IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- We began by collecting the SpaceX launch data, which was available through an API and GET requests or from webscraping of a wiki page.
- We replaced missing data with the mean and created a class column for landing outcomes (0 = failure, 1 = success).
- We performed exploratory data analysis (EDA) with charts, graphs and SQL queries.
- We created maps using folium to see the proximity of launch sites to other features.
- We created classification models and tested their accuracy with the aid of train-test-split and the confusion matrix.

Introduction

- SpaceX is a company that provides rockets to other companies who wish to launch things into various orbits around earth.
- SpaceX saves money by landing the rockets first stage and then reusing it.
- SpaceX would like to develop an algorithm to predict if the landing of the rockets first stage will be successful.
- Questions that we will explore:
  - Are there any correlations between any variables and probability of success?
  - Do any launch sites have higher success rates?
  - Do the different booster types affect success rates?
  - Does payload mass affect success rate?
  - What are the most successful orbit types?
  - Can we develop a classification algorithm to predict success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API / Get Requests
  - Webscraping
- Perform data wrangling
  - Missing payload values replaced with payload mean
  - Created a landing outcome label for success outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary (cont)

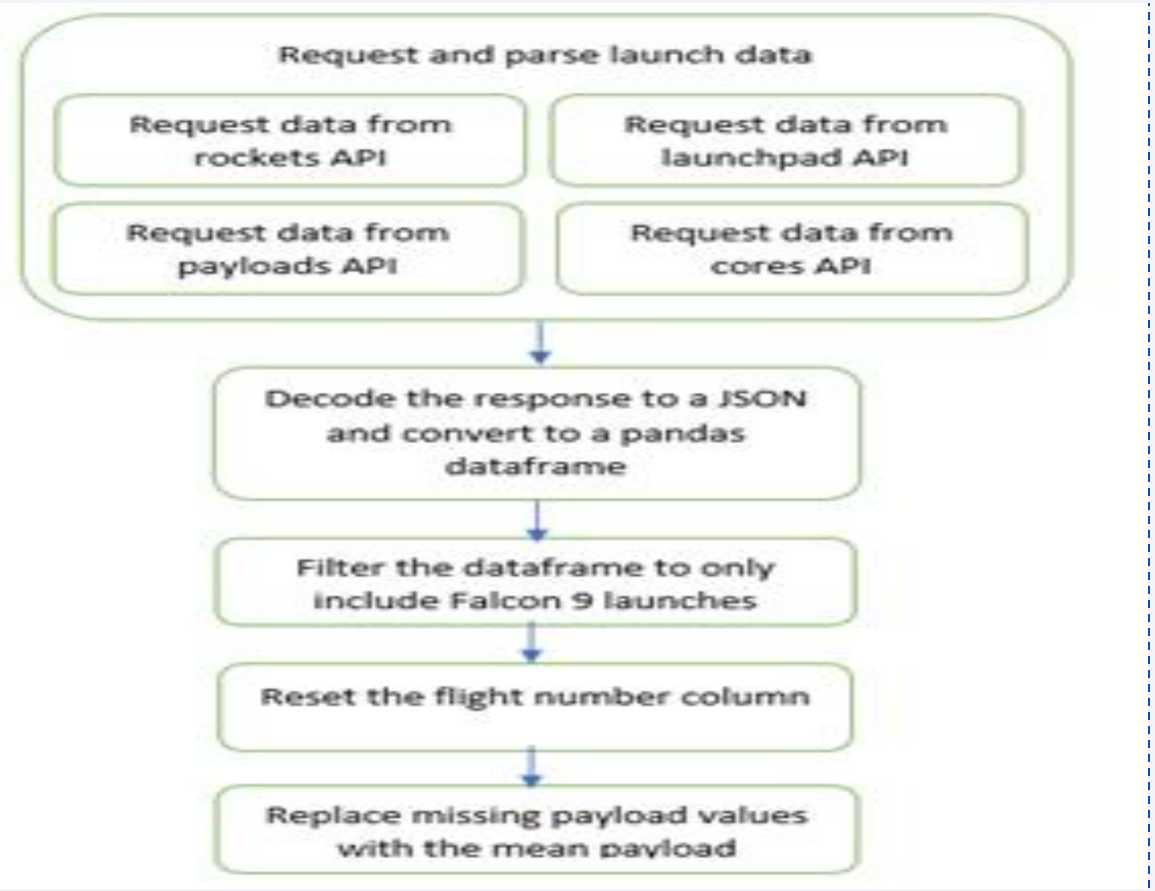
- Perform predictive analysis using classification models
  - Split the data with train-test-split
  - Use the training data to train each model
  - Predict the result for the test data and compare to the actual results
  - Calculate the accuracy of each model and explore the confusion matrix

Data Collection

- Data was collected with two methods
  - Using an API and GET Requests
  - Using webscraping and Beautiful Soup

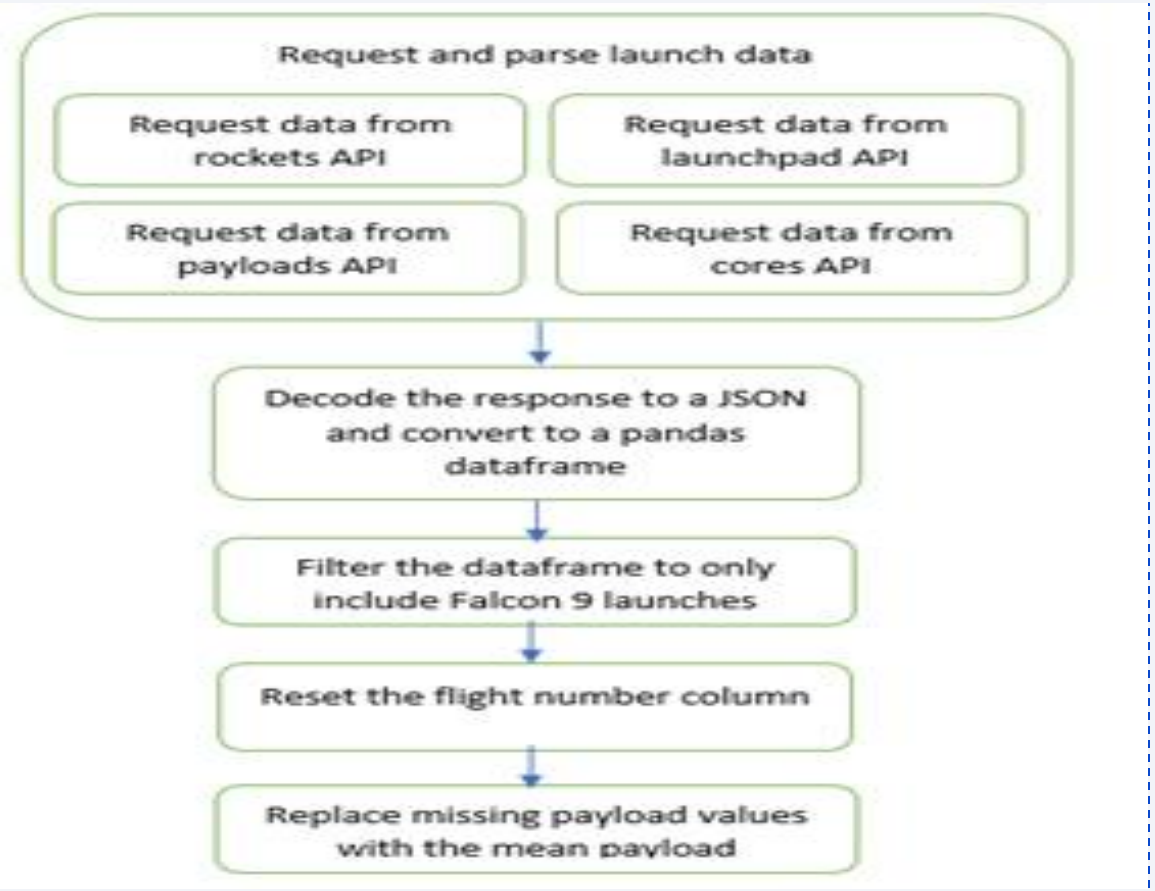
Data Collection – SpaceX API

- Request and parse the SpaceX launch data using a GET request
  - Request data from rockets API
  - Request data from launchpad API
  - Request data from payloads API
  - Request data from cores API
- Decode the response to a Json and then convert to a Pandas dataframe
- data = pd.json_normalize(response.json())



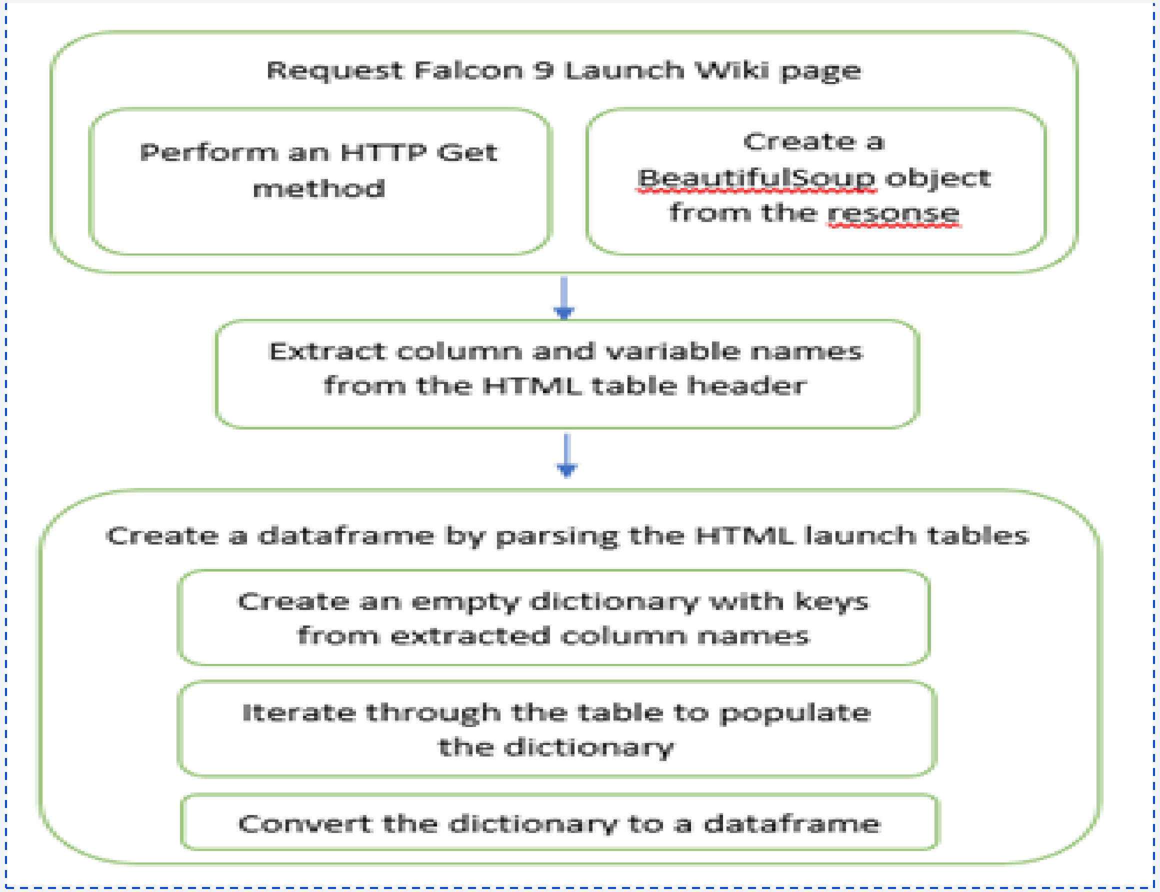
Data Collection – SpaceX API

- Filter the dataframe to only include Falcon 9 launches
  - data_falcon9 = df[df['BoosterVersion']!='Falcon 1']
- Reset the flight number column
- Replace the missing payload values with the payload mean
- For more details: https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/spacex-data-collection-api.ipynb



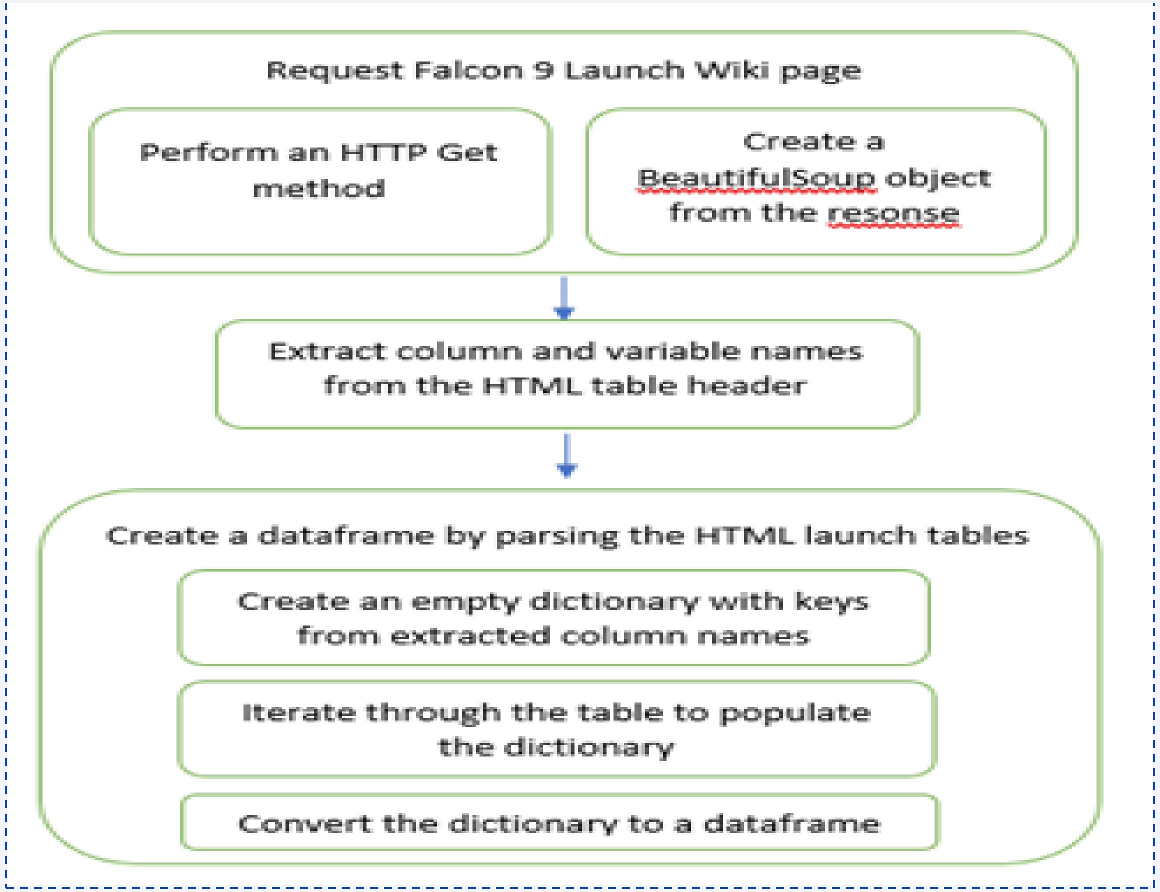
Data Collection - Scraping

- Request Falcon 9 Launch Wiki page from its URL
  - Perform an HTTTP Get method to request the page
  - Create a BeautifulSoup object from the response
- Extract column and variable names from the HTML table header



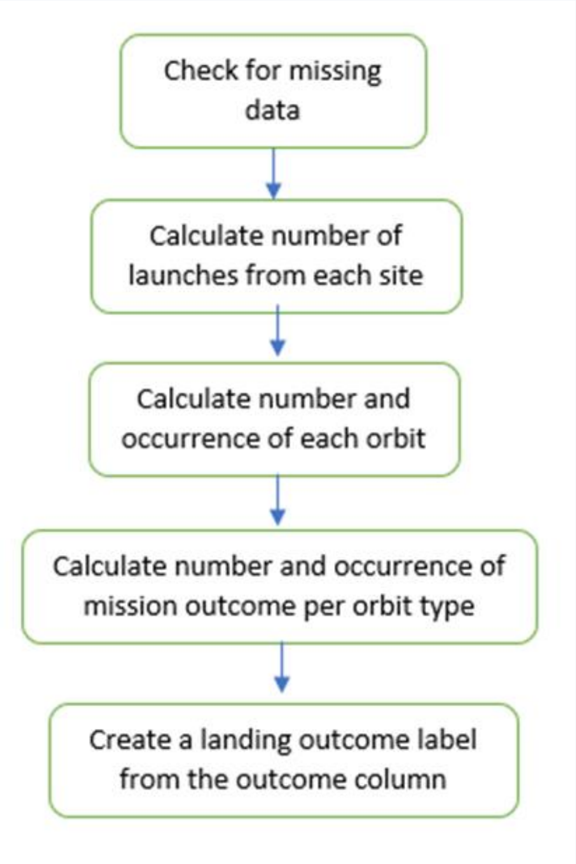
Data Collection - Scraping

- Create a dataframe by parsing the HTML launch tables
  - Create an empty dictionary with keys from the extracted column names
  - Iterate through the table to populate the dictionary
  - Convert the dictionary to a dataframe
- For more details: https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping-SpaceX.ipynb



Data Wrangling

- First, we verify that no data is missing
- Calculate the number of launches from each site
  - df['LaunchSite'].value_counts()
- Calculate the number and occurrence of each orbit
  - df['Orbit'].value_counts()
- Calculate the number and occurrence of mission outcome per orbit type
  - landing_outcomes = df['Outcome'].value_counts()
- Create a landing outcome label from the outcome column
  - 0 is a bad outcome, 1 is a successful landing
- We can see the success rate is 0.667
- For more details: https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/Data%20Wrangling%20Lab.ipynb



EDA with Data Visualization

For data visualization, we looked at the following plots:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Success Rate by Orbit Type
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Year vs. Success Rate

• From each plot, we are able to see relationships between the variables

EDA with Data Visualization

- The EDO Data Visualization notebook can be viewed at: https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Display the unique launch sites
- Display 5 records with launch site beginning with 'CCA'
- Calculate the total payload mass for NASA (CRS)
- Calculate the average payload mass for F9 v1.1
- Find the date of first successful landing on ground pad
- List the booster names with success in drone ship and payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried maximum payload mass. A subquery was used.

EDA with SQL

- List the records for failure in drone ship landings for the year 2015.
- List the rank of landing outcomes between 2010-06-04 and 2017-03-20, in descending order

- For full results of SQL queries:
  - https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The following objects were added to a folium map
  - Circles – Launch Sites
  - Markers – used as labels
  - Marker Clusters – used launches by failure or success
  - Added a marker and line from a launch site to the nearest coastline, railroad and highway
- The completed folium map allows is to investigate any relationship between launch sites and proximity to other locations, in addition to whether or not this might affect the success rate of landings.
- To view the full notebook with folium https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/Interactive%20Folium%20Map.ipynb

Build a Dashboard with Plotly Dash

- We built a dashboard that displays a pie-chart and a scatter plot.
- The pie chart can show percentages of success for all sites, or can show percentage of success and failure at each individual site.
- The scatter plot shows payload vs success for either all sites or for an individual site.
- Is a drop down that allows selection of sites and a slider bar that sets a range of payloads to display for the scatterplot.
- The full notebook may be viewed on github: https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/spacex_dash_app.py

Results

- The SpaceX program has been increasing its success rate as the year increases.
- The launch site are in isolated locations.
- Rockets with heavier payloads have been less successful in landings.
- Some launch sites have higher success rates than others.
- Some orbit types have very successful landing rates.

Predictive Analysis (Classification)

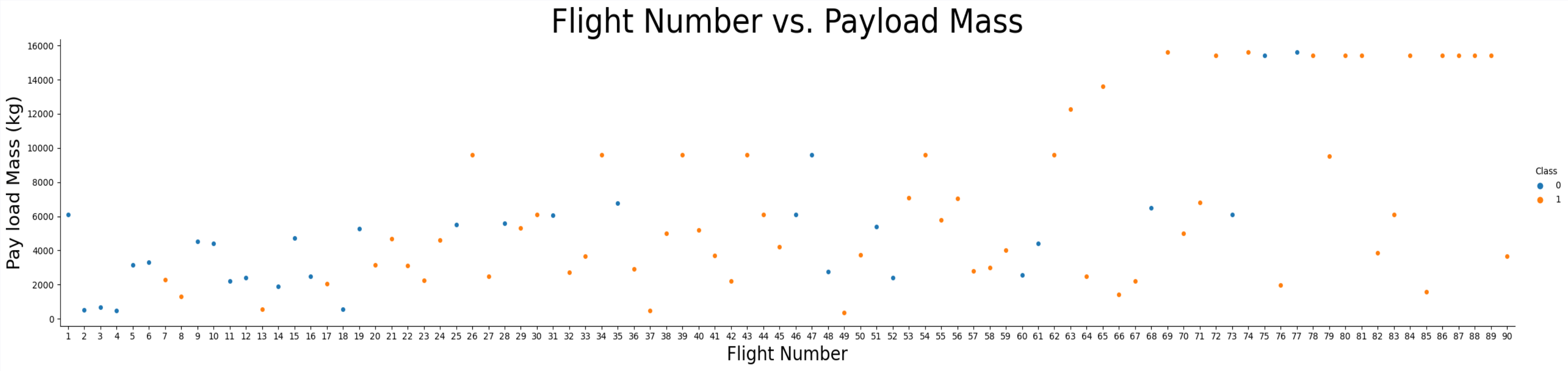
- Classification was performed with 4 models
  - Logistic Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbors
- The complete notebook may be viewed on github https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/Predictive%20Analysis.ipynb

Section 2

Insights drawn from EDA

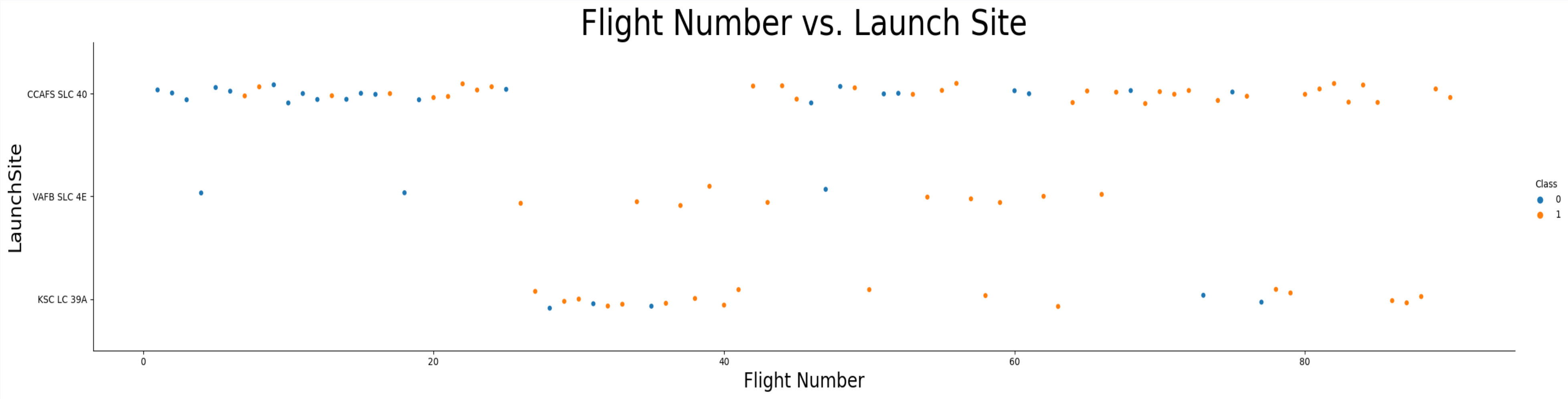
Flight Number vs. Payload Mass

- Flight number vs. Payload Mass – Overlayed with Class
  - This plot shows us that as flight number increases, the success rate increases.
  - We also see that higher payload mass decreases the success rate.



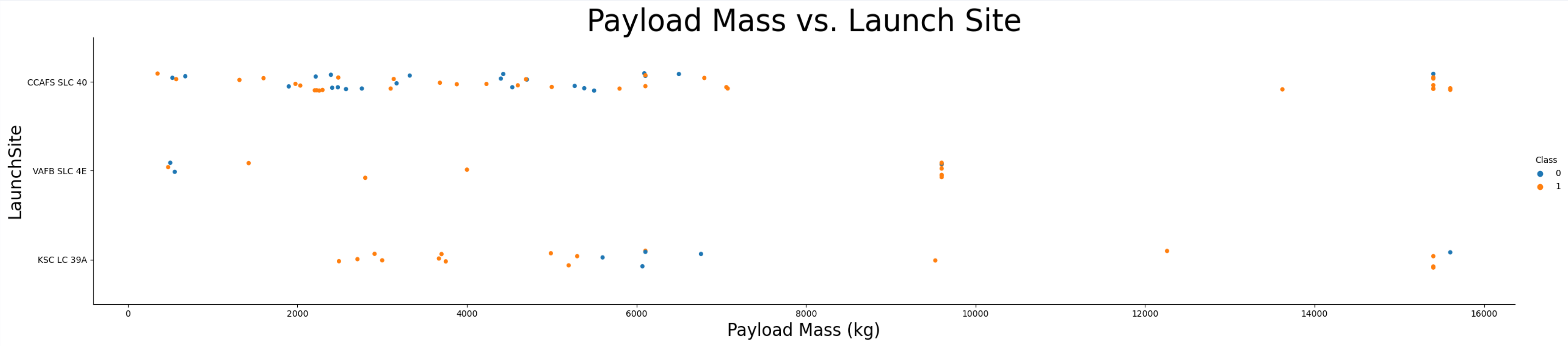
Flight Number vs. Launch Site

- Flight number vs. Launch Site – Overlayed with Class
  - This plot shows us that as flight number increases, more launch sites are being used.
  - We see that some of the launch sites have a higher success rate.
  - We can see that the launch site CCAFS SLC 40 has had the most launches.



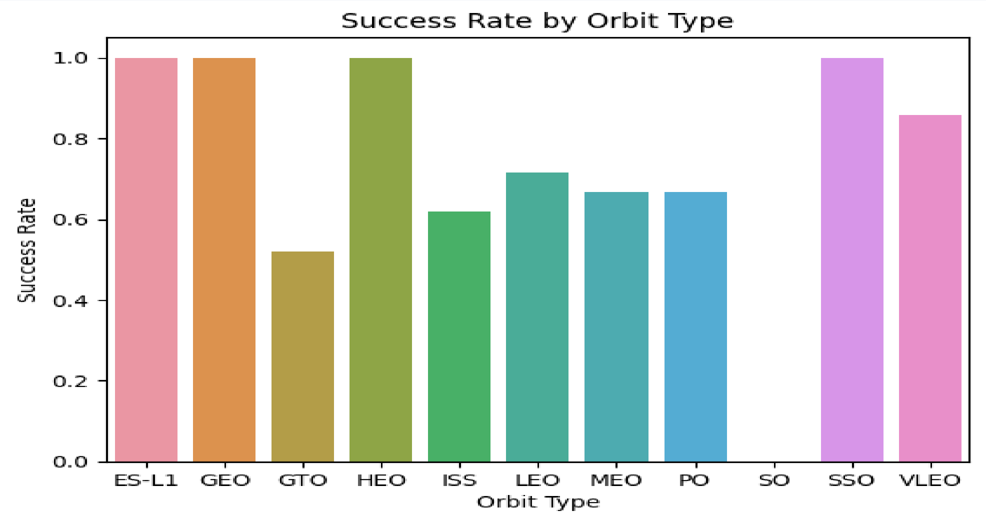
Payload vs. Launch Site

- Payload Mass vs. Launch Site – Overlayed with Class
  - This plot shows us that the heaviest payloads are not launched from VAFB SLC 4E.
  - We see that the launch site CCAFS SLC 40 is used for all ranges of payload mass.
  - We can see that the launch site CCAFS SLC 40 has had the most launches.



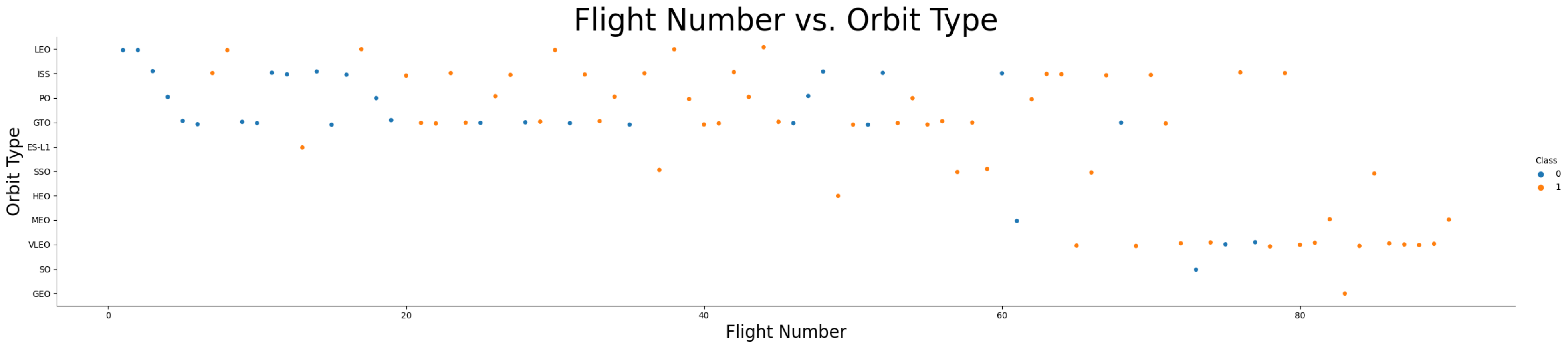
Success Rate vs. Orbit Type

- Success Rate by Orbit Type
  - Several orbit types have a successful landing rate of 1.0.
  - Some of the orbit type have a low success rate.
  - The orbit type SO has not had a successful landing.



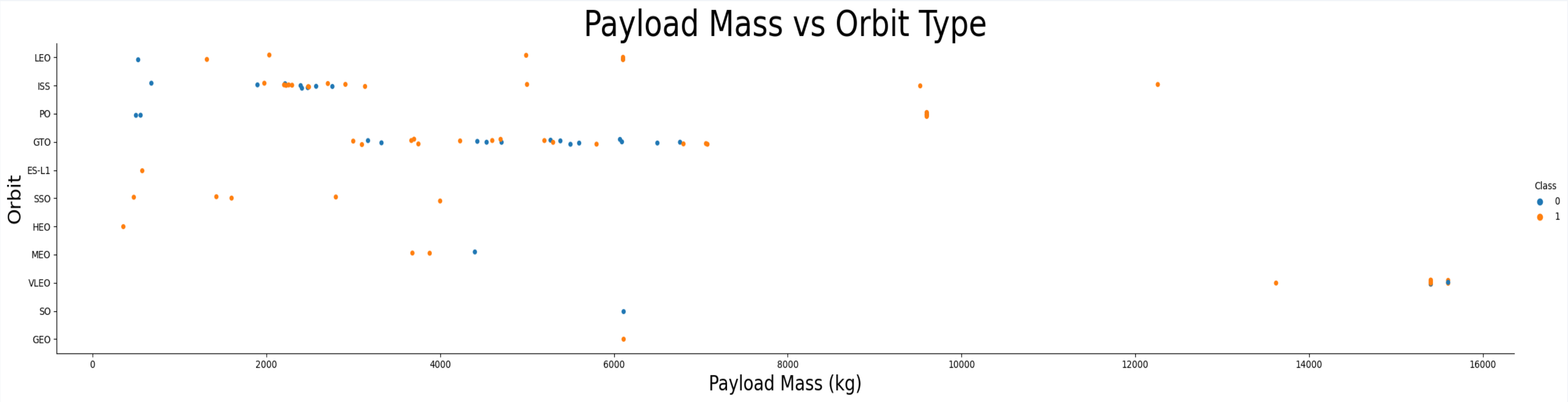
Flight Number vs. Orbit Type

- Flight Number vs. Orbit Type – Overlayed with Class
  - For LEO, later flights were more successful.
  - Many of the orbit types were not attained early on in the program.



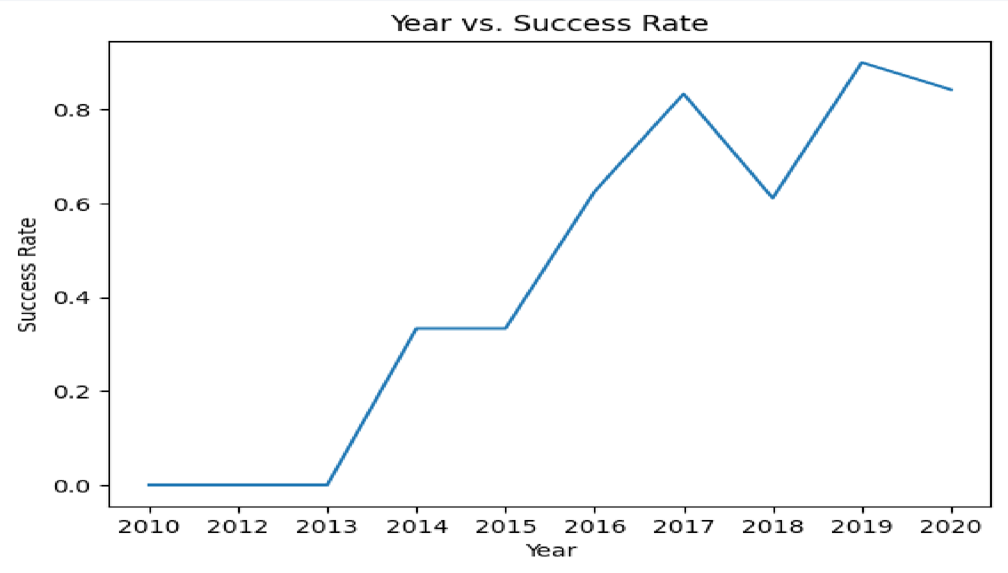
Payload vs. Orbit Type

- Payload Mass vs. Orbit Type – Overlayed with Class
  - For PO, LEO and ISS, higher payload launches have a higher success rate.
  - The success rate for GTO seems random when related to payload mass.



Launch Success Yearly Trend

- Year vs. Success Rate
  - Overall, there has been an increase in the success rate from 2013 to 2020.
  - There was a slight decrease in success rates in 2018 and 2020.



All Launch Site Names

- The following are the unique launch site names

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- There a four unique launch sites.

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

Launch Site Names Begin with 'CCA'

- Below are five records where the launch site name begins with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
2010-08-12	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 (para
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No at
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No at
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No at

```
%sql select * from SPACEXTABLE where "Launch_site" like 'CCA%' limit 5
```

Total Payload Mass

- We were interested in the total payload mass carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = "NASA (CRS)"

Average Payload Mass by F9 v1.1

- We were interested in the average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = "F9 v1.1"

First Successful Ground Landing Date

- It was important to know the date of the first successful landing outcome on a ground pad.

min("Date")
2015-12-22

%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below is the list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = "Success (drone ship)"
and "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
```

Total Number of Successful and Failure Mission Outcomes

- It was important to know the total number of successful and failure mission outcomes

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

%sql select "Mission_Outcome", count(*) as total from SPACEXTABLE group by "Mission_Outcome" order by "Mission_Outcome"

Boosters Carried Maximum Payload

- To the right is the list of the names of the boosters which have carried the maximum payload mass

%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)

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

- Below is the list of the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

```
%sql select substr(Date,6,2) as Month, "Booster_Version", "Launch_Site" from
SPACEXTABLE where substr(Date,1,4)='2015' and "Landing_Outcome" = "Failure (drone ship)"
```

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

- To the right is the ranking of 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

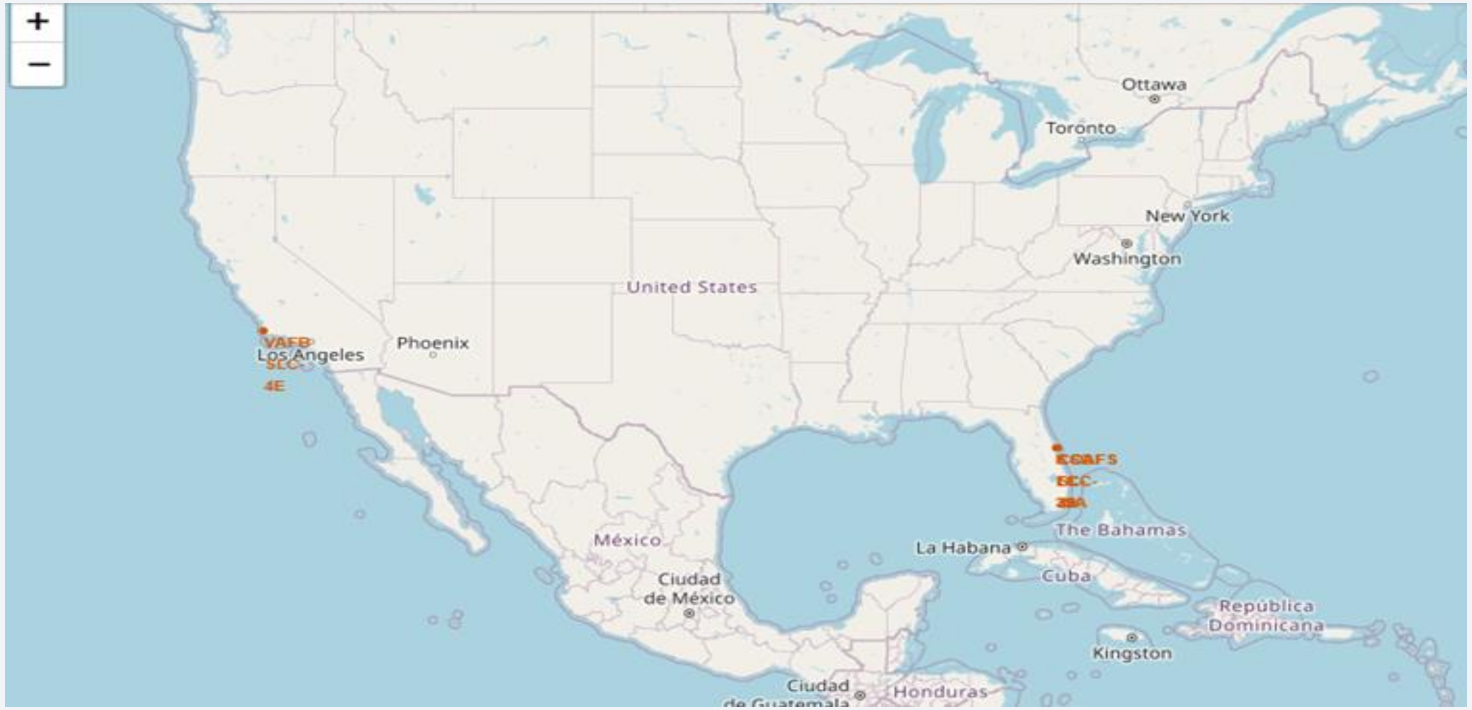
%sql select "Landing_Outcome", count(*) as total from SPACEXTABLE group by "Landing_Outcome" order by total desc

Landing_Outcome	total
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Section 3

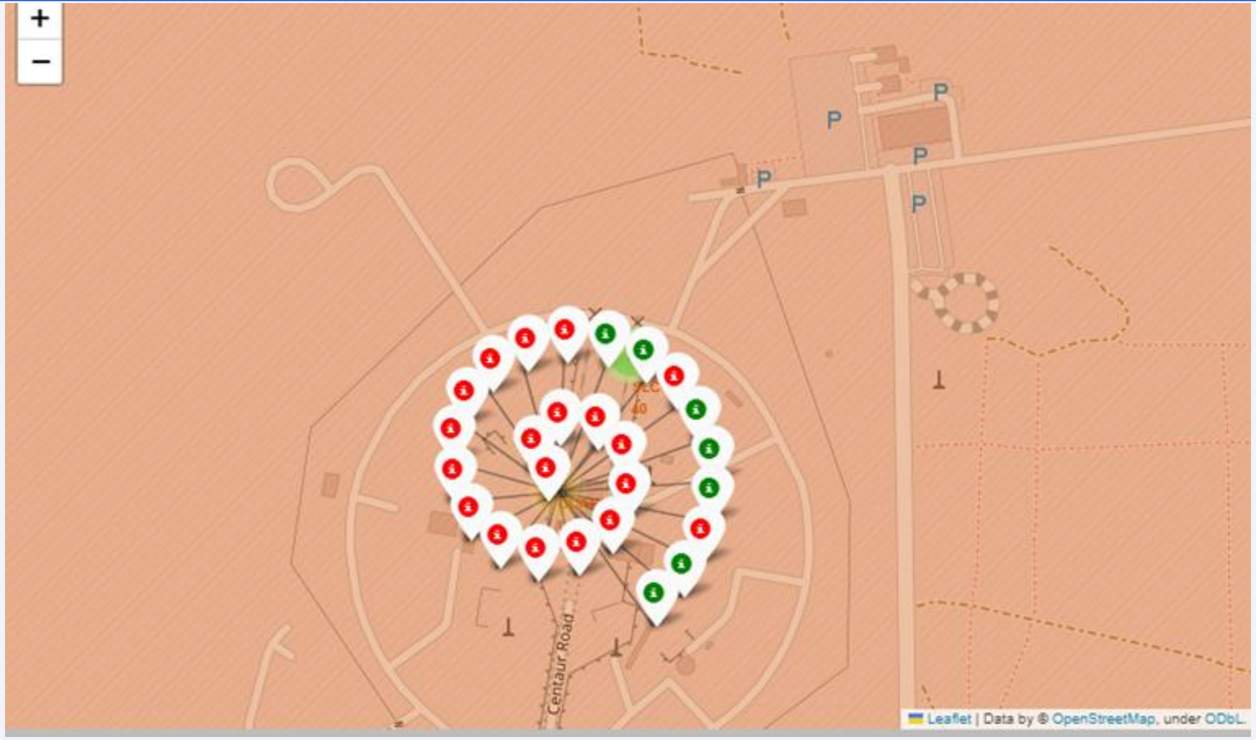
Launch Sites Proximities Analysis

Locations of All Launch Sites



- From this map, we can see that each launch site is located on either the east or west coast of the United States. Specifically, they are in California and Florida.

Clusters with Color-Coded Outcomes



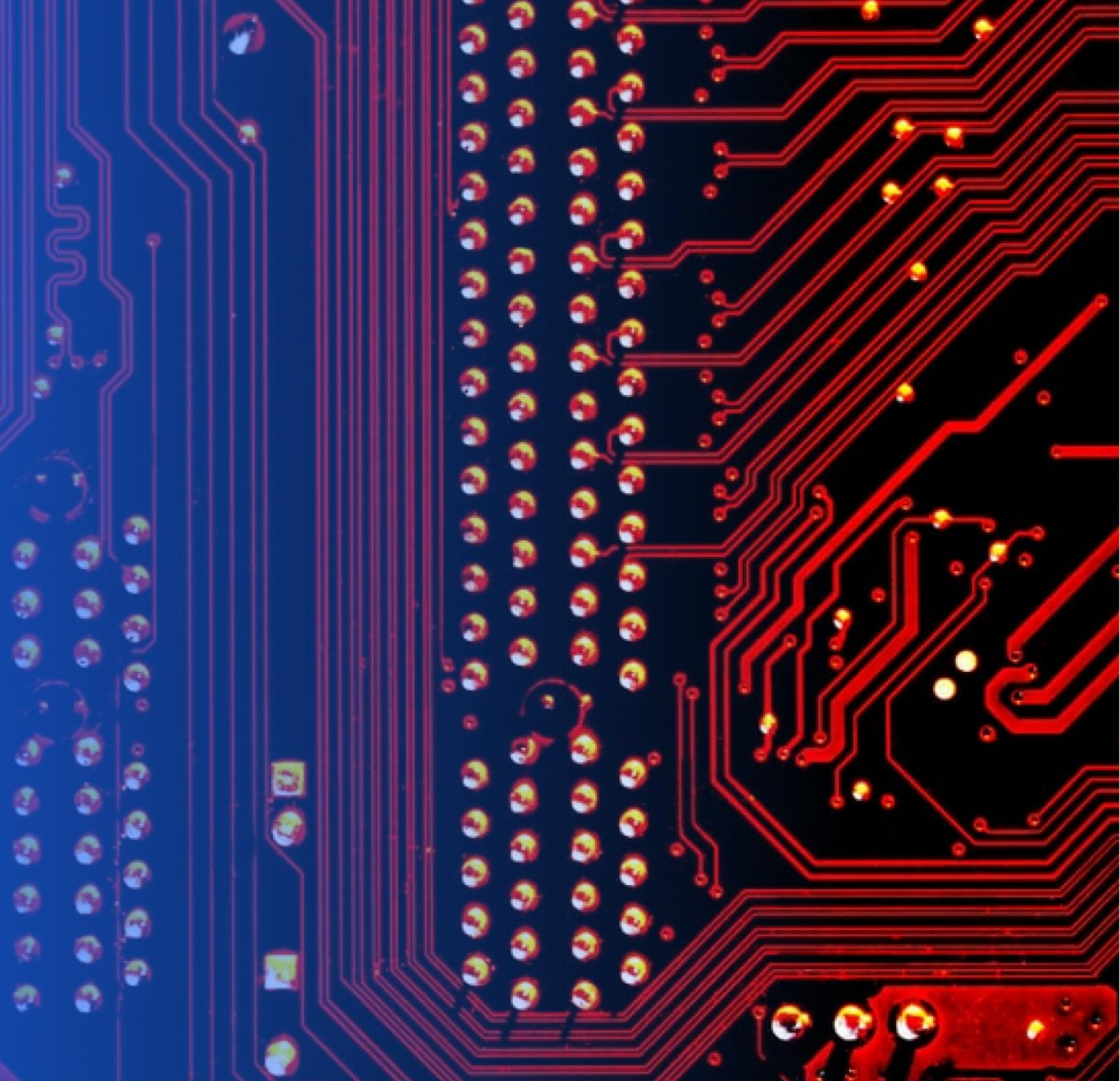
- This zoomed-in map shows successful (green) and unsuccessful (red) landing outcomes for rockets launched from CCAFS LC-40.

Proximity to Coastlines and Highways

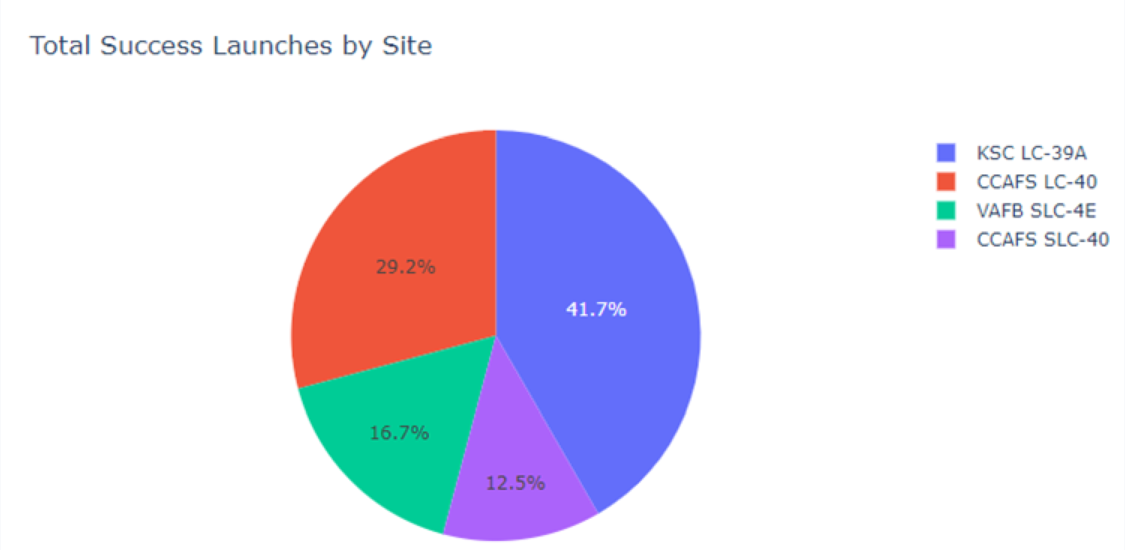
- Launch site CCAFS SLC-40 is 1.41 km from the nearest coast and 0.58 km to the nearest highway.
- We can see that the launch sites are located in close proximity to coastlines, but there are not many roads or railroads nearby.

Section 4

Build a Dashboard with Plotly Dash

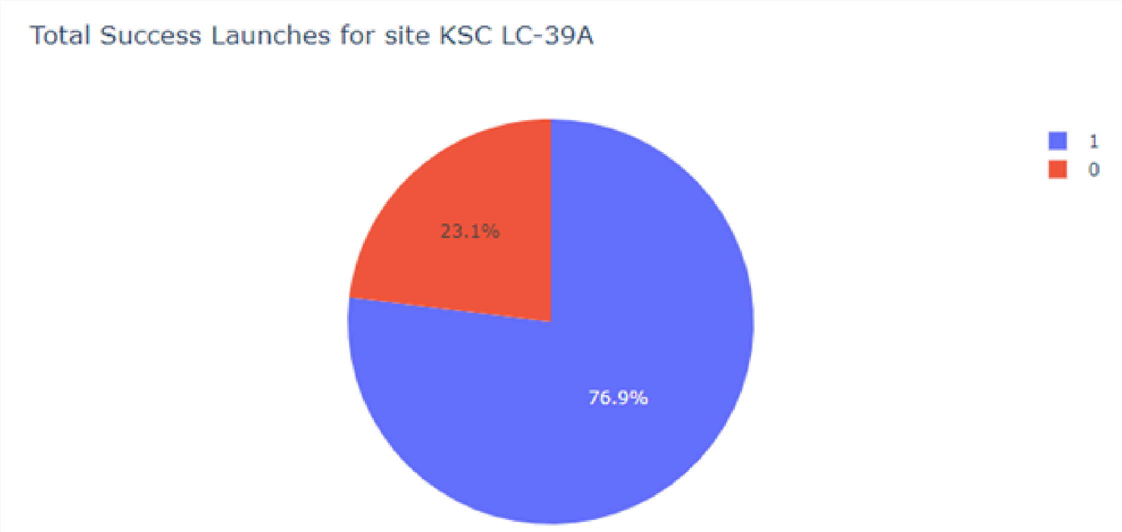


Successful Launches by Site



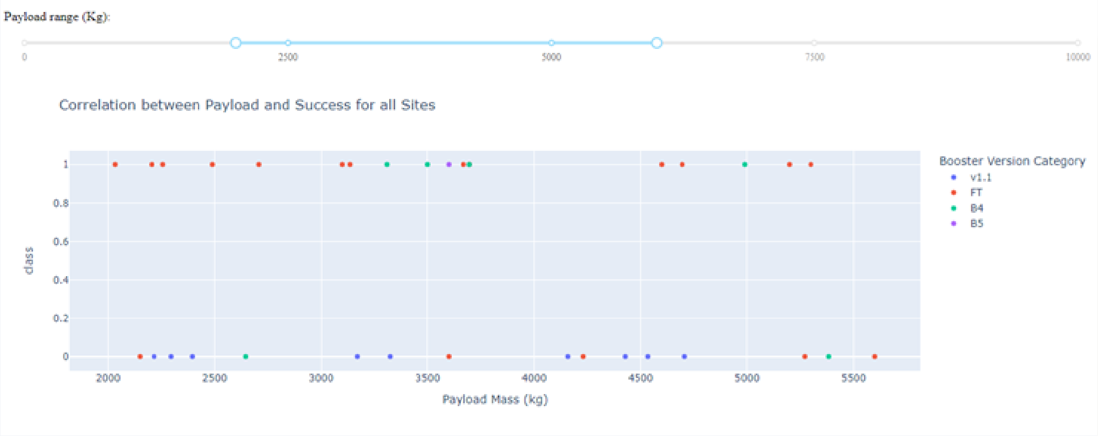
- The pie chart shows us that some sites have had more successful launches than others.
- KSC LC-39A has had the largest number of successful launches, and CCAFS SLC-40 has had the lowest number of successful launches.

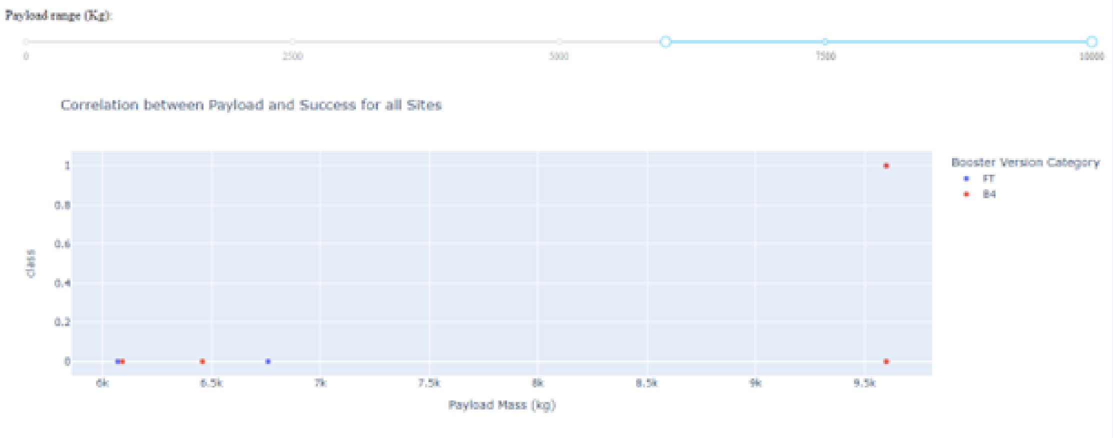
Success Ratio



- Launch site KSC LC-39A has had the highest ratio of success (76.9%).

Using the Range Slider





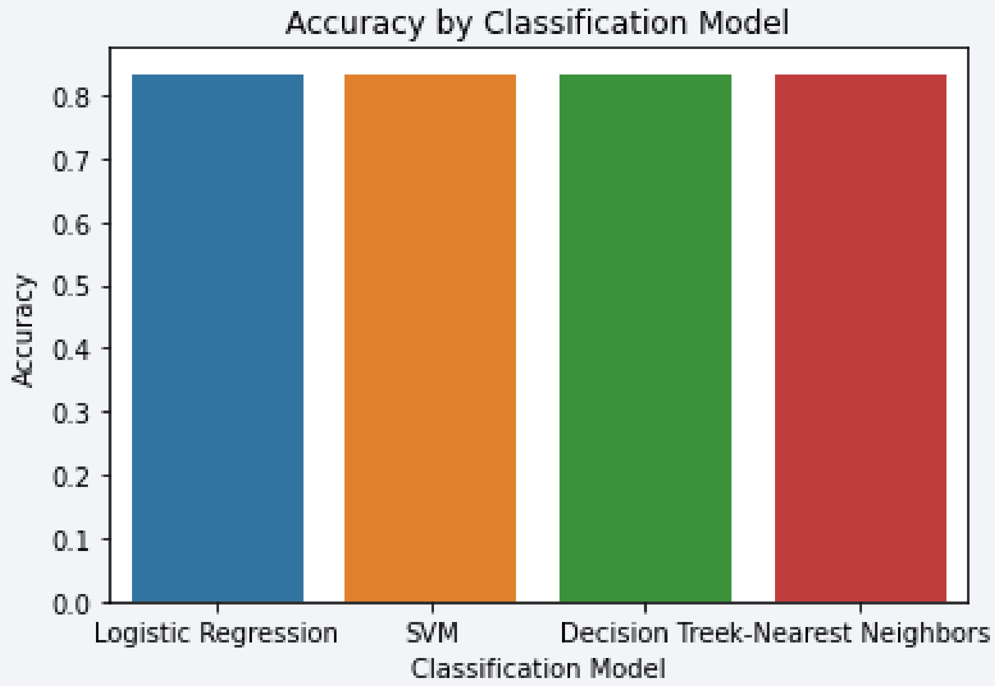
- For payloads ranging between 2000 kg and 6000 kg, booster version FT has a high success rate.
- For payloads over 6000 kg, there has only been one successful landing outcome.

Section 5

Predictive Analysis (Classification)

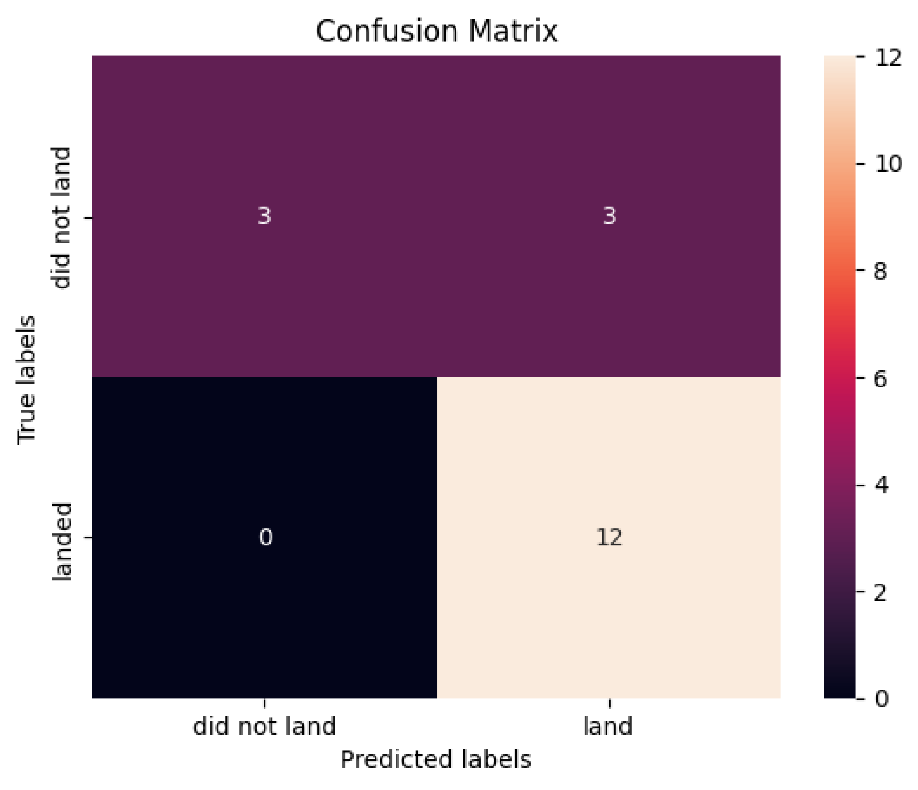
Classification Accuracy

- Classification was performed with 4 models
    - Logistic Regression
    - SVM
    - Decision Tree
    - K-Nearest Neighbors
- Each model had the same accuracy (0.8333)



Confusion Matrix

- The classification models are very accurate when predicting rockets that actually landed successfully.
- The classification models are only 50% when predicting outcomes for rockets that did not land successfully.



Conclusions

- The SpaceX program has been increasing its success rate as the year increases.
- The launch site are in isolated locations.
- Rockets with heavier payloads have been less successful in landings.
- Some launch sites have higher success rates than others.
- Some orbit types have very successful landing rates.

Appendix

All python notebooks can be viewed on github

https://github.com/craiggore630/Applied_Data_Science_Capstone/blob/main/spacex-data-collection-api.ipynb

Thank you!