IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Beatriz Calzadilla>
<June 19th>

Outline

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

Executive Summary

Summary of methodologies

- Data collected via API and web scraping
- Exploratory Data Analysis (EDA) using Pandas and SQLite
- Geographic visualization with Folium
- Visualization of key information in an interactive dashboard
- Predictions using machine learning models.

Summary of all results

Among all evaluated models, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) achieved the highest predictive accuracy.

Introduction

Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

Problems you want to find answers

This project aims to predict the success of a rocket landing based on key variables such as payload mass, booster model, and achieved orbit.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Api
  - Web scrapping
- Perform data wrangling
  - A new column ‘Class’ was added, to work as labels
- 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

- Via API

  A connection was made with the SpaceX site, so data could be retrieved and used to set up the first version of the dataset

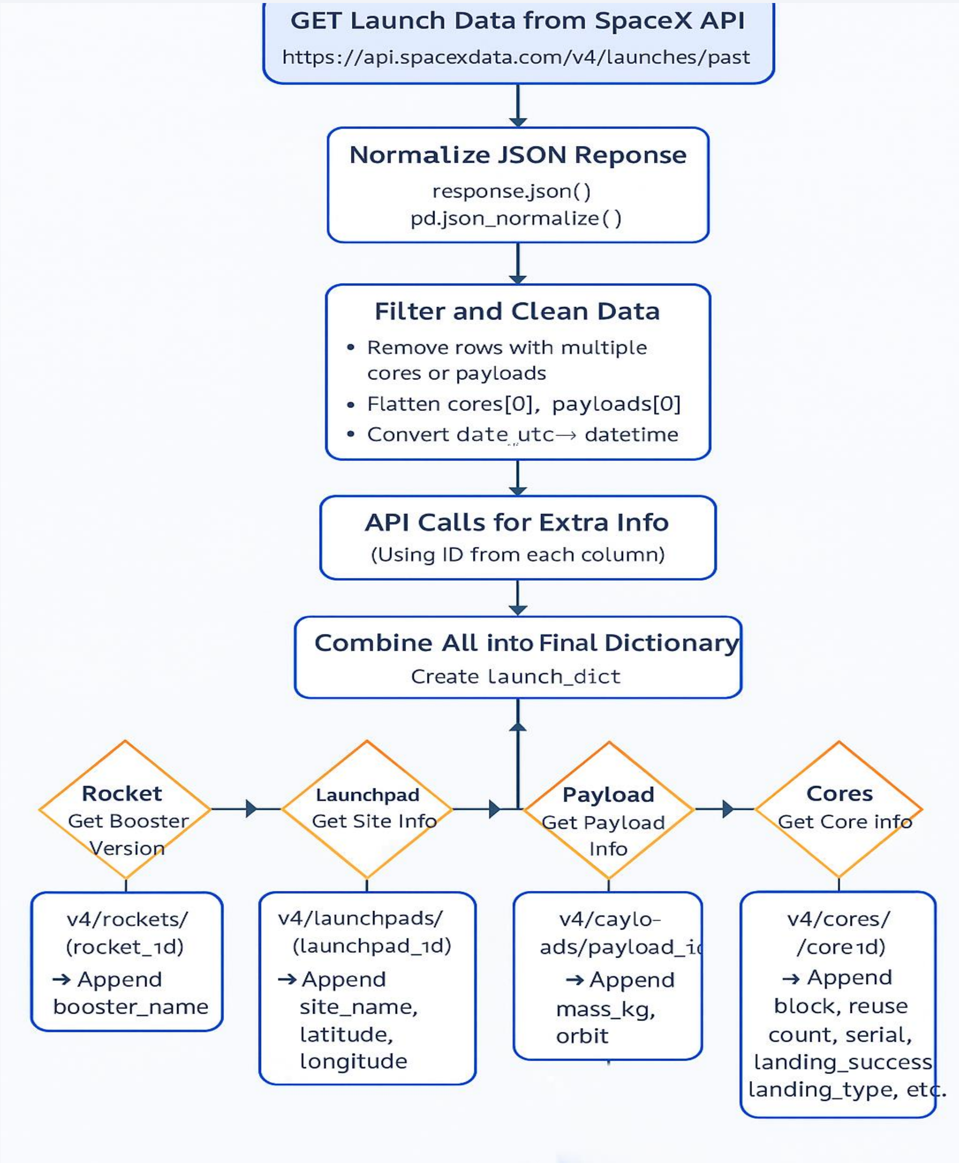
- Via Web Scrapping

  The SpaceX Wikipedia page was scrapped to finish assembling the first version of the dataset

Data Collection – SpaceX API

Launch data was retrieved using the SpaceX REST API. The JSON response was normalized and filtered to remove rows with multiple cores or payloads. Additional information was obtained via API calls using identification numbers, and the resulting data was compiled into a structured dataset for analysis.

https://github.com/betica07/DSc_CapstoneProject_IBM.git



GET Launch Data from SpaceX API
https://api.spacexdata.com/v4/launches/past

Normalize JSON Reponse
response.json()
pd.json_normalize()

Filter and Clean Data
- Remove rows with multiple cores or payloads
- Flatten cores[0], payloads[0]
- Convert date utc→ datetime

API Calls for Extra Info
(Using ID from each column)

Combine All into Final Dictionary
Create launch_dict

Rocket
Get Booster Version

Launchpad
Get Site Info

Payload
Get Payload Info

Cores
Get Core info

v4/rockets/
(rocket_1d)
→ Append
booster_name

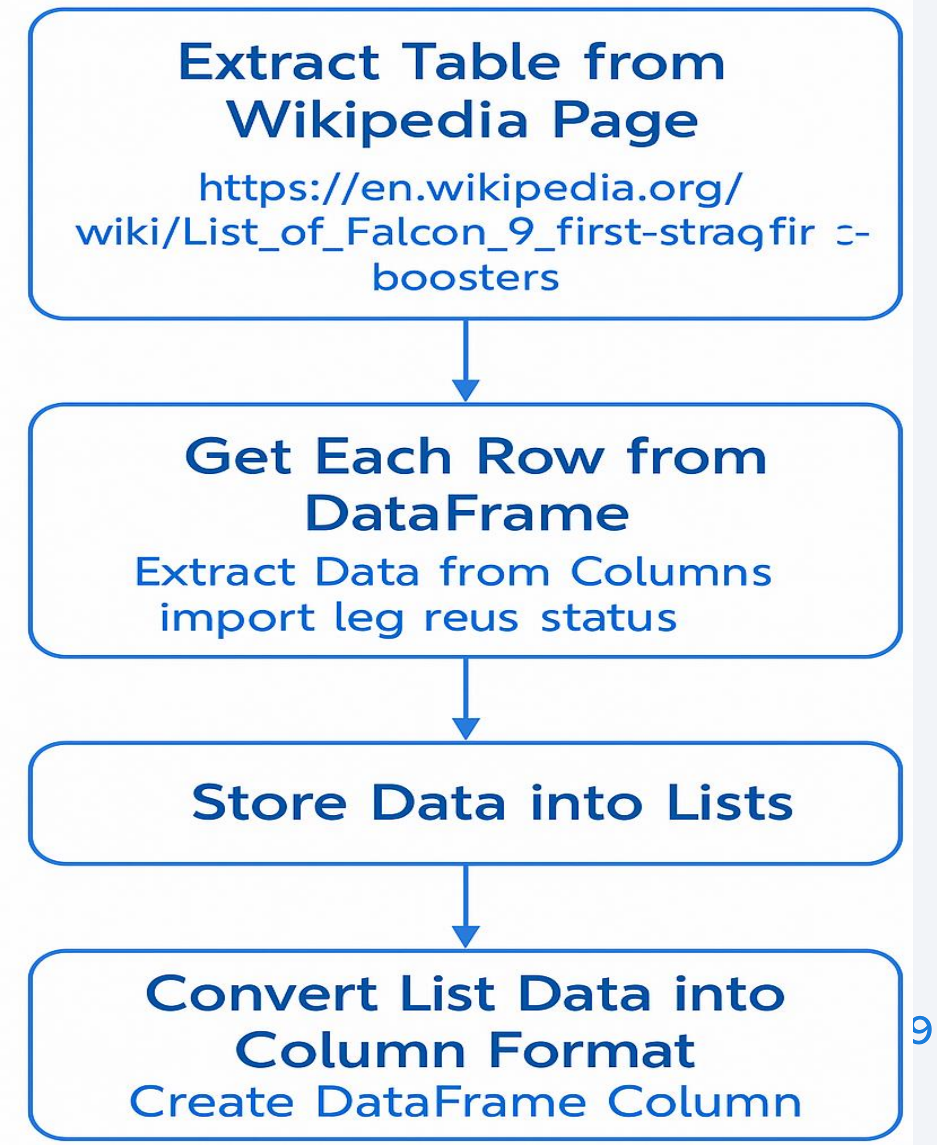
v4/launchpads/
(launchpad_1d)
→ Append
site_name,
latitude,
longitude

v4/caylo-
ads/payload_i
→ Append
mass_kg,
orbit

v4/cores/
/core1d)
→ Append
block, reuse
count, serial,
landing_success
landing_type, etc.

Data Collection - Scraping

- HTML elements were identified using class attributes.
  requests.get() and BeautifulSoup were used to extract and parse launch record data from the SpaceX site.
  Relevant content (launch site, date, payload mass, booster version, etc.) was extracted and stored in a structured dataframe.
- https://github.com/betica07/DSc_CapstoneProject_IBM.git



Extract Table from Wikipedia Page

https://en.wikipedia.org/wiki/List_of_Falcon_9_first-straqfir c-boosters

Get Each Row from DataFrame

Extract Data from Columns import leg reus status

Store Data into Lists

Convert List Data into Column Format

Create DataFrame Column

Data Wrangling

- Data collected from API and web scraping was first inspected to identify missing or duplicate values.
- Rows with missing target values were removed, and categorical variables were encoded using one-hot encoding.
- Features were renamed for consistency, and column types were adjusted to enable model training.
- The final dataset was balanced and cleaned to improve prediction accuracy in later stages.

EDA with Data Visualization

Several charts were plotted to explore the relationship between payload mass, launch site, orbit type, and mission success. A scatter plot charts was used to assess the correlation between payload mass and launch site, as well as flight number and orbit type. A bar chat was used to visualize the correlation between orbit type and success rate and a line chart to do a follow up of the success rate over the years. These visualizations supported pattern recognition and variable selection for predictive modeling.

https://github.com/betica07/DSc_CapstoneProject_IBM.git

EDA with SQL

- Queried and counted successful launches grouped by launch site.
- Analyzed the total number of launches from the site with the highest success rate.
- Retrieved missions with payloads over 4000 kg targeting Low Earth Orbit (LEO).
- Filtered and ranked launches by success and payload mass, limiting to the top 5 entries.
- Grouped data by booster version and launch site to identify the most frequent combinations.

https://github.com/betica07/DSc_CapstoneProject_IBM.git

Build an Interactive Map with Folium

- Markers were added for each SpaceX launch site using their respective coordinates to visualize launch locations.
- Circle markers were included to highlight the geographical boundaries of launch sites.
- Popups with site names were attached to each marker for easy identification.
- The purpose of these objects is to offer an intuitive and interactive view of the spatial distribution of SpaceX launch activity, aiding in spatial analysis and contextual understanding.

https://github.com/betica07/DSc_CapstoneProject_IBM.git

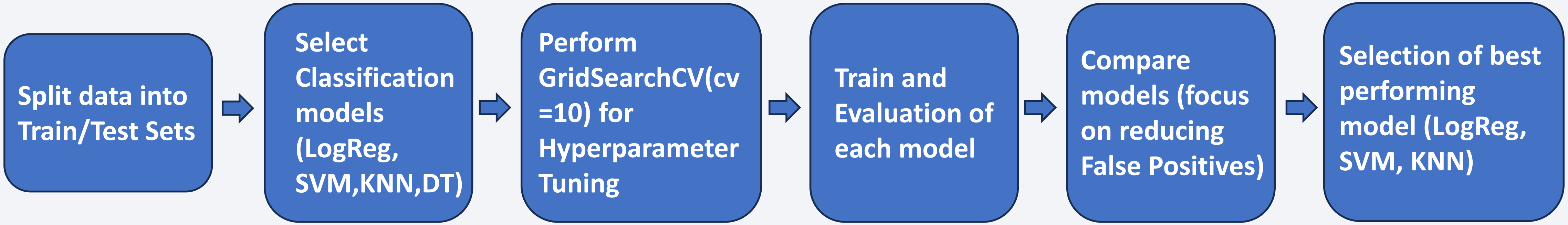
Build a Dashboard with Plotly Dash

- A pie chart was added to display the total number of successful launches by site. When a specific launch site is selected, the chart shows the distribution between successful and failed launches.
- A scatter plot was included to show the correlation between payload mass and launch outcome, with filtering options by site and payload range.
- These plots and interactions were included to allow dynamic exploration of launch success rates across different sites and payloads, enabling pattern discovery through interactive visual analysis.

https://github.com/betica07/DSc_CapstoneProject_IBM.git

Predictive Analysis (Classification)

- Multiple classification models were trained and evaluated using GridSearchCV with 10-fold cross-validation to identify the best-performing algorithm.
- Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) demonstrated the highest accuracy and lowest number of false positives, making them the most suitable choices for the task.
- In contrast, the Decision Tree model showed weaker performance, likely due to overfitting and limited generalization on small datasets.



Split data into Train/Test Sets

Select Classification models (LogReg, SVM,KNN,DT)

Perform GridSearchCV(cv =10) for Hyperparameter Tuning

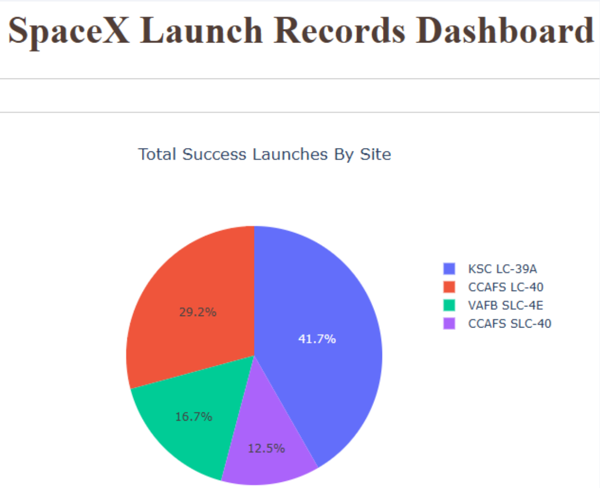
Train and Evaluation of each model

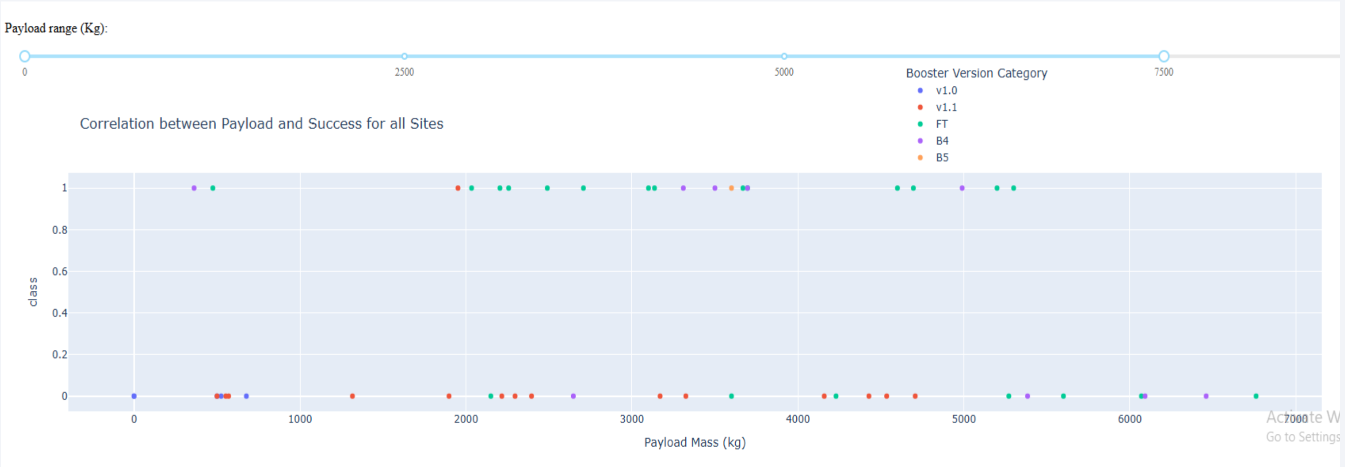
Compare models (focus on reducing False Positives)

Selection of best performing model (LogReg, SVM, KNN)

Results

- The site with the largest number of successful launches is KSC LC-39A, with 10 successful launches out of 13 total.
- The site with the highest launch success rate is also KSC LC-39A, achieving a success rate of 76.9%.
- The payload range with the highest launch success rate is 3000–4000 kg.
- The payload ranges with the lowest launch success rates are 0–1000 kg and 6000–7000 kg.
- The Falcon 9 Booster version with the highest launch success rate is FT.

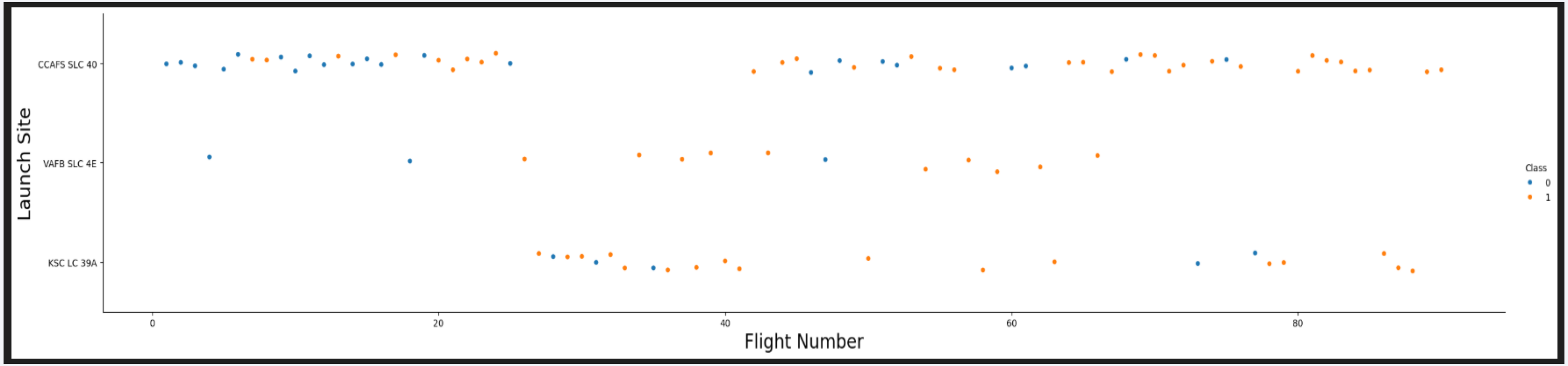




Section 2

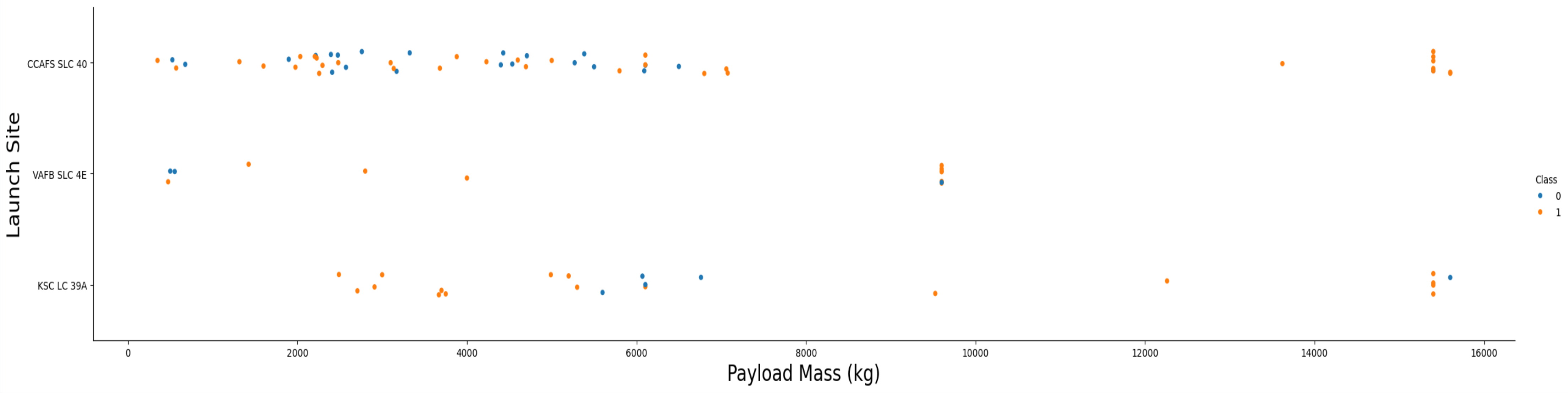
Insights drawn from EDA

Flight Number vs. Launch Site



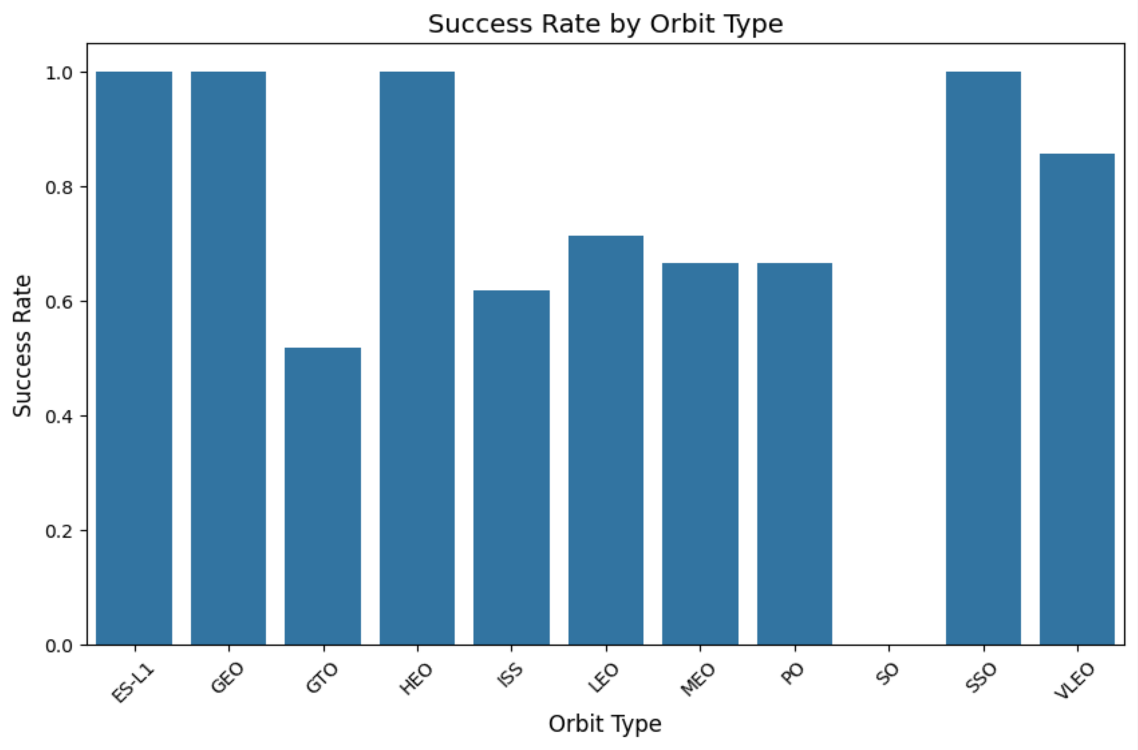
CCAFS SLC-40 shows a high launch frequency and appears to have a comparatively higher proportion of successful missions

Payload vs. Launch Site



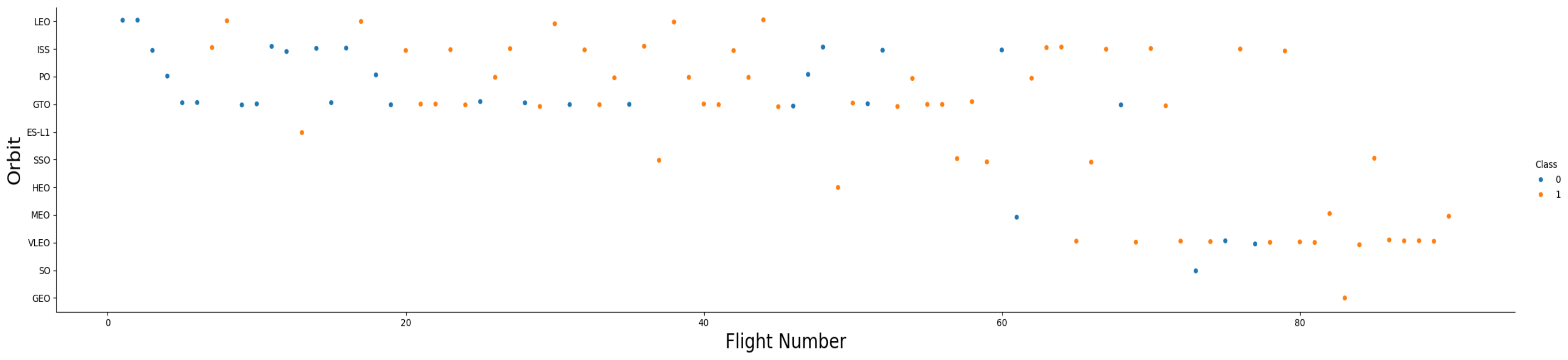
A larger number of launches originated from the CCAFS SLC-40 site, where most of the successful missions were concentrated. However, when analyzing success rate, the VAFB SLC-4E and KSC LC-39A sites show a higher success ratio, with approximately 77% of their launches resulting in successful landings.

Success Rate vs. Orbit Type



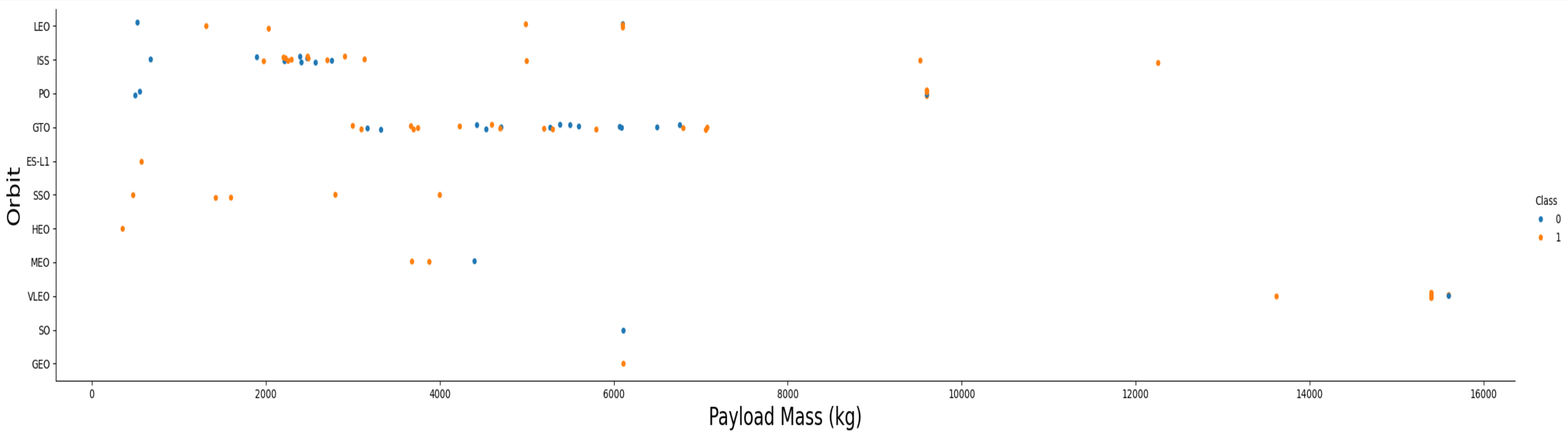
These orbits — HEO, GEO, ES-L1, and SSO — show a 100% success rate, likely due to their use in critical and well-planned missions. They are stable or strategically important trajectories, commonly used for observation, communication, or research, often supported by mature technologies and high-quality control.

Flight Number vs. Orbit Type



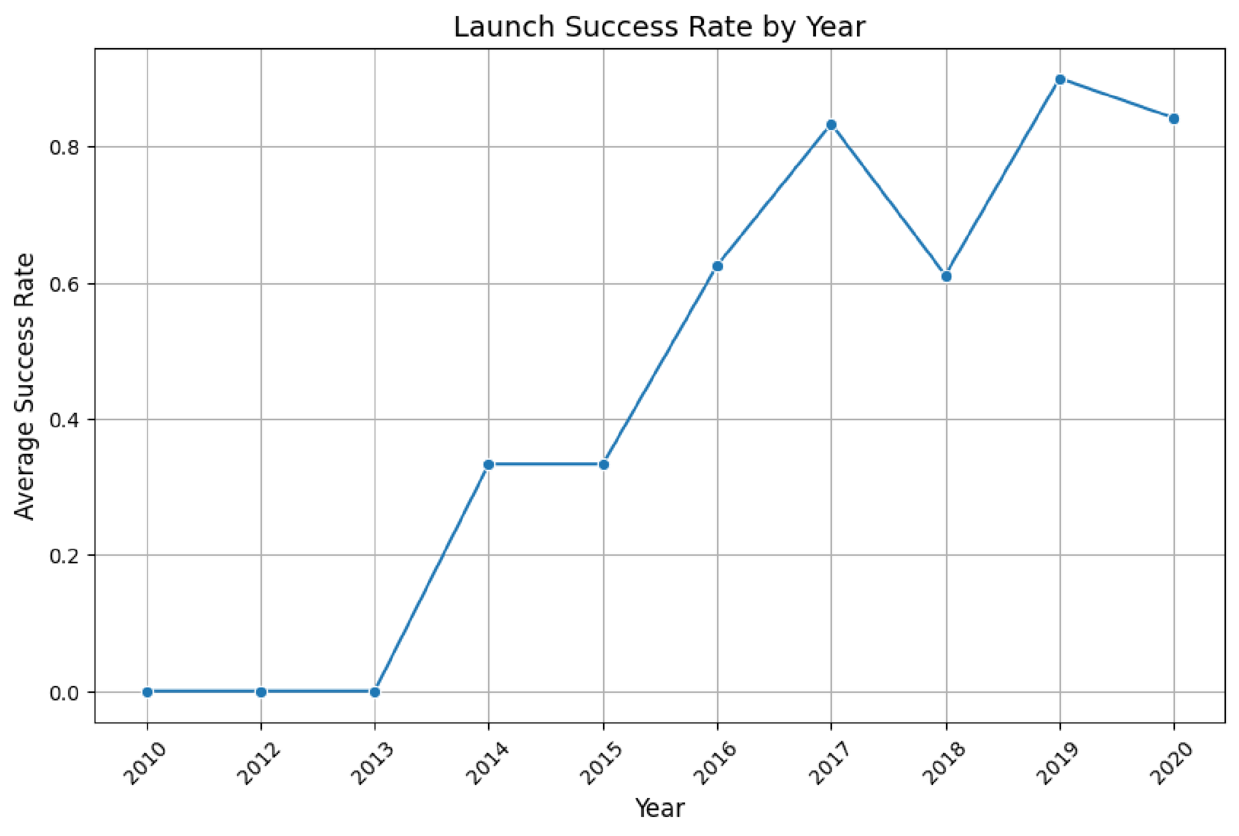
Most launches were directed toward GTO, PO, ISS, and LEO orbits. Among these, GTO registered the highest number of successful launches, while VLEO stands out as the orbit type with the best success rate.

Payload vs. Orbit Type



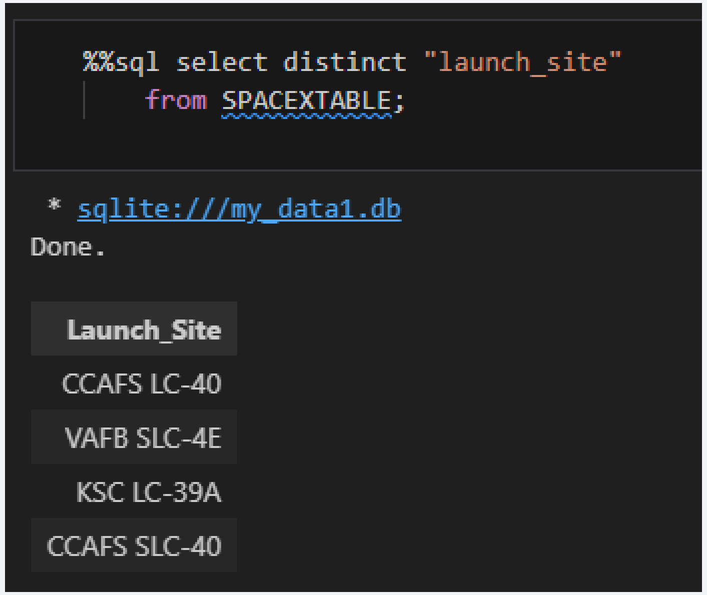
- Most successful launches occurred with light to mid-range payloads directed to LEO, SSO, and HEO orbits, showing consistent reliability.
- GTO and GEO handled heavier payloads, with GTO being the most frequently used high-orbit and still maintaining a high success rate.

Launch Success Yearly Trend



This chart shows a significant increase in the success rate over the years, highlighting the progress and technological advancement achieved over time

All Launch Site Names



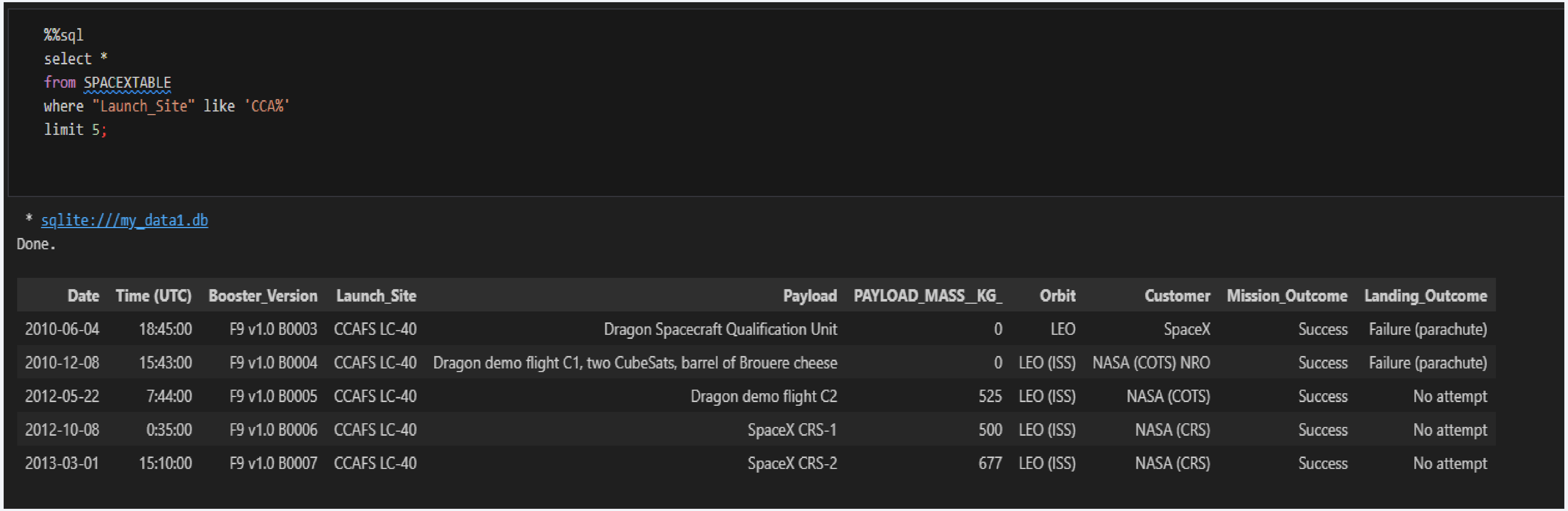
```
%%sql select distinct "launch_site"
    from SPACEXTABLE;
```

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

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

A query was executed to retrieve all distinct launch site names from the dataset, revealing three unique launch locations used by SpaceX."

Launch Site Names Begin with 'CCA'



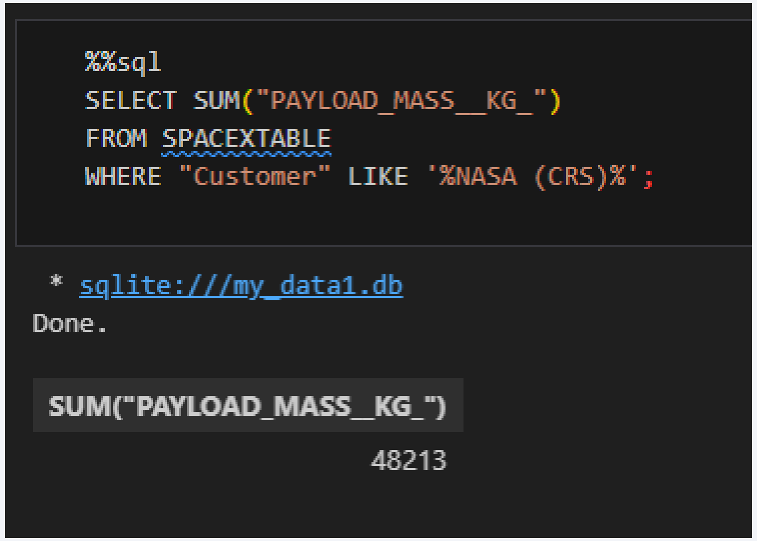
```
%%sql
select *
from SPACEXTABLE
where "Launch_Site" like 'CCA%'
limit 5;
```

```
 * 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

A SQL query was used to retrieve launch records where the site name starts with 'CCA', displaying early missions launched from CCAFS LC-40 between 2010 and 2013

Total Payload Mass

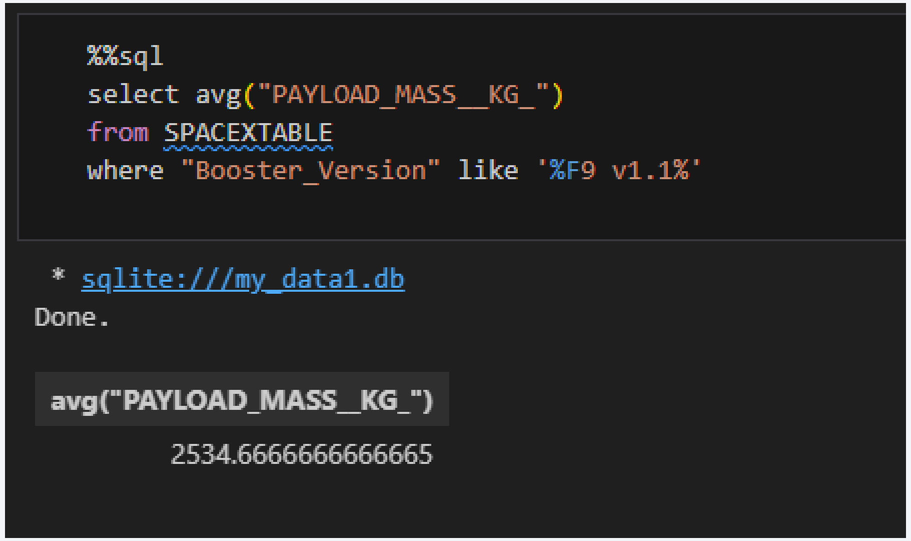


```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_")
FROM SPACEXTABLE
WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

SUM("PAYLOAD_MASS__KG_")
48213

Average Payload Mass by F9 v1.1

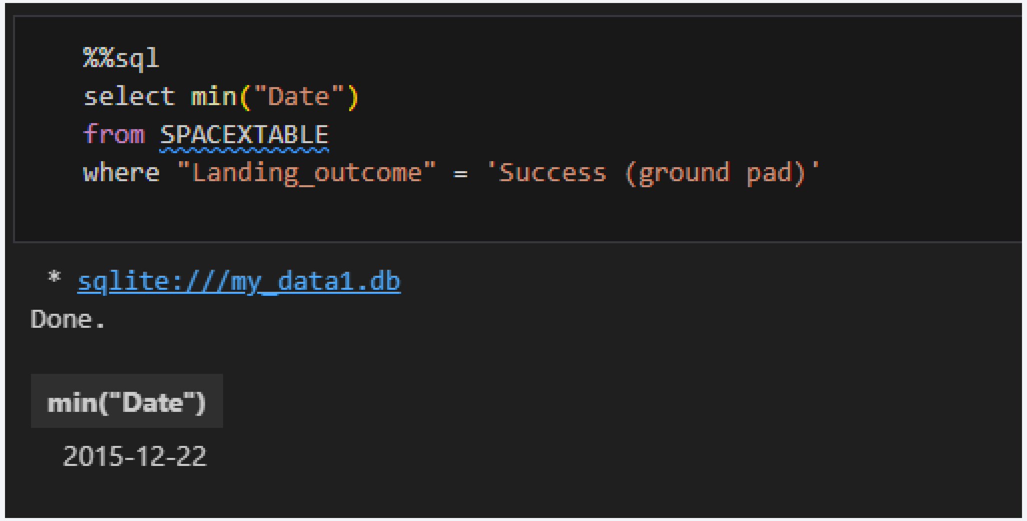


```
%%sql
select avg("PAYLOAD_MASS__KG_")
from SPACEXTABLE
where "Booster_Version" like '%F9 v1.1%'
```

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

avg("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

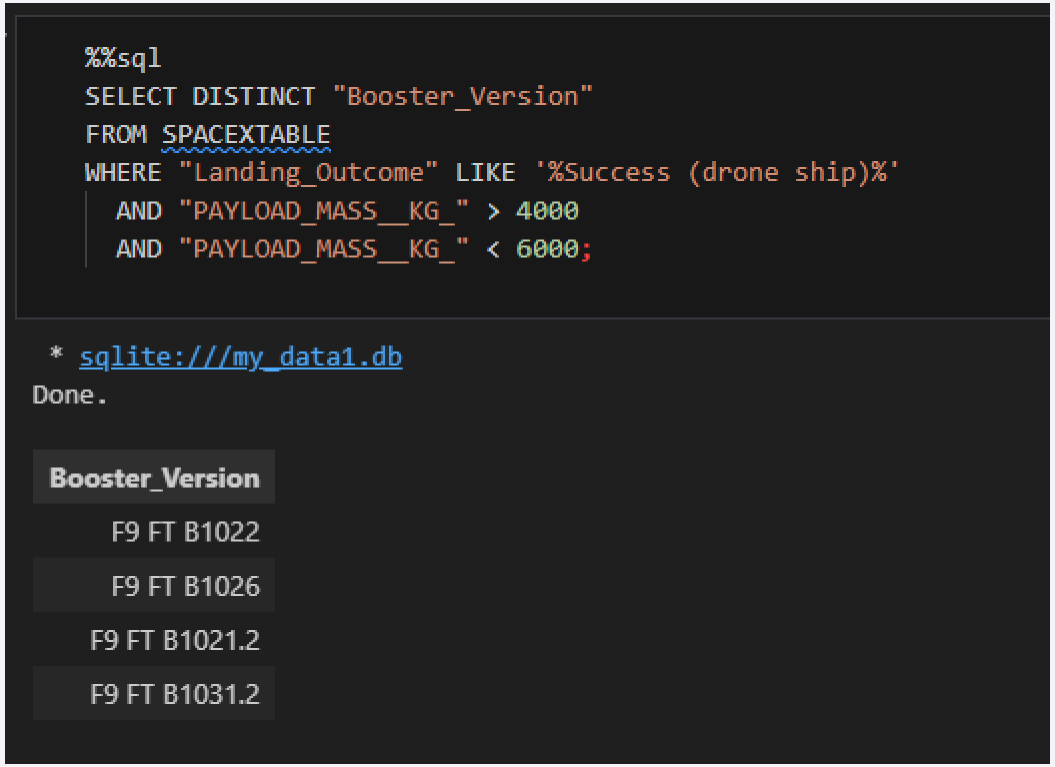


```
%%sql
select min("Date")
from SPACEXTABLE
where "Landing_outcome" = 'Success (ground pad)'
```

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

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

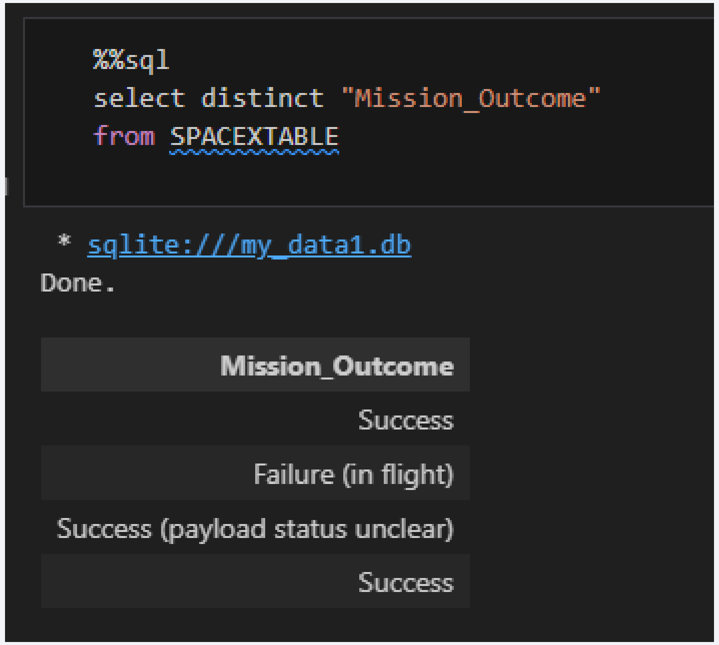


```
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE '%Success (drone ship)%'
  AND "PAYLOAD_MASS__KG_" > 4000
  AND "PAYLOAD_MASS__KG_" < 6000;
```

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

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

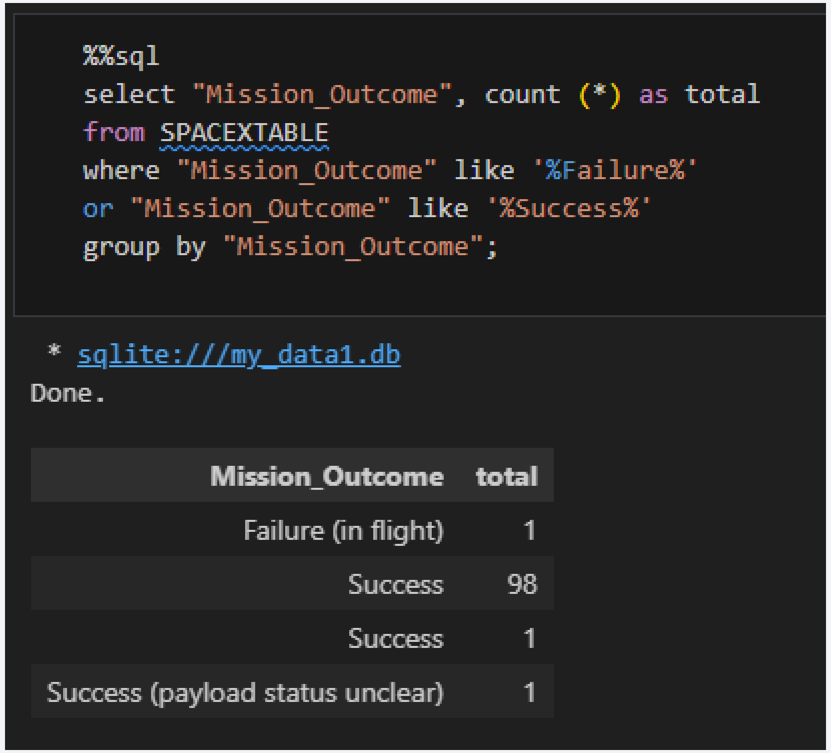
Total Number of Successful and Failure Mission Outcomes



```
%%sql
select distinct "Mission_Outcome"
from SPACEXTABLE
```

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

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

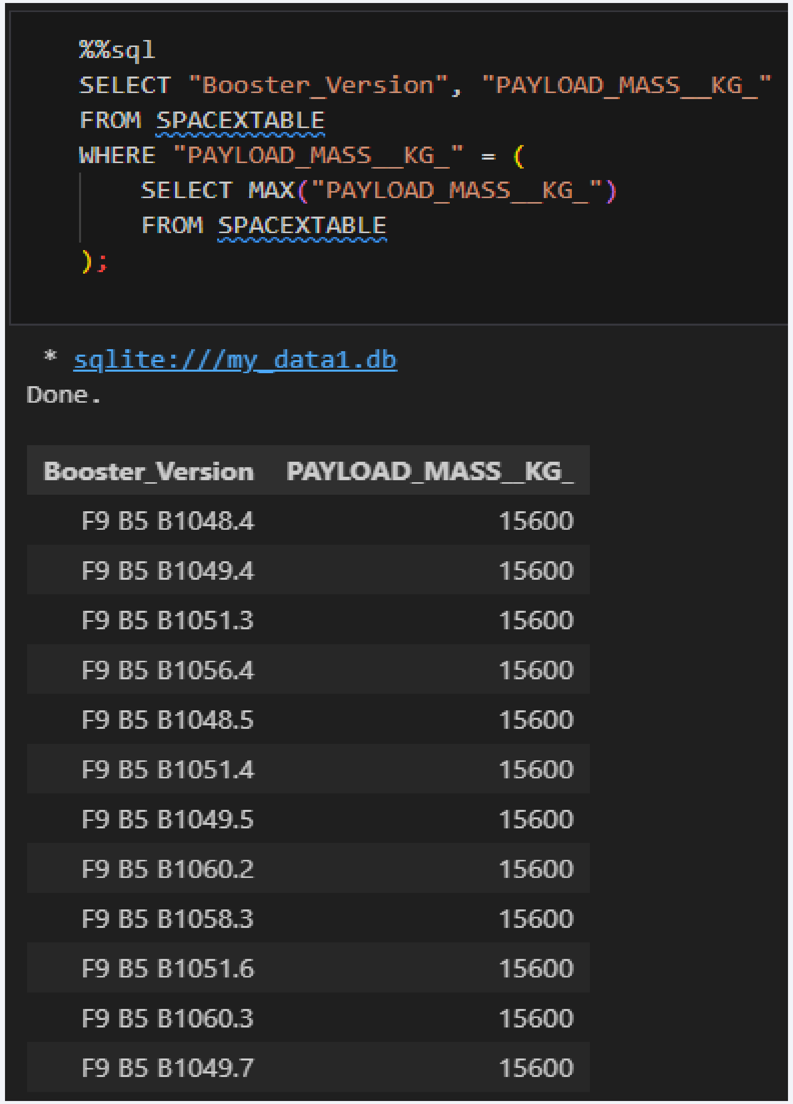


```
%%sql
select "Mission_Outcome", count (*) as total
from SPACEXTABLE
where "Mission_Outcome" like '%Failure%'
or "Mission_Outcome" like '%Success%'
group by "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

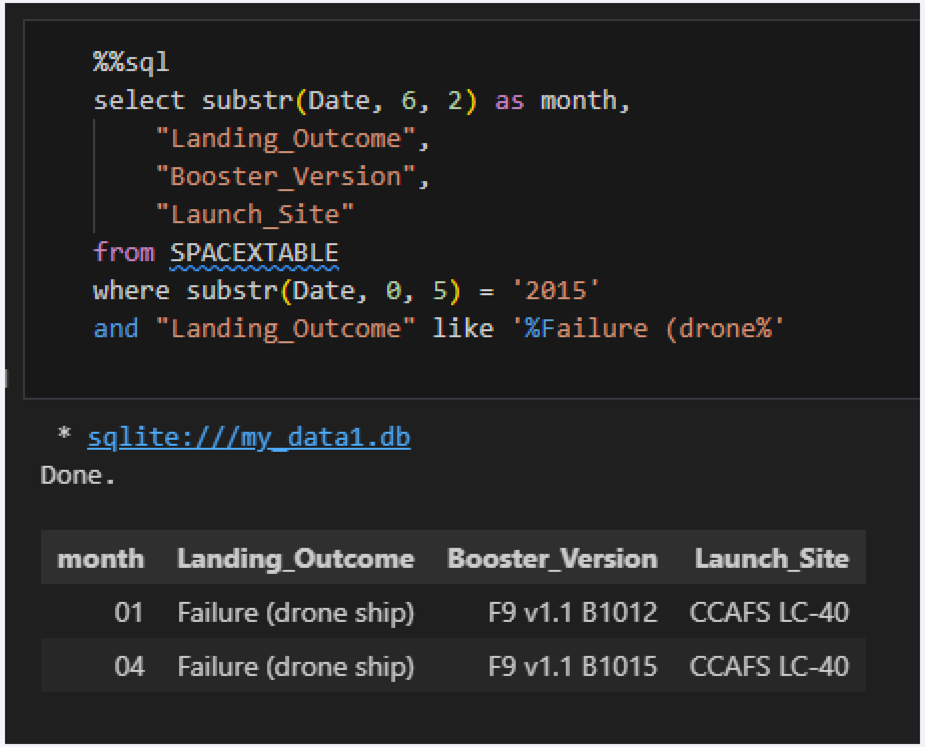


```
%%sql
SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

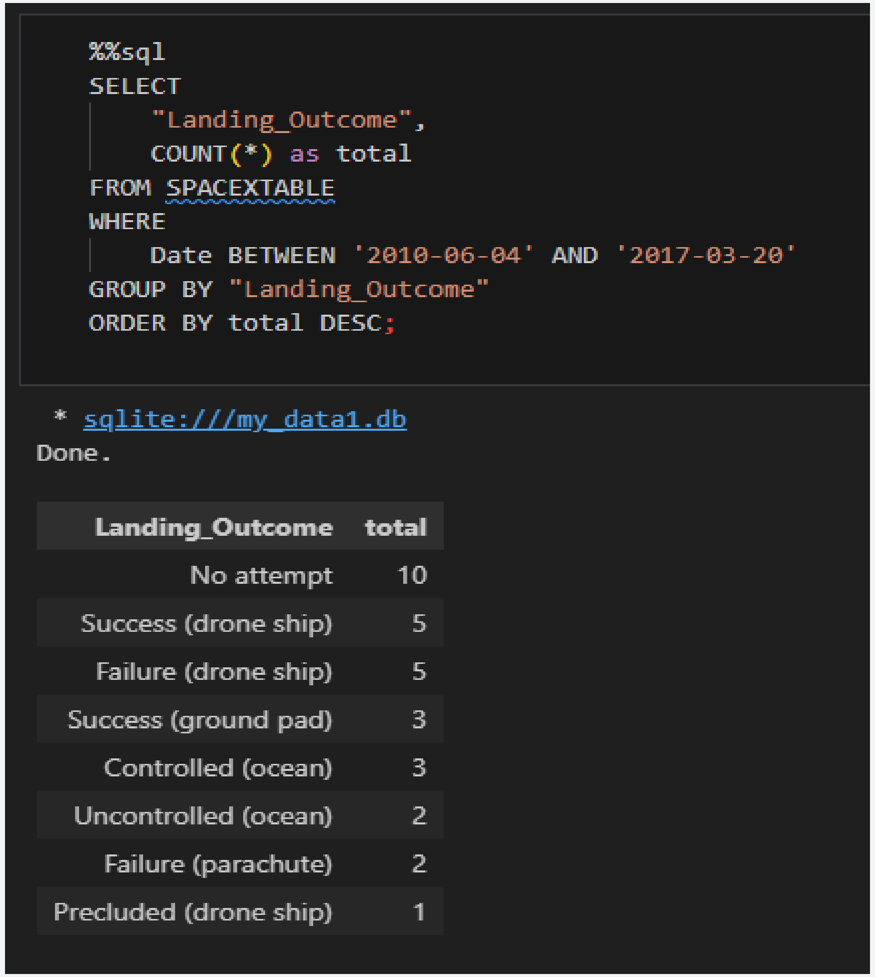


```
%%sql
select substr(Date, 6, 2) as month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
from SPACEXTABLE
where substr(Date, 0, 5) = '2015'
and "Landing_Outcome" like '%Failure (drone%'
```

```
 * 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

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



```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) as total
FROM SPACEXTABLE
WHERE
    Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY total DESC;
```

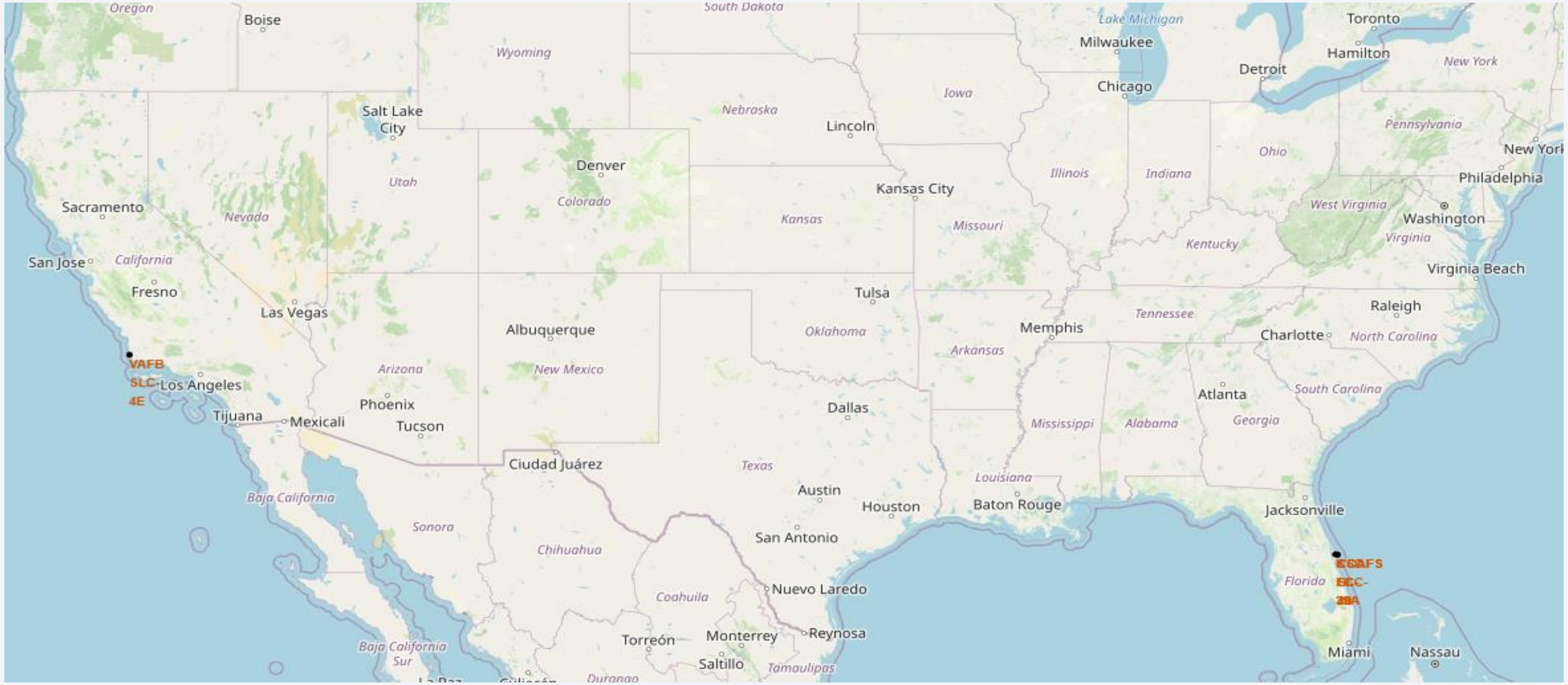
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	total
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

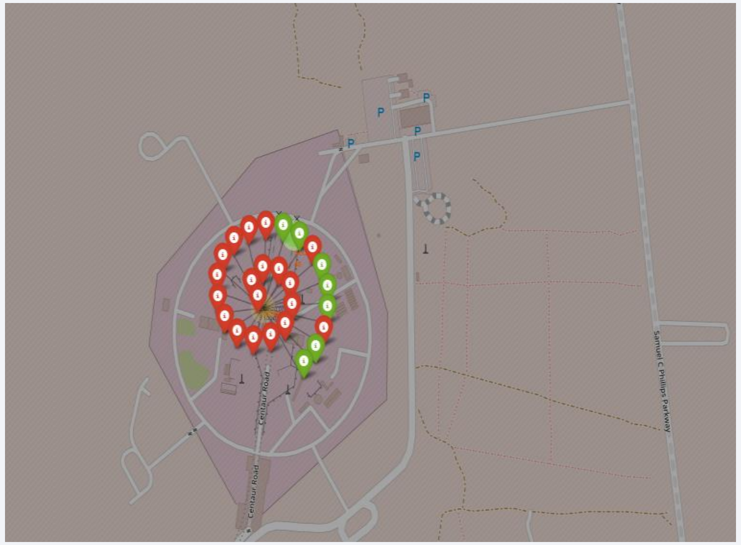
Section 3

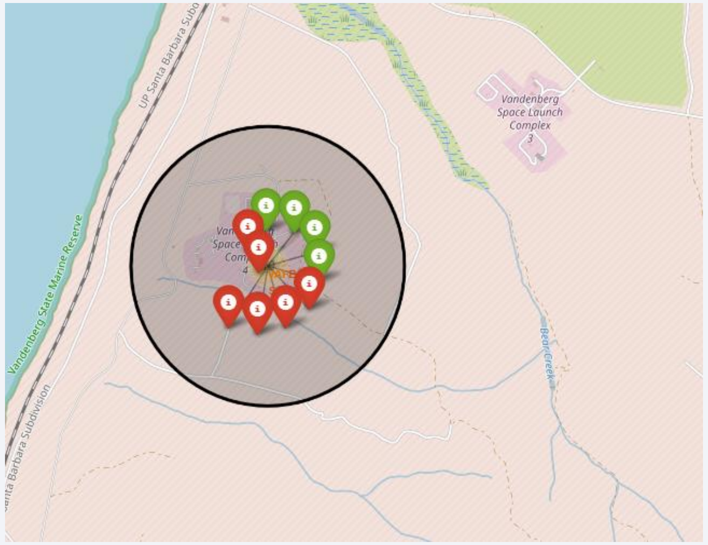
Launch Sites Proximities Analysis

< Launch sites' location markers map>



<Folium Map Screenshot 2>


P
Centaur Road
Samuel C Phillips Parkway

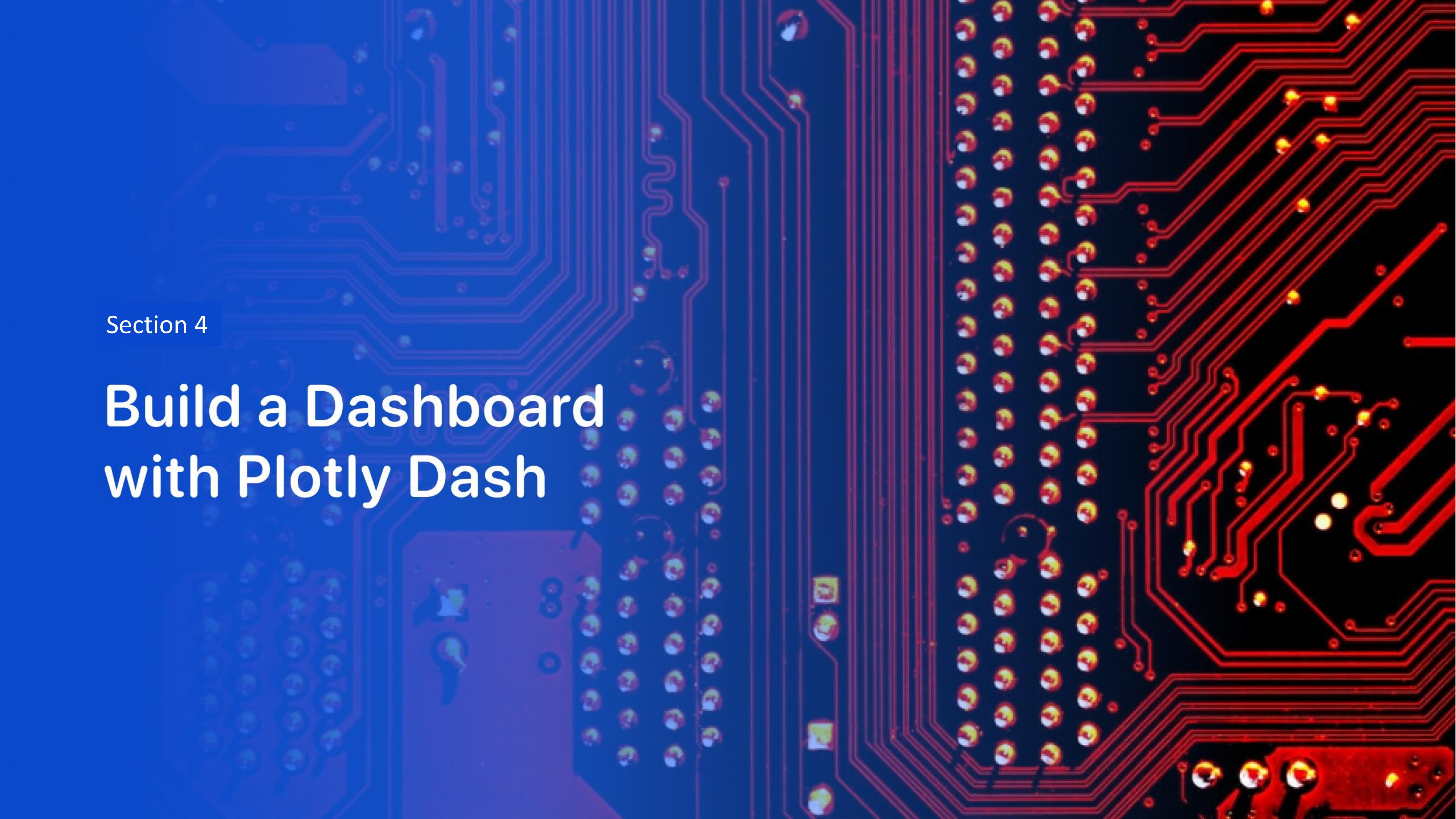

UP Santa Barbara Subdivision
Vandenberg Space Launch Complex 3
Vandenberg State Marine Reserve
Bear Creek
Santa Barbara Subdivision

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

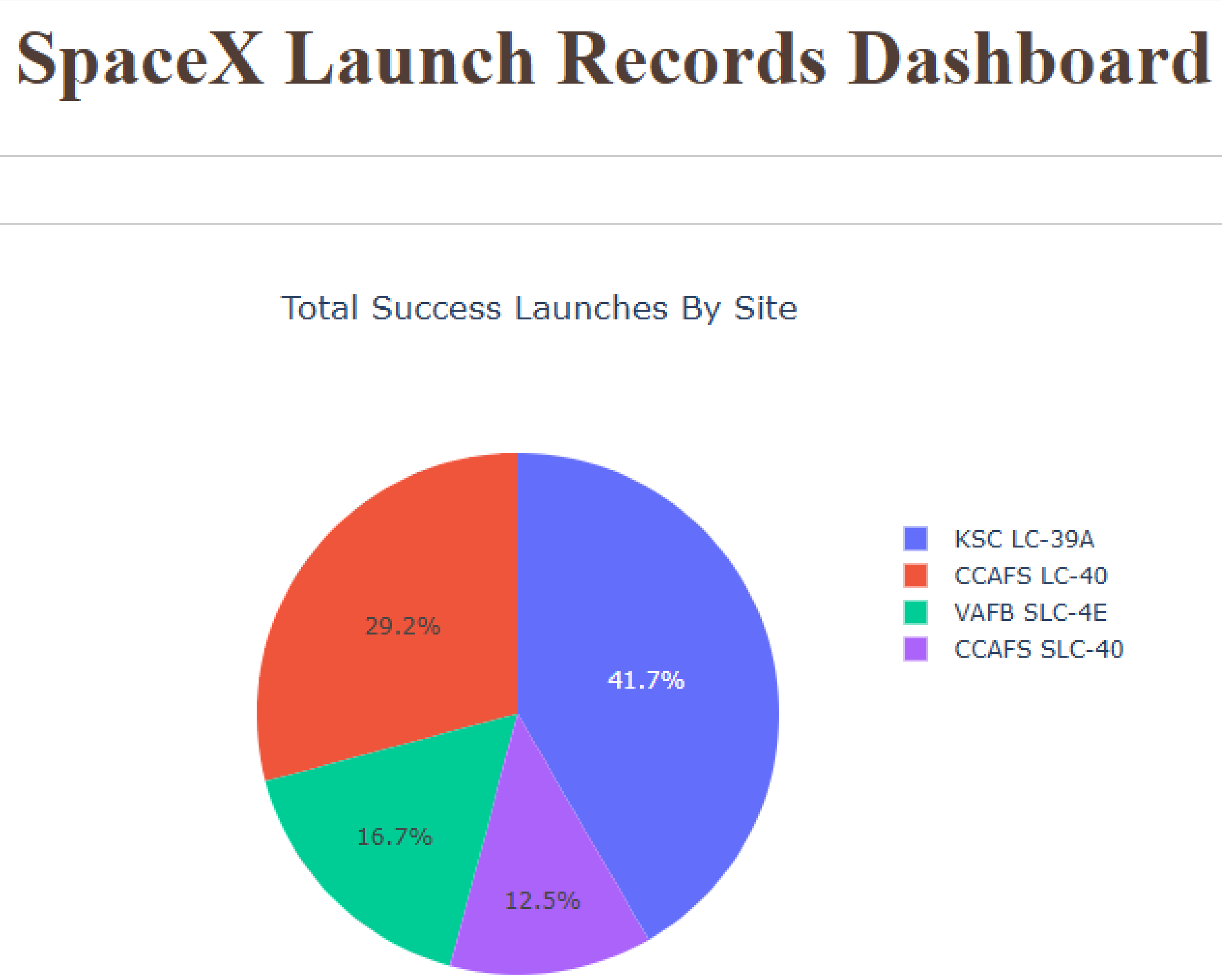
- Explain the important elements and findings on the screenshot



Section 4

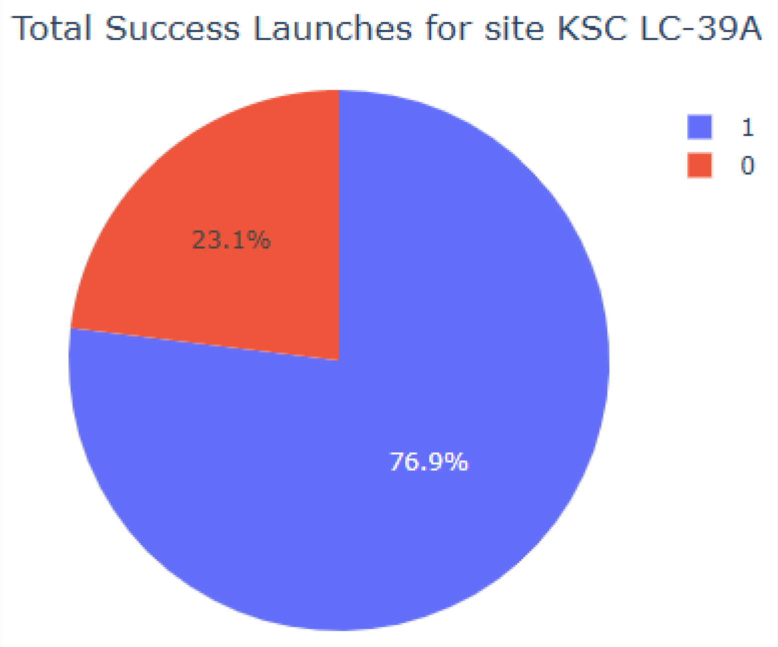
Build a Dashboard with Plotly Dash

< Pie chart “Launch success count”>

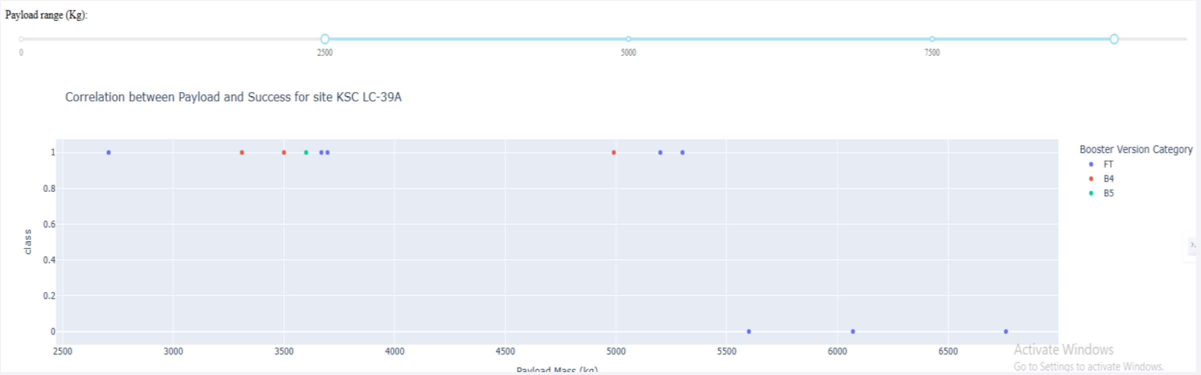


SpaceX Launch Records Dashboard

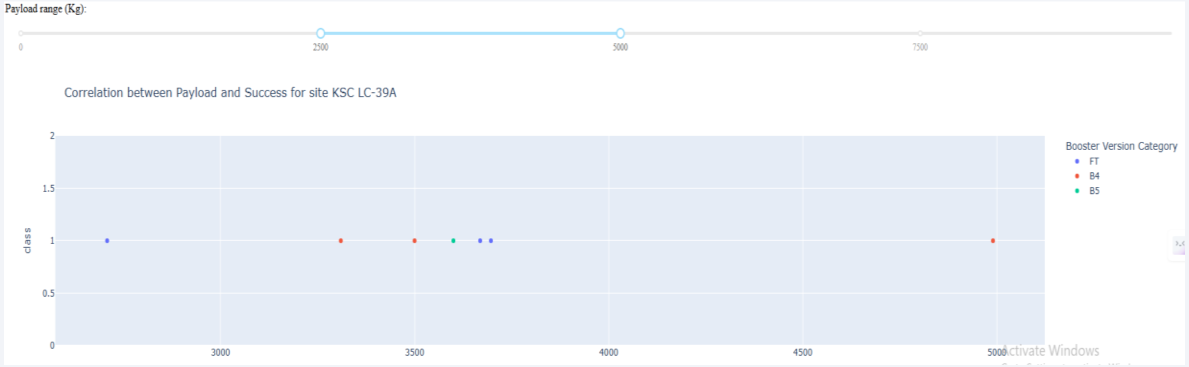
<Chart “Highest success ratio”>


Total Success Launches for site KSC LC-39A
1
0
23.1%
76.9%

<Chart Payload and Launch outcome>



Payload range (Kg):
0
2500
5000
7500
Correlation between Payload and Success for site KSC LC-39A
class
1
0.8
0.6
0.4
0.2
0
2500
3000
3500
4000
4500
5000
5500
6000
6500
Payload Mass (kg)
Booster Version Category
FT
B4
B5
Activate Windows
Go to Settings to activate Windows.

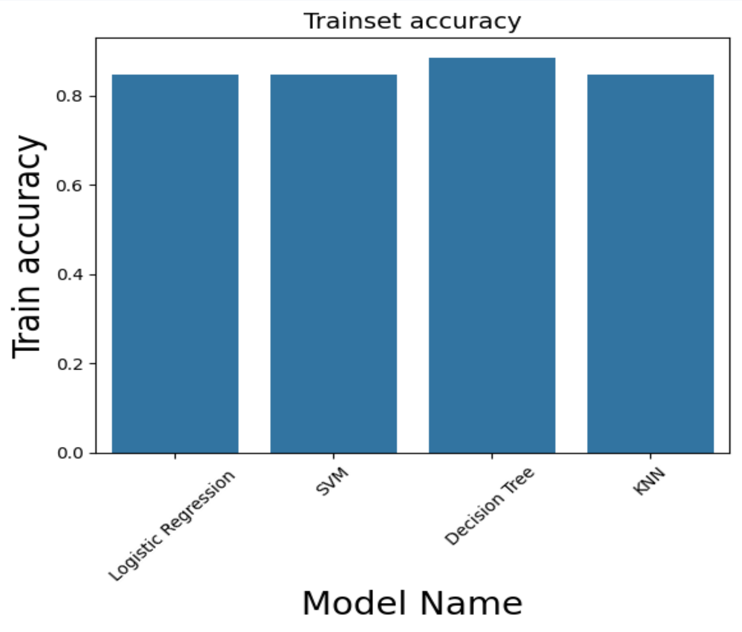


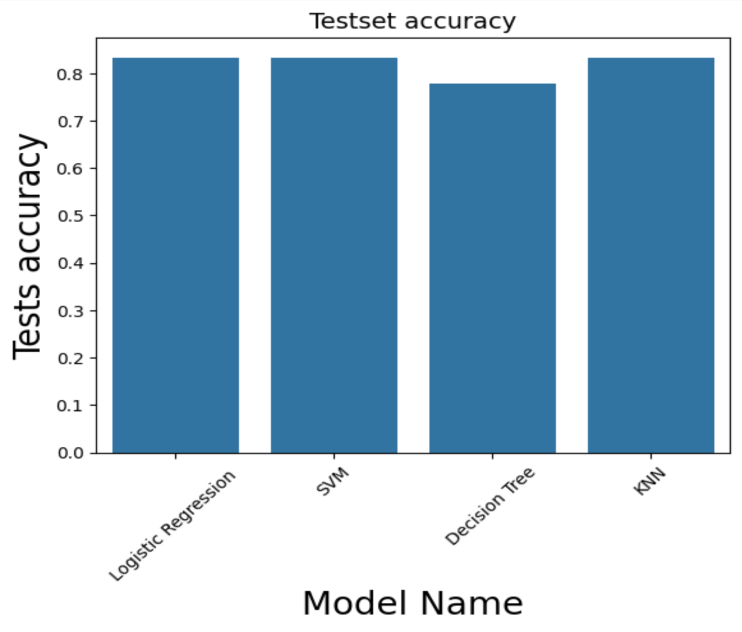
Payload range (Kg):
0
2500
5000
7500
Correlation between Payload and Success for site KSC LC-39A
class
2
1.5
1
0.5
0
3000
3500
4000
4500
5000
Booster Version Category
FT
B4
B5
Activate Windows

Section 5

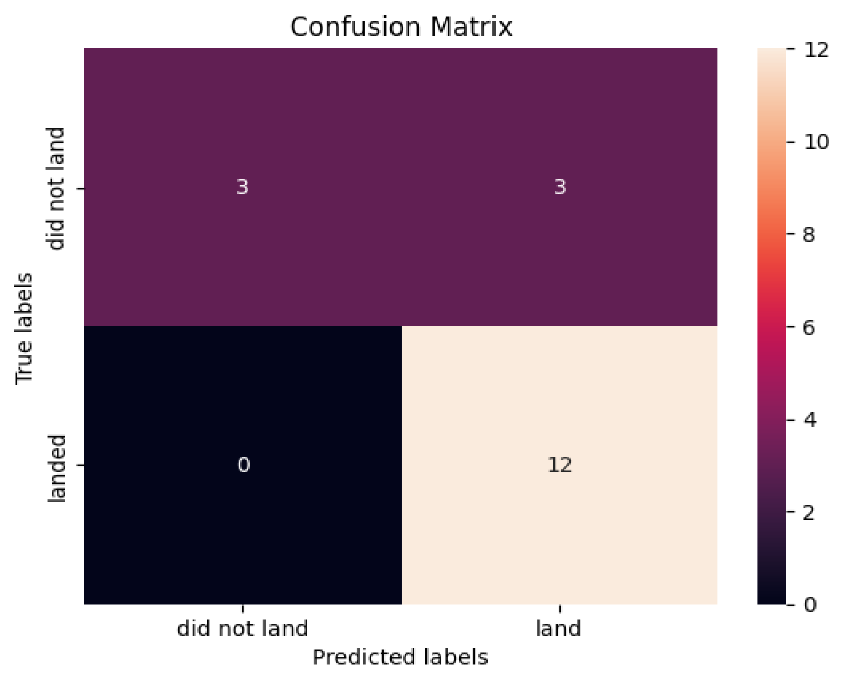
Predictive Analysis (Classification)

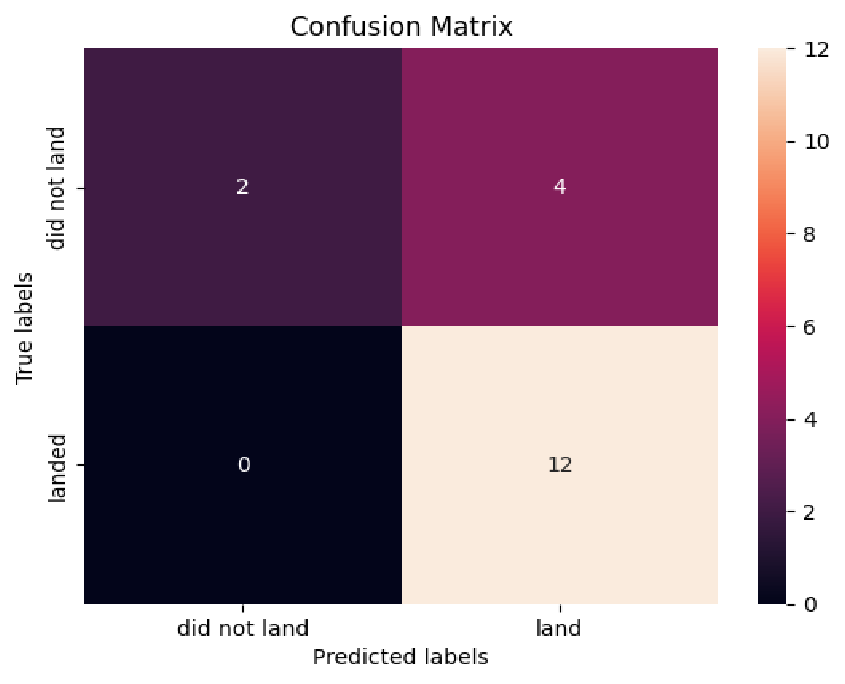
Classification Accuracy





Confusion Matrix





Logistic regression, SVM and KNN

Decision Tree

Conclusions

- KSC LC-39A stands out as the top-performing launch site, both in number of successful launches and in success rate (76.9%).
- The most successful payload range is 3000–4000 kg, while the Falcon 9 FT booster version achieved the highest success rate.
- Among the four models analyzed, Logistic Regression, SVM, and KNN exhibit similar performance in terms of classification results, as shown by their confusion matrices. All three models achieved comparable predictions for successful landings and made a similar number of misclassifications for failures. While none of them delivers a perfect solution, they represent some of the most effective options available given the current data and modeling context.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!