IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonathan

10/01/2026

Outline

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

Executive Summary

Summary of Methodologies

Historical SpaceX launch data was collected from the SpaceX REST API and Wikipedia. The data was cleaned, processed, and explored using exploratory data analysis techniques and data visualization. Key features such as launch site, payload mass, orbit type, and mission outcome were analyzed. Multiple machine learning models, including Logistic Regression, Decision Trees, and Support Vector Machines, were trained and evaluated using cross-validation to predict first-stage booster landing success.

Summary of Results

The analysis showed that payload mass, orbit type, and launch site have a significant impact on booster landing success. Among the tested models, the optimized Logistic Regression model achieved the best overall performance, with high predictive accuracy on the test dataset. These results demonstrate that machine learning can effectively predict booster landing outcomes and support cost estimation and mission planning.

Introduction

Project Background and Context

SpaceX has significantly reduced the cost of space missions by successfully reusing Falcon 9 rocket boosters. Predicting whether the first-stage booster will land successfully is crucial for estimating launch costs and improving mission planning. This project uses historical SpaceX launch data to analyze key factors influencing booster landing outcomes and to build predictive models.

Problems to Find Answers

- What factors most strongly influence the success of Falcon 9 first-stage landings?
- How do launch site, payload mass, orbit type, and mission parameters affect launch outcomes?
- Can machine learning models accurately predict whether a booster will land successfully?

Section 1

Methodology

Methodology

Executive Summary

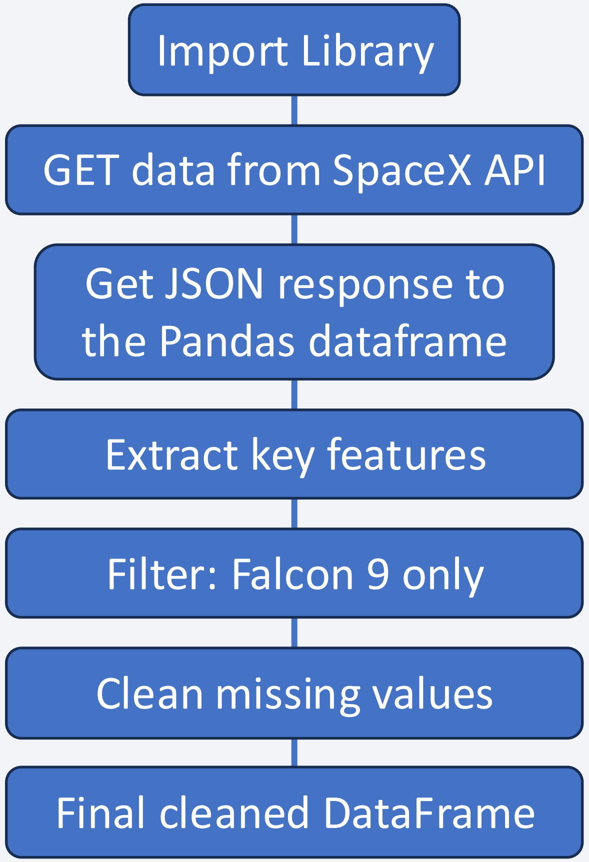
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data were collected from IBM Cloud storage using python retrieving the data from a csv file usiing pandas
- The data collection process involved obtaining launch data from the **SpaceX REST API** and supplementing it with historical information scraped from **Wikipedia**. The API provided structured details such as flight number, launch site, payload mass, orbit type, and mission outcome, while web scraping ensured completeness of earlier records. The raw data was then cleaned by handling missing values, standardizing formats, and encoding launch outcomes into binary variables. Finally, the datasets were merged into a single structured **Pandas DataFrame**, preparing the data for exploratory analysis and machine learning modeling.

Data Collection – SpaceX API

- **Import Libraries**
- *“Load Python libraries for data handling and HTTP requests”*
- **Retrieve Data via REST API**
- *“Send GET request to SpaceX API endpoint for past launches”*
- **Convert JSON to DataFrame**
- *“Flatten JSON response into a structured table using pandas.json_normalize()”*
- **Extract Key Features**
- *“Select relevant launch attributes: flight number, rocket type, payload mass, orbit, launch site, date, etc.”*
- **Filter to Falcon 9 Launches**
- *“Focus analysis on Falcon 9 by filtering the dataset”*
- **Clean Missing Values**
- *“Handle NaNs in PayloadMass by replacing them with the column mean”*
- **Final Cleaned Dataset**
- *“Ready for exploratory data analysis and further modeling”*

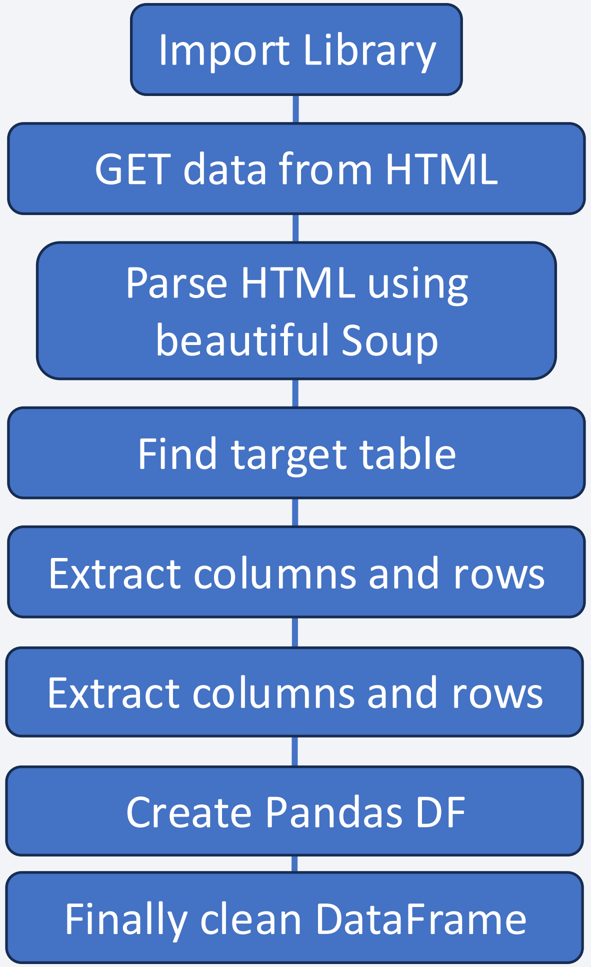


Github link : [Github_Link]

Data Collection - Scraping

- We start by **importing the necessary Python libraries**, including requests to fetch web pages, BeautifulSoup to parse HTML, and pandas to structure the data.
- Next, we **define the target URL** containing the data, then **send an HTTP GET request** to download the HTML content of the page.
- Using **BeautifulSoup**, we **parse the HTML** to locate the target table. We then **extract rows and columns**, storing the values in Python lists.
- Finally, we **convert these lists into a pandas DataFrame**, **clean and format the data** by renaming columns and fixing types, and optionally **export the dataset to a CSV** for further analysis.
- This workflow ensures that raw HTML from the web is transformed into a clean, structured dataset ready for exploratory data analysis and modeling.

Github link : Github_Link



Data Wrangling

- In the Data Wrangling phase, we refine our SpaceX dataset to make it **analysis-ready**:
- We start with the combined launch data collected from the SpaceX REST API and web scraping, then **clean and transform key fields**.
- We calculate statistics such as the **number of launches per site** and the **occurrence of each orbit type and mission outcome**, using methods like .value_counts() to summarize categorical patterns in the data.
- We create a **binary landing outcome label**, where a successful first-stage landing is coded as 1 and unsuccessful as 0, transforming raw text outcomes into a format suitable for machine learning.
- Additional transformations include handling missing values, filtering or recoding variables, and preparing the dataset for subsequent exploratory analysis and modeling.
- The result is a cleaned, consistently labeled dataset that supports deeper analysis and prediction tasks in the later stages of the project.

Github link : Github_Link

EDA with Data Visualization

- **Exploratory Data Analysis & Visualization (Slide)**
- In the EDA notebook, we plotted a series of visualizations to better understand relationships in the SpaceX launch dataset:
- **Scatter plots** to explore correlations such as *Flight Number vs. Payload Mass* and *Payload Mass vs. Orbit* to identify trends and patterns between variables.
- **Bar charts** to compare categorical distributions like *Launch Site vs. Success Rate* and *Orbit vs. Success* to see which categories had higher outcomes.
- **Line charts** to show trends over time, for example how success rates or payload characteristics evolve across years.
- These chart types were chosen to **reveal relationships, compare categories, and visualize trends** to support further analysis and model decisions.

GitHub URL : Github link

EDA with SQL

- **SQL Queries Summary (Slide)**
- Queried **all unique launch site names** from the SpaceX dataset.
- Retrieved launch site records beginning with specific text patterns.
- Calculated the **total payload mass** carried by boosters (e.g., NASA missions).
- Found the **average payload mass** for a specific booster version (e.g., F9 v1.1).
- Identified the **earliest date** of a successful first-stage landing on a ground pad.
- Counted the **total number of successful and failed mission outcomes**.
- Listed **booster versions with successful drone ship landings** and payload mass within a range.
- Ranked records by landing outcomes and examined temporal patterns (e.g., failures vs. successes over time).
- *These queries help extract insights on operational trends, payload performance, and mission outcomes from the SpaceX database.*
- GitHub URL : [Github link]()

Build an Interactive Map with Folium

- **Markers & Circles:**
  We added **circle markers** at each SpaceX launch site using their latitude and longitude to visually indicate the **location of each launch facility** on the map. Circles and text labels make the map easier to interpret geographically.
- **Marker Clusters:**
  We used **MarkerCluster** with colored markers to show **launch outcomes at each site**. This helps quickly identify which sites have higher success rates.
- **Distance Lines (PolyLines):**
  We drew **lines between launch sites and nearby landmarks** to visualize **proximities and spatial relationships**, aiding analysis of geographic influences on launch operations.
- **Distance Markers:**
  We placed **text markers** at these nearby points to display **distance values** (in kilometers) directly on the map, making spatial interpretation intuitive.
- **Purpose:** These objects turn raw coordinates into an **interactive spatial visualization** that reveals launch site locations, success patterns, and geographic proximities enhancing insight beyond static data tables.

- GitHub URL : Github link

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

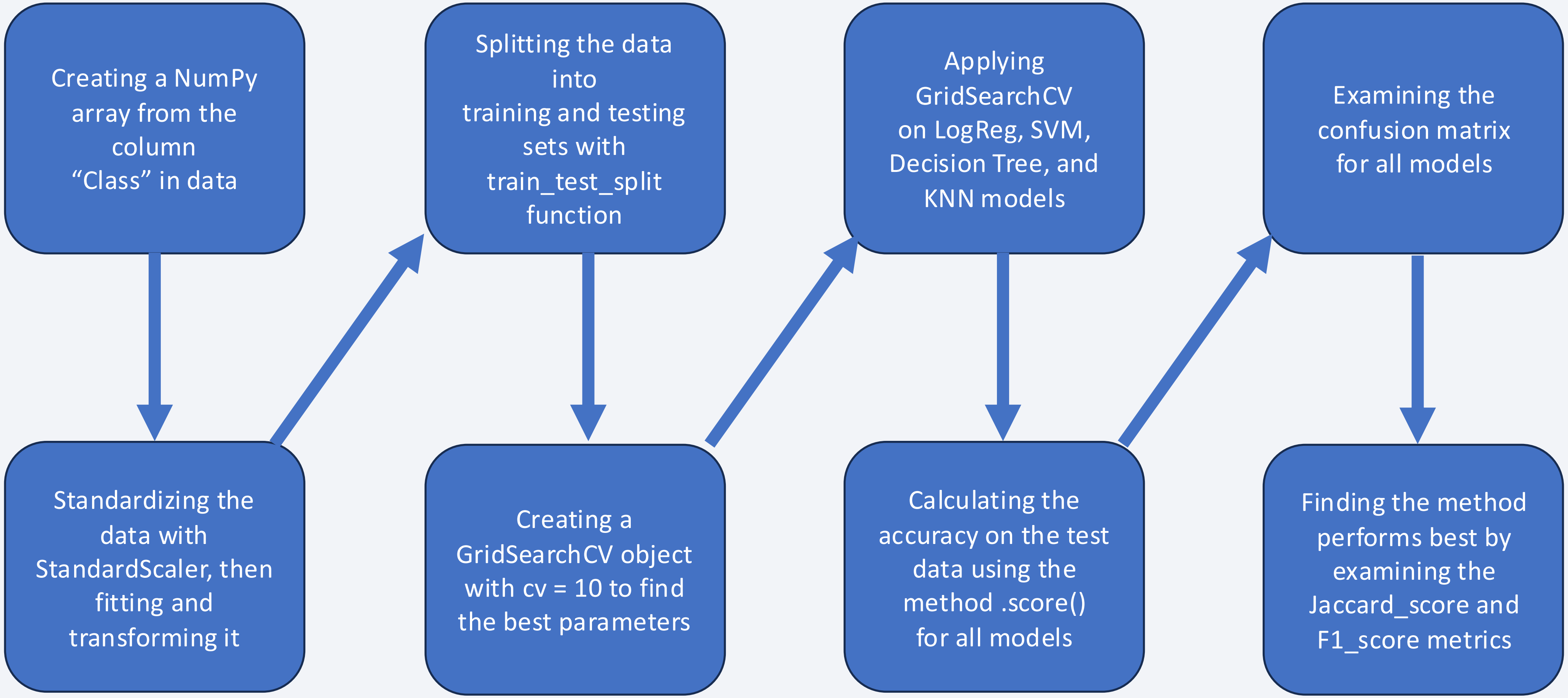
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL : Github link

Predictive Analysis (Classification)



Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

GitHub URL : Github link

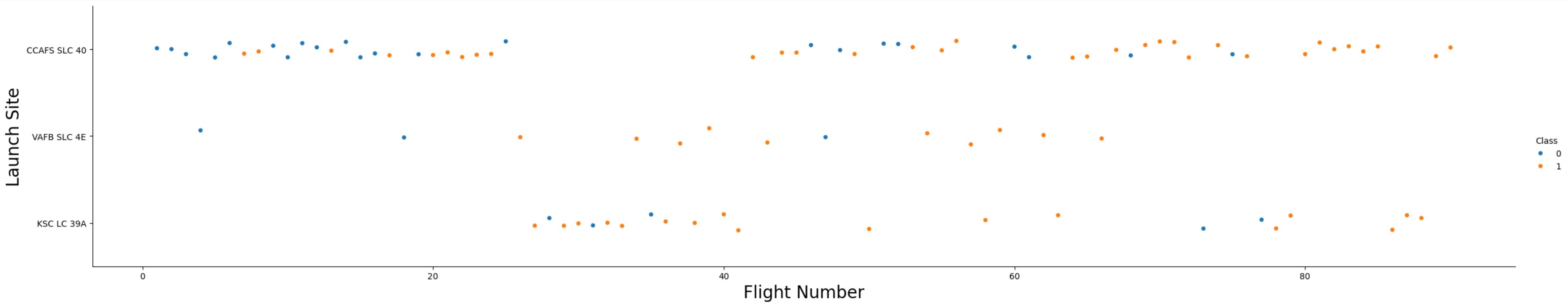
Results

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

Section 2

Insights drawn from EDA

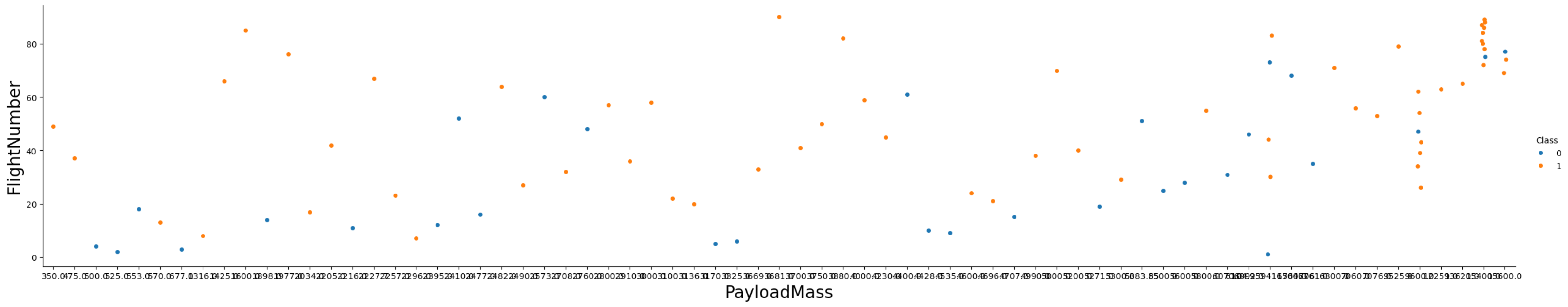
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

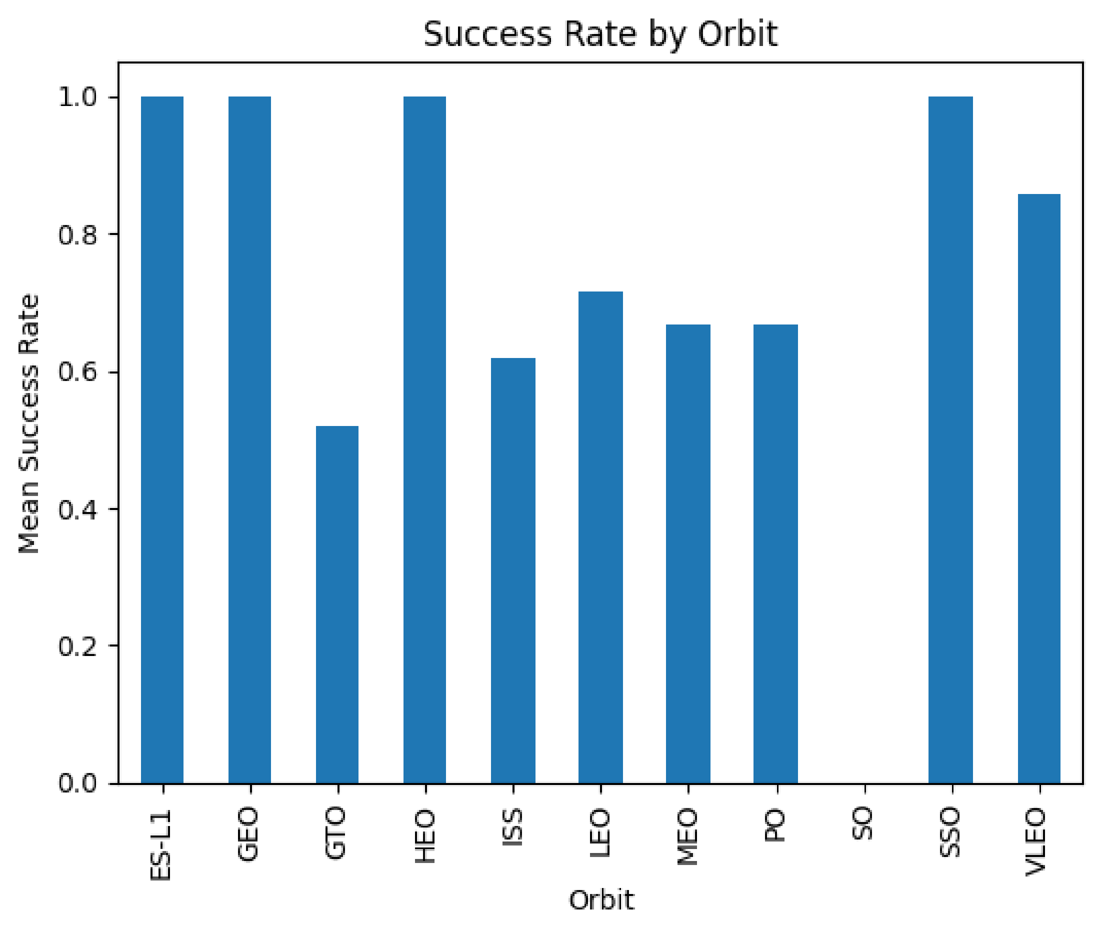
Success Rate vs. Orbit Type

Explanation:

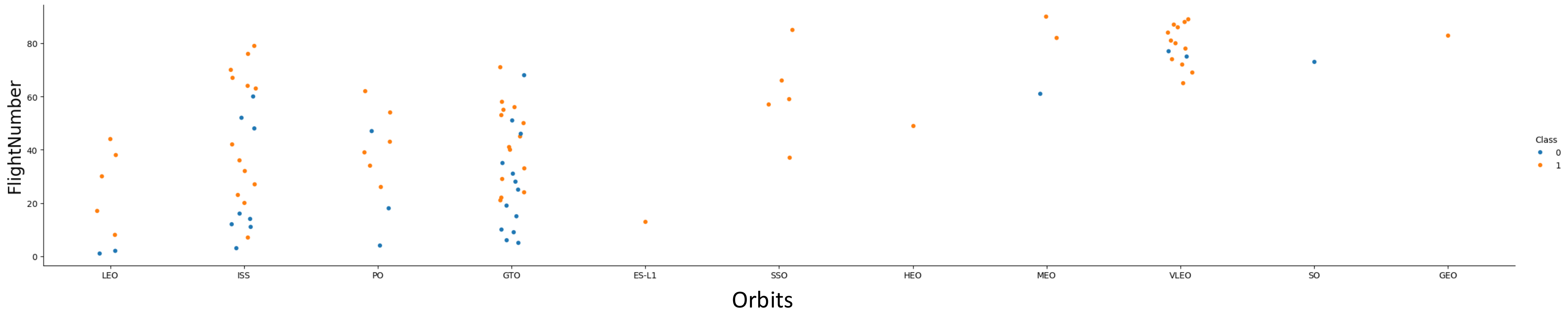
• Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO

• Orbits with 0% success rate:
- SO

• Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO



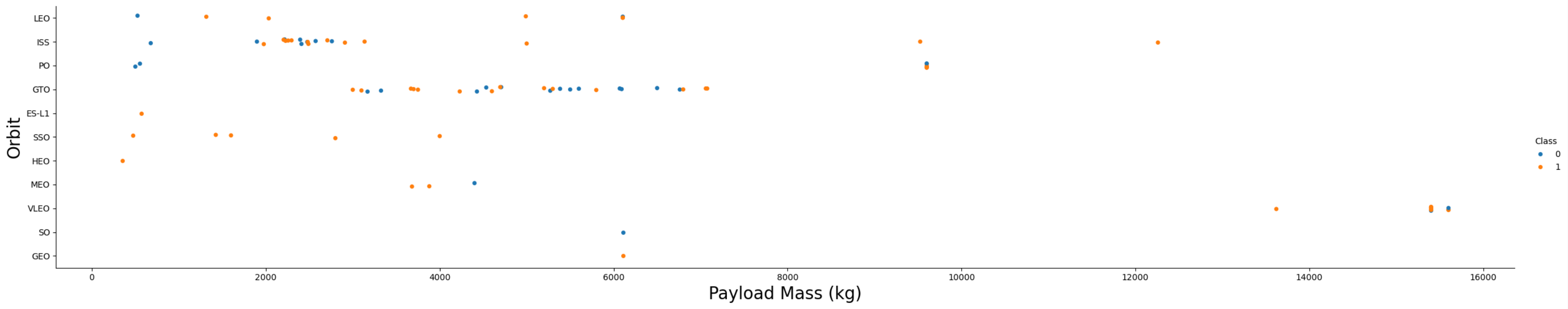
Flight Number vs. Orbit Type



Explanations :

- In the LEO orbit the Success appears related to the number of flights;
- on the other hand, there seems to be no relationship between flight
- number when in GTO orbit.

Payload vs. Orbit Type



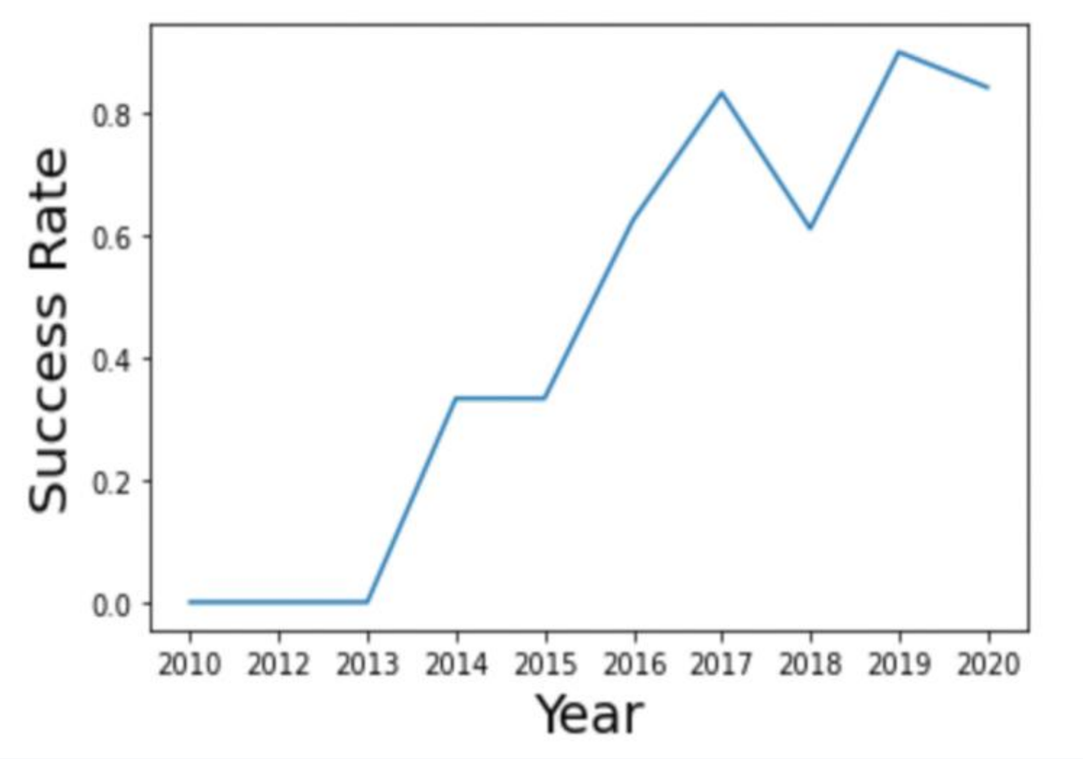
Explanation :

The graph illustrates a strong positive correlation between orbit number and payload mass.
As the orbit number increases from 8 to 20, there is a clear corresponding increase in the available payload capacity. This trend suggests that higher orbit assignments are associated with the ability to carry significantly more mass. Therefore, mission planning can leverage this relationship to optimize payload delivery based on orbital requirements.

Launch Success Yearly Trend

Explanations :

- This graph shows an overall positive trendline with an increase of the success rate launch over the year starting from 2013 until 2020 with a small drop in 2018 that however recover in 2019

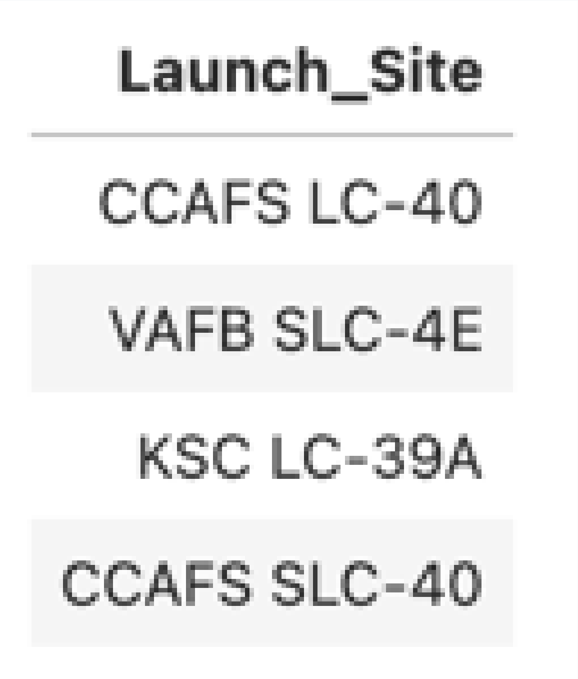


All Launch Site Names

All the unique launch site have been retrieved using the command :

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

Therefore, displaying the names of the unique launch sites in the space mission.



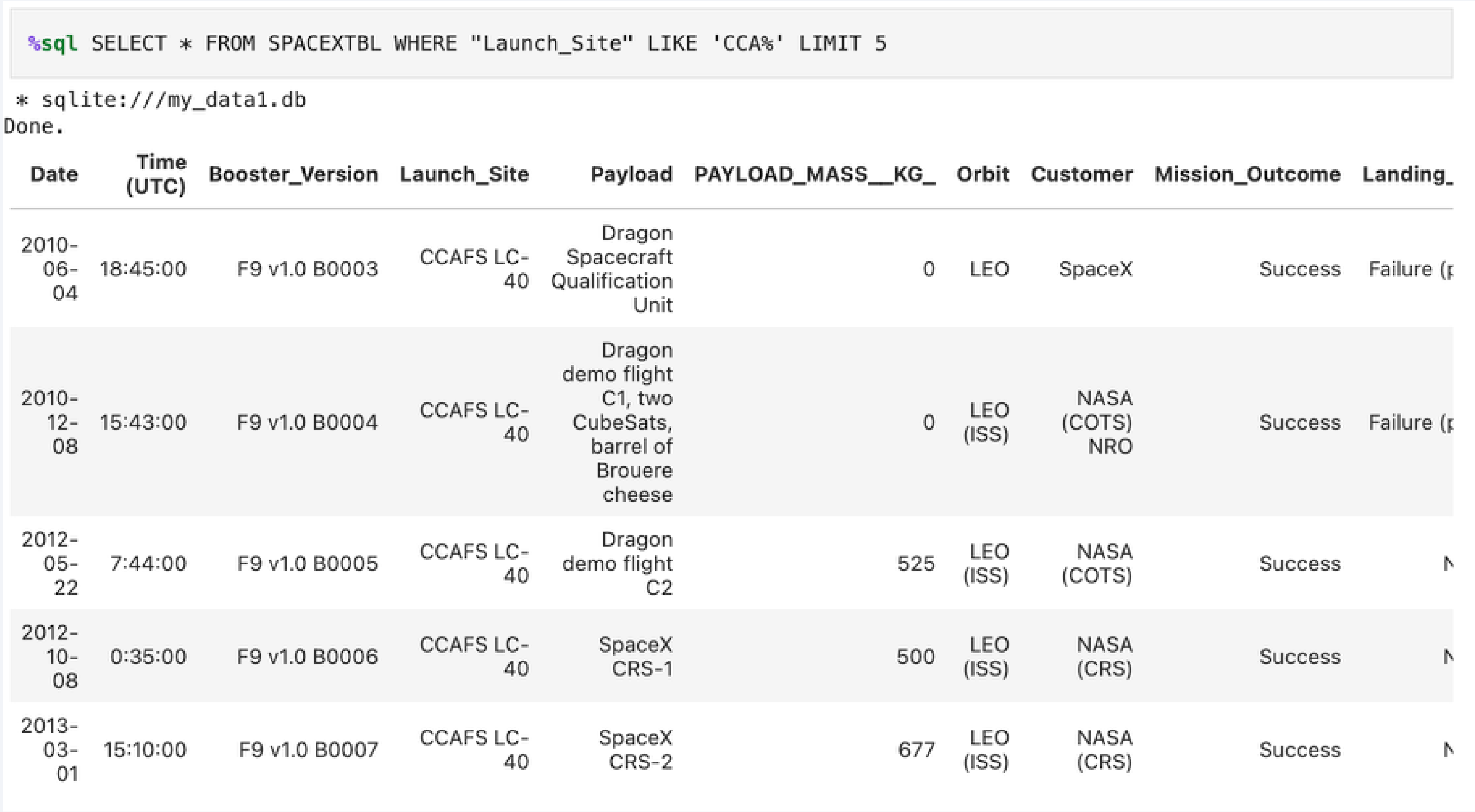
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Using the sql command :

%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5

We have successfully retrieved the 5 launch-site begin with 'CCA'



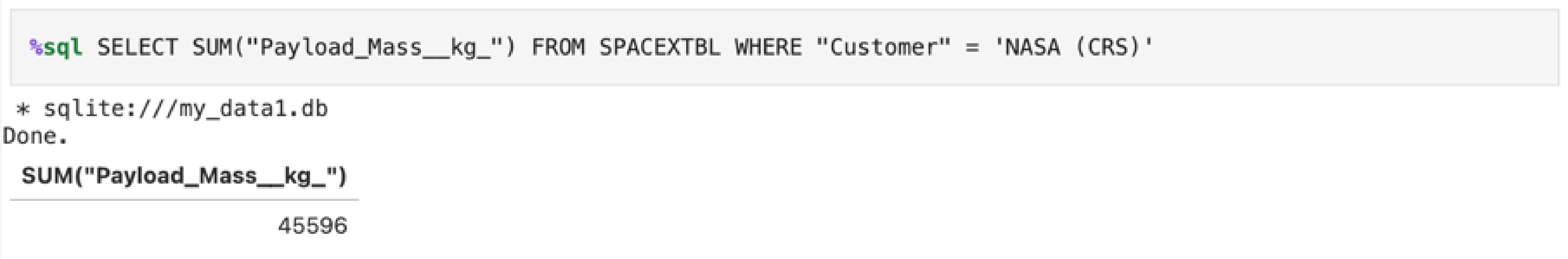
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

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

Total Payload Mass

The total payload mass for NASA was calculated using the SQL query shown below :



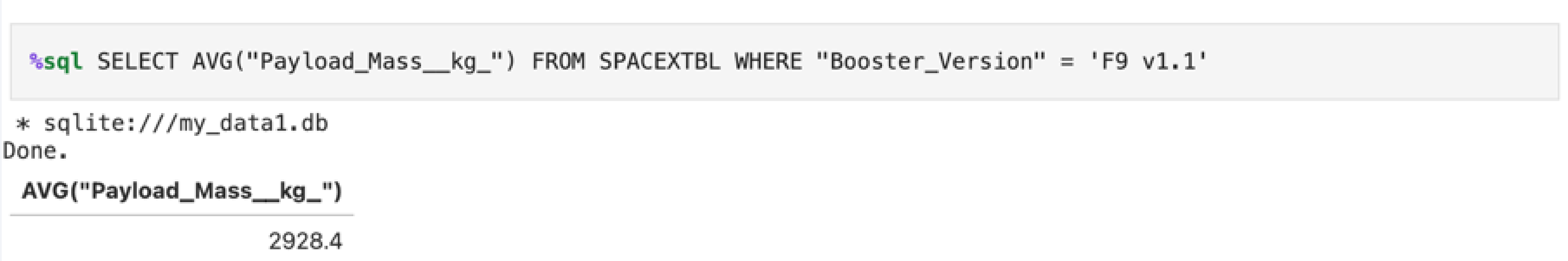
```
%sql SELECT SUM("Payload_Mass__kg_") FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
 SUM("Payload_Mass__kg_")
--------------------------
                    45596
```

The query filtered the data to include only missions where NASA was the customer. Finally, the sum of the payload mass was computed to produce the final result.

Average Payload Mass by F9 v1.1

The average payload mass for booster wersion F9 v1.1 was calculated using the SQL query shown below :



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

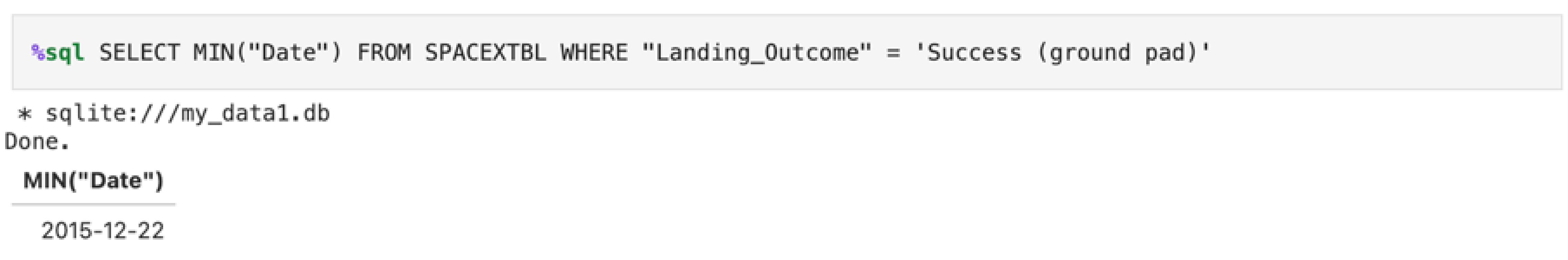
```
 * sqlite:///my_data1.db
Done.
```

AVG("Payload_Mass__kg_")
2928.4

The query filtered the data to include only missions where the booster version was the F9 v1.1. Finally, the average of the payload mass was computed to produce the final result.

First Successful Ground Landing Date

The first successful landing ground was found using the SQL query shown below :



```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)'
```

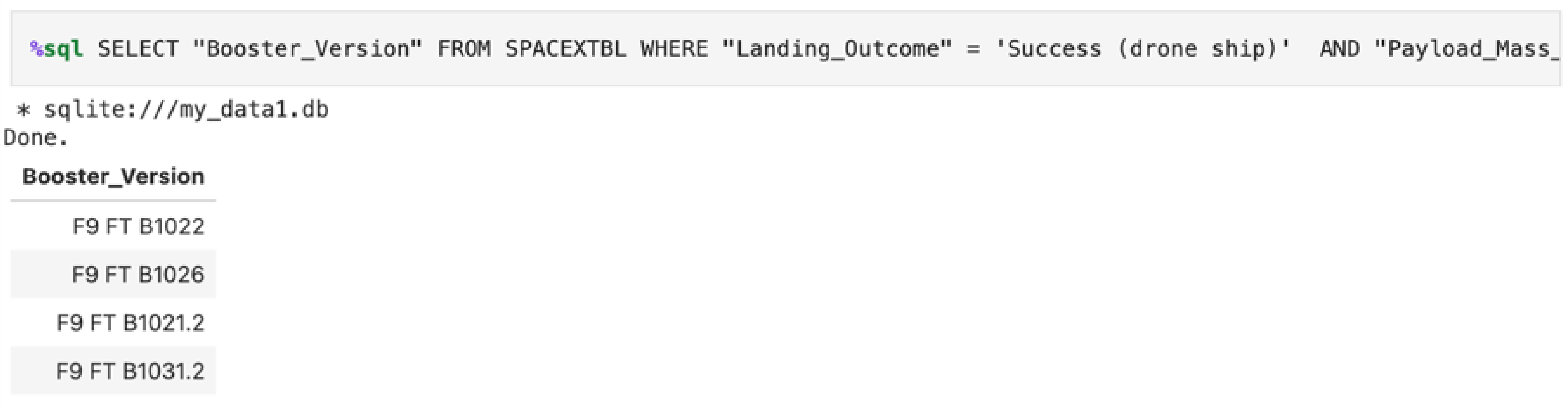
```
 * sqlite:///my_data1.db
Done.
```

MIN("Date")
2015-12-22

The query filtered the data to retrieve the first missions based on the date and filtered with the criteria of a Success ground pad to produce the result.

Successful Drone Ship Landing with Payload between 4000 and 6000

The first successful drone ship landing was found using the SQL query shown below :



```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)'  AND "Payload_Mass_
```

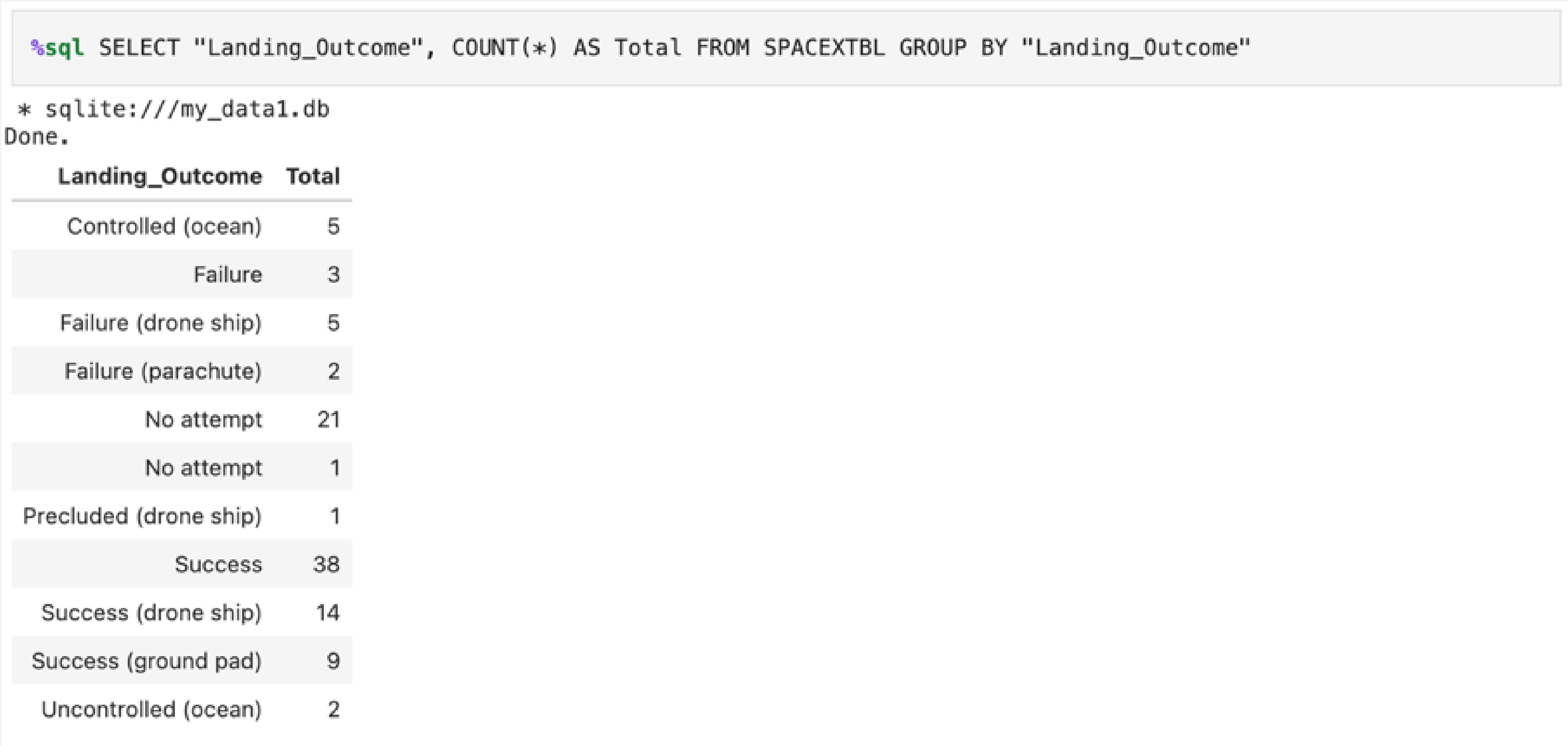
```
 * sqlite:///my_data1.db
Done.
```

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

The query filtered the data to retrieve the first missions based on the date and filtered with the criteria of a Success ground pad to produce the result.

Total Number of Successful and Failure Mission Outcomes

The number of successful and failure mission outcome was retrieved using the SQL query shown below :



```
%sql SELECT "Landing_Outcome", COUNT(*) AS Total FROM SPACEXTBL GROUP BY "Landing_Outcome"
```

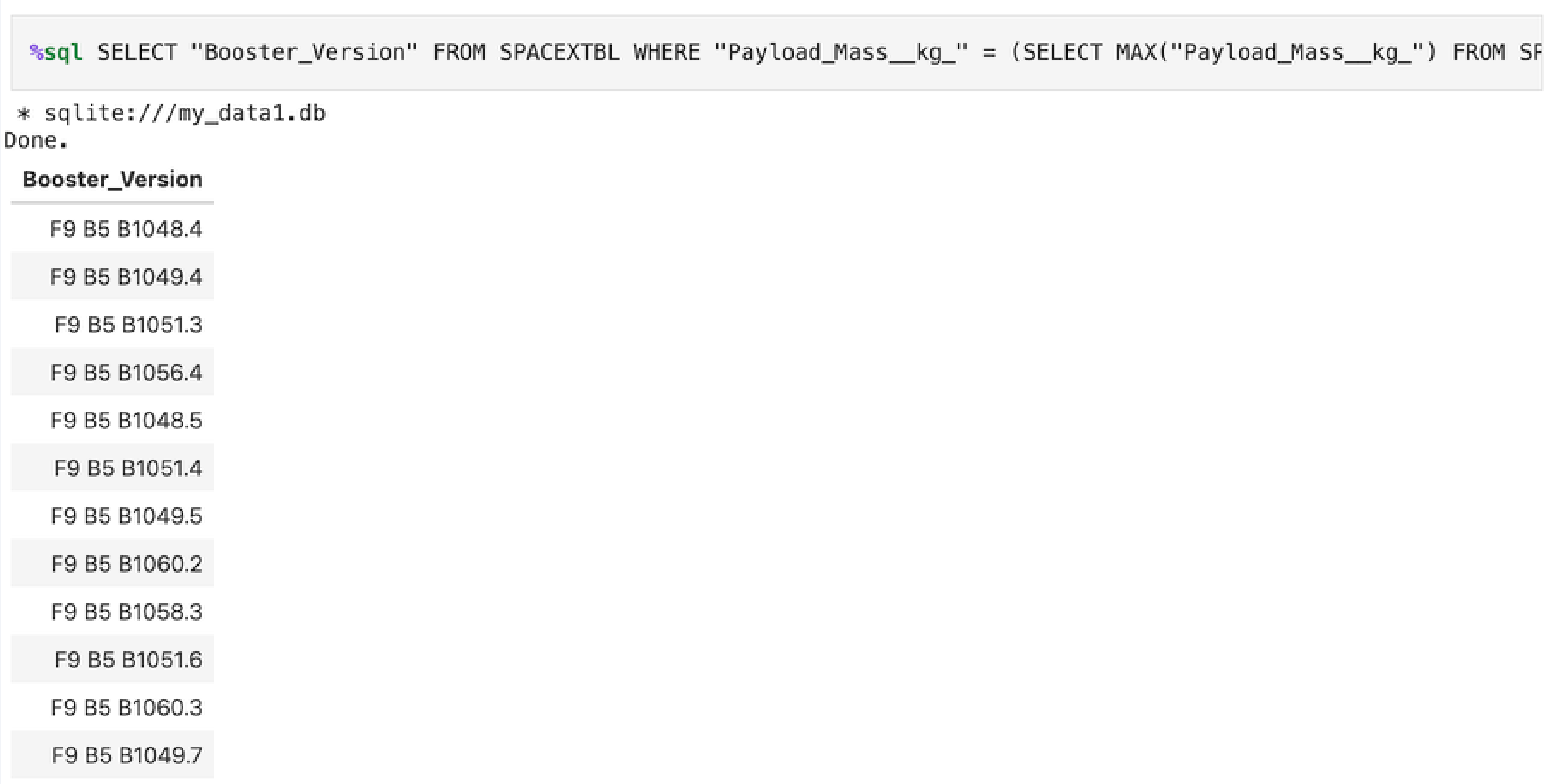
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Total
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

The query filtered the data to count the landing outcome and group them to produce the result.

Boosters Carried Maximum Payload

The Boosters version that carried the maximum load were retrieved using the SQL query shown below :



```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SP
```

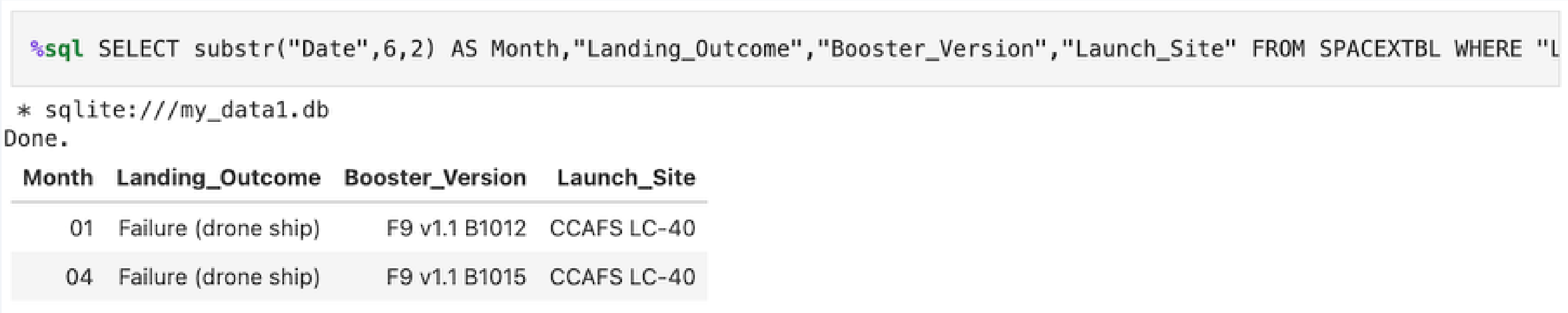
```
 * sqlite:///my_data1.db
Done.
```

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

The query filtered the data to select the booster than had the maximum number as a payload from the database to produce the result.

2015 Launch Records

The SQL query bellow listed the records which displayed the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.



```
%sql SELECT substr("Date",6,2) AS Month,"Landing_Outcome","Booster_Version","Launch_Site" FROM SPACEXTBL WHERE "L
```

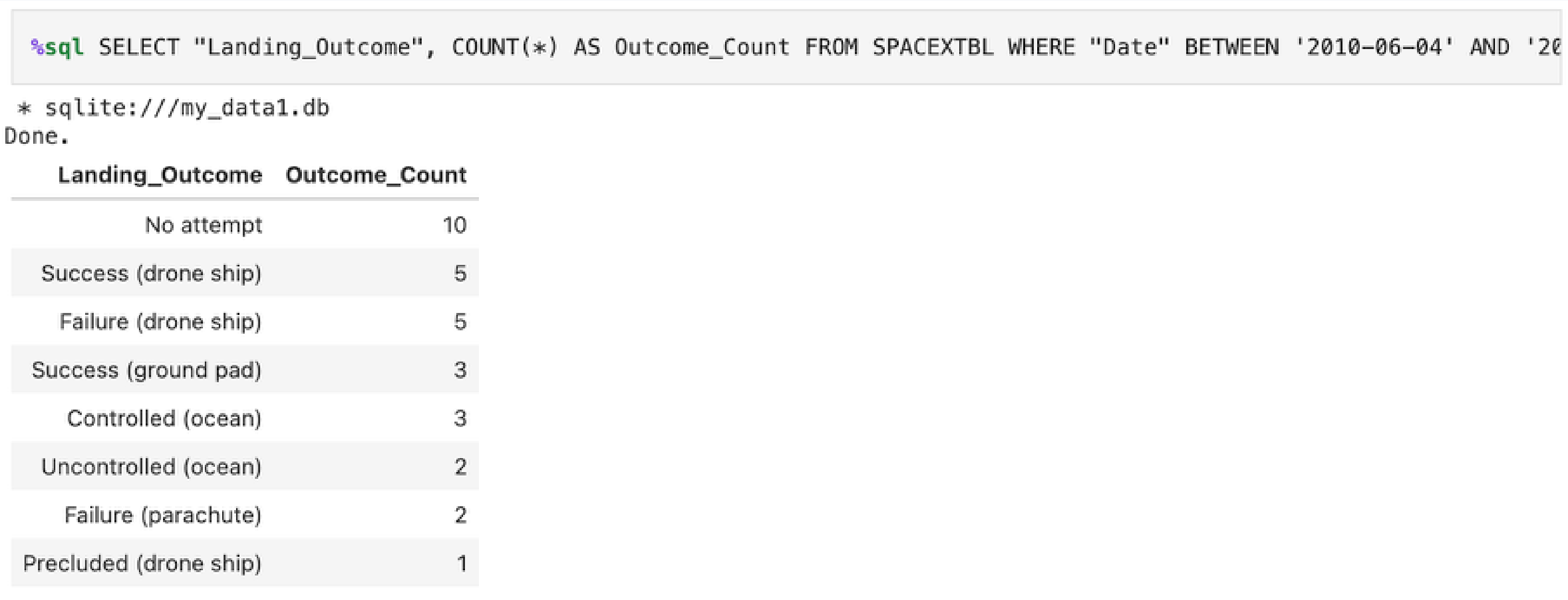
```
 * sqlite:///my_data1.db
Done.
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The query filtered the data to retrieve the first missions based on the date and filtered with the criteria of a Success ground pad to produce the result.

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

The first successful landing ground was found using the SQL query shown below :



```
%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTBL WHERE "Date" BETWEEN '2010-06-04' AND '20
```

```
* 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

This table shows the Ranking count of the landing outcomes such as Failure or Success… between the date 2010-06-04 and 2017-03-20 in descending order.

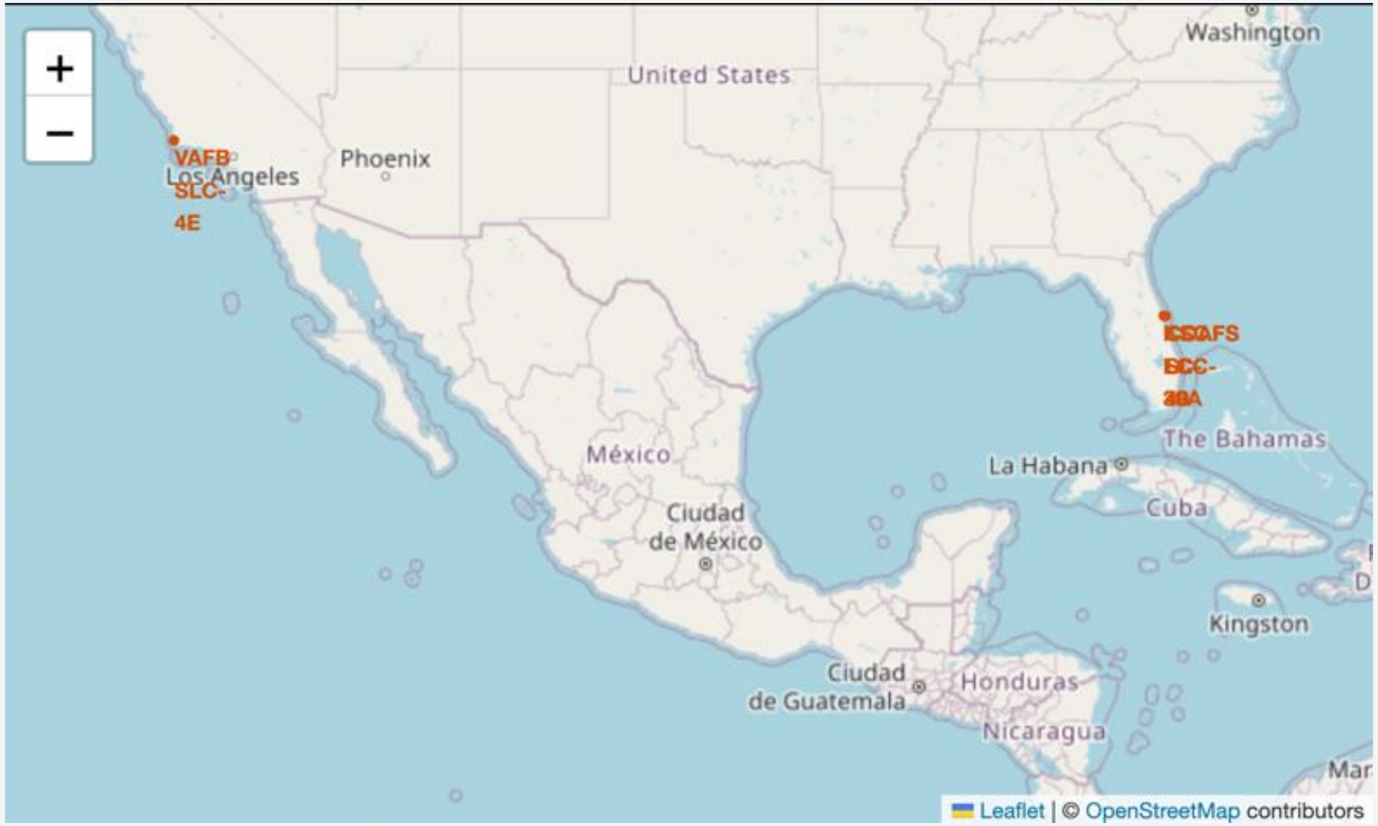
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

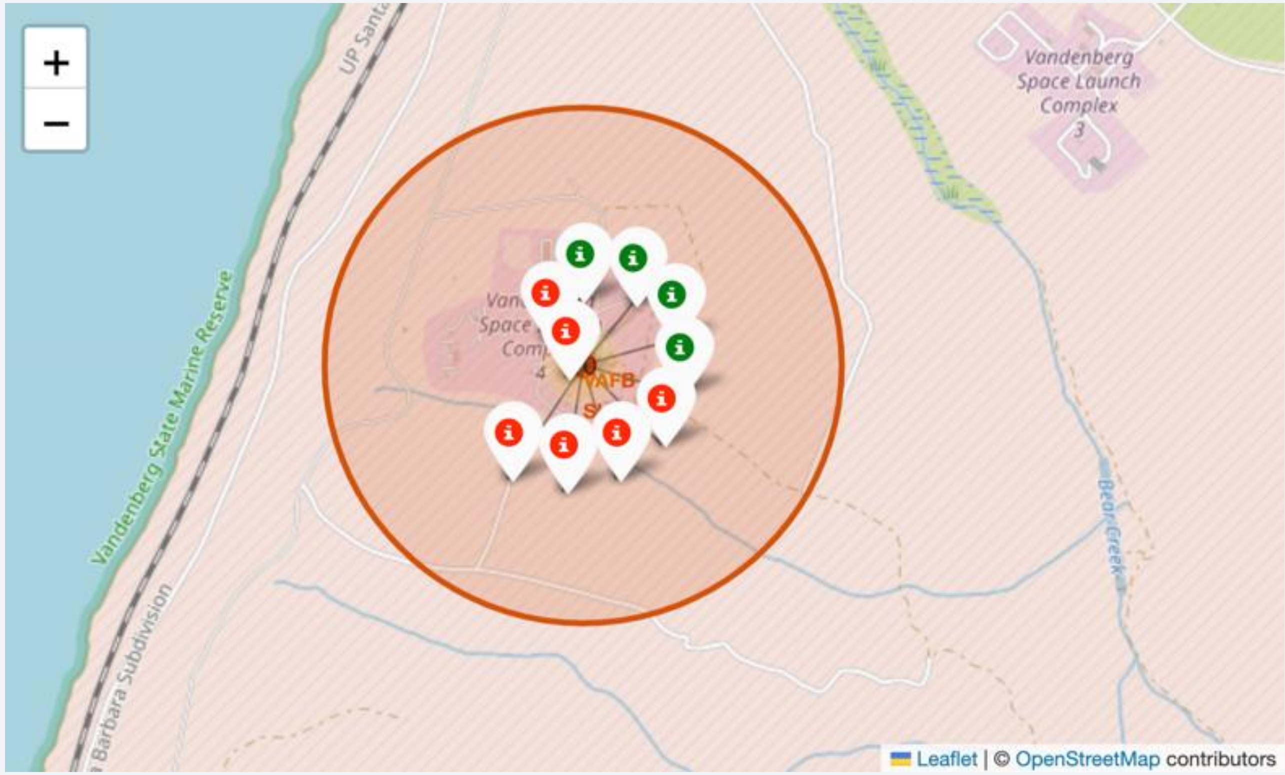
Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Folium map : Launch Site location

<Folium Map Screenshot 2>


+
−
UP Santa
Vandenberg State Marine Reserve
Barbara Subdivision
Vandenberg Space Launch Complex 3
Van Space Comp 4
VAFB S
Bear Creek
Leaflet | © OpenStreetMap contributors

Folium map : Launch Site markers of success and failed launch

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- **Green Marker** = Successful Launch
- **Red Marker** = Failed Launch

Launch Site VAFB SLC 4E has a very high Success Rate.

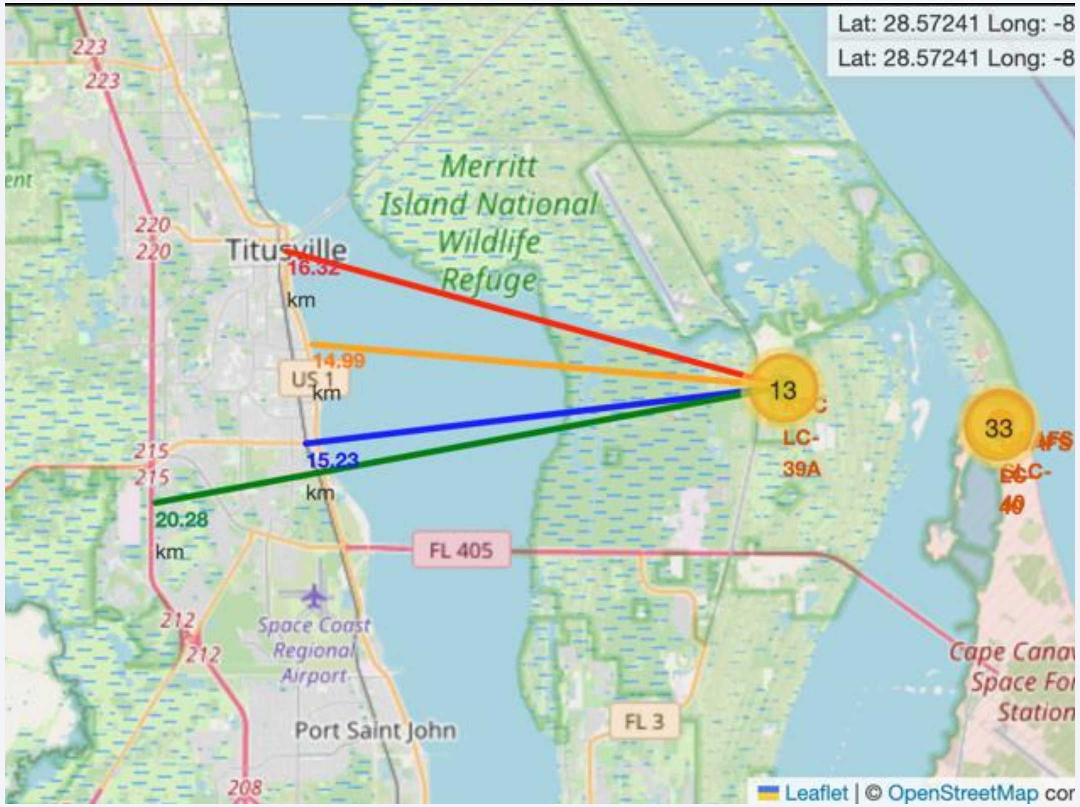
<Folium Map Screenshot 3>

• Operating a visual analysis of the launch site KSC LC-39A we can see that it is:

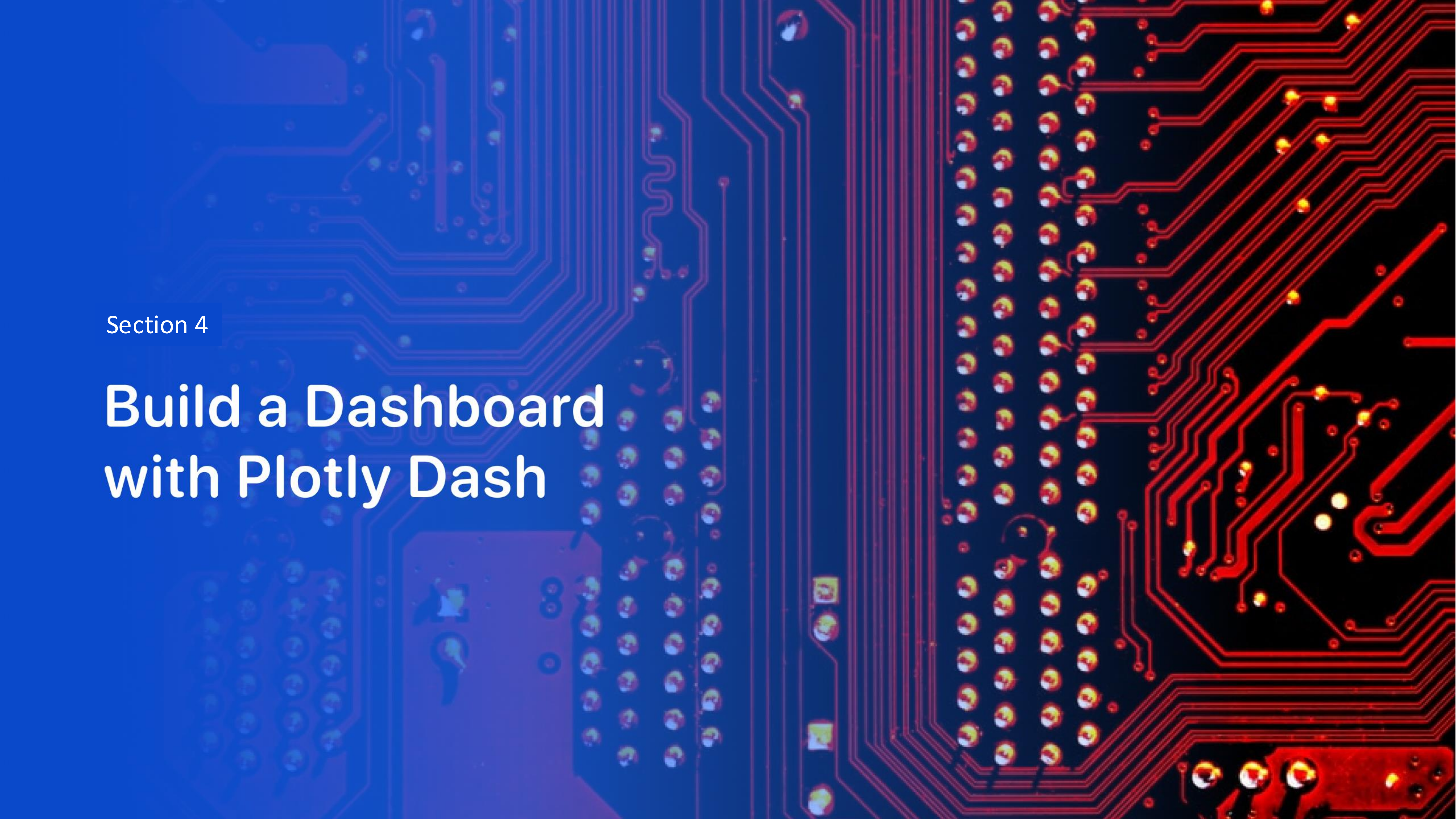
- close to railway (**15.23 km**)
- close to highway (**20.28 km**)
- close to coastline (**14.99 km**)

• Also the launch site KSC LC-39A is pretty close to its closest city Titusville (16.32 km).

• **Failed rocket** with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



Folium map : Launch Site distance

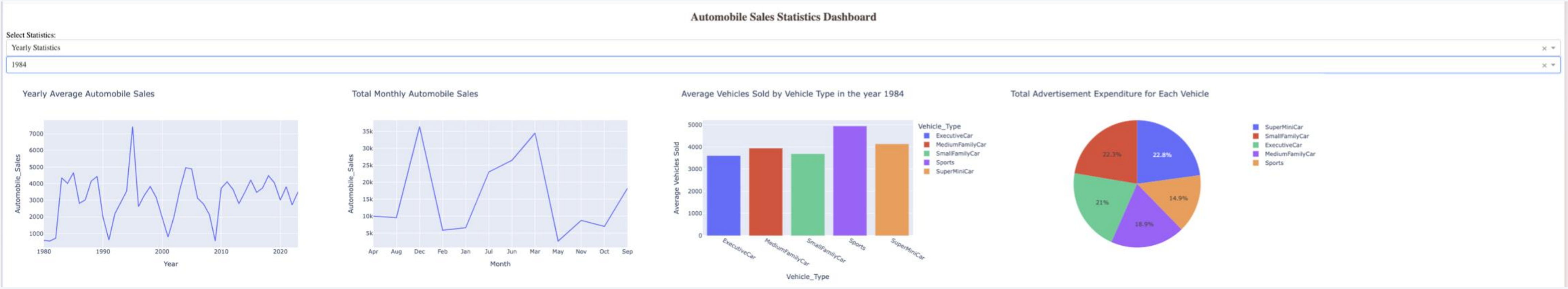


Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

This Ploty dashboard presents a visual breakdown of automobile sales performance and associated advertising investments.



Automobile Sales Statistics Dashboard

Select Statistics:

Yearly Statistics

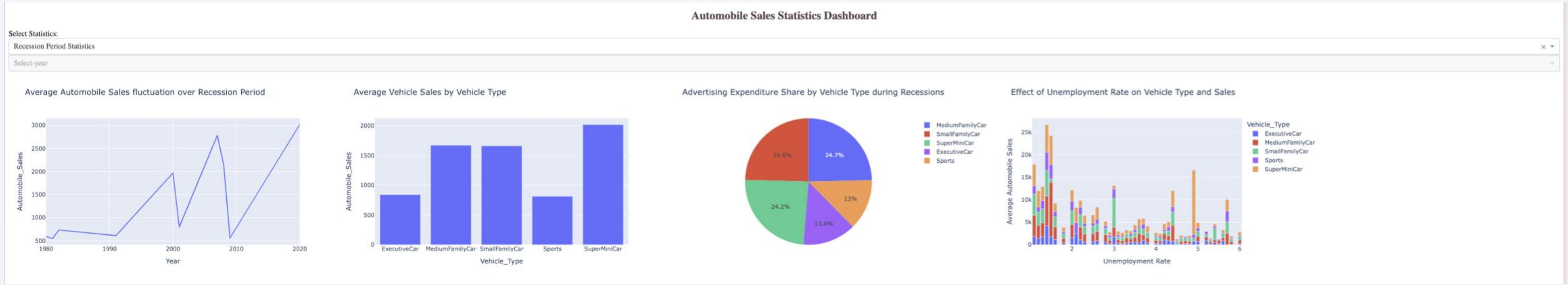
1984

- **Yearly Average Automobile Sales:** Highlights overall sales trends across the observed period.
- **Total Monthly Automobile Sales:** Provides granular monthly performance insights.
- **Average Vehicles Sold by Vehicle Type in 1984:** Compares sales distribution across different vehicle categories for a historical year.
- **Total Advertisement Expenditure for Each Vehicle:** Correlates marketing spend with vehicle types to assess investment alignment.

Supporting charts and data visualizations help identify patterns, seasonality, and the impact of advertising on sales outcomes.
This integrated view supports strategic decision-making for sales targeting and marketing budget allocation.

<Dashboard Screenshot 2>

This section examines how key economic indicators, specifically recessions and unemployment rates, influence automobile sales performance. The analysis focuses on three interconnected areas:



Automobile Sales Statistics Dashboard

Select Statistics:

Recession Period Statistics

Select-year

- **Average Automobile Sales Fluctuation over Recession Period:** A trend line or chart illustrating how total sales volume rises and falls in correlation with official recession periods, highlighting the sensitivity of the industry to broader economic downturns.
- **Advertising Expenditure Share by Vehicle Type during Recessions:** A breakdown showing how marketing budgets are reallocated among different vehicle types (e.g., sedans, SUVs, trucks) during economic contractions, revealing strategic shifts in promotional focus.
- **Effect of Unemployment Rate on Vehicle Type and Sales:** An analysis visualizing the relationship between rising unemployment rates and the sales performance of specific vehicle segments. This identifies which vehicle types are most resilient or vulnerable to changes in consumer employment stability.

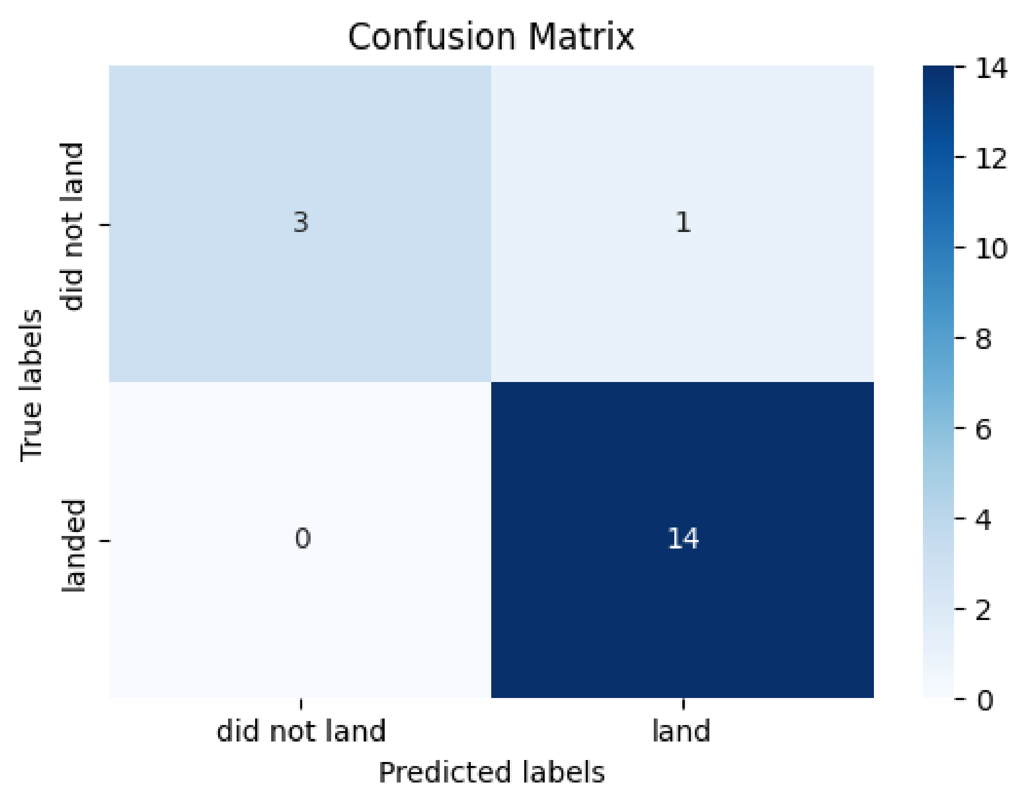
Together, these visualizations provide a data-driven foundation for understanding market dynamics during economic stress and for planning resilient sales and advertising strategies.

Section 5

Predictive Analysis (Classification)

Confusion Accuracy and Matrix

We used **svm_cv** and **tree_cv** as well as **GridsearchCV** to create two predictive model and find the best parameter which gave us this prediction matrix having only one false positive giving us a confusion matrix accuracy of 94,44% and a Cross-validated training accuracy of 78%



Conclusions

- The **Decision Tree model** proved to be the most effective algorithm for this dataset.
- **Launches with lower payload masses** tend to achieve better outcomes than those carrying heavier payloads.
- Most **launch sites are located near the Equator** and are all situated **close to coastal areas**.
- The **launch success rate has steadily increased over time**.
- **KSC LC-39A** records the **highest launch success rate** among all sites.
- Launches to **ES-L1, GEO, HEO, and SSO orbits** achieved a **100% success rate**.

Appendix

Special Thanks to :

All coursera instructors
IBM professional certificate
Coursera®

Thank you!