IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steffen Liedtke
27.05.2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

- SpaceX stands out as the leading entity in the commercial space era, substantially reducing the costs of space travel. The company markets its Falcon 9 rocket launches via its website for $62 million, a significant cut from the $165 million charged by other providers. This reduction in cost is primarily due to SpaceX's capability to reuse the first stage of the rocket. Hence, by predicting whether the first stage will land successfully, we can estimate the launch's cost. Utilizing publicly available data and machine learning models, our objective is to forecast whether SpaceX will reuse the first stage of the rocket.

Problems you want to find answers

- In what ways do factors like the mass of the payload, the launch site, the number of flights, and the type of orbit influence the successful landing of the first stage?
- Has there been a rise in successful landings over the years?
- What is the most suitable algorithm for binary classification in this particular context

Section 1

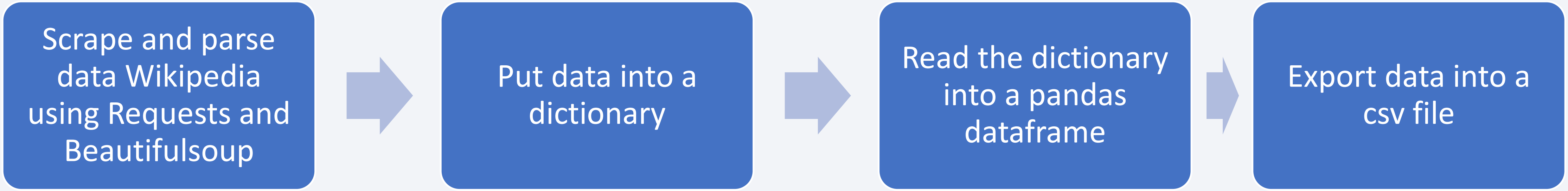
Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Using SpaceX Rest API
    - Using Web Scrapping from Wikipedia
- Perform data wrangling
    - Filtering the data
    - Dealing with missing values
    - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Construction, optimization, and assessment of classification models to guarantee the most optimal results

Data Collection – Web Scraping



Scrape and parse data Wikipedia using Requests and Beautifulsoup

Put data into a dictionary

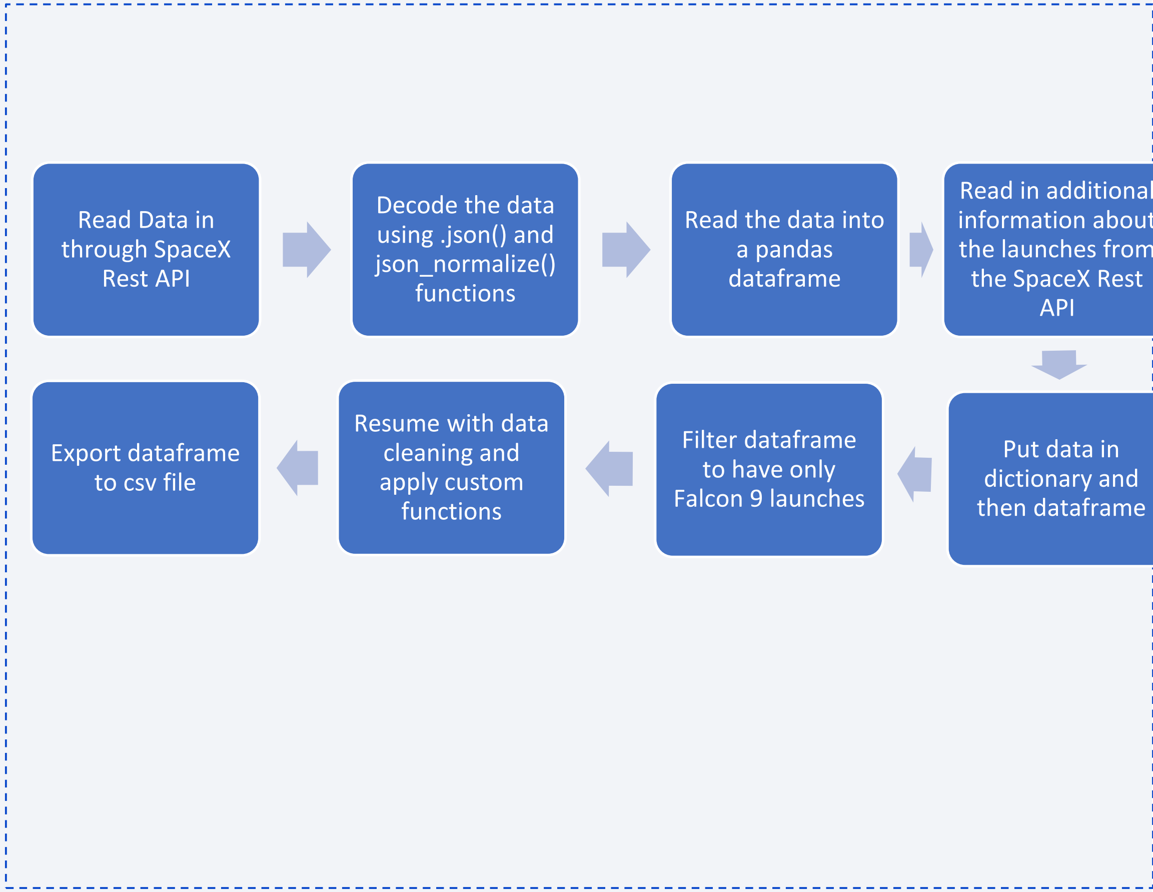
Read the dictionary into a pandas dataframe

Export data into a csv file

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Github URL:
- https://github.com/SteffenLi/IBM-Course/blob/master/jupyter-labs-spacex-data-collection-api.ipynb



Read Data in through SpaceX Rest API

Decode the data using .json() and json_normalize() functions

Read the data into a pandas dataframe

Read in additional information about the launches from the SpaceX Rest API

Put data in dictionary and then dataframe

Filter dataframe to have only Falcon 9 launches

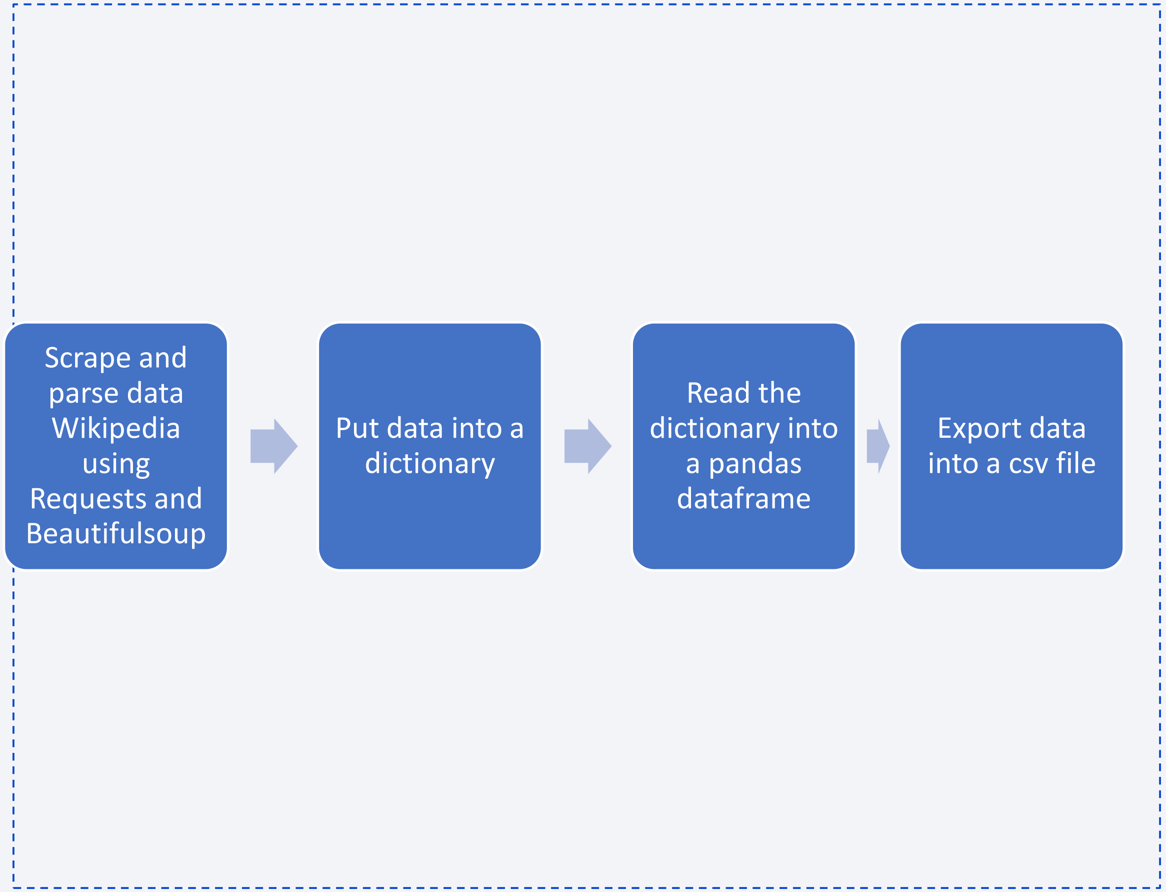
Resume with data cleaning and apply custom functions

Export dataframe to csv file

Data Collection - Scraping

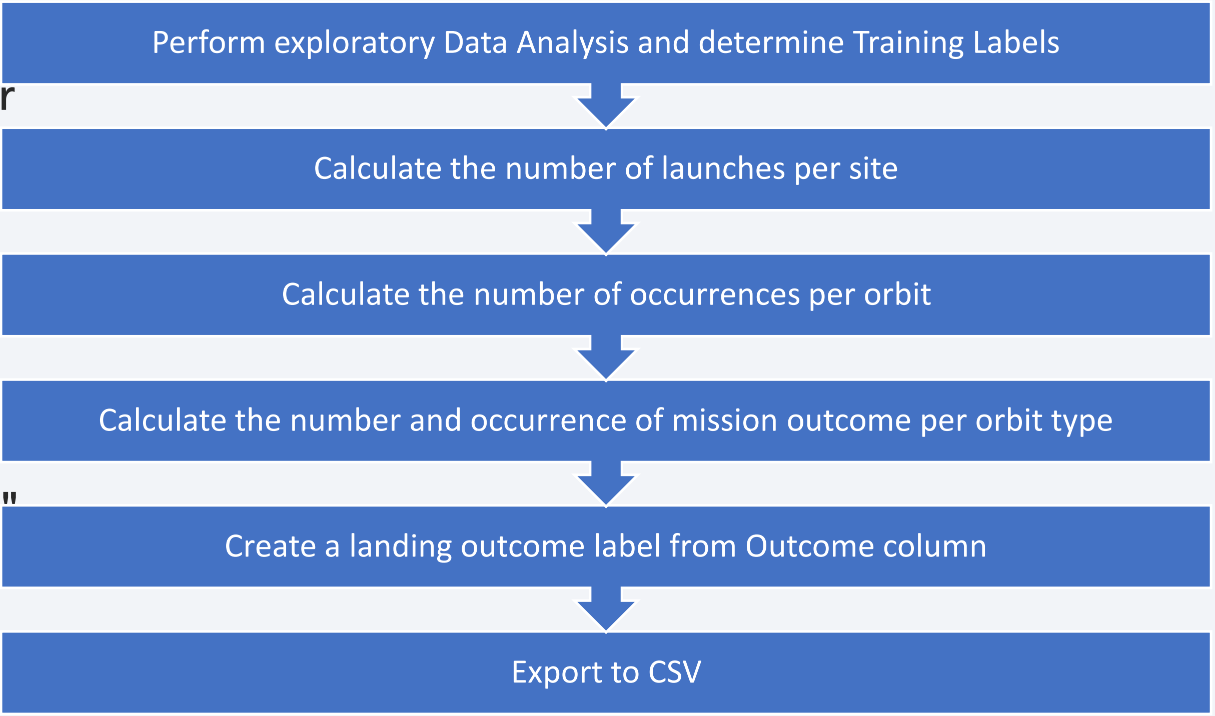
- Present your web scraping process using key phrases and flowcharts

- Github URL:
- https://github.com/SteffenLi/IBM-Course/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb


Scrape and parse data Wikipedia using Requests and Beautifulsoup
Put data into a dictionary
Read the dictionary into a pandas dataframe
Export data into a csv file

Data Wrangling

- The dataset includes numerous instances where the booster did not achieve a successful landing. Sometimes, despite ar attempt being made, the landing failed due to various mishaps. For instance, "True Ocean" implies a successful landing in a designated ocean region, while "False Ocean" signifies an unsuccessful attempt to land in a specified ocean region. Similarly, "True RTLS" indicates a successful landing on a ground pad, while "False RTLS" suggests an unsuccessful landing attempt on a ground pad. "True ASDS" signifies that the landing on a drone ship was successful, and "False ASDS" denotes a failed landing attempt on a drone ship. These outcomes have been primarily converted into Training Labels, where "1" indicates a successful booster landing and "0" stands for an unsuccessful landing.



Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches per site

Calculate the number of occurrences per orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Export to CSV

- Github URL:

https://github.com/SteffenLi/IBM-Course/blob/master/Spacex%20Data%20Wrangling.ipynb

EDA with Data Visualization

- **Scatter plots**
- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Scatter plots illustrate the relationships between different variables. If a correlation exists, these plots can be utilized in machine learning models.

- **Bar charts**
- Comparison of categories:
- Launch Site
- Orbit Type
- Bar charts are used to compare discrete categories and showcase the relationship between the specific categories and a measured value.
- **Line chart**
- Success Rate Yearly Trend
- Line charts depict trends in data over time, particularly useful for time series analysis. The "Success Rate Yearly Trend" chart shows the variation in success rates over the years.

- **Github Link**

https://github.com/SteffenLi/IBM-Course/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Unique Launch Sites: Display the names of the launch sites involved in the space mission.
- Launch Sites starting with 'CCA': Show 5 records of launch sites that begin with the string 'CCA'.
- Payload Mass by NASA (CRS): Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Average Payload Mass (Booster Version F9 v1.1): Determine the average payload mass carried by booster version F9 v1.1.
- First Successful Ground Pad Landing: List the date when the first successful landing outcome was achieved on a ground pad.
- Successful Drone Ship Landings (Payload Range): Identify boosters that successfully landed on a drone ship with a payload mass greater than 4000 and less than 6000.
- Total Mission Outcomes: List the total number of successful and failure mission outcomes.
- Booster Versions with Maximum Payload: Identify the booster versions that carried the maximum payload mass.
- Failed Drone Ship Landings (2015): List the failed landing outcomes in drone ships, along with their booster versions and launch site names for the months in 2015.
- Landing Outcomes Ranking (2010-06-04 to 2017-03-20): Rank the count of landing outcomes (Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017, in descending order.

Github URL:

https://github.com/SteffenLi/IBM-Course/blob/master/EDA%20SQL.ipynb

Build an Interactive Map with Folium

Markers of all Launch Sites:

- NASA Johnson Space Center: Placed a marker with a circle, popup label, and text label on the map, representing the NASA Johnson Space Center. The marker was positioned using latitude and longitude coordinates to indicate its start location.
- Launch Sites: Added markers with circles, popup labels, and text labels for all launch sites on the map. The latitude and longitude coordinates were utilized to display the geographical locations of the launch sites and their proximity to the Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Success and Failed Launches: Incorporated colored markers on the map to represent successful launches (green) and failed launches (red). The markers were grouped using a marker cluster to facilitate identification of launch sites with relatively high success rates.

Distances between a Launch Site to its proximities:

- Distance Visualization: Incorporated colored lines on the map to display the distances between the launch site KSC LC-39A (used as an example) and its proximities, such as the railway, highway, coastline, and closest city.

Github URL

https://github.com/SteffenLi/IBM-Course/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List

- Implemented a dropdown list that allows the uer to select a specific launch site.

Pie Chart showing Success Launches (All Sites/Certain Site)

- Introduced a pie chart to visually represent the total count of successful launches across all sites. If a specific launch site is selected from the dropdown list, the pie chart also displays the success versus failed launch counts for that particular site

Slider of Payload Mass Range

- Incorporated a slider control that enables the user to select a desired payload mass range.

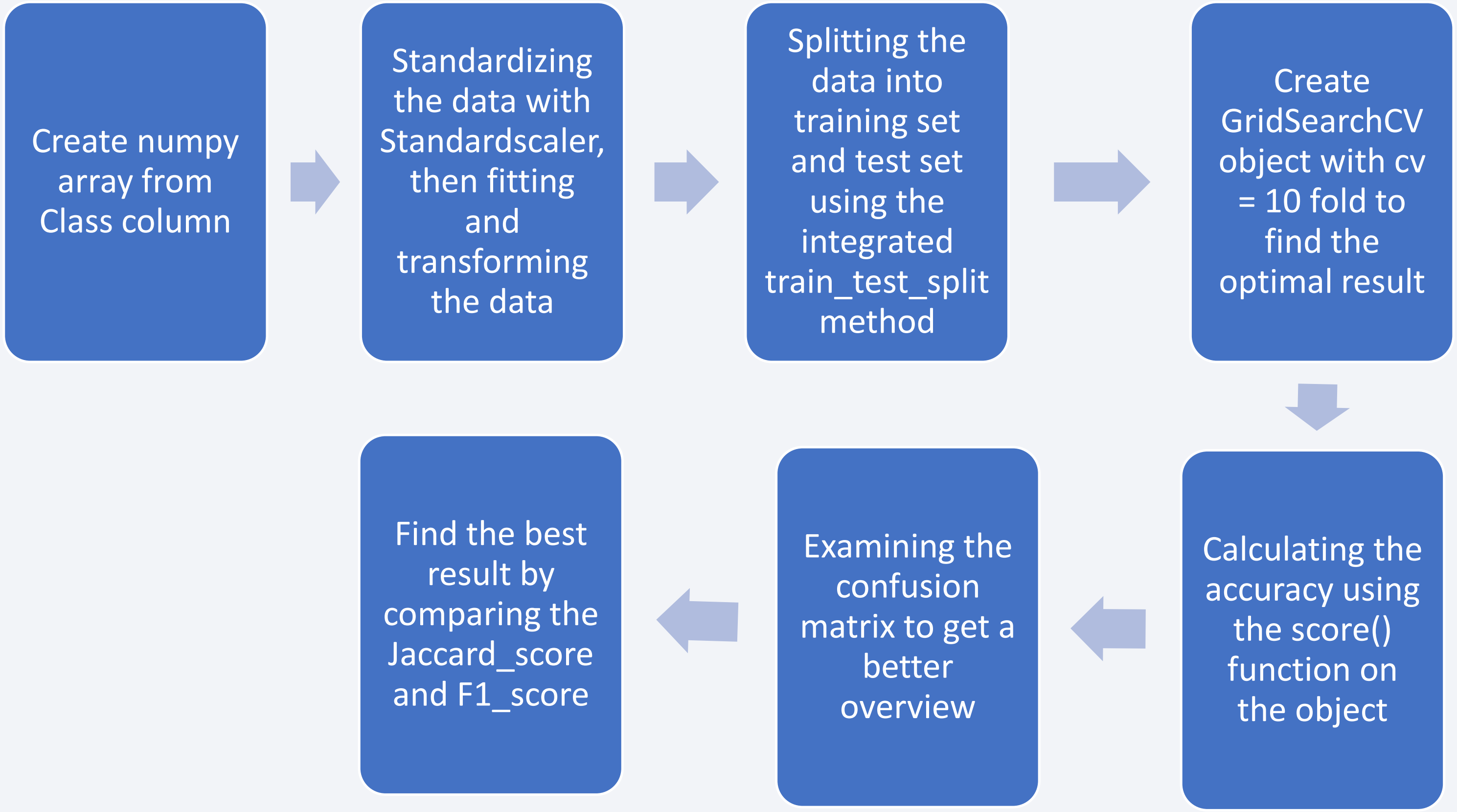
Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Generated a scatter chart illustrating the relationship between payload mass and launch success rate. This chart compares different booster versions and their corresponding payload masses, providing insights into how payload mass influences the success rate of the launches

Github URL:

https://github.com/SteffenLi/IBM-Course/blob/master/dash%20script.py

Predictive Analysis (Classification)



Create numpy array from Class column

Standardizing the data with Standardscaler, then fitting and transforming the data

Splitting the data into training set and test set using the integrated train_test_split method

Create GridSearchCV object with cv = 10 fold to find the optimal result

Calculating the accuracy using the score() function on the object

Examining the confusion matrix to get a better overview

Find the best result by comparing the Jaccard_score and F1_score

Github URL:

https://github.com/SteffenLi/IBM-Course/blob/master/Spacex%20Machine%20Learning.ipynb

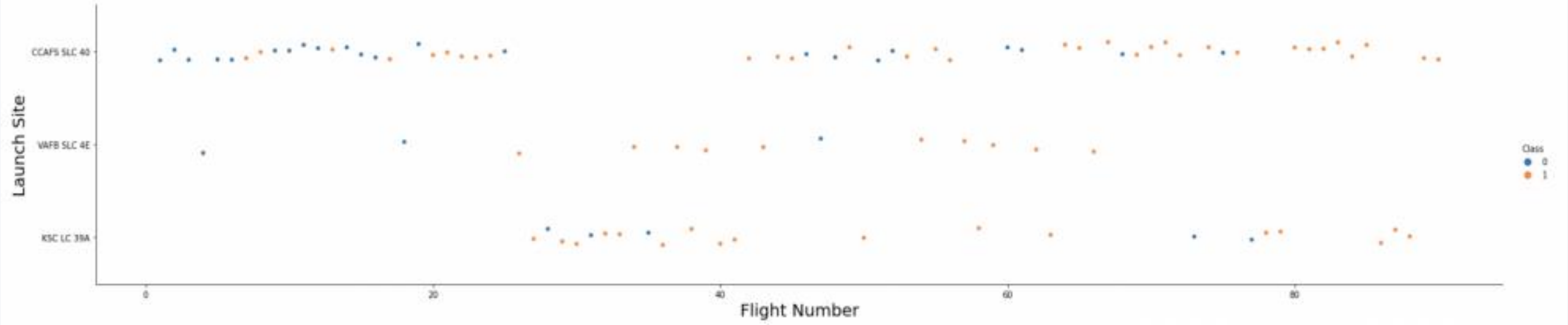
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation

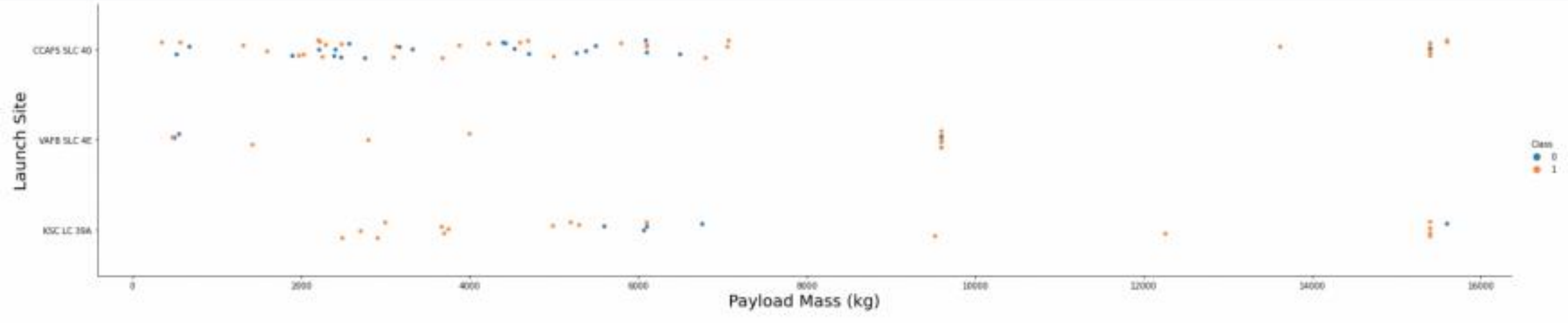
• Historical trend: The earliest flights in the dataset resulted in failures, while the latest flights achieved a higher success rate, indicating an improvement over time.

• Launch site distribution: The CCAFS SLC 40 launch site accounted for approximately half of all launches recorded in the dataset.

• Higher success rates: VAFB SLC 4E and KSC LC 39A launch sites exhibited comparatively higher success rates, suggesting better performance in terms of successful launches.

• Assumption of increasing success rate: It is reasonable to assume that each new launch has a higher rate of success, implying a positive trend of improving success rates over time.

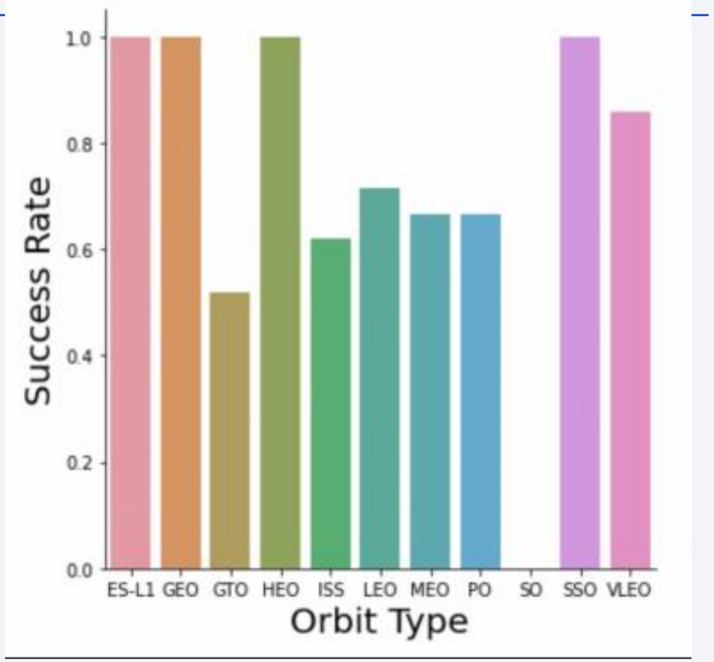
Payload vs. Launch Site



Explanation

- Historical trend: The earliest flights in the dataset resulted in failures, while the latest flights achieved a higher success rate, indicating an improvement over time.
- Launch site distribution: The CCAFS SLC 40 launch site accounted for approximately half of all launches recorded in the dataset.
- Higher success rates: VAFB SLC 4E and KSC LC 39A launch sites exhibited comparatively higher success rates, suggesting better performance in terms of successful launches.
- Assumption of increasing success rate: It is reasonable to assume that each new launch has a higher rate of success, implying a positive trend of improving success rates over time.

Success Rate vs. Orbit Type



Explanation

Orbits with 100% success rate

- ES-L1
- GEO
- HEO
- SSO

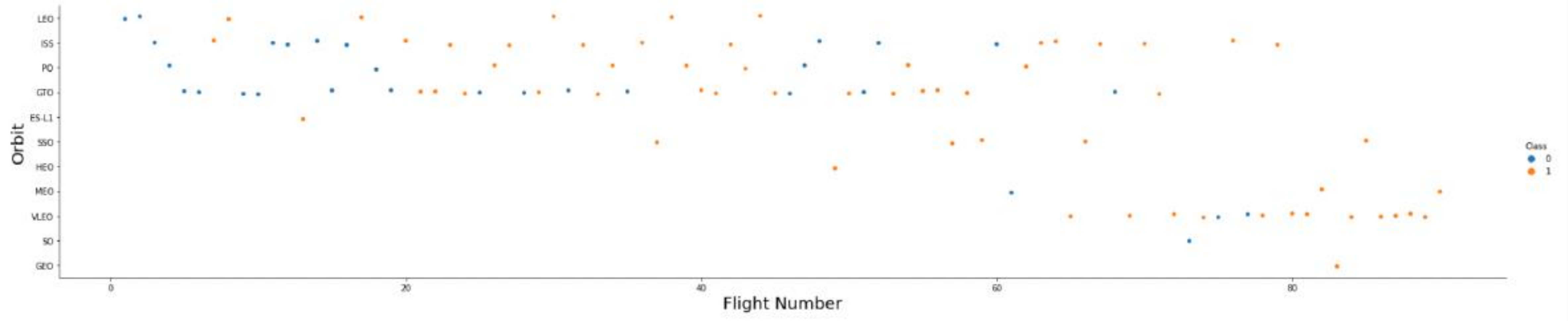
Orbits with 0% success rate

- SO

Orbits with success rate between 50% and 85%

- GTO
- ISS
- LEO
- MEO
- PO

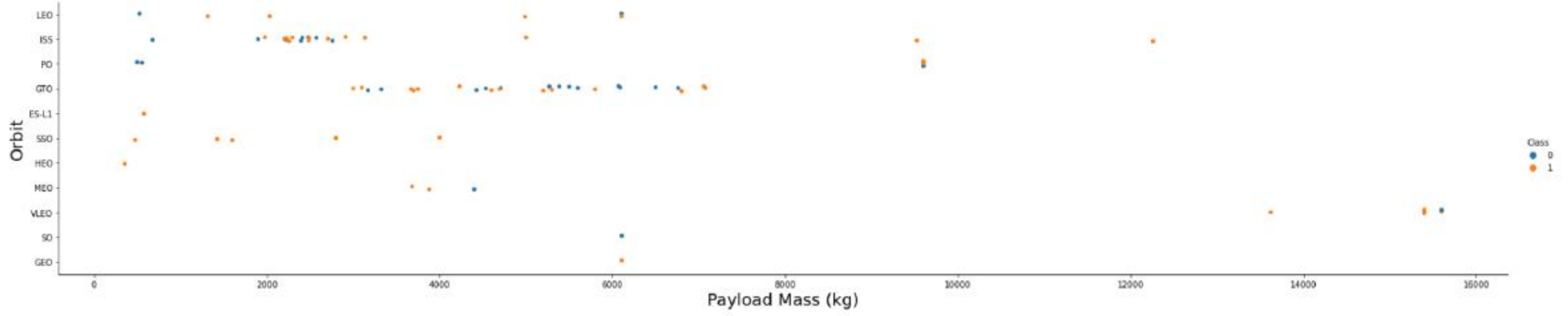
Flight Number vs. Orbit Type



Explanation

- In the LEO orbit, there appears to be a relationship between the success rate and the number of flights. As the number of flights increases, the success rate also tends to increase.
- On the other hand, in the GTO orbit, there doesn't seem to be a noticeable relationship between the flight number and the success rate. The success rate does not show a consistent pattern or correlation with the number of flights in this orbit.

Payload vs. Orbit Type



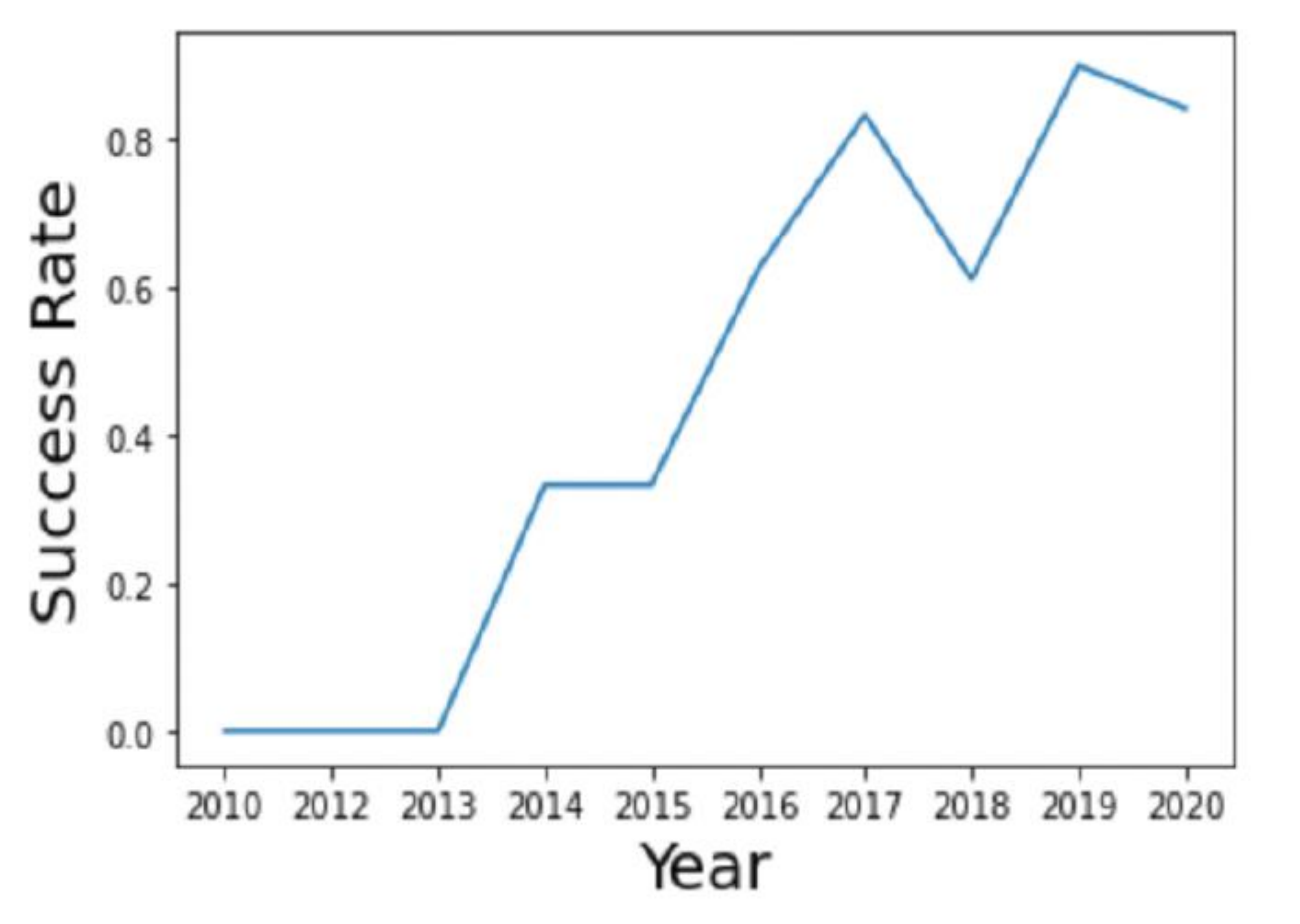
Explanation

- Heavy payloads have a negative influence on GTO orbits, meaning that as the payload mass increases, the success rate tends to decrease in GTO orbits.
- However, heavy payloads have a positive influence on GTO and Polar LEO (ISS) orbits. In these orbits, as the payload mass increases, the success rate tends to increase as well..

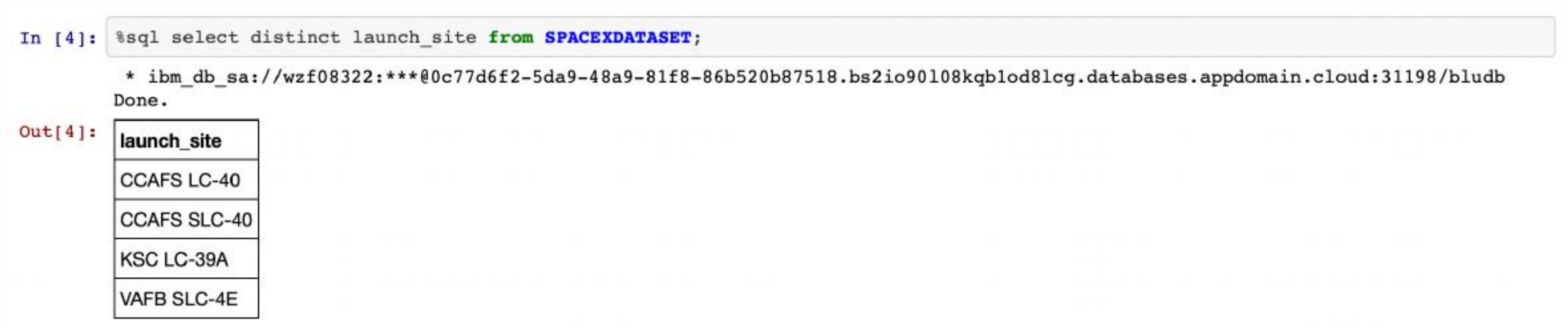
Launch Success Yearly Trend

Explanation

- The success rate has shown a continuous increase from 2013 until 2020.



All Launch Site Names



```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

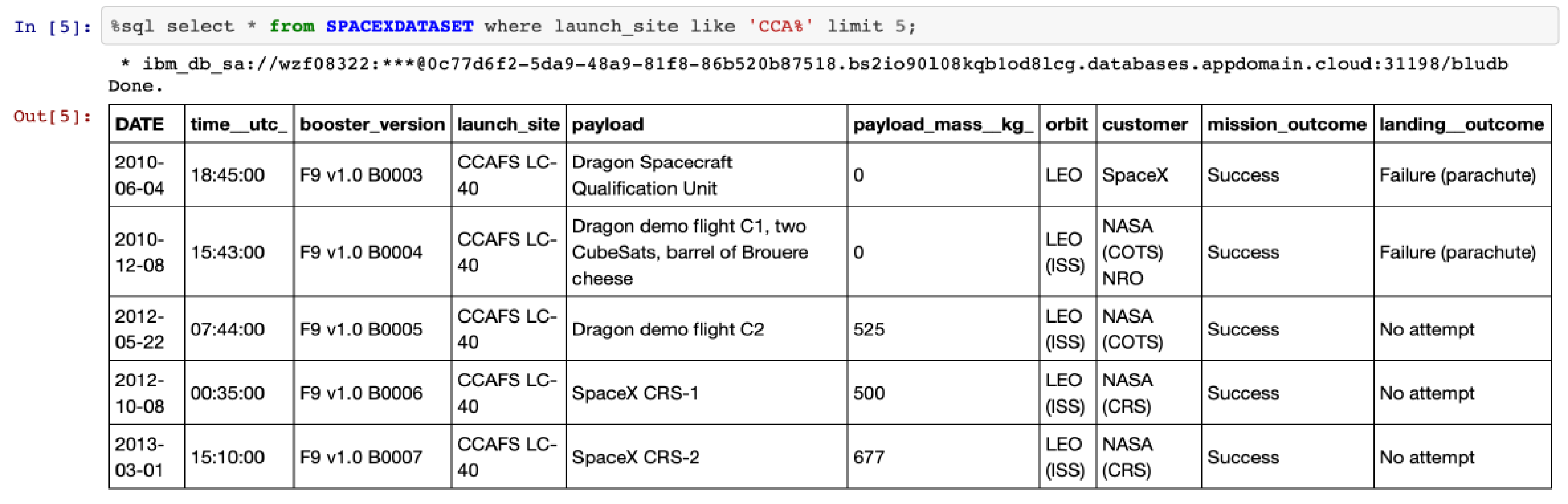
Out[4]:

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

Explanation

- Unique Launch Sites: Showing the names of the distinct launch sites involved in the space mission.

Launch Site Names Begin with 'CCA'



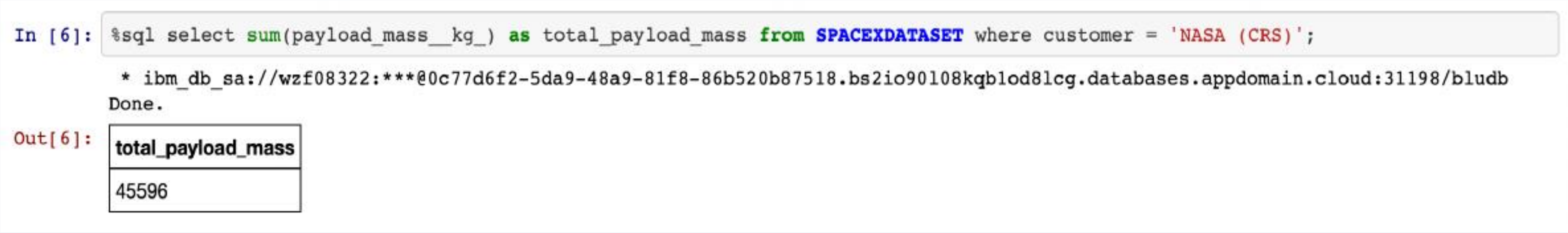
```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[5]:
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Explanation

- Unique Launch Sites: Showing the names of the distinct launch sites involved in the space mission.

Total Payload Mass



```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

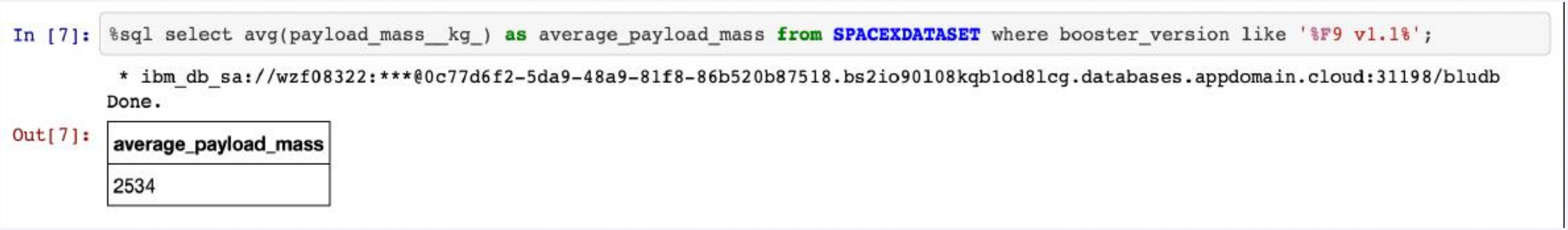
Out[6]:

total_payload_mass
45596

Explanation

- Using SQL to give me the sum (total) of the payload mass from NASA (CRS)

Average Payload Mass by F9 v1.1



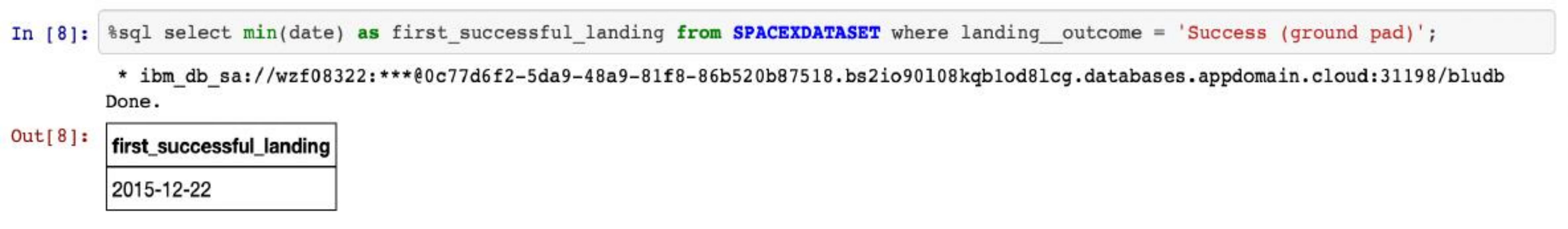
```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
        Done.

Out[7]:
```

average_payload_mass
2534

Explanation

- This query calculates the average payload mass by filtering the space mission records for the specific booster version 'F9 v1.1' and then calculates the average using the AVG() function..

First Successful Ground Landing Date



```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

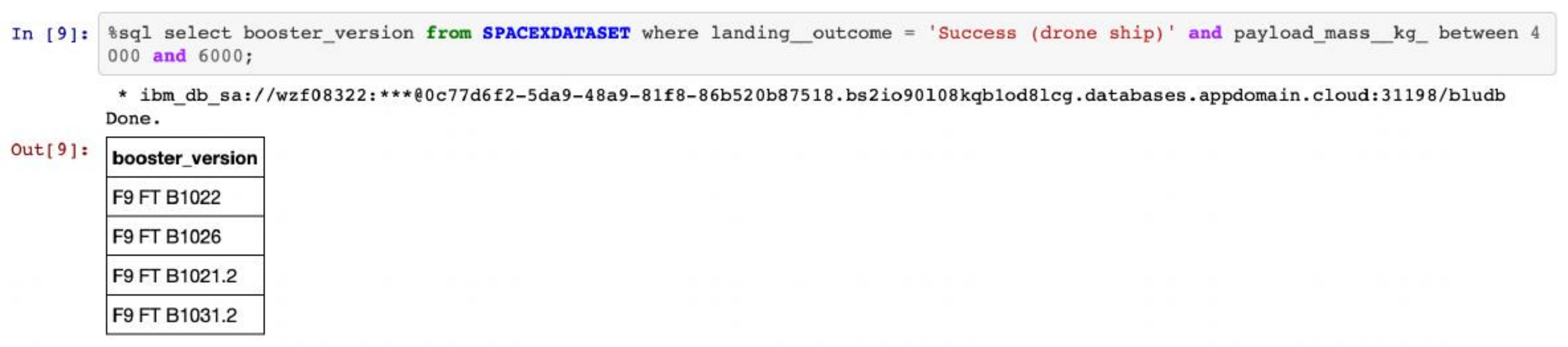
Out[8]:

first_successful_landing
2015-12-22

Explanation

- This query retrieves the minimum (earliest) date from the space mission records where the landing outcome is recorded as 'Success (ground pad)'. It provides the date when the first successful landing on the ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000



In [9]:
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

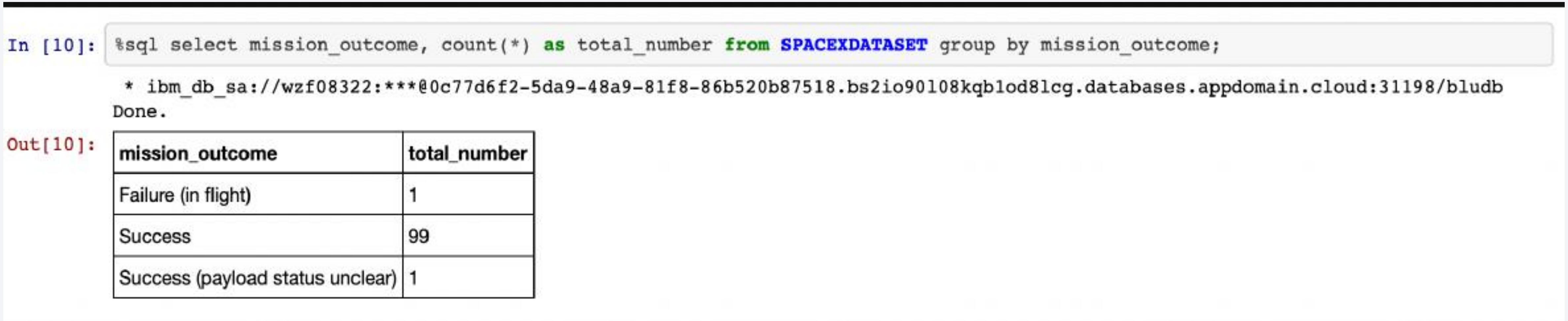
Out[9]:

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

Explanation

- This query retrieves the names of the boosters from the space mission records where the landing outcome is recorded as 'Success (drone ship)' and the payload mass is within the specified range of greater than 4000 and less than 6000.

Total Number of Successful and Failure Mission Outcomes



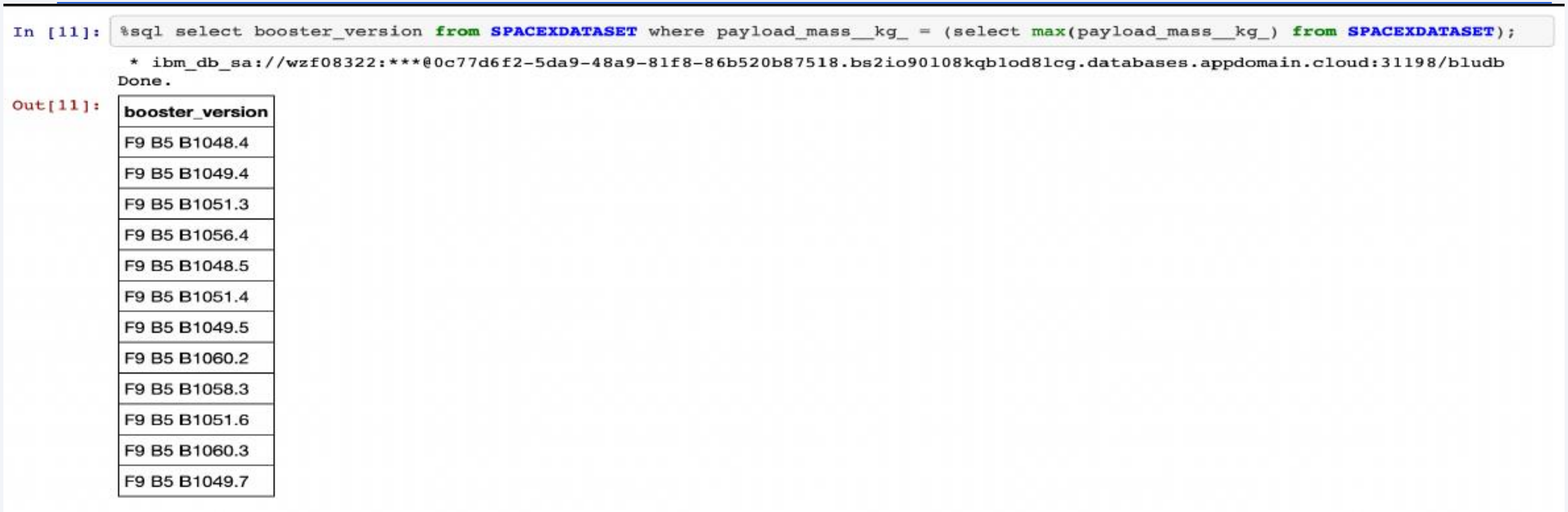
```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[10]:
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation

- This query counts the occurrences of each unique landing outcome (such as 'Success' or 'Failure') in the space mission records and provides the total count for each outcome. The result includes the landing outcome and the corresponding count.

Boosters Carried Maximum Payload



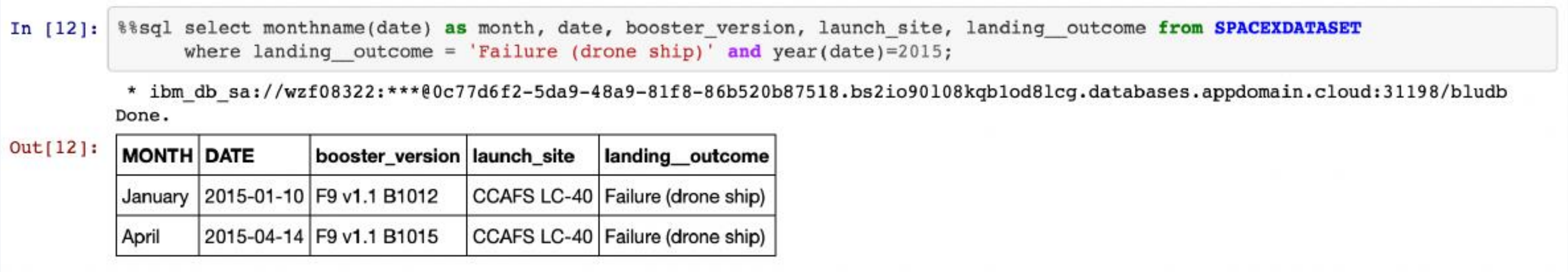
```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[11]:
```

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

Explanation

- This query retrieves the names of the booster versions from the space mission records where the payload mass is equal to the maximum payload mass found in the dataset. It provides the names of the booster versions that have carried the highest payload mass

2015 Launch Records



```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
                where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

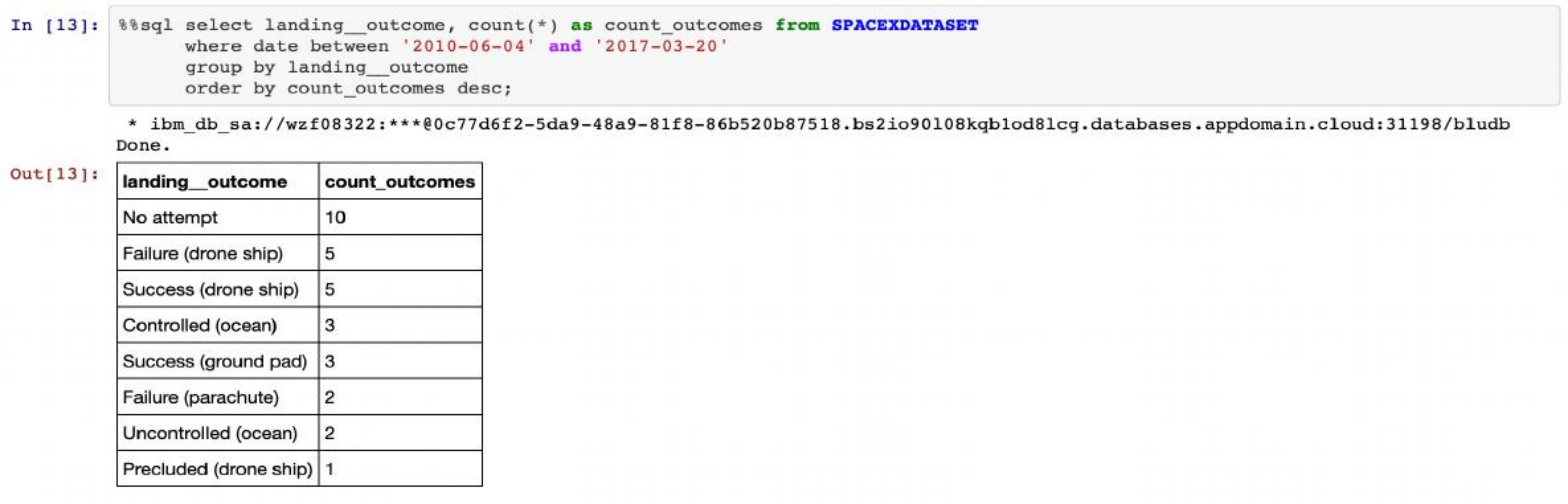
Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation

- This query retrieves the landing outcomes, booster versions, and launch site names from the space mission records where the landing outcome is a failure on a drone ship and the date falls within the year 2015. The result will display the corresponding landing outcomes, booster versions, and launch site names for the failed missions in the specified time period.

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



```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Explanation

- This query retrieves the landing outcomes and their respective counts from the space mission records, limited to the specified date range. The results are then grouped by the landing outcome and ordered in descending order based on the count. This provides a ranking of the landing outcomes based on their occurrence frequency within the given time period.

Section 3

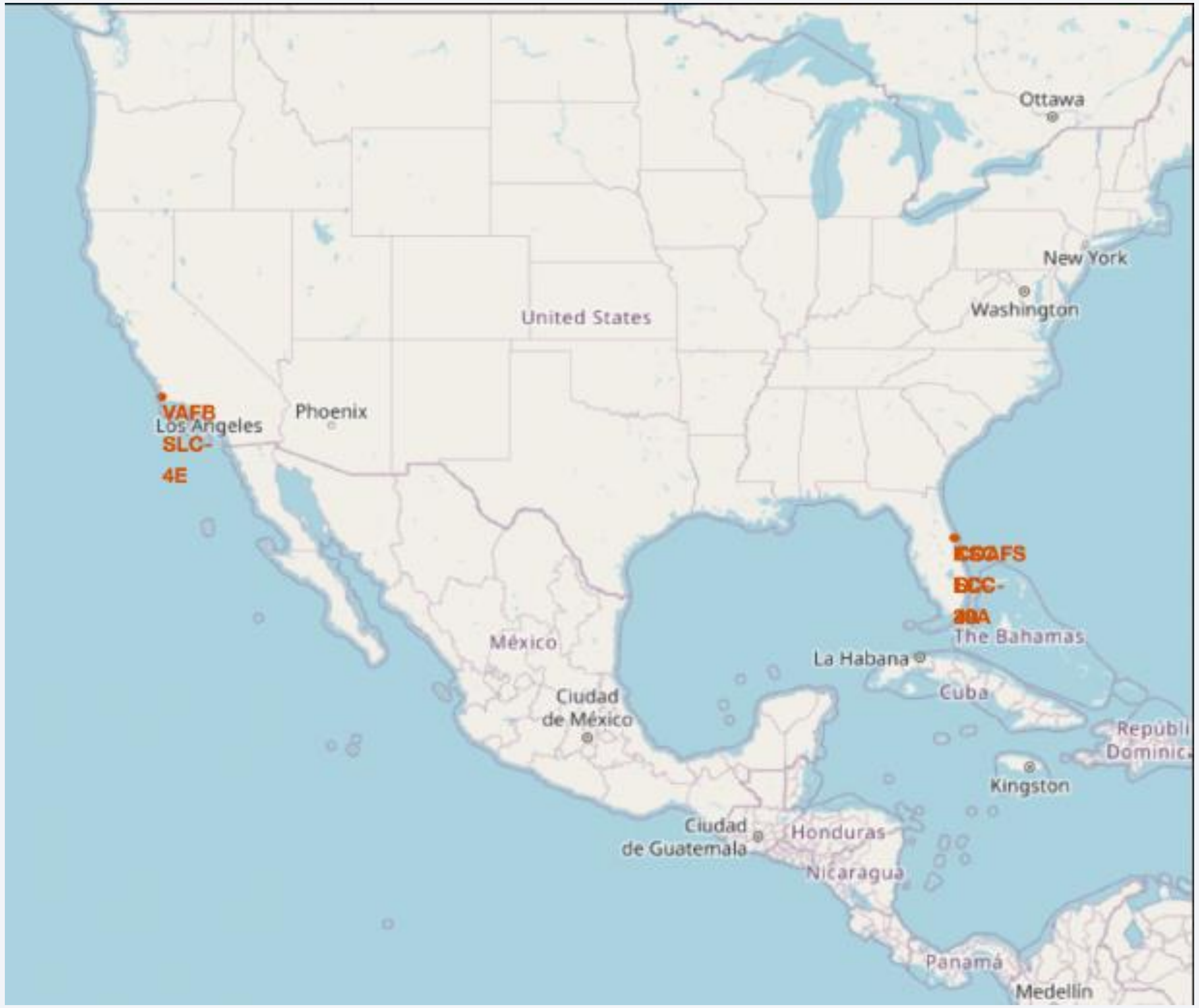
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

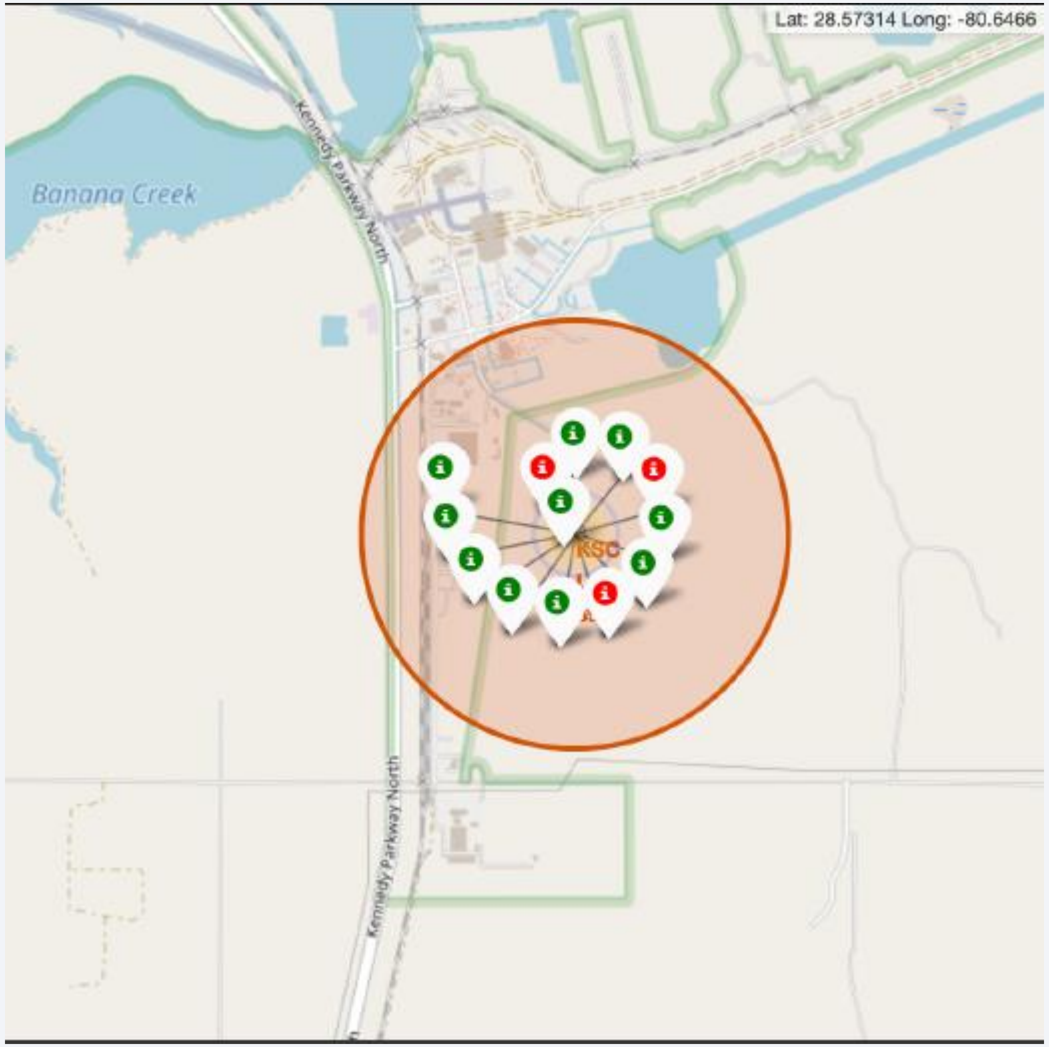
Explanation

- Proximity to Equator: Most launch sites are located near the Equator line. The reason behind this is that the Earth's rotation is fastest at the Equator, moving at a speed of 1670 km/hour. When a spacecraft is launched from the Equator, it already inherits this high speed from the Earth's rotation. This speed helps the spacecraft maintain the necessary velocity to stay in orbit, thanks to the principle of inertia.

- Coastal Location: All launch sites are strategically situated in close proximity to coastlines. Launching rockets towards the ocean helps minimize the risk of debris falling or explosions occurring near populated areas. By directing launches over water, the potential danger to human populations is significantly reduced.



<Folium Map Screenshot 2>



Explanation

- Color-coded markers: The color-labeled markers provide a convenient way to identify launch sites with relatively high success rates.

Green Marker: Indicates a successful launch.

Red Marker: Represents a failed launch.

- Launch Site KSC LC-39A: This particular launch site, KSC LC-39A, stands out with a notably high success rate, indicating its exceptional performance and reliability.

<Folium Map Screenshot 3>

Explanation

- Visual analysis of launch site KSC LC-39A reveals its proximity to various features:

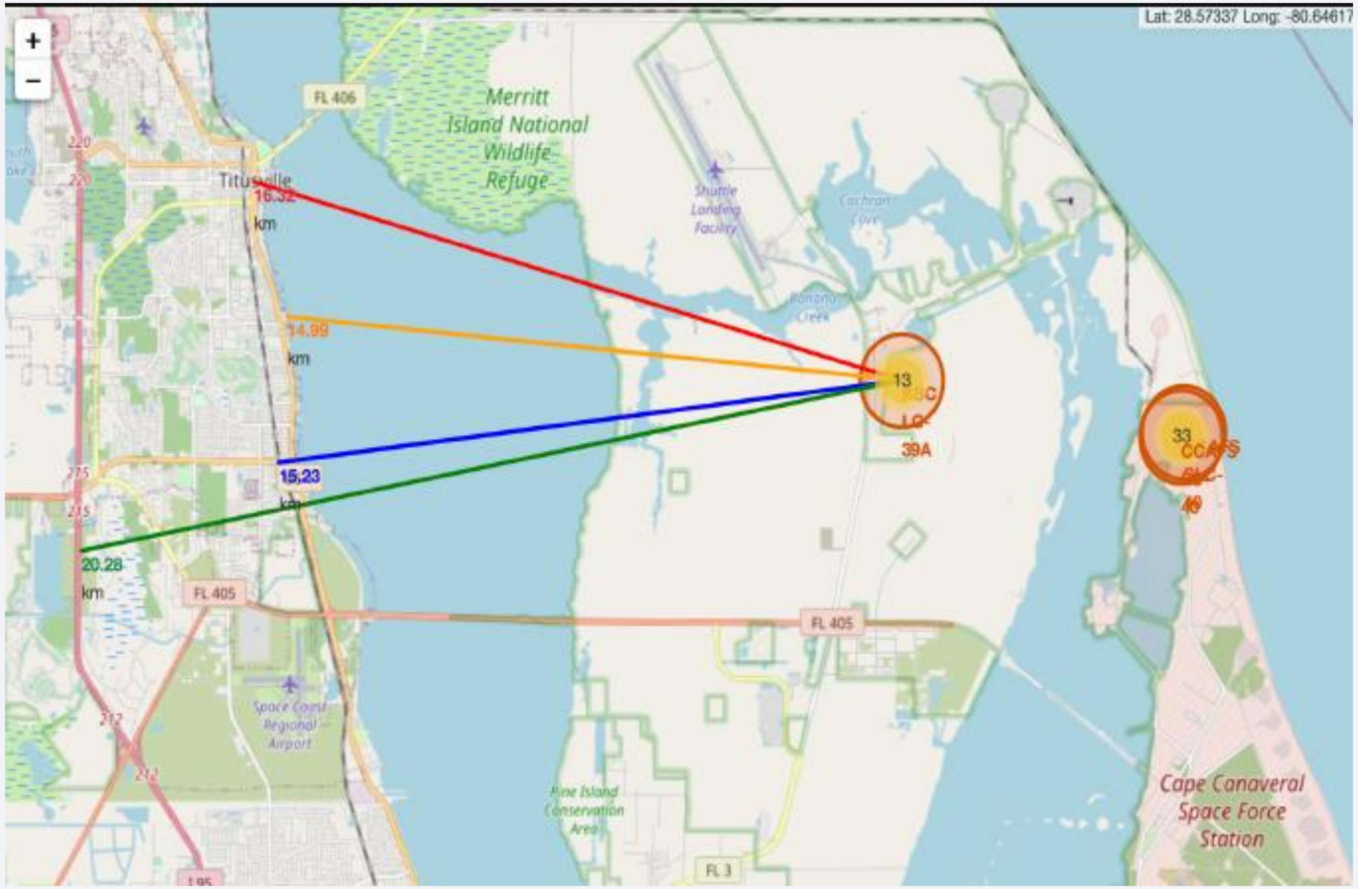
Railway: The launch site is relatively close to a railway, with a distance of 15.23 km.

Highway: It is also in close proximity to a highway, located approximately 20.28 km away.

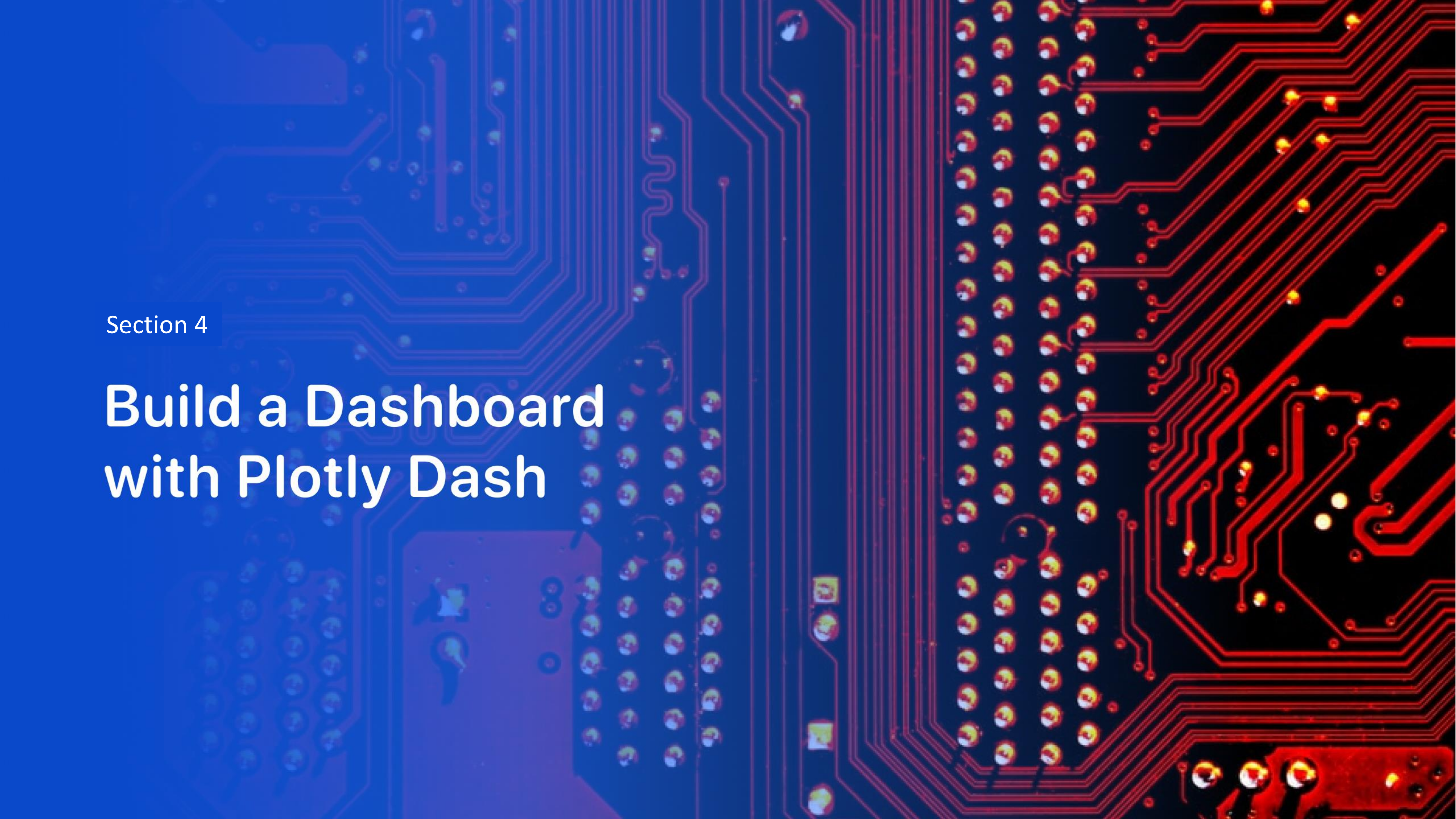
Coastline: The launch site is relatively close to the coastline, with a distance of approximately 14.99 km.

- Close proximity to Titusville: KSC LC-39A is also relatively close to its closest city, Titusville, located approximately 16.32 km away.

- Consideration of safety: Failed rockets, due to their high speed, can cover distances of 15-20 km within seconds. This poses a potential danger to populated areas. Hence, launching rockets towards the ocean and minimizing the risk of debris falling near populated areas becomes crucial in ensuring safety.



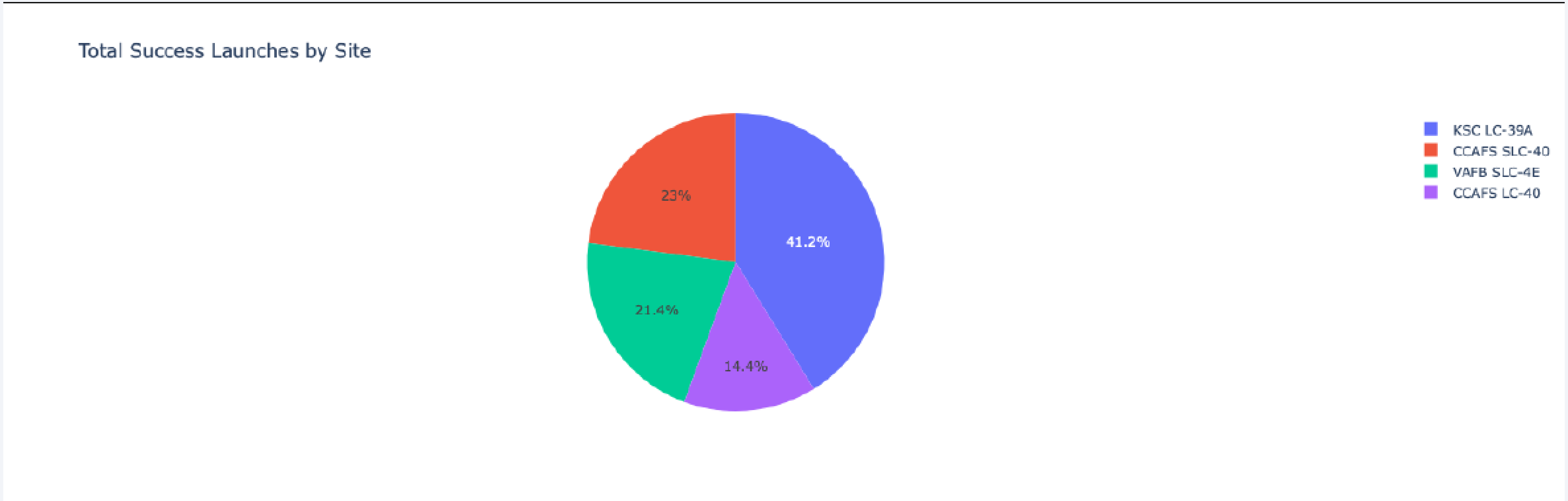
Lat: 28.57337 Long: -80.64617
+
−
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
KSC LC-39A
33
CCAFS SLC-40
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
FL 406
FL 405
FL 3
Space Coast Regional Airport
Cape Canaveral Space Force Station



Section 4

Build a Dashboard with Plotly Dash

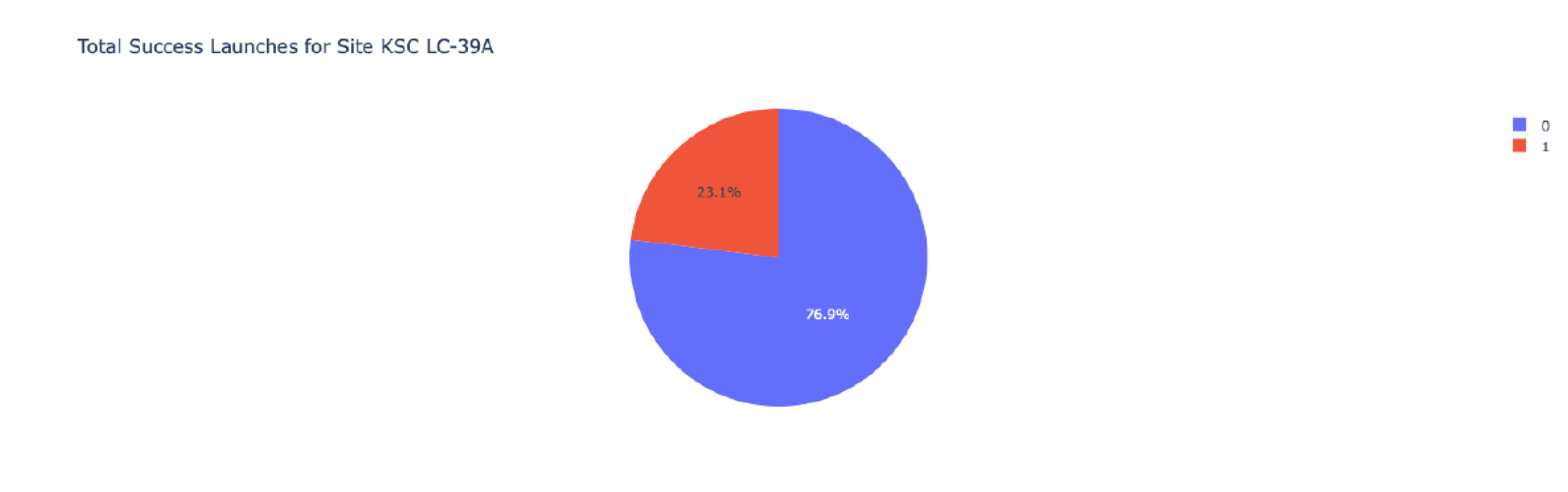
Total Success Launches by Site



Explanation

- The pie chart provides clear evidence that among all the launch sites, KSC LC-39A stands out with the highest number of successful launches.

Launch site with highest success rate



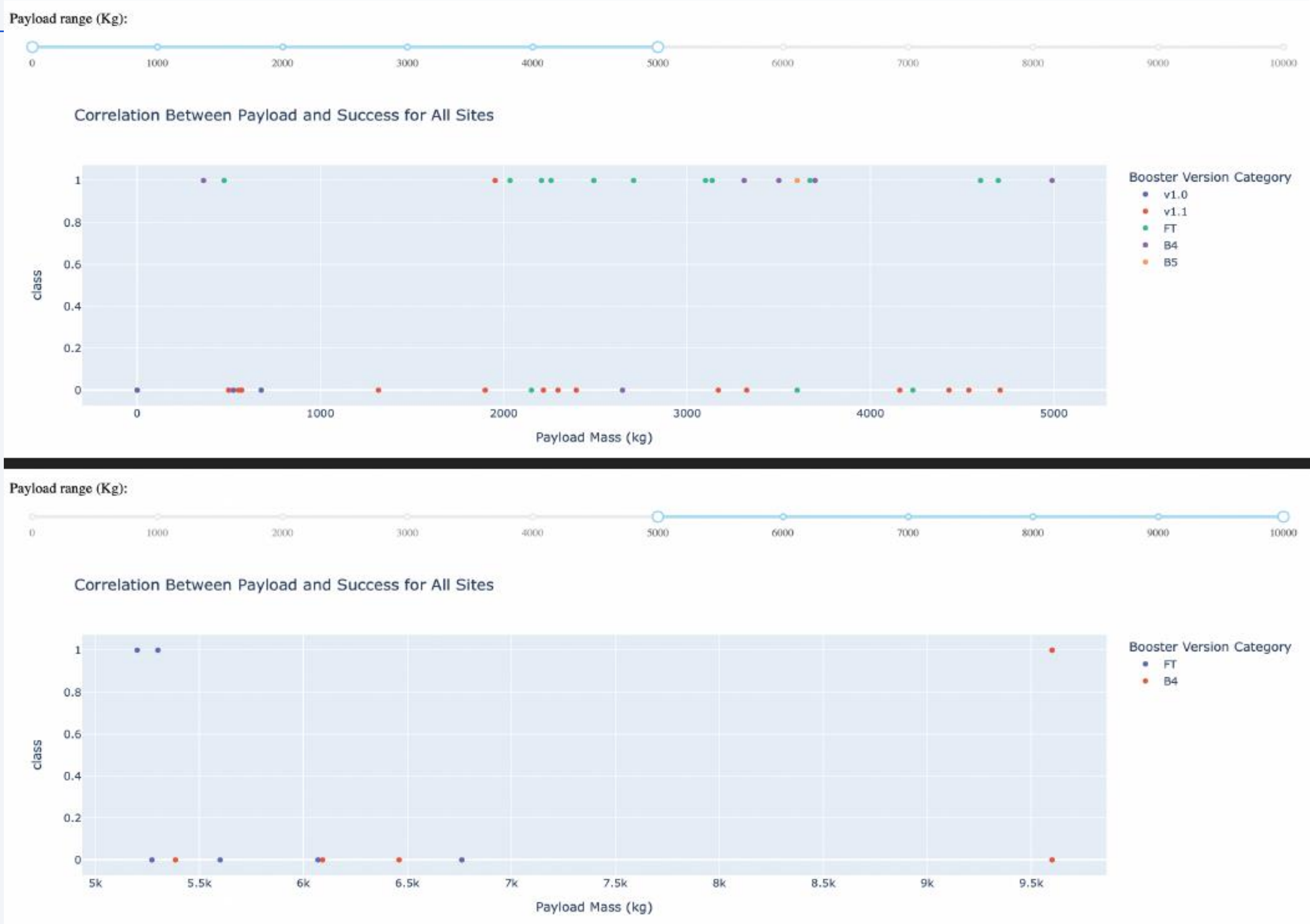
Explanation

- KSC LC-39A boasts the highest launch success rate of 76.9%, having achieved 10 successful landings compared to only 3 failed landings.

<Dashboard Screenshot 3>

Explanation

- The charts clearly indicate that payloads ranging from 2000 kg to 5500 kg exhibit the highest success rate.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

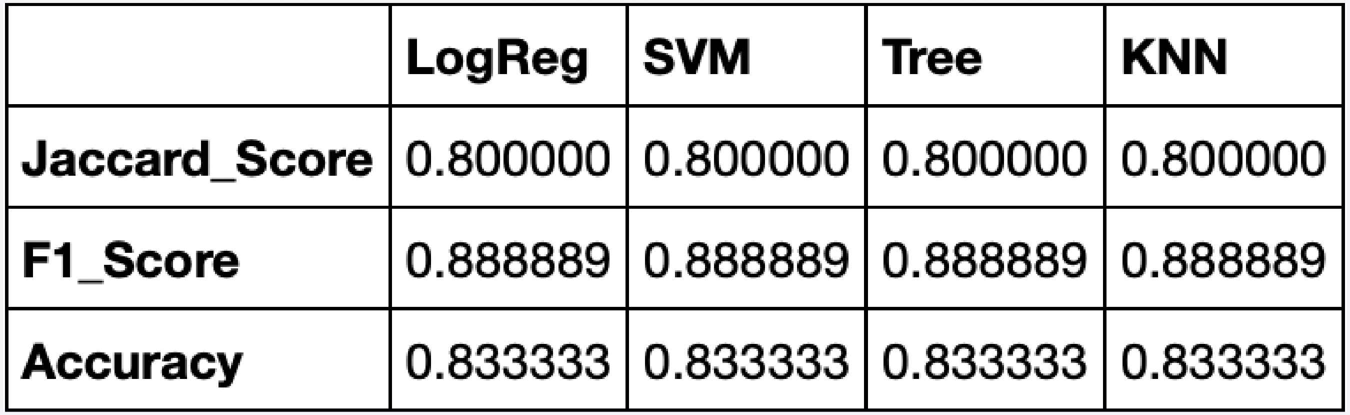
Explanation

- • Test Set evaluation: Based on the scores of the Test Set, it is inconclusive to determine which method performs the best.

- • Sample size impact: The similarity in Test Set scores could be attributed to the small size of the test sample, which consisted of only 18 samples. To address this, we evaluated all methods using the entire Dataset.

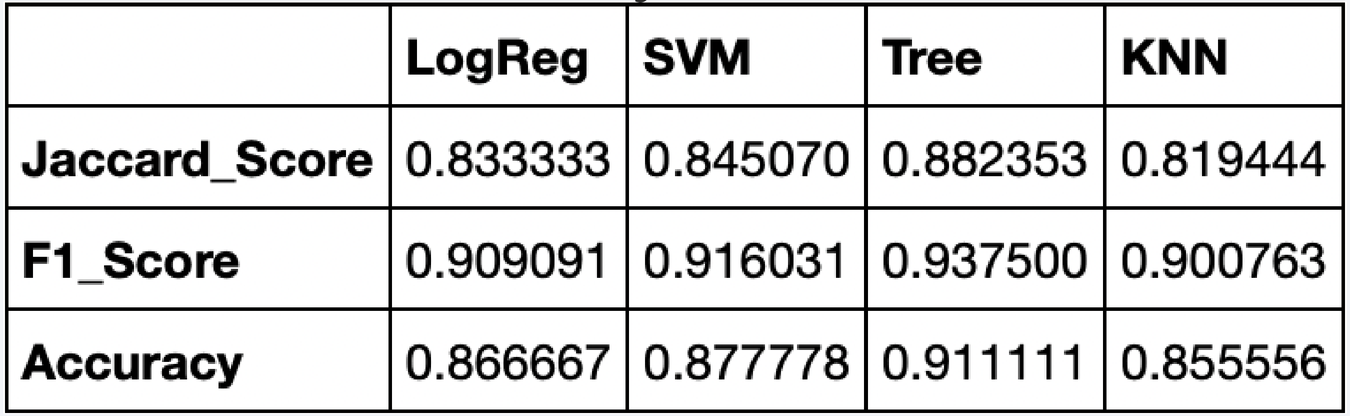
- • Dataset evaluation: Evaluating the scores of the whole Dataset confirms that the Decision Tree Model is the best model. This model not only achieves higher scores but also demonstrates the highest accuracy compared to other methods.

Scores and Accuracy Test Set



	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

Scores and Accuracy Entire Data Set

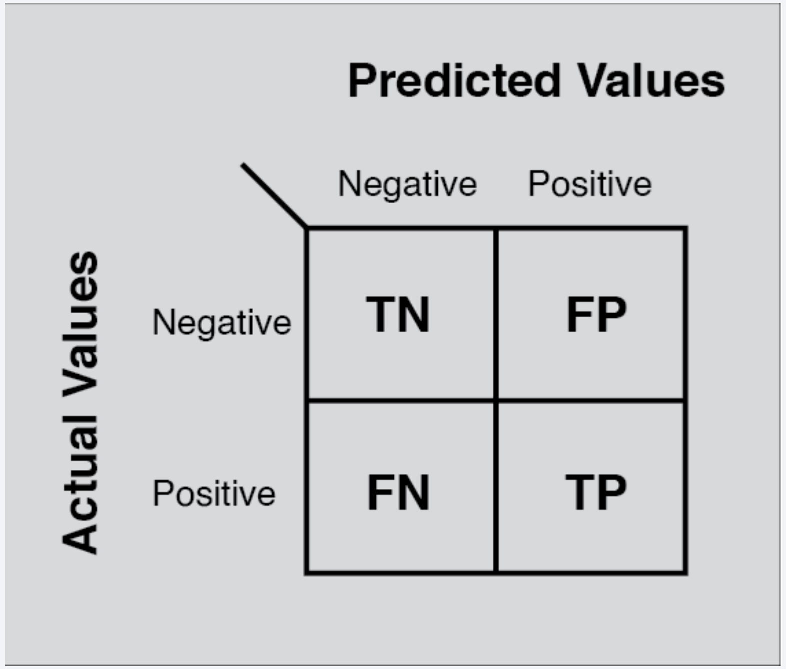


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation

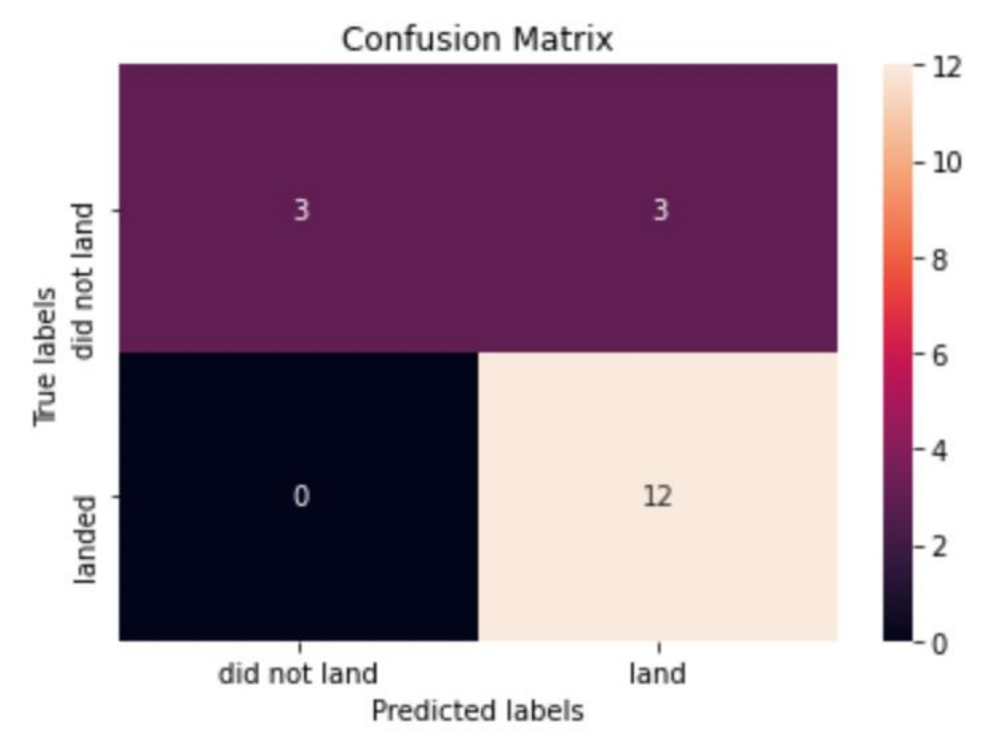
- Upon analyzing the confusion matrix, it becomes evident that logistic regression is capable of distinguishing between the different classes. However, a notable issue arises in the form of false positives, which constitutes the primary challenge in the classification process.



Predicted Values

Actual Values

	Negative	Positive
Negative	TN	FP
Positive	FN	TP



Conclusions

- The Decision Tree Model is the most suitable algorithm for this dataset, delivering the best performance.
- Launches with lower payload masses exhibit better results compared to launches with larger payload masses.
- Most launch sites are located near the Equator line, and all sites are in close proximity to the coast.
- The success rate of launches has shown a consistent increase over the years.
- KSC LC-39A stands out with the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!