IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gia Busyra
27 April 2024

Outline

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

Executive Summary

Summary of methodologies

This project wants to determine the factors for a successful rocket landing. This is the following methodologies were used:

- Collect data using REST API and web scraping techniques
- Wrangle data to create a success/fail outcome variable
- EDA and data visualization techniques to find insights
- Calculate and aggregate the data using SQL
- Explore launch site success rates and proximity to geographical markers using Folium
- Visualize the launch sites with the most success and successful payload ranges
- Build Models to predict landing outcomes using logistic regression, SVM, decision tree, and KKN

Summary of all results

EDA:

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics: Most launch sites are near the equator, and all are close to the coast

Predictive Analytics: All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Project background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the International Space Station, launching a satellite constellation that provides internet access, and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Problems

- What factor best affects successful landing?
- How is the rate of successful landings over time?
- What is the best predictive model for a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data using REST API and web scraping techniques
- Perform data wrangling
  - Filtering the data, handling missing values, and applying one hot encoding to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predict landing outcomes using classification models. Tune and evaluate models to find the best model and parameters

Data Collection – SpaceX API

- Request data response from SpaceX API
- Decode response using .json() and convert to a data frame using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create a dictionary from the data
- Create a data frame from the dictionary
- Filter data frame to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to CSV file

Data Collection - Scraping

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from the HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create data frame from the dictionary
- Export data to CSV file

Data Wrangling

- Perform EDA and determine data labels
- Calculate:
    - # of launches for each site
    - # and occurrence of orbit
    - # and occurrence of mission outcome per orbit type
- Create a binary landing outcome column (dependent variable)
- Export data to CSV file

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- View the relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

EDA with SQL

- Names of unique launch sites
- 5 records where the launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- Date of first successful landing on the ground pad
- Names of boosters that had success landing on drone ships and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions that have carried the max payload
- Failed landing outcomes on drone ship, their booster version, and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Build an Interactive Map with Folium

- Added a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch site coordinates with a popup label showing its name using its latitude and longitude coordinates
- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rate
- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites. Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range. Allow user to select payload mass range
- Pie Chart Showing Successful Launches. Allow users to see successful and unsuccessful launches as a percentage of the total
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version. Allow users to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score and Accuracy

Results

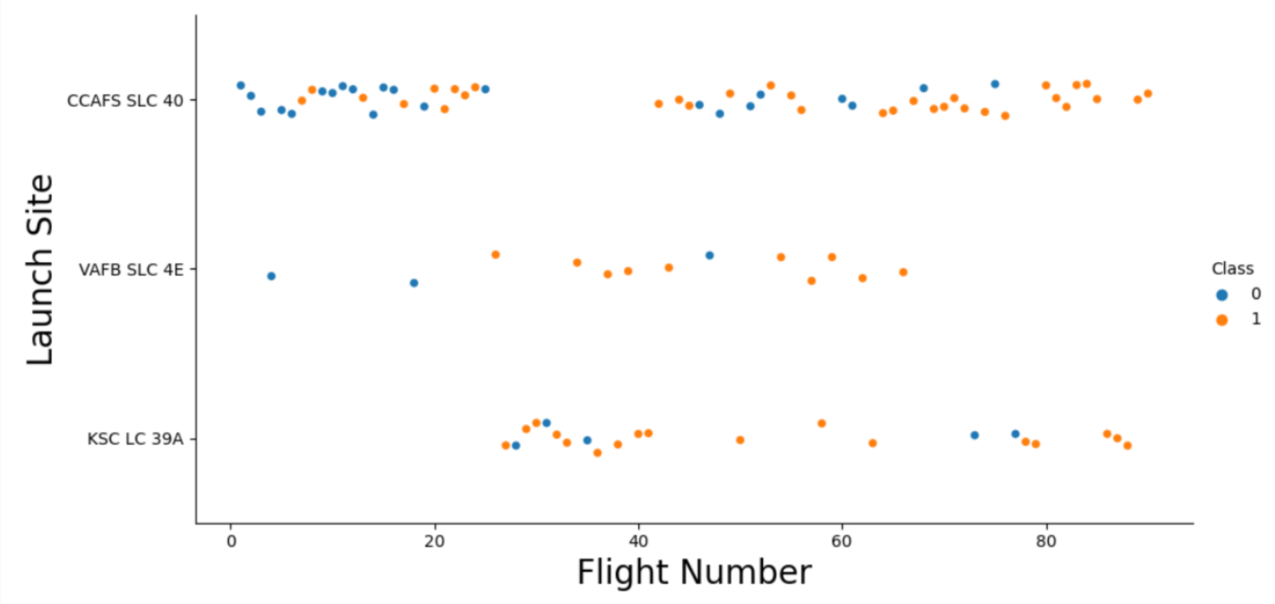
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

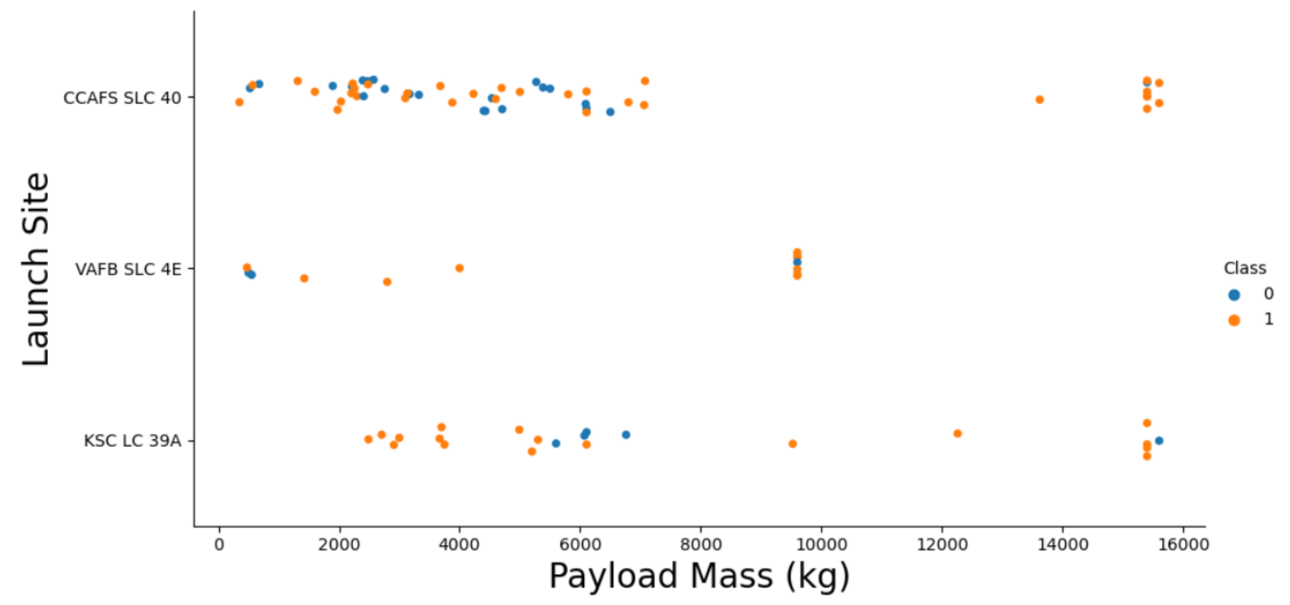
Flight Number vs. Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



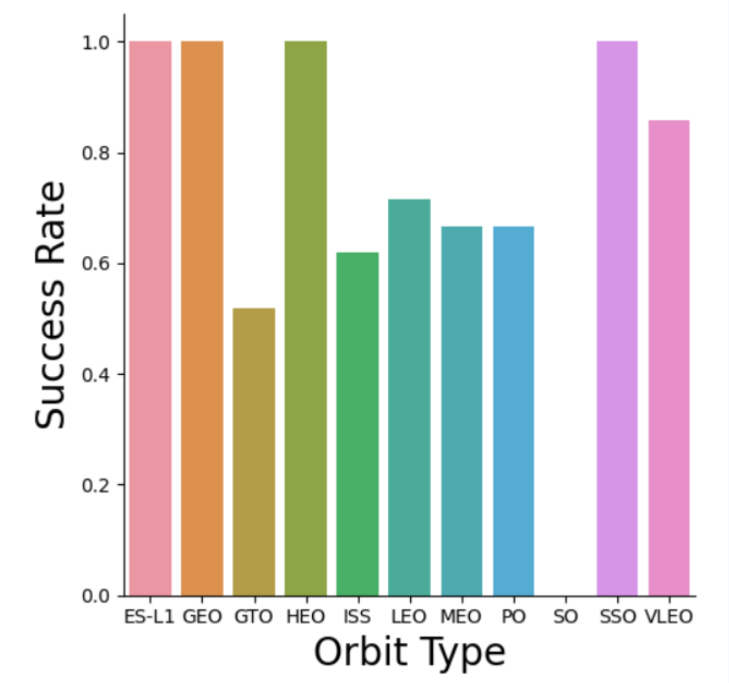
Payload vs. Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



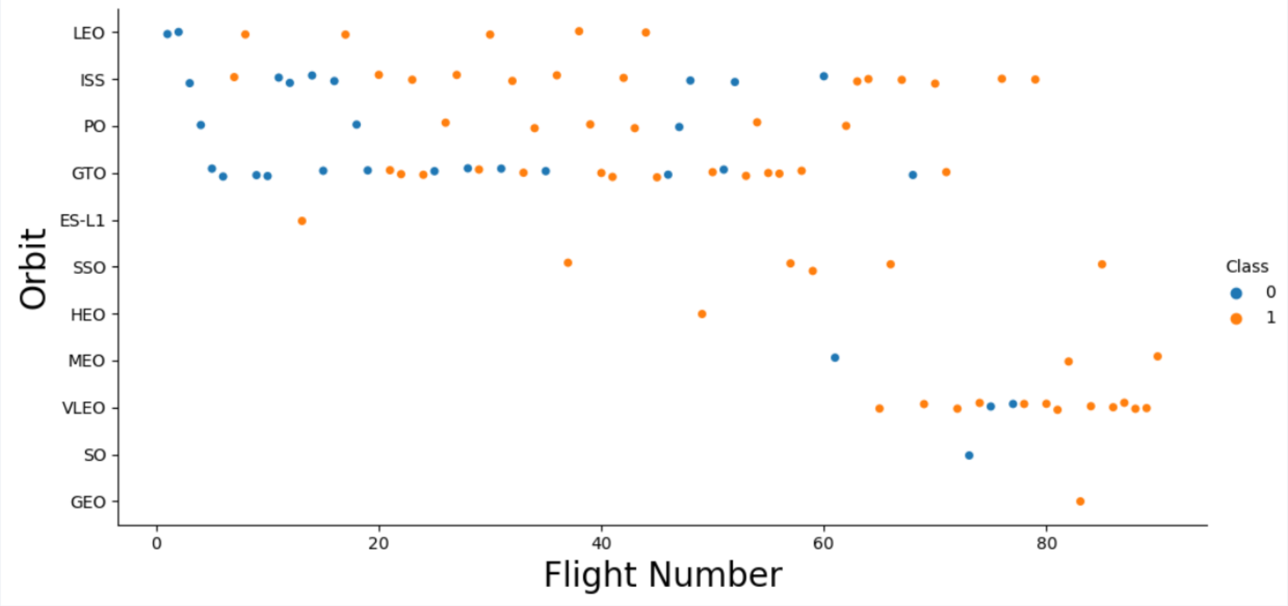
Success Rate vs. Orbit Type

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



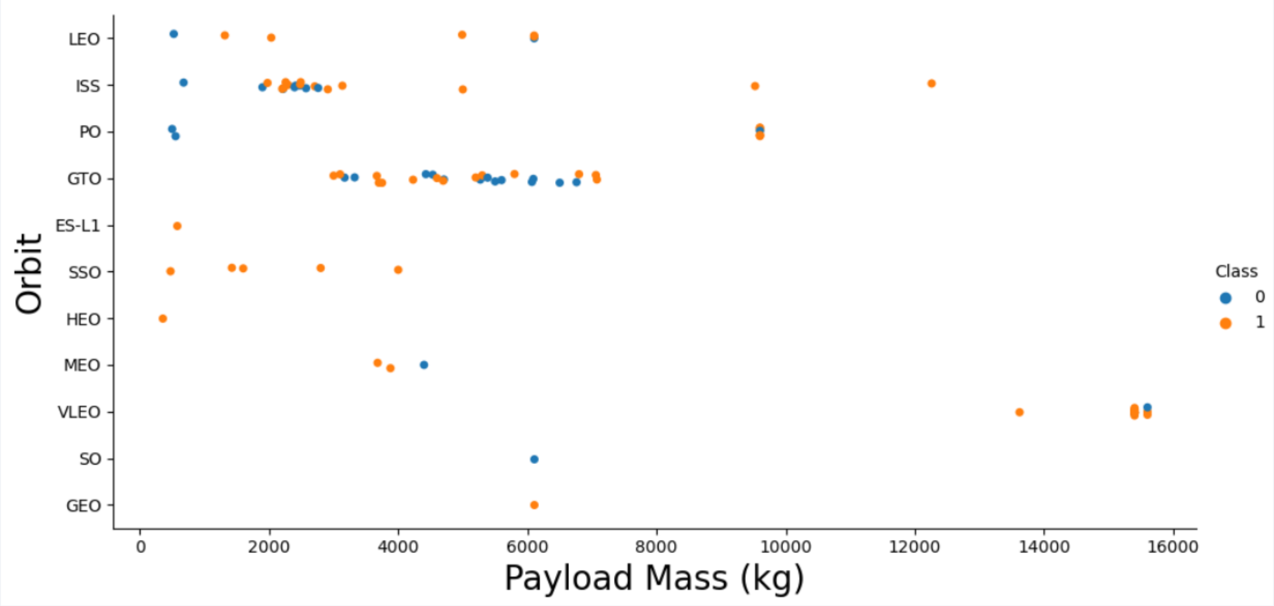
Flight Number vs. Orbit Type

- The success rate typically increases with the number of flights for each orbit
- VLEO has high success rate in higher flights
- The GTO orbit, however, does not follow this trend



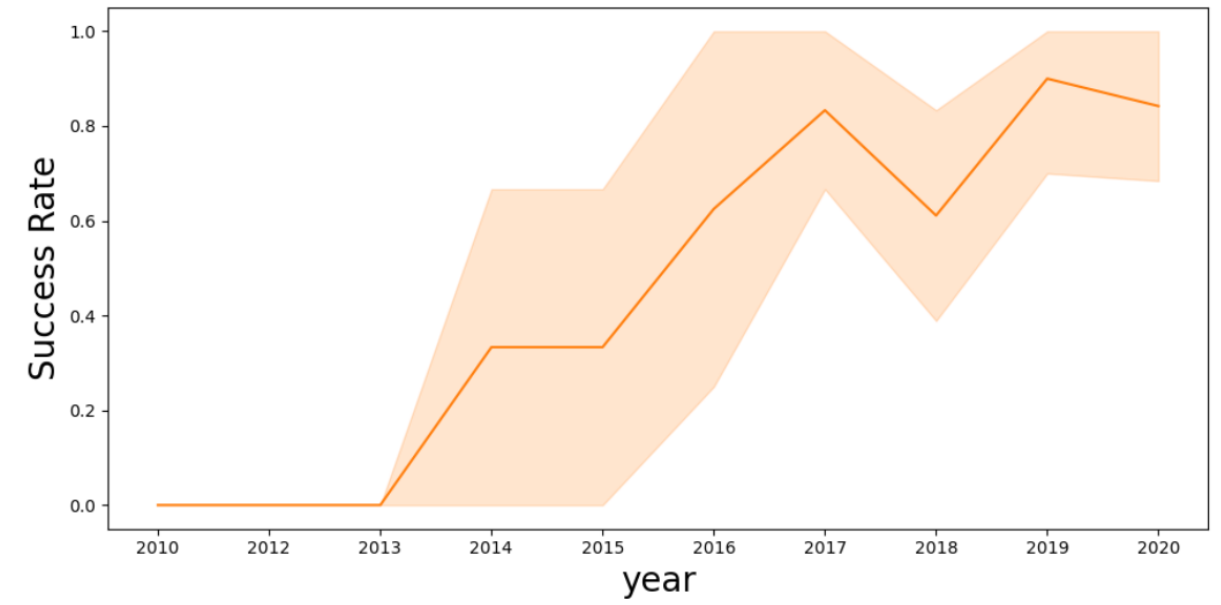
Payload vs. Orbit Type

- VLEO has the heaviest payloads
- The GTO orbit has mixed success with heavier payloads



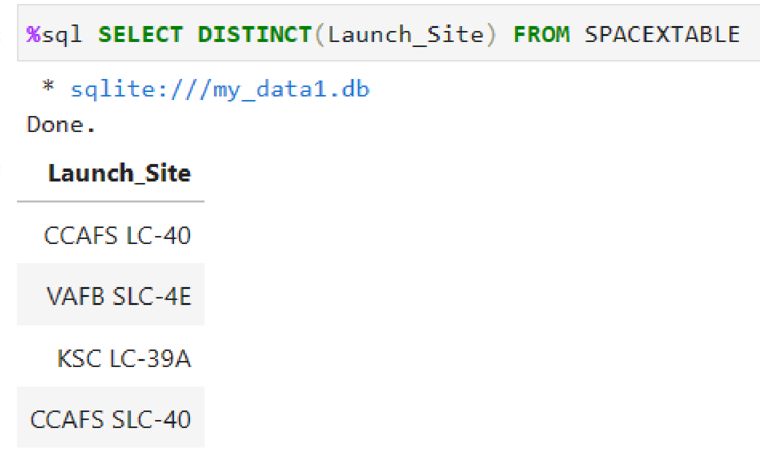
Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

Launch Site Names



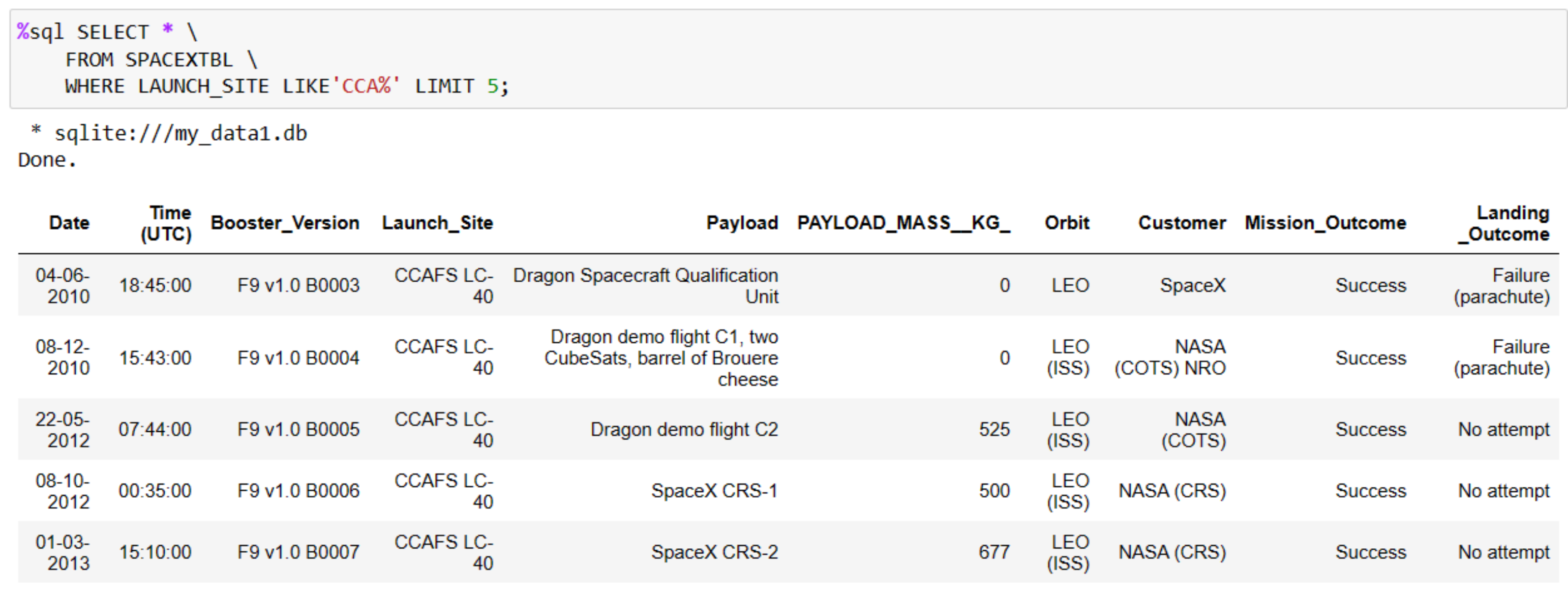
```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

```
 * 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'

Launch Site Names with start “CCA”



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

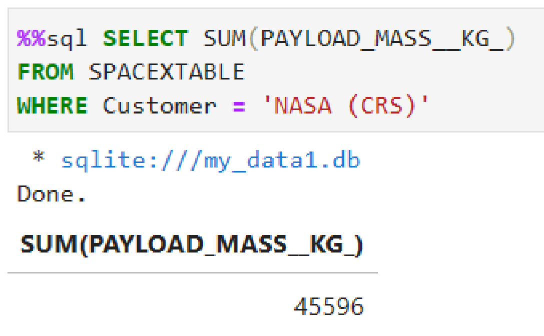
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total Payload Mass

45,596 kg (total) carried by boosters launched by NASA (CRS)



```
%%sql SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

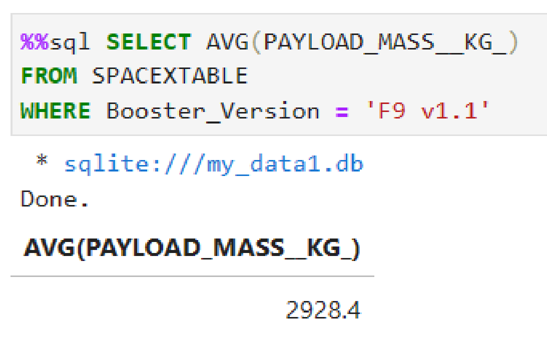
```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Average Payload Mass by F9 v1.1

2,928 kg (average) carried by booster version F9 v1.1



```
%%sql SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1'
```

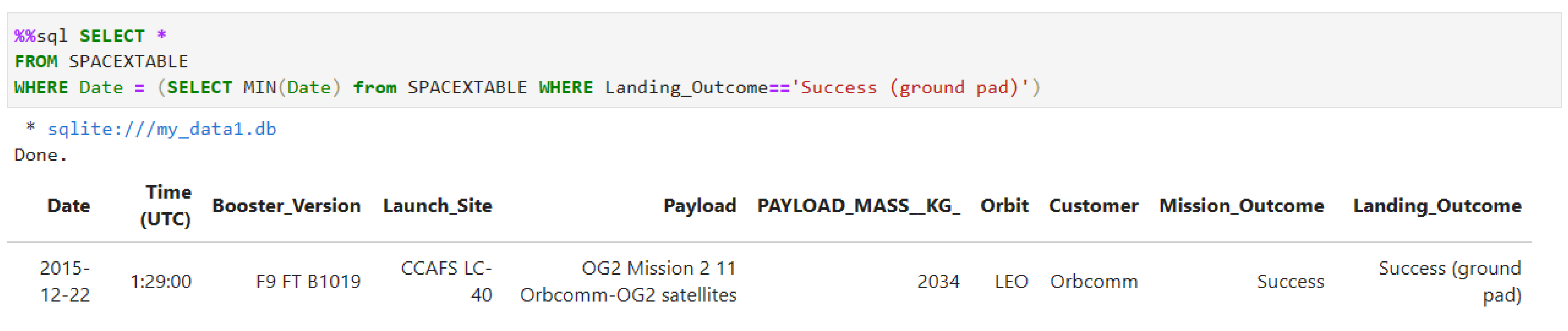
```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

First Successful Ground Landing Date

Successful landing at ground pad at 2015-12-22



```
%%sql SELECT *
FROM SPACEXTABLE
WHERE Date = (SELECT MIN(Date) from SPACEXTABLE WHERE Landing_Outcome=='Success (ground pad)')
```

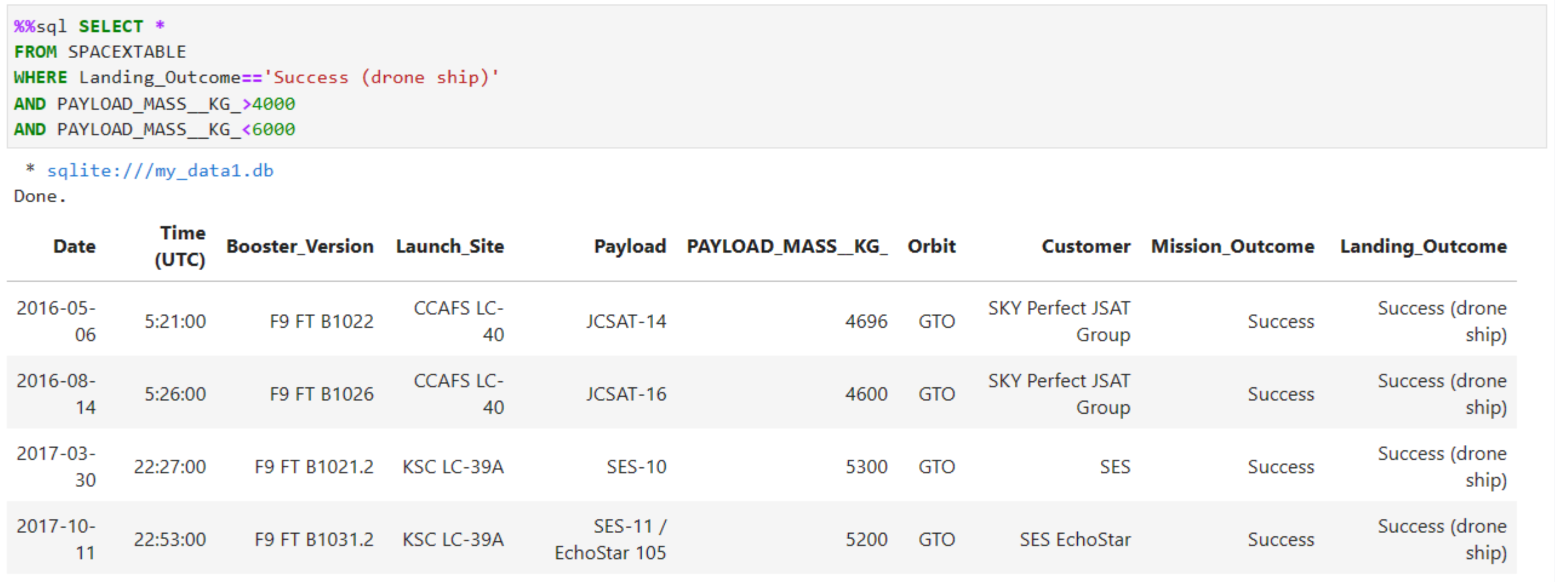
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

Successful Drone Ship Landing with Payload between 4000 and 6000

4 payloads including JSCAT-14, JSCAT-16, SES-10, and SES-11 / EchoStar 105



```
%%sql SELECT *
FROM SPACEXTABLE
WHERE Landing_Outcome=='Success (drone ship)'
AND PAYLOAD_MASS__KG_>4000
AND PAYLOAD_MASS__KG_<6000
```

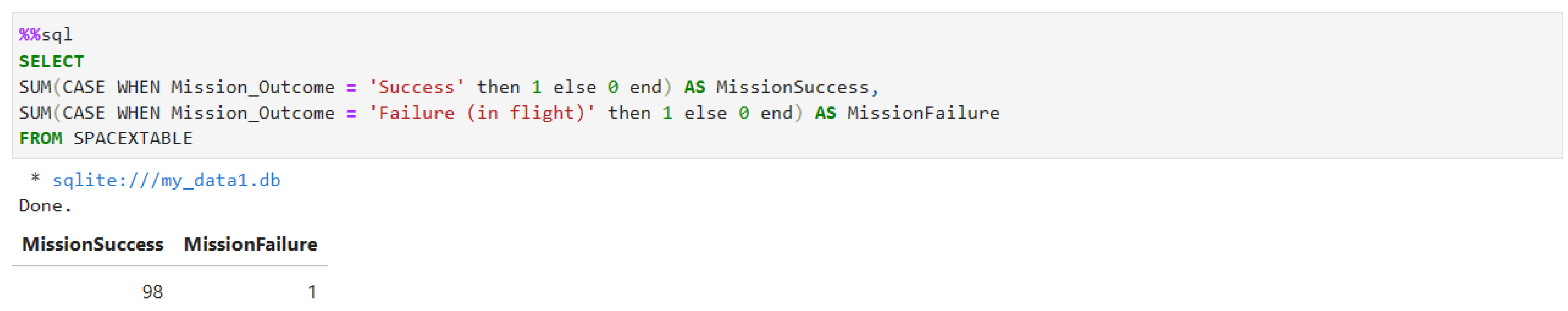
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failure Mission Outcomes

There are 99 Success Missions and 1 Failed Mission



```
%%sql
SELECT
SUM(CASE WHEN Mission_Outcome = 'Success' then 1 else 0 end) AS MissionSuccess,
SUM(CASE WHEN Mission_Outcome = 'Failure (in flight)' then 1 else 0 end) AS MissionFailure
FROM SPACEXTABLE
```

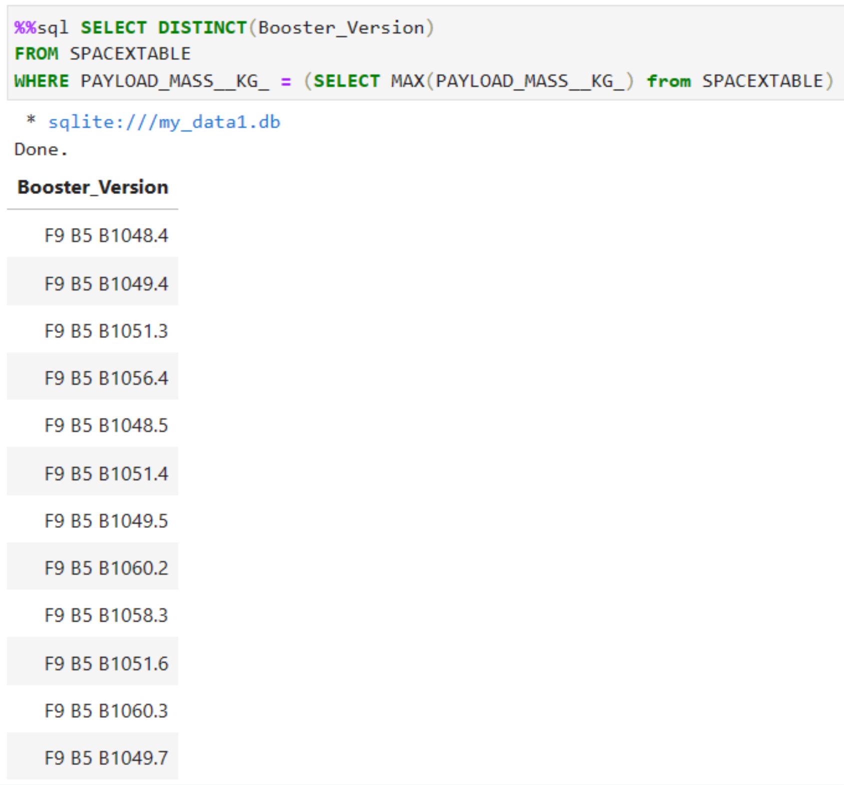
```
 * sqlite:///my_data1.db
Done.
```

MissionSuccess	MissionFailure
98	1

Boosters Carried Maximum Payload

Boosters Carried Maximum Payload

12 Payloads. Including 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



```
%%sql SELECT DISTINCT(Booster_Version)
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

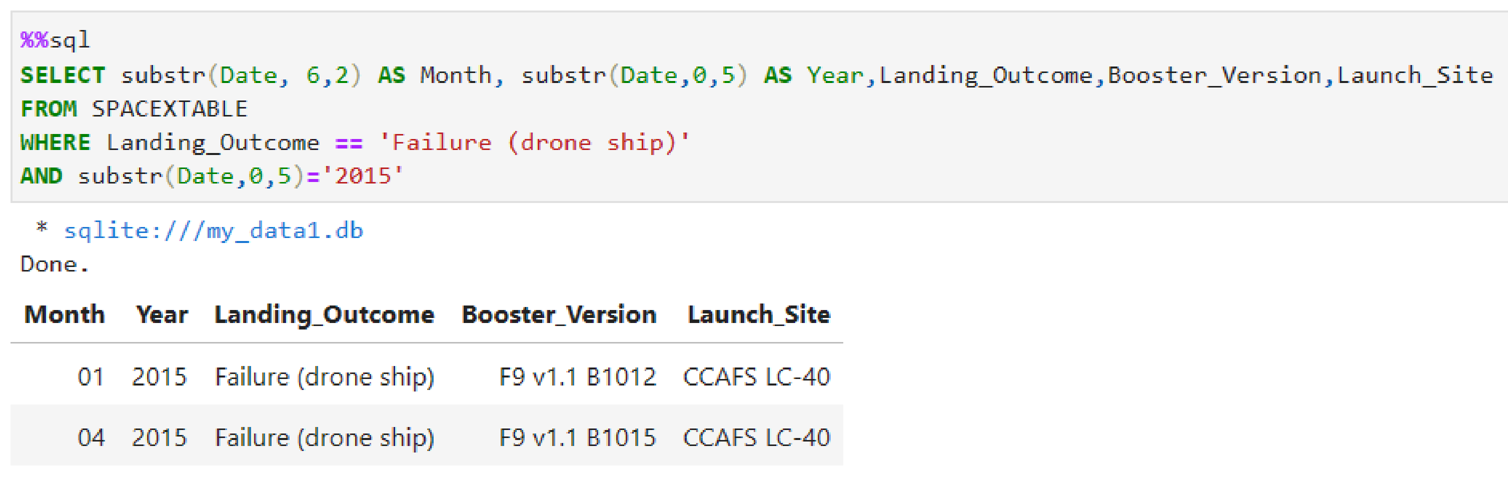
```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

2 records in 2015 with status on Landing Outcome is Failure (drone ship) and Booster Version: F9 v1.1 B1012 and F9 v1.1 B1015

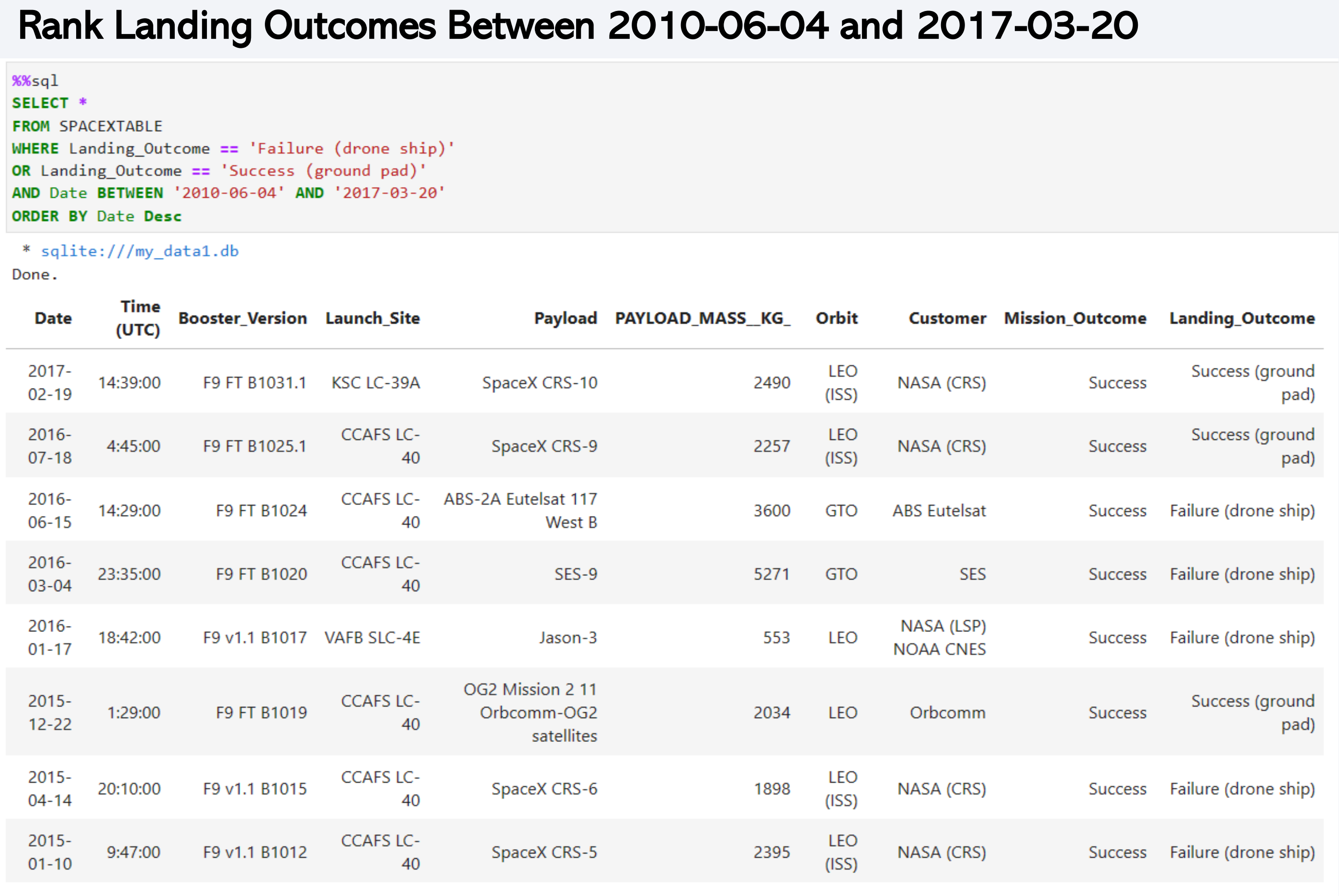


```
%%sql
SELECT substr(Date, 6,2) AS Month, substr(Date,0,5) AS Year,Landing_Outcome,Booster_Version,Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome == 'Failure (drone ship)'
AND substr(Date,0,5)='2015'
```

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

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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



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

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Landing_Outcome == 'Failure (drone ship)'
OR Landing_Outcome == 'Success (ground pad)'
AND Date BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY Date Desc
```

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

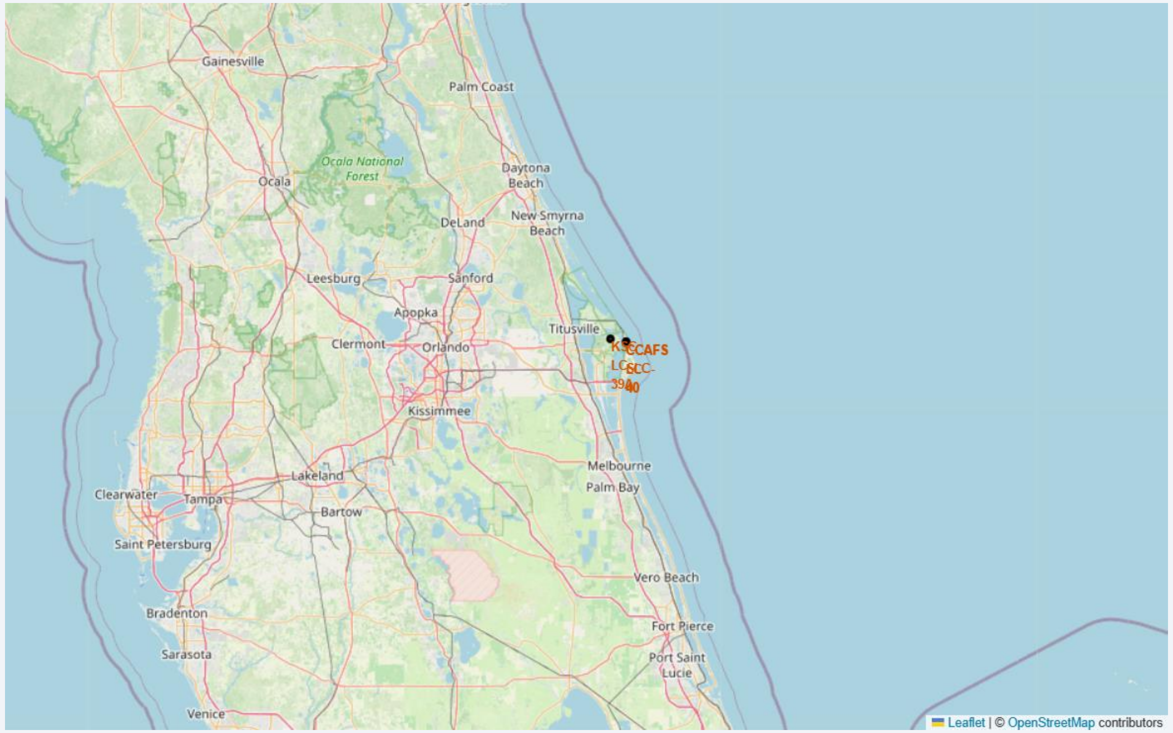
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-01-17	18:42:00	F9 v1.1 B1017	VAFB SLC-4E	Jason-3	553	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

Section 3

Launch Sites Proximities Analysis

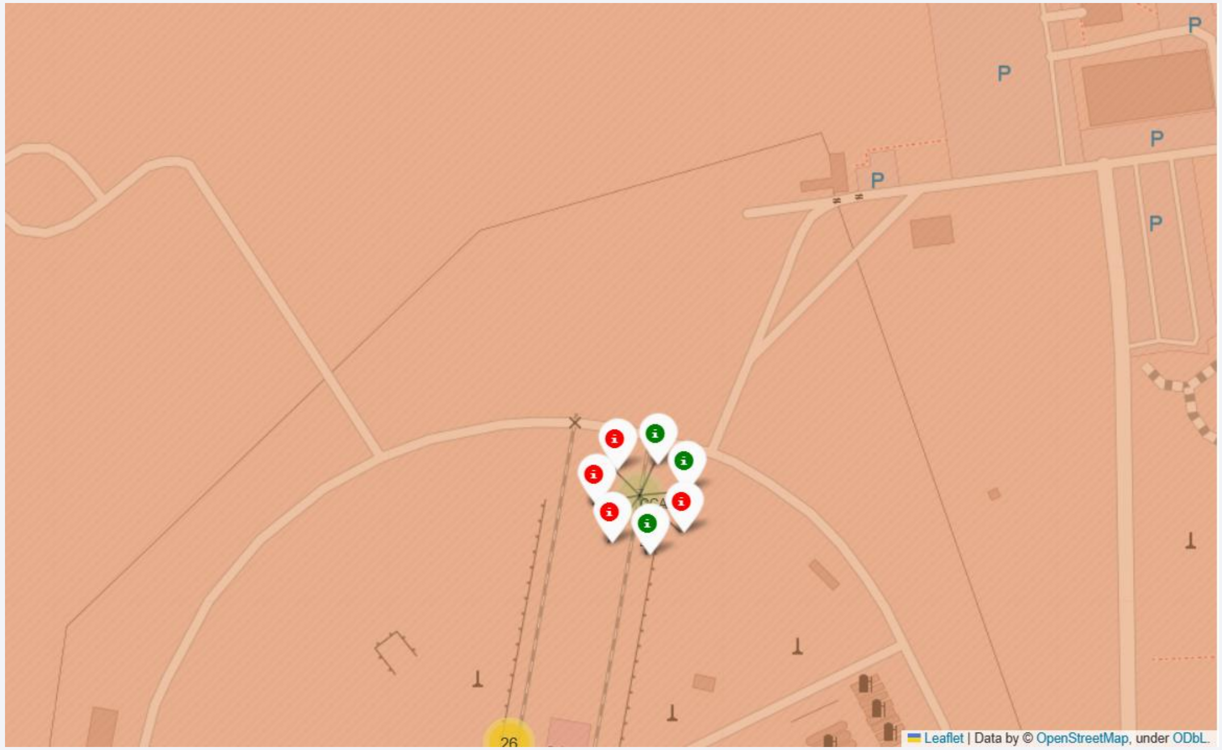
<Folium Map Screenshot 1>

Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



<Folium Map Screenshot 2>

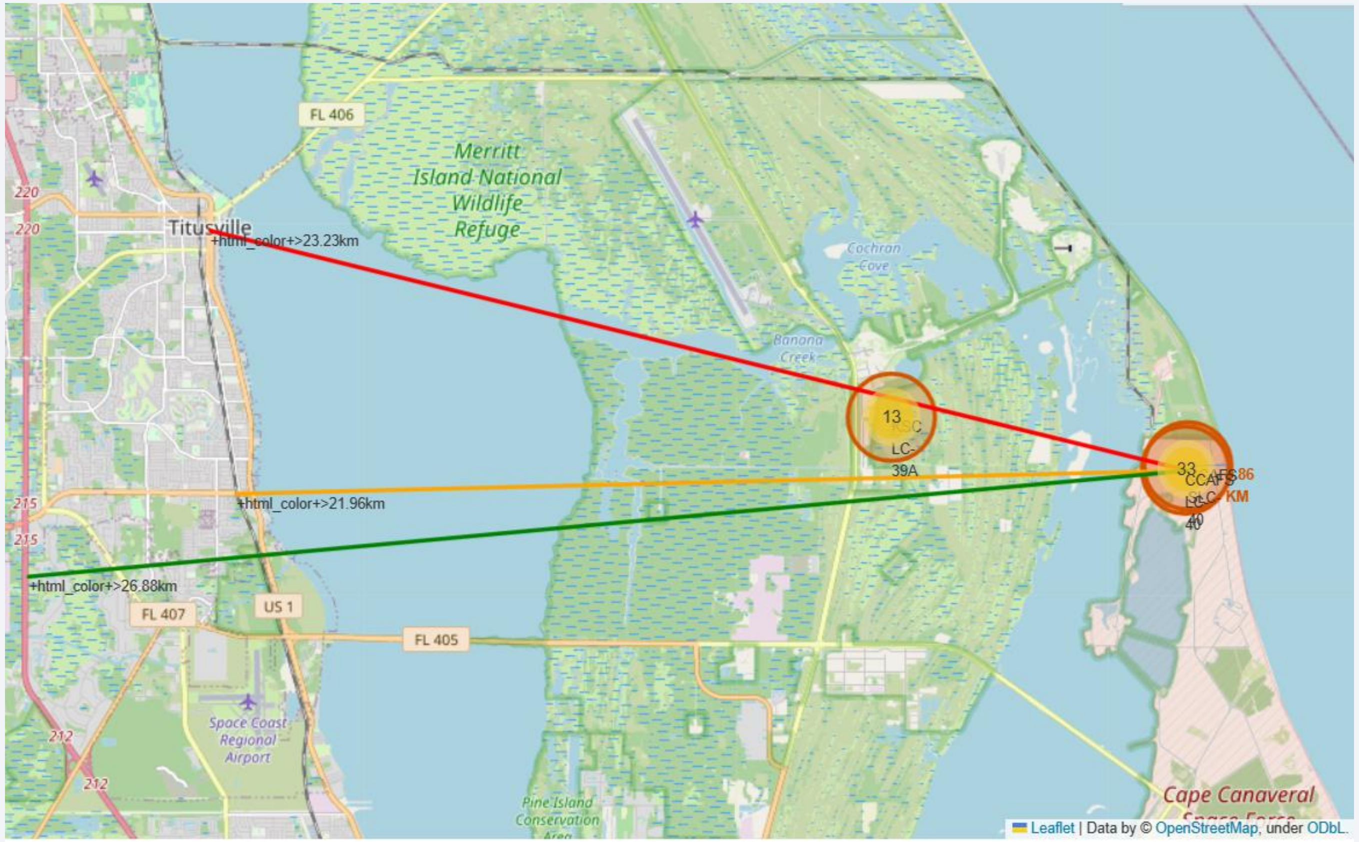
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



<Folium Map Screenshot 3>

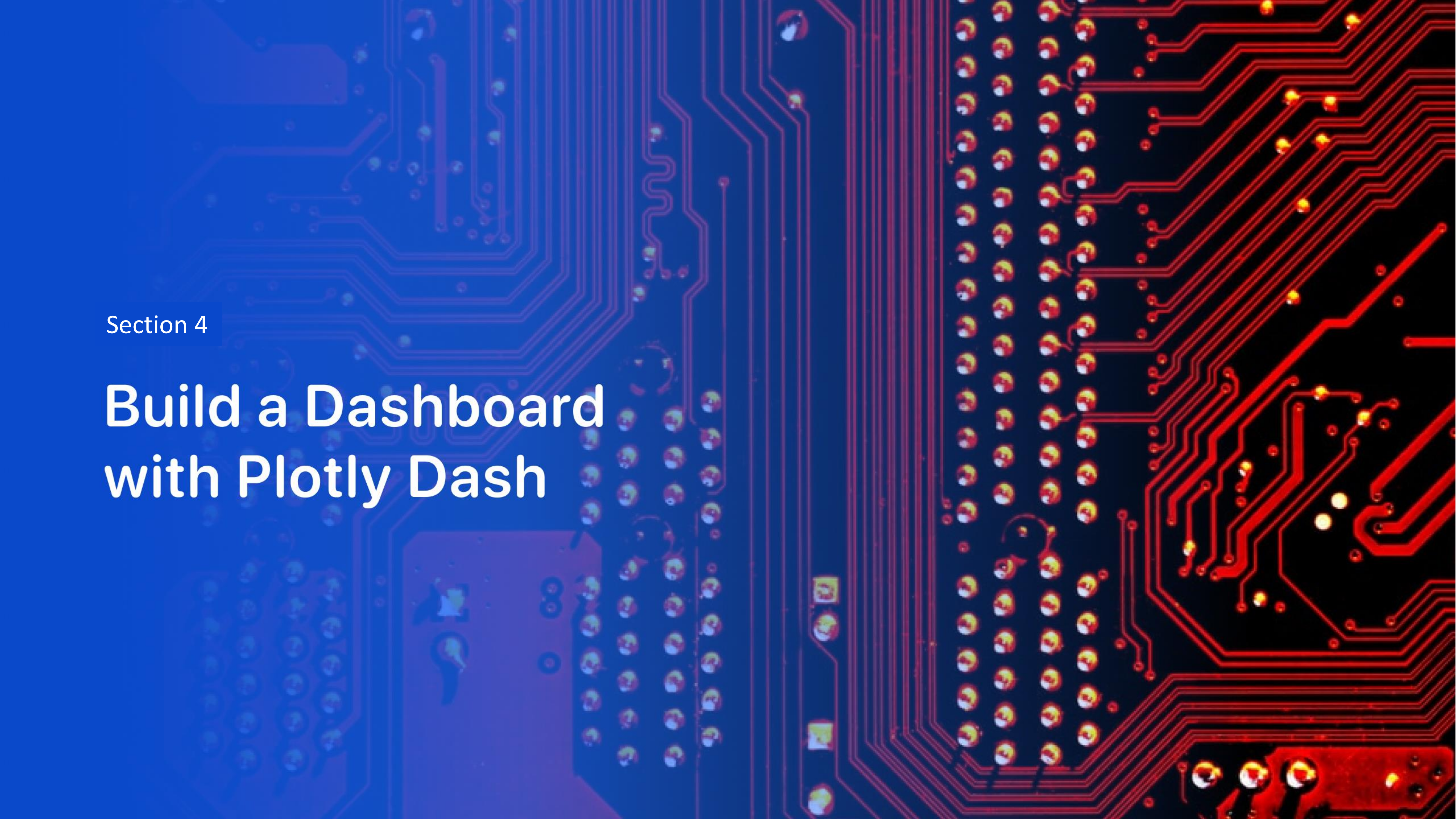
CCAFS SLC-40

- 86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway
- Coasts: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- Safety / Security: needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- Transportation/Infrastructure and Cities: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.



Section 4

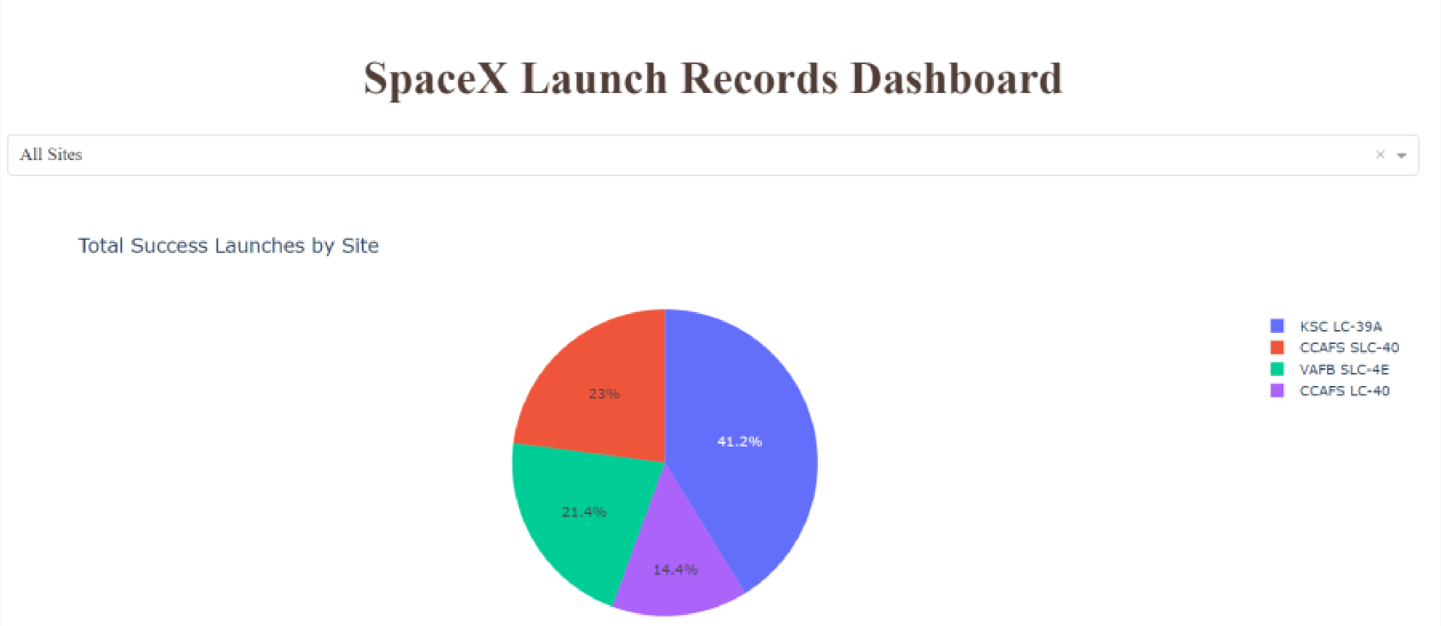
Build a Dashboard with Plotly Dash



<Dashboard Screenshot 1>

Success as Percent of Total

KSC LC-39A has the most successful launches amongst launch sites (41.2%)

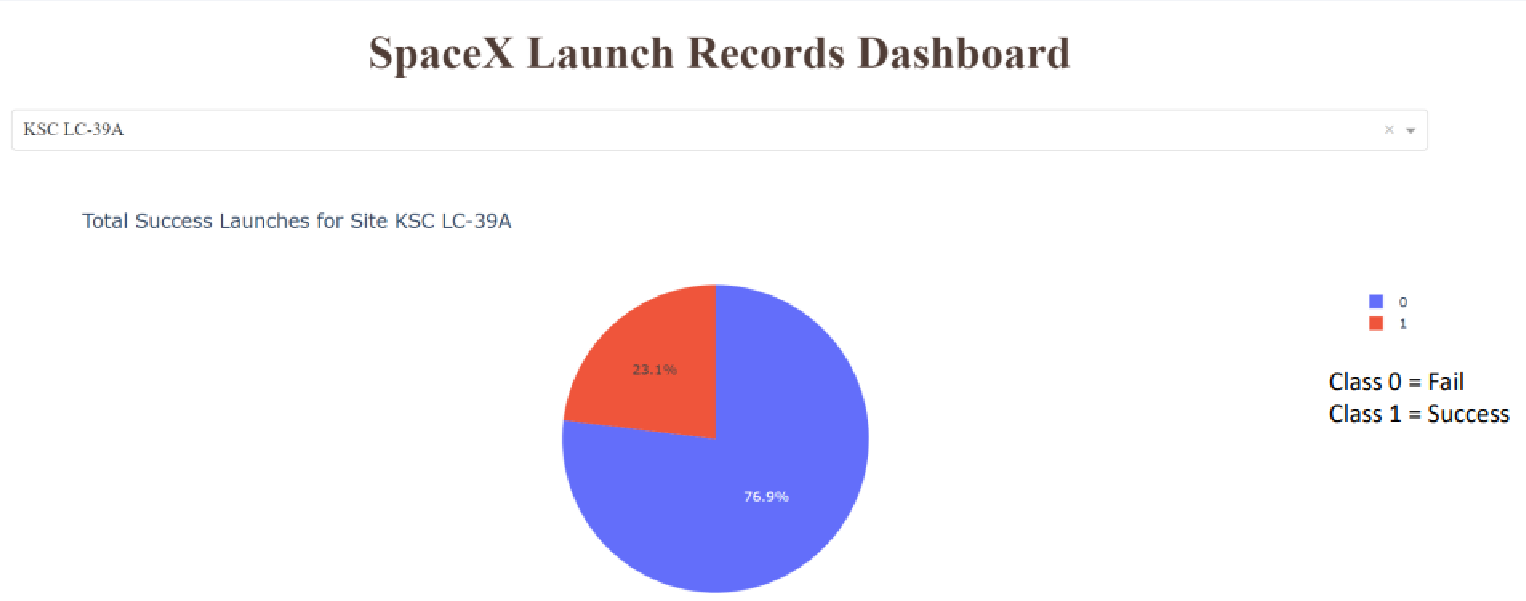


SpaceX Launch Records Dashboard

<Dashboard Screenshot 2>

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



SpaceX Launch Records Dashboard

KSC LC-39A

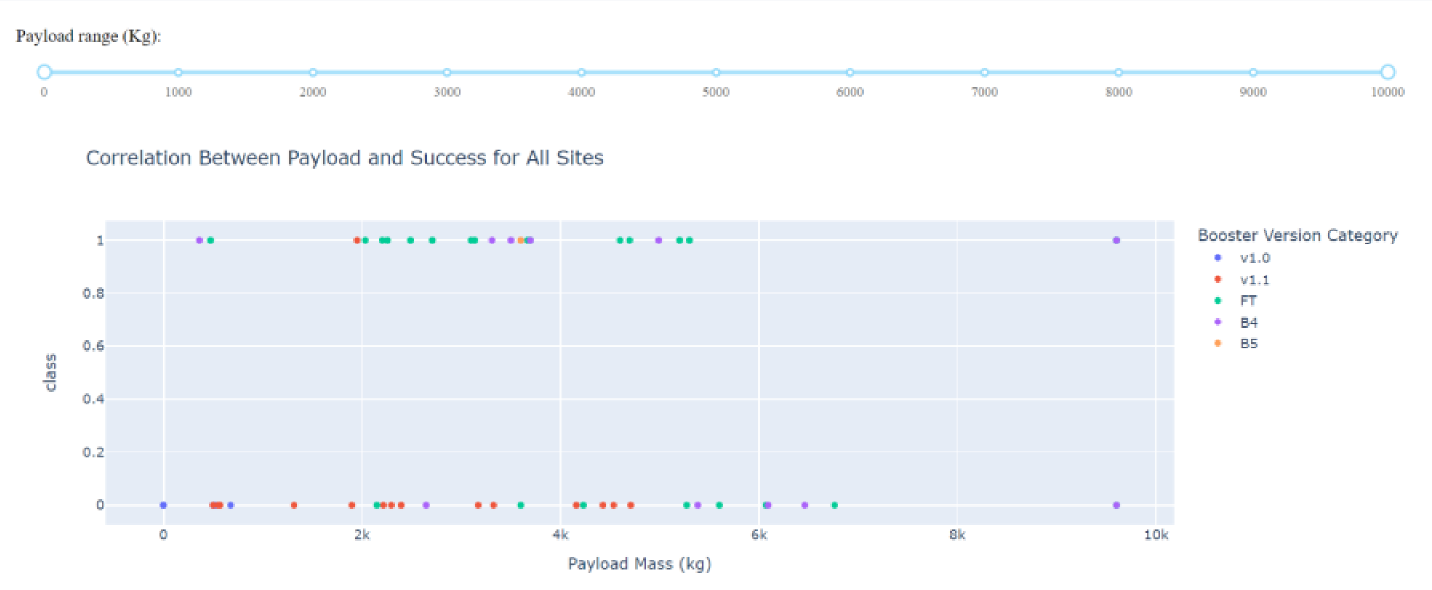
Total Success Launches for Site KSC LC-39A

Class 0 = Fail
Class 1 = Success

<Dashboard Screenshot 3>

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

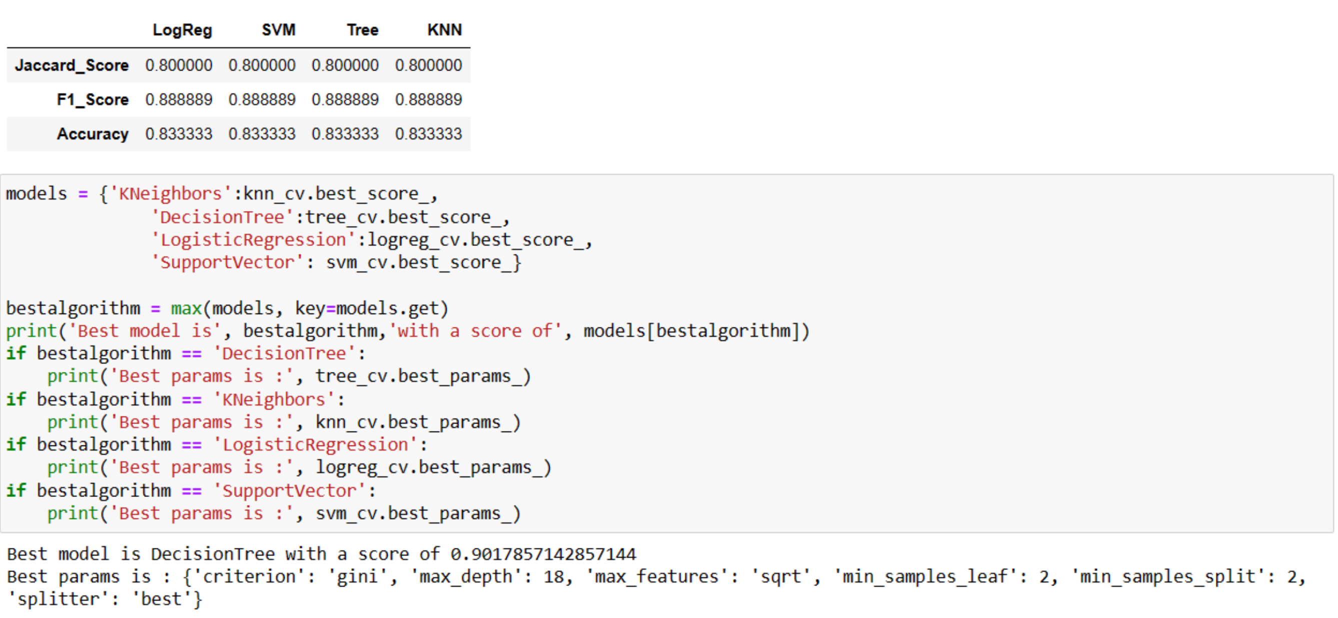


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at .best_score_
- .best_score_ is the average of all cv folds for a single combination of the parameters



	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

```
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 0.9017857142857144
Best params is : {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2,
'splitter': 'best'}
```

Confusion Matrix

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative

Precision = TP / (TP + FP)

- 12 / 15 = .80

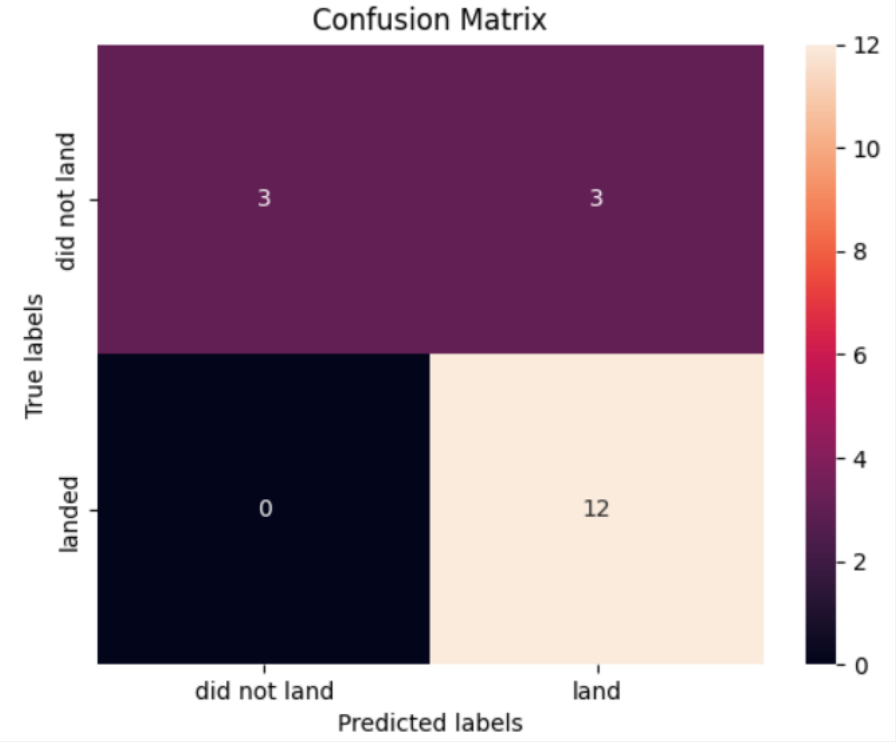
Recall = TP / (TP + FN)

- 12 / 12 = 1

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- 2 * (.8 * 1) / (.8 + 1) = .89

Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming
- Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- Coast: All the launch sites are close to the coast
- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!