IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paige Knittel
May 3, 2023

Outline of the Capstone Project

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

Executive Summary

- **Part 1:** collect (read) and clean data from the wiki page
- **Part 2:** sort and transform information into a workable database using web scraping
- **Part 3:** explore data and begin to notice patterns
- **Part 4:** execute SQL queries and answer questions to further learn about the dataset
- **Part 5:** use visitations to further gain information
- **Part 6:** use folium to create map specific visuals
- **Part 6 - dash:** create a dashboard to visualize data and observe trends
- **Part 7:** use machine learning to make informed calculations from data

Through this project, we aim to discover patterns that make the first stage of SpaceX's rockets reusable. As this is the most expensive stage of a launch, reusability causes them to be so cost effective.

It was found that improved technology over time makes predictions more accurately, and certain orbits have very high or low recovery rates.

Introduction

- SpaceX is able to often reuse the first stage of a rocket launch
- This step is the most expensive, so being able to reuse this step can decrease the cost
- Before we dive into our search, we want to sift through what kind of data we have and stumble upon some patterns,
- Let’s take a look!

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from a SpaceX REST API that records launch data
- Perform data wrangling
  - We clean, sort, edit, and reformat data to make it the most useful for future visualizations and calculations
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We used predictive analysis to see if the first step would be reusable; we use the train_test_split function, a few different machine learning methods, and conduct error analysis

Data Collection

- First, a few url's were collected from SpaceX and we convert them to a .csv file.
- Then, we used web scraping to collect information of certain types to be grouped accordingly such as whether or not a launch was successful.
- The useful data was put into a data frame to be conveniently manipulated throughout the rest of analysis.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Part 1 code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%201.ipynb

How the information began

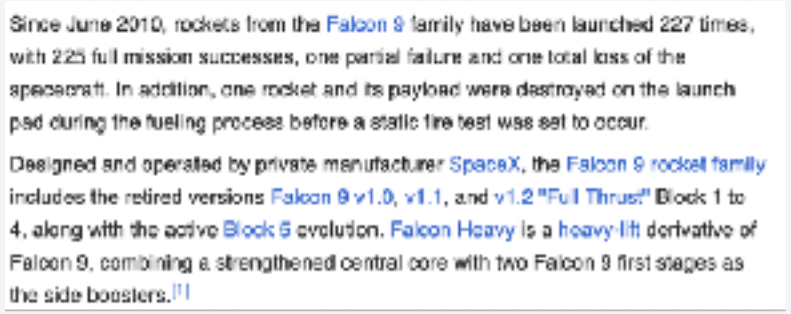
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	F
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	

How the information was at the end of this part

Data Collection - Scraping

- Data from a wikipedia page was read and processed.

- Part 2 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%202.ipynb



Since June 2010, rockets from the Falcon 9 family have been launched 227 times, with 225 full mission successes, one partial failure and one total loss of the spacecraft. In addition, one rocket and its payload were destroyed on the launch pad during the fueling process before a static fire test was set to occur.

Designed and operated by private manufacturer SpaceX, the Falcon 9 rocket family includes the retired versions Falcon 9 v1.0, v1.1, and v1.2 "Full Thrust" Block 1 to 4, along with the active Block 5 evolution. Falcon Heavy is a heavy-lift derivative of Falcon 9, combining a strengthened central core with two Falcon 9 first stages as the side boosters.[1]

How the information began

```
# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
```

A code snapshot

Data Wrangling

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts().to_frame()
```

	Orbit
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
HEO	1
ES-L1	1
GEO	1
SO	1

Code Snippet

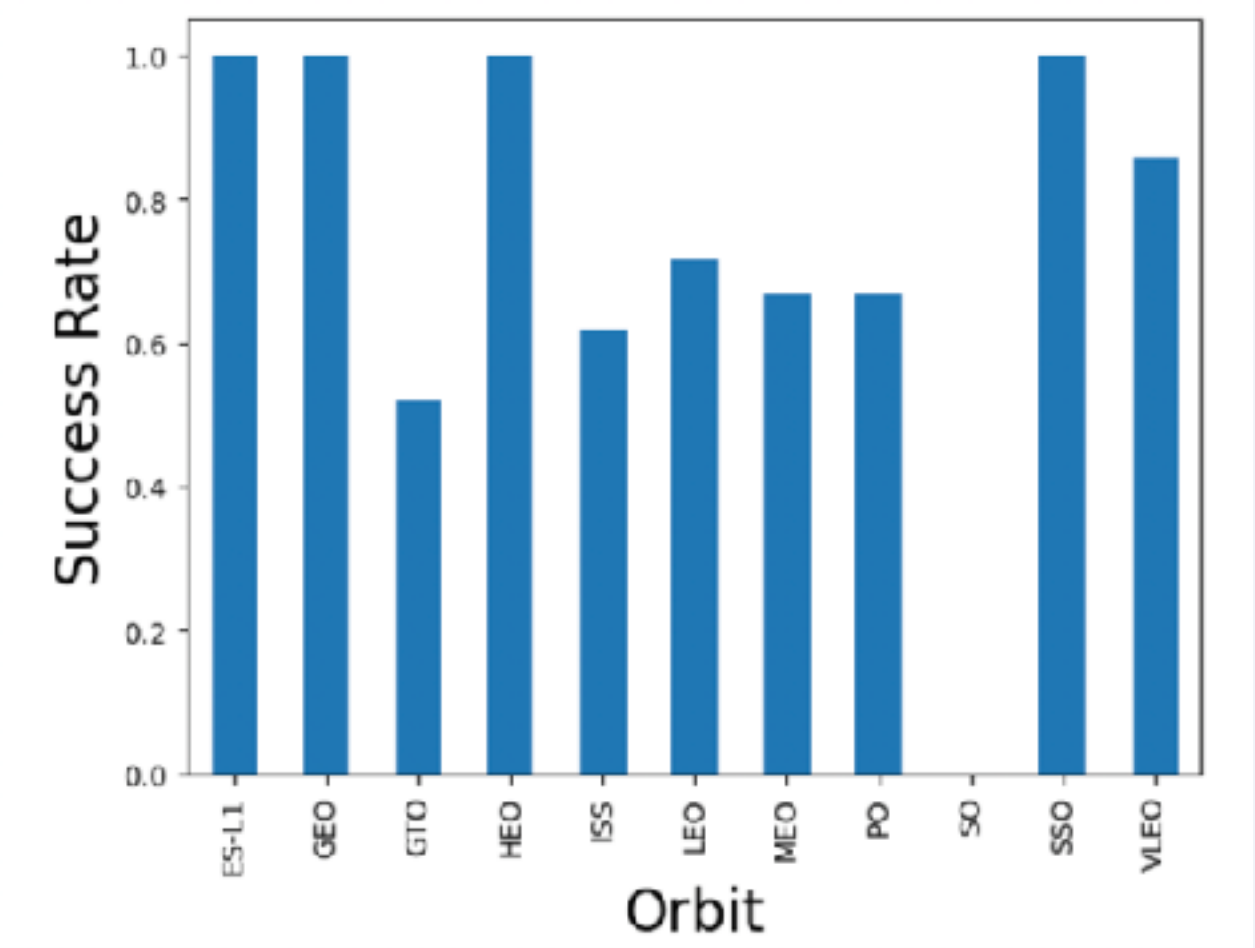
- To begin processing data, we checked for any missing information
- We checked if all data types were what we expected them to be
- Information began to be explored to see if there is anything off balance or unusual
- Part 3 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%203.ipynb

EDA with SQL

- Some questions were posed to encourage exploration of patterns using magic SQL in Python
- The rest of this project was conducted in Python
- Part 4 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%204.ipynb

EDA with Data Visualization

- We did a series of scatter and bar plots to hopefully stumble upon some patterns
- It shows that in time, launches become increasingly successful
- Part 5 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%205.ipynb



Build an Interactive Map with Folium

- Using Folium map, launch sites with circles and markers were placed
- Lines were used to show patterns to nearby land features such as shores, roads, and cities where we noticed launch sites tend to be near the shore and a bit away from cities.
- These visuals allowed us to easily spot these patterns
- Part 6 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%206.ipynb

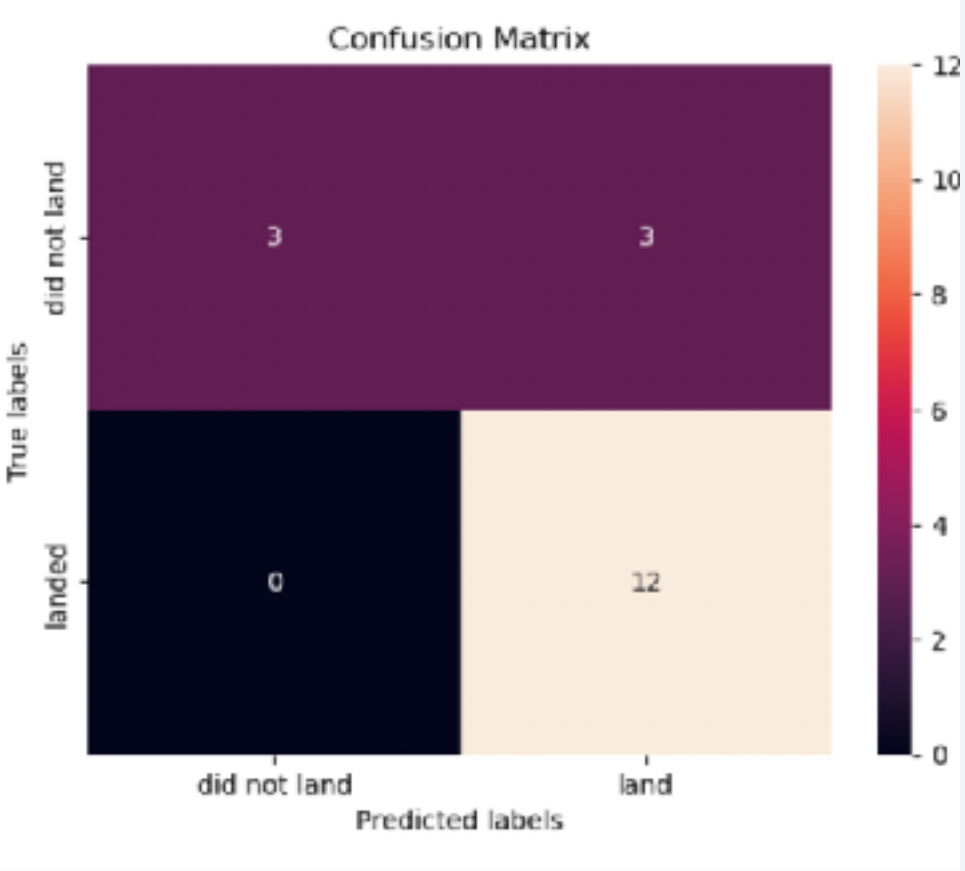
Build a Dashboard with Plotly Dash

- The dashboard created allowed for an interactive visualization of some of the most useful data we have
- We could see all launch sites together or take a look at each individually
- A pie chart and scatter plot were used
- Part 6 Dashboard: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%206-Dash.py

Predictive Analysis (Classification)

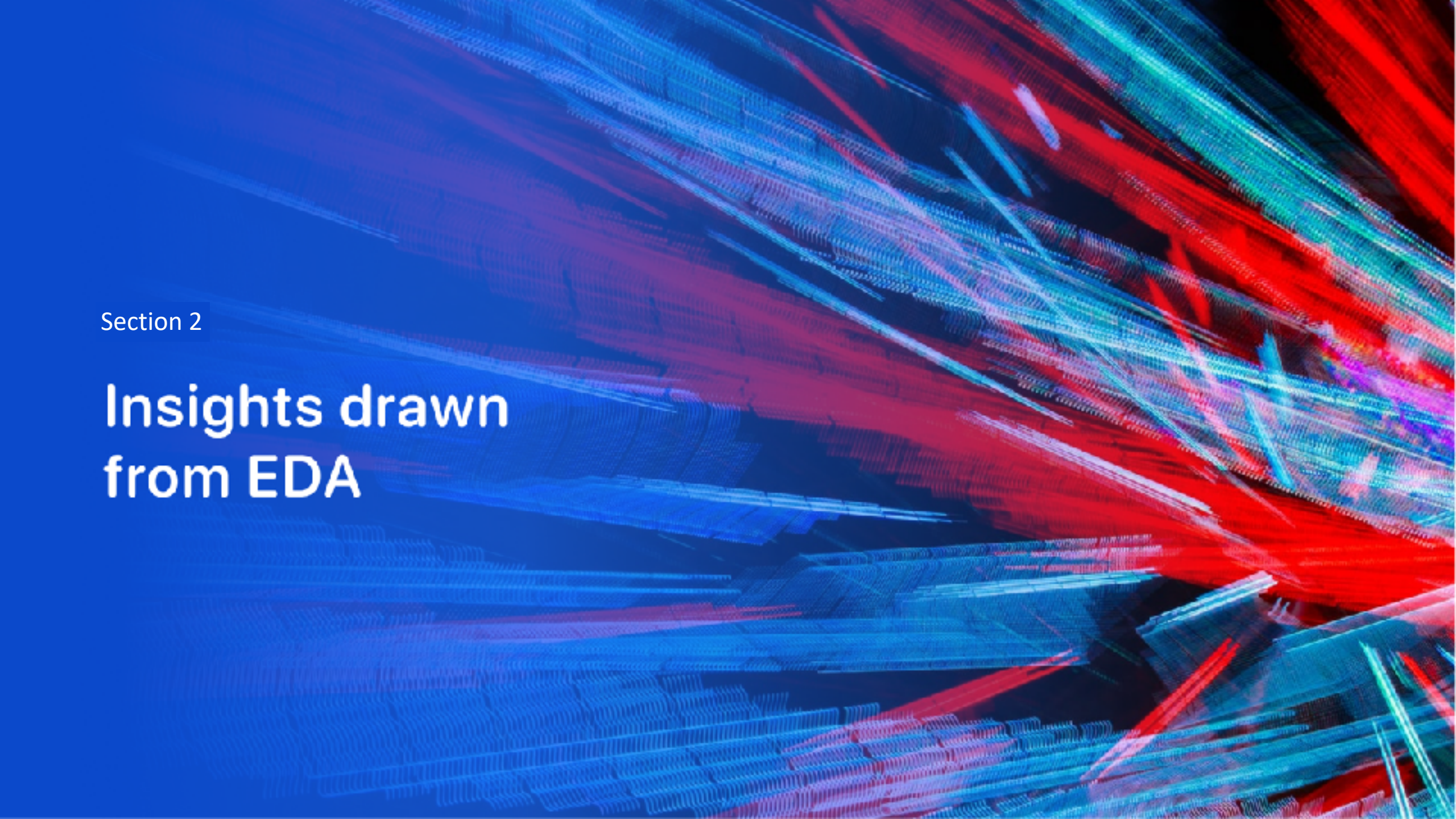
- Finally, we used important machine learning models to make predications about successful launches using our data.
- We conducted error analysis to see that most methods had around the same error.
- Part 7 Code: https://github.com/plknit00/Final-Project-IBM/blob/main/Part%207.ipynb

Results

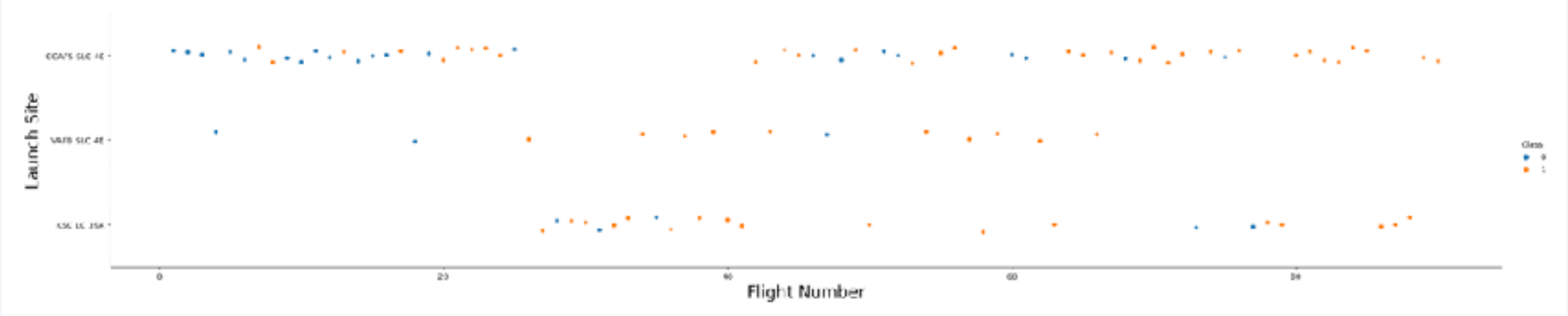


A confusion matrix for SVM data

- We used the following methods listed with their accuracy on test data
  - Logistic Regression, 0.833
  - SVM - Support Vector Machines, 0.833
  - Decision Tree, 0.778
  - k-Nearest Neighbors, 0.833

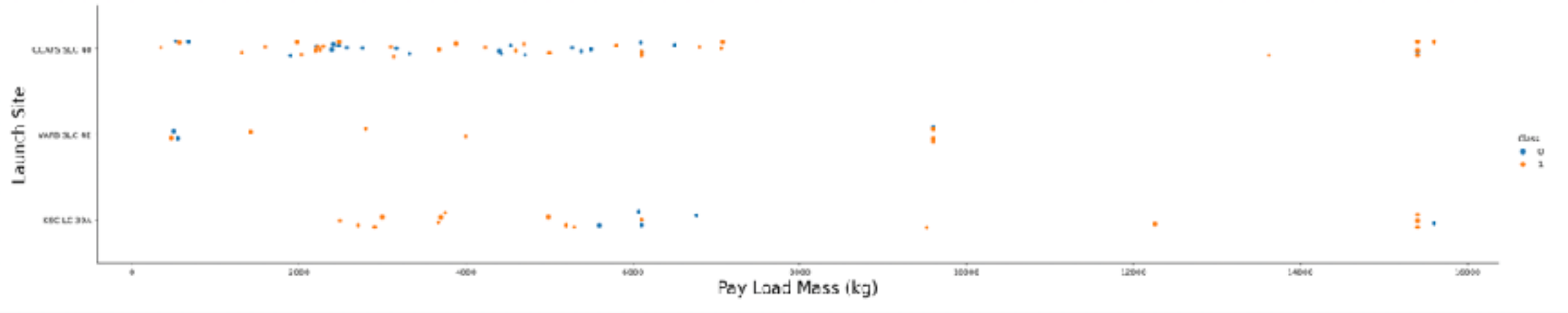

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



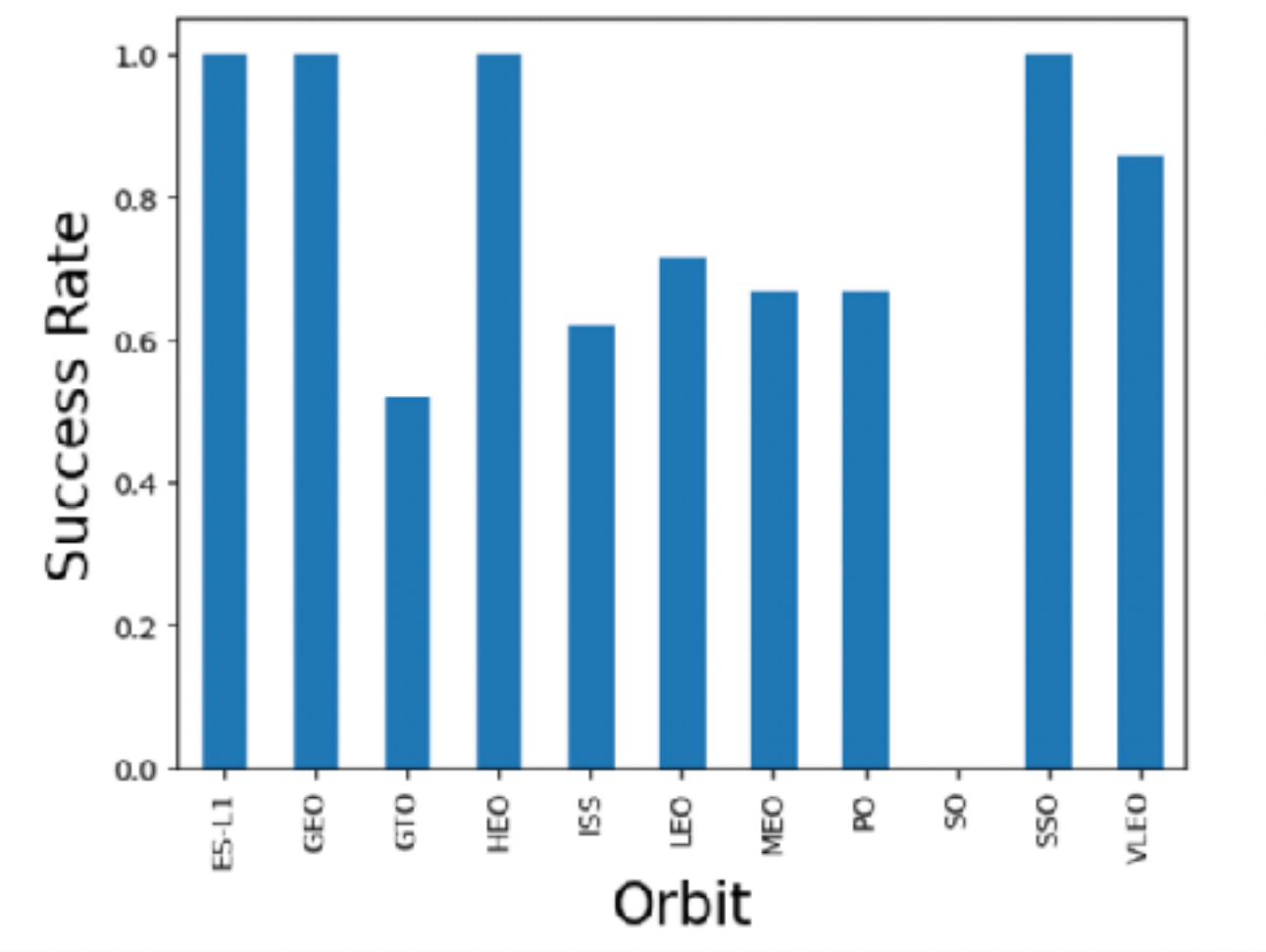
This chart depicts where flights have launched from over time as the larger the flight number the more recent the flight is. The blue dots were not successful recoveries of the first step but the orange were successful. We can notice a pattern of more success the later time gets.

Payload vs. Launch Site



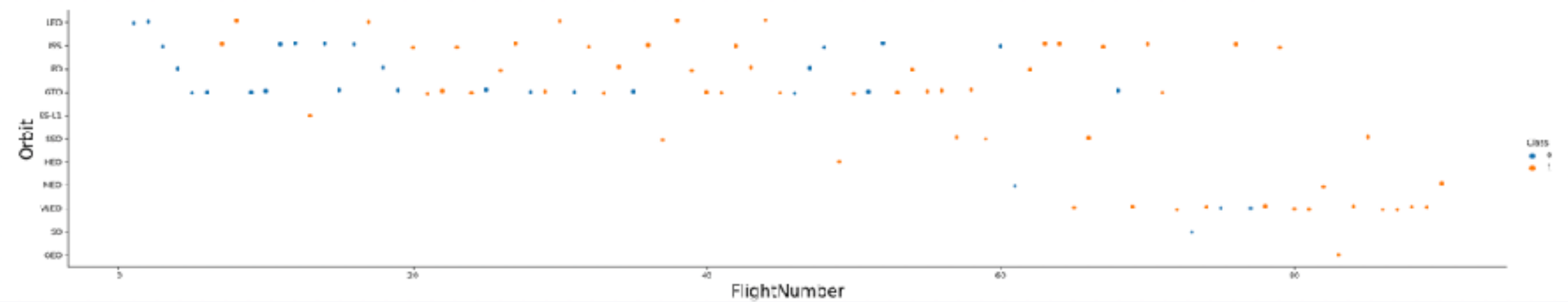
Here we have the mass that the rockets can support at various launch sites. The blue dots were not successful recoveries of the first step but the orange were successful. It appears those that can carry a lot tend to be successful.

Success Rate vs. Orbit Type



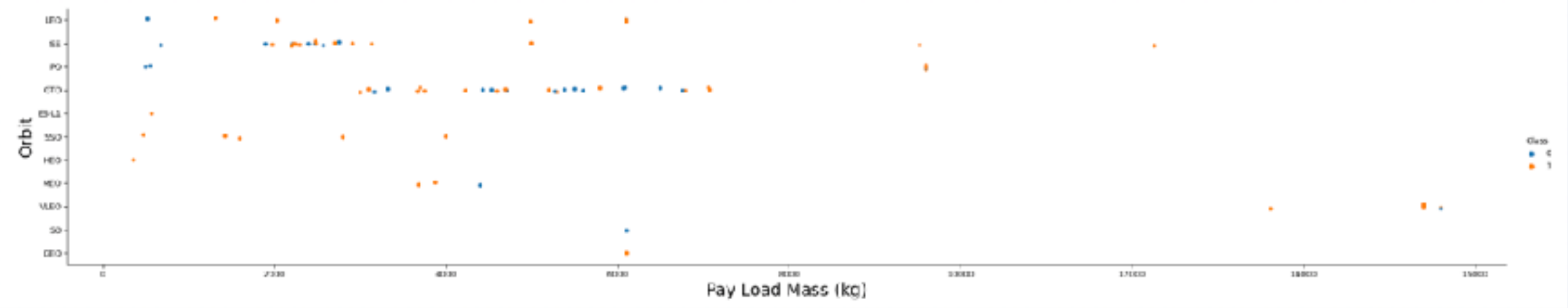
This chart displays that there are several orbits with 100% success rate, one with 0% success rate, and a few in the middle.

Flight Number vs. Orbit Type



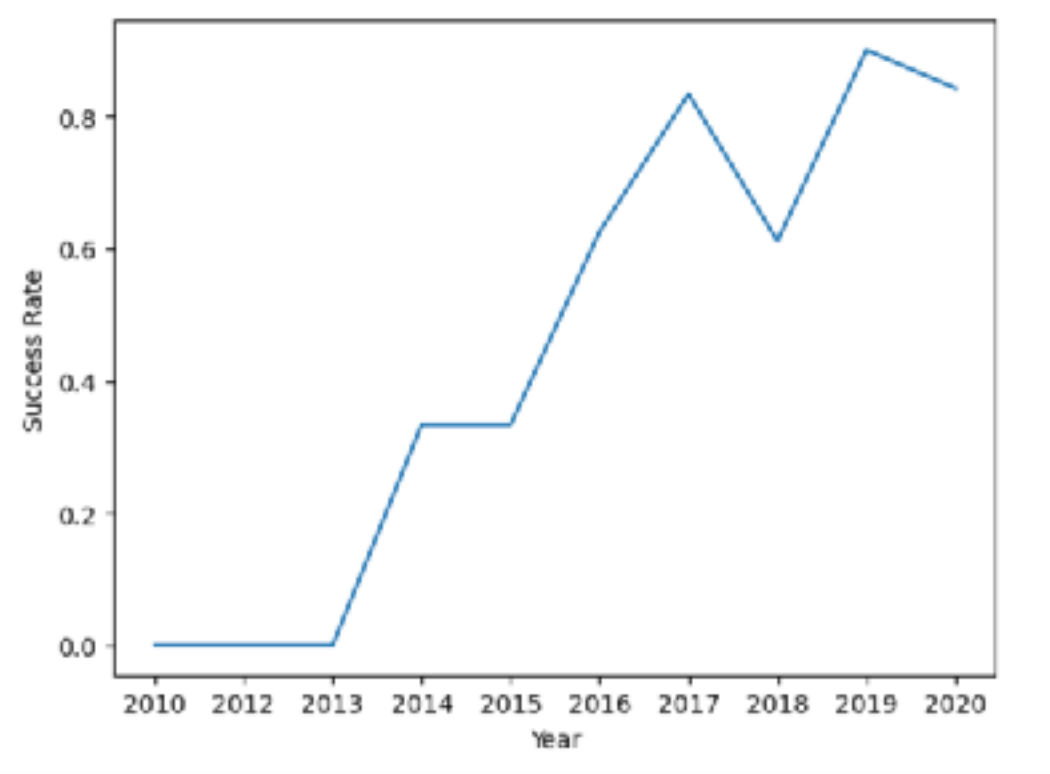
This chart depicts different orbits rockets have taken over time as the larger the flight number the more recent the flight is. The blue dots were not successful recoveries of the first step but the orange were successful. We can notice a pattern of more success the later time gets and there is a shift into using different orbits in more recent years.

Payload vs. Orbit Type



Here we have the mass that the rockets can support at various launch sites. The blue dots were not successful recoveries of the first step but the orange were successful. It appears those that can carry a lot tend to be successful. It appears the orbits further towards the horizontal axis tend to have a higher success rate

Launch Success Yearly Trend



As time goes on and technology becomes more advances, we have an upward trend overall in success rate.

All Launch Site Names

- Find the names of the unique launch sites
- We only want one instance of each launch site, so we use the distinct method

```
%%sql
SELECT DISTINCT LaunchSite
FROM SPACEXTBL
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA’
- We use the like method to use as an exact comparator only for the first three values of the string

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LaunchSite like 'CCA%'
LIMIT 5
```

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- We want to sum over the payload mass only where the launch site is NASA

```
%%sql
SELECT sum(payload_mass__kg_)
FROM SPACEXTBL
WHERE LaunchSite = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- We want the average payload only where the version is F9 v1.1

```
%%sql
SELECT avg(payload_mass__kg_)
FROM SPACEXTBL
WHERE version = 'F9 v1.1'
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- We want the earliest (minimum) date where the outcome is true, successful

```
%%sql
SELECT min(date)
FROM SPACEXTBL
WHERE Outcome like 'True%'
```

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
- We want the booster only with payloads between 4000 and 6000

```
%%sql
SELECT BoosterVersion
FROM SPACEXTBL
WHERE payload_mass__kg_ BETWEEN 4000 AND 6000
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- We want to count all instances of the outcome being true - successful
- We can change True to False to count the failures

```
%%sql
SELECT COUNT(*)
FROM SPACEXTBL
WHERE Outcome like 'True%'
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- We want the boosters with the maximum payload so we must compute the max payload separately

```
%%sql
SELECT BoosterVersion, payload_mass_kg
FROM SPACEXTBL
WHERE payload_mass__kg_ =
    (SELECT Max(payload_mass_kg)
    FROM SPACEXTBL)
```

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We want lots of information from the year 2015 including the month thus the use of substr to extract information from the date variable

```
%%sql
SELECT substr(Date,4,2),
    Outcome,
    BoosterVersion,
    LaunchSite
FROM SPACEXTBL
WHERE substr(Date,7,4) = '2015'
    AND Outcome like 'False%'
```

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

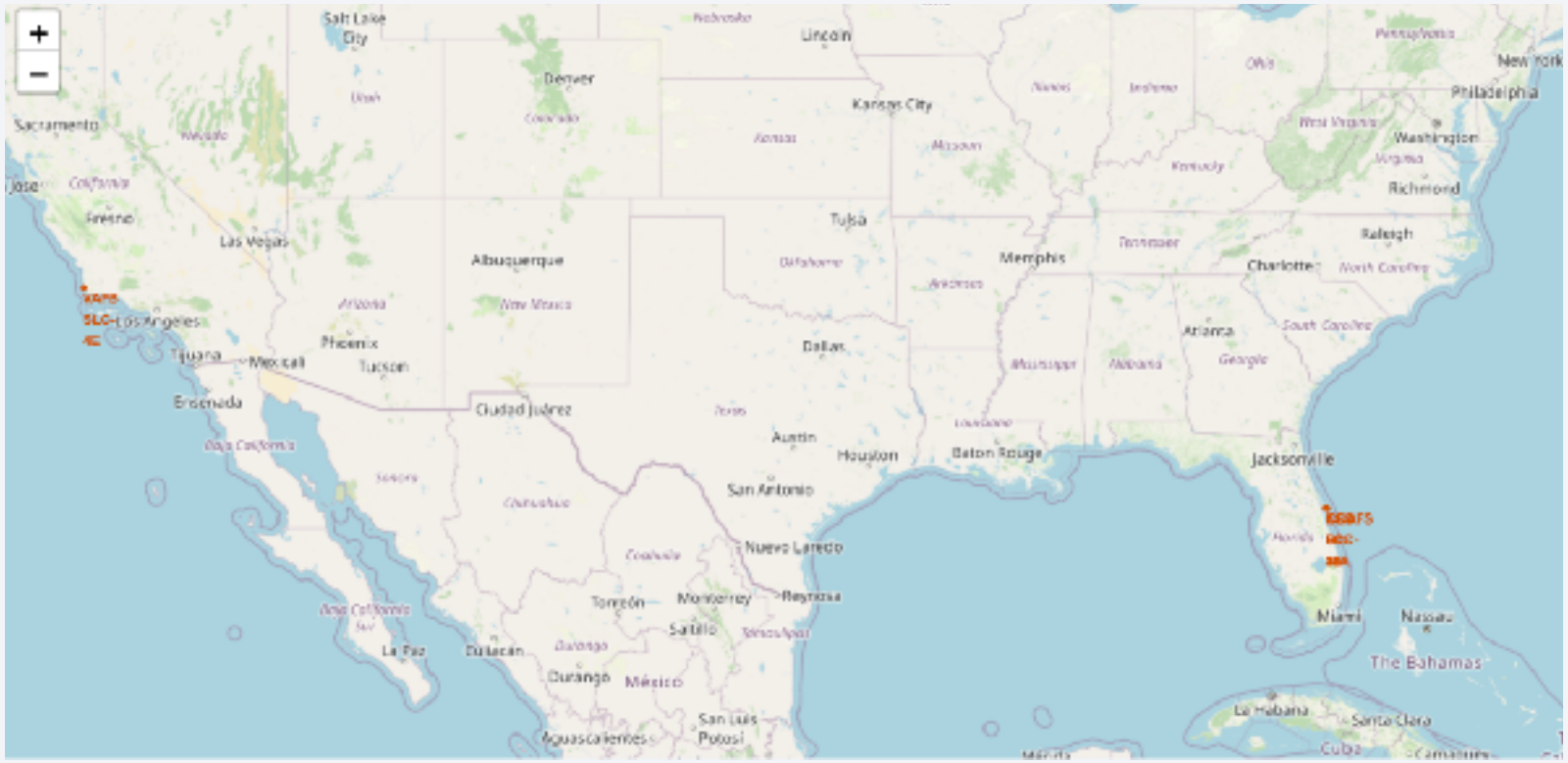
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- We want to select dates within our range and then order them based upon being the most recent

```
%%sql
SELECT FlightNumber
FROM SPACEXTBL
WHERE date BETWEEN '04-06-2010' AND '20-03-2017'
ORDER BY date DESC
```

Section 3

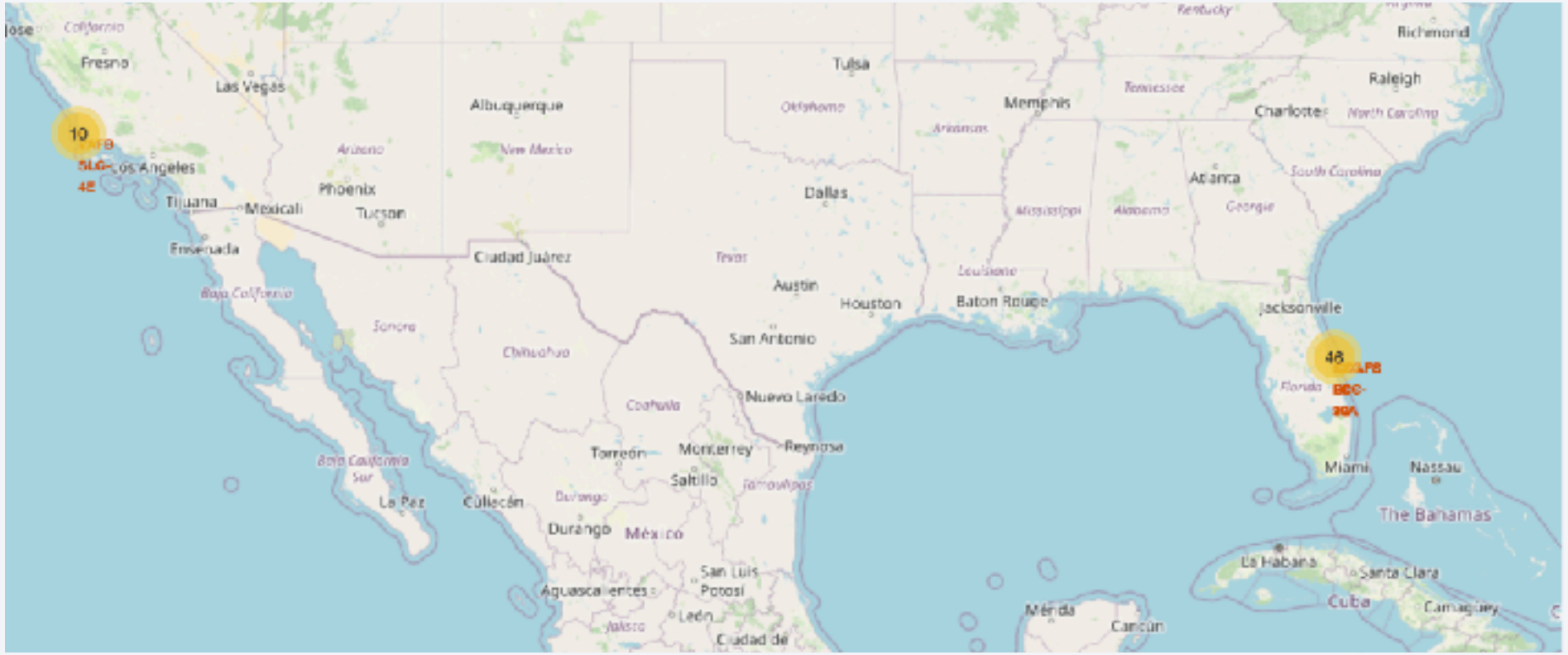
Launch Sites Proximities Analysis

Creating Markers



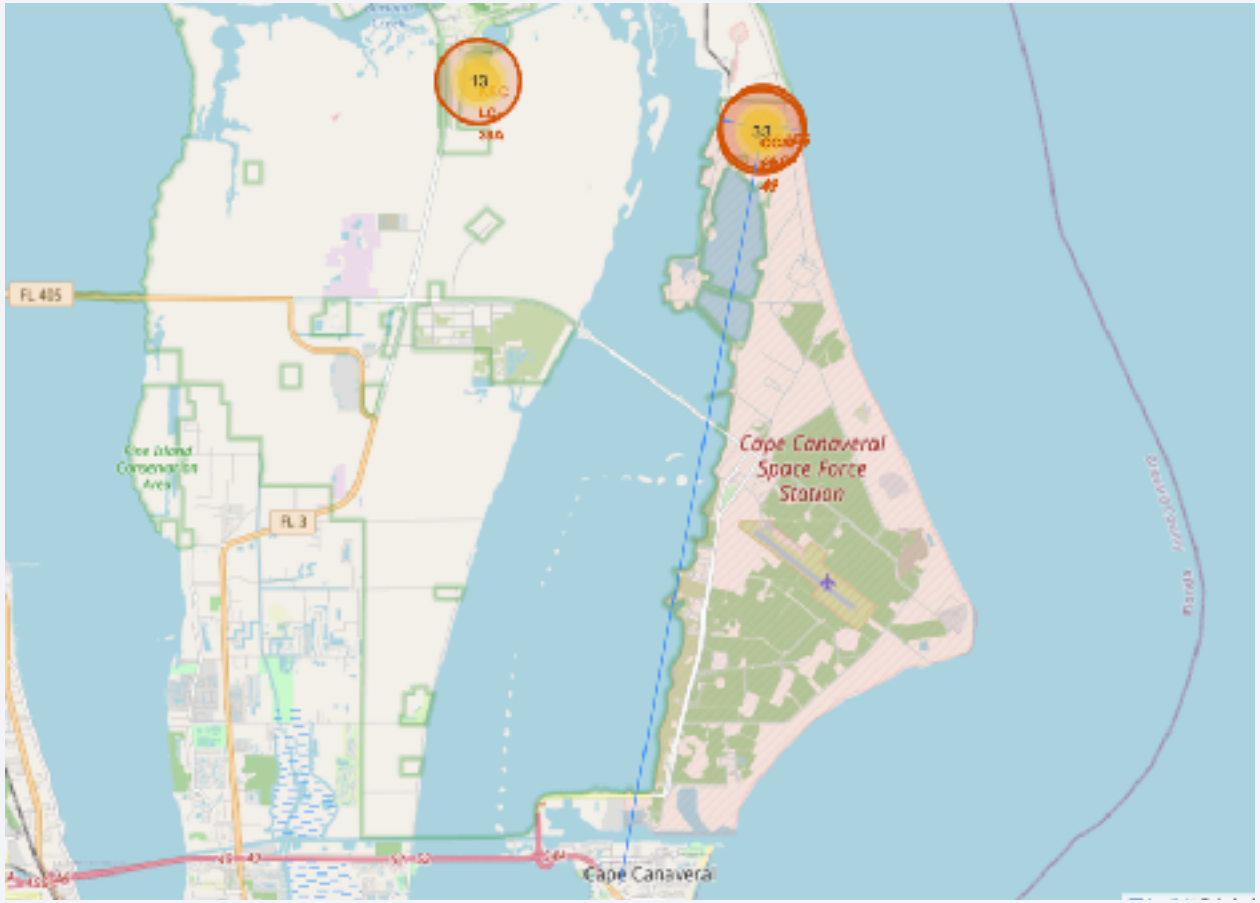
Markers at different launch sites were created so we could observe features around the areas.

Adding Success Counts



We added counts of launches and specifically color coded whether or not they were successfully at each launch location.

Observing Distances to Different Features

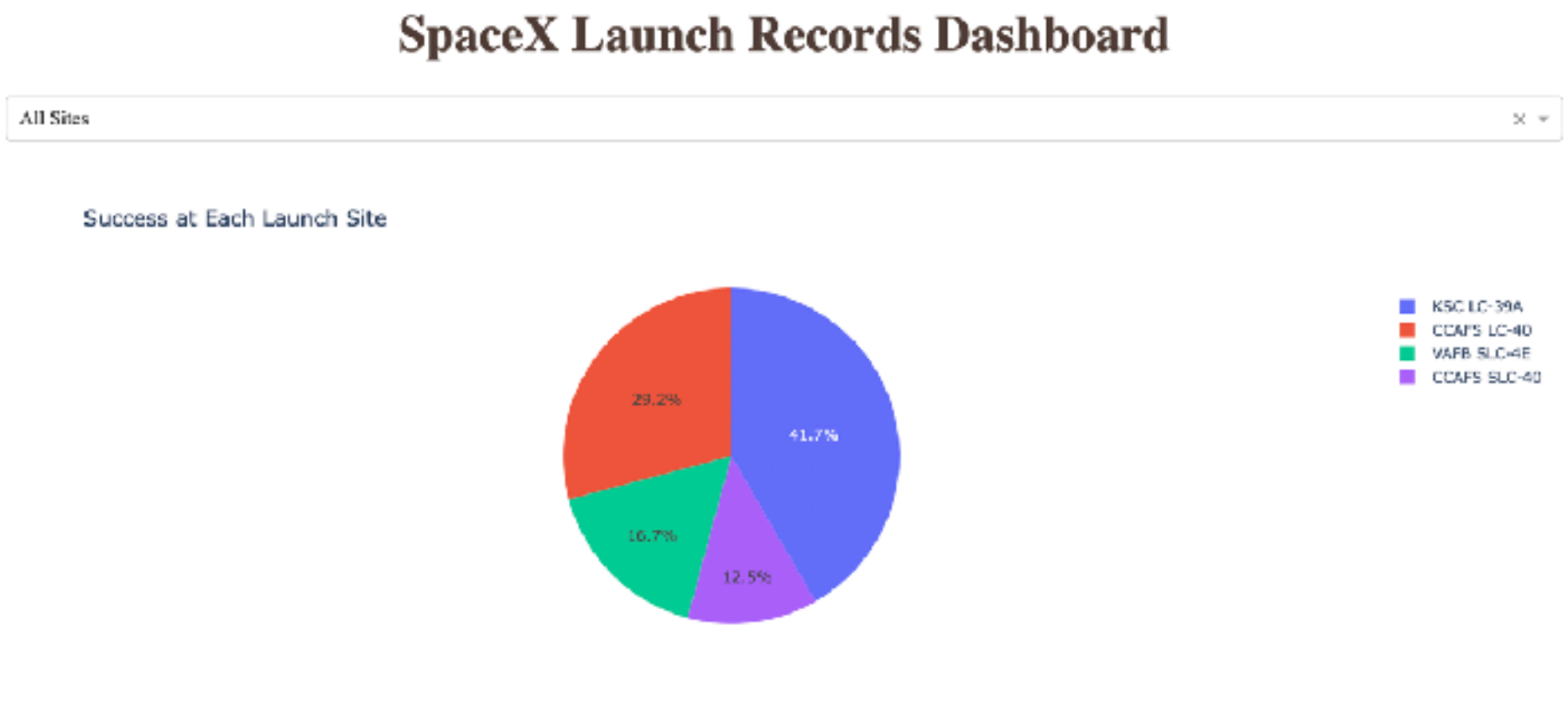


We mapped distances between roads, railroads, coastlines, and nearest cities to observe patterns in launch site locations.

Section 4

Build a Dashboard with Plotly Dash

Success Rate at Different Launch Sites

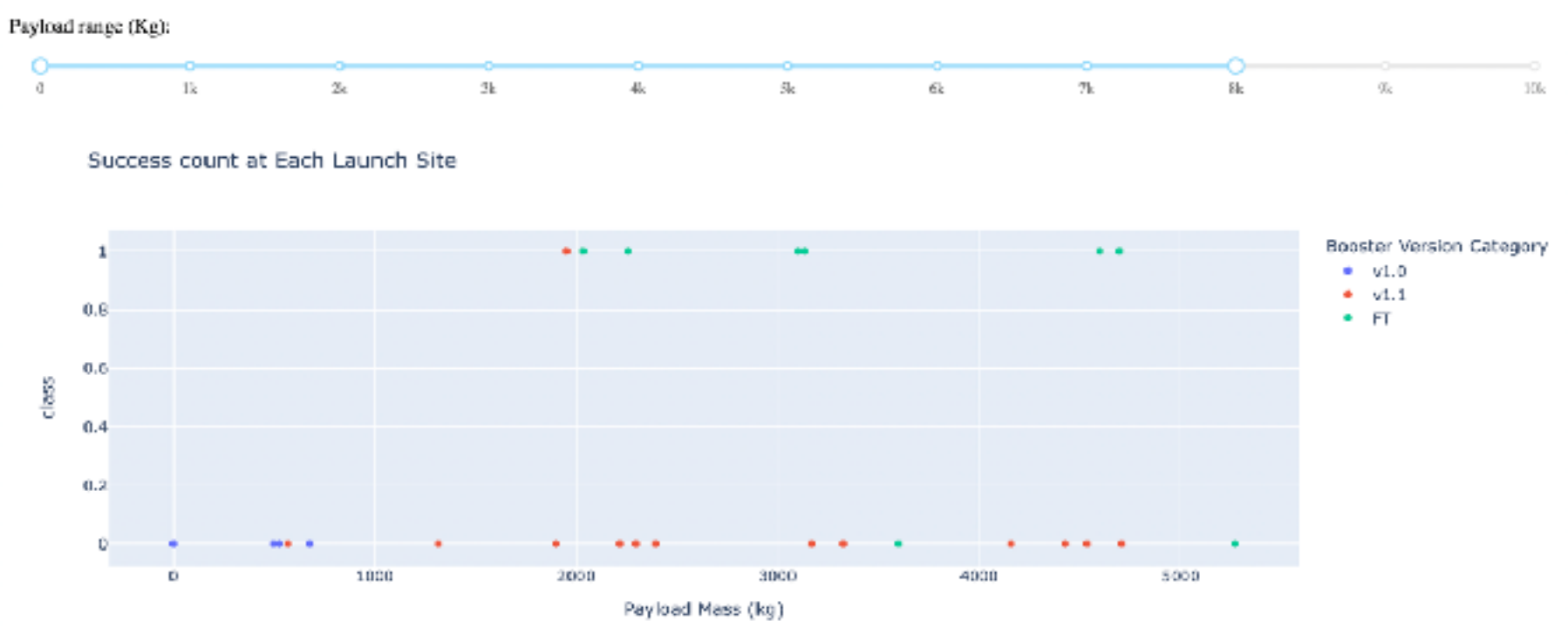


In this dashboard, we selected all launch sites and we compare the proportion of launches at each site.

Most Successful Launch Site

- I was not able to get this chart to function properly. But, we would’ve seen a pie chart for the launch site with the largest slice of pie labeled as successful launch.

Payload vs. Launch Outcome



We can see in this portion of our dashboard that we are able to use a slider to choose the ranges of payload we are interested in seeing appear in this plot.

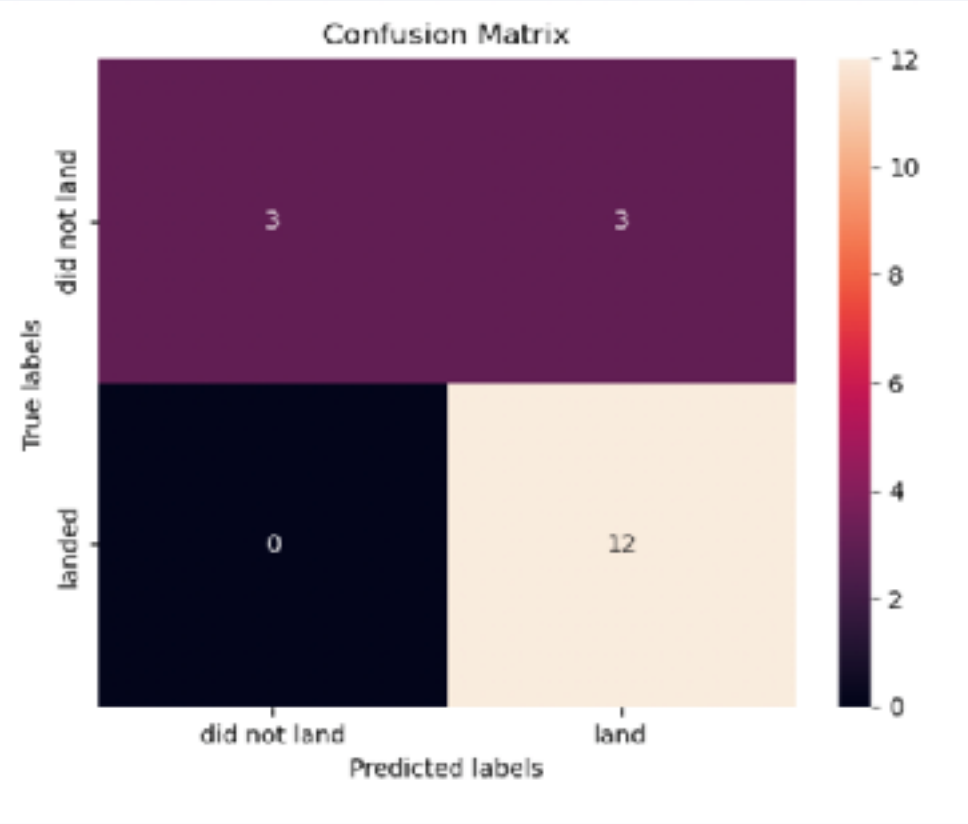
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- KNN, Logistic regression, and SVM models all had the same accuracy within testing data. Because our test sample was not very large (18 elements), this is not surprising. All had an accuracy of 83.3%
- Decision tree method was slightly less accurate in this case.

Confusion Matrix



This is an identical confusion matrix we obtained for all three methods mentioned on the previous slide.

Conclusions

- There are multiple models for successfully predicting the success of a launch with over an 80% accuracy rate.
- As technology improves in time, we are able to have a higher success rate and find better orbit paths for these launches.
- SpaceX is working in the right direction and will become more even more cost efficient in time as their success rate in reusing the first step becomes higher.

Appendix

- The link below contains all of my files for this project.
- I used information from the previous IBM labs and videos in this Data Science course
- https://github.com/plknit00/Final-Project-IBM

Thank you!