IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sumet (Ben) Viwatmanitsakul

December 23, 2023

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

- Project objective: Predict the successful landing of the Falcon 9 rocket's first stage for estimating launch costs. Data was collected from SpaceX REST API & and Wikipedia. Binary training labels visualized using Folium & Plotly Dash. Classification models are used for forecasting and confusion Matrix is used for accuracy determination. Methodology integrates data collection, wrangling, exploratory analysis, interactive visualization & and predictive modeling.
- The Decision Tree Model had the highest accuracy score of 0.888. The number of flights doesn't have a direct correlation with the success rate. Launch sites and payload mass relationship remains unclear, particularly at CCAFS SLC 40.
- KSC's LC-39A has the highest success rate at 41.7%, with an impressive success rate ratio of 76.9%. In contrast, CCAFS SLC-40 has the lowest success rate at 12.5%, with a 57.1% success ratio. These insights show the varying success rates at different launch sites, highlighting the importance of site-specific factors in the overall success of Falcon 9 first stage landings.

Introduction

SpaceX offers Falcon 9 rocket launches on its website for $62 million, whereas other providers charge over $165 million, largely due to SpaceX's ability to reuse the first stage.

The project aims to predict the successful landing of Falcon 9's first stage, which would determine the cost of the launch.

Section 1

Methodology

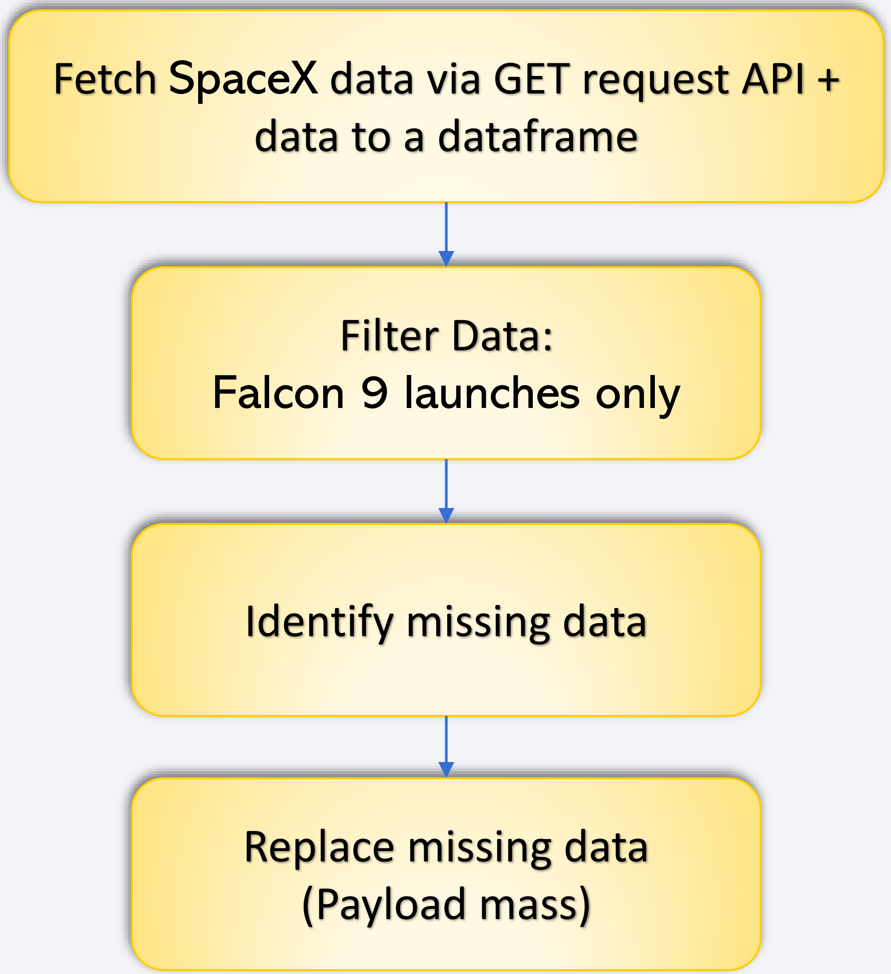
Methodology

Executive Summary

- Data collection methodology:
  - Gathered data through the SpaceX REST API and Web Scraping Falcon 9 and Falcon Heavy Launches Records - Wikipedia (see appendix)
- Perform data wrangling
  - Convert landing outcomes to training labels: 1 for success, 0 for failure.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data, Split into training/testing sets, Train model, Find best hyperparameters with Grid Search, Evaluate with Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors, and Output the Confusion Matrix to determine best accuracy

Data Collection

- Request and parse the SpaceX launch data, including Booster Name, Payload and orbit, Launch sites, Outcome and other Info, using the GET request API and apply the collected data to a dataframe.
- Filter data to only include Falcon 9 launches.
- Identify the missing data from each column.
- Replace the missing Payload mass values with the mean value.



Fetch SpaceX data via GET request API + data to a dataframe

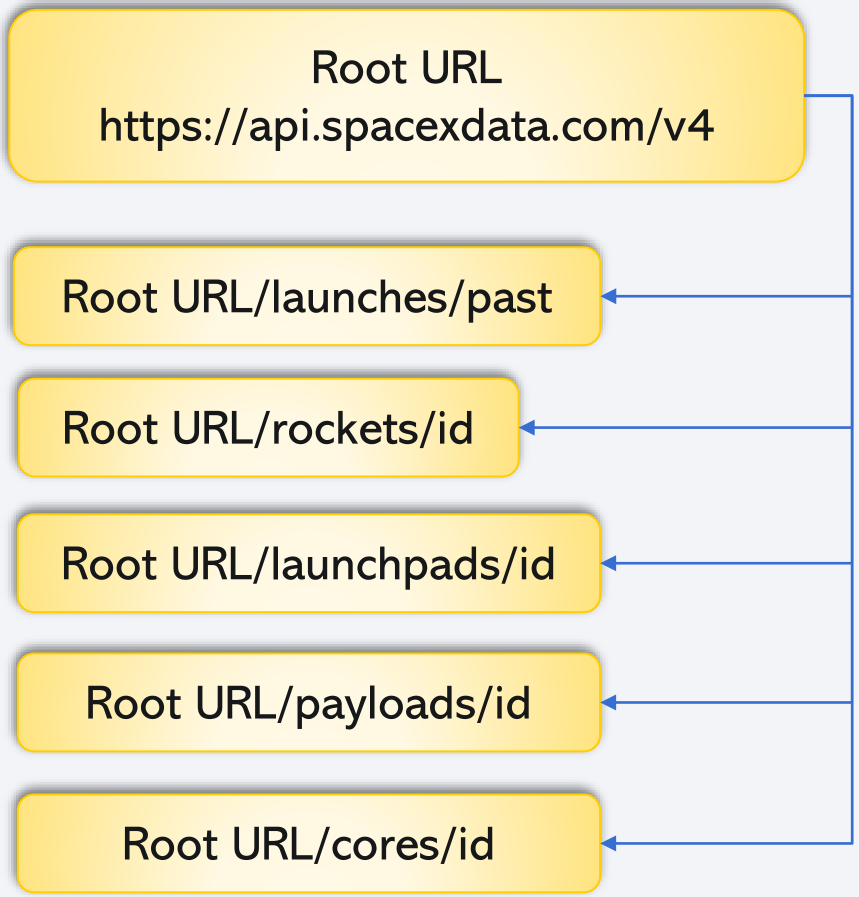
Filter Data:
Falcon 9 launches only

Identify missing data

Replace missing data (Payload mass)

Data Collection – SpaceX API

- Request the launch data for SpaceX, call the GET request API endpoints: *https://api.spacexdata.com/v4/launches/past*.
- Request the Booster version data, use the rocket col and call the GET request API endpoints : (*https://api.spacexdata.com/v4/rockets/[id]*).json().
- Request the Launch Site data, use the launchpad col and call the GET request API endpoints: (*https://api.spacexdata.com/v4/launchpads/[id]*).json().
- Request the Payload + Orbit data, use the payloads col and call the GET request API endpoints: (*https://api.spacexdata.com/v4/payloads/[id]*).json()
- Request the Cores data, using the cores col and call the GET request API endpoints: (https://api.spacexdata.com/v4/cores/[id]).json()
- GitHub URL of the completed notebook: [SpaceX API calls notebook]



Root URL
https://api.spacexdata.com/v4

Root URL/launches/past

Root URL/rockets/id

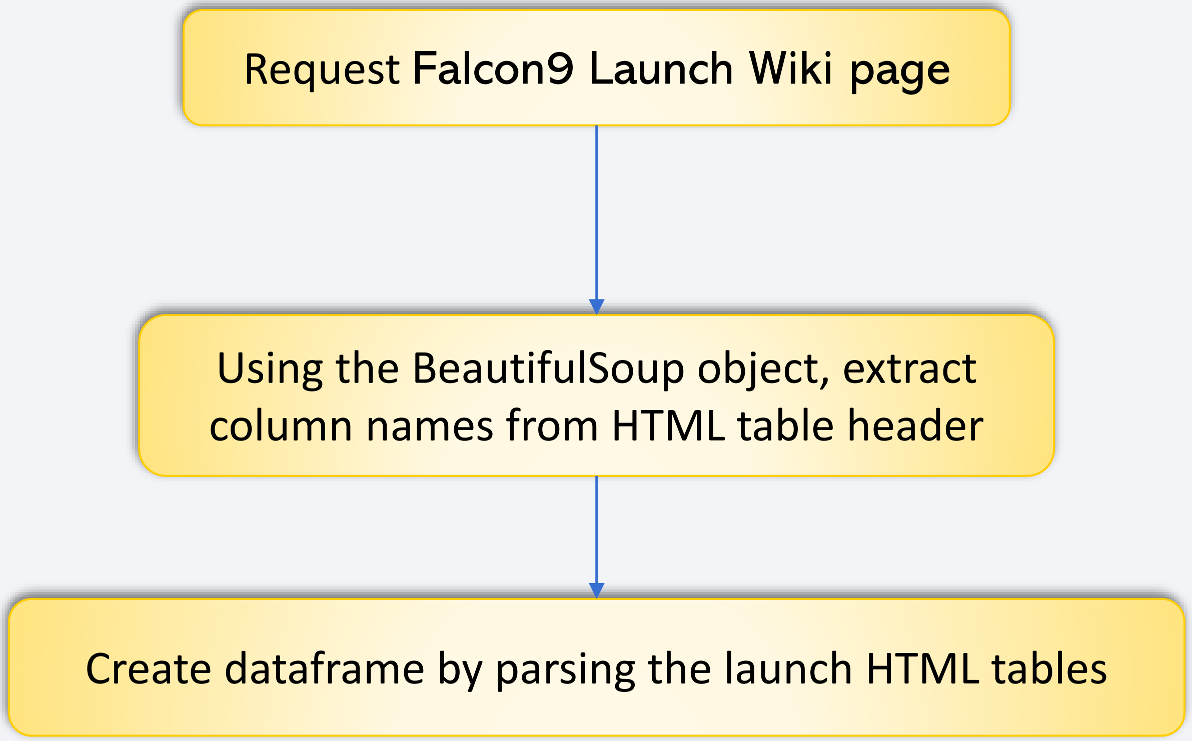
Root URL/launchpads/id

Root URL/payloads/id

Root URL/cores/id

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Extract all column names from the HTML table (3rd table) header using the BeautifulSoup library.
- Create a dataframe by parsing the launch HTML tables
- GitHub URL of the completed notebook: Web scraping notebook



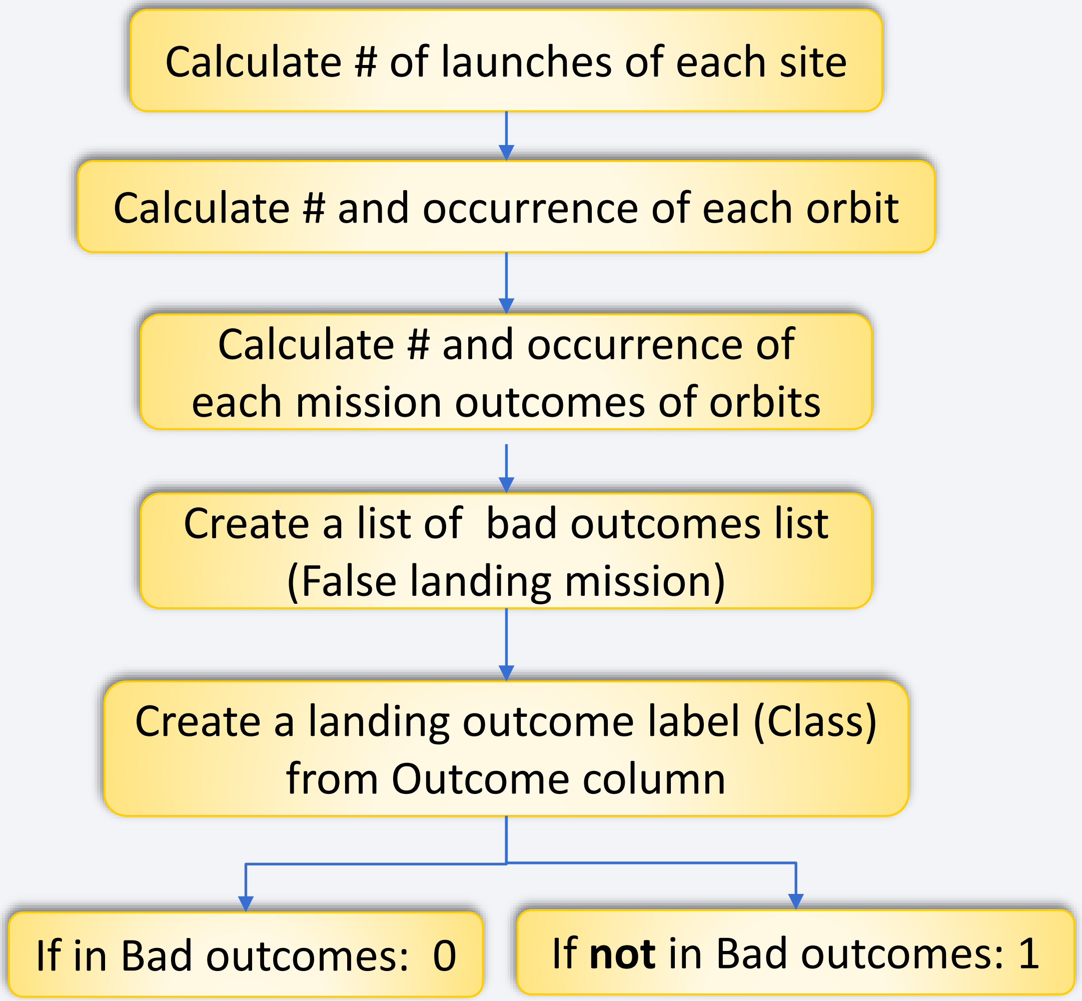
Request Falcon9 Launch Wiki page

Using the BeautifulSoup object, extract column names from HTML table header

Create dataframe by parsing the launch HTML tables

Data Wrangling

- The goal is to convert landing outcomes to training labels: 1 for success, 0 for failure.
- First read data into dataframe
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcomes of the orbits.
- Create a list of bad outcomes (False landing outcome)
- Create a landing outcome label (named Class) from the Outcome column: If the outcome value is in the bad outcomes list, assign 0, otherwise assign 1.
- GitHub URL of the completed notebook: [Data wrangling notebooks]()



Calculate # of launches of each site

Calculate # and occurrence of each orbit

Calculate # and occurrence of each mission outcomes of orbits

Create a list of bad outcomes list (False landing mission)

Create a landing outcome label (Class) from Outcome column

If in Bad outcomes: 0

If **not** in Bad outcomes: 1

EDA with Data Visualization

- The charts that were plotted are a Scatter plot, Bar chart, and Line plot.
- Scatter plot is used to show the relationship between 2 data variables.
  - FlightNumber vs. PayloadMass, Flight Number vs. Launch Site, Payload vs. Launch Site, FlightNumber vs. Orbit type, and Payload vs. Orbit type
- Bar chart is used to show the relationship between success rate and orbit type.
- Line plot is used to get the average launch success yearly trend
- GitHub URL of the completed notebook: EDA with data visualization notebook

EDA with SQL

- 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 in the ground pad was achieved.
- List the names of the boosters that have success in drone ships and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failed mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass. (using a subquery).
- List the records that will display the month names, failure outcomes in drone ship, booster versions, and launch_site for the months in the year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.
- GitHub URL of the completed notebook: EDA with SQL notebook

Build an Interactive Map with Folium

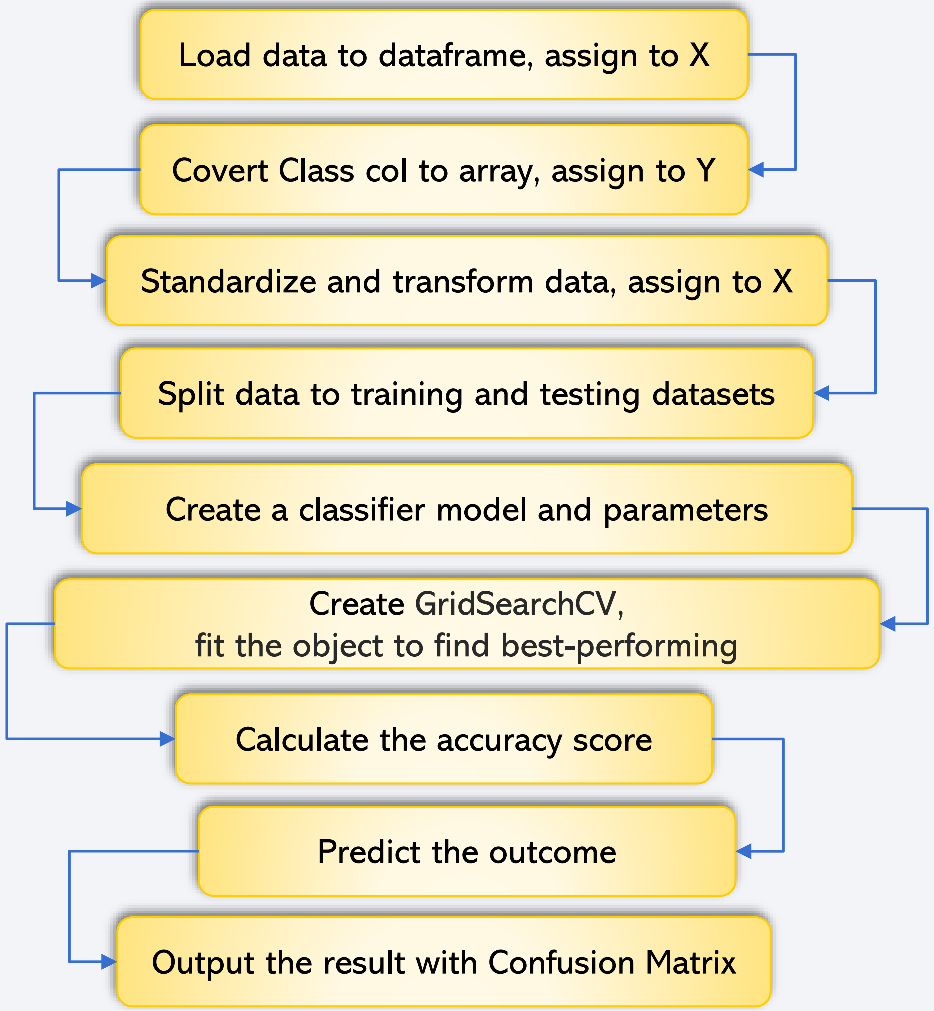
- The map objects that were created and added to each folium are Markers, Marker Clusters, Circles, and Lines.
- Markers pinpoint locations on a map representing a point of interest.
- Marker clusters group nearby markers into a single icon, improving map readability when dealing with a large number of markers.
- Circles represent areas on a map with a defined center point and radius, useful for visualizing boundaries.
- Lines are used to represent connections or paths between points or connections between different locations on the map.
- GitHub URL of the completed notebook: Interactive map with Folium map

Build a Dashboard with Plotly Dash

- The dashboard featured a pie chart, scatter plot, dropdown list, and range slider for interaction.
- The pie chart shows the total successful launches count for all sites if a specific launch site was selected, show the Success vs. Failed counts for the site.
- The scatter chart shows the correlation between payload and launch success.
- The dropdown menu will allow the user to select a launch site, with "All sites" being the default selected value.
- The range slider will allow the user to select the payload range.
- GitHub URL of the completed notebook: Plotly Dash lab

Predictive Analysis (Classification)

- Load data to the dataframe and assign it to variable X.
- Convert value in "Class" column to array and assign it to variable Y.
- Standardize the data in X and reassign it to X using the transform object.
- Use train_splite function to split the X/Y data into 80% training and 20% testing data sets.
- Create a classifier object (ex: Logistic Regression, Support Vector Machine, Decision Tree classifier, and K nearest neighbors) object and parameters for each classifier.
- Create a GridSearchCV object and fit the object with training data to find the best-performing classification model.
- Calculate the accuracy score using the score method.
- Predict the outcome with the test data and output the result using the Confusion Matrix.
- GitHub URL of the completed notebook: [Predictive analysis lab]()



Load data to dataframe, assign to X

Covert Class col to array, assign to Y

Standardize and transform data, assign to X

Split data to training and testing datasets

Create a classifier model and parameters

Create GridSearchCV,
fit the object to find best-performing

Calculate the accuracy score

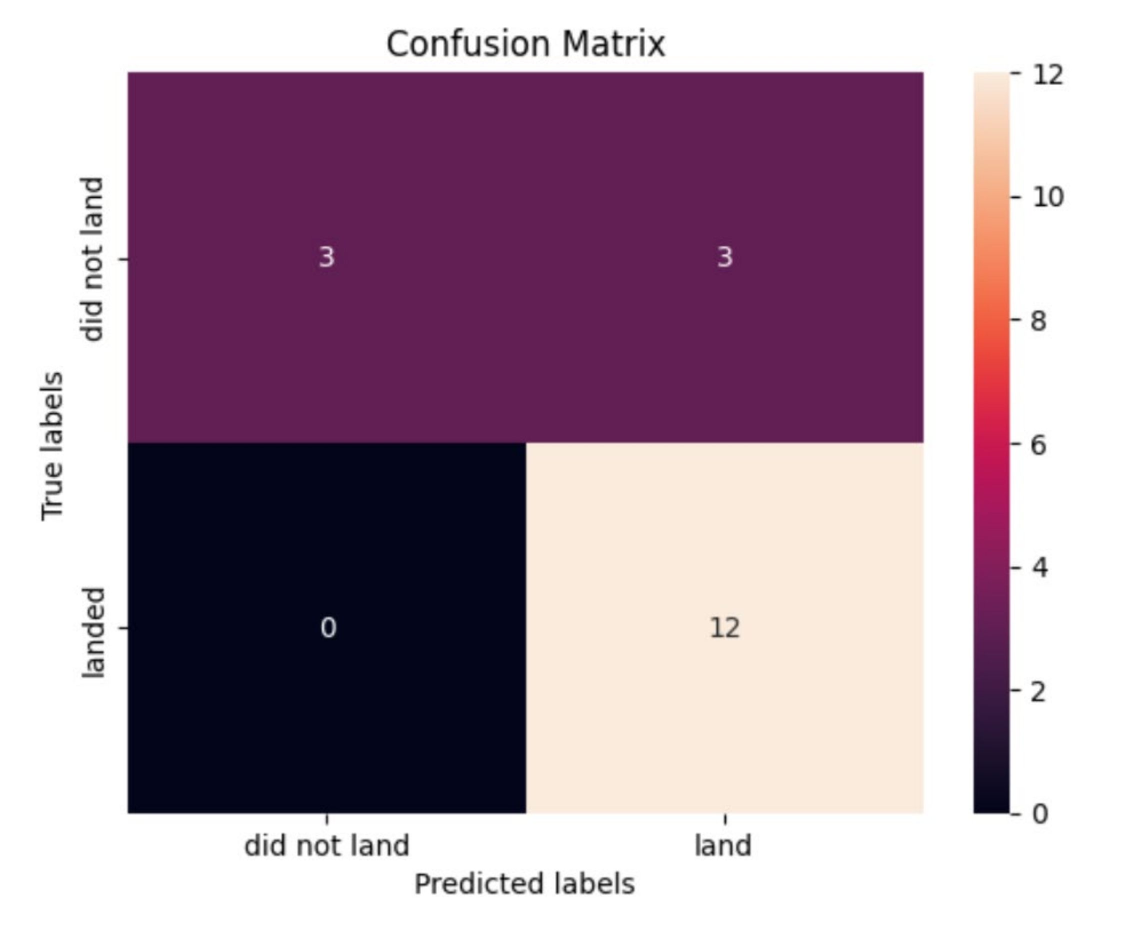
Predict the outcome

Output the result with Confusion Matrix

Results

Logistic Regression

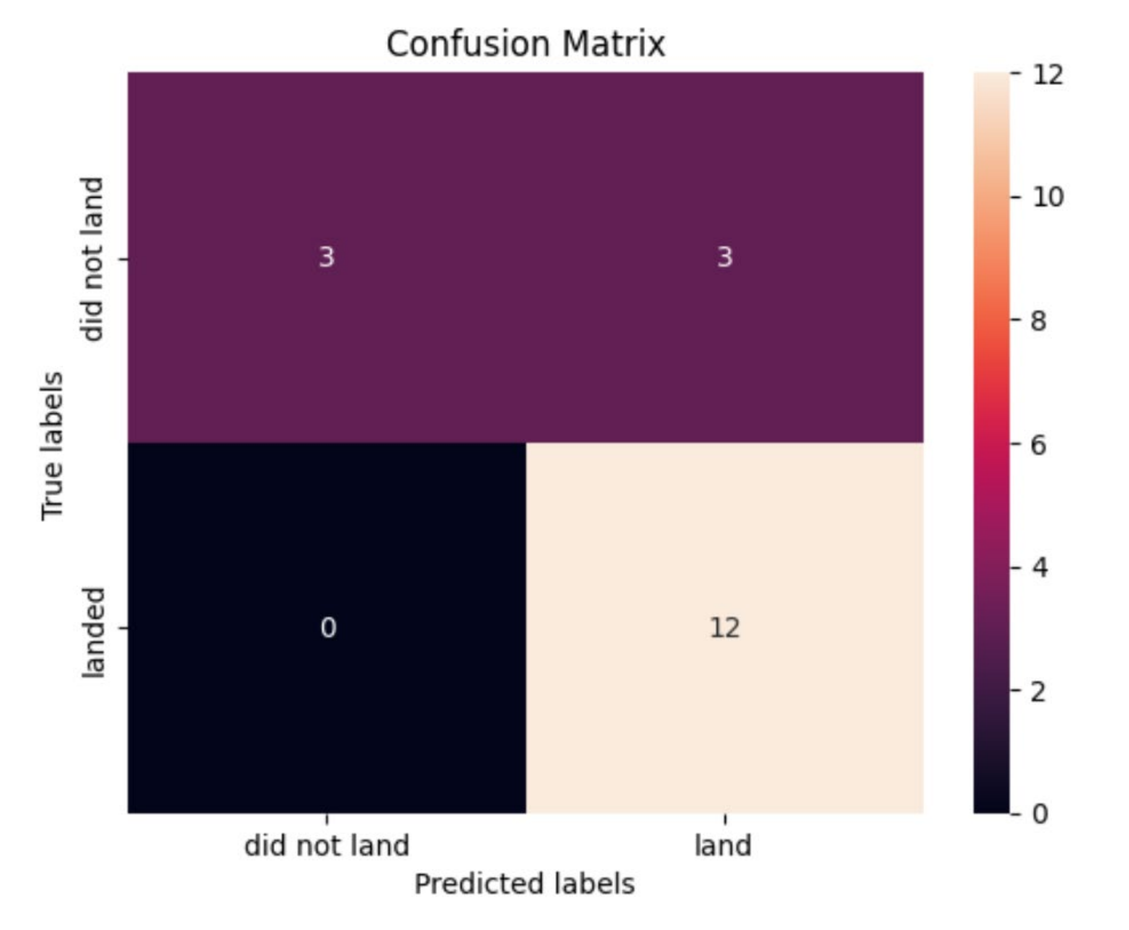
- Hpyerparmetrs:
  'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs’
- Accuracy : ~0.846
- Score: ~0.833
- Predictive analysis results: high false positive rate.



Results

Support Vector Machine (SVM)

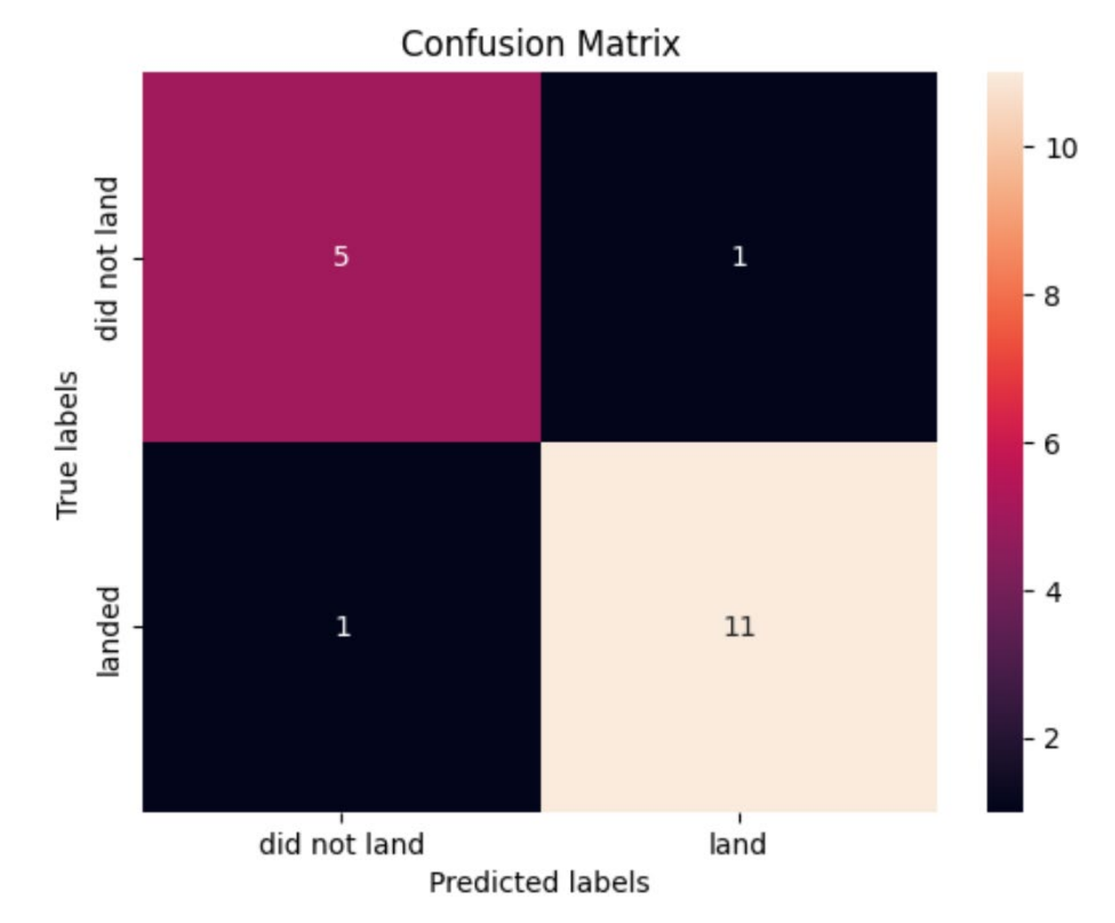
- Turned Hpyerparmetrs: 'C': 1.0, 'gamma': ~0.0316, 'kernel': 'sigmoid'
- Accuracy : ~0.848
- Score: ~0.833
- Predictive analysis results: high false positive rate.



Results

Decision Tree Classifier

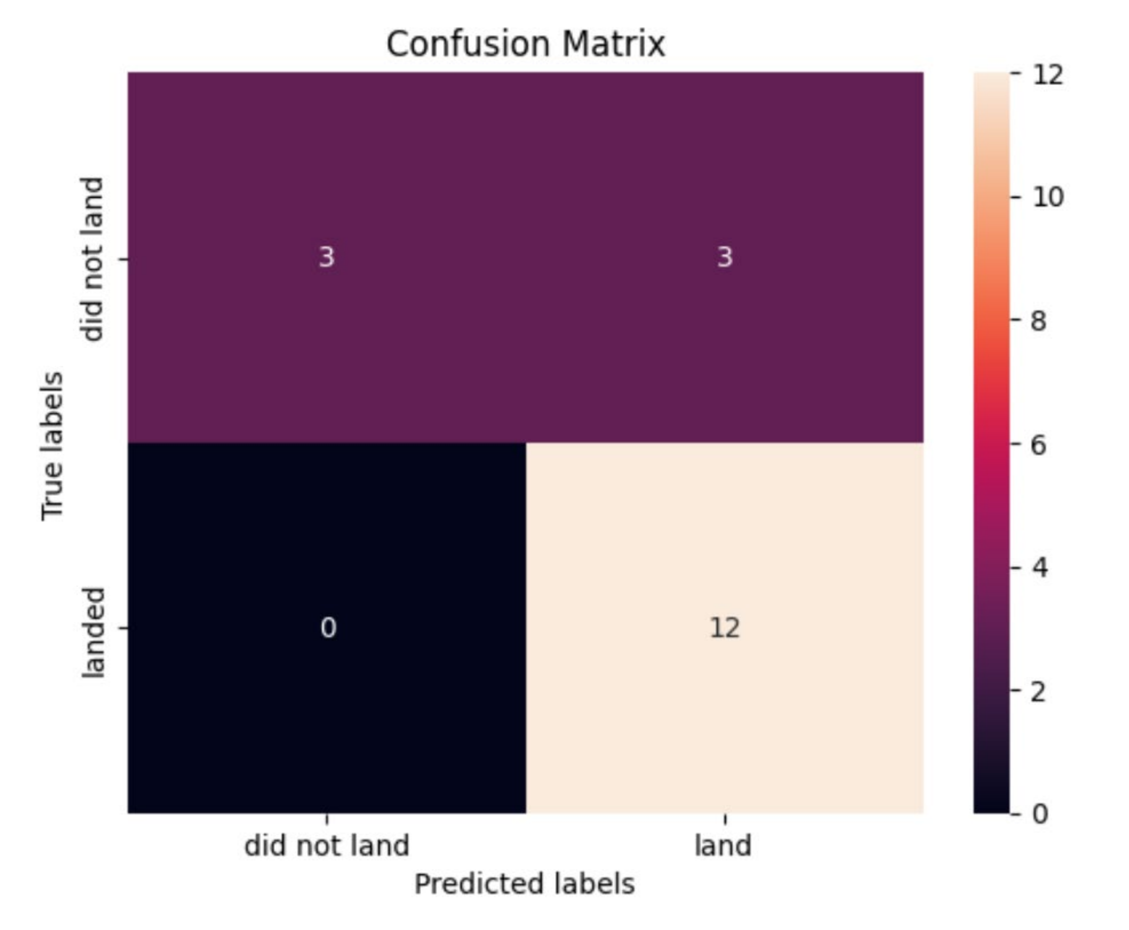
- Turned Hpyerparmetrs: 'criterion': 'gini', 'max_depth': 6, 'max_features': log2', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'
- Accuracy : ~0.889
- Score: ~0.888
- Predictive analysis results: low false positive and low false negative rate.
- The accuracy score is best among the other 3 models.



Results

K Nearest Neighbors

- Turned Hpyerparmetrs:
  'algorithm': 'auto', 'n_neighbors': 10, 'p’: 1
- Accuracy : ~0.848
- Score: ~0.833
- Predictive analysis results: high false positive rate.

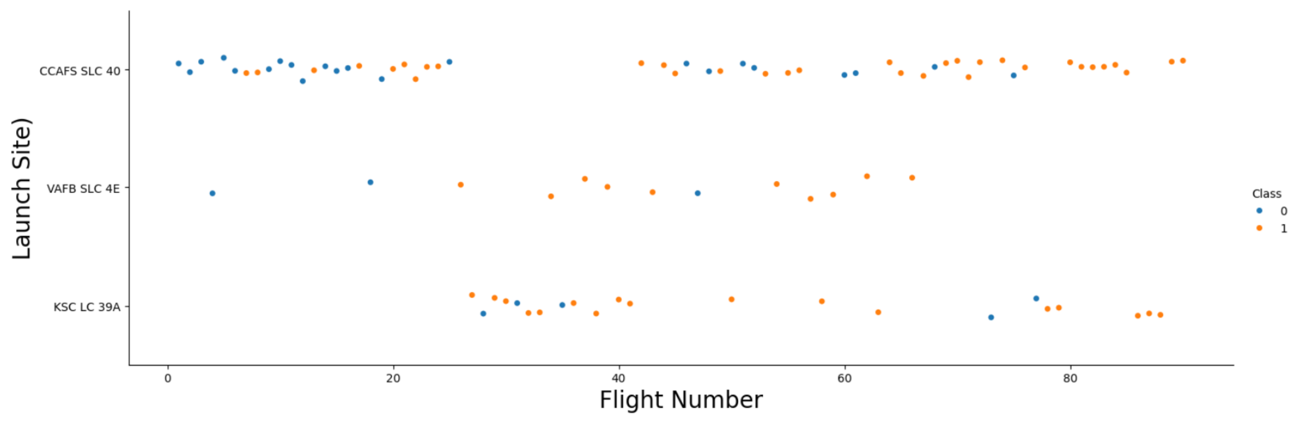


Section 2

Insights drawn from EDA

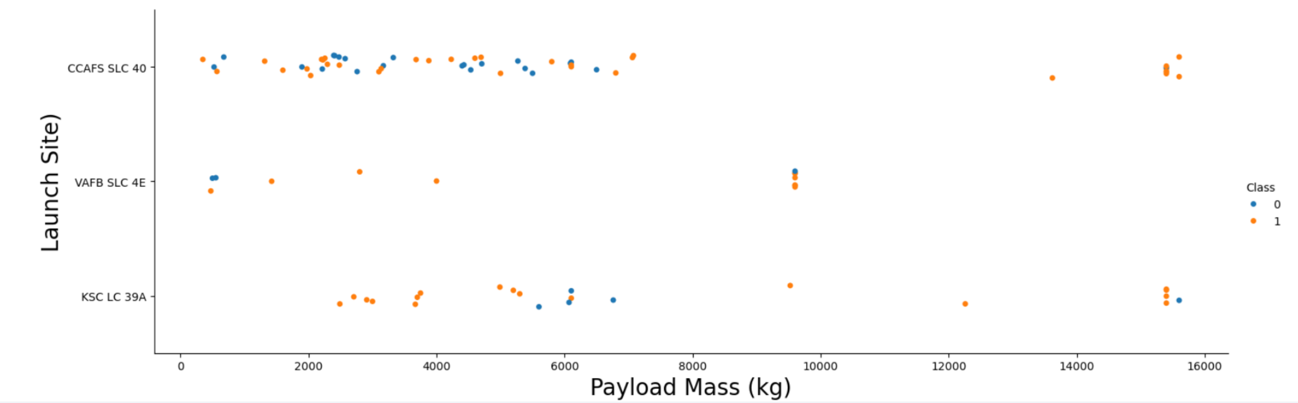
Flight Number vs. Launch Site

- CCAFS SLC 40 launchsite appears to have the highest number of flights.
- VAFB SLC 4E launchsite appears to have the lease number of flights.
- The success rate doesn’t appear to relate to the number of flights.



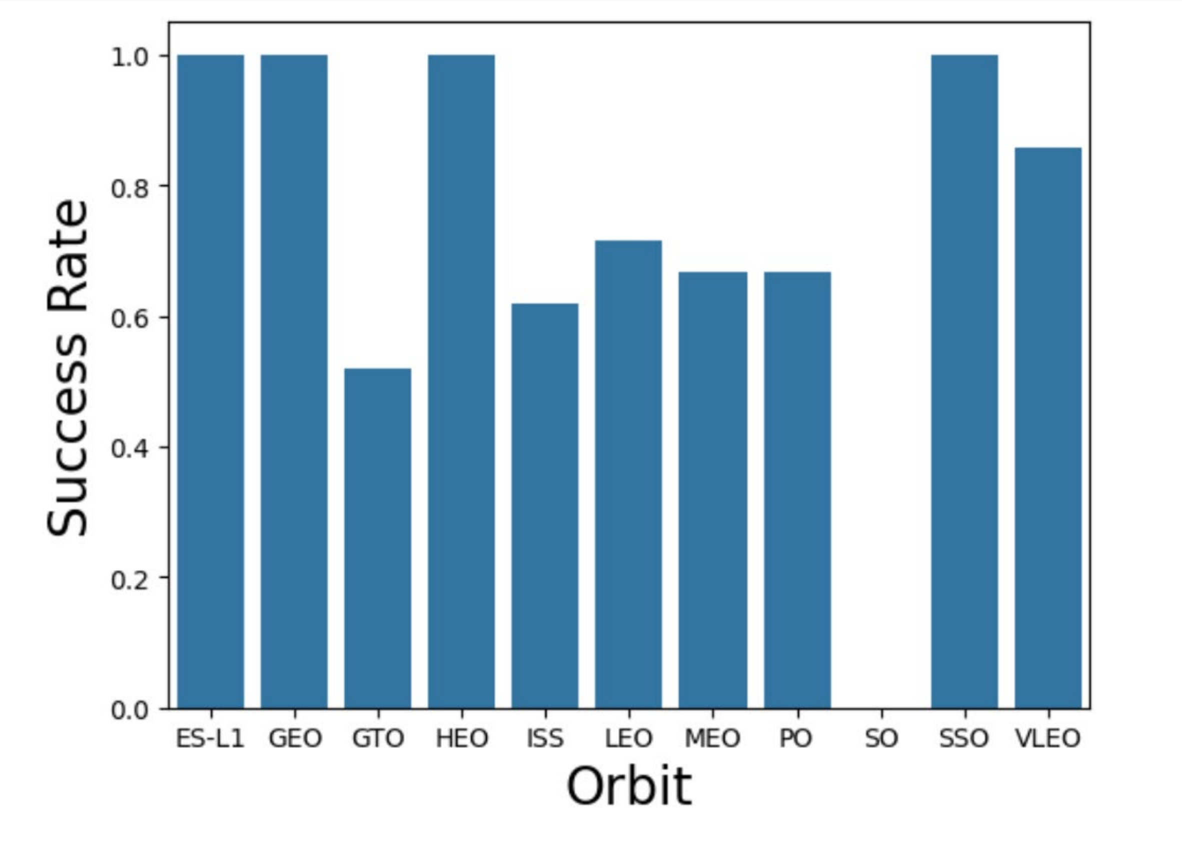
Payload vs. Launch Site

- VAFB-SLC launchsite has no rockets launched greater than 10000kg.
- KSC LC 39 launchsite had more failed launch rates with payload mass between 5000-7000kg and closer to 16000kg.
- CCAFS SLC 40 launchsite and Payload mass do not correlate with the successful rates.



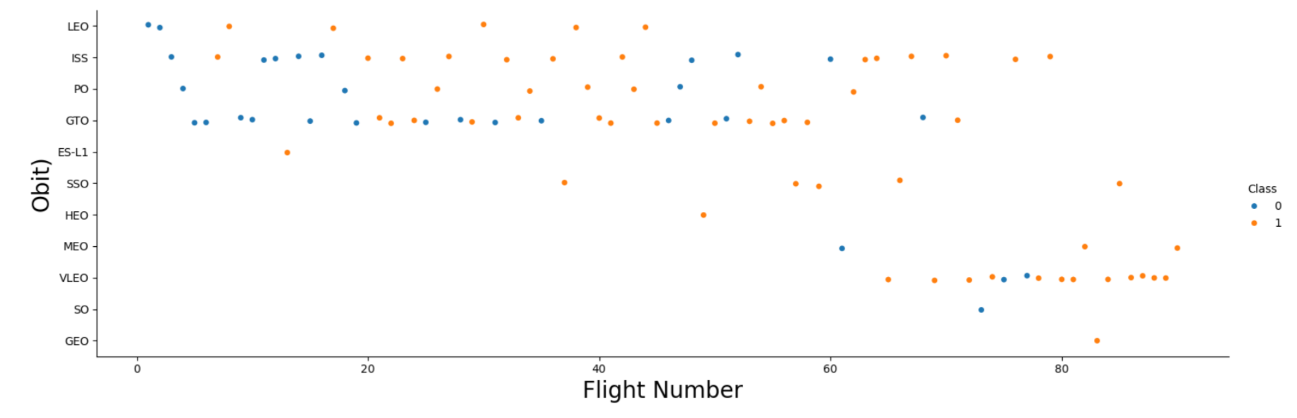
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO obits have a high success rate.
- GTO orbit has the lowest successful rate.
- SO orbit has no successful rate.



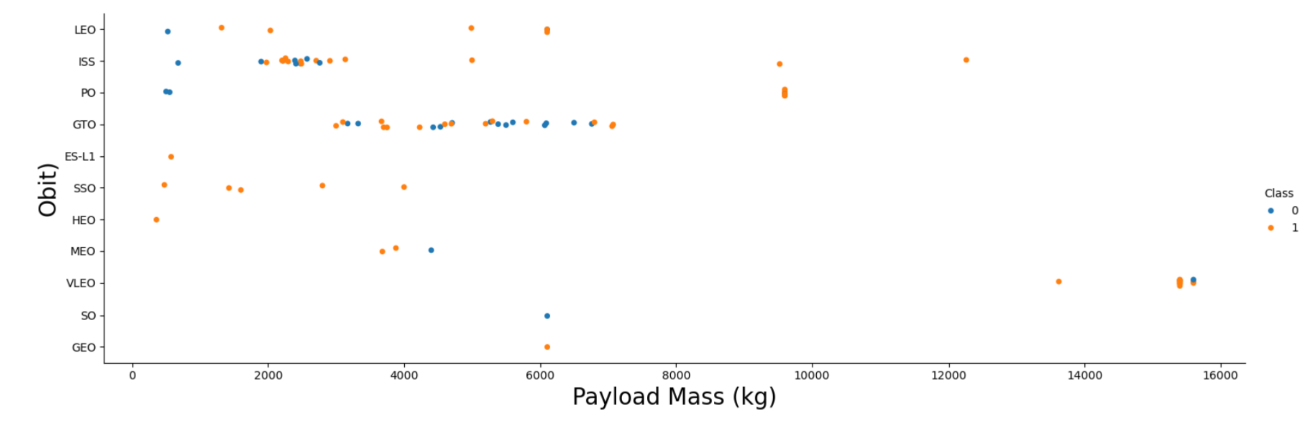
Flight Number vs. Orbit Type

- Success rates in LEO orbit seem to be closely linked to the frequency of flights.
- But not relationship between the flight number for GTO obit.



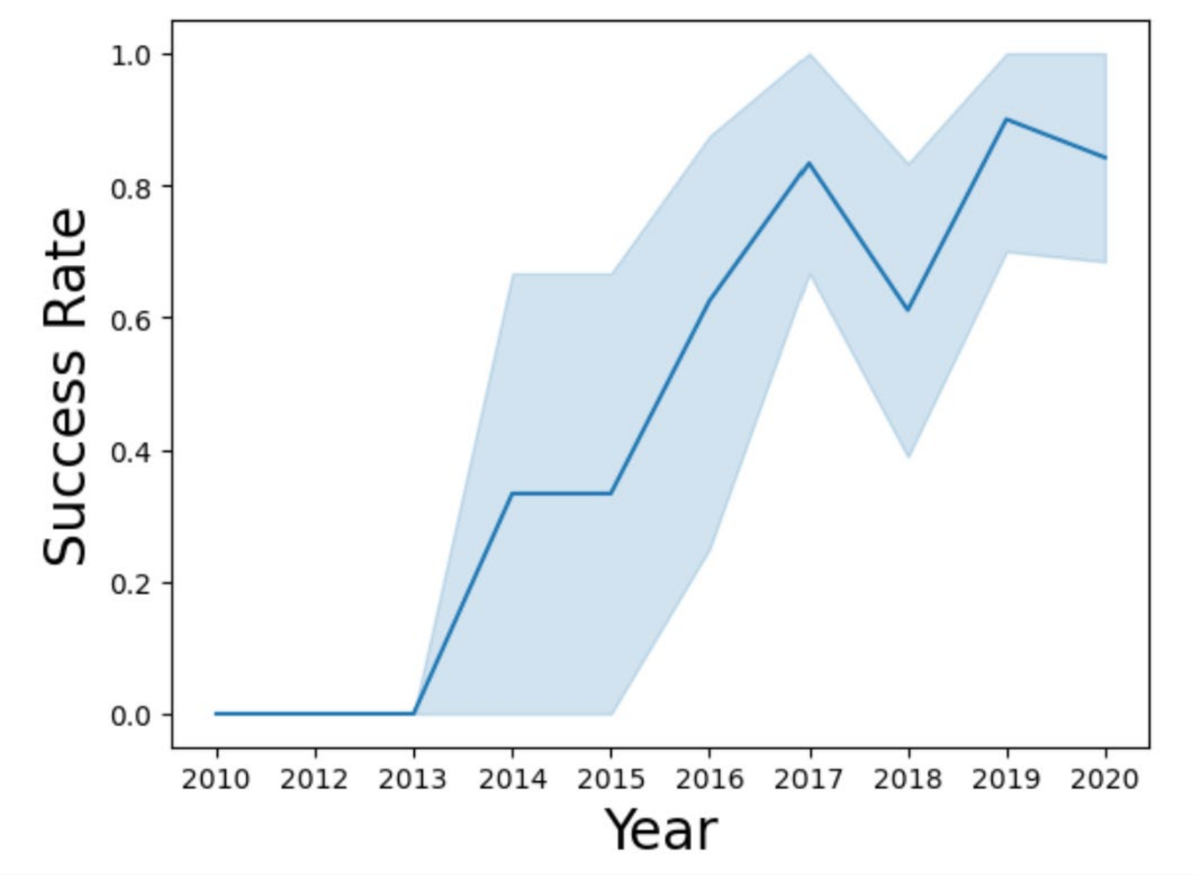
Payload vs. Orbit Type

- With heavy payloads the landing successful rate is higher or Polar, LEO, and ISS orbits.
- It is unclear how the successful landing rate is related to the GTO orbit and payload mass.



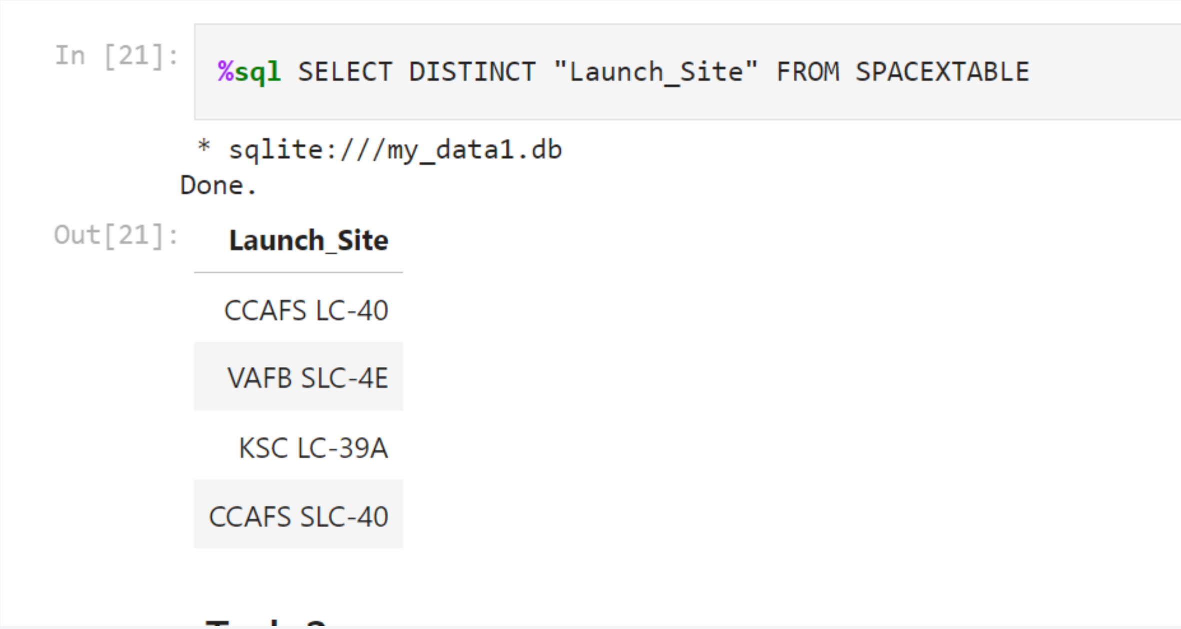
Launch Success Yearly Trend

- The success rate has increased since 2013 till 2000.
- But it went down between 2017-2018



All Launch Site Names

- Find the names of the unique launch sites
- Here are the 4 unique launchsites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40



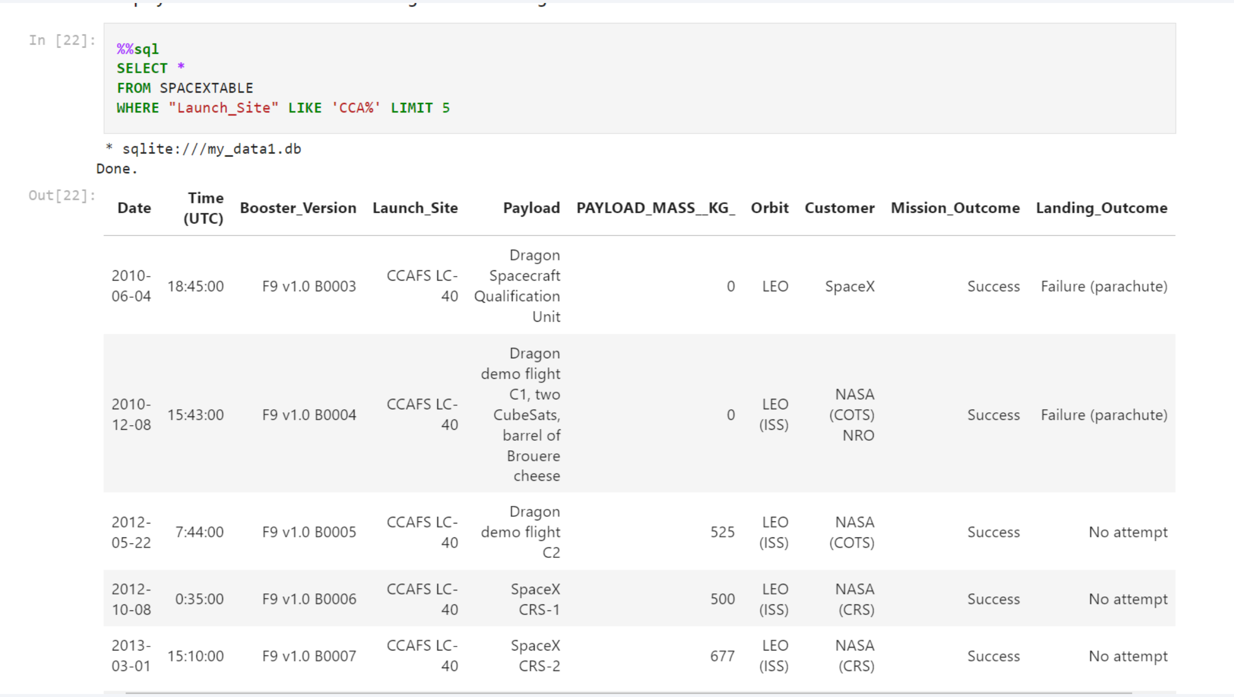
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * 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'

- Find 5 records where launch sites begin with CCA.
- The result shows only 5 records with launch sites starting with CCA.

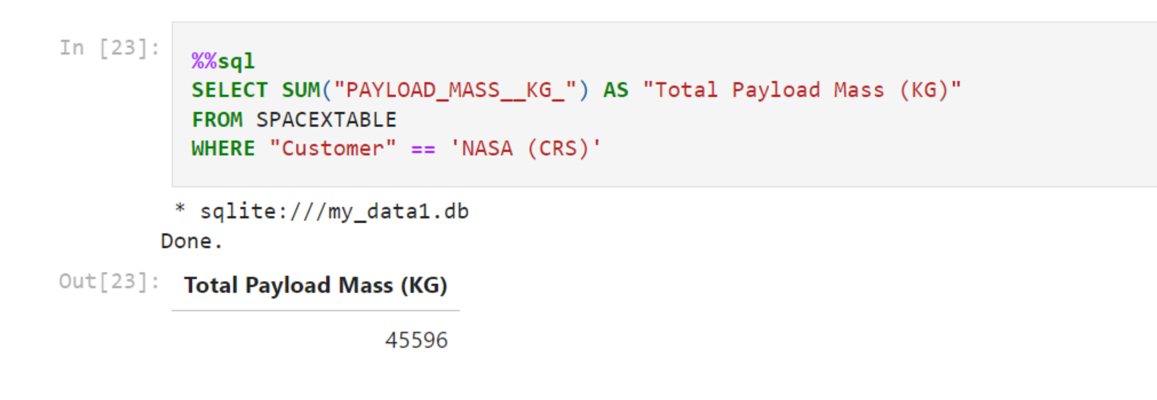


```
In [22]: %%sql
         SELECT *
         FROM SPACEXTABLE
         WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5

          * sqlite:///my_data1.db
         Done.
Out[22]:
```

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

- Calculate the total payload carried by boosters from NASA.
- The result is 45596 Kg.



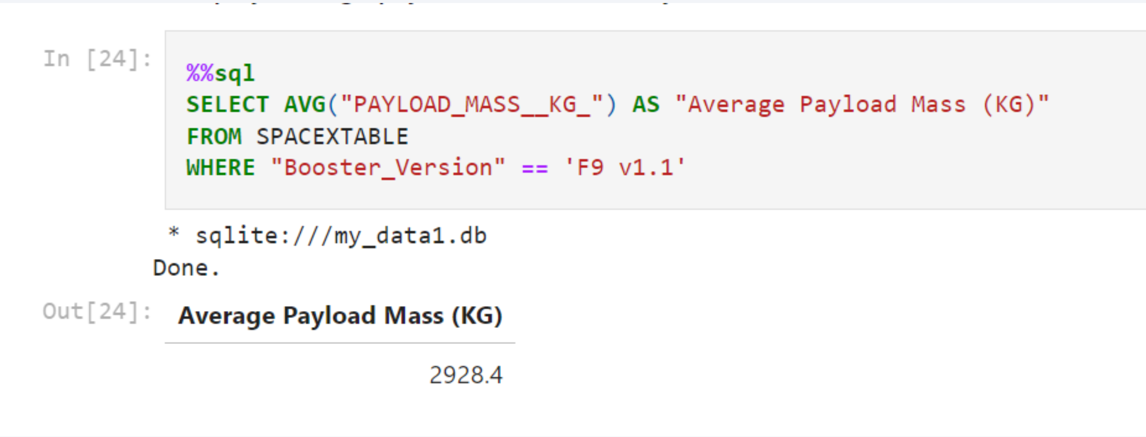
```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass (KG)"
FROM SPACEXTABLE
WHERE "Customer" == 'NASA (CRS)'
```

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

Total Payload Mass (KG)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1.
- The result is 2928.4 Kg.

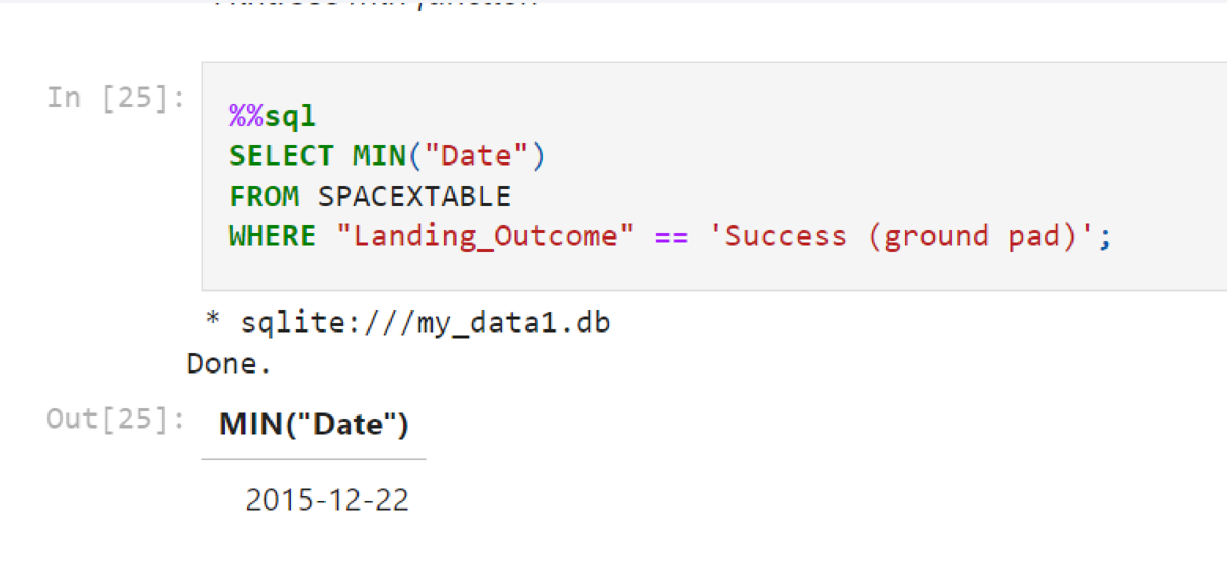


```
In [24]:  %%sql
          SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass (KG)"
          FROM SPACEXTABLE
          WHERE "Booster_Version" == 'F9 v1.1'

           * sqlite:///my_data1.db
          Done.
Out[24]:
```

Average Payload Mass (KG)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad.
- The result is December 22, 2015.



In [25]:

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

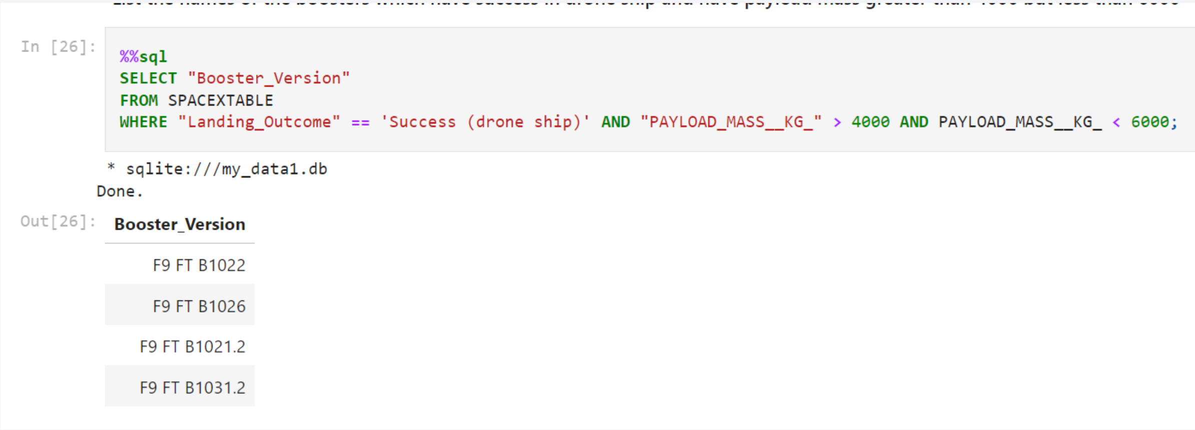
```
 * sqlite:///my_data1.db
Done.
```

Out[25]:

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- Here are the results:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2



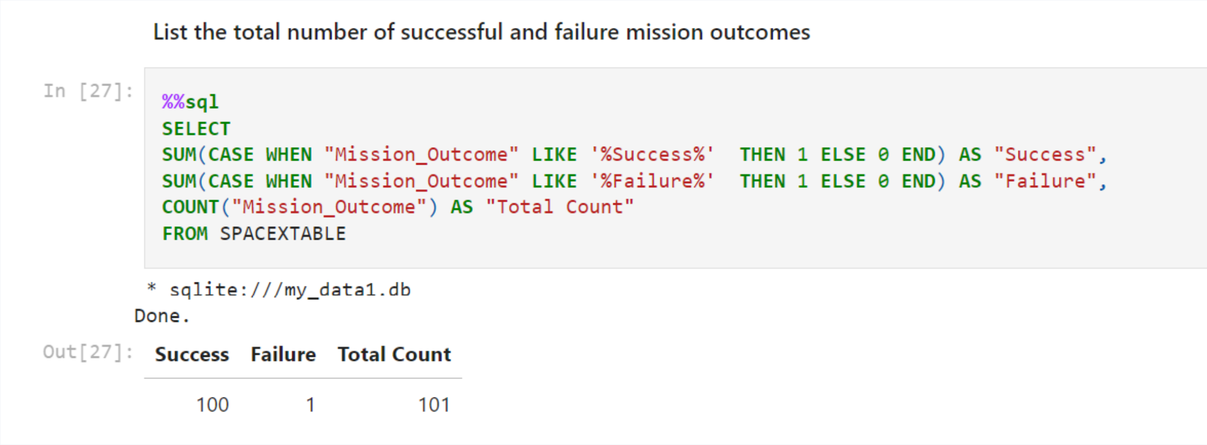
```
In [26]: %%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" == 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND PAYLOAD_MASS__KG_ < 6000;

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

Out[26]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes.
- The results are:
  - Total count = 101
  - Success = 100
  - Failure = 1



List the total number of successful and failure mission outcomes

In [27]:

```
%%sql
SELECT
SUM(CASE WHEN "Mission_Outcome" LIKE '%Success%'  THEN 1 ELSE 0 END) AS "Success",
SUM(CASE WHEN "Mission_Outcome" LIKE '%Failure%'  THEN 1 ELSE 0 END) AS "Failure",
COUNT("Mission_Outcome") AS "Total Count"
FROM SPACEXTABLE
```

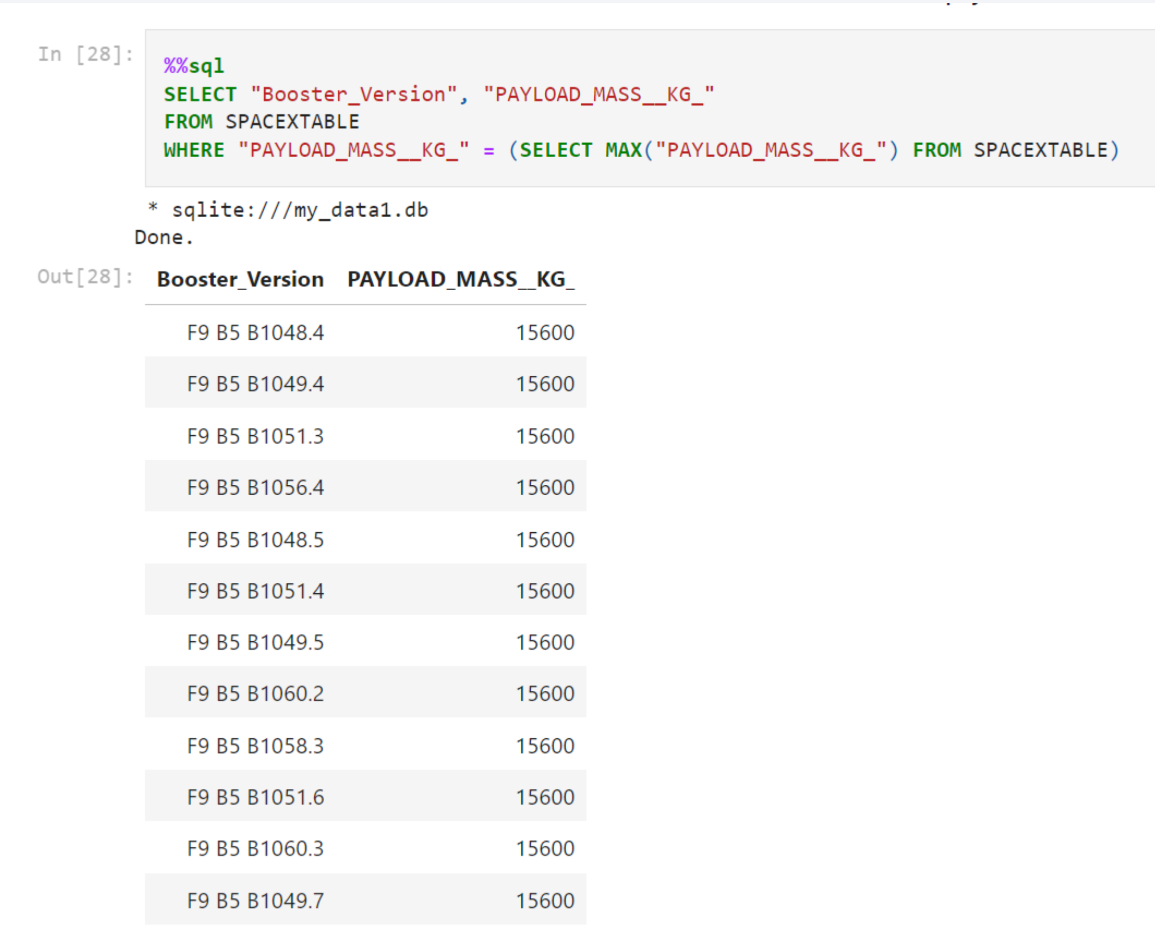
```
 * sqlite:///my_data1.db
Done.
```

Out[27]:

Success	Failure	Total Count
100	1	101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass.
- The maximum payload mass is 15600 Kg.
- There are 12 boosters carrying the maximum payload mass.



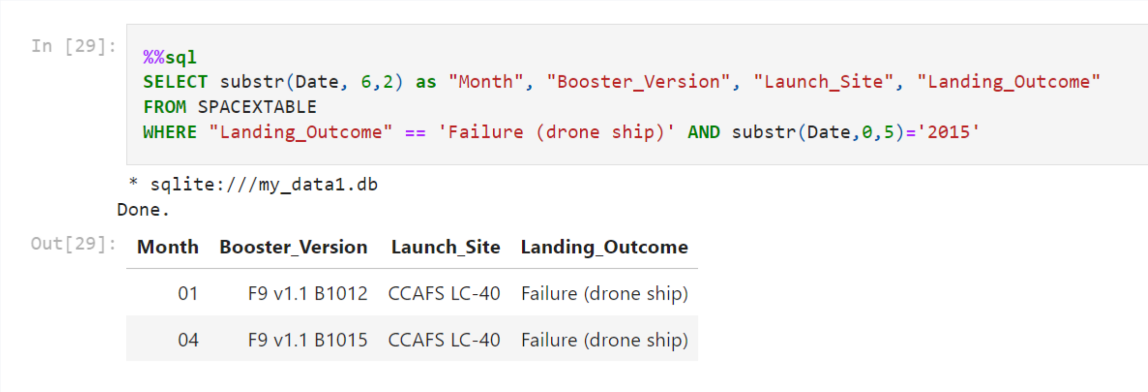
```
In [28]: %%sql
         SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
         FROM SPACEXTABLE
         WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)

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

Out[28]:

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- There are 2 records with failed landing outcomes in 2015.



In [29]:

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

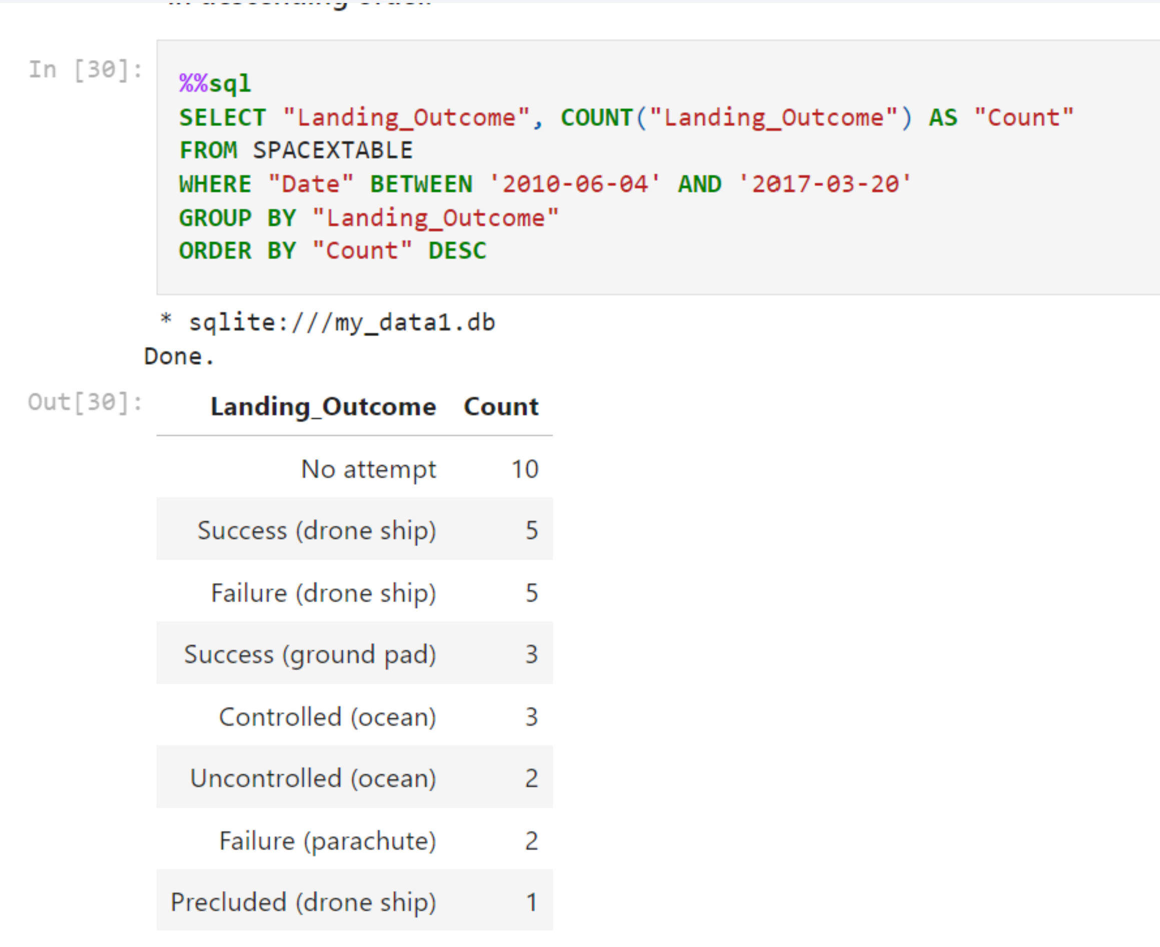
```
 * sqlite:///my_data1.db
Done.
```

Out[29]:

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

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.
- The highest rank in this period is 10 for No attempt.
- The lowest rank is 1 for Precluded (drone ship).
- Both Success and Failure (drone ship) have equal rank counts of 5 in the specified period.



In [30]:

```
%%sql
SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS "Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC
```

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

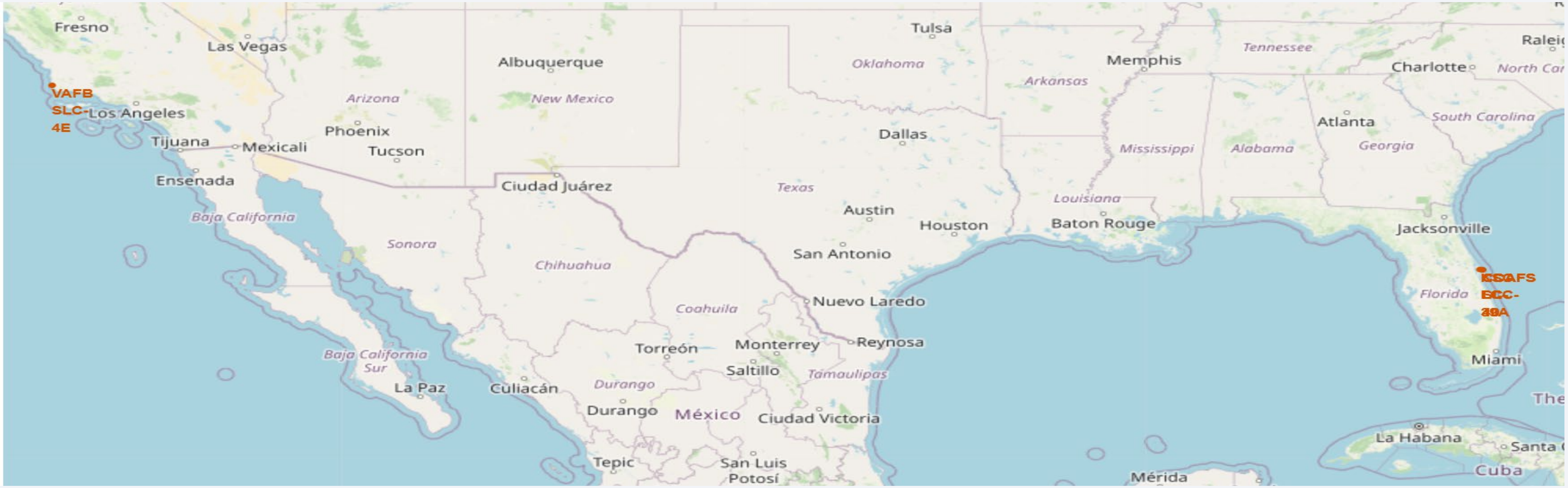
Out[30]:

Landing_Outcome	Count
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

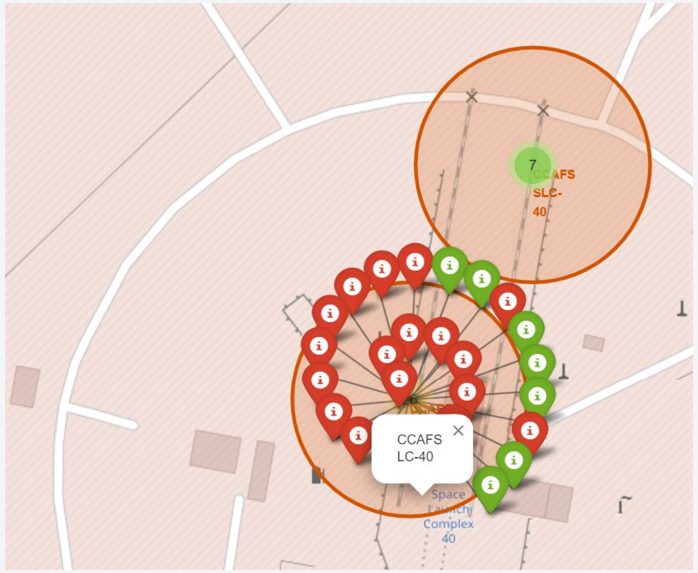
All Launch Sites

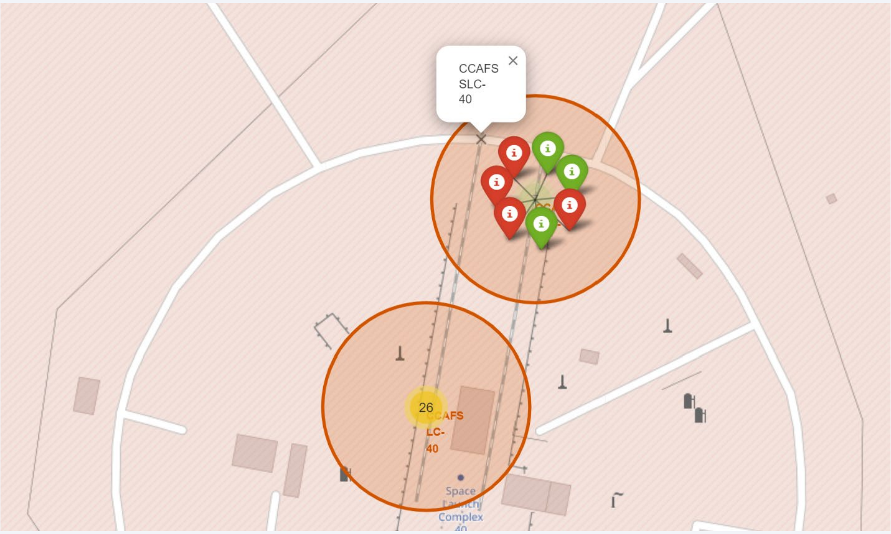


- VAFB SLC-4E is in CA.
- CCAFS SLC-40, CCAFS LC-40, and KSC LC-39A are located in FL.

Launchsites and Outcomes

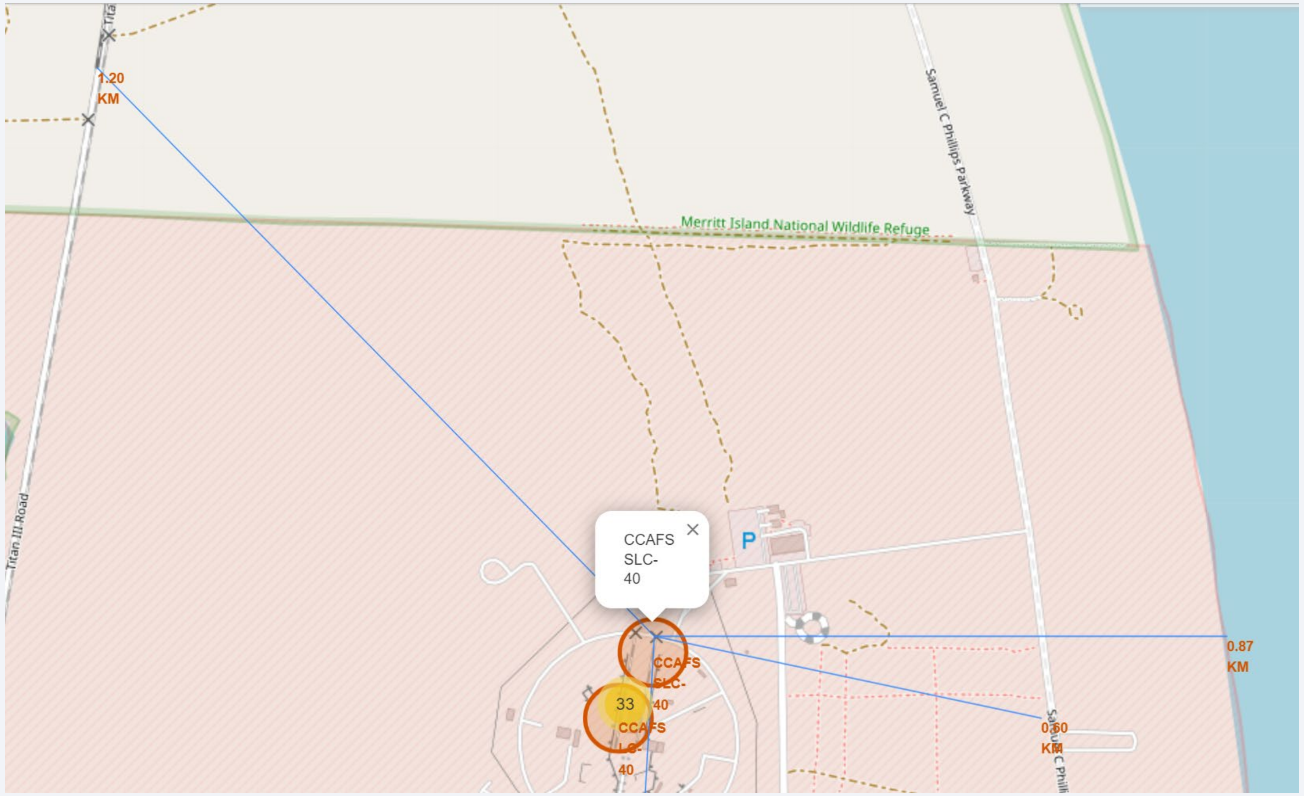
- CCAFS LC-40 launchsite has 26 outcomes
  - 7 outcomes are successful (Green marker)
  - 21 outcomes are failure (Red marker)
- CCAFS SLC-40 launchsite has 7 outcomes
  - 3 outcomes are successful (Green marker)
  - 4 outcomes are failure (Red marker)





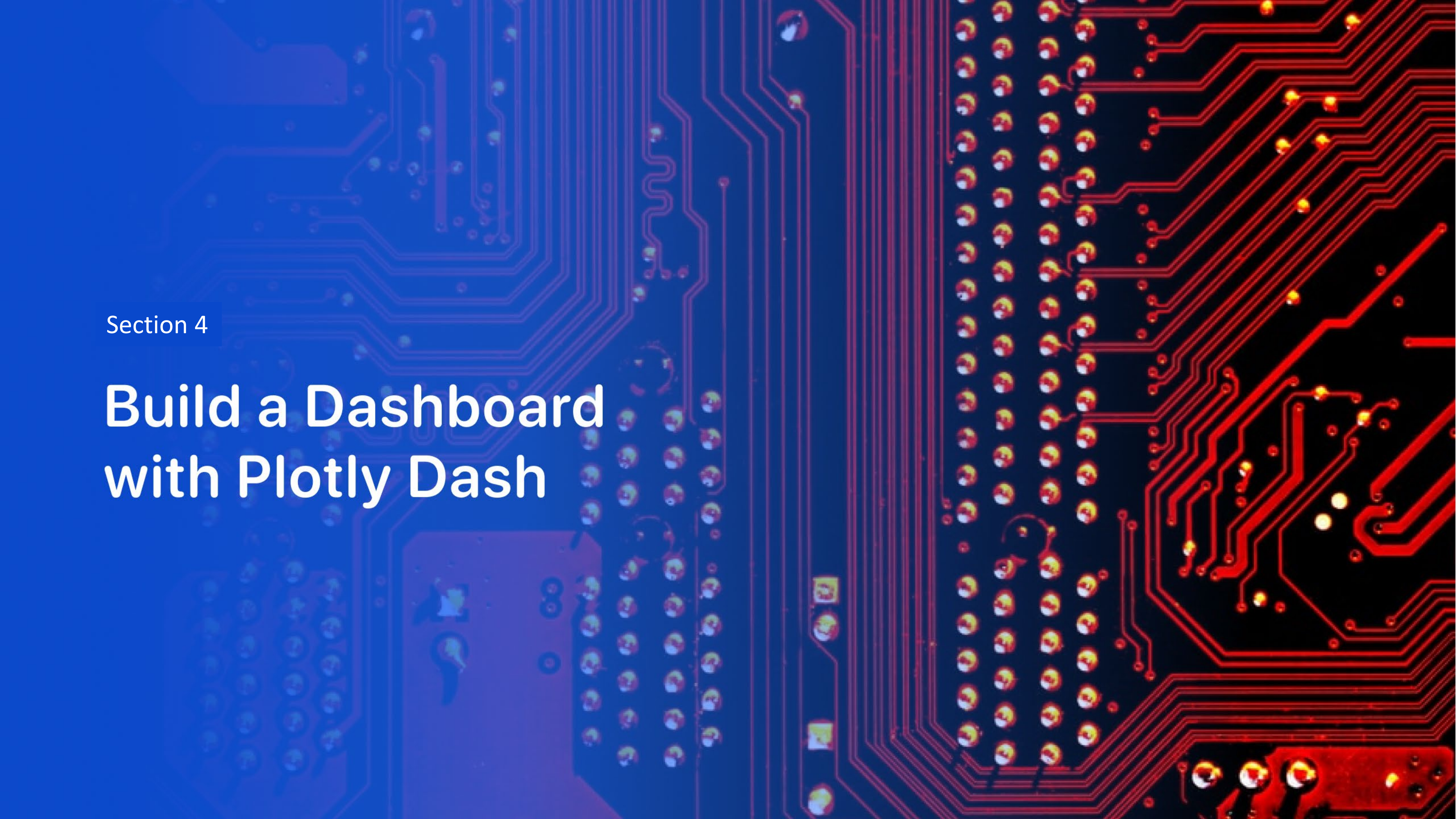
Launch Site to Its Proximities

- CCAFS SLC-40 is located approximately
  - 0.87km from the coastline
  - 0.30km from the highway
  - 1.20km from the railway
- Launch sites are situated at a greater distance from the city



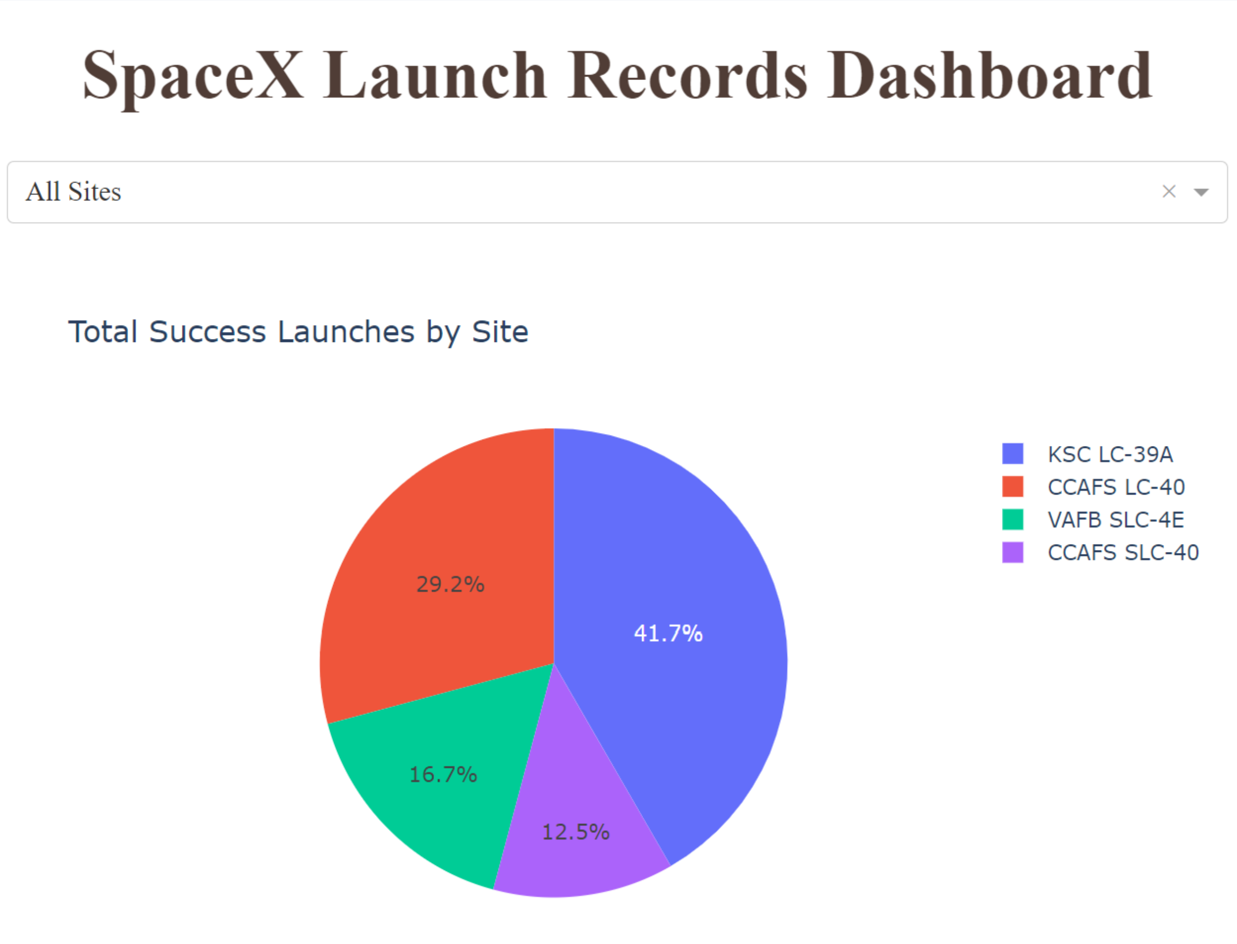
Section 4

Build a Dashboard with Plotly Dash



Total Success Launches by Site

- KSC LC-39A appears to have the highest success rate at 41.7%.
- While CCAFS SLC-40 has the lowest successful rate at 12.5%.



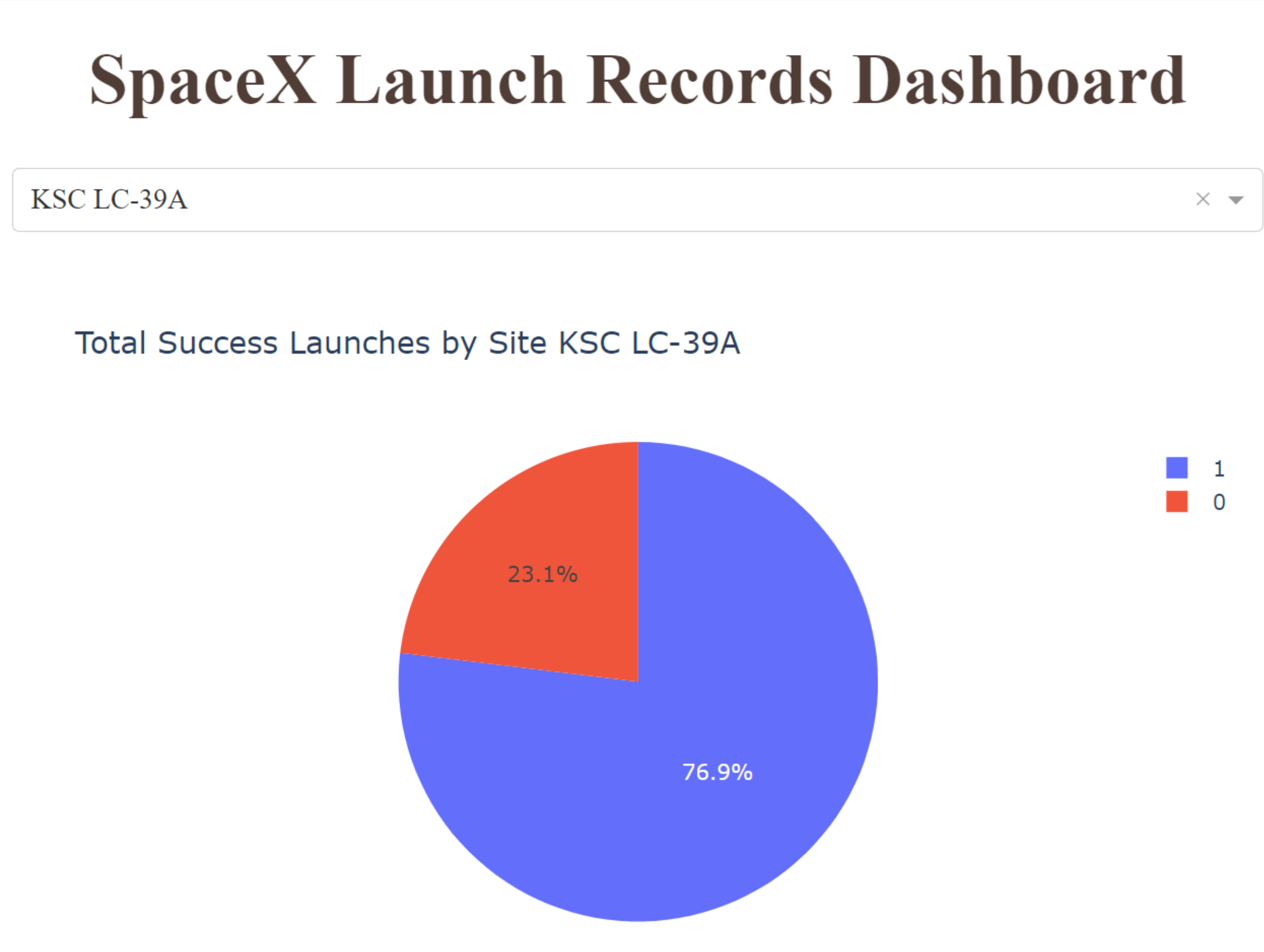
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

Highest Launch Success Ratio

- KSK LC-39A has the highest launch success ratio.
- 76.6% success vs 23.1% failure rate.

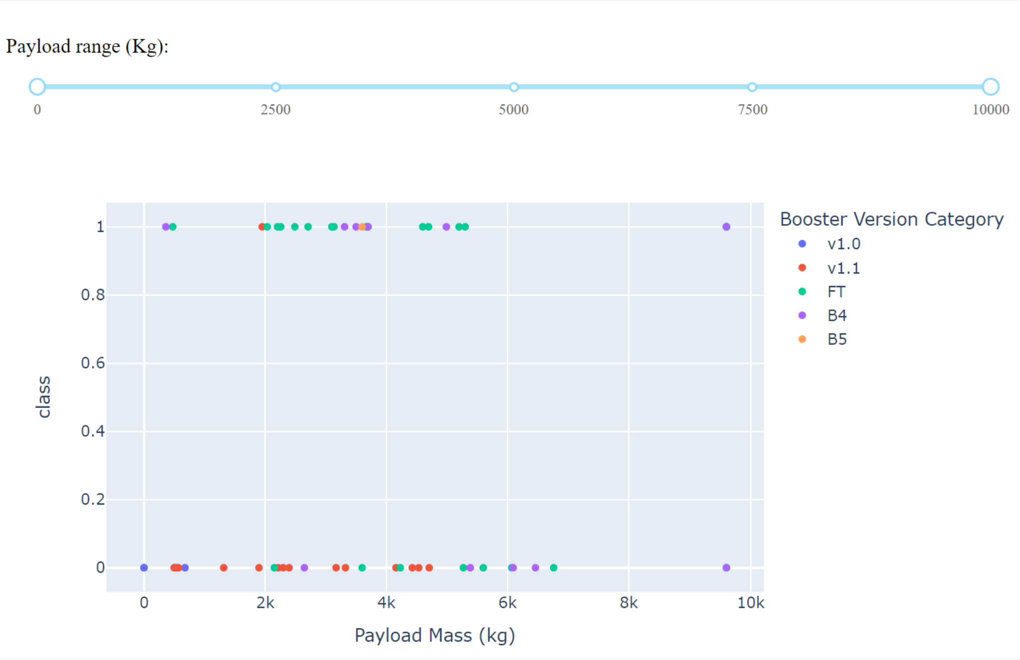


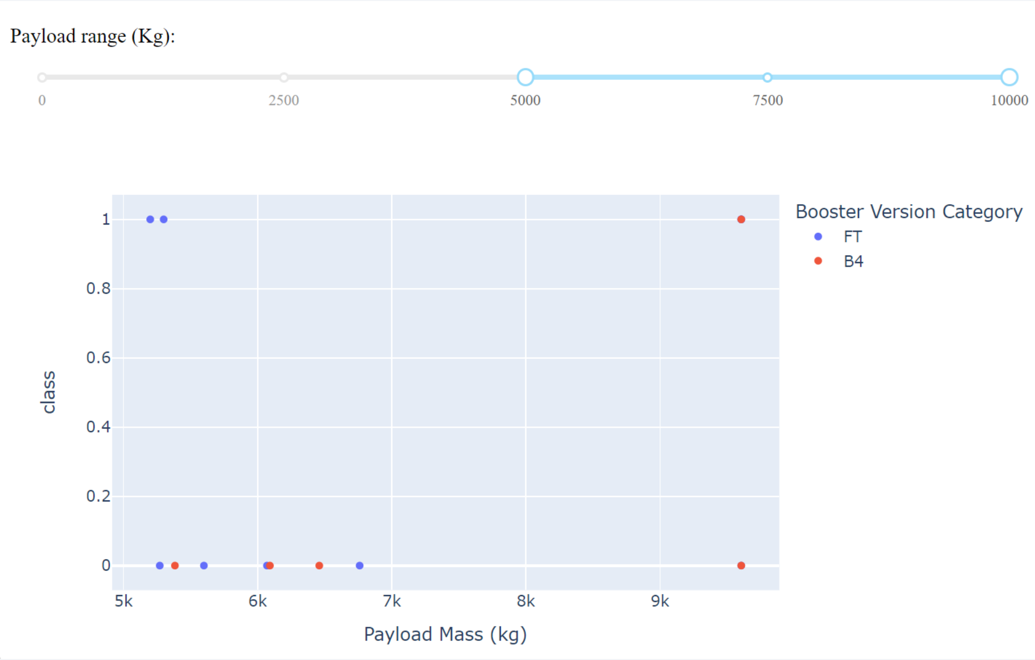
SpaceX Launch Records Dashboard

KSC LC-39A × ▾

Total Success Launches by Site KSC LC-39A

<Dashboard Screenshot 3>





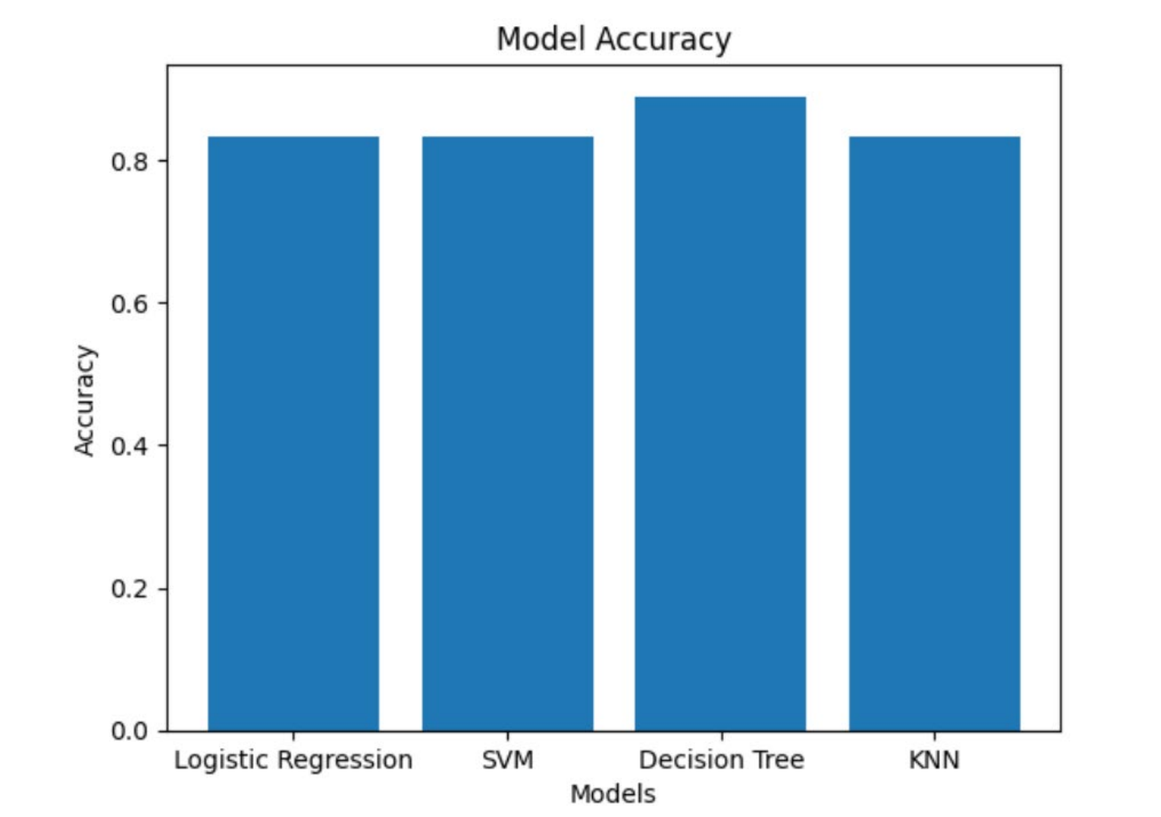
- Within the full range of 10000kg payload mass, Booster Version FT appears to have the largest successful rate up to 5300kg.
- In contrast Booster Version V1.1 appears to have the largest failure rate.
- In the 5000kg to 10000kg range, only Booster Version FT and B4 have some successful missions.
- Booster Version B4 has both success and failure at the same payload mass of 9600kg. The

Section 5

Predictive Analysis (Classification)

Classification Accuracy

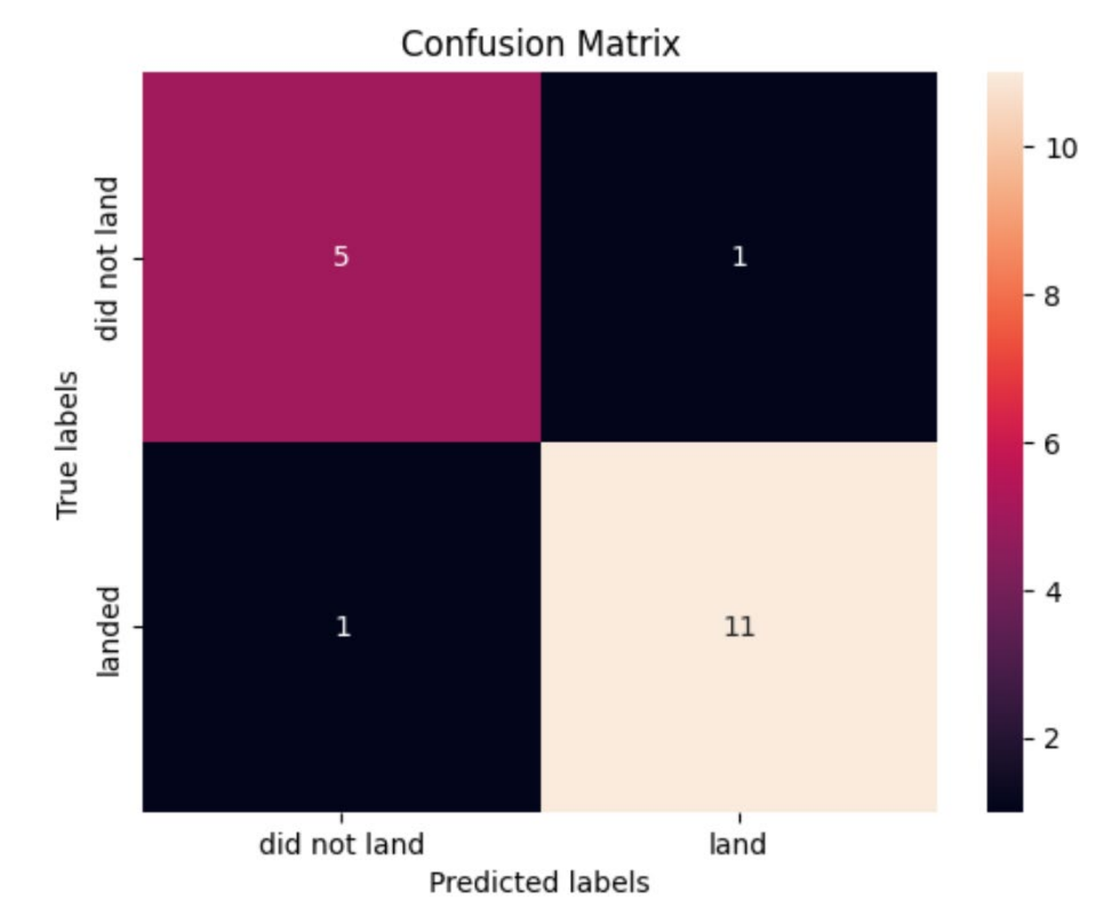
- The Decision Tree model appears to have the highest accuracy score of ~0.888.
- While the others have an accuracy score of ~0.833.



Confusion Matrix

Decision Tree Classifier

- Decision Tree Classifier appears to perform the best with an accuracy score ~0.888
- Predictive analysis results: low false positive and low false negative rate.
- The accuracy score is best among the other 3 models.



Conclusions

- The Decision Tree Model appears to perform the best with an accuracy score of 0.888.
- It seems that the number of flights does not have a direct correlation with the success rate.
- The correlation between launch sites and payload mass is unclear, particularly for the CCAFS SLC 40 launch site.
- KSC's LC-39A has the highest success rate at 41.7%, with a 76.9% success rate ratio.
- In contrast, CCAFS SLC-40, has the lowest successful rate of 12.5%, with a 57.1% success ration.

Appendix

- Notebooks:
  - GitHub URL: SpaceX API calls notebook
  - GitHub URL: Web scraping notebook
  - GitHub URL: Data wrangling notebooks
  - GitHub URL: EDA with data visualization notebook
  - GitHub URL: EDA with SQL notebook
  - GitHub URL: Interactive map with Folium map
  - GitHub URL: Plotly Dash lab
  - GitHub URL: Predictive analysis lab

Thank you!