IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ryan Hill
11 August 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

IBM Developer

SKILLS NETWORK

Executive Summary

- Predict the success of a launch and the success of the recovery of the first stage in order to determine pricing strategies for future clients
- Determine the key factors in determining where to locate a launch site

Summary of methodologies

- Data collection & wrangling using APIs, BeautifulSoup, and One Hot Encoding
- EDA with Visualizations and SQL
- Interactive analysis with Folium Maps and Dash
- Predictive analysis using Decision Tree Classifier, SVM, KNN, & Logistic Regression
- Summary of all results

✓ Decision Tree Classification model performed the best on the test group

✓ Interactive analytics using Folium Maps concluded that proximity to the coast, interstate hwys, railroads, and urban areas are key factors for site location

✓ Visualizations provided key insights when performing EDA

Introduction

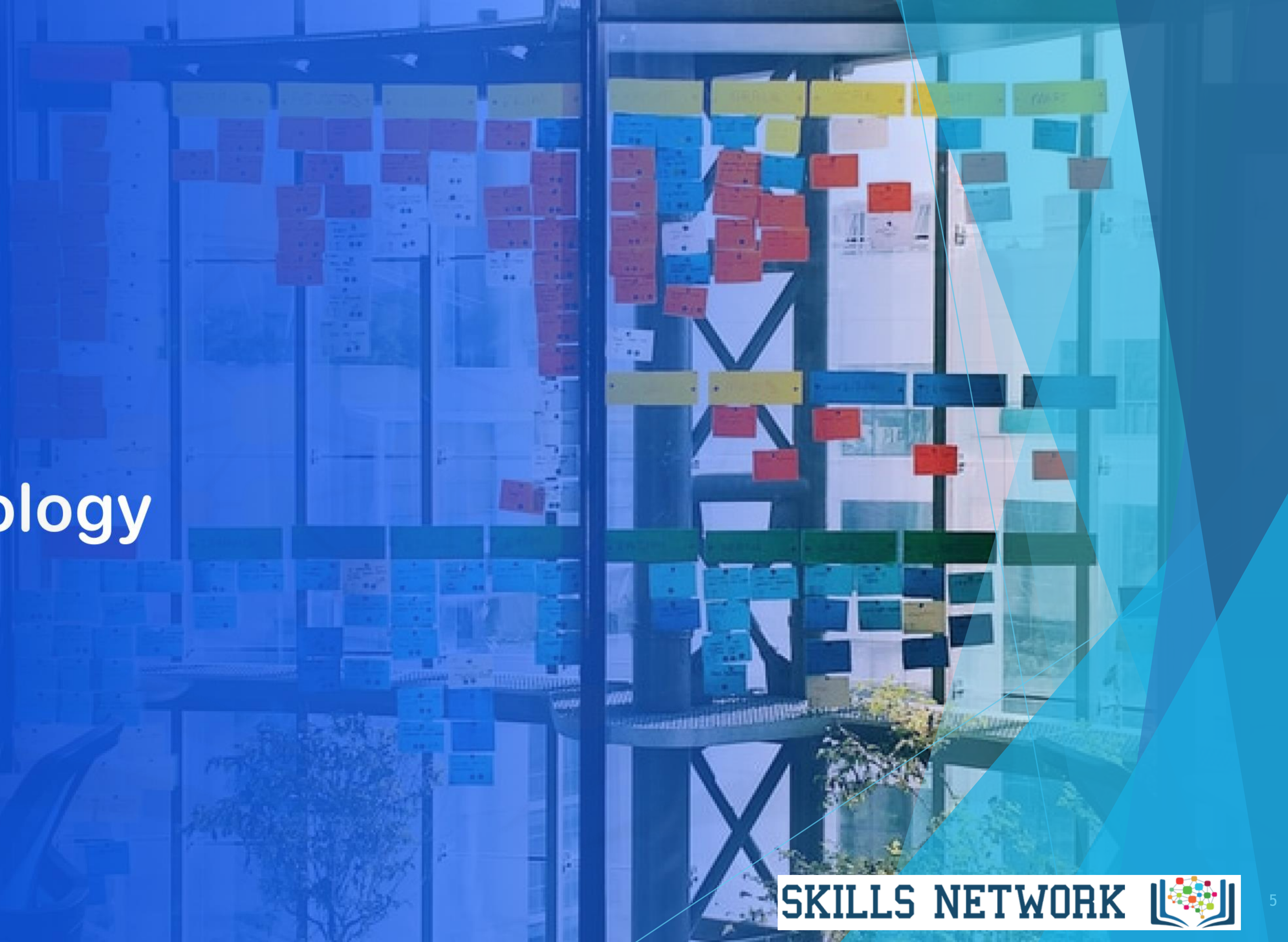
The commercial Space Industry has become one of the fastest growing industries in the 21st Century. As technology has rapidly evolved in both the computing power of computers and rocketry, the cost of getting a payload into Space has decreased precipitously, leading to a commercial rush to Space tourism and other private sector possibilities. This project evaluates how a new competitor into the market can analyze currently available data in order to answer fundamental questions on how to survive in the industry.

Key Questions to Answer:

- Where should a new competitor locate their launch facilities?
- How much should a new competitor charge per launch/payload?
- Predict the probability of recovering the first stage to improve cost estimates
- What factors are most important in determining a successful launch?

Section 1

Methodology



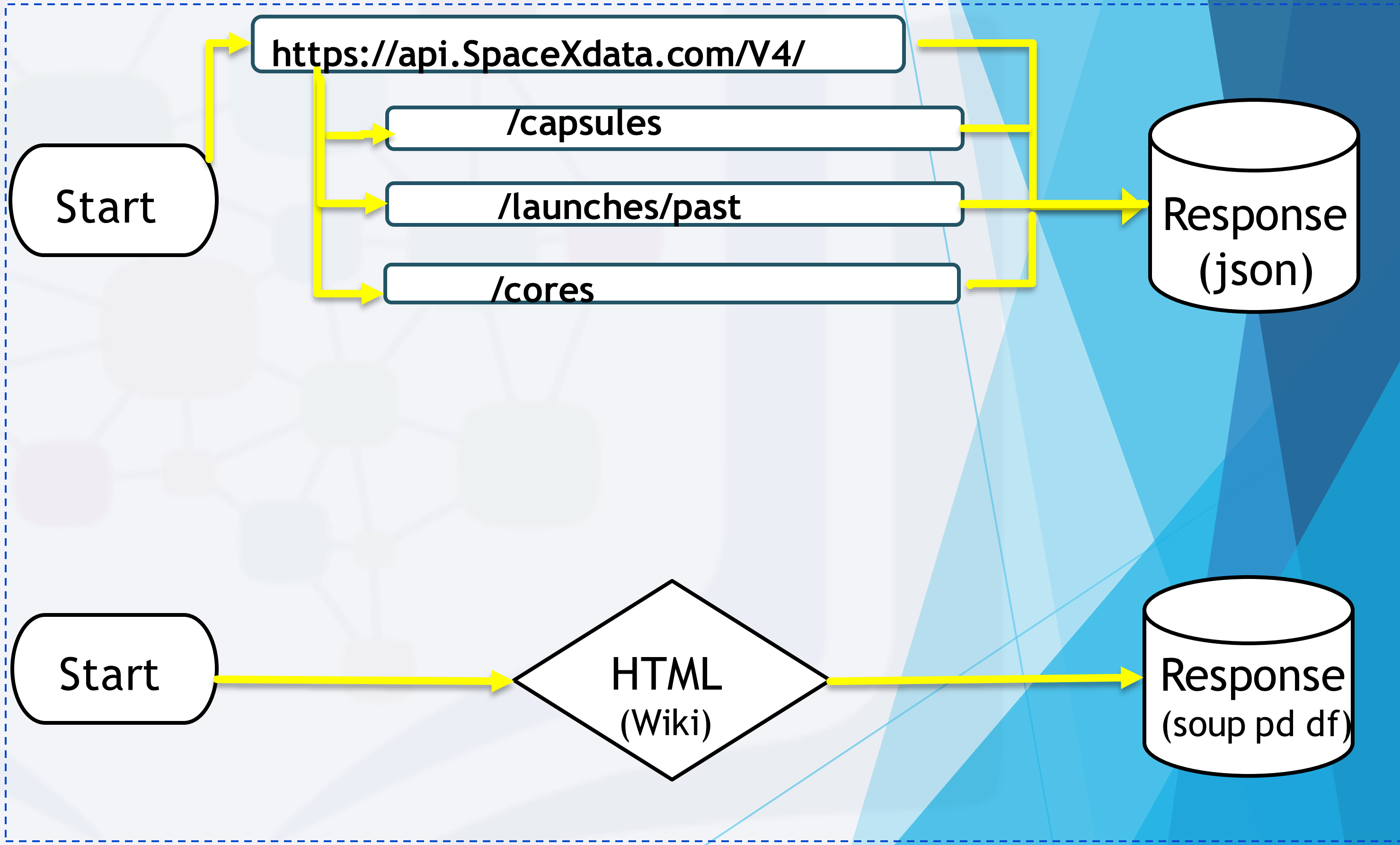
Executive Summary

- **Data collection methodology**: Data was collected using a SpaceX API to send a get request to a defined URL or by using the HTML Web Scraping API BeautifulSoup
- **Perform data wrangling**: Data was first Normalized, then filtered for desired information, sampled to explore what further wrangling needed to be performed including the removal of null values, and finally put into a searchable data frame. One Hot Encoding was used to perform these tasks.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**

Determine the model that performs the best

Data Collection – SpaceX API & BeautifulSoup

- Collect Data using SpaceX REST API using the "Get" request from 4 endpoints on a SpaceX URL and normalize into a data frame.
- Collect Data using HTML websites and BeautifulSoup and parse data into a panda data frame.
- External Reference:
- https://github.com/RMHUNC/Space-Y/blob/682bdd1631419b129111edc9ecd4034bc7128a13/spacex-data-collection-api.ipynb



IBM Developer

SKILLS NETWORK

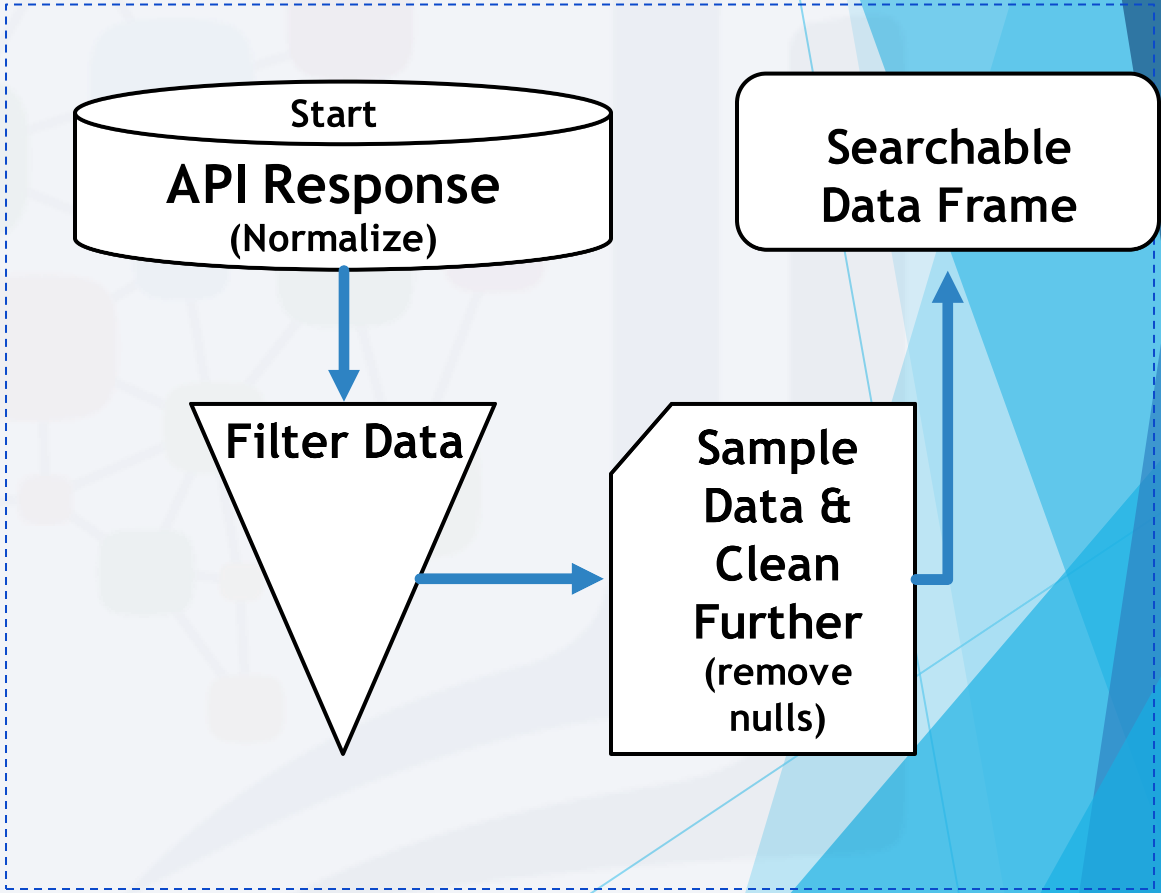
Data Collection – Wrangling

- Scraping (cleaning) Process:

1. Normalize the Data
2. Filter the Data
3. Sample the Data & Clean further (deal w/nulls)
4. Aggregate Data into Searchable DF

- External Reference:

- https://github.com/RMHUNC/Space-Y/blob/a5e9bbf2ebaea82112d6f515a61fdca55242cfb3/spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

Charts & Graphs Plotted

Scatterplot of Launch Number by Payload Mass overlayed by Success or Failure

- This was done to see if the first stage was more likely to land successfully as the flight number increased, and how payload mass affected the likelihood of a successful first stage return.

Flight Number Versus Launch Site

- To see if there is a correlation between flight number and launch site

Payload Size Versus Launch Site:

- To see if payload size determined the launch site

Average Success Rate Versus Orbit (Bar):

- To see if some Orbits have a higher rate of success

Flight Number Versus Orbit Type

- To see if the company decided to focus on a particular orbit type over time

Payload Mass Versus Orbit Type:

- To see if there is a relationship between the payload mass and orbit type

Annual Launch Success Rate Over Years(Line):

- To see if the company became more successful at launches over time

External Reference:

https://github.com/RMHUNC/Space-Y/blob/682bdd1631419b129111edc9ecd4034bc7128a13/EDA%20with%20Visualizations.ipynb

EDA with SQL: Summary of Queries

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome on a ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass using a subquery
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, and launch site for the months in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- External Reference:

➢ https://github.com/RMHUNC/Space-Y/blob/682bdd1631419b129111edc9ecd4034bc7128a13/jupyter-labs-eda-sql-coursera_sqllite(2).ipynb

Build an Interactive Map with Folium

Map objects included on the Folium Interactive Map:

- Marker Icons/labels: Showing Points of Interest: Railway, Highway, City, Coast
- Marker Circles: Showing launch sites
- Marker Clusters: Showing successful and failed launches by launch site
- PolyLines showing Distance between points

The objects included on the Folium Map are used to identify the locations of successful and failed launches, as well as their proximity to relevant cost saving and safety points to assist in determining where a new competitor should look to locate.

- https://github.com/RMHUNC/Space-Y/blob/682bdd1631419b129111edc9ecd4034bc7128a13/LaunchSiteLocationAnalysis.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Dash Graphs:

- Pie Chart showing the Total Number of Successful Launches Count
- Pie Chart showing the Percentage of Successful versus Unsuccessful Launches by Launch Location
- Scatter plot with payload slider showing Successful versus Unsuccessful Launch by Payload Overlaid by Rocket Booster Version

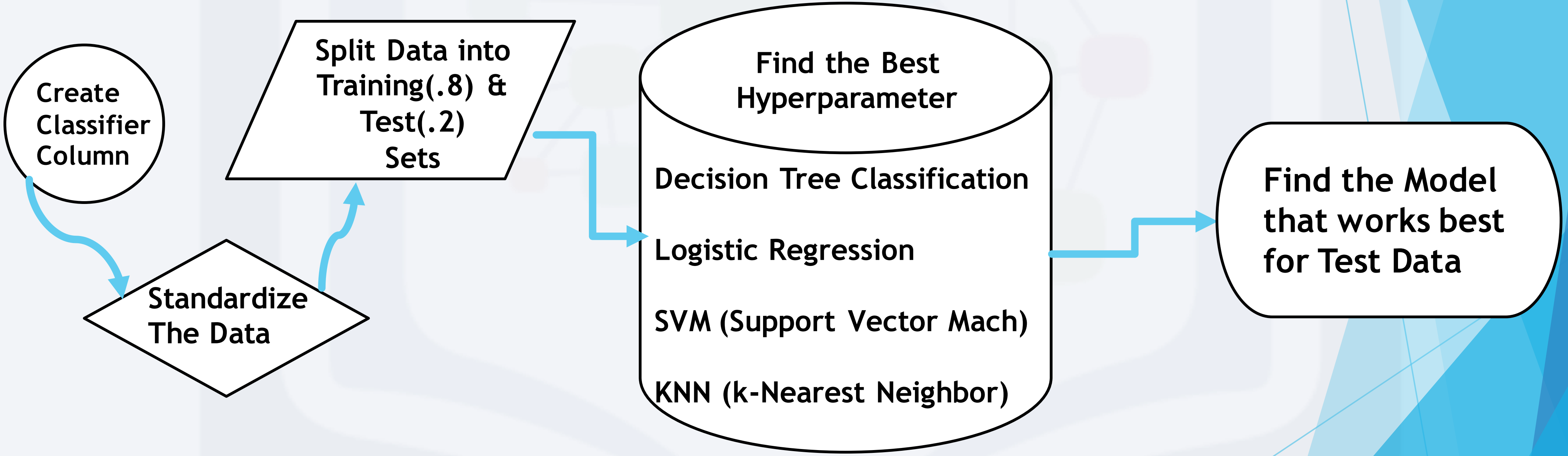
These interactive graphs allow an end user to analyze what booster versions are the most successful and to determine the relationship between payload and launch success, and between launch site and launch success.

External Reference:

https://github.com/RMHUNC/Space-Y/blob/f6e91322649d74ea1e5c7b36bf614979314a104f/spacex_dash_appfinal.py

Predictive Analysis (Classification)

- A classifier was added to the Data, then the Data was standardized
- The Split_Train_Test function was used on Decision Tree Classification, Logistic Regression, KNN, & SVM. The Confusion Matrix and accuracy for all models were compared and then the models were cross compared to determine the best model.



Create Classifier Column

Standardize The Data

Split Data into Training(.8) & Test(.2) Sets

Find the Best Hyperparameter

Decision Tree Classification

Logistic Regression

SVM (Support Vector Mach)

KNN (k-Nearest Neighbor)

Find the Model that works best for Test Data

https://github.com/RMHUNC/Space-Y/blob/3310efec1955e31028f417ab6abef4492460531e/SpaceX%20Machine%20Learning%20Analysis%20Capstone-Final.ipynb

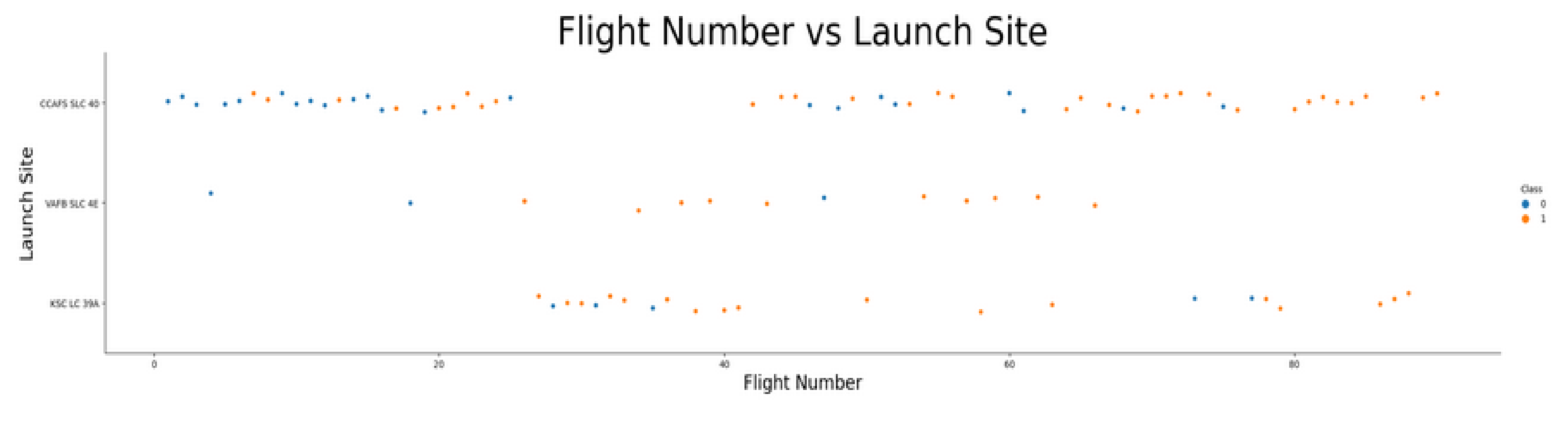
Results

- The Success rates for SpaceX launches is directly proportional to the launch number over time. They become more successful at launches over time.
- Launch pad KSC LC 39A had the most successful launches
- Heavier payloads have a lower success rate than lighter payloads
- Launches to ES-L1, GEO, HEO, & SSO Orbits were the most successful
- Average Payload Mass has increased over time
- Vandenburg Air Force Base (VAFB) is the least frequently used launch site
- The Decision Tree Classification Model worked best on predicting the outcome of the test group with an accuracy score of 86%.

Section 2

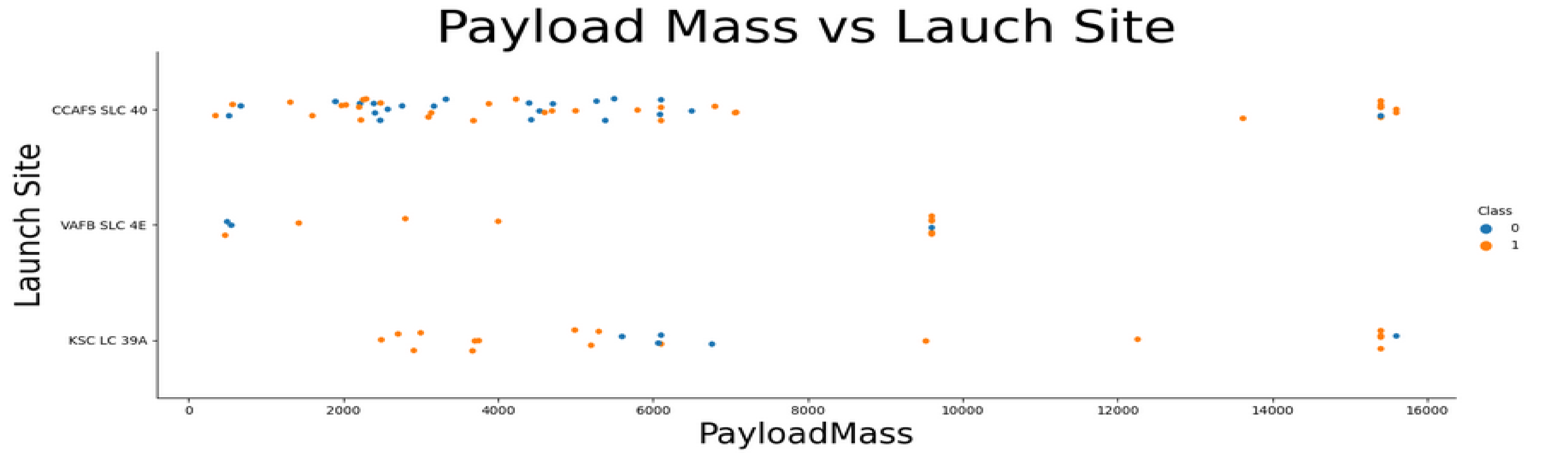
Insights drawn from EDA

Flight Number vs. Launch Site



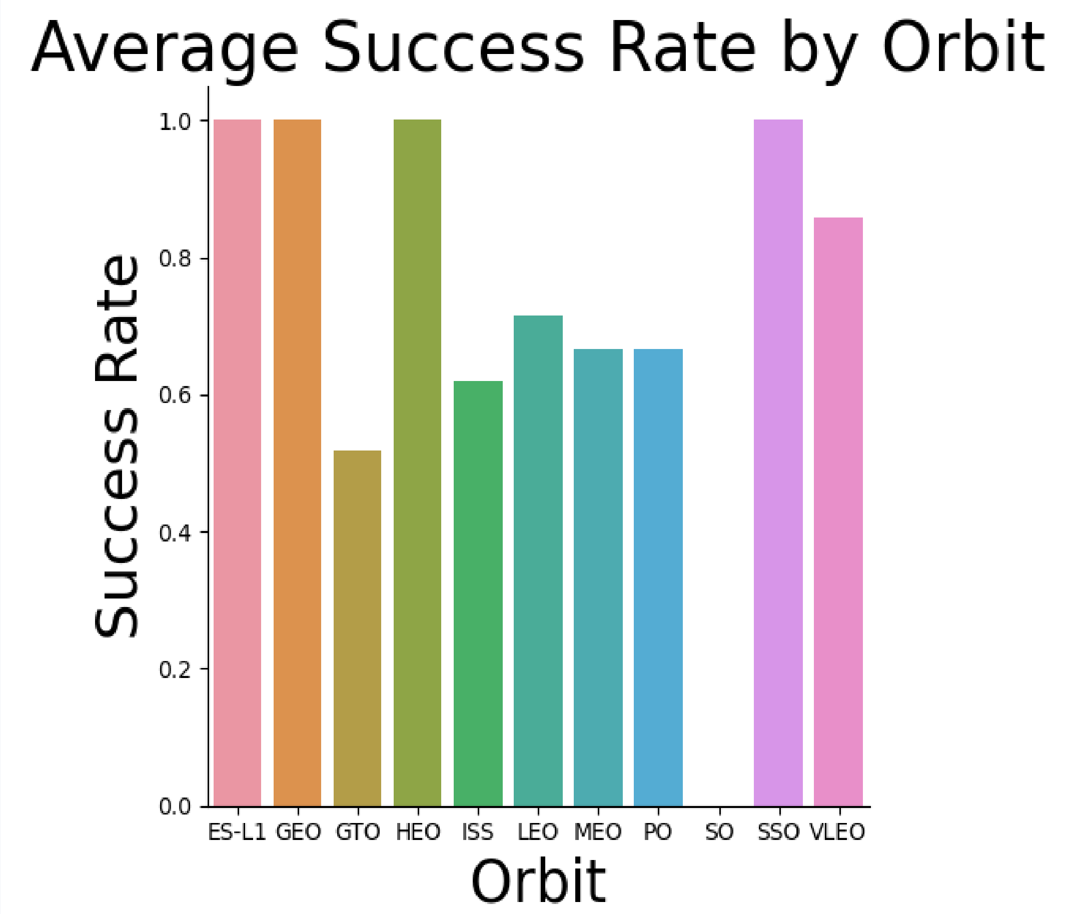
- It is clear that Vandenburg Air Force Base is infrequently used and has a high failure rate
- Cape Canaveral Space Force Center handles the majority of all launches

Payload vs. Launch Site



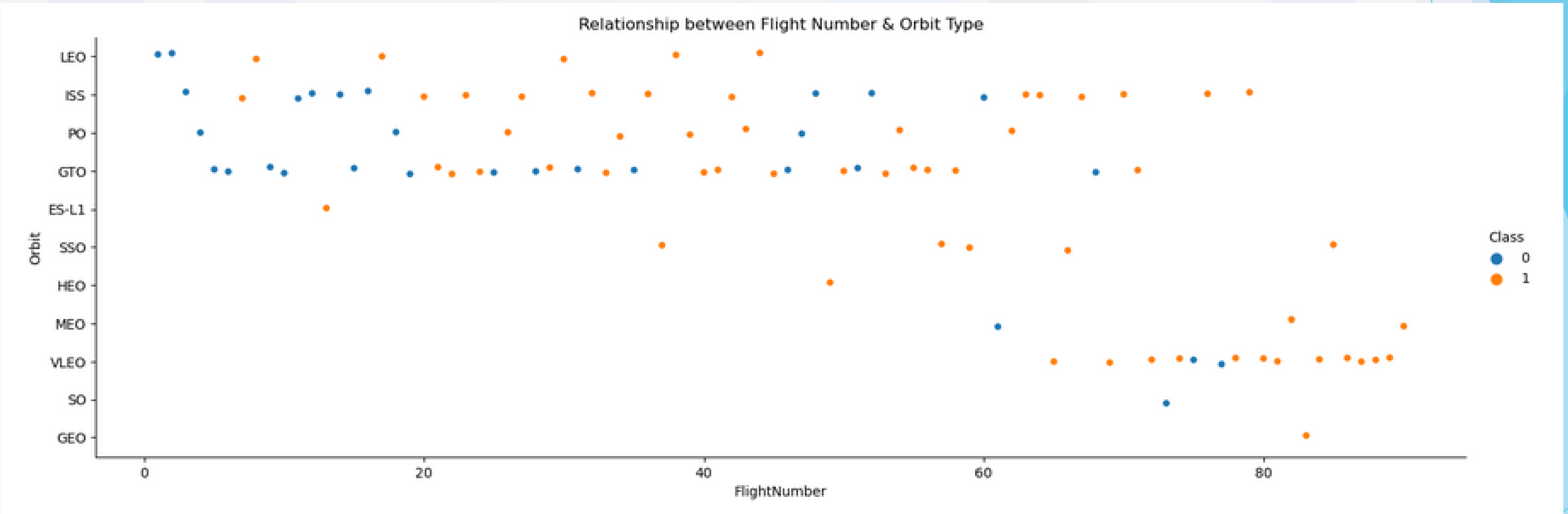
- Kennedy Space Center is used predominately for Payload launches between 2000-6500kg
- Vandenburg AFB is used predominately for low Payload launches

Success Rate vs. Orbit Type



- ➢ The ES-L1, GEO, HEO, & SSO Orbits had the highest success rate of all the Orbits.
- ➢ The SO had the lowest success rate at 0%, followed by the GTO at around 50%

Flight Number vs. Orbit Type

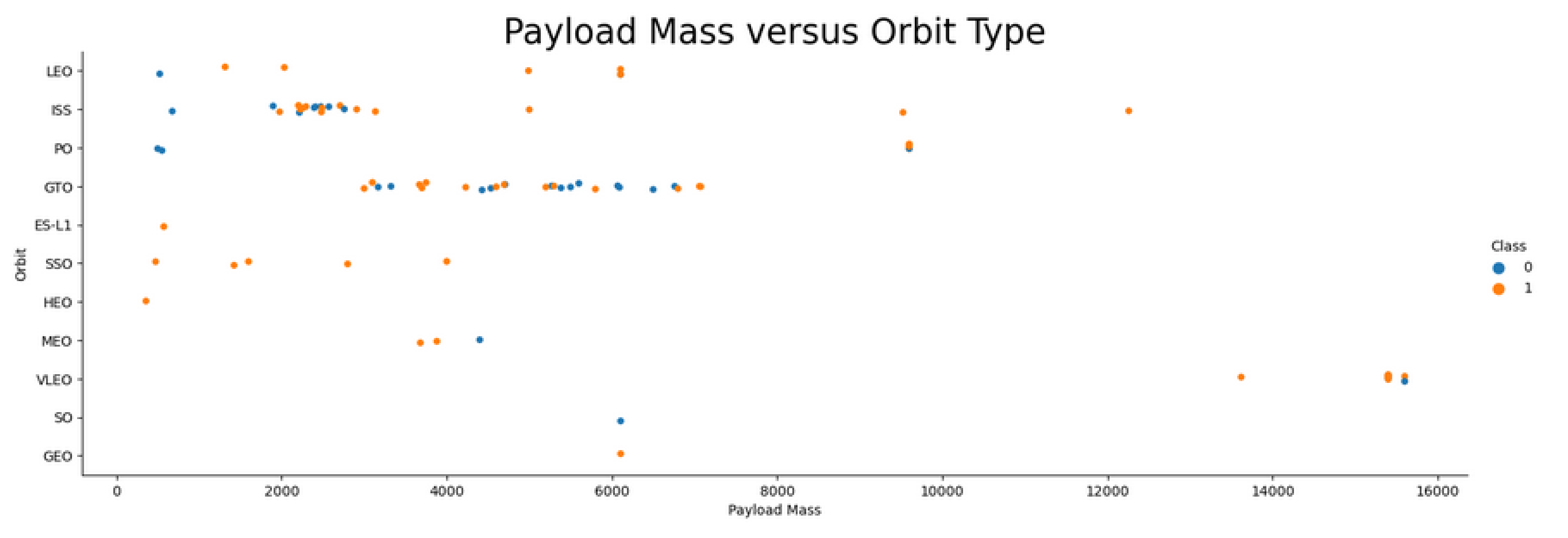


- Over time there is a trend towards launches to VLEO, perhaps because they are profitable

IBM Developer

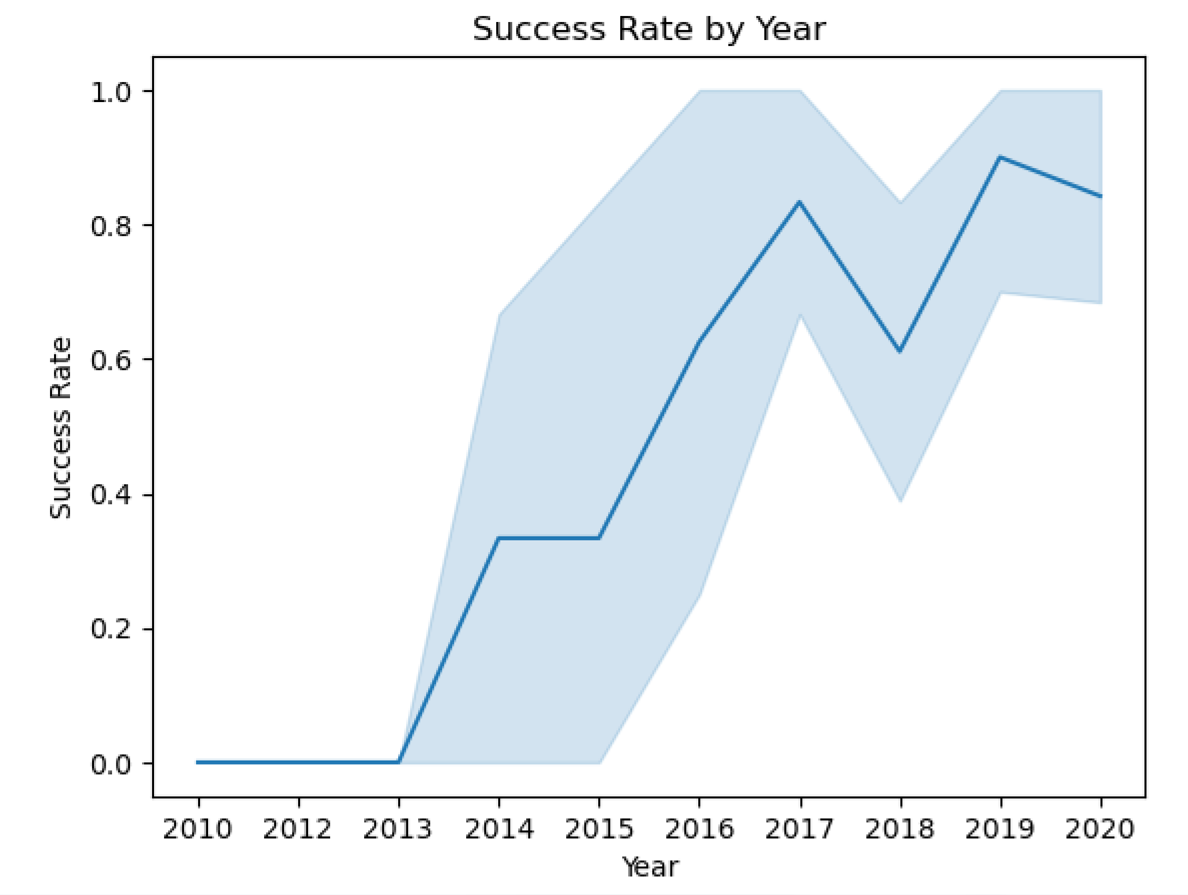
SKILLS NETWORK

Payload vs. Orbit Type



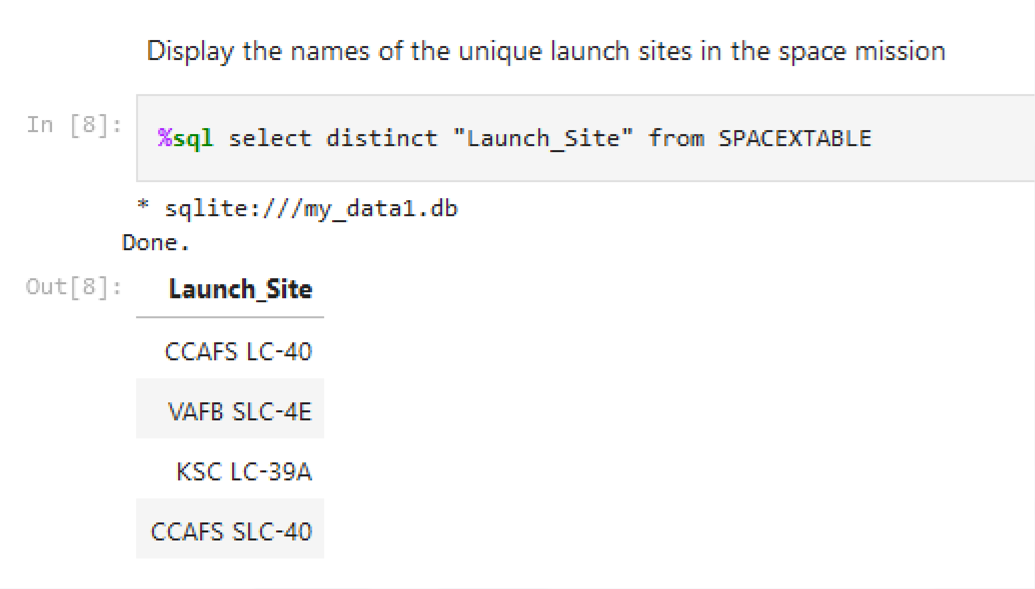
- Payloads to ISS tend to be between 2,000-3,000kg
- Payloads to GTO tend to be between 3,000-7,000kg

Launch Success Yearly Trend



SpaceX's Launch Success Rate Greatly improved over time

All Launch Site Names



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

```
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

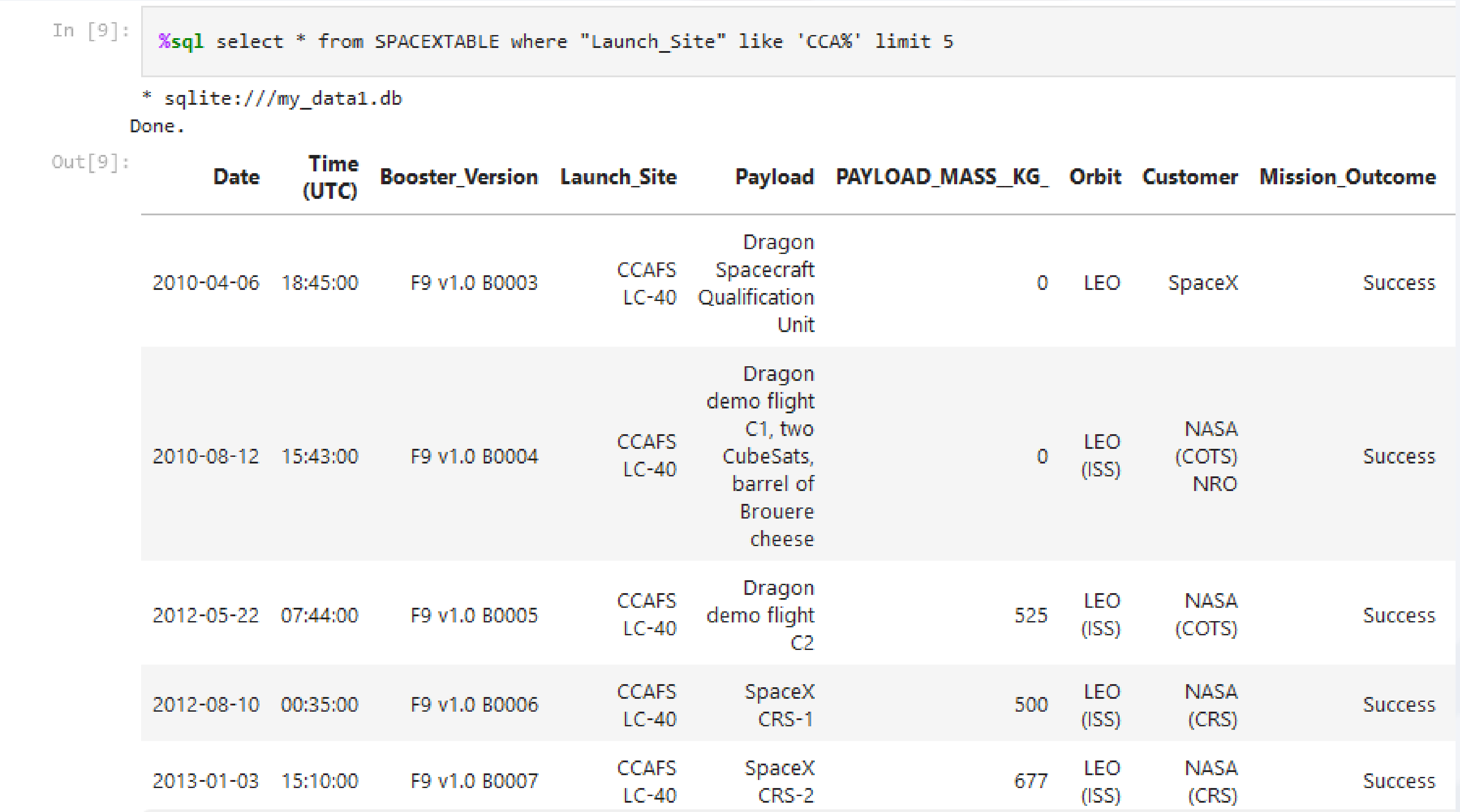
CCAFS LC-40
Cape Canaveral Space Force Center

VAFB SLC-43
Vandenburg Air Force Base

KSC LC-39A
Kennedy Space Center

CCASF SLC-40
Cape Canaveral Space Force Center

Launch Site Names Begin with 'CCA'



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

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

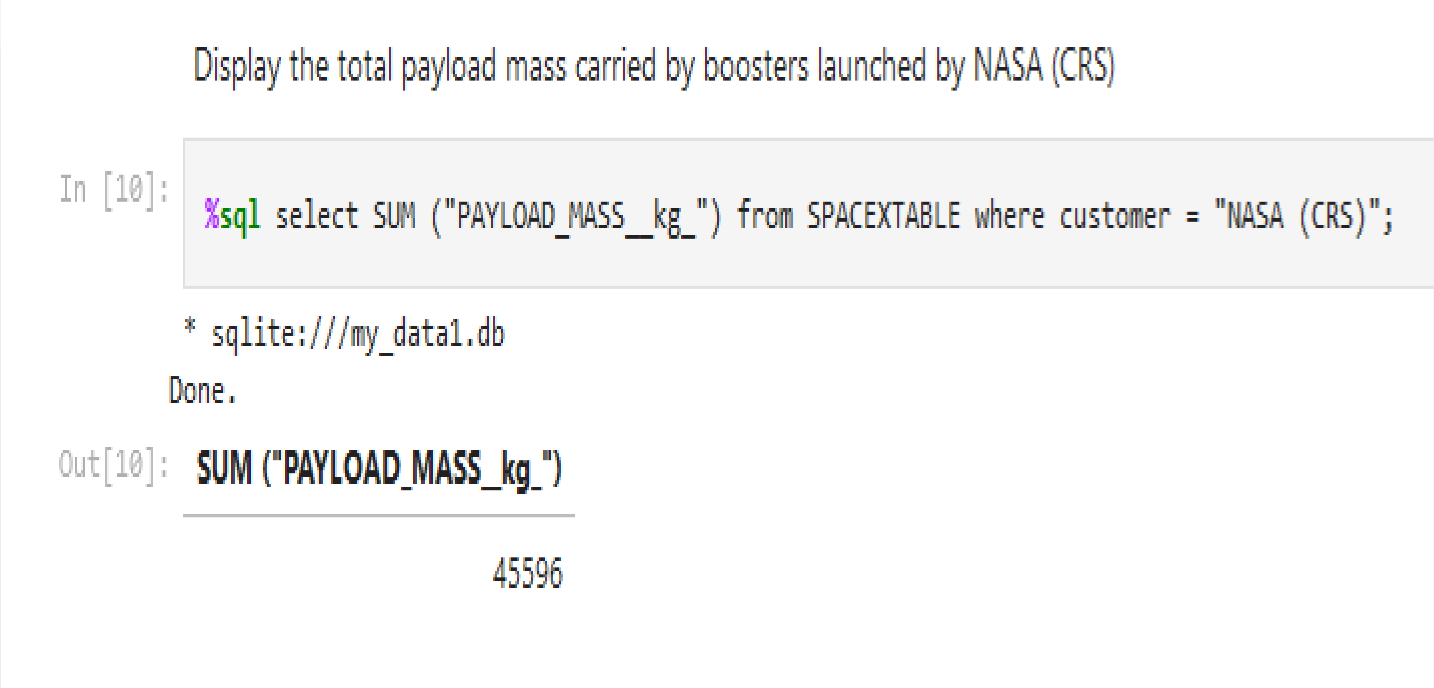
Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-08-12	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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Display of 5 records where the launch site begins with the string “CCA”

Narrow the results by using “limit”

Total Payload Mass



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

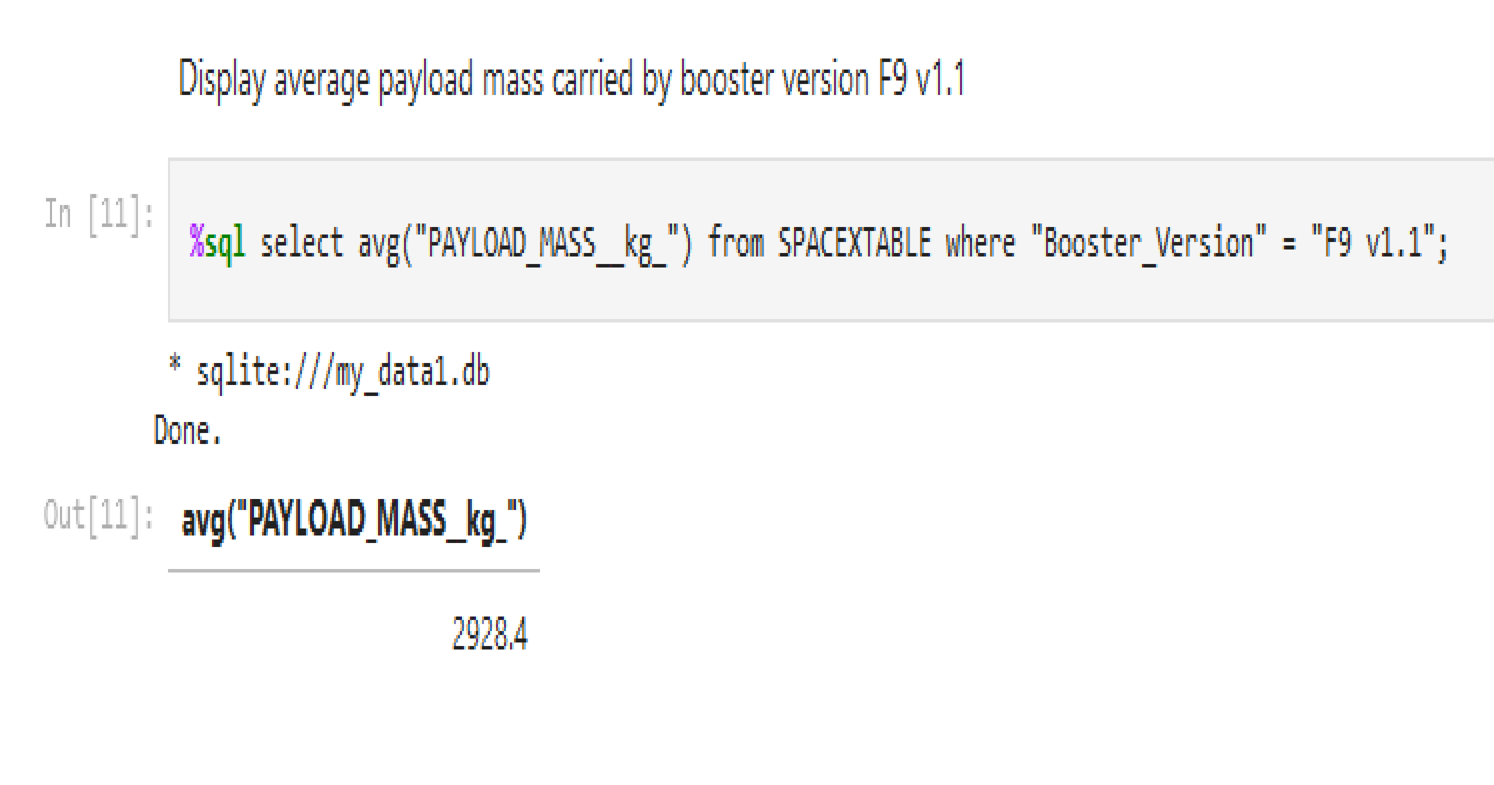
```
In [10]: %sql select SUM ("PAYLOAD_MASS__kg_") from SPACEXTABLE where customer = "NASA (CRS)";
```

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

```
Out[10]: SUM ("PAYLOAD_MASS__kg_")
         ----------------------------
                            45596
```

The total mass launched by NASA(CRS) was 45,596kg

Average Payload Mass by F9 v1.1



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

```
In [11]: %sql select avg("PAYLOAD_MASS__kg_") from SPACEXTABLE where "Booster_Version" = "F9 v1.1";
```

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

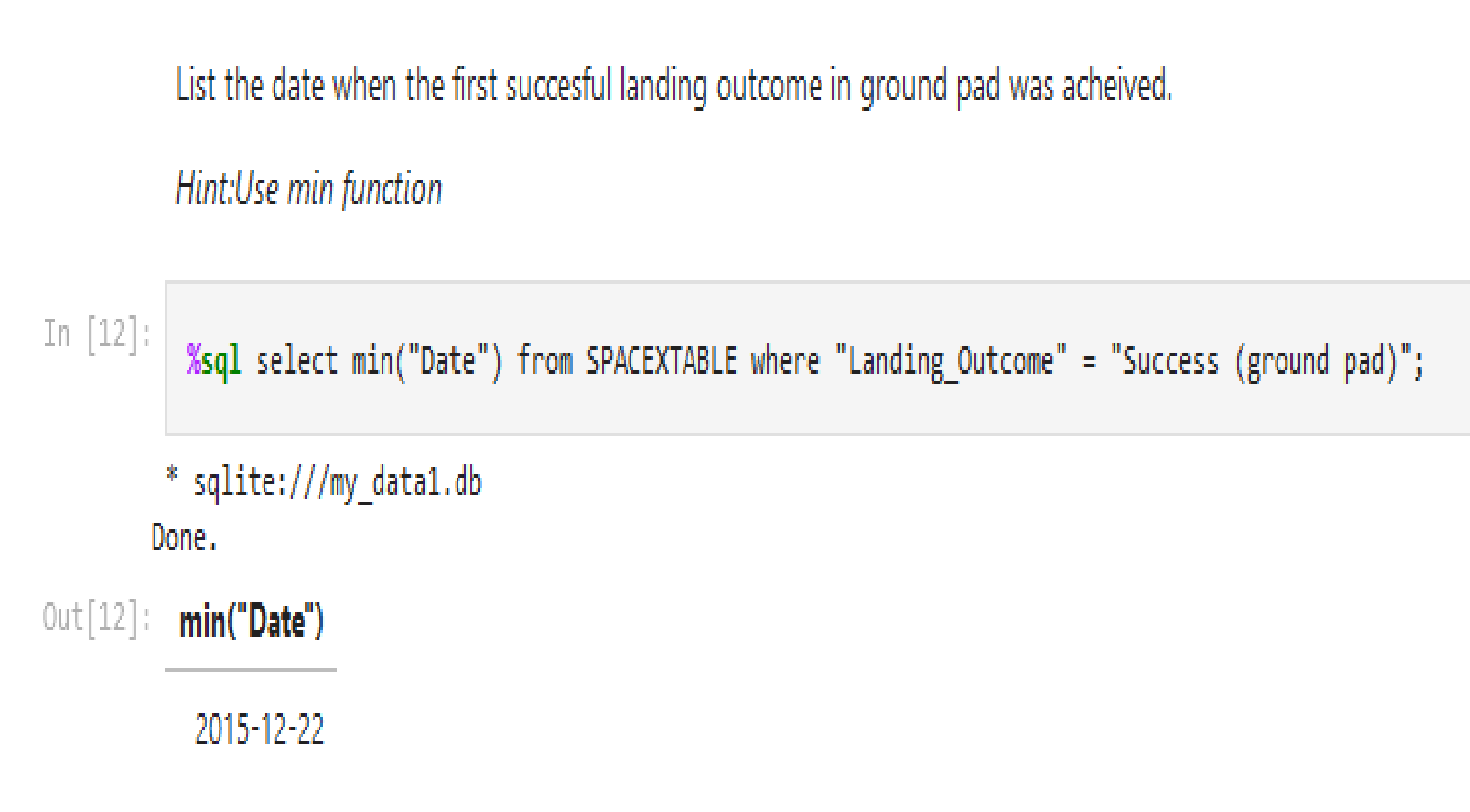
```
Out[11]: avg("PAYLOAD_MASS__kg_")
                          2928.4
```

The average payload mass carried by booster F9v1.1 was 2,928.4 kg

IBM Developer

SKILLS NETWORK

First Successful Ground Landing Date



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

Hint:Use min function

In [12]:

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";
```

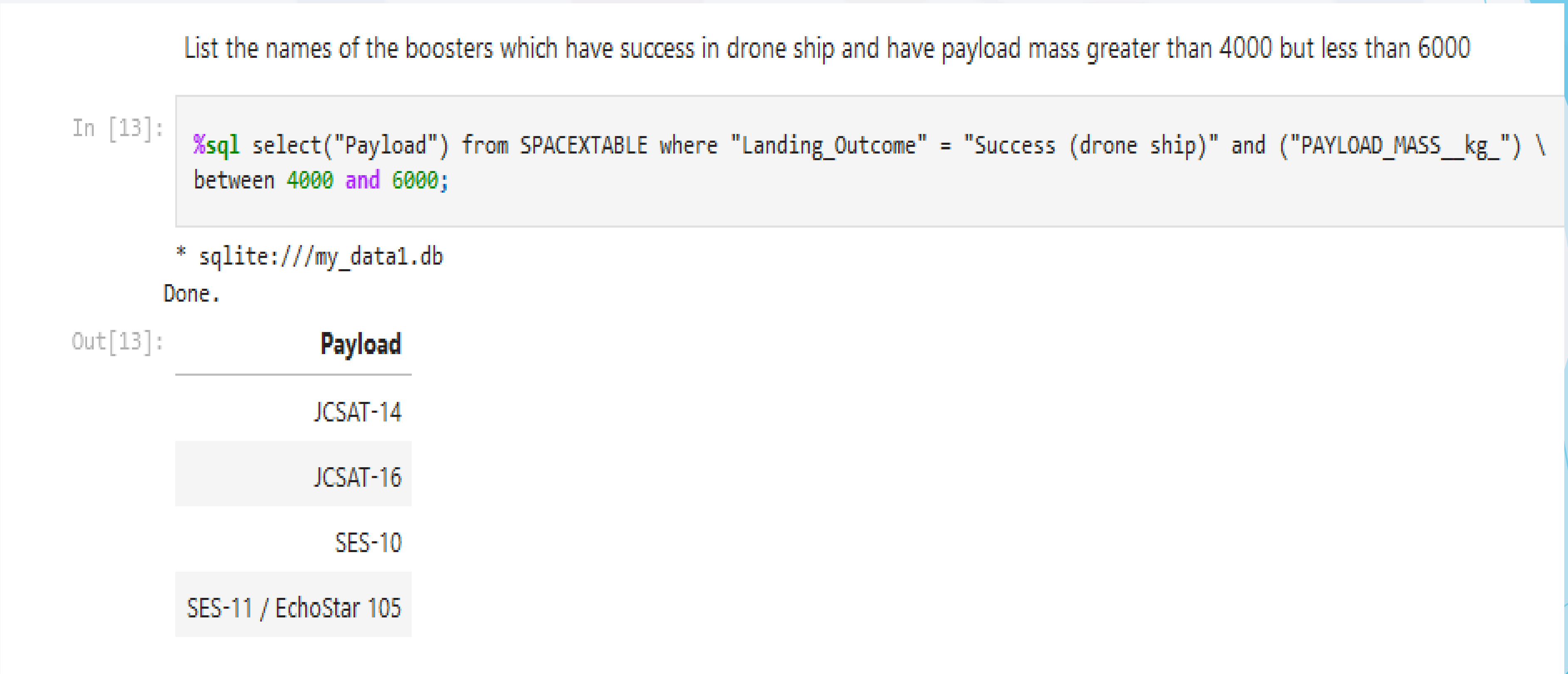
```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

min("Date")
2015-12-22

The first successful landing on a ground pad occurred on Dec. 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000



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

In [13]:

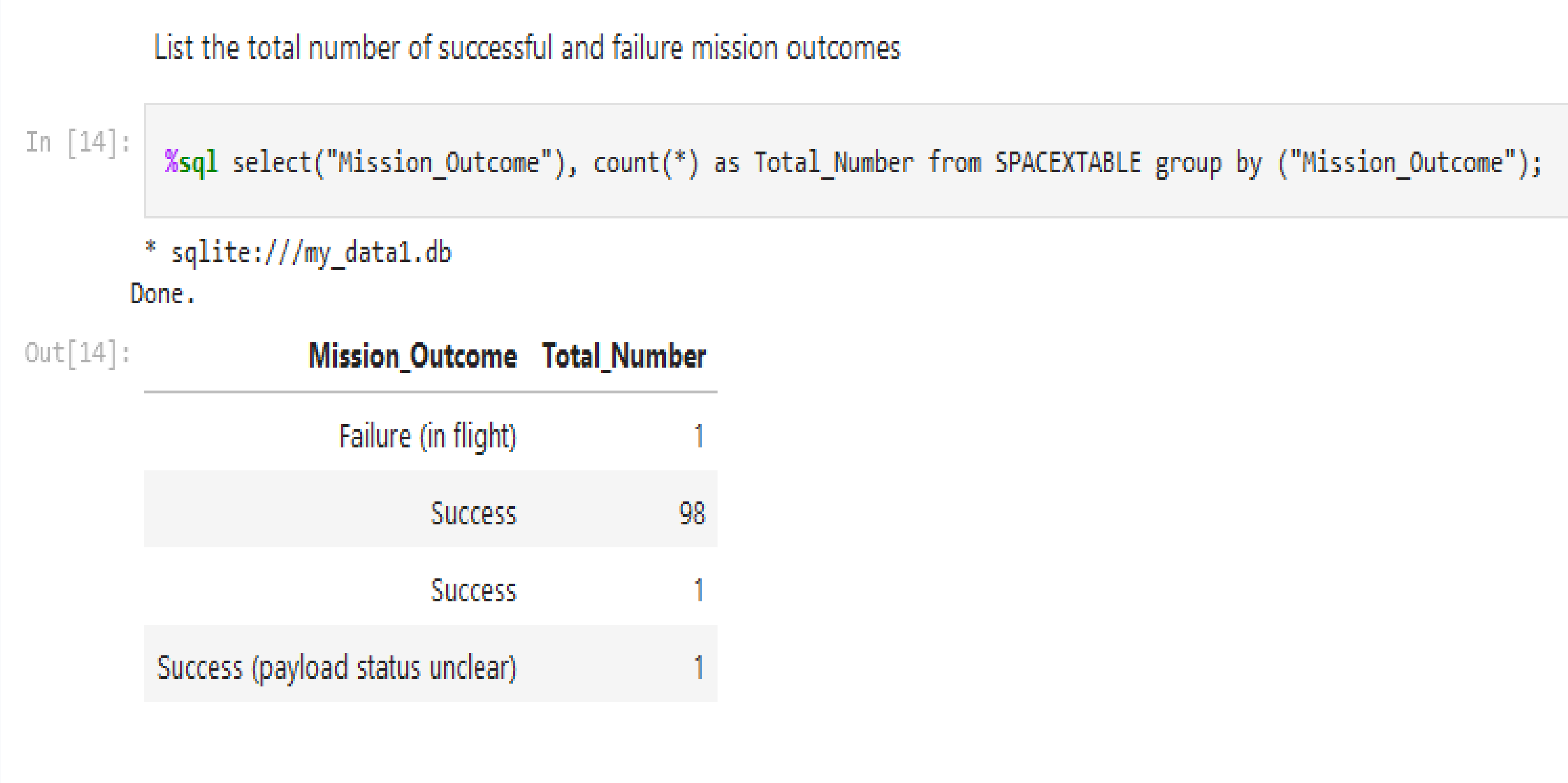
```
%sql select("Payload") from SPACEXTABLE where "Landing_Outcome" = "Success (drone ship)" and ("PAYLOAD_MASS__kg_") \
between 4000 and 6000;
```

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

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

In [14]:

```
%sql select("Mission_Outcome"), count(*) as Total_Number from SPACEXTABLE group by ("Mission_Outcome");
```

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

Out[14]:

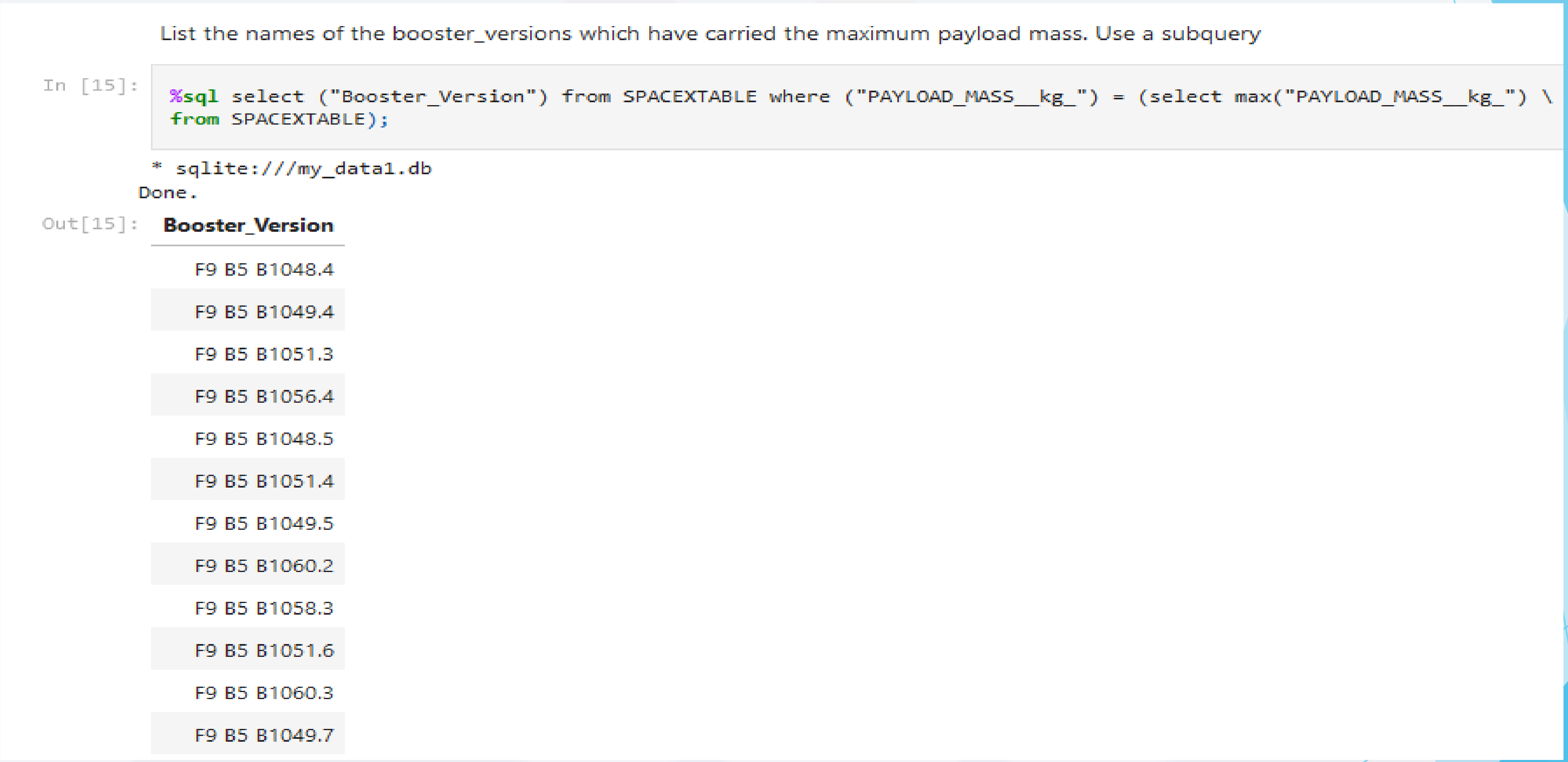
Mission_Outcome	Total_Number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful mission outcomes is 100 with only 1 failure

IBM Developer

SKILLS NETWORK

Boosters Carried Maximum Payload



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [15]:

```
%sql select ("Booster_Version") from SPACEXTABLE where ("PAYLOAD_MASS__kg_") = (select max("PAYLOAD_MASS__kg_") \
from SPACEXTABLE);
```

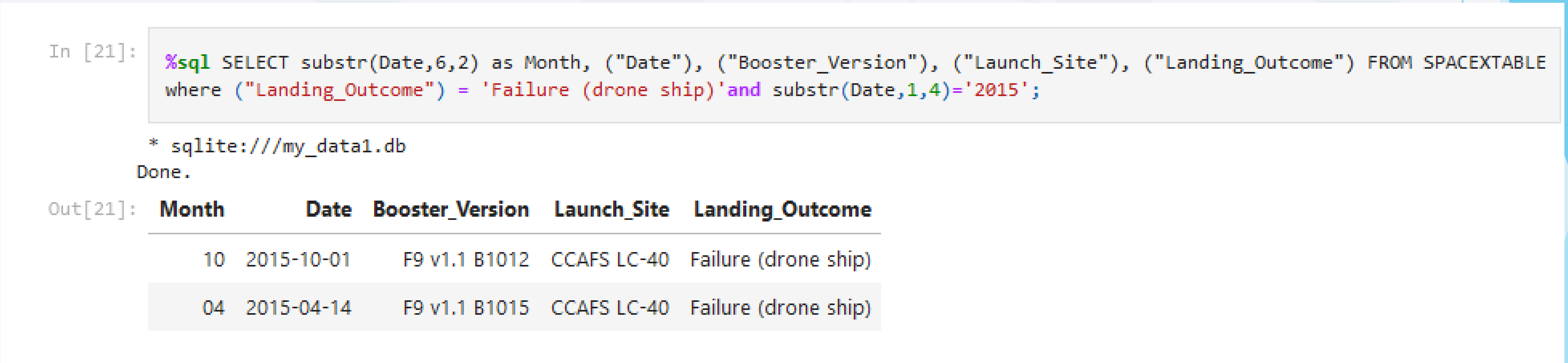
```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

Using a subquery, one can see that several booster versions have carried the maximum payload.

2015 Launch Records



In [21]:

```
%sql SELECT substr(Date,6,2) as Month, ("Date"), ("Booster_Version"), ("Launch_Site"), ("Landing_Outcome") FROM SPACEXTABLE
where ("Landing_Outcome") = 'Failure (drone ship)'and substr(Date,1,4)='2015';
```

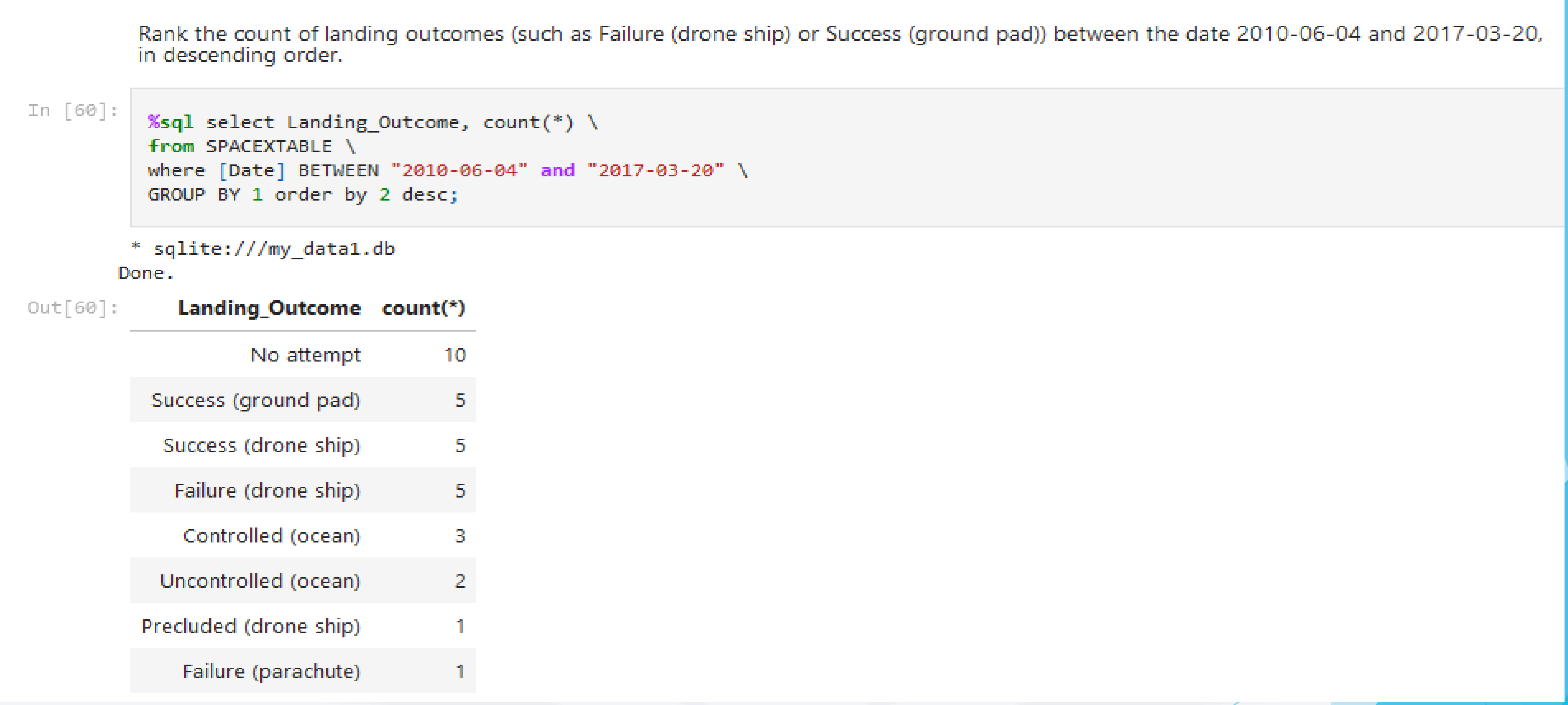
```
 * sqlite:///my_data1.db
Done.
```

Out[21]:

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Display of the month names, failure landing outcomes in drone ship, booster versions, launch site, and for the months in year 2015

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



Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
In [60]:  %sql select Landing_Outcome, count(*) \
          from SPACEXTABLE \
          where [Date] BETWEEN "2010-06-04" and "2017-03-20" \
          GROUP BY 1 order by 2 desc;
```

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

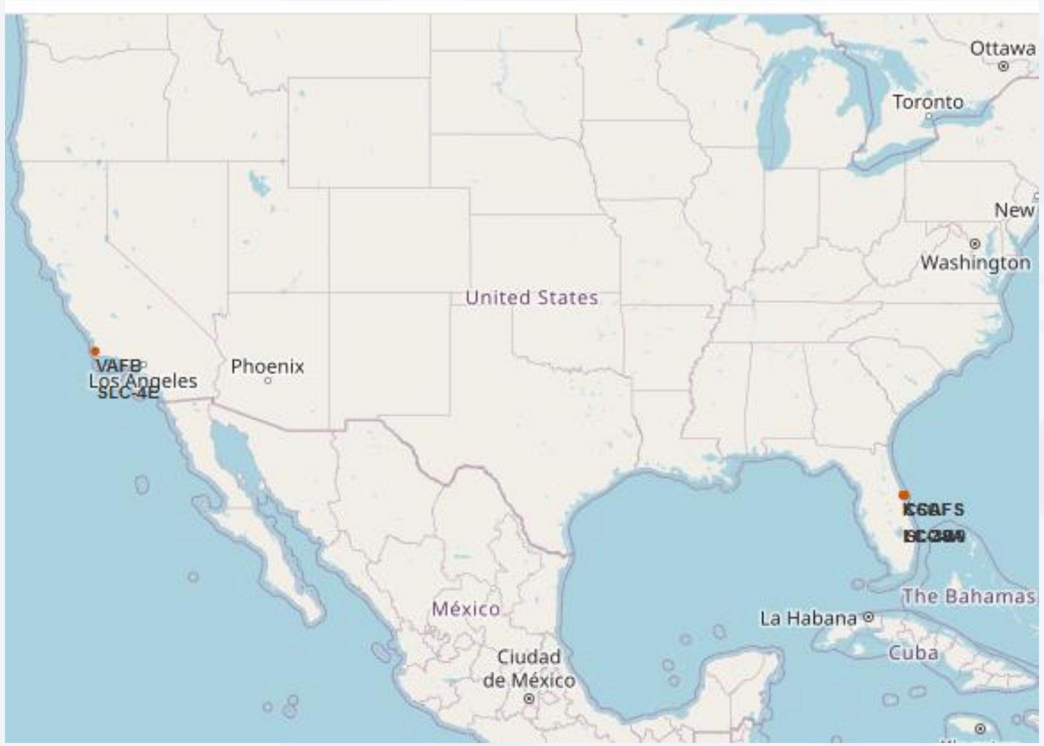
Out[60]:

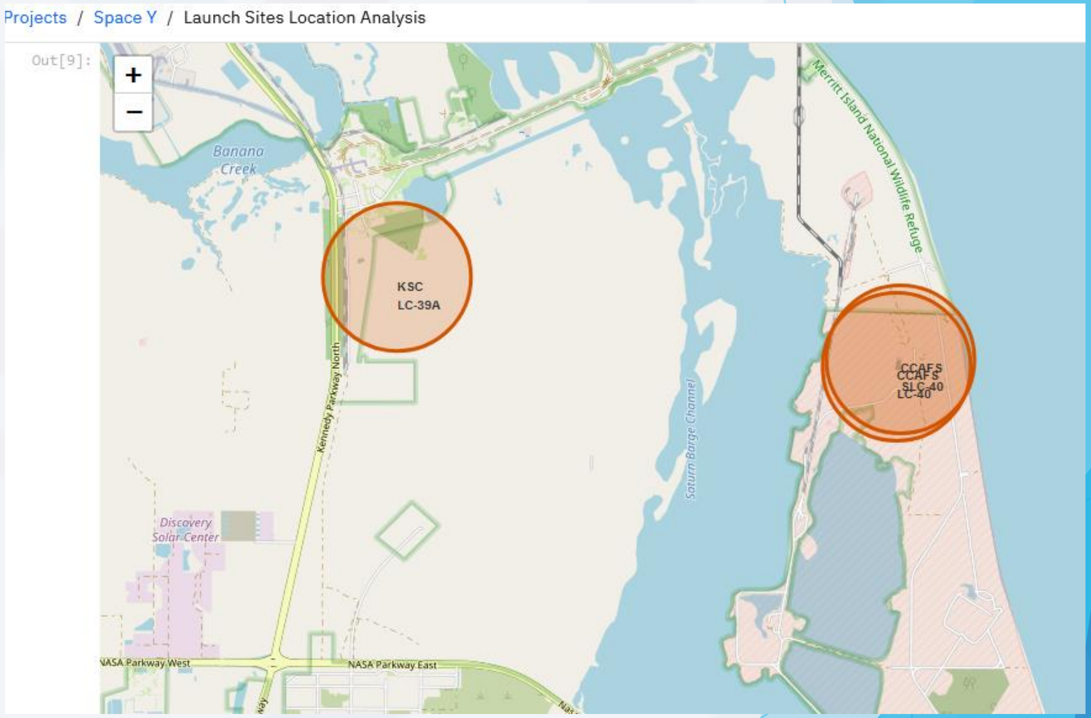
Landing_Outcome	count(*)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

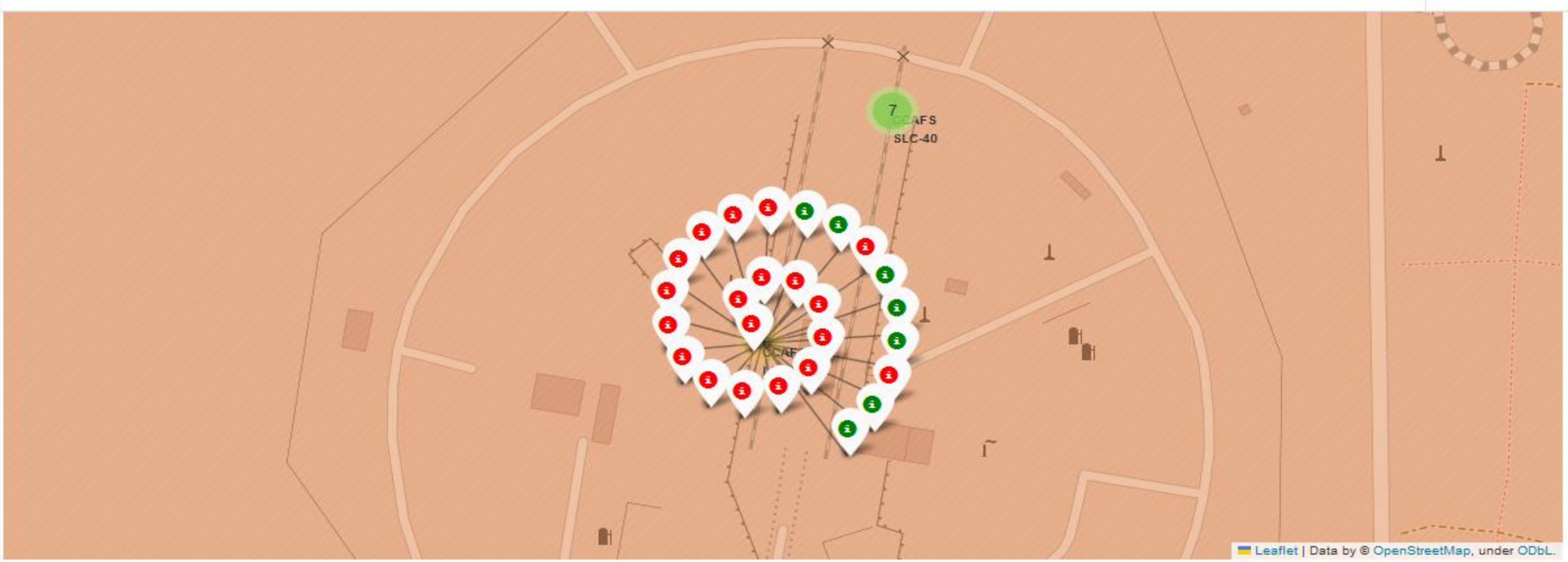
Folium Maps Showing all 4 Launch Sites




Projects / Space Y / Launch Sites Location Analysis

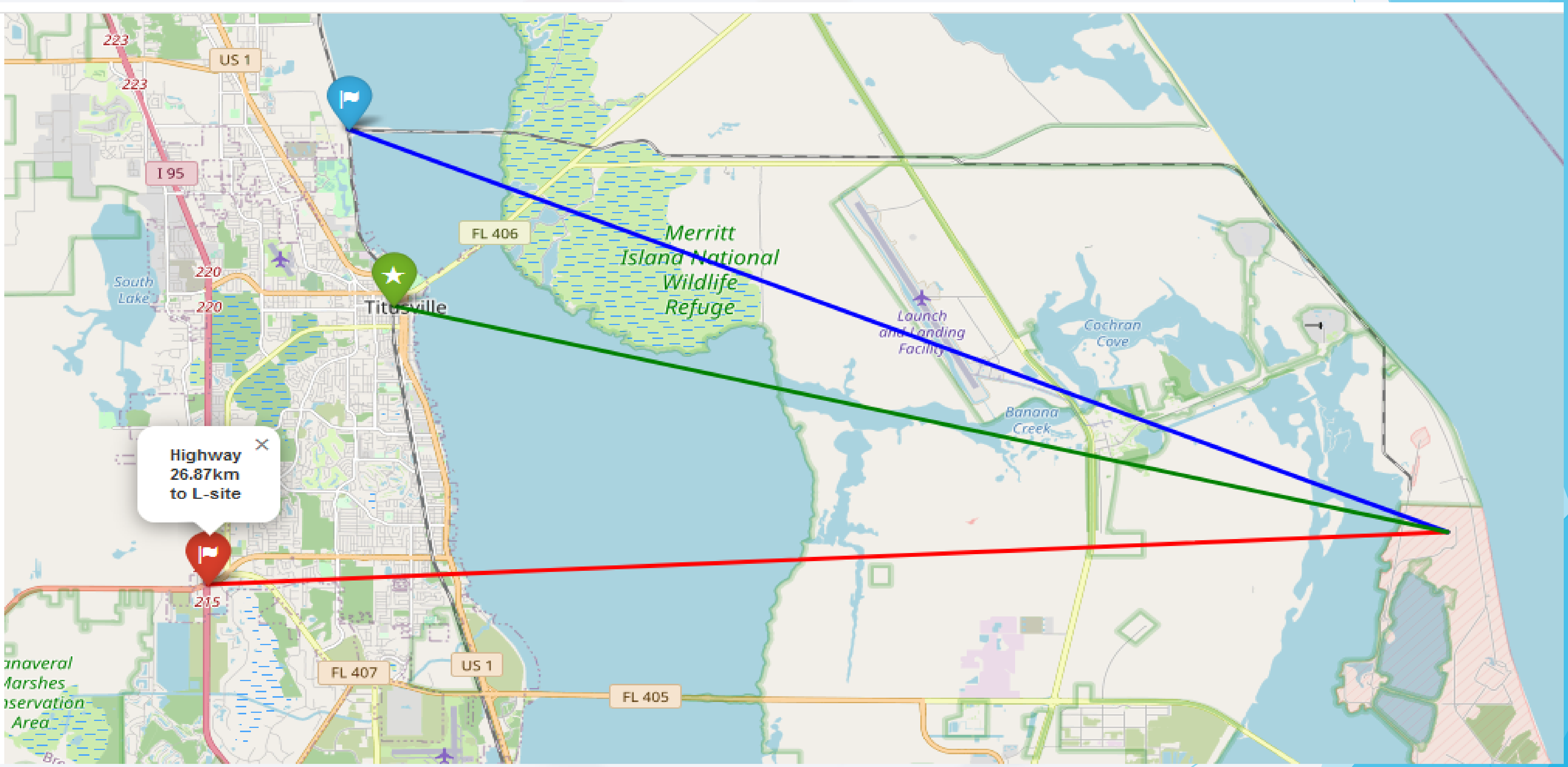
Out[9]:

Map of Success vs Failures of a Launch Site



This screenshot shows each instance of a successful (green) or failed (red) launch at launch site CCASF- LS40. It is important to be able to see the success and failure rates at each launch site to see if any conclusions can be made.

Proximity of Relevant Geographical Features to a Launch Site



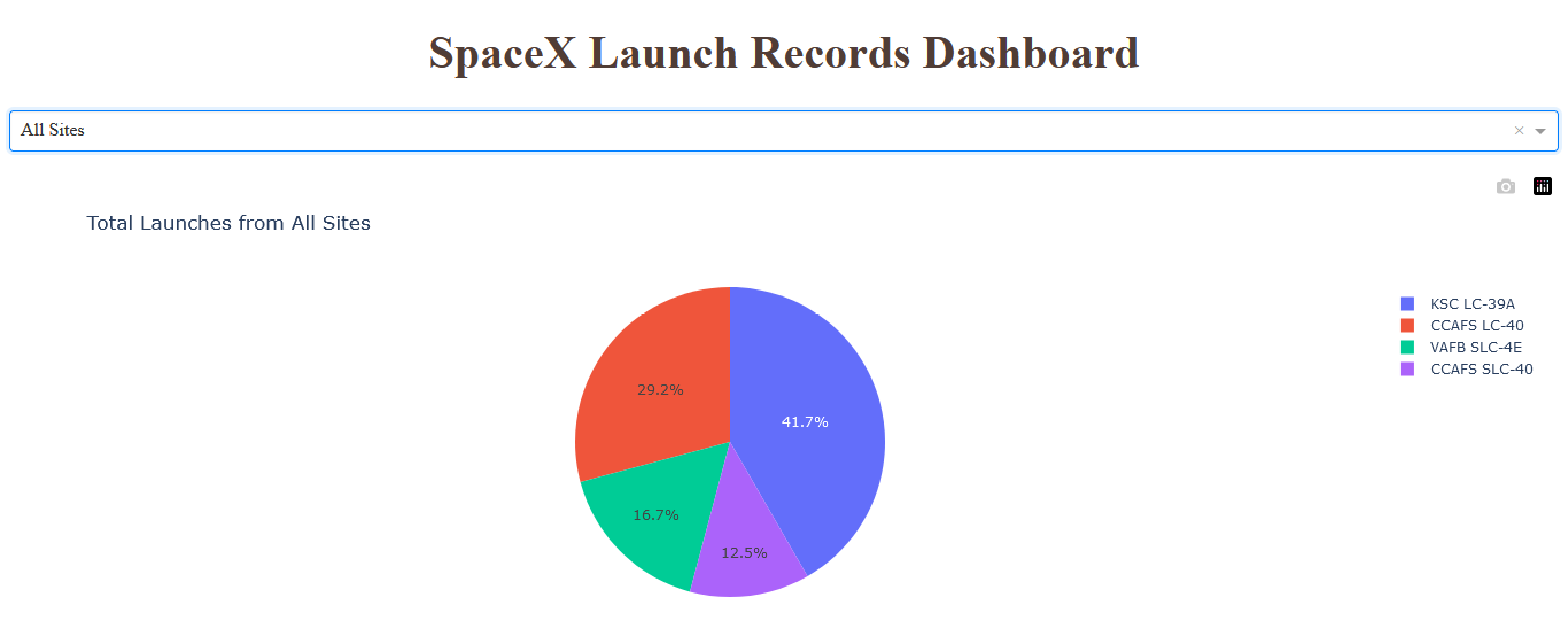
IBM Developer

SKILLS NETWORK

Section 4

Build a Dashboard with Plotly Dash

Total Successful Launches by Launch Site



SpaceX Launch Records Dashboard

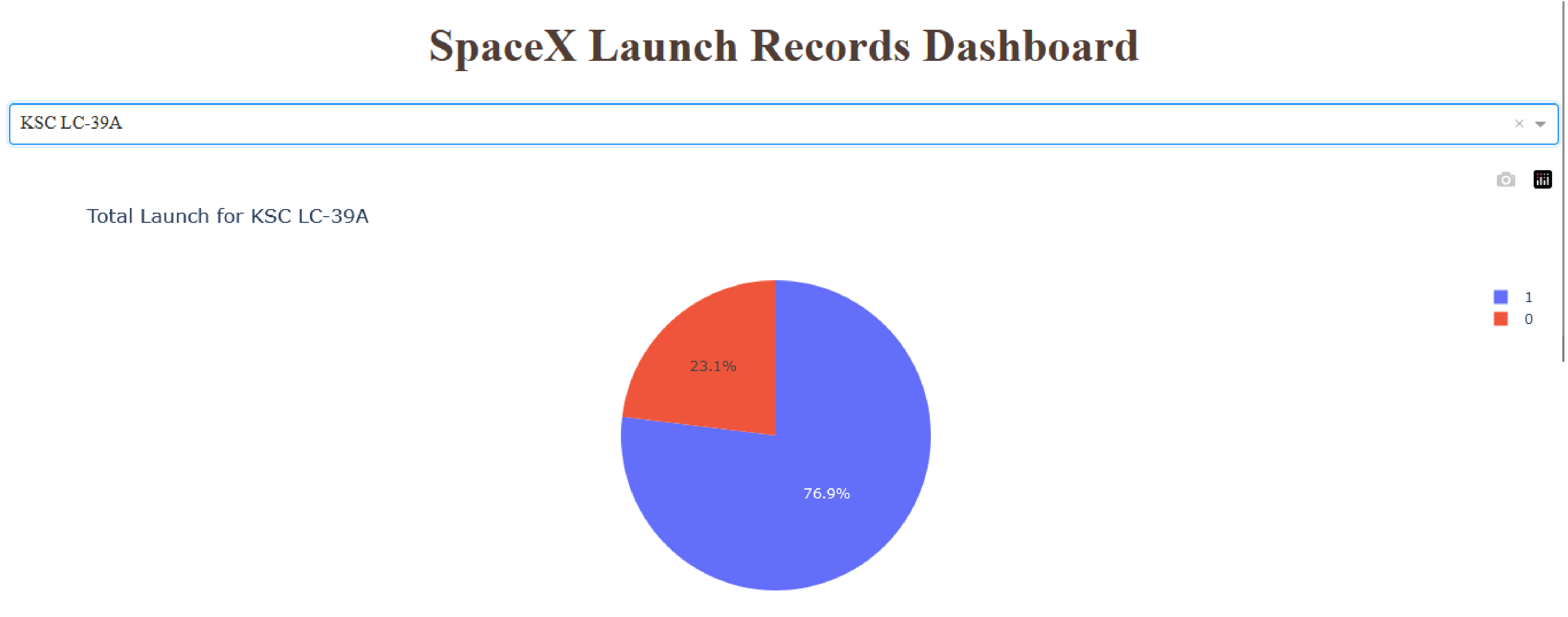
All Sites

This chart tells one which launch sites have the most successful number of launches

IBM Developer

SKILLS NETWORK

Launch Site with the Highest Percentage Success Rate



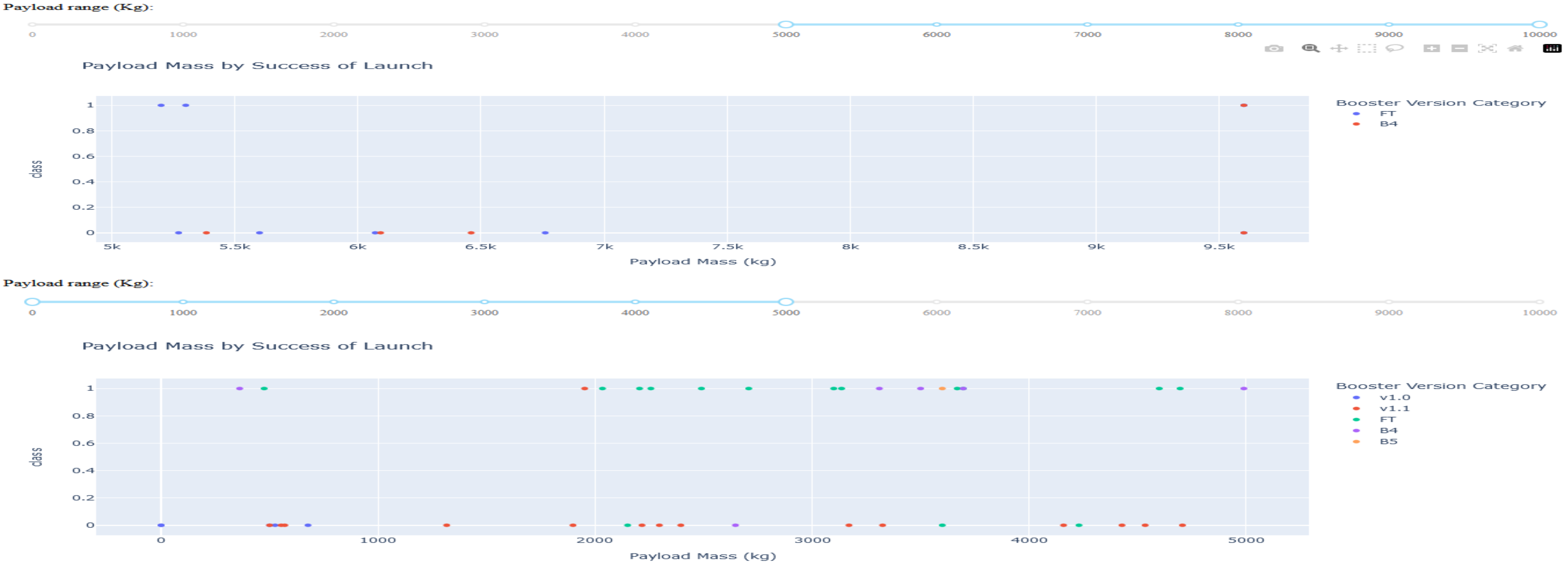
SpaceX Launch Records Dashboard

KSC LC-39A

Total Launch for KSC LC-39A

Kennedy Space Center’s LC-39A had the highest percentage of successful launches at 76.9%

Success Rate of Booster Version by Payload



The FT Booster performs best in both the >5000kg Payload & in the <5000kg Categories

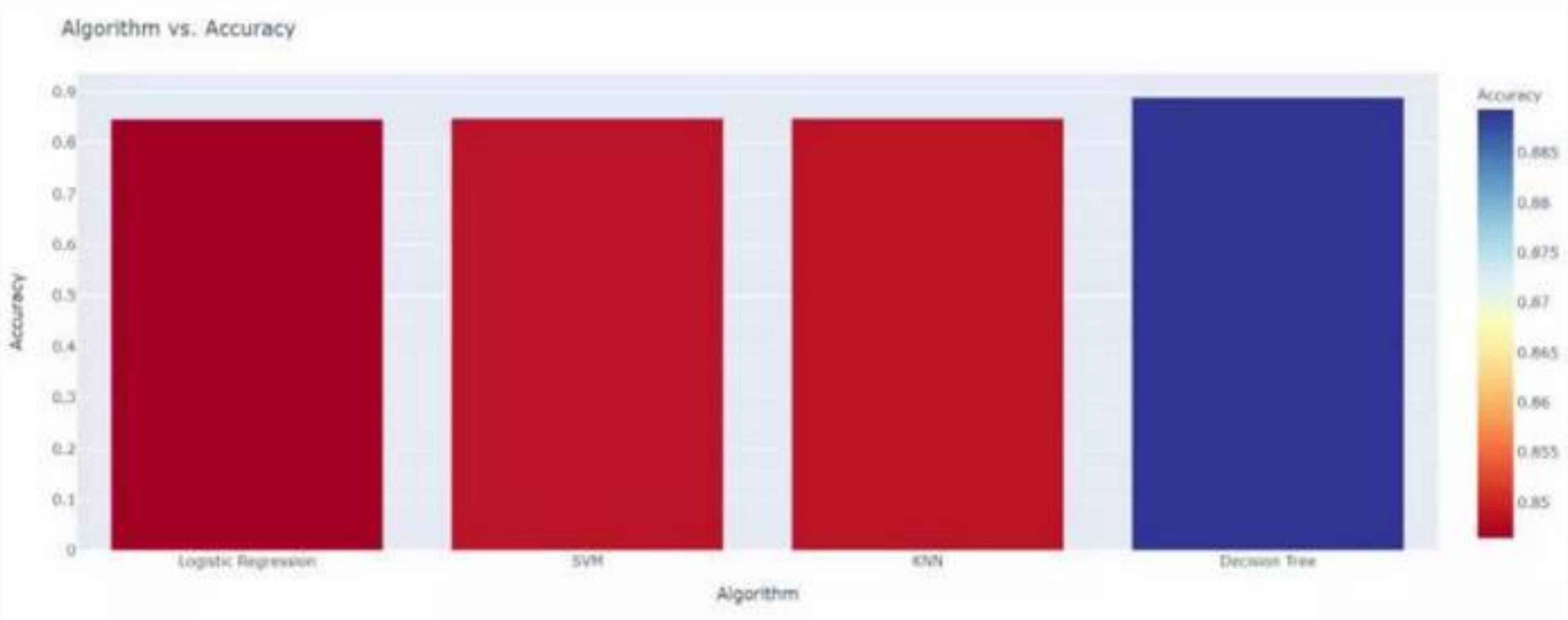
Section 5

Predictive Analysis (Classification)

IBM Developer

SKILLS NETWORK

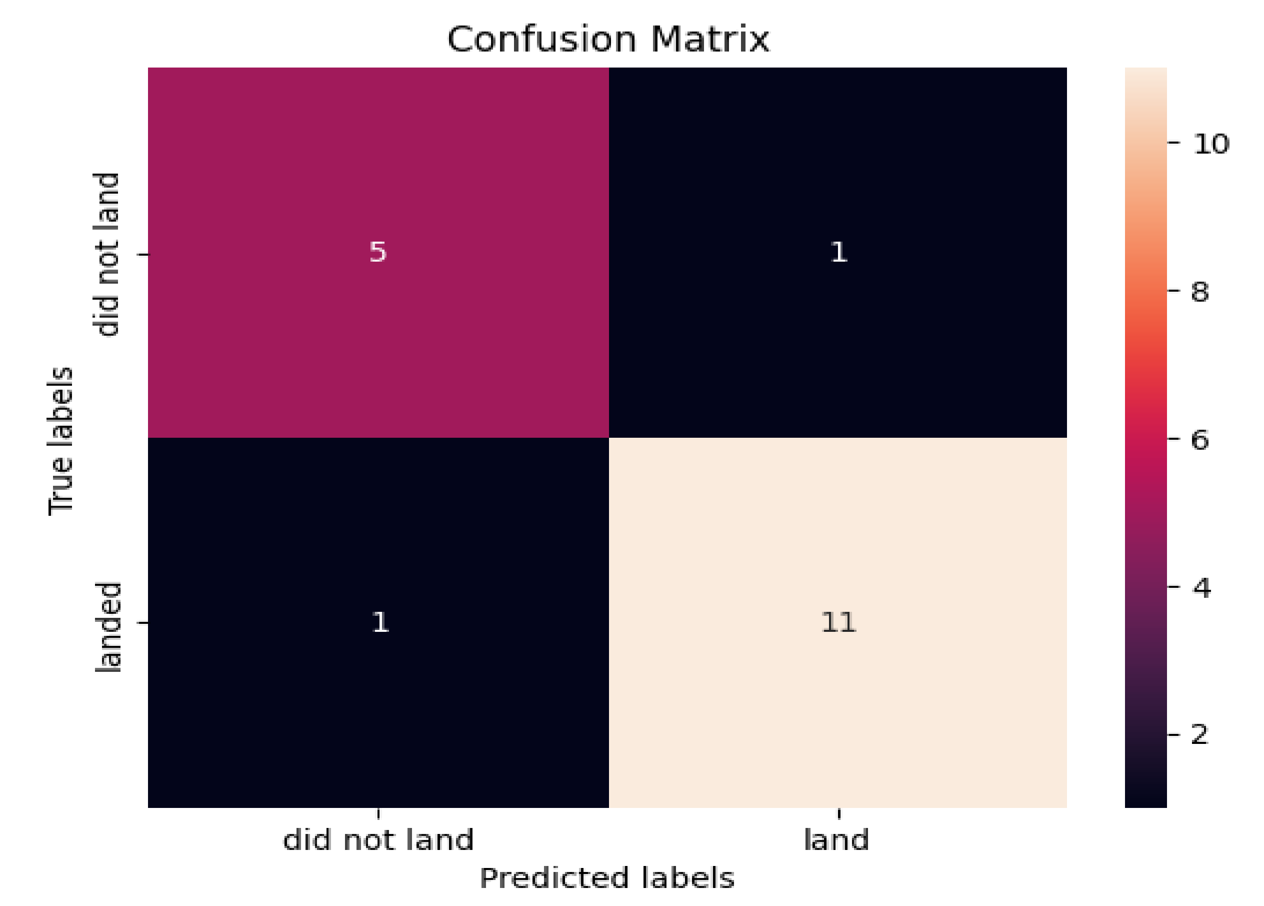
Classification Accuracy



The Decision Tree Classification model had the highest accuracy rate at 88.8%

Confusion Matrix

Decision Tree Confusion Matrix



- The decision tree classification model performed the best on the test set
- It resulted in only 1 False Positive & 1 False Negative
- It accurately predicted 5 out of 6 failures
- It accurately predicted 11 out of 12 successful outcomes
- It’s accuracy rate was 86.9% when paired with the highest reliability rate

Conclusions

- The FT Booster has the highest success rate in both low and high Payload Mass launches indicating that it is very reliable and should be used whenever possible.
- Launch site location should be near the coast for safety reasons, and close to railways, interstate highways, and urban centers for economic reasons.
- SpaceX has learned from their mistakes as their launch success rate as improved greatly over time
- The SO and GTO Orbits should be avoided as they may not be profitable until more analysis is conducted into their high failure rate
- One important factor NOT taken into account in this project was the weather conditions during booster recovery. Since it can play an important role on recovery success, further analysis should be done which includes weather as a variable for recovery success

Appendix

- All codes and analysis can be found on my GitHub SpaceY repository
- https://github.com/RMHUNC/Space-Y
- I would like to take all the Instructors and Teaching Assistants at IBM for all the hard work they put into teaching the Data Science Professional Certification Program!

Thank you!