IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

NGUYEN MINH LONG

2023/12/29

Outline

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

Executive Summary

- This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- Problems you want to find answers :
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be fulfilled to ensure a successful landing program.

Introduction

- Predict whether the Falcon 9 first stage will land successfully or not.
- Purpose of this project
  - Determine if the first stage will land.
  - Determine the cost of a launch.
  - Information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Collected data will be guaranteed to be in the correct format through the API.
    - Example: Booster names – https://api.spacexdata.com/v4/rockets/
- Perform data wrangling
    - The data also collects historical Falcon 9 launch records from a Wikipedia page titled ”List of Falcon 9 and Falcon Heavy launches“.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - "Logistic Regression", "SVM", "Decision Tree", "KNN“ & Confusion Matrix.

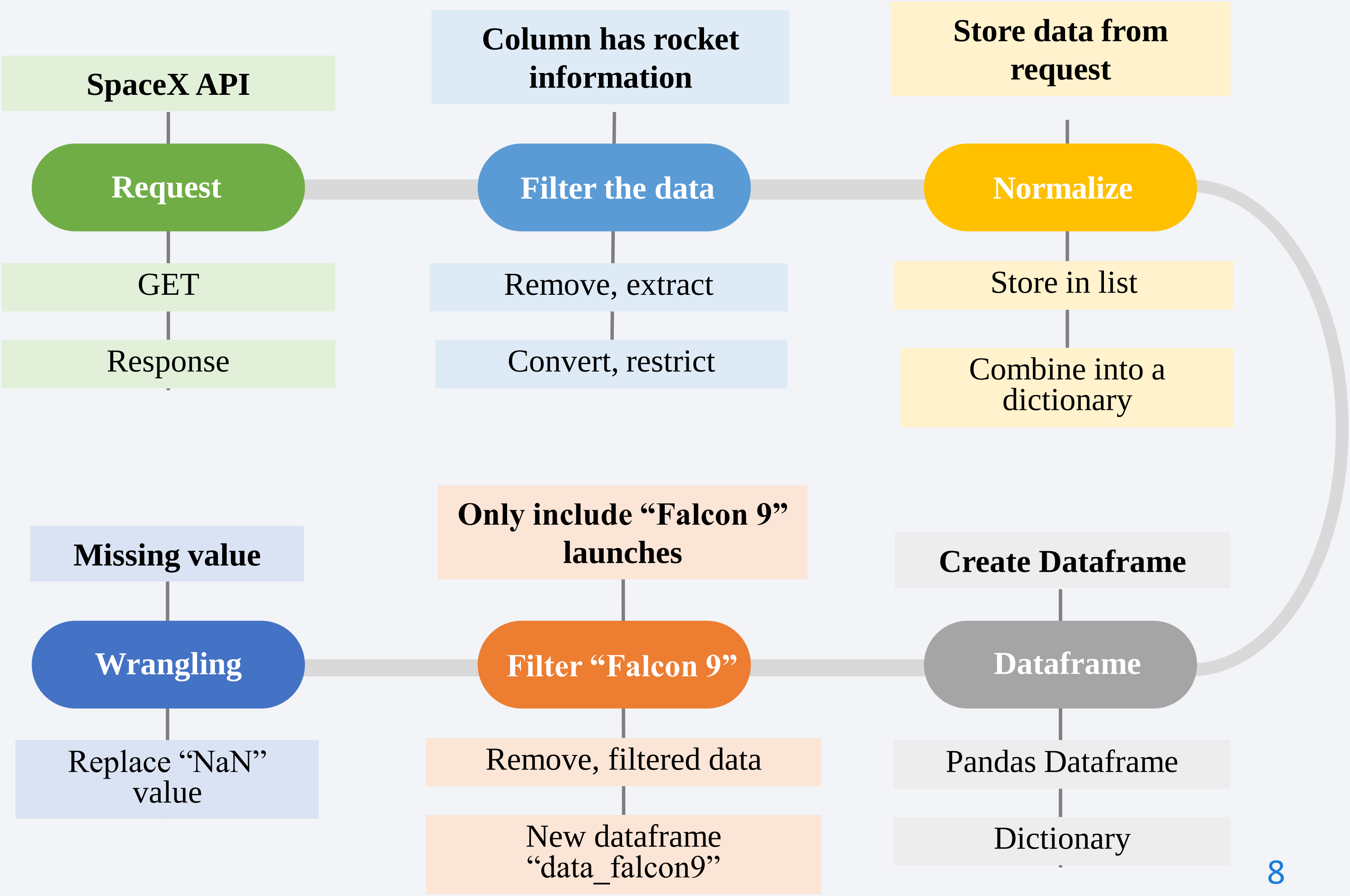
Data Collection

- Data collection process involved a combination of API requests from SpaceX public API and web scraping data from a table in SpaceX's Wikipedia entry.
- SpaceX API Data Columns:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins.
  - Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Wikipedia Webscrape Data Columns:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

- Data Collection – SpaceX API Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%201/jupyter-labs-spacex-data-collection-api.ipynb



SpaceX API

Request

GET

Response

Column has rocket information

Filter the data

Remove, extract

Convert, restrict

Store data from request

Normalize

Store in list

Combine into a dictionary

Create Dataframe

Dataframe

Pandas Dataframe

Dictionary

Only include “Falcon 9” launches

Filter “Falcon 9”

Remove, filtered data

New dataframe “data_falcon9”

Missing value

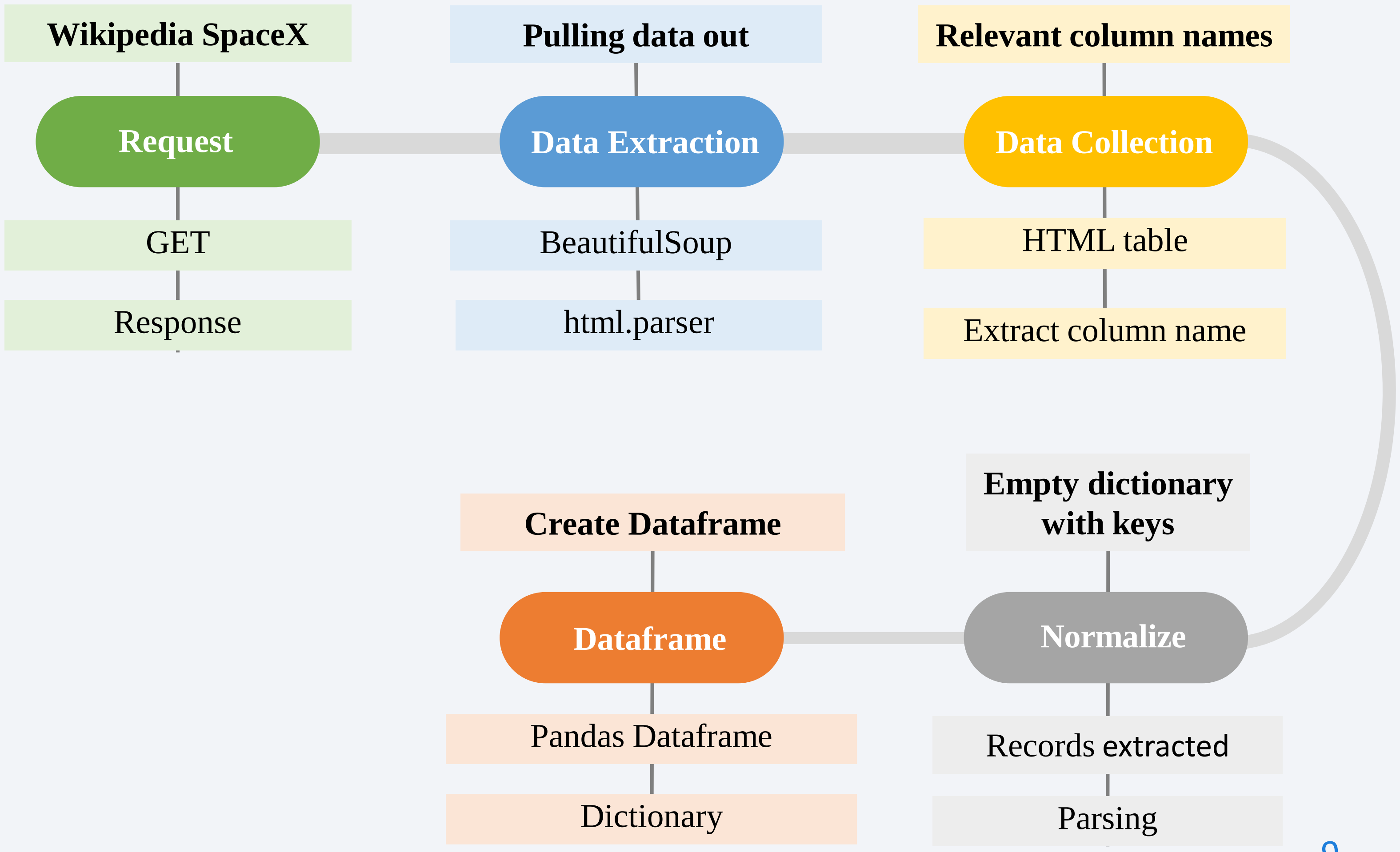
Wrangling

Replace “NaN” value

Data Collection - Scraping

- Data Collection – Scraping Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%201/jupyter-labs-webscraping.ipynb



Wikipedia SpaceX

Request

GET

Response

Pulling data out

Data Extraction

BeautifulSoup

html.parser

Relevant column names

Data Collection

HTML table

Extract column name

Empty dictionary with keys

Normalize

Records extracted

Parsing

Create Dataframe

Dataframe

Pandas Dataframe

Dictionary

Data Wrangling

- There are several different cases where the booster did not land successfully (True Ocean, True RTLS, …).
  - We will mainly convert into Training Labels
    - “1” means the booster successfully landed (True Ocean, True RTLS, …).
    - “0” means it was unsuccessful (False Ocean, False RTLS, …).
- For each value of the Outcome column, we will assign a value of 0 or 1.
- Finally, combine the sequence 0 and 1 into the dataset.
  - Example: Value “True Ocean” will correspond to the value "1" of the new "Class" column.

Data Wrangling Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%201/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- In this project, I use scatter plot, bar plot, and line plot. Most of which use scatter plot.
- Determining a successes or failed landing depends on many factors, so using a scatter plot is most appropriate when comparing the relationship between two variables.
  - Example: FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, ...

EDA with data visualization Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%202/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- In this project, the SQL statements used include:
  - Loaded data set into IBM DB2 Database.
  - Queried using SQL Python integration.
  - Queries were made to get a better understanding of the dataset.
  - Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes.

EDA with SQL Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium maps mark all launch sites, using Circle and Marker. Also including successful or failed launches at each Launch Site. Finally, there are examples of proximity to key locations: Coastline, Railway, Highway, and City using Polyline.
- Finding an optimal location for building a launch site certainly involves many factors.

Interactive map with Folium Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%203/lab_jupyter_launch_site_location.ipynb

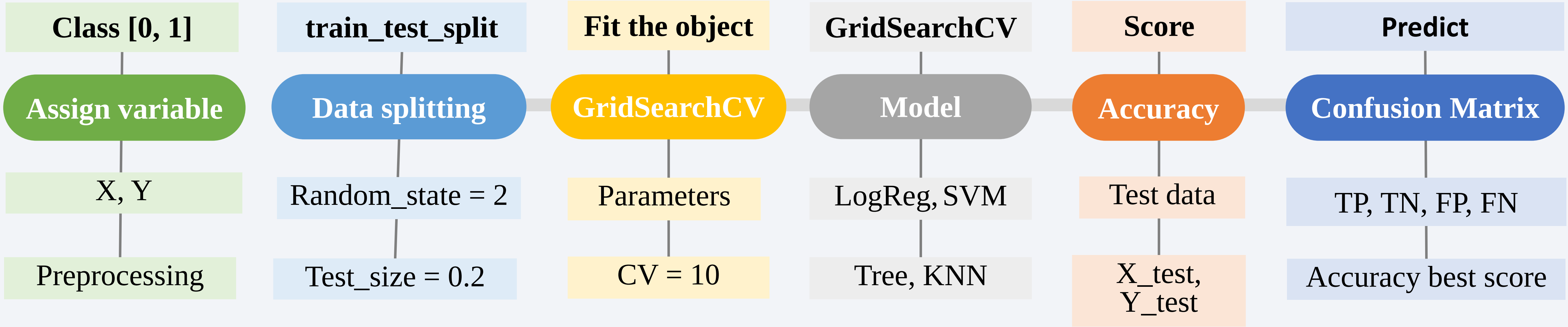
Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
  - A pie chart can be selected to show the distribution of successful landings across all launch site. It can also be used to display the number of successes or failures (in percentage) of each launch site.
  - Scatter plot takes two inputs: all launch site or each launch site with payload. Payload mass can be adjusted to a value within a range of 0 – 10,000 kg.
- The pie chart is used to visualize launch site success rate.
- Scatter charts can help show how success varies by launch site, payload volume, and booster version category.

Dashboard

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%203/spacex_dash_app.py

Predictive Analysis (Classification)

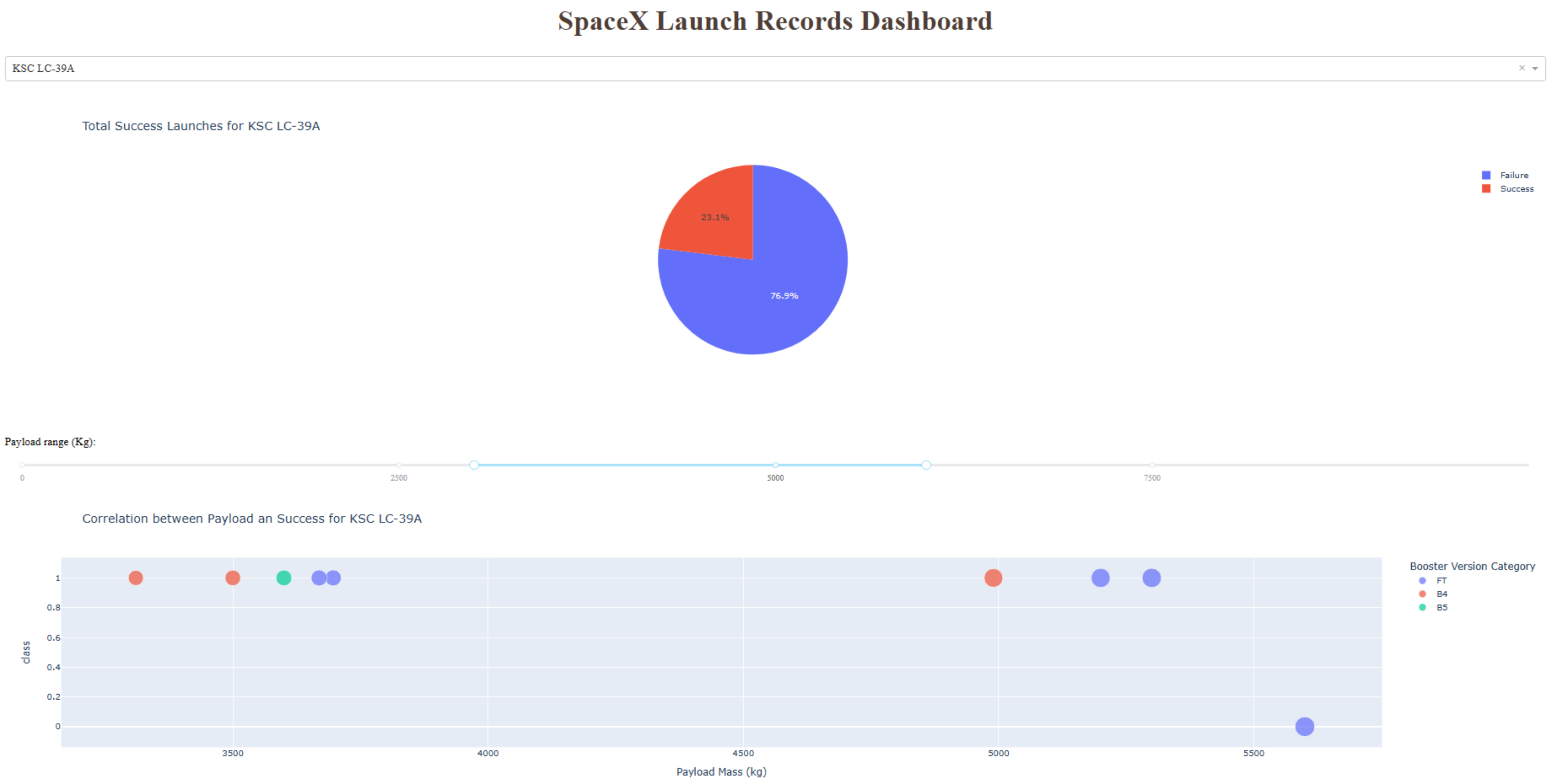


Class [0, 1]	**train_test_split**	**Fit the object**	**GridSearchCV**	**Score**	**Predict**
Assign variable	**Data splitting**	**GridSearchCV**	**Model**	**Accuracy**	**Confusion Matrix**
X, Y	Random_state = 2	Parameters	LogReg, SVM	Test data	TP, TN, FP, FN
Preprocessing	Test_size = 0.2	CV = 10	Tree, KNN	X_test, Y_test	Accuracy best score

Predictive analysis Notebook

https://github.com/Nguyen-Minh-Long/IBM_Data_Science/blob/main/10.%20Applied%20Data%20Science%20Capstone/Week%204/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

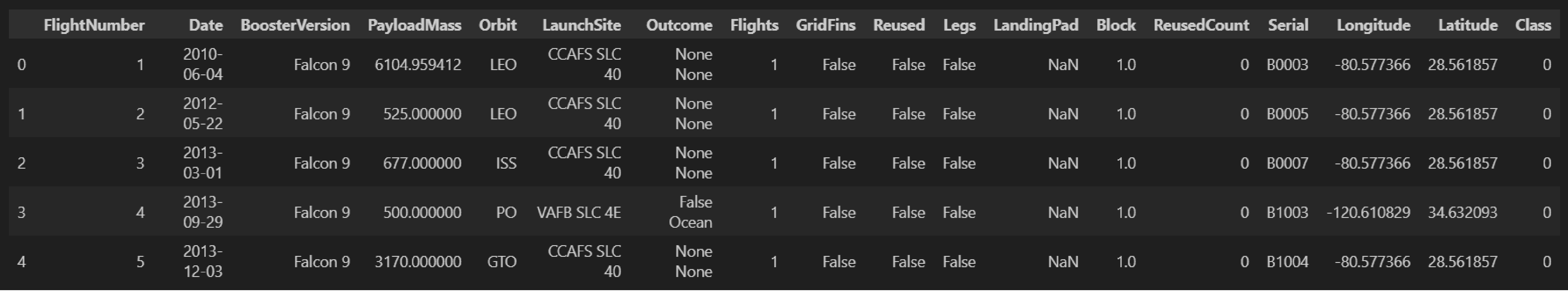


SpaceX Launch Records Dashboard

KSC LC-39A

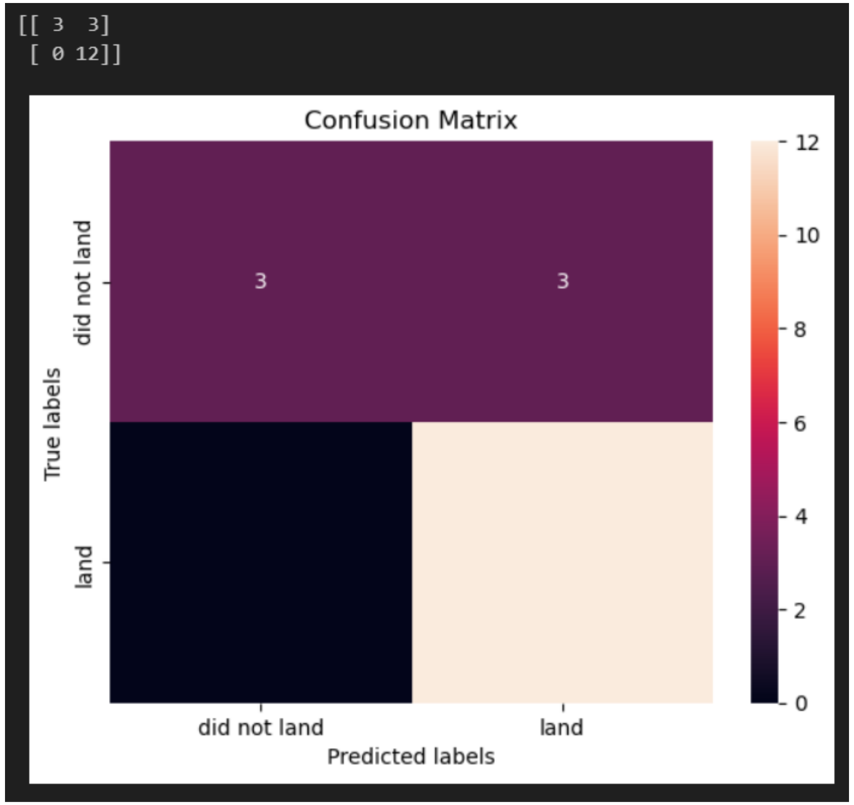
Payload range (Kg):

Figure 1: Interactive analytics



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Figure 2: Exploratory data analysis results



```
[[ 3  3]
 [ 0 12]]
```

Figure 3: Predictive analysis results – Logistic Regression

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

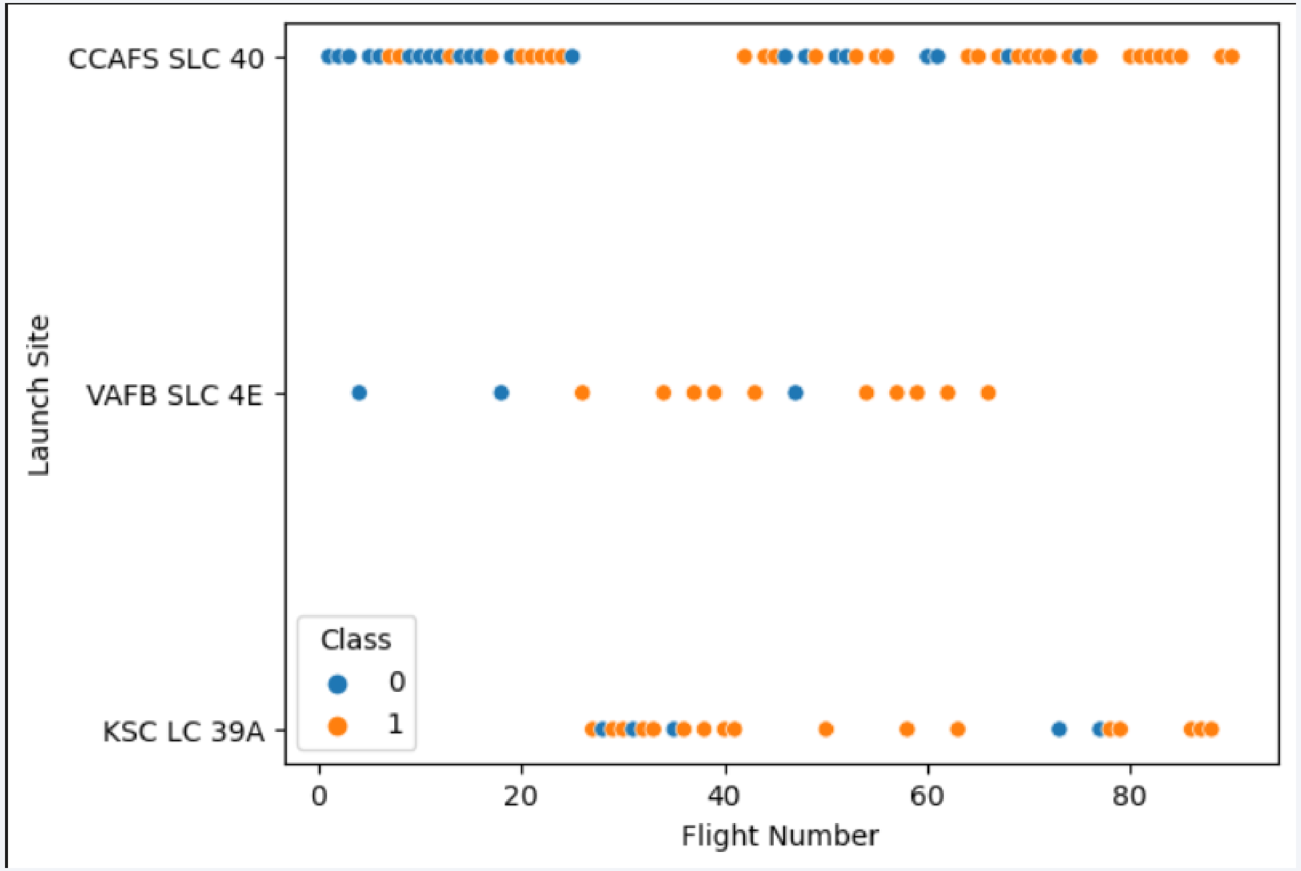


Figure 4: Scatter Plot Flight Number vs. Launch Site

- It seems that as the number of flights increases, the success rate also increases.
- Especially from the 20th flight onwards is a breakthrough period, significantly increasing the success rate.
- In addition, CCAFS SLC 40 is the place with the most launches.

Payload vs. Launch Site

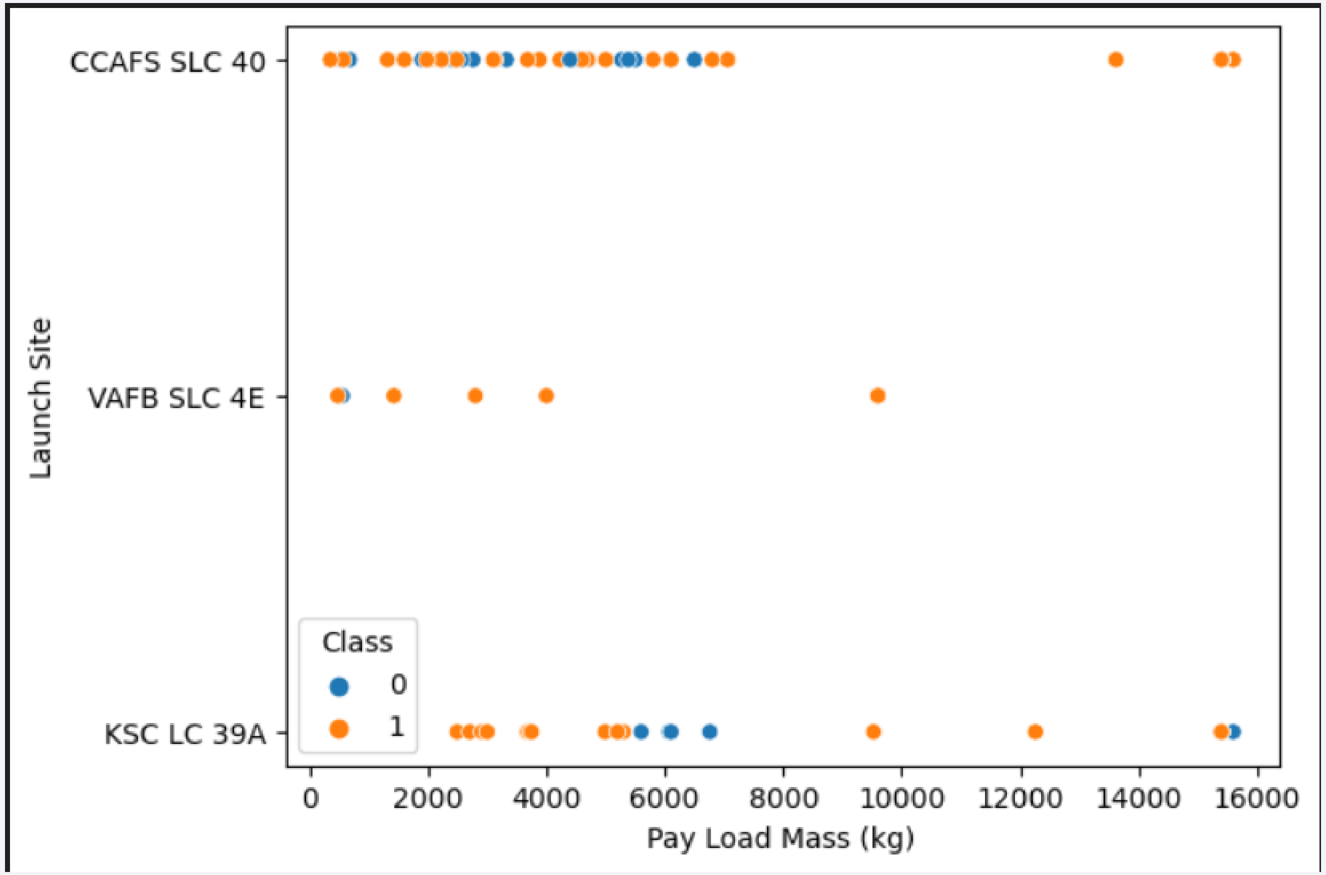


Figure 5: Scatter Plot Payload vs. Launch Site

- Most of the payload mass is in the range of 0 – 8,000 kg.
- The VAFB-SLC Launch Site there are no rockets launched for heavy payload mass (greater than 10,000).

Success Rate vs. Orbit Type

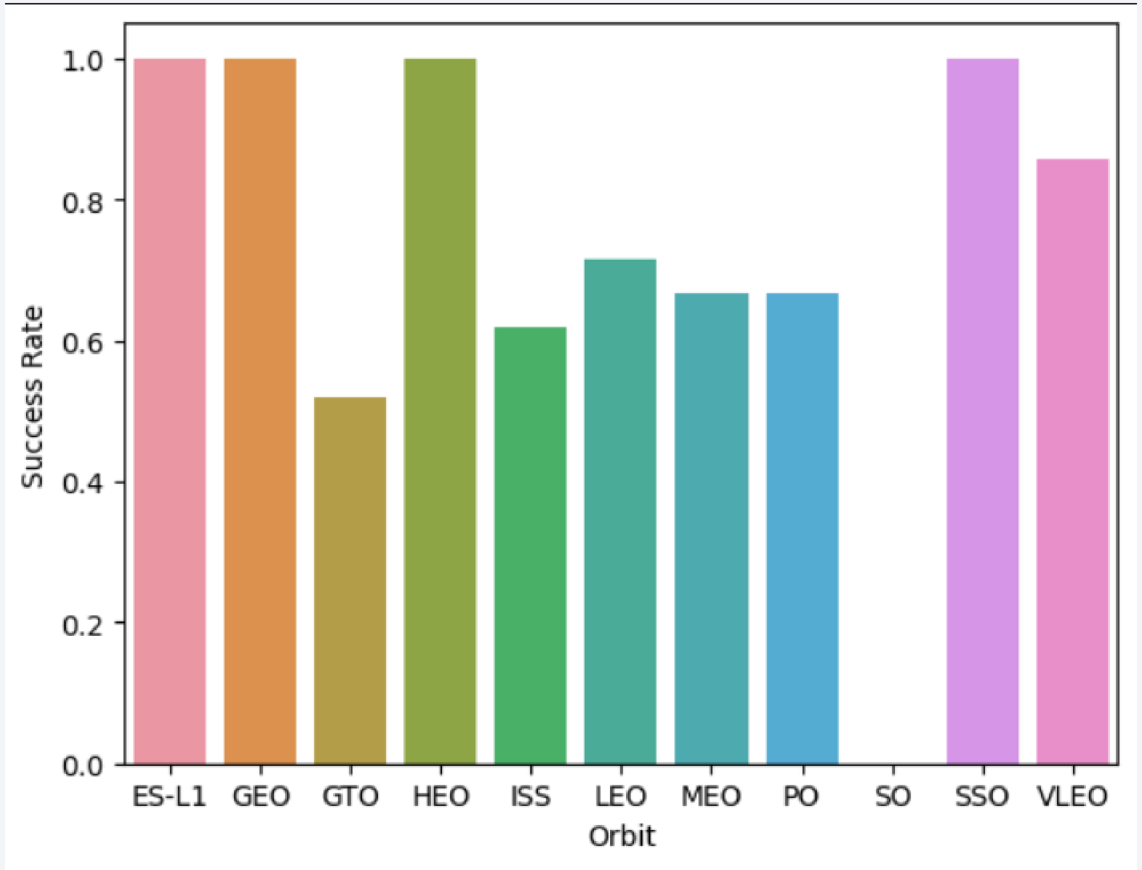


Figure 6: Bar Plot Success Rate vs. Orbit Type

- Orbits with a success rate of up to 100% ES-L1, GEO, HEO, SSO.
- VLEO has decent success rate and attempts, about 86%.
- GTO has the around 50%. success rate but largest sample
- SO has 0% success rate.

Flight Number vs. Orbit Type

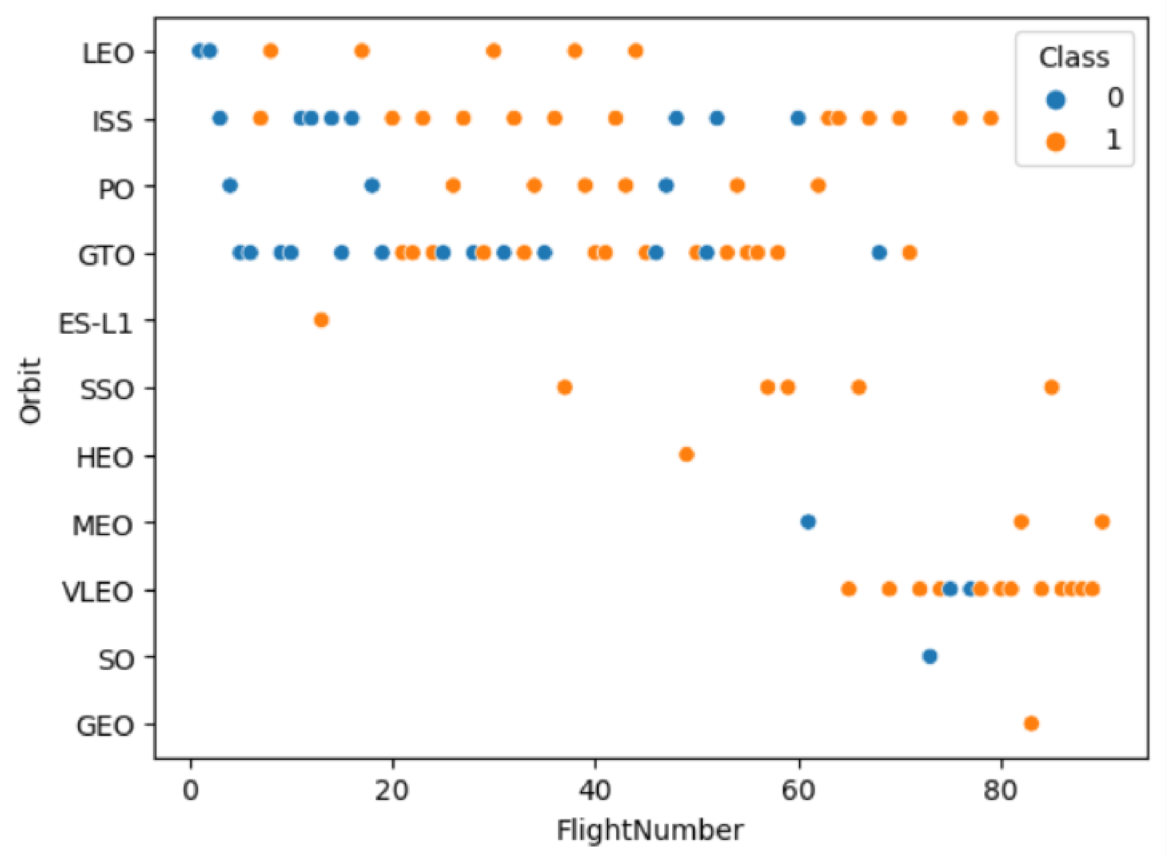


Figure 7: Scatter Plot Flight Number vs. Orbit Type

- 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

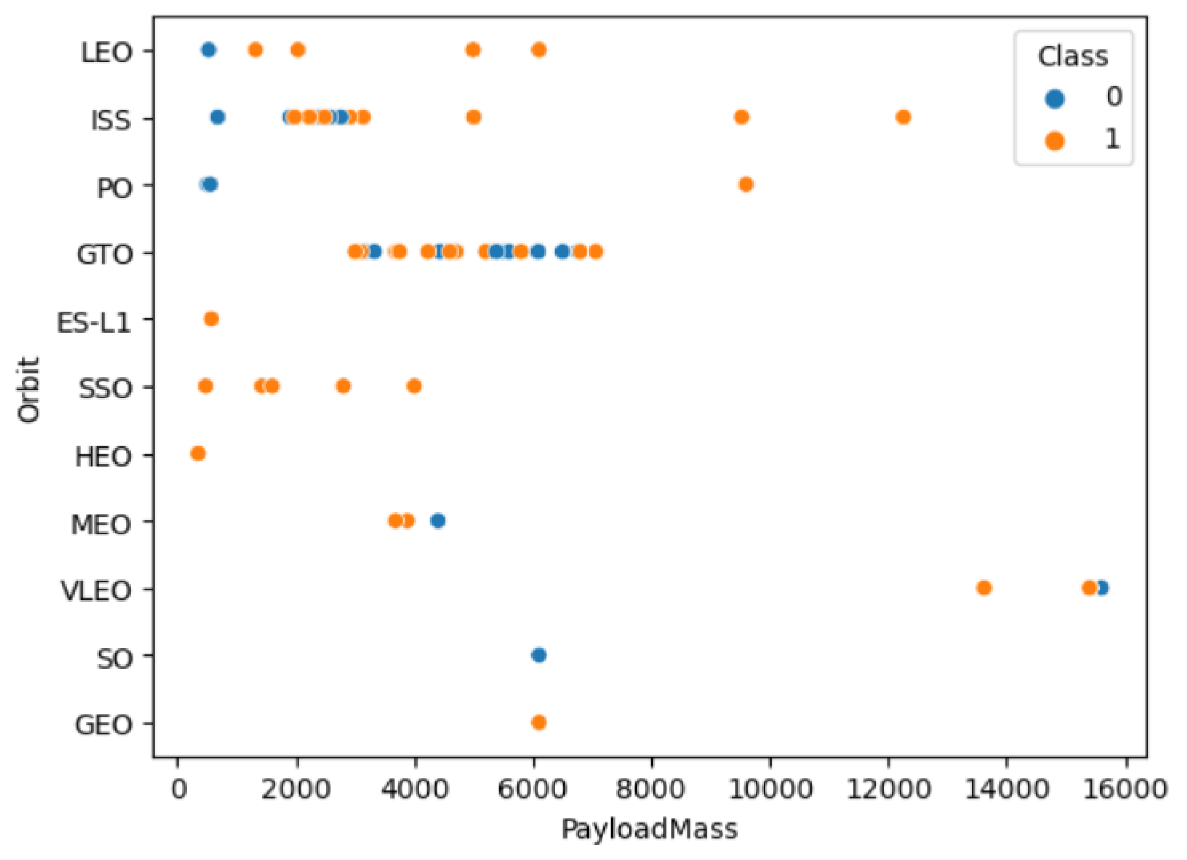


Figure 8: Scatter Plot Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

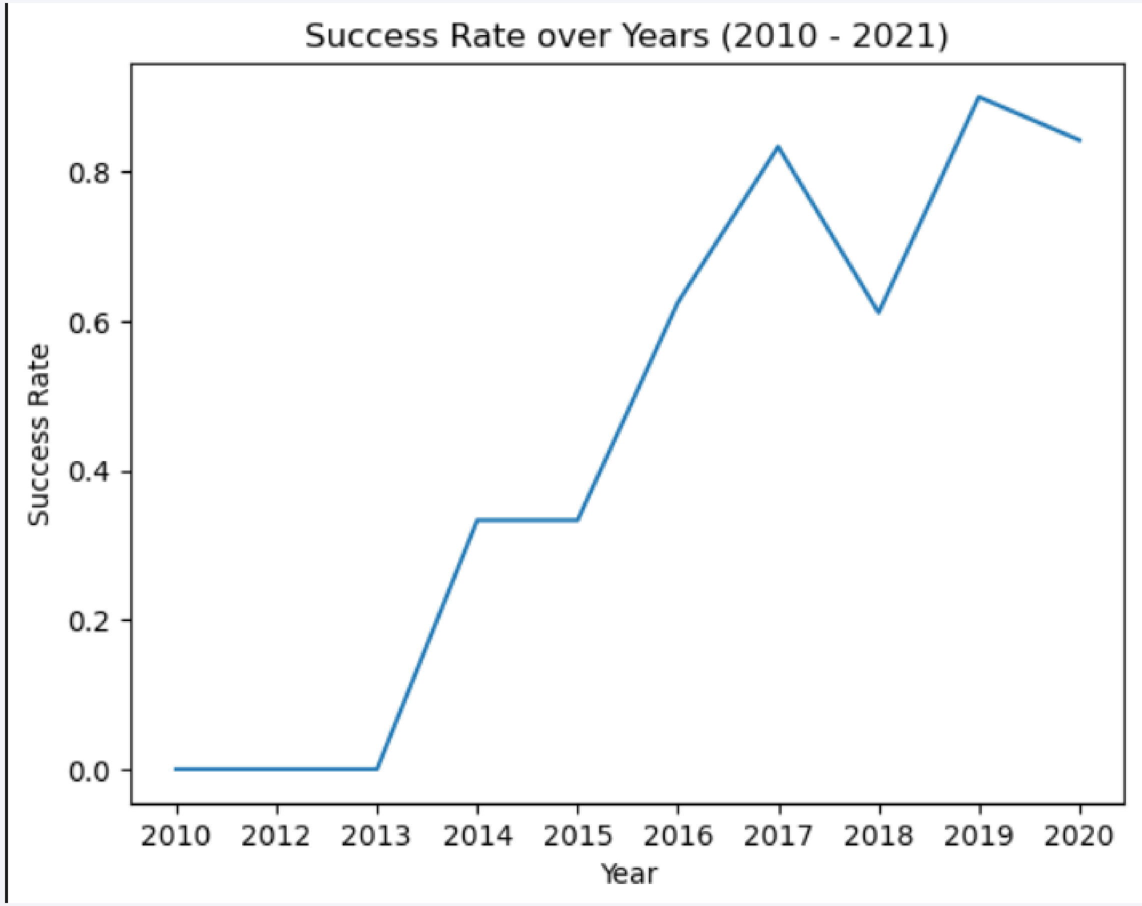
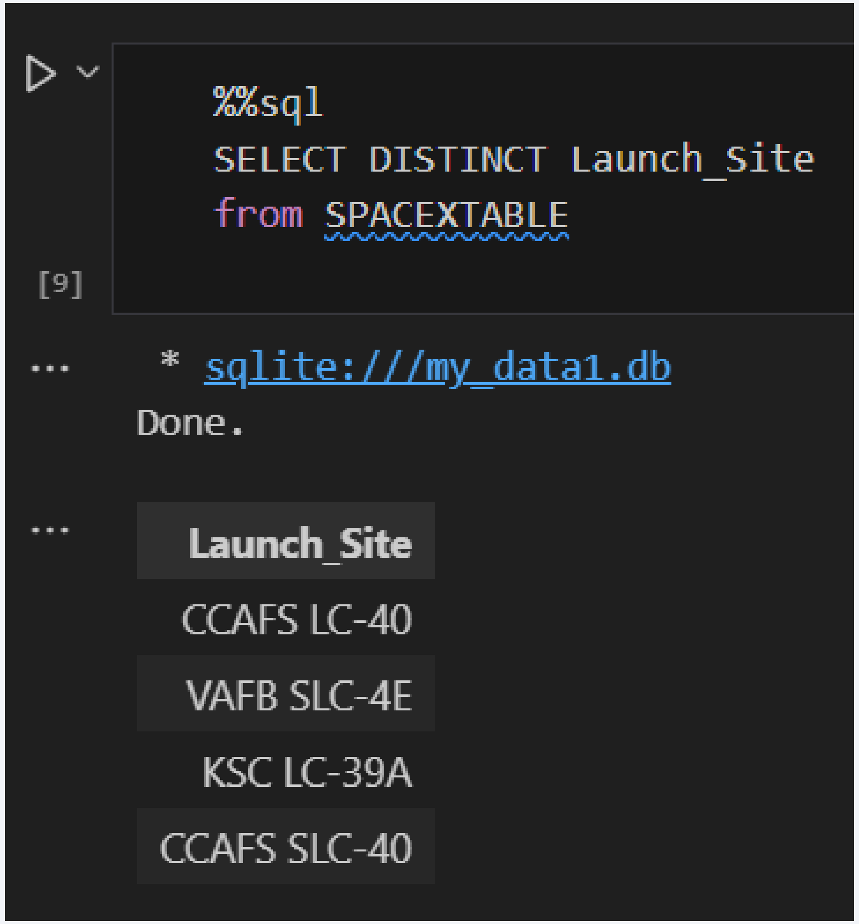


Figure 9: Line Plot Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2017 (stable in 2014).
- After 2015 it started increasing.
- Slight decrease in 2018.

All Launch Site Names



```
%%sql
SELECT DISTINCT Launch_Site
from SPACEXTABLE
```

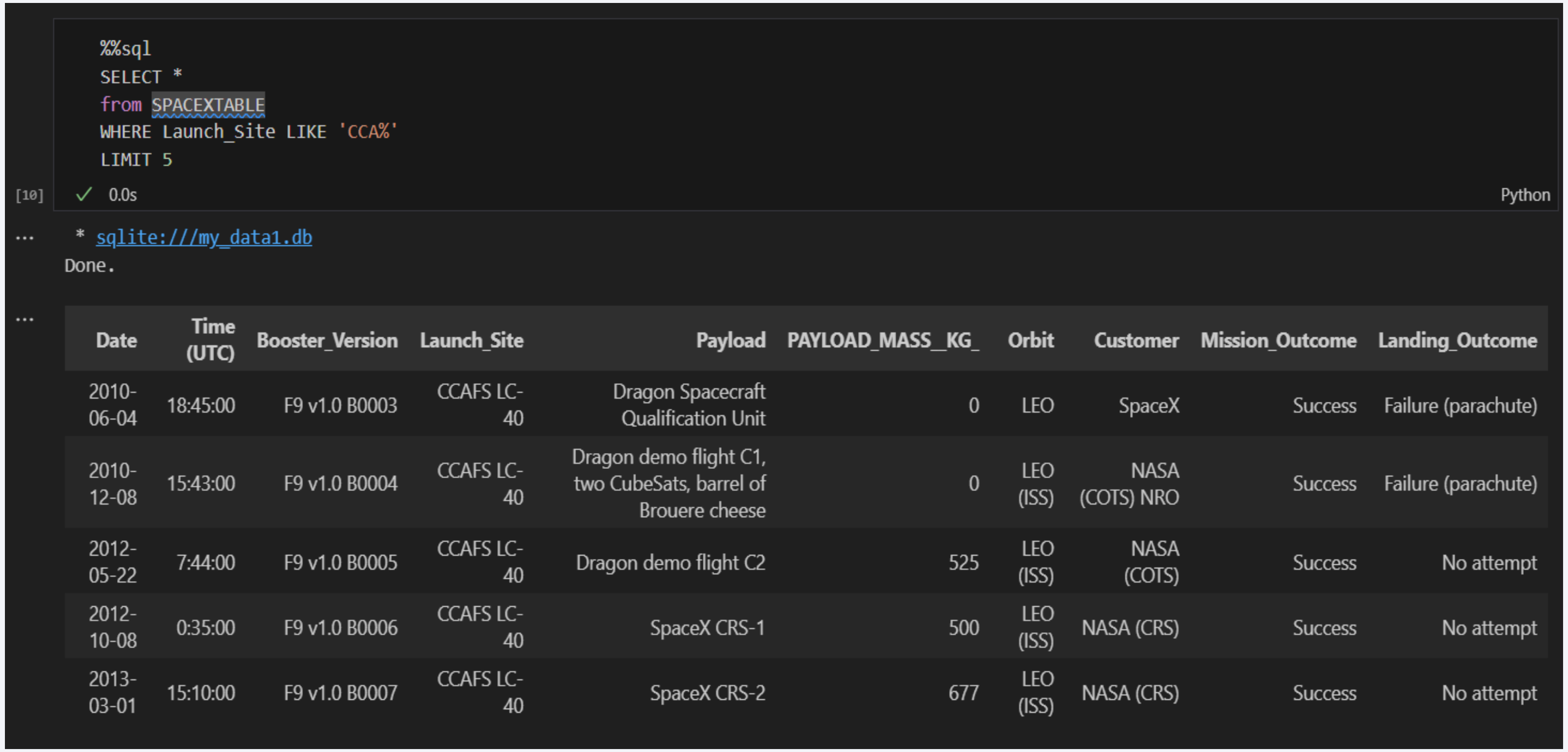
```
[9]
 * sqlite:///my_data1.db
Done.
```

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

Figure 10: All Launch Site Names

- The SELECT DISTINCT statement is used to return only distinct (different) values.
- The result has 4 values:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'



```
%%sql
SELECT *
from SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

```
[10]  ✓ 0.0s                                                    Python
```

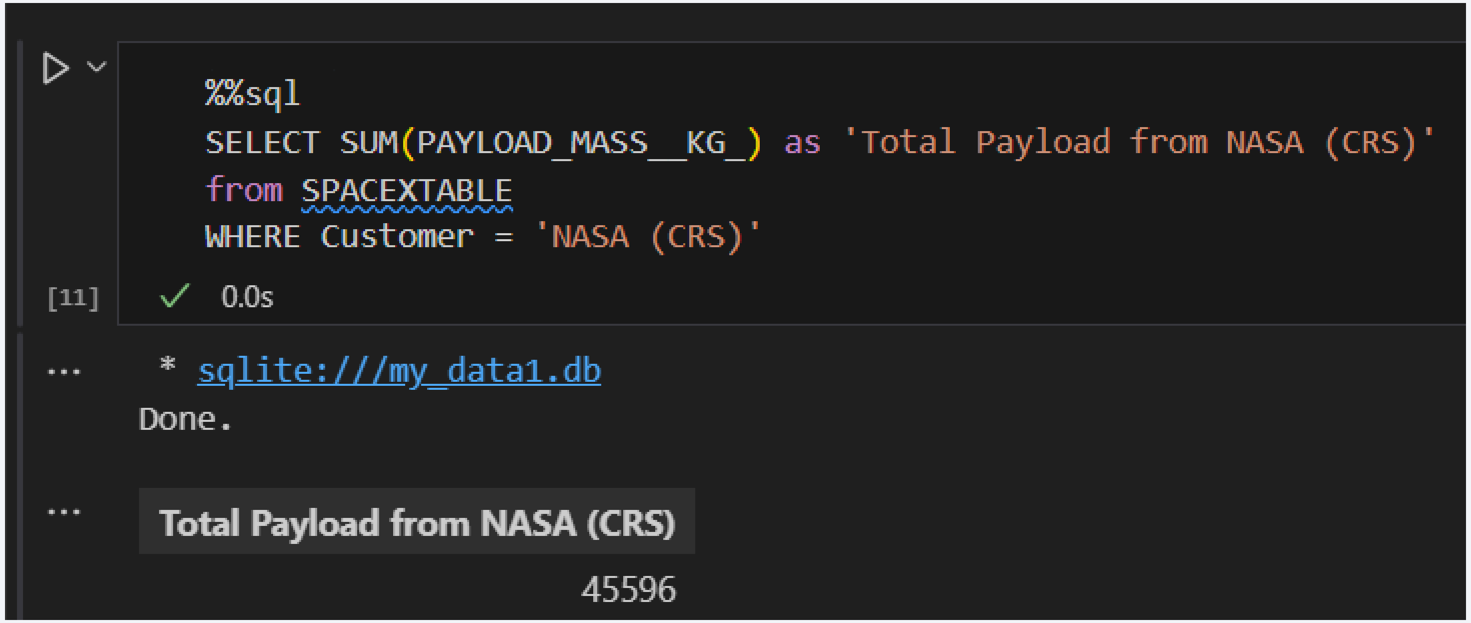
```
* sqlite:///my_data1.db
Done.
```

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

Figure 11: Record with Launch Site Names Begin with 'CCA'

- Use the LIKE Operator to find strings starting with “CCA”.
- Limit the results displayed using the LIMIT Operator.

Total Payload Mass



```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as 'Total Payload from NASA (CRS)'
from SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

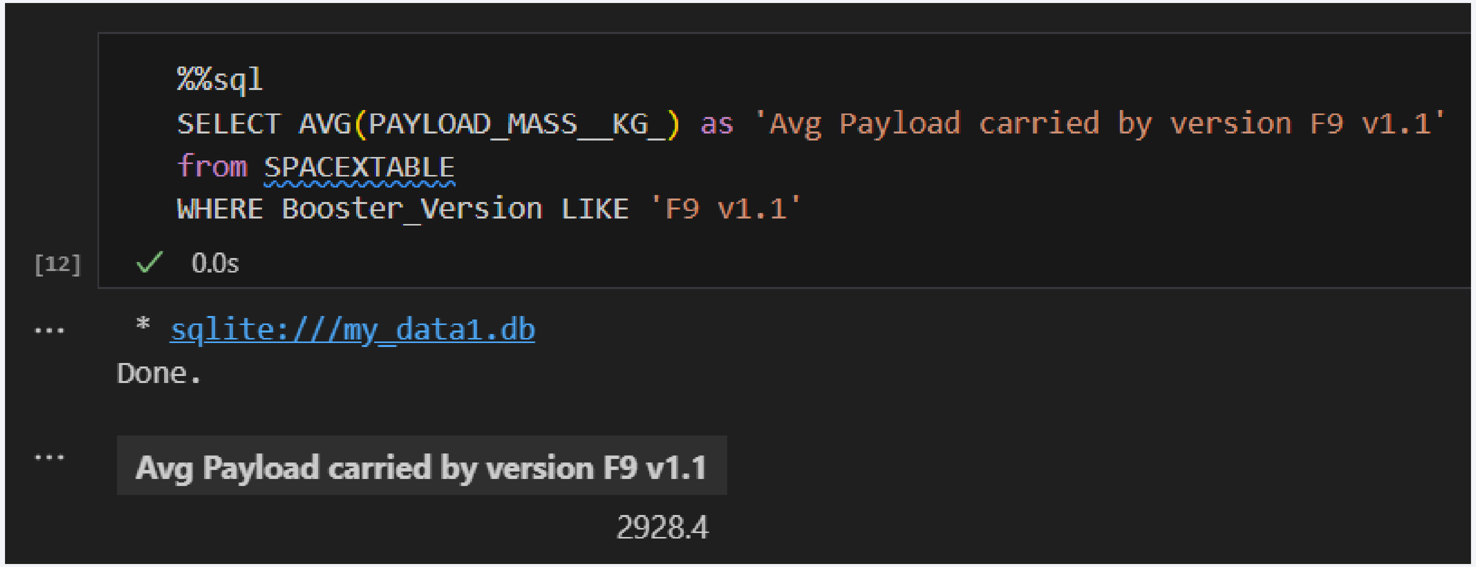
```
 * sqlite:///my_data1.db
Done.
```

Total Payload from NASA (CRS)
45596

Figure 12: Total Payload Mass

- The SUM() function returns the total sum of a numeric column.
- The result is 45,596 Kg.

Average Payload Mass by F9 v1.1



```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as 'Avg Payload carried by version F9 v1.1'
from SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1'
```

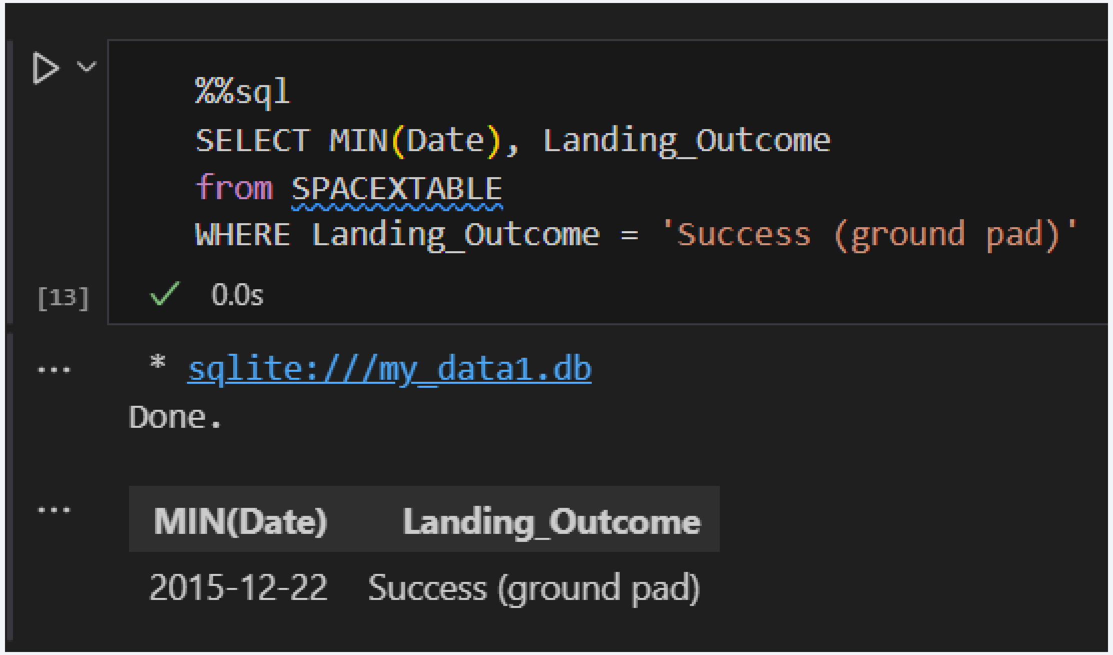
```
[12]  ✓  0.0s

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

Avg Payload carried by version F9 v1.1
2928.4

Figure 13: Average Payload Mass by F9 v1.1

- The AVG() function returns the average value of a numeric column.
- Like Operator used to search for booster version name “F9 v1.1”.

First Successful Ground Landing Date



```
%%sql
SELECT MIN(Date), Landing_Outcome
from SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
```

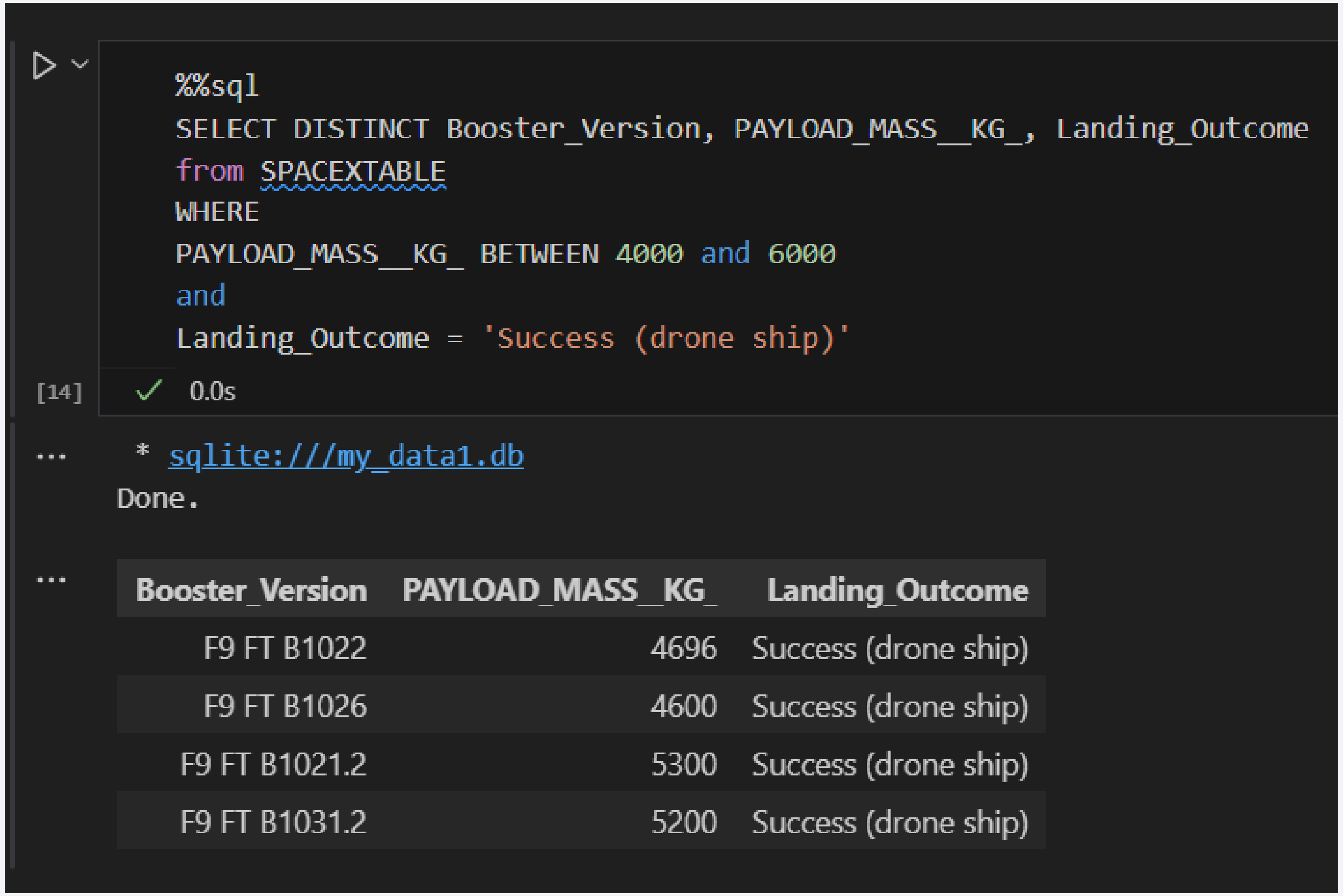
```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

Figure 14: First Successful Ground Landing Date

- The MIN() function returns the smallest value of the selected column.
- The condition here are successful landing “Success” with landing on the ground “ground pad”.
- The results obtained are December 12, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000



```
%%sql
SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_, Landing_Outcome
from SPACEXTABLE
WHERE
PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
and
Landing_Outcome = 'Success (drone ship)'
```

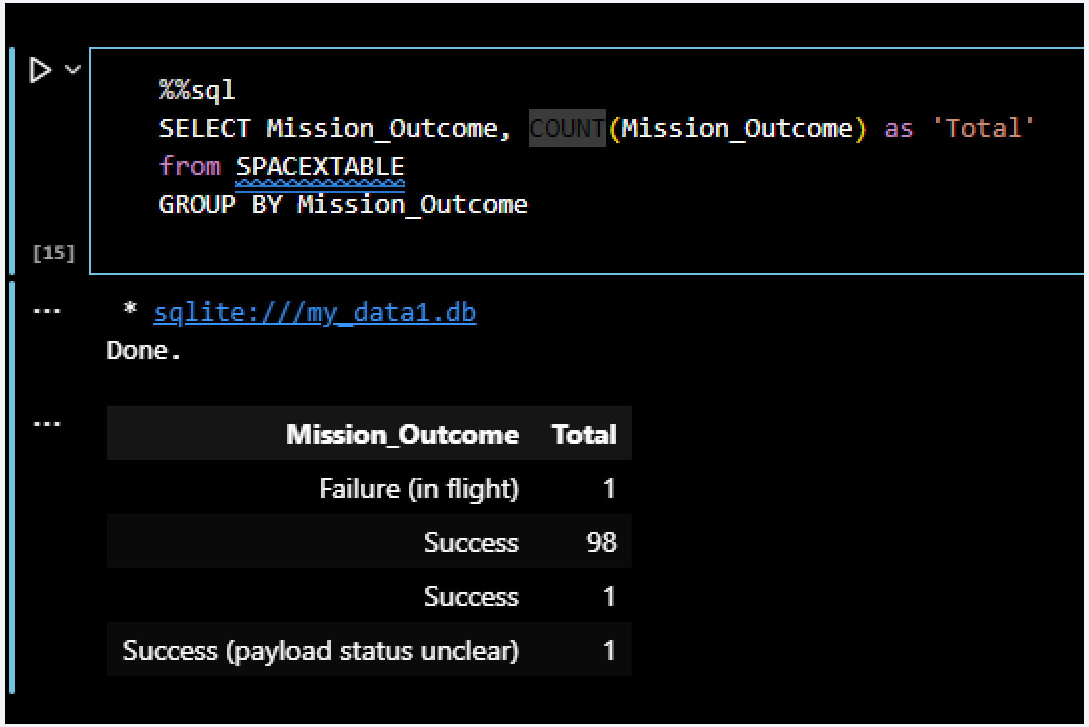
```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Figure 15: Successful Drone Ship Landing

- The AND operator is used to filter records based on more than one condition.
- First condition, have payload mass greater than 4,000 but less than 6,000.
- Second condition, landing success on the drone ship .

Total Number of Successful and Failure Mission Outcomes



```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) as 'Total'
from SPACEXTABLE
GROUP BY Mission_Outcome
```

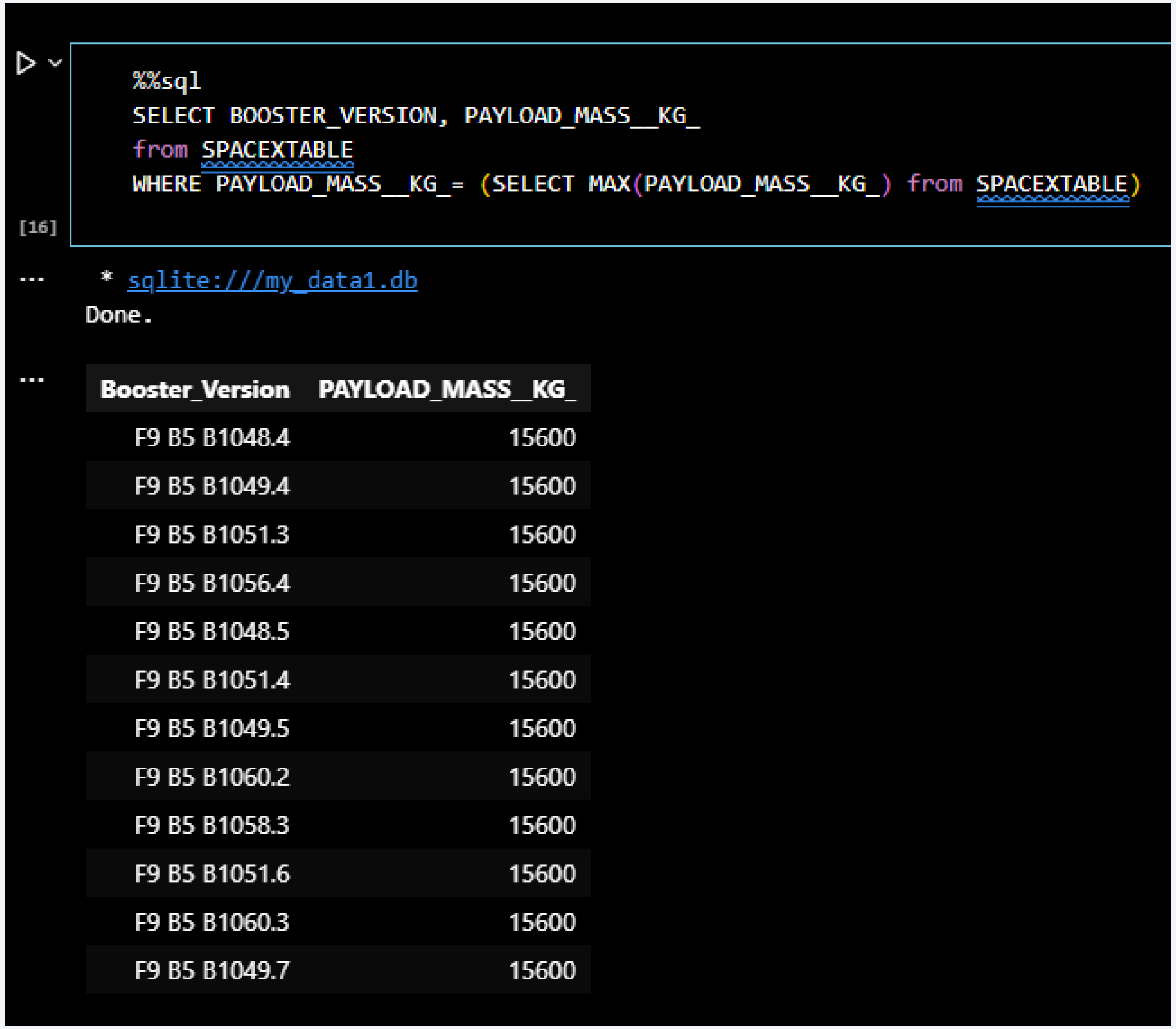
```
 * sqlite:///my_data1.db
Done.
```

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

Figure 16: Total Number of Successful and Failure

- The COUNT() function returns the number of rows that matches a specified criterion.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.

Boosters Carried Maximum Payload



```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_
from SPACEXTABLE
WHERE PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

[16]

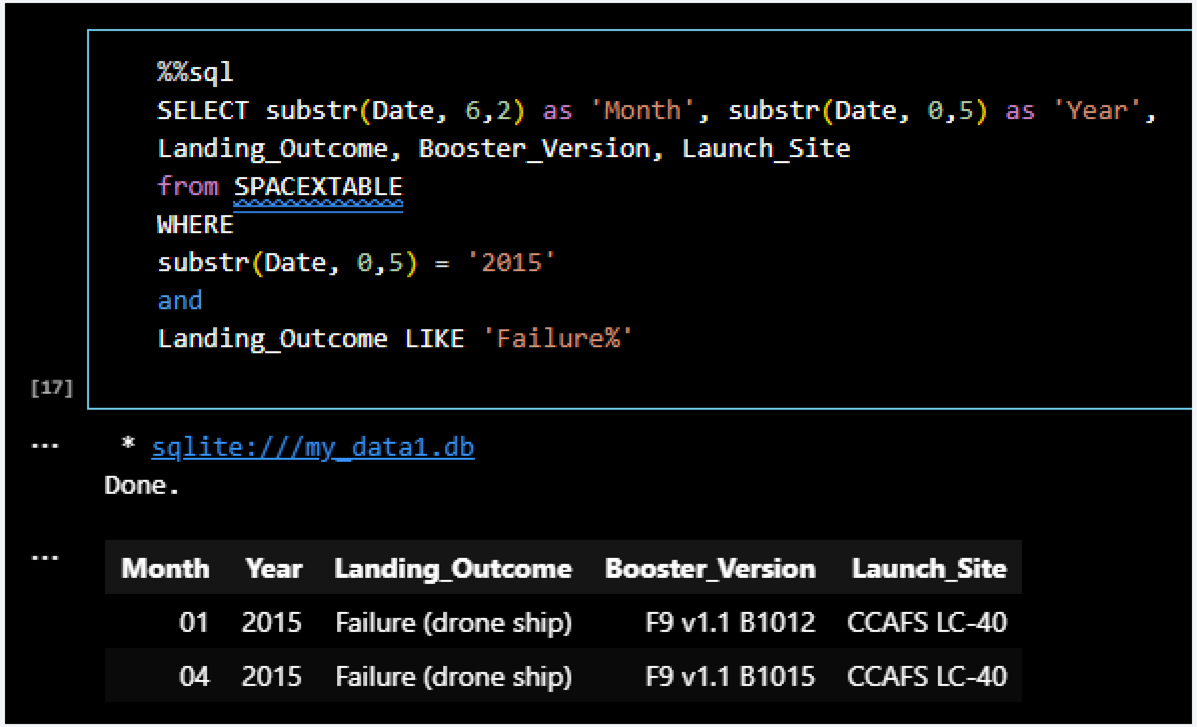
```
 * sqlite:///my_data1.db
Done.
```

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

Figure 17: Boosters Carried Maximum Payload

- By using subquery, the result obtained from subquery is 15,600 kg.
- Returning to the main clause, we will list booster names with payload mass reaching a maximum value of 15,600kg.

2015 Launch Records



```
[17]
%%sql
SELECT substr(Date, 6,2) as 'Month', substr(Date, 0,5) as 'Year',
Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
WHERE
substr(Date, 0,5) = '2015'
and
Landing_Outcome LIKE 'Failure%'
```

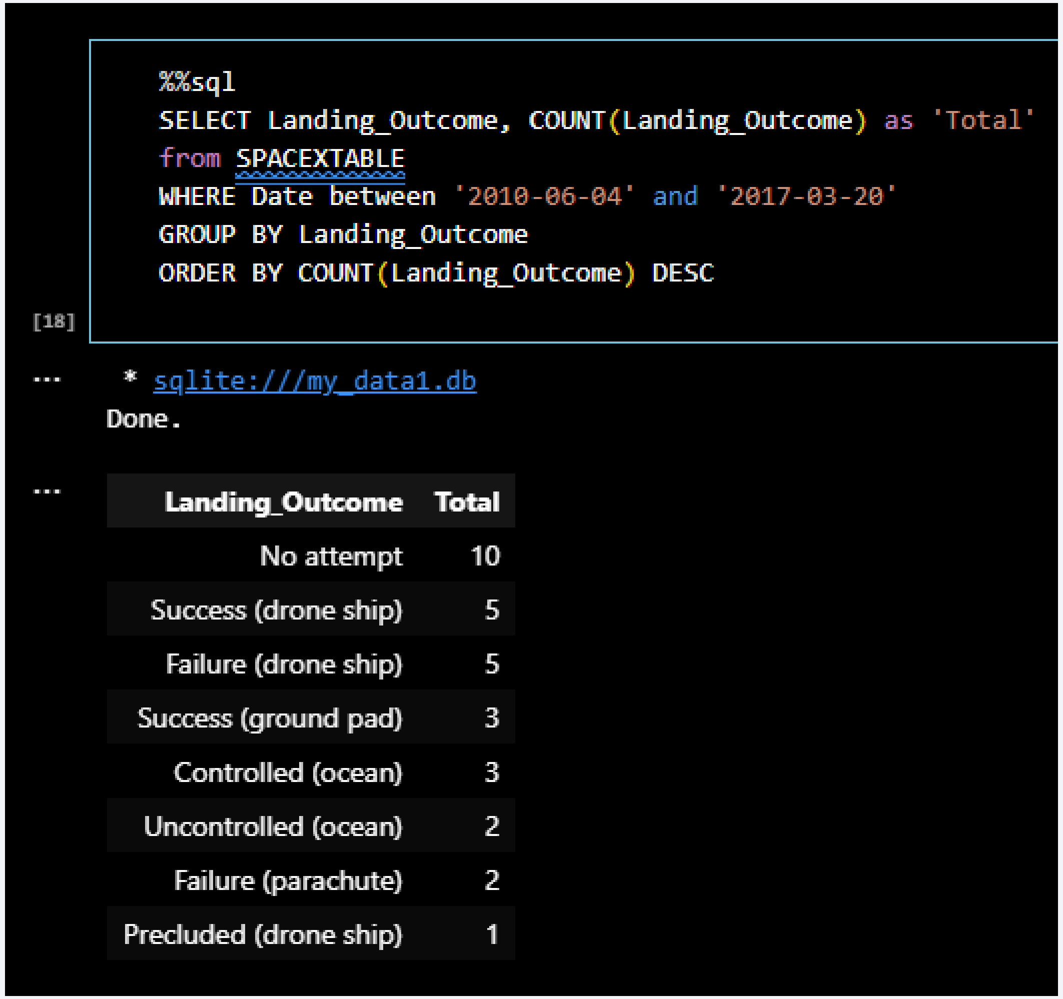
```
 * sqlite:///my_data1.db
Done.
```

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Figure 18: 2015 Launch Failure Records

- SQLite does not support month names. So, using substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.
- The first condition is that the launch must be in 2015.
- And the second condition is that the launch is classified as a failure.

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



[18]

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) as 'Total'
from SPACEXTABLE
WHERE Date between '2010-06-04' and '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) 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

Figure 19: Rank Landing Outcomes

- The BETWEEN operator selects values within a given range.
- The DESC command is used to sort the data returned in descending order.
- The condition here is that the experimental data collection date must be between June 4, 2010 and March 20, 2017.

Section 3

Launch Sites Proximities Analysis

Marked Launch Sites

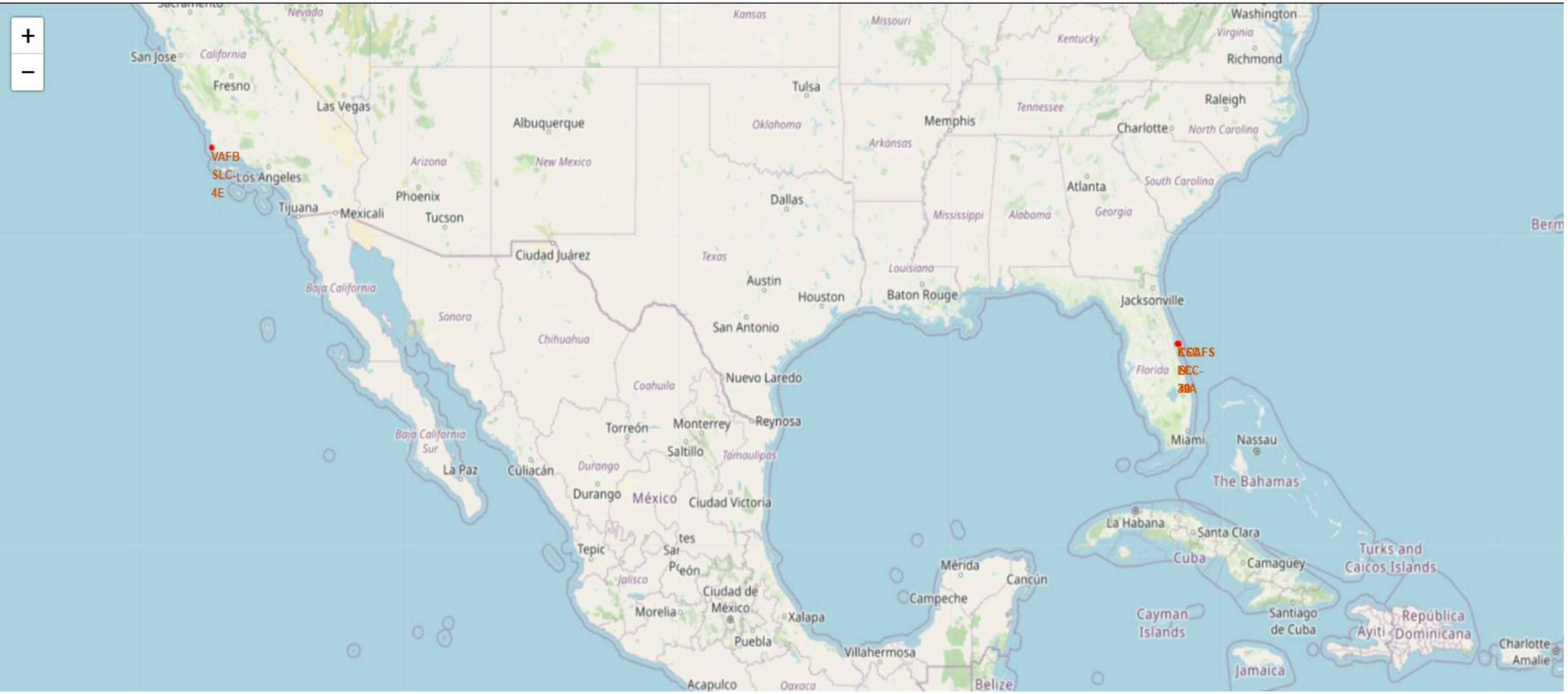
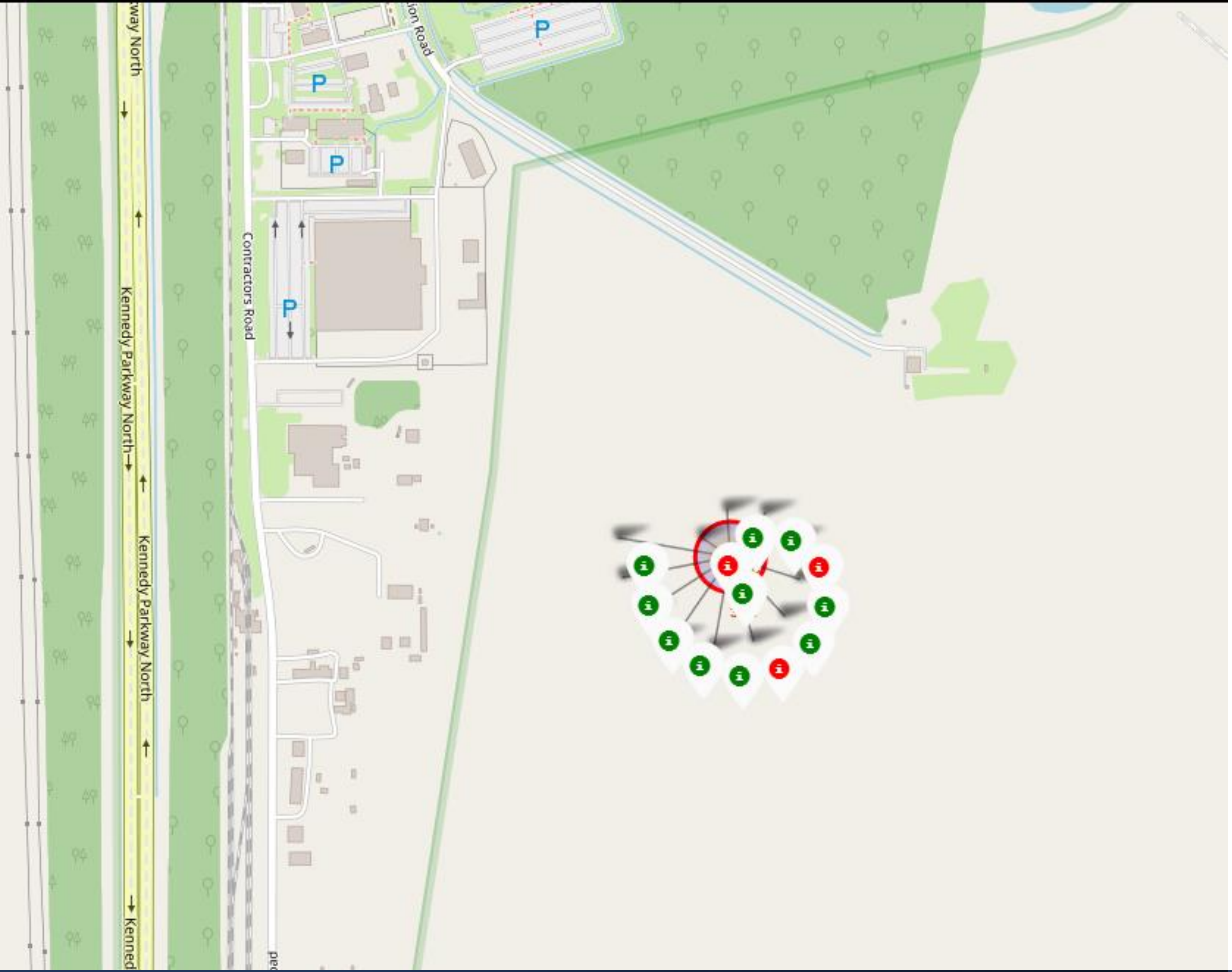


Figure 20: Folium map with marked launch sites

- There are a total of 4 launch sites in the map shown.
- Particularly, the 3 launch sites, “CCAFS LC-40”, “CCAFS SLC-40”, “KSC LC-39A”, are located quite close to each other, so the names of launch site are overlapped.



Color-labeled Launch Outcomes

- This is KSC LC-39A launch site.
- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

Figure 21: Folium map color-labeled launch outcomes

Proximities

- Using VAFB SLC-4E as an example, launch sites are very close to railways for large part and supply transportation (1,51 km).
- Also close to highways for human and supply transport (0,85 Km).
- And especially far away from city areas to avoid affecting the city and its people (14,15 km).

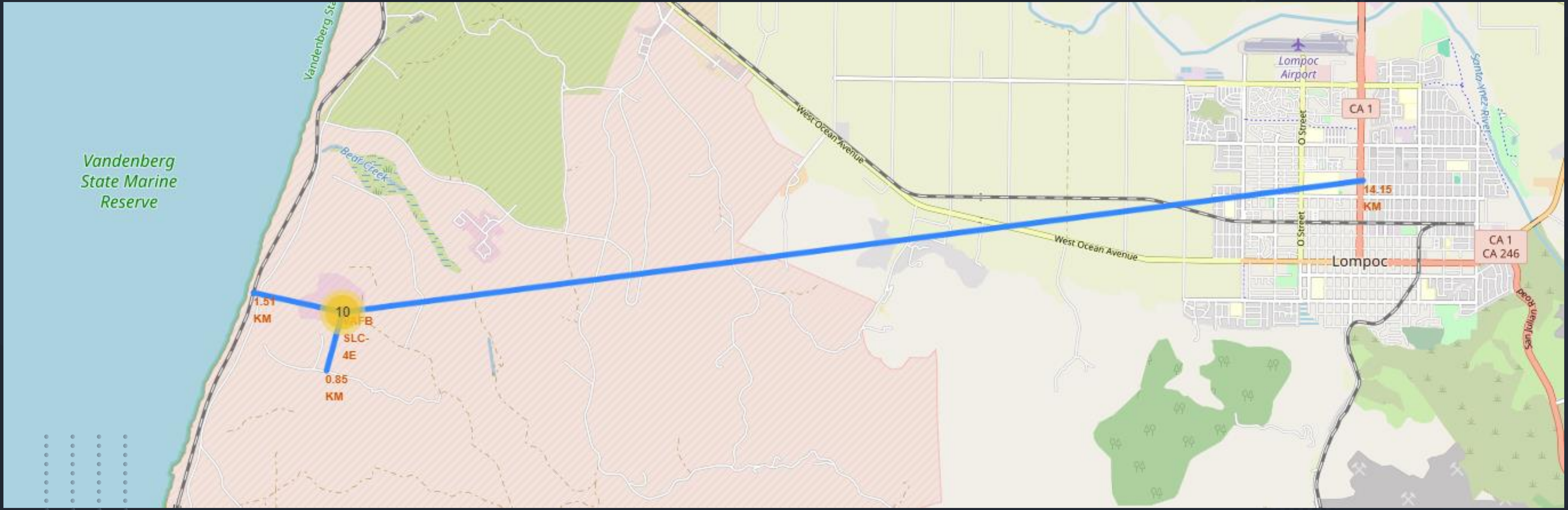
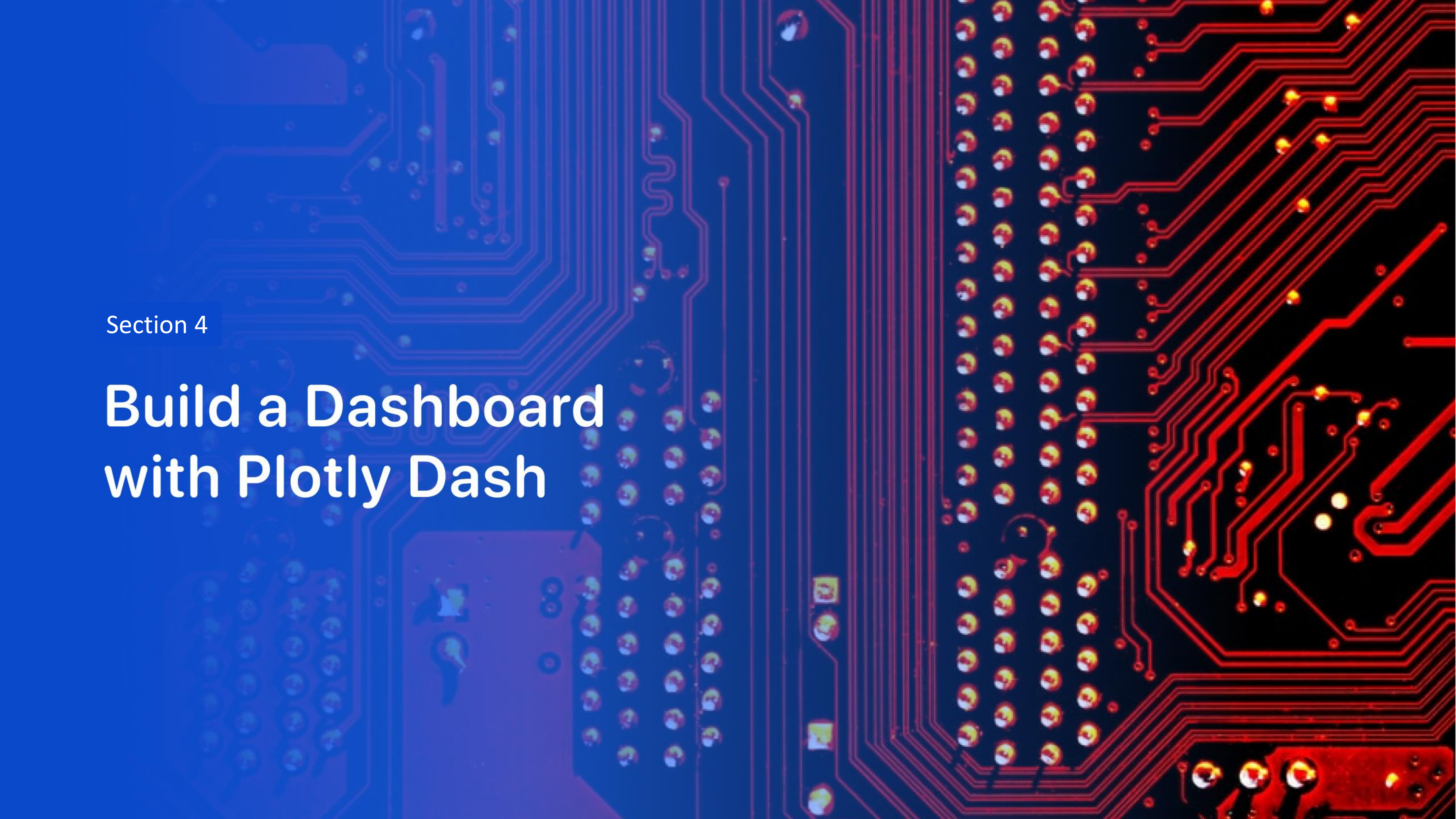


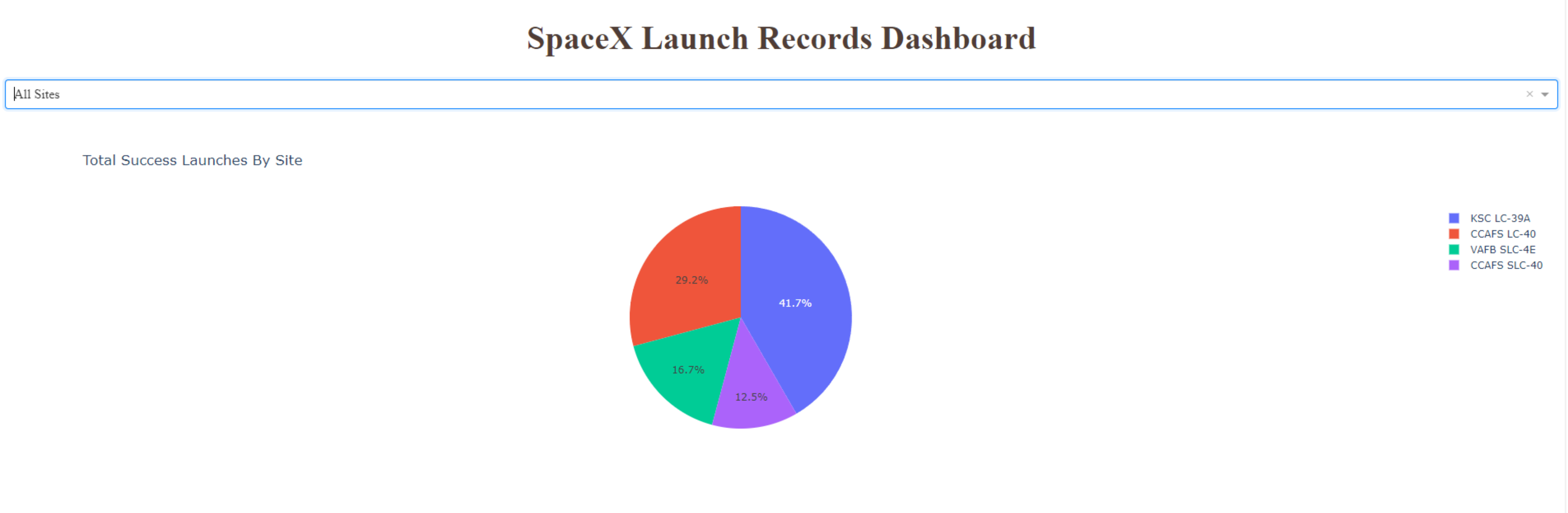
Figure 22: Folium map with Proximities



Section 4

Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



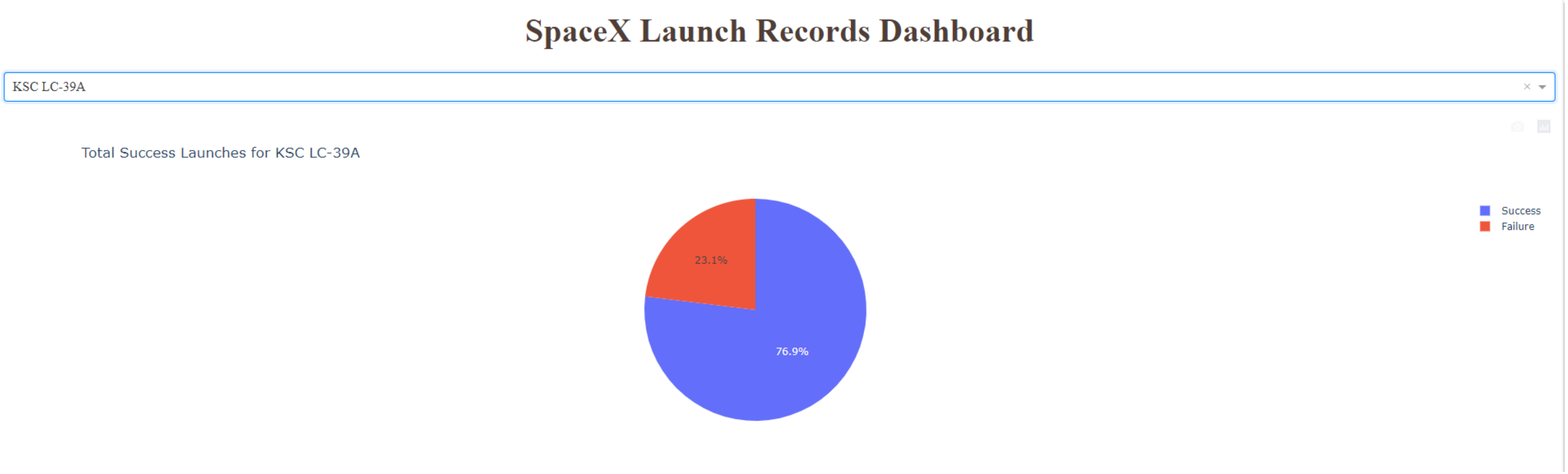
SpaceX Launch Records Dashboard

All Sites

Figure 23: All launch site success rate

- This is the distribution of successful landings across all launch sites.
- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same landing success rate.
- VAFB has the smallest successful landing rate.

Launch Site Highest Success Rate



SpaceX Launch Records Dashboard

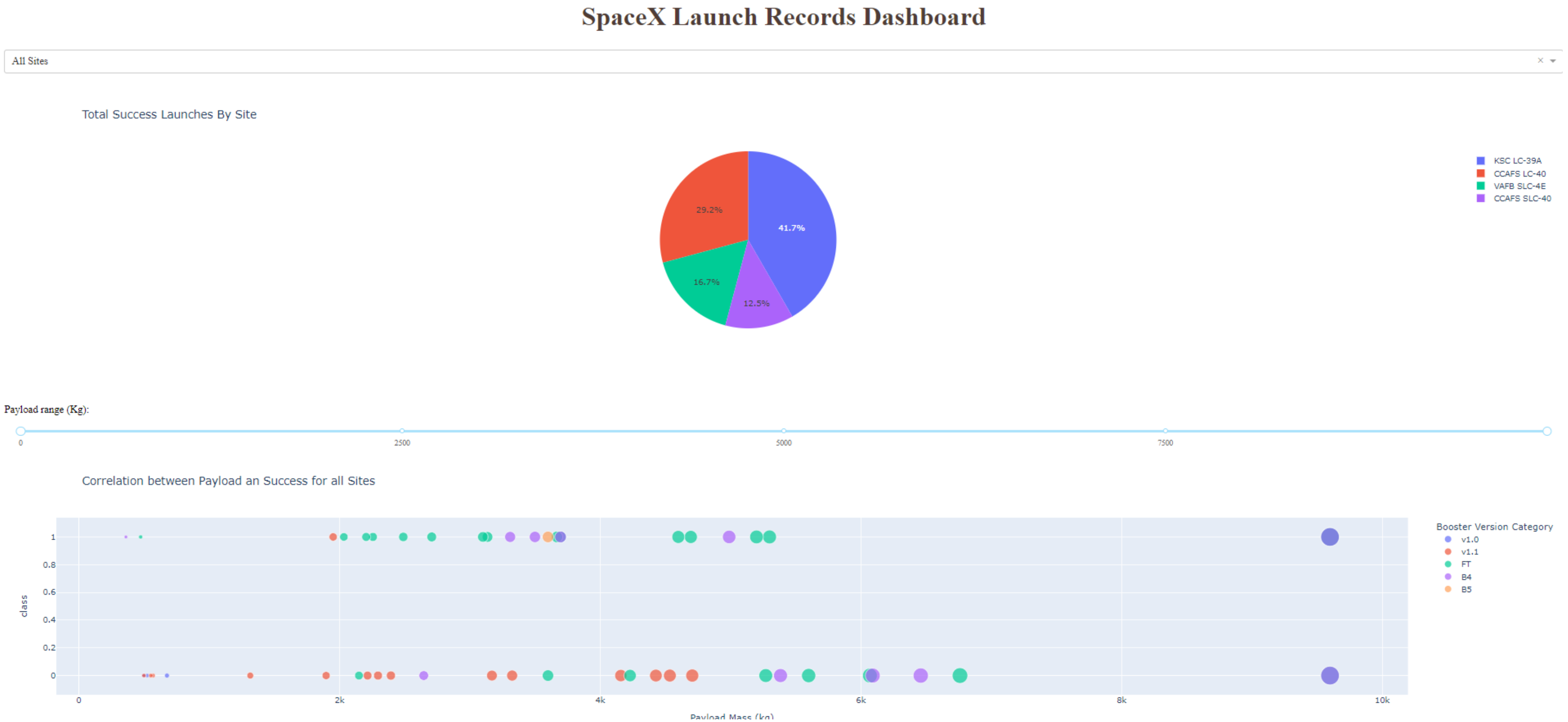
KSC LC-39A

Total Success Launches for KSC LC-39A

Figure 24: Launch Site has the highest success rate

- KSC LC-39A has the highest success rate with 10 successful landing and 3 failed landing

Payload Mass vs. Success vs. Booster Version Category



SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

Payload range (Kg):

Correlation between Payload an Success for all Sites

Figure 25: Scatter plot Payload Mass vs. Success Rate vs. Booster Version Category

- The Dashboard has a Payload slider with values set from 0 - 10,000 kg.
- The class represents 1 for a successful landing and 0 for a failed landing.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The model with the highest performs is “Decision Tree” with value of 87.5%.

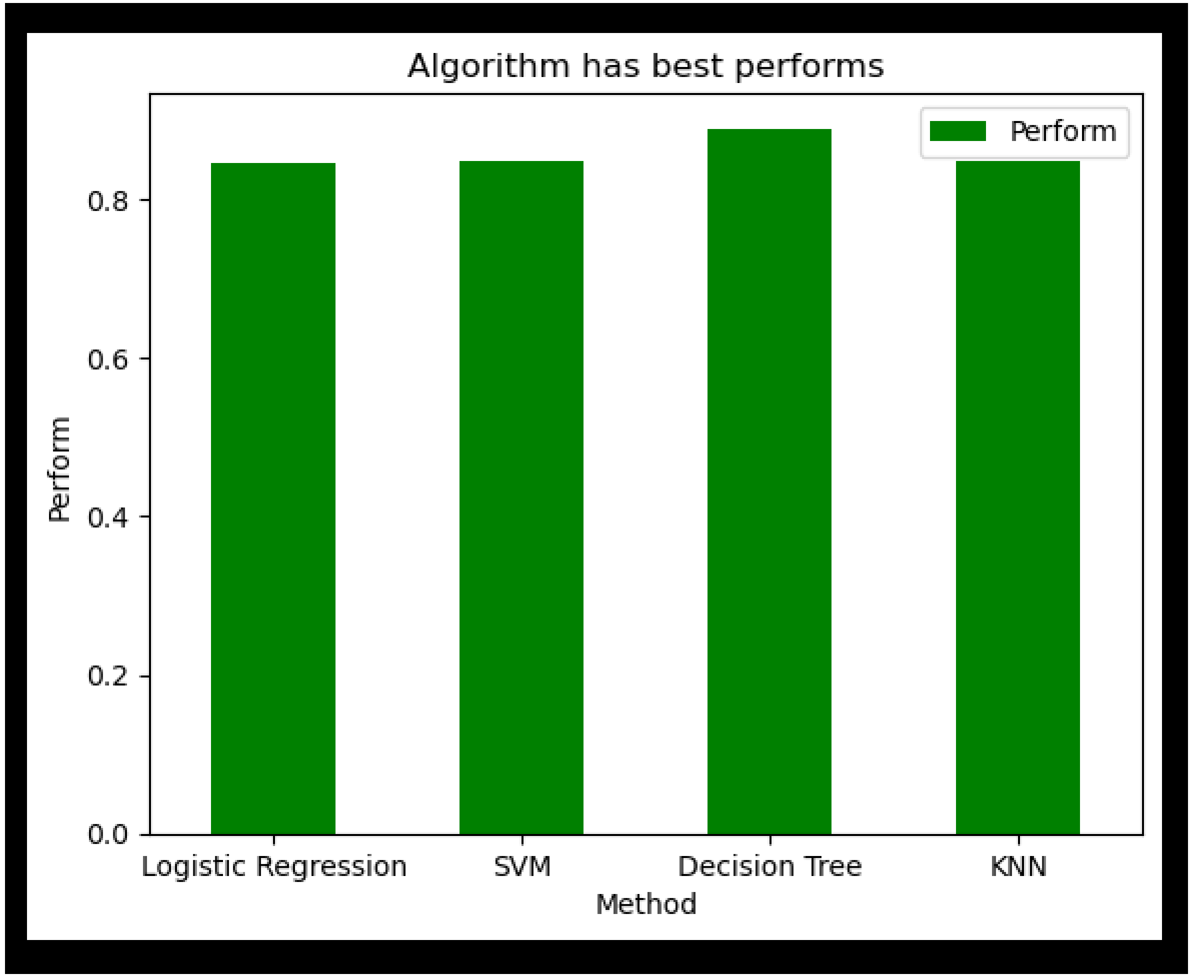
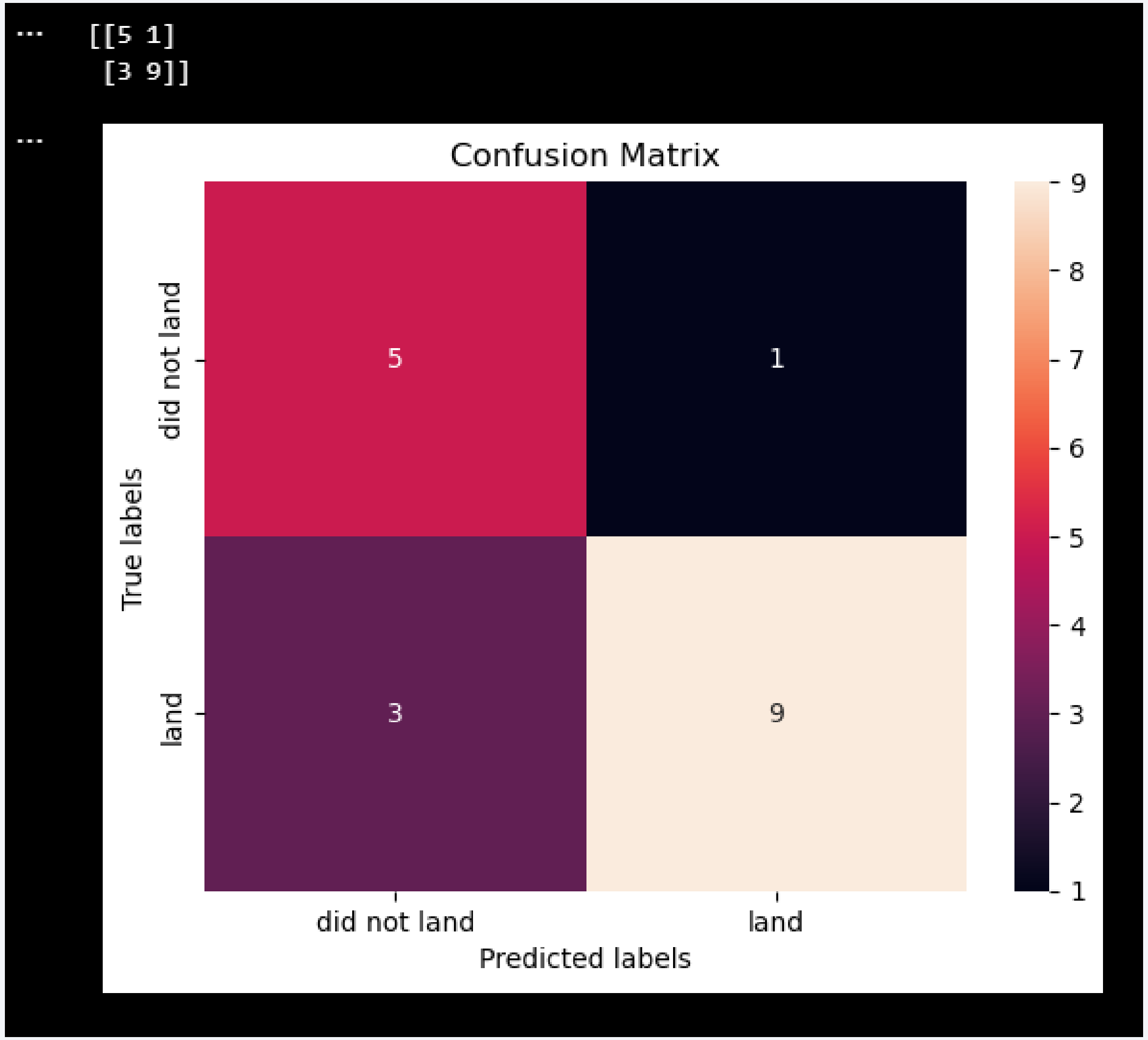


Figure 26: Bar Plot Model Accuracy

Confusion Matrix



```
...   [[5 1]
       [3 9]]

...
```

Figure 27: Confusion Matrix

- The values obtained are:
  - True Positive: 9
  - True Negative: 3
  - False Positive: 5
  - False Negative: 1

Conclusions

- 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.
- After finished the project we can conclude that:
  - The larger the flight amount at a launch site, the greater the success rate at a launch site.
  - Launch success rate started to increase in 2013 till 2020.
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
  - KSC LC-39A had the most successful launches of any sites.
  - The Decision tree classifier is the best machine learning algorithm for this task.

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!