IBM Developer
SKILLS NETWORK

Data Science and the Space Race: An Analytical Journey Beyond Earth

Kristi Zhupa
18.06.2023

Outline

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

Executive Summary

- **Summary of methodologies**
    - Data Collection through API
    - Data Collection with Web Scraping
    - Data Wrangling
    - Exploratory Data Analysis with SQL
    - Investigating Data through Data Visualization
    - Interactive Visual Analytics with Folium
    - Predictive Modeling with Machine Learning
- **Summary of all results**
    - Results from Exploratory Data Analysis
    - Visual Evidence of Interactive Analysis
    - Predictive Analytics

Introduction

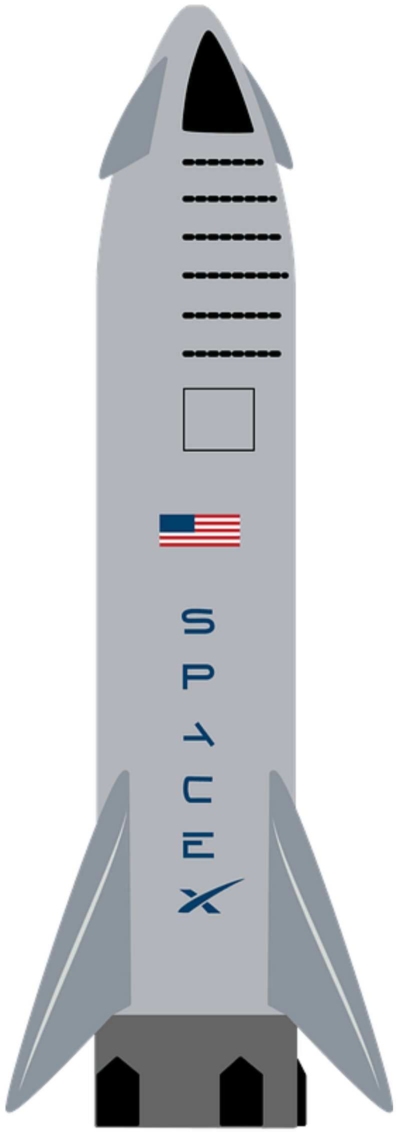
Background

SpaceX, a key player in the space sector, aims to make space travel more affordable for everyone. Their significant achievements include dispatching spacecraft to the International Space Station, setting up a satellite system to provide internet access, and undertaking manned space missions. The affordable price tag of their launches ($62 million each) is largely due to SpaceX's unique ability to reuse the first stage of its Falcon 9 rocket. In contrast, other providers that can't reuse this part have costs that can reach over $165 million per launch. If we can predict whether the first stage will land successfully, we can estimate the cost of each launch. To do this, we can use available public data and machine learning models to anticipate whether SpaceX, or a rival company, will be able to reuse the first stage of the rocket.

Topics for Exploration

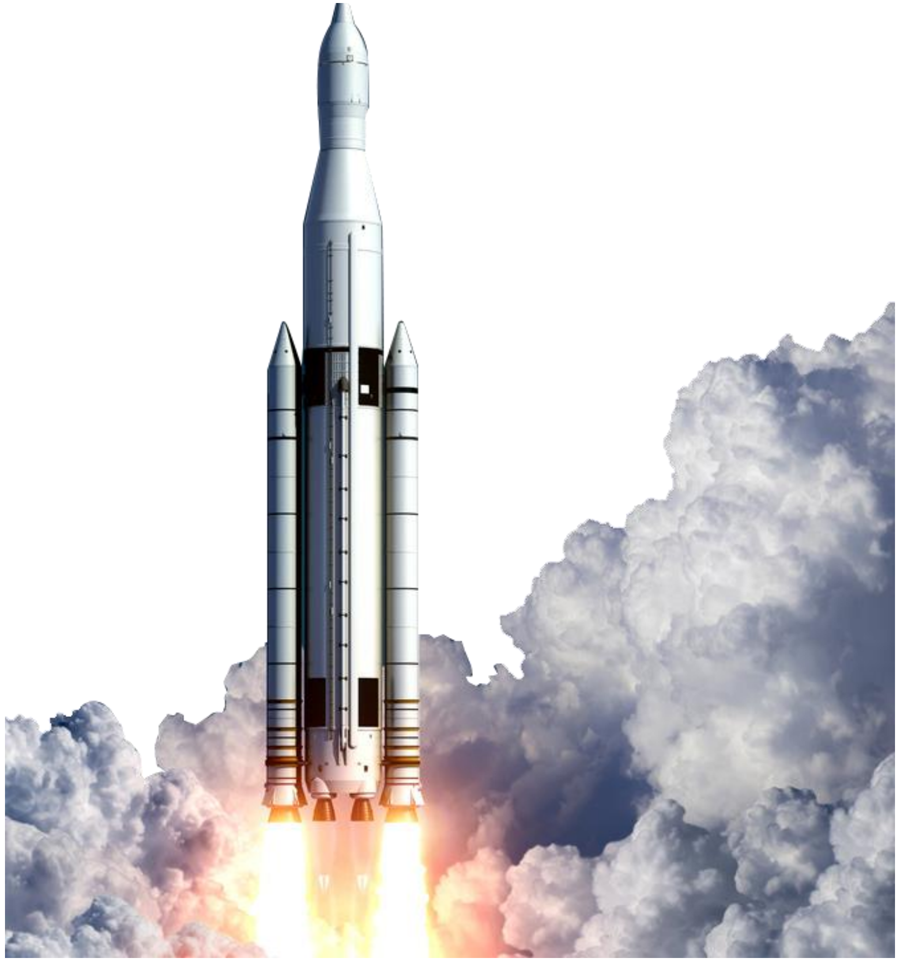
- How do factors such as payload mass, launch site, the number of flights, and orbits contribute to the success of the first stage landing?
- Is there an increase in the success rate of landings over the years?
- Which algorithm proves to be the most efficient for binary classification in this context?

Methodology



Methodology

- Data collection using both the SpaceX REST API and web scraping techniques, which includes gathering data from Wikipedia.
- Data wrangling tasks that include filtering out irrelevant data, handling missing values, and applying one hot encoding to prepare for binary classification.
- Conducting exploratory data analysis (EDA) utilizing SQL and various data visualization methods.
- Implementing interactive visual analytics through tools such as Folium and Plotly Dash to enhance understanding of the data.
- Executing predictive analysis by constructing, tuning, and evaluating various classification models for best results.



Data Collection

- The data collection involved using both API requests from SpaceX's REST API and web scraping from SpaceX's Wikipedia page.
- These two methods were crucial for getting a full picture of the launch information for in-depth analysis.
- Data columns from SpaceX REST API included: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Data columns from Wikipedia Web Scraping included: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

- The process initiates by soliciting rocket launch data from the SpaceX API.
- The received response is decoded into a JSON format and subsequently transformed into a pandas dataframe using .json() and .json_normalize() functions.
- Additional information regarding the launches is gathered from the SpaceX API through the application of custom functions.
- The collected information is then organized into a dictionary, which is then transformed into a dataframe.
- The dataframe is filtered to isolate Falcon 9 launches, focusing the analysis on this specific launch vehicle.
- Any missing values in the payload mass are addressed by replacing them with the mean value of the existing data.
- The completed, cleaned dataframe is then exported to a CSV file for potential further utilization and analysis.

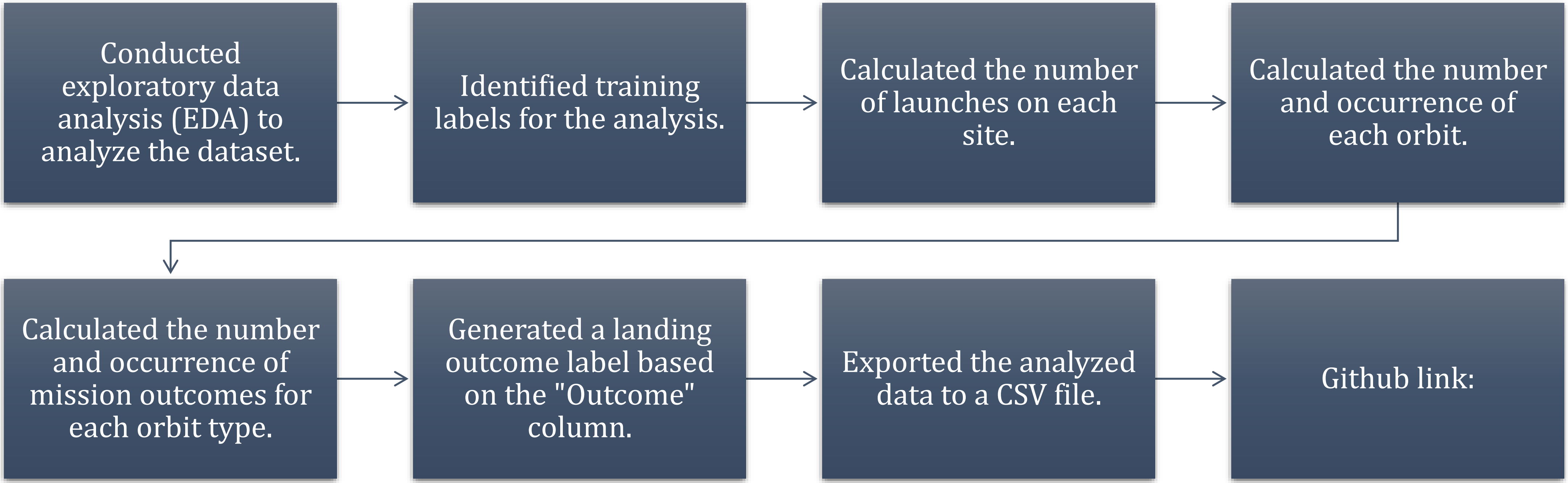
Github Url: Spacex-data-collection-api

Data Collection - Scraping

- Data concerning Falcon 9 launches is retrieved from Wikipedia.
- The HTML response is used to create a BeautifulSoup object.
- Column names are extracted from the HTML table header.
- Data is collected through the parsing of HTML tables.
- This collected data is organized into a dictionary.
- A dataframe is subsequently generated from this dictionary.
- Finally, the prepared dataframe is exported to a CSV file for future use and analysis.

Github Url: Spacex-data-collection-api

Data Wrangling



Conducted exploratory data analysis (EDA) to analyze the dataset.

Identified training labels for the analysis.

Calculated the number of launches on each site.

Calculated the number and occurrence of each orbit.

Calculated the number and occurrence of mission outcomes for each orbit type.

Generated a landing outcome label based on the "Outcome" column.

Exported the analyzed data to a CSV file.

Github link:

Github Url: Spacex-data_wrangling

EDA with Data Visualization

- Throughout the data exploration process, a range of visualizations were employed to investigate relationships between variables. The analysis involved examining the correlation between flight number and launch site, as well as payload and launch site. Additionally, an exploration of the success rate for each orbit type, the connection between flight number and orbit type, and the yearly trend in launch success were conducted. To effectively present these findings, the following charts were generated:

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit Type
- Orbit Type vs. Success Rate
- Success Rate Yearly Trend

Github Url: Exploring-preparing-data

EDA with SQL

- The SpaceX dataset was loaded into a PostgreSQL database within the Jupyter notebook environment.
- EDA was conducted using SQL to extract valuable insights from the data. Through the formulation of queries, various aspects were explored, including:
- Determining the names of unique launch sites involved in space missions.
- Calculating the total payload mass carried by boosters launched by NASA (CRS).
- Calculating the average payload mass carried by booster version F9 v1.1.
- Obtaining the total count of successful and failed mission outcomes.
- Identifying failed landing outcomes on drone ships, along with their corresponding booster versions and launch site names.

Github Url: Eda-sql

Map with Folium

Launch Site Markers:

- Placed a circle at the NASA Johnson Space Center's coordinates, along with a popup label displaying its name using the latitude and longitude coordinates.

- Added circles at the coordinates of all other launch sites, accompanied by popup labels indicating their names using latitude and longitude coordinates.

Map Visualization using Folium:

- Created a map using Folium to visualize the launch sites and their associated markers.

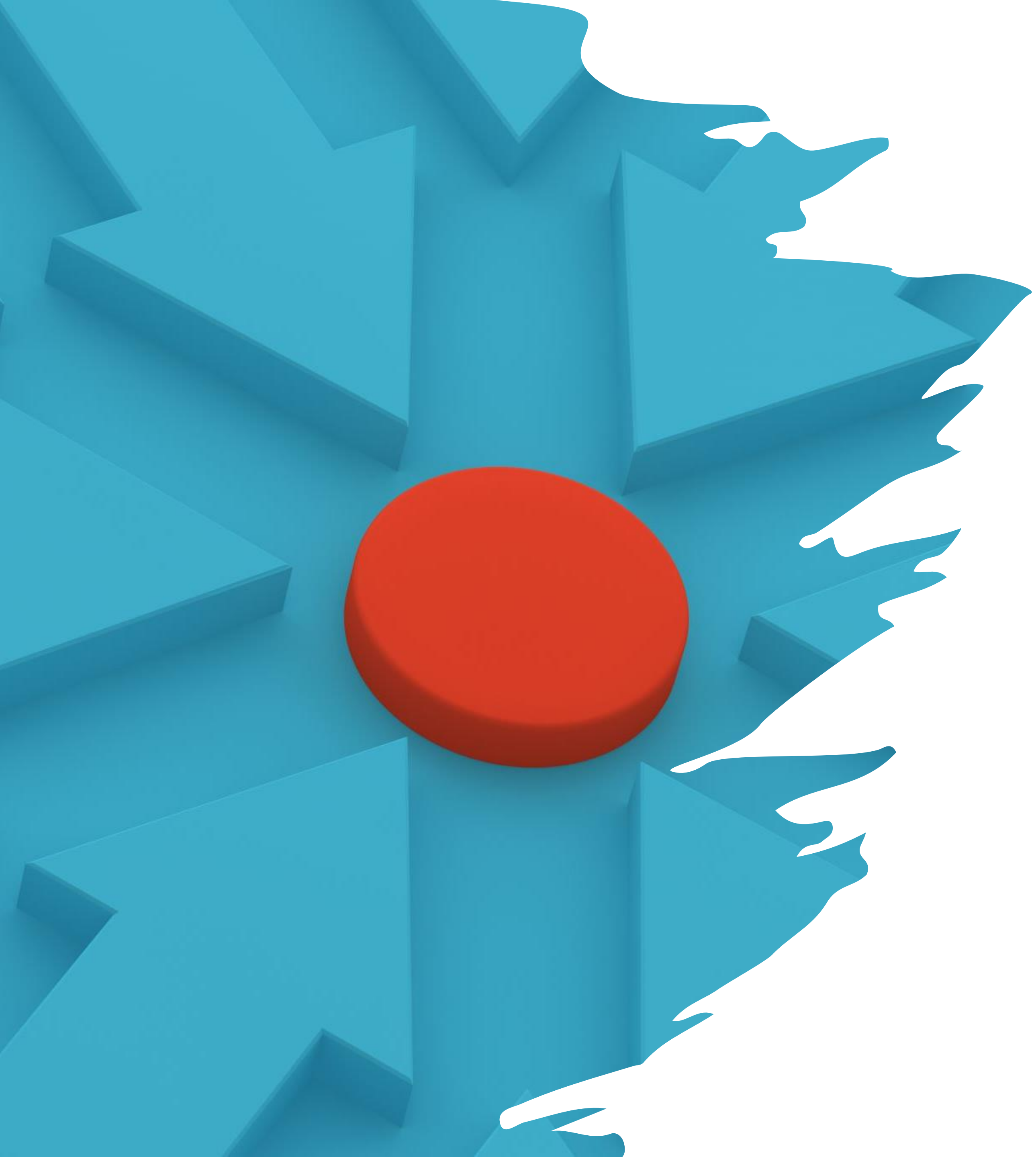
Launch Outcome Markers:

- Included colored markers on the map to represent successful launches (**green**) and unsuccessful launches (**red**) at each launch site, providing a visual representation of the success rates at different launch sites.

Distances and Proximity:

- Corporated colored lines on the map to depict the distances between launch site CCAFS SLC-40 and its nearest coastline, railway, highway, and city, highlighting the proximity of the launch site to these features.

Github URL: Analysis-folium



Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Implemented a dropdown list that allows users to select a Launch Site.

Pie Chart showing Successful Launches:

- Introduced a pie chart to display the total count of successful launches for all sites. If a specific Launch Site is chosen, it shows the distribution of successful versus failed launches for that site.

Slider for Payload Mass Range:

- Incorporated a slider that enables users to select a desired range of Payload Mass.

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

- Added a scatter chart to visualize the relationship between Payload Mass and Success Rate for different Booster Versions.

Github URL: Spacex_dash

Predictive Analysis (Classification)

In the predictive analysis process, the following steps were performed:

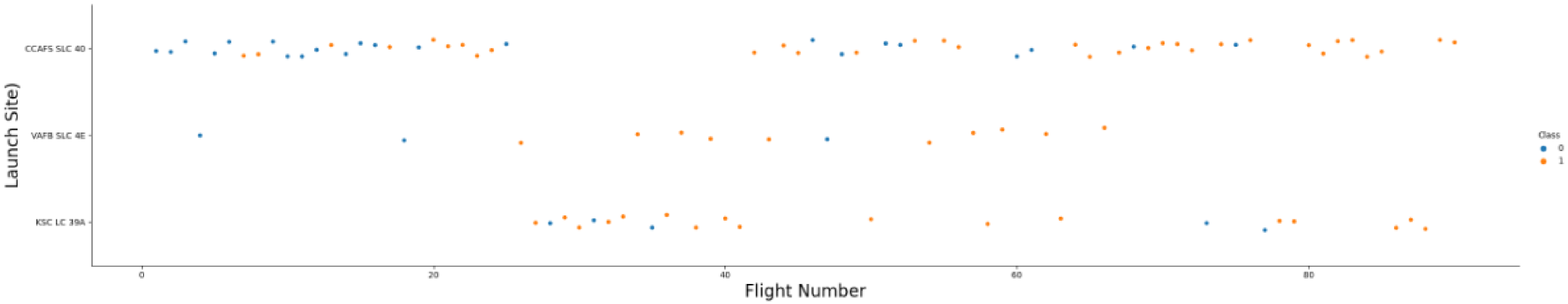
- The "Class" column was converted into a NumPy array.
- Data standardization was carried out using the StandardScaler, where the data was fitted and transformed.
- The dataset was split into training and testing sets using the train_test_split function.
- To optimize the model's parameters, a GridSearchCV object was created with a cross-validation value of 10.
- Different algorithms including logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and K-Nearest Neighbor (KNeighborsClassifier()) were applied using the GridSearchCV object.
- The accuracy of each model was calculated on the test data using the .score() method.
- The confusion matrix was examined for each model to evaluate the classification performance.
- The best model was determined based on metrics such as Jaccard Score, F1 Score, and Accuracy.

Github Url: [Machine_Learning_Prediction]

Results

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

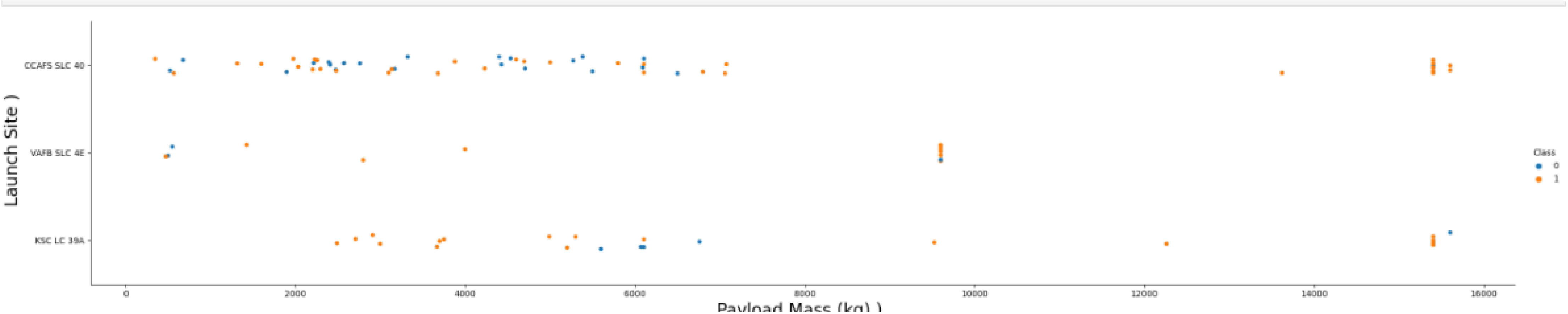
EDA with Visualization



Flight Number vs. Launch Site

Based on the analysis conducted, the following conclusions can be drawn:

- The analysis reveals that earlier flights had a lower success rate, indicated by the color **blue** representing failures.
- In contrast, later flights exhibited a higher success rate, as denoted by the color orange symbolizing successful launches.
- Therefore, based on these observations and analysis, it can be inferred that **newer launches tend to have a higher success rate.**



Payload vs. Launch Site

- There is a **positive** correlation between payload mass (kg) and the success rate. Typically, as the payload mass increases, the success rate also tends to increase.

- Most launches with a payload greater than 7,000 kg were successful, indicating a higher success rate for heavier payloads.

- KSC LC 39A has achieved a 100% success rate for launches with a payload less than 5,500 kg, suggesting a high level of reliability for smaller payloads at this launch site.

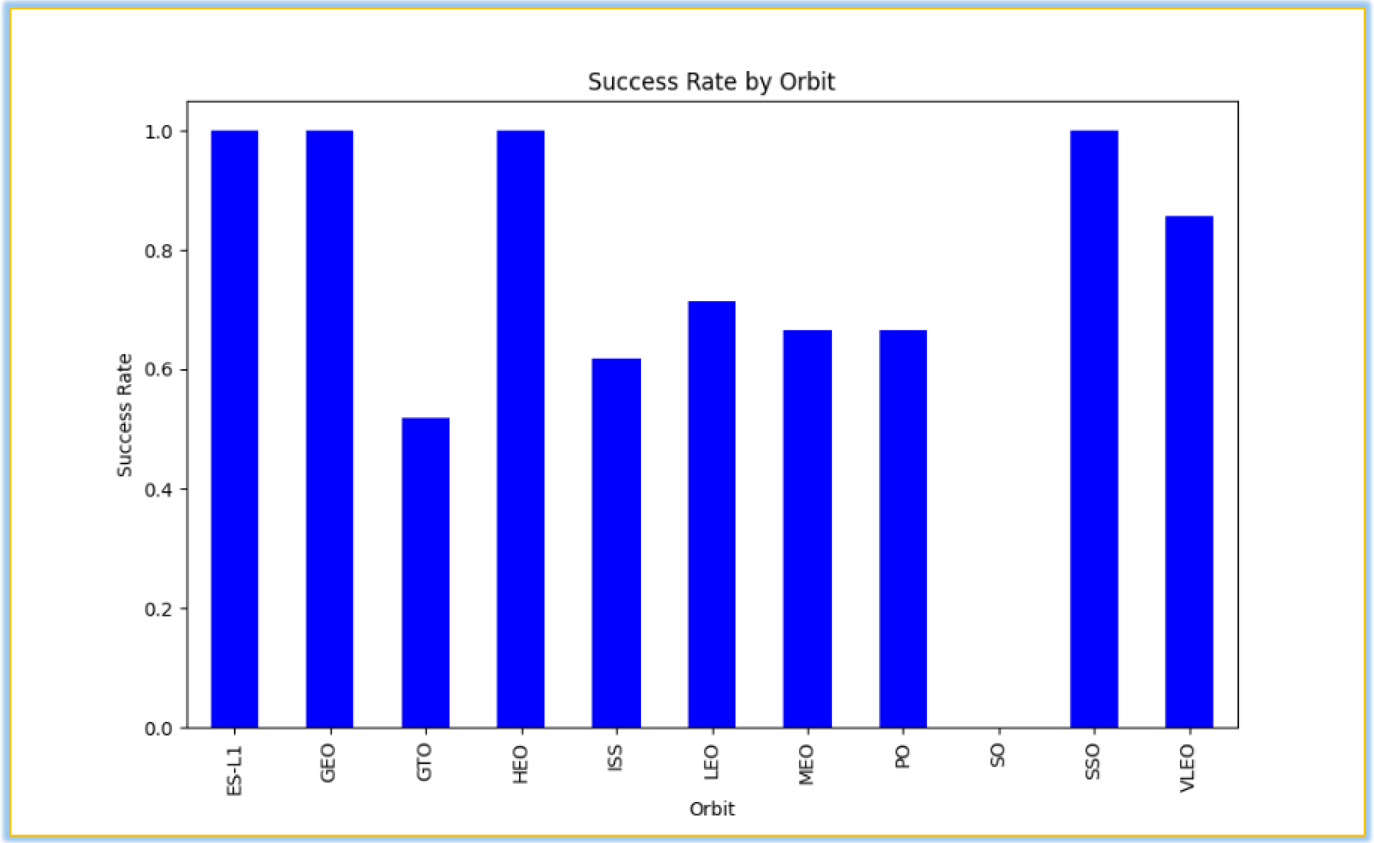
- It was observed that VAFB SLC 4E has not conducted any launches with a payload greater than approximately 10,000 kg, indicating a limitation in handling heavier payloads at this specific launch site.

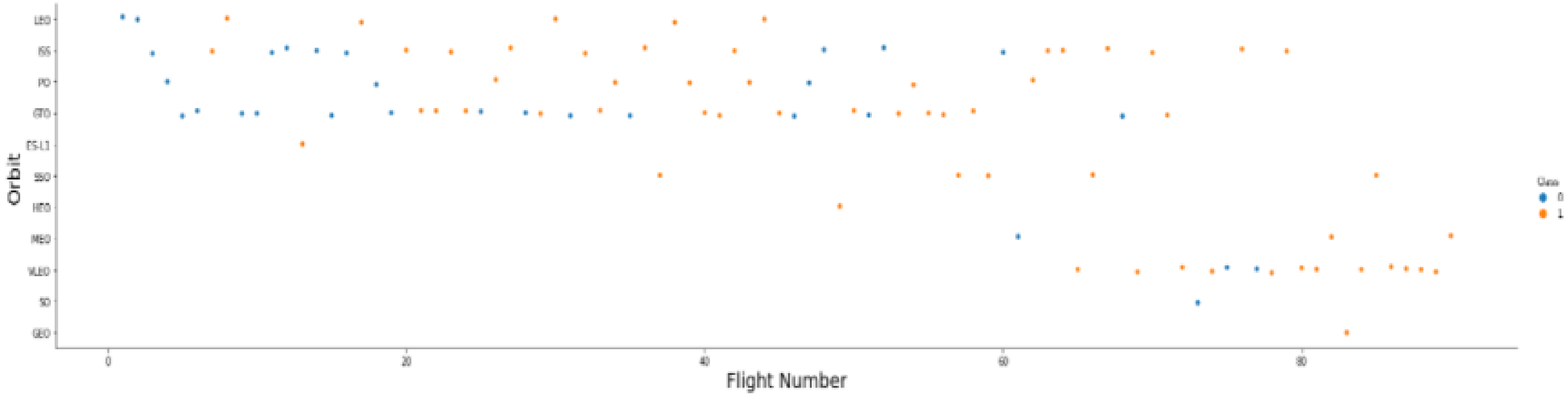
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits achieved a 100% success rate. This indicates that all launches targeting these orbit types were successful.

- GTO, ISS, LEO, MEO, and PO orbits exhibited success rates ranging between 50% and 80%. This suggests that launches targeting these orbit types had moderate success rates, with some variations in the outcome.

- The SO orbit type recorded a 0% success rate. This indicates that no successful launches were achieved for this specific orbit type.



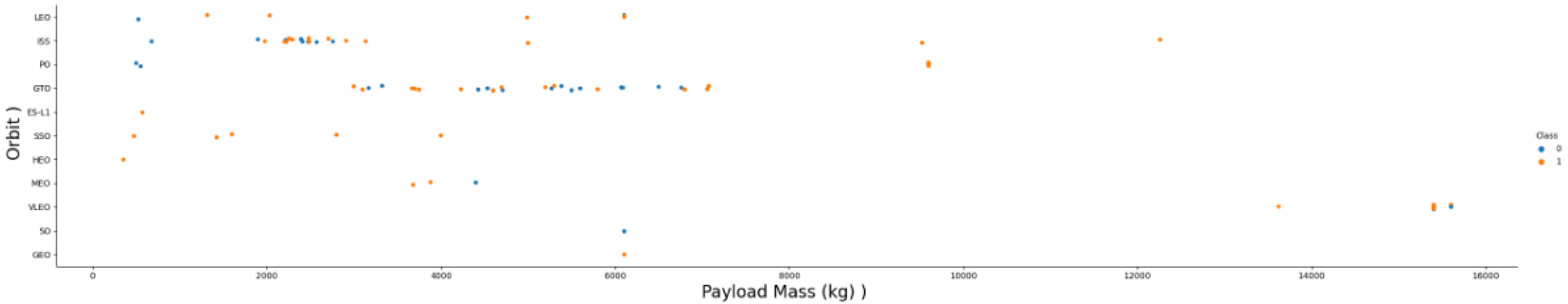


Flight Number vs. Orbit Type

- In general, the success rate tends to increase as the number of flights for each orbit increases. This indicates that with more flights, there is a higher likelihood of achieving successful launches for a particular orbit.

- This relationship is particularly evident for the Low Earth Orbit (LEO), where the success rate shows a consistent upward trend as the Flight Number increases. This suggests a positive correlation between the number of flights and the success rate for the LEO orbit.

- However, it is noteworthy that the Geostationary Transfer Orbit (GTO) does not follow this trend. The success rate for the GTO orbit does not exhibit a consistent increase with the Flight Number.



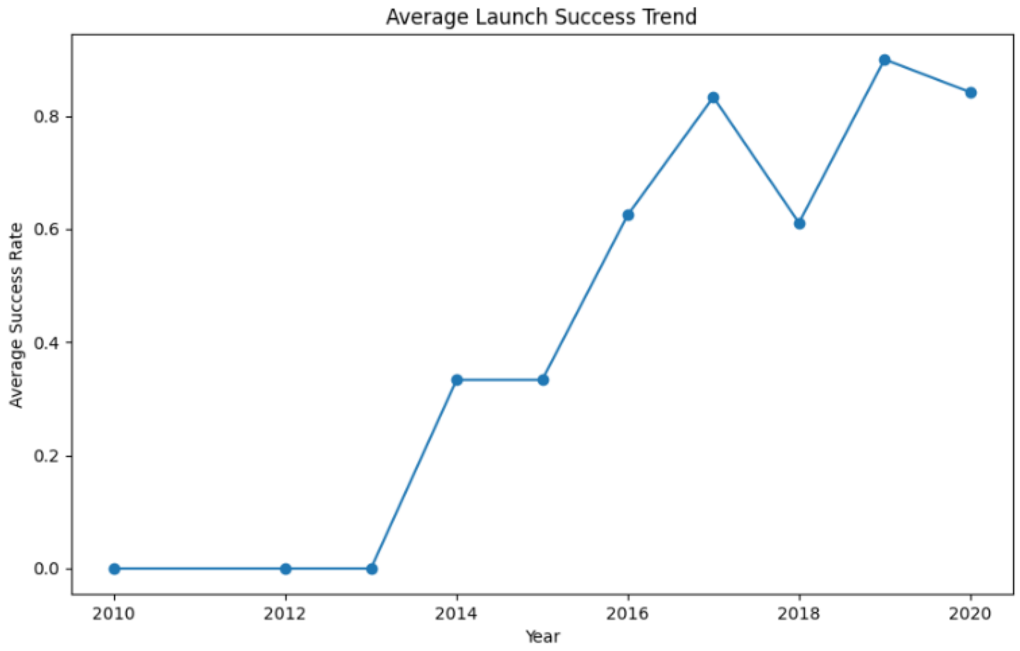
Payload vs. Orbit Type

- Heavy payloads tend to perform better in Low Earth Orbit (LEO), International Space Station (ISS), and Polar Orbit (PO). This suggests that these orbit types are well-suited for handling and accommodating heavier payloads, resulting in a higher success rate for launches with such payloads.

- However, when it comes to the Geostationary Transfer Orbit (GTO), the success rate with heavier payloads is more mixed. This indicates that the GTO orbit may present certain challenges or factors that affect the success of launches with heavier payloads, resulting in a less consistent success rate for this orbit type.

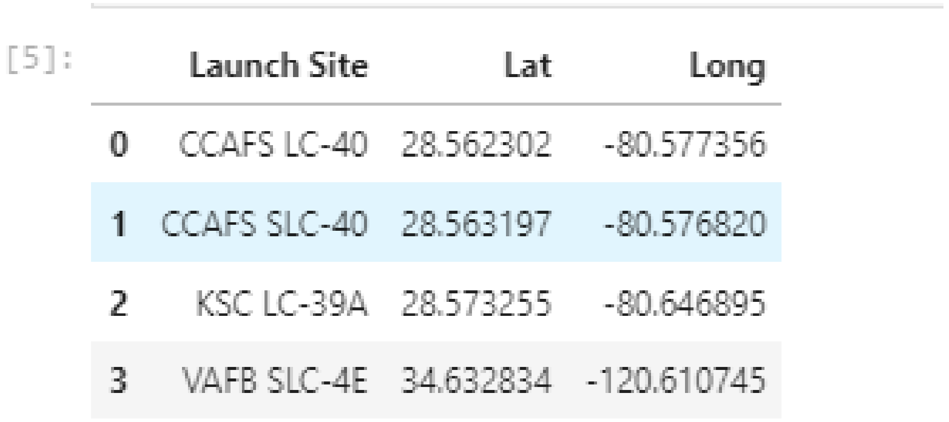
Launch Success Yearly Trend

- The success rate displayed **positive progress** during the periods of 2013-2017 and 2018-2019, indicating an **upward trend** in successful launches.
- However, there was a **decrease** in the success rate between 2017-2018 and from 2019-2020, suggesting a **temporary setback** in launch success.
- Despite these fluctuations, the **overall trend** reveals **significant improvement** in the success rate since 2013, reflecting **notable advancements** in mission outcomes over the years.



All Launch Site Names

- The DISTINCT keyword was employed to display only the unique launch sites from the SpaceX data.

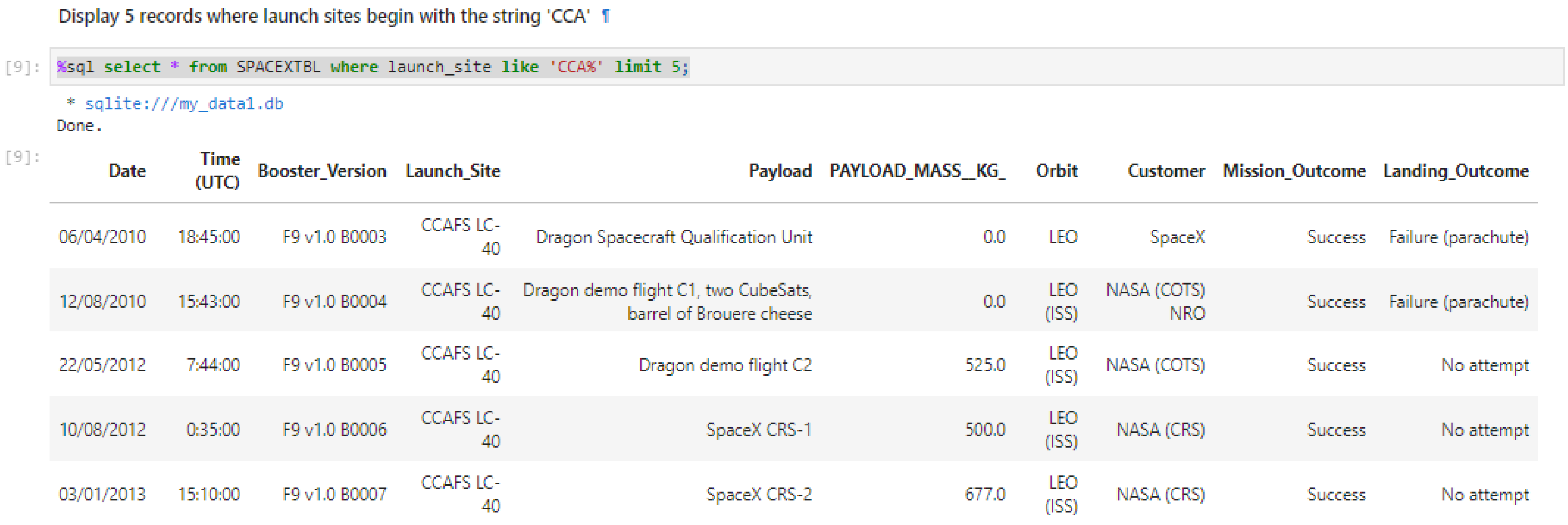


[5]:

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`



Display 5 records where launch sites begin with the string 'CCA' ¶

```
[9]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

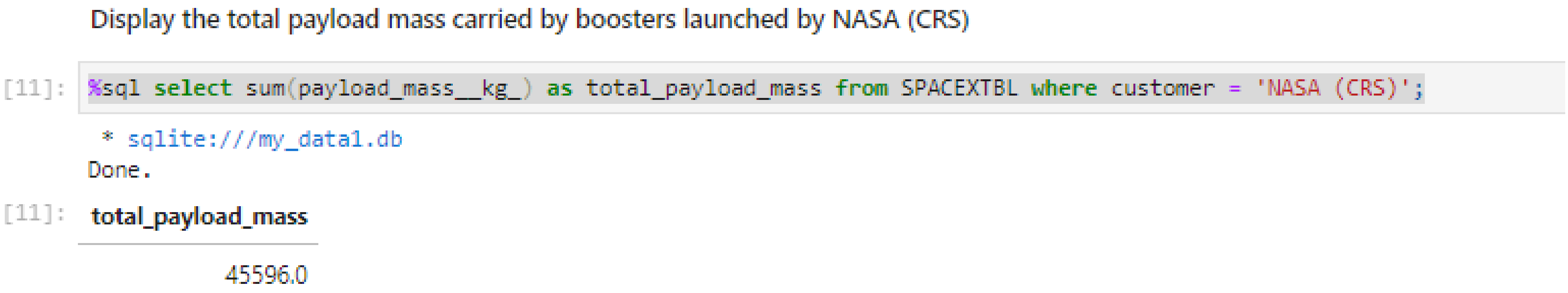
```
 * sqlite:///my_data1.db
Done.
```

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by NASA boosters was calculated as 45596.

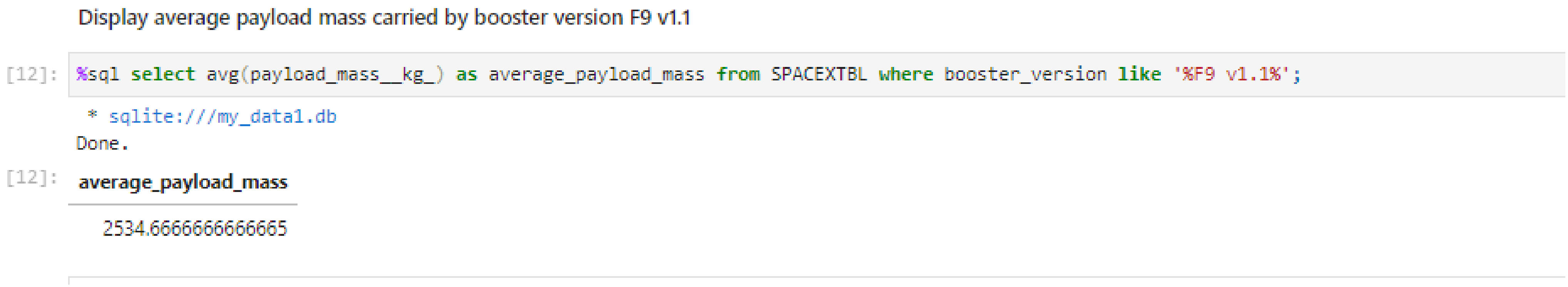


Display the total payload mass carried by boosters launched by NASA (CRS)

```
[11]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.

[11]: total_payload_mass
                 45596.0
```

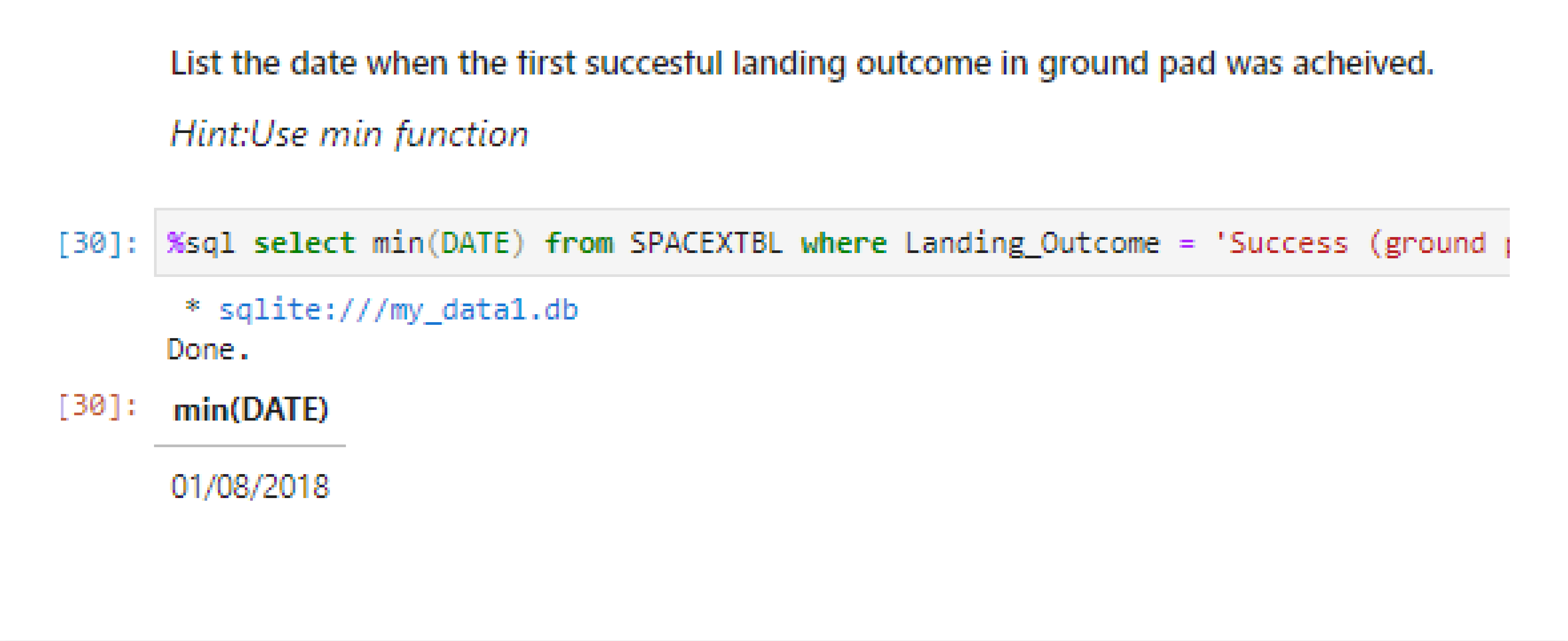
Average Payload Mass by F9 v1.1

The average payload mass carried by the booster version F9 v1.1 was calculated to be 2534,6.



Display average payload mass carried by booster version F9 v1.1

```
[12]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';

 * sqlite:///my_data1.db
Done.

[12]: average_payload_mass
        2534.6666666666665
```



List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[30]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground p
```

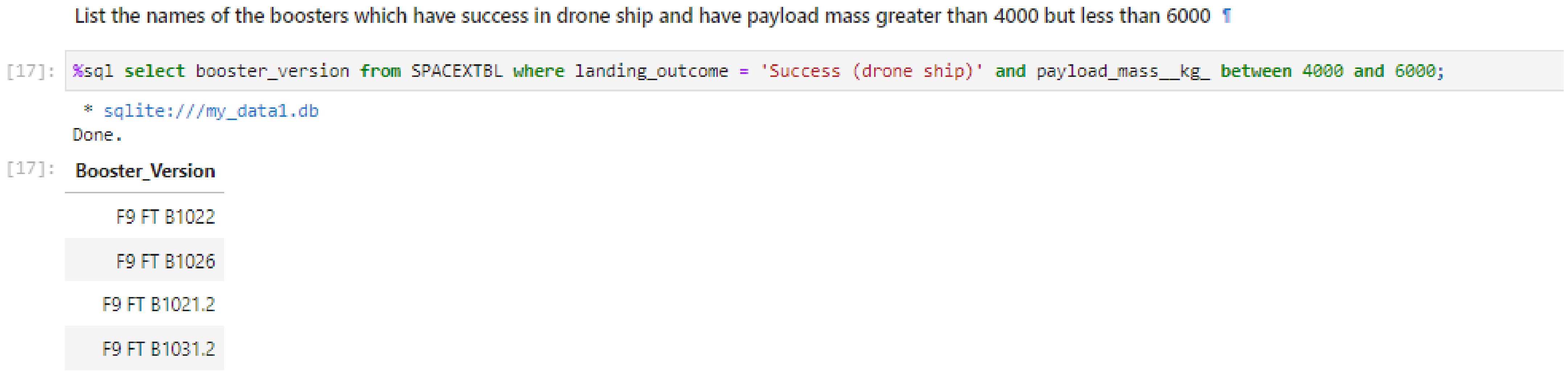
```
 * sqlite:///my_data1.db
Done.
```

[30]:

min(DATE)
01/08/2018

First Successful Ground Landing Date

- The observation revealed that the first successful landing outcome on a ground pad occurred on 01/08/2018.



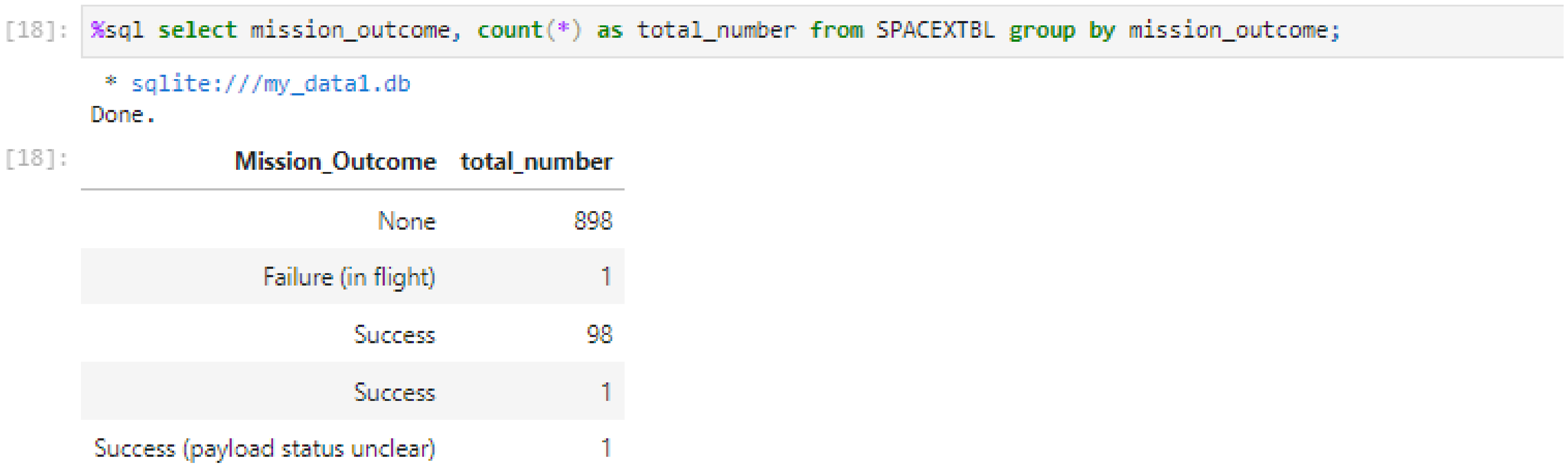
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 ¶

```
[17]: %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
[17]: Booster_Version
        F9 FT B1022
        F9 FT B1026
      F9 FT B1021.2
      F9 FT B1031.2
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- By utilizing the **WHERE** clause, we filtered for boosters that successfully landed on a drone ship. Additionally, we applied the **AND** condition to identify cases where the landing was successful and the payload mass fell within the range of greater than 4,000 but less than 6,000.



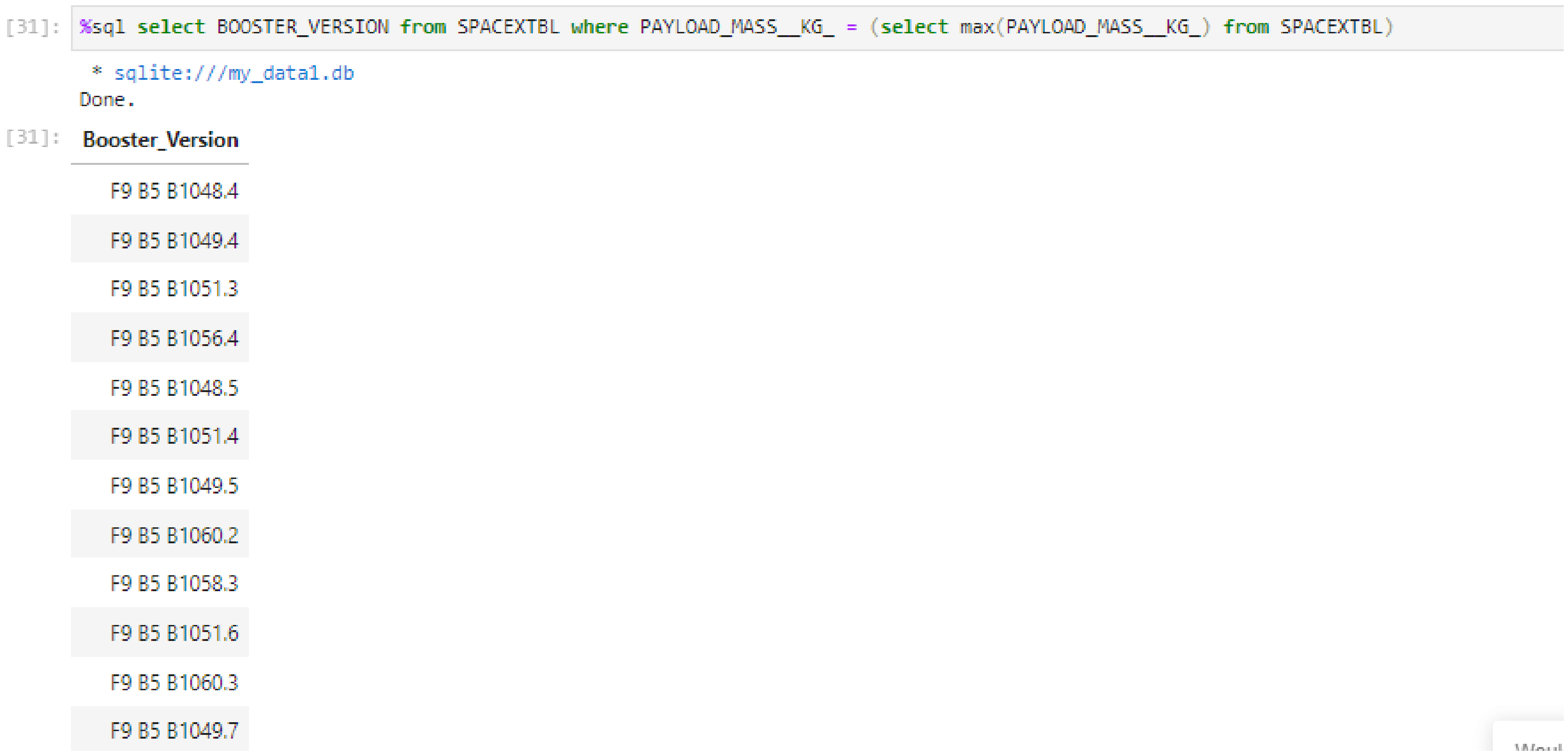
```
[18]: %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;

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

[18]:

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Total Number of Successful and Failure Mission Outcomes

- Out of the observed cases, there was **a single failure** during flight. On the other hand, there were **98 successful** flights. Additionally, there was **one** instance where the success of the flight was noted, but the status of the payload remained unclear.



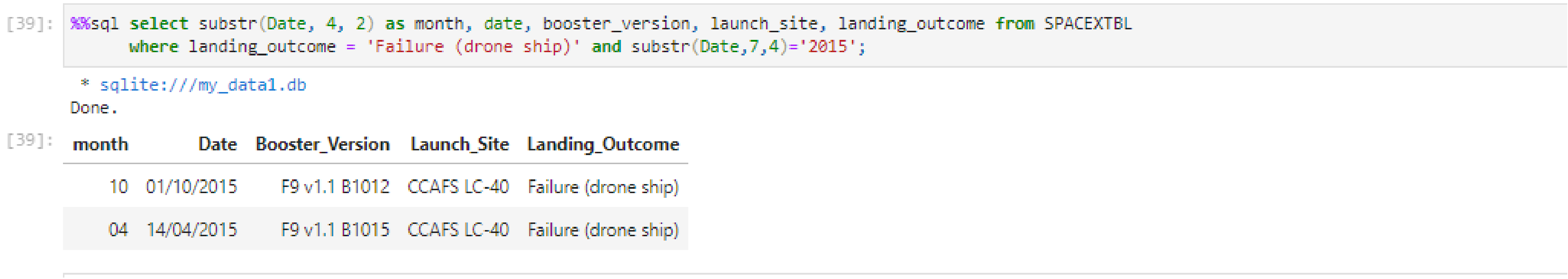
```
[31]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

[31]:

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

Boosters Carried Maximum Payload

- The booster that carried the maximum payload was determined using a subquery in the **WHERE** clause and the **MAX()** function.



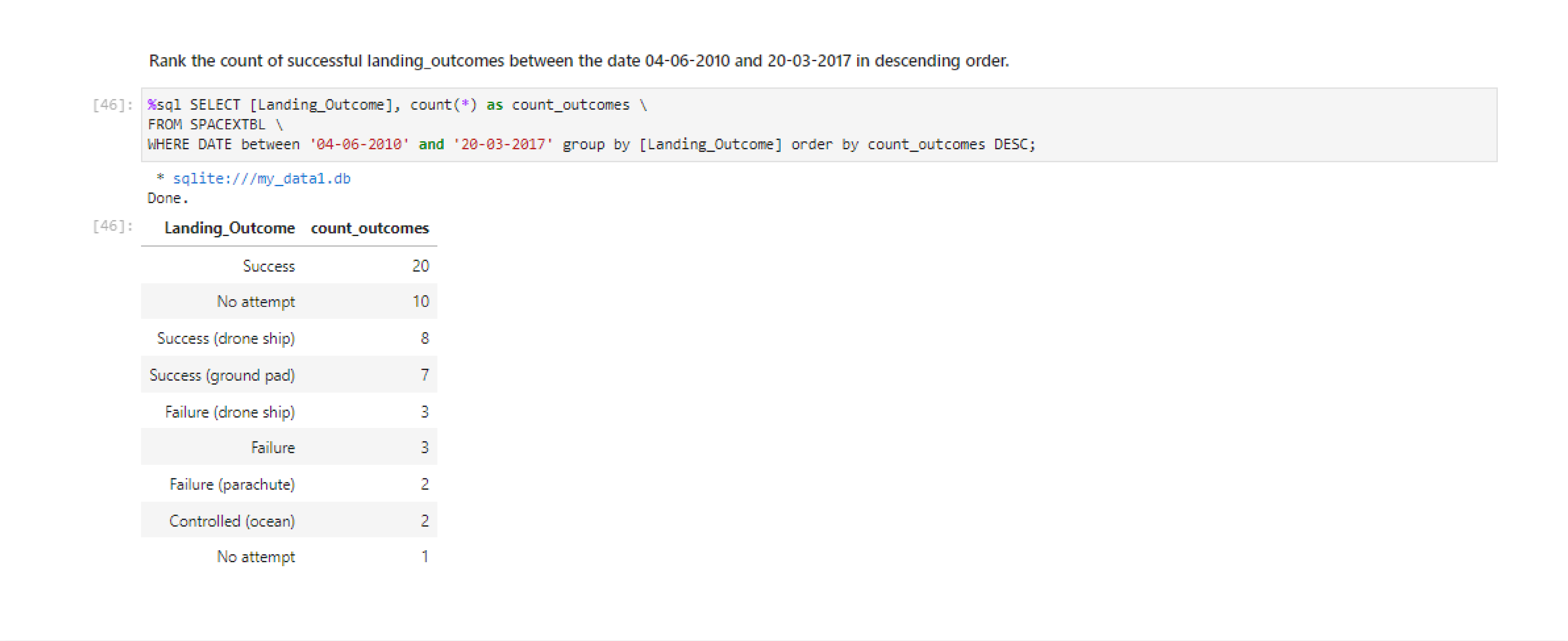
```
[39]: %%sql select substr(Date, 4, 2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTBL
          where landing_outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015';

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

[39]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Failed Landings on Drone Ship

- The data is presented, including the month, date, booster version, launch site, and landing outcome.



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[46]: %sql SELECT [Landing_Outcome], count(*) as count_outcomes \
      FROM SPACEXTBL \
      WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

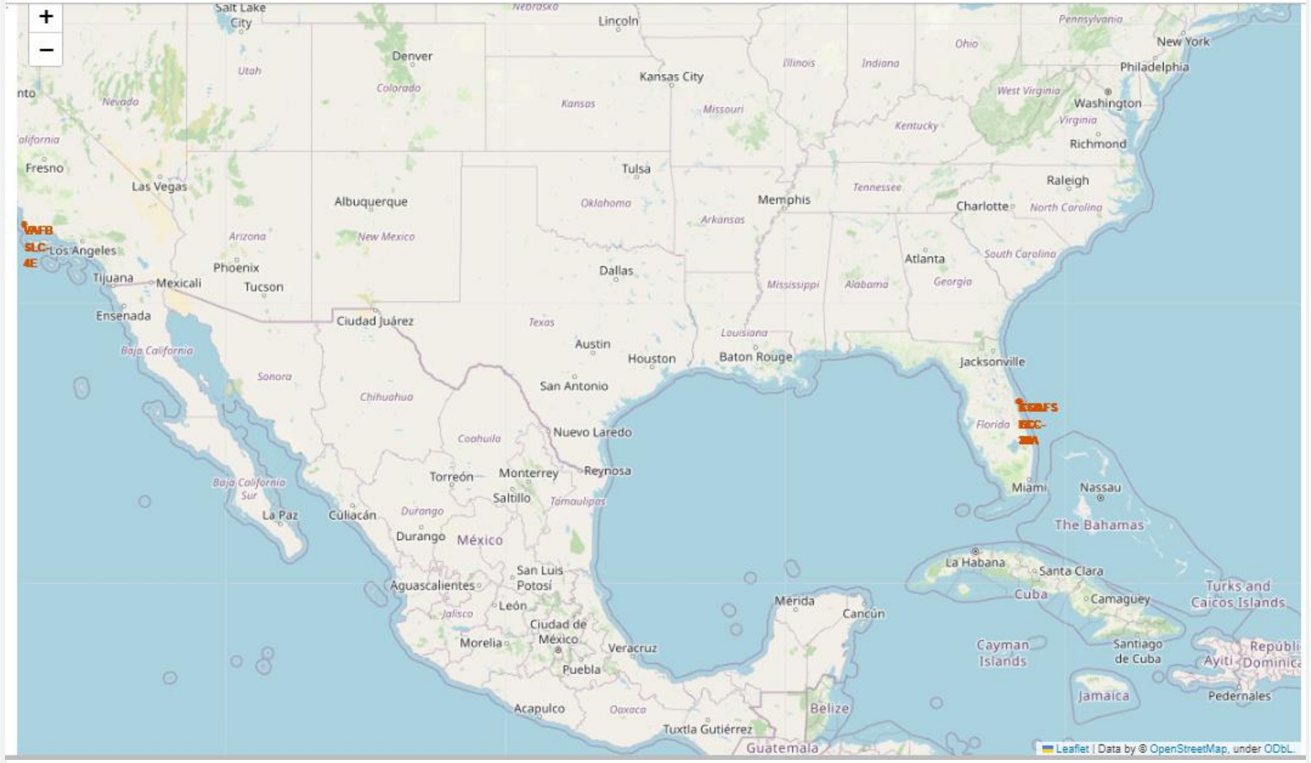
[46]:

Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

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

- The count of landing outcomes between June 4, 2010, and March 20, 2017, is provided in descending order.

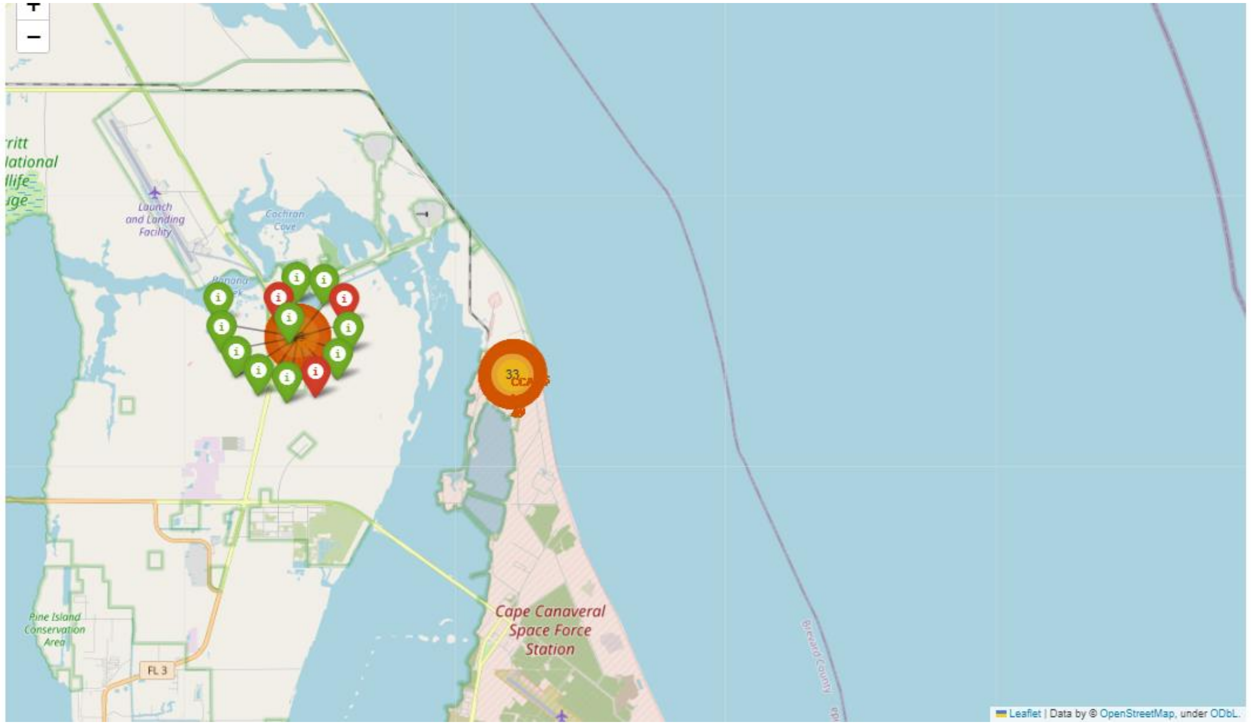
Launch Site Analysis



Launch Sites

- Most launch sites are located near the Equator due to the higher speed of the Earth's rotation at the Equator. Objects at the Equator are already moving at a significant speed, and when a spacecraft is launched from the Equator, it retains this speed due to inertia. This high speed helps the spacecraft maintain sufficient velocity to stay in orbit.

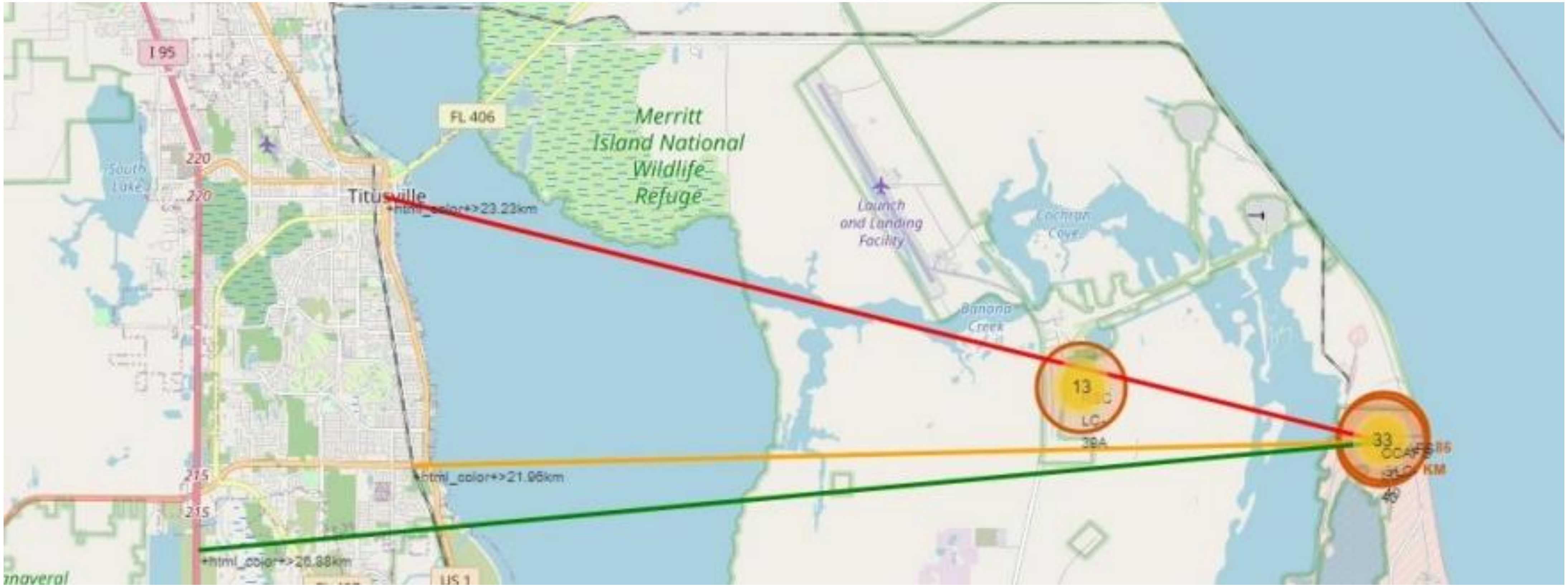
- Additionally, all launch sites are situated close to the coast. Launching rockets towards the ocean reduces the risk of debris falling or exploding near populated areas, ensuring safety during launches.


+
−
ritt
lational
llife
uge
Launch and Landing Facility
Cochran Cove
Banana
33
CCAFS
Cape Canaveral Space Force Station
Pine Island Conservation Area
FL 3
Brevard County
Leaflet | Data by © OpenStreetMap, under ODbL.

Color-labeled launch outcomes on the map

- Launch records are represented on the map with color-labeled markers, enabling easy identification of launch sites with comparatively high success rates.

- **Green** markers indicate successful launches.
- **Red** markers indicate failed launches.

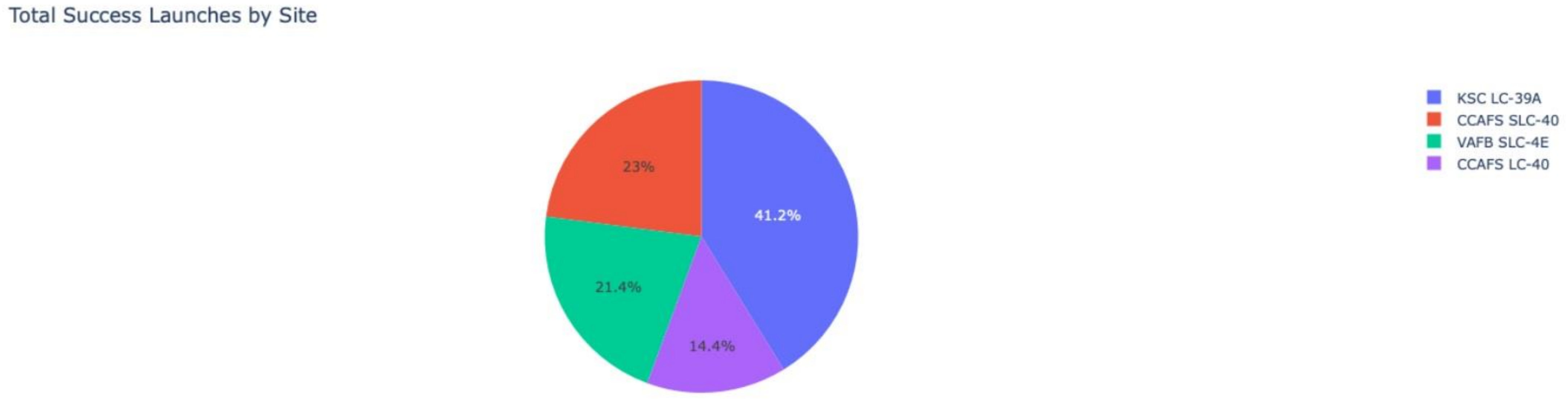


Distance to Landmarks

- The CCAFS SLC-40 launch site is located at the following distances from nearby landmarks:

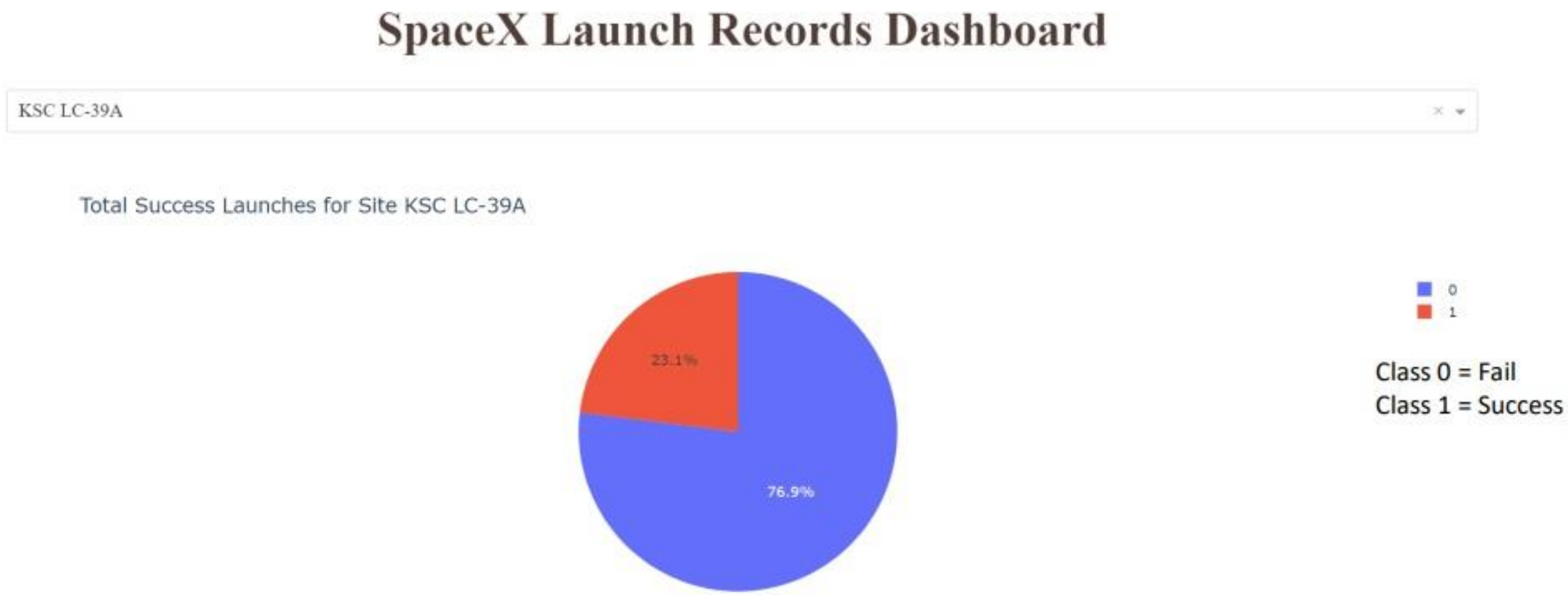
- Nearest coastline: 0.86 km
- Nearest railway: 21.96 km
- Nearest city: 23.23 km
- Nearest highway: 26.88 km

Dashboard with Plotly



Launch success count for all sites

KSC LC-39A has the highest percentage of successful launches among all the launch sites, accounting for 41.2% of the total successful launches.



SpaceX Launch Records Dashboard

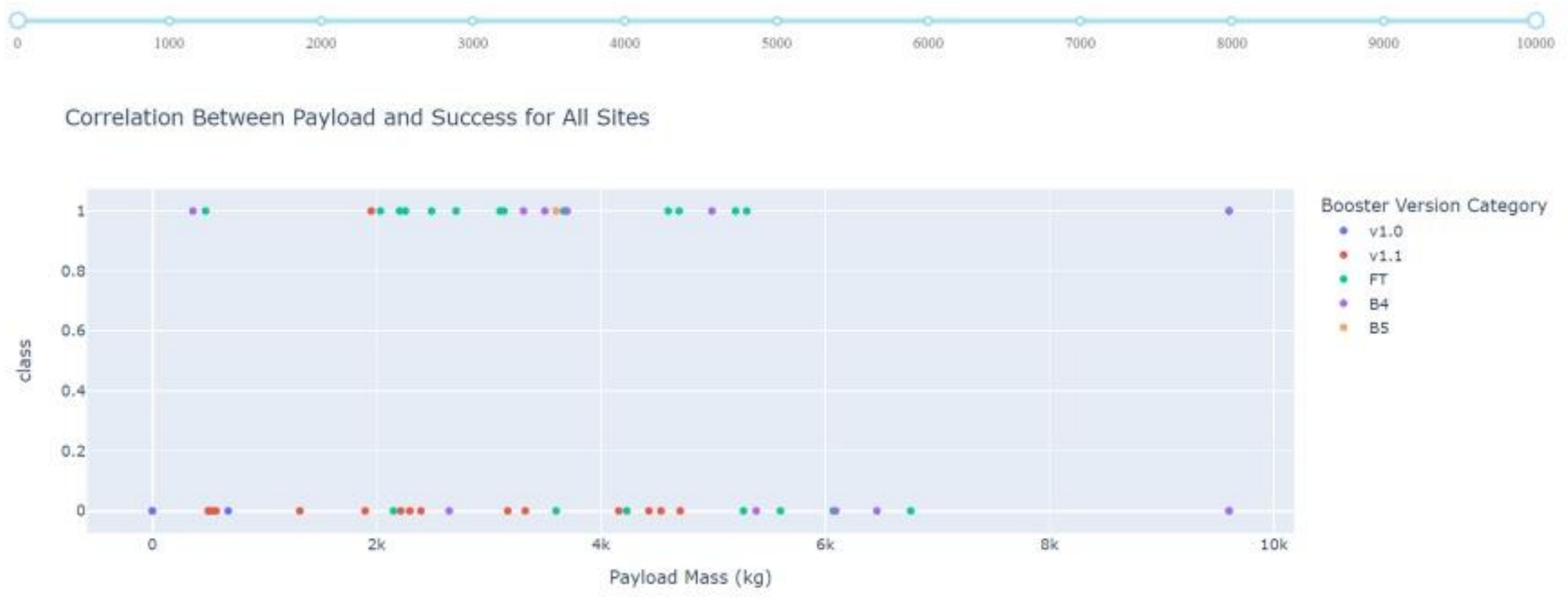
KSC LC-39A

Total Success Launches for Site KSC LC-39A

Class 0 = Fail
Class 1 = Success

Launch site with highest launch success ratio

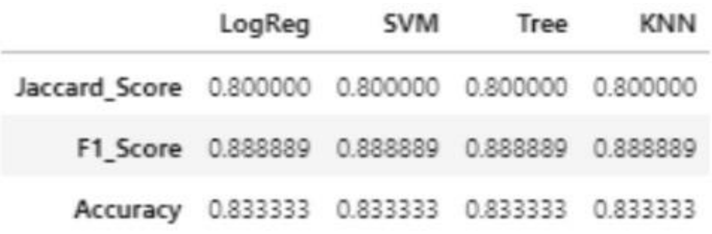
KSC LC-39A has the highest success rate among all the launch sites, with a success rate of 76.9%. Out of a total of 13 launches from this site, 10 were successful, while 3 ended in failure.



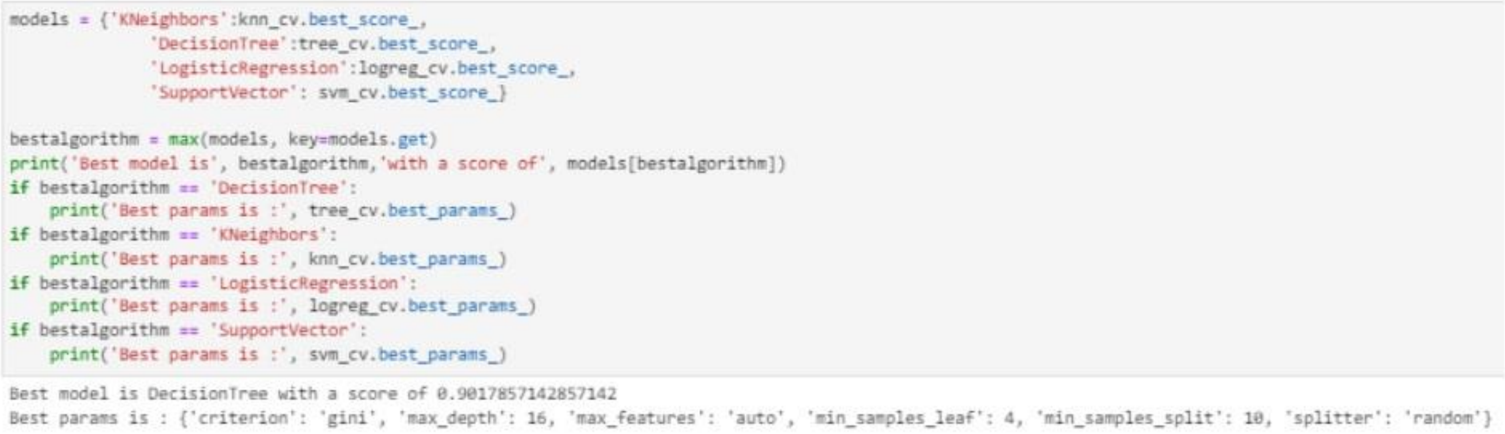
Payload Mass and Success

- By analyzing the data based on booster versions, the following observations can be made:
- Payloads ranging between 2,000 kg and 5,000 kg exhibit the highest success rate.
- The success rate is represented by a value of 1, indicating a successful outcome, while a value of 0 represents an unsuccessful outcome.

Predictive analysis (Classification)



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

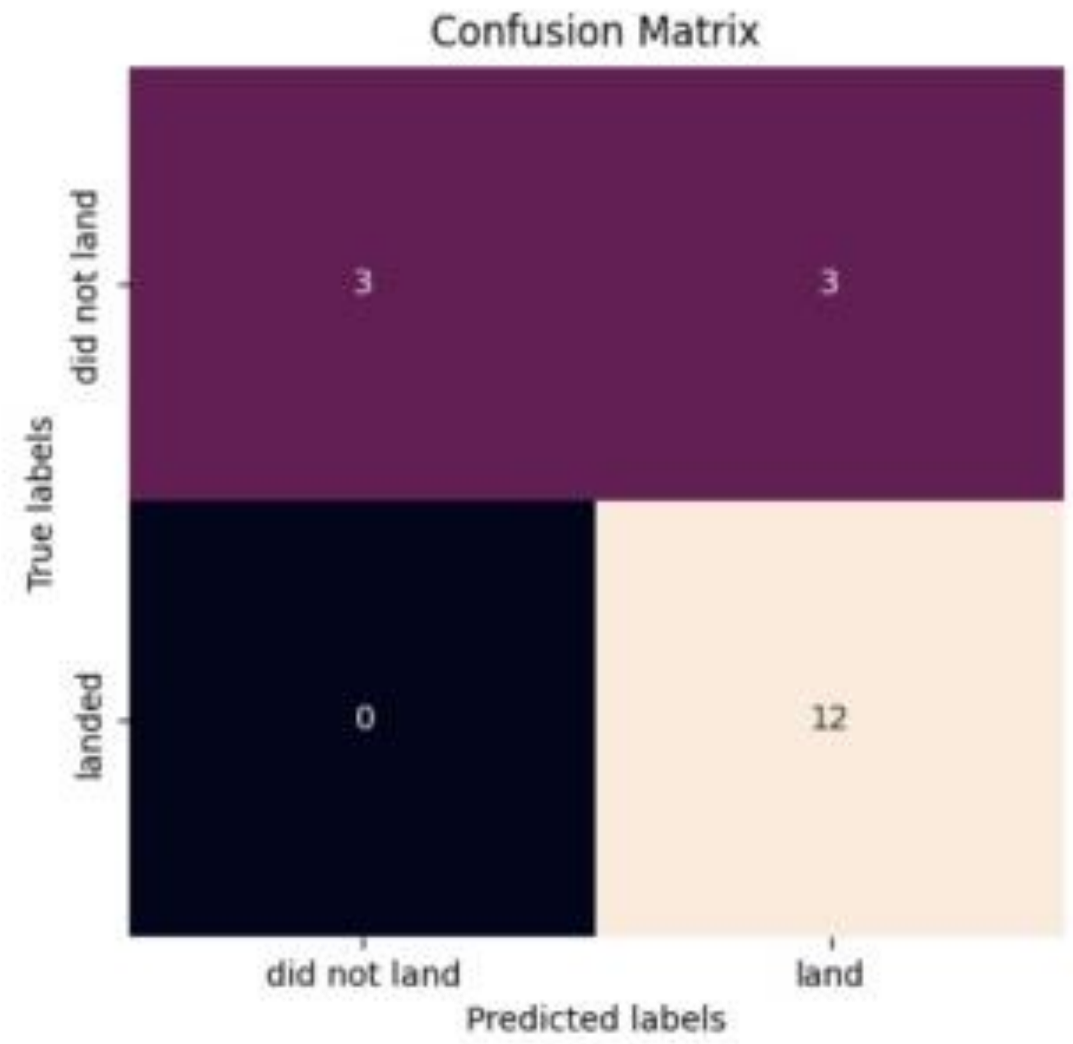


```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9017857142857142
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

Classification Accuracy

Overall, all the models exhibited similar performance levels and achieved similar scores and accuracy. However, when considering the .best_score_ metric, the Decision Tree model slightly outperformed the other models, indicating its superior performance compared to the rest in terms of predictive accuracy.



Confusion Matrix

- Confusion Matrix Outputs:

- True positive: 12
- True negative: 3
- False positive: 3
- False negative: 0

- Upon analyzing the confusion matrix, it is evident that the major issue lies in the occurrence of **false positives**.

Conclusions

Based on the analysis conducted, the following conclusions can be drawn:

- The **success rate** at a launch site tends to increase with the number of flights conducted at that site. A larger flight amount correlates with a higher success rate.
- From 2013 to 2020, there was a noticeable **increase in the launch success rate**, indicating an overall improvement in successful launches during this period.
- **Orbits** ES-L1, GEO, HEO, SSO, and VLEO exhibited the highest success rates among all the orbits considered, suggesting their suitability for successful missions.
- **KSC LC-39A** emerged as the launch site with the highest number of successful launches compared to other sites, indicating its reliability and success in launching missions.
- Among the various **machine learning algorithms explored**, the Decision Tree classifier demonstrated superior performance and is considered the best algorithm for the given task.
- **Payload mass.** Higher payload mass resulted in a higher success ration overall.

These conclusions highlight the relationship between flight amount and success rate, the temporal trend of increasing success rates, the success rates of different orbits, the performance of launch sites, and the effectiveness of the Decision Tree classifier for the analysis at hand.

Thank You!