Winning the Space Race with Data Science

Nikos Neveskalos
10/03/2023

Executive Summary

In this report, we attempt to predict if the Falcon 9 first stage will land successfully, using the launch data from SpaceX's A.P.I. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used by our company, Space Y, to bid against SpaceX for a rocket launch. From the data we obtained with the A.P.I., we performed exploratory data analysis to find some (if any) patterns in the data as well as some preliminary insights about how each important variable would affect the success rate. Finally, we determined what would be the label for training supervised models and determined the model with the best accuracy.

Table Of Contents

Introduction

The commercial space age is here, as companies are making space travel affordable for everyone. Perhaps the most successful of all is SpaceX. The main reason of their success is that SpaceX's rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. Consequently, if we can determine if the first stage will land, we can determine the cost of a launch. This is very important if Space Y wants to be able to compete with SpaceX and outbid them for rocket launches in the future.

Therefore, how can we predict if SpaceX will reuse the first stage? Using exploratory data analysis and machine learning models, we can figure out what variables, such as launch site, payload mass and orbit, impact the success of a launch.

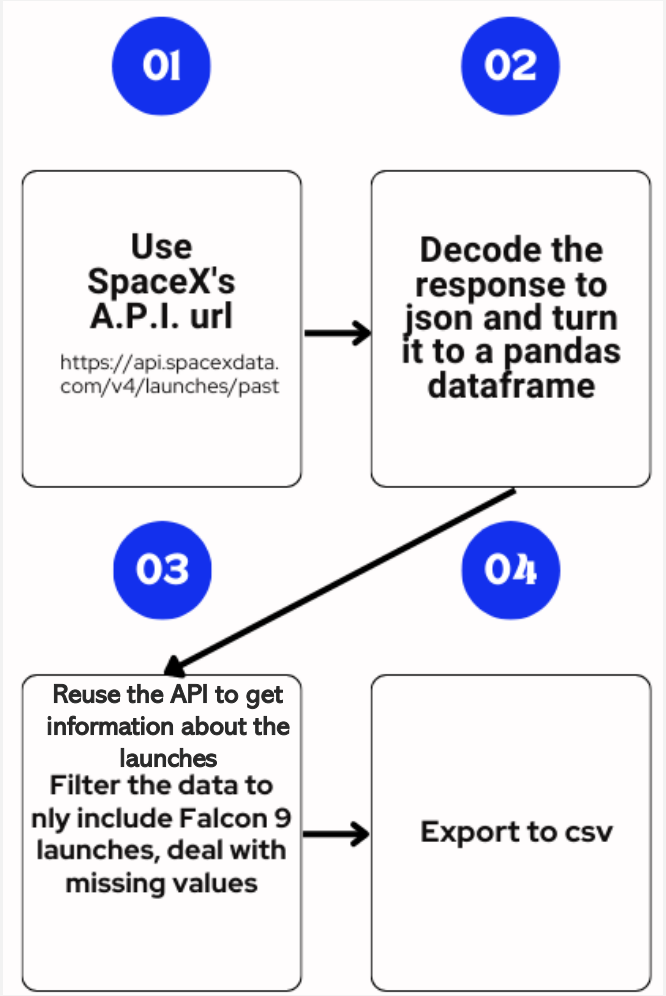
Section 1
Methodology

Methodology

- The data collection and analysis for this project were carried out using the Python programming language.
- The launch data was collected using SpaceX's A.P.I., as well as web scraping launch records from Wikipedia.
- After fitting the launch data to a dataframe, missing values in the " Payload Mass" variable were replaced with the mean of the variable.
- A new variable, "Class", was created to account for the outcome of the landing of the first stage of the rocket.
- Visualisations such as scatter plots and bar charts, analysis with SQL and interactive visual analytics were used for preliminary insights.
- The dataset was standardised and split into a training dataset and a testing dataset.
- Models for logistic regression, support vector machine, decision tree and k-nearest-neighbour were created, tuned and tested, with all models achieving similar accuracy scores, but decision trees being the best, with an accuracy score of 87%.

Data Collection – SpaceX API – Initial Data Wrangling

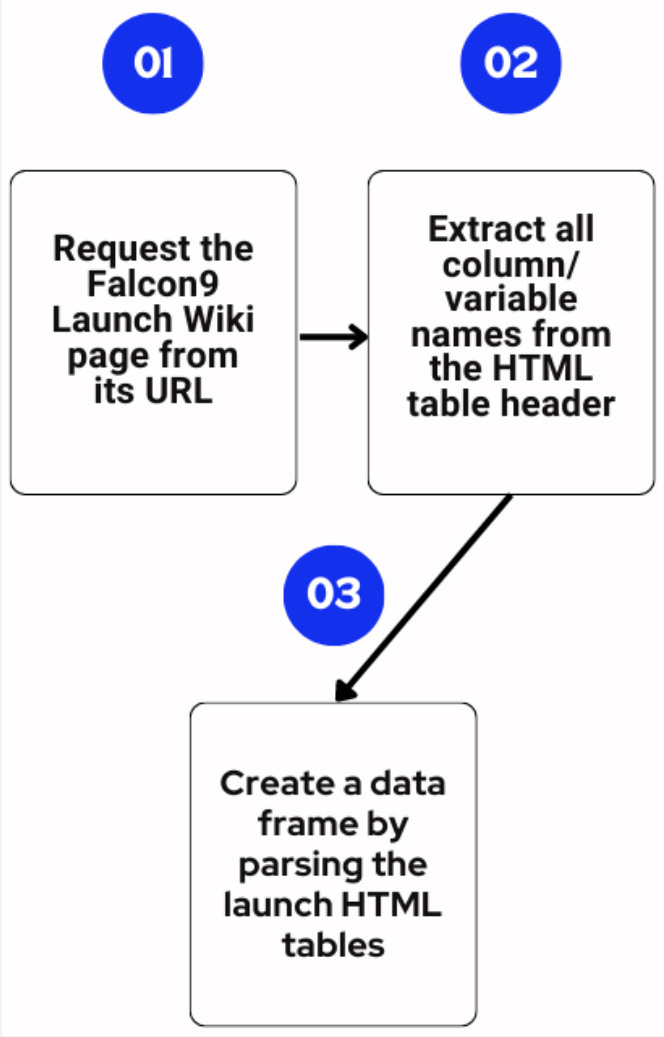
- Using SpaceX's A.P.I., we make an HTTP request to get the data.
- We decode the response content as a json using .json() and turn it into a Pandas dataframe using .json_normalize().
- We use the API again to get information about the launches using the IDs given for each launch. Since the data includes launches of other rockets, we filter it to only keep the Falcon 9 launches and to deal with missing values.



Data Collection - SpaceX API - Initial Data Wrangling

Data Collection - Scraping

- Using a Wikipedia page titled *"List of Falcon 9 and Falcon Heavy launches"*, we will be performing web scraping to collect Falcon 9 historical launch records.
- We perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
- We create a BeautifulSoup object from the HTML response.
- We extract the column names of the table.
- We create an empty dictionary and fill it up with launch records extracted from table rows.
- We create a dataframe from the dictionary and export it to csv



Data Collection - Scraping

Data Wrangling

- After the initial data wrangling process during data collection, in order to filter out launches of other spacecraft and deal with missing values, we will now use data wrangling for exploratory data analysis.
- We identify the variable types of the dataset, number of launches on each site, orbit types and occurrences of each orbit type, different mission outcomes and occurrences of mission outcome per orbit type.
- Finally, and more importantly, we create a landing outcome label from “Outcome” column, a new variable column named “Class”, a binary variable with value 0 for unsuccessful landing and 1 for successful landing.
- This new variable will allow us to perform further EDA in the future by comparing landing outcomes for other variables such as launch site or payload mass.

Data Wrangling

EDA with Data Visualization

- Having now created the new variable “Class”, determining the landing outcome of each launch, we can visualize the relationships between other important variables, and figure out the landing outcomes.
- We will use scatter plots for *“FlightNumber” vs “PayloadMass”, “FlightNumber” vs “LaunchSite”, “PayloadMass” vs “LaunchSite”, “FlightNumber” vs “Orbit”* and *“PayloadMass” vs “Orbit”.* Scatter plots are useful for discovering trends between variables.
- We will use a bar chart to visualise the success rate of each orbit type, to see if certain orbit types have better success rates.
- Finally, using a line chart for Date and Class, we can plot the success rate of launches over the years. The above charts will be analysed in further detail later in this presentation.

EDA with Data Visualization

EDA with SQL

For EDA with SQL, the following queries were used:

- Manes of the unique launch sites
- 5 launch sites whose name begins with the string “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of the first successful landing in ground pad
- Names of the successful boosters in drone ship with payload mass between 4 and 6 thousand kg
- Total number of successful and failed mission outcomes
- Names of booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions and launch site names in year 2015
- Rank of the count of landing outcomes between 2010-06-04 and 2017-03-20

EDA with SQL

Build an Interactive Map with Folium

- When creating an interactive map using Folium, the following map objects were added:

1. The locations of the launch sites as markers
2. The launch outcomes for each site, as colour-labelled markers depending on outcome
3. Lines to the closest shoreline, highway and railroad to each launch site, including distance markers.

- The coloured-labelled markers were especially important, since they allowed us to easily see which launch sites had higher success rates.
- The other objects provide more detail and make the map easier to interpret. The above maps will be analysed in further detail later in this presentation.

Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash

- When building a dashboard with Plotly Dash, the following plots, graphs and interactions were added: 1)

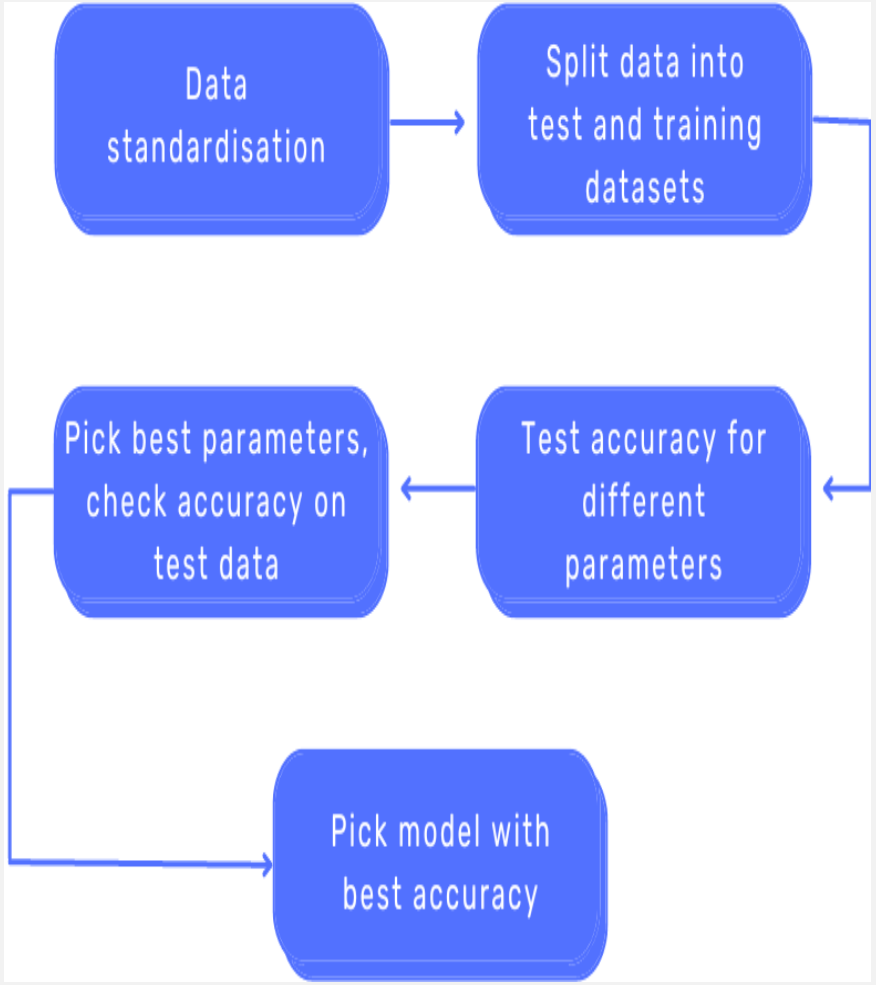
1. A Launch Site Drop-down Input Component, allowing us to view the success rate for each launch site separately, as well as all of them together with the help of pie charts,
2. A scatter plot of launch outcome vs payload mass, with hue for booster version,
3. A slider for the scatter plot, allowing as to select the desired payload mass range.

- The scatter plot allows us to get an idea of how different payloads and boosters might impact the success of a launch. The above dash will be analysed in further detail later in this presentation.

Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

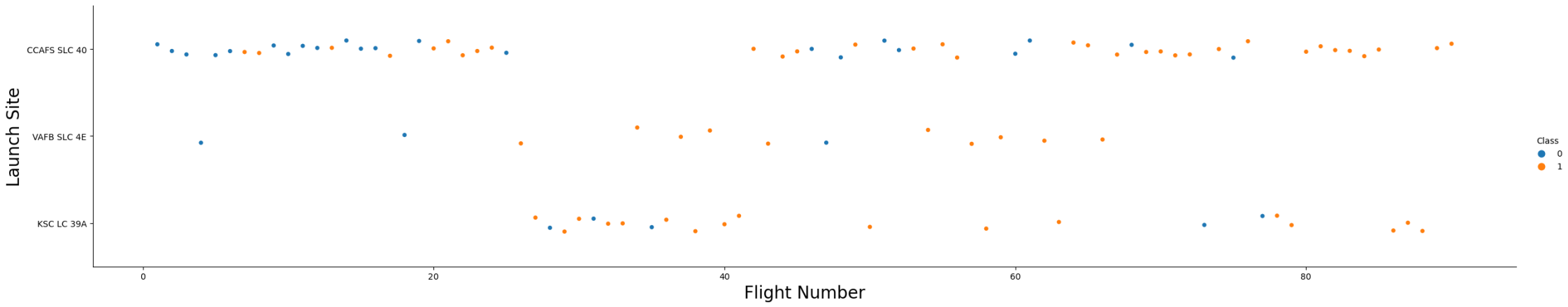
- Using the dataset we created earlier, we build and test different classification models, to see which one might better predict the outcome of a launch.
- Models for logistic regression, SVM, KNN and classification trees were tested.
- Firstly, the data in the “Class” column were standardised.
- Then, the data were split into a training dataset and a test dataset.
- Each model was tested for different parameters and the ones with the better accuracy were selected and their scores on the test data were tested.
- Finally, a confusion matrix was plotted for each model. It was discovered that a Decision tree had the best accuracy of 87% (and 94% on the test data), meaning that the model could successfully predict the outcome of a launch 87 out of 100 times, a satisfactory rate.



Predictive Analysis (Classification)

Section 2
Results

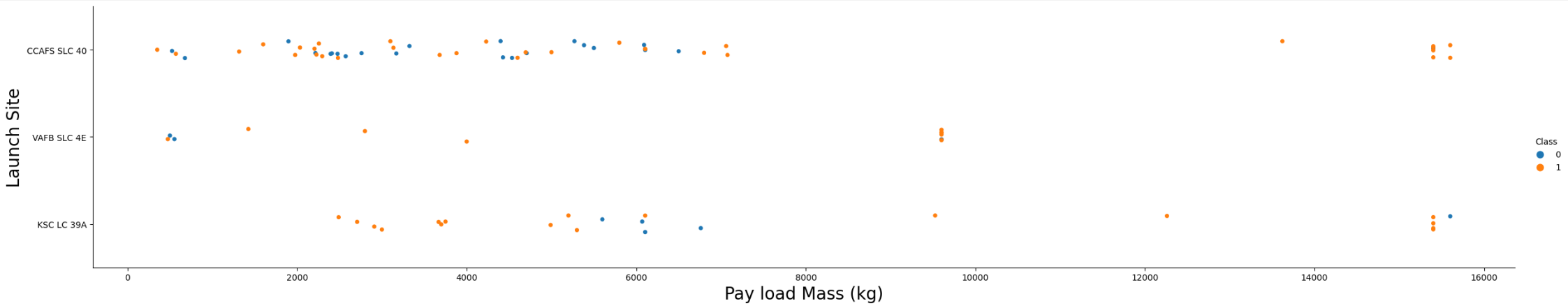
Flight Number vs. Launch Site



Insights

- The first 6 flights were unsuccessful
- The last 13 flights were successful
- Success rate has been increasing with time
- The CAAFS SLC 40 site has the most launches, especially earlier on; but also, the most failures

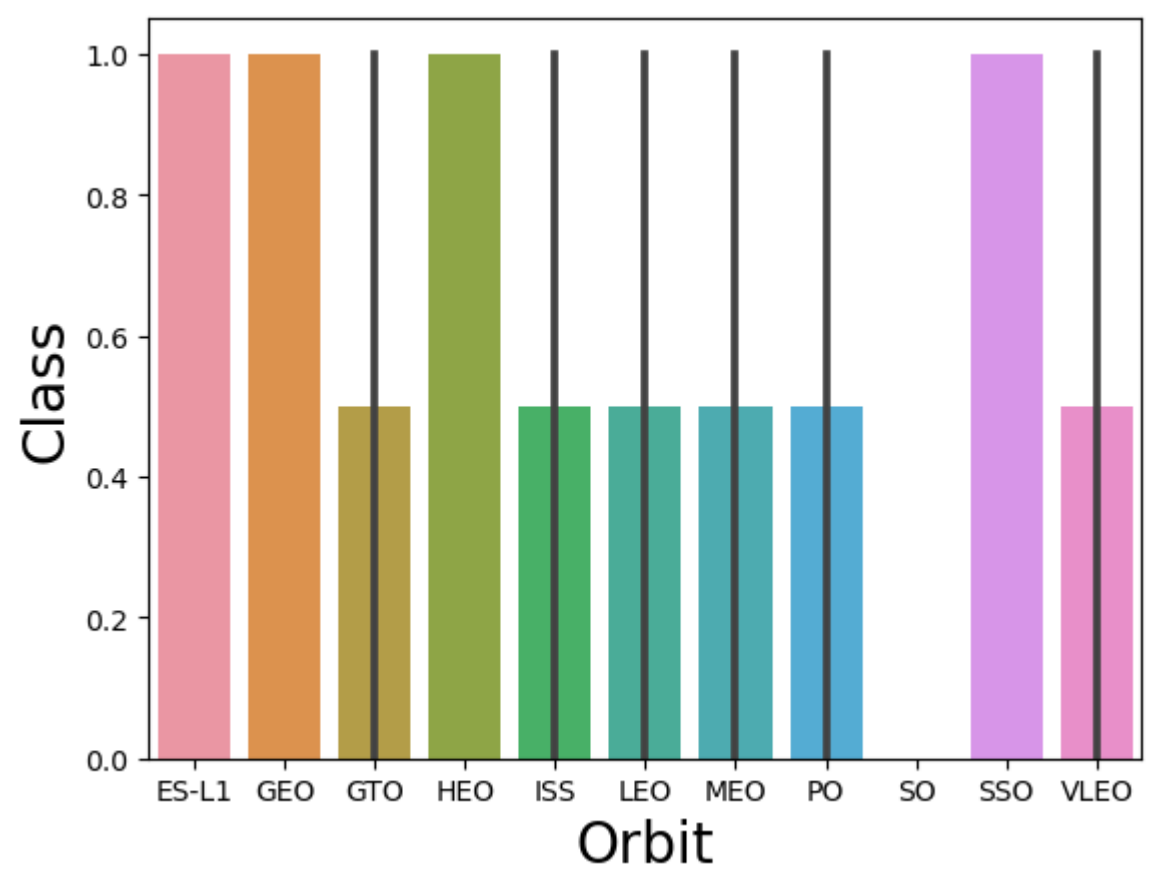
Payload vs. Launch Site



Insights

- Higher payload mass seems to indicate higher success rate
- Launches above 7000kg are mostly successful
- No launches above 10000kg have taken place at VAFB SLC 4E site

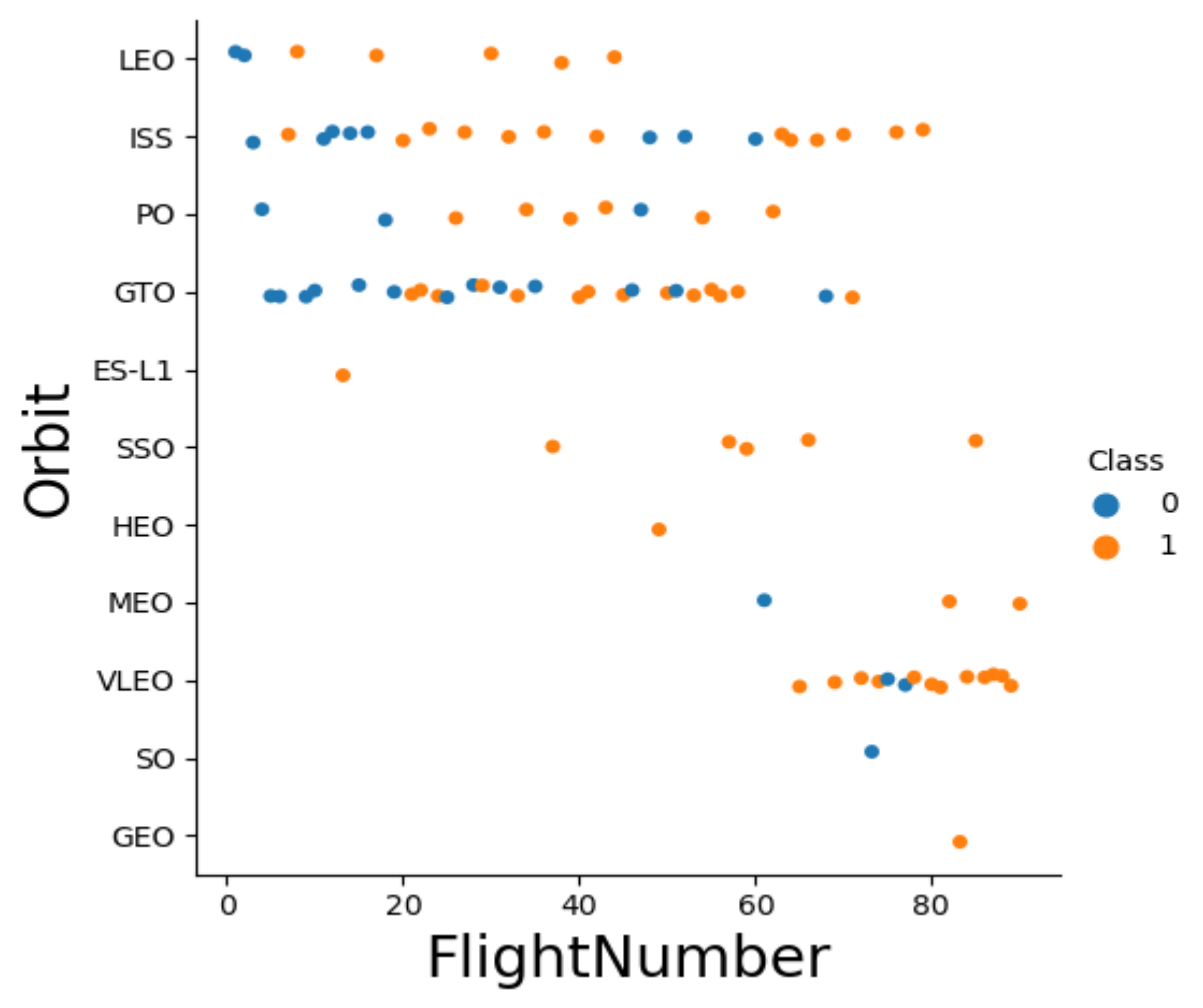
Success Rate vs. Orbit Type



Insights

- Orbits with 100% success rate are ES-L1, GEO, HEO and SSO
- SO orbit has 0% success rate

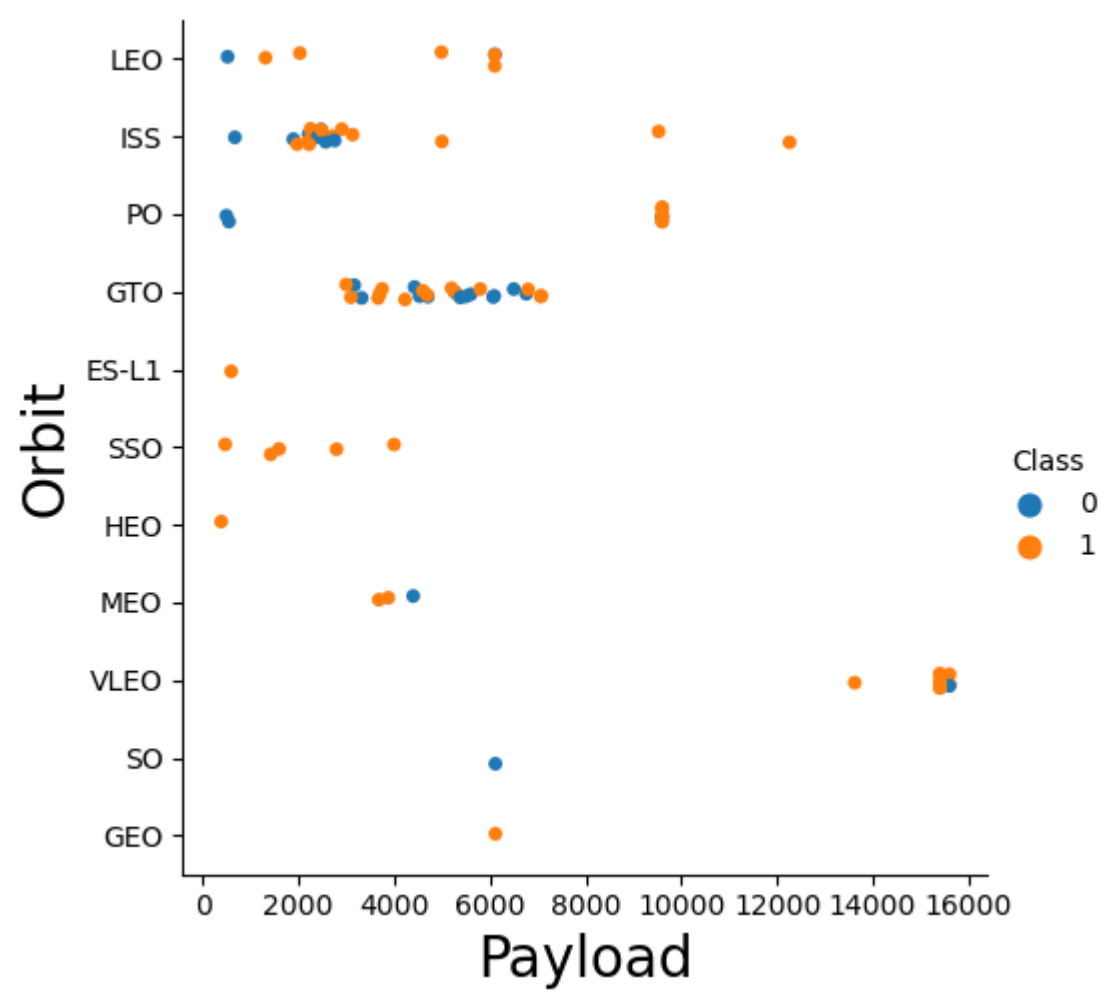
Flight Number vs. Orbit Type



Insights

- Success rate has improved over time for all orbit types

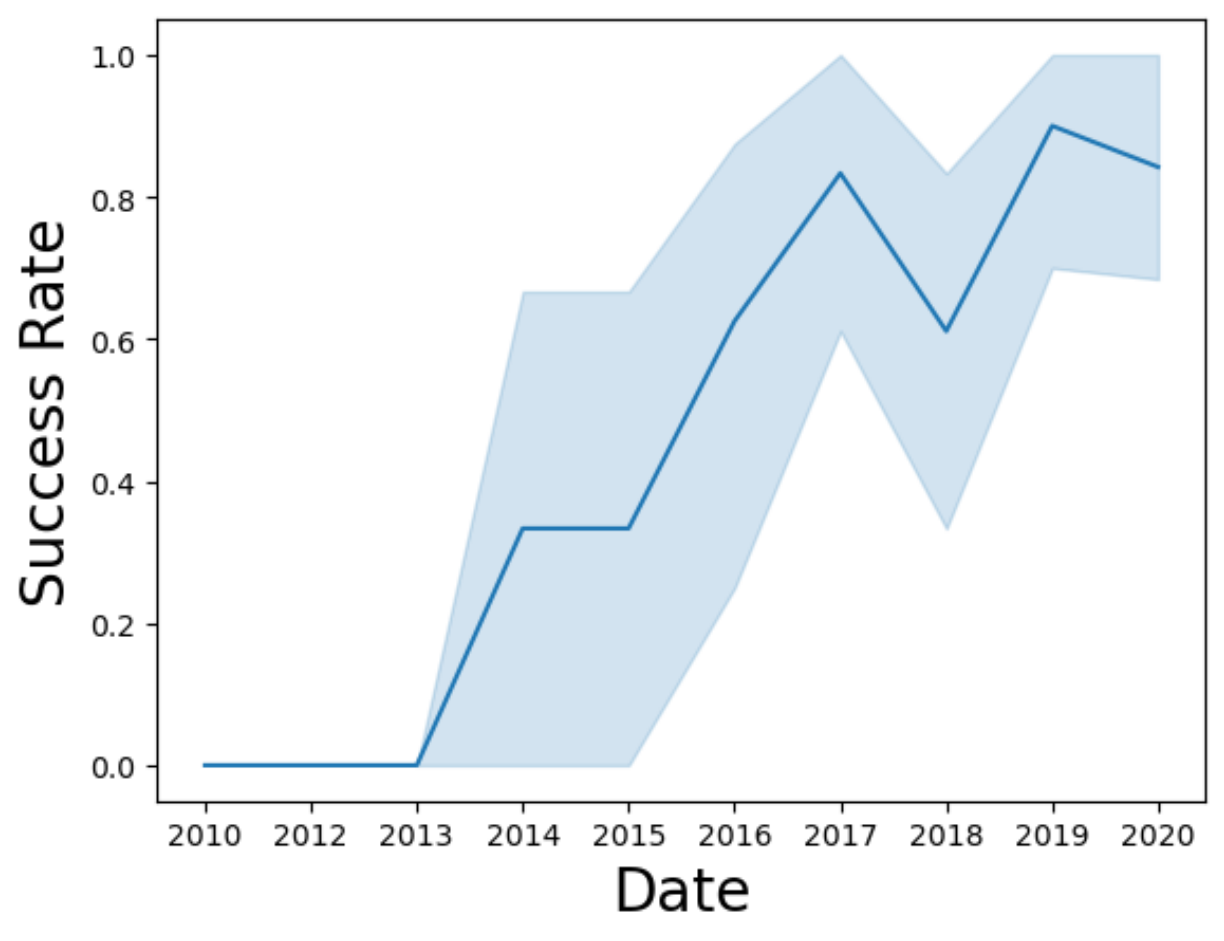
Payload vs. Orbit Type



Insights

- ISS have the widest range of payload mass
- VLEO orbits only have high payload mass
- SSO, HEO and MEO orbits only have small payload mass

Launch Success Yearly Trend



Insights

- Over time, success rate has been steadily getting better
- With the exception of years 2018 and 2020, where success rate slightly dipped

All Launch Site Names

%sql select distinct launch_site from SPACEXDATASET;

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

- Displaying the names of the unique launch sites

Launch Site Names Begin with 'CCA'

%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

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

- 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

total_payload_mass
45596

- Displaying the total (sum) payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

average_payload_mass
2534

- Displaying the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

first_successful_landing
2015-12-22

- Displaying the date when the first successful landing outcome in ground pad was achieved. Interestingly, it was 5 whole years after the first launch

Successful Drone Ship Landing with Payload between 4000 and 6000

%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

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

- Displaying the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

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

- Displaying the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

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

- Displaying the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
    where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

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

- Displaying the failed landing outcomes in drone ship for the year 2015, their booster versions, and launch site names

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
    where date between '2010-06-04' and '2017-03-20'
    group by landing__outcome
    order by count_outcomes desc;
```

landing__outcome	count_outcomes
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

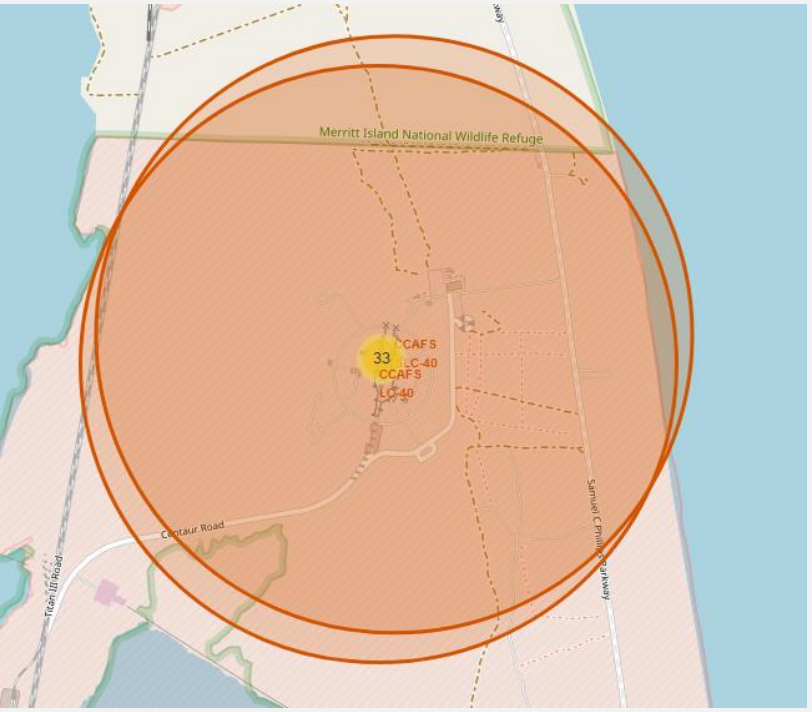
- Displaying the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order.
- We see that for most failed outcomes, no landing attempt was made
- The most successes were ground pad landings

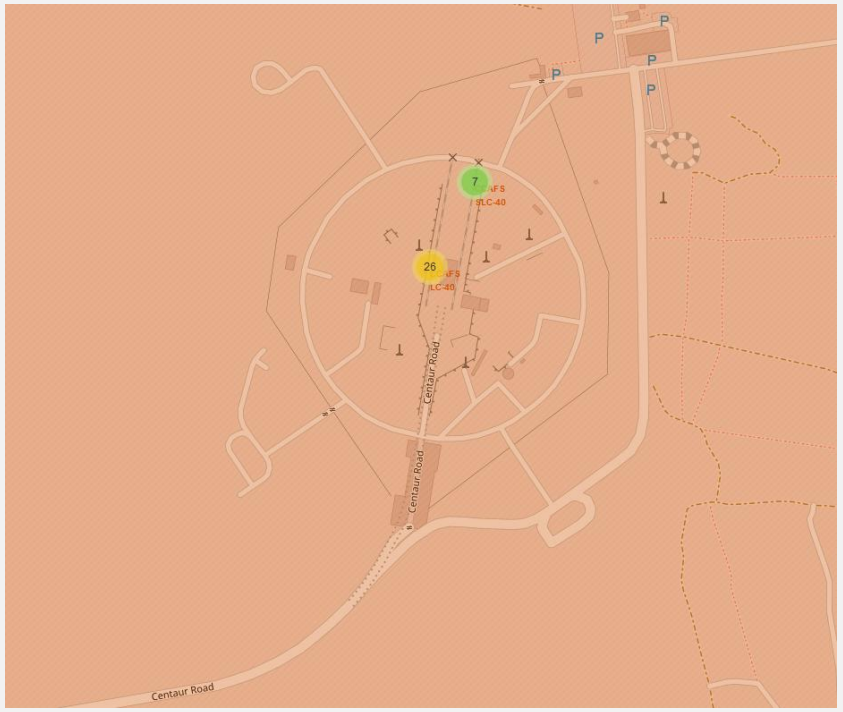
Launch Site Markers on a Global map

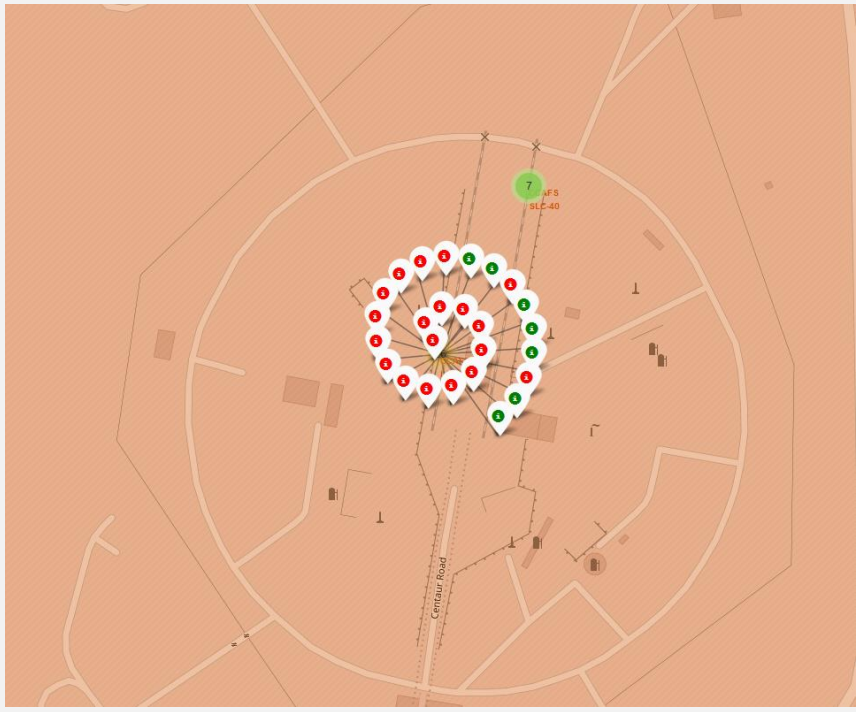
Insights

- All launch sites are in close proximity to the ocean
- All launch sites are far from major population centres such as L.A. Or Miami

Colour-Labelled Launch Outcomes



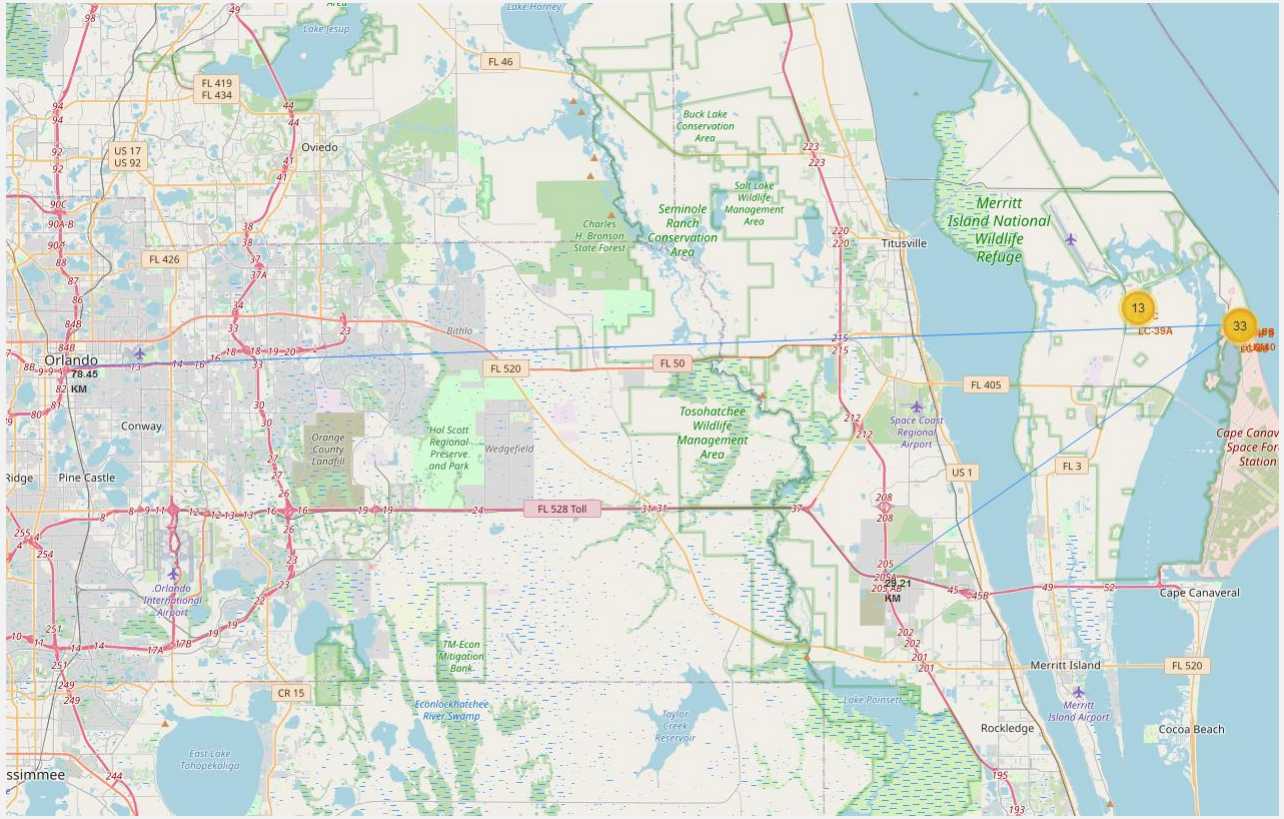


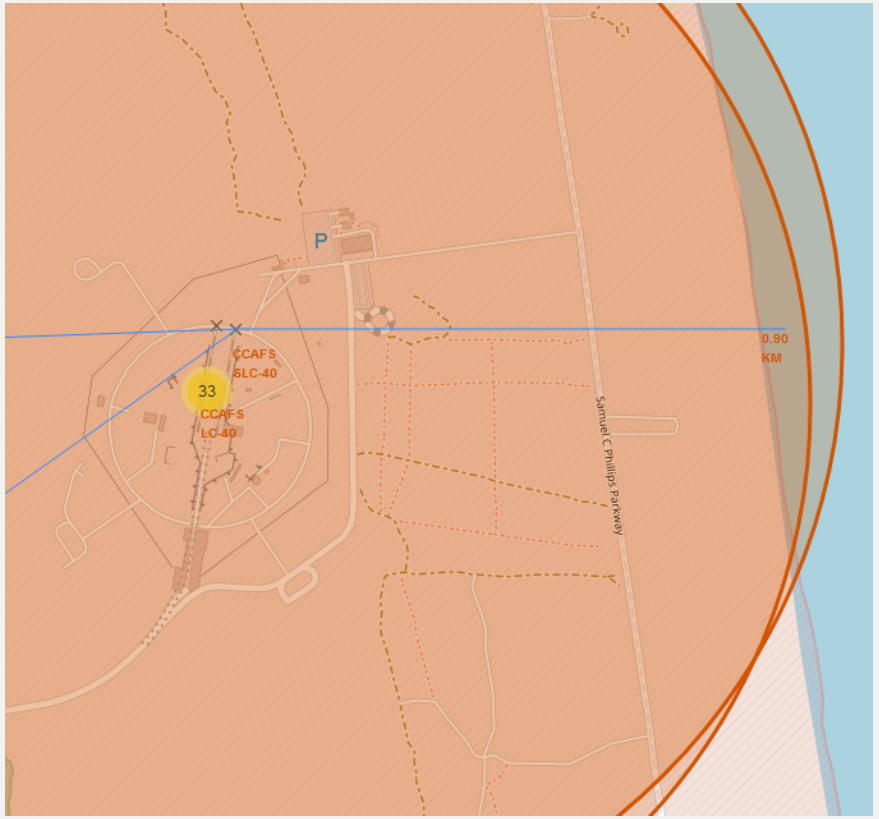


Insights

- Example of colour-labelled markers of launch outcomes for launch site CCAFS LC 40
- Green indicates a successful launch while red a failed one

Launch Site Proximities

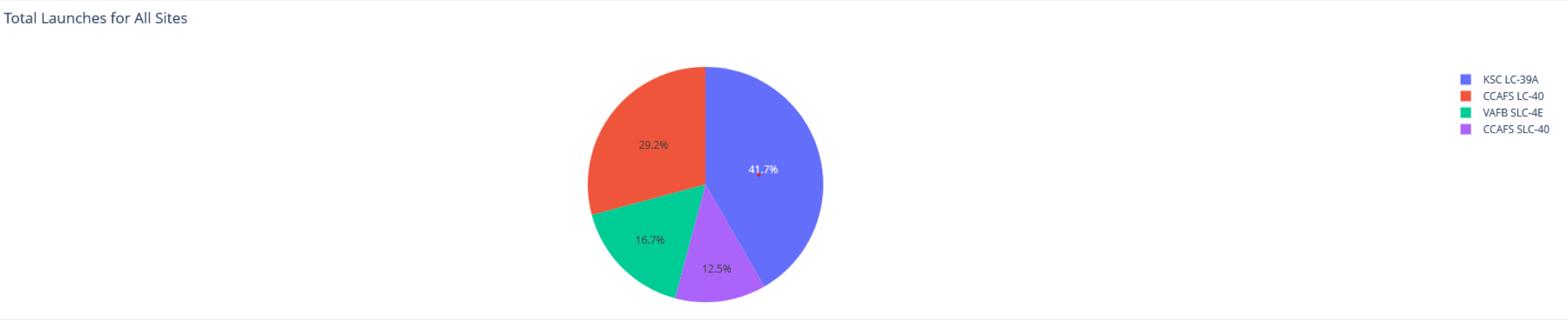




Insights

- Proximities of launch site CCAFS LC 40 to major cities, highways and the shoreline
- The launch site is very close to the shore, but far from major highways and the city of Orlando

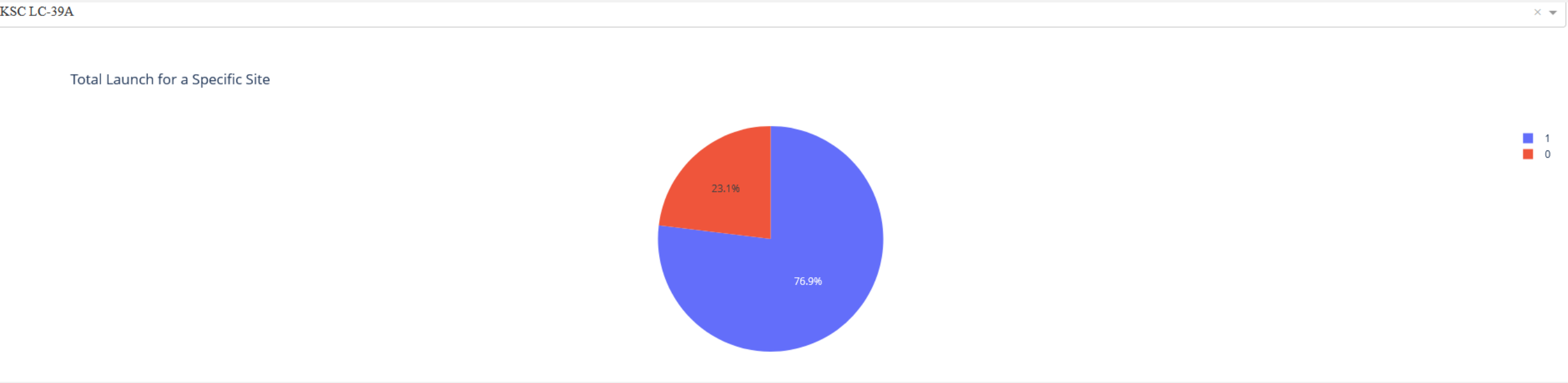
Dashboard – Successful Launches by Launch Site



Insights

- Launch Site KSC LC-39A has the best success rate
- Launch Site CCAFS SLC-40 has the lowest success rate

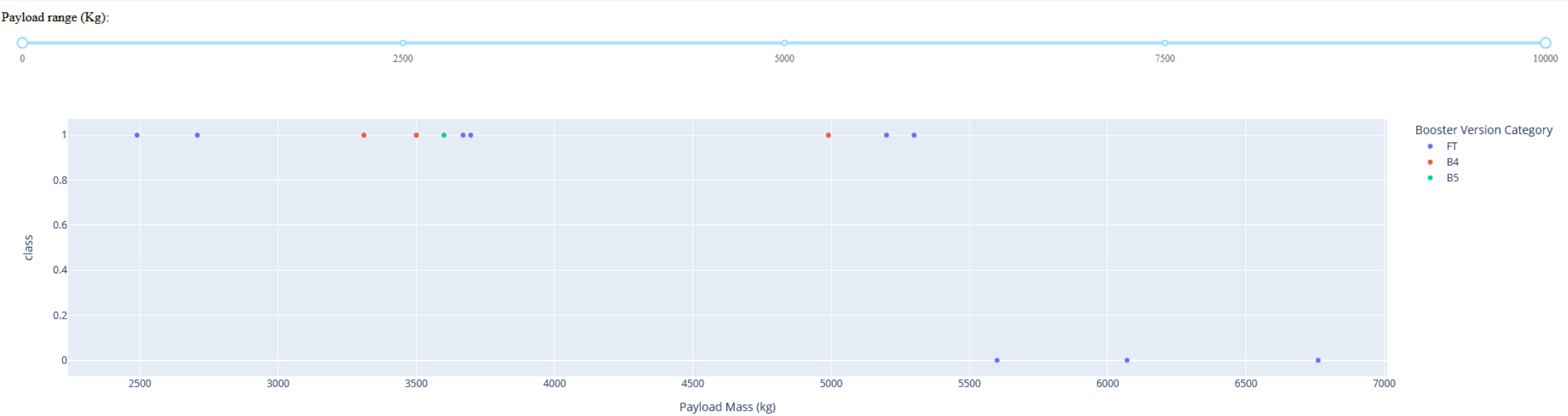
Dashboard – Launch Site KSC LC-39A



Insights

- Launch Site KSC LC-39A has the best success rate, with 76.9%

Dashboard – Payload Mass (Whole Range)



Insights

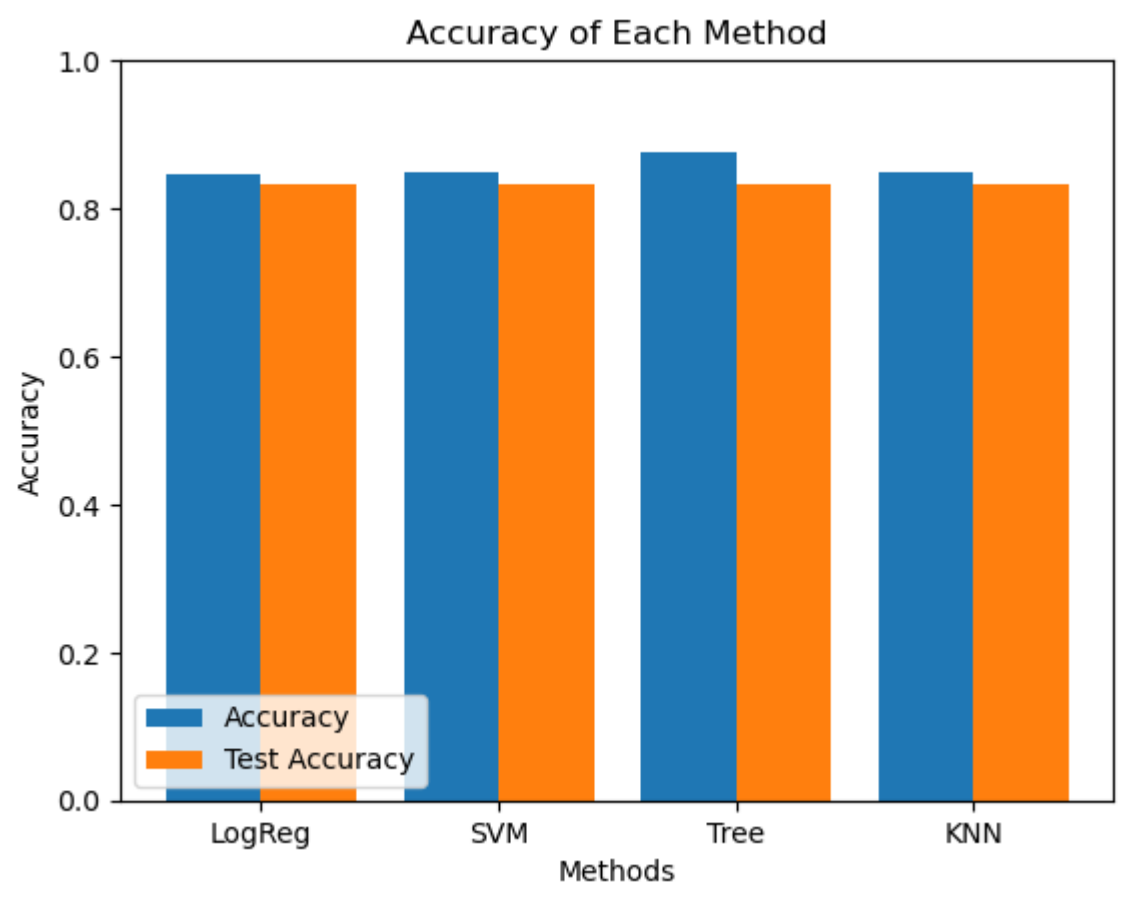
- Focusing on the whole range of payloads, we see that higher payloads lead to unsuccessful launches for these booster variations...

Dashboard – Payload Mass (2500-5000kg)

Insights

- ...while payloads between 2500 and 5000 kg only lead to successful launches

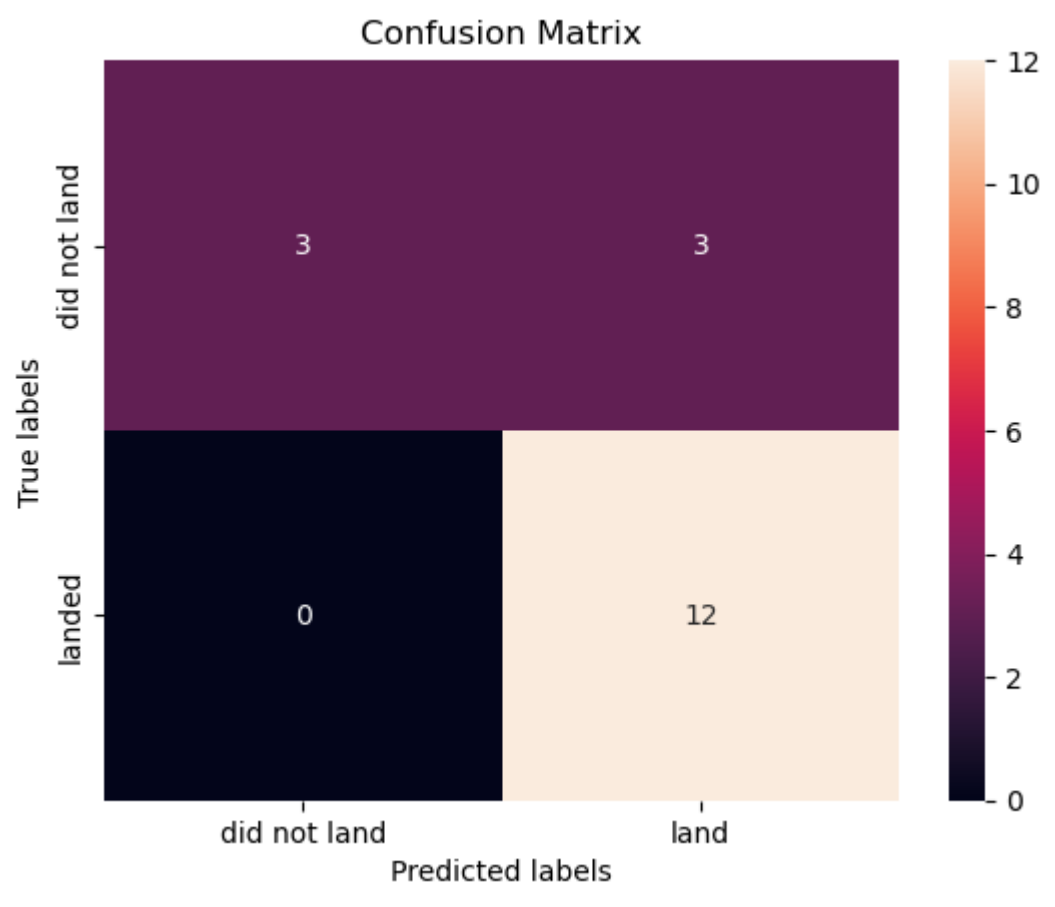
Classification Accuracy



Insights

- Decision Tree model seems to have the highest accuracy, with 87%
- All models performed similarly on the test data (possibly due to small sample size)

Confusion Matrix



Insights

- Because all the models have the same test accuracy, the confusion matrixes are the same
- The main issue for our models is False Positives

Conclusions

- The success of a mission seems to depend on a number of factors such as launch site, orbit type and payload mass, but the most important factor by far is the number of launches and the knowledge gained in each one
- Launch Site KSC LC-39A has the highest success rate, but more data is needed to explain why
- Orbit types HEO, SSO, GEO and ES-L1 are the least risky
- Higher payloads seem to indicate higher success rates
- Decision Tree model works the best for predicting the outcome of future SpaceX missions. As time goes on and more launches happen, the model can further be enhanced with the additional data
- SpaceX learned through trial and error and a lot of hard work... So SpaceY should be prepared to do the same if we want to be able to compete

Appendix

Special thanks to the instructors of this course for their helpful insight and knowledge

Thank you for your time!

Nikos Neveskalos
10/03/2023