IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Gonzalez
20-05-2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling- Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- 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

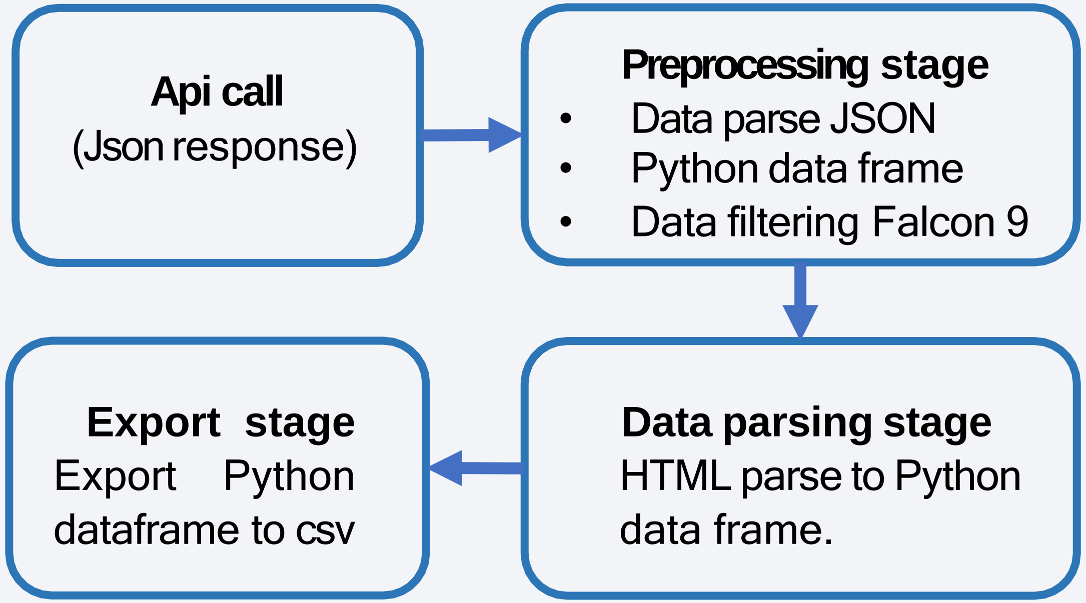
- The data was collected using various methods.
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis

Data Collection – SpaceX API

- The outcome of the data collection was a data frame with critical features that allowed further analysis which included date, longitude, latitude, launch site, and outcome.

- The link to the notebook is [here in Github]()

SpaceX API Data Collection general schema



Api call
(Json response)

Preprocessing stage
- Data parse JSON
- Python data frame
- Data filtering Falcon 9

Data parsing stage
HTML parse to Python data frame.

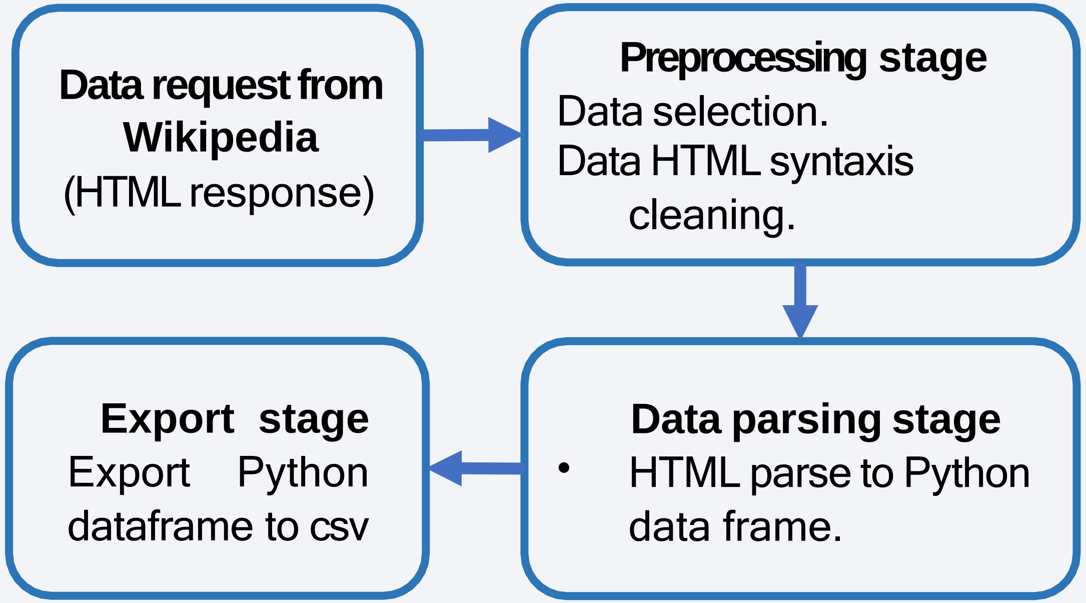
Export stage
Export Python dataframe to csv

Data Collection - Scraping

- The recollecting data process from Wikipedia required several lines of code for a correct noise-cleaning.

- The notebook with the web scraping data collection process can be found here (here in Github).

Web scraping Data Collection general schema



Data request from Wikipedia
(HTML response)

Preprocessing stage
Data selection.
Data HTML syntaxis cleaning.

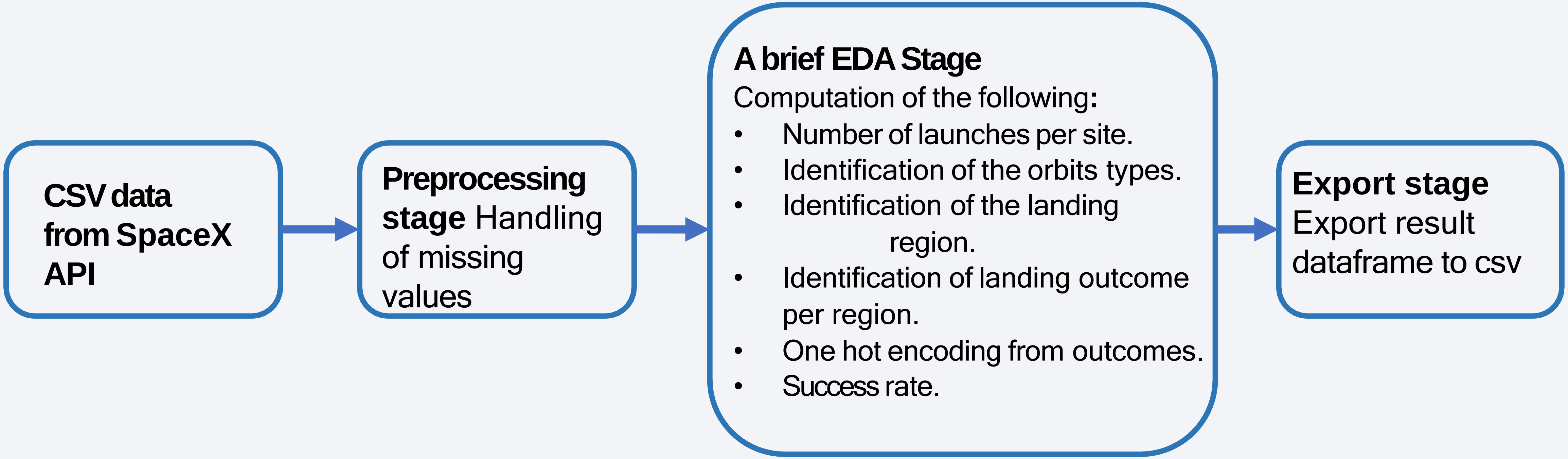
Data parsing stage
- HTML parse to Python data frame.

Export stage
Export Python dataframe to csv

Data Wrangling

- Given the csv data from the SpaceX API data collection, analyses to launch facilities, orbits, and mission outcomes features were made.
- The aim goal of the analysis was to provide accurate labeling of the outcomes in a 1/0 fashion, where 1 means the booster successfully landed, and 0 meaning unsuccessful landing.
- Also, from the analysis was possible to identify the number of launches per site, the name and number of orbits in which the launch was made, and the types of possible outcomes for a Falcon 9 mission.
- A 66% of success rate was calculated as well.
- The link to the notebook is here in Github

Data Wrangling



CSV data from SpaceX API

Preprocessing stage Handling of missing values

A brief EDA Stage
Computation of the following**:**

- Number of launches per site.
- Identification of the orbits types.
- Identification of the landing region.
- Identification of landing outcome per region.
- One hot encoding from outcomes.
- Success rate.

Export stage
Export result dataframe to csv

EDA with Data Visualization

- A total of seven charts were plotted as part of the Exploratory Data Analysis: five scatter plots, one line plot, and one bar plot.
- Generally speaking, the graph type selected for each feature relationship analysis was directly related to the necessity of **understanding the connection** among **three or two features**.
- The scatter plots give us an insight of the relationship among the **landing outcome** (1/0 feature) and **the duos** (Payload Mass, Flight Number), (Flight Number, Launch Site), (Payload Mass, Launch Site), (Flight Number, Orbit), and (Payload Mass, Orbit).
  - The bar plot lets us analyze the relationship between success rate and orbit type. Lastly, a line chart shows the relationship between year and average success rate.
- The EDA with Data Visualization notebook can be found here in Github

EDA with SQL

- Further Exploratory Data Analyses were made by means of SQL-queries in DB2 to a Falcon 9 data set stored in IBM DB2.
- The queries results show us the following information.
  - The names of the unique launch sites in the space mission.
  - Showed 5 records where launch sites begin with the string 'CCA'.
  - The total payload mass carried by boosters launched by NASA (CRS).
  - The average payload mass carried by booster version F9 v1.1.
  - The date when the first successful landing outcome in ground pad was achieved.
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - The total number of successful and failure mission outcomes

- The EDA with SQL notebook can be found here in Github

Build an Interactive Map with Folium

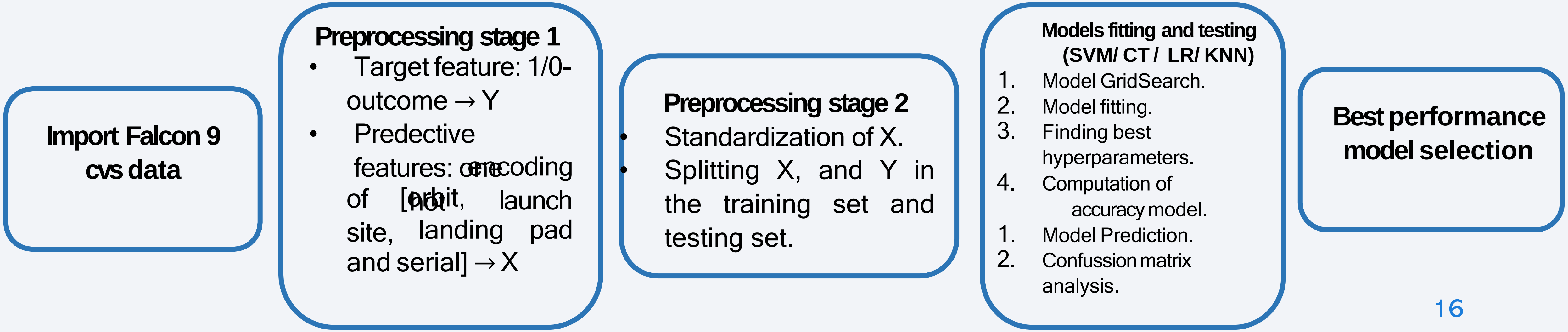
- The main goal plotting Falcon 9 data with Folium was identify launch sites, success and failed launches for site, and calculate the distances between launch sites and its proximities.
- A Folium-object marker allowed us to identify launch sites, a Folium-object circle made it possible to see the count of successful and failed launches in situ, and lines show us a visual representation between the launch sites and proximities as railways, cities, etc.
- The Folium notebook can be found here in Github

Build a Dashboard with Plotly Dash

- A plotly dashboard with two interactive charts were made: a pie chart and a scatter plot gave us insight between the duos (landing outcome, site) and (payload, launch success), respectively.
- The chart interactivity was achieved through a site dropdown list and a payload range slider, allowing us to obtain data insights quickly and efficiently.
- The Dash notebook can be found here in Github

Predictive Analysis (Classification)

- SVM, classification tree, logistic regression and KNN were tasted in order to found the best 1/0-outcomes feature classifier.
- A standard mythology for fitting and testing the classification models were carried out. The detailed procedure is explained in the following schema.
- The Predictive Analysis notebook can be found here in Github



Import Falcon 9 cvs data

Preprocessing stage 1

- Target feature: 1/0-outcome → Y
- Predective features: one encoding of [hot orbit, launch site, landing pad and serial] → X

Preprocessing stage 2

- Standardization of X.
- Splitting X, and Y in the training set and testing set.

Models fitting and testing (SVM/ CT / LR/ KNN)

1. Model GridSearch.
2. Model fitting.
3. Finding best hyperparameters.
4. Computation of accuracy model.
1. Model Prediction.
2. Confussion matrix analysis.

Best performance model selection

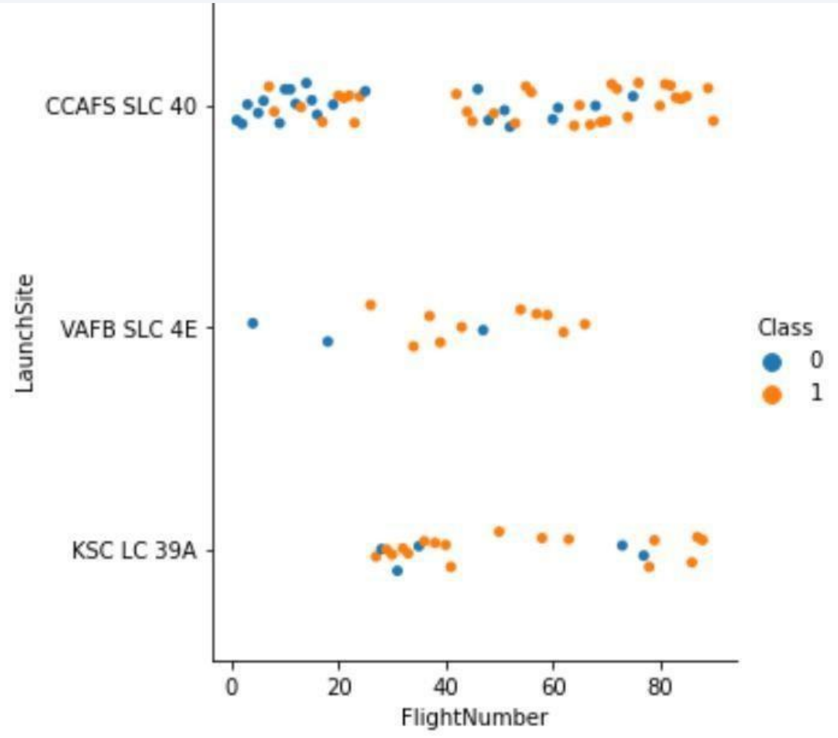
Results

- Broadly speaking, the results of the exploratory data analysis shows the following:
  - There’s a yearly increasing trend in the success landing outcome;
  - There’s a 66% general success rate;
  - Different launch sites have different success rates;
  - VAFB-SLC launch site haven’t rockets launched for heavy payload mass;
  - The orbits ES-L1, GEO, HEO, and SSO have an average success rate of 100%; excluding SO landing sites with a 0% of success rate, generally speaking, all landing sites have an average success greater than 50%.
- The best performance models classifier were the logistic regression, SVM, and KNN, each with a score of 83%.

Section 2

Insights drawn from EDA

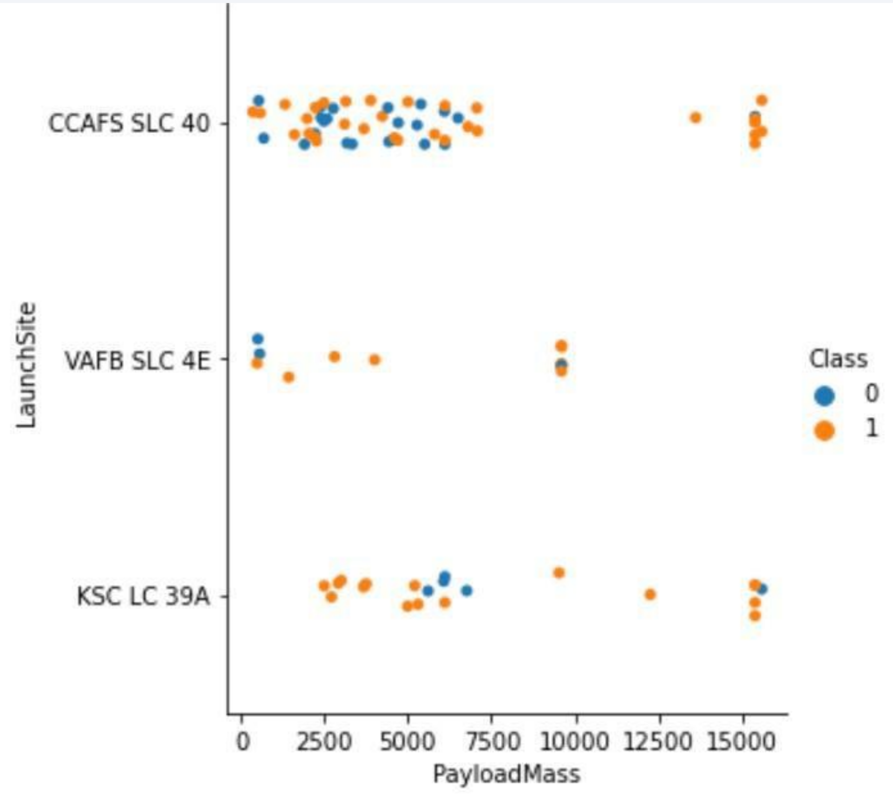
Flight Number vs. Launch Site



- In the case of the Launch Site CCAFS SLC 40 it's possible to successful nota landings rise in the number of corresponding with an increase in the flight number.

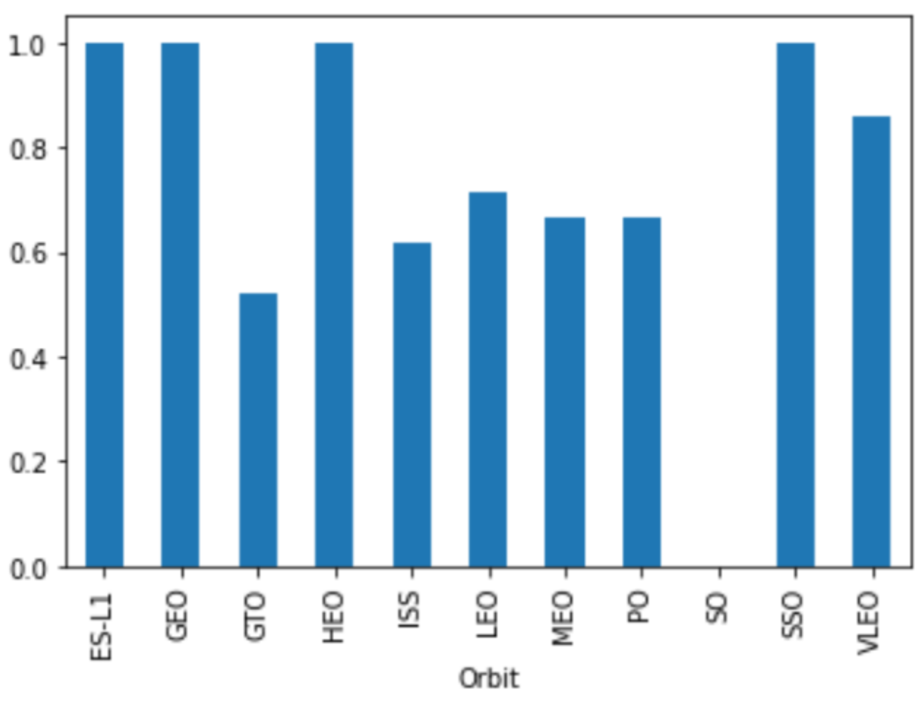
- In contrast, we can notice a general failure outcome in the Launch Sites VAFB SLC 4E and KSC LC 39A.

Payload vs. Launch Site



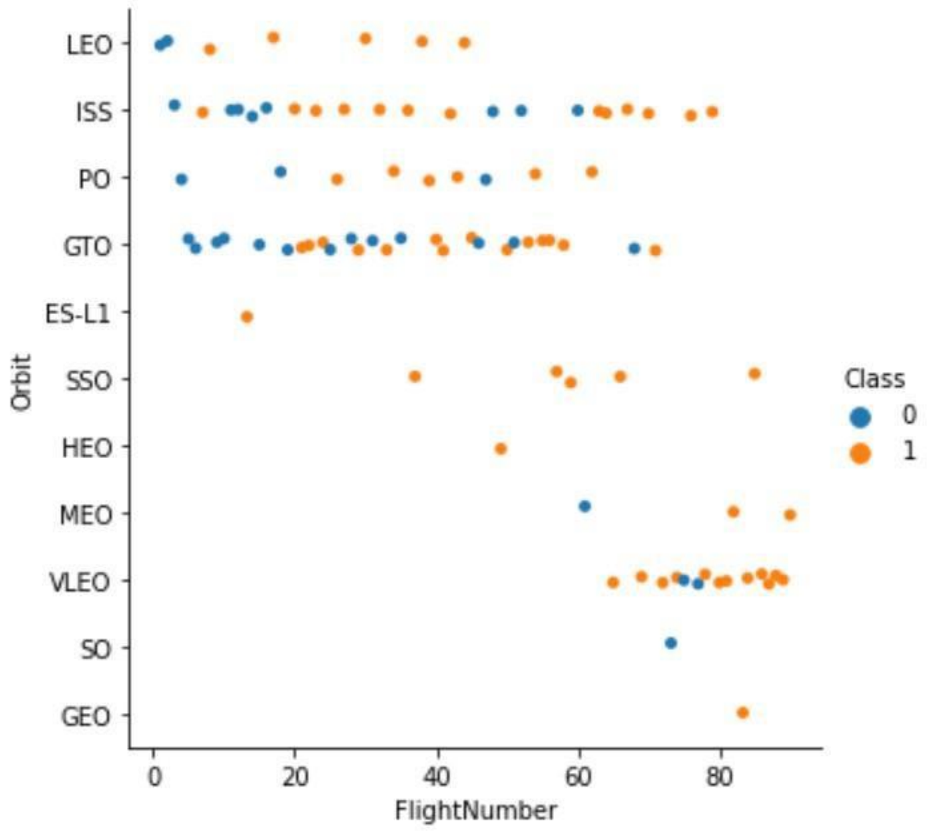
- For the VAFB-SLC 4E launch site there are no rockets launched for heavy pay load mass (greater than 10000)

Success Rate vs. Orbit Type



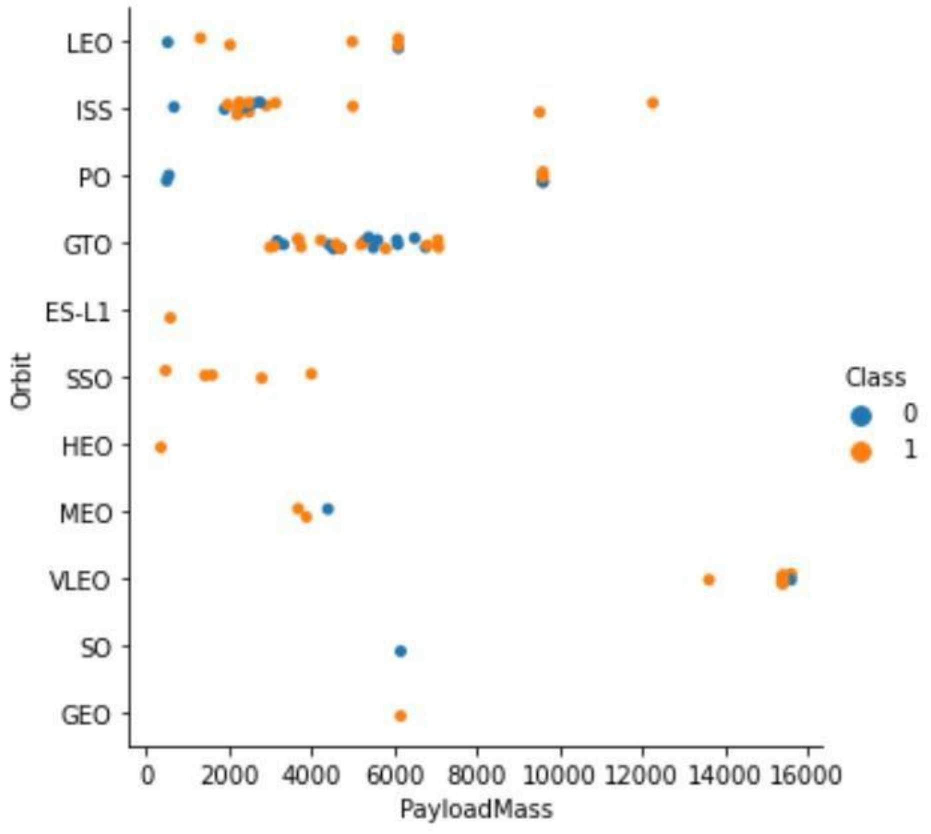
- The orbits ES-L1, GEO, HEO, and SSOhave an average success rate of 100%; excluding SO landing sites with a 0% of success rate, generally speaking, all landing sites have an average success greater than 50%.

Flight Number vs. Orbit Type



- The LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit

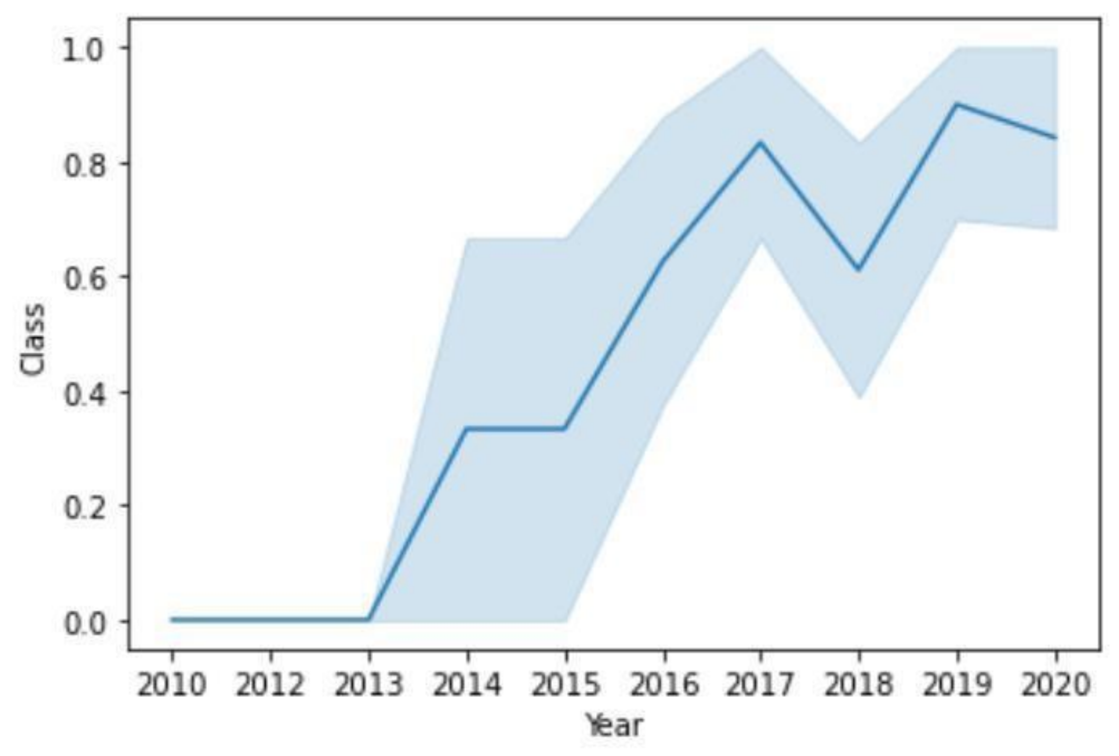
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO, and ISS.

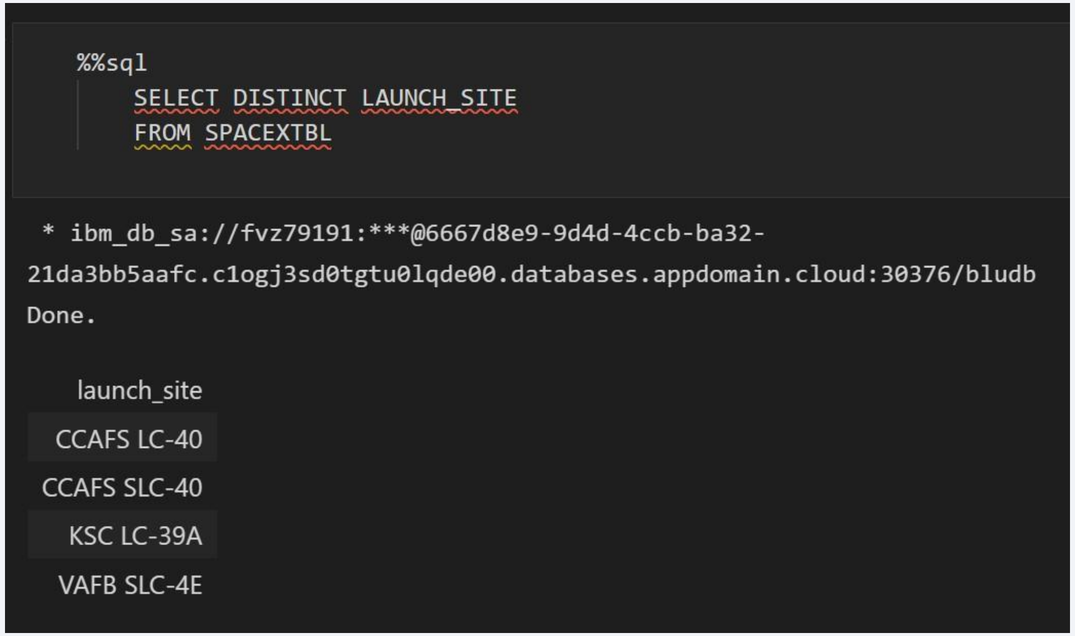
- For GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names



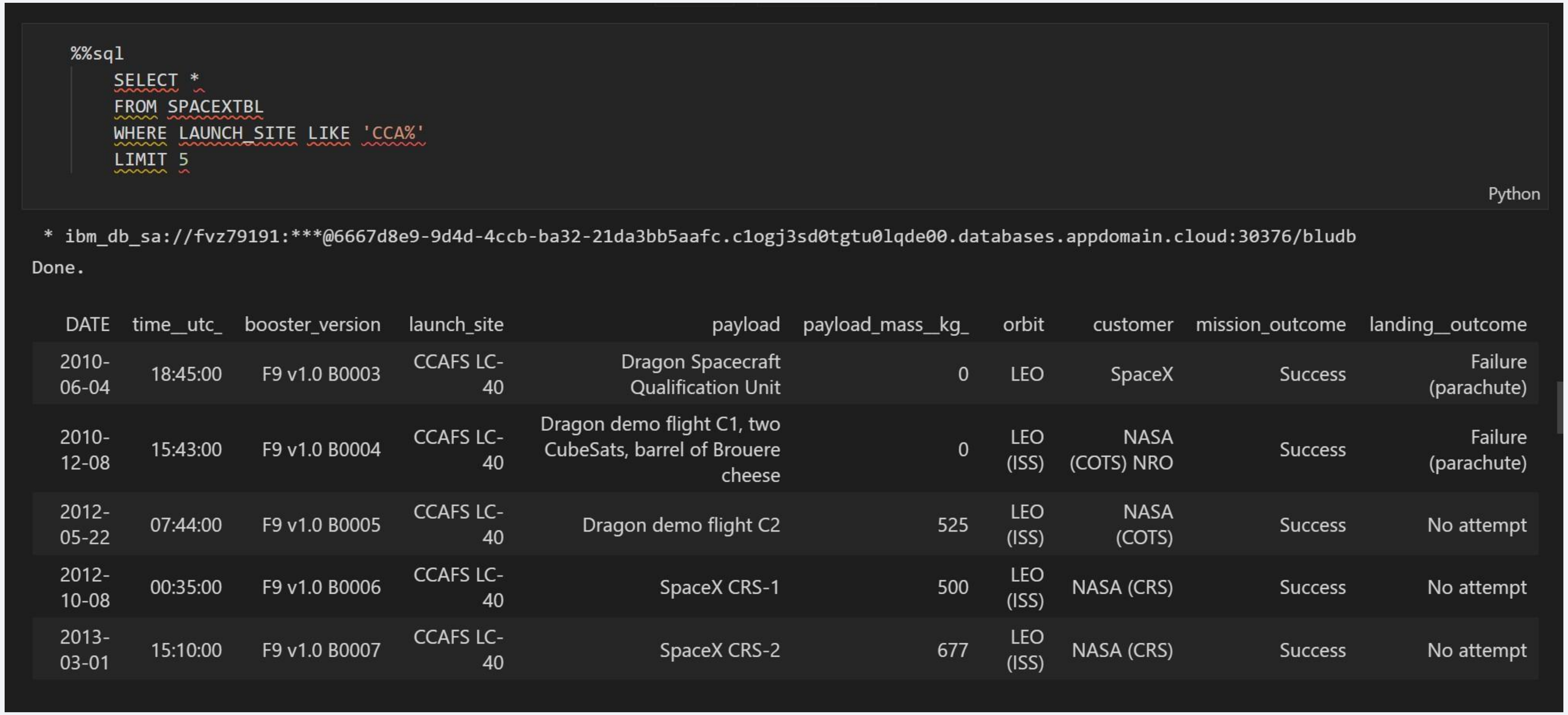
```
%%sql
    SELECT DISTINCT LAUNCH_SITE
    FROM SPACEXTBL
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-
21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

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

- The names of the unique launch sites are CCAFS LC-40, CCAFS SLC- 40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'



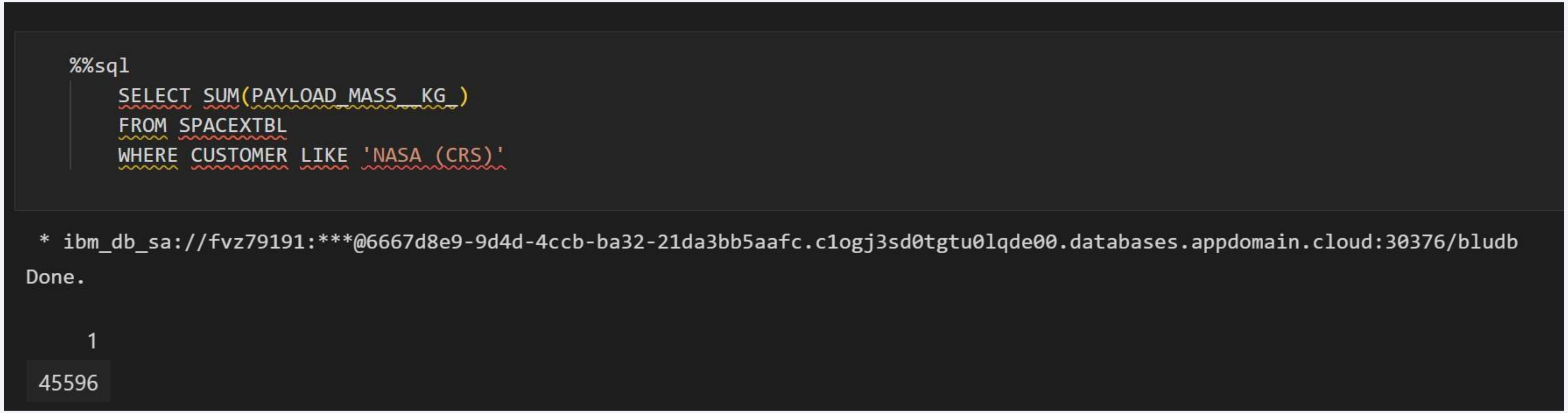
```
%%sql
    SELECT *
    FROM SPACEXTBL
    WHERE LAUNCH_SITE LIKE 'CCA%'
    LIMIT 5
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/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

- The image shows 5 records where launch sites begin with ‘CCA’.

Total Payload Mass



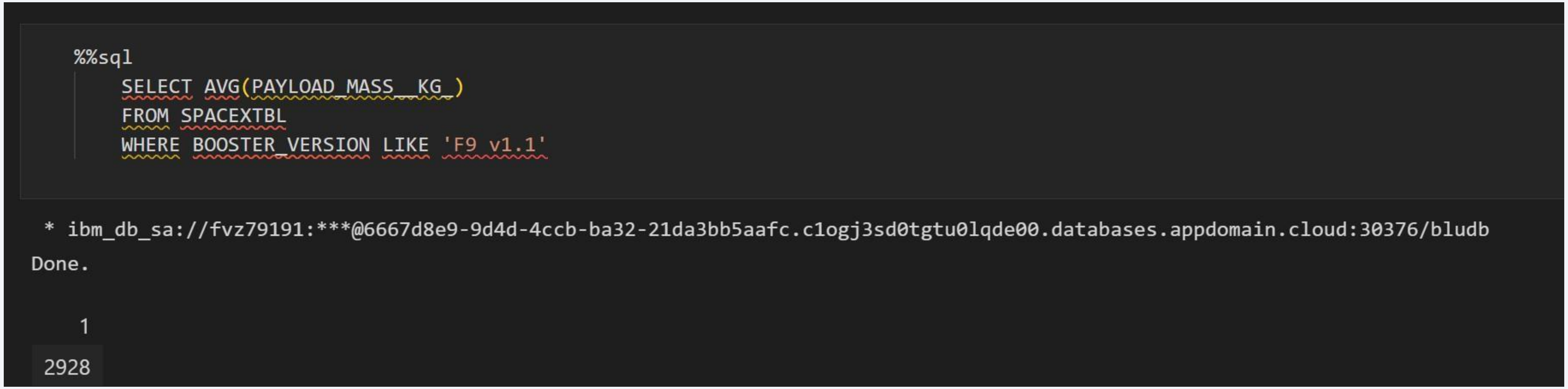
```
%%sql
    SELECT SUM(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
    WHERE CUSTOMER LIKE 'NASA (CRS)'
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

     1
 45596
```

- The total payload carried by boosters from NASA is 45,596

Average Payload Mass by F9 v1.1



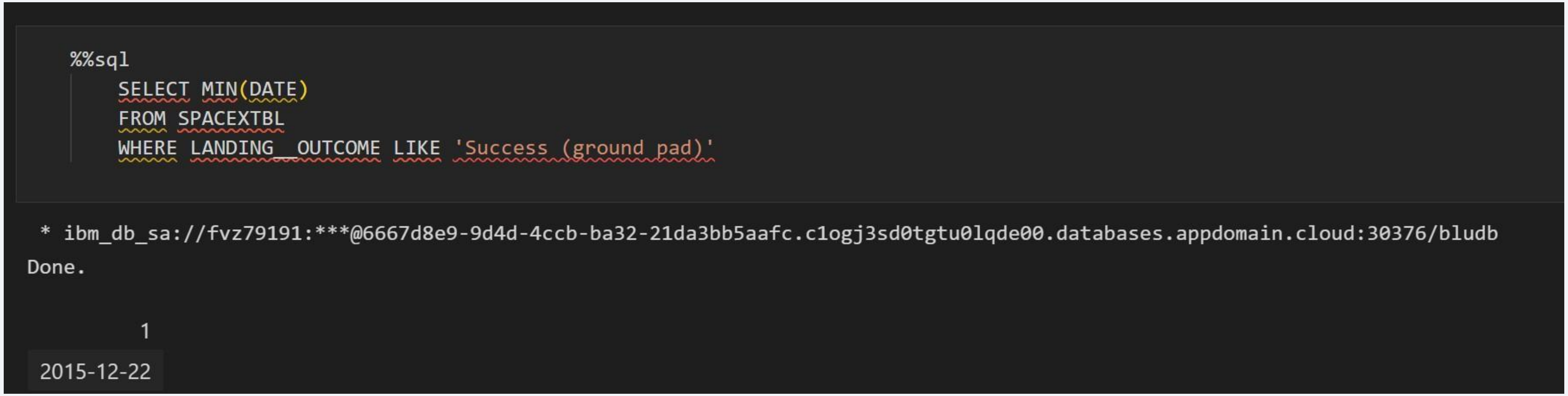
```
%%sql
    SELECT AVG(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
    WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

    1
 2928
```

- The average payload mass carried by booster version F9 v1.1is 2,928 kg

First Successful Ground Landing Date



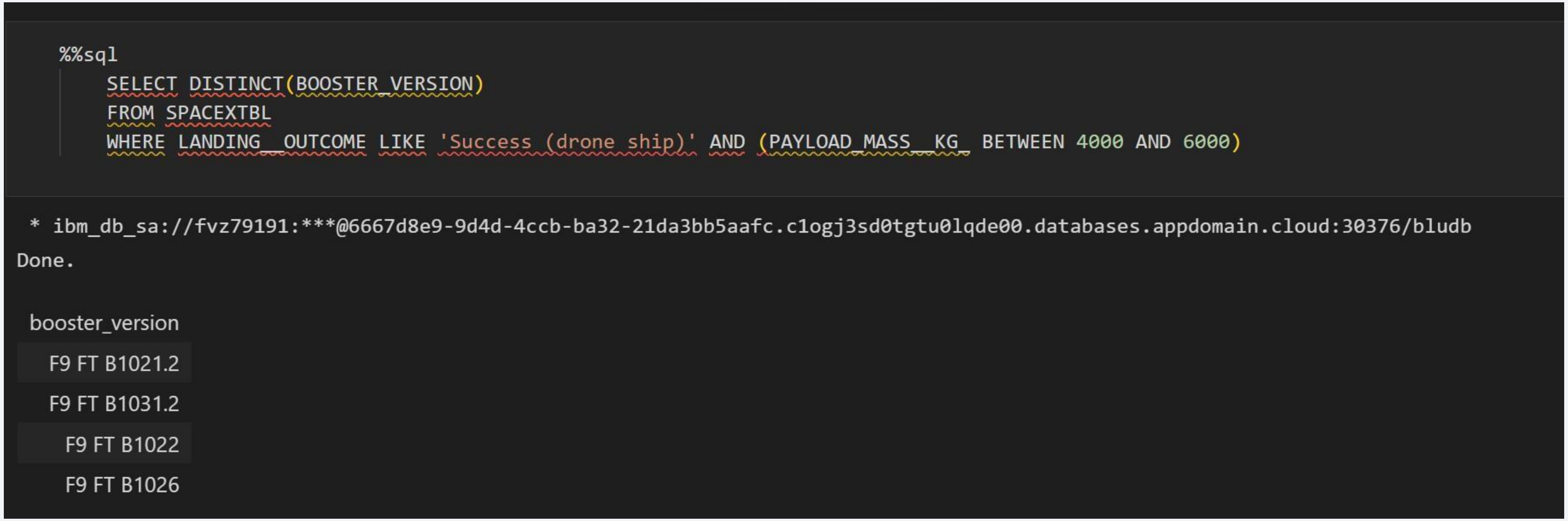
```
%%sql
    SELECT MIN(DATE)
    FROM SPACEXTBL
    WHERE LANDING__OUTCOME LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.

         1
2015-12-22
```

- The date of the first successful landing outcome on ground pad is 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000



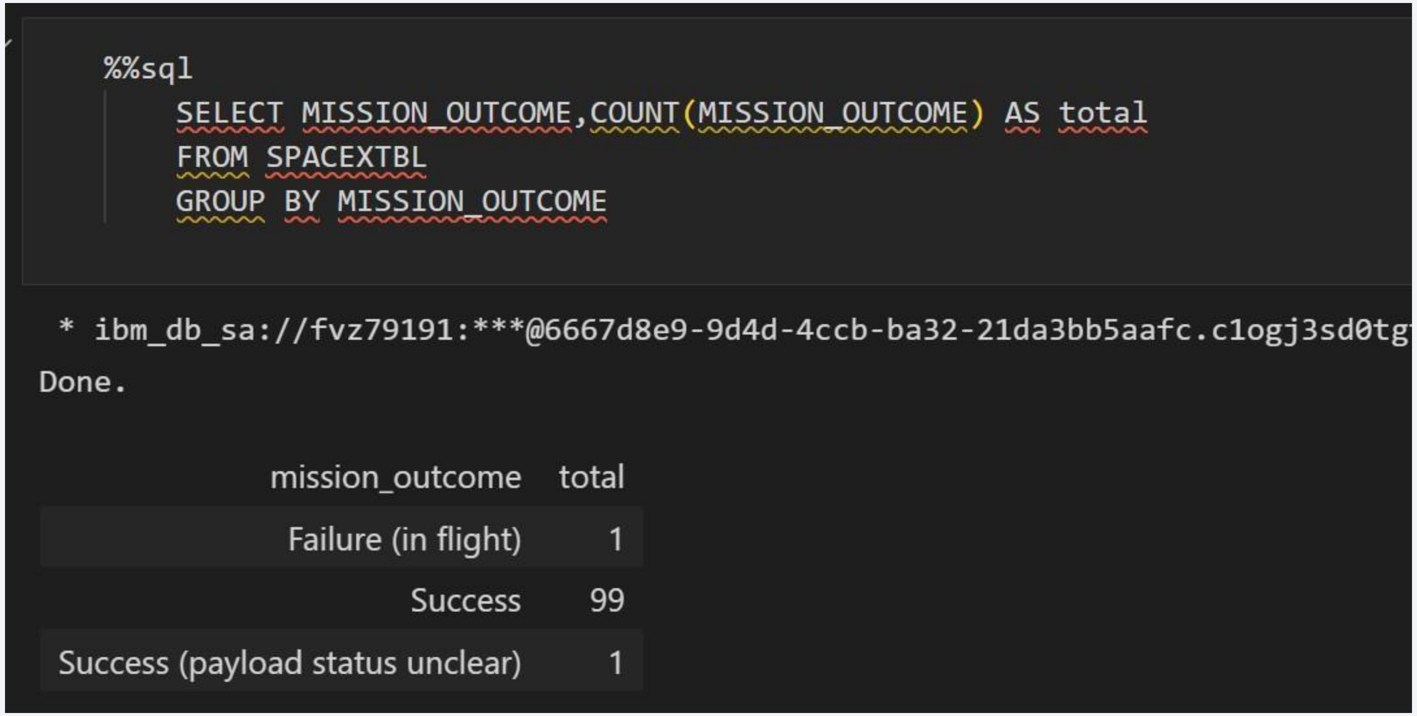
```
%%sql
    SELECT DISTINCT(BOOSTER_VERSION)
    FROM SPACEXTBL
    WHERE LANDING__OUTCOME LIKE 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/bludb
Done.
```

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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, and F9 FT B1026

Total Number of Successful and Failure Mission Outcomes



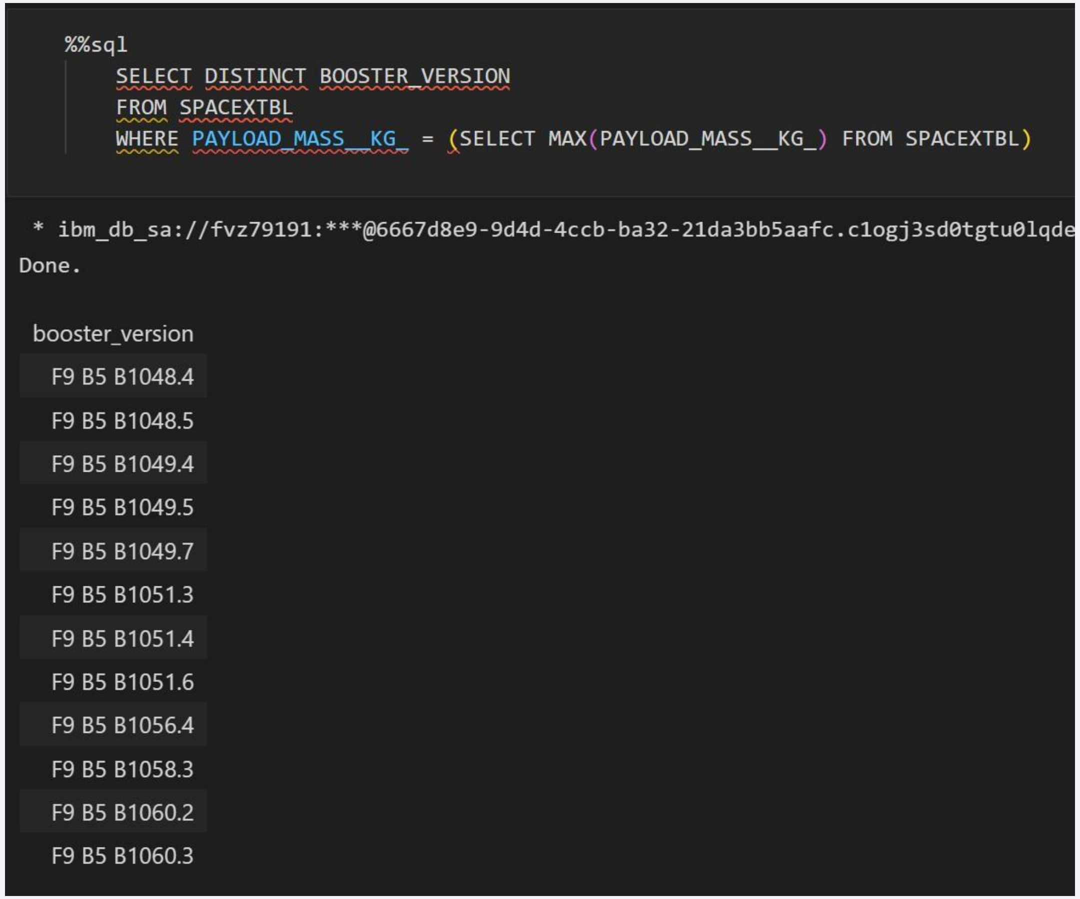
```
%%sql
    SELECT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) AS total
    FROM SPACEXTBL
    GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tg
Done.
```

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

- The total number of successful and failure mission outcomes is the following. Failure 1, success 99, and success (payload status unclear) 1.

Boosters Carried Maximum Payload



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

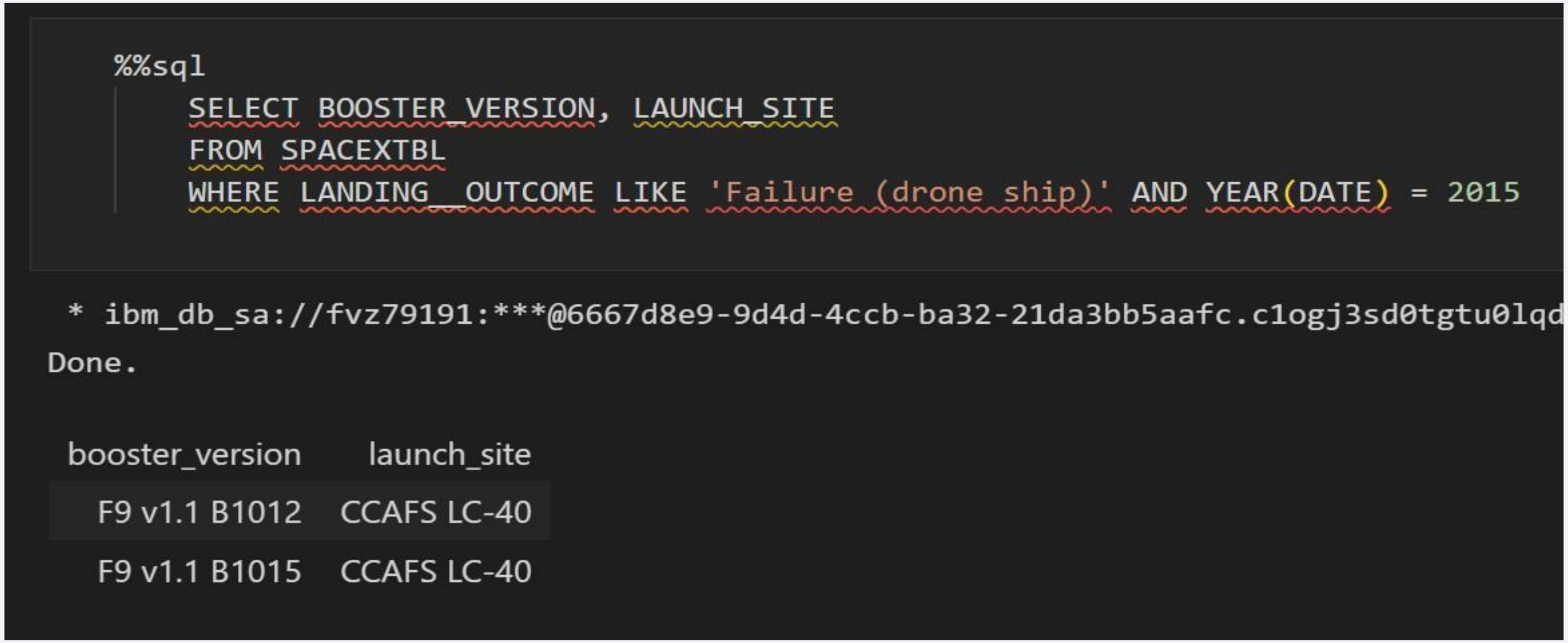
```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde
Done.
```

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

- The names of the booster which have carried the maximum payload mass are the following:

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

2015 Launch Records



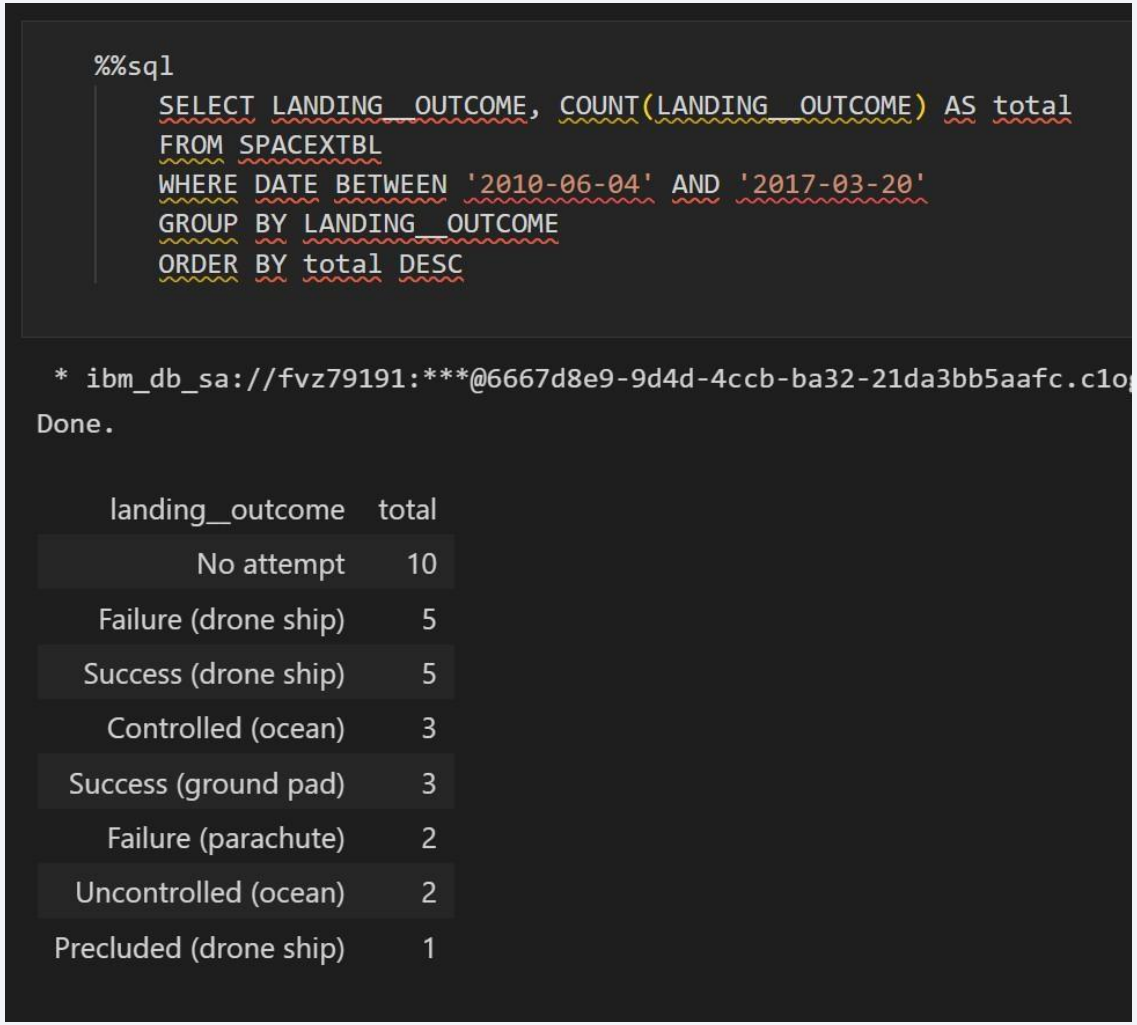
```
%%sql
    SELECT BOOSTER_VERSION, LAUNCH_SITE
    FROM SPACEXTBL
    WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqd
Done.

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are the F9 v1.1 B1012 with launch site CCAFS LC-40 and F9 v1.1 B1015 CCAFS LC-40, respectively.

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



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

```
 * ibm_db_sa://fvz79191:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1o
Done.
```

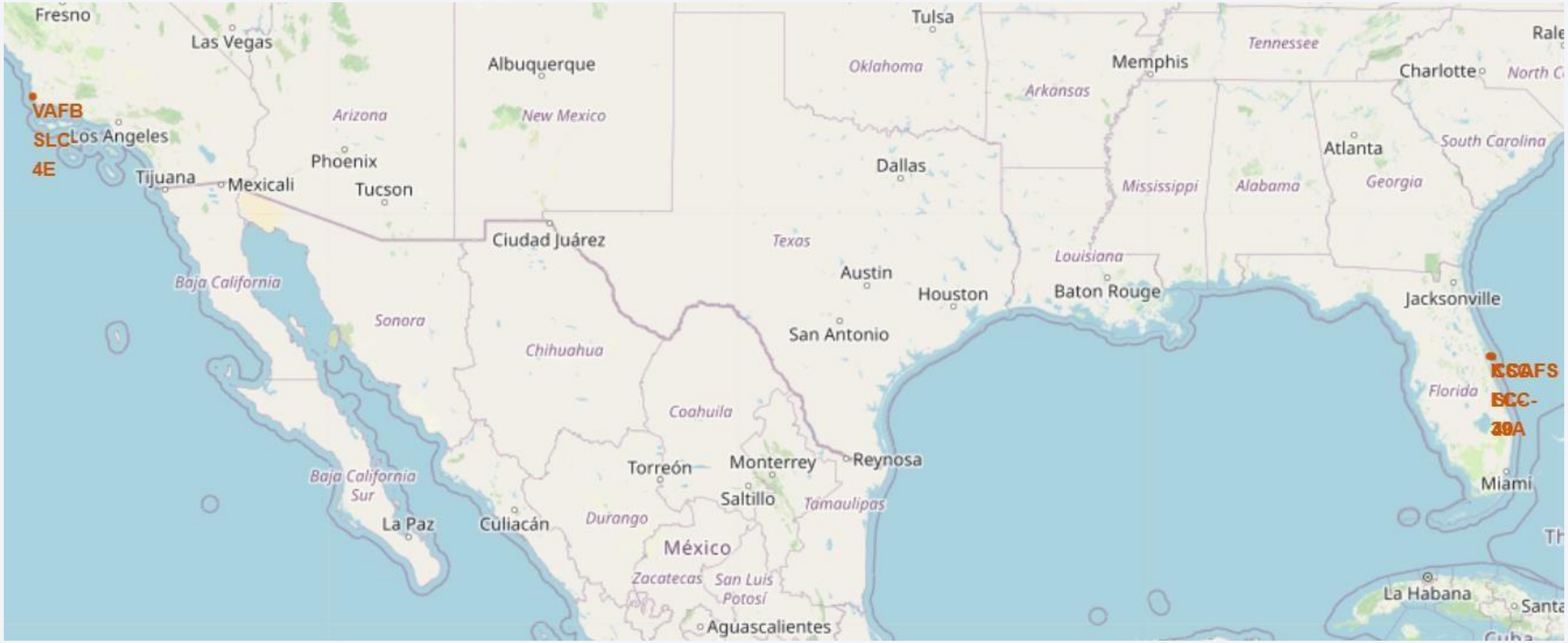
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

The rank count of landing outcomes (such as Failure (drone ship) or Success (groun pad)) between the date 2010-06-04 and 2017- 03-20, in descending order are shown in the image.

Section 3

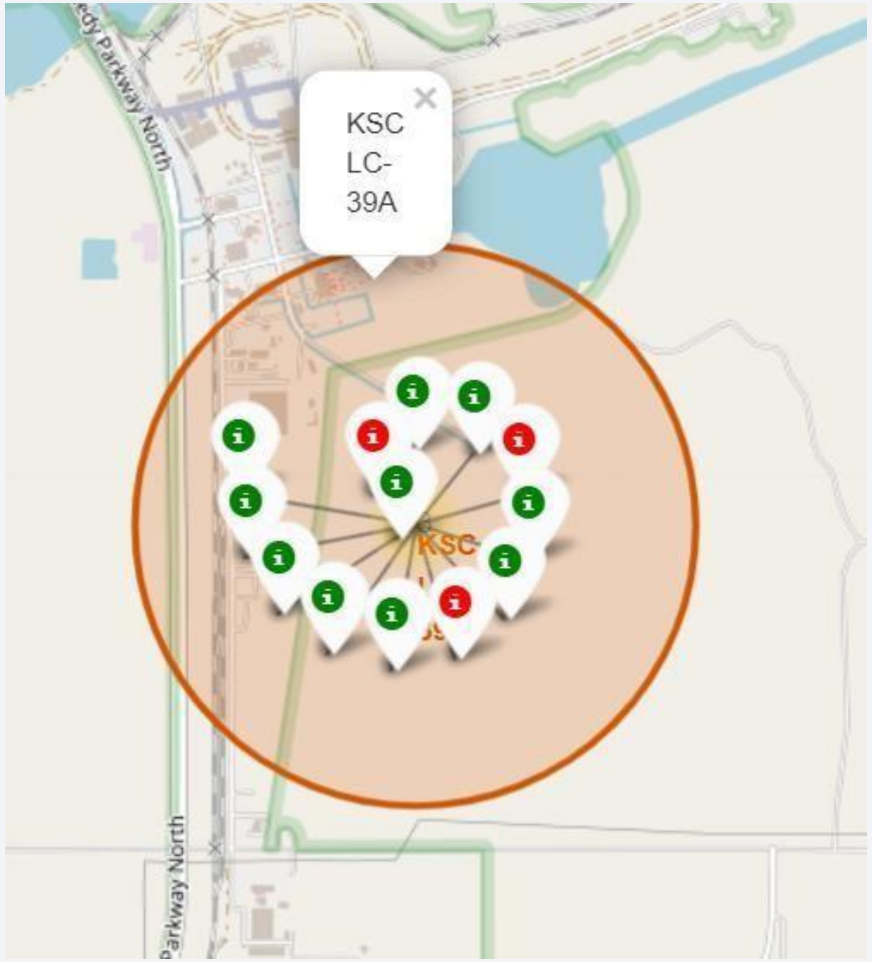
Launch Sites Proximities Analysis

Launch sites in Folium Map



- It’s possible to notice two cluster launch sites in Florida and California.

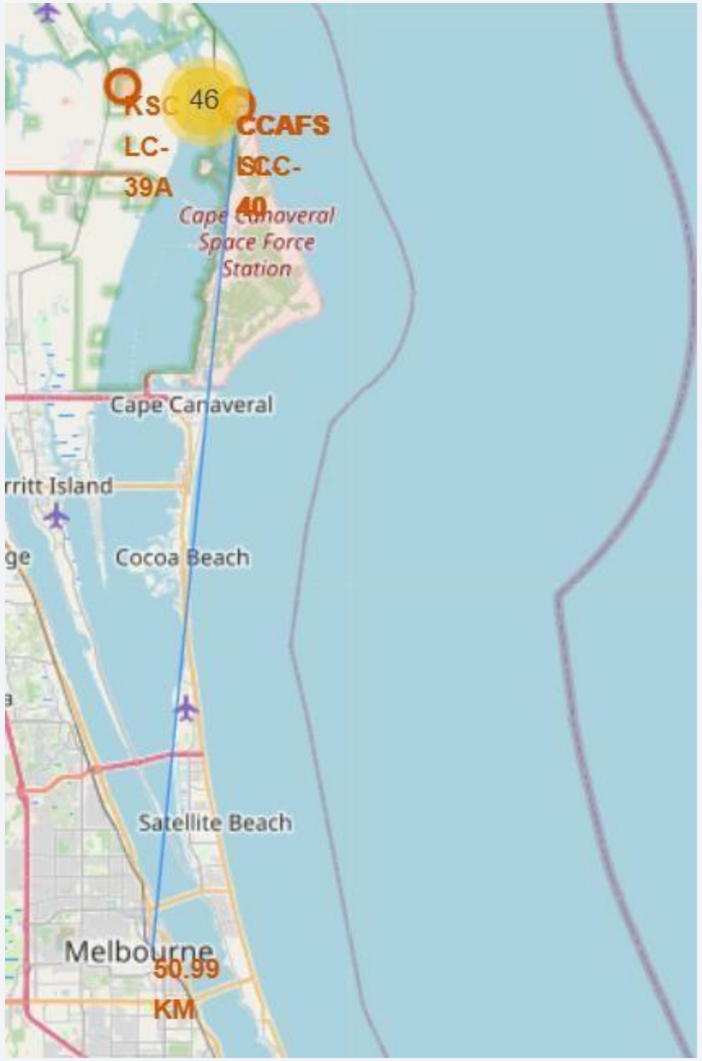
Outcome feature per launch site with Folium Map



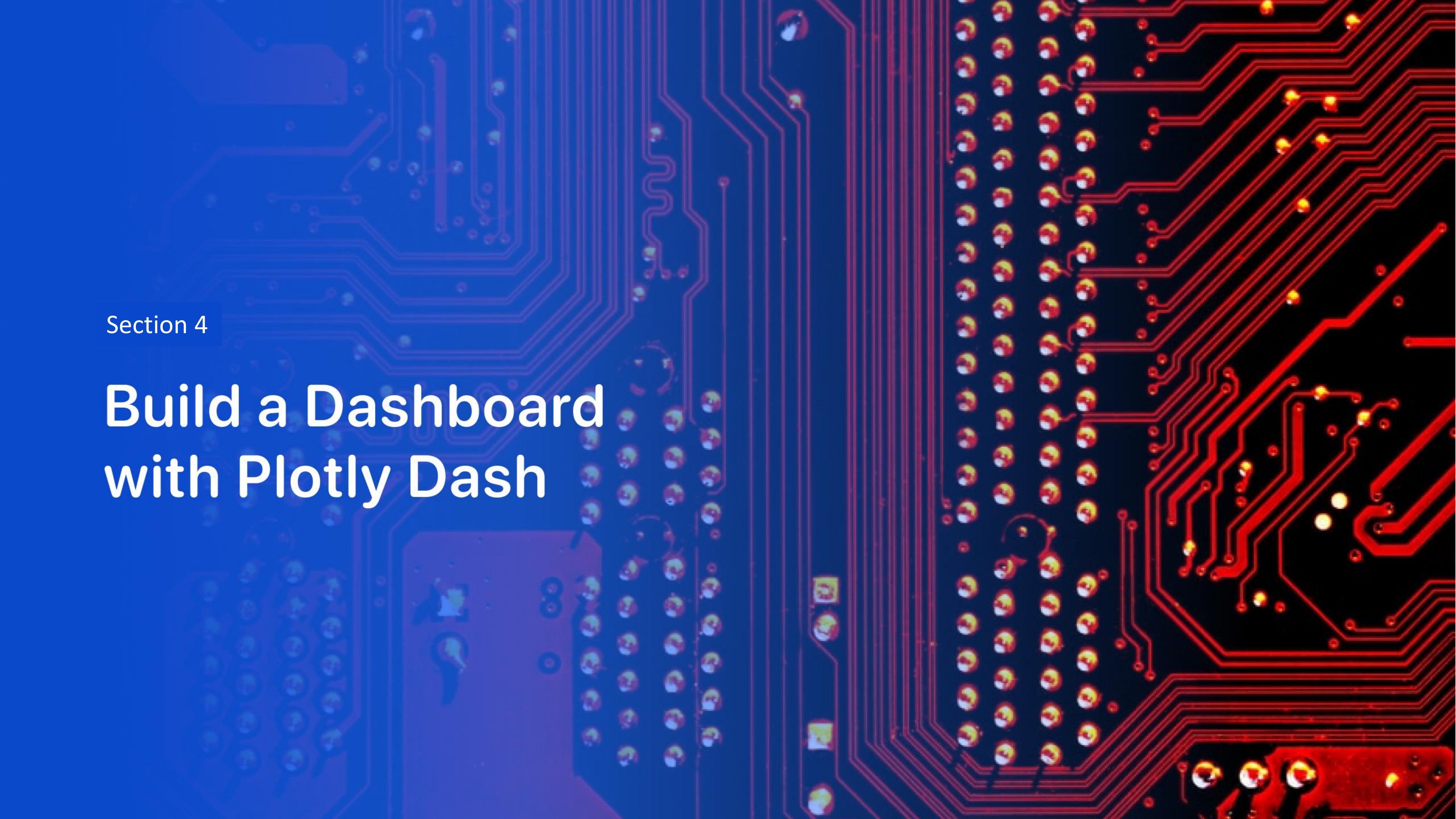
- Folium allows us to visually analyze the outcome feature per launch site.

- For instance, for the launch site KSC LC-39A in Florida, there are 10 successful landings and 3 failed landings

Distance between CCAFS SLC-40 launch site and its proximities



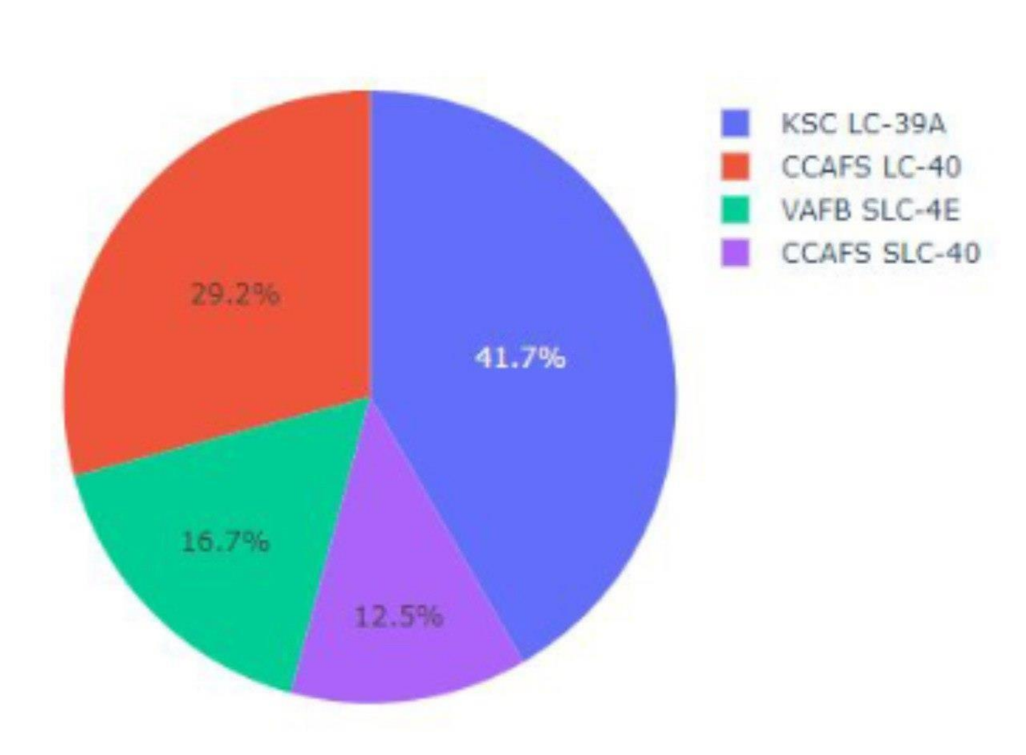
- The selected area for the distance analysis between the CCAFS launch site was Melbourne city.
- The distance between CCAFS and Melbourne city is 50.99 km.



Section 4

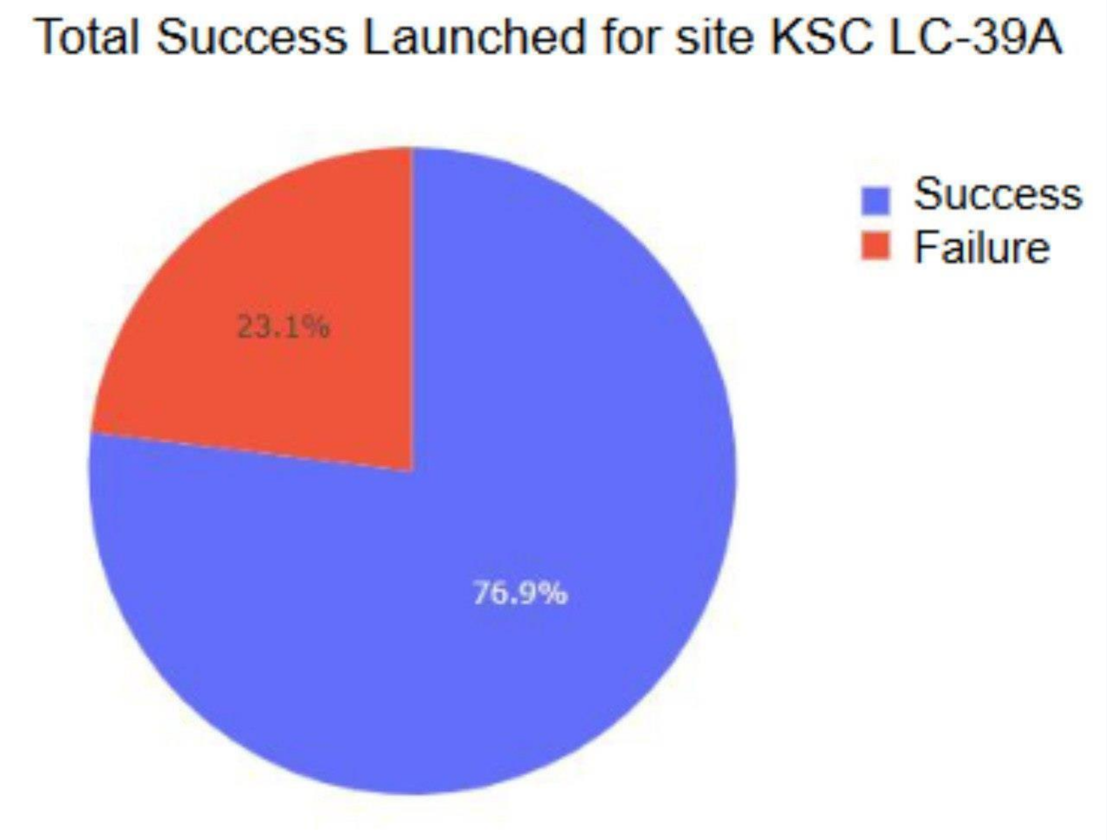
Build a Dashboard with Plotly Dash

Total Success Launches by site



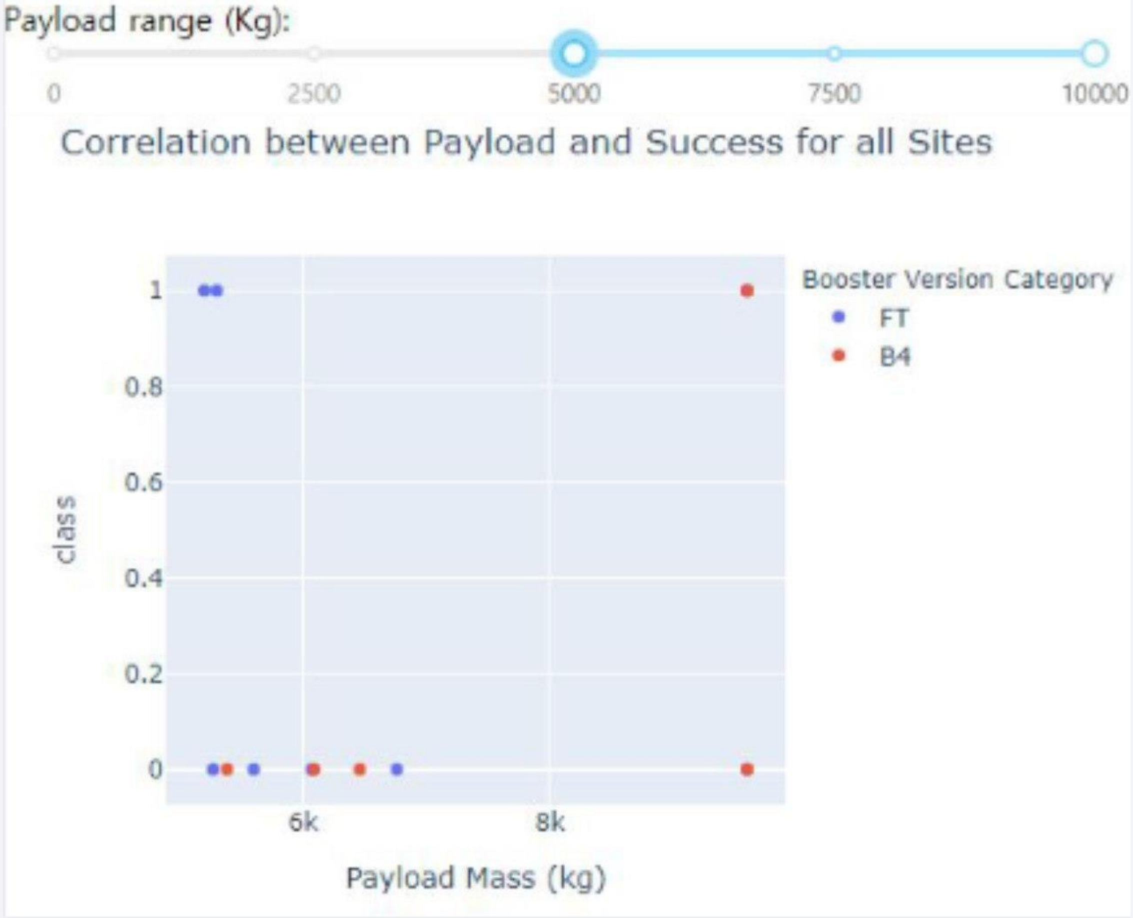
- Through the pie chart, it’s possible to determine that the highest launch success rate is KSC LC-39 A with 41.7%, followed by CCAFS LC-40 with 29.2% and at the bottom the VAFB SLC- 4E, and CCAFS SLC-40 with 16.7% and 12.5%, respectively

KSC LC-39A total Success Launches



- Through the pie chart is possible to determine that KSC LC-39A has a 76.9% success rate

Correlation between payload and success for all sites

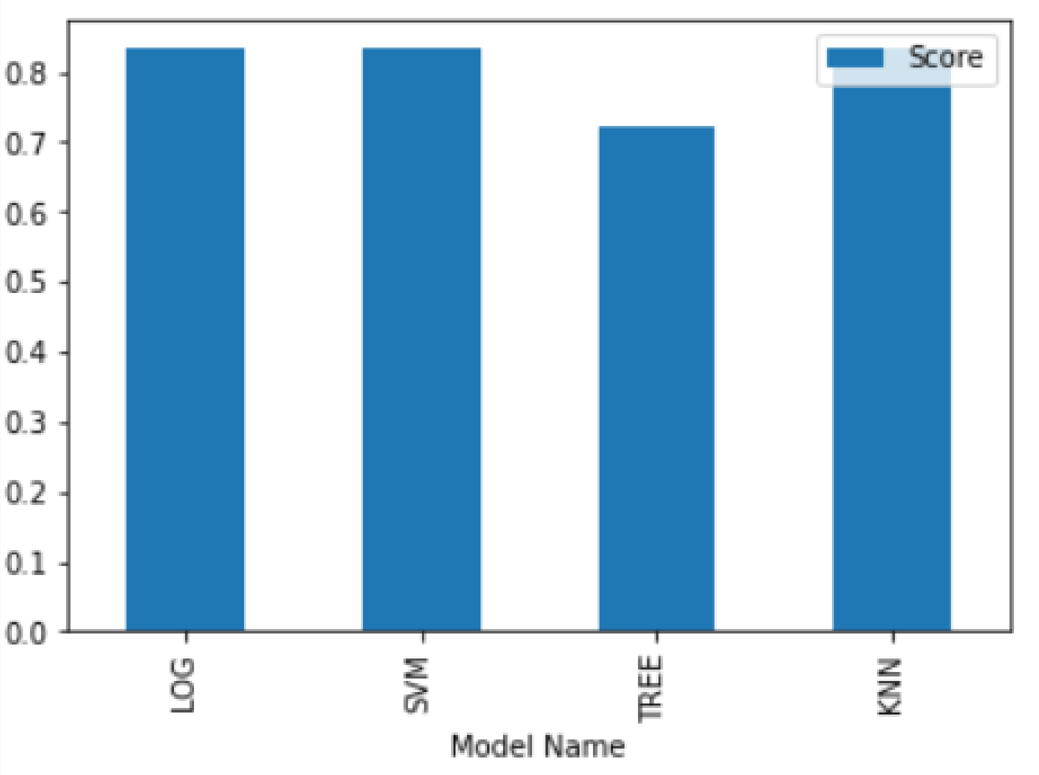


- The plot shows the correlation between payload and success for all sites with a [5000,10000] range.
- With heavier payload mass it appears to end in an unsuccessful landing

Section 5

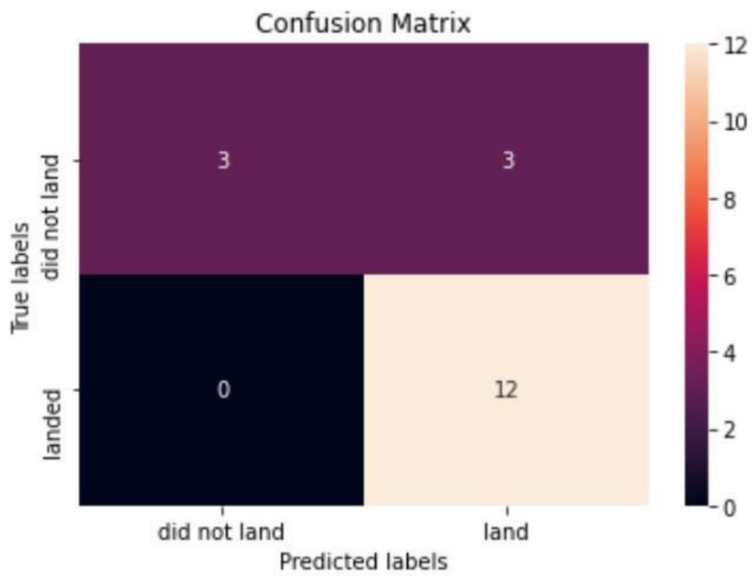
Predictive Analysis (Classification)

Classification Accuracy

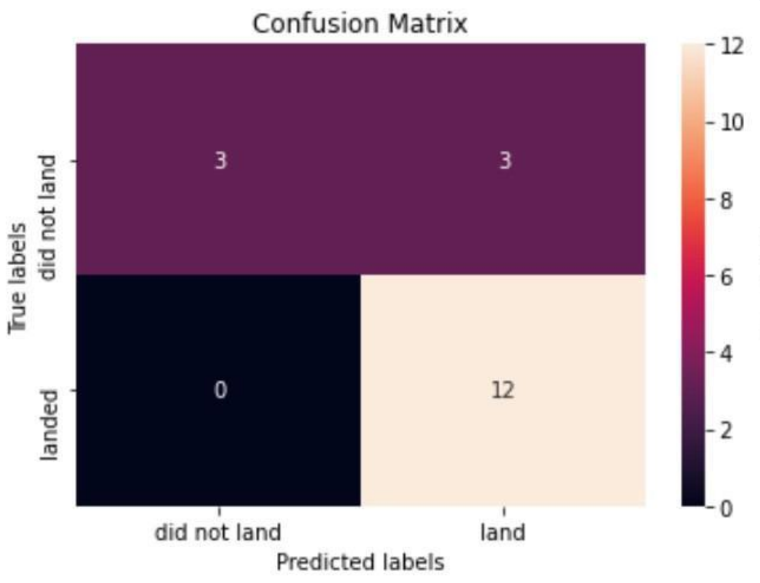


- The best performance models classifier were the logistic regression, SVM, and KNN, each with a score of 83%.

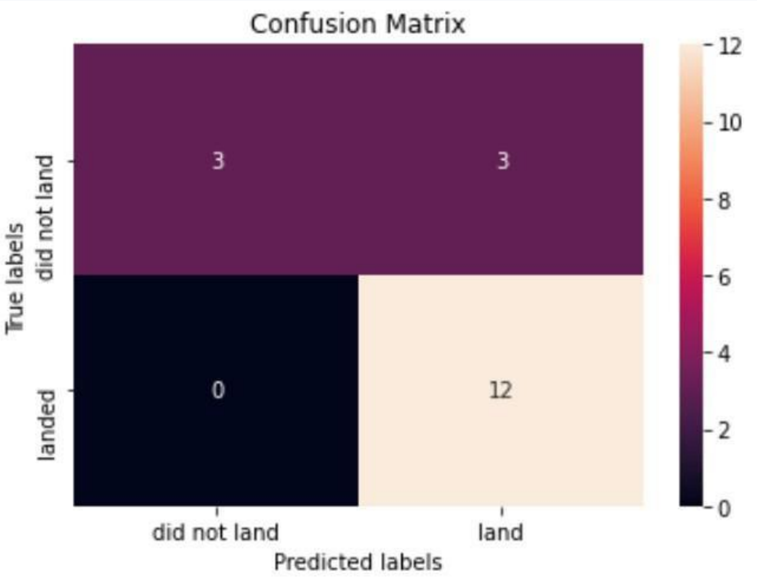
Confusion Matrix



LR results



SVM results



KNN results

- The confusion matrix for the models LR, SVM, and KNN are the same, with 3 TP, 3 FP, 0 FN, and 12 TN.

Conclusions

- Except in very specific cases, payload mass does not appear to be significantly related to a successful outcome.
- The analysis indicates a strong correlation between orbit type and flight number with unsuccessful launches.Certain orbits, such as ES-L1, GEO, HEO, and SSO, exhibit a consistent trend of successful landings, each with an average success rate of 100%.
- Based on the visual data provided by Folium and the dashboard, there is a noticeable relationship between the launch site and the landing outcome.Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are suitable classification models for predicting landing success, considering features such as orbit, launch site, landing pad, and Falcon 9 serial number.

Appendix

- The repository for this project is available here in Github

Thank you!