IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

George Y Li
2023.07.28

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Building interactive maps with Folium
  - Building a Dashboard with Plotly Dash
  - Machine Learning Prediction Analyisis
- Summary of all results
  - EDA results
  - Interactive Analytics
  - Predictive Analysis

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million.
- This is because SpaceX can **reuse** its first stage.
- If the first stage of the Falcon 9 can land, we can determine the cost of a launch, and this information can be used if alternate companies wants to bid against SpaceX for a rocket launch.
- Instead of relying on complicated rocket science, our goal is to predict if the Falcon 9 first stage will land successfully using **Data Science**.

Section 1

Methodology

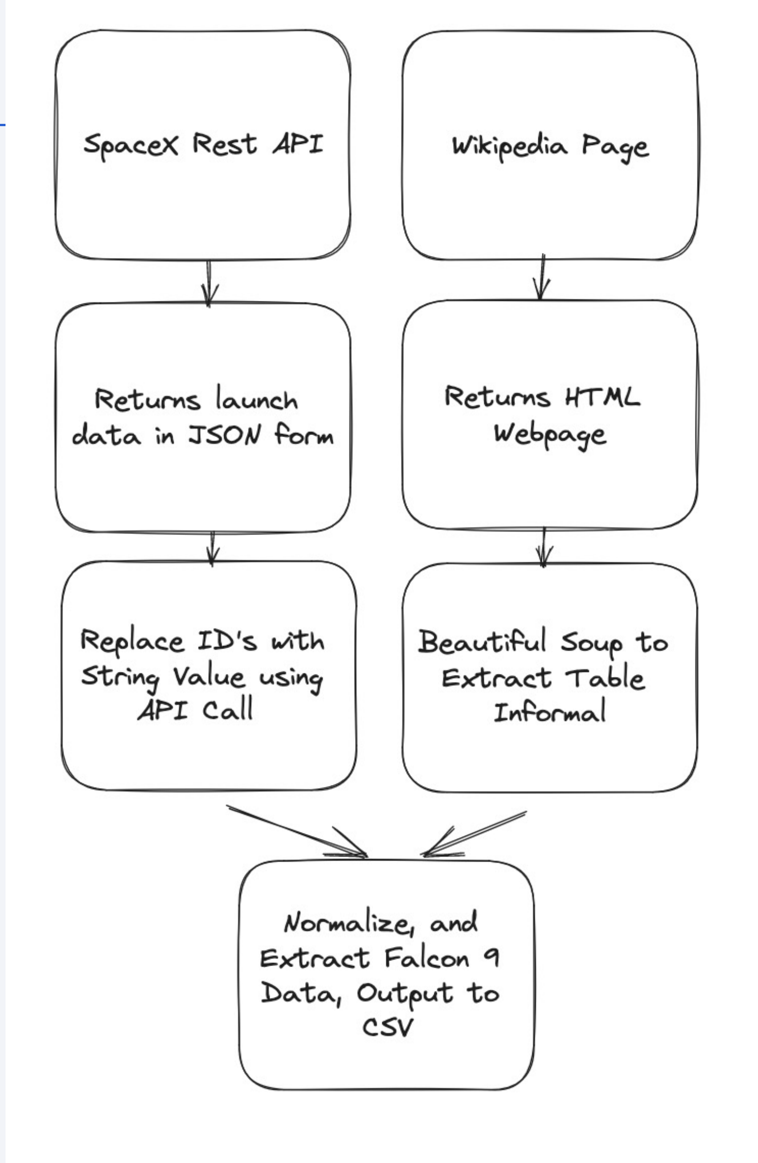
Methodology

Executive Summary

- Data collection methodology:
  - REST API calls with SpaceX API located at https://api.spacexdata.com/v4/
  - Webscraping Wikipedia page at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Data was processed using One-Hot Encoding data fields for Machine Learning algorithms and data cleaning of null values and irrelevant data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models such as LR, KNN, Decision Trees, SVM.

Data Collection

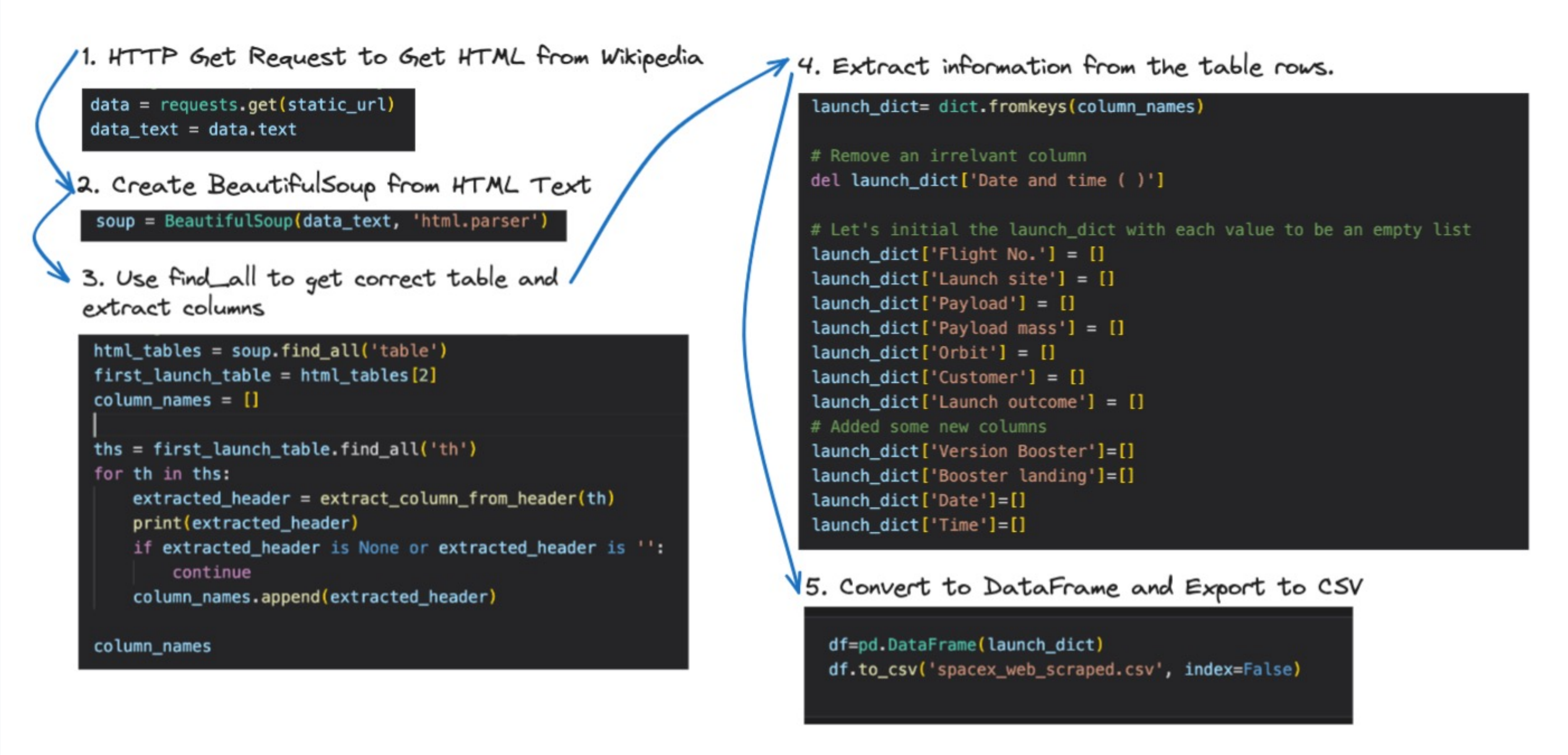
- Data Set 1: SpaceX API
  - SpaceX launch data gathered from SpaceX Rest API.
  - This API provides information about launches, including rocket type, payload delivered, launch specifications, landing specifications and more.
- Data Set 2: Web scraping Wikipedia
  - Beautiful Soup is used to parse Falcon 9 launch data from a Wikipedia article.



Data Collection – SpaceX API

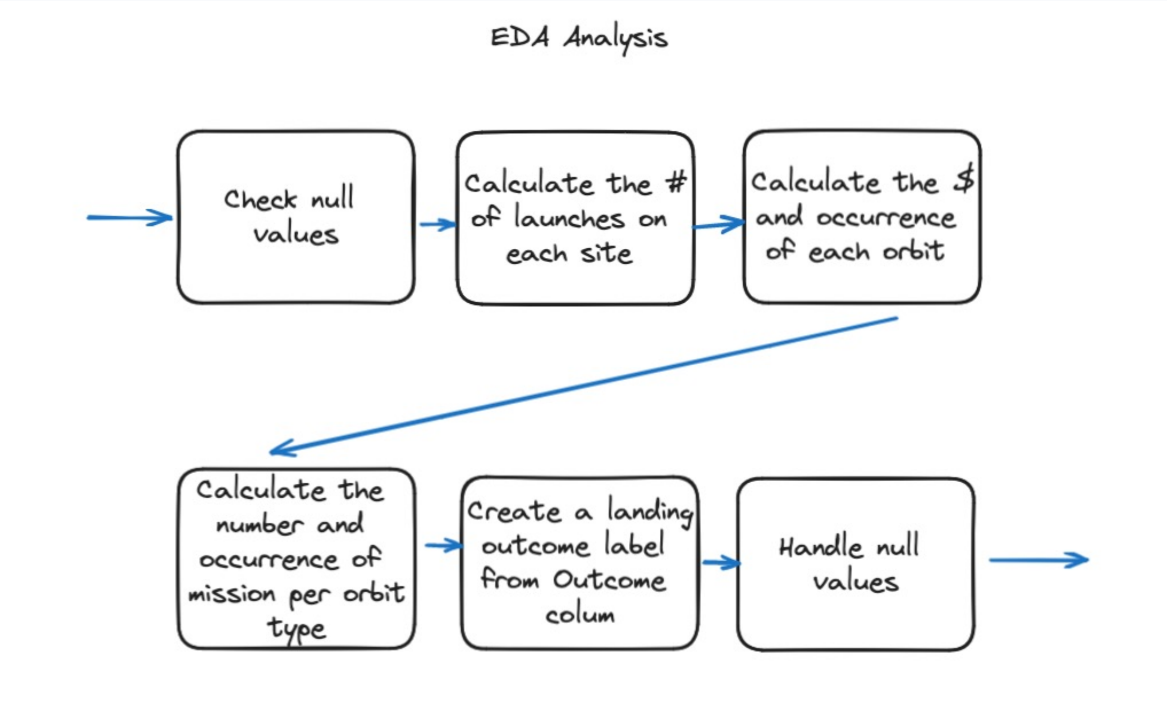
GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/01%20jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/02%20jupyter-labs-webscraping.ipynb

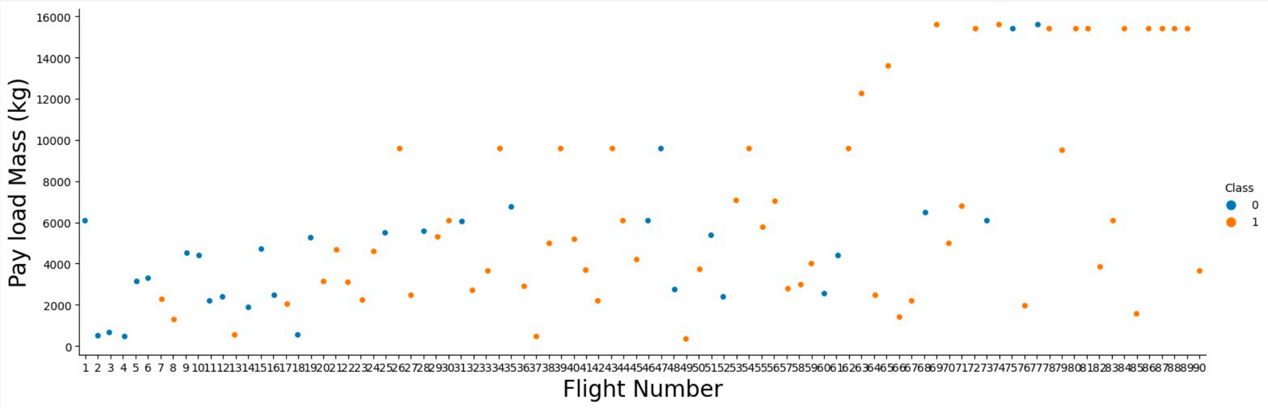
Data Wrangling

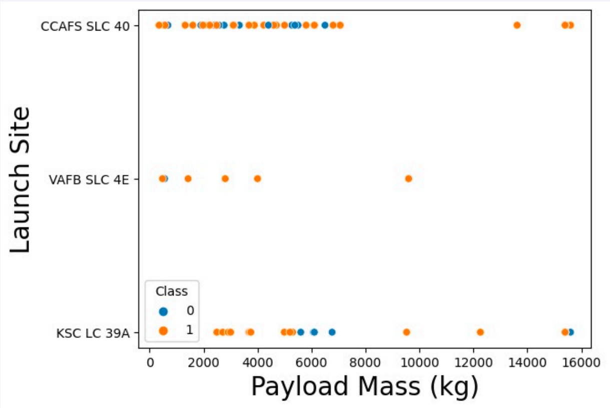


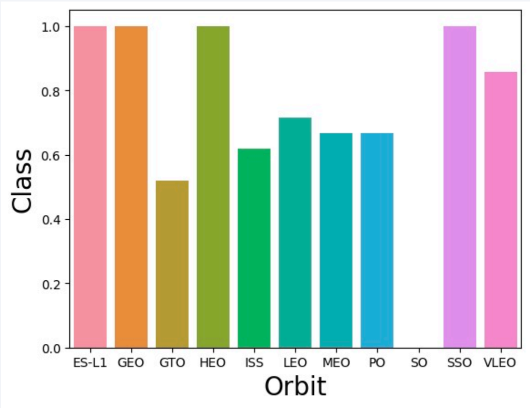
The main goal is to convert outcomes, labeled with “True ASDS”, “False ASDS,” etc into training labels with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful

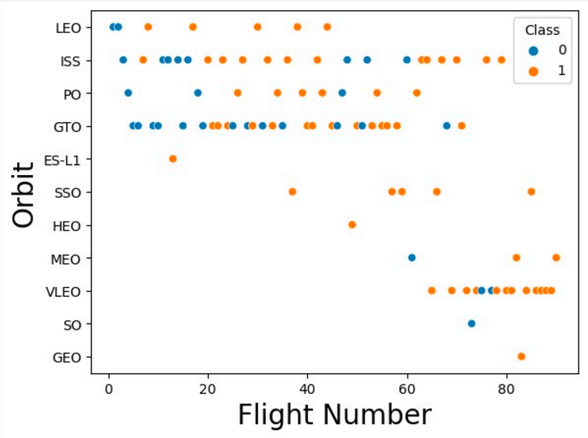
GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/03%20labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

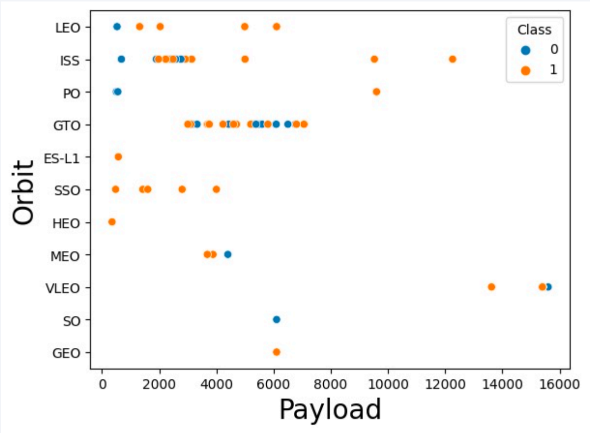
EDA with Data Visualization

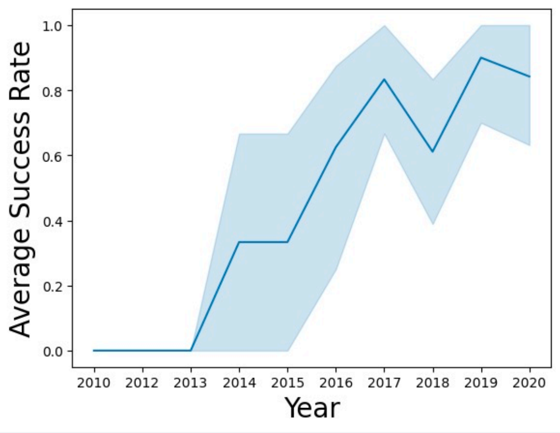












We compare a variety of variables to find correlations.

GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/05%20Assignment%3A%20Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

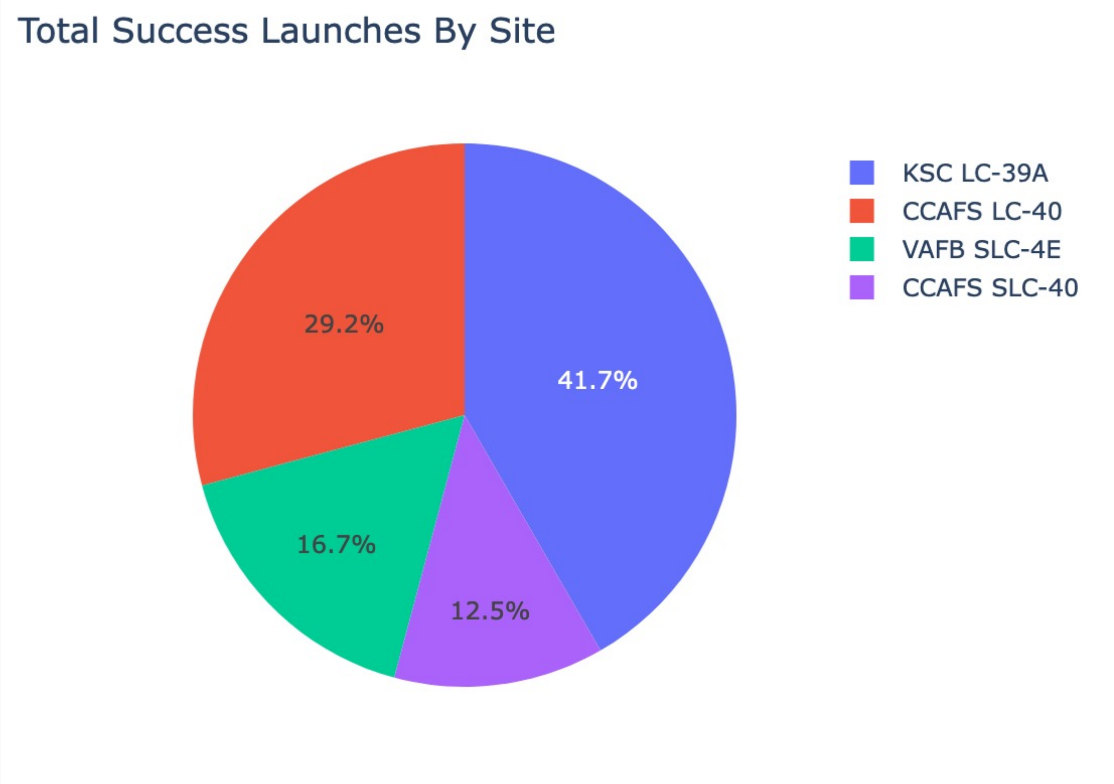
- SQL Queries Performed
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites began with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in 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. Use a subquery.
  - List the records which will display the months, failure landing outcomes, booster versions, launch sites for the months in year 2015.
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/04%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

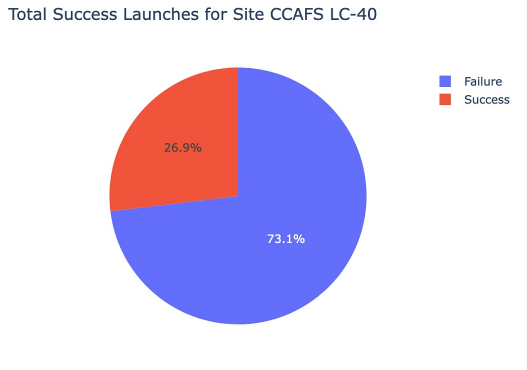
Build an Interactive Map with Folium

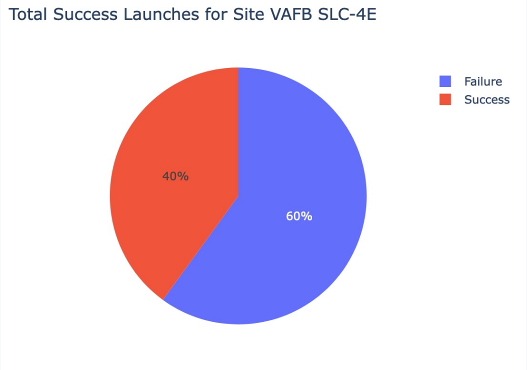
Map objects, including markers, circles, and lines have been added to a Folium Map with the aim of finding an optimal location for successful launches.

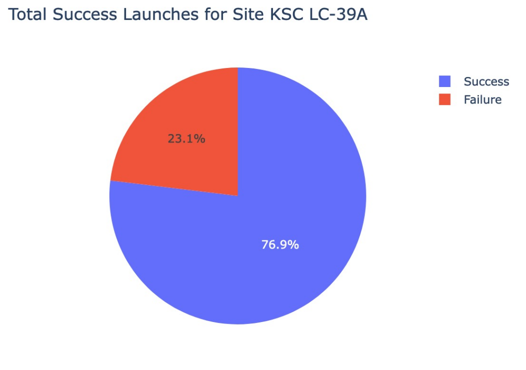
GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/06%20lab_jupyter_launch_site_location.jupyterlite.ipynb

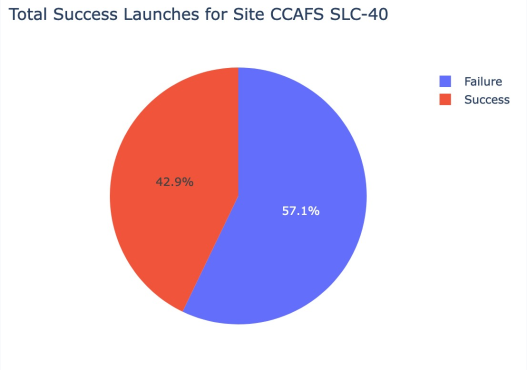
Build a Dashboard with Plotly Dash







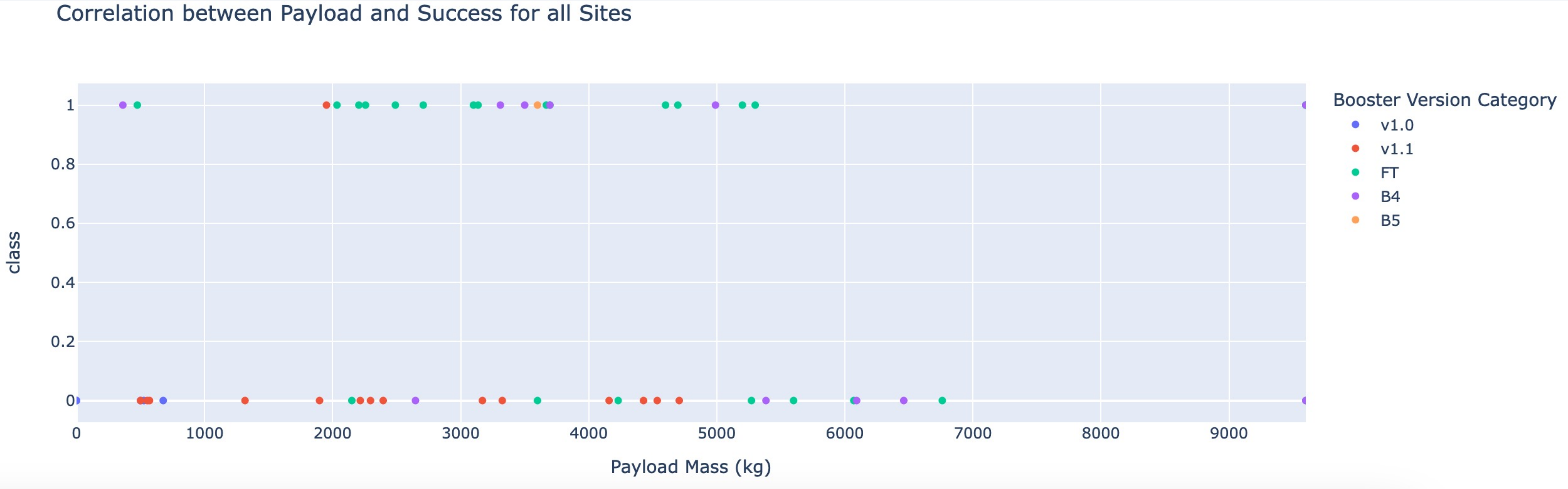




Plotly Dash gives us interactive visualizations. For example, here we can see success rates at all sites and at individual sites.

GitHub: https://github.com/AkaCoder404/data-science-learning/blob/main/IBM_Certificate_Coursera/Unit10/07%20spacex_dash_app.py

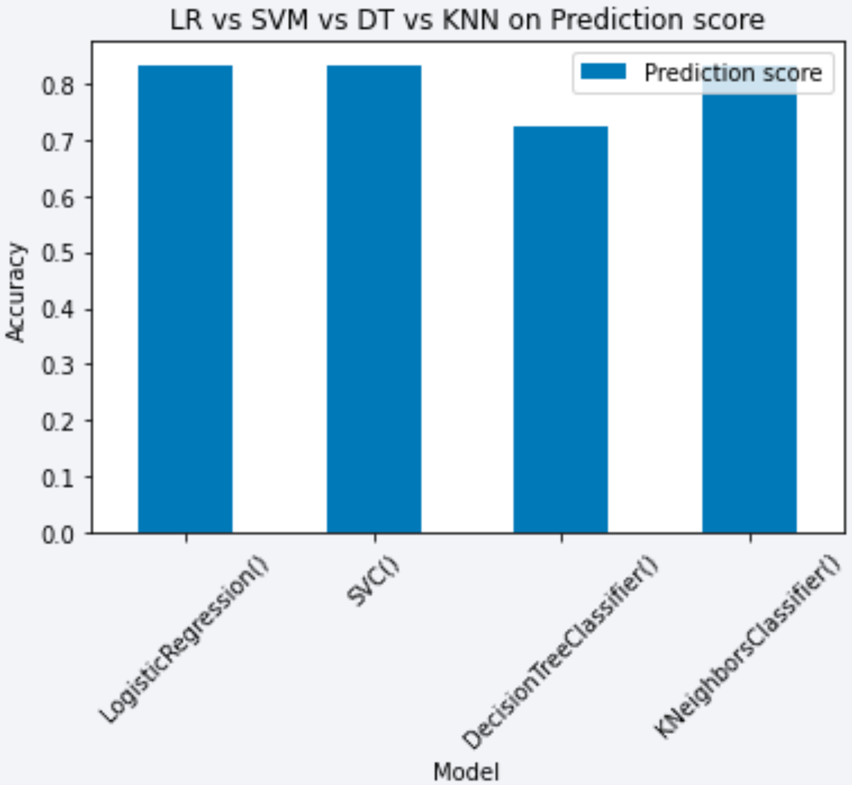
Build a Dashboard with Plotly Dash



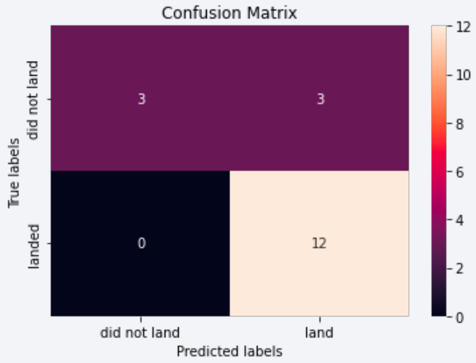
We also visualize the correlation between Payload mass and Success.

Predictive Analysis (Classification)

- The SVM, KNN, and LR models achieved the highest accuracy at 83.33%.



LogisticRegression



DecisionTree

SVM

KNeighbors

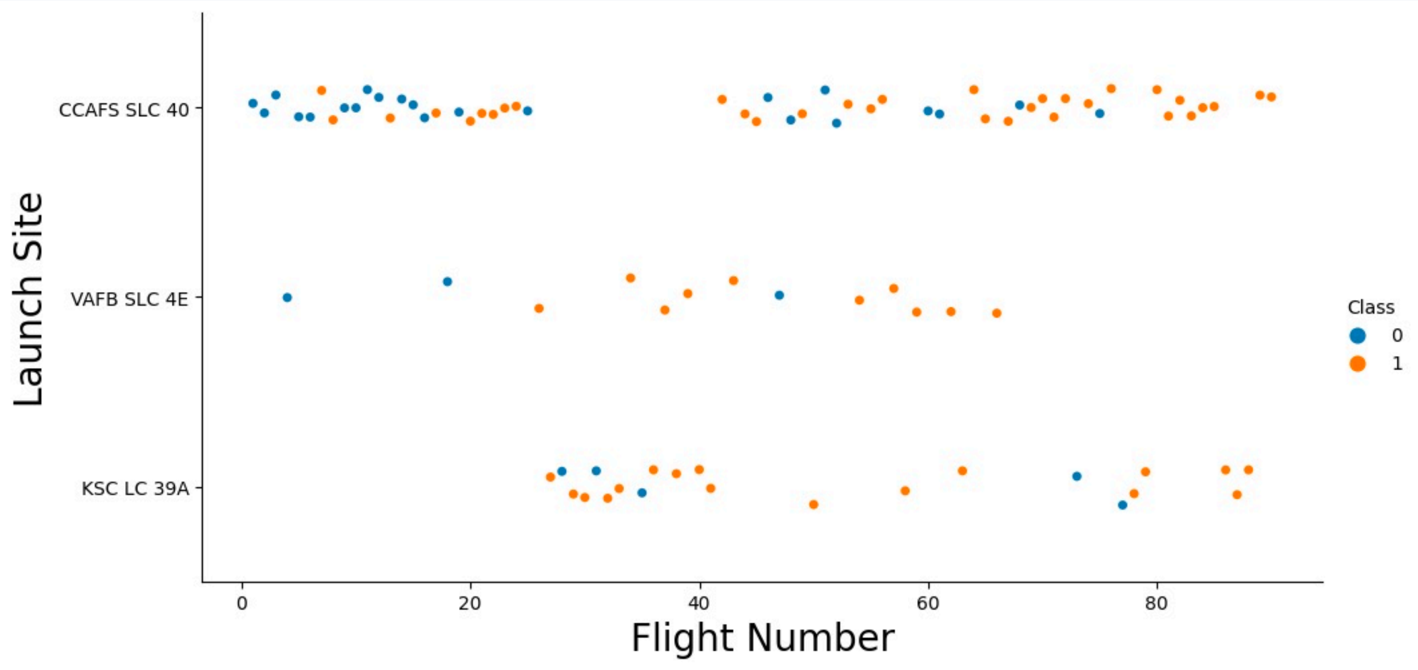
Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than heavier payloads.
- The success rates for SpaceX launches is directly proportional to years. They will eventually perfect the launches.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Section 2

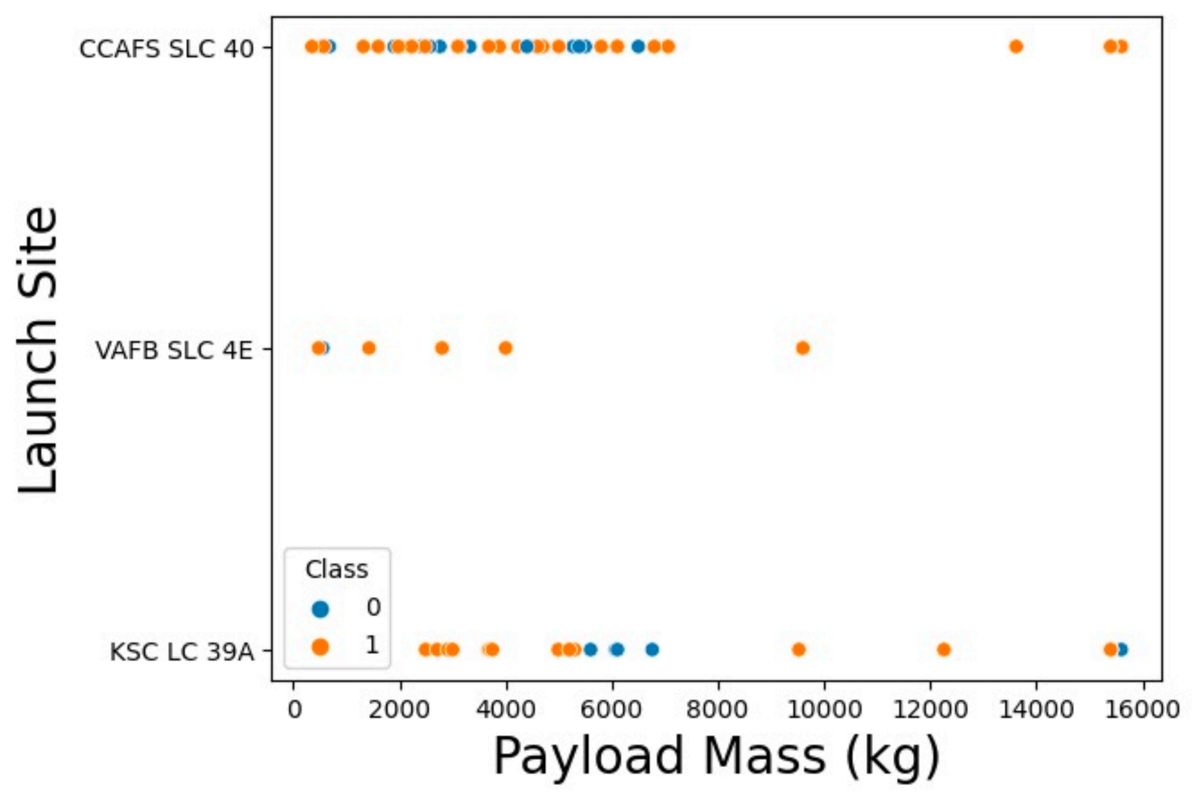
Insights drawn from EDA

Flight Number vs. Launch Site



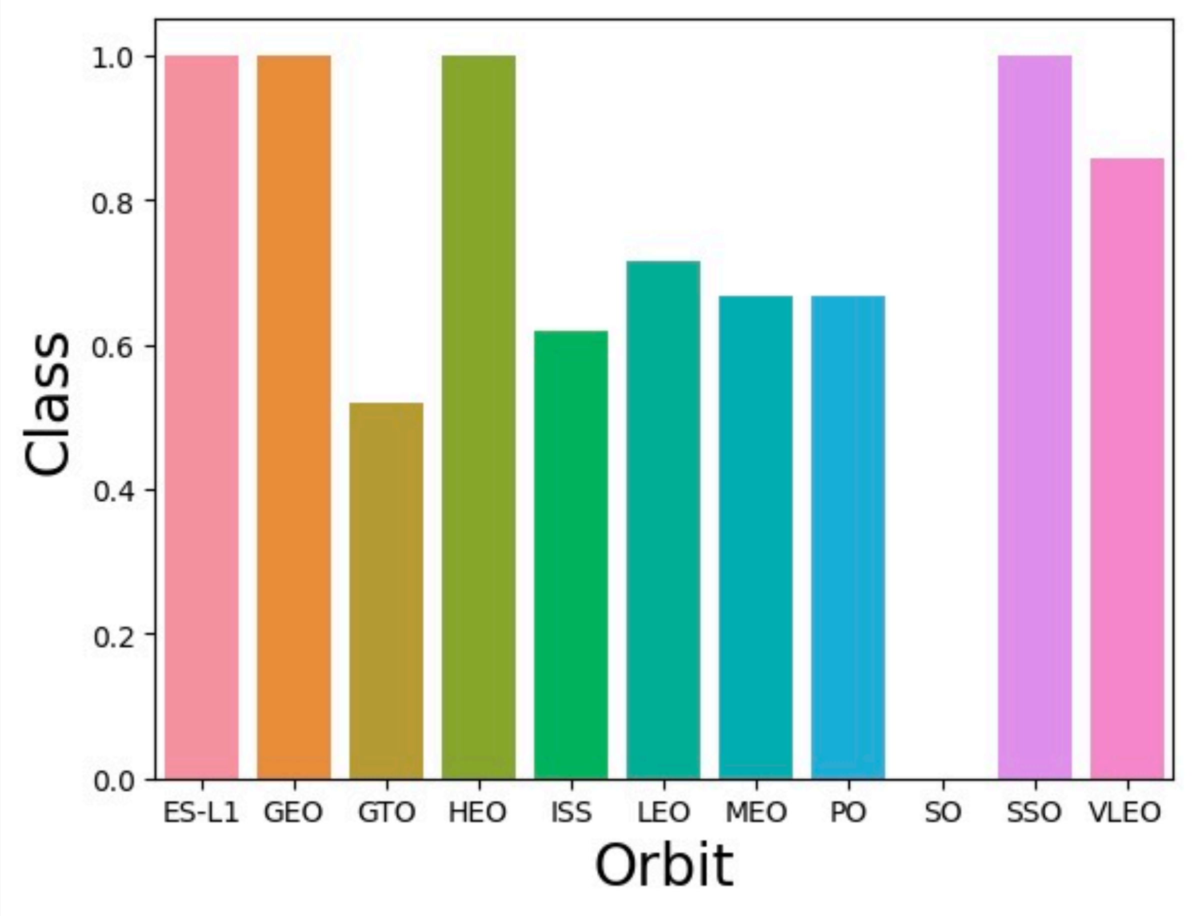
- There are more and more success as time goes on (the more amount of flights.
- There are a lot more flights taken at CCAFS SLC 40.

Payload vs. Launch Site



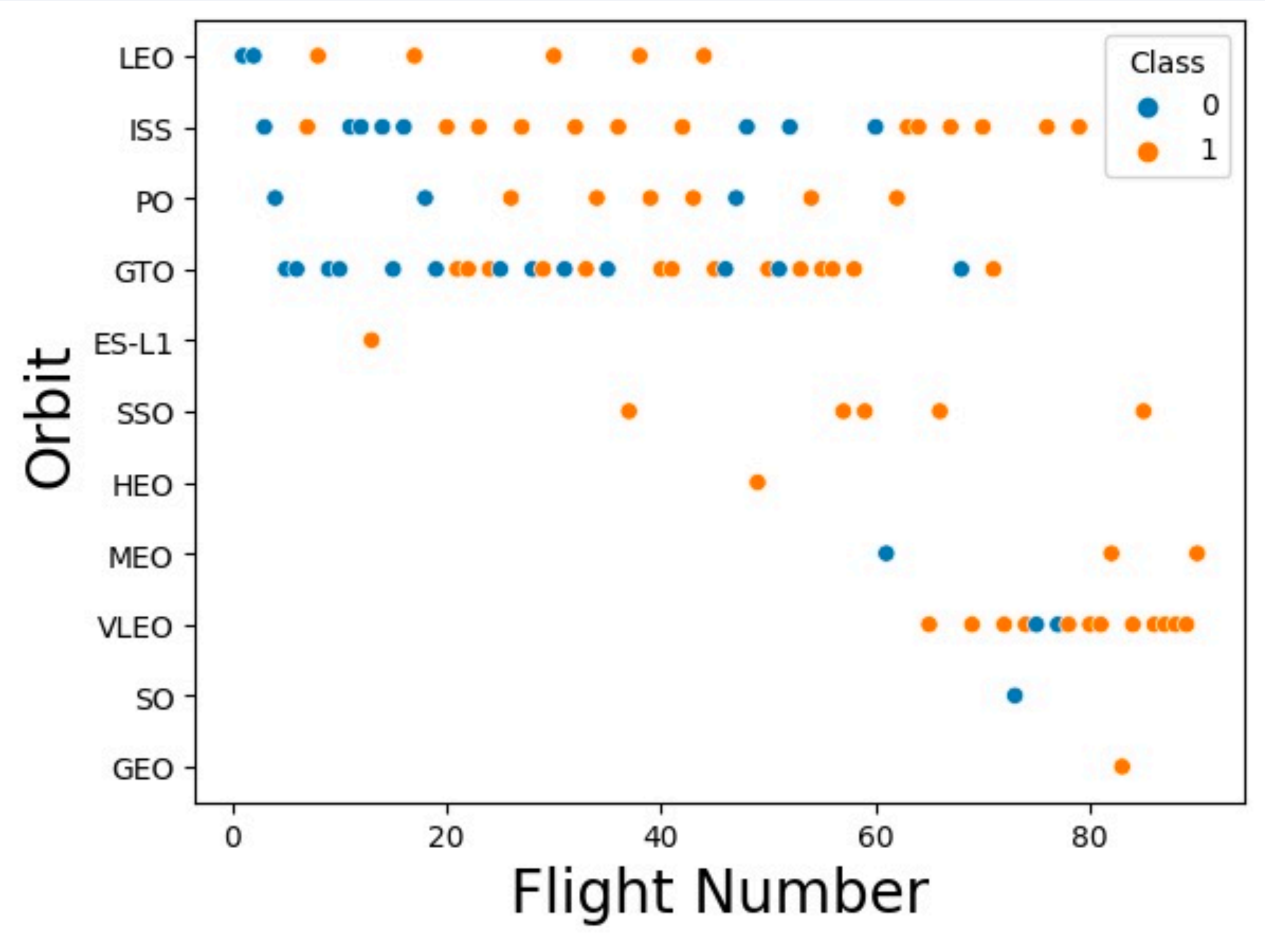
- CCAFS SLC 40 majority of launches are below 7000.
- WAFB SLC 4E does not launch payloads greater than 10000.

Success Rate vs. Orbit Type



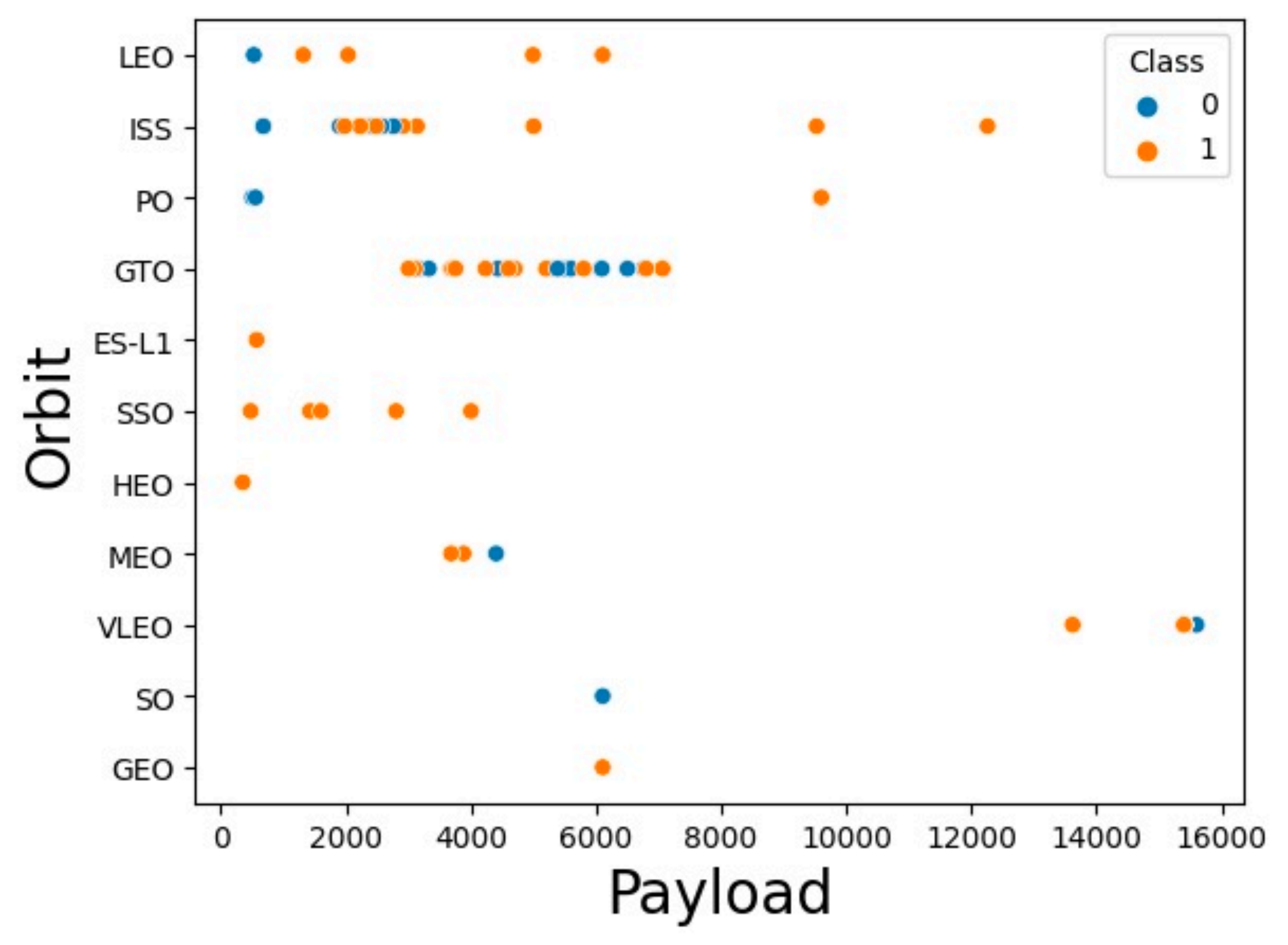
- Orbit types ES-L1, GEO, HEO, and SSO have perfect success rates.
- Orbit type GTO has the lowest.

Flight Number vs. Orbit Type



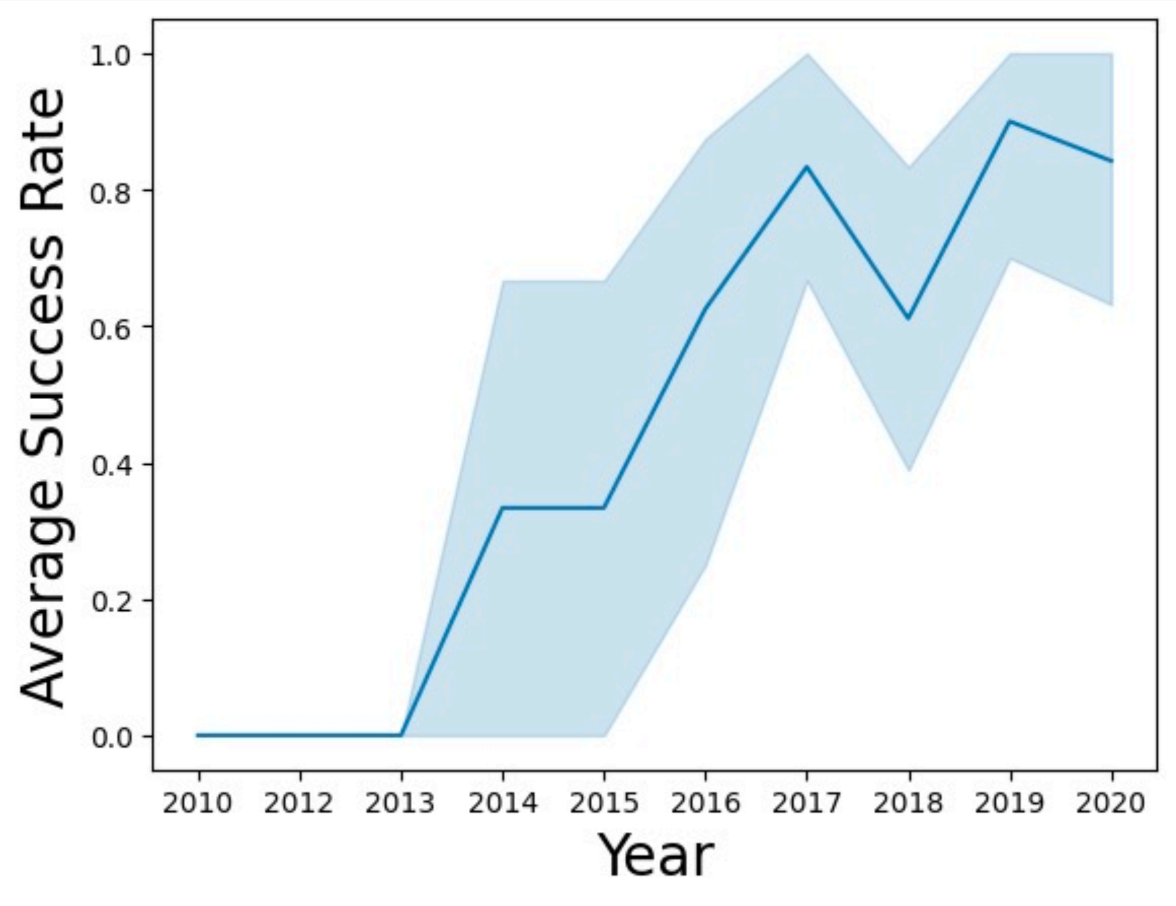
- There is a trend of more VLEO launches in the later years.
- All SSO launches have been successful.

Payload vs. Orbit Type



- There is a strong correlation between Orbit Type ISS and Payload at Payload of 2000.
- There is also a strong correlation between GTO and Payload range of 4000-8000.

Launch Success Yearly Trend



- Launch success rate has increased significantly since 2013.
- There was a sharp dip in 2018, but stabilizes back by 2019.

All Launch Site Names

- Find the names of the unique launch sites.
- There are a total of 4 different launch sites.

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

```
 * 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 spacextbl 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	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- Find 5 records where launch sites begin with `CCA`

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- The total payload mass is 45596.0 KG

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

✓ 0.0s

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

Total_Payload_Mass
45596.0

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") as "AVG_Payload_Mass_by_F9_v1.1"
FROM spacextbl
WHERE "Booster_Version" LIKE "F9 v1.1"
```

✓ 0.0s

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

AVG_Payload_Mass_by_F9_v1.1
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing on ground pad was at 2018-01-08

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

✓ 0.0s

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

Date
01/08/2018

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
- The booster names are F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2

```
%%sql
SELECT "Booster_Version" FROM spacextbl
WHERE "Landing_Outcome" = "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
```

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

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
- There are 100 total values.

```
%%sql
SELECT "Mission_Outcome", COUNT("Mission_Outcome") as "Count"
FROM spacextbl
GROUP BY "Mission_Outcome"
```

✓ 0.0s

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

Mission_Outcome	Count
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass.
- The booster versions are listed here.

```
%%sql
SELECT "Booster_Version"
FROM spacextbl
WHERE "PAYLOAD_MASS__KG_" is (SELECT MAX("PAYLOAD_MASS__KG_") FROM spacextbl)
```

```
 * 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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT substr(Date,4,2) as "Month", "Booster_Version", "Landing_Outcome", "Launch_Site"
FROM spacextbl
WHERE substr(Date,7,4) is "2015"
```

✓ 0.0s

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

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

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
- We can see that there are many landing outcomes of each type between these two dates.

```
%%sql
SELECT "Date", "Landing_Outcome", COUNT("Landing_Outcome") AS "Count"
FROM spacextbl
WHERE substr(Date,7)||substr(Date,4,2)||substr(Date,1,2)
BETWEEN '20100604' AND '20170320'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC
```

✓ 0.0s

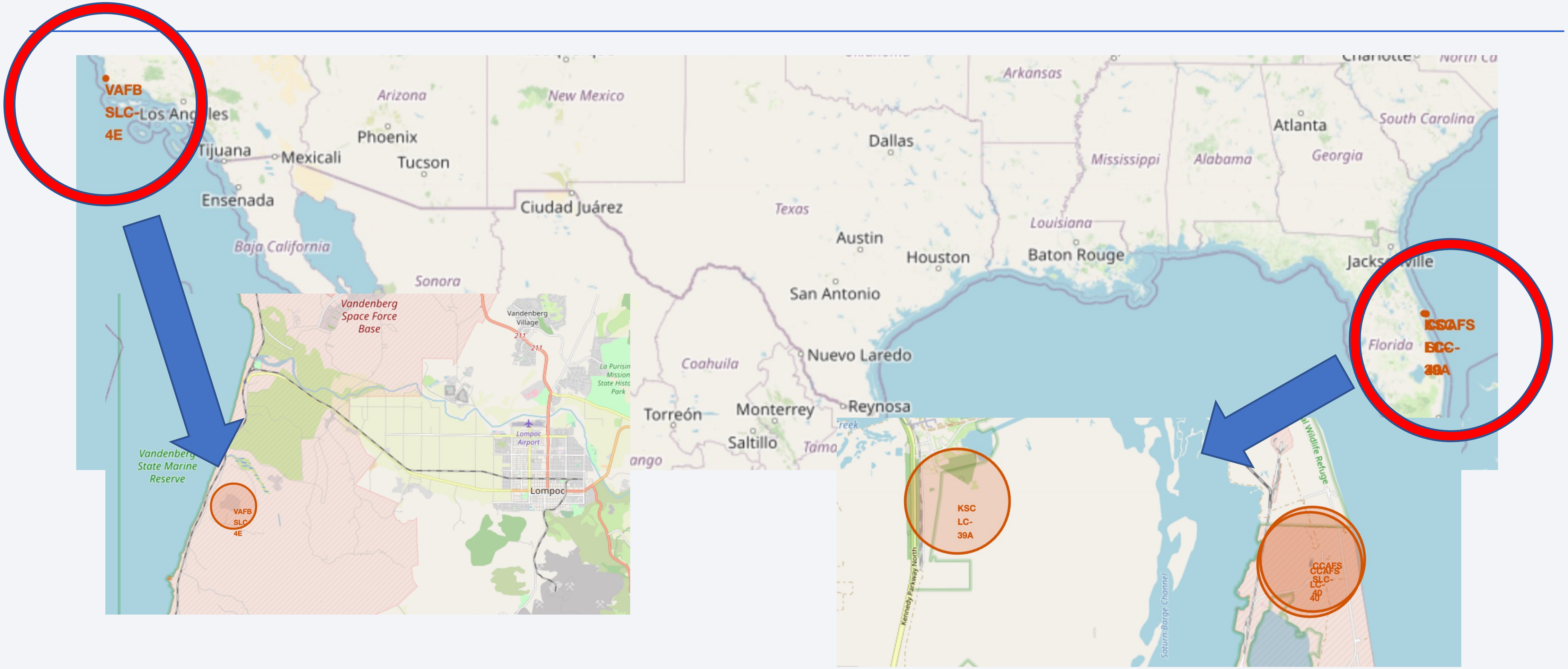
```
* sqlite:///my_data1.db
Done.
```

Date	Landing_Outcome	Count
22/05/2012	No attempt	10
22/12/2015	Success (ground pad)	5
04/08/2016	Success (drone ship)	5
01/10/2015	Failure (drone ship)	5
18/04/2014	Controlled (ocean)	3
29/09/2013	Uncontrolled (ocean)	2
28/06/2015	Precluded (drone ship)	1
12/08/2010	Failure (parachute)	1

Section 3

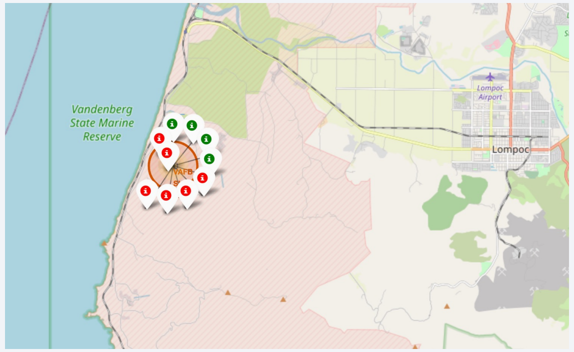
Launch Sites Proximities Analysis

All launch sites marked on a global map

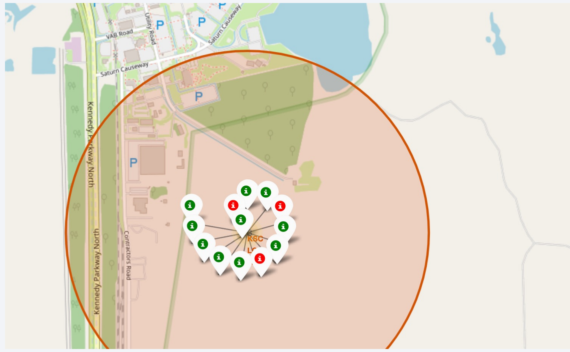


There are four launch sites, three in Florida, one in California

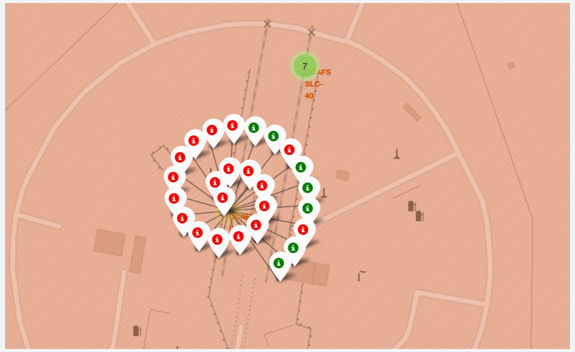
Success/Failed Outcomes Marked on the Map



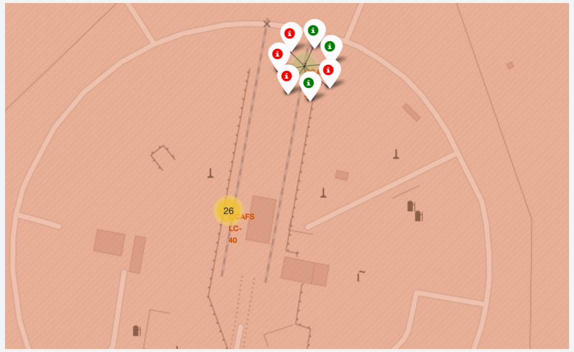
VAFB SLC-4E



KSC LC-39A



CCAFS-LC40



CCAFS-SLC40

- This shows failed/success outcomes at the four different launch sites.

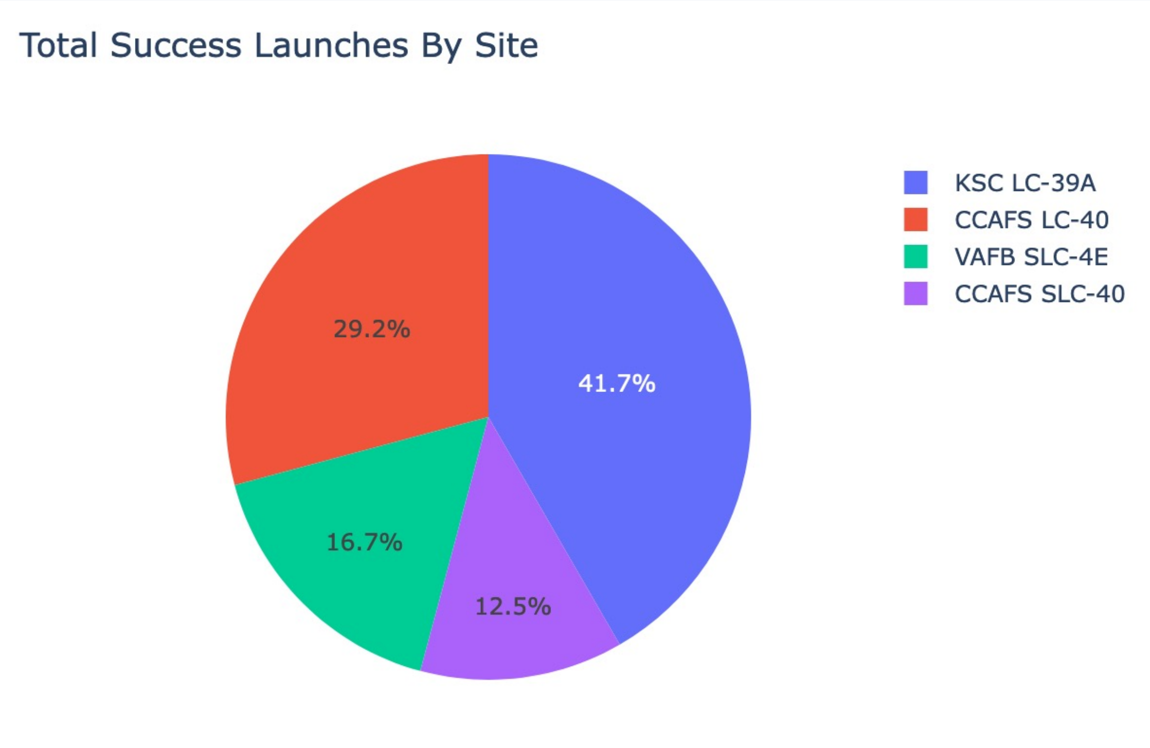
Distance between Launch Sites to its Proximities

- With Folium and a Distance calculation functions using Longitude and Latitude, we can also show distance between the Launch Sites and its Proximities. Here, for example, we show its distance from CCAFS LC-40 Launch site to the nearest Highway and Coast.

Section 4

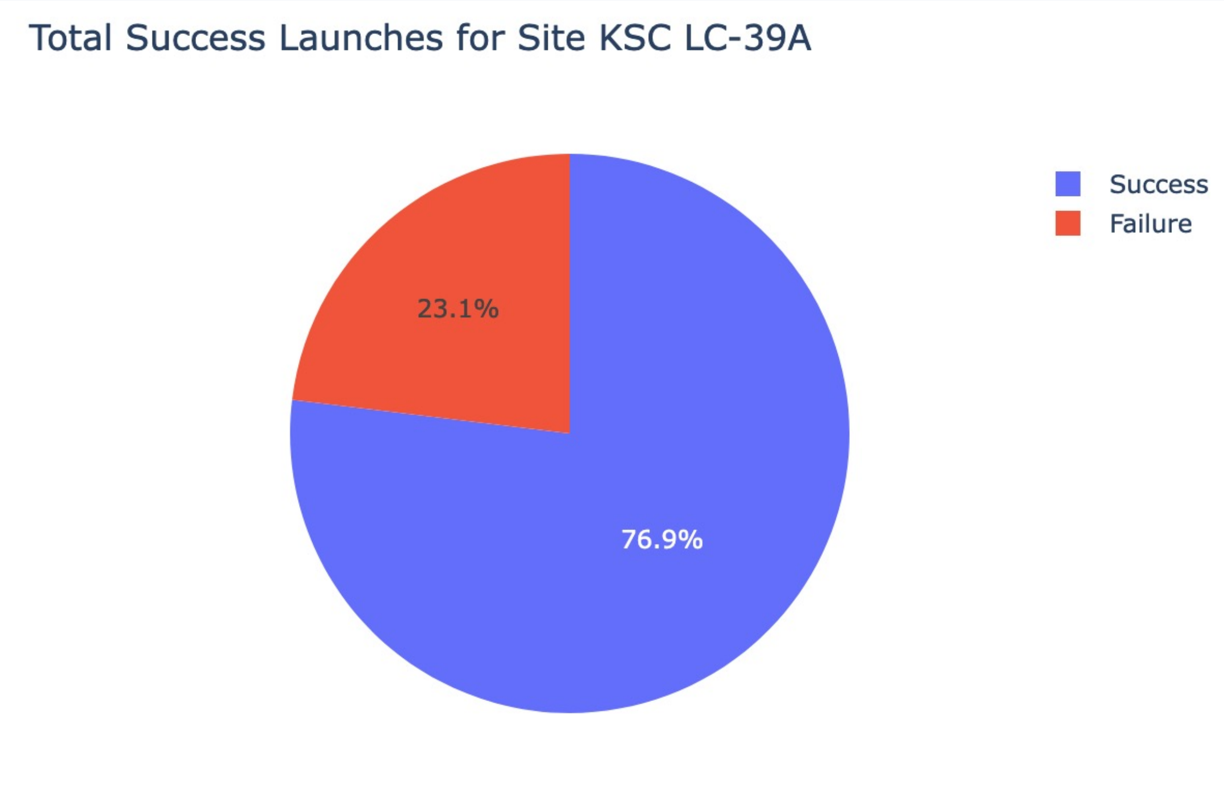
Build a Dashboard with Plotly Dash

Total Successful Launches by All Sites



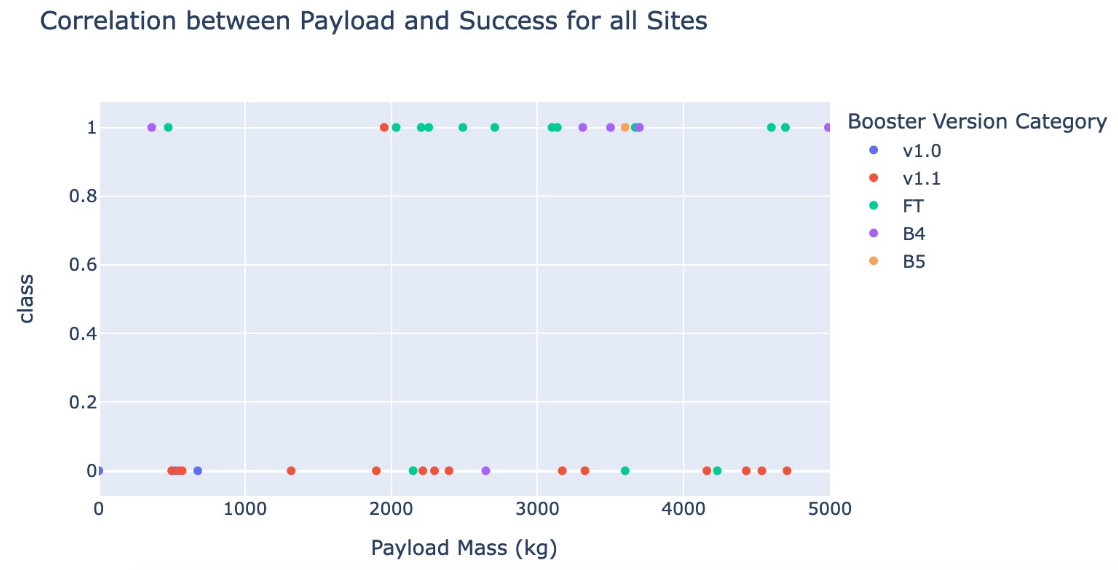
- We can see that KSC LC-39A had the most successful launches among all the sites.
- The fewest was CCAFS SLC-40.

Success Rate by Site

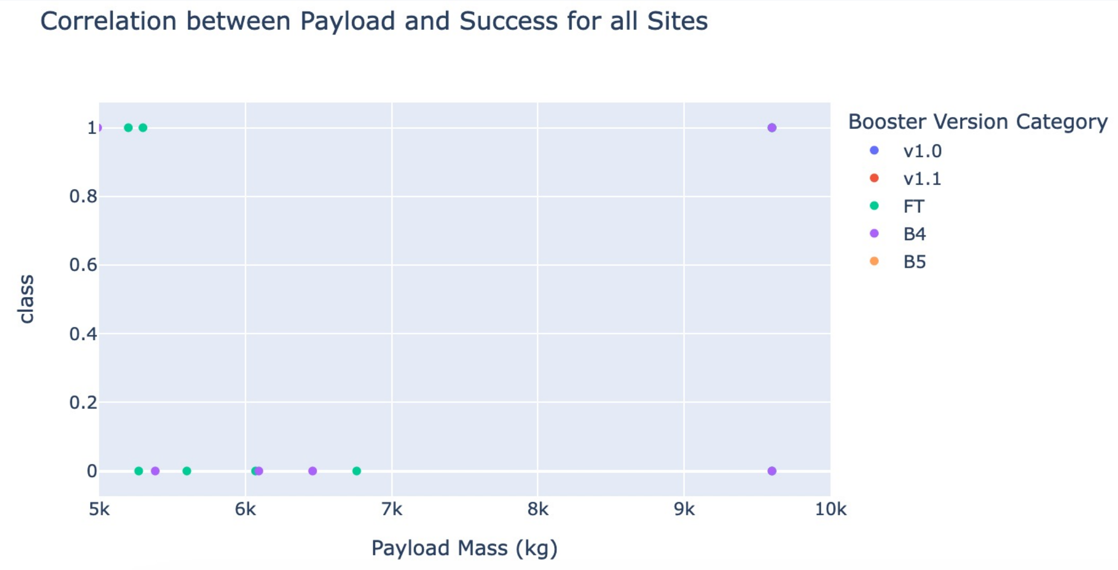


- KSC LC-39A had a success rate of 76.9%.

Payload vs Launch Outcome



0 to 5K KG



5K to 10 KG

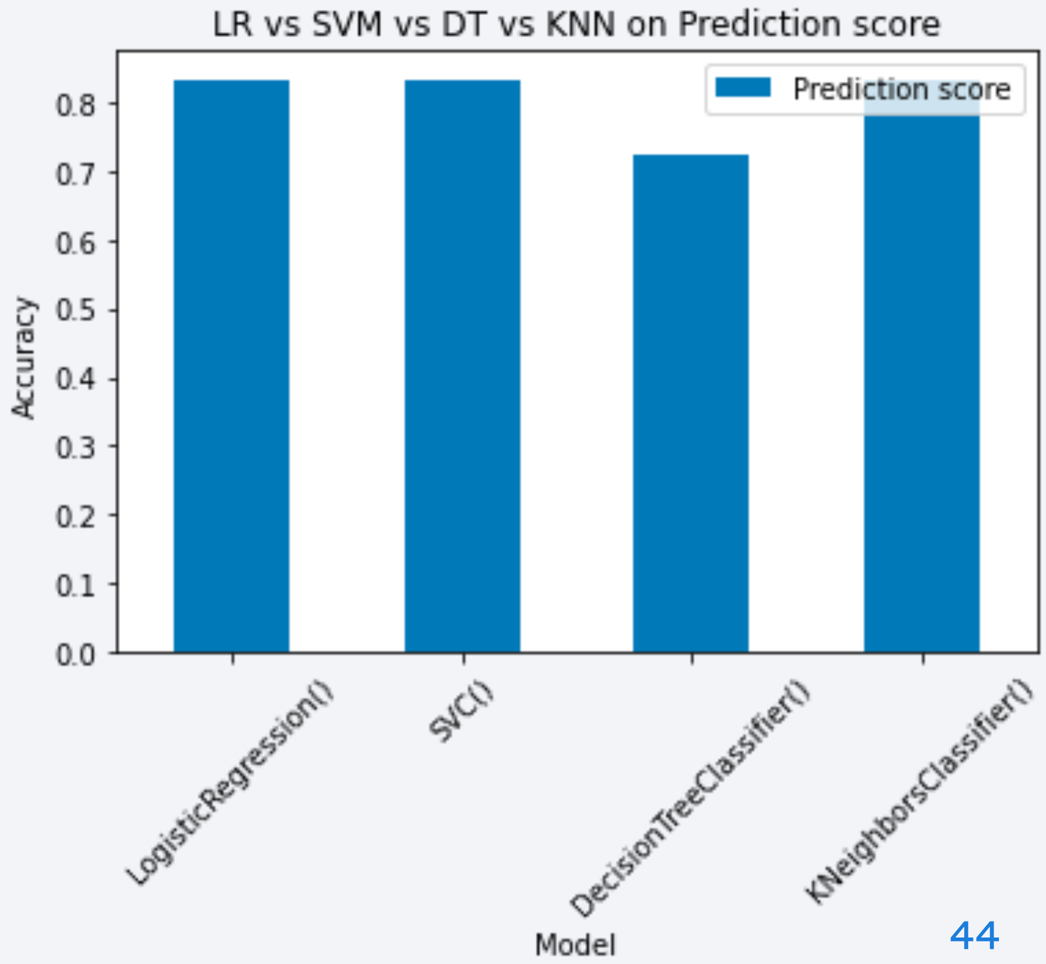
- The success rates for lower weighted payloads is higher than for heavy weighted payloads.

Section 5

Predictive Analysis (Classification)

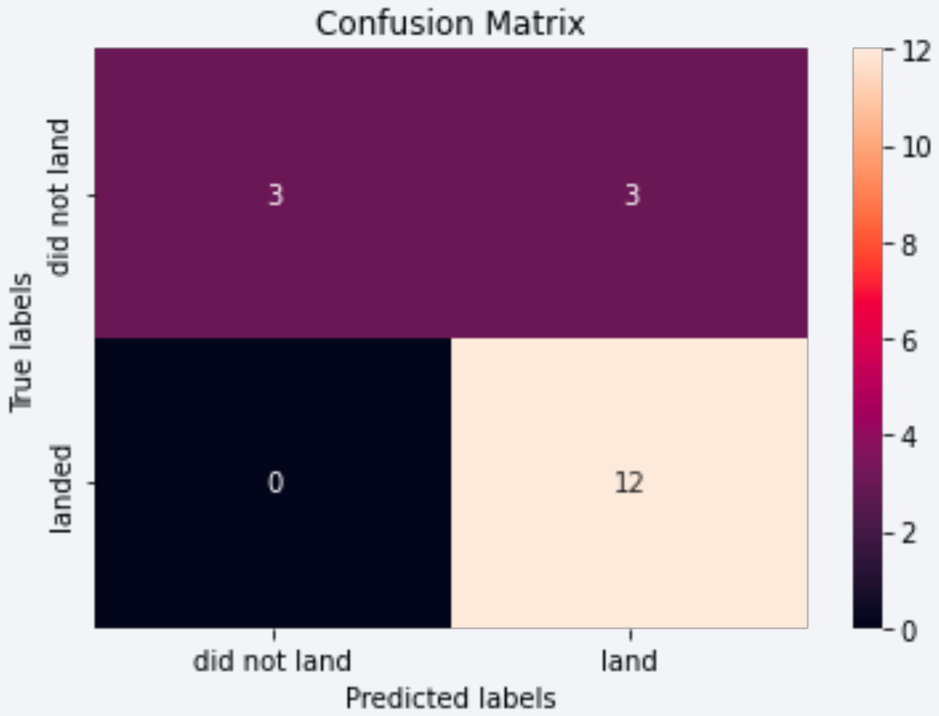
Classification Accuracy

- We see that Logistic Regression, SVM, and K-Neighbor Classification methods have the highest accuracy.



Confusion Matrix

- Assuming that
  - Positive: ‘land’
  - Negative: ‘did not land’
- There is a true positive of 12.
- There is a true negative of 3.
- There is a false positive of 3. This is the major problem.
- There are no false negatives.



Conclusions

- The are a couple conclusions that we can make…
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than heavier payloads.
- The success rates for SpaceX launches is directly proportional to the number of flights taken. In the future, there will be higher and higher success rates.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.
- Thus, given the parameters of a launch, we can make conclusions about the success rate.

Appendix

- Project from IBM Professional Certificate Applied Capstone Project https://www.coursera.org/learn/applied-data-science-capstone
- All the code can be found at this repository at https://github.com/AkaCoder404/data-science-learning/tree/main/IBM_Certificate_Coursera/Unit10

Thank you!