Winning Space Race with Data Science

<Marcial Galván Sosa>
<12 December 2024>

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
    - SpaceX API
    - Wikipedia web scraping
  - Data Wrangling
  - Exploratory Data Analysis
    - SQL
    - Seaborn Visualization + Featured Engineering
  - Interactive Dashboard
  - Predictive Analysis
- Summary of all results
  - SpaceX is still improving it successful landing rate
  - The development of a reusable first stage is mandatory in order to compete with them.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; 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.

- **Will it land?** The low cost the strategy is based on the reuse the first stage of Falcon 9, therefore if we can determine if the first will land, we can determine the cost of a launch. Therfore The aim of this project is, therefore, predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

In the elaboration of the current report we follow these steps:

- Data collection: Data collected from two main sources: SpaceX api and Wikipedia page about Falcon 9 Launches
- Data wrangling: Clean and classify the data, added a new feature 'landing_outcom' and apply one hot encoding to get data ready for modeling.
- Exploratory data analysis (EDA): Using visualization tools and SQL
- Interactive visual analytics: Using Folium and Plotly Dash
- Predictive analysis with classification models: Evaluation of 4 classification models: Logistic Regression, Support Vector Machine, Classification Trees and K-Nearest Neightbours

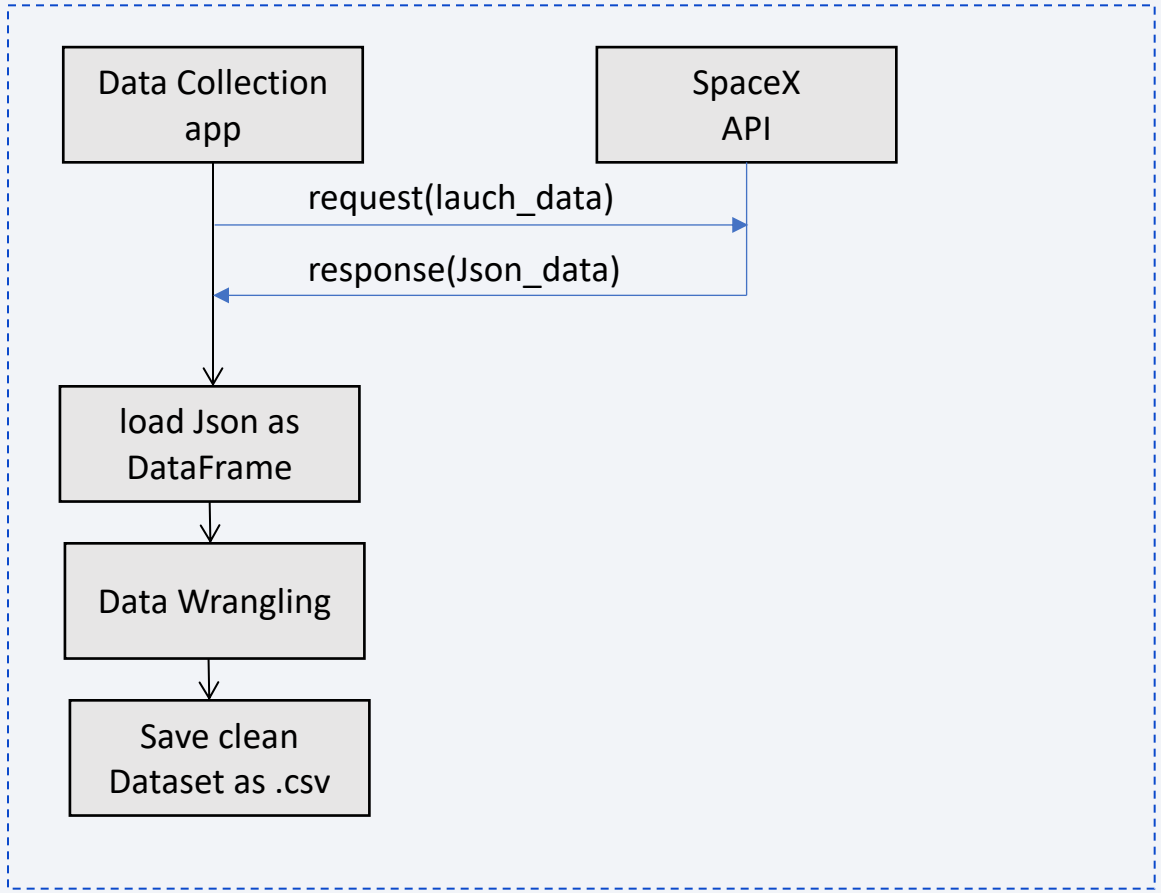
Data Collection

- Sources of Data:
  - SpaceX Api (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping from Wikipedia page about Falcon 9 Lauches (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

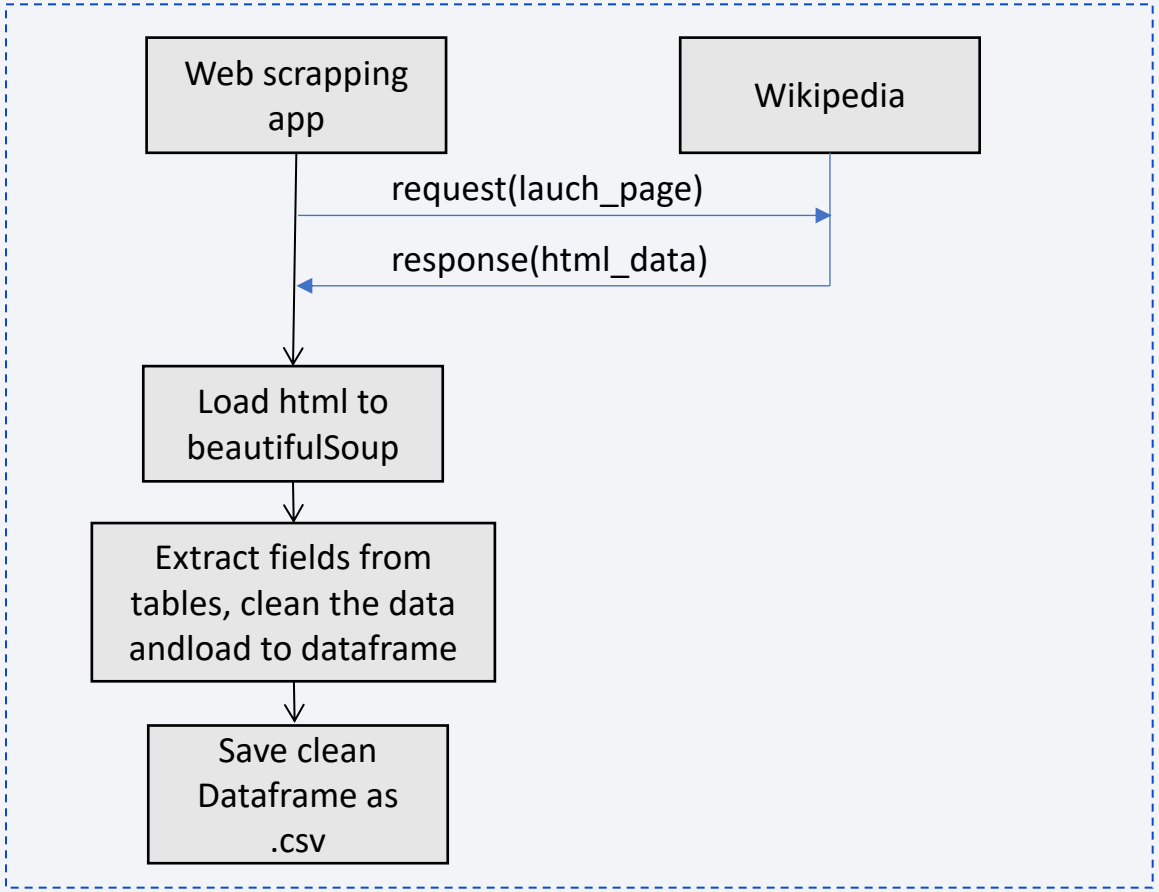
- Request launch data trought SpaceX API
- Load Json response to pandas Dataframe
- Data Wrangling
- Save clean data to the working dataset in .csv format

- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/01_SpaceX_Data_Collection_SpaceX_API.ipynb



Data Collection - WebScraping

- Request html data from Wikipedia
- Load data into BeautifulSoup
- Parsing launch html data
- Clean data
- Load data as pandas Dataframe
- Save dataframe as .csv
- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/02_SpaceX_Data_Collection_webscraping.ipynb



Data Wrangling

- Perform exploratory data analysis and determining Training Labels
- Exploratory data about
  - launches per site
  - Occurrence of each orbit
  - Occurrence of mission outcome of the orbits
- Creating the target label Class with values
  - 0 : First stage did not land successfully
  - 1 : First stage landed successufully
- GitHub: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/03_SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- The followed Charts were plotted:
  - Relationship between flight number and Launch Site
    - Allows to see the successful and failed launch by launch site
  - Relationship between Payload Mass and Launch site.
    - To observe if there is any relationship between launch sites and their payload mass
  - Relationship between success rate of each orbit type
    - Visually check if there are any relationship between success rate and orbit type
  - Relationship between Flight Number and Orbit type
    - To observe when and how each orbit was targeted
  - Relationship between payload Mass and Orbit Type
    - To reveal the relationship between payload Mass and orbit type
  - Launch success yearly trend
    - To get the average launch success trend

- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/05_SpaceX-EDA-data-visualization.ipynb

EDA with SQL (1/2)

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string ‘CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the data when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the Booster Versions which have carried the Maximun Payload Mass.

EDA with SQL (2/2)

- List records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
- 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

- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/04_SpaceX-EDA-SQLlite.ipynb

Build an Interactive Map with Folium

- Based on Folium Maps we added Markers, cluster markers, circle and polylines to enhance the data interpretation
- We added the objects in order to have geospatial information about:
  - Launch sites
  - Launch sites dependent of success/failures
  - Distance from launch sites to key infraestructures

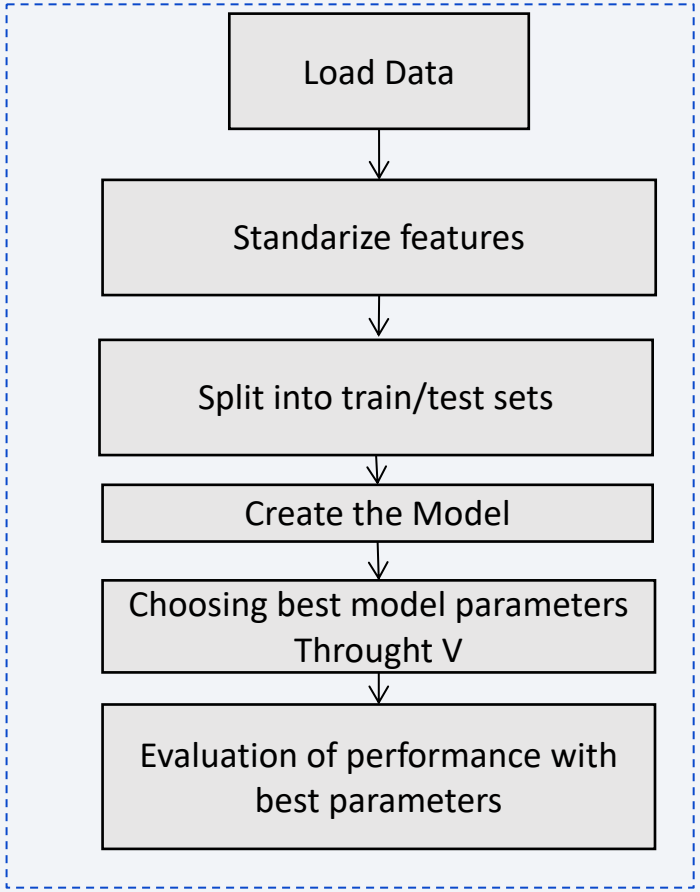
- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/06_SpaceX_Launch_Site_location_folium.ipynb

Build a Dashboard with Plotly Dash

- With Plotly Dash we use these components:
  - Dropdown menu
  - Piechart
  - Scatter plot
- Trought the dropdown menu we will select the launch site subject to analysis, showing the successs/failure proportions of launches from that launch site, we can select 'All' to have a global overview
  At the same time in the scatter view we see the detail about Payload Mass bring to orbit and the booster that make it possible, as well as the success/failure result.
- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/07_SpaceX_Dashboard_with_Plottly.py

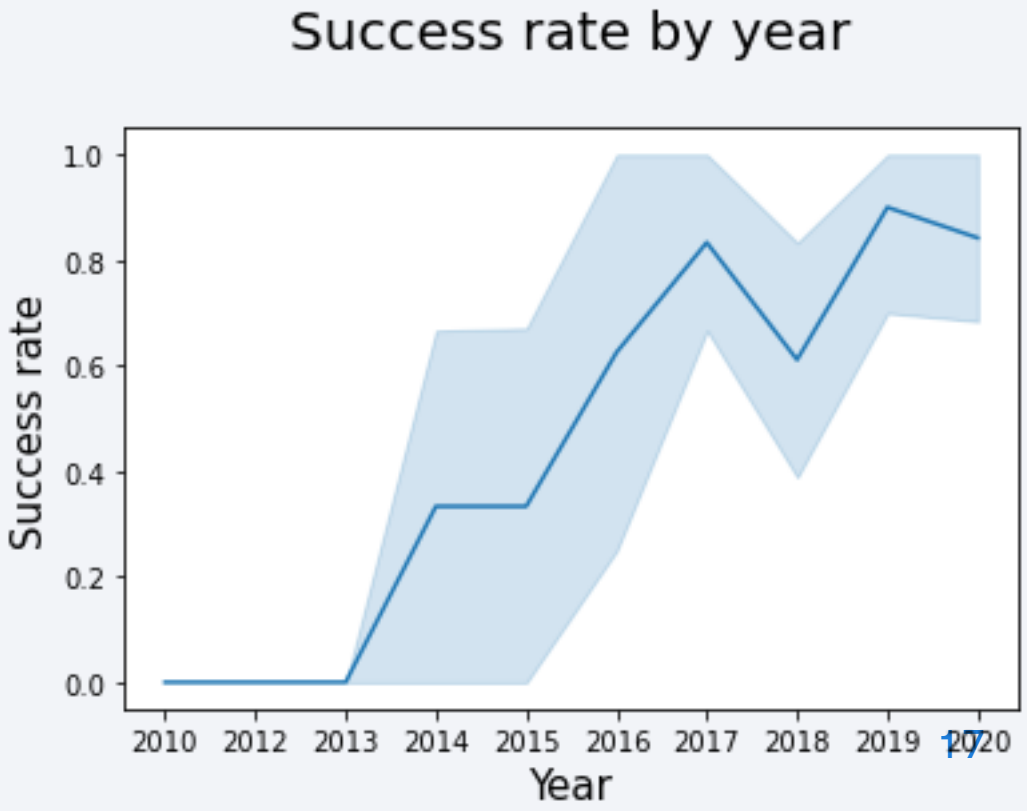
Predictive Analysis (Classification)

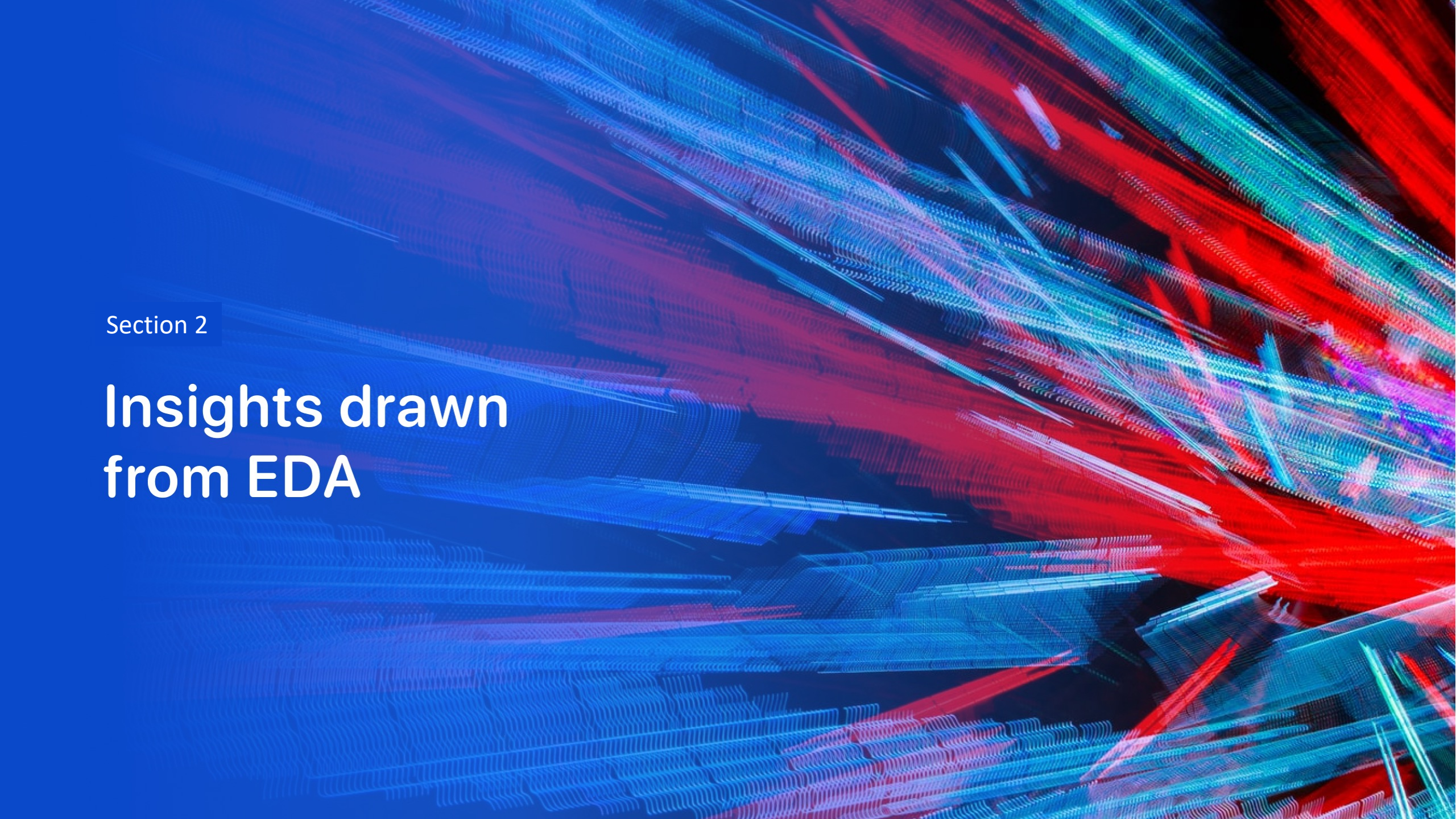
- After extracted from the dataset our features and target variable, the features are standardized and the data is divided into training and testing sets
- We build an train 4 classifications models (Logistic Regression, Support Vector Machine, Decission tree and K-nearest neighbour) and testing them using gridsearchCV to obtain the best parameters for each one.
- Evaluation of perform of the best parameters by using a Conffussion Matrix
- Final step is the comparision of all the models.
- Github: https://github.com/narcuak/Applied-Data-Sciencie-Capstone-IBM-SkillsBuild/blob/main/labs/08_SpaceX_Machine_Learning_Prediction.ipynb



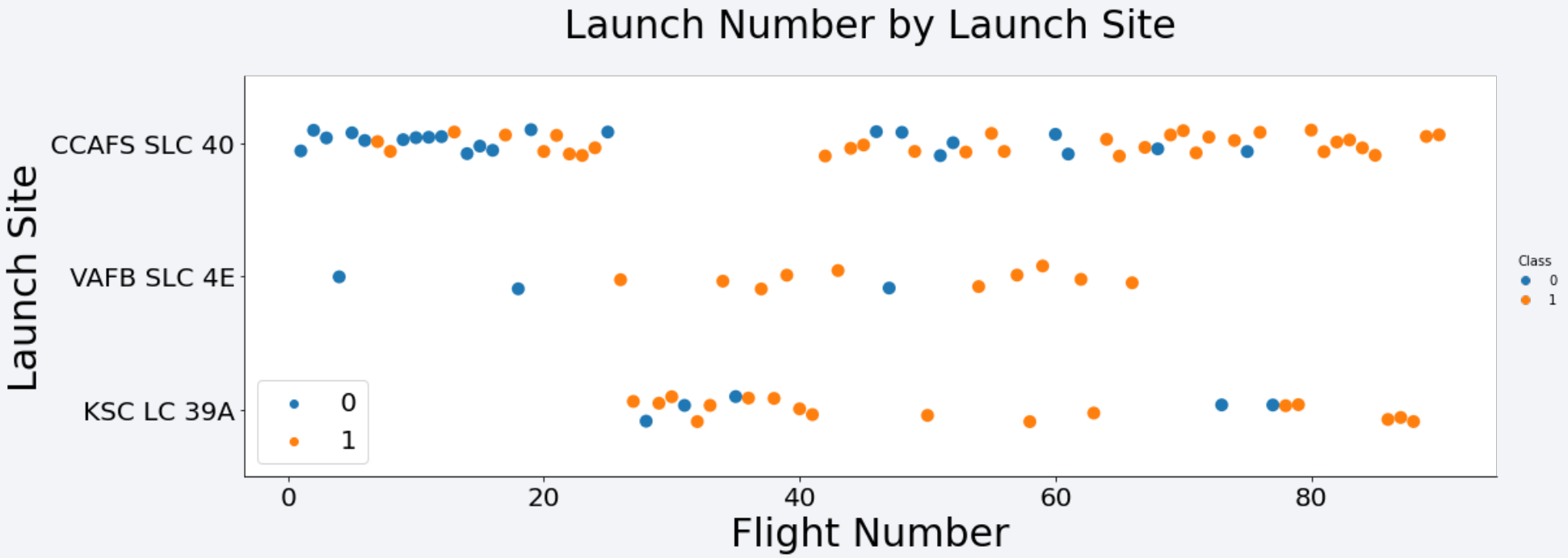
Results

- Exploratory data analysis results
  - Success rate is still increasin
  - There is a 100% success rate for launches to ES-L1, GEO, HEO, and SSO orbits
  - Launchs brings cargo up to 10.000 kg
- Interactive analytics
  - All lauch sites are in the coast lines
  - Lauch pad are strategically positioned avoiding important infrastrctures
- Predictive analysis results
  - All the predictive models performs the same.



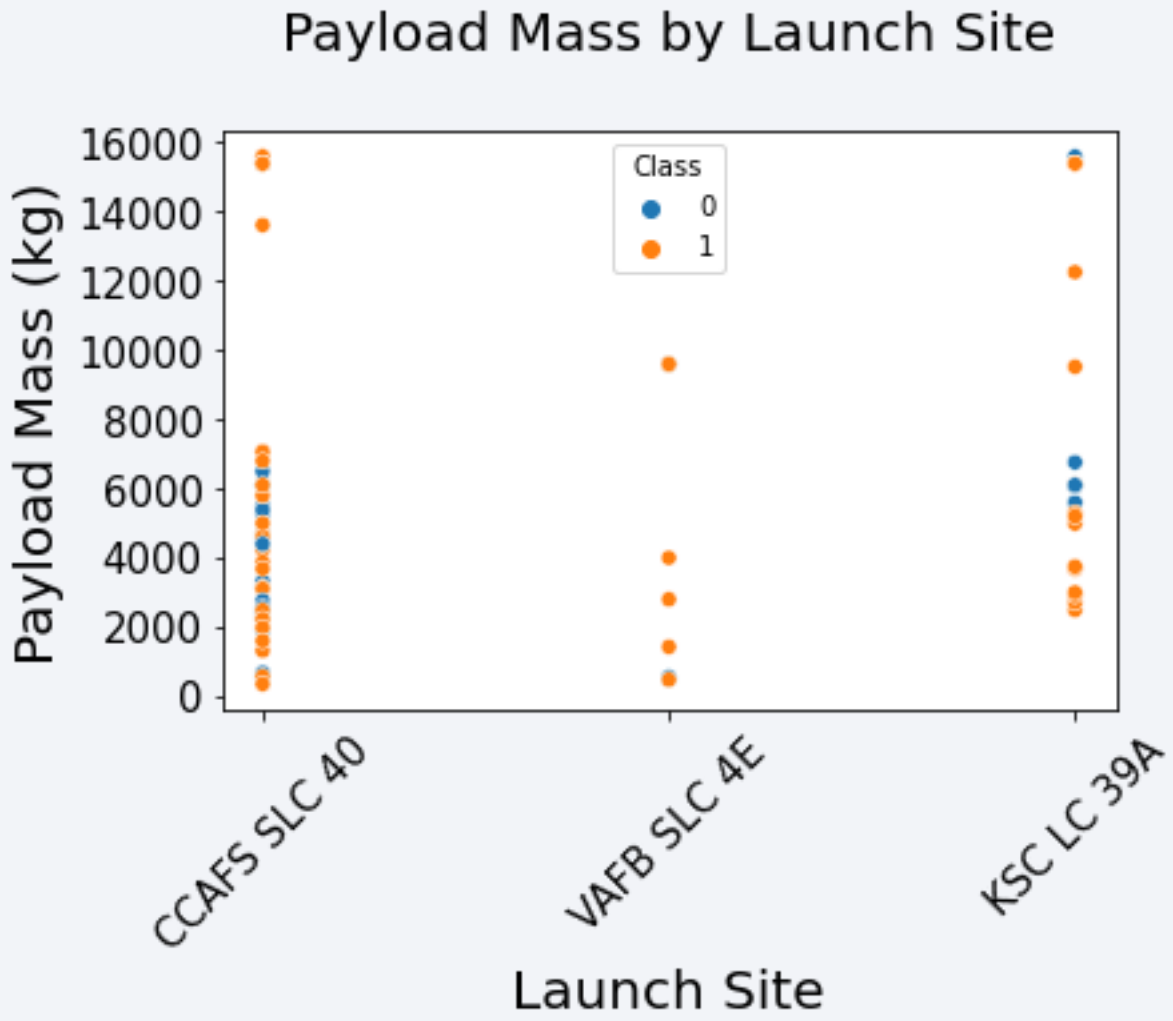

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



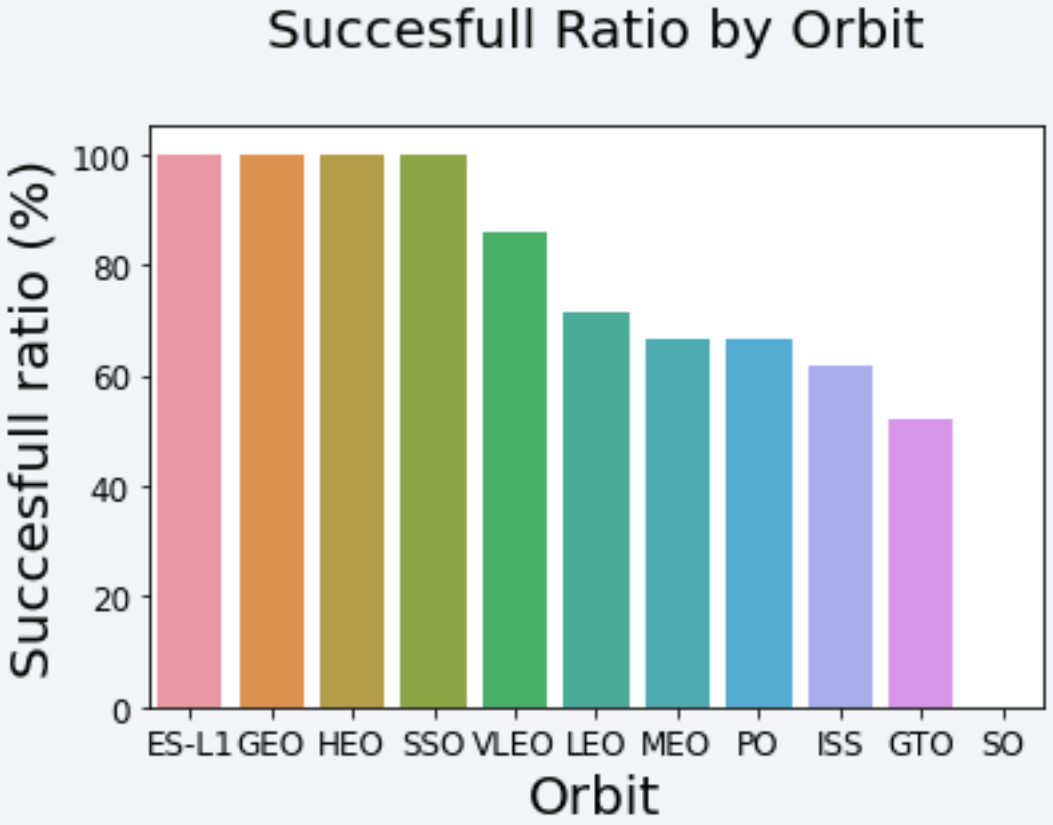
- The number of successful launches increases with the number of flights, which is expected as we are considering the entire history of launches since the beginning. We can see that the initial launch site was CCAFS SLC 40.
- Given the progression, we expect future launches to have an even higher success rate across all Launch Sites

Payload vs. Launch Site



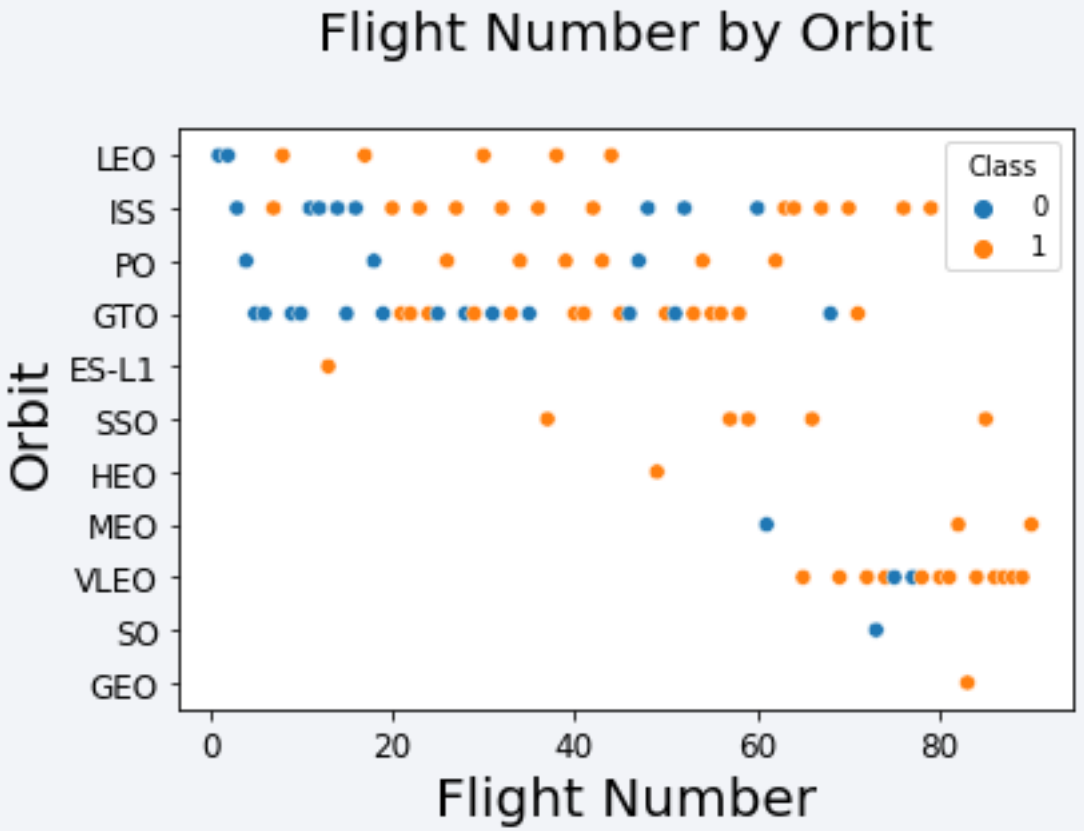
- Most launches with payloads over 8,000 kg have been successful.
- Most launches from VAFB SLC 4E have been successful, although it does not launch payloads over 10,000 kg.

Success Rate vs. Orbit Type



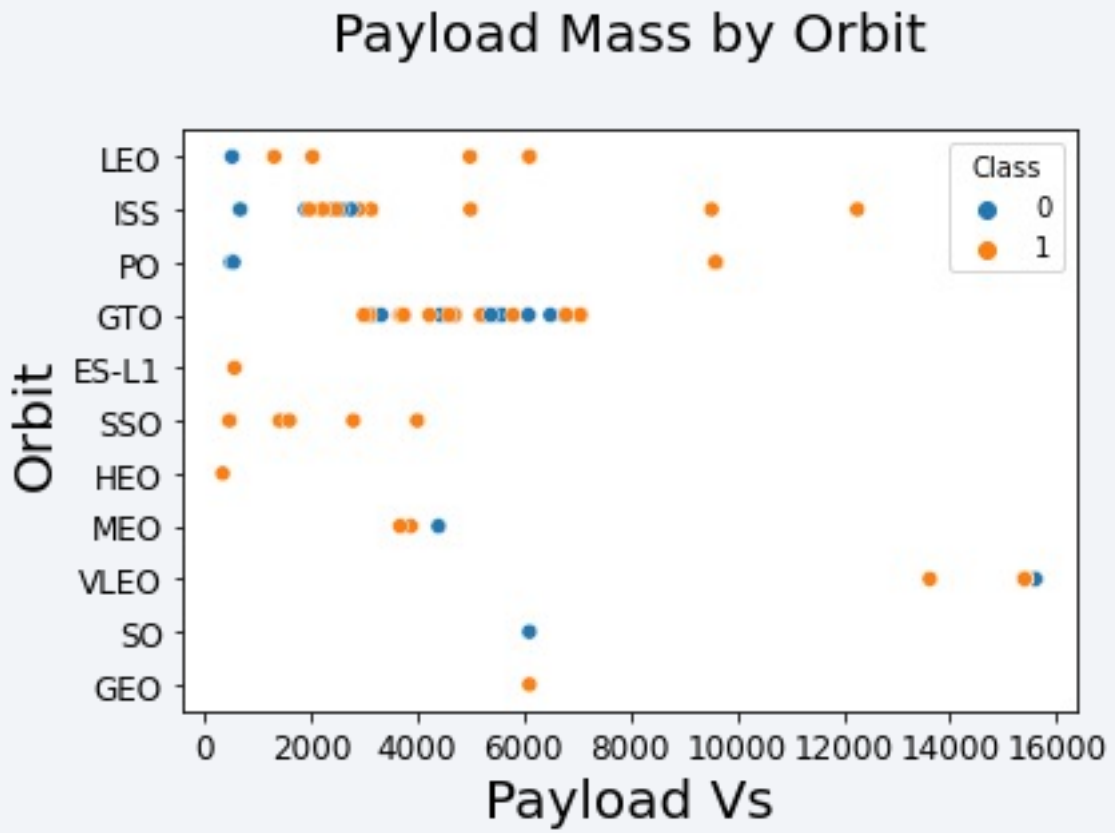
- There is a 100% success rate for launches to ES-L1, GEO, HEO, and SSO orbits
- Lowest succesfull orbit is GTO
- SO Orbit has 0% success.

Flight Number vs. Orbit Type



- As we have seen, the success rate increases for each orbit with the number of flights.
- The SSO, HEO, MEO, VLEO, SO, and GEO orbits were only attempted after gaining a certain level of experience, starting from flight 38.
- Last attempted orbir is GEO

Payload vs. Orbit Type



- The heaviest payloads (>8000kg) have only been launched to the VLEO, PO and ISS orbit.
- The orbit with the most consistent launch weight is GTO

Launch Success Yearly Trend

- Since 2013, the success rate has increased over the years, along with a decrease in uncertainty, although a slight decline is observed in 2018.
- It is expected that this rate will continue to improve over time as technology and expertice advances.

All Launch Site Names

```
%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

- We found 4 different launch sites:
  - CCAFS SL-40
  - CCAFS SLC 40
  - VAFB SLC 4E
  - KSC LC 39A

Launch Site Names Begin with '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

- 5 records where launch sites begin with `CCA` are showed, all of them to LEO orbit, mainly demo flights

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from spacextable
```

Python

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

sum(PAYLOAD_MASS__KG_)
619967

- The total payload carried by boosters from NASA are of 619967 Kg

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from spacextable
where booster_version like 'F9 v1.1%'
```

Python

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- average payload mass carried by booster version F9 v1.1 is 2534.6 kg

First Successful Ground Landing Date

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

Python

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

min(date)
2015-12-22

- First successful landing outcome on ground pad was on 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version, payload_mass__kg_ from
spacextable where landing_outcome = 'Success (drone
ship)' and PAYLOAD_MASS__KG_>4000 and
PAYLOAD_MASS__KG_<6000
```

Python

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_outcome, count(mission_outcome) as outcome_total from spacextable group by mission_outcome
```

Python

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

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

- Total number of succesful missions:
  - 99 + 1 with payload status unclear
- Failure missions:
  - 1

Boosters Carried Maximum Payload

```
%sql select Booster_Version  from spacextable where payload_mass__kg_ = (select
max(PAYLOAD_MASS__KG_) from spacextable)
```

✓ 0.1s Python

```
 * 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

- Names of the booster which have carried the maximum payload mass:
  - 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

```
%sql select date,strftime('%m', date) as launch_month, substr(Date, 6,2) as
month_2, substr(date, 0,5) as year,booster_version,landing_outcome,launch_site
from spacextable where landing_outcome = 'Failure (drone ship)' and strftime('%Y',
date) = '2015'
```

Python

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

Date	launch_month	month_2	year	Booster_Version	Landing_Outcome	Launch_Site
2015-01-10	01	01	2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
2015-04-14	04	04	2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- There were two failed landing_outcomes in drone ship:
  - Booster F9 v1.1 B1012 from CCAFS LC-40
  - Booster F9 v1.1 B1015 from CCAFS LC-40

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

```
%sql select landing_outcome, count(landing_outcome) from
spacextable where (date <= '2017-03-20' and date >=
'2010-06-04') group by landing_outcome order by count
(landing_outcome) desc
```

✓ 0.0s Python

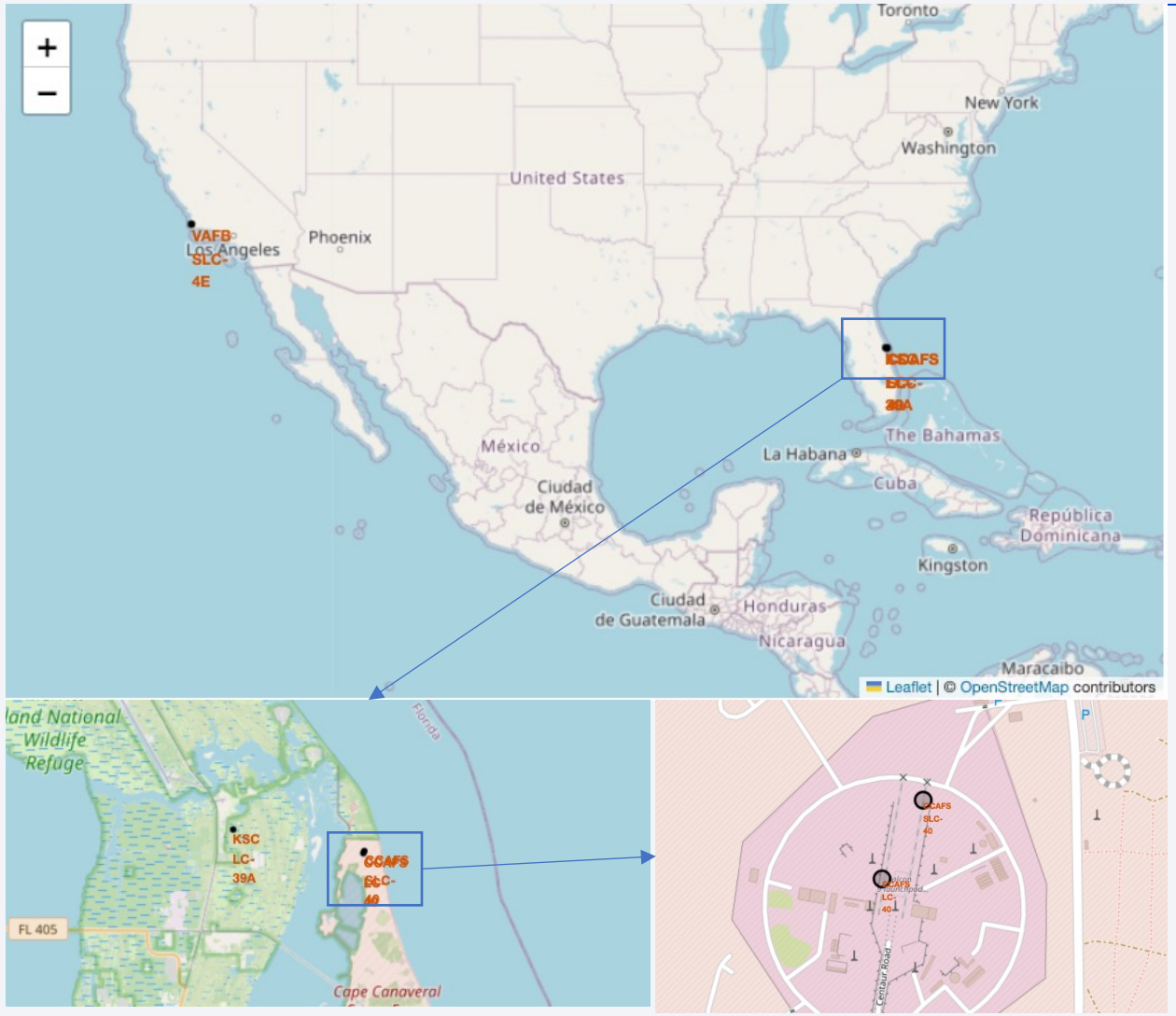
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count(landing_outcome)
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

This is the initial development period where we can see how the system is being tested for landing on both drone ships, on land, and with parachutes, as well as occasions when recovery was not attempted.


Section 3
Launch Sites
Proximities Analysis

Folium: Launch Sites locations around the world



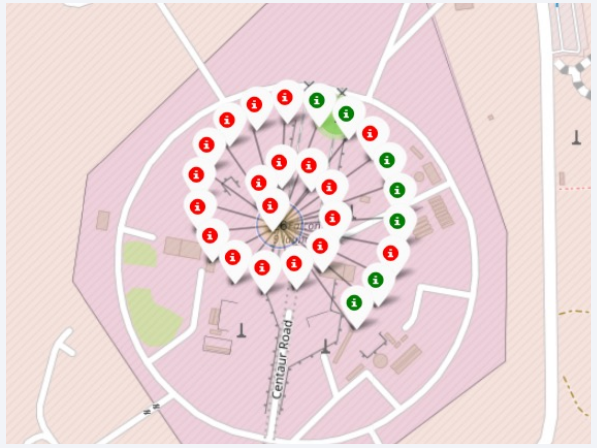
- We can see that the launch sites are located on the coast of the United States, one on the west coast and three on the east coast. This location is suitable in terms of safety for the launches. Additionally, they try to be positioned as close to the equator as possible, an optimal location for launch efficiency.

Folium :Success/Failed Launchs for Each Site

VAFB SLC-4E

KSC LC-39A

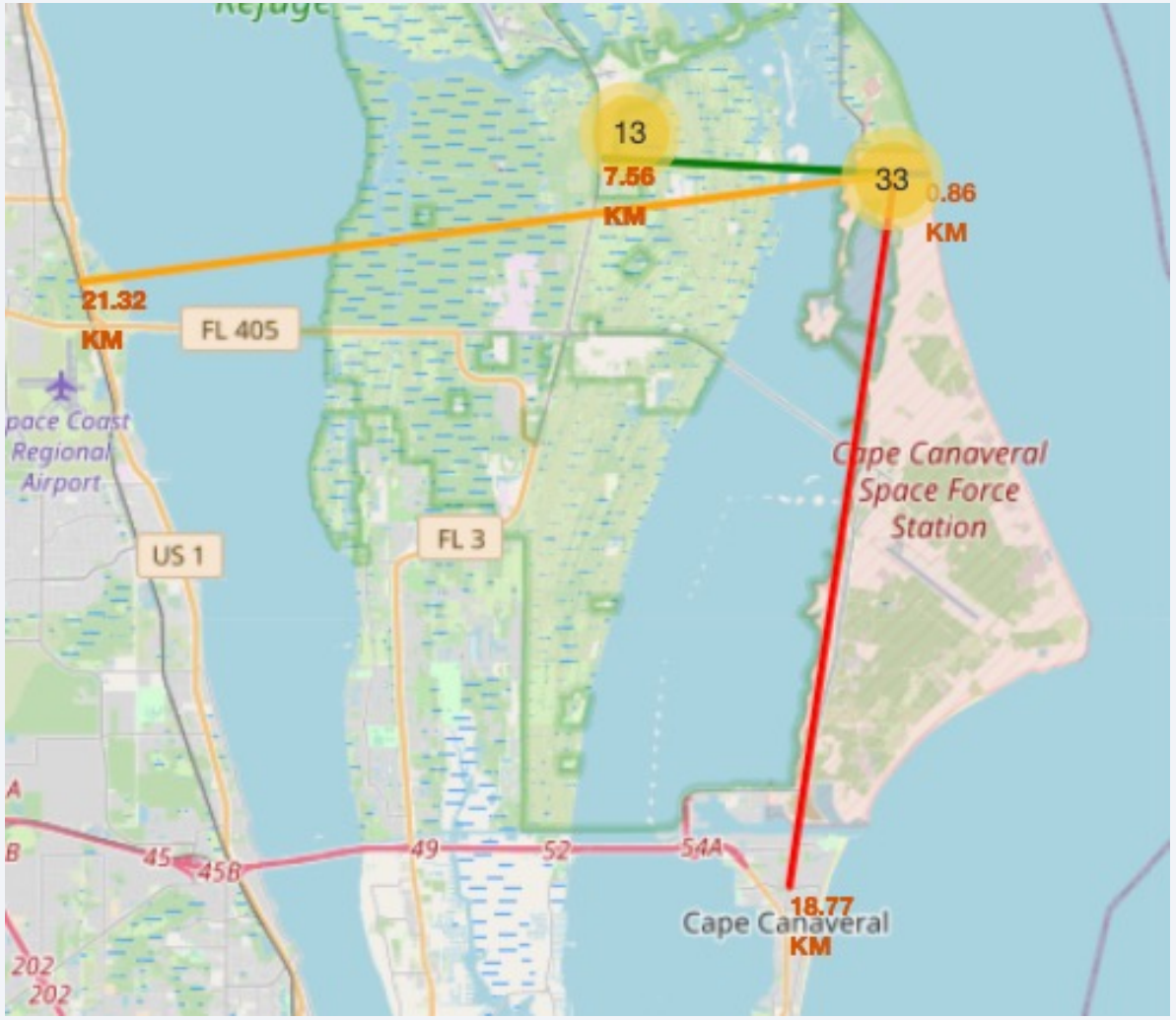
CCAFS LC-40



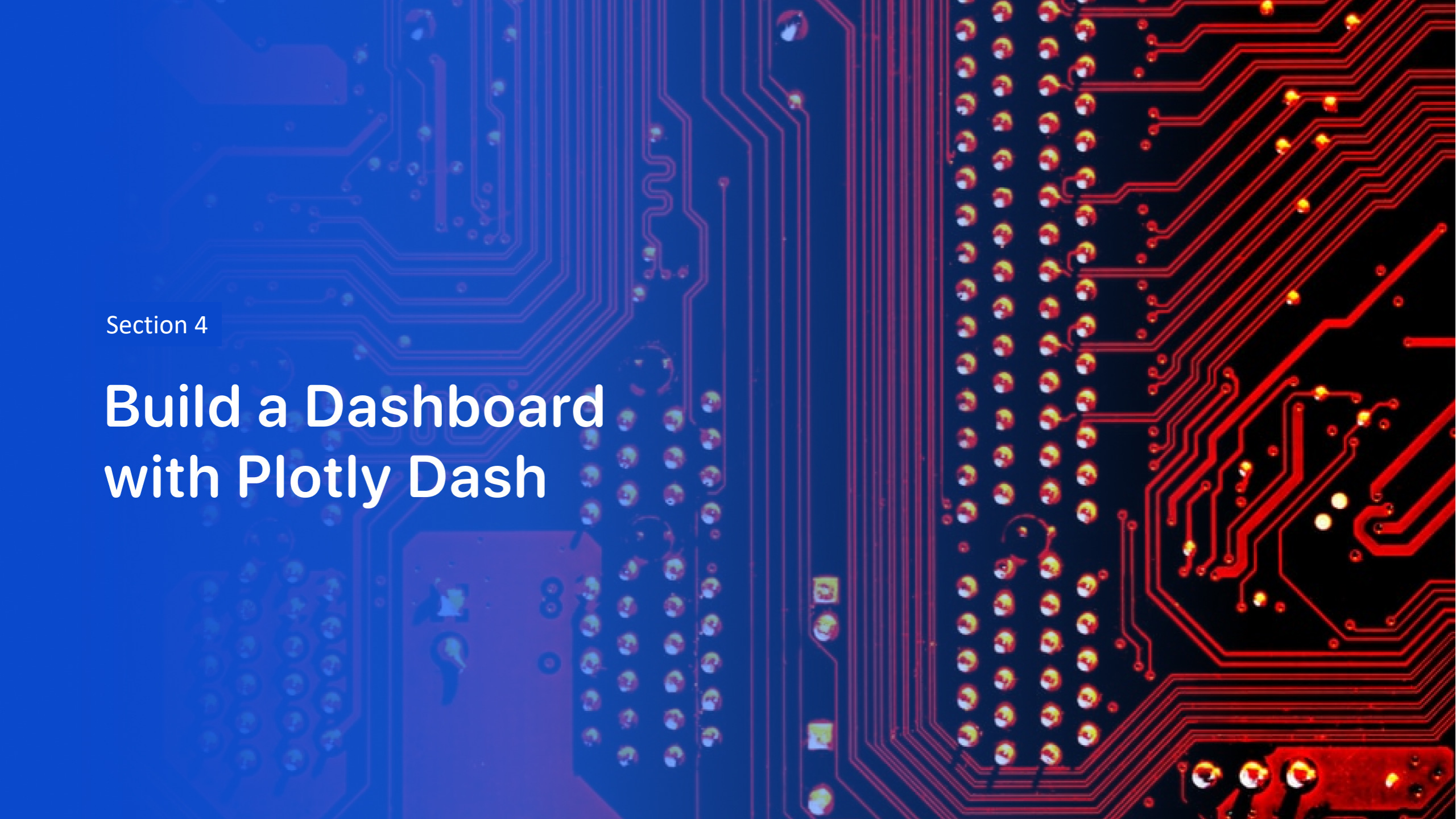
CCAFS LC-40

- Clusters of markers have been deployed around the launch sites.
- The color code is: green for success, red for failure.
- The launch site with more launches is CCAFS LC-40

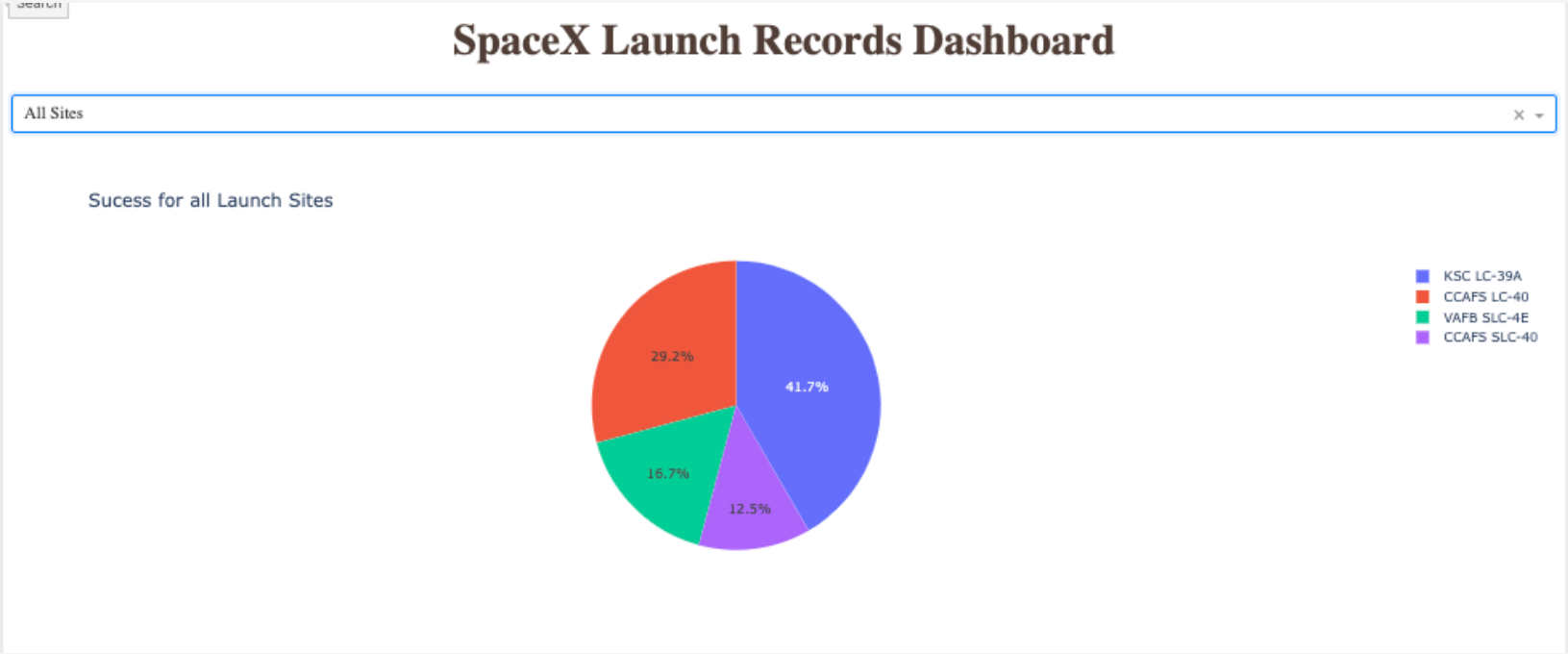
Distances Between Launchsite CCAFS LC-40 to its proximities



- For safety reasons, distances to the nearest key locations to the launch site CCAFS LC-40 have been established:
  - Nearest coast (blue) 0.86 km
  - Nearest city (red) 18.77km
  - Nearest railway (green) 7.56 km
  - Nearest highway (orange) 21.32 km

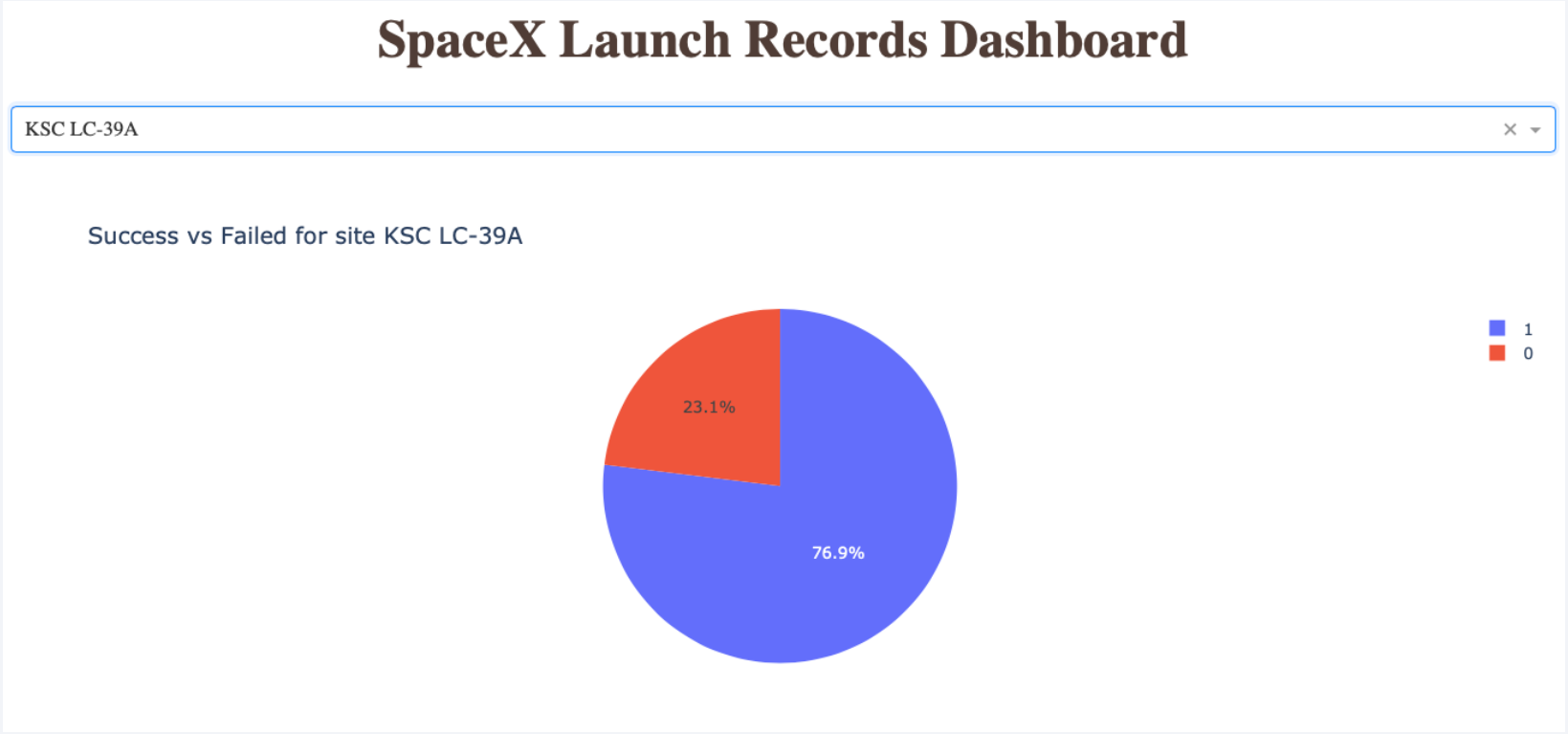

Section 4
Build a Dashboard with Plotly Dash

Successful rate for all Launch Sites



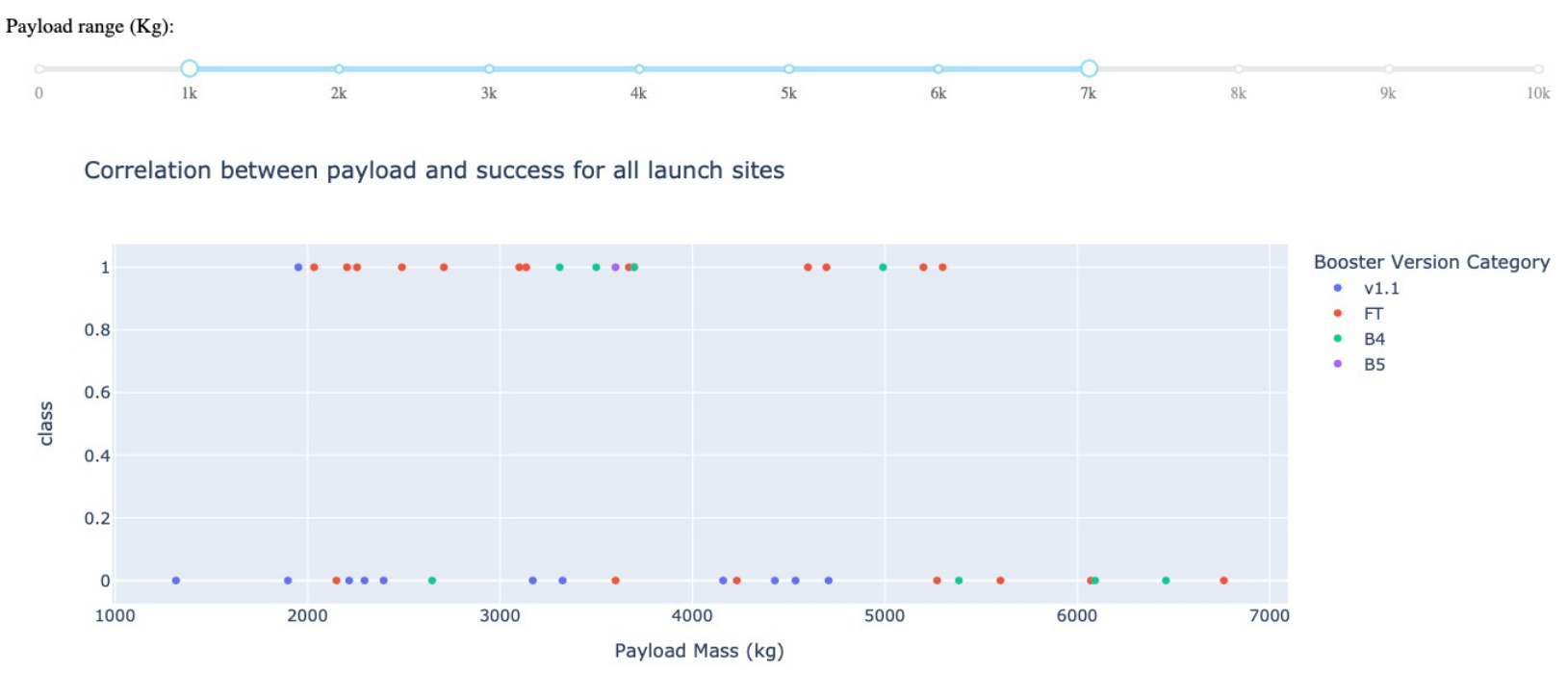
- The success rates of all launch sites are shown.
  - KSC LC-39A has most of the successful landing with 41.7%
  - CCAFS SLC-40 has the worst with 12.5 %

Success for the best launch site KSC LC-39A



The best success rate is complish by launches from KSC LC-39A with 76.9% success with only 23.1% failure.

<Correlation between payload and success>



- In the chart, we can see, for all launch sites, the successful and failed launches with payloads between 1,000 and 7,000 kg (1 = success, 0 = fail). The chart is categorized by the booster version.
- Most fails are from F9 v1.1 launches

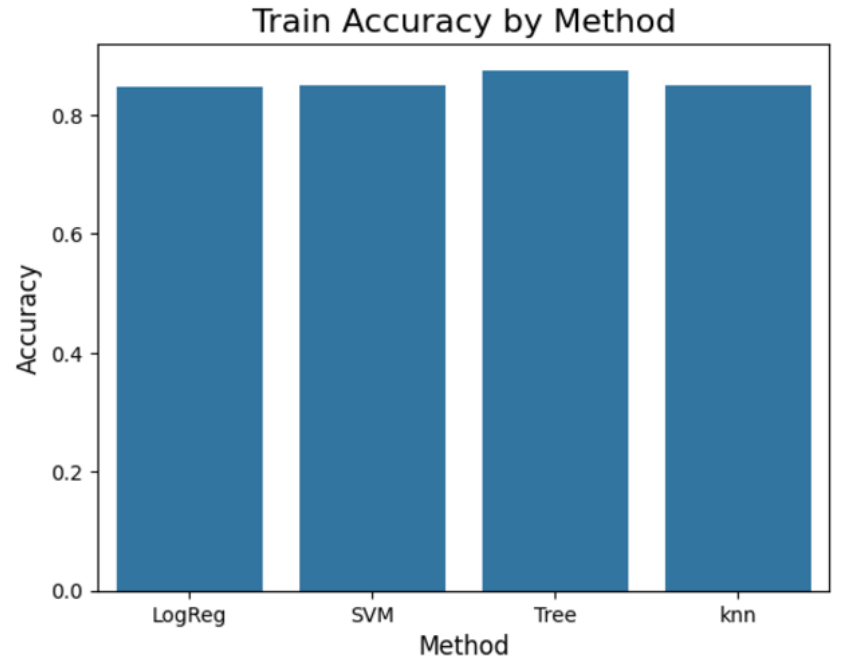
Section 5

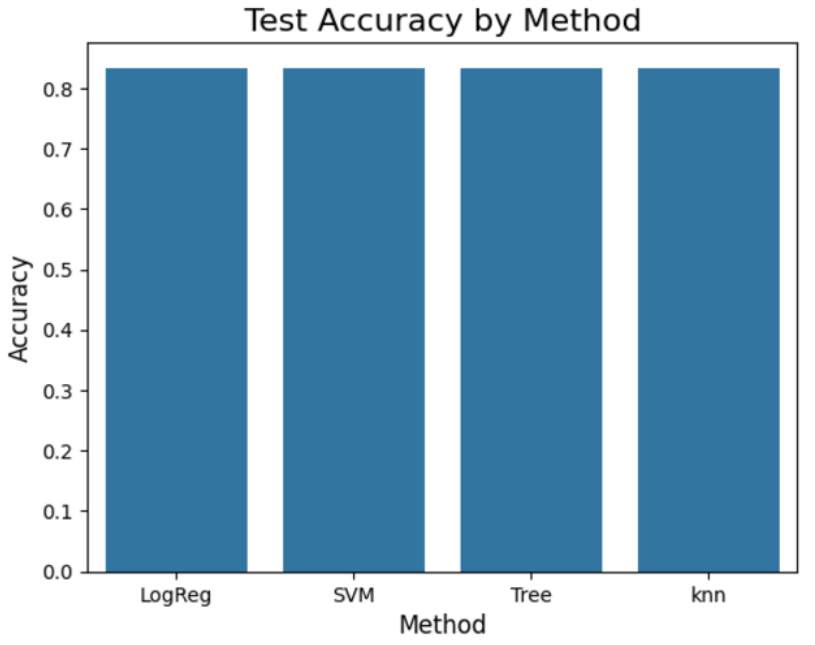
Predictive Analysis (Classification)

Classification Accuracy

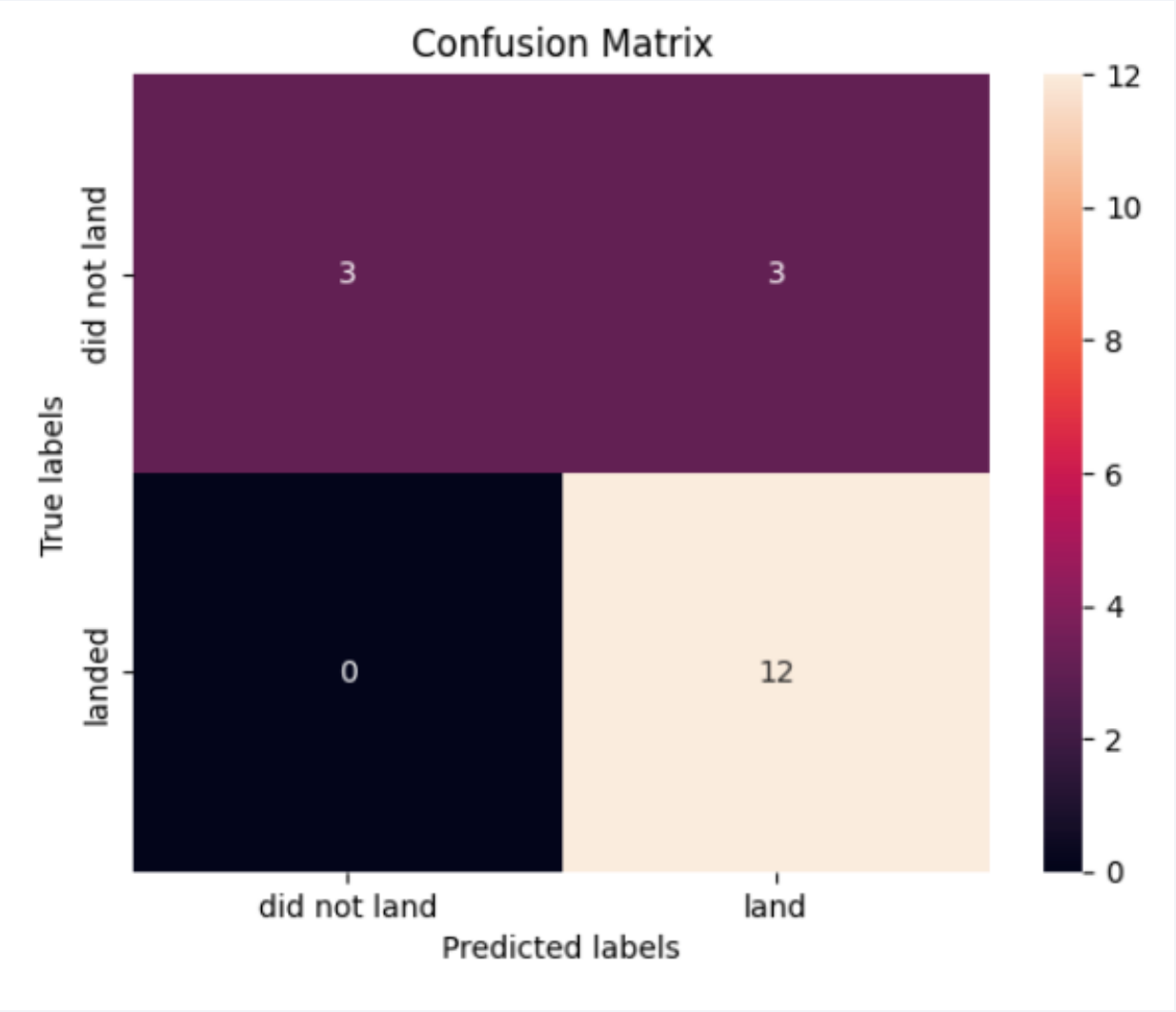
	Method	train accuracy	test accuracy
0	LogReg	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Tree	0.875000	0.833333
3	knn	0.848214	0.833333

- All methods tested achieved the same test accuracy, although Decission tree has a slight advantange in the training set.





Confusion Matrix



- Here we have the confusion matrix of the Decision Tree.
- The algorithm is able to correctly predict all successful landings but misclassifies failed landings, showing a 50% error rate. As we will see in the conclusions, this is most likely due to the characteristics of the dataset.

Conclusions

- We conducted this study to estimate the possibilities a company has, in terms of costs, to bid against SpaceX for a rocket launch.

- Despite an initial phase with a higher failure rate, the success rate of landings has increased over the course of the launches. Given the progression, we expect future launches to have an even higher success rate across all Launch Sites. This implies that they will likely be able to further reduce the price per launch.

- The methods tested for predicting the first-stage landing all yield similar results, successfully predicting successful landings but having a high error rate for failed ones. This may be due to two factors:
    1. The success rate is variable over time and currently evolving, which causes the data to be somewhat biased toward success.
    2. Only a small set of features has been considered out of all those that may truly affect the launch.

- SpaceX has not only demonstrated the viability of its proposal but has also gained experience over the years, positioning itself, thanks to its new approach, as the leading provider of space launches.

- **Our Recommendation:**
    - If any company wants to compete with SpaceX, the path to forward is to develop a first-stage landing system to reduce the launch cost, otherwise, under the current technologic developtment it will be impossible .

Thank you!