IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Qassam Sarmad
20-5-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analytics (EDA)
  - Predictive Analysis
- Summary of all results
  - EDA Results
  - Geospatial results
  - Dashboard
  - Predictive Analysis

Introduction

- SpaceX's Falcon 9 rockets cost significantly lesser to launch primarily because SpaceX can land the first stages of the rockets and reuse them
- Not all attempted landing of Falcon 9's first stage is a success
- A predictive model was made to determine if the landing of the rocket's first stages will be successful
- This can help other companies in their bids in using Falcon 9 for their launches
- It will also enable Space X in identifying the gaps in their launches that result in unsuccessful landing attempts

Section 1

Methodology

Executive Summary

- Data collection
  - Using GET commands on Space X APIs
  - Web scraping from Wikipedia
- Perform data wrangling
  - Using mean/average of the values to fill in missing data
  - Converting text data into numerical for effective prediction
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using multiple models and selecting the most effective model based on its accuracy

Data Collection

Using Space X REST APIs

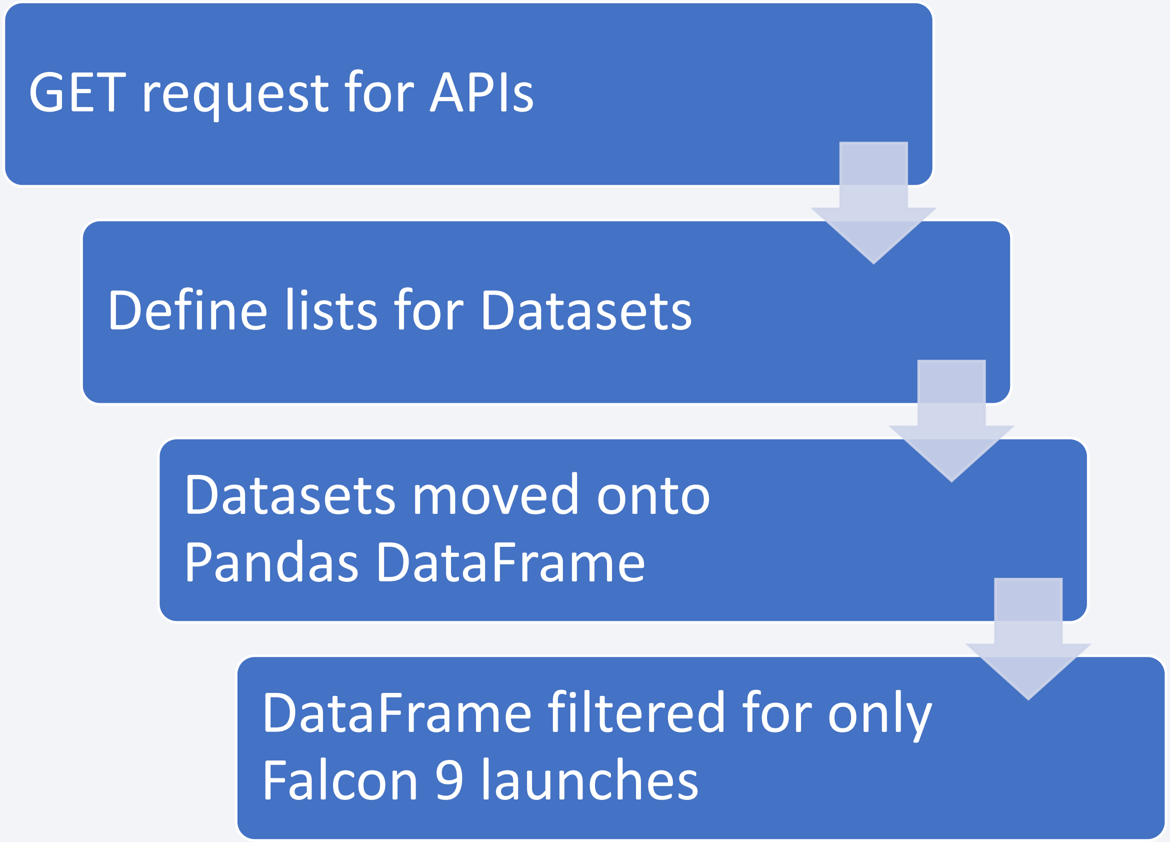
1. GET commands used to gather data
2. The responses received in .json were converted into dataframes using Pandas

Webscraping

1. BeautifulSoup object created from a static HTML website
2. The dictionary in the Soup was converted to Pandas DataFrame

Data Collection – SpaceX API

- GET command used to get response from REST API
- Response converted to .json file and then to Pandas DataFrame
- Lists defined for the data to be stored
- Lists used in dictionary to build the dataset
- Pandas DataFrame constructed from the dictionary
- DataFrame filtered to include only Falcon 9 launches
- Missing values of Payload Mass replaced with the mean value pf Payload Mass


GET request for APIs

Define lists for Datasets

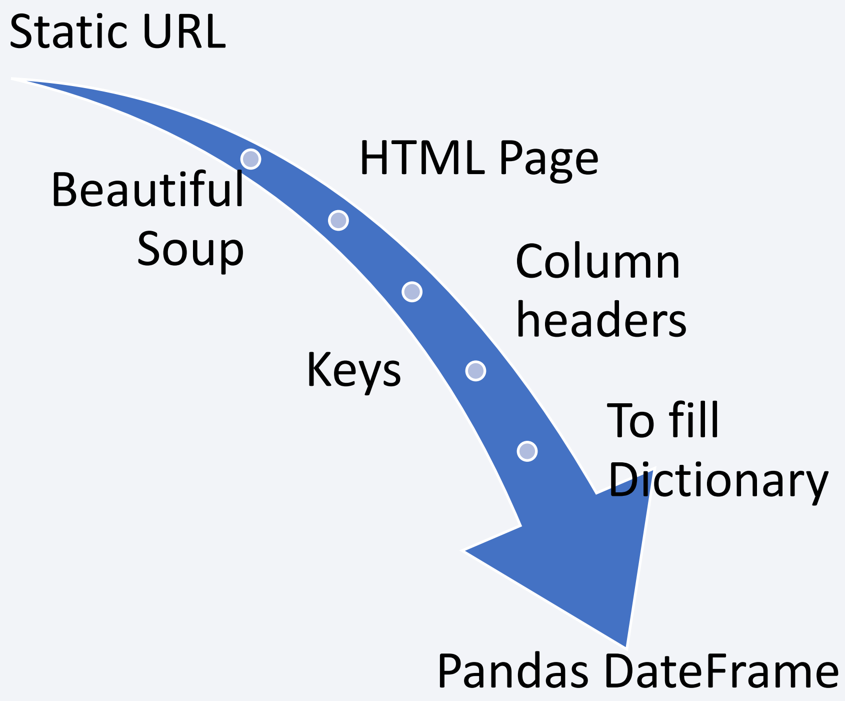
Datasets moved onto Pandas DataFrame

DataFrame filtered for only Falcon 9 launches

GitHub Link

Data Collection – Web Scraping

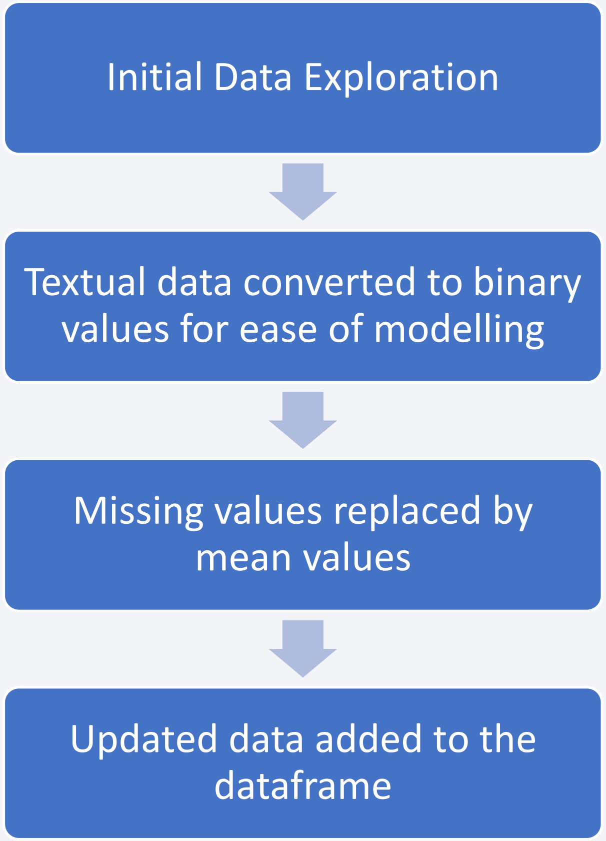
- Static URL extracted from a HTML page
- BeautifulSoup object created from the HTML response
- Column headers extracted from the table located on the page
- Column names were used as keys in a dictionary
- Functions and logic were used to fill the dictionary values
- Dictionary converted to Pandas Dataframe



GitHub Link

Data Wrangling

- Initially the data was explored to determine:
  - Number of Launches
  - Number of occurrence for each orbit
  - Landing outcomes
- Successful and unsuccessful landings were equated to binary values
- These binary values were added onto the data frame
- Missing values were also removed by adding mean values of those columns



Initial Data Exploration

Textual data converted to binary values for ease of modelling

Missing values replaced by mean values

Updated data added to the dataframe

GitHub Link

EDA with Data Visualization

- Scatter Charts are used to visual relation between 2 numberic variables. In this project they were used to:
  - Orbit type and Flight numbers
  - Flight number and launch site
  - Payload and orbit type
  - Payload and launch site
- Bar charts are useful in comparison between 2 or more categorical variables. In this project it was used to see:
  - Success rate and orbit type
- Line graphs useful in seeing the change in the relationship between 2 numerical values over time. In this project they were used to:
  - Success rate and year

GitHub Link

EDA with SQL

SQL queries were used on the dataset to gather the following information:

- Unique launch sites
- Launch sites with the string ‘CCA’
- Total payload mass carried by NASA’s boosters
- Average payload mass carried by Falcon 9 boosters
- Date of the first successful ground landing
- Names of boosters successful with a payload mass between 4000 and 6000 kg
- Total number of successful and unsuccessful missions
- Boosters that carried the max payloads
- Details of boosters which failed landing outcomes

GitHub Link

Interactive Geospatial Map on Folium

1. All launch sites marked on the map
2. Numbers of successful and unsuccessful launches for each site added on the map
3. Distance of each launch site to their proximities measured and marked

GitHub Link

Interactive Dashboard with Plotly Dash

1. Pie chart made showing the total number of successful charts per site
2. Scatter plot was made to show the correlation between the landing outcome and the payload mass

Predictive Analysis (Classification)

- Machine learning algorithms were first shortlisted
- The data was split into training and testing sets
- The object was fitted to the parameters and the model was trained for each algorithm
- Confusion matrices were plotted for each model during the evaluation step
- Accuracy scores were used to determine the best performing model

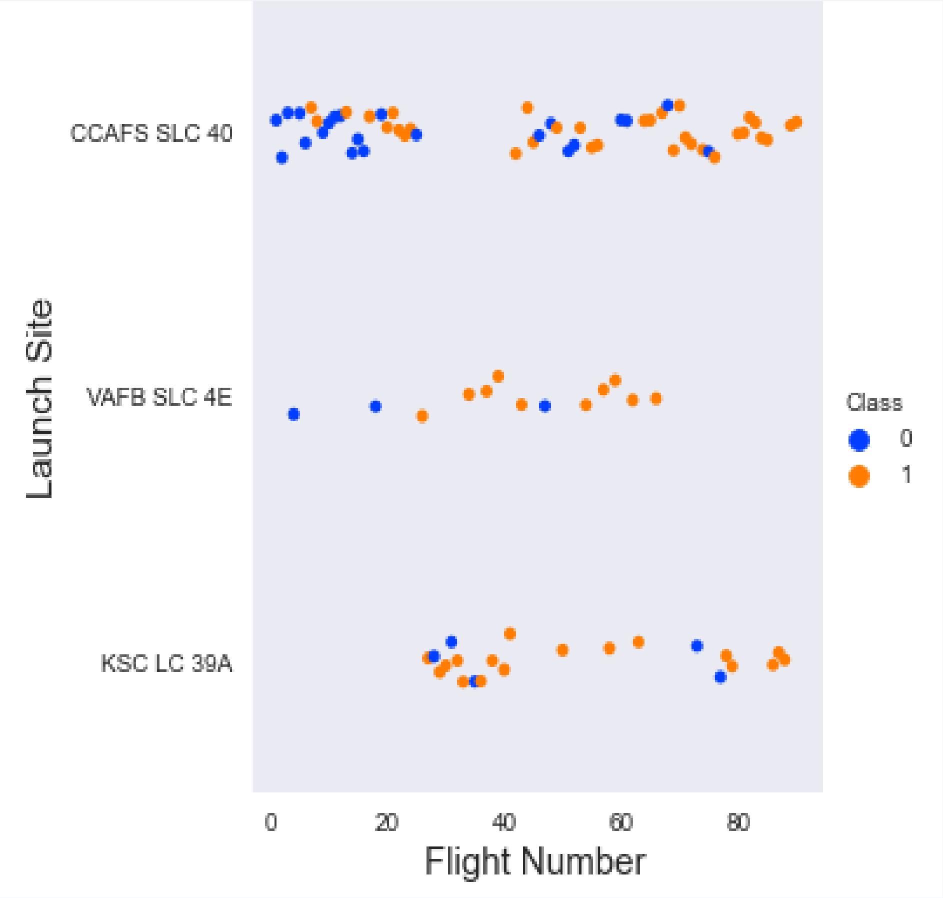
GitHub Link

Section 2

Insights drawn from EDA

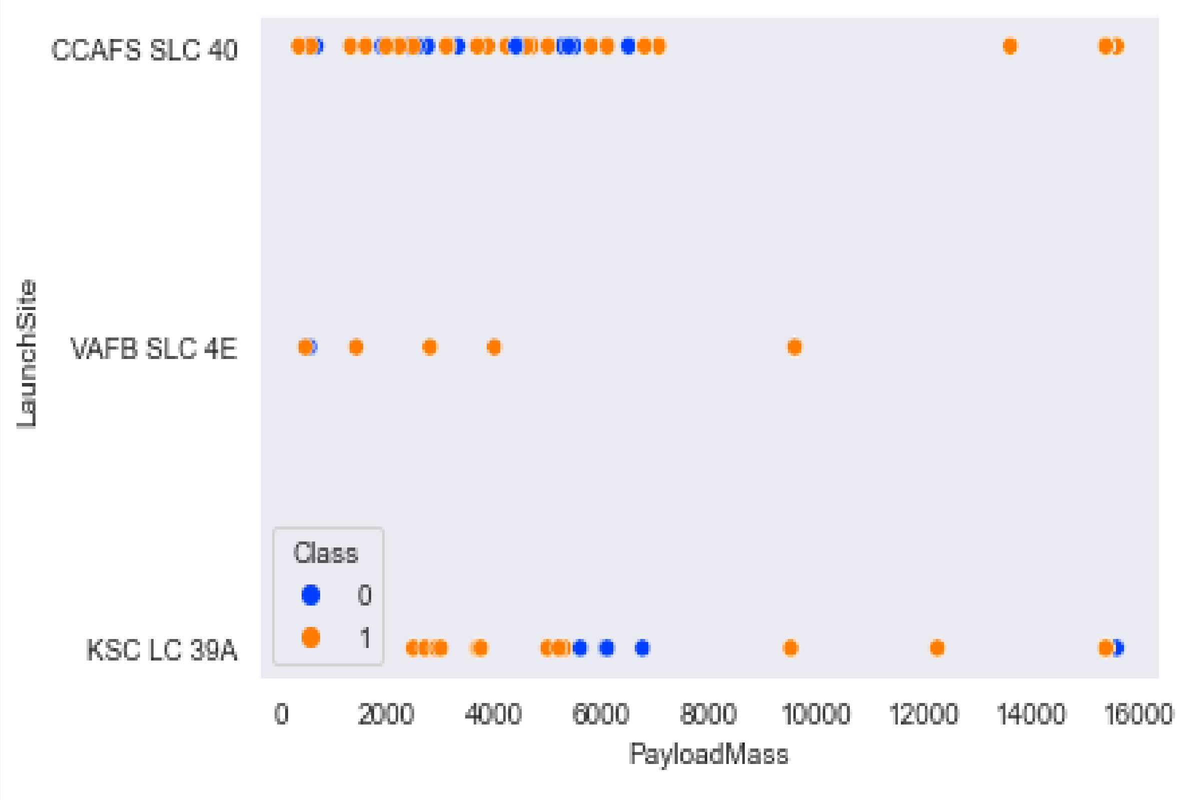
Flight Number vs. Launch Site

- Earlier flights can be seen to be unsuccessful from all sites
- CCAFS SLC 40 can be observed as the oldest and most used site



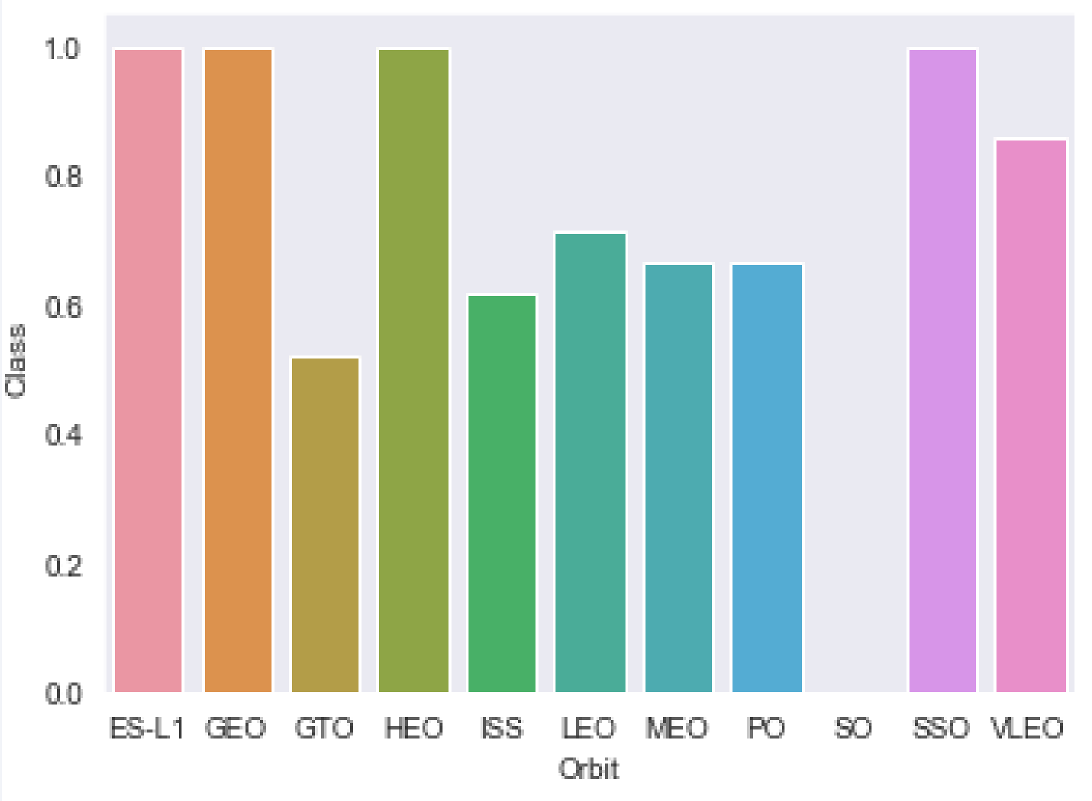
Payload vs. Launch Site

- The average mass of payload seems to be somewhere around 2000 to 8000 kg
- No clear correlation can be seen between these two parameters



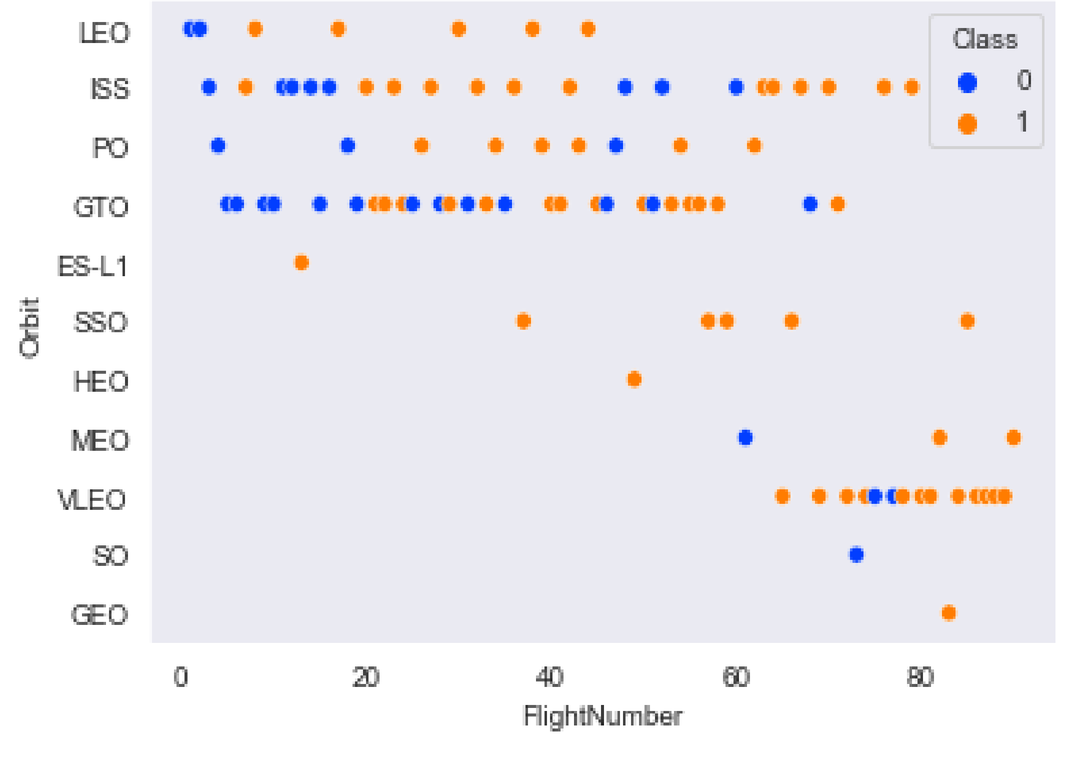
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO exhibit perfect success rates
- SO has a 0% success rate



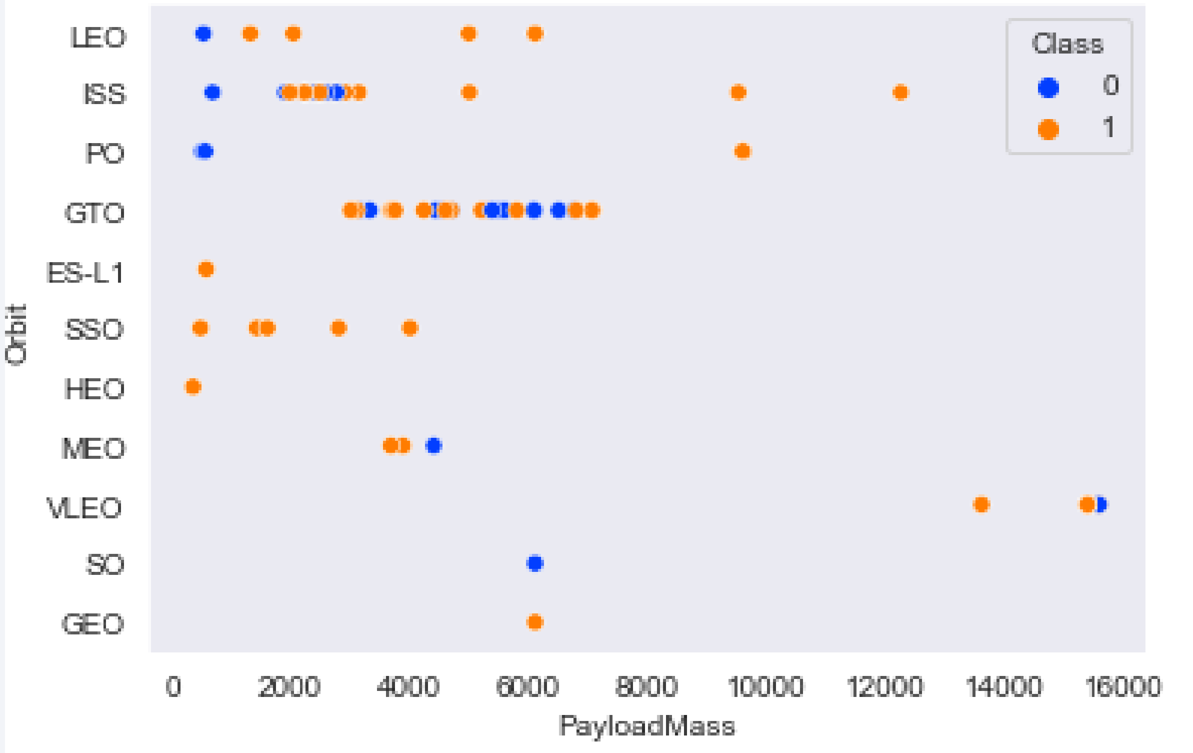
Flight Number vs. Orbit Type

- This can be used to better explain the column chart discussed for Success rate vs Orbit type
- GEO, HEO and ES L-1 have only 1 flight hence the 100% success rate is not as impressive
- SSO has 5, 100% successful flights
- Earlier flights saw lower success rates compared to later ones



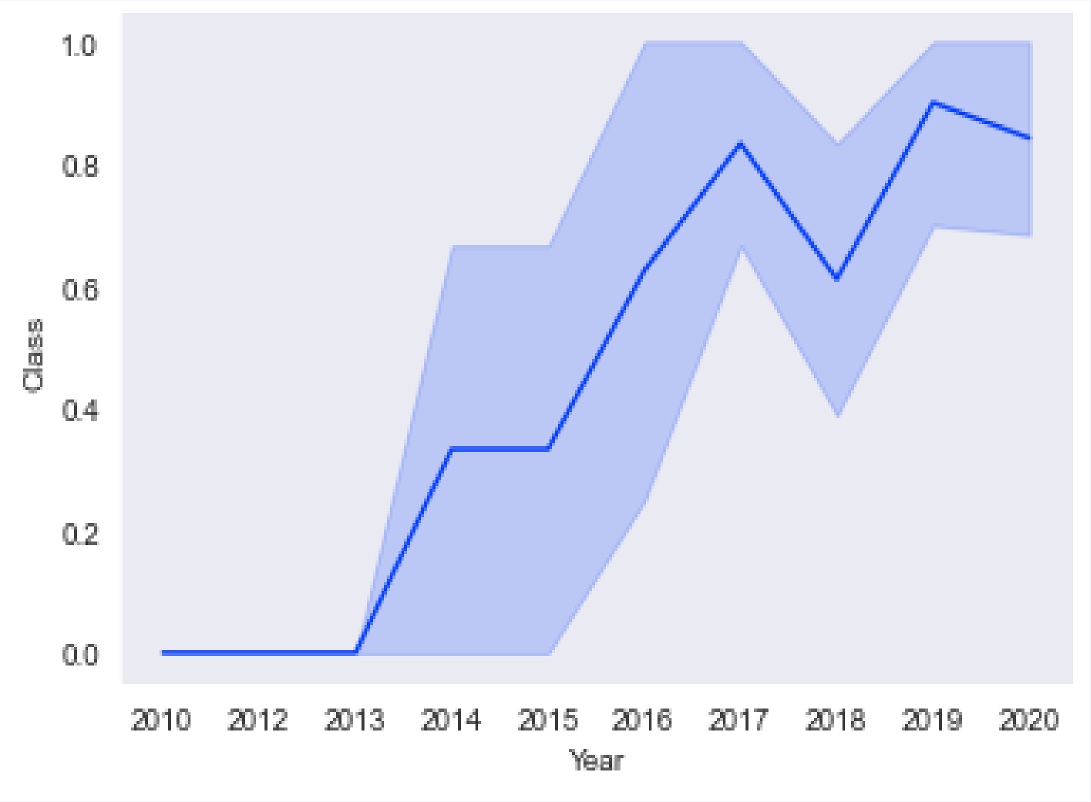
Payload vs. Orbit Type

- Heavier payloads can be seen to be more successful
- Overall, the relationship is very scattered and its difficult to form a correlation



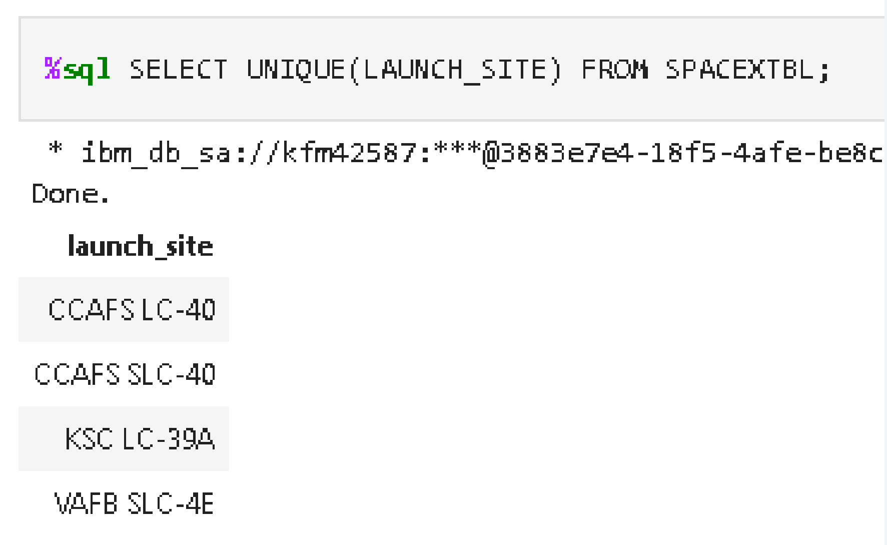
Launch Success Yearly Trend

- Initial years can be seen to be very unsuccessful up till 2013
- Afterwards there is a trend of general increase in the success
- Sharp dip can be observed in 2018 but things seem to be back on track afterwards



All Launch Site Names

- UNIQUE command only returns the unique launch sites from the database



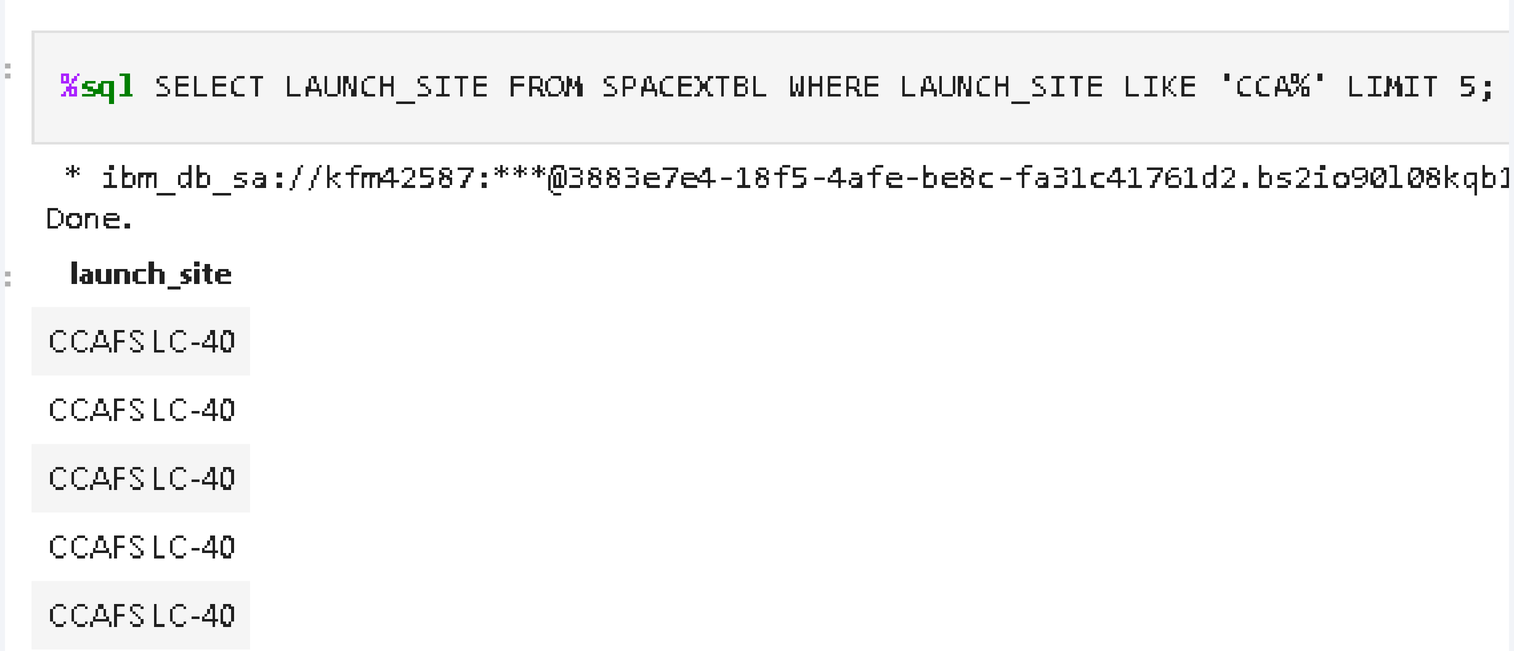
```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c
Done.
```

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

Launch Site Names Begin with 'CCA'

- ‘%’ is a wildcard that is used in SQL, ‘CCA%’ only gets entries starting with CCA
- LIMIT is used to only fetch a certain number of entries, in this case, 5



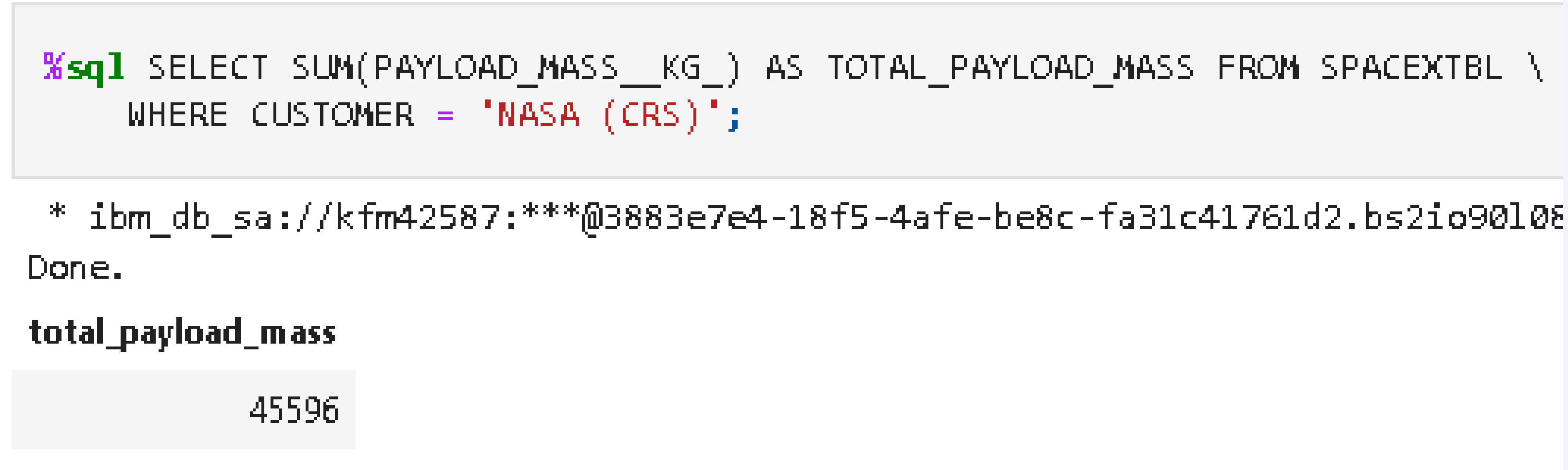
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1
Done.
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- SUM is a simple function that was used to sum the payload masses in the database



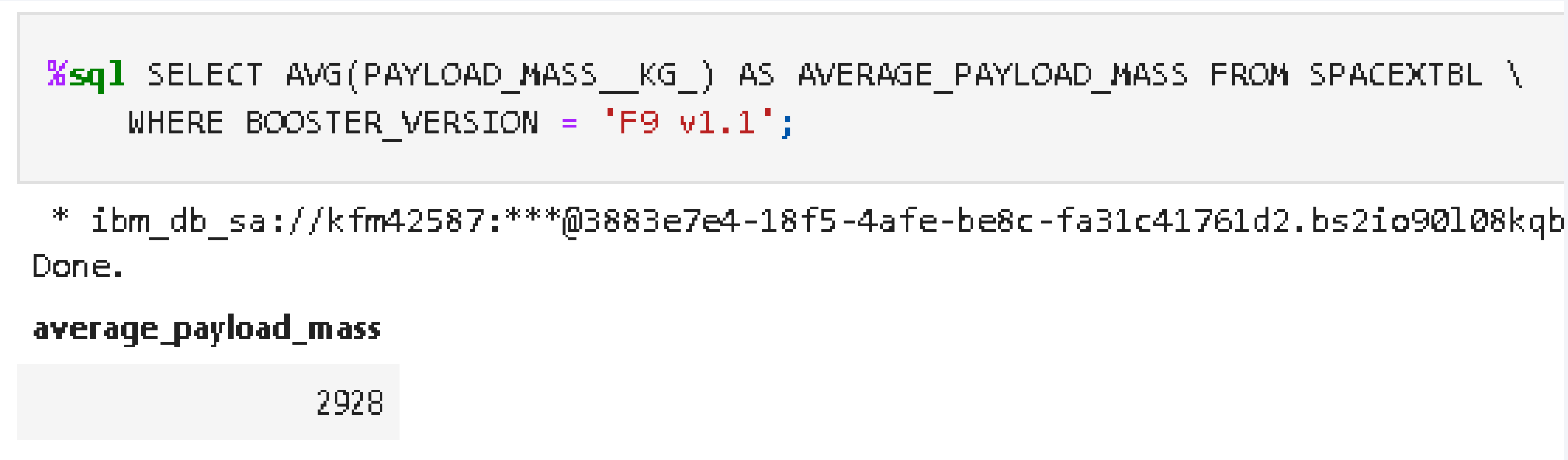
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- AVG is used to calculate the average of any given column
- WHERE function is used to the check the condition to see that only the entries having ‘F9 v1.1’ are used for the calculation



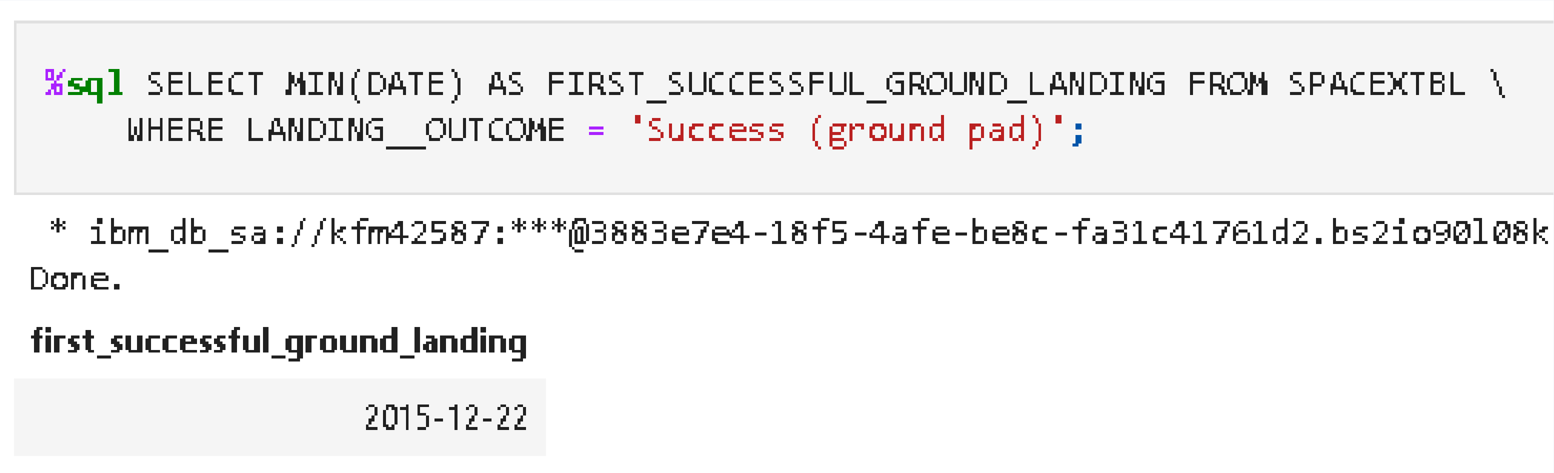
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb
Done.
```

average_payload_mass
2928

First Successful Ground Landing Date

- MIN argument to used to fetch the earliest date
- WHERE function is used to check the successful condtion



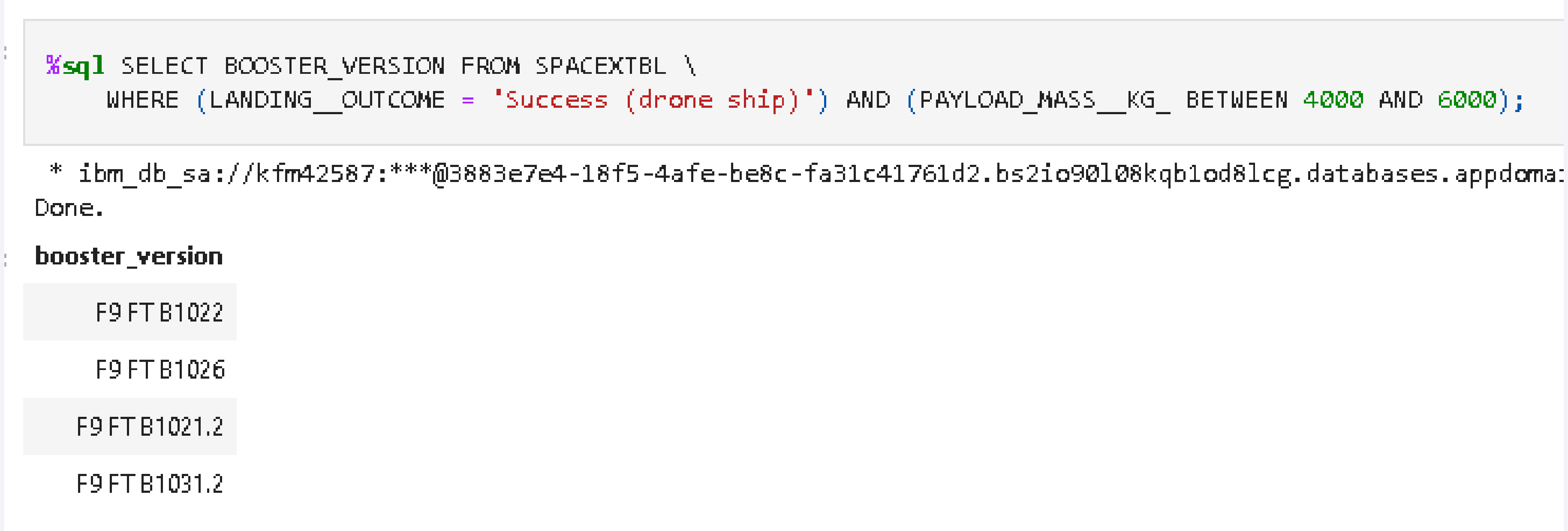
```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08k
Done.
```

first_successful_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE is used to set the condition in the argument, Success in this case
- AND function is used to make sure both conditions are fulfilled
- BETWEEN is used to set the range of the payload mass



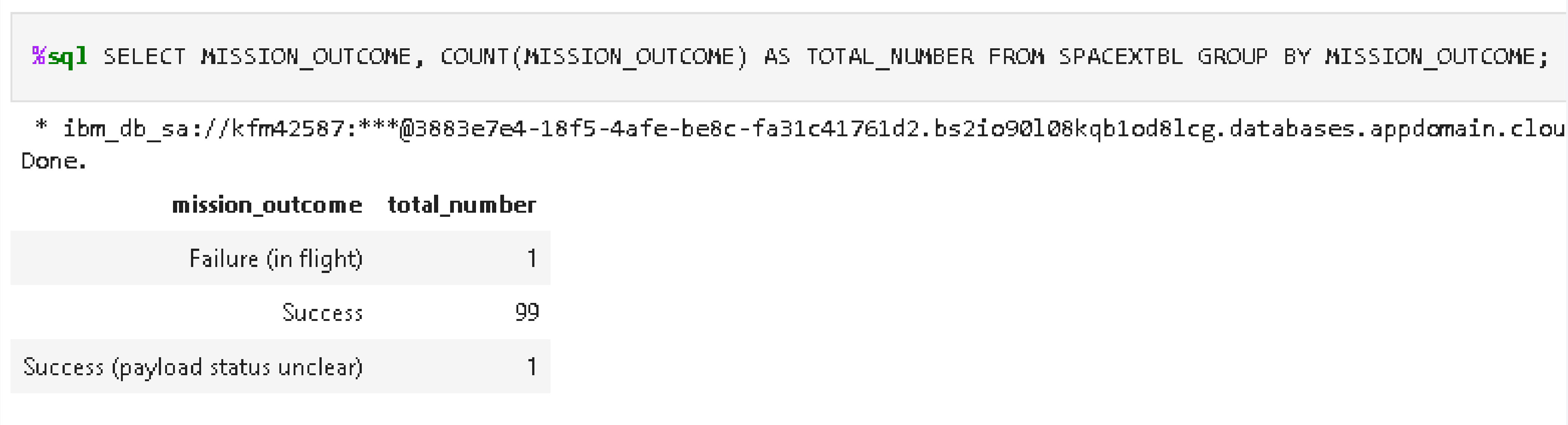
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdoma
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- COUNT is used to count the total number of occurrences against an argument in the database
- AS is used to store the values under a new header
- GROUP BY is used to arrange certain values in a separate table



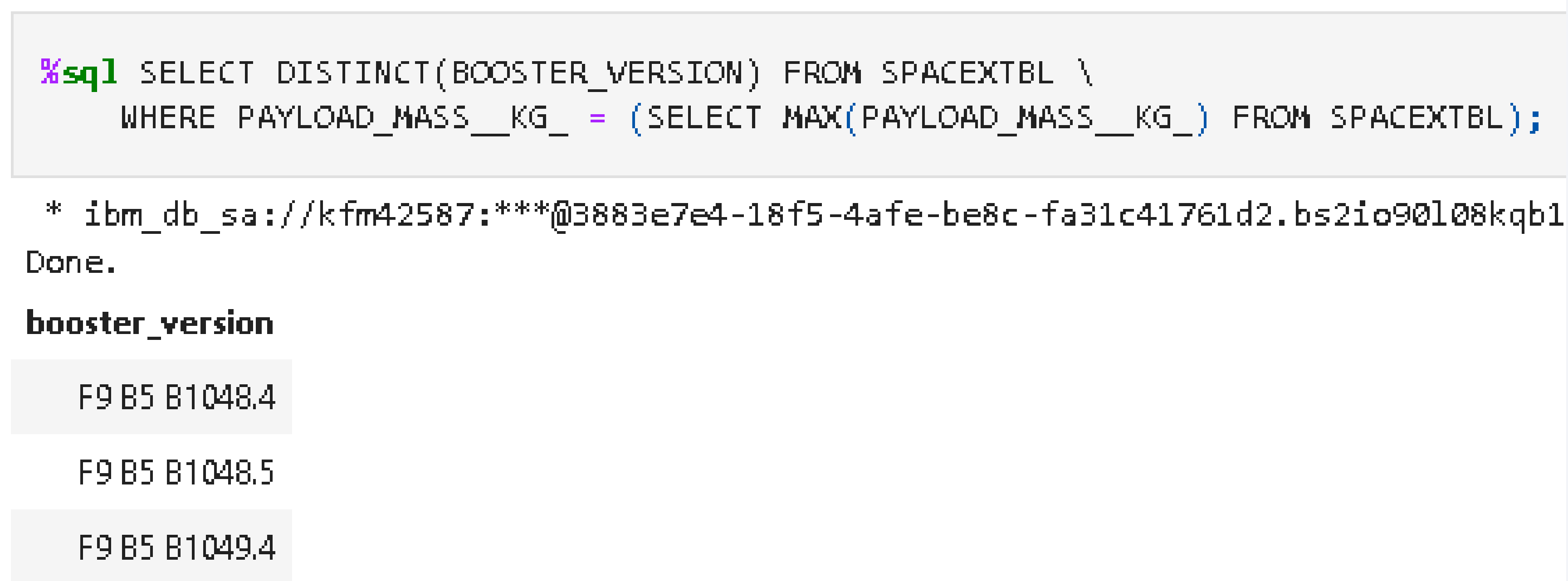
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.clou
Done.
```

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

Boosters Carried Maximum Payload

- DISTINCT is used so that the query does not count the same entries more than once
- The second SELECT statement in the brackets show that a sub-query was used in this



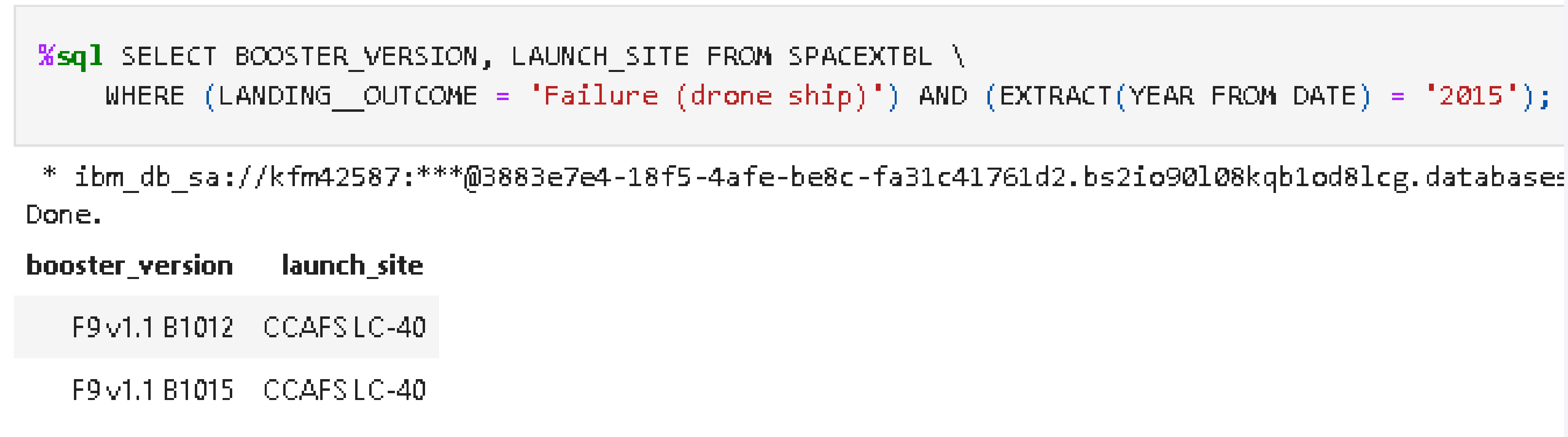
```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4

2015 Launch Records

- The launch records for the year 2015 were tabulated using this query where the landing was unsuccessful



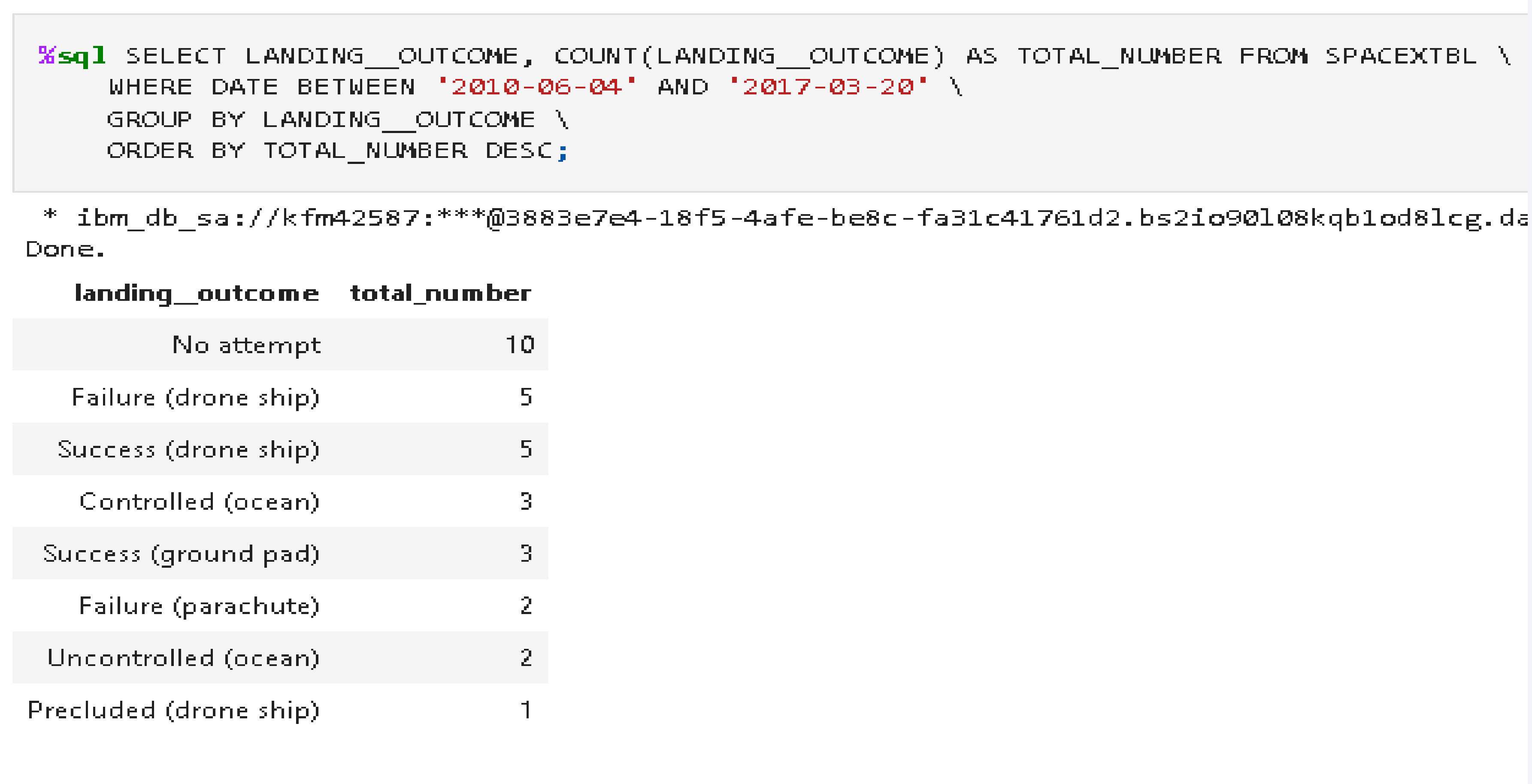
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015');
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The query shows the landing outcomes between the given dates and presents them in a table as can be seen



```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \
     WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
     GROUP BY LANDING__OUTCOME \
     ORDER BY TOTAL_NUMBER DESC;
```

```
 * ibm_db_sa://kfm42587:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.da
Done.
```

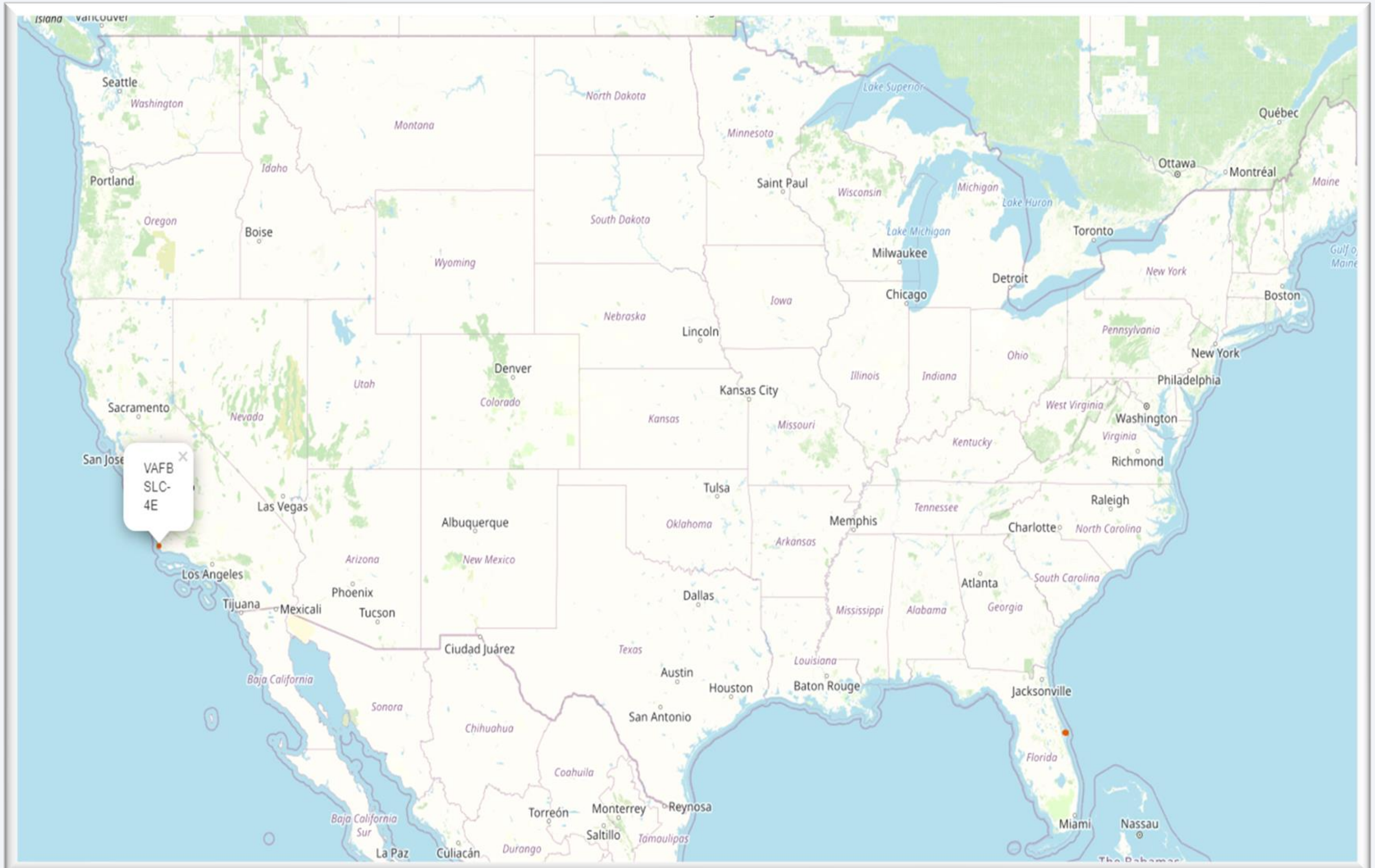
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

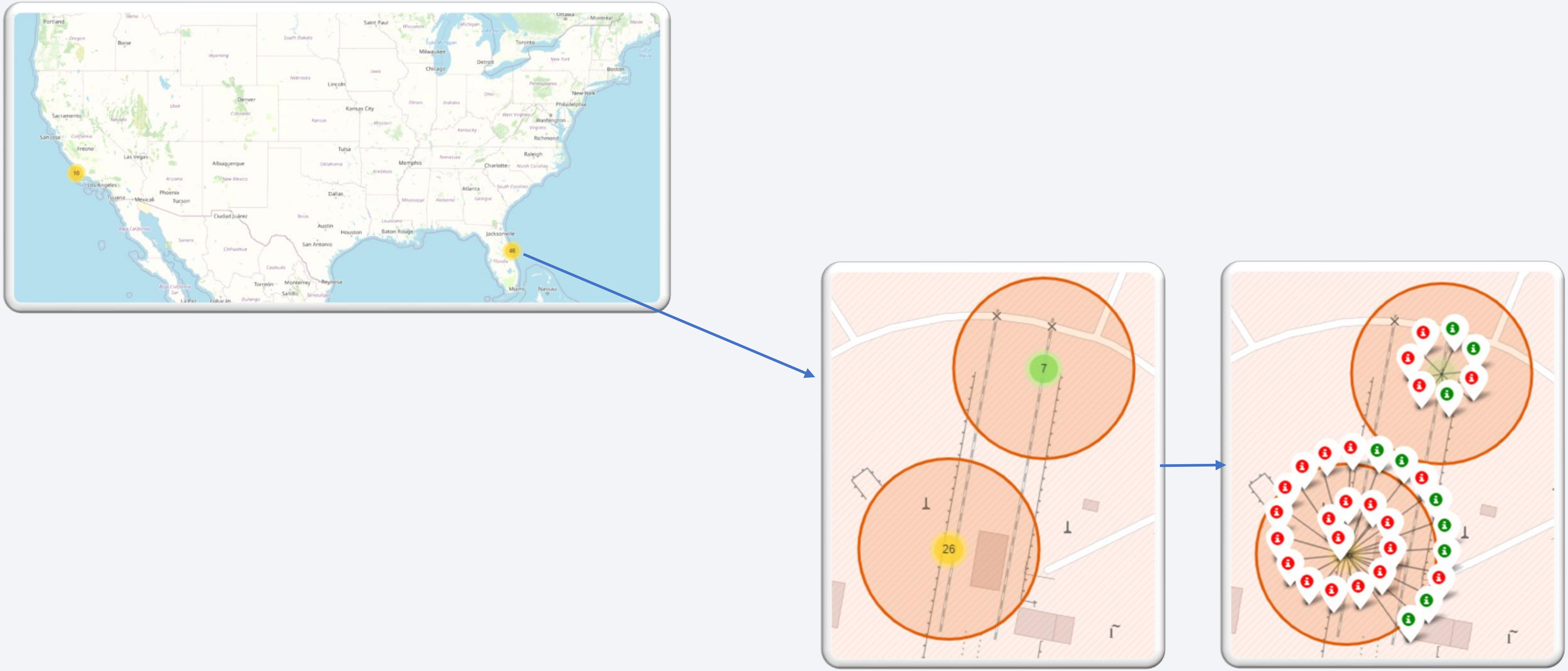
Launch Sites Proximities Analysis

All Launch Sites

- All sites are located on USA’s coast



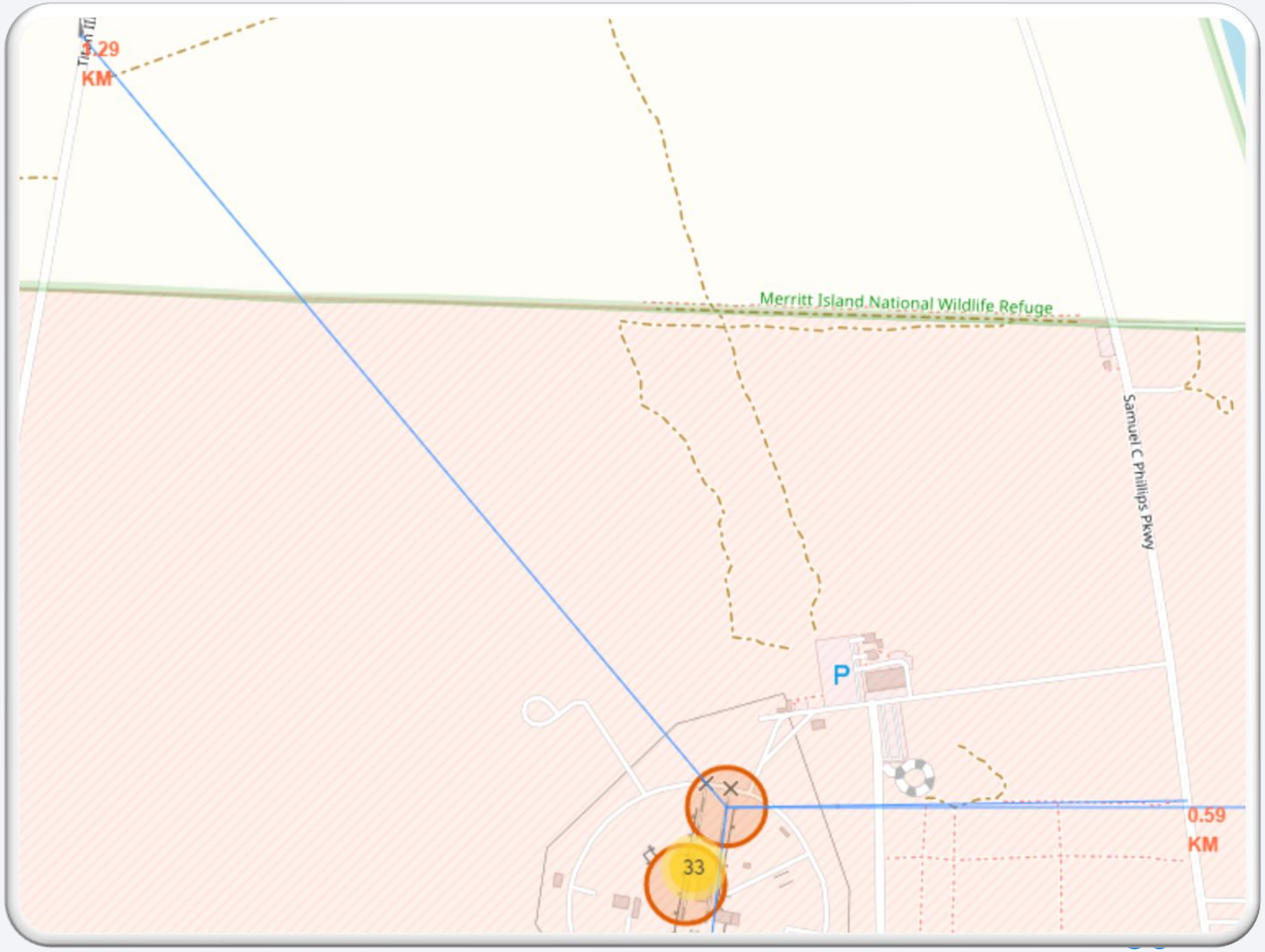
Success and Failures in Launch Sites

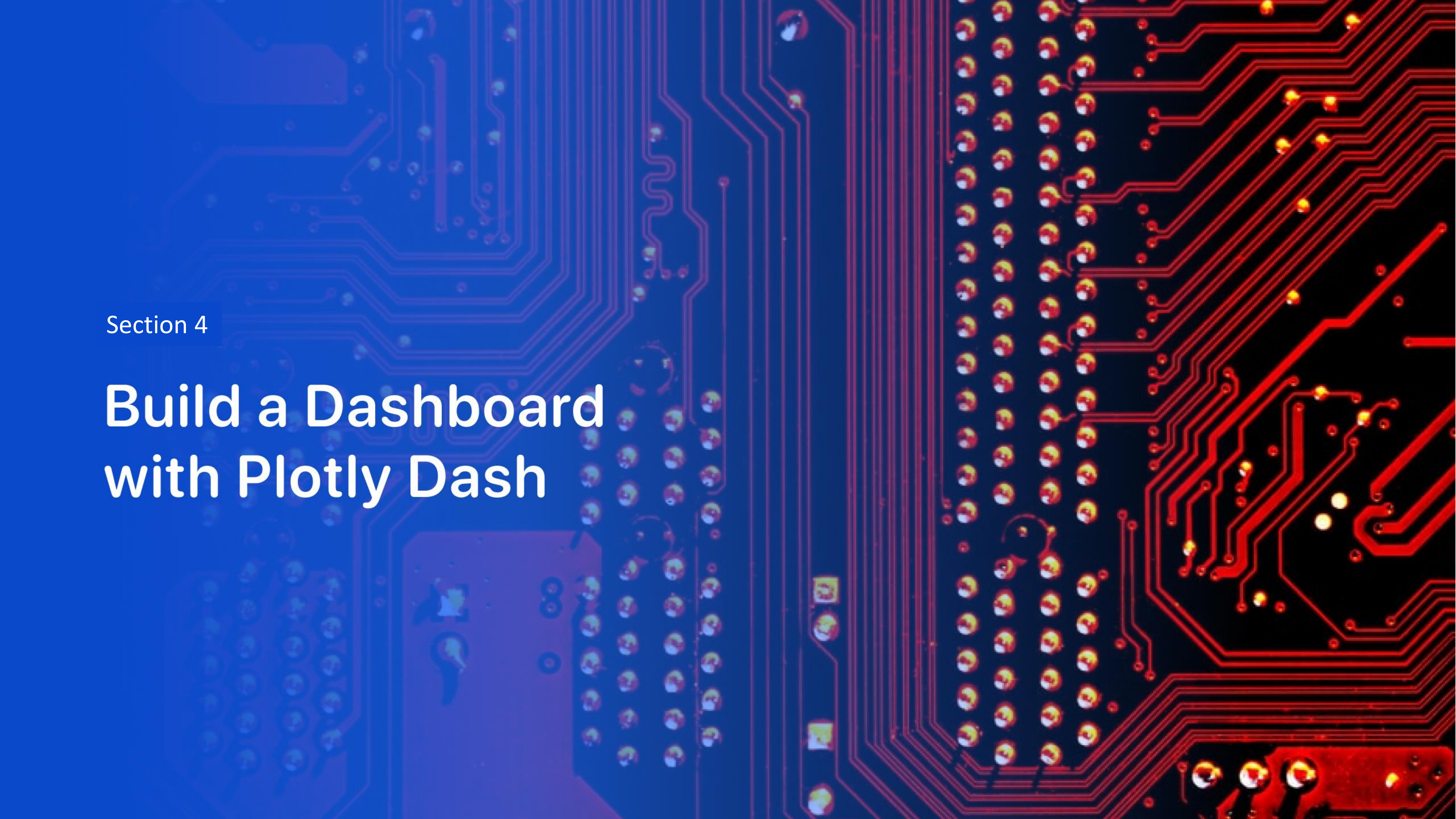


CCAFS SLC-40

Proximity to Points of Interest

- Distances marked in red
- Blue lines show the direction of the launch site to point of interest



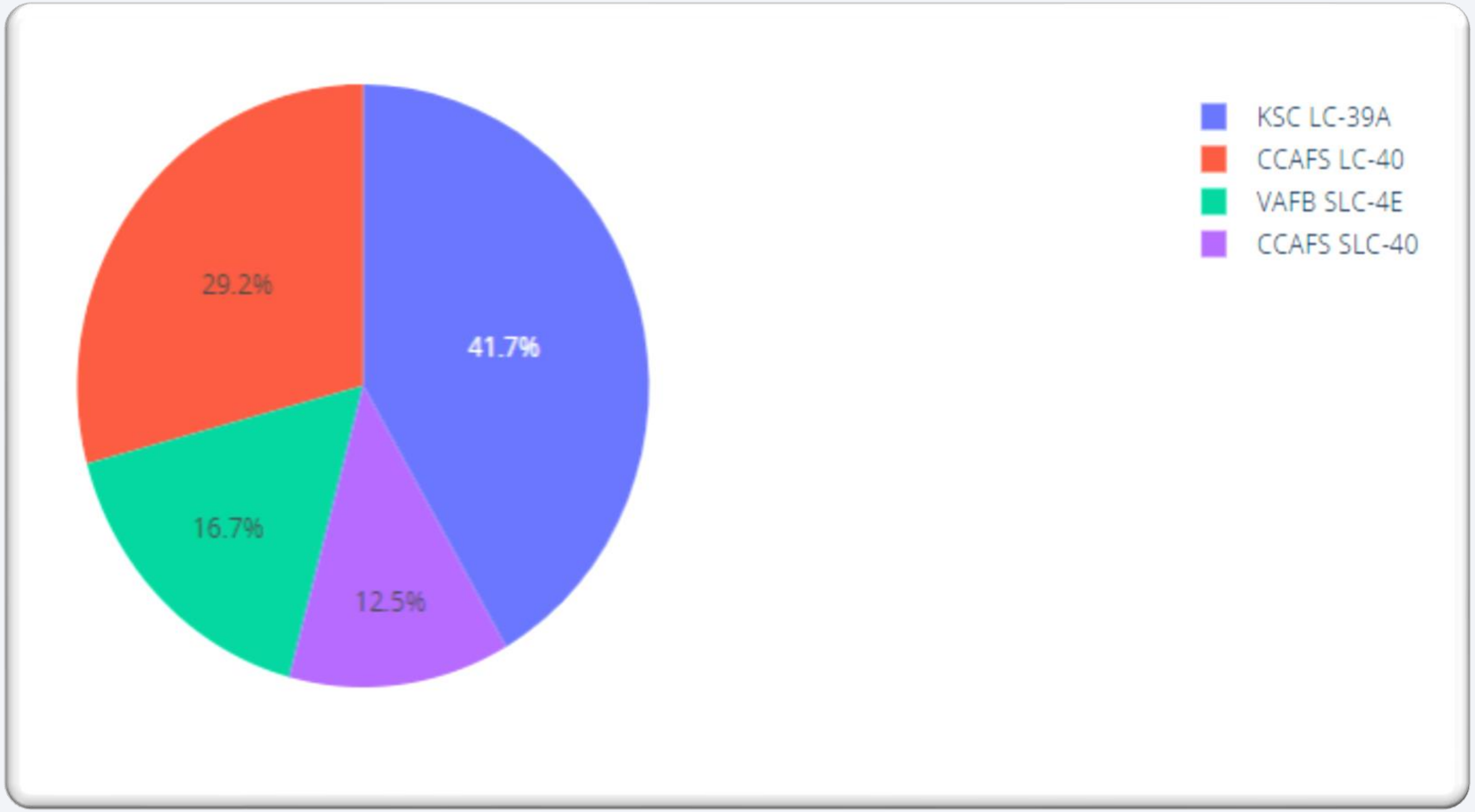


Section 4

Build a Dashboard with Plotly Dash

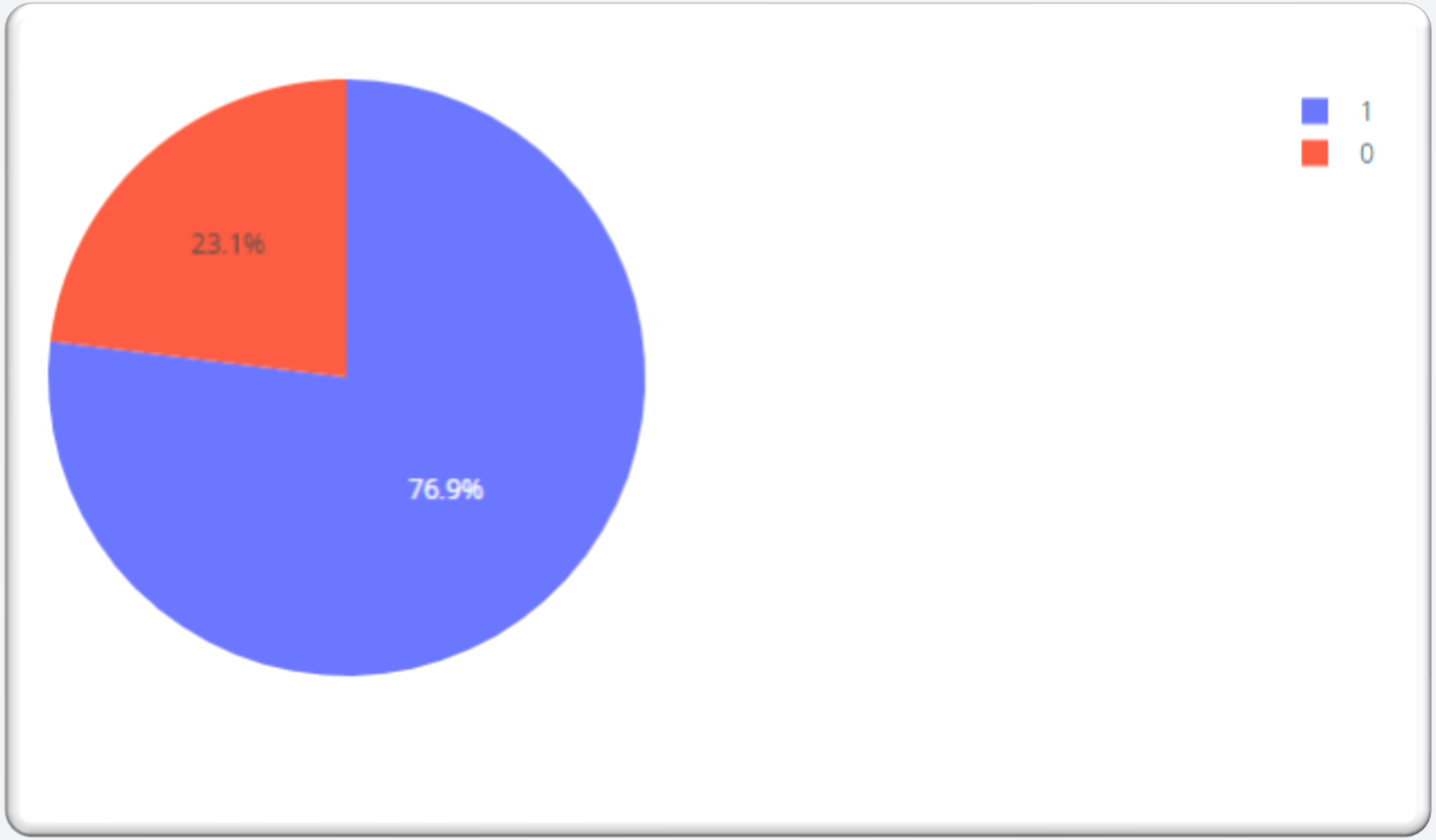
Successful Launches for All Sites

Pie chart showing the division of successful launches between all sites



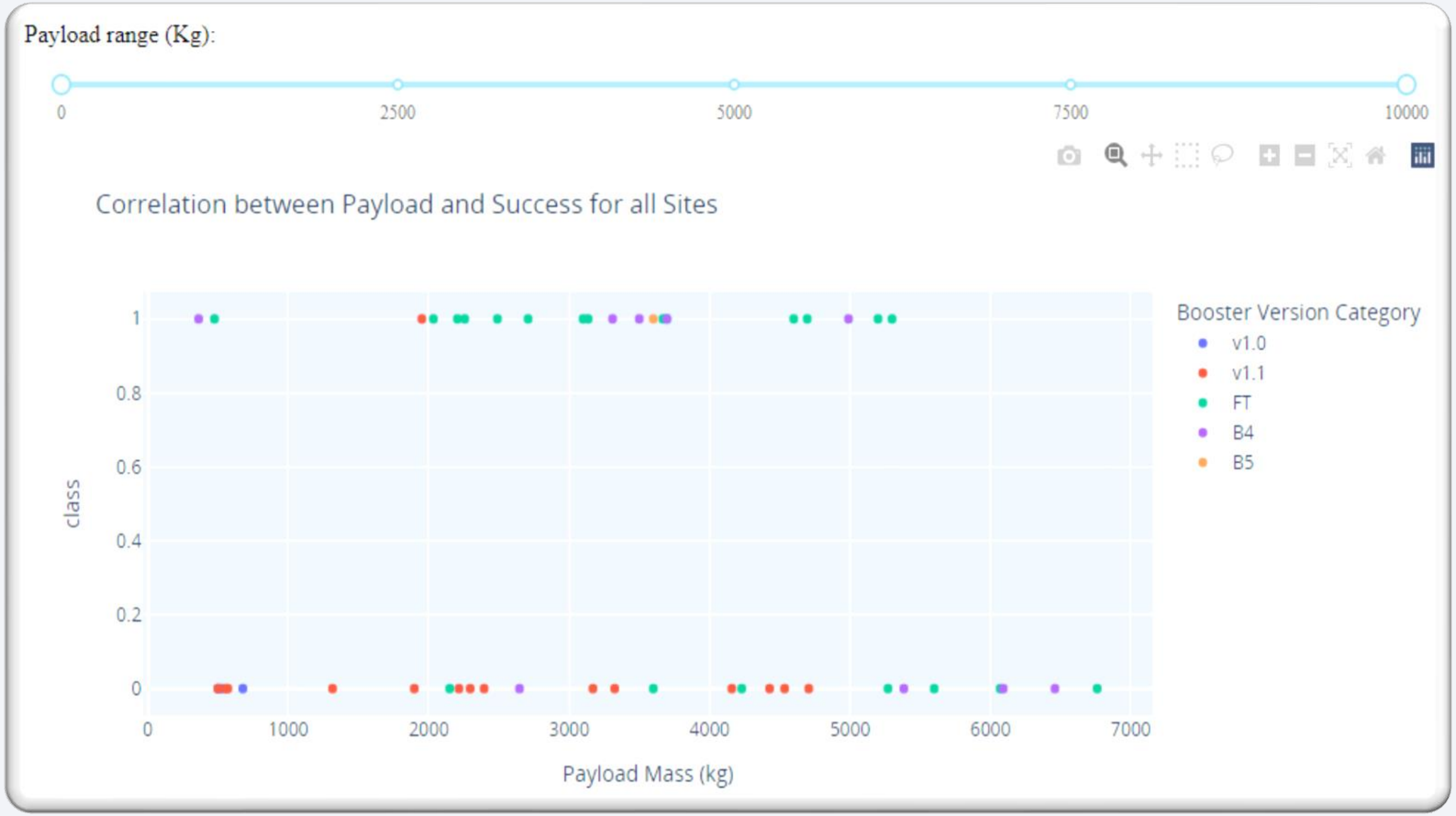
Launch Success Distribution for KSC LC 39-A

Distribution of launch success for the most successful site KSC LC-39A



Payload Mass vs Launch Outcome for All Sites

The graph shows the correlation between the mass of the payload and the outcome of the launch

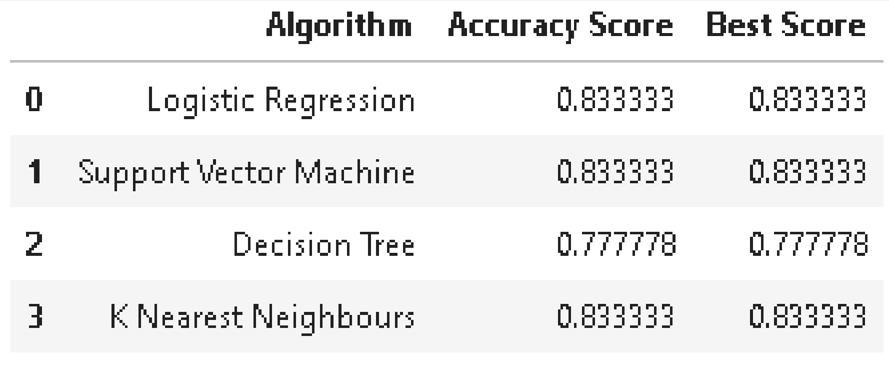


Section 5

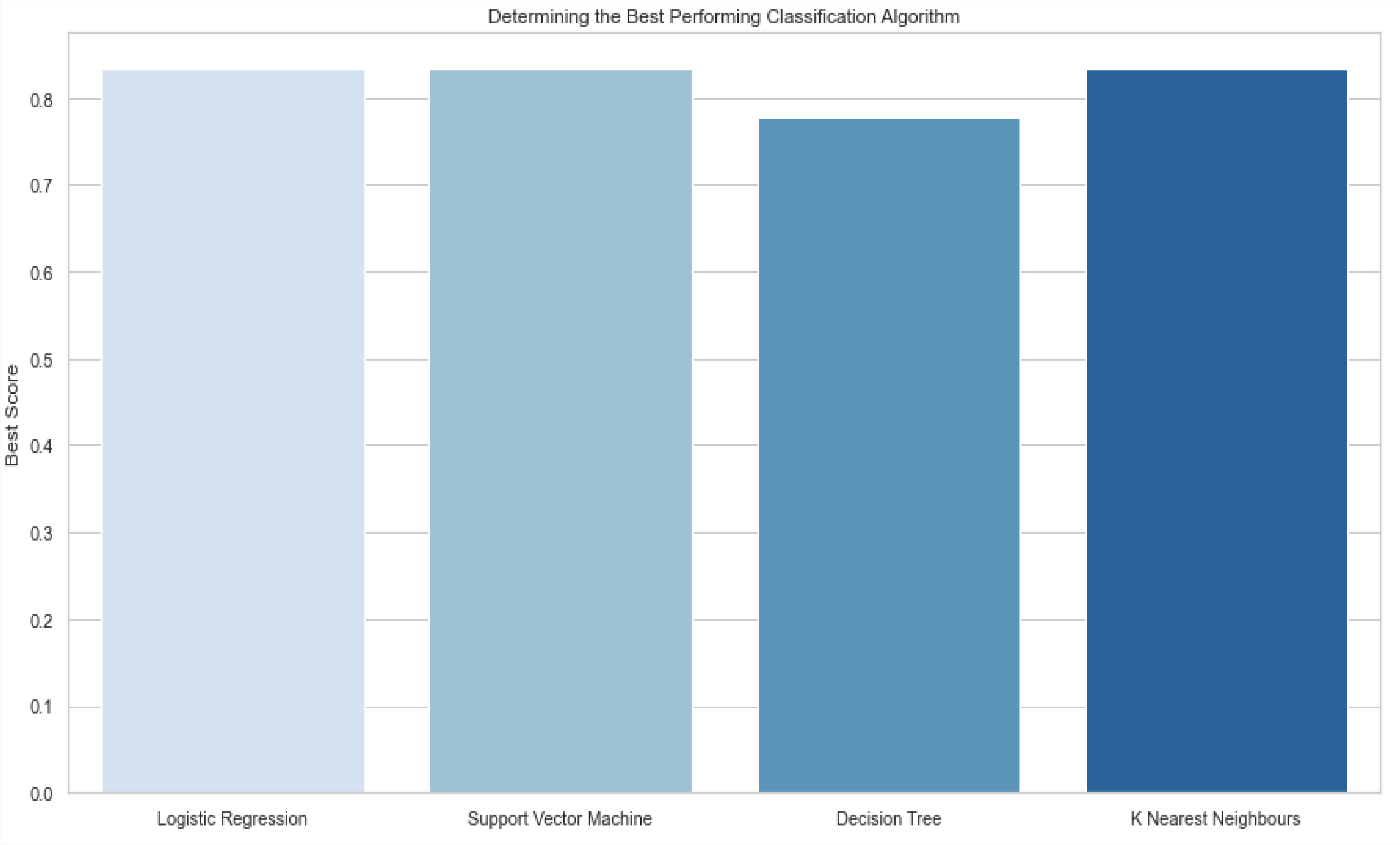
Predictive Analysis (Classification)

Classification Accuracy

- The accuracy scores and best scores for Logistic Regression, Support Vector Machine and K Nearest Neighbors is the same
- They are all feasible models for this project

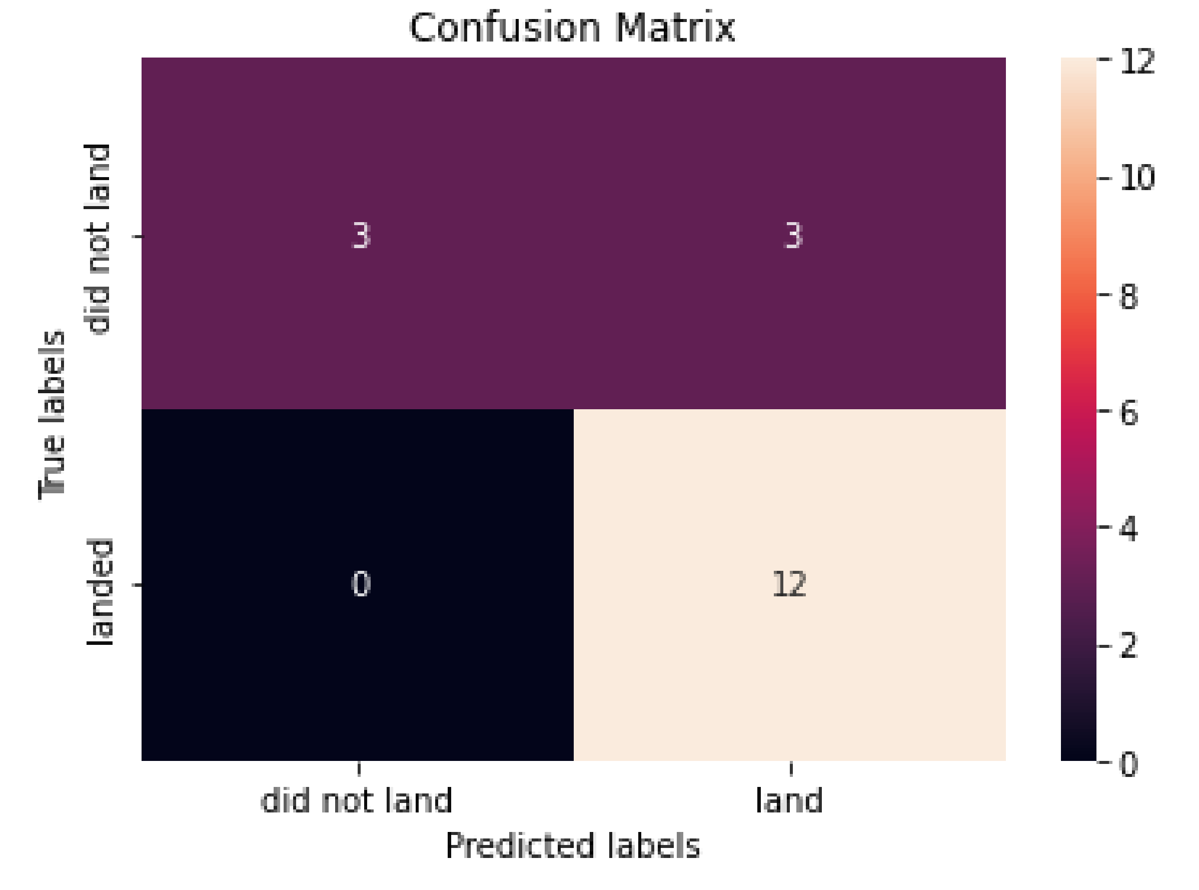


	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.833333
1	Support Vector Machine	0.833333	0.833333
2	Decision Tree	0.777778	0.777778
3	K Nearest Neighbours	0.833333	0.833333



Confusion Matrix

- The confusion matrix is the same for 3 best performing model
- As can be seen from the matrix, only 3 values (top right) give a false positive.
- Rest of the 15 results are correctly classified



Conclusions

- With time the success rate of landings has been steadily increasing. Specially after 2013
- Orbit type SSO has had 5 successful landings, making it 100% successful
- Launch site KSC LC-39A is the site with the most successful launches (76.9%)
- Distribution of payload is not a very good parameter to measure the success of launches
- Logistic Regression, Support Vector Machine and K Nearest Neighbors are equally good models can either can be used in this case

Thank you!