IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SHI YUN FOO
11 Feb 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection via API and web scraping
  - Exploratory Data Analysis(EDA) with SQL and data visualization
  - Visual analytics with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis

- Summary of all results
  - EDA results
  - Interactive map and dashboard
  - Predictive results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars while other providers cost upward of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. Therefore, this project is to predict if the first stage will land successfully which can be used to determine the cost of a launch to compete against SpaceX.

- Problems you want to find answers
  - What are the main characteristics of a successful or failed landing?
  - What are the effects of each variables on the outcome of a landing?
  - What are the conditions which will allow SpaceX to achieve the best landing success rate?

Section 1

Methodology

Methodology

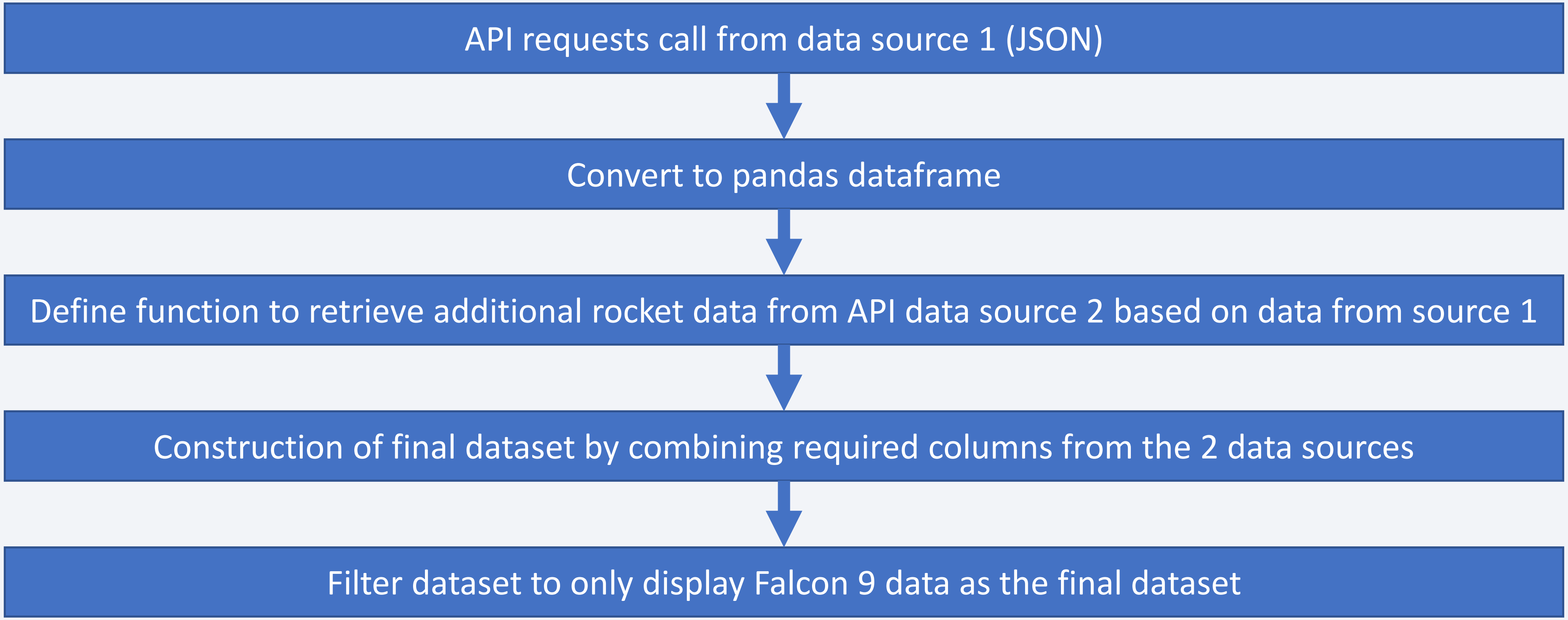
Executive Summary

- Data collection methodology:
  - HTTP requests to get data from SpaceX API via Python requests library
  - Web scrapping of Wikipedia via beautifu lsoup library
- Perform data wrangling
  - Replacing null pay load mass with mean mass
  - Creation of class column to clearly identify successful vs failure landing as training label
- 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 – SpaceX API

API Data Source 1: https://api.spacexdata.com/v4/launches/past
API Data Source 2: https://api.spacexdata.com/v4/rockets

GitHub Link



API requests call from data source 1 (JSON)

Convert to pandas dataframe

Define function to retrieve additional rocket data from API data source 2 based on data from source 1

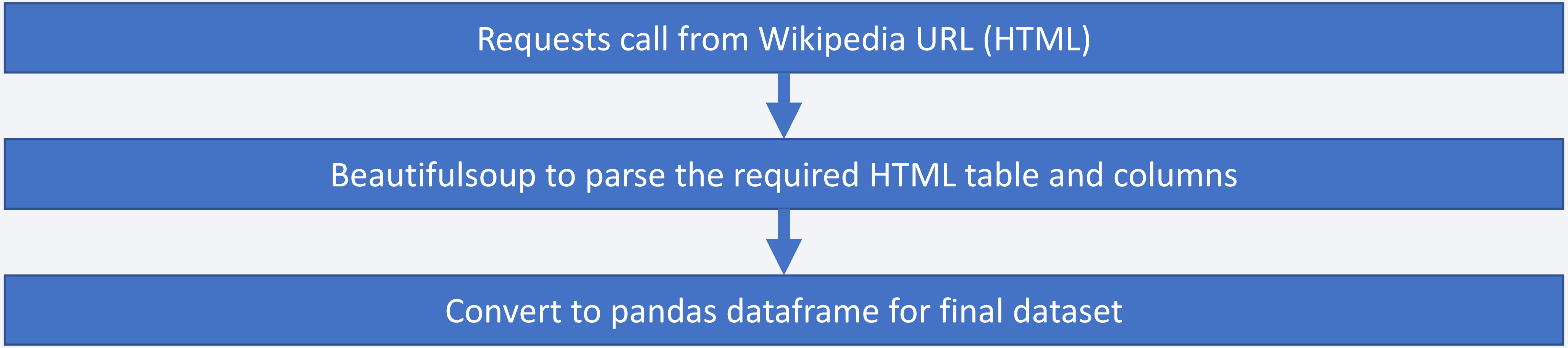
Construction of final dataset by combining required columns from the 2 data sources

Filter dataset to only display Falcon 9 data as the final dataset

Data Collection - Scraping

Wikipedia Data Source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

GitHub Link



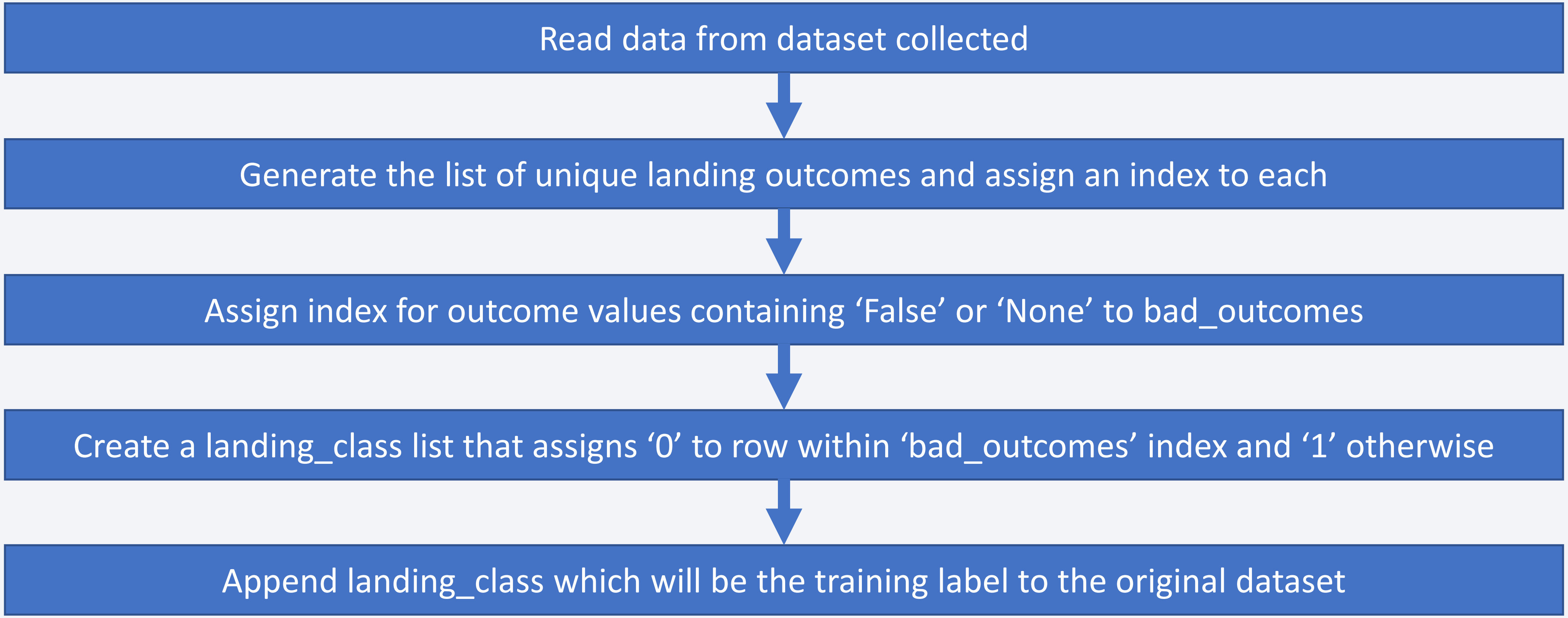
Requests call from Wikipedia URL (HTML)

Beautifulsoup to parse the required HTML table and columns

Convert to pandas dataframe for final dataset

Data Wrangling

GitHub Link



Read data from dataset collected

Generate the list of unique landing outcomes and assign an index to each

Assign index for outcome values containing ‘False’ or ‘None’ to bad_outcomes

Create a landing_class list that assigns ‘0’ to row within ‘bad_outcomes’ index and ‘1’ otherwise

Append landing_class which will be the training label to the original dataset

EDA with Data Visualization

GitHub Link

- Scatter Plot was used for visualization of relationship between 2 variables and can identify the strength, direction and pattern of the relationship.
  - Flight Number vs Pay Load
  - Flight Number vs Launch Site
  - Flight Number vs Orbit
  - Pay Load vs Orbit

- Bar Chart was used for comparing values across categories
  - Mean of success rate for different Orbit

- Line Chart was used for data trending across a time period
  - Success rate per year from 2010 to 2020

EDA with SQL

GitHub Link

SQL queries were written to display the following information

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA’
3. Total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. Date when the first successful landing outcome in ground pad was achieved
6. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Total number of successful and failure mission outcomes
8. Names of the booster_versions which have carried the maximum payload mass
9. Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Rank 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

Build an Interactive Map with Folium

GitHub Link

- Folium map object was created and centered on NASA Johnson Space Center at Houson, Texas. The following markers were added onto the map

  - Red circle at NASA Johnson Space Center's coordinate with label showing its name
  - Red circles at each launch site coordinates with label showing launch site name
  - Group of points were clustered to display data for multiple launches for the same site with Green icon for successful and Red for failed landing
  - Lines to show distance between launch site to key locations (coastline, highway, railway, city)

These information can help to visualize the geospatial distribution of the launch sites with their corresponding landing success rate and the potential relationship between the environment/location vs. success rate.

Build a Dashboard with Plotly Dash

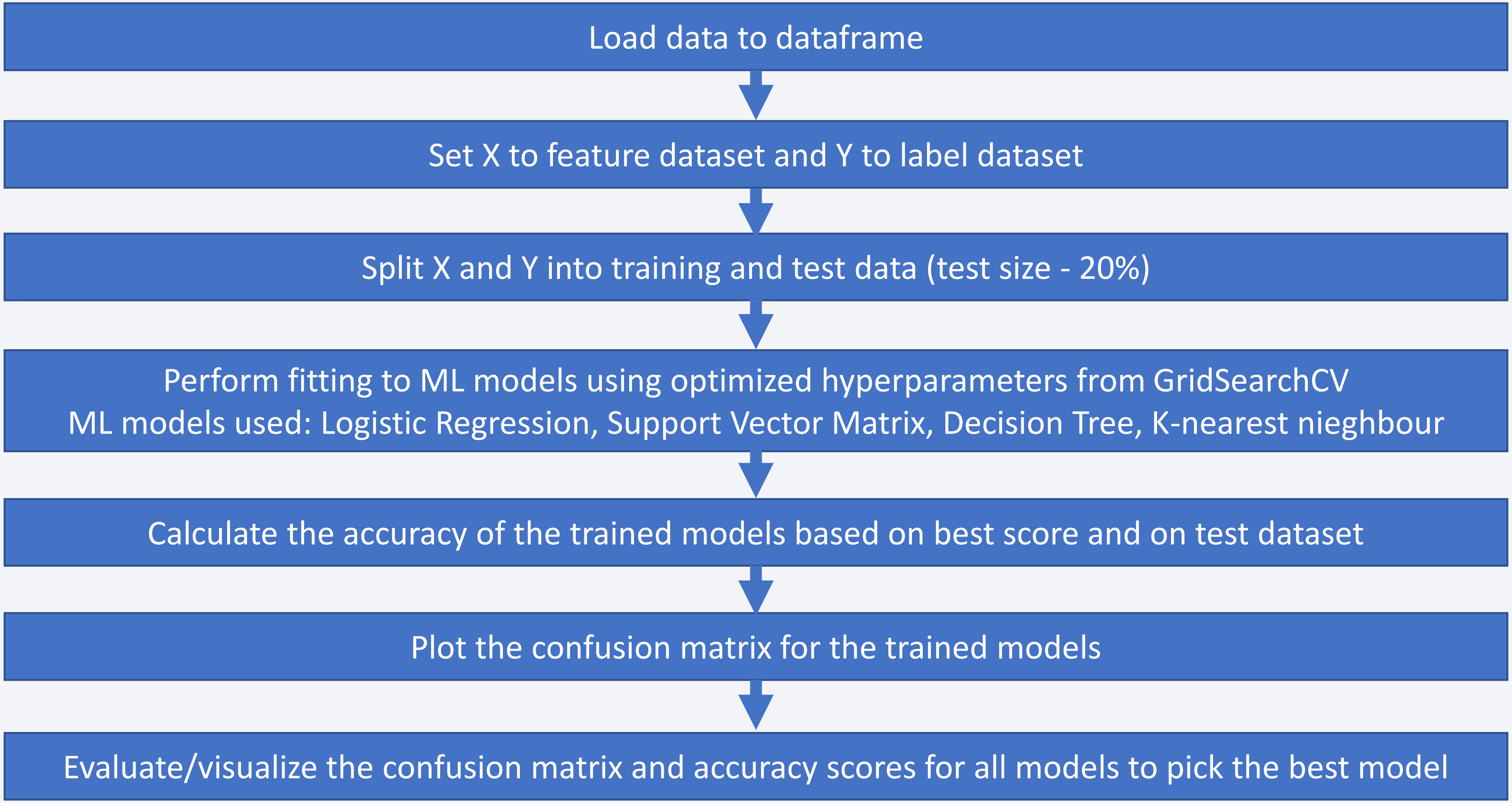
GitHub Link

The following components were added to the Plotly Dash dashboard

- Dropdown selection box – To allow user to select a specific launch site or all launch sites
- Pie Chart – To visualize the percentage of successful vs failed landing
- Range Slider – To allow user to specify the range of payload mass
- Scatter Chart – To visualize the relationship between landing outcome vs payload mass

Predictive Analysis (Classification)

GitHub Link



Load data to dataframe

Set X to feature dataset and Y to label dataset

Split X and Y into training and test data (test size - 20%)

Perform fitting to ML models using optimized hyperparameters from GridSearchCV
ML models used: Logistic Regression, Support Vector Matrix, Decision Tree, K-nearest nieghbour

Calculate the accuracy of the trained models based on best score and on test dataset

Plot the confusion matrix for the trained models

Evaluate/visualize the confusion matrix and accuracy scores for all models to pick the best model

Results

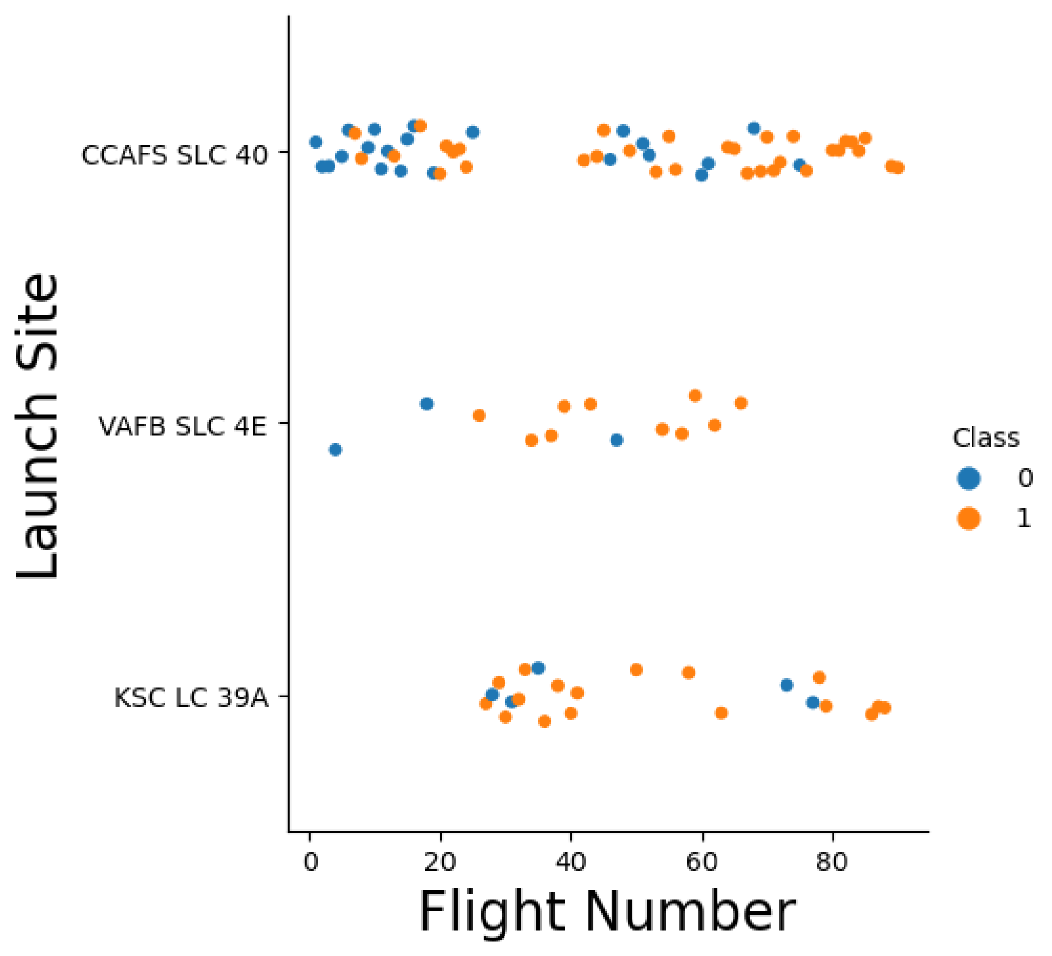
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

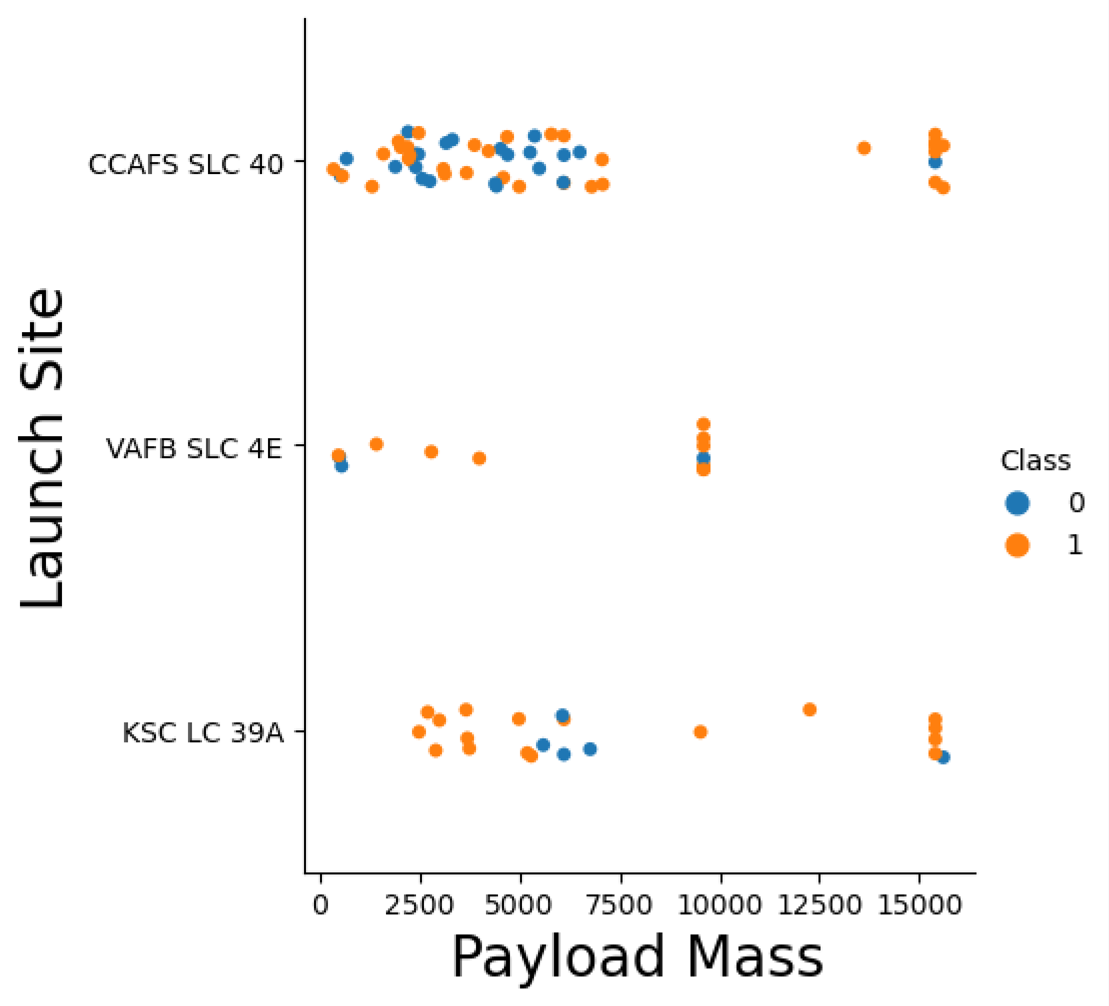
Scatter plot:



There is an increase in success rate with higher flight number for CCAFS and VAFB, however it is not as obvious for KSC

Payload vs. Launch Site

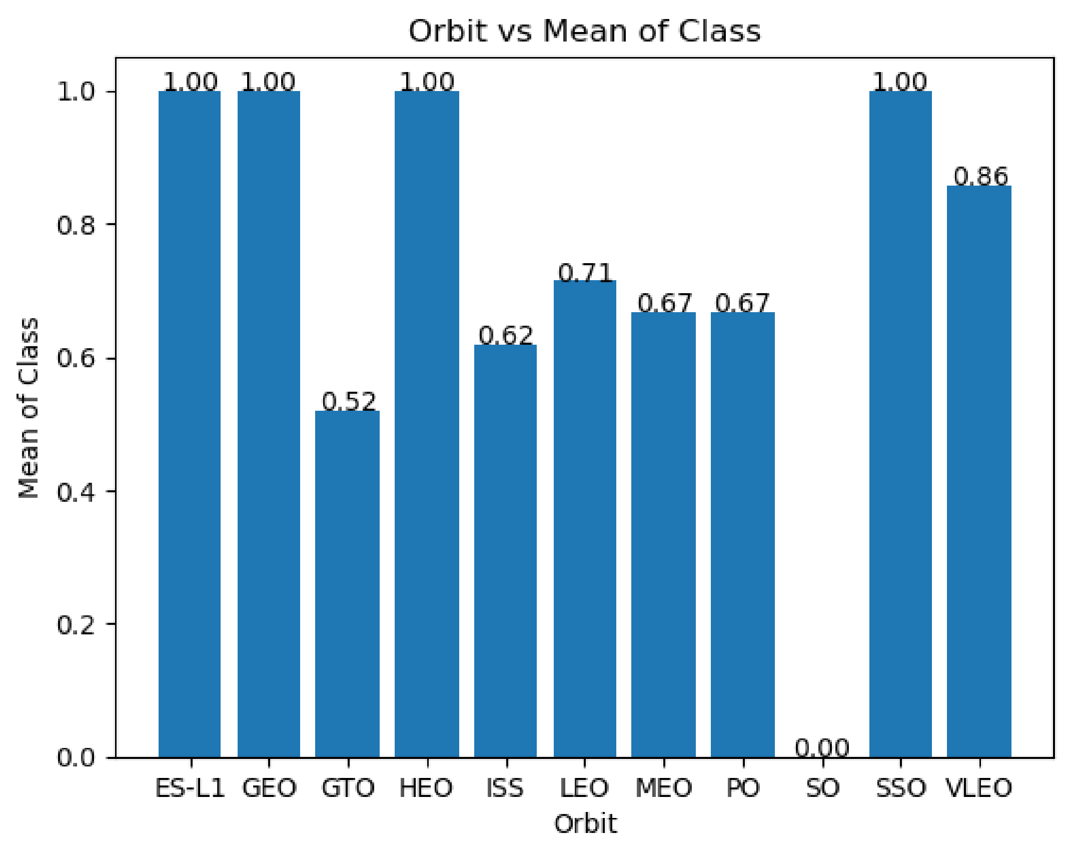
Scatter plot:



There isn’t a clear relationship between payload mass and success rate of landing

Success Rate vs. Orbit Type

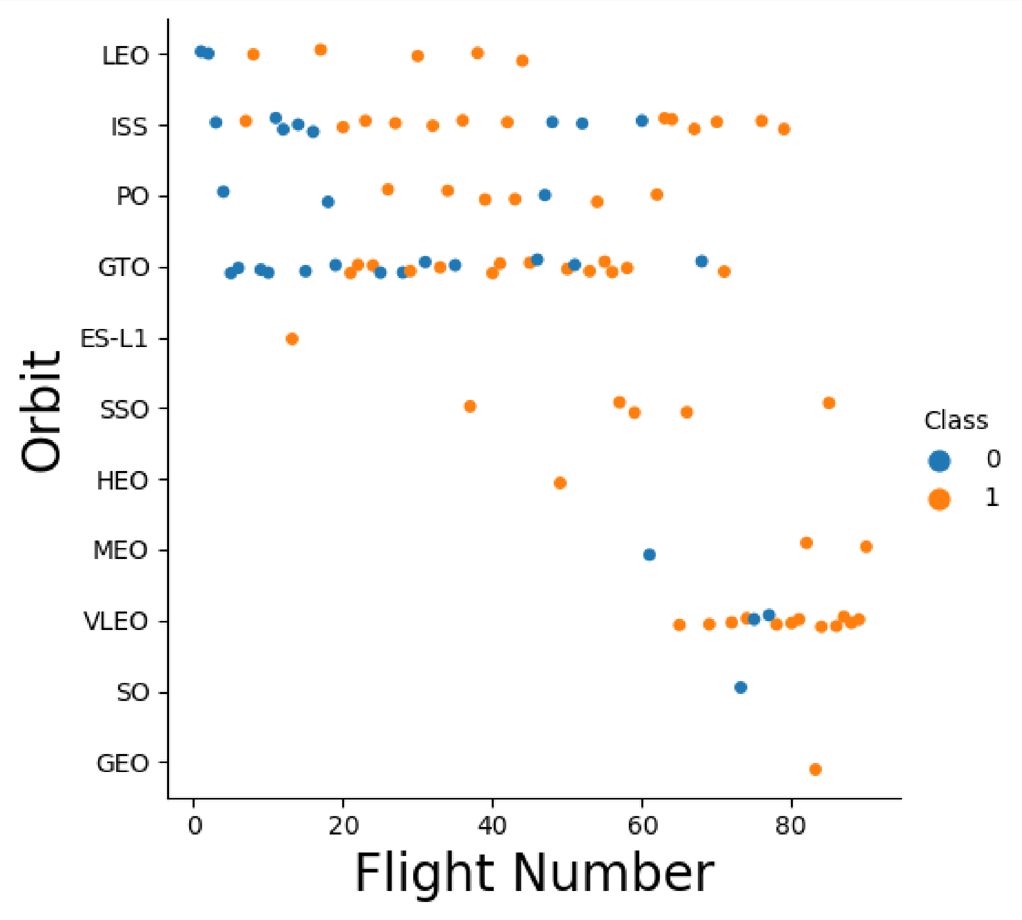
Bar Chart:



- ES-L1, GEO, HEO and SSO have 100% successful landing but ES-L1, GEO and SSO has only 1 launch each and hence it is not representative
- Among the Orbits with higher number of launches (>5), VLEO has the highest success rate and GTO has the lowest success rate

Flight Number vs. Orbit Type

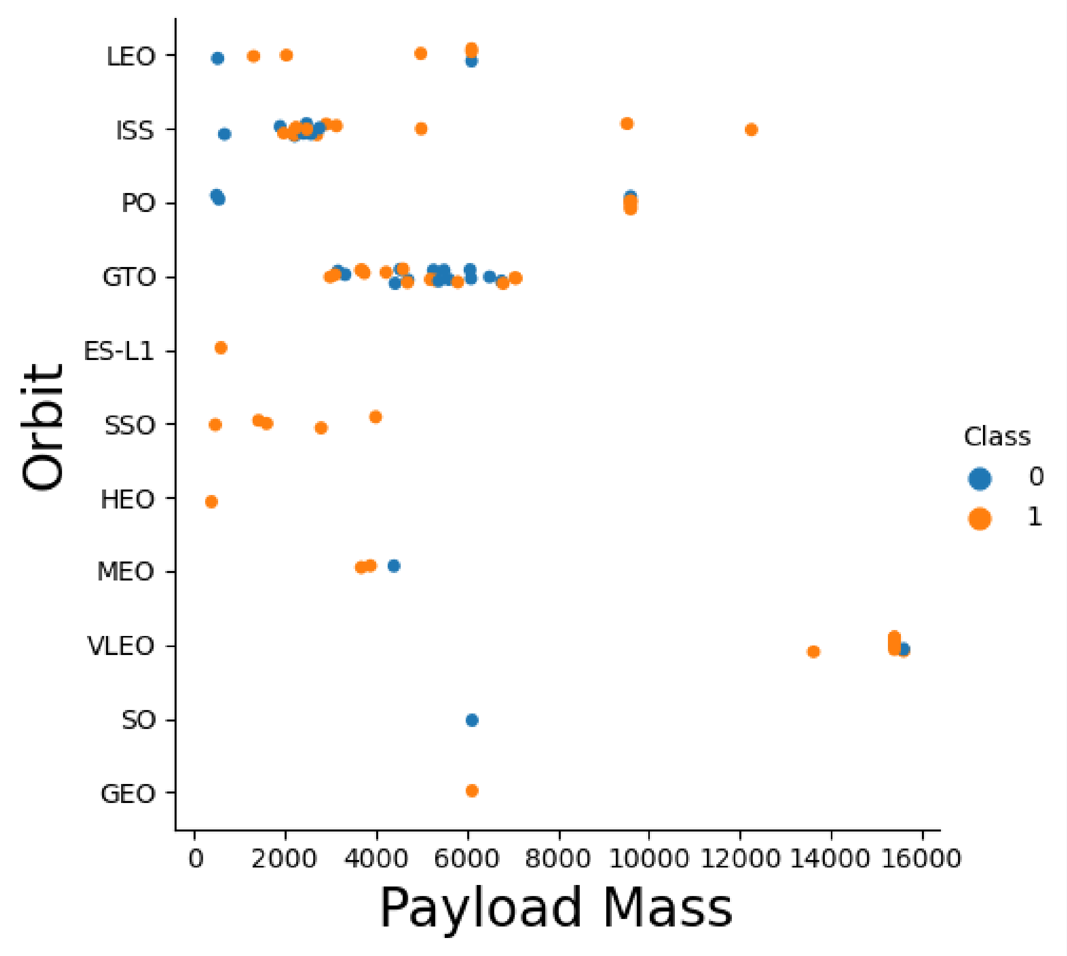
Scatter plot:



- There seem to be some relationship between flight number and success rate for LEO
- Most of the failures were between flight 0 to 20 when the technology was not as advanced

Payload vs. Orbit Type

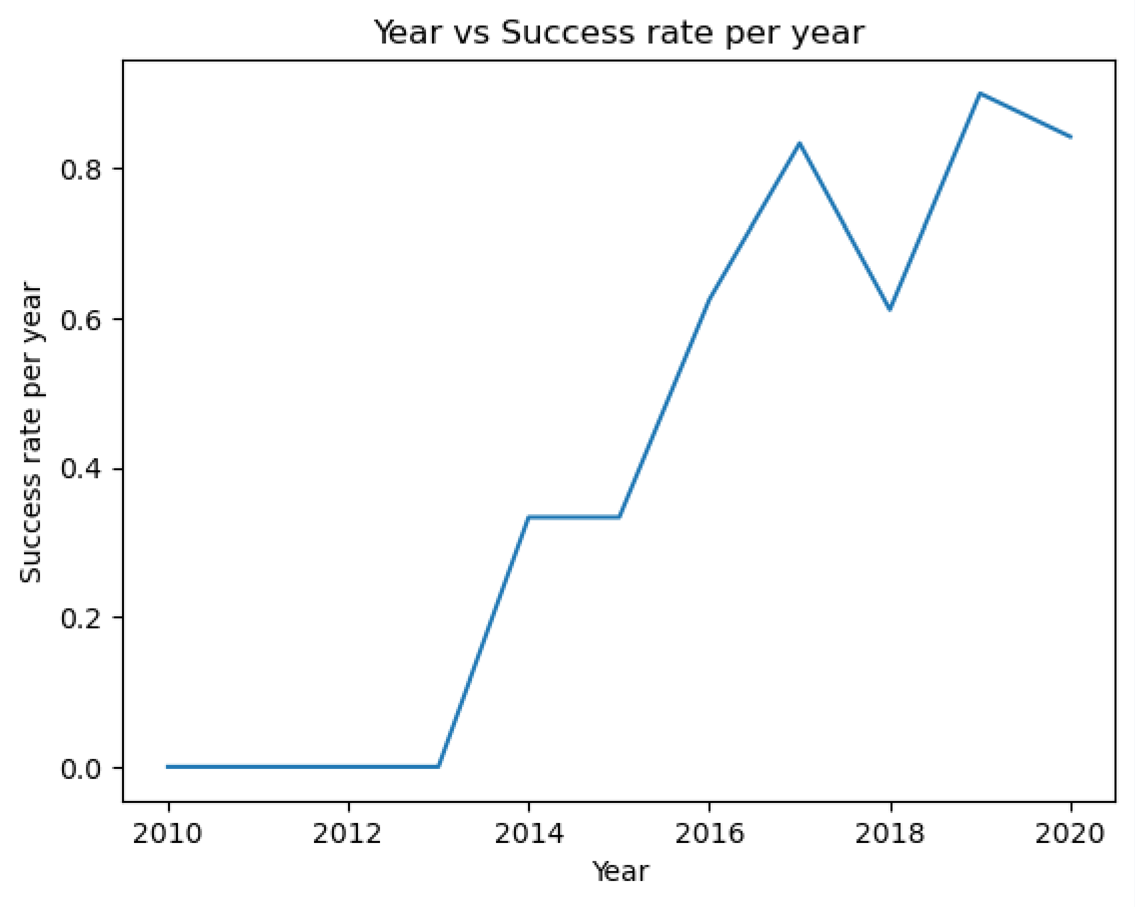
Scatter plot:



- There is a higher success rate at higher payload mass for ISS. However, this relationship cannot be clearly seen for the other Orbits

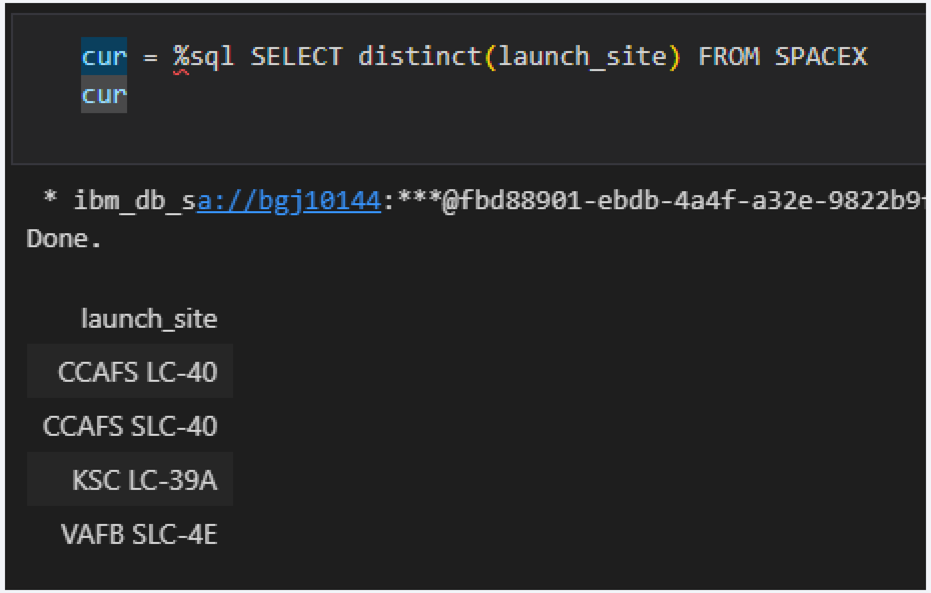
Launch Success Yearly Trend

Line Chart:



- The success rate starts to increase from 2013 with a dip in 2018 and 2020. Generally, the success rate is increasing over the years.

All Launch Site Names



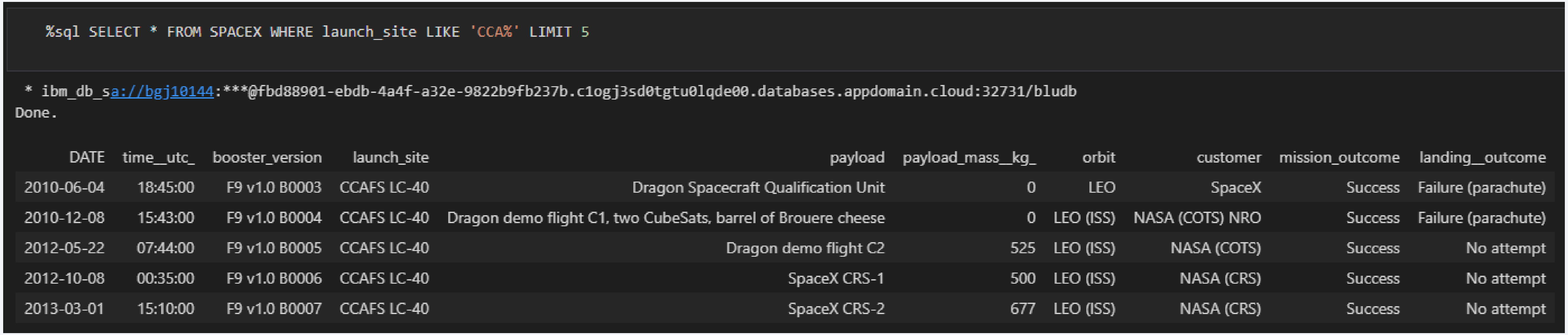
```
cur = %sql SELECT distinct(launch_site) FROM SPACEX
cur
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9
Done.
```

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

- List of all unique launch site. There are only 4 launch sites.

Launch Site Names Begin with 'CCA'



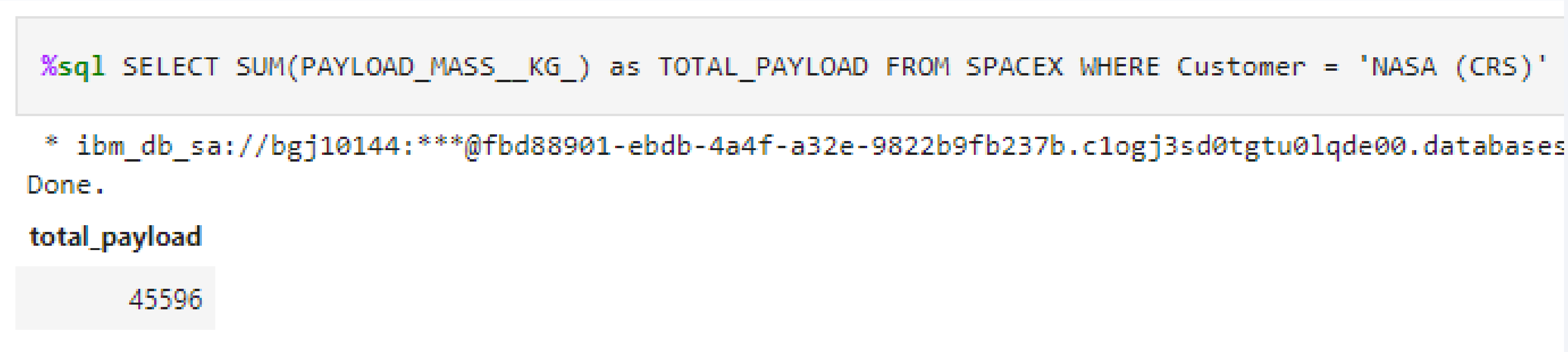
```
%sql SELECT * FROM SPACEX WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	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

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

Total Payload Mass



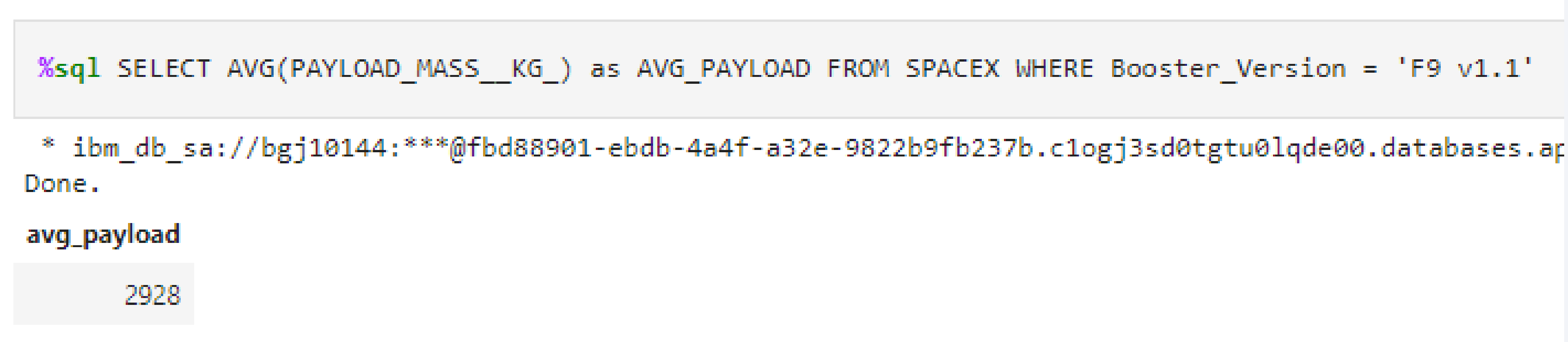
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD FROM SPACEX WHERE Customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases
Done.
```

total_payload
45596

- The total payload mass launched by NASA is 45596 KG

Average Payload Mass by F9 v1.1



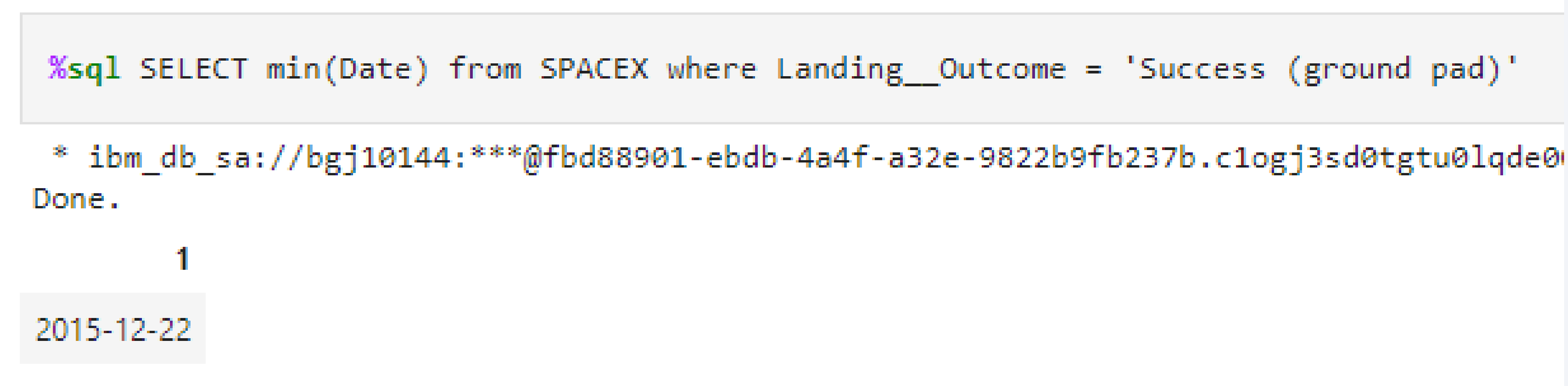
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as AVG_PAYLOAD FROM SPACEX WHERE Booster_Version = 'F9 v1.1'
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.ap
Done.
```

avg_payload
2928

- The average payload mass carried by booster version F9 v1.1 is 2928 KG

First Successful Ground Landing Date



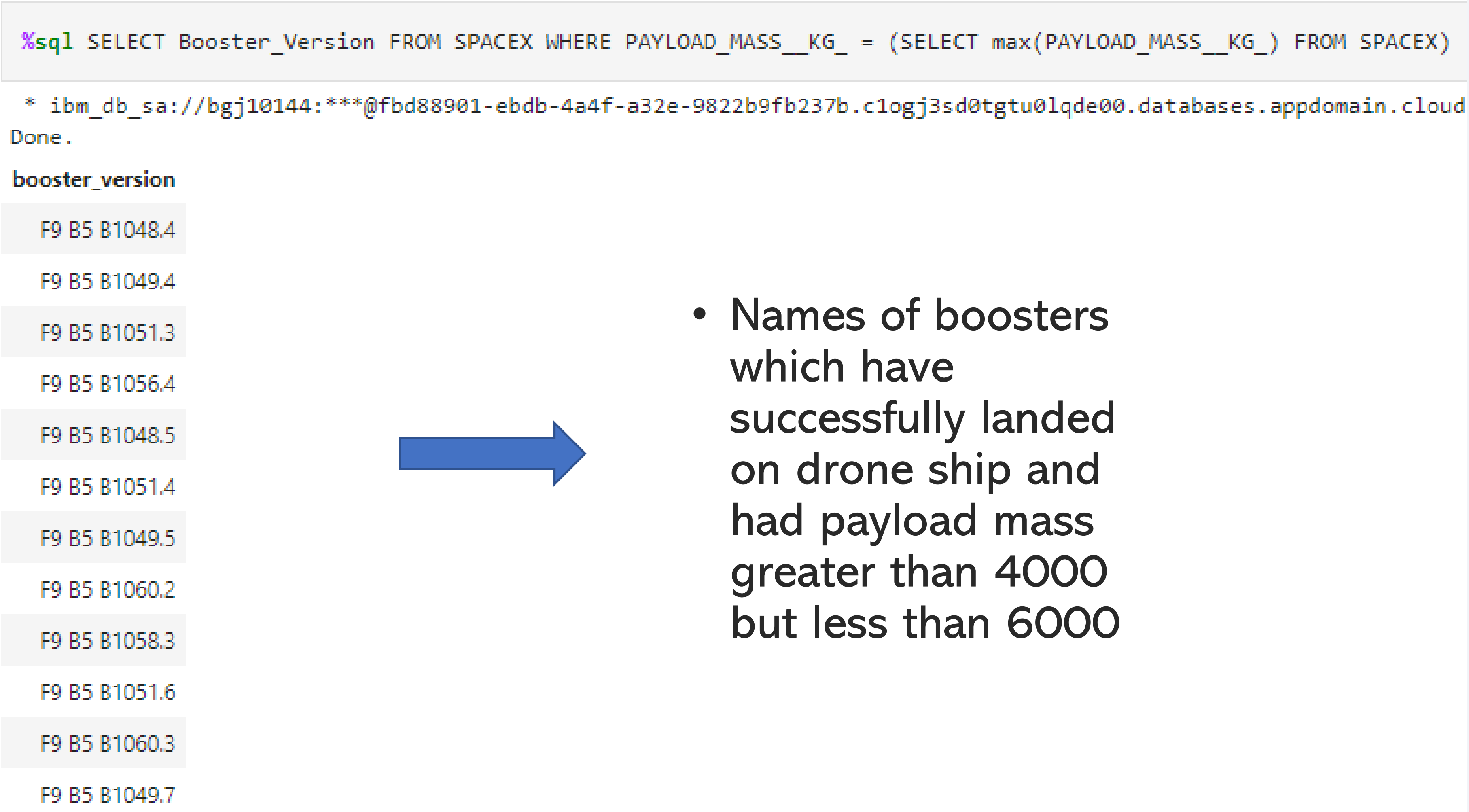
```
%sql SELECT min(Date) from SPACEX where Landing__Outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde0
Done.
```

1
2015-12-22

- The first successful landing outcome on ground pad occurred on 22 Dec 2015

Successful Drone Ship Landing with Payload between 4000 and 6000



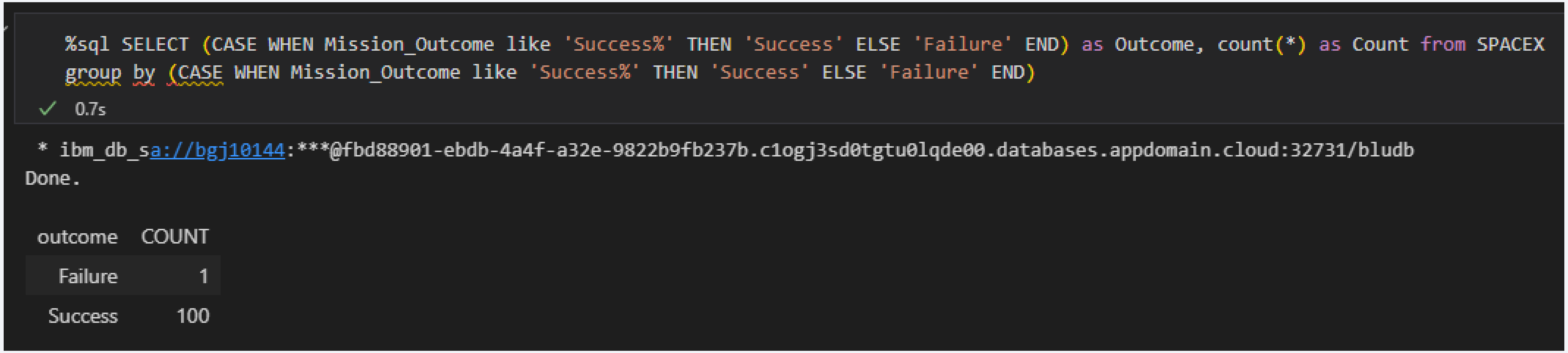
```
%sql SELECT Booster_Version FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEX)
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
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

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



```
%sql SELECT (CASE WHEN Mission_Outcome like 'Success%' THEN 'Success' ELSE 'Failure' END) as Outcome, count(*) as Count from SPACEX
group by (CASE WHEN Mission_Outcome like 'Success%' THEN 'Success' ELSE 'Failure' END)
```

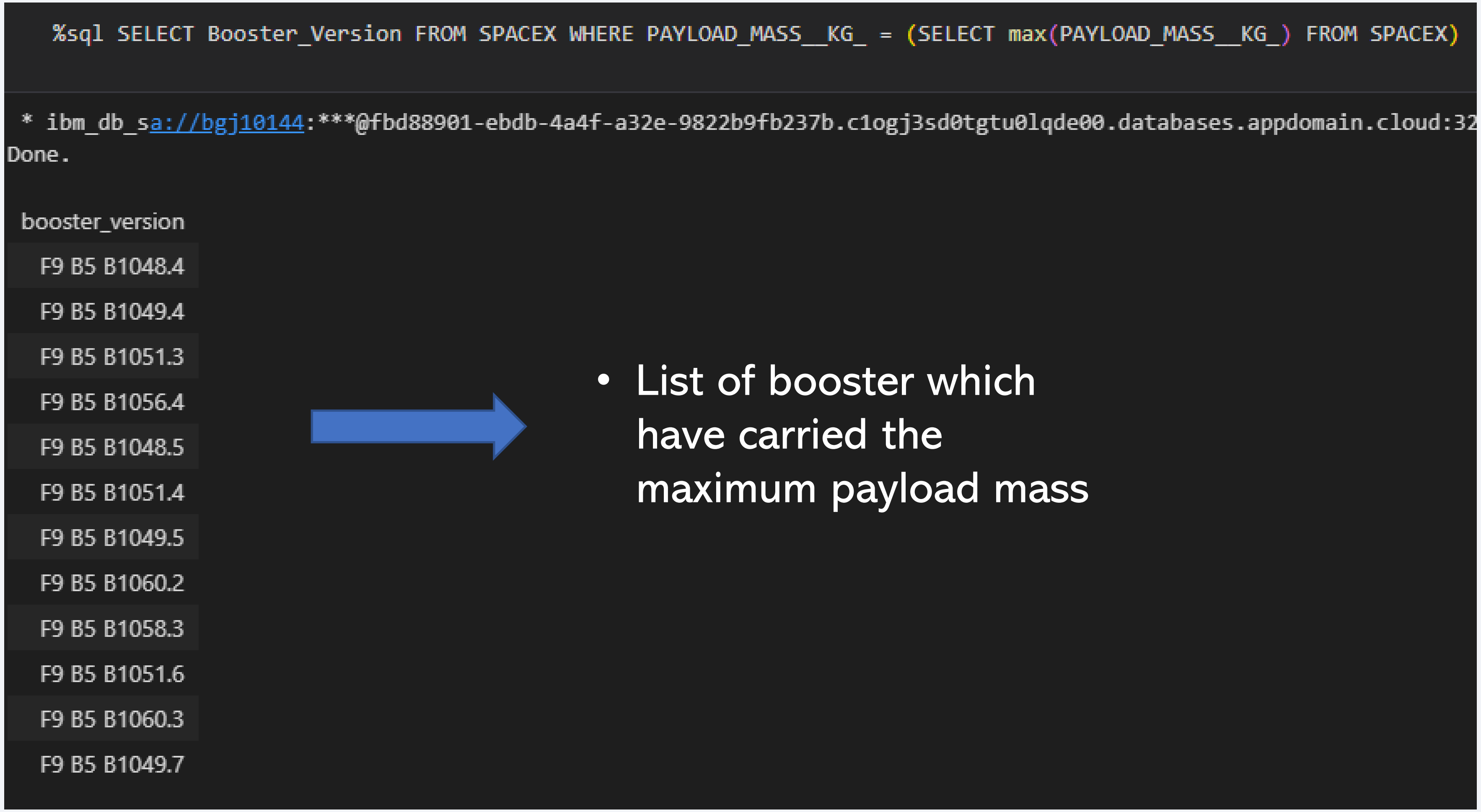
✓ 0.7s

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

outcome	COUNT
Failure	1
Success	100

- There are 1 failure and 100 success outcomes

Boosters Carried Maximum Payload



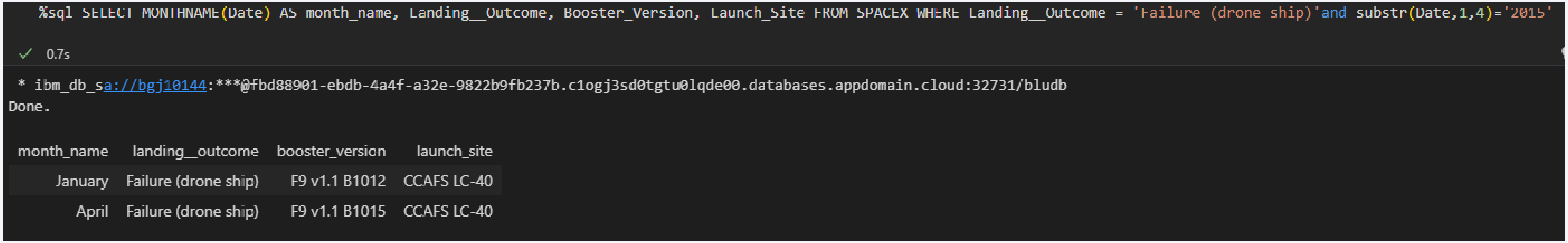
```
%sql SELECT Booster_Version FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEX)
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32
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

- List of booster which have carried the maximum payload mass

2015 Launch Records



```
%sql SELECT MONTHNAME(Date) AS month_name, Landing__Outcome, Booster_Version, Launch_Site FROM SPACEX WHERE Landing__Outcome = 'Failure (drone ship)'and substr(Date,1,4)='2015'
```

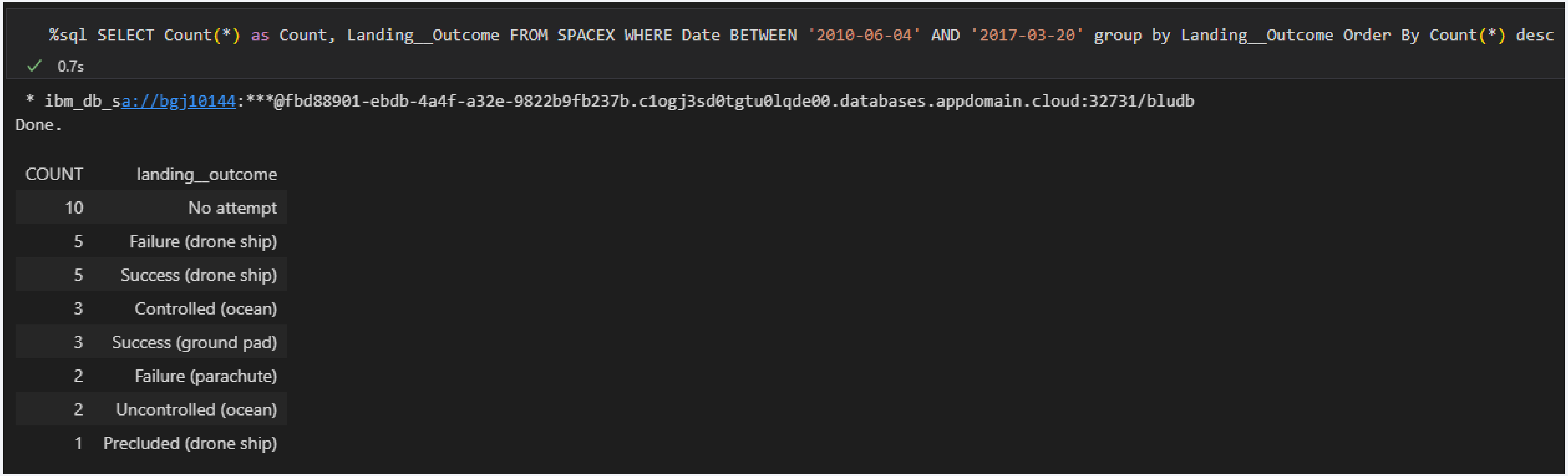
✓ 0.7s

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

month_name	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There are 2 failed landing outcomes in drone ship in year 2015

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



```
%sql SELECT Count(*) as Count, Landing__Outcome FROM SPACEX WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' group by Landing__Outcome Order By Count(*) desc
```

```
 * ibm_db_sa://bgj10144:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

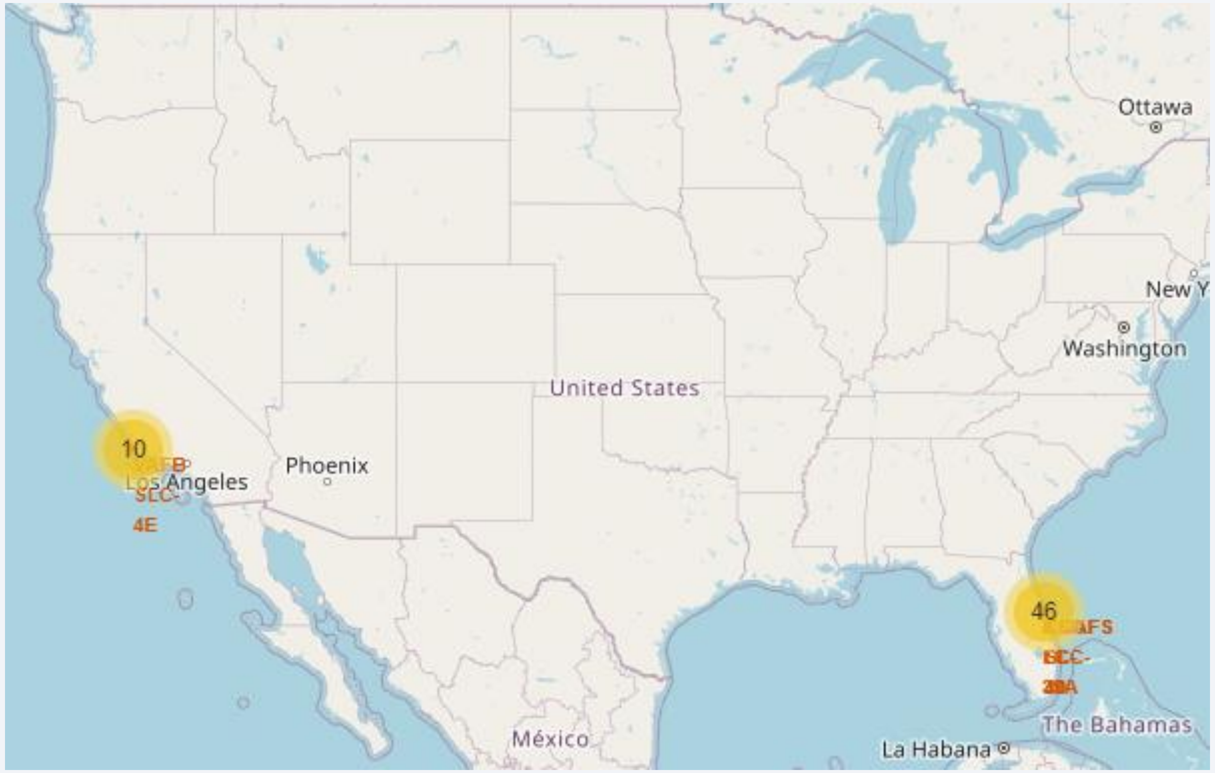
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

- The count of the different landing outcomes was calculated and Failure (drone ship) and Success (drone ship) has the highest count.

Section 3

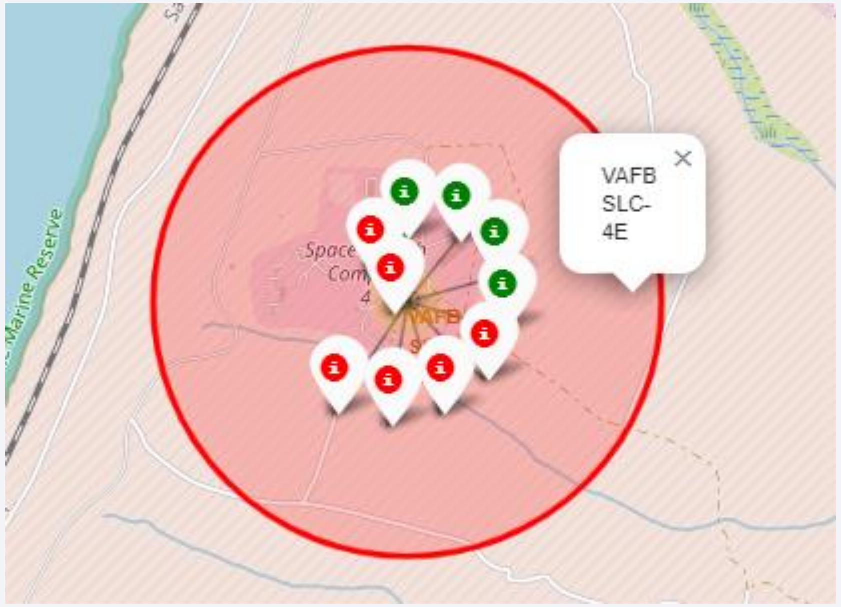
Launch Sites Proximities Analysis

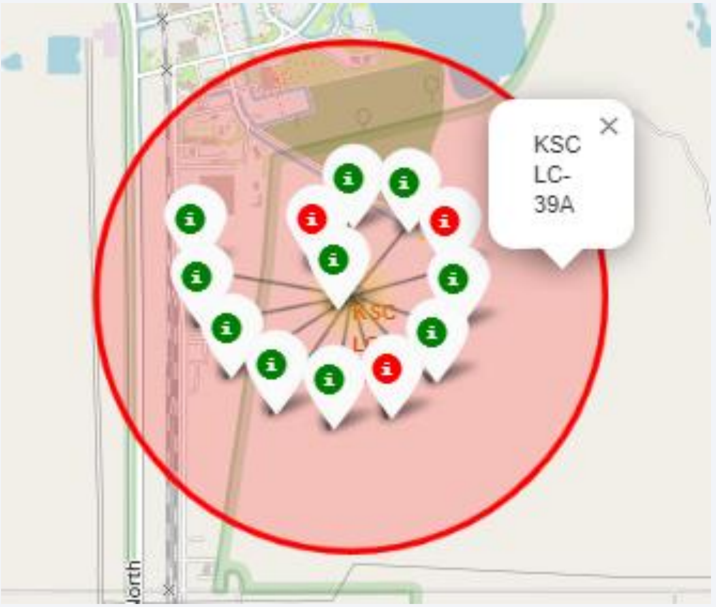
Folium Map Overall SpaceX Geolocation

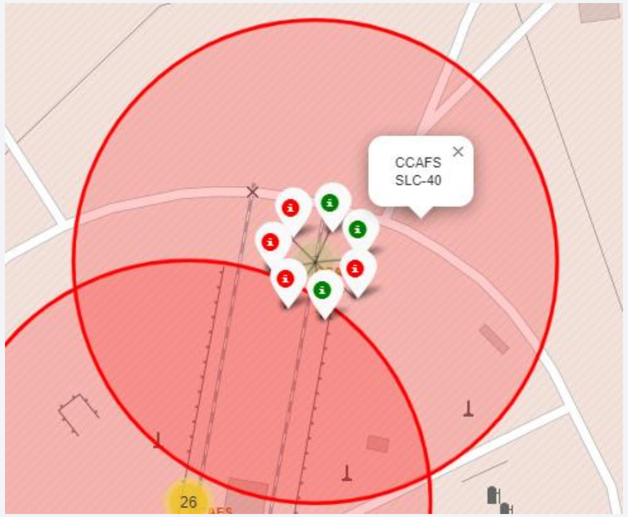


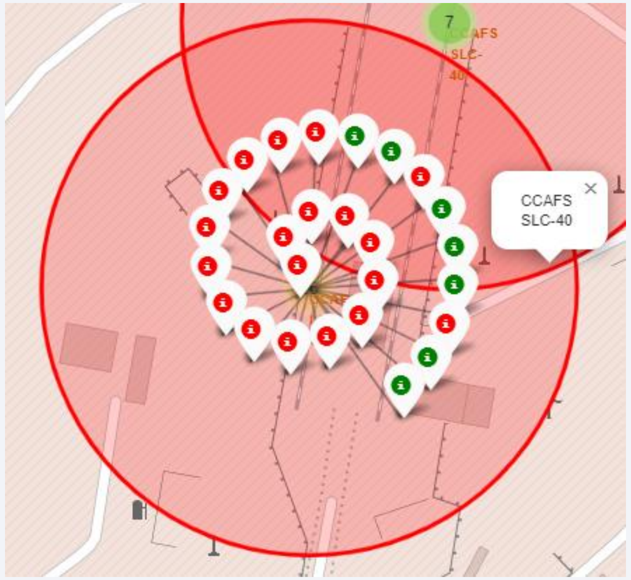
- The launch sites are located near to coast lines

Folium Map – Launch site success rate



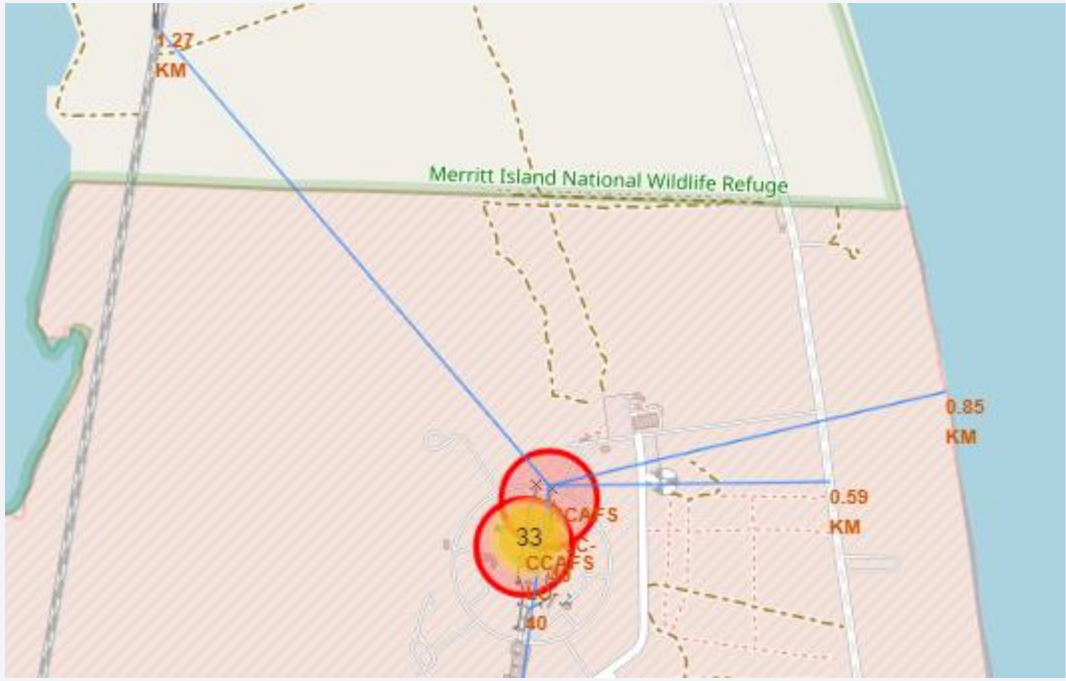


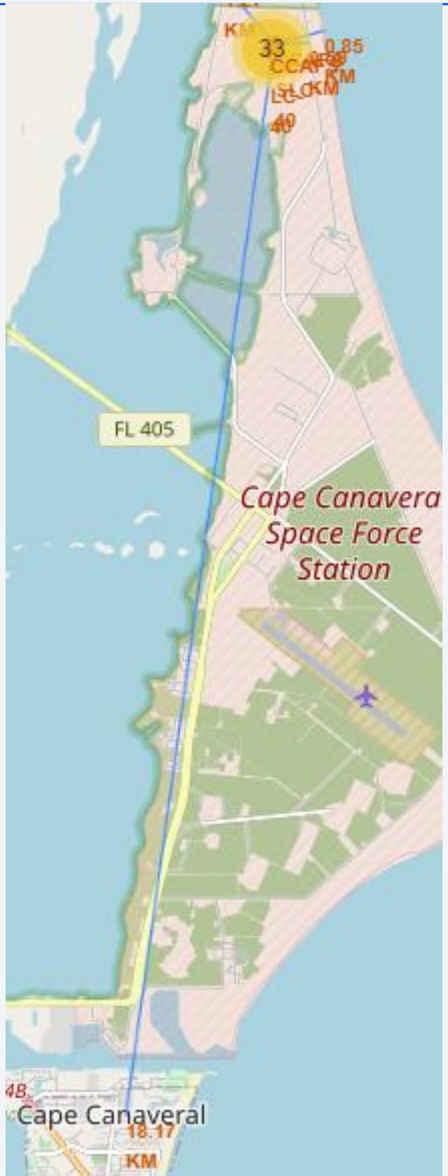




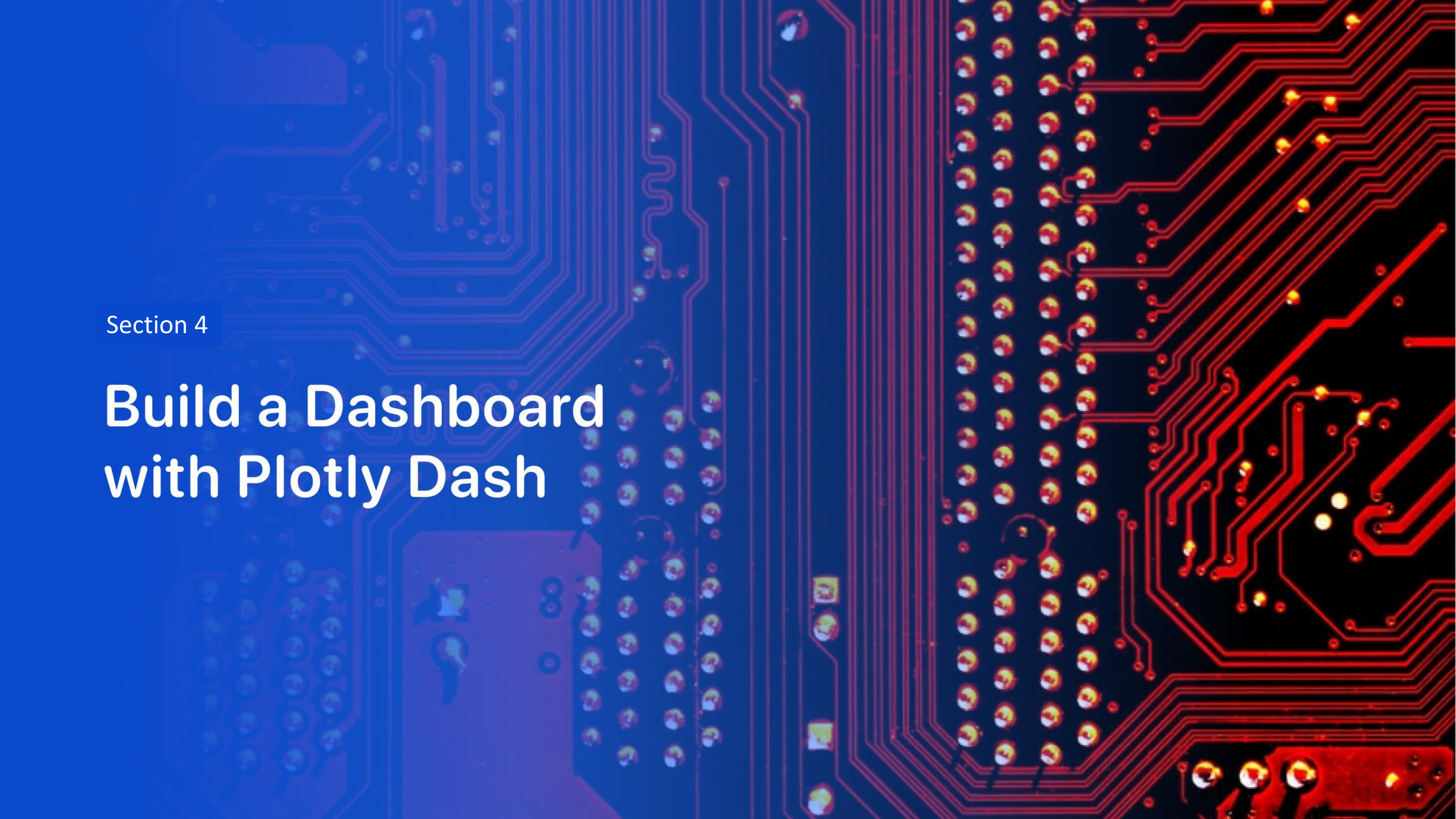
- KSC LC-39A has the highest success rate
- CCAFS SLC-40 has the most number of launches

Folium Map – Proximities of launch site





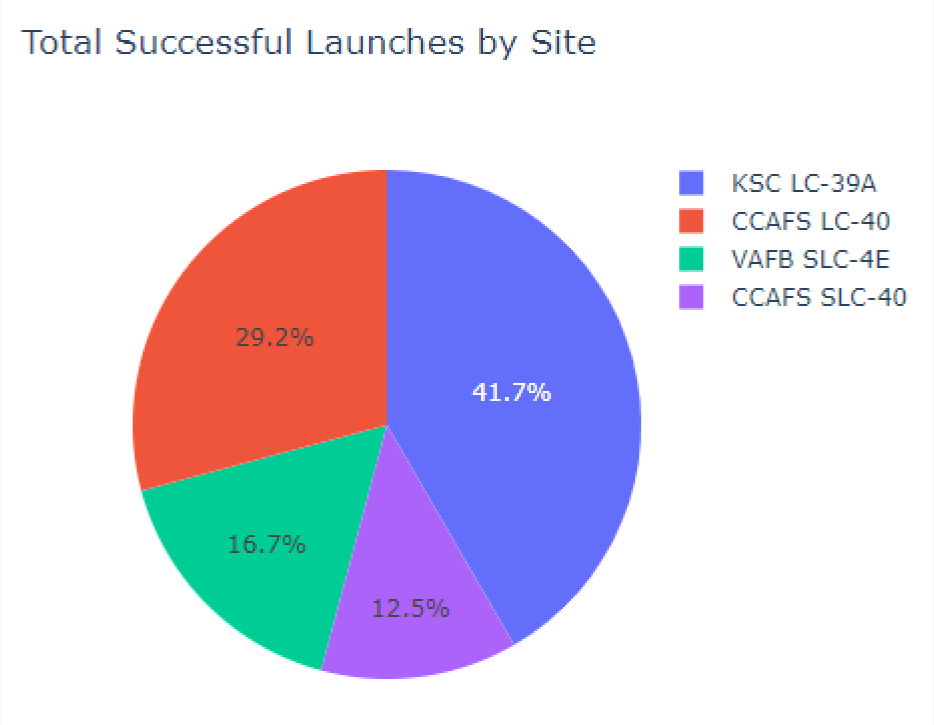
The launch site (CCAFS SLC-40) is close to railways, highways and coastline, but far away from cities



Section 4

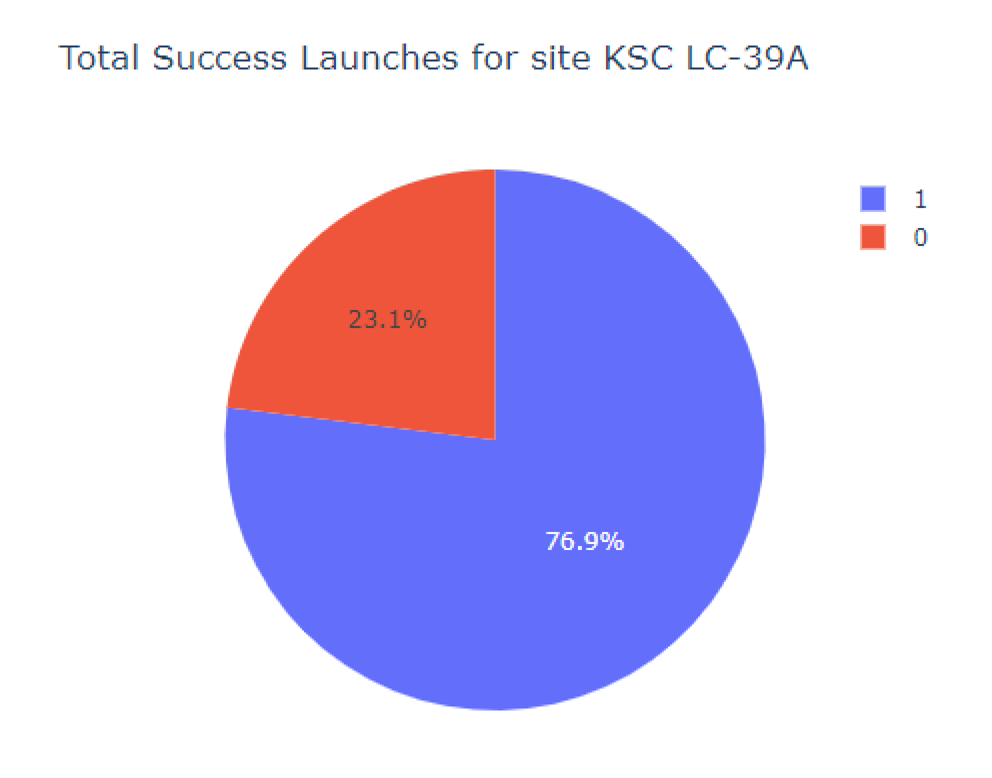
Build a Dashboard with Plotly Dash

Successful Launches by Site



- KSC LC-39A accounts for the highest number of successful launches and CCAFS SLC-40 has the lowest number

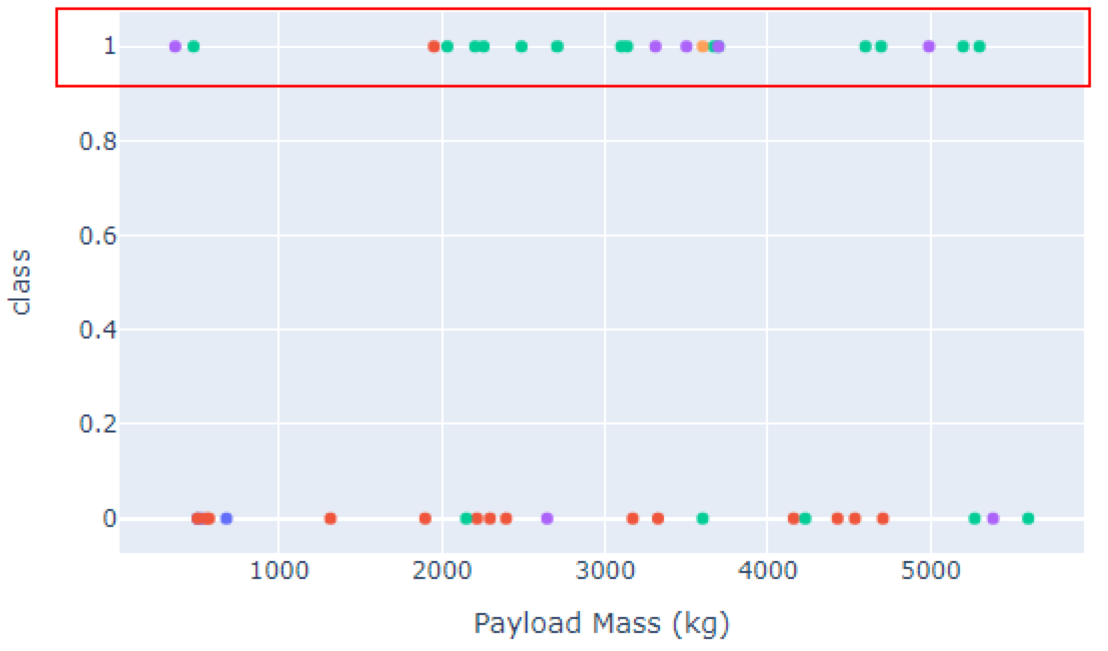
Success rate for KSC LC-39A



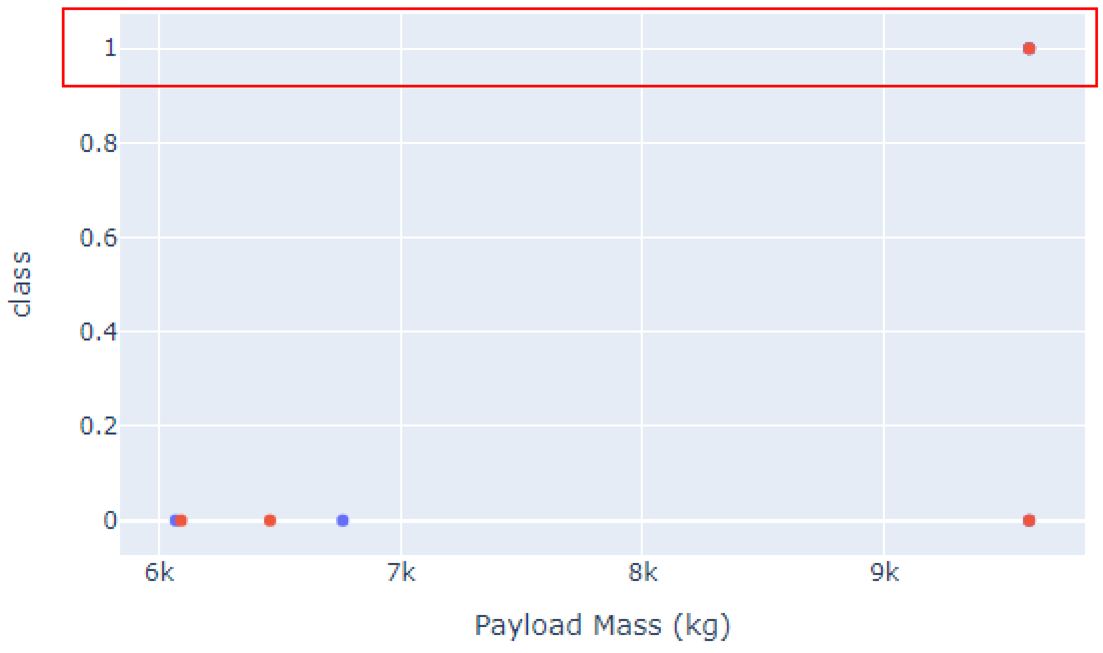
- KSC LC-39A has a success rate of 76.9%

Distribution of success/failures across payload mass

Payload: 0 to 6000 KG



Payload: 6000 to 10000 KG

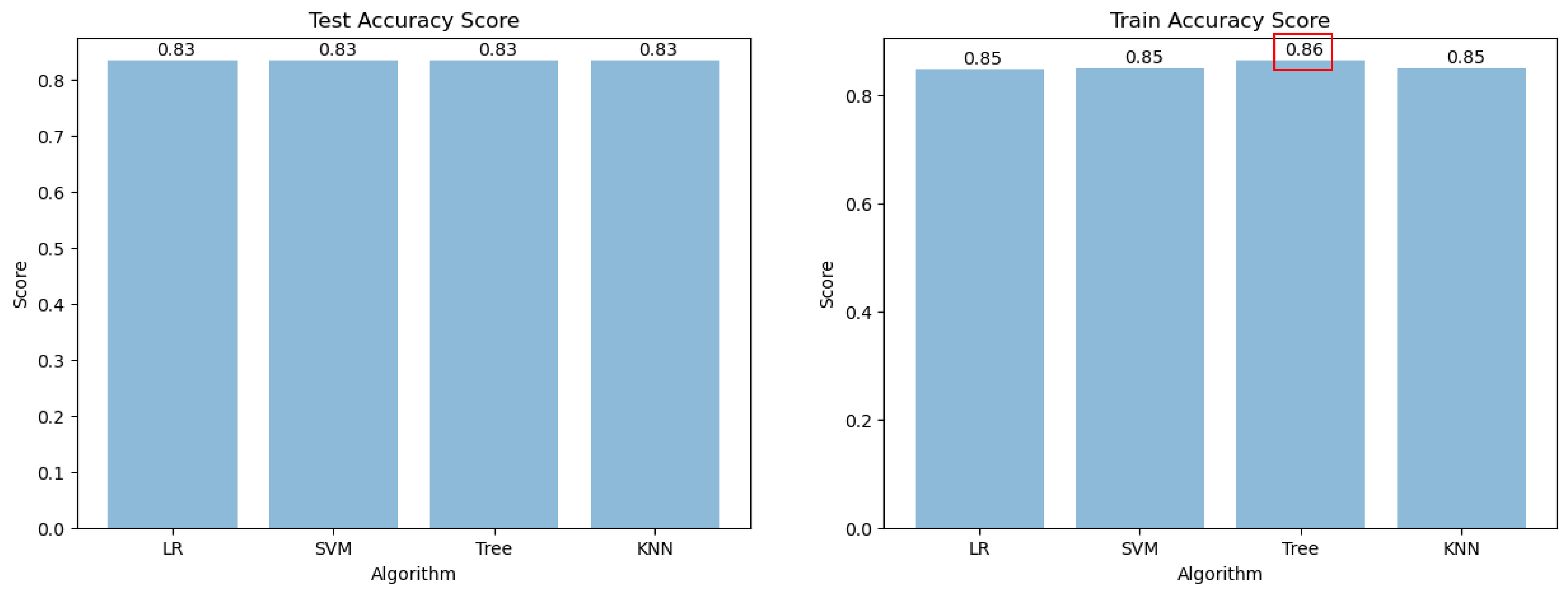


There are only a few launches with payload above 6000 KG.
There seems to be a higher occurrences of successful landing at payload below 6000 KG.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

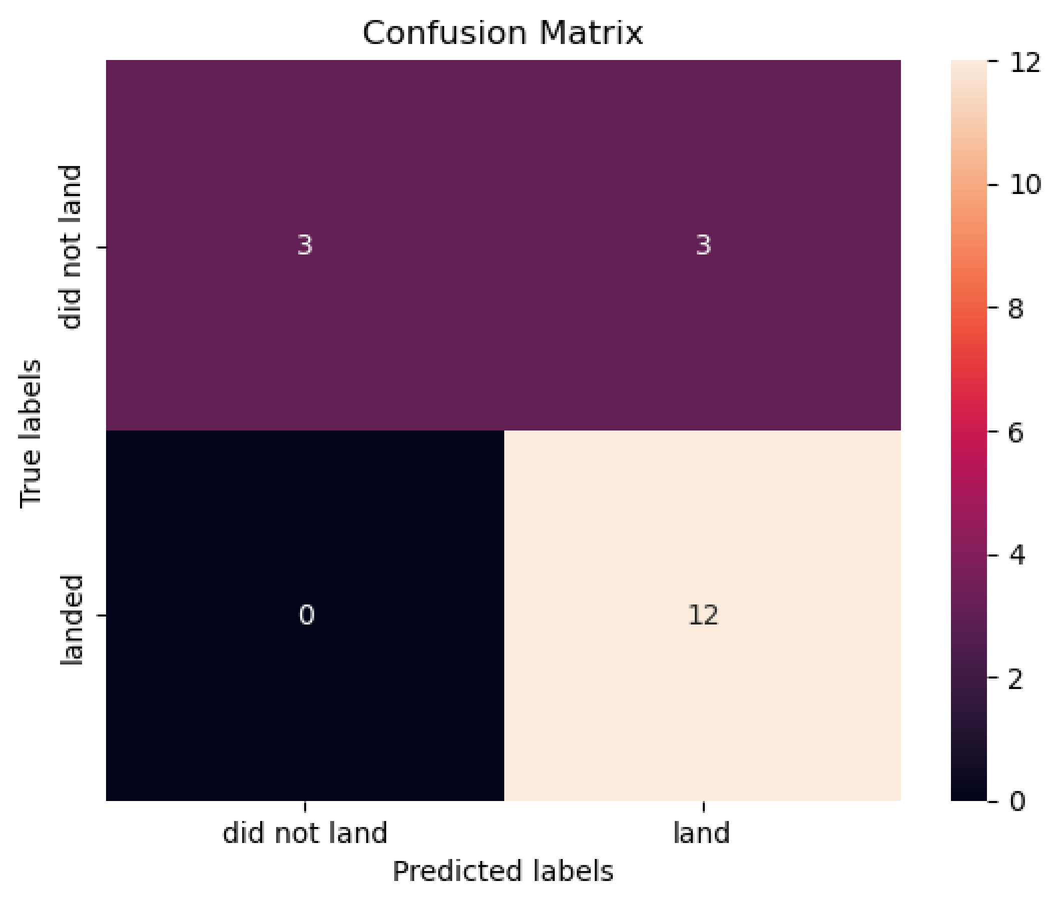


From the bar charts above, we can see that all models have the same test accuracy and decision tree model has a slight edge in scoring 0.01 higher than the other models for train accuracy.

All models performed very similarly and the best model would be decision tree.

Confusion Matrix

Confusion Matrix for Decision Tree Model



The model is able to distinguish between the different classes with some risk of having false positives

Conclusions

- There is a general trend that the success rate increases with increasing flight number and over years. This is probably due to the advancement in technology and methods over the years based on lessons learnt from the previous flights.
- Generally lower payload mass also contributes to higher success rate, probably due to an easier rocket control with lower overall mass. However, ISS has a higher success rate at higher payload mass.
- Among the Orbits with higher number of launches (>5), VLEO has the highest success rate and GTO has the lowest success rate. Further study on VLEO can be performed to deep dive on the success factors.
- There is a preference to locate launch sites near coast lines and away from the cities so that sea landing could be planned easier and for the safety purposes.
- All classification models performed very similarly with Decision Tree performing slightly better for train accuracy. Hence Decision Tree model could be adopted.

Thank you!