IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yoi Chang Calderon
10-08-2023

Outline

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

Executive Summary

- In this work Data Science Methodologies were used in order to implement Machine Learning Algorithms to the SpaceX Falcon-9 rocket data set with the purpose of predicting if the first stage of this rocket will land successfully.

- The Machine Learning models implemented achieved a mean accuracy of 0.84 over all the data set, whereas the mean Machine Learning models scores on test data obtained was 0.944.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone.
- Perhaps the most successful is SpaceX due this company can make rocket launches at relatively inexpensive prices.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; while other providers cost upwards of 165 million dollars each.
- Much of the savings is because SpaceX can reuse the first stage of its Falcon 9 rocket.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch

Section 1

Methodology

Methodology

Executive Summary

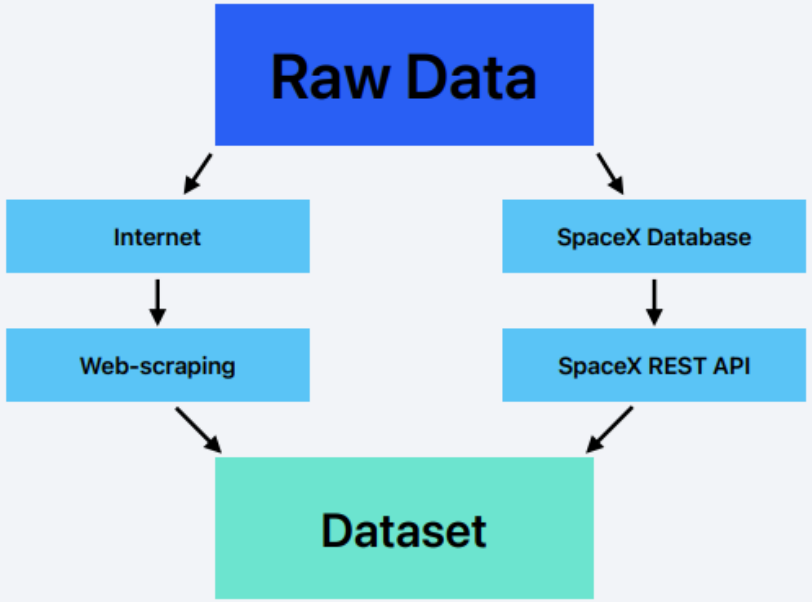
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The raw data needed for the project is available either on the web or in the SpaceX databases.

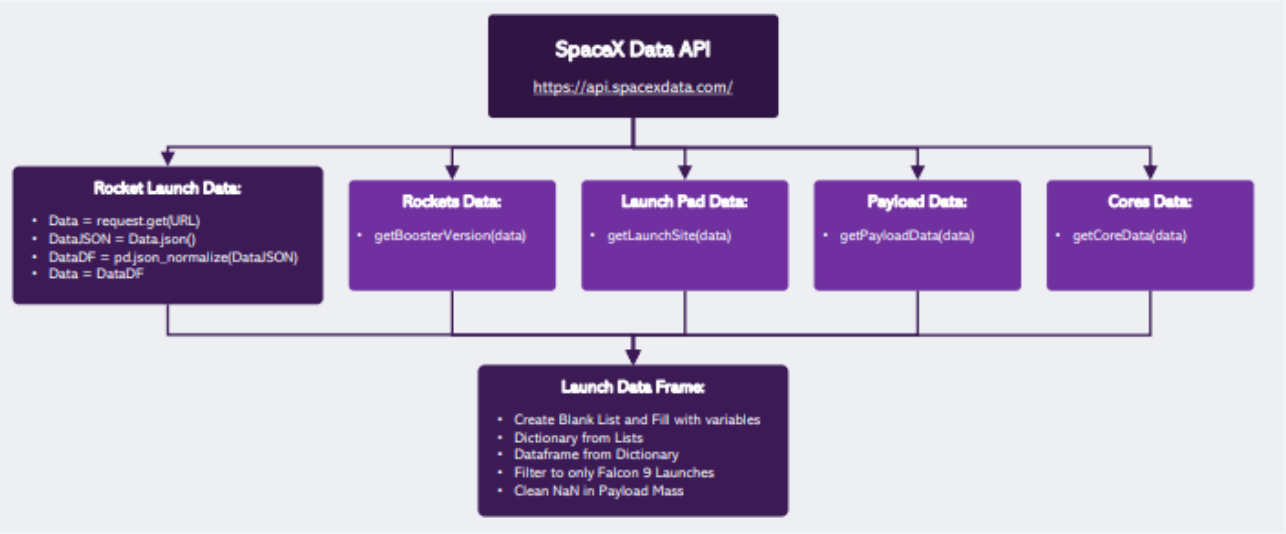
- So basically there are two ways of getting the data for this project:

    1. By making web-scraping in the related Wiki pages
    2. Using the SpaceX REST API



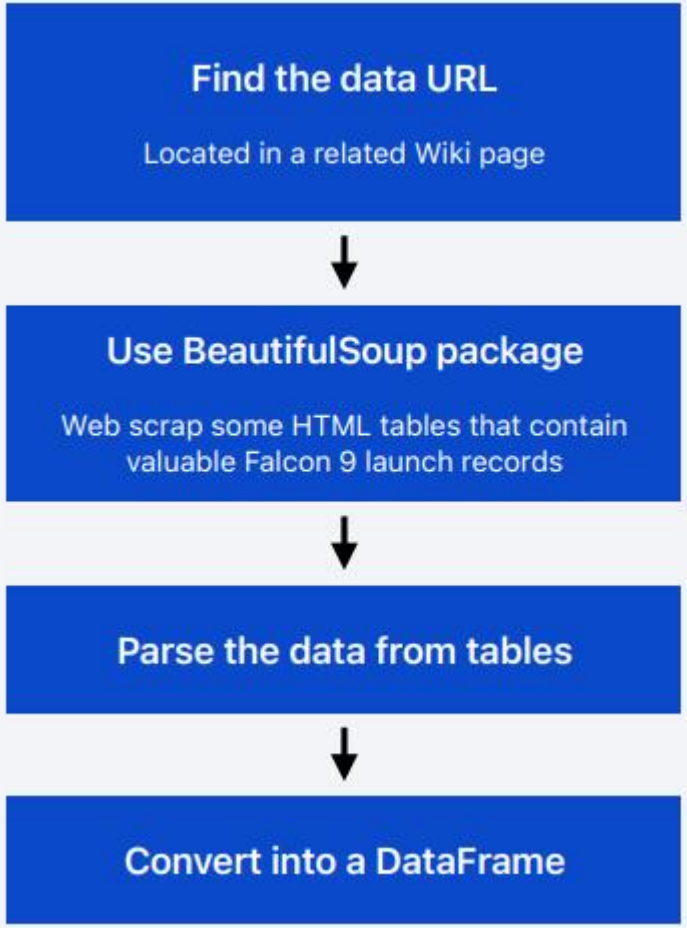
Data Collection – SpaceX API

- To use the SpaceX API effectively first we need to find the API endpoints or URL and then we can use the request libraries to obtain the launch data. This result can be viewed by calling the .json() method. Our response will be in the form of a JSON, specifically a list of JSON objects. To convert this JSON to a dataframe, we can use the json_normalize function.
- GitHub Link for the Jupyter Notebook: Lab 1: Collecting the data



Data Collection - Scraping

- • Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. We will be using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records. Then we'll need to parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis.
- GitHub Link for the Jupyter Notebook: Lab 1: Collecting the data



Data Wrangling

- Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

- In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- GitHub Link: **Lab 2: Data wrangling**

EDA with Data Visualization

- The following charts were plotted (see Section 2) in order to have an overview of the relation between the features used and the success rate of the different launches:
    - Relationship between Flight Number and Launch Site
    - Relationship between Payload and Launch Site
    - Relationship between success rate of each orbit type
    - Relationship between Flight Number and Orbit type
    - Relationship between Payload and Orbit type
    - Launch success yearly tren

GitHub Link: Notebook for Peer Assignment

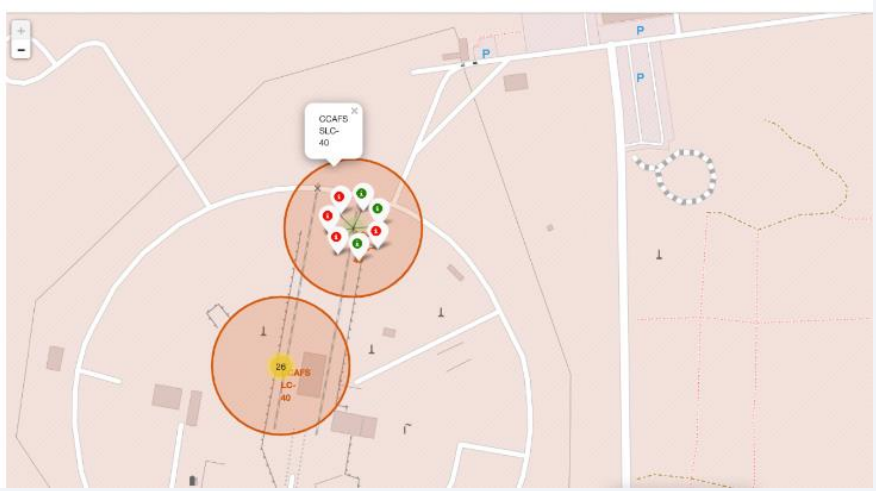
EDA with SQL

- Some SQL queries were performed in order to understand the SpaceX dataset:
  - Display the names of 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 the boosters launched by NASA (CRS)
  - Display the average payload mass carried by the 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 total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass.
  - Rank the count of successful landing outcomes between 2010 an 2017

GitHub Link: SQL Notebook

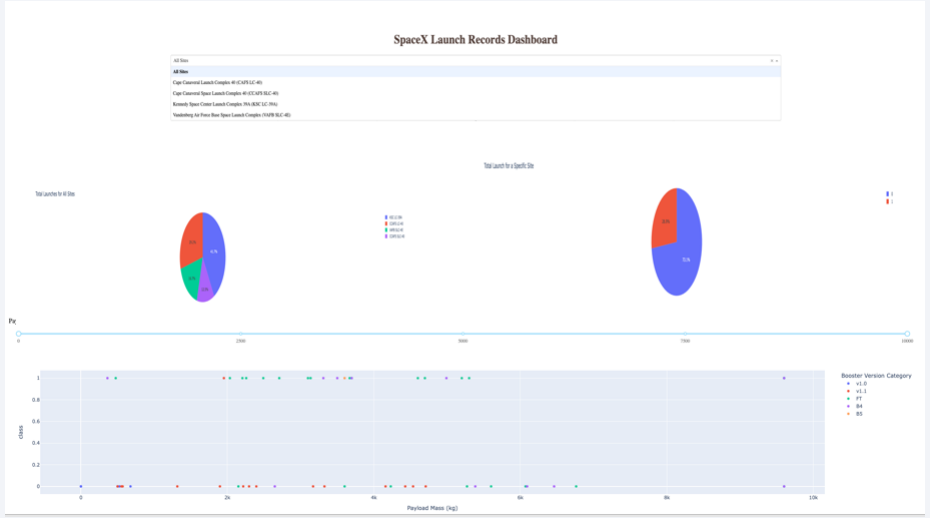
Build an Interactive Map with Folium

- • The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- • The tasks made for building the interactive map in order to find some geographical patterns about launch sites were:
  - • Mark all launch sites on the map
  - • Mark the success/failed launches for each site on the map
  - • Calculate the distances between a launch site to its proximities.
- GitHub Link: [Locations Analysis with Folium]



Build a Dashboard with Plotly Dash

- The following visuals that were used in the dashboard were intended to give some insights of the site with largest successful launches and highest launch success rate as well for the different F9 boosters and the payloads ranges with highest and lowest success rate:
  - Launch site drop-down input component
  - Total success launches by site pie chart
  - Range slider to select payload



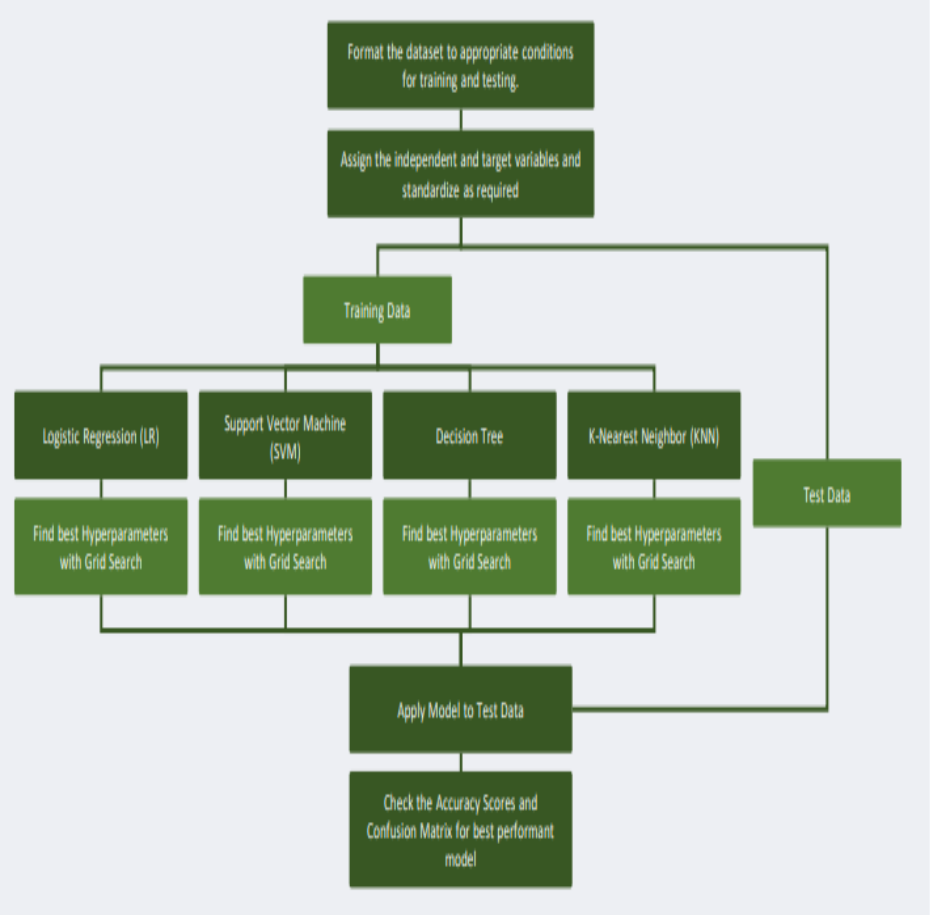
  - Scatter plot of Payload Mass (kg) vs Success rate per booster version

Predictive Analysis (Classification)

• Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.

• This will include:

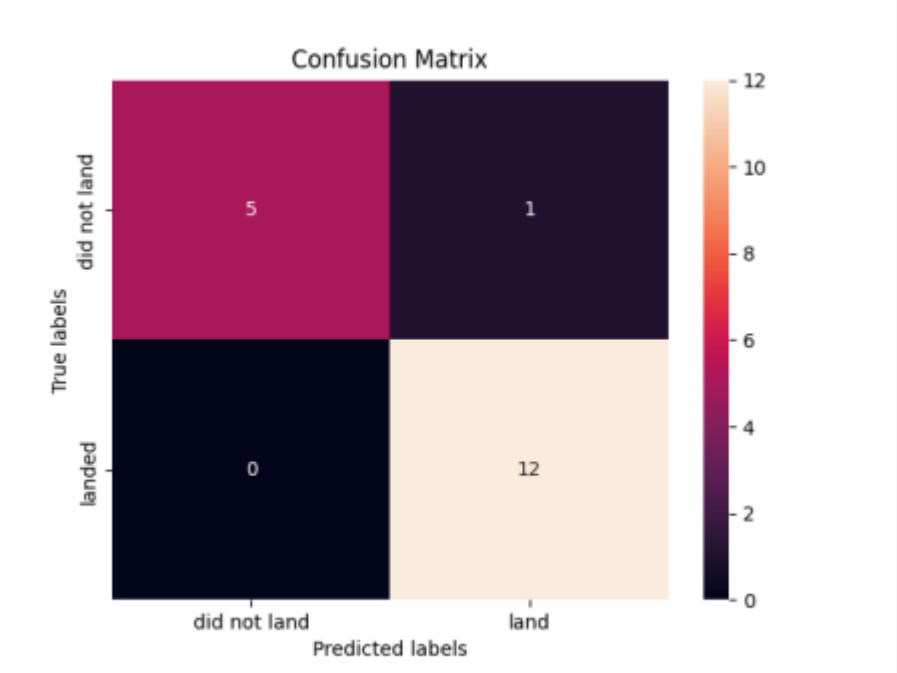
- Preprocessing, allowing us to standardize our data, and Train_test_split, allowing us to split our data into training and testing data.
- Train the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
- Using the best hyperparameter values, we will determine the model with the best accuracy using the training data.
- Test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
- Finally, output the confusion matrix to evaluate the model efficiency.



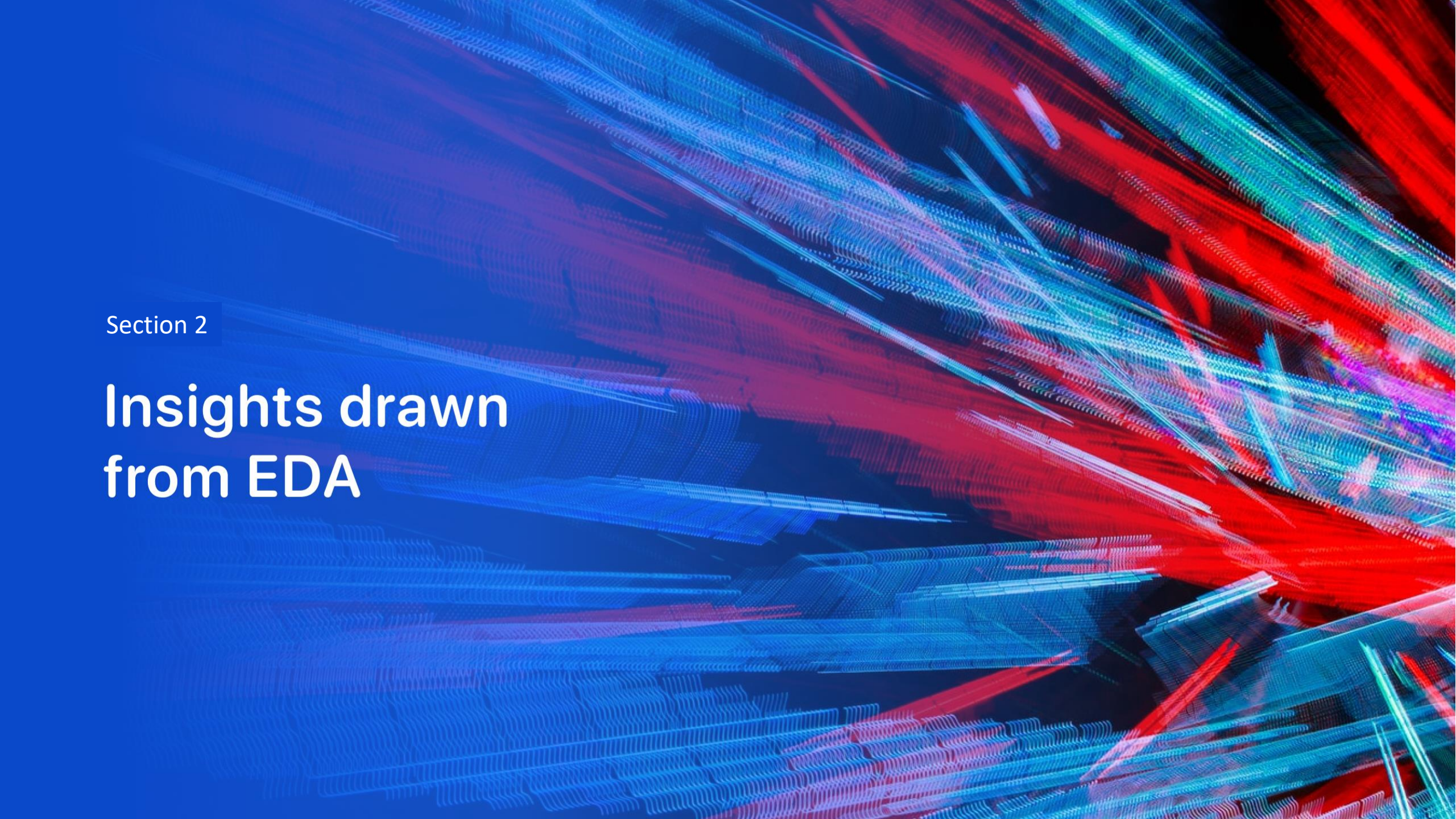
GitHub Link: [Landing Prediction]

Results

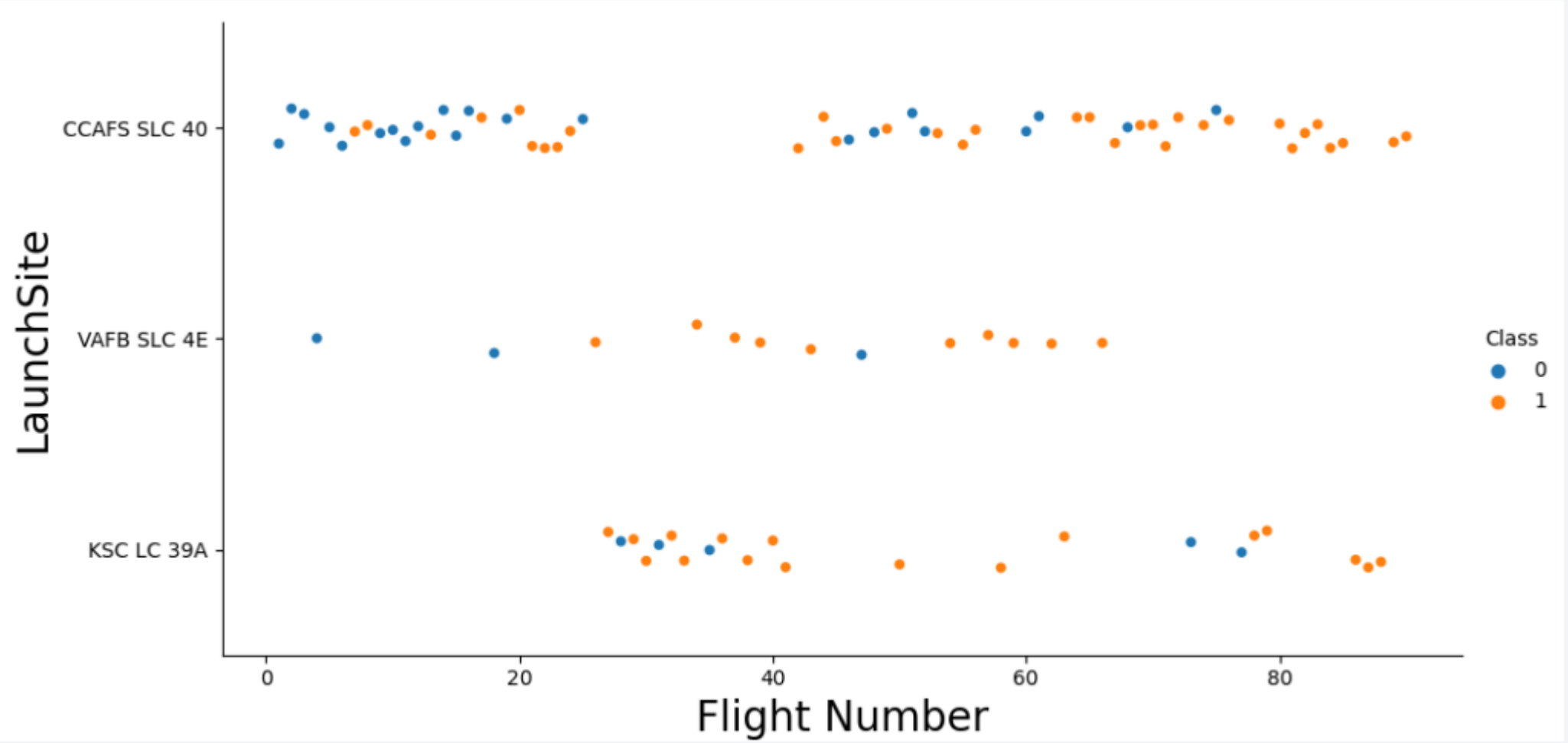
- The accuracy of each model applied to the data obtained was:
- Logistic Regression: 0.82
- SVM: 0.82
- Decision Tree: 0.88
- KNN: 0.84
- While the R2 scored of all ML models applied was 0.94



GitHub Link: **Landing Prediction**

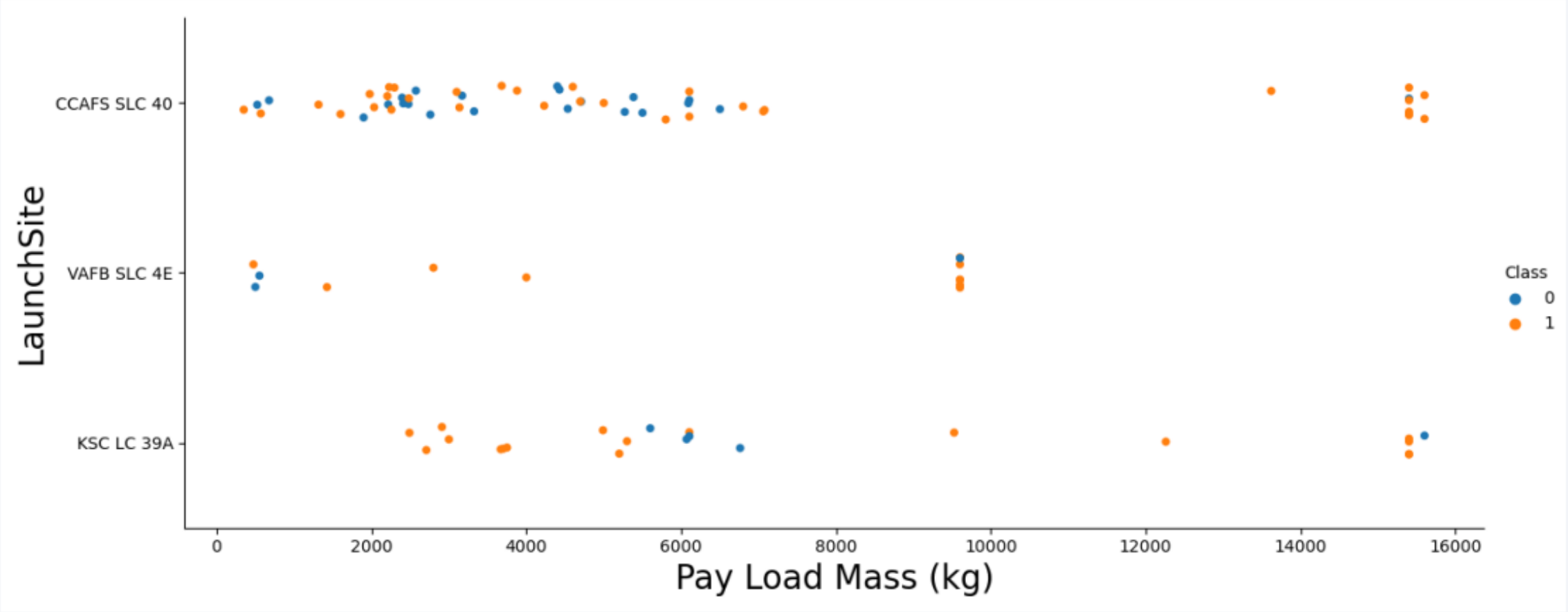

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



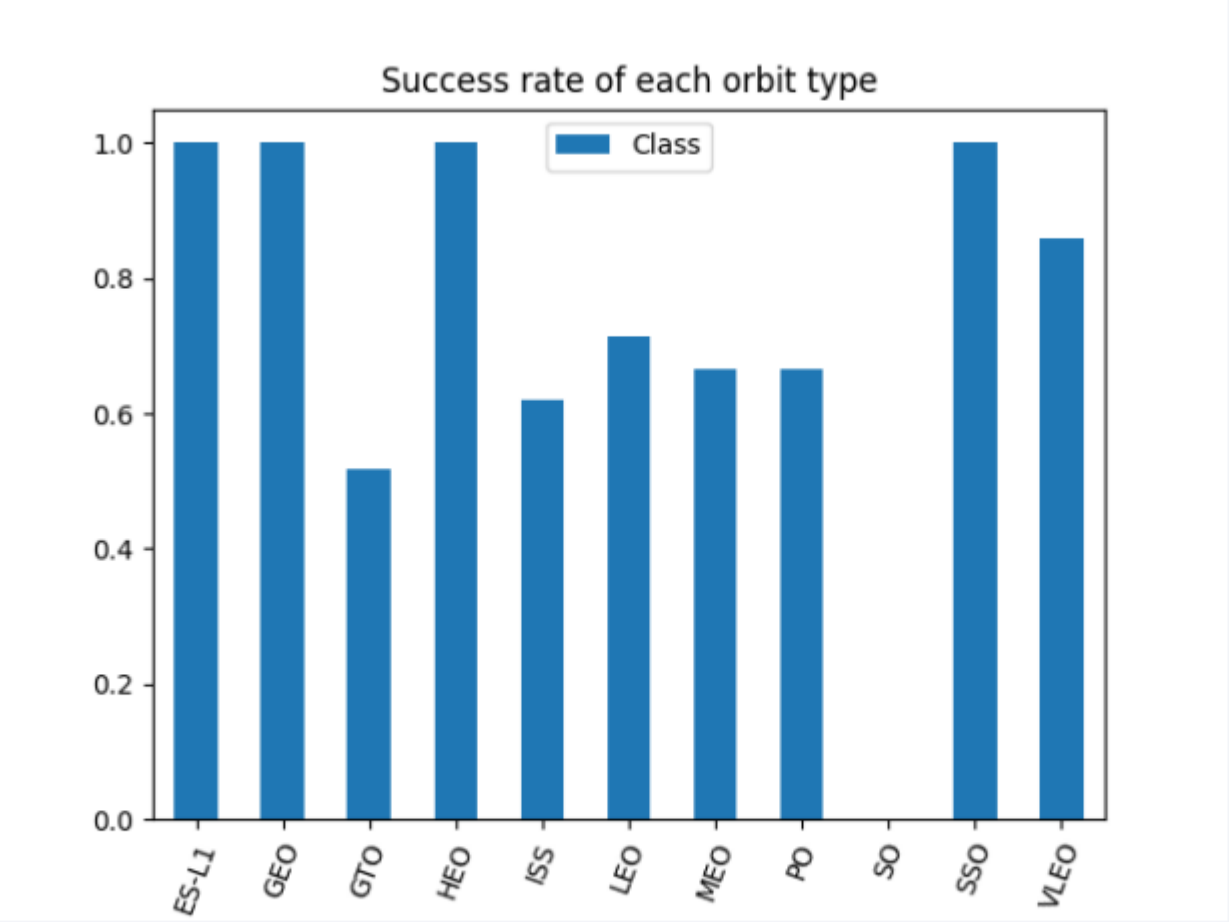
We can see that most of the successful launches (class = 1) were fater flight number 17.

Payload vs. Launch Site



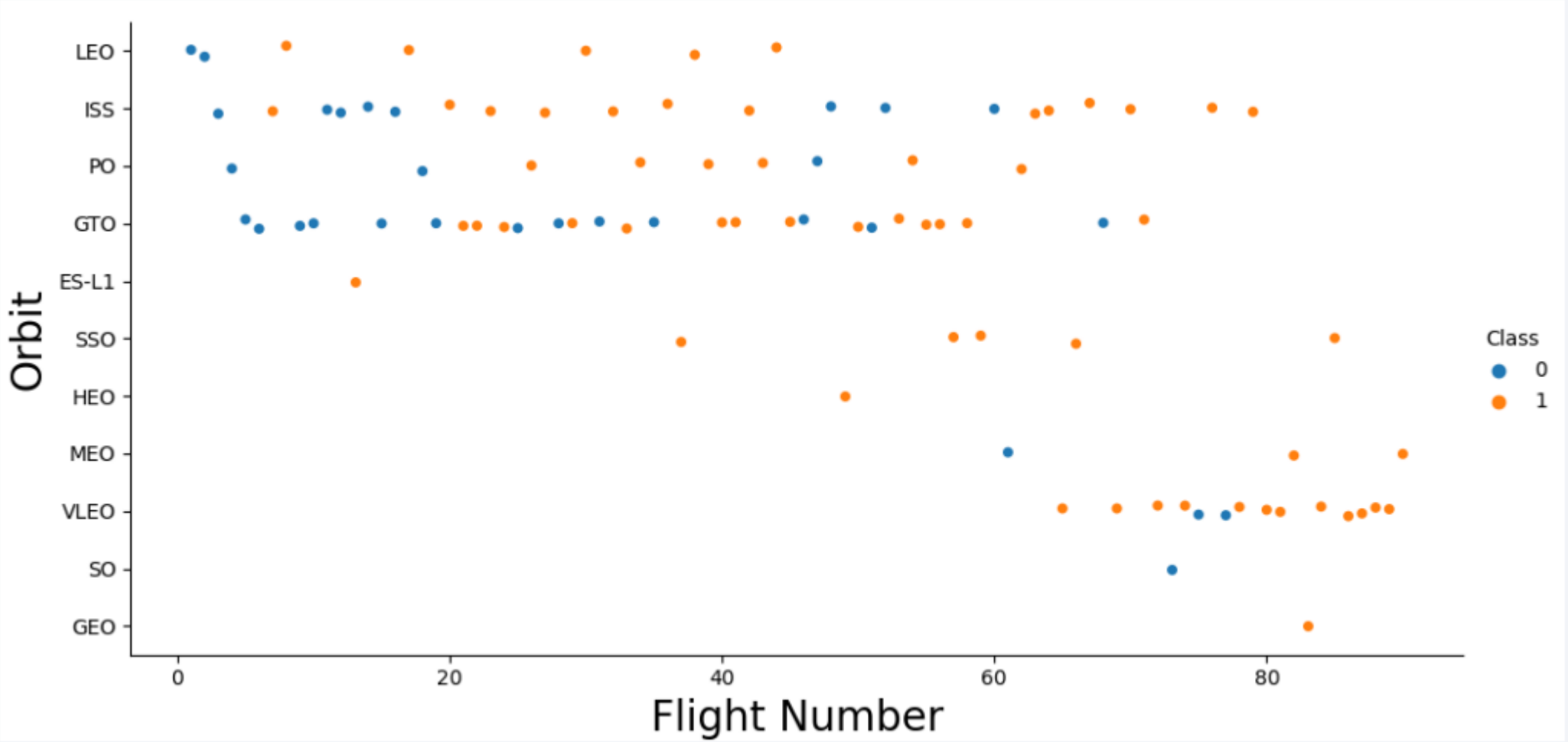
If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type



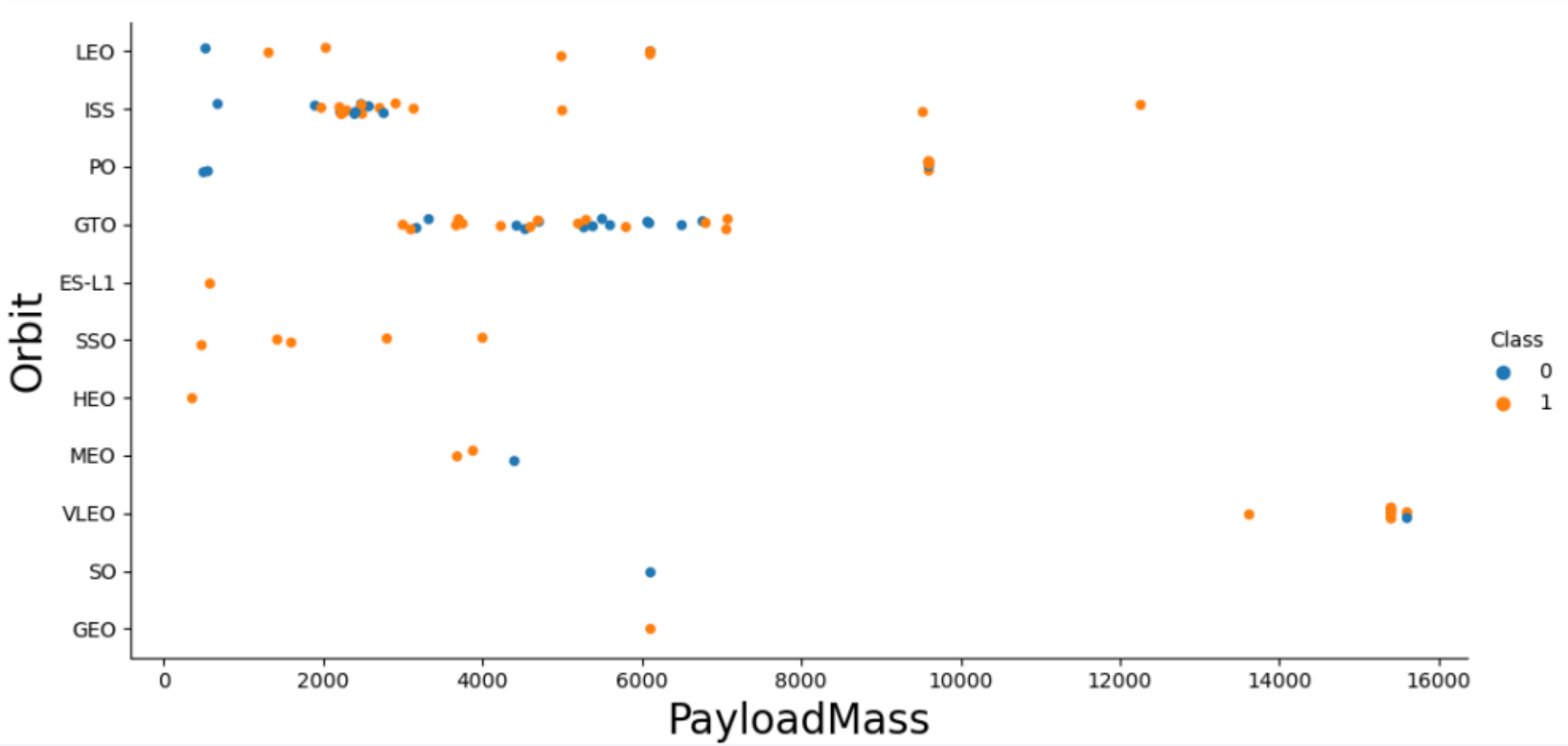
It is interesting how ES-L1, GEO, SSO and HEO launched orbits have 100% succes rate.

Flight Number vs. Orbit Type



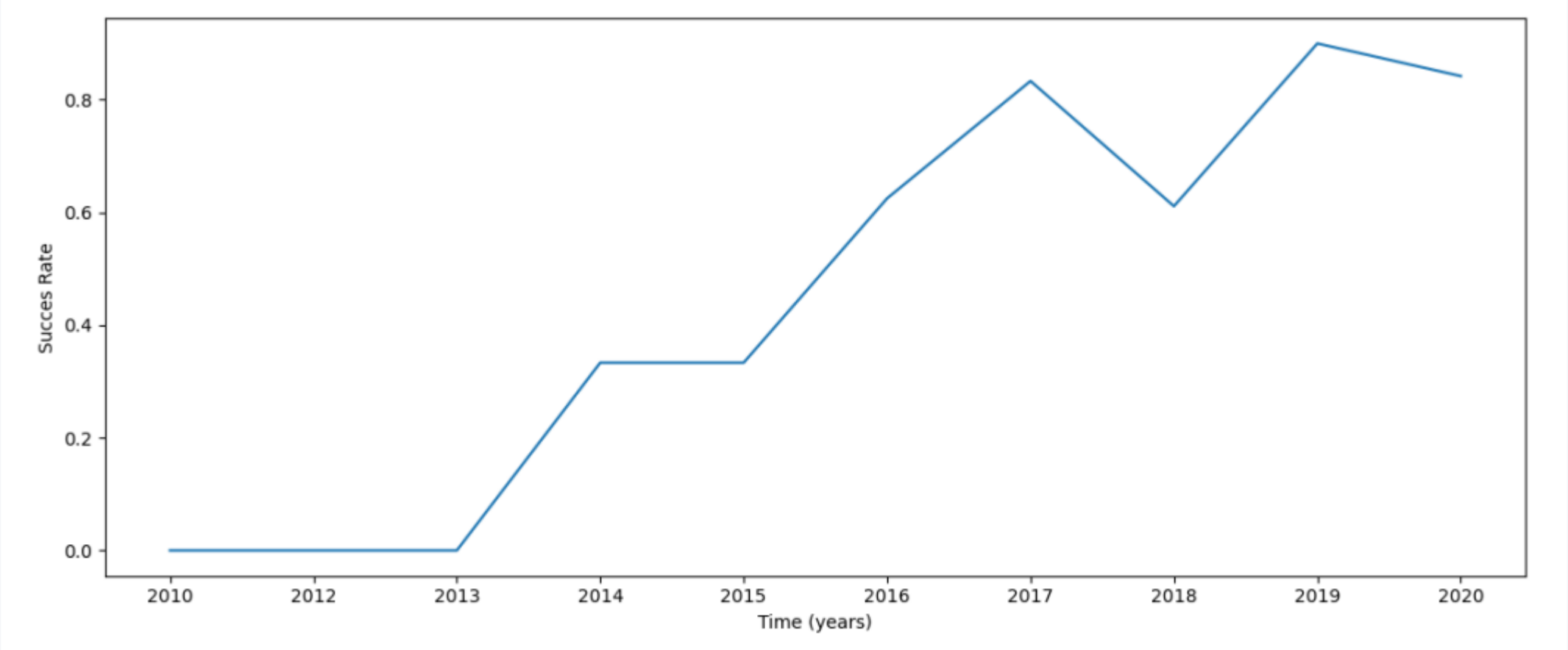
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



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

All Launch Site Names

```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

I used the DISTINCT function to find the different values in Launch_site column

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
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	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

As I don’t not know exactly how the Launch Site is called I used the LIKE function

Total Payload Mass

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Payload LIKE '%CRS%'
```

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

SUM(PAYLOAD_MASS__KG_)
111268.0

Same as previous task but now I add the SUM function to the query

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

We can see by this calculation that most of the Launches where kinda ”lightweight”

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
select Date
from SPACEXTBL
where Landing_Outcome = 'Success (ground pad)' LIMIT 1
```

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

Date
22/12/2015

Despite SpaceX started launching in 2010 they took 5 years to achive its first succesful landing in ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

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, PAYLOAD_MASS__KG_
from SPACEXTBL
where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select count(Landing_Outcome) as Successful, (select count(Landing_Outcome) from SPACEXTBL where Landing_Outcome like 'Failure%') as Failure
from SPACEXTBL
where Landing_Outcome like 'Success%'
```

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

Successful	Failure
61	10

There is a 6:1 relation between successful and unsuccessful launches.

Boosters Carried Maximum Payload

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

F9 B5 boosters were designed for heavy mass launches

2015 Launch Records

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

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

+ Code + Markdown

```
%%sql
select Date, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL
where Date like '%2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

Date	Landing_Outcome	Booster_Version	Launch_Site
01/10/2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14/04/2015	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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
select Landing_Outcome, Count (*) as number_of_launch_outcome
from SPACEXTBL
where Date between '04/06/2010' and '20/03/2017' group by Landing_Outcome
ORDER BY number_of_launch_outcome DESC
```

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

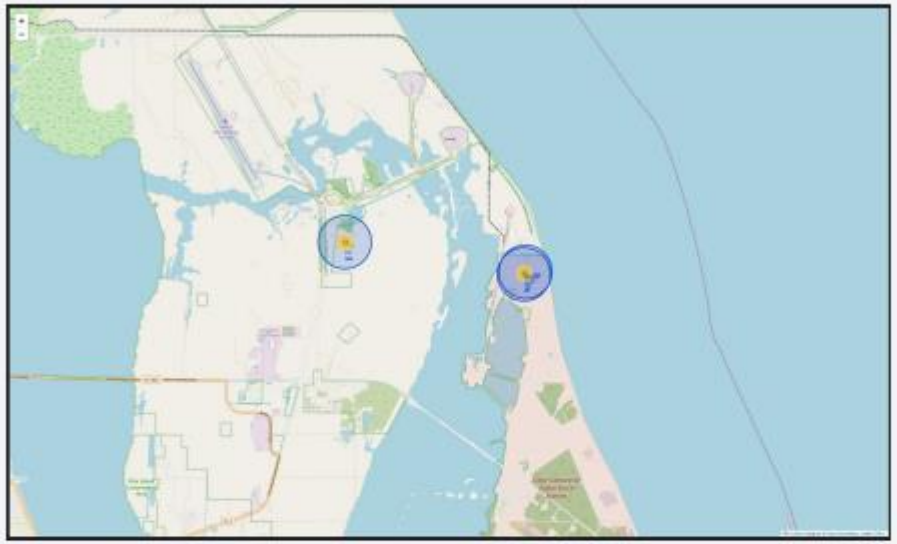
Landing_Outcome	number_of_launch_outcome
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Most of the ladings were successful

Section 3

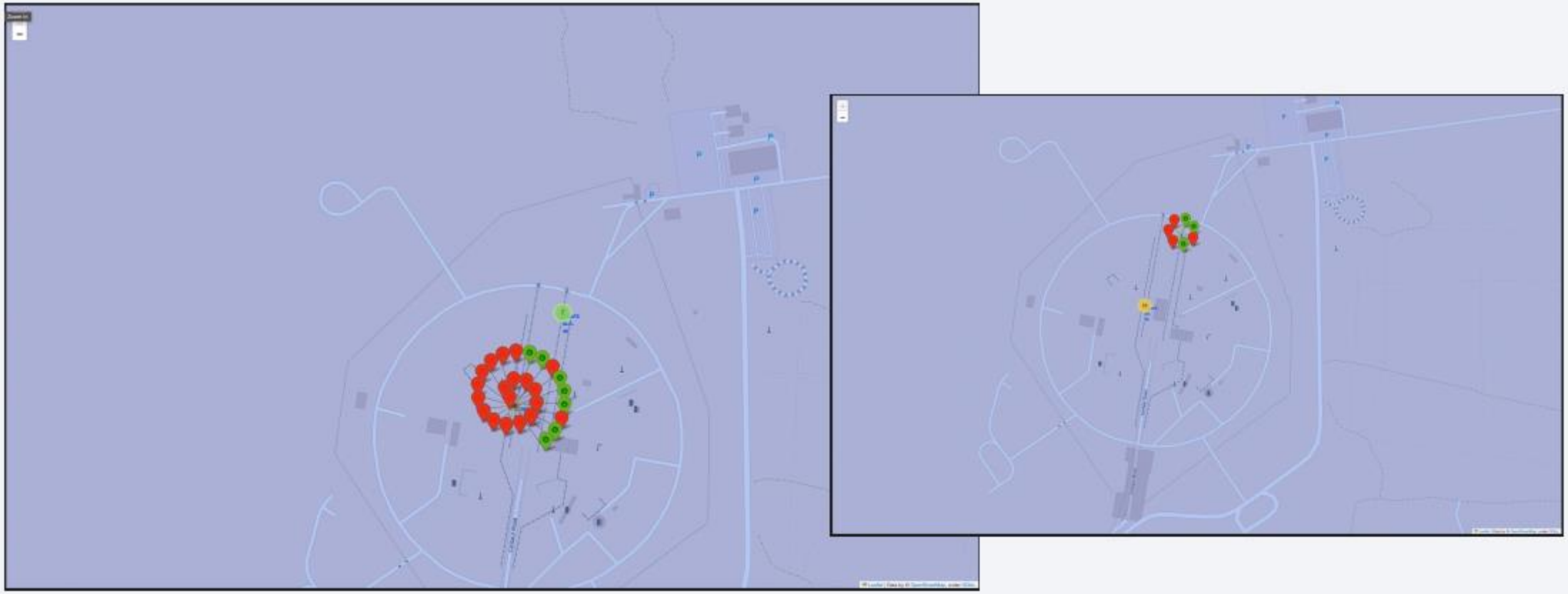
Launch Sites Proximities Analysis

Folium Map Screenshot 1



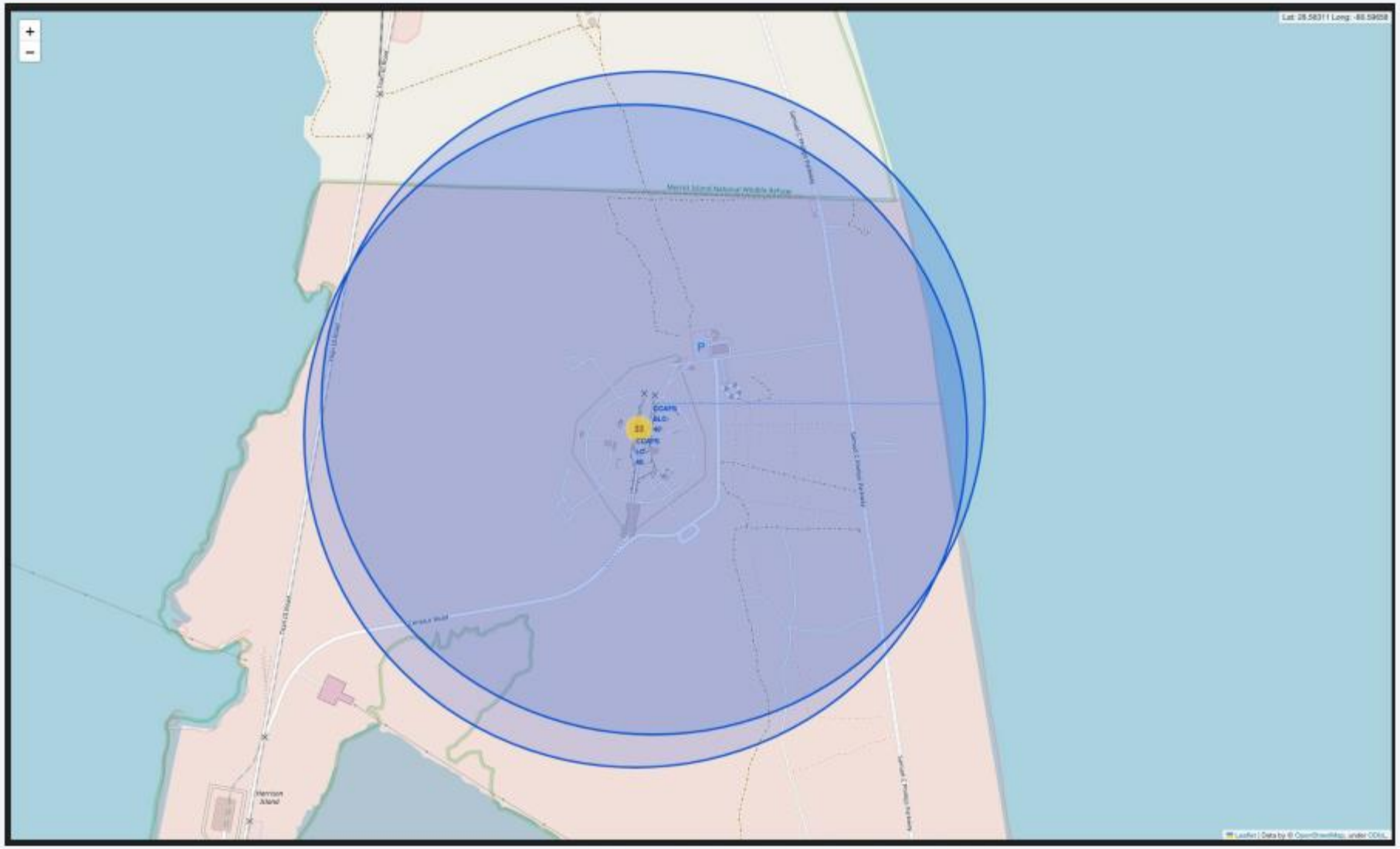
I created a map where the launch sites of the Falcon 9 rocket are marked

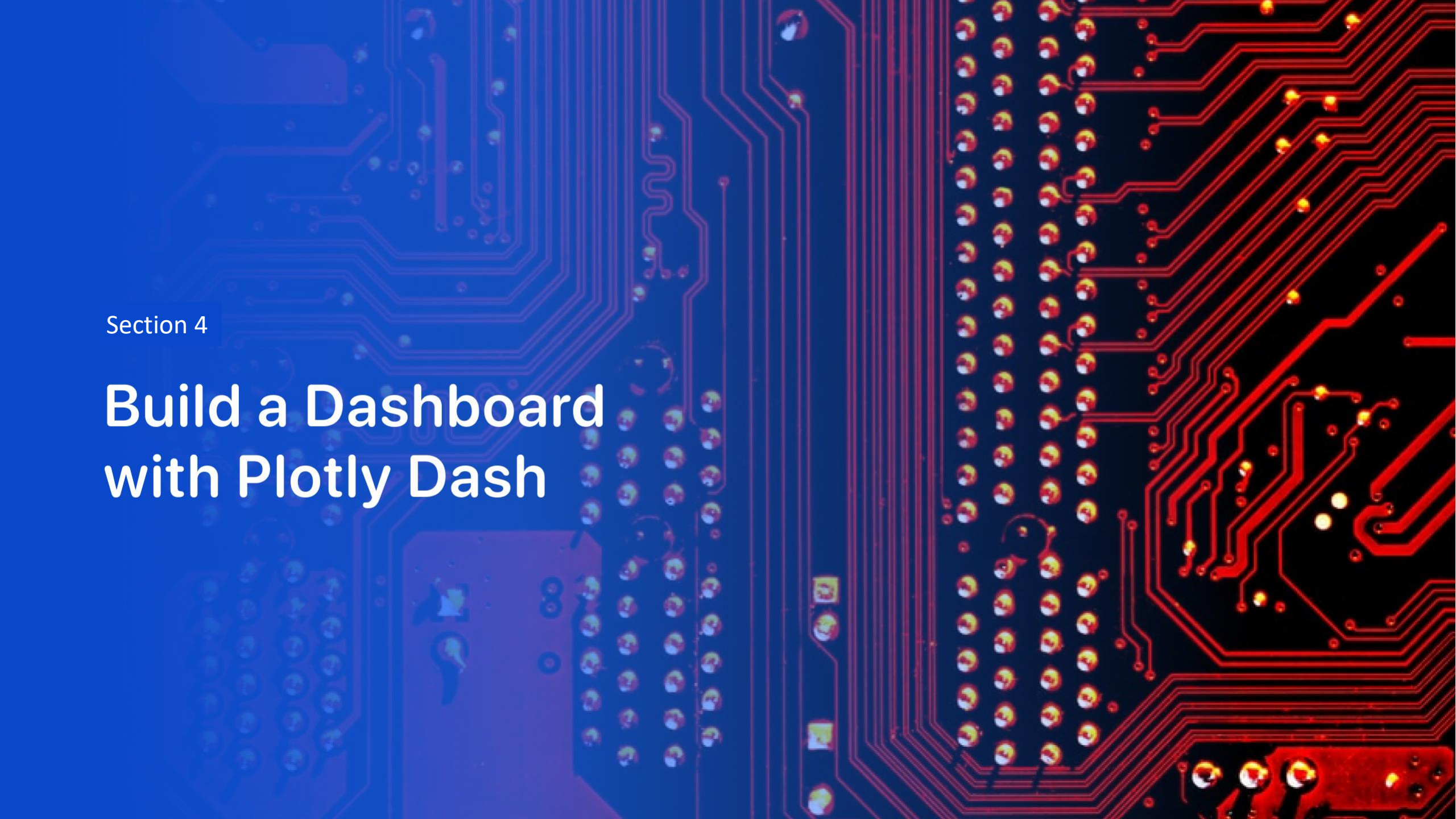
Markers of successful and failed launches at each site



CCAFS LC-40 and CCAFS SCL-40 luaunch sites at the left and right respectively

Distance between a launch site to its proximites

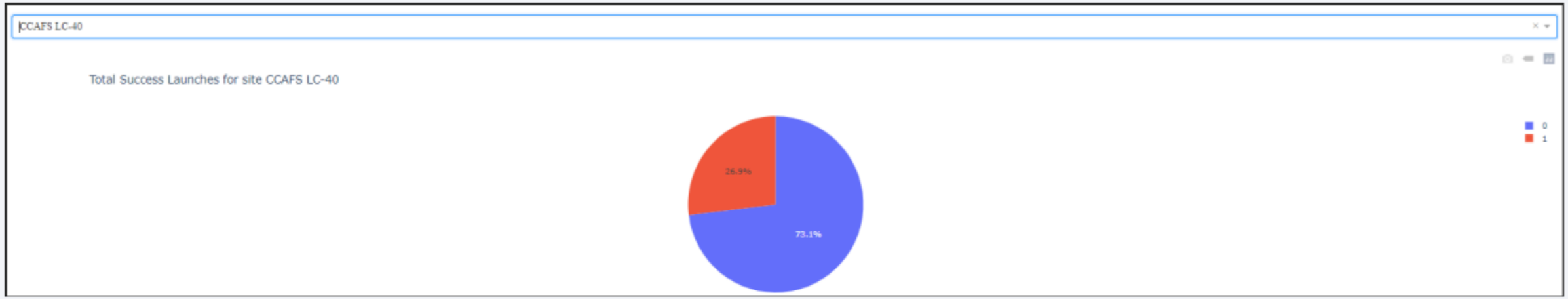



Section 4
Build a Dashboard with Plotly Dash

Total Launches for All Sites

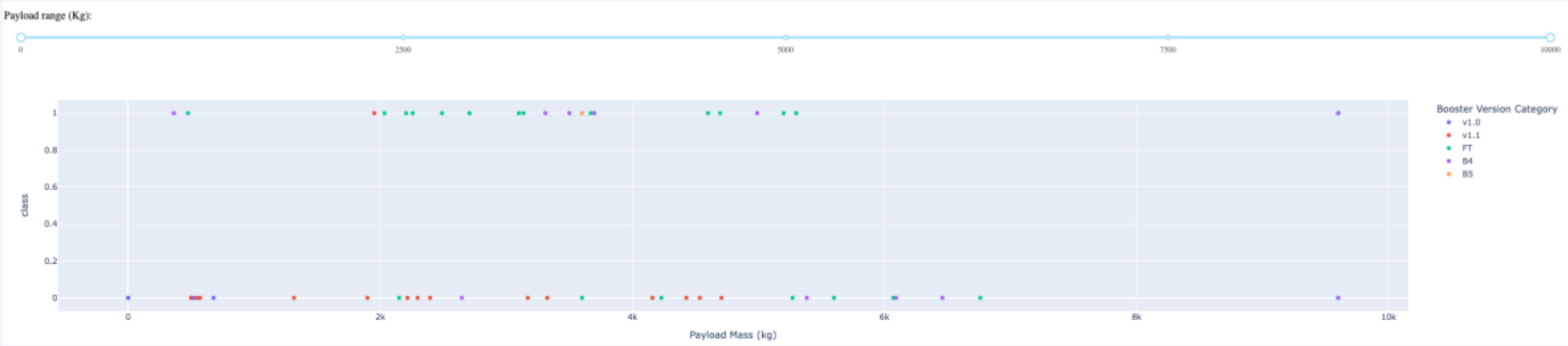
Wee can see that KSC LC-39A launch sited has most of the launches of the Falcon 9 Rocket

Total Success Launches for site CCAFS LC-40



We can see that this launch site has more than 70% success rate

Dashboard Screenshot 3



Payload Mass range between 2k and 4k has the most succes rate pero Booster

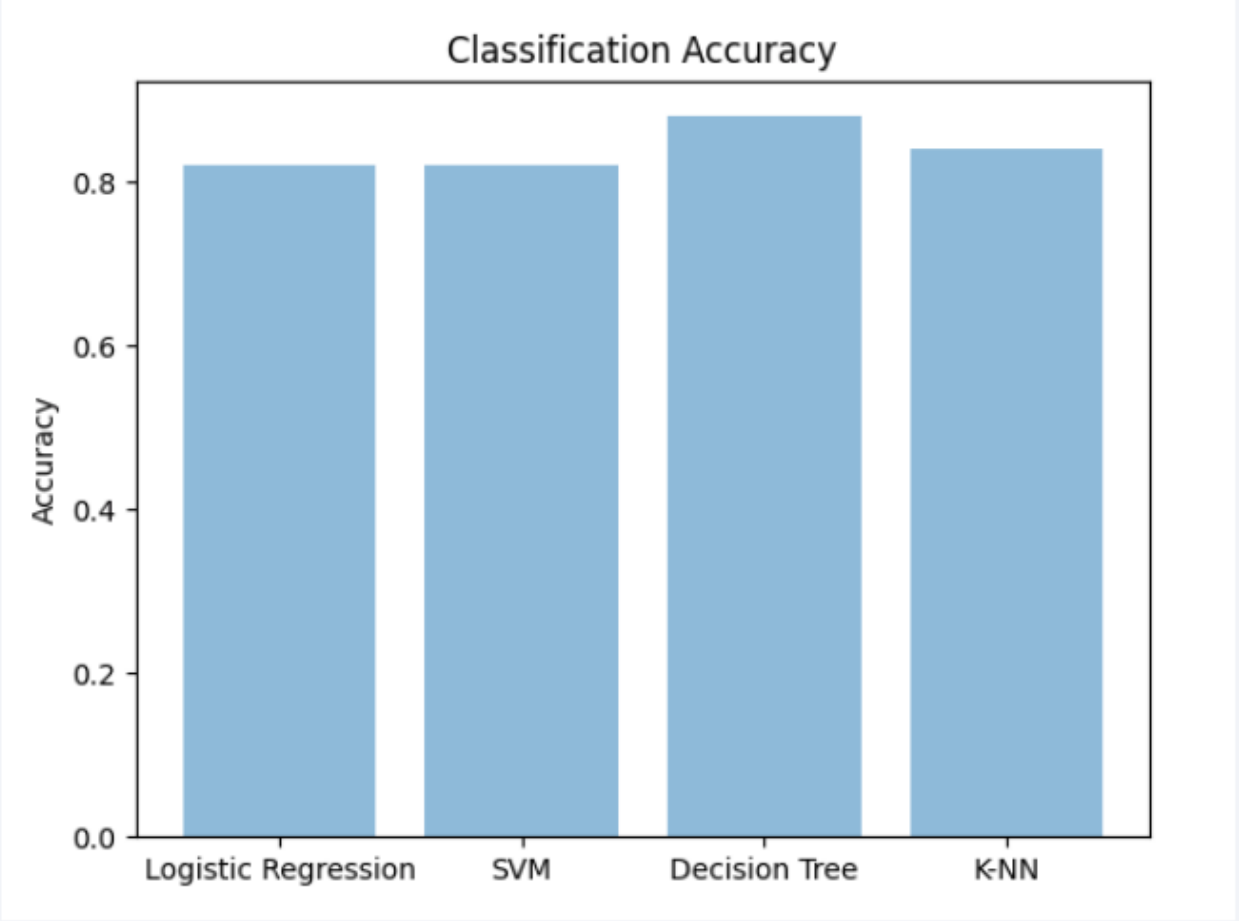
Section 5

Predictive Analysis (Classification)

Classification Accuracy

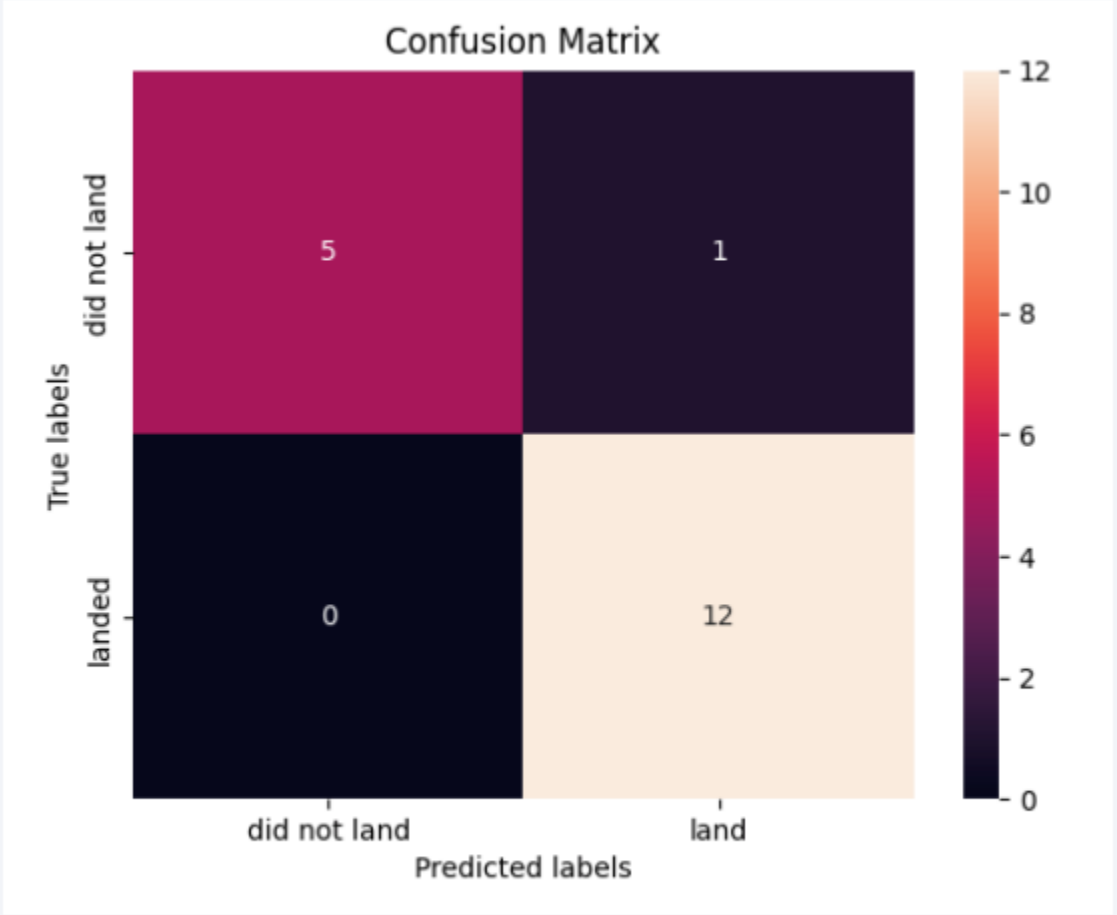
- The right bar chart shows the final accuracy of each classification model.

- We can see that despite there is not much difference between all the classification models, Decision Tree model is the one with the highest accuracy.



Confusion Matrix

- In the right panel we can see the confusion matrix of the decision tree model
- We can see that theres only one value that wasn't well predicted, and that's why this model has a R2 score of 0.94 for the predicted values.



Conclusions

- Despite all the different classification models have the same R2 score value (0.94) on test data, the bar chart in page 43 shows clearly that the decision tree model has a slightly better approach for the purpose of this project.
- Decision Tree model may have better accuracy due the behave of the data showed in section 2 and mainly because we have a lot of independent variables to work with (~80 features).
- Because of the number of features is relatively high, an entropy-approach like classification method is more suitable in order to model this data. That's why Decision Tree turned out to have the best accuracy

Appendix

- Data Collection
- Data Collection – SpaceX API
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Predictive Analysis (Classification)
- Build a Dashboard with Plotly Dash

Thank you!