IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason McIntire
December 30, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodologies Used:

- Data collection
  - Rest API
  - Web scraping
- Data Wrangling
- Exploratory Data Analysis
- Data Visualization
  - Folium (maps)
  - Plotly Dash (dashboard)
- Predictive Analysis (Machine Learning)

Summary of all results:

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

Introduction

This project was launched to review historical SpaceX launch data to determine how we can predict the success of future launches.

In particular, I want to discover the best model to be used for accurately predicting the success of a given future launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected from the SpaceX API And by scraping data from SpaceX Falcon 9 launch Wikipedia page
- Perform data wrangling
    - We ensured there were no missing values, and created a new column that indicated the success or failure (class) of each launch, based on Outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Using SciKit-Learn to determine best classification model to calculate the best model to predict launch success

Data Collection

Historical launch data was collected from the SpaceX API And by scraping data from SpaceX Falcon 9 launch Wikipedia page.

SpaceX API
https://api.spacexdata.com/v4

- Launch history
- Booster version
- Launchpad Location
- Payload Size
- Core Specs

Wikipedia Tables
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

HTML table with all data needed for data analysis

Data Collection – SpaceX API

GET https://api.spacexdata.com/v4/launches/past

GET https://api.spacexdata.com/v4/rockets

GET https://api.spacexdata.com/v4/launchpads

GET https://api.spacexdata.com/v4/payloads

GET https://api.spacexdata.com/v4/cores

After retrieving a collection of all past launches, subsequent data was retrieved to get:

- Booster version
- Launchpad location
- Payload size
- Cores

View Jupyter Notebook on GitHub

Data Collection - Scraping

Data was collected via web scraping from:

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Convert data output from Wikipedia to BeautifulSoup object

↓

Locate the required table from within the BeautifulSoup object

↓

Extract column names

↓

Parse HTML table into a dictionary object

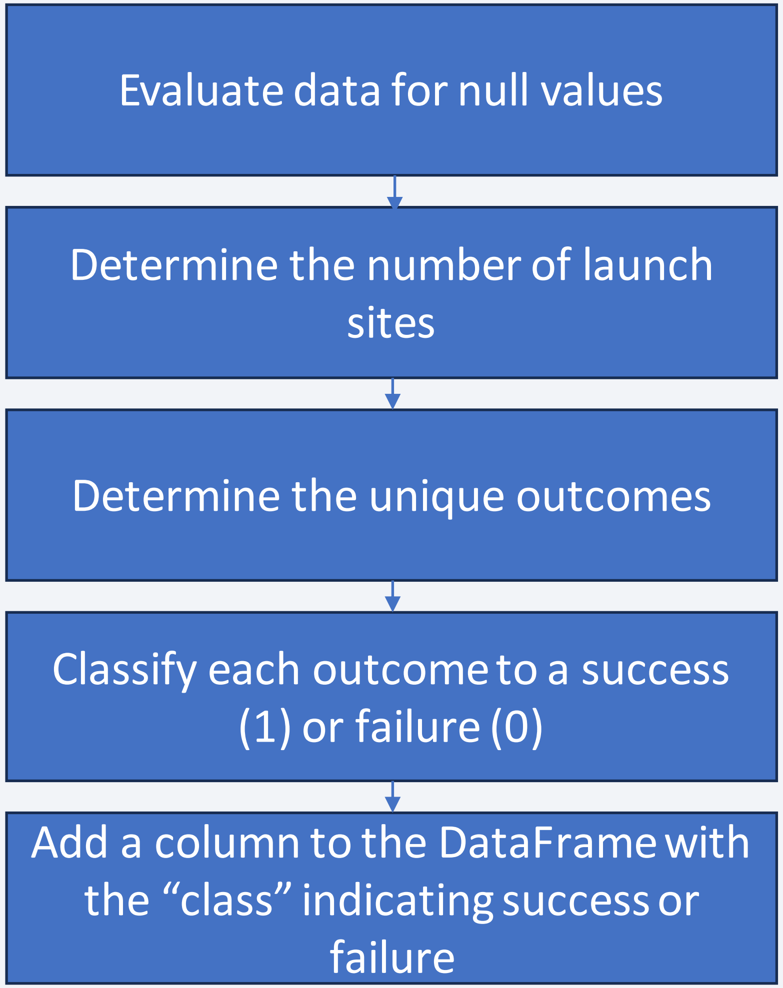
↓

Generate a pandas DataFrame from the dictionary object

View Jupyter Notebook on GitHub

Data Wrangling

This flowchart demonstrates the process we undertook when wrangling the data for this project



[View Jupyter Notebook on GitHub]

EDA with Data Visualization

During the EDA process, we plotted the following:

Chart Type	Feature 1	Feature 2	The Purpose of This Chart is to show…
Scatter	Flight Number	Payload Mass	The payload mass changed over time, as the flight numbers are sequential
Scatter	Flight Number	Launch Site	The use of launch sites has changed over time, as the flight numbers are sequential
Scatter	Payload Mass	Launch Site	If some launch sites were chosen based on their payload mass
Bar	Orbit	Success Rate	The success rate of launches based on their orbit
Scatter	Orbit	Flight Number	The flight orbit has changed over time, as the flight numbers are sequential
Scatter	Payload Mass	Orbit	The occurrence of launch payload mass in each orbit
Line	Date	Class	The success rate by month, in chronological order

[View Jupyter Notebook on GitHub]

EDA with SQL

The following queries were used in EDA:

Select distinct landing outcomes

Select 5 launches in sites whose name starts with “CCA”

Select total payload mass where NASA was the customer

Select average payload mass for v1.0 Falcon 9 launches

Select date of the earliest successful ground pad launch

Select boosters that have successfully landed on a drone ship with a payload mass between 4000 and 6000 kg

Select the count of missions with each mission outcome

Select boosters that have carried the maximum payload

Select month, landing outcome, booster version, and launch site for launches in 2015

Select outcomes and their count in descending order between June 4, 2010 and March 20, 2013

View Jupyter Notebook on GitHub

Build an Interactive Map with Folium

Folium Map Objects:

- Launch Sites
  - CircleMarker – marks each launch site with a pin-style marker with a circle around the point.
  - Marker – Provides a popup description of the marker when clicked on
  - MarkerCluster – a group of markers at the launch site that show the successful (green) and failed (red) launches at that site when clicked
- Landmarks
  - DivIcon – a label that shows the distance that was measured from the nearest launch site to this position
  - PolyLine – a line connecting the launch site to the landmark

[View Jupyter Notebook on GitHub](#)

Build a Dashboard with Plotly Dash

Dashboard Characteristics (Features and Interactions):

- Dropdown - options to view all sites, or select a specific site
- Pie chart – automatically updates when the Launch Site dropdown is changed.
  - When all sites are selected, it will compare the successful launches for the different sites
  - When a specific site is selected, it will compare success vs failure for that site
- Slider – allows restricting the payloads that will be shown on the scatter plot
- Scatter Plot – automatically updates when the Launch Site dropdown OR the Payload Slider are changed. It displays the launches for the given site(s) and the payloads within the range selected

[View Jupyter Notebook on GitHub]

Predictive Analysis (Classification)

Model Selection was done by comparing accuracy among different estimators with a scaled, common training and testing data set. Accuracy was also compared visually by producing a confusion matrix.

Estimators used:

- Logistic Regression
- SVC
- Decision Tree
- K-Nearest Neighbor

Capture "y" as the class column in the data

↓

Capture and scale "X" as the other features of the data

↓

Split training (80%) and testing data (20%)

↓

Test each estimator using GridSearchCV

↓

Visually confirm the confusion matrix to determine the most accurate

[View Jupyter Notebook on GitHub]

Results

The following slides will show the following results:

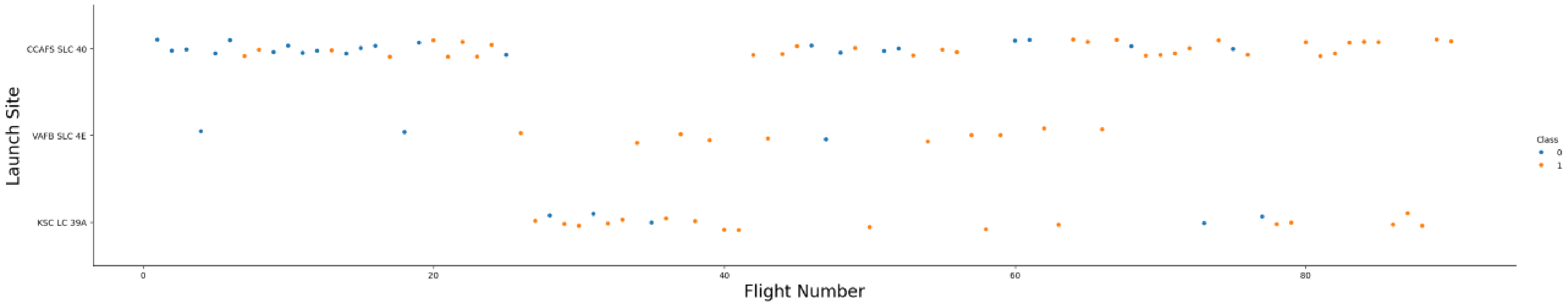
EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

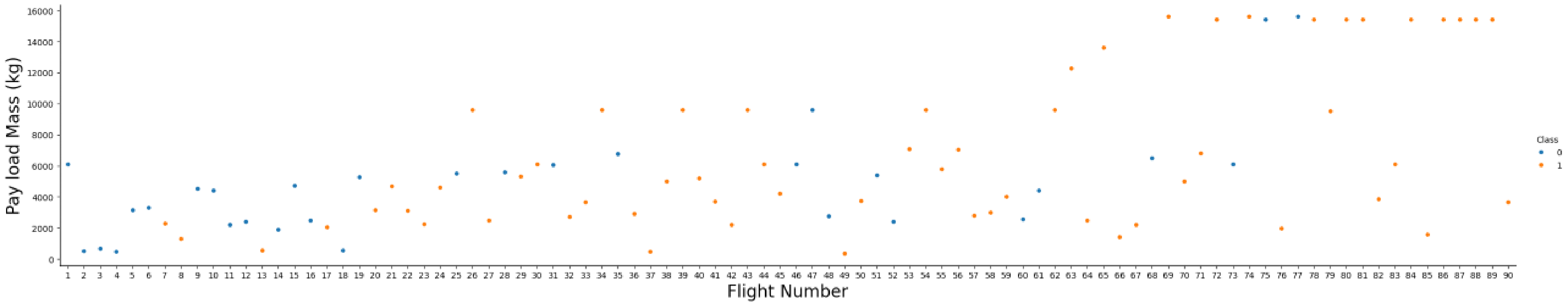
Section 2

Insights drawn from EDA



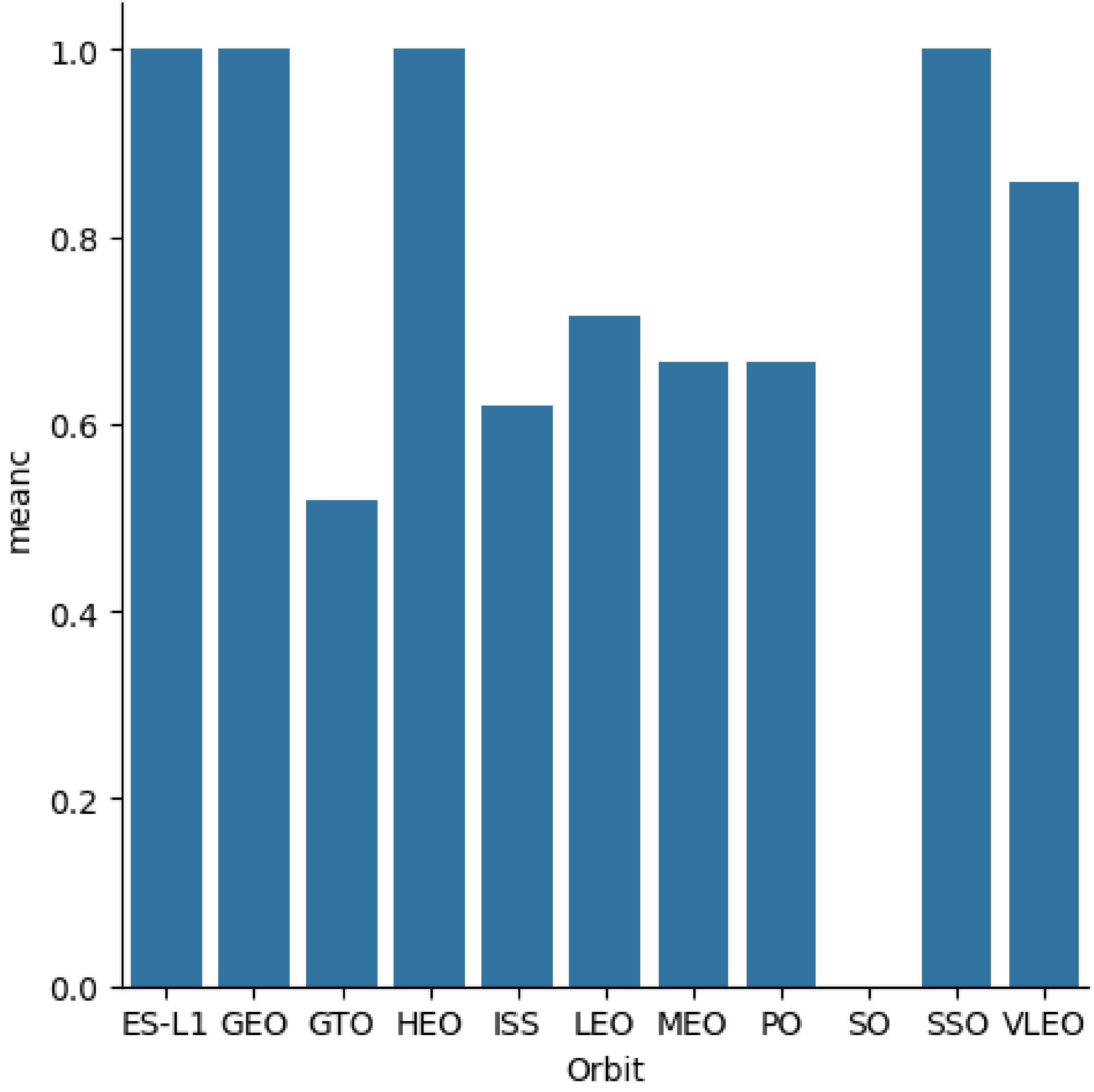
Flight Number vs. Launch Site

This scatter plot shows the occurrence of launches and their location in chronological order.



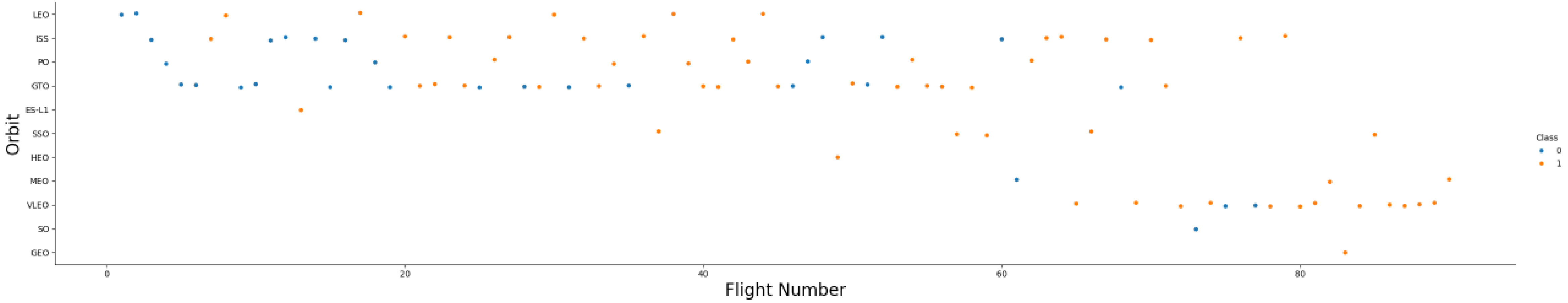
Payload vs. Launch Site

This scatter plot shows the size of the rocket pay load in chronological order.



Success Rate vs. Orbit Type

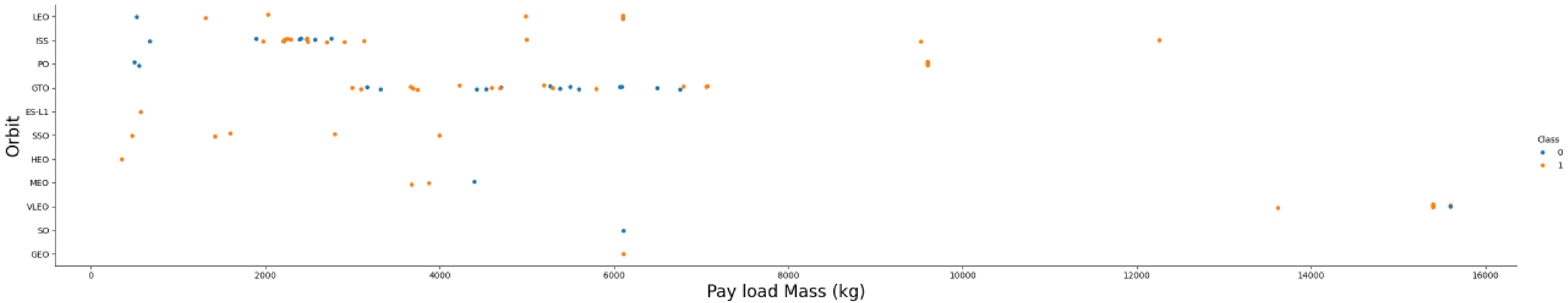
This chart shows the success rate of each type of launch orbit, where 1.0 is 100%



Flight Number vs. Orbit Type

This scatter plot shows the orbit of each flight in chronological order, with the color indicating whether the launch had succeeded (red) or failed (blue)

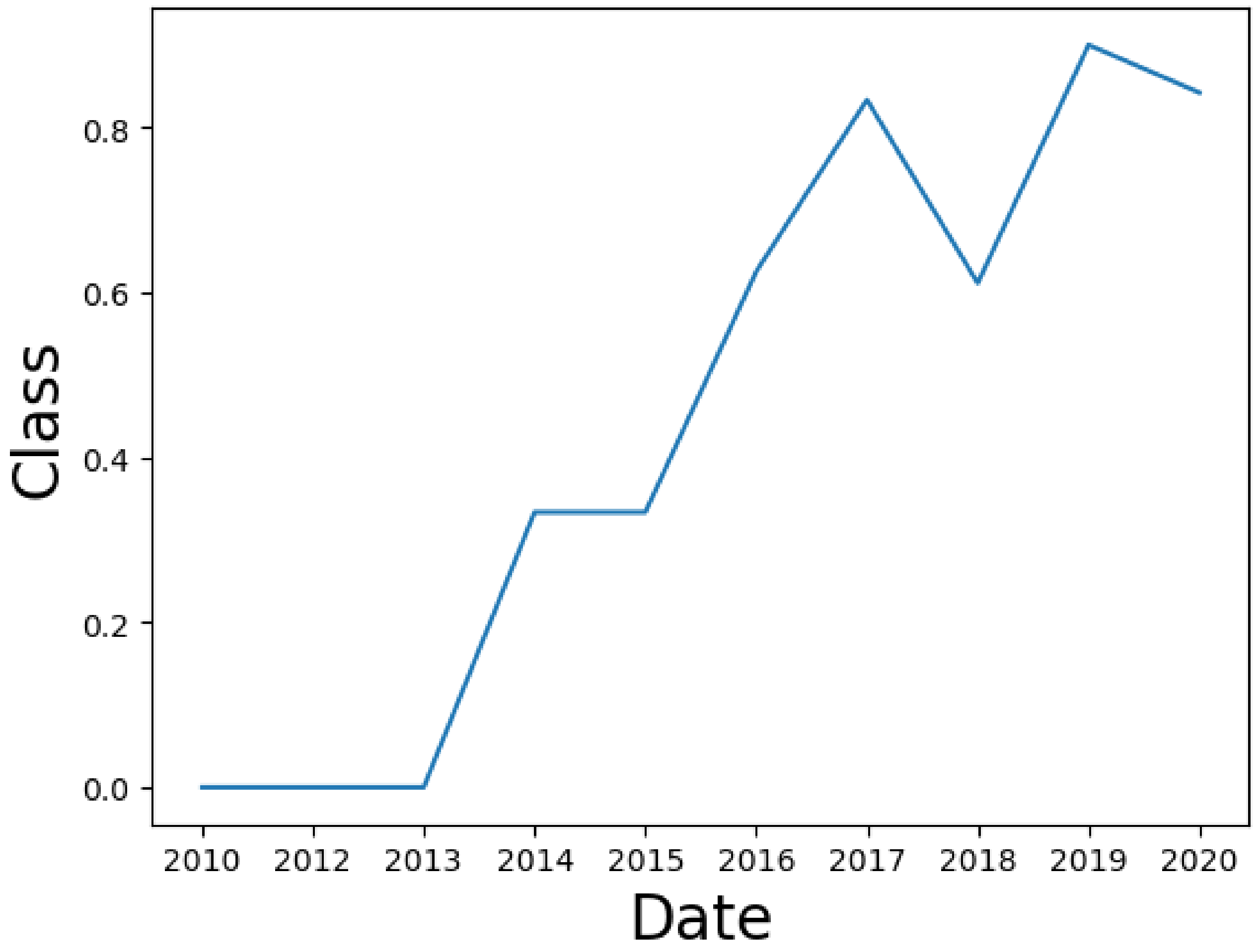
Payload vs. Orbit Type



This scatter plot shows the orbits achieved by different payload masses.

Launch Success Yearly Trend

This chart shows that the success rate of launches has increased over time, despite a noticeable dip in 2018



All Launch Site Names

This query shows a unique list of Launch Site names in the data

```
[8]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

[8]:

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
```

[12] Python

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

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

This query shows five rows whose Launch Site name begins with CCA

Total Payload Mass

Task 3

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") as total FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

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

total
45596

This query calculates and returns the total payload mass for rockets launched for NASA

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.0%"
```

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

AVG(PAYLOAD_MASS__KG_)
340.4

This query calculates and returns the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

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

Hint:Use min function

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

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

MIN("Date")
2015-12-22

This query locates the date of the first successful landing outcome on ground pad

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

```
%%sql
SELECT Booster_Version, Landing_Outcome, 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	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000tion here

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

This query calculates the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

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

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

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

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

This query lists the names of the booster which have carried the maximum payload mass

2015 Launch Records

This query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5) = '2015'
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- This query ranks 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

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

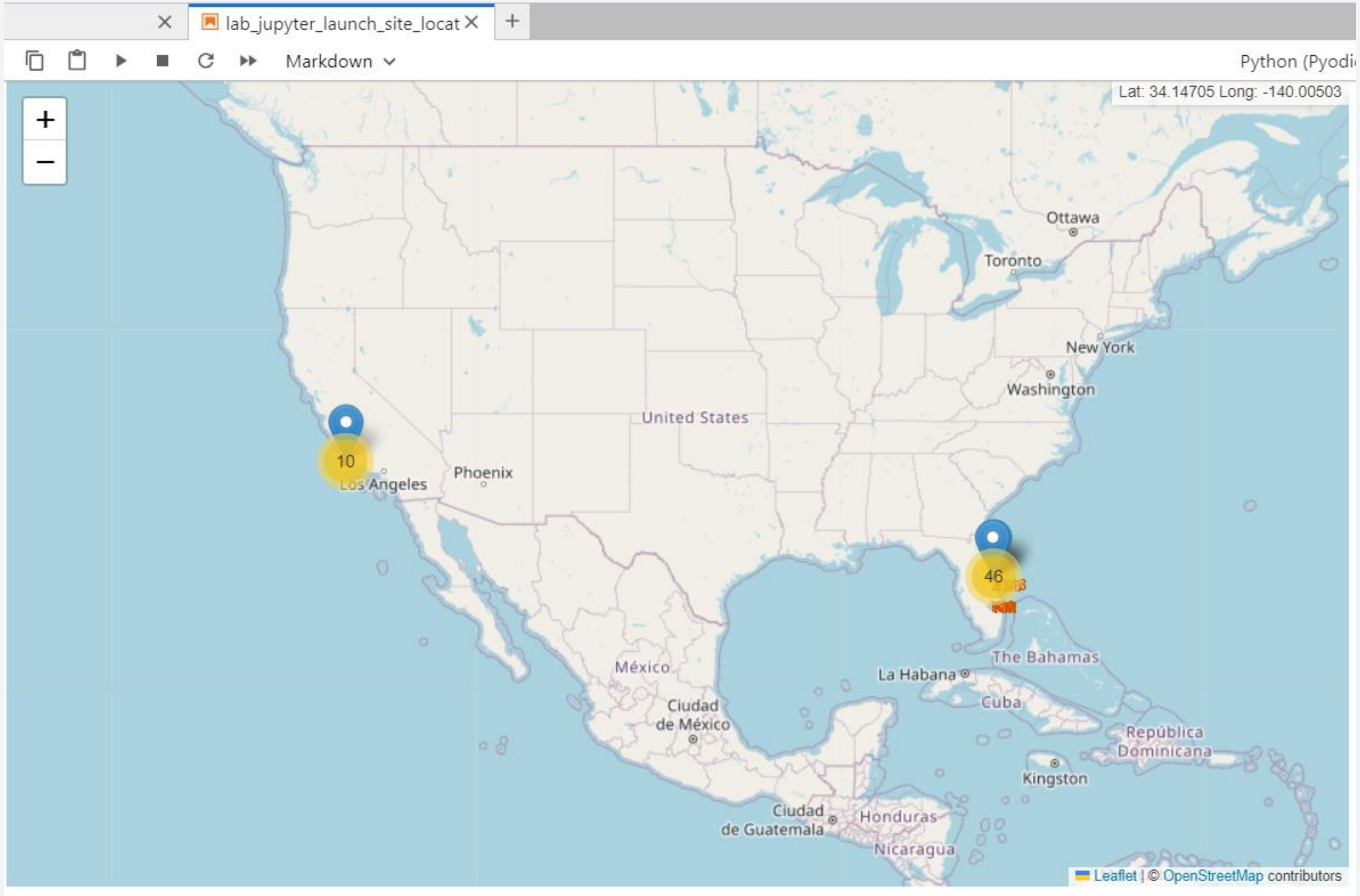
```
 * sqlite:///my_data1.db
Done.
```

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

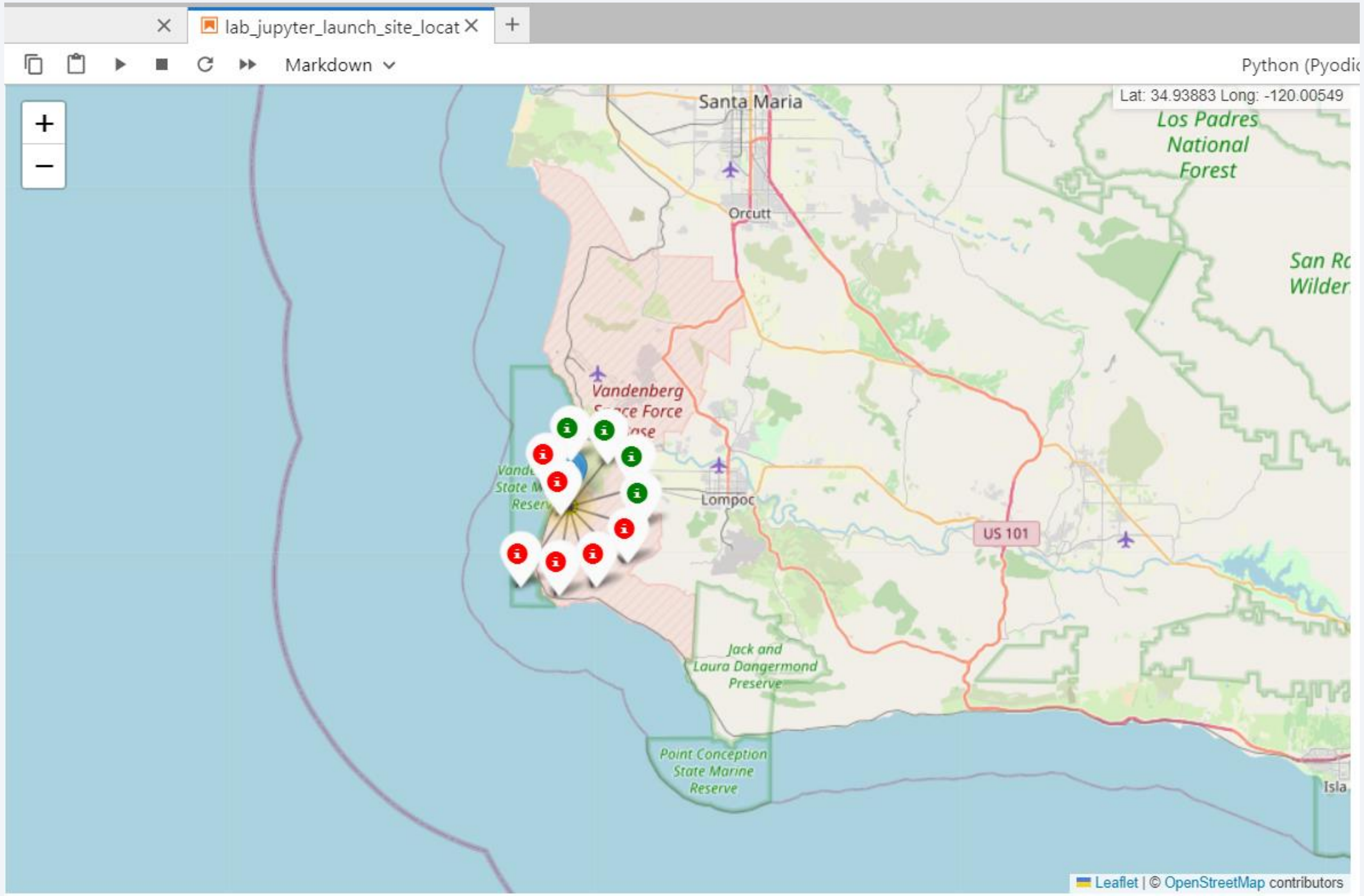
All SpaceX Launch Sites



The zoomed-out view of the generated Folium map shows:

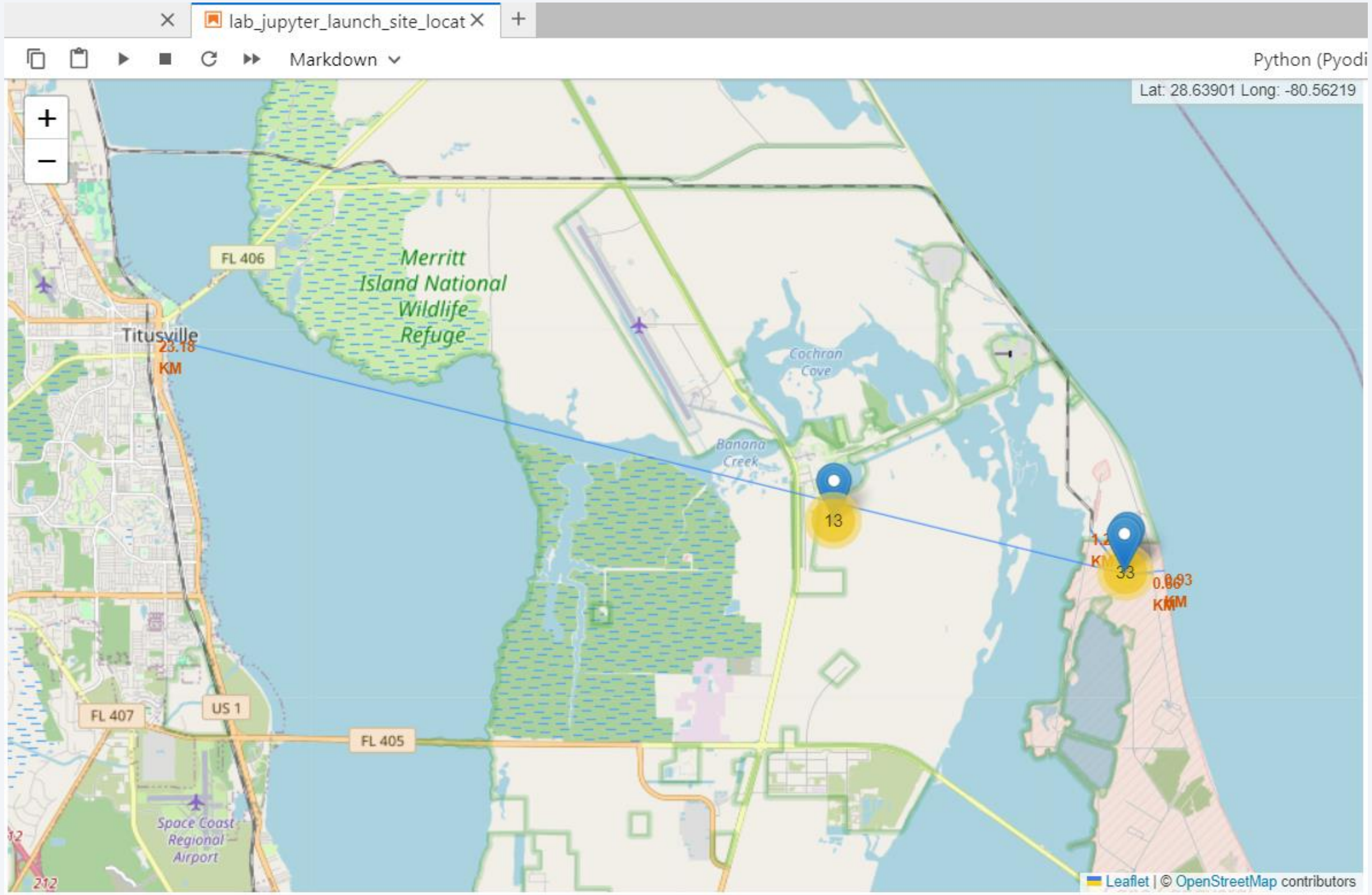
- The location of sites where SpaceX launches initiated
- The number of launches at each site

Launch Outcomes of Vandenburg Space Force Base



The attached map shows the launch outcomes of Vandenburg Space Force Base

Nearest Landmarks to Cape Canaveral Space Center



The attached map shows distances from Cape Canaveral Space Center to the nearest:

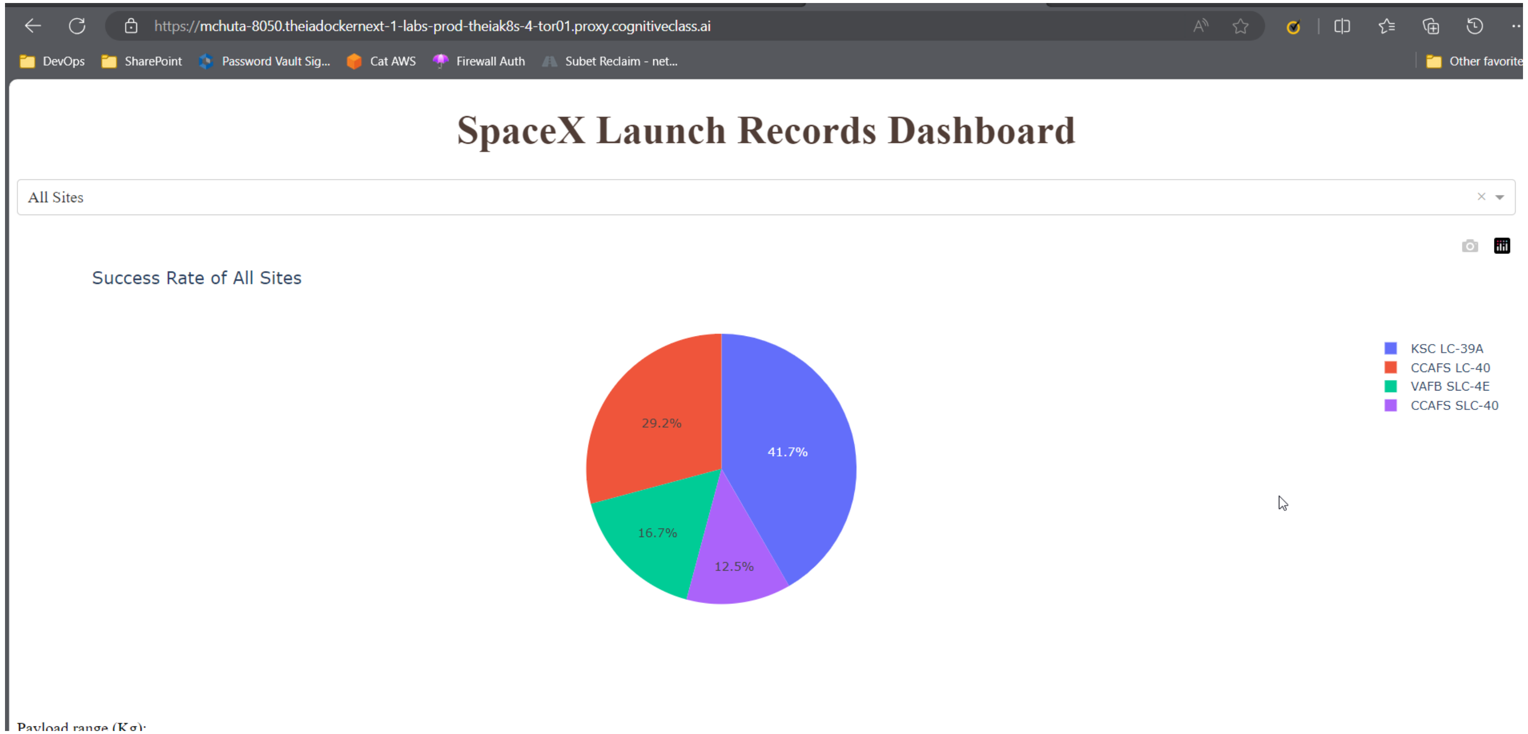
- Coastline (0.93km)
- Highway (0.66km)
- Railroad (1.26km)
- City (23.18km)

These demonstrate that there is significant infrastructure to support the launch sites, but they are not located near cities.

Section 4

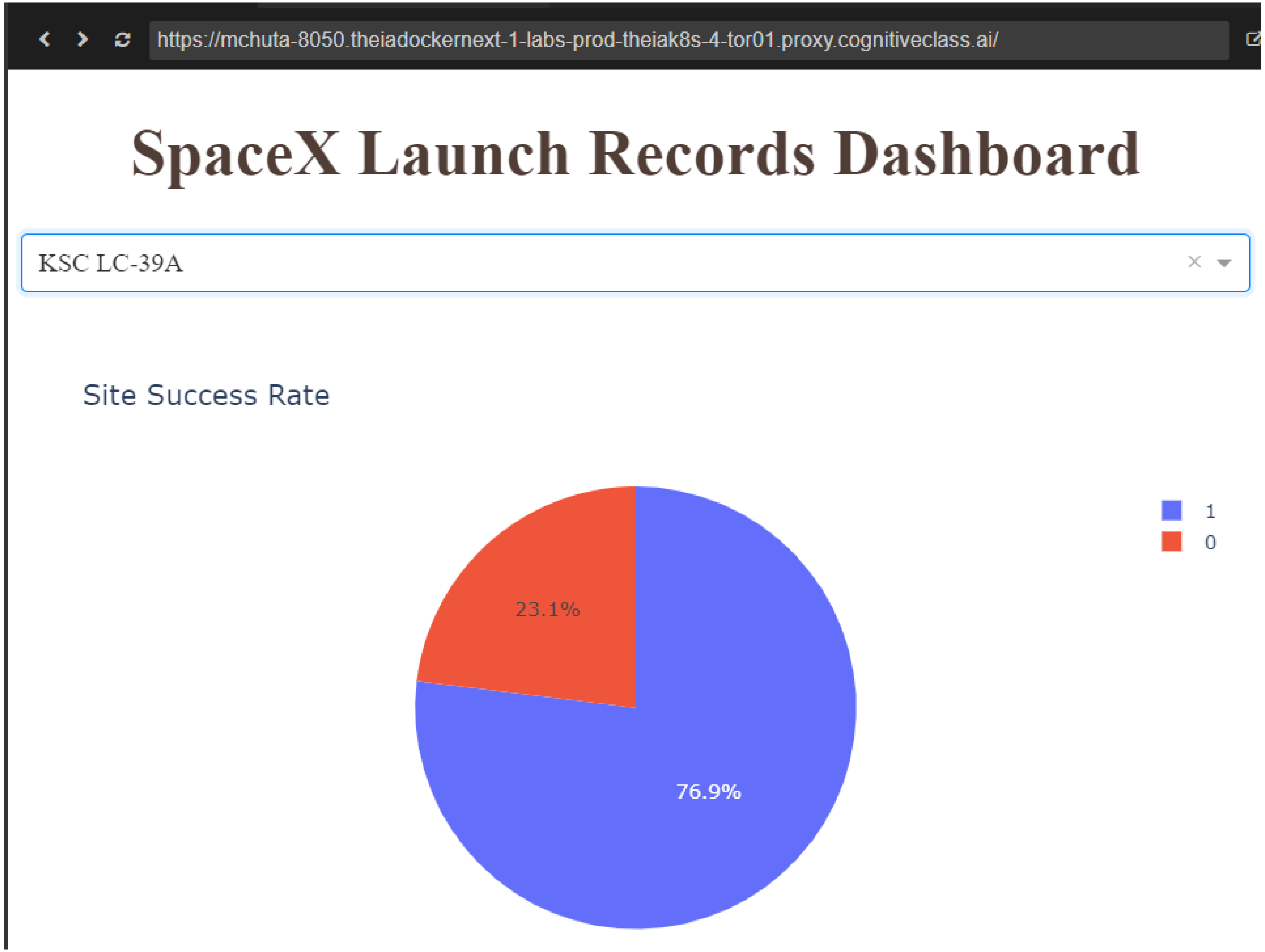
Build a Dashboard with Plotly Dash

Success Rate of All Sites



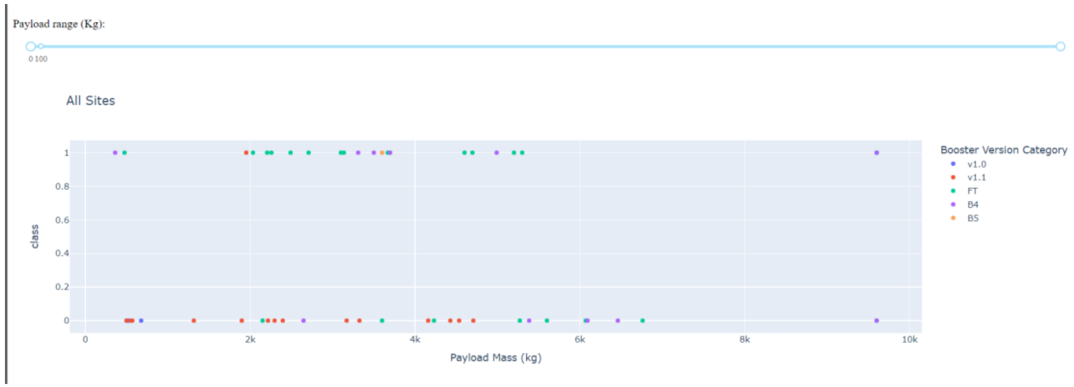
The data show that the Kennedy Space Center (KSC LC-39A) had the most successful launches

Most Successful Site – Pie Chart



The data show that the Kennedy Space Center (KSC LC-39A) had the most successful rate of launches.

Payload Mass (kg) vs Success Rate by Booster Version



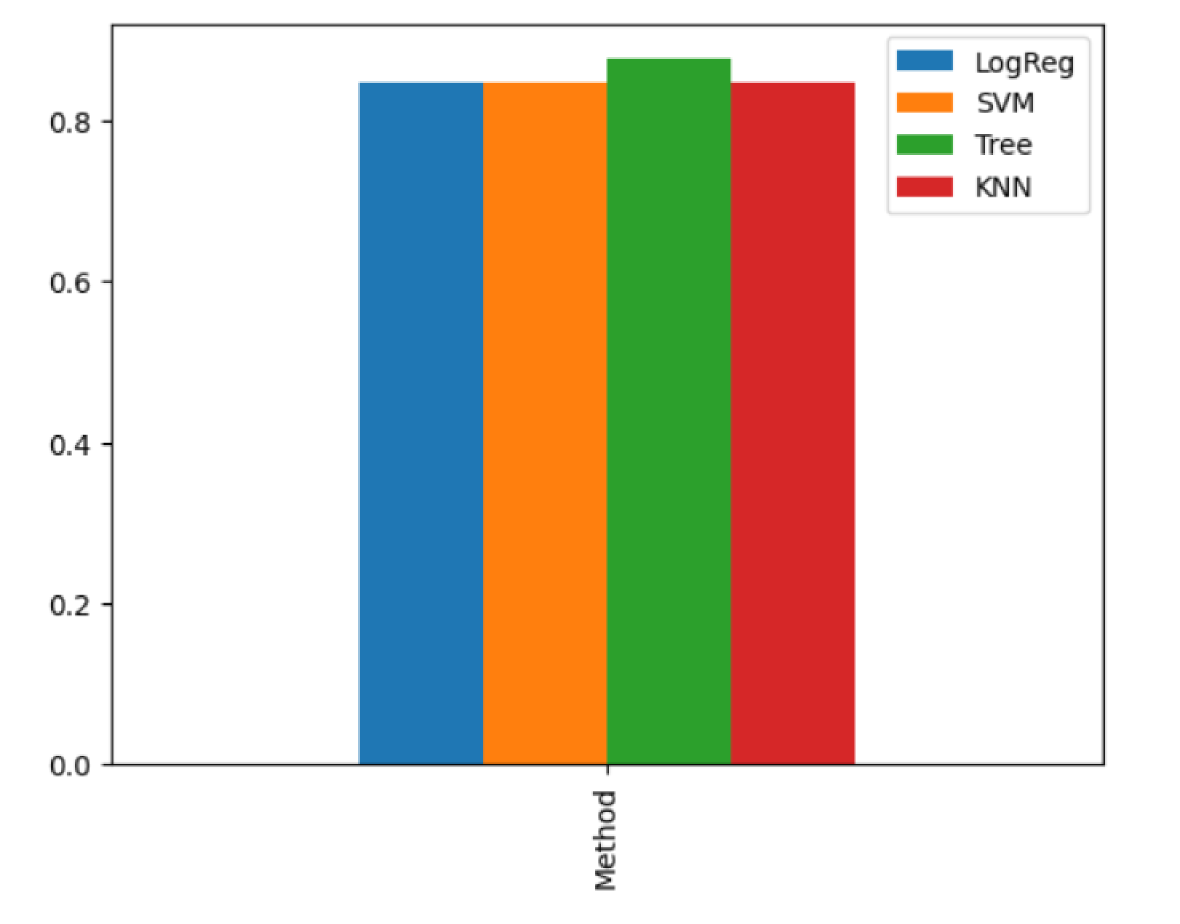
The data show the relationship between Payload Mass and Success Rate. They also show that there are no launches for the v1.0 Booster Version with Payload Mass between 2000kg and 7000kg

Section 5

Predictive Analysis (Classification)

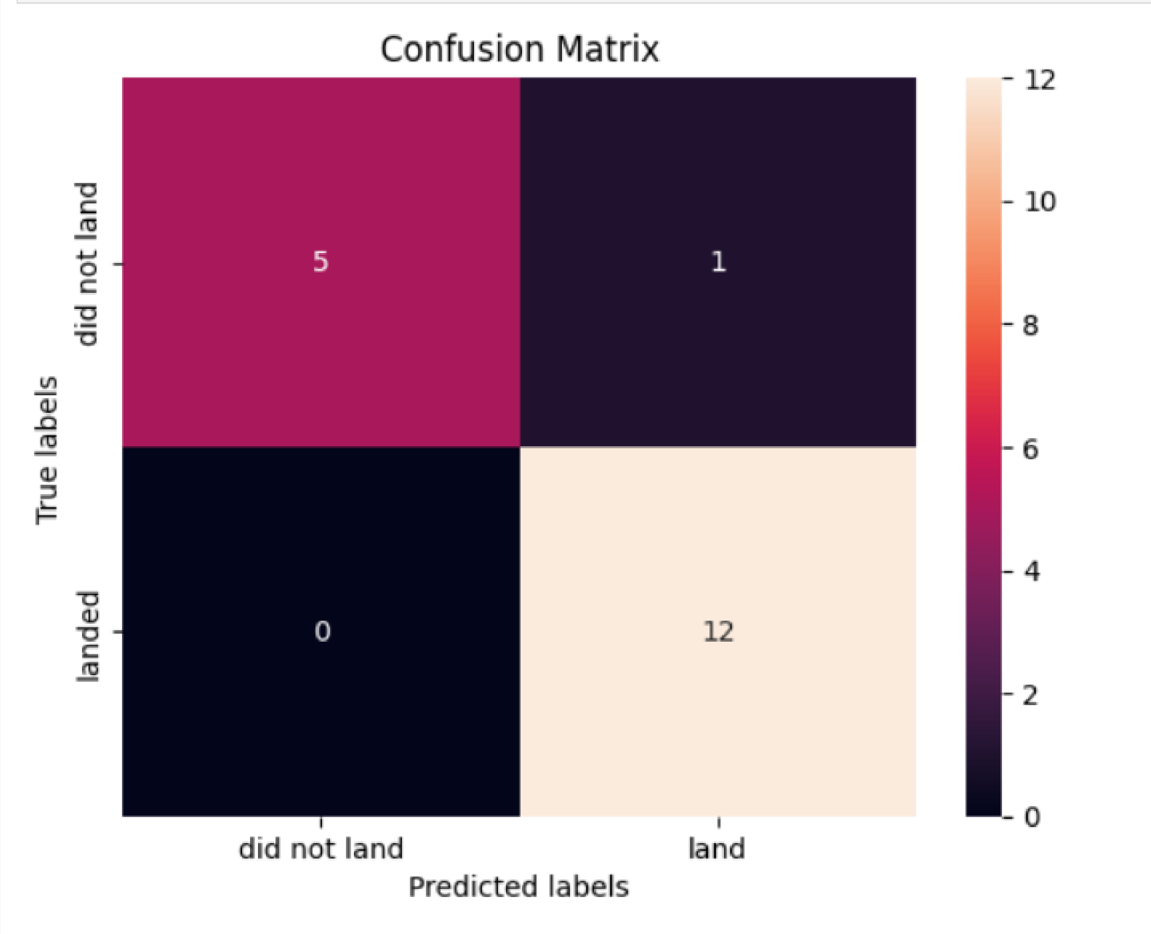
Classification Accuracy

The classification method with the highest accuracy was the Decision Tree, as shown in the bar chart to the right.



Confusion Matrix

The classification method with the most accurate confusion matrix is shown here. It was produced by the Decision Tree method.



Conclusions

- SpaceX launches were most success from Kennedy Space Center in Florida
- Launch sites are supported by heavy infrastructure (railroad and highways), near coastlines, and far from nearby population centers (cities)
- A decision tree machine learning classification method was found to be the most accurate in predicting successful missions and performed the best on our test data.
- Except for a slight dip in 2018, launches have been increasingly successful over time.
- NASA has sent over 45 thousand kilograms of cargo into space by using SpaceX rockets

Thank you!