IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Santi Wongprasoet

July 19th , 2023

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

# Executive Summary

**Summary of methodologies**

This research aims to find the factor for successful rocket landing and the model to predict the result. The methodologies has shown below

- **Data collection**: use the SpaceX REST API and Web Scraping
- **Data wrangling**: prepare data and create success/unsuccessful outcome variable
- **Exploratory data analysis**: data visualization techniques and SQL. Factors: payload, launch site, and flight number
- **Folium**: launch site success rate and proximity to geographical markers
- **Plotly Dash**: launch sites with success launches and successful payload ranges
- **Classification Models**: predict landing outcome using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

# Executive Summary (Cont.)

**Results**

**Exploratory data analysis results**

- New launches have a higher success rate
- KSC LC-39A has the highest successful rate with 76.9%
- 100% success rate: ES-L1, GEO, HEO, and SSO

**Interactive analytics results**

- All launch sites are in very close proximity to the coast to the Equator line

**Predictive analysis results**

- The decision tree model is the best model

# Introduction

**Background**

SpaceX 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. 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 SpaceX for a rocket launch

**Explore**

- The relation of payload mass, launch site, number of flight, and orbits to first-stage landing success
- Historical successful landing rate
- Predictive model for successful landing

Section 1

# Methodology

# Methodology

- Data collection:
  - Using the SpaceX REST API and Web Scraping
- Data wrangling
  - Prepare data by filtering the data, handling missing value, and apply one hot encoding
- Exploratory data analysis (EDA)
  - Using visualization and SQL
- Interactive visual analytics
  - Using Folium and Plotly Dash
- Perform predictive analysis
  - Classification models will be used to predict landing outcome
  - Tune and evaluate models to find the best model and parameters

# Data Collection - SpaceX API

The steps to collect the data from SpaceX API

- **Request data** (rocket launch data) from SpaceX API
- **Decode the response content** as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
- **Get information** about the launches by using the API and the IDs given for each launch
- **Create dictionary** from data set
- **Create dataframe** from dictionary
- **Filter the dataframe** to only include Falcon 9 launches
- **Replace missing values** with mean for the PayloadMass using the .mean()
- **Export data** to csv file

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/01%20Data%20Collection%20API.ipynb**

# Data Collection - Web Scraping

The steps to collect the data from web scraping

- **Request data** from the Falcon9 Launch Wikipedia page
- **Create a BeautifulSoup object** from the HTML response
- **Extract all column names** from the HTML table header
- **Parsing the launch HTML tables** and collect data into dictionary
- **Create a data frame** from dictionary
- **Export data** to csv file

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/02%20Web%20Scraping.ipynb**

# Data Wrangling

The steps to data wrangling

- **Perform EDA (Exploratory Data Analysis)** to find some patterns in the data
- **Determine label** for training supervised models
- **Calculate**
  - the number of launches on each site
  - the number and occurrence of each orbit
  - the number and occurrence of mission outcome per orbit type
- **Create binary label** from a landing outcome column
- **Export data** to csv file

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/03%20Data%20Wrangling.ipynb**

# Data Wrangling Cont.

### Landing Outcome

- **True Ocean:** the mission outcome was successfully landed to a specific region of the ocean
- **False Ocean**: mission outcome was unsuccessfully landed to a specific region of the ocean
- **True RTLS**: the mission outcome was successfully landed to a ground pad
- **False RTLS**: the mission outcome was unsuccessfully landed to a ground pad
- **True ASDS**: the mission outcome was successfully landed to a drone ship
- **False ASDS**: the mission outcome was unsuccessfully landed to a drone ship
- **None ASDS** and **None None**: a failure to land

### Outcome conversion

- 1 for a successful landing, 0 for an unsuccessful landing

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/03%20Data%20Wrangling.ipynb**

# EDA with Data Visualization

**Charts**

- Flight Number and Payload Mass (kg)
- Flight Number and Launch Site
- Flight Number and Orbit Type
- Payload Mass (kg) and Launch Site
- Payload Mass (kg) and Orbit Type
- Success rate and Orbit Type

**Analysis**

- **Consider relationship by scatter plot,** which is useful for machine learning if there is a relation.
- **Show comparisons with bar charts** among discreate categories and a measured value.

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/05%20EDA%20Data%20Visualization.ipynb**

# EDA with SQL

**Display:**

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

**List:**

- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in year 2015.
- count of landing outcomes or Success between the date 2010-06-04 and 2017-03-20, in descending order.

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/04%20EDA%20SQL.ipynb**

# Build an Interactive Map with Folium

**Map objects:**

**Mark all launch sites on a map**

- Create a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name based on its latitude and longitude coordinates.
- Create a red circle at all launch sites coordinates with a popup label showing its name based on its latitude and longitude coordinates.

**Mark the launch outcomes for each site on the map with color**

- If a launch was successful, then we use a green marker and if a launch was failed, we use a red marker

**Distances between a launch site to its proximities**

- Show a colored line between a launch site to its proximities such as closest city, railway, highway, and nearest coastline.

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX/blob/main/06%20Interative%20Visual%20Analysis%20with%20Folium.ipynb**

# Build a Dashboard with Plotly Dash

**Launch Site Drop-down Input Component**

- Select one specific launch site

**Pie chart visualizing launch success counts**

- Show the success and unsuccessful launches for the selected site as percent of total

**Range slider of payload mass**

- Select pay load mass range

**Scatter chart displaying payload mass vs success rate by booster version**

- Show how payload correlated with successful outcomes

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX/blob/main/07%20Interactive%20Dashboard%20with%20Ploty%20Dash.py**

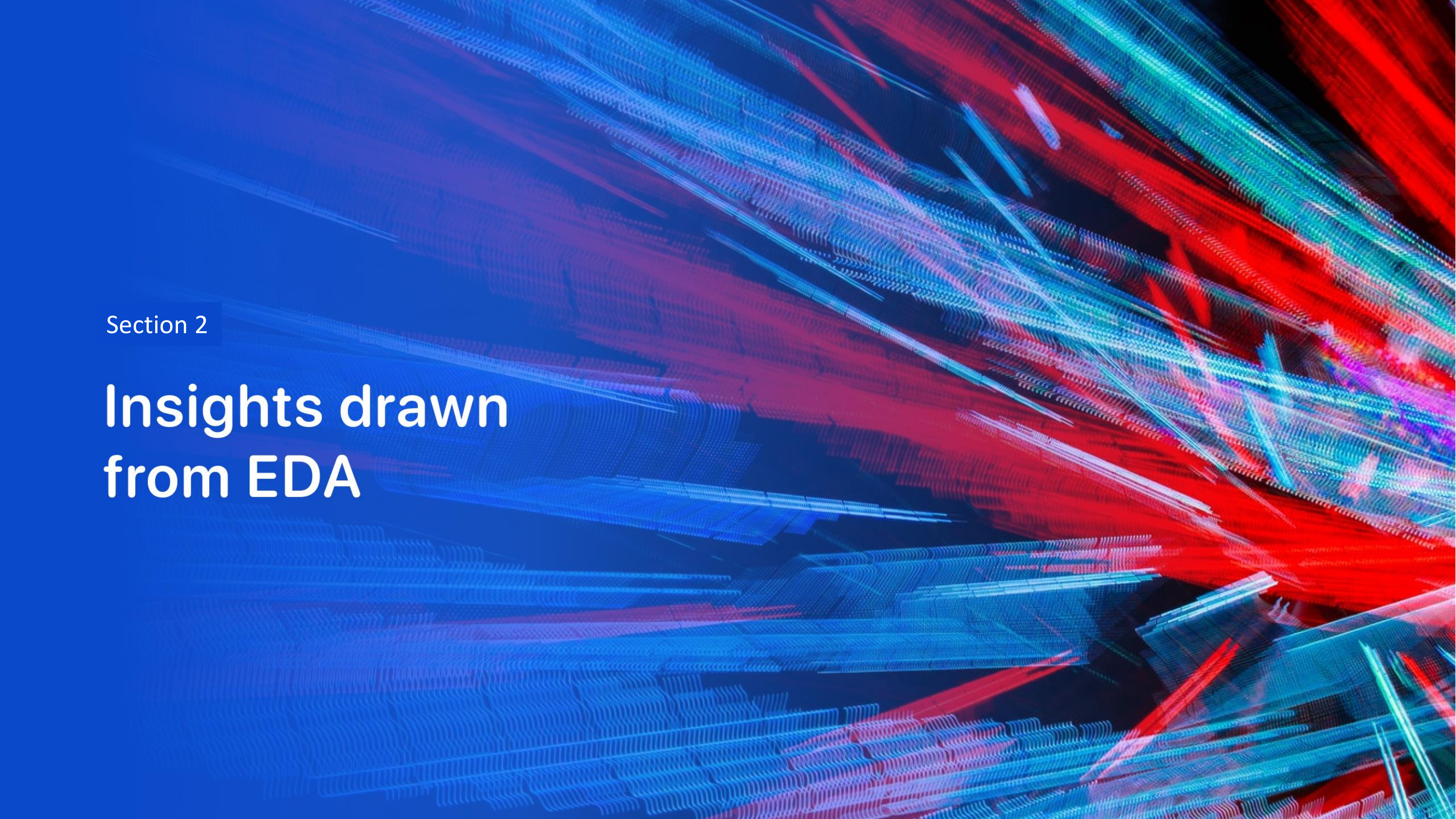
# Predictive Analysis (Classification)

- **Create a NumPy** array from the column Class in data
- **Standardize the data** using StandardScaler to fit and transform the data.
- **Split the data** into training and test data using train_test_split function
- **Create a GridSearchCV** object with cv = 10 to find the best parameters
- **Apply** GridSearchCV on different **algorithms**: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and k-nearest neighbors (KNeighborsClassifier())
- **Calculate the accuracy** of all models on the test data using the .score():
- **Plot confusion matrix** for all models
- **Find the best model** using accuracy

**https://github.com/NOTST/IBM-Data-Science-Capstone-SpaceX-/blob/main/08%20Machine%20Learning%20Prediction.ipynb**

# Results

- **Exploratory data analysis results**
  - New launches have a higher success rate
  - KSC LC-39A has the highest successful rate with 76.9%
  - 100% success rate: ES-L1, GEO, HEO, and SSO
- **Interactive analytics results**
  - All launch sites are in very close proximity to the coast to the Equator line
- **Predictive analysis results**
  - The decision tree model is the best model

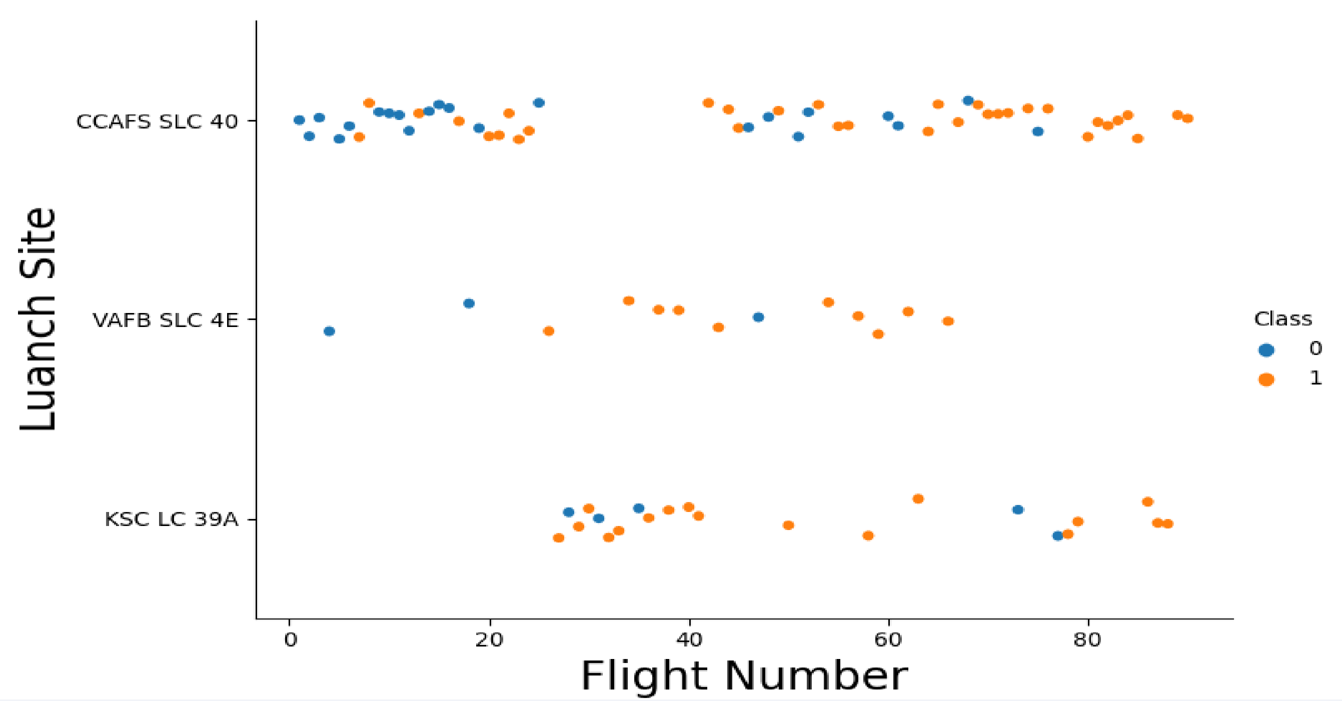


Section 2

# Insights drawn from EDA

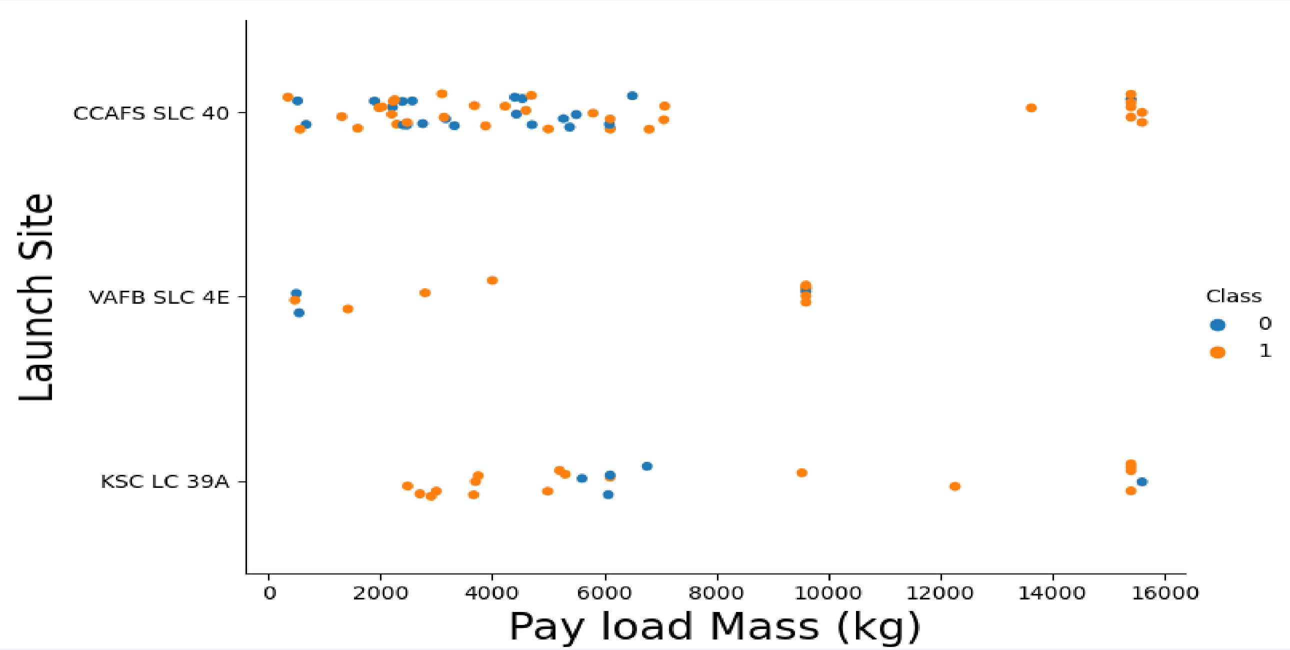
# Flight Number vs. Launch Site

- The more flight number, the higher success rate
- New launches have a higher success rate
- CCAFS SLC 40 launch site has launched a half of all launches approximately
- VAFB SLC 4E and KSC LC 39A launch sites had high success rate



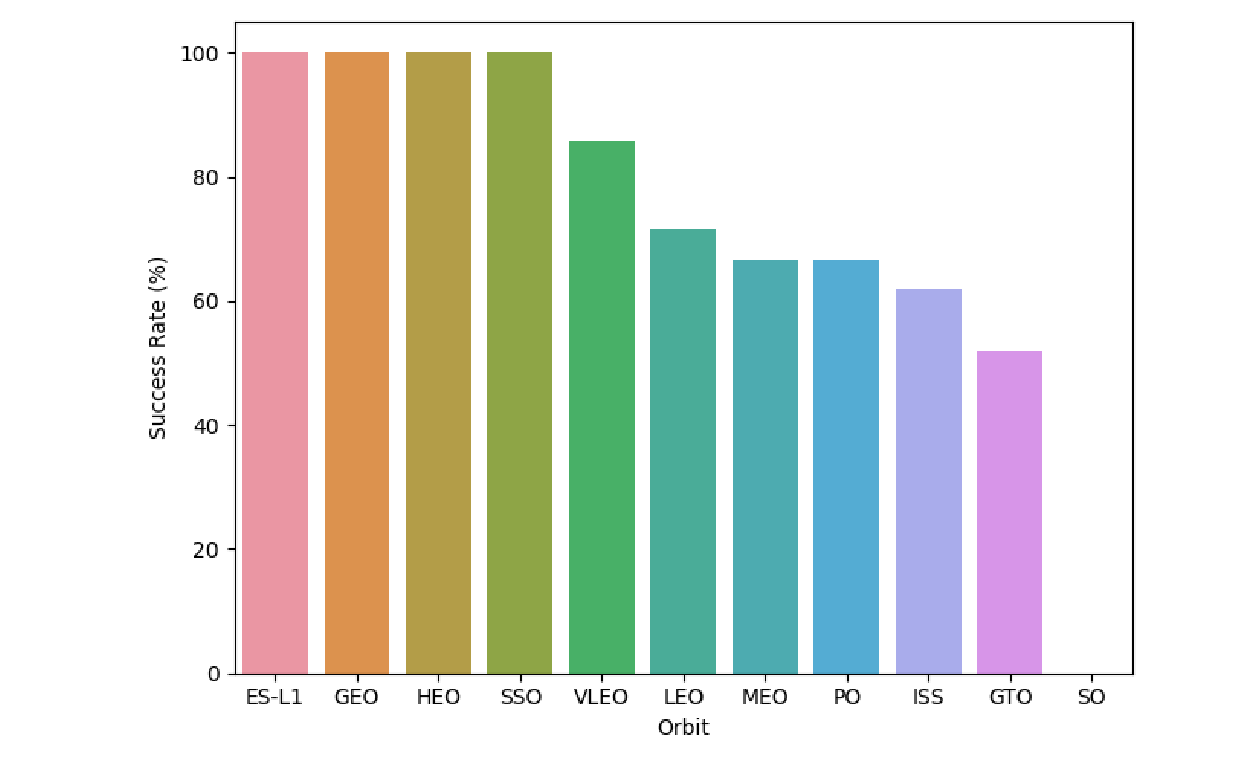
# Payload vs. Launch Site

- The higher payload mass (kg), the higher success rate
- Mostly, payload mass with more than 7000 kg had higher success rate
- VAFB SLC 4E has all launched less than 10,000 kg
- 100% success rate at KSC LC 39A launch site has payload mass less than 5000 kg



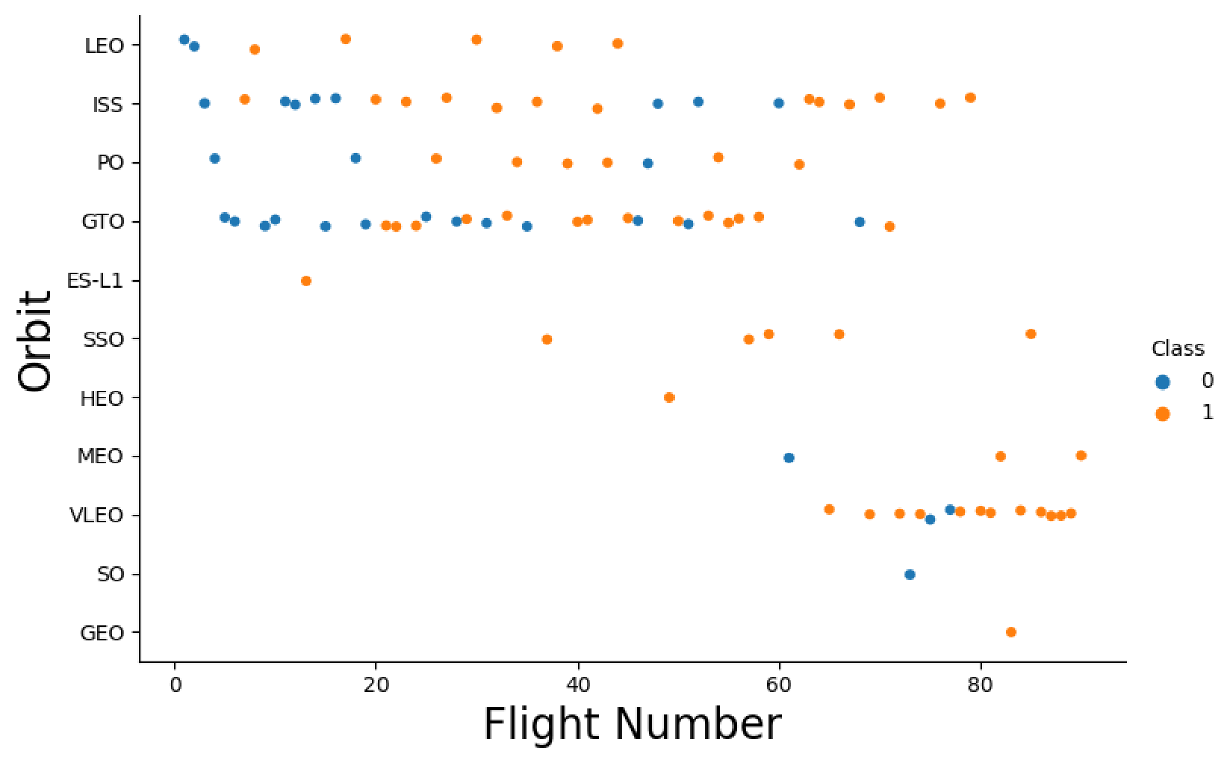
# Success Rate vs. Orbit Type

- 100% success rate: ES-L1, GEO, HEO, and SSO
- 85% success rate: SSO
- 50 - 80% success rate: LEO, MEO, PO, ISS, and GTO
- 0% success rate: SO



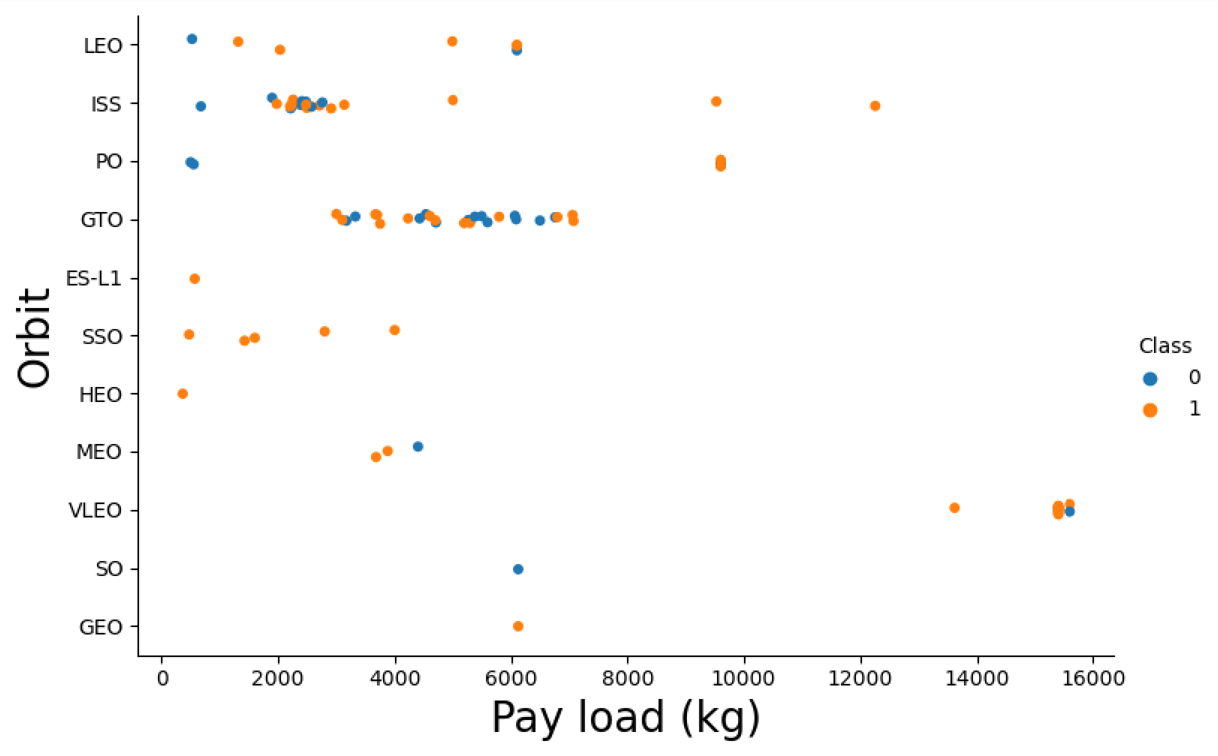
# Flight Number vs. Orbit Type

- Commonly, the  success rate increases with the number of flights for each orbit
- LEO orbit: The success rate appears related to the number of flights
- GTO orbit: The success rate seems to be no relationship between flight number



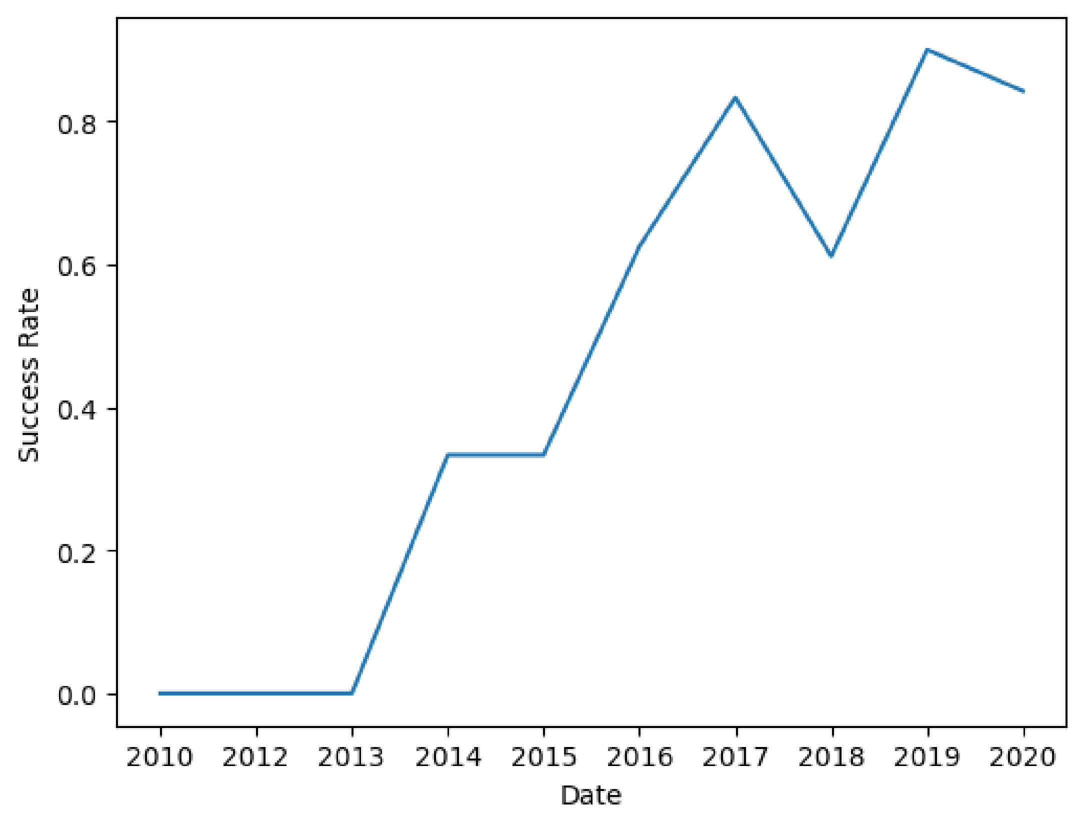
# Payload vs. Orbit Type

- With heavy payloads, the successful landing rate are more for PO, LEO and ISS.
- Success rate of GTO seems to be no relationship between payload mass

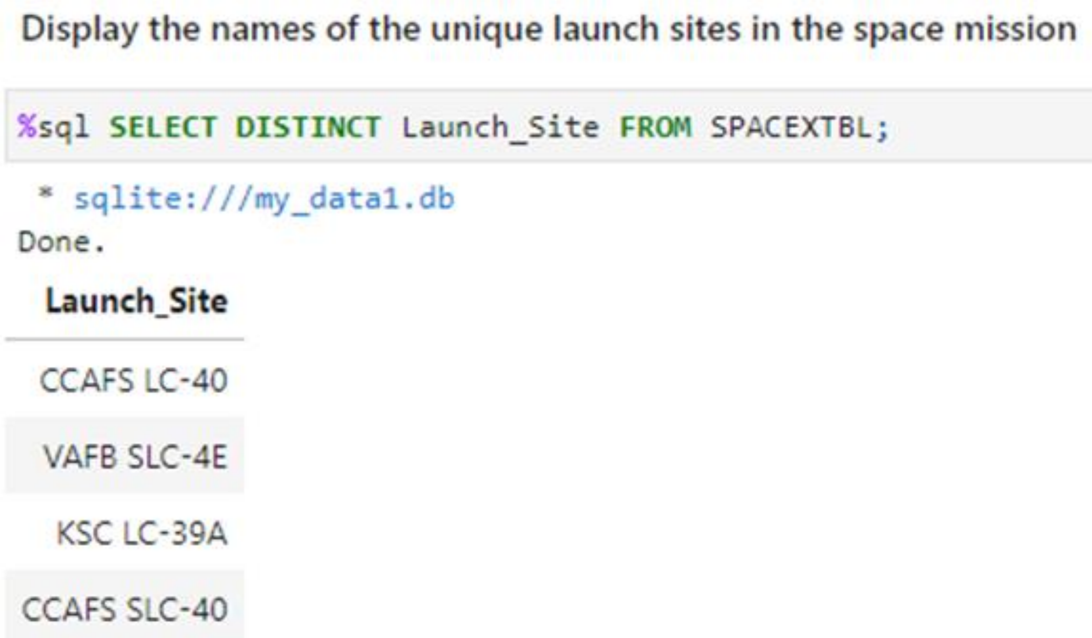


# Launch Success Yearly Trend

- Overall, the success rate kept increasing since 2013 till 2020
- The success rate have dropped in 2017-2018 and 2019-2020



# All Launch Site Names



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

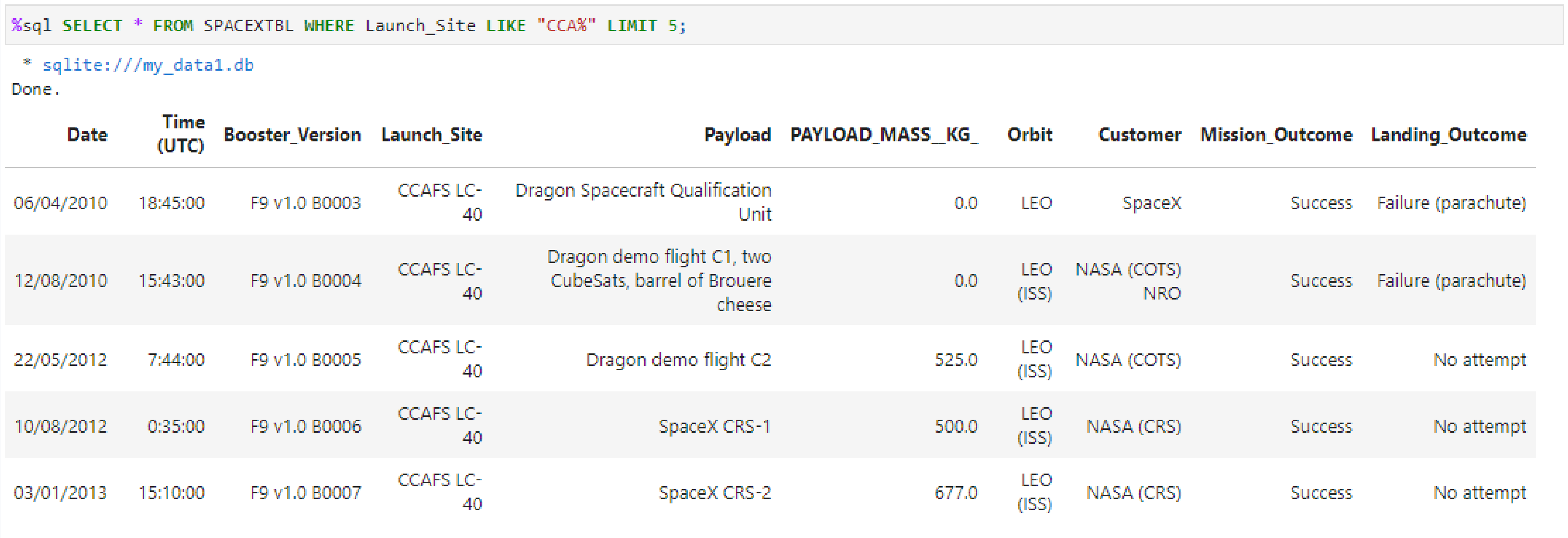
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

# Launch Site Names Begin with 'CCA'

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



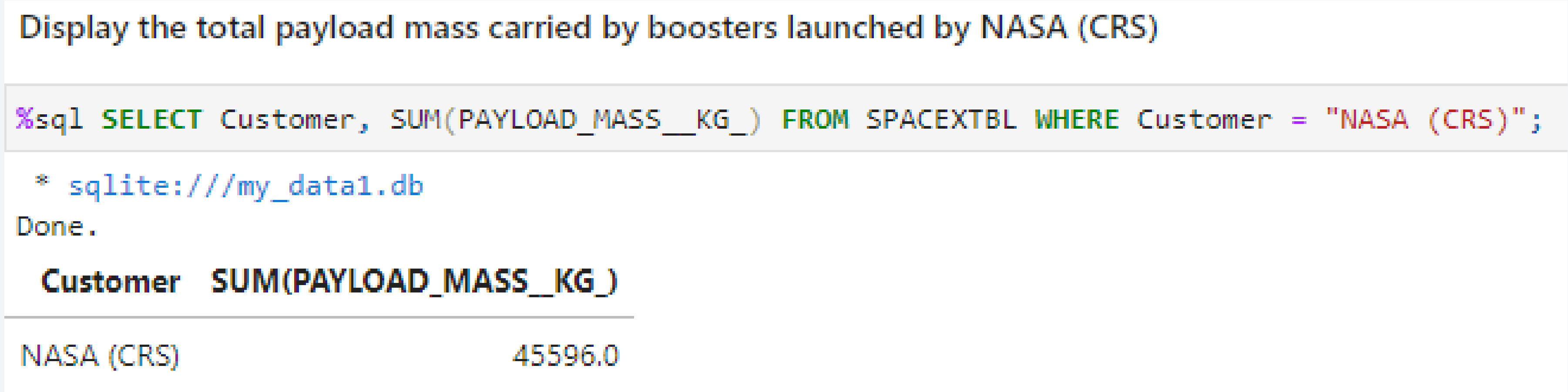
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 06/04/2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0.0 | LEO | SpaceX | Success | Failure (parachute) |
| 12/08/2010 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0.0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 22/05/2012 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525.0 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 10/08/2012 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 03/01/2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677.0 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- Total payload mass of 45,596 kg carried by boosters launched by NASA (CRS)



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

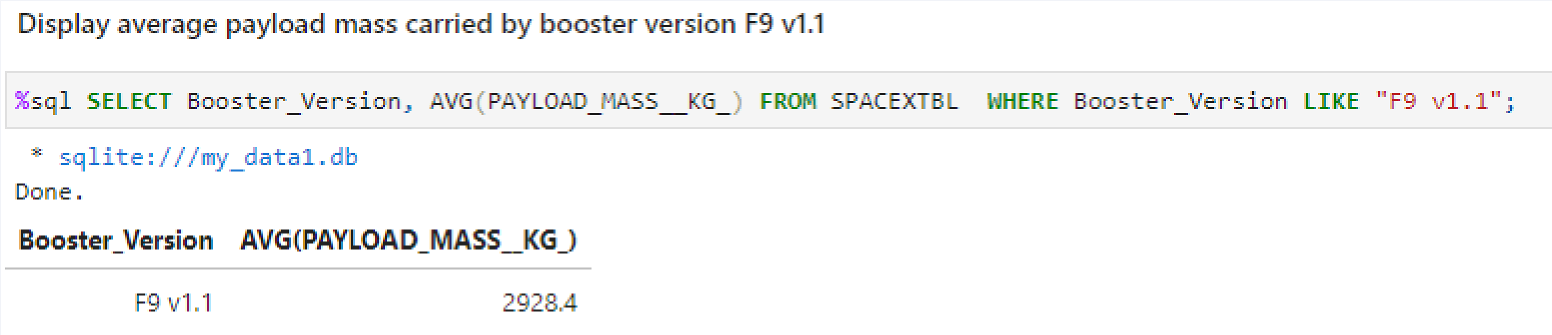
```
%sql SELECT Customer, SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

```
 * sqlite:///my_data1.db
Done.
```

| Customer | SUM(PAYLOAD_MASS__KG_) |
|---|---|
| NASA (CRS) | 45596.0 |

# Average Payload Mass by F9 v1.1

- Average payload mass of 2928 kg carried by booster version F9 v1.1



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

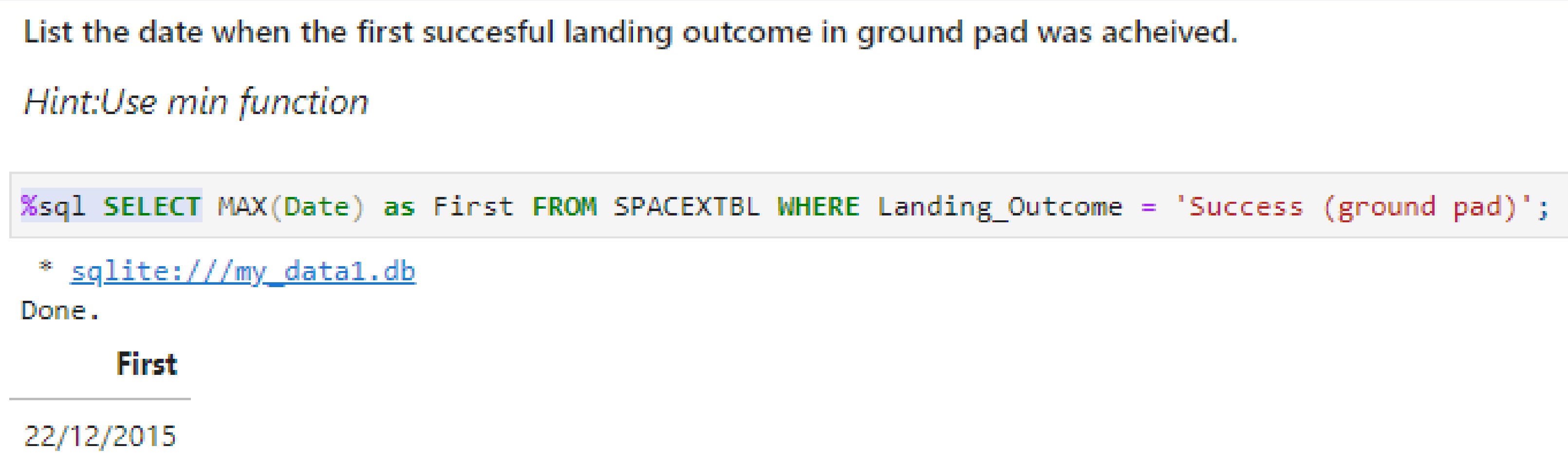
```
%sql SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL  WHERE Booster_Version LIKE "F9 v1.1";
```

```
 * sqlite:///my_data1.db
Done.
```

| Booster_Version | AVG(PAYLOAD_MASS__KG_) |
|---|---|
| F9 v1.1 | 2928.4 |

# First Successful Ground Landing Date

- The first successful landing outcome on ground pad: 22 December 2015



List the date when the first succesful landing outcome in ground pad was acheived.

*Hint:Use min function*

```
%sql SELECT MAX(Date) as First FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

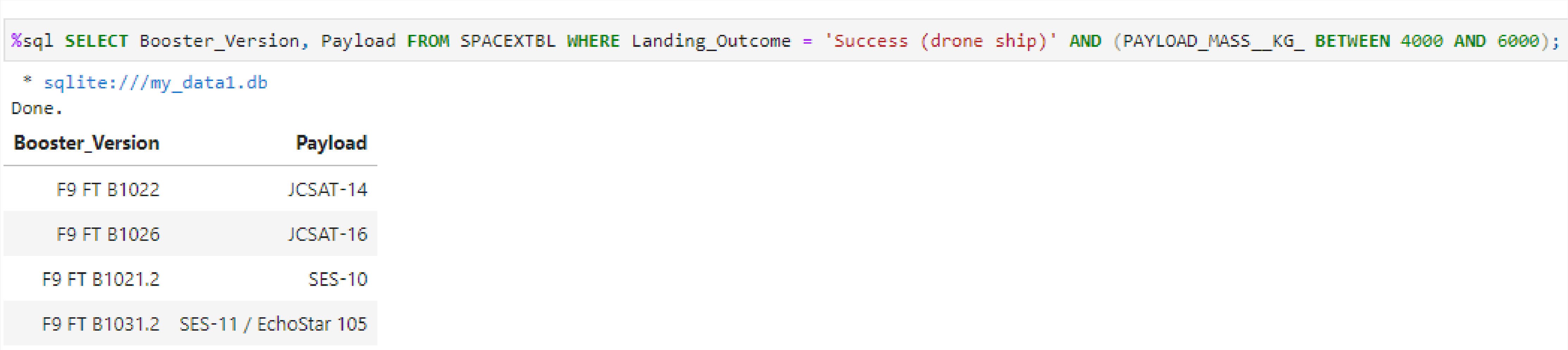
```
 * sqlite:///my_data1.db
Done.
```

| First |
|---|
| 22/12/2015 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 :

- JCSAT-14
- JCSAT-16
- SES-10
- SES-11 / EchoStar 105



```
%sql SELECT Booster_Version, Payload FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

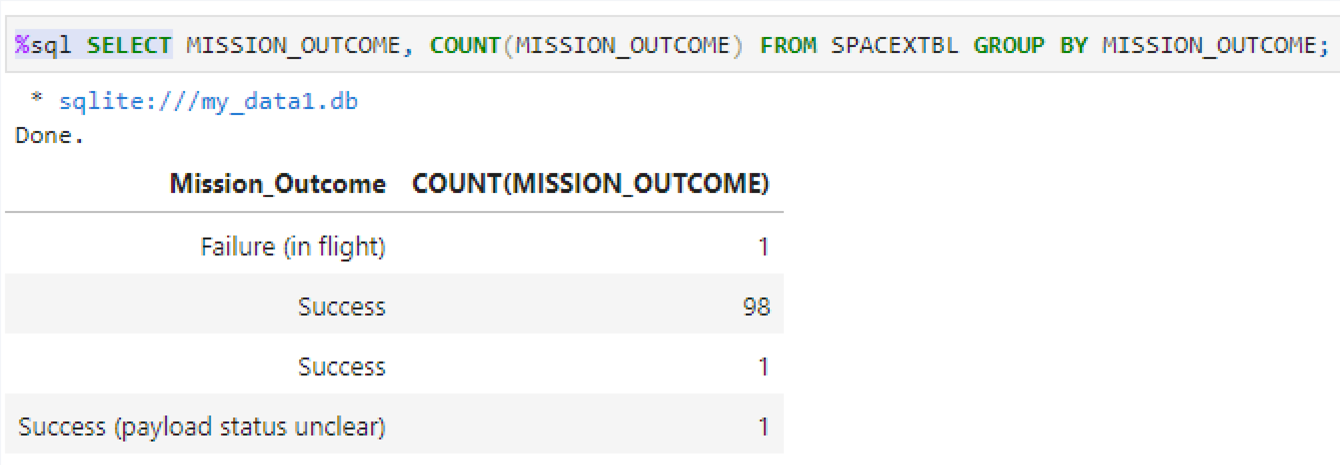
```
 * sqlite:///my_data1.db
Done.
```

| Booster_Version | Payload |
|---:|---:|
| F9 FT B1022 | JCSAT-14 |
| F9 FT B1026 | JCSAT-16 |
| F9 FT B1021.2 | SES-10 |
| F9 FT B1031.2 | SES-11 / EchoStar 105 |

# Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

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



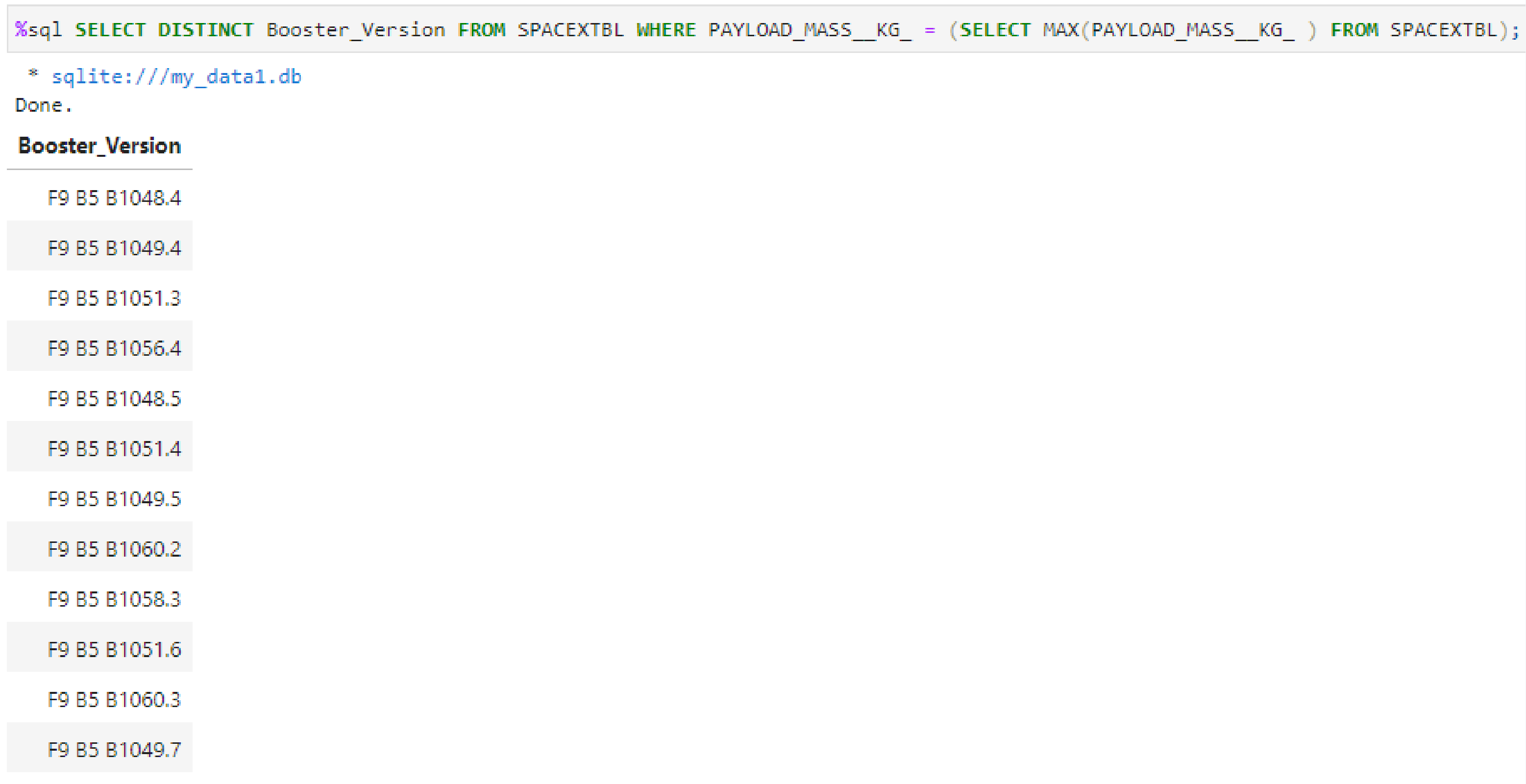
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

| Mission_Outcome | COUNT(MISSION_OUTCOME) |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

The booster which have carried the maximum payload mass



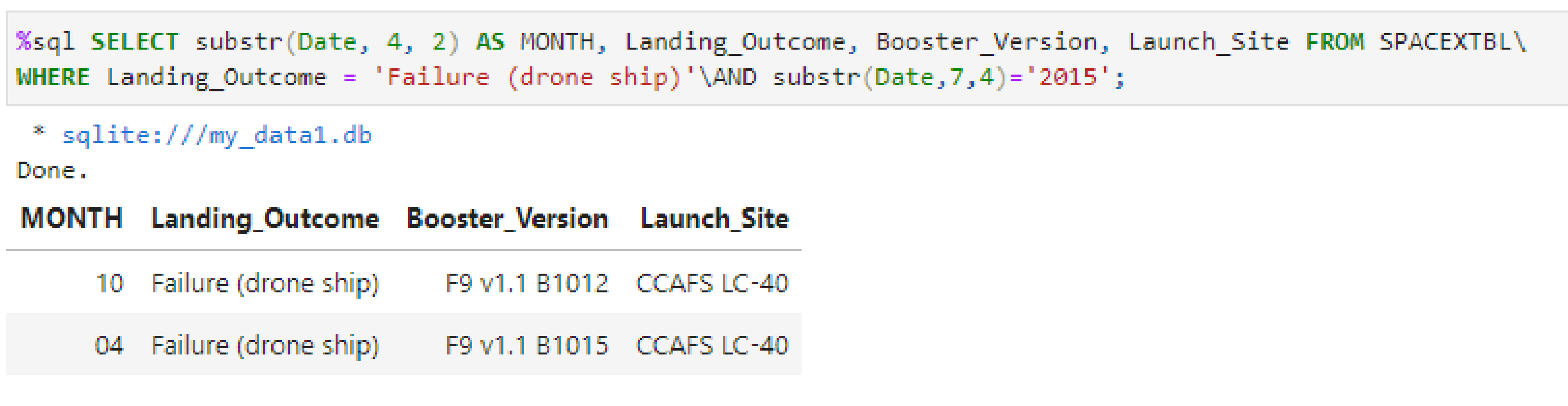
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_ ) FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

| Booster_Version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1049.4 |
| F9 B5 B1051.3 |
| F9 B5 B1056.4 |
| F9 B5 B1048.5 |
| F9 B5 B1051.4 |
| F9 B5 B1049.5 |
| F9 B5 B1060.2 |
| F9 B5 B1058.3 |
| F9 B5 B1051.6 |
| F9 B5 B1060.3 |
| F9 B5 B1049.7 |

# 2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site for in year 2015



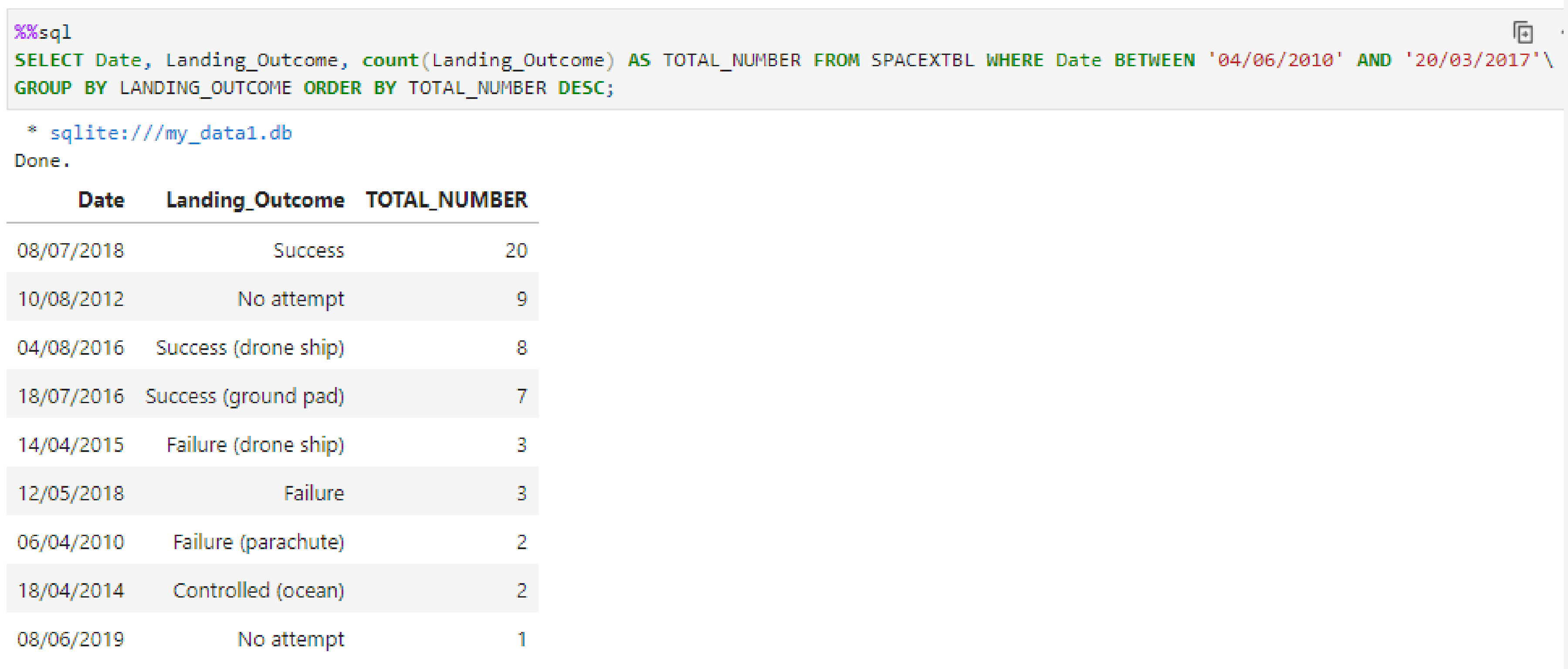
```
%sql SELECT substr(Date, 4, 2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL\
WHERE Landing_Outcome = 'Failure (drone ship)'\AND substr(Date,7,4)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

| MONTH | Landing_Outcome | Booster_Version | Launch_Site |
|---:|---|---:|---:|
| 10 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

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

- Count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order



```
%%sql
SELECT Date, Landing_Outcome, count(Landing_Outcome) AS TOTAL_NUMBER FROM SPACEXTBL WHERE Date BETWEEN '04/06/2010' AND '20/03/2017'\
GROUP BY LANDING_OUTCOME ORDER BY TOTAL_NUMBER DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

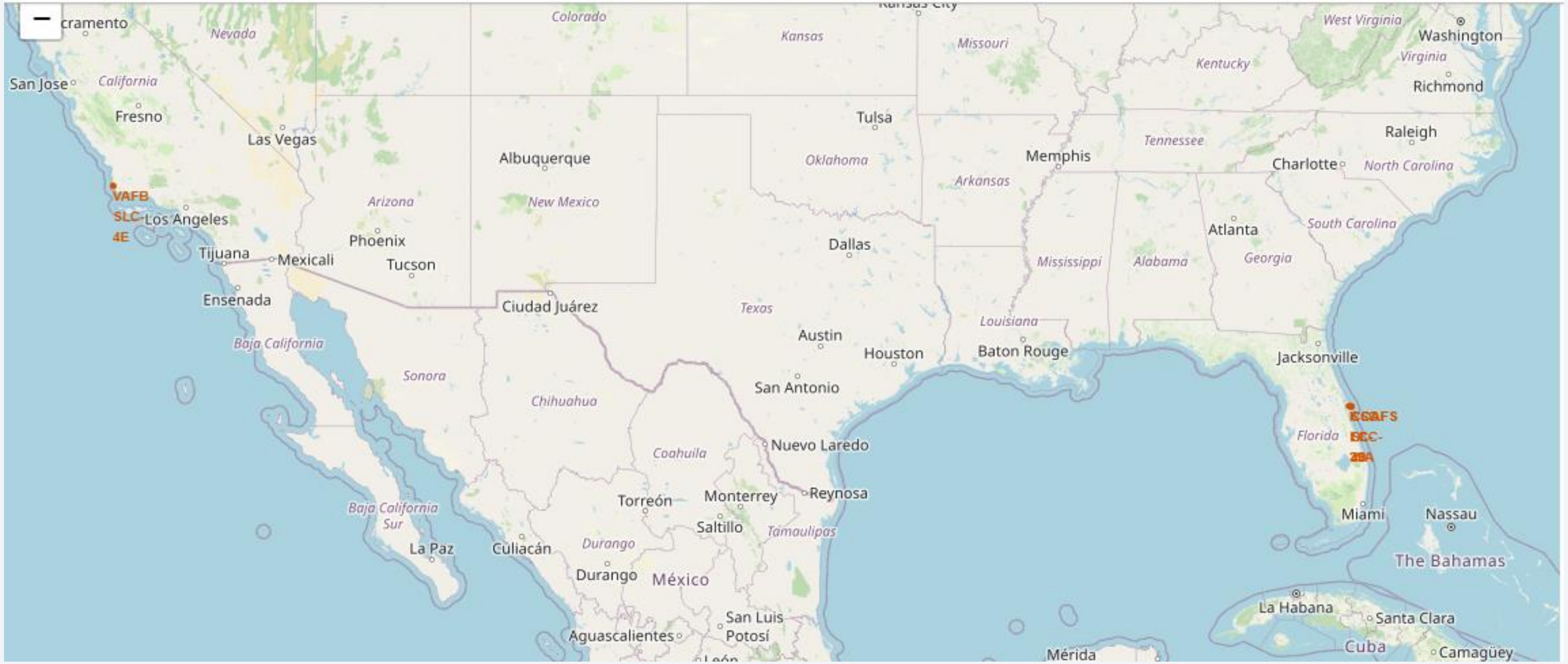
| Date | Landing_Outcome | TOTAL_NUMBER |
|---|---|---|
| 08/07/2018 | Success | 20 |
| 10/08/2012 | No attempt | 9 |
| 04/08/2016 | Success (drone ship) | 8 |
| 18/07/2016 | Success (ground pad) | 7 |
| 14/04/2015 | Failure (drone ship) | 3 |
| 12/05/2018 | Failure | 3 |
| 06/04/2010 | Failure (parachute) | 2 |
| 18/04/2014 | Controlled (ocean) | 2 |
| 08/06/2019 | No attempt | 1 |

Section 3

# Launch Sites Proximities Analysis

# Launch sites' location with marker

- All launch sites are in very close proximity to the coast to the Equator line
- The closer launch site to the Equator line has additional natural booster that helps save the fuel cost.

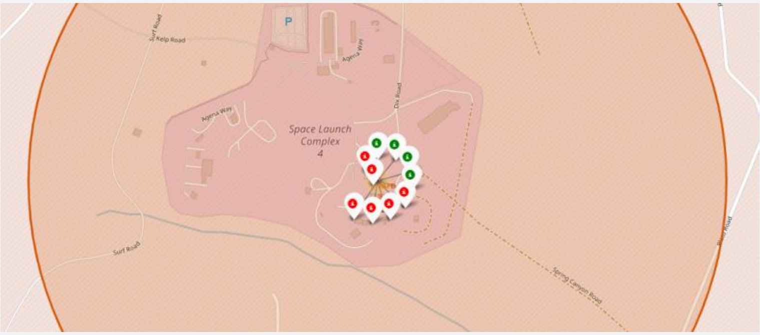


# Color-labeled launch outcomes

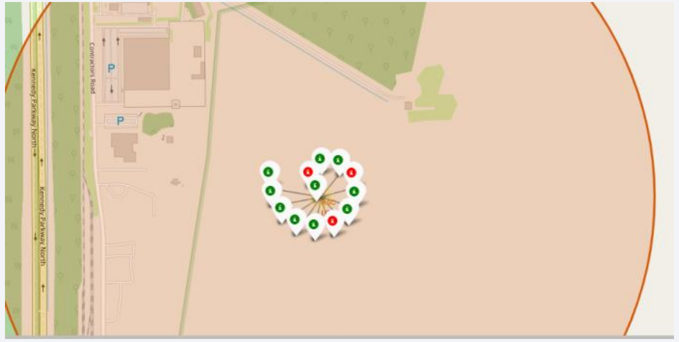
launch outcomes:

- Green for successful launches
- Red for unsuccessful launches

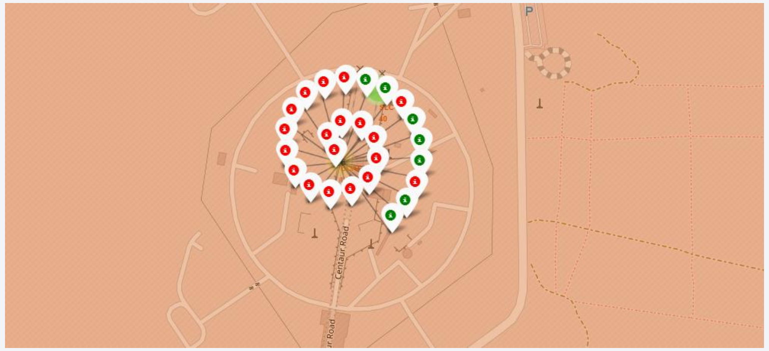
**VAFB SLC-4E** with 40% success rate



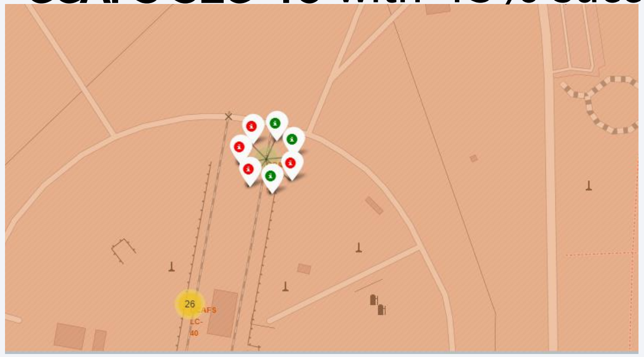
**KSC LC-39A** with 77% success rate



**CCAFS LC-40** with 27% success rate



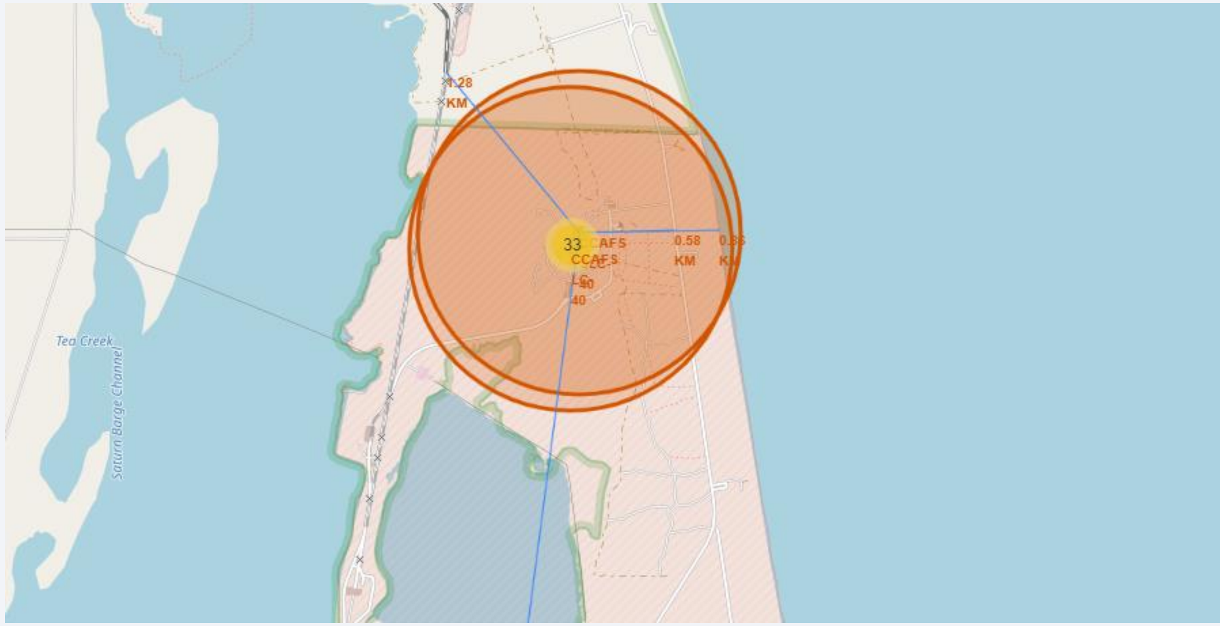
**CCAFS SLC-40** with 43% success rate

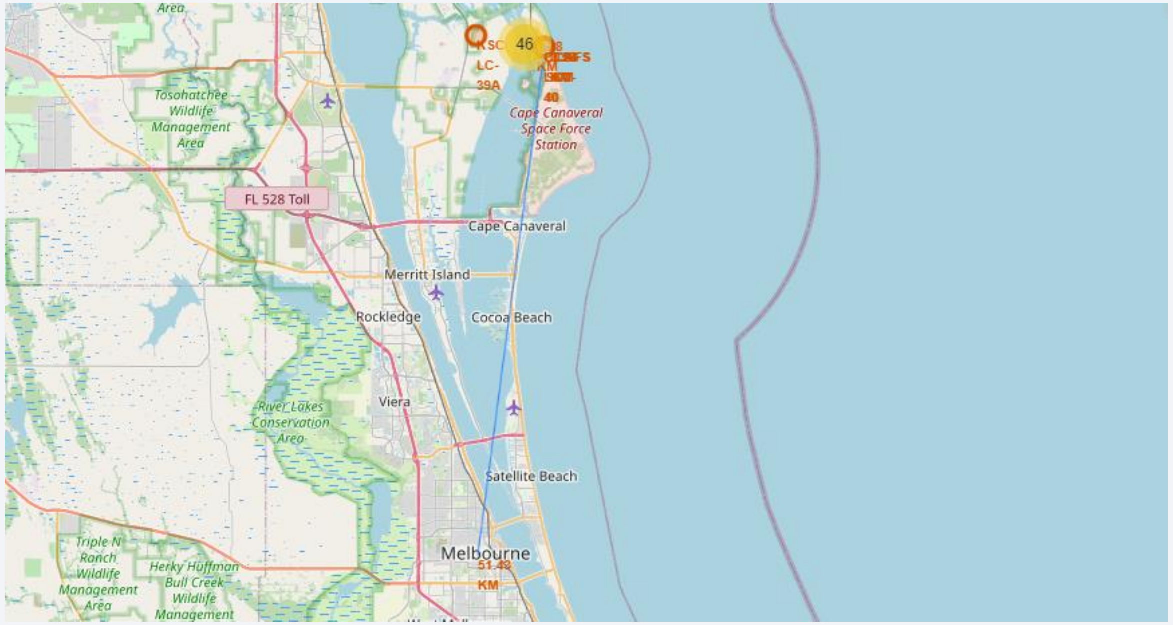


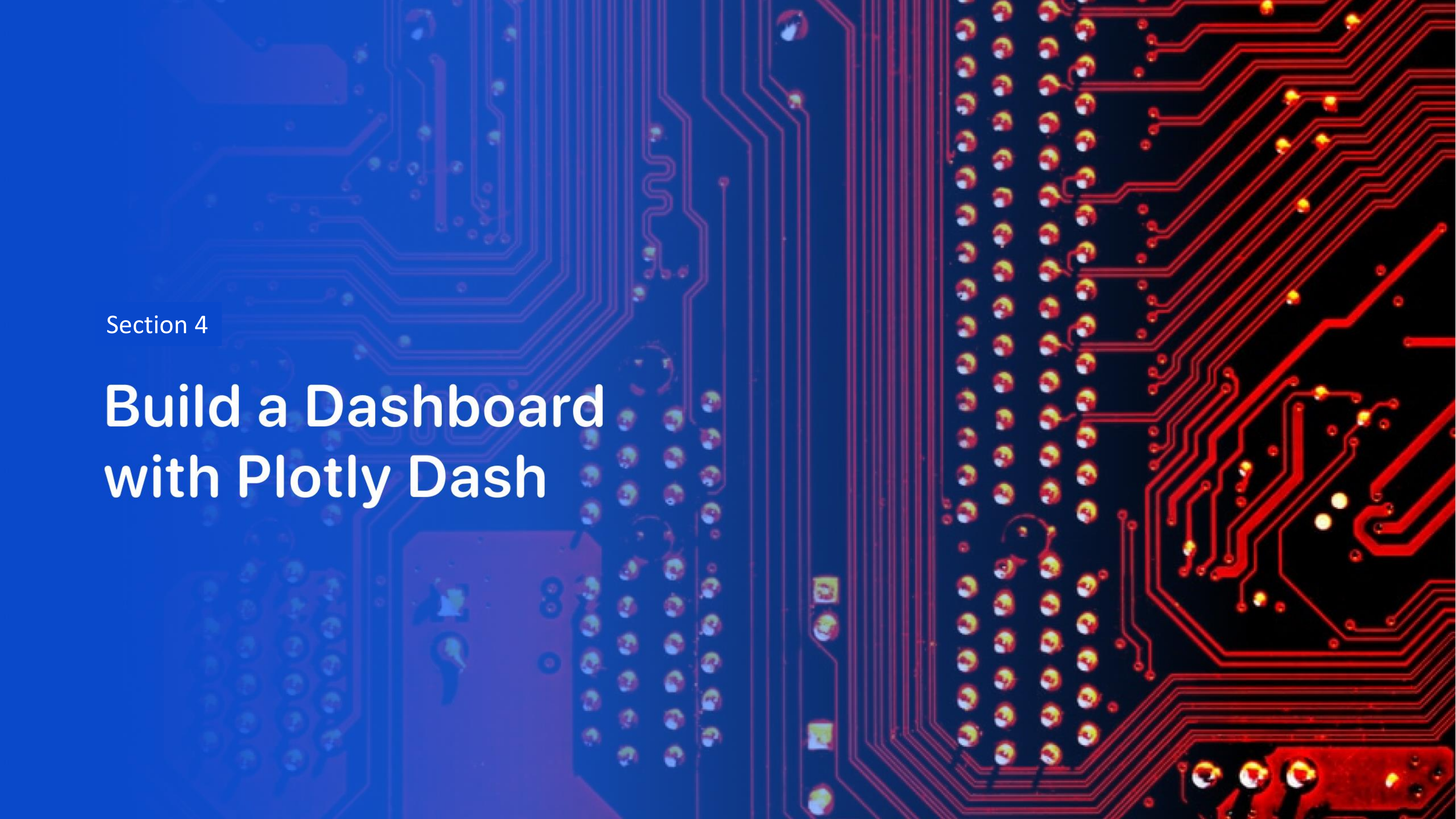
# Distance to Proximities

Launch site to its proximities

- 0.58  km from nearest coastline
- 1.28  km from nearest railroad
- 51.43  km from Melbourne




46
KSC
LC-
39A
CCAFS
40
Cape Canaveral Space Force Station
Tosohatchee Wildlife Management Area
FL 528 Toll
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
Viera
River Lakes Conservation Area
Satellite Beach
Melbourne
51.43
KM
Triple N Ranch Wildlife Management Area
Herky Huffman Bull Creek Wildlife Management



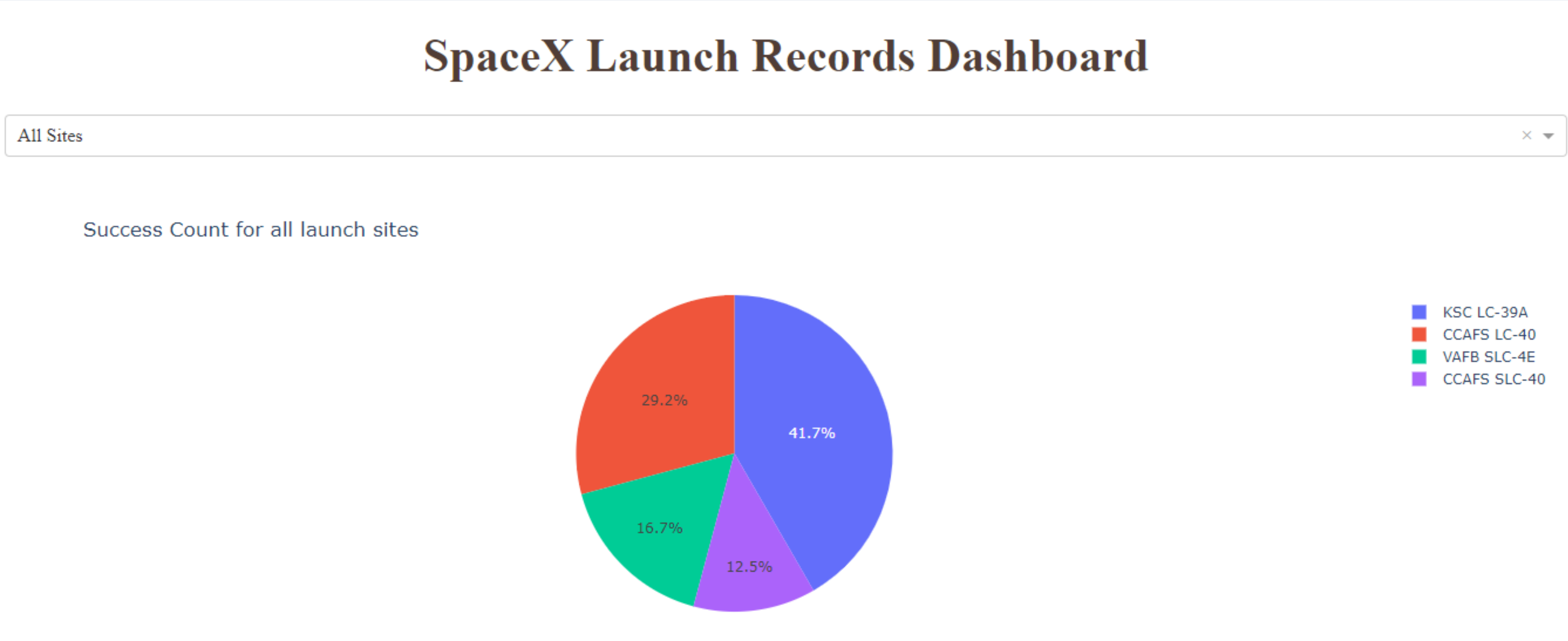
Section 4

# Build a Dashboard with Plotly Dash

# Launch Success for all sites

Successful percent of total launches

- KSC LC-39A has the most successful percent with 41.7%



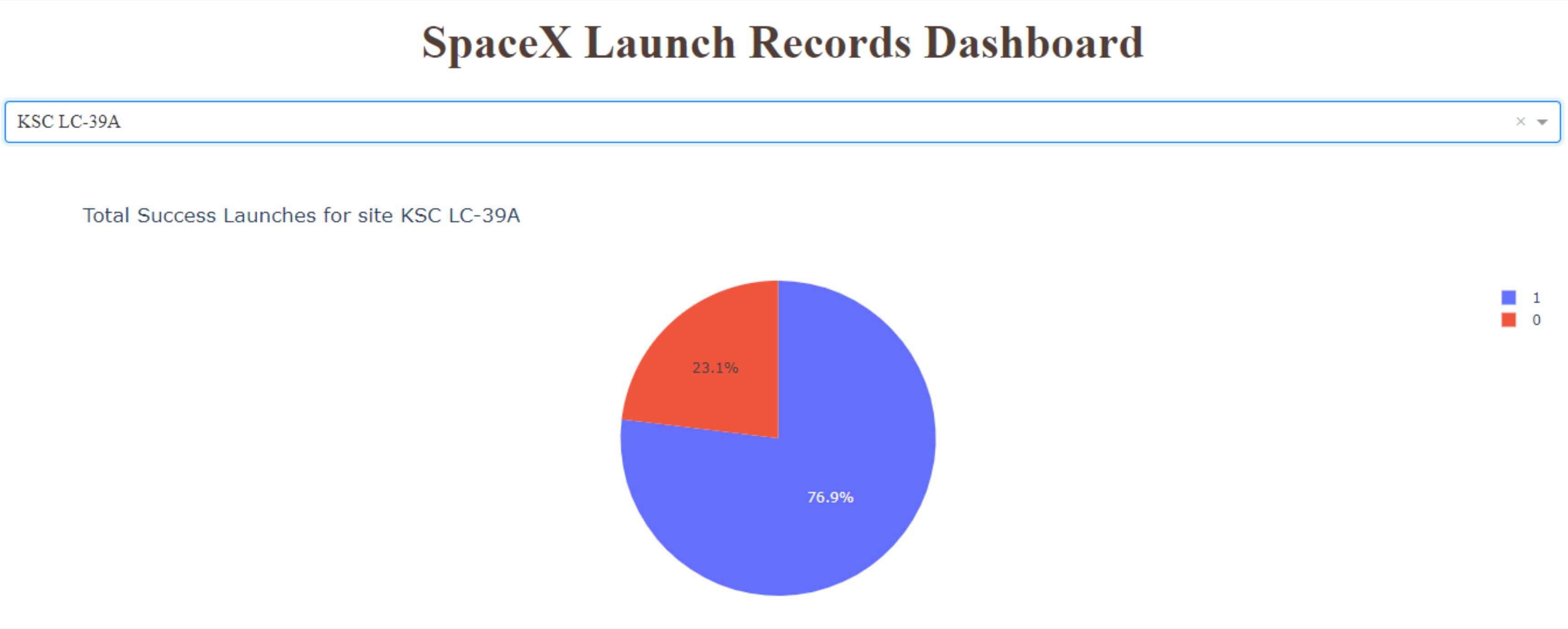
SpaceX Launch Records Dashboard

All Sites

# Launch success of KSC LC-39A

Successful percent of total launches

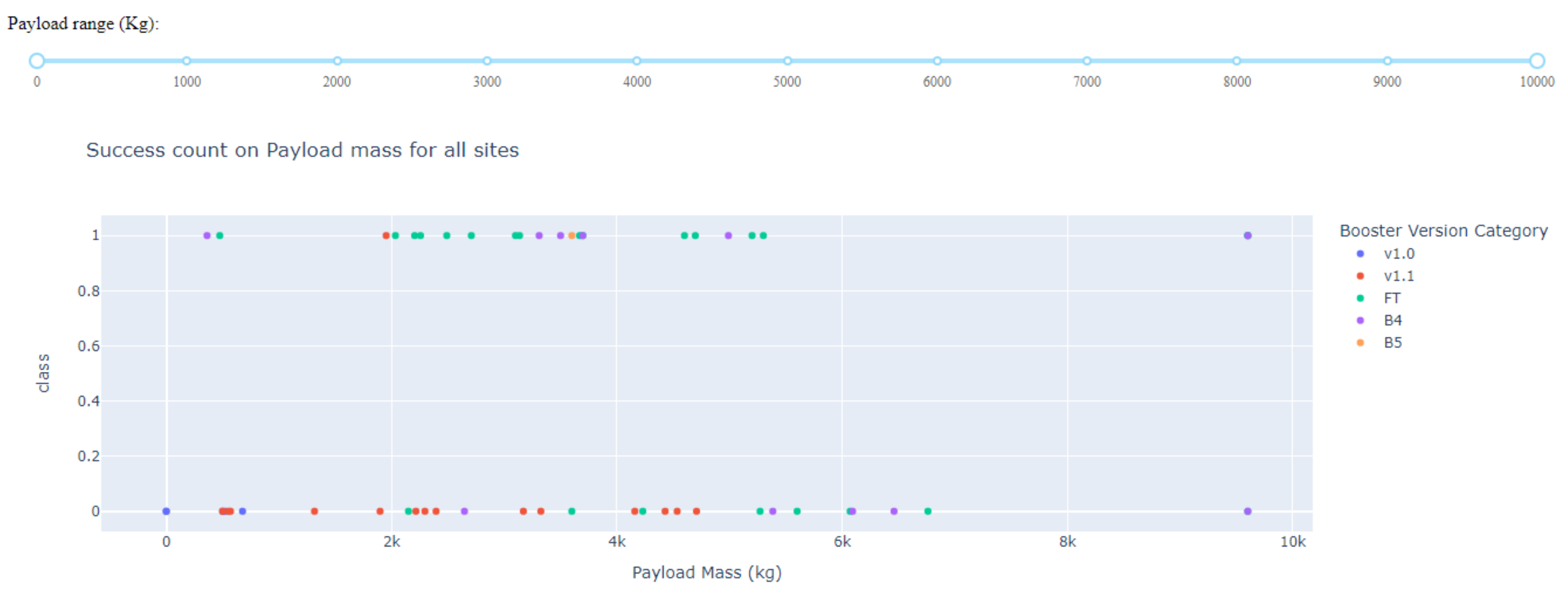
- KSC LC-39A has the highest successful rate with 76.9%
- 10 successful launches of 13 total launches



SpaceX Launch Records Dashboard

# Payload Mass and Launch Success

- Payloads in range 2000 kg and 5000 kg have the highest success rate
- FT booster version has the highest success rate

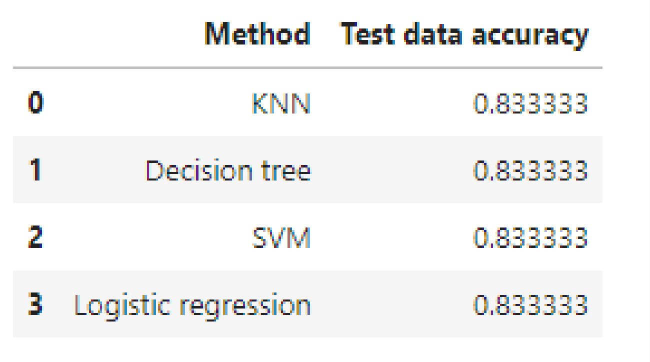


Section 5

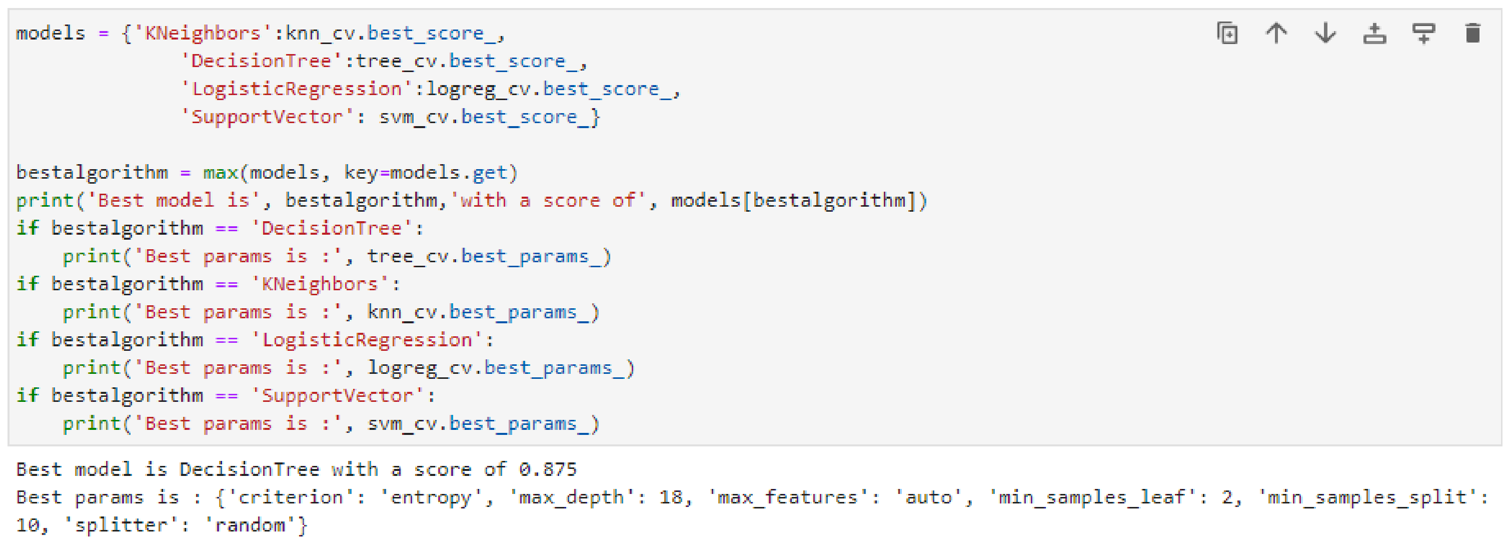
# Predictive Analysis (Classification)

# Classification Accuracy

- Landing outcome prediction using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)
- All built classification models have the same test data accuracy. It is possible that comes from small dataset.
- The decision tree model is the best model when using .best_score_



|   | Method | Test data accuracy |
|---|---|---|
| 0 | KNN | 0.833333 |
| 1 | Decision tree | 0.833333 |
| 2 | SVM | 0.833333 |
| 3 | Logistic regression | 0.833333 |



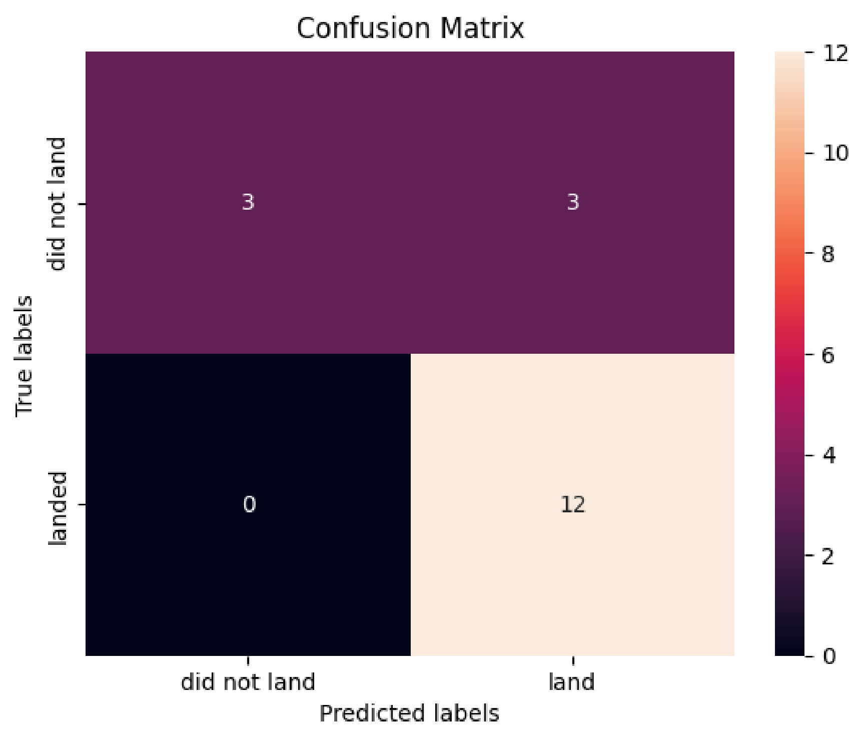
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'entropy', 'max_depth': 18, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split':
10, 'splitter': 'random'}
```

# Confusion Matrix - Decision tree

- Confusion matrix model is used to define the performance of a classification algorithm
- Outcomes:
  - 12 True Positive
  - 3 True Negative
  - 3 False Positive
  - 0 False Negative
- Precision: TP/(TP + FP) = 12/15 = 0.80
- Recall: TP/(TP + FN) = 12/12 = 1
- F1 score: 2 * (Precision * Recall) / (Precision + Recall)

  = 2*(0.8+1)/(0.8+1) = 0.89
- Accuracy: (TP + TN) / (TP+TN + FP + FN) = 0.83



# Conclusions

- **Model performance:** All built classification models have the same test data accuracy with the decision tree model slightly outperforming
- **Launch sites:** All launch sites are in very close proximity to the coast to the Equator line
- **Launch success:** improves over time
- **KSC LC-39A:** has the highest successful rate with 76.9%
- **Orbits:**100% success rate are ES-L1, GEO, HEO, and SSO
- **Payload Mass:** The successful landing rate increases with heavy payloads

# Thank you!