IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diogo Pereira Da Silva
Dias Grilo
21/12/2023

Outline

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

Executive Summary

Summary of Methodologies

- **Data Collection**: Utilized API calls and web scraping techniques to gather relevant data on Falcon 9 rocket launches.
- **Data Cleaning and Wrangling**: Implemented data cleaning and wrangling methods to refine the dataset.
- **Exploratory Data Analysis (EDA)**: Conducted thorough EDA using various types of graphs.
- **Visualization with Plotly and Dash**: Used Plotly and Plotly Dash to create interactive charts, enhancing the interpretability and engagement of the data analysis results.
- **Predictive Modeling**: Applied several machine learning models, including logistic regression, decision trees, k-nearest neighbors, and support vector machines. This diverse approach allowed for a robust comparison of model performance.

Summary of Results

The predictive models yielded valuable insights into the factors influencing the successful landing of Falcon 9's first stage. Key findings include:

- **Model Performance**: Each model displayed varying degrees of accuracy, with specific strengths and weaknesses highlighted through the analysis.
- **Cost Implications**: The results provided a clear understanding of how successful landings correlate with overall launch costs, reaffirming the economic advantage of SpaceX's reusable rocket strategy.
- **Strategic Insights for Competitors**: The findings offer actionable intelligence for companies aiming to compete with SpaceX. Understanding the likelihood of successful landings allows for more informed bidding strategies and cost management in the competitive space launch market.
- **Visualization Impact**: The use of interactive charts facilitated a deeper and more engaging exploration of the data, making the findings accessible to a broader audience, including stakeholders without a technical background.

Introduction

The project revolves around SpaceX's Falcon 9 rocket, particularly focusing on the first stage landing aspect. SpaceX has disrupted the space industry with its cost-effective launch services, primarily due to its pioneering approach in reusing the first stage of the Falcon 9 rocket. We would like to predict when a first stage will successfully land or not based on factors in the data we collect.

Problems To Answer

- Predictive Accuracy of Landing Success
- Influential Factors in Landing Success
- Competitive Analysis
- Model Selection and Justification
- Visualization and Communication

Section 1

Methodology

Methodology

Data collection methodology:

- The data was collected through API calls and Web Scraping.

Perform data wrangling

- The data was pre-processed with one-hot encoding and casting to the correct data types. Data normalization was also used on the independent variables before model fitting and evaluation.

Perform exploratory data analysis (EDA) using visualization and SQL

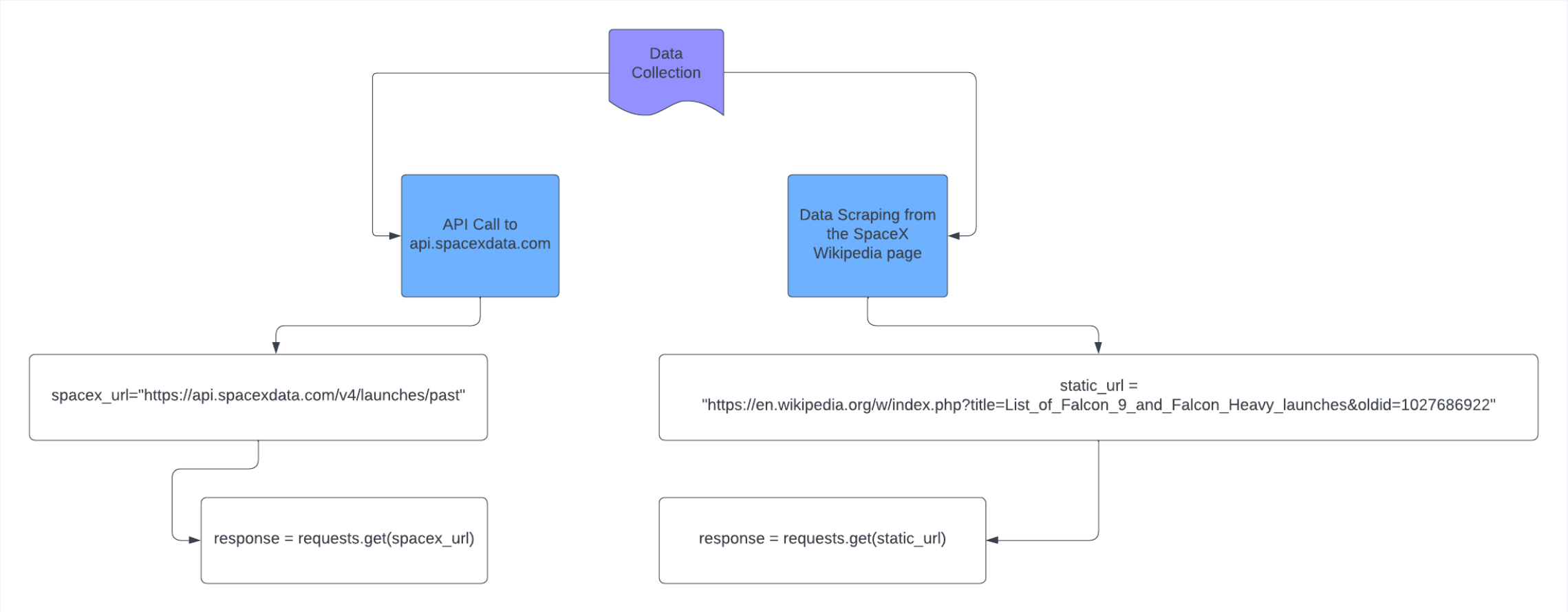
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

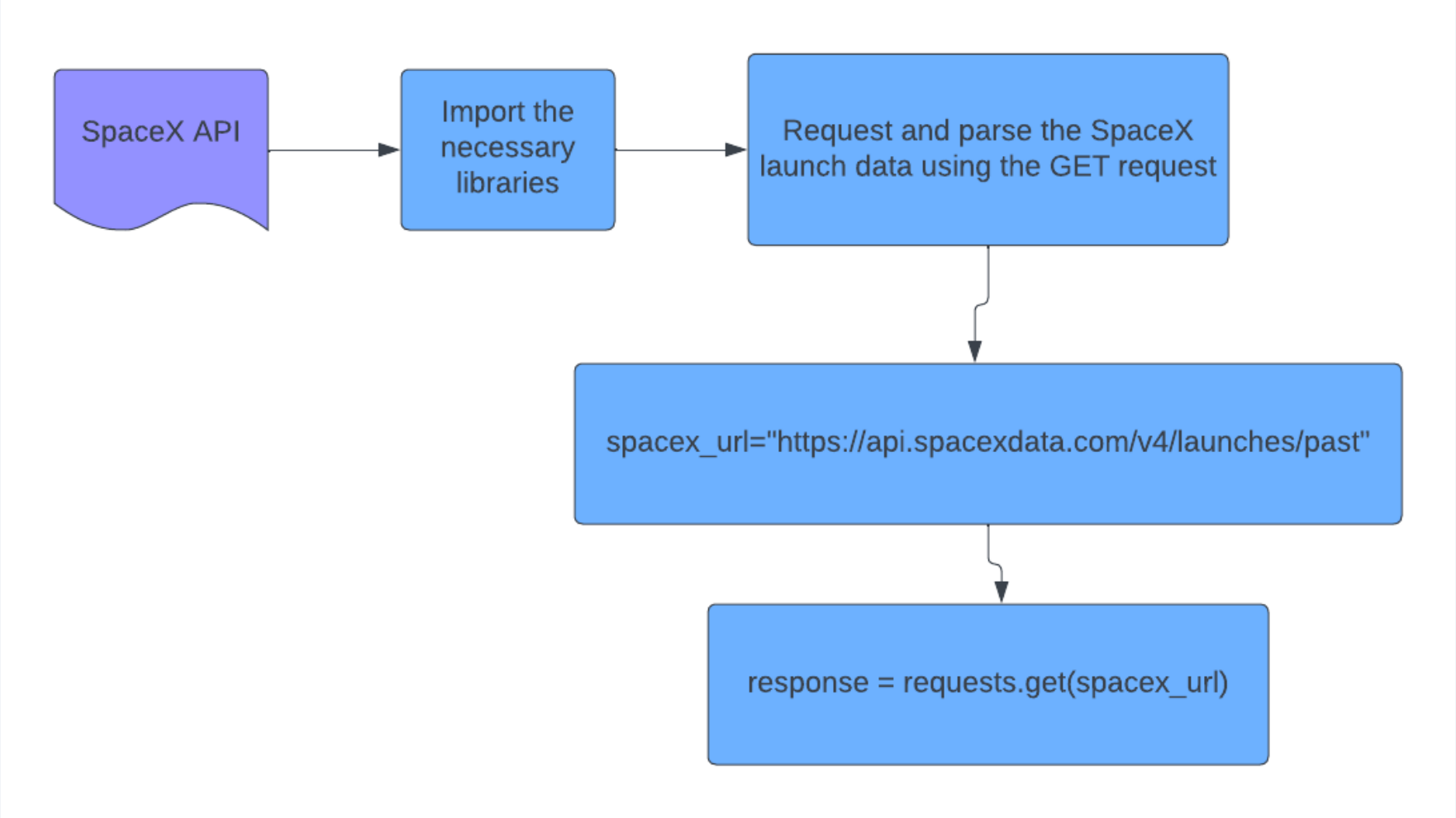
- Building a classification model involves several steps. Building a model involves first choosing what type to use. We have several options include Linear Regression, Linear Regression, KNN nearest neighbour, and CVM. We can used GridSearchCV to find the best parameters for our models, fine-tuning them and making sure we have the best parameters. Afterwards we will use the accuracy score in order to be able to pick the best model type.

Data Collection

- Data collection was done through API calls to the SpaceX database and also Webscraping the SpaceX Wikipedia webpage.

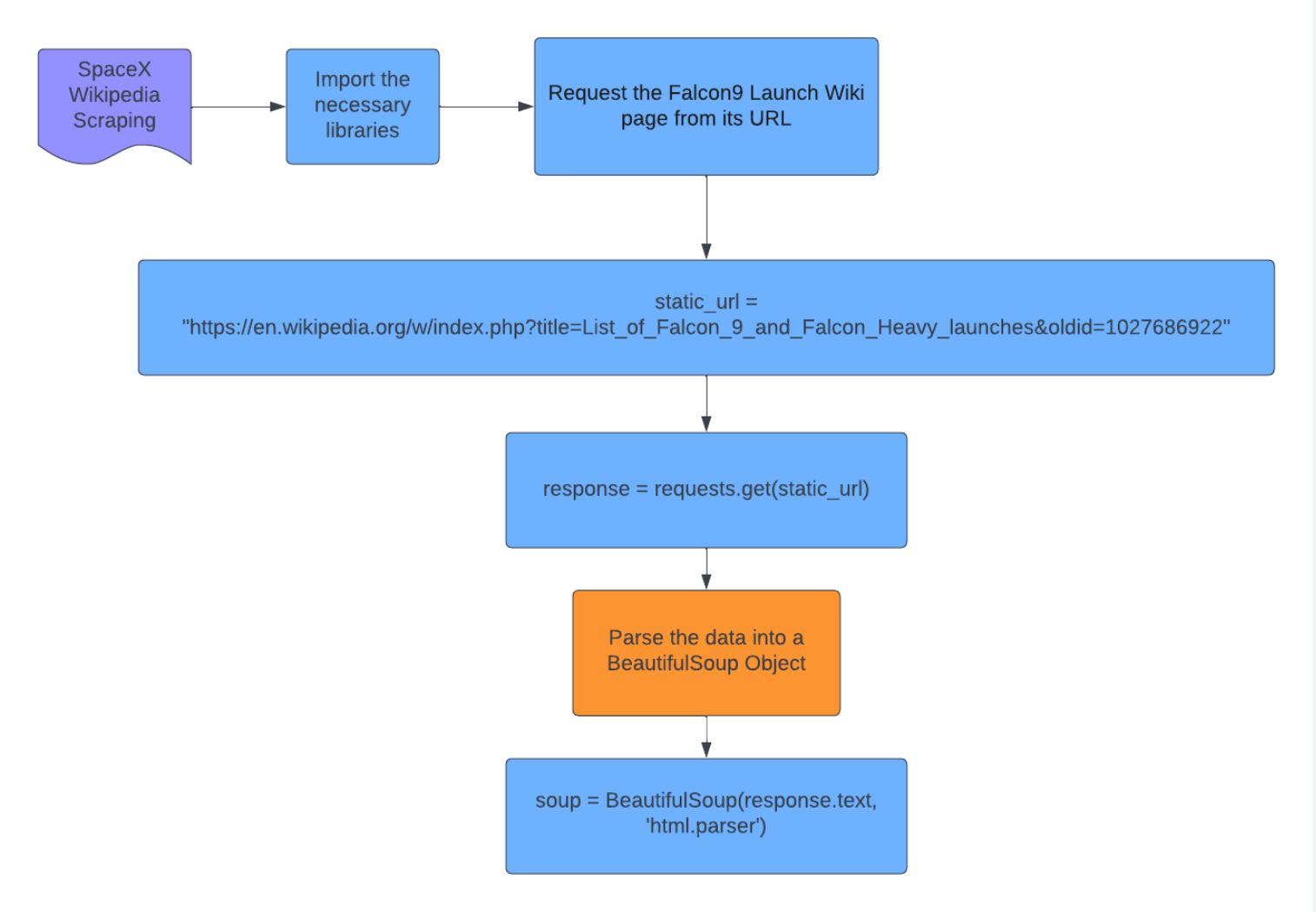


Data Collection – SpaceX API



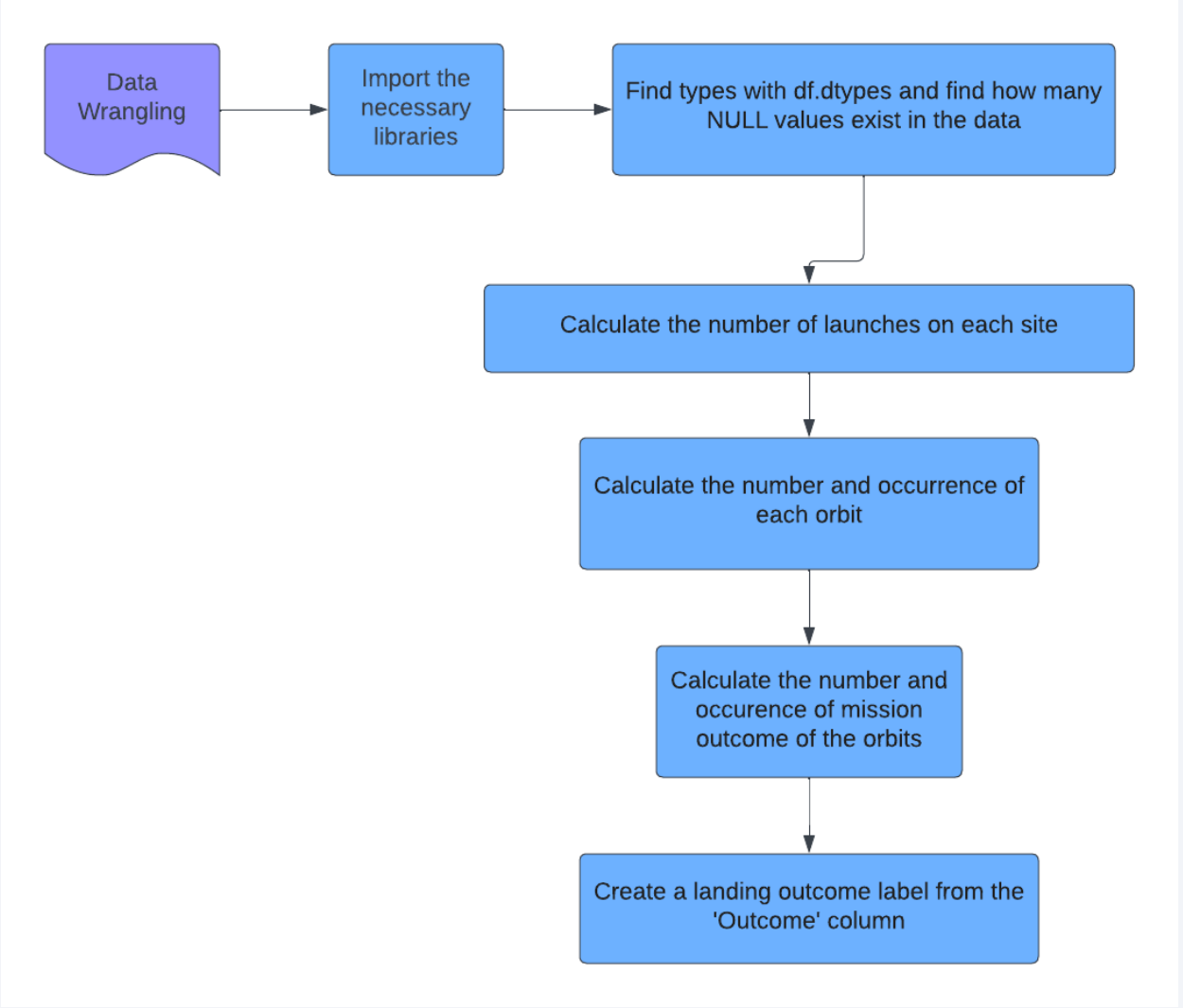
GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/API_Calls.ipynb

Data Collection - Scraping



GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/WebScraping.ipynb

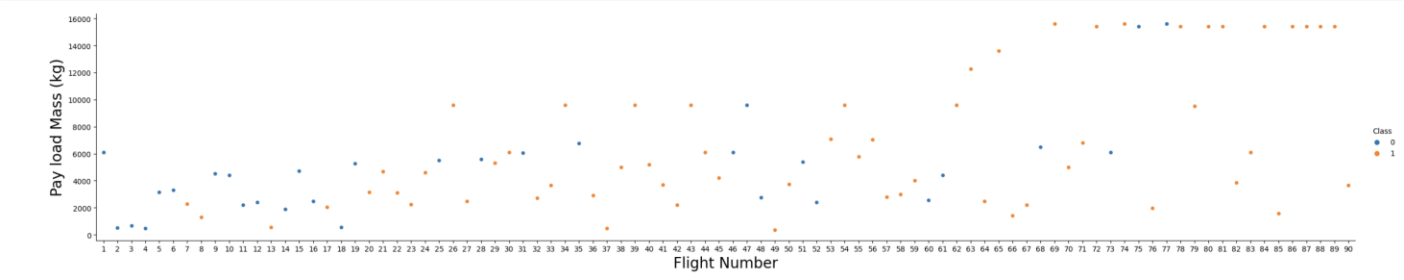
Data Wrangling



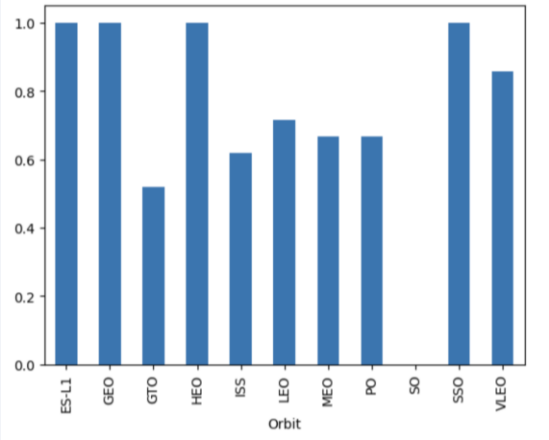
GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/Data%20Wrangling.ipynb

EDA with Data Visualization

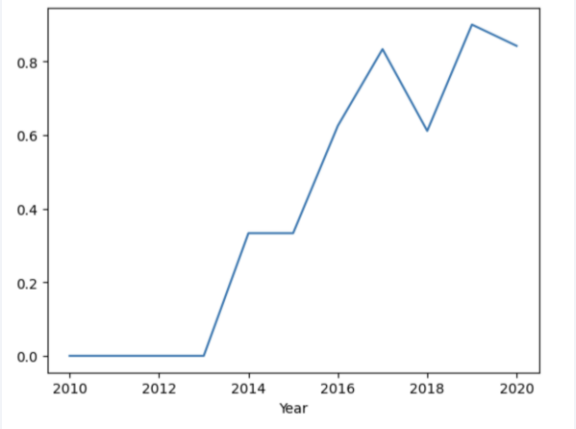
1. I first used some scatter plots to see the relationship between variables like for example this scatter plot with the Flight Number and Payload Mass and how they affect if the first stage of the rocket successfully landed back down or not



2.I also used a bar chart to group all the launch sites into groups and see each of their first stage rocket success rates



3. Finally a line chart was used to check the change in success rates through the years

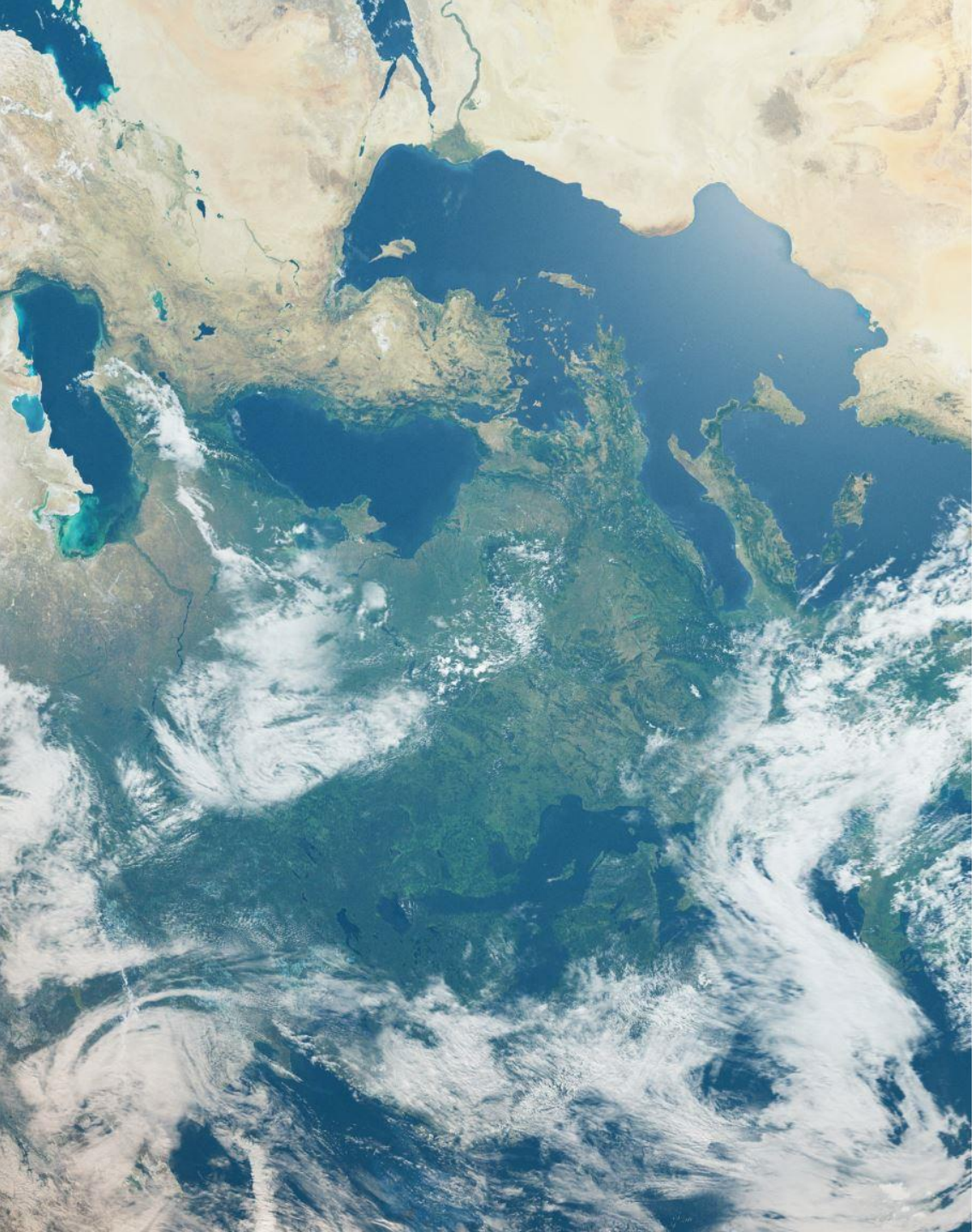


GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/SQL.ipynb

- Created a table from the original data with no Null values.
- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begin with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome in ground pad was achieved.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster versions which have carried the maximum payload mass. Used a subquery.
- Listed the records which displayed the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Ranked 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.



Build an Interactive Map with Folium

- Marked where the NASA Johnson Space Center's is with a popup label showing its name.
- Created and added a folium.Circle and a folium.Marker for each launch site on the site map.
- Added the launch outcome markers for each site, and saw which sites have the highest success rates.
- Added a MousePosition on the map to get the coordinates for a mouse over a point on the map. As such, while you are exploring the map, you can easily find the coordinates of any points of interests.
- Marked down a point on the closest coastline using the MousePosition and calculated the distance between the coastline point and a launch site.
- Created a folium.Marker to show the distance
- Drew a PolyLine between a launch site to the selected coastline point

Github: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/Folium.ipynb

Build a Dashboard with Plotly Dash

- I created two interactive graphs in Plotly Dash - one pie chart that depicts the success rate percentages for different Falcon 9 launch sites and one scatter plot that depicts the spread of successful and unsuccessful launches based on the rocket Payloads.
- I added these graphs because they offer a clear and interactive view on factors that correlate highly to a Falcon 9's booster landing success rate.

GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/Building%20a%20Plotly%20Dash%20Application

Predictive Analysis (Classification)

- I built 4 different models to choose from and try to predict a successful stage 1 landing: Logistic Regression, Decision Tree, and a Support Vector Machine.
- I evaluated each model based on each of the model's cross-validated score.
- I improved the models by using GridSearchCV with a set of different parameters that it could choose from to find the ones that fit each of the models the best.
- I finally chose the model that had the highest cross-validated score.

GitHub Link: https://github.com/diogodiasgrilo/Data-Science-Notebooks/blob/main/IBM_Data_Science_Capstone_Project/Model%20Building.ipynb

Results

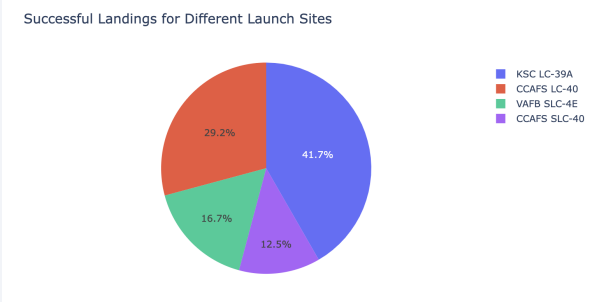
Exploratory data analysis results:

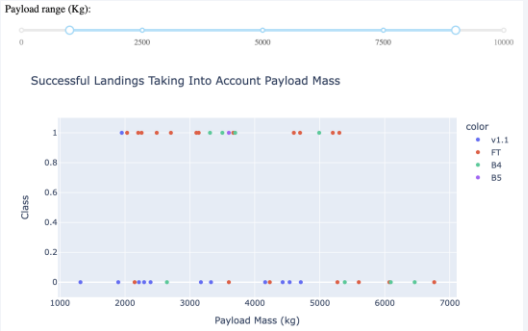
- Launch site CCAFS SLC 40 had the highest number of launches out of any launch site.
- The GTO orbit type was the highest used orbit type from out of all the launches.
- True ASDS was the highest occurring mission outcome out of all the launch outcomes.
- As the flight number increased, the likelihood of the mission being a success also steadily increased.
- For the VAFB-SLC launch site, there were no rockets launched for payload masses greater than 10,000 kg.
- ES-L1, GEO, HEO and SSO orbit types had the highest booster landing success rates.
- In the LEO orbit, the success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number to success rate when in GTO orbit.
- With heavy payloads the successful landing or positive landing rates are highest for Polar, LEO and ISS orbit types.
- The success rates since 2013 have kept increasing until 2017 (stable in 2014) and after 2015 they started increasing again.

Predictive analysis results:

- Flight Number, Payload Mass, Orbit Type, Launch Site, Landing Pad, and Serial number have the highest impact on whether a Falcon 9 reusable rocket will land. These were the data sections used to train and use the predictive models for this specific project.

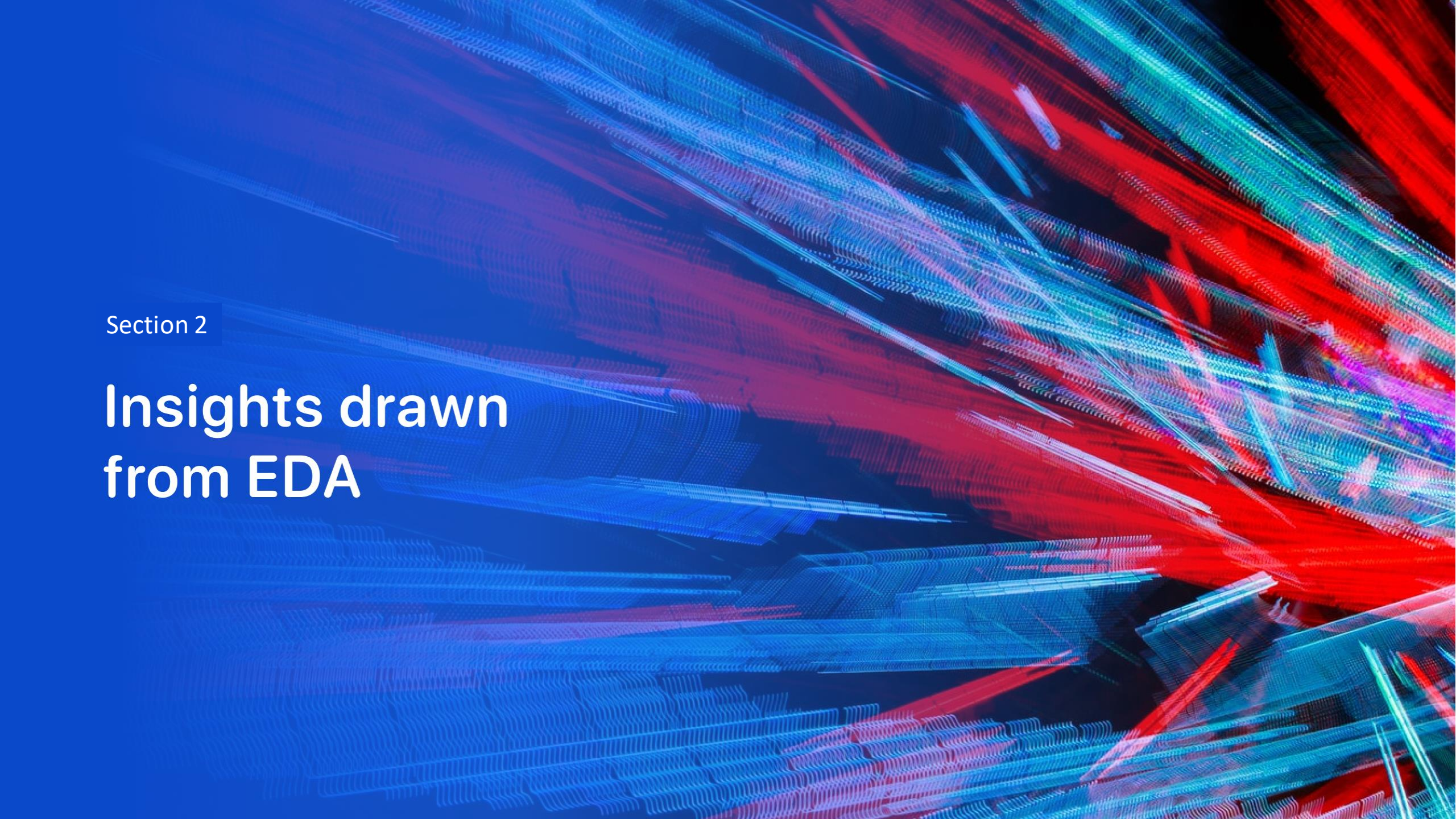
Interactive analytics demo in screenshots:



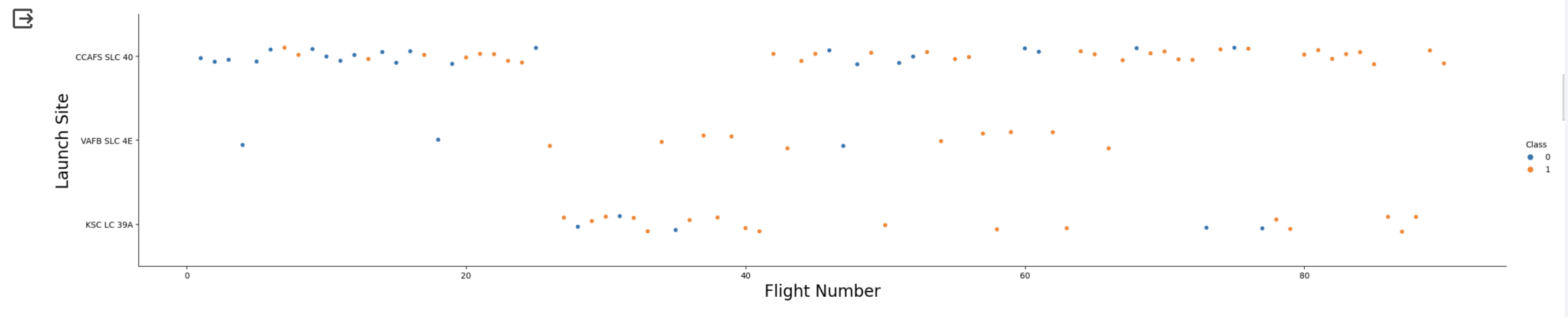


Section 2

Insights drawn from EDA

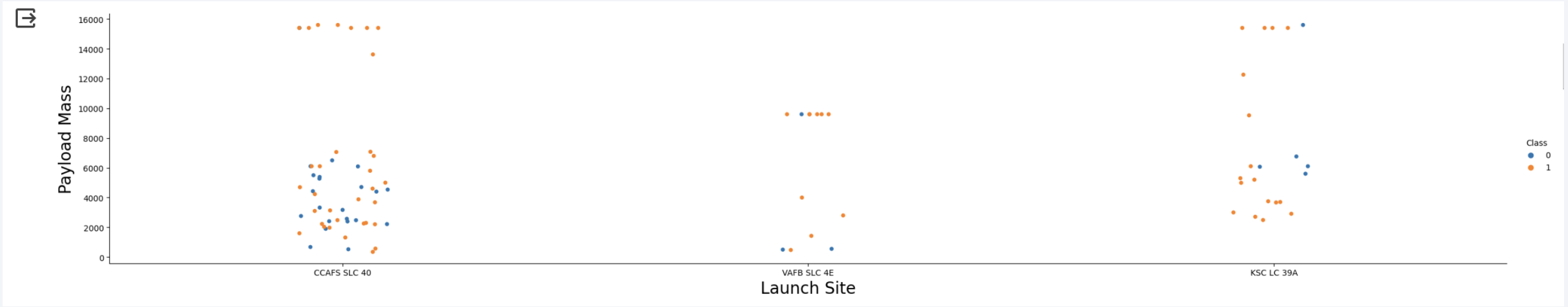


Flight Number vs. Launch Site



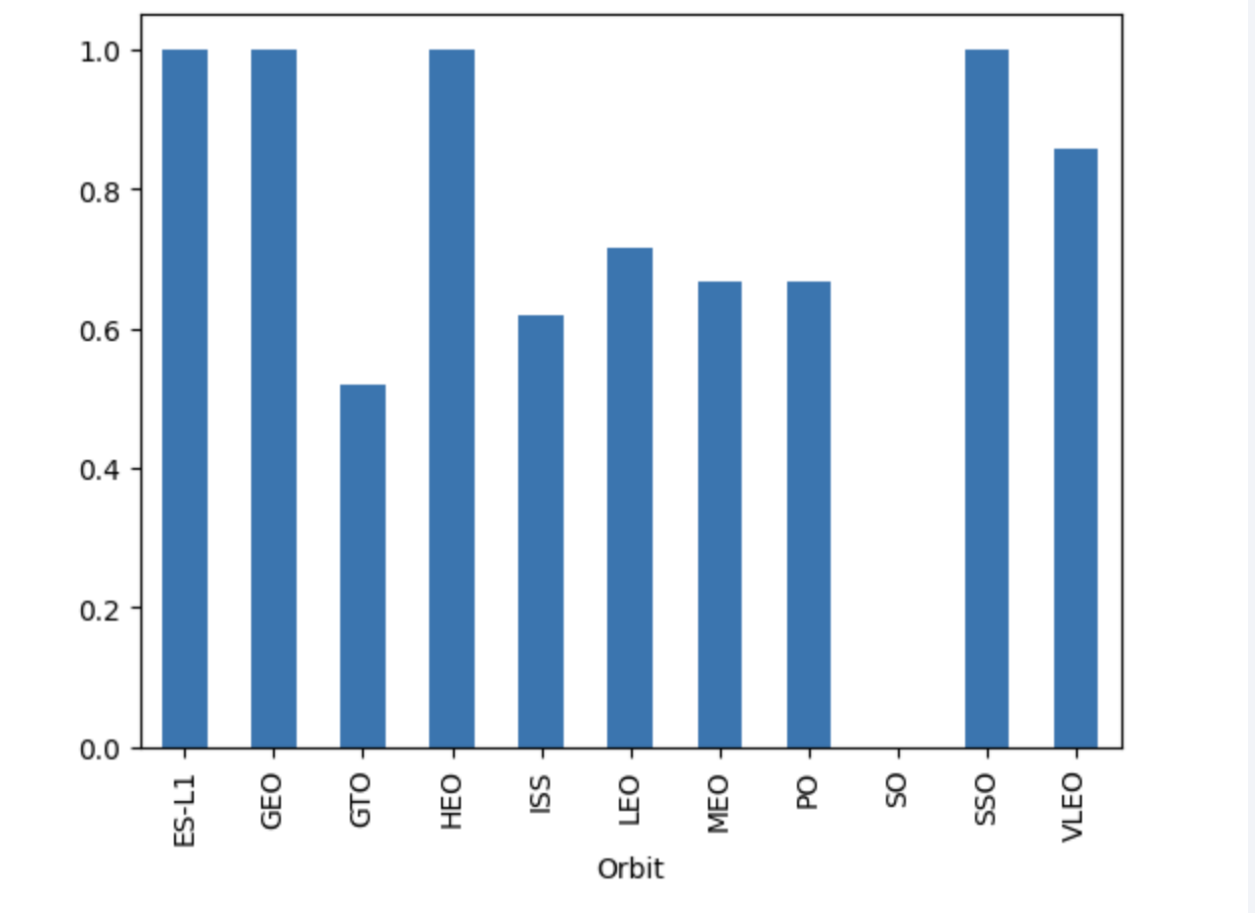
- The launch site with the most launches was CCAFS SLC 40 and as the flight number goes up, all launch sites become more and more successful with their launches.

Payload vs. Launch Site



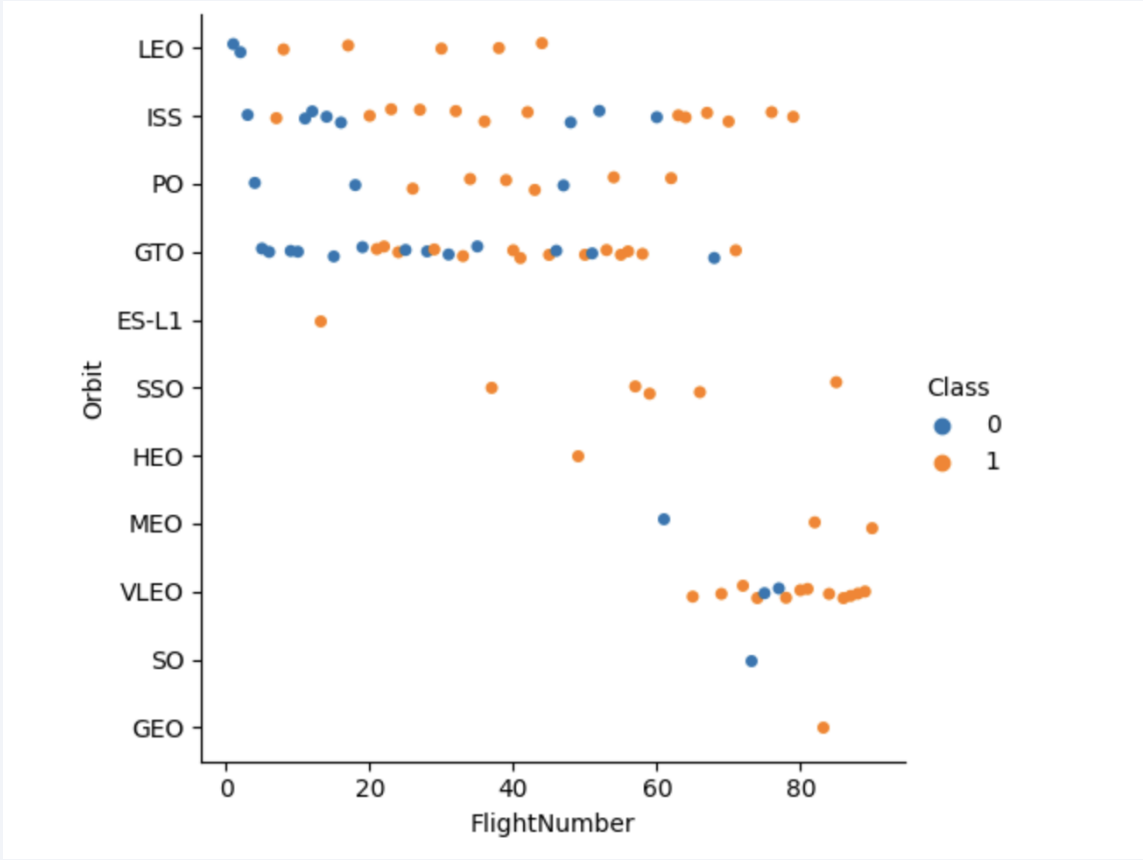
- The greatest payloads were launched out of the CCAFS SLC 40 launch site and all of them, and most heavier payloads overall, did extremely well.

Success Rate vs. Orbit Type



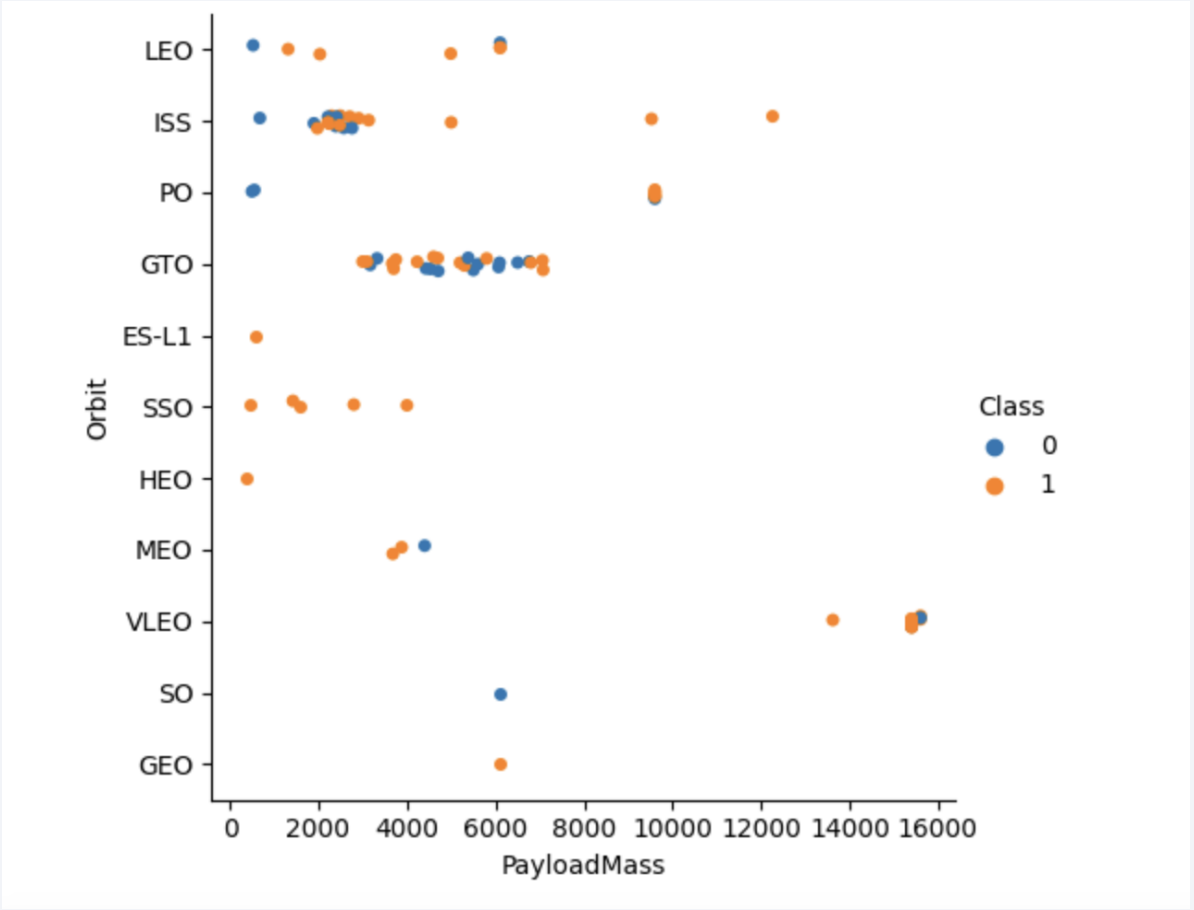
- The most successful orbits overall were the ES-L1, GEO, HEO, and the SSO orbit types.

Flight Number vs. Orbit Type



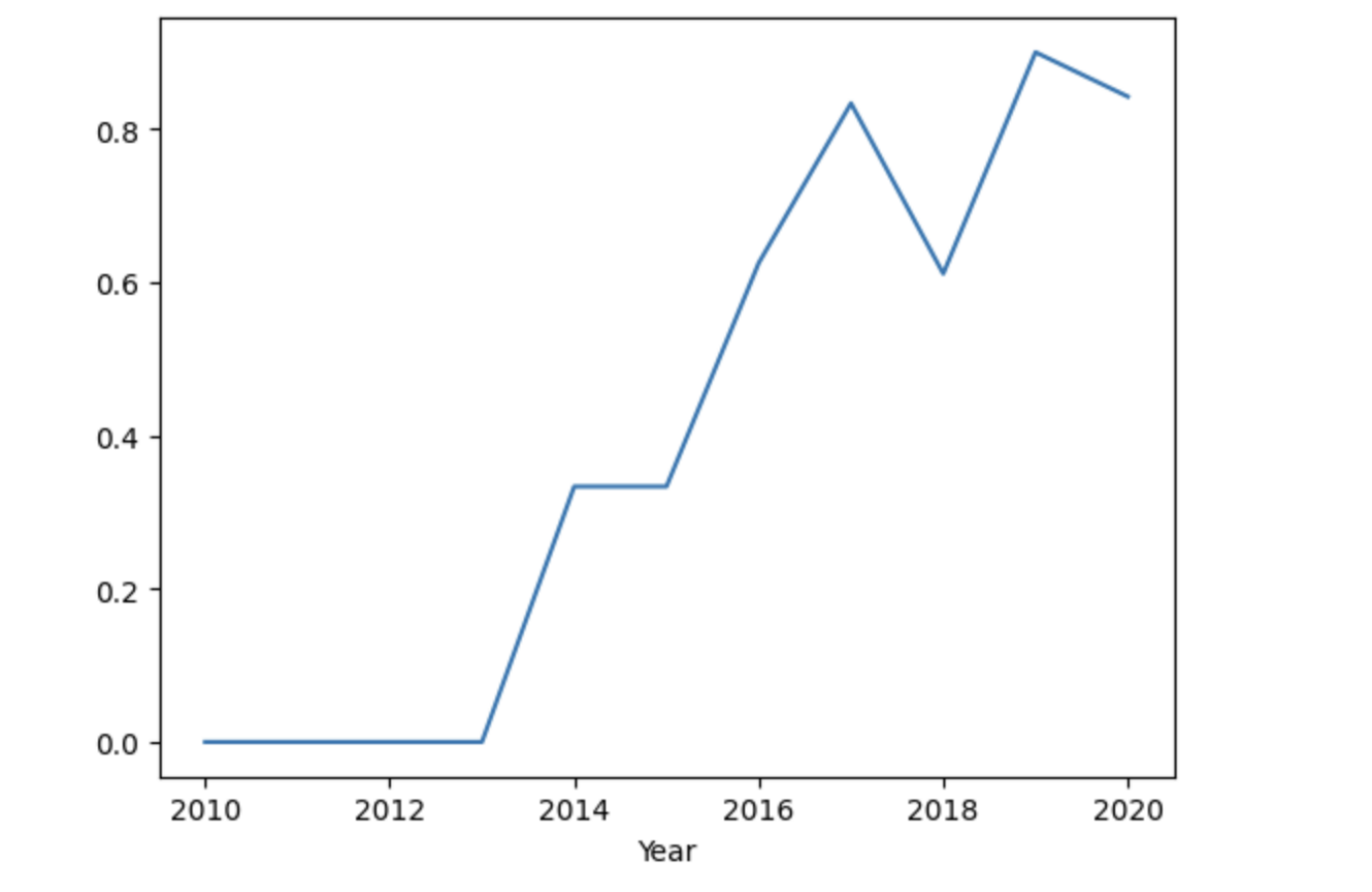
- As the flight number goes up, every orbit type starts performing much better.

Payload vs. Orbit Type



- Orbit types like ES-L1, SSO and HEO do well with low payload masses, and the best orbits for higher payloads are ISS and VLEO.

Launch Success Yearly Trend



- As the years go by, the success rate starts to steadily increase, with only a small decrease in 2017.

All Launch Site Names

- The unique names of the launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

```
1 %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

- We select only the 'DISTINCT' launch site names from the table using the query above. That way we make sure we get no duplicates.

Launch Site Names Begin with 'CCA'

```
1 %sql SELECT Launch_Site FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- By sing the 'WHERE' keyword in our SQL query we are able to find a pattern in the 'Launch_Site' column. The pattern we are looking for is names that begin with 'CCA'. We can do this by using the 'LIKE' keyword and then the pattern in this case 'CCA%' where % means that anything can follow but the word must start with CCA. At the end of our query we use the 'LIMIT' keyword to only show the first 5 occurrences of these target names.

Total Payload Mass

```
1 %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
SUM(PAYLOAD_MASS__KG_)
619967
```

- By using the aggregate function SUM() in this query we are able to add up all the values in a specific column of our choosing, in this case the total payload mass.

Average Payload Mass by F9 v1.1

```
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

- By using the aggregate function AVG() in this query we are able to find the average of all the values in a specific column of our choosing, in this case the total payload mass. Then, we can use the keyword 'WHERE' to specify that we only want values where the 'Booster_Version' is 'F9 v1.1'.

First Successful Ground Landing Date

```
[15]  1 %sql SELECT MIN(DATE) AS 'First Successful Landing Date' FROM SPACEXTABLE WHERE Landing_Outcome = 'Success';
```

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

First Successful Landing Date
2018-07-22

- By using the aggregate function MIN() in this query we are able to find the minimum value of all the values in a specific column of our choosing, in this case the date. Then, we can use the keyword 'WHERE' to specify that we only want values where the 'Landing_Outcome' was a 'Success'. This will give us the smallest date in these values which will in turn be the first successful landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000 were Merah Putih, Es hail 2, SSO-A, Nusantara Satu, RADARSAT Constellation, GPS III-03, ANASIS-II, and GPS III-04.

```
1 %sql SELECT Payload FROM SPACEXTABLE WHERE Landing_Outcome = 'Success' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

Payload
Merah Putih
Es hail 2
SSO-A
Nusantara Satu, Beresheet Moon lander, S5
RADARSAT Constellation, SpaceX CRS-18
GPS III-03, ANASIS-II
ANASIS-II, Starlink 9 v1.0
GPS III-04 , Crew-1

- In this query we use the 'WHERE' keyword to only select values that had a successful landing and then we use the 'AND' keyword to specify yet another condition. This next condition uses the 'BETWEEN' keyword to state that the values also need to have payloads between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
1 %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE;
```

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

COUNT(Mission_Outcome)
101

- By using the aggregate function COUNT() in this query we are able to count the number of values in a specific column of our choosing, in this case in the 'Mission_Outcome' column. This will give us the total number of missions, failed and successful.

Boosters Carried Maximum Payload

- The names of the boosters which have carried the maximum payload mass were Starlink 1 v1.0, Starlink 2 v1.0, Starlink 3 v1.0, Starlink 4 v1.0, Starlink 5 v1.0, Starlink 6 v1.0, Starlink 7 v1.0, Starlink 11 v1.0, Starlink 12 v1.0, Starlink 13 v1.0, Starlink 14 v1.0, and Starlink 15 v1.0.

```
1 %sql SELECT DISTINCT Payload FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Payload
Starlink 1 v1.0, SpaceX CRS-19
Starlink 2 v1.0, Crew Dragon in-flight abort test
Starlink 3 v1.0, Starlink 4 v1.0
Starlink 4 v1.0, SpaceX CRS-20
Starlink 5 v1.0, Starlink 6 v1.0
Starlink 6 v1.0, Crew Dragon Demo-2
Starlink 7 v1.0, Starlink 8 v1.0
Starlink 11 v1.0, Starlink 12 v1.0
Starlink 12 v1.0, Starlink 13 v1.0
Starlink 13 v1.0, Starlink 14 v1.0
Starlink 14 v1.0, GPS III-04
Starlink 15 v1.0, SpaceX CRS-21

- By using a subquery we are able to retrieve data that will be used as a condition or value in the main query. In this case, we are only retrieving the payload names of payloads that carried the maximum payload mass in the whole table. The subquery is used to retrieve the maximum payload and compare it to the values to be shown.

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015 were: 'January - Failure (drone ship) - F9 v1.1 B1012 - CCAFS LC-40', and 'April - Failure (drone ship) - F9 v1.1 B1015 - CCAFS LC-40'.

- By using the 'CASE' keyword we are able to correctly convert each month's number to the name of the month like requested. This then allows us to use the substr() function to select only the months that had failed launches in the year '2015'. Another import detail to notice is that we have to use 2 'AND' keywords in order to be able to narrow down to only failed landings that were specifically on a drone ship.

```
%%sql
SELECT
    CASE
        WHEN substr(Date, 6, 2) = '01' THEN 'January'
        WHEN substr(Date, 6, 2) = '02' THEN 'February'
        WHEN substr(Date, 6, 2) = '03' THEN 'March'
        WHEN substr(Date, 6, 2) = '04' THEN 'April'
        WHEN substr(Date, 6, 2) = '05' THEN 'May'
        WHEN substr(Date, 6, 2) = '06' THEN 'June'
        WHEN substr(Date, 6, 2) = '07' THEN 'July'
        WHEN substr(Date, 6, 2) = '08' THEN 'August'
        WHEN substr(Date, 6, 2) = '09' THEN 'September'
        WHEN substr(Date, 6, 2) = '10' THEN 'October'
        WHEN substr(Date, 6, 2) = '11' THEN 'November'
        WHEN substr(Date, 6, 2) = '12' THEN 'December'
    END AS Month,
    Landing_Outcome AS 'Landing Outcome',
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2015'
    AND Landing_Outcome LIKE 'Failure%'
    AND Landing_Outcome LIKE '%drone ship%';
```

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

Month	Landing Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) as Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

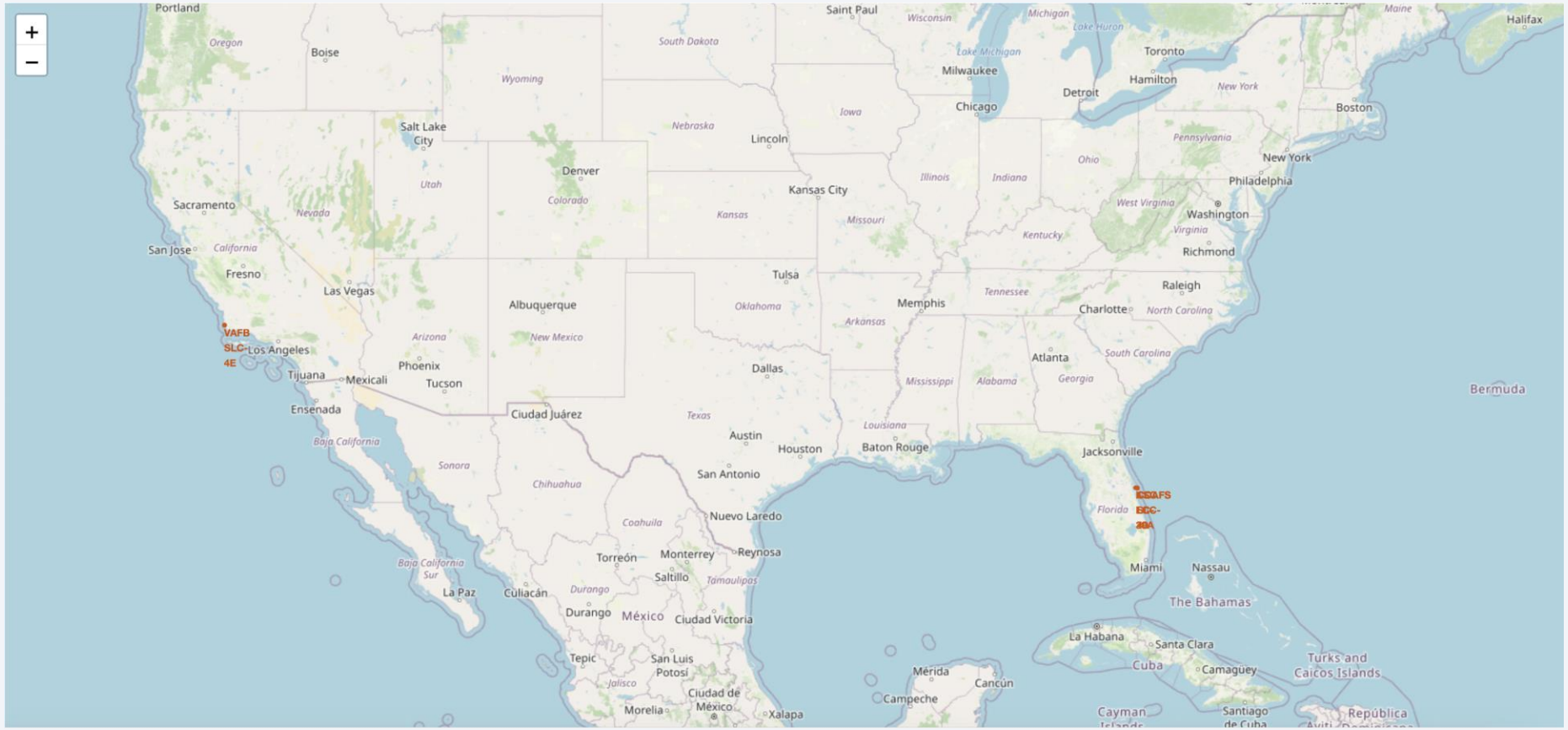
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	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

- To rank each landing outcome between 2010-06-04 and 2017-03-20 we need to use the 'GROUP BY' keyword to group the values by their landing outcomes. We can then specify that we want the ordering to be done by the sum of the outcomes for each landing outcome and displayed in descending order.

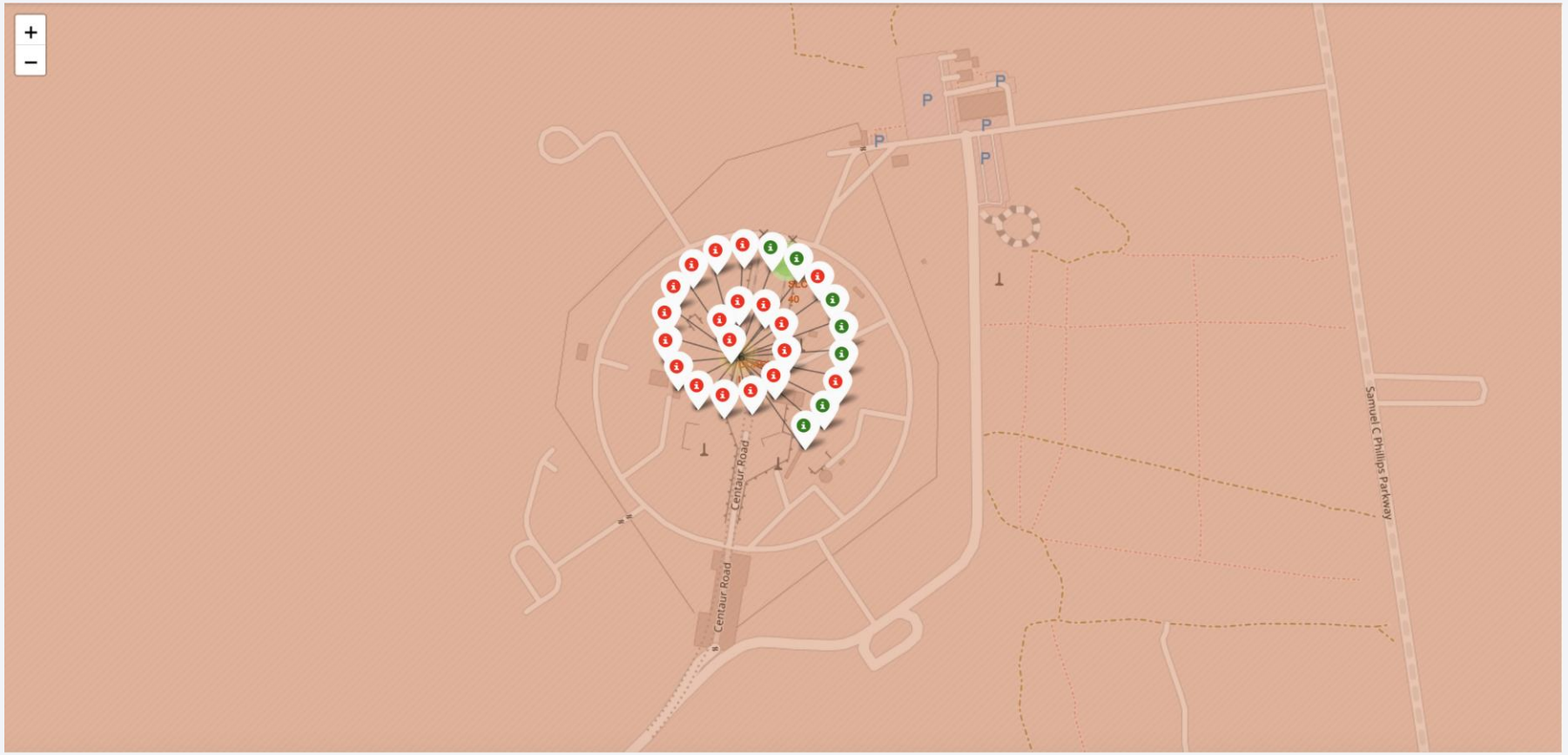

Section 3
Launch Sites
Proximities Analysis

SpaceX Falcon 9 Launch Sites



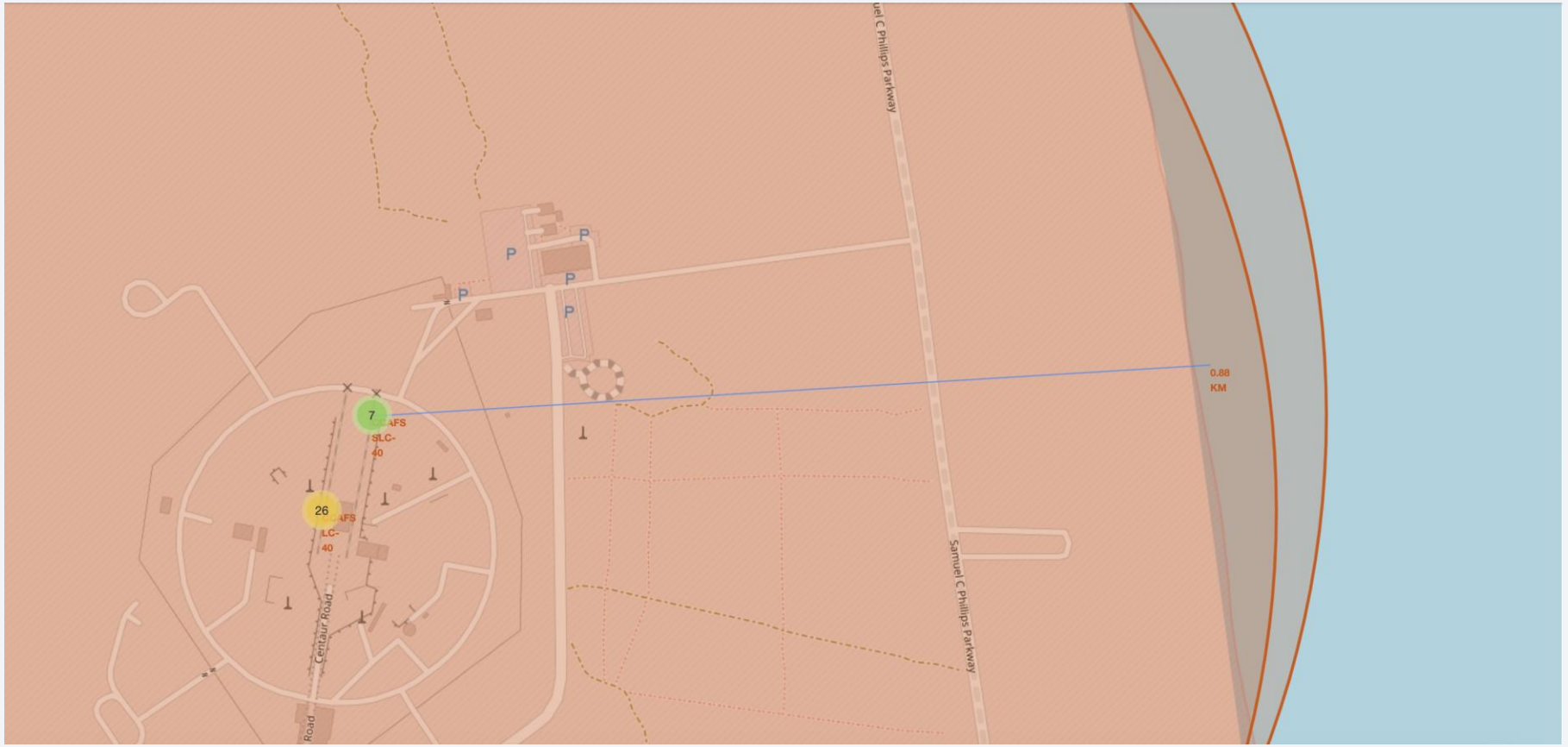
- As we can see from the map, SpaceX has three launch sites in central Florida and one in California. This means that they are quite spread out and we can learn insights on how this can affect the success of the launches.

Marked Landing Outcomes at Each Launch Site

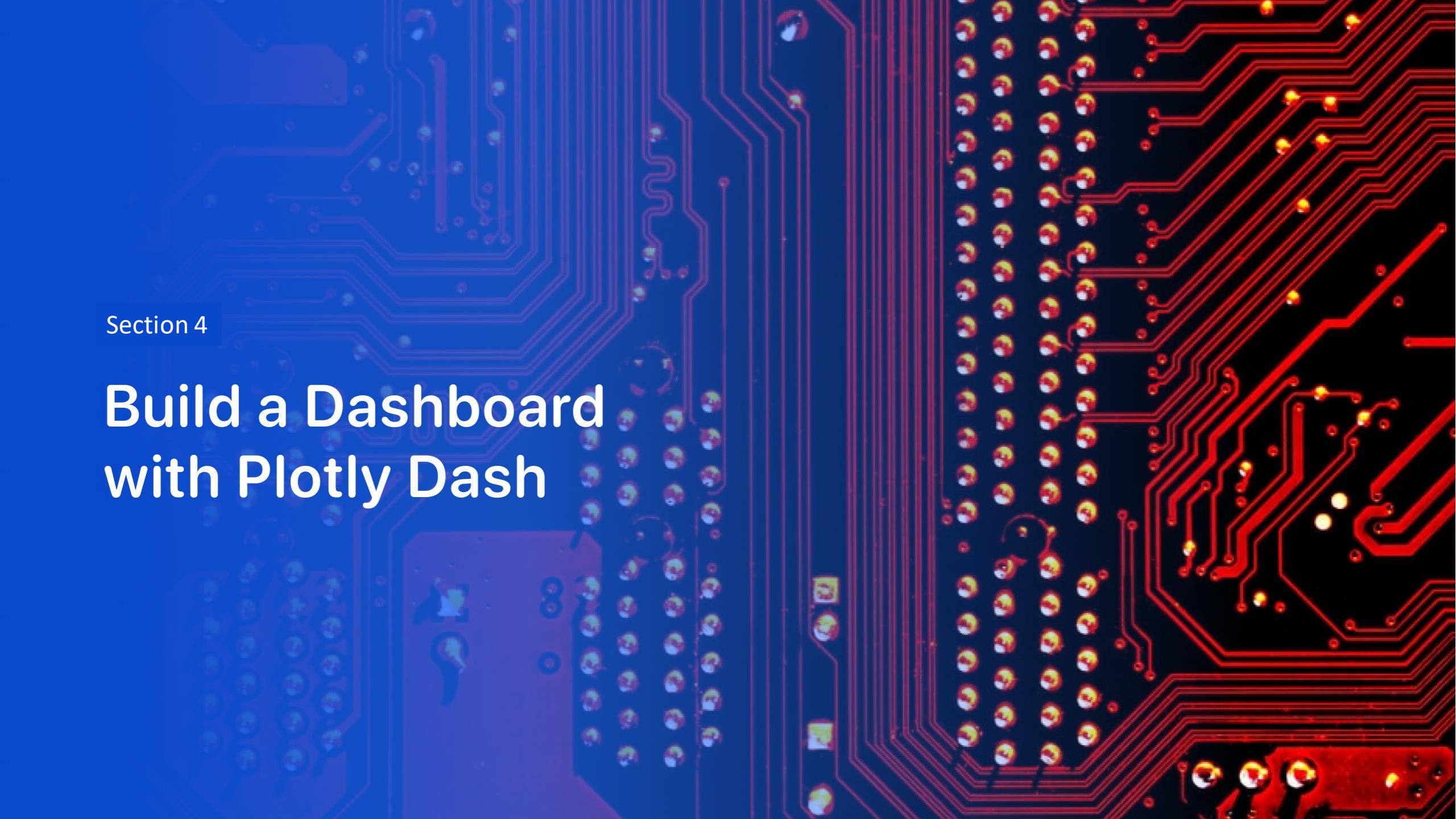


- In this Folium map I marked every launch site with its individual launches. Each red mark means a failed launch and each green mark means a successful launch. This allows us to easily zoom into each of the launch sites and be able to quickly and visually see the outcomes of the launches.

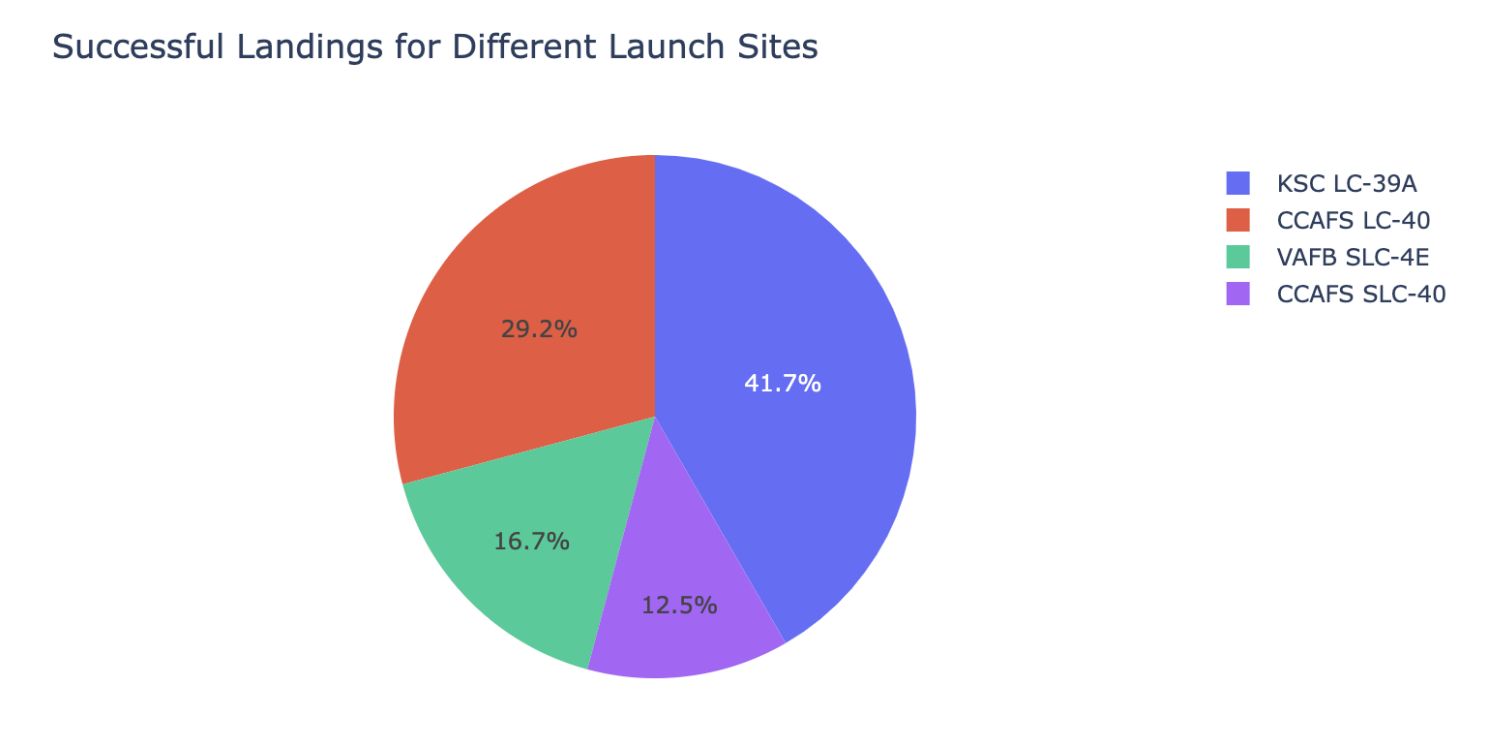
Proximity of **Landmarks** to Launch Sites



- This Folium map has markings detailing how far away a specific landmark is to a launch site, in this case the nearest coastline. We can visually see it by the blue line displayed and at the position of the coastline a mark indicating how far away it is from the launch site. In this case the closest coastline is 0.88 km away.

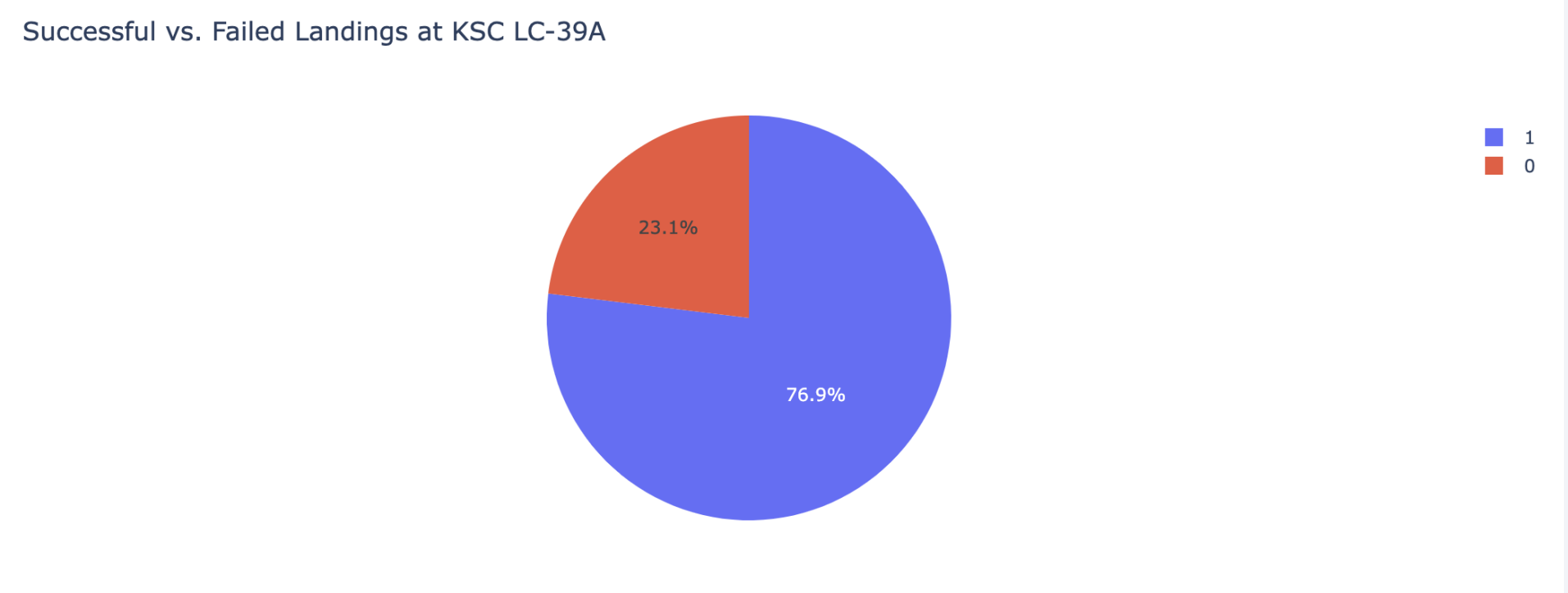

Section 4
Build a Dashboard with Plotly Dash

Launch Success Percentages for all Launch Sites



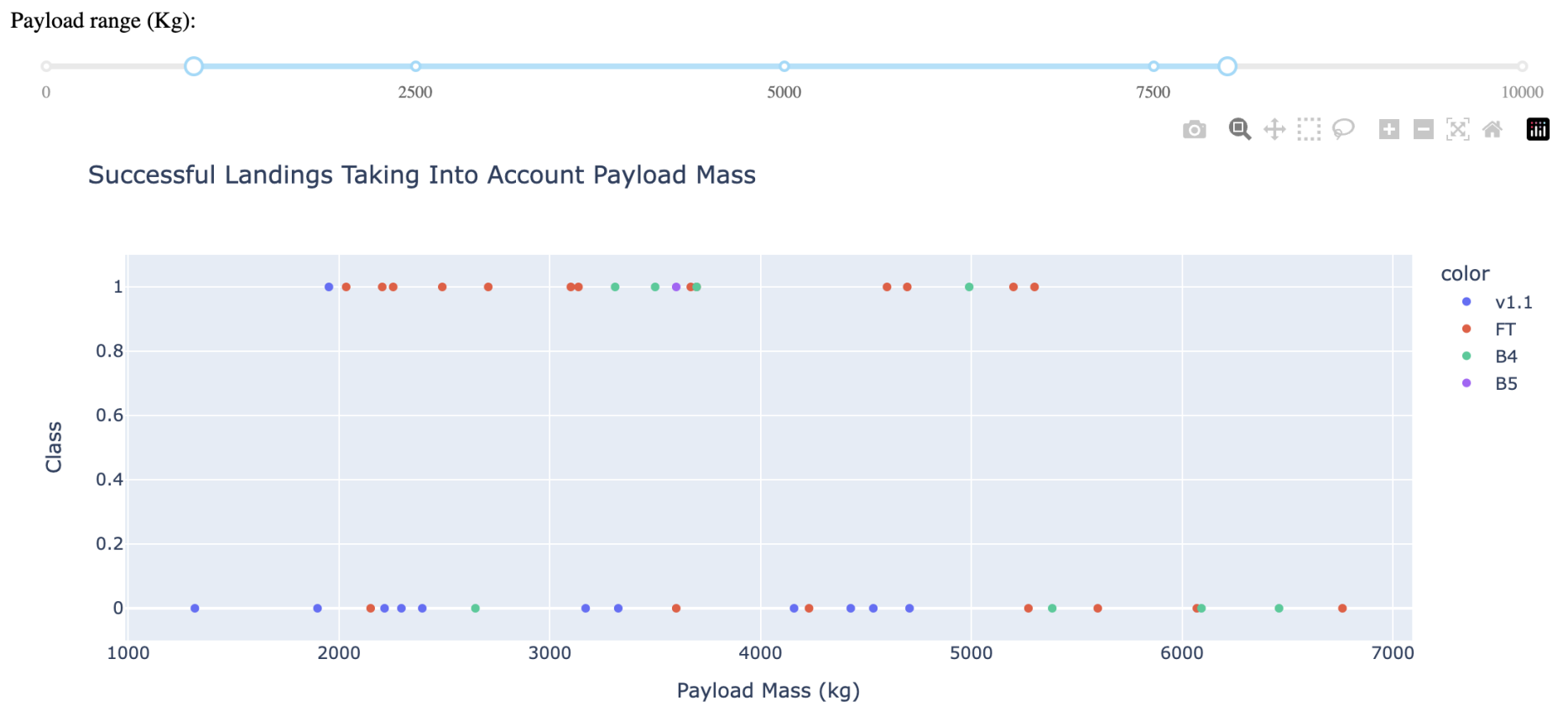
- With this view of the interactive pie chart made with Plotly Dash we are able to visually see the different success percentages of each of SpaceX's launch sites. We can quickly tell that KSC LC-39A has the highest success rate out of all the sites.

Success Rate Pie Chart For Top Launch Site



- This pie chart represents the top SpaceX launch site. It is split up into the percentages of successful launches and failed launches. As we can see here the successful launches clearly outweigh the failures.

Scatter Plot of Payload Vs Success Rate

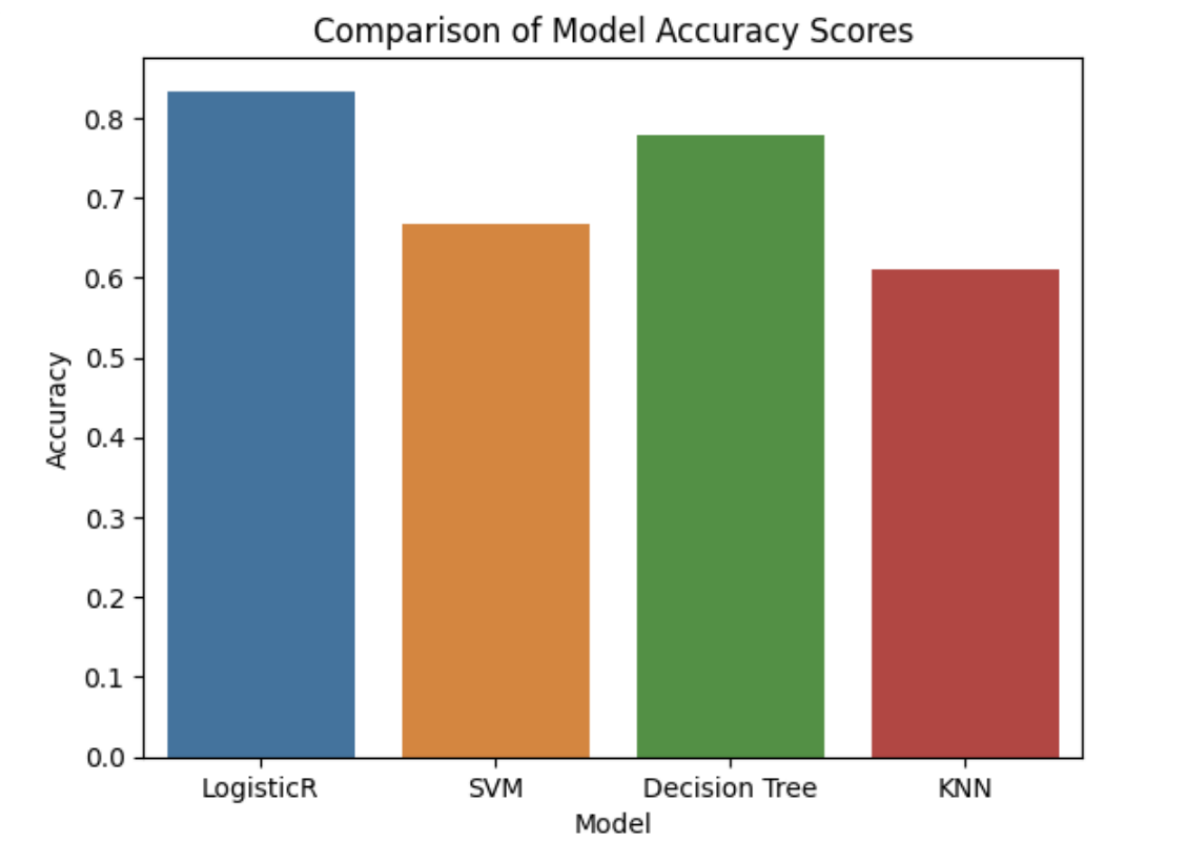


- There are several points to note on this graph. The payload range that is the most successful is between 1952 kg and 5300 kg and the booster version that is the most successful is the FT version. It is a great interactive graph because with the intuitive slider you can quickly change the graph to only show specific kg ranges, allowing for a customized learning experience.

Section 5

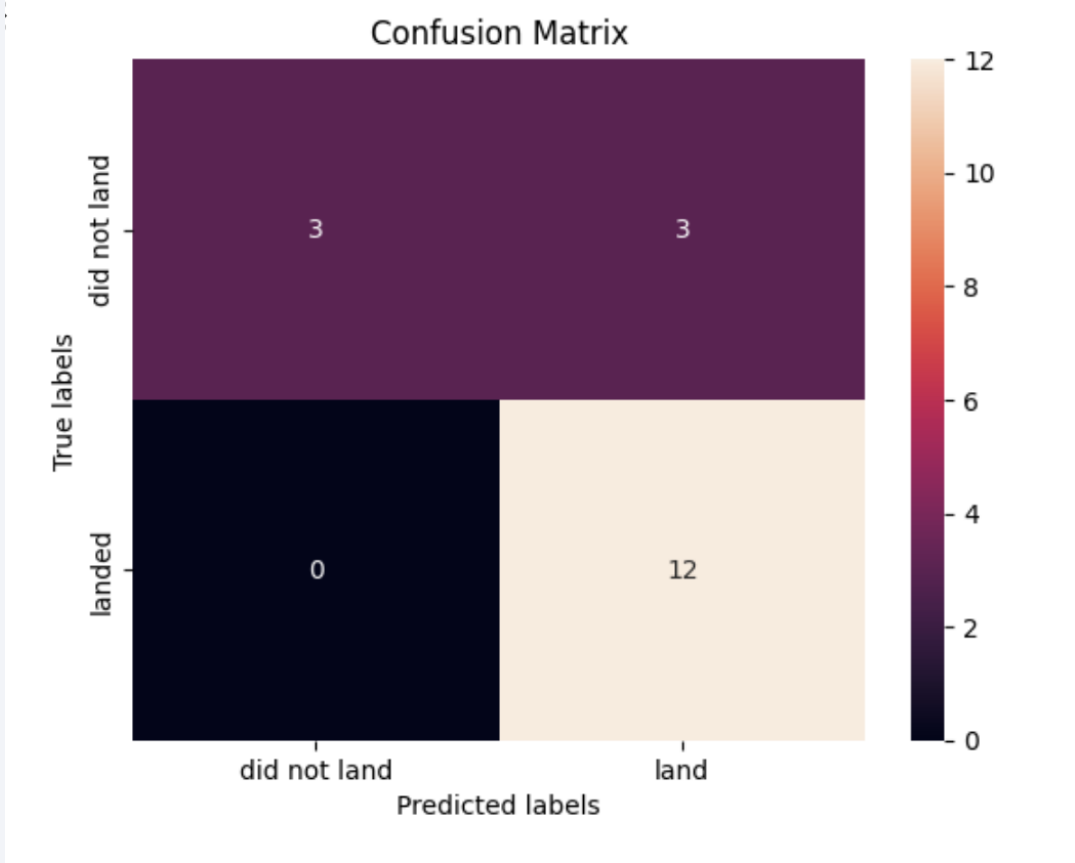
Predictive Analysis (Classification)

Classification Accuracy



- I tested several different models aiming to choose one that was the best at predicting our target – if the first stage rocket of the SpaceX launch will land or not. I test 4 different models including Logistic Regression, Support Vector Machines, Decision Tree Classifiers and K Nearest Neighbors. Out of these four the Logistic Regression model outranked all the others and from my observations as seen in the graph above this is the best predictive model based on the models' accuracy scores.

Confusion Matrix



- This is the confusion matrix for our best performing model – Logistic Regression. A confusion matrix is a table that summarizes the performance of a classification model by showing the number of true positive, true negative, false positive, and false negative predictions. It helps assess the model's accuracy, precision, recall, and other performance metrics. We can see here for example that our Logistic Regression model did extremely well at predicting when the rocket landed successfully.

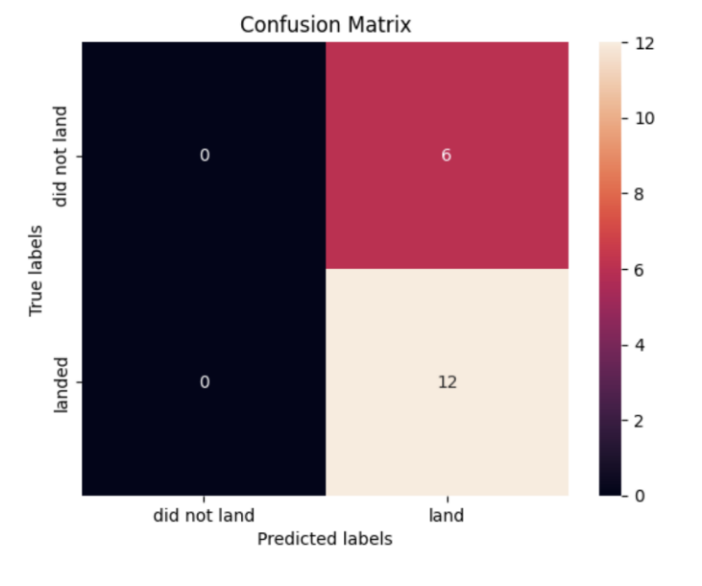
Conclusions

- Launch site CCAFS SLC 40 had the highest number of launches out of any launch site.
- The GTO orbit type was the highest used orbit type from out of all the launches.
- True ASDS was the highest occurring mission outcome out of all the launch outcomes.
- As the flight number increased, the likelihood of the mission being a success also steadily increased.
- For the VAFB-SLC launch site, there were no rockets launched for payload masses greater than 10,000 kg.
- ES-L1, GEO, HEO and SSO orbit types had the highest booster landing success rates.
- In the LEO orbit, the success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number to success rate when in GTO orbit.
- With heavy payloads the successful landing or positive landing rates are highest for Polar, LEO and ISS orbit types.
- The success rates since 2013 have kept increasing until 2017 (stable in 2014) and after 2015 they started increasing again.
- Flight Number, Payload Mass, Orbit Type, Launch Site, Landing Pad, and Serial number have the highest impact on whether a Falcon 9 reusable rocket will land. These were the data sections used to train and use the predictive models for this specific project.

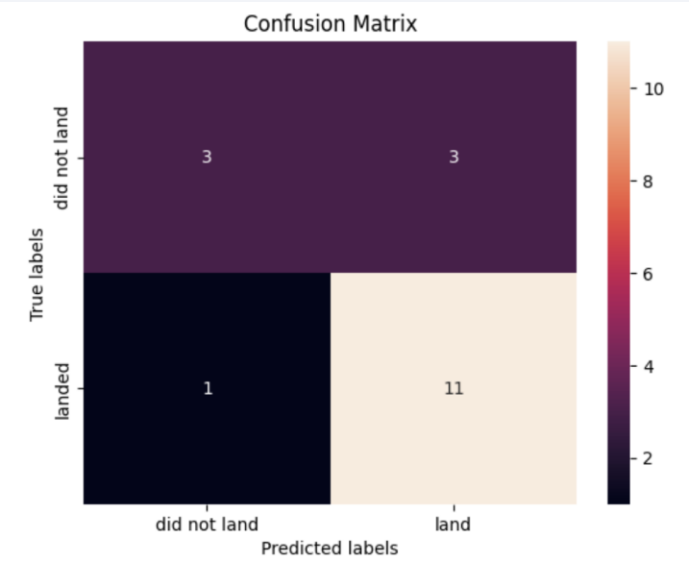
Appendix

- Confusion matrixes for the 3 other predictive models:

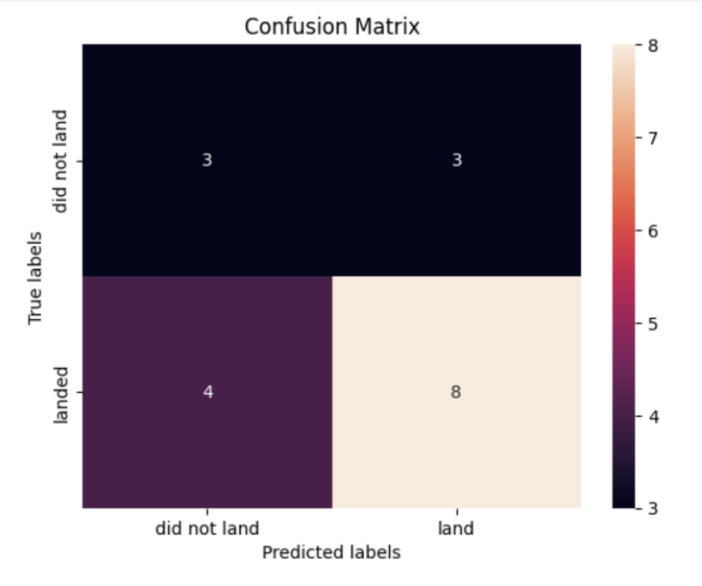
Support Vector Machine



Decision Tree Classifier



K Nearest Neighbors



Thank you!