IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

K. Y.
12/24/2023

Outline

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

Executive Summary

- Summary of methodologies

Collect data
- SpaceX API
- Web scarping

Prepare and explore data
- Create outcome variable
- Dummy code predictors
- Fill missing values
- Examine data patterns using data visualization
- Understand the geographic location and successful rate for launching sites

Analyze data
- Use multiple types of model to predict the outcome, including logistic regression, support vector machine, decision tree, and K-nearest neighbor
- Compare model performance and make recommendation

- Summary of all results
  - SpaceX improved its successful rate over time
  - Launching rockets with greater pay load mass are more likely to be a success
  - All launching sites are close to coastline
  - All models performed similarly on the test data, with slightly better performance for the decision tree model

Introduction

- Space X advertises Falcon 9 rocket launches with a cost of 62 million dollars. Other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage. If we can determine if the first stage will land, we can tell the cost of a launch. This information can be used to help other companies who want to bid against space X for a rocket launch. This purpose of the project is to use data science tools to obtain relevant public information and predict if the first stage will land successfully and Space X will reuse the first stage.
- The purpose of this project is to answer the following questions
  - How are pay load mass, launch site, number of flights, and orbits associated with launch success?
  - How did the launch successful rate change over time?
  - Which model provide the best prediction of the landing outcome?

Section 1

Methodology

Methodology

Collect Data
- Space REST API
- Web scraping

↓

Wrangle data
- Select variables and fill missing values
- One hot encoding

↓

Explore data
- EDA with SQL
- Data visualization

↓

Visualize data
- Folium
- Plotly Dash

↓

Build models
- Predicting landing outcome
- Classification models
- Tune and evaluate model

Data Collection

Obtain data through Space X API using the GET request → Decode response content as a Json and turn it into a Panda framework → Clean data, check missing values, and fill in missing values when needed → Performed web scraping from Wikipedia for Falcon 9 launch records using BeautifulSoup → Extracted launch records from html table, parse table elements, and save a Pandas data frame for next step

Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/taijimao/testrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- GitHub URL of the completed web scraping notebook: https://github.com/taijimao/testrepo/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Load data

- Load data
- check the header to confirm data are loaded correctly

Understand data

- Check % of missing values
- Understand data type
- Check launch sites, orbits, mission outcomes

Generate new variables

- Create a landing outcome label
- Save to a csv file

- GitHub URL of the completed data wrangling related notebooks: https://github.com/taijimao/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter plot

- Pay load mass vs Flight number
- Launch Site vs Flight number
- Launch Site vs Pay load mass

Bar chart

- Success rate for each orbit

Scatter plot

- Flight number vs Orbit
- Pay load mass vs Class

Line chart

- Launch success yearly trend

- GitHub URL of the completed EDA with data visualization notebook: https://github.com/taijimao/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Display name of unique launch site
- Display 5 records of launch sites that begin with the string “CCA”
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List 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 booster_versions which have carried the maximum payload mass using a subquery
- List records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
- Rank the count of landing outcomes (such as failure (drone ship) or Success (ground pad) between date 2010-06-04 and 2017-03-20 in descending order

- GitHub URL of the completed EDA with SQL notebook: https://github.com/taijimao/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for launch sites
  - NASA Space center
  - All launch sites
- Colored markers for launch outcomes
  - Successful (green) and unsuccessful (red) launches
- Distances between a launch site to proximities
  - Colored lines to show distance between a launch site and its proximity to the nearest coastline, highway, and city

- GitHub URL of the completed interactive map with Folium map: https://github.com/taijimao/testrepo/blob/main/lab_jupyter_launch_site_location%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Dropdown list with launch sites

↓

Pie chart showing the % of successful vs failed launches

↓

Slider for users to choose payload mass range

↓

Show a scatter plot between payload mass vs successful rate by booster version for site(s) chosen

- GitHub URL of the completed Plotly Dash lab: https://github.com/taijimao/testrepo/blob/main/jupyter_labs_interactive_data_dashboard.ipynb

Predictive Analysis (Classification)

1. Create Outcome from class column using NumPy
2. Standardize predictors X
3. Split data into training and test sets
4. Create a GridSearchCV object for parameter optimization with cv = 10
5. Apply GridSearchCV on different types of models: logistic regression, support vector machine, decision tree, K-Nearest Neighbor
6. Calculate accuracy on the test data using .score for all models
7. Evaluate model fit using the confusion matrix
8. Identify the best model using model performance measures

- GitHub URL of the completed predictive analysis lab: https://github.com/taijimao/testrepo/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

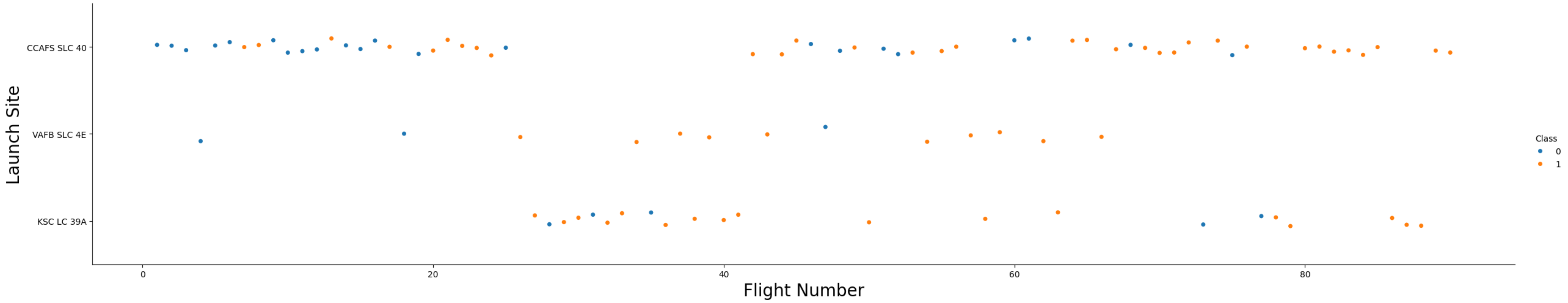
Results

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

Section 2

Insights drawn from EDA

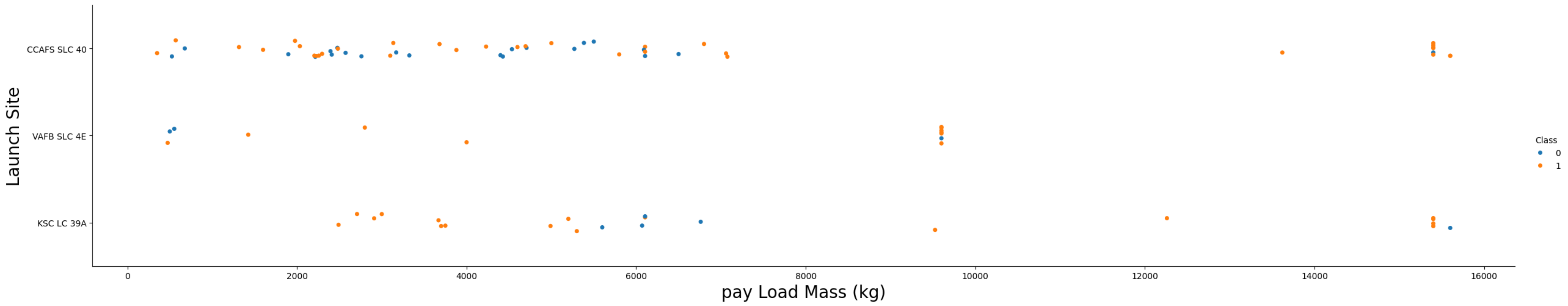
Flight Number vs. Launch Site



Major findings:

- Earlier launches had a lower success rate
- Success rate increased as SpaceX accumulate experience
- Site CCAFS SLC40 was used to launch about half of rockets
- Site VAFB SLC 4E and KSC LC 39A had higher success rates

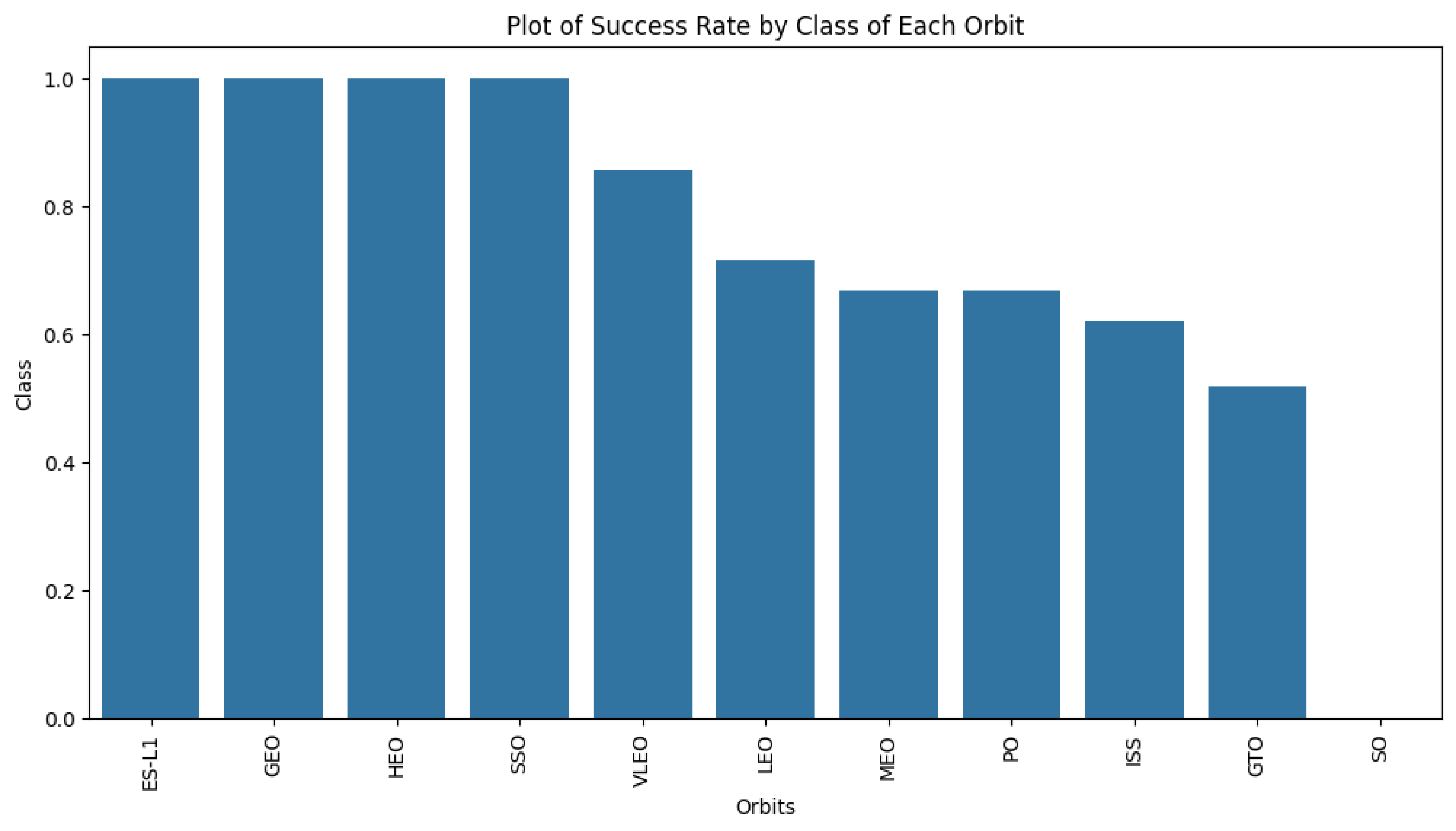
Payload vs. Launch Site



Major findings:

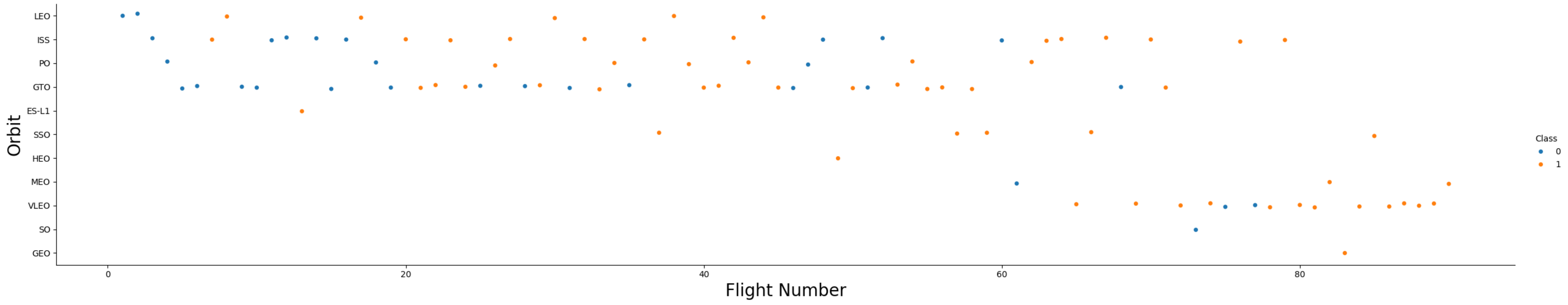
- Most launches with a pay load higher than 8,000 kg were successful
- Launches with a pay load higher than 8,000 kg had a higher successful rate than those lighter than 8,000 kg
- Site CCAFS SLC 40 has been used to launch most of the rockets lighter than 8,000 kg, whereas site VAFB SLC 4E has not been used to launch any rocket heavier than 10,000 kg

Success Rate vs. Orbit Type



- Four orbits had 100% successful rate: ES-L1, GEO, HEO, and SSO
- Six orbits had 50%-80% successful rate: VLEO, LEO, MEO, PO, ISS, and GTO
- SO had 0 successful rate

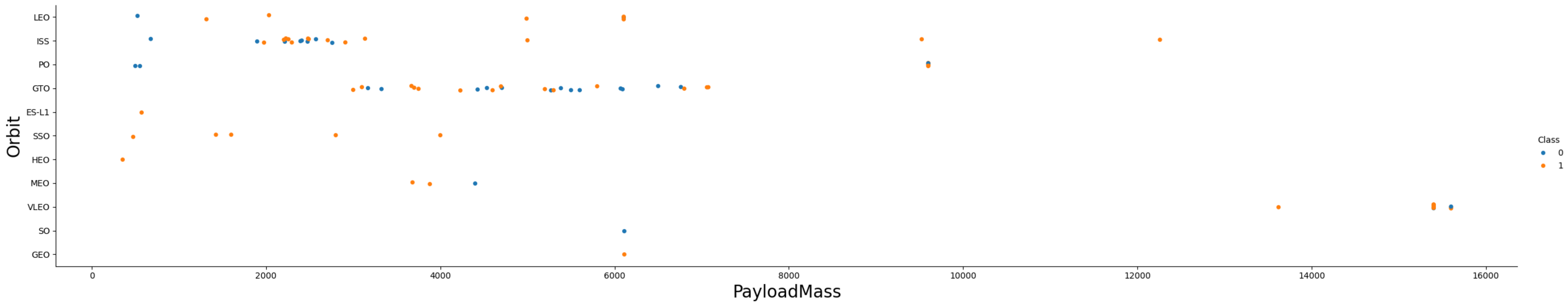
Flight Number vs. Orbit Type



Major findings:

- Overall, successful rate increased as flight number increased
- LEO followed this pattern closely while GTO did not
- Orbit ES-L1, SO, HEO, and GEO had only one launch. Careful when interpreting the successful rate for these orbits

Payload vs. Orbit Type



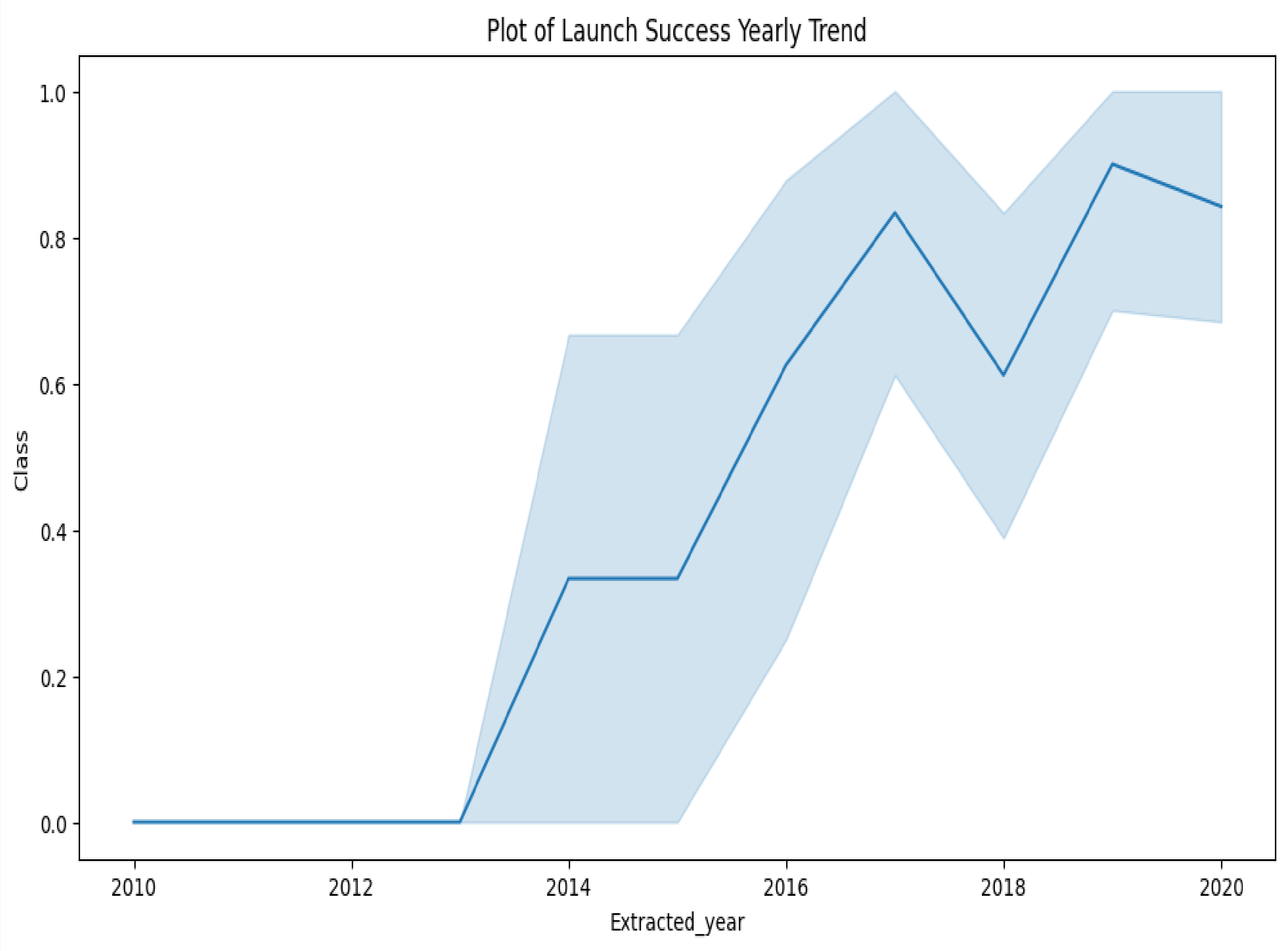
Major findings:

- Orbit ISS had a higher successful rate when launching rockets with pay load mass greater than 3,000 KG than launching rockets with pay load mass smaller than 3,000 KG
- Orbit GTO had mixed experience when launching rockets with varying pay load mass values

Launch Success Yearly Trend

Major findings:

- Successful rate remained at 0 before 2013
- Successful rate increased since 2013
- It dipped between 2017 and 2019
- Successful rate reached its highest point at 90% in 2019 and declined since then



All Launch Site Names

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

```
[7]  %sql select distinct "Launch_site" from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

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

Python

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

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

- The total pay load mass is 45,596 kg

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg

```
%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

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

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

min(DATE)
2015-12-22

- The dates of the first successful landing outcome on ground pad is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 and landing_outcome = 'Success (drone ship)'
```

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

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

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

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

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

count(MISSION_OUTCOME)
99

- The total number of successful and failure mission outcomes is 99

Boosters Carried Maximum Payload

12 boosters which have carried the maximum payload mass:

- 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

```
%sql select 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

```
%sql SELECT substring(Date, 6, 2), substr(Date, 0, 5), Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0 , 5) = '201
```

Python

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

substring(Date, 6, 2)	substr(Date, 0, 5)	Booster_Version	Launch_Site
01	2015	F9 v1.1 B1012	CCAFS LC-40
04	2015	F9 v1.1 B1015	CCAFS LC-40

- There are two records with failed landing_outcomes in drone ship in 2015. One with a booster version F9 v1.1 B1012 and the other with a booster version F9 v1.1 B1015. Both were launched from CCAFS LC-40.

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as Landing_Outcome_Count from SPACEXTBL where (Landing_Outcome like "Success%" OR Landing_Outcome like "Failure%") and (date
```

Python

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

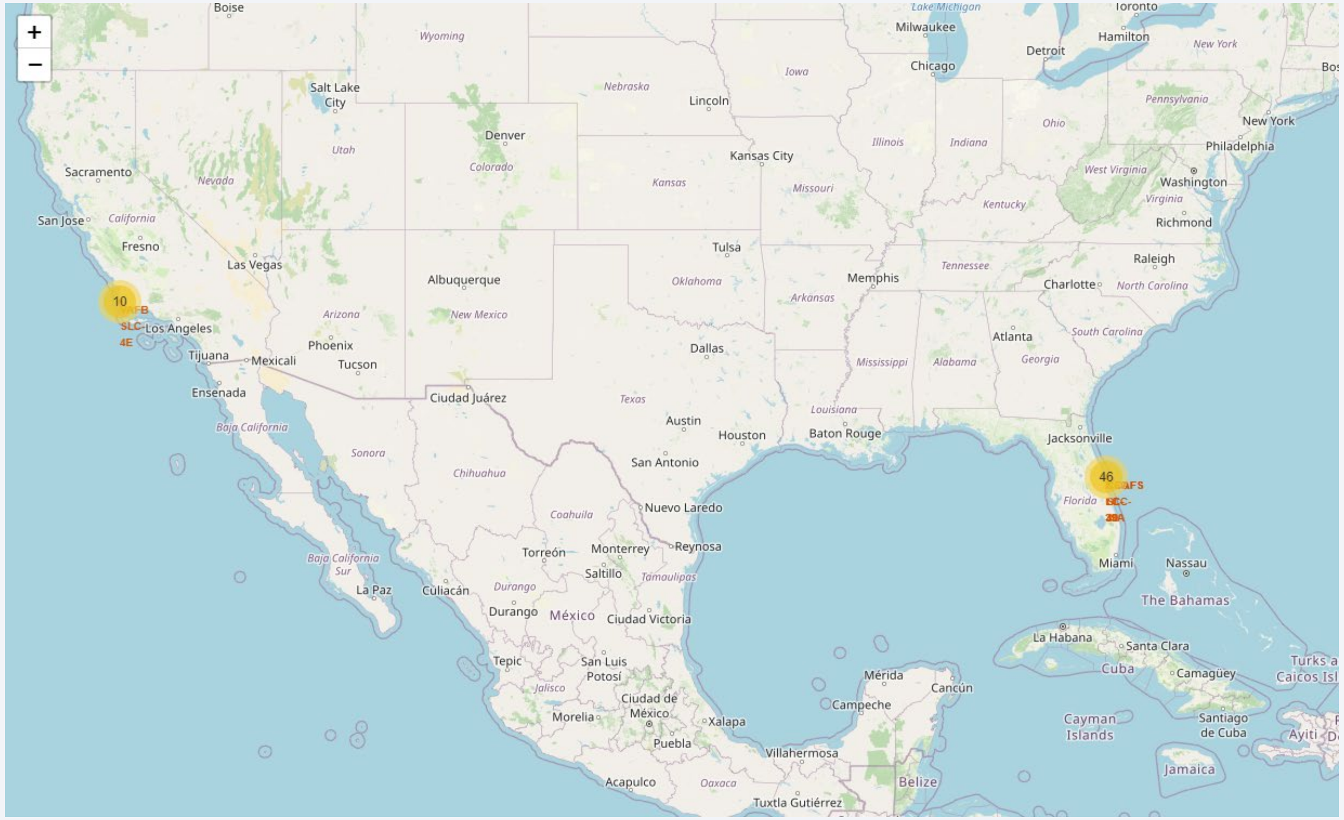
Landing_Outcome	Landing_Outcome_Count
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Failure (parachute)	2

- Based on the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, SpaceX had 5 success in drone ship, 3 success in ground pad, 5 failure in drone ship, and 2 failure in parachute

Section 3

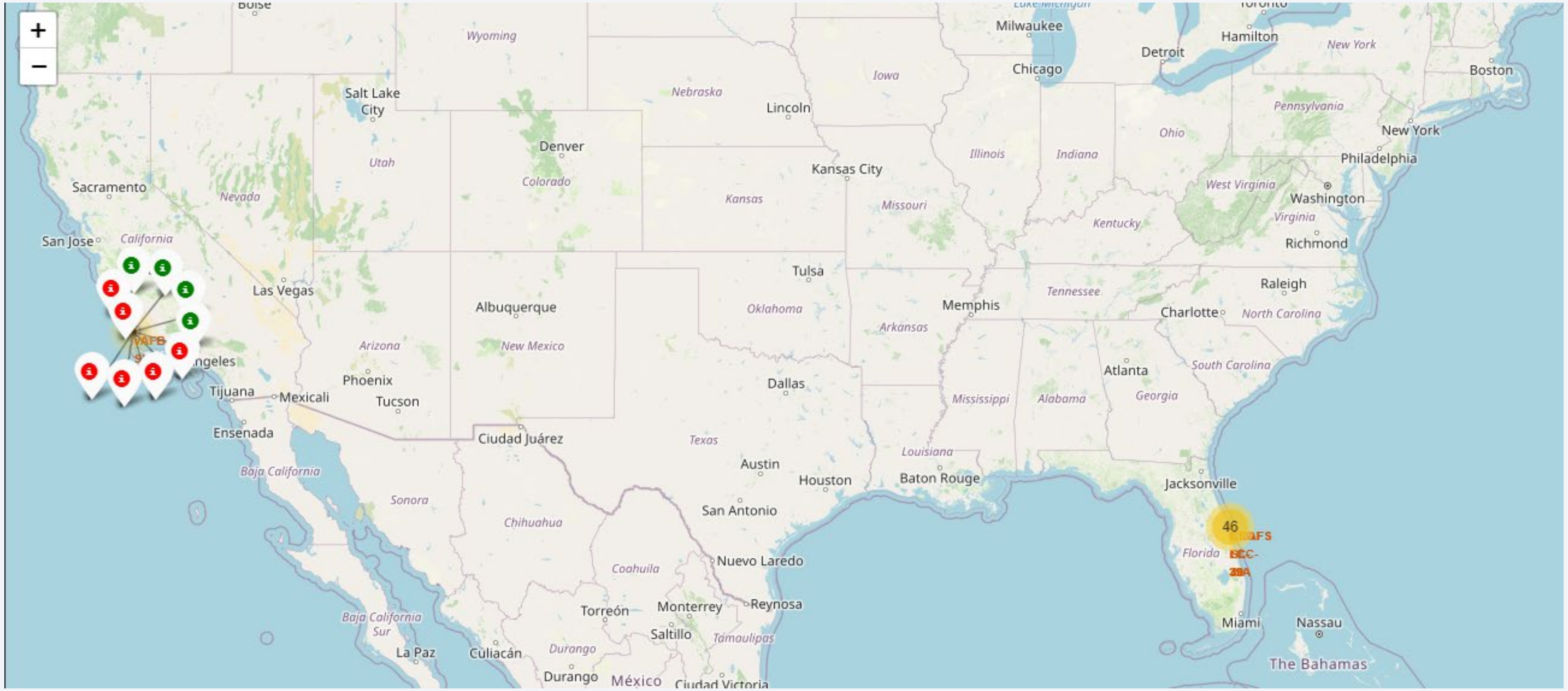
Launch Sites Proximities Analysis

Overview of launch sites



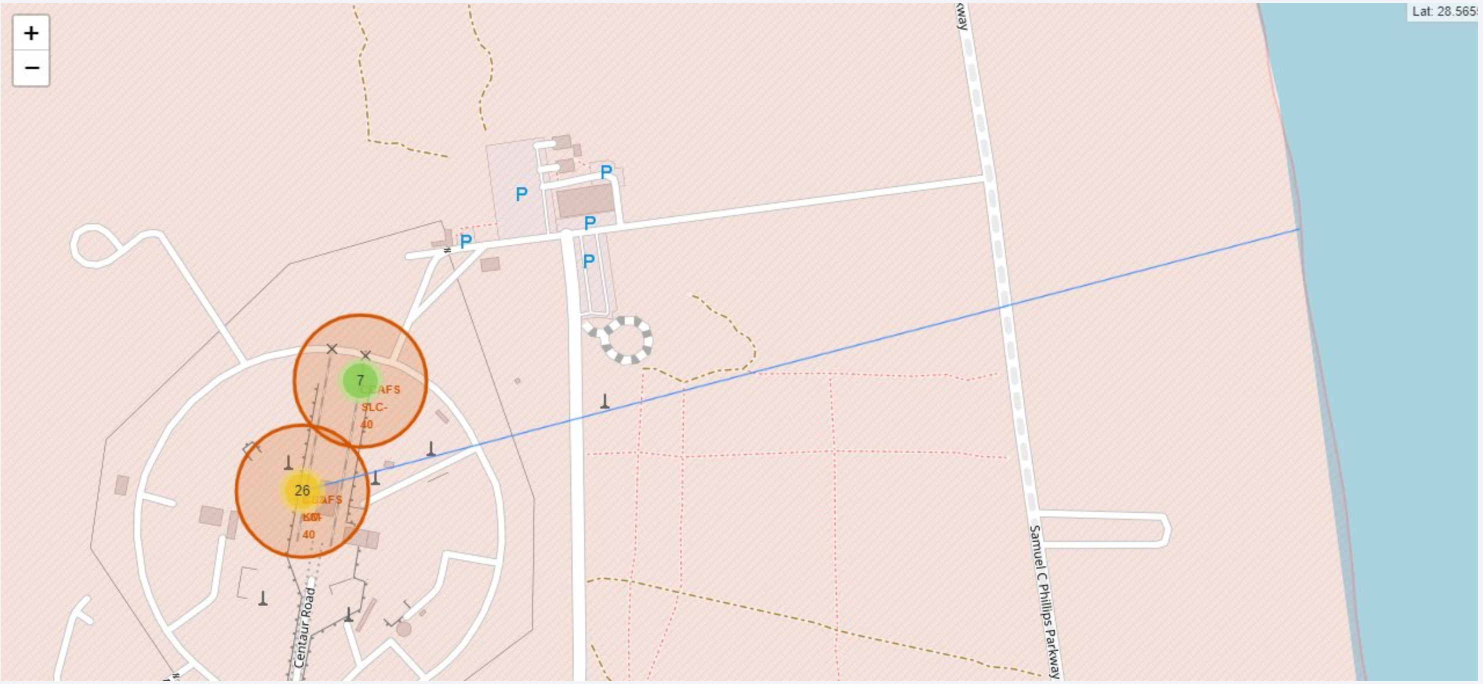
- the generated folium map shows there are two cluster of launch sites for Space X. One on the west coast in CA and the other on the east coast in Florida. Both of them are close to the coast line.

Launch outcomes for the site in CA



- Launch sites in CA had 6 failures (red) and 4 success (green)

Distance to coastline

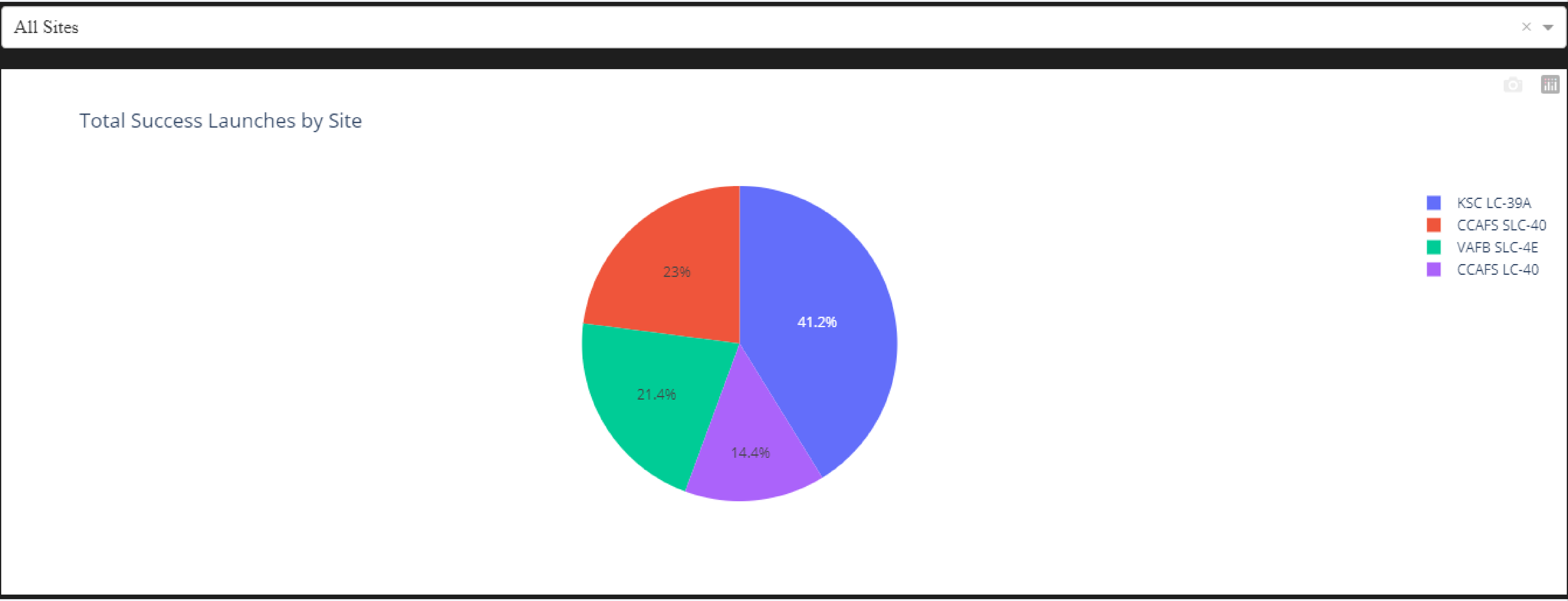


- Launch sites are close to coastline
- This helps minimize any potential impact on people and property
- It may also make it easier to use sea water to cool down the launching pad afterwards

Section 4

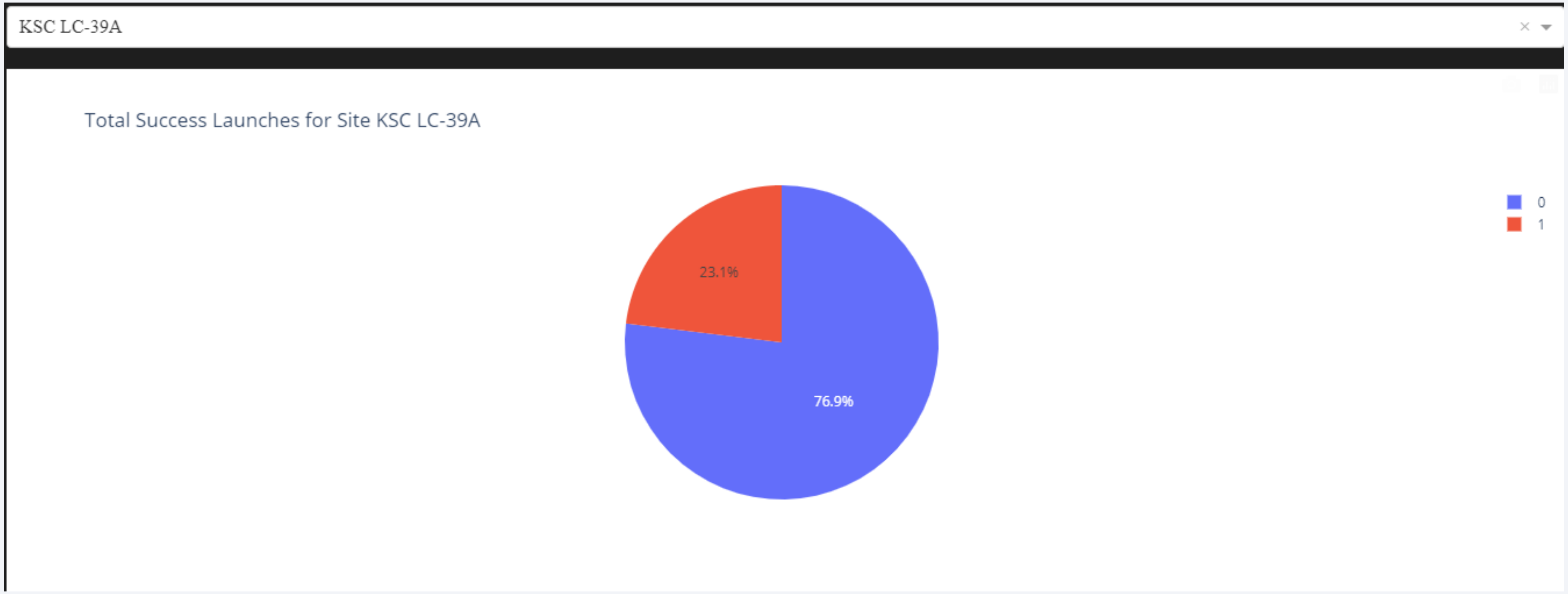
Build a Dashboard with Plotly Dash

Successful launch by site



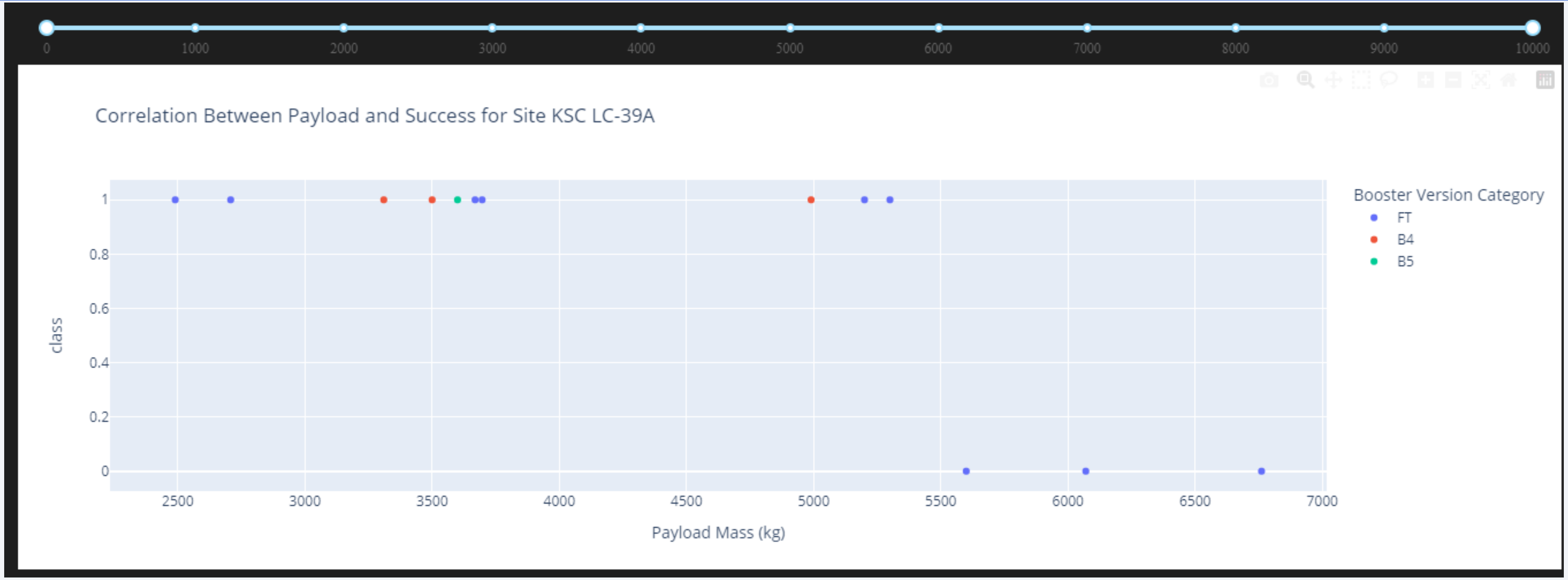
- Four launching sites varied in their share of successful launches
- 41.2% of successful launches were launched from KSC LC-39A

Launch success for KSC LC-39A



- KCS LC-39A's successful rate is 76.9%

<Dashboard Screenshot 3>

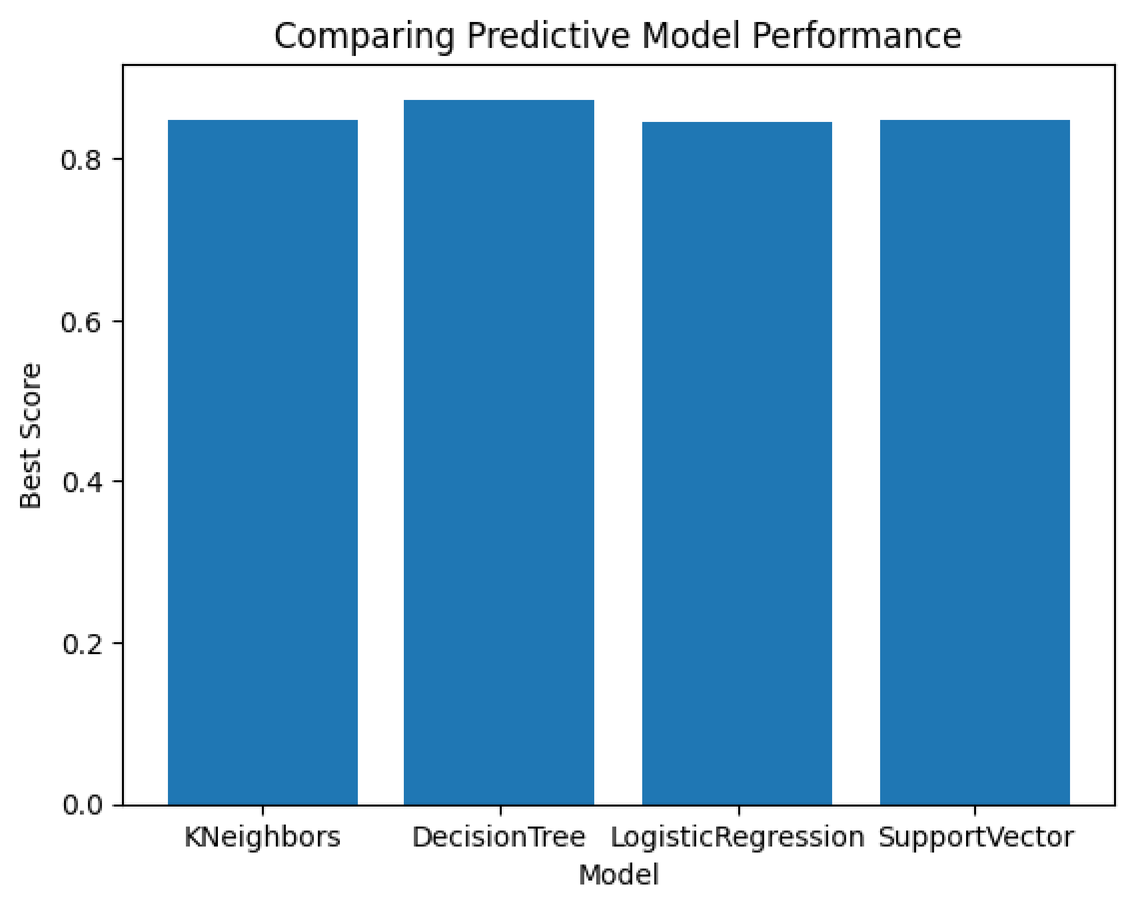


- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

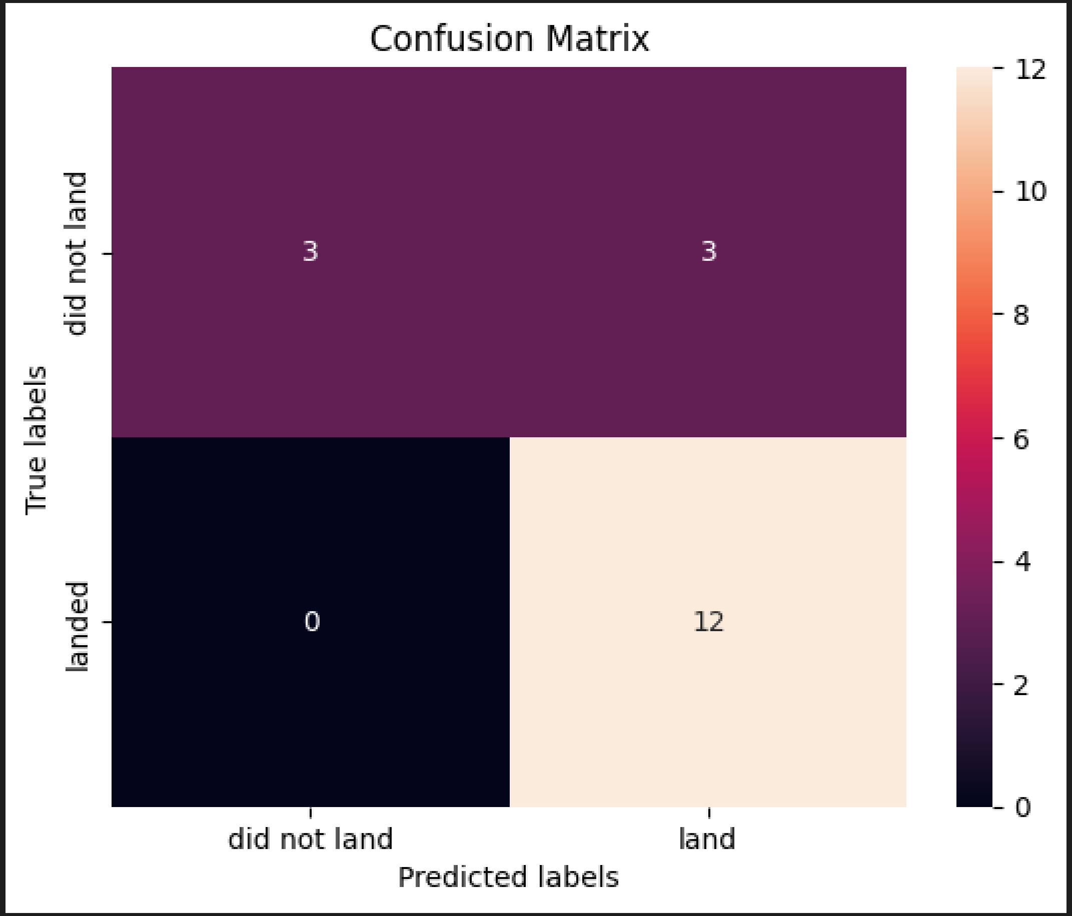


- Different types of models had similar performance
- The decision tree model performed slightly better than other types of models

Confusion Matrix

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

✓ 0.1s



- With the predictors we had in this exercise, the decision tree model predicted all 12 successful landing outcomes and 3 failure landing outcomes. It predicted incorrectly for three failed landing outcomes.

Conclusions

- Different types of models performed similarly on the test data. The decision tree model performed slightly better than other types of models
- All launch sites are close to coastline, which minimize potential impact on people and property and adds convenience to use natural resources to cool down the launching pad
- Space X increase its successful rate over time
- Pay load mass is positively related to successful rate, but sites vary
- Launch sites and orbits vary in the number of launches observed. It is important to continue monitoring and understand the pattern of successful launch of rockets by launching site, orbit, and pay load mass

Thank you!