IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adolfo Godoy Araos
March 14th, 2024

Outline

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

Executive Summary

- Data was gathered from the publicly available SpaceX API and the SpaceX Wikipedia page. A new column called "Class" was created to categorize successful landings. Various exploratory techniques were employed, and only the relevant columns were selected as features for further analysis. The analysis showed important relationships between variables and identified the location of the launch sites. Categorical variables were converted into numerical format to prepare the data for modeling. Subsequently, the data was standardized, and GridSearchCV was utilized to identify the optimal parameters for the machine learning models. The accuracy scores of all models were visualized to gauge their performance.

- The analysis yielded four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Interestingly, all models exhibited a similar accuracy rate of approximately 83.3%. However, it was noted that all models tended to overpredict successful landings. This suggests a potential need for additional data to refine the models and improve their predictive accuracy. Further data collection efforts could enhance the robustness and reliability of the machine learning models in predicting SpaceX's first stage reuse outcomes.

Introduction

- Companies are offering space travels to customers, and Space X is one of the biggest players related to space lunches. Their Falcon 9 rocket lunches cost around 62 million dollars, while this number is way greater for other competitors.
- This is because of the Space X can reuse the first stage.
- A new company Space Y wants to join in the business, and the only way they have to do this is to predict which rockets are going to land safely and be reuse in the first stage.
- The question that we have to answer is. Given some features of the rocket, is it going to land safely?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Use of SpaceX API and SpaceX Wikipedia Page.
- Perform data wrangling
  - Adjusting variable types and labeling each lunch as a successful landing or unsuccessful landing.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create different Machine Learning Models and Tune them using GridSearchCV

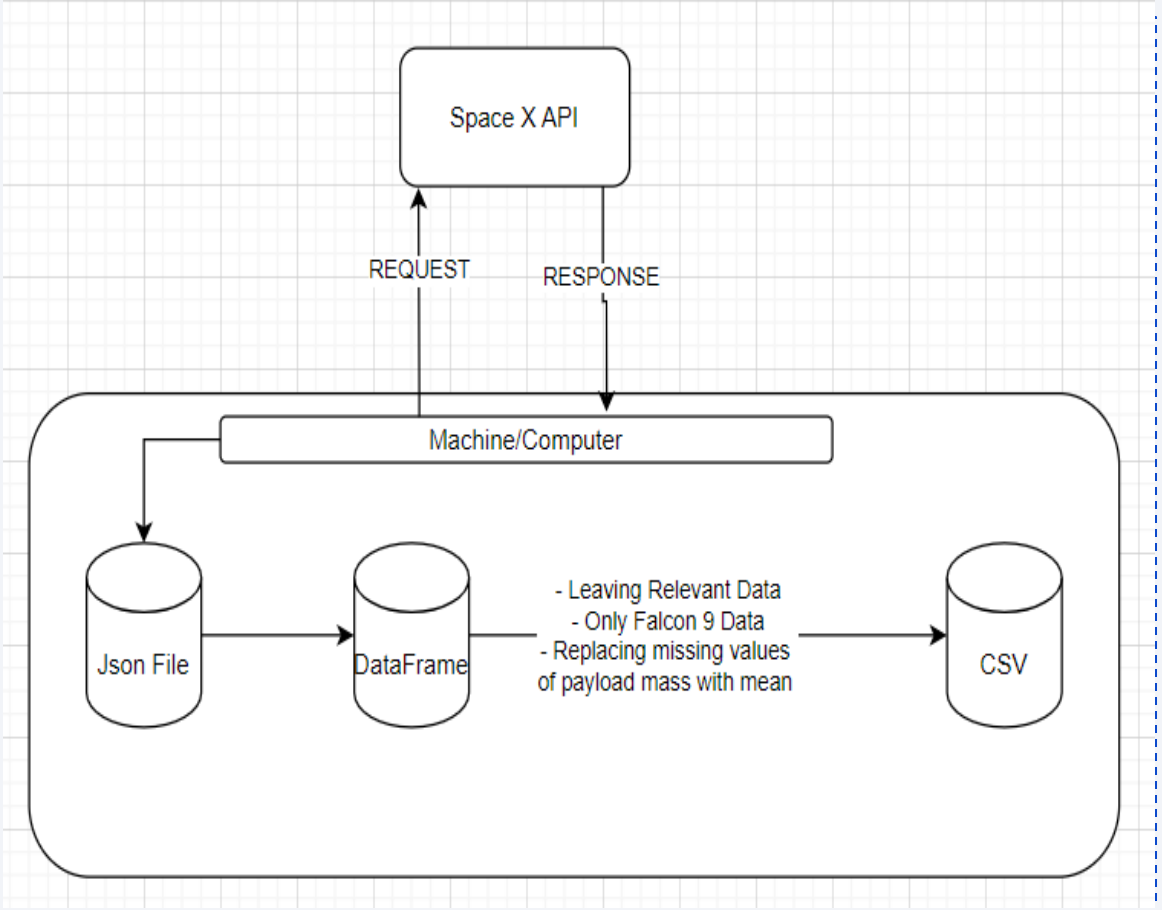
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

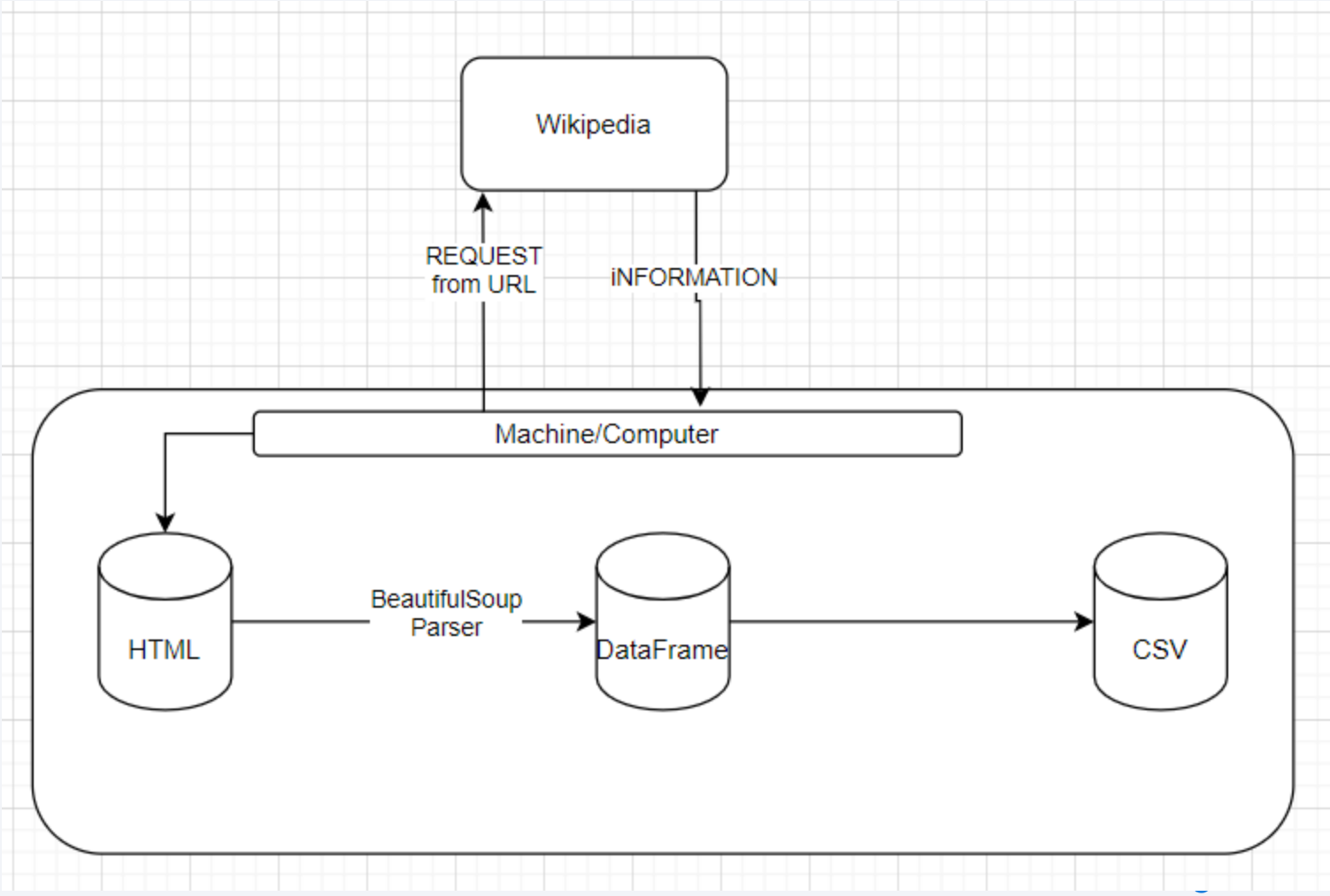
- First we request the data from the API and it gives us the response. Then we translate this information into a dataframe, we make some adjustments and finally export the information a csv file.

- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%201/Data%20Collection/Data-Collection-API.ipynb



Data Collection - Scraping

- First, we call use the URL to request for the information. Then we search for the columns we need, and using BeautifulSoup, turn into a dictionary. Finally we translate this into a dataframe and then a csv.
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%201/Data%20Collection/Data-Collection-Wikipedia.ipynb



Data Wrangling

- First, we **calculate** the launches per each site, orbit and possible outcome.
- Then we **label** whether the launch had a successful (1) or unsuccessful (0) landing.

- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%201/Data%20Wrangling/Data-Wrangling.ipynb

EDA with Data Visualization

- During the exploration, multiples chart were plotted. Some of the most important ones were:
  - Success Rate vs Orbit: To identify if some orbits have better chances than others.
  - Success Rate vs Year: To identify the improvement within the years.
  - Orbit vs Flight Nunber: To search for a relationship between those factors.
  - Obit vs Payload Mass: To understad which orbits are realted with major Mass Payloads
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%202/EDA%20with%20Data%20Visualization/Data_Visualization.ipynb

EDA with SQL

- Some of the queries performed were:
  - Show Distinct launch sites
  - Show 5 records of launch sites that begin with CCA
  - Show the total mass carried
  - Show the avg mass carried per record
  - Show the total successful and failed outcomes
  - Show first date of a successful ground pad achieved.
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%202/EDA%20with%20SQL/SQL.ipynb

Build an Interactive Map with Folium

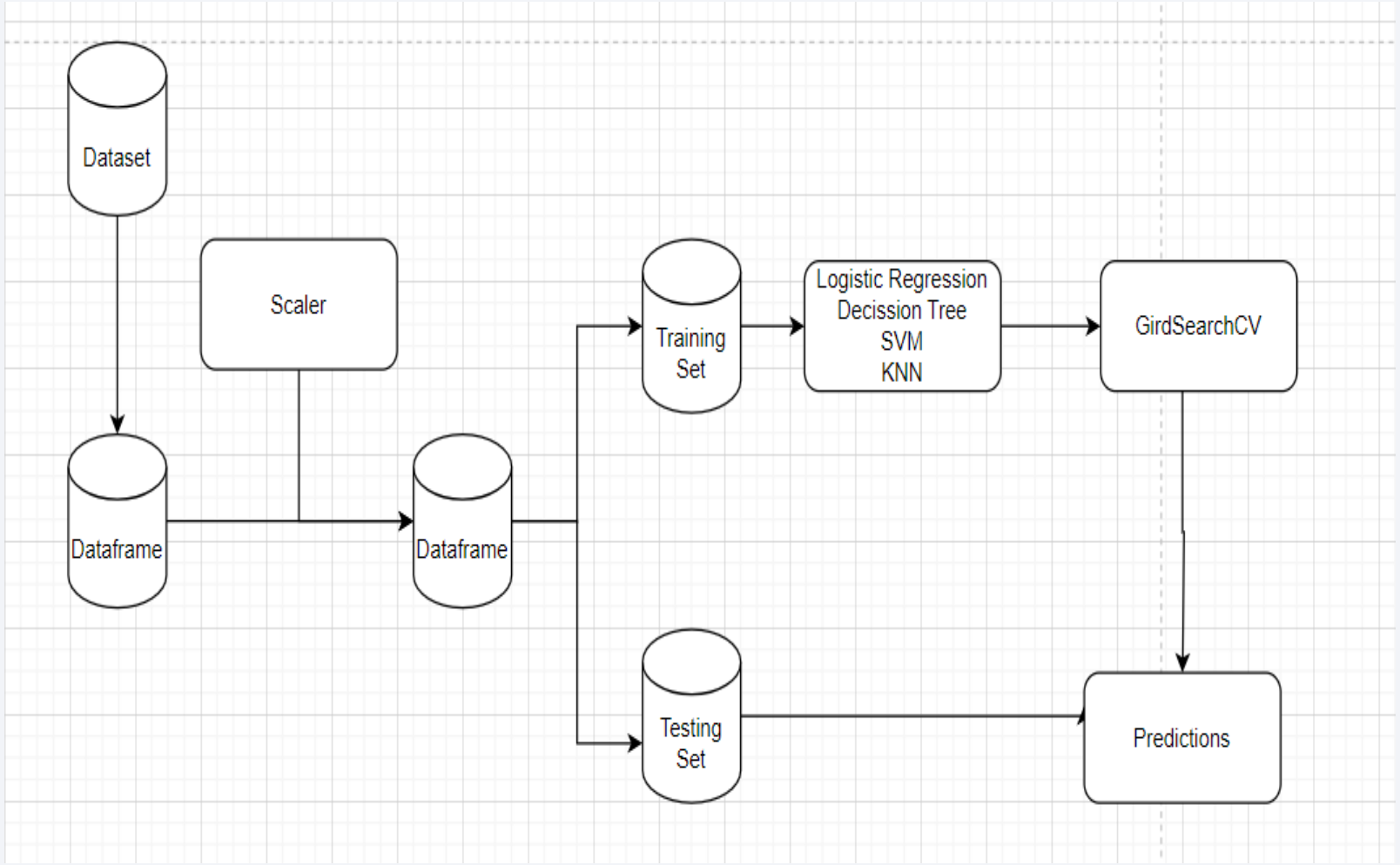
- A map was created in Folium, showing with a mark the launch sites in the US. It also shows the successful and unsuccessful landing sites, and how close they are to some other points like the city, coast or highway.
- It allowed us to identify if the relationship between the probability of success vs the geography of the launch.
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%203/Interactive%20Visual%20Analytics%20with%20Folium/Interactive_Analysis_Folium.ipynb

Build a Dashboard with Plotly Dash

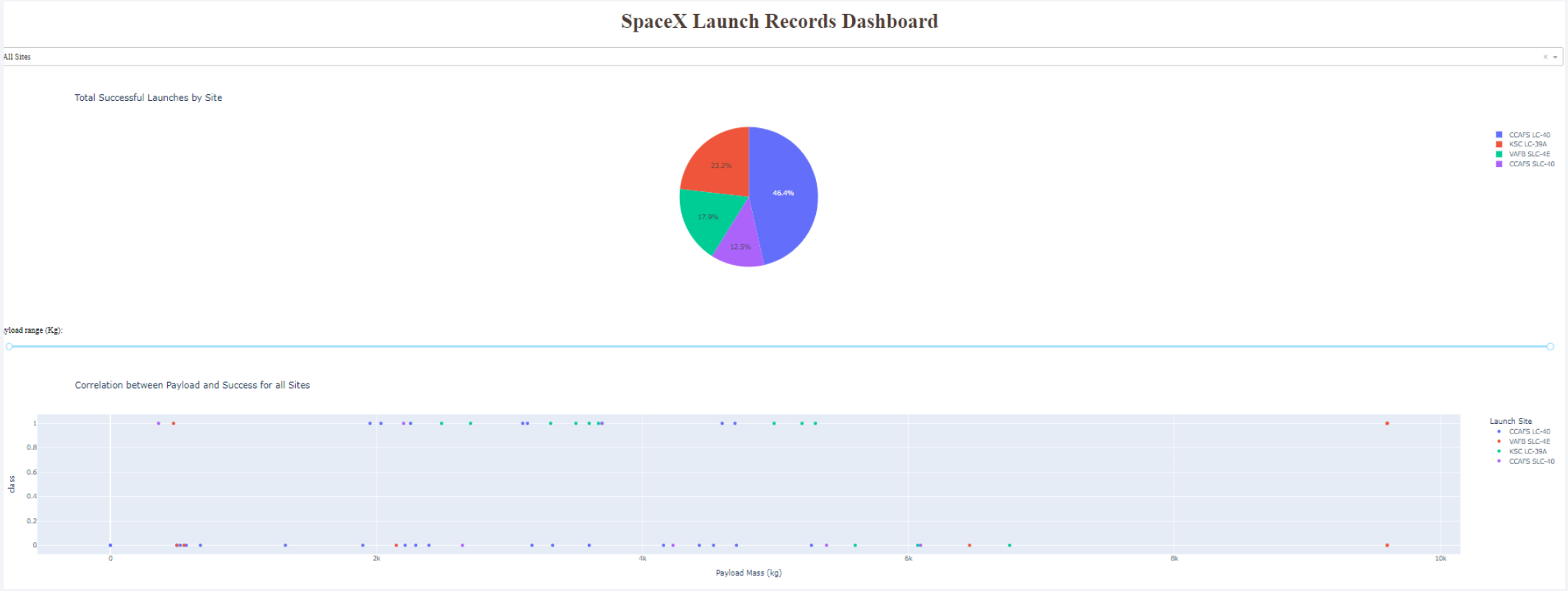
- There we 2 important charts in the dashboard.
- The first one is a pie chart that allows us to see the distribution of successful landing in each launch site.
- H secund on is the scatter plot, that takes into account the site and the payload mass
- Explain why you added those plots and interactions
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%203/Interactive%20Dashboard%20with%20ploty%20dash/Ploty_Dashborad.py

Predictive Analysis (Classification)

- We prepare the data, then training multiple models, and tunning their parameters using GridSearchCV.
- Finally, we use the best performing model with the test set.
- https://github.com/aagodoy1/Applied-Data-Science-Capstone/blob/main/Week%204/Machine_Learning_Prediction.ipynb



Results

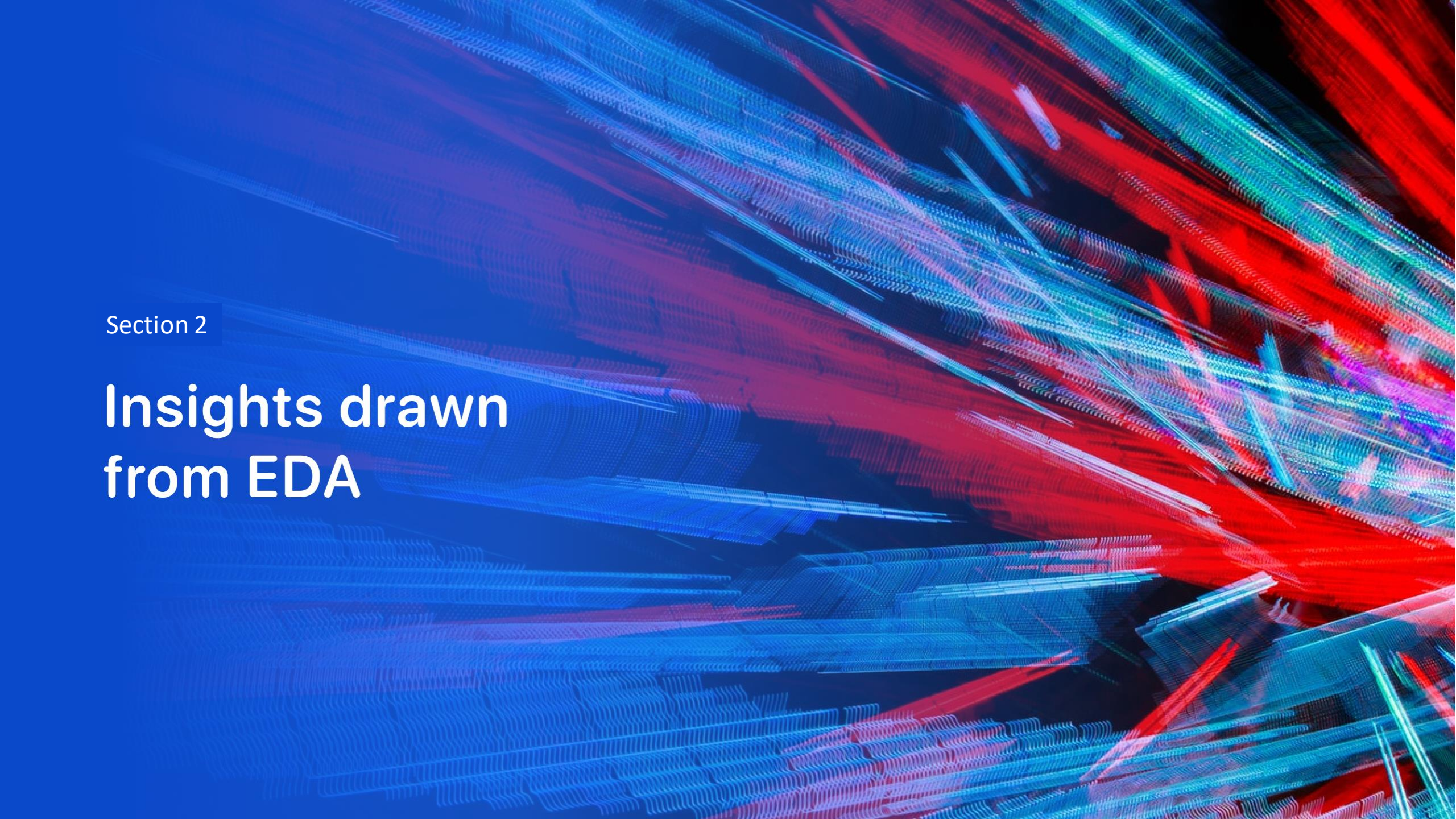


The exploratory data analysis showed the number of successful landings per launch site. Then we could see that CCAF-SCL 40 is the launch site with highest success rate.

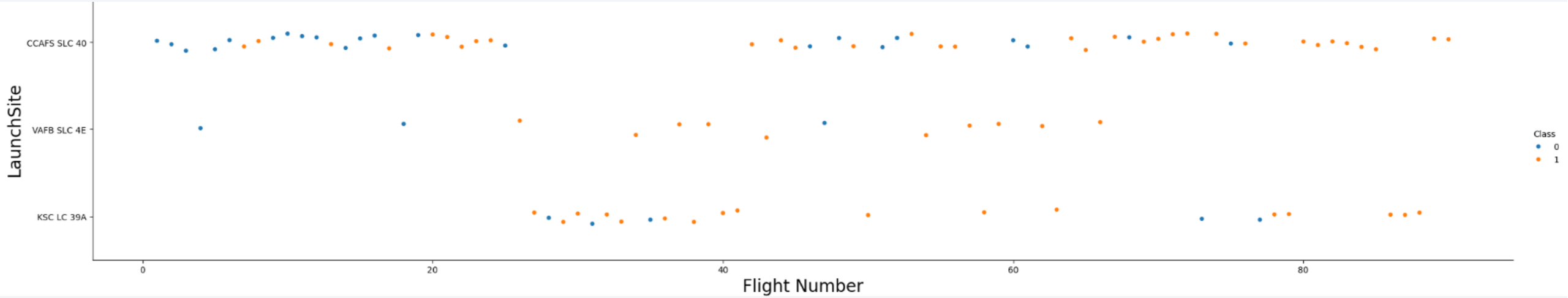
On the other hand, all the models had a 83.33% accuracy, and the over estimate the successful landings.

Section 2

Insights drawn from EDA



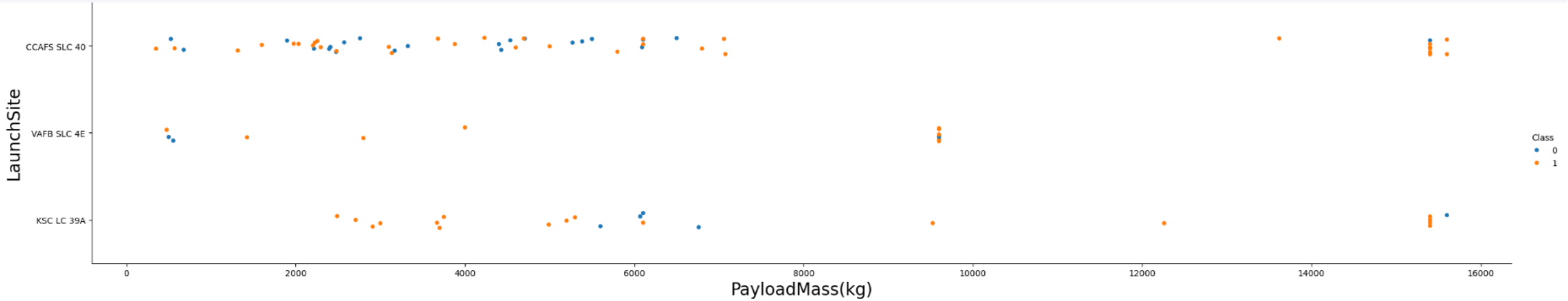
Flight Number vs. Launch Site



Blue dots indicate unsuccessful launch, while orange dots show the opposite.

The Launch Site VAFB has less than 15 records, however most of them were successful.

Payload vs. Launch Site

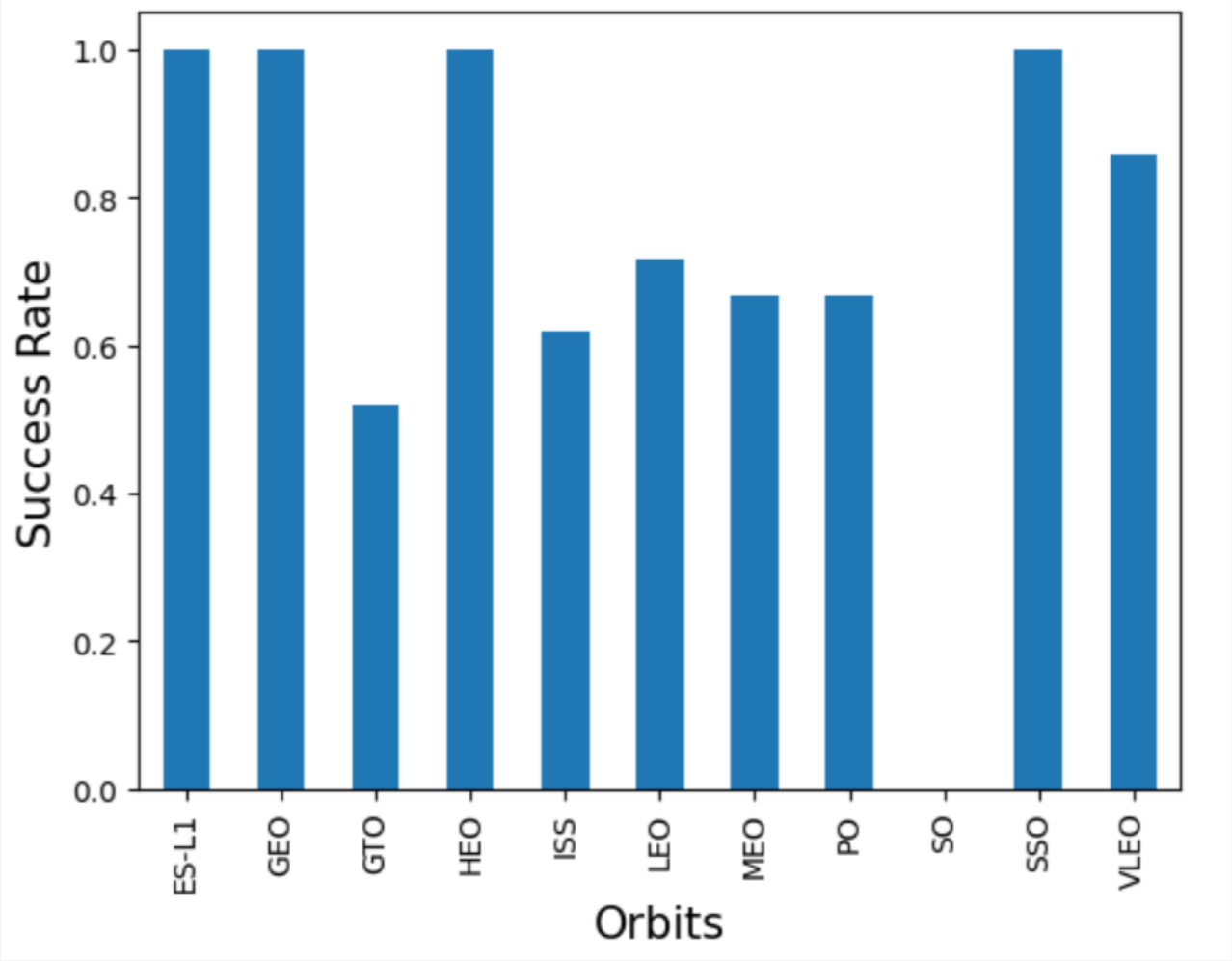


Note that VAFB Launch Site doesn't reach the 10.000 kg in its flights. On the other hand, the sites CCAAFS and KSF barely have flights between 7.000 kg and 15.000 kg.

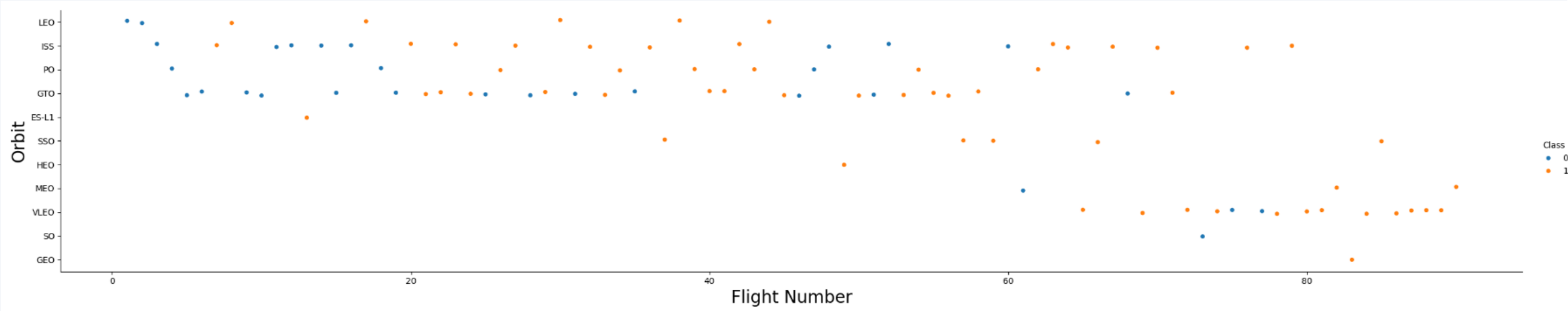
Success Rate vs. Orbit Type

The orbits ES-L1, GEO, HEO and SSO have 100% of success rate. While SO, has 0%.

Finally, is the only other orbit that has less than 60% success rato.

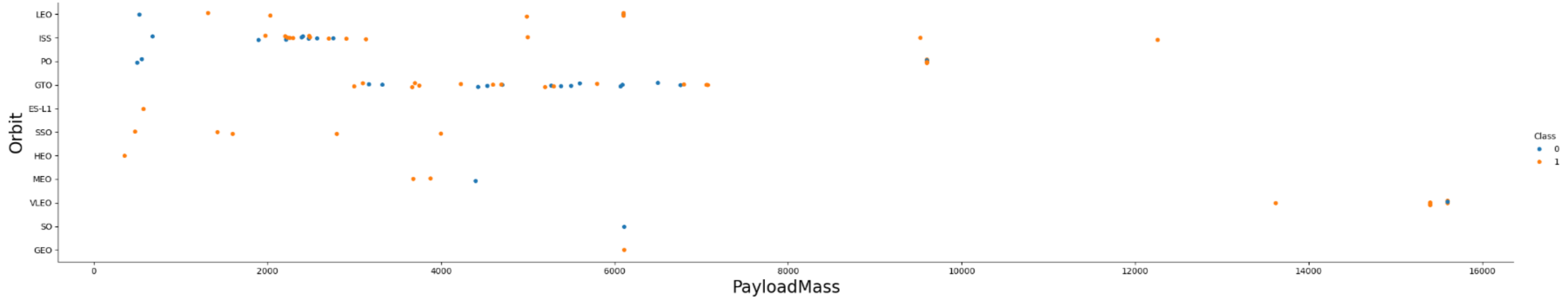


Flight Number vs. Orbit Type



The first flights were to the Orbits LEO, ISS, PO and GTO. Meanwhile, the newest flights has been to the orbits MEO and VLEO.

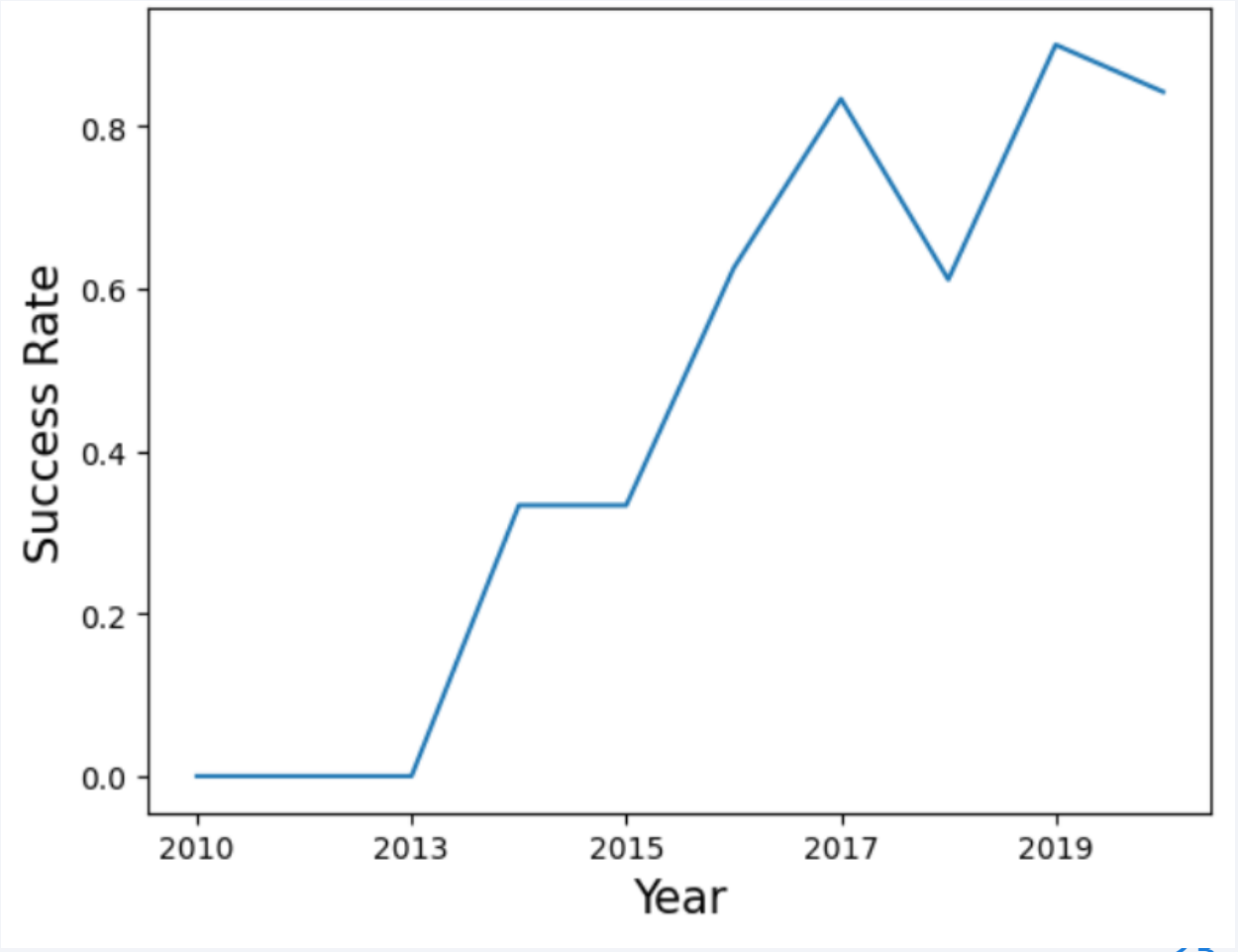
Payload vs. Orbit Type



The orbit VLEO is the destination of the flights with major Payload Mass. The orbit ISS present flights mostly between 2.000 and 3.000 kg.

Launch Success Yearly Trend

One might think that every year the success rate should be better. However, we can see that this is not the same, for example in 2018 the success rate was less than in 2017.



All Launch Site Names

```
%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

There only 4 launch sites

Launch Site Names Begin with 'CCA'

We can see the 5 elements needed. 2 of them have 0 payload_mass.

```
12]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No at
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No at
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No at

Total Payload Mass

The total payload carried by boosters from NASA is 45.596 kg.

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

```
%sql select sum(PAYLOAD_MASS__KG_)   from SPACEXTABLE where customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2.928 kg

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'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is 2015-12-22

```
%sql select min(date) from SPACEXTABLE where landing_outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is the next one:

```
%sql select distinct Payload from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ betw
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes is the next one:

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group by 1
```

```
 * 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

The list of names of the booster which have carried the maximum payload mass is only the Staling 14

```
%sql select distinct Payload from SPACEXTABLE A join (select Booster_Version, sum(PAYLOAD_MASS__KG_) as suma from SPAC
```

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

Payload
Starlink 14 v1.0, GPS III-04

2015 Launch Records

The list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is the next one:

```
ion, launch_site from SPACEXTABLE where substring(Date, 0, 5) = '2015' AND landing_outcome like 'Failure (drone ship)%'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	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

- The ranking 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 is:

```
%sql select LANDING_OUTCOME, count(*) as count from SPACEXTABLE where Date >= '2010-06-04' AND Date <= '2017-03-20' GRO
```

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

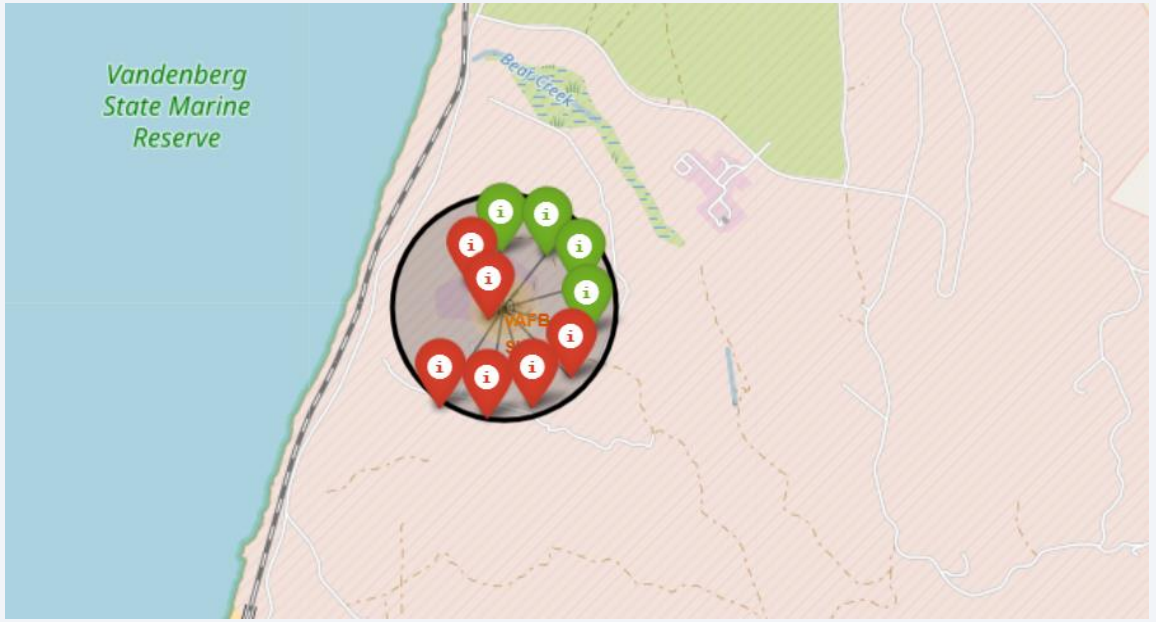
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites

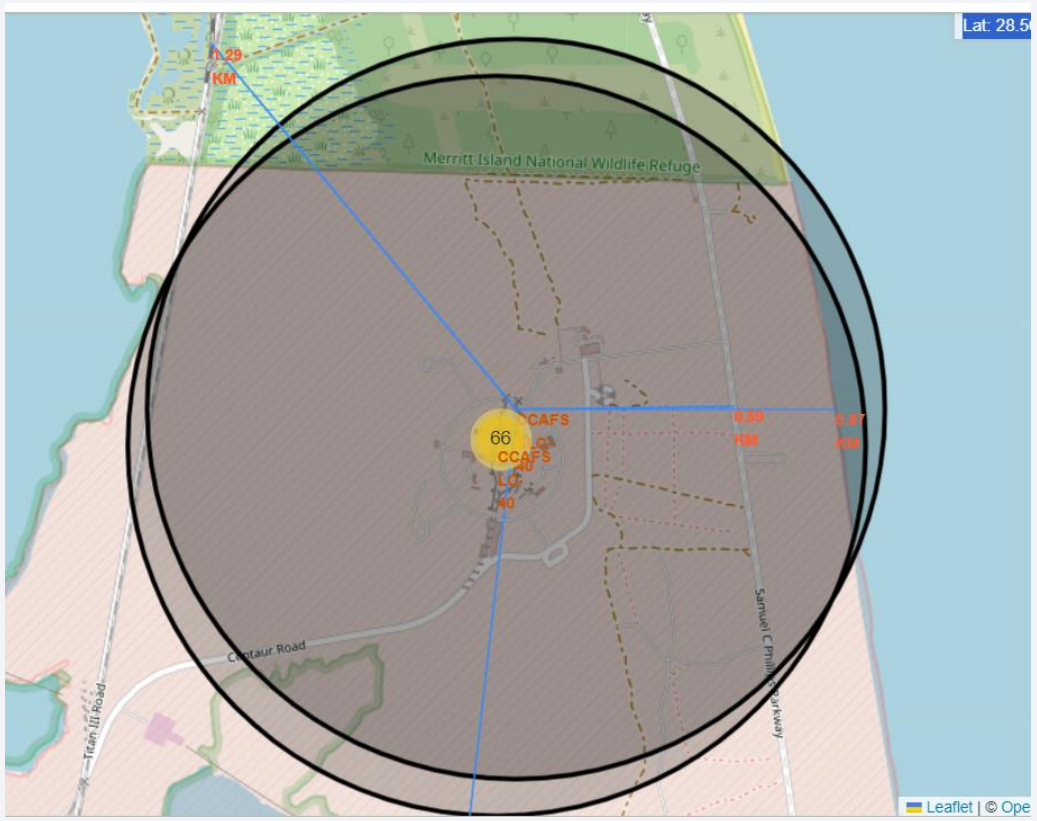
Most of the launch sites are in Florida, while only one is in California

Success Color Markers

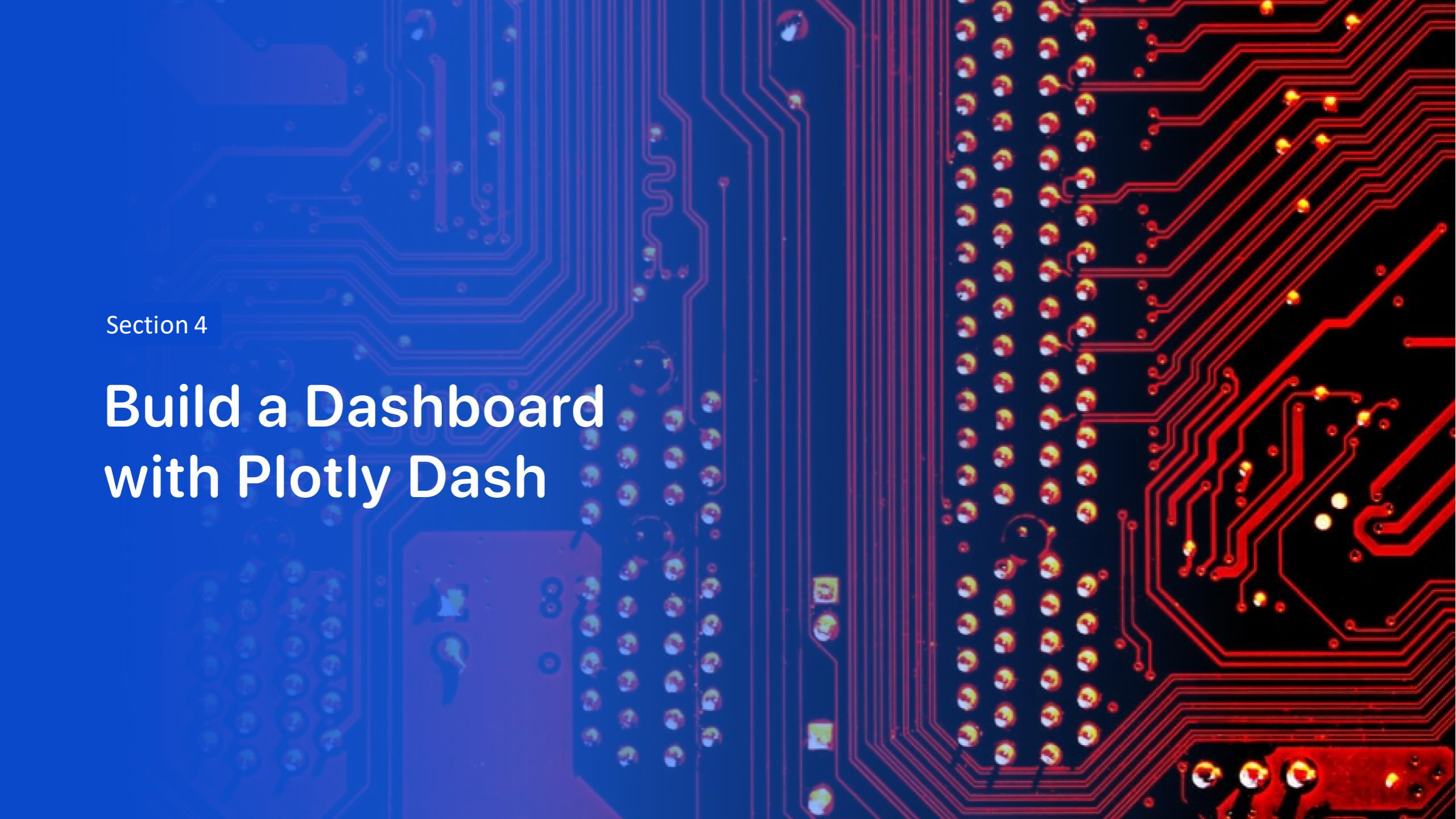


If we click each Launch Site, we can see the landing sites too, where those in red failed and those in green were successful.

Distance Markers from Important Sites



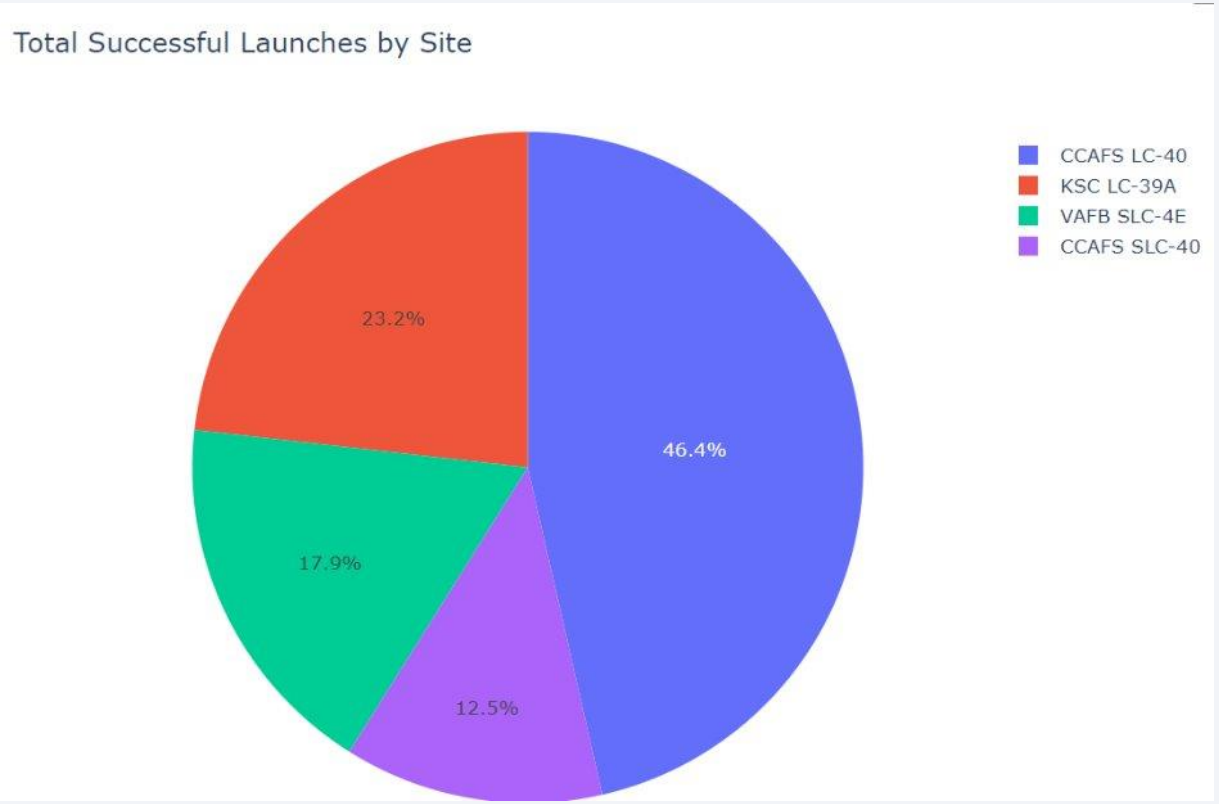
We also can see which important sites are close from the launch sites, and the distance to them.


Section 4
Build a Dashboard with Plotly Dash

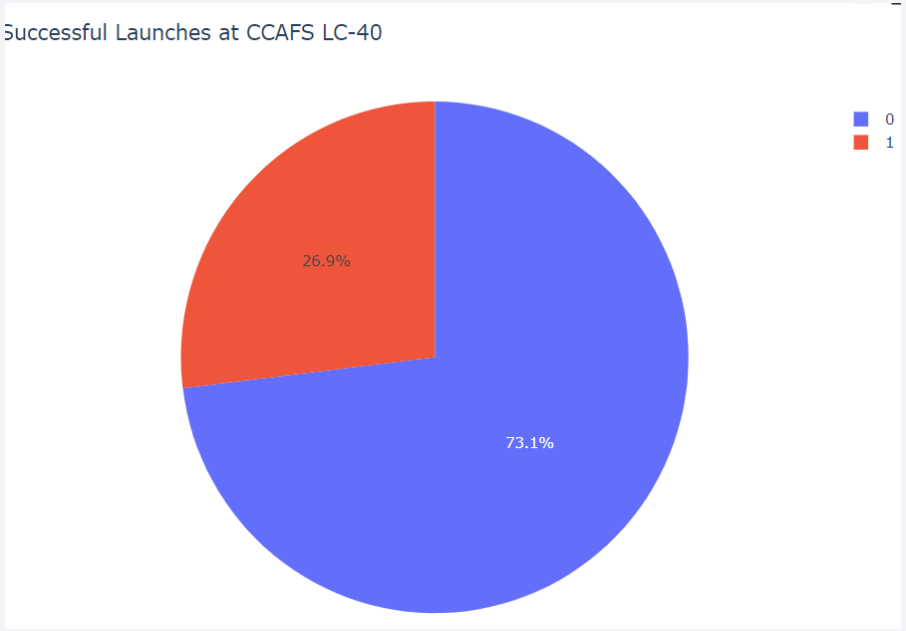
Distribution of successful launch sites

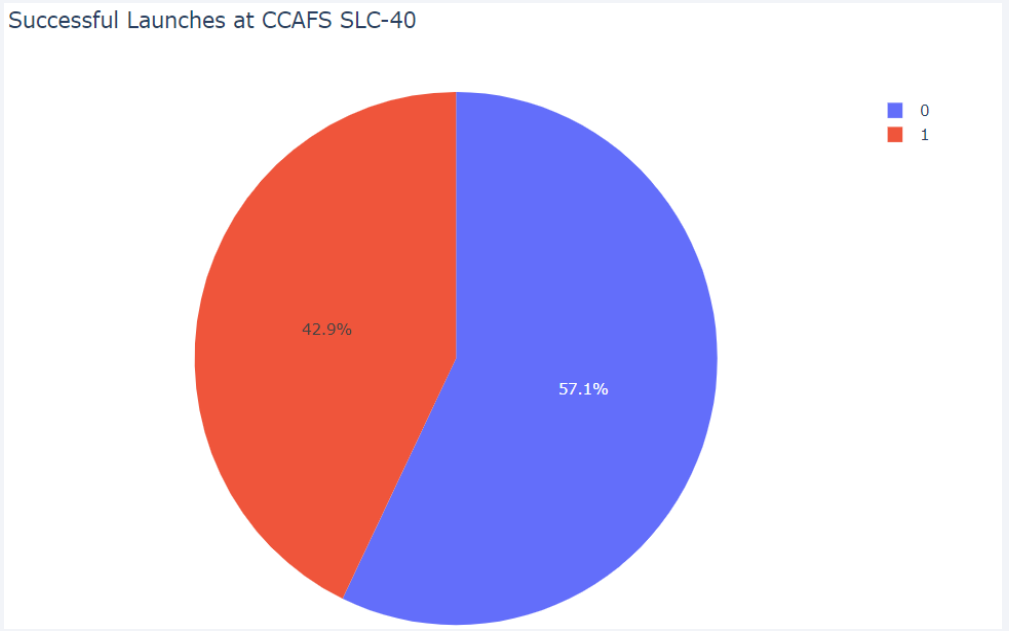
We can see that the Launch Site CCAFS LC is the one with more successful Launches.

The smallest one is CCAFS SLC, which can be explain by the proximity that we saw previously.



Comparation of Successful Rate

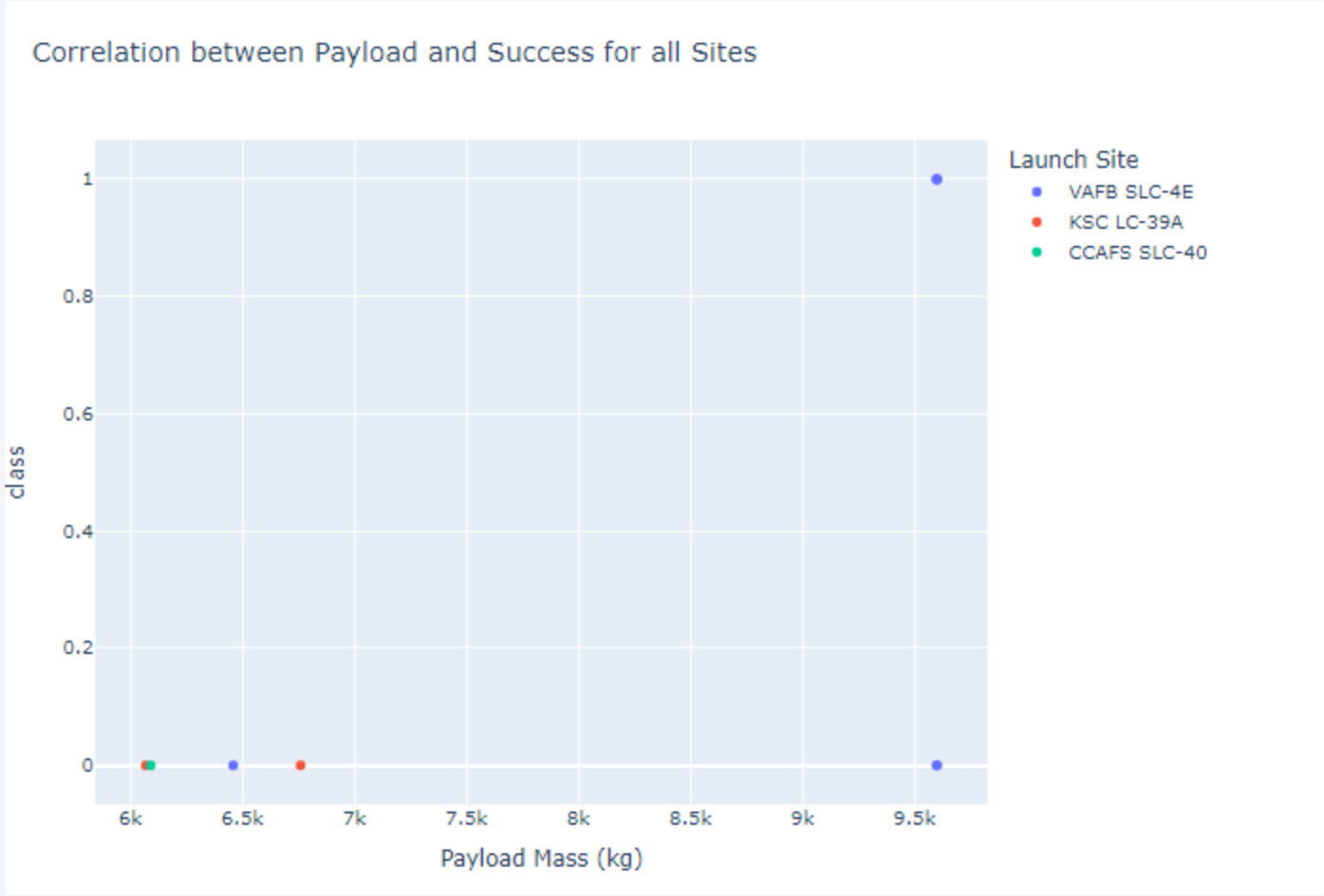




As we saw before, LC-40 has the most successful Launches and SLC-40 the least. However, is we take a closer look, we can see that LC-40 only have 27% of successful rate, vs the 43% successful rate of SLC-40.

Payload Mass vs Success

The most Interesting Insight moving the bar of payload mass, is that CCAFS LC-40 has no attempts of more that 6.000 kg. Meanwhile, when we look the amount of flights over the 6.000 kg, VAFB SCL-4E is the one that leads this graph.

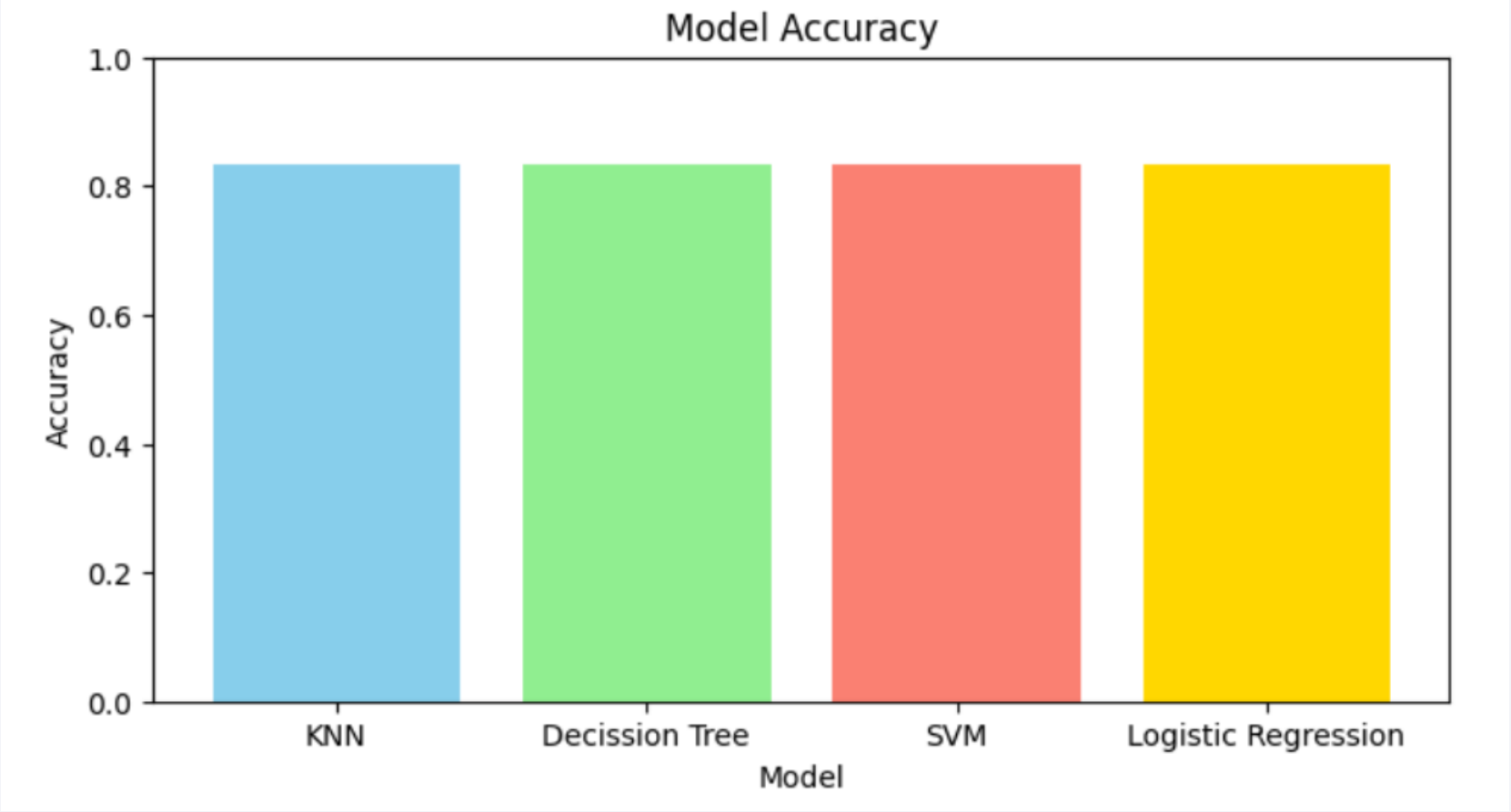


Section 5

Predictive Analysis (Classification)

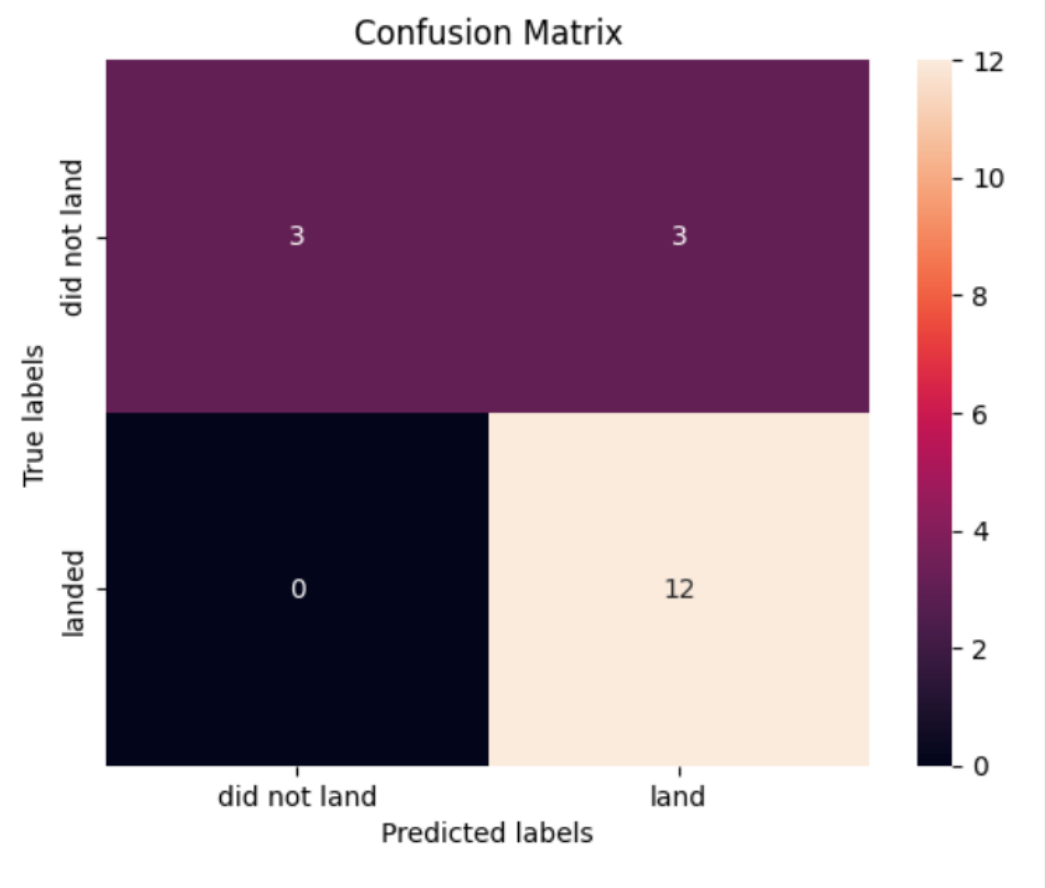
Classification Accuracy

All the models show the same accuracy, with 83.33%



Confusion Matrix

Since all the models have the same accuracy, this confusion Matrix works for all of them. Here we can see that it has no false negative, however is have the same true negative values as false positive, which might indicate that the model is **unbalance and over predicts successful landings.**



Conclusions

- The exploratory analysis showed that the best orbits to go to are ES-L1, GEO, HEO and SSO, since they have 100% success rate.
- 2019 was the year with the highest average success rate of all years.
- The launch site with highest success rate is CCAFS-SCL40
- Most of the flight have less than 8.000 kg payload mass. However, those that have more than that, present a high success rate.
- All the models had 83.33% accuracy, and all of them over predicted the successful landings, which might be a problem when estimating the cost of the project.

Appendix

If you would like to see the entire notebook, please go to this address:

https://github.com/aagodoy1/Applied-Data-Science-Capstone/tree/main

Thank you!