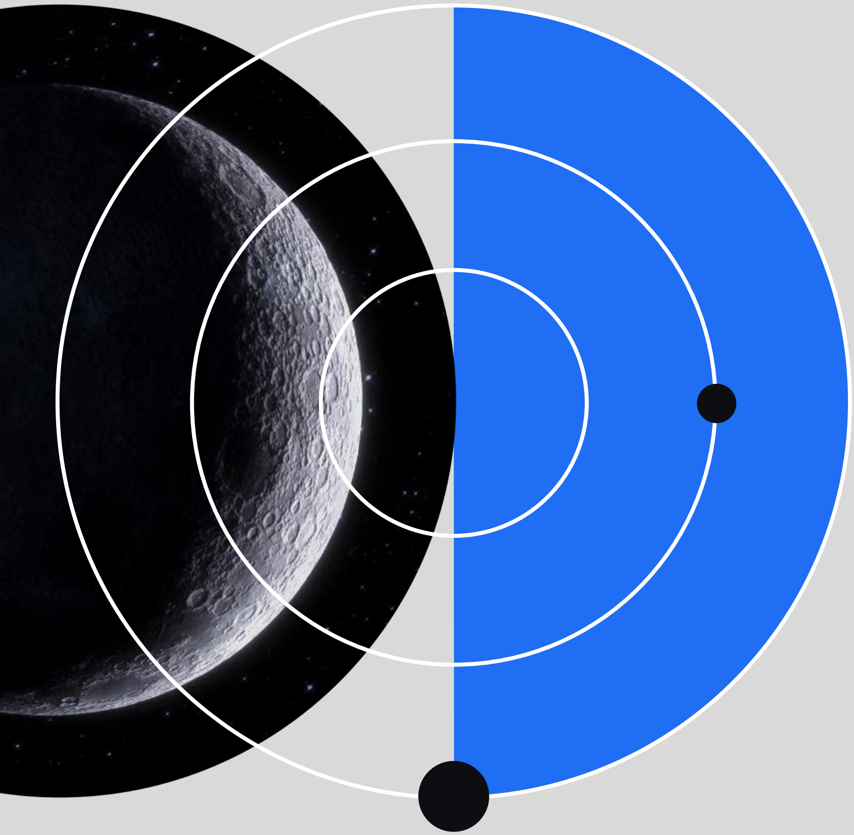


# WINNING SPACE RACE

## WITH DATA SCIENCE

Name: Nhat Minh Nguyen

Date: 7/02/2023

# TABLE OF CONTENTS

## 01 Executive summary

- Summary of methodologies
- Summary of all result

## 02 Introduction

- Project background and context
- Question to be answered

## 03 Methodology

- Data collection
- EDA
- Interactive map, dashboard
- Predictive analysis

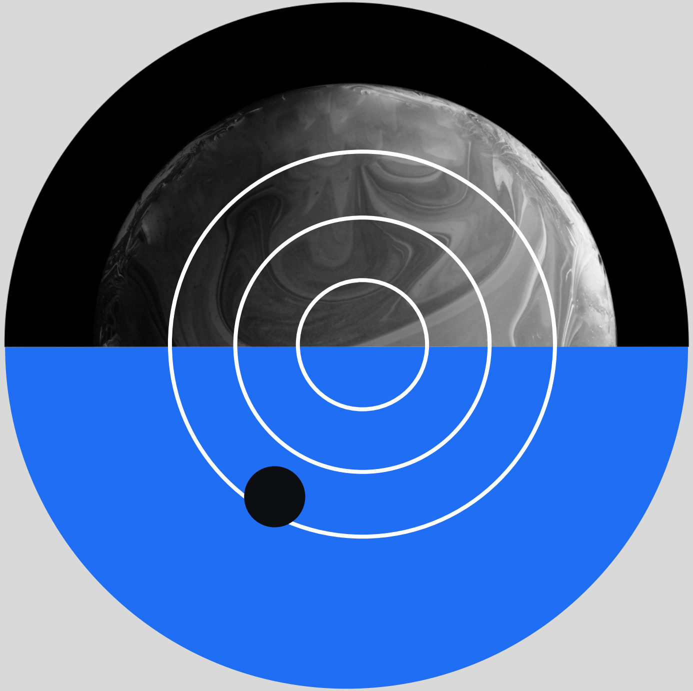
## 04 Result

- EDA result
- Interactive analytics demo
- Predictive analysis result

## 05 Conclusion

# Executive summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory data analysis (EDA) with SQL
  - Exploratory data analysis with visualization
  - Visual analytics and dashboard
  - Predictive analysis using classification
- Summary of all results
  - EDA result
  - Predictive analysis result



# Introduction

## ❖ Project background and context

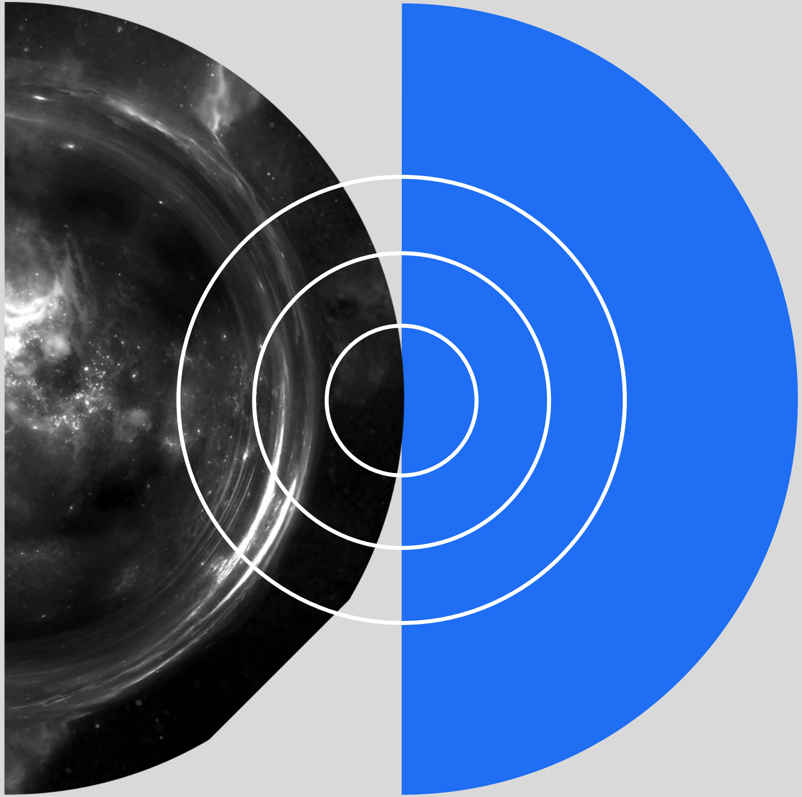
SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. If we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if the first stage of the Falcon 9 will land or not.

## ❖ Problems to find answers

- ✓ How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- ✓ Does the rate of successful landings increase over the years?
- ✓ What is the best model that can be used for the classification task?

## SECTION 1

# METHODOLOGY



# Methodology

- **Data collection methodology**
  - Data was collected from public source (Wikipedia) using web scrapping
- **Data wrangling**
  - Filtering the data
  - Handle missing values
  - Apply One hot encoding on 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**
  - Preprocessing the data
  - Find best hyperparameter using Grid Search
  - Evaluate models performance

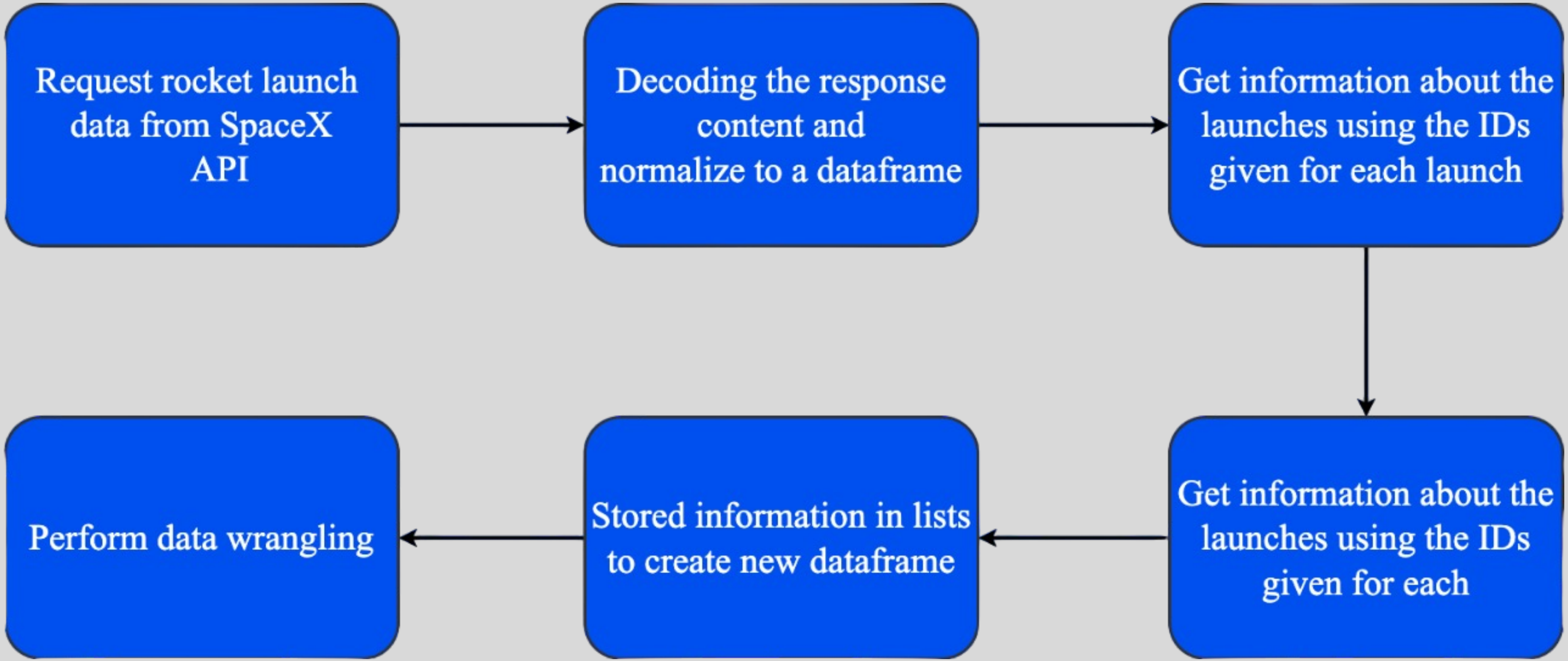
# Data collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

- **Data Columns are obtained by using SpaceX REST API:** FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- **Data Columns are obtained by using Wikipedia Web Scraping:** Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

# Data collection – SpaceX API



Request rocket launch data from SpaceX API

Decoding the response content and normalize to a dataframe

Get information about the launches using the IDs given for each launch

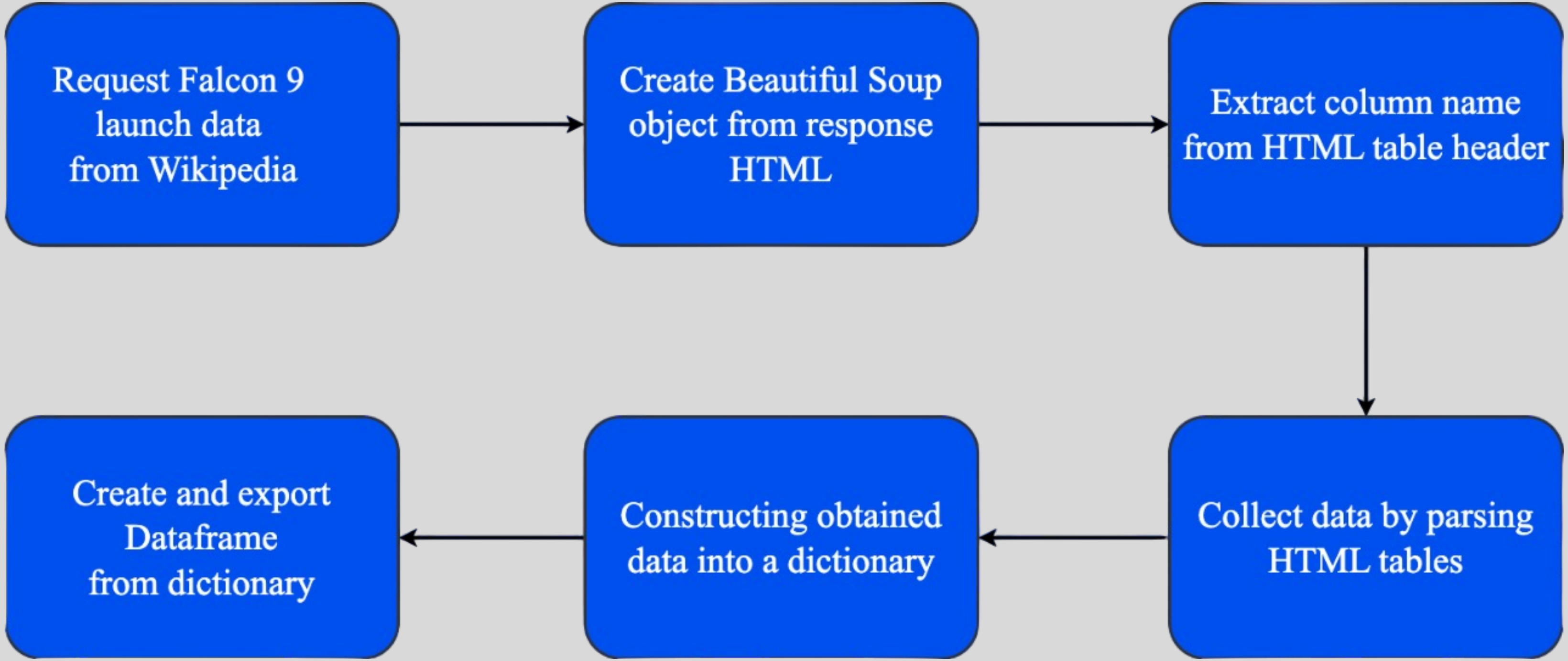
Get information about the launches using the IDs given for each

Stored information in lists to create new dataframe

Perform data wrangling

*Data collection - API*

# Data collection – Web scrapping



Request Falcon 9 launch data from Wikipedia

Create Beautiful Soup object from response HTML

Extract column name from HTML table header

Collect data by parsing HTML tables

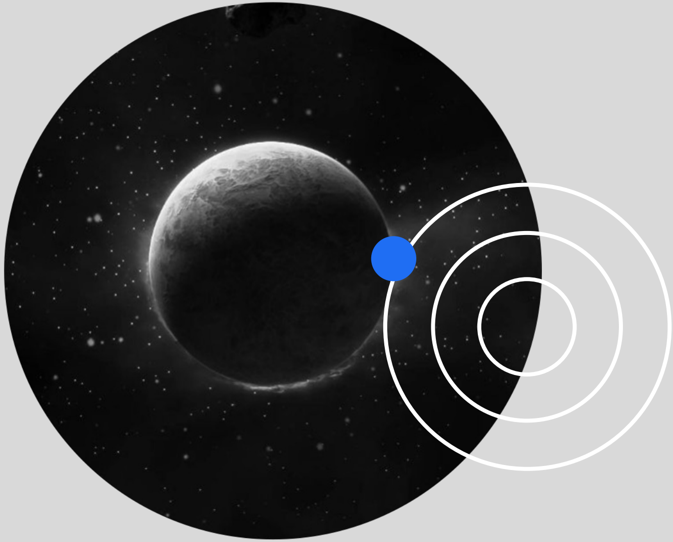
Constructing obtained data into a dictionary

Create and export Dataframe from dictionary

*Data collection - Web scrapping*

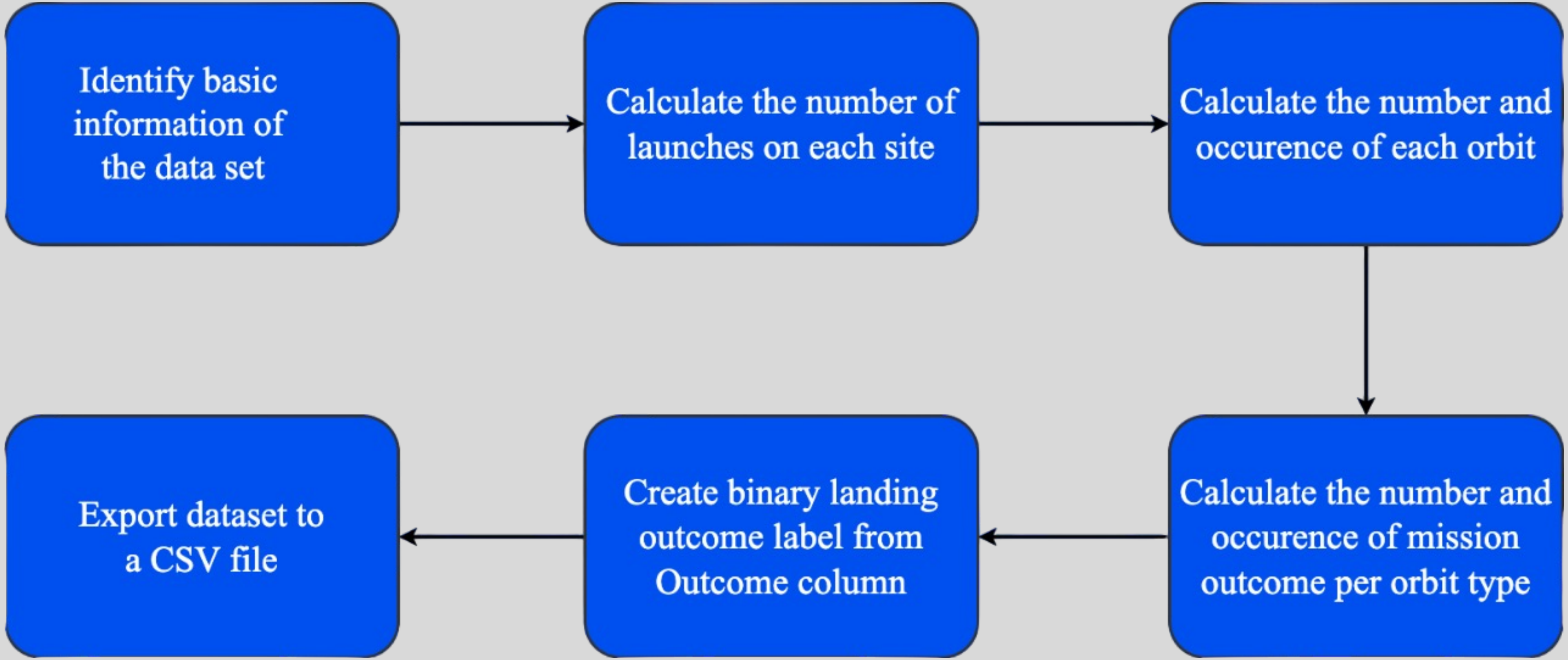
# Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with "1" means the booster successfully landed, "0" means it was unsuccessful.



*Data wrangling*

# Data Wrangling – Flow chart



Identify basic information of the data set

Calculate the number of launches on each site

Calculate the number and occurence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create binary landing outcome label from Outcome column

Export dataset to a CSV file

*Data wrangling*

# EDA with data visualization

## Scatter plot

To check **relationship** between variables

## Bar chart

To present **frequency** of categorical variable

## Line chart

To show **trend** in data overtime

### Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

*EDA with visualization*

# EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List 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
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions and launch site names for the months 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

*EDA with SQL*

# Build an interactive map with Folium

❖ **Markers of all Launch Sites:**

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

❖ **Colored Markers of the launch outcomes for each Launch Site:**

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

*Interactive visual analytics*

# Build a dashboard with Plotly Dash

- **Launch Sites Dropdown List**

  Added a dropdown list to enable Launch Site selection.

- **Pie Chart showing Success Launches (All Sites/Certain Site)**

  Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

- **Slider of Payload Mass Range**

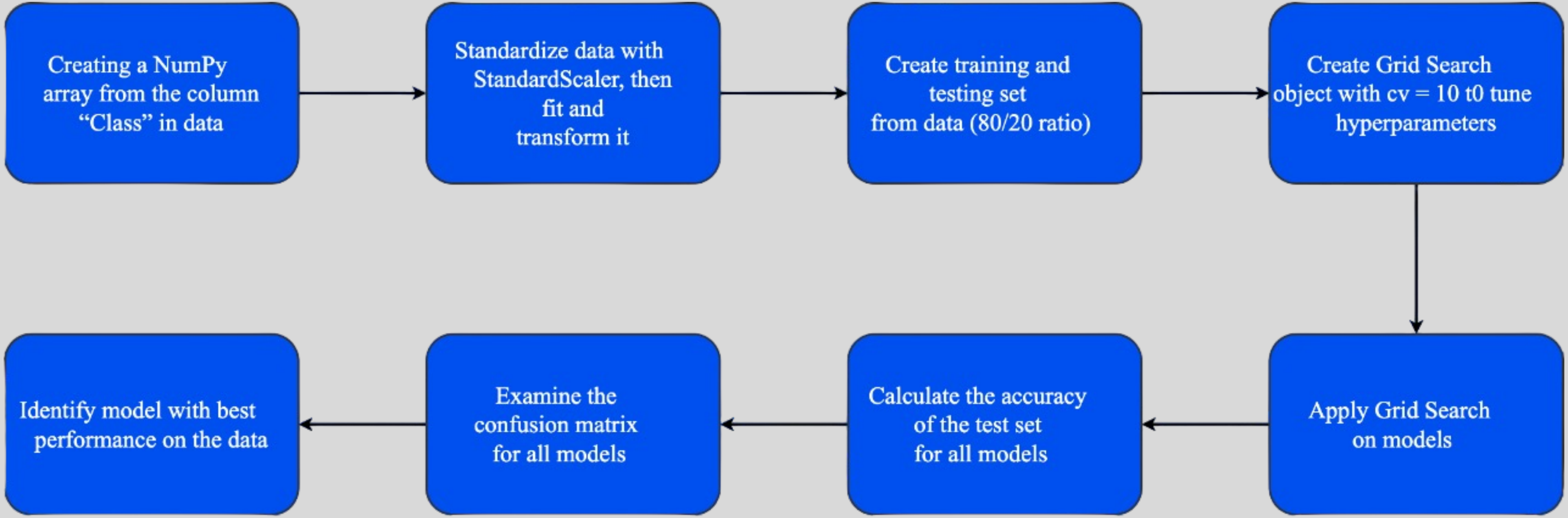
  Added a slider to select Payload range.

- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions**

  Added a scatter chart to show the correlation between Payload and Launch Success.

***Dashboard***

# Predictive analysis - Classification



Creating a NumPy array from the column "Class" in data

Standardize data with StandardScaler, then fit and transform it

Create training and testing set from data (80/20 ratio)

Create Grid Search object with cv = 10 t0 tune hyperparameters

Apply Grid Search on models

Calculate the accuracy of the test set for all models

Examine the confusion matrix for all models

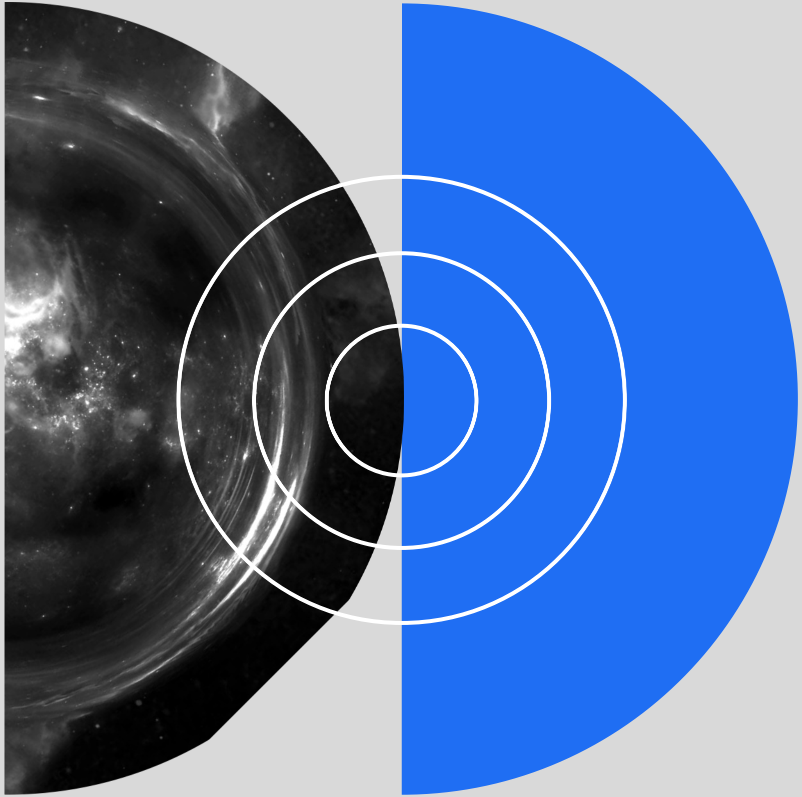
Identify model with best performance on the data

*Machine learning prediction (Classification)*

# SECTION 2

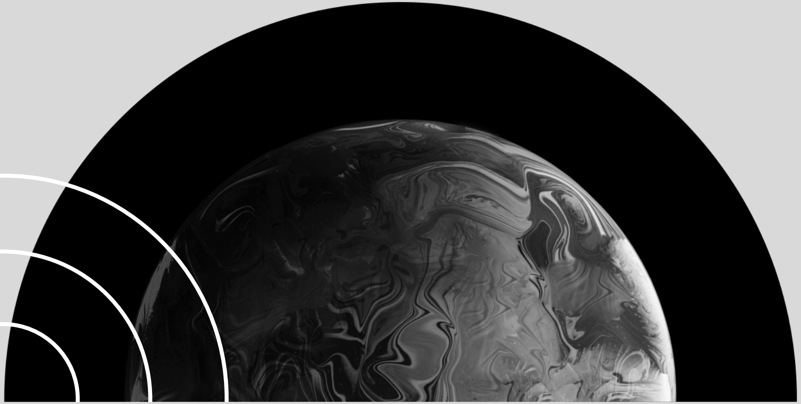
# RESULT

- Insights drawn from EDA
- Interactive analytics demo in screenshots
- Predictive analysis results



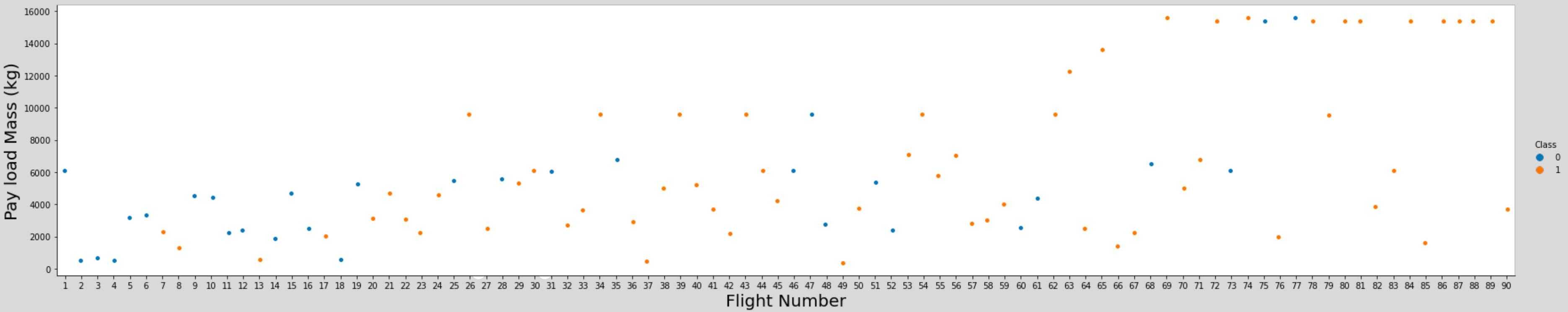
# Insights drawn from EDA

- EDA with visualization
- EDA with SQL



# Visualization Result

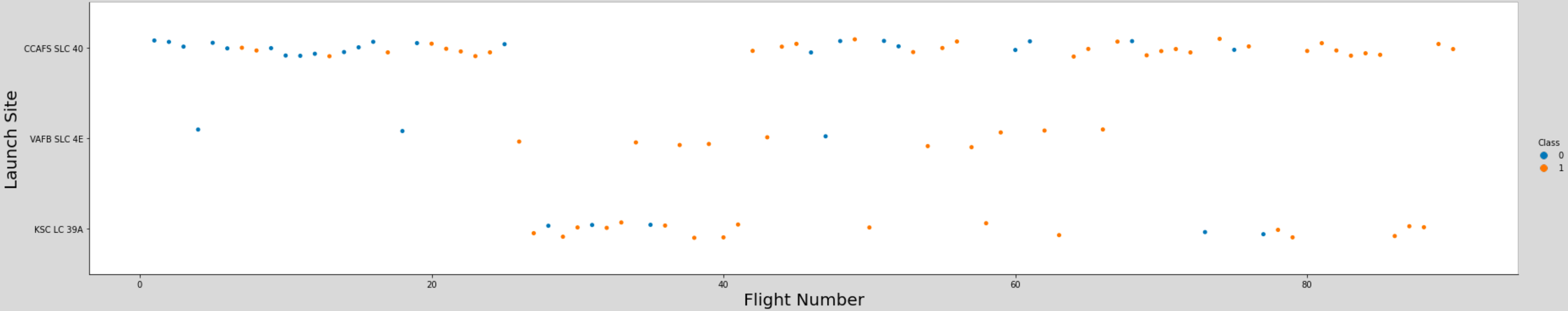
## Flight number vs Payload mass



As Flight Number increase, the first stage is more likely to land successfully. Also the more massive the payload, the less likely the first stage will return

# Visualization Result

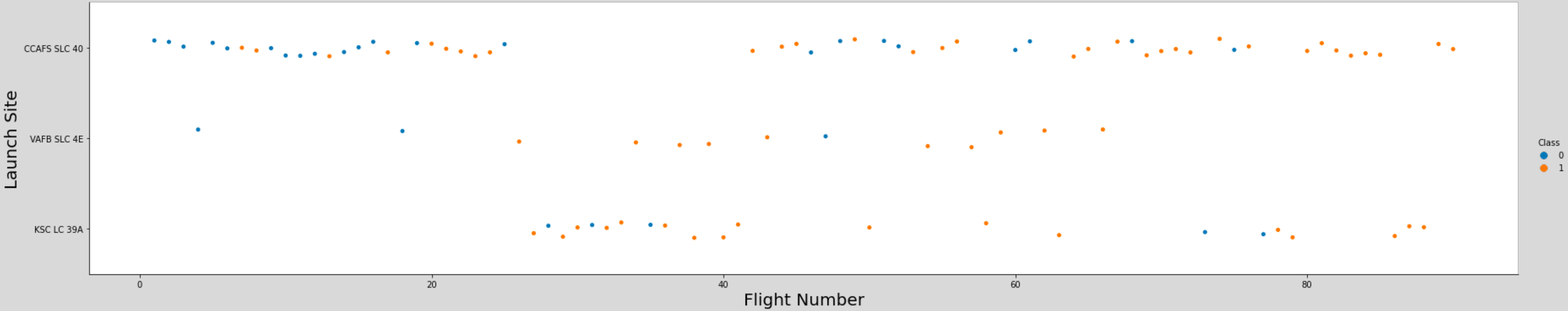
## Flight number vs Launch Site



- The earliest flights failed all while the latest flights succeced all
- The CCAFS SLC 40 launch site has about half of all launches.
- It can be assumed that each new launch has a higher rate of success.

# Visualization Result

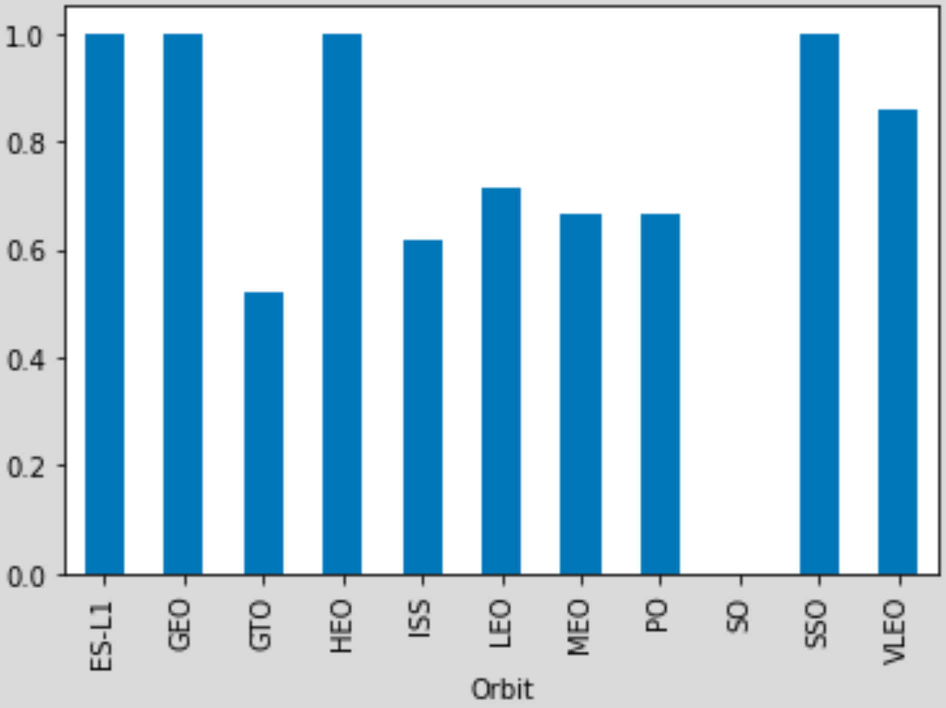
## Payload vs Launch Site



- For every launch site the higher the payload mass, the higher the success rate.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg

# Visualization Result

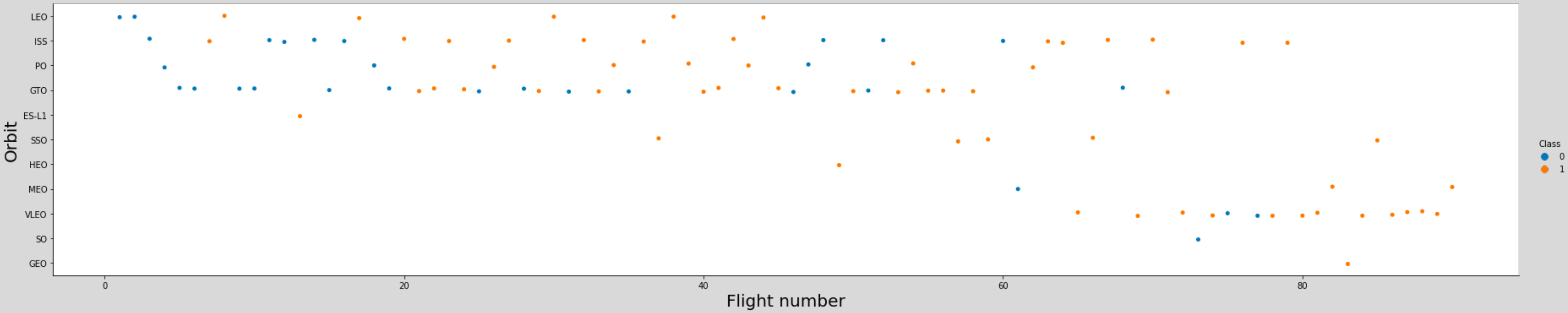
## Success rate vs Orbit type



- Orbit with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbit with 0% success rate: SO

# Visualization Result

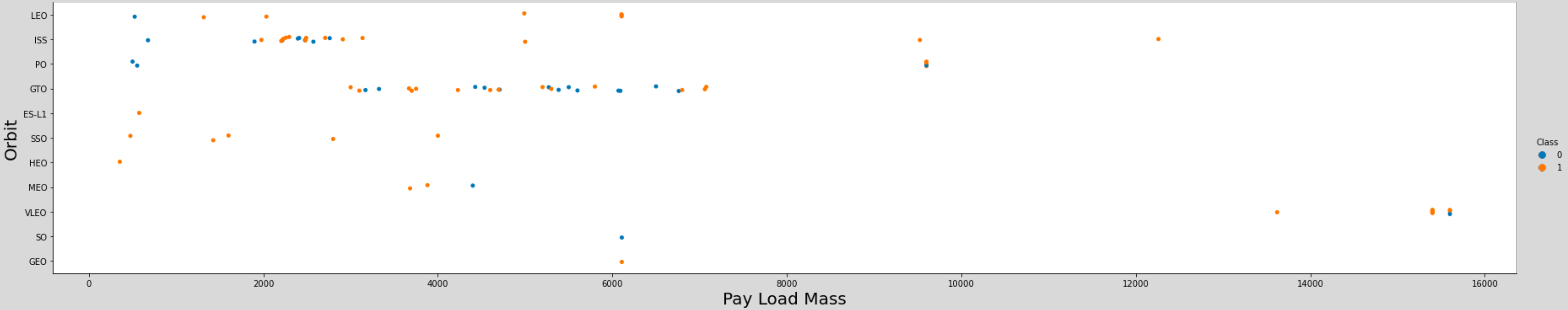
## Flight Number vs Orbit type



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.

# Visualization Result

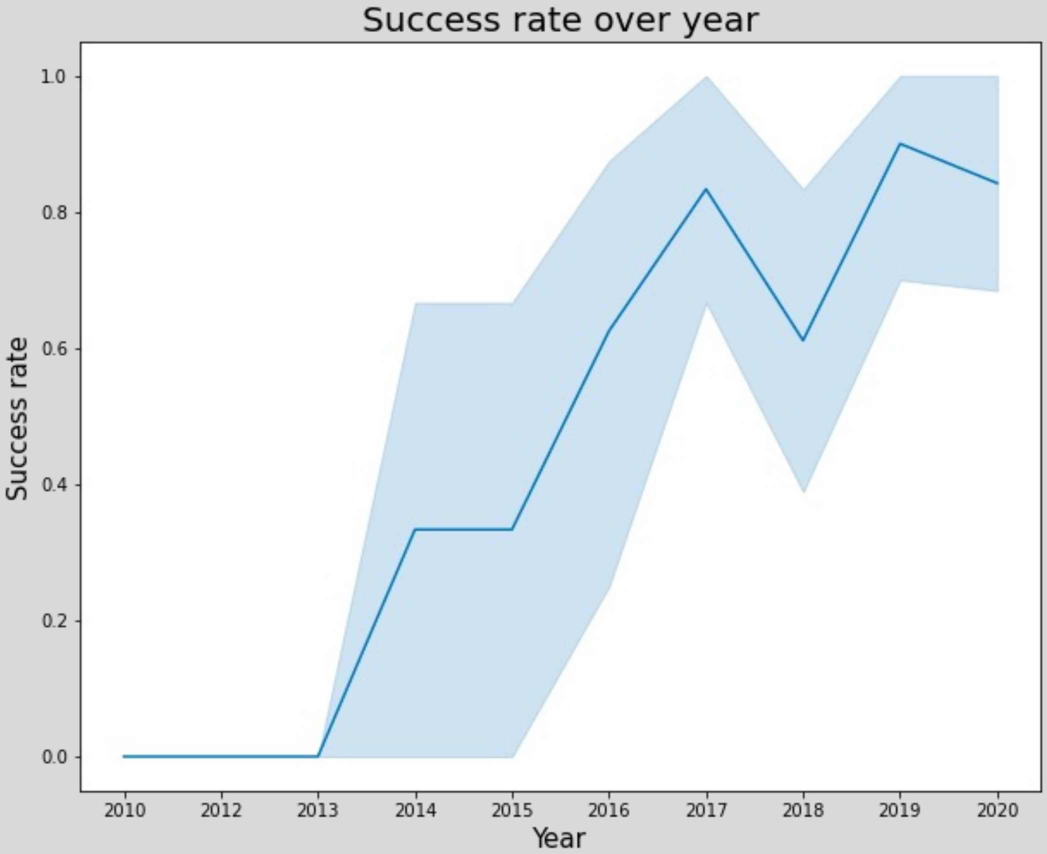
## Pay load mass vs Orbit type



With heavy payloads the successful landing or positive landing rate are more for PO (Negative), LEO and ISS (Positive).

However for GTO we cannot distinguish this well as both positive and negative landing rate are both there.

# Visualization Result

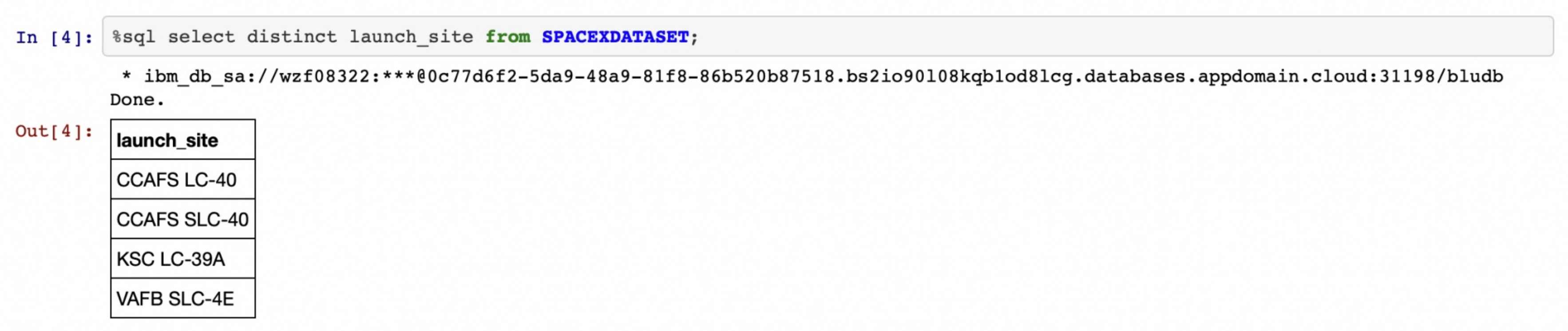


The success rate ***since 2013***

kept increasing till 2020.

# EDA with SQL Result

## All launch sites name



```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

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

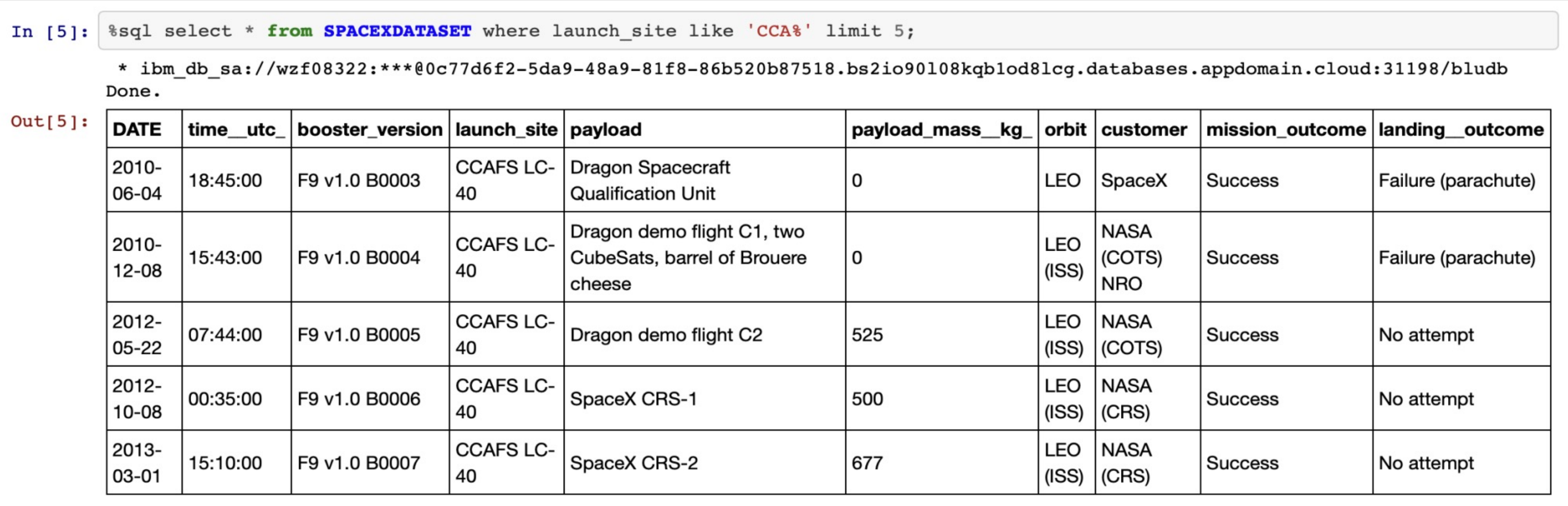
Out[4]:

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

**Explanation:** Display the names of the unique launch site.

# EDA with SQL Result

## Launch site names begin with “CCA”



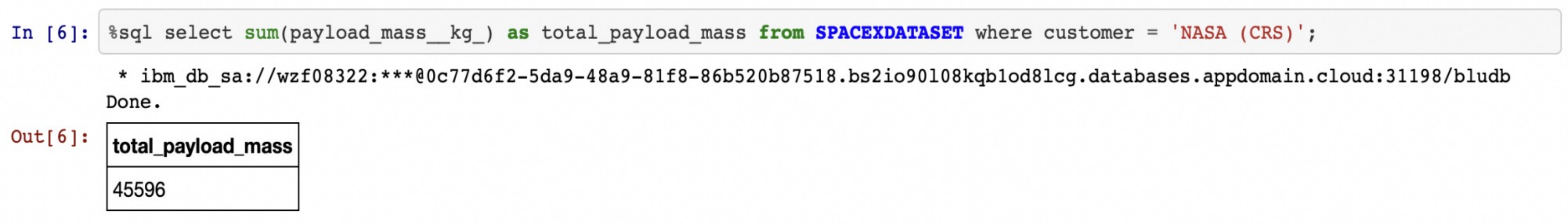
```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

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

| 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 |

**Explanation:** Display 5 records where launch sites begin with the string 'CCA'.

# EDA with SQL Result

## Total payload mass



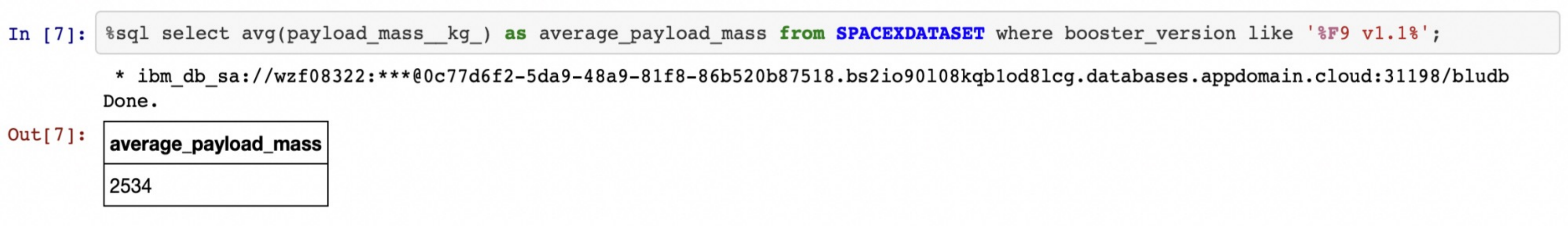
```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

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

Out[6]:
```

| total_payload_mass |
|---|
| 45596 |

**Explanation:** Display the total payload mass carried by boosters launched by NASA (CRS).

# EDA with SQL Result

## Average payload mass by F9 v1.1



```
In [7]: %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/bludb
        Done.
```

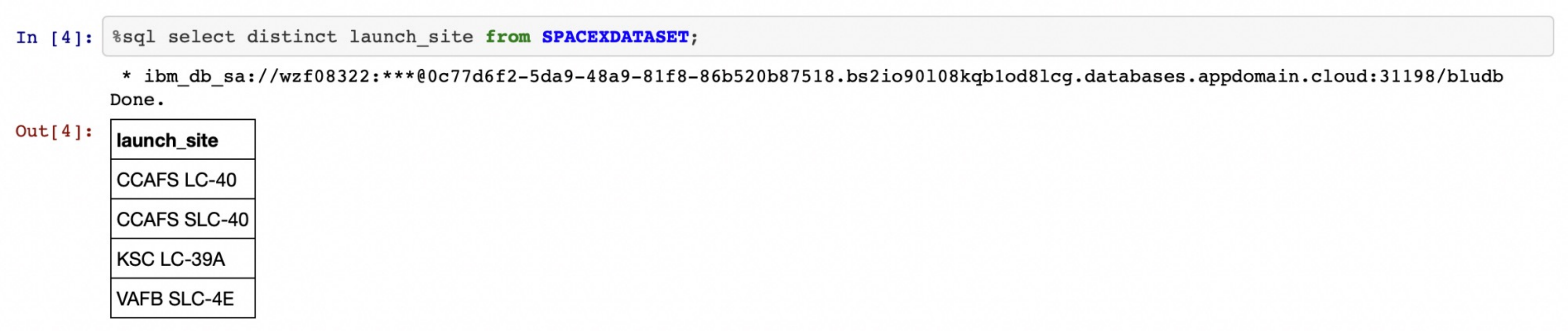
Out[7]:

| average_payload_mass |
|---|
| 2534 |

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

# EDA with SQL Result

## First successful ground landing date



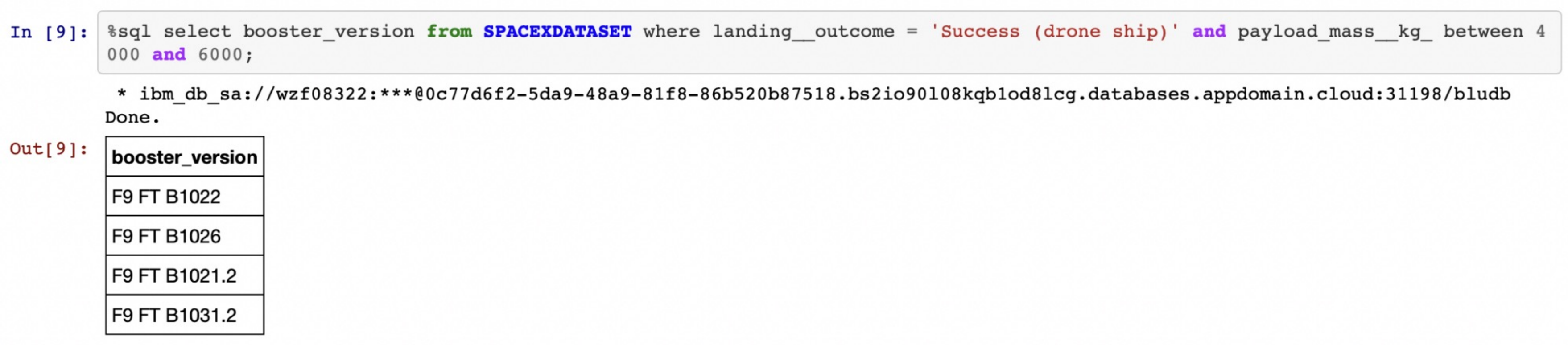
```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[4]:
```

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

**Explanation:** List the date when the first successful landing outcome in ground pad was achieved.

# EDA with SQL Result

## Successful drone ship landing with payload between 4000 and 6000



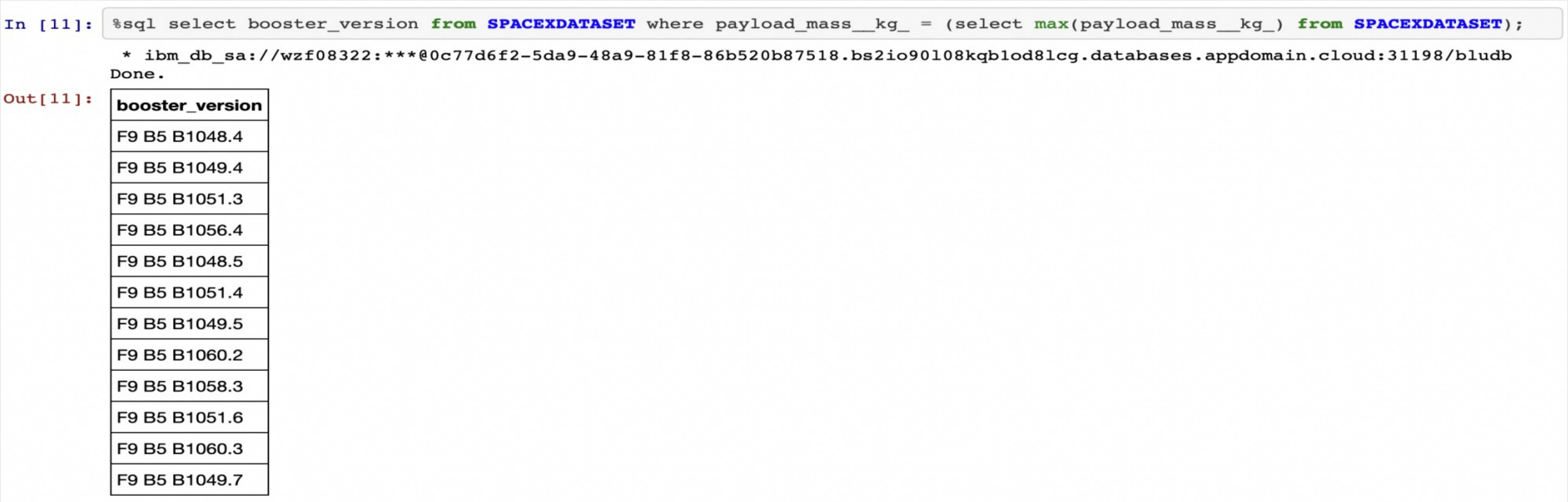
```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[9]:
```

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

**Explanation:** List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

# EDA with SQL Result

## Boosters carried maximum payload



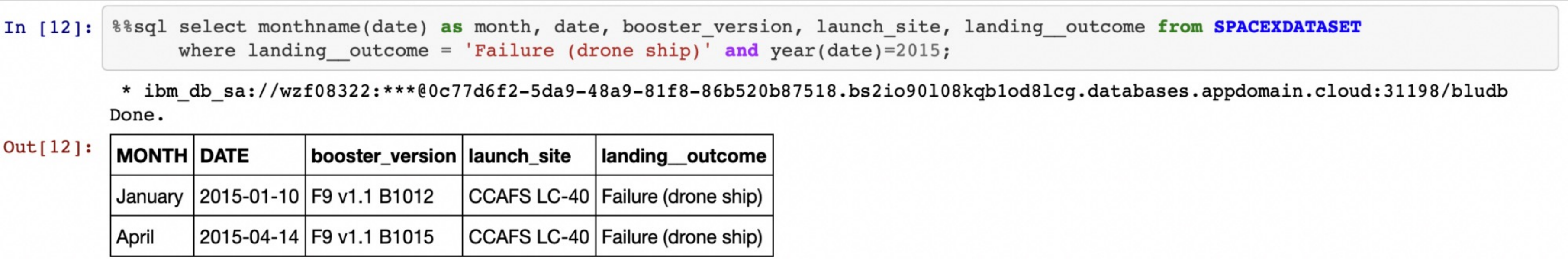
```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[11]:
```

| 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 |

**Explanation:** List the names of the booster versions which have carried the maximum payload mass.

# EDA with SQL Result

## 2015 launch records



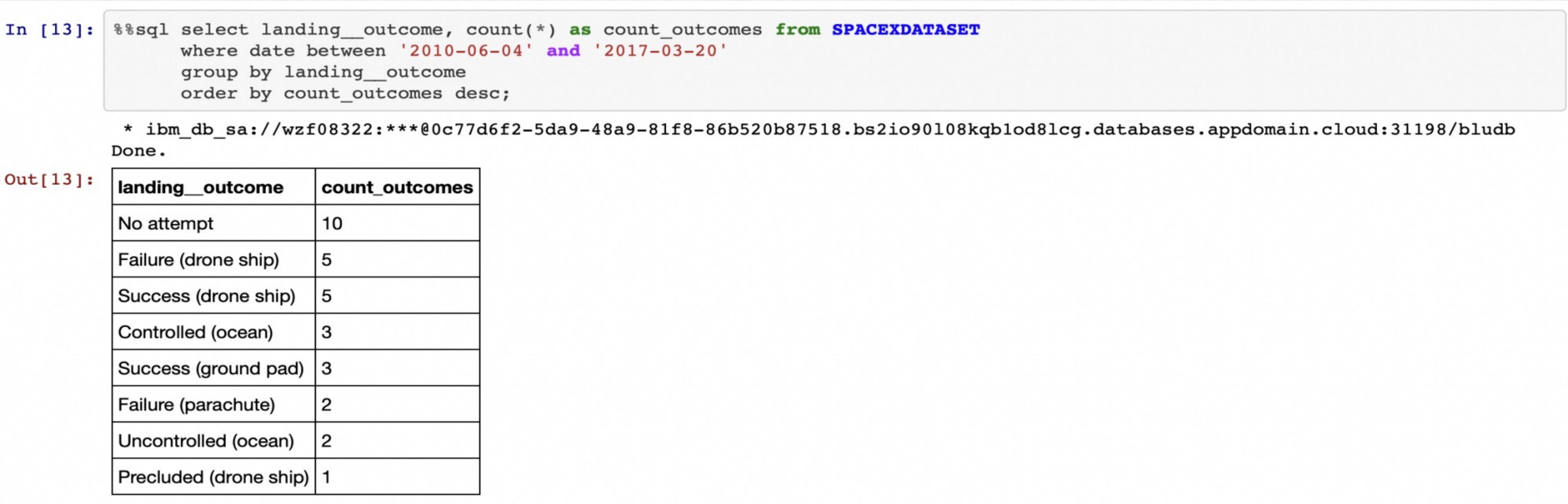
```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[12]:
```

| 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) |

**Explanation:**List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

# EDA with SQL Result

## Rank success count between 2010-06-04 and 2017-03-20



```
In [13]: %%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;

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

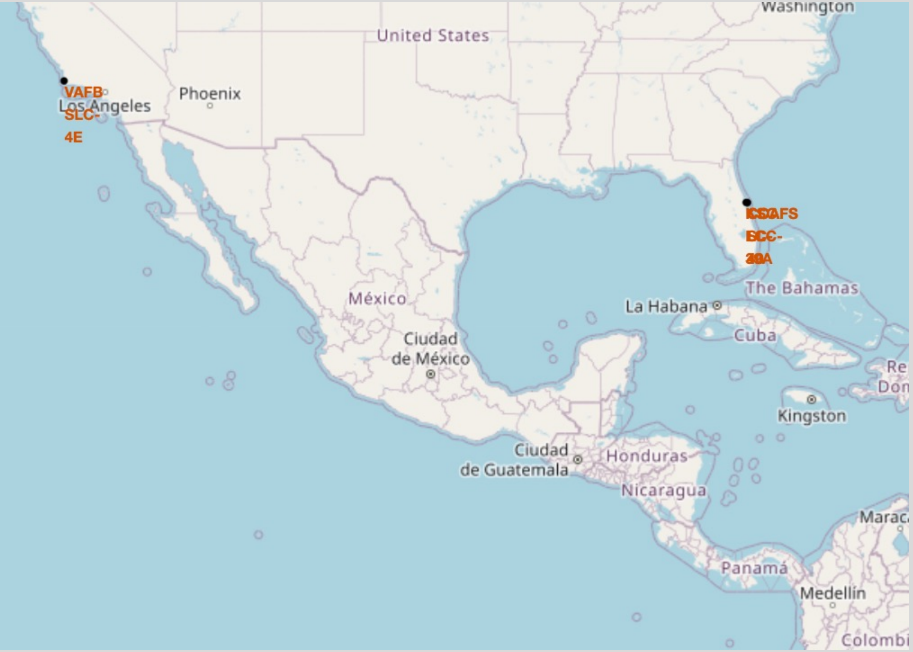
Out[13]:

| 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 |

**Explanation:** 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.

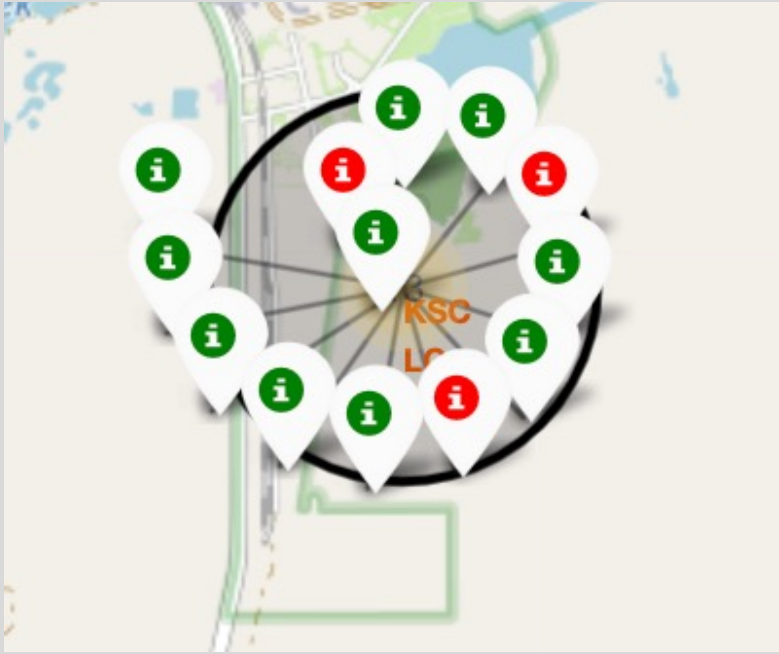
# Interactive analytics with Folium

# All launch sites’ location markers on a global map



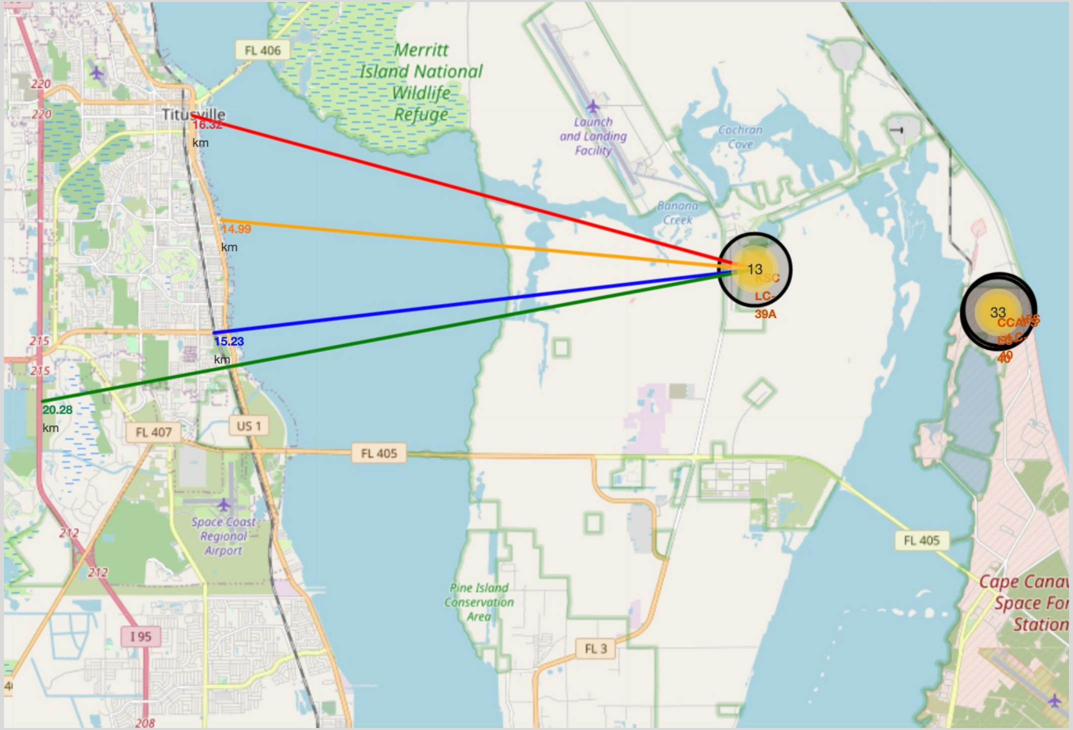
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimize the risk of having any debris dropping or exploding near people.

# Color-labeled launch records on the map



- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates, Green Marker for Successful Launch and Red Marker for Failed Launch.
- Launch Site KSC LC-39A has a very high Success Rate.

# Distance from the launch site KSC LC-39A to its proximities



FL 406
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
Launch and Landing Facility
Cochran Cove
Banana Creek
13
KSC LC-39A
33
CCAFS
40
220
215
FL 407
US 1
FL 405
Space Coast Regional Airport
212
I 95
208
Pine Island Conservation Area
FL 3
Cape Canaveral Space Force Station

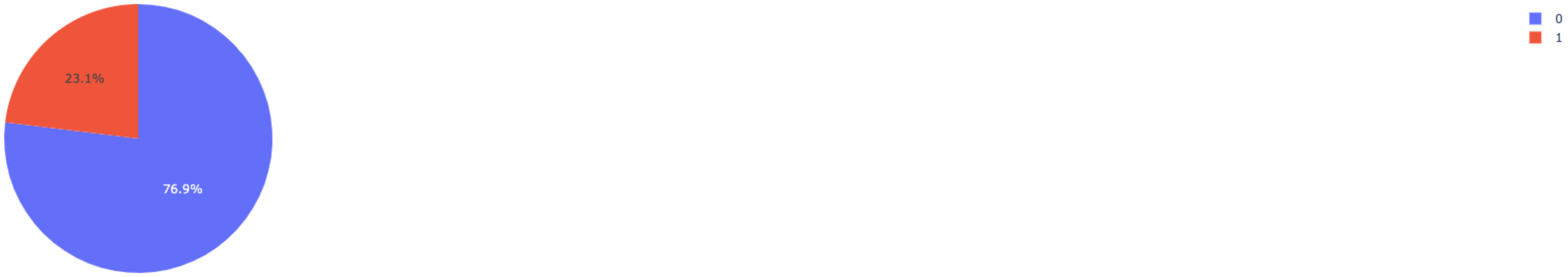
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

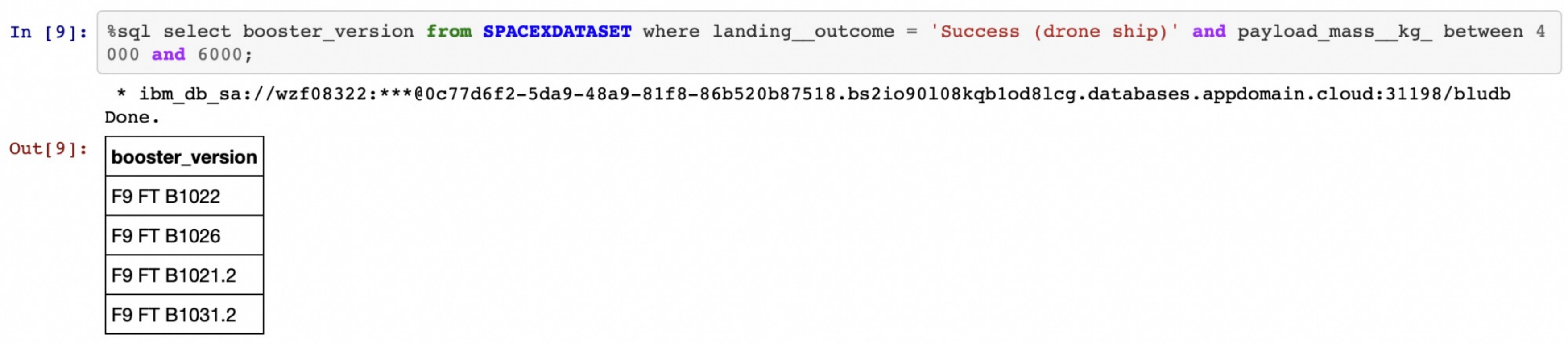
# Build a Dashboard with Plotly Dash

# Launch site with highest launch success ratio



**Explanation:** KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

# Successful drone ship landing with payload between 4000 and 6000



In [9]:

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

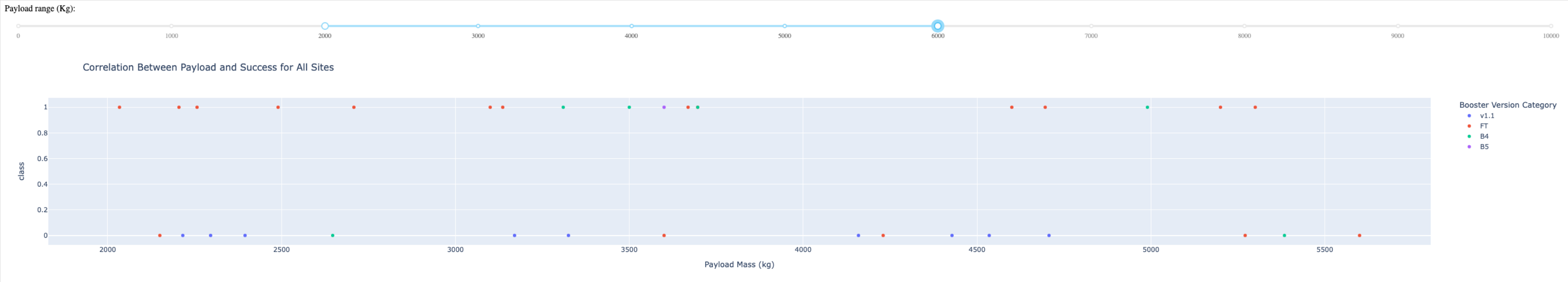
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

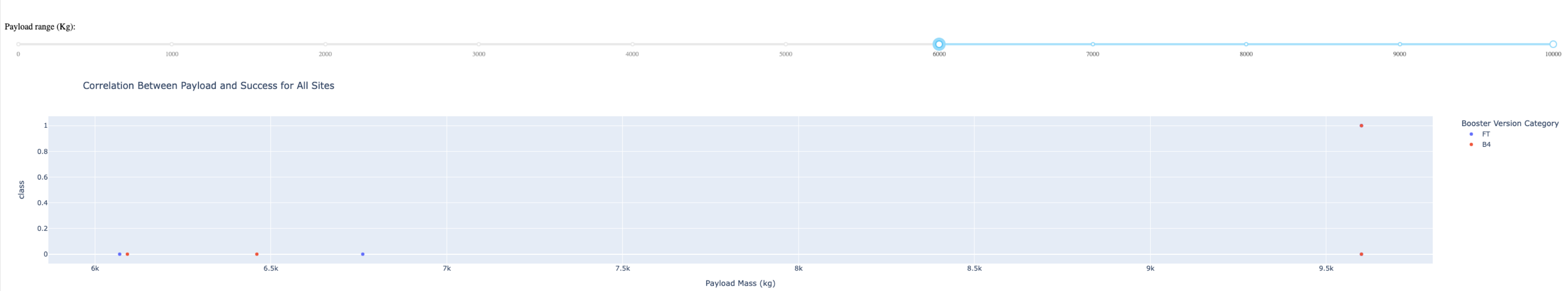
| booster_version |
|---|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

**Explanation:** Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

# Payload mass vs launch outcome for all sites



Payload range (Kg):



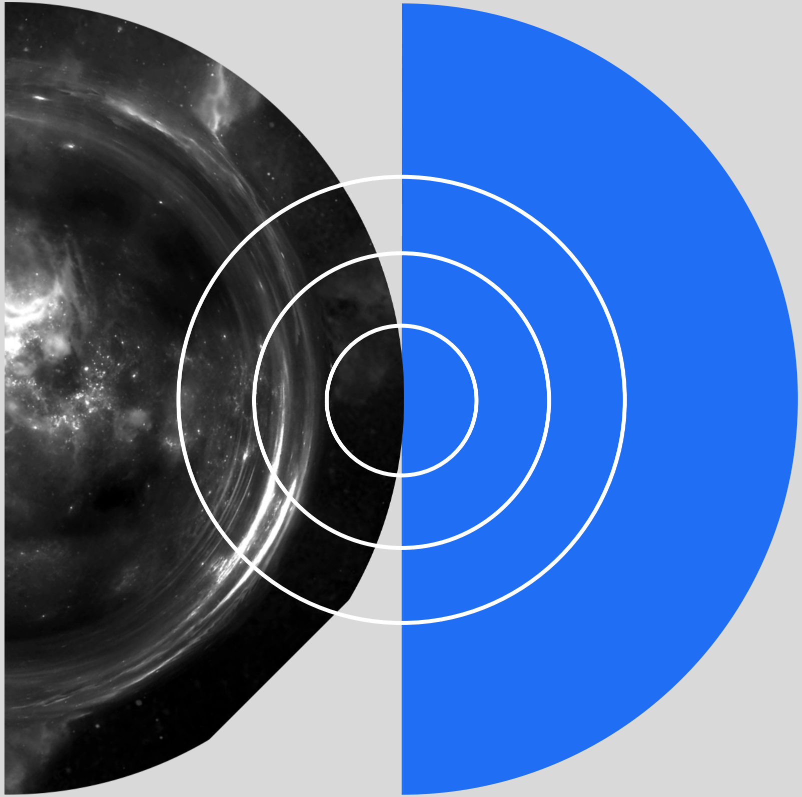
Payload range (Kg):

**Explanation:** Payload between 2000 and 6000 kg have the highest success rate (16), while payload between 6000 and 10000 have the least (1).

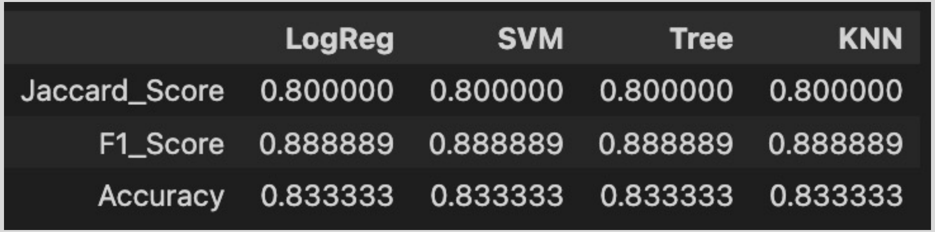
# SECTION 3

# PREDICTIVE ANALYSIS

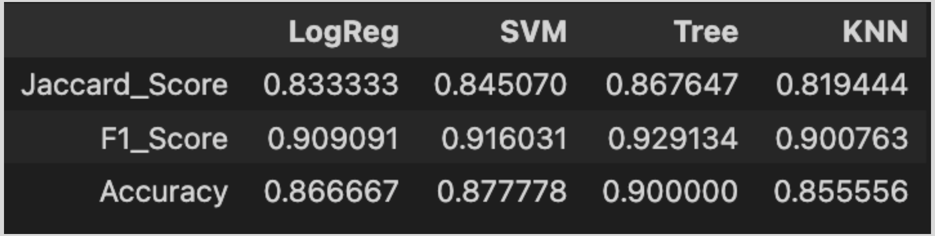
- Classification accuracy
- Confusion matrix



# Classification accuracy



| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy | 0.833333 | 0.833333 | 0.833333 | 0.833333 |



| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.833333 | 0.845070 | 0.867647 | 0.819444 |
| F1_Score | 0.909091 | 0.916031 | 0.929134 | 0.900763 |
| Accuracy | 0.866667 | 0.877778 | 0.900000 | 0.855556 |

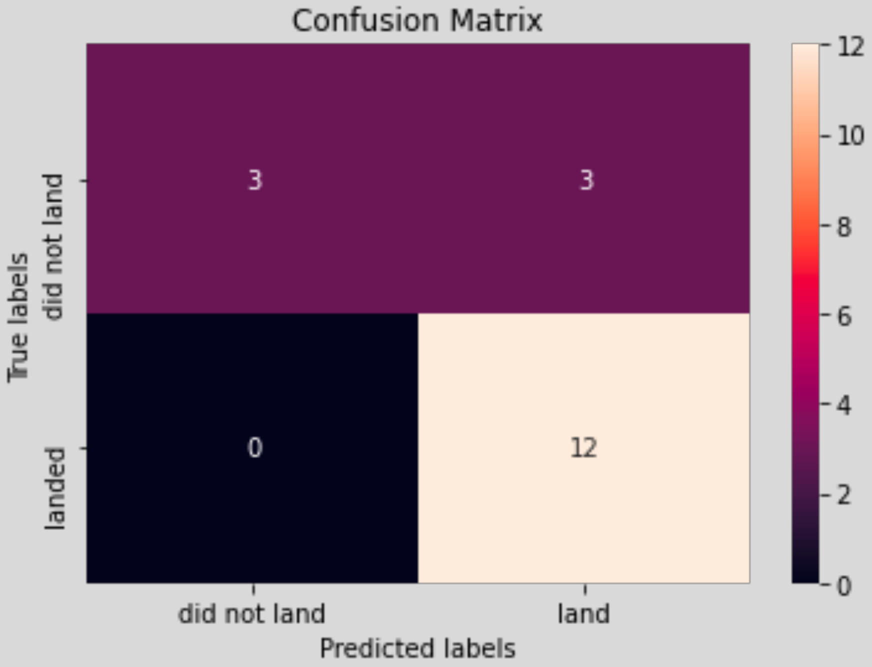
## Score of test set

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

## Score of the whole dataset

- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

# Confusion matrix of decision tree model



- 3 launches predicted not land and actually did not land
- 3 launches predicted to be landed and in fact did not land
- 0 launches were predicted not landed and were actually landed
- 12 launches predicted to be landed and were actually landed

# CONCLUSION

- Decision Tree Model is the best model to predict whether or not the first stage of the Falcon 9 will landed.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- 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.

# Thank you!