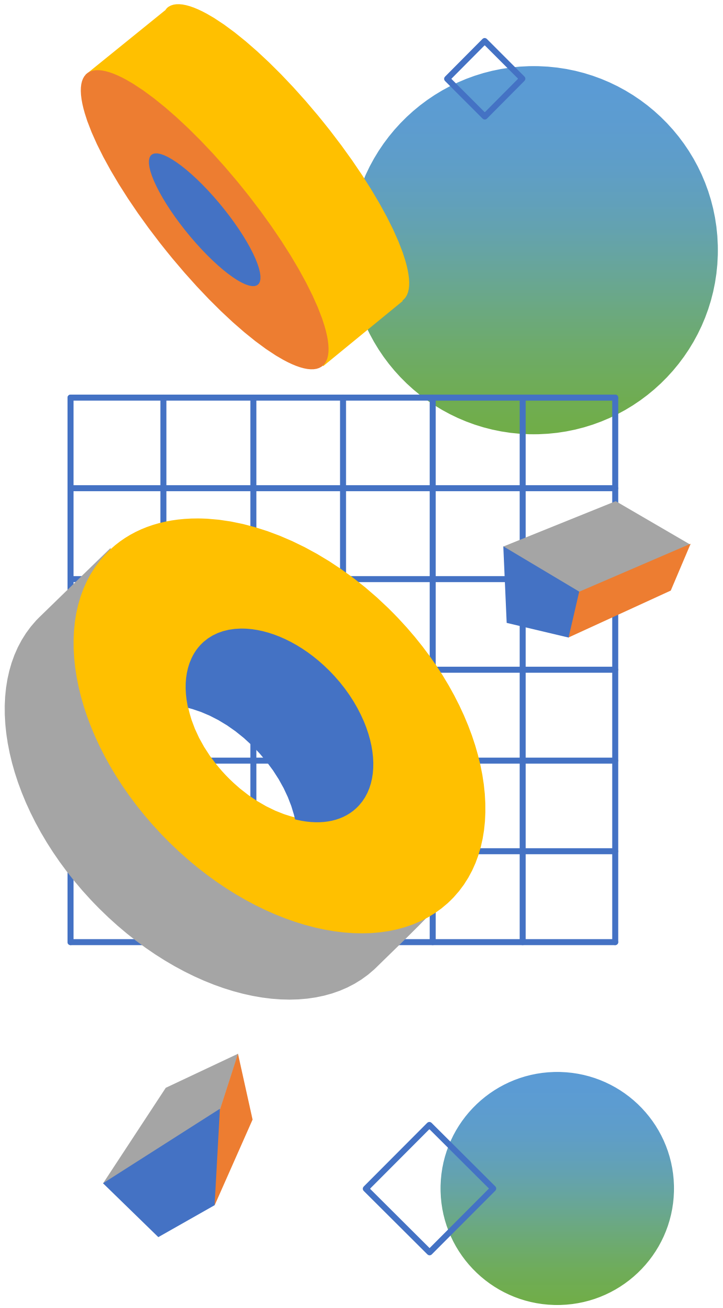
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MOHIT SINGH AIRY
8/8/2024



OUTLINE

01 **EXECUTIVE SUMMARY**

02 **INTRODUCTION**

03 **METHODOLOGY**

04 **RESULTS**

05 **CONCLUSION**

06 **APPENDIX**

01 - EXECUTIVE SUMMARY

Summary of methodologies-

❑ Data Collection
❑ Data Wrangling
❑ Exploratory Data analysis with SQL
❑ Exploratory Data analysis with Data Visualization
❑ Interactive Visual Analysis Using Folium
❑ Dashboarding with Plotly Dash
❑ Predictive Analysis using different Classification Models

Summary of all results-

- **Model Performance**: Decision Tree Model is the most effective for predicting landing success.
- **Launch Site Success**: KSC LC-39A has the highest success rate for Falcon 9 landings.
- **Trend Analysis**: Success rates have improved over time, particularly for lighter payloads.
- **Impact of Technology**: Advancements in technology and experience contribute to better landing outcomes.

02 - Introduction

- Project background and context

SpaceX, founded in 2002, has transformed the space industry with its reusable Falcon 9 rocket, cutting launch costs significantly. This innovation, particularly the ability to reuse the rocket's first stage, is a key factor in reducing space travel expenses.Predicting the success of first-stage landings is essential for maximizing cost savings and operational efficiency. This project analyzes SpaceX's launch data to determine how factors like payload mass, launch site, and orbit type affect landing success. By applying machine learning, we aim to improve prediction accuracy and support SpaceX's mission to make space travel more affordable.

- Problems you want to find answers

➢ How do variables such as payload mass, launch site, number of flights, and orbit type affect the success of first-stage landings?

➢ Does the rate of successful landings increase over the years?

➢ What is the best classification algorithm for predicting the success of first-stage landings?

➢ Which launch site has the highest success rate for first-stage landings?

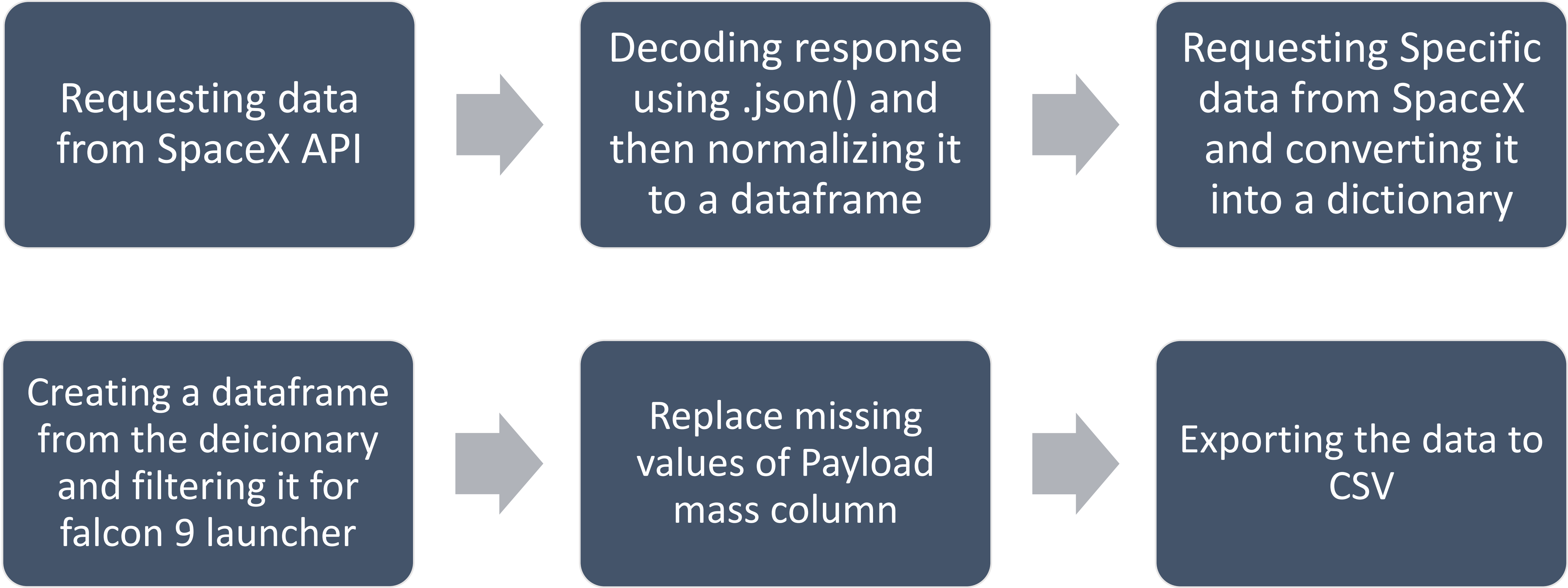
Section 1

Methodology

Executive Summary

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

Data Collection



Requesting data from SpaceX API

Decoding response using .json() and then normalizing it to a dataframe

Requesting Specific data from SpaceX and converting it into a dictionary

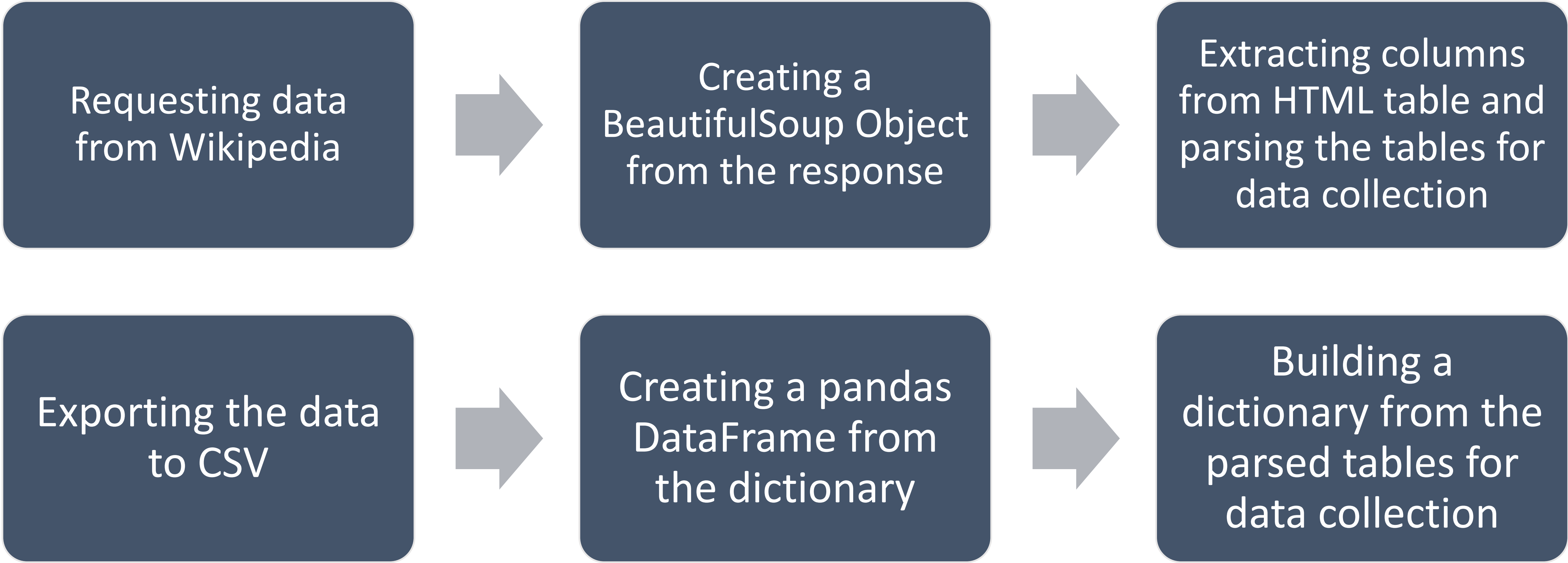
Creating a dataframe from the deicionary and filtering it for falcon 9 launcher

Replace missing values of Payload mass column

Exporting the data to CSV

LINK : SpaceX data collection API

Data Collection - Scraping



Requesting data from Wikipedia

Creating a BeautifulSoup Object from the response

Extracting columns from HTML table and parsing the tables for data collection

Building a dictionary from the parsed tables for data collection

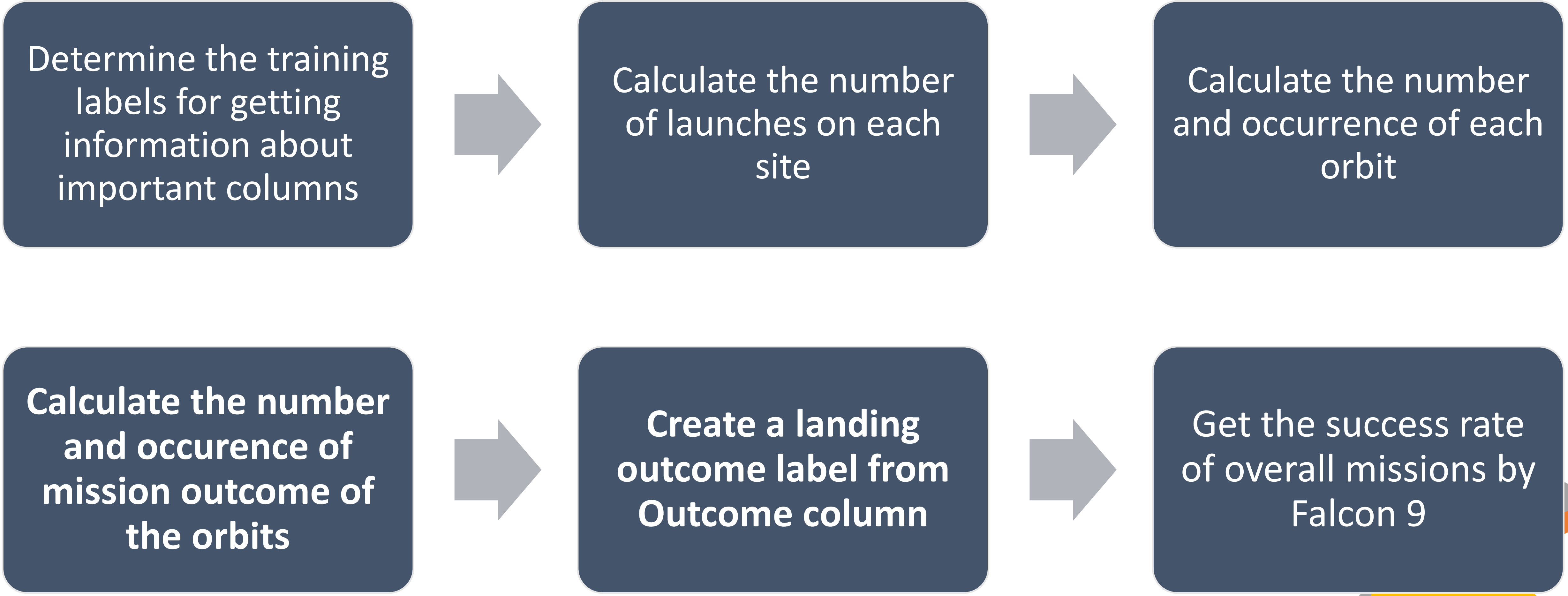
Creating a pandas DataFrame from the dictionary

Exporting the data to CSV

LINK : SpaceX Data Web Scraping

Data Wrangling

- **Cleaning:** Missing values were handled by replacing them with calculated means where applicable. Outliers and inconsistencies were identified and addressed to ensure data accuracy.
- **Encoding:** Categorical variables, such as launch site and orbit type, were converted into numerical formats using one-hot encoding to facilitate binary classification.
- **Label Creation:** The launch outcomes were transformed into binary labels for classification—“1” for successful landings and “0” for unsuccessful ones.



Determine the training labels for getting information about important columns

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Get the success rate of overall missions by Falcon 9

LINK : [SpaceX Data Wrangling]()

EDA with Data Visualization

Charts Plotted during Data Visualization

- Flight Number Vs Launch Site Catplot : *To get an overview of number of flights on each launchsite while observing the ratio of successful missions and failed missions on each site.*
- Payload Mass Vs Launch Site Catplot: *To know the capability of each site with successful missions with different payloads*
- Orbit Type Vs Success Rate Barplot: *To Visualize the relationship between success rate of each orbit type.*
- Flight number Vs Orbit type scatter plot: *To visualize which orbit type is most and least common on which these missions success rate is affected*
- Payload Mass Vs Orbit Type scatter plot: *To visualize how payload mass affects success rate on each Orbit Type*
- Yearly Launches and Success rate LinePlot: *To know how the success rate changes over the years.*

LINK : EDA with Data Visualization

EDA with SQL

SQL Queries Performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

LINK : EDA with SQL

Build an Interactive Map with Folium

1. **Markers:**
   1. **Purpose:** Markers were added to represent the locations of various launch sites.
   2. **Details:** Each marker was labeled with the name of the launch site and included a popup with additional details. This helps users visually locate and identify launch sites on the map.
2. **Circles:**
   1. **Purpose:** Circles were used to highlight the geographic area around each launch site.
   2. **Details:** The circles, with varying sizes and colors, indicate the proximity of launch sites to key landmarks such as the equator and coastlines. This visualization helps assess the strategic location of each site.
3. **Colored Markers for Launch Outcomes:**
   1. **Purpose:** To distinguish between successful and failed launches visually.
   2. **Details:** Green markers indicated successful launches, while red markers represented failed launches. This color-coding allows users to quickly identify patterns in launch success rates.
4. **Lines:**
   1. **Purpose:** To show distances between a specific launch site (e.g., KSC LC-39A) and its nearby landmarks.
   2. **Details:** Lines were drawn to connect the launch site to nearby railways, highways, coastlines, and cities. This visualization helps understand the potential risk zones and logistical considerations for each site.

LINK ：FOLIUM INTERACTIVE MAP NOTEBOOK

Build a Dashboard with Plotly Dash

Plots/objects added in the map

1-Pie Chart:

•**Purpose:** Display the total count of successful versus failed launches for all sites or a specific site selected by the user.

•**Details:** This chart helps users quickly compare the success rates across different sites and visualize overall performance.

2-Slider for Payload Mass Range:

•**Purpose:** Allow users to filter the data based on payload mass.

•**Details:** The slider enables users to adjust the range of payload masses and observe how success rates vary with different payload sizes, facilitating a more granular analysis.

3-Scatter Chart:

•**Purpose:** Show the correlation between payload mass and success rates for different booster versions.

•**Details:** This chart provides insight into how varying payload masses impact landing success across different rocket versions, helping identify trends and patterns.

4-Dropdown List for Launch Sites:

•**Purpose:** Enable users to select and focus on specific launch sites.

•**Details:** By selecting a launch site from the dropdown, users can view site-specific data and metrics, such as success rates and payload mass distributions

LINK : Dash App python program

Predictive Analysis (Classification)

1. <u>Model Building:</u>

- **Initial Models:** Built several classification models including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- **Data Preparation:** Standardized data using StandardScaler and split into training and test sets using train_test_split.

2. <u>Model Evaluation:</u>

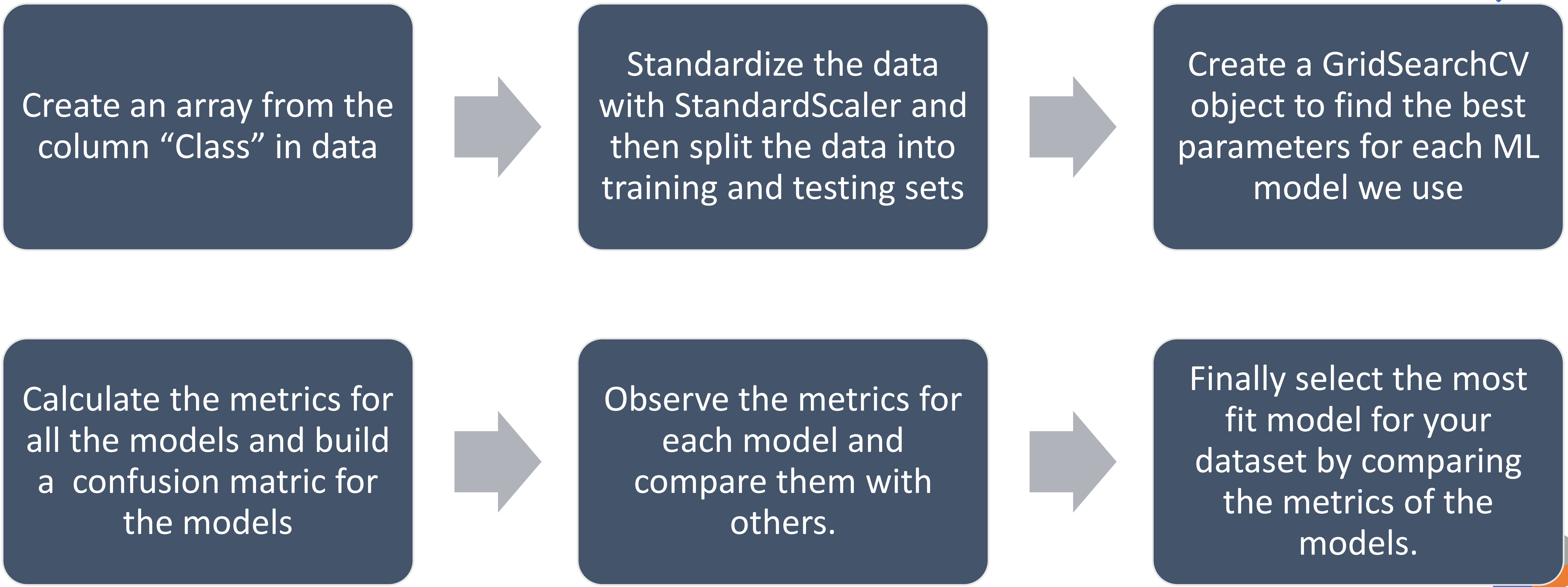
- **Metrics Used:** Evaluated models using metrics such as accuracy, Jaccard score, and F1 score.
- **Confusion Matrix:** Analyzed confusion matrices to understand false positives and negatives for each model.

3. <u>Model Improvement:</u>

- **Hyperparameter Tuning:** Used GridSearchCV with cross-validation to find the best parameters for each model.
- **Model Comparison:** Compared performance across models to identify strengths and weaknesses.

4. <u>Best Performing Model:</u>

- **Selection:** Based on overall performance metrics and accuracy scores, the Decision Tree model was identified as the best performer.
- **Final Evaluation:** Conducted final tests to confirm the model's accuracy and robustness on the entire dataset.



Create an array from the column “Class” in data

Standardize the data with StandardScaler and then split the data into training and testing sets

Create a GridSearchCV object to find the best parameters for each ML model we use

Calculate the metrics for all the models and build a confusion matric for the models

Observe the metrics for each model and compare them with others.

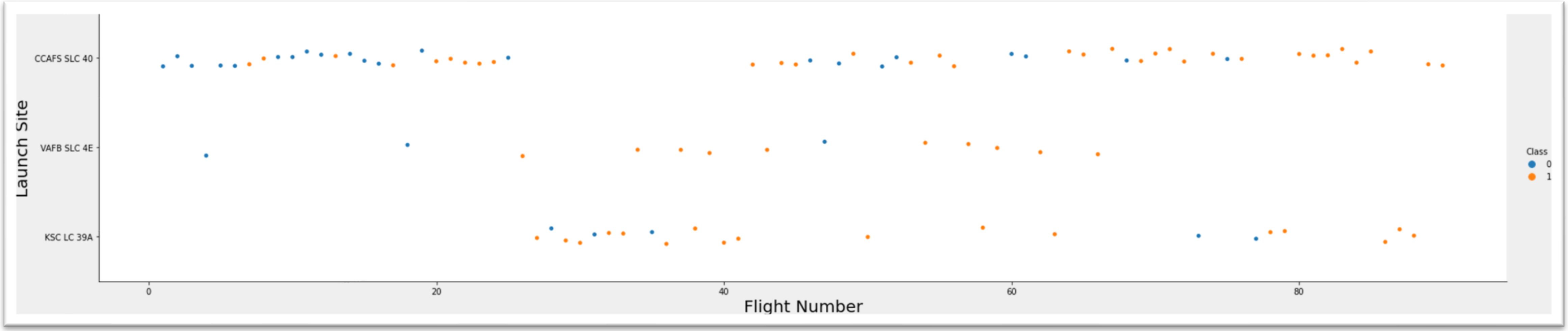
Finally select the most fit model for your dataset by comparing the metrics of the models.

LINK :Predictive Analysis Notebook

Section 2

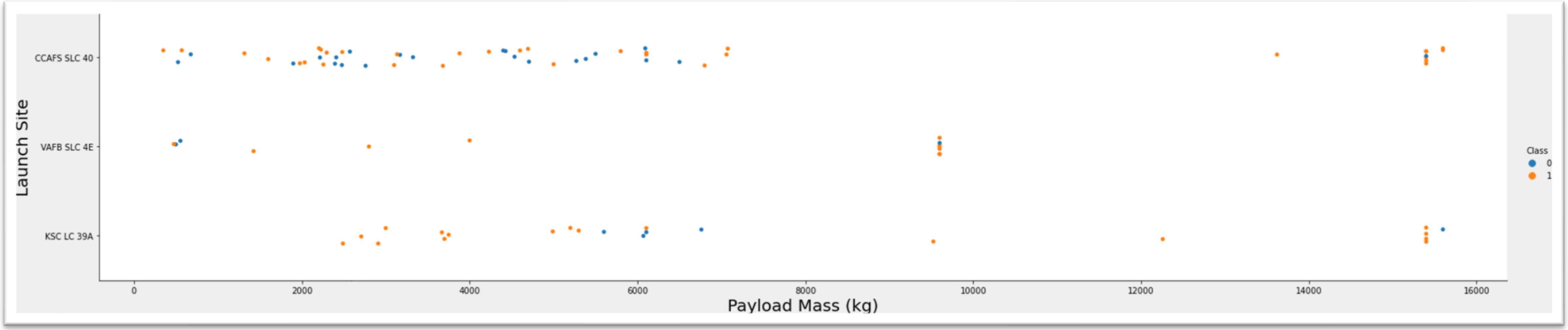
Insights drawn from EDA

Flight Number vs. Launch Site



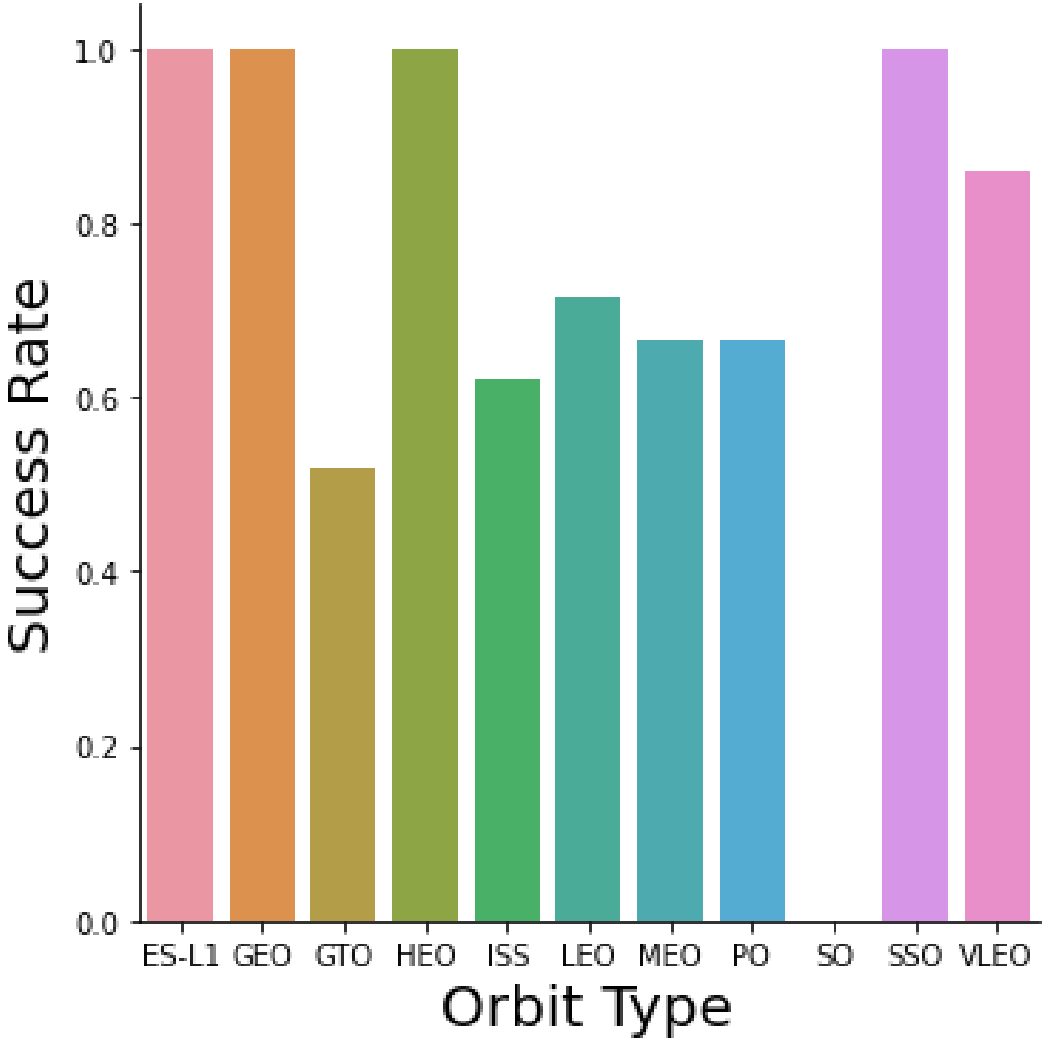
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



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

Success Rate vs. Orbit Type



•Orbits with 100% success rate are:

- ES-L1
- GEO
- HEO
- SSO

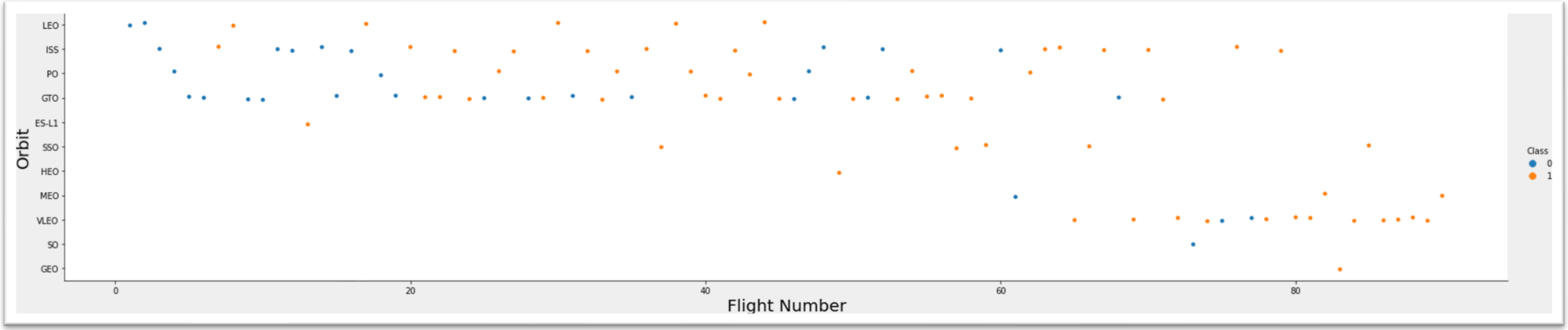
•Orbits with 0% success rate are:

- SO

•Orbits with success rate between 50% and 85%:

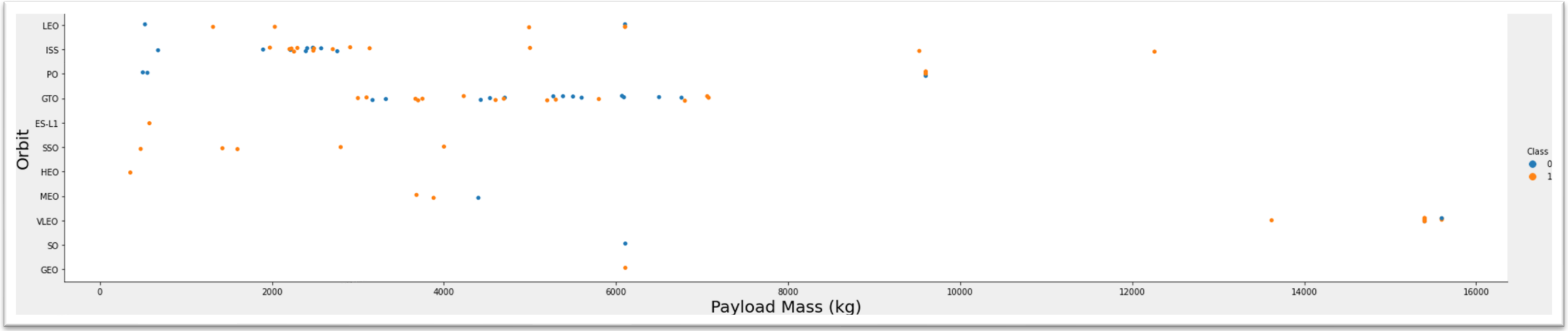
- GTO
- ISS
- LEO
- MEO
- PO

Flight Number vs. Orbit Type



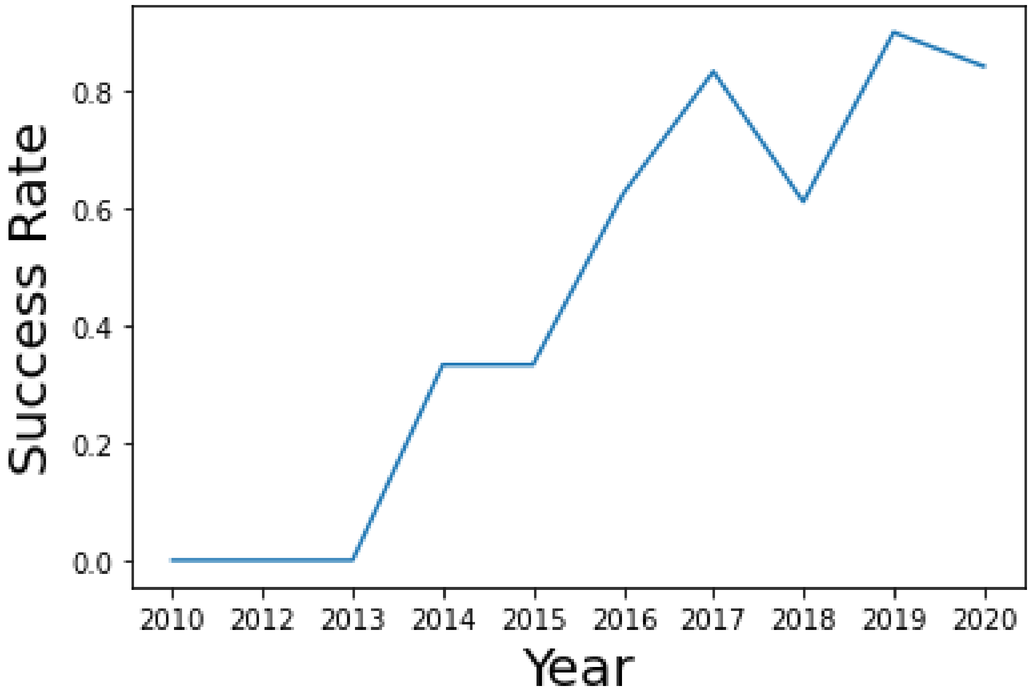
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



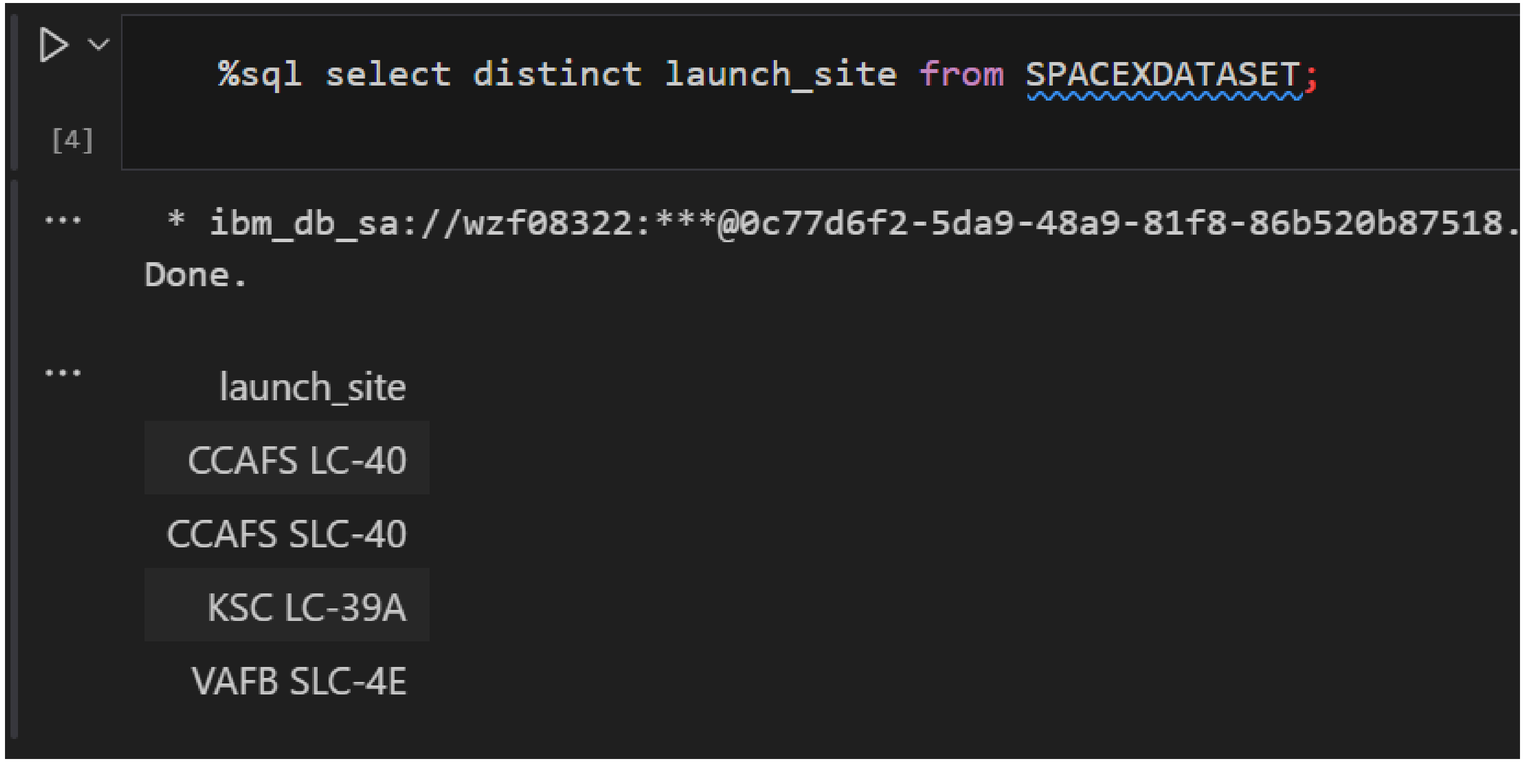
You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names



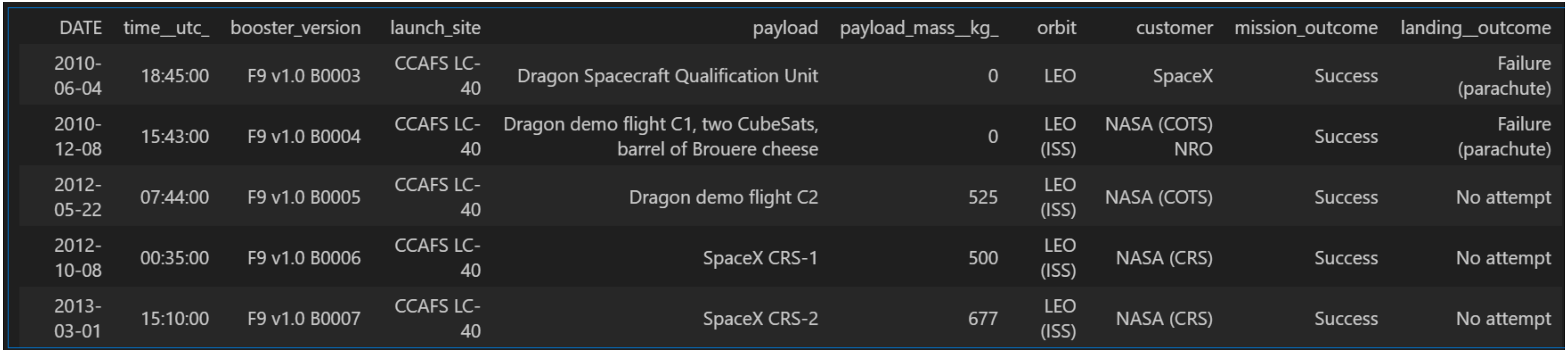
```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.
Done.
```

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

- Used the Distinct Function in my SQL Query to get the list of all Launch Site Names

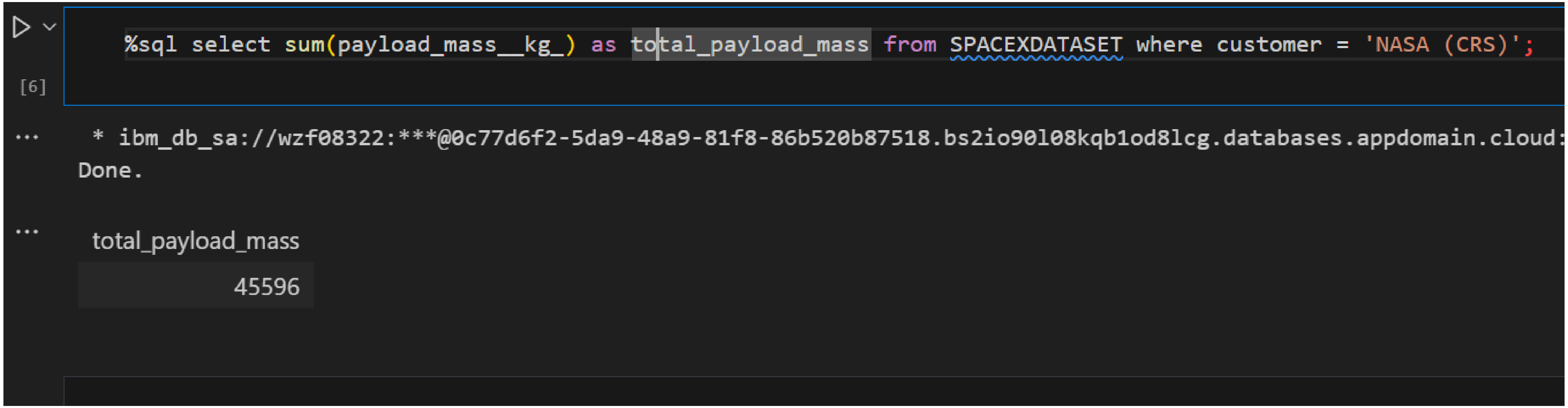
Launch Site Names Begin with 'CCA'



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

There are two different Launch Sites that start with ‘CCA’ which are CCAFS LC-40 and CCAFS SLC-40

Total Payload Mass



```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

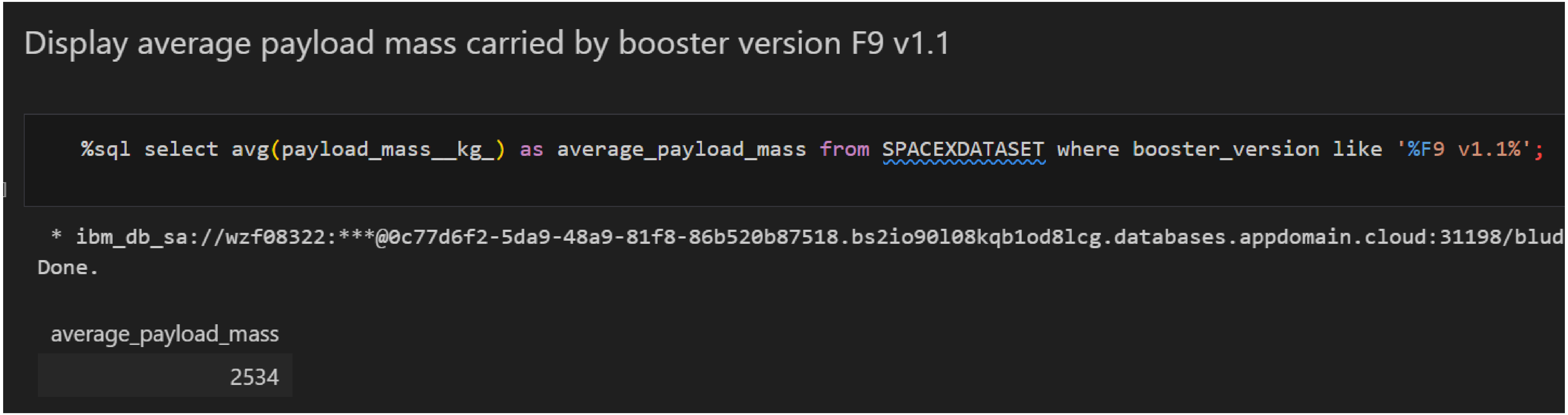
[6]

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

total_payload_mass
45596

The total Payload mass carried by boosters of NASA is 45596

Average Payload Mass by F9 v1.1



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

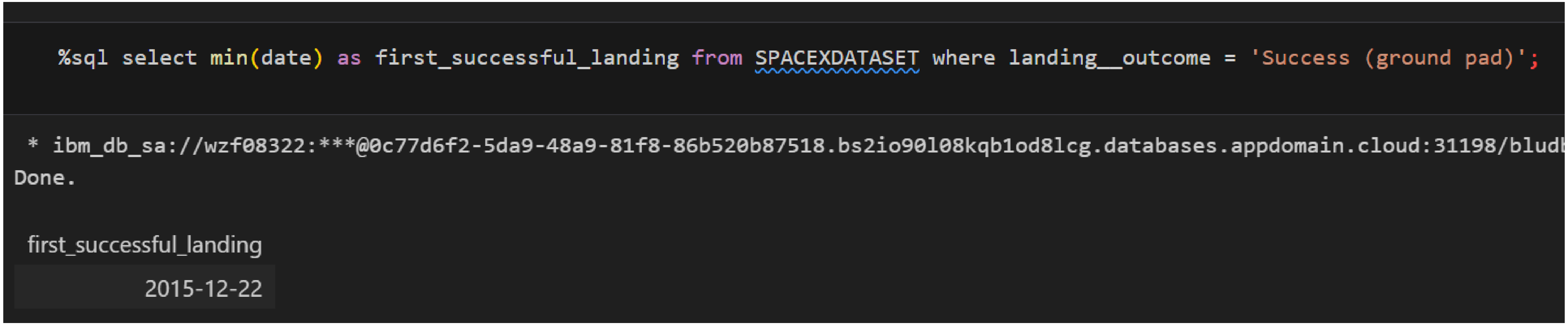
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

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

average_payload_mass
2534

The Average Payload Mass carried by F9 v1.1 is 2534

First Successful Ground Landing Date



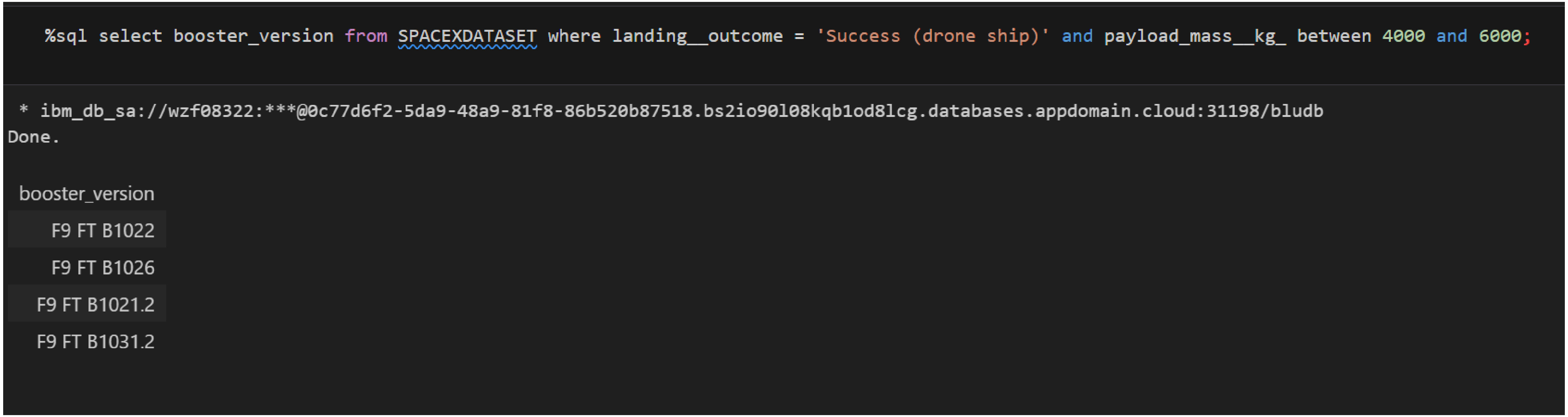
```
  %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

The First Successful Ground Landing was on 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



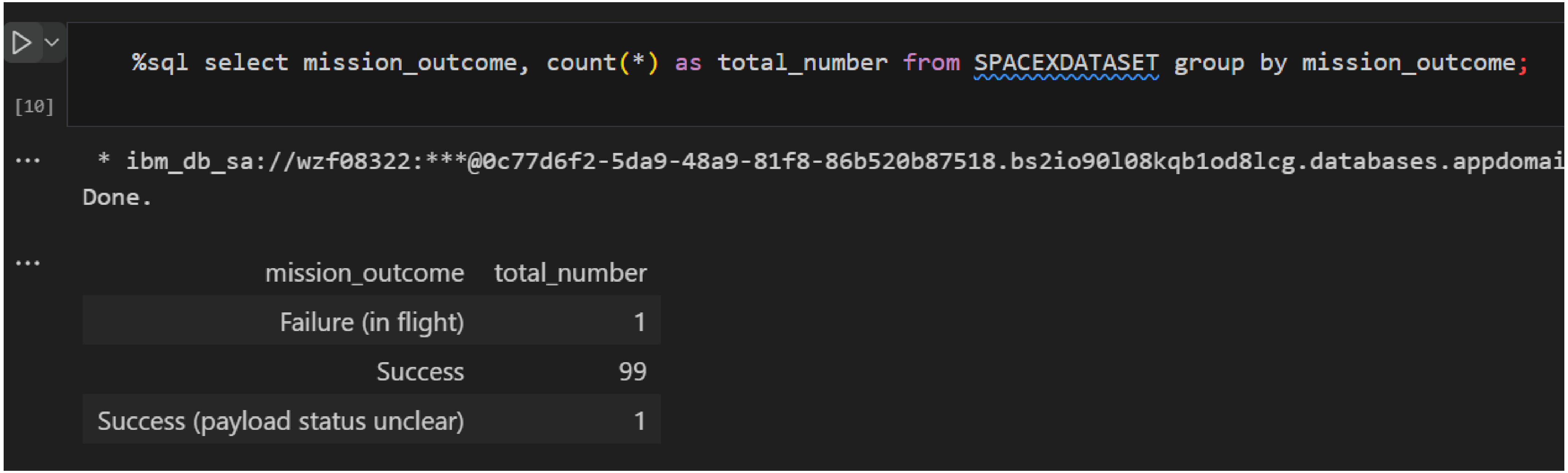
```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

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

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

Total Number of Successful and Failure Mission Outcomes



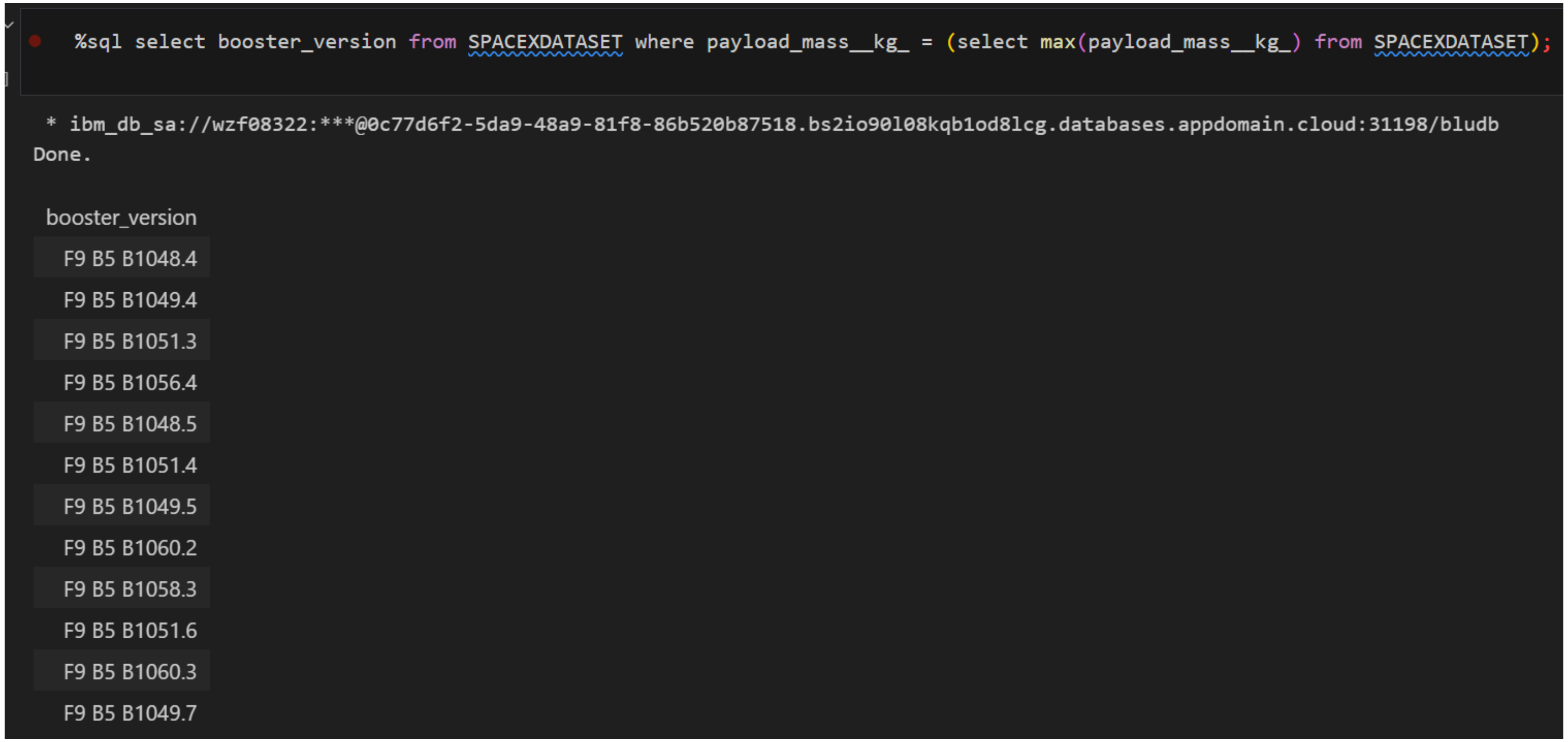
```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

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

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

Displaying the total number of successful and failure mission outcomes with 1 inflight failure, 99 Success and 1 success where payload status is unclear

Boosters Carried Maximum Payload



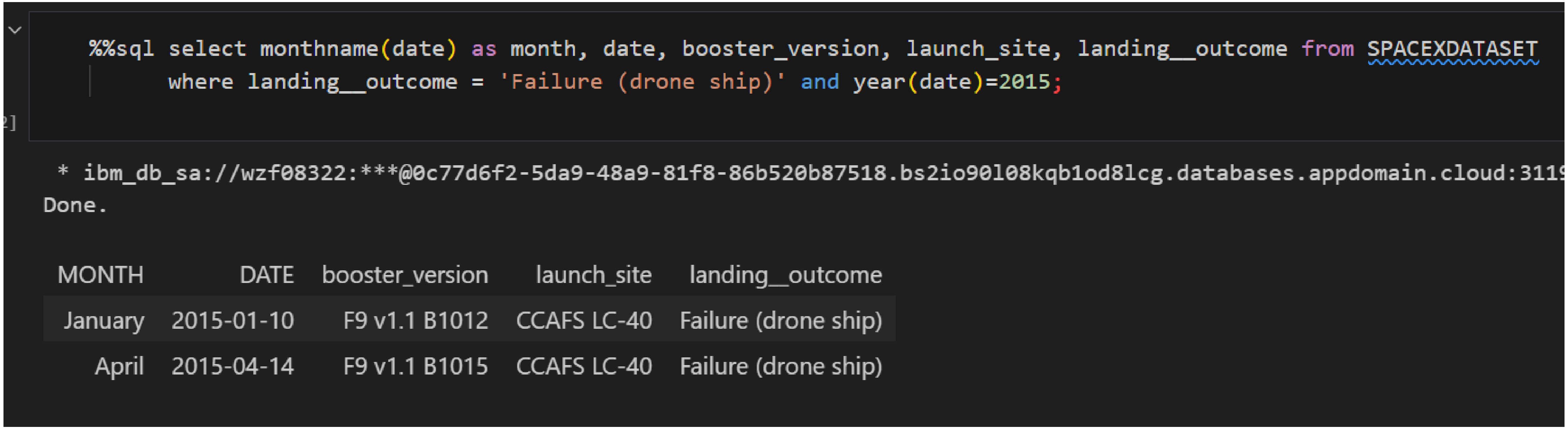
```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

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

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Displaying the names of the booster versions which have carried the maximum payload mass. Use a subquery

2015 Launch Records



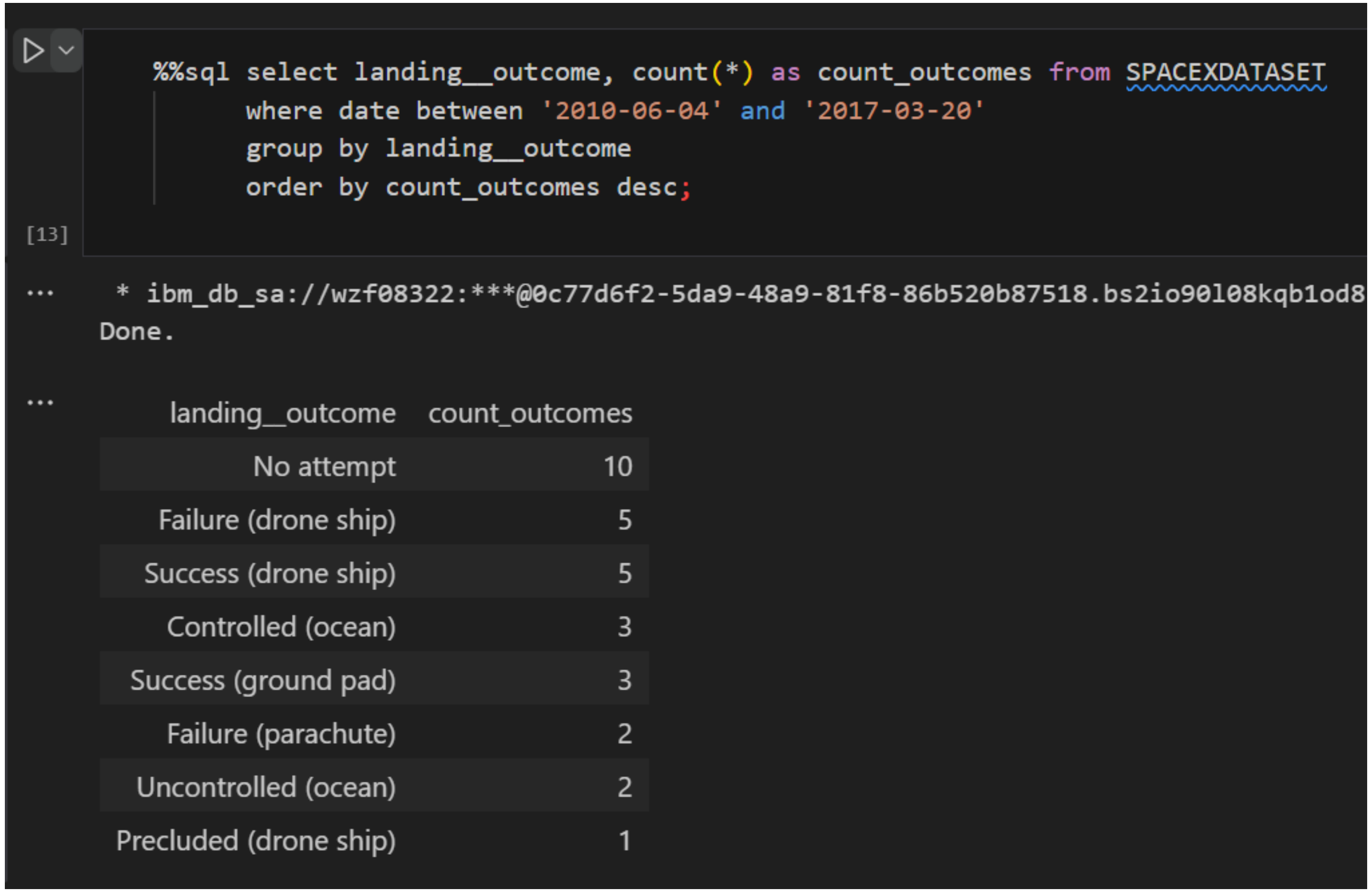
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
Done.
```

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)

Displaying the failed landing outcomes in drone ship, their booster versions, and launch site names for the in year 2015

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;
```

[13]

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8
Done.
```

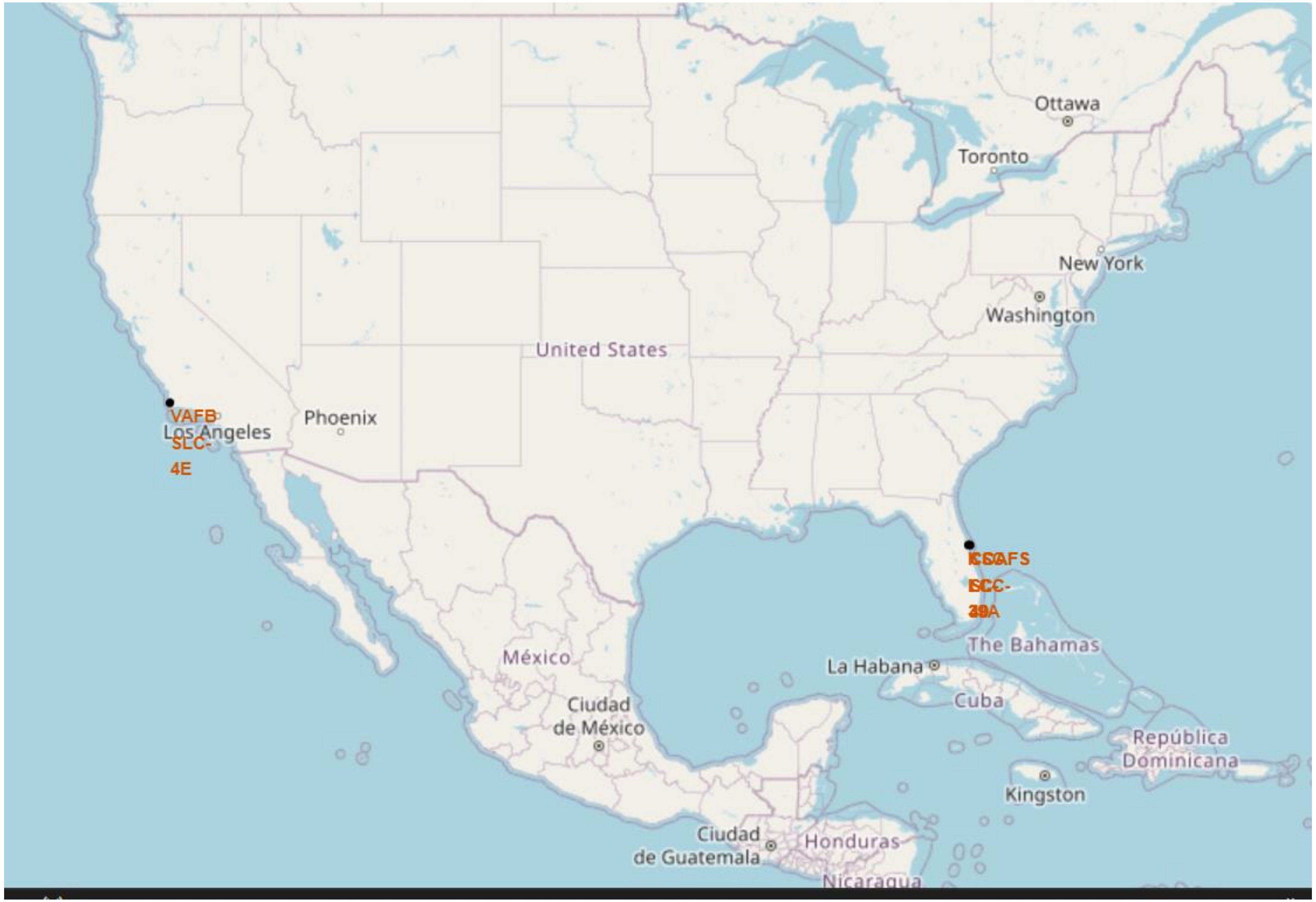
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Section 3

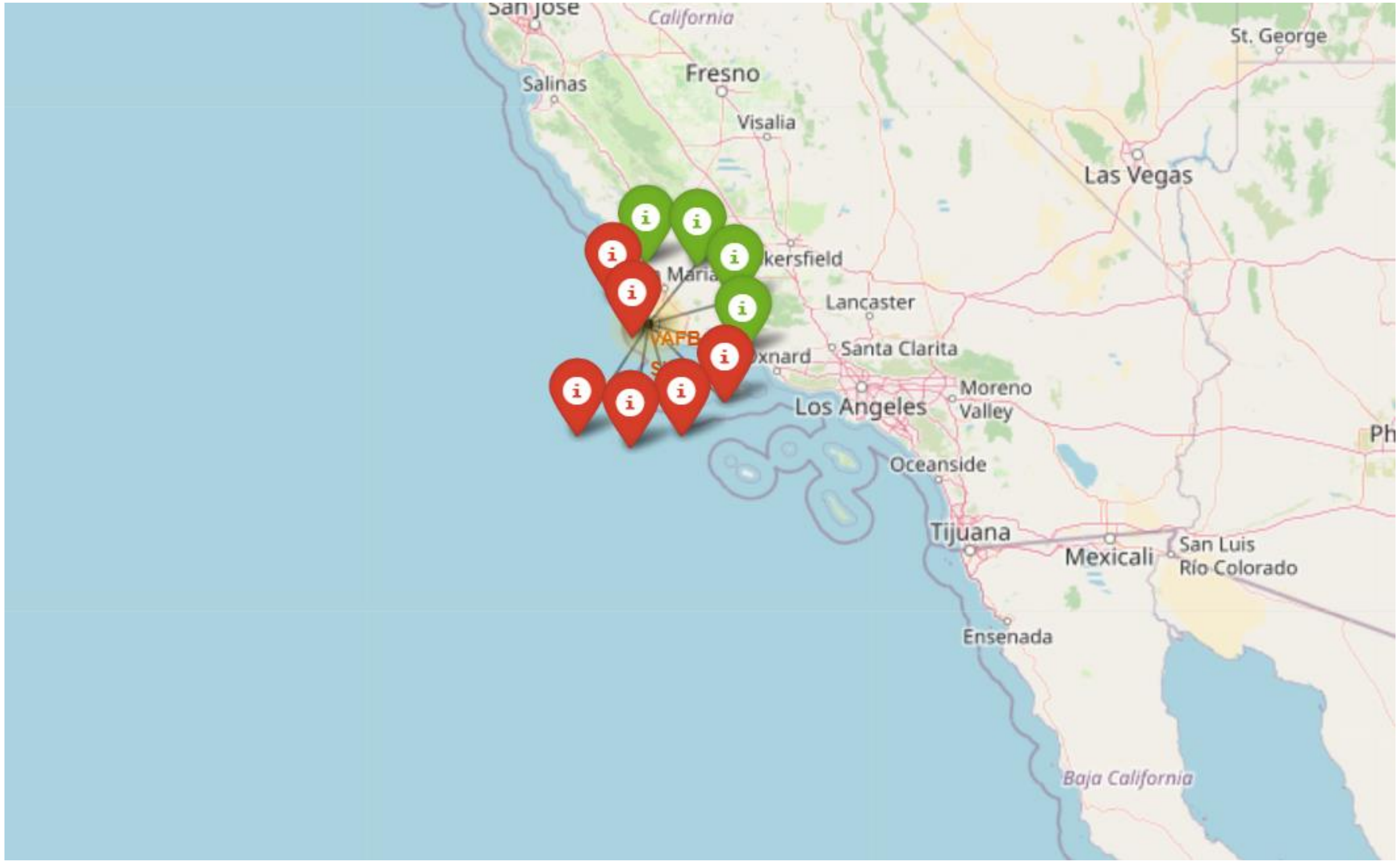
Launch Sites Proximities Analysis

Location of all Launch Sites



The Launch Sites are present in the East (Florida) and West Coast(California) of the United States of America

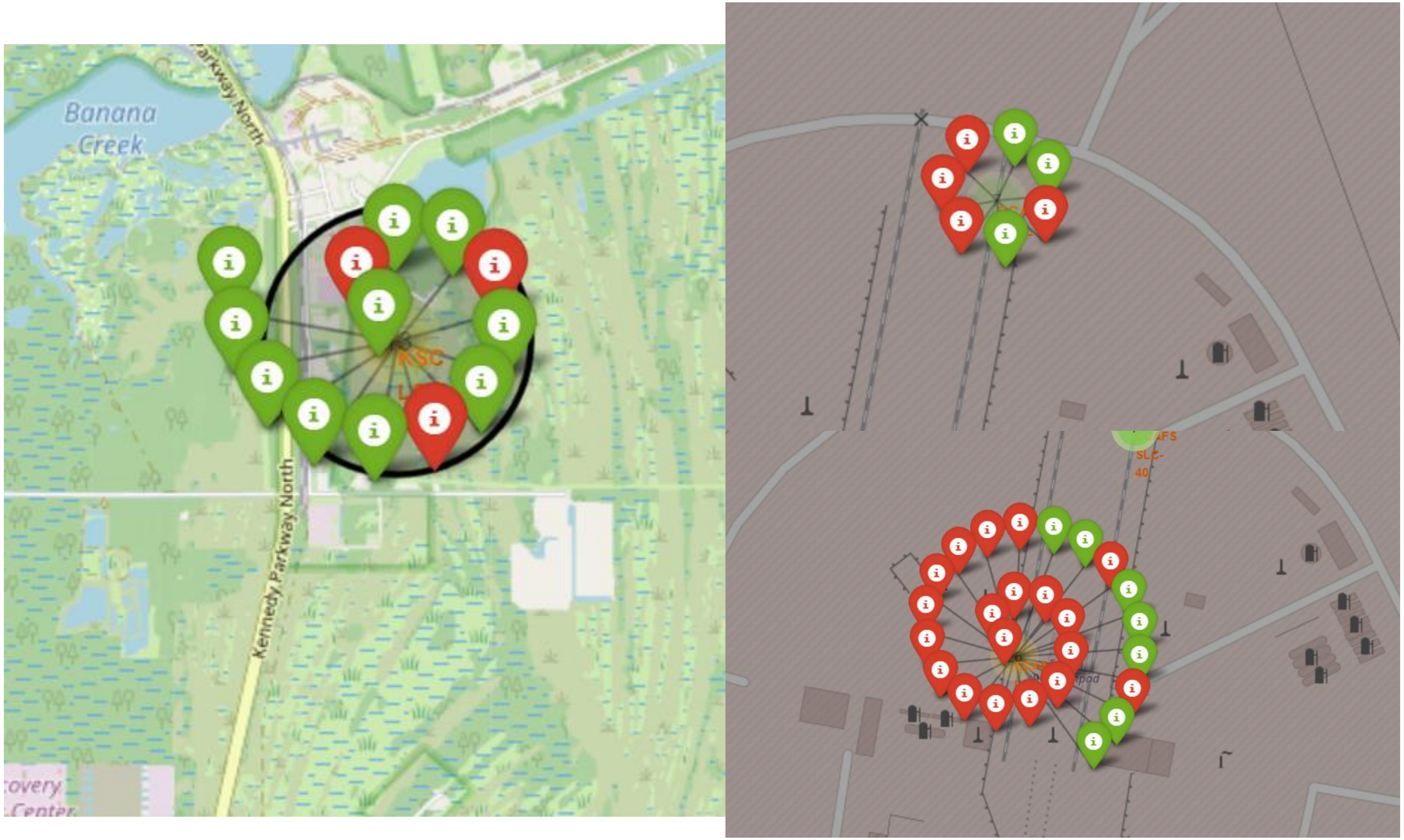
Color Labelled Launch Outcomes



West Coast Launch Outcomes

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

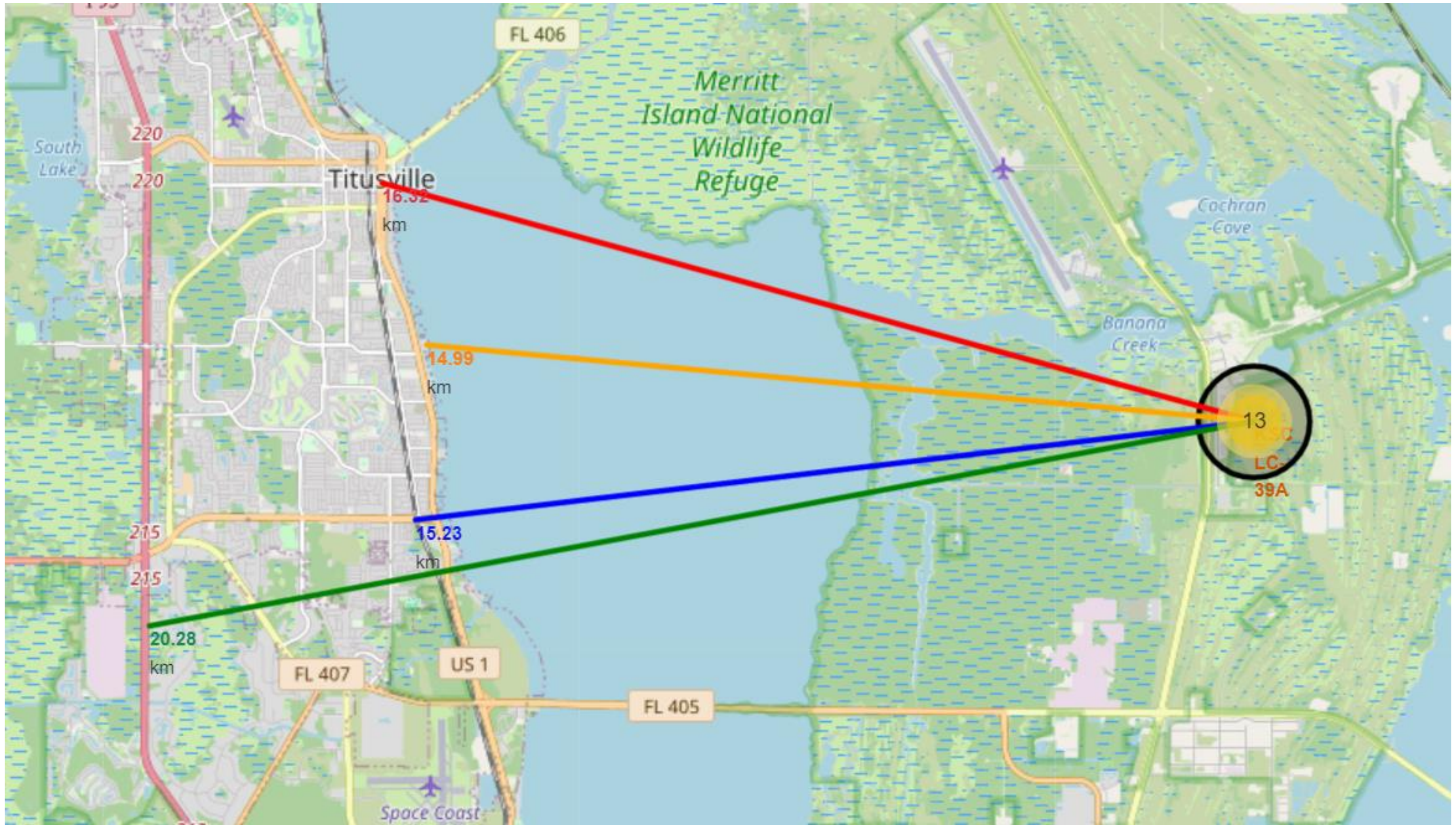
- Green Marker = Successful Launch
- Red Marker = Failed Launch



East Coast Launch Outcomes

- Launch Site KSC LC-39A has a very high Success Rate

Launch Site Distance to Landmarks

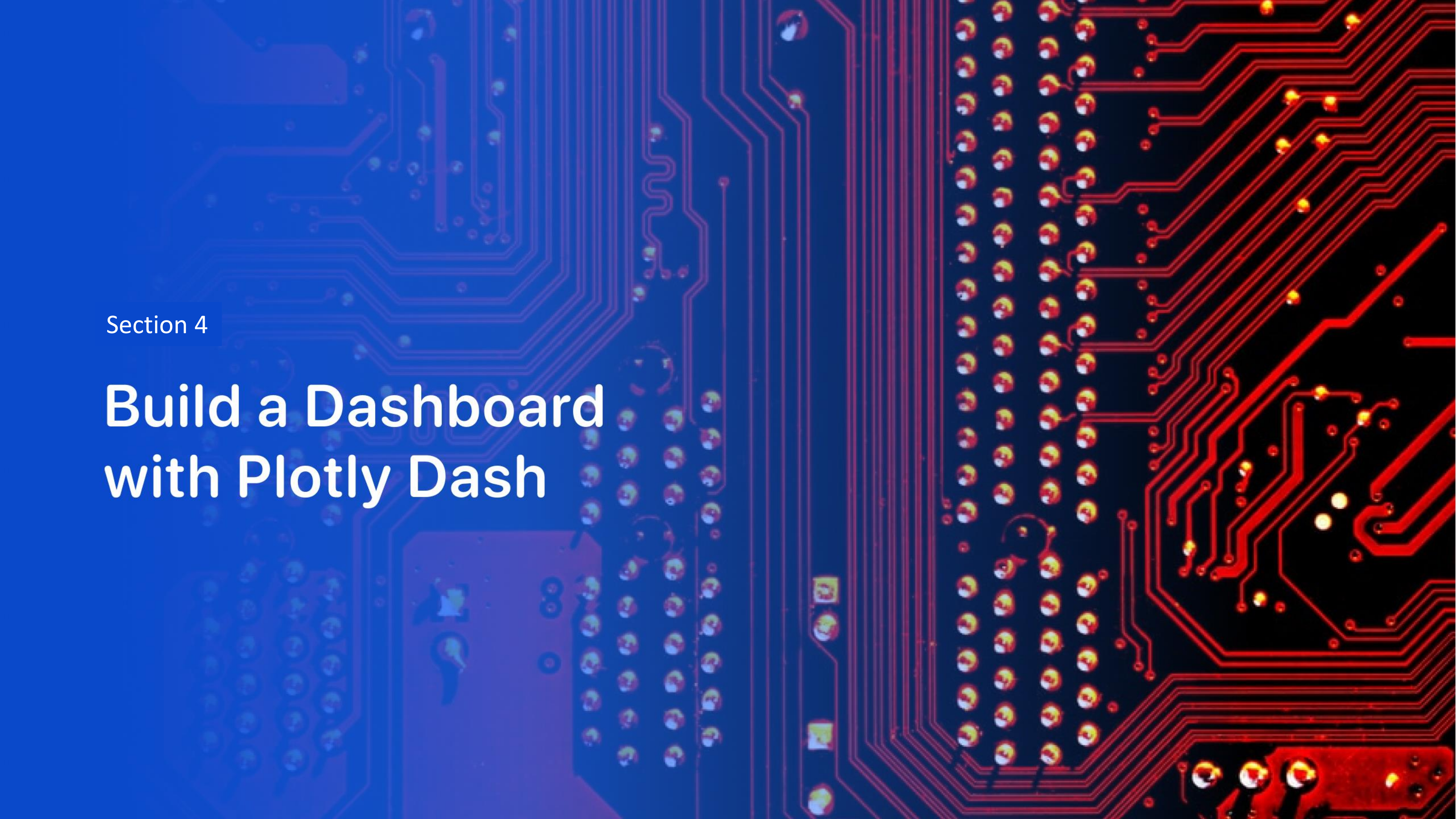


From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

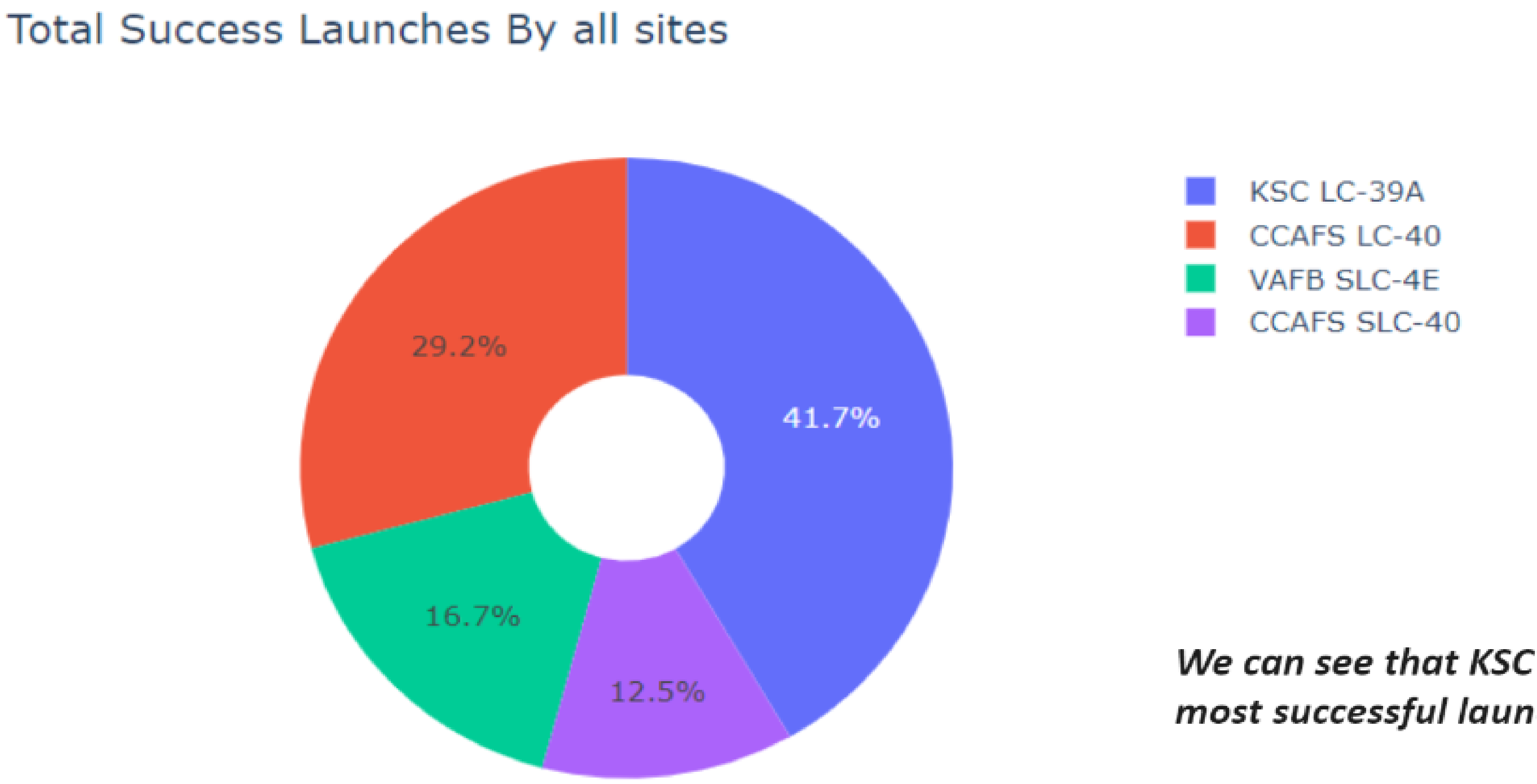
- 15.23 km from the railway
- 20.28 km from the highway
- 14.99 km from the coastline
- 16.32 km from the closest city which is potentially dangerous for the city.

Section 4

Build a Dashboard with Plotly Dash



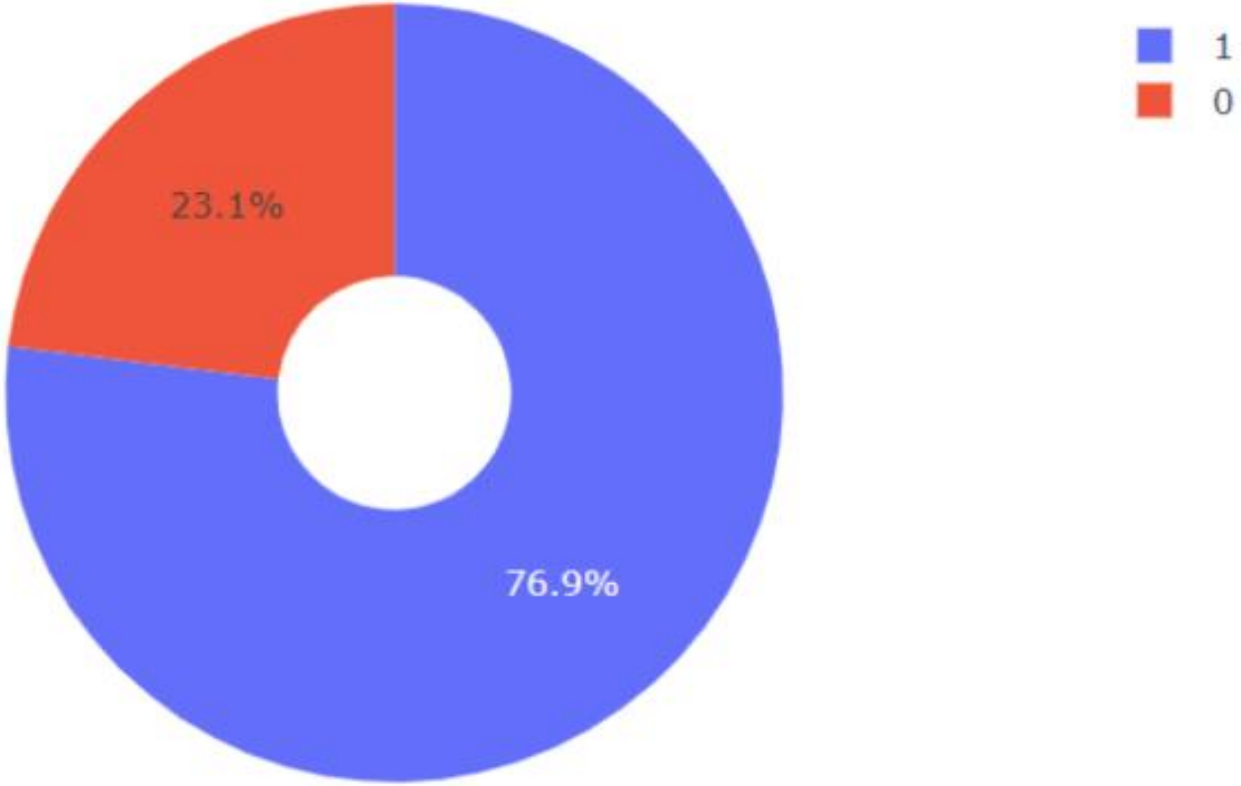
Launch Success count for each site



Total Success Launches By all sites

We can see that KSC LC-39A had the most successful launches from all the sites

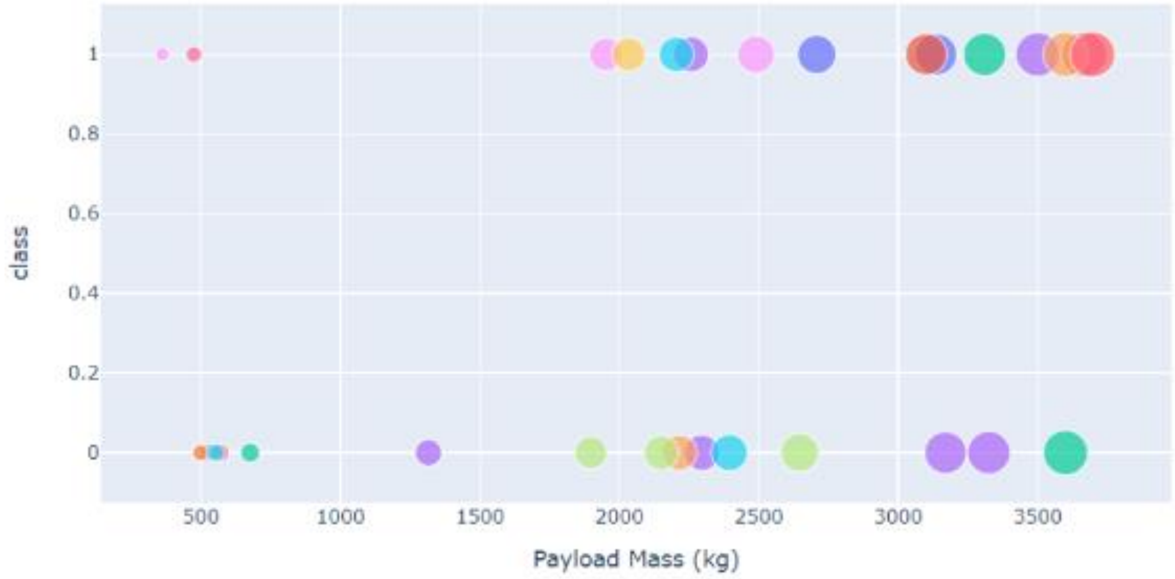
Launch Site with Highest Success Rate



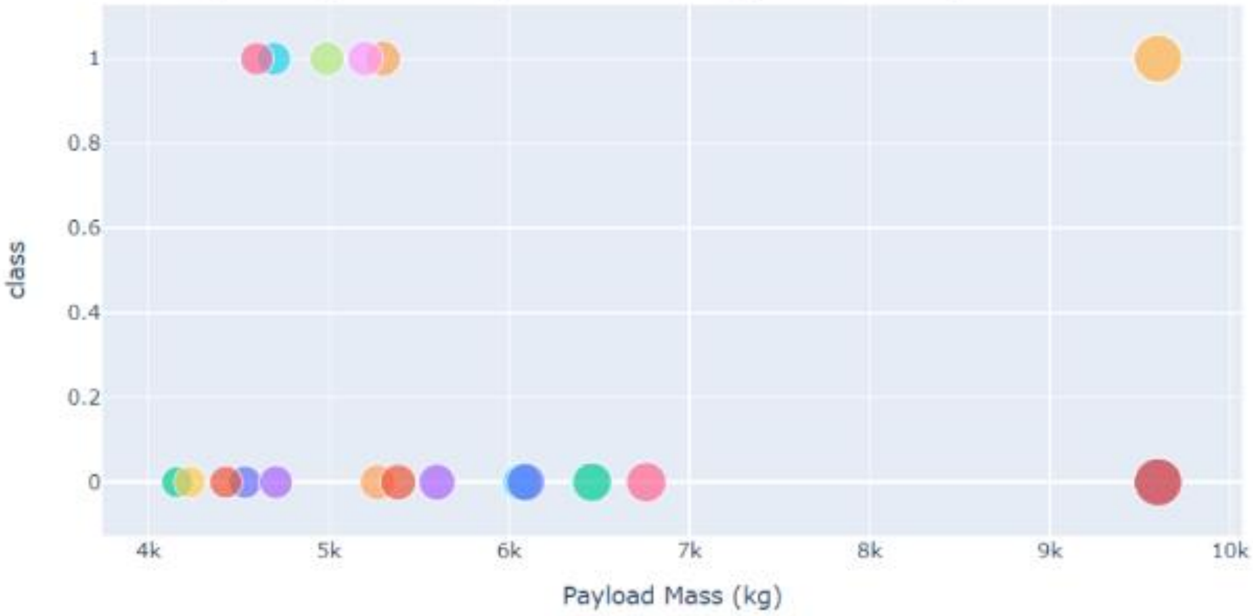
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Low Weighted Payload 0kg – 4000kg



Heavy Weighted Payload 4000kg – 10000kg



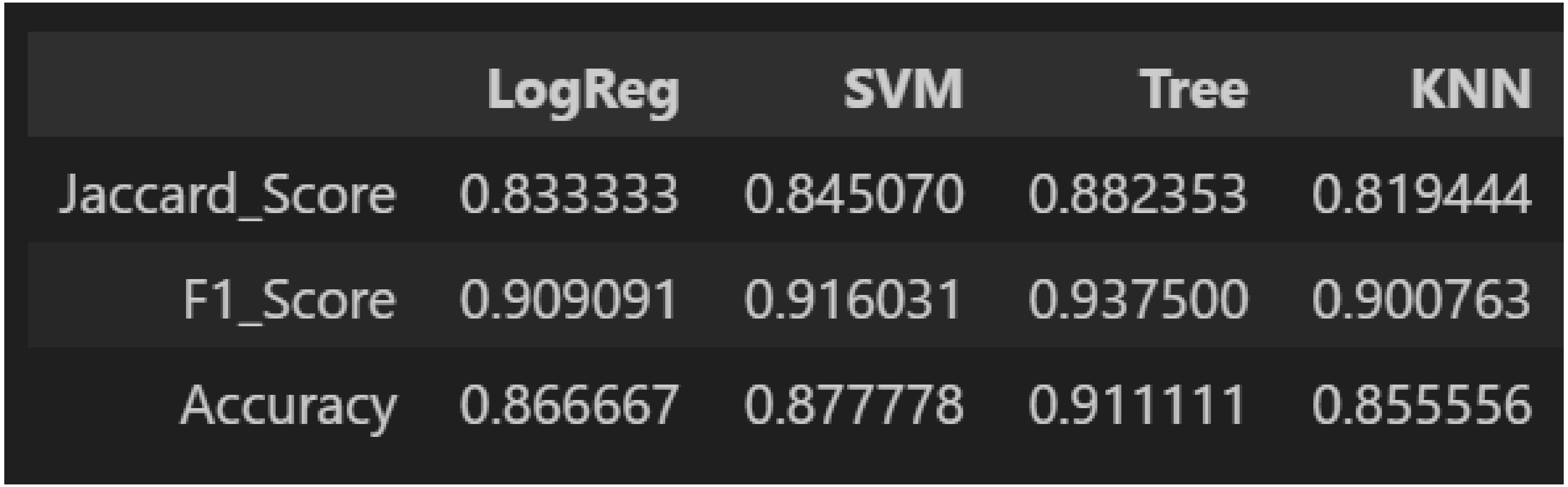
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

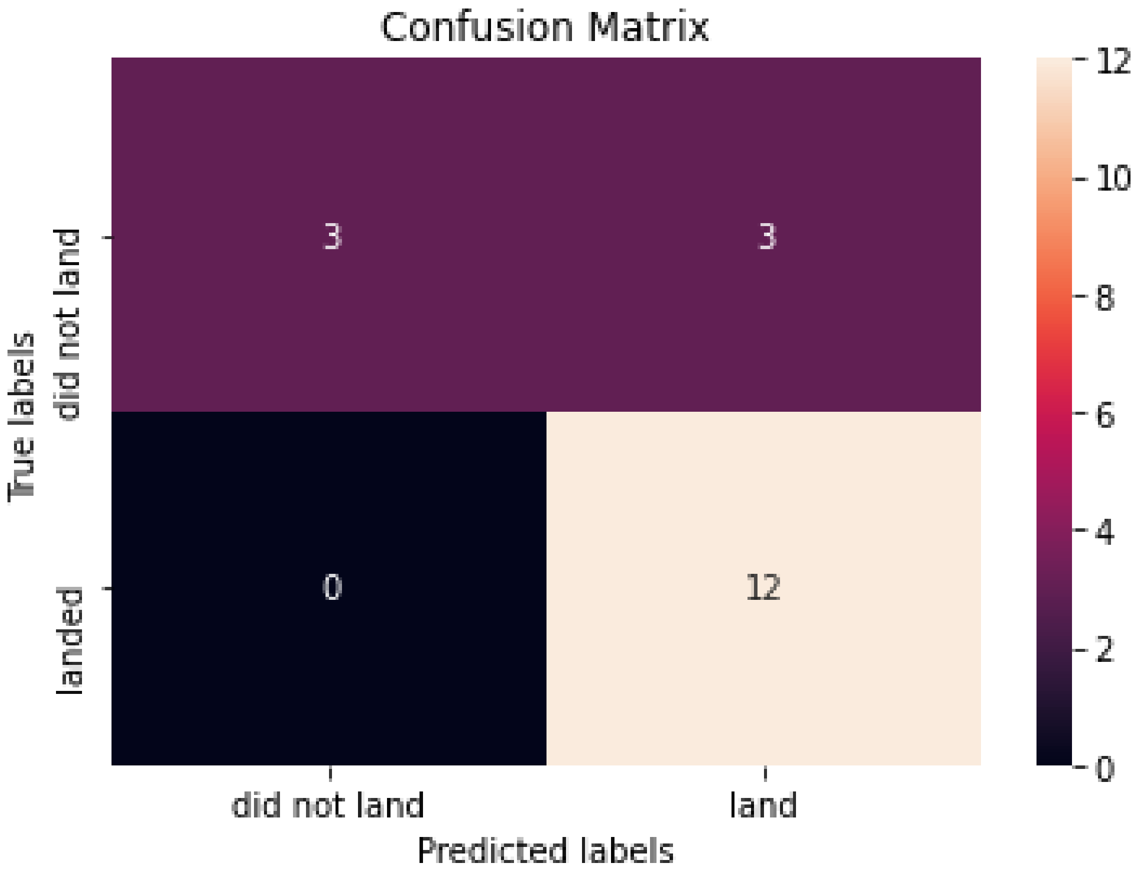
Metrics based on Entire Dataset predictions



	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

The decision tree classifier is the model with the highest classification accuracy

Confusion Matrix



- Examining the confusion matrix, we see that Decision Tree Classifier can distinguish between the different classes.
- We see that the major problem is false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions and Insights

- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- KSC LC-39A has a city in close proximity which may pose danger for the population.
- Launch success rate started to increase in 2013 till 2020.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Decision Tree Model is the best algorithm for this dataset as it has the highest accuracy and performs well for this SpaceX data.

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!