IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chris McAlpine
February 18, 2022

Outline

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

Executive Summary

Methodology

- Data was collected through the use of API and web scraping
- Exploratory Data Analysis (EDA) was done using SQL and visualizations
  - Includes interactive visual analytics and predictive analytics via classification models
- Variable relationships were identified, and the overall data analysis provides a strong picture of what contributes to the most successful launches

Results

- The most successful launches are those that take place at site KSC LC-39A, with a payload mass between 2,000-4,000 kg, with recovery by drone ship

Introduction

Project background and context

- The goal of this project is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully
- SpaceX can save upwards of $100MM on a rocket launch relative to competition if it can reuse the first stage
- By determining if the first stage will land or not, we can figure out the cost of a launch
- This info can be used if another company wants to bid against SpaceX for a launch

Problems you want to find answers for

- What factors (operating conditions, orbit, payload size, etc.) allow for a successful landing?
- How do the relationships of different factors influence the landing success rate?

Section 1

Methodology

Methodology

Data collection methodology:

- Data was collected by using **get requests** to the SpaceX API
- Web scraping was done with **BeautifulSoup** to collect Falcon 9 historical launch records with the goal of extracting a HTML table, then parsing that and converting to a Pandas dataframe

Perform data wrangling

- Determined the label for trained supervised models by converting mission outcomes into training labels

Perform EDA using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Broke dataset into test/train splits, then ran models (LR, SVM, Decision Tree, kNN) to determine best parameter sets for each model, and the accuracy of each

Data Collection

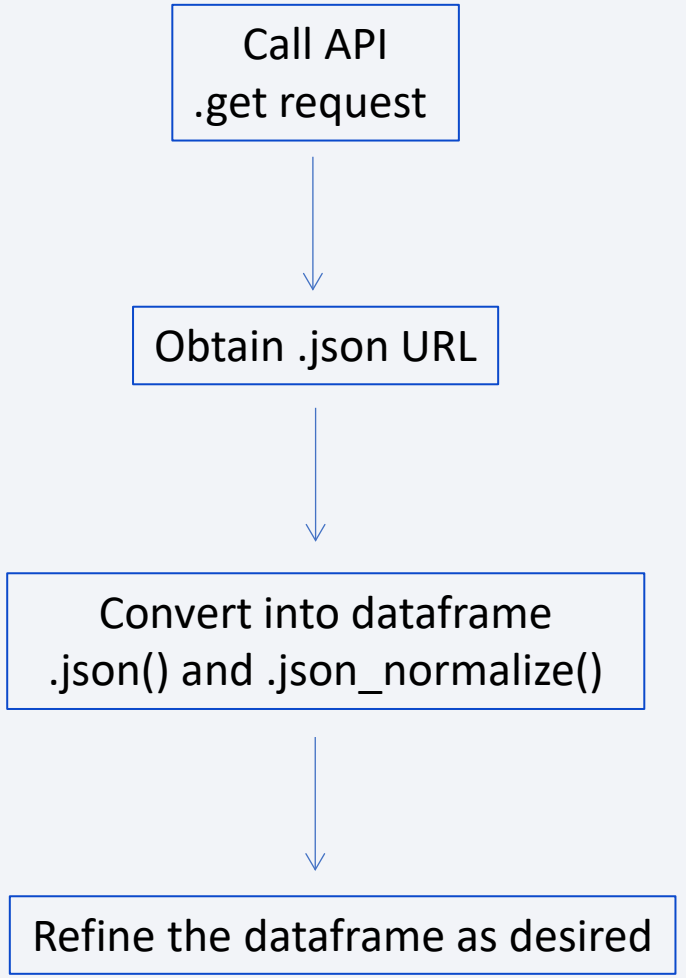
The data collection process went as follows:

- Defined a series of helper functions to extract API info using IDs in the launch data
- Identified correct URL, then issued the .**get()** request
- Used **.json()** and **.json_normalize()** to turn contents into a dataframe
- Took subset of desired dataframe features and removed unnecessary rows
- Cleaned up dataset by filling in missing payload values using **.mean()**
- Dataset exported to CSV file (**.to_csv**)for further use
- Also used BeautifulSoup to webscrape historical Falcon 9 launch records
    - Extracted HTML table, cleaned it, then converted it to a Pandas dataframe

GitHub link to notebook

Data Collection – SpaceX API

- The get request was used to the SpaceX API
- Data wrangling and cleaning followed to produce a useful dataframe




```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

Check the content of the response

```
print(response.content)
```

```
b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"link
s":{"patch":{"small":"https://images2.imgbox.com/3c/0e/T8iJcSN3_o.png","large":"http
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-D
S0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

Now we decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()

```
# Use json_normalize meethod to convert the json result into a dataframe
response.json()

data = pd.json_normalize(response.json())
```

Using the dataframe data print the first 5 rows

```
# Get the head of the dataframe
data.head()
```

```
data_falcon9 = data[data['BoosterVersion']!='Falcon 1']
```

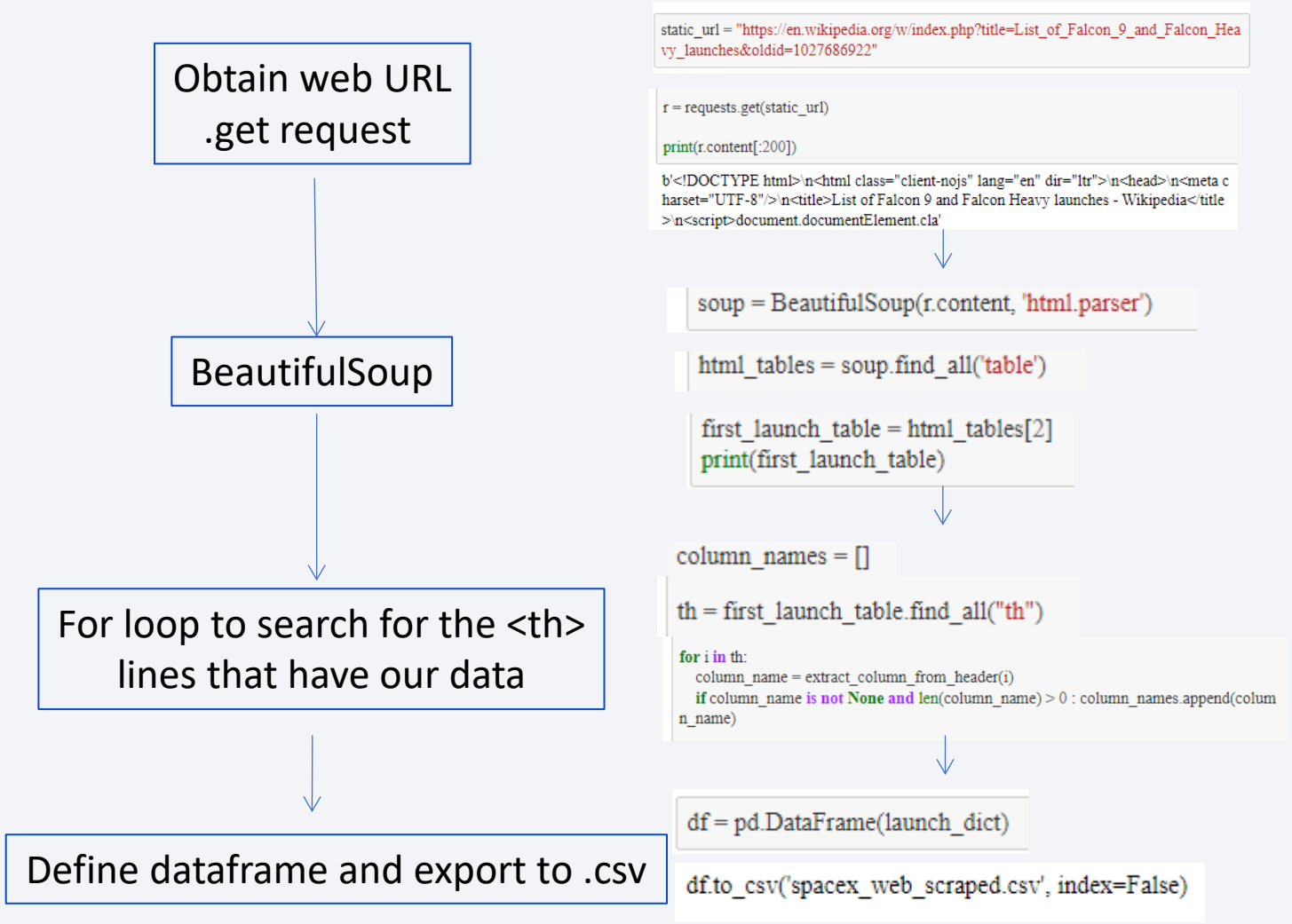
```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

GitHub link to notebook

Data Collection - Scraping

- Web scraping was done with BeautifulSoup to collect Falcon 9 historical launch records from a Wikipedia page

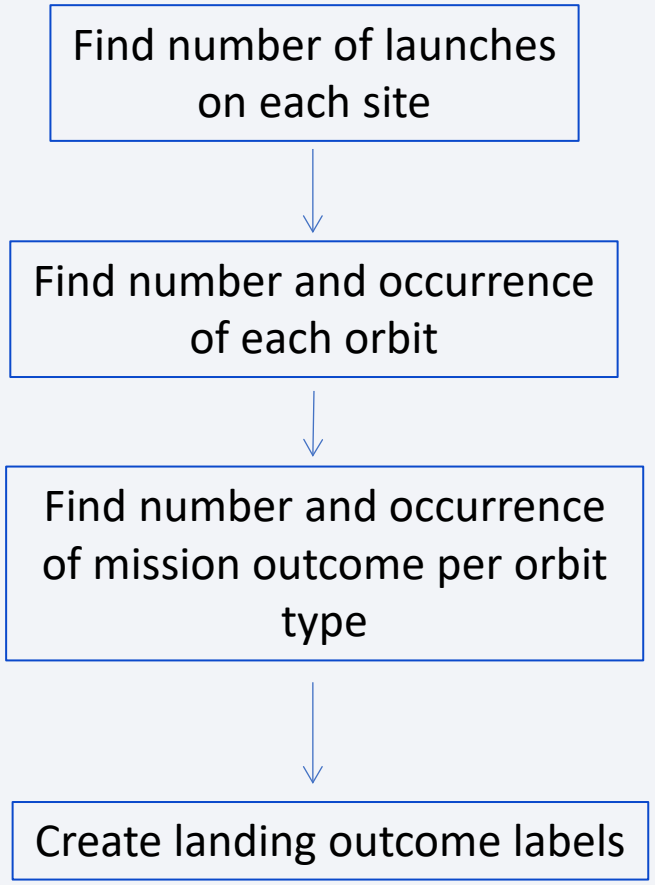
- The goal was to extract a Falcon 9 launch records HTML table, then parse the table and convert it to a Pandas dataframe



GitHub link to notebook

Data Wrangling

- EDA was performed to determine the label for trained supervised models
- Converted mission outcomes into training labels where 1 was a successful booster landing and 0 was unsuccessful
- Exported results to a .csv file




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

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

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

```
conditions = [
    (df['Outcome'] == 'None None'),
    (df['Outcome'] == 'False Ocean'),
    (df['Outcome'] == 'False RTLS'),
    (df['Outcome'] == 'False ASDS'),
    (df['Outcome'] == 'None ASDS')
]

choices = [0,0,0,0,0]

df['landing_class'] = np.select(conditions, choices, default = 1)

df.head(15)
```

GitHub link to notebook

EDA with Data Visualization

Multiple plots were generated to assess variable relationships (results in Section 2):

- Scatter plots
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
- Column Chart
  - Success Rate vs. Orbit Type
- Line Plot
  - Success Rate vs. Time

GitHub link to notebook

EDA with SQL

The following SQL queries allow for more insight into the data (results in Section 2):

- All launch site names, including a specific query for sites beginning with ‘CCA’
- The total payload mass, and the average payload mass for the F9 v1.1 booster
- The date of the first successful ground landing
- The successful drone ship landings when the payload is between 4K-6K kg
- The total number of successful mission outcomes
- The boosters that carry the maximum payload
- Launch records for 2015, and ranked landing outcomes between 6/4/2010 and 3/20/2017

GitHub link to notebook

Build an Interactive Map with Folium

Folium, a Python package, was used to generate an interactive map, allowing us to:

- See the launch locations for all the SpaceX launch sites
  - Identified as a circle
- See the success/failure launch outcomes at each site, as well as the quantity of launches at each site
  - Launches identified by either a green marker (Success) or a red marker (Failure)
  - Quantity identified as a numbers within the location circle
- Determine the distance from the launch sites to potential hazards (coastline, major highway, railroad, etc.)
  - Distance shown as a blue line

Maps can be found in Section 3

GitHub link to notebook

Build a Dashboard with Plotly Dash

Plotly Dash, a python framework, was used to create a dashboard that allows us to:

- Add plots/graphs that help summarize the data
- View the distribution of the total successful launches w.r.t. each of the four launch sites
- View the success and failure launch rates at each of the four sites individually
- Compare the payload mass against the launch outcome for all sites

- The plots are useful because a visual representation can oftentimes show a clear comparison or outcome from a lot of complex data

GitHub link to notebook

Predictive Analysis (Classification)

- The process shown below is the same for each model (LR, SVM, Decision Tree, kNN)

Standardize the data

```
Y = data['Class'].to_numpy()
Y
```

```
array([0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1])
```

```
transform = preprocessing.StandardScaler()
X = np.array(transform.fit(X).transform(X))
X[0:5]
```

Split data into train/test

```
in, X_test, Y_train, Y_test
```

```
n, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2, random_state=2)
Train set:', X_train.shape, Y_train.shape)
Test set:', X_test.shape, Y_test.shape)
```

```
set: (72, 83) (72,)
t: (18, 83) (18,)
```

Use GridSearchCV to determine optimal parameters to use

```
neters ={'C':[0.01,0.1,1],
    'penalty':['l2'],
    'solver':['lbfgs']}
```

```
istic Regression using GridSearchCV
neters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge

g = LogisticRegression()
g_cv = GridSearchCV(logreg, parameters, cv = 10)
g_cv.fit(X_train, Y_train)
```

Run model using best parameters and training data → Run model using test data → Find model accuracy Plot confusion matrix

Best LR Parameters to Use and Accuracy on Validation Data

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

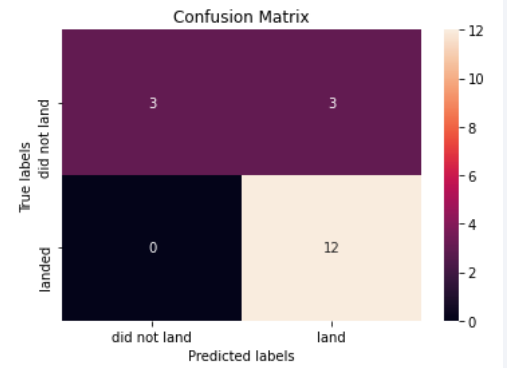
Accuracy

```
logreg2 = LogisticRegression(C = 0.01, penalty = 'l2', solver = 'lbfgs')
logreg2.fit(X_train, Y_train)

print("score", logreg2.score(X_test, Y_test))
```

```
score 0.8333333333333334
```

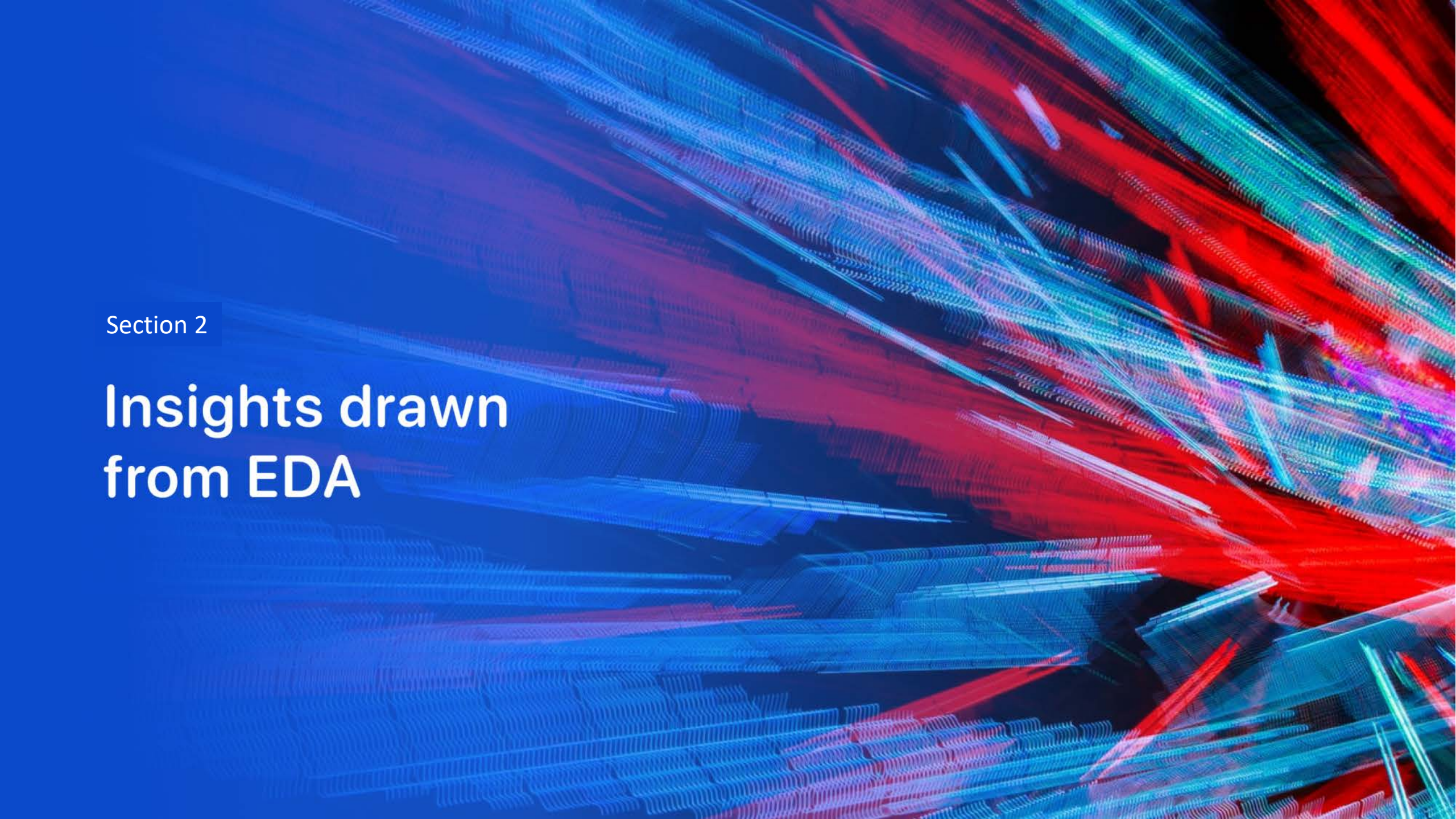
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



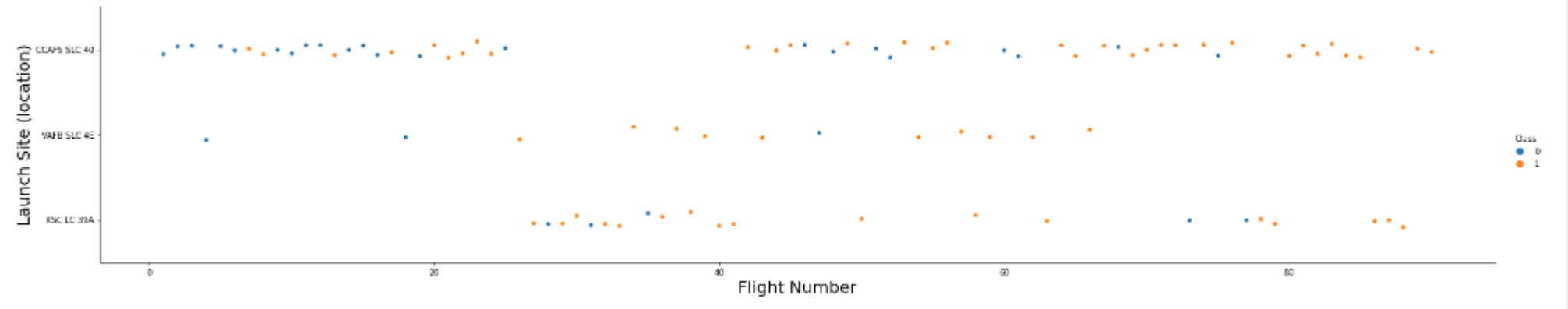
Results

Some of the important results from analysis performed (detailed in Sections 2-5) are:

- Orbits ES-L1, GEO, HEO, and SSO have had 100% launch success so far
- The annual trend of successful launches has increased almost every year, and the overall trend is consistently positive since testing began
- Recovery via drone ship is the most successful landing outcome observed from the provided timeline
- Test site KSC LC-39 has the highest success rate of the four sites
- Lighter payload masses are more successful, especially in the 2,000-4,000 kg range

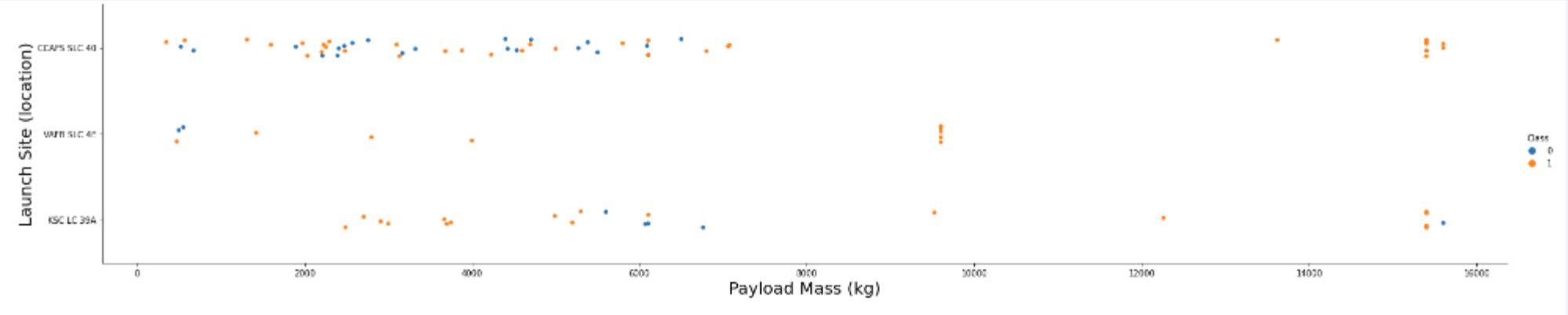

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- Nearly all of the initial 25 tests were from launch site CCAFS SLC 40 (92%)
- 64% of the initial tests failed
- 75% of the next 16 tests were from launch site KSC LC 39A
- No tests were run at site VAFB LSC 4E after launch 66

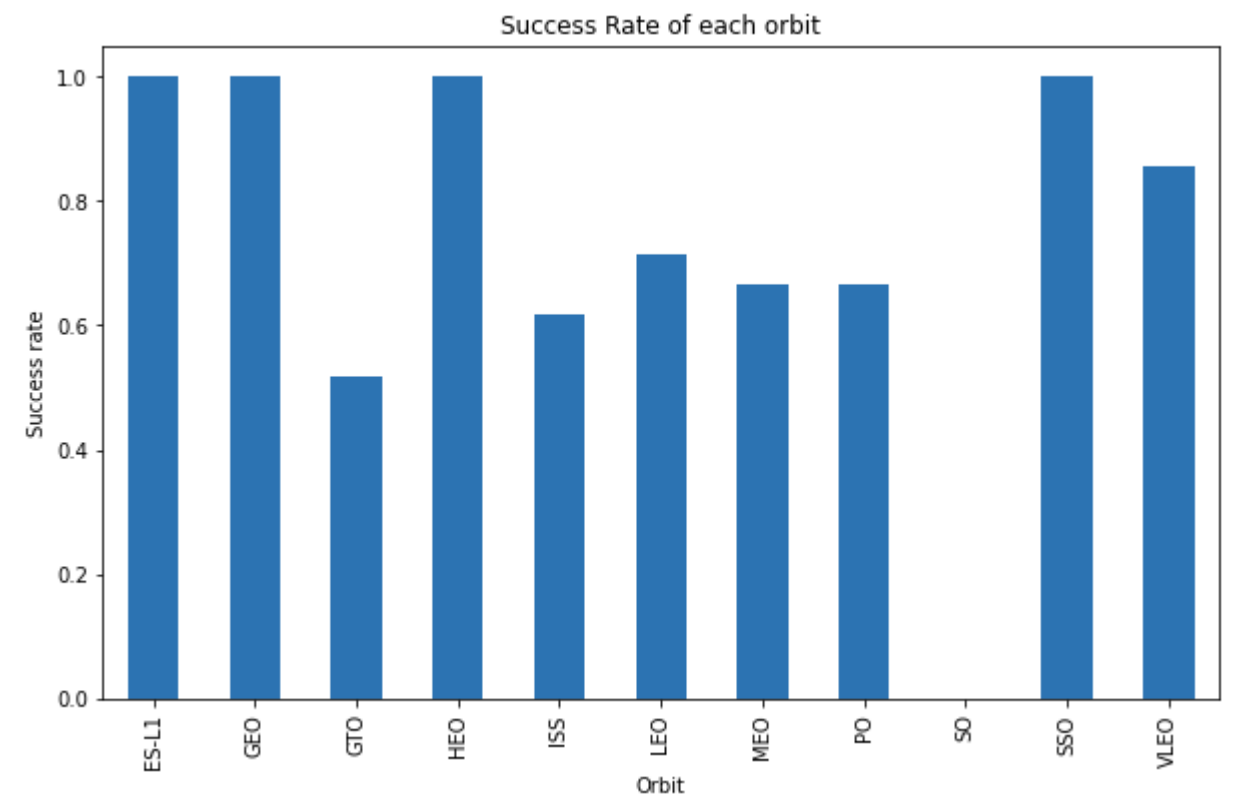
Payload vs. Launch Site



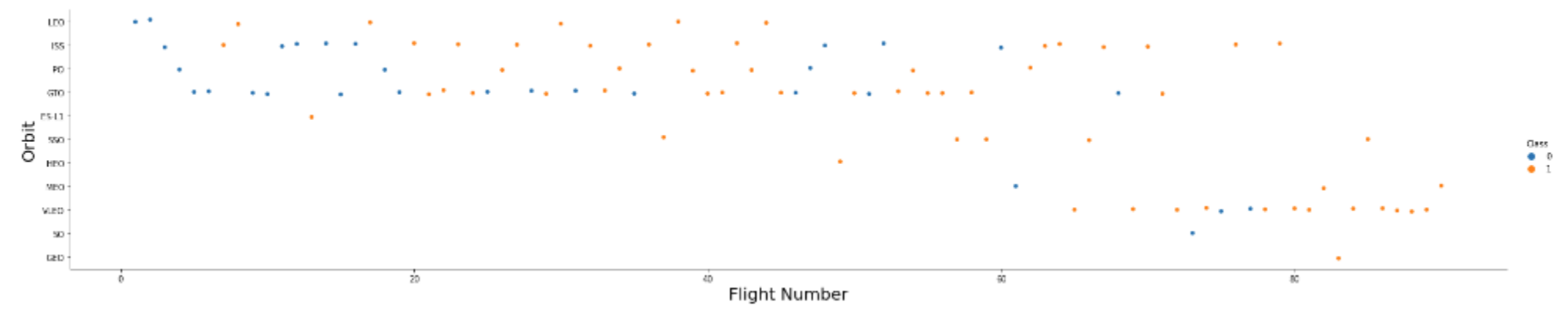
- Almost all of the tests where the payload mass was greater than 8,000 kg were successful
- Clustering of tests occur at specific payload masses over 8,000 kg, indicating repeated tests took place at specific sites with specific weights

Success Rate vs. Orbit Type

- 100% of tests were successful in orbits ES-L1, GEO, HEO, and SSO
  - Note that the quantity of tests in each specific orbit is not taken into account

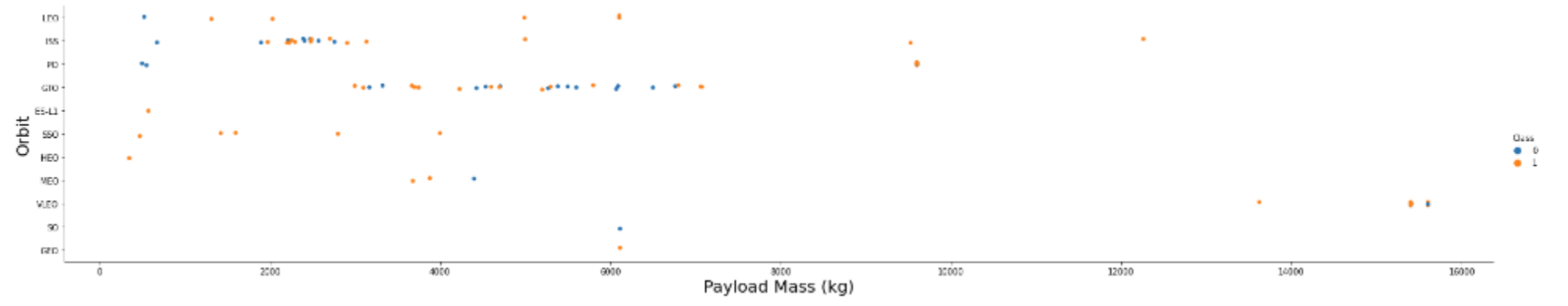


Flight Number vs. Orbit Type



- A majority of the later flights were in orbit VLEO
- A larger percentage of later flights succeeded, implying that improvements were made as flight testing continued on

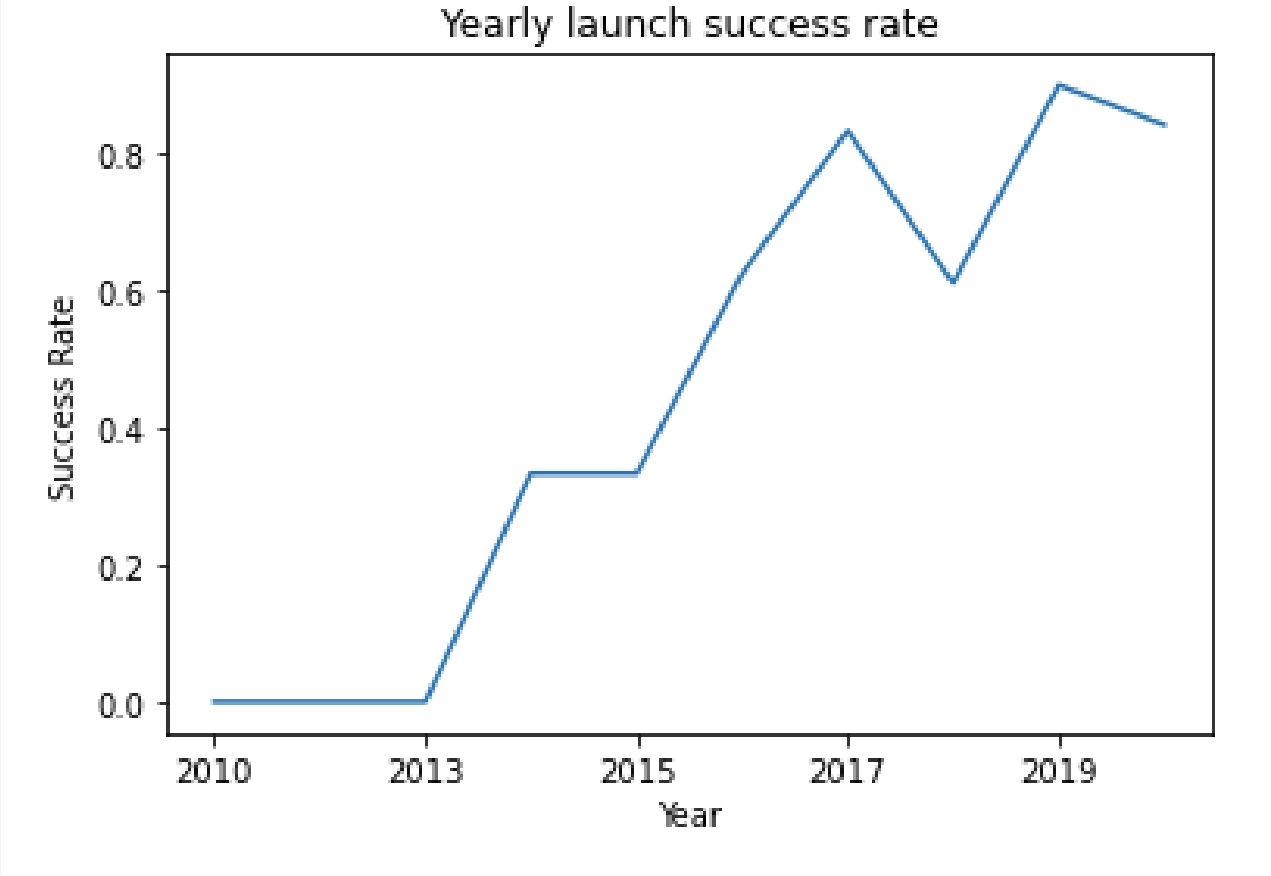
Payload vs. Orbit Type



- Heavier payload masses had a negative impact on success in GTO orbit
- Heavier payload masses had a positive impact on success in ISS orbit

Launch Success Yearly Trend

- On average, annual success rates have increased, from 0% in 2013 to ~80% in 2020
- Launces in 2018 were ~20% less successful than previous year. Could look for factors that explain this change.



All Launch Site Names

```
%sql SELECT DISTINCT launch_site FROM spacextbl
```

- Selected the **DISTINCT** names under the launch_site header found in table spacextbl

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

Launch Site Names Beginning with 'CCA'

```
%sql SELECT * FROM spacextbl WHERE launch_site like 'CCA%' LIMIT 5
```

- Found first 5 records where launch sites begin with `CCA` by selecting rows in table spacextbl where the sites under header launch_site began with the letters CCA

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

```
%sql SELECT sum(payload_mass__kg_) as "Total payload mass (kg) from NASA (CRS)" FROM spacextbl WHERE customer = 'NASA (CRS)'
```

- The total payload carried by boosters from NASA is 45,596 kg
- Result obtained by summing up the numbers in column “payload_mass_kg_” that had NASA (CRS) listed as the unique customer

Total payload mass (kg) from NASA (CRS)
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(payload_mass__kg_) as "Avg. payload mass (kg) from booster version F9 v1.1" FROM spacextbl WHERE booster_version = 'F9 v1.1'
```

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg
- Number obtained by averaging all the payload mass values from the table for entries where the booster version was F9 v1.1

Avg. payload mass (kg) from booster version F9 v1.1
2928

First Successful Ground Landing Date

```
%sql SELECT min(DATE) as "Date of first successful landing outcome in ground pad" FROM spacextbl WHERE landing__outcome = 'Success (ground pad)'
```

- December 22, 2015 was the date of the first successful landing outcome on ground pad
- Number obtained by selecting the MIN date (aka the earliest date) from the table where the landing outcome was a success on ground pad

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

```
%sql SELECT booster_version FROM spacextbl WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ between 4000 and 6000
```

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are shown in the above table
- Booster names obtained by querying the table to find when the landing outcome was a success (drone ship) and the payload mass was between 4,000kg and 6,000kg

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(*) as Count FROM spacextbl GROUP BY mission_outcome
```

- The total number of successful and failure mission outcomes are shown in the table above
- 100 of the 101 missions were a success (99%)
- Numbers obtained by counting the mission outcomes from the table and grouping by the specific types of outcome

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

Boosters Carrying Maximum Payload

```
%sql SELECT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) from spacextbl)
```

- The F9 B5 booster has carried the max payload mass
- A subquery was used in SQL to produce the result, where the booster that had the max payload value was selected from the table

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 booster_version, launch_site, landing__outcome, DATE FROM spacextbl WHERE landing__outcome = 'Failure (drone ship)' AND year(DATE) = '2015'
```

- The failed landing outcomes via drone ship, their booster versions, and launch site names for the year 2015 are shown in the table

- Selected the three specific columns where the landing outcome failed (drone ship) and the year was 2015

booster_version	launch_site	landing__outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

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

```
%sql SELECT landing__outcome, COUNT(*) as count FROM spacextbl WHERE DATE between '2010-06-04' and '2017-03-20' GROUP BY landing__outcome ORDER BY count desc
```

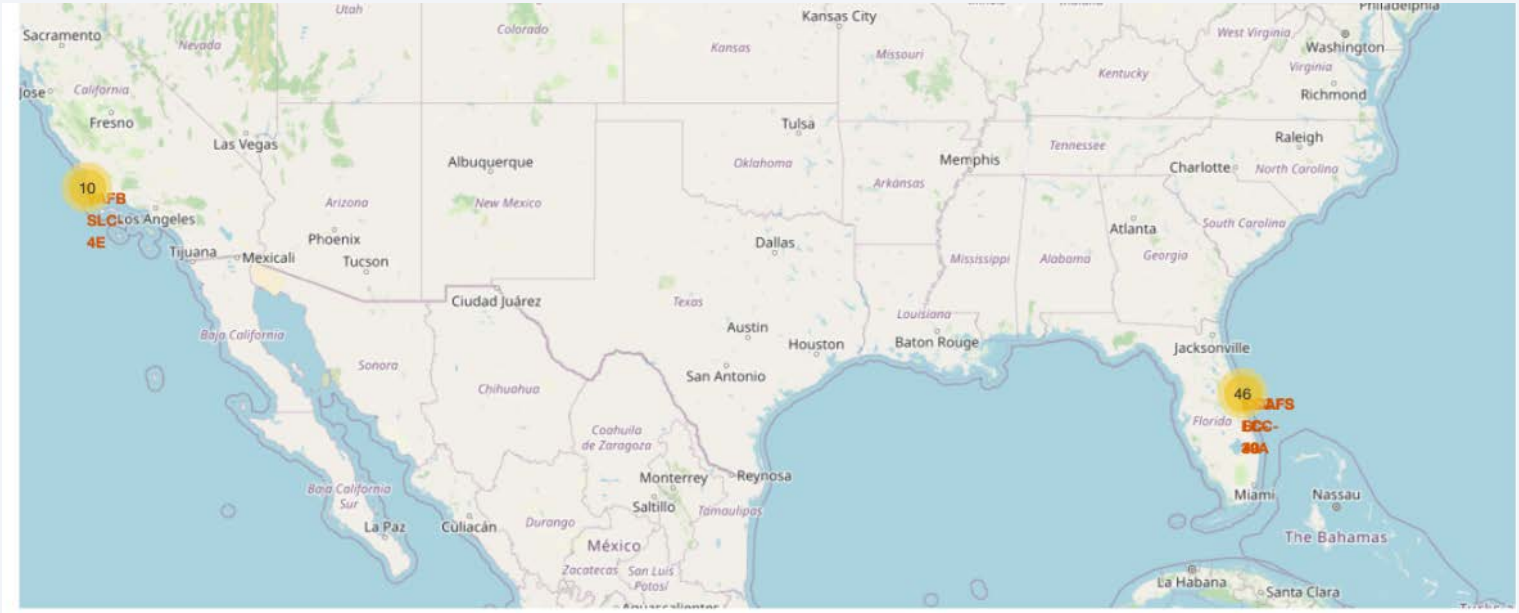
- The count of landing outcomes 2010-06-04 and 2017-03-20 are shown in the table in descending order
- Used a SQL query containing **COUNT(*)**, **WHERE**, **GROUP BY** and **ORDER BY** to achieve desired table

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

Section 3

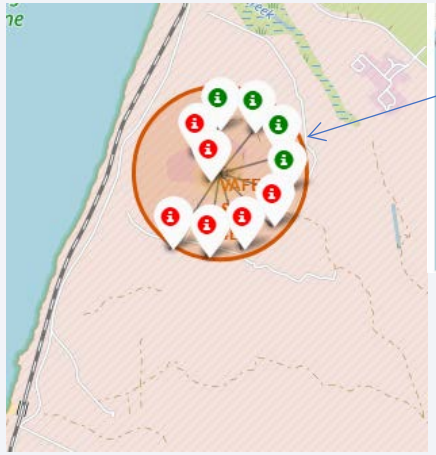
Launch Sites Proximities Analysis

Launch Site Locations

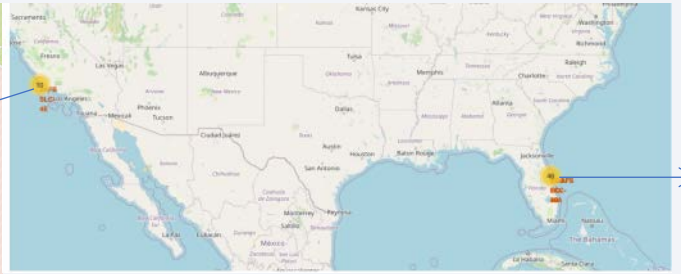


- The only SpaceX launch site locations on the global map are found in the United States
  - Sites are located near the ocean in California and Florida

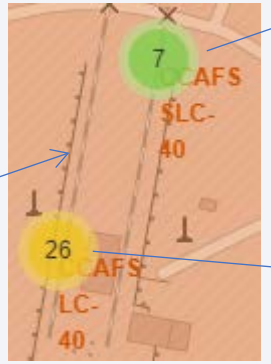
Color Labeled Launch Outcomes on the Map

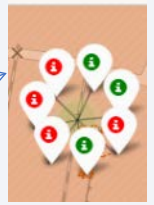


10: VAFB SLC-4E



13: KSC LC-39A



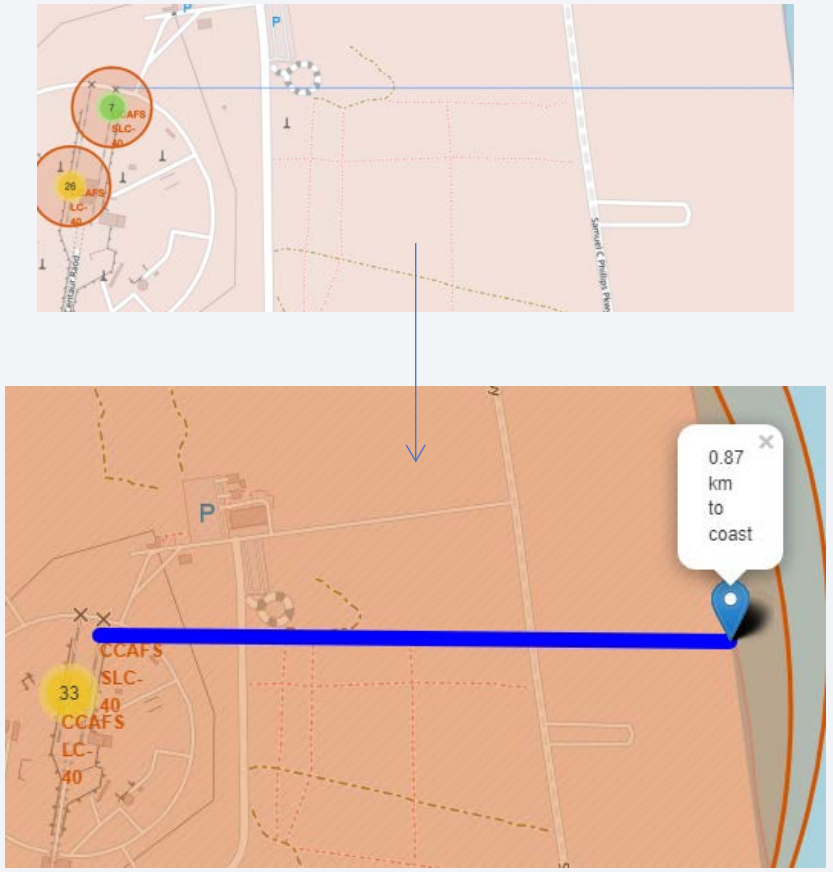


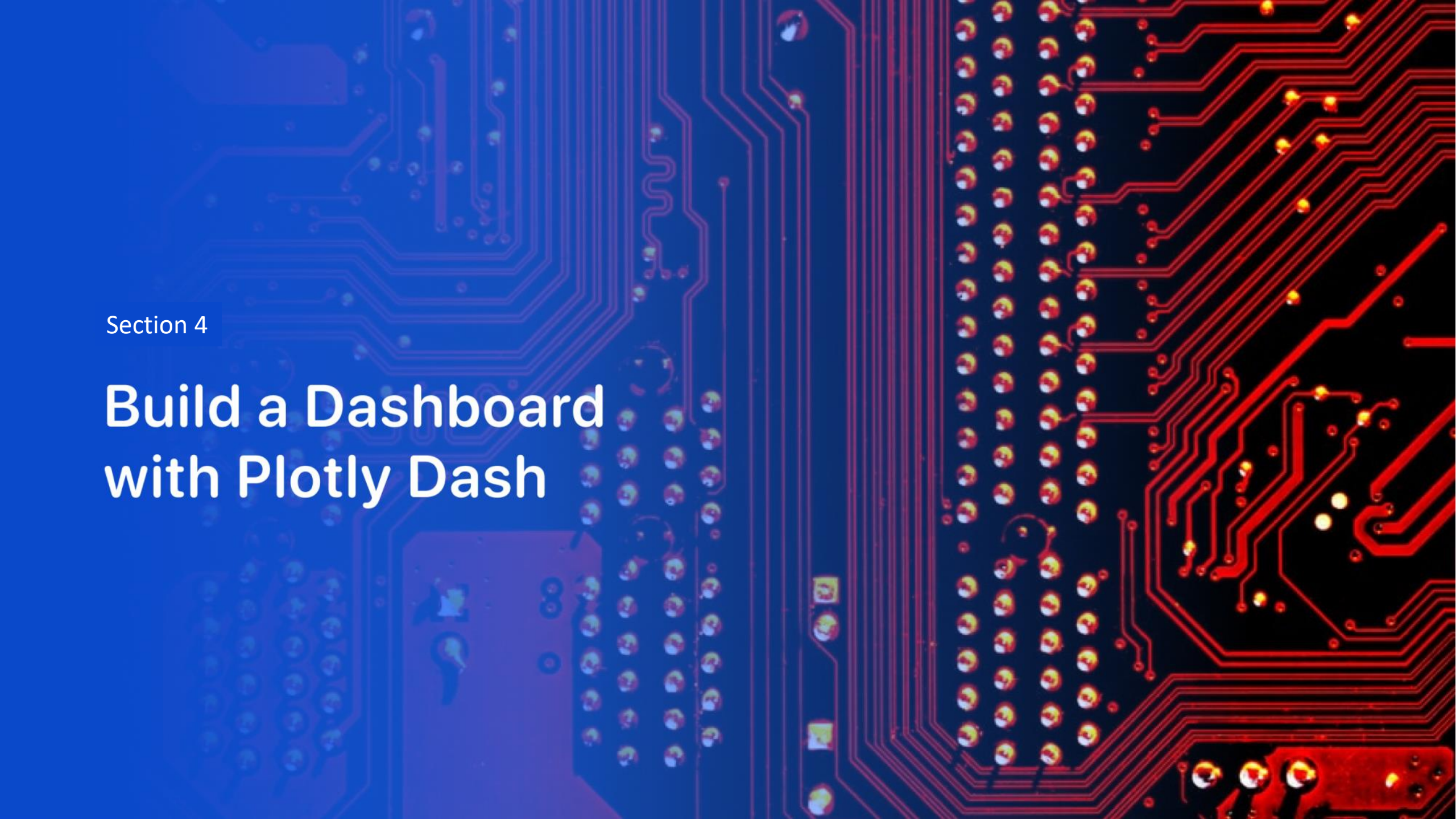
7: CCAFS SLC-40
26: CCAFS LC-40

- Successful launches are the Green markers at the sites
- Failed launches are the Red markers at the sites

Florida Launch Sites Distance to Coastline

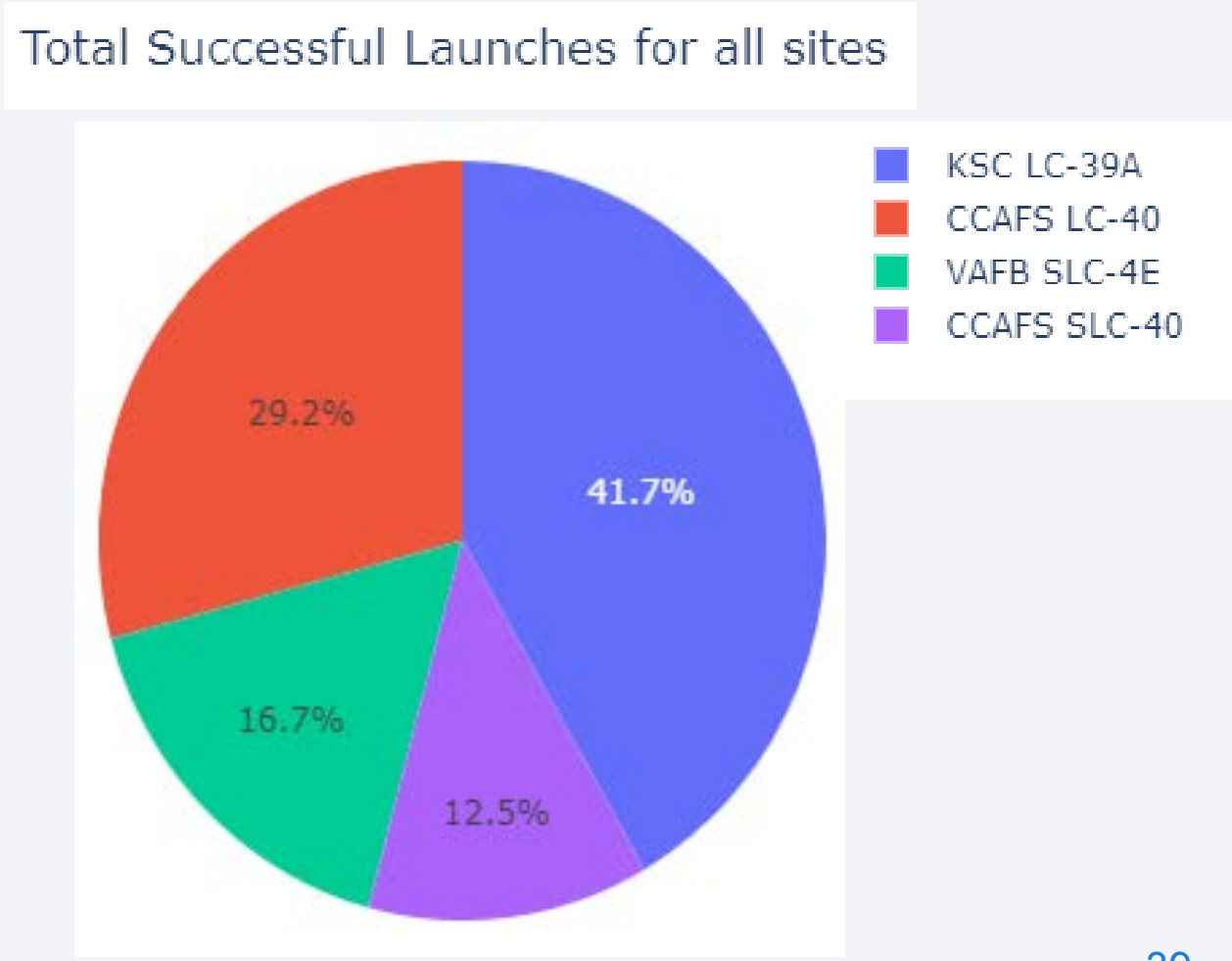
- Both the California and the Florida launch sites are not in extremely close proximity to major highways or major cities
- The California site is close to a railroad and a coastline
- The Florida sites are also close to a coastline
  - This makes sense as one of the more successful ways to bring the first stage back after landing is via drone ship
- Shown is the distance between the nearest Florida launch site and the coastline
  - Chose to represent the Florida sites because the majority of launch sites are in this area




Section 4
Build a Dashboard with Plotly Dash

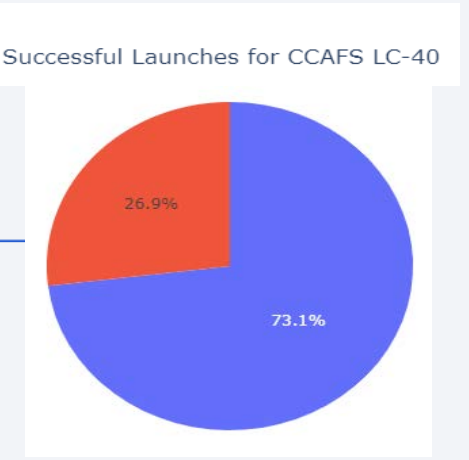
Total Successful Launches for All Sites

- The pie chart shows the percent breakdown, by site contribution, to the total successful launches
- Site KSC LC-39A has the highest percentage of successful SpaceX launches
- Site CCAFS SLC-40 has the lowest percentage of successful launches

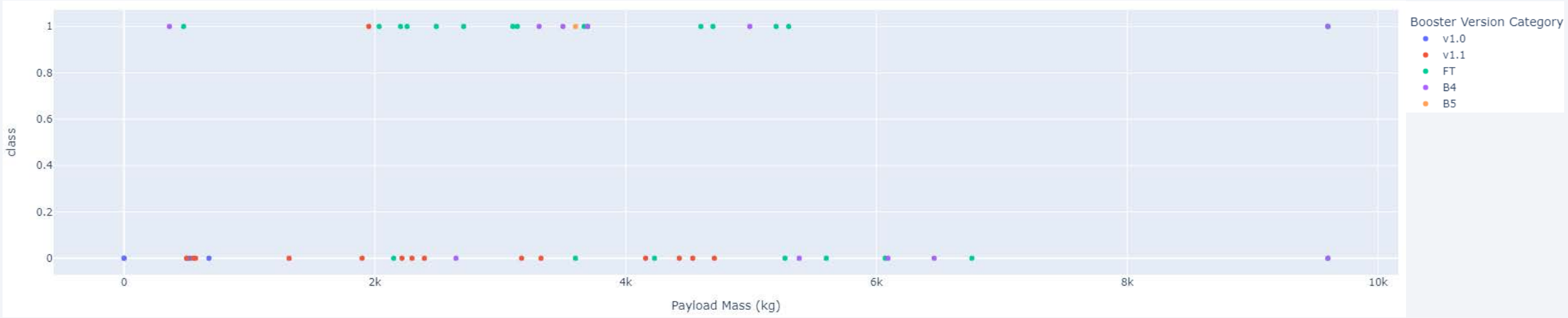


Success/Failure Breakdown for Each Launch Site

- Launch site KSC LC-39A has the highest launch success rate at 76.9%
  - This site accounted for 41.7% of all successful launches
- Launch site CCAFS SLC-40 had the lowest launch success rate at 57.1%
- A positive note that all launch sites had more successful launches than failures



Payload vs. Launch Outcome for all sites



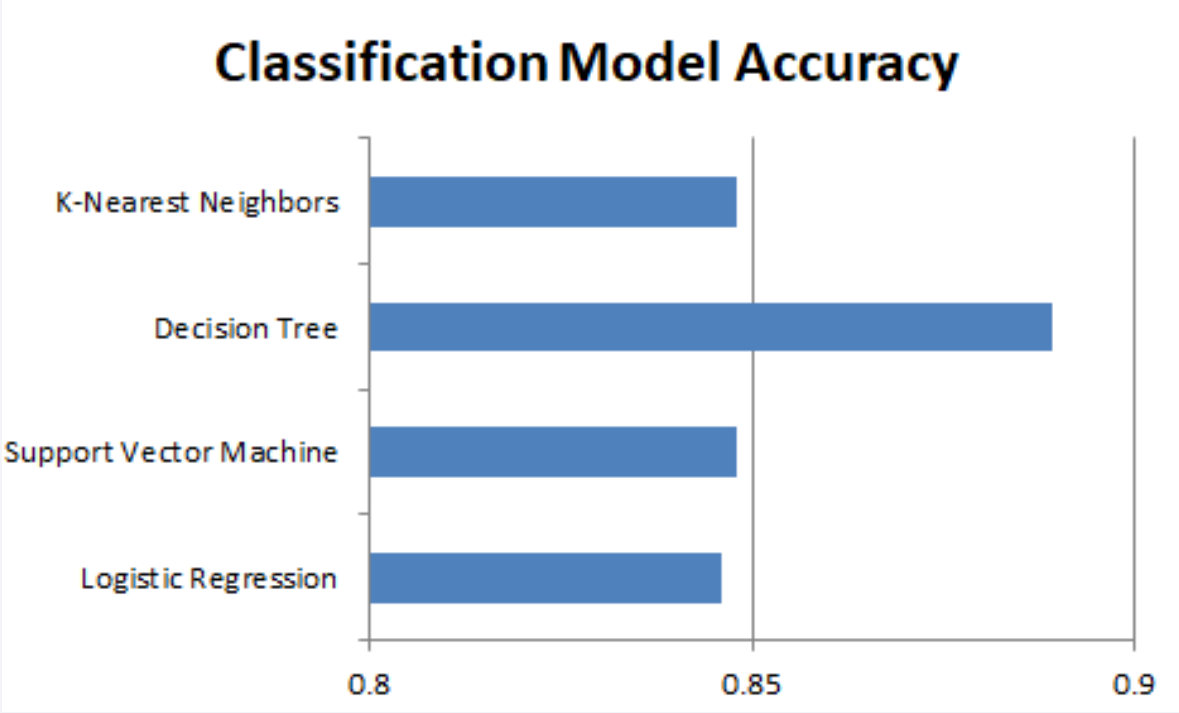
- Payload masses between 2,000 – 4,000 kg appear to have the most success
  - This makes in part because a lighter first stage should be easier to recover
- Booster version FT has the most successful launch outcomes, while booster v1.1 has the least successful outcomes.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

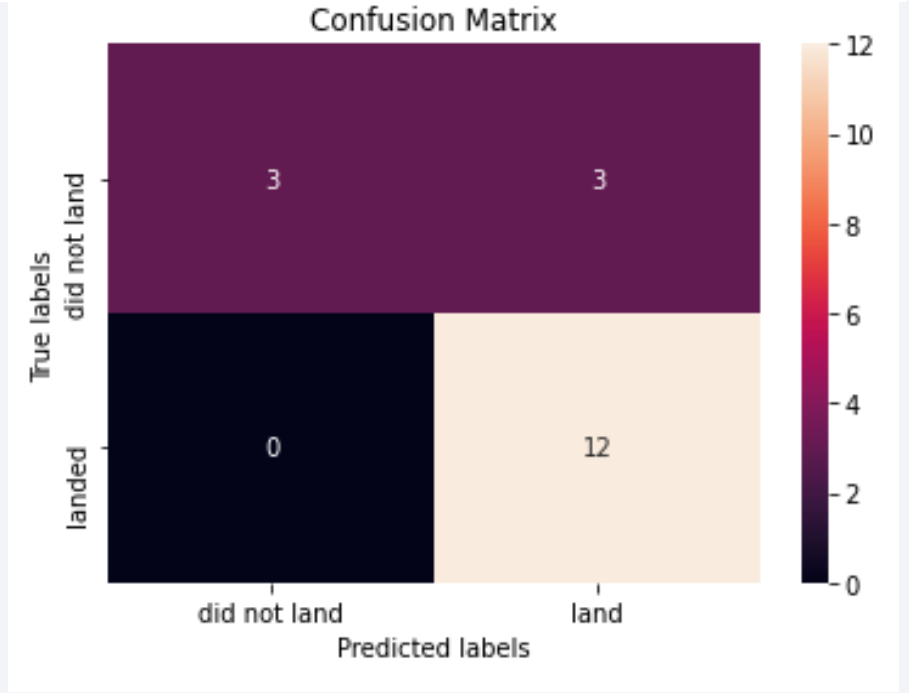
- The Decision Tree method performs best, as the model (89%) is the most accurate, with the others around 85% accuracy
- Accuracy on the validation data when model is tuned to use the best parameters
- Each classification model returned 83.3% accuracy on the test data



Confusion Matrix

- The confusion matrix for the decision tree classifier is shown
- The matrix does show that the decision tree classifier can distinguish between different classes
- However, notice the false positives where the model was predicted to land three times when it actually didn’t land

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



Conclusions

After analyzing the data in multiple ways, the most successful launches have the following characteristics:

- Are conducted in orbits ES-L1, GEO, HEO, and SSO
- Are performed after 2017, implying that this year's test should succeed at least 60% of the time (once in last four years), and closer to 80% or more (three of last four years)
- The most successful landing outcome is recovery by drone ship
- The highest chance of a successful launch is to run the test at site KSC LC-39A
- Payload masses between 2,000 – 4,000 kg are most successful

Further studies should include finding the price savings for a successful test

Thank you!