IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marion McGowan
June 6th 2022

Outline

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

Executive Summary

- We have collected data from public SpaceX API and by scrapping a table from the SpaceX Wikipedia page.
- We derived a new variable ‘Class’ which denotes all the successful landings.
- We performed EDA (Exploratory Data Analysis) using SQL and data visualisations using Plotly and Folium
- We performed predictive analysis to reach our results

Introduction

- The aim of this project is to predict if the Falcon 9 first stage will land successfully.
- We investigated the relationship between the various variables and the outcome of the mission (ie. Did the Falcon 9 first stage land successfully?)
- We investigated multiple classification models to predict the outcome of Falcon 9 launches

Section 1

Methodology

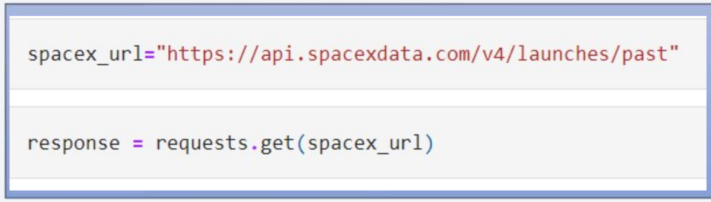
Methodology

- Data collection methodology:
  - SpaceX API
  - Webscraping Wikipedia table
- Perform data wrangling
  - Clean data, removing missing or irrelevant data to ensure all data is relevant to the Falcon 9. Create a label called ‘Class’ to identify successful outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

Github Link

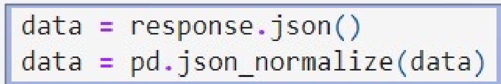
Request and Parse the launch Data using GET Request



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

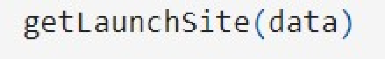
```
response = requests.get(spacex_url)
```

Normalize and Filter Data



```
data = response.json()
data = pd.json_normalize(data)
```

- Remove data for launches where cores > 1
- Use custom functions to return variables, like Launch Site →

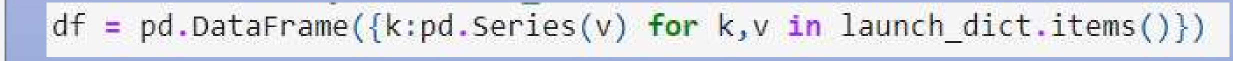


```
getLaunchSite(data)
```

- All columns were added to a dictionary, launch_dict

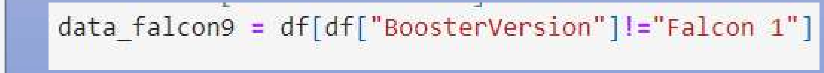
Create DataFrame

- The dictionary was used to create a DataFrame



```
df = pd.DataFrame({k:pd.Series(v) for k,v in launch_dict.items()})
```

- Filter Data Frame to include only Falcon 9 Launches



```
data_falcon9 = df[df["BoosterVersion"]!="Falcon 1"]
```

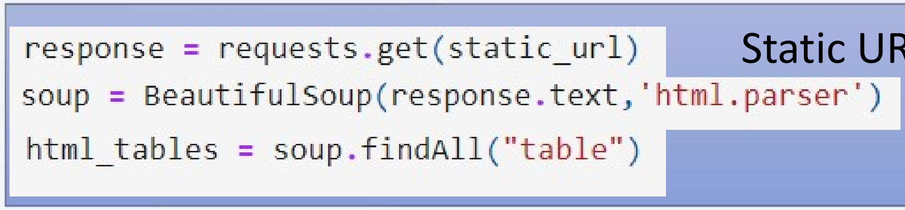
- Export to CSV

Classified as Public (Green)

Data Collection - Scraping

Github Link

Use GET request to create a BeautifulSoup object

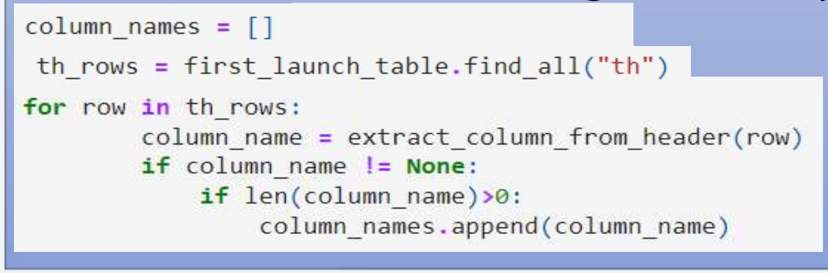


```
response = requests.get(static_url)
soup = BeautifulSoup(response.text,'html.parser')
html_tables = soup.findAll("table")
```

Static URL = [Wikipedia Link]()

Parse BeautifulSoup data

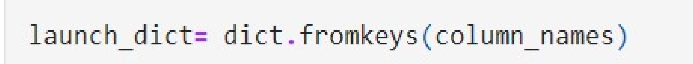
Extract column names using a FOR loop



```
column_names = []
 th_rows = first_launch_table.find_all("th")
for row in th_rows:
        column_name = extract_column_from_header(row)
        if column_name != None:
            if len(column_name)>0:
                column_names.append(column_name)
```

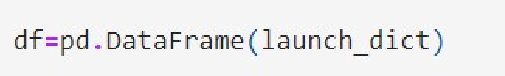
Create DataFrame

Create a dictionary launch_dict



```
 launch_dict= dict.fromkeys(column_names)
```

After launch_dict was populated with data, we created our df



```
df=pd.DataFrame(launch_dict)
```

Data Wrangling

Github Link

- Investigate data types, value counts and missing data points. Launches missing payload mass were assigned the mean payload mass.
- Derived a new variable, Class, which represents the success/failure of booster landing. Bad landing outcomes are assigned a Class of 0, as shown in the table below.

Class	Outcome	Landing Outcomes
0	Bad Outcome	False ASDS, False Ocean, False RTLS, None ASDS, None None
1	Good Outcome	True ASDS, True RTLS, True Ocean

- This will be our training label, with Class == 1 identifying launches where the Falcon 9 first stage landed successfully

EDA with SQL

Github Link

SQL was used to explore the data further. We used clauses and aggregation functions to gain insights, including:

- Fetch the names of the unique launch sites in the space mission
- Find the total and average payload mass carried by Booster Versions
- List the date of the first successful landing outcome in ground pad
- 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 maximum payload mass.
- List the failed landing_outcomes in drone ship, in the year 2015
- Get the count of each landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

EDA with Data Visualization

Github Link

Several scatter plots were created to determine if there was any noticeable correlation between the variables plotted.

1. Flight Number vs. Launch site scatter chart
2. Payload vs. Launch site scatter chart
3. Flight Number vs. Orbit type scatter chart
4. Payload vs. Orbit type scatter chart

When plotting scatter charts, the data points were coloured by the launch outcome (or Class), to visualise the relationship between these variables and successful launches.

Mean Success Rate vs. Orbit type bar chart was created to investigate relationship between these variables.

A trend line was also drawn to observe the pattern of successful launches over a number of years

Build an Interactive Map with Folium

Github Link

A Folium Interactive Map was created and all launch sites were identified on the map, with a circle and marker added to each location

The variable Class was used to add a ClusterMarker to each launch site. This helped us visualise the successes and failures at each launch site.

The map allowed us to identify infrastructure required for launch sites, and the distance from proximities to the site

This will help us identify the criteria for a successful launch site

Build a Dashboard with Plotly Dash

Github Link

I built a Dashboard using Plotly Dash to display the following interactive visualizations:

- Pie Chart illustrating the proportion of successful launches by launch site, or of all sites
- Scatter Plot showing the relationship between Payload Mass and Launch Outcome for different Payload Masses

Predictive Analysis (Classification)

Github Link

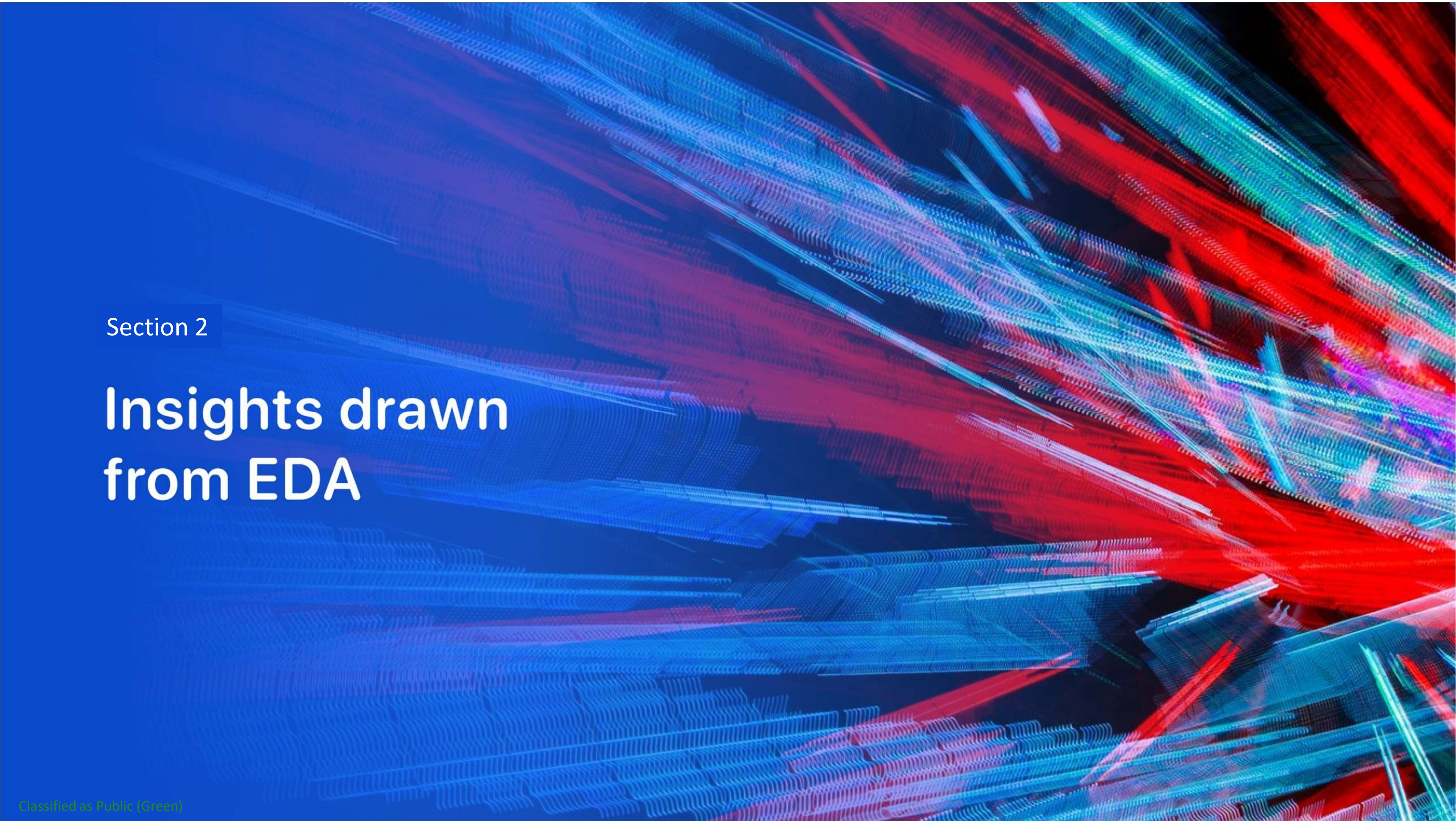
- Standardised the data and split it into test and train datasets
- Test size was 0.2 and we had 18 samples
- Set parameters and created a GridSearchCV object for the classification method
- Fit our data to the object and trained the model
- Checked the accuracy of each model

Classification Methods
Logistic Regression
Support Vector Machine
Decision Tree Classifier
K Nearest Neighbours

Results

In this presentation, we will look at:

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

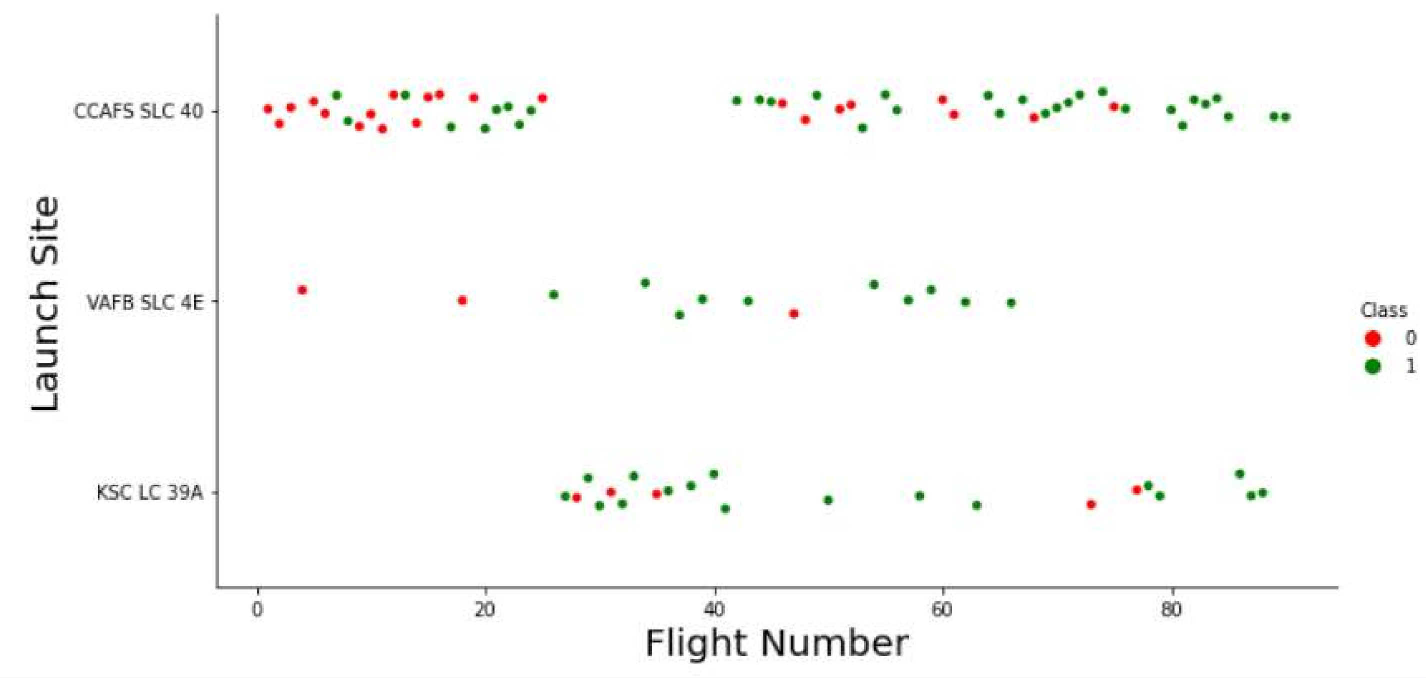


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

The majority of launches occur at CCAFS SLC 40, but it appears that the KSC LC-39A site might have the higher proportion of successful launches across all sites

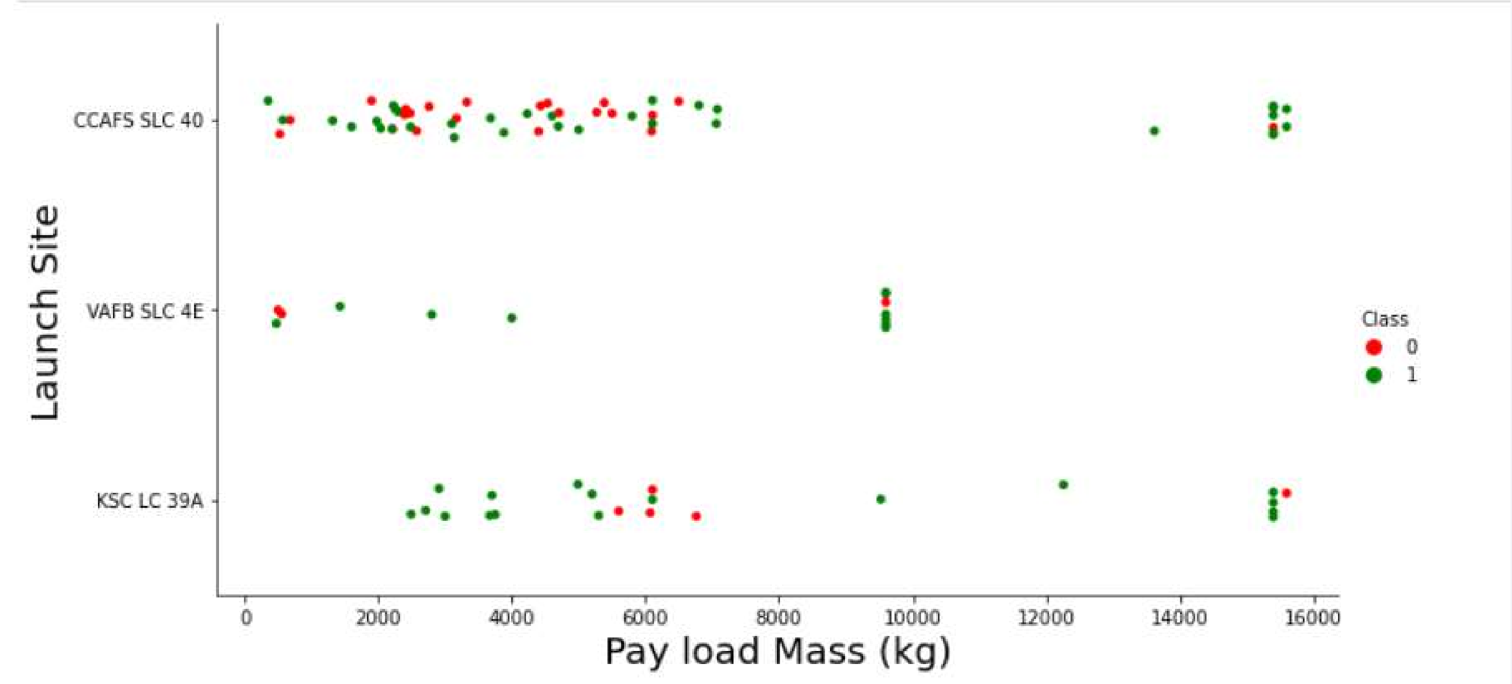


Classified as Public (Green)

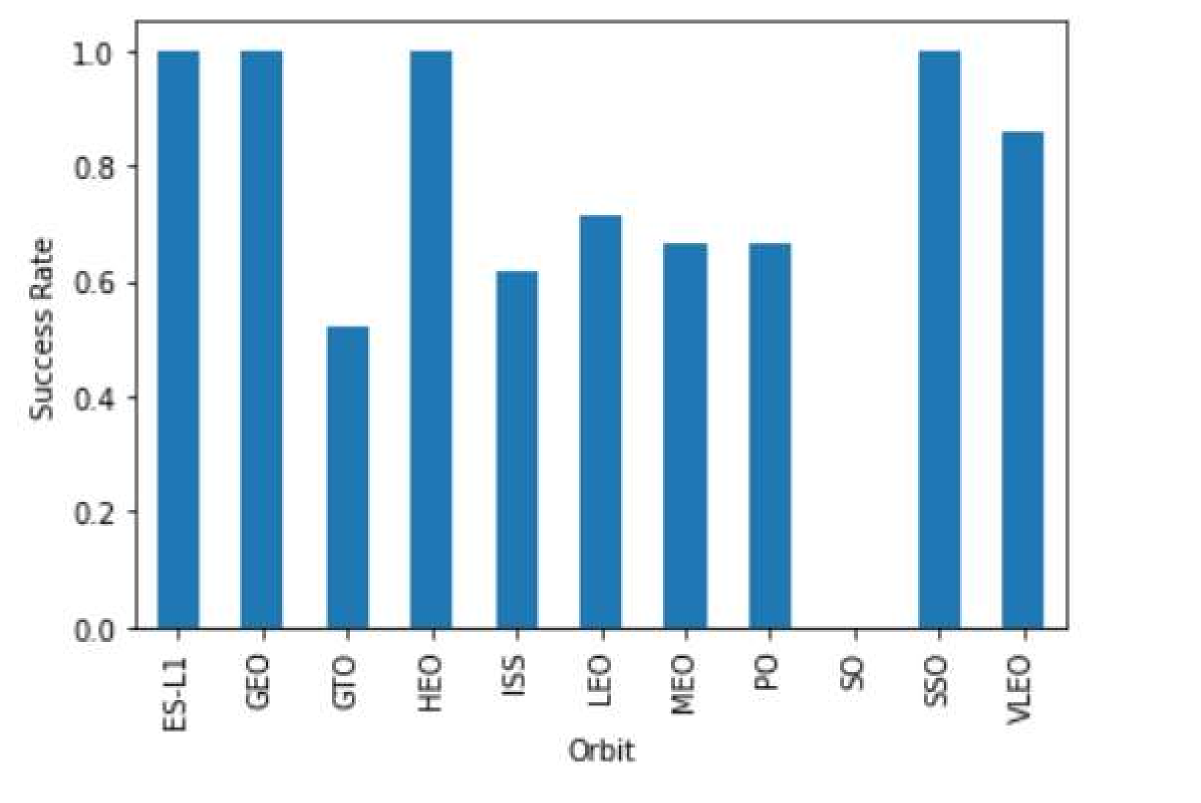
Payload vs. Launch Site

Most launches have a Payload Mass of <6000kg

It is important to note, that there is only a small number of data points for launches >6000kg in our dataset



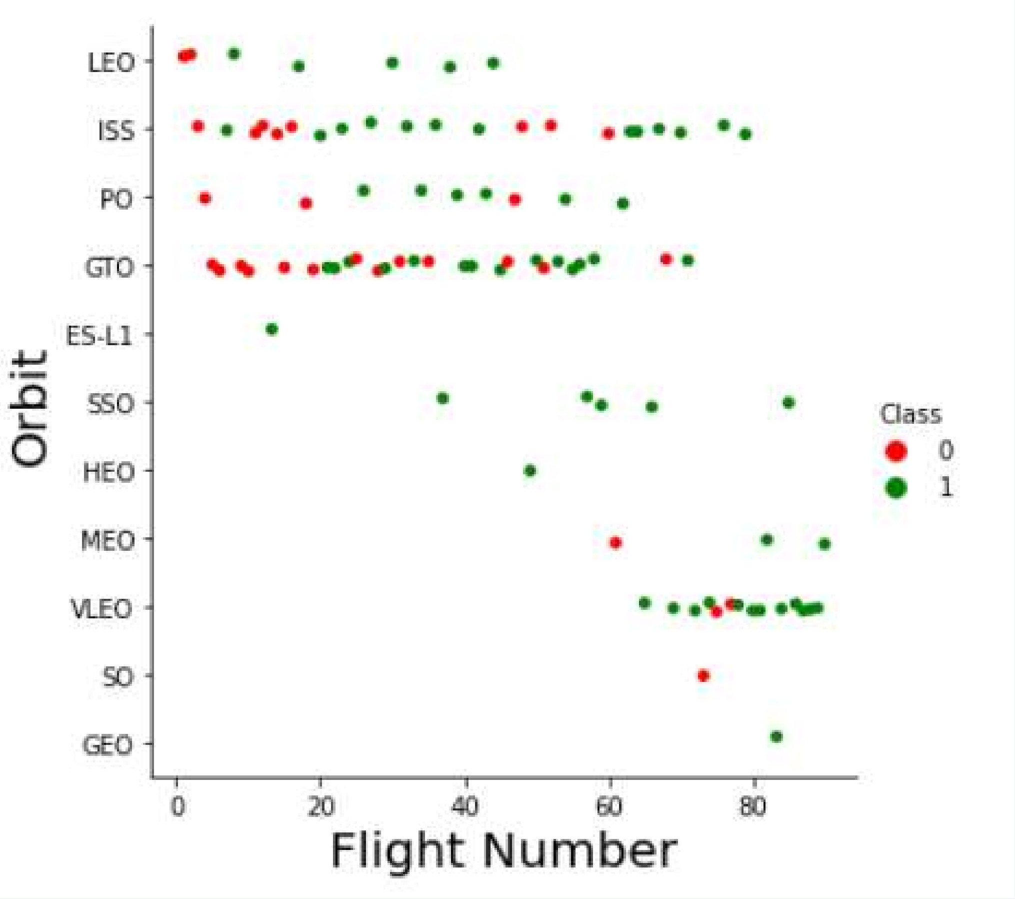
Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO are orbit types with a 100% landing outcome success rate
- However, we'll see on the next slide that this might be a result of the very small number of launches using these orbits

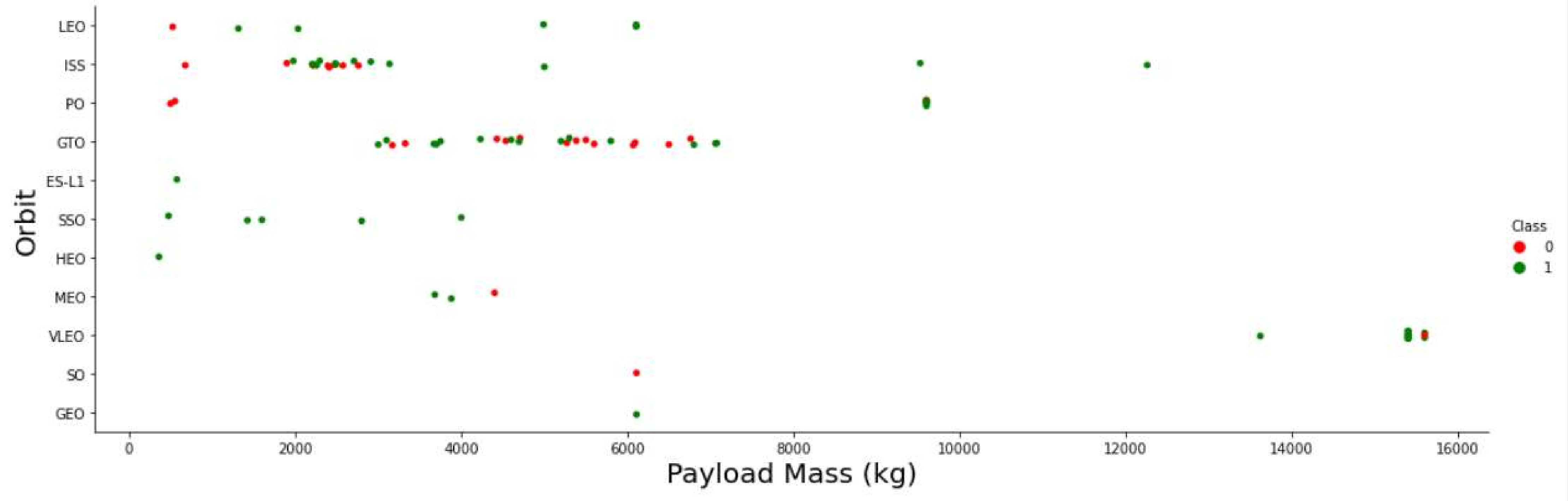
Flight Number vs. Orbit Type

- Note that our DataFrame was created in such a way that Flight Number directly corresponds to Date of Launch
- GTO, ISS, and VLEO Orbits appear to be the most common orbit types
- They have all had failures, but they also have had more successful outcomes that orbit types with a 100% success rate



Classified as Public (Green)

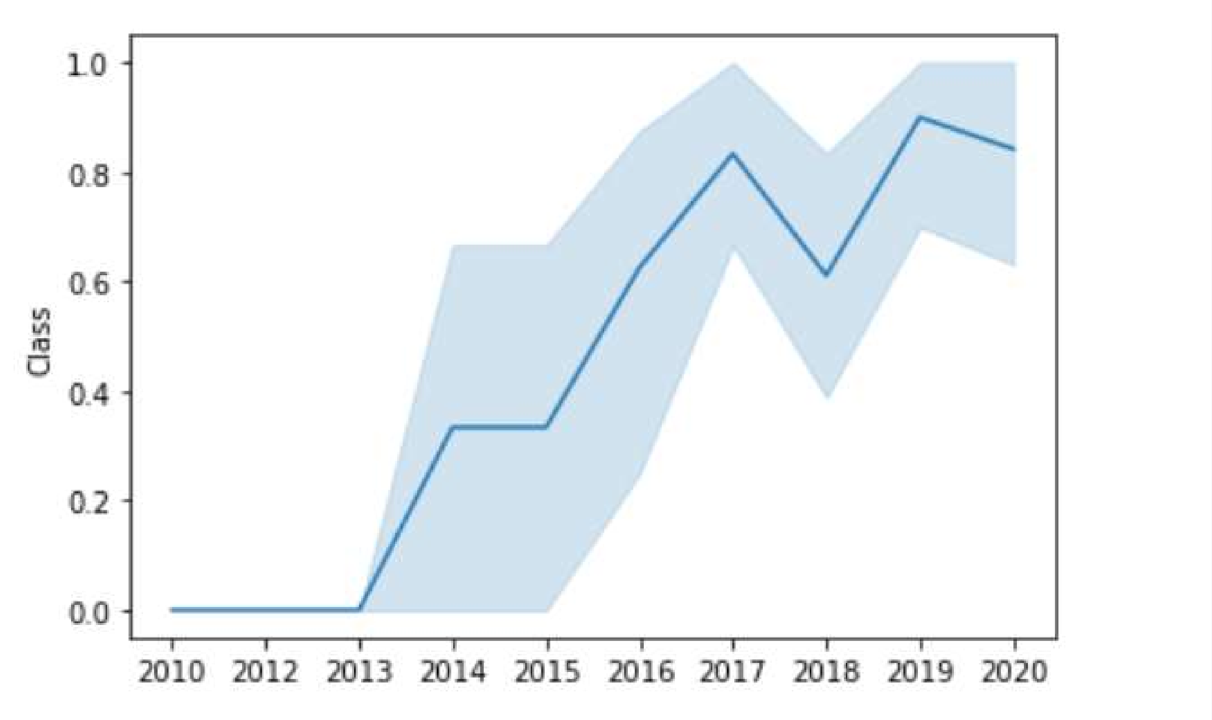
Payload vs. Orbit Type



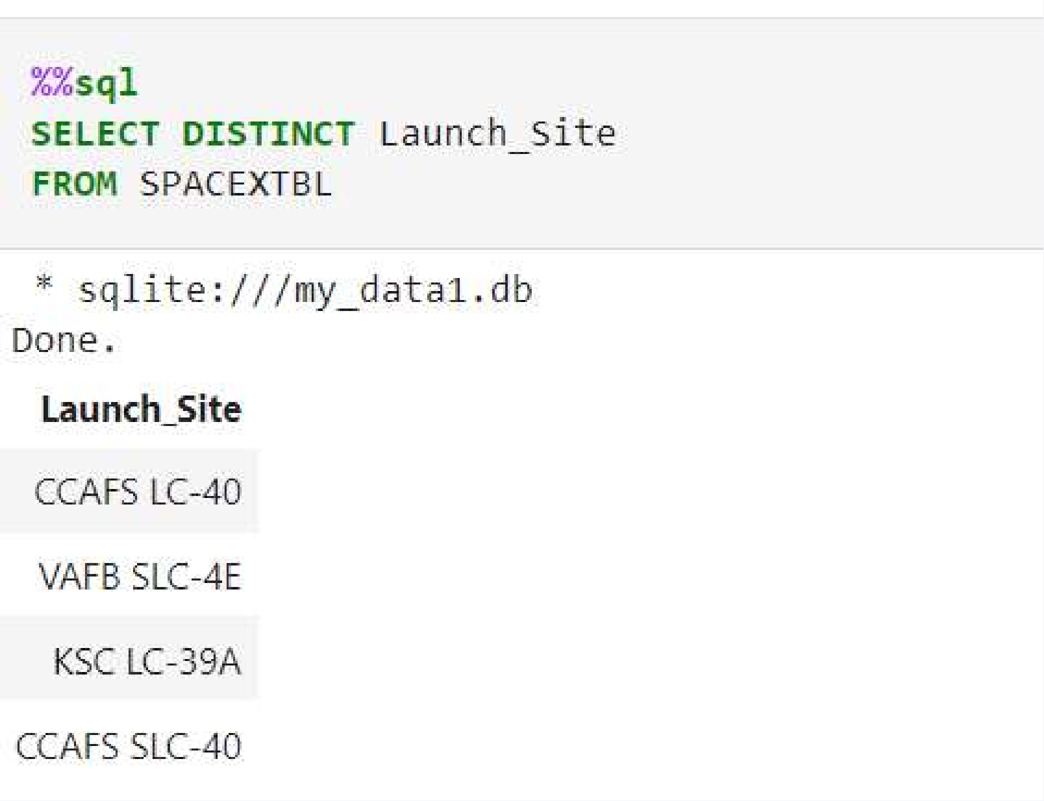
It appears that smaller payloads are more likely to be successful, but there are not a lot of data points with payload mass >7000kg

Launch Success Yearly Trend

- The success rate of launches increased between 2013 and 2020
- It is not a stable increase. We can see several dips in the percentage of successful launches, in 2018 and 2020



All Launch Site Names



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

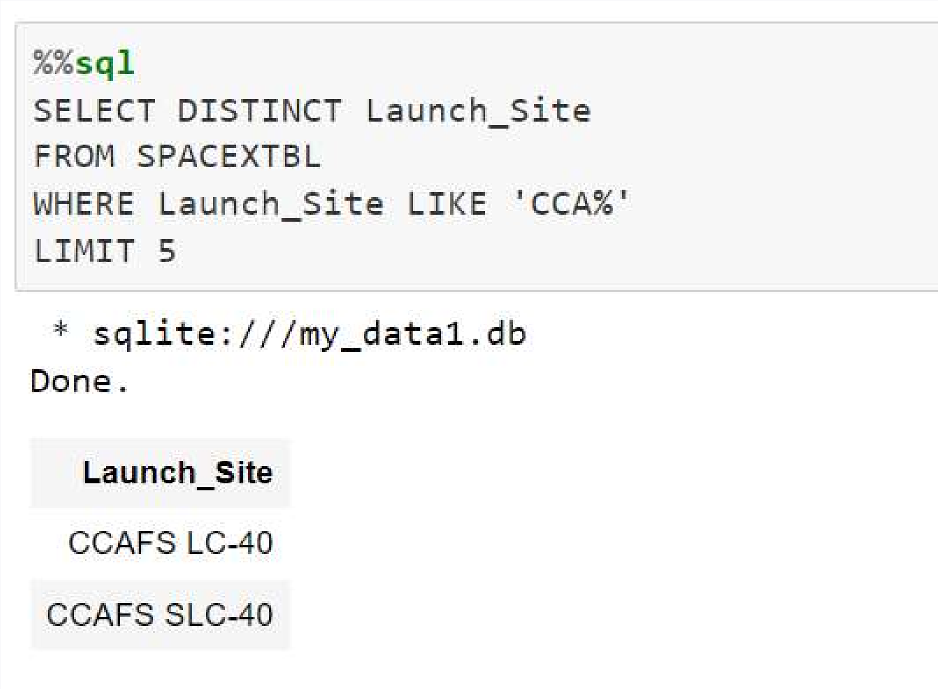
```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

SpaceX uses four launch sites

- CCAFS LC-40
- VAFB SLC-42
- KSC LC-39A
- CCAFS SLC-40

Launch Site Names Begin with 'CCA'



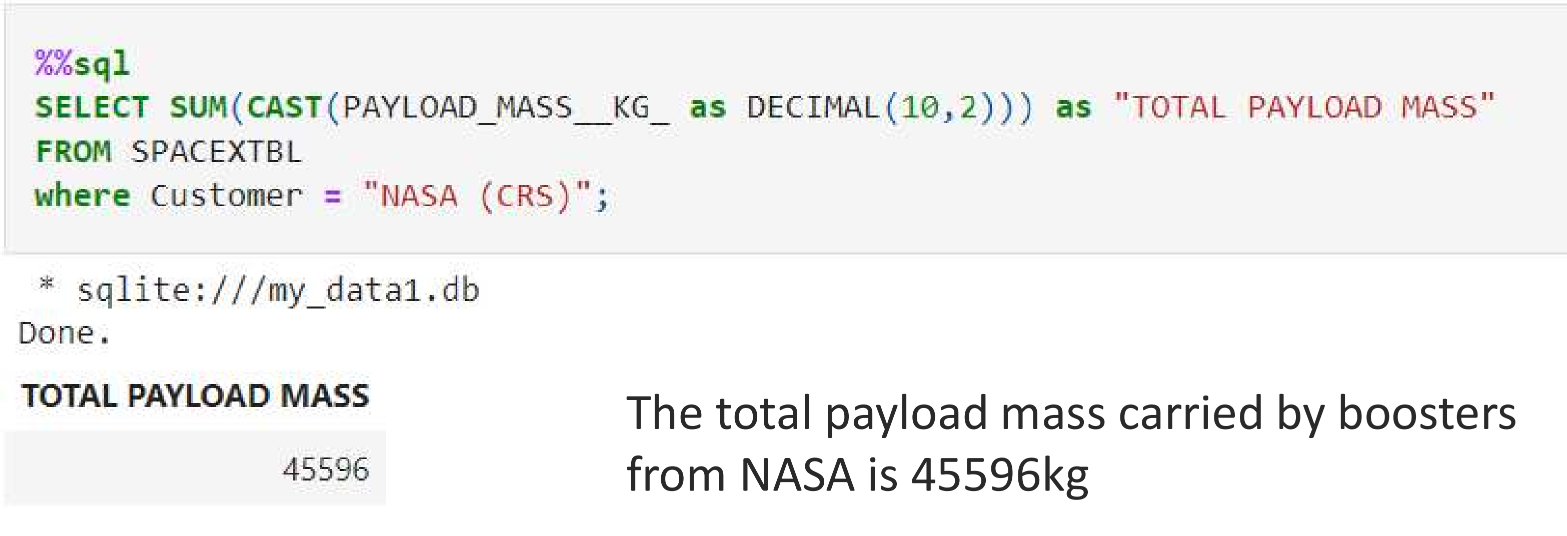
```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

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

Launch_Site
CCAFS LC-40
CCAFS SLC-40

- Queried launch site names beginning with ‘CCA’
- I looked at the first 5 records, before specifying distinct values only
- There are two launch sites that begin with ‘CCA’

Total Payload Mass



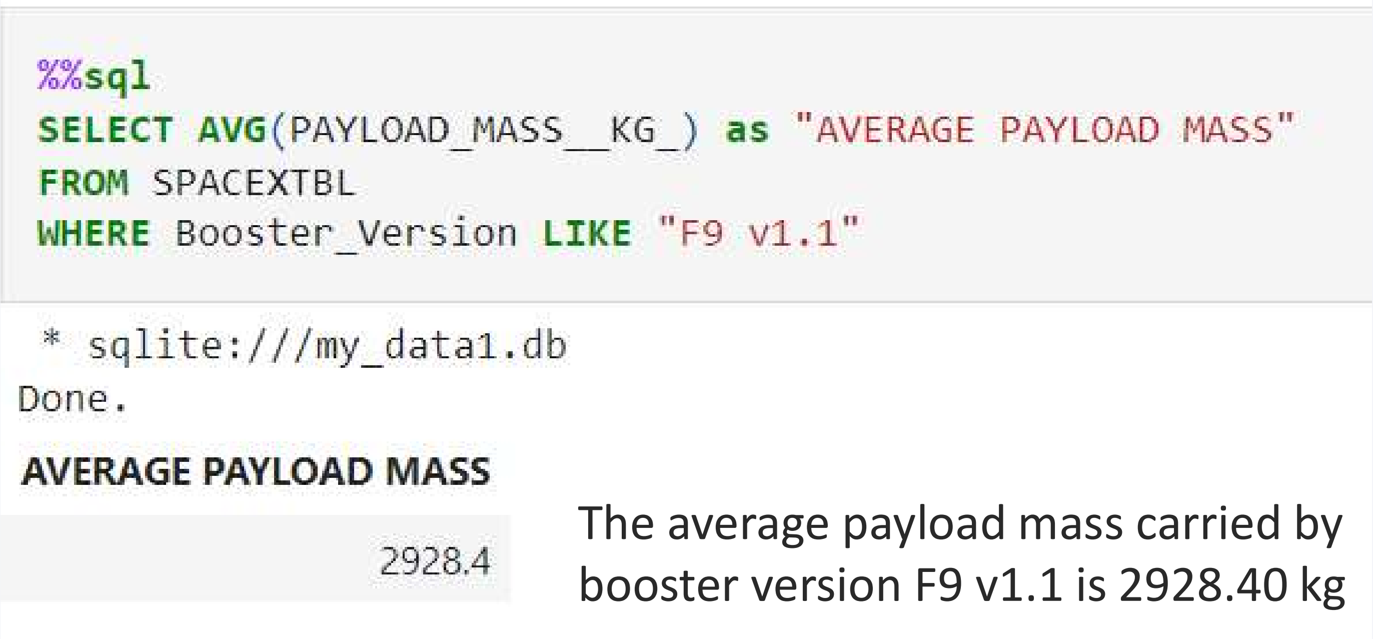
```
%%sql
SELECT SUM(CAST(PAYLOAD_MASS__KG_ as DECIMAL(10,2))) as "TOTAL PAYLOAD MASS"
FROM SPACEXTBL
where Customer = "NASA (CRS)";
```

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

TOTAL PAYLOAD MASS
45596

The total payload mass carried by boosters from NASA is 45596kg

Average Payload Mass by F9 v1.1



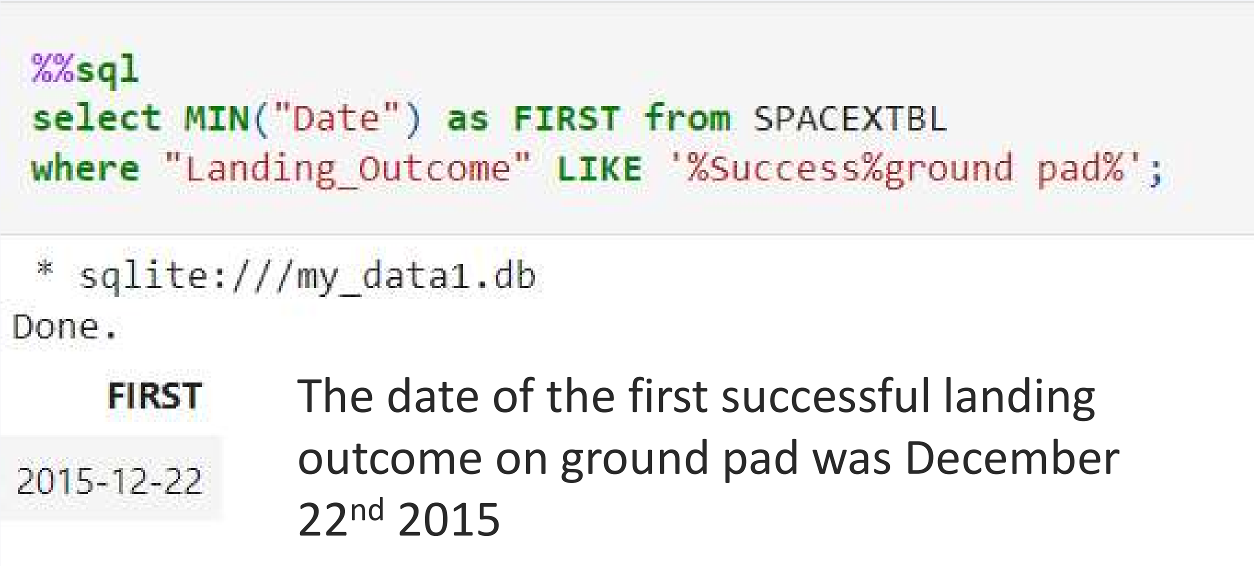
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as "AVERAGE PAYLOAD MASS"
FROM SPACEXTBL
WHERE Booster_Version LIKE "F9 v1.1"
```

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

AVERAGE PAYLOAD MASS
2928.4

The average payload mass carried by booster version F9 v1.1 is 2928.40 kg

First Successful Ground Landing Date



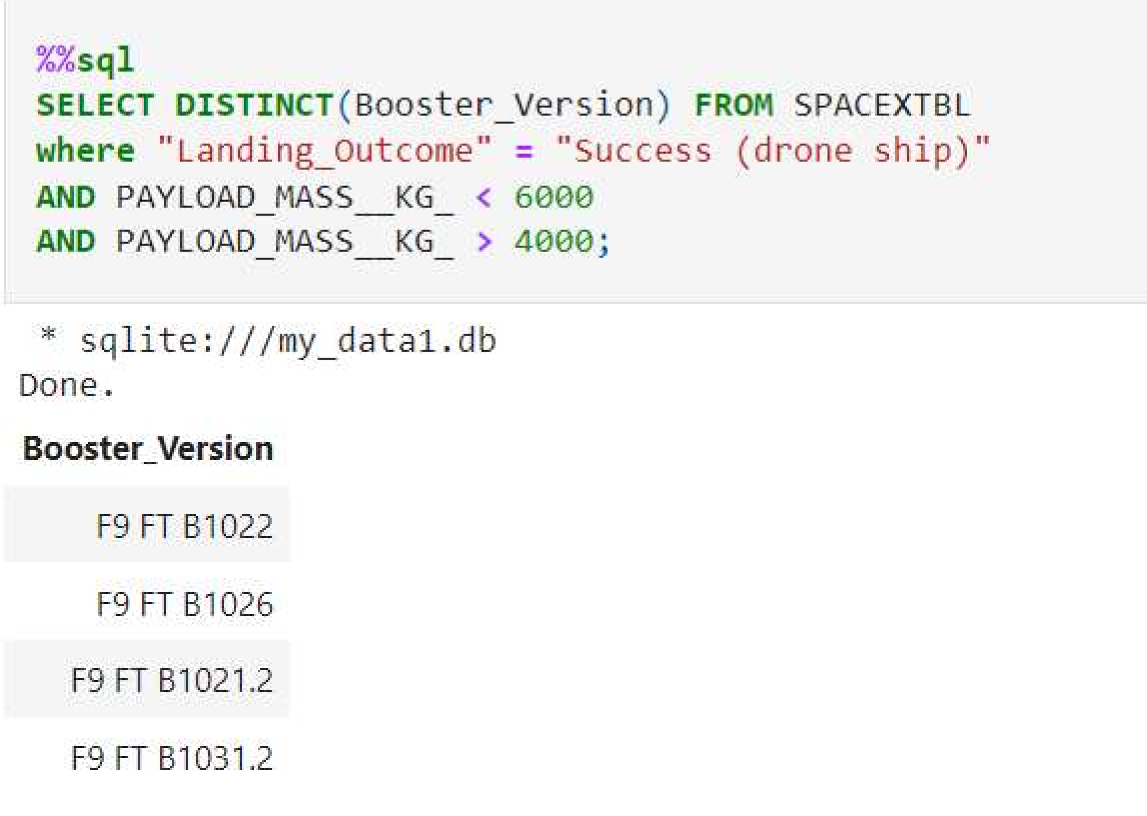
```
%%sql
select MIN("Date") as FIRST from SPACEXTBL
where "Landing_Outcome" LIKE '%Success%ground pad%';
```

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

FIRST
2015-12-22

The date of the first successful landing outcome on ground pad was December 22nd 2015

Drone Ship Landings with Payload between 4000 and 6000



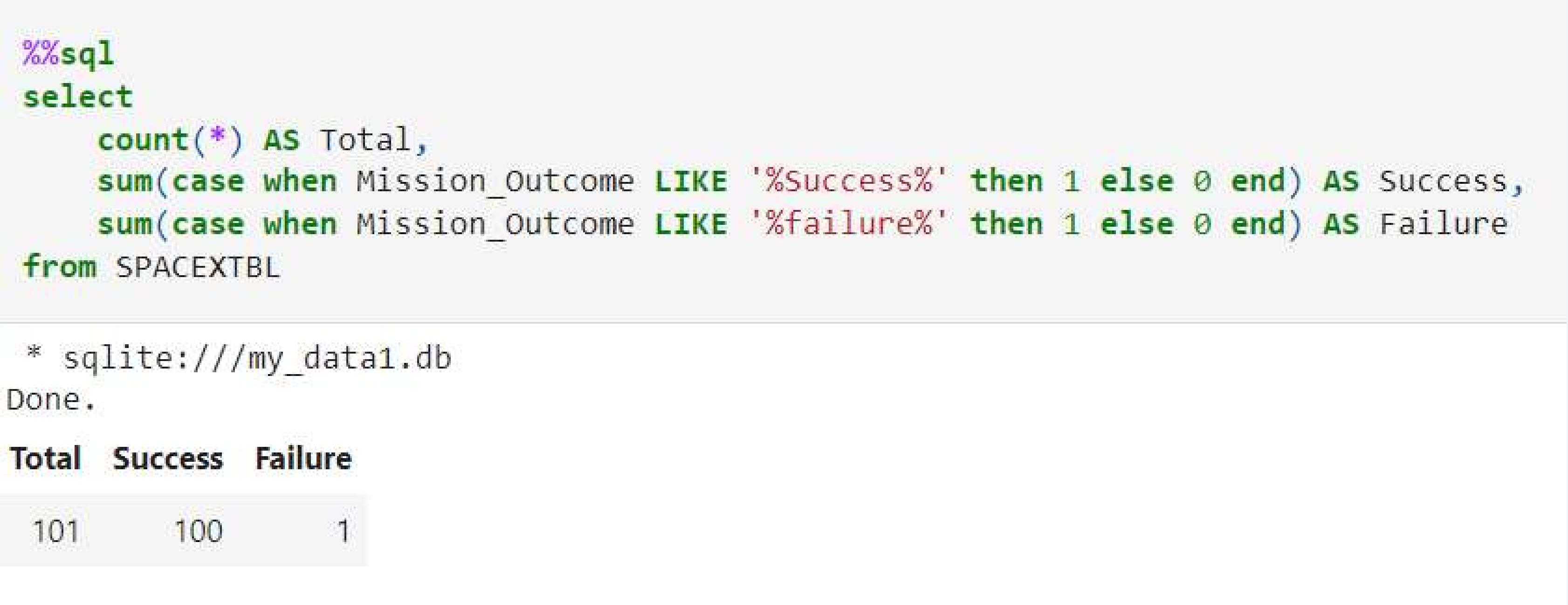
```
%%sql
SELECT DISTINCT(Booster_Version) FROM SPACEXTBL
where "Landing_Outcome" = "Success (drone ship)"
AND PAYLOAD_MASS__KG_ < 6000
AND PAYLOAD_MASS__KG_ > 4000;
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Four Booster Versions had a successful drone ship landing with payload between 4000kg and 6000kg
- Note that they are all FT Booster Versions

Total Number of Successful and Failure Mission Outcomes



```
%%sql
select
    count(*) AS Total,
    sum(case when Mission_Outcome LIKE '%Success%' then 1 else 0 end) AS Success,
    sum(case when Mission_Outcome LIKE '%failure%' then 1 else 0 end) AS Failure
from SPACEXTBL
```

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

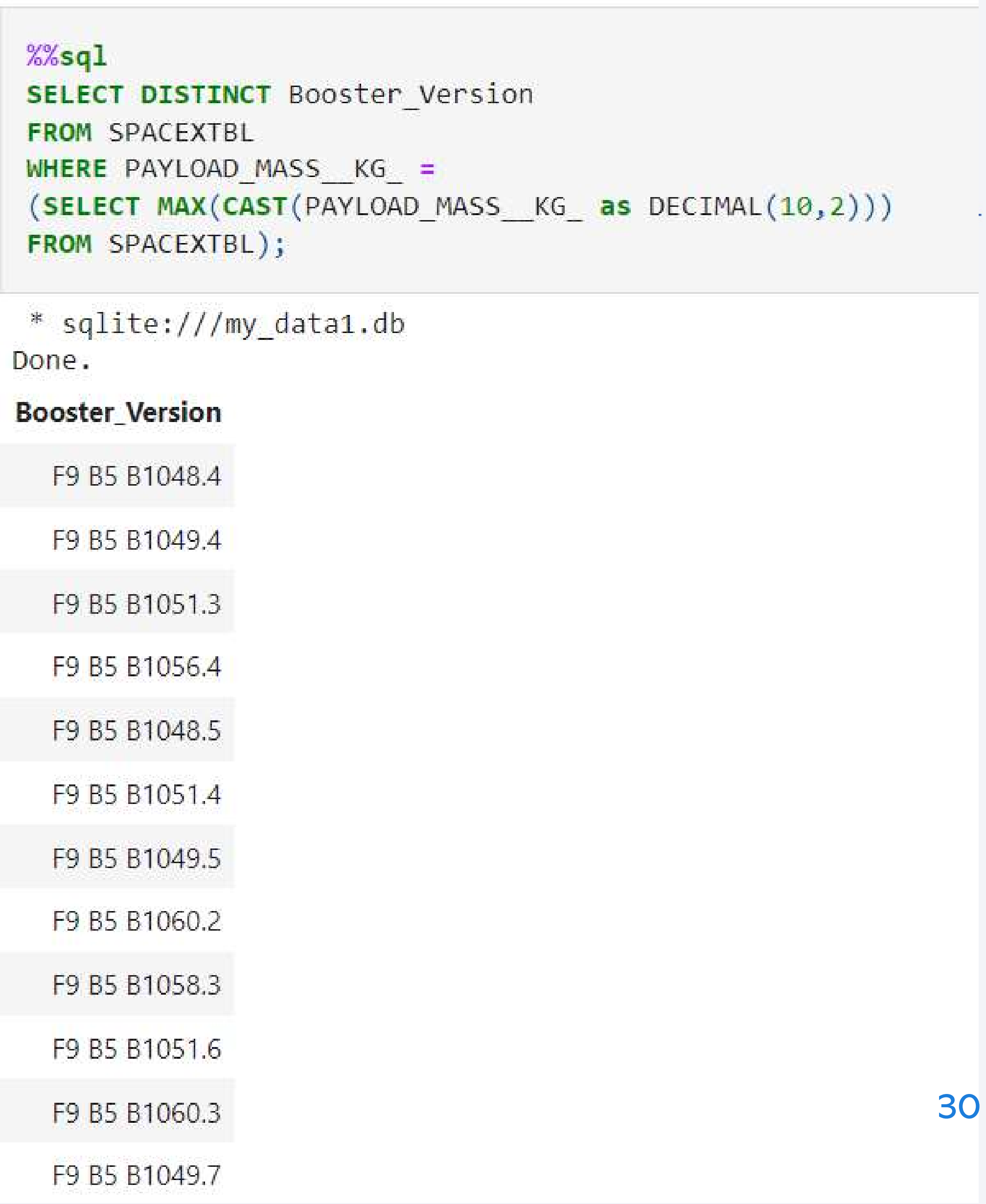
Total	Success	Failure
101	100	1

There has only one launch where mission outcome failed

Boosters Carried Maximum Payload

These Booster Versions have carried the maximum payload

Note that they are all B5 booster versions

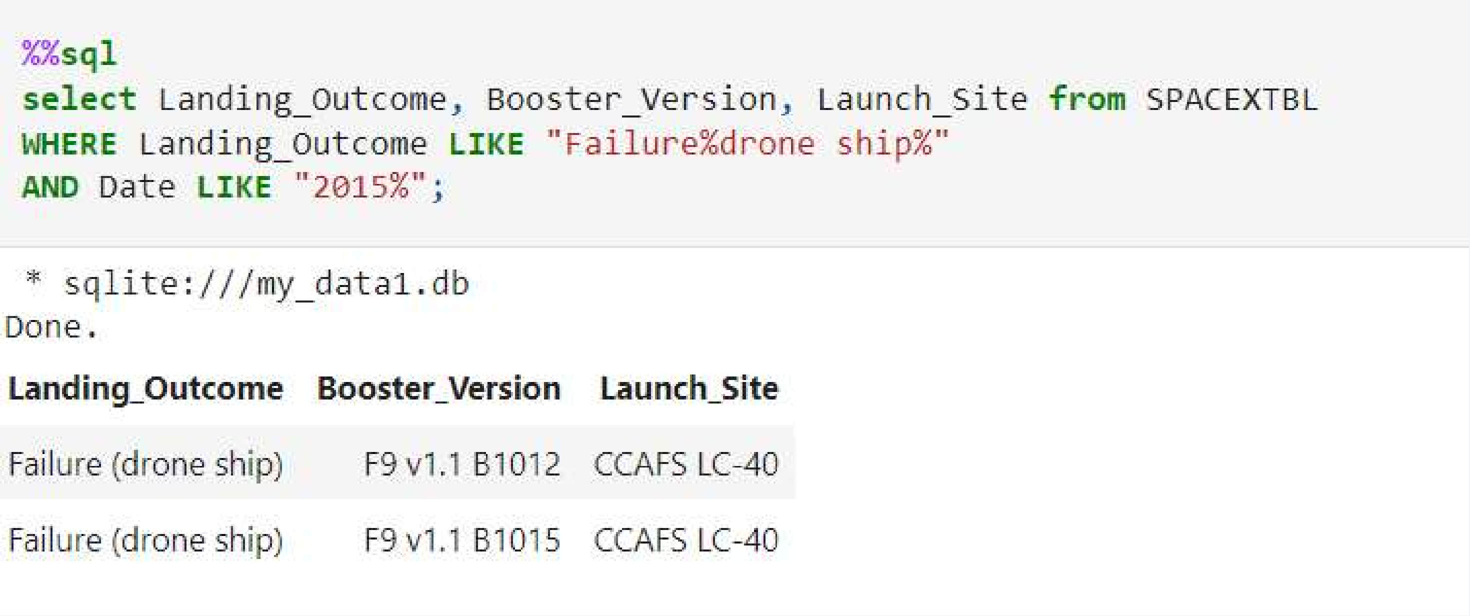


```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(CAST(PAYLOAD_MASS__KG_ as DECIMAL(10,2)))
FROM SPACEXTBL);
```

```
 * 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

2015 Launch Records



```
%%sql
select Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
WHERE Landing_Outcome LIKE "Failure%drone ship%"
AND Date LIKE "2015%";
```

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

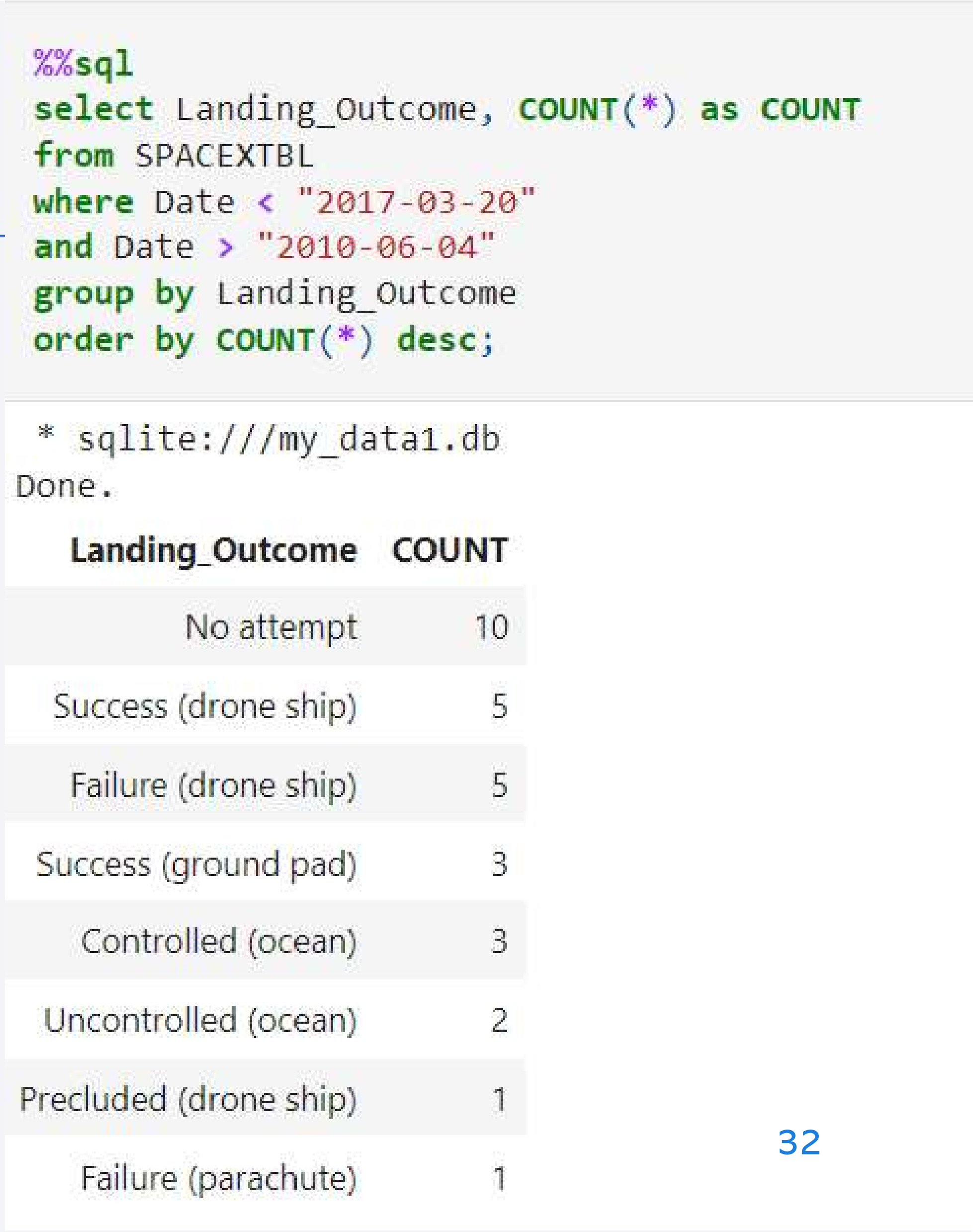
Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The above launches, failed to land the drone ship in 2015
- They were using F9 v1.1 boosters and launching from CCAFS LC-40

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

Looking at a rank of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

- Most commonly occurring outcome was ‘No landing attempted’



```
%%sql
select Landing_Outcome, COUNT(*) as COUNT
from SPACEXTBL
where Date < "2017-03-20"
and Date > "2010-06-04"
group by Landing_Outcome
order by COUNT(*) desc;
```

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

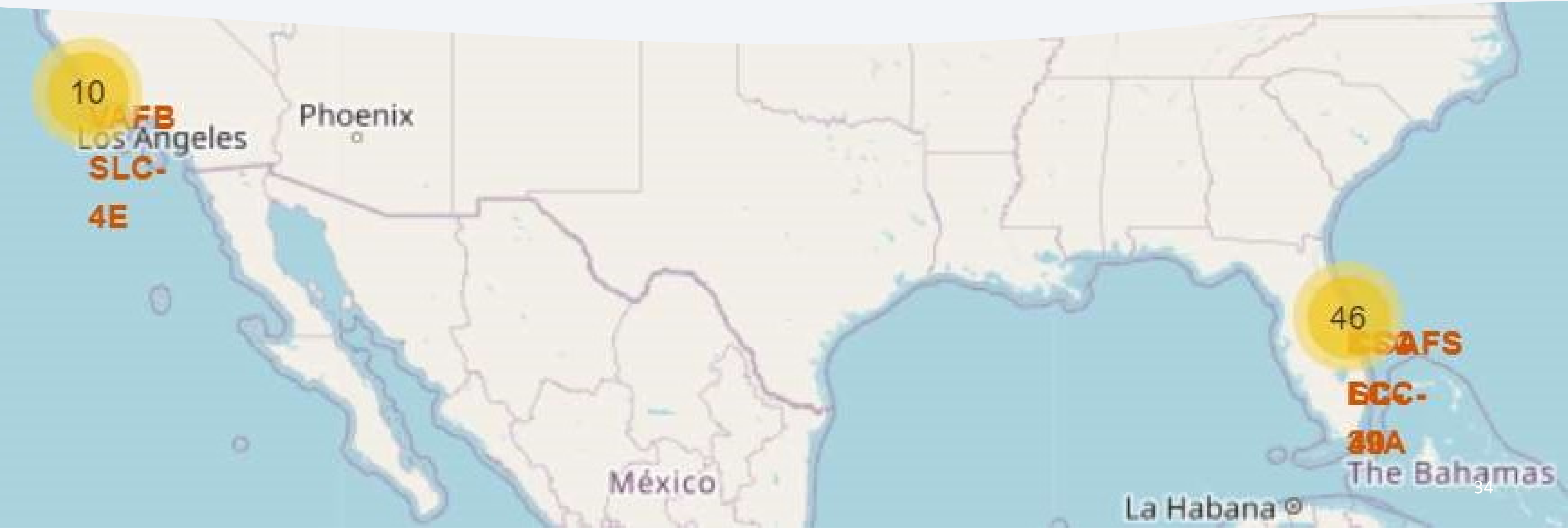
Landing_Outcome	COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

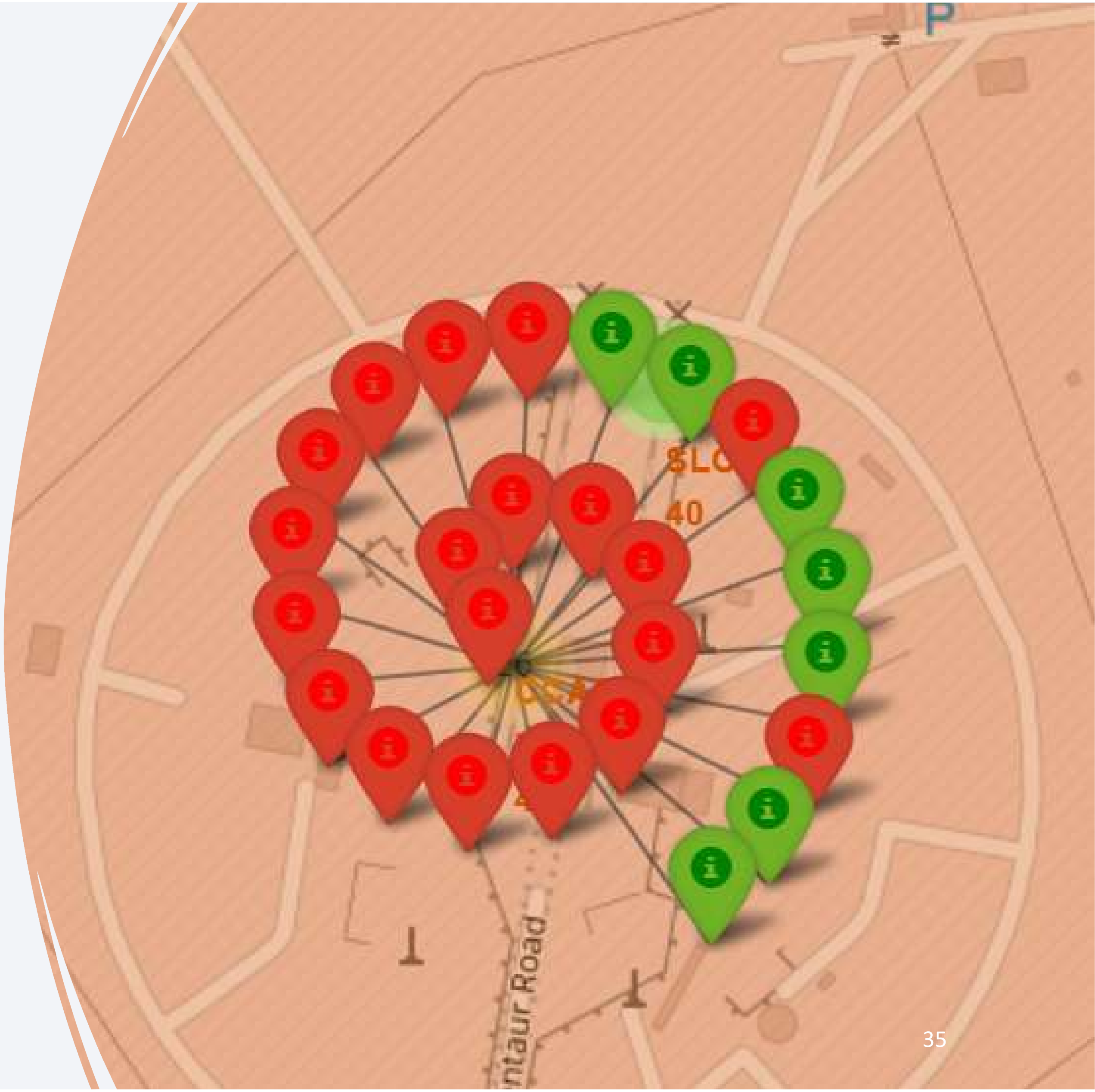
Map of All Launch Sites

- Currently all launch sites are located close to the coastline
- Its important to note that three of SpaceXs' launch sites are located in very close proximity to one another in Florida. They account for 46 out of the 56 rows of data that we have on launches



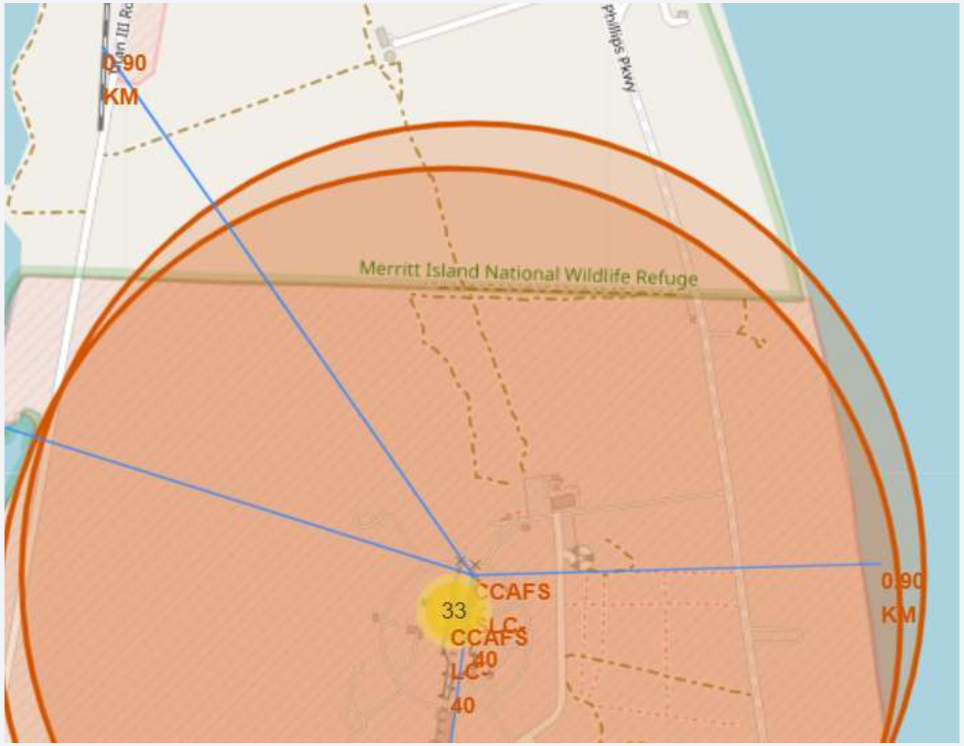
Map of Launch Outcomes per Site

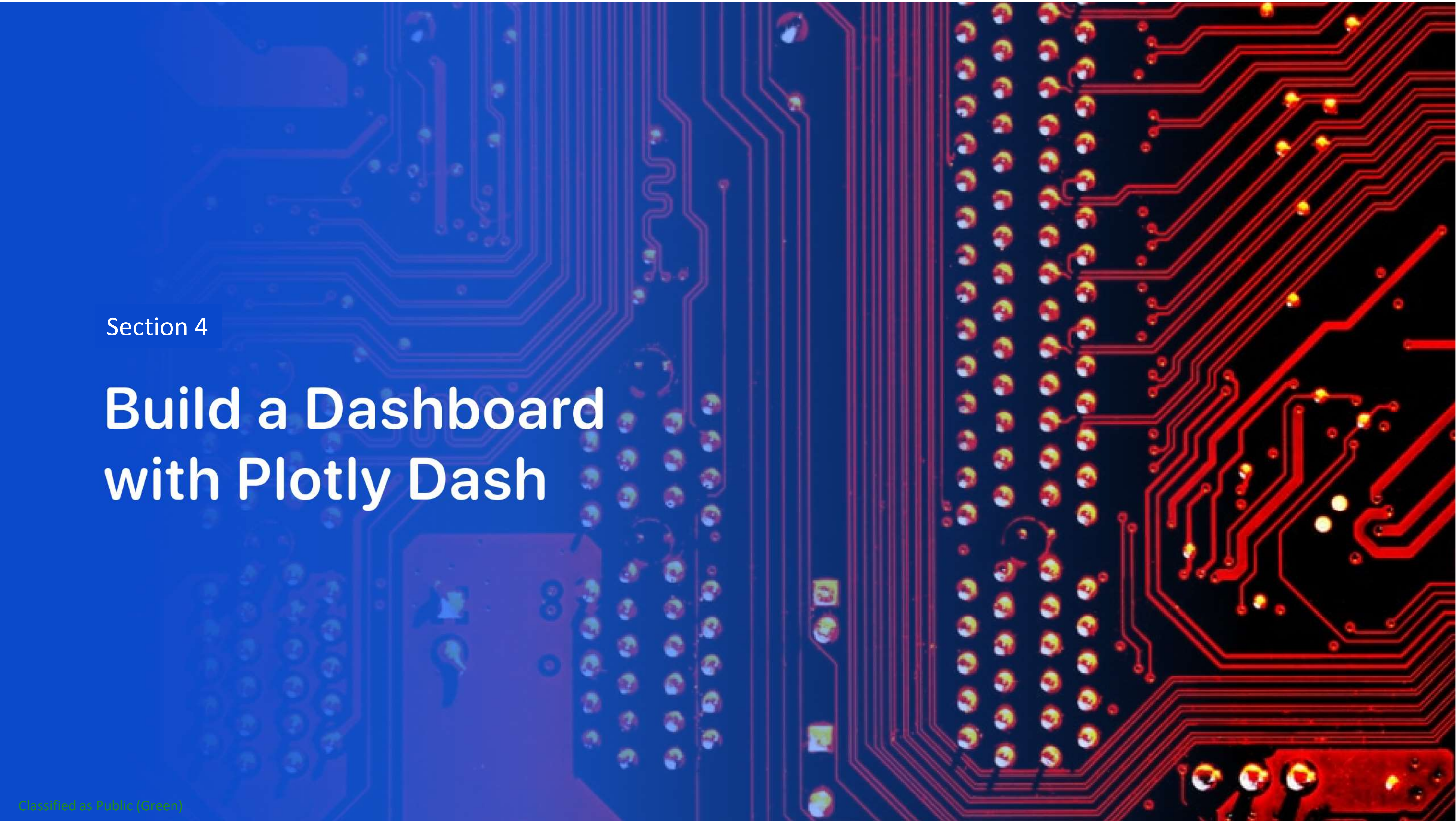
- I zoomed in and looked at the options for each site
- For the purposes of this presentation, I will look at the site, CCAFS LC-40; this site accounts for the most launches in our dataset



Launch Site Proximities

- Looking again at CCAFS LC-40, we can see that this launch site is located less than 1km from the coastline and a railway line



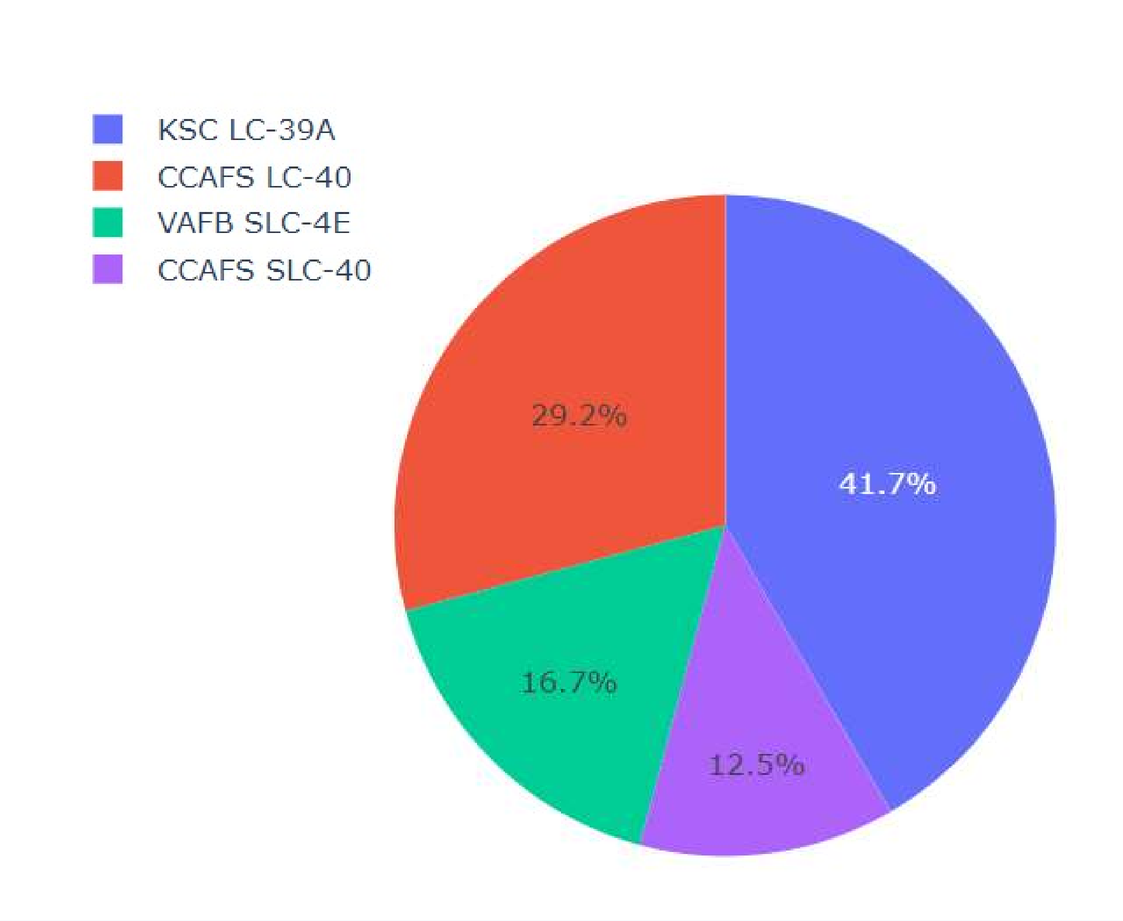


Section 4

Build a Dashboard with Plotly Dash

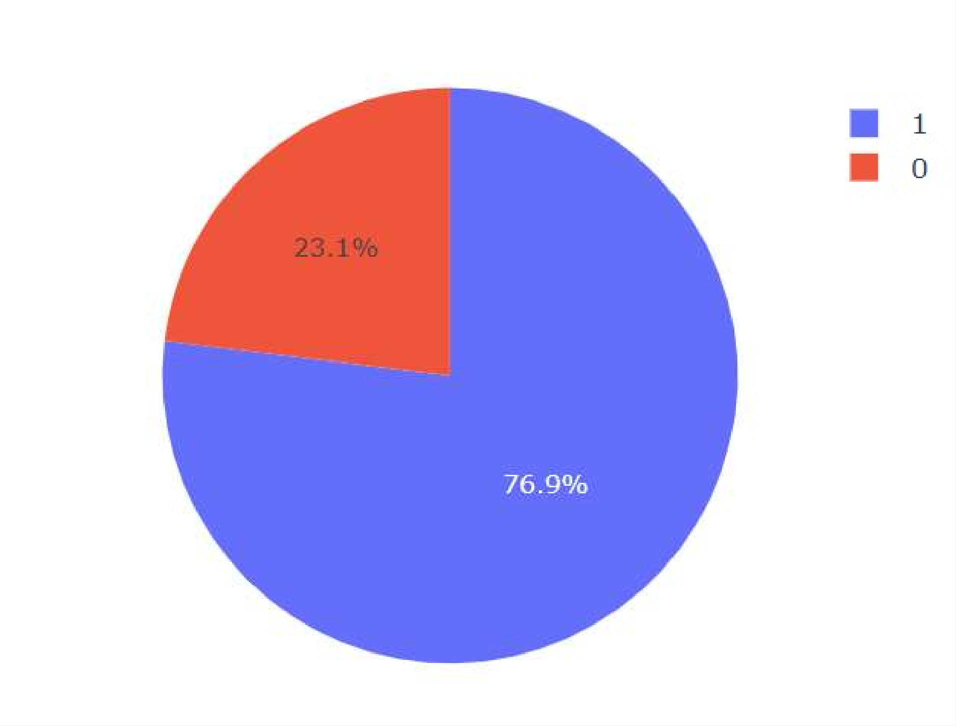
Total Successful Launches by Site

- The majority of successful launches occur at KSC LC-39A
- 41.7% of successful launches occur at KSC LC-39A
- CCAFS LC-40, the launch site with the next highest proportion of successful launches, accounted for only 29.2% of successful launches



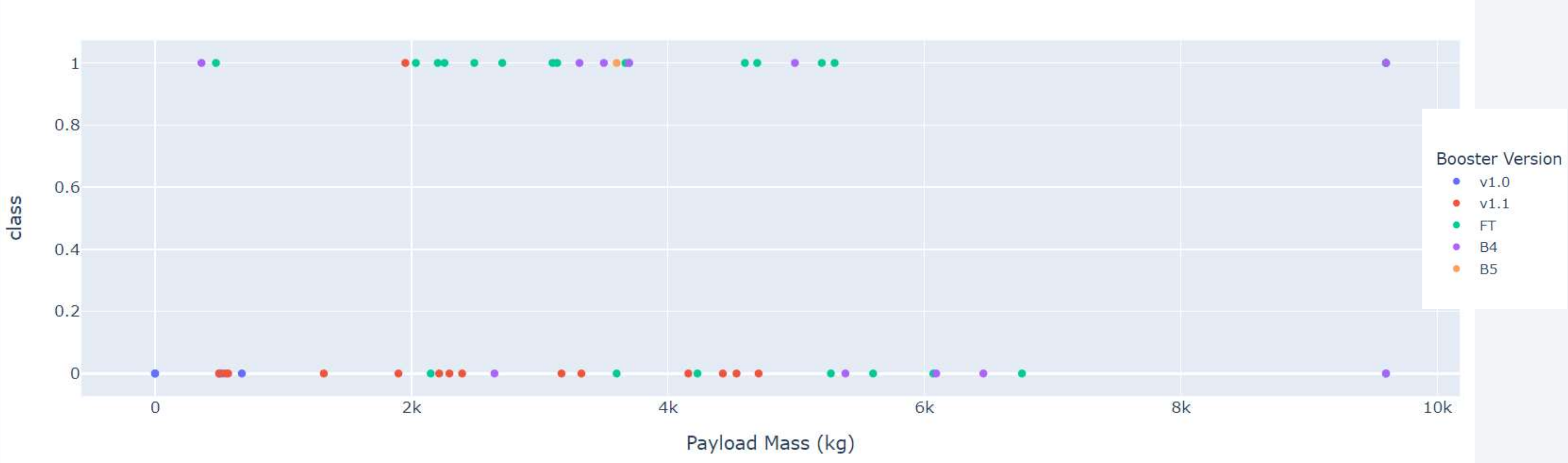
Total Successful Launches for KSC LC-39A

Looking at only KSC LC-39A, we can see that 76.9% of launches at that site are successful



Classified as Public (Green)

Payload vs. Launch Outcome scatter plot



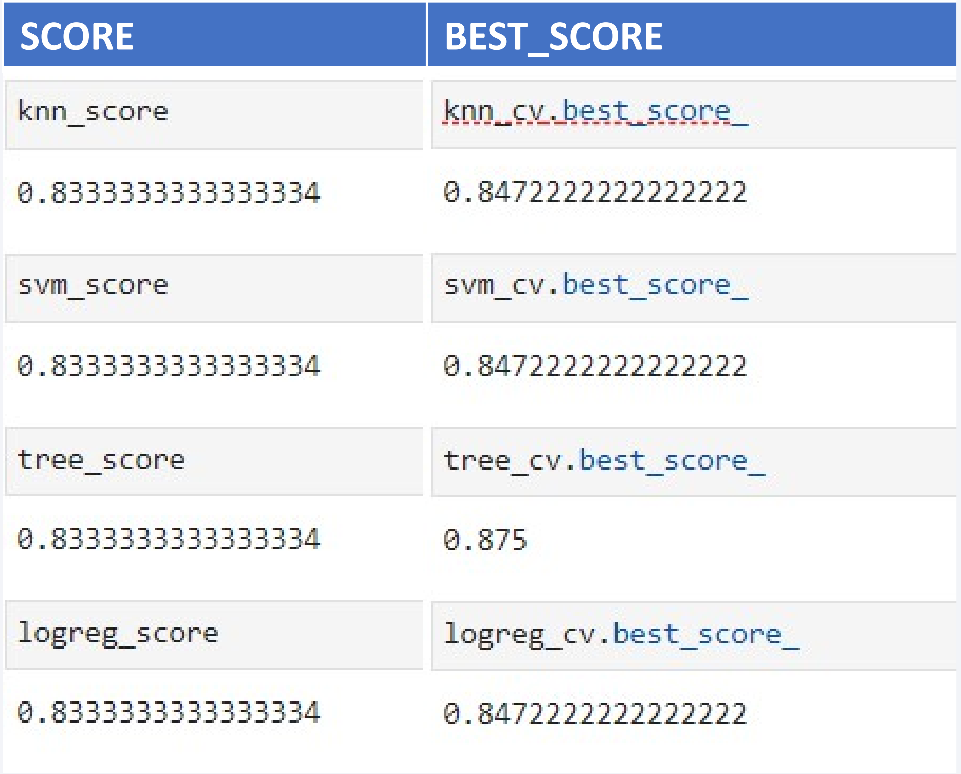
- v1.0 and v1.1 boosters have a very poor success rate
- FT and B4 boosters appear more likely to be successful
- Most launches have a payload mass between 2000kg and 6000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

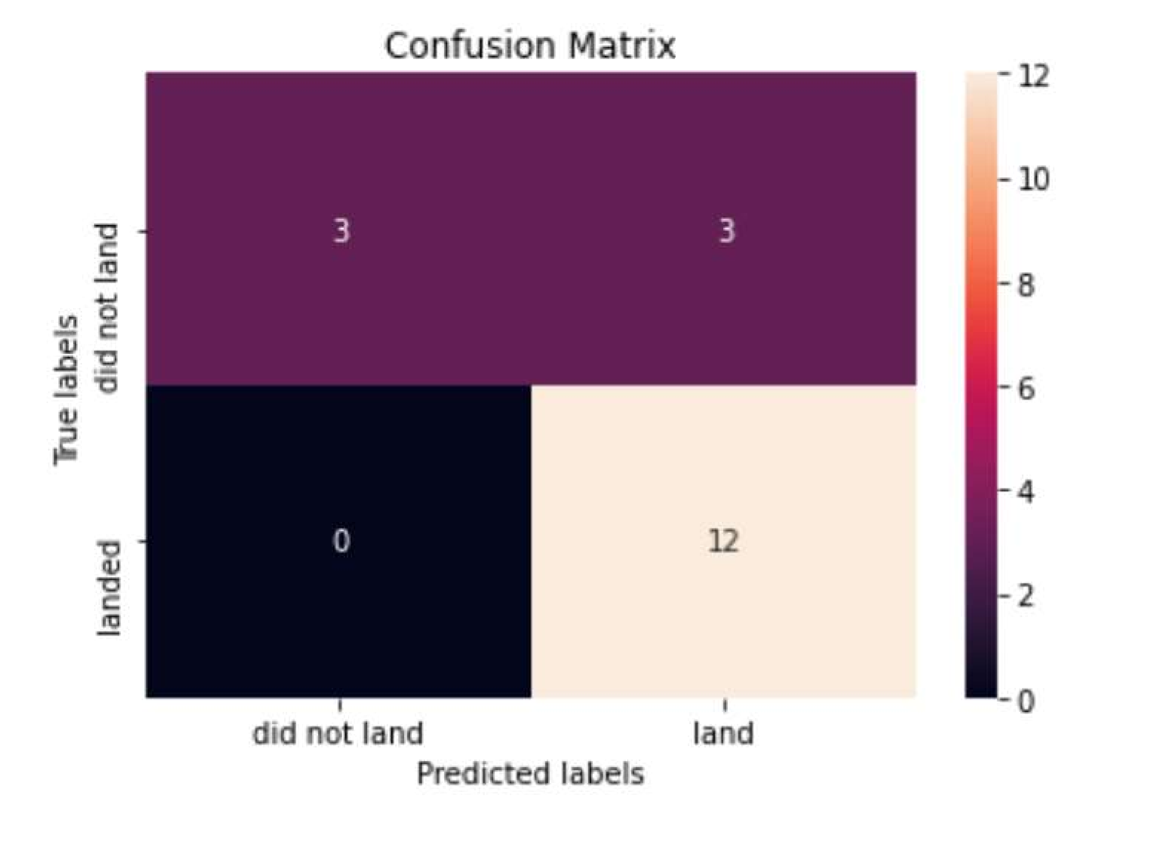
- All models gave the same degree of accuracy, using the method score;
  0.8333334
- This is likely due to the very small data set
- Looking at the best_score_ (the mean cross-validated score of the best_estimator) we see that Decision Tree Algorithm is receiving a higher score.
- So we will proceed with the Decision Tree predictive model



SCORE	BEST_SCORE
knn_score	knn_cv.best_score_
0.8333333333333334	0.8472222222222222
svm_score	svm_cv.best_score_
0.8333333333333334	0.8472222222222222
tree_score	tree_cv.best_score_
0.8333333333333334	0.875
logreg_score	logreg_cv.best_score_
0.8333333333333334	0.8472222222222222

Confusion Matrix

- On this run, the Decision Tree model successfully predicted 15 of the 18 landing outcomes correctly
- There were three false positive predictions



Conclusions

- Payloads with smaller mass are more likely to be successful.
- The majority of successful launches occur at KSC LC-39A.
- The Tree Classifier Algorithm is current the best classification model for predicting launch outcome.
- This is a very small data set, and we have very few data points for payloads >7000kg or particular orbits that appear to have a high success rate.
- We should continue to monitor new data as it becomes available and update our models accordingly.

Thank you!