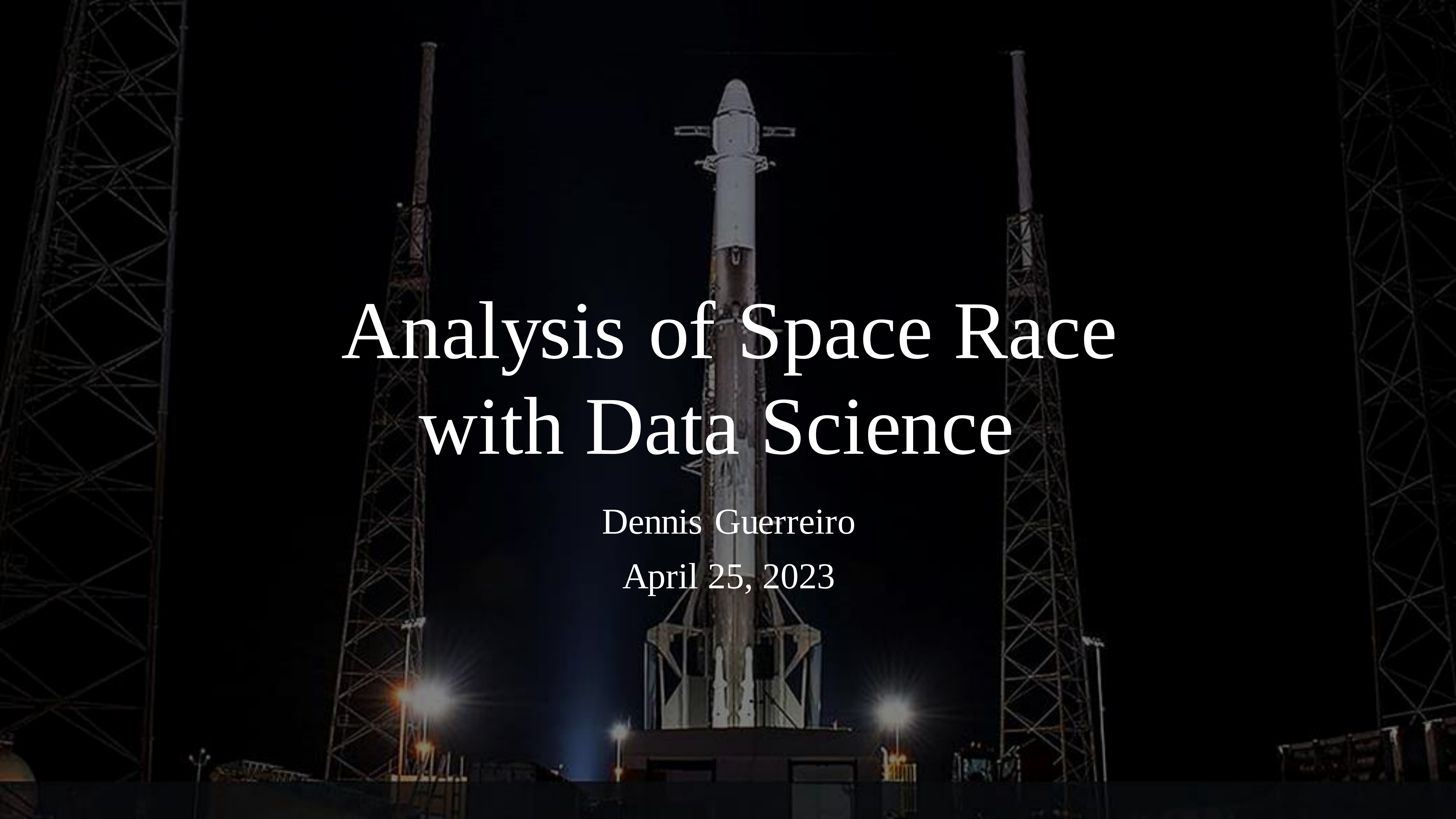


Analysis of Space Race with Data Science

Dennis Guerreiro

April 25, 2023

• Executive Summary

In this project we will apply the following comprehensive data science methodology:

- Data collection
- Data wrangling
- Exploratory data analysis
- Data visualization
- Model development
- Model evaluation

Using this Framework we can predict successfully the outcome of a launch, which is the main objective of the project.

Introduction

- Overview of the project:

  SpaceY will competes with other Companies that are making space travel affordable for everyone such as Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX. However, we focus on SpaceX because it has the lowest launch costs compared to its competitors. SpaceX Advertises Falcon 9 rockets 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 SpaceX's Falcon 9 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 for Space Y company to bid against SpaceX for a rocket launch.

- Problems to be solved:

  Determine if SpaceX will reuse the first stage

- # Methodology

The sequence of steps that have been followed to complete this project:

1. Data Collection using SpaceX API and Web Scraping method
2. Data Wrangling using Pandas and Numpy
3. Exploratory Data Analysis using Pandas, Numpy, and SQL
4. Data visualization using Matplotlib, Seaborn, Folium, and Plotly Dash
5. Perform Predictive Analysis using Classification Models:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K-nearest neighbors (KNN)

○ Data Collection, Space API

The data was obtained using the **SpaceX API**

- ✓ This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

- ✓ I made a get request to the SpaceX API. I got the launch data using a series of functions and the following endpoints:

	Endpoints	Purposes
URL: **https://api.spacexdata.com/v4/**	/capsules	Booster name
	/cores	Outcome of the landing. Type of landing. Number of flights.
	/launches/past	Name of the launch site Longitude, and Latitude

- ✓ Next, I decoded the response content as a Json using .json() function call and turn it into a DataFrame using .json_normalize().

- ✓ Then, I cleaned the data, checked for missing values, and applied the mean function to fill the missing data.

○ Data Collection, Web Scraping

Another popular data source for obtaining Falcon 9 Launch data is Web Scraping related Wiki pages.

- ✓ I used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- ✓ Then I extracted all column/variable names from the HTML table header
- ✓ I created a DataFrame by parsing the launch HTML tables

We have combined the SpaceX API and the Web Scraping method to obtain the data we need. Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or no.

Please, see Appendix for specific information about collecting data using SpaceX API and Web Scraping.

○ Data Wrangling

In this step of the process, I expanded the data exploratory analysis to:

- ✓ Identify and calculate the percentage of the missing values in each attribute.
- ✓ Calculate the number of launches on each site using the method **value_counts()**
- ✓ Calculated the number of launches on each site and the occurrence of each orbit.
- ✓ Determine the training labels for training supervised models
- ✓ We created landing outcome label from Outcome column.

Takeaway:
A new column called "Class" was added to the DataFrame. This variable will represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully.

Lastly, we calculated the mean of the values in the Class column of the DataFrame to determine the success rate. The success rate is equal to 0.666...

Please, see Appendix for specific information about Data Wrangling

○ Exploratory Data Analysis

We continue with the Exploratory Data Analysis process that we saw in the previous step (Data Wrangling) This time we expanded the EDA process to perform an important step of this project, the Data Feature Engineering.

This step is necessary to improve the machine learning models and includes the following tasks:

- ✓ Select the features that will be used in success prediction in the future module.
- ✓ Create dummy variables to categorical columns using the function get_dummies and features dataframe to apply OneHotEncoder to the column Orbits, LaunchSite, LandingPad, and Serial
- ✓ Cast all numeric columns to float64

Please, see Appendix for specific information about EDA with SQL

○ EDA with SQL

- We connected to a Db2 Database called bludb using port 31505
- We loaded the dataset into the corresponding table in a Db2 database.
- We executed SQL queries to answer important questions such as:
  - The name the names of the unique launch sites in the space mission.
  - Display 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
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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.

Please, see Appendix for specific information about EDA with SQL.

○ Data Visualization

Next, we combine exploratory data analysis with python libraries for interactive visualization analysis.

We build various graphs, line chart, bar chart, scatter plot to identify patterns and relationships between variables:

✓ For example, to discovered some preliminary relationship between the launch site and success rates, we build a scatterplot graph using the **Matplotlib** and **Seaborn** Python data visualization libraries.

✓ Also, we used **Folium** libraries to find some geographical patterns about launch sites. Using folium we was able to do the following tasks:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
- Marker colors based on the class value: **green tag represents a successful launch, and red tag represents a failed launch**

✓ We build an interactive dashboard using **Plotly Dash**.

The dashboard includes:

- Pie Chart: shows the total launches from each site
- Scatter plot: The correlation between payload and mission Outcome

Please, see Appendix for specific information

○ Predictive Analisys, Classification Models

In order to build the classification models we used Scikit-Learn library functions.

The machine learning prediction phase include the following steps:

- ✓ Load the required libraries and Define Auxiliary Functions.
- ✓ Standardize the data with sklearn. function StandardScaler()
- ✓ Split into training data and test data.
- ✓ Build different machine learning models (Logistic Regression, SVM, Decision Tree, and KNN)
- ✓ Fit the models on the training set
- ✓ Tune the models using GridSearchCV, feature engineering, and algorithm tunning
- ✓ Find the best performing classification model
- ✓ Evaluate the models based on their accuracy scores and confusion matrix

Please, see Appendix for specific information

- # Results

In this project, the goal was to predict if the SpaceX Falcon 9 first stage would land successfully using several machine learning classification algorithms.

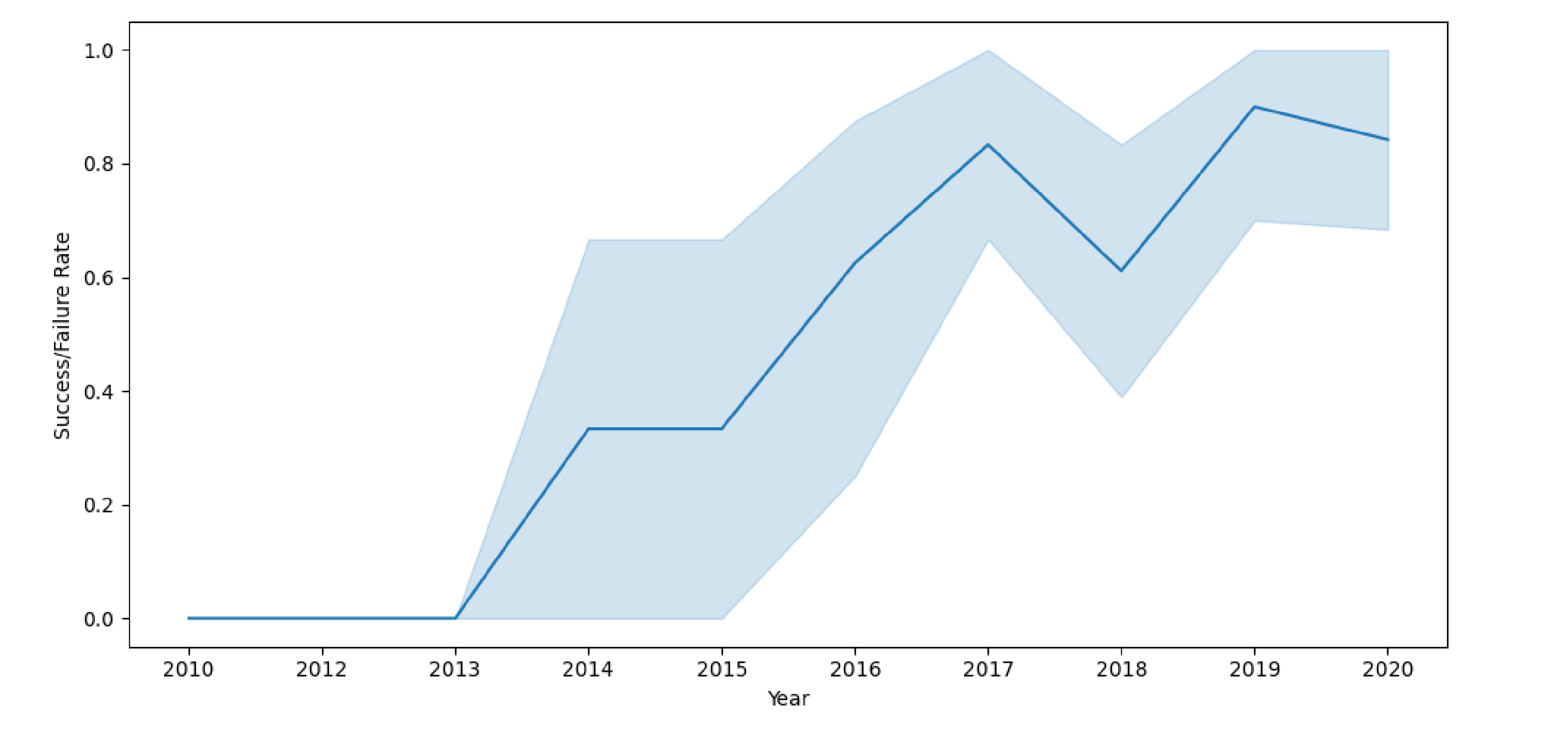
The main steps in this project included data collection, wrangling, exploratory data analysis as well as model development and evaluation of each model.

Next, we will present the results in the form of observations

Launch Success Yearly Trend

Required EDA with visualization results
Observation 1

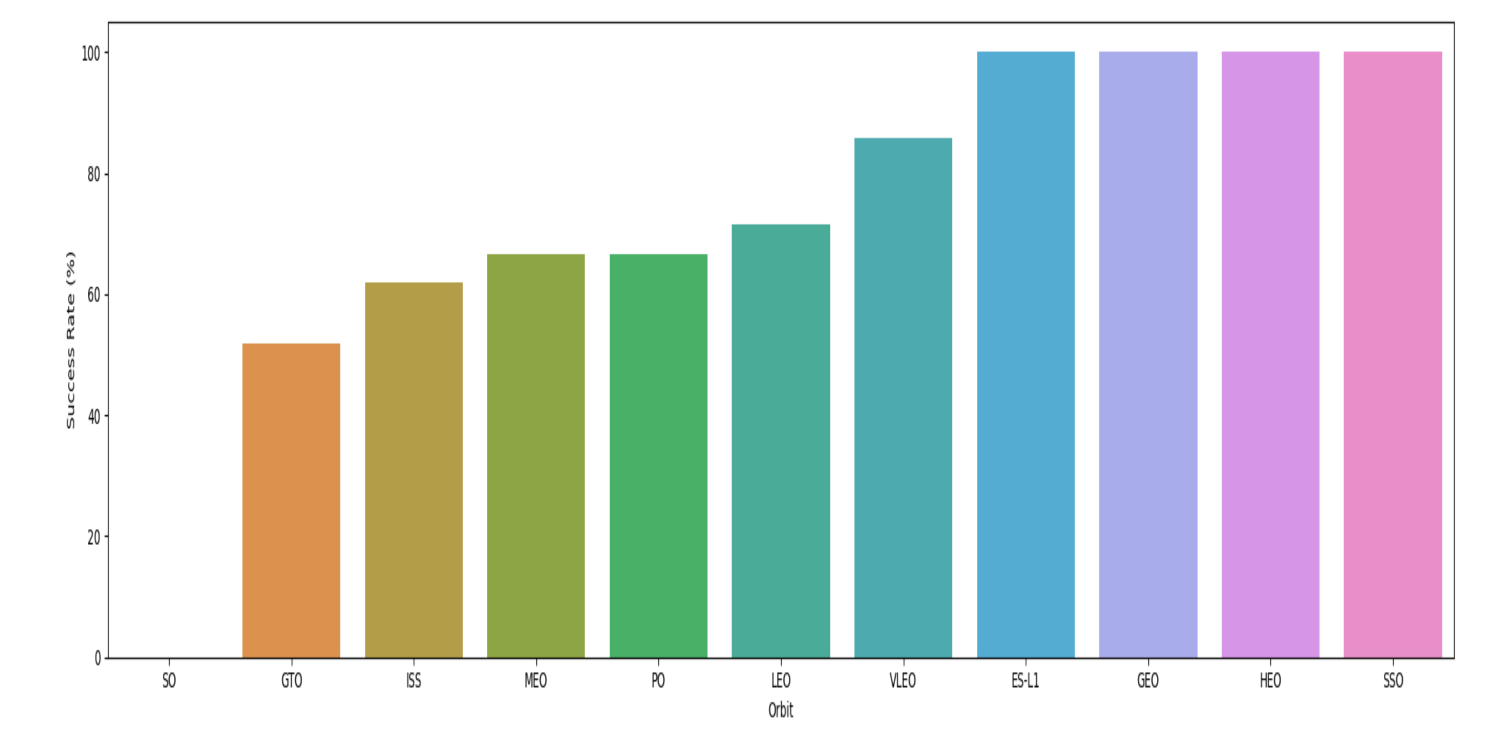
- ✓ Sucess rate since 2013 kept increasing till 2020



Success Rate of each Orbit

Required EDA with visualization results
Observation 2

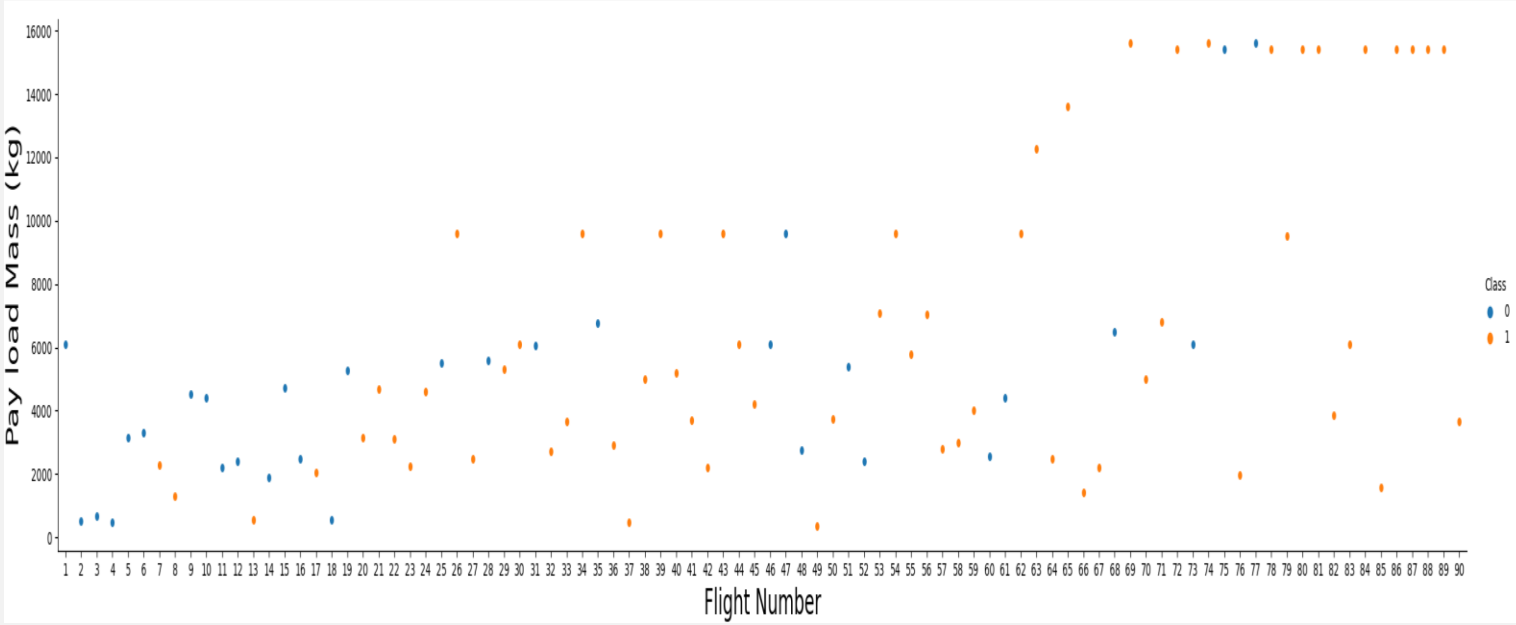
- ✓ A total of 4 orbits (ES-L1, GEO, HE0, SSO) have a 100% success rate
- ✓ The SO orbit has the least success rate among the orbits



FlightNumber vs. PayloadMass

Required EDA with visualization results
Observation 3

- ✓ We see that as the flight number increases, the first stage is more likely to land successfully
- ✓ The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



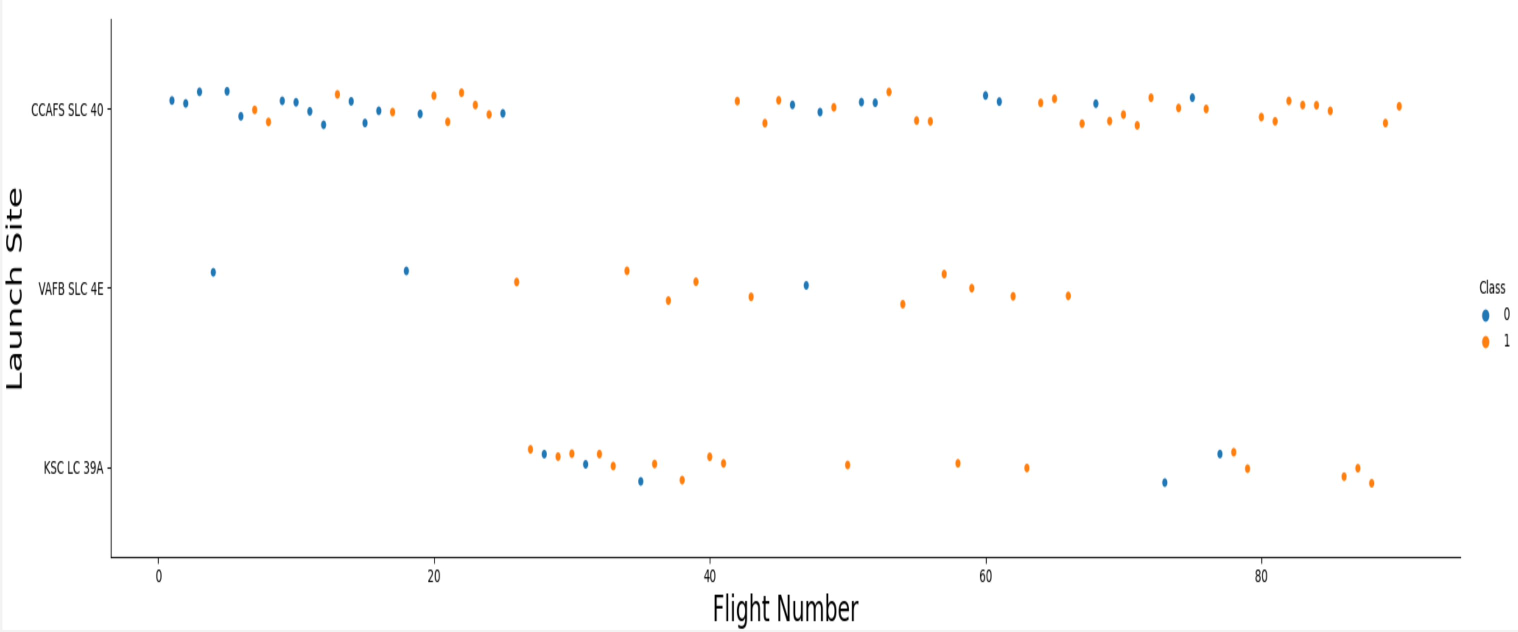
FlightNumber vs. Launch Site

Required EDA with visualization results
Observation 4

- ✓ We see that different launch sites have different success rates:

  CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%

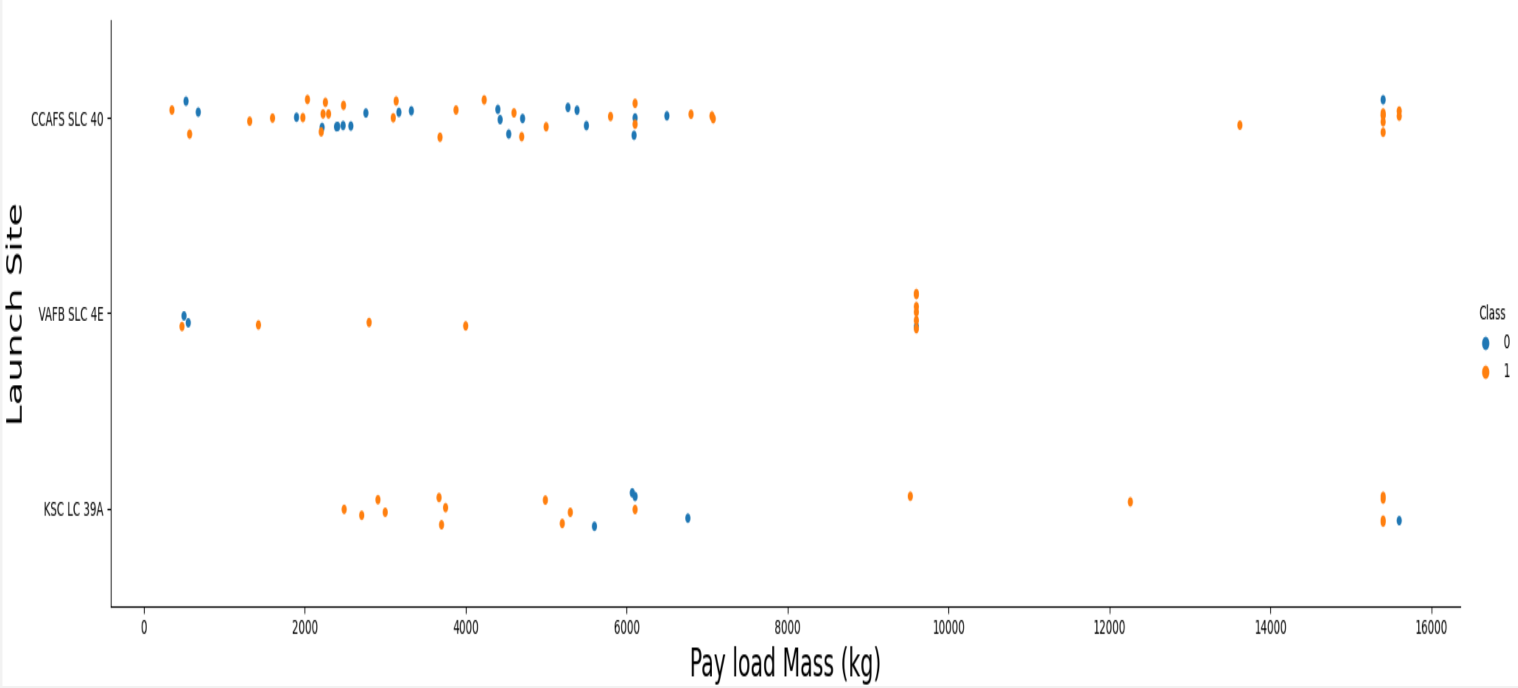
- ✓ The larger the flight amount at a launch site, the greater the success rate at a launch site.



Payload vs. Launch Site

Required EDA with visualization results
Observation 5

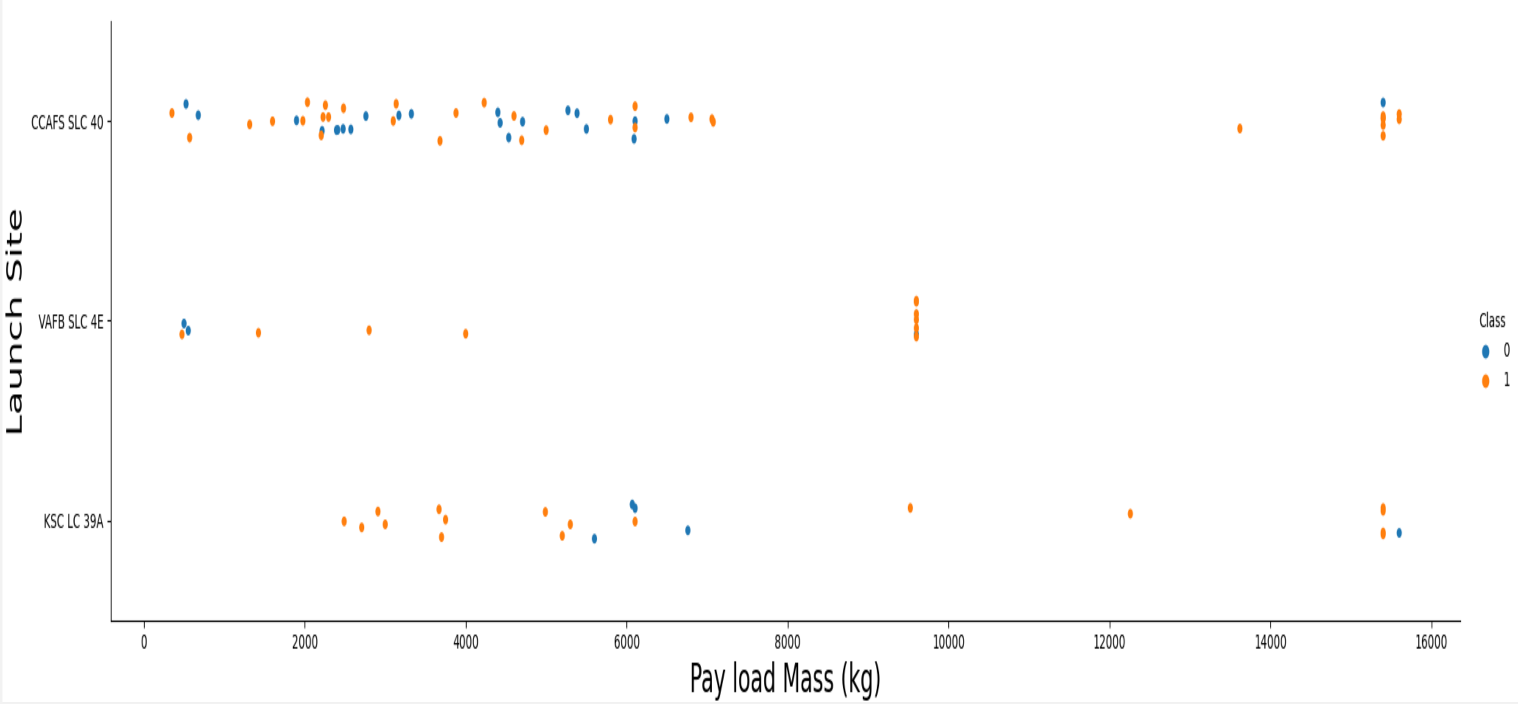
- ✓ We observed from the chart that there are no rockets launched for heavy payload mass (greater than 10000) at the VAFB-SLC launch site.



FlightNumber vs. Orbit type

Required EDA with visualization results
Observation 6

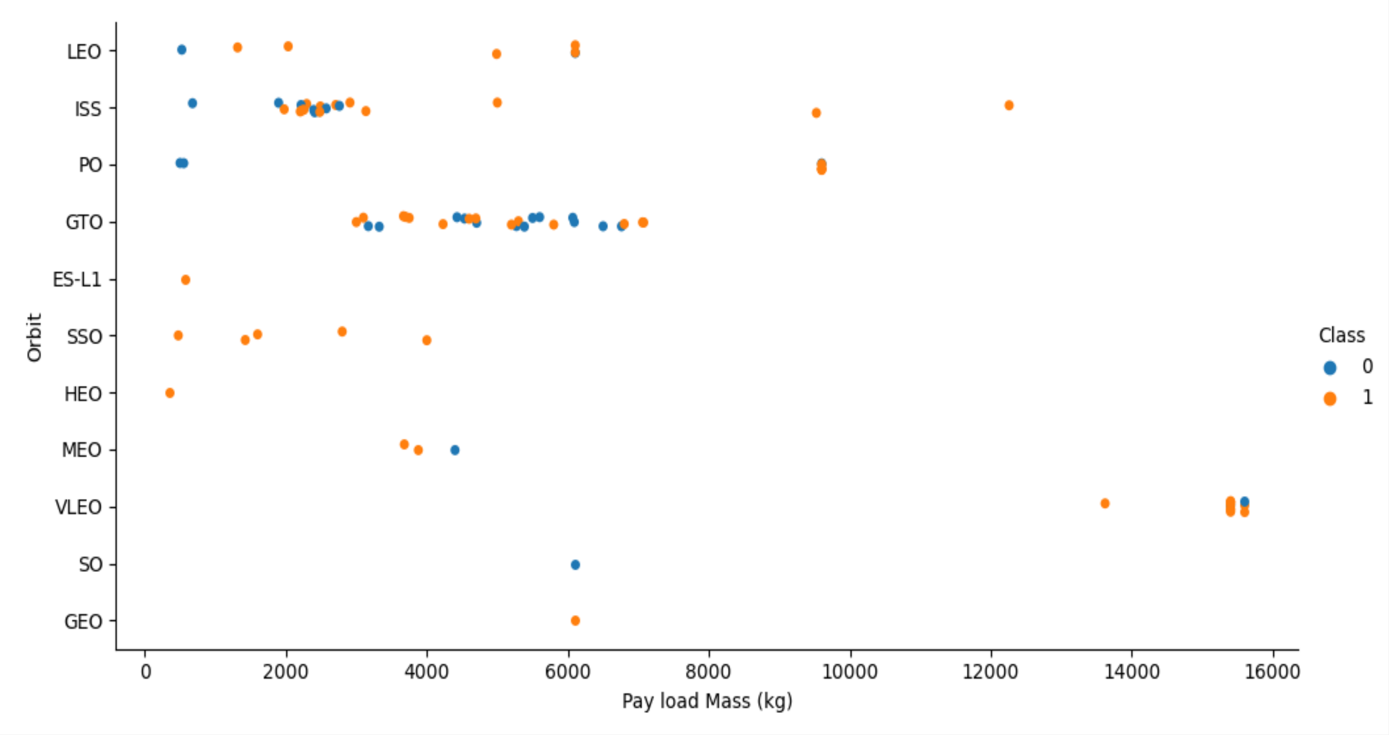
- ✓ According to the chart, in the LEO orbit, the Success appears related to the number of flights
- ✓ on the other hand, there seems to be no relationship between flight number when in GTO orbit



Payload vs. Orbit type

Required EDA with visualization results
Observation 7

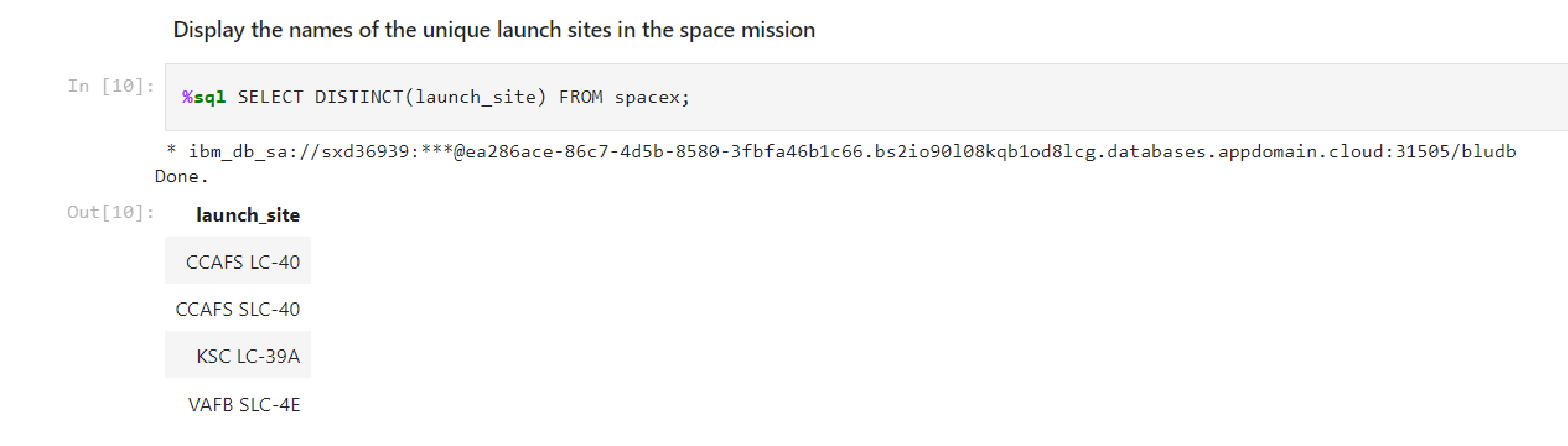
- ✓ We can observed that with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- ✓ for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here



EDA with SQL

Required EDA with SQL results
Observation 1

✓ The name of the unique launch sites in the space mission



Display the names of the unique launch sites in the space mission

```
In [10]: %sql SELECT DISTINCT(launch_site) FROM spacex;
```

```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

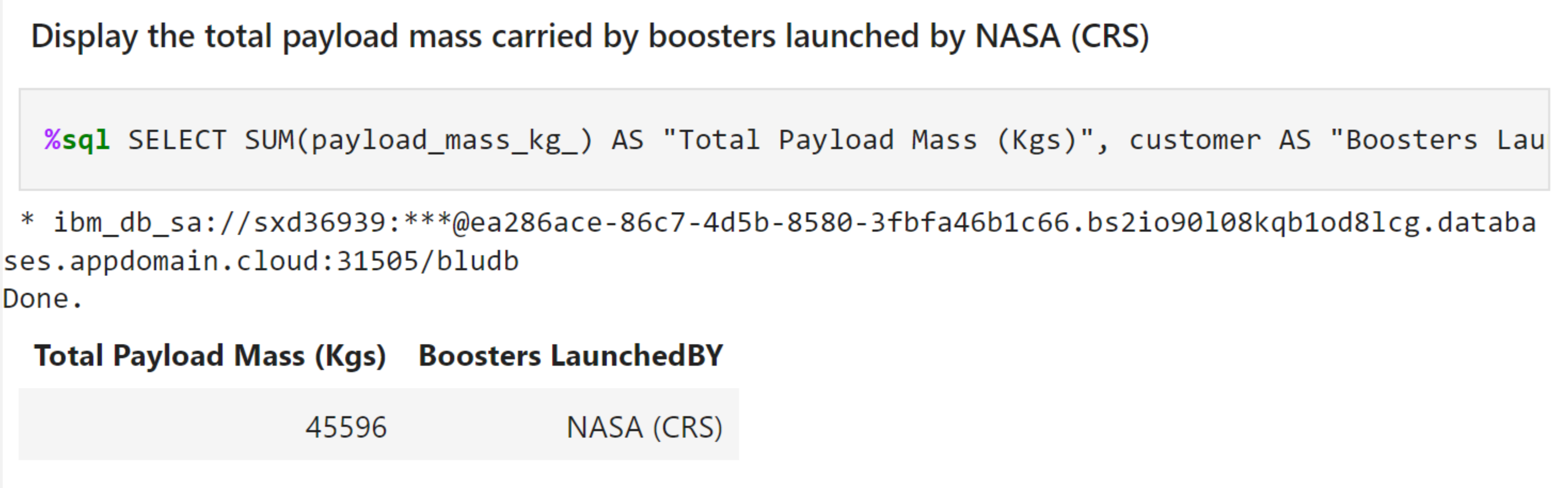
Out[10]:

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

EDA with SQL

Required EDA with SQL results
Observation 2

- ✓ Total payload mass carried by boosters launched by NASA CRS



Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass_kg_) AS "Total Payload Mass (Kgs)", customer AS "Boosters Lau
```

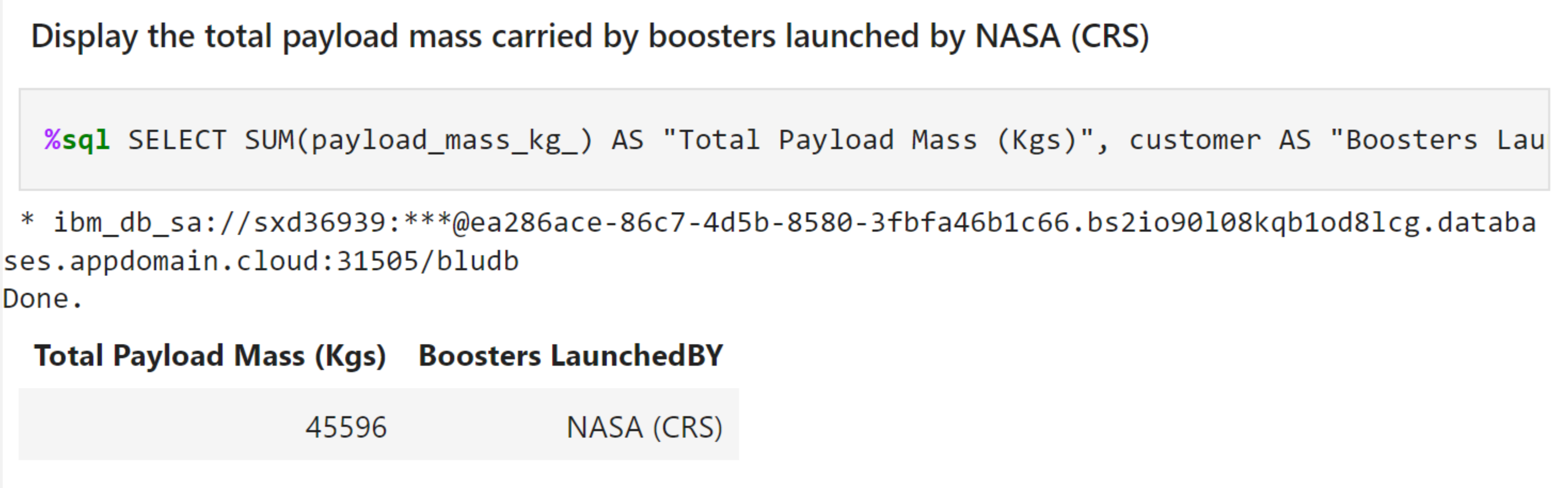
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

Total Payload Mass (Kgs)	Boosters LaunchedBY
45596	NASA (CRS)

EDA with SQL

Required EDA with SQL results
Observation 3

✓ The Average payload mass carried by booster version F9 v1.1



Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass_kg_) AS "Total Payload Mass (Kgs)", customer AS "Boosters Lau
```

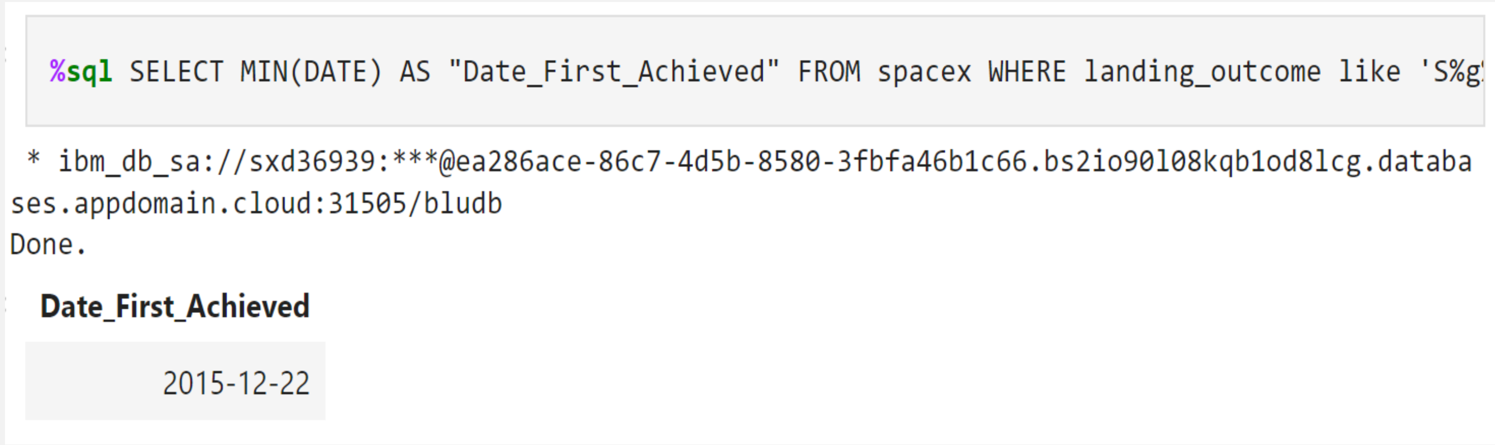
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

Total Payload Mass (Kgs)	Boosters LaunchedBY
45596	NASA (CRS)

EDA with SQL

Required EDA with SQL results
Observation 4

✓ The date when the first successful landing outcome in ground pad was acheived.



```
%sql SELECT MIN(DATE) AS "Date_First_Achieved" FROM spacex WHERE landing_outcome like 'S%g
```

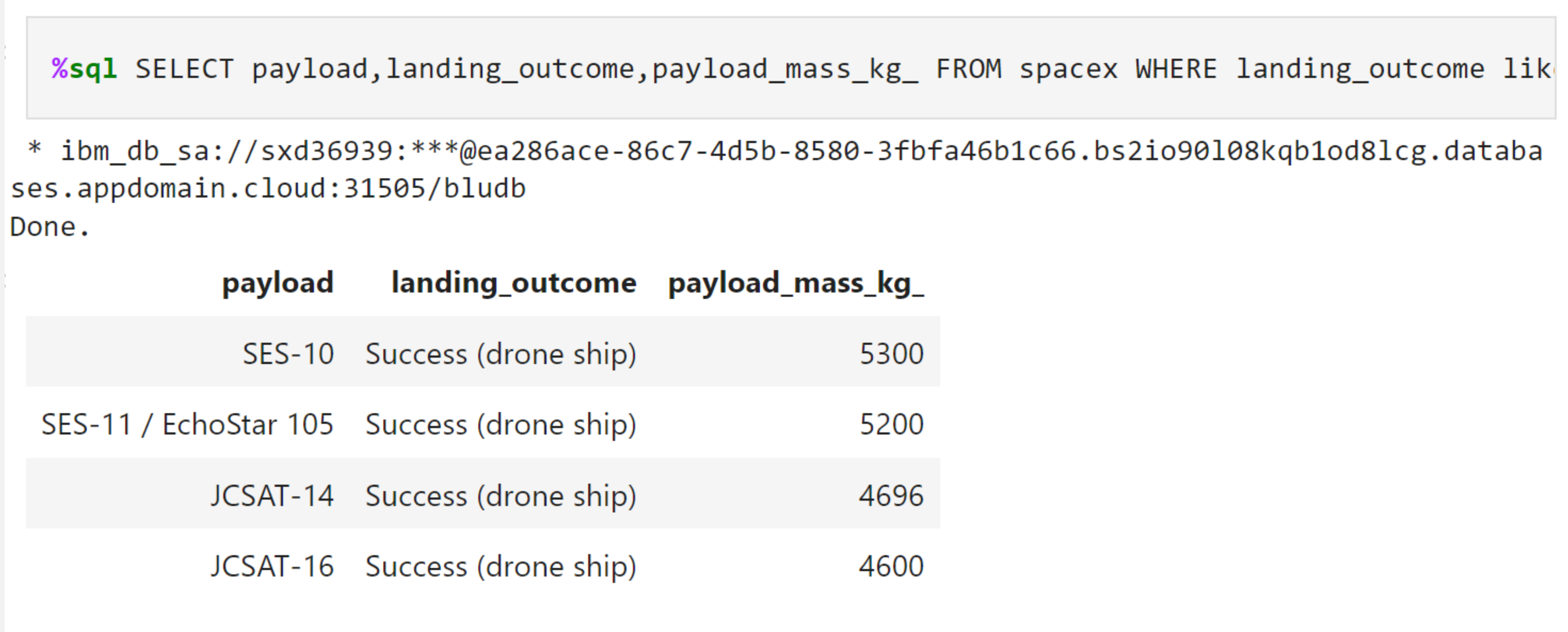
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

Date_First_Achieved
2015-12-22

EDA with SQL

Required EDA with SQL results
Observation 5

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



```
%sql SELECT payload,landing_outcome,payload_mass_kg_ FROM spacex WHERE landing_outcome lik
```

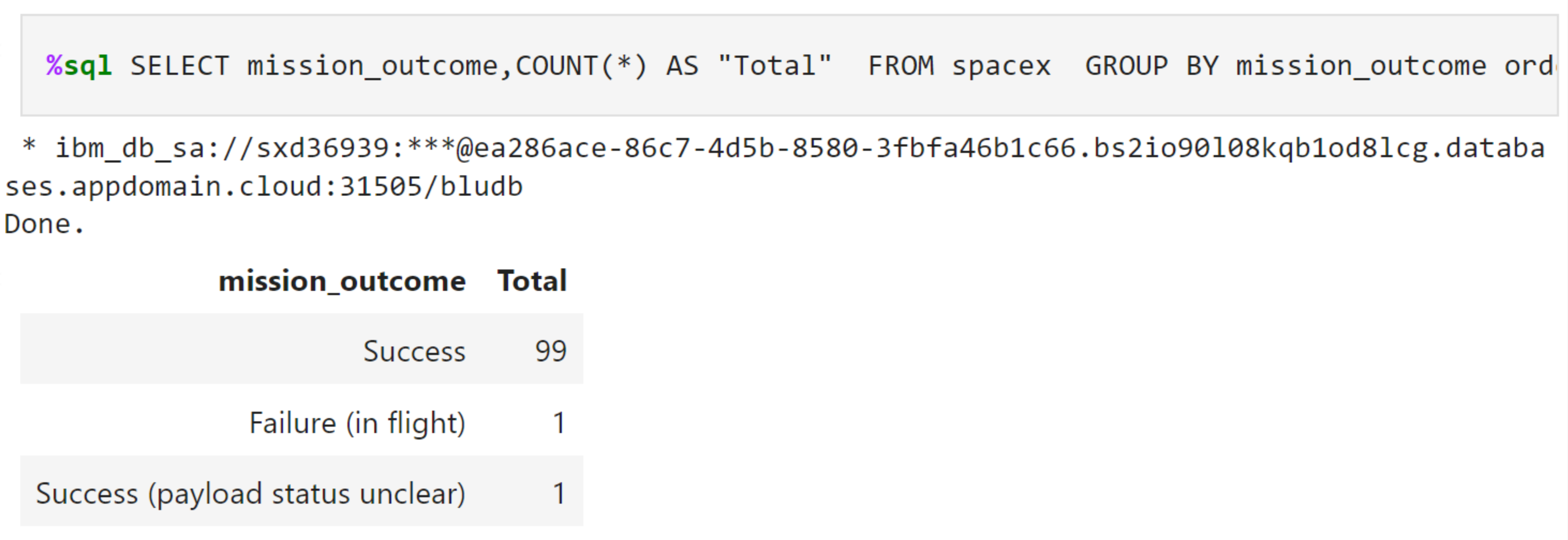
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

payload	landing_outcome	payload_mass_kg_
SES-10	Success (drone ship)	5300
SES-11 / EchoStar 105	Success (drone ship)	5200
JCSAT-14	Success (drone ship)	4696
JCSAT-16	Success (drone ship)	4600

EDA with SQL

Required EDA with SQL results
Observation 6

- ✓ List the total number of successful and failure mission outcomes



```
%sql SELECT mission_outcome,COUNT(*) AS "Total"  FROM spacex  GROUP BY mission_outcome ord
```

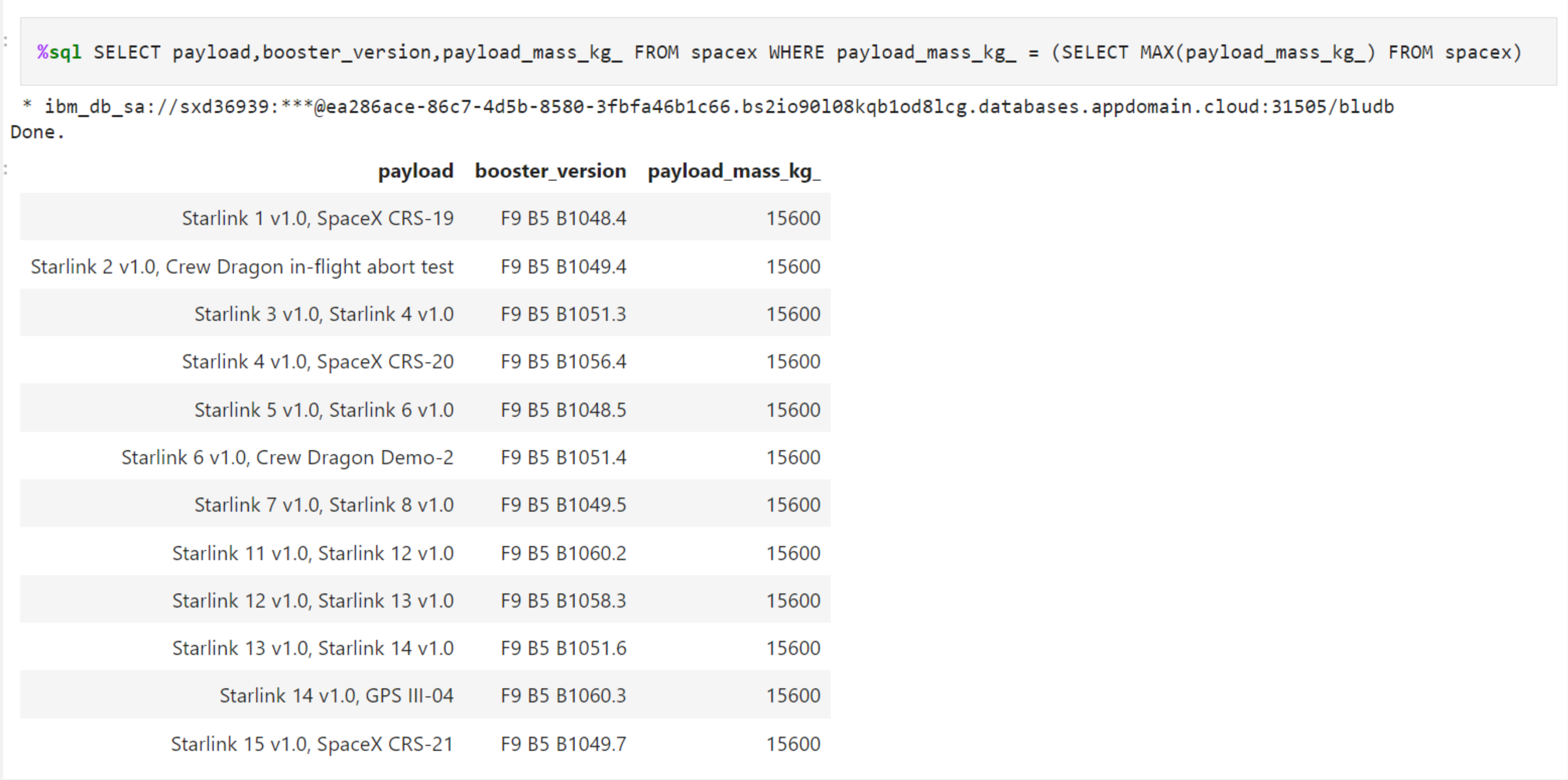
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

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

EDA with SQL

Required EDA with SQL results
Observation 7

✓ The names of the booster_versions which have carried the maximum payload mass. Use a subquery



```
%sql SELECT payload,booster_version,payload_mass_kg_ FROM spacex WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM spacex)
```

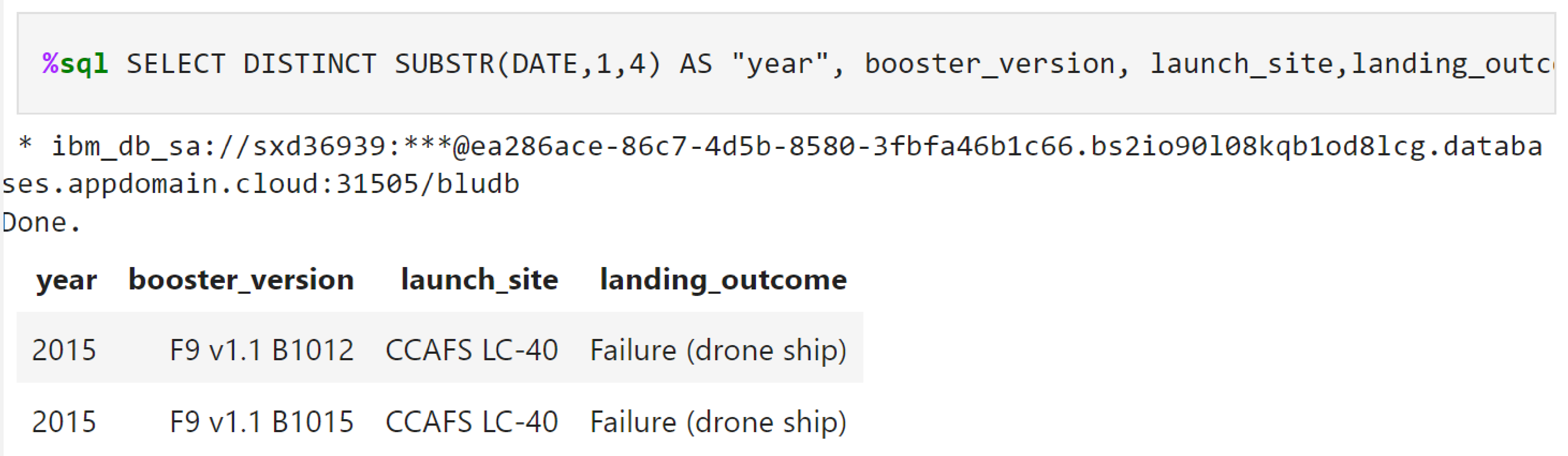
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

payload	booster_version	payload_mass_kg_
Starlink 1 v1.0, SpaceX CRS-19	F9 B5 B1048.4	15600
Starlink 2 v1.0, Crew Dragon in-flight abort test	F9 B5 B1049.4	15600
Starlink 3 v1.0, Starlink 4 v1.0	F9 B5 B1051.3	15600
Starlink 4 v1.0, SpaceX CRS-20	F9 B5 B1056.4	15600
Starlink 5 v1.0, Starlink 6 v1.0	F9 B5 B1048.5	15600
Starlink 6 v1.0, Crew Dragon Demo-2	F9 B5 B1051.4	15600
Starlink 7 v1.0, Starlink 8 v1.0	F9 B5 B1049.5	15600
Starlink 11 v1.0, Starlink 12 v1.0	F9 B5 B1060.2	15600
Starlink 12 v1.0, Starlink 13 v1.0	F9 B5 B1058.3	15600
Starlink 13 v1.0, Starlink 14 v1.0	F9 B5 B1051.6	15600
Starlink 14 v1.0, GPS III-04	F9 B5 B1060.3	15600
Starlink 15 v1.0, SpaceX CRS-21	F9 B5 B1049.7	15600

EDA with SQL

Required EDA with SQL results
Observation 8

✓ List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



```
%sql SELECT DISTINCT SUBSTR(DATE,1,4) AS "year", booster_version, launch_site,landing_outc
```

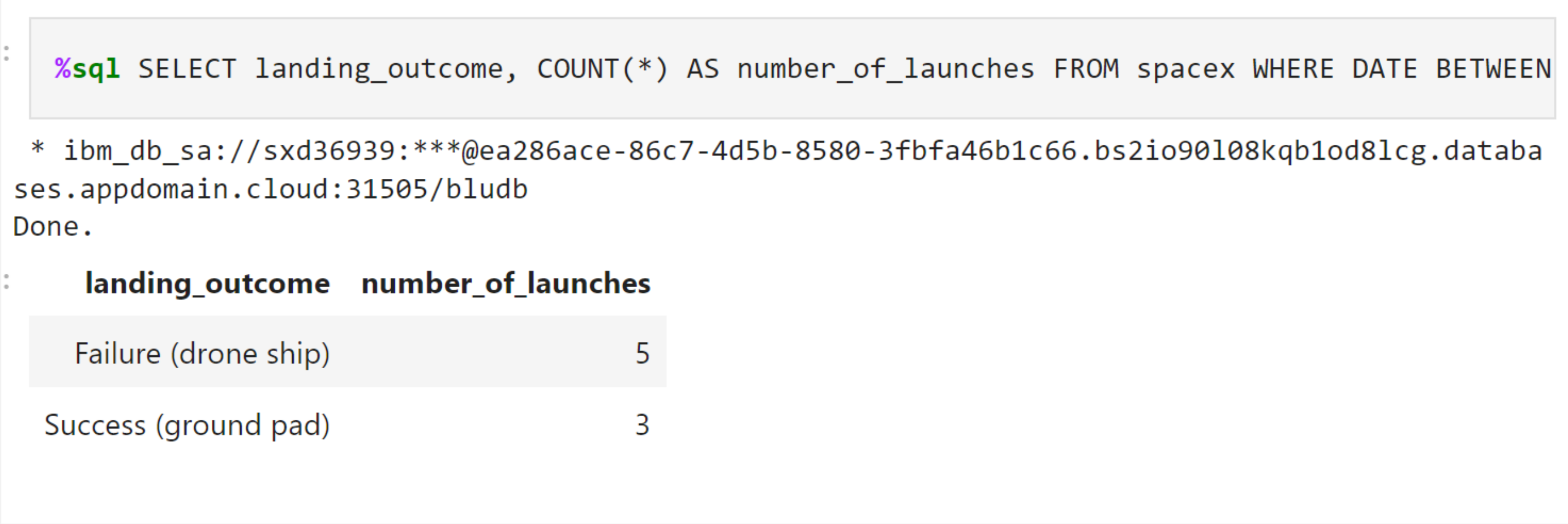
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

year	booster_version	launch_site	landing_outcome
2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

EDA with SQL

Required EDA with SQL results
Observation 9

- ✓ 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



```
%sql SELECT landing_outcome, COUNT(*) AS number_of_launches FROM spacex WHERE DATE BETWEEN
```

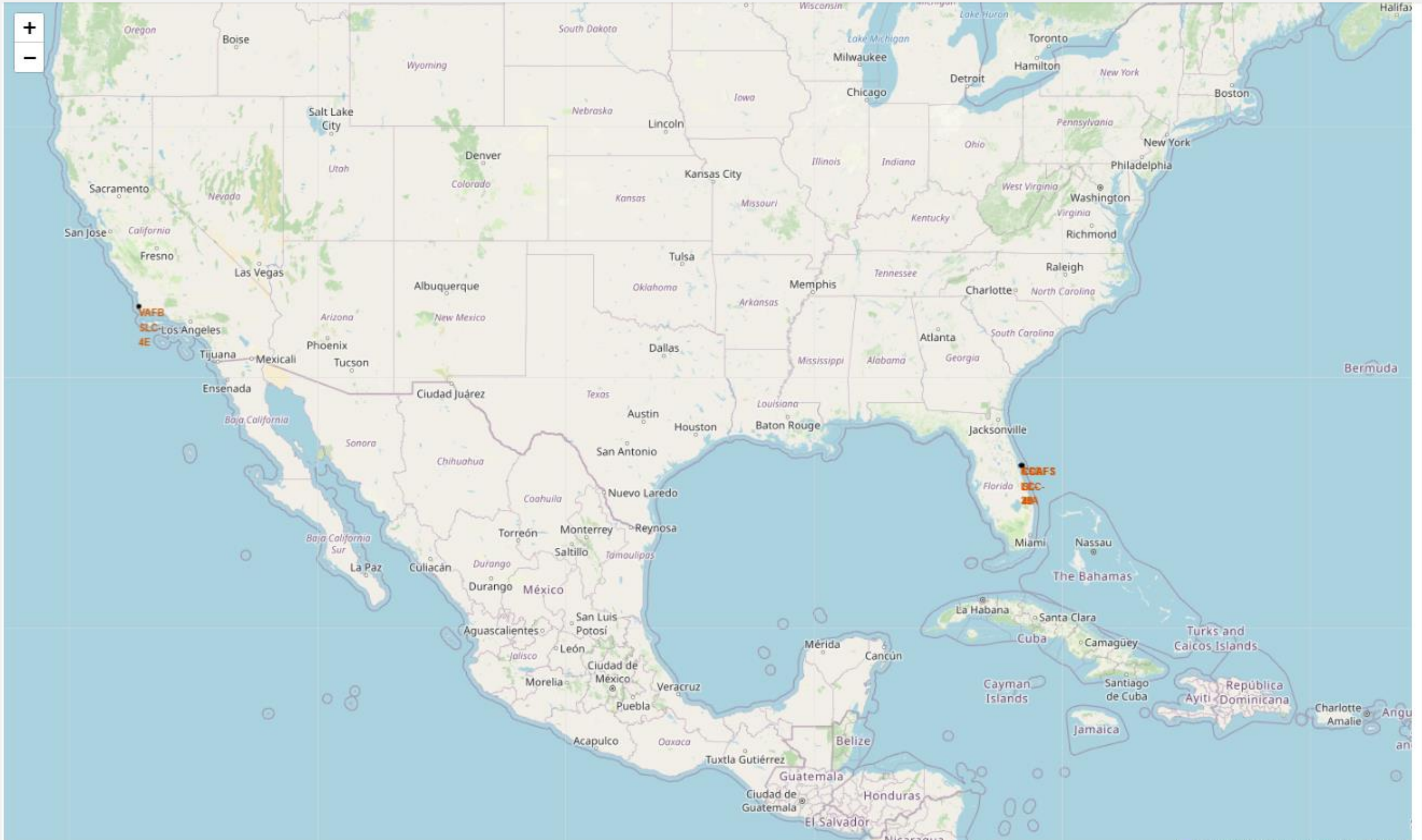
```
 * ibm_db_sa://sxd36939:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databa
ses.appdomain.cloud:31505/bludb
Done.
```

landing_outcome	number_of_launches
Failure (drone ship)	5
Success (ground pad)	3

Interactive Maps using Folium

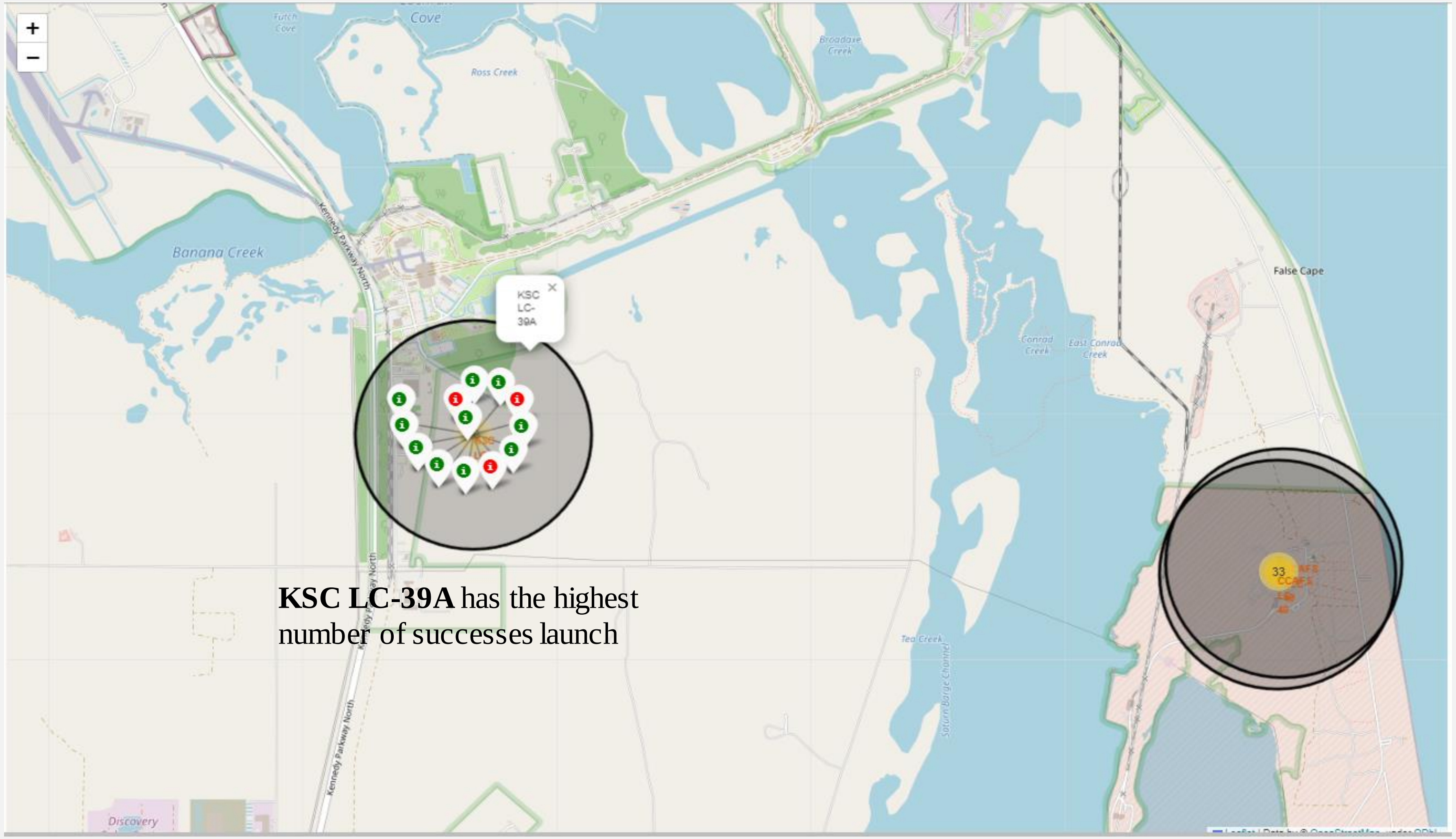
Required interactive map with Folium results
Observation 1

All launch sites on map



Interactive Maps using Folium

Required interactive map with Folium results
Observation 2



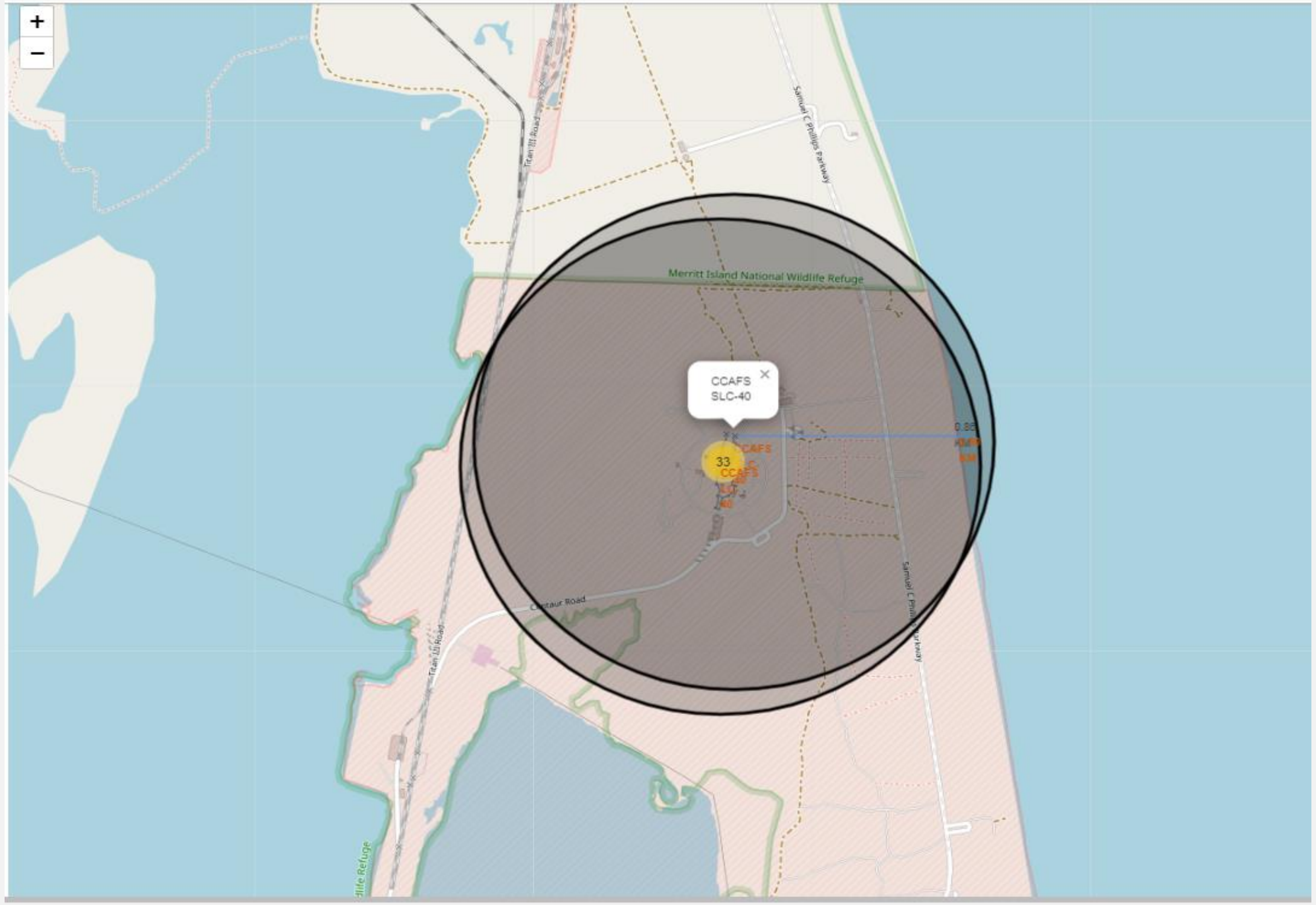
KSC LC-39A has the highest number of successes launch

Interactive Maps using Folium

Required interactive map with Folium results

Observation 3

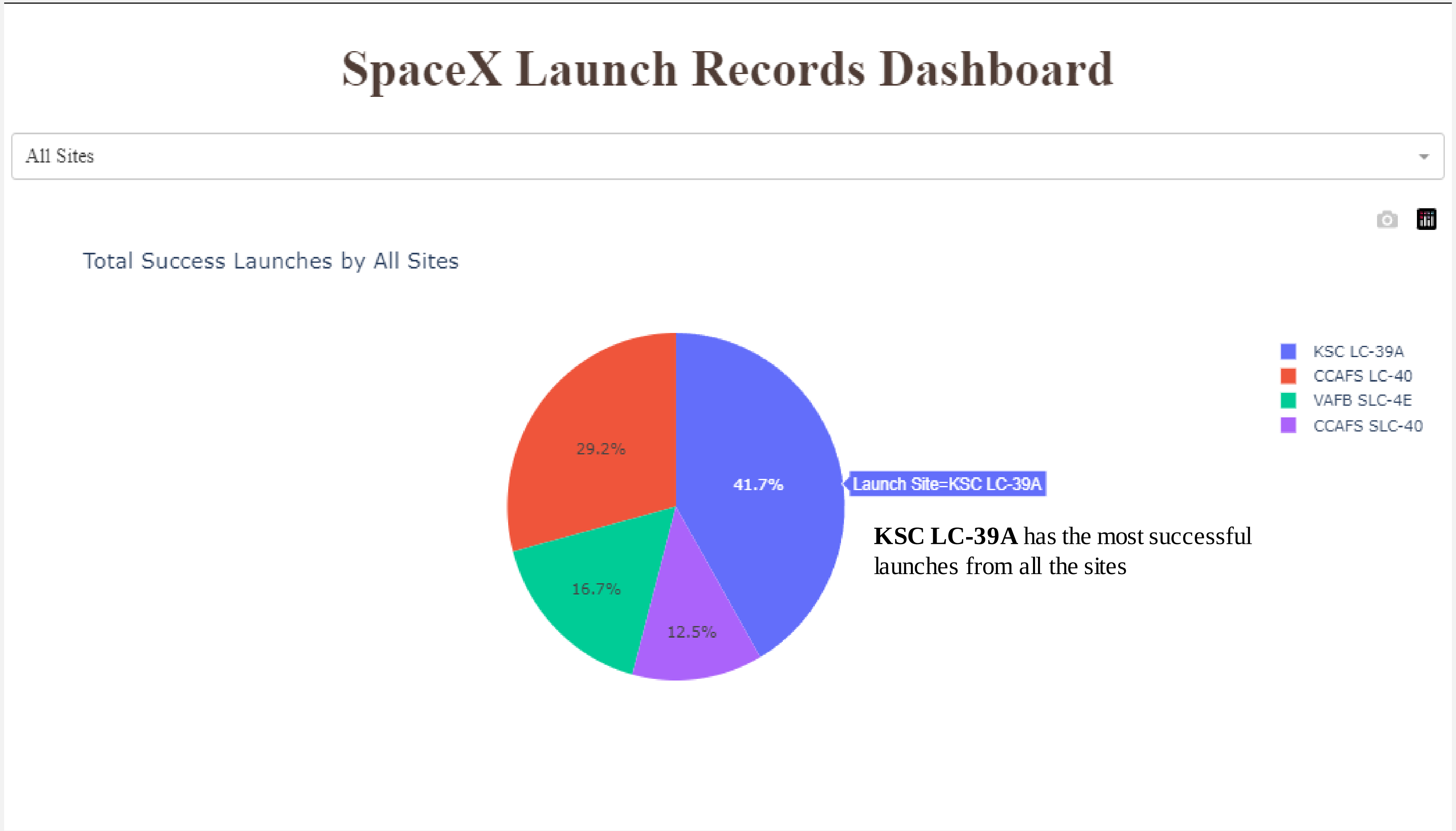
Calculated distance between the CCAFS SLC-40 launch site and the nearest coastline



+
−
Merritt Island National Wildlife Refuge
Samuel C Phillips Parkway
Titan III Road
Centaur Road
CCAFS
SLC-40
33
CCAFS
0.88

Pie Chart with Plotly Dash

Required Plotly Dash results
Observation 1



SpaceX Launch Records Dashboard

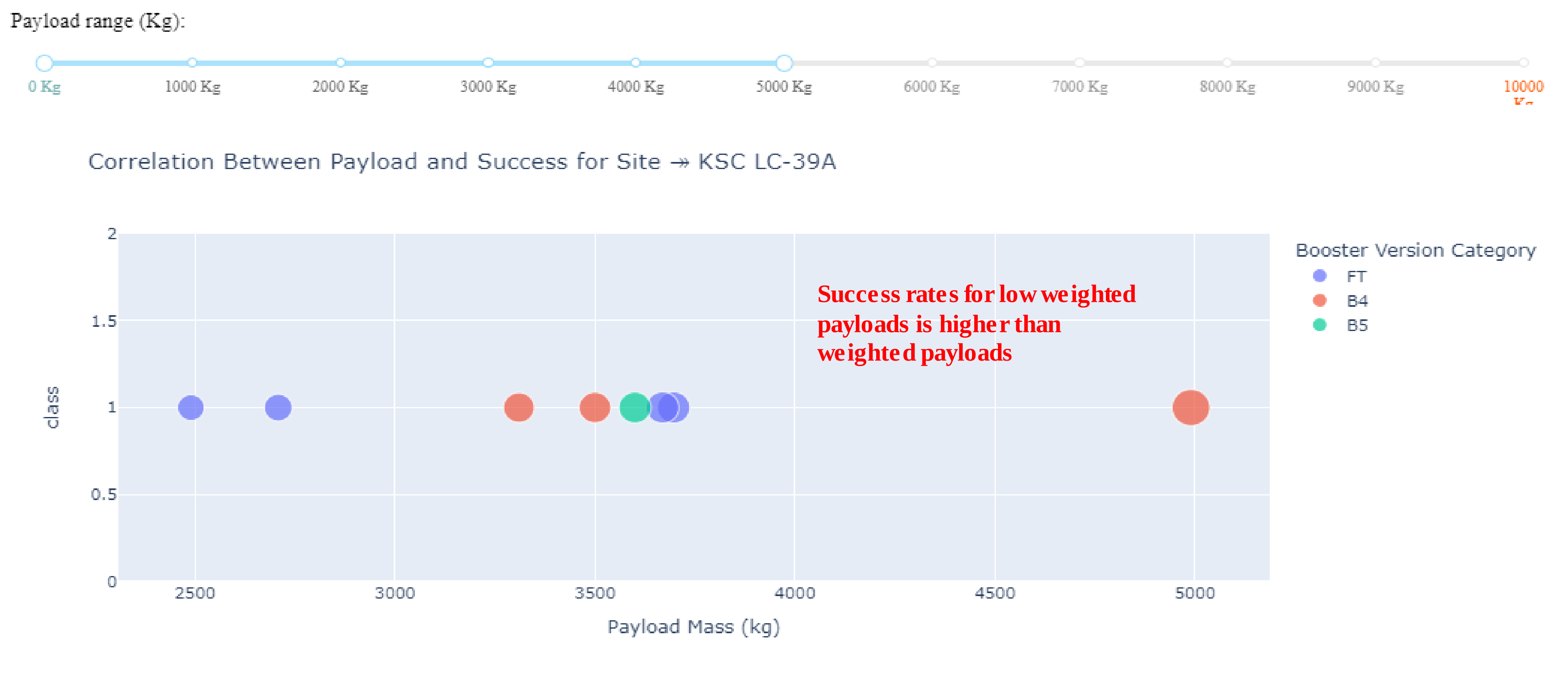
All Sites

KSC LC-39A has the most successful launches from all the sites

Scatter plot with Plotly Dash

Required Plotly Dash results
Observation 2

Payload range 0kg –5000kg

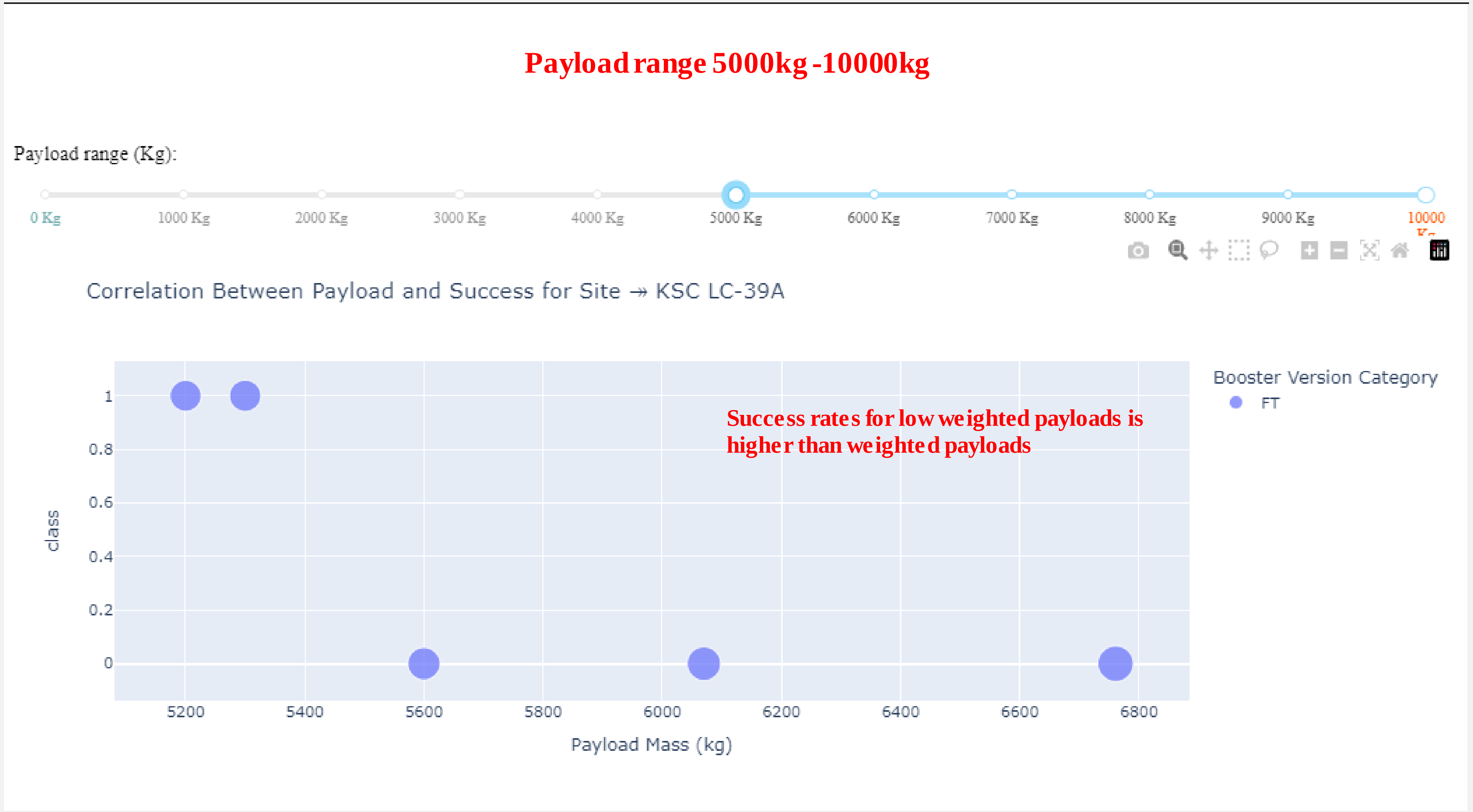


Payload range (Kg):

Success rates for low weighted payloads is higher than weighted payloads

Scatter plot with Plotly Dash

Required Plotly Dash results
Observation 3



Payload range 5000kg -10000kg

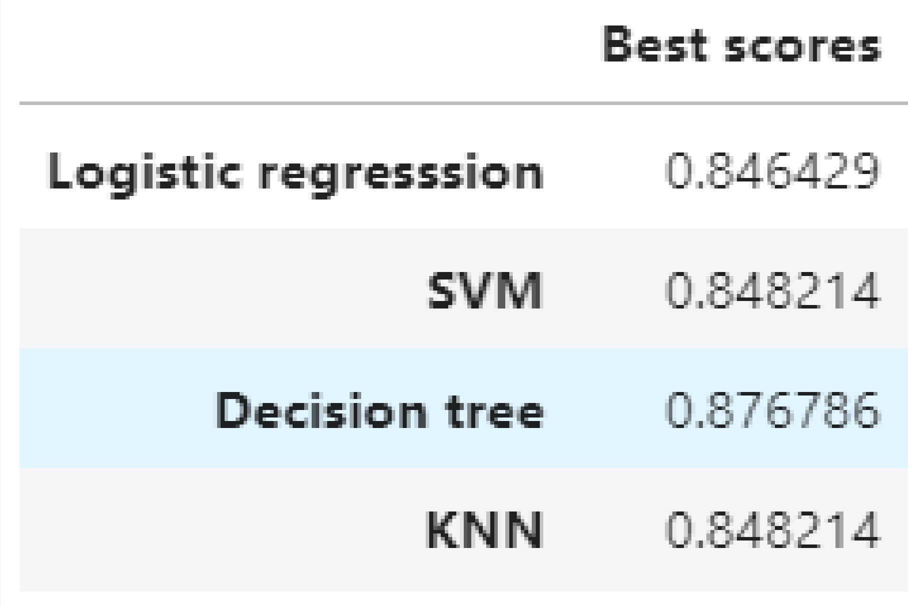
Payload range (Kg):

Success rates for low weighted payloads is higher than weighted payloads

Classification Accuracy

Required predictive analysis results
Observation 1

✓ The Decision Tree classifier is the model with the highest classification accuracy.



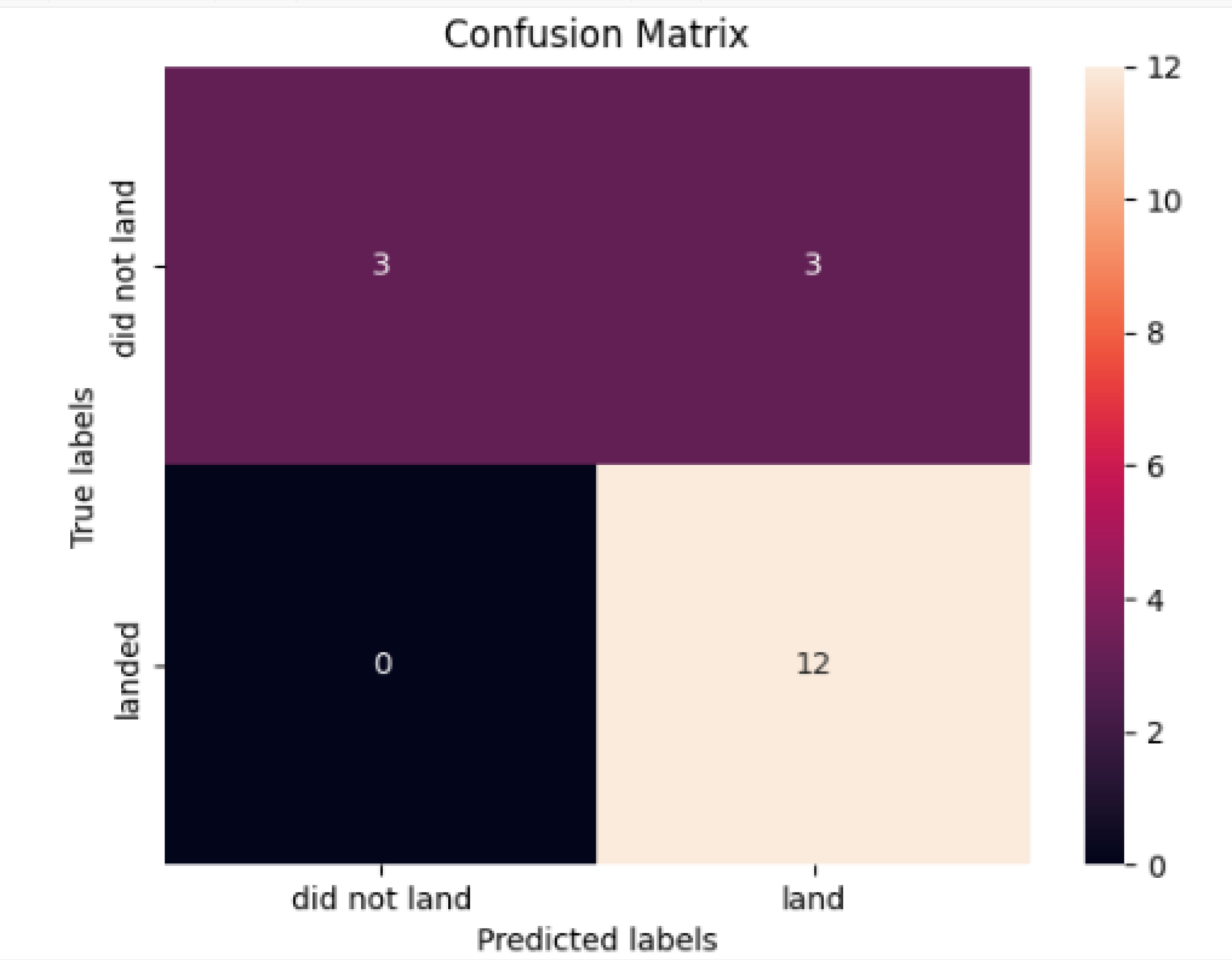
	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.876786
KNN	0.848214

Predictive Analysis

Decision tree

Required predictive analysis results
Observation 2

- ✓ Confusion Matrix is used to measure the performance of the classification models.
- ✓ Examining the confusion matrix of the Decision Tree model, we see that logistic regression can distinguish between the different classes.
- ✓ We see that the major problem is false positives, unsuccessful landing marked as successful landing by the classifier.
- ✓ The confusion matrix of the other models are identical in terms of prediction.



Conclusions

In conclusion, we have successfully analyzed the dataset and built a model that can predict the final mission Outcome based on its features. We have also identified the most important features that impact the mission Outcome. The Decision Tree model has an accuracy of 87%, which is good enough for practical use..

Appendix

Links to the notebooks:

Data Collection using SpaceX API:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-Data Collection-API.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Data Collection using Web Scraping:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-Web Scraping-from-Wikipedia.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Data Wrangling:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-Data Wrangling.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Exploratory Data Analysis with Visualization:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-EDA-Vizualization-Matplotlib-Pandas.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Exploratory Data Analysis using SQL:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-eda-sql.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Predictive Analysis, Classification Models:
SpaceX-Falcon-9-first-stage-Landing-Prediction/SpaceX-Machine_Learning-Prediction.ipynb at master · dennisgue/SpaceX-Falcon-9-first-stage-Landing-Prediction (github.com)

Thank you!