

# Data Science Capstone Project

---

**Bayu Wilson**

**August 25, 2021**

# Outline

---



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

# Executive Summary

---



- Methodologies
  - SpaceX REST API and Beautiful Soup to collect launch data
  - Exploratory Data Analysis (EDA) with visualization which found interesting correlations between success and payload mass/time/launch pad.
  - Used SQL to find that SpaceX is improving its success rate over time
  - Used folium maps to notice how all the launch sites are in the southern part of USA (closer to equator) and on the coasts near the oceans
  - Created a dashboard to help visualize and determine which landing pad had the highest success rate
  - Applied predictive analysis to build the best classification model: TREE
- The most important factors in predicting a safe first-stage landing are year of launch and launch pad. Using the TREE methodology provides the best predictions for successful launches.

# Introduction

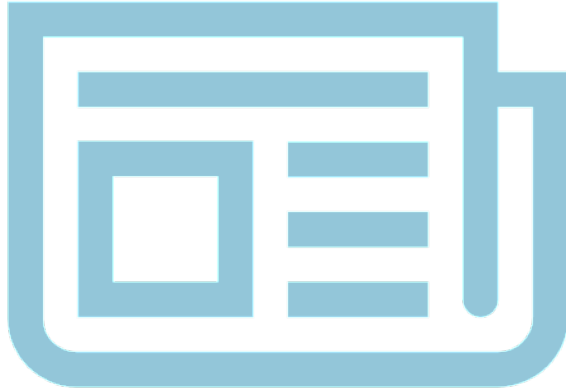
---



- Society is entering a new age where space travel is becoming more technologically feasible. Though a major obstacle that must be addressed is the astronomical cost of space travel. One solution used by SpaceX is reusing rockets (first stage) by landing them back safely after launch. Unfortunately, the first stage still fails to land occasionally which would nullify this money-saving solution. **The purpose of this project is to determine what factors allow for a safe first stage landing.** A safe landing means the rocket can be reused which would decrease the cost of rocket launches compared to single-use rockets.

# Methodology

---



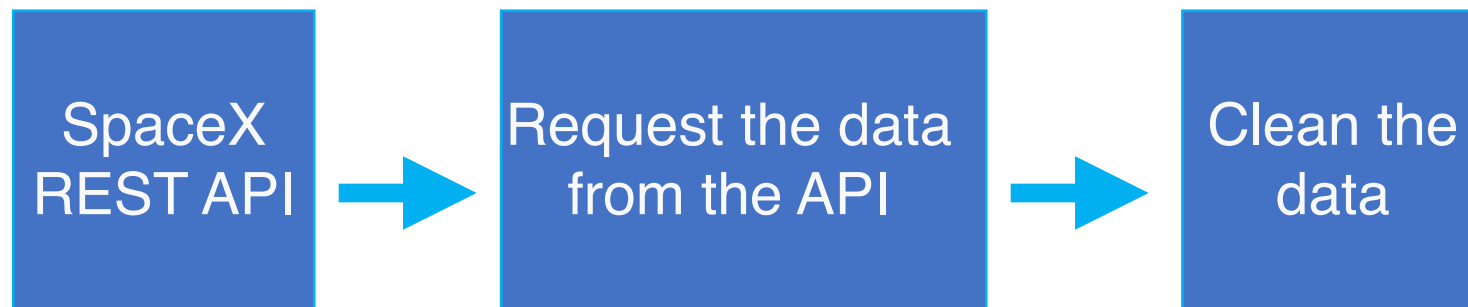
- Data collection methodology:
  - Data was collected using GET requests to the SpaceX REST API as well as web scraping using BeautifulSoup
- Perform data wrangling
  - Used pandas data frames, visualizations, and SQL to perform exploratory data analysis (EDA)
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# Methodology

# Data collection

---

- In order to gather SpaceX data about the launches, we use the SpaceX Rest API
- I used the `requests` library and used a GET request to get the the rocket launch data from the API
- Then I turned the data file into a pandas correctly-labeled data frame, filtered the data frame to only include Falcon 9 launches and then dealt with missing values by replacing them with the mean of the column



# Data collection – SpaceX API

In this slide I show how the data was requested and collected from the SpaceX API

[Github URL Hyperlink](#)

Request the data from the API

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

Convert json file into a dataframe, then clean up the data frame and make it easy to understand

```
data = pd.json_normalize(response.json())
...
launch_dict_df = pd.DataFrame(launch_dict)
```

Filter the data frame to only include Falcon 9 launches

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

Use mean of PayloadMass column to fill in the missing values

```
mean_falcon9 = data_falcon9['PayloadMass'].mean()
data_falcon9 = data_falcon9['PayloadMass'].replace(to_replace=np.nan, value=mean_falcon9)
```



# Data collection – Web scraping

Present your web scraping process use  
key phrases and flowcharts

Github URL Hyperlink

Requested the Falcon9 Launch Wiki page from its URL  
using BeautifulSoup

```
r = requests.get(static_url)
soup = BeautifulSoup(r.text, 'html.parser')
```

Extract all column/variable names from the HTML table header

```
column_names = []

for i in first_launch_table.find_all('th'):
    name = extract_column_from_header(i)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

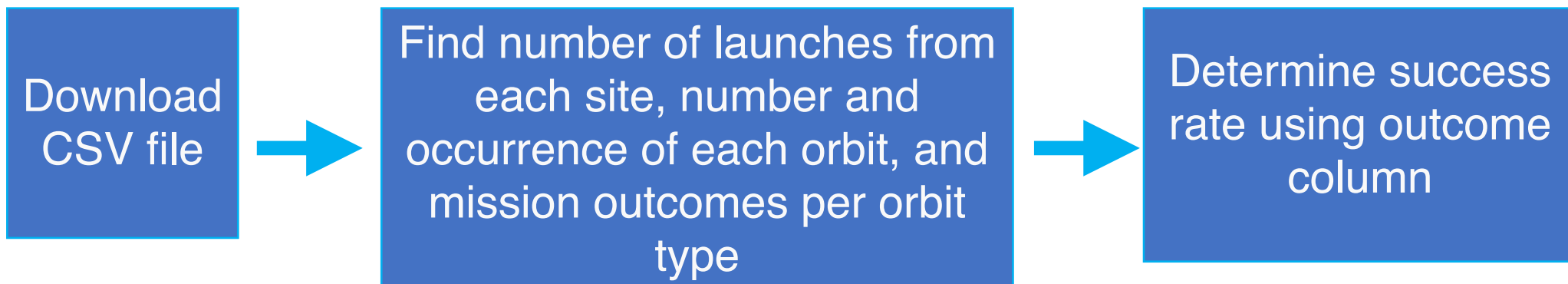
Create a data frame by parsing the launch HTML tables

*In the code I looped through each independent launch variable to  
create a launch dictionary which is easily converted to a data frame.  
This data frame is then exported to a CSV file to be used in the  
subsequent section.*

# Data wrangling

---

- Performed Exploratory Data Analysis (EDA) to help find patterns in data as well as inform our decisions regarding ML models.
- Found that there are 3 launch sites with the most launches from “CCAFS SLC 40”. Combining the success rate of all the launches, it was found that ~67% of outcomes were successful (rocket landed safely).
- **Github URL Hyperlink**



# EDA with data visualization

---

- In the first few plots, it appeared that larger mass payloads and higher flight numbers are more likely to land successfully. To delve deeper into this hypothesis we checked if this holds true for each orbit type. In the LEO orbit the Success appears related to the flight number but there seems to be no relationship between flight number when in GTO orbit. Also, heavy payloads appear to decrease the success rate on GTO orbits and increase the success rate on Polar LEO (ISS) orbits. Though what seemed to be the **most important factor is time**. As time went on, the success rate flew at an upward trajectory in the majority of cases. This is seen in the final yearly trend figure.
- **Github URL Hyperlink**

# EDA with SQL

---

- The SpaceX data table was stored in the cloud and then I used IBM Db2 on Cloud to write and execute the SQL queries
- I wrote various SQL queries to explore the data. I displayed unique launch sites, found average payload mass by a booster version, and even the counts of each type of landing outcome. I noticed that we sampled the data between ~2010 and ~2017. Prior to that the success rate is lesser and after that the success rate is higher. This means that we are looking at a section of time when SpaceX is mature enough to improve its success rate but it had not perfected it yet.
- **Github URL Hyperlink**

# Build an interactive map with Folium

---

- Found that all the launch sites are in the southern part of USA (closer to equator) and on the coasts near the oceans
- Made map easier to read using marker clusters
- Noticed how launch sites are nearby railways and far away from cities
- **Github URL Hyperlink**

# Build a Dashboard with Plotly Dash

---

- KSC LC-39A had the most successful launches while CCAFS SLC-40 had the least amount of successful launches
- Of all 13 of KSC LC-39A launches, 76.9% of them were successful
- It seems that often the B5 and v1.1 boosters versions failed while the FT version often succeeded. At masses greater than 5500 kg all boosters failed, even FT, in this mass range
- **Github URL Hyperlink**

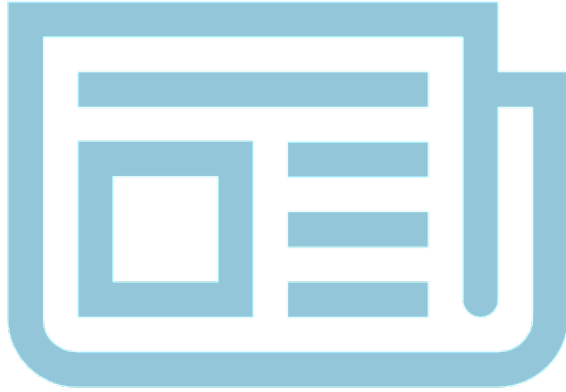
# Predictive analysis (Classification)

---

- The TREE method has the highest classification accuracy at about ~87.6%
- The model is 100% accurate if the true label is 'landed'. Though it has some trouble knowing the outcome if the true outcome is 'did not land'. Here the problem is false-positives... predicting it landed when it actually did not land
- **Github URL Hyperlink**

# Results

---

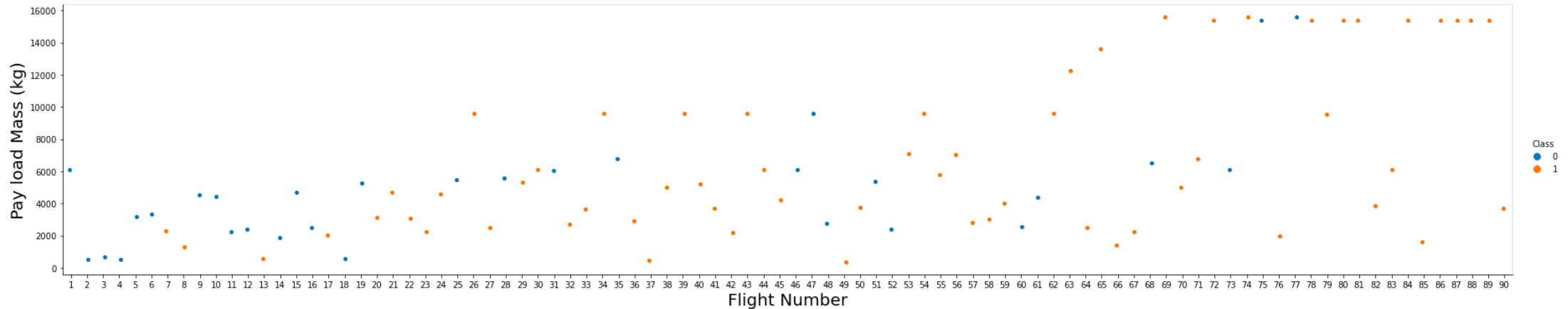


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



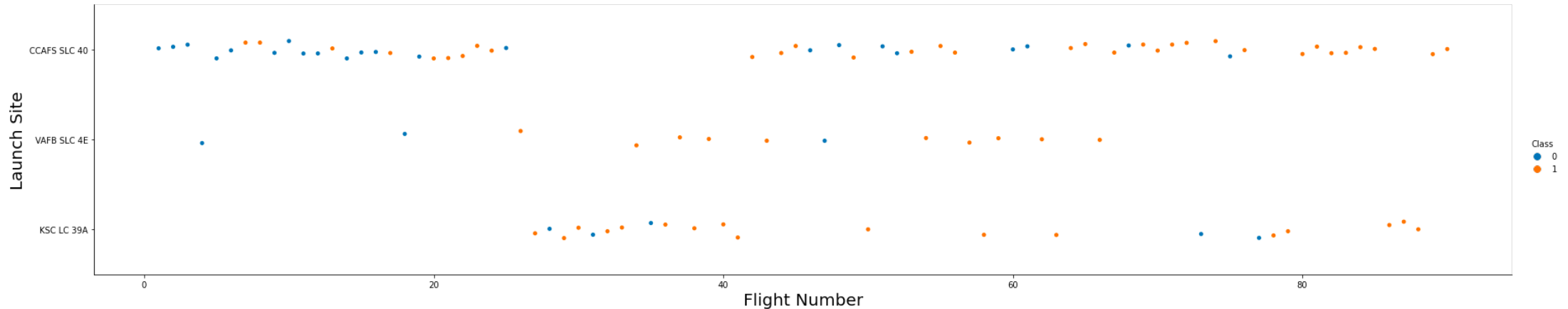
# EDA with Visualization

# Payload mass vs flight number



- In this figure we have **payload mass** on vertical axis and **flight number** on horizontal axis. Blue dots represent first stage landing failure and orange dots represent success. It is apparent that **larger mass payloads and higher flight numbers are more likely to land successfully**.
- Github URL Hyperlink

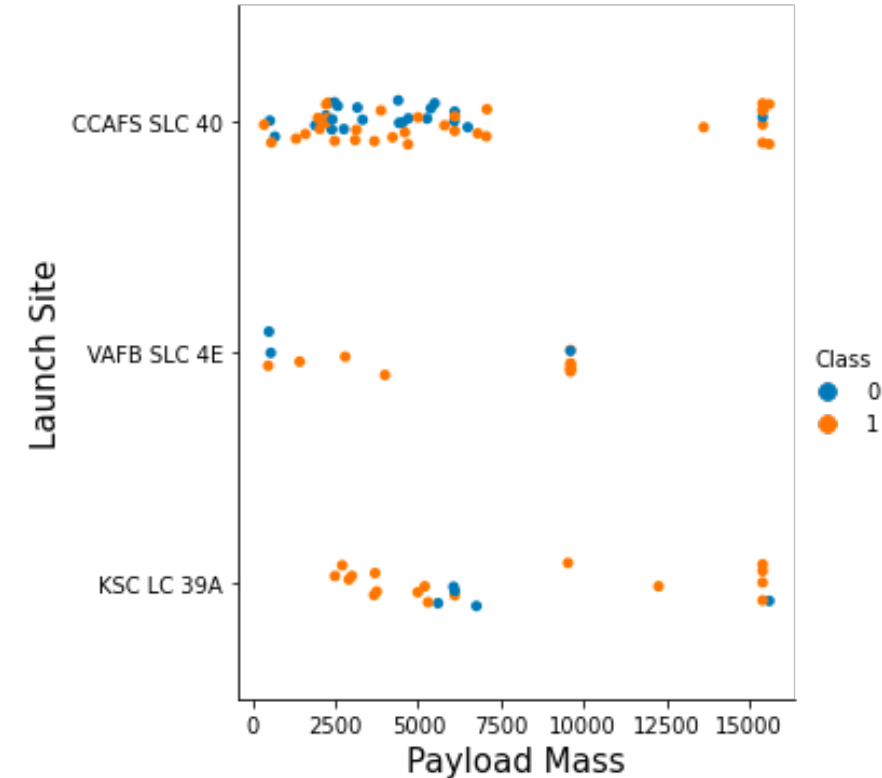
# Flight Number vs. Launch Site



- In this figure we have **launch site** on vertical axis and **flight number** on horizontal axis. Blue dots represent first stage landing failure and orange dots represent success. It is clear that **higher flight number means it's more likely that the landing is successful**. Launch site does not seem to be a significant factor though.
- **Github URL Hyperlink**

# Payload vs. Launch Site

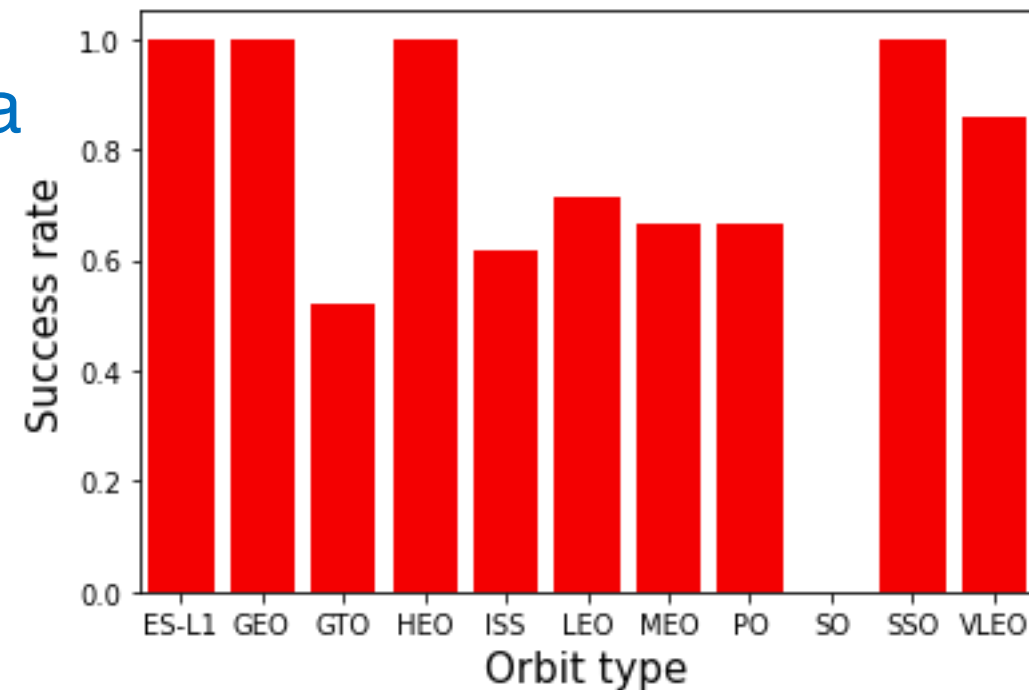
- In this figure we have **launch site** on vertical axis and **payload mass** on horizontal axis. Blue dots represent first stage landing failure and orange dots represent success. For all three launch site, it appears that **higher payload masses are more successful**.
- **Github URL Hyperlink**



# Success rate vs. Orbit type

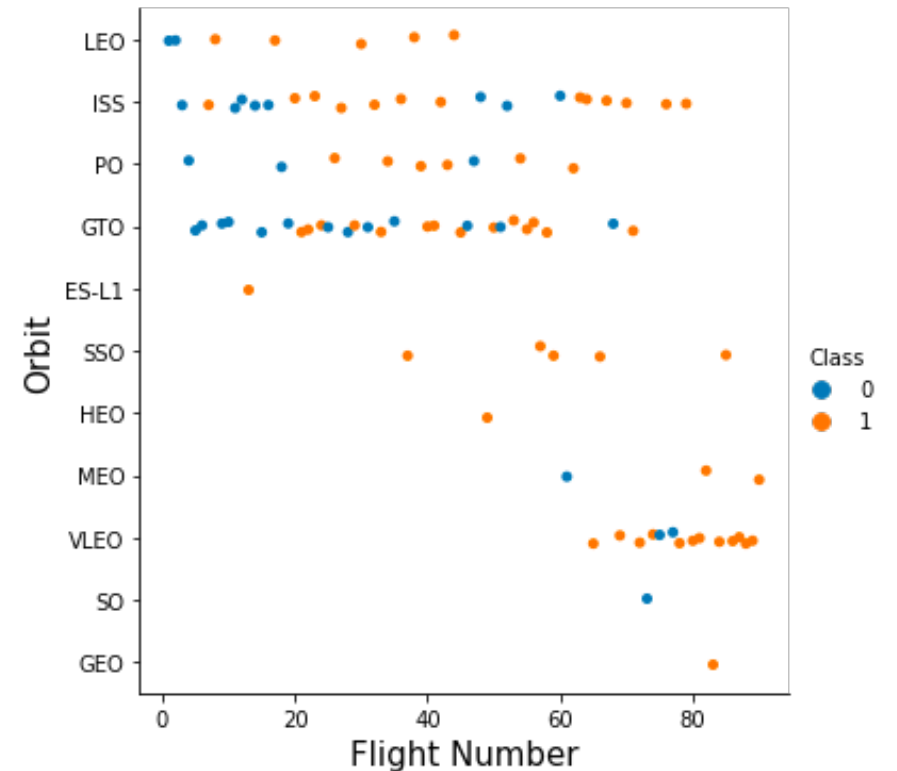
---

- In this figure we have **success rate** on vertical axis and **orbit type** on horizontal axis. Most orbit types have a 100%, ~60%, or 0% success rate
- **Github URL Hyperlink**



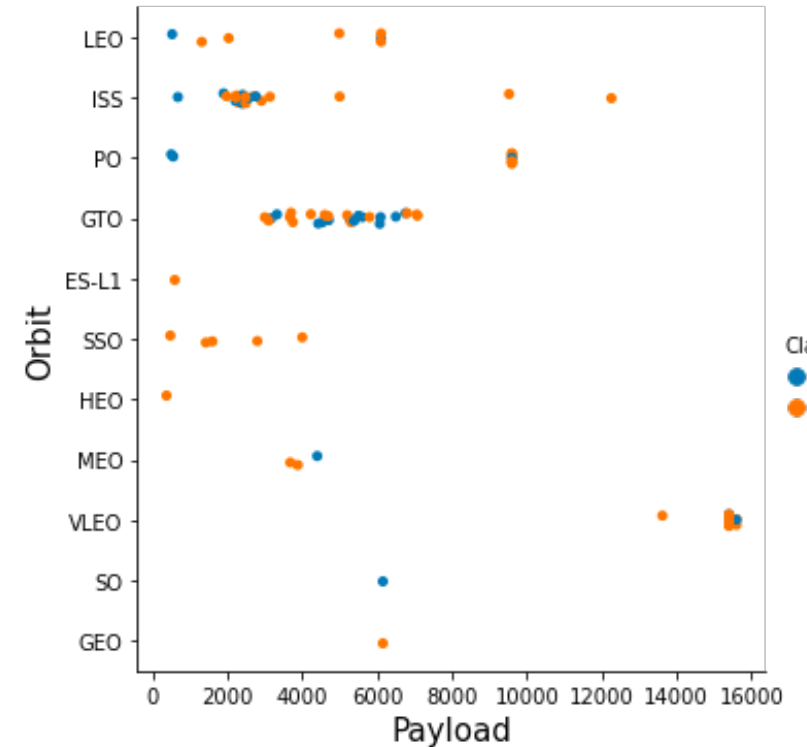
# Flight Number vs. Orbit type

- In this figure we have **orbit type** on vertical axis and **flight number** on horizontal axis. In the LEO orbit the Success appears related to the flight number but there seems to be no relationship between flight number when in GTO orbit.
- **Github URL Hyperlink**



# Payload vs. Orbit type

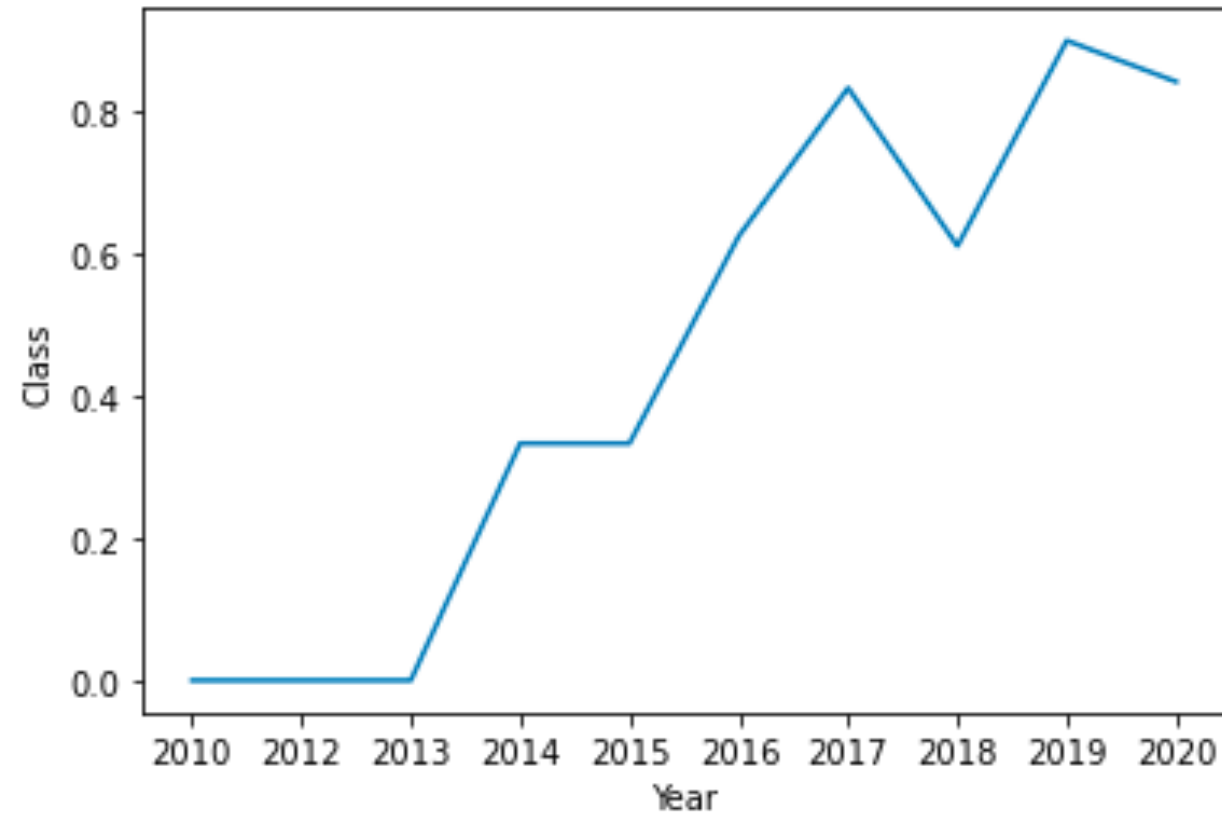
- In this figure we have **orbit type** on vertical axis and **payload mass** on horizontal axis. Heavy payloads appear to decrease the success rate on GTO orbits and increase the success rate on Polar LEO (ISS) orbits
- **Github URL Hyperlink**



# Launch success yearly trend

---

- In this figure we have **success rate** on vertical axis and **year of launch** on horizontal axis. The success rate has been increasing from 2013 until 2020
- **Github URL Hyperlink**



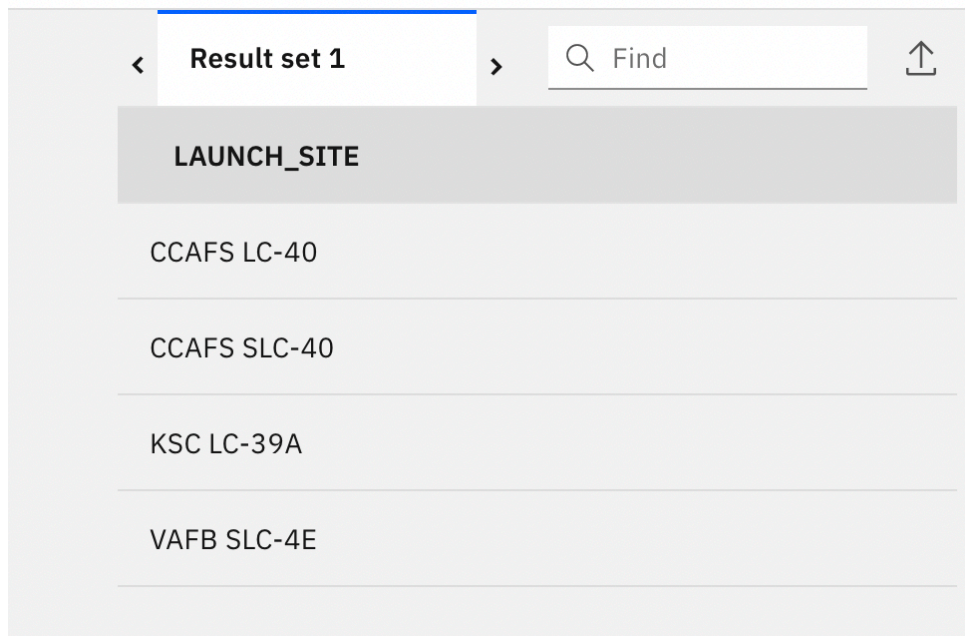


# EDA with SQL

# All launch site names

---

```
SELECT UNIQUE(LAUNCH_SITE) FROM NEWSPACEXTABLE;
```



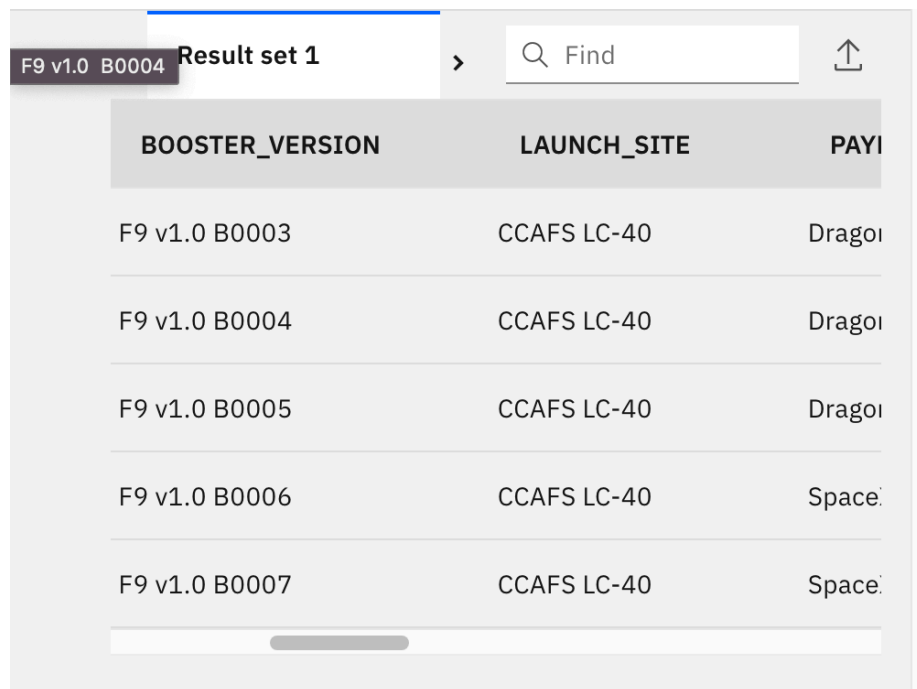
The screenshot shows a database query result interface. At the top, there is a tab labeled 'Result set 1' and a search bar with the text 'Find'. Below the search bar, the column header 'LAUNCH\_SITE' is displayed. The table contains four rows of data, each representing a unique launch site name.

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

Selected unique names  
of launch site using  
**UNIQUE** sql function

# Launch site names begin with `CCA`

```
SELECT * FROM NEWSPACEXTABLE  
WHERE LAUNCH_SITE LIKE 'CCA%'  
LIMIT 5;
```



The screenshot shows a database interface with a table titled 'Result set 1'. The table has three columns: 'BOOSTER\_VERSION', 'LAUNCH\_SITE', and 'PAYLOAD'. It displays five rows of data, all with 'CCAFS LC-40' as the launch site. The first three rows have a payload of 'Dragon', and the last two have a payload of 'SpaceX'.

BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD
F9 v1.0 B0003	CCAFS LC-40	Dragon
F9 v1.0 B0004	CCAFS LC-40	Dragon
F9 v1.0 B0005	CCAFS LC-40	Dragon
F9 v1.0 B0006	CCAFS LC-40	SpaceX
F9 v1.0 B0007	CCAFS LC-40	SpaceX

Found 5 launch sites that begin with 'CCA' using WHERE (...) LIKE (...) condition as well as LIMIT function

# Total payload mass

---

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM NEWSPACEXTABLE  
WHERE CUSTOMER='NASA (CRS)';
```

Total payload mass carried  
by NASA (CRS) boosters is  
45596 kg

# Average payload mass by F9 v1.1

---

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM NEWSPACEXTABLE  
WHERE BOOSTER_VERSION='F9 v1.1';
```

Average payload mass using F9 v1.1 is 2928 kg

# First successful ground landing date

---

```
SELECT MIN(DATE) FROM NEWSPACEXTABLE  
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

First successful landing outcome on ground pad was 2015-12-22

# Successful drone ship landing with payload between 4000 and 6000

---

```
SELECT BOOSTER_VERSION FROM NEWSPACEXTABLE  
WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND  
(PAYLOAD_MASS__KG_>4000) AND  
(PAYLOAD_MASS__KG_<6000);
```

Result set 1		Find	↑
BOOSTER_VERSION			
F9 FT B1022			
F9 FT B1026			
F9 FT B1021.2			
F9 FT B1031.2			

Found boosters which satisfy payload mass condition and that are successful on drone ship using WHERE (...) AND (...)

# Total number of successful and failure mission outcomes

---

```
SELECT COUNT(MISSION_OUTCOME) AS Successes FROM NEWSPACEXTABLE  
WHERE MISSION_OUTCOME LIKE 'Success%'
```

```
SELECT COUNT(MISSION_OUTCOME) AS Failures FROM NEWSPACEXTABLE  
WHERE MISSION_OUTCOME LIKE 'Failure%'
```

Found the total number of successful and and failing mission outcomes using WHERE (...) LIKE (...). There were 100 successful outcomes and 1 failure mission outcomes.



# Boosters carried maximum payload

```
SELECT BOOSTER_VERSION FROM NEWSPACEXTABLE  
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM  
NEWSPACEXTABLE)
```

Result set 1		Find	
BOOSTER_VERSION			↕
F9 B5 B1048.4			
F9 B5 B1049.4			
F9 B5 B1051.3			
F9 B5 B1056.4			
F9 B5 B1048.5			
Result set is truncated, only the first 12 rows have been loaded. Select "View all loaded data" on the right top of the result to view all loaded rows.			
		More	

Found name of boosters that carry the maximum payload mass using a subquery

# 2015 launch records

```
SELECT LANDING__OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM  
NEWSPACEXTABLE  
WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (DATE LIKE '2015%')
```

< Result set 1 >		Find	↑
LANDING__OUTCOME ↑↓	BOOSTER_VERSION		
Failure (drone ship)	F9 v1.1 B1012		
Failure (drone ship)	F9 v1.1 B1015		

Made list of failed landing outcomes, booster versions and launch sites for 2015

# Rank success count between 2010-06-04 and 2017-03-20

```
SELECT COUNT(LANDING__OUTCOME) AS COUNTS, LANDING__OUTCOME FROM NEWSPACEXTABLE
WHERE DATE > '2010-06-04' AND DATE < '2017-03-20'
GROUP BY LANDING__OUTCOME ORDER BY LANDING__OUTCOME DESC
```

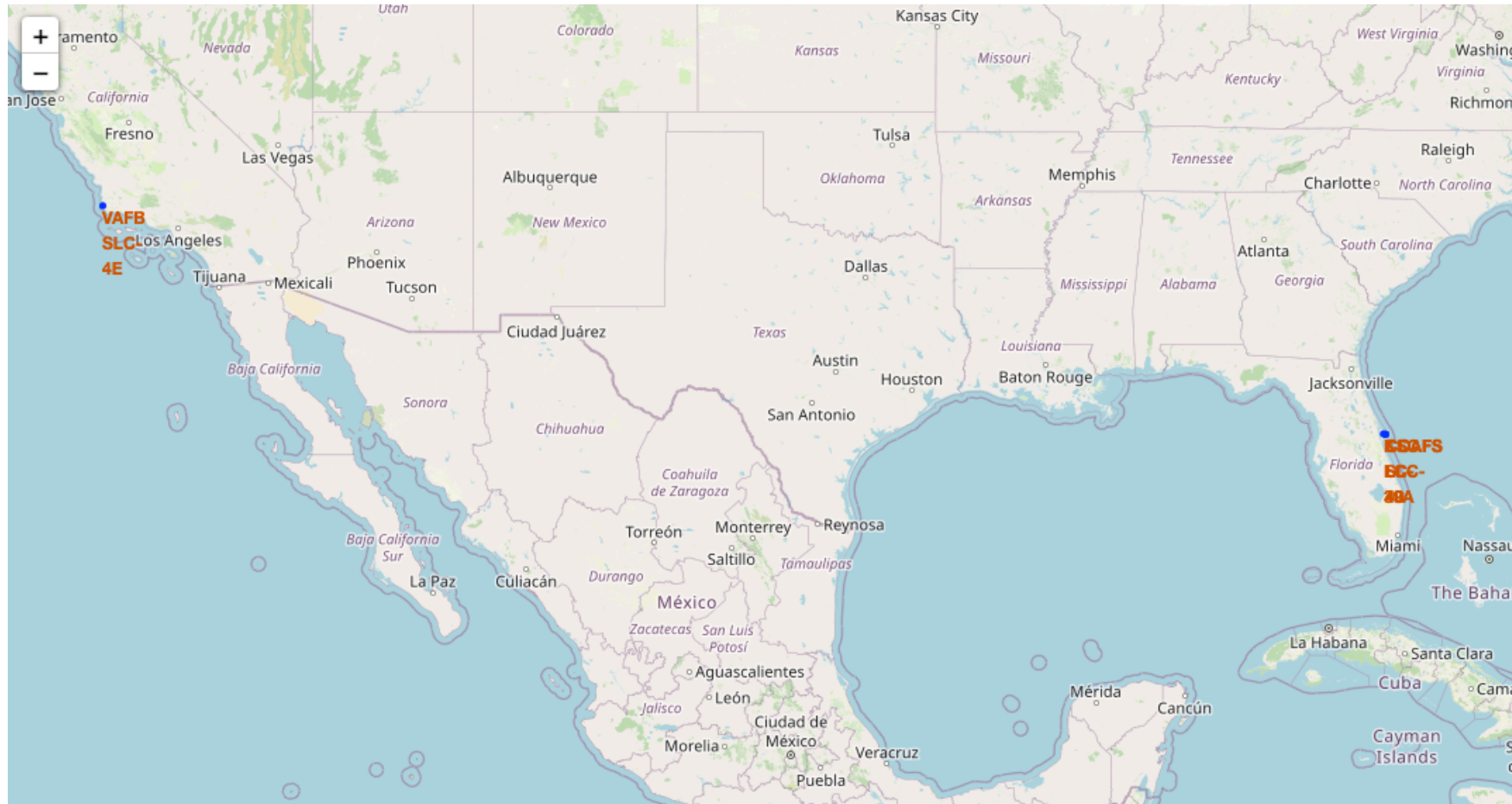
COUNTS	LANDING__OUTCOME
2	Uncontrolled (ocean)
5	Success (ground pad)
5	Success (drone ship)
1	Precluded (drone ship)
10	No attempt

Result set is truncated, only the first 5 rows have been loaded

Grouped the types of landing outcomes in a time frame to compare the counts of each outcome. There were a lot of “No attempt” landing outcomes

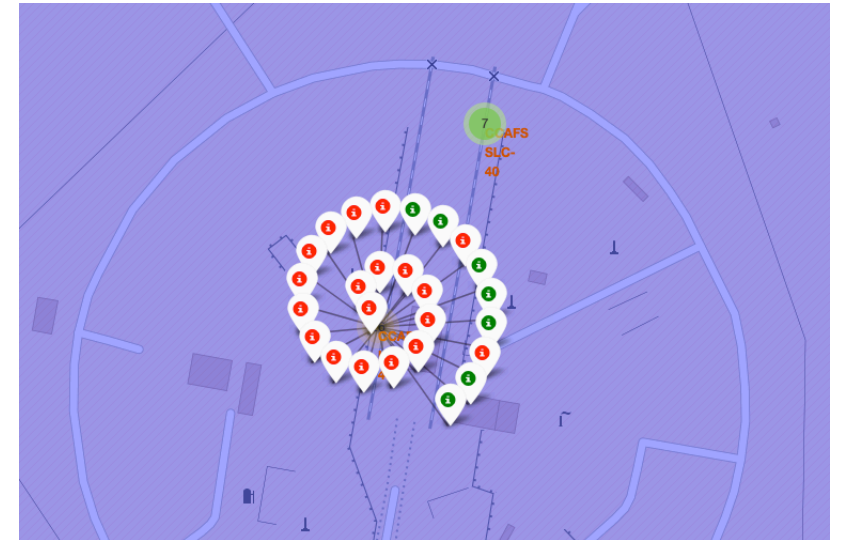
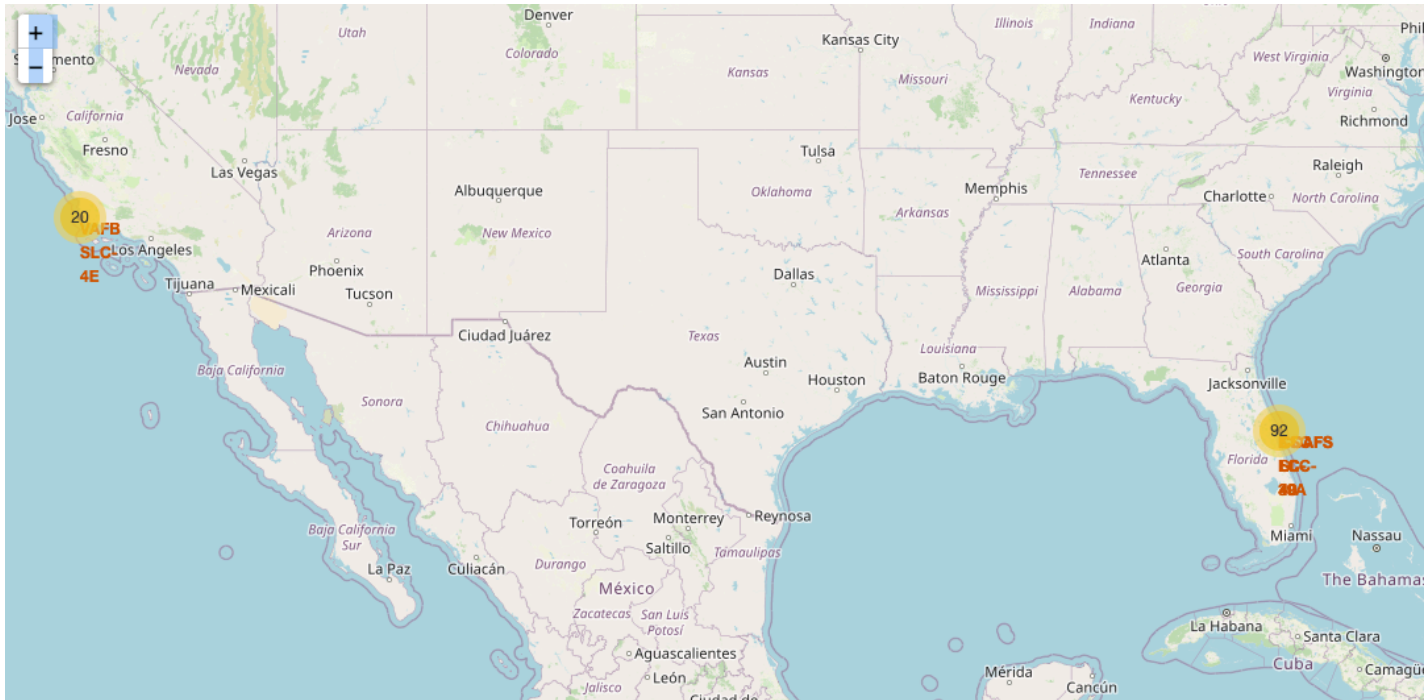
# Interactive map with Folium

# All launch site locations



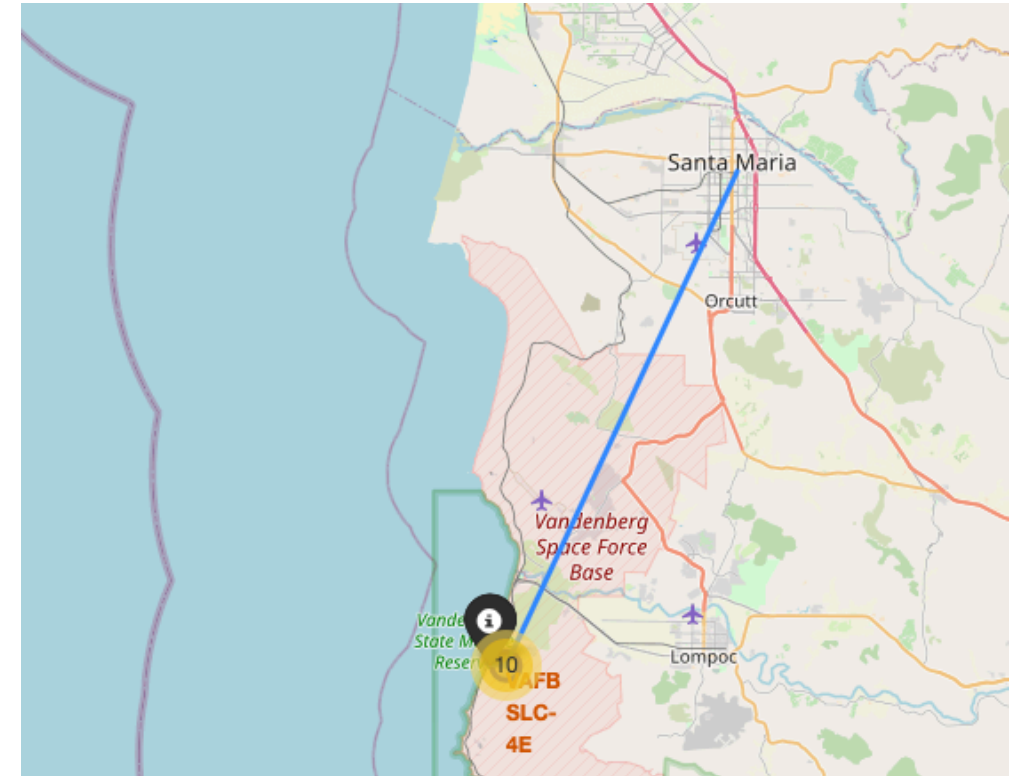
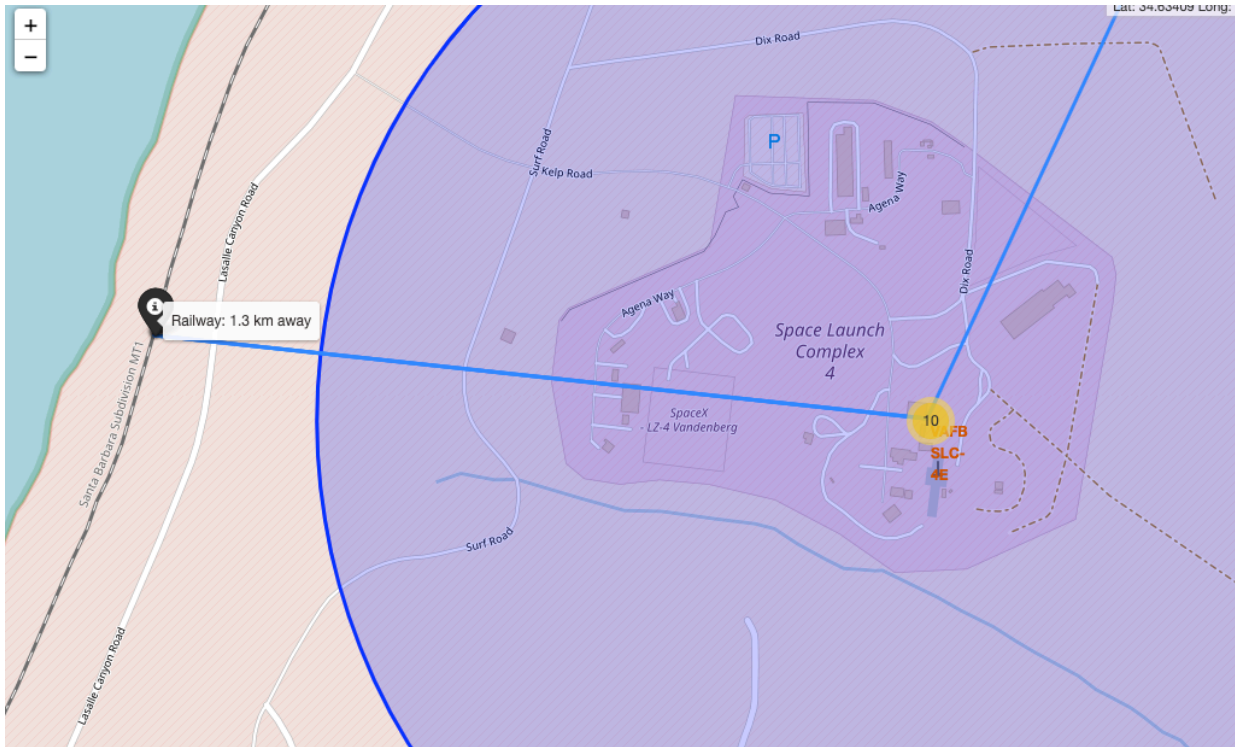
- All the launch sites are in the southern part of USA (closer to equator) and on the coasts near the oceans

# Using marker clusters



- Now the map is easier to understand and more user friendly.

# Proximity of launch site to railway and Santa Maria



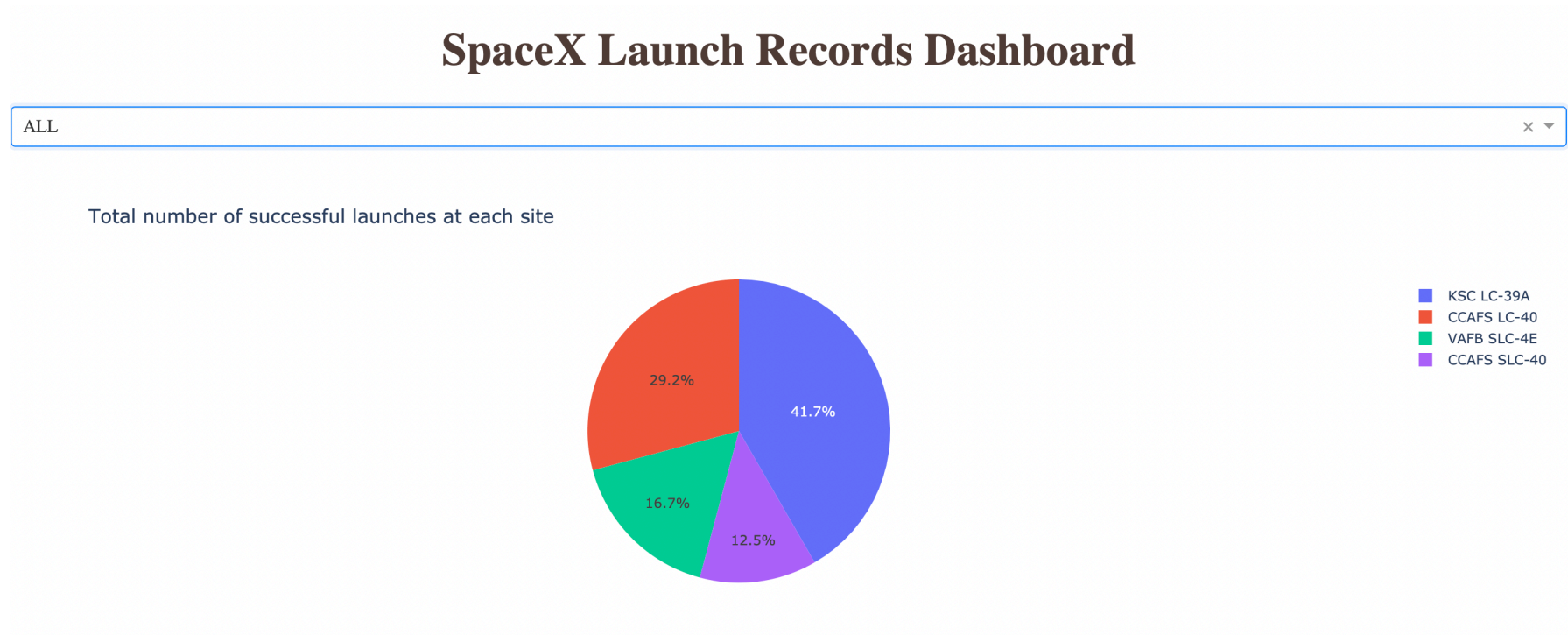
- The launch site is ~1.3 km from a railway and ~39 km from the nearest large city, Santa Maria



# Build a Dashboard with Plotly Dash

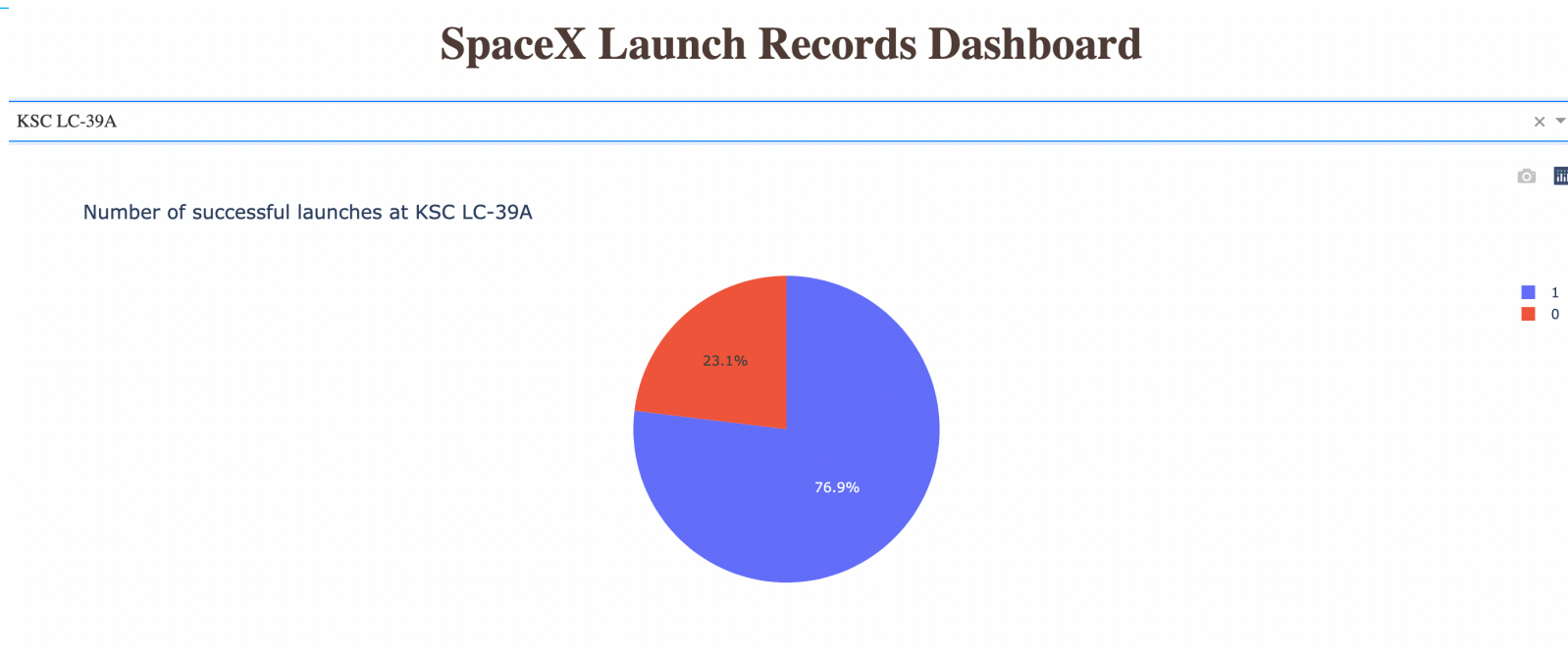


# Total number of successful launches at each site



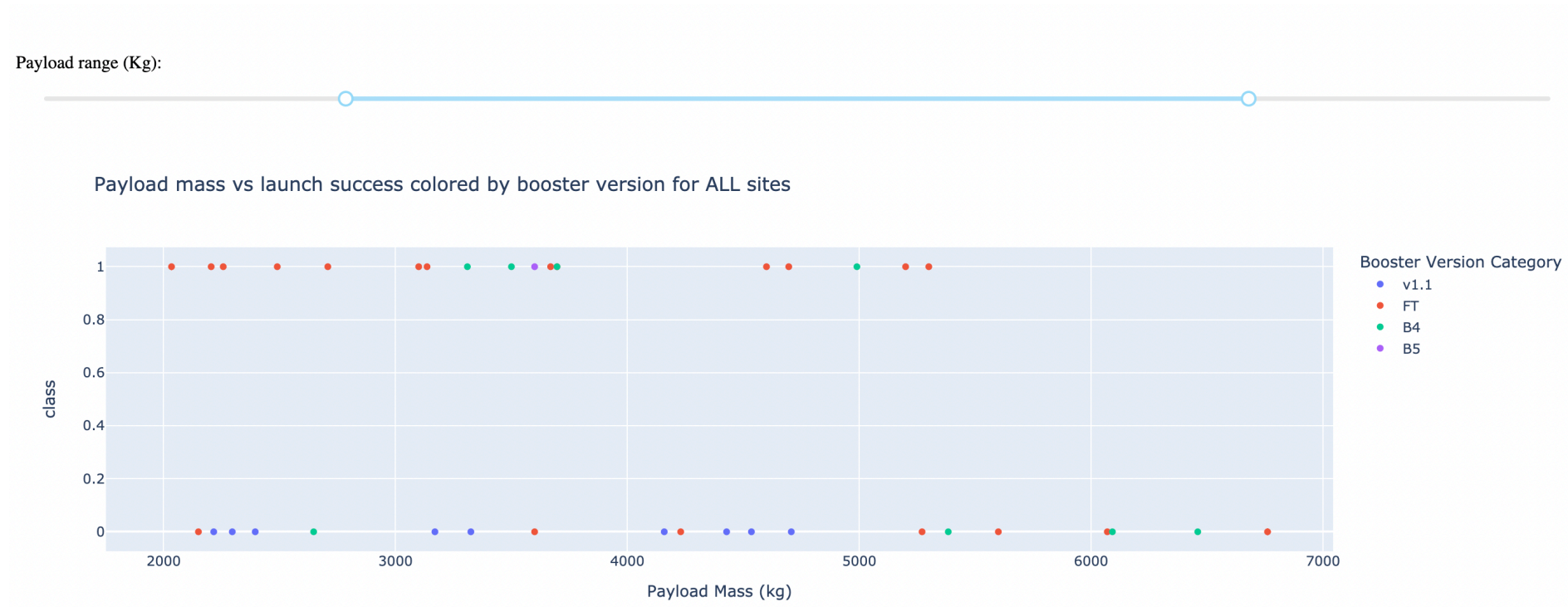
- KSC LC-39A had the most successful launches while CCAFS SLC-40 had the least amount of successful launches

# Number of successful launches at KSC LC-39A



- KSC LC-39A had the most successful launches. Of all 13 of its launches, 76.9% of them were successful

# Payload vs launch success colored by booster version for ALL sites

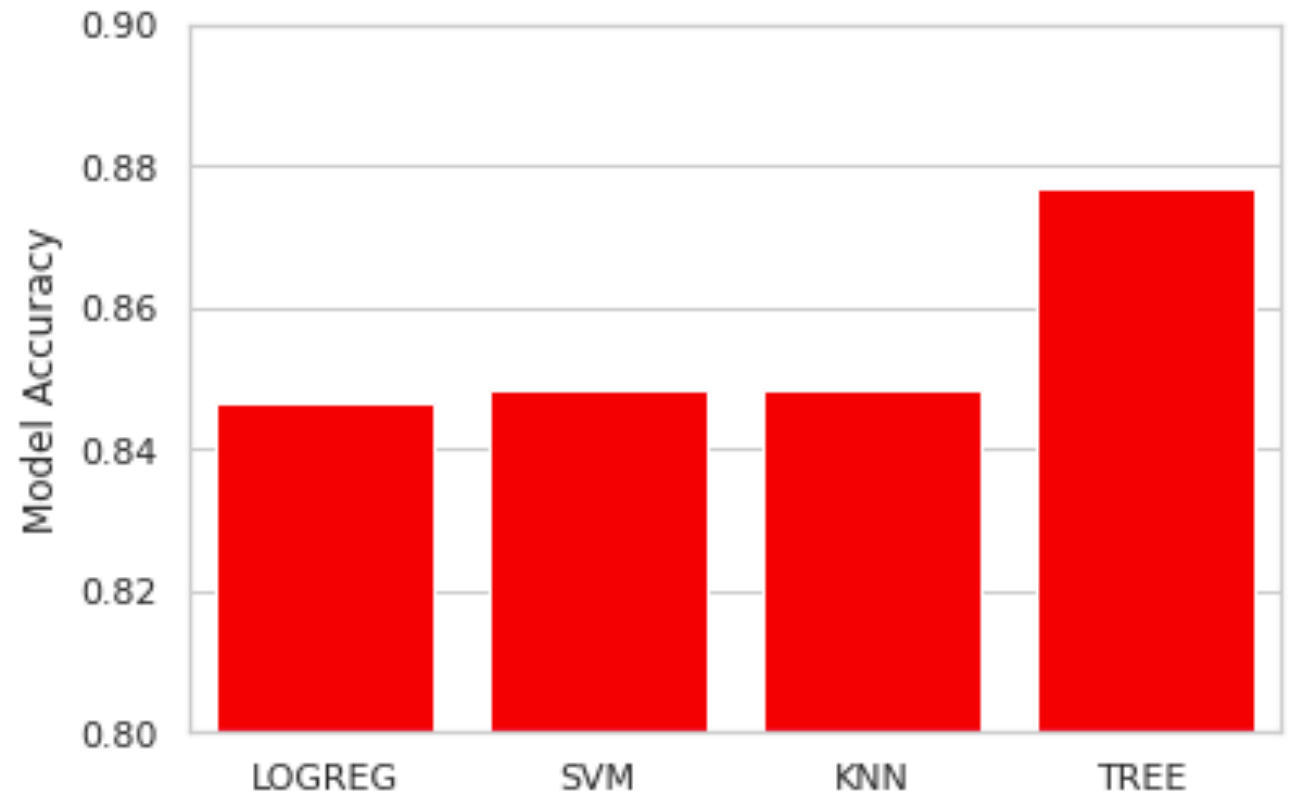


- It seems that often the B5 and v1.1 boosters versions failed while the FT version often succeeded. At masses greater than 5500 kg all boosters failed, even FT, in this mass range

# Predictive analysis (Classification)

# Classification Accuracy

The TREE method has the highest classification accuracy at about ~87.6%

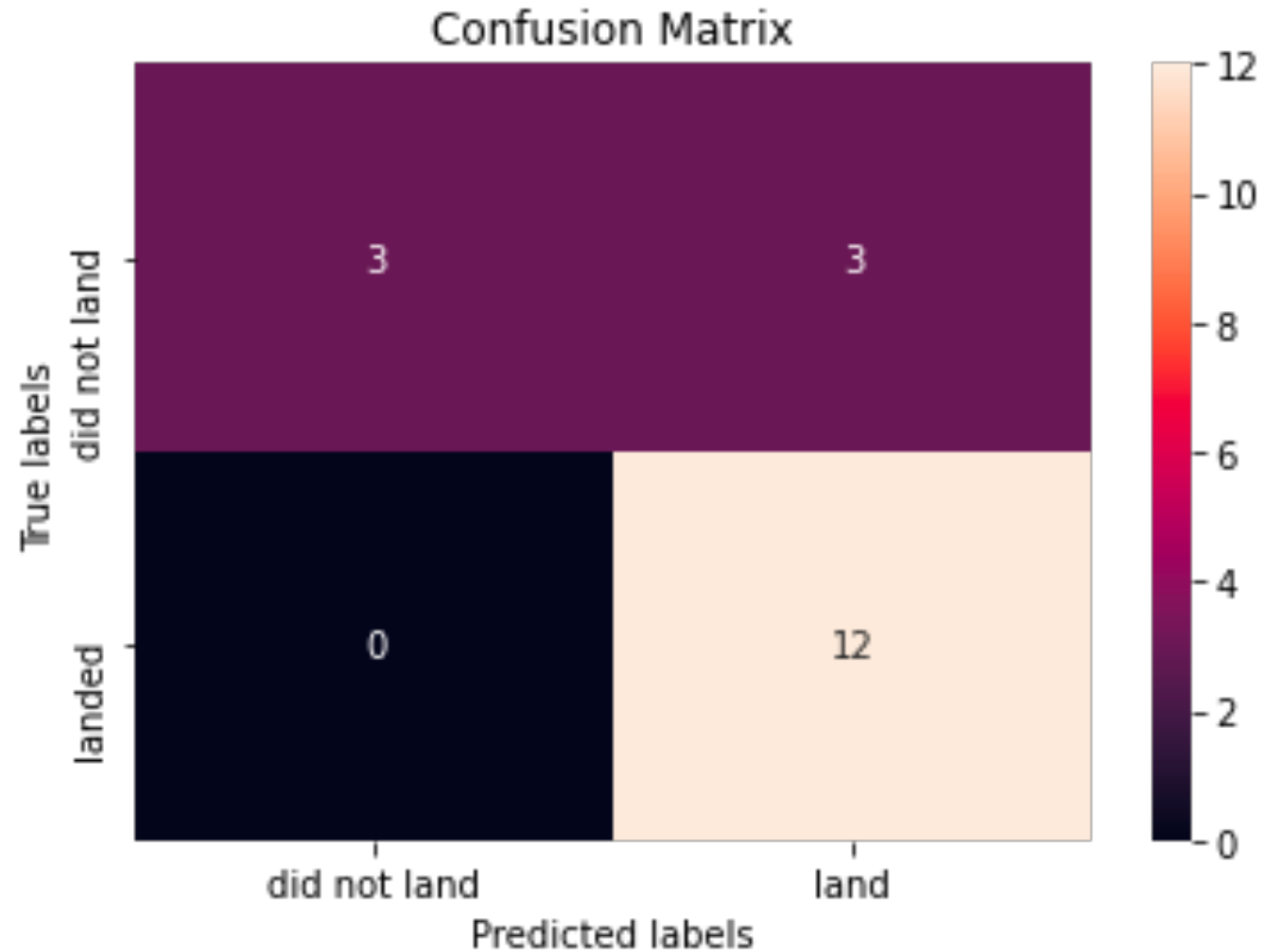


Here are the best tuned hyperparameters:

```
{'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

# Confusion Matrix

The model is 100% accurate if the true label is 'landed'. Though it has some trouble knowing the outcome if the true outcome is 'did not land'. Here the problem is false-positives... predicting it landed when it actually did not land



# CONCLUSION

---



- The most obvious trend found was that success rate correlates positively with time
- Out of four launch pads, KSC LC-39A had the most successful launches. Of all 13 of its launches, 76.9% of them were successful
- TREE is the best method to use with 87% classification accuracy, though it struggles with false-positives