IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Gomes
07.04.2023

Outline

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

Executive Summary

- In this project I utilized exploratory data analysis to find out how rocket launches from SpaceX have evolved over the last few years, with a focus on whether rockets managed to land and thus be recovered. I had a look at how SpaceX utilizes different launch and landing locations and sends their rockets with different payloads and into different orbits.
- The results showed a great increase of the success rate of landing since 2013. The data also showed that only specific payloads are sent into some of the different orbits. I was able to create several models to predict the success of landing a rocket.

Introduction

- The background idea for this experiment was that I wanted to better understand how SpaceX, a big player on the current space race, managed to improve their systems over the last years and what success rate they can achieve on landing their rockets for reuse.

- More specifically I also wanted to find out if the payload or the rocket model, or even the launch site could have impact on the landing success rate. I also wanted to try to develop a model that could successfully predict whether a rocket will land or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected through web scraping by accessing the SpaceX API. The API was accessed through the requests package and then converting the .json data into a pandas data frame.
- The data had to be cleaned to be readable in a data frame format. This was done through filtering for specific words and information within each attribute column.
- After leaving only the relevant columns in the data frame, I filtered the data to include only Falcon 9 flights.

Data Collection – SpaceX API

- API was access with the requests package and the .get() function.

- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/fbde66567e3e0d957073b9bb3fb010488e94b132/notebook_1_spacex-data-collection-api.ipynb

```
In [7]: spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
In [8]: response = requests.get(spacex_url)
```

Check the content of the response

```
In [9]: print(response.content)
```


b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"sh
tps://images2.imgbox.com/94/f2/NN6Ph45r_o.png","large":"https://images2.imgbo
{"campaign":null,"launch":null,"media":null,"recovery":null},"flickr":{"small"
ebcast":"https://www.youtube.com/watch?v=0a_00nJ_Y88","youtube_id":"0a_00nJ_Y
196-spacex-inaugural-falcon-1-rocket-lost-launch.html","wikipedia":"https://e

Data Collection - Scraping

- Web scraping from the API was achieved through normalization of the .json data into a pandas data frame

- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/fbde66567e3e0d957073b9bb3fb010488e94b132/notebook_1_spacex-data-collection-api.ipynb

```python
# Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url)
data = pd.json_normalize(response.json())
```

Data Wrangling

- Data wrangling involved having an exploratory look at data of rocket launches. I investigated the different available launch sites, as well as different orbits available.
- I investigated the different outcomes of landing attempts and created a new column ‘Class’ based on this data. Bad outcomes were classified as 0 and good outcomes as 1. This was achieved through list comprehension.
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/b8ce59766d9ff75631f6dd72f3180e650de689c3/notebook_2_spacex-data_wrangling.ipynb

EDA with Data Visualization

- Several plots were used to visualize the data:
  - If launch sites and orbits changed with increasing flight number
  - If payload mass has any relation with the orbit of launch site
  - If different orbits have different success rates of landing
  - How the landing success rate increased over the years
- There exploratory efforts were made with the goal of finding curious data.
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/6e7956c307976b5983f5bb6e7805d9d4eb9a14de/notebook_3_spacex-eda-dataviz.ipynb

EDA with SQL

- I utilized SQL queries to explore some of the data regarding SpaceX's launches, including data about payload masses and booster versions. I also looked at data in different time intervals.
- Every query was simple and included the SELECT column FROM spacex-table. Many basic functions were used, such as AVG(), MAX(), MIN() or COUNT().
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/68442a0cbb69256bfecd62c4b73040d62a384de0/notebook_4_spacex-eda-sql.ipynb

Build an Interactive Map with Folium

- With the folium module for python, I created markers and circles to identify launch sites on the world map. I also added lines to display distances between launch sites and coastlines.
- The goal of this experiment was to have a better idea from where SpaceX is launching and to see the location of water bodies in proximity of launch sites, since many landings occur on the sea.
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/7165b93acf3092c5e24df781263b8bbac1603a8e/notebook_5_spacex-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

- Created a dashboard to make data visualization more interactive. Added plots showing successful landings depending on launch site, payload mass and booster type. Added interactivity to allow selection of specific launch site and payload mass range.
- These plots were used to explore the relationship between launch success rates and launch sites
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/281652812a1836462dcae92011290d9c5417e9a5/notebook_6_spacex_dash_app.py

Predictive Analysis (Classification)

- Created predictive models to see if landings can be predicted. Tried different machine learning models and used GridSearchCV to try out different parameters. Analyzed each model's success by calculating accuracy, score and a confusion matrix.
- Model types were logistic regression, decision tree, support vector machine and k nearest neighbors. Each model was fit with previously scaled and split training data and predicted the results for test data. The GridSearchCV method selected the best parameters for each model.
- https://github.com/pecko-g/IBM-Data-Science-Capstone/blob/3d98da0678ea19cbf28be0b08d5cd411128b2cee/notebook_7_spaceX_machine_learning_prediction.ipynb

Results

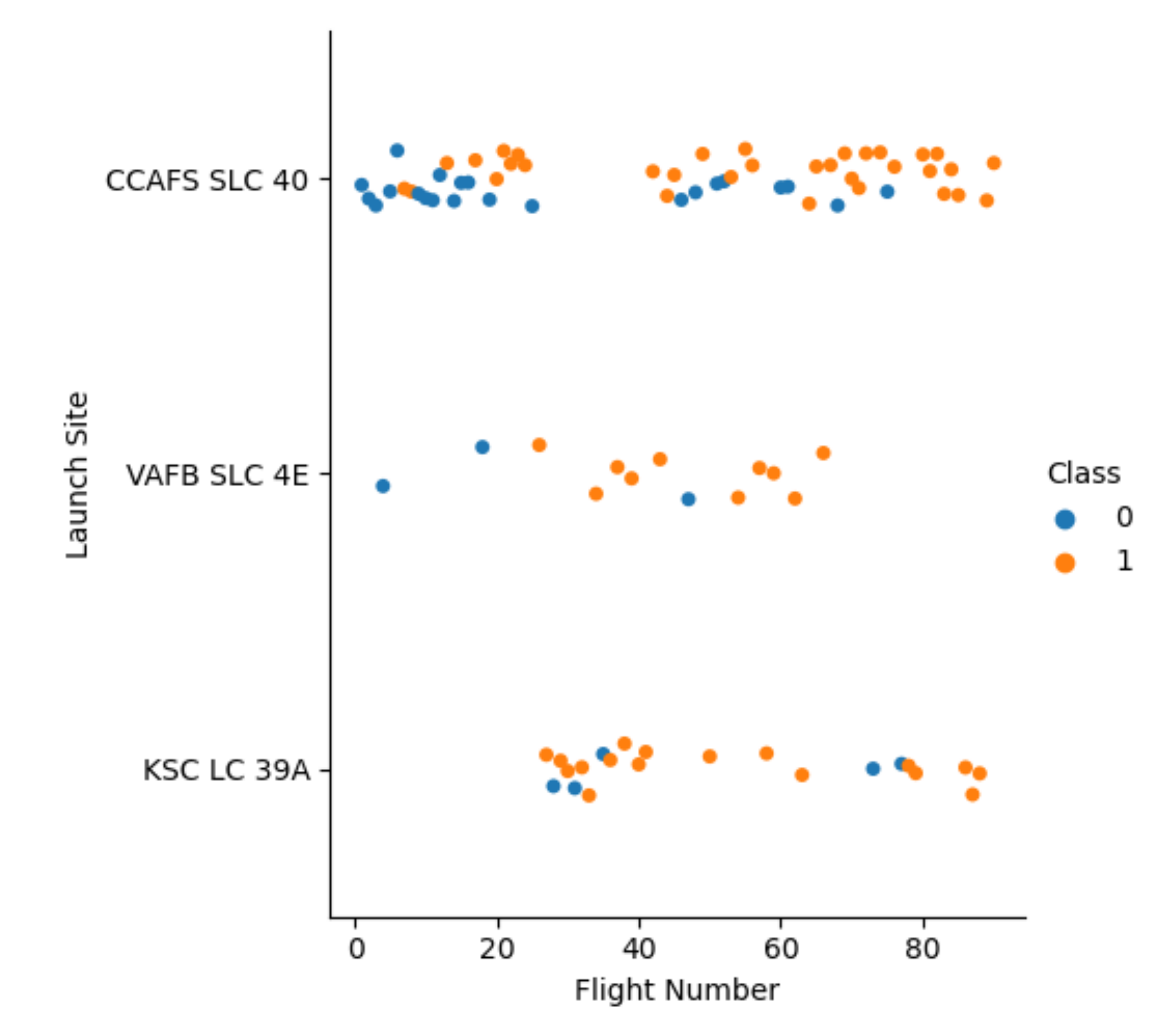
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

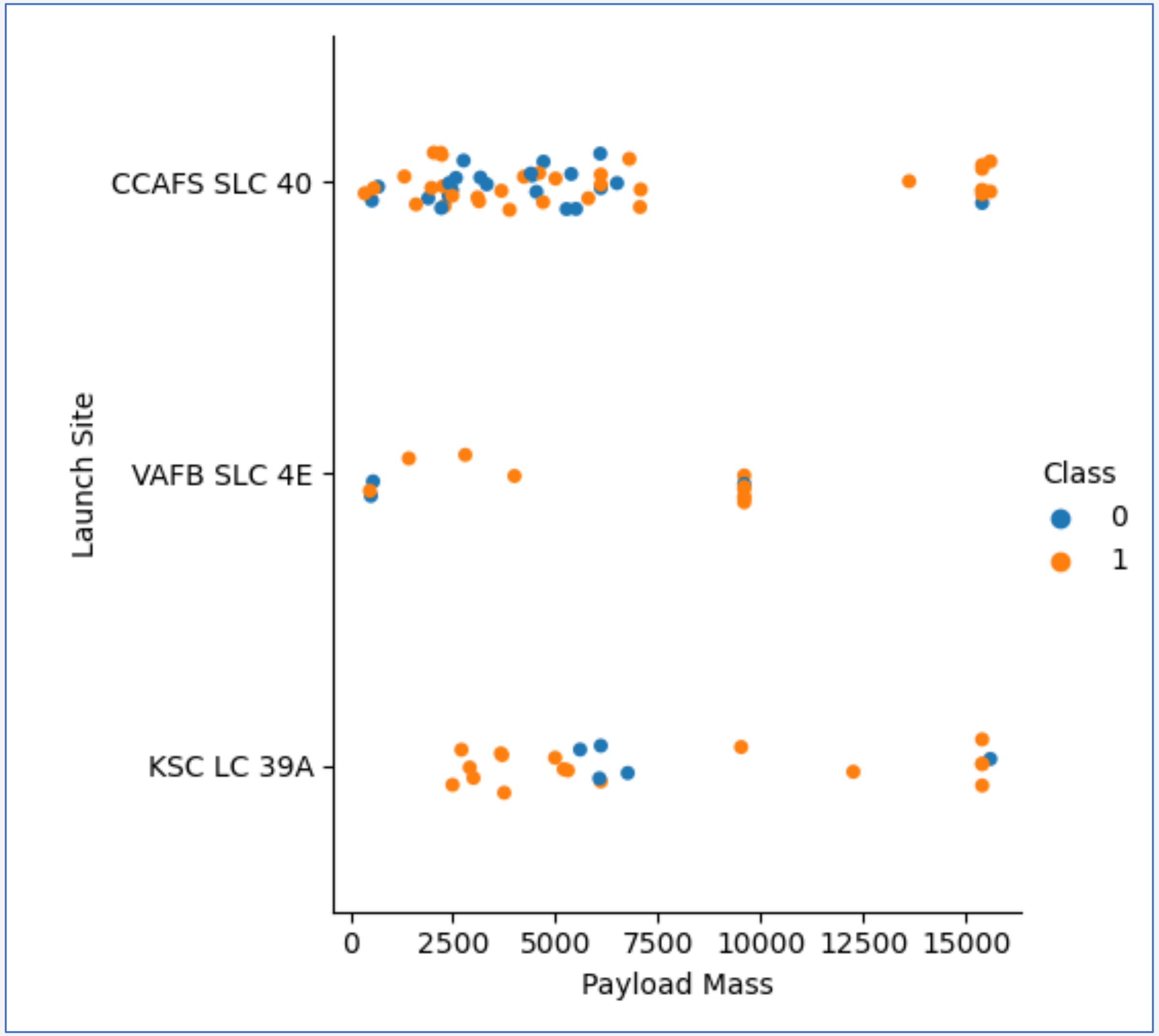
Flight Number vs. Launch Site

- We can observe that most launches came from launch site CCAFS SLC 40 and the fewest from VAFB SLC 4E.
- We can also observe that lately there haven't been any launches from VAFB SLC 4E and that launch site KSC LC 39A only started being used around flight number 30.



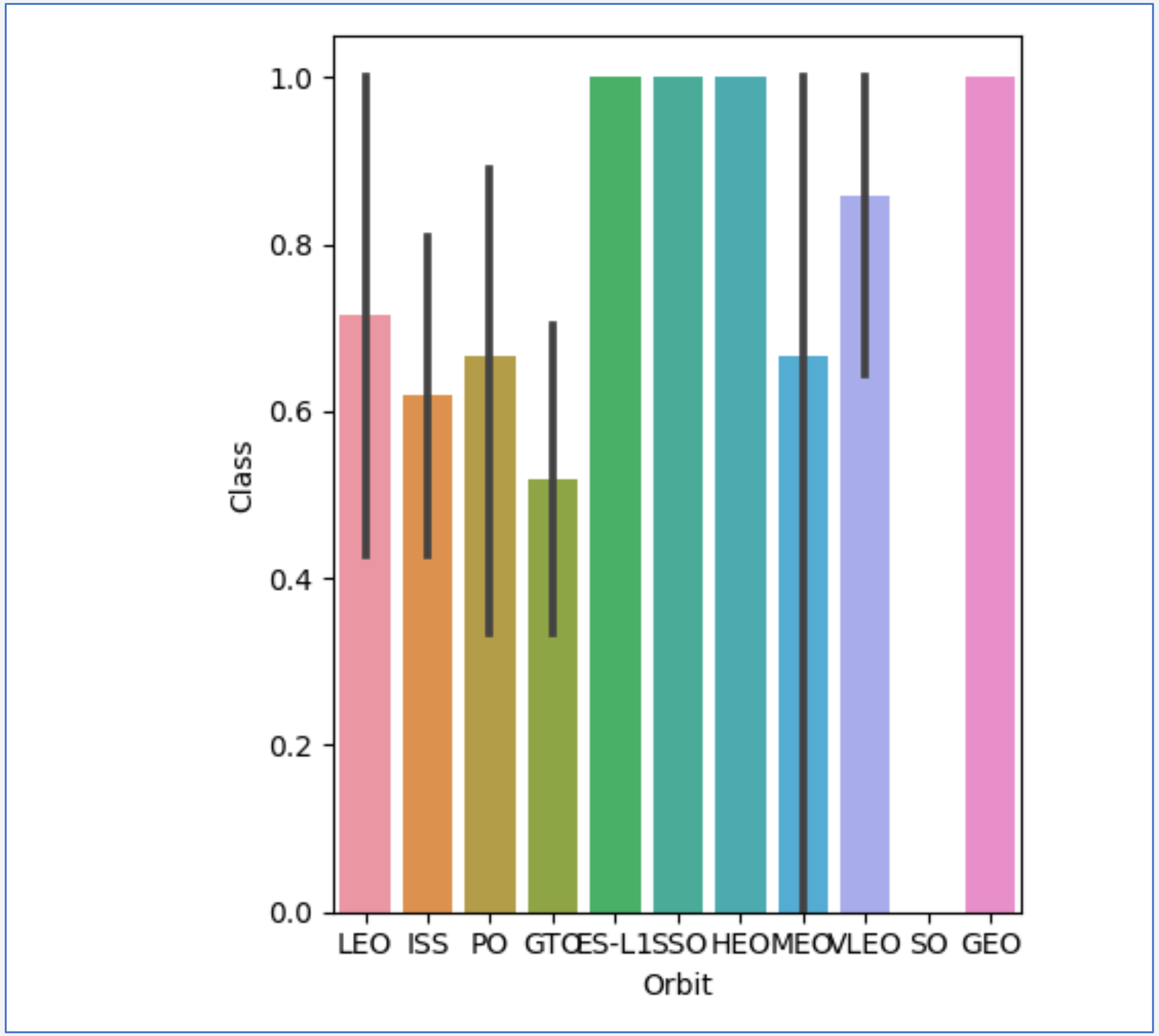
Payload vs. Launch Site

- Most payloads are under 7500 kg.
- Payloads over 10000kg don‘t launch from launch site VAFB SLC 4E.



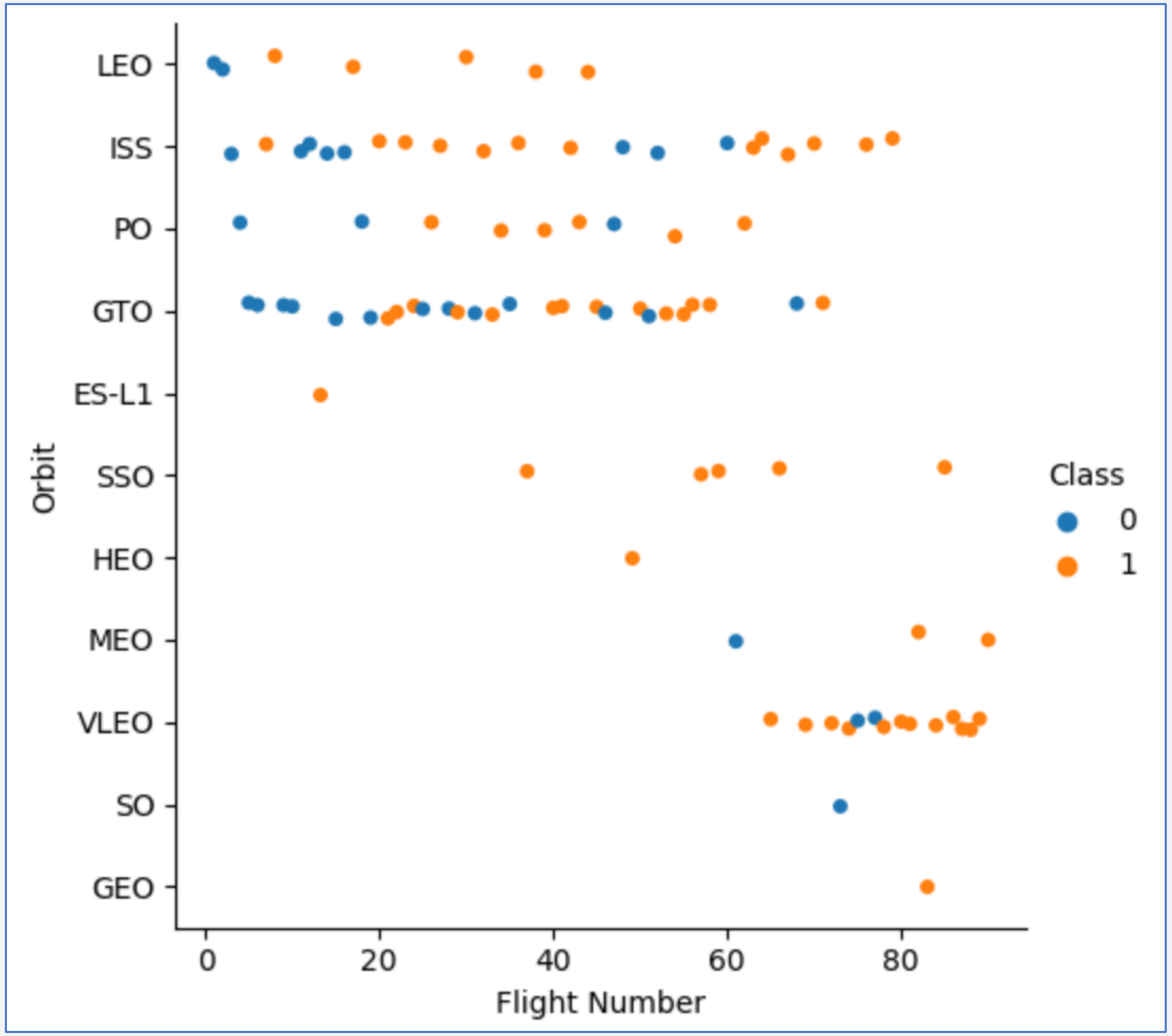
Success Rate vs. Orbit Type

- Orbits ES-L1, SSO, HEO and GEO have a landing success rate of 100%, while orbit SO hasn't had any successful landings.



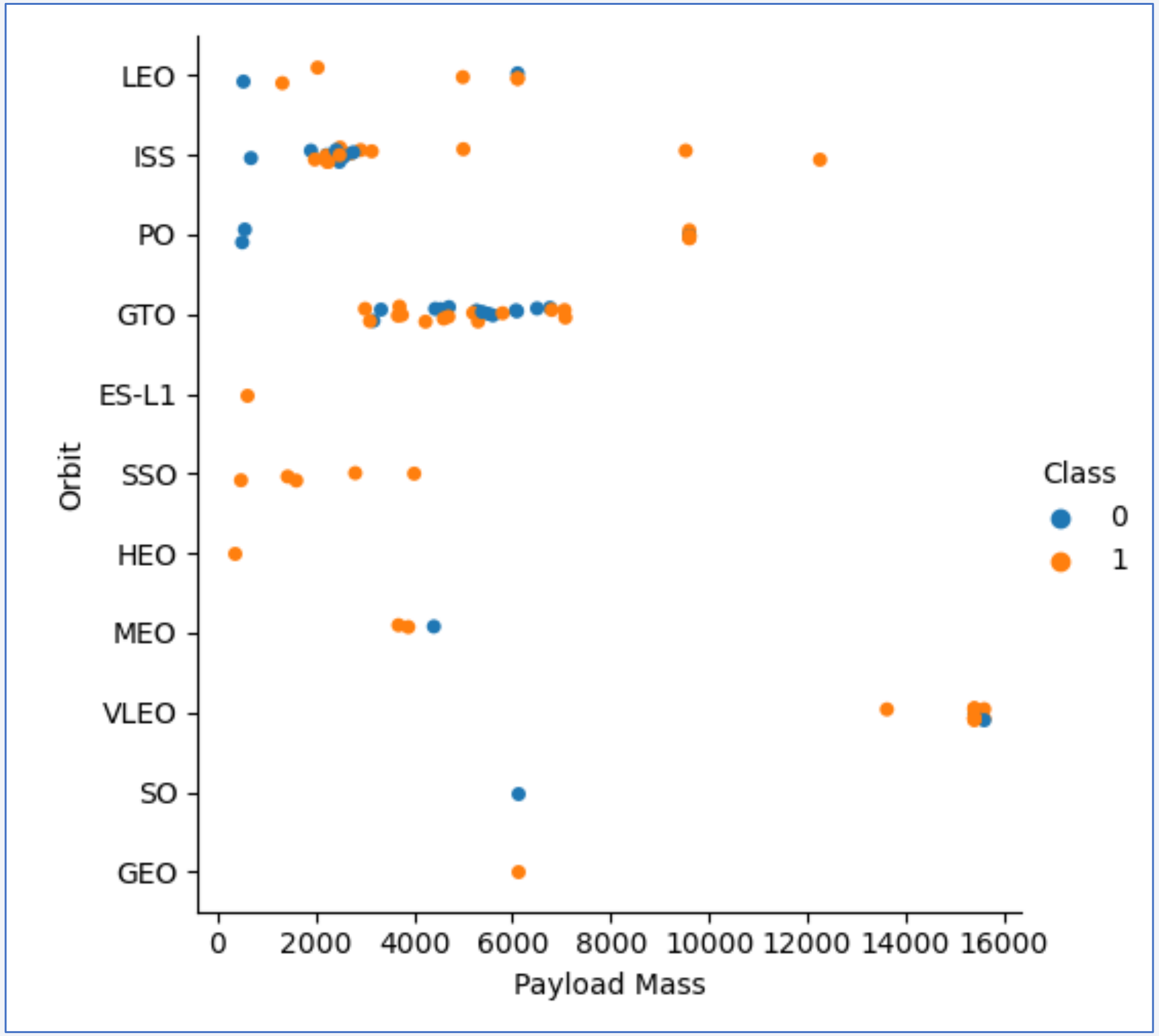
Flight Number vs. Orbit Type

- We observe that until around flight number 70 most launches were sent into orbits LEO, ISS, PO and GTO.
- Some orbits have only been used once.
- Lately the majority of launches goes into orbit VLEO.
- We also see that for orbit LEO, success increased with increasing flight number, but that is not the case for orbit GTO



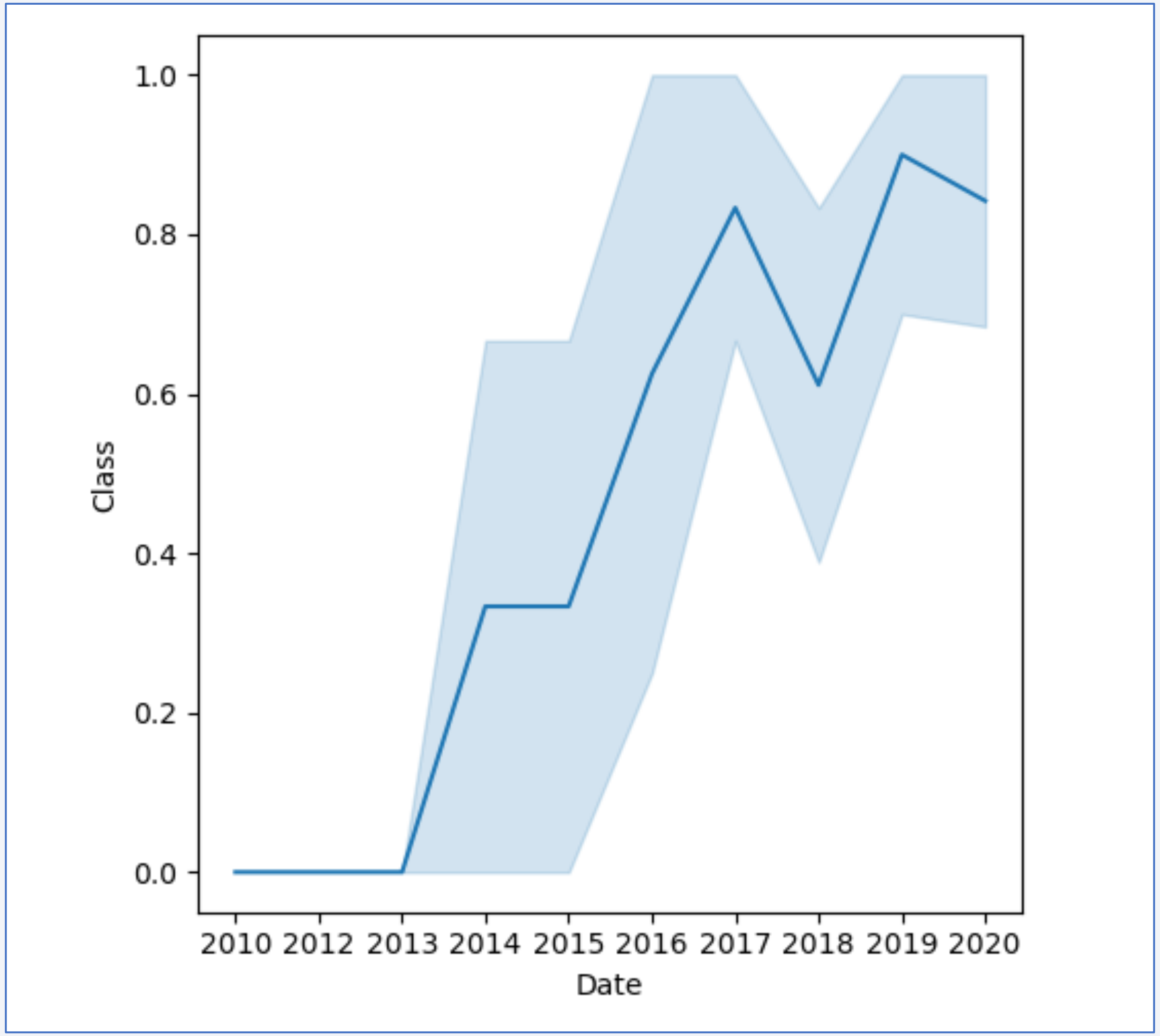
Payload vs. Orbit Type

- We observe that the heaviest payloads are sent to orbit VLEO.
- The launch to some orbits has greater success of landing if the payload is heavier, such as PO, LEO and ISS.
- The orbit GTO has no relationship between payload mass and landing success rate.



Launch Success Yearly Trend

- Landings have had increased success since 2013 and continue on a rising trend, although there was a small decrease in success in 2018.



All Launch Site Names

- Query used:
- SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E
- There are 4 distinct launch sites.

Launch Site Names Begin with 'CCA'

- Query used:
- SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
  - 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
- This query shows the first five launches from site CCAFS LC-40.

Total Payload Mass

- Query used:
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
  - 45596
- This query shows the total payload mass that was carried so far by SpaceX rockets.

Average Payload Mass by F9 v1.1

- Query used:
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';
  - 2928
- The average payload mass carried by rockets with booster version F9 v1.1

First Successful Ground Landing Date

- Query used:
- SELECT MIN(DATE) FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)';
  - 2015-12-22
- Query shows the date of the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query used:
- SELECT BOOSTER_VERSION FROM SPACEX WHERE (LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000);
    - F9 FT B1022
    - F9 FT B1026
    - F9 FT B1021.2
    - F9 FT B1031.2
- These are the boosters that successfully landed on a drone ship while carrying a payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- Query used:
- SELECT SUM(LANDING__OUTCOME LIKE 'Success%') AS LAND_SUCCESS, SUM(LANDING__OUTCOME LIKE 'Fail%') AS LAND_FAIL FROM SPACEX;
  - LAND_SUCCESS 61
  - LAND_FAIL 10
- Out of 71 launches, 61 landed successfully and only 10 failed to land. That is a success rate of 86%.
- This query presents a problem though, since I only considered outcomes with the words ‘success’ or ‘failure’, while there are other positive and negative outcomes as well.

Boosters Carried Maximum Payload

- Query used:
- SELECT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
  - 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
- These are all the booster versions that carried maximum payload weight.

2015 Launch Records

- Query used:
- SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE (DATE LIKE '2015%' AND LANDING__OUTCOME LIKE 'Fail%');
  - LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
  - Failure (drone ship) F9 v1.1 B1012 CCAFS LC-40
  - Failure (drone ship) F9 v1.1 B1015 CCAFS LC-40
- There were 2 failed landing attempts in the year 2015. Both rockets were launched from CCAFS and attempted landing on a drone ship. Booster versions were different.

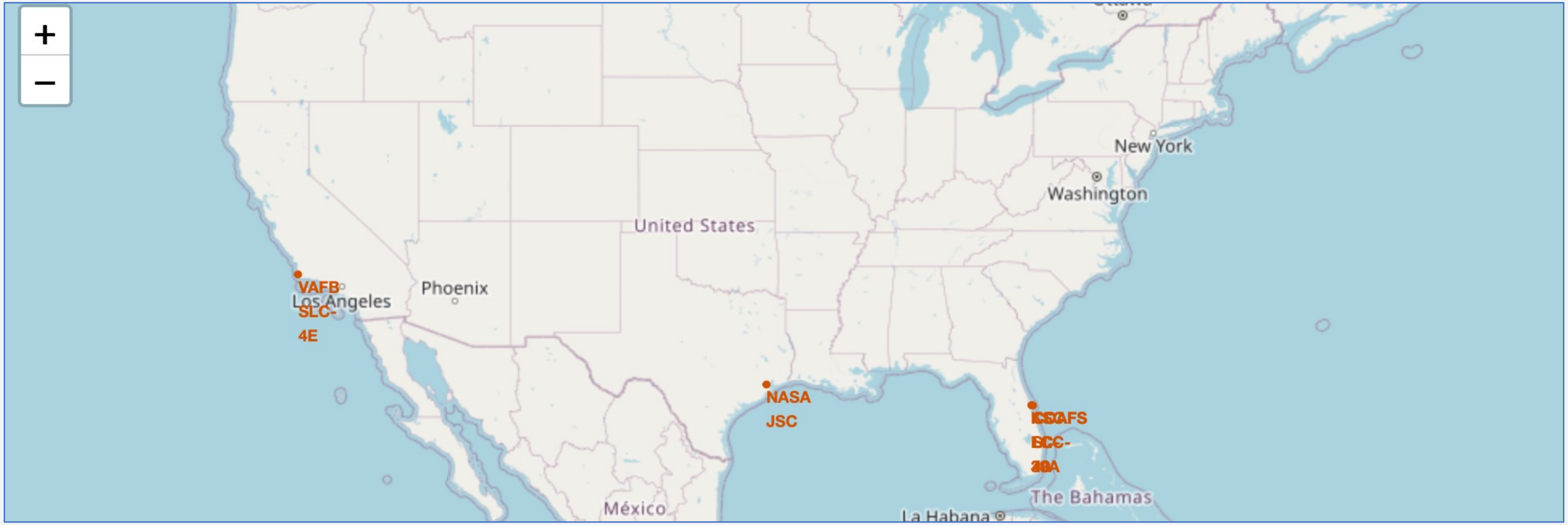
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Query used:
- SELECT LANDING__OUTCOME,COUNT(LANDING__OUTCOME) FROM SPACEX WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(LANDING__OUTCOME) DESC;
  - 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
- The most frequent landing outcome was ‘no attempt’ with 10 flights, followed by ‘failure (drone ship)’ and ’success (drone ship)’ with 5 each.

Section 3

Launch Sites Proximities Analysis

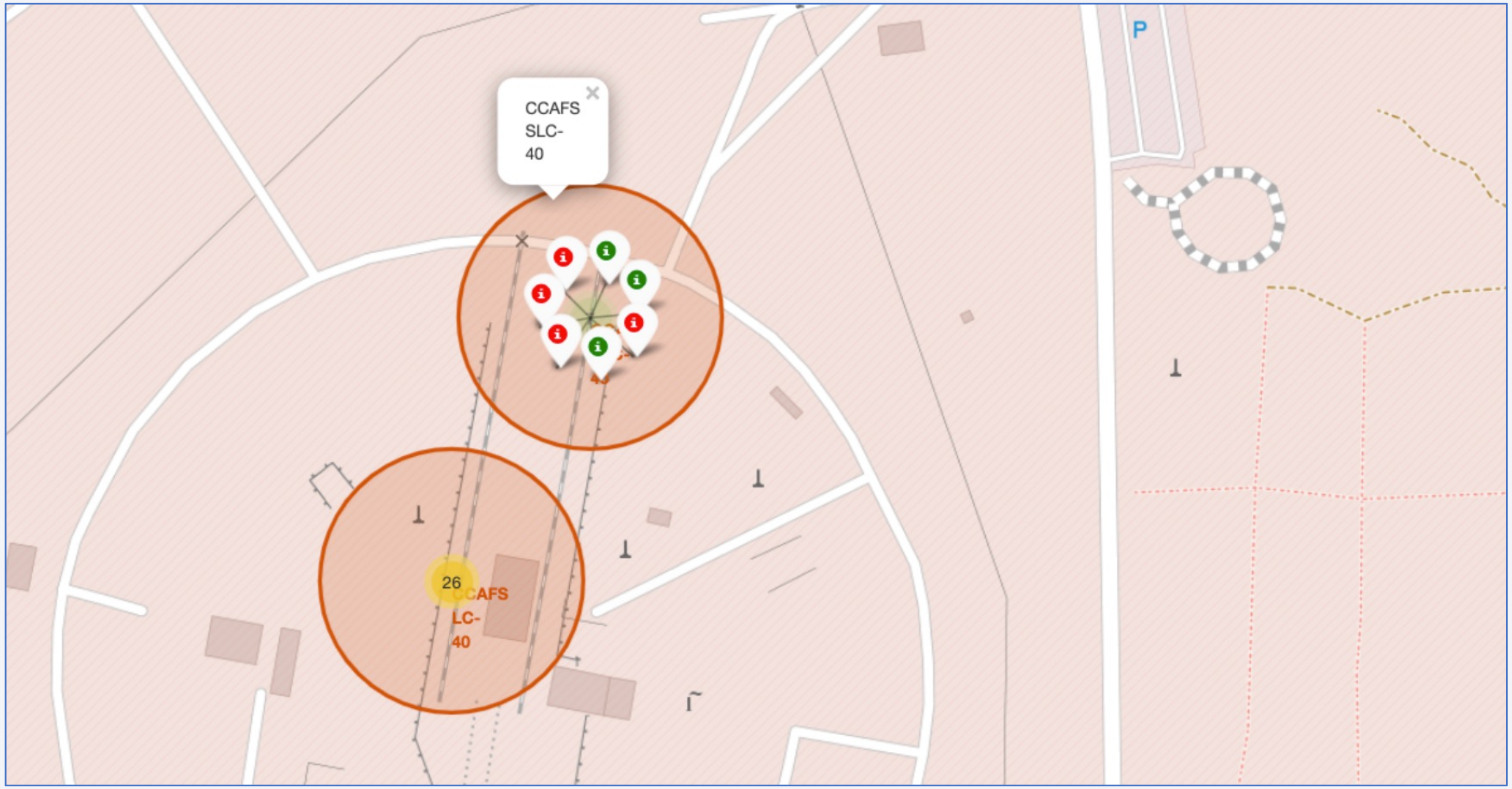
SpaceX Launch Sites



- This map shows the location of the SpaceX launch sites, as well as the launch site from NASA for comparison. We can observe that the launch sites are in the US States of Florida and California.

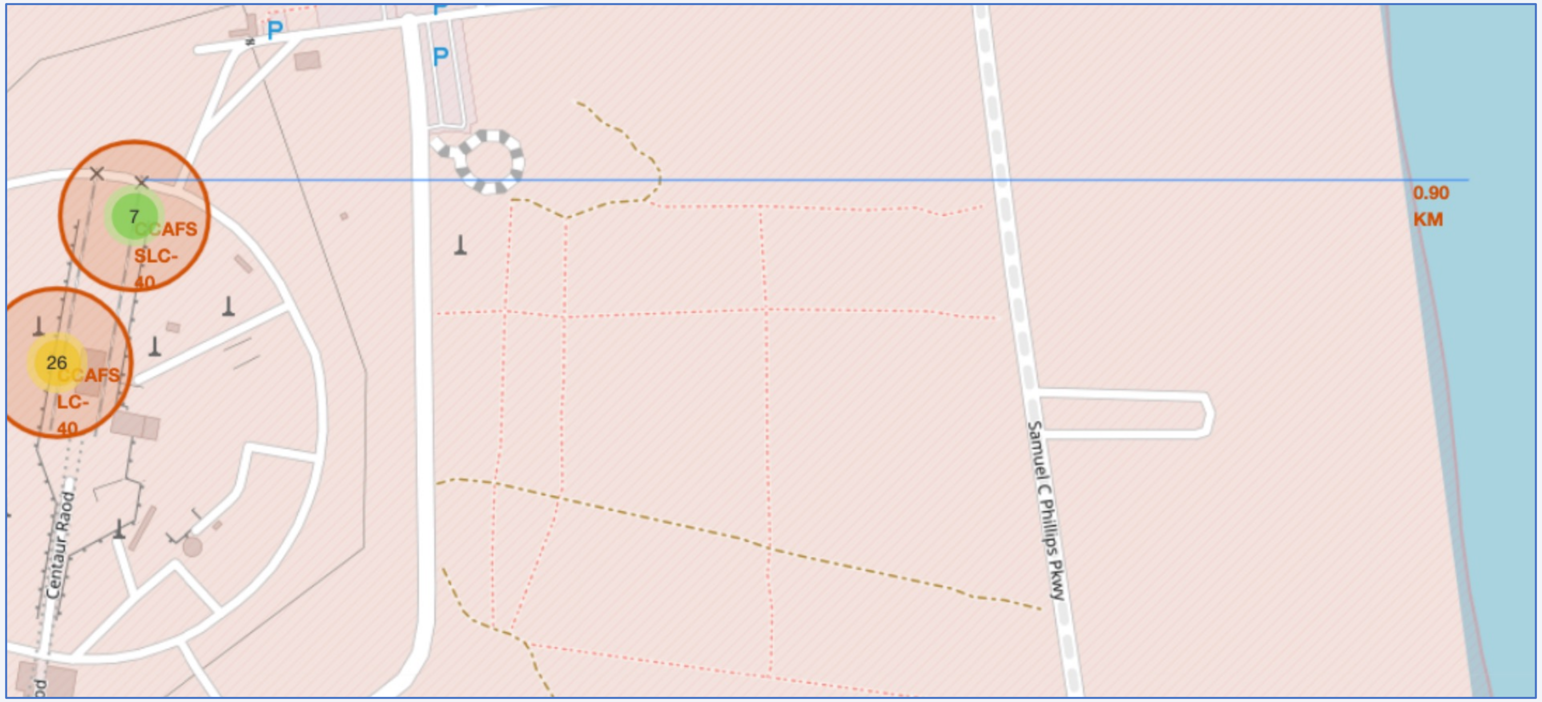
Launch Site Dependent Landing Success

- This image shows the different landing outcomes (positive vs negative) coming from launch site CCAFS SLC-40.
- Unfortunately, I couldn't get my map to work and have no other data to show.



Proximity of Launch Sites to Coastline

- This image shows the closest coastline to the launch site CCAFS SLC-40, with the respective distance.
- Unfortunately, I couldn't get my map to work and have no other data to show.
- Proximity of SpaceX's launch sites to the coastline is important, since many landings occur on the sea.

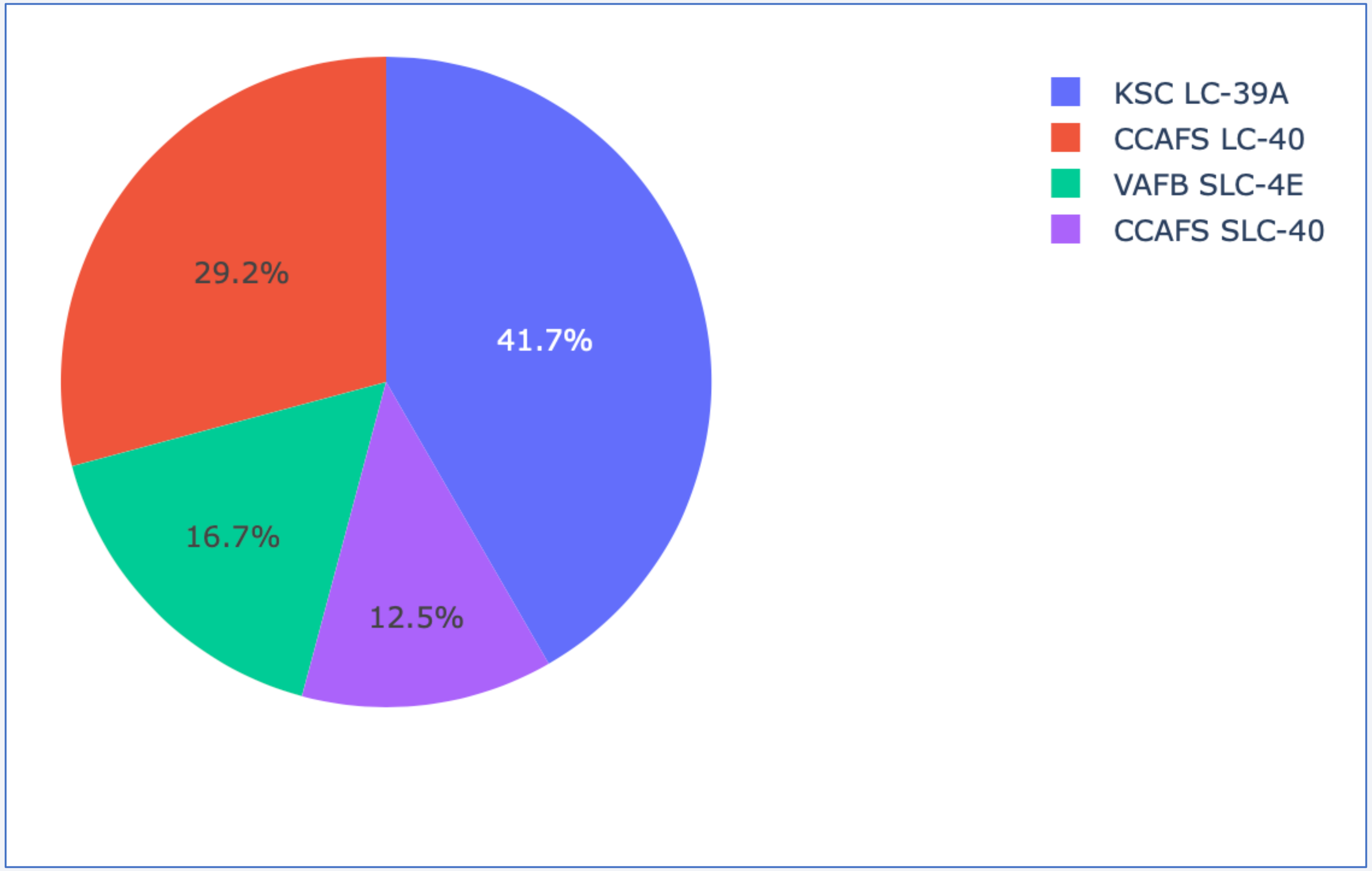


Section 4

Build a Dashboard with Plotly Dash

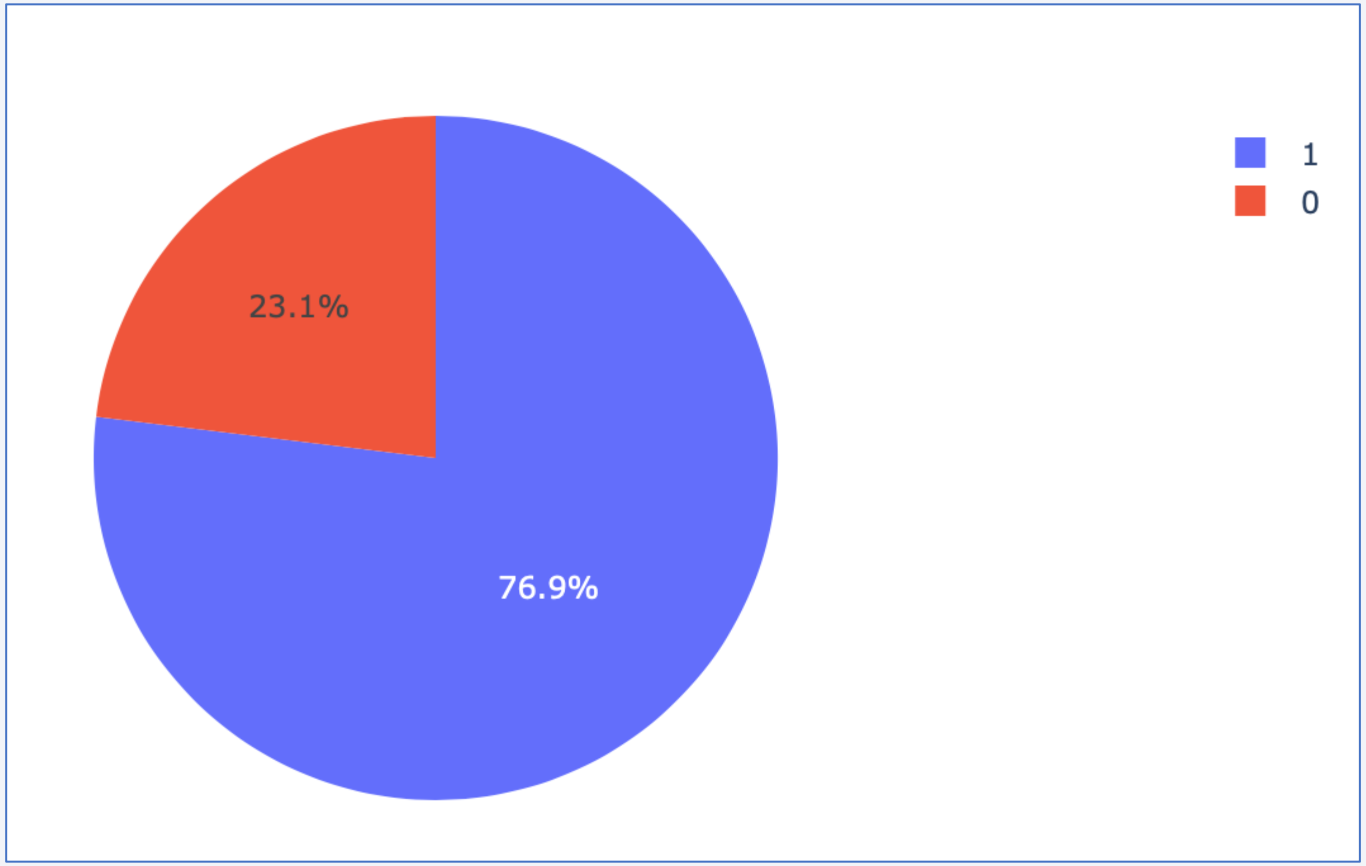
Successful Landings for All Launch Sites

- This pie chart shows us that most successful landings come from launches at site KSC LC-39A.



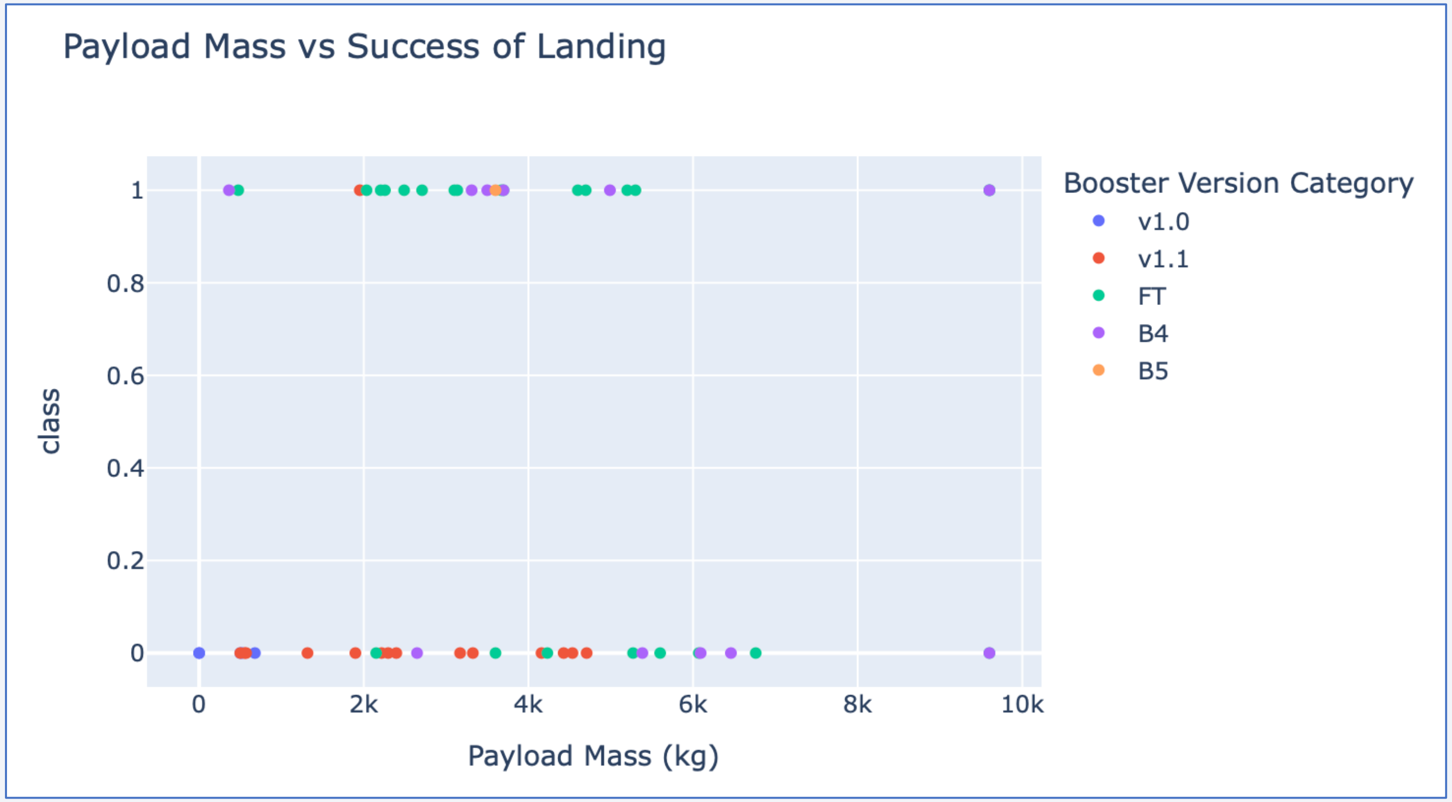
Successful Landings from Launches at KSC LC-39A

- When we zoom in on the launch site KSC LC-39A, we discover that 77% of landings were successful.



Payload Mass vs Success of Landing

- This scatter plot shows the payload mass and the success rate of the landings for all launches.
- 1 is a successful landing and 0 is a failed landing.

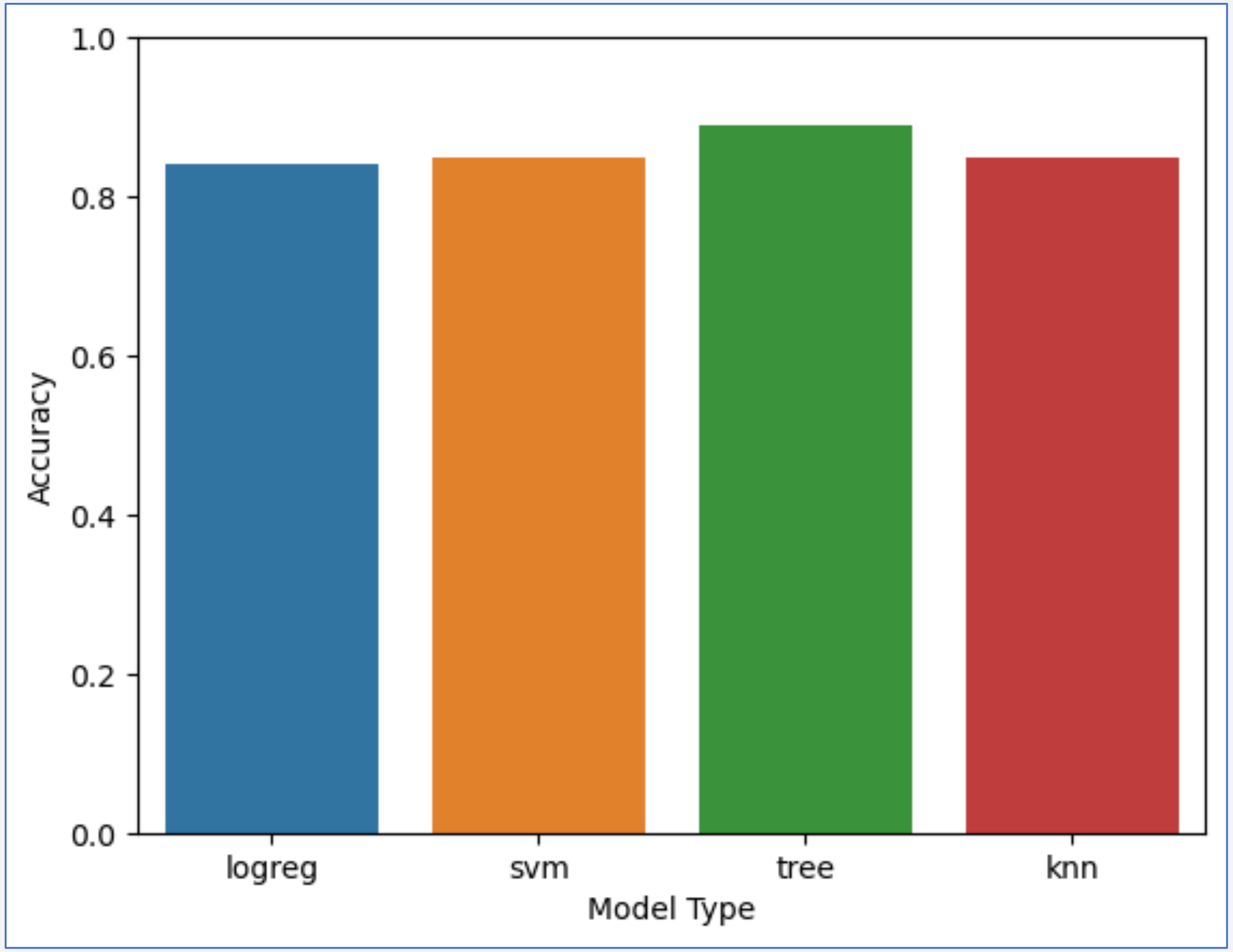


Section 5

Predictive Analysis (Classification)

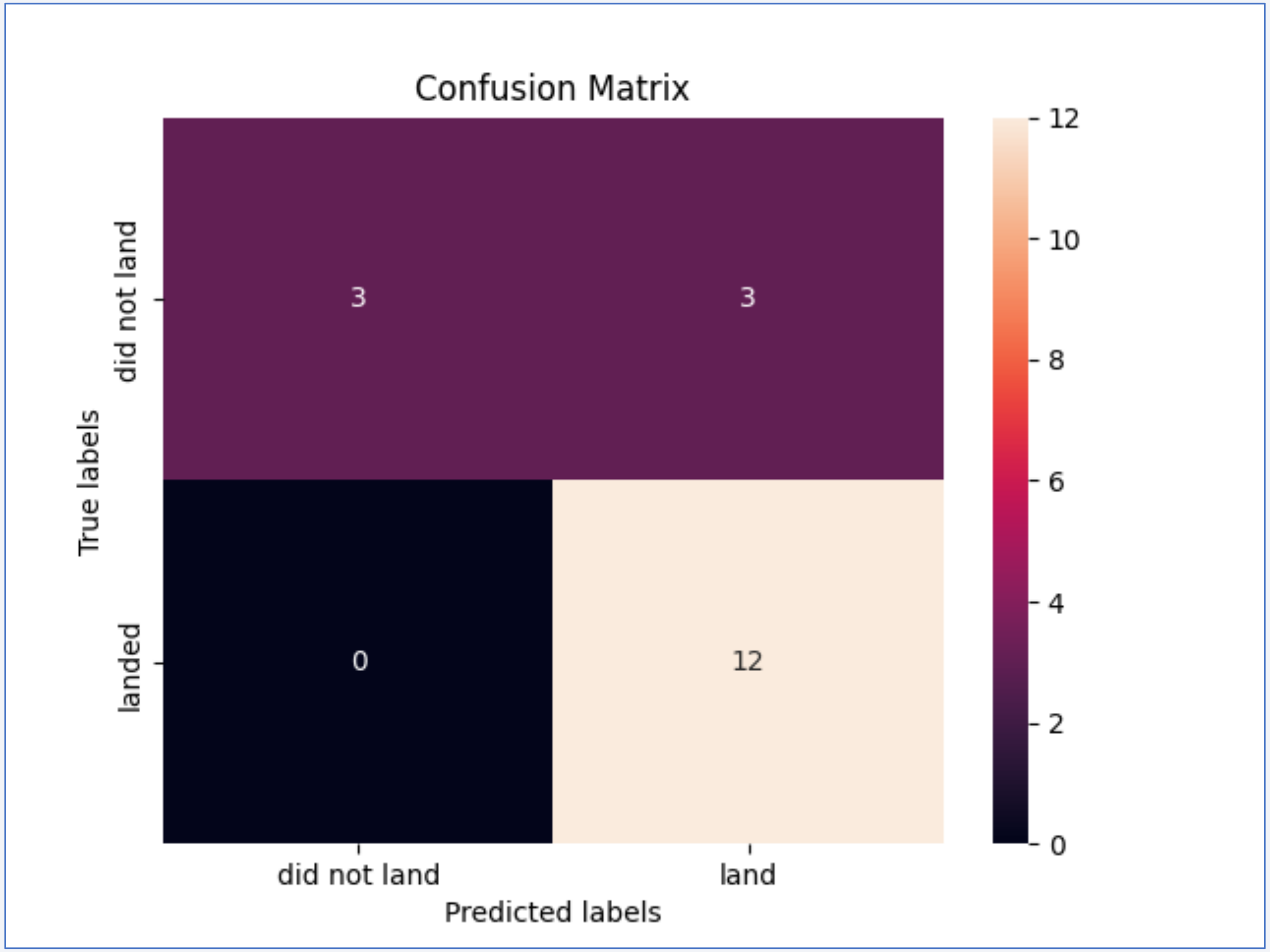
Classification Accuracy

- While all models had very similar accuracy, just over 80%, the decision tree model was slightly better.



Confusion Matrix

- This is the confusion matrix for our selected decision tree model.
- The model has predicted in total 15 successful and 3 failed landings.
- The model has correctly predicted all successful landings, but has falsely predicted 3 failed landings as successful, so it predicted 3 so-called false positives.



Conclusions

- SpaceX is one of the strongest players in the current space race. It revolutionized the industry by constructing rockets that can land after being used and thus can be reused, saving a lot of resources and money.
- The success of the rocket landings has increased since 2013 and so far, achieved a peak of 90% success in 2019.
- There are 4 different launch sites and several different orbits that have been used by SpaceX rockets.
- For some of the orbits, the success rate of the landing of the rocket is greater than for others, although some orbits are used much more than others.
- Payload mass seems to have no impact on whether the rocket will land or not and a big range of payload weights have been transported.

Thank you!