IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Melissa Larsen
April 17, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Data Visualization
  - Data Analysis
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Data Analysis Results
  - Predictive Analysis Results

Introduction

- Project Background and Context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems to be Answered
  - Does the rate of successful landings change over time?
  - What is the cost of a launch?
  - Will SpaceX reuse the first stage?
  - What other variables affect the success of the first stage landing?

Section 1

Methodology

Methodology

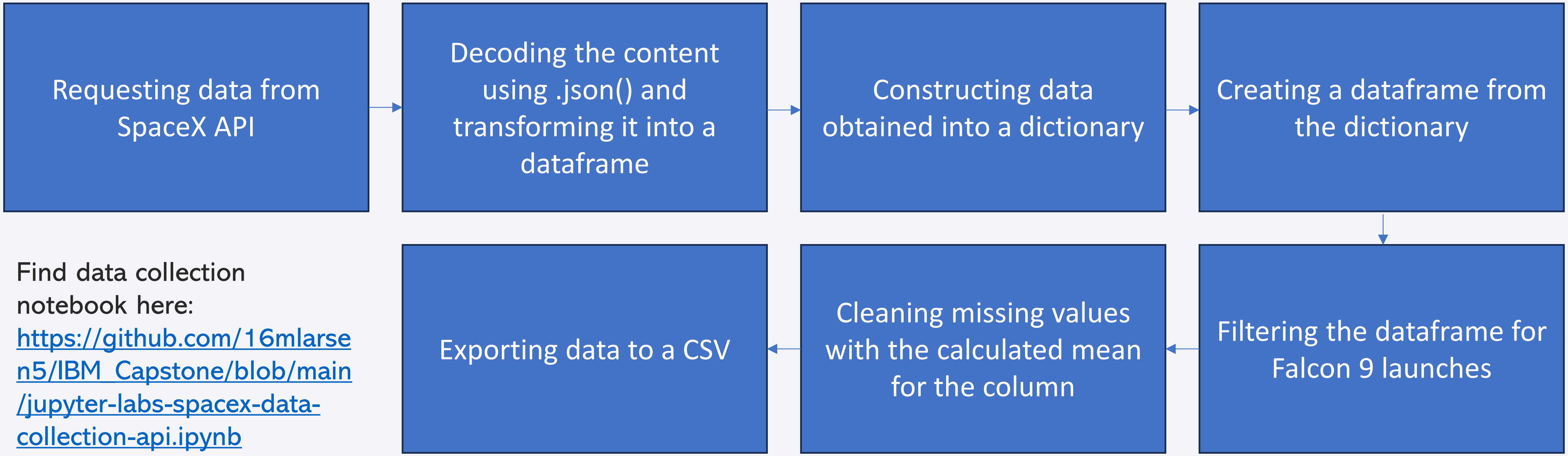
Executive Summary

- Data collection methodology:
  - Use SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models

Data Collection

- Dataset Collection
  - Combination of API requests from SpaceX REST API and web scrapping data from a table on SpaceX's Wikipedia page
  - Both datasets combined provide a more detailed look into launches
    - Data obtained from SpaceX Rest API:
      - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Landing Pad, Block, Reused, Grid Fins, Legs Serial, Longitude, Latitude
    - Data obtained from Wikipedia Table:
      - Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



Find data collection notebook here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Requesting data from Wikipedia

Creating a BeautifulSoup object from the response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

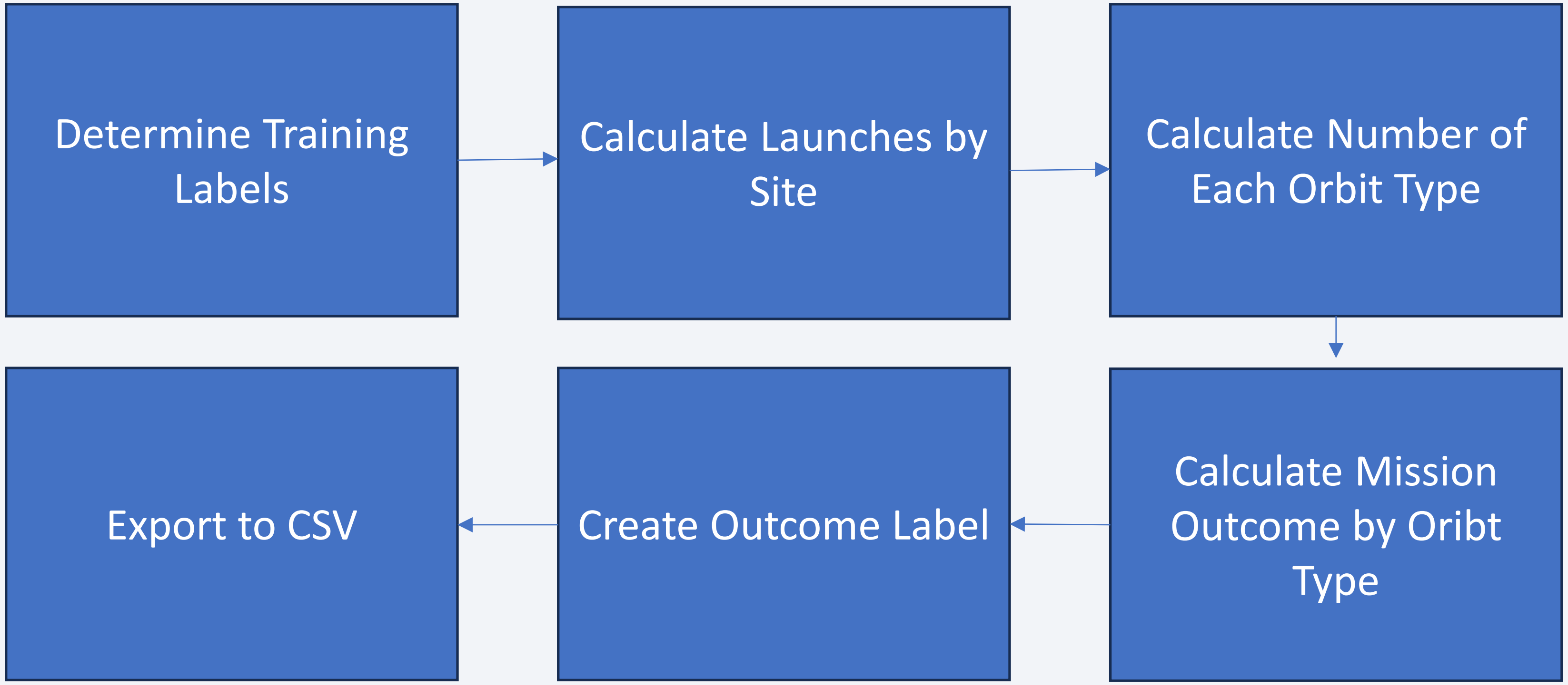
Creating a dataframe from the dictionary

Exporting the data to a CSV

Find data collection notebook here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

- The outcome of the booster landing successfully or not is converted into Training Labels. A 1 was set to indicate the booster successfully landed while a 0 meant it was an unsuccessful landing.
- Find data wrangling notebook here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Several different chart types were plotted to show the relationships between variables. Charts that were plotted include:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type and Success Rate
- The type of charts used were scatter plots, bar charts, and line graphs. The scatter plots showed relationship between variables, the bar charts helped compare categories, and the line graphs showed data trends over time.
- Find the EDA with Data Visualization Notebook here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/edadataviz.ipynb

EDA with SQL

SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes as Failure (drone ship) or Success (ground pad)

- The EDA with SQL notebook can be found here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers of Launch Sites:

- Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center.
- Markers with Circle, Popup Label and Text Label of all Launch Sites
- Colored markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Coloured Lines to show distances between the Launch Site and Railways, Highways, Coastlines and Closest City.

Interactive Visual Analytics Notebook can be found here:
https://github.com/16mlarsen5/IBM_Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard had a dropdown list, pie chart, slider, and scatter chart
  - Dropdown list with launch site selection
  - Pie chart showing successful launch count
  - The slider allows the user to select payload range.
  - Scatter plot of payload mass and success rate by booster versions
- The Dash App can be found here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
  - create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data
- You can find the predictive analysis notebook here: https://github.com/16mlarsen5/IBM_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

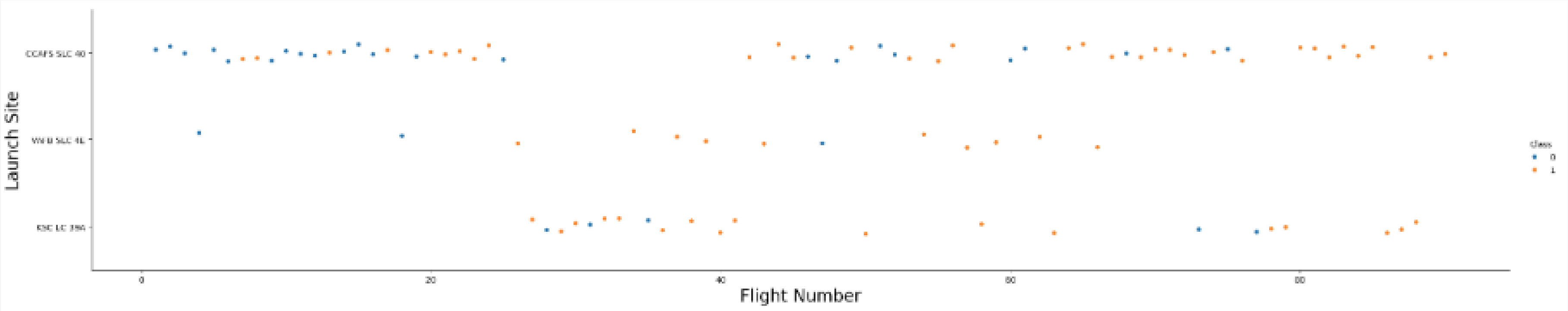
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

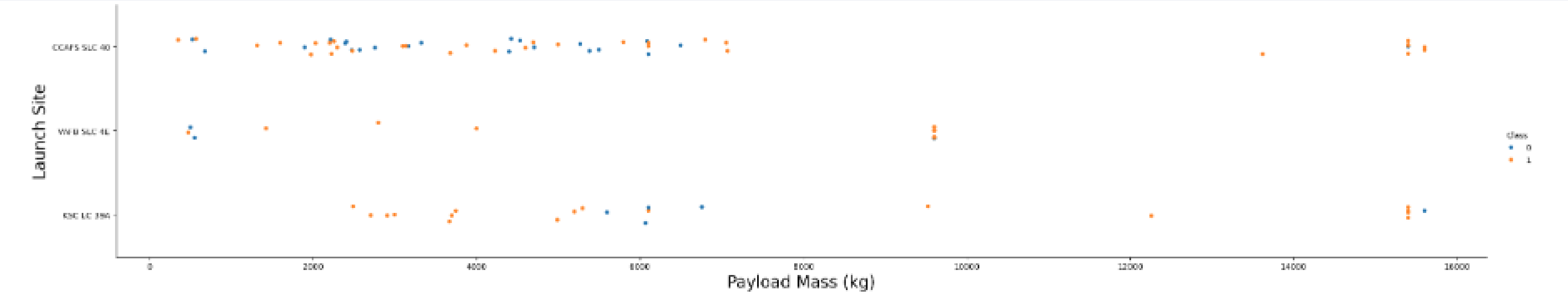
Insights drawn from EDA

Flight Number vs. Launch Site



- The smaller flight numbers mostly failed
- The CCAFA ALC 40 has the most failures but also the most launches
- The other two launch sites have a higher success rate

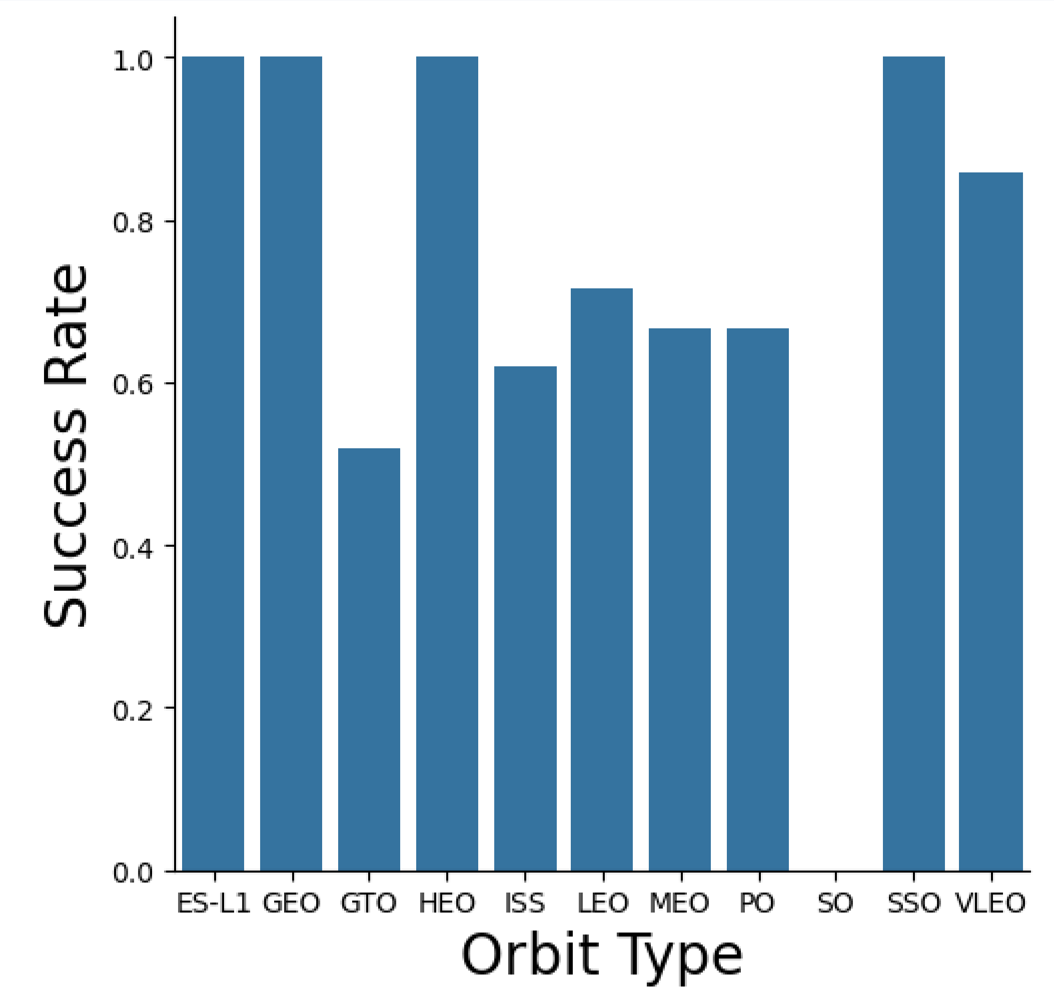
Payload vs. Launch Site



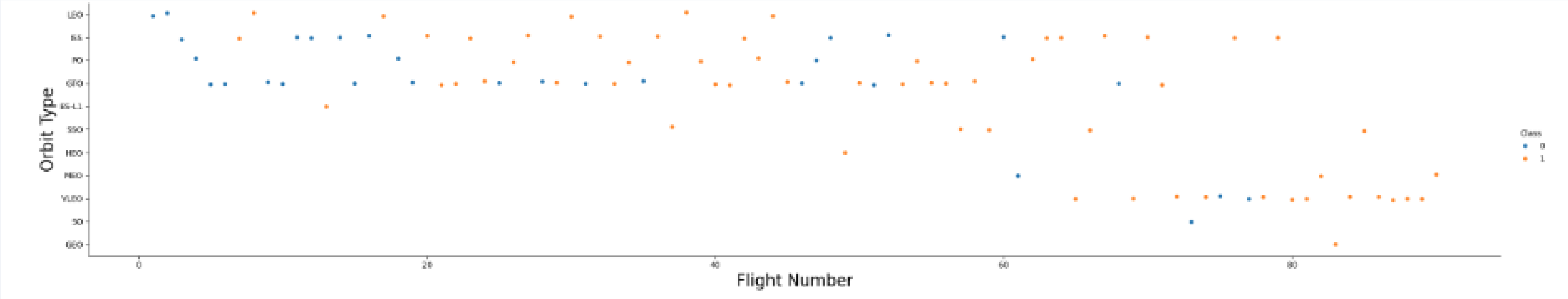
- The majority of launches were below 8000 kg mass
- The launches that were higher had a higher success rate
- KSC LC 39A has a 100% success rate below 5000 kg mass

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO all have a 100% success rate
- SO has a 0% success rate

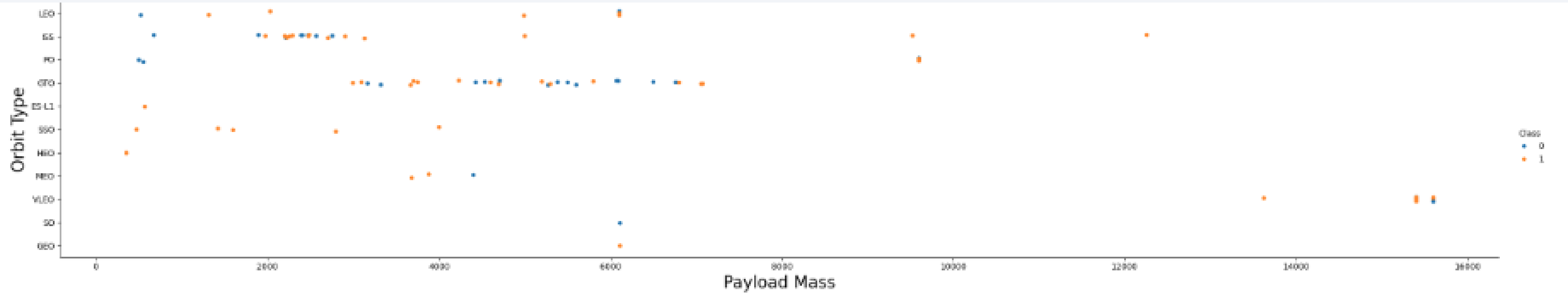


Flight Number vs. Orbit Type



- Smaller flight numbers mostly failed
- LEO, ISS, PO, and GTO had lower flight numbers while the other types didn't have flights until later

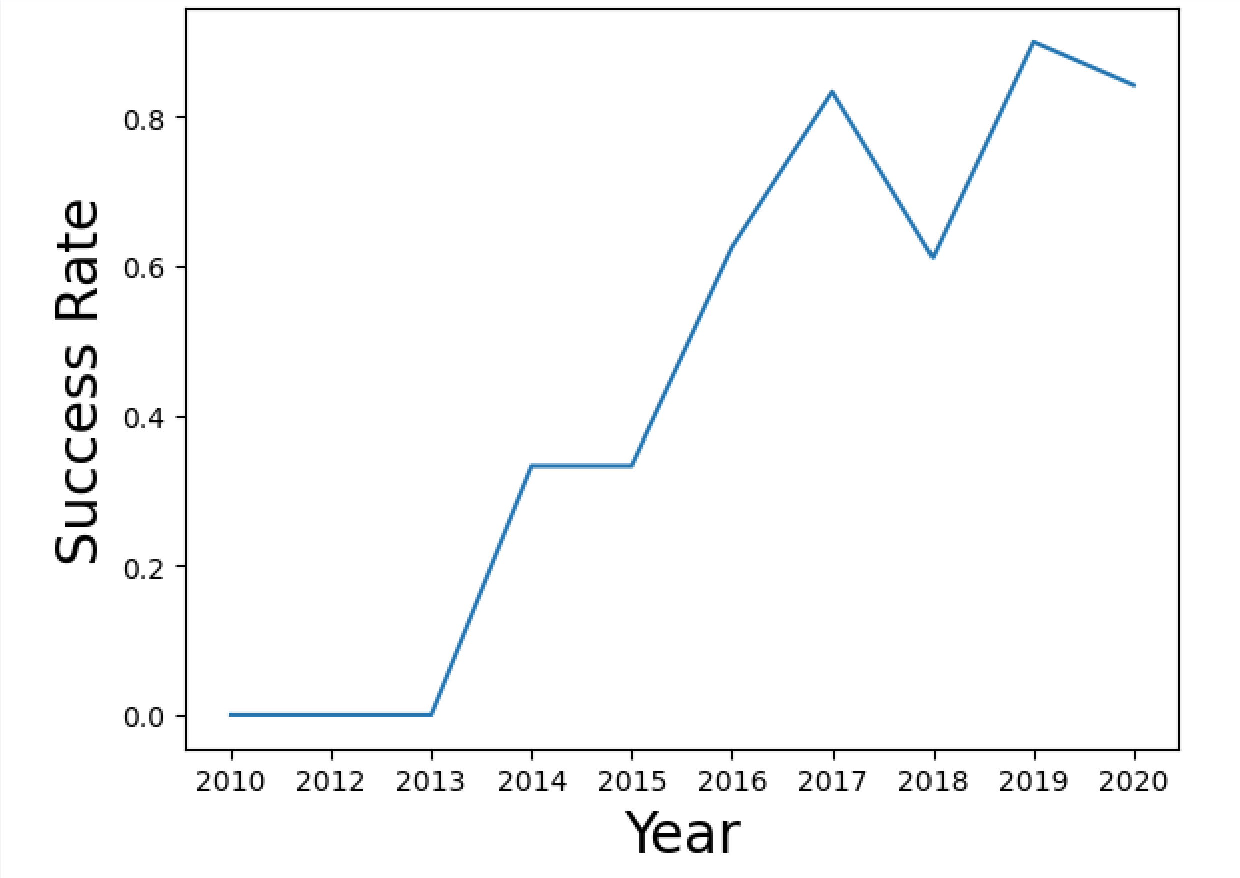
Payload vs. Orbit Type



- Most had a payload mass of below 8000 kg
- VLEO had a higher mass payload for their flights

Launch Success Yearly Trend

- Between 2013 and 2017 the success rate had a steady increase
- In 2018 there was a slight decline in success rate
- The success rate went back to increasing after 2019



All Launch Site Names

- Names of the unique launch sites
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

Total Payload Mass

```
In [14]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';

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

Out[14]:

total_payload_mass
45596

- Calculate the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [15]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

Out[15]:

average_payload_mass
2534.6666666666665

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

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [16]:

```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

Out[16]:

first_successful_landing
2015-12-22

- The first successful landing outcome on ground pad happened on 12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [17]:
```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4(
```

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

Out[17]:

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

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [18]:
```
%sql select Mission_Outcome, count(*) as total_number from SPACEXTABLE group by Mission_Outcome;
```

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

Out[18]:

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

- There was 1 in flight failure
- There were 99 successes
- There was 1 success where the payload status was unclear

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

In [19]:
```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

Out[19]:

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

- There were two failed landing_outcomes in drone ship in year 2015

In [22]:
```
%%sql select substr(Date, 6,2) as month, date, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTABLE
      where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

Out[22]:

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

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

In [23]:

```
%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
      where Date between '2010-06-04' and '2017-03-20'
      group by Landing_Outcome
      order by count_outcomes desc;
```

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

Out[23]:

Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- Launch sites are located in Florida and California
- The map is marked to show successes
- The launch sites are in close proximity

Colored Markers Show Success Rate

- The color markers identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site CCAFS SLC – 40 has a lower success rate.

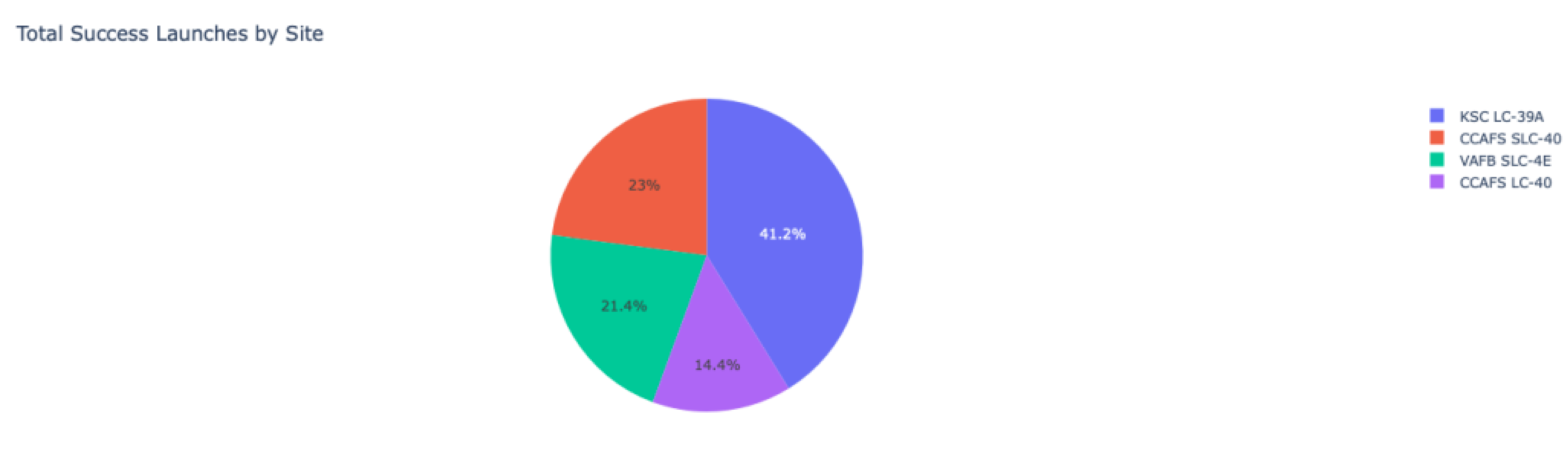
Distance From Launch Site

- The map shows it is 0.9 km from CCAFS SLC – 40 and the railway point

Section 4

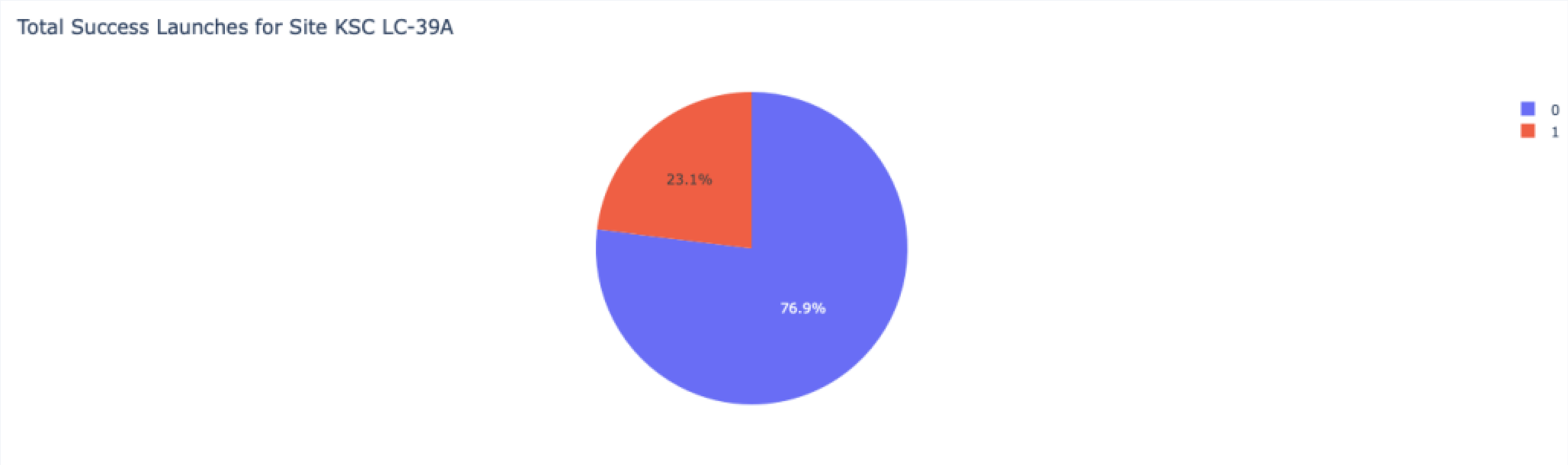
Build a Dashboard with Plotly Dash

Launch Success by Launch Site



- KSC LC – 39 A has the most successful launches

Launch Site with Highest Success Ratio



- KSC LC-39A has the highest launch success rate

Payload Mass and Launch Outcome by Site

- Most successes occur between 2000 and 5500 kg payload mass

Section 5

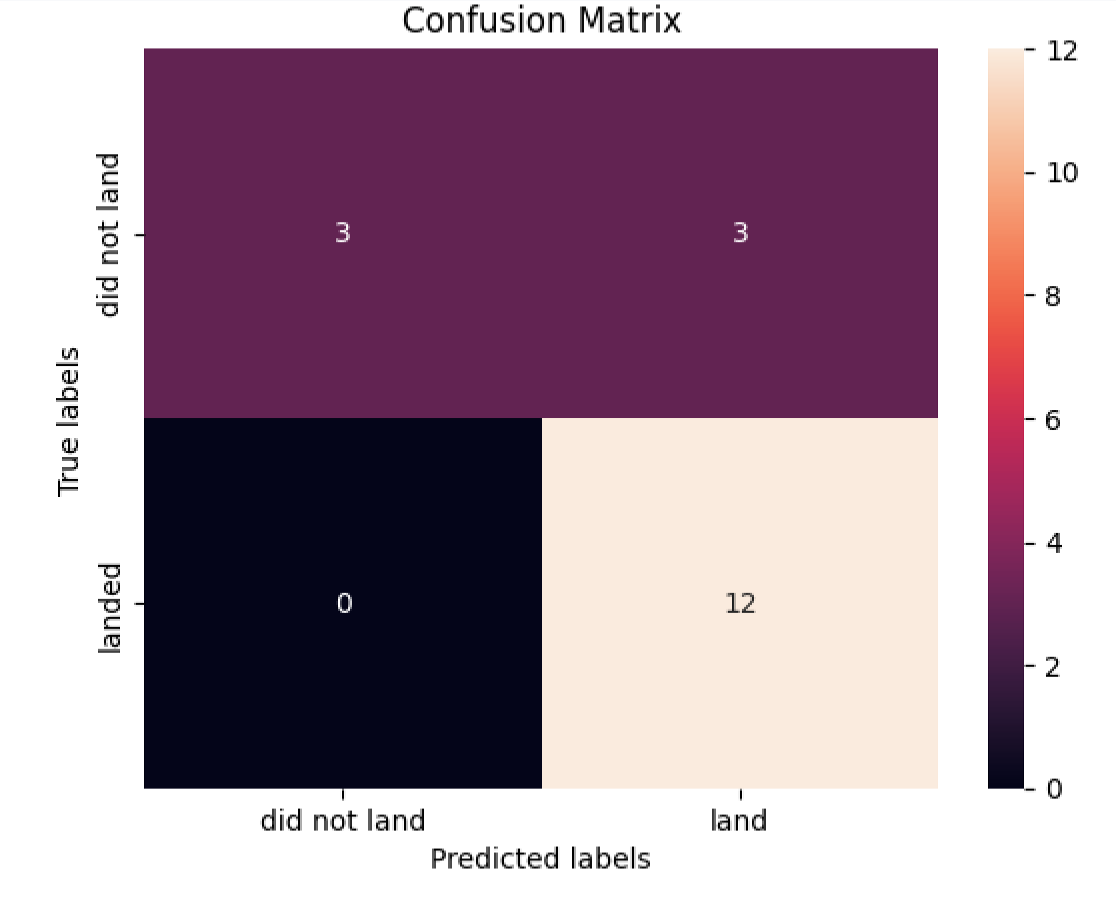
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Confusion Matrix

- The confusion matrix shows that true positives and true negatives seem to be okay but we need work in the area of false positives



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Overall GitHub site for all notebooks: https://github.com/16mlarsen5/IBM_Capstone/tree/main

- THANK YOU

Thank you!