IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrew Gomoll
December 2022

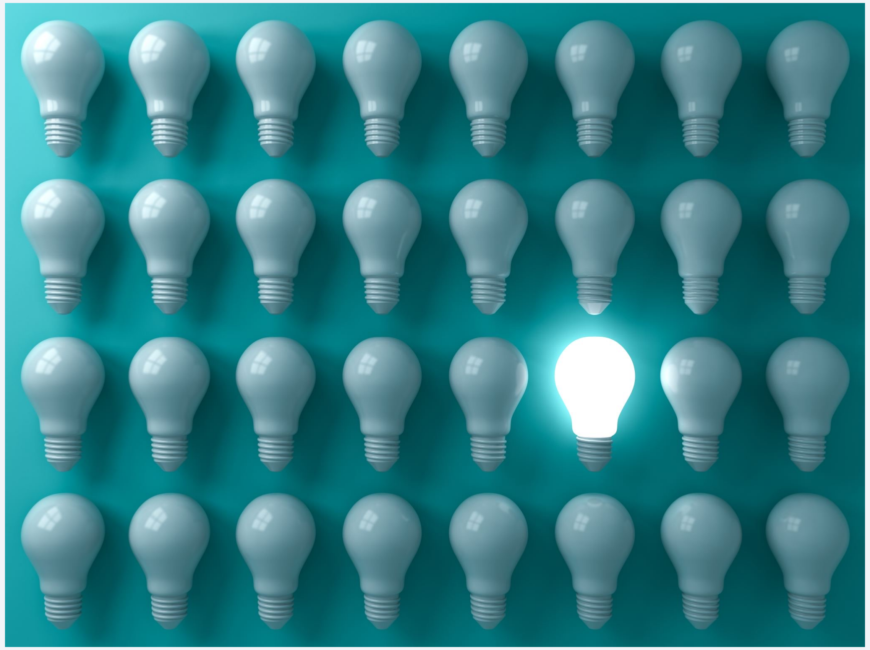
Outline

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

Executive Summary

- Summary of methodologies
  - Data was accessed using standard extraction techniques
  - Trends were explored using wrangling and visualization techniques
- Summary of all results
  - Success rates generally improved over time as more launches were conducted
  - Although several Machine Learning models were tested, all returned moderate accuracy scores
  - Accuracy may be improved by further evaluating the feature set for confounding relationships

Introduction



- In recent years private and public companies have launched into the extra-terrestrial products and travel market creating a commercial space age.
- For example, Virgin Galactic has offered some space flights, Rocket Labs provides small satellites, and Blue Origin manufactures reusable rockets.
- In the same vein, Space X has sent spacecraft to the international space station as well as other manned missions to space.
- Space X has been able to execute more launches than other providers as their signature rockets are reusable, allowing for relatively inexpensive missions.
- Advertised at $62 million per launch, the Space X missions undercut competing providers with more expensive launches at $162 million.
- This analysis aims to more accurately predict the price of each launch using data wrangling, analysis, and machine learning to predict the reusability of Space X rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

Data Collection

- Data was collected using several methodologies
    - Get request to Space X API
    - Web scraping Space X records from Wikipedia
    - Connecting to a database to extract CSV files

Data Collection – SpaceX API

- First, data was collected by by sending a request to the Space X API to collect data. The data was then cleaned and formatted for use.
- The API response was sent as a json containing a large volume of data. From this data, variables were created to list of information of interest and then converted to a pandas DataFrame.
- The data was then filtered to focus on Falcon9 launches and eliminate missing data.
- https://github.com/agomoll/Space_X_Analysis/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Second, launch data was collected by scraping wiki Space X records from the web using Beautifulsoup.
- The data was then parsed into a dictionary and then stored in a Pandas DataFrame.
- https://github.com/agomoll/Space_X_Analysis/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

- Data processing:
  - Calculate the percentage of the missing values in each attribute
  - Identify the datatype of each attribute
  - Calculate the number of launches on each site
  - Calculate the number of launches to each orbit level
  - Calculate the number of outcome types
  - Create Outcome Classification column for successful (1) and unsuccessful (0) classifiers
  - Determine Success Rate
- https://github.com/agomoll/Space_X_Analysis/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summary of plotted charts:
  - Flight number vs Launch Site scatter plot
  - Payload vs Launch Site scatter plot
  - Success Rate vs Orbit Type bar chart
  - Flight Number vs. Orbit Type
  - Payload vs Orbit Type
  - Launch Success yearly trend

- https://github.com/agomoll/Space_X_Analysis/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Summary of SQL queries:
  - Display unique launch sites
  - Display Launch sites beginning with 'CCA'
  - Display total mass carried by NASA (CRS)
  - Display average mass carried by booster version F9 v1.1
  - Display date of first successful landing on a ground pad
  - List names of boosters with success landing on drone ships with payload greater than 4000kg but less than 6000kg
  - Display the total number of missions, both success and failures
  - List the booster versions that have carried the max payload mass
  - 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 the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- https://github.com/agomoll/Space_X_Analysis/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To create interactive map experiences, the following features were added to map visualizations using Folium:
  - Circles – shape and color make it easier to locate on map
  - Map Labels – show brief description of site name
  - Popup labels - deploy when site is clicked to better show site info
  - Launch Markers – color code success or failed stage one landing
  - Launch site to obstacle lines with calculations – show distance to landing obstacles
- https://github.com/agomoll/Space_X_Analysis/blob/master/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Created a web application dashboard using Plotly Dash
- Features include:
  - Drop Down Menu to enable launch site selection
  - Pie Chart showing total successful launches for all sites
    - If a specific site is selected the pie chart shows the successes vs failures at that site
  - Slider bar to further filter by payload
  - Scatter chart displaying Payload vs Launch Site Success with Booser Versions
- https://github.com/agomoll/Space_X_Analysis/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Summary of classification model testing:
  - Create a NumPy array from the target data column ‘Class’ and assign to variable ‘Y’
  - Standardize the feature dataset using Standard Scalar and assign to variable ‘X’
  - Use train_test_split() to divide the data into training and testing sets
  - Test a variety of models predictive ability with the available data
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree Classifier
    - K Nearest Neighbors
- https://github.com/agomoll/Space_X_Analysis/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - It appears that Launch sites VAFB and KSC had higher landing success rates than CCAFS
  - However, it also appears that landing success at CCAFS improved dramatically in later flights
  - Overall, success increased over time with increasing payloads

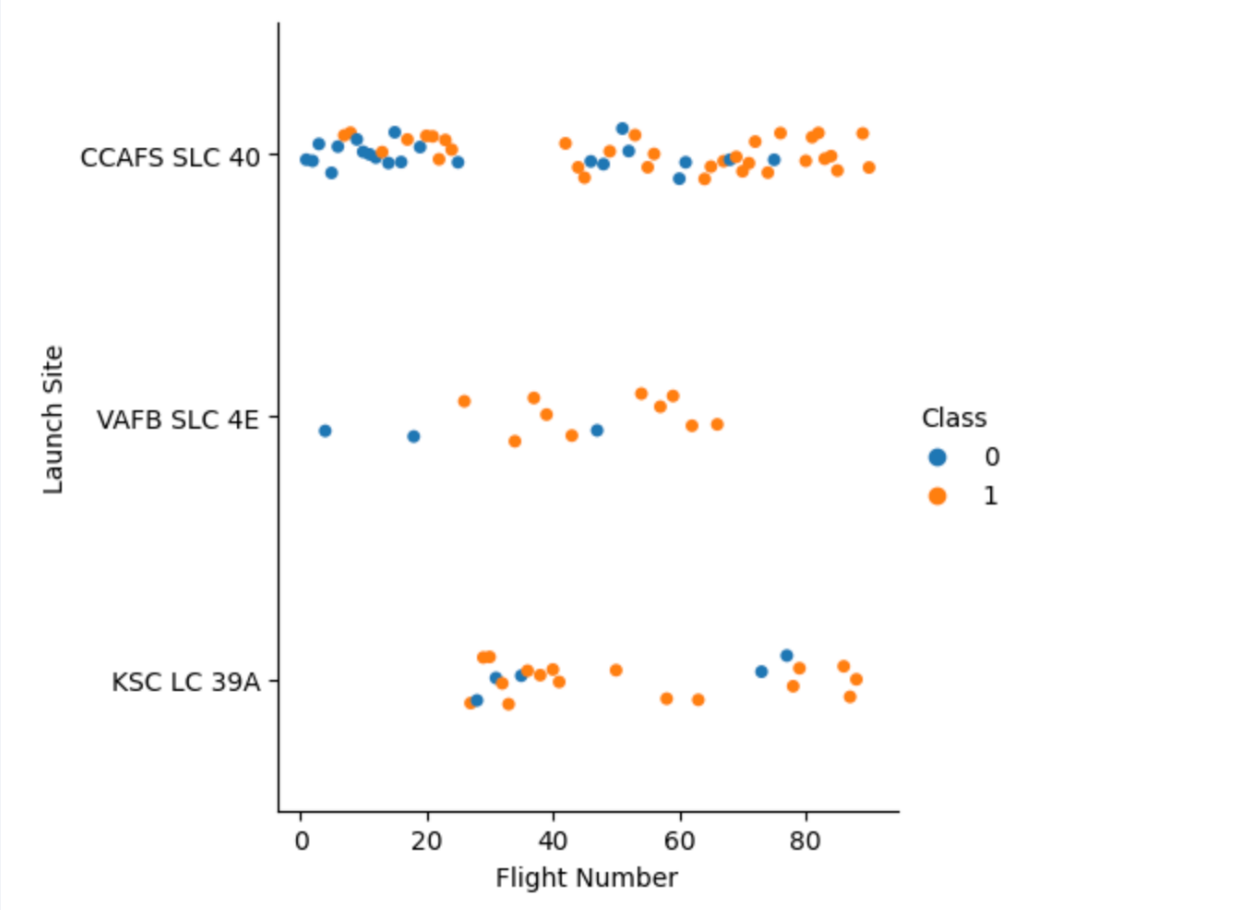
- Predictive analysis results
  - All models tested returned an accuracy score of 83%
  - Of the 18 launches in the test set, 12 successful launches were accurately predicted while 3 false positives were predicted.

Section 2

Insights drawn from EDA

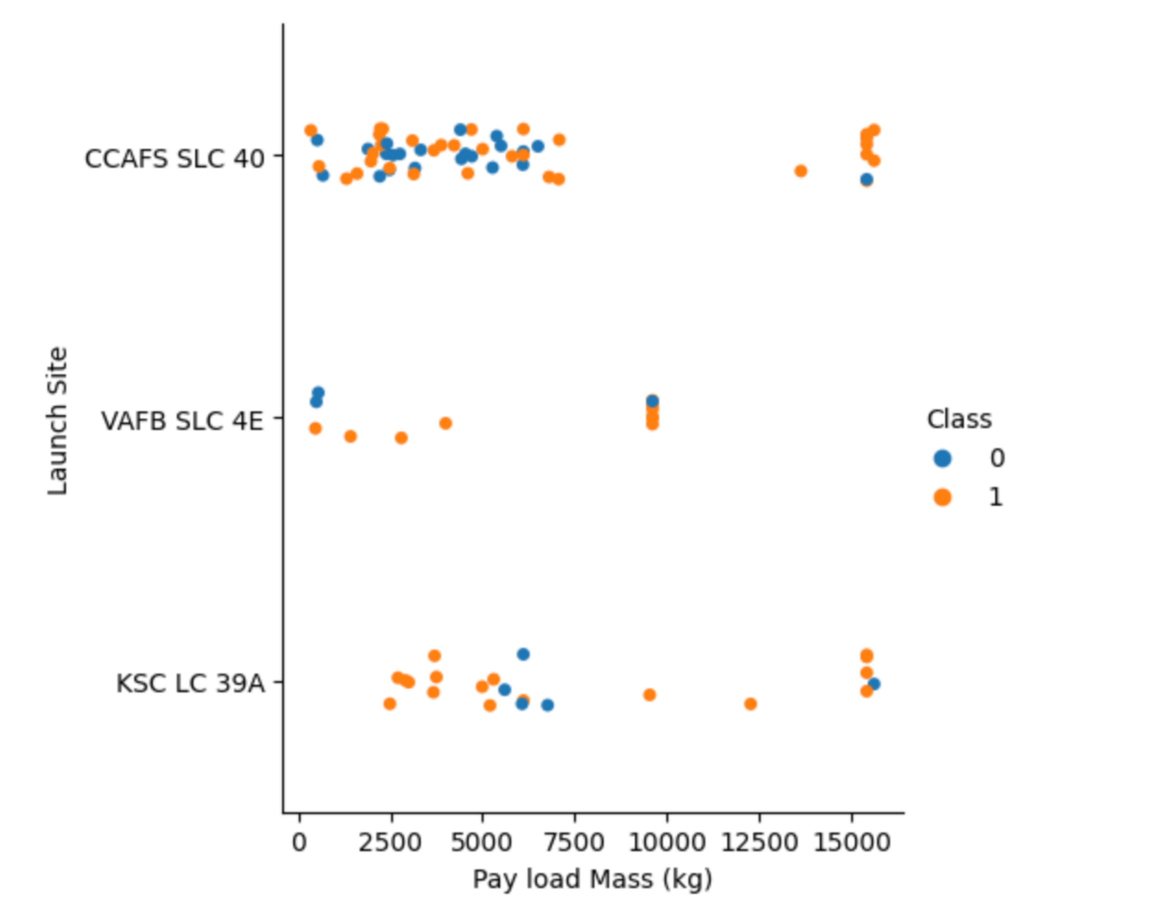
Flight Number vs. Launch Site

- It appears that Launch sites VAFB and KSC had higher landing success rates than CCAFS
- However, it also appears that landing success at CCAFS improved dramatically in later flights



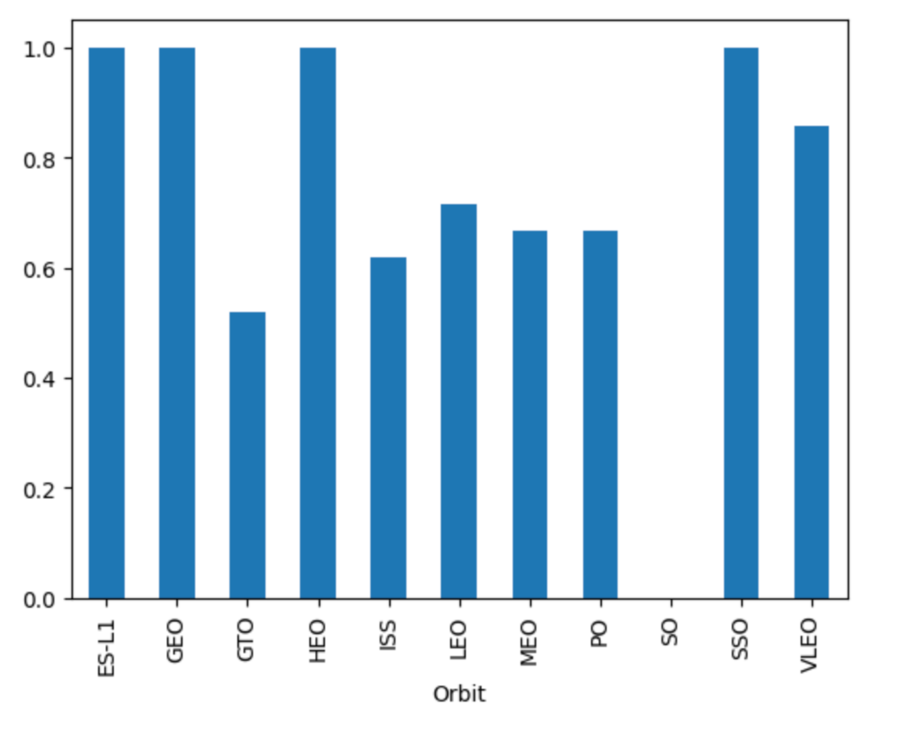
Payload vs. Launch Site

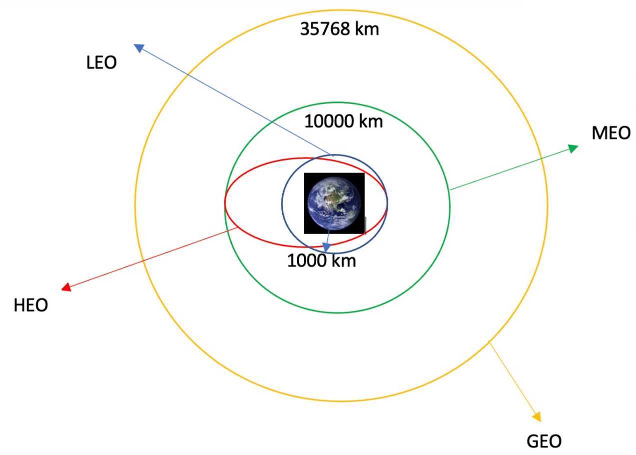
- CCAFS had a higher percentage of success with Heavier payloads
- VAFB did not have any payloads over 10,000 kg
- Successful landing for KSC varied across payloads



Success Rate vs. Orbit Type

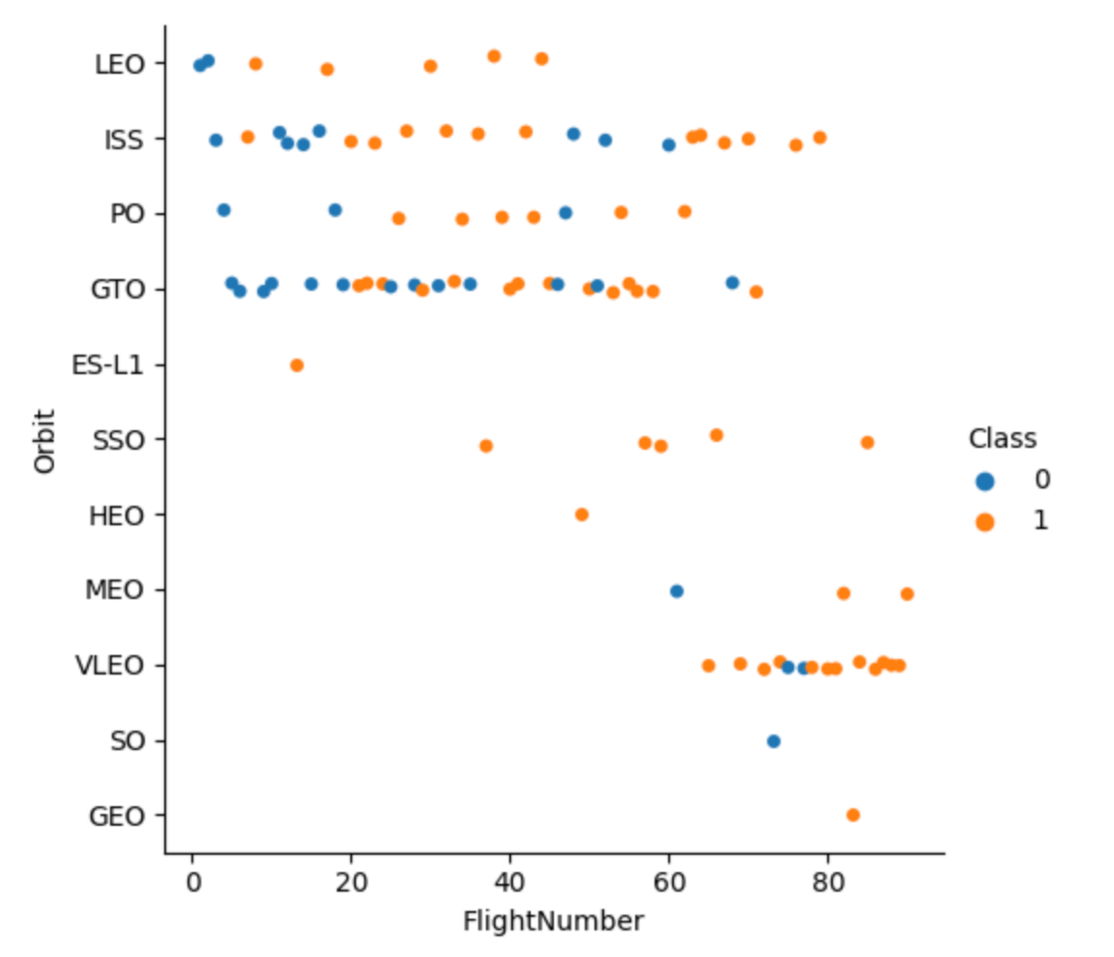
- It appears that the highest rates of landing success occurred for launches to:
    - ES-L1
    - GEO
    - HEO
    - SSO
    - VLEO
- Given the variable nature of these orbits, it is unclear whether a relationship exists





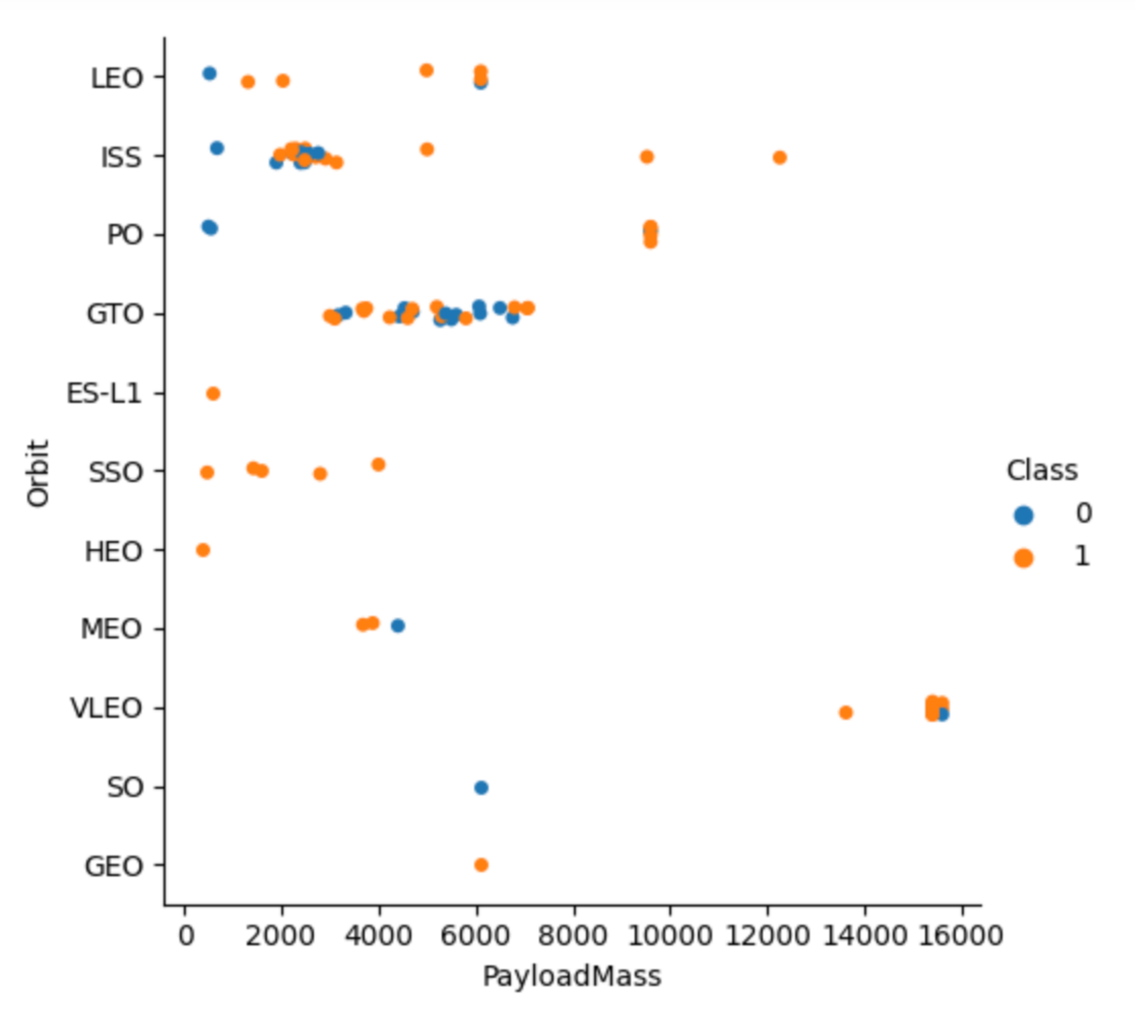
Flight Number vs. Orbit Type

- It appears that early flights were focused on LEO, ISS, PO, and GTO orbits with varied success
- Later flights explored more orbit types with increasing success



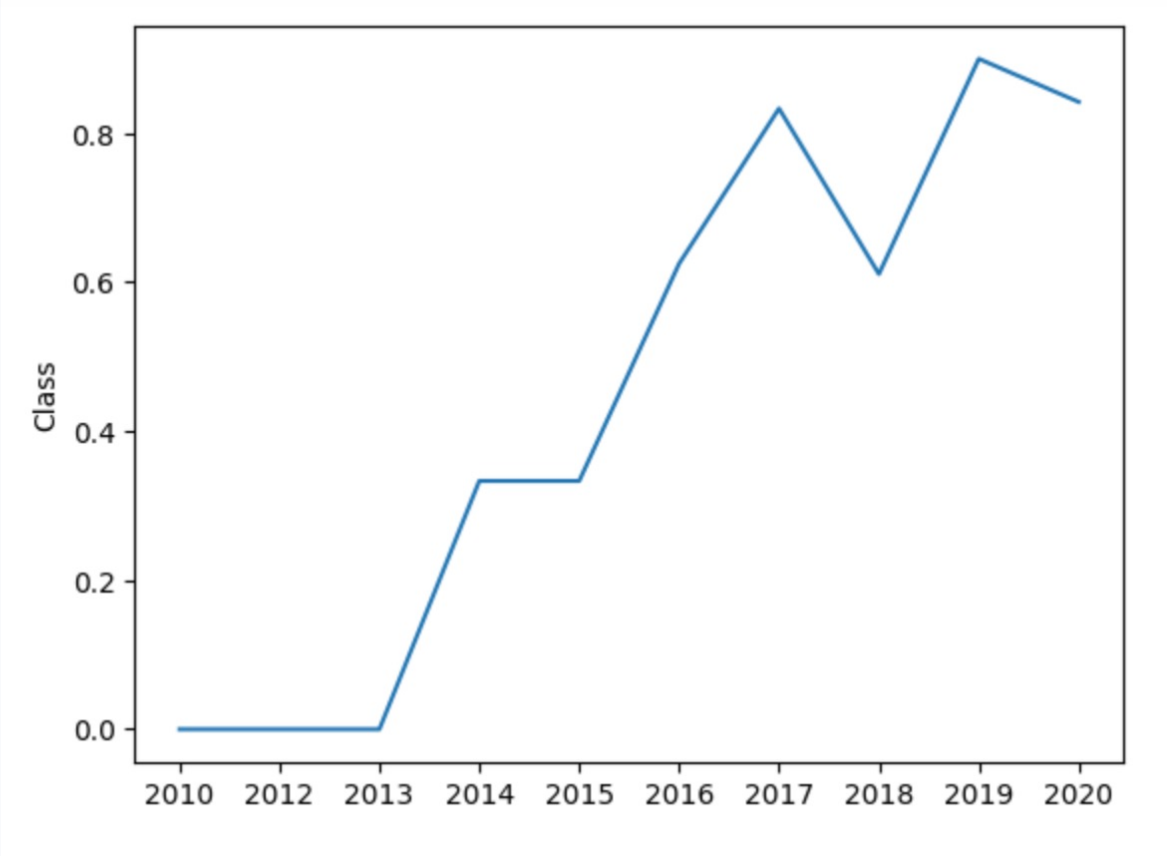
Payload vs. Orbit Type

- It appears that lighter payload masses (<10K kg) were the focus of earlier orbit targets.
- Later flights had success with heavier payloads, particularly at VLEO



Launch Success Yearly Trend

- Overall, the success rate increased since 2013



All Launch Site Names

- Find the names of the unique 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'

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

```
%sql SELECT * from SPACEXTBL WHERE ("Launch_Site") LIKE 'CCA%' LIMIT (5);
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE (Customer) = 'NASA (CRS)';
```

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE (Booster_Version) = 'F9 v1.1';
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE ("Landing _Outcome") = 'Success (ground pad)';
```

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

MIN("Date")
01-05-2017

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

```
%sql SELECT DISTINCT("Booster_Version") FROM SPACEXTBL WHERE ("Landing _Outcome") = 'Success (drone ship)' AND ("PAYLOA
```

```
 * 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

```
%sql SELECT COUNT("Mission_Outcome") FROM SPACEXTBL ;
```

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

COUNT("Mission_Outcome")
101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT DISTINCT("Booster_Version") FROM SPACEXTBL WHERE ("PAYLOAD_MASS__KG_") = (SELECT MAX("PAYLOAD_MASS__KG_") F
```

```
 * 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", substr(Date,7,4) AS "Year", "Landing _Outcome", "Booster_Version", "Launch_
WHERE "Landing _Outcome" = 'Failure (drone ship)'
AND substr(Date,7,4)='2015'
```

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

Month	Year	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	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

```
%%sql SELECT "Landing _Outcome", COUNT(*) AS "Count"
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE "%Success%"
AND "Date" BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "Landing _Outcome"
ORDER BY "Count";
```

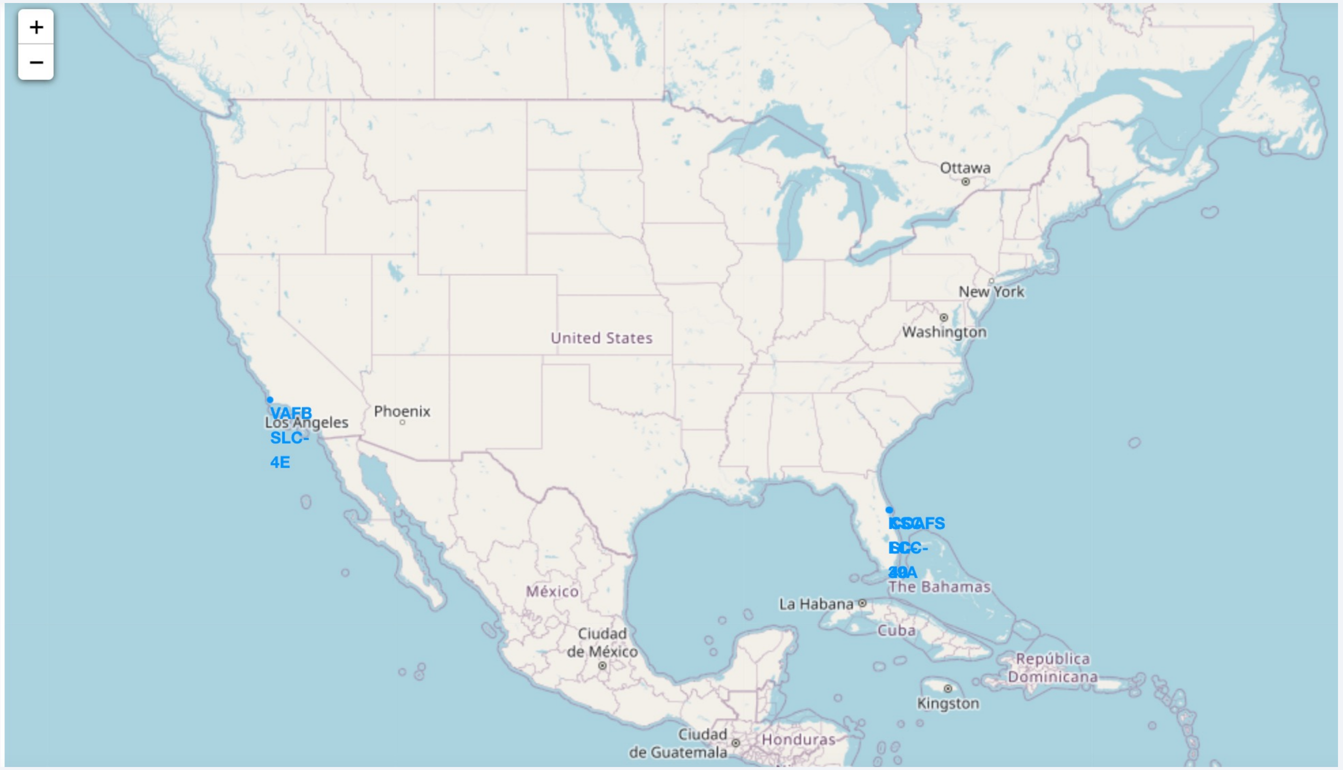
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	Count
Success (ground pad)	6
Success (drone ship)	8
Success	20

Section 3

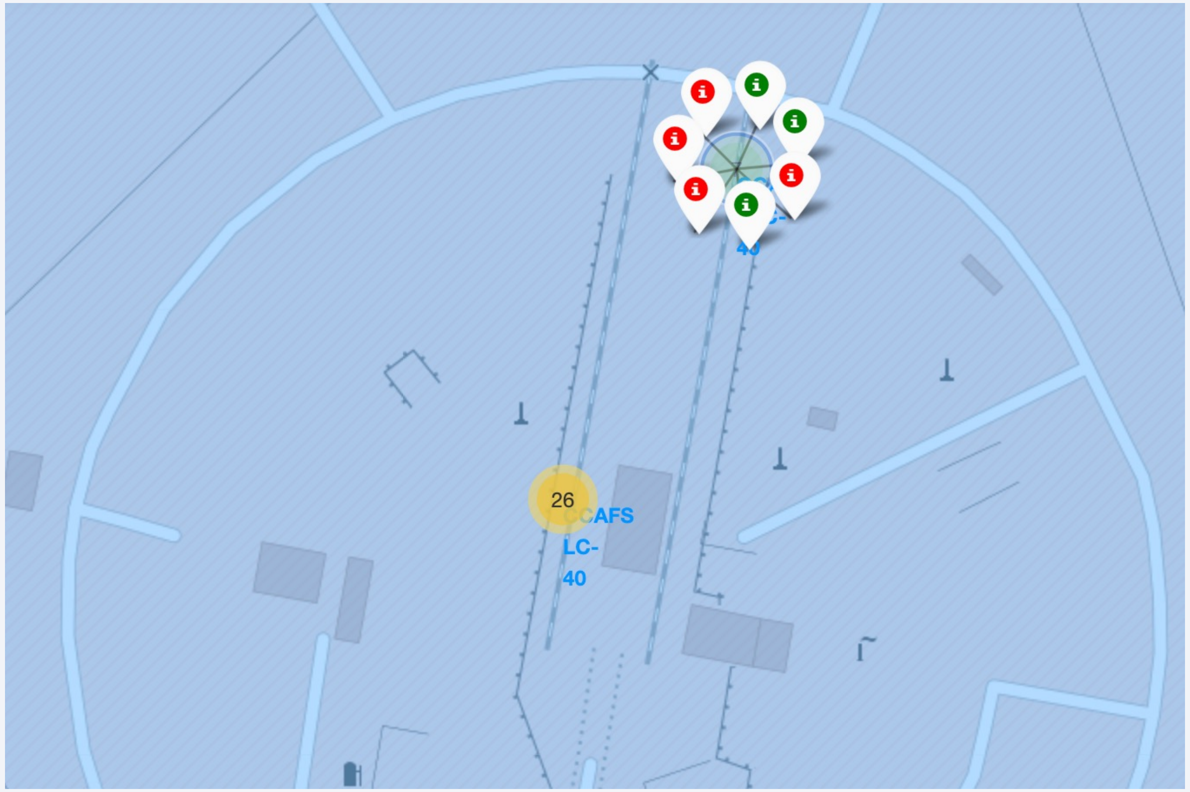
Launch Sites Proximities Analysis

All Launch Site Locations



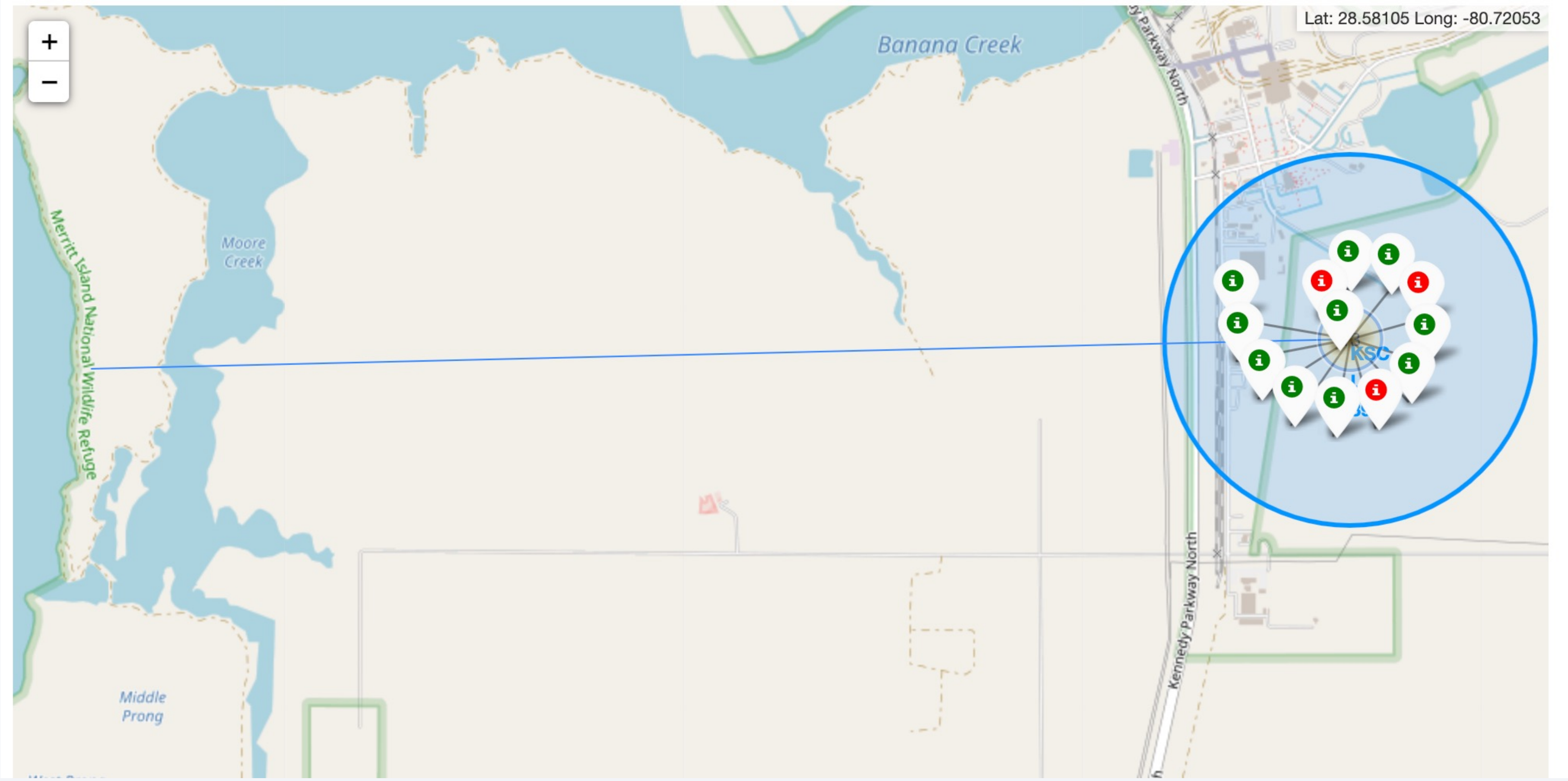
- Launch sites of interest are located in California and Florida

Launch Site Outcome Markers



- The display shows a launch site where 3 of the 7 launches had successful landings

Distance to Proximities

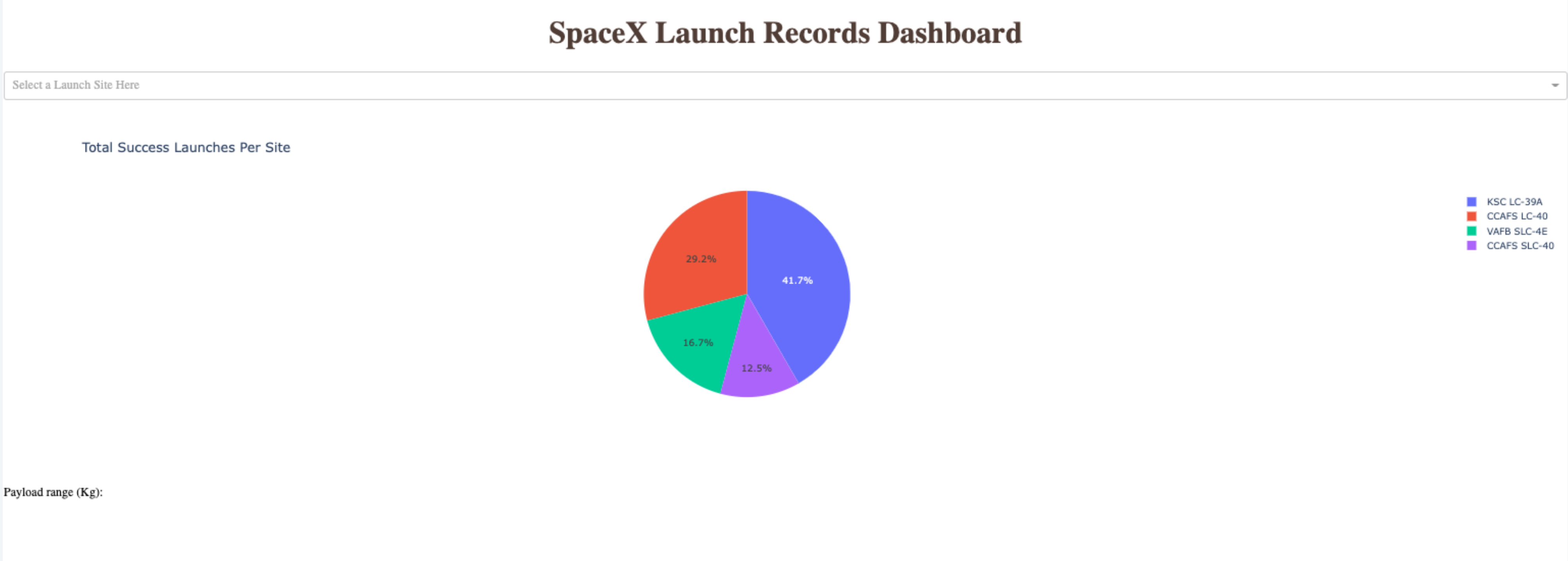


- Showing Distance to eastern coastline from Launch Site KSC LC-39A

Section 4

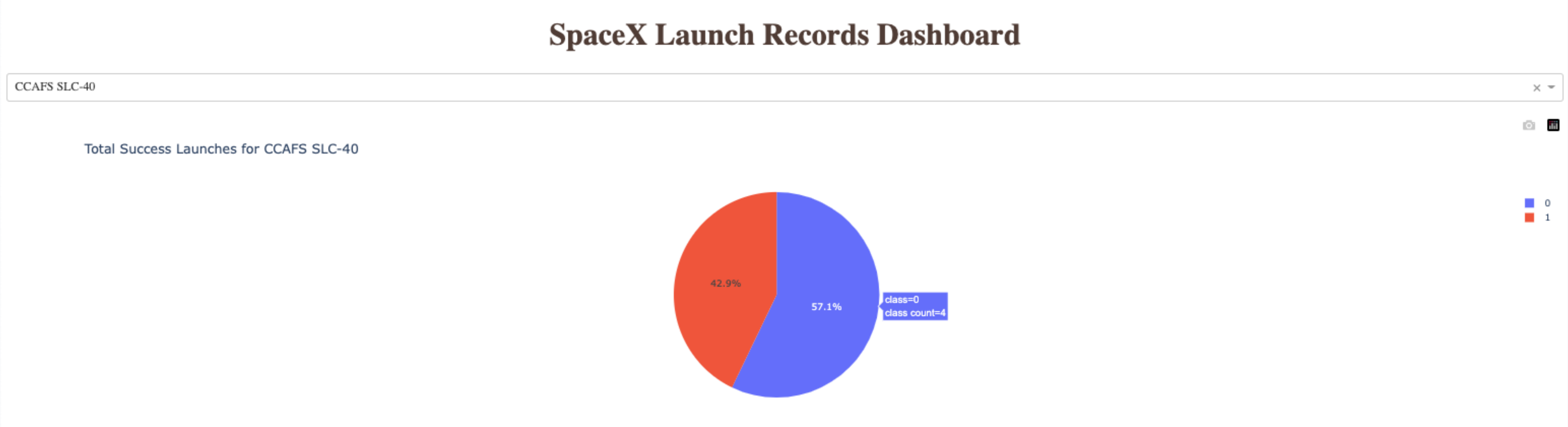
Build a Dashboard with Plotly Dash

Success Records for All Sites



- Across all sites, KSC LC-39A accounted for the highest percentage of successes followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40

Highest Success Ratio



- For Launch Site CCAFS SLC-40, 42.9% of the launches were successful, while 57.1% of the launches were unsuccessful

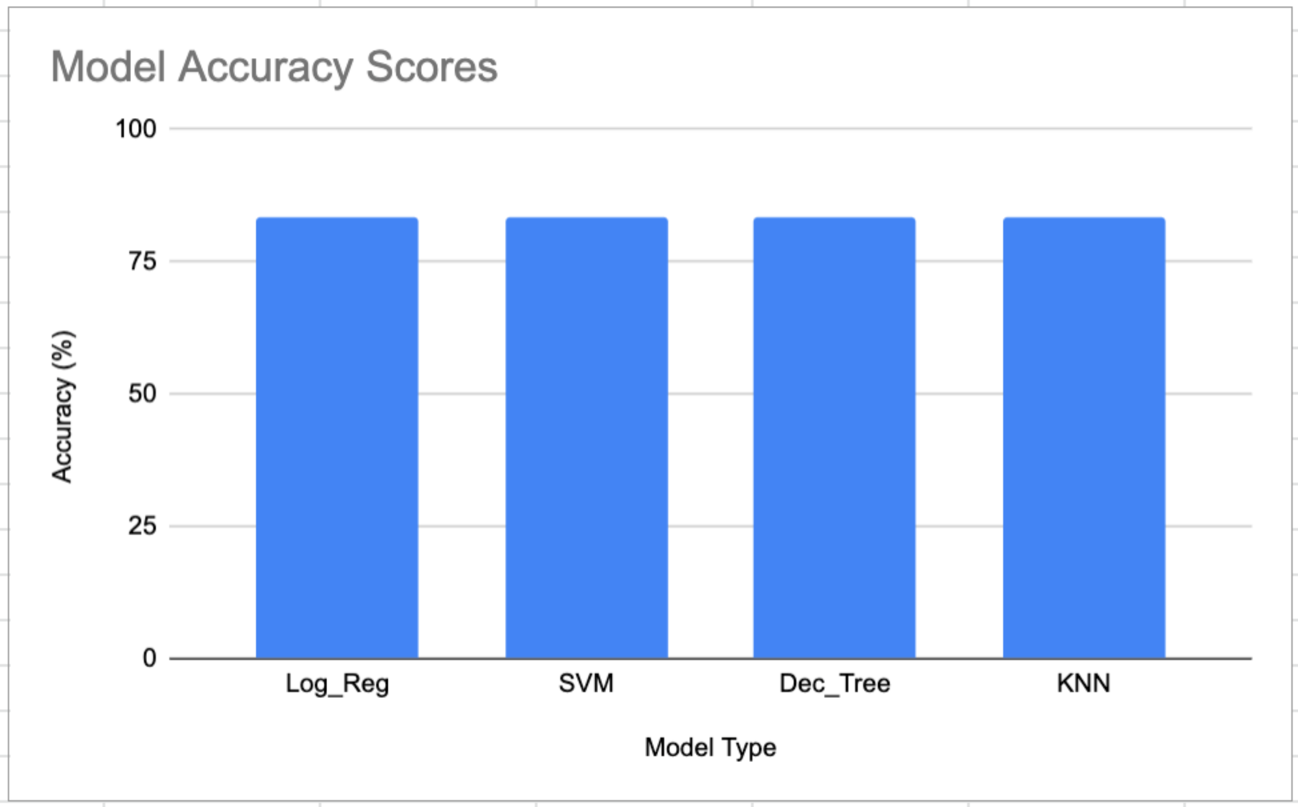
Count of Successful Launches by Payload Mass

- It appears that Booster version v1.1 had the lowest success rate across all payload ranges while version FT had the highest success, especially between 2000kg and 5000kg

Section 5

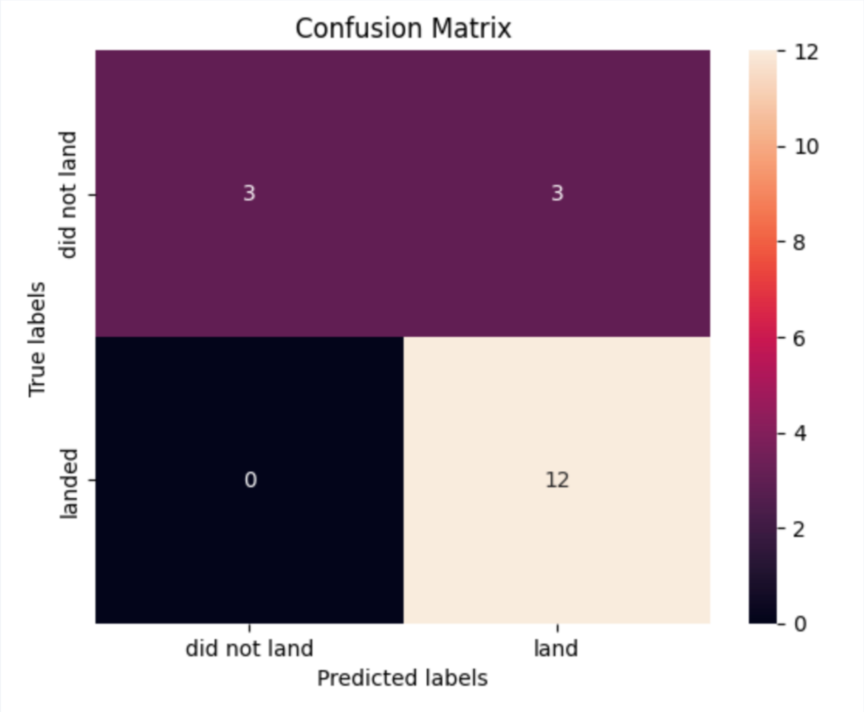
Predictive Analysis (Classification)

Classification Accuracy



- All models resulted in similar accuracy scores of 83%

Confusion Matrix



- The confusion matrix shows that of the 18 launches in the test set:
  - 12 successful landing scenarios were accurately predicted
  - 3 false positive predictions were made
  - 3 accurate predictions were made for unsuccessful landings

Conclusions

- Overall, stage one landing success increased over time.
- Machine learning model accuracy was moderately successful at predicting the successful landings, although there was a noticeable rate of false positive predictions.
- The dataset was relatively small with only 18 launches in the test set
- Accuracy may be improved by evaluating the feature set. It is possible that confounding features need to be removed

Thank you!