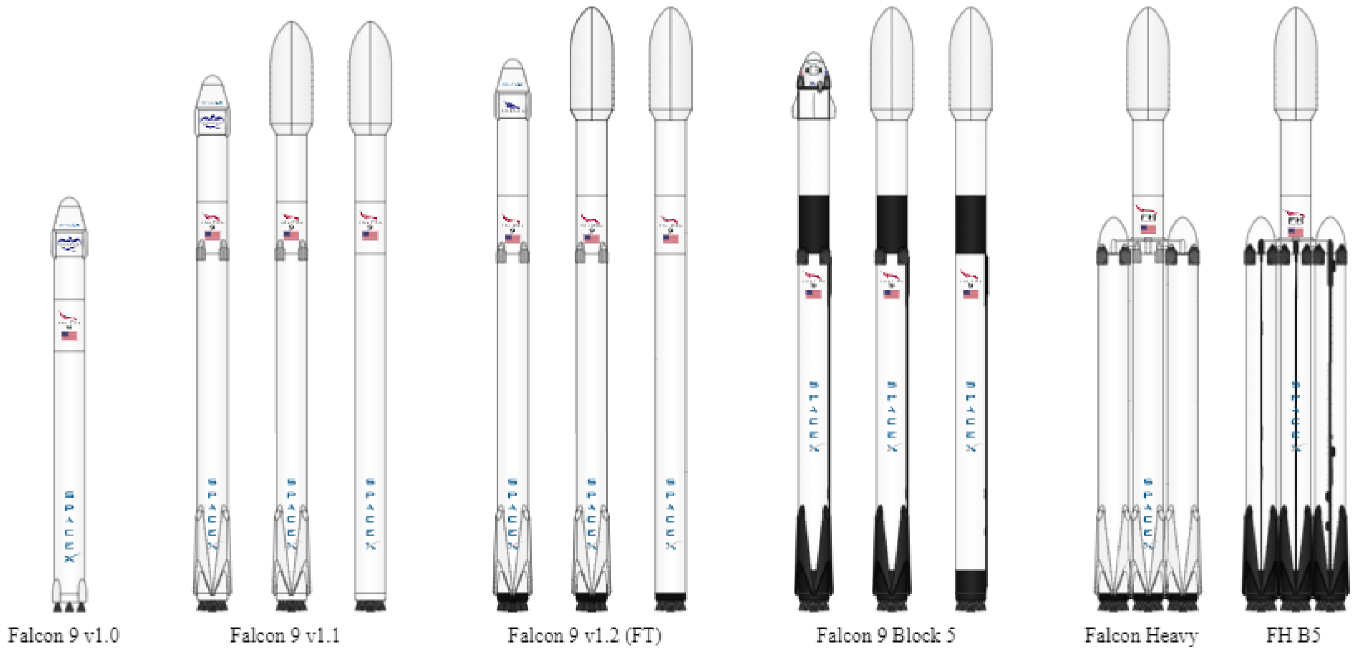
IBM Data Science Capstone Project

SPACEX

Chunyun Wang
22.12.2024

Outline

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



Executive Summary

The project attempts to predict SpaceX Falcon 9 first stage landing. This information can be used to determine the cost of a launch. The methodologies used for data collection, data wrangling, exploratory data analysis (EDA) and summary of all results are described as follows.

Summary of methodologies

- Data collection (SpaceX API & web scraping)
- Data wrangling (create success/fail outcome variable)
- EDA
  - data visualization with python and SQL
  - building an interactive map with Folium
  - building a dashboard with Plotly Dash
- Predictive analysis (Classification)
  - logistic regression
  - support vector machine (SVM)
  - decision tree
  - K-nearest neighbor (KNN)

Summary of all results

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

Introduction

Project 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.

Objective

Predict whether SpaceX Falcon 9 first stage will successfully land.

Key Questions

- What factors influence landing success?
- How accurate can predictions be using historical data?

Methodology

Methodology

- **Data collection**

  using SpaceX API request and web scraping

- **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**

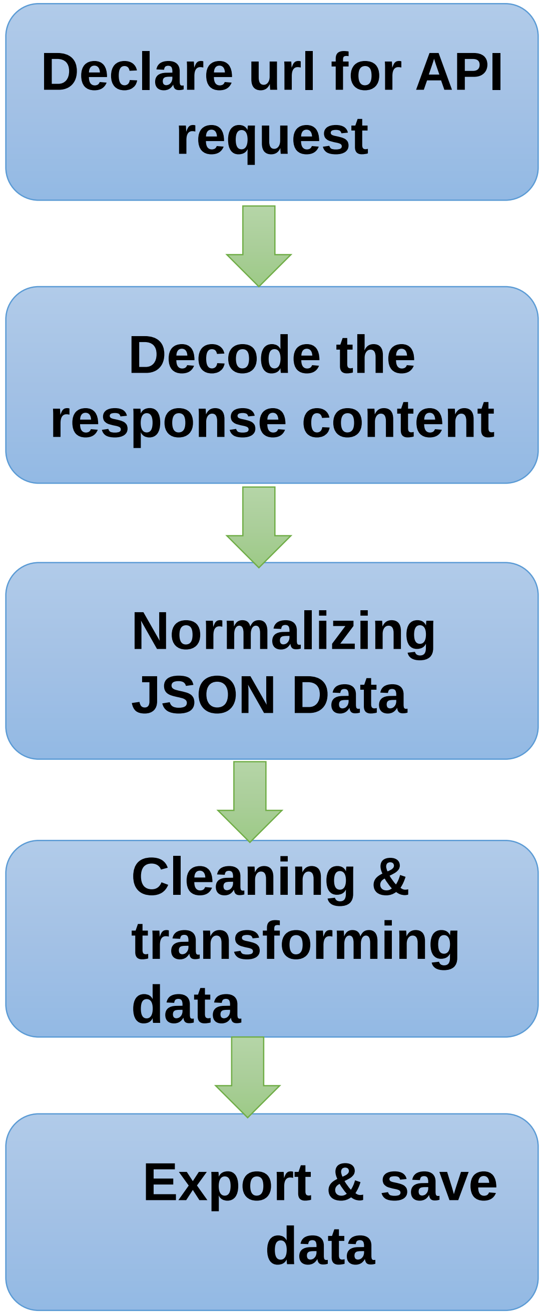
  to predict landing outcomes using classification models, tune and evaluate models to find best model and parameters

Data Collection – SpaceX API

Key Steps:

- **API request**
  use requests.get() to fetch JSON data
- **Decode the response content**
  use .json() to parse the API response
- **Normalizing JSON Data**
  use json_normalize from Pandas to convert structured JSON data into a flat table format suitable for analysis
- **Cleaning and transforming the data**
  - ✓ data wrangling with API
  - ✓ Create data frame
  - ✓ filtering
  - ✓ handling missing data
- **Export and save data**

GitHub URL of the completed SpaceX API calls notebook



Declare url for API request

Decode the response content

Normalizing JSON Data

Cleaning & transforming data

Export & save data

Data Collection – Web Scraping

Key Steps

- **Fetch the webpage**

  Use requests.get() request HTML content from Wikipedia to get Falcon 9 launch data

- **Parse the html**

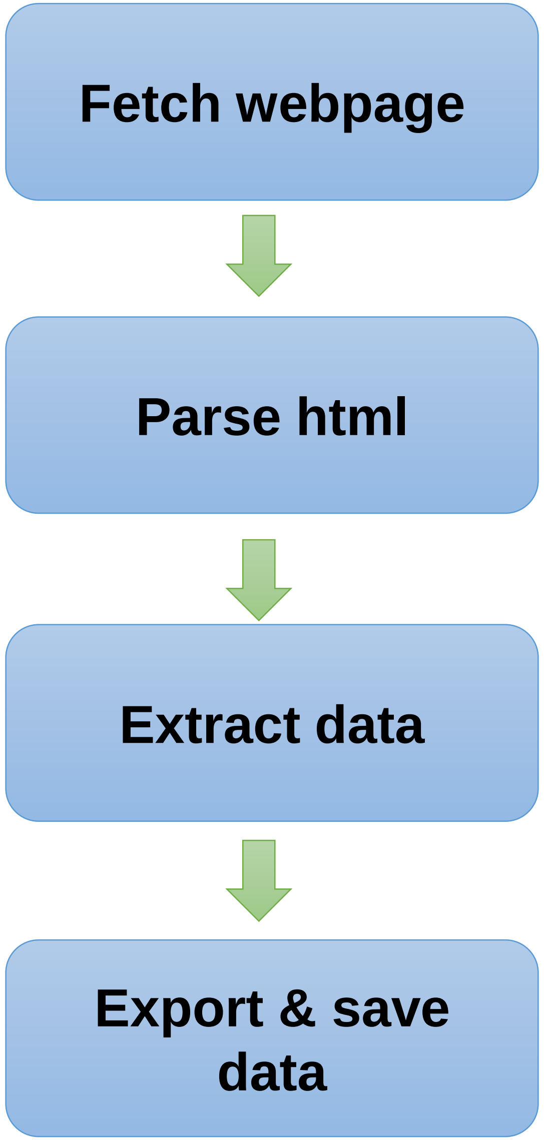
  ✓ Create BeautifulSoup object from a response text content to initialize the Parser

  ✓ Use find_all() method to extract all column/variable names from the HTML table header

- **Extract data**

  Iterate through elements to extract data and create a data frame by parsing the launch html tables

- **Export and save data to a csv file**



GitHub URL of the completed web scraping notebook

Data Wrangling

Steps

- **Perform EDA**
- **Calculate**
  - ✓ Number of launches on each site
  - ✓ Number and occurrence of each orbit
  - ✓ Number and occurrence of mission outcome of the orbits
- **Target variable conversion**
  - ✓ Convert landing outcome column into binary classes
- **Export data to csv file**

Outcome	**Description**	**Class**
True Ocean	Booster successfully landed in a specific region of the ocean	1 (Good Outcome)
False Ocean	Booster failed to land in the designated ocean region	0 (Bad Outcome)
True RTLS	Booster successfully landed on a ground pad	1 (Good Outcome)
False RTLS	Booster failed to land on a ground pad	0 (Bad Outcome)
True ASDS	Booster successfully landed on a drone ship	1 (Good Outcome)
False ASDS	Booster failed to land on a drone ship	0 (Bad Outcome)
None ASDS	No landing attempt on drone ship or failed to land	0 (Bad Outcome)
None None	No landing attempt or failed to land altogether	0 (Bad Outcome)

GitHub URL of my completed data wrangling related notebooks

EDA with Data Visualization

Visualizations

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit Type vs Success Rate
- Flight Number vs Orbit Type
- Payload Mass vs Orbit
- Yearly Success Rate Trend

Insights

- **Scatter plots** show clearly **relationship** between variables**,** certain payload masses and orbit types have higher landing success rates. These relationships can be used in machine learning model.

- **Bar charts** show comparisons among discrete categories, which benefit the **feature engineering** for **machine learning** to predict success rate.

- **Line charts** show the success rate keeps increasing since 2013 till 2020 (time series).

GitHub URL of EDA with data visualization notebook

EDA with SQL

Performed SQL queries

Display

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

List

- Date of first successful landing on ground pad
- Names of boosters successfully landing on drone ship and have payload mass between 4,000 and 6,000
- Total number of successful and failed missions
- Names of booster versions which 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)

GitHub URL of completed EDA with SQL notebook

Build an Interactive Map with Folium

- **Markers Indicating Launch Sites**
  - ✓ Added marker with circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates.
  - ✓ Added markers with circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates.

- **Colored Markers of Launch Outcomes**

  Added colored markers with styled labels of successful (green) and unsuccessful (red) launches at each launch site, using Marker Cluster to identify which launch sites have relatively high success rates.

- **Distances Between a Launch Site to Proximities**

  Added colored lines to show distance between launch site CCAFS SLC 40 and its proximity to the nearest coastline, railway, highway and city.

GitHub of completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- **Dropdown List with Launch Sites**

  Allow user to select all launch sites or a specific launch site
- **Pie Chart Showing Successful Launches (All Sites/Certain Site)**

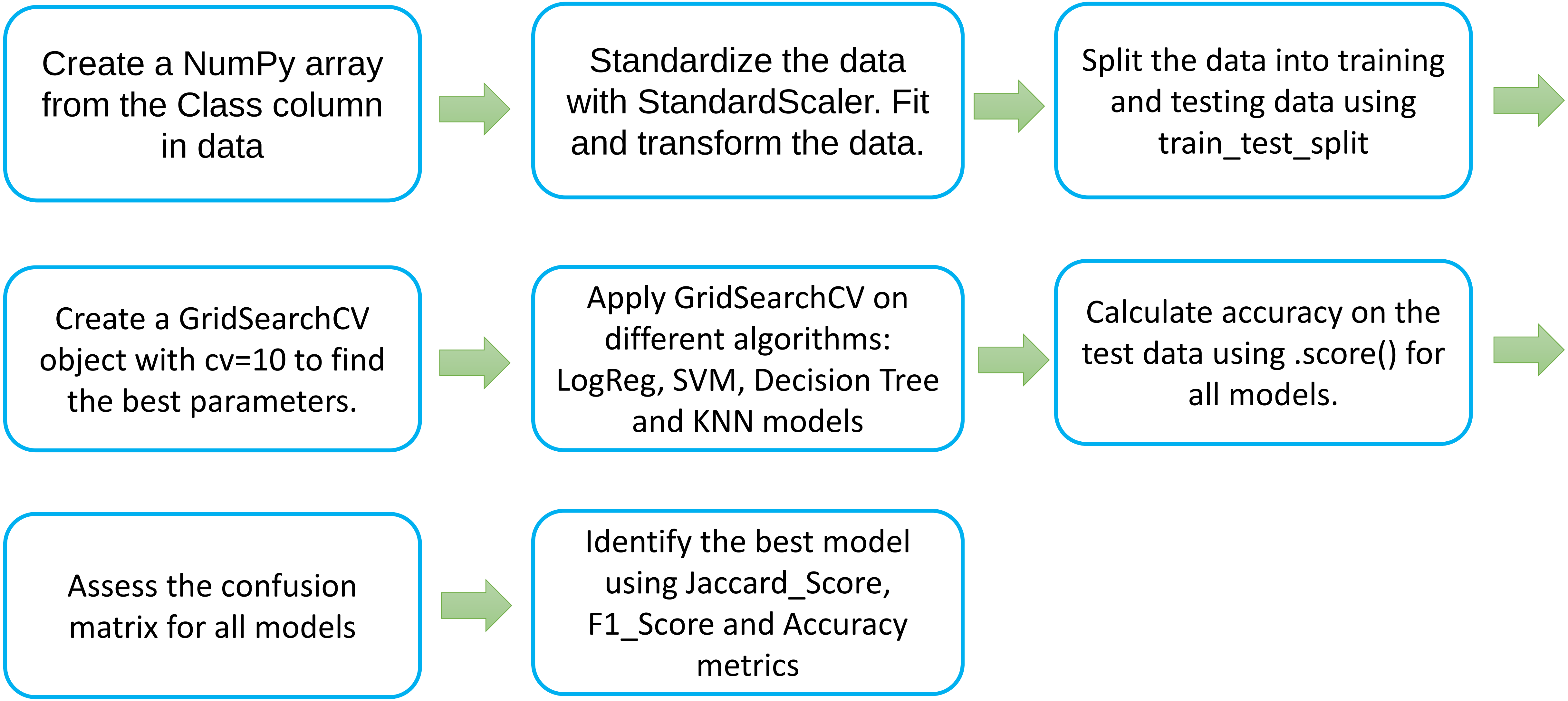
  show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- **Slider of Payload Mass Range**

  Allow user to select payload mass range
- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**

  Allow user to see the correlation between Payload and Launch Success

GitHub URL of completed Plotly Dash lab

Predictive Analysis (Classification)



Create a NumPy array from the Class column in data

Standardize the data with StandardScaler. Fit and transform the data.

Split the data into training and testing data using train_test_split

Create a GridSearchCV object with cv=10 to find the best parameters.

Apply GridSearchCV on different algorithms: LogReg, SVM, Decision Tree and KNN models

Calculate accuracy on the test data using .score() for all models.

Assess the confusion matrix for all models

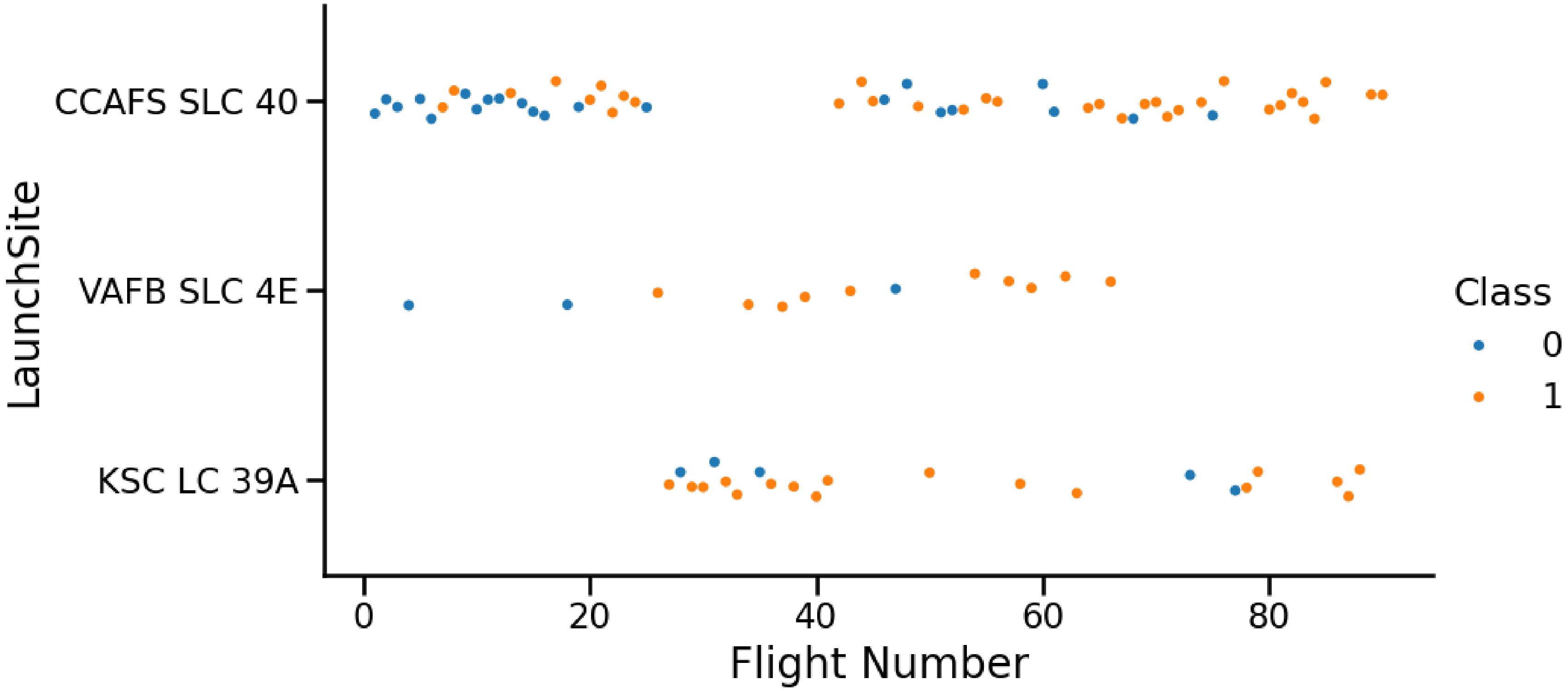
Identify the best model using Jaccard_Score, F1_Score and Accuracy metrics

GitHub URL of completed predictive analysis

Results

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

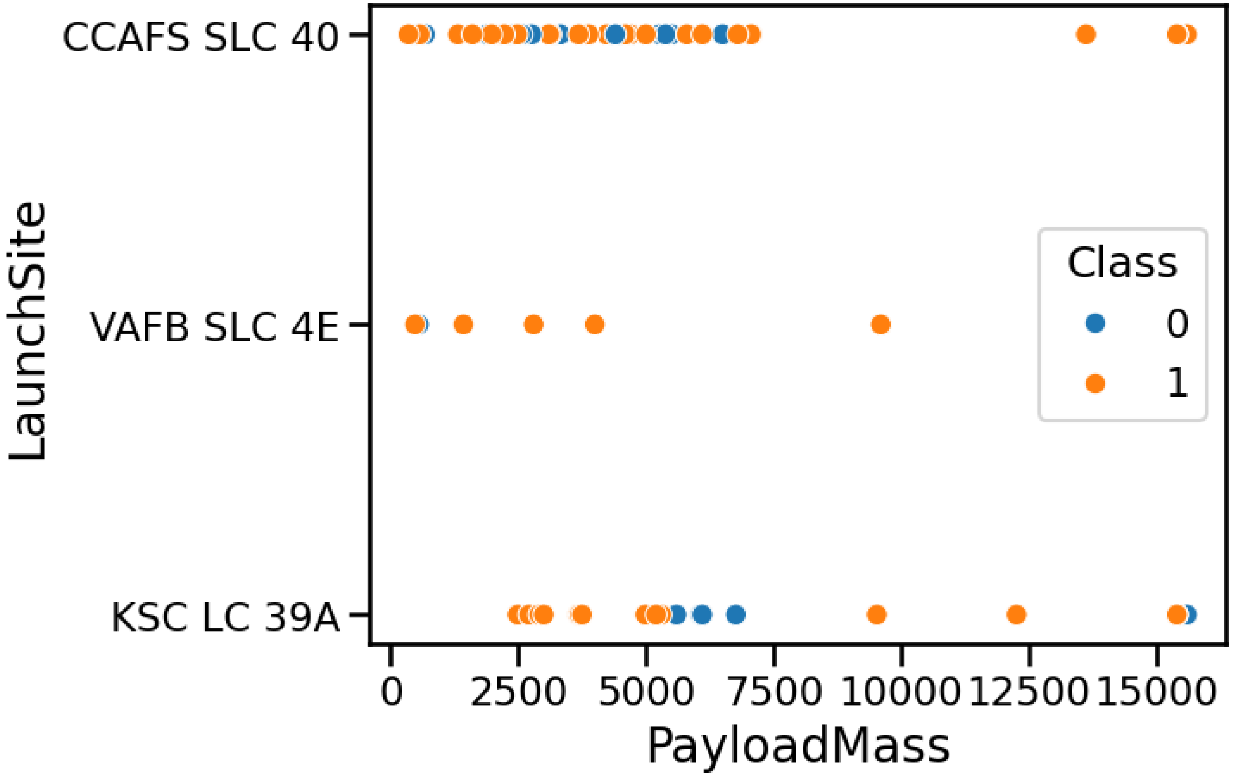
Flight Number vs. Launch Site



Explanation

- Earlier flights had a lower success rate and later flights had a higher success rate.
- CCAFS SLC 40 launch site has about a half of all launches, while VAFB SLC 4E and KSC LC 39A launch sites have higher success rates.
- New launch seems has a higher success rate.

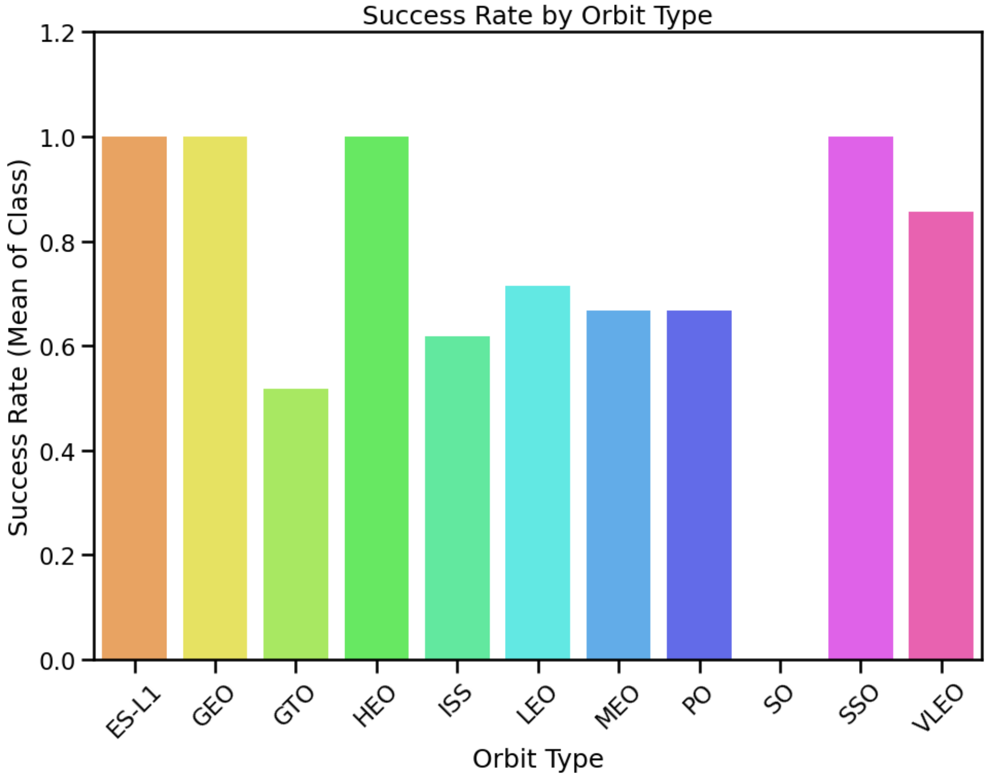
Payload vs. Launch Site



Explanation

- Obviously, the higher the payload mass, the higher the success rate.
- Most launces with a payload greater than 7,000 kg were successful.
- VAFB SLC 4E has a 100% success rate for launches when payload mass over 1,000 kg, while no over 10,000 Kg rockets launched at this site.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

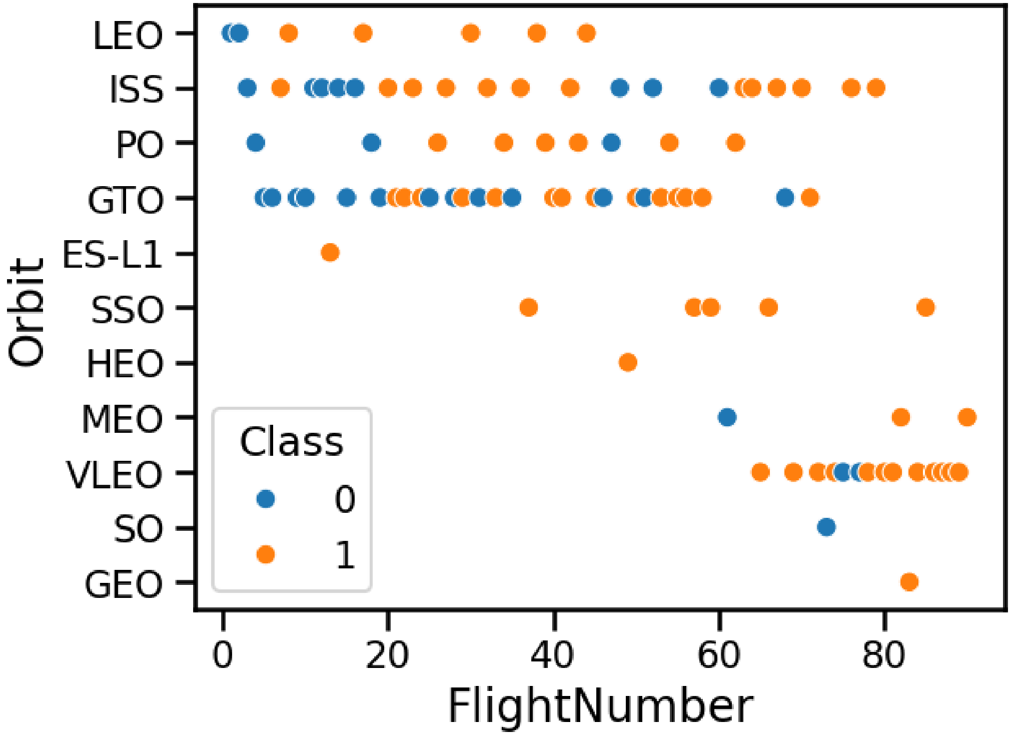
Success Rate vs. Orbit Type



Explanation

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

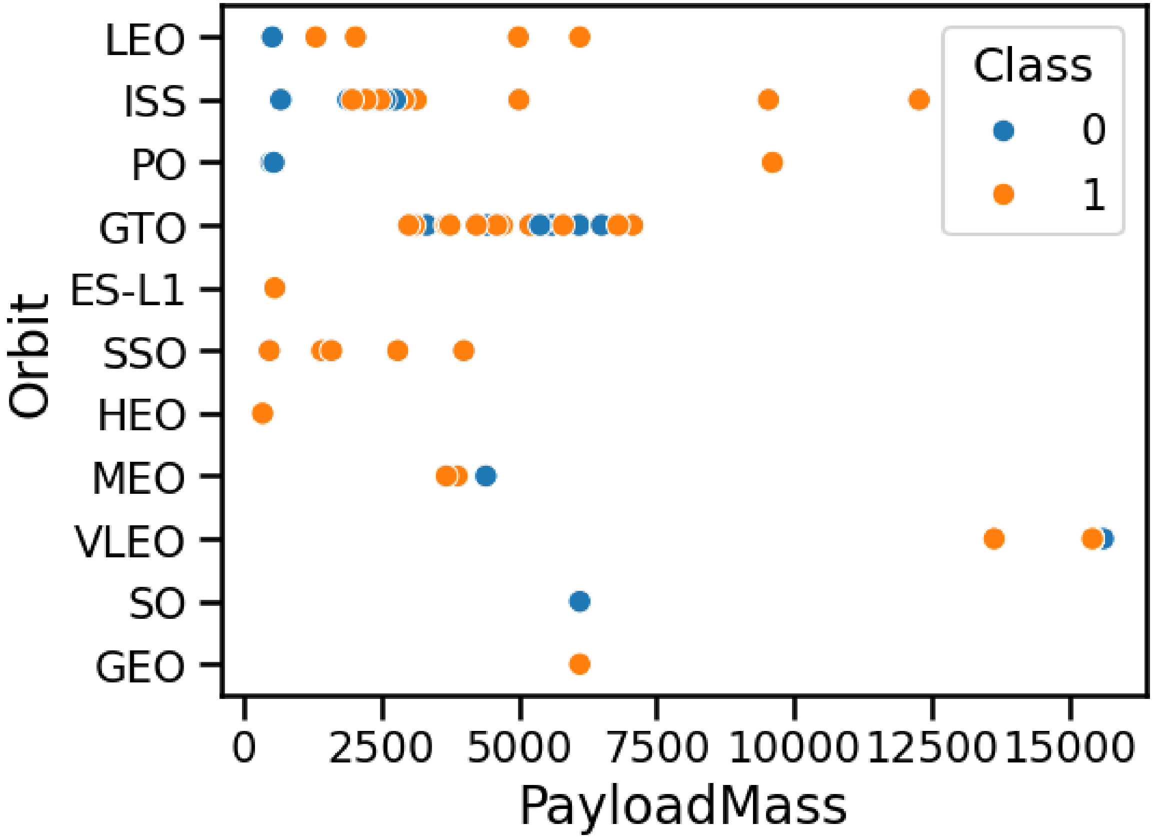
Flight Number vs. Orbit Type



Explanation

- In the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

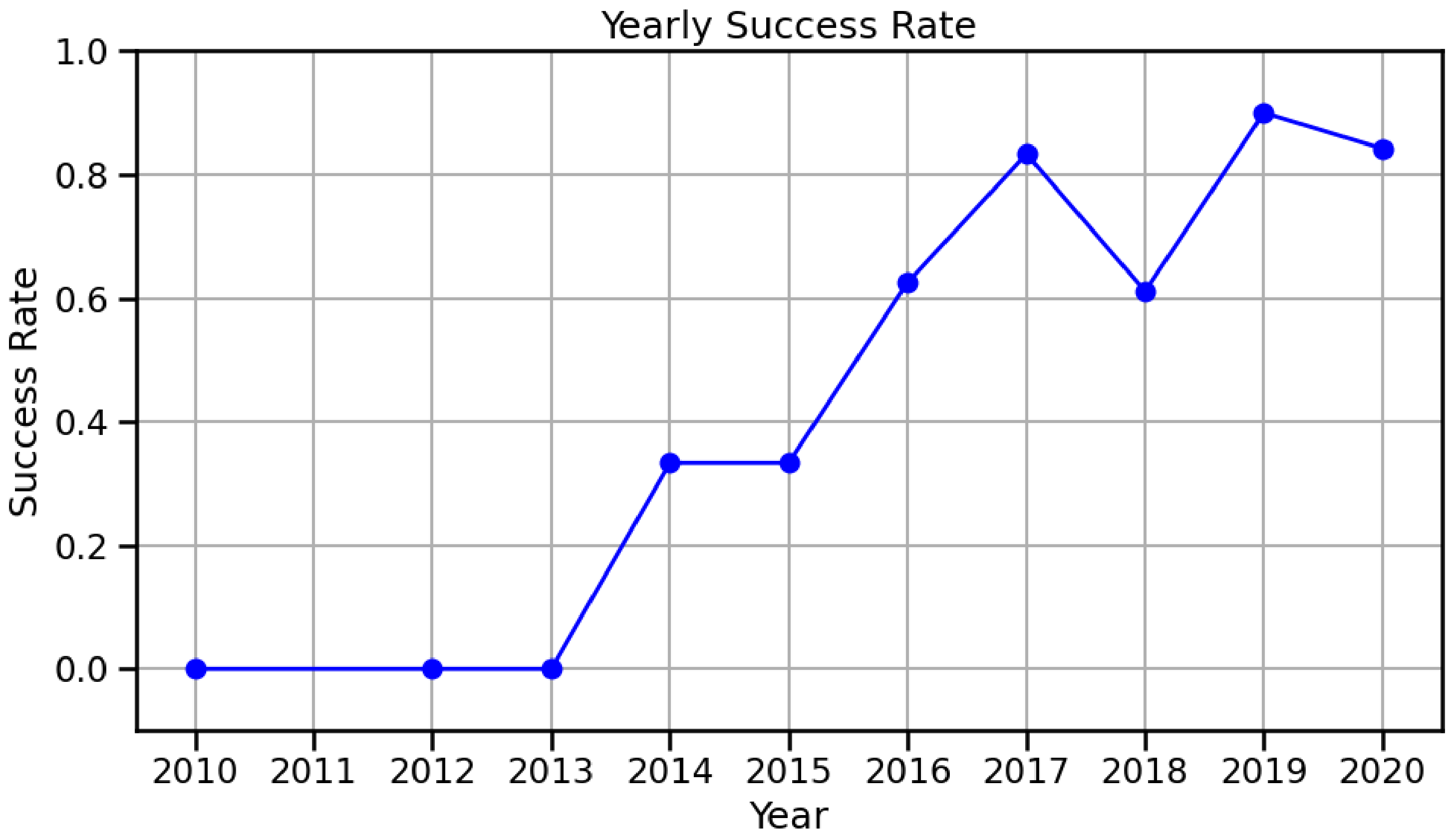
Payload vs. Orbit Type



Explanation

- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

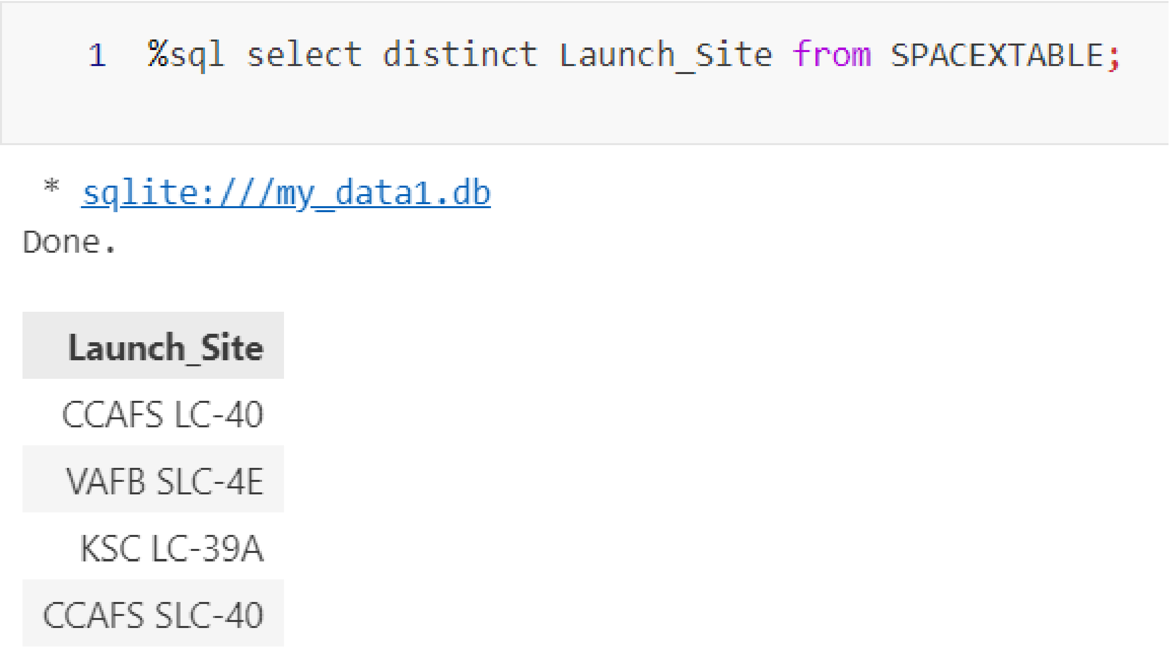
Launch Success Yearly Trend



Explanation

The success rate since 2013 kept increasing till 2020.

All Launch Site Names

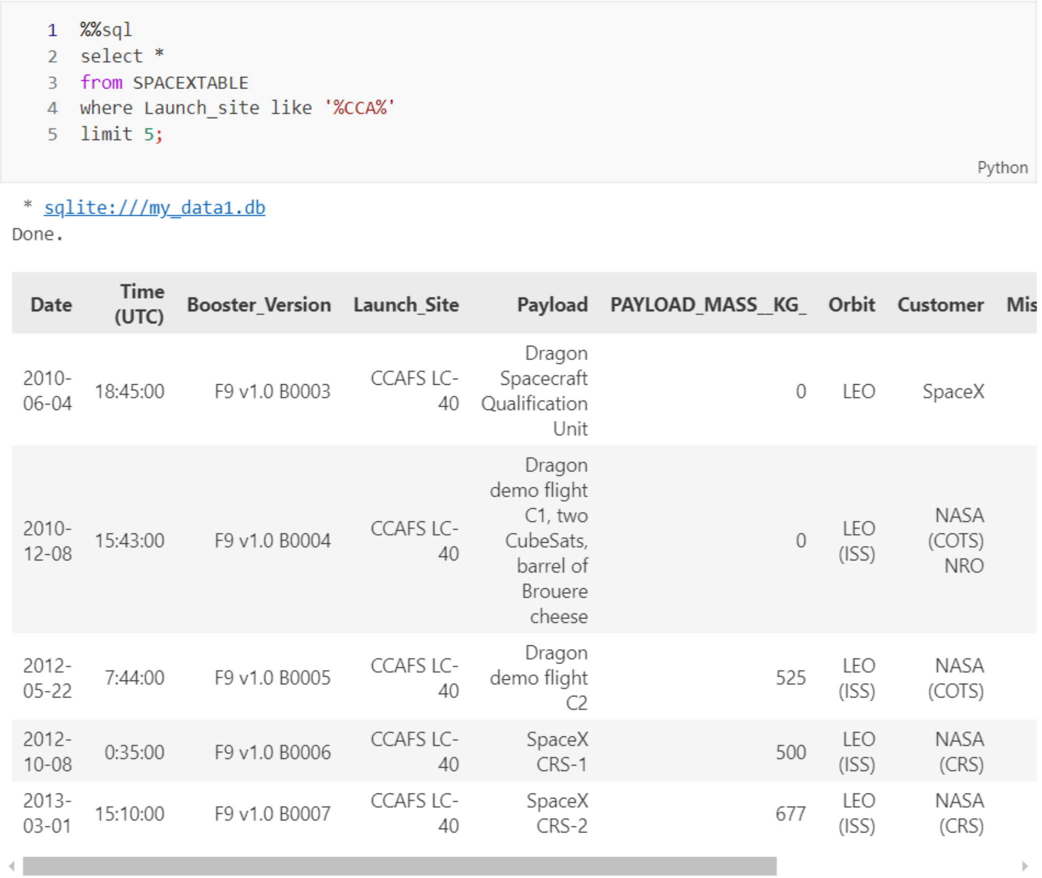


```
%sql select distinct Launch_Site from SPACEXTABLE;
```

```
 * 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'



```
%%sql
select *
from SPACEXTABLE
where Launch_site like '%CCA%'
limit 5;
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mis
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

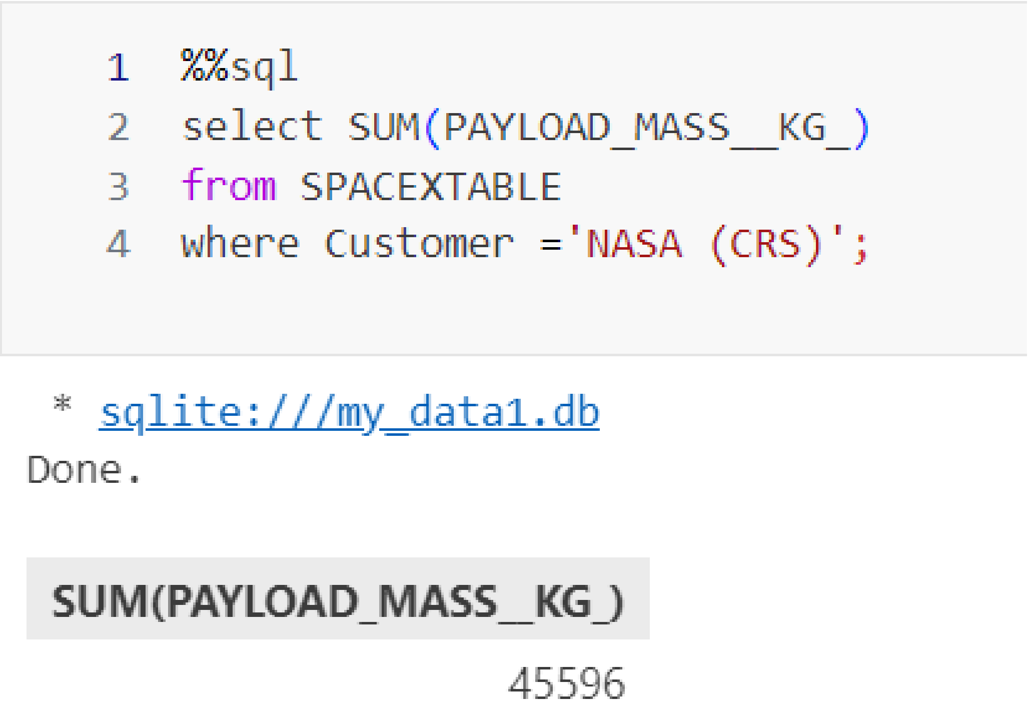
Explanation

Displaying the names of the unique launch sites in the space mission.

Explanation

Displaying 5 records where launch sites begin with the string 'CCA’. These 5 launches performed in LEO orbit, and four of them were from NASA.

Total Payload Mass



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

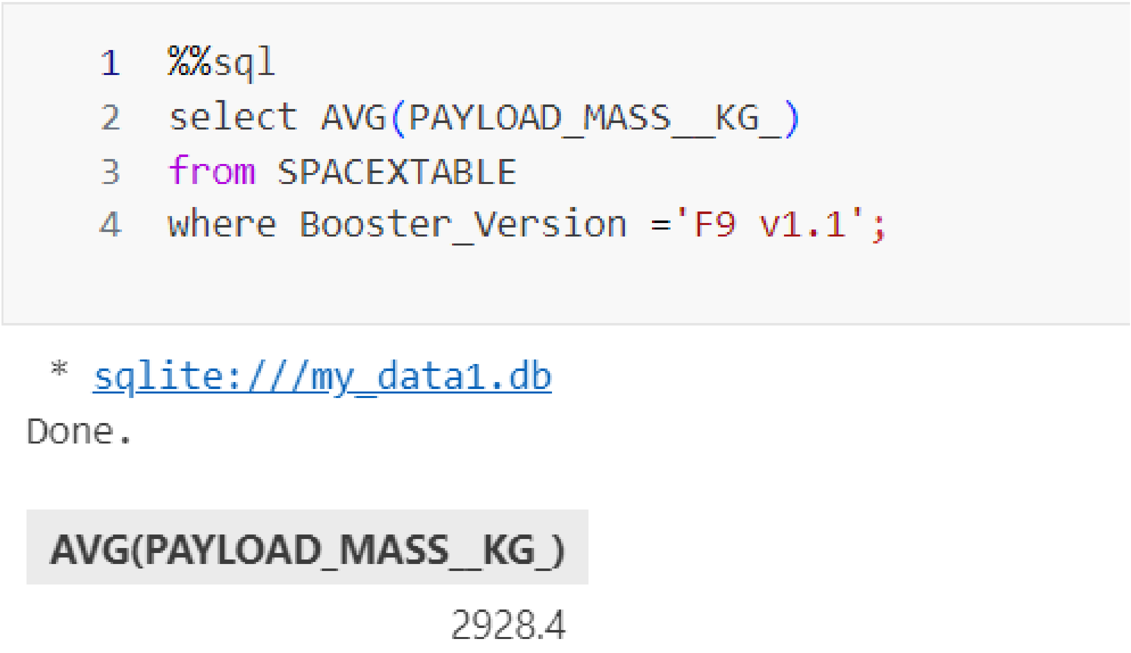
```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Explanation

The total payload mass carried by boosters launched by NASA (CRS) is 45596 Kg.

Average Payload Mass by F9 v1.1



```
%%sql
select AVG(PAYLOAD_MASS__KG_)
from SPACEXTABLE
where Booster_Version ='F9 v1.1';
```

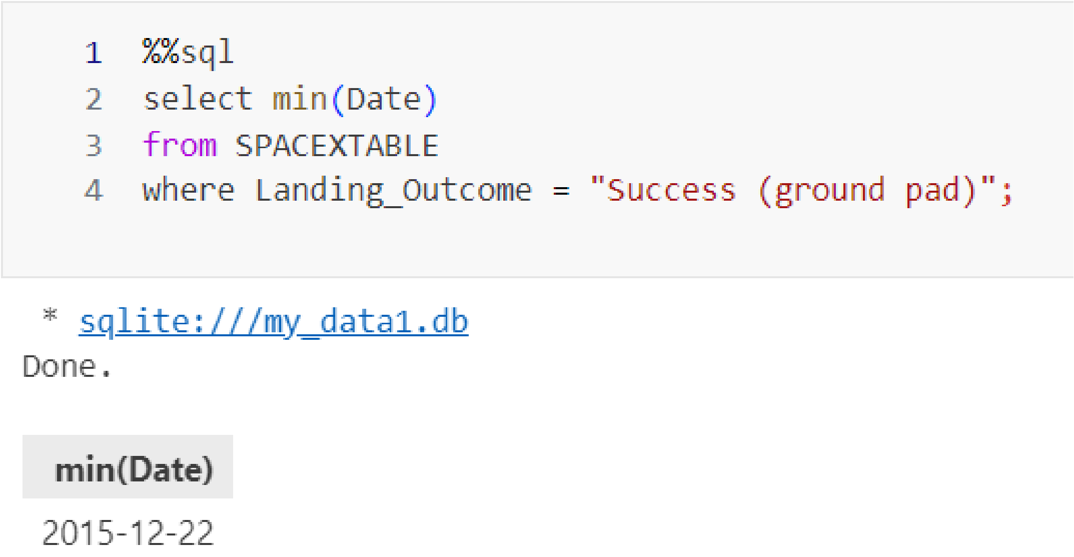
```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Explanation

The average payload mass carried by booster version F9 v1.1 is 2928.4 Kg.

First Successful Ground Landing Date

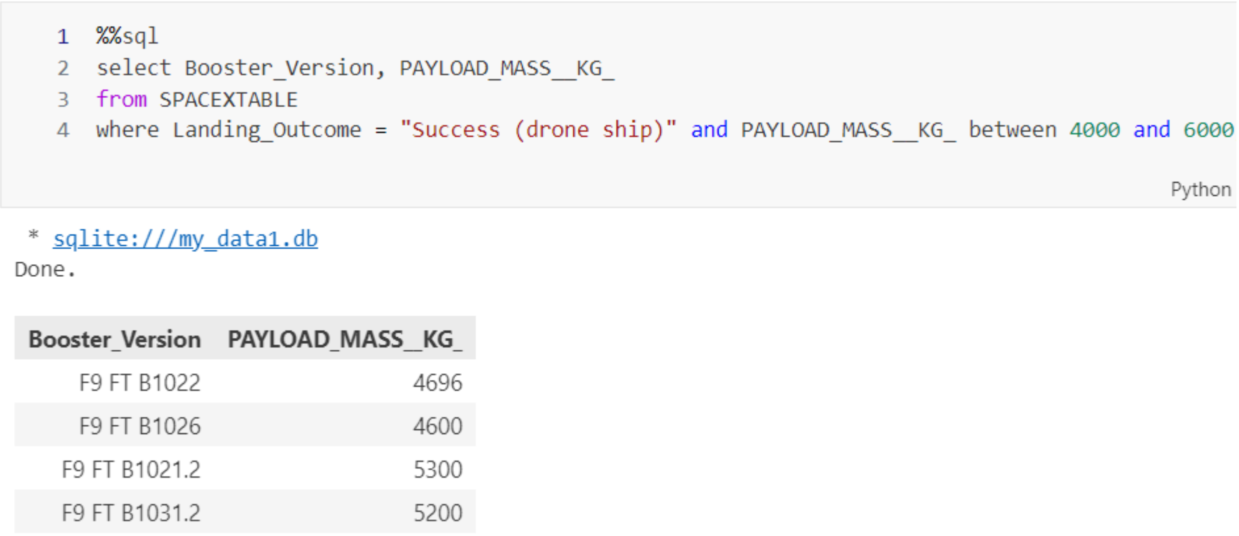


```
%%sql
select min(Date)
from SPACEXTABLE
where Landing_Outcome = "Success (ground pad)";
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTABLE
where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

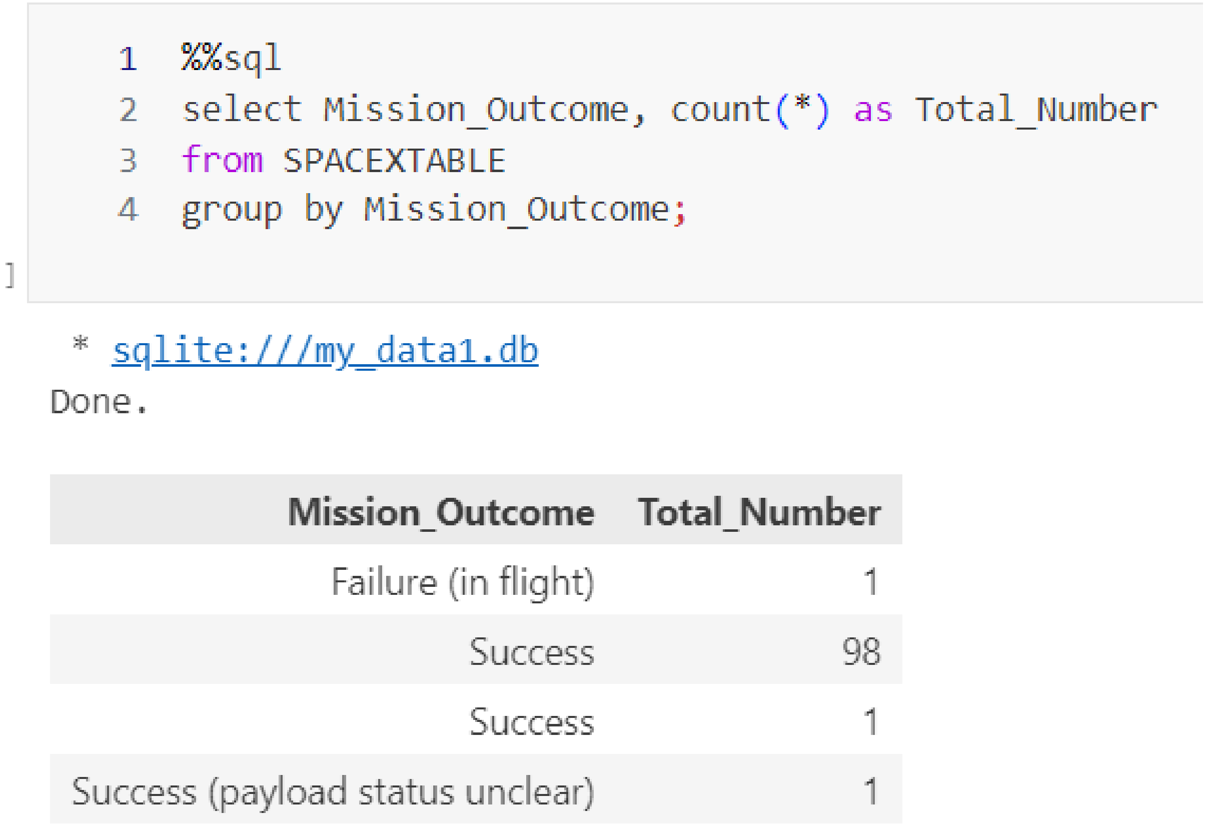
Explanation

On December 22, 2015, the first successful landing outcome in ground pad was achieved.

Explanation

The 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



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

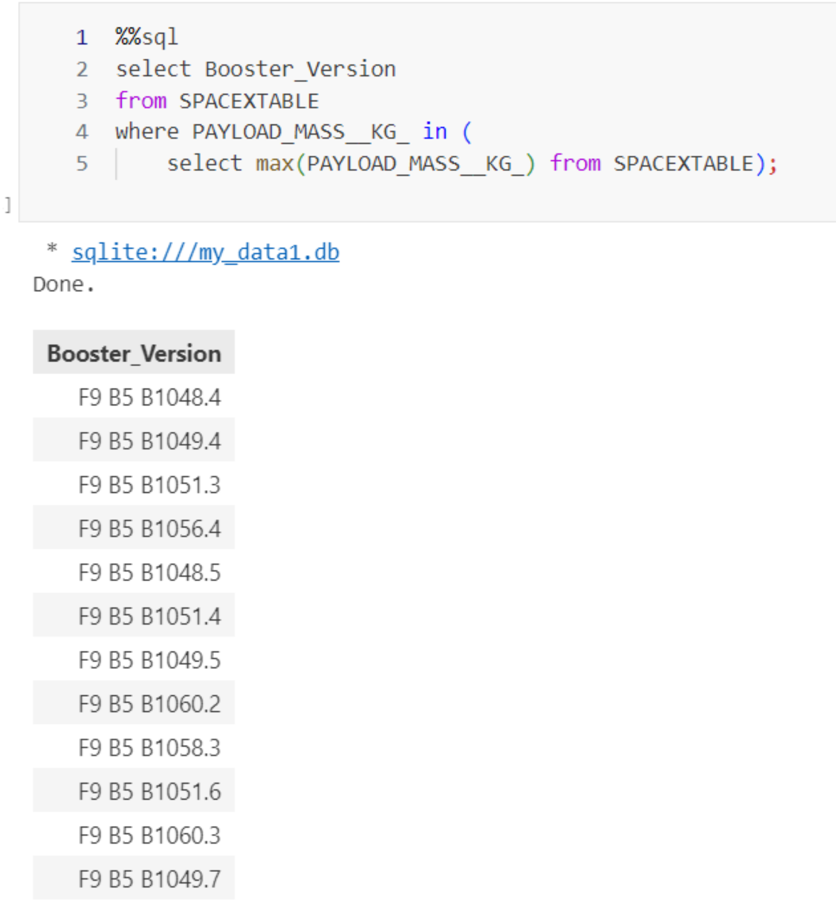
```
 * sqlite:///my_data1.db
Done.
```

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

Explanation

- The total number of successful mission outcomes is 100.
- The total number of failure mission outcomes is 1.

Boosters Carried Maximum Payload



```
%%sql
select Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_ in (
    select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

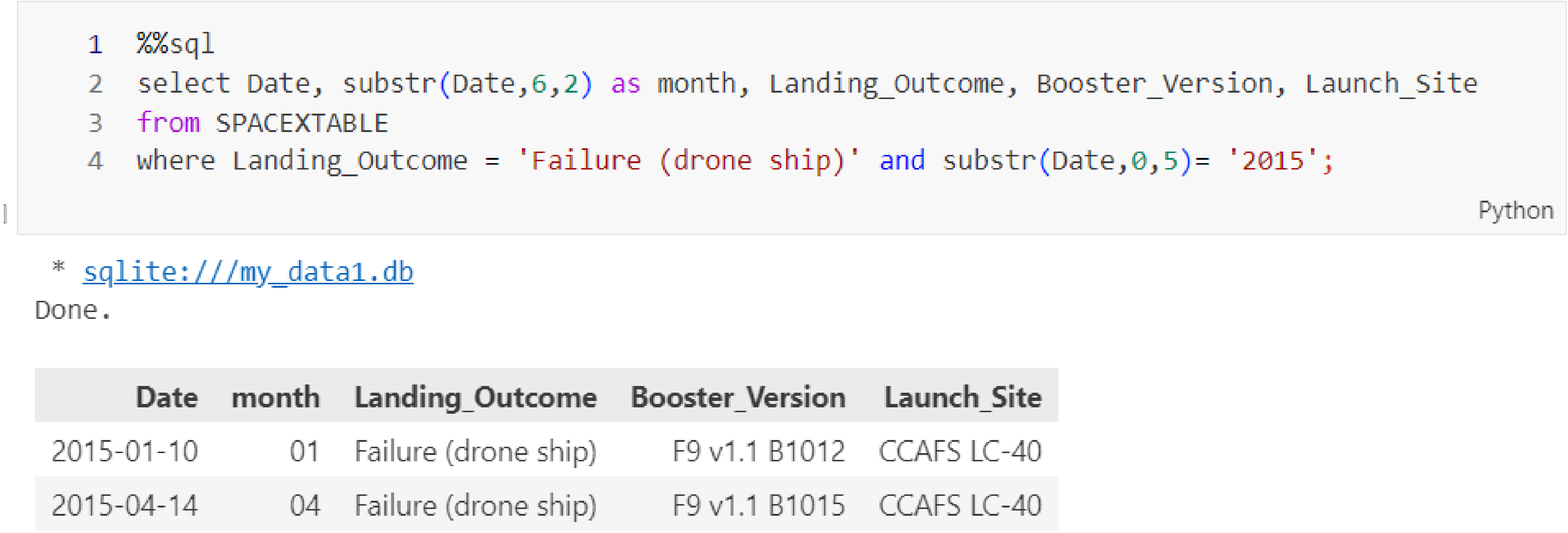
```
 * sqlite:///my_data1.db
Done.
```

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

Explanation

These boosters carry the maximum payload.

2015 Launch Records



```
%%sql
select Date, 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';
```

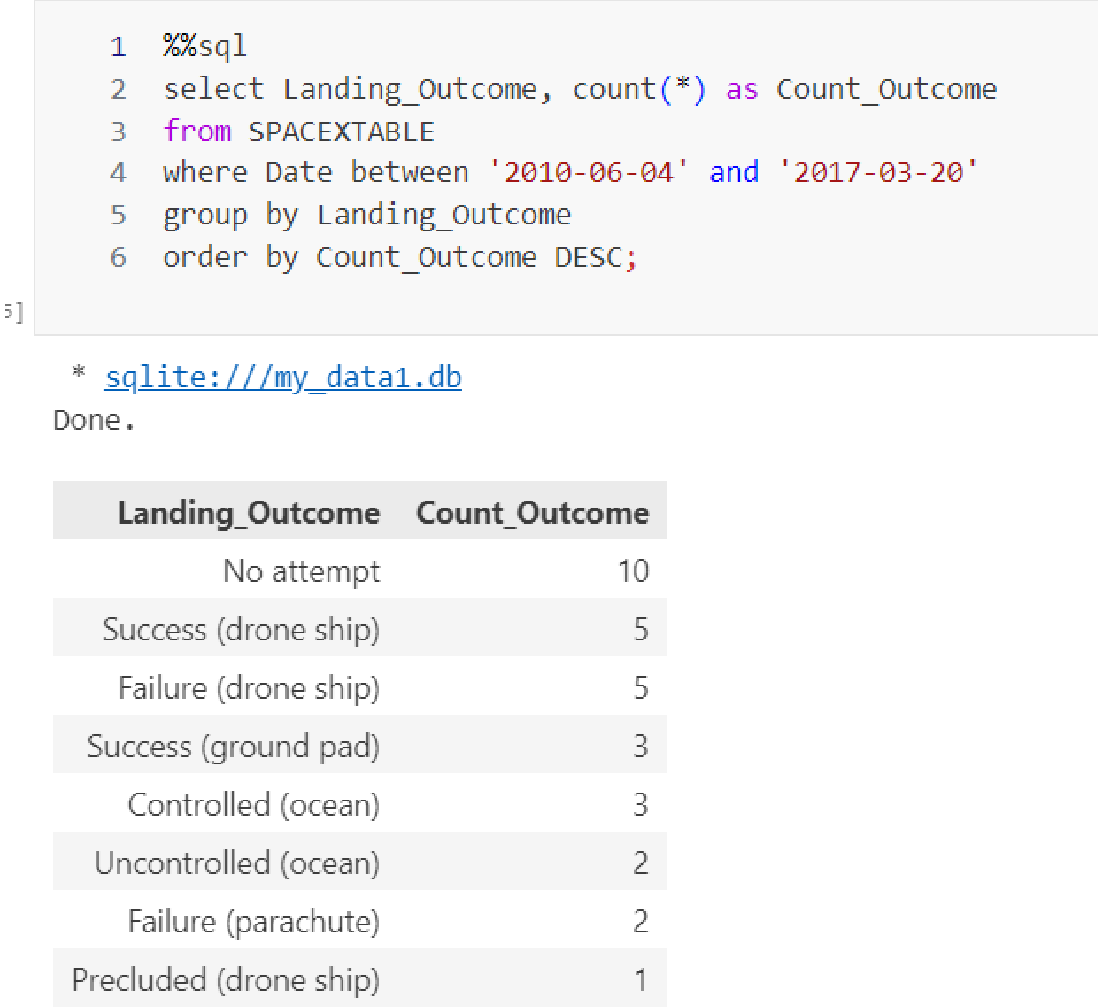
```
 * sqlite:///my_data1.db
Done.
```

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

Explanation

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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



```sql
%%sql
select Landing_Outcome, count(*) as Count_Outcome
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Count_Outcome DESC;
```

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

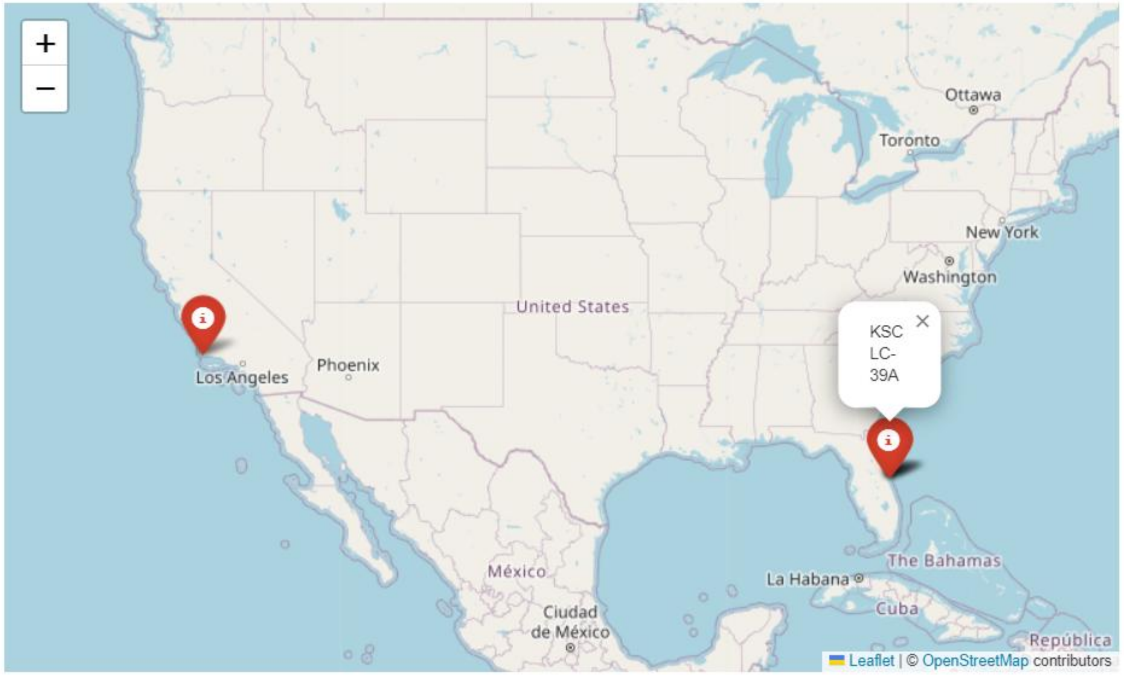
Landing_Outcome	Count_Outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanation

Ranking 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.

Location of all launch sites on a map

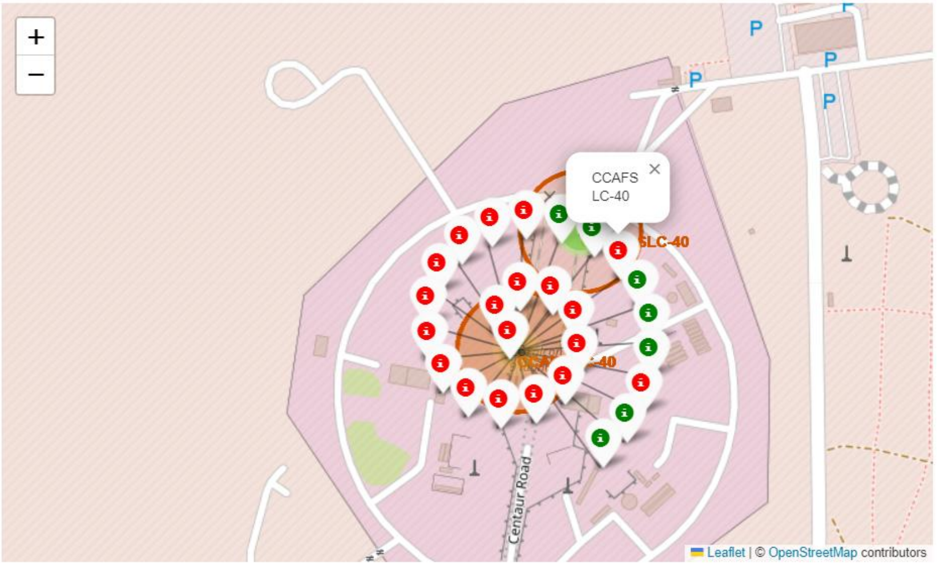
All launch sites are in proximity to the Equator line and are in very close proximity to the coast.

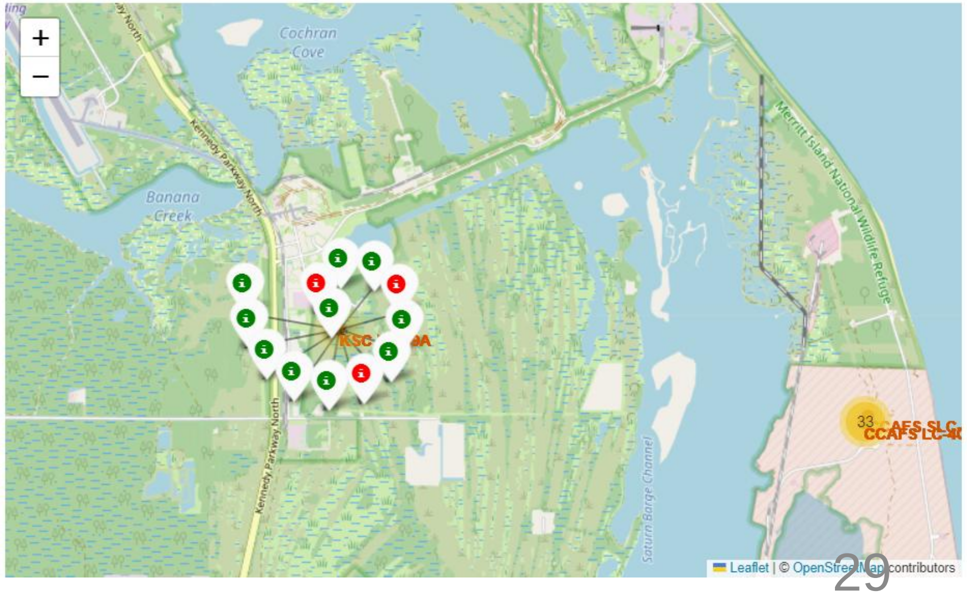


Explanation:

- Due to the rotational speed of earth, 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, which helps save the cost of putting in extra fuel and boosters.
- An overwater trajectory of rockets launch ensures that any debris or failed stages fall safely into the ocean, minimizing risks to populated areas.

Colour-labeled launch outcome on the map





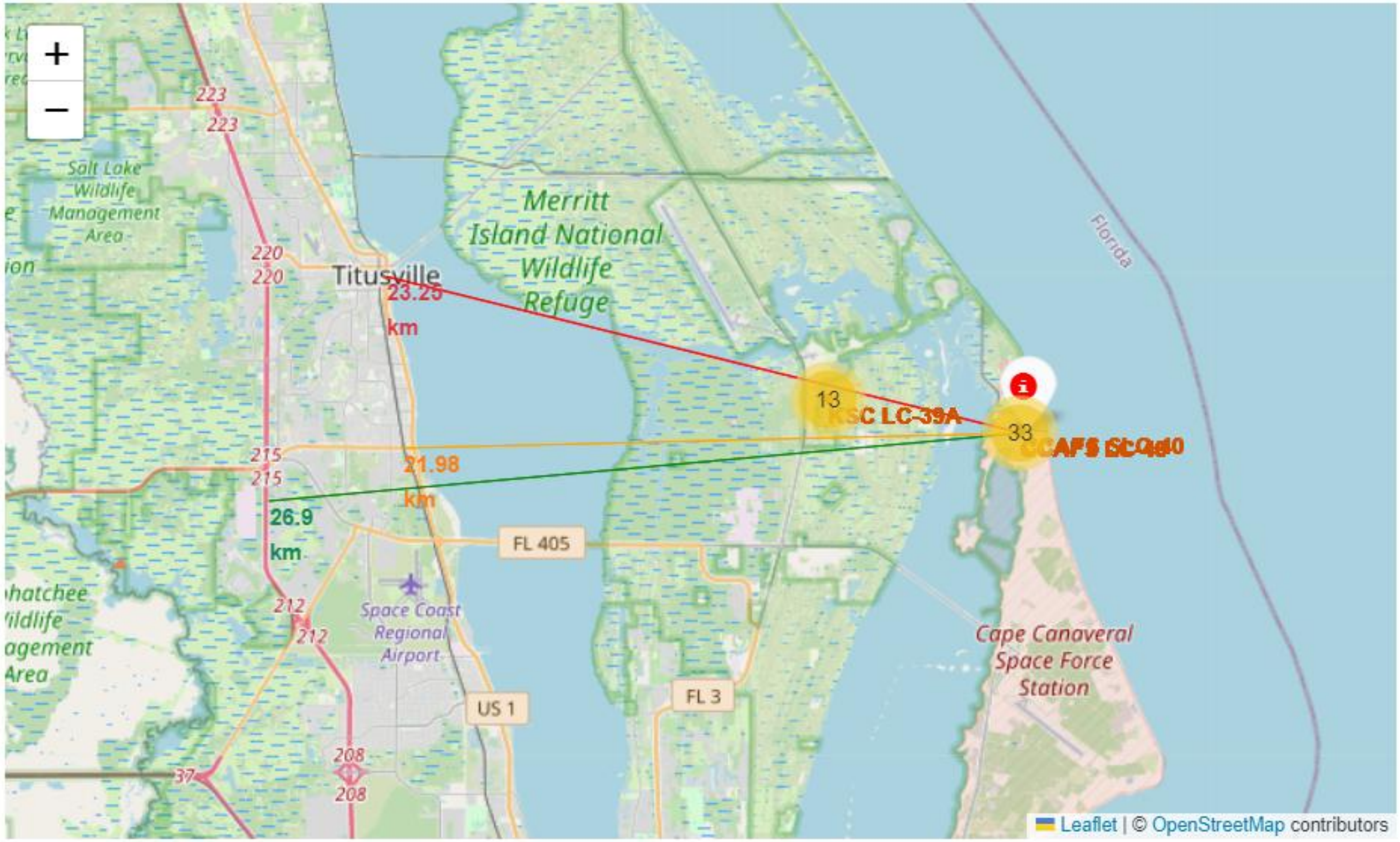
Explanation:

- The colour-labeled markers (green: success; red: fail) clearly show the relatively success rate at each launch site.
- Launch Site KSC LC-39A has the highest success rate.

Distance from the launch site CCAFS LC-40 to its proximities

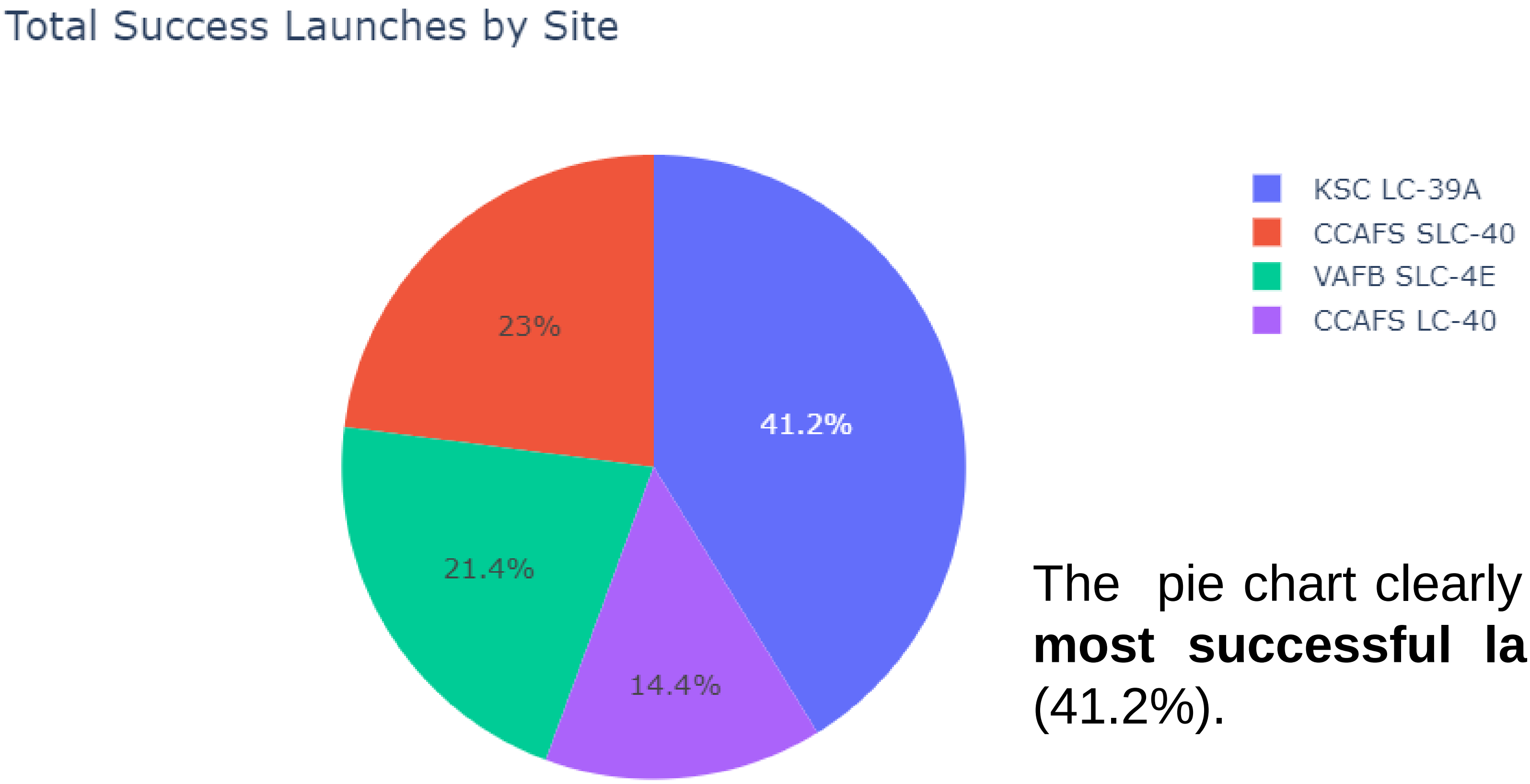
CCAFS LC-40

- 0.88 km from nearest coastline
- 21.98 km from nearest railway
- 23.25km from nearest city Titusville
- 26.90 km from nearest highway



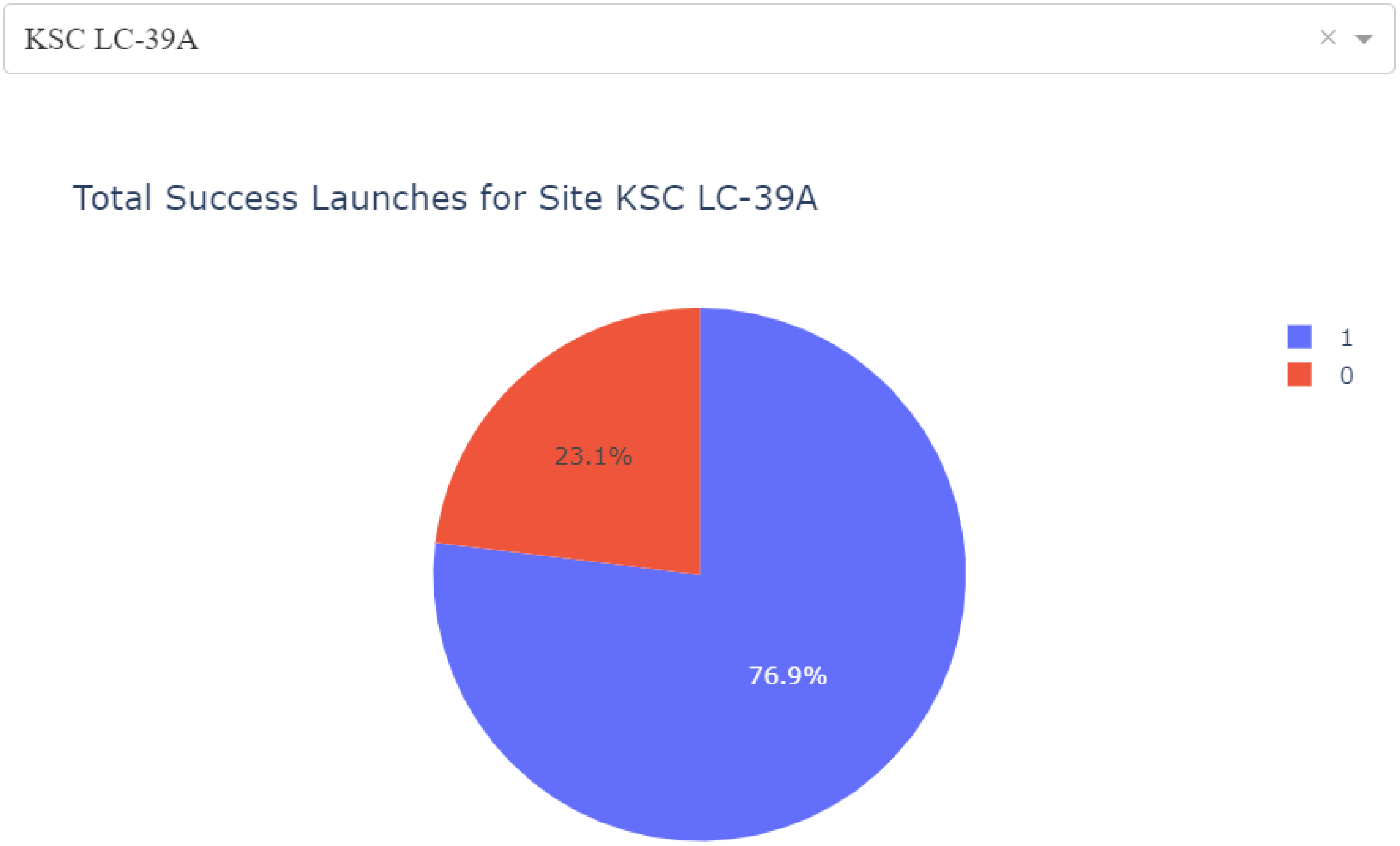
+
−
Salt Lake Wildlife Management Area
Titusville
23.25 km
Merritt Island National Wildlife Refuge
13
KSC LC-39A
33
CCAFS SLC-40
21.98 km
26.9 km
FL 405
Space Coast Regional Airport
US 1
FL 3
Cape Canaveral Space Force Station
Florida
223
220
215
212
208
37
Leaflet | © OpenStreetMap contributors

Total Success Launches rate by Site



The pie chart clearly shows that **KSC LC-39A** has the **most successful launches** among all launch sites (41.2%).

Success Launches for Site KSC LC-39A

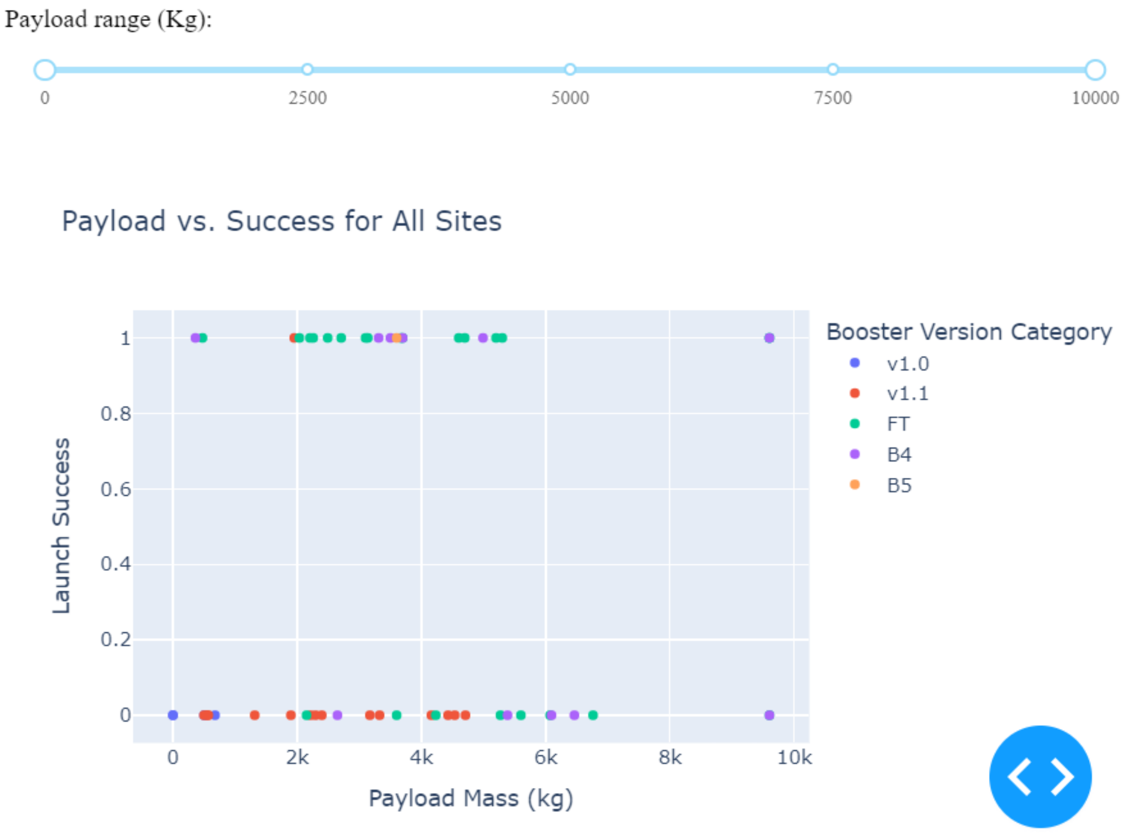


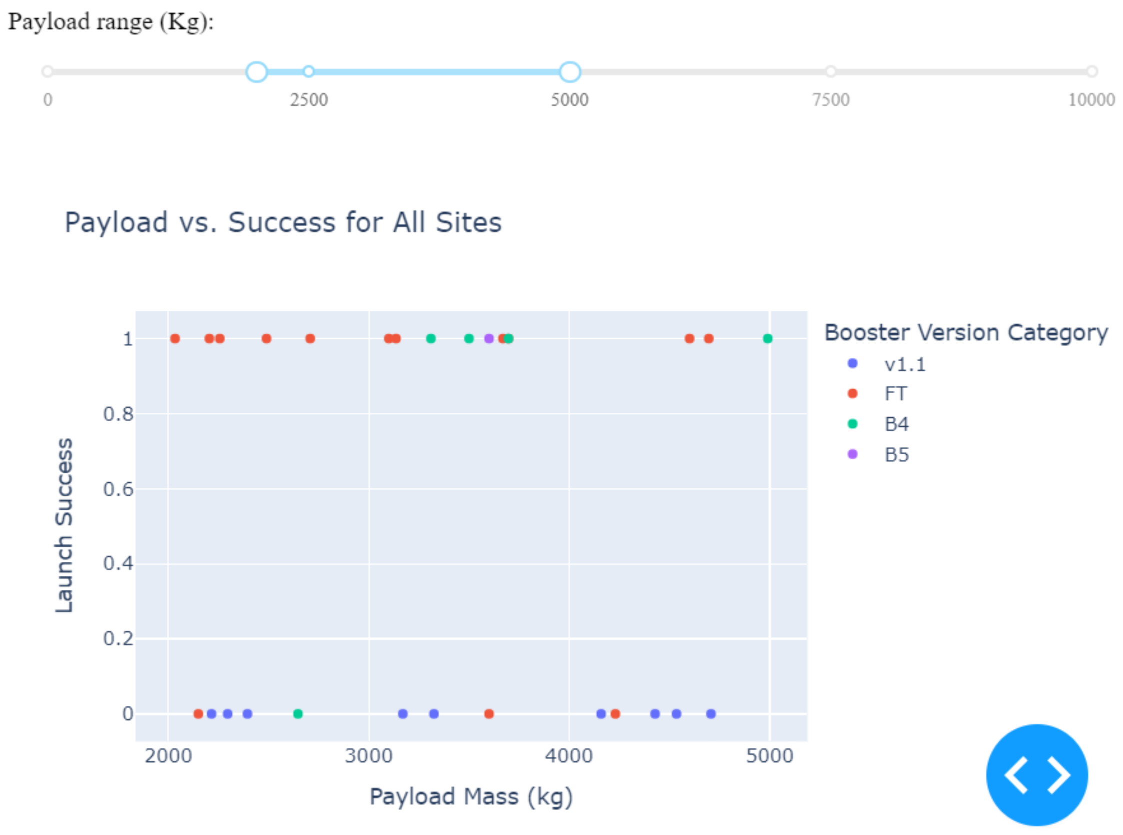
KSC LC-39A

Total Success Launches for Site KSC LC-39A

KSC LC-39A has the **highest launch success rate (76.9%)** with 10 successful and only 3 failed landings.

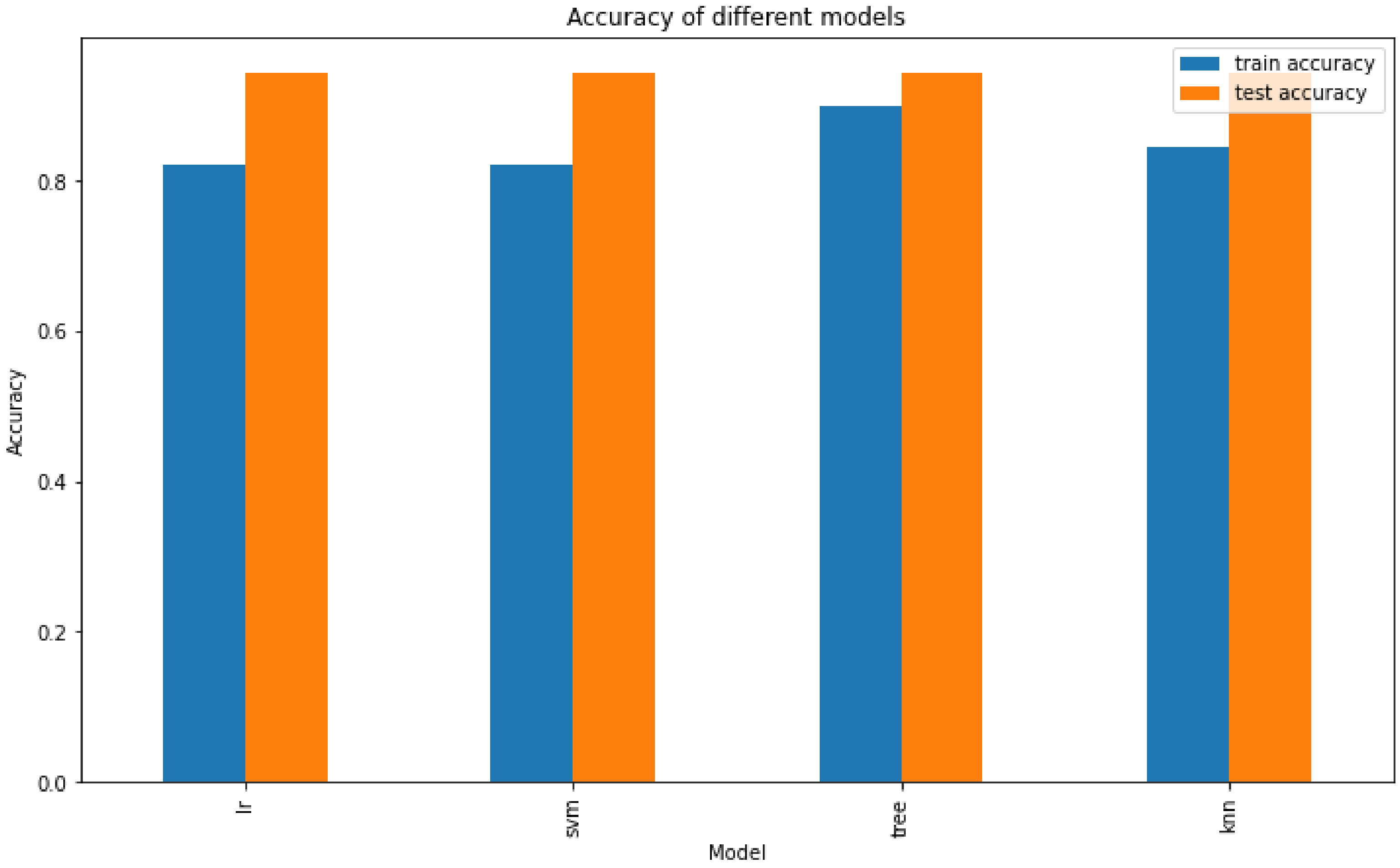
Payload Mass vs. Success for all sites by Booster Version





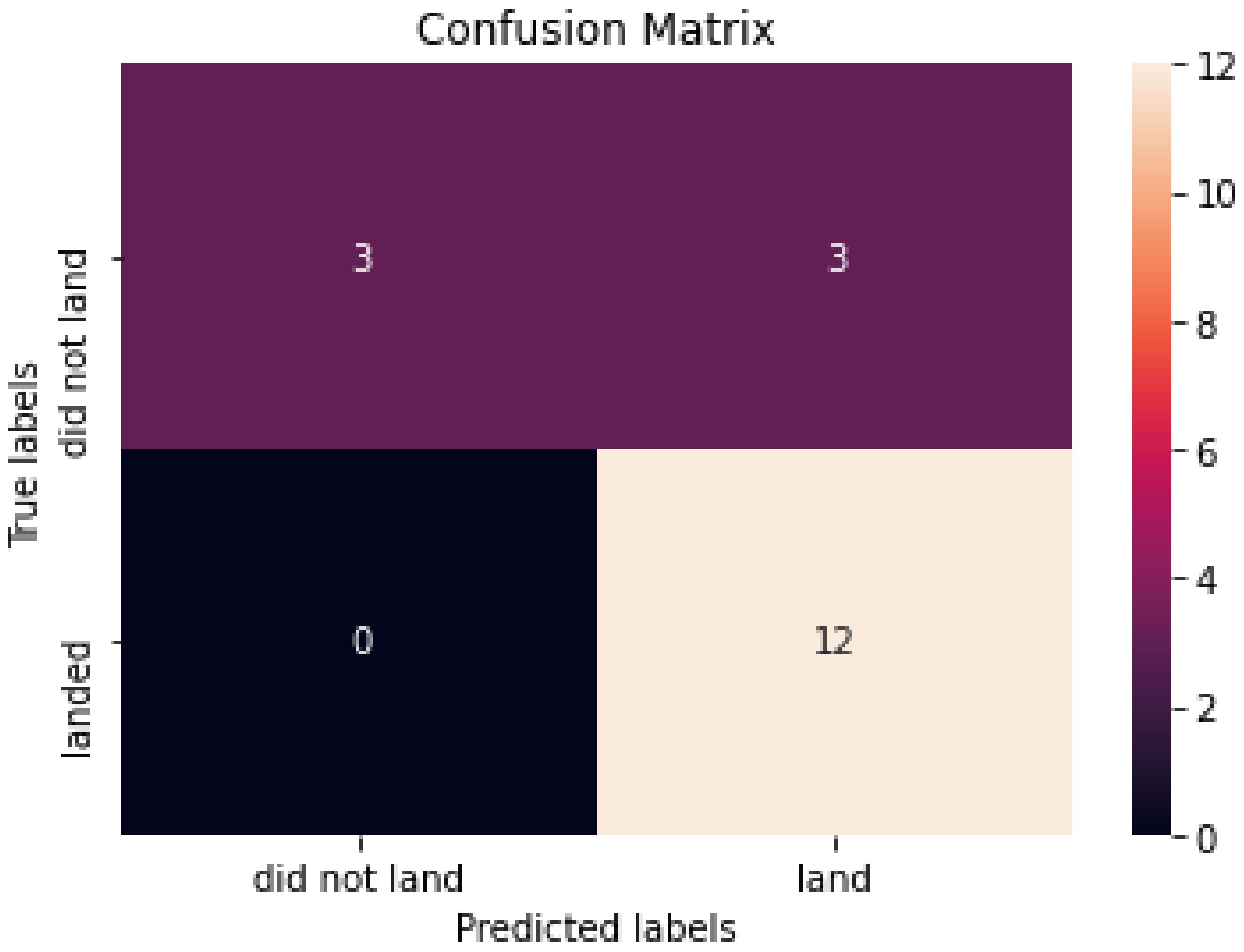
- Payloads **between 2,000 kg and 5,000 kg** have the **highest success rate**.
- **FT booster** has **the highest success rate**.

Classification Accuracy



Decision Tree model has the highest classification accuracy (train accuracy: 0.90, test accuracy: 0.94).

Confusion Matrix



- Decision Tree model can distinguish between the different classes.
- The major problem is false positives (Type 1 error).

Conclusions

- Decision Tree Model performs the best for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Across all launch sites, the higher the payload mass, the higher the success rate.
- A larger dataset will be helpful to build on the predictive analytics results.

Appendix

My GitHub repository for this assignment

Special Thanks to:

Instructors

Coursera

IBM

Thank you!