IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Mario Grandi

21 Aug 2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection and Wrangling
  - Exploratory Data Analysis with:
    - Data Visualisation
    - SQL
    - Folium interactive Map
  - Machine Learning prediction

- Summary of all results
  - Valuable data has been collected from public sources
  - Features of interest to predict the successfulness of rocket launches have been identified using Exploratory Data Analysis
  - Predictive analysis shows impact of important characteristics and identifies highest performing model

# Introduction

- In this project the viability of rocket company Y is evaluated to compete against Space X.
  - SpaceX is a leading commercial space company known for its cost-effective Falcon 9 rocket launches, which are significantly cheaper due to the reusability of the first stage.
    - Has successfully sent spacecraft to the International Space Station
    - Has deployed the Starlink satellite internet constellation
    - Continues to conduct manned space missions.
- The following criteria have been identified to make company Y competitively viable:
  - Estimation of success rate of first stage landings based on variables such as Payload, launch site, number of flights, orbits etc…
  - Identify best algorithm to perform binary classification and predict launch outcome

Section 1

# Methodology

# Methodology Overview

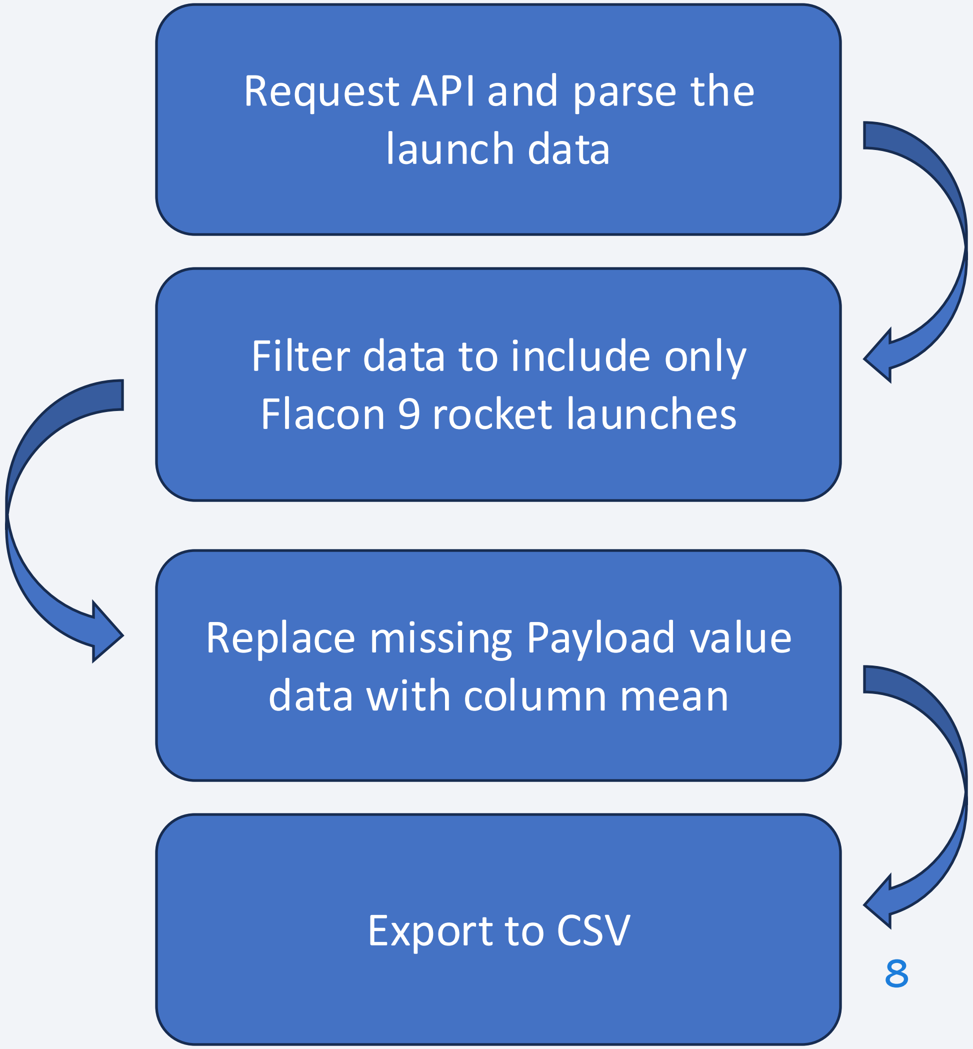
- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scraping from public information sources
- Perform data wrangling
  - Filtering data
  - Handling missing values
  - Using One Hot encoding to prepare data for binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning, and evaluation of classification models to ensure highest predictive performance

# Data Collection

- Falcon9 launch data is collected using a combination of methods and sources to obtain a comprehensive dataset that could be used to perform a more detailed analysis.
  - SpaceX REST API (https://api.spacexdata.com/v4/rockets/) is used to obtain:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
  - Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) is used to obtain:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch, outcome, Version Booster, Booster landing, Date, Time

# Data Collection – SpaceX API

- SpaceX provides a public API for accessing and using their data
- Data is collected, cleaned, and used according to the diagram
- GitHub URL : SpaceX API data Collection



Request API and parse the launch data

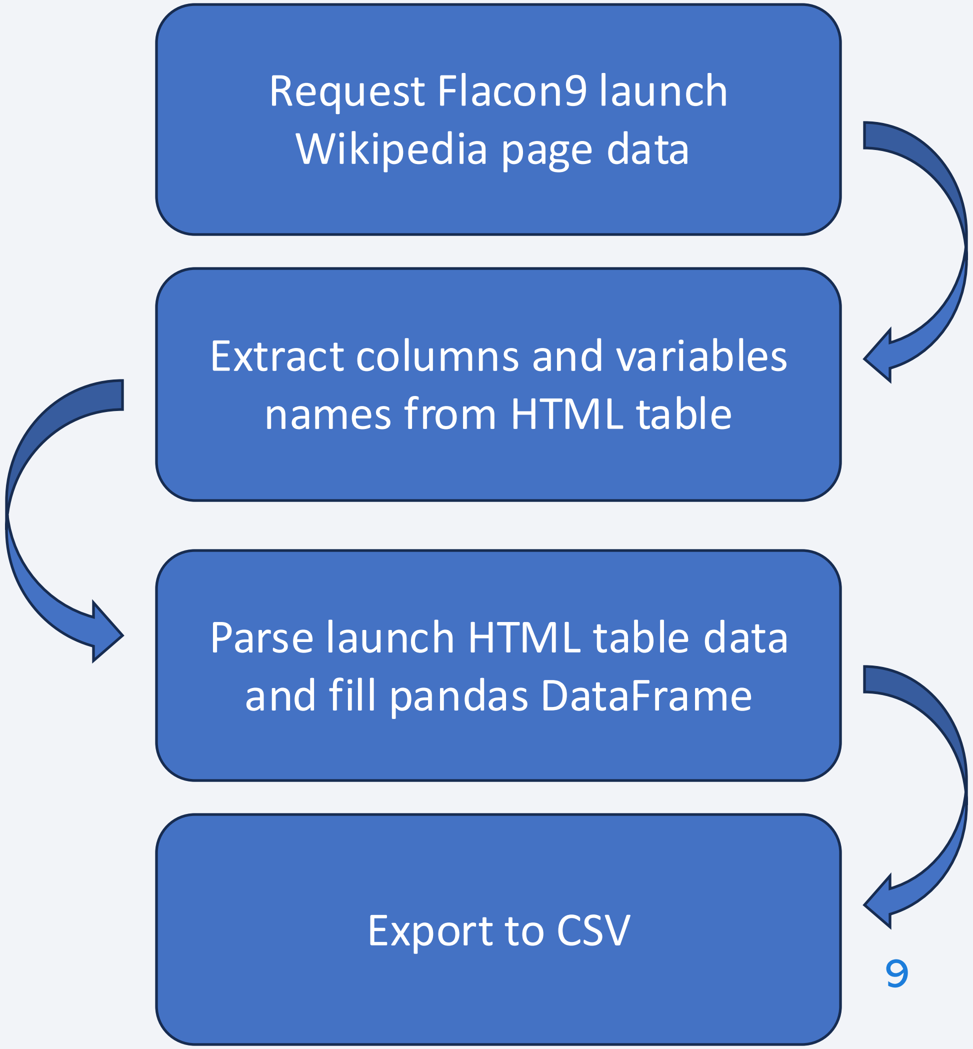
Filter data to include only Flacon 9 rocket launches

Replace missing Payload value data with column mean

Export to CSV

# Data Collection - Scraping

- Scraped data on SpaceX Falcon9 launches from public source (Wikipedia)
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose
- GitHub URL: Web Scraping



Request Flacon9 launch Wikipedia page data

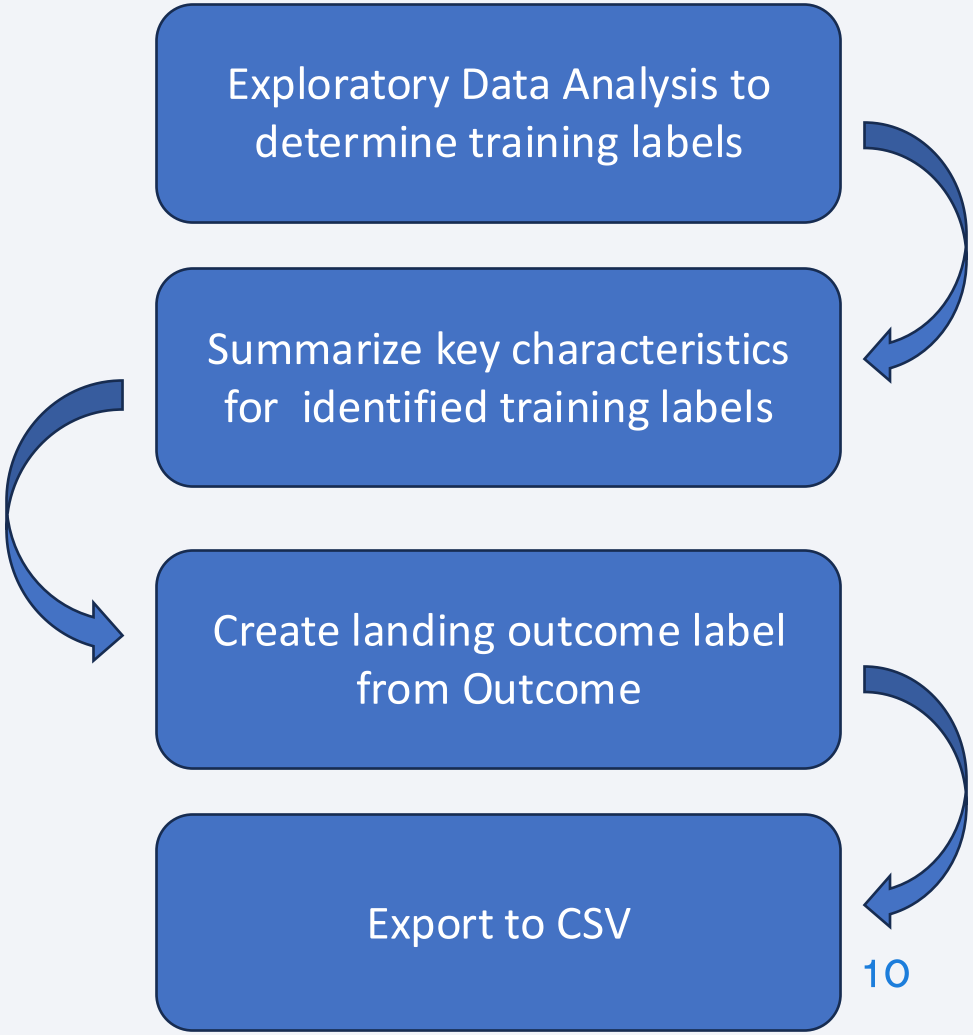
Extract columns and variables names from HTML table

Parse launch HTML table data and fill pandas DataFrame

Export to CSV

# Data Wrangling

- A set of training labels is determined based on an initial EDA
- Collected data contains different success and fail states for the landings of the Falcon9 booster.
- The different landing outcomes are converted to a binary training label column named *outcomes* containing 1 if booster landed successfully and 0 if booster landing failed.
- GitHub URL: Data Wrangling


Exploratory Data Analysis to determine training labels

Summarize key characteristics for  identified training labels

Create landing outcome label from Outcome

Export to CSV

# EDA with Data Visualization

- Generated Charts:
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Orbit type vs Success Rate
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Success rate yearly trend

- GitHub URL: EDA with Data Visualization

- Scatter plots are used to determinte continuous trends between variables. Any exisitng relationships could be used by a machine learning model to predict outcomes.
- Bar charts show comparison between discrete categories to show the relationship between the category to a measured value.
- Line charts are used to show data trends over time (time series)

# EDA with SQL

- SQL queries:
  - Display launch site names
  - Display top 5 records with launch site name starting with 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - 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.
  - 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

- GitHub URL: EDA with SQL

# Build an Interactive Map with Folium

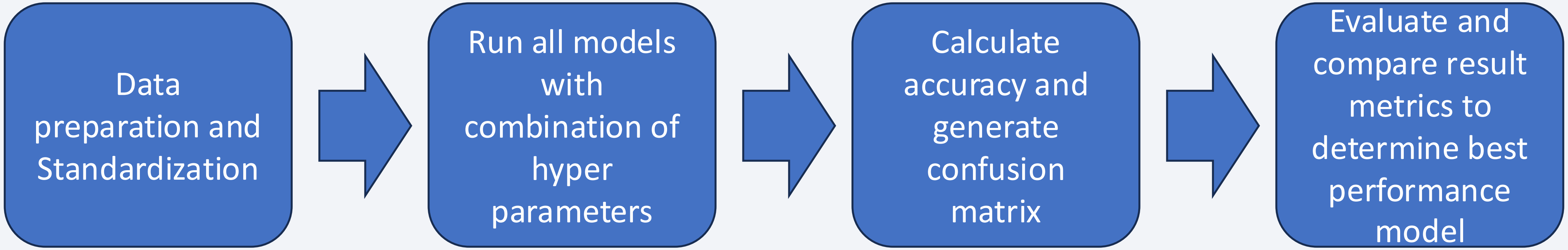
- Markers for all launch sites:
  - Added Circle, popup label and text label for NASA Johnson Space Center and all Falcon9 launch sites using longitude and latitude coordinate information
- Red/Green Markers:
  - Added Red and Green marker to represent the Failed and Successful Falcon9 launches at each site. This was used to identify sites with relatively high success rate.
- Distance Markers with color lines:
  - Added Distance markers and lines between launch site CCAFS SLC-40 to closest coastline, railway, highway, and city.
- GitHub URL: Folium Interactive Map

# Build a Dashboard with Plotly Dash

- Interactive Dashboard includes:
  - Dropdown list for site selection
  - Pie Chart displaying successful launches for each and all sites
  - Slider bar to select custom Payload Mass ranges to investigate
  - Scatter chart of Payload Mass vs mission outcome for different booster versions with Payload mass range selected according to slider. Success rate is also shown for this given Payload mass range.
- GitHub URL: Plotly Dash

# Predictive Analysis (Classification)

- Four classification models have been built, trained, evaluated, and compared:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbour


Data preparation and Standardization
Run all models with combination of hyper parameters
Calculate accuracy and generate confusion matrix
Evaluate and compare result metrics to determine best performance model

- GitHub URL: Predictive Analysis

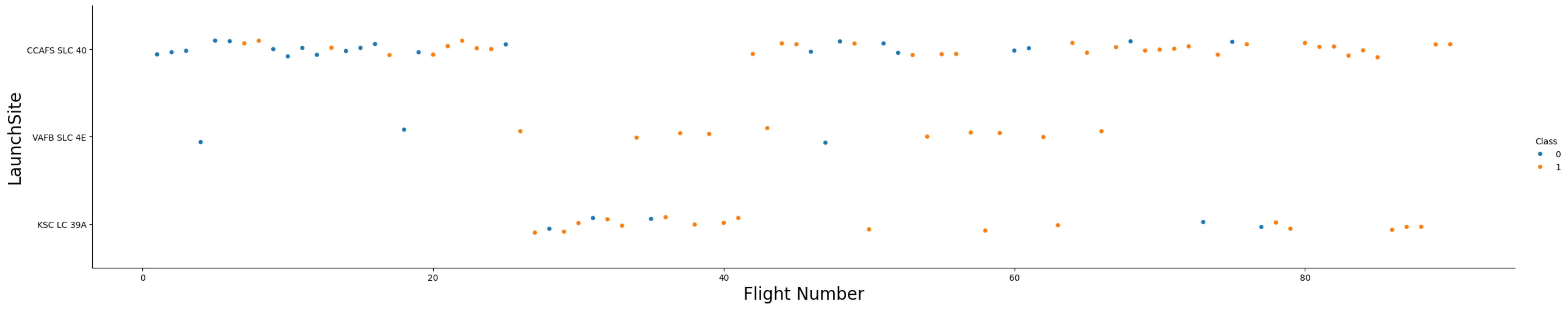
# Results

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

Section 2

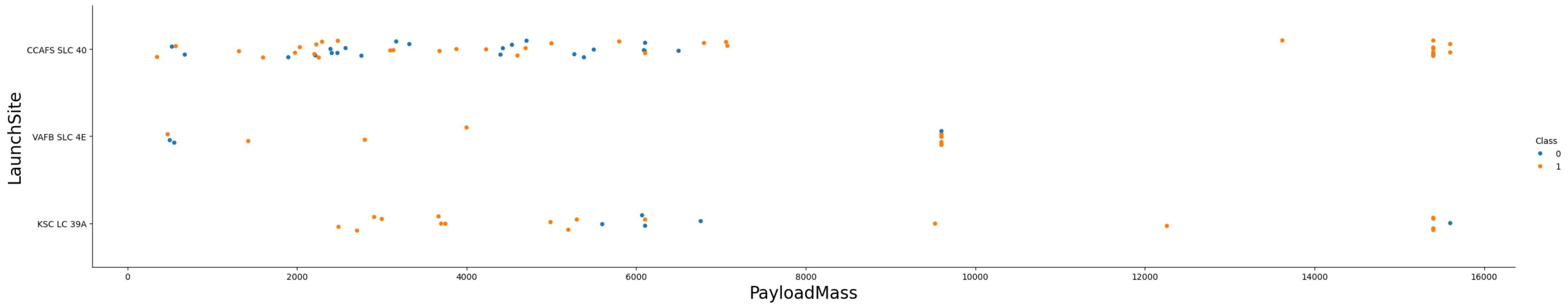
# Insights drawn from EDA

# Flight Number vs. Launch Site



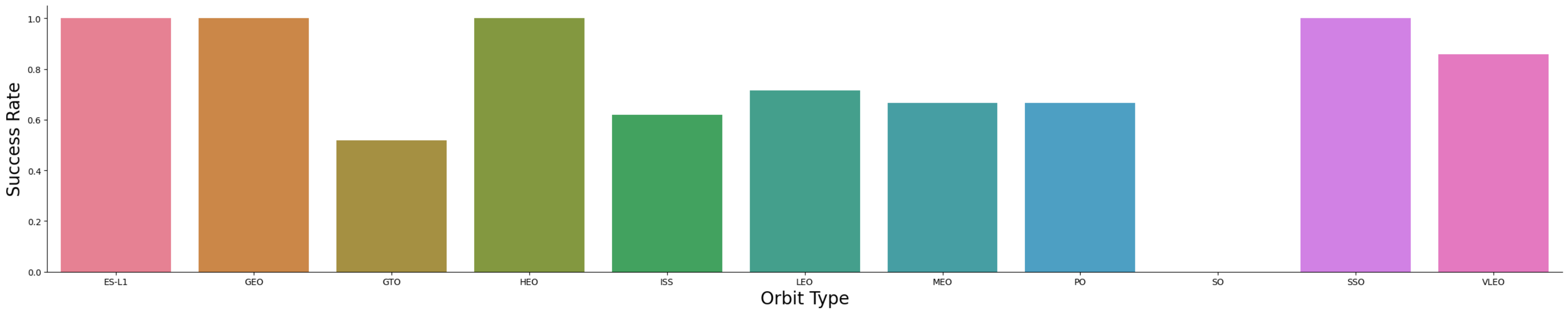
- CCASFS SLC 40 is the launch site with the highest number of launches followed by KSC LC 39A and VABFB SLC 4E
- Due to the higher number of launches, CCASFS SLC 40 is the site with the highest number of successful launches. However, KSC LC 39A if found to be the launch site with the highest success outcome, at it possesses the largest number of successful out of total launches.
- It is important to note that the launch success rate increases with the Flight Number, indicating that successful launches become more common over time throughout all launch sites.

# Payload vs. Launch Site



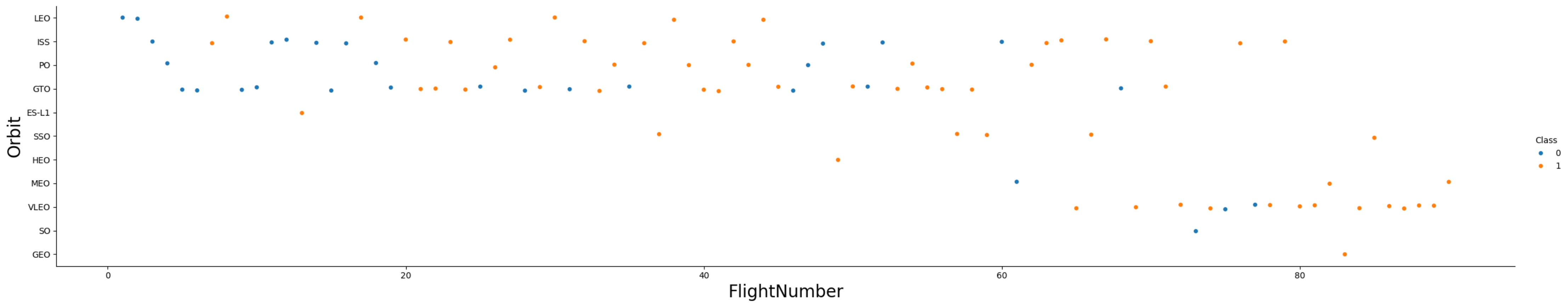
- There are no launches at site VAFB SLC 4E for Payloads above 1 ton
- Success rate seems to be higher for payloads above 7000 kg in all sites
- KSC LC 39A has 100% success rate for payloads below 5500 kg.

# Success Rate vs. Orbit Type



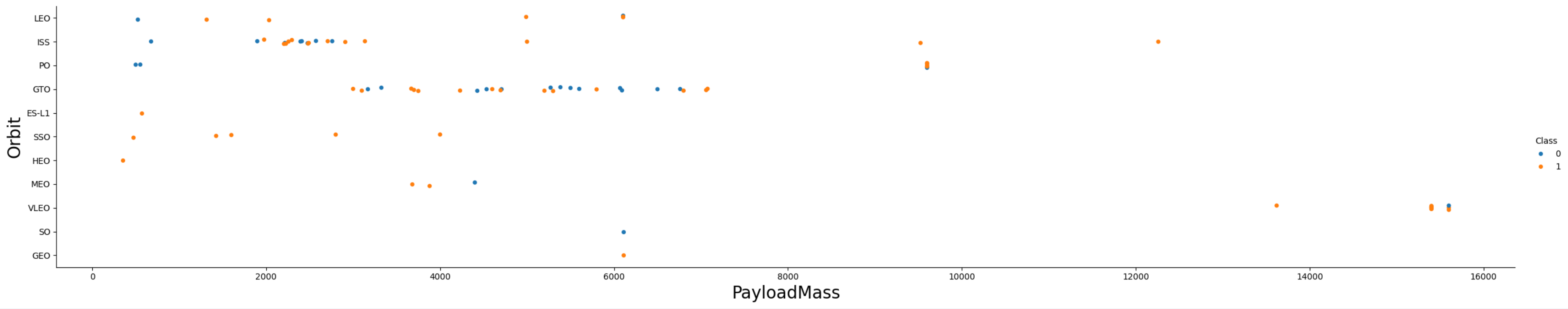
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rate
- Orbit SO has 0% success rate
- Orbits GTO, ISS, LEO, MEO,PO, and VLEO have success rates between ~50% and 85%

# Flight Number vs. Orbit Type



- LEO orbit success is related to the number of Flights
- GTO orbit has no relation between number of flights and success rate.

# Payload vs. Orbit Type

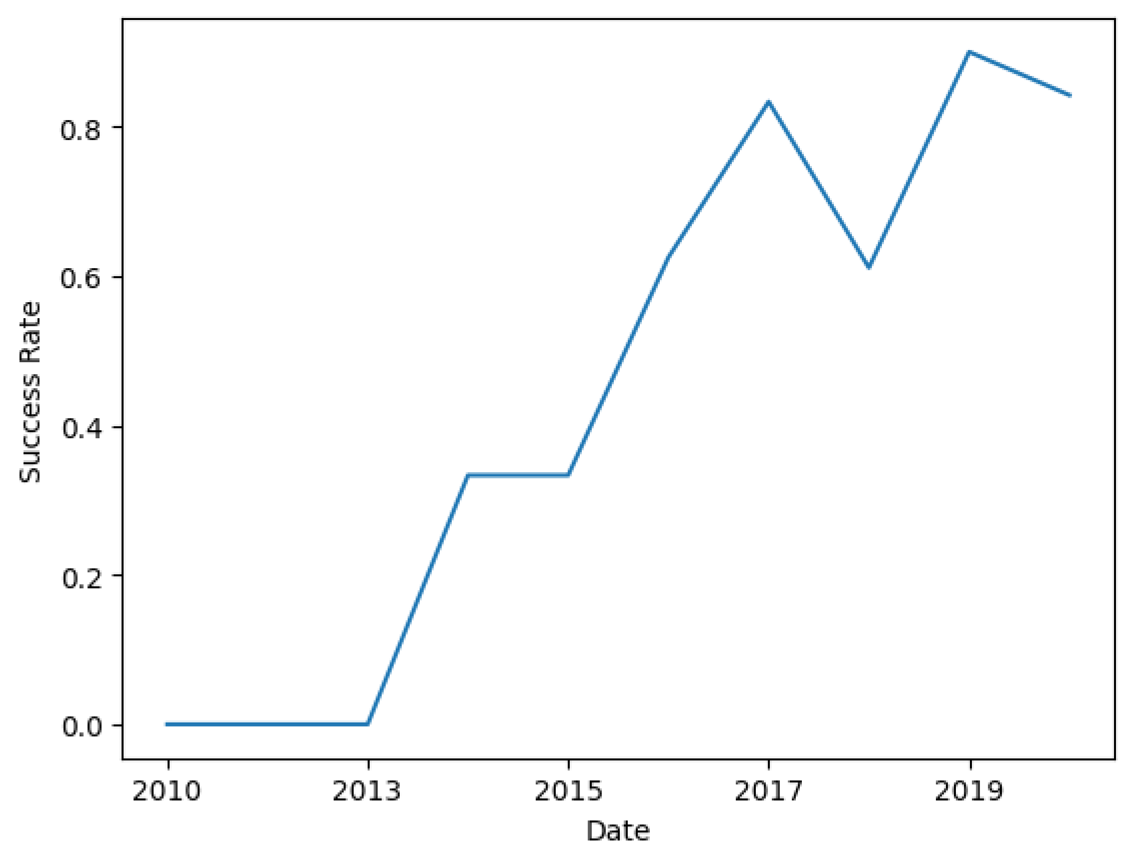


- Successful landing rates are higher for Polar, LEO and ISS at higher payloads.

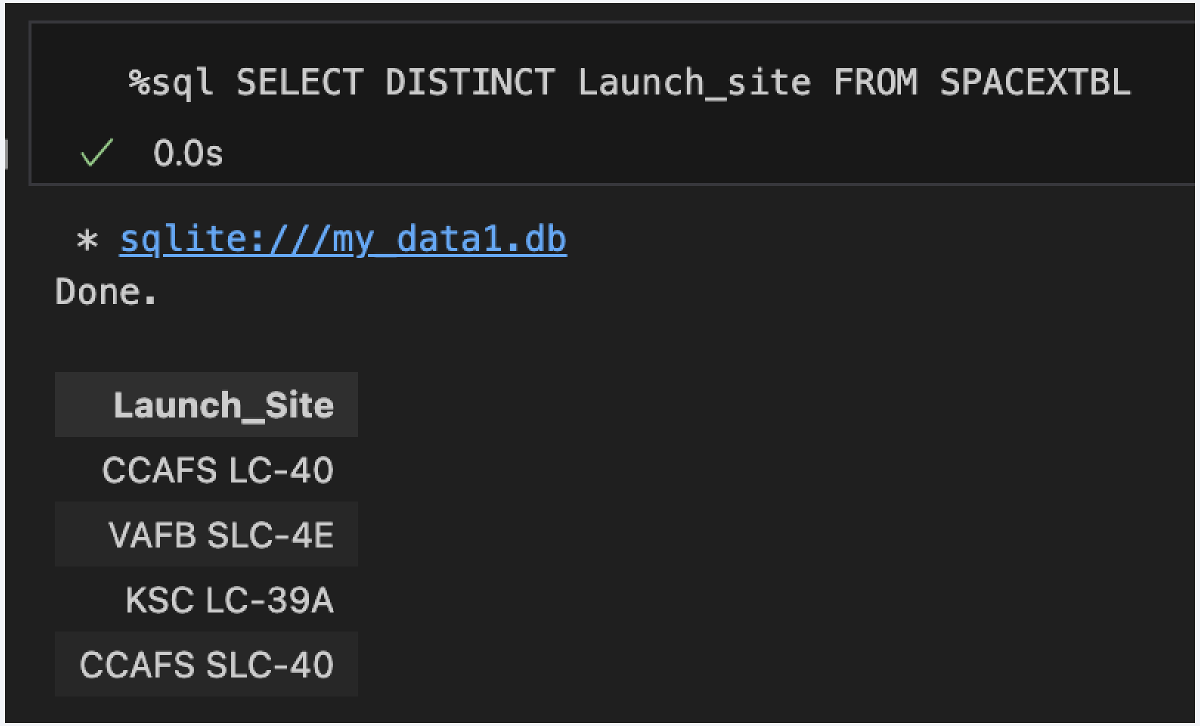
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

# Launch Success Yearly Trend

- Success rate has steadily increased since 2013 up to 2020



# All Launch Site Names



```
%sql SELECT DISTINCT Launch_site FROM SPACEXTBL
```

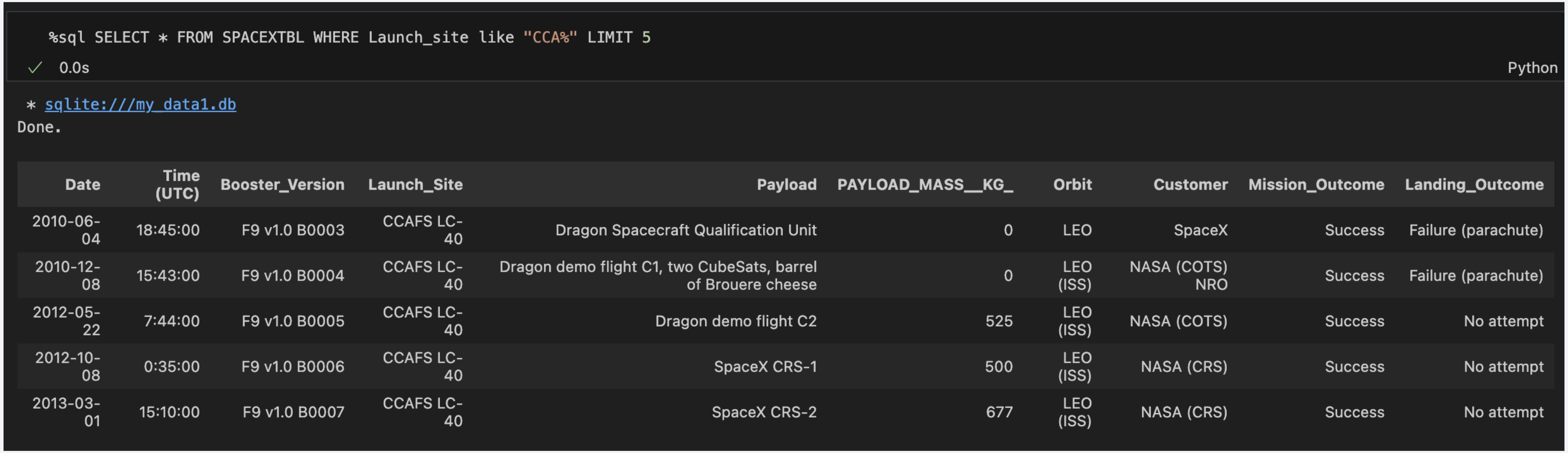
✓ 0.0s

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

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

- Displaying names of unique launch sites from SpaceX space mission data

# Launch Site Names Begin with 'CCA'



```
%sql SELECT * FROM SPACEXTBL WHERE Launch_site like "CCA%" LIMIT 5
```

✓ 0.0s

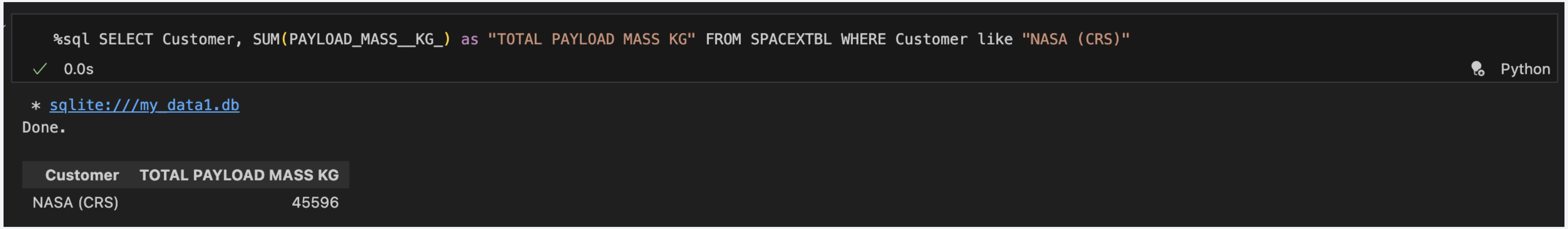
Python

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

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 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) |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

- Displaying first 5 data entry points for launch site name starting with 'CCA'

# Total Payload Mass



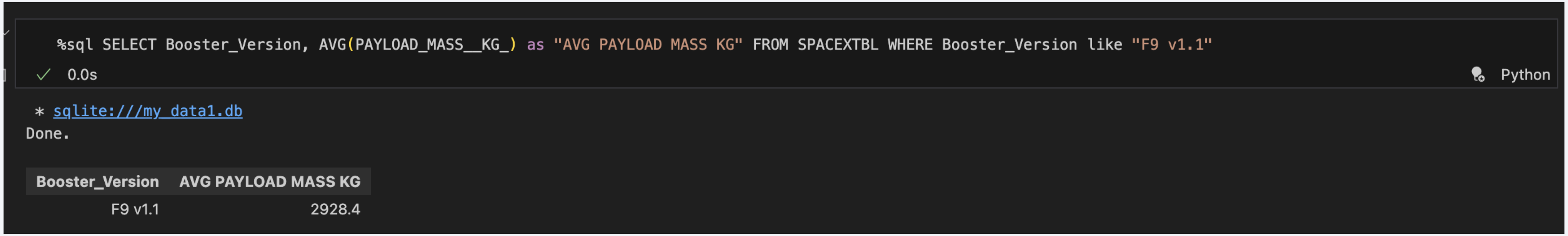
```
%sql SELECT Customer, SUM(PAYLOAD_MASS__KG_) as "TOTAL PAYLOAD MASS KG" FROM SPACEXTBL WHERE Customer like "NASA (CRS)"
```

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

| Customer | TOTAL PAYLOAD MASS KG |
|---|---|
| NASA (CRS) | 45596 |

- Displaying total payload mass in Kg carried by boosters launched by NASA (CRS) by summing all payload masses from this customer.

# Average Payload Mass by F9 v1.1



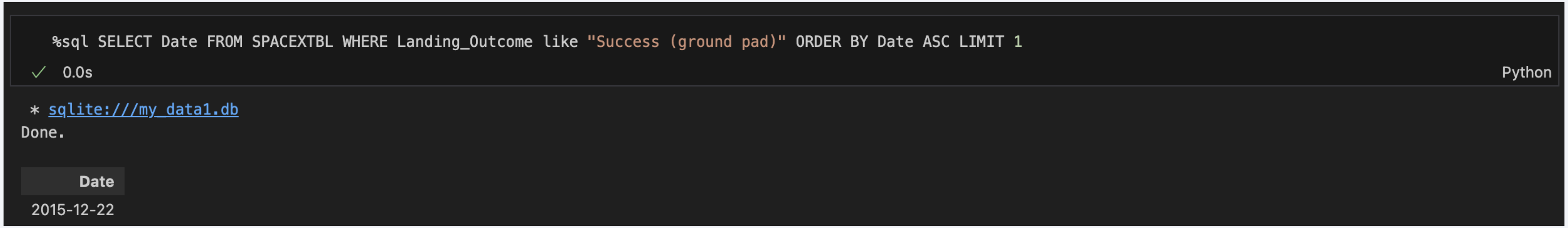
```
%sql SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) as "AVG PAYLOAD MASS KG" FROM SPACEXTBL WHERE Booster_Version like "F9 v1.1"
```

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

| Booster_Version | AVG PAYLOAD MASS KG |
|---|---|
| F9 v1.1 | 2928.4 |

- Displaying average payload mass in Kg carried by booster version "F9 V1.1" by performing the average of all payload masses carried by this booster version.

# First Successful Ground Landing Date



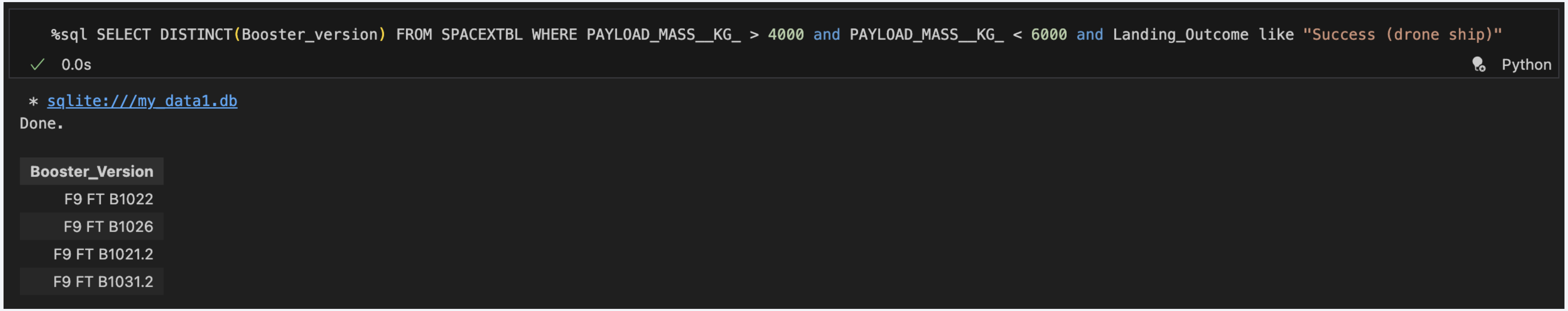
```
%sql SELECT Date FROM SPACEXTBL WHERE Landing_Outcome like "Success (ground pad)" ORDER BY Date ASC LIMIT 1
```

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

| Date |
|---|
| 2015-12-22 |

- Displaying the date for the first successful landing outcome on ground pad by selecting the lowest date value for this landing outcome.

# Successful Drone Ship Landing with Payload between 4000 and 6000



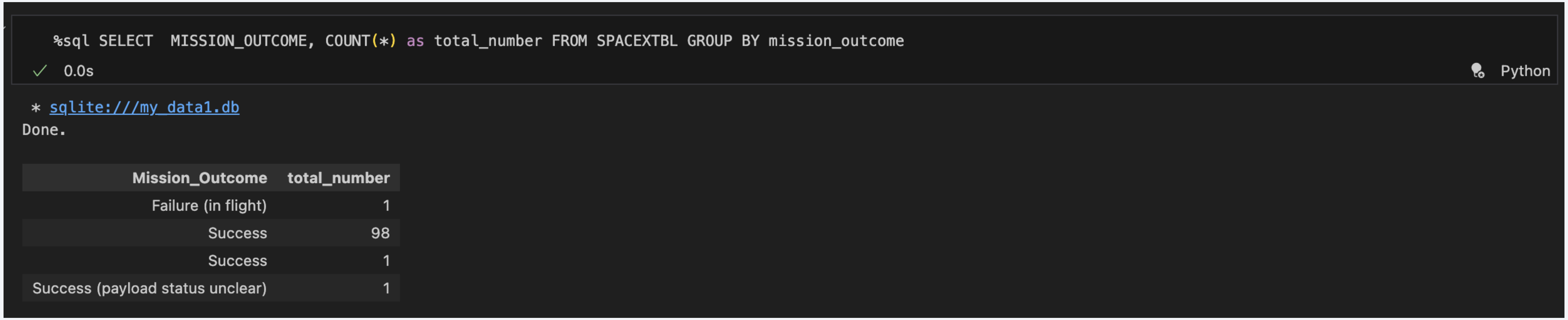
```
%sql SELECT DISTINCT(Booster_version) FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 and Landing_Outcome like "Success (drone ship)"
✓  0.0s                                                                                                                      Python

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

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

- Displaying the booster versions for that have successfully landed on a drone ship with a payload mass between 4000 kg and 6000 kg.

# Total Number of Successful and Failure Mission Outcomes



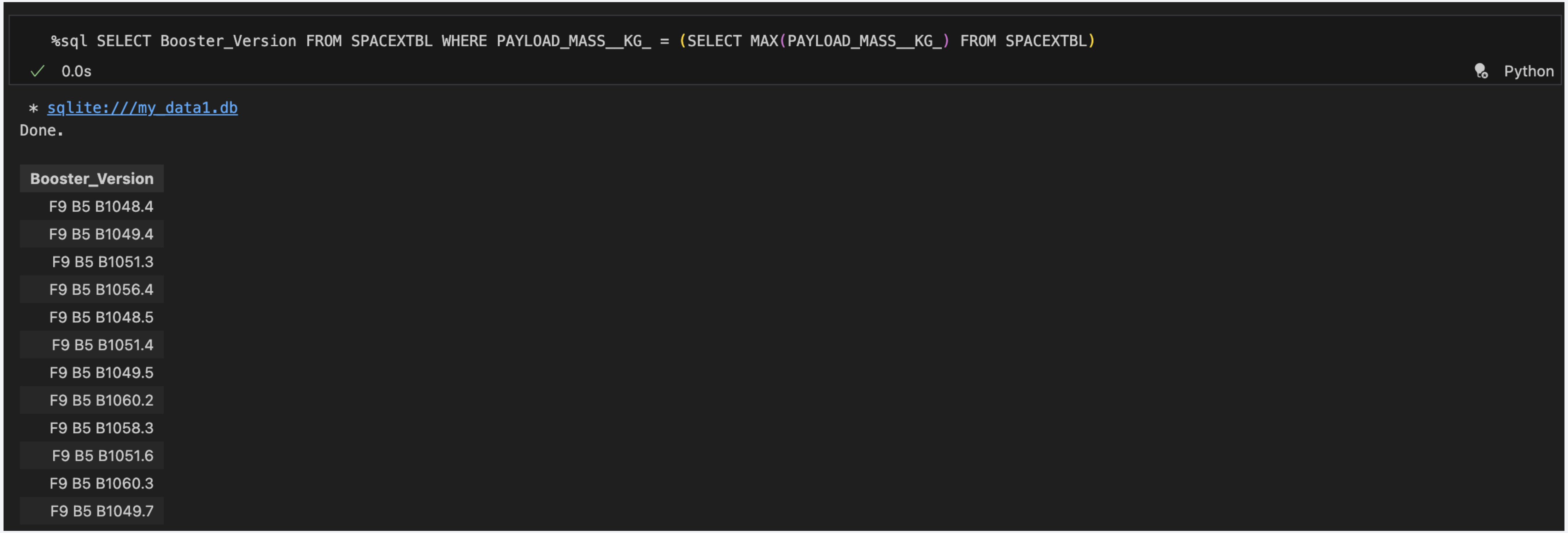
```
%sql SELECT  MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY mission_outcome
```

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

| Mission_Outcome | total_number |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

- Displaying the number of successful and failed missions.
- An error in the formatting of one of the Success values in the mission outcome causes an error in the GROUP BY command which results in the code not grouping this value.

# Boosters Carried Maximum Payload



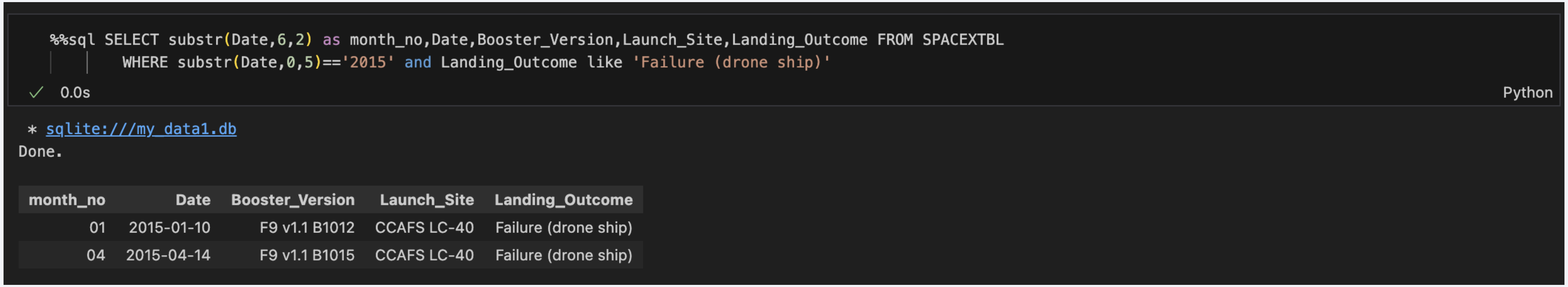
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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
```

- Displaying all booster versions that have carried the maximum payload mass.
- A subquery is used to determine the maximum payload value and is used as a filter value in the SQL query

# 2015 Launch Records



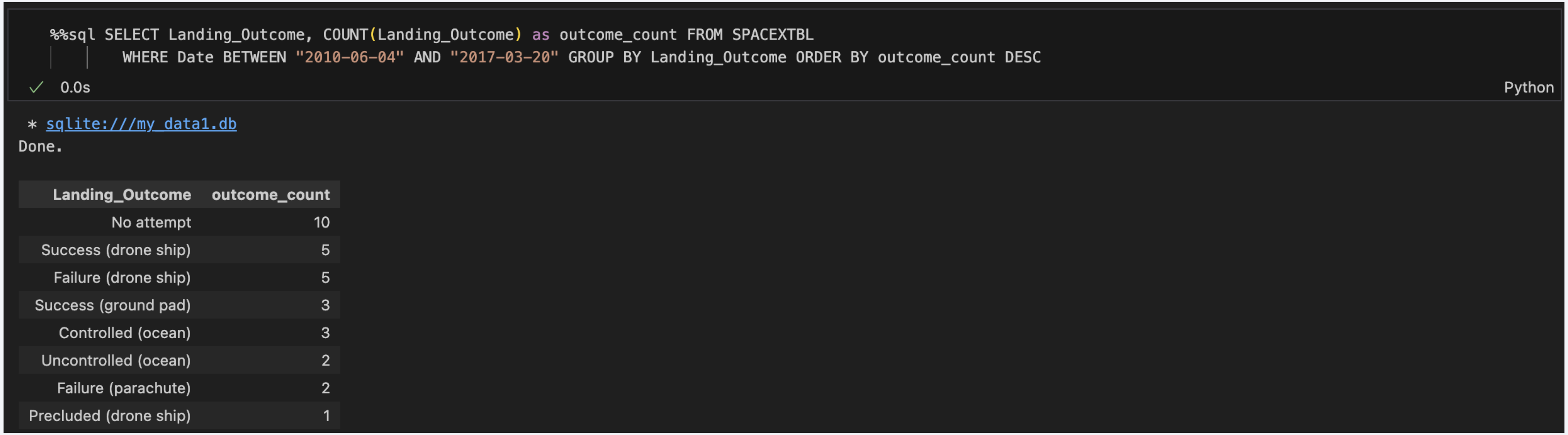
```
%%sql SELECT substr(Date,6,2) as month_no,Date,Booster_Version,Launch_Site,Landing_Outcome FROM SPACEXTBL
        WHERE substr(Date,0,5)=='2015' and Landing_Outcome like 'Failure (drone ship)'
```

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

| month_no | Date | Booster_Version | Launch_Site | Landing_Outcome |
|---|---|---|---|---|
| 01 | 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

- Displaying the month, date, booster version, launch site, and failed landing outcomes for landings on drone ship in 2015
- Only two of this type of failures were identified in 2015, both within 6 months of each other

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



```
%%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) as outcome_count FROM SPACEXTBL
        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 |

- Ranking the counts of landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order.

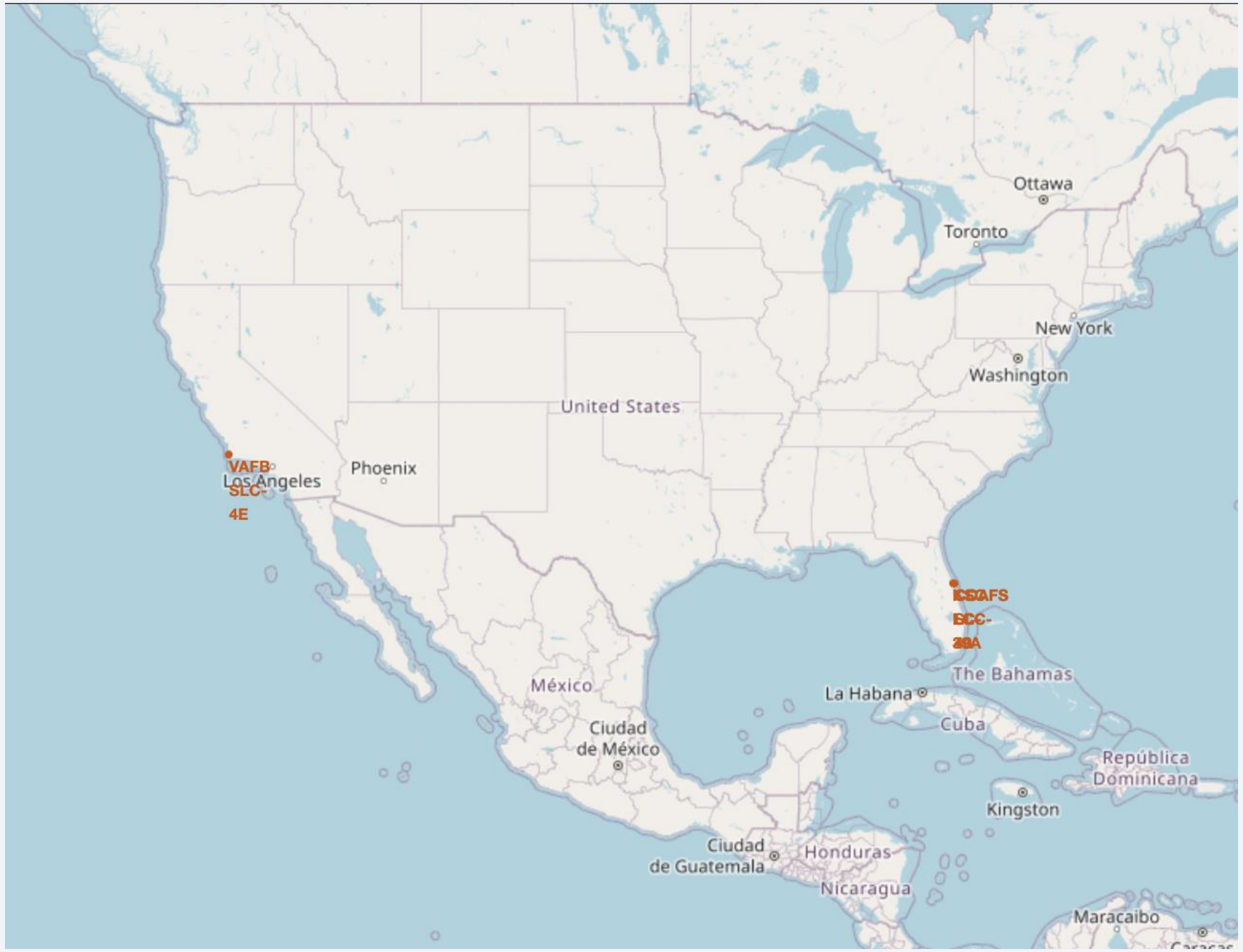
Section 3

# Launch Sites Proximities Analysis

# Map of launch sites

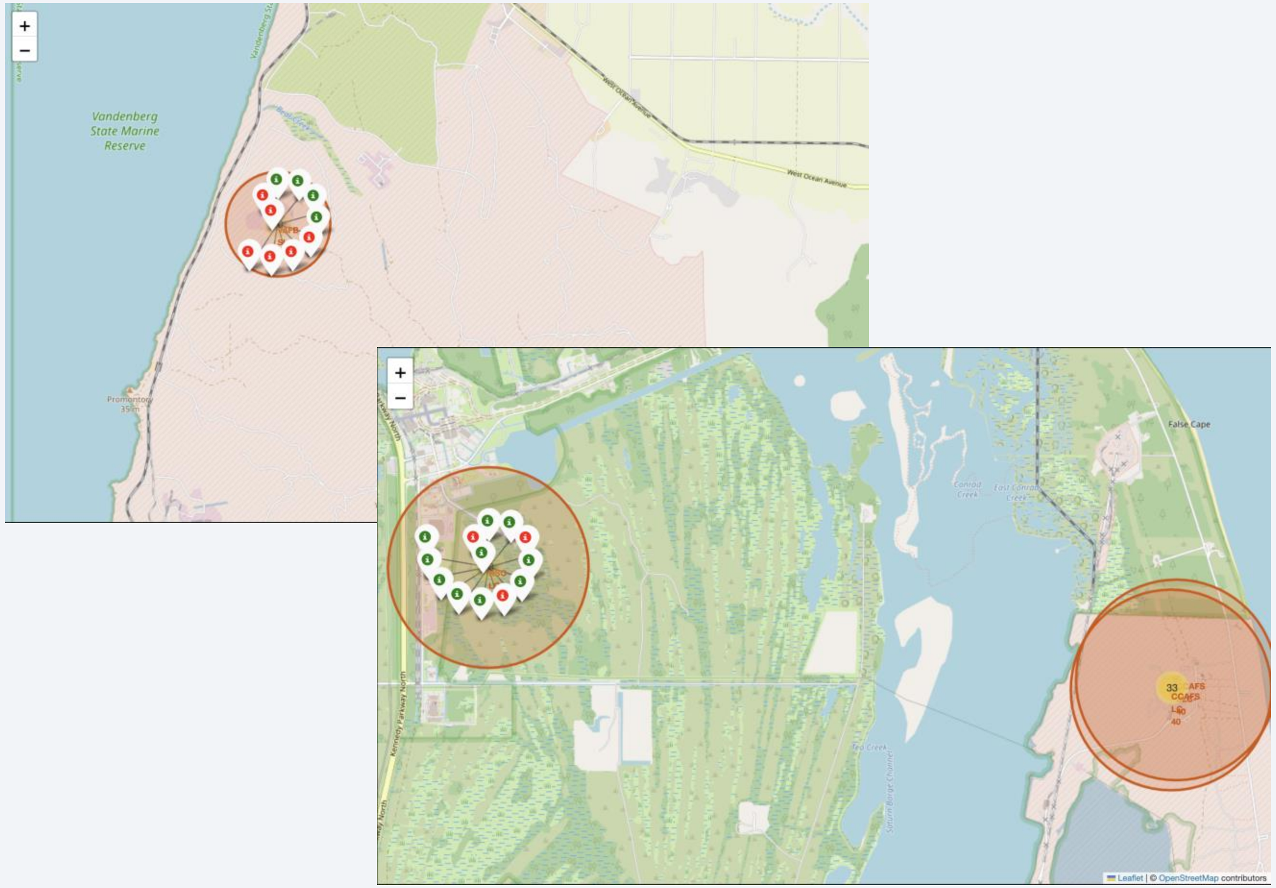
- All launch sites are within 35 degrees of the equator, with CCAFS LC-40 being the closest to the equator at latitude 28.562302 degrees. Launch sites are usually situated at the closest point possible to Equator line, because anything on the surface of the Earth at the equator is already moving at the maximum speed (1670 kilometers per hour), thus reducing the amount of energy required by the rocket to achieve the escape velocity.

- Launch sites are also located in proximity to the coast, with VAFB SLC-4E being the only site in the West coast of the USA. Placing launch sites on the coast is done for safety reasons, so that in case of catastrophic failure, the rocket can be safely dumped back into the sea without risking damaging property with falling debris or hurting people with explosions.



# Color-labeled launch outcomes

- From the colored label markers it is easy to identify which launch sites have a relatively high launch success.
  - Green: Successful launch
  - Red: Failed launch

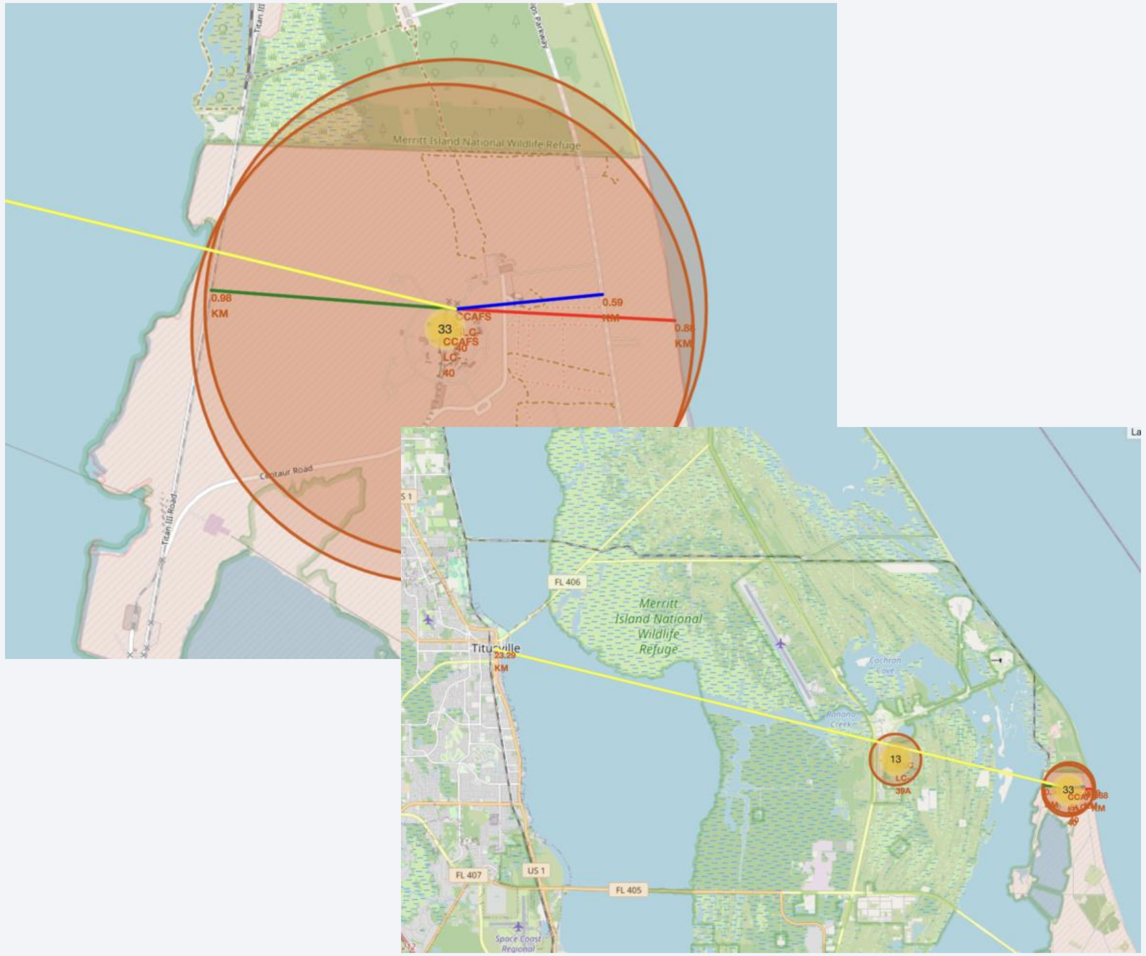


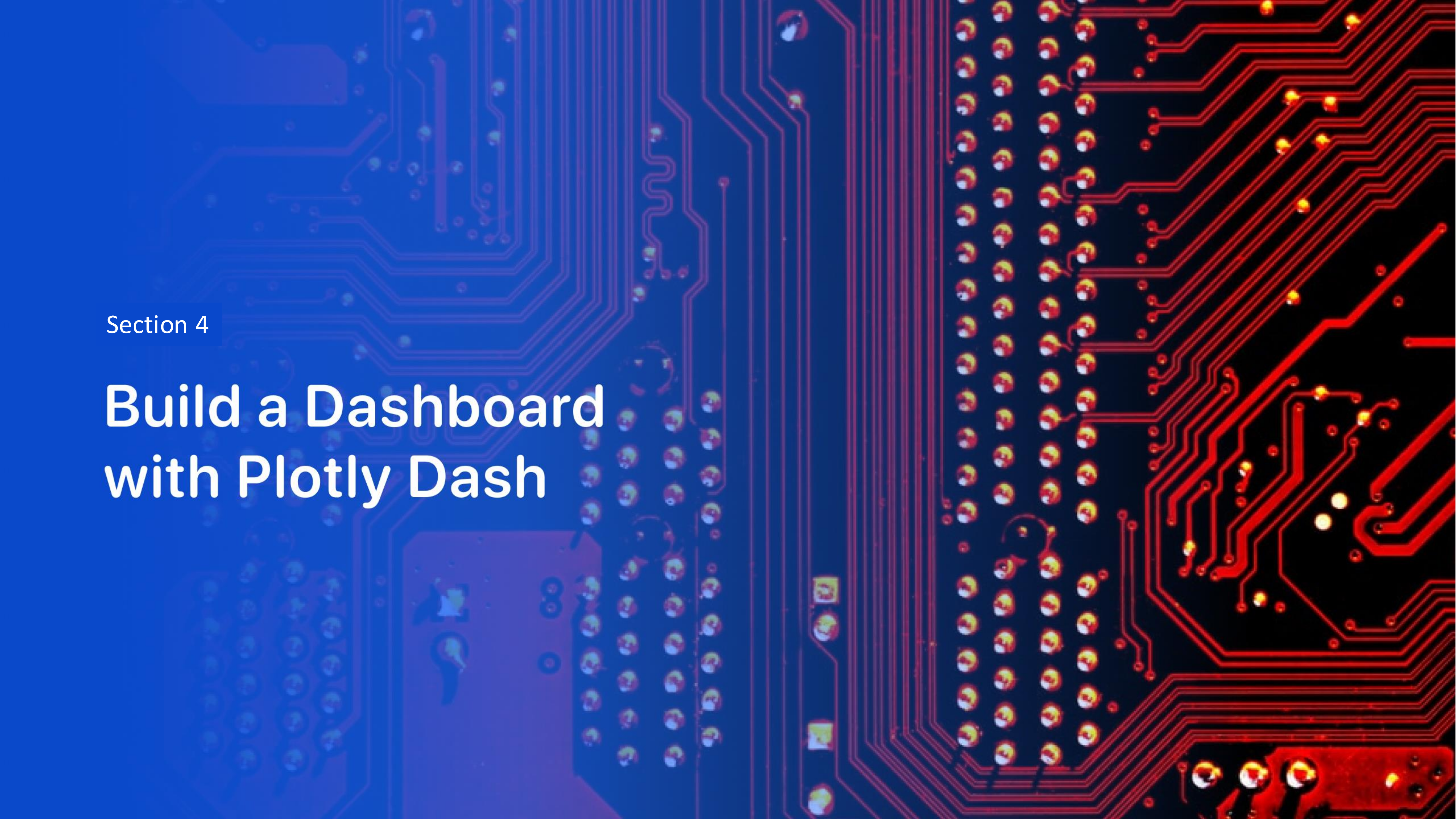
Vandenberg State Marine Reserve
False Cape
33
AFS
CCAFS
40
Leaflet | © OpenStreetMap contributors

# Logistics and Safety for launch site CCAFS SLC 40

A visual analysis of launch site CCAFS SLC 40 reveals that:

- The nearest Railway is found at 0.98km from launch site. This is quite close to the launch site (<1km) as trains can be halted while launches are being performed.
- The nearest Highway was found at 0.59km from launch site. This is quite close to the launch site (<1km) as highway can be closed while launches are being performed.
- The nearest coastline was fund at 0.88km from launch site. This is quite close to the launch site (<1km) as it provides a safety mechanism for rockets to fail safely into the ocean if something goes wrong.
- The closest city was found to be at 23.29km from launch site. This is quite far from the launch site in order to provide safety buffer to local residents. Although it may seem close when considering that a rocket may travel 15-20 km in a few seconds (making the proximity to the city potentially dangerous), it is important to also consider that the launches are performed in the opposite direction to the city, towards the ocean.

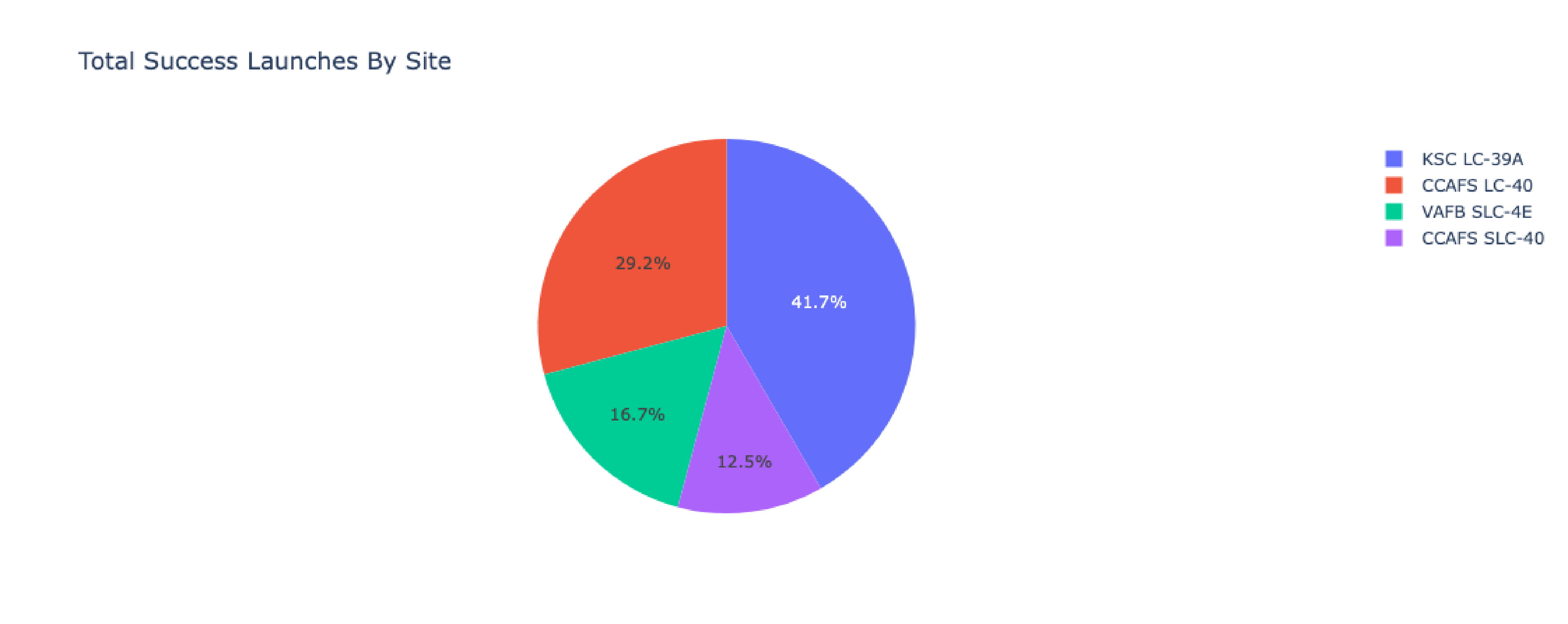




Section 4

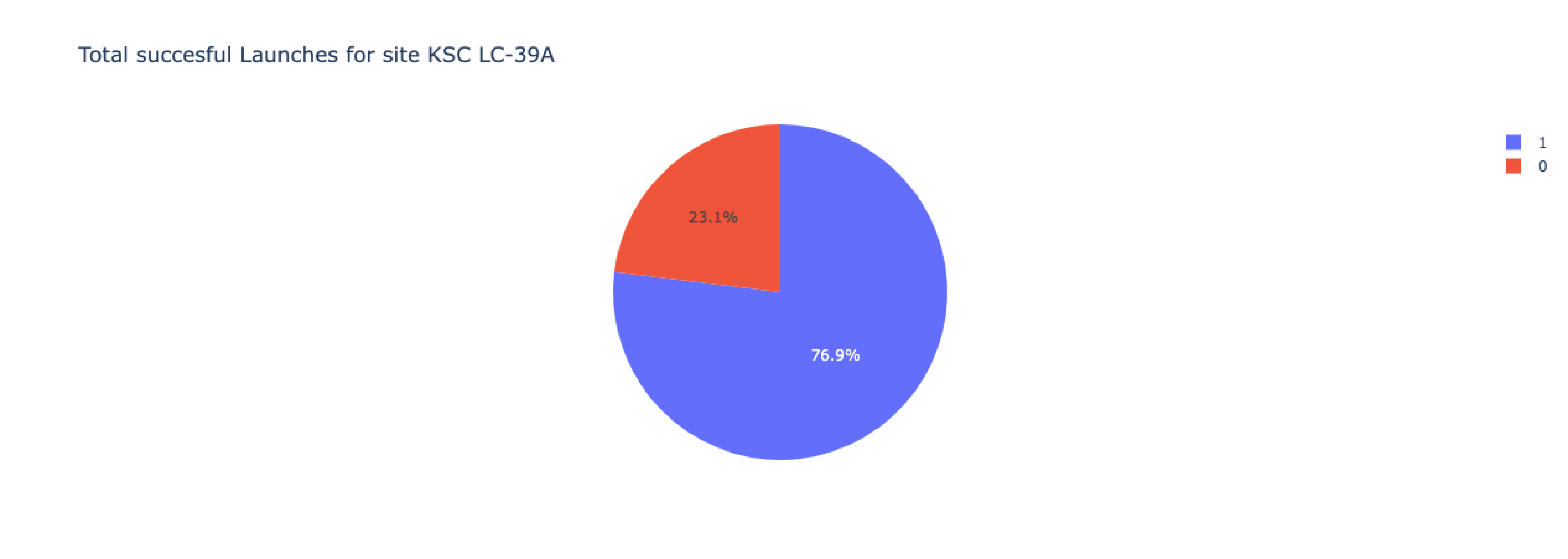
# Build a Dashboard with Plotly Dash

# Total success launches by site



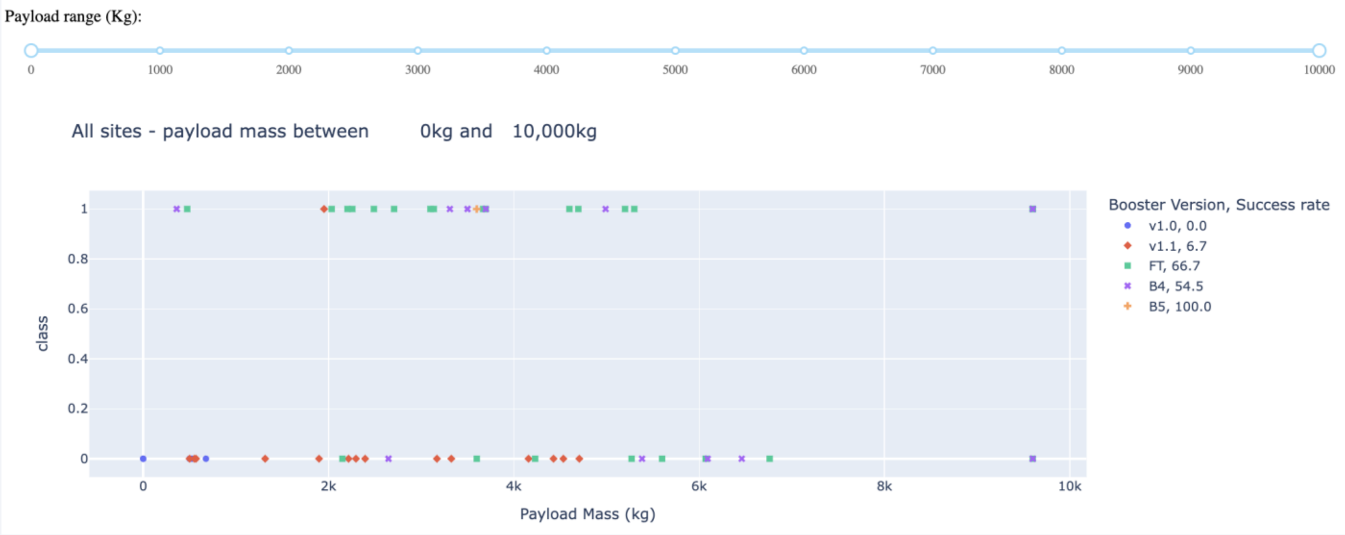
- Chart shows the importance of launch site on launch success
- KSC LC-39A is found to be the most successful launch site overall

# Launch Success ratio at site KSC LC-39A



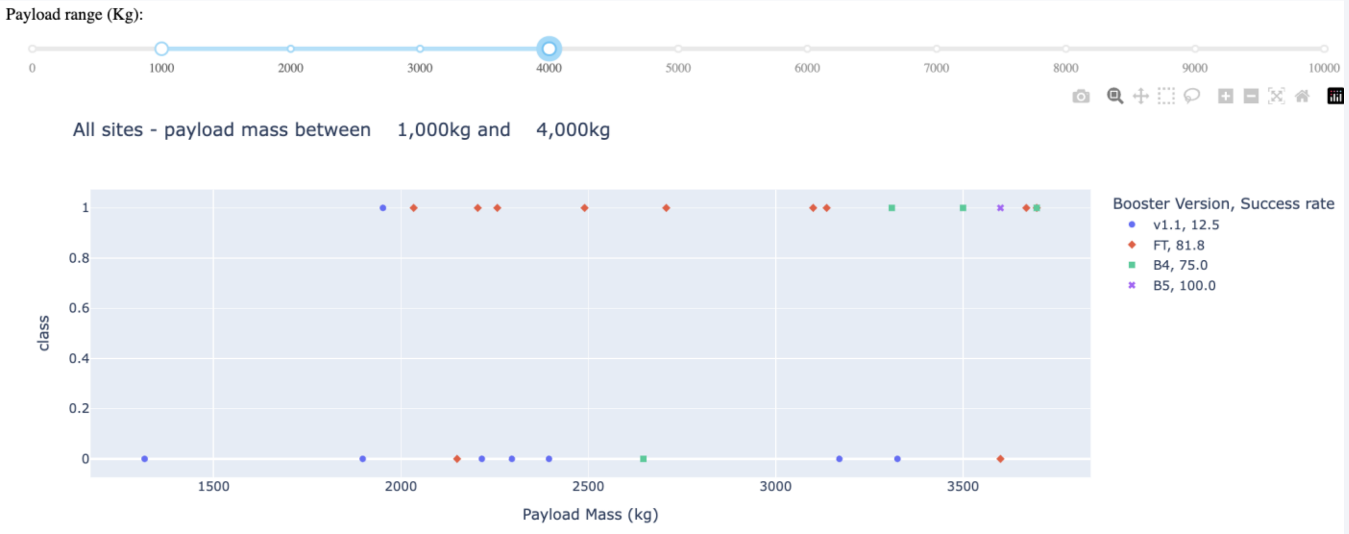
- Pie Chart for launch success rates at site KSC LC-39A shows a 76.9% success rate

# Payload vs launch outcome



Payload range (Kg):

All sites - payload mass between 0kg and 10,000kg



Payload range (Kg):

All sites - payload mass between 1,000kg and 4,000kg

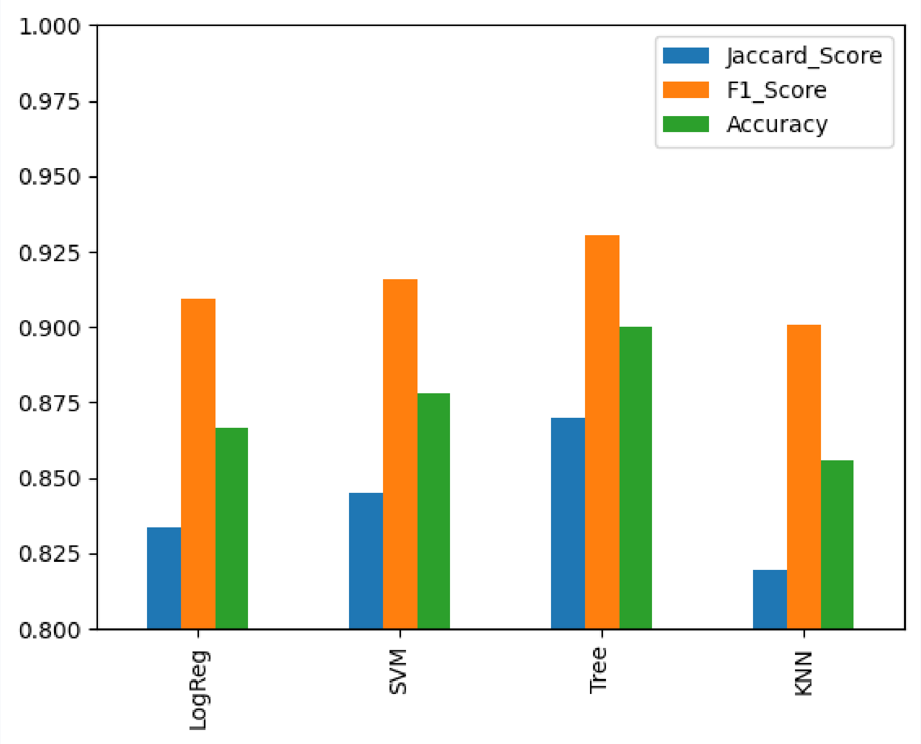
- The chart shows the outcome against the payload for each booster version
- The highest overall success rates across all booster versions is found to be for payload mass ranging between 1000 kg and 4000 kg

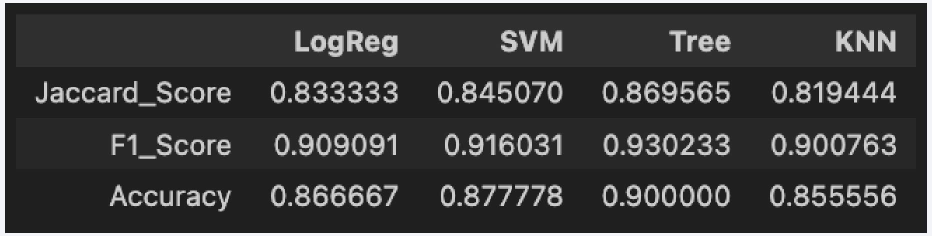
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Model with highest classification accuracy is found to be the Decision Tree Classifier
  - Scores on all data:
    - Accuracy = 90.0%
    - F1 Score = 93.0%
    - Jaccard Score = 86.9%

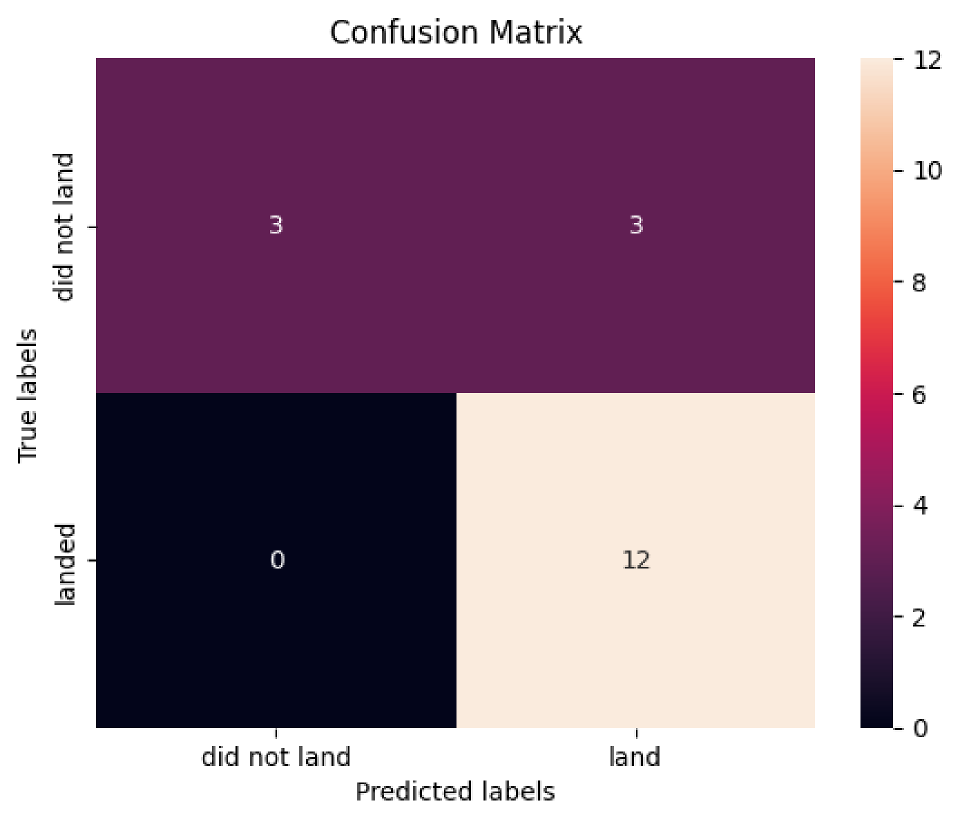




| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.833333 | 0.845070 | 0.869565 | 0.819444 |
| F1_Score | 0.909091 | 0.916031 | 0.930233 | 0.900763 |
| Accuracy | 0.866667 | 0.877778 | 0.900000 | 0.855556 |

# Confusion Matrix

- Confusion matrix shows that the Decision Tree Classifier model:
  - possess very strong true positive predictive powers (predicting correctly all 12 true landed labels)
  - Only has 50% accuracy at predicting true negative. It has an issue with predicting false positives.

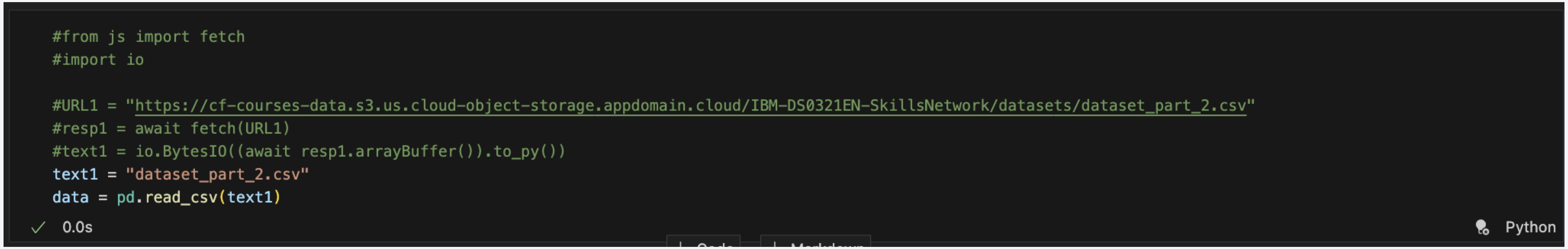


# Conclusions

- Data sources from SpaceX and public data have been collected and analysed
- To compete with SpaceX, the following conclusions has been derived:
  - Highest success launches are from launch site KSC LC-39A.
  - Launch sites will be located near coastal lines and away from cities.
  - Launches with Payload mass between 1000 kg and 4000 kg are more sucessful over all booster version.
  - Success rate of landings increases over time, and over rocket evolutions.
  - Orbits ES-L1, GEO, HEO, and SSO have 100% success rate.
  - Decision Tree Classifier is the model that shows the best performance at predicting launches outcomes.

# Appendix

- Some code snippets have been edited in order for the code to run on a local machine. In these cases, the URL was commented out and the csv was downloaded with wget.
  The csv was subsequently loaded from local memory using the same pandas read_csv command.



```
#from js import fetch
#import io

#URL1 = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv"
#resp1 = await fetch(URL1)
#text1 = io.BytesIO((await resp1.arrayBuffer()).to_py())
text1 = "dataset_part_2.csv"
data = pd.read_csv(text1)
```

- Folium world map may not show on certain web browsers on GitHub. Please try using a different web browser or downloading and running the Jupyter Notebook locally, if required.

## Thank you!