Winning Space Race with Data Science

David Bell
05/06/2022

Outline

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

Executive Summary

Summary Of Methodology

- Collect data through web scraping and API calls (to corroborate data)
- Perform exploratory data analysis (EDA) using both SQL and visualisations
- Build an interactive map to present locational information about launch sites
- Build an interactive Dashboard for presentation
- And finally, build and evaluate a machine learning model to make a prediction

Summary of all results

- Smaller payload launches have a higher success rate.
- KSC LC-39A has the highest success rate for previous launches.
- The success rate of launches has been generally increasing since 2010.
- Launches with the objective of reaching ES-L1, SSO, HEO and GEO orbits have previously had a 100% success rate.
- The Decision Tree model is the best classification method for predicting the outcome of the Falcon 9 booster landing, with an accuracy of 0.875.

Introduction

SpaceX 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 SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

In this research I will attempt to determine answers to the following problems:

- How will the impact of various parameters will affect the outcome?
- Is there a correlation between launch sites and success rates?
- Is it possible to predict if the Falcon 9 first stage will land successfully based on historical data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection Sources:
  - SpaceX API [1]
  - Web Scraping of SpaceX Wikipedia site [2]
- Perform data wrangling
  - Process the outcome of each launch into distinct class variables are used to train the a predictive model (0 = unsuccessful, 1 = successful)
  - The data is standardised and a dataframe produced
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model.
  - Data split into testing and training sets and then used in 4 classification algorithms which are evaluated to determine the best algorithm to use

Data Collection

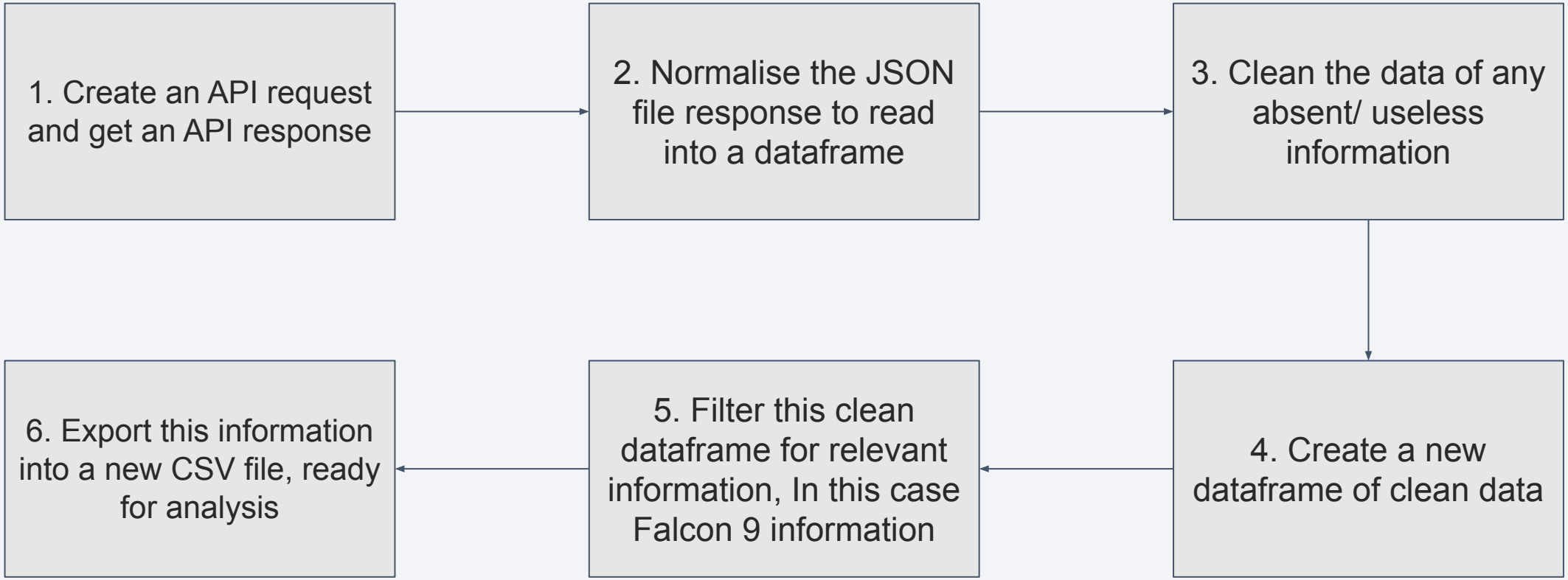
For this project, two methods were used to collect data

- Data was requested directly from the SpacX API
- Data was scraped from the SpaceX Falcon Project Wikipedia

This was done to compare the two datasets and ensure the correct data was being collected.

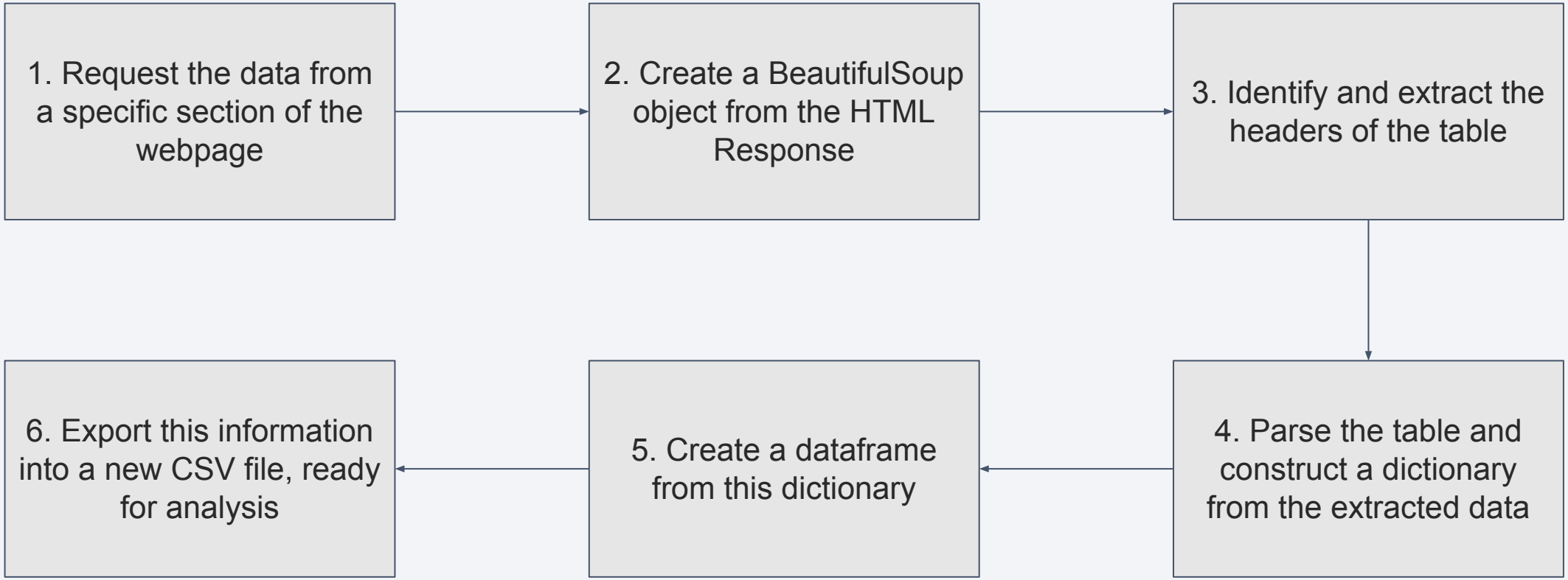
The data was found to be the same in both instances and the API source data was used for analysis.

Data Collection – SpaceX API



[Github Data Collection via API]

Data Collection – Scraping



[GitHub Data Collection via Web Scraping]

Data Wrangling

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;

- True Ocean means the mission outcome was successfully landed to a specific region of the ocean
- 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.

GitHub Data Wrangling

EDA with Data Visualization

- Flight Number vs Payload Mass - Scatter Plot
- Flight Number vs Launch Site - Scatter Plot
- Payload Mass vs Launch Site - Scatter Plot
- Orbit Type vs Success Rate - Bar Chart
- Flight Number vs Orbit Type - Scatter Plot
- Payload Mass vs Orbit Type - Scatter Plot
- Orbit Type vs Annual Success Rate - Line Chart

Scatter plots were used to **identify relationships** between the variables for use in the machine learning algorithm.
Bar charts can show **direct comparisons** between categorical variables.
A Line chart was used to **identify a trend** in the data over a given timeframe.

GitHub EDA with Data Visualisations

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster version 9 v1.1
- List the date of the first successful landing outcome on ground pad was achieved
- List the names of boosters which have had success on drone ship and hav payload mass between 4000 and 6000 kg
- List the total number of both successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes on drone ships, with their booster versions and launch site names for the months of 2015
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Github EDA with SQL

Build an Interactive Map with Folium

Labelled markers were added to the map to indicate the locations of all launch sites. The site locations were:

- Cape Canaveral Space Launch Complex 40 (CCAFS LC-40)
- Vandenberg Space Launch Complex 4 (VAFB SLC-4E)
- Kennedy Space Centre (KSC LC-39A)
- Cape Canaveral Space Force Station (CCAFS SLC-40)

Colour coded cluster markers were added to the map to show the success/failed launches for each site.

- Green means Success
- Red Means Failure

The directions and distances from the nearest transport links to the launch sites were calculated and plotted on the map.

- Closest City
- Closest Railway
- Closest Highway

Github Folium Map

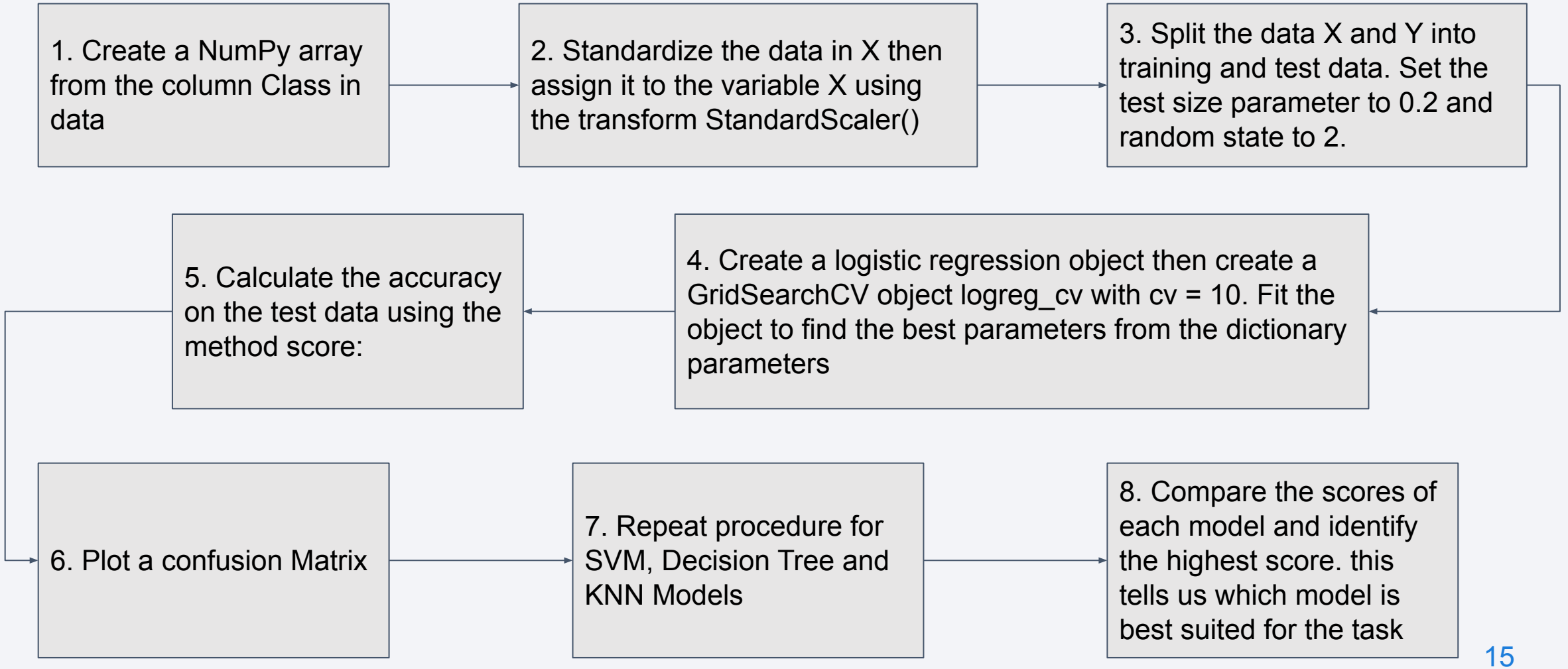
Build a Dashboard with Plotly Dash

To easily convey the insights drawn from the project, a dashboard was developed to include the following:

- Dropdown list to allow user to select information regarding specific launch sites
- Pie Chart to show successful launch rates at each site
- Slider to allow the user to refine their search depending on the payload mass
- Scatter chart of the payload mass vs Success rate for different booster versions

GitHub Plotly Dashboard Notebook

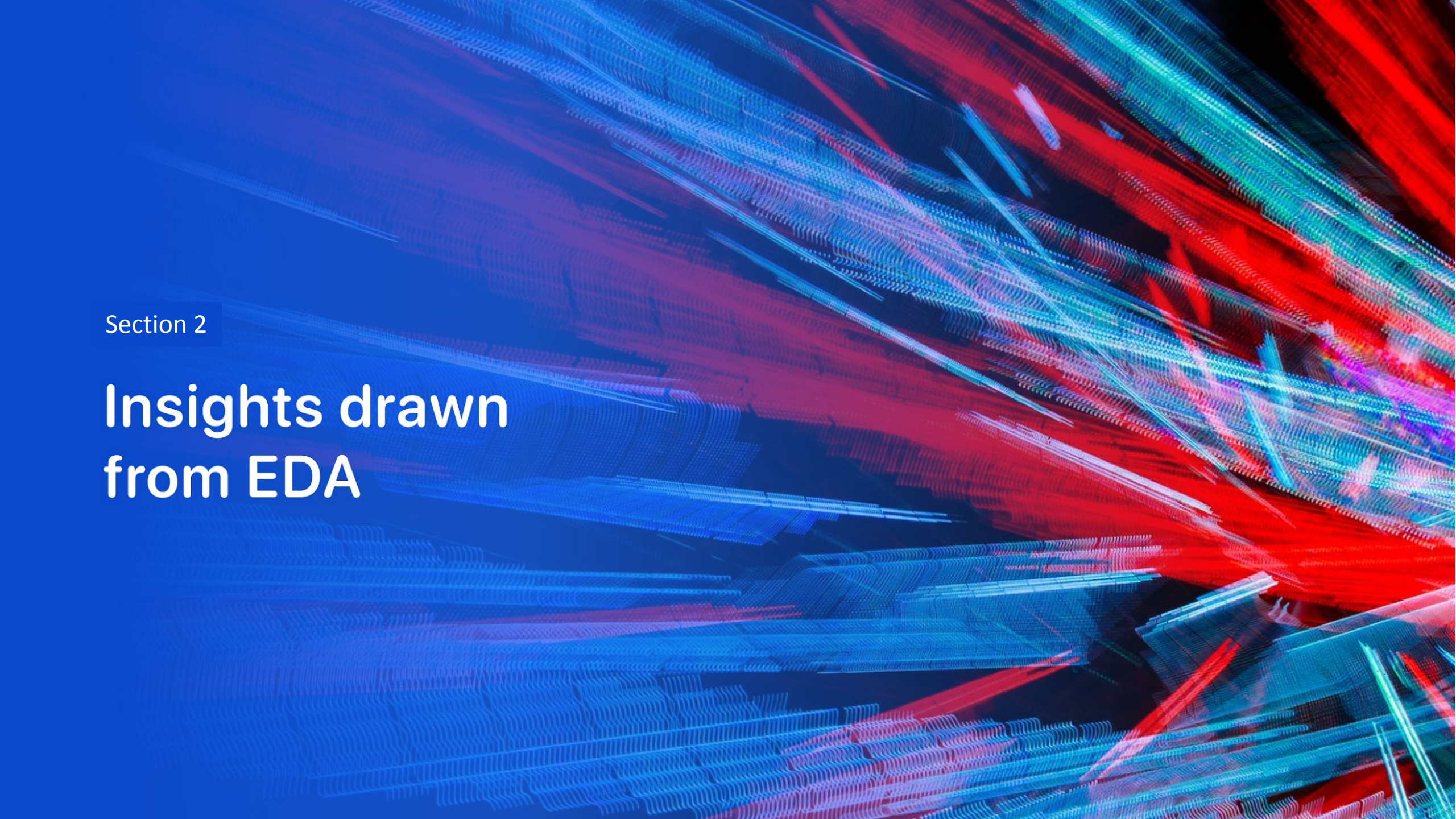
Predictive Analysis (Classification)



[Github ML Prediction]

Section 2

Insights drawn from EDA

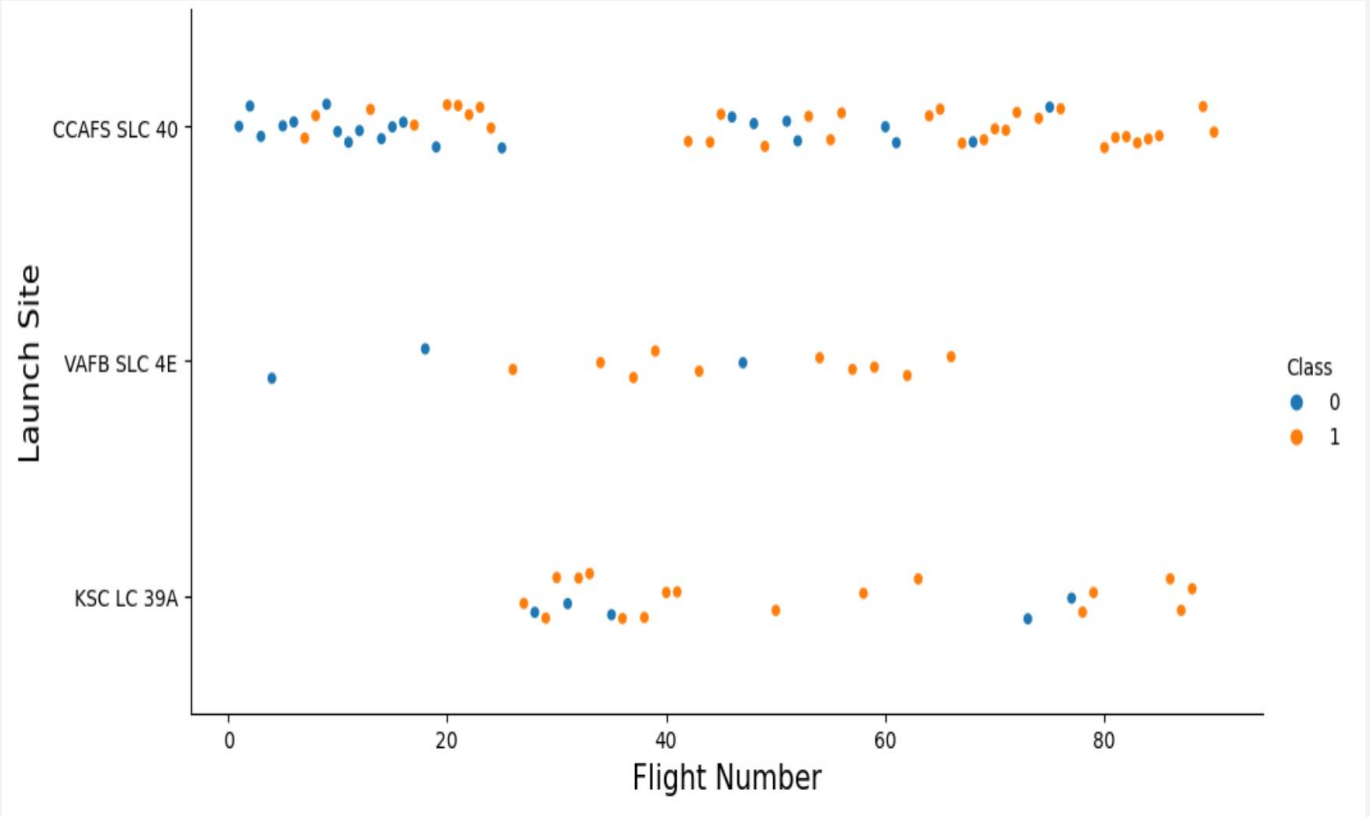


Flight Number vs. Launch Site

This graph shows the flight number and which site it was launched from.

Most launches came from the CCAFS SLC 40 launch site

KSC LC 39A has the highest success rate of the launch sites



The Class legend indicates if the launch was successful or not

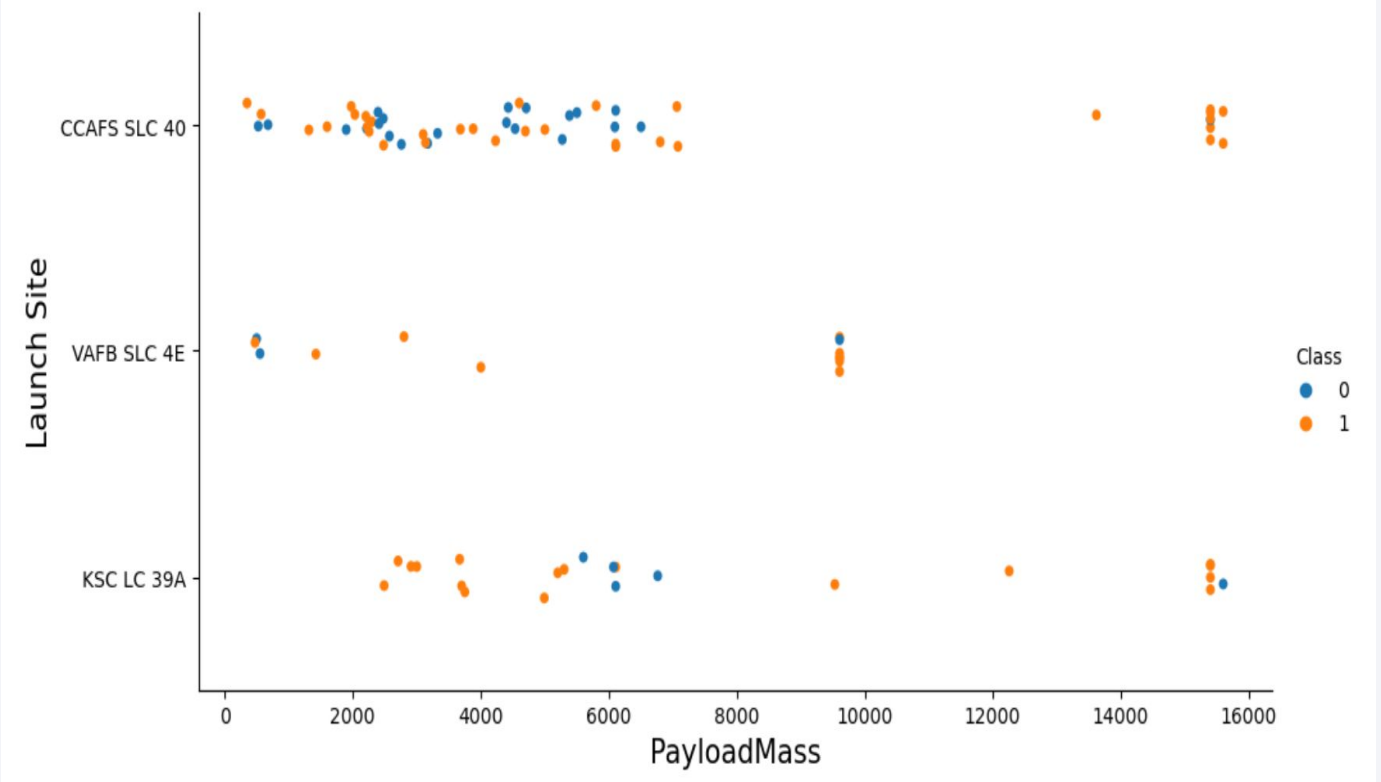
- 0 is a failure
- 1 is a success

Payload vs. Launch Site

This graph shows the the payload mass in kg and which site it was launched from.

As the payload increases, the success rate appears to increase.

This could be due to early launches having a higher failure rate so a smaller payload



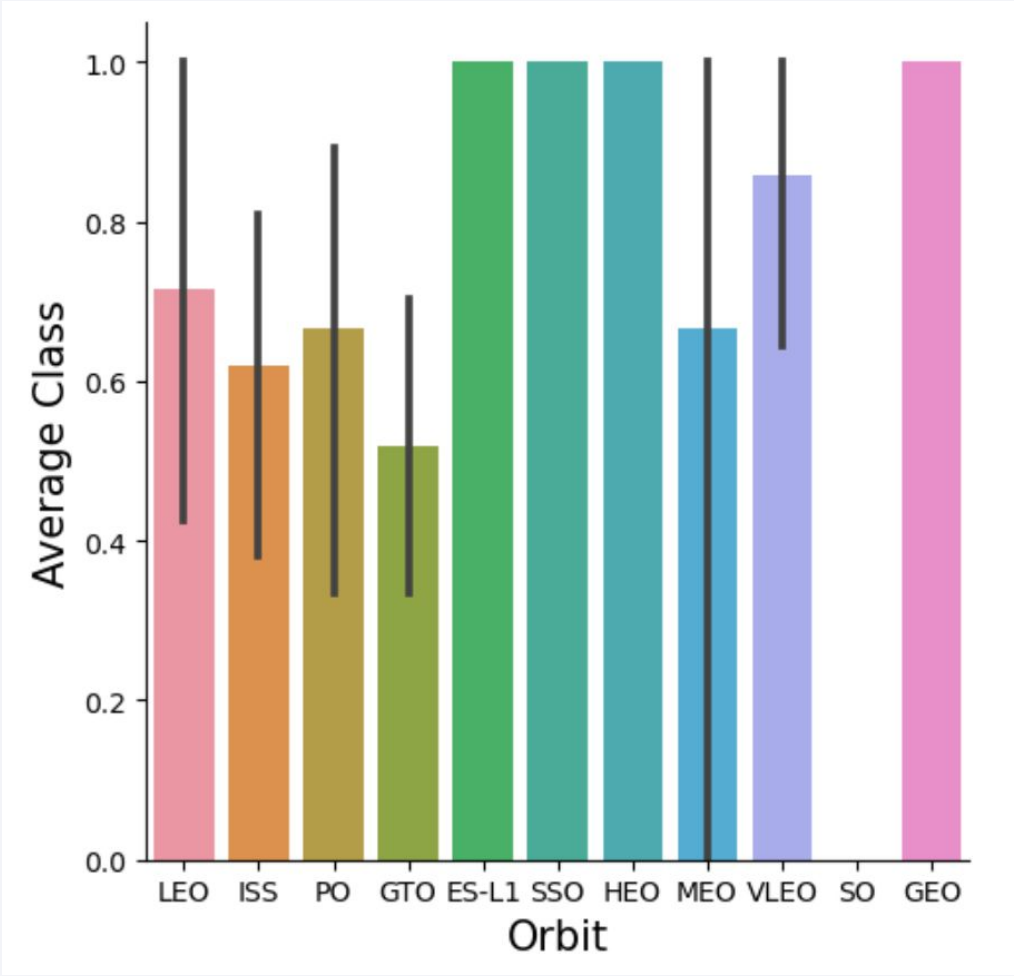
The Class legend indicates if the launch was successful or not

- 0 is a failure
- 1 is a success

Success Rate vs. Orbit Type

This graph shows the average success rate for different orbital objectives for different launches.

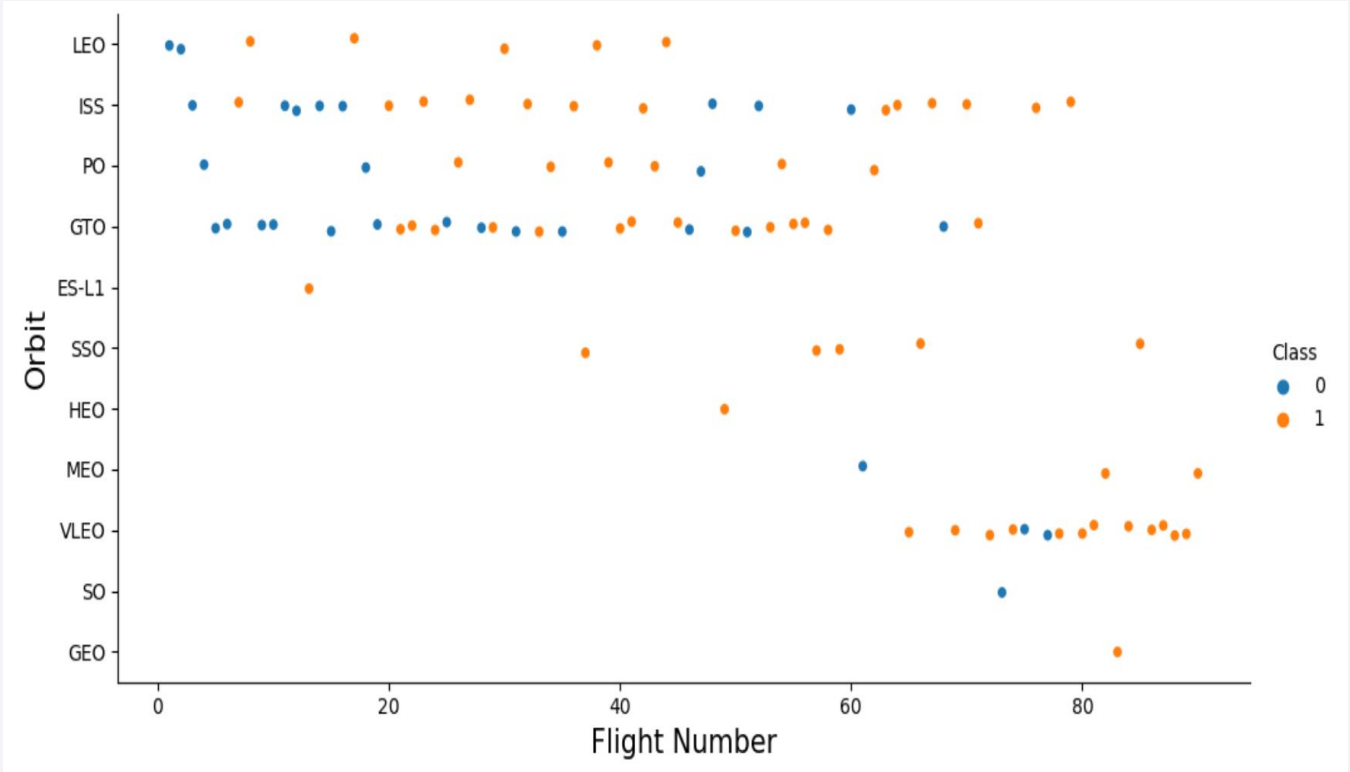
- ES-L1, SSO, HEO and GEO orbits had a 100% success rate
- All launches of SO orbital objectives resulted in failure



Flight Number vs. Orbit Type

This graph shows which flights had specified orbital objectives.

No relationship between flight number and Orbital objective can be determined from this data



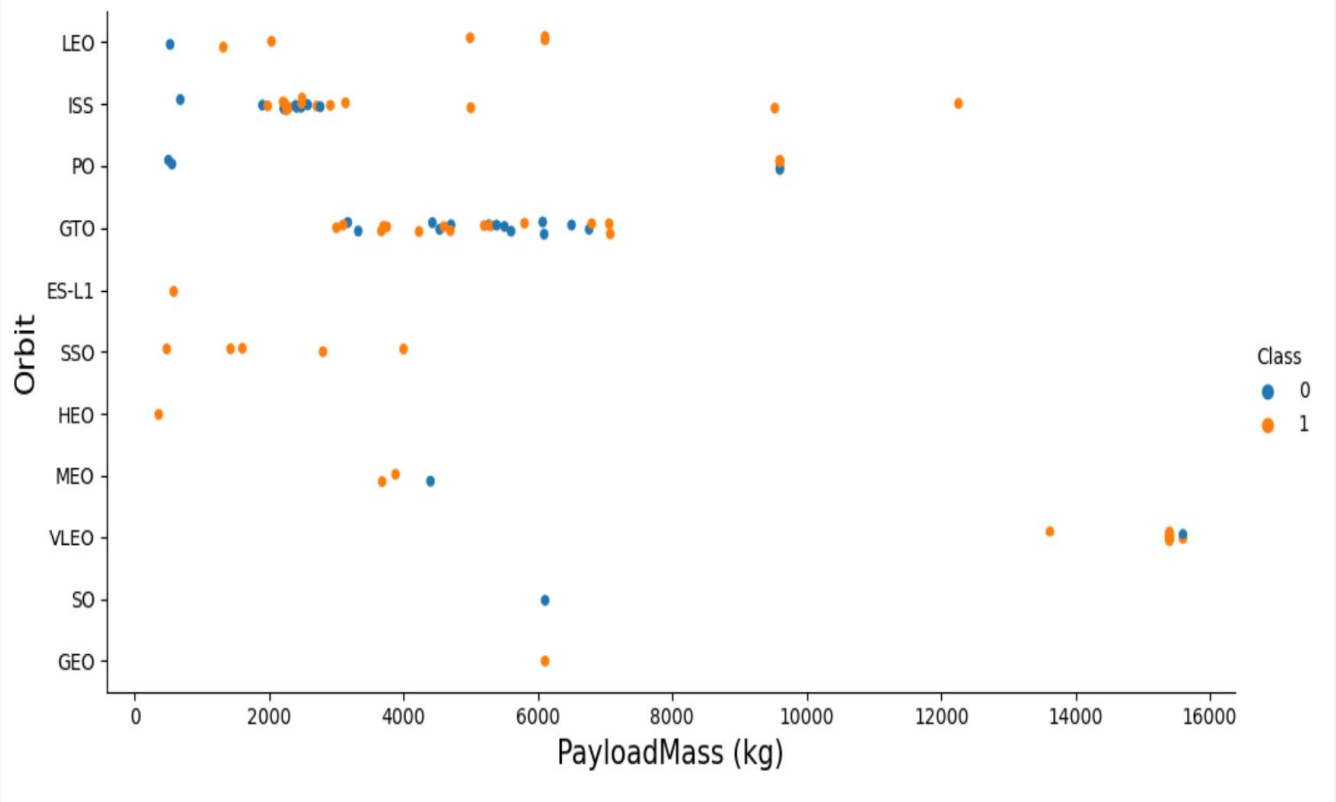
The Class legend indicates if the launch was successful or not

- 0 is a failure
- 1 is a success

Payload vs. Orbit Type

This graph shows the payload of the launch and the orbital objective.

- SSO Orbits have the highest success rates for smaller payloads
- SO orbits have the lowest success rate overall

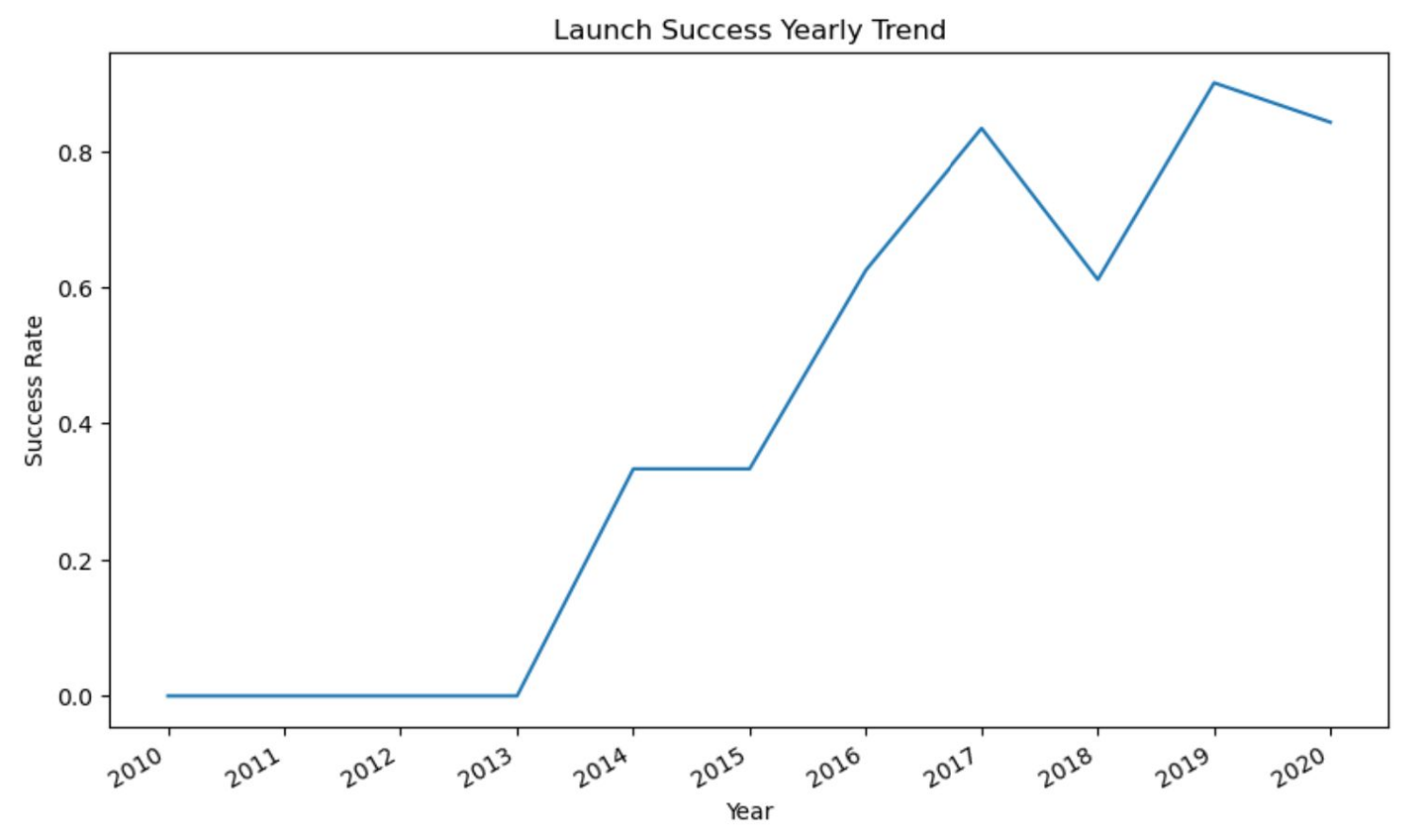


The Class legend indicates if the launch was successful or not

- 0 is a failure
- 1 is a success

Launch Success Yearly Trend

This graph shows an overall increasing trend in the success rate of launches since 2010. This could be due to analysis from the previous missions being used to optimise the next flight



All Launch Site Names

The distinct() command returns only the unique rows from the table

```
%%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

Launch Site Names Begin with 'CCA'

Limit is the command used to limit the search to the first X number of entries that satisfy the search parameters

```
%%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

Total Payload Mass

Total Payload is given by the Sum of the entries in the Payload Mass column, and the search parameter was for the customer of NASA

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL PAYLOAD (KG) BY NASA (CRS)" FROM SPACEXTEL WHERE "CUSTOMER" == "NASA (CRS)";
```

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

TOTAL PAYLOAD (KG) BY NASA (CRS)
45596.0

Average Payload Mass by F9 v1.1

The average payload mass for entries with the Booster Version F9 v1.1

```
%%sql
SELECT "Booster_Version", AVG("PAYLOAD_MASS__KG_") AS "AVERAGE PAYLOAD (KG)"
    FROM SPACEXTBL
        WHERE "Booster_Version" LIKE "F9 v1.1";
```

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

Booster_Version	AVERAGE PAYLOAD (KG)
F9 v1.1	2928.4

First Successful Ground Landing Date

The first successful date is given by the earliest entry with the landing outcome as “Success”

```
%%sql
SELECT MIN("Date") AS "DATE OF FIRST SUCCESSFUL LANDING"
    FROM SPACEXTBL
        WHERE "Landing_Outcome" LIKE "Success";
```

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

DATE OF FIRST SUCCESSFUL LANDING
01/07/2020

Successful Drone Ship Landing with Payload between 4000 and 6000

Search for the booster version which had successful landings on a drone ship with a payload between 4000-6000 kg

```
%%sql
SELECT DISTINCT("Booster_Version") AS  "Booster Version", "Landing_Outcome" AS "Landing Outcome"
    FROM SPACEXTBL
        WHERE "PAYLOAD_MASS__KG_"
            BETWEEN 4000 AND 6000 AND "Landing_Outcome" LIKE 'Success (drone ship)';
```

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

Booster Version	Landing Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Count() provides a number of entries that satisfy the specific search parameters
- this was then grouped by outcome for display purposes

```
%%sql
SELECT "Mission_Outcome" AS "MISSION OUTCOME", count(Mission_Outcome) as "Total"
    FROM SPACEXTBL
    GROUP BY "Mission_Outcome";
```

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

MISSION OUTCOME	Total
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

A list of unique booster versions that carried the maximum payload

```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ AS "PAYLOAD CARRIED"
    FROM SPACEXTBL
        WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

Booster_Version	PAYLOAD CARRIED
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

2015 Launch Records

A list of failed outcomes for the booster version F9 v1.1 for the year 2015. the month in which these failures took place was found

```
%%sql
SELECT Landing_Outcome AS "LAUNCH OUTCOME",
    Booster_version AS "BOOSTER VERSION",
    Launch_site AS "LAUNCH SITE",
    SUBSTR(Date, 4, 2) AS "MONTH"
        FROM SPACEXTBL
            WHERE SUBSTR(Date,7,4)='2015' AND Landing_Outcome = "Failure (drone ship)"
                ORDER BY "MONTH" ASC;
```

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

LAUNCH OUTCOME	BOOSTER VERSION	LAUNCH SITE	MONTH
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	10

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
SELECT Landing_Outcome AS "Landing Outcome", COUNT(Landing_Outcome) AS "Total"
    FROM SPACEXTBL
        WHERE Date BETWEEN '04-06-2010' AND '20-03-2017'
            GROUP BY Landing_Outcome
            ORDER BY "Total" DESC;
```

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

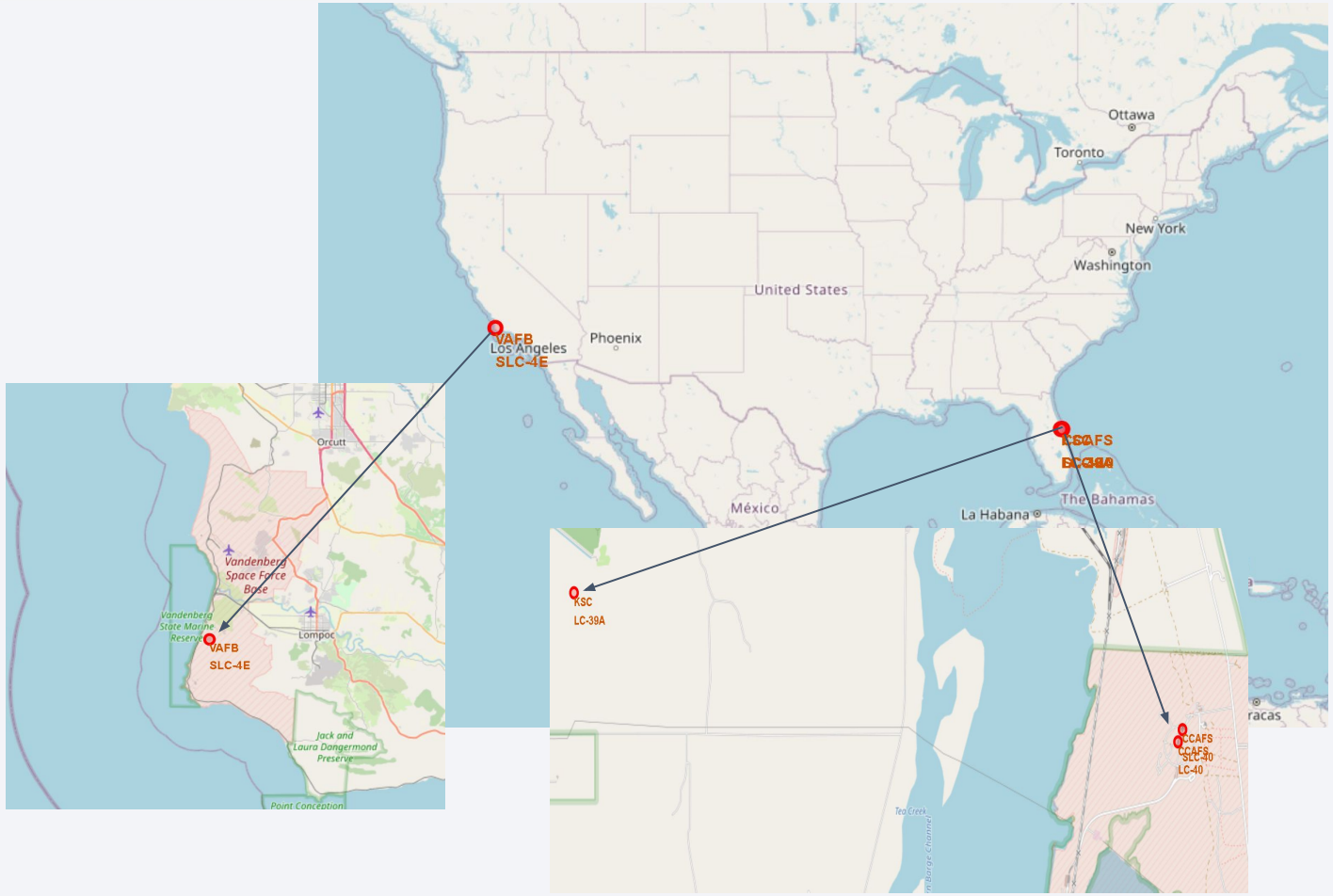
Landing Outcome	Total
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

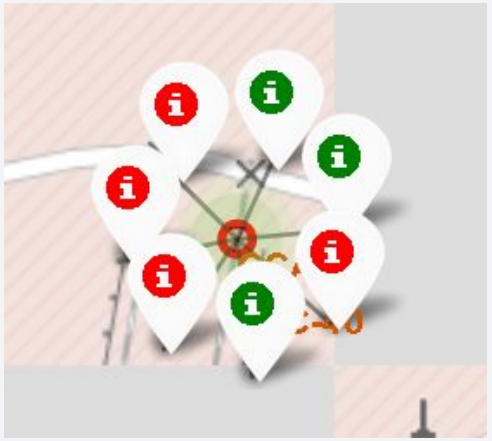
Launch Site Locations

The launch sites are all located close to the equator for optimum weather conditions.

Coastal locations mean any failures in the launch should reduce debris impact on land

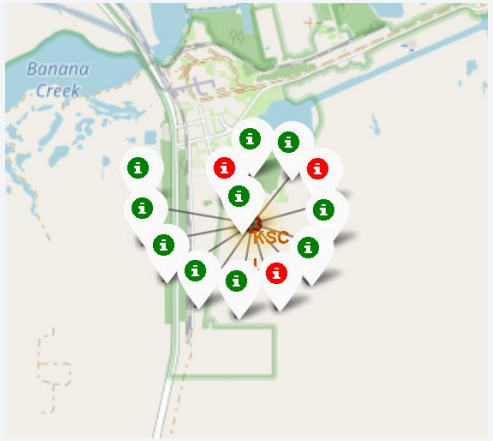


Colour Coded Launches



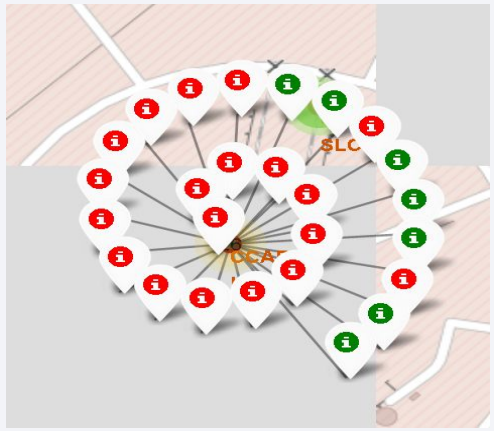
CCAFS SLC-40

3 Successes
4 Failures



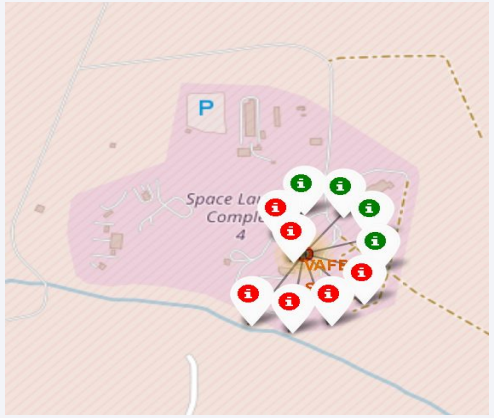
KSC LC 39A

10 Successes
3 Failures



CCAFS LC-40

7 Successes
19 Failures

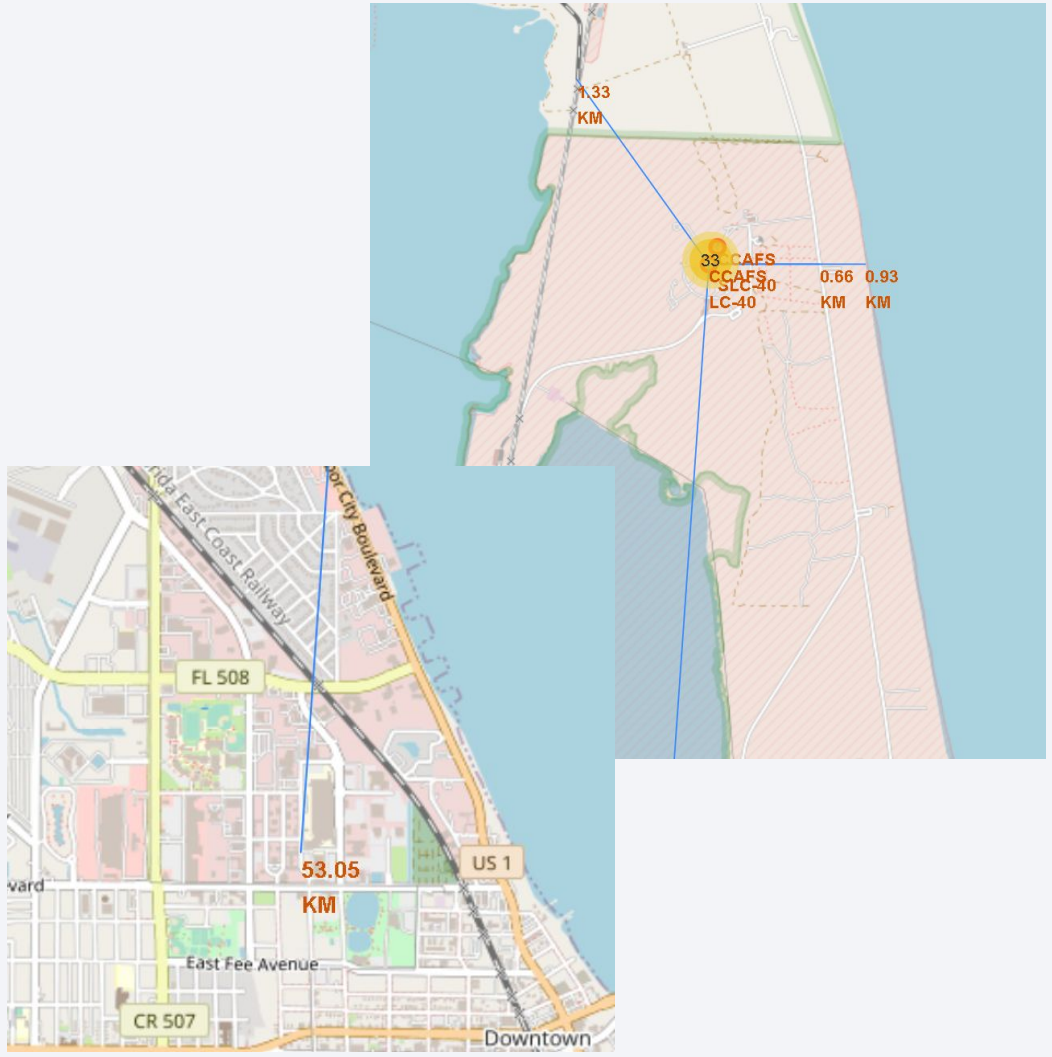


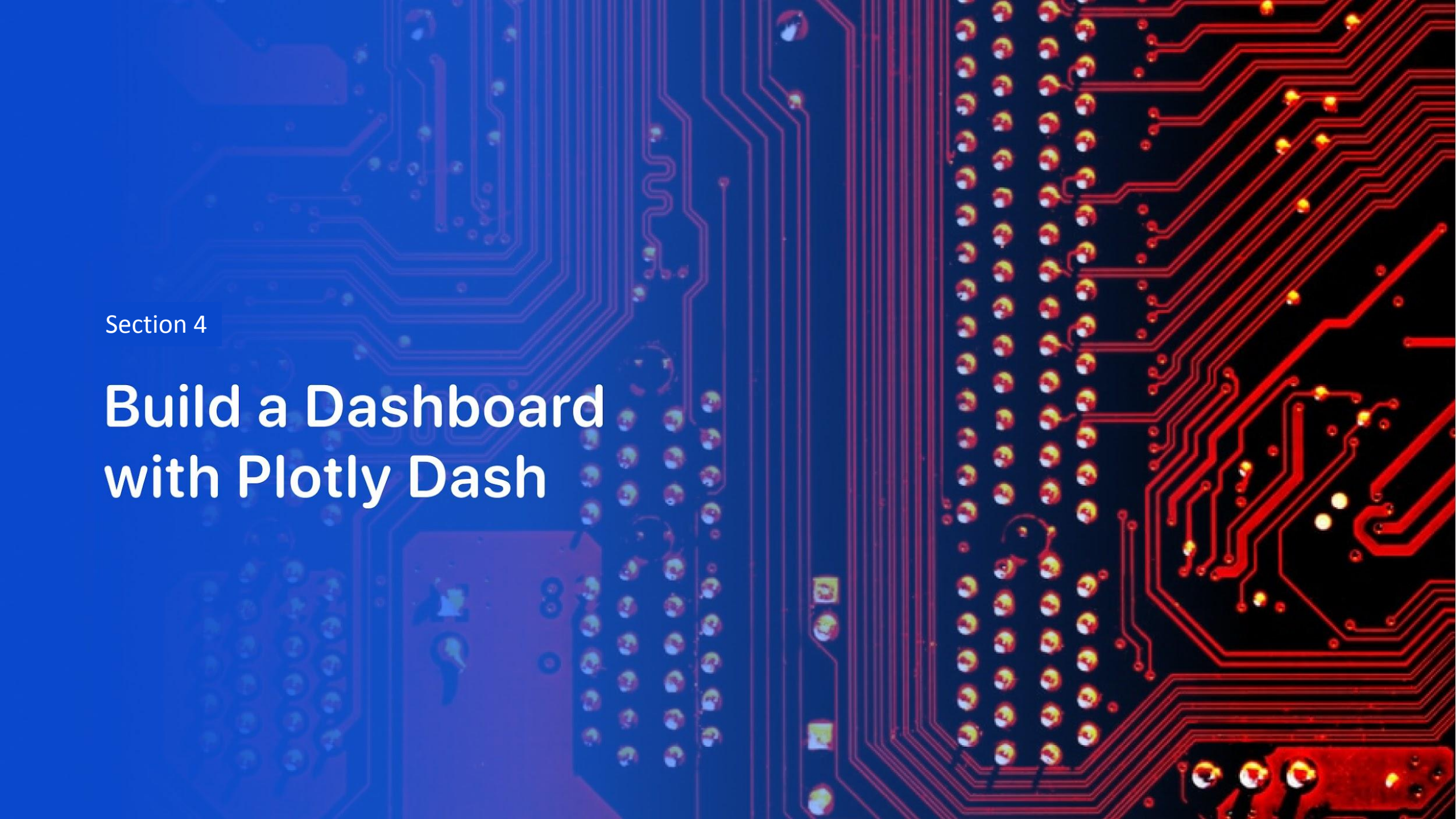
VAFB SLC-fE

4 Successes
6 Failures

Relative Distances

- The nearest railway is 1.33km
- The nearest highway is 0.66km
- The coastline is 0.93km
- The nearest City is Florida (53 km)



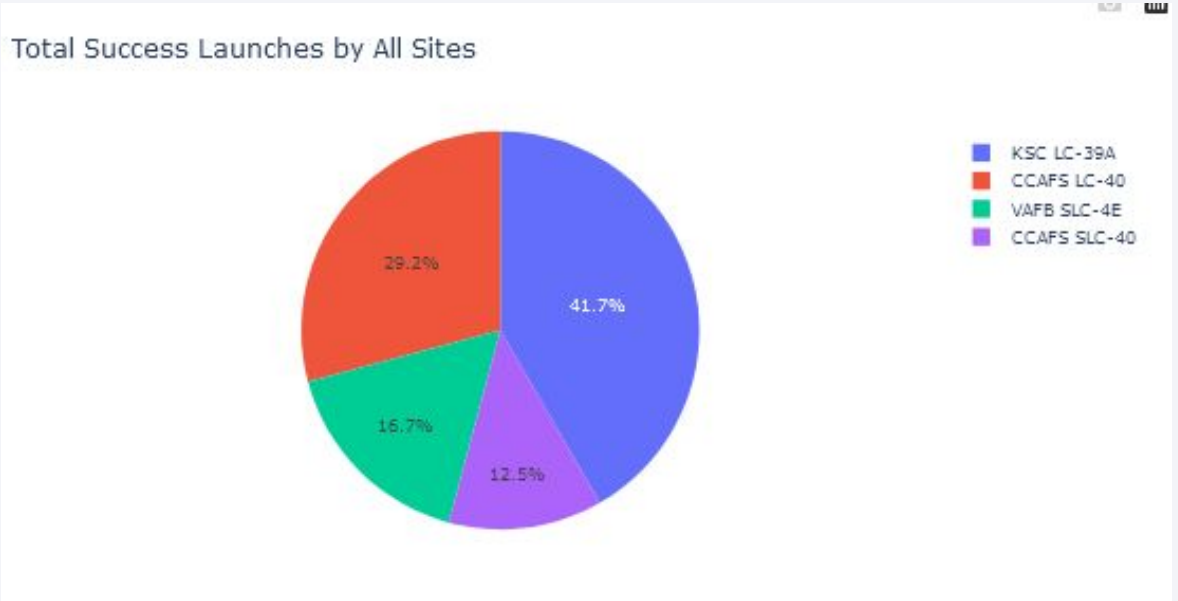

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

This pie chart shows the percentages of the total launches that took place from the 4 sites

Kennedy Space Centre (KSC LC 39A had the most launches with almost half.

CCAFS SLC - 40 had the fewest.



KSC LC-39A Launches

This pie chart shows the proportion of successful to unsuccessful launches for KSC LC 79A Launch Site

Blue indicates successful launch

Red indicates a failure

Payload Mass vs Launch Outcome for All Sites

These charts show the average success rates for payloads between 0kg and 3000kg, and 4000kg and 7000kg.

The colour indicates the Booster Version used for that payload

Section 5

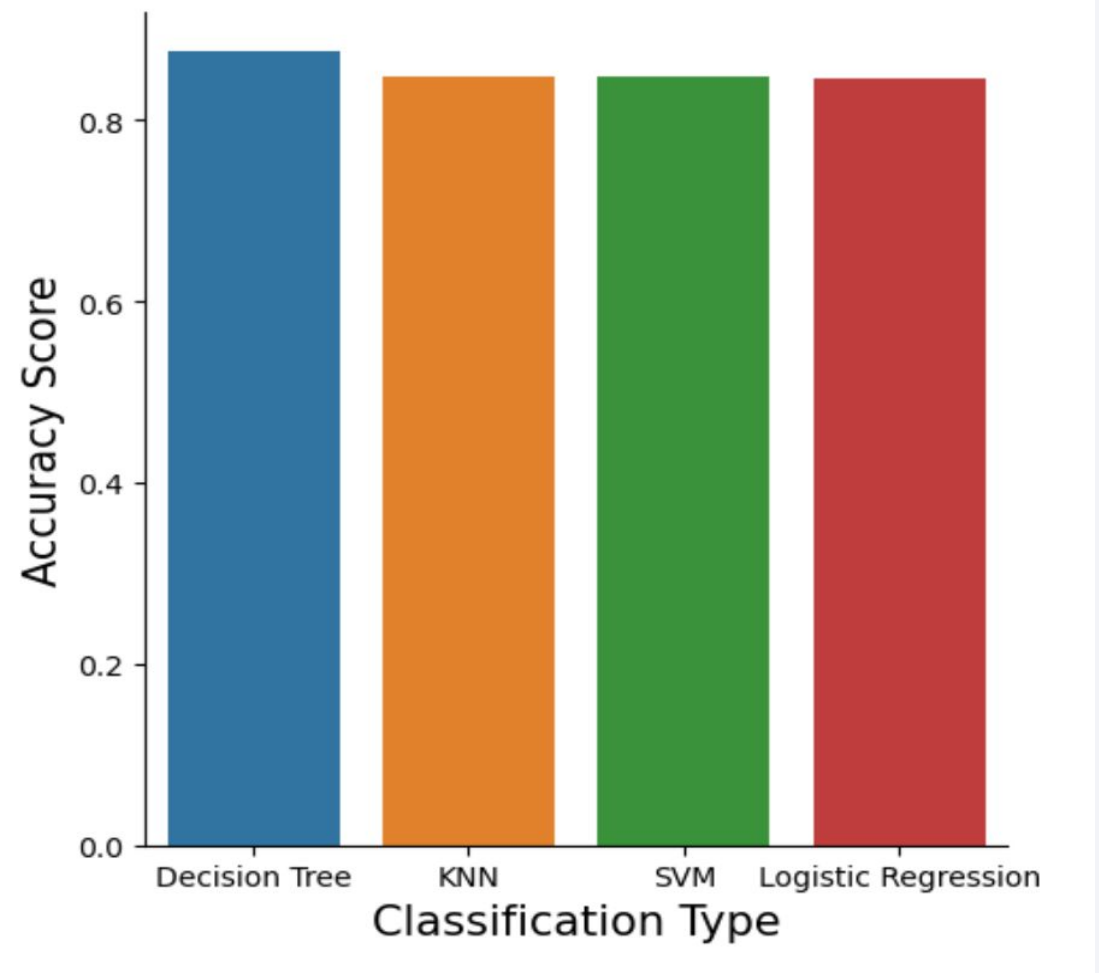
Predictive Analysis (Classification)

Classification Accuracy

The Decision Tree method of Classification was found to be have the highest accuracy score when predicting an outcome based on the test data with a value of 0.875

The Test Data Score appears identical for the all models.

	Classification Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

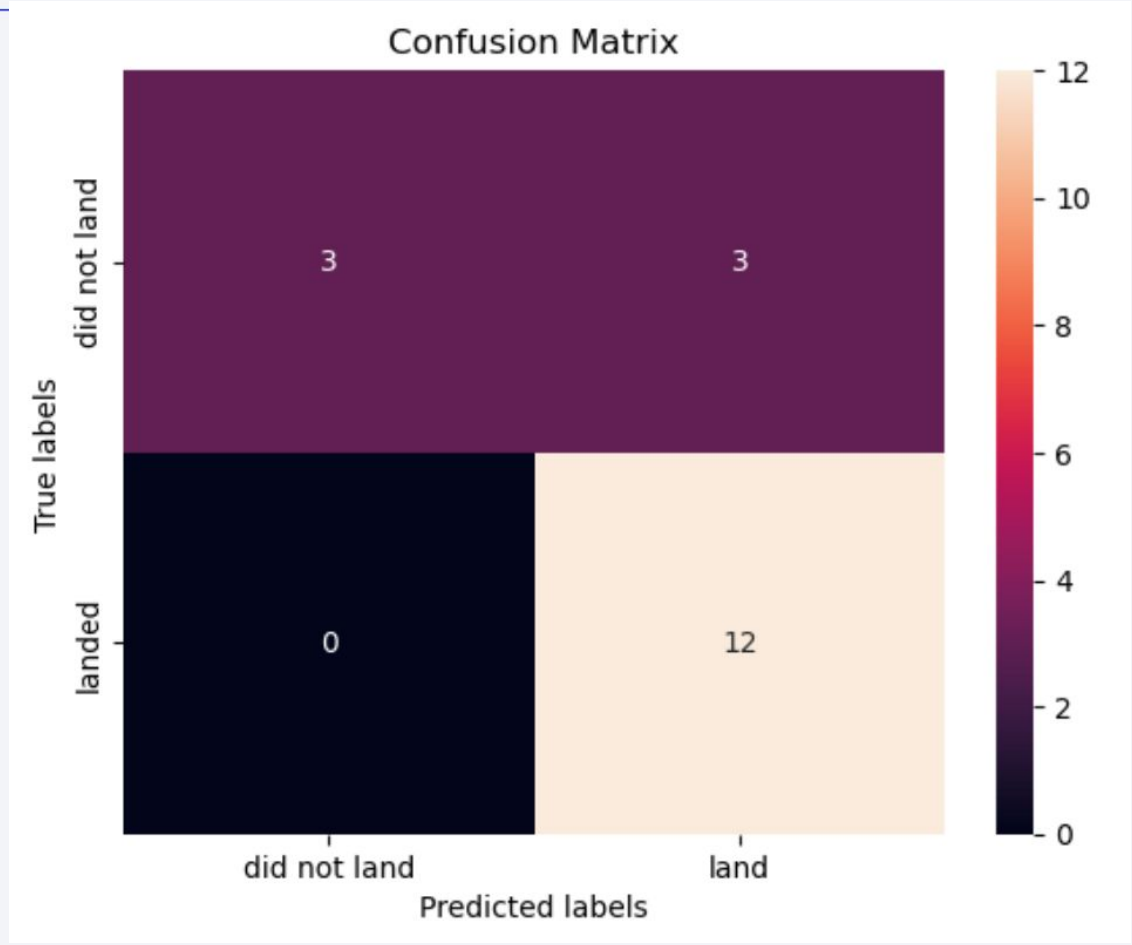


Confusion Matrix - Decision Tree

The highest accuracy model was found to be the decision tree with an accuracy score of 0.875. this means the model correctly predicted the outcome 87.5% of the time.

Of the 18 predictions made the model correctly predicted correctly that 12 would land (True Positive) and 3 would not (True Negative).

The model incorrectly predicted 0 would not land when they did (False Positive), and 3 would land when in reality they didn't (False Negative). This gives an misclassification error of 16.5%.



Conclusions

- Smaller payload launches have a higher success rate.
- KSC LC-39A has the highest success rate for previous launches.
- The success rate of launches has been generally increasing since 2010.
- Launches with the objective of reaching ES-L1, SSO, HEO and GEO orbits have previously had a 100% success rate.
- The Decision Tree model is the best classification method for predicting the outcome of the Falcon 9 booster landing, with an accuracy of 0.875.

Improvements

- Broader Data set is required and further refinement of models to increase the accuracy of predicting success rates.
- Investigation into why multiple runs of the model gives different outcomes occasionally

Appendix

GitHub Repository Link

https://github.com/Genesis273/Falcon-9-Capstone-Project/tree/main

Data Sources

1. https://api.spacexdata.com/v3/launches - SpaceX API

2. https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches - Falcon Heavy Launches

Thank you!