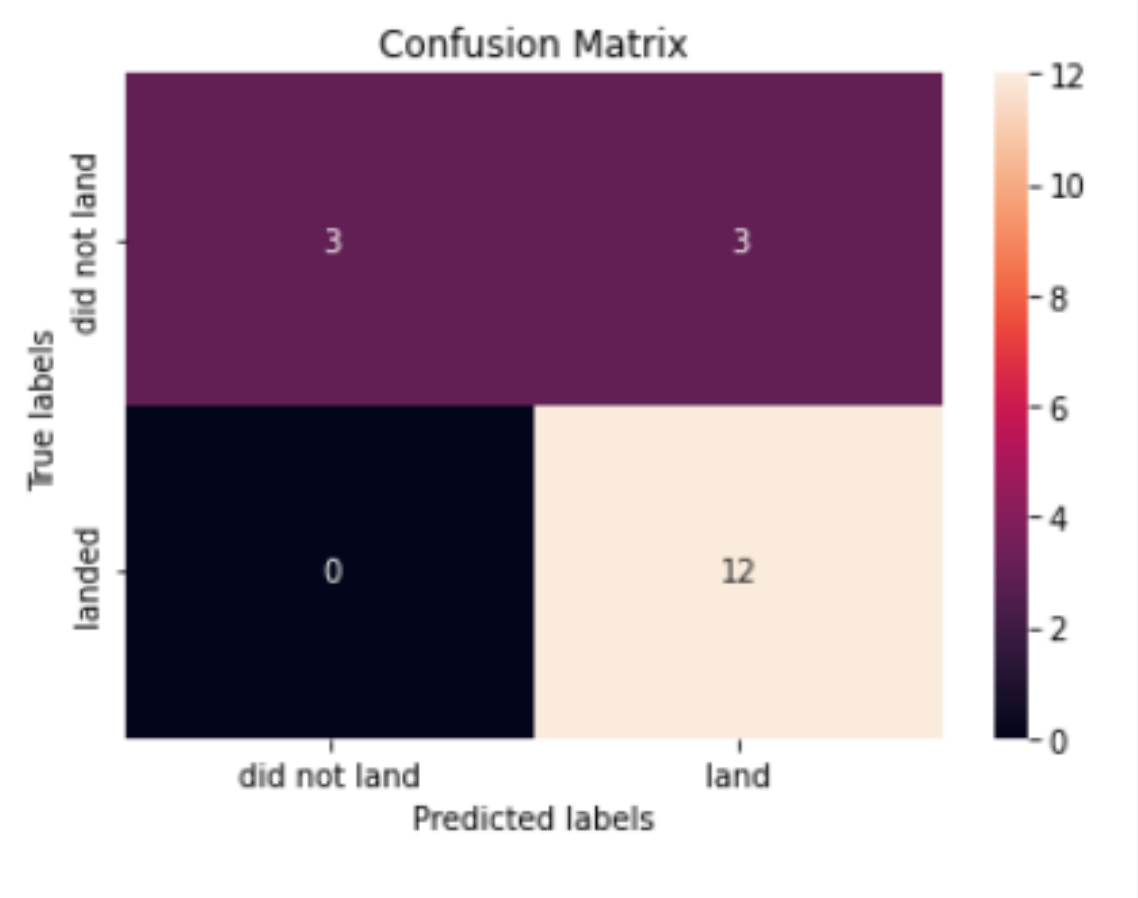
Classification Accuracy

```
Log reg : acc score  0.8333333333333334
SVM : acc score  0.8333333333333334
Tree class: acc score  0.8333333333333334
k nearest neigh : acc score  0.8333333333333334
```



All 4 models have the same accuracy score of 0.8334

Confusion Matrix



All 4 methods have the same confusion matrix results (doing so they have the same precision, recal, f1score and also the same AUC_ROC cur ve).
All 4 models can distinguish between different classes :
-True Positive, True Negative and False Negative have been predicted very well

The problem is with False Positive, where 3 launches that not land have been predicted as success landing

Conclusions

The main goal being to be able to predict if the 1st stage will land successfully, we must analyze which attributes give best chances for a successful landing :

- The Launch Site is a very important predictor, KSC LC-39A being the launch site with the highest success rate 77%, it should be prioritize as launch site for compagnies. particularity of this launch site should be analyze, the success of this launch site is it related to proximities ? Or is it just coincidence with the following attributes ?
- The Payload mass is also very important. Heavypayload (greater than 10000KG) are the best choice for chance of successful outcome. Otherwise we should choose payload between 2000 and 6000 KG.
- Booster Version Category is "FT" should be also prioritize when the payload is under 5500KG because the success rate for this booster with this payload have the highest success rate 76%.
- Orbit type that should be prioritize is SSO with 100% success rate on their 5 launch attempt.

Our build models predict good the landing outcome but have to be improve for the false negative prediction.

Thank you!