IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Rachel
26 June 2022

# Outline

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

# Executive Summary

Data science methodology was applied to develop a model that predicted if the first stage of the SpaceX falcon 9 would land successfully and can therefore be re-used. Past launch data was obtained from using the SPACEX REST API and web scraping Wikipedia. It was cleaned and formatted. Exploratory data analysis identifies key attributes that would effect the mission outcomes. These were launch site, payload, orbit type, flight number and year of launch. Four predictive models were tested Logistic Regression, SVM, Decision Tress, K-Nearest Neighbour. The decision tree was the most accurate model, with a test model accuracy of 0.888.

# Introduction

A new company Space Y wants to be competitive in commercial space market. To do this they needs to estimate the cost of each launch. The primary driver is the lunch cost is if the first stage of the rocket land successfully and can be re-used. Using their competitor SpaceX past launch data, the primary goal is to develop a predictive model which answers: Will the first stage of the Falcone 9 land successfully?

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

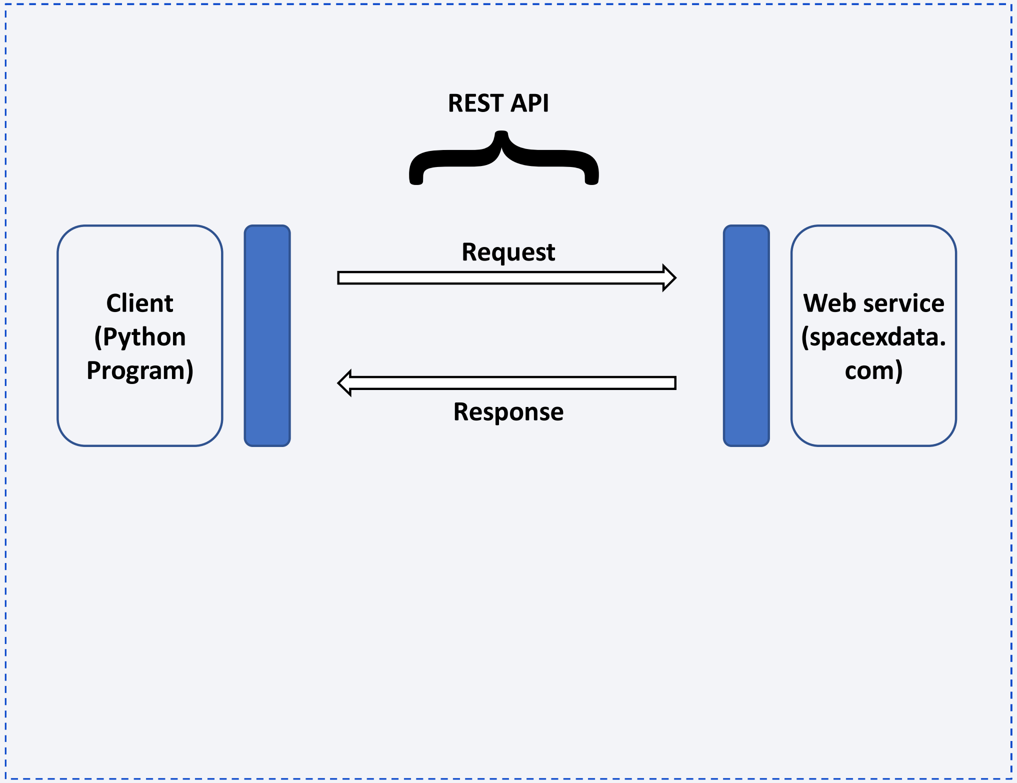
# Data Collection

- Data was collected using two methods
    - SpaceX REST API
    - Web scraping from Wikipedia

# Data Collection – SpaceX API

- SpaceX REST API was used to access data form the endpoint
  - https://api.spacexdata.com/v4/rockets/launches/past
- IDs from the /launches/past were cross referenced using addition endpoints (rocket, payload, launchpad, cores) to get a complete dataset
- The final data obtained was
  - FlightNumber
  - Date
  - BoosterVersion
  - PayloadMass,
  - Orbit
  - LaunchSite
  - Outcome
  - Flights
  - GridFins
  - Reused
  - Legs
  - LandingPad
  - Block
  - ReusedCount
  - Serial
  - Longitude
  - Latitude

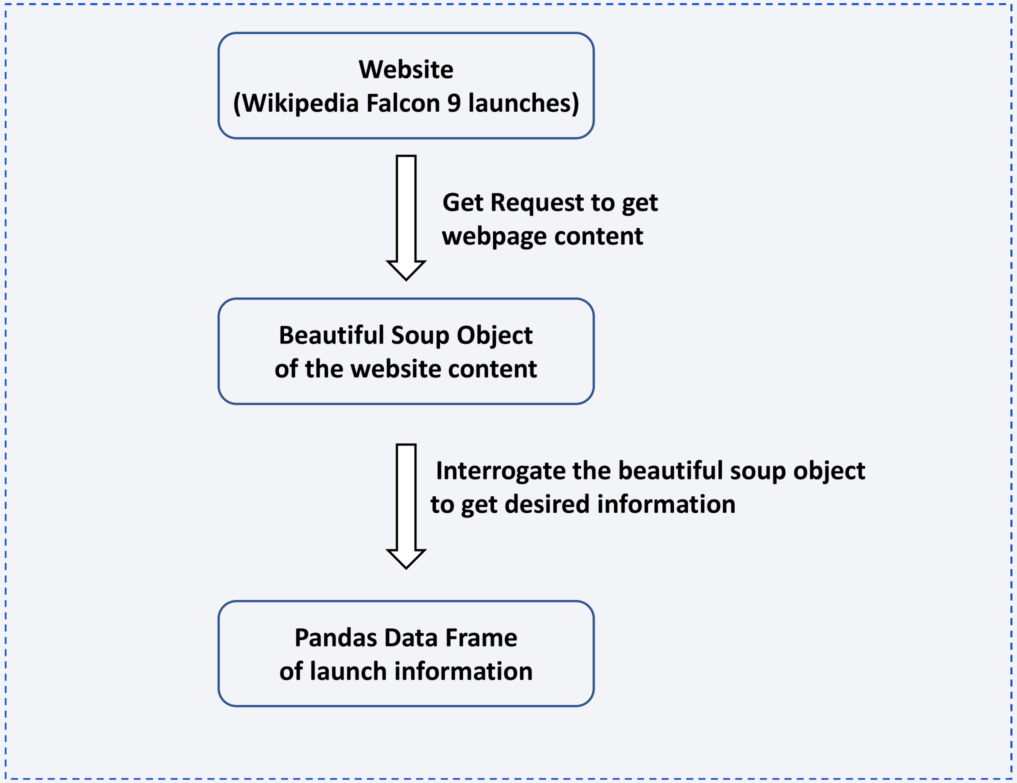
- https://github.com/Rachie-M/Assignment/blob/main/01-jupyter-labs-spacex-data-collection-api.ipynb



# Data Collection - Scraping

- Beautiful Soup waws used to get information from a Wikipedia page on Falcon 9
  - "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
- The data was scraped and cleaned from the HTML table and stored in a Pandas data frame

- https://github.com/Rachie-M/Assignment/blob/main/02-jupyter-labs-webscraping.ipynb

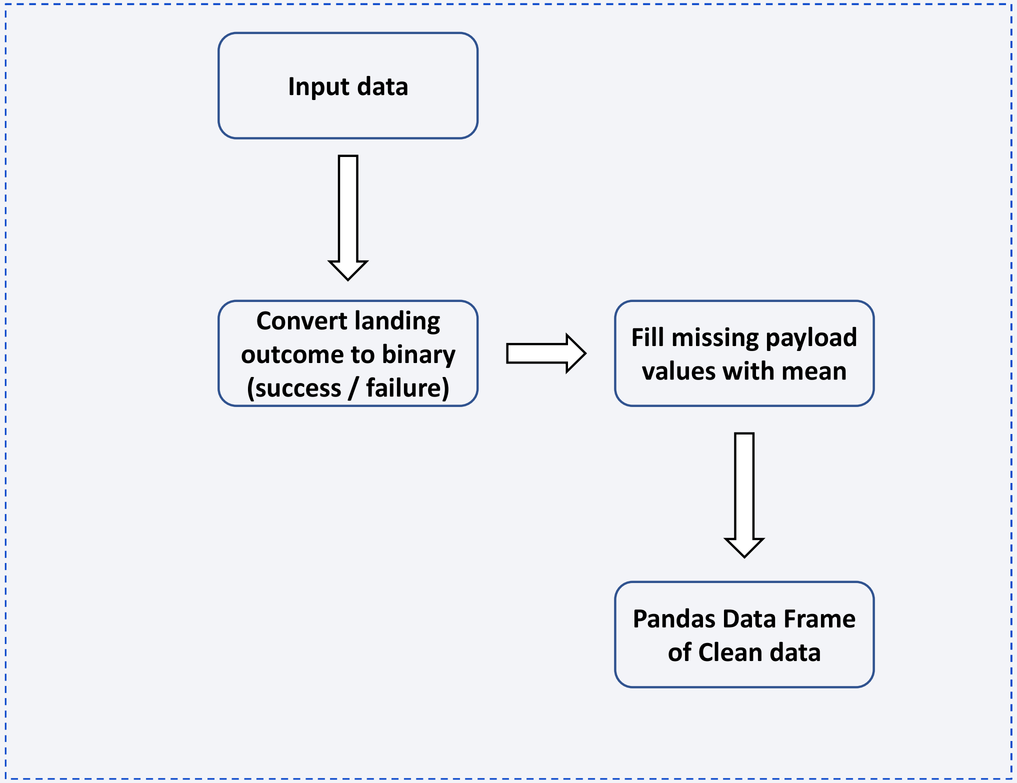

Website
(Wikipedia Falcon 9 launches)
Get Request to get webpage content
Beautiful Soup Object of the website content
Interrogate the beautiful soup object to get desired information
Pandas Data Frame of launch information

# Data Wrangling

- Initially looked at the distribution of the data
  - Data type of each column
  - Count number of launches for each launch sites
  - Count number of launches for each orbit types

- Clean the data
  - Create a new column ‘class’ which converts the Outcome variable to binary (success / failure)
  - Looked at missing values and replaced missing payload values with mean payload

- https://github.com/Rachie-M/Assignment/blob/main/03-labs-jupyter-spacex-Data%20wrangling.ipynb


Input data
Convert landing outcome to binary (success / failure)
Fill missing payload values with mean
Pandas Data Frame of Clean data

# EDA with Data Visualization

- EDA was performed to become familiar with the data and determine what attributes have an effect on landing success
- EDA was primarily done through looking at figures (and not statistical analysis)
- Charts that were plotted where
  - Payload Mass vs Flight Number
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - The success rate of each orbit type
  - Flight Number vs Orbit type
  - Payload vs Orbit type
  - The yearly trend of launch success

- https://github.com/Rachie-M/Assignment/blob/main/05-jupyter-labs-eda-dataviz.ipynb

# EDA with SQL

- SQL queries allow you to interrogate the data while it’s still in the database. This means you can retrieve only the data you want
- A summary of the SQL queries you performed
    - Identify unique launch sites
    - Identify launch sites that begin with 'CCA’
    - Determine the total payload mass carried by boosters launched by NASA (CRS)
    - Determine the average payload mass carried by booster version F9 v1.1
    - Determine the date of the first successful landing outcome to ground pad
    - List the names of the boosters which have success in drone ship and have payload mass > 4000 but < 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  launch failure outcomes that involved a drone ship for 2015
    - Determine the number of launches for each unique landing outcomes between the date 04-06-2010 and 20-03-2017 and rack accordingly

- https://github.com/Rachie-M/Assignment/blob/main/04alternative-COMPLETED-jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

- Spatial analysis was carried out by creating maps with folium
- These maps showed
  - The unique launch sites as circles on the map
  - Using marker clusters, each launch was coloured depending on if it was a success or failure
  - Distance between launch sites and major feature (coast, railways, highways and cities) was calculated and plotted as a line

- https://github.com/Rachie-M/Assignment/blob/main/06-COMPLETED_lab_jupyter_launch_site_location.ipynb

# Build a Dashboard with Plotly Dash

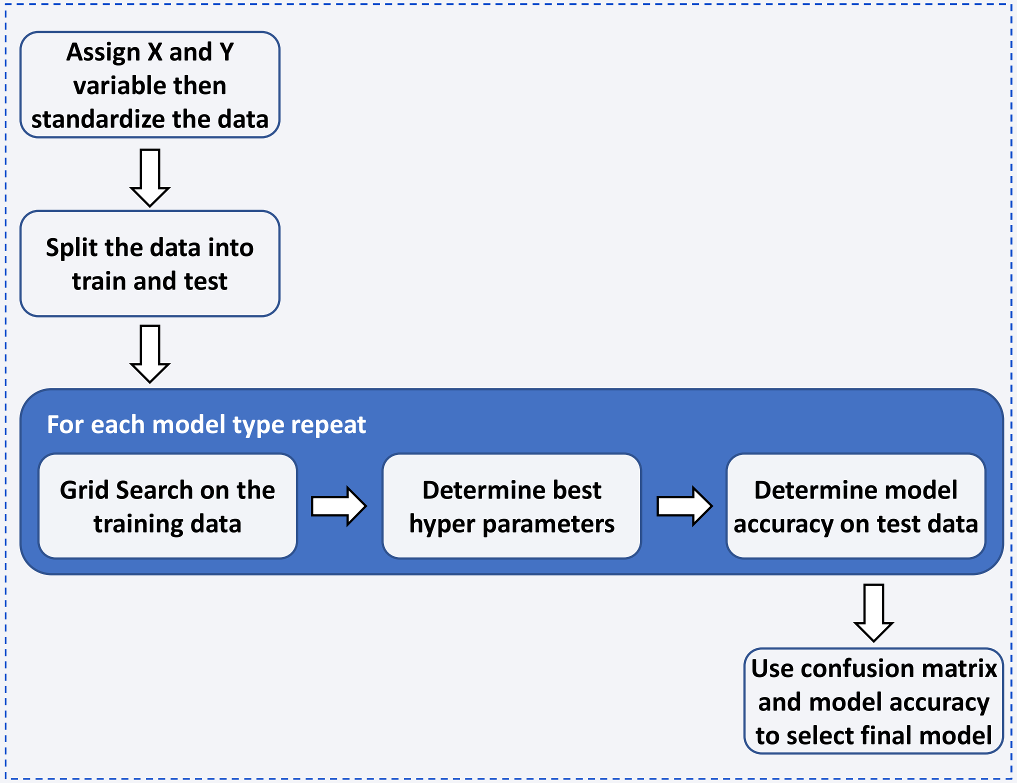
- An interactive dashboard was generated using Ploty Dash
- The interactive feature where
  - A drop down to select all site or an individual launch location
  - A slider to choose payload mass in kg
- The plots/graphs on the dashboard were
  - Pie chart showing the launch success/failure
  - Scatter plot of the payload in kg vs if the launch was a success or not
- Having these plots interactve allow you to investigate if site location and payload effect the launch success

- https://github.com/Rachie-M/Assignment/blob/main/07-spacex_dash_app.py

# Predictive Analysis (Classification)

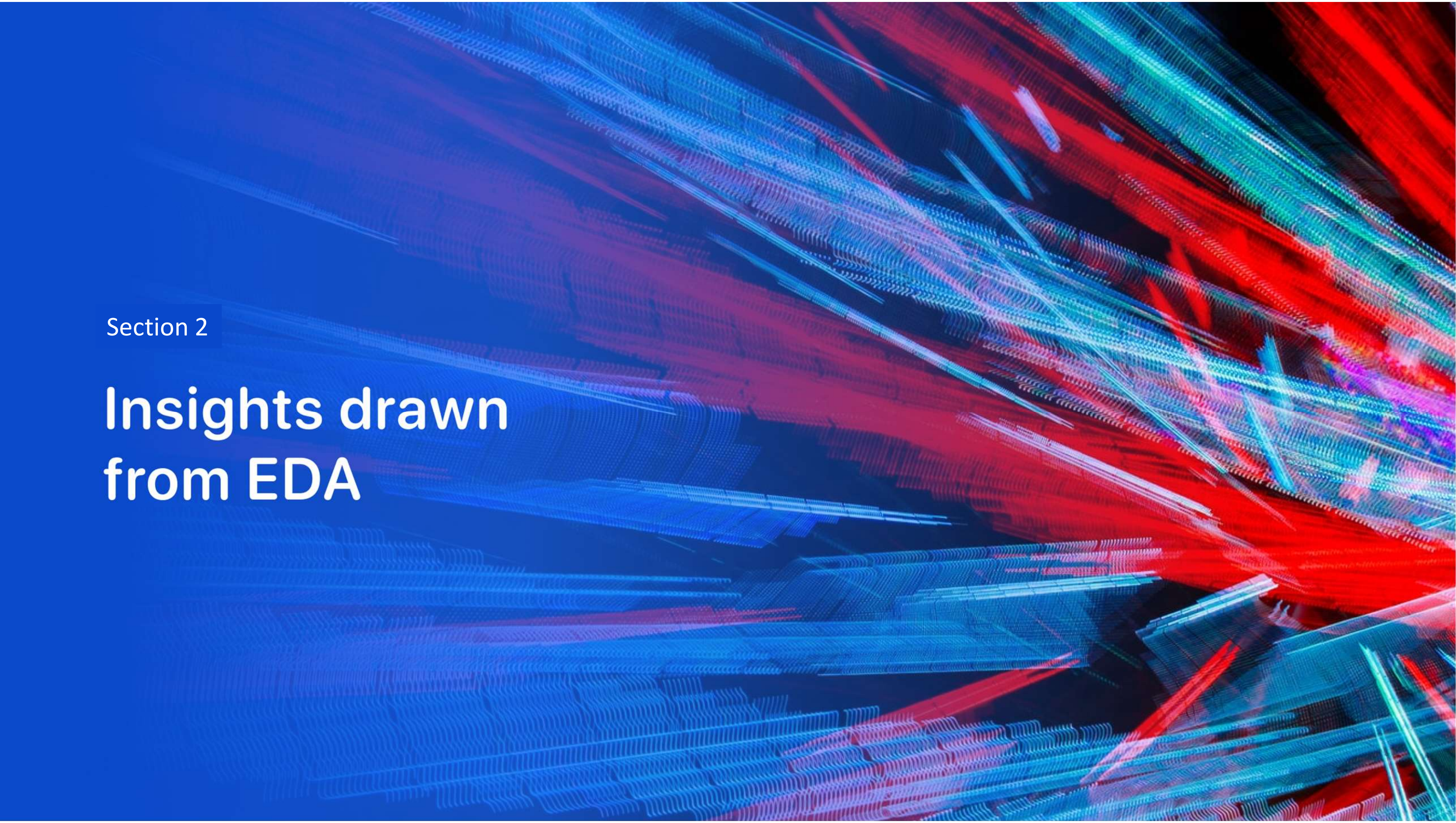
- Four predictive model were generated
  - Logistic Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbour
- The workflow used is depicted in the diagram
- A model accuracy score was used to determine which model was the best at predicting launch outcomes
- For consistency and comparison purposes the same train and test data was used for each model

- https://github.com/Rachie-M/Assignment/blob/main/08-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb


Assign X and Y variable then standardize the data
Split the data into train and test
For each model type repeat
Grid Search on the training data
Determine best hyper parameters
Determine model accuracy on test data
Use confusion matrix and model accuracy to select final model

# Results

- Exploratory data analysis show the attributes that have an effect of the launch outcome and the parameters used to bulid the model are:
  - FlightNumber
  - PayloadMass
  - Flights
  - Block
  - ReusedCount
  - Orbit
  - LaunchSite
  - LandingPad
  - Serial
  - GridFins
  - Reused
  - Legs
- Interactive analytics
  - KSC LC-39A is the site with the most successful launch mission and if the payload is <5000 kg then it has 100% success rate
- Predictive analysis results
  - The best model was the decision tress which returned a test model accuracy of 0.888

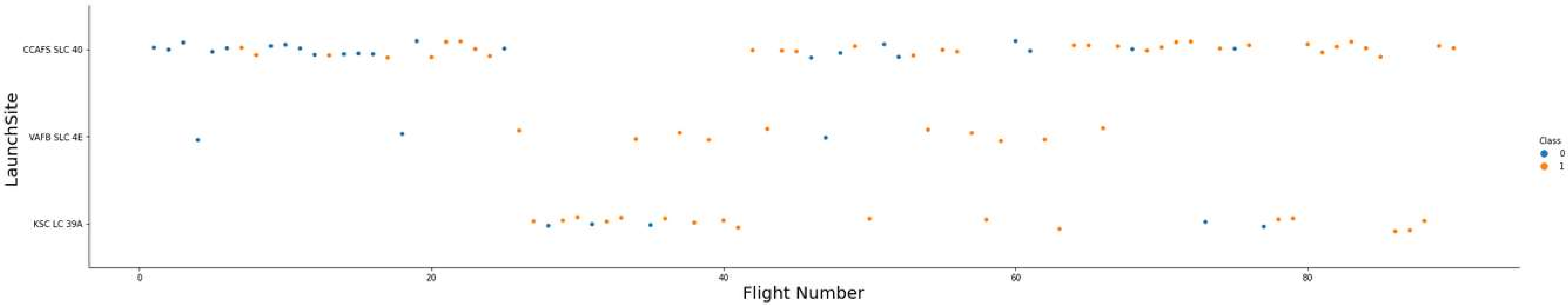


Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

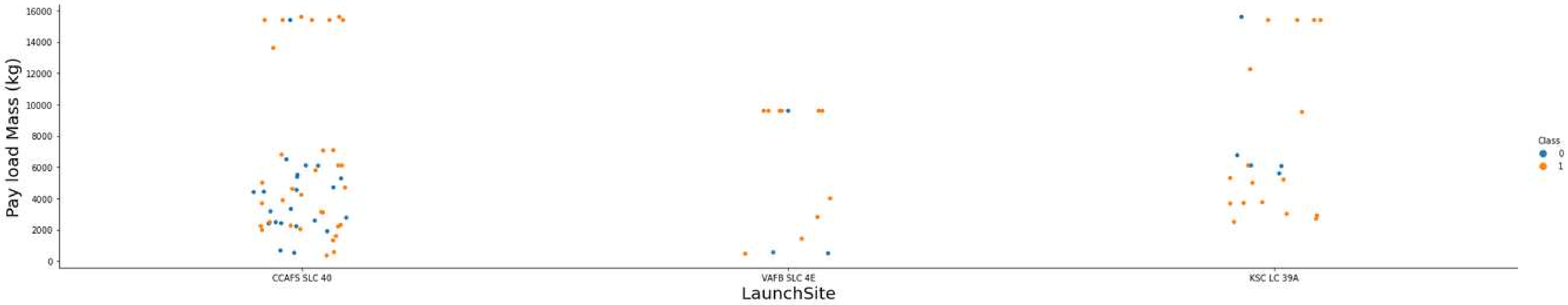
Flight Number vs. Launch Site



- The success of missions goes up as the flight number goes up
- CCAFS SLC 40 is the most used launch site, followed by KSC LC 39A then VAFB SLC 4E
- The most unsuccessful mission are launched from CCAFS SLC 40

# Payload vs. Launch Site

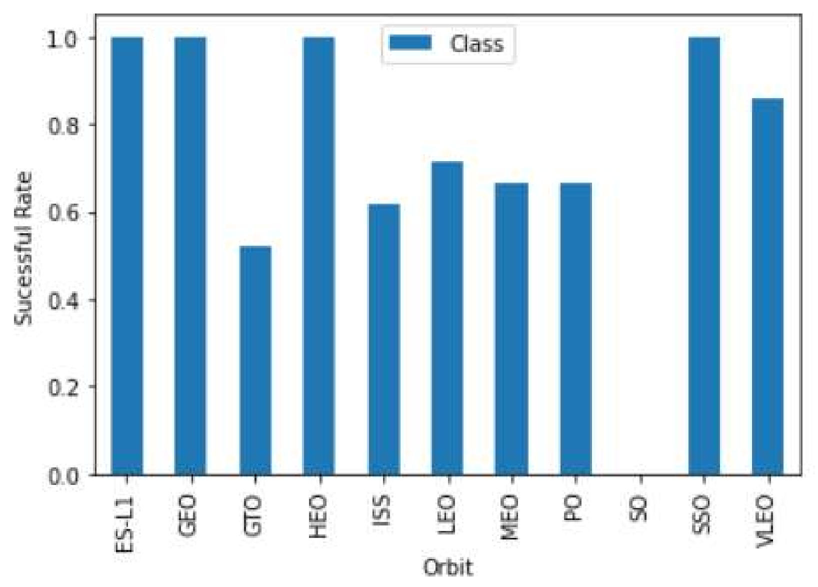
Flight Number vs. Launch Site



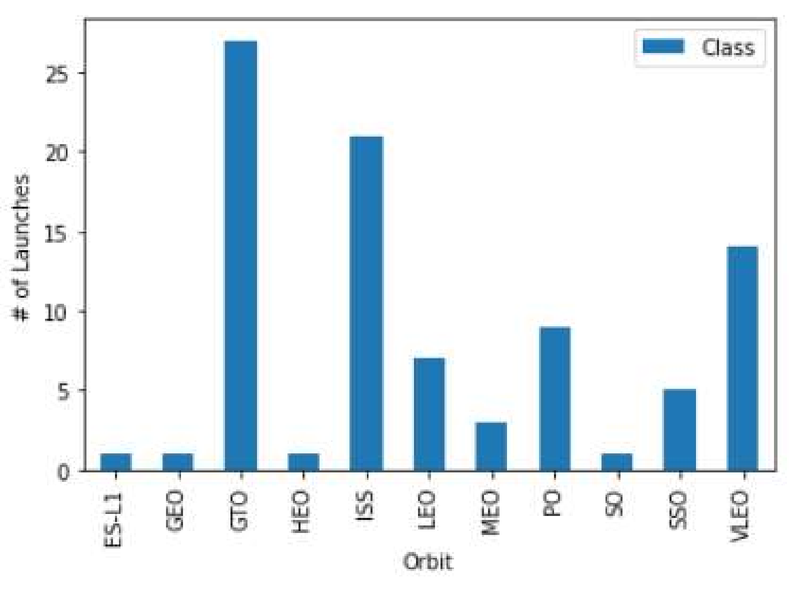
- If the payload is above 9000 kg then the mission is more likely to be successful
- Below a payload of 9000 kg sites KSC LC 39A and VAFB SLC 4E have more successful mission then CCAFS SLC 40

# Success Rate vs. Orbit Type

Success Rate of Each Orbit

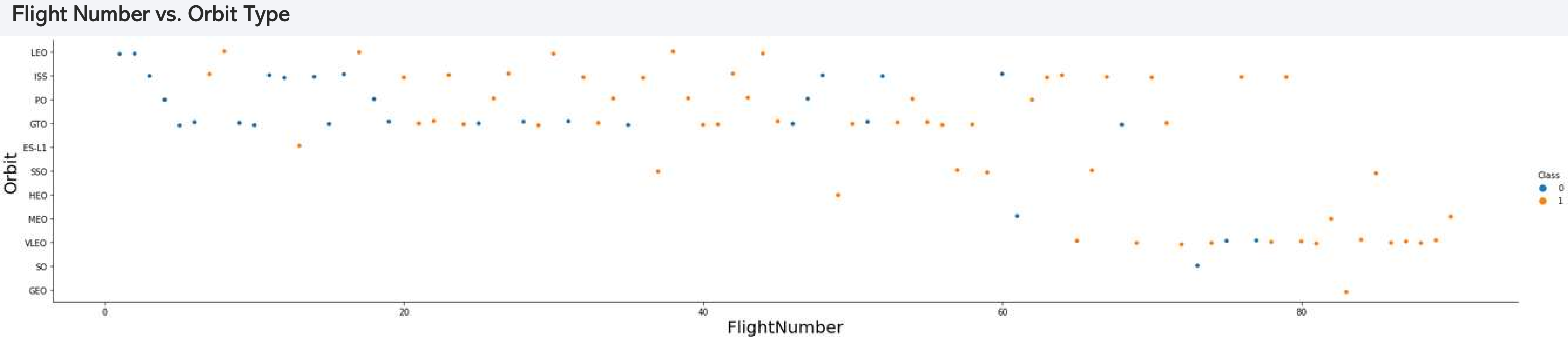


Number of Launches for Each Orbit



- The most common orbit is GTO, ISS and VLEO with VLEO having the heighst success rate
- ES-L1, GEO, HEO, SSO all have a 100% success rate however there has been only 1 launch for the for ES-L1, GEO, HEO and 5 for SSO
- The SO orbit hasn’t had any successful missions

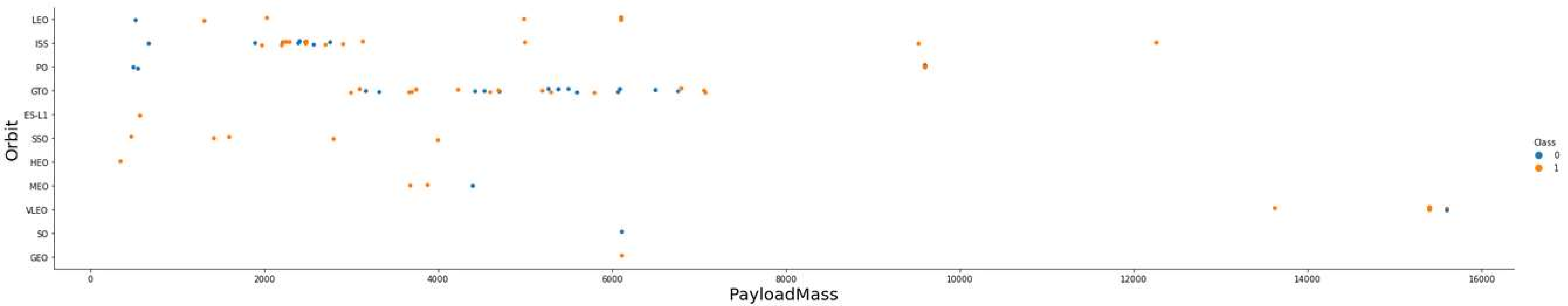
# Flight Number vs. Orbit Type



- As flight numbers increase there’s a shift to VLEO and SSO orbit which are more successful
- The success of ISS and GTO increases as flight number increases

# Payload vs. Orbit Type

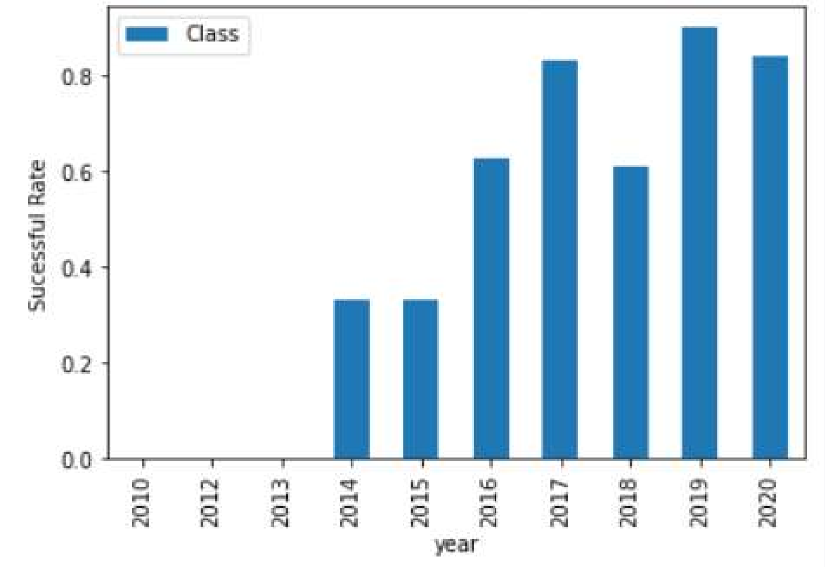
Flight Number vs. Launch Site



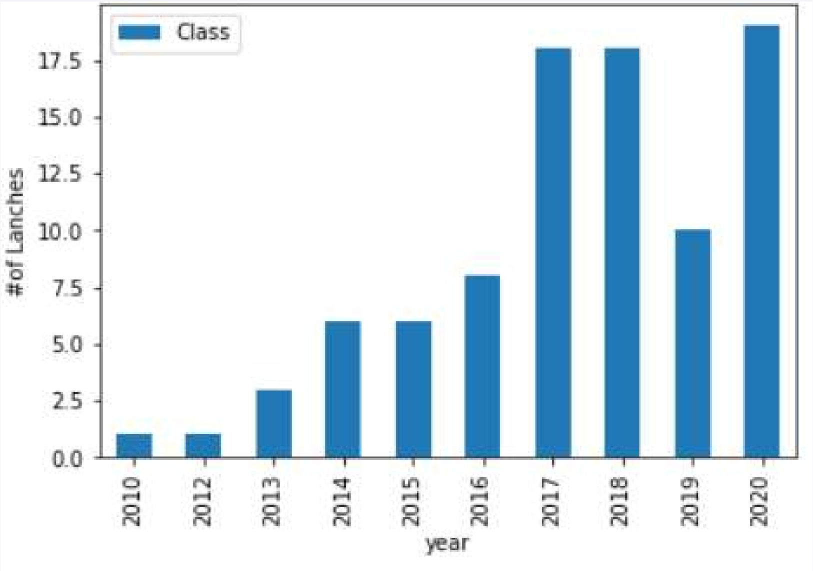
- VLEO, PO and ISS orbits have the highest payload with increase success rate compared to if they had a lower payload
- LEO, ISS and PO are all unsuccessful for payloads <1000
- ES-L1, SSO, HEO and MEO are all successful for payloads <4000
- GTO mixes mission success regardless of payload

# Launch Success Yearly Trend

Success Rate per Year



Number of Launches per Year

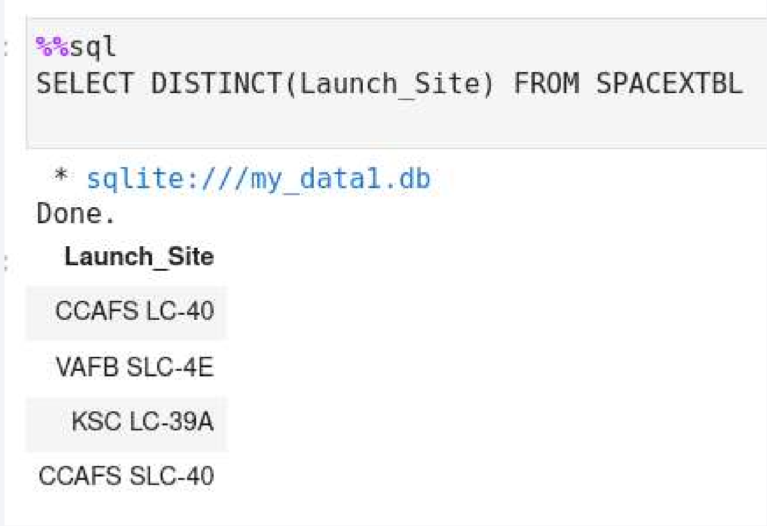


- 2010 where when missions started with the first successful mission in 2014
- As the years increase the number of mission attempts increases and the success rate goes up
- The best success rate was in 2019

# All Launch Site Names

- Find the names of the unique launch sites

- DISTINCT() is used to find the unique launch sites
- There are 4 unique launch sites



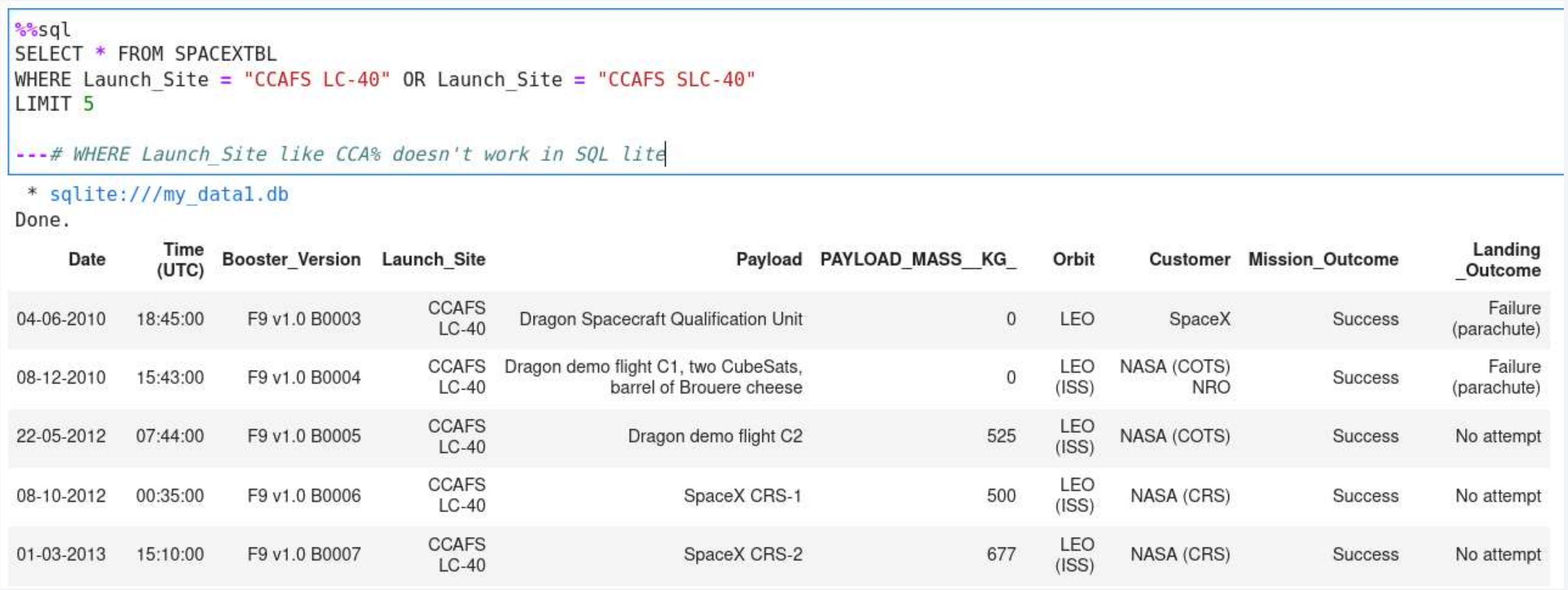
```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

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

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

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- SQLite didn’t let me use the like and % commands in SQL
- The launch sites starting with CAA are CAAFS LC-40 and CAAFS SLC-40



```
%%sql
SELECT * FROM SPACEXTBL
WHERE Launch_Site = "CCAFS LC-40" OR Launch_Site = "CCAFS SLC-40"
LIMIT 5

---# WHERE Launch_Site like CCA% doesn't work in SQL lite
```

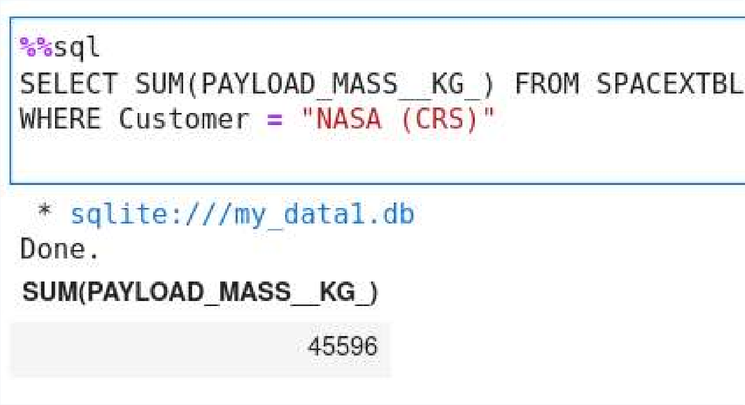
```
 * sqlite:///my_data1.db
Done.
```

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- Calculate the total payload carried by boosters from NASA

- The total payload is obtained using SUM() command
- The total payload using NASA boosters is 45596kgs
- Note: this doesn’t in the NASA (CRS) – Kacific partnership



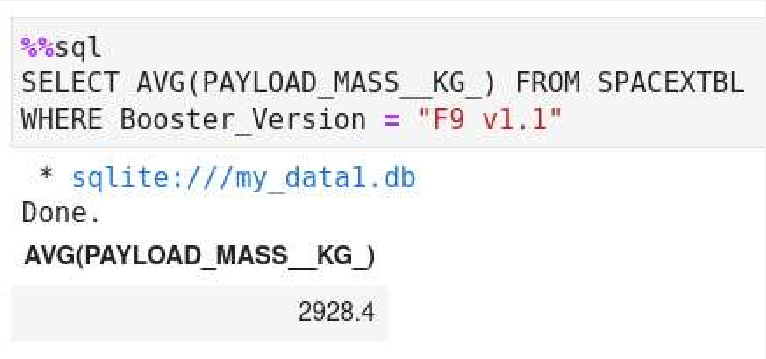
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
SUM(PAYLOAD_MASS__KG_)
                 45596
```

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

- The average payload is obtained using AVG() command
- The average payload using F9 v1.1 boosters is 2928.4kgs



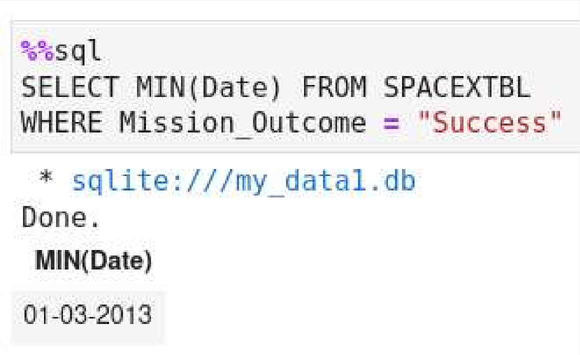
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE Booster_Version = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
AVG(PAYLOAD_MASS__KG_)
                2928.4
```

# First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

- MIN() is just to find the first successful mission
- The first successful mission is in 2013



```
%%sql
SELECT MIN(Date) FROM SPACEXTBL
WHERE Mission_Outcome = "Success"
```

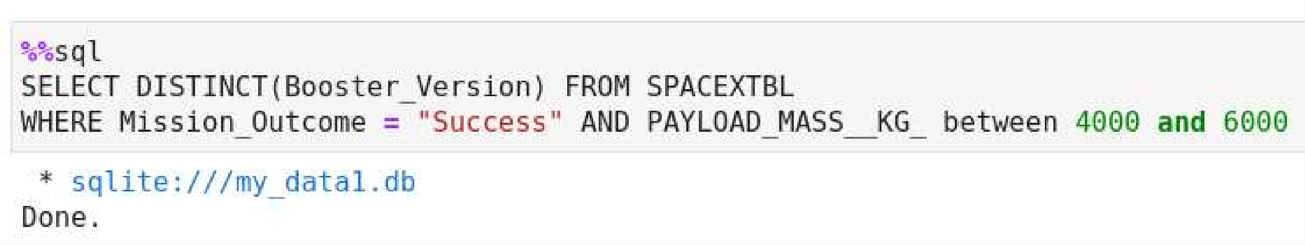
```
 * sqlite:///my_data1.db
Done.
```

| MIN(Date) |
|---|
| 01-03-2013 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

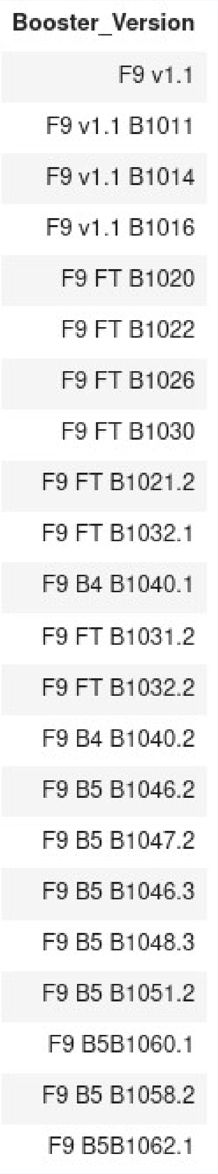
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- DISTINCT() is used to find the unique Booster versions
- There are multiple variation of the F9 booster that have successful missions for payload between 4000 and 6000 kgs



```
%%sql
SELECT DISTINCT(Booster_Version) FROM SPACEXTBL
WHERE Mission_Outcome = "Success" AND PAYLOAD_MASS__KG_ between 4000 and 6000

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

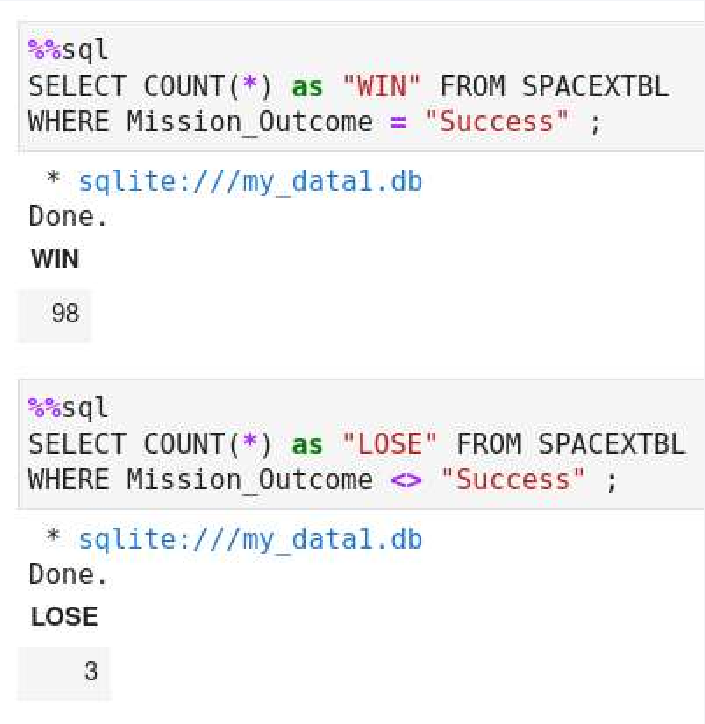


| Booster_Version |
|---|
| F9 v1.1 |
| F9 v1.1 B1011 |
| F9 v1.1 B1014 |
| F9 v1.1 B1016 |
| F9 FT B1020 |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1030 |
| F9 FT B1021.2 |
| F9 FT B1032.1 |
| F9 B4 B1040.1 |
| F9 FT B1031.2 |
| F9 FT B1032.2 |
| F9 B4 B1040.2 |
| F9 B5 B1046.2 |
| F9 B5 B1047.2 |
| F9 B5 B1046.3 |
| F9 B5 B1048.3 |
| F9 B5 B1051.2 |
| F9 B5B1060.1 |
| F9 B5 B1058.2 |
| F9 B5B1062.1 |

# Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

- COUNT() was used to determine the number of successful and failed missions
- There were 98 successful missions and 3 failures



```
%%sql
SELECT COUNT(*) as "WIN" FROM SPACEXTBL
WHERE Mission_Outcome = "Success" ;
```

```
 * sqlite:///my_data1.db
Done.
WIN
 98
```

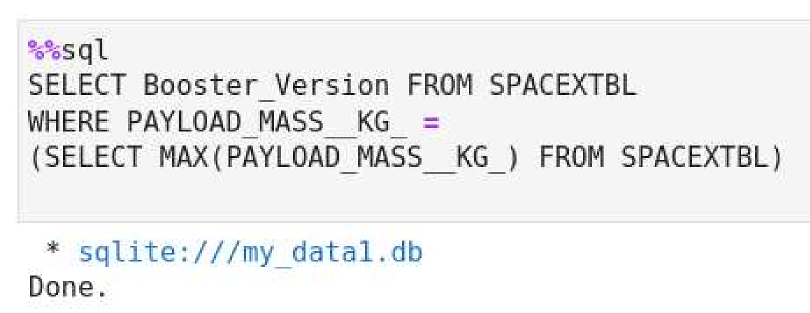
```
%%sql
SELECT COUNT(*) as "LOSE" FROM SPACEXTBL
WHERE Mission_Outcome <> "Success" ;
```

```
 * sqlite:///my_data1.db
Done.
LOSE
   3
```

# Boosters Carried Maximum Payload

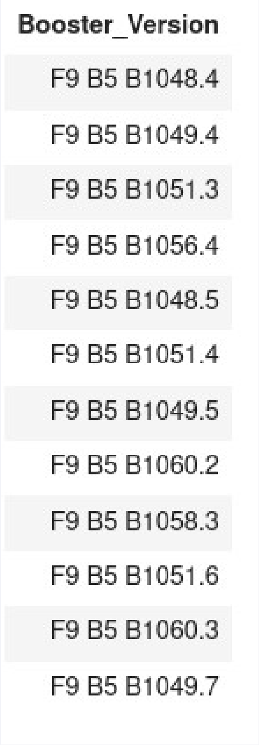
- List the names of the booster which have carried the maximum payload mass

- A combination of subquery and MAX() are used to determine which boosters have carried the max payload
- There are 12 booster versions that have carried the max payload



```
%%sql
SELECT Booster_Version FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

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

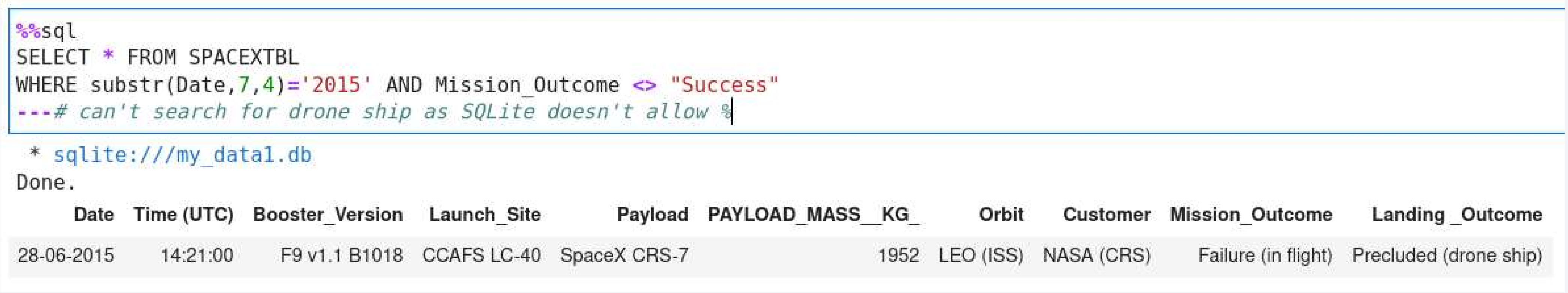


| Booster_Version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1049.4 |
| F9 B5 B1051.3 |
| F9 B5 B1056.4 |
| F9 B5 B1048.5 |
| F9 B5 B1051.4 |
| F9 B5 B1049.5 |
| F9 B5 B1060.2 |
| F9 B5 B1058.3 |
| F9 B5 B1051.6 |
| F9 B5 B1060.3 |
| F9 B5 B1049.7 |

# 2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- A WHERE clause is used to look at failure outcomes in 2015
- There was one failed outcome involving a drone ship in 2015



```
%%sql
SELECT * FROM SPACEXTBL
WHERE substr(Date,7,4)='2015' AND Mission_Outcome <> "Success"
---# can't search for drone ship as SQLite doesn't allow %
```

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

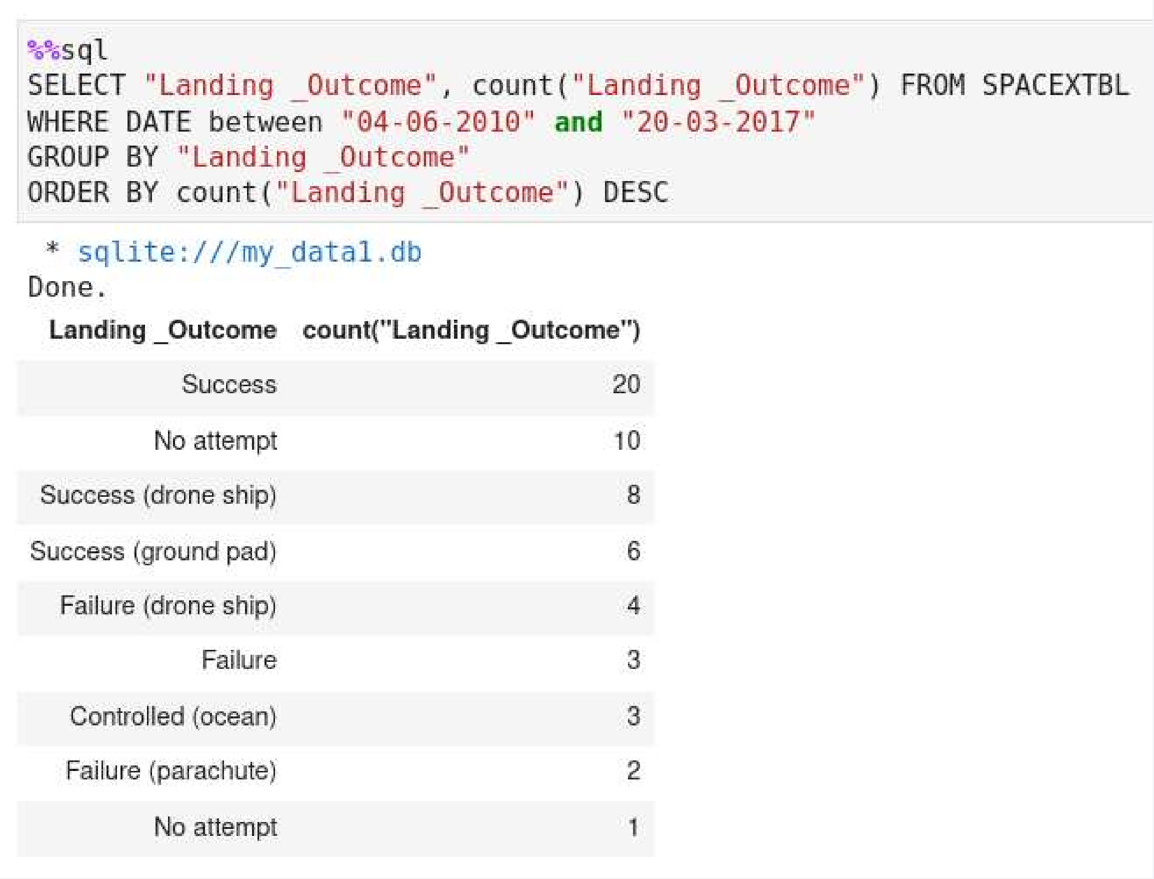
| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 28-06-2015 | 14:21:00 | F9 v1.1 B1018 | CCAFS LC-40 | SpaceX CRS-7 | 1952 | LEO (ISS) | NASA (CRS) | Failure (in flight) | Precluded (drone ship) |

# 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

- GROUP BY and ORDER BY where used to rank landing outcomes

- Bwtween 2010-06-04 and 2017-03-20 the most common mission outcome was Success



```
%%sql
SELECT "Landing _Outcome", count("Landing _Outcome") FROM SPACEXTBL
WHERE DATE between "04-06-2010" and "20-03-2017"
GROUP BY "Landing _Outcome"
ORDER BY count("Landing _Outcome") DESC
```

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

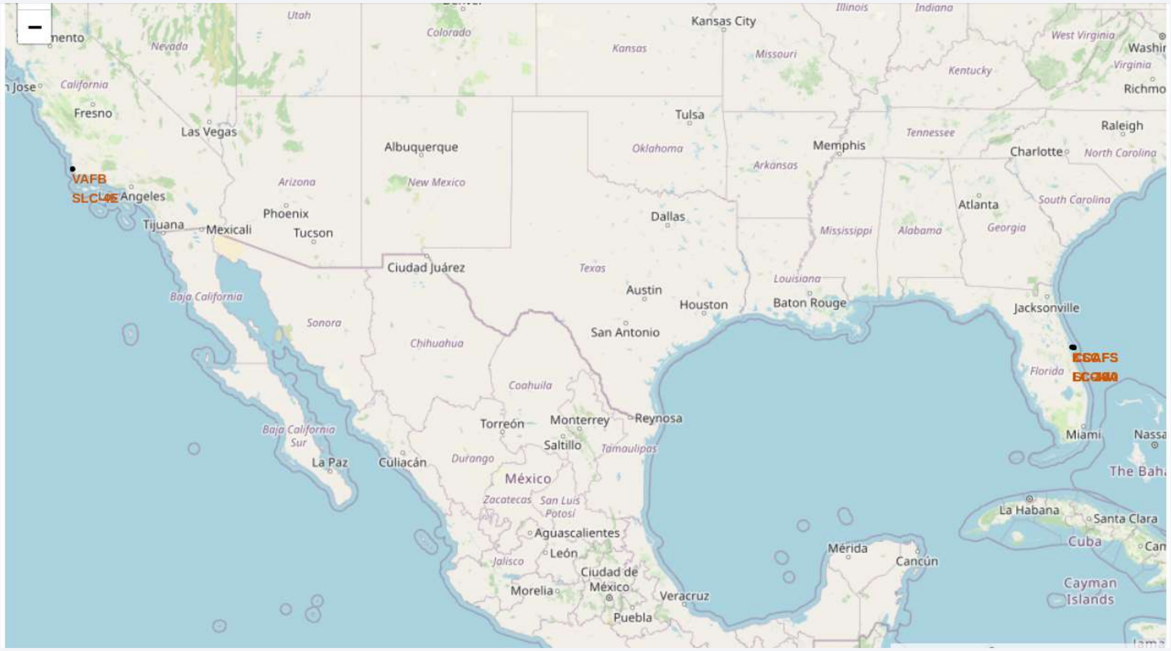
| Landing _Outcome | count("Landing _Outcome") |
|---:|---:|
| Success | 20 |
| No attempt | 10 |
| Success (drone ship) | 8 |
| Success (ground pad) | 6 |
| Failure (drone ship) | 4 |
| Failure | 3 |
| Controlled (ocean) | 3 |
| Failure (parachute) | 2 |
| No attempt | 1 |

Section 3

# Launch Sites Proximities Analysis

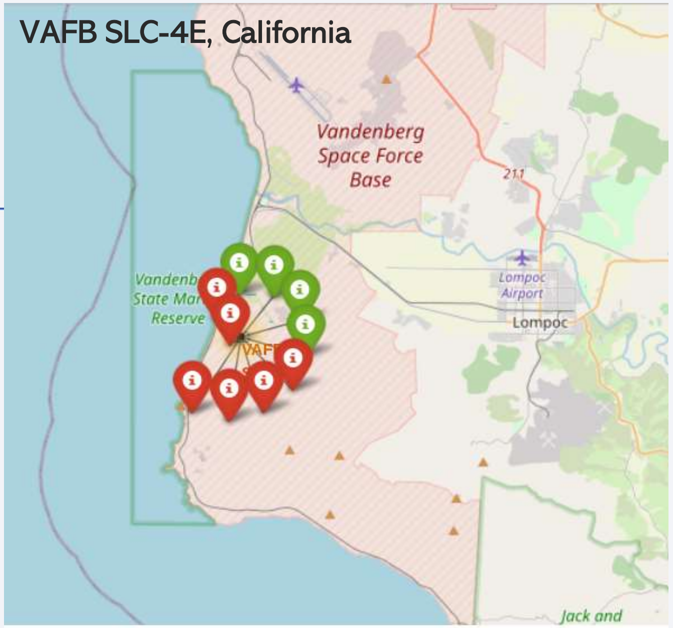
# Launch Site Locations

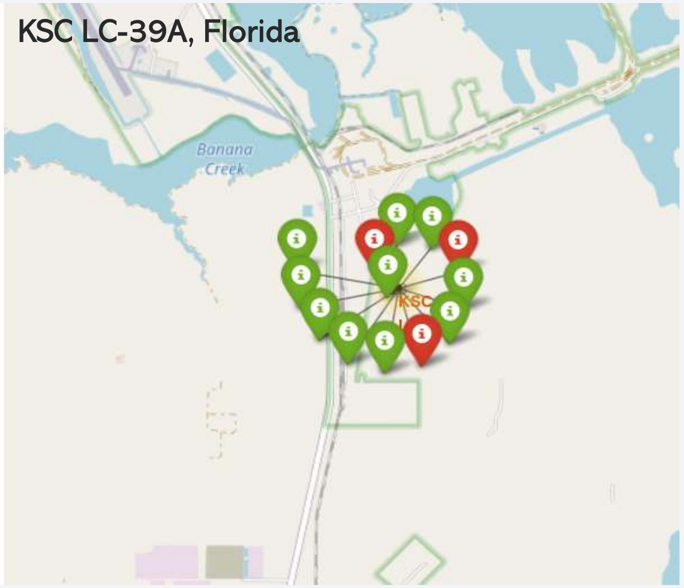
- There are 4 unique launch sites
- One (VAFB SLC-4E) is in California,
- Three (KSC LC-39A, CCAFS LC-40, CCAFS SLC-40) are in close proximity to each other in Florida

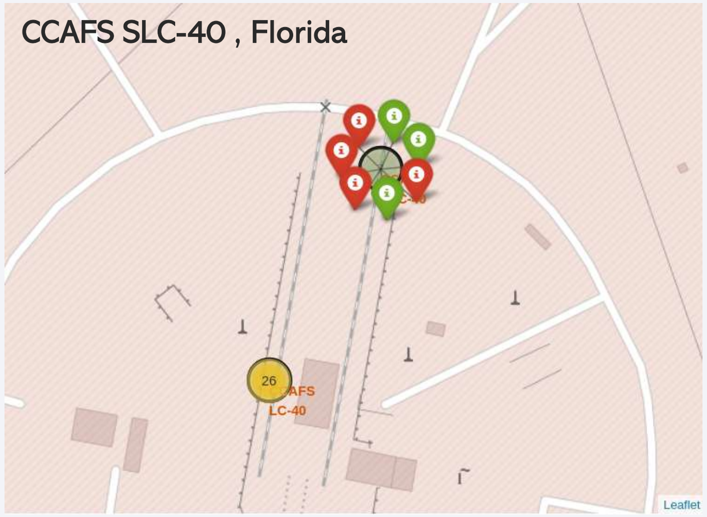


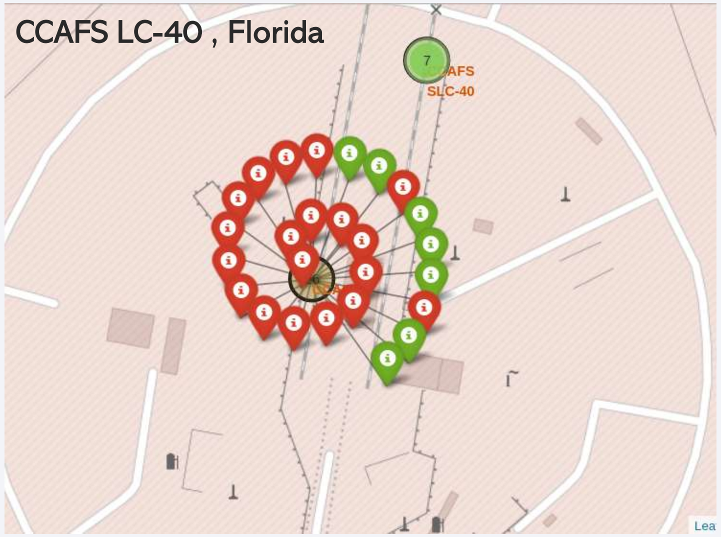
# Launch Success per Launch Site

- Markers indicate launch outcome (Green-success, Red-failure) at each site
- Most launches where at CCAFS LC-40, and CCAFS has the least launches
- KSC LC-39A as the most successful launches


VAFB SLC-4E, California

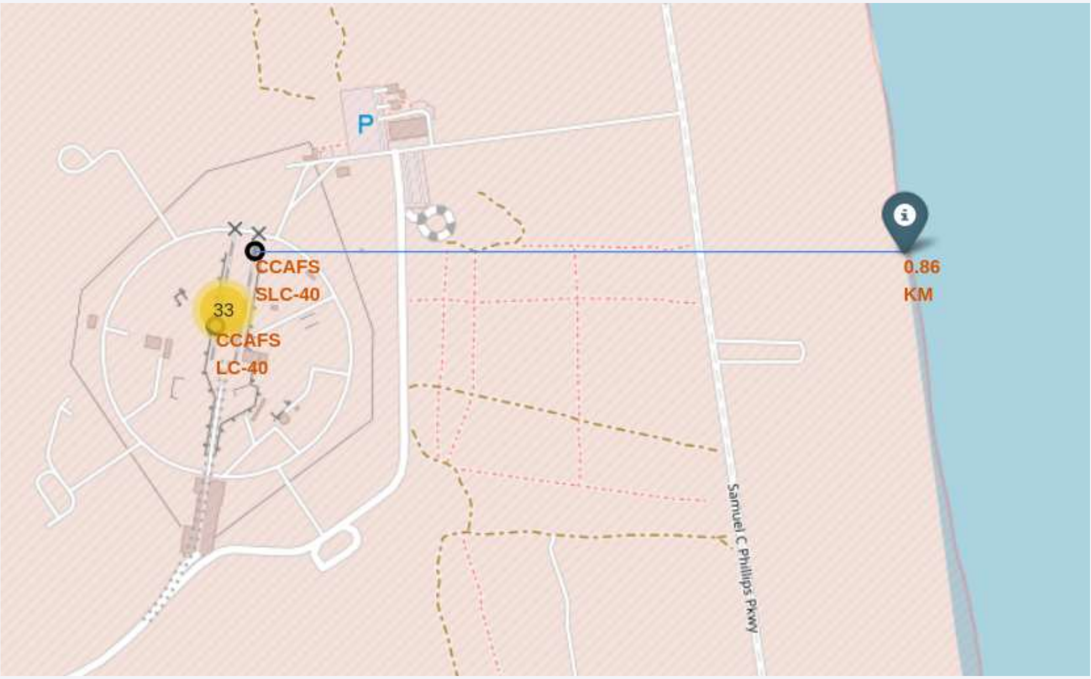

KSC LC-39A, Florida

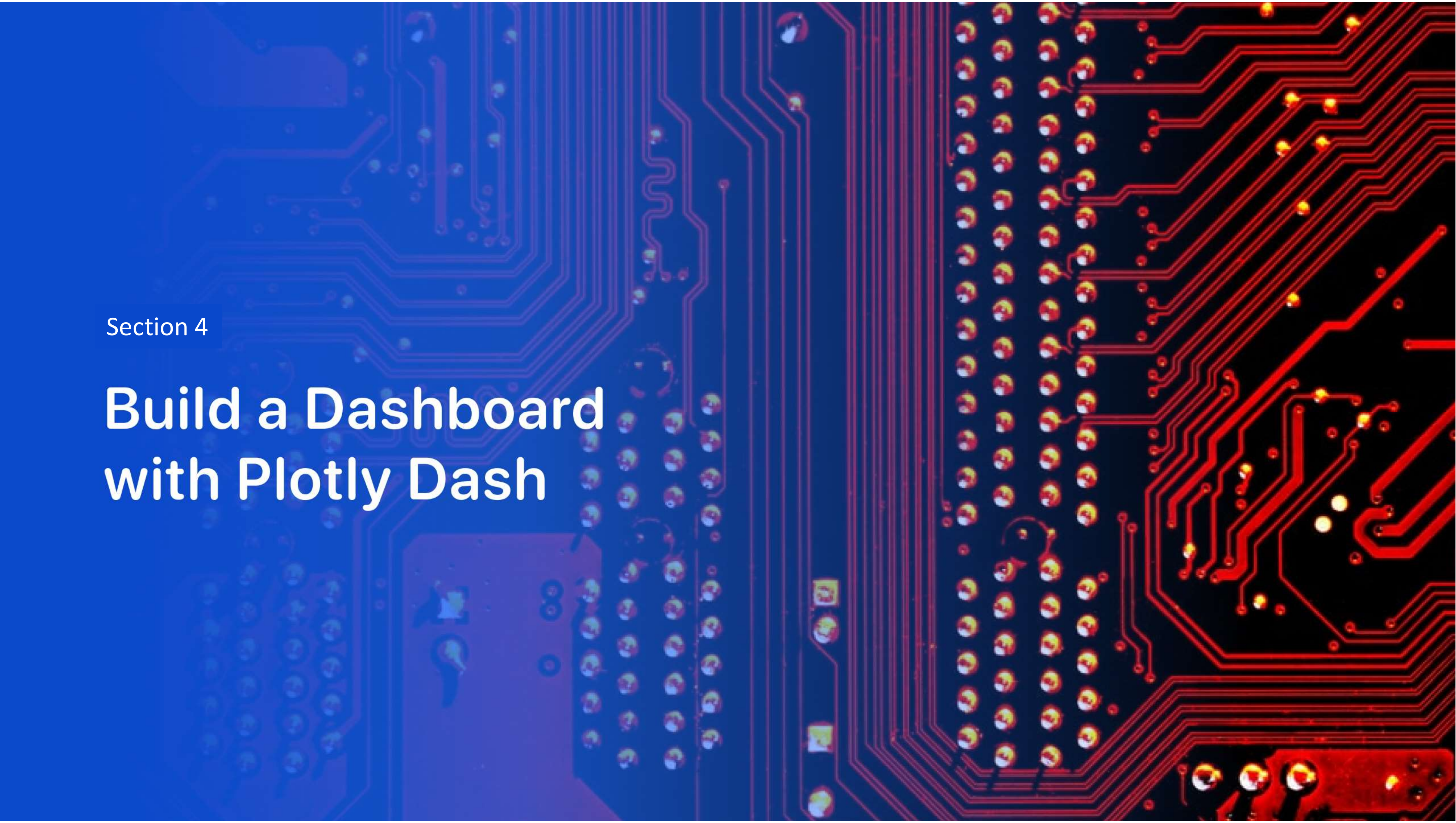

CCAFS SLC-40 , Florida


CCAFS LC-40 , Florida

# Site Proximity to Urban Features

- CCAFS SLC-40 is 0.86 km from the cost
- In California, VAFB SLC-4E is on the coast and in the proximity of the city Lompoc and Santa Barbara MT1 railway
- For three launch sites in Florida, the closest city is Titusville, KSC LC-39A is the closest to a railway line and freeway while CCAFS SLC-40 is the closest to the cost

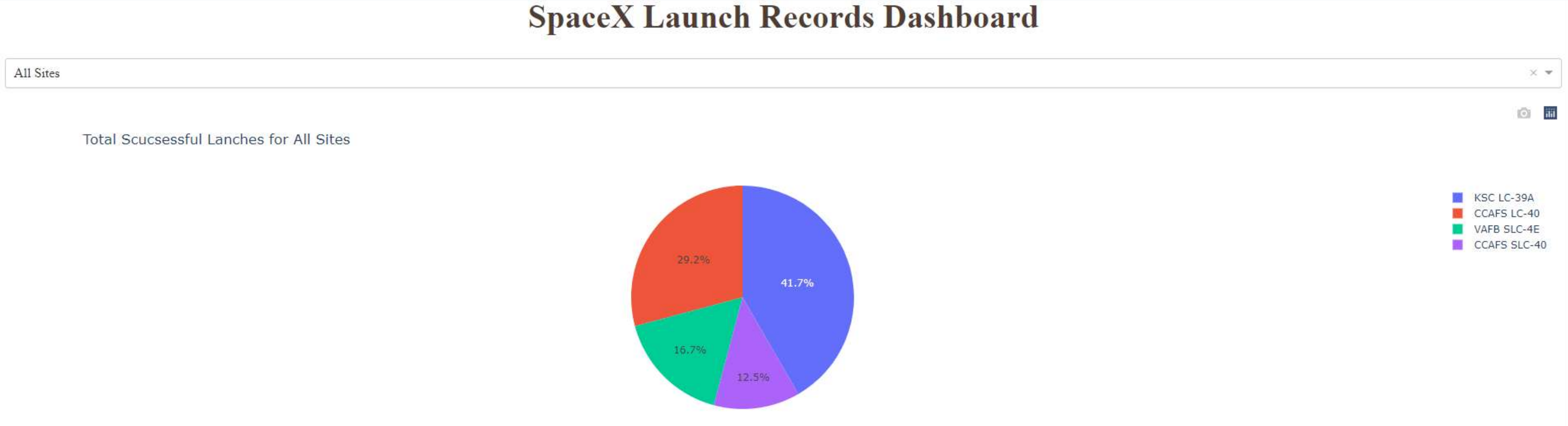




Section 4

# Build a Dashboard with Plotly Dash

# Successful Launches for All Sites



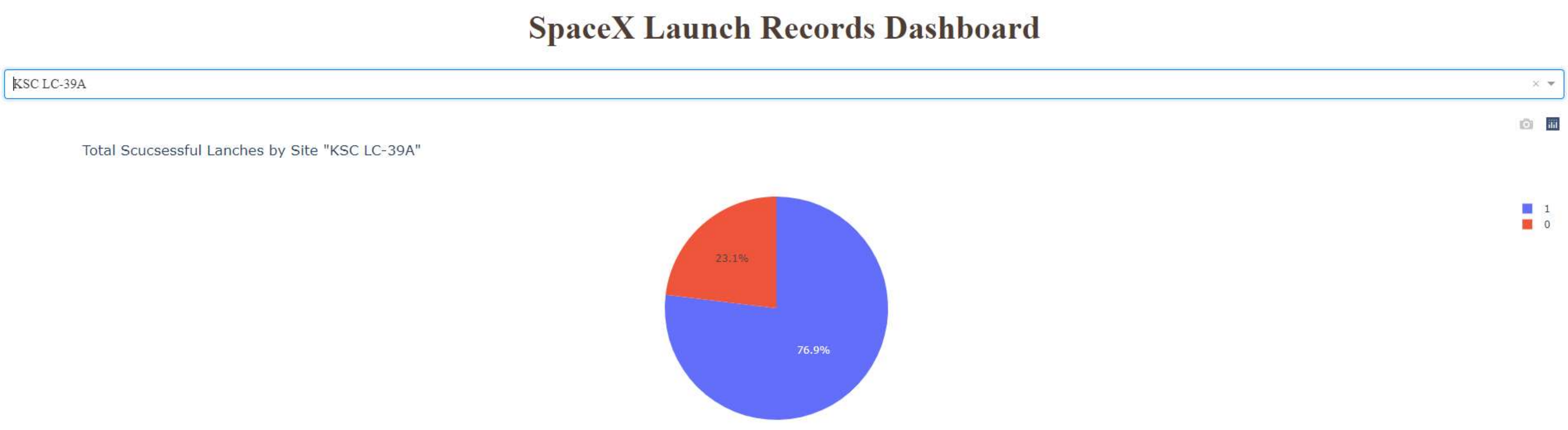
SpaceX Launch Records Dashboard

All Sites

Total Scucsessful Lanches for All Sites

- KSC LC-39-A is the site with the most successful launches at 41.7%
- CCAFS SLC-40 is the site with the least successful launches at 12.5%

# Mission Outcomes for KSC LC-39A



SpaceX Launch Records Dashboard

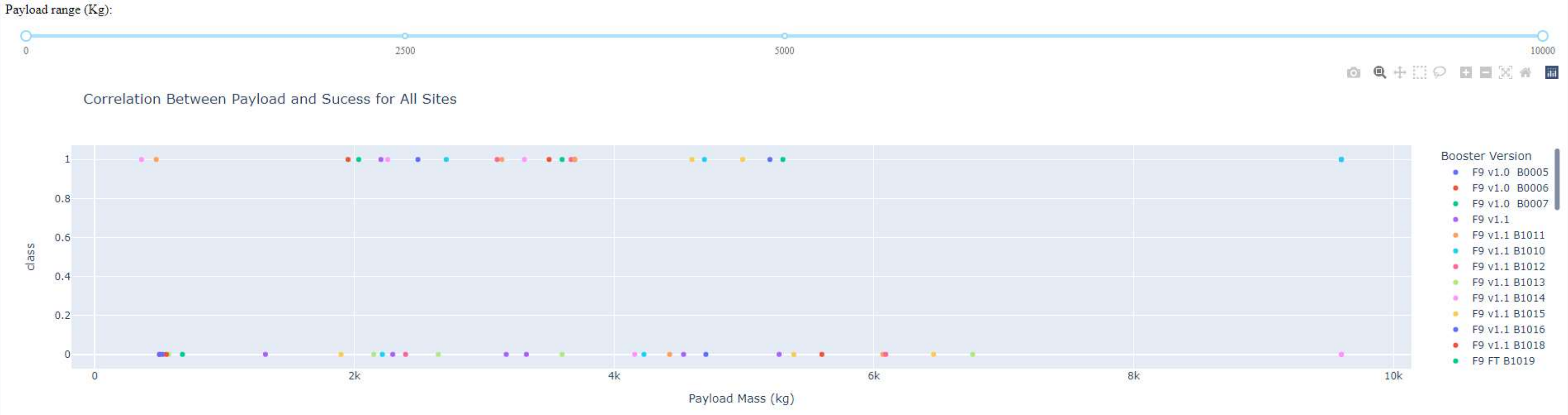
KSC LC-39A

Total Scucsessful Lanches by Site "KSC LC-39A"

- KSC LC-39-A is the site with the most successful launches at 41.7%
- At KSC LC-39-A, 76.9% of missions are successful and 23.1% of missions are unsuccessful

# Successful Launches for All Sites

All Sits and All Payloads

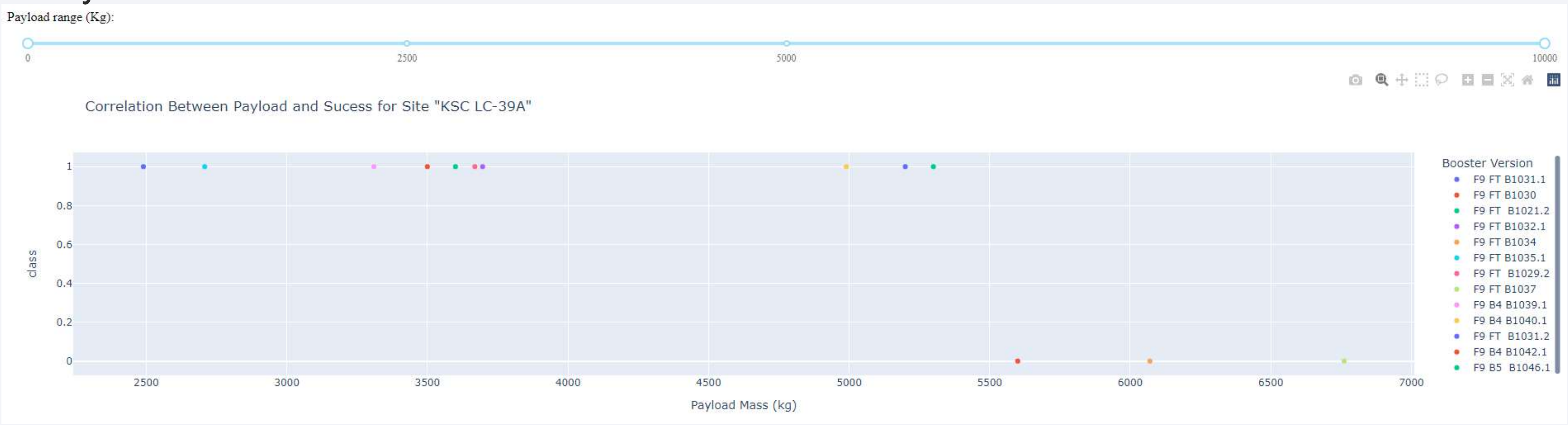


- There is a mix of missions outcome when payload is < 5000kg
- When the payload is > 5000kg the mission is more likely to be a success

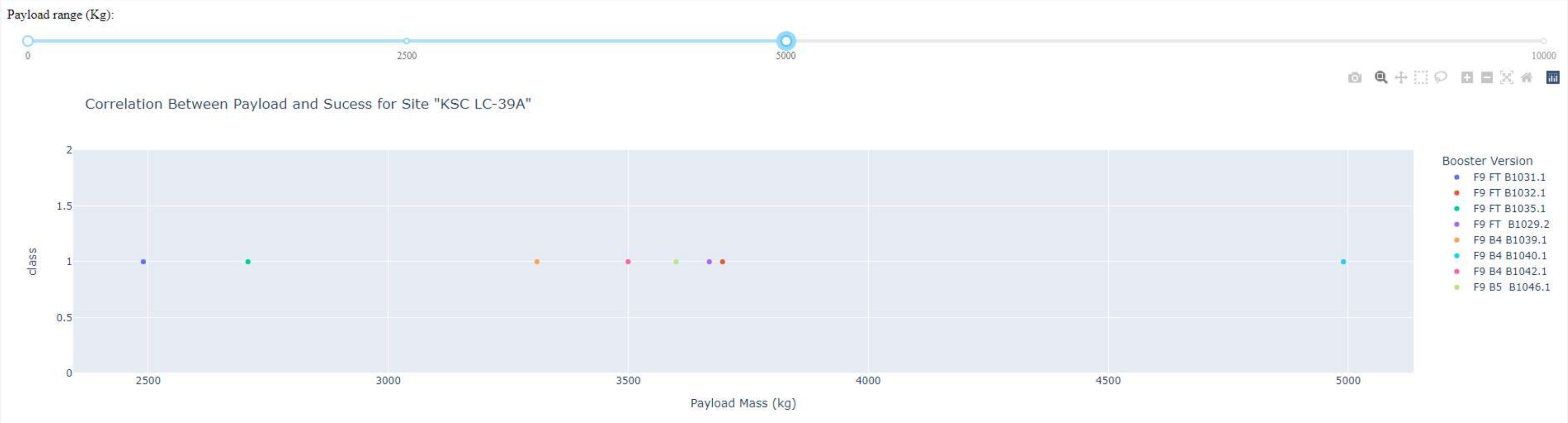
# Payload for KSC LC-39-A

- At KSC LC-39A, adjusting the payload to be between 0 and 5000 kg give a 100% successful mission outcomes

## All Payloads



## Payloads between 0 and 5000 kg



Section 5

# Predictive Analysis (Classification)

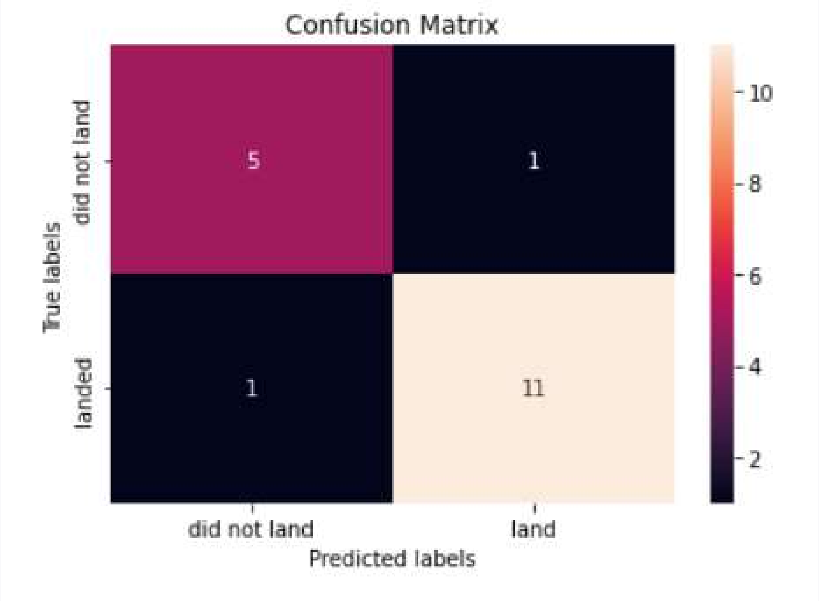
# Classification Accuracy

- Logistic Regression, SVM, K-Nearest Neighbour all displayed similar model accuracy 0.833 on the test data
- The best model accuracy on the test data was 0.888 from the decision tress model with the hyper parameters
  - Criteron – entropy
  - Max_depth - 12
  - Max_features – auto
  - Min_samples_leaf – 4
  - Min_samples_split – 2
  - Splitter - best

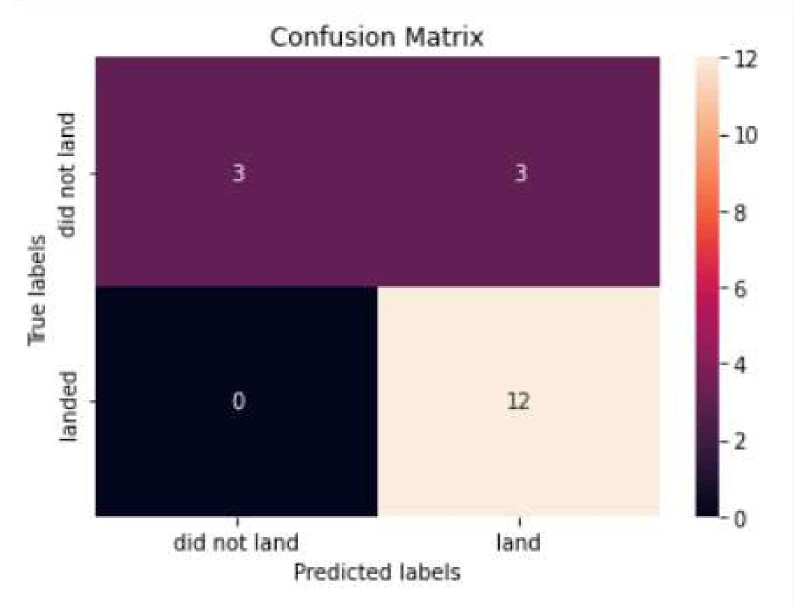
| PModel Type | Training Accuracy | Test Accuracy |
|---|---|---|
| Logistic Regression | 0.846 | 0.833 |
| SVM | 0.848 | 0.833 |
| Decision Tree | 0.889 | 0.888 |
| K-Nearest Neighbour | 0.848 | 0.833 |

# Confusion Matrix

Decision Tress Test Confusion Matrix



SVM Test Confusion Matrix



- The decision tress confusion matrix show that it only produced on False Negative and one False positive
- The SVM confusion matrix (indicative of k-nearest neighbour and Logistic Regression as well) show that these models are 100% accurate at predicting if the first stage landed, however it's a 50-50 chance of predicting the first stage did not land correctly

# Conclusions

- The best model was the decision tress which returned a test model accuracy of 0.888
  - The other tested model are 100% accurate at predicting if the mission is a success however are only 50% accurate at predicting is the mission is a failure
- KSC LC-39A is the site with the most successful launch mission and if the payload is <5000 kg then it has 100% success rate
- As the years increase the number of mission attempts increases and the success rate goes up
- As flight numbers increase there’s a shift to VLEO and SSO orbit which are more successful

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

# Thank you!