IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeffrey Oczek
December 29, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection and Wrangling (with Requests, Beautiful Soup, Pandas)
  - Exploratory Data Analysis and Visualization (with SQL, Pandas, Matplotlib, Seaborn, Plotly)
  - Interactive Analytics Dashboard (with Plotly Dash)
  - Geographical Analysis (with Folium)
  - Machine Learning Classification (with Scikit-Learn)
- Summary of all results
  - The landing success rate improved over time
  - 3 of 4 ML classification models gave the same result for the given test data
  - Further feature engineering may be necessary to improve model accuracy

Introduction

- Project background and context
  - The goal of this project is to understand the factors that make for a successful landing of the first stage of a SpaceX Falcon 9 rocket launch. Since the cost of a launch is mainly affected by if the rocket can land, we've investigated historical launch data to explore these factors. Based on this historical data, we've also
- Problems you want to find answers
  - Understand the factors that influence a successful landing of the first stage
  - Predict which landings will be successful based on known factors

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using REST APIs and web scraping
- Perform data wrangling
  - Data wrangling was performed with Pandas, particularly extracting the landing class label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four binary classification models were compared; evaluation was done on a test set using an accuracy score

Data Collection

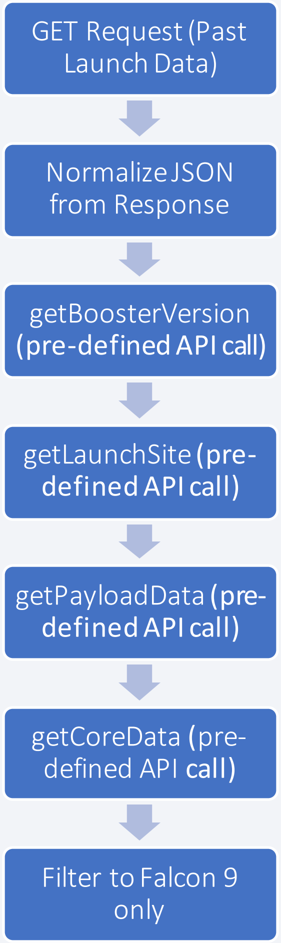
- Data collection was done by:
  - Connecting to SpaceX's REST APIs (over HTTP/GET using Requests library)
    - Multiple calls over the API to get different components of the data
    - Results are JSON
  - Web Scraping the HTML from Wiki page on SpaceX Launch Data
    - An algorithm to parse the Wiki table data was used
- Results were put in a Pandas DataFrame object
- The following pages contain flowcharts of the collection process

Data Collection – SpaceX API

- The Data Collection process relies on repeated calls to SpaceX rest API as shown in the flowchart
- The JSON results from the initial request are processed and fed to additional functions for data retrieval
- The results are truncated to include only Falcon 9 launches

- GitHub URL of the SpaceX API calls notebook:

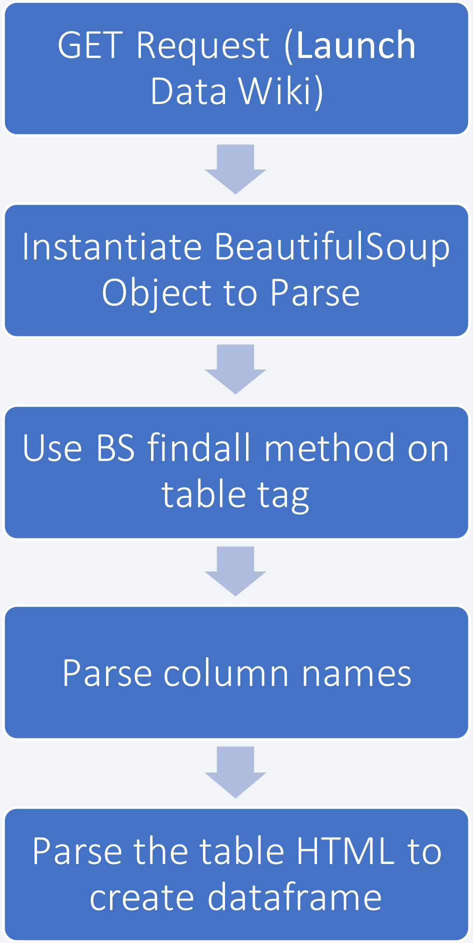
https://github.com/joczek/ibmdscapstone/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

- An initial GET request is used to retrieve the HTML from the Wiki page
- The BeautifulSoup library is used to parse the HTML
- Some additional steps to parse the table format are used
- Results are stored in a Pandas dataframe
- GitHub URL of the web scraping notebook:

https://github.com/joczek/ibmdscapstone/blob/main/Web%20Scraping.ipynb



Data Wrangling

- Data Wrangling was performed to inspect and clean the data
- A binary class label is added based on landing outcomes
- This process was primarily performed using the Pandas library

- GitHub URL of data wrangling related notebook:

https://github.com/joczek/ibmdscapstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- EDA with Data Visualization was primarily performed using the Matplotlib, Seaborn and Plotly libraries
- Charts plotting relationships between the features, such as flight number, orbit type, year, launch site, landing outcome, were used to identify meaningful trends in the data
- GitHub URL of EDA with data visualization notebook:

https://github.com/joczek/ibmdscapstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Queries for EDA included:
    - Distinct launch sites
    - Records with Launch Site starting with CCA
    - Total Payload Mass for Customer: NASA (CRS)
    - Average Payload Mass carried by booster version F9 v1.1
    - Date of first successful landing outcome on ground pad
    - Names of boosters with success on drone ship, payload mass between 4000 and 6000
    - Total number of success and failure mission outcomes
    - Booster names that have carried maximum payload mass
    - Display month, year, failure landing outcomes in drone ship, booster version, launch site for 2015 launches
    - Rank of landing outcomes between date of 04-06-2010 and 20-03-2017
- GitHub URL of EDA with SQL notebook:

https://github.com/joczek/ibmdscapstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- The Folium map had the objects
  - Markers – These were text labels that displayed geographical launch site information, or icons that displayed the launch successes and failures at each location
  - Circles – These are to highlight areas on the map, and as parent objects to popup objects
  - Lines – The line on the map is illustrating the shortest path from one of the launch sites to the nearest coast.
- These items help get a visual understanding of how SpaceX launches per site occur, and broader context for these sites like proximity to other important areas
- GitHub URL of completed interactive map with Folium map:

https://github.com/joczek/ibmdscapstone/blob/main/Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard consisted of:
  - A dropdown selector with each of the four launch sites and an 'All sites' option.
    - This applied a filter to the charts below it
  - A Payload Mass range selector
    - This applied a filter to the bottom chart (scatter plot)
  - A Pie Chart showing landing success data
    - When the Launch Site dropdown was 'All Sites', the pie chart showed the percentage of successful launches that each Launch site had out of all successful launches between the four sites
    - When the Launch Site dropdown was a specific Launch site, it showed the percentage of successful and unsuccessful launches from that site

- GitHub URL of Plotly Dash lab:

https://github.com/joczek/ibmdscapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared. The algorithms were
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - K-Nearest Neighbors
- The data was split into a train/test split with a 20% test size.
- Final hyperparameters were optimized for each model by applying grid search with cross validation (k-fold CV, k=10).
- To determine the best model, each of the four models were scored by fitting the model to the training data, predicting the output labels on the test set, then comparing the predicted test labels to the true test labels (using accuracy score, and visually analyzing with a confusion matrix).
- GitHub URL of predictive analysis lab:

https://github.com/joczek/ibmdscapstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

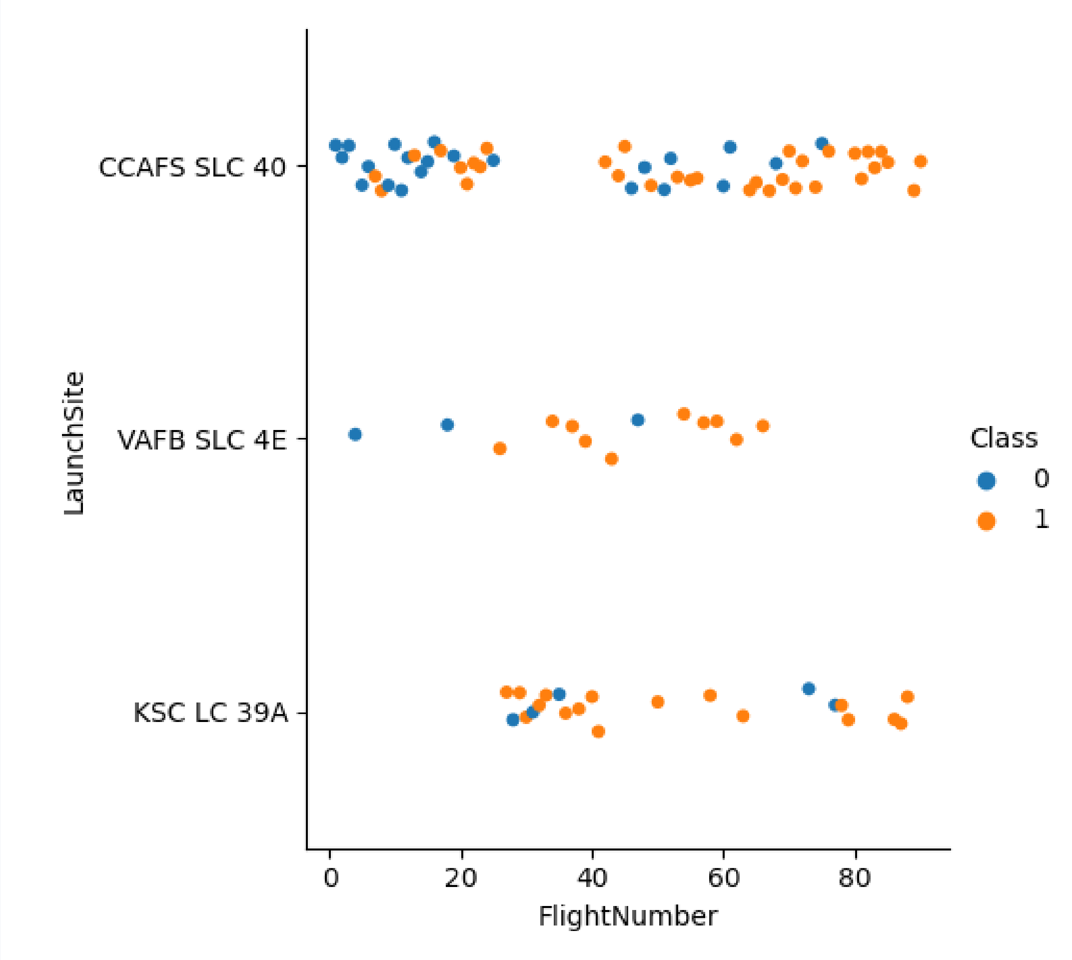
- Exploratory data analysis results
  - Timeframe is a significant factor in landing success rate
  - Perhaps not all flights have been engineered to require relanding success
    - (a follow-up question to investigate: have higher payload mass flights placed less importance on the landing outcome?)
- Interactive analytics demo in screenshots
  - Screenshots are presented in Section 4
- Predictive analysis results
  - Model accuracy needs to be improved if accurate predictions are necessary

Section 2

Insights drawn from EDA

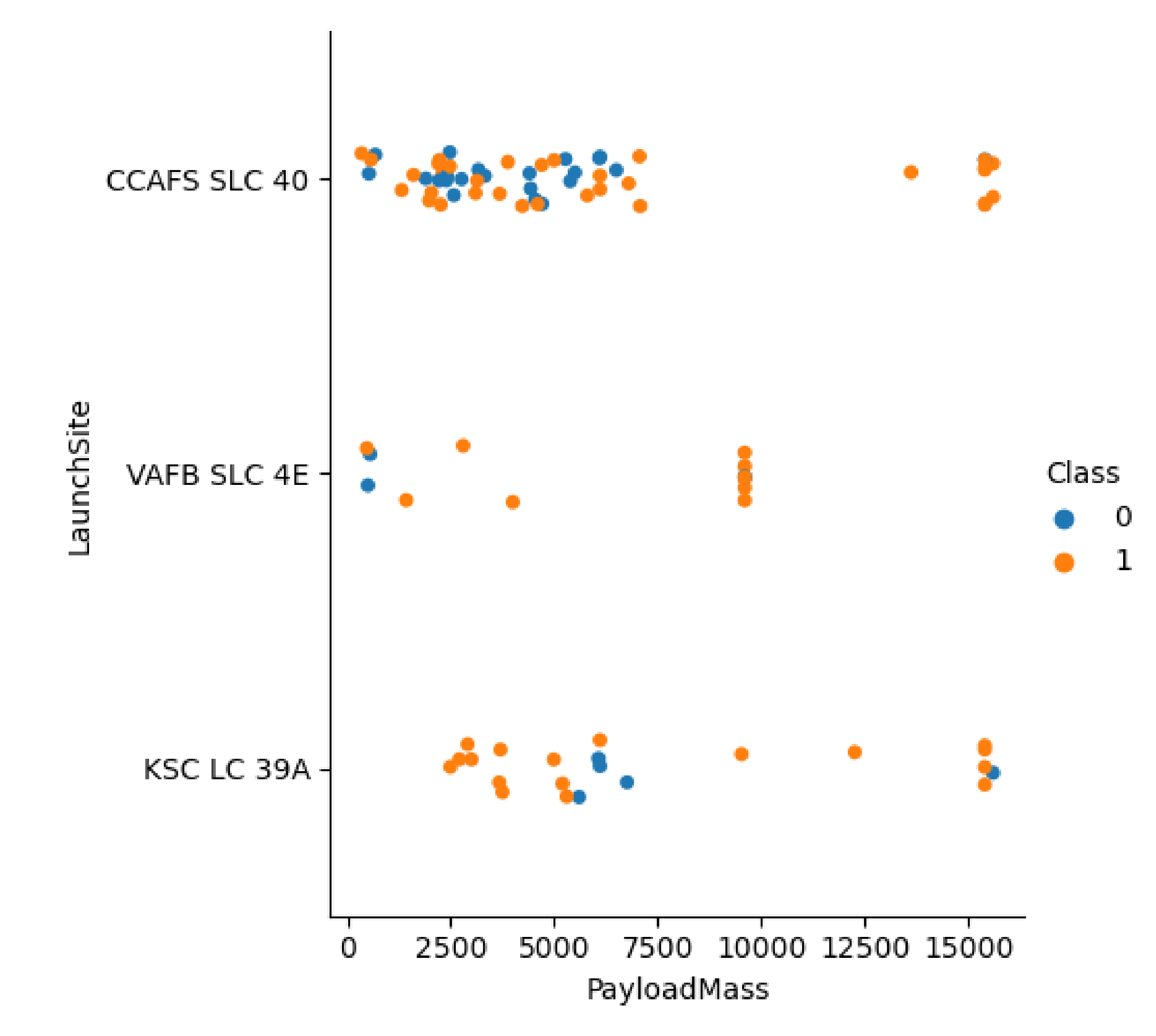
Flight Number vs. Launch Site

- CCAFS SLC 40 is responsible for the most launches
- KSC LC 39A only started with launches after about 20 flights
- VAFB SLC 4E has the fewest number of launches



Payload vs. Launch Site

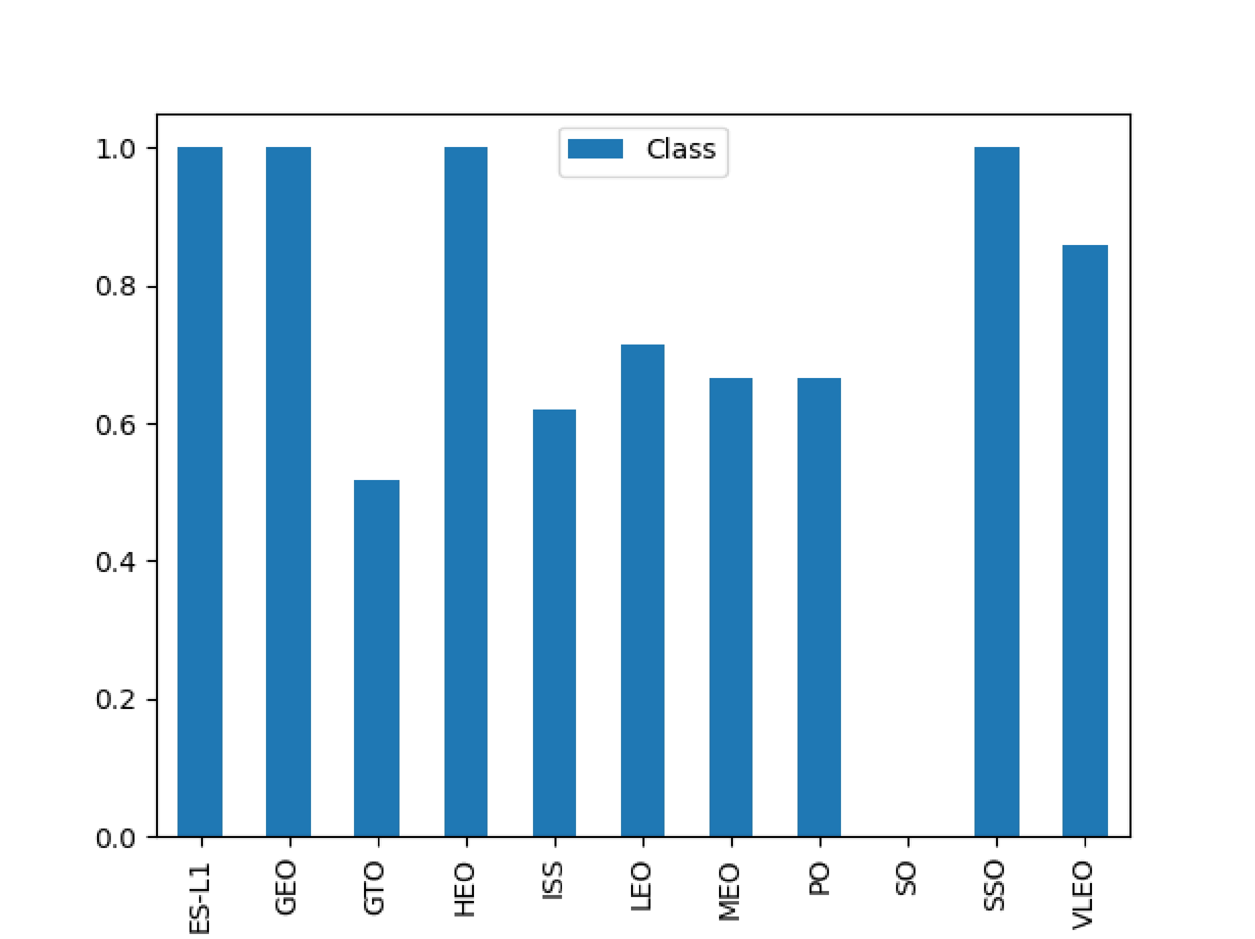
- CCAFS SLC-40 has most low (0-7500kg) payload masses
- VAFB SLC 4E has almost all launches with ~10,000kg payload mass
- CCAFS SLC-40 also had the most launches in general



Success Rate vs. Orbit Type

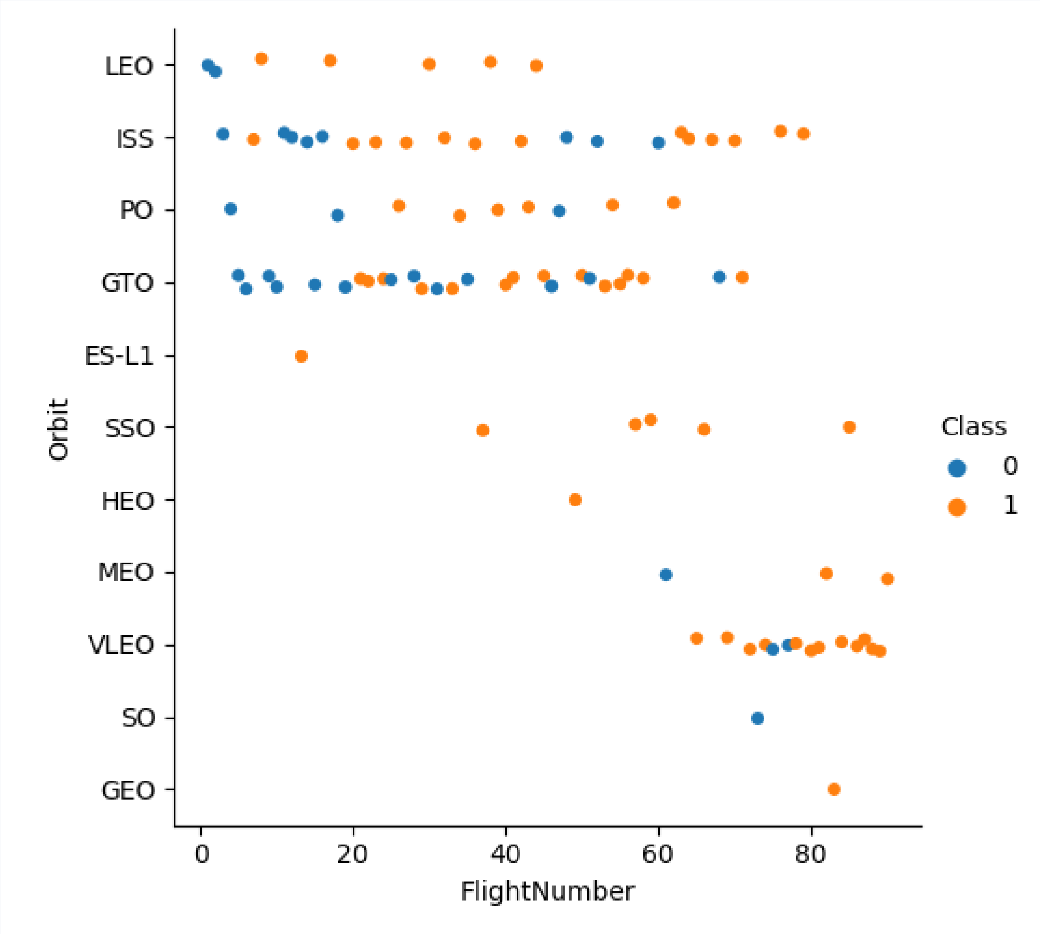
- There are 4 orbit types with perfect success rates (ES-L1, GEO, HEO, SSO)

- There is 1 orbit type with a 0 success rate (SO)



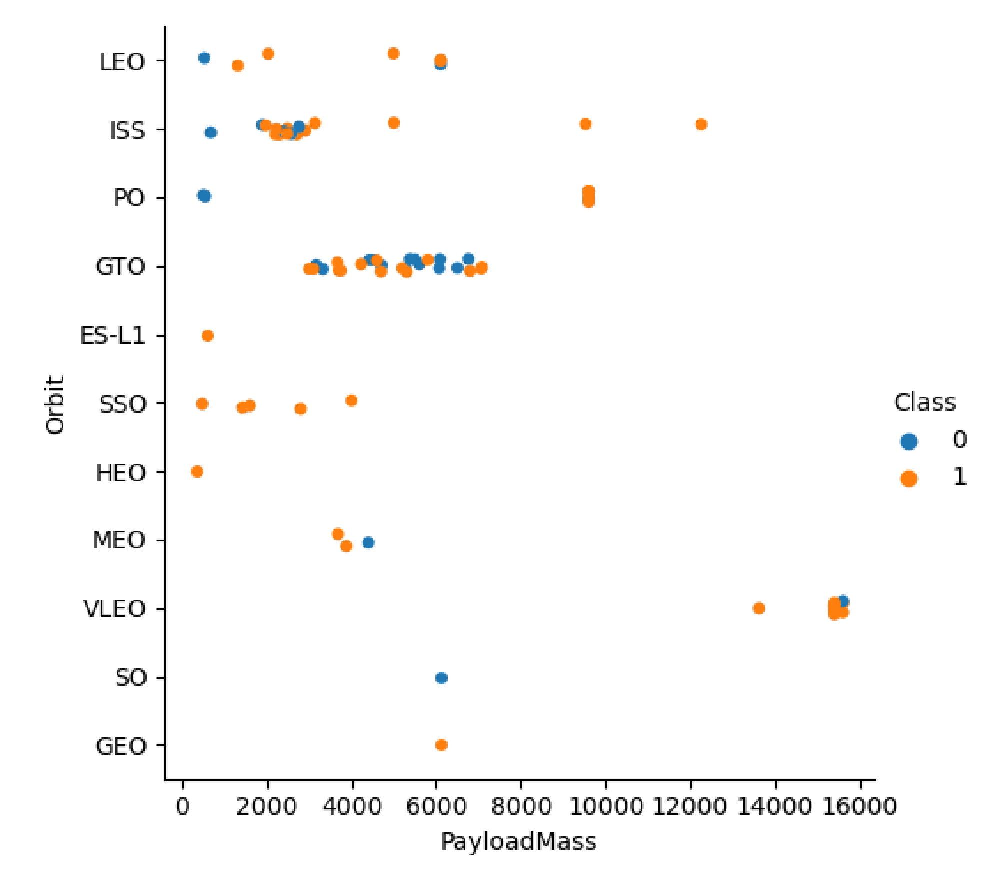
Flight Number vs. Orbit Type

- We can see that successes improve as flight number increases.
- If we use magnitude of flight number as a proxy for time/experience, it is obvious that successes are increasing over time



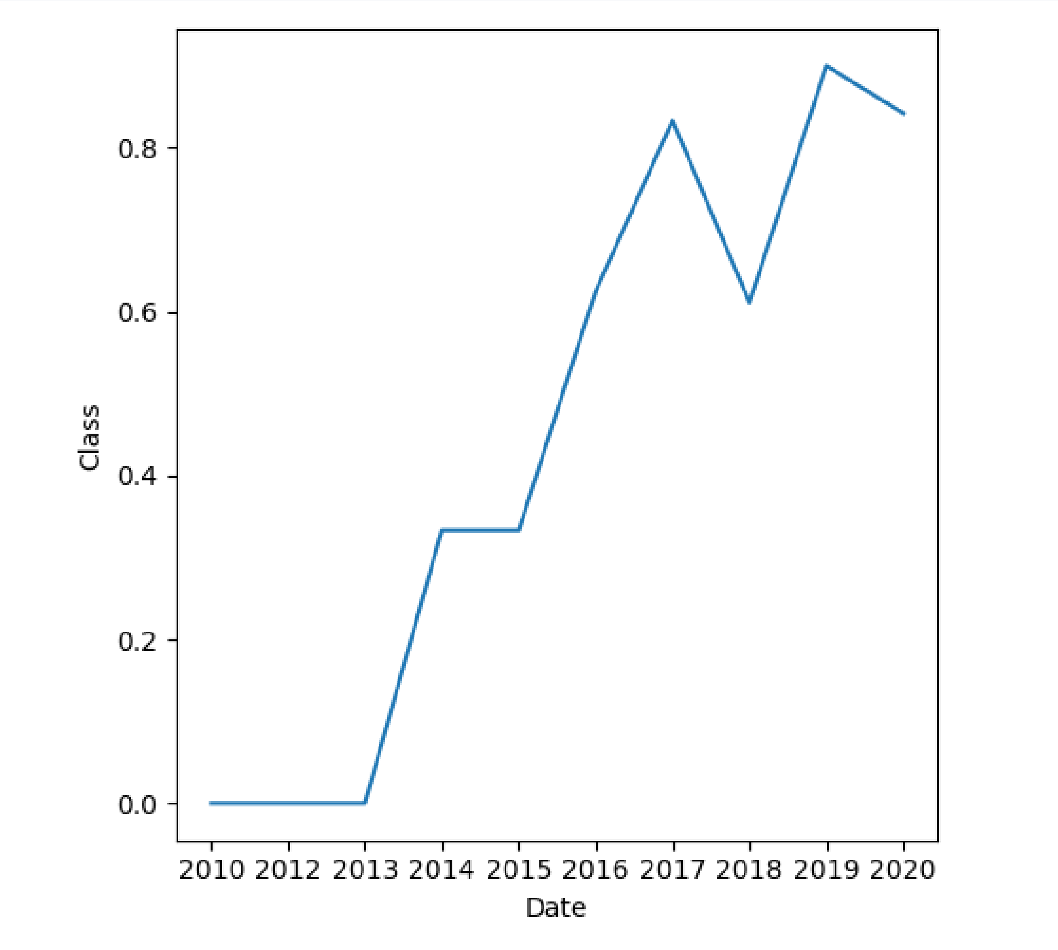
Payload vs. Orbit Type

- The only orbit type that has 100% failure is SO. Some have a 100% success rate, such as ES-L1, SSO, HEO, GEO.
- In general, we can see that lower payload masses are almost all of the successes (there is only one success visible from 8000kg and above)



Launch Success Yearly Trend

- The yearly success rate mostly increases over time, but with some downward fluctuation compared to the previous years in 2018 and 2020.



All Launch Site Names

- Find the names of the unique launch sites
- This query shows the four launch sites in the dataset. It limits to only unique values rather than all rows (via the DISTINCT keyword)

```
In [8]: %sql select distinct Launch_site from SPACEXTBL

 * sqlite:///my_data1.db
Done.

Out[8]:
```

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`
- This query uses string wildcard functionality (the LIKE operator in the WHERE clause) to return rows that begin with CCA.

In [11]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Out[11]:

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
- This query uses the sum(x) function to map the values in the PAYLOAD_MASS__KG_ field to an aggregated sum row.

```
In [20]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)"

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

Out[20]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

- This query is a straightforward SELECT query. The avg(x) function takes all values in this column (limited by the WHERE clause criteria) and returns one 'mean' row.

```
%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
- This is a straightforward query that assesses the first successful landing with the 'Success (ground pad)' value in the Landing Outcome field. The min(x) function returns the minimum value for the Date field under this criteria.

```
In [30]: %sql select min(Date) from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)'

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

Out[30]:

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
- This query relies on the DISTINCT keyword and BETWEEN operator to select these booster_version values

```
In [34]: %%sql
select distinct Booster_Version
from SPACEXTBL
where [Landing _Outcome] = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000

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

Out[34]:

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 2 text values that show 'Success' that look similar in the first query. The one with 1 result has extra whitespace, but can be cleaned with the trim function.

```
%sql select Mission_Outcome, count(*) from SPACEXTBL 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

```
%%sql
select trim(Mission_Outcome) as Mission_Outcome_Trimmed,
count(*)
from SPACEXTBL
group by Mission_Outcome_Trimmed
```

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

Mission_Outcome_Trimmed	count(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The unique booster versions are listed here by using the DISTINCT keyword in the SQL query.
- The maximum payload mass is determined by subquery referenced in the outer queries WHERE clause.

```
%%sql
select distinct Booster_Version
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (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
- There are two records with this landing_outcome in 2015. One in January, and one in April. The Launch site is the same for both. The booster version also differs slightly (F9 v1.1 B1012 vs B1015)

```
%%sql
select substr(Date,4,2) as month,
substr(Date,7,4) as year,
Booster_Version,
launch_site
from SPACEXTBL
where [Landing _Outcome] = 'Failure (drone ship)'
and year = '2015'
```

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

month	year	Booster_Version	Launch_Site
01	2015	F9 v1.1 B1012	CCAFS LC-40
04	2015	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, in descending order

```
%%sql
select [Landing _Outcome],
count(*) as outcome_count
from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017'
group by [Landing _Outcome]
order by outcome_count desc;
```

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

Landing _Outcome	outcome_count
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

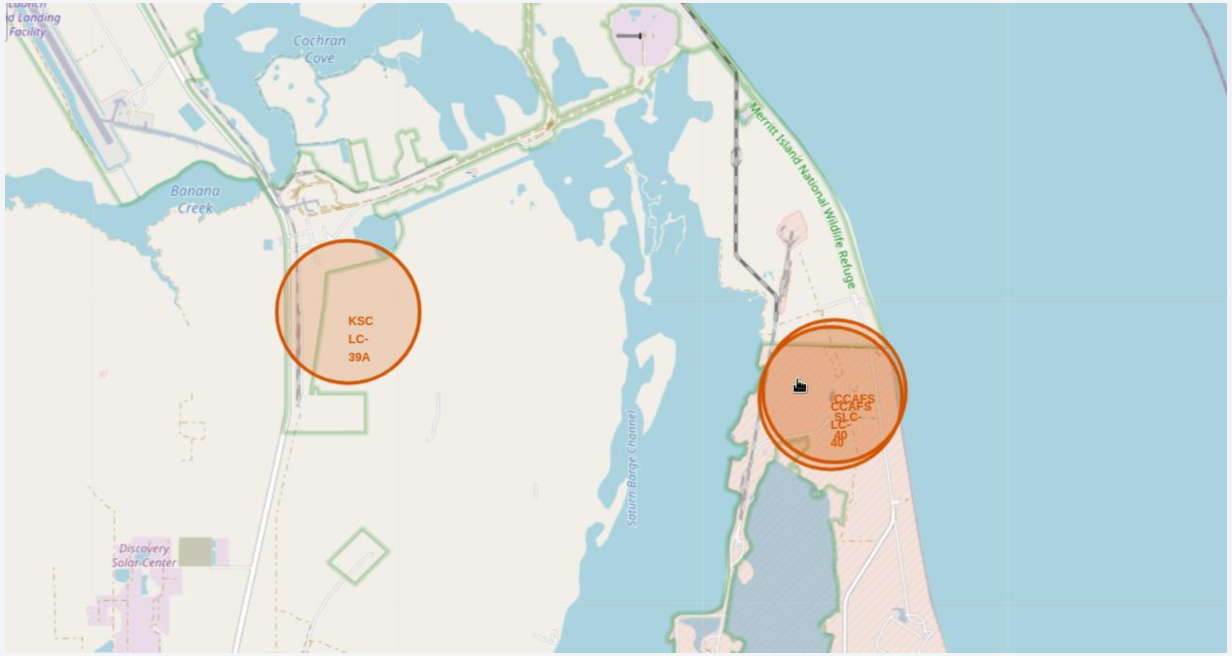
- The query used is a basic SELECT query with GROUP BY
- A landing outcome labeled as 'Success' has the highest count

Section 3

Launch Sites Proximities Analysis

Global Summary of Launch Site Locations

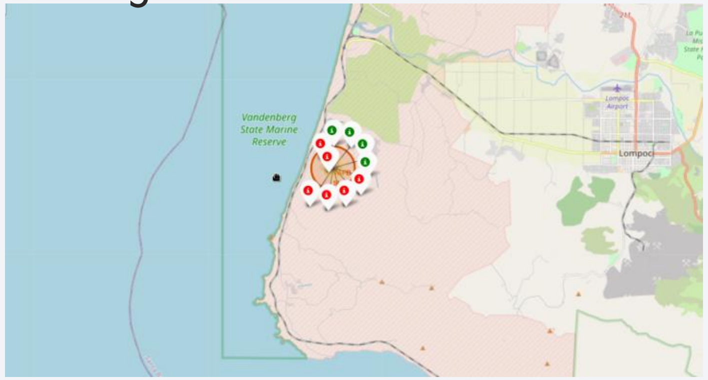
- Pictured are the launch sites used for SpaceX launches
- The 2nd image is zoomed in to make the 3 clustered sites appear more clearly



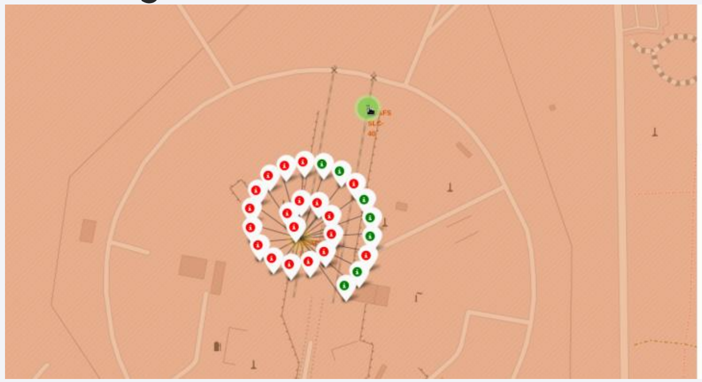
Successful and Unsuccessful Landings by Site

- Green Markers Indicate Success
- Red Markers Indicate Failure

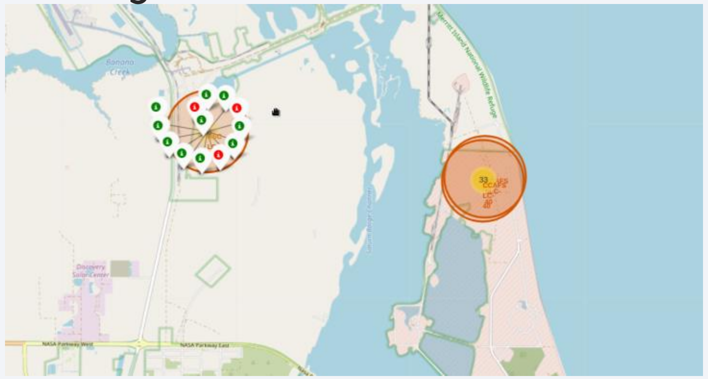
Landings at VAFB SLC-4E



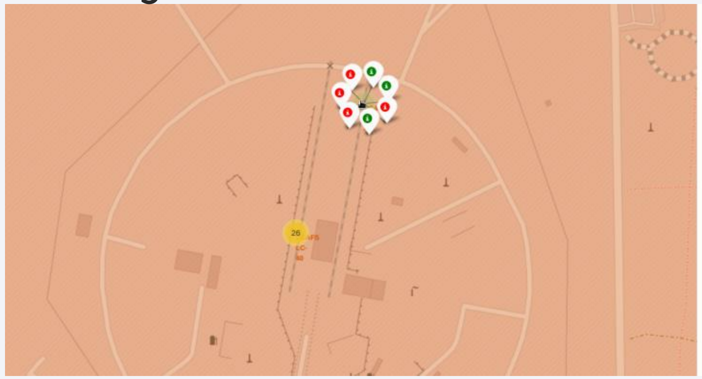
Landings at CCAFS LC-40



Landings at KSC LC-39A

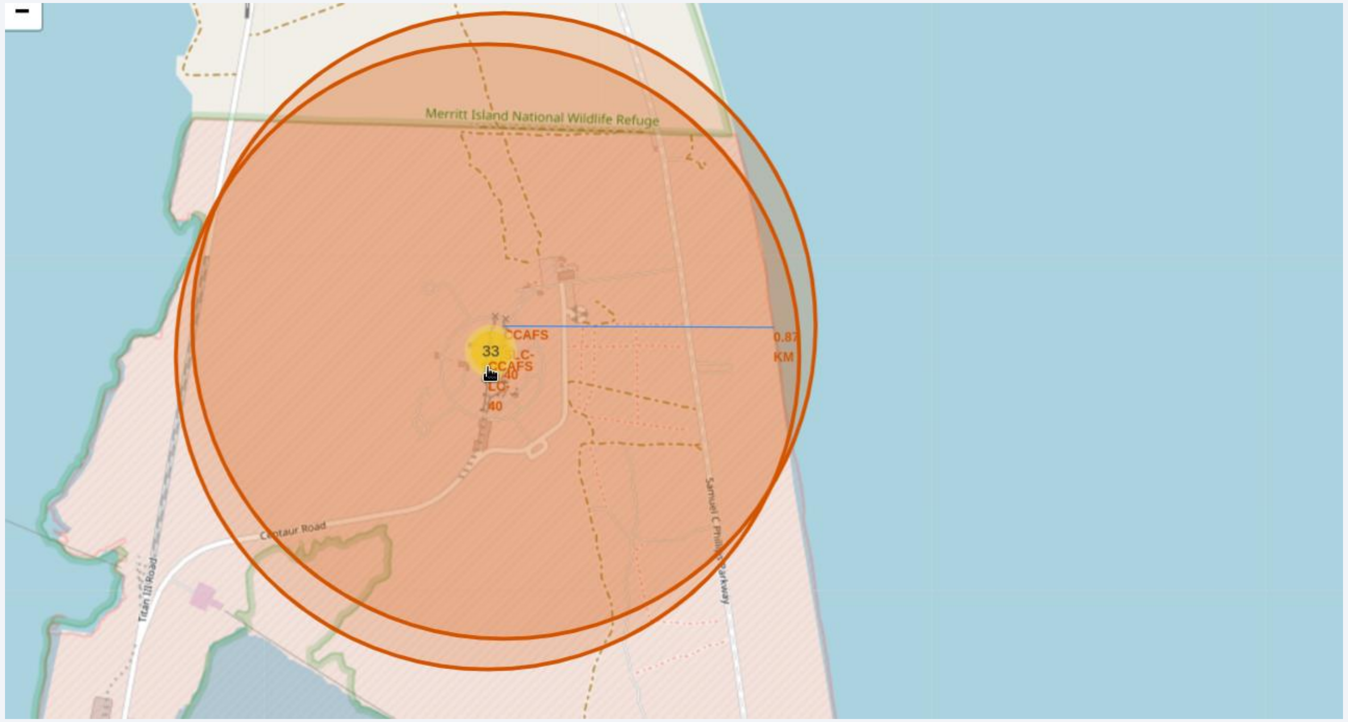


Landings at CCAFS SLC-40



Proximity to Coastline

- The map shows the coastline is less than 1km from the launch site CCAFS SLC-40
- The launch site is close to a the coastline
- Each of the launch sites are close to a coast and a railway
- The launch sites are generally further away from cities

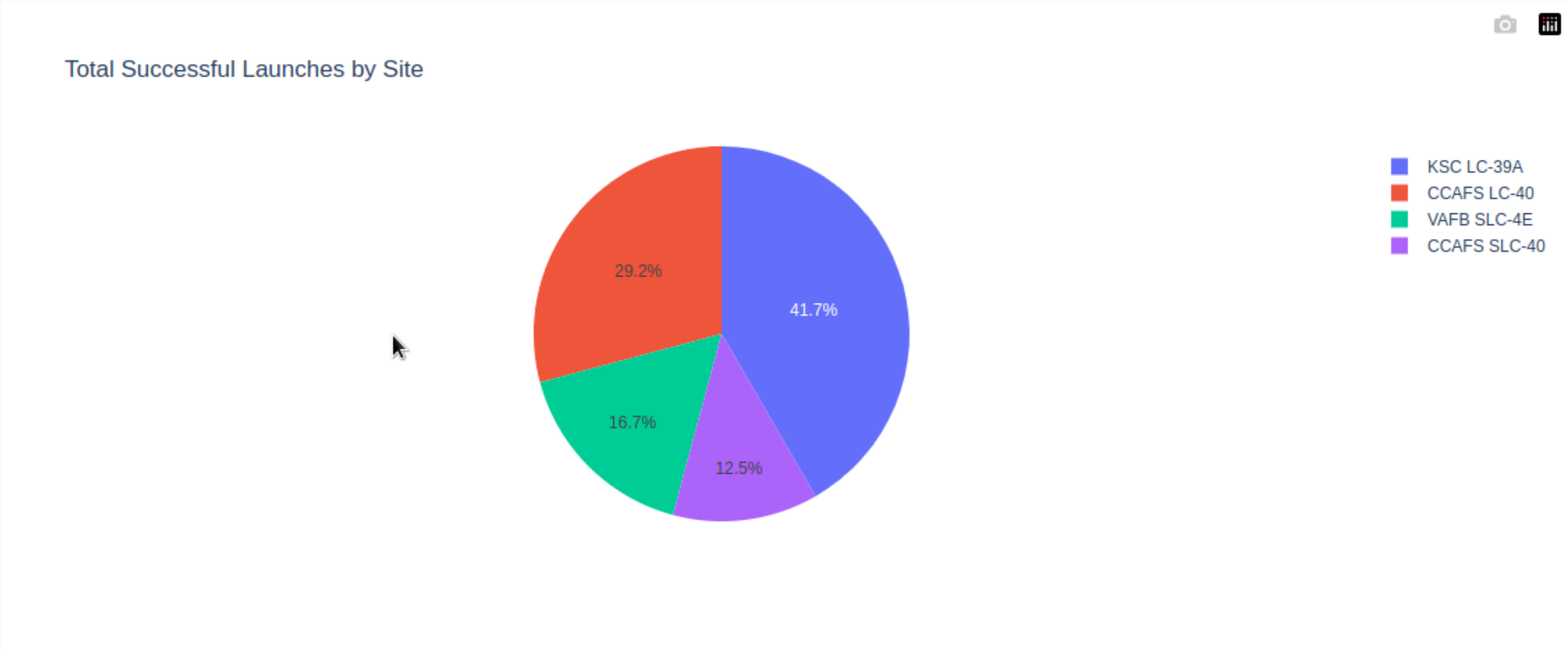


Section 4

Build a Dashboard with Plotly Dash

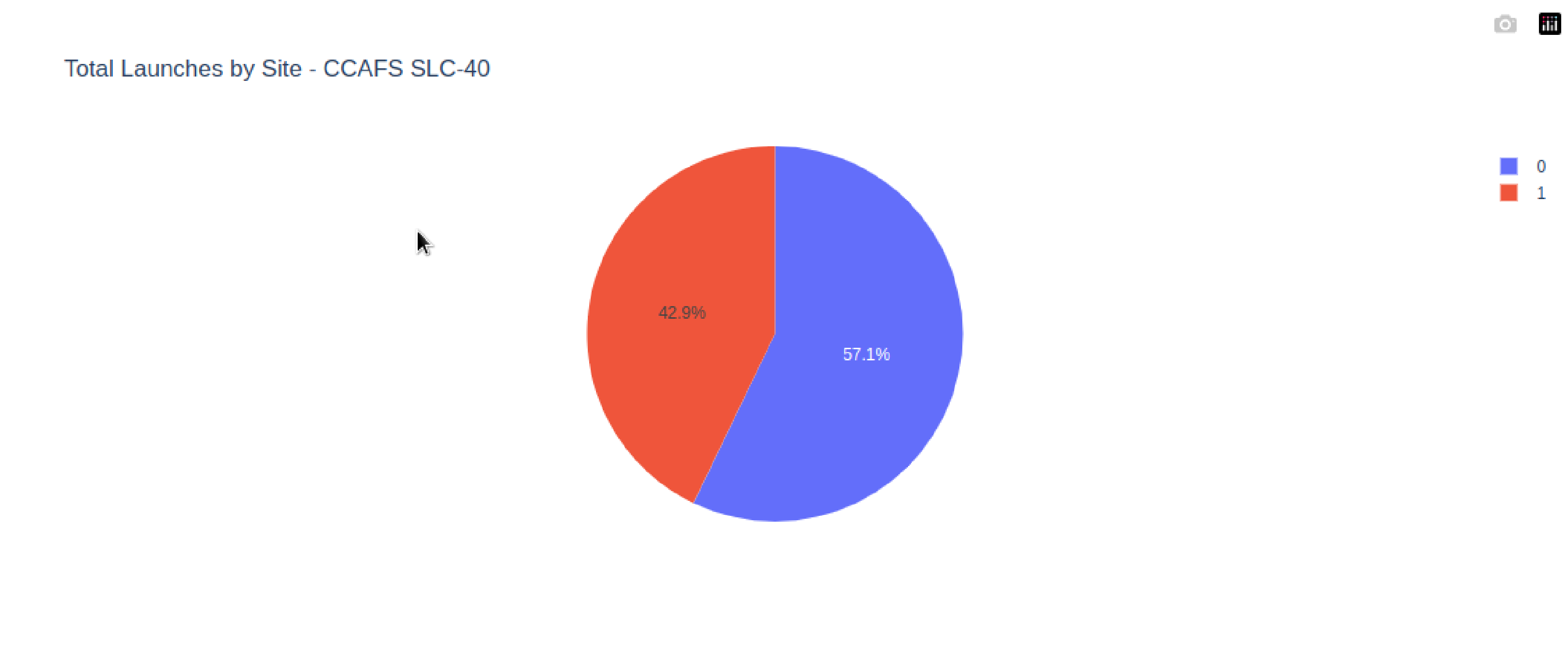
Successful Landings by Launch Site

- The total number of successful landings is pictured below
- The site KSC LC-39A has the highest number of successful landings, however, it is not the launch site with the highest success rate (Successful launches over total launches)



Launch Successes for CCAFS SLC-40

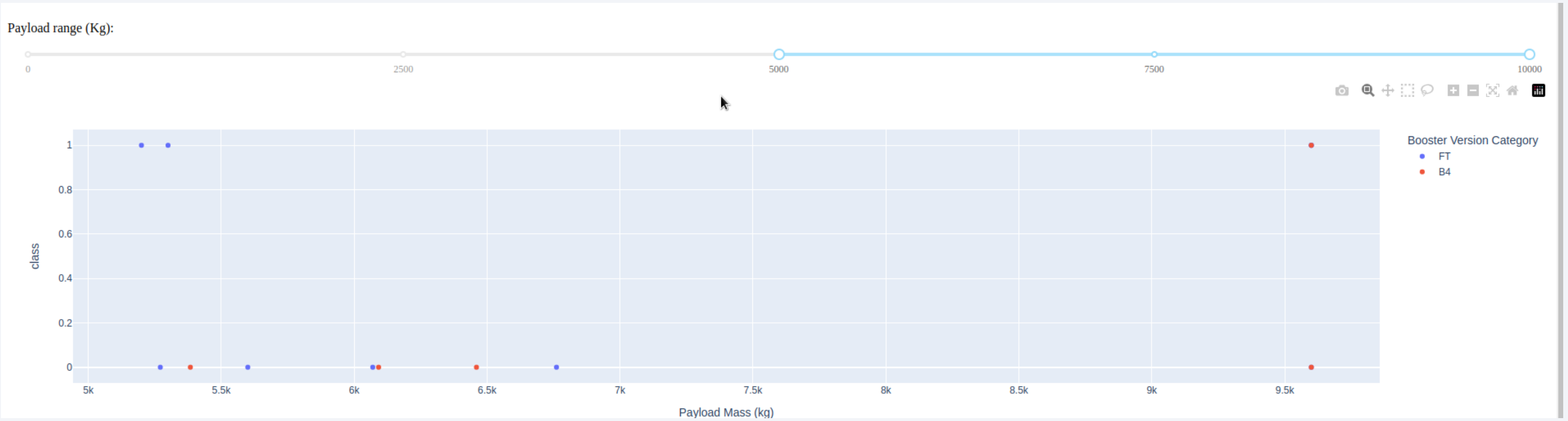
- The launch site with the highest landing success rate is CCAFS SLC-40
- The success rate is 42.9%



Flight Successes by Payload Mass Category

The success rate of flights with payload masses below 5000kg is visibly higher, and there is a variety of booster versions across this category

The success rate of flights with Payload masses above 5000kg is much lower, and there are only 2 booster versions in this range

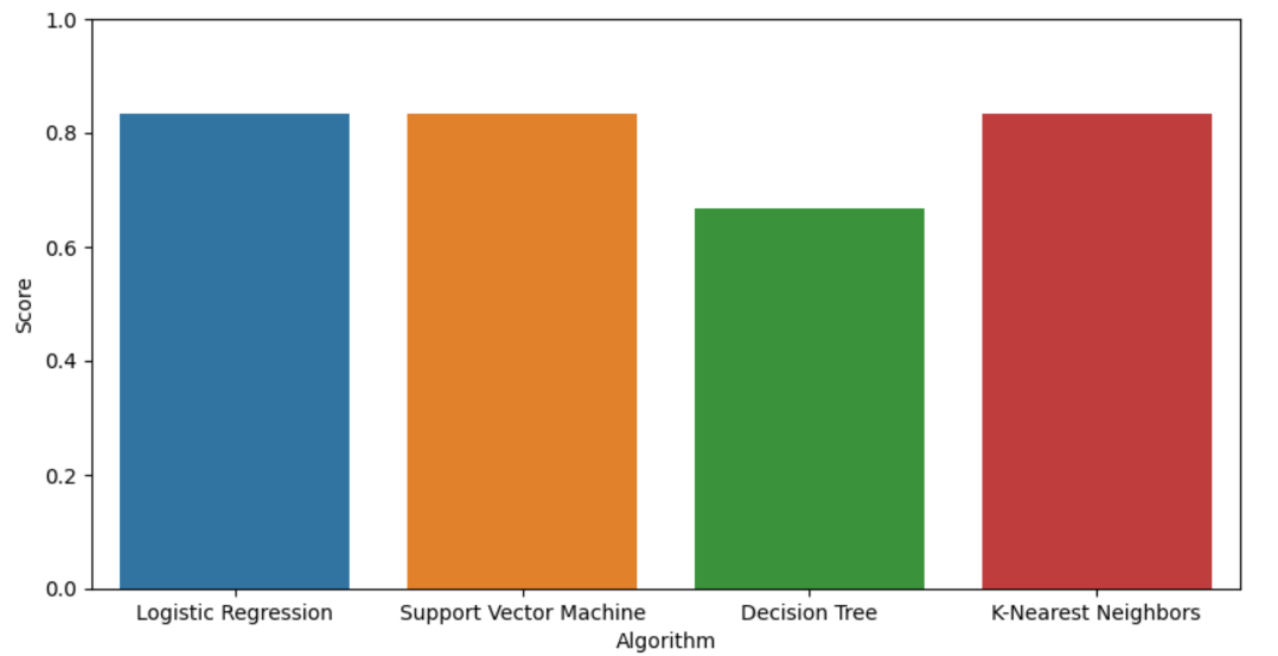


Section 5

Predictive Analysis (Classification)

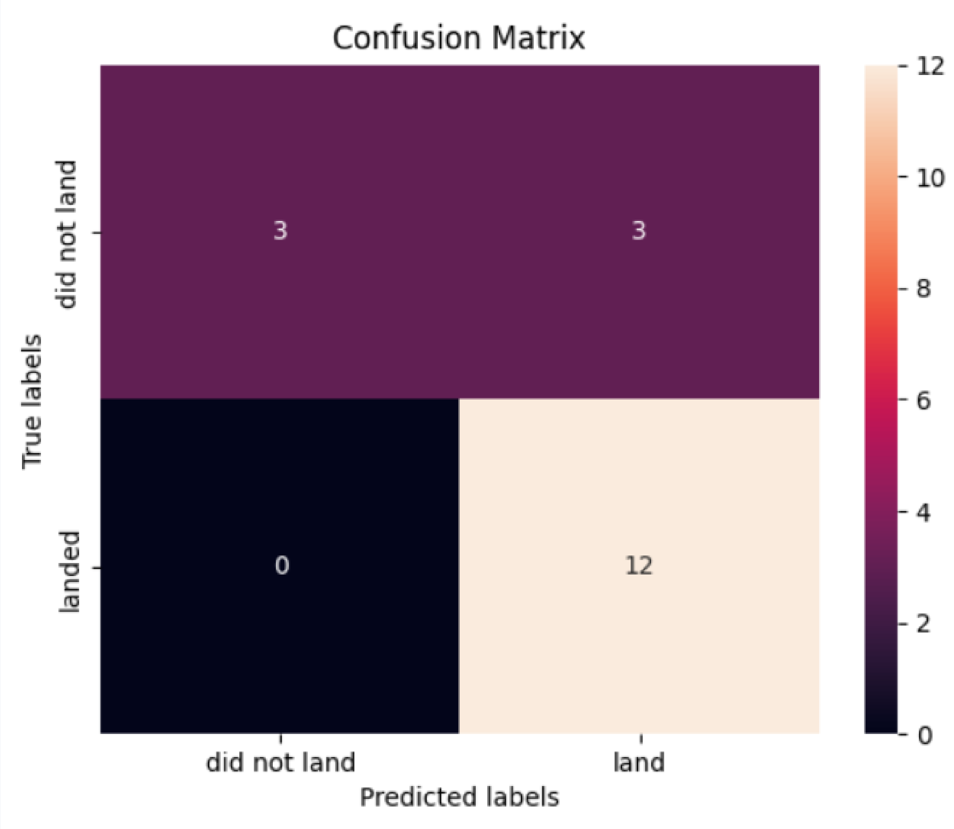
Classification Accuracy

- 3 out of 4 of the models were tied for the highest classification accuracy on the test set
- The worst performing model in this case was the decision tree classifier



Confusion Matrix

- 3 out of the 4 learning algorithms (KNN, SVM, Logistic Regression) produced the same confusion matrix



Conclusions

- Launches with lower payload mass have better success rates
- Success rates improved dramatically over time
- The ML classification algorithms performed similarly – the Decision Tree had the highest accuracy score on the training data, but this likely overfit, as it did the worst on the test data

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!