SPACE X, DATA SCIENCE

Sara De La Llama

Dec 30th, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- SpaceX faces the strategic challence of accurately estimating the cost of Falcon 9 launches, a factor that depends heavily on wether the rocket,s first stage can be successfully recovered after lunch. Improving this estimation requires understanding the conditions under which a landing is likely to succeed.

- This project analyzes historical Falcon9 launch and landing data to identify the key variables associated with successful first-stage recovery. Using publicly available data, the study focuses unrecovering patterns that explain how mission characteristic influence landing outcomes.

- Launch data were collected through the SpaceX API and subsequently processed and explored using data analysis tools such as pandas, SQL and Folium. Exploratory Data Analysis (EDA) and interactive visualizations were used to highlight relantionships between launch parameters and landing success.

- Based on these insights, a classification model was developed to predict the probability of a successful first-stage landing. The model considers factors such as orbit type, launch site, landing pad, and booster serial number, providing a data-driven approach to support cost estimation and mission planning for Falcon9 launches.

Introduction

- The growing demand for ore cost-efficient space missions has made launch cost optimization a key challenge in the aerospace industry. For SpaceX, the cost of a Falcon 9 mission is strongly influenced by wether the rocket,s first stage can be successfully recovered after launch.
- Accurately predicting landing success is therefore essential for both operational planning and cost estimation. This study focuses on identifying the main factors that contribute to a successful Falcon 9 fist-stage landing by analysing historical launch and recovery data.
- The project begins with data collection from publicly available sources, followed by an exploratory data analysis using tools such as pandas, SQL and Folium to uncover relevant patterns. Finally, a classification model is developed to estimate the probability of landing success for future launches.

Methodology

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA), SQL
- Interactive Visual Analytics with Folium and Plotly Dash
- Predictive Analysis with classification models

Data Collection

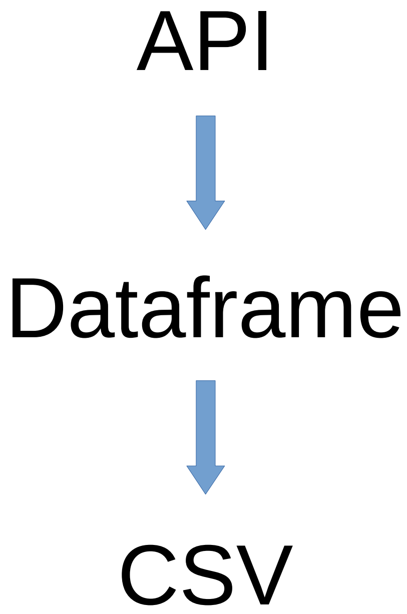
- This project uses public data to analyze Falcon 9 launch and landing outcomes.

- The main data source is the SpaceX public API, which provides structured launch information.

- Additional historical data were collected from Wikipedia to complement and validate the dataset.

Data Collection-SpaceX API

- Launch data were retrieved from the SpaceX API in JSON format.
- Relevant variables were parsed into a Python dataframe for analysis.
- The cleaned dataset was exported to CSV format for further use.



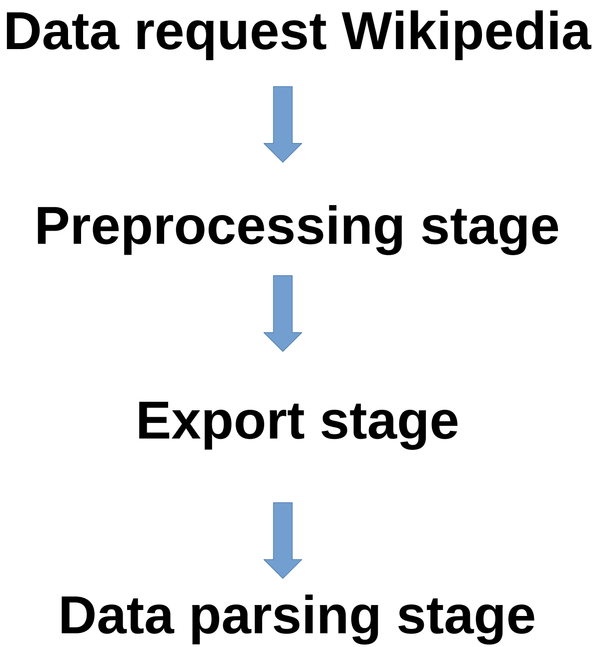
API

Dataframe

CSV

Data Collection-Web Scraping

- Historical Falcon 9 data were extracted from Wikipedia using web scraping techniques.
- HTML content was cleaned and parsed into a structured Python dataframe.
- The processed data were exported to CSV format and merged with API data.



Data request Wikipedia

Preprocessing stage

Export stage

Data parsing stage

Data Wrangling

- Falcon 9 launch data were prepared for analysis using structured CSV files.
- Mission outcomes were encoded as binary labels (1 = successful landing, 0 = failure).
- Launch sites, orbit types, and mission results were summarized.
- An overall landing success rate of approximately 66% was obtained.

Exploratory Analysis with SQL

- SQL queries were executed on Falcon 9data stored in IBM DB2.

- Queries explored launch sites, payload mass, booster versions, and outcomes.

- Temporal and categorical patterns in mission success were identified.

Exploratory Data Visualization

- Multiple visualizations were created to explore feature relationships.

- *Scatter plots analyzed landing success against payload, flight number, and orbit.*

- Bar and line charts highlighted success trends by orbit type and year.

Interactive Map (Folium)

- Launch locations were visualized using interactive maps.

- Successful and failed landings were displayed by site.

- Distances to nearby infrastructure were calculated and visualized.

Interactive Dashboard (Plotly Dash)

- An interactive dashborad was developed using Plotly Dash.
- Users explored landing outcomes by launch site and payload range.
- Filters enabled fast and flexible data exploration.

Predictive Analysis (Classification)

- Several classification models were evaluated (SVM, Logistic Regression, KNN, Decision Tree).
- Features were encoded and standardized before training.
- Models were compared to select the best-performing classifier.

- Data Preparation
- Model Training
- Model Selection

Results

Exploratory Analysis Insights

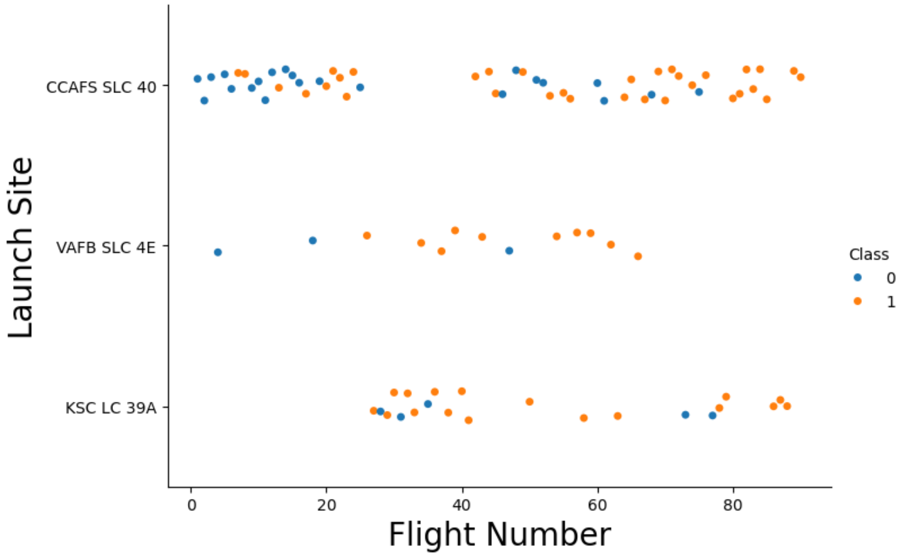
- Landing success shows a clear improvement over time.
- Overall first-stage recovery rate is approxemately **66%**.
- Launch performance varies significantly across launch sites.
- Heavy payload missions were not launched from the VAFB-SLC.
- Certain orbits (ES-L1, GEO, HEO, SSO) achieved consistently high landing success.

Predictive Modeling

- Logistic Regression, SVM and KNN achieved the highest accuracy.
- Best classification performance reached **83% accuracy.**

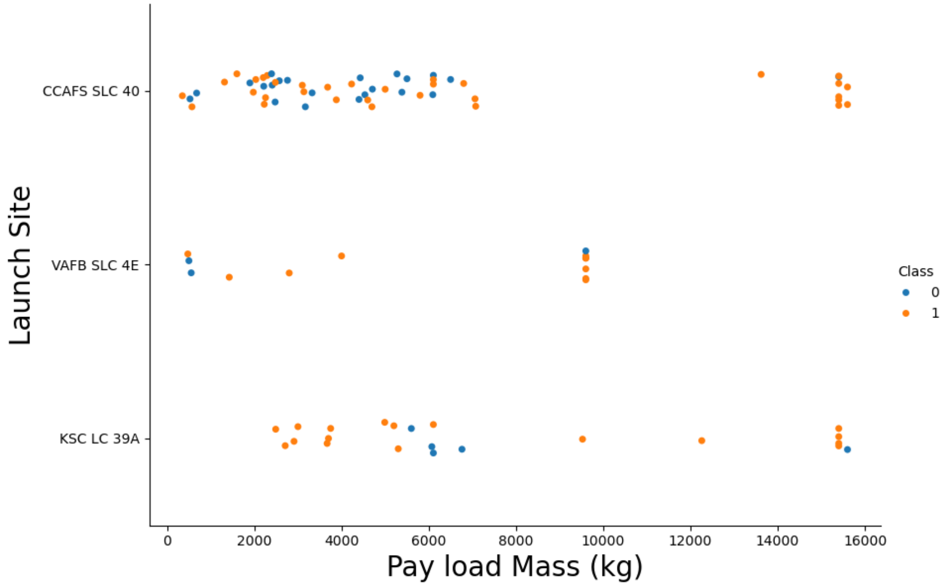
EDA

Flight Number and Launch Site



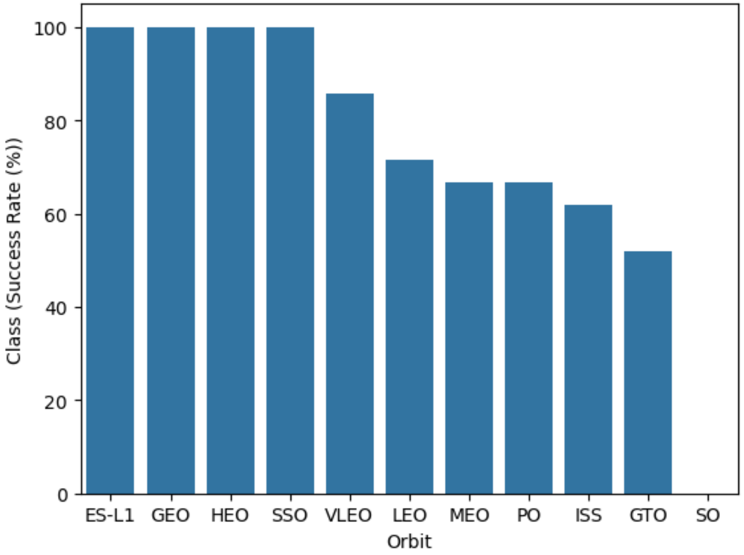
- Landing success increases as the number of Falcon 9 flights grows, especially at CCAFS SLC-40 and KSC LC-39A.
- VAFB SLC-4E shows fewer launches and more variability in landing outcomes.

Payload Mass and Launch Site



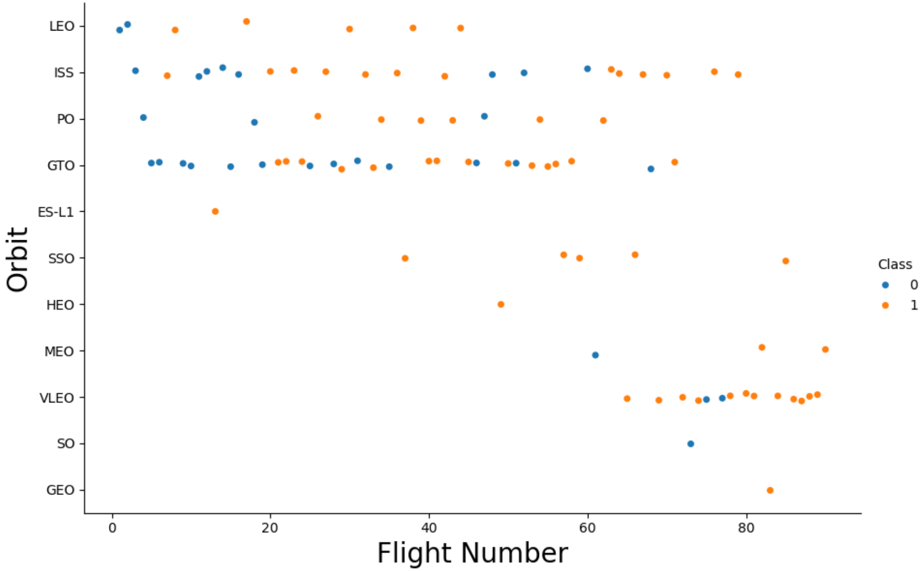
- Succesful landings are achieved across multiple payload ranges, with higher consistency at CCAFS SLC-40 and KSC LC-39A compared to VAFB SLC-4E.

Success rate of each orbit type



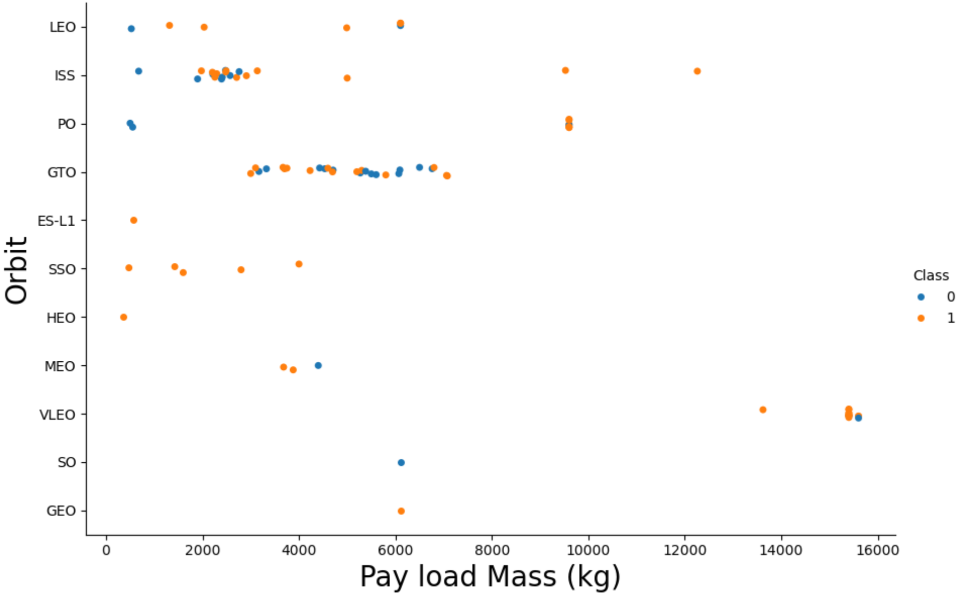
- Landing success varies by orbit, with near-perfect rates for ES-L1, GEO, HEO and SSO and lower peformance for more demanding orbits such as GTO and SO.

- Flight Number and Orbit type



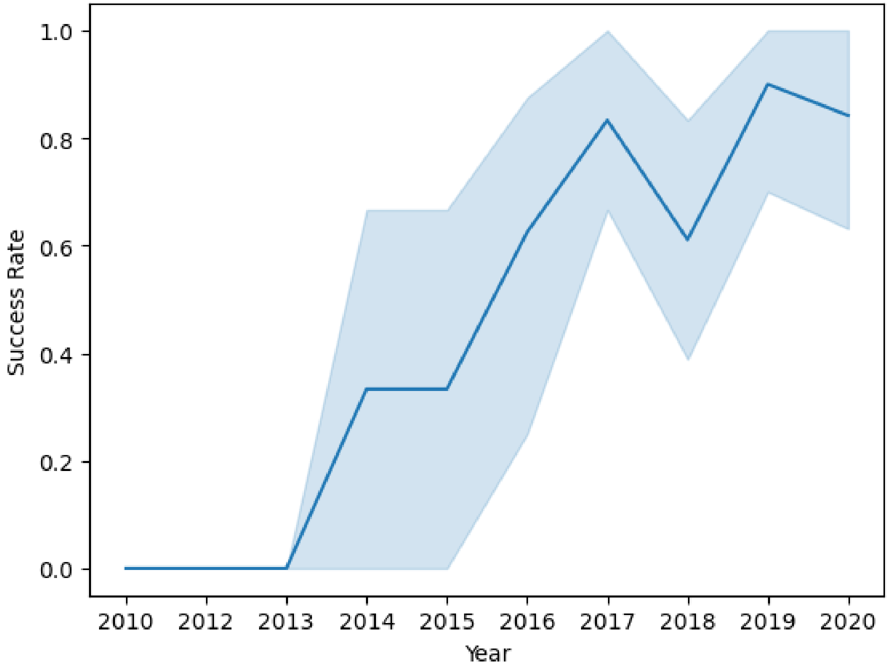
- As flight experience increases, successful landings become more frequent across most orbit types, indicating a learning and optimization effect over time.

- Payload mass and Orbit type



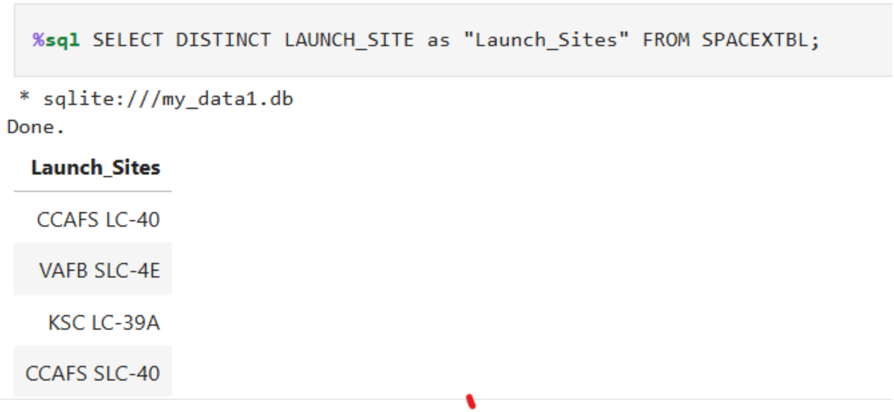
- Successful recoveries are achieved across different payload masses, though heavier payloads are concentrated in specific orbits such as GTO, where outcomes are more variable.

Launch Success Yearly Trend



- Success rates are near zero until 2013, then rise sharply from 2014 onward and, despite some volatility, stabilize at high levels from 2017 onward.

Launch Site Names



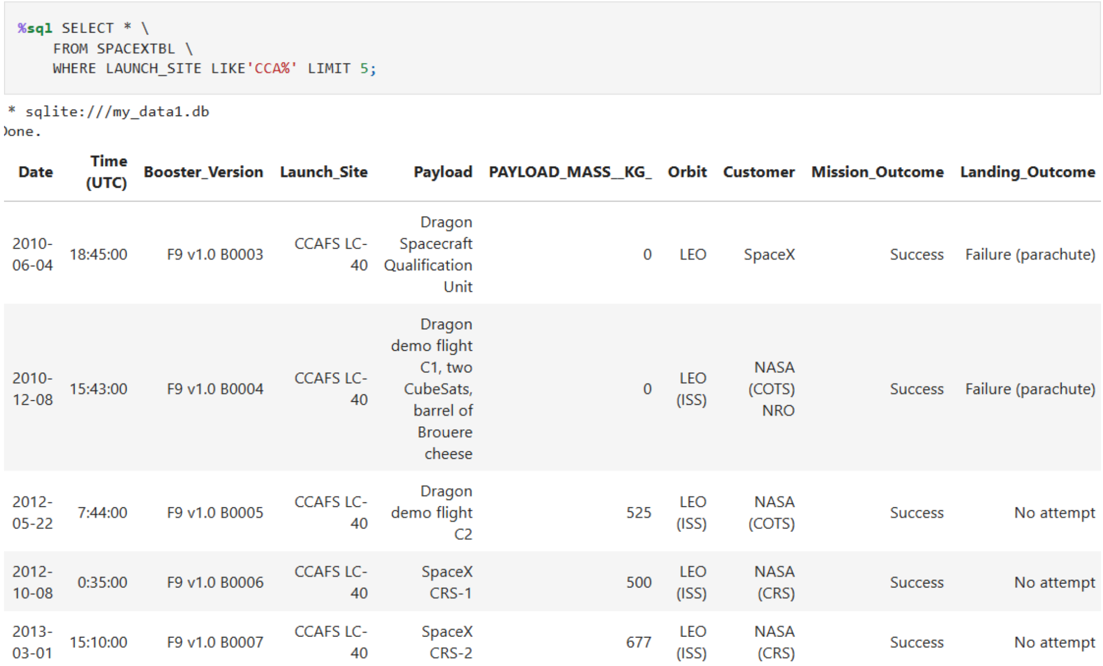
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

- The dataset includes multiple distinct launch sites.

Launch Site Names begin with ´CCA`



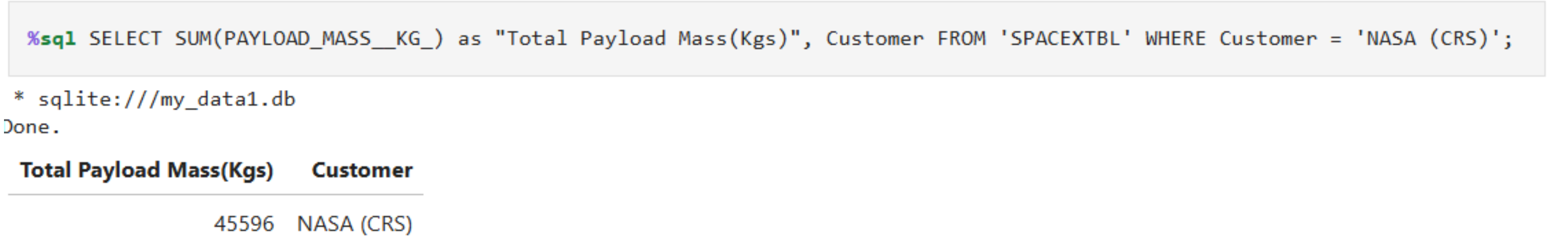
```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
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

- Filtering launch sites that start with “CCA” returns records mainly from CCAFS LC-40.

Total Payload mass



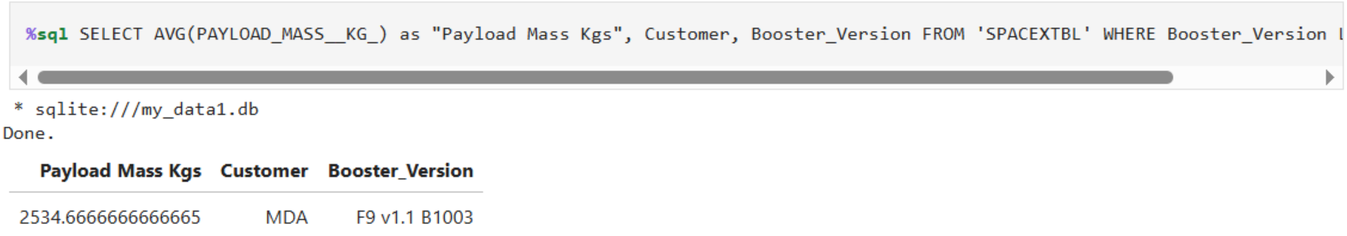
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

- The query calculates the total payload mass delivered for NASA CRS missions by summing payload mass across those launches.

- Average payload mass

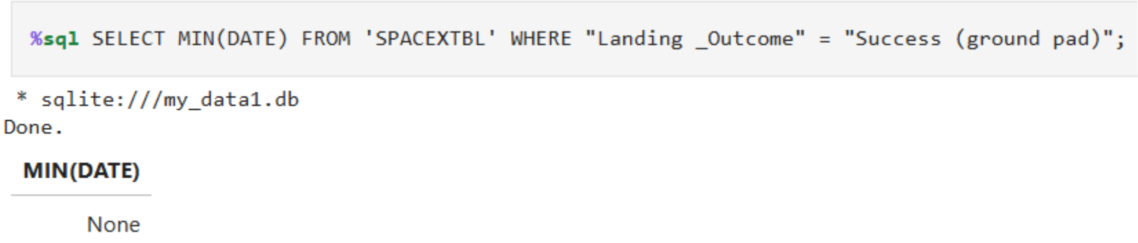


```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version L
```

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

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

- First succesful landing outcome in ground date

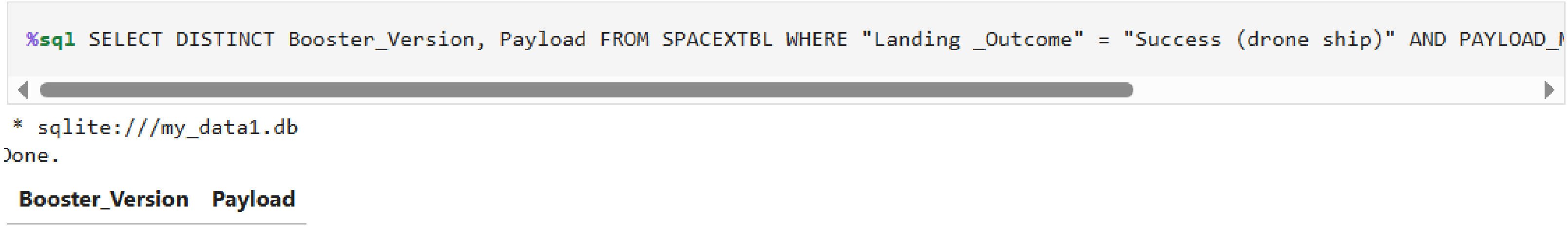


```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
```

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

MIN(DATE)
None

- Drone ship and have payload mass between 4000 and 6000

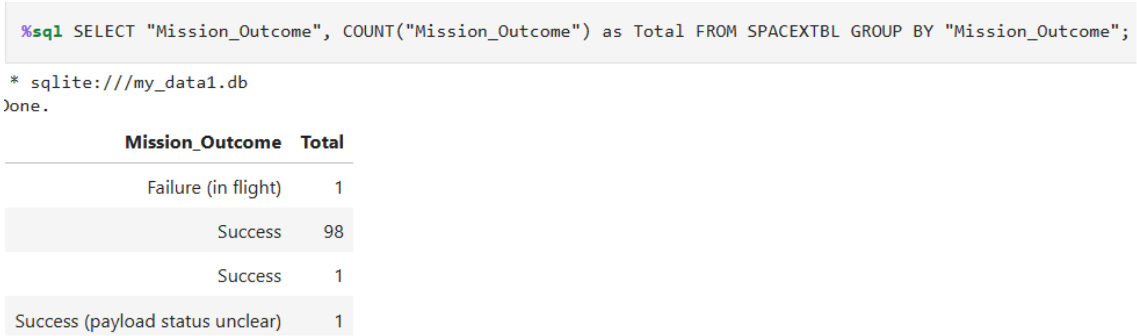


```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_M
```

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

Booster_Version	Payload

- Total number of successful and failure mission outcomes



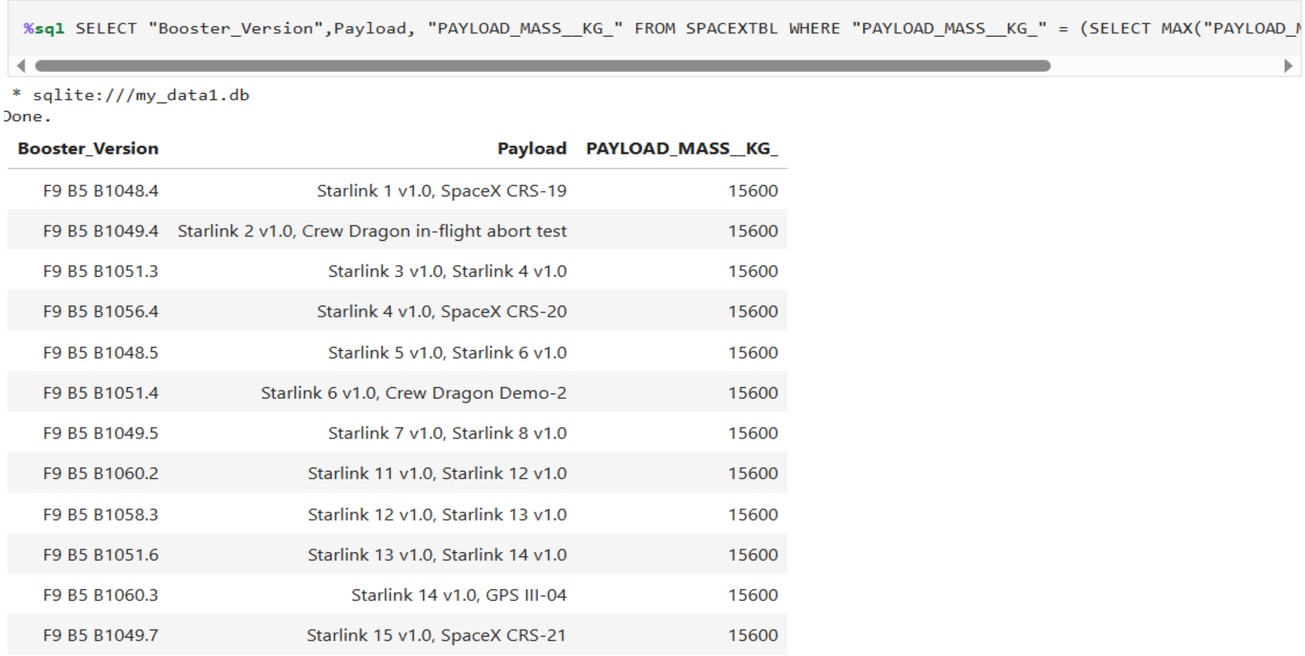
```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Failure: 1.
- Success: 98.
- Success (payload status unclear): 1.

- Boosters carried maximum payload mass

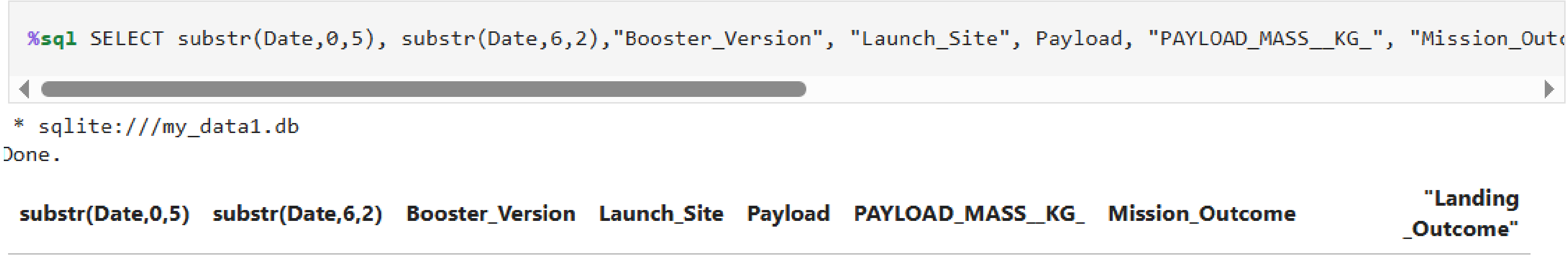


```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_
```

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

- Launch Records 2015

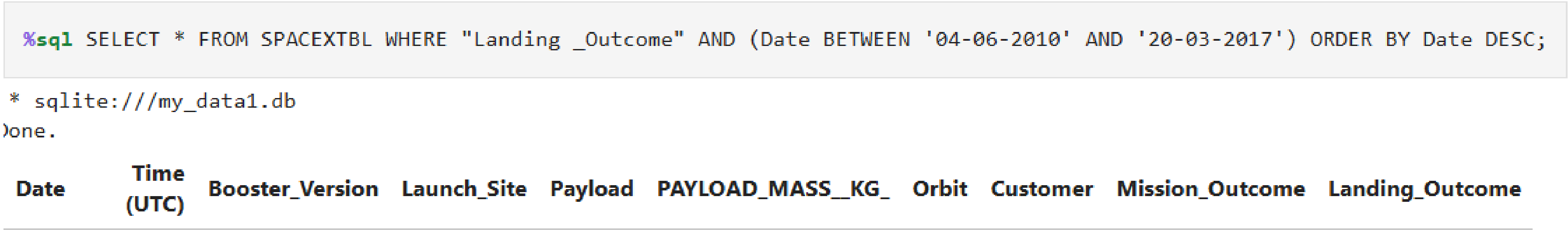


```
%sql SELECT substr(Date,0,5), substr(Date,6,2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Outc
```

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

substr(Date,0,5)	substr(Date,6,2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	"Landing _Outcome"

- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20

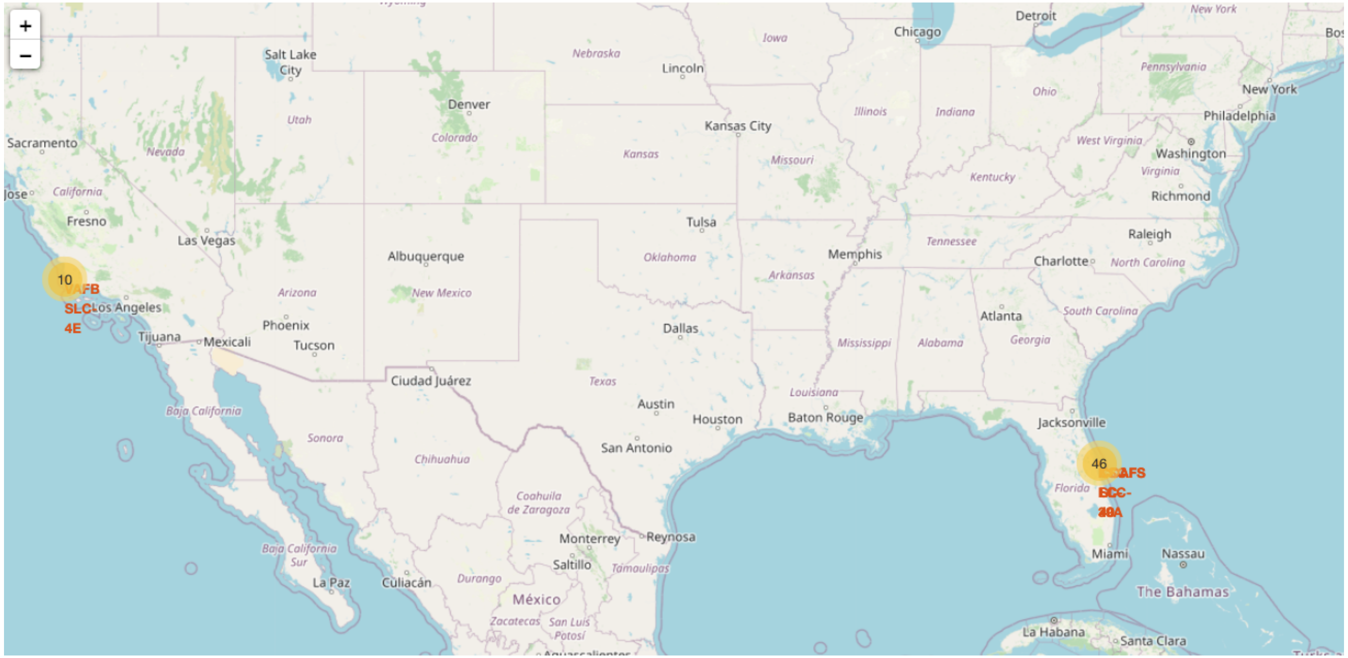


```
%sql SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" AND (Date BETWEEN '04-06-2010' AND '20-03-2017') ORDER BY Date DESC;
```

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

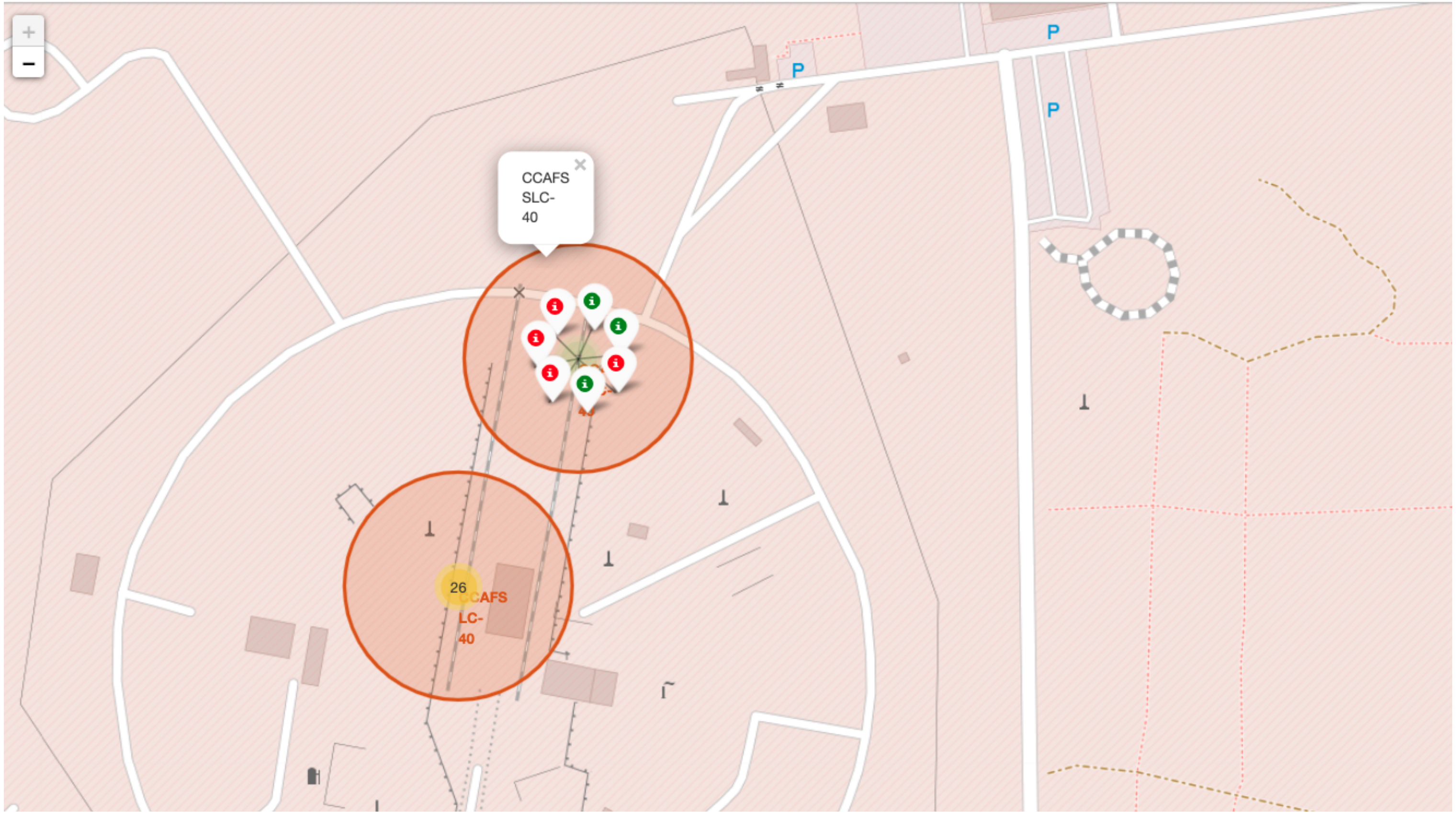
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome

Launch Site with Folium

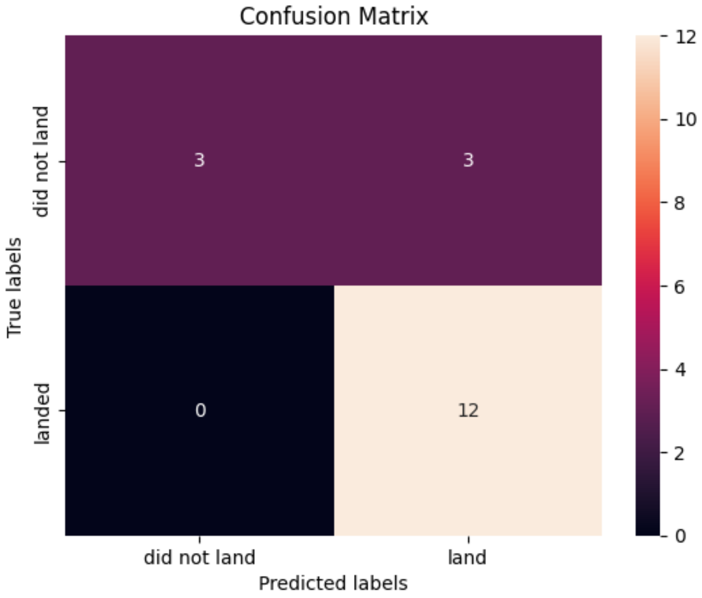


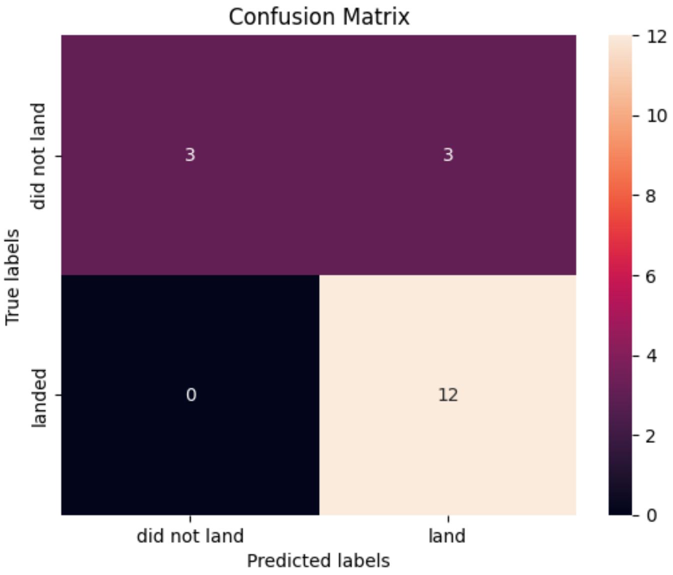
- Two cluster launch sites in Florida and California

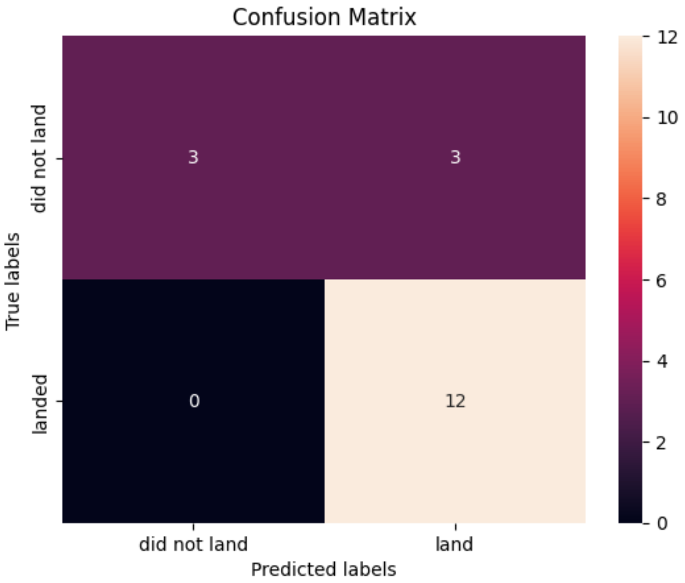
Outcome feature per launch site



Predictive Analysis (Classification)







- The Confussion Matrix are the same for this models with: 3 TP, 3 FP, 0 FN and 12 TN.

Conclusions

- Exploratory analysis indicates that payload mass has limited predictive power for landing success, except in a small number of edge cases where its influence becomes noticeable.
- Failure events are not randomly distributed: mission sequence (flicht number) and orbital class emerge as relevant explanatory variables, showing a tronger association with unsuccessful landings.
- Orbit type appears to be a high-impact categorical feature. In particular, missions targeting ES-L1, GEO, HEO and SSO show consistently positive outcomes, achieving a perfect historical landing success rate within the analyzed dataset.
- Spatial visualization using Folium maps, together with dashboard-level aggregation reveals non-trivial patterns linking launch sites to landing performance, suggesting location-dependent operational effects.
- From modeling perspective, supervised classification approaches-including Logistic Regression, Support Vector Machines and K-Nearest Neighbors-demonstrate strong suitability for this problem. When fed with features such as orbit category, launch complex, landing zone and Falcon 9 booster identifier, these models are able to effectively estimate the probabilty of a successful landing.

THE END