IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sarah Spalding
September 16, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Data Visualization
  - SQL Exploratory Data Analysis
  - Folium Interactive Map
  - Plotly Dash Dashboard
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Results
  - Analytics
  - Predictive Analysis

Introduction

- Background:
  - SpaceX Falcon 9 rocket launches cost 62 million dollars
  - Other providers cost upwards of 165 million dollars
  - SpaceX is able to save money by reusing the first stage
- Goal:
  - If we can determine if the first stage will successfully land, we can determine the cost of the launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping Wikipedia
- Perform data wrangling
  - Null values were removed, and one hot encoding was used on categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, K Nearest Neighbor, Support Vector Machine, and Decision Tree models were build and evaluated for the best classifier

Data Collection

- Space X launch data was collected from the SpaceX REST API
- URL: https:// api.spacexdata.com/v4/launches/past

Request JSON from API → Load JSON Response → Save result as Pandas Dataframe

- Space X launch data was also collected from the SpaceX Wikipedia page
- This was done by webscraping using beautiful soup

Use HTTP get method to request wiki page → Create beautiful soup object from HTML response → Format extracted data into a Pandas Dataframe

Data Collection – SpaceX API

Get response from API

```
spacex_url = https://api.spacexdata.com/v4/launches/past
response = requests.get(spacex_url)
```

Convert response to a JSON

```
results = json.loads(response.text)
data = pd.json_normalize(results)
```

Apply function to clean data

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

Filter dataframe then export to a csv

```
data_falcon9.to_csv('dataset_part1.csv', index = False)
```

Assign the lists to a dictionary then dataframe

```
launch_dict = {'FlightNumber': list(data['flight_number']),'Date': list(data['date']),'BoosterVersion':BoosterVersion,'PayloadMass':PayloadMass,'Orbit':Orbit,'LaunchSite':LaunchSite,'Outcome':Outcome,'Flights':Flights,'GridFins':GridFins,'Reused':Reused,'Legs':Legs,'LandingPad':LandingPad,'Block':Block,'ReusedCount':ReusedCount,'Serial':Serial,'Longitude': Longitude,'Latitude': Latitude}
data = pd.DataFrame.from_dict(launch_dict)
```

https://github.com/sspalding/Data-Science-Labs/blob/2697b9be6a107994db626ce037c033b120bc1a43/Course10_Data%20Science%20Capstone/Course10Lab1_Collecting%20the%20Data.ipynb

Data Collection - Scraping

Get Response from HTML

```
response =
requests.get(static_url).text
```

Create beautifulsoup object

```
soup =
BeautifulSoup(response,'html5l
ib')
```

Find Tables

```
html_tables =
soup.find_all('table')
```

Get Column Names

```
column_names = []
for row in
first_launch_table.find_all(
'th'):
    name =
extract_column_from_header(r
ow)
    if (name != None and
len(name)>0):

column_names.append(name)
```

Create Dictionary

```
launch_dict=
dict.fromkeys(column_names)
```

Append data to keys

Convert dictionary to dataframe

```
df = pd.DataFrame(launch_dict)
```

Save Dataframe as CSV

```
df.to_csv('spacex_web_scraped.cs
v', index = False)
```

https://github.com/sspalding/Data-Science-Labs/blob/2697b9be6a107994db626ce037c033b120bc1a43/Course10_Data%20Science%20Capstone/Course10Lab2_Data%20Collection%20and%20Webscraping.ipynb

Data Wrangling

Check Null Values

```
df.isnull().sum()/df.count()*100
```

Calculate the Number of Launches from Each Site

```
df['LaunchSite'].value_counts()
```

Calculate the Occurrence of Mission Outcome Per Orbit Type

```
landing_outcomes =
df['Outcome'].value_counts()
```

Calculate the Occurrence of Each Orbit

```
df['Orbit'].value_counts()
```

Create a Landing Outcome Label from the Outcomes Column

```
landing_class = []
for elem in df['Outcome']:
    if elem in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
df['Class']=landing_class
```

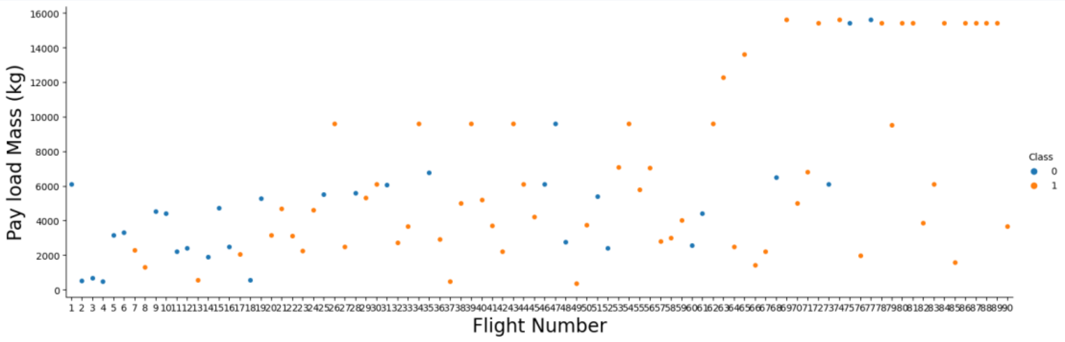
Save to CSV

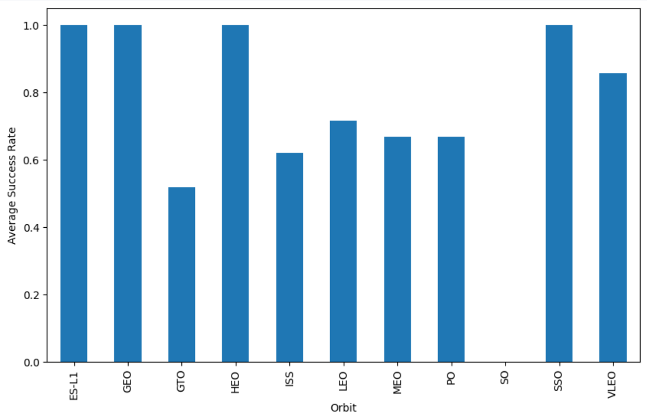
```
df.to_csv('dataset_part2.csv',
index = False)
```

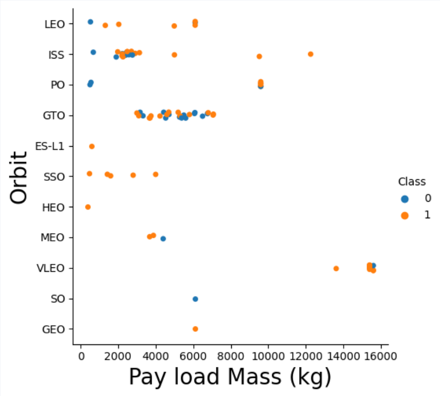
https://github.com/sspalding/Data-Science-Labs/blob/2697b9be6a107994db626ce037c033b120bc1a43/Course10_Data%20Science%20Capstone/Course10Lab3_DataWrangling.ipynb

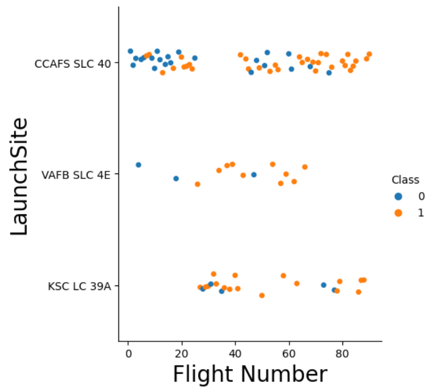
EDA with Data Visualization

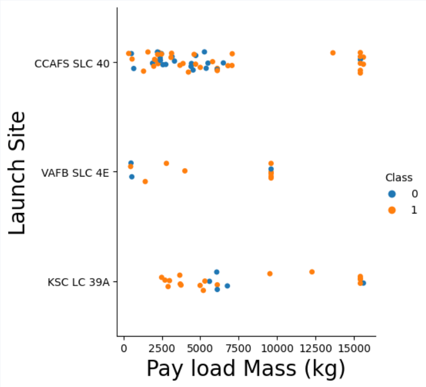
The plots shown below were used to identify which conditions had any correlation to the success of the stage 1 booster landing

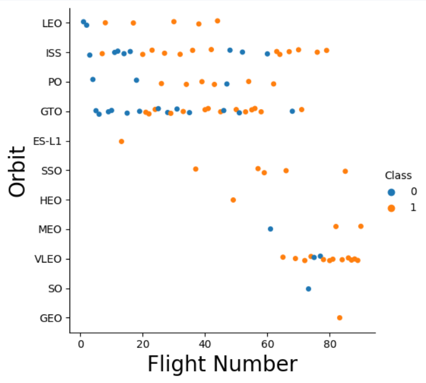


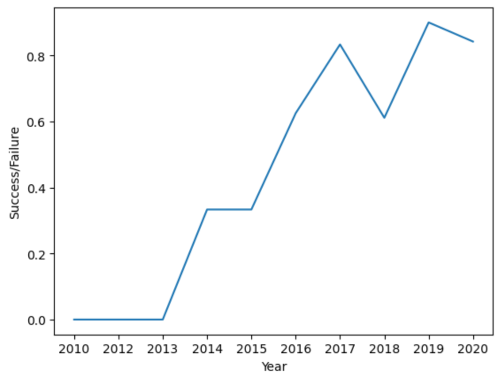












https://github.com/sspalding/Data-Science-Labs/blob/2697b9be6a107994db626ce037c033b120bc1a43/Course10_Data%20Science%20Capstone/Course10Lab5_Exploring%20and%20Preparing%20Data.ipynb

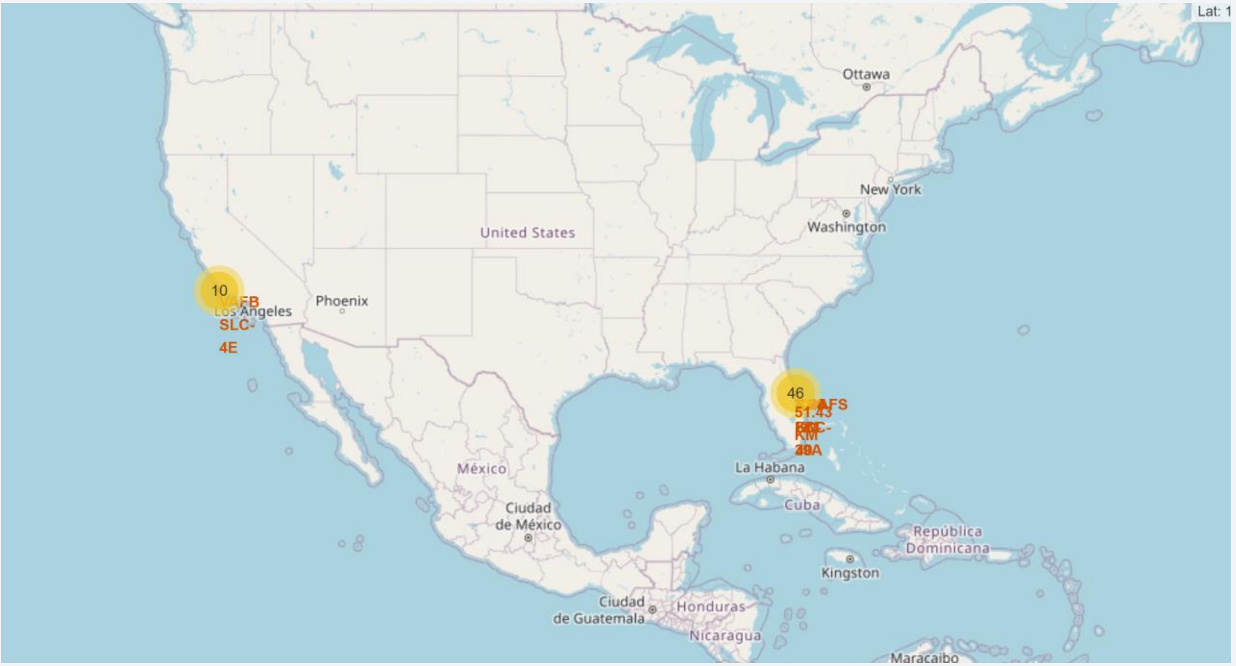
EDA with SQL

- SQL Queries Performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass, use a subquery
  - 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.

https://gitLub.com/sspalding/Data-Science-Labs/blob/2697b9be6a107994db626ce037c033b120bc1a43/Course10_Data%20Science%20Capstone/Course10Lab4_Exploratory%20Data%20Analysis.ipynb

Build an Interactive Map with Folium

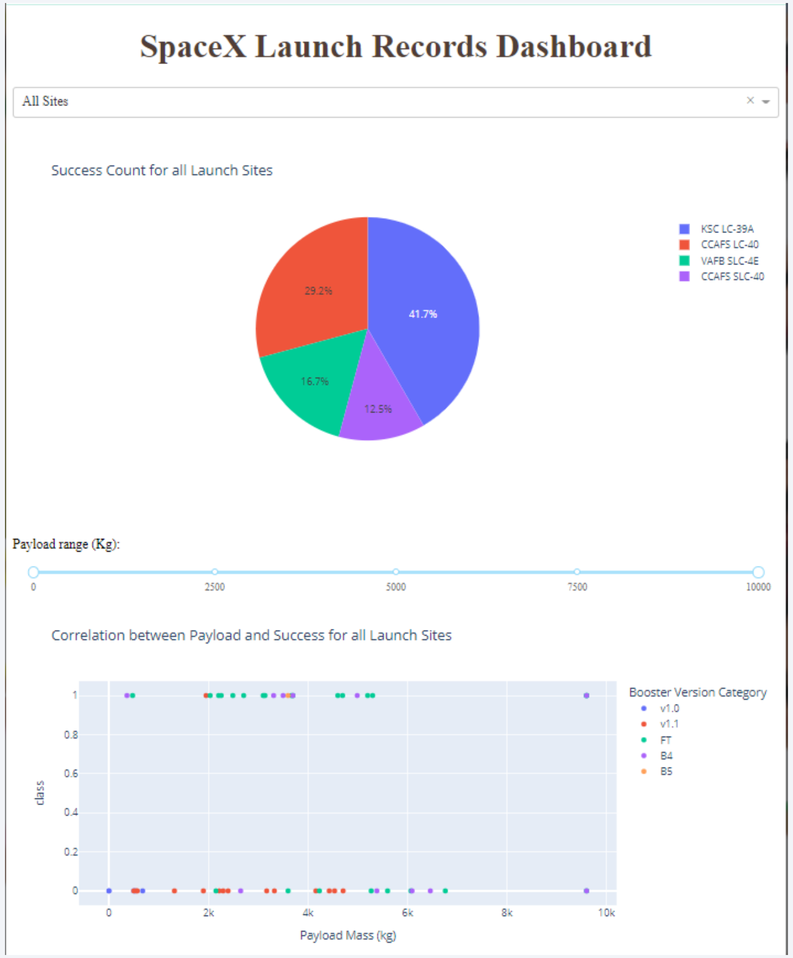
- Map markers were added to the map with the goal of finding the launch site with the highest booster landing
- Marker Clusters were added to group the launches from specific sites to make the map less crowded
- A polyline between the launch site and the nearest coastline, city, railway and highway were added



https://github.com/sspalding/Data-Science-Labs/blob/72c0218c3c57cbdb7883c62a5056df6f3f1a2837/Course10_Data%20Science%20Capstone/Course10Lab6_Location%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

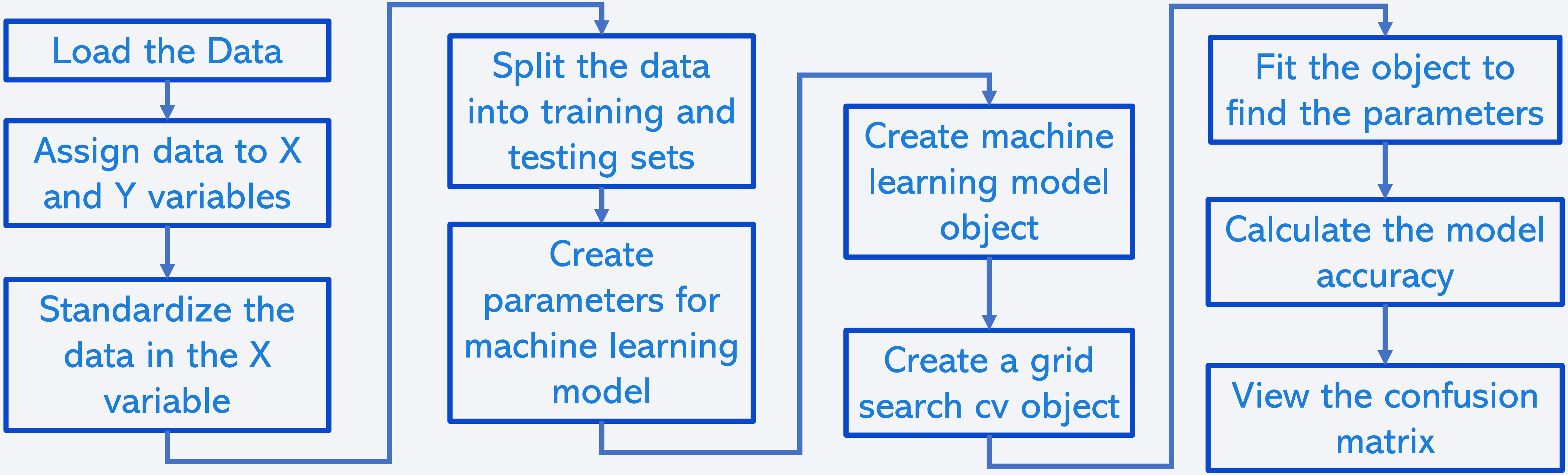
- A drop-down menu was added to the dashboard to allow users to choose the launch site(s) they wish to see daya on
- An interactive pie chart was added to show users the total successful launch count for their selection from the drop-down menu
- A slider was added to allow users to choose the payload size
- A scatter plot was added to show users the correlation between their chosen payload and launch success for different sites



https://github.com/sspalding/Data-Science-Labs/blob/72c0218c3c57cbdb7883c62a5056df6f3f1a2837/Course10_Data%20Science%20Capstone/Course10Lab7_Dashboard%20Application%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- The below process was performed for Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors machine learning models
- The accuracy of the model was compared to determine which was the best



https://github.com/sspalding/Data-Science-Labs/blob/72c0218c3c57cbdb7883c62a5056df6f3f1a2837/Course10_Data%20Science%20Capstone/Course10Lab7_Machine%20Learning.ipynb

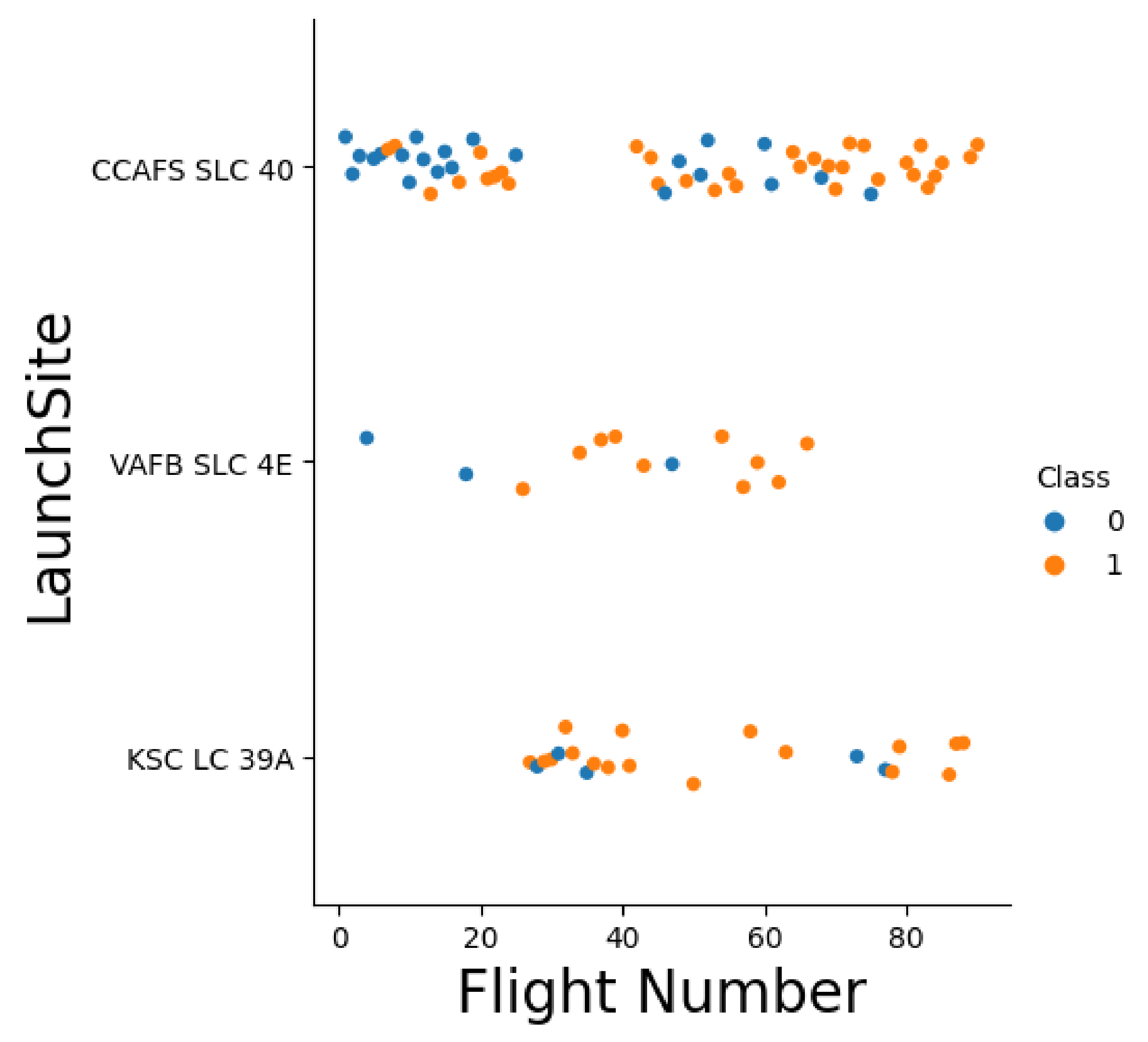
Results

- The success rate of booster landings is correlated to the year, it is predicted that as time goes on their will be less unsuccessful launches
- The launch site with the highest success rate was KSC LC 39A
- The orbits ES L1, GEO, HEO, and SSO had the highest success rates
- Lower weighted payloads had more successful booser landings then higher weighted payloads
- All four machine learning models performed the same in terms of prediction accuracy

Section 2

Insights drawn from EDA

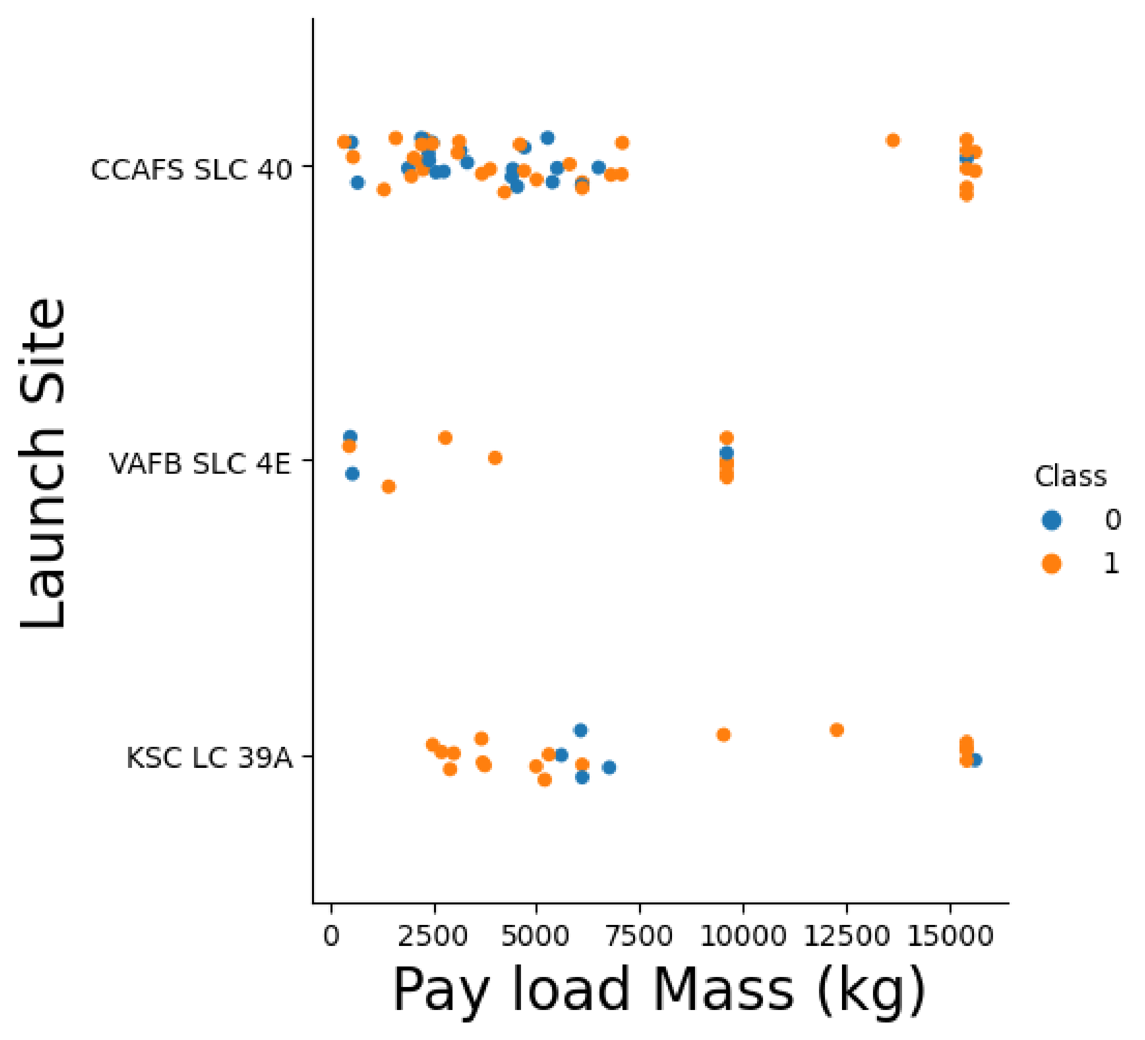
Flight Number vs. Launch Site



- Launch site CCAFS SLC 40 had the most launches

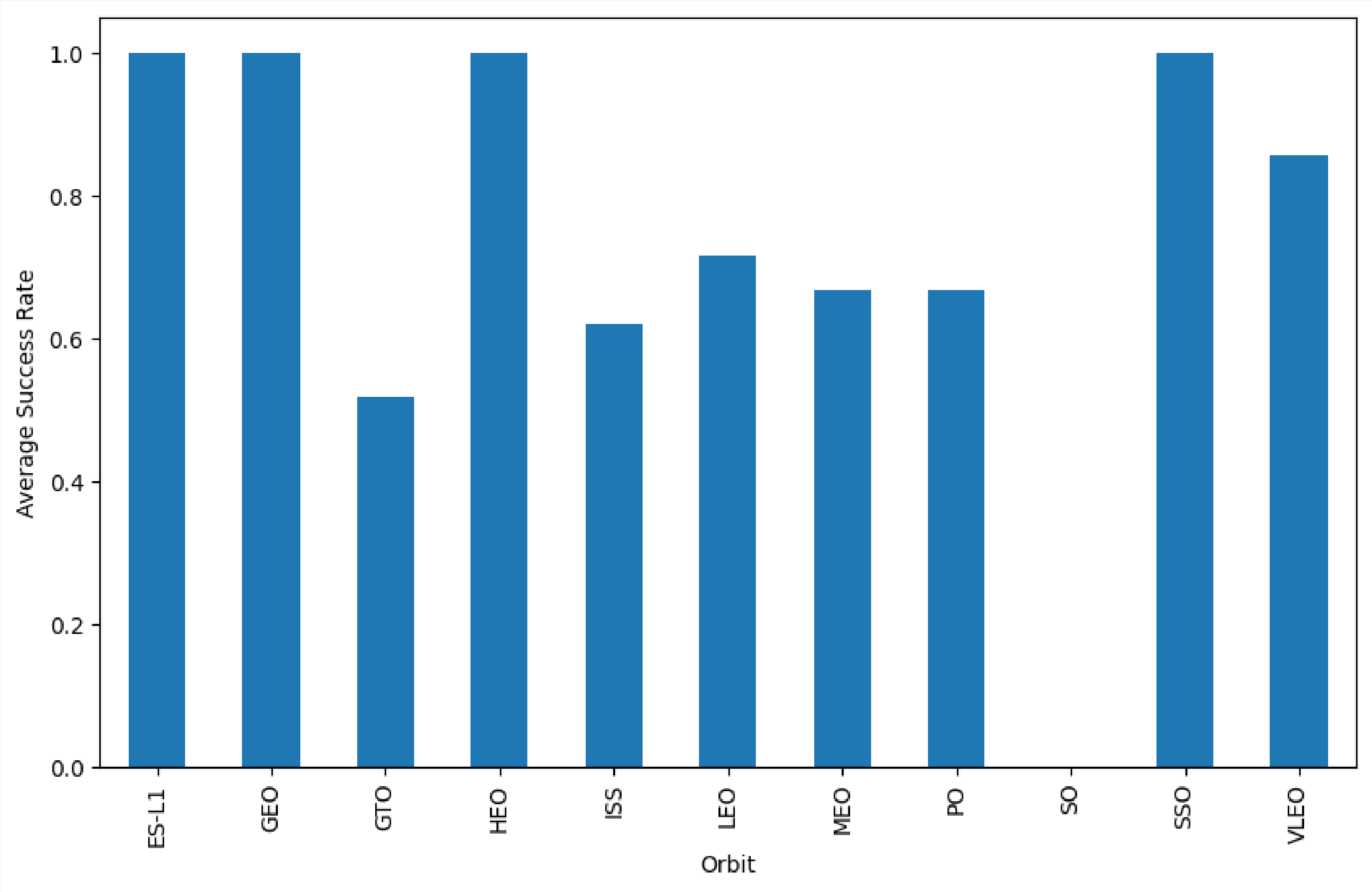
Payload vs. Launch Site

- Launch site VAFB SLC 4E had the payloads with the lowest weight



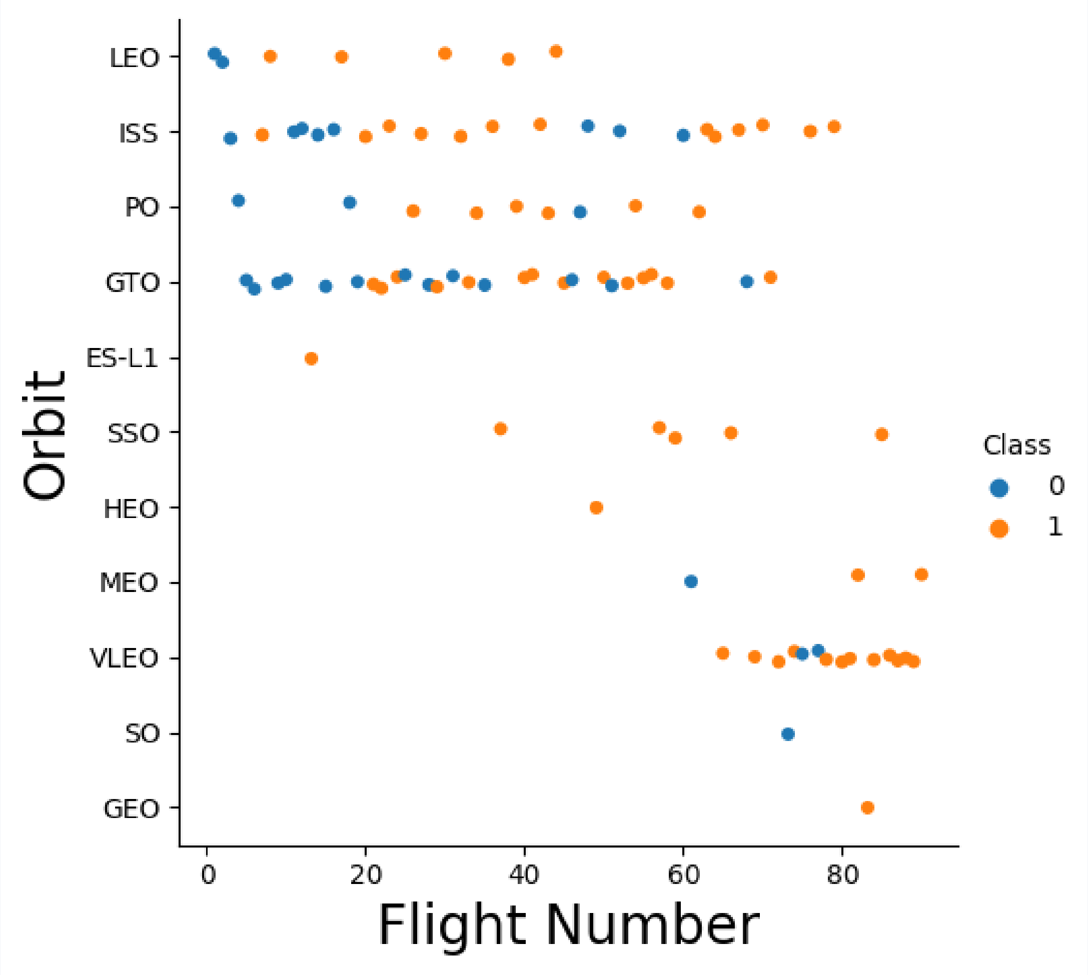
Success Rate vs. Orbit Type

- The launches going to orbits ES L1, GEO, HEO, and SSO had the highest booster landing success rate
- The launches going to orbit SO had the lowest booster landing success rate



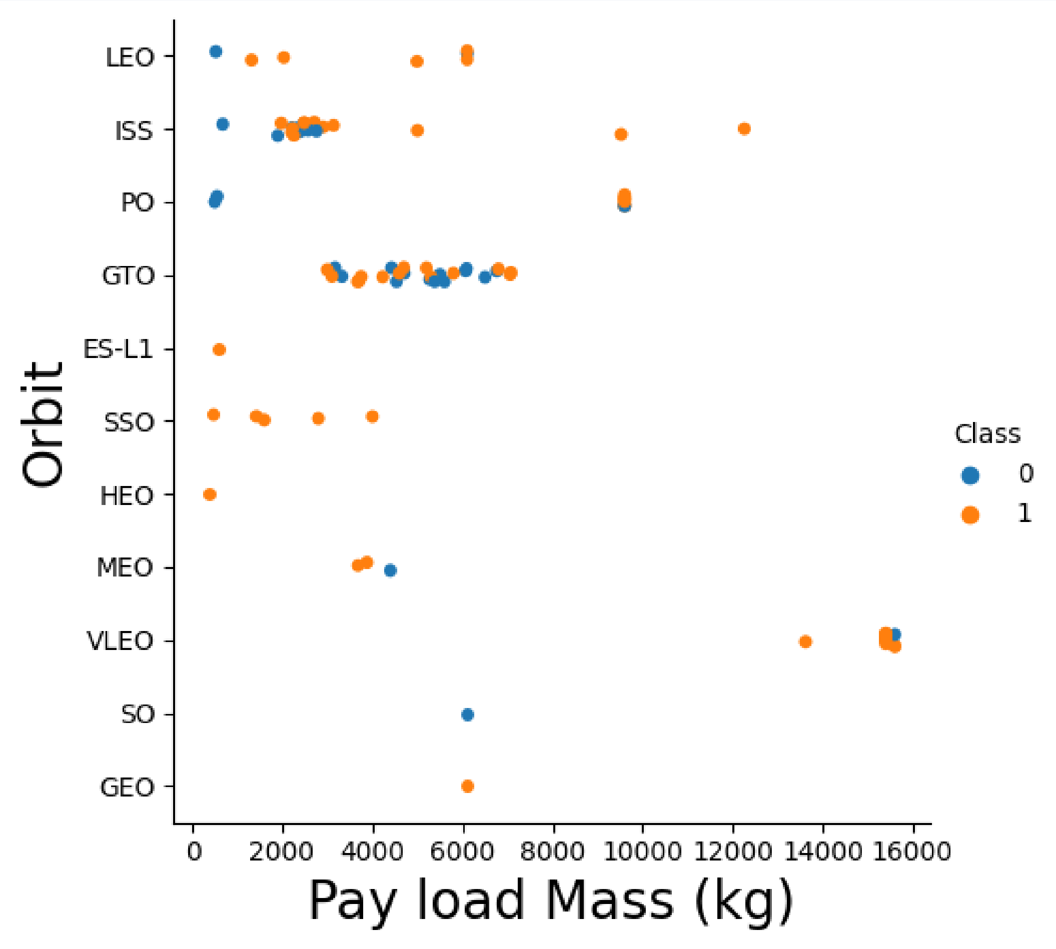
Flight Number vs. Orbit Type

- Recently, more launches have been going to the VLEO orbit



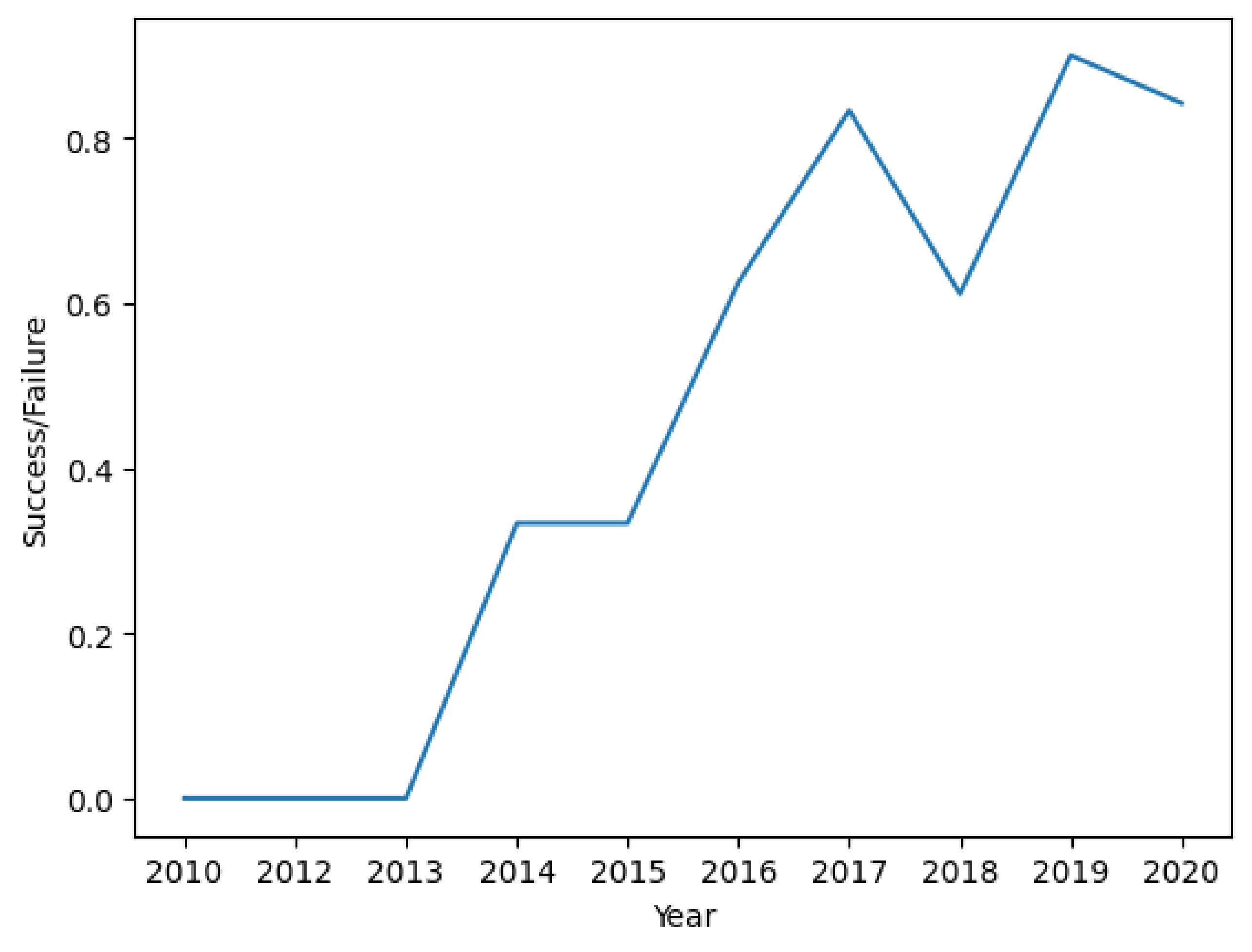
Payload vs. Orbit Type

- Most payloads going to the ISS weigh between 2000 and 4000 kg
- Most payloads going to the GTO orbit weigh between 2000 and 8000 kg



Launch Success Yearly Trend

- The success rate has been generally increasing over the years



All Launch Site Names

- The launch site names were queried using the following code

```
%sql select unique(LAUNCH_SITE)
from SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- The 5 records where launch sites begin with `CCA` were found using the following query

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- The total payload carried by boosters from NASA was 45,596
- It was found using the following query

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where
"customer" = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2,928
- It was found using the following query

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where
BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 2015-12-22
- The following query was used to find it

```
%sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME =
'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were found by the following query

```
%sql select BOOSTER_VERSION from SPACEXTBL
where LANDING__OUTCOME = 'Success (drone
ship)' and PAYLOAD_MASS__KG_ between 4000 and
6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes was found using the following query

```
%sql select MISSION_OUTCOME,
count(MISSION_OUTCOME) from
SPACEXTBL group by
MISSION_OUTCOME;
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass was found using the following query

```
%sql select BOOSTER_VERSION from
SPACEXTBL where PAYLOAD_MASS__KG_ =
(select max(PAYLOAD_MASS__KG_) from
SPACEXTBL);
```

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 were found using the following query

```
%sql select MONTH(DATE),
LANDING__OUTCOME,
BOOSTER_VERSION, LAUNCH_SITE
from SPACEXTBL where
EXTRACT(YEAR FROM DATE) and
LANDING__OUTCOME = 'Failure
(drone ship)';
```

1	landing__outcome	booster_version	launch_site
1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
1	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
3	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
6	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

- The ranked 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 were found using the following query

```
%sql select LANDING__OUTCOME from SPACEXTBL where DATE
between '2010-06-04' and '2017-03-20' order by DATE
desc;
```

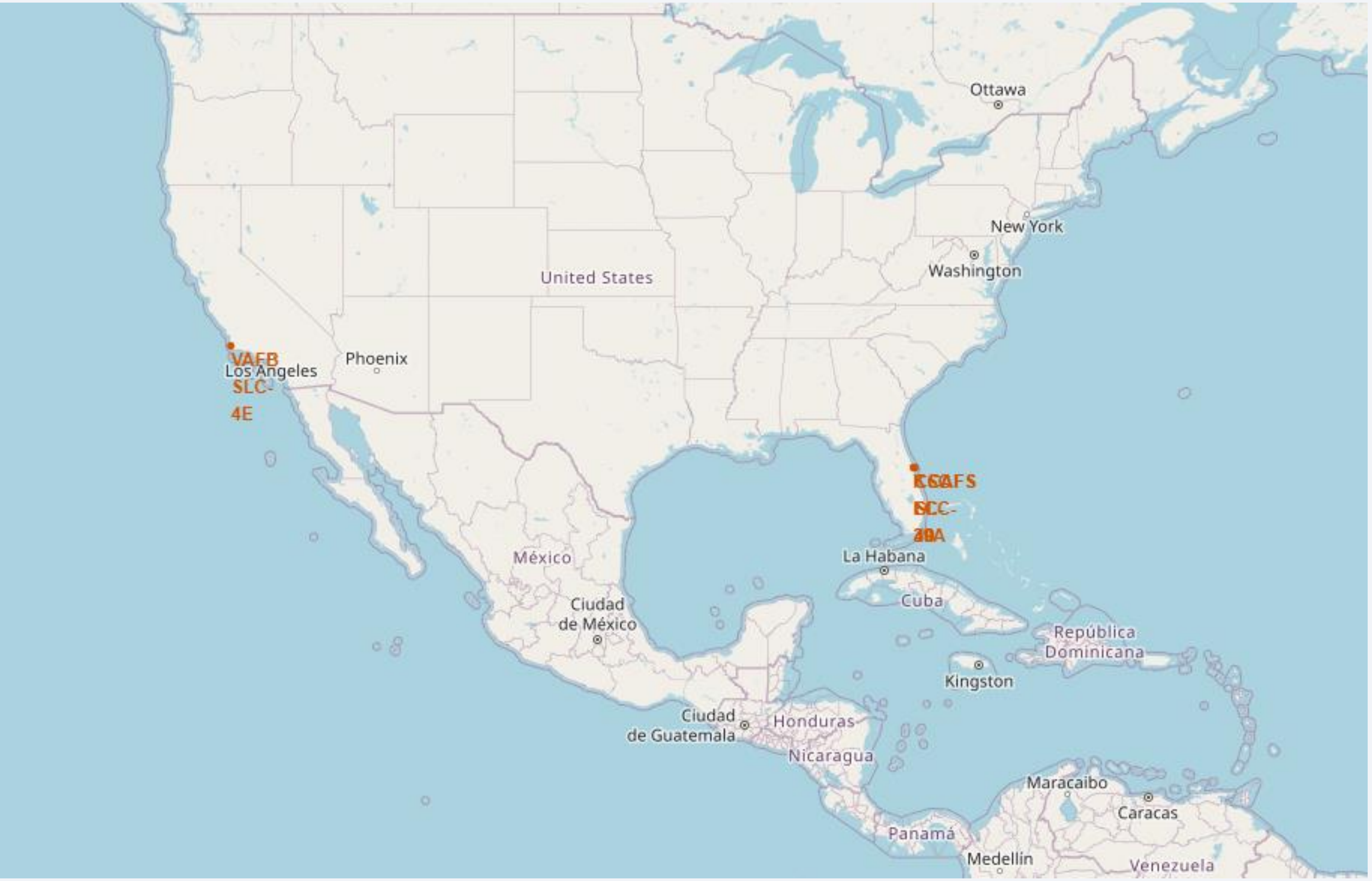
landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 3

Launch Sites Proximities Analysis

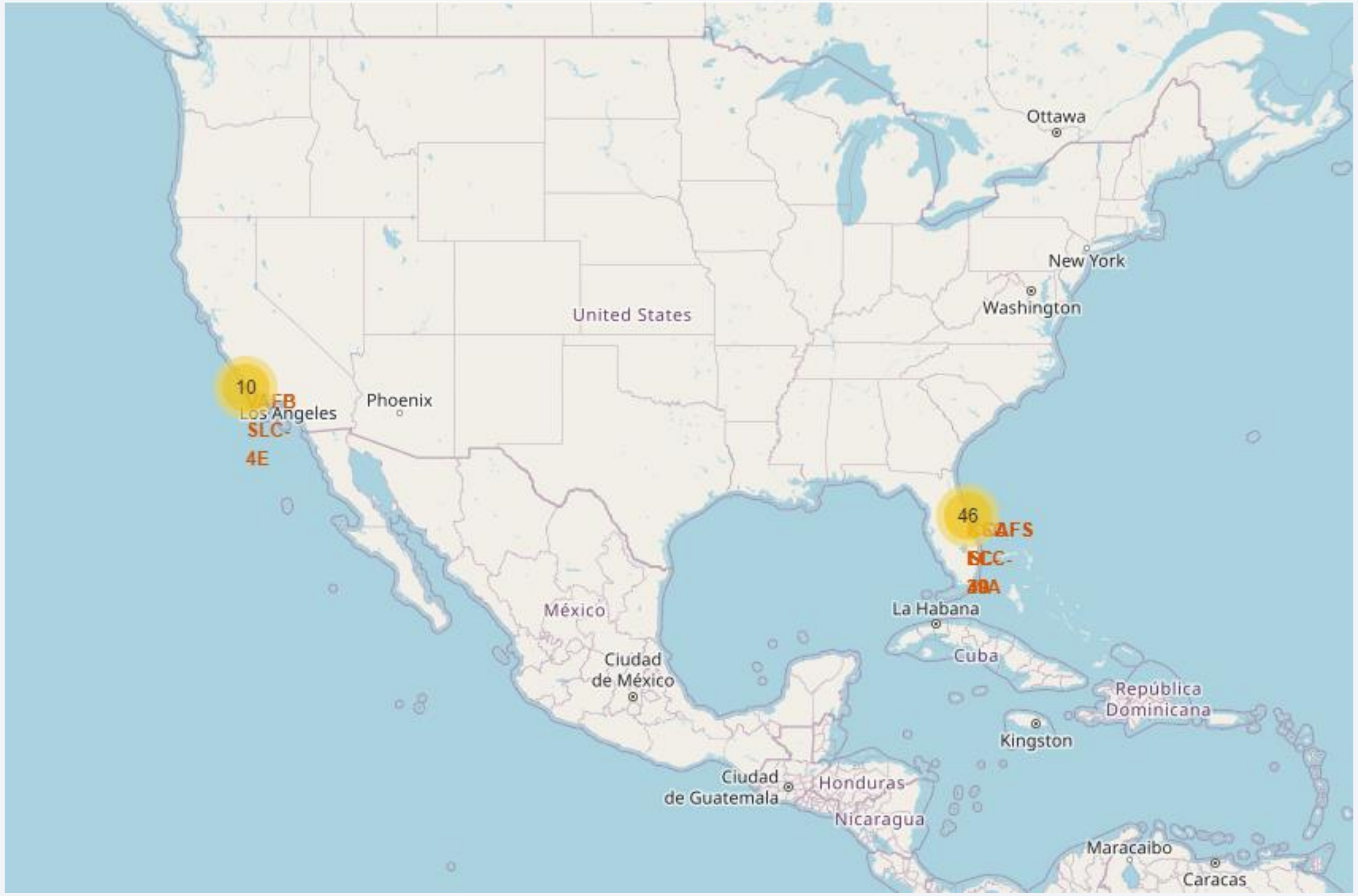
All Launch Site Locations

- Four launch sites are shown on the map:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E



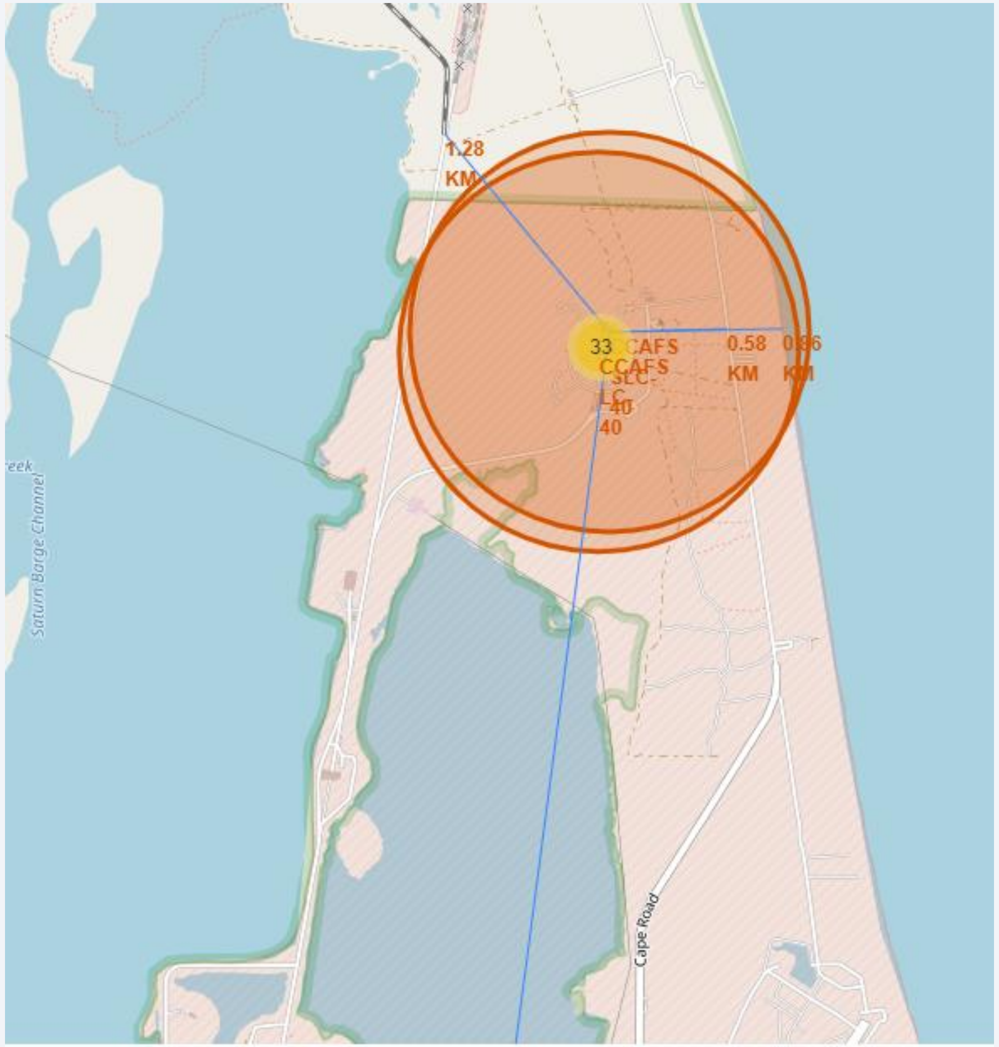
Mapped Launch Outcomes

- The launch success rate was added to the map as color, where green was successful, and red was failure



Launch Site Map with Proximities to Railways, Highways and Coastlines

- The blue line is the proximity line to the locations of interest

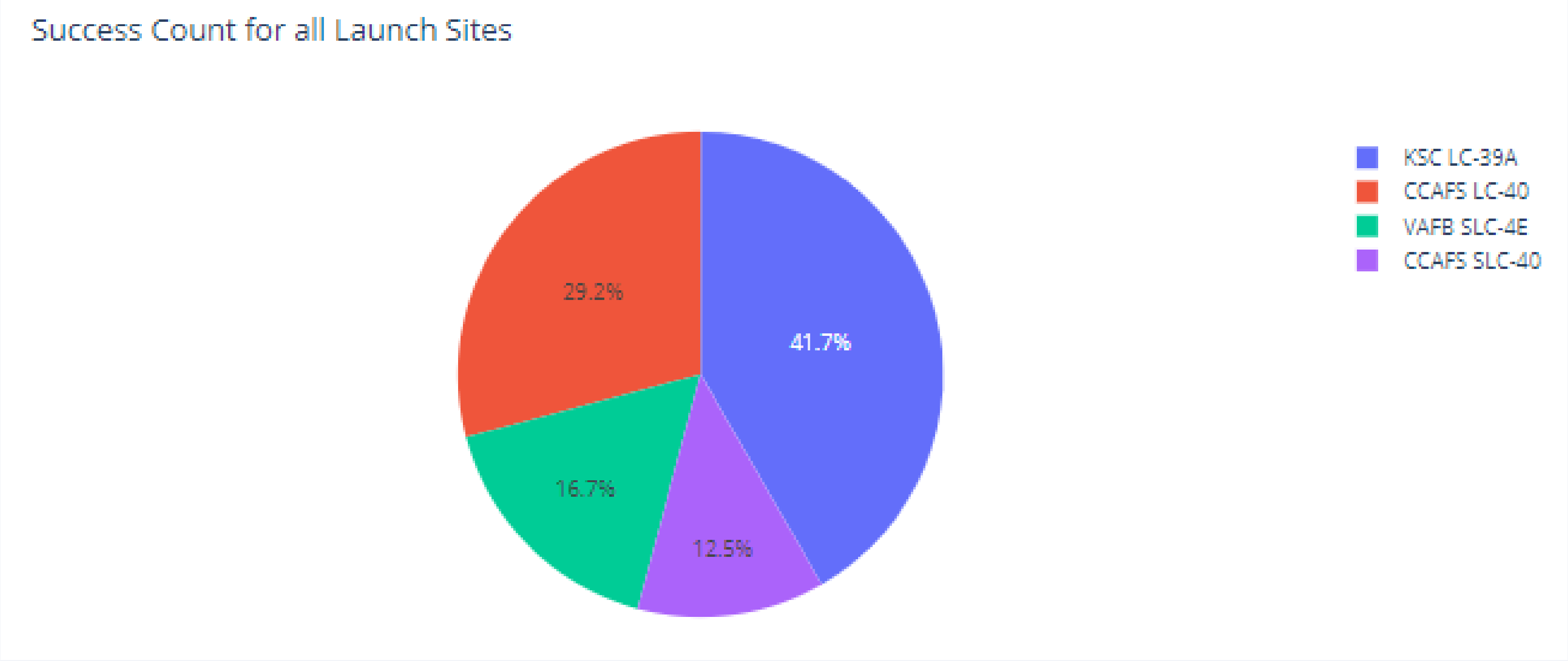


Section 4

Build a Dashboard with Plotly Dash

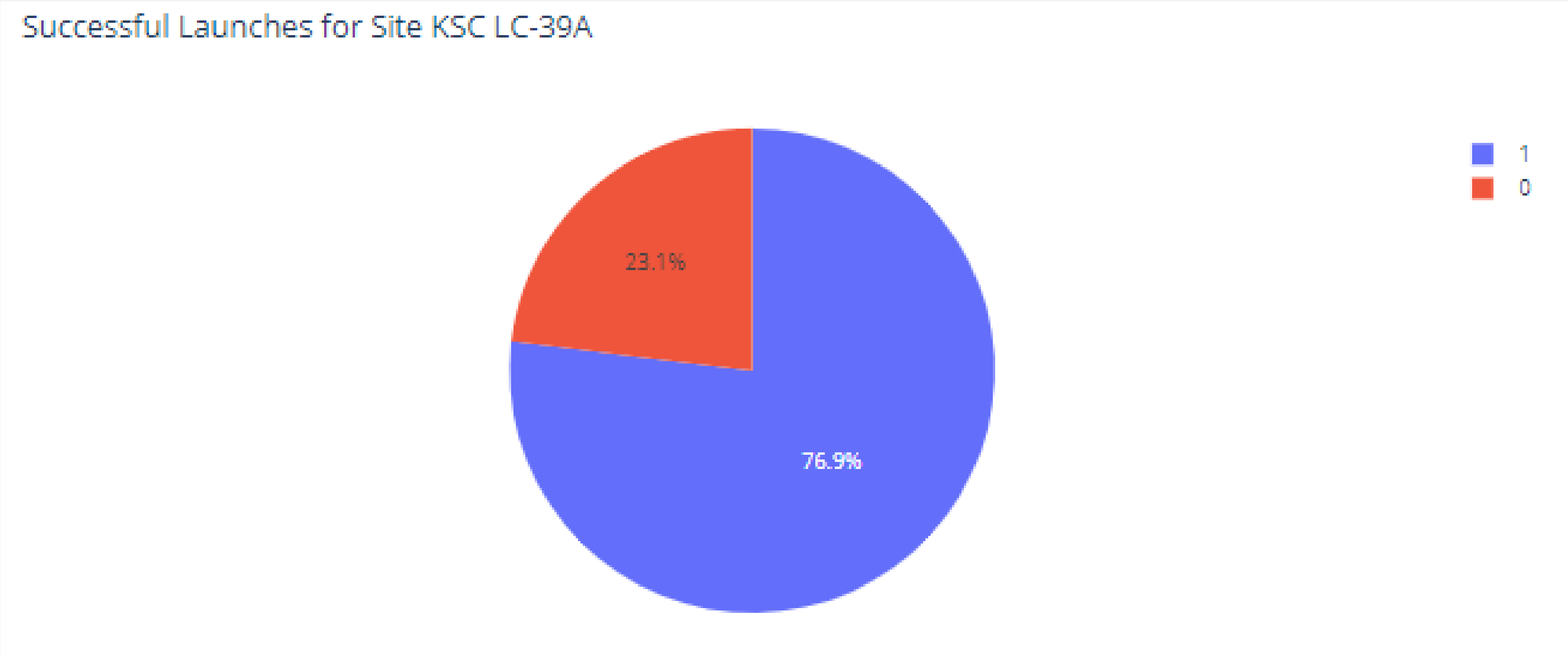
Pie Chart of Success Counts for all Launch Sites

- Hovering over a slice of the pie chart will display more information about that slice



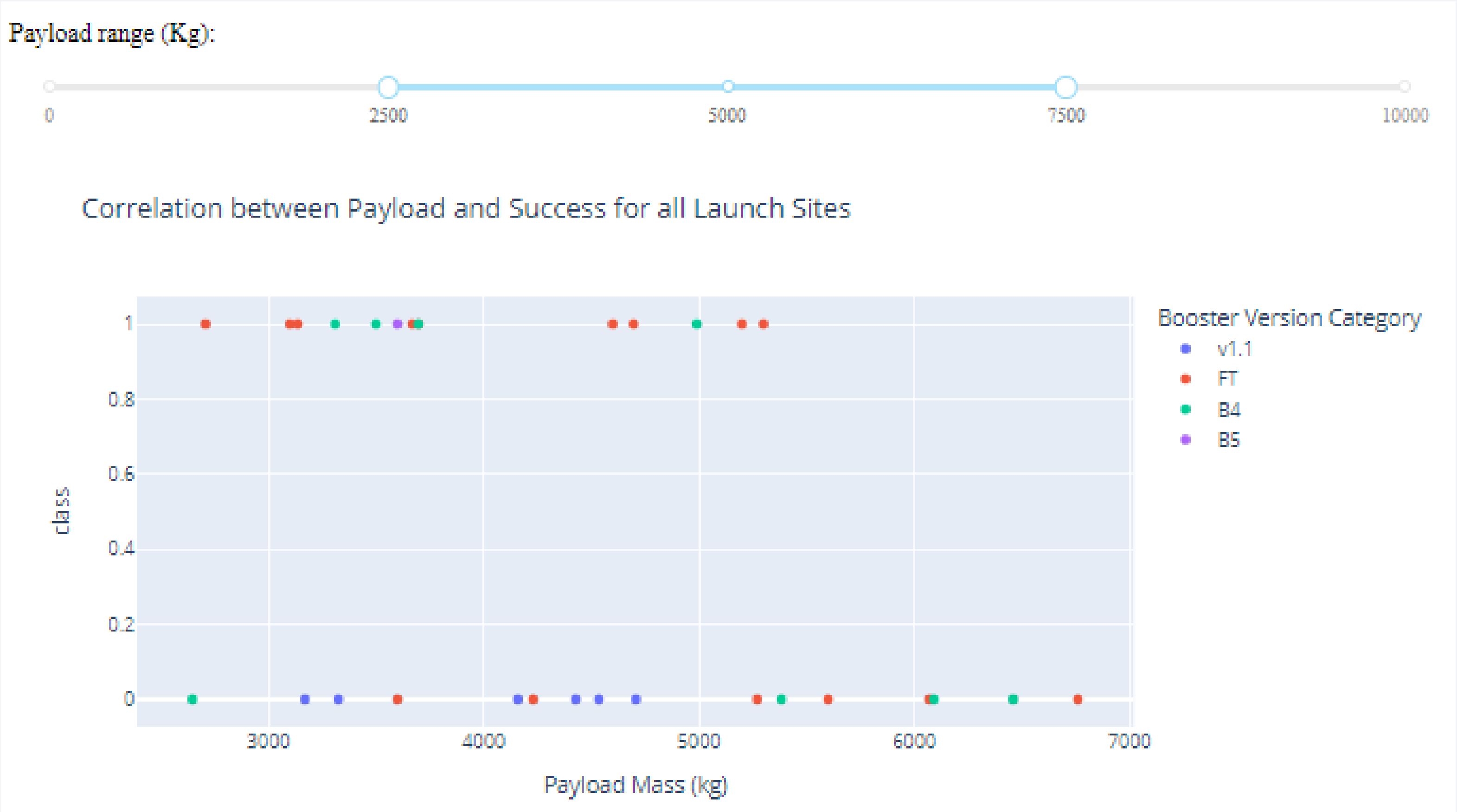
Pie Chart of Successful Launches from KSC LC-39A

- KSC LC-39A had the highest success rate of all the launch sites
- 1 indicates a success, 0 indicates a failure



Payload vs Launch Outcome for all sites

- Adjusting the payload range slider changes what is displayed in the scatter plot
- This chart show a range of 2500 to 7500 kg
- Hovering over individual data points displays more information about that data point

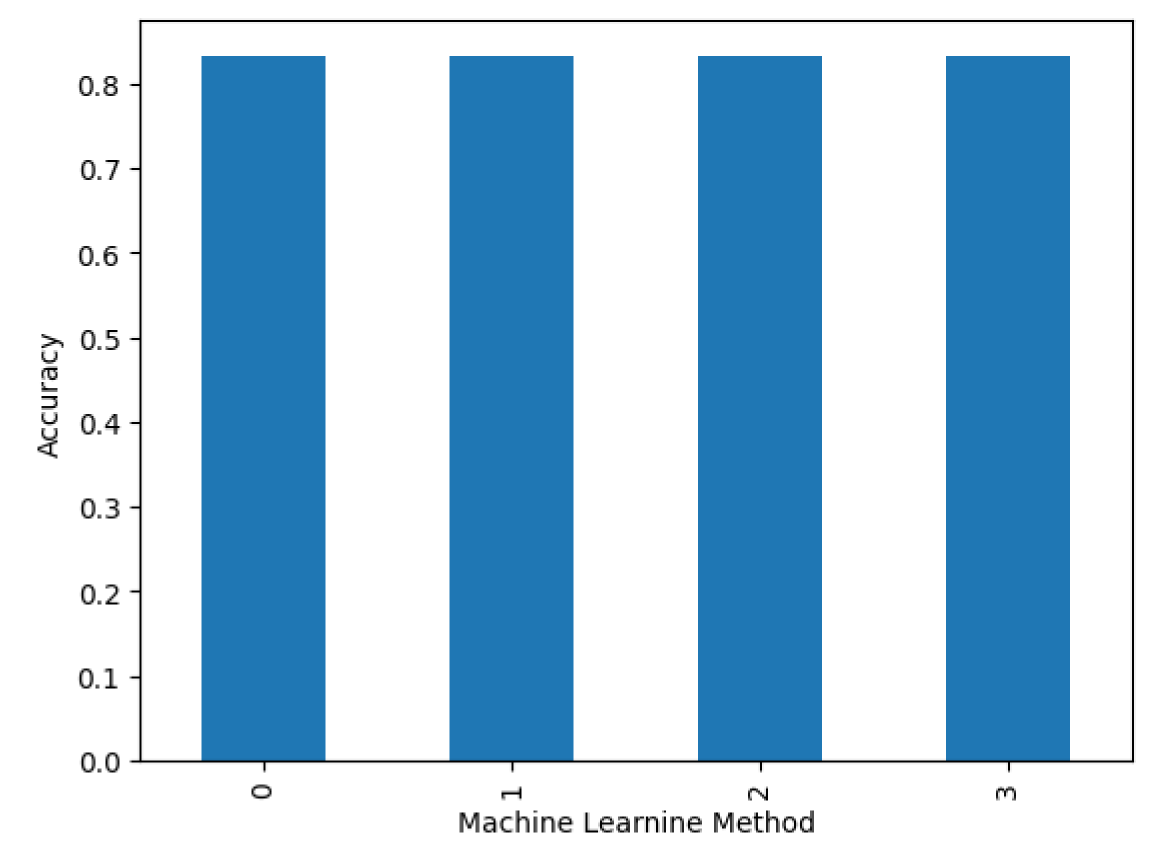


Section 5

Predictive Analysis (Classification)

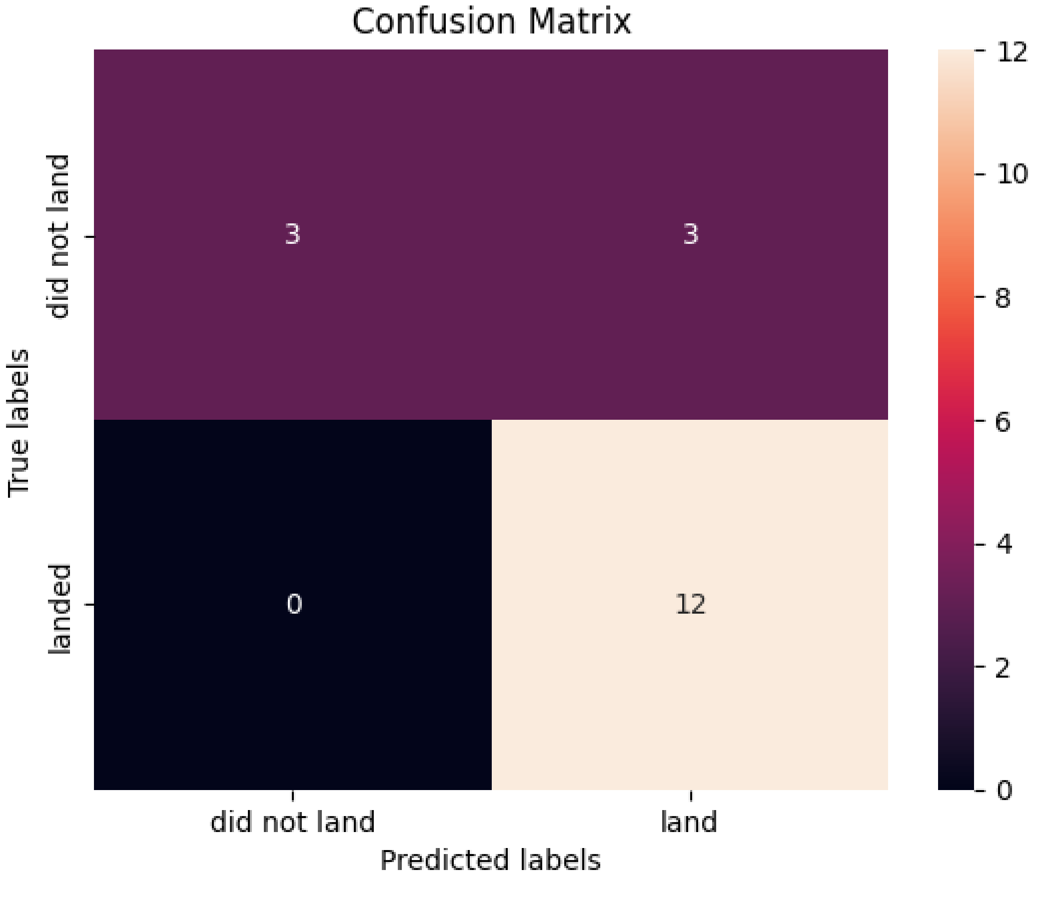
Classification Accuracy

- All models had the same accuracy of 83.3%



Confusion Matrix

- All models had the same confusion matrix
- The confusion matrix shows there were 3 true positives, 3 false negatives, 0 false positives, and 12 true positives



Conclusions

- All four machine learning models performed the same in terms of prediction accuracy
- The conditions to get the highest likelihood of a successful booster landing are:
  - Launch from KSC LC 39A
  - Have a lower weighted payload
  - Go to the GEO, HEO, SSO, or ES L1 Orbit
- Over time the success rate has increased and will most likely continue to do so

Thank you!