IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jennifer Remington
January 17, 2024

Outline

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

Executive Summary

There is significant cost savings to rocket launches when the first stage of the launch can be reused. The following methodologies where used to determine the factors that impact the success of a launch and ultimately predict whether a launch will be successful in order to realize those cost savings.

- **Data Collection** via SpaceX API and web scraping of launch data related Wiki pages
- **Data Wrangling** to convert landing outcomes to a classification variable (0 or 1)
- **Exploratory Data Analysis** using SQL and Visualization techniques including a Folium Map
- **Interactive Visual Analytics** via a Plotly Dashboard
- **Predictive Analysis** via a machine learning pipeline

Key Findings

- Launch site KSC LC-39A has the highest success rate for launches
- Launch success rates have been improving since 2013
- The type of orbit can impact your success rate
- All the predictive models tested performed with the same accuracy which was above 80%

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each. Much of this cost savings is because SpaceX can reuse the first stage. If we can predict if the first stage will land and be reused, we can better determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX.

To be able predict the success of the first stage we need to answer the following questions:

- What factors impact the success of the first stage launch?
- Given those factors, what predictive model can determine with the most accuracy whether the launch will be a success or failure?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data was collected via SpaceX API and response was converted from JSON objects to a dataframe via json_normalize()
  - Falcon 9 launch data was collected via web scraping related Wiki pages with BeautifulSoup and transformed into a useful dataset
- Perform data wrangling
  - Converted landing outcome to classification variable (0 or 1) and added it as a column in the dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into testing and training data
  - Created an object for each predictive model being tested
  - Created a GridSearchCV object for each predictive model being tested
  - Fit the training data
  - Used the score() function to find the accuracy of the test data

Data Collection

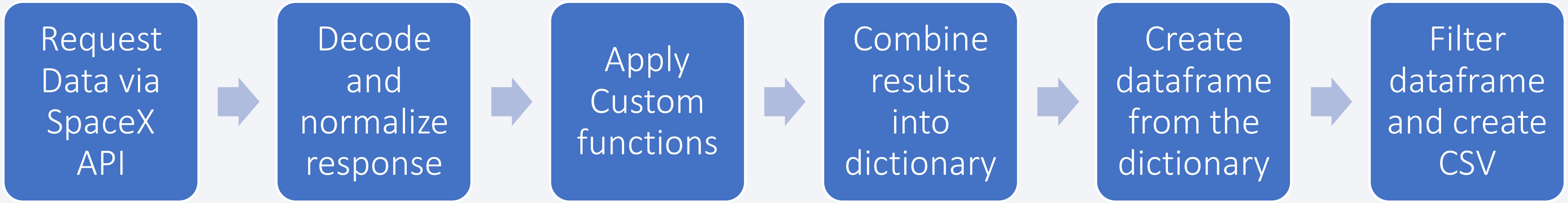
Data Collection via SpaceX API:

- Request Data from SpaceX API using requests.get()
- Decode and normalize response using json_normalize()
- Apply Custom functions to get more meaningful information related to the booster version, payload, launch sites, and the landing
    - getBoosterVersion()
    - getLaunchSite()
    - getPayloadData()
    - getCoreData()
- Combine results into dictionary
- Create dataframe from the dictionary
- Filter dataframe and create CSV

Data Collection via Web Scraping:

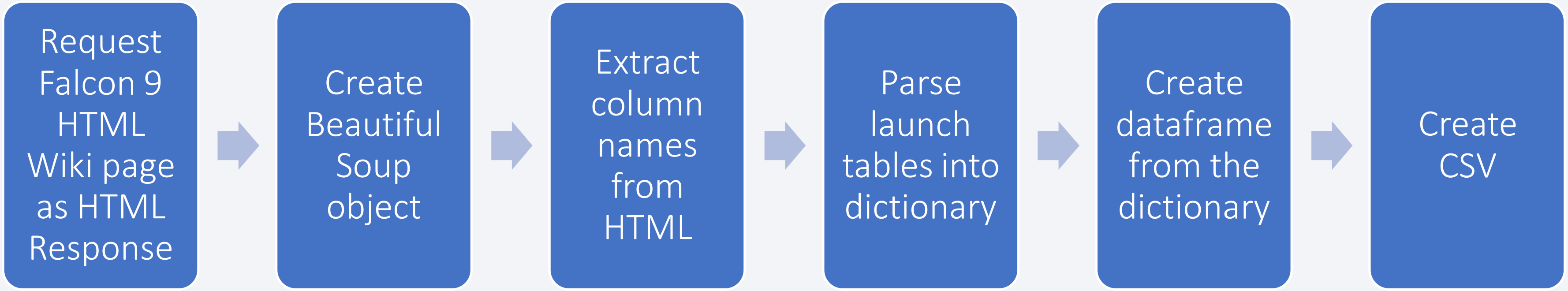
- Request Falcon 9 HTML Wiki page as HTML Response
- Create Beautiful Soup object
- Extract column names from HTML
- Parse launch tables into dictionary
- Create dataframe from the dictionary
- Create CSV

Data Collection – SpaceX API


Request Data via SpaceX API
Decode and normalize response
Apply Custom functions
Combine results into dictionary
Create dataframe from the dictionary
Filter dataframe and create CSV

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Request Falcon 9 HTML Wiki page as HTML Response

Create Beautiful Soup object

Extract column names from HTML

Parse launch tables into dictionary

Create dataframe from the dictionary

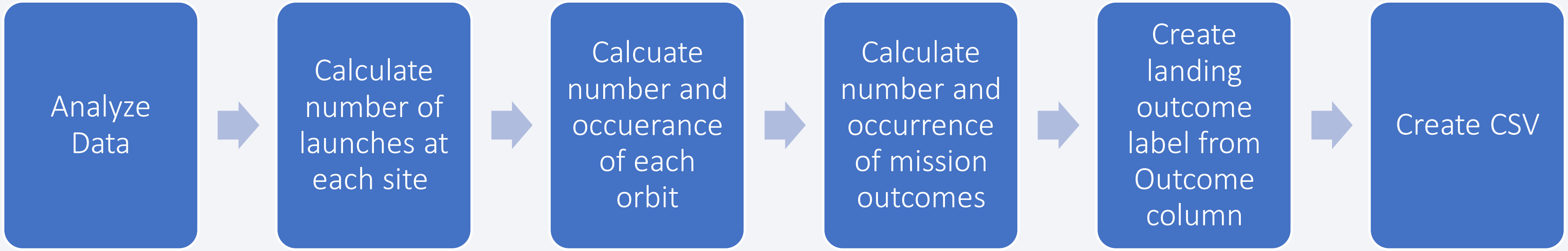
Create CSV

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

First, we explored the data set including:
Number of launches at each site
Number and occurrence of each type of orbit
Number and occurrence of mission outcomes

Ultimately we decided to convert those outcomes into Training Labels where 1 means the booster successfully landed 0 means it was unsuccessful.



Analyze Data

Calculate number of launches at each site

Calcuate number and occuerance of each orbit

Calculate number and occurrence of mission outcomes

Create landing outcome label from Outcome column

Create CSV

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

EDA with Data Visualization

- We created the following scatter plots with an overlay of the launch outcome to discover how or if these feature relate to each other and/or the launch outcome.
    - Flight Number vs. Payload Mass
    - Flight Number vs. Launch Site
    - Launch Site vs. Payload Mass
    - Flight Number vs. Orbit Type
- We created a bar chart to see which orbits have the highest success rate
- We plotted a line chart to see the launch success trend over time

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

EDA with SQL

We executed the following SQL queries to further understand the SpaceX data set:

- List the unique launch sites in the space mission
- Show 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 of the first successful landing outcome in ground pad
- List the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the boosters that have carried the maximum payload mass
- List the months, failure landing_outcomes in drone ship, booster versions, launch_site for the year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Build an Interactive Map with Folium

We built an interactive map with Folium and added the following to it:

- Circles and markers to show the location of each launch site on a world map and determine if they have anything in common
- Marker cluster showing color coded markers for all launch records to get a visual representation of the success/failure outcomes when zoomed in on the various launch sites
- MousePosition to get coordinate for a mouse over a point on the map
- Distance Marker at the nearest coastline to one of the launch sites
- Polyline from that point to the launch site to illustrate the site's proximity to the coastline
- Distance Marker at the nearest city to another one of the launch sites
- Polyline from that point to the launch site to illustrate the site's proximity to the coastline

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Build a Dashboard with Plotly Dash

We built a dashboard with Plotly Dash and added the following to it:

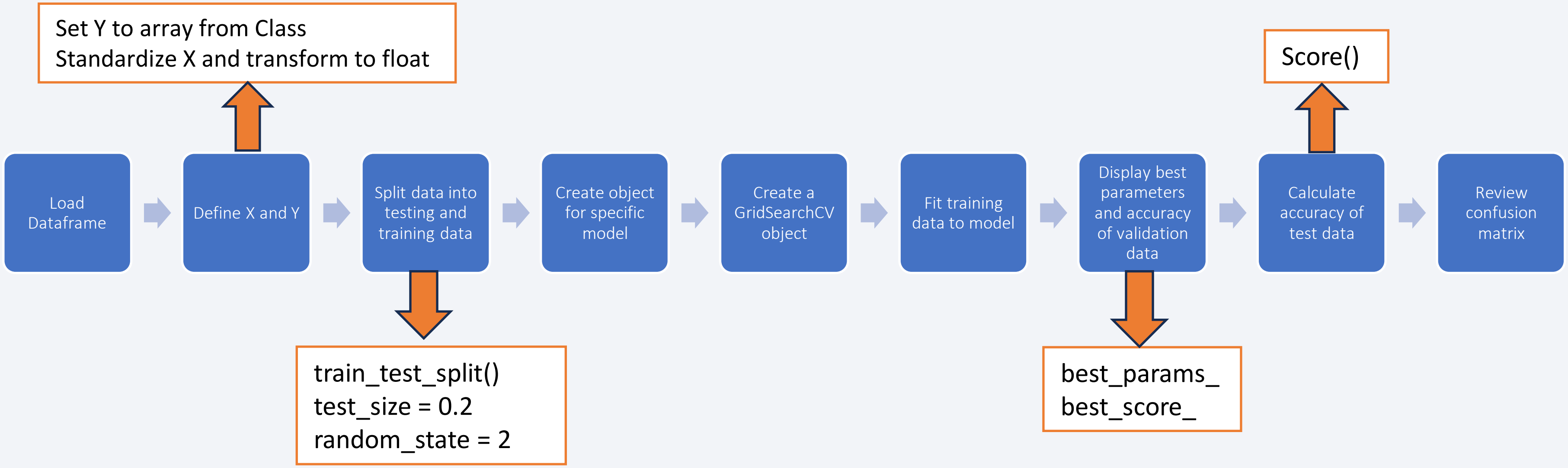
- A dropdown to allow the user to select between all launch sites or show data for an individual launch site
- A pie chart to show the success rate for each launch site or all launch sites combined depending on what the user has selected in the drop down
- A range slider to allow the user to select the entire payload mass range or just certain portions of the range
- A scatter plot showing the payload mass relationship to the launch outcome (class) for each booster version category

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Predictive Analysis (Classification)

Built the following predictive models and reviewed the performance of each via the process depicted below:

Logistic Regression
Support Vector Machine
Decision Tree
K Nearest Neighbors



Load Dataframe

Define X and Y

Set Y to array from Class
Standardize X and transform to float

Split data into testing and training data

train_test_split()
test_size = 0.2
random_state = 2

Create object for specific model

Create a GridSearchCV object

Fit training data to model

Display best parameters and accuracy of validation data

best_params_
best_score_

Calculate accuracy of test data

Score()

Review confusion matrix

Detailed information can be found here:
https://github.com/jenremgit/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Results

Exploratory data analysis results

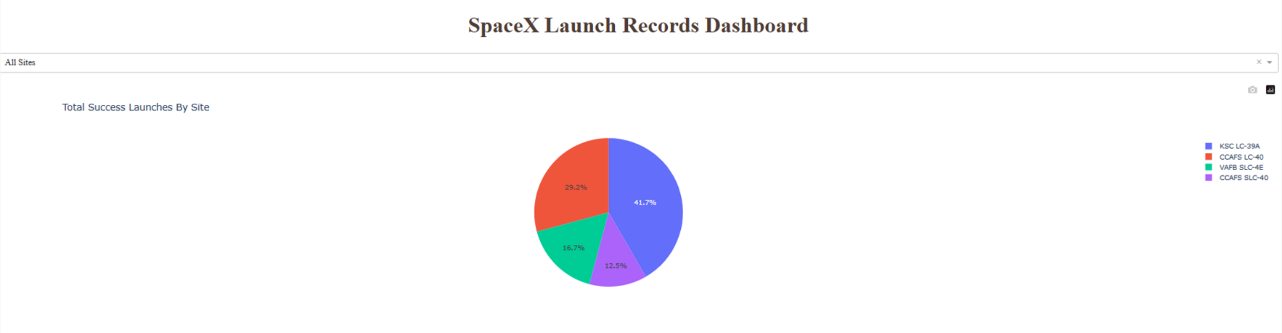
- KSC LC-39A has the highest success rate for launches
- Launch success rates have been improving since 2013
- The type of orbit can impact your success rate

Interactive analytics

- Launch sites in this country are as near the equator as possible
- Launch sites are close to the coast while keeping their distance from railways, highways and cities
- Launches are more successful with a lower payload mass

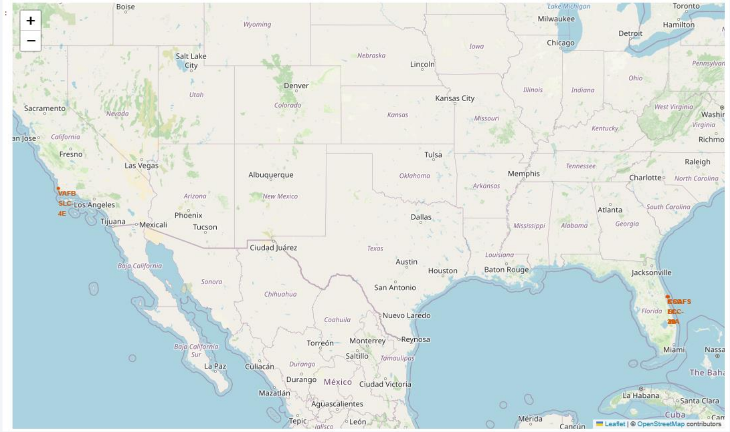
Predictive analysis results

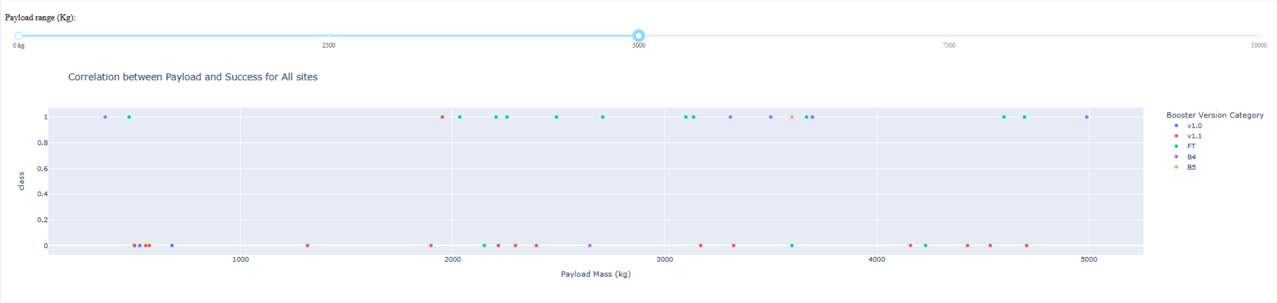
- The following predictive models all perform with an accuracy just above 80% (.8333333333333334)
  - Logistic regression
  - Support Vector Machine
  - Tree Classification
  - K Nearest Neighbors



SpaceX Launch Records Dashboard

All Sites





Payload range (Kg):

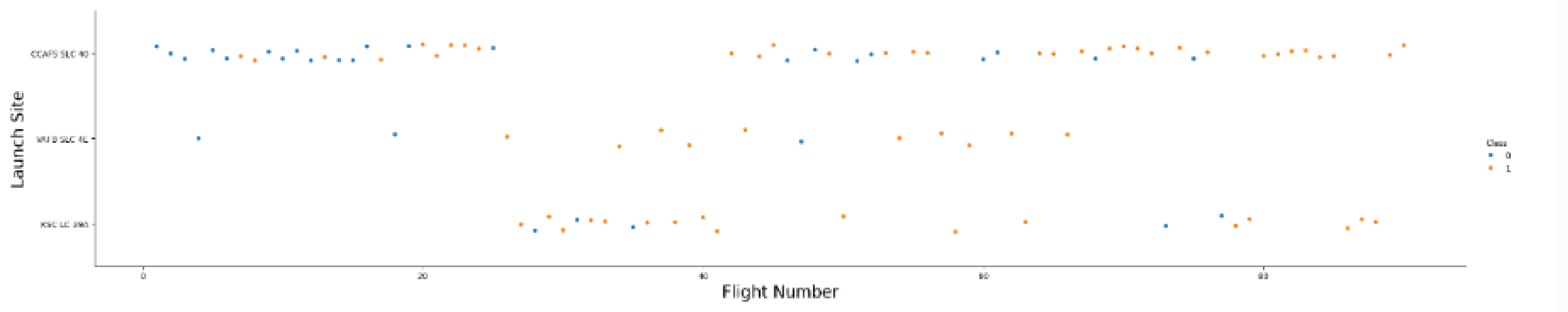
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

This scatter plot of Flight Number vs. Launch Site shows us:

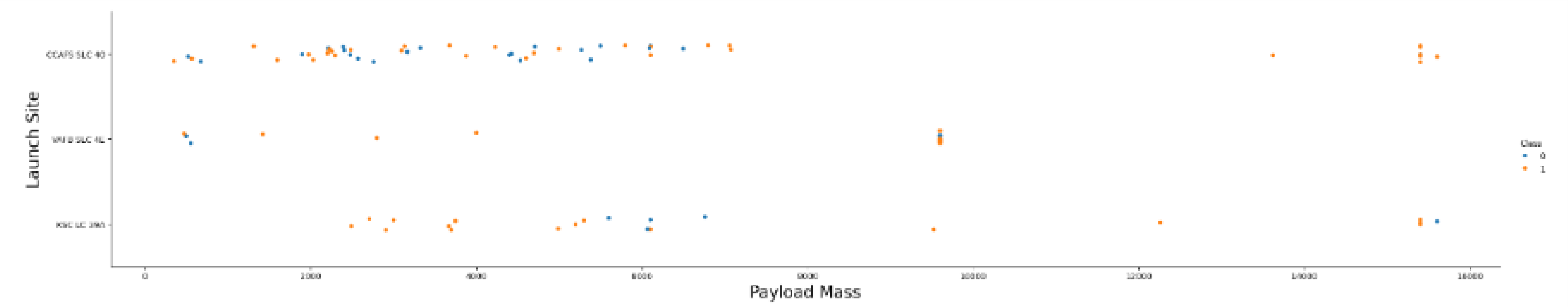
- the CCAFS SLC-40 site has more launches than the other sites
- the VAFB SLC 4E and KSC LC 39A sites have a higher success rate
- as the flight number increases the success rate increases



Payload vs. Launch Site

This scatter plot of Payload Mass vs Launch Site shows us:

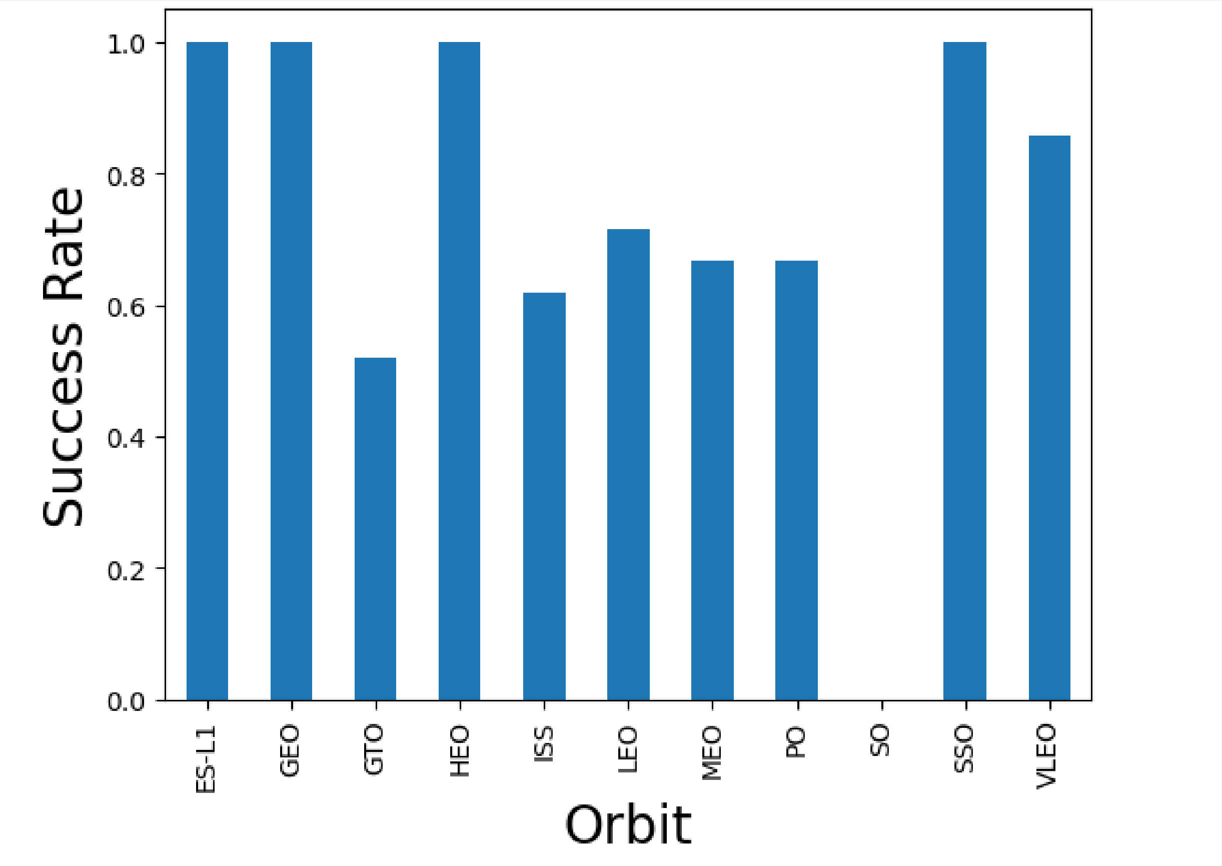
- the VAFB-SLC site has no launches at a payload heavier than 10000
- most launches have a payload mass < 8000
- the success rate of the launches is higher at higher payloads



Success Rate vs. Orbit Type

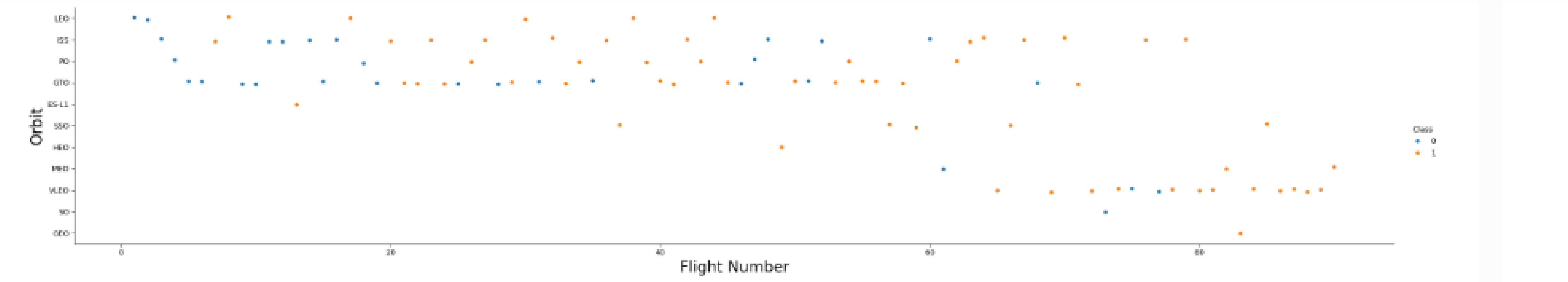
The orbits with the highest success rate are:

- ES-L1
- GEO
- HEO
- SSO



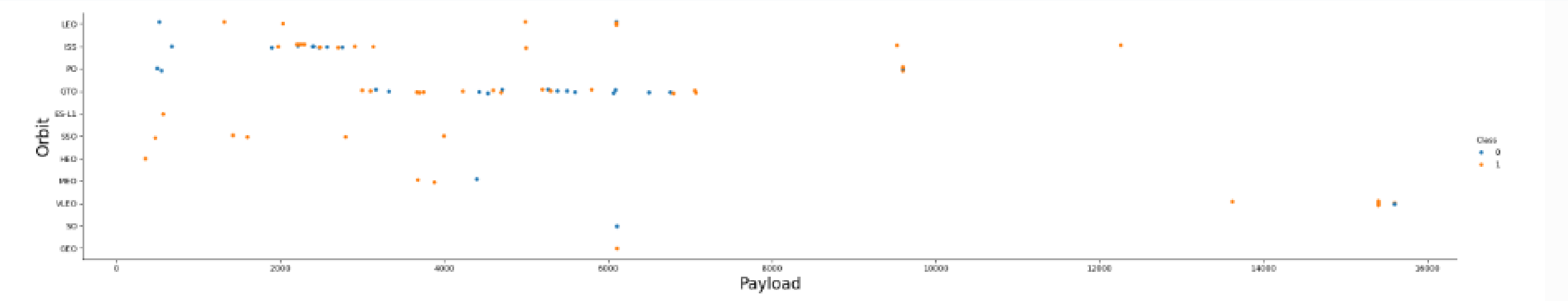
Flight Number vs. Orbit Type

We created a scatter chart to see if there is a relationship between Flight Number and Orbit Type. We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



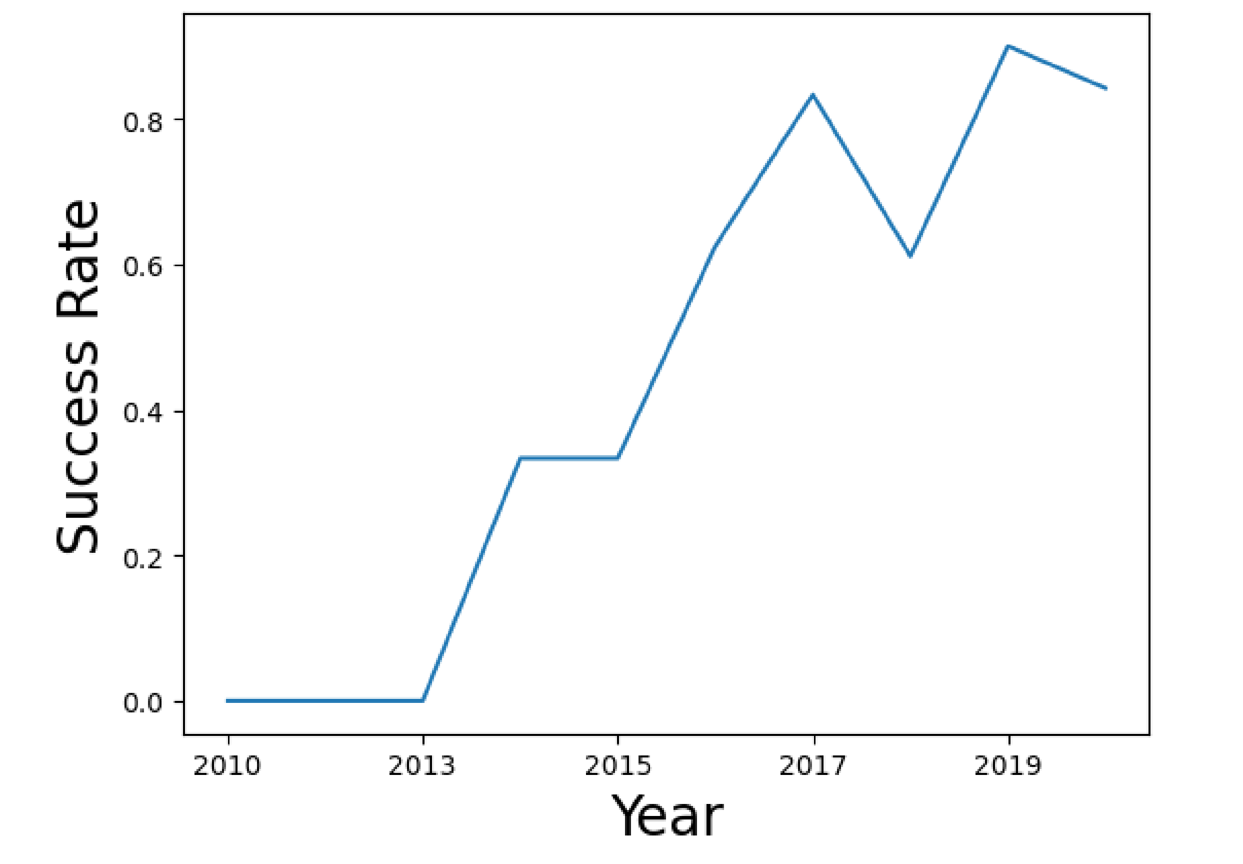
Payload vs. Orbit Type

We created a scatter chart to see the relationship between Payload and Orbit Type. We can see that Polar, LEO and ISS have higher success rates at higher payloads. However GTO has a mixture of success and failure even at the higher payloads. We can also see that only VLEO, ISS and Polar have very high payloads (over 8000).



Launch Success Yearly Trend

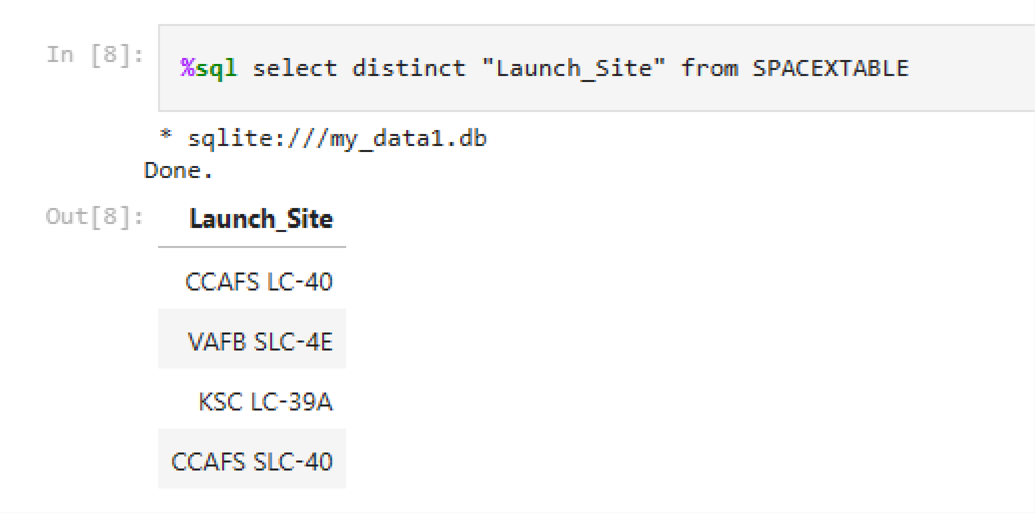
We plotted a line chart to see the launch success trend over time and can see that the success rate continually increases starting in 2013.



All Launch Site Names

We used a **select distinct** query to find the unique launch site names.

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



In [8]:

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

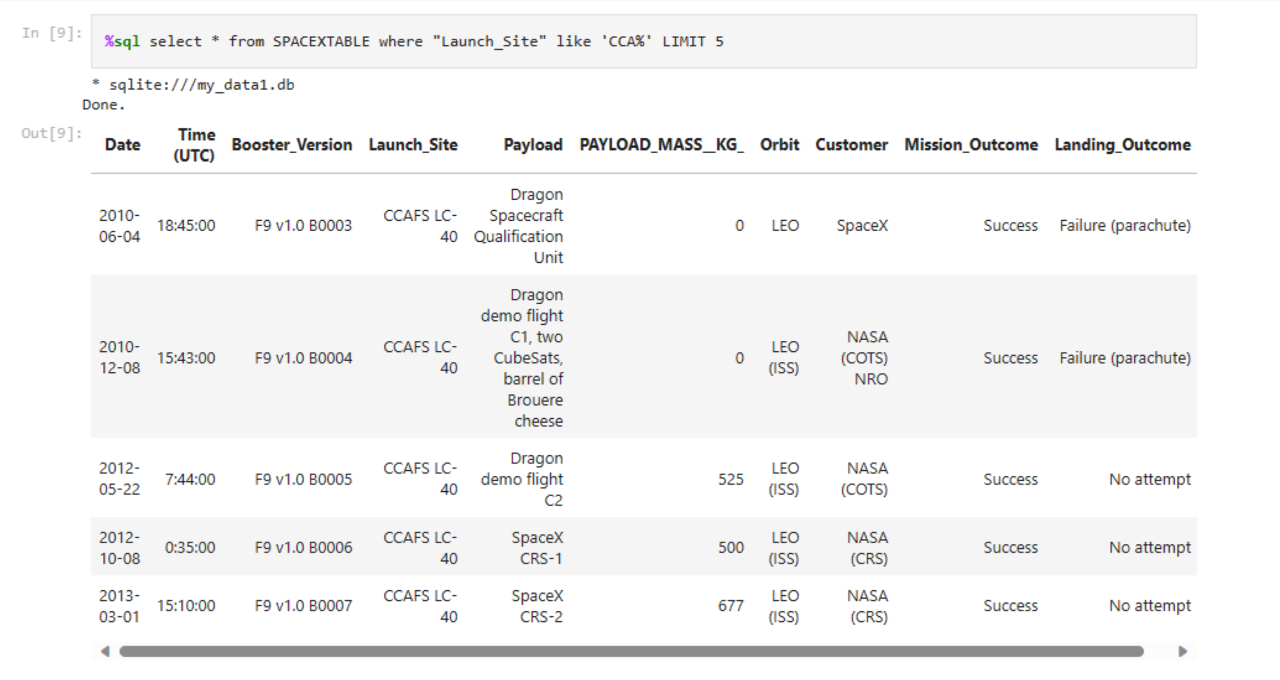
```
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

Launch Site Names Begin with 'CCA'

We used the **LIKE** and **LIMIT** qualifiers in the WHERE clause to list 5 records where the launch site name begins with "CCA".



```
In [9]: %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' LIMIT 5
```

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

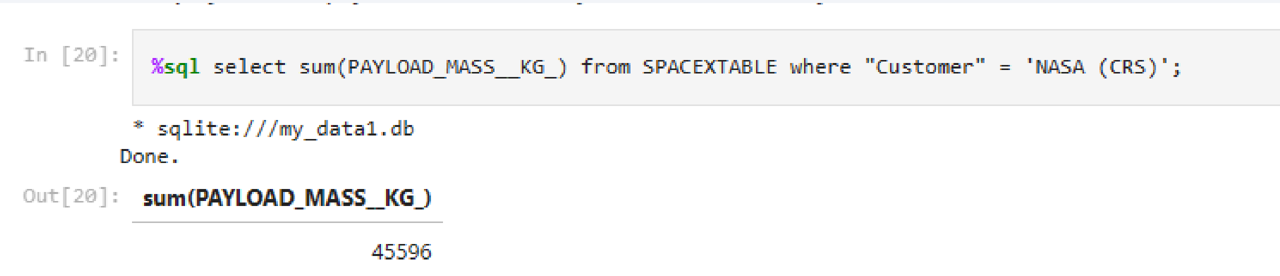
Out[9]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

We used the **sum()** function and **"="** operator to calculate the total payload mass of all boosters from NASA.

45596



```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = 'NASA (CRS)';
```

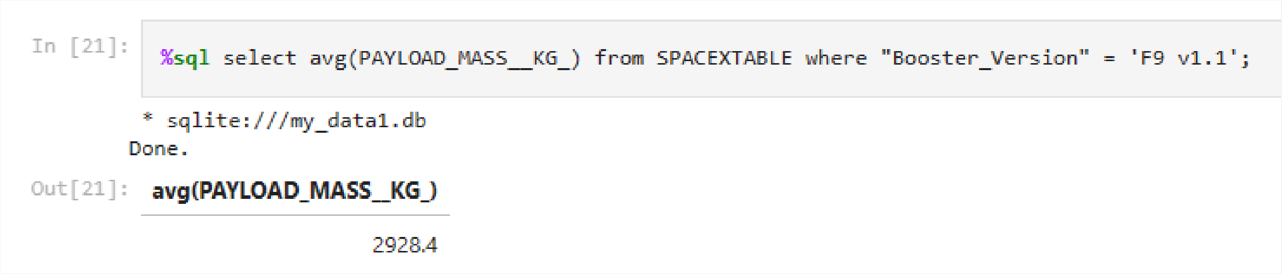
```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

We used the **avg()** function and **"="** operator to calculate the average payload mass carried by booster version F9 v1.1.

2928.4

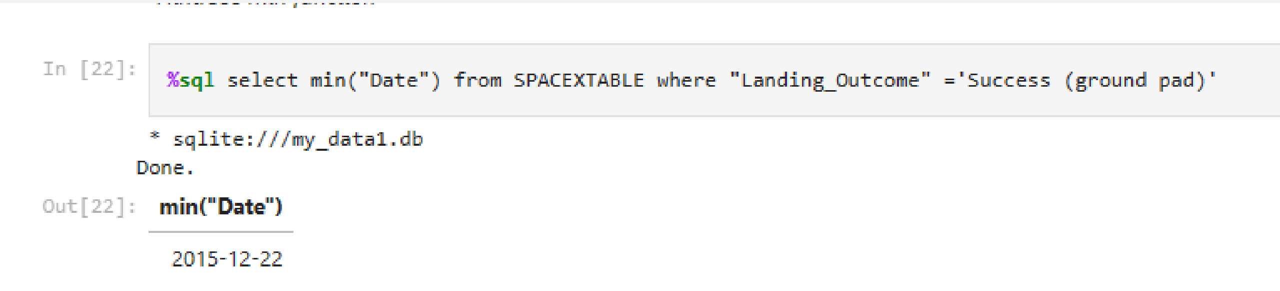


```
In [21]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = 'F9 v1.1';

          * sqlite:///my_data1.db
         Done.

Out[21]: avg(PAYLOAD_MASS__KG_)
                         2928.4
```

First Successful Ground Landing Date

We used the **min()** function and **"="** operator to find the date of the first successful landing on ground pad.

2015-12-22

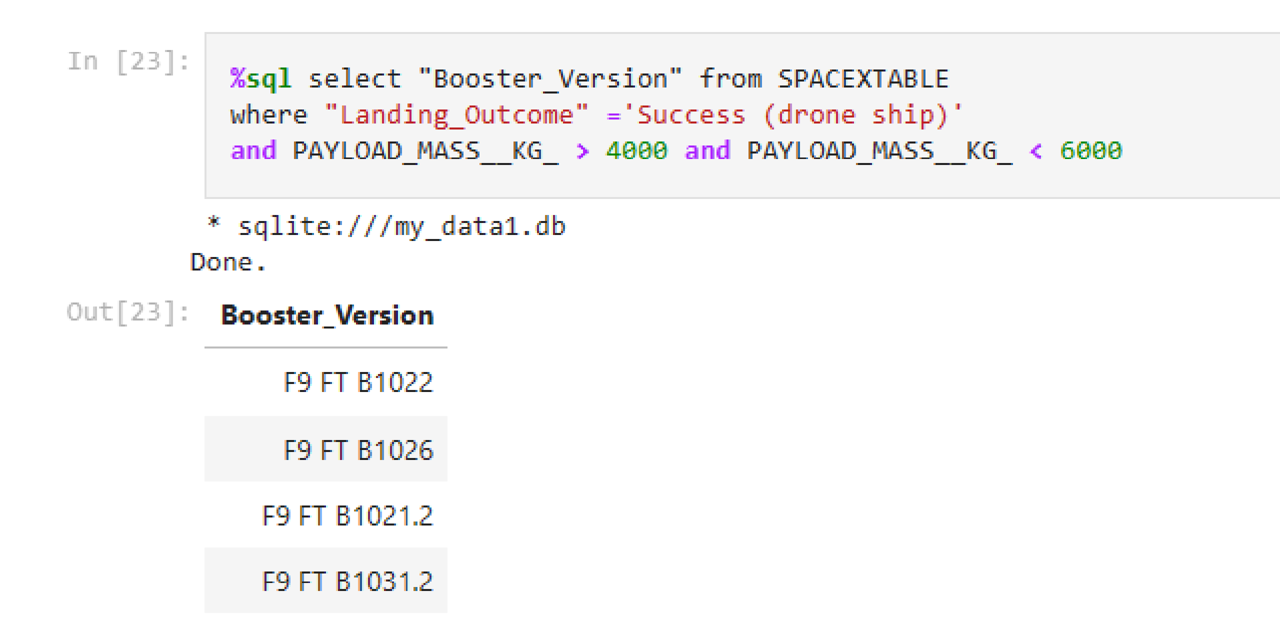


```
In [22]:  %sql select min("Date") from SPACEXTABLE where "Landing_Outcome" ='Success (ground pad)'

  * sqlite:///my_data1.db
 Done.
Out[22]:  min("Date")
          2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

We used > and < in the where clause to find the names of the boosters that successfully landed on drone ship with a payload mass between 4000 and 6000.

F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2



```
In [23]: %sql select "Booster_Version" from SPACEXTABLE
         where "Landing_Outcome" ='Success (drone ship)'
         and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

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

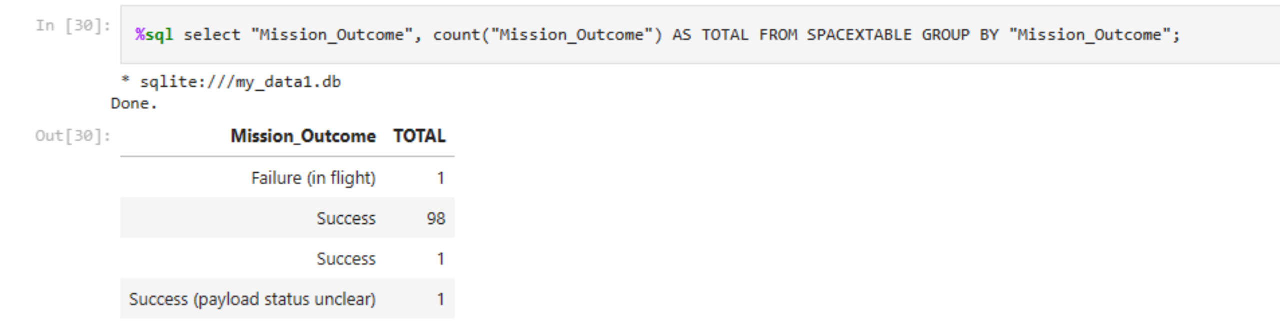
Out[23]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

We used the **count()** function and **GROUP BY** to calculate the total number of successful and failure mission outcomes. **Note:** There was a slight data issue in terms of a single successful record that has a type-o. Accounting for that, the actual totals are:

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



In [30]:
```
%sql select "Mission_Outcome", count("Mission_Outcome") AS TOTAL FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

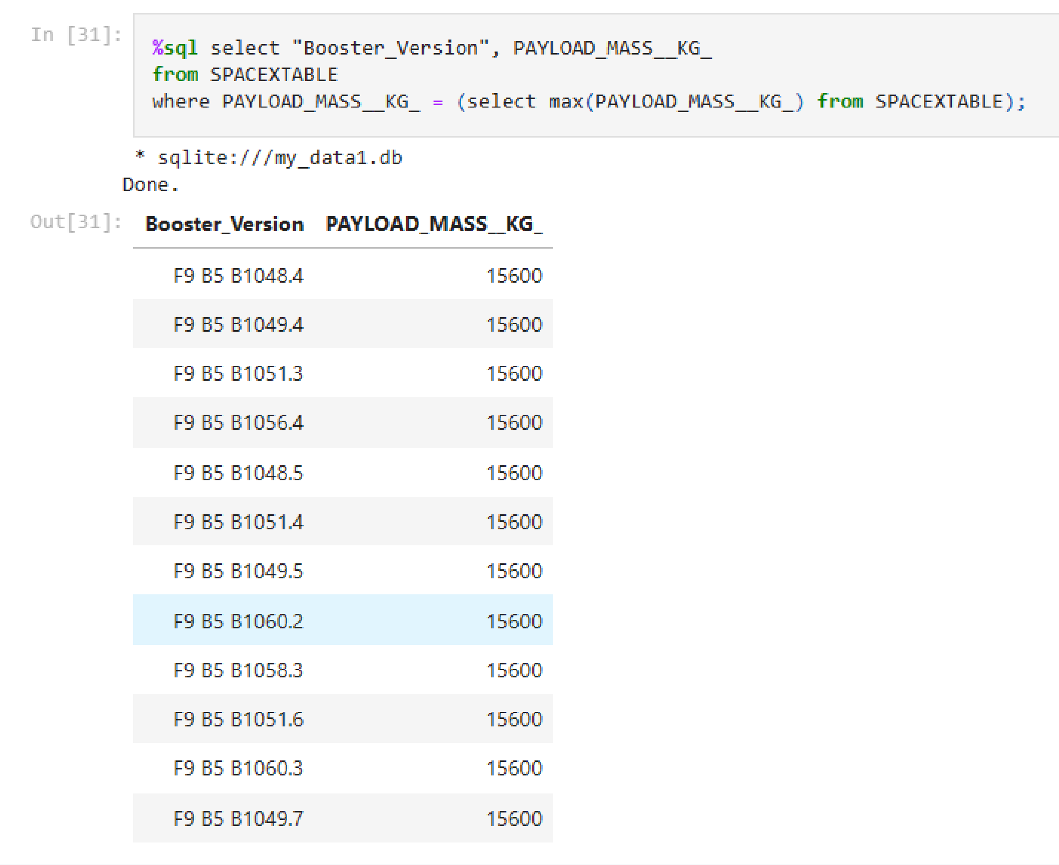
```
 * sqlite:///my_data1.db
Done.
```

Out[30]:

Mission_Outcome	TOTAL
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

We used a **subquery** to list the names of the boosters that have carried the maximum payload mass of 15600.



In [31]:

```
%sql select "Booster_Version", PAYLOAD_MASS__KG_
from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

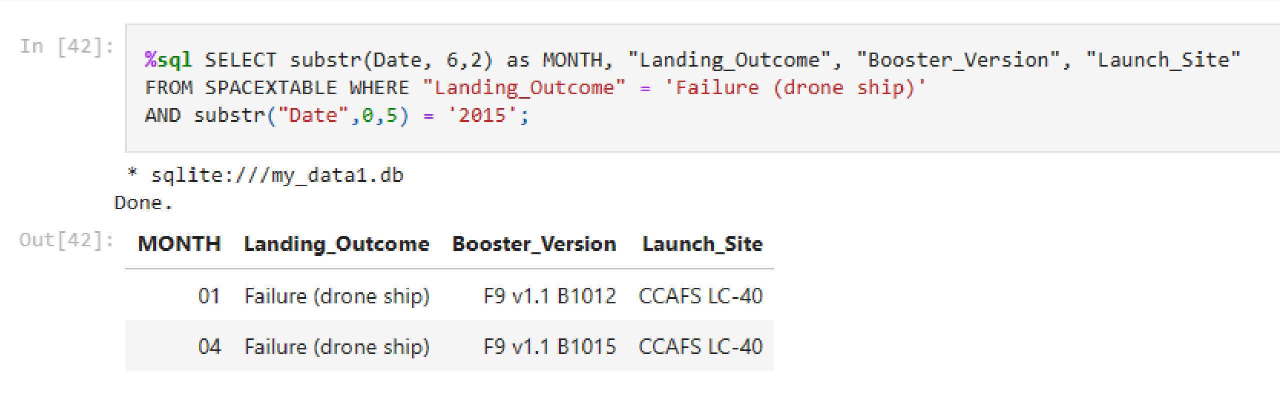
```
 * sqlite:///my_data1.db
Done.
```

Out[31]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

We used the **substr()** function to list the failed outcomes in drone ship for the months in 2015.



In [42]:

```
%sql SELECT substr(Date, 6,2) as MONTH, "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)'
AND substr("Date",0,5) = '2015';
```

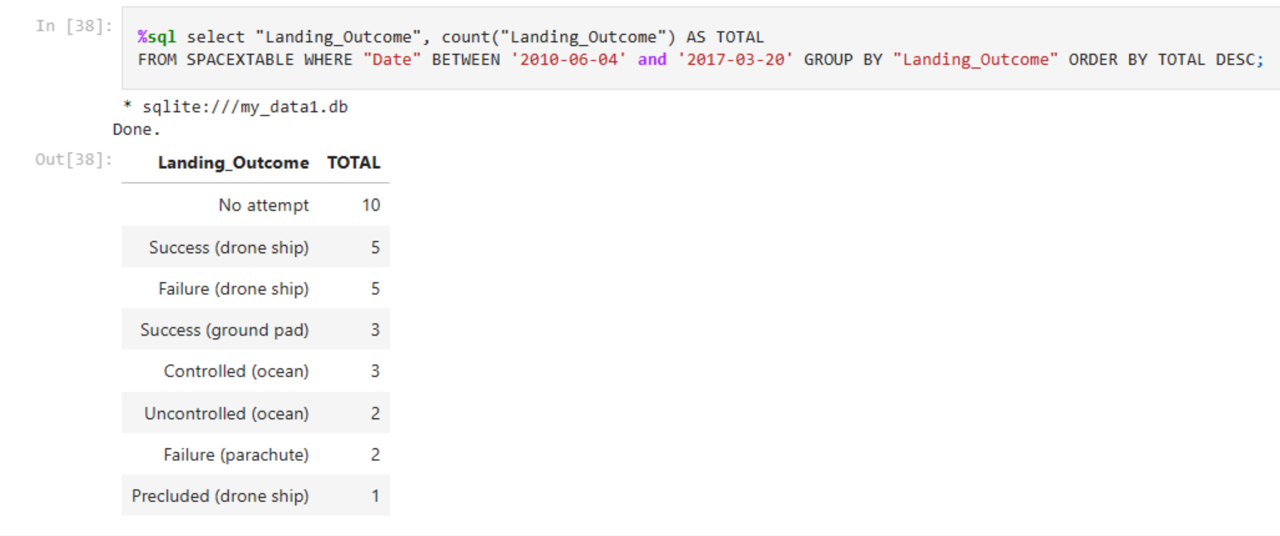
```
 * sqlite:///my_data1.db
Done.
```

Out[42]:

MONTH	Landing_Outcome	Booster_Version	Launch_Site
01	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

We used **count()**, **GROUP BY**, **BETWEEN** and **ORDER BY** to rank the types of landing outcomes in descending order.



```
In [38]: %sql select "Landing_Outcome", count("Landing_Outcome") AS TOTAL
         FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' and '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY TOTAL DESC;

 * sqlite:///my_data1.db
Done.
Out[38]:
```

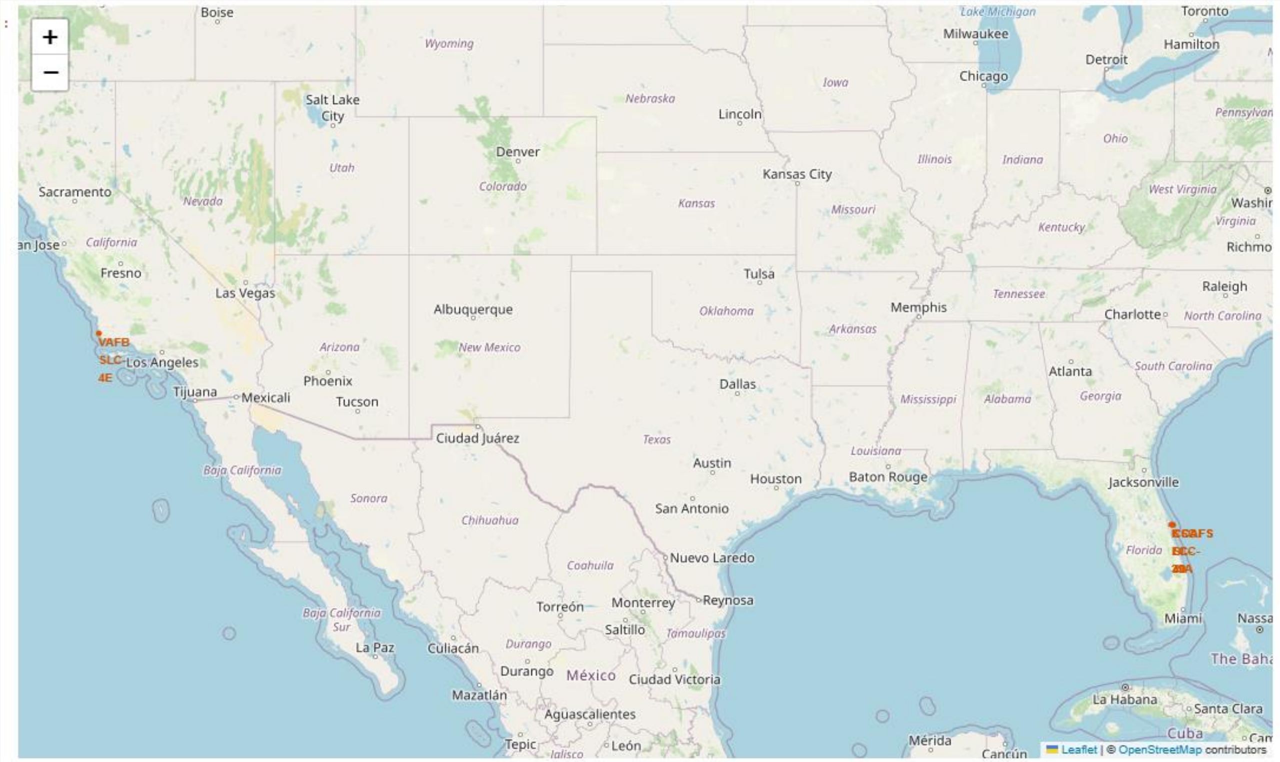
Landing_Outcome	TOTAL
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

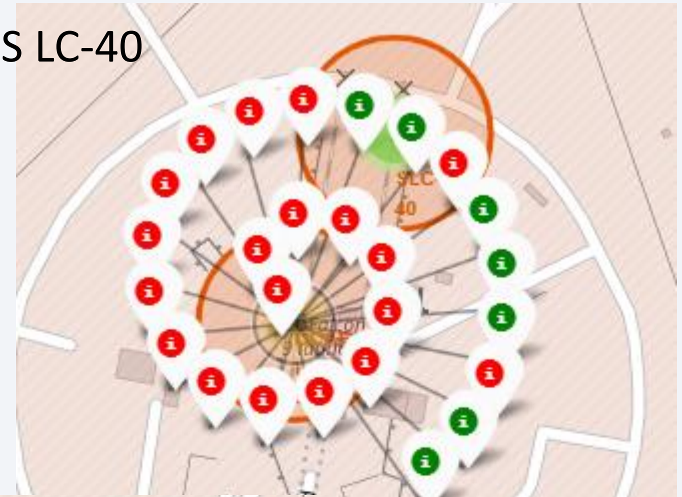
This Folium world map shows that SpaceX launch sites are located in the United States of America in locations within that country that are closest to the equator and close to coastlines.



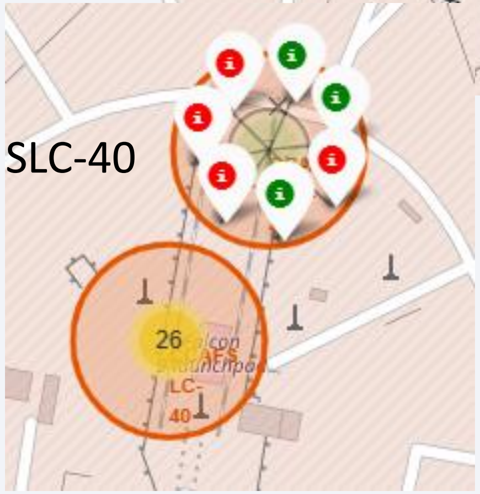
Launch Outcomes at Sites

With the green markers representing successful launches and the red markers representing failed launches, it's easy to see on these drill-down's of the Folium map that the CCAFS LC-40 site had the most launches overall and that the KSC LC-39A had the highest success rate.

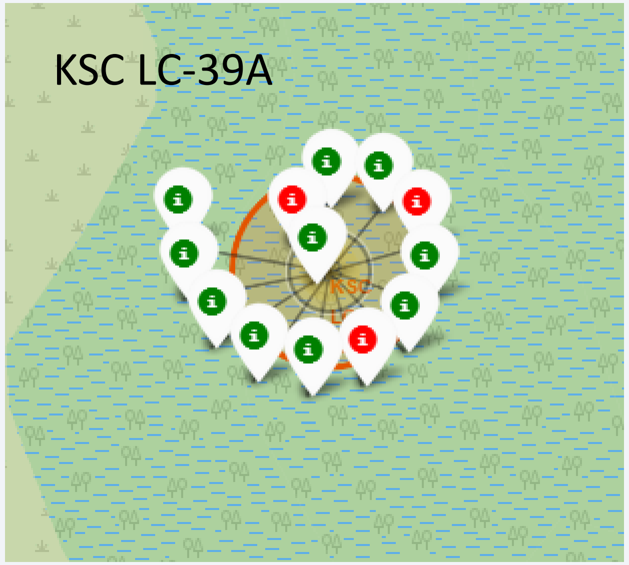
CCAFS LC-40



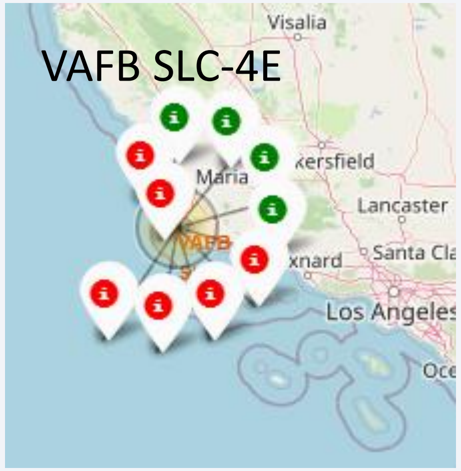
CCAFS SLC-40



KSC LC-39A

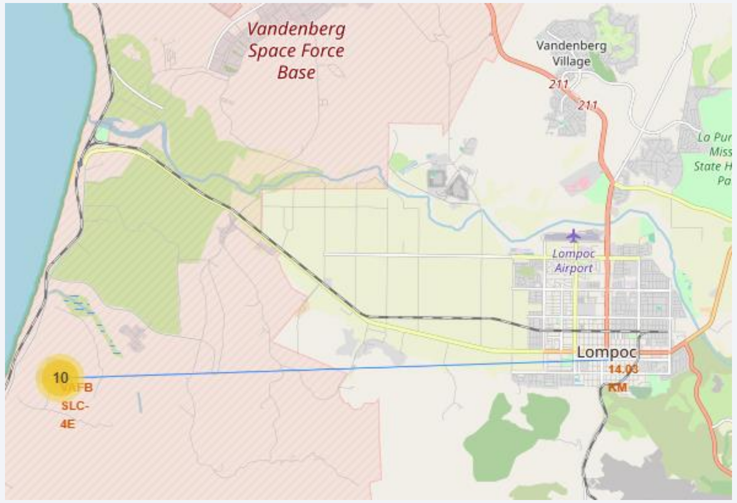


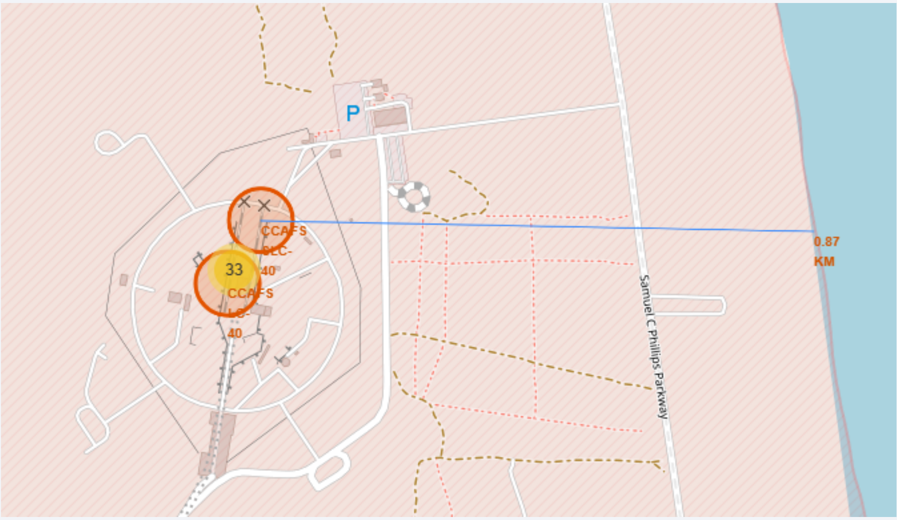
VAFB SLC-4E



Launch Site Proximities

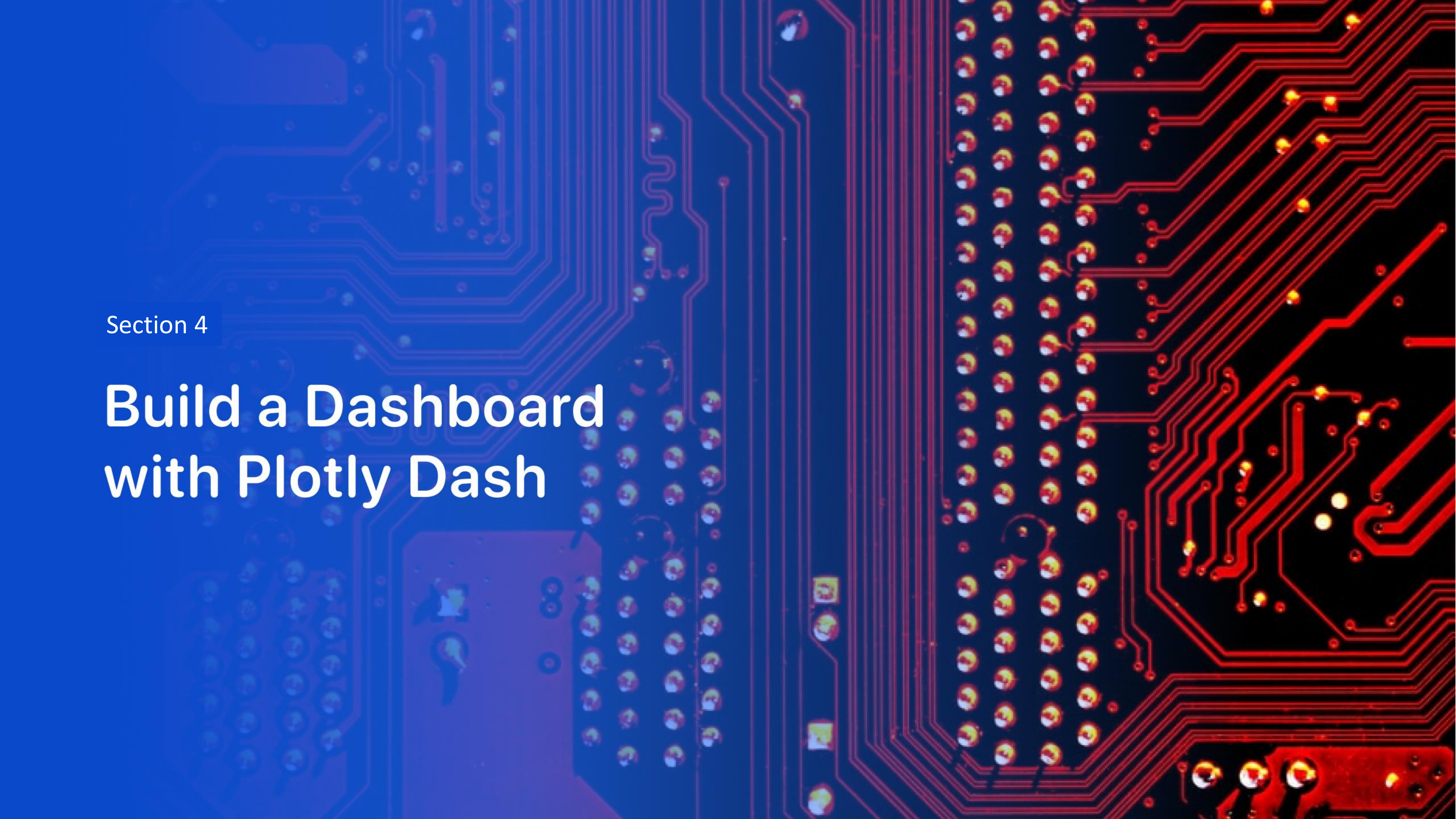
When distances to various proximities we can see that launch sites are located very near coastlines and yet keep a certain distance from cities.


Vandenberg Space Force Base
Vandenberg Village
211
211
La Purisima Mission State Historic Park
Lompoc Airport
Lompoc
14.03 KM
10
VAFB SLC-4E



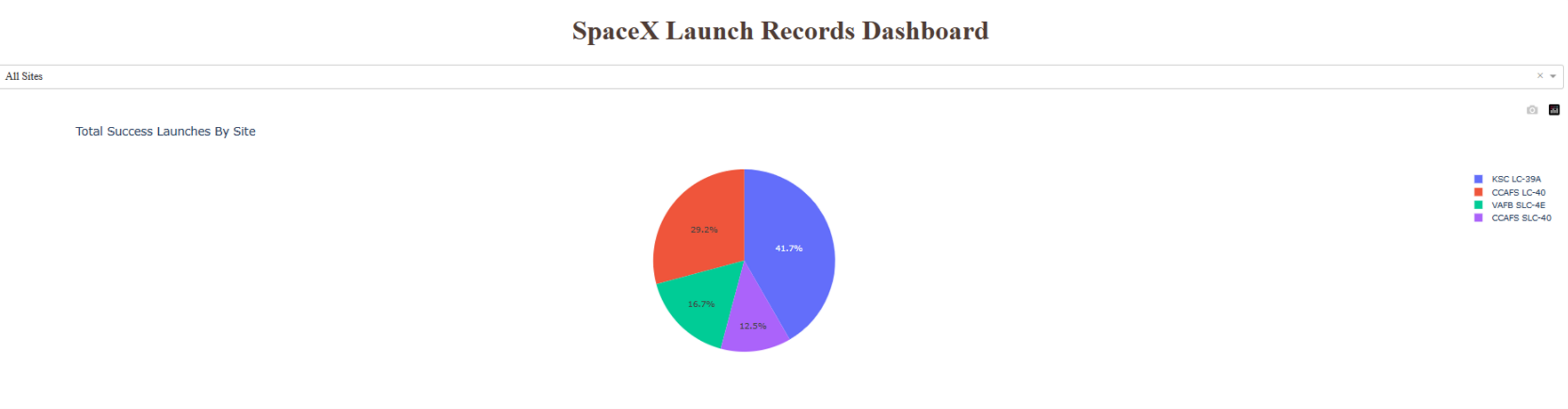
Section 4

Build a Dashboard with Plotly Dash



SpaceX Successful Launches by Launch Site

The pie chart on this dashboard shows that of all the launch sites, KLC LC-39A had the most successful launches.

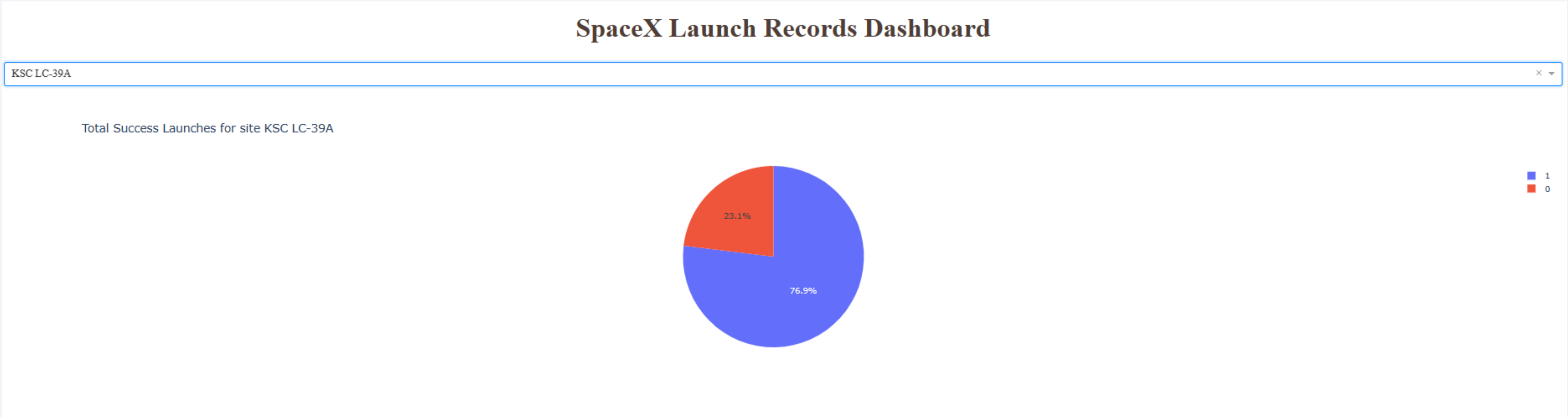


SpaceX Launch Records Dashboard

All Sites

Launch Site with Highest Success Ratio

The pie chart on this dashboard shows the site KSC LC-39A has a success ration of 76.9% which is the highest of all the launch sites.



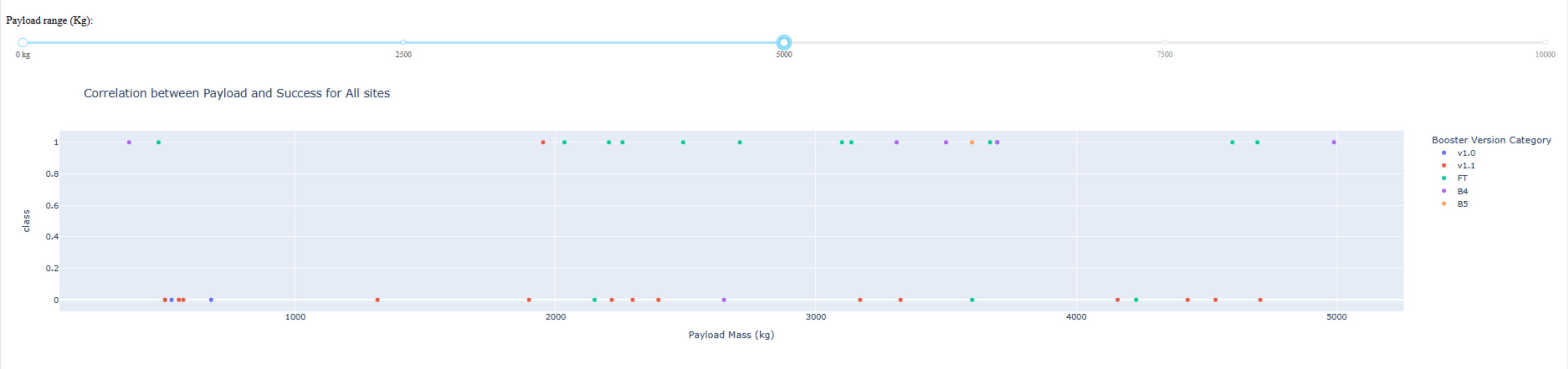
SpaceX Launch Records Dashboard

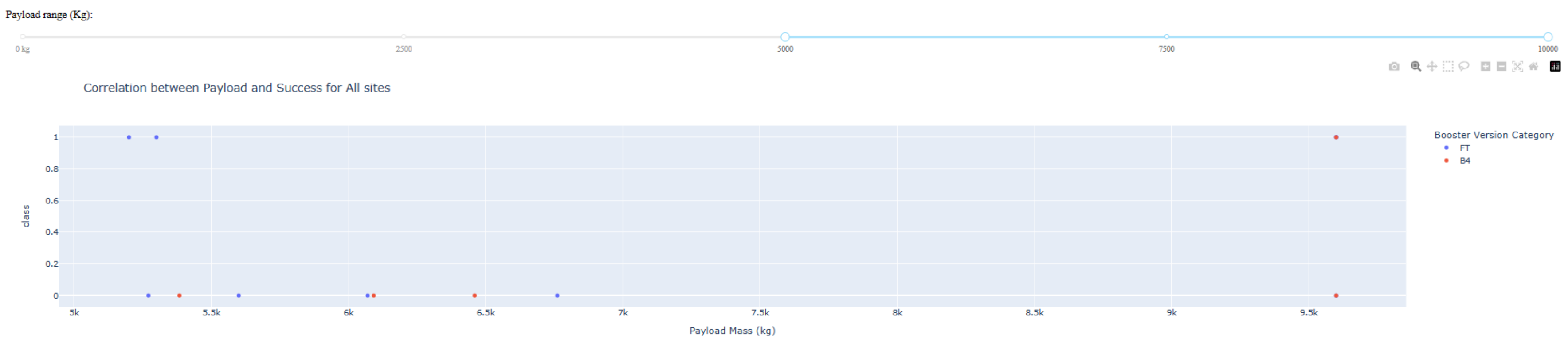
KSC LC-39A

Total Success Launches for site KSC LC-39A

Launch Outcome based on Payload Range

By comparing the low payload range (0-5000) to the high payload range (5000-10000) we can see that the most launches occur with low payload mass and that there are more successful launches in that lower payload mass range.





Section 5

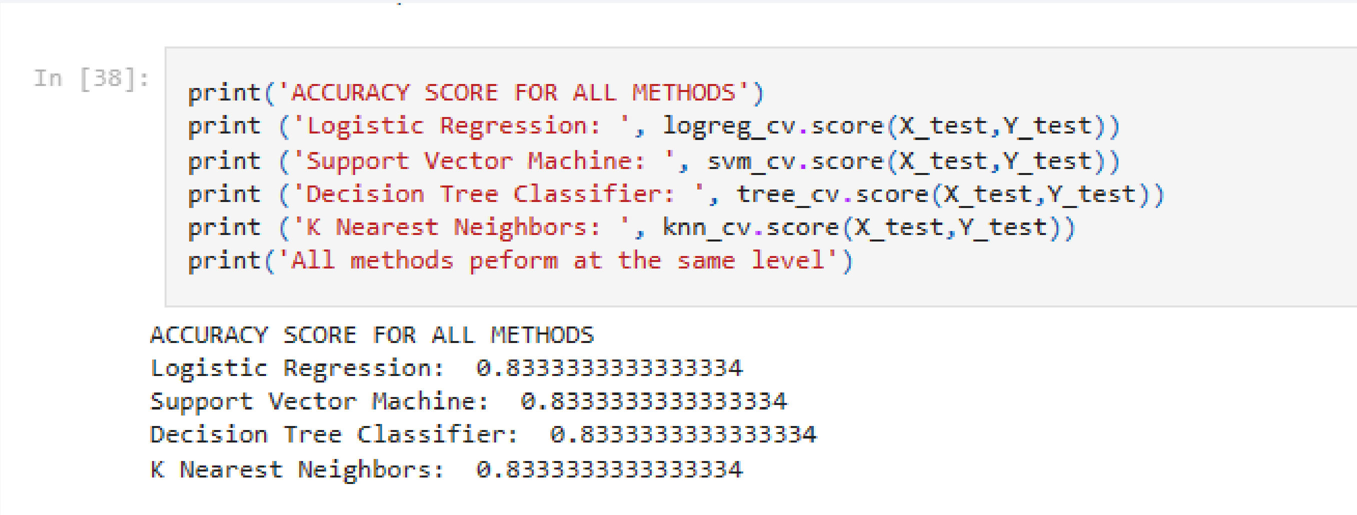
Predictive Analysis (Classification)

Classification Accuracy

After training and testing four different prediction models, each model performed with the same accuracy:

.8333333333333334

Any of these four models could be used to predict the success of launch.



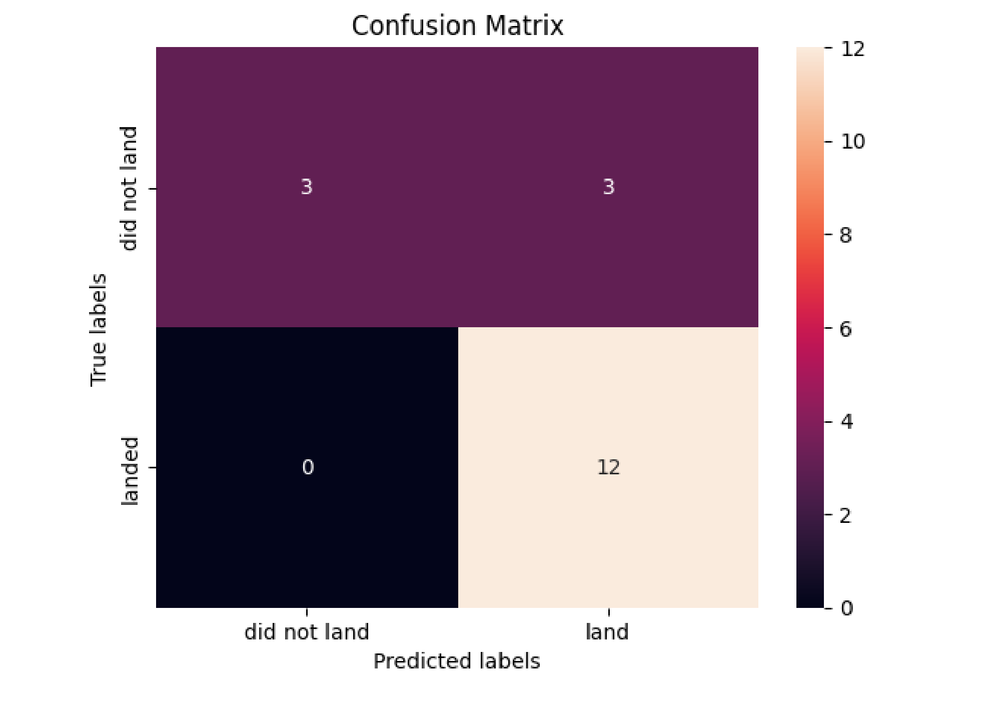
In [38]:

```
print('ACCURACY SCORE FOR ALL METHODS')
print ('Logistic Regression: ', logreg_cv.score(X_test,Y_test))
print ('Support Vector Machine: ', svm_cv.score(X_test,Y_test))
print ('Decision Tree Classifier: ', tree_cv.score(X_test,Y_test))
print ('K Nearest Neighbors: ', knn_cv.score(X_test,Y_test))
print('All methods peform at the same level')
```

```
ACCURACY SCORE FOR ALL METHODS
Logistic Regression:  0.8333333333333334
Support Vector Machine:  0.8333333333333334
Decision Tree Classifier:  0.8333333333333334
K Nearest Neighbors:  0.8333333333333334
```

Confusion Matrix

The Confusion Matrix for all four models that were tested looked the same. The main issue showing on the matrix is that 3 launches that were predicted to land did not land.



Conclusions

- Launch success rates are improving over time.
- Launch site KSC LC-39A has the highest success rate for launches.
- The type of orbit can impact the success of your launch with ES-L1 , GEO , HEO , and SSO having the best success rate.
- The following predictive models can predict with an accuracy just above 80% (.8333333333333334) whether a launch will be a success.
  - Logistic regression
  - Support Vector Machine
  - Tree Classification
  - K Nearest Neighbors

Thank you!