IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

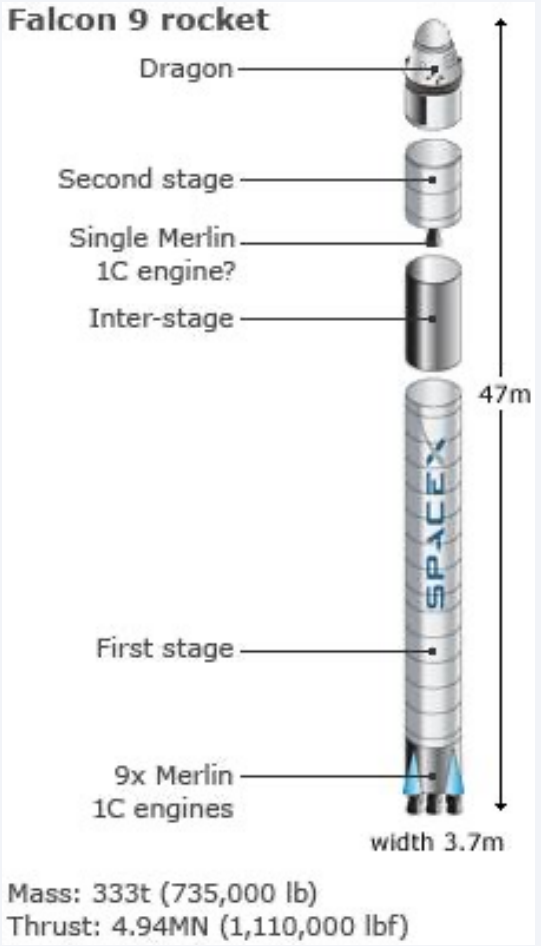
Beatrice Costanza Marrano
28th July 2023

Outline

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

Executive Summary

- **Summary of project:** In this project, we will take the role of a data scientist working for the new rocket company SpaceY. More in details, we will predict if the «Falcon 9» first stage will land successfully and consequently the price of each launch. In fact SpaceX advertises «Falcon 9» rocket launches on its website with a cost of 62 million dollars, while other providers cost upward of 165 million dollars. This big difference is due to the fact that SpaceX can reuse the first stage if it lands successfully.
- **Summary of methodologies:** To accomplish this we will gather data about SpaceX and create some visualization and dashboards for our team. Data will be gathered in 2 ways: with the SpaceX REST API call and with Web Scraping of some Wiki Tables.
- **Summary of all results:** The insights we found with this analysis is that the different launch sites have different success rates, and different maximum payload masses they carry; there are 4 main orbits that have a very high success rate (0.9-1.0), while all the others are around 0.6; only some orbtits' success rate is linked to the Flight number; the success rate has kept increasing from 2013 to 2020. Furthermore, the application of the machine learning classification models, showed that the best model to predict if the first stage will land successfully is the decision tree classification model.



Introduction

- Project background and context:
  - The company SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers cost upward of 165 million dollars. This big difference is due to the fact that SpaceX can reuse the first stage if it lands successfully. In this project we will try to predict if the Falcon 9 first stage will land successfully in order to know if the cost of a launch will be reduced thanks to its reuse.

- Project questions:
  - Will the first stage land successfully or not?
  - What will then be the price of a launch?

Section 1

Methodology

Methodology

Executive Summary

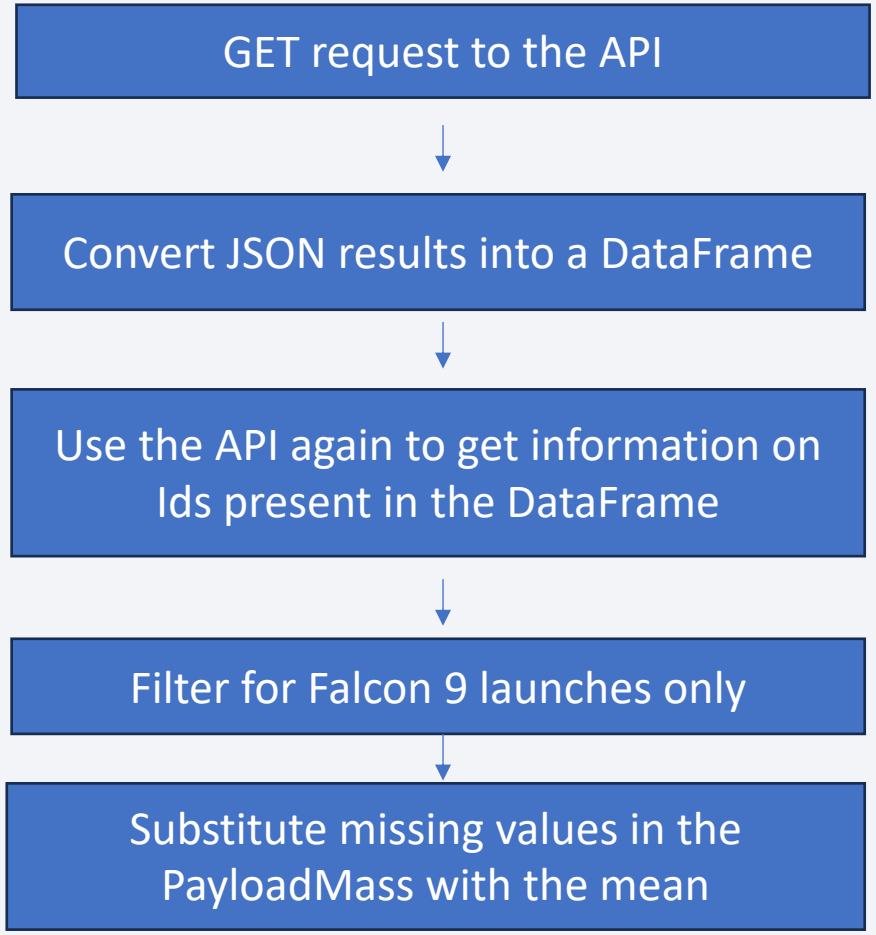
- Data collection methodology:
  - SpaceX REST API, Web Scraping
- Perform data wrangling
  - Create datasets for analysis, filter, clean data and deal with Nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluating different classification models

Data Collection

In order to get all information on different Launches the **Data Collection** process will consist of 2 phases:

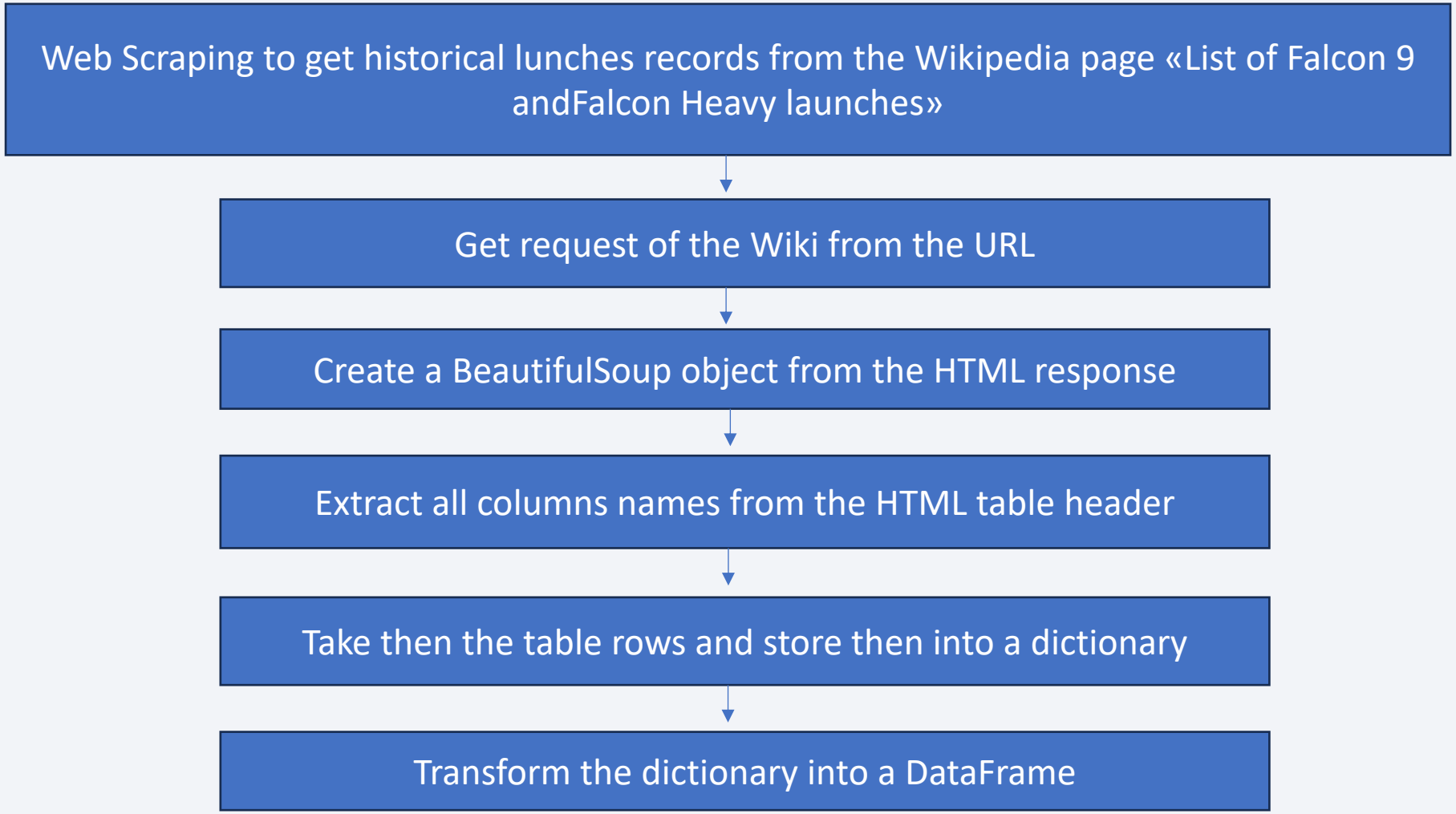
- Gather data with SpaceX REST API
- Gather data with Web Scraping using BeautifulSoup from Wiki Tables

Data Collection – SpaceX API



Data source: https://github.com/beatricemarrano/spacey/blob/main/Data%20Collection%20API.ipynb

Data Collection – Scraping



Data source: https://github.com/beatricemarrano/spacey/blob/main/Data%20Collection%20Web%20Scraping.ipynb

Data Wrangling

Perform some exploratory data analysis (EDA) to lfind some patterns in the data and determine the label or training the classification models

Make some analysis of the number and typesof landing outcomes

Create a binary Class label indicating 1 for a successful landing and 0 otherwise

Data source: https://github.com/beatricemarrano/spacey/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The main type of chart that has been used as been the Categorical Scatterplot (sns.catplot) that helped us understanding the relation between a numerical variable and one or more categorical variables like Orbit and Launch Site. This type of plot allows to color the dots with a categorical variable like the binary Class variable, allowing us to understand also how the relationship between the 2 variables are also linked with the final outcome (1/0 for successful or unsuccessful landing).

Data source: https://github.com/beatricemarrano/spacey/blob/main/EDA%20with%20Visualizations.ipynb

EDA with SQL

- Find the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Find the total payload mass carried by boosters launched by NASA (CRS)
- Find the average payload mass carried by booster version F9 v1.1
- Find the data of the first succesful landing outcome in ground pad
- Boosters names which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Find the total number of successful and failure mission outcomes for every type of outcome
- Booster_versions which have carried the maximum payload mass
- Display information on month names, failure landing_outcomes in drone ship ,booster versions and launch_site for the months in year 2015.
- Order (desc) the count of landing outcomes happened between the date 2010-06-04 and 2017-03-20

Data source: https://github.com/beatricemarrano/spacey/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We first created a map centered at the NASA Johnson Space Center; then we added a circle at the Launch sites coordinates to detect where the 4 launch sites were located; then more specifically we wanted to mark on the map the success/failed launches using respectively a green or red marker; finally we computed the distances between a launch site and the closes coastline, the closest city and highway or railway if present. To do this we used instead a line that connected the launch site with the selected place (city, coast etc) using the mouse position tool to get the coordinate of the place pointed by the mouse.
- Adding these object makes the analysis more straight forward and the insights more easy to find by looking at distances and places visually instead of looking at numeric values of distances and coordinates.

Data source: https://github.com/beatricemarrano/spacey/blob/main/Interactive%20visualizations.ipynb

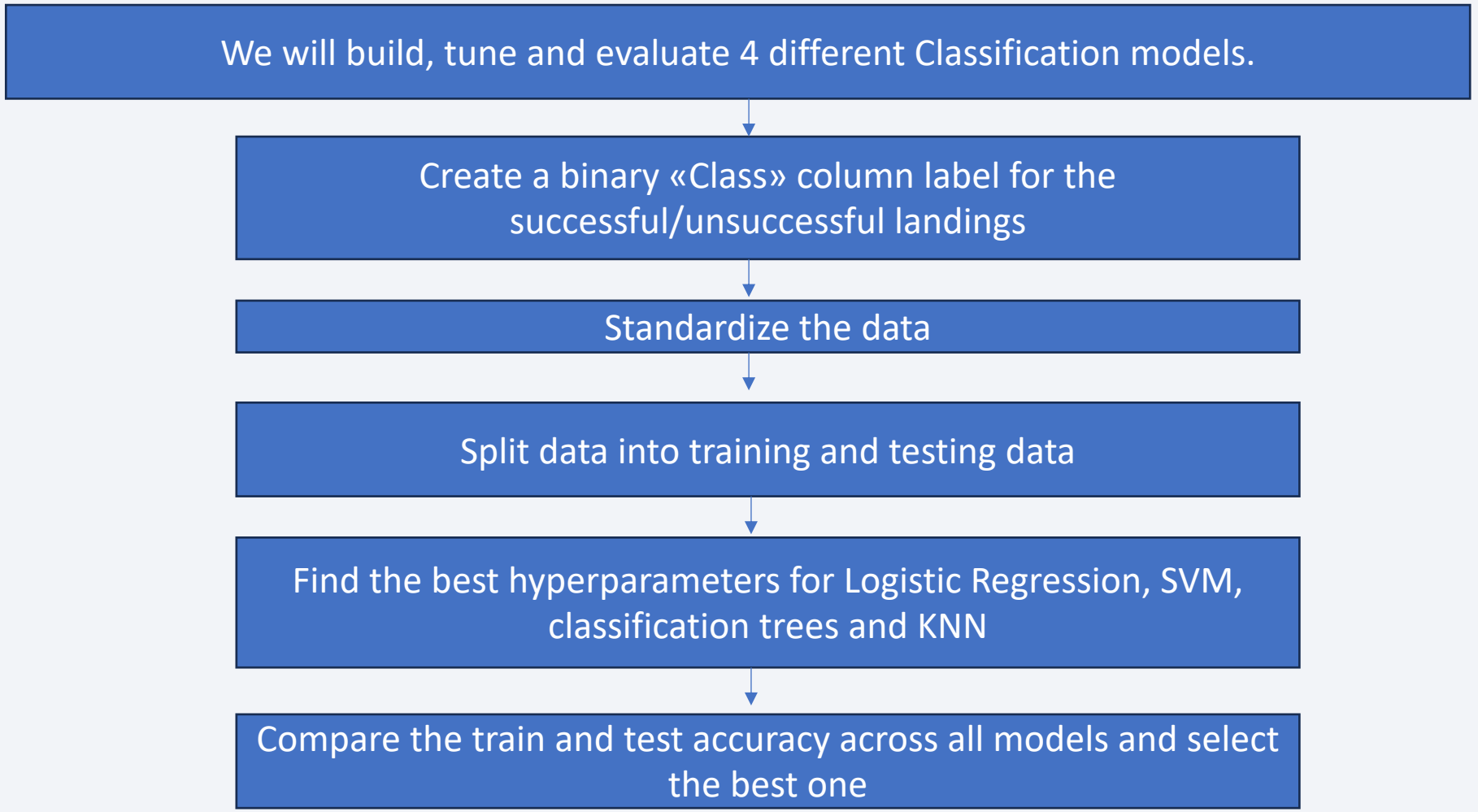
Build a Dashboard with Plotly Dash

- The Dashboard created with Plotly Dash has been structured in two parts. The top part of it consists of a dropdown menu that allows the selection of a specific launch site or of all the sites; then if all sites are selected the pie chart will show the % of total success rate distributed across the different sites, allowing for a clear understanding of the sites with the highest and lowest success rates. In instead a specific site is selected the pie chart will be divided in 2 showing the % of success and failed launches (0/1). The lower part of the dashboard shows a range filter selection of the Payload mass that ranges from 0 to 10000 kg and based on the selected range and also based on the choice of the dropdown menu filter on the Launch site, there will be a scatterplot showing the payload mass vs the Class outcome and the dots have been colored by the Booster version type to allow the user to get more information on this relationships.

- These plots and interactions allow the users of the dashboard to go deep into discovering insights for all the combinations of measures and categorical variables such as the launch site dropdown selection and the payload mass range filter selection. An interactive dashboard instead of a static visualization or plot allows the user to discover personally what the data scientist has foud out in an intuitive and user friendly way.

Data source: https://github.com/beatricemarrano/spacey/blob/main/dash_interactivity.py

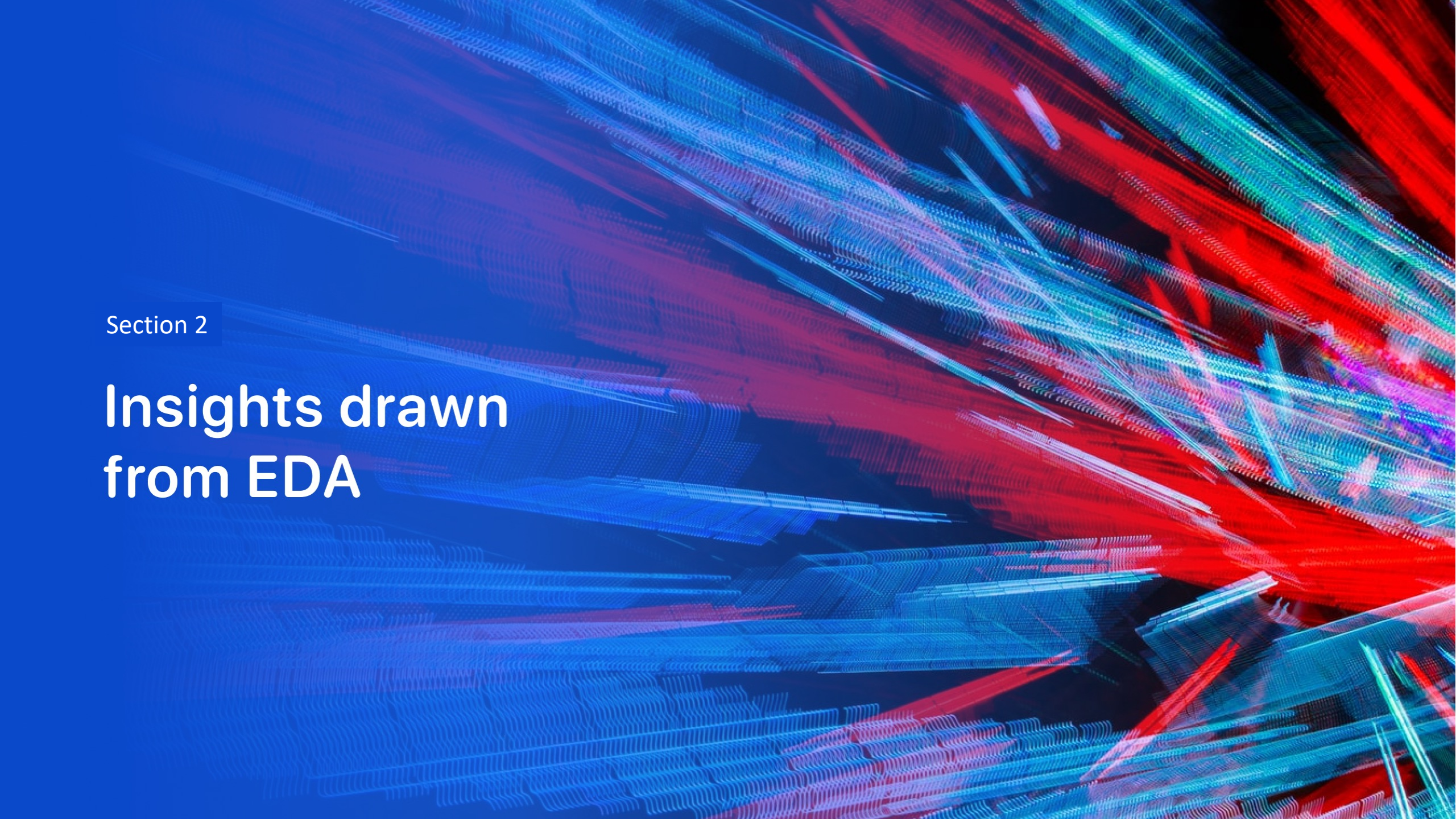
Predictive Analysis (Classification)



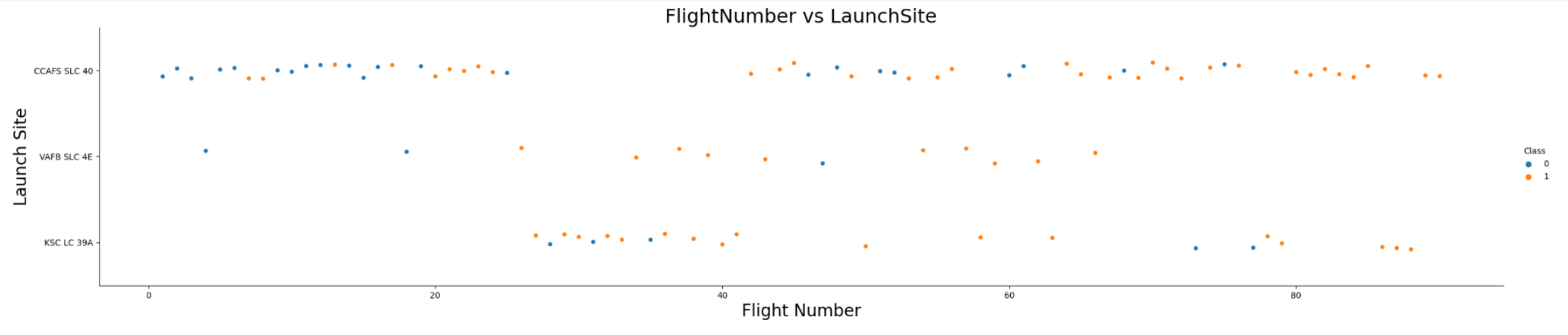
Data source: https://github.com/beatricemarrano/spacey/blob/main/Machine%20Learning%20Predictions.ipynb

Results

- Exploratory data analysis results: The exploratory data analysis gave us many information related to every feature and also on the inter features relationships. Among the main finding we have:
  - There are 4 different launch sites
  - For the VAFB SLC 4E there are no rockets lunched for payload mass higher than 10000
  - There are 4 Orbits with a very high average success rate
  - Only for some orbits the success appears to be related to the flight number
  - Overall the success rate has always grown from 2013 to 2020.
- Interactive analysis: The interactive analysis showed geographical patterns linked to each launch site and their success rate in a visual and intuitive way.
- Predictive analysis results: The classification analysis showed that the best model is the Decision Tree Classifier with a train accuracy of 88% and a test accuracy of 94%.

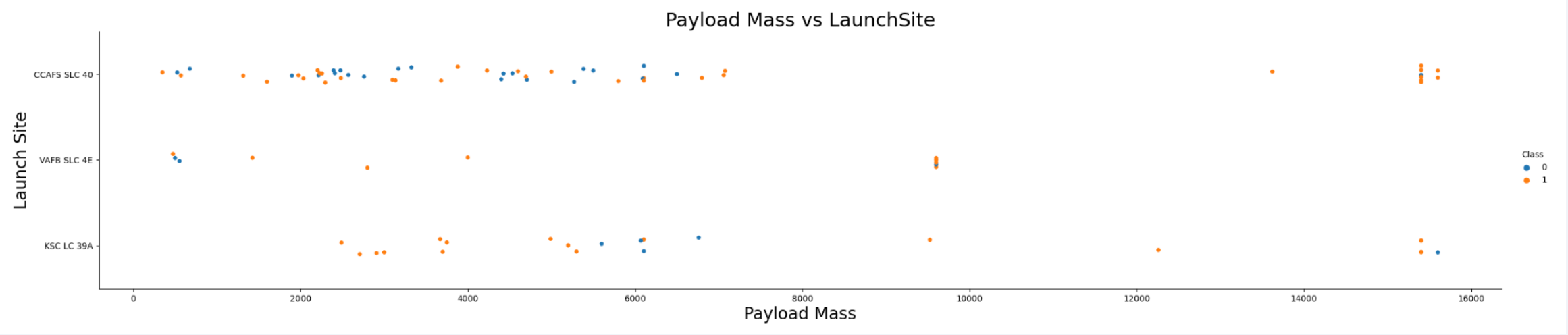

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- We can notice how the VABF SLC 4E site has maximum flight number around 67 will the others have higher flight numbers and almost all successful landings but fewer compared to the other sites and that the higher the flight number the more successful the landing is.

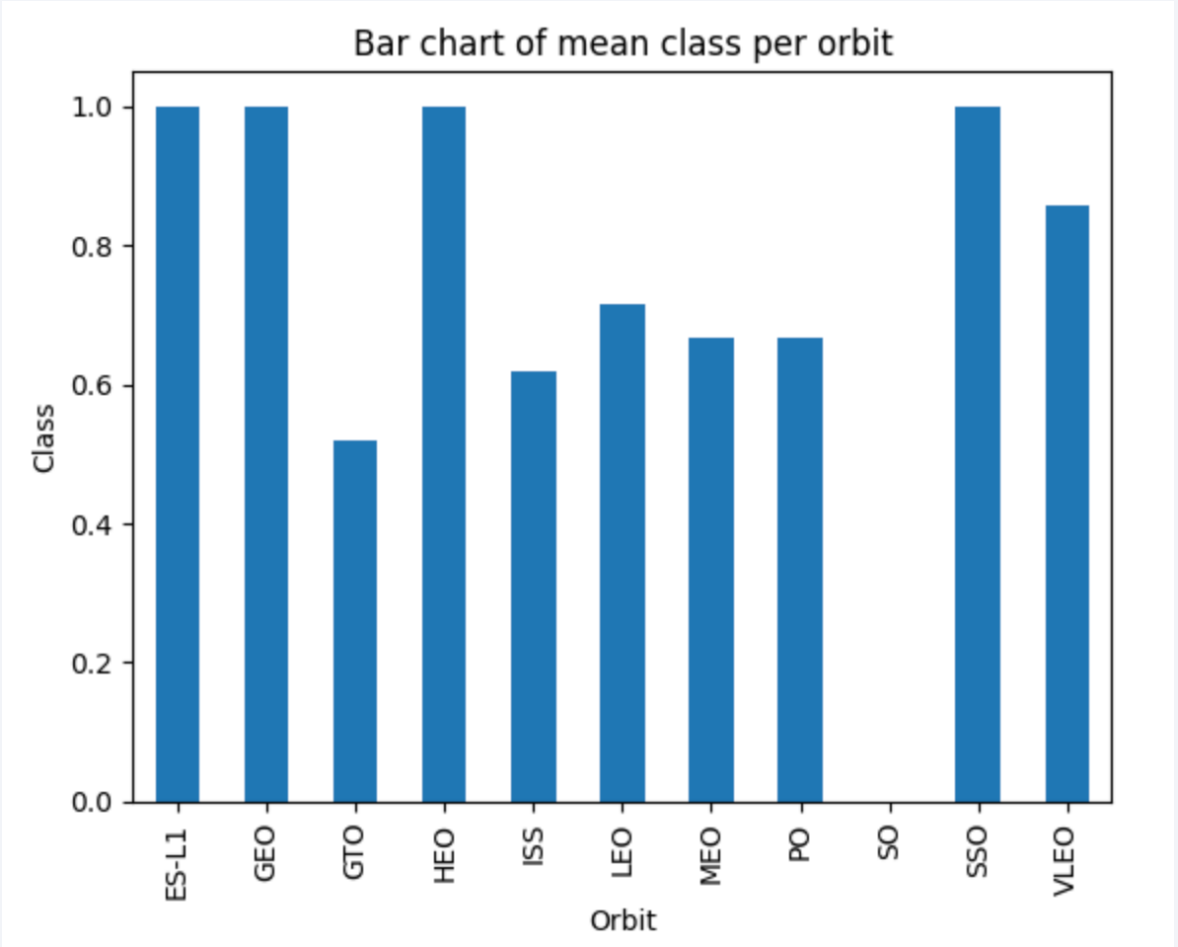
Payload vs. Launch Site



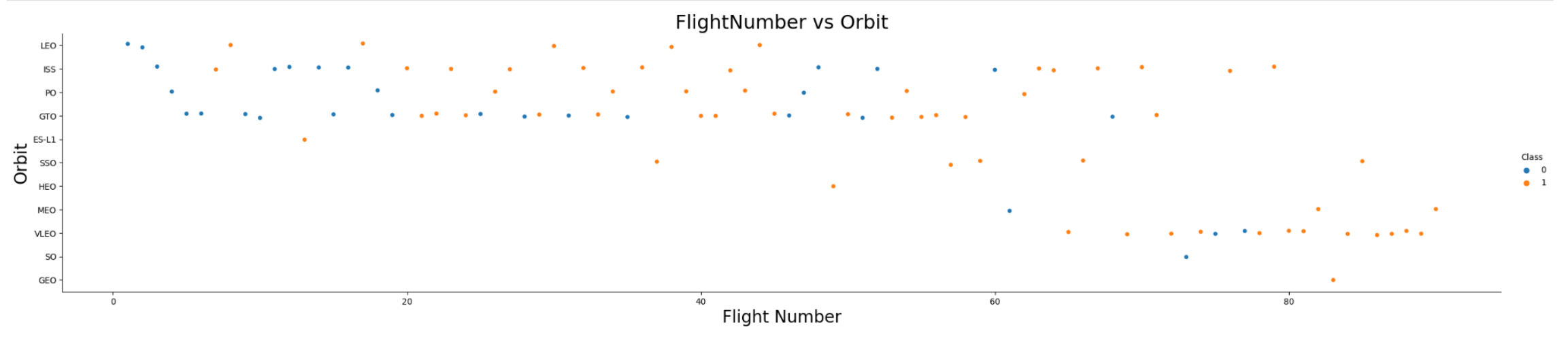
- We can notice that for the VAFB SLC 4E there are no rockets lunched for payload mass higher than 10000.

Success Rate vs. Orbit Type

- We can notice how there are 4 Orbits with a very high average success rate (between 0.9 and 1), while all the others are around 0.6 and the SO is at 0.

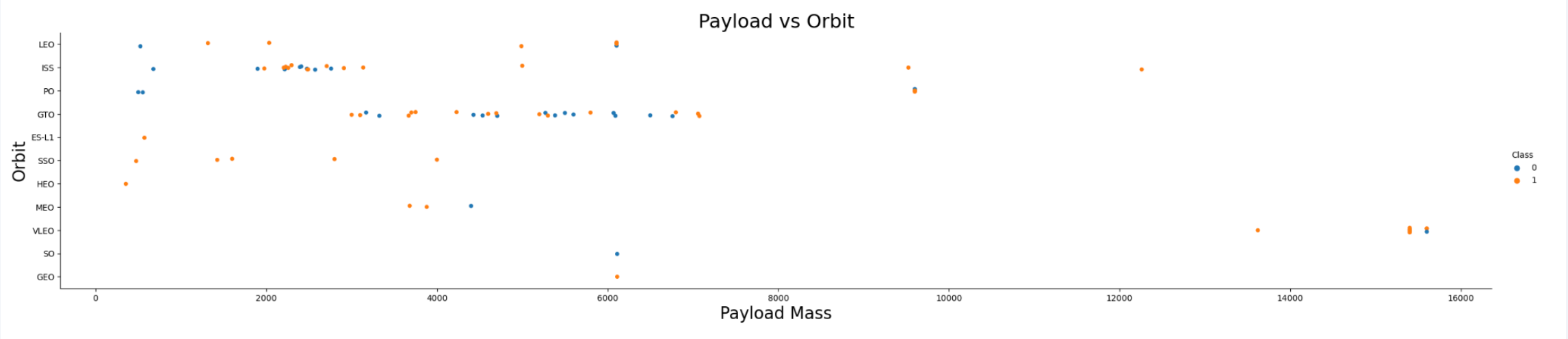


Flight Number vs. Orbit Type



- We can notice that in the LEO orbit the success appears to be related to the flight number but there seems to be no relationship between flight number and GTO orbit.

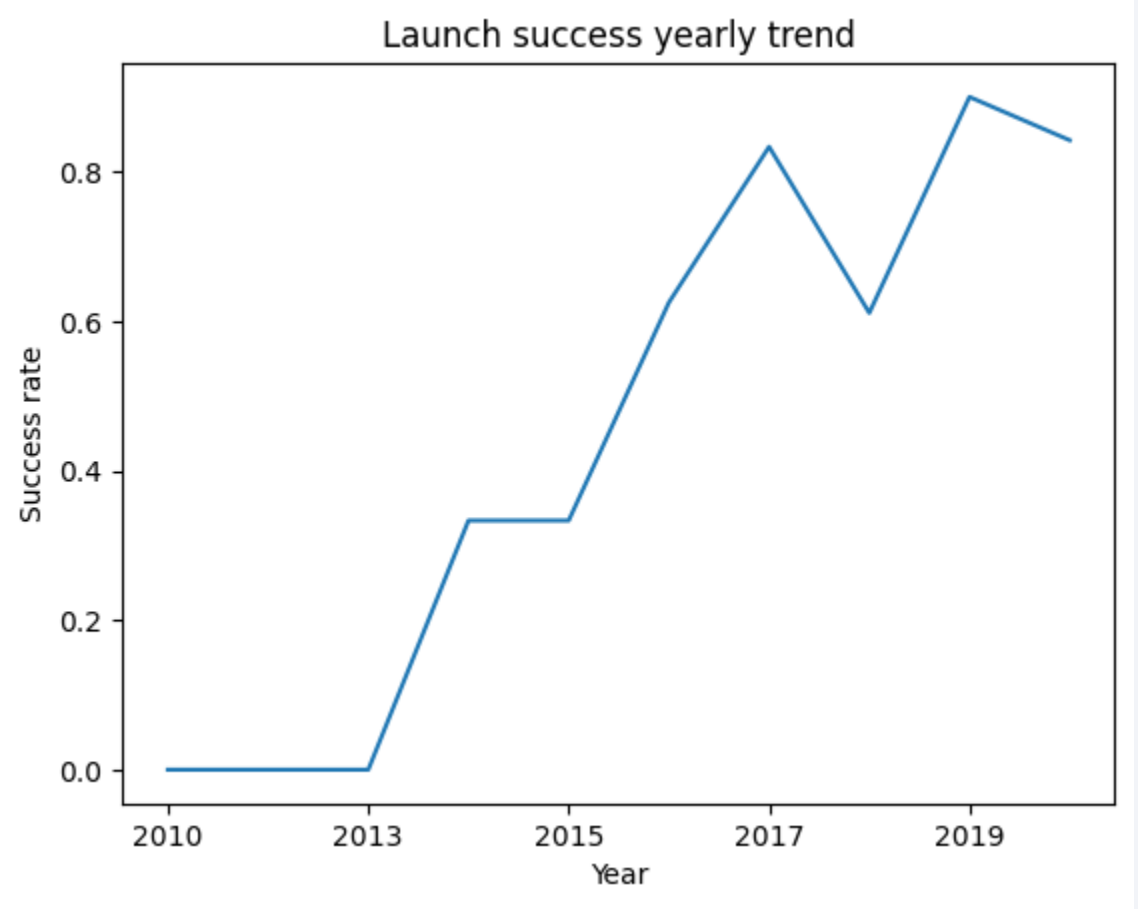
Payload vs. Orbit Type



- We can notice how for a high payload mass, the successful landing are more for Polar, LEO and ISS.

Launch Success Yearly Trend

- We can see that the success rate has always grown from 2013 to 2020, it just had a short decreasing period in 2017-2018.



All Launch Site Names

- Find the names of the unique launch sites
- We can find the names of the launch sites by selecting the distinct values of the Launch_Site column.

```
%sql SELECT distinct Launch_Site FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- We can find this by filtering for Launch Site like “CCA%” and limiting the output rows to 5 using the word LIMIT

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site lIke 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- We can find it by summing the payload mass WHERE the Customer is NASA (CRS)

```
%sql SELECT sum(PAYLOAD_MASS__KG_) as Tot_playload_mass FROM SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

Tot_playload_mass
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- We can find it by averaging the payload mass WHERE the booster version is F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) as Avg_payload_mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

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

Avg_payload_mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- We find it by selecting the MIN date where the Landing outcome is successful on ground pad

```
%sql SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- We can do it by selecting the distinct Booster versions where the landing outcome is successful on drone ship and the payload mass is between 4000 and 6000

```
%sql SELECT distinct Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- We can find it be selecting the distinct mission outcomes and counting how many we have for each category by grouping by the outcomes.

```
]: %sql SELECT distinct Mission_Outcome, count(*) as Tot FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- We can find it by selecting the distinct boosters where the payload mass is equal to the maximum payload among all recordings in the table. This is accomplished using a subquery to find the max(payload).

```
%sql SELECT distinct Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FR
```

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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We can find it by selecting all the columns we want to display and using the WHERE condition on the Date selecting only the 2015.

```
%sql SELECT substr(Date, 4, 2) as month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE sub
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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
- We can find it by selecting the outcomes and counting them by grouping by outcomes and ordering by descending counts and using the where condition on the time period requested.

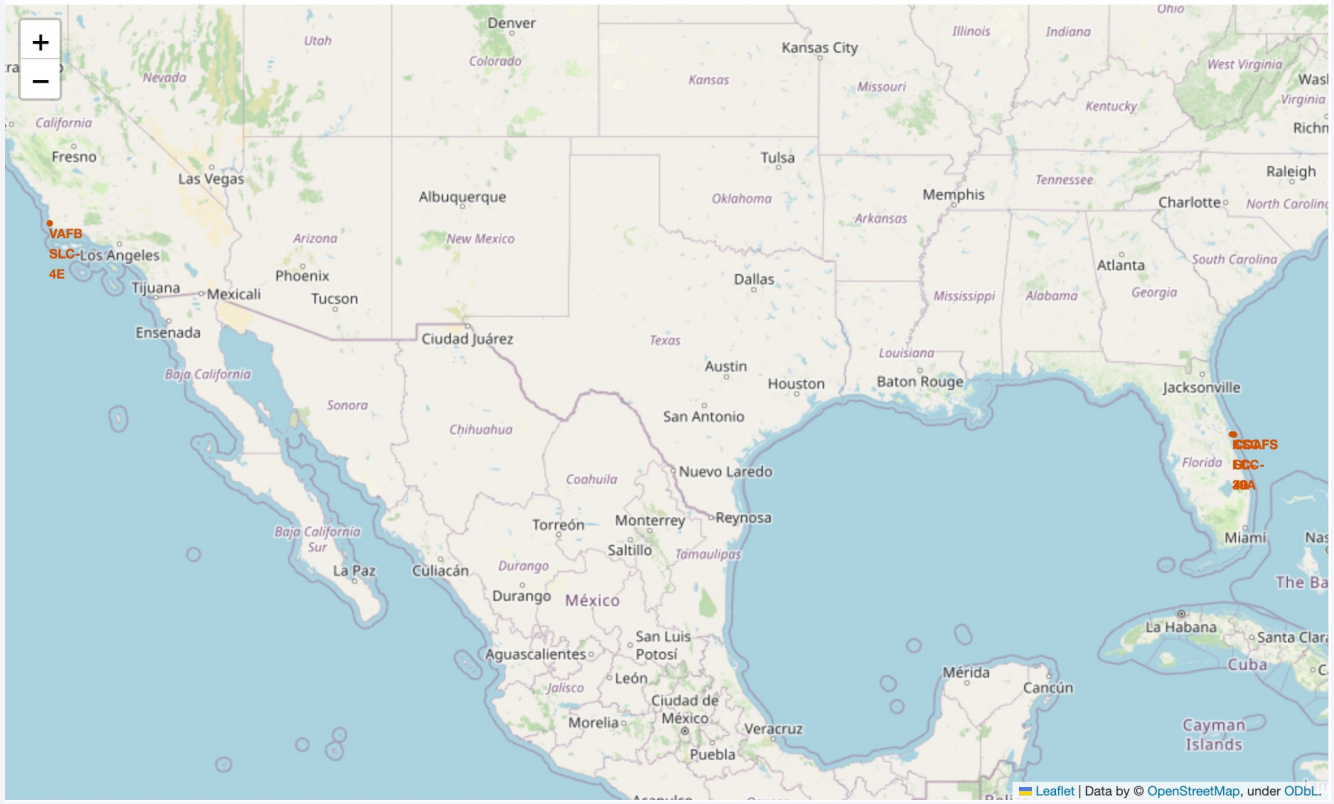
```
%sql SELECT Date, Landing_Outcome, count(Landing_Outcome) FROM SPACEXTABLE WHERE Date>= '04/06/2010' and Date <='
```

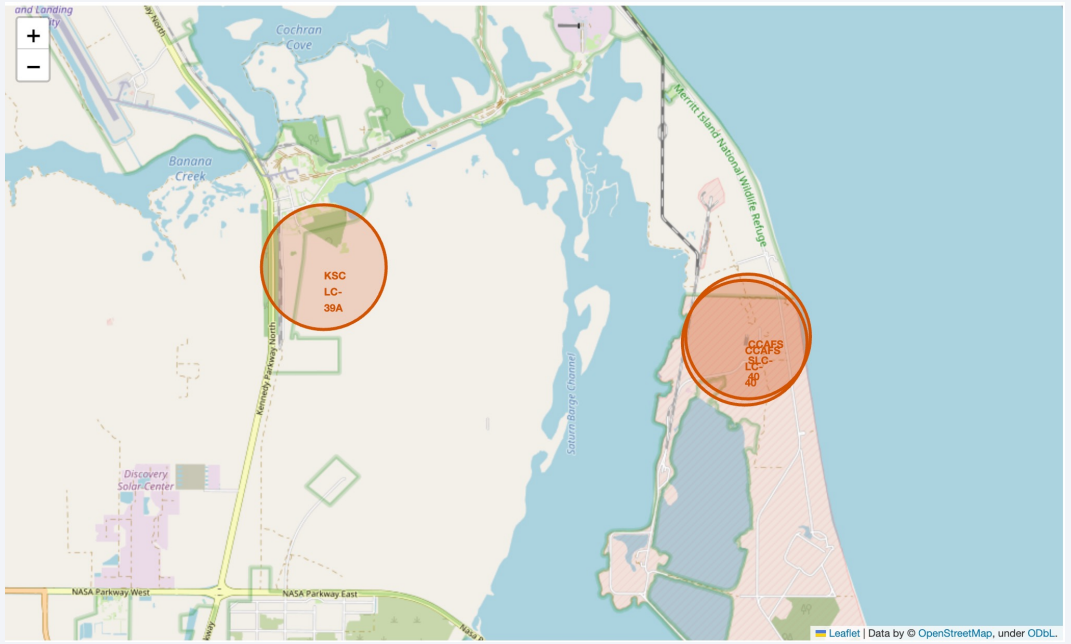
```
 * sqlite:///my_data1.db
Done.
```

Date	Landing_Outcome	count(Landing_Outcome)
08/07/2018	Success	20
10/08/2012	No attempt	9
04/08/2016	Success (drone ship)	8
18/07/2016	Success (ground pad)	7
14/04/2015	Failure (drone ship)	3
12/05/2018	Failure	3
06/04/2010	Failure (parachute)	2
18/04/2014	Controlled (ocean)	2
08/06/2019	No attempt	1


Section 3
Launch Sites
Proximities Analysis

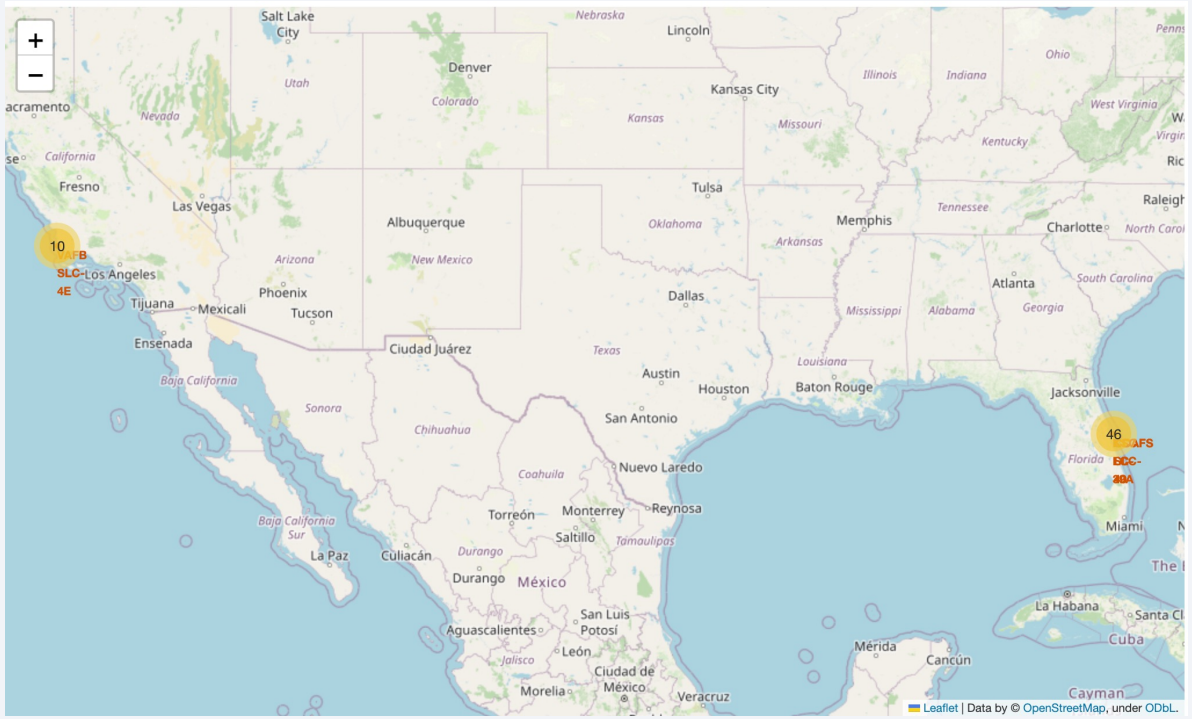
Mark all launch sites on the map

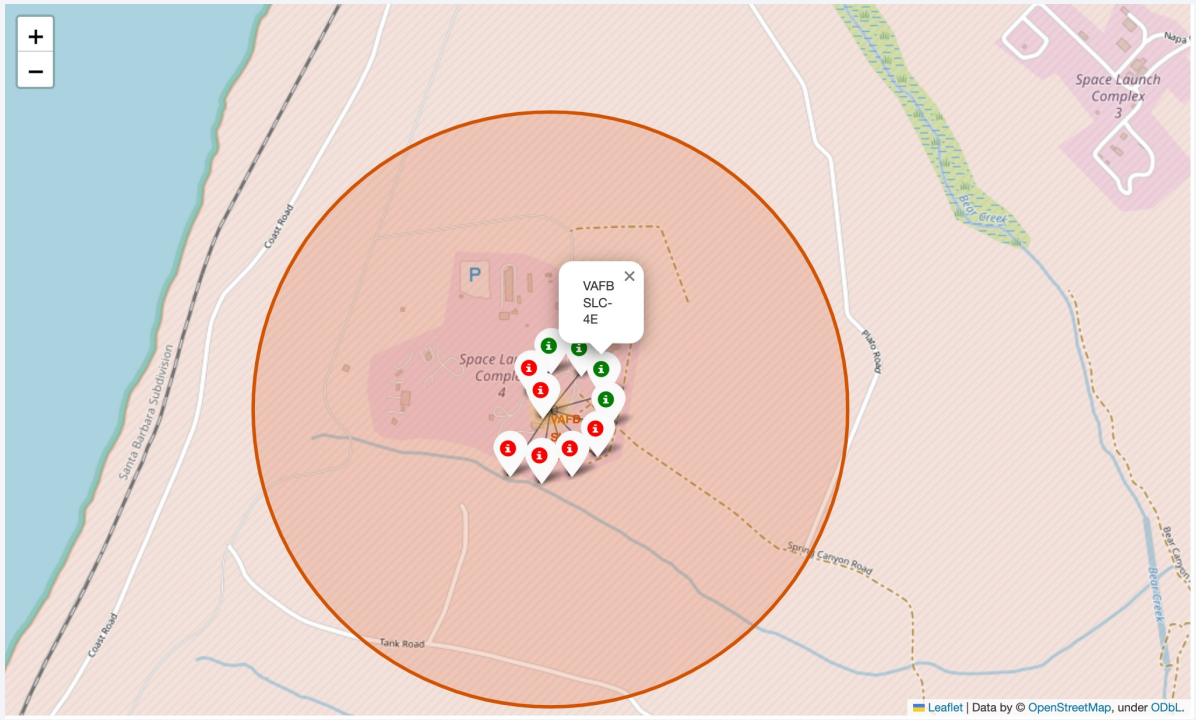




- The majority of the launch sites are in Florida quite close to the Equator line. There is one site that instead is in California, close to Los Angeles, slightly far away from the Equator line and on the western coast.
- All 4 launch sites are on the coast, 3 on the eastern coast and very close to each other and one on the western coast.

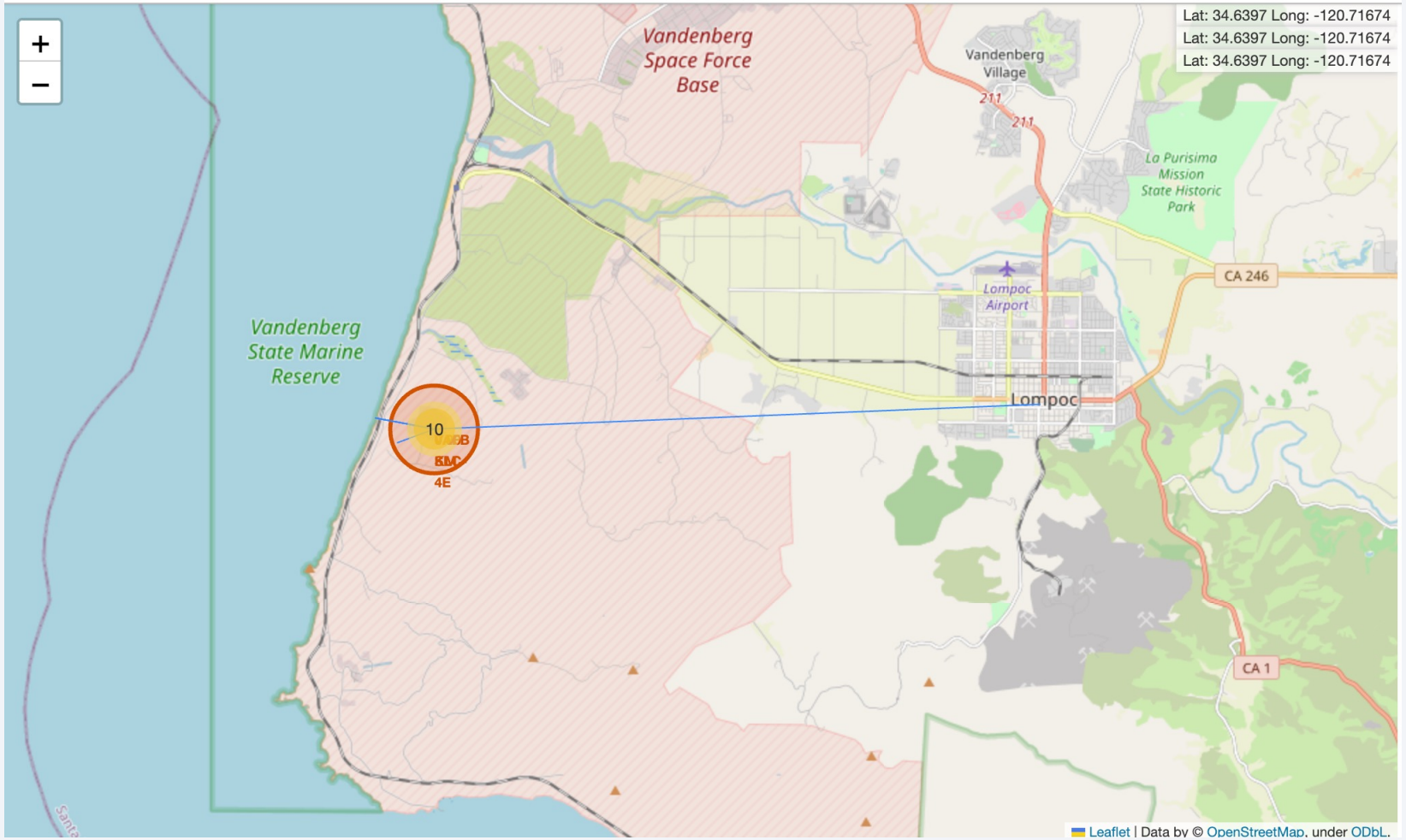
Mark success and failed launches for every site



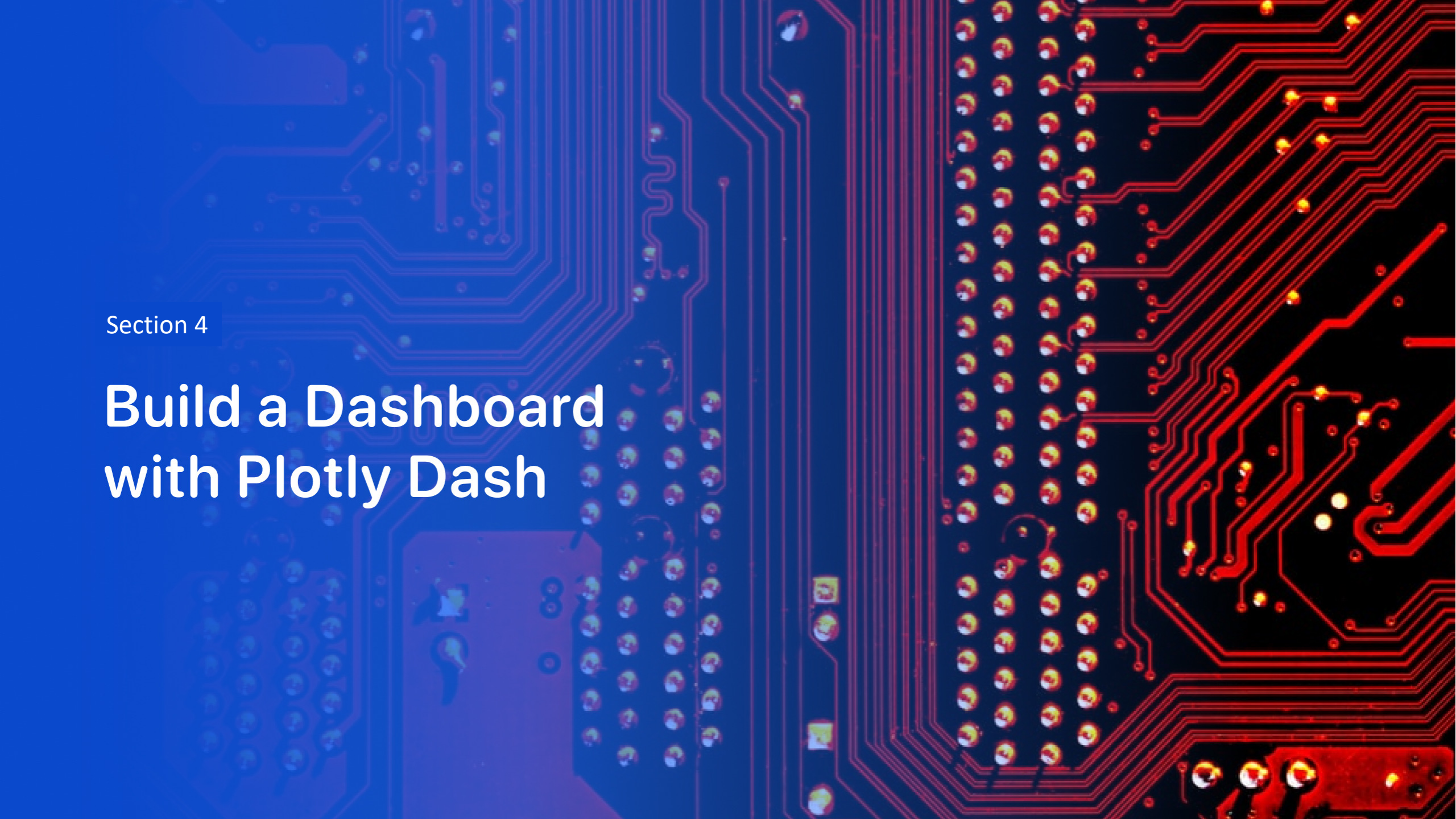


- We can notice both the total number of lunches for each site as can be seen in a yellow circle and if we zoom in, for example on the VAFB SLC 4E site we can see that 4 out of 10 launches were successful
- The KSC LC-39A has 10 successes out of 13; the CCAFS LC-40 has only 7 out of 26 successes and CCAFS SLC-40 has 3 out of 7 successes.

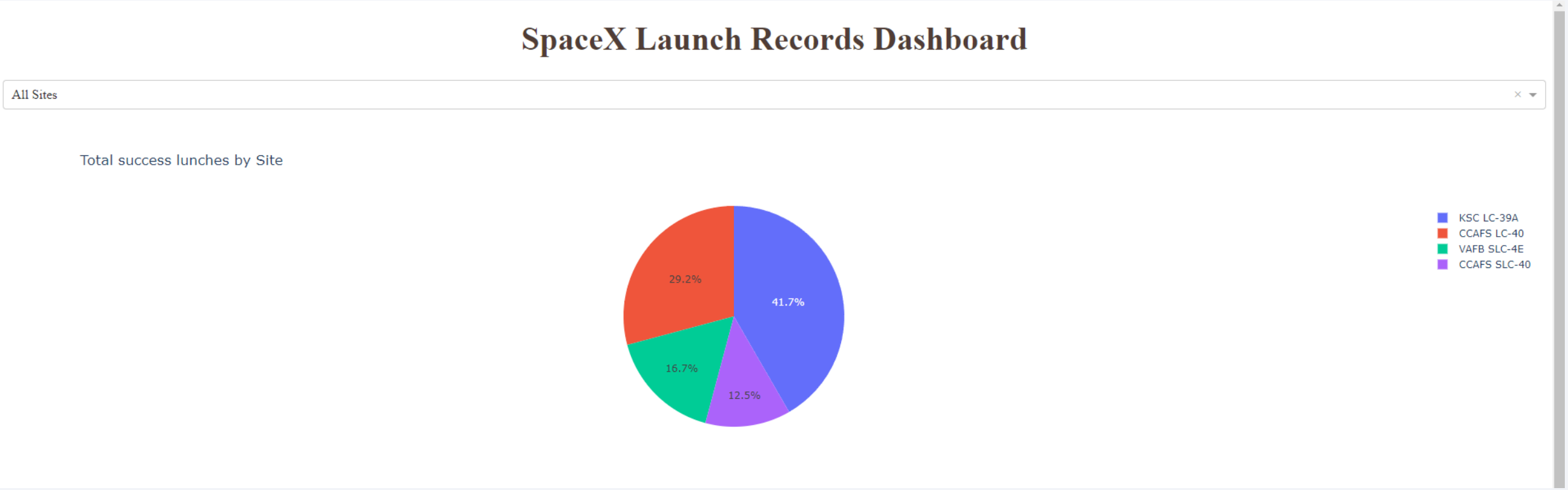
Compute distances from launch sites to proximities



- Here we are looking at the VAFB SLC-4E site and we can see that the displayed lines indicate the distance from the site to the closest coastline, railway and to the closest city that is Lampoc.
- This allows us to see where the sites are located respectively to their proximities and therefore if there are any geographical patterns linked to each launch site.

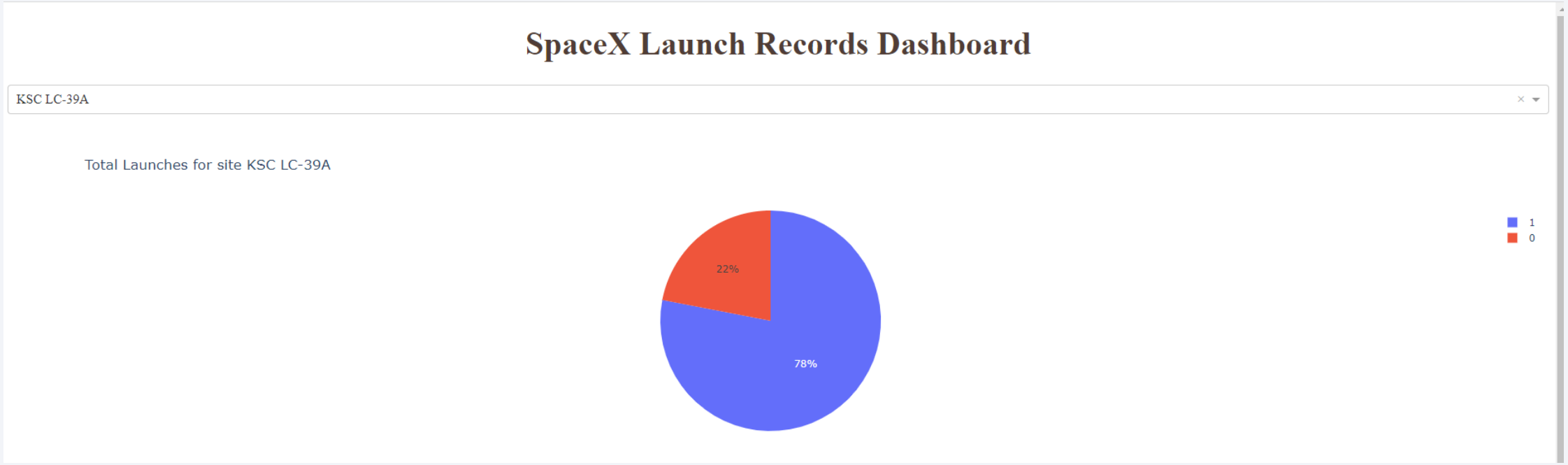

Section 4
Build a Dashboard with Plotly Dash

Successful lunches by Site



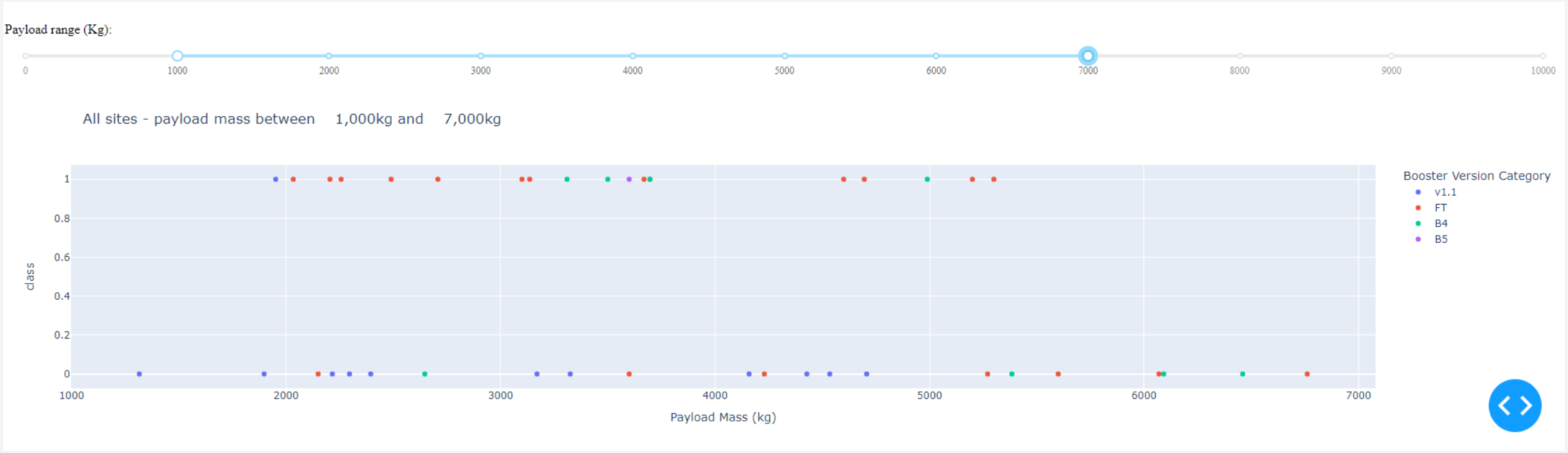
- The launch site with the highest success rate is the KSC LC-39A
- The launch site with the lowest success rate is the CCAFS SLC-40

Success rate for the KSC LC-39A site



- The KSC LC-39A is the site with the highest success rate with a 78% of successful landing and only 22% of unsuccessful landings.

Payloads vs. Launch Outcome scatter plot



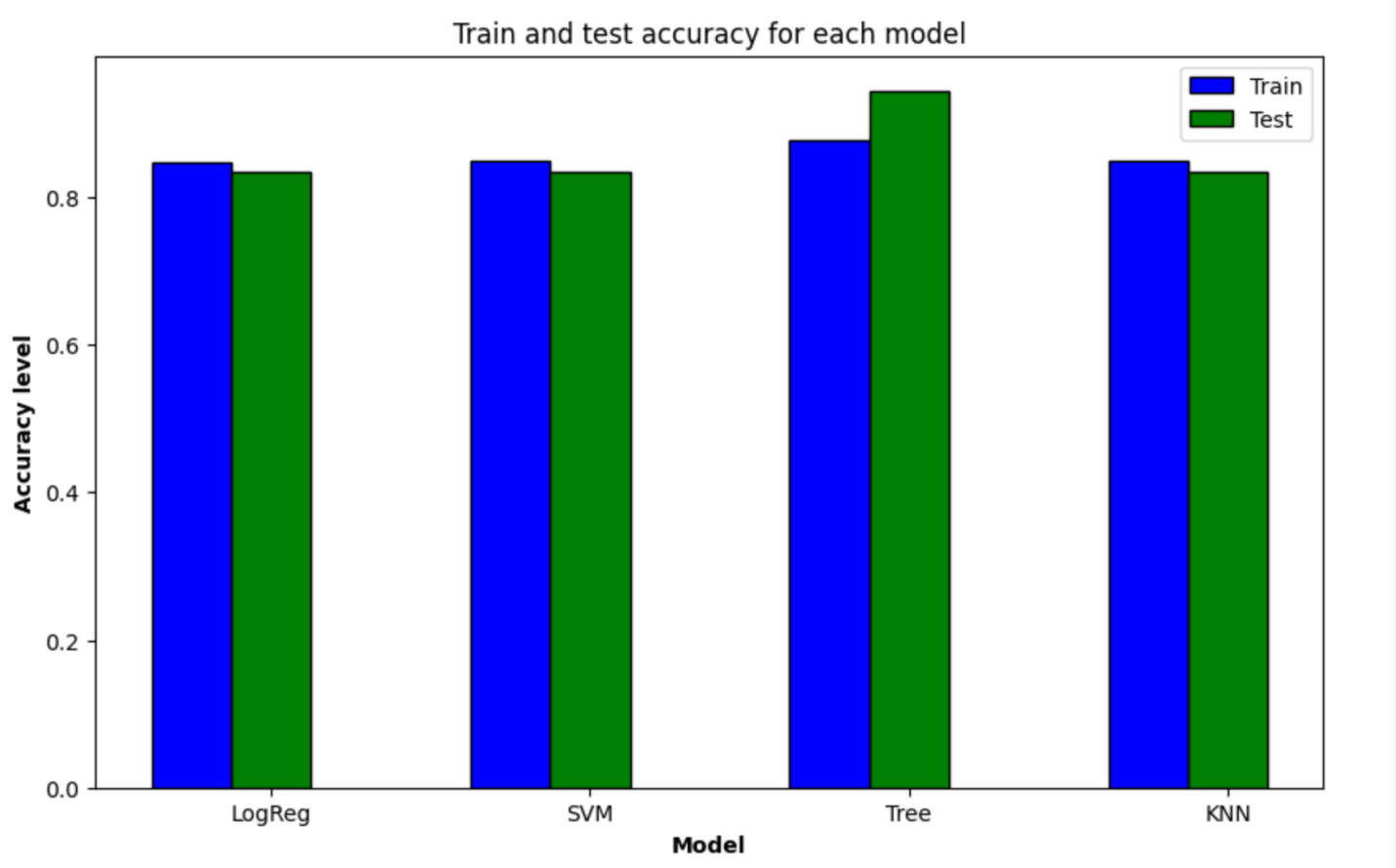
- This scatter plot shows the relationship between Payload and Launch Outcome for all sites, with a range of payload selected going form 1000 kg to 7000kg. It is also possible to see the dots colored by booster version.
- The FT booster version is the one with the highest success rate while the v1.1 is the one with the lowest.
- The higher the payload mass and the lower the number of successes.

Section 5

Predictive Analysis (Classification)

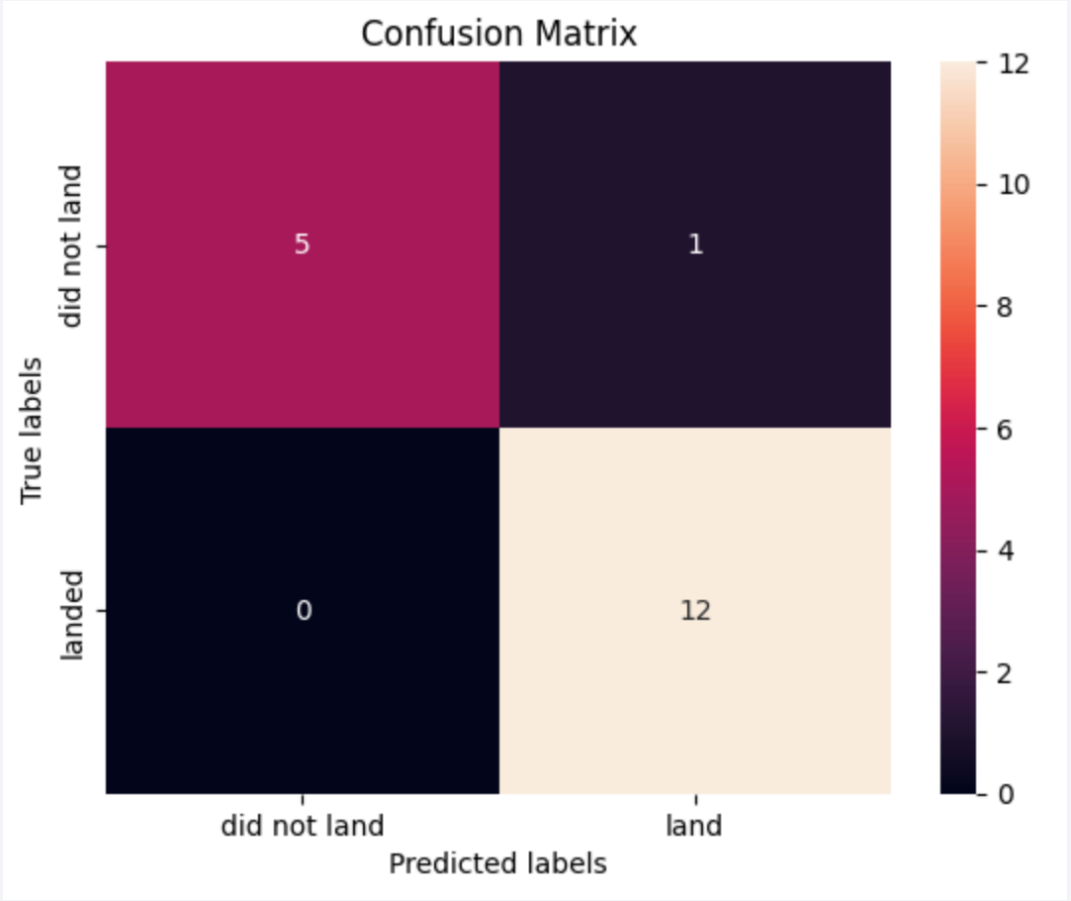
Classification Accuracy

- The best performing method is the Decision tree Classifier because all the tested methods have the same Test Accuracy of 94% but the highest train accuracy is given by the Decision tree with a 88%. The other methods are also very close to the best model, just with a slightly lower train accuracy.



Confusion Matrix

- Confusion matrix of the best performing model: Decision Tree
- We see that it can distinguish between the different classes.
- Only 1 launch that did not land successful has been predicted as landed correctly.



Conclusions

- The best launch site is KSC LC-39A with 78% of successful landings.
- The success rate has always been growing over time from 2013 to 2020.
- Each site has a different maximum payload to carry.
- In order to predict the successful launches we can use the decision tree classifier since it had the best train and test accuracy.
- There are 4 Orbits with a very high average success rate around 90%, while the others are all around 60%.
- The FT booster version is the one with the highest success rate while the v1.1 is the one with the lowest.
- By being able to predict a success/failed landing of the first stage, we will know the price of the launch.

Appendix

- GitHub link to the repository: https://github.com/beatricemarrano/spacey/tree/main

Thank you!