# Applied Data Science Capstone Project

Papawin Charoenchaipiyakul

# Executive Summary

- The data collected from SpaceX API and SpaceX Wikipedia. Then create a new label that classifies successful landings. The data explored by using SQL, visualization, folium maps, and dashboards. And use GridSearchCV to find best parameters by gathered relevant columns, create categorical dummies, and standardized data.
- By using 4 models : Logistic regression, SVM, Decision Tree, and KNN. All model accuracy are about 83%.

# Introduction

- Background
  - SpaceX is a provider of space transportation with best pricing due to the ability to recover part of rocket after launch
  - SpaceY is another company that want to compete with SpaceX
- Problem
  - SpaceY want to use machine learning to predict successful landings

# Methodology

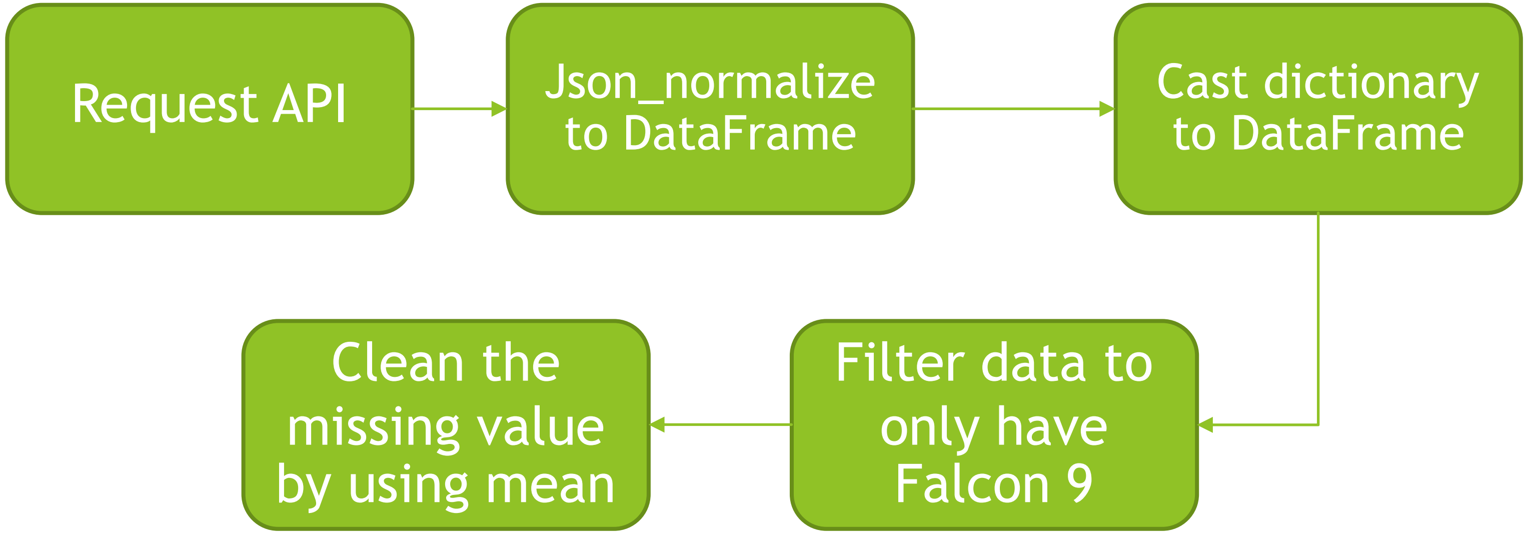
# Methodology

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

# Data Collection

- Data collected by
  - request API from SpaceX API
  - Web scarping from SpaceX Wikipedia page.

# Data Collection - API



Request API

Json_normalize to DataFrame

Cast dictionary to DataFrame

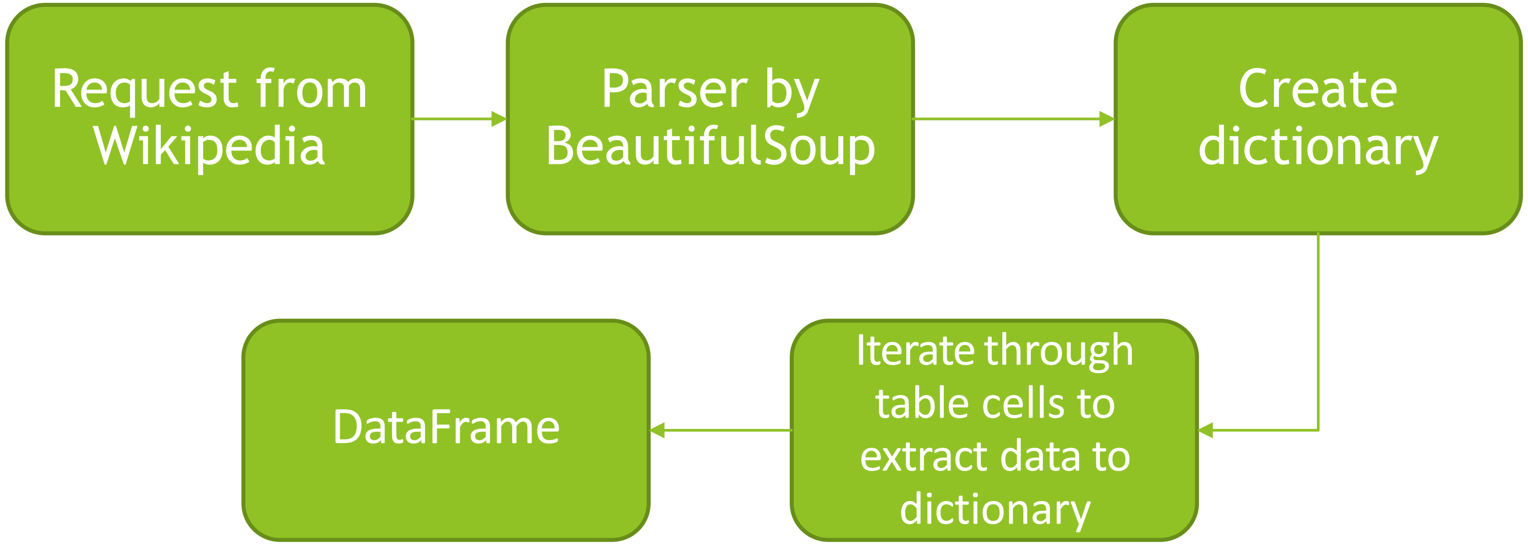
Filter data to only have Falcon 9

Clean the missing value by using mean

Github URL:

https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/Data%20Collection%20API%20Lab.ipynb

# Data Collection – Web Scraping



Github URL:

https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

# Data Wrangling

- Create a label with landing outcomes ( 1 = success and 0 = failure)
- Outcome has two components: Mission Outcome and Landing Location
  - True ASDS, True RTLS and True Ocean -> 1
  - None None, False ASDS, Non ASDS, False Ocean and False RTLS -> 0

Github URL:
https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/Data%20Wrangling.ipynb

# EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Scatter plots, line charts and bar plots were used to compared relationships between variables to decide if a relationship exists.

Github URL:
https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

# EDA with SQL

- Loaded data into IBM DB2 then queries using SQL Python integration

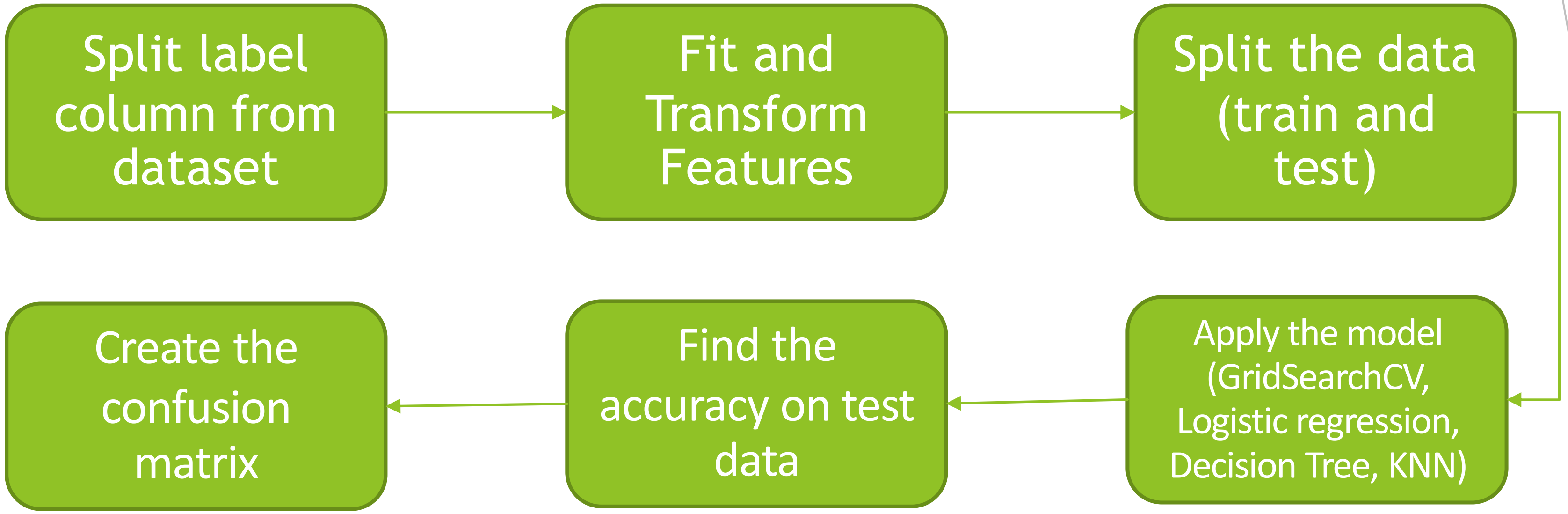
Github URL:
https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/EDA%20with%20SQL.ipynb

# Build an interactive map with Folium

- Folium map marks Launch Sited, landing outcomes and distance to key locations.
- This allows to visualized landing outcomes related to location.

Github URL:
https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

# Predictive analysis (Classification)

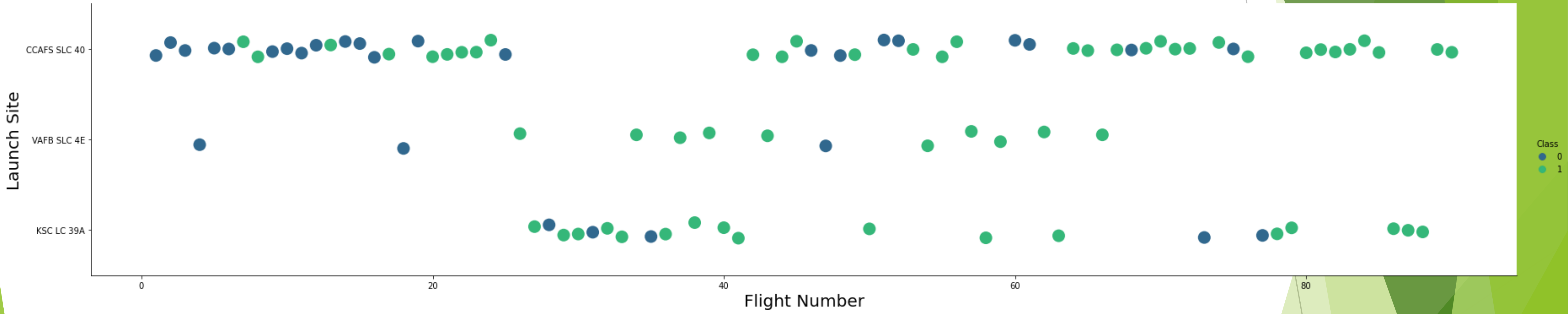

Split label column from dataset
Fit and Transform Features
Split the data (train and test)
Apply the model (GridSearchCV, Logistic regression, Decision Tree, KNN)
Find the accuracy on test data
Create the confusion matrix

Github URL:

https://github.com/alotofp/AppliedDataScienceCapstone/blob/main/Machine%20Learning%20Prediction.ipynb

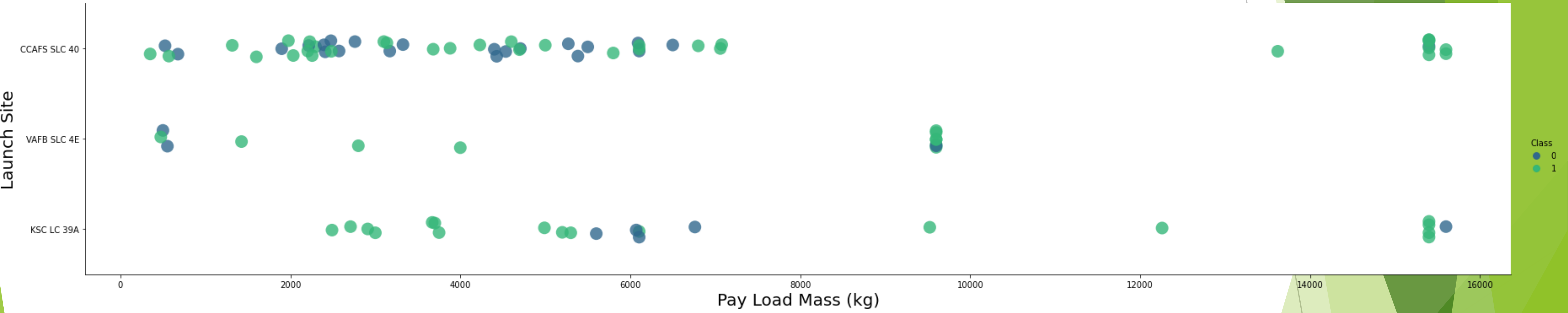
# Insights drawn from EDA

# Flight Number vs. Launch Site



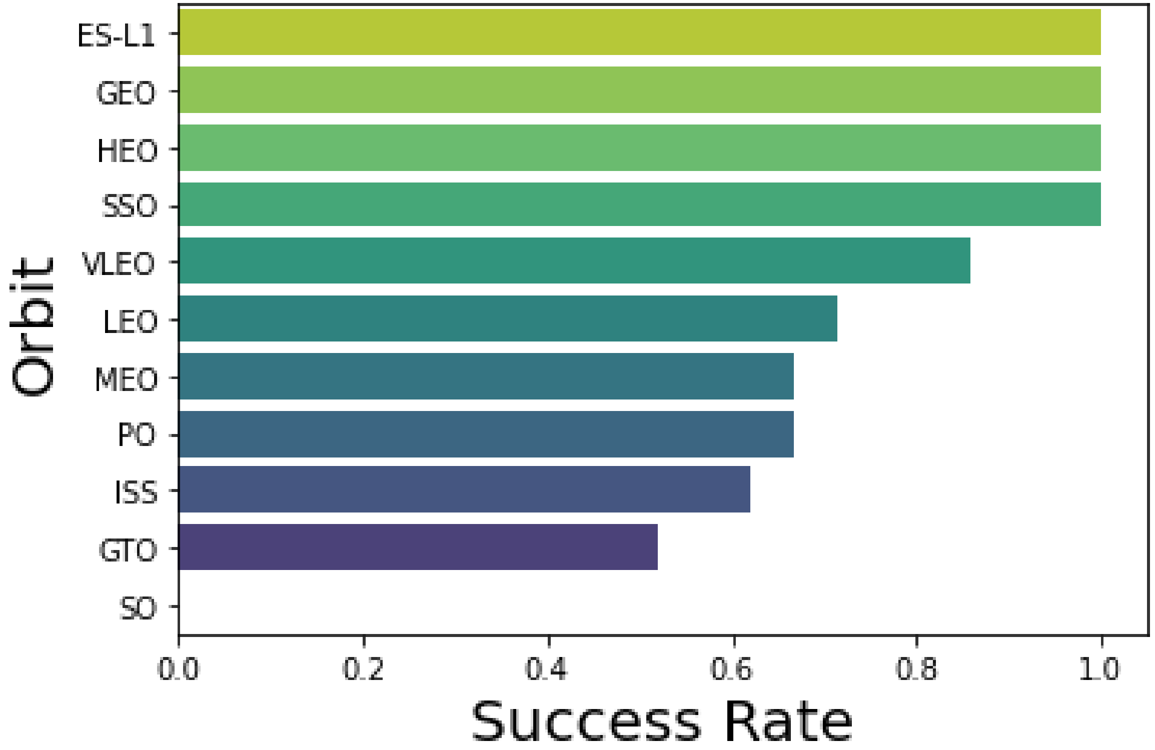
- Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

# Payload vs. Launch Site



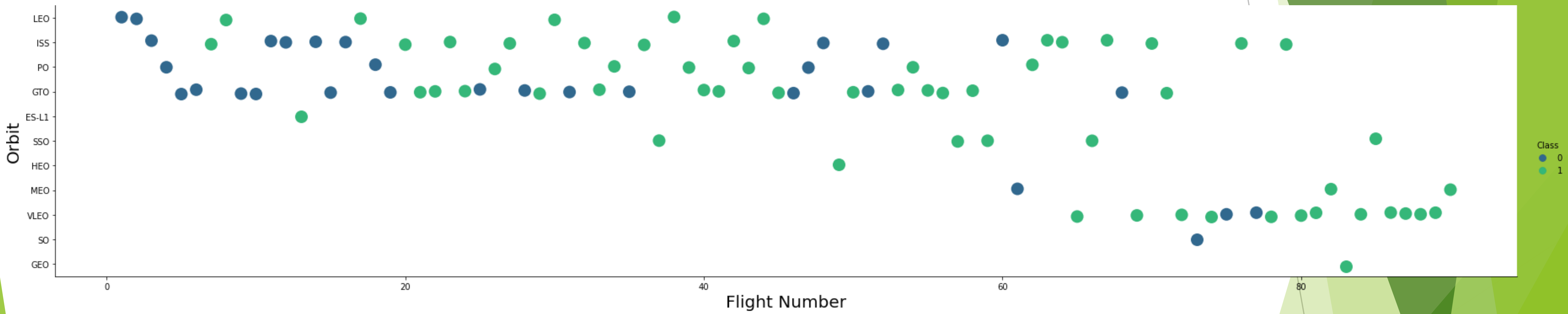
- Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

# Success Rate vs. Orbit Type



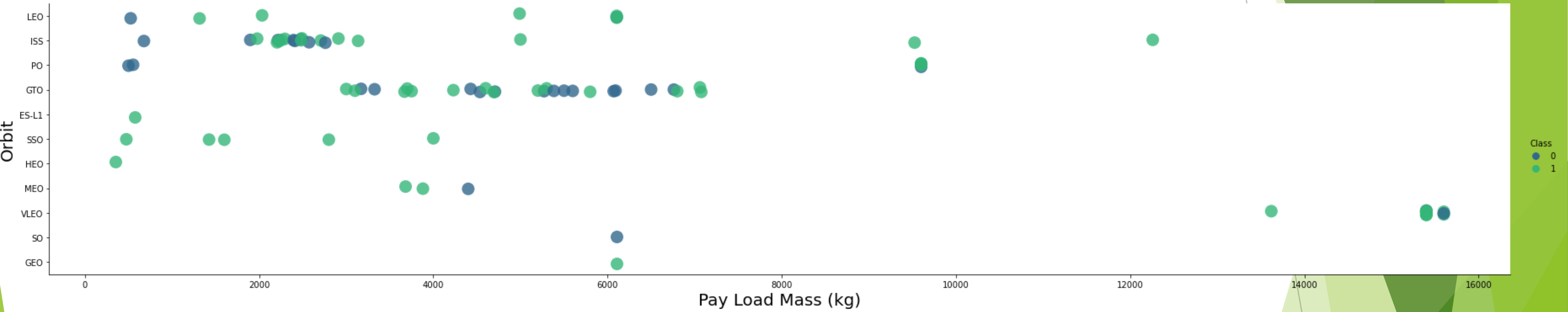
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

# Flight Number vs. Orbit type



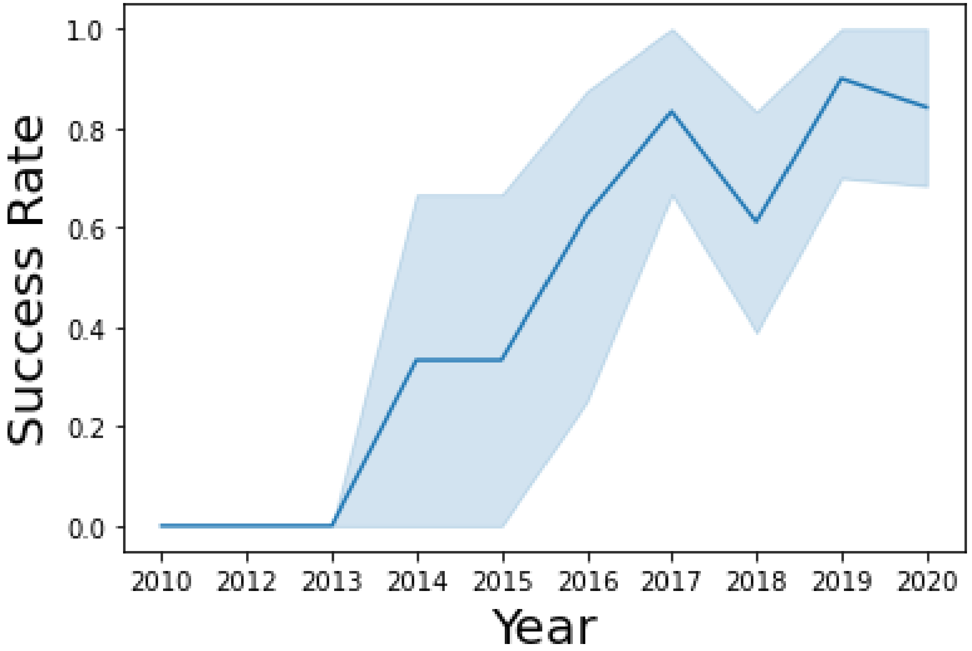
- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

# Payload vs. Orbit type



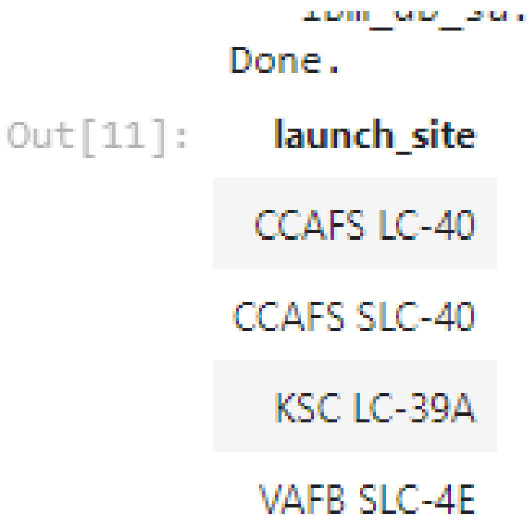
- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

# Launch Success Yearly Trend



- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%

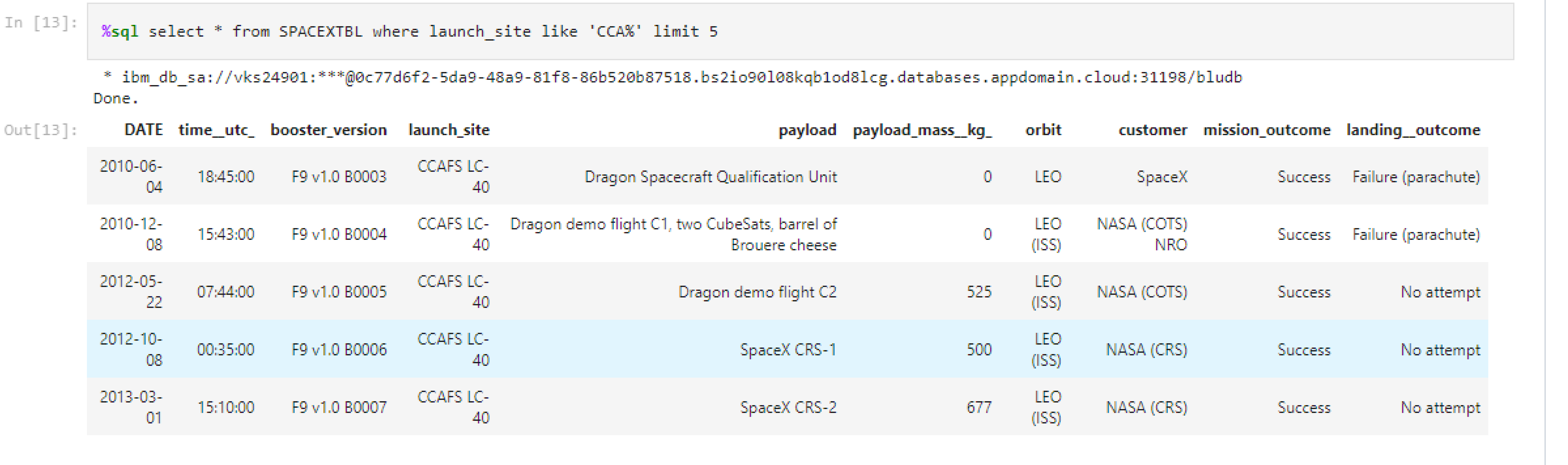
# All Launch Site Names



```
        ibm_db_sa.
        Done.
Out[11]:    launch_site
           CCAFS LC-40
          CCAFS SLC-40
            KSC LC-39A
          VAFB SLC-4E
```

- Launch site names :
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

# Launch Site Names Beginning with `CCA`



```
In [13]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

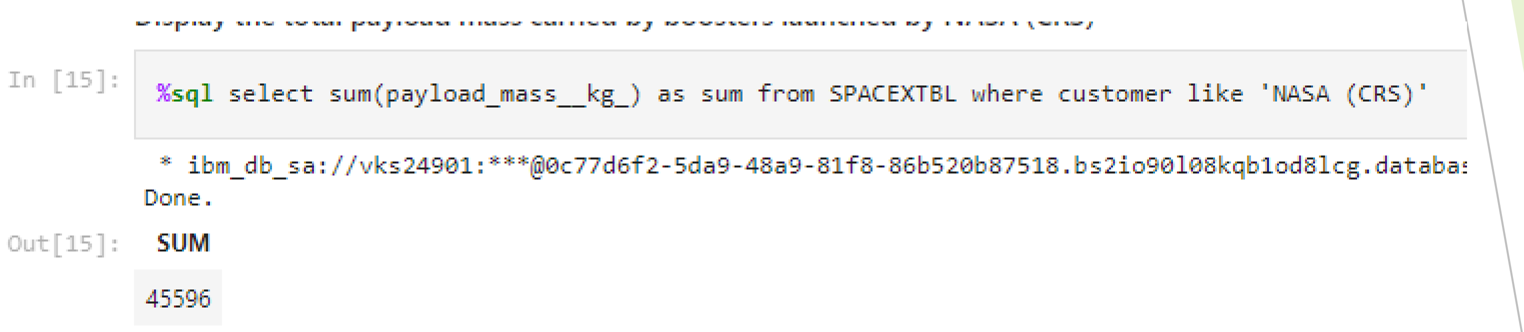
```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

| 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 |

- First five entries  in database with  Launch Site name  beginning with CCA.

# Total Payload Mass from NASA



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

In [15]:

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

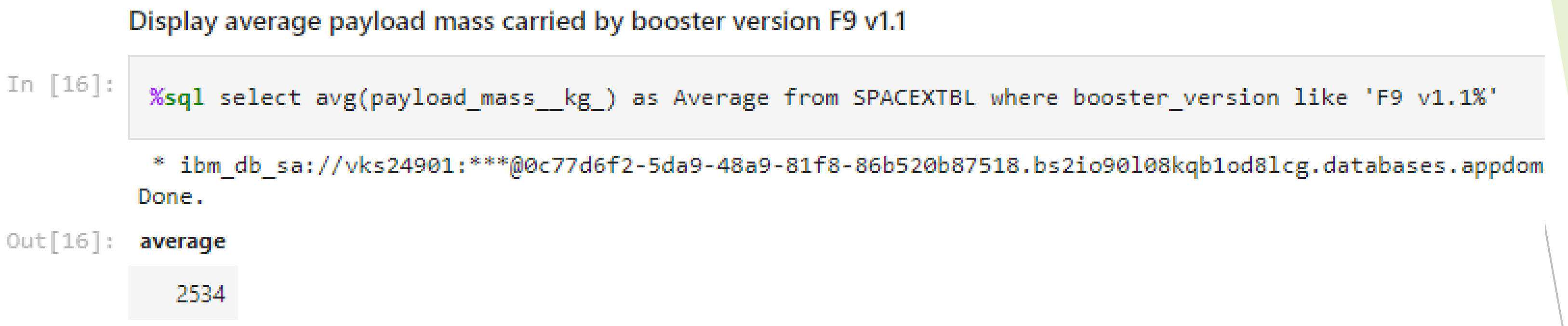
```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databa
Done.
```

Out[15]:

| SUM |
|---|
| 45596 |

- This query sums the total payload  mass in kg where NASA was the customer.

# Average Payload Mass by F9 v1.1



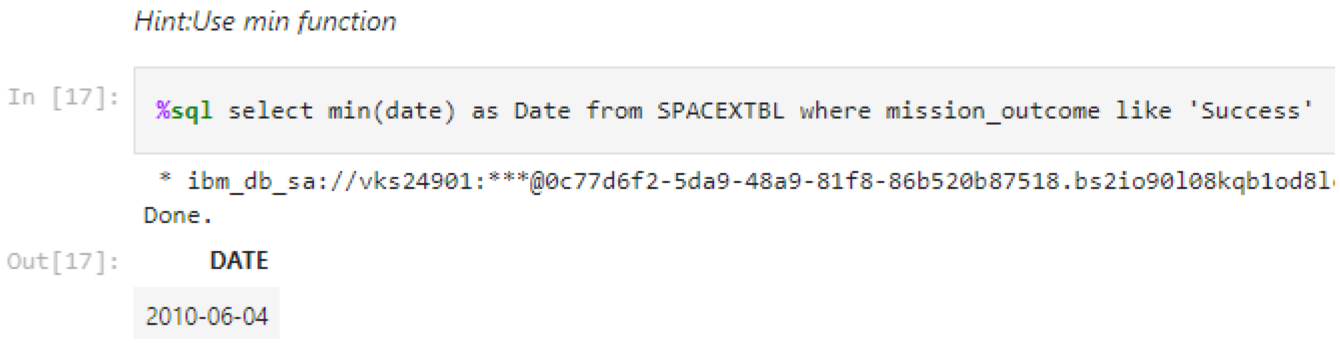
Display average payload mass carried by booster version F9 v1.1

```
In [16]:  %sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'

           * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdom
          Done.
Out[16]:  average
             2534
```

- This query calculates the  average payload mass or  launches which used  booster version F9 v1.1

# First Successful Ground Pad Landing Date



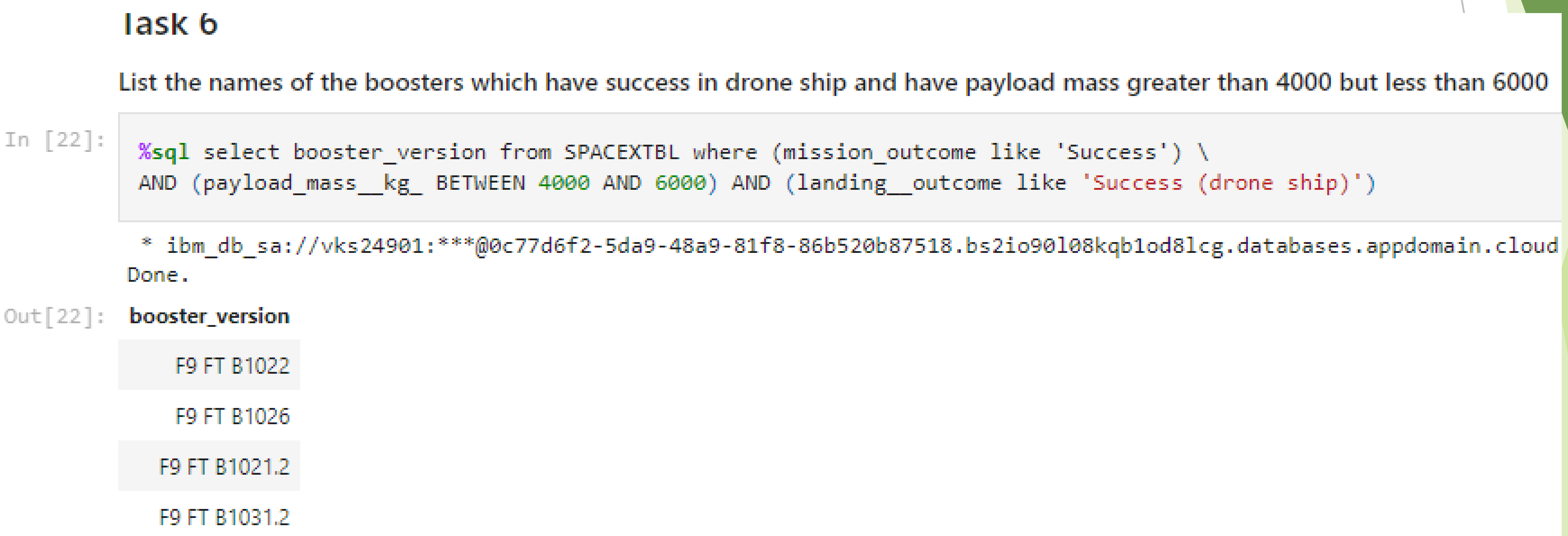
*Hint:Use min function*

```
In [17]:  %sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'

 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l
Done.
Out[17]:      DATE
          2010-06-04
```

- This query returns the first successful ground pad landing date.

# Successful Drone Ship Landing with Payload  Between 4000 and 6000



### lask 6

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

In [22]:

```
%sql select booster_version from SPACEXTBL where (mission_outcome like 'Success') \
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

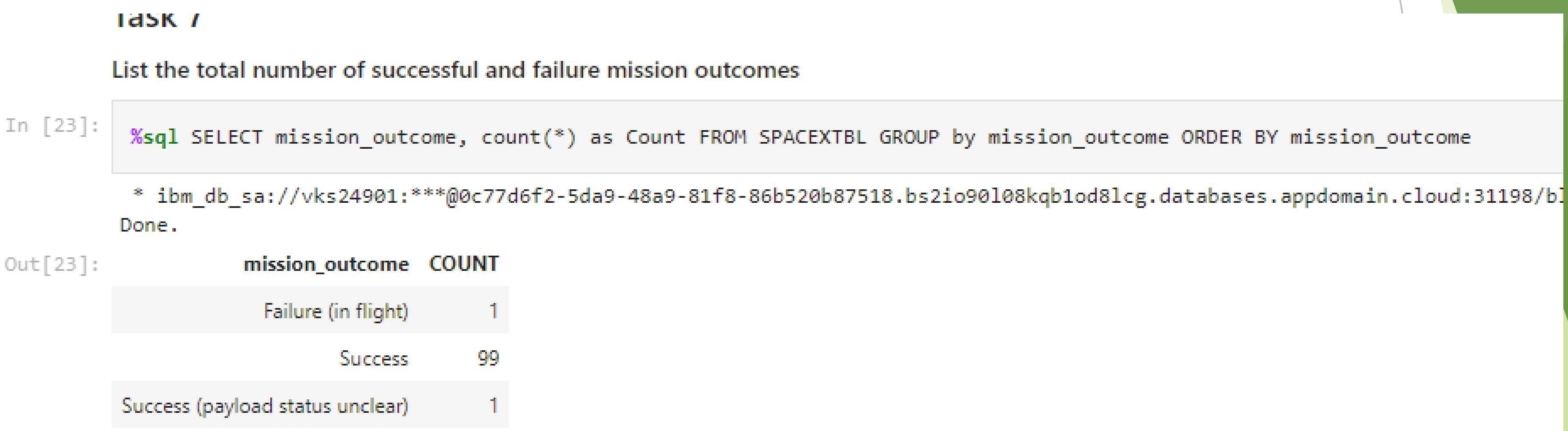
```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud
Done.
```

Out[22]:

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

- This query returns the four  booster versions that had  successful drone ship landings  and a payload mass between  4000 and 6000.

# Total Number of Each Mission Outcome



### Task 7

List the total number of successful and failure mission outcomes

In [23]:

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

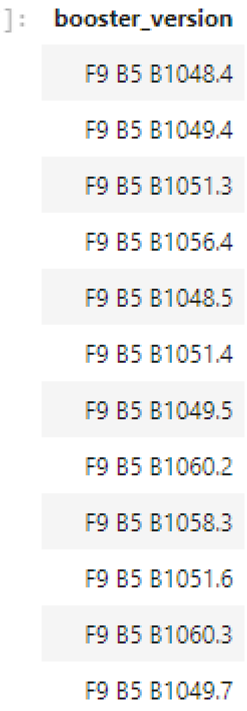
```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bl
Done.
```

Out[23]:

| mission_outcome | COUNT |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- This query returns a count of each mission outcome.

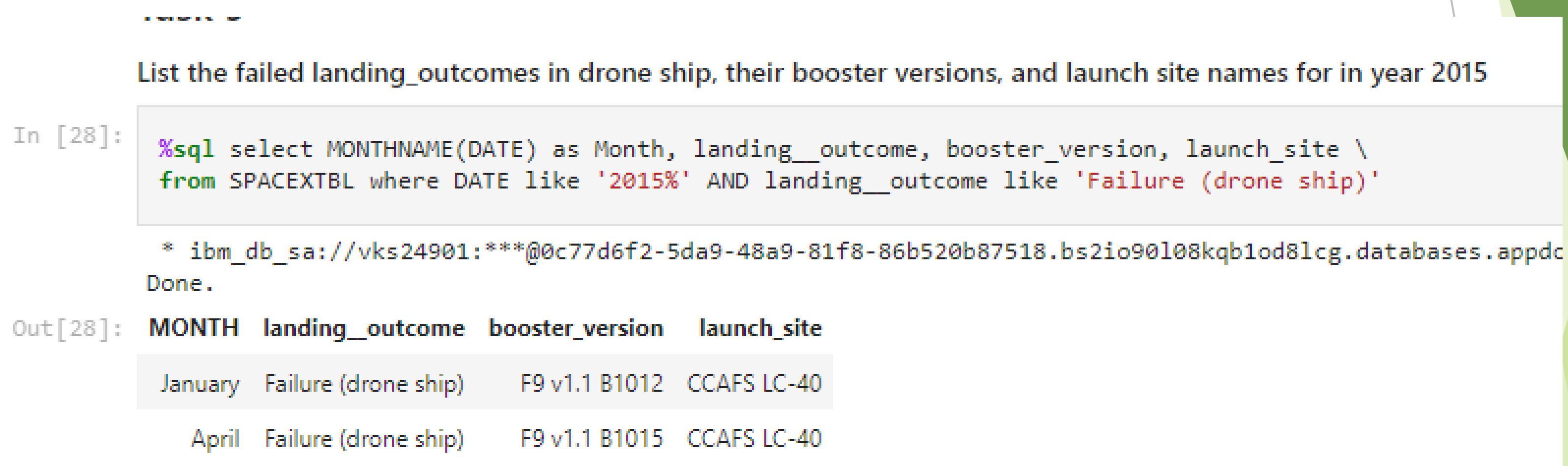
# Boosters that Carried Maximum Payload



| 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 |

- This query returns the booster versions that carried the highest payload mass.
- These booster versions are very similar and  all are of the F9 B5 B10xx.x variety.

# 2015 Failed Drone Ship Landing Records



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

In [28]:

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site \
from SPACEXTBL where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

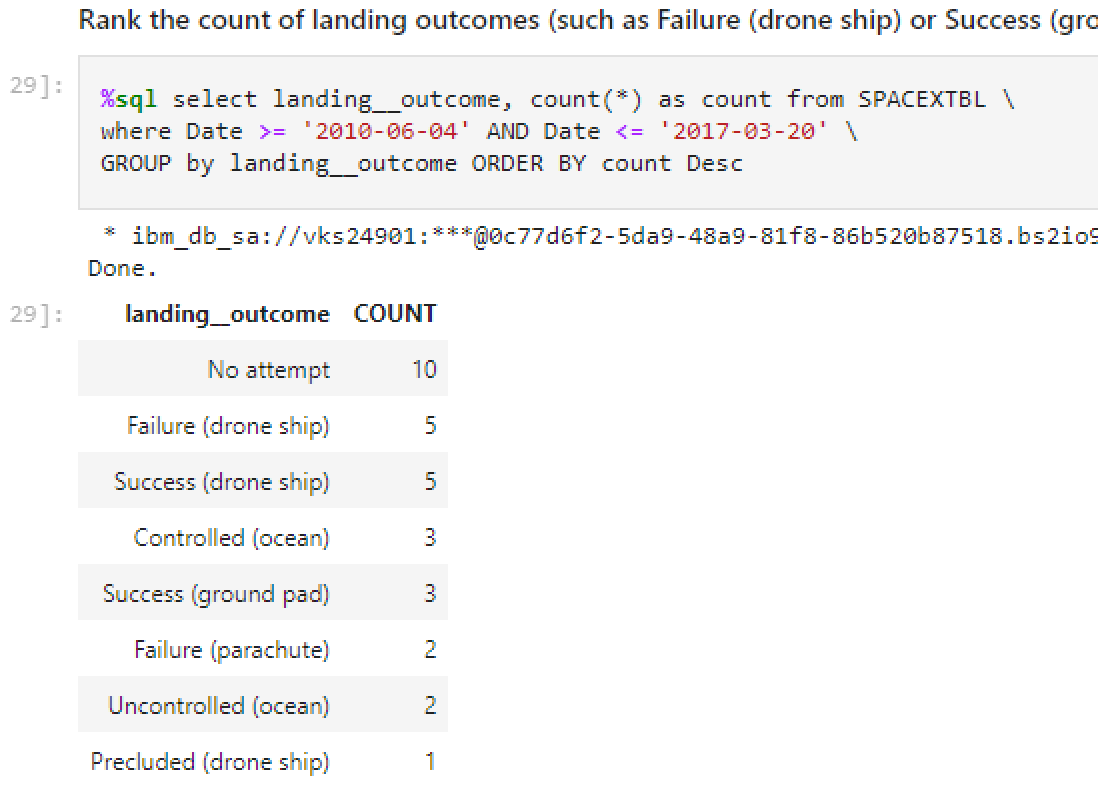
```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdc
Done.
```

Out[28]:

| MONTH | 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 |

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

# Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20



Rank the count of landing outcomes (such as Failure (drone ship) or Success (gro

```
29]:  %sql select landing__outcome, count(*) as count from SPACEXTBL \
      where Date >= '2010-06-04' AND Date <= '2017-03-20' \
      GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://vks24901:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io
Done.
```

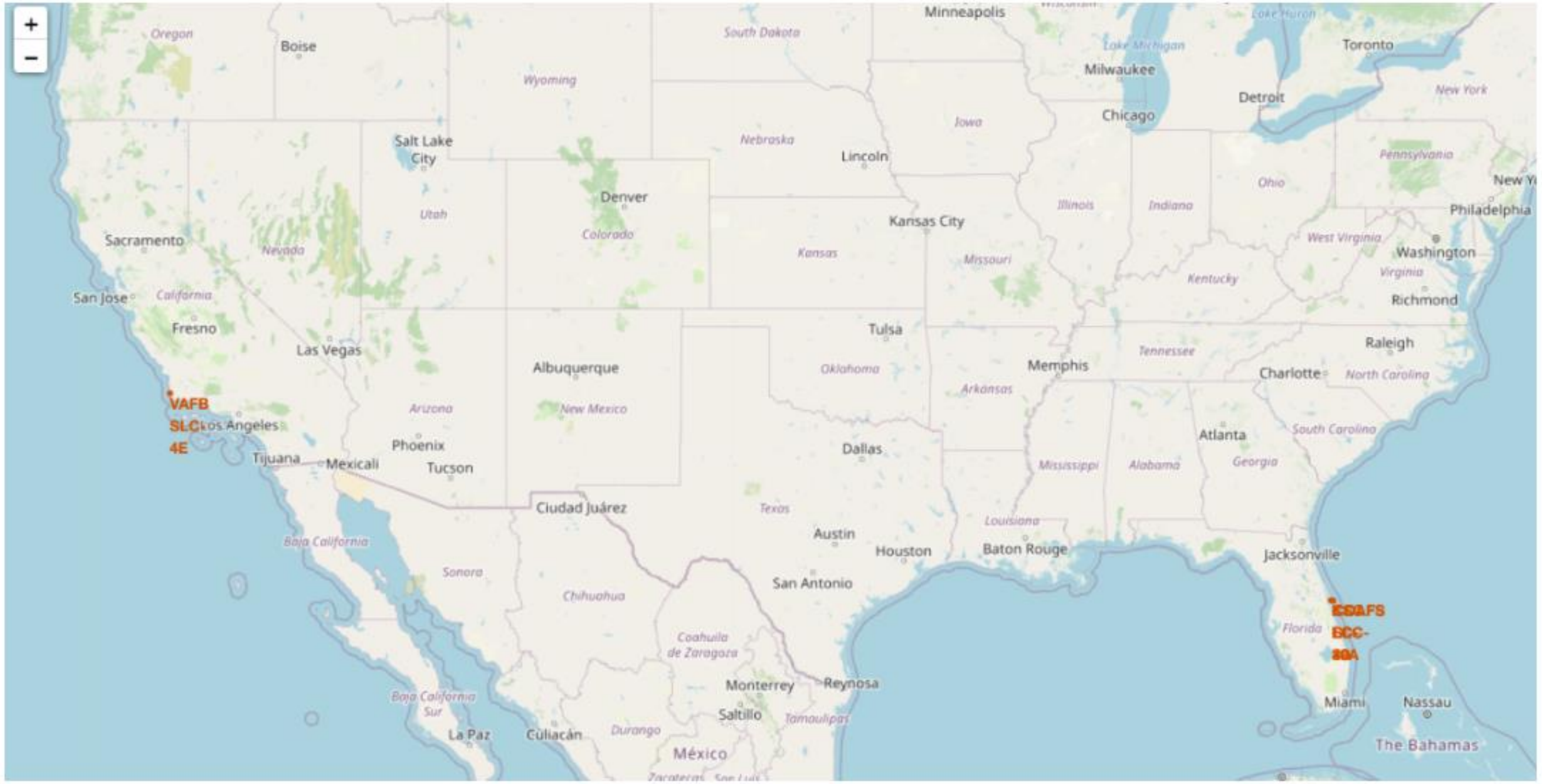
29]:

| landing__outcome | COUNT |
|---|---|
| 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 |

- This query returns a list of successful landings  and between 2010-06-04 and 2017-03-20  inclusively.

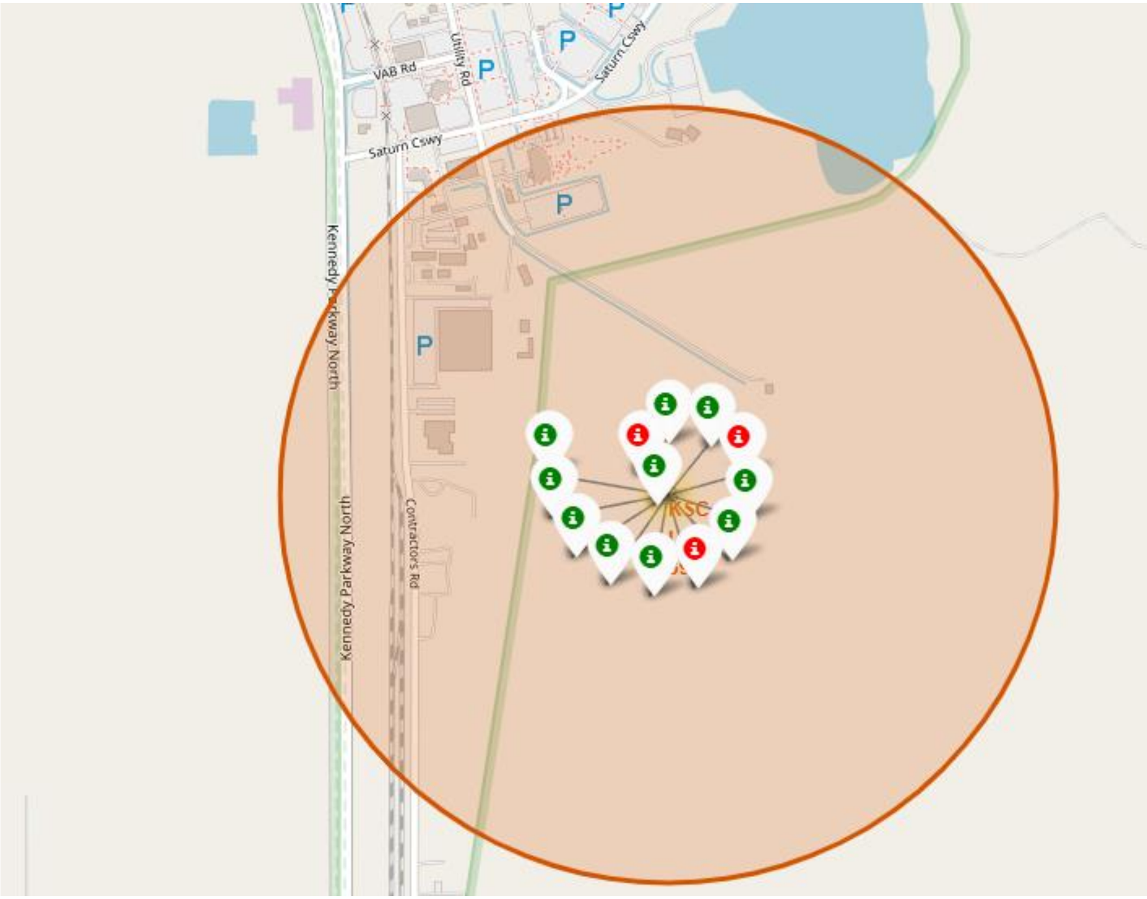
# Launch Sites Proximities Analysis

# Launch Sites Locations



- The map shows all launch sites relative USA map.

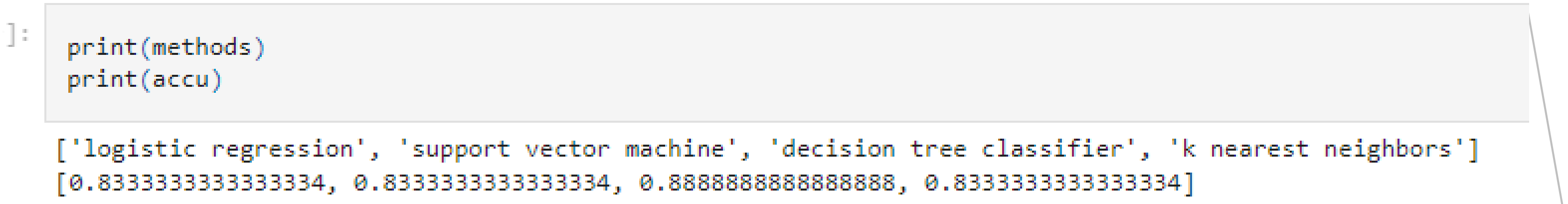
# Color-Coded Launch Markers



- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

# Predictive Analysis (Classification)

# Classification Accuracy

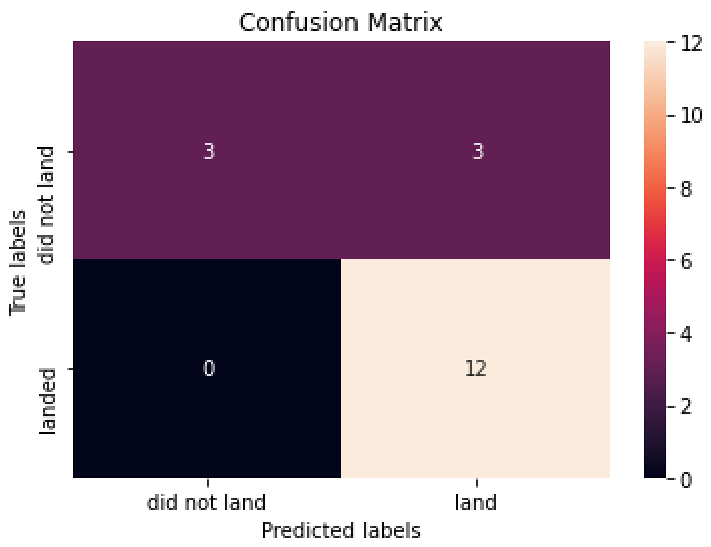


```
print(methods)
print(accu)
```

```
['logistic regression', 'support vector machine', 'decision tree classifier', 'k nearest neighbors']
[0.8333333333333334, 0.8333333333333334, 0.8888888888888888, 0.8333333333333334]
```

- All models have almost the same accuracy around 83%.
- It should be noted that the test data size is relatively small with only 18 samples.
- We need more data to determine the best model.

# Confusion Matrix



- Since all models performed the same for the test set, the confusion matrix is the same across all models.
- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landing.

# Conclusion

- Problem: develop a machine learning model to predict the landing outcomes
- Objective: create a model that can use to predict the landing outcome
- Results: created models accuracy are around 83%
- Data collected from SpaceX API and SpaceX Wikipedia
- If possible more data should be collected to better determine the best machine learning model  and improve accuracy

# Appendix

Github repository URL: https://github.com/alotofp/AppliedDataScienceCapstone