IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zhongran Ou
10/11/2021

Outline

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

Executive Summary

- The goal of this project is to predict the successfulness of a rocket landing, as well as gaining insights about factors which may affect the outcome of a rocket launch, by studying into past launch data from Space X.
- After analyzing our data, we are confident that we have build an accurate model to predict the outcome of a rocket landing and have gained comprehensive insights that will help our future launches.

Introduction

- As we are entering commercial space age, Space Y wish to be the pinnacle among competitions.
- The reason we set up this project, is to gain information which will help us understand what makes a successful rocket landing.
- Our main problem is that we do not have an existing way to predict rocket landing for our company.
- To do this, we have studied through previous launch data from our competitor Space X, and we have found fairly useful information.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - We collect our data from official Space X data REST API.
- **Perform data wrangling**
  - We reformat and reselect our received data, create a new dataframe for better exploration and analysis.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - We create graphs and queries to further explore our data.
- **Perform interactive visual analytics using Folium and Plotly Dash**
  - Gather geographical data for better understanding on the location of each launch site.
- **Perform predictive analysis using classification models**
  - We use several machine learning algorithm to build our final predictive model.

Data Collection

- We collect our data from the Space X launch data REST API, which contains useful information we need for prediction.
- We initially store our data as csv, then convert it to a dataframe to be used on Jupyter Notebook for better data processing.

- Flow chart next page…

Data Collection – SpaceX API

- API collection work flow:

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Get request

↓

Receive response in JSON

↓

Normalize JSON file to dataframe

↓

Specify formats (date)

↓

Select columns we want

↓

Create new dataframe

↓

Ready for wrangling

Data Collection - Scraping

- Wiki page web scrapping work flow:

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/jupyter-labs-webscraping.ipynb

Get request to Space X Wiki URL

↓

BeautifulSoup4 to identify HTML tabs.

↓

Find all ‘tables’

↓

Iterate through ‘<th>’

↓

Get columns and column names

↓

Convert to dataframe

Data Wrangling

- We explore our collected data as follows:
    1. Explore the counts of each column, its count, as well as the amount of empty entries (null).
    2. Explore the types of orbits we have in our data.
    3. Recognize which columns indicates the landing successfulness.
    4. Based on the columns ‘Outcome’, we generate a new binary column ‘class’, for 1=success otherwise 0=fail for each row.
    5. Get the overall successfulness in percentage = 0.66
    6. Store newly processed data for later analysis.

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To determine the effects of each factor on landing outcome, we have plotted:
    - **Flight Number VS Launch Site** ------ explore relationship effect on outcome
    - **Payload VS Launch Site** ------ explore relationship effect on outcome
    - **Success rate VS Orbit type** ------ explore effect of orbit on outcome
- We discovered that orbit type seems to have huge impact on landing outcome, hence we further explore possible relational effects caused by orbit type:
    - **FlightNumber VS Orbit type** ------ explore relationship effect on outcome
    - **Payload VS Orbit type** ------ explore relationship effect on outcome
    - **Launch success VS Year** ------ explore yearly trend

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

To further explore our data, we have ran several queries:

- Get unique launch sites we have in our data.
- Get 5 record with launch sites start with ‘CCA’
- Get total payload mass launched by NASA (CRS)
- Get average payload mass by Falcon 9 v1.1
- Get dates which ground pad landing were successful
- Get boosters which successfully landed on drone ship while carrying 4000 to 6000 payload mass.
- Total number of missions
- Boosters that have carried the maximum payload mass.
- Get all failed landing on drone ship, along with its booster ver and launch site in 2015.
- Rank the number of outcome type between 2010-06-04 to 2017-03-20.

GitHub Link: https://github.com/simonou99/spacex_landing_prediction/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- To study geo-effects on launch sites, we have created several Folium maps:
  - Show all launch sites ------ explore locational distribution
  - Mark outcome for each launches ------ explore effects of launch sites
  - Calculate distances between launch sites and proximities ------ explore if proximities alter the effect of launch sites on landing outcome

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We have also created a dashboard to show percentage and distribution of landing successful rate:
  - Pie chart showing percentage of successful rate of each launch site.
  - Individual pie chart to show the percentage of outcome for each launch site.
  - Scatter plot for class against payload mass, for showing the effect of payload mass on successful rate.
  - Individual scatter plot for class against payload mass.

GitHub Link:

https://github.com/simonou99/spacex_landing_prediction/blob/main/dash1.py

Predictive Analysis (Classification)

- After exploring our data, we have decided to use **Logistic Regression, KNN, SVM** and **Decision Tree** algorithms to build our model.
- We first cast our dataframe into numpy array for performing machine learning.
- We then perform standardization on our data which reassign X in to floats and numbers.
- We train-test split our data into 80% train data and 20% test data.
- Finally we run each algorithm and create confusion matrixes to compare the accuracy of each model.
- We have finally decided that the **Decision Tree** model will be our **best predictive** model with average accuracy=0.85 and **accuracy=0.88 with the best estimation parameter**:

  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random’}

GitHub Link:
https://github.com/simonou99/spacex_landing_prediction/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

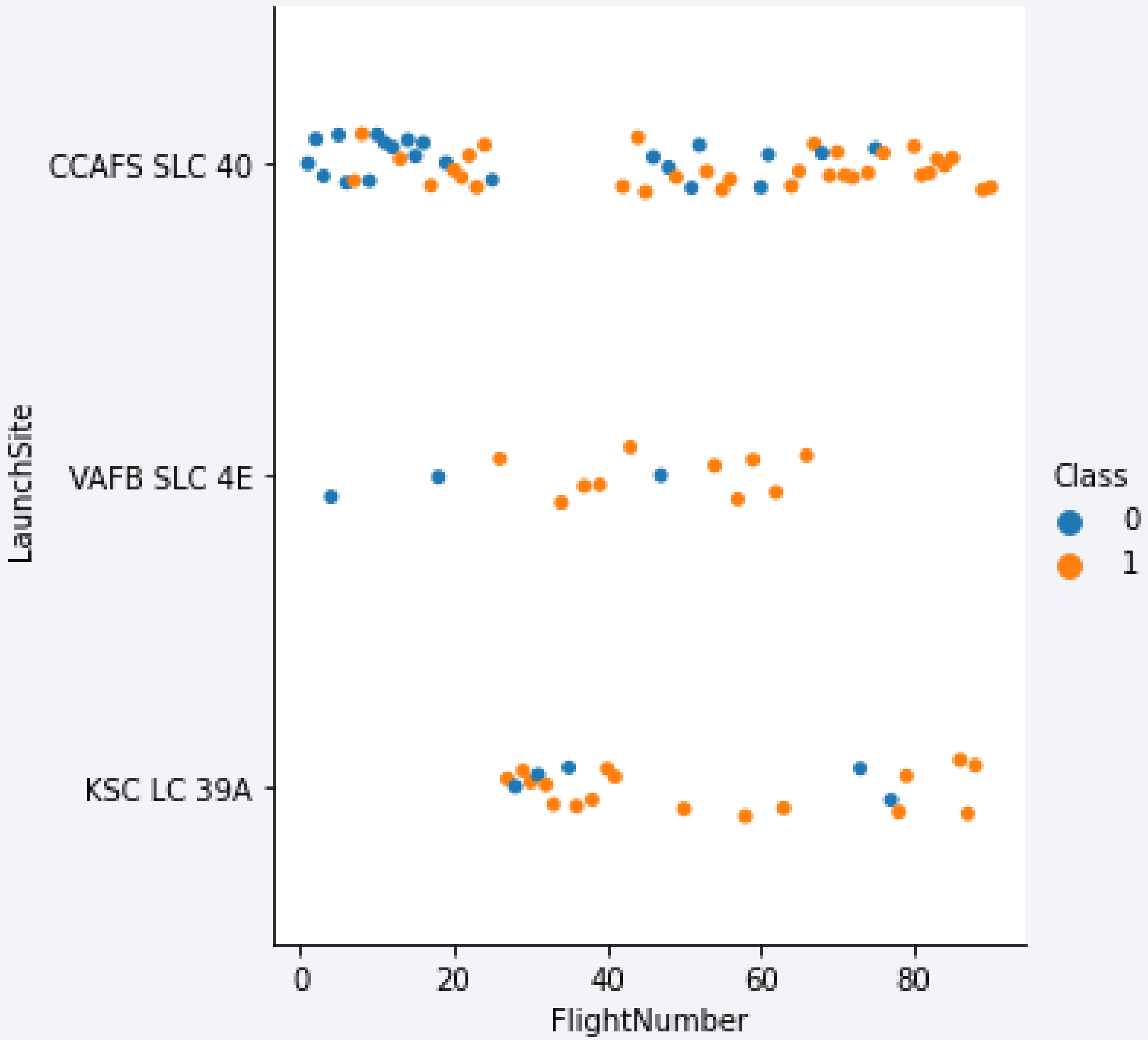
- Following Pages:
    - Exploratory data analysis results
    - Interactive analytics demo in screenshots
    - Predictive analysis results

Section 2

Insights drawn from EDA

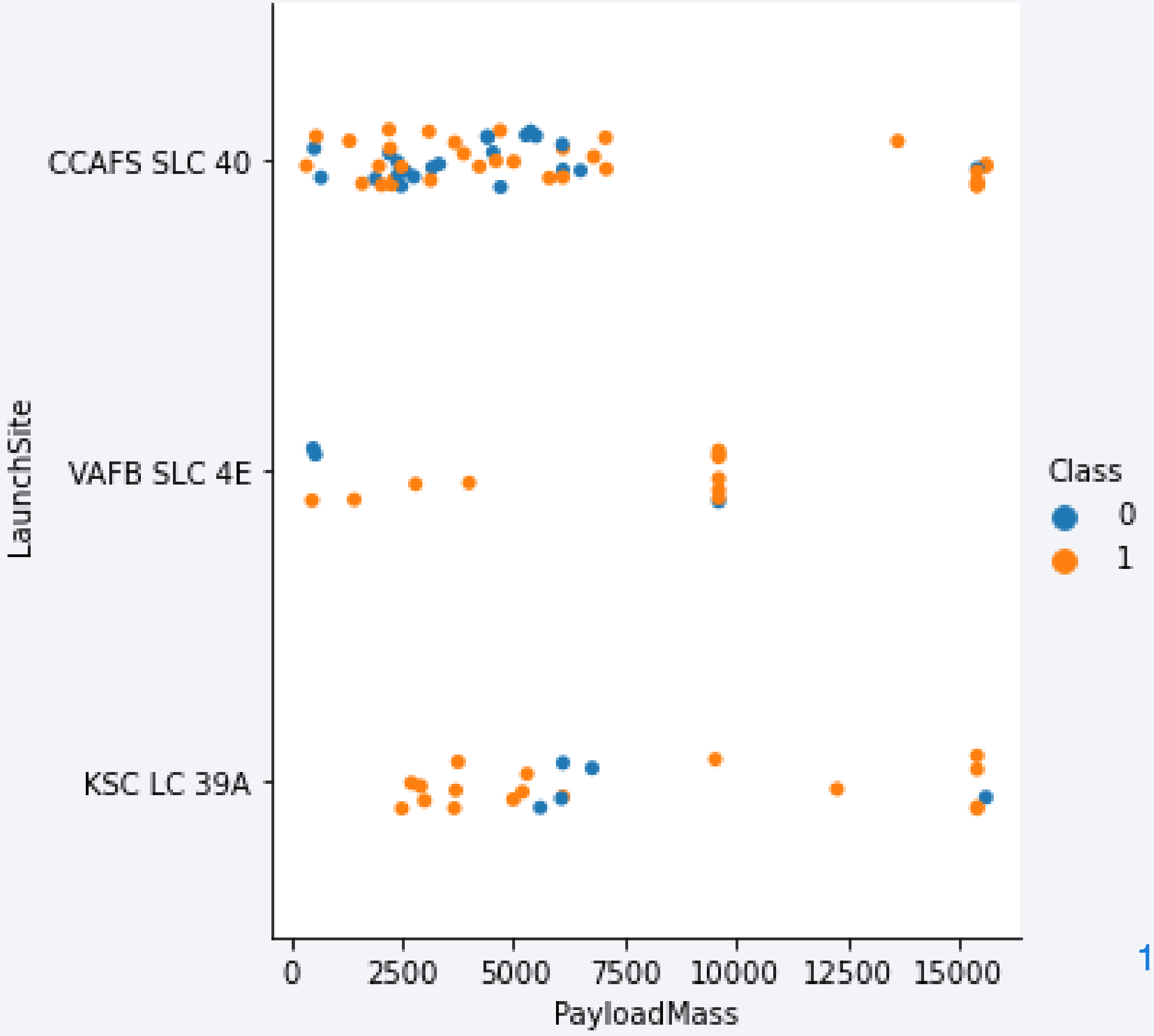
Flight Number vs. Launch Site

- From the plot, there seems to be little to no correlation with landing outcome.
- However, do notice that CCAFS SLC 40 have the most failed landing records. Also, VAFB SLC 4E seems to have the best success to fail ratio, but there are less records than the other 2 launch sites.



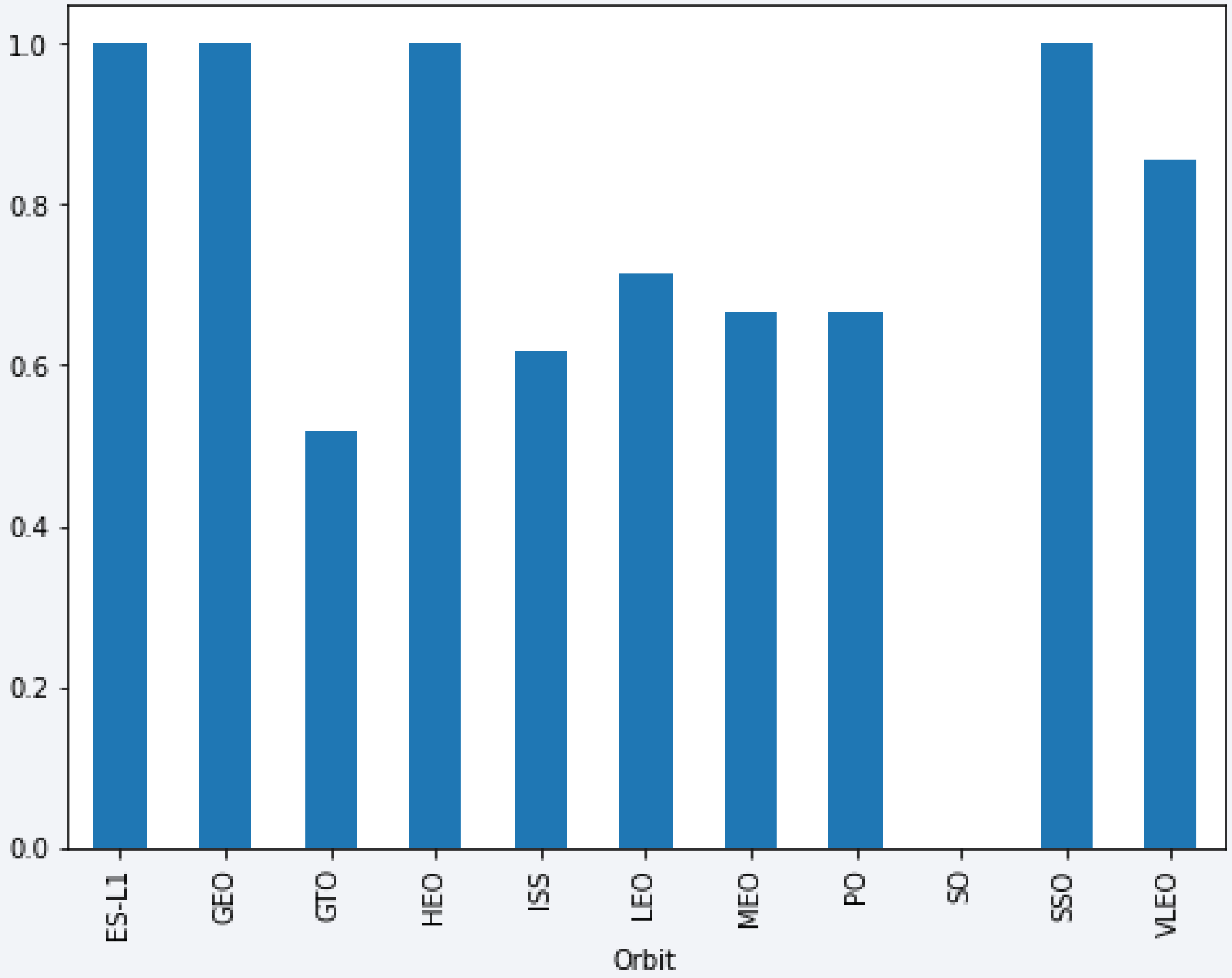
Payload vs. Launch Site

- There are significant clustering behavior in the plot.
- There are significantly more successful landing records for PayloadMass above 9000.
- Landing at KSC LC 39A seems to perform better if the payload mass is less than 5000.
- CCAFS SLC 40 seems to perform better if payload mass is above 7000.



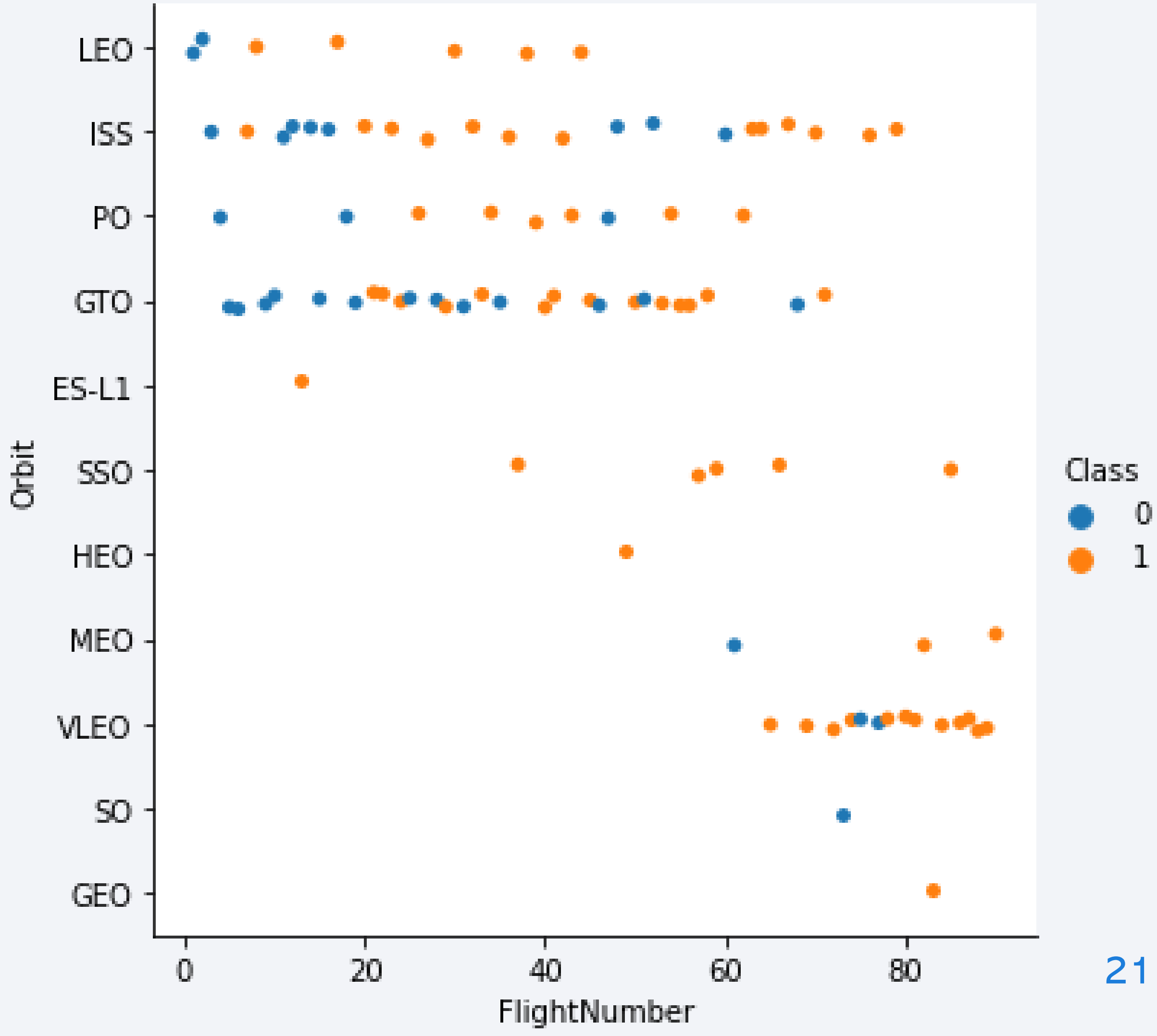
Success Rate vs. Orbit Type

- Orbits: ES-L1, GEO, HEO and SSO seem to have perfect successful rate.
- SO has a success rate of 0.
- Other orbits may have a success rate between 0.5 and 0.7.



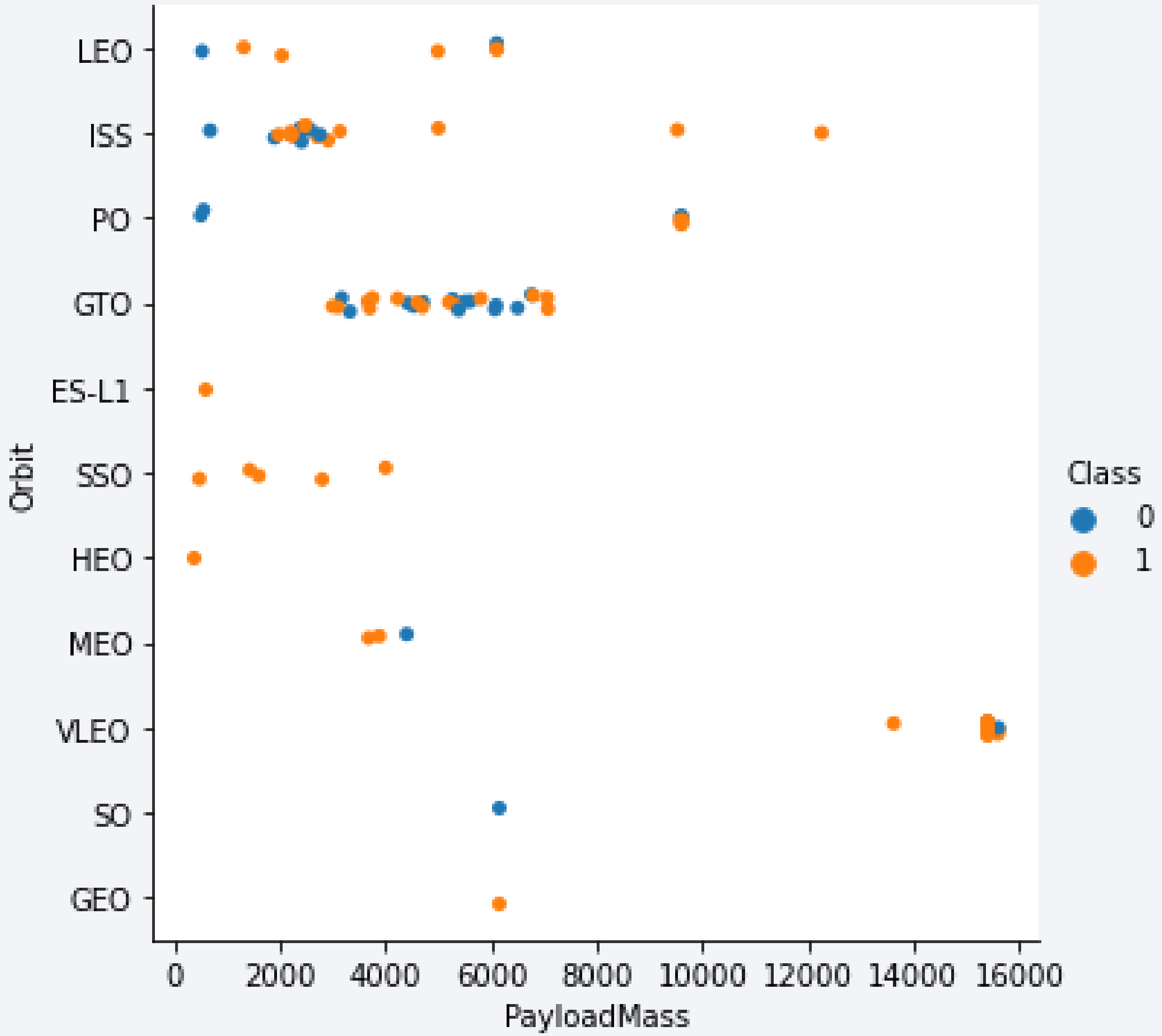
Flight Number vs. Orbit Type

- Successful rate seems to decrease significantly if FlightNumber is below 20 and below 40 if the orbit is GTO.
- Othewise, the relationship FlightNumber vs orbit shows little effect on landing outcome.



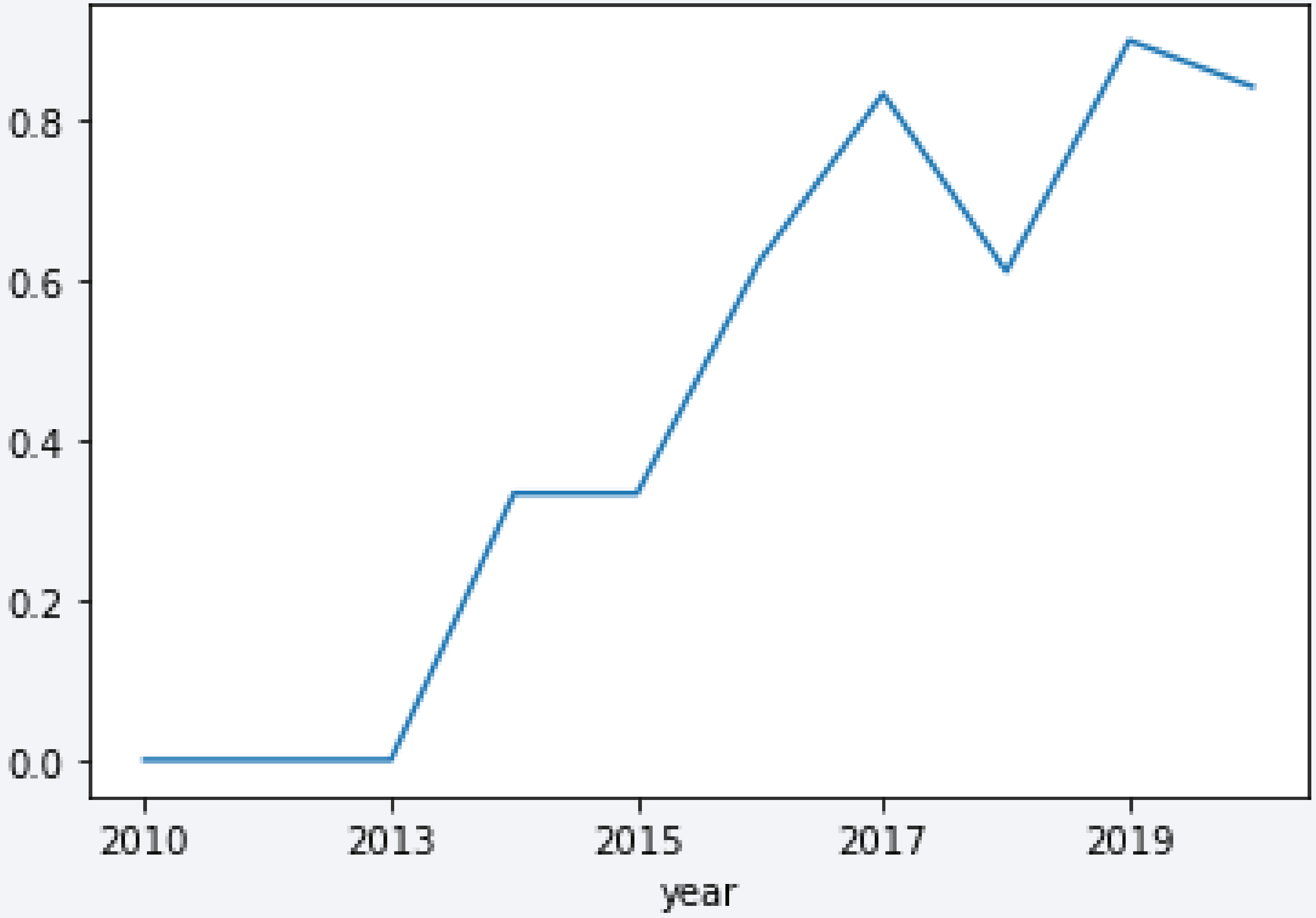
Payload vs. Orbit Type

- For LEO, ISS and PO, success rate seems to increase with payload mass, and has almost 0 success rate at small payload mass below 1000.
- Compared to previous plots, ES-L1, SSO and HEO do not seem to be able to launch great payload mass.



Launch Success Yearly Trend

- Overall trend shows us we have better success rate as time goes on.
- However, there is a huge drop in success rate at around 2018.
- We also have a decreasing success rate after 2019 to the end of record, which we may need to be caution.



All Launch Site Names

- We have 4 unique launch sites in the data.

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

Launch Site Names Begin with 'CCA'

- Query shows that CCA series launch sites have always successful mission outcome, but no successful landing outcome.

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

total_payload_kg_by_nasa_crs
45596

Average Payload Mass by F9 v1.1

average_payload_by_f9_v1p1
2534

First Successful Ground Landing Date

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are in total 4 drone ships that match the condition.

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

Total Number of Successful and Failure Mission Outcomes

- There are in total 100 successful mission outcomes with only 1 failed mission.

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are in total 12 boosters which had carried the maximum payload.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- There are 2 failed landing records in the year 2015, with both launched at the same site CCAFS LC-40.

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

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

- The most outcome category is ‘No attempt’ which there are 10 records.
- The least will be ‘Precluded’, and there is only 1 record.
- There are in total 8 successful and 7 failed landing records within the selected time.

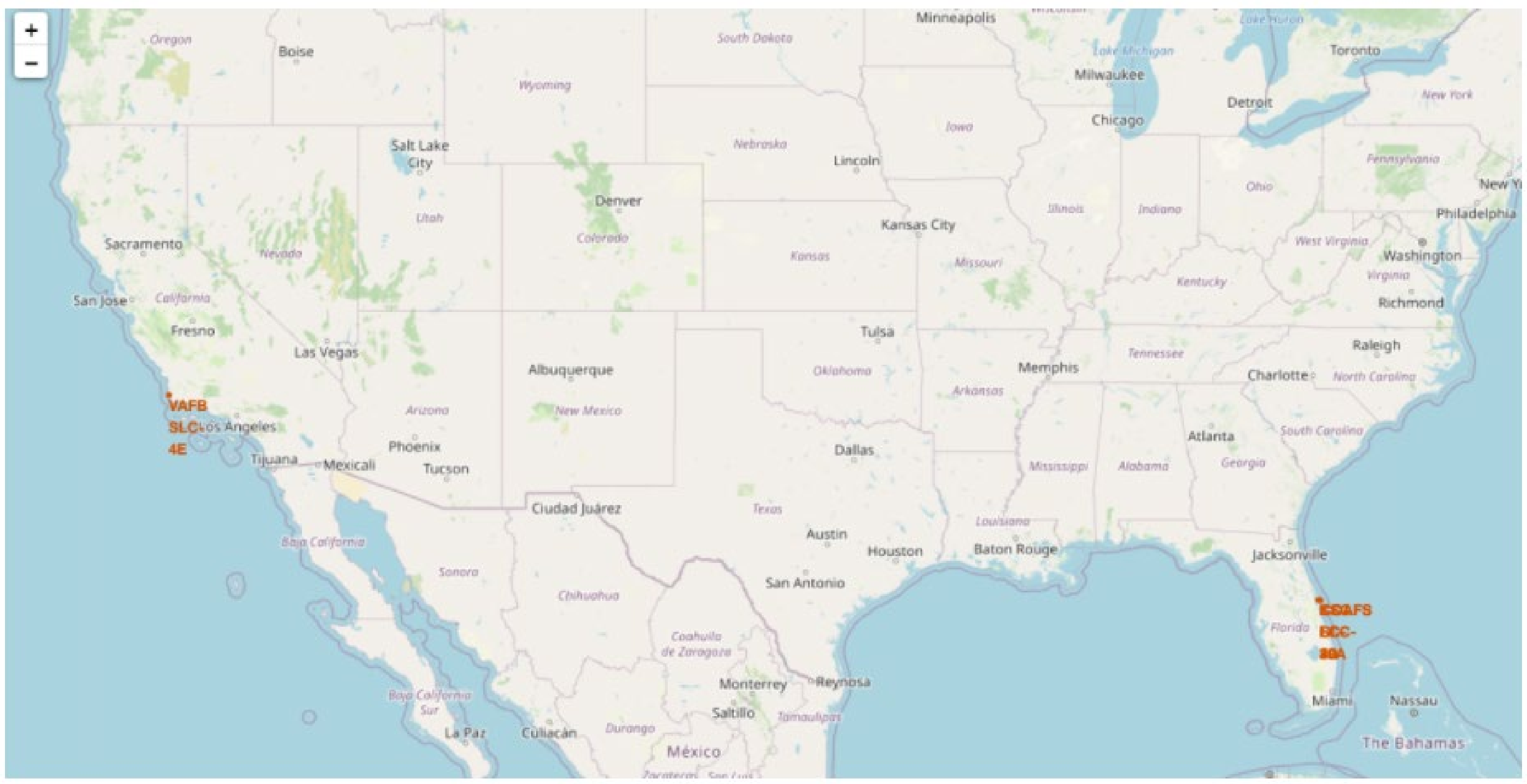
landing__outcome	num_of_landings
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

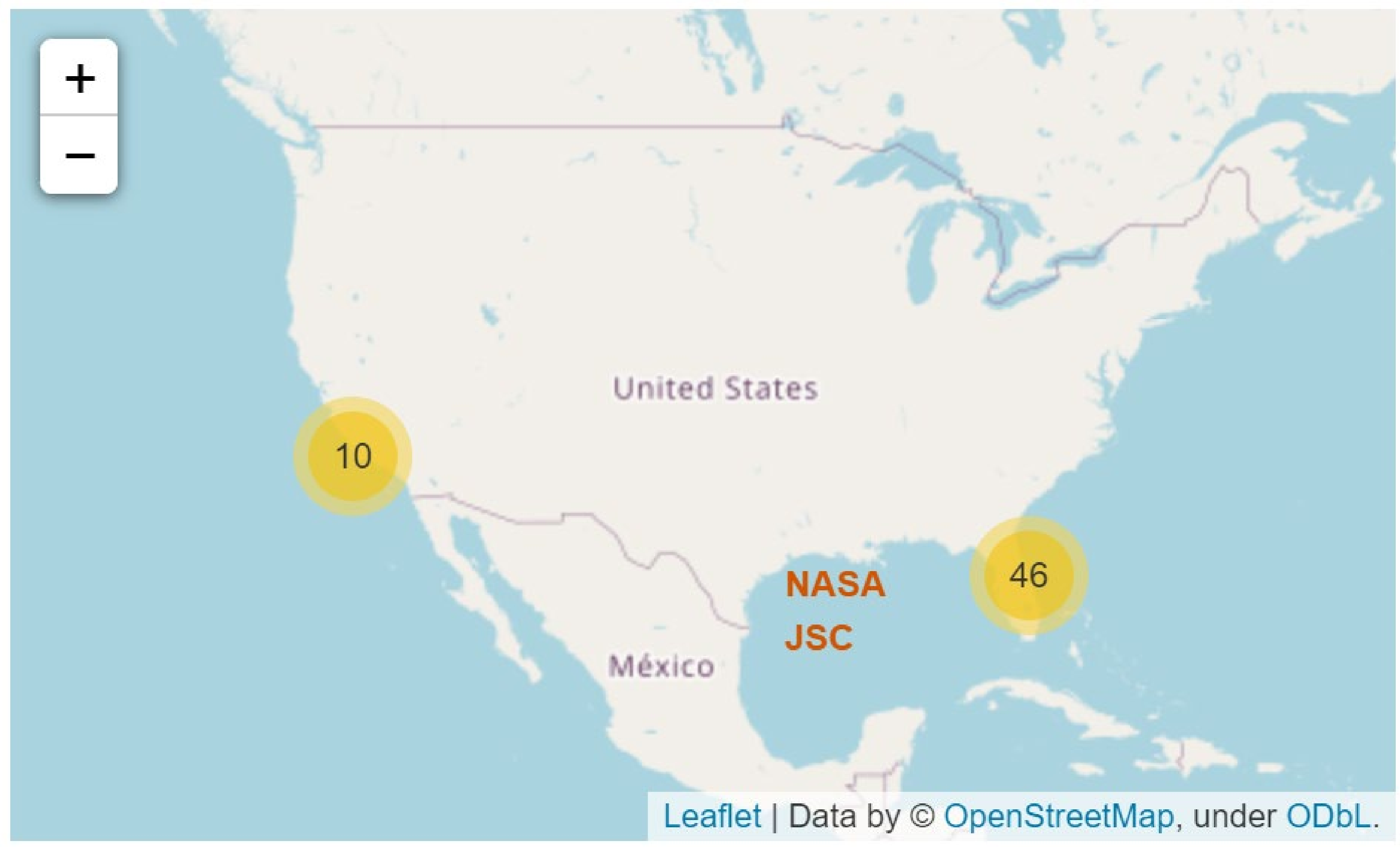
Map Location of All Launch Sites

- Launch sites are generally built close to the equator line.
- There is only 1 site at the west coast, while most are located at the east coast in Florida.



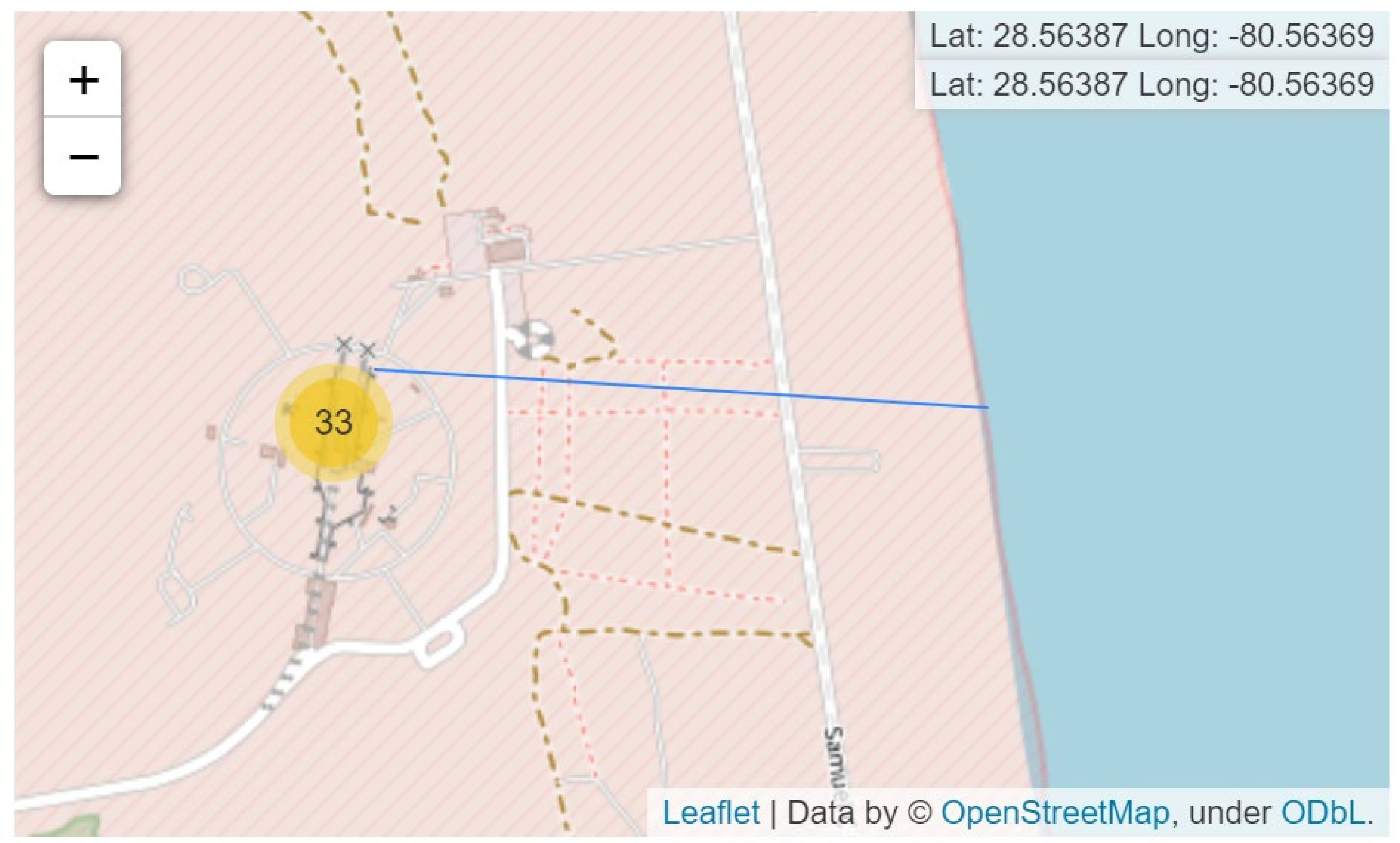
Map Locations of Landing Outcomes for Each Site

- There are 46 records in the east coast and 10 in the west.
- For further exploration with zoom-in function, please visit the GitHub page.



Proximity Distance to Launch Site on Map

- Launch site 7 with the closest proximity (coast).
- With distance = 1.0 km

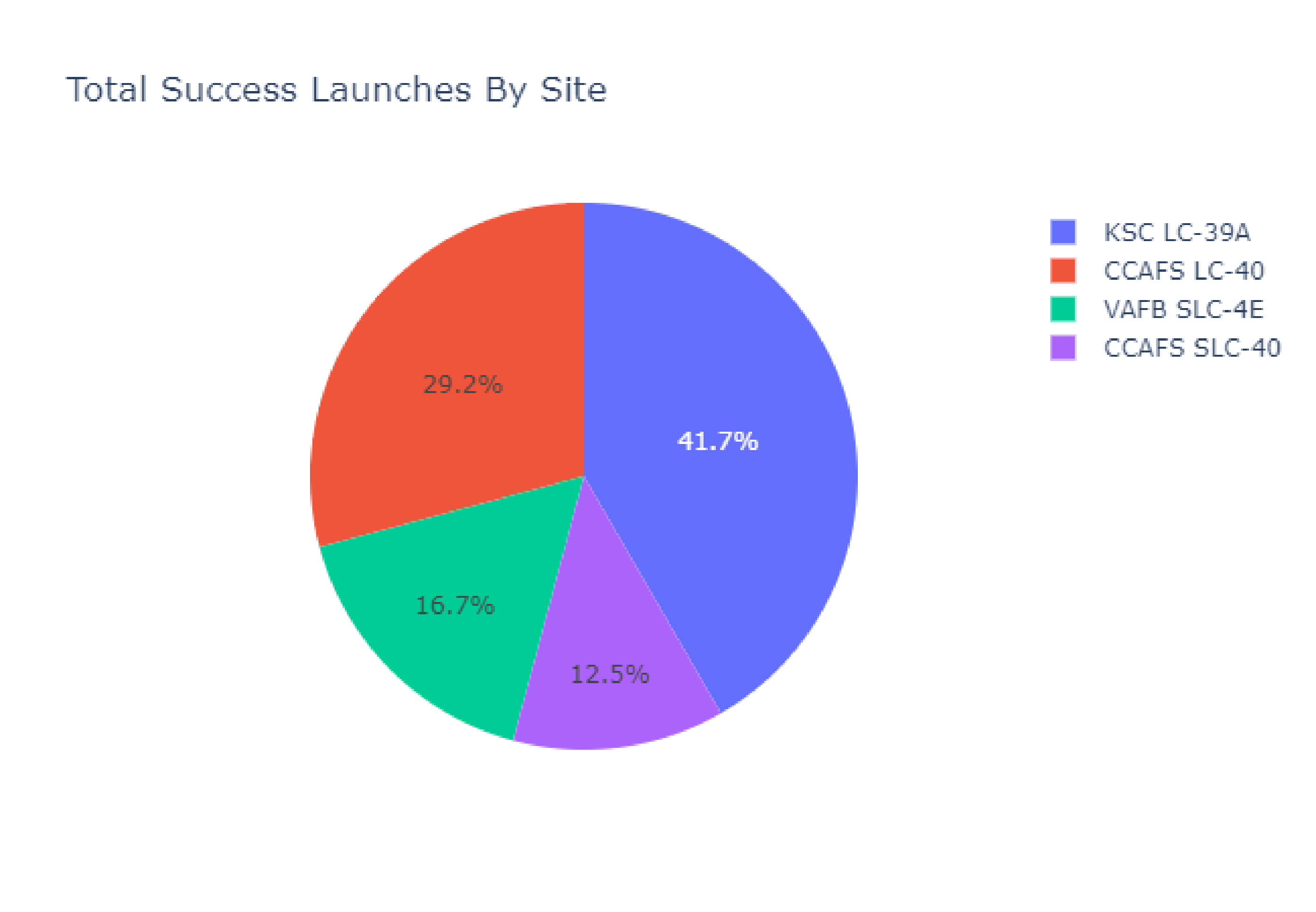


Section 5

Build a Dashboard with Plotly Dash

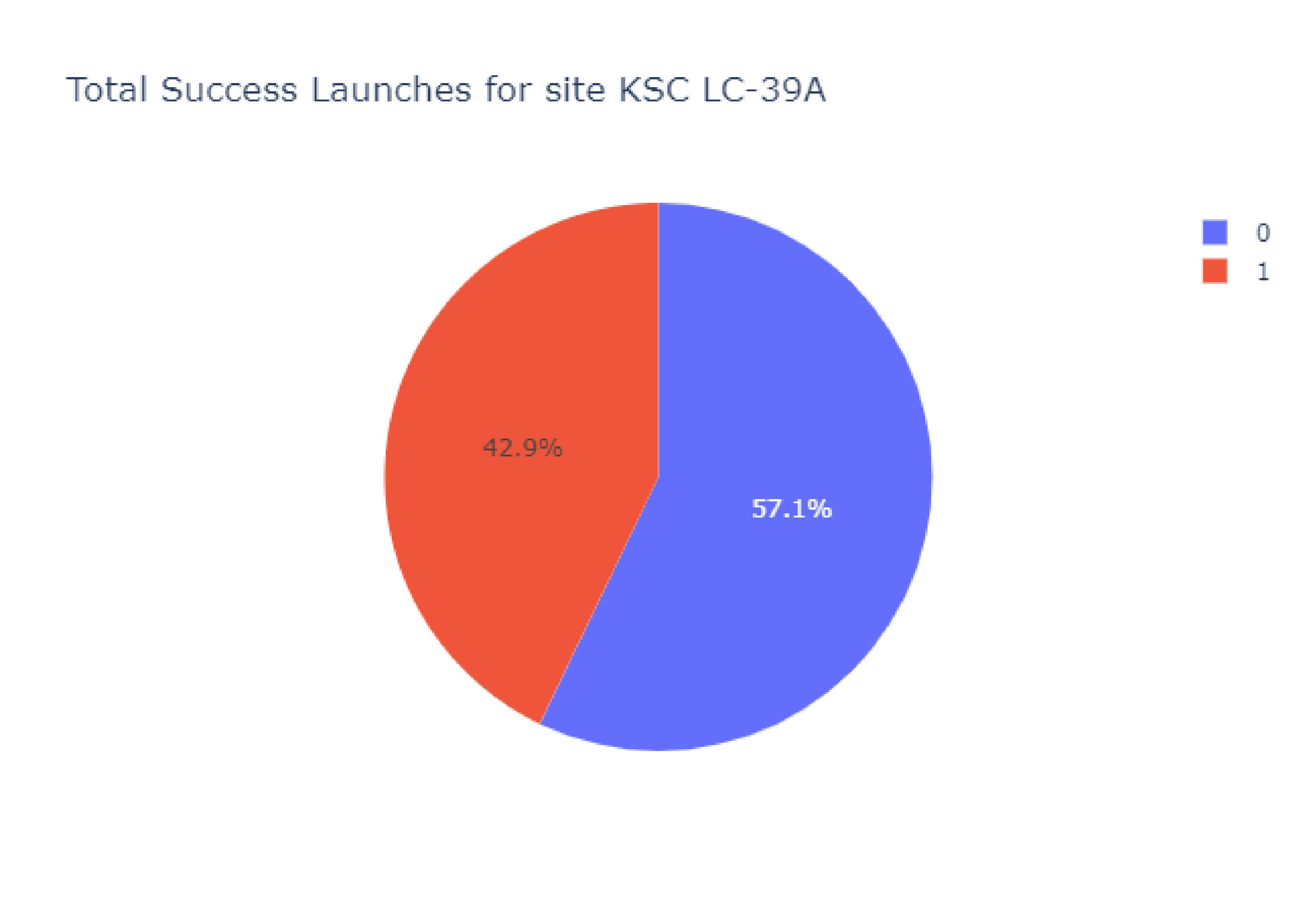
Launch Success Count for All Sites

- From the chart, site KSC LC-39A has the most success count, while CCAFS SLC-40 has the least.

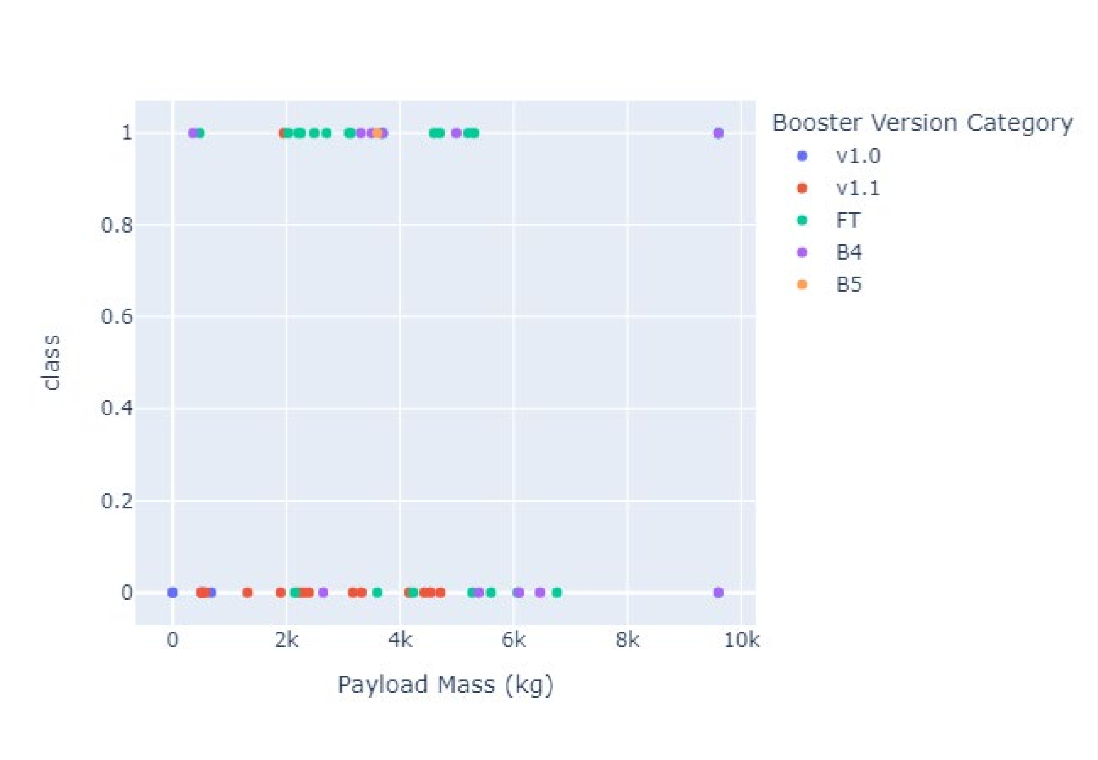


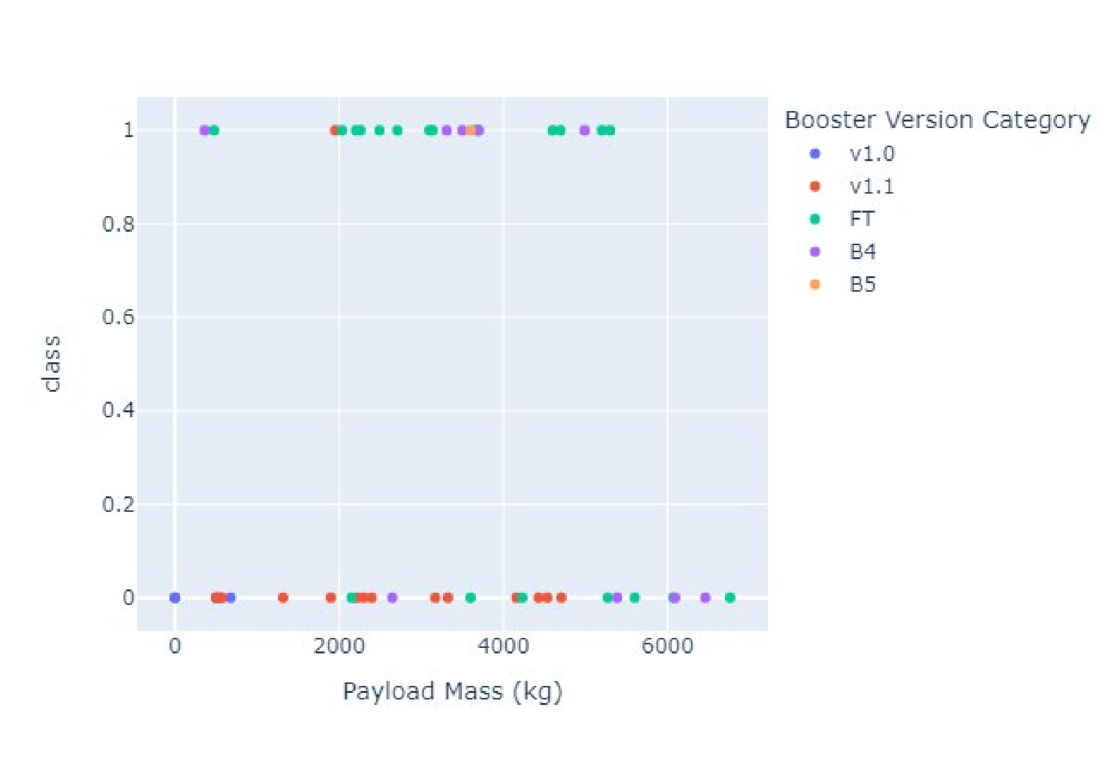
KSC LC-39A Launch Success Rate

- While KSC LC-39A may have the most success counts, however it actually have more failed landing than successful landing, with a success rate of 42.9%.



Payload VS Landing Outcome for All Boosters





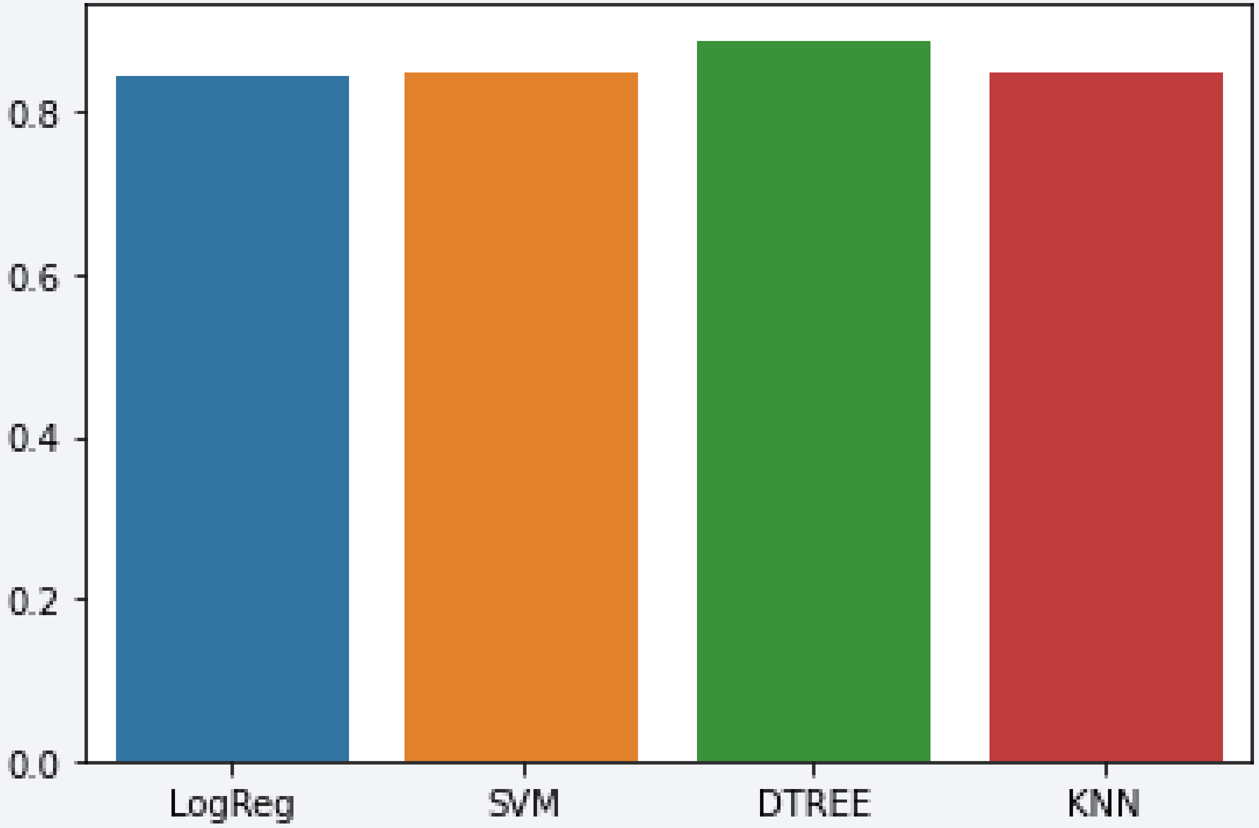
- Since there are only 2 records above 7000k, we make a slice.
- Bewlow 7000k for all boosters,v1.1 seems to have the most failed landing counts, while FT have the best successful landing counts.

Section 6

Predictive Analysis (Classification)

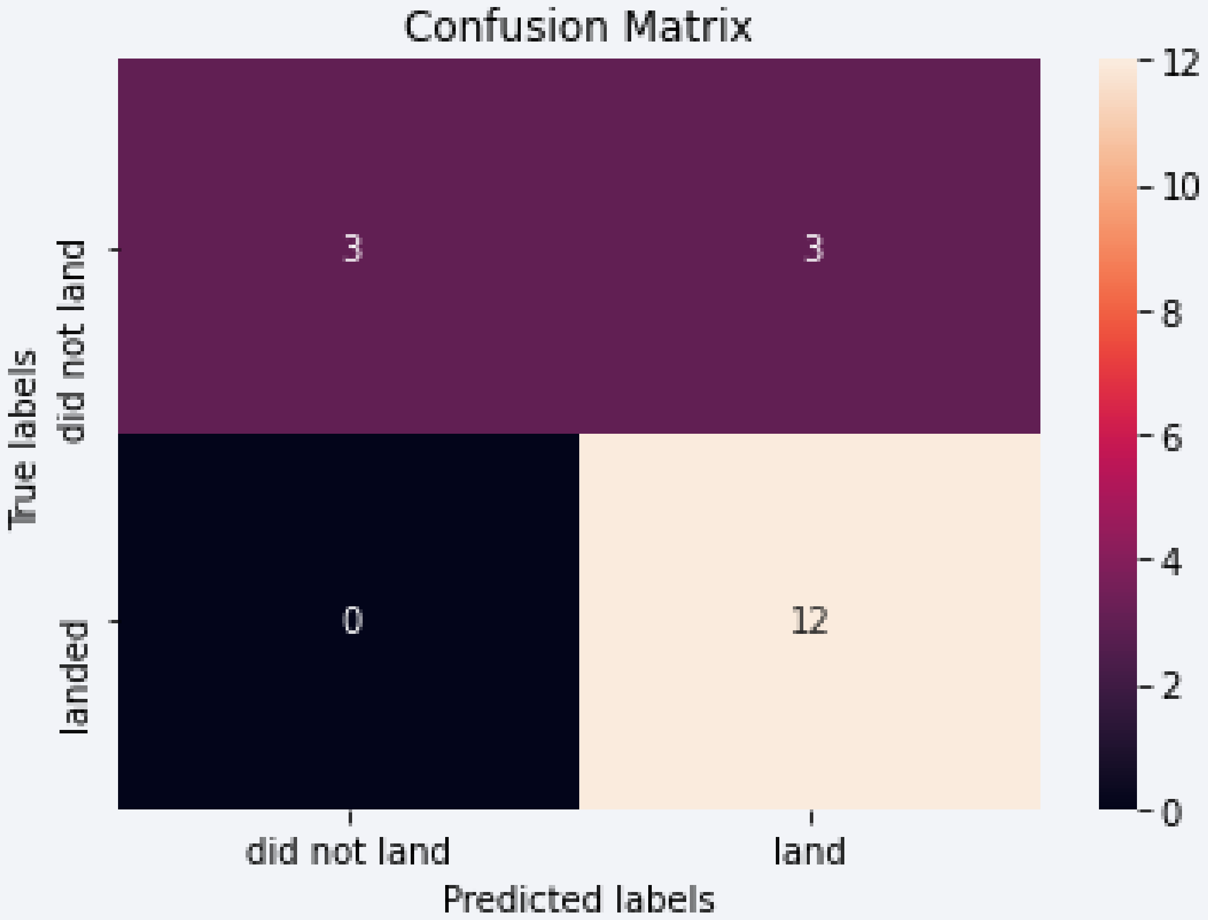
Classification Accuracy

- After building and testing our models, we are confident that **Decision Tree** model should be the best predictive model, with an accuracy of 0.88 with parameters:
  - {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}



Confusion Matrix for Decision Tree

- The decision tree confusion matrix shows that we have almost 0 false-positive rate, and very high true-positive rate, which concludes that we have a very sufficient predictive model.



Conclusions

- Orbit type and payload mass have the most impact on landing outcome.
- When choosing launch sites, CCA sites have a higher successful rate if the payload mass is above 10000.
- ES-L1, GEO, HEO and SSO generally will have very good successful rate.
- There is a decrease in successful rate after 2019 to present date.
- Decision Tree algorithm with parameters given in the previous slides, should be best predictive model.

Appendix

- All codes, notebooks, processed data and reports can be found at:

  https://github.com/simonou99/spacex_landing_prediction

Thank you!