SpaceX Launch First Stage Data Analysis

SRIKRISHNAN R
11/29/2024

Outline

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

Executive Summary

- Summary of Methodologies
- The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:
  • **Collect** data using SpaceX REST API and web scraping techniques
  • **Wrangle** data to create success/fail outcome variable
  • **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
  • **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
  • **Explore** launch site success rates and proximity to geographical markers
  • **Visualize** the launch sites with the most success and successful payload ranges
  • **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K -nearest neighbors (KNN)

- Results
- Exploratory Data Analysis:
  • Launch success has improved over time
  • KSC LC-39A has the highest success rate among landing sites
  • Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate
- Visualization/Analytics:
  • Most launch sites are near the equator, and all are close to the coast
- Predictive Analytics:
  • All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

METHODOLOGY

Methodology

Steps

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Explore data via EDA with SQL and data visualization techniques
- Visualize the data using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models
- Tune and evaluate models to find best model and parameters

Data Collection – API Call

Steps

- Request data from SpaceX API (rocket launch data)
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/01_SpaceX_Data_Collection.ipynb

Data Collection – Web Scraping

Steps

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary
- Export data to csv file
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/02_SpaceX_Web_Scraping.ipynb

Data Wrangling

Steps

- Perform EDA and determine data labels
- Calculate:
  - Number of launches for each site
  - Number and occurrence of orbit
  - Number and occurrence of mission outcome per orbit type
- Create binary landing outcome column (dependent variable)
- Export data to csv file
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/03_SpaceX_Data_Wrangling.ipynb

Landing Outcomes:

- **True Ocean**: mission outcome had a successful landing to a specific region of the ocean
- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS:** meant the mission had a successful landing on a ground pad
- **False RTLS**: represented an unsuccessful landing on a ground pad
- **True ASDS:** meant the mission outcome had a successful landing on a drone ship
- **False ASDS:** represented an unsuccessful landing on drone ship
- **Outcomes converted** into 1 for a successful landing and 0 for an unsuccessful landing

EDA With Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists.
- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/05_SpaceX_EDA_DataVisualization.ipynb

EDA With SQL

Display:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- 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 which had success landing on drone ship and have payload mass greater than 4,000 but less than 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)
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/04_SpaceX_EDA_SQL.ipynb

Folium Map

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red 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 of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/06_SpaceX_FoliumMaps.ipynb

Plotly Dash Dashboard

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Slider of Payload Mass Range

- Allow user to select payload mass range

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/07_SpaceX_Plotly_Dashoard.py

ML Predictive Analytics

Charts

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree, and K-Nearest Neighbor
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model
- https://github.com/rsk1707/IBM-DataScience-Capstone-SpaceY/blob/main/08_SpaceX_ML_Predictions.ipynb

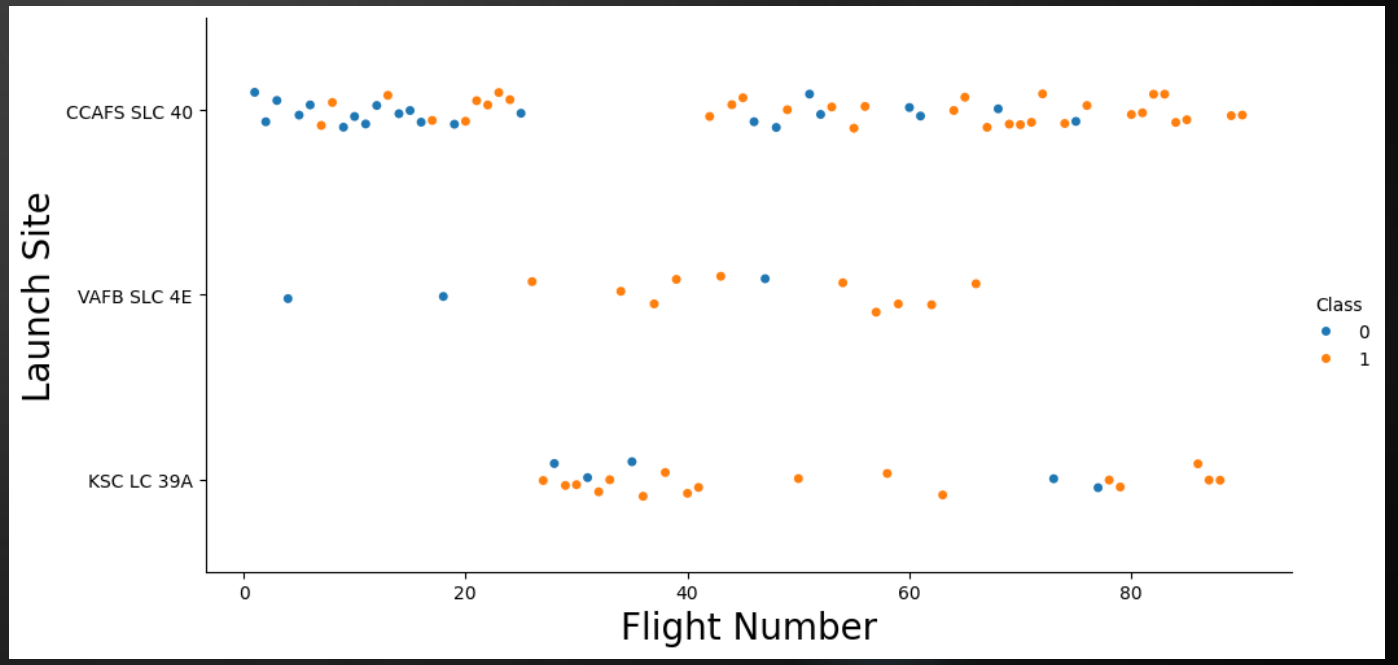
RESULTS

Results Summary

- Exploratory Data Analysis:
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual Analytics
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- Predictive Analytics
  - Decision Tree model is the best predictive model for the dataset

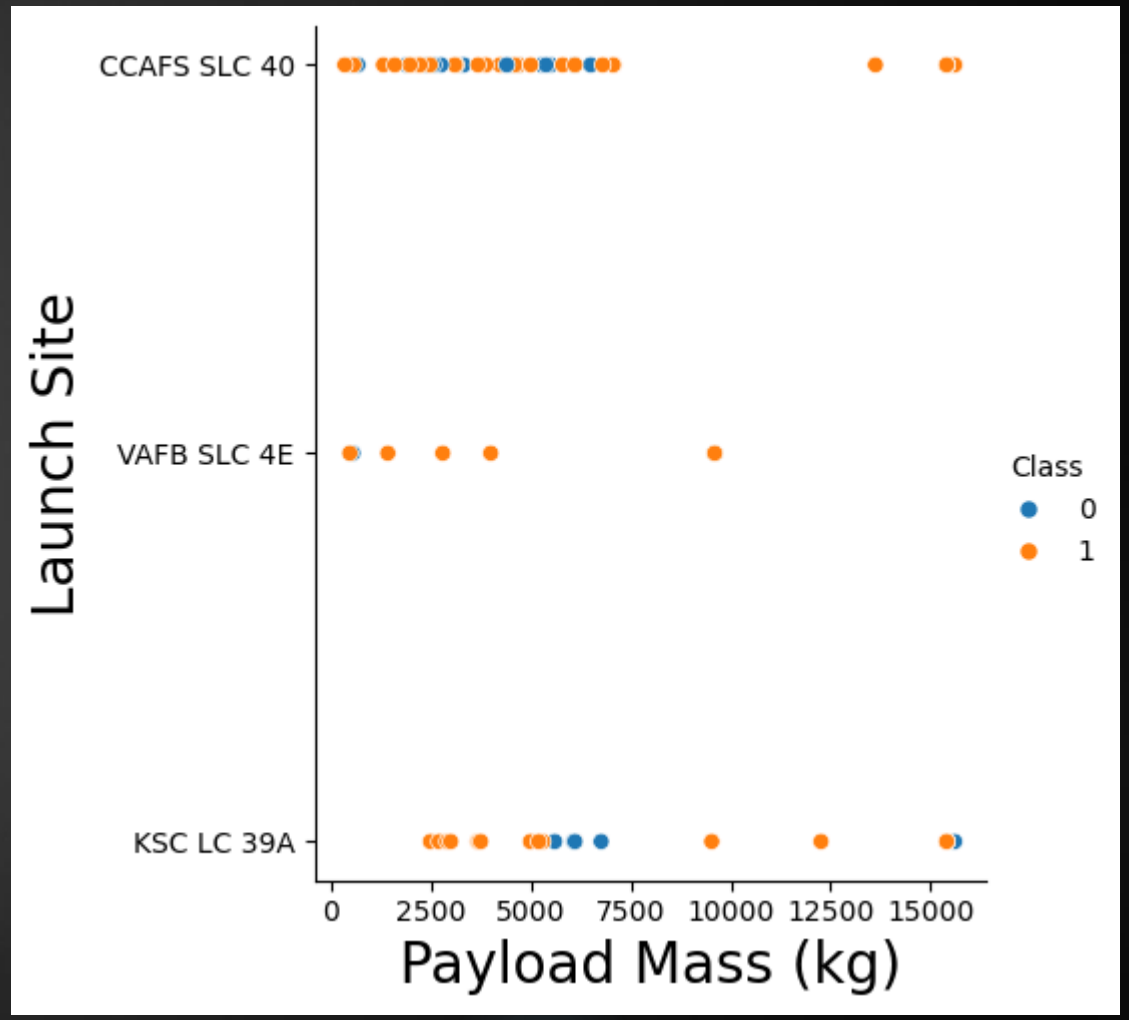
Flight Number vs Launch Site

- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



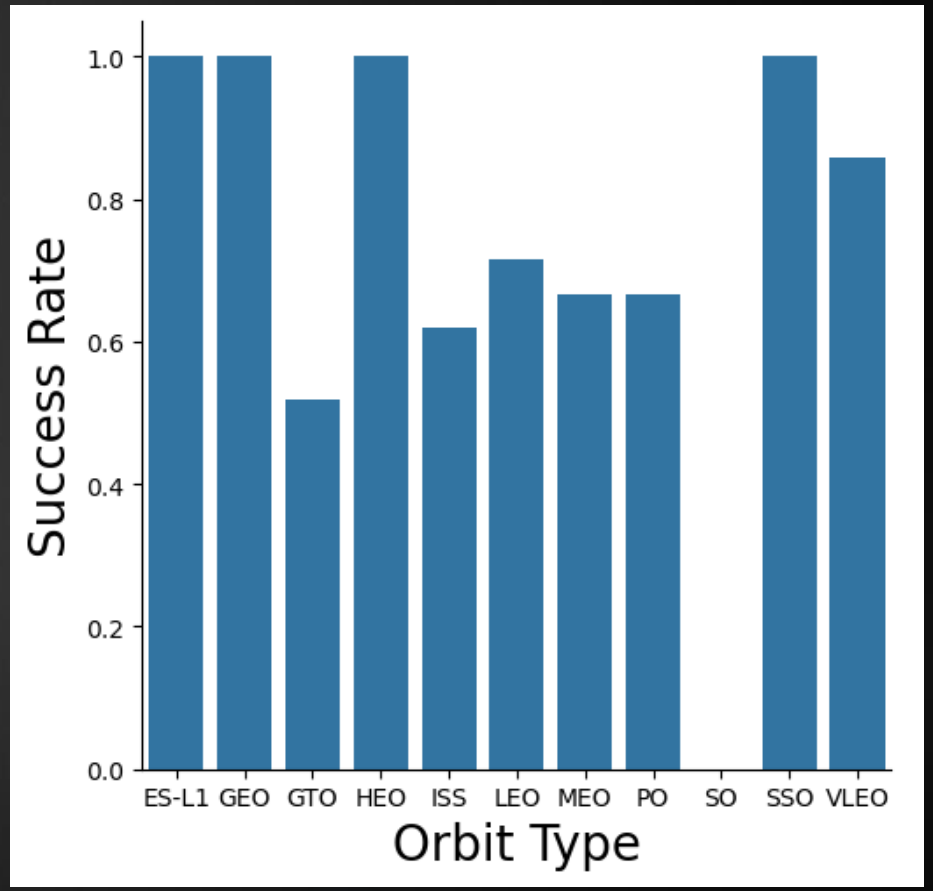
Payload vs Launch Site

- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SLC 4E has not launched anything greater than ~10,000 kg



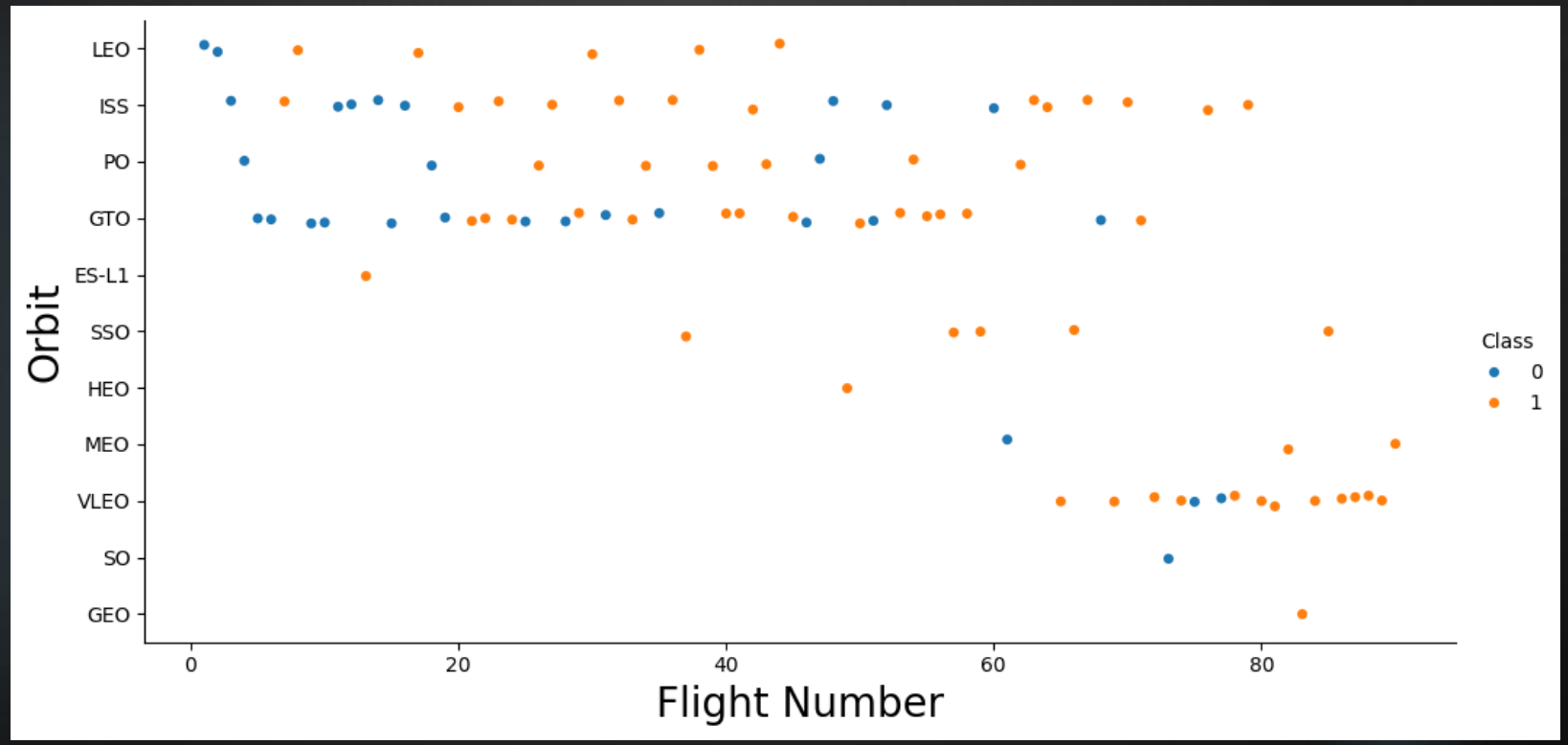
Success Rate by Orbit

- ES-L1, GEO, HEO and SSO have 1.0 success rate
- SO has 0.0 success rate
- Others have intermediate values



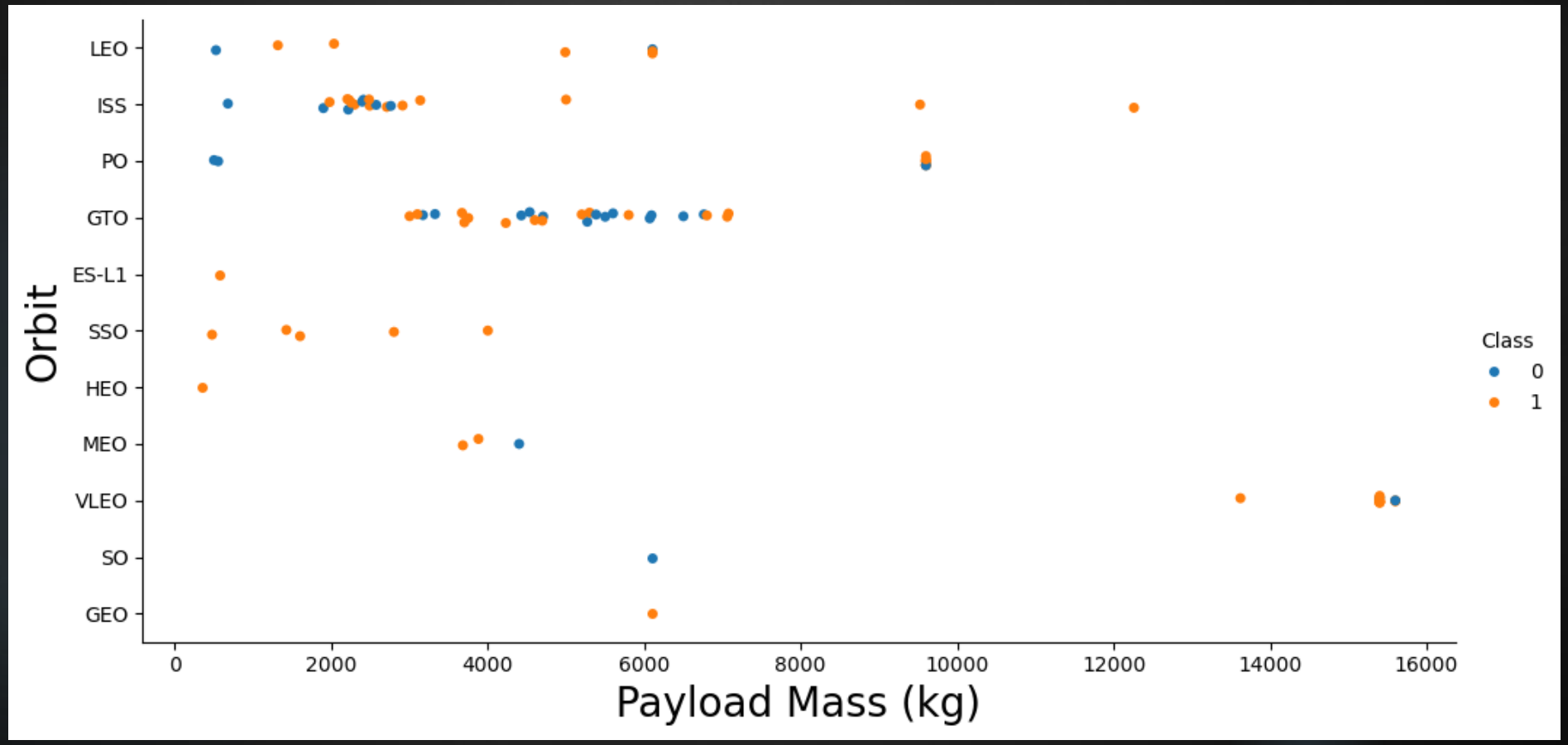
Flight Number vs Orbit

- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



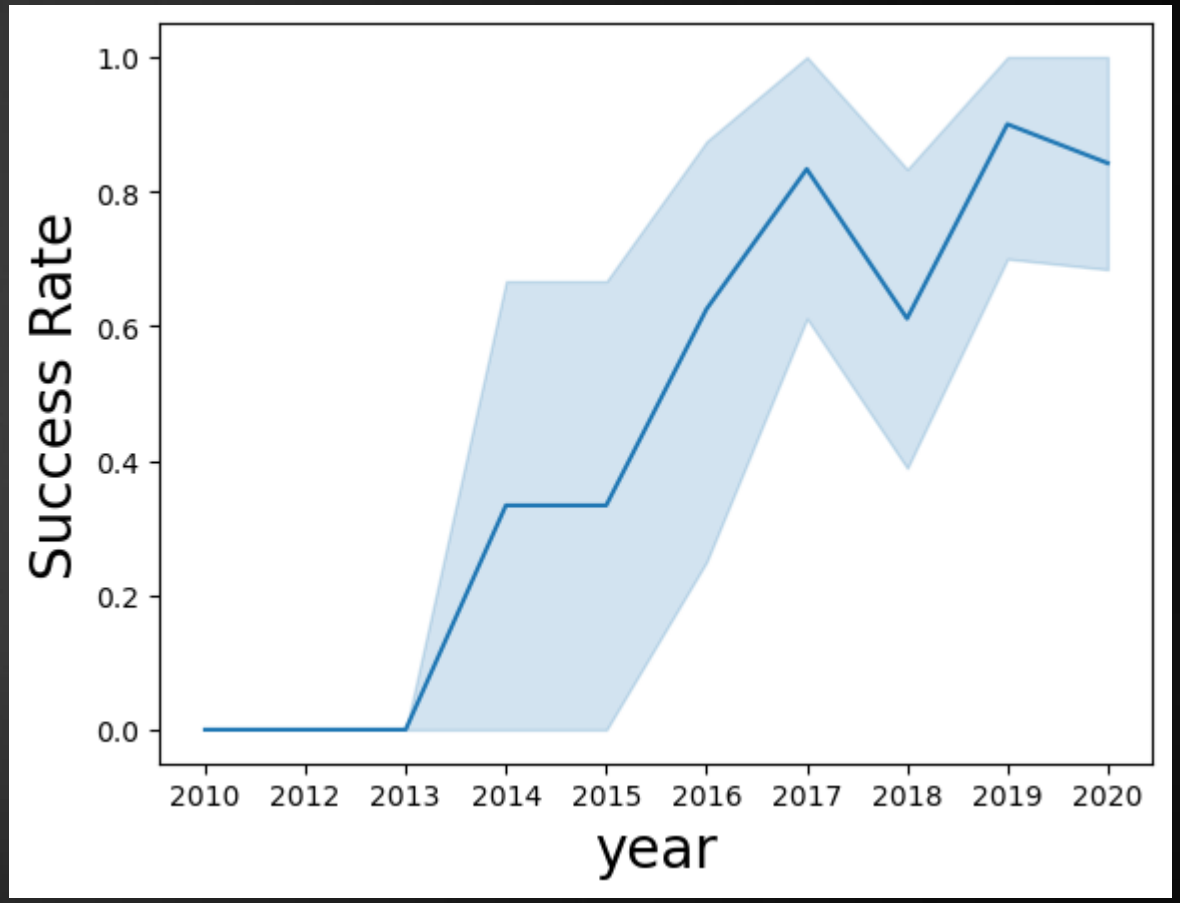
Payload Mass vs Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



Launch Site Information

```
In [13]: %sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[13]:

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

LAUNCH SITE NAMES BEGINNING WITH CCA

LAUNCH SITE NAMES

```
In [12]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Payload Information

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

TOTAL PAYLOAD MASS – 45596 KG

```
%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

AVERAGE PAYLOAD MASS BY BOOSTER VERSION F9 V1.1 – 2928.4 KG

Landing & Mission Information

```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

First Successful Ground Landing Date – Dec 22nd 2015

Booster Versions of Drone Ship Landings between 4000-6000KG Payload

```
%sql SELECT PAYLOAD FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as TOTAL_NUM FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;
```

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

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

Total Number of Successful and Failure Mission Outcomes:
99 Success, 1 Failure in flight, 1 Success with payload status unclear

Boosters Carrying Maximum Payload

```
%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

Version Numbers:

- 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

Failed Landings on Drone Ships in 2015

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

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

Count of Successful Landings between June 4th 2010 and March 20th 2017

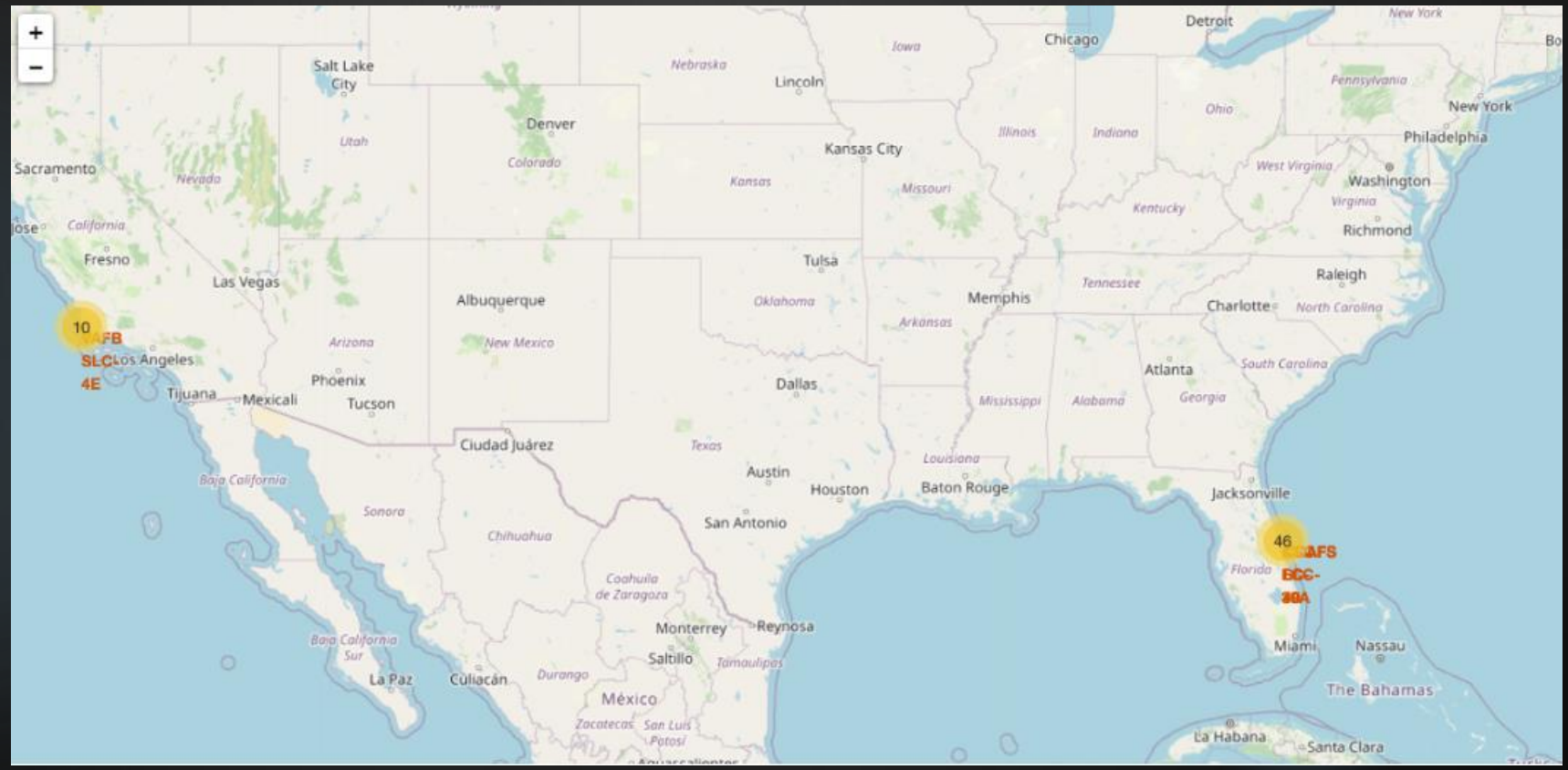
```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

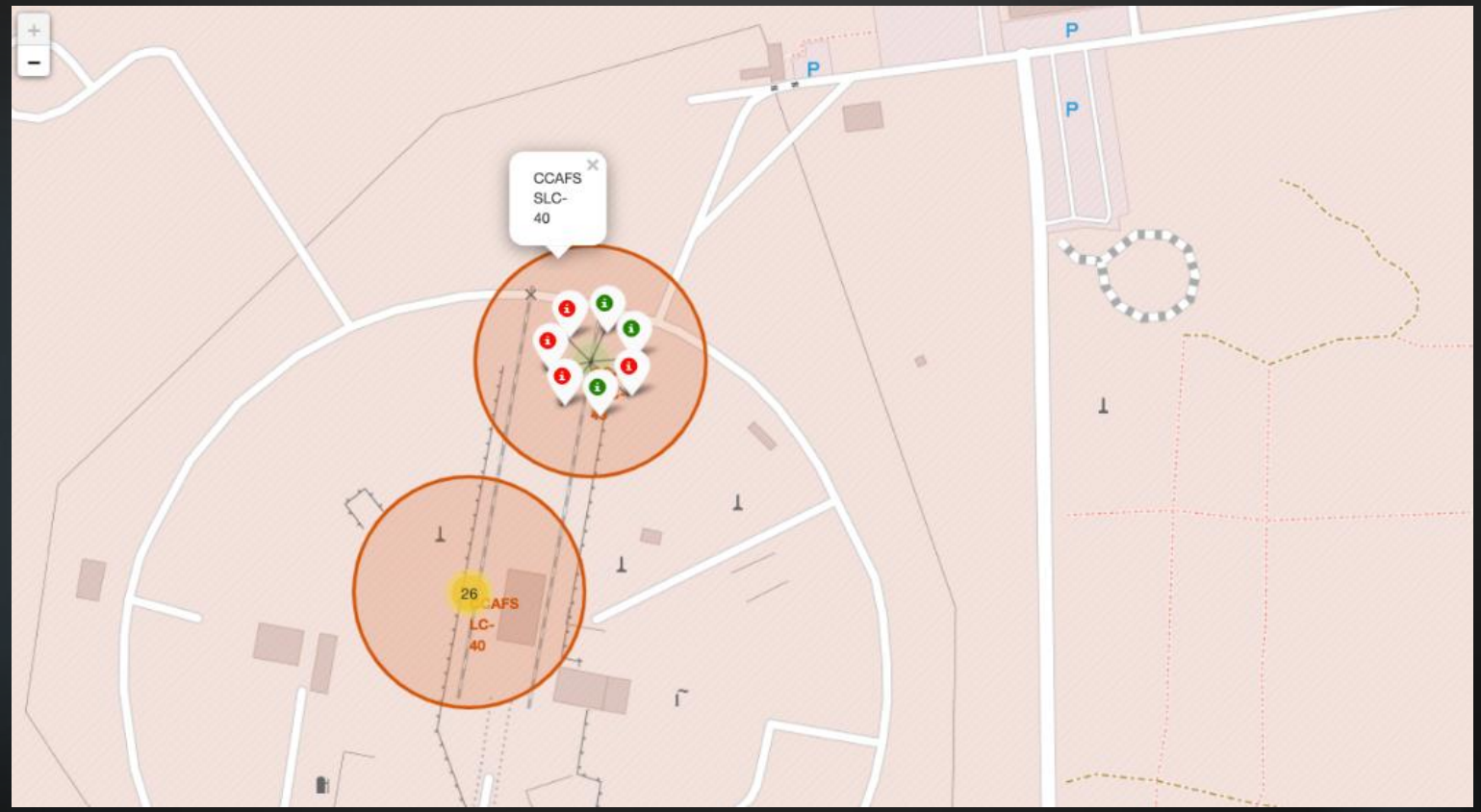
Launch Site Analysis

Near Equator: 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 - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



Site-Wise Launch Outcome

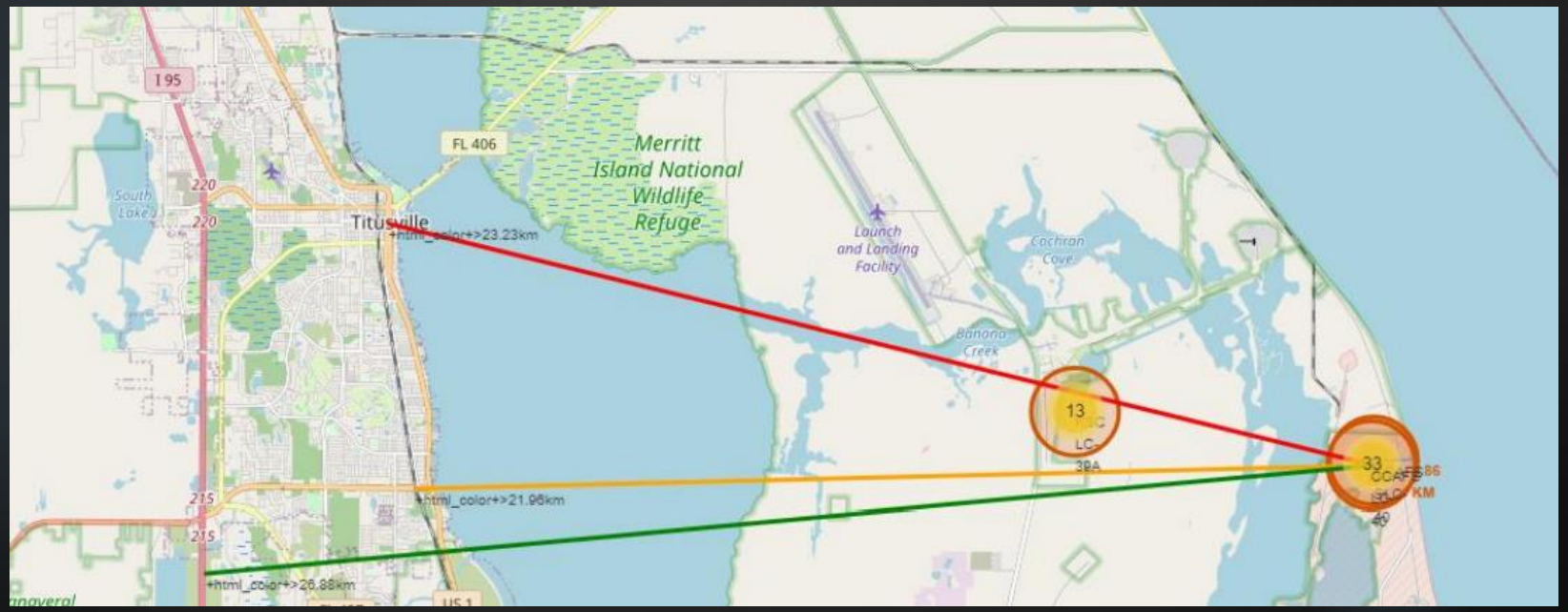
- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Launch Site Proximities

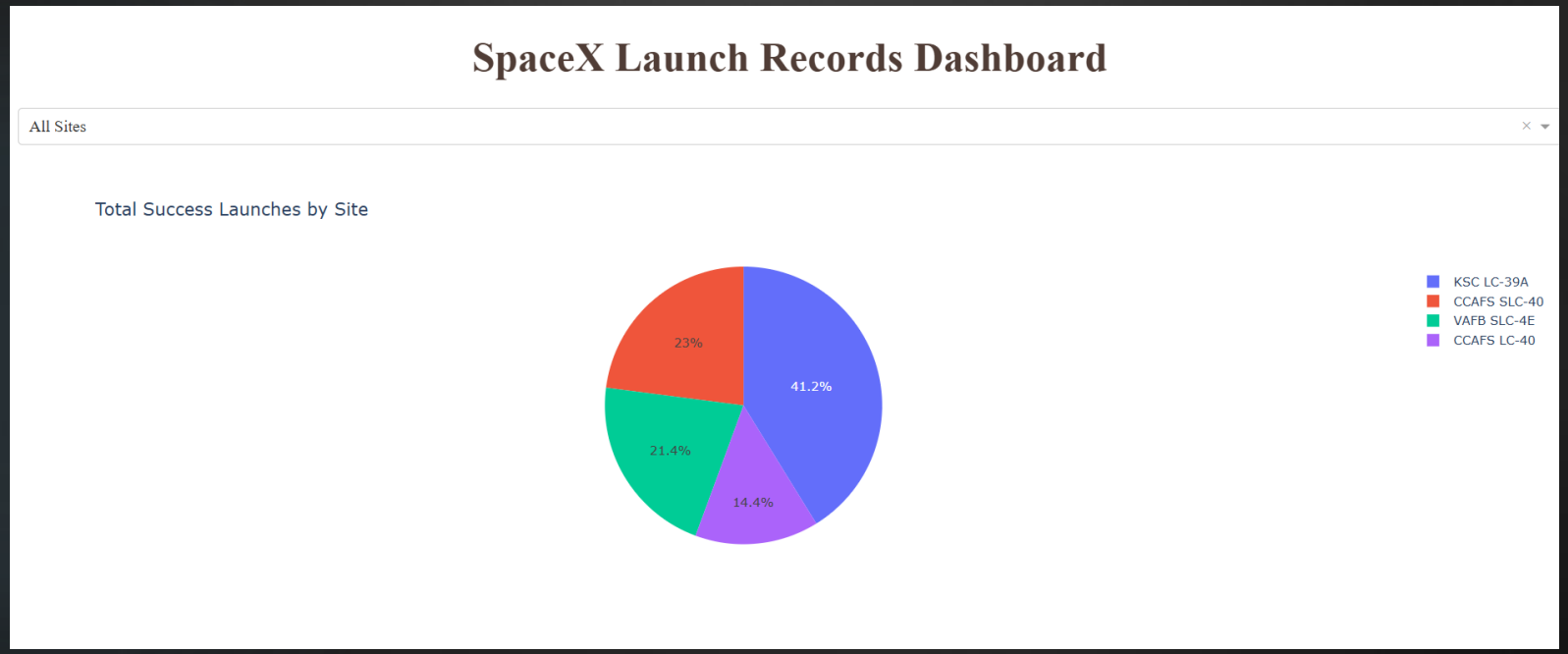
CCAFS SLC-40

- .86 km from nearest coastline 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway
- Launch sites are typically close to the coast line and far from nearest railway, city and highway



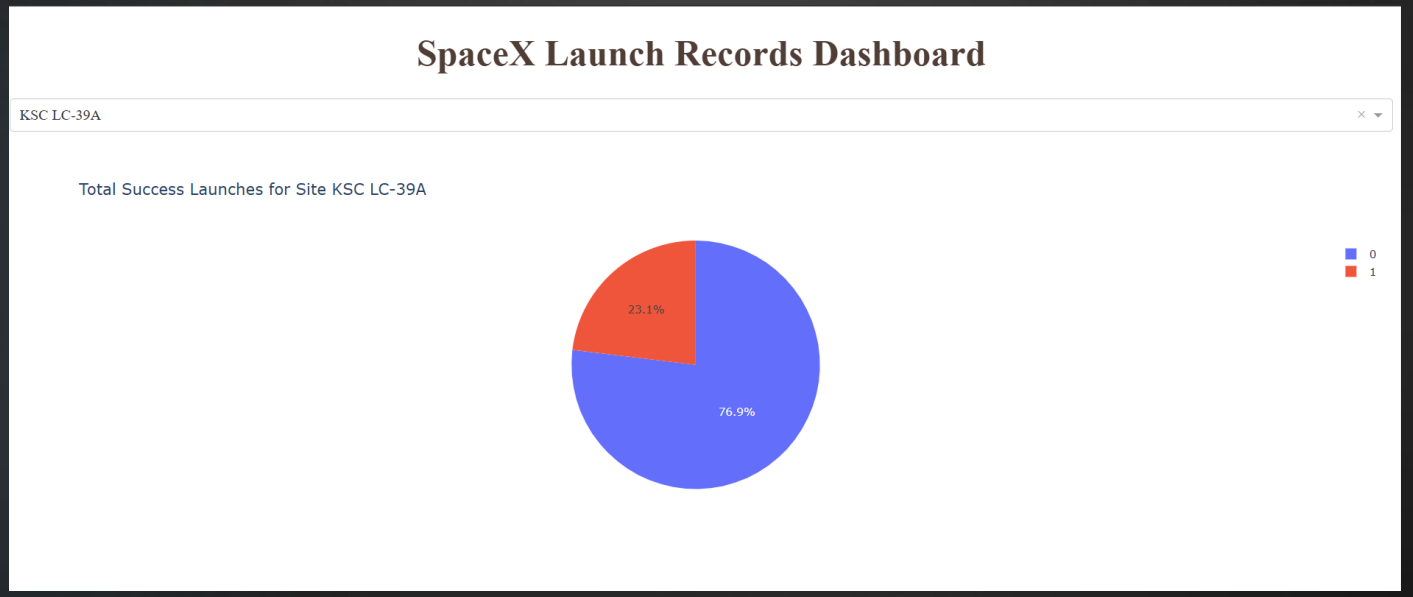
Plotly Dashboard – Site that Contributed the Most Successful Launches

- KSC LC-39A 41.2%



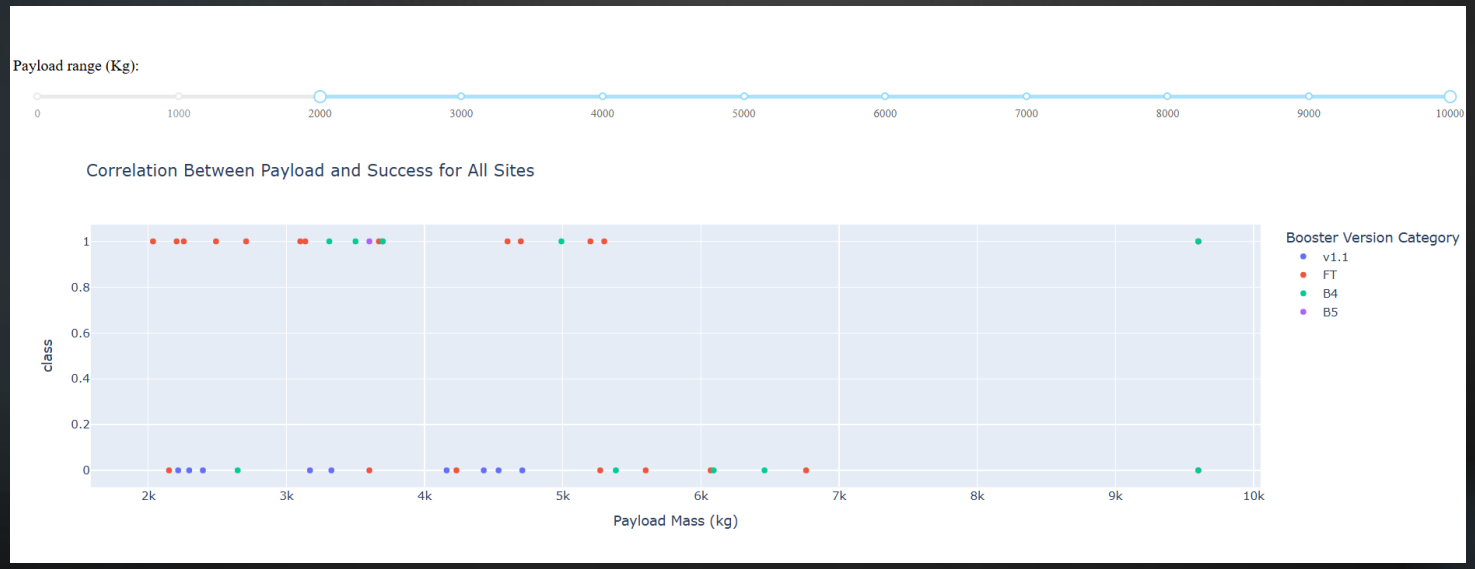
Plotly Dashboard – Most Successful Site

- KSC LC-39A 76.9%



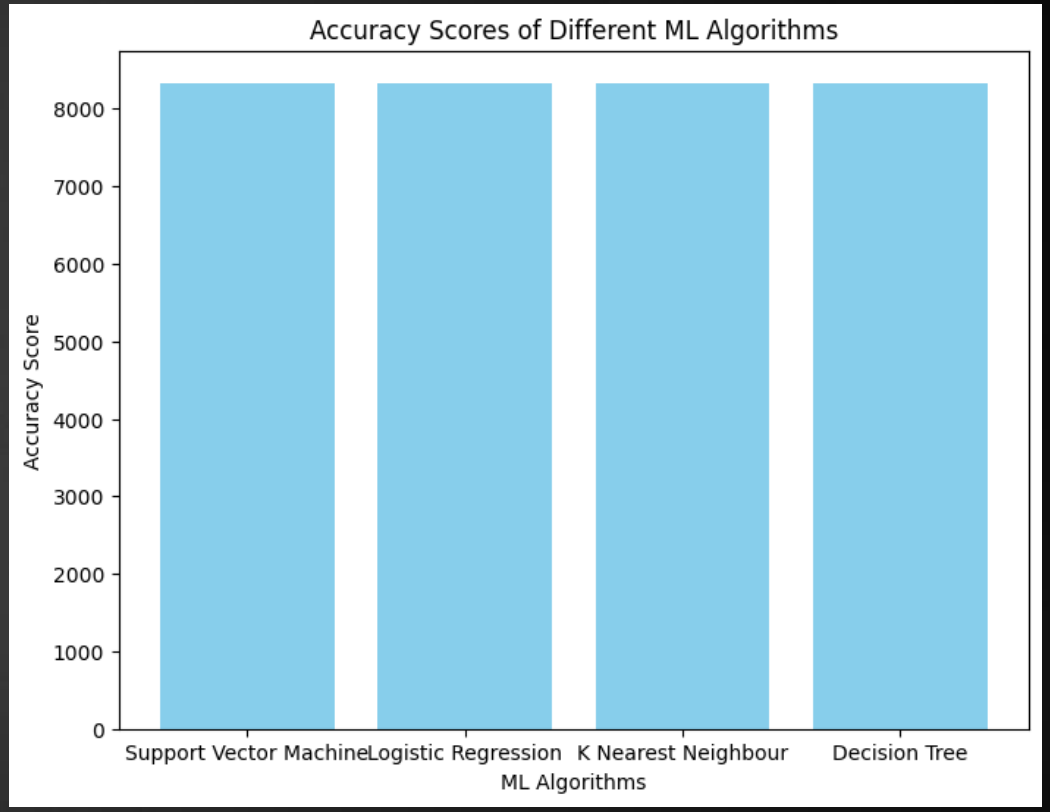
Plotly Dashboard – Payload Mass and Booster Version With Success Rate

- Which payload range(s) has the highest launch success rate? 2000-6000 KG
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? FT 50%



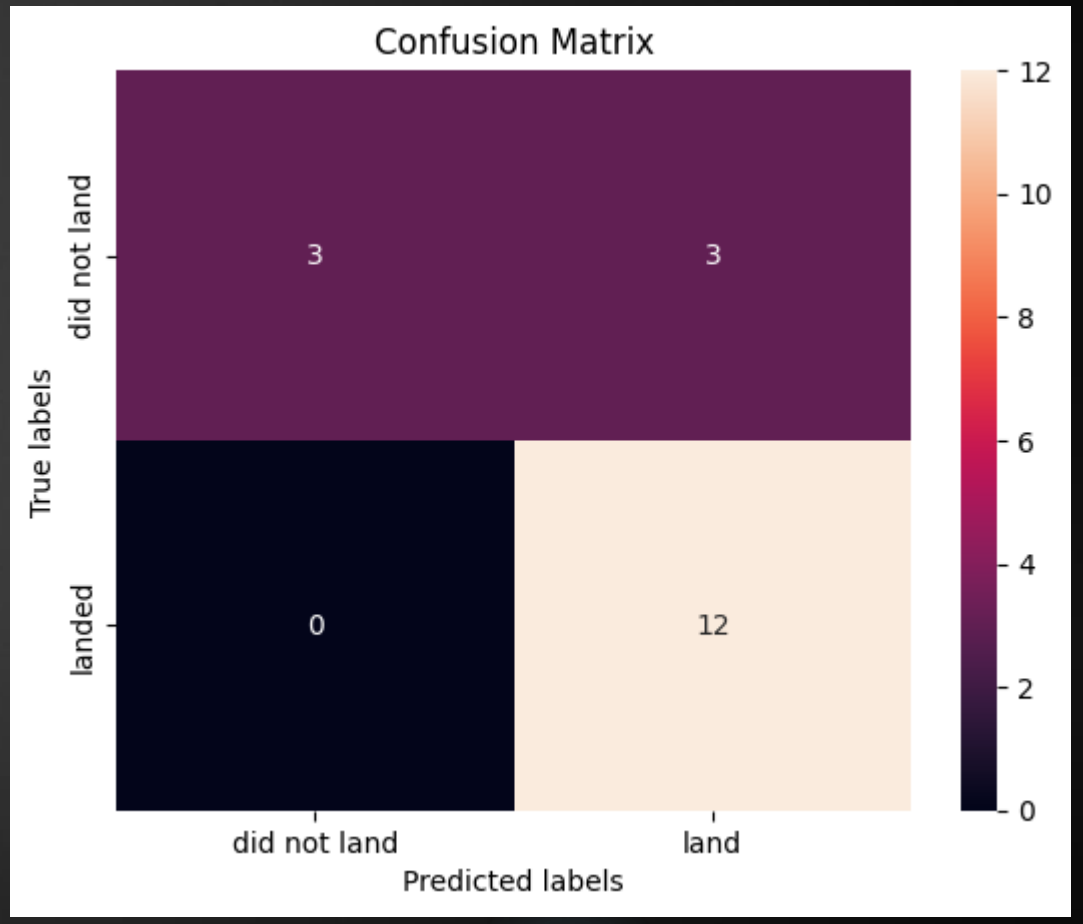
ML Predictive Analytics – Algorithm Accuracies

- 4 models were tried, and all had the same accuracy score and identical confusion matrices
- The algorithms were: SVM, Logistic Regression, KNN and Decision Trees and all had an accuracy of 83.33%



ML Predictive Analytics – Confusion Matrix

- All algorithms had 3 false positives



Conclusion

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming in the training data
- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast
- **Launch Success**: Increases over time • KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate •
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

THANK YOU