IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diego Sánchez
19/08/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection methodology
  - Perform data wrangling
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

SpaceX has gained worldwide attention for a series of historic milestones.

It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

In this presentation we are going to show the process to determine if the first stage will land successfully and its cost.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data collection was made by making a get request to the SpaceX API using the library ‘requests’. After that, we decoded the response content as a Json file and turned it into a dataframe.

- Perform data wrangling

  Once we had all the data, we filtered the dataframe to only include Falcon9 launches. Then, we dealt with missing values in some columns replacing them with the average value of that 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

  In the final stage, we applied four machine learning methods to predict if the first stage of a launch will land successfully.

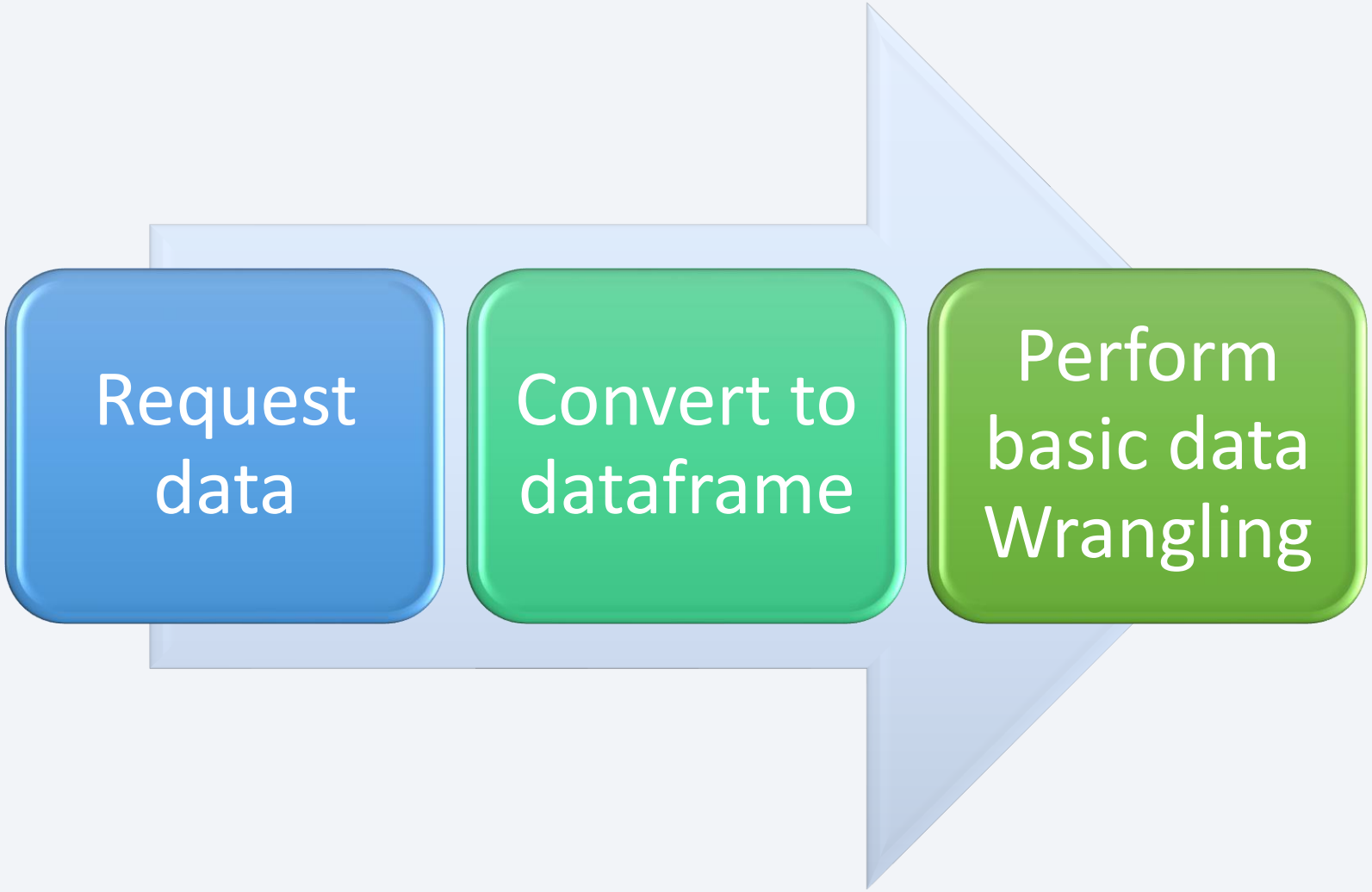
Data Collection

Data collection was made by making a get request to the SpaceX API using the library ‘requests’. We saved the requested information in a ‘response’ object. This object was a Json file, so we converted it to a Pandas dataframe. After that, we selected only the information we wanted from this first dataframe and created a new data frame with information about DATE, FLIGHT NUMBRER, BOOSTER VERSION, PAYLOAD MASS, LAUNCH SITE, OUTCOME, ORBIT, etc.

We are only interested on FALCON 9 launches, so we filtered the data to only include FALCON 9 records. We observed that some columns had missing values, so we replaced them with the average value of that column.

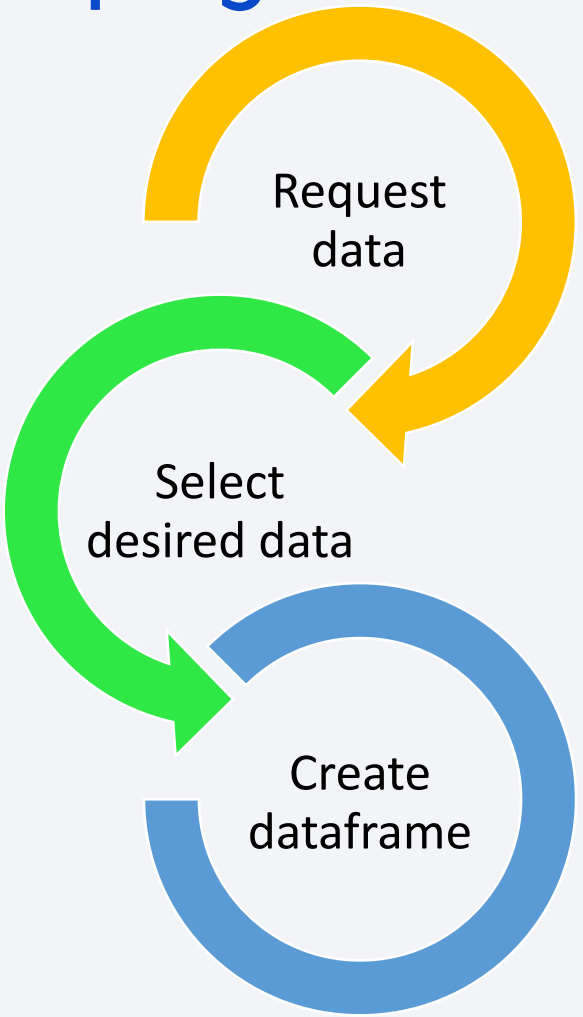
We also performed web scrapping to collect FALCON 9 historical launch records from Wikipedia page ‘List of Falcon 9 and Falcon Heavy Launches’. We extracted the FALCON 9 table using BeautifulSoup library. Then, we created a dataframe: first we got the column names of FALCON 9 table and then we filled up a dictionary with Keys=column names and their values. With this dictionary, we created a new dataframe for launch records.

Data Collection – SpaceX API



GitHub URL:

Data Collection - Scraping



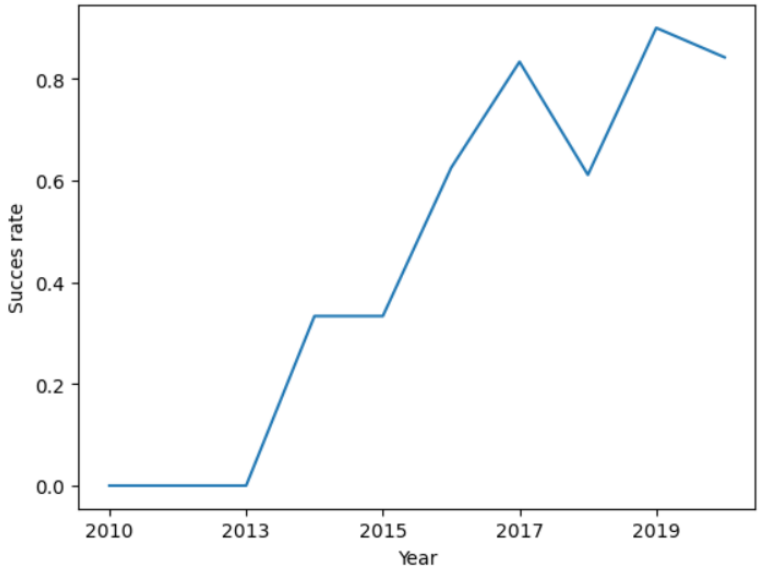
GitHub URL:

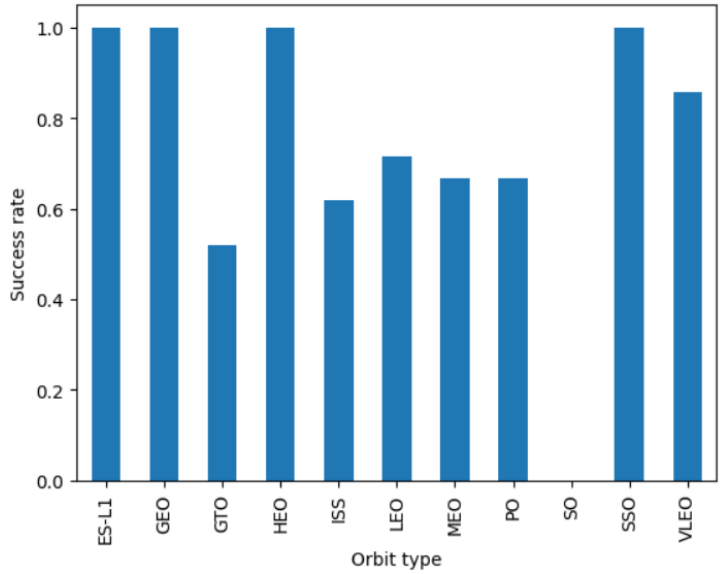
Data Wrangling

Once we had all the data, we filtered the dataframe to only include Falcon9 launches. Then, we dealt with missing values in some columns replacing them with the average value of that column.

After that we calculated the number of launches at each site, the orbit occurrence, the mission outcome occurrence and we created a new column called ‘class’ which we filled up with 1 (success) and 0 (fail) values for mission outcome

GitHub URL





EDA with Data Visualization

We plotted some scatter plots to see the relationship between Flight Number and Payload Mass, Flight Number and Launch Site, Launch Site and Payload Mass, Flight Number and Orbit type, Payload Mass and Orbit. In addition, que plotted a bar chart to see the success rate of different orbits and a line chart to observe the success rate over the years.

We plotted these graphs to find the most relevant variables in the success of a launch.

GitHub URL:

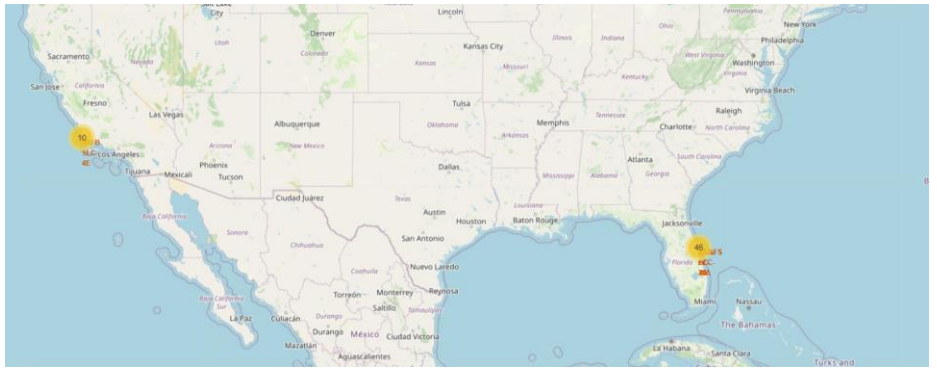
EDA with SQL

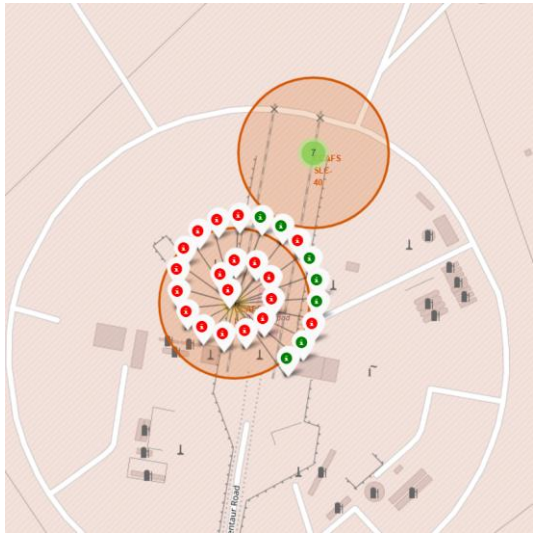
We performed some SQL querys to find some insights of our data:

- We displayed the names of unique launch sites
- We displayed the average payload mass carried by booster version F9 v1.1
- We found the date of the first successful landing outcome in ground pad
- We calculated the total number of successful and failure mission outcomes
- We listed the booster version which carried the maximum payload mass
- We listed all failure records in 2015 and all records outcomes between June 2010 and March 2017

GitHub URL:

Build an Interactive Map with Folium





In order to see the successful and failure outcomes in each location, we created a map with Folium library. On this map we marked the four launch sites with a red circle, then we added the number of total launches performed at each location and if those launches were successful or not. Using an interactive map allows as to easily select the location with the biggest success rate.

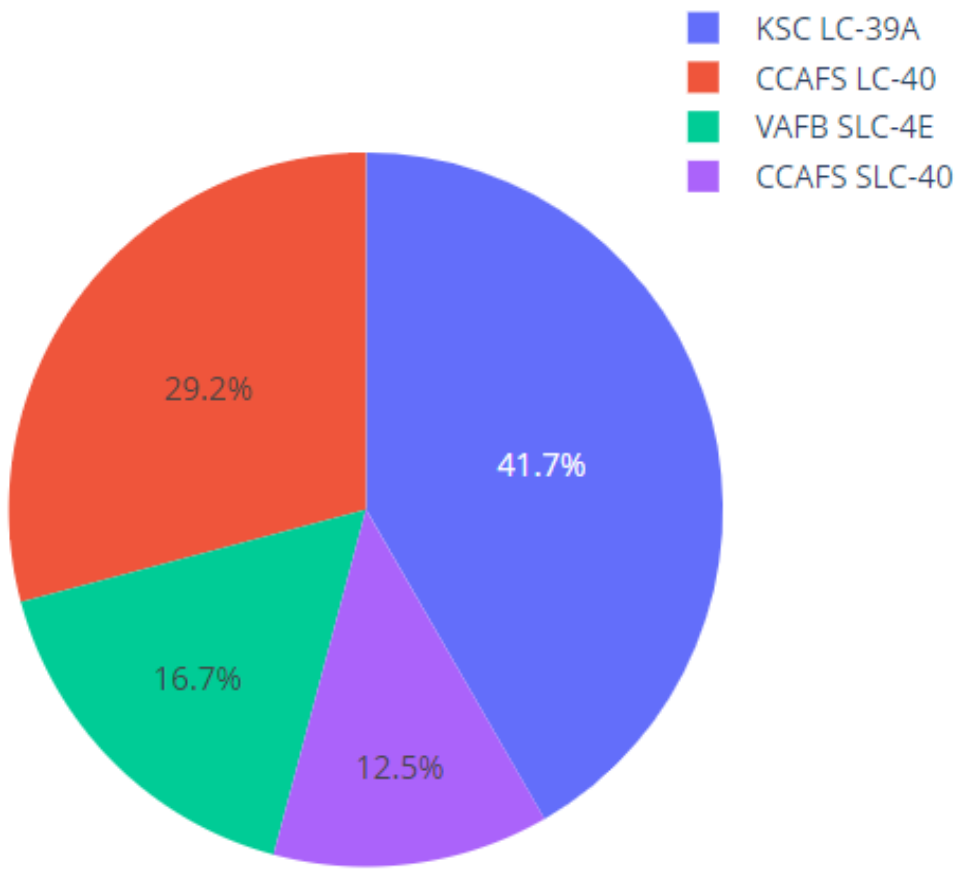
GitHub URL:

Build a Dashboard with Plotly Dash

After building a map, we created an interactive application to visualize the data. On this app we added two types of plots: success rate of each and all locations pie chart and a scatter plot showing the relation between success rate and payload mass for different values of payload mass.

We added this graphs because launch site and payload mass are the two variables that most affect the success or failure of a launch.

GitHub URL



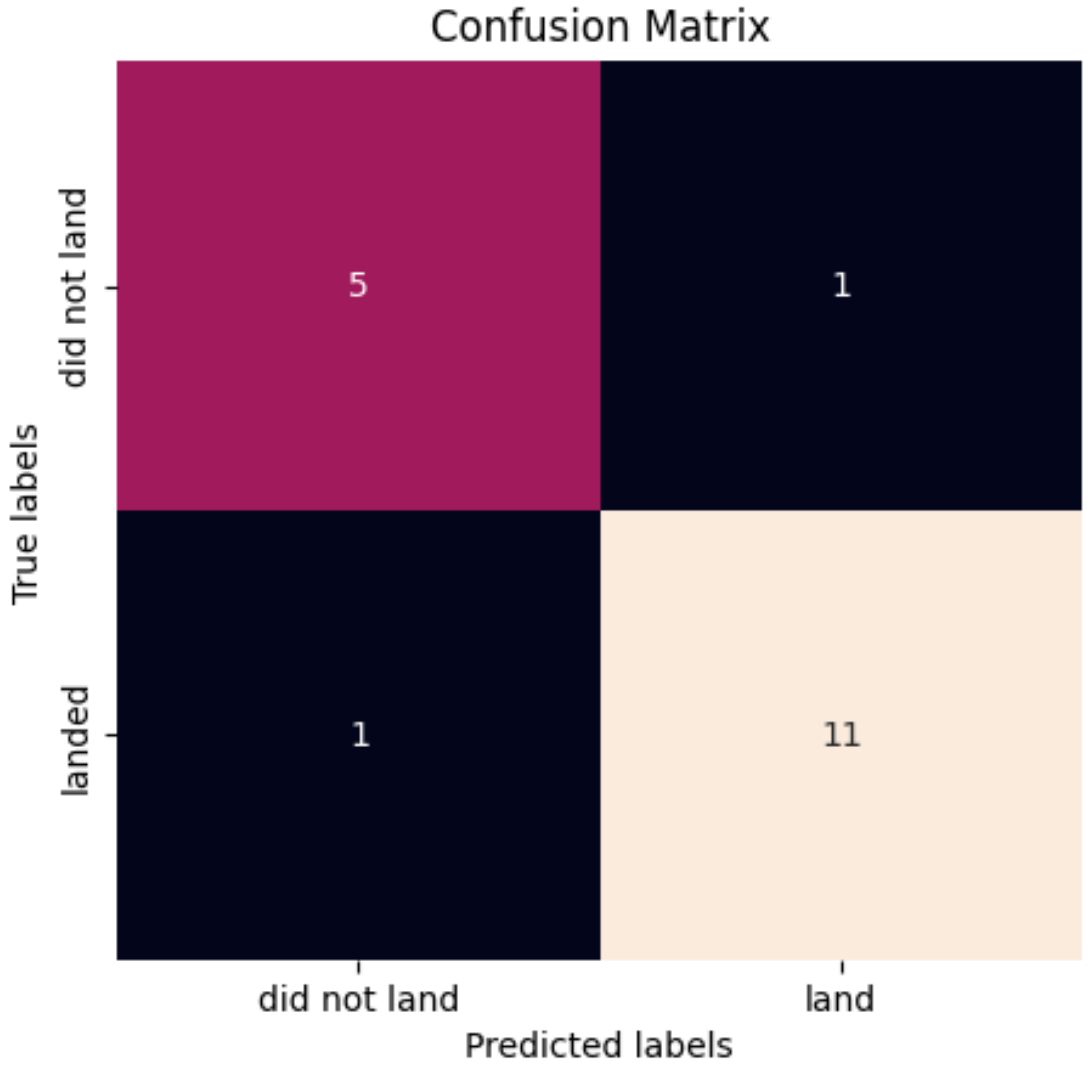
Predictive Analysis (Classification)

The main objective of this presentation is to find the if a landing will be successful or not. To find this we built four machine learning models (logistic regression, SVM, decision tree and KNN). We divided the data in two parts: class column (1 if success, 0 if fail) and the rest of columns. After that, we splitted our data in train and test set, built the models using GridSearch.

We trained the models with some parameters, found best parameters and accuracy for each model and finally we plotted the confusion matrix. This matrix shows if the model predicts the correct outcomes.

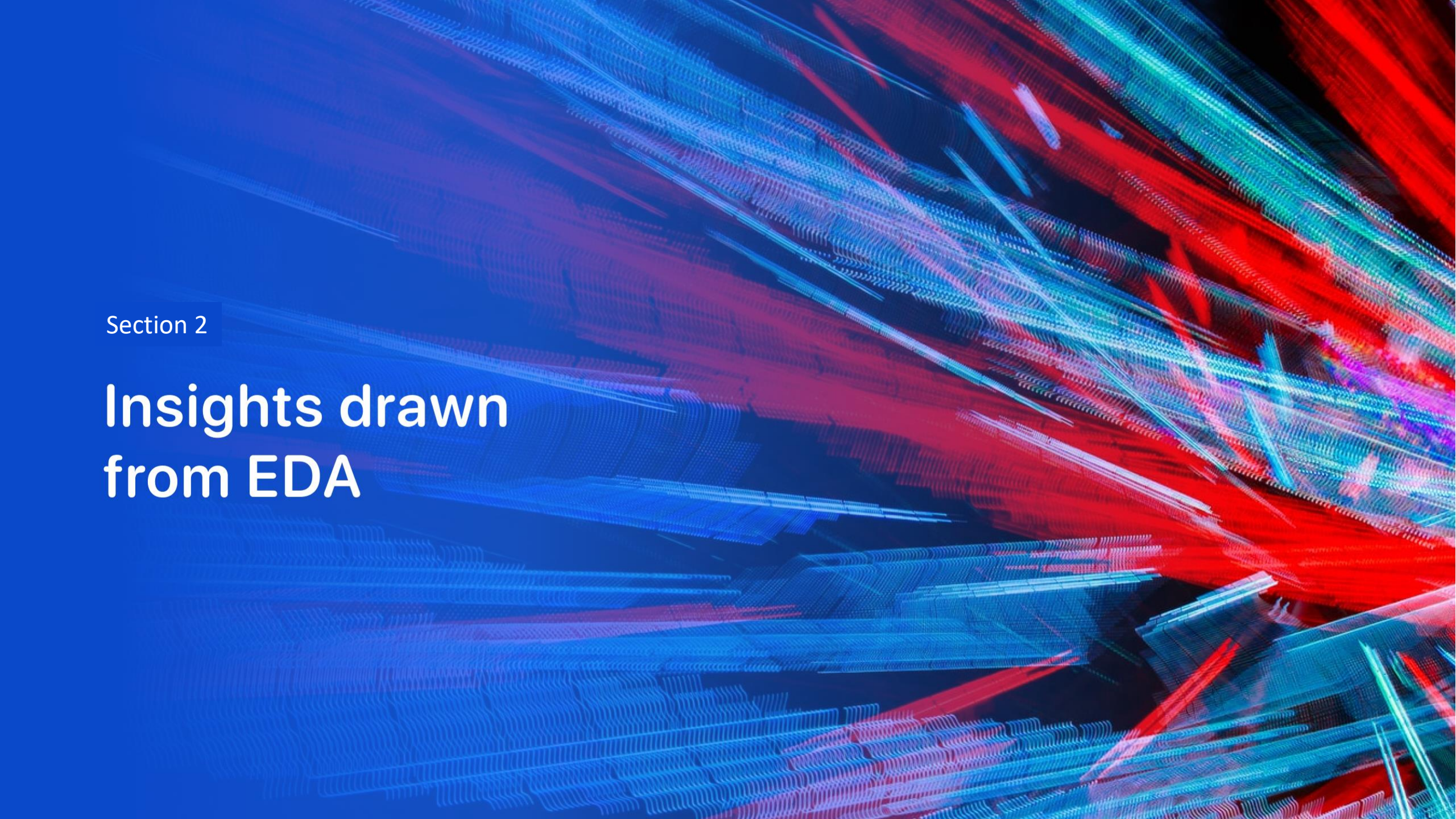
The matrix with the biggest number of true positives corresponds to the best predictive model.

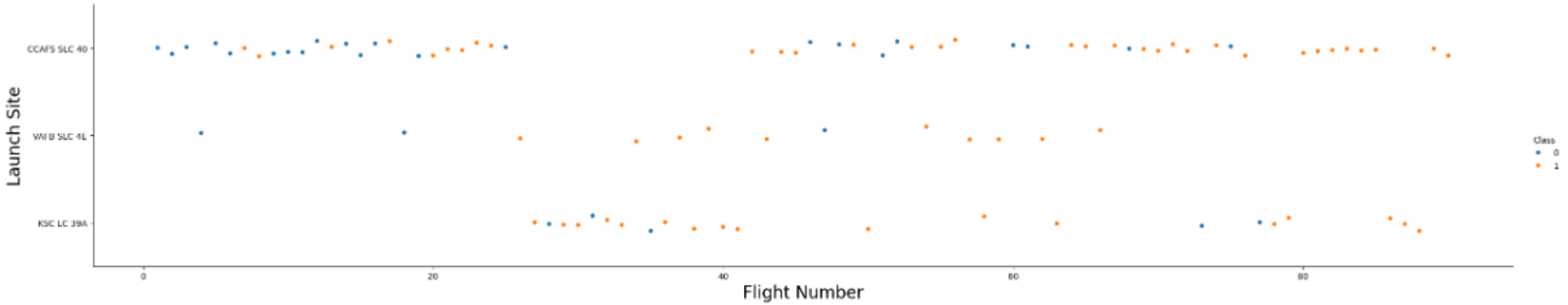
GitHub URL



Results

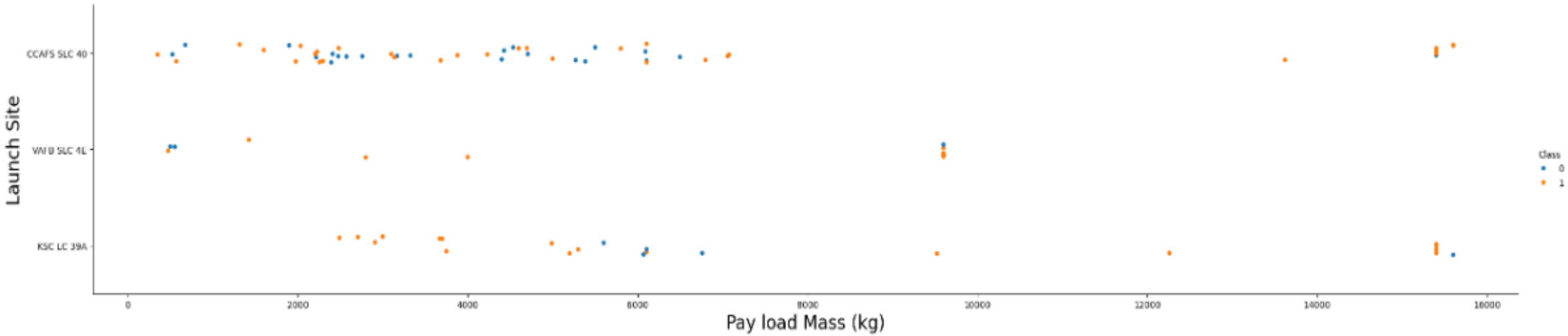
- Exploratory data analysis results
  - Scatter plots, bar chart and line chart showing relations between different variables
- Interactive analytics demo in screenshots
  - Interactive map and web app to get insights in success rate
- Predictive analysis results
  - Results and accuracy of four machine learning models


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

We can see that 23 of the first 25 flights were performed at CCAFS SLC 40 location and nearly 61% of them failed, so SpaceX decided to change the location to KSC LC 39A and VAFB SLC 4L for next 18 flights. The success rate was better in those locations, so they decided to try again CCAFS SLC 40 to see if the success rate would improve there. And it made it as we can see at the end of the chart (flight number 80 and next). It is clear that tests performed in KSC LC 39A and VAFB SLC 4L helped to improve the success rate in the preferred location (CCAFS SLC 40).

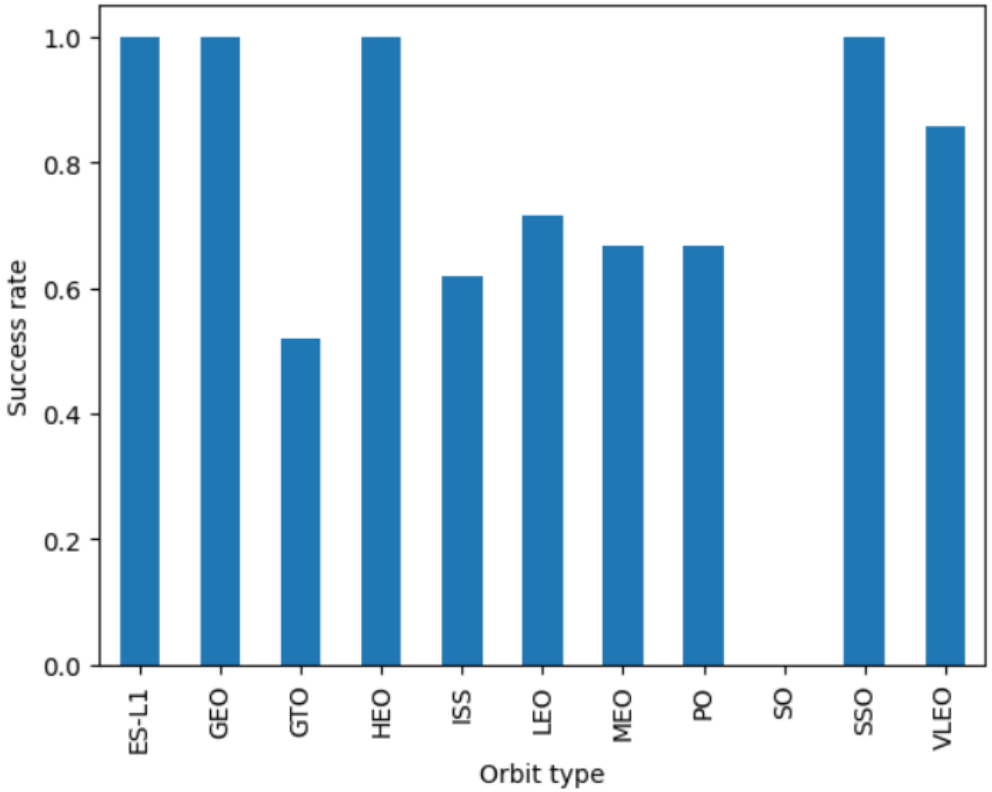


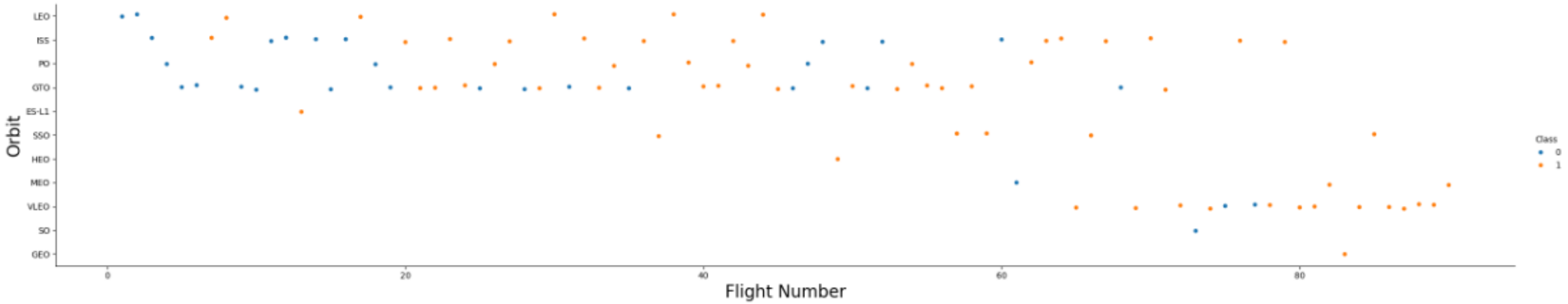
Payload vs. Launch Site

If we observe this chart, we find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000). In addition, we see that mostly light payload and heavy payload mass launches are performed in CCAFS SLC 40 launch site.

Success Rate vs. Orbit Type

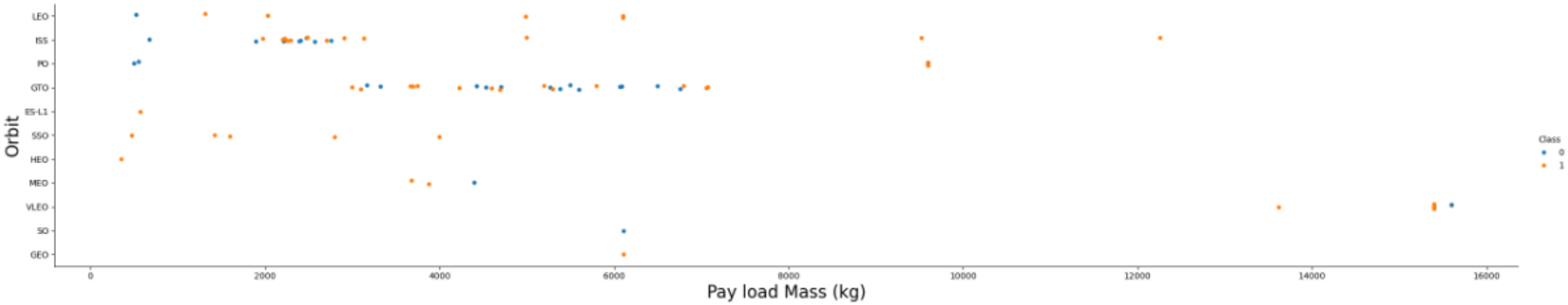
If we observe this bar chart, we find that ES-L1, GEO, HEO and SSO orbits have the biggest success rate.





Flight Number vs. Orbit Type

We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success. In addition, we can see that VLEO orbit was the preferred one after 60 flights with a big success rate.



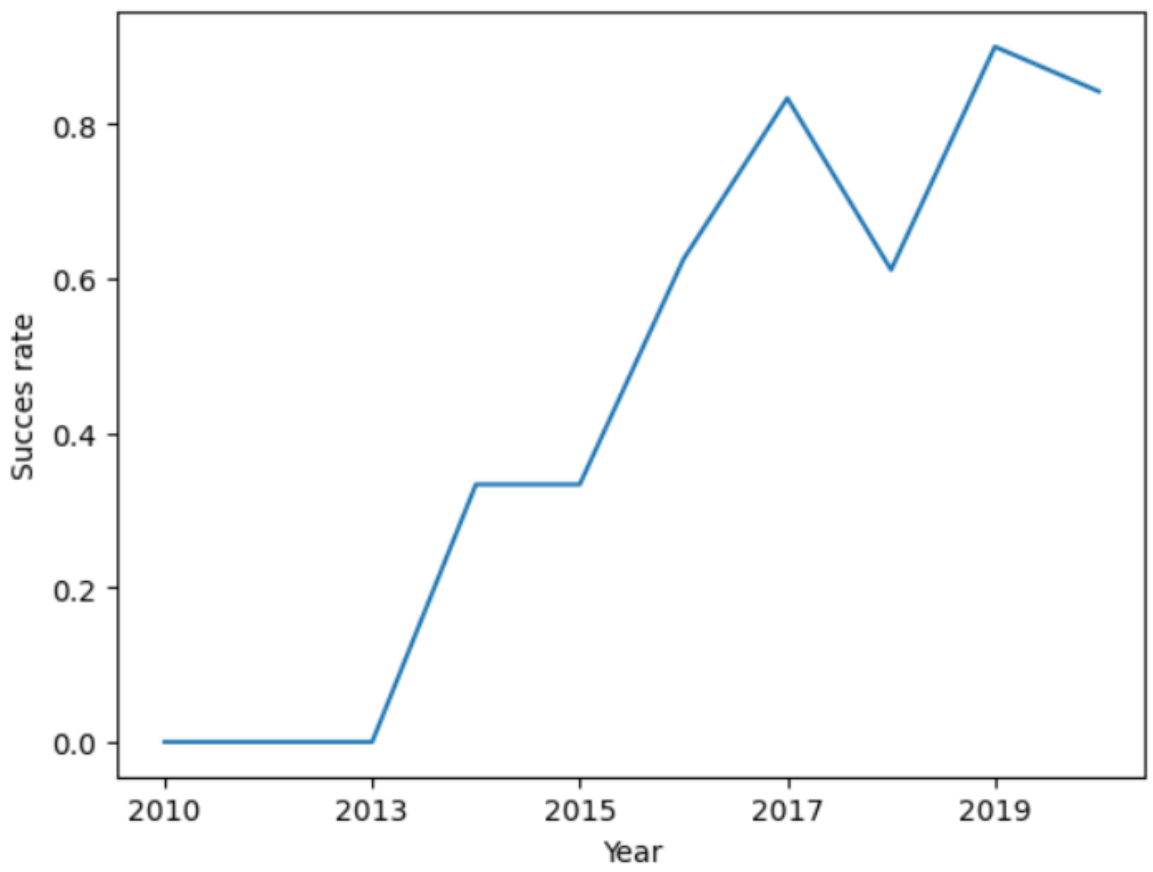
Payload vs. Orbit Type

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

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

We can observe that the success rate since 2013 kept increasing till 2020.



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

```
%%sql

SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

✓ 0.0s

```
* sqlite:///my_data1.db
Done.
```

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

All Launch Site Names

We use DISTINCT on column Launch_Site to select all the Launch sites without repetition

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

```
%%sql

SELECT * FROM SPACEXTABLE
    WHERE Launch_Site LIKE 'CCA%'

LIMIT 5
```

✓ 0.0s Python

* sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Launch Site Names Begin with 'CCA'

- We use WHERE LIKE syntaxis on column Launch_Site to select launch sites which begin with string ‘CCA’ and we add LIMIT 5 to show only five records.

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
    WHERE Customer='NASA (CRS)'
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Total Payload Mass

- We use SUM on PAYLOAD__MASS_KG_ column to display the total payload mass and we apply WHERE on Customer column to sum only payload mass carried by NASA (CRS)

Average Payload Mass by F9 v1.1

- We use AVG on PAYLOAD__MASS_KG_ column to display the average payload mass and we apply WHERE on Booster_Version column to display only payload mass carried by F9 v1.1

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

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
    WHERE Booster_Version='F9 v1.1'
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We use MIN on DATE column to display the first successful ground landing and we apply WHERE to Landing_Outcome column to select the required successful landing outcome.

```
%%sql

SELECT MIN(DATE) FROM SPACEXTABLE
    WHERE Landing_Outcome='Success (ground pad)'
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2015-12-22

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

```
%%sql

SELECT Booster_Version FROM SPACEXTABLE
    WHERE Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- We select all records on Booster_Version column WHERE Landing_Outcome is successful drone ship and payload mass BETWEEN 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

- We SELECT Mission_Outcome column and count all records on that column. Then we GROUP them BY Mission_Outcome result. We see that only one mission failed and a hundred were successful.

List the total number of successful and failure mission outcomes

```
%%sql

SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE
GROUP BY Mission_Outcome
```

✓ 0.0s

```
* sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- We SELECT Booster_Version column and we want records with maximum payload mass, so we apply a subquery on PAYLOAD_MASS__KG_, the subquery says that PAYLOAD_MASS__KG_ must be equal to the maximum PAYLOAD_MASS__KG_ in the table. Then we group the booster version by name.

List the names of the booster_versions which have carried the maximum

```
%%sql

SELECT Booster_Version FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
GROUP BY Booster_Version
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

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

```
SELECT substr(Date, 6,2) as MONTH, DATE, Booster_Version, Launch_Site, Landing_Outcome
    FROM SPACEXTABLE
WHERE Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015'
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

MONTH	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

2015 Launch Records

- We select MONTH, DATE, Booster_Version, Launch_Site and Landing_Outcome columns to have a full insight on failure drone ship outcomes in 2015. For MONTH and year select we need to use the ‘substr’ commands provided on lab.

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

- Finally, We SELECT Landing_Outcome column and count all records on that column with DATE between 2010-06-04 and 2017-03-20. We group the results by Landing_Outcome result and order them in descending order in count.

```
%%sql

SELECT Landing_Outcome, COUNT(*) as COUNT FROM SPACEXTABLE
WHERE DATE between '2010-06-04' and '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT DESC
```

✓ 0.0s

```
 * sqlite:///my_data1.db
Done.
```

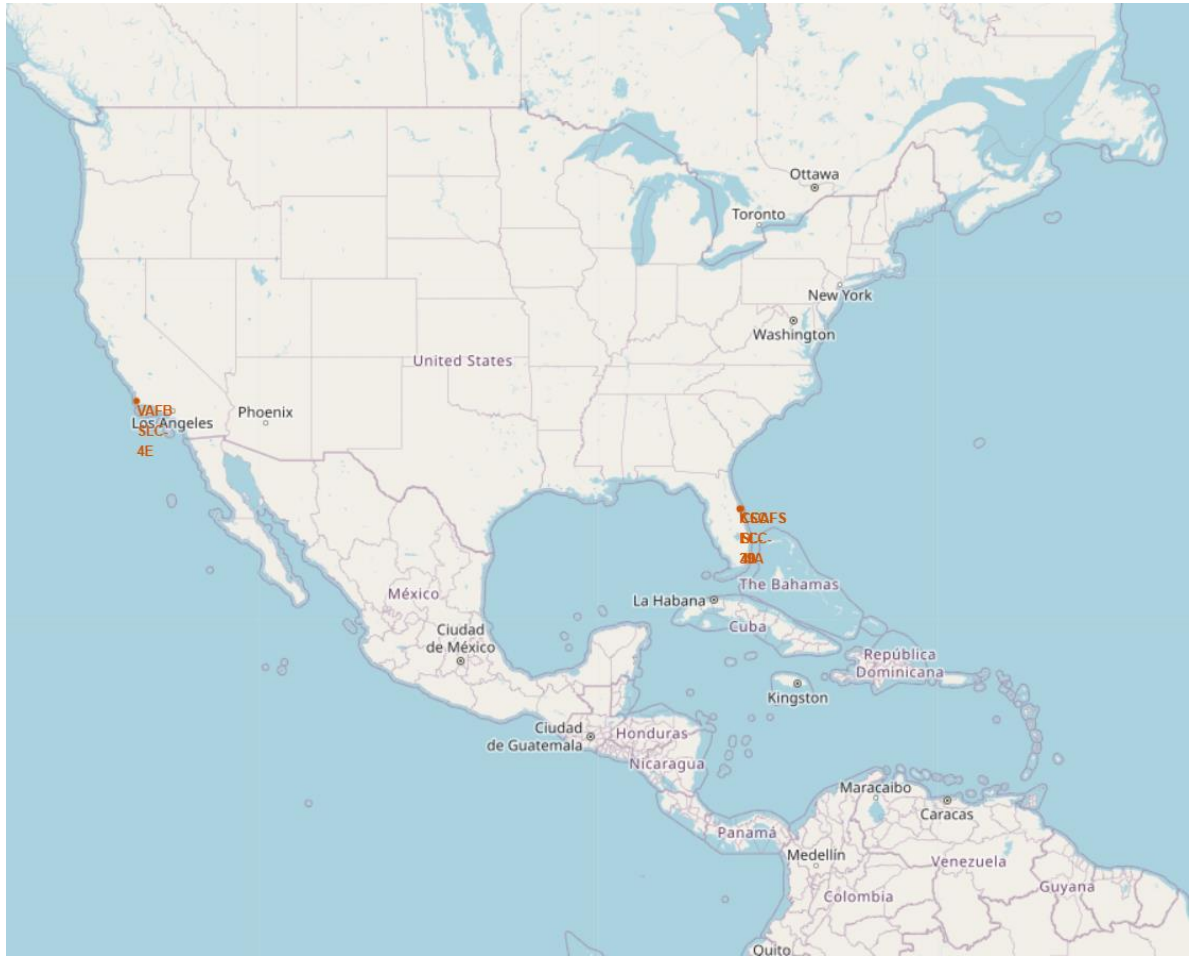
Landing_Outcome	COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

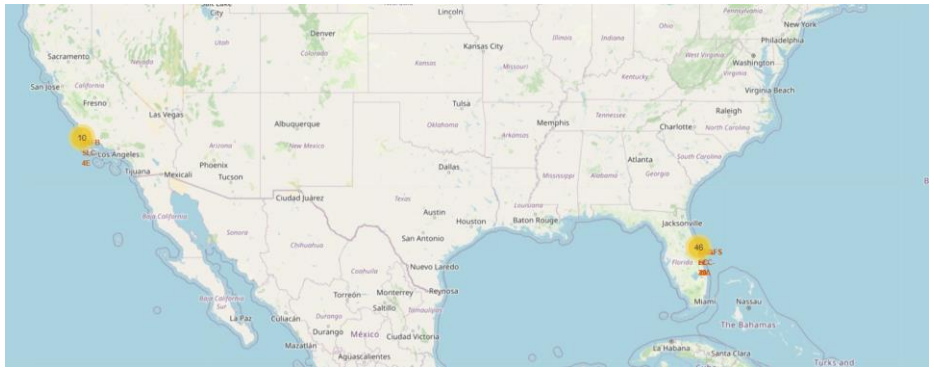
Launch Sites Proximities Analysis

Map of launch site locations on USA

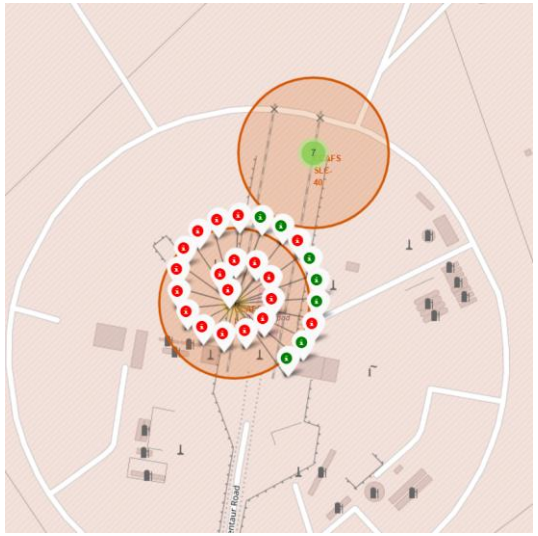
- We have four launch site locations: three located in Florida (east coast) and one located in California (west coast). Those located in Florida are in Cape Canaveral Space Force Station.

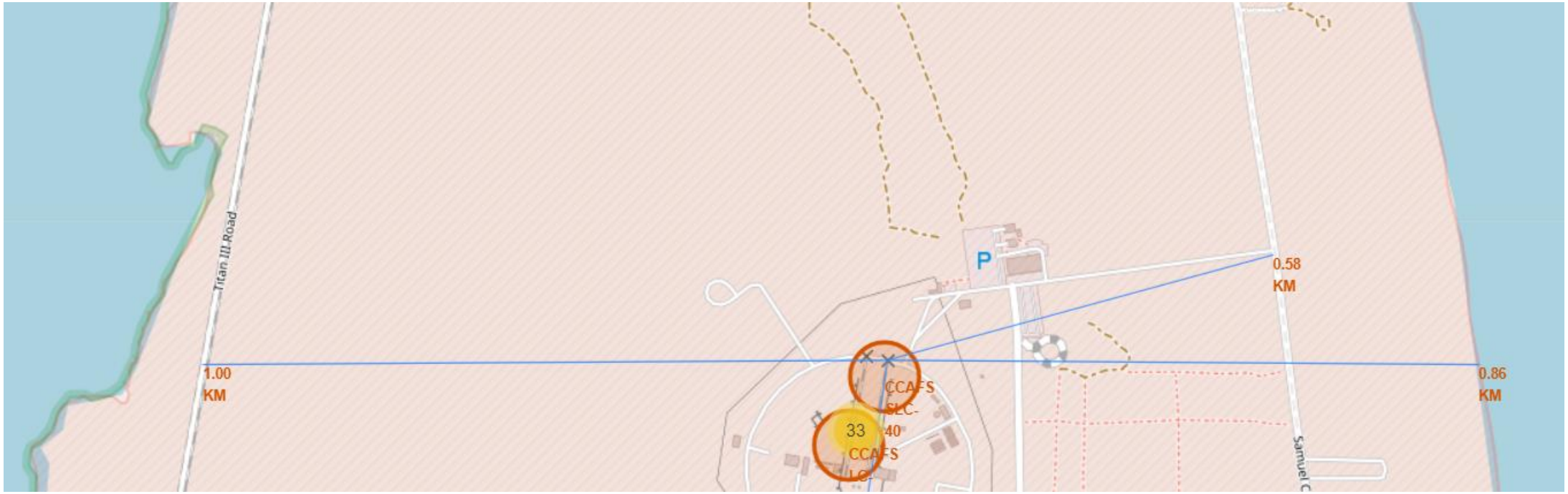


Map of launch records in each location



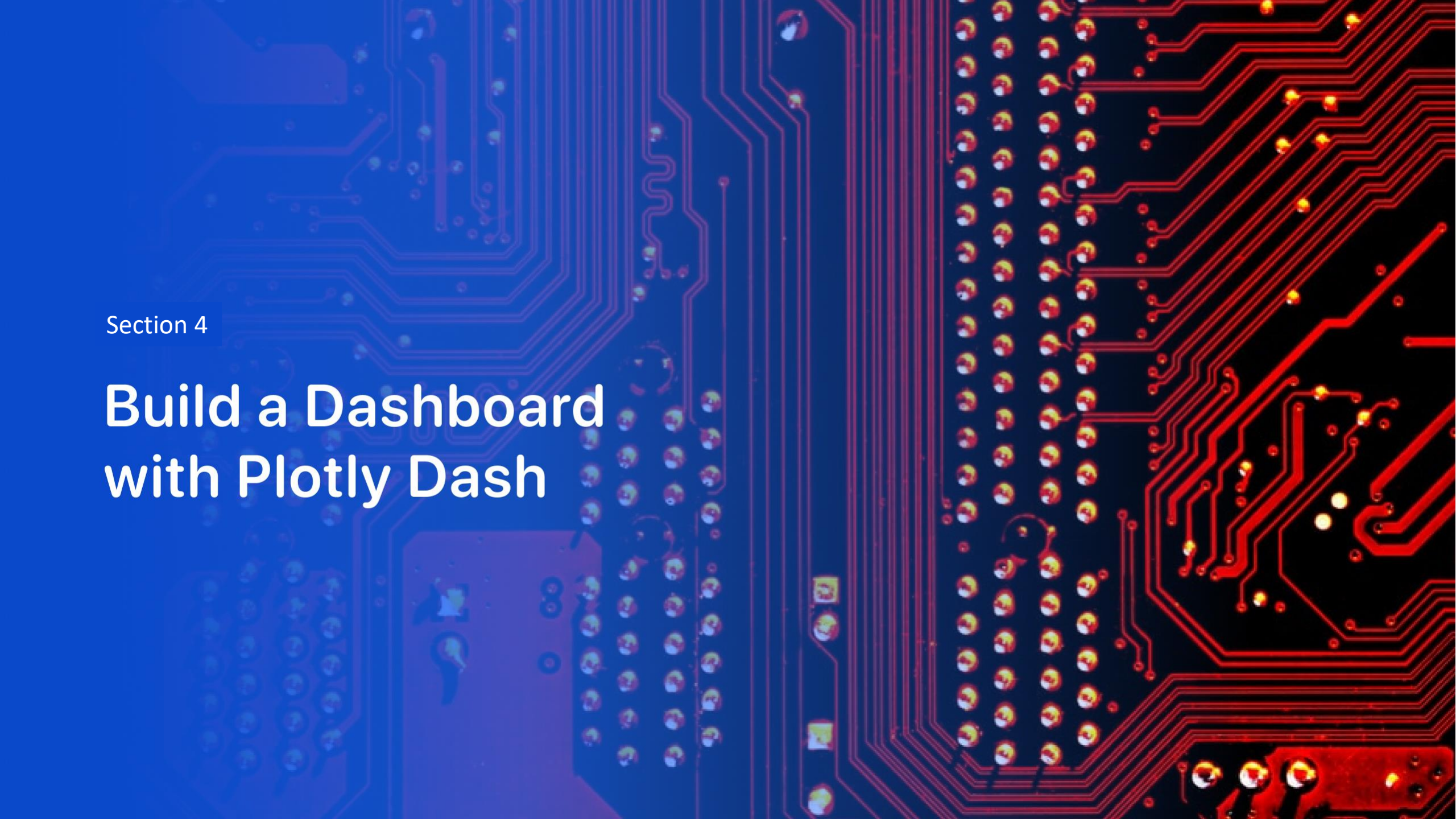
- In this screenshots we can see the number of launch records in each location. There are 10 records on west coast dependency and 46 on east coast. If we zoom in each location we see if a launch had a successful outcome or not.

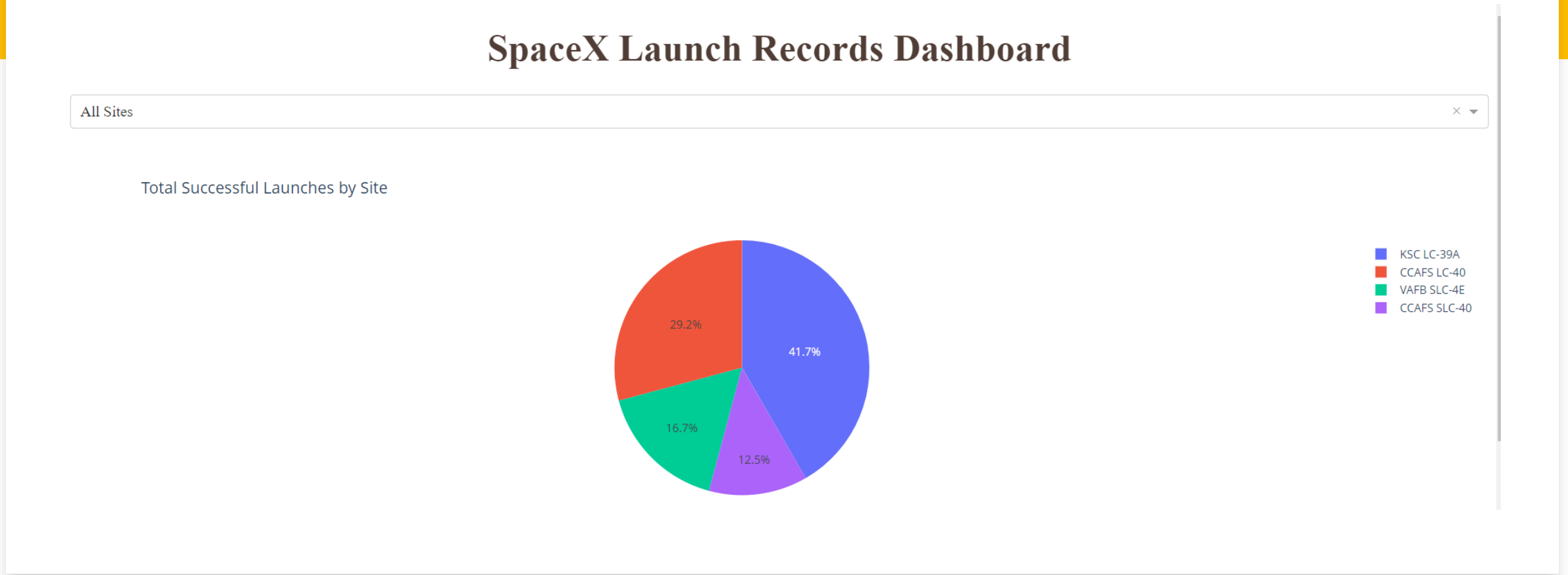




Map of proximities and distances of a launch site

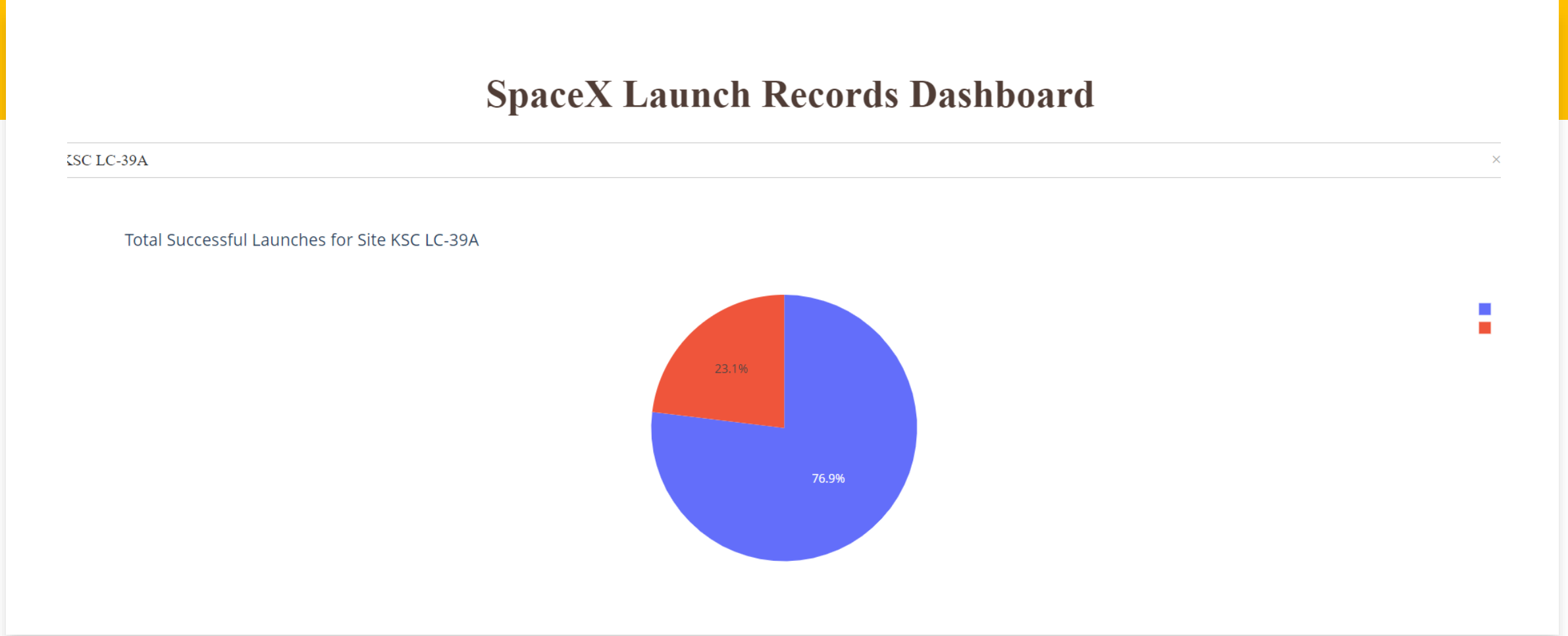
- In this screenshot we observe the distance between CCAFS SLC-40 launch location to its proximities such as line coast (0.86 km) or highway (1 km).


Section 4
Build a Dashboard with Plotly Dash



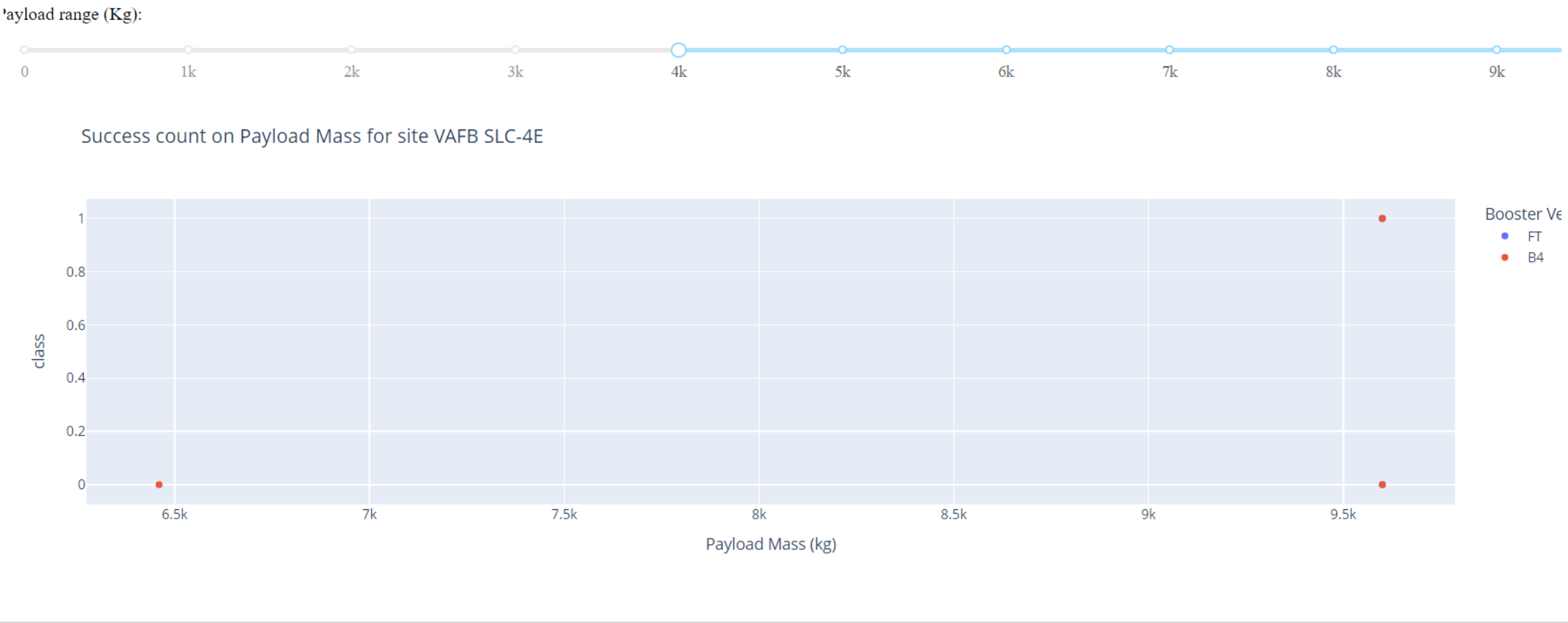
Total Successful Launches by Site

- We observe that KSC LC-39A has the biggest successful launches, followed by CCAFS LC-40 while CCAFS SLC-40 has the lowest successful rate.



Total successful launches for site KSC LC-39A

- We observe that KSC LC-39A has a 76.9% success rate on launches and 23.1% on failure. So, almost 8 out of 10 launches will be successful.

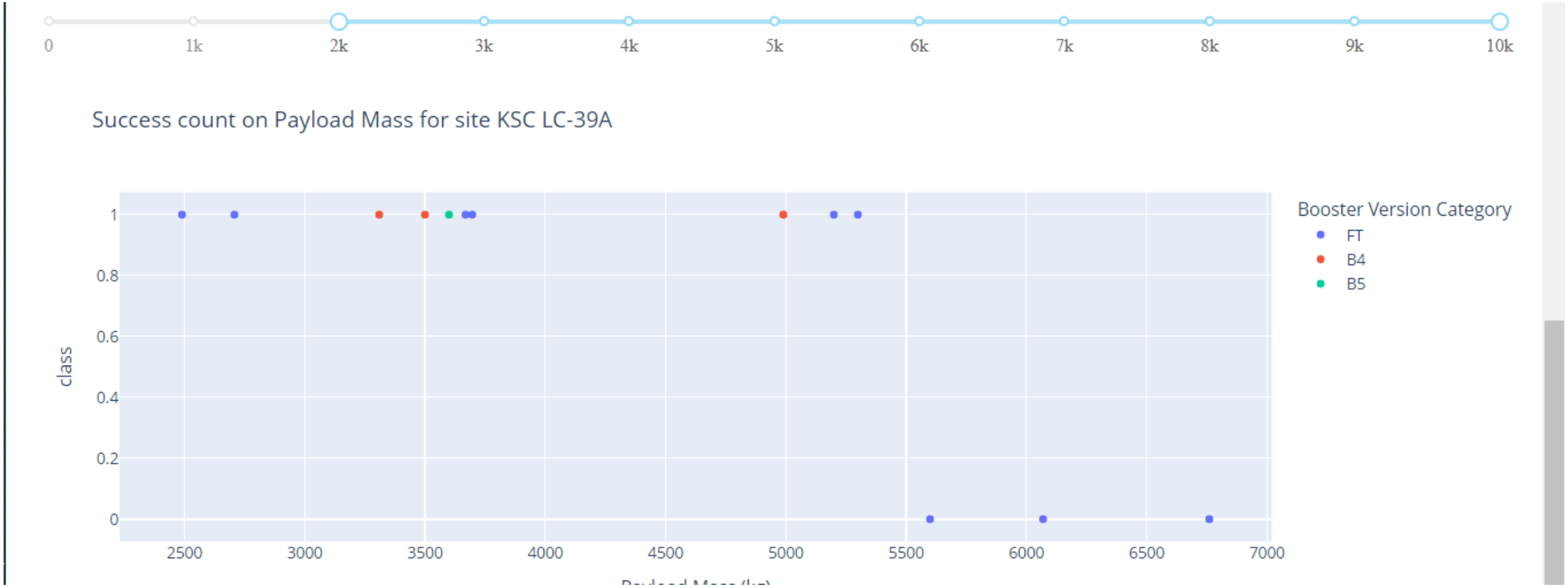


Payload vs Launch Outcome scatter plot VAFB SLC-4E

- We observe that only B4 Booster Category launches were performed with a payload mass greater than 4000 in VAFB SLC-4E site. Only one was successful.

Payload vs Launch Outcome scatter plot CCAFS LC-40

- We observe that v1.1 and FT Booster Category launches were performed with a payload mass greater than 1000 in CCAFS LC-40 site. FT one has a bigger successful rate than v1.1.



Payload vs Launch Outcome scatter plot KSC LC-39A

- We observe that FT, B4 and B5 booster category launches were performed ins KSC LC-39A launch site. Light payload mass (less than 5500) has a 100% success rate.

Payload vs Launch Outcome scatter plot CCAFS SLC-40

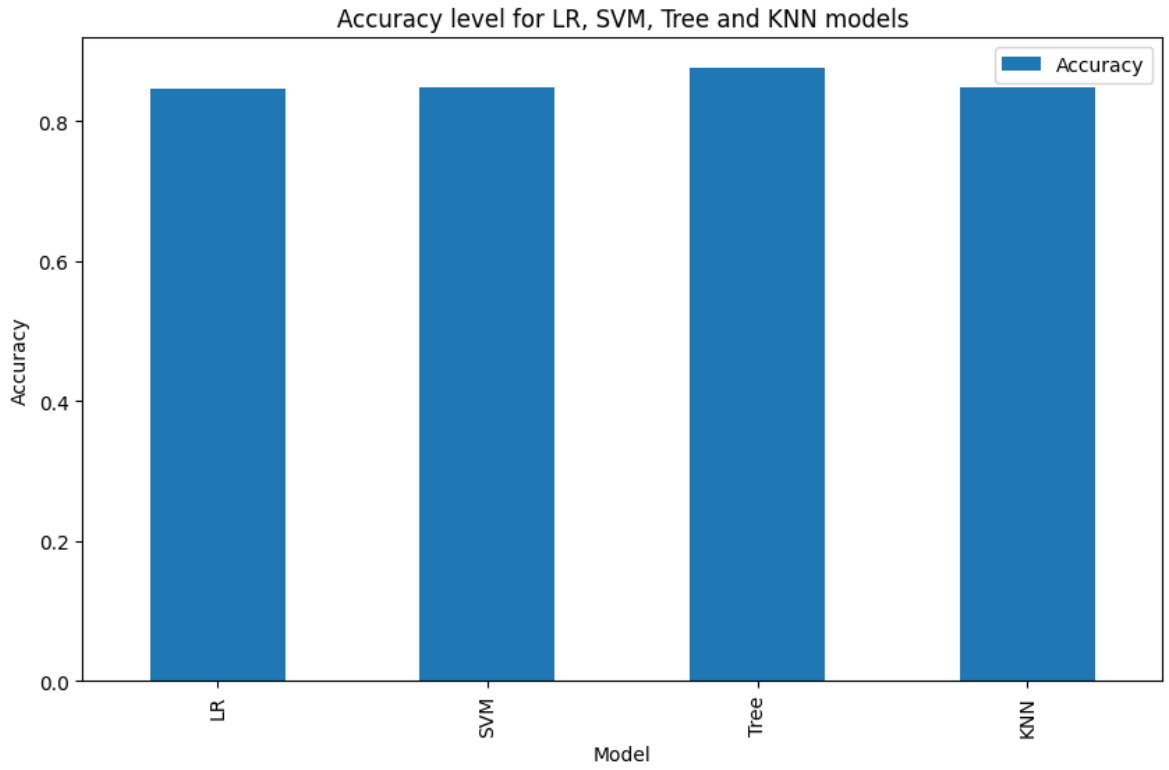
- We observe that B4 and FT Booster Category launches were performed with a payload mass greater than 2000 in CCAFS SLC-40 site. B4 one has the biggest failure rate (three out of four launches failed, a 75%).

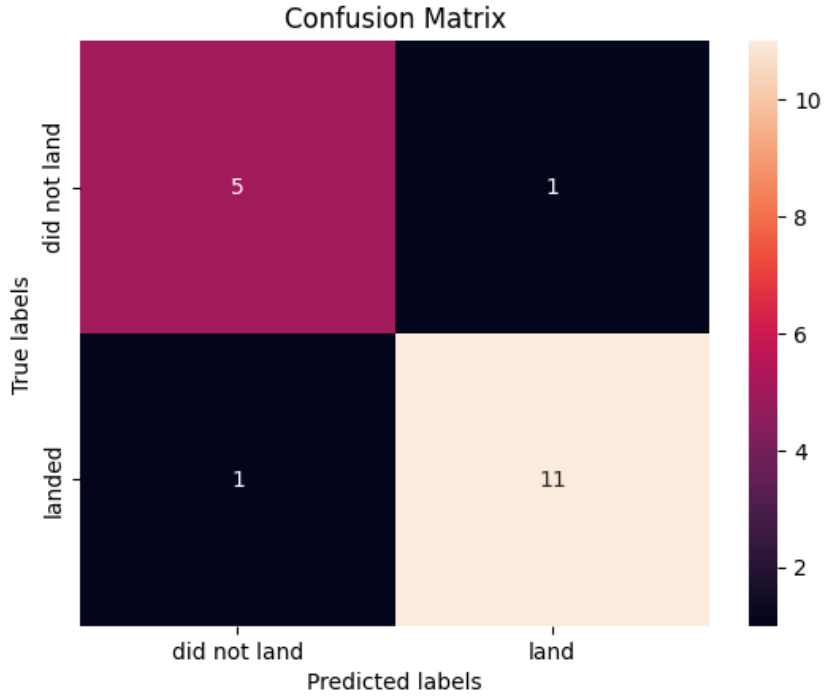
Section 5

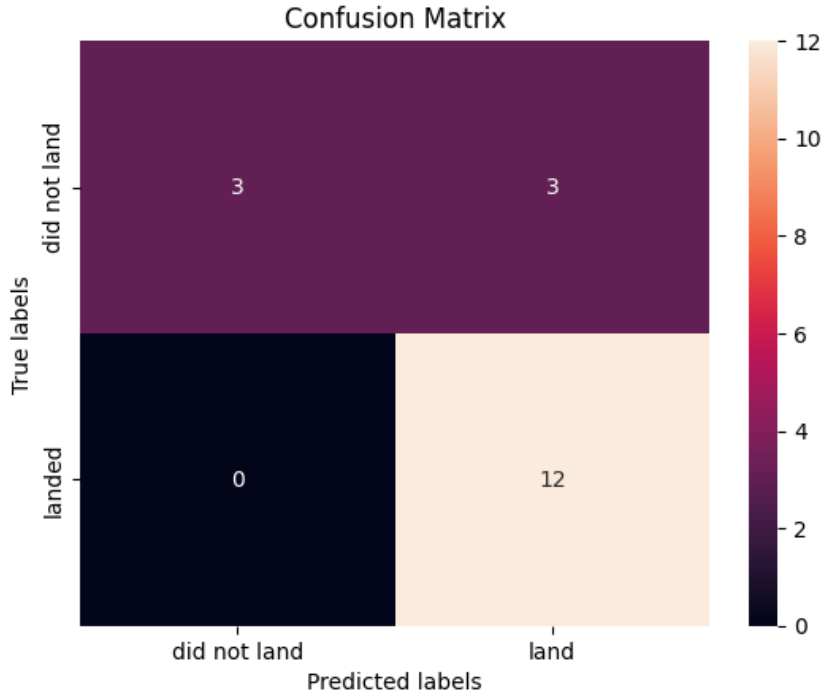
Predictive Analysis (Classification)

Classification Accuracy

All methods perform in a similar way but the one with the best accuracy is **decision tree**.







Confusion Matrix

- Left matrix is the confusion matrix of decision tree model, the one with best accuracy. This matrix has 16 true positives (11 on land label and 5 on did not land label) and 2 false positives. If we compare this matrix with the other three model matrix (right matrix), we observe this is the one with the biggest number of true positives. That's why this method hast the best accuracy.

Conclusions

- Payload Mass and Launch Site are the two variables which most affect a launch cost.
- KSC LC-39A launch site has a 76.9% success rate, the biggest one.
- Light payload mass (less than 5500 kg) has a 100% success rate in KSC LC-39A launch site.
- ES-L1, GEO, HEO and SSO orbits have the biggest success rate.
- The best model to predict if a land will be successful or not is decision tree model.

Appendix

If you want to know more about the code used to get the results, we showed in his presentation, please refer to GitHub repository to visualize the notebooks we created.

Link:

Thank you!