IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Qianli Wu
08/22/2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scraping, API calls, database queries
- Perform data wrangling
  - Data cleaning, handling missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build: Logistic Regression, SVM, Decision Trees, KNN
  - Tune: GridSearchCV for hyperparameter optimization
  - Evaluate: Accuracy, best parameters, and model performance.

Data Collection

- Identify data sources
  - Select official SpaceX website, Wikipedia, and external APIs as primary sources
- Web scraping
  - Extract data tables, launch details, and metadata from web pages
- API data extraction
  - Use REST APIs (e.g., SpaceX API).
- Data integration
  - Combine data from web scraping and API calls.
- Data storage
  - Store the collected data in CSV files and SQL databases.

Data Collection – SpaceX API

- API Endpoint Identification
- API Request Setup
- Data Retrieval
- Data Parsing
- Data Storage

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Identify Target Websites
- Inspect HTML Structure
- Set Up Web Scraper
- Parse HTML Content
- Data Cleaning and Storage

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Load Raw Data
- Data Cleaning
- Data Transformation
- Data Aggregation and Grouping
- Data Integration

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

•Explore relationships: Scatter plots.
•Compare categories: Bar charts.
•Track trends: Line charts.
•Geographical insights: Map visualization.
•Proportions: Pie charts.
•Dynamic exploration: Interactive plots (Plotly Dash).

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Unique Launch Sites
- Launches by Location
- Total Payload by Customer
- First Successful Ground Pad Landing
- Booster Versions with Specific Payload Mass
- Mission Outcome Count
- Boosters with Maximum Payload
- Failures by Month and Year
- Rank Landing Outcomes
- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers
  - To pinpoint the exact locations of SpaceX launch sites and provide context through interactive pop-ups with relevant site information.
- Circles:
  - To visually distinguish different launch sites, provide a sense of scale, and highlight regions around the launch areas.
- Lines:
  - To visually distinguish different launch sites, provide a sense of scale, and highlight regions around the launch areas.

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart (Success-Pie-Chart)
- Scatter Plot (Success-Payload-Scatter-Chart)
- Dropdown Menu (Site Selection)
- Range Slider (Payload Range)

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

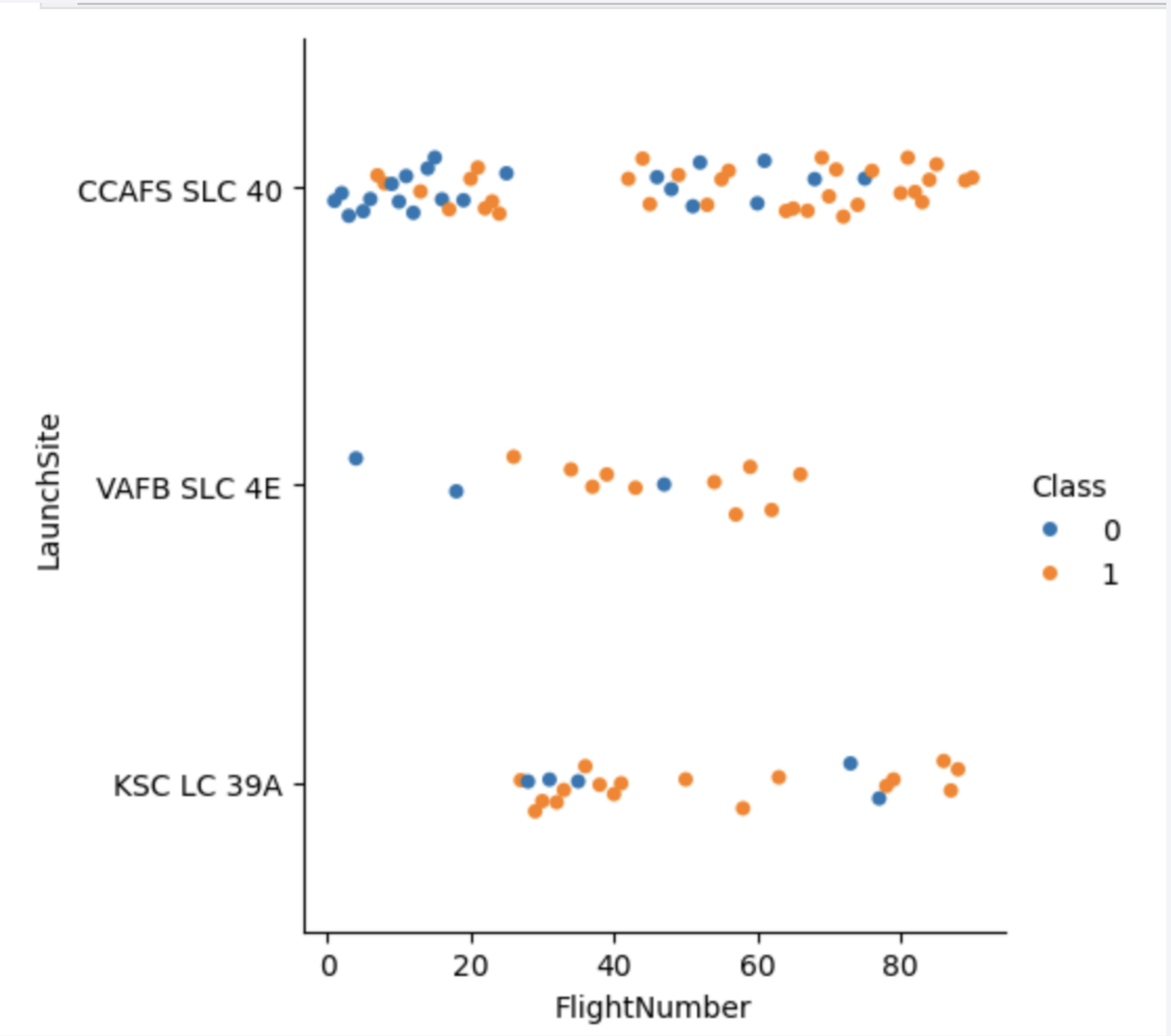
- Data Preparation
- Model Building
- Model Evaluation
- Model Improvement
- Best Model Selection

- https://github.com/Tsuuunade/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Section 2

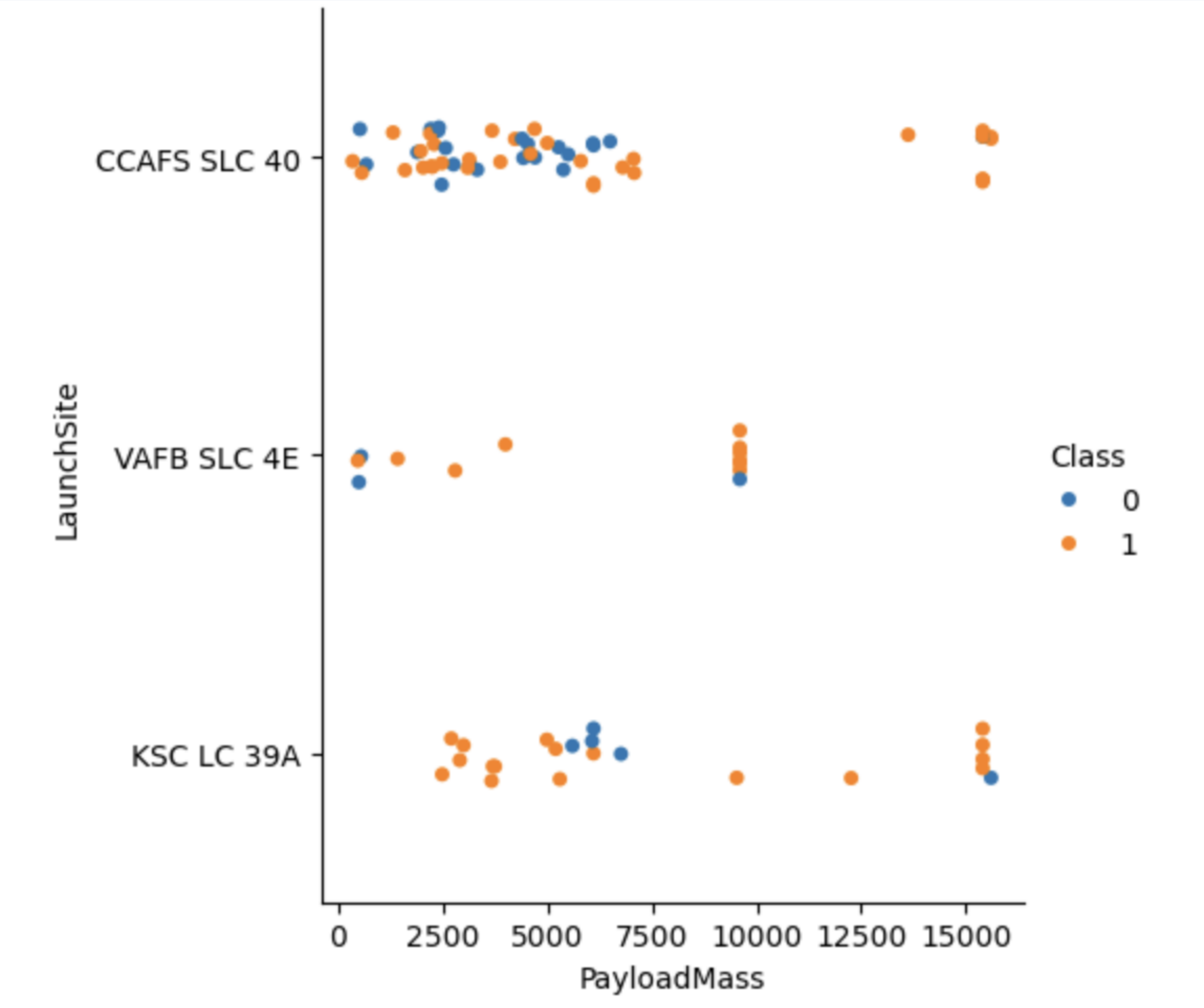
Insights drawn from EDA

Flight Number vs. Launch Site



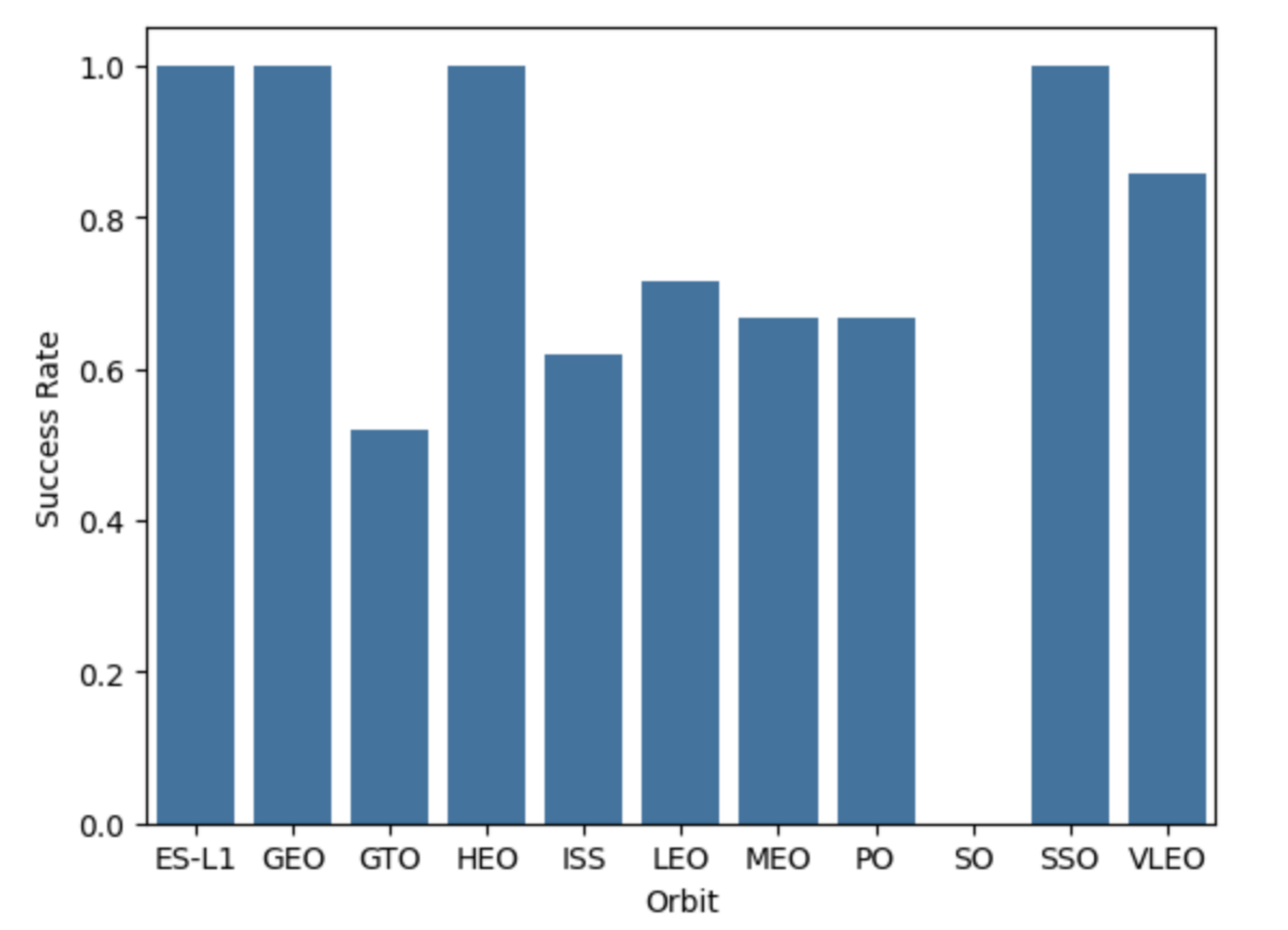
- This scatter plot shows the relationship between flight number and launch site for SpaceX launches, with the data points categorized by the success (Class = 1) or failure (Class = 0) of the mission. Each launch site (CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A) has different patterns of successes and failures, which can be observed across varying flight numbers. The distribution suggests that certain sites have had more consistent success rates (e.g., KSC LC 39A), while others show a more mixed performance.

Payload vs. Launch Site



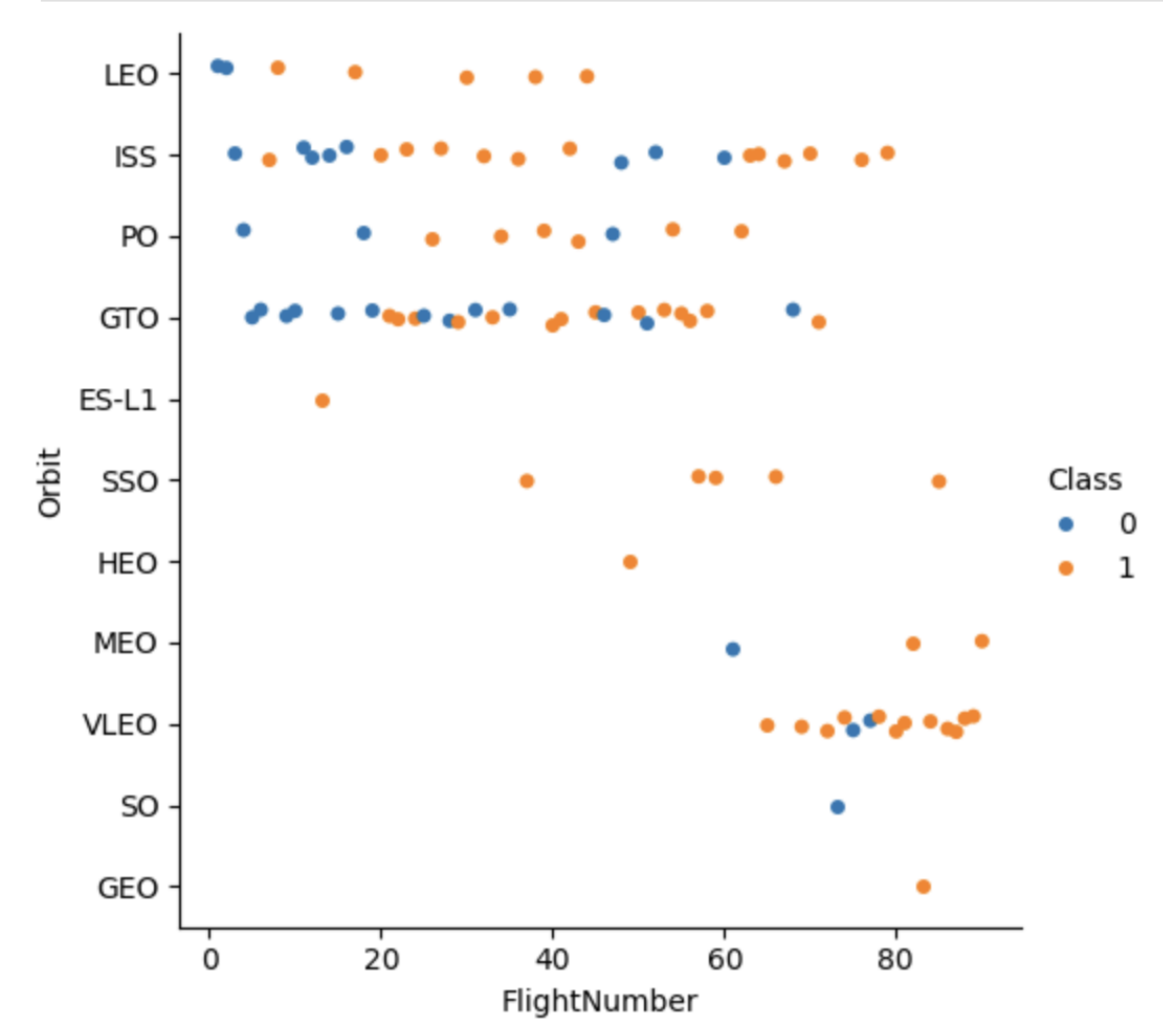
- Higher payload masses are more frequently launched from CCAFS SLC 40 and KSC LC 39A.Success (Class = 1) and failure (Class = 0) are distributed across different payload ranges; however, CCAFS SLC 40 shows successful launches at both low and high payload masses, while VAFB SLC 4E has fewer data points concentrated in a narrow payload range.
- Show the screenshot of the scatter plot with explanations

Success Rate vs. Orbit Type



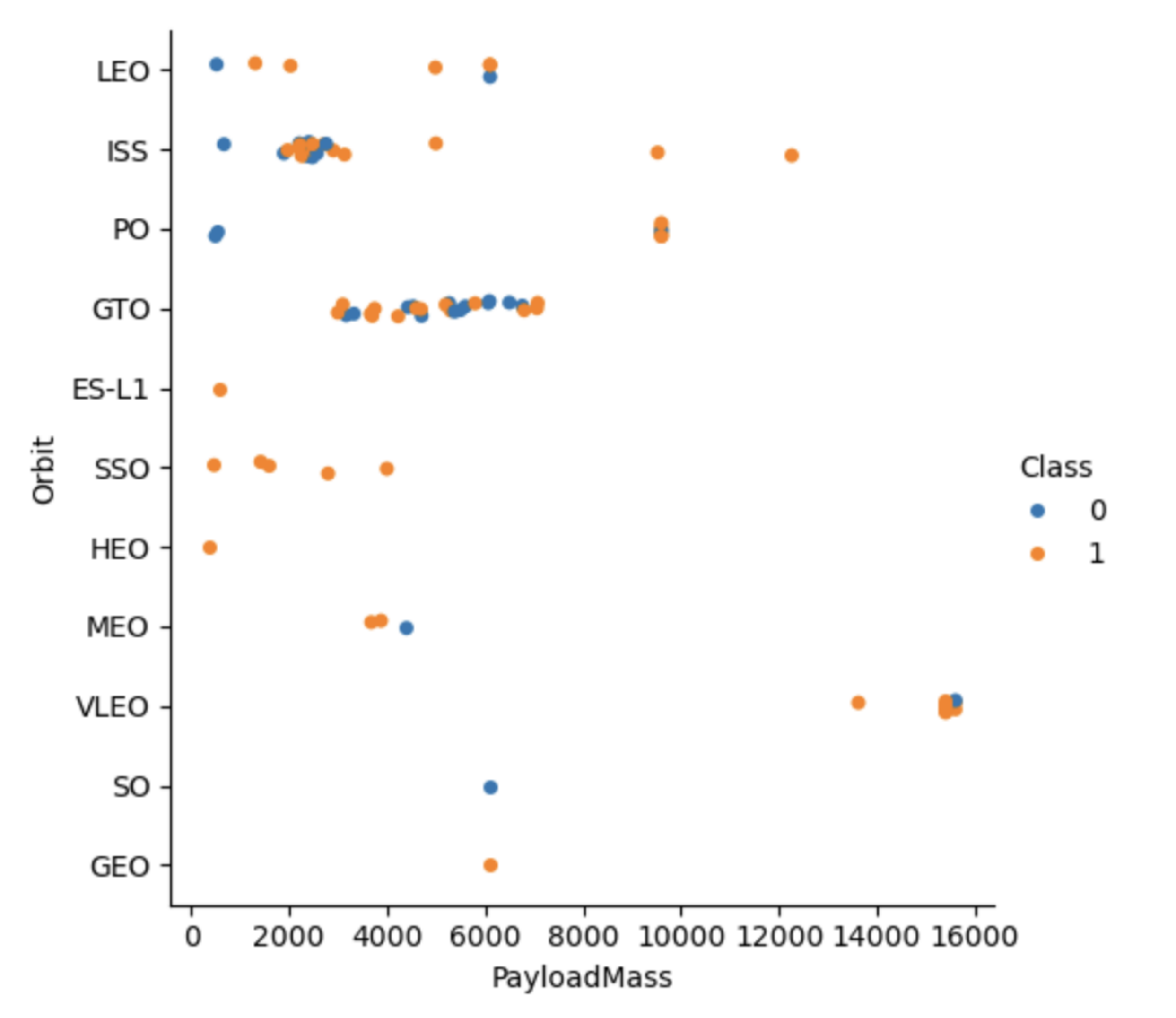
- Orbits like ES-L1, GEO, LEO, SSO, and VLEO have the highest success rates (close to 1.0).Orbits such as GTO and HEO have lower success rates, indicating more challenging missions.

Flight Number vs. Orbit Type



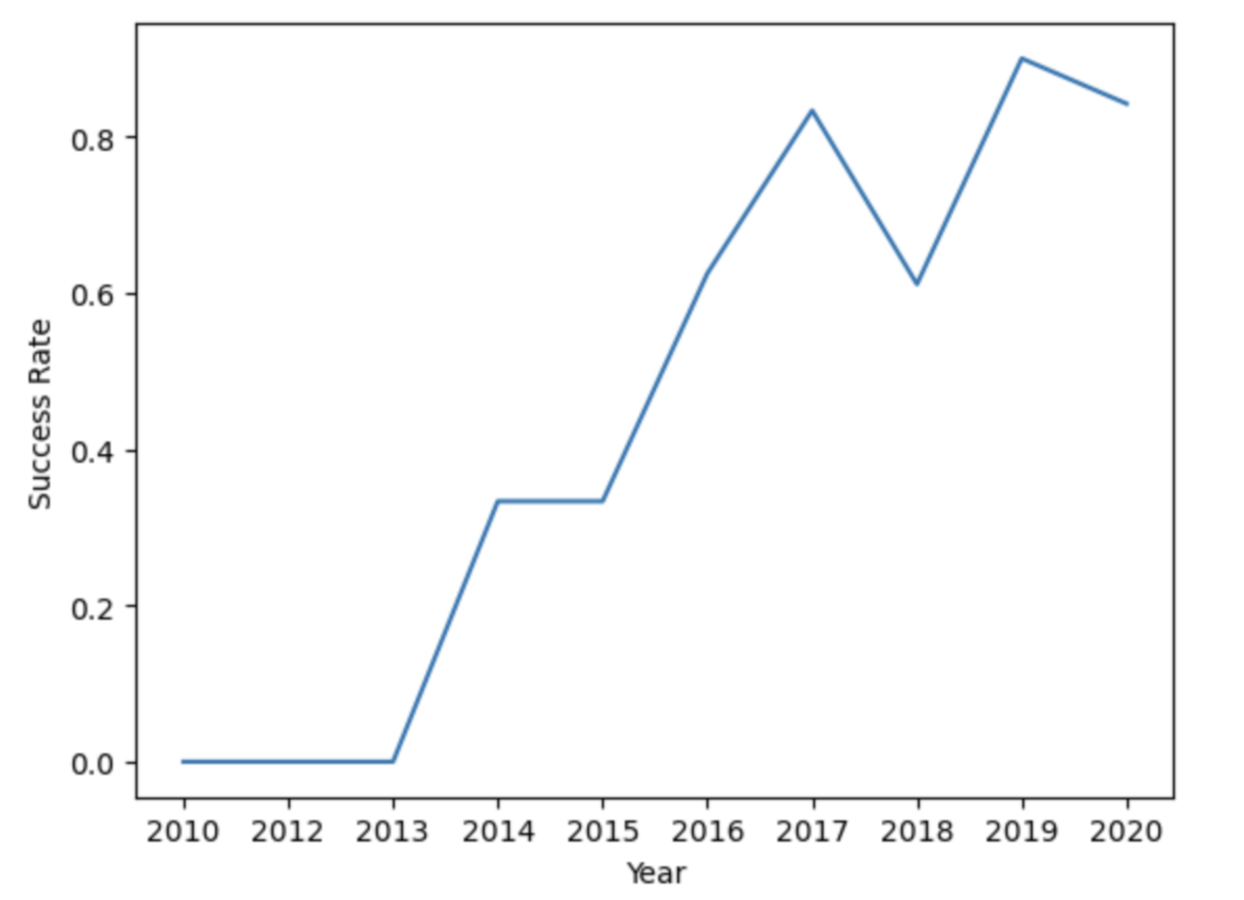
- The graph shows varying success rates across different orbits, with orbits like LEO and ISS having more successful outcomes, while orbits like GTO and HEO have a mix of successes and failures.

Payload vs. Orbit Type



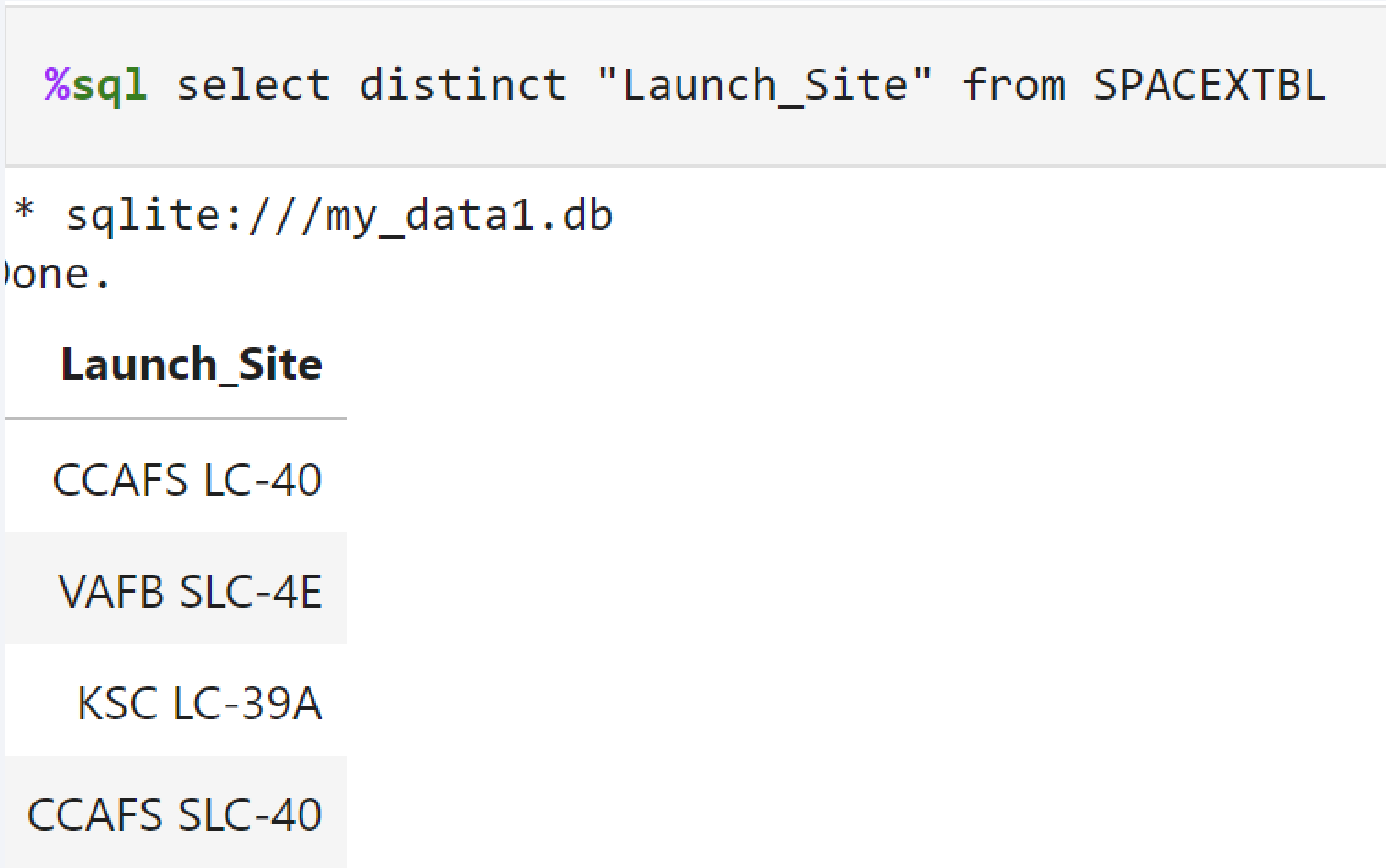
- Lower payload masses (less than 4000 kg) are observed across various orbits, with a relatively balanced mix of successes and failures.
- Higher payload masses (greater than 10000 kg) are associated with fewer orbits, mostly GTO, and exhibit more successful outcomes (Class 1).

Launch Success Yearly Trend



- The line chart illustrates the success rate of launches over time, showing a clear upward trend from 2013 to 2020. There is a significant increase in success rate starting around 2014, peaking in 2019. This indicates overall improvement in the performance and reliability of the launches over the years.

All Launch Site Names

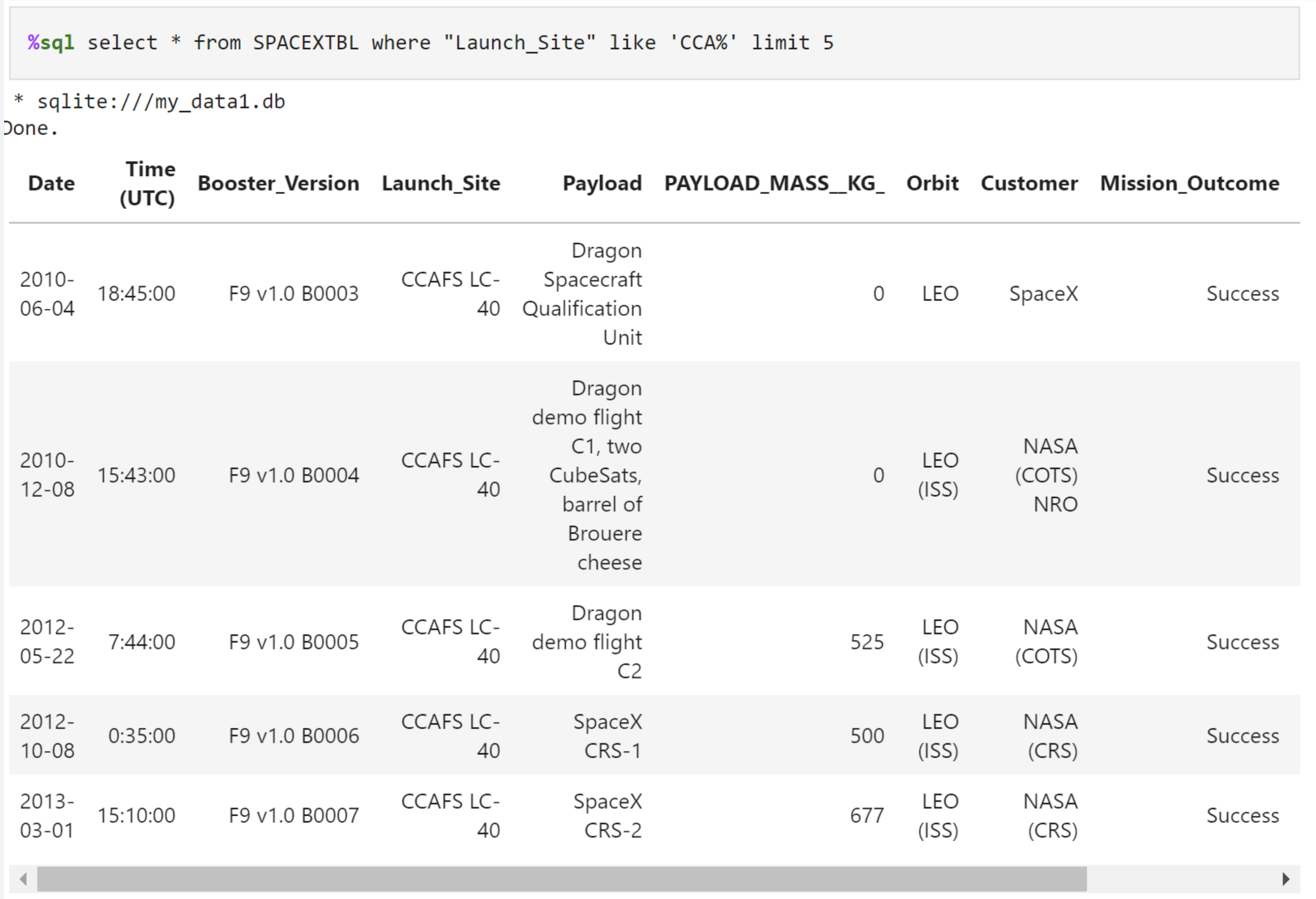


```
%sql select distinct "Launch_Site" from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

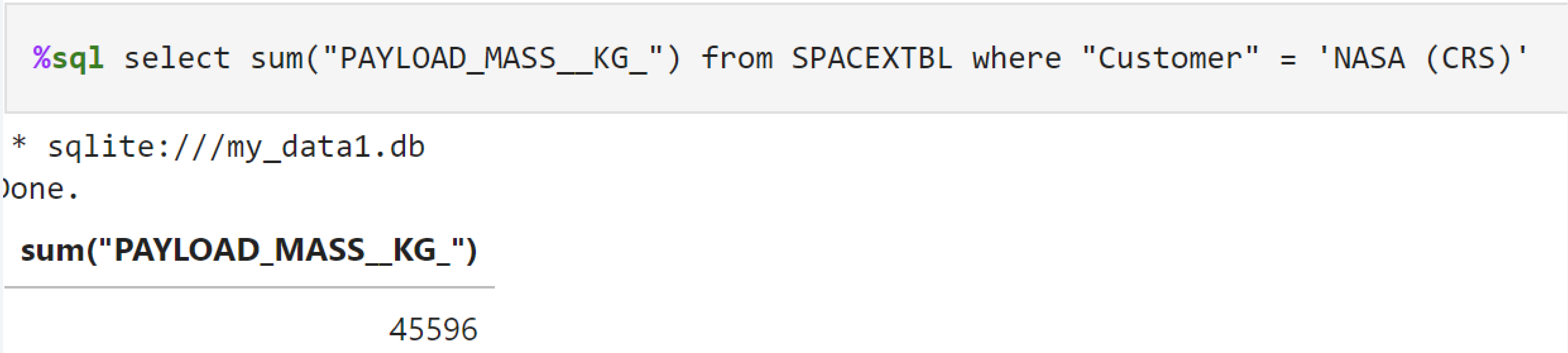


```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

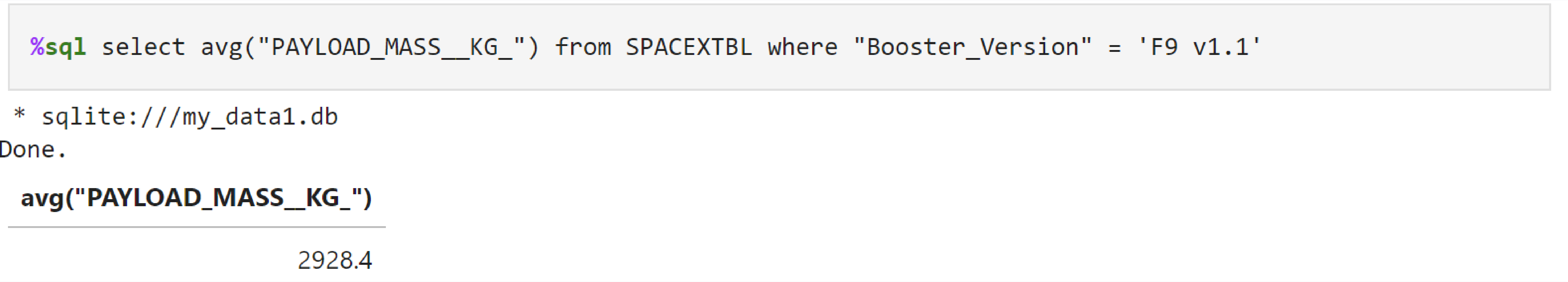


```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" = 'NASA (CRS)'
```

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

sum("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

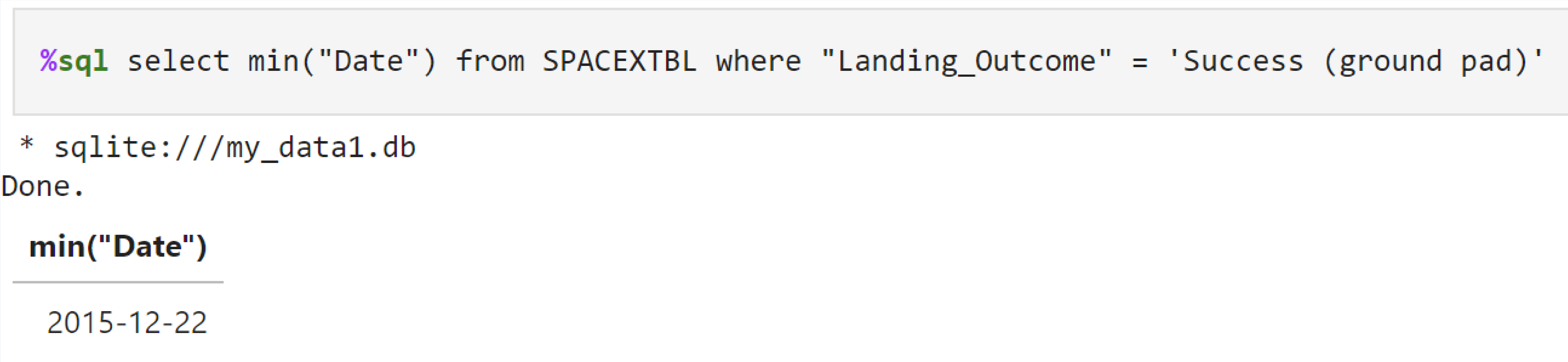


```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" = 'F9 v1.1'
```

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

avg("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

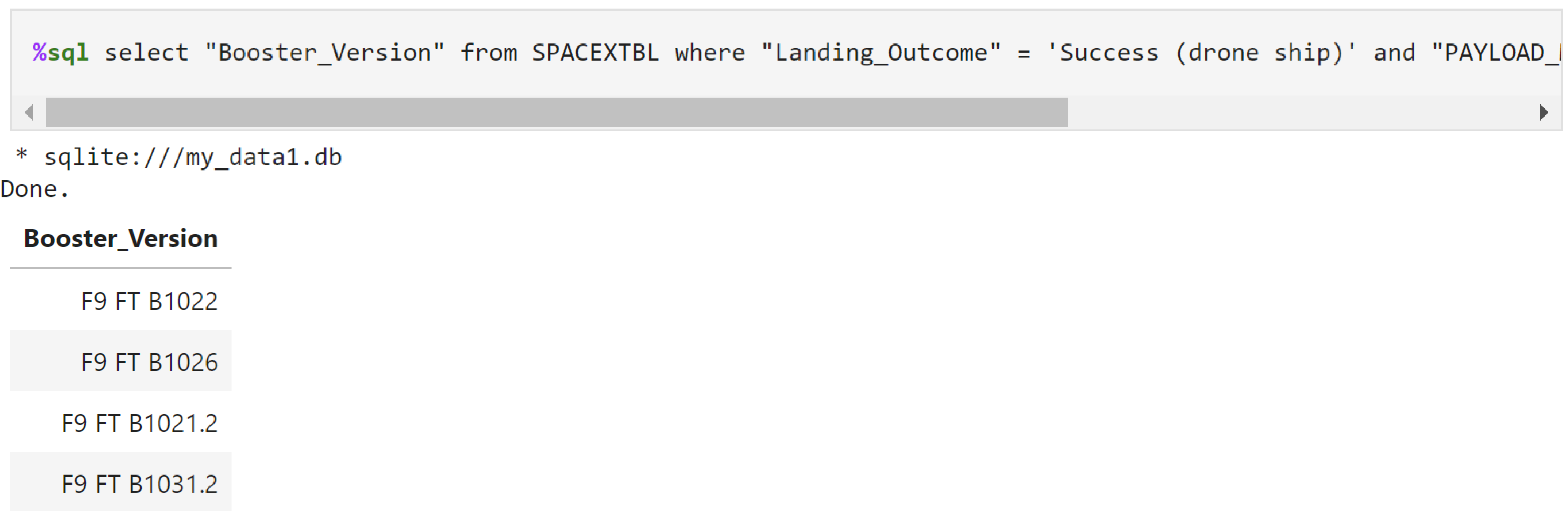


```
%sql select min("Date") from SPACEXTBL where "Landing_Outcome" = 'Success (ground pad)'
```

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

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

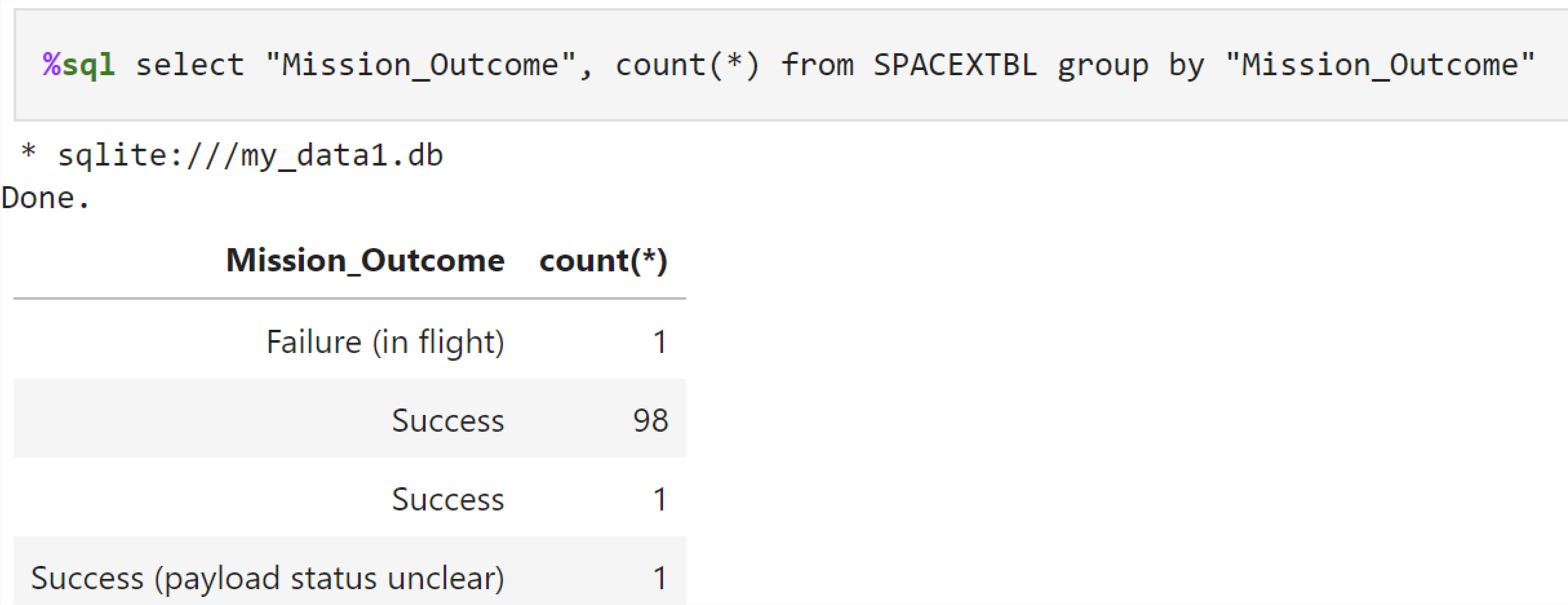


```
%sql select "Booster_Version" from SPACEXTBL where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_
```

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

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

Total Number of Successful and Failure Mission Outcomes

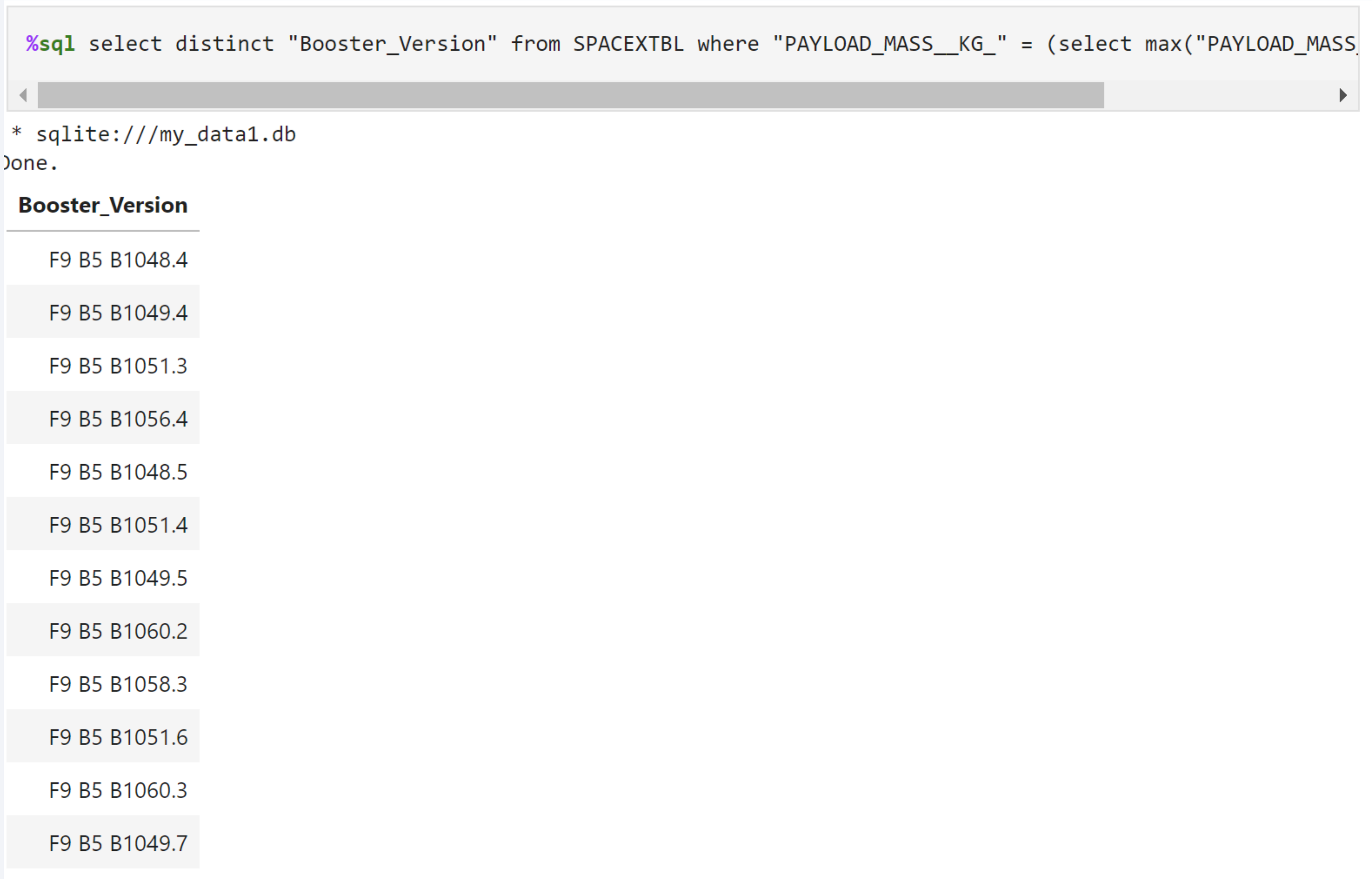


```
%sql select "Mission_Outcome", count(*) from SPACEXTBL group by "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

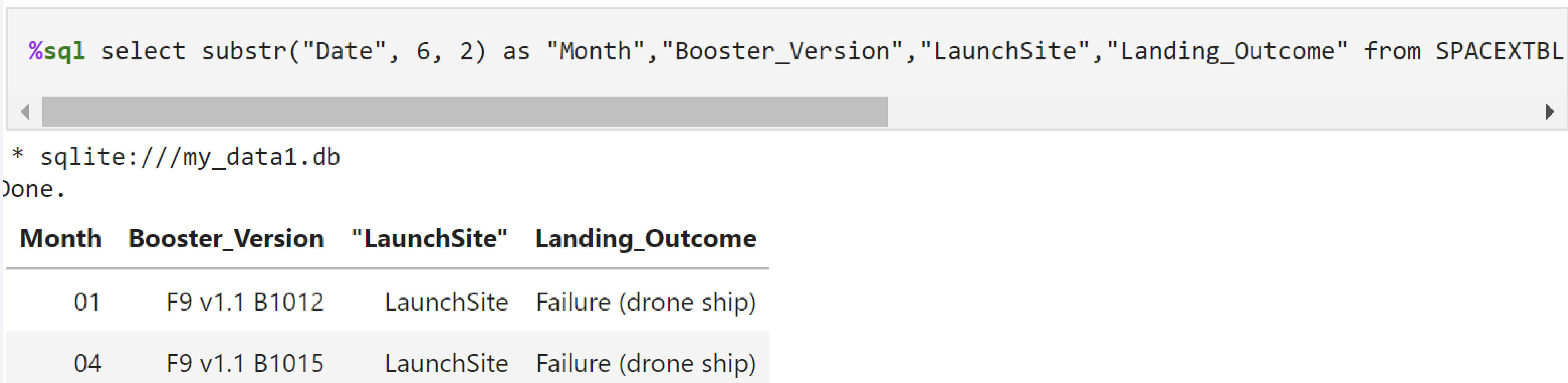


```
%sql select distinct "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS_
```

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

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

2015 Launch Records

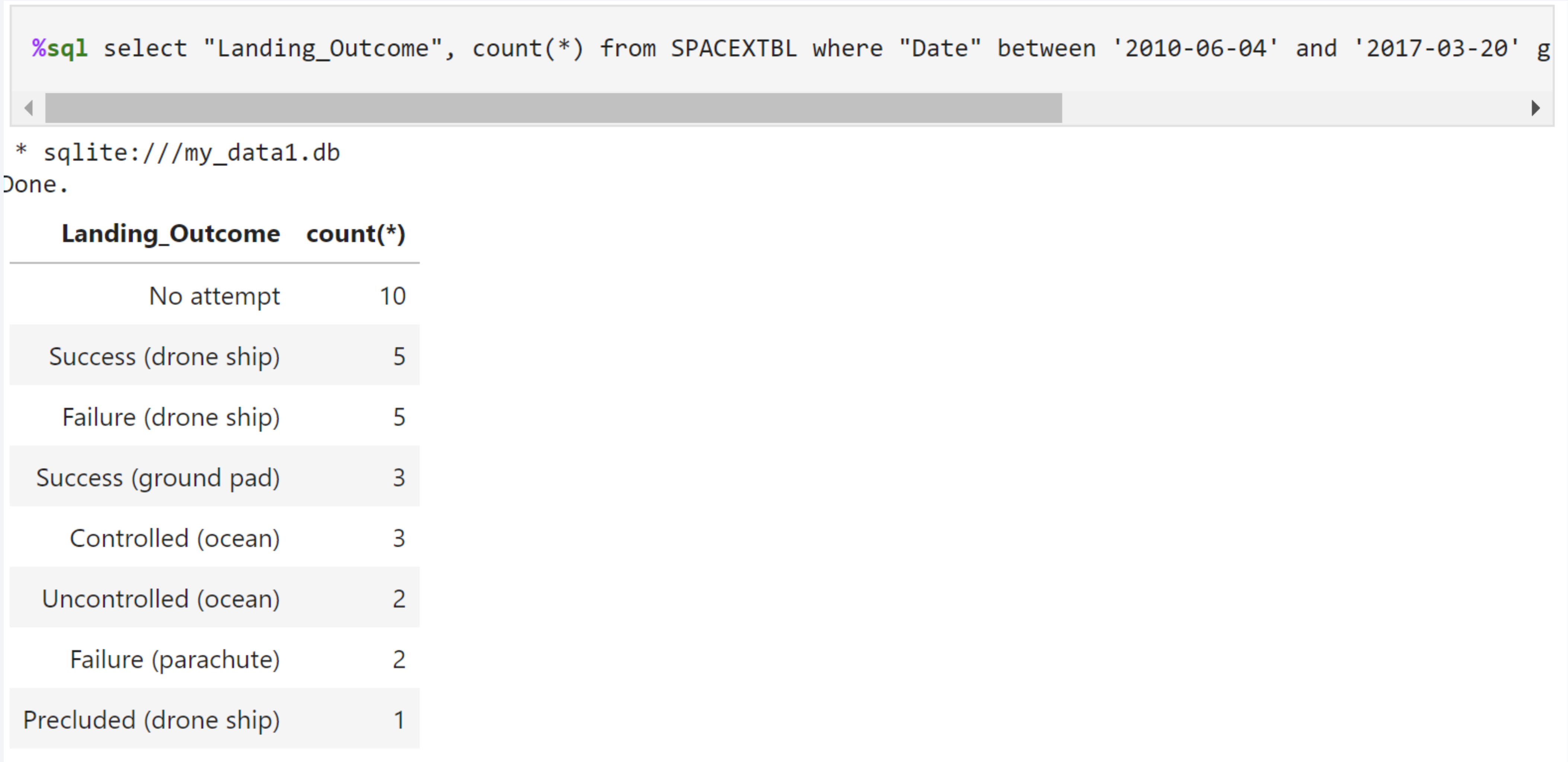


```
%sql select substr("Date", 6, 2) as "Month","Booster_Version","LaunchSite","Landing_Outcome" from SPACEXTBL
```

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

Month	Booster_Version	"LaunchSite"	Landing_Outcome
01	F9 v1.1 B1012	LaunchSite	Failure (drone ship)
04	F9 v1.1 B1015	LaunchSite	Failure (drone ship)

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



```
%sql select "Landing_Outcome", count(*) from SPACEXTBL where "Date" between '2010-06-04' and '2017-03-20' g
```

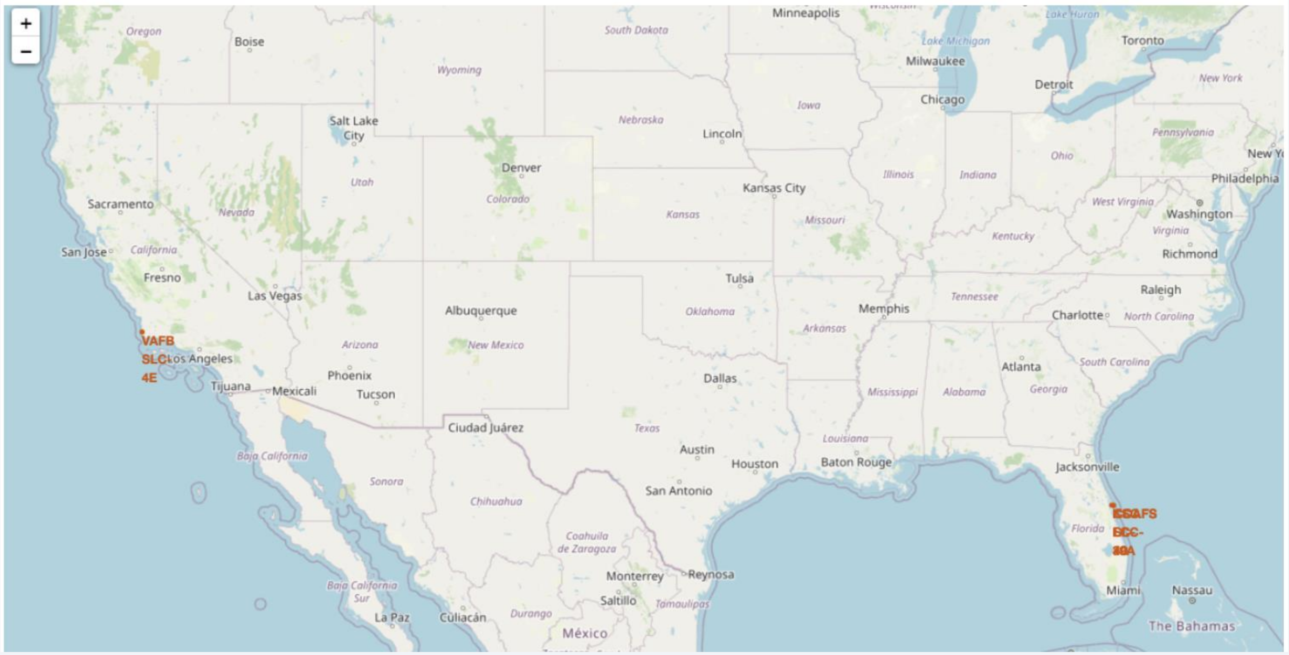
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

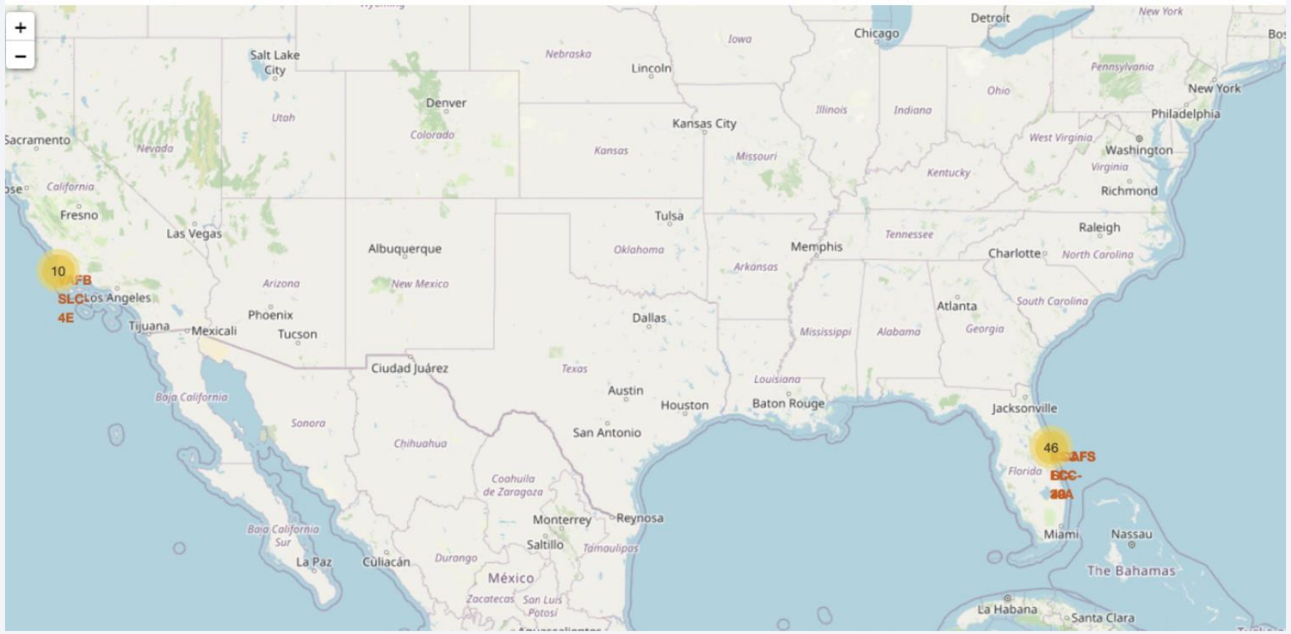
Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites with Location Markers



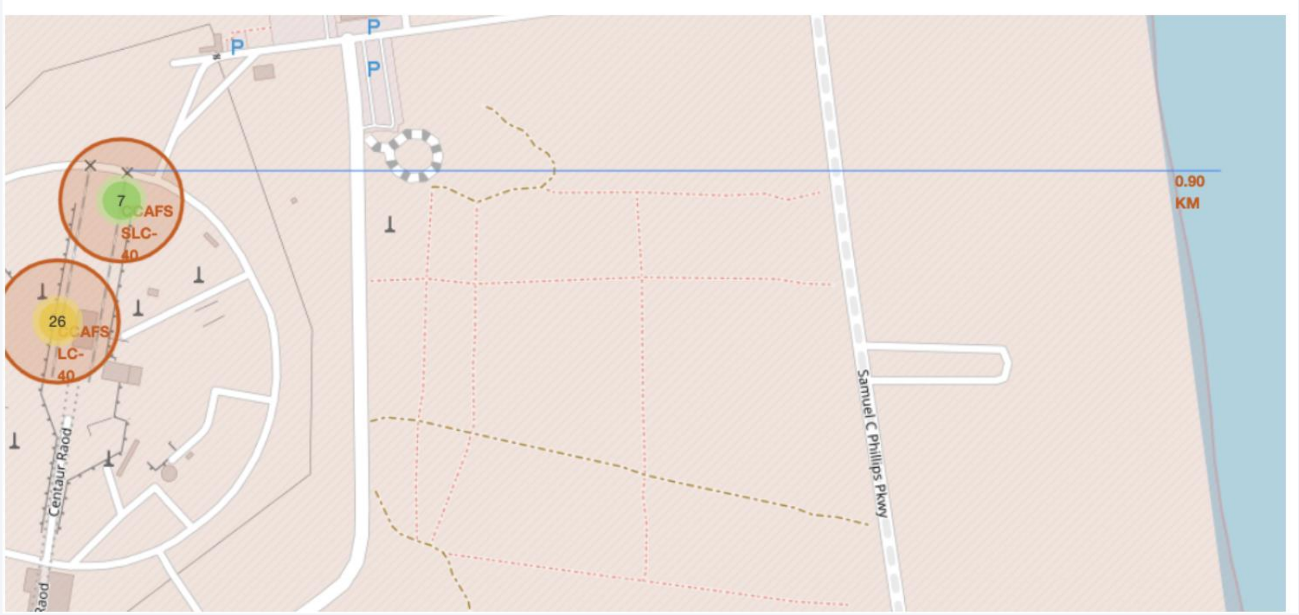
- Not all launch sites are in close proximity to the Equator. For example, the "VAFB SLC 4E" site is located in California, USA, far from the Equator. The other sites, such as "CCAFS SLC-40" and "KSC LC-39A," are in Florida, which is closer to the Equator but not directly on it.
- all the launch sites are in very close proximity to the coast. This is likely due to safety and logistical reasons, allowing rockets to be launched over the ocean, minimizing the risk to populated areas in case of an accident during launch.

Launch Sites with Color-Coded Launch Outcomes

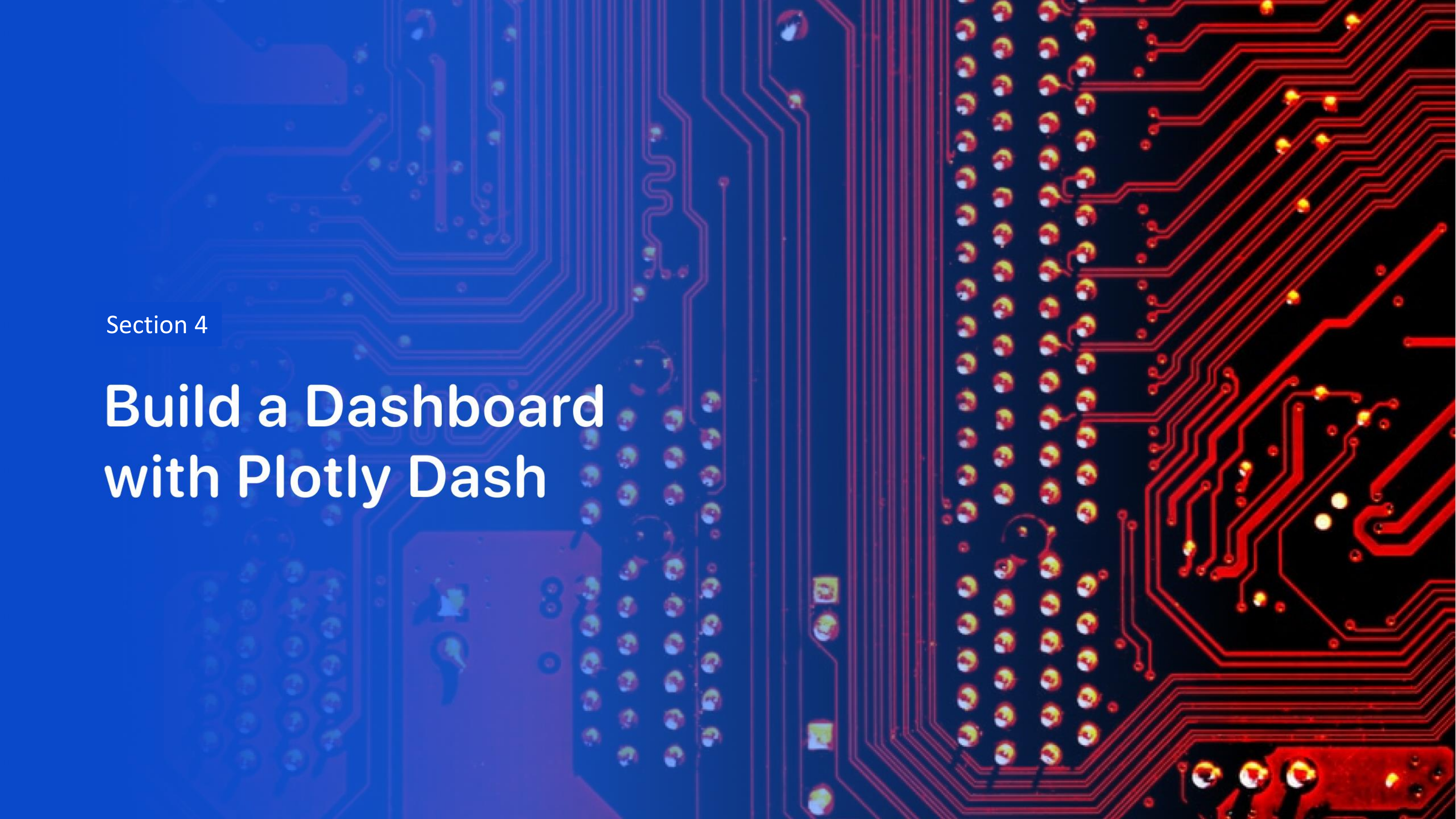


- Most launches occur from the sites in Florida ("CCAFS SLC-40" and "KSC LC-39A"), with the majority being successful, indicated by a greater number of orange markers.
- The "VAFB SLC 4E" site in California also shows a few launches, with a mixed outcome of successes and failures.
- The clustering of markers helps visualize the frequency and outcomes of launches from specific sites, showing that Florida sites have the highest launch activity and success rate.

Launch Site Proximity Analysis: CCAFS SLC-40 to Key Infrastructure



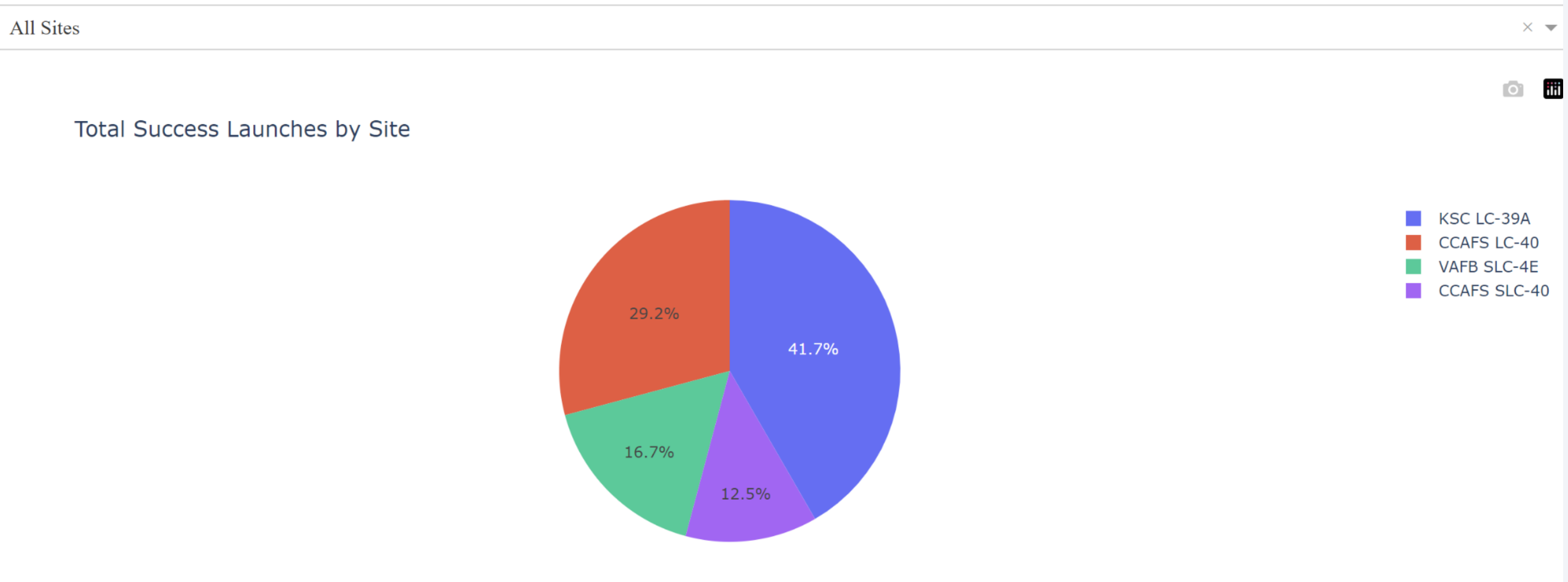
- The close distance to the coast supports the strategic choice for rocket launches, offering a clear trajectory over the ocean.
- The site's location near roadways facilitates efficient transportation and logistics operations, crucial for launch preparations and post-launch activities.
- The visual and quantitative representation of proximities provides insight into the strategic placement of launch sites in relation to infrastructure and safety considerations.



Section 4

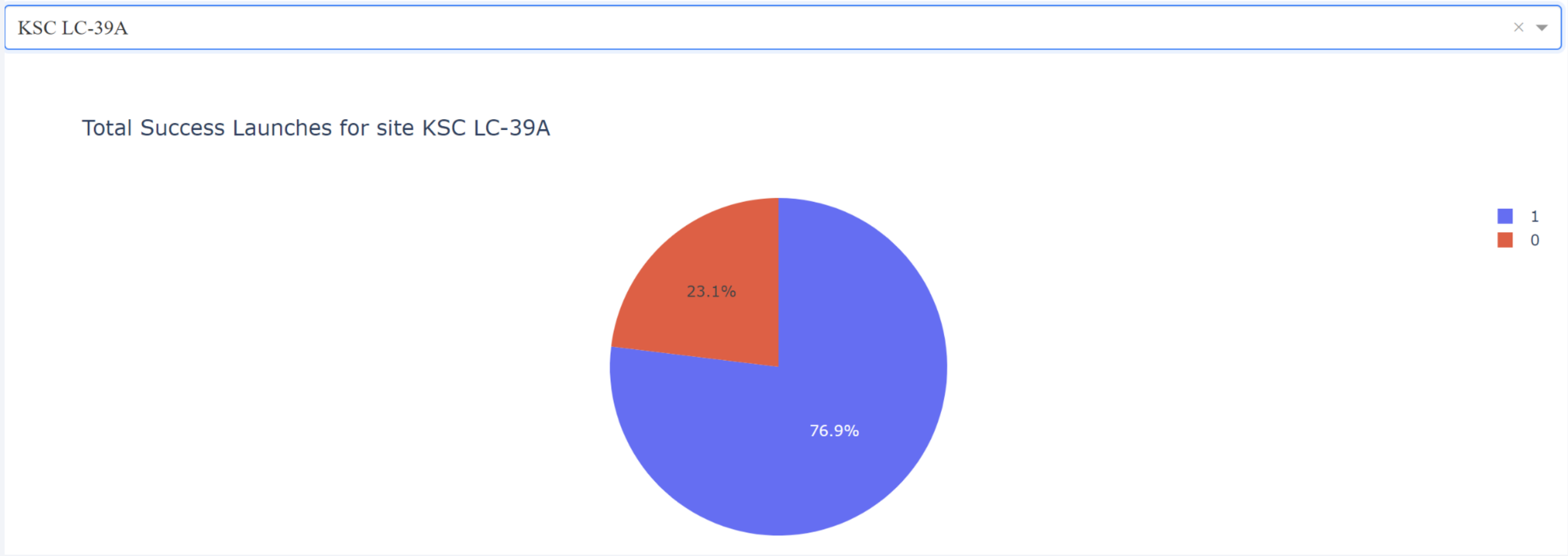
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- KSC LC-39A has the most significant share of successful launches, suggesting it might be a preferred site for frequent or complex missions.
- VAFB SLC-4E and CCAFS LC-40 also contribute considerably, indicating their importance in launch operations.

Launch Success Distribution for KSC LC-39A

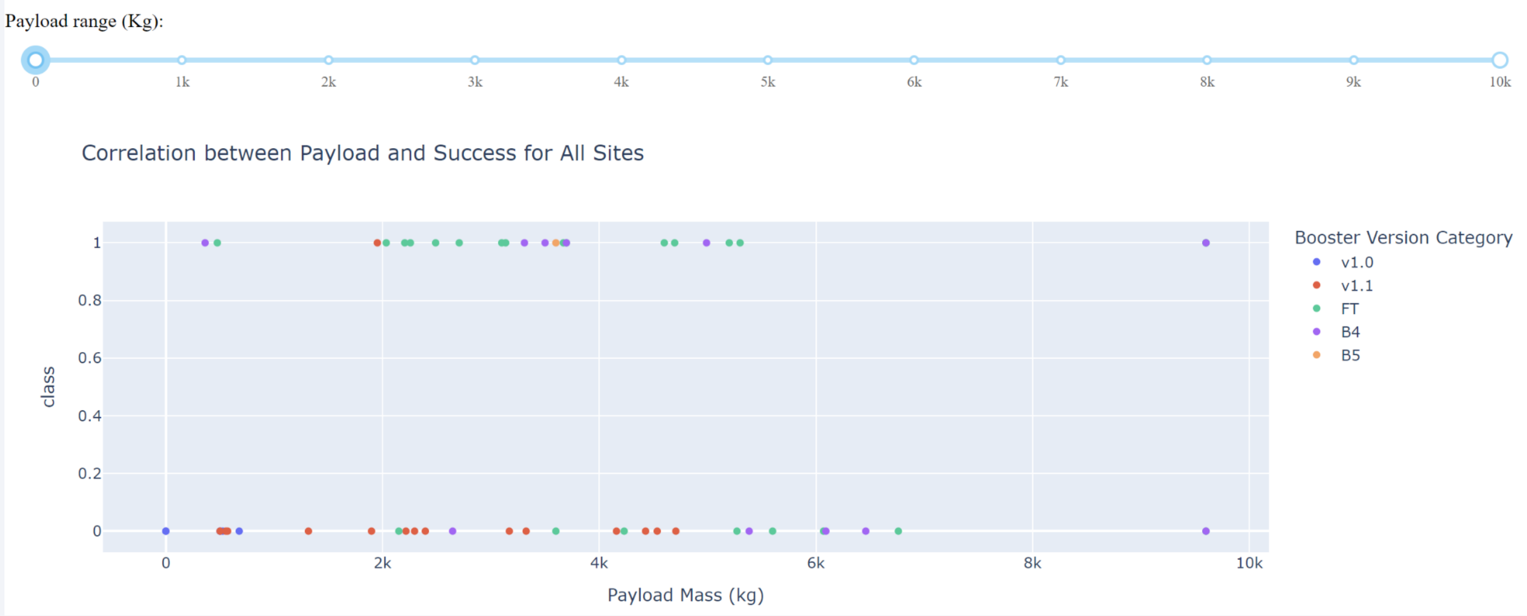


KSC LC-39A × ▾

Total Success Launches for site KSC LC-39A

- The site KSC LC-39A has a high launch success ratio, with 76.9% of its launches being successful.
- The failure rate is comparatively lower at 23.1%, indicating strong performance metrics for this launch site.

Correlation between Payload Mass and Launch Success for All Sites

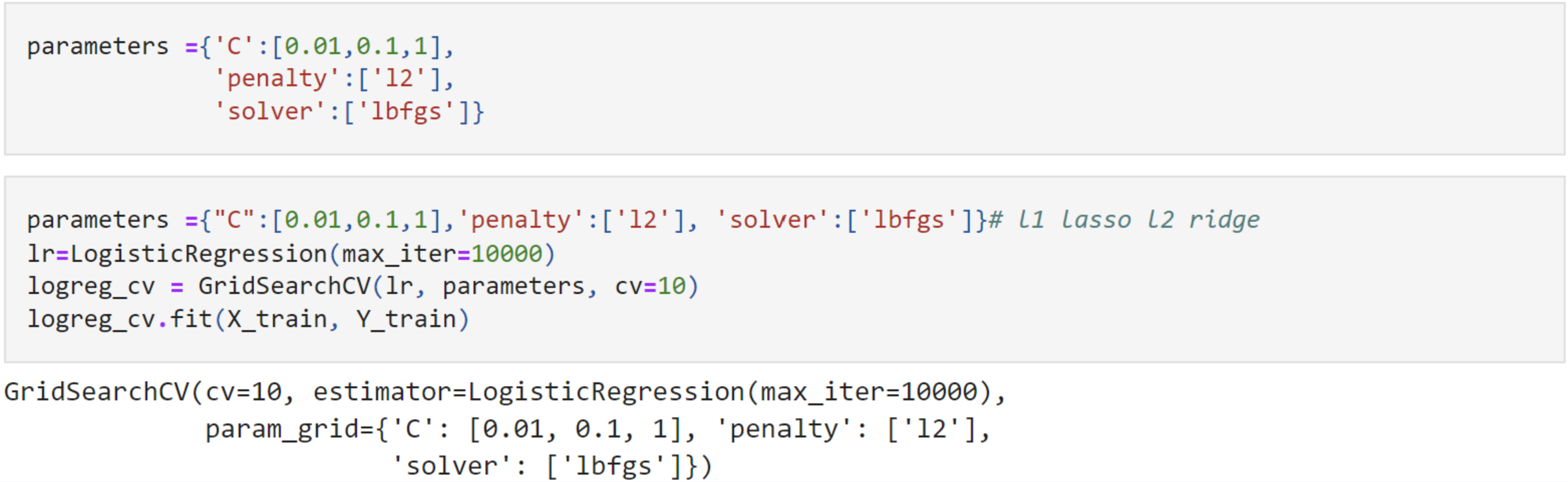


- Most launches with payloads below 4,000 kg have high success rates, particularly for booster versions B4, FT, and B5.
- For payloads over 4,000 kg, the success rate becomes more variable, but certain booster versions like FT and B5 show consistent success.
- The plot helps identify which payload ranges and booster versions are more likely to result in successful launches, aiding in strategic planning for future missions.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

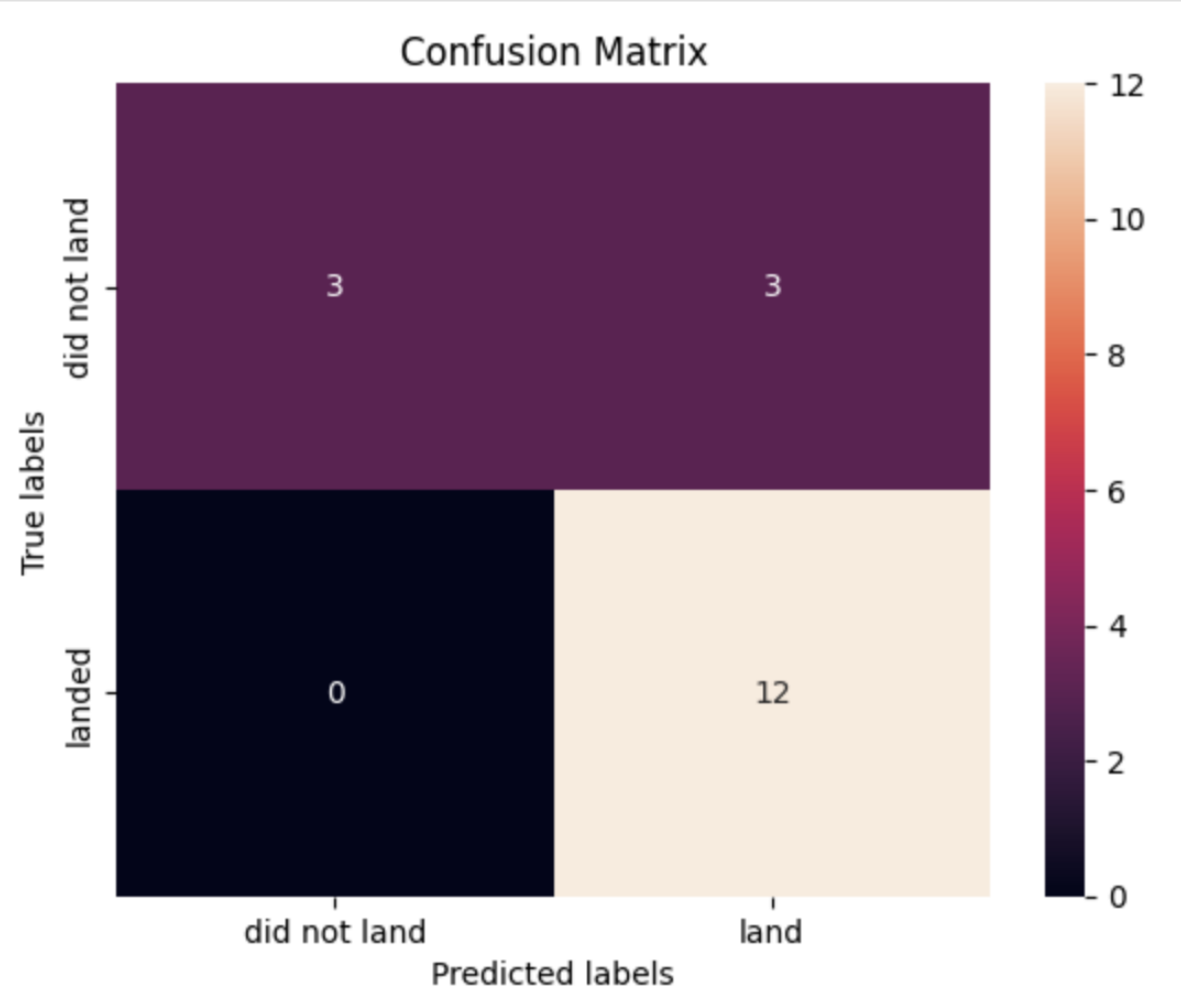


```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression(max_iter=10000)
logreg_cv = GridSearchCV(lr, parameters, cv=10)
logreg_cv.fit(X_train, Y_train)
```

```
GridSearchCV(cv=10, estimator=LogisticRegression(max_iter=10000),
             param_grid={'C': [0.01, 0.1, 1], 'penalty': ['l2'],
                         'solver': ['lbfgs']})
```

Confusion Matrix



- The model successfully identified 12 out of 12 instances where the rocket landed.
- No cases were misclassified where the model predicted a rocket would not land, but it actually did land.
- The model demonstrates strong performance in predicting successful landings with no false negatives, but there are some false positives that need attention. The high true positive rate indicates the model's effectiveness in identifying successful landings.

Conclusions

- Key Findings:
  - Higher success rates at specific launch sites (e.g., KSC LC-39A).
  - Success more likely with specific payload ranges and booster versions.
  - Proximity to the coast associated with higher success rates.
- Recommendations:
  - Optimize future launches using successful booster versions (e.g., FT, B5).
  - Focus on payload ranges under 4,000 kg.
  - Prioritize launch sites with higher success rates (e.g., KSC LC-39A).
- Overall Conclusion:
  - The combined approach of data collection, visualization, and predictive modeling effectively identifies key factors for launch success.
  - Insights gained can guide future SpaceX launch operations for improved success rates.

Appendix

- url for github:
- https://github.com/Tsuuunade/Applied-Data-Science-Capstone.git

- You can see all figures and code here.

Thank you!