IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kyle Winnaar
01 April 2023

Outline

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

Executive Summary

Summary of methodologies used

- Data was collected from the public SpaceX API, and the SpaceX Wikipedia page
- The data from these various sources were wrangled, and indicator variables (1 – success, 0 – failure) were added to the data.
- The data was explored using interactive dashboards (Folium and Plotly Dash) and database (SQL) queries.
- Machine learning models were trained to predict the outcome of the landings.

- Summary of all results
  - The models produced similar results, with an accuracy of 83.33 %. The model tended to overpredict successes, leading to False Positives.

Introduction

Background and Context

1. The company “SpaceY” would like to compete with SpaceX for contracts for commercial space travel.
2. The bulk of the cost comes from the first-stage of the rocket.
3. SpaceX has addressed this issue by **reusing** the first-stage rocket.
4. How much would a single launch cost?

What are we interested in predicting?

1. Can we estimate the cost of a launch using previous SpaceX data?
2. How likely is the first-stage of the rocket to be recovered?
3. Can we predict mission successes/failures based on some input parameters?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and scraped from the Wikipedia website.
- Perform data wrangling
  - Landing outcomes were categorized as either success (1), or failure (0) using one-hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various classification models were built, tuned, and evaluated.

Data Collection

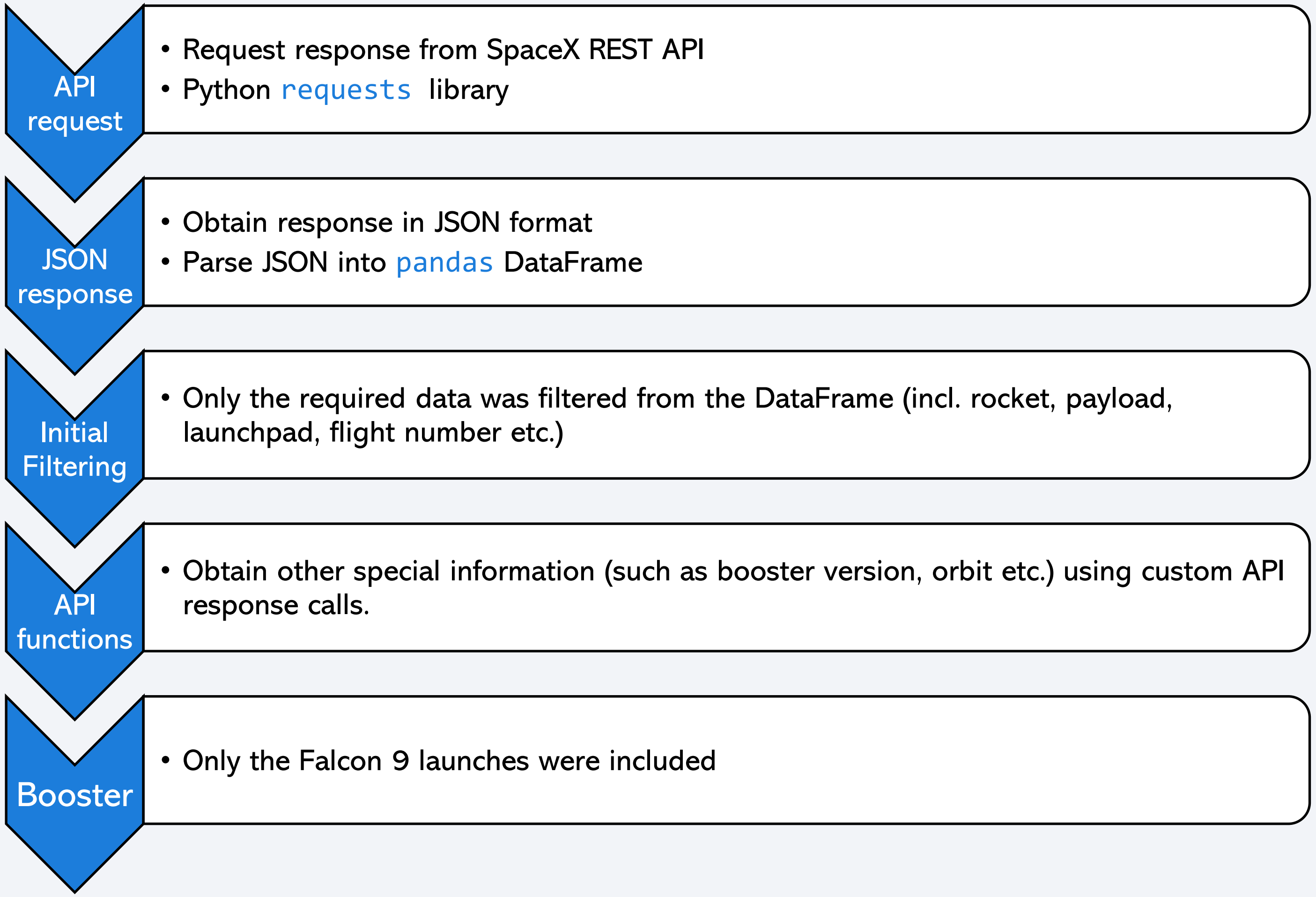
Data collection using the SpaceX REST API

- Data was collected using the `requests` library in Python. The request returned JSON data containing information such as booster version, payload mass etc. The data was normalized before saving as .CSV file for later use.

Data collection using web scraping

- Data was collected by scraping from the SpaceX Wikipedia page using the `beautifulsoup` library in Python. The response was the raw HTML in string format. The HTML was then further parsed to obtain the data within the tables to obtain data such as launch site, orbit reached, launch outcome and booster landing outcome The flight and rocket identifiers were used to link the tables. The data was normalized before saving as .CSV file for later use.

Data Collection – SpaceX API



API request

- Request response from SpaceX REST API
- Python `requests` library

JSON response

- Obtain response in JSON format
- Parse JSON into `pandas` DataFrame

Initial Filtering

- Only the required data was filtered from the DataFrame (incl. rocket, payload, launchpad, flight number etc.)

API functions

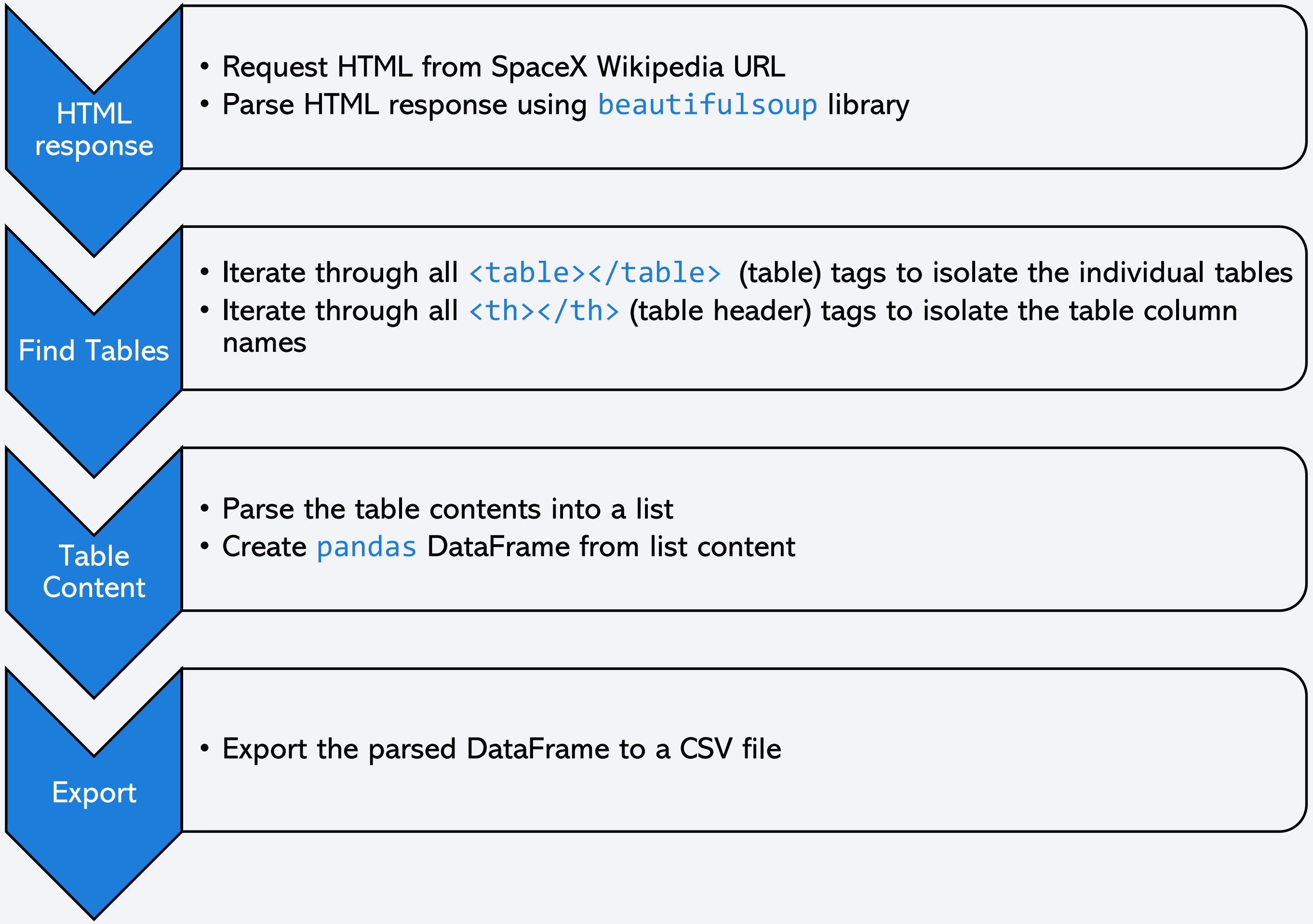
- Obtain other special information (such as booster version, orbit etc.) using custom API response calls.

Booster

- Only the Falcon 9 launches were included

See Jupyter Notebook (SpaceX API)

Data Collection – Web Scraping



HTML response

- Request HTML from SpaceX Wikipedia URL
- Parse HTML response using `beautifulsoup` library

Find Tables

- Iterate through all `<table></table>` (table) tags to isolate the individual tables
- Iterate through all `<th></th>` (table header) tags to isolate the table column names

Table Content

- Parse the table contents into a list
- Create `pandas` DataFrame from list content

Export

- Export the parsed DataFrame to a CSV file

See Jupyter Notebook (web scraping)

Data Wrangling

Mission Outcomes

- The mission outcomes were segmented into a **binary class**, based on the **outcome type**.
  - Any outcome containing "`False`", or "`None`" was classified as a 0 (unsuccessful)
  - Any other outcome was classified as a 1 (success)
  - The binary class (known as one-hot encoding) was applied to the data set, as this would be used as the categorical data classifier for future machine learning models.
- The data set was also checked for null values, to avoid introducing bias into the models.
- The data was grouped by **Launch Site** to count how many launches occurred per launch site.
- The data was grouped by **Orbit Type** to count how many launches occurred per launch site.

See Jupyter Notebook (data wrangling)

EDA with Data Visualization

Data visualization:

- Flight Number vs Launch Site (categorized by first-stage rocket recovery outcome):
  - Determine at which launch sites most failures occurred
  - Determine most used launch site, and obtain timeline of launch sites used in chronological order
- Payload Mass vs Launch Site (categorized by first-stage rocket recovery outcome):
  - Determine range in which most successes/failures occurred
  - Determine launch sites and mission outcomes at which larger payloads were observed
  - Determine minimum and maximum payload mass ranges for each launch site
- Orbit Type vs Rate of Success (success of first-stage rocket recovery):
  - Determine orbit(s) at which the rate of success (of first-stage rocket recovery) were the smallest/largest
- Flight Number vs Orbit Type (categorized by Mission Outcome):
  - Determine the least/most orbits achieved
  - Determine at which orbit types the most failures occurred
- Payload Mass vs Orbit Type (categorized by Mission Outcome):
  - Determine minimum and maximum payload mass ranges for each orbit type
  - Determine any relationship between payload mass, orbit type, and the rate of success of the mission outcome
- Year vs Rate of Success:
  - Determine the overall improvements to the reliability of the recovery of the first-stage rocket

See Jupyter Notebook (EDA with data visualisation)

EDA with SQL

SQL Queries performed:

- List the **unique** launch sites
- List **5 records** where launch sites **begin with** the string **“CCA”**
- List the **total payload mass** carried by boosters launched by **NASA (CRS)**
- List **average payload mass** carried by booster version **“F9 v1.1”**
- List the **date** when the **first successful landing outcome** in ground pad was achieved
- List the names of the boosters which have success in drone ship **and** have payload mass in the range (4000; 6000)
- List the **total number** of successful and failure mission outcomes
- Using a **subquery**, list the names of the booster versions which have carried the **maximum payload mass**
- List the failed landing outcomes in drone ship, their booster versions, and launch site names **for the year 2015**
- **Rank the count of landing outcomes** between 4 October 2010 and 20 March 2017, in **descending** order¶

- See Jupyter Notebook (EDA with SQL queries)

Build an Interactive Map with Folium

- Folium maps were used to visualize the following:
  - The total number of launches
  - The location (geographically) of the launch sites
  - The number of successful and failed launches, per launch site respectively
  - The proximity of the launch sites to major infrastructure
- `Markers` objects were placed at the launch site locations using the latitude and longitude.
- `MarkerCluster` objects were used to indicate the number of successful and unsuccessful
- `Polyline` objects were used to show proximity of the launch sites to major infrastructure (such as coastlines, railways, highways, city centers etc.)

- See Jupyter Notebook (Folium dataviz)

Build a Dashboard with Plotly Dash

- Dashboard plots:
  - A scatter plot of payload mass vs mission outcome (grouped by booster version) to determine whether payload mass and booster version had any effect on the mission outcome
  - A pie chart of the number of total successful mission outcomes to quickly determine which launch site was the most utilized
  - A pie chart of the number of successful and unsuccessful mission outcomes for a given launch site to determine which launch sites were most effective at producing mission success
- Interactions:
  - Dropdown menu to filter by launch site
  - Range slider to filter by payload mass to within a specific range

See Python script (Plotly Dash app)

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

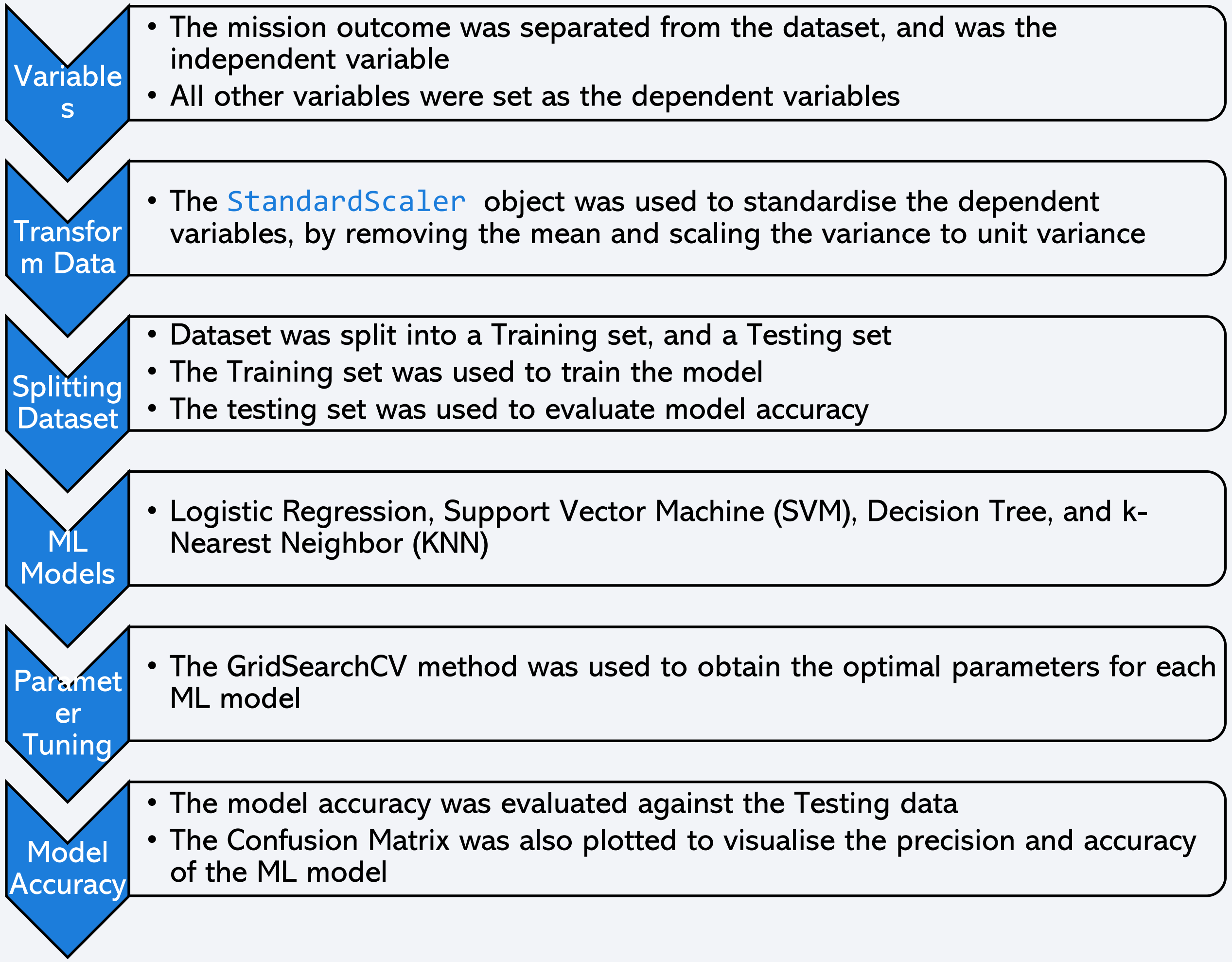
Data sets

- The data set was split into a training set and a testing set
- The models were trained on the training set
- The model accuracy was evaluated on the testing set

Classification models

- Four models were chosen:
  1. Logistic Regression
  2. Support Vector Machine (SVM)
  3. Decision Tree
  4. k-Nearest Neighbor (KNN)
- A Grid Search Method was used to tune the hyper-parameters, and find the optimal parameters for each model

See Jupyter Notebook (Machine Learning)



Variables

- The mission outcome was separated from the dataset, and was the independent variable
- All other variables were set as the dependent variables

Transform Data

- The `StandardScaler` object was used to standardise the dependent variables, by removing the mean and scaling the variance to unit variance

Splitting Dataset

- Dataset was split into a Training set, and a Testing set
- The Training set was used to train the model
- The testing set was used to evaluate model accuracy

ML Models

- Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbor (KNN)

Parameter Tuning

- The GridSearchCV method was used to obtain the optimal parameters for each ML model

Model Accuracy

- The model accuracy was evaluated against the Testing data
- The Confusion Matrix was also plotted to visualise the precision and accuracy of the ML model

Results

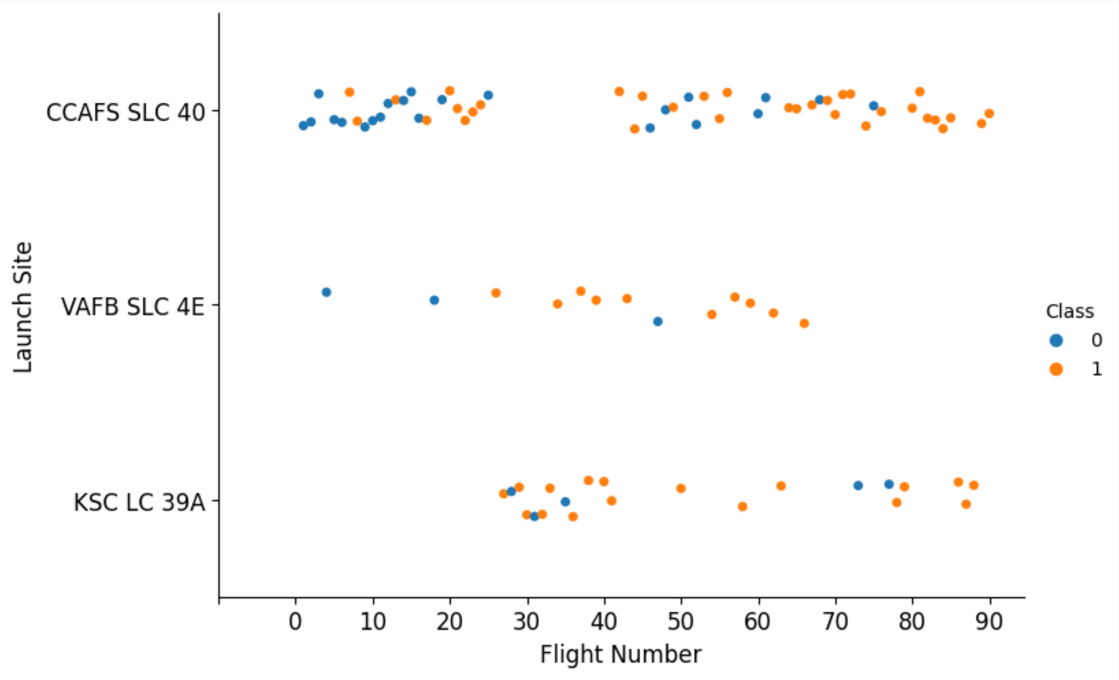
- The following results will be shown as follows:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs Launch Site

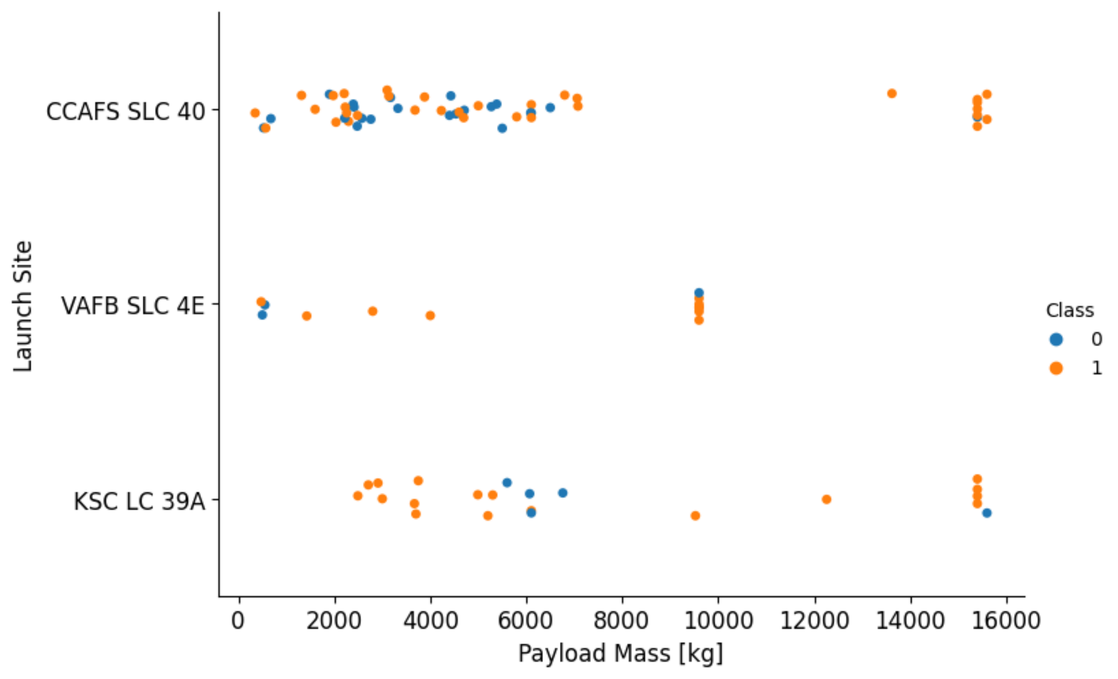
Flight Number vs Launch Site



1. Initially, most tests were performed at **CCAFS SLC 40**.
    a. Many early failures were experienced before some successes were observed.
    b. CCAFS SLC 40 was used the most frequently for launches.
2. Mostly success were observed in the later test flights, across all launch sites.
3. At the end, only CCAFS SLC 40 and KSC LC 39A launch sites were used.
    a. Is this due to some particular factor, e.g. permissible payload or mission profile restrictions?.
4. The launch which was utilised the least is **VAFB SCL 4E**.
    a. Is this due to mission profile restrictions, or perhaps other test system requirements?
5. The launch site which showed the least frequent failures is **KSC LC 39A**.
    a. Is this due to improvements in system reliability?

Payload vs Launch Site

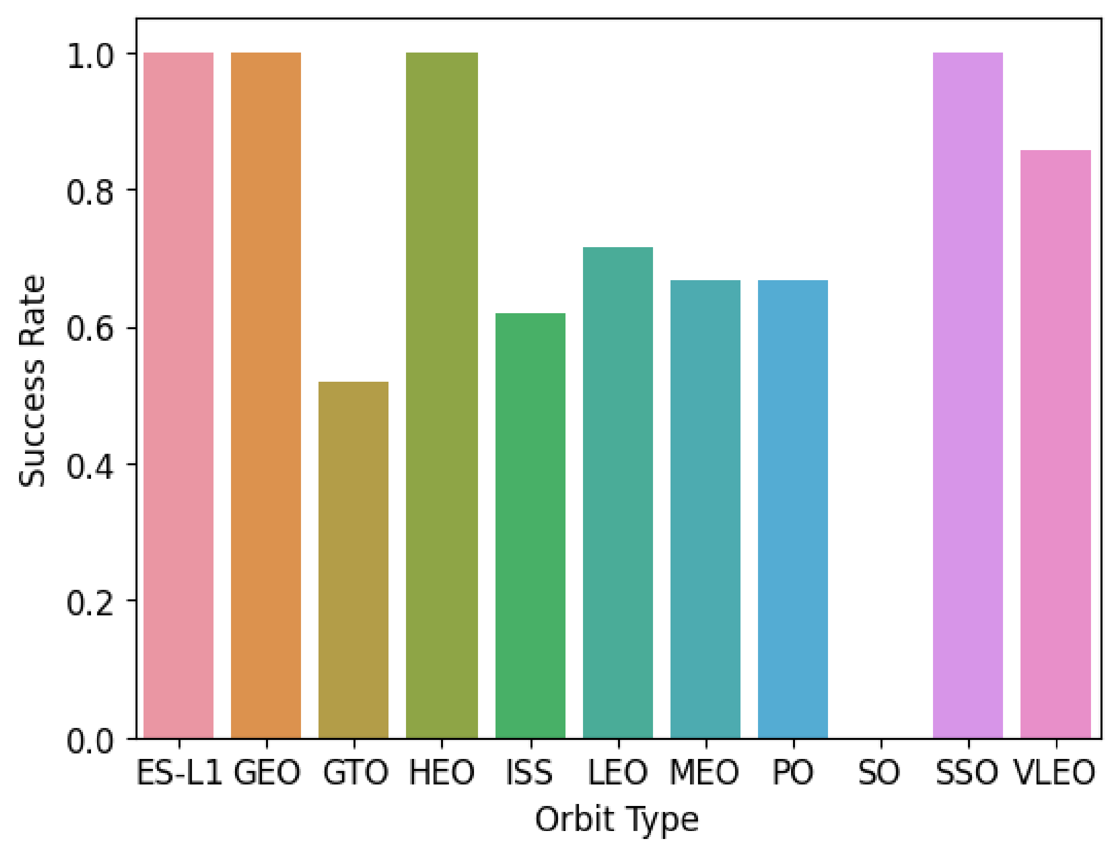
Payload Mass vs Launch Site



1. The majority of launches had lower payload mass (i.e. less than 8000 kg). This is expected as the vehicle is still undergoing developmental testing.
2. Nearly all launches above 14 000 kg at **CCAFS SCL 40** and **KSC LC 39A** launch site experienced a positive mission outcome.

Success Rate vs Orbit Type

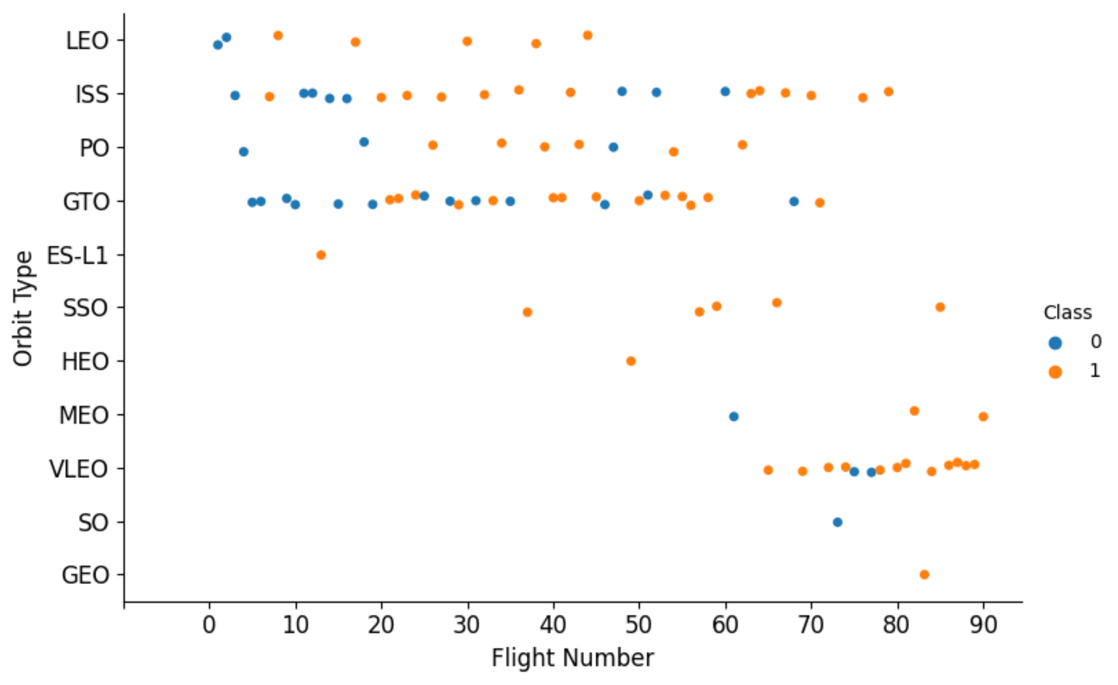
Success Rate vs Orbit Type



1. Orbit types ES-L1, GEO, HEO and SSO produce the largest success rates.
   a. ESL-1, GEO and HEO only had **one** launch, so it cannot be said that the success is due to those locations.
   b. SSO produced **four** successes.
2. Orbit type **GTO** produces the lowest (most accurate) success rate.
   a. Several launches occurred at GEO. Is this due to the orbit type itself?
3. Orbit type **SO** produced the lowest (actual) success rate, as it observed a failure from its single flight.
4. Orbit types **GTO**, **ISS**, **LEO**, **MEO**, and **PO** are probably the most accurate representation of the actual success rate, due to their having the most observed launches.

Flight Number vs Orbit Type

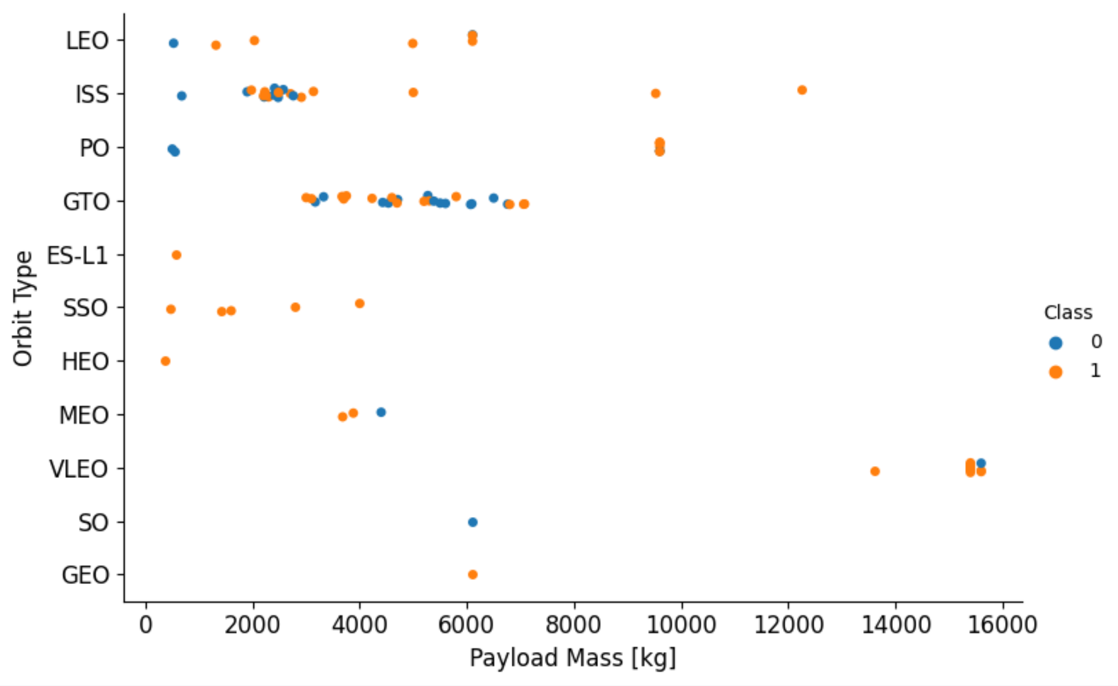
Flight Number vs Orbit Type



1. As expected, mission success occurred more often in later missions than in the beginning.
2. It is also observed that there were more numerous launches in the later stages of the program. This is expected, as mission confidence generally improved over time.

Payload vs Orbit Type

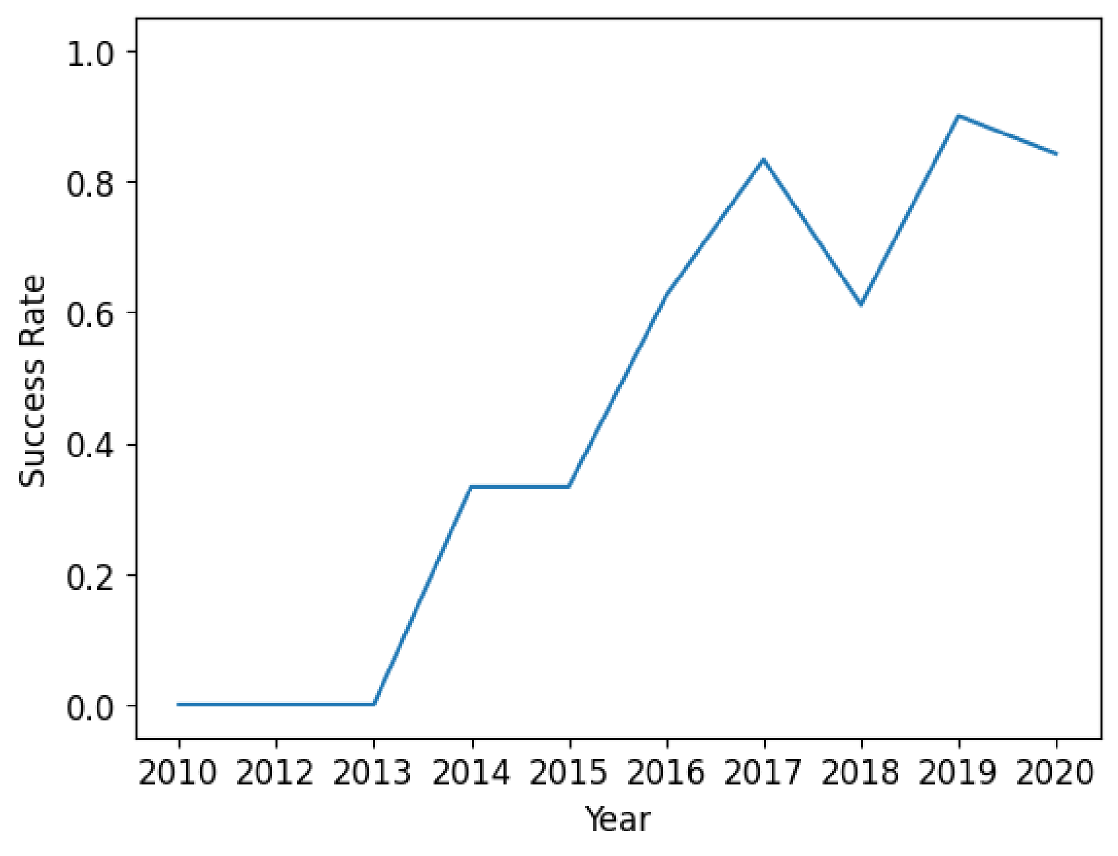
Payload Mass vs Orbit Type



1. There does not appear to be any correlation between payload mass and orbit type. Perhaps this was due to another system-level requirement?

Launch Success Yearly Trend

Launch Success Rate Yearly Trend

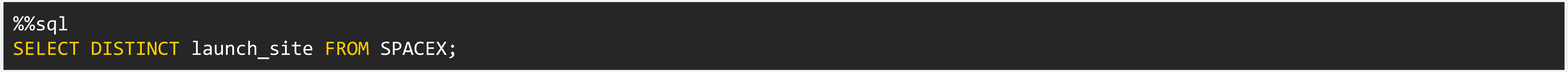


1. Significant improvements in system reliability began in 2013
2. 2018 was the only year in which there was a **decrease** in success rate.
   a. Is this due to the number of launches take that year?
3. Overall success rate reached a maximum of approx. 90% between 2010-2020.

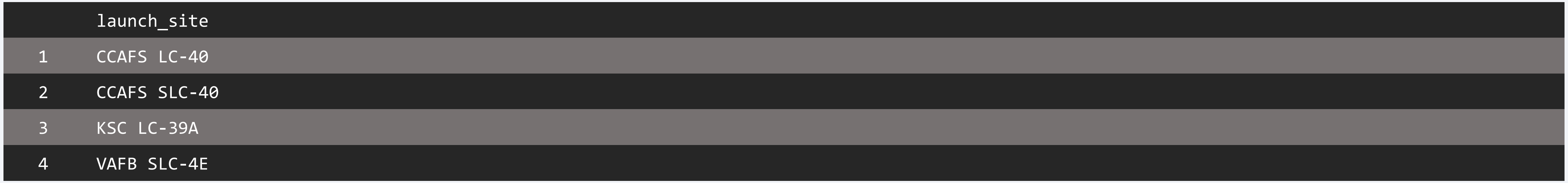
All Launch Site Names

Find the names of the unique launch sites:

- The DISTINCT keyword was used to obtain unique values



```
%%sql
SELECT DISTINCT launch_site FROM SPACEX;
```

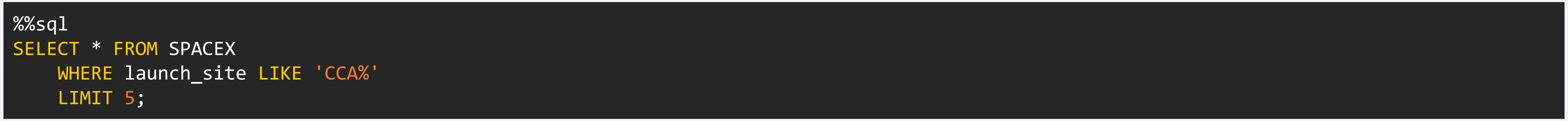


```
        launch_site
   1    CCAFS LC-40
   2    CCAFS SLC-40
   3    KSC LC-39A
   4    VAFB SLC-4E
```

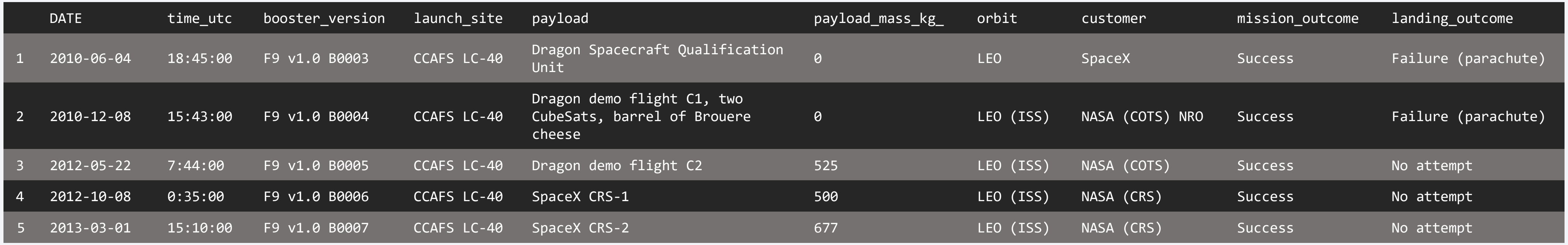
Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`:

- The LIMIT keyword was used to limit the return to 5 records



```
%%sql
SELECT * FROM SPACEX
    WHERE launch_site LIKE 'CCA%'
    LIMIT 5;
```

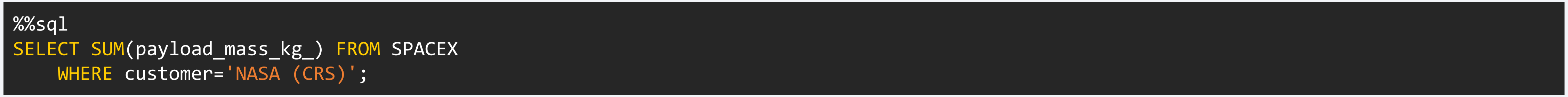


	DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
3	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
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

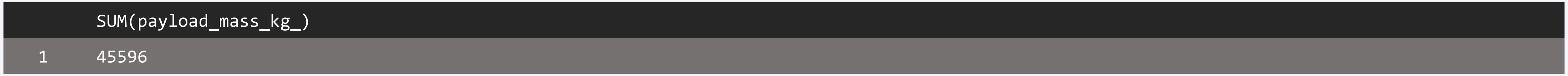
Total Payload Mass

Calculate the total payload carried by boosters from NASA (CRS)

- The SUM keyword was used to sum all payload masses together records
- The search was only limited to the customer ‘NASA (CRS)’



```
%%sql
SELECT SUM(payload_mass_kg_) FROM SPACEX
    WHERE customer='NASA (CRS)';
```

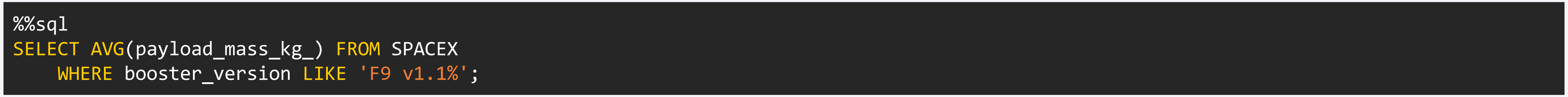


```
        SUM(payload_mass_kg_)
  1     45596
```

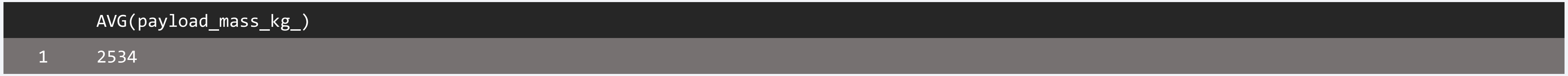
Average Payload Mass by F9 v1.1

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

- The AVG keyword was used to average all payload masses together records
- The search was only limited to booster versions of ‘F9 v1.1%’



```
%%sql
SELECT AVG(payload_mass_kg_) FROM SPACEX
    WHERE booster_version LIKE 'F9 v1.1%';
```

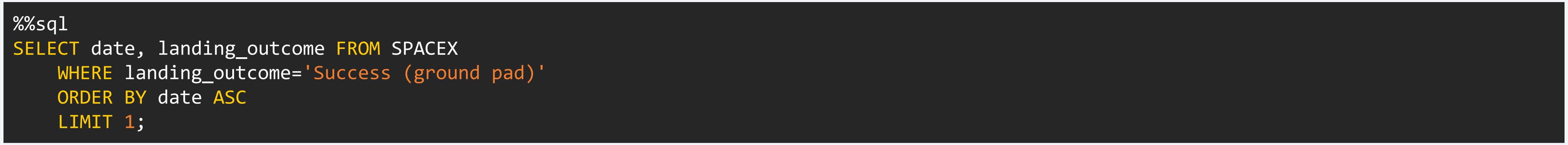


```
        AVG(payload_mass_kg_)
  1     2534
```

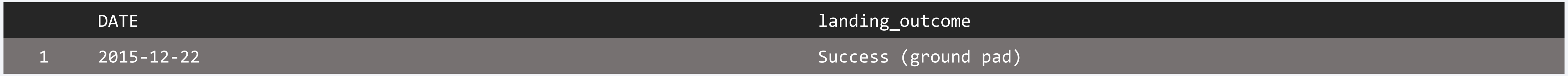
First Successful Ground Landing Date

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

- The ORDER BY keyword was used to all dates, with the ASC keyword used to sort the dates in ascending order
- The search was only limited to landing outcomes of ‘Success (ground pad)’
- The LIMIT keyword was used to find the first record



```
%%sql
SELECT date, landing_outcome FROM SPACEX
    WHERE landing_outcome='Success (ground pad)'
    ORDER BY date ASC
    LIMIT 1;
```

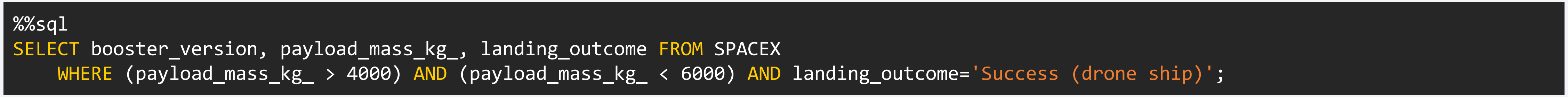


	DATE	landing_outcome
1	2015-12-22	Success (ground pad)

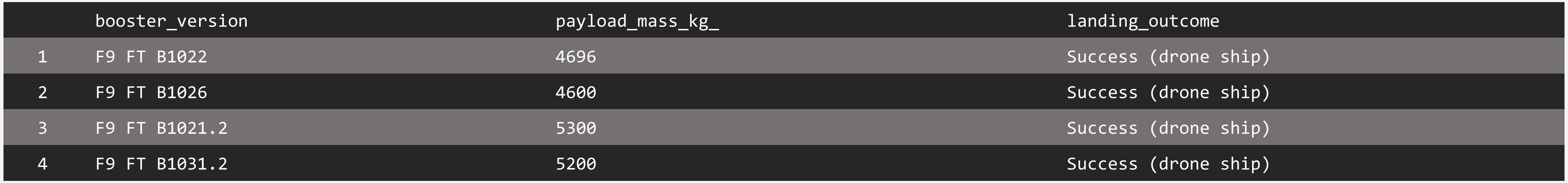
Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The AND keyword was used to chain inequalities together
- Inequalities were used to segment by payload mass and landing outcome



```
%%sql
SELECT booster_version, payload_mass_kg_, landing_outcome FROM SPACEX
    WHERE (payload_mass_kg_ > 4000) AND (payload_mass_kg_ < 6000) AND landing_outcome='Success (drone ship)';
```

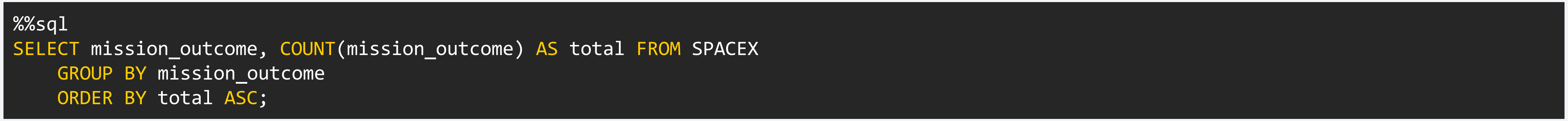


	booster_version	payload_mass_kg_	landing_outcome
1	F9 FT B1022	4696	Success (drone ship)
2	F9 FT B1026	4600	Success (drone ship)
3	F9 FT B1021.2	5300	Success (drone ship)
4	F9 FT B1031.2	5200	Success (drone ship)

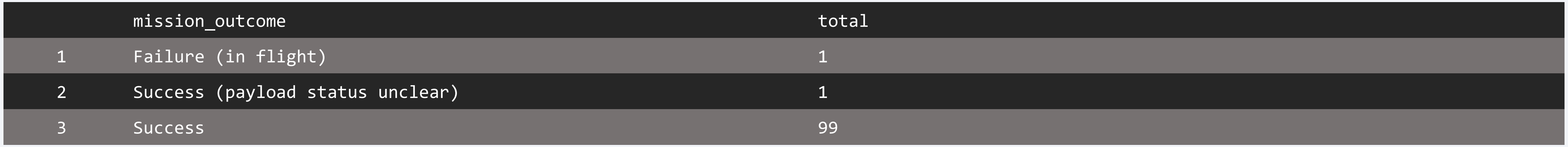
Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

- The COUNT keyword was used to tally the numbers
- The GROUP BY keyword was used to group the mission outcomes into distinct categories
- The AS keyword was used to give the tally count a name



```
%%sql
SELECT mission_outcome, COUNT(mission_outcome) AS total FROM SPACEX
    GROUP BY mission_outcome
    ORDER BY total ASC;
```

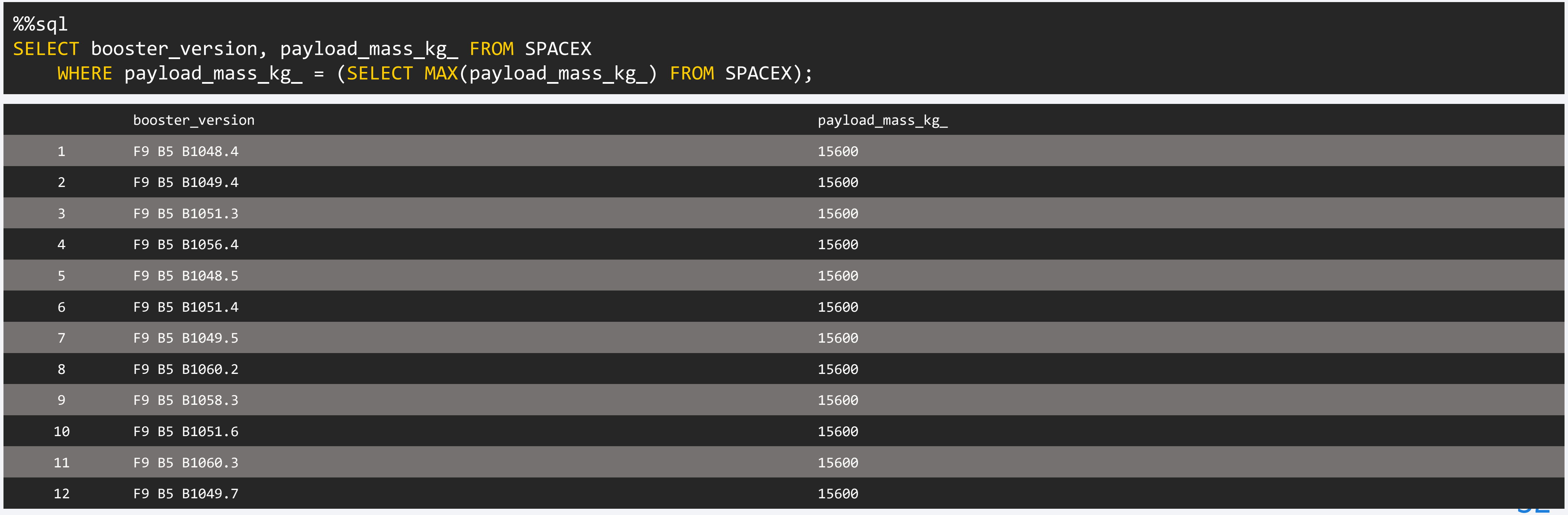


	mission_outcome	total
1	Failure (in flight)	1
2	Success (payload status unclear)	1
3	Success	99

Boosters Carried Maximum Payload

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

- The MAX keyword was used to calculate the maximum payload mass among all records
- A SELECT MAX(payload_mass_kg_) subquery was used to limit the search



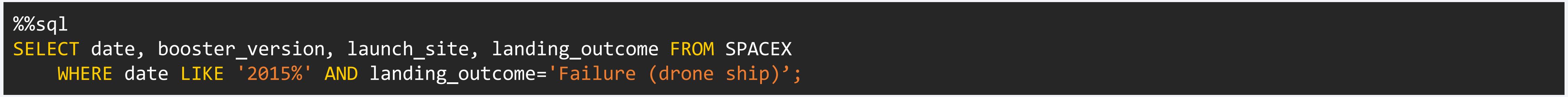
```
%%sql
SELECT booster_version, payload_mass_kg_ FROM SPACEX
    WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM SPACEX);
```

	booster_version	payload_mass_kg_
1	F9 B5 B1048.4	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1051.3	15600
4	F9 B5 B1056.4	15600
5	F9 B5 B1048.5	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1049.5	15600
8	F9 B5 B1060.2	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1051.6	15600
11	F9 B5 B1060.3	15600
12	F9 B5 B1049.7	15600

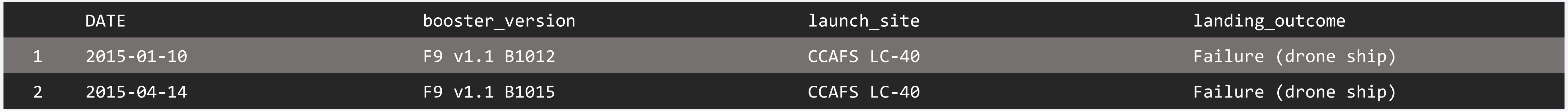
2015 Launch Records

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

- The LIKE keyword was used to create a regular expression to filter out all dates in 2015



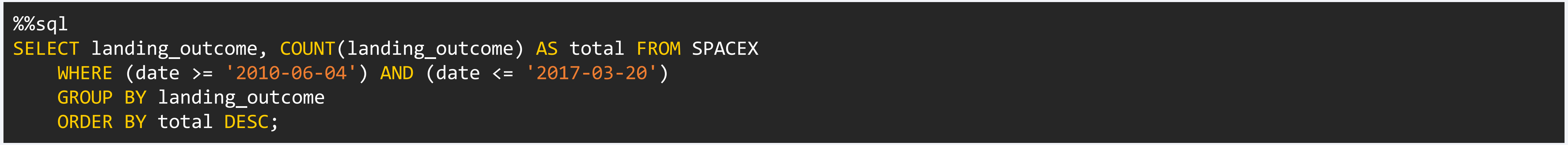
```
%%sql
SELECT date, booster_version, launch_site, landing_outcome FROM SPACEX
    WHERE date LIKE '2015%' AND landing_outcome='Failure (drone ship)’;
```



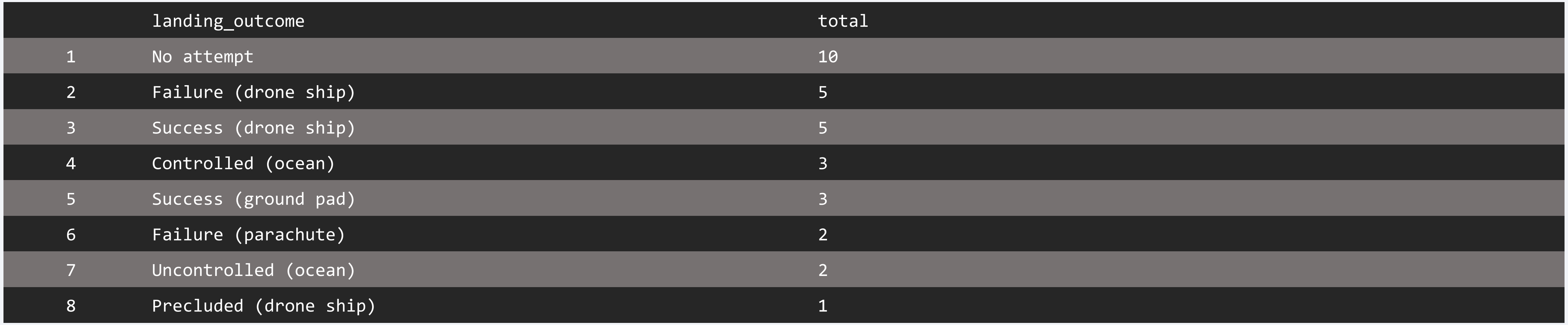
	DATE	booster_version	launch_site	landing_outcome
1	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order



```
%%sql
SELECT landing_outcome, COUNT(landing_outcome) AS total FROM SPACEX
    WHERE (date >= '2010-06-04') AND (date <= '2017-03-20')
    GROUP BY landing_outcome
    ORDER BY total DESC;
```



	landing_outcome	total
1	No attempt	10
2	Failure (drone ship)	5
3	Success (drone ship)	5
4	Controlled (ocean)	3
5	Success (ground pad)	3
6	Failure (parachute)	2
7	Uncontrolled (ocean)	2
8	Precluded (drone ship)	1

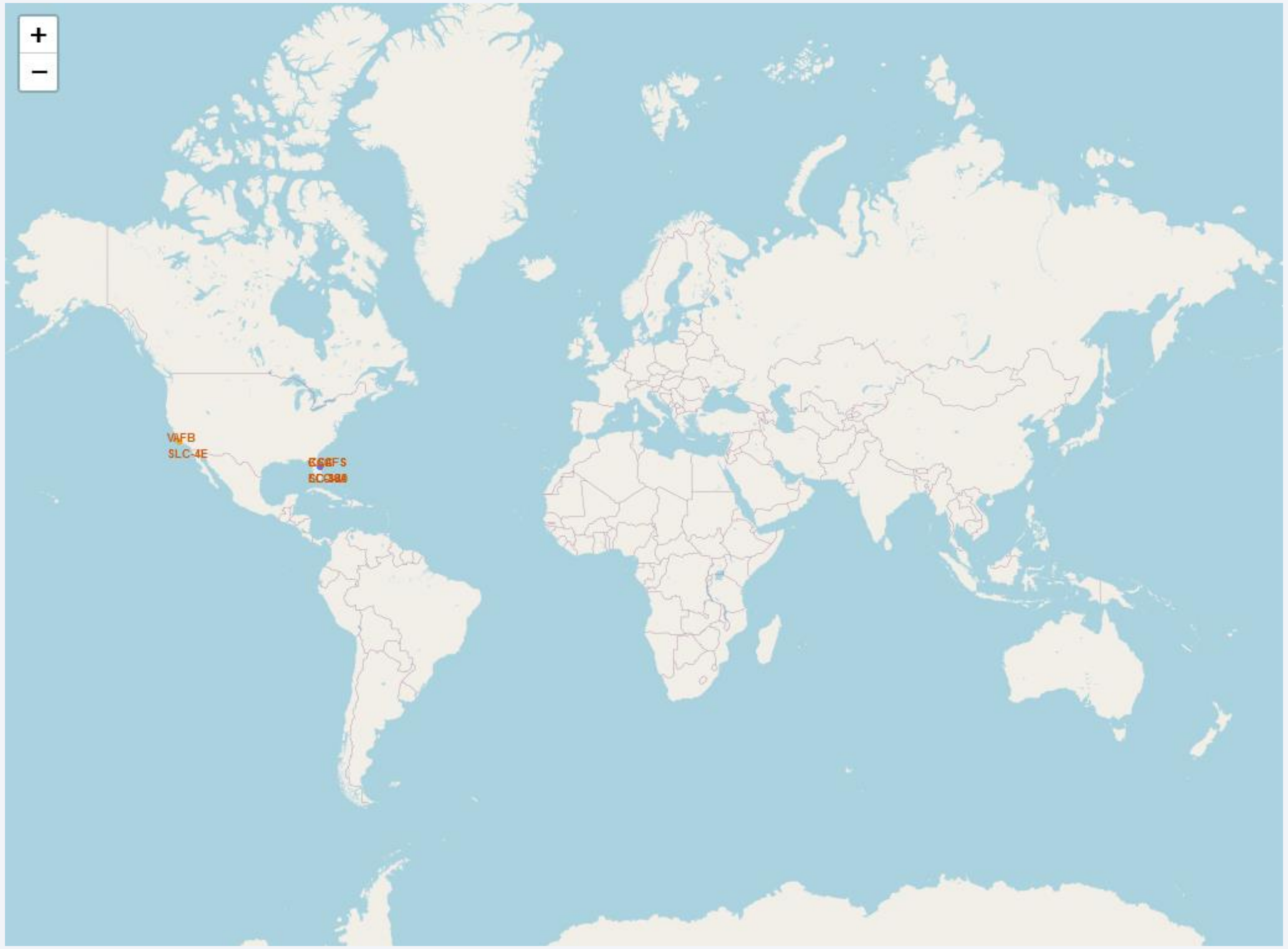
Section 3

Launch Sites Proximities Analysis

Folium Map: Launch Site Locations

Findings:

- All launches sites are located in North America, near the coastlines of the Atlantic and Pacific Ocean respectively.
- **VAFB SLC-4E** is located North of Los Angeles.
- **KSC LC 39-A**, **CCAFS LC-40** and **CCAFS SLC-40** are located at Cape Canaveral.
- The launch sites are located within reasonable distance to road, freight, and rail infrastructure, but are located far enough from the civilian population.



Folium Map: Color-coded Launch Outcomes

All launches at each location are shown in Fig 1.

Fig 2. details the launch outcomes, coded by color (green/success, red/failure)

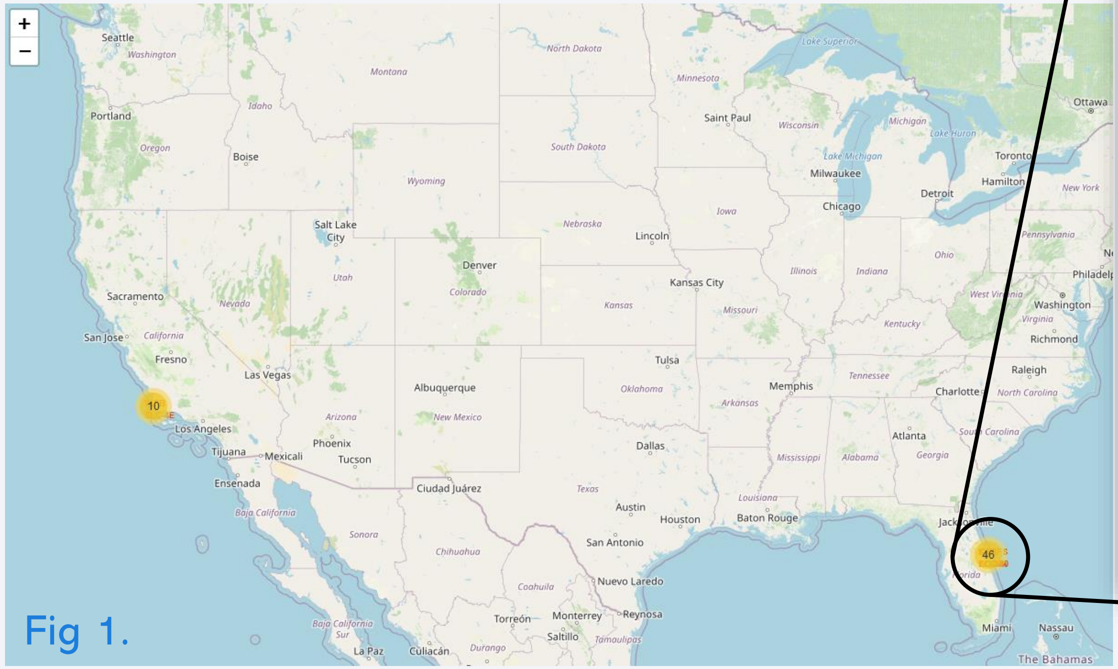


Fig 1.

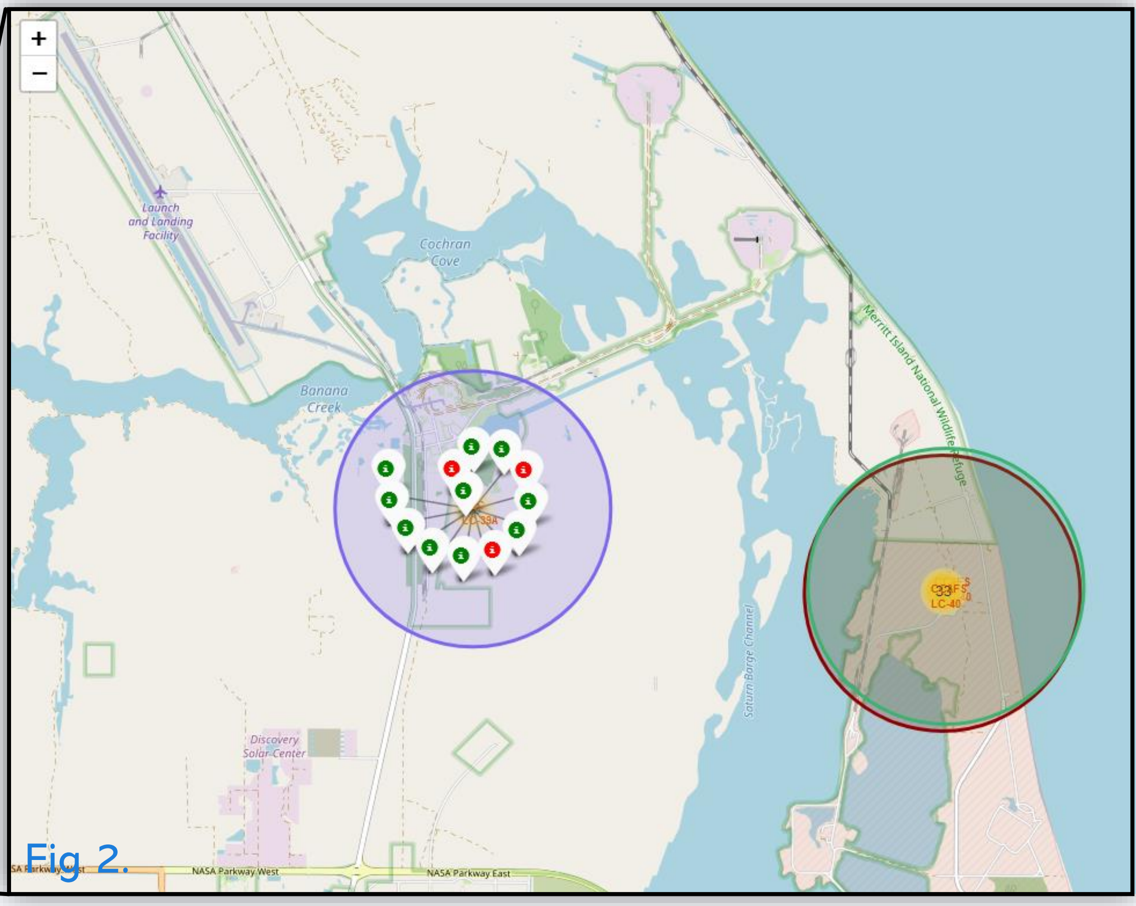


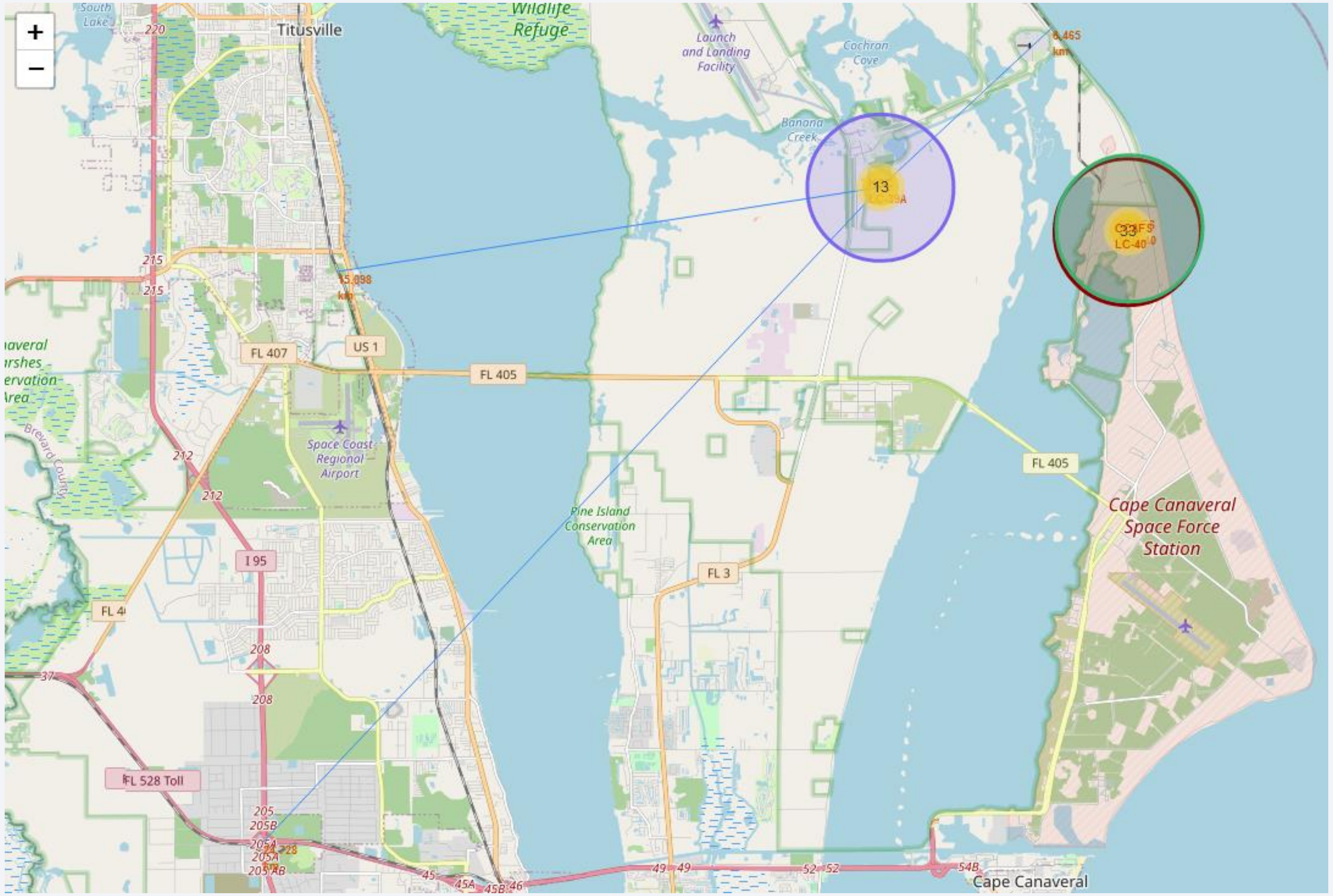
Fig 2.

Folium Map: Proximity to Nearby Infrastructure

Major infrastructure near **KSC LC-39A** is shown in the figure (right).

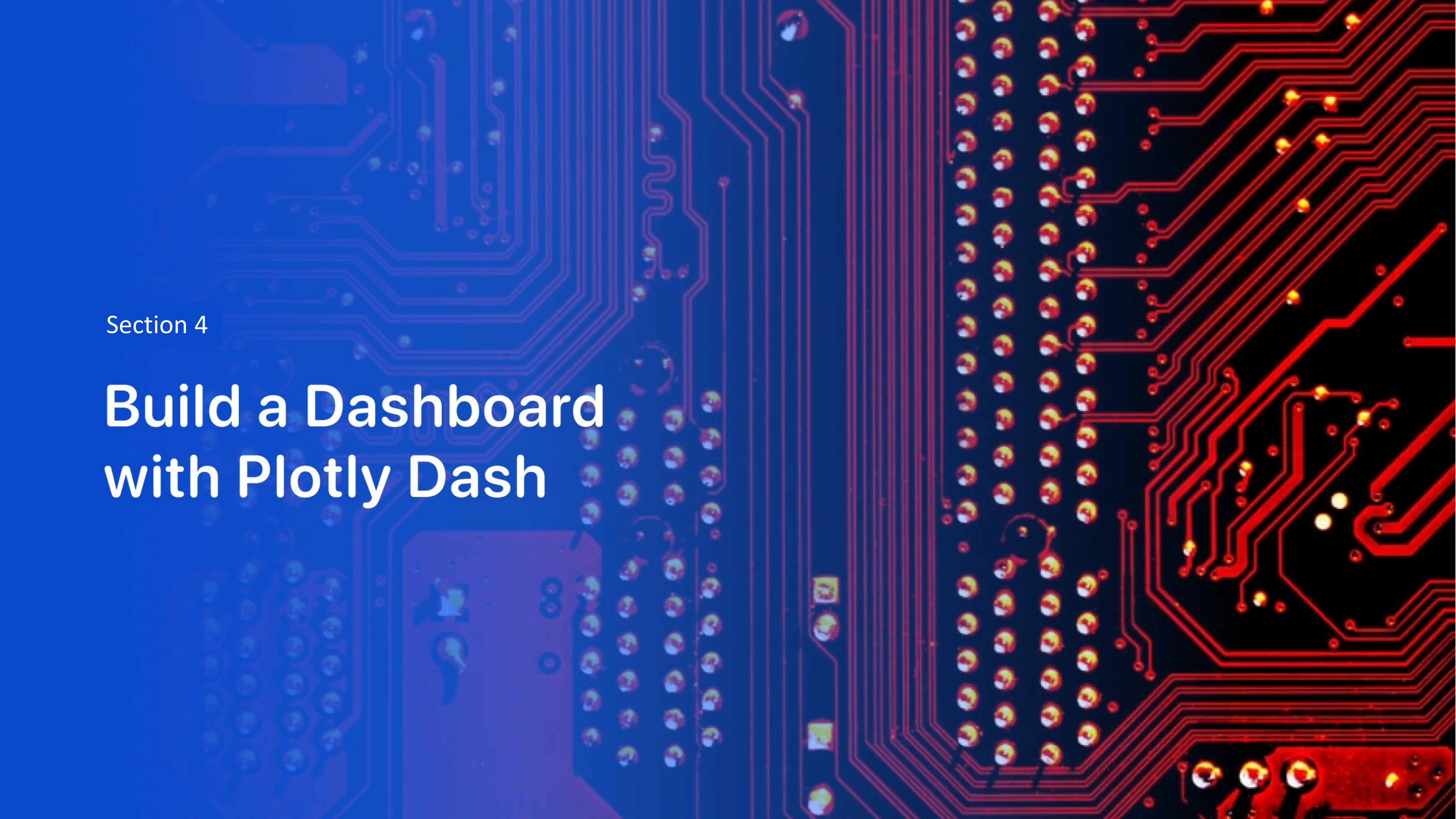
Proximity to major infrastructure:

- Highway (24.73 km)
- Railway (15.10 km)
- Coastline (6.47 km)



Section 4

Build a Dashboard with Plotly Dash



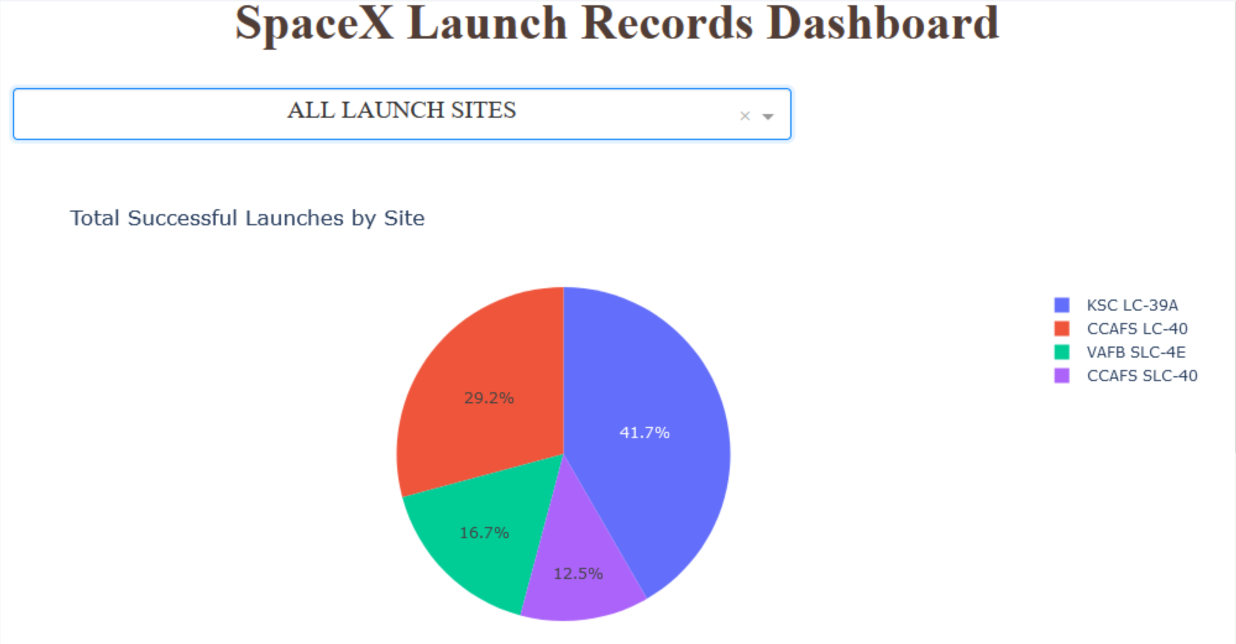
Dashboard – Total Successful Launches by Site

Layout

- Launch sites can be selected by the dropdown menu
- The proportion of total successful launches are segmented and shown in the pie chart

Findings:

- KSC LC 39-A has the most successful launches
- CCAFS SLC-40 has the least successful launches



SpaceX Launch Records Dashboard

ALL LAUNCH SITES

Total Successful Launches by Site

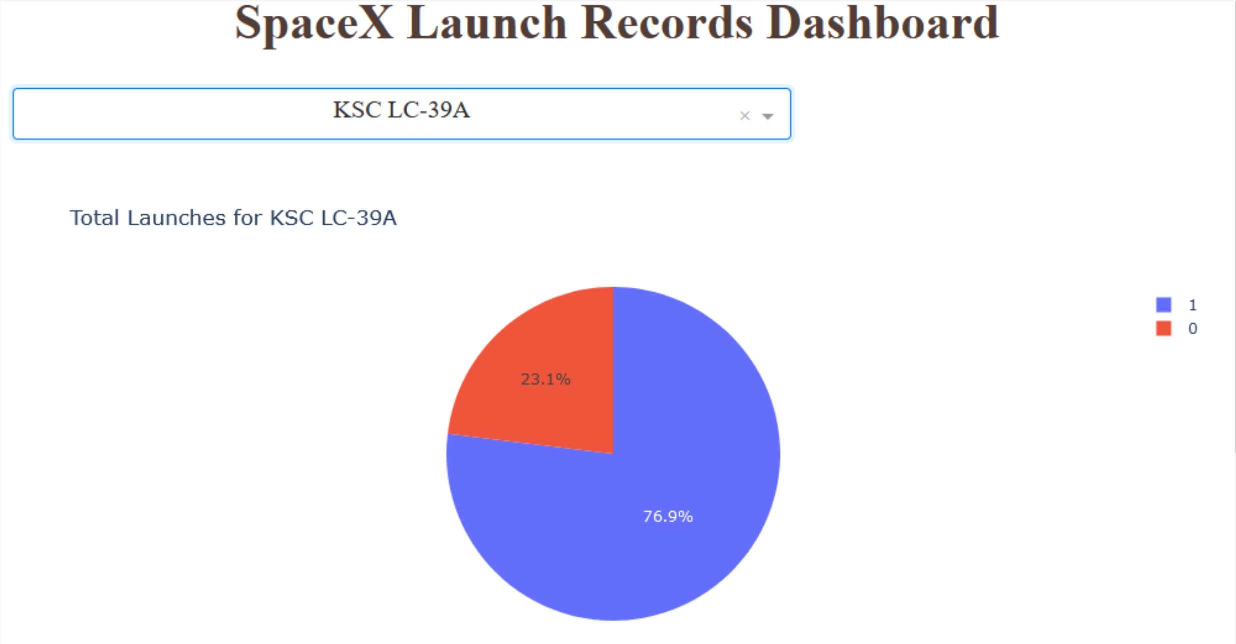
Dashboard – Most Successful Launch Site

Layout

- Once a particular launch site is selected, the pie chart is segmented into the proportion of successful and failed launches

Findings:

- Approx. 77 % of launches at KSC LC-39A resulted in success, i.e., 23 % resulted in failure



SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for KSC LC-39A

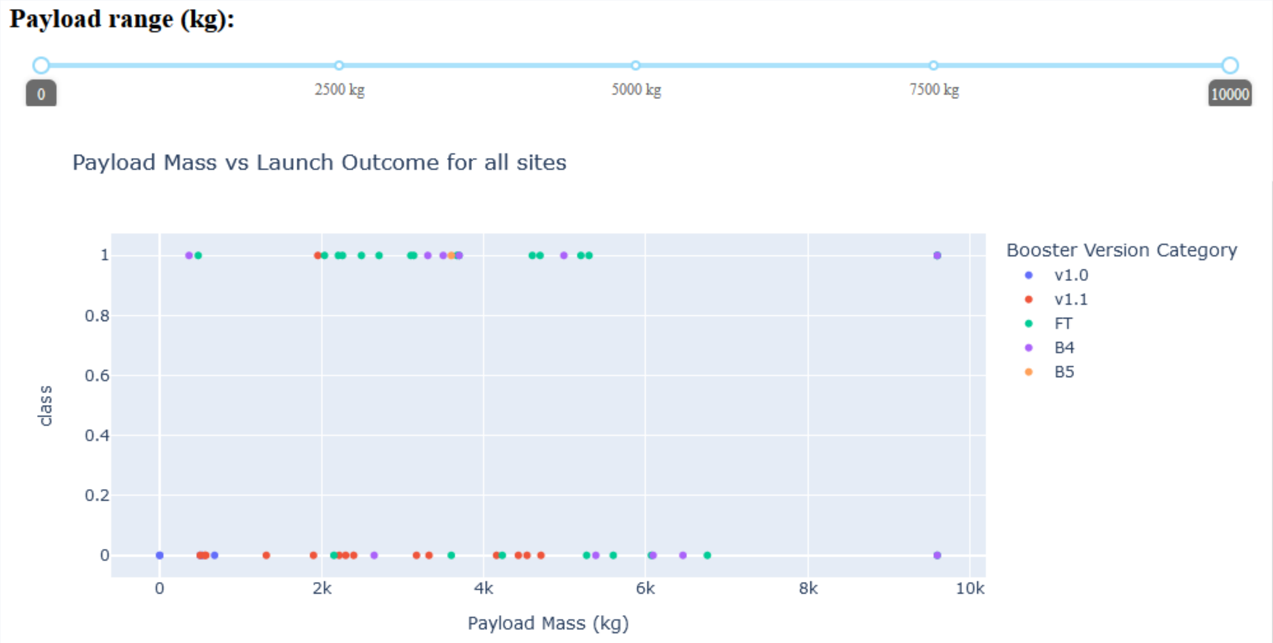
Dashboard – Payload Mass vs Landing Outcome

Layout

- The slider allows the user to filter by a payload mass range
- The scatter plot is also connected to the dropdown menu

Findings:

- The highest launch landing success is within the range of (1 900, 3 700) kg
- The FT booster appears to show the highest landing success rate



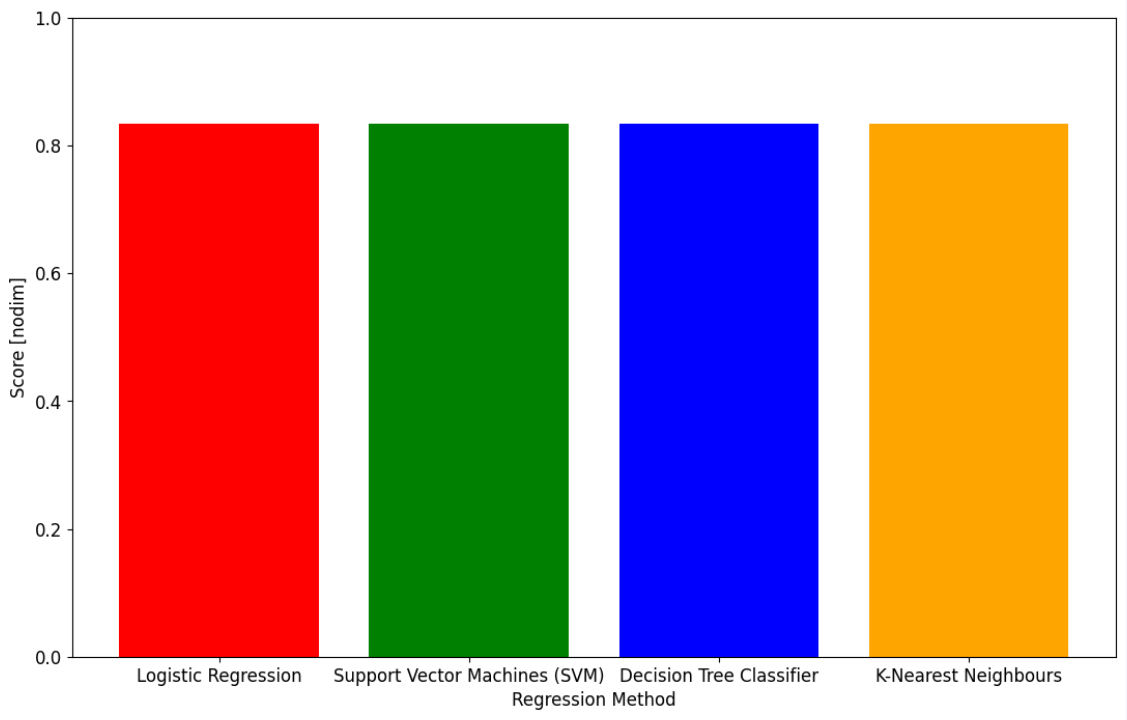
Section 5

Predictive Analysis (Classification)

Classification Accuracy

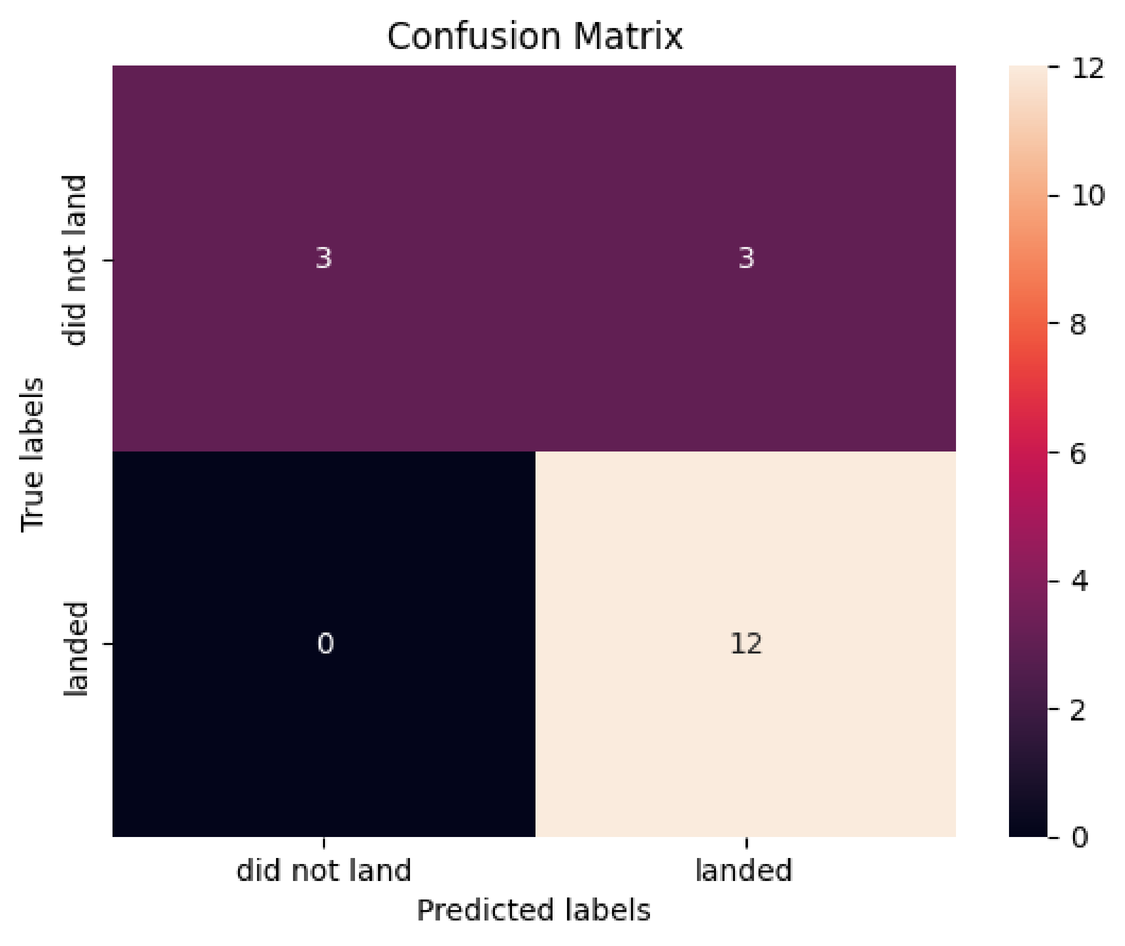
- Four models were used:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. K-Nearest Neighbors (KNN)
- All models produced the same accuracy of **83.33 %.**
- This is likely due to:
    - The limited data used for testing, and
    - The random seed used when splitting the date

Regression score based on test data



Confusion Matrix

- All models produced the same confusion matrix.
- The models correctly predicated all successful landings (i.e., True Positives).
- The model only correctly predicted half of all unsuccessful landings.
- The model tends to produce **false positives** (i.e., predicted a successful landing when an unsuccessful landing occurred).



Conclusions

- The aim of this work was to assist in the prediction of the outcome of Stage 1 landings during rocket launches at SpaceX.
- EDA was performed using SQL queries, a Folium-based notebook, and a Plotly Dash web app.
- Significant improvement in success rate began in 2013. 2018 was the only year which saw a decrease in launch success rate.
- The highest landing success occurs in the range of (1 900, 3 700) kg.
- Four machine learning models were used. All models were able to predict the landing outcomes of Falcon 9 rocket launches with 83.33 % accuracy. The models tended to overpredict (False Positive) successful launches.
- Orbit types **GTO**, **ISS**, **LEO**, **MEO**, and **PO** are probably the most accurate representation of the actual success rate, due to their having the most observed launches.

Appendix

- Check out my GitHub repo

Thank you!