IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Maureen Jarau
15th Dec 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **The methodologies used to collect and analyze data:**
  - Data Collection using web scrapping and SpaceX REST API
  - Exploratory Data Analysis (EDA) involving data wrangling, data visualization and interactive visual analytics with Folium
  - Machine Learning (ML) prediction
- **Summary of all results**
  - Relevant data were successfully obtained from public sources
  - EDA helped to determine which features and methods are the best to predict the success in each launch.
  - ML prediction provided the best model to predict the success of Falcon9 launch

Introduction

- SpaceX, an American private rockets and spacecraft manufacturer is currently the leading company in the space travel revolution.

- The company offers low rocket launches such as Falcon 9 as low as because they can US$62M; while other providers cost upward of US$165M per launch.

- The low cost is mainly due to the capability of SpaceX to reuse the first stage by a successful landing mission to be utilized for the next launch. Continuous landing success will contribute to a significant reduction in launching costs for the company.

- As a Data Scientist of a new startup rivaling company (SpaceY), the objective of this project is to create a Machine Learning model to predict the landing outcome of the first stage. This project will provide insights into determining the reasonable cost to bid against SpaceX in rocket launching.

Introduction

Problems to be solved:

- Identify all features that contribute to landing outcome
- Determine the relationship between each variable and how it influences the landing outcome
- The best features/conditions required to enhance the probability of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
  - Webscrapping(https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - Data was processed using one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

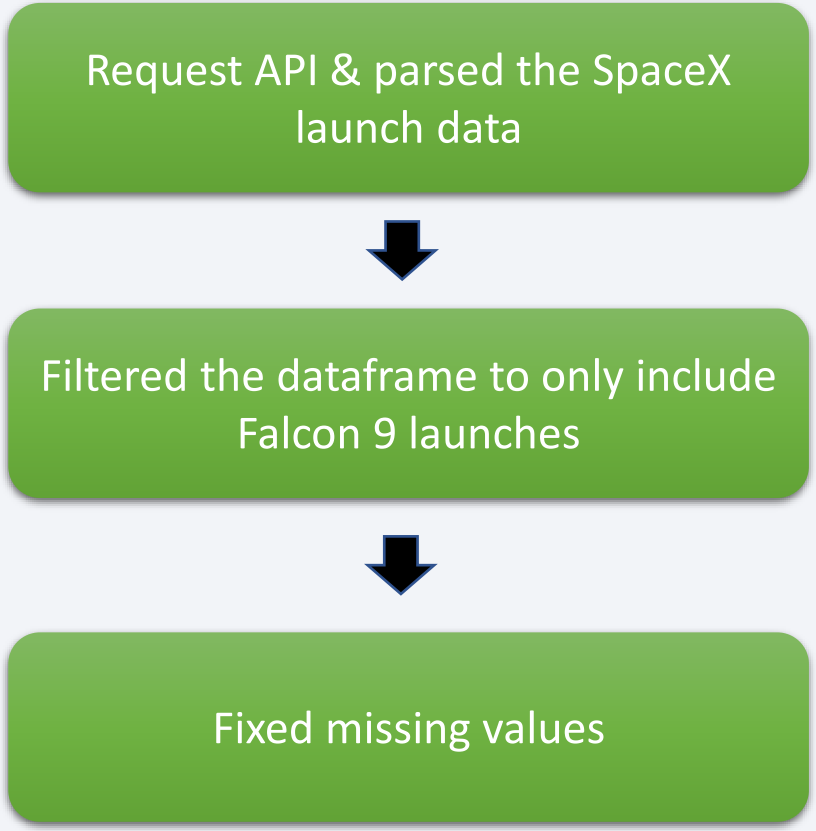
- Perform predictive analysis using classification models
  - Collected data obtained until this step were normalized, split into training and test data sets and evaluated using 4 different models (Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors).
  - The accuracy of each model was evaluated using different combinations of parameters.

Data Collection

- Data collection is a process of identifying and gathering available data resources to answer business questions and evaluate the outcomes.
- The dataset was collected by SpaceX REST API and Web Scrapping from Wikipedia.

Data Collection – SpaceX API

- SpaceX shared a public API where the data can be obtained and used
- The flowchart shows the data collection process by SpaceX API



Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/e2833705d6a0a992b4b1638c4914126d35d3c1f7/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Data from SpaceX launches were obtained from Wikipedia from its URL

- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/68b9ebd0c19d3ac9be526ced6bc140a3a0228429/jupyter-labs-webscraping.ipynb

Request Falcon 9 launch on Wiki page

↓

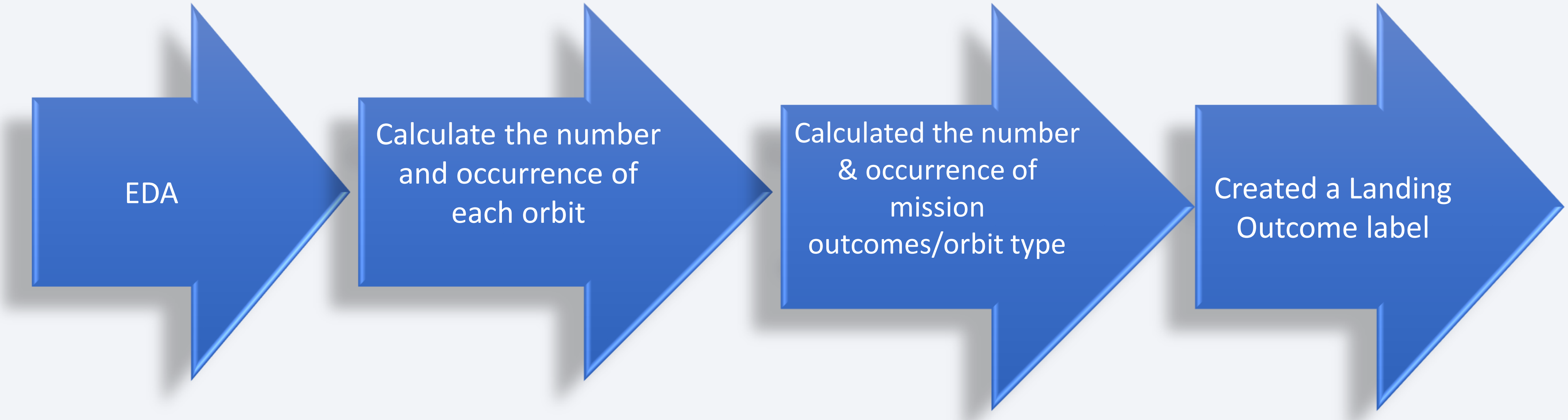
Extracted all column/variable names from the HTML table header

↓

Created a dataframe by parsing the launch HTML tables

Data Wrangling

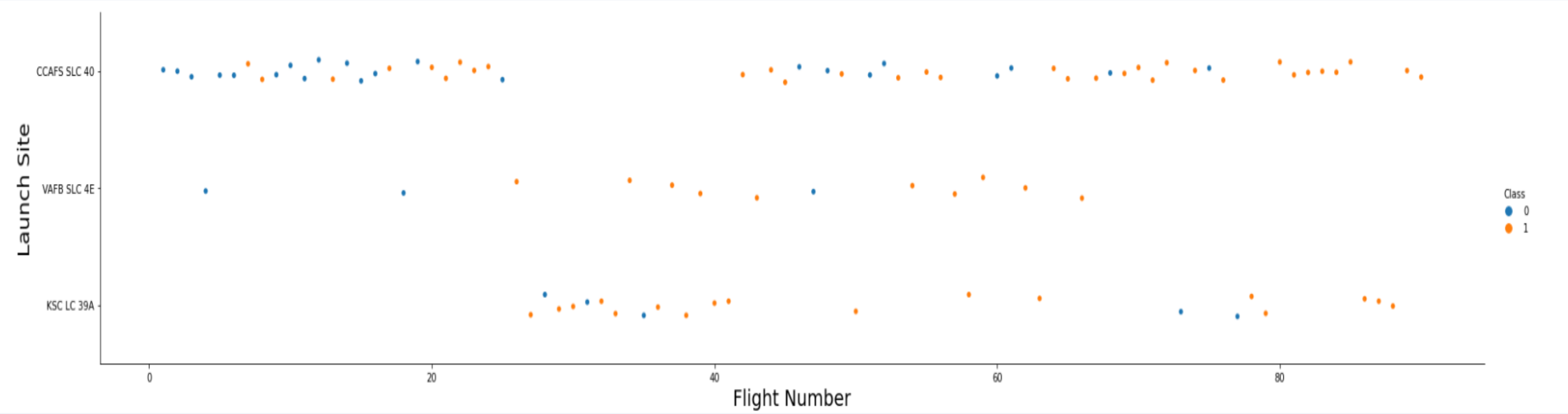
- Data Wrangling involves cleaning and unifying messy and complex datasets for easy access and Exploratory Data Analysis (EDA)
- EDA was first performed, followed by summaries of launches per site (i.e. calculations of occurrences of each orbit & mission outcomes per orbit)
- Lastly, the landing outcome column was added to the dataframe column.

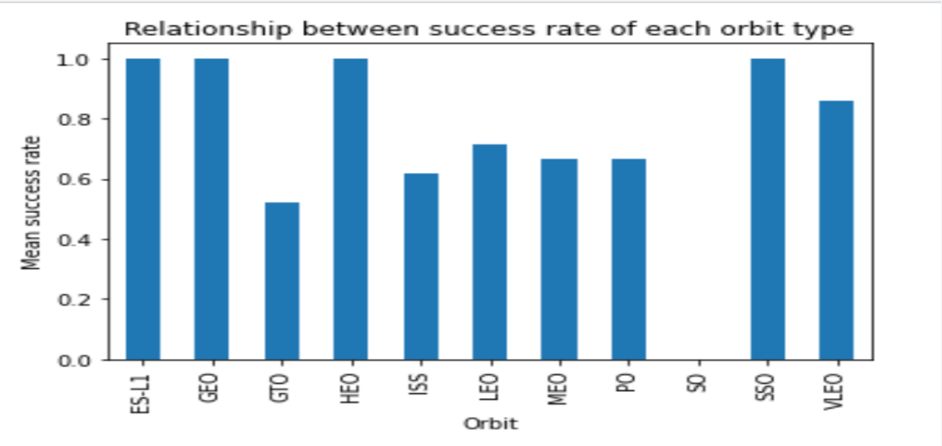


- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/5a5211871ca9230a074f4d183d660409d036b2a4/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

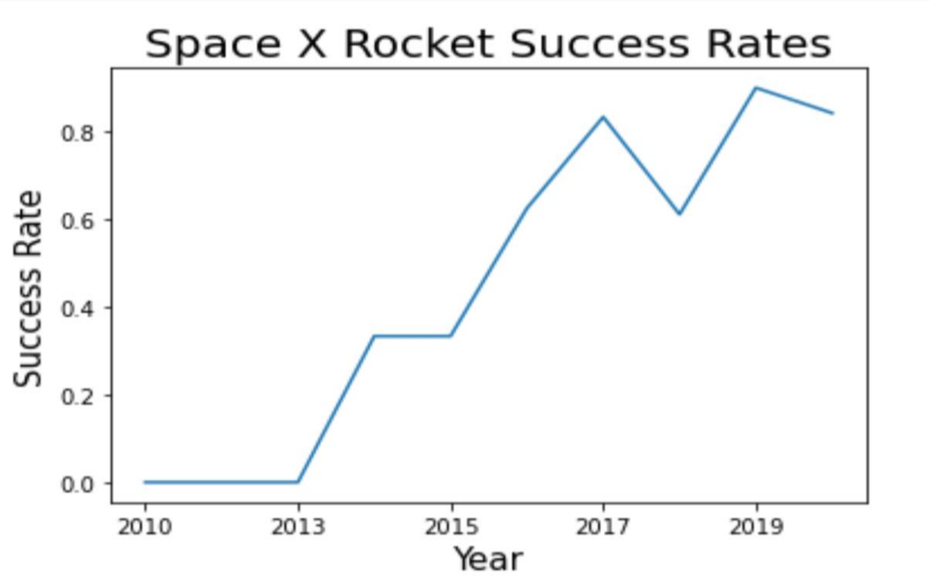
- We used scatter graph and barplots to visualize the relationship between attributes:-
  - ➢ Payload vs Flight Number
  - ➢ Flight Number vs Launch site
  - ➢ Payload vs Launch site
  - ➢ Flight vs Orbit
  - ➢ Orbit vs Payload





EDA with Data Visualization

- We used line graph to present the trend of the success rate launch over time.



- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/ff39b37eec36002b17823edabc4bc3babe329968/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed as part of EDA in this project:-

- Display the names of the launch sites
- Display the name of 5 launch sites begin with string ‘CCA’
- Display the total payload mass carried by booster launch by NASA (CRS)
- Display the average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in the ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass >4000 but <6000
- List the total number of successful and failed mission outcomes
- List the name of the booster_versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions and launch sites names for the year 2015
- Rank the count of landing outcomes/success between 2010-06-04 and 2017-03-20 in descending order

- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/330c42dd649d2463cfb5e3240ab9b7c6000ee694/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The exact latitude and longitude coordinates of each launch sites were used
- Map objects i.e markers, circles, lines and marker clusters were created and added to Folium maps
  - Markers - to show points such as launch sites
  - Circles – to highlight areas around specific coordinates
  - Lines – to show distances between 2 coordinates
  - Marker clusters – to show groups of events that occurred in each coordinate e.g launch sites

- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/36211afaf4edd8900e887a7ae4b78340e5a125b2/lab_jupyter_launch_site_location_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- SpaceX Launch Record Dashboard was built to enable users to interact with the visual data in one place.
- The dashboard consists of:-
  - Pie chart – to show the percentage of successful launches of selected launch site(s)
  - Scatter plots – to show the relationship between Payload Mass (Kg) and landing outcome of selected launch site(s)
- Those plots and interactions enable users to visually analyze the relationship between payload and launch site(s) thus enable to users to determine the best site to launch a rocket.
- Source: https://github.com/mjarau/Applied-Data-Science-Capstone-SpaceX/blob/069bb77c893fe7487bdd50dbdf2da574103d205f/Build%20a%20Dash%20Application%20with%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

Model Building

- Clean dataset is used at this stage.
- Pandas and Numpy were used
- Transformed and split train/test dataset
- 4 types of ML classifier models were selected: LogReg, Decision Tree, SVM, KNN
- Set parameters & GridSearchCV object, fit to the dataset

↓

Model Evaluation

- Accuracies of each ML classification model were tested.
- Tuned hyperparameters were determined
- Confusion matrix plotted

↓

Model Improvement

- Feature Engineering and Algorithm Tuning were utilized for this stage

↓

Model Selection

- ML classifier model with the highest accuracy will be selected for use in prediction analysis

Results

The results are divided into 3 sections:-

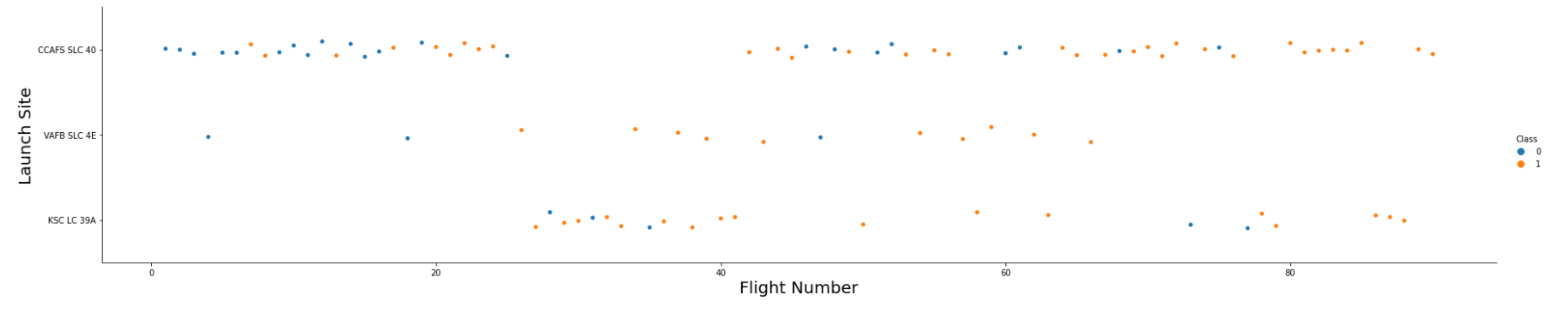
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

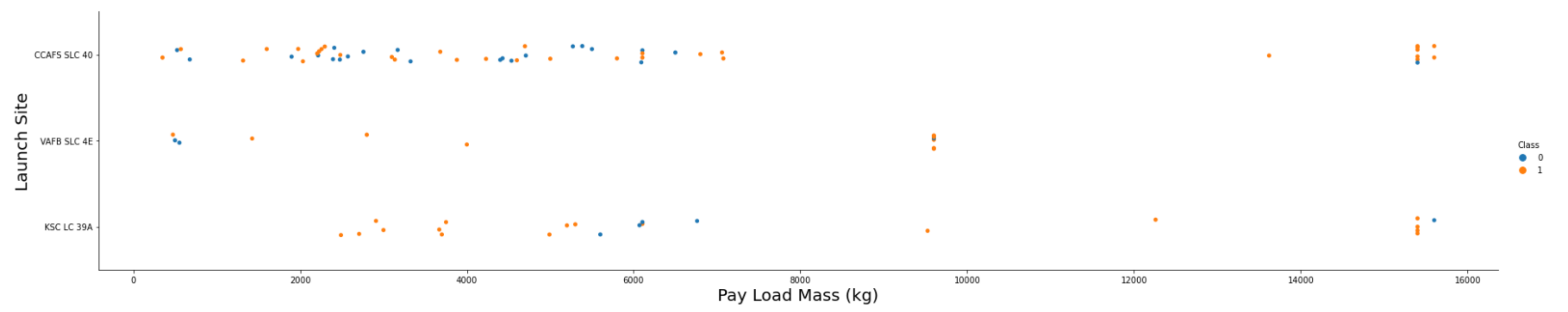
Flight Number vs. Launch Site

- The scatter plot shows the more flight frequency of the launch site, the higher the success rate will be.
- The best launch site is CCAF5 SLC-40, followed by VAFB SLC 4E



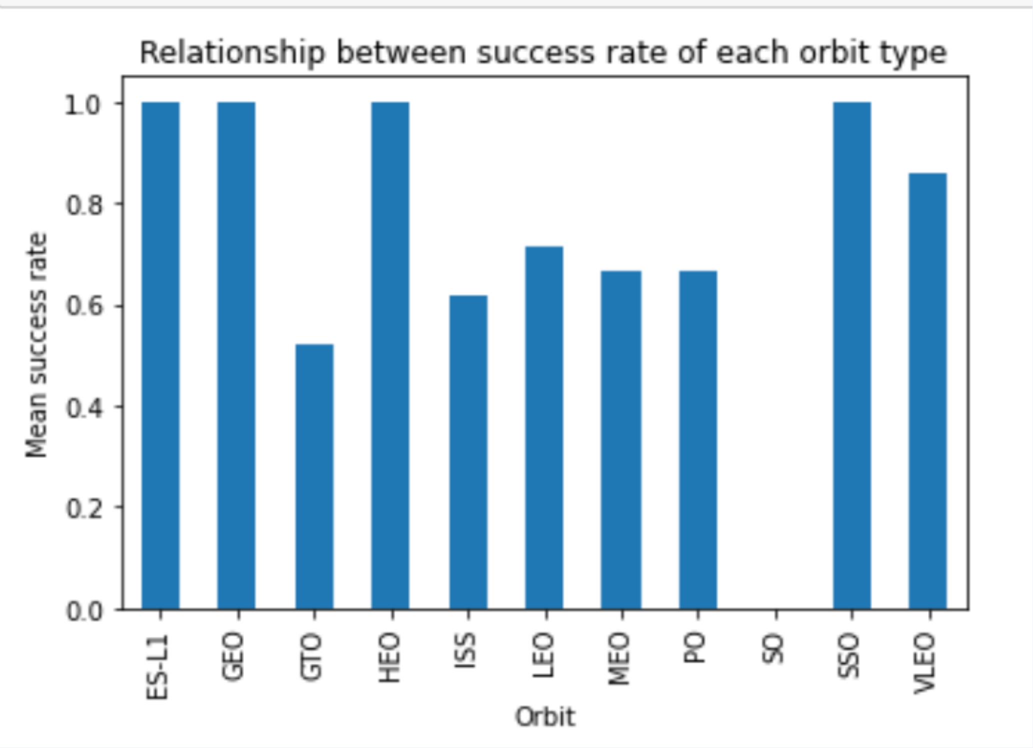
Payload vs. Launch Site

- The scatter plot below shows Payload of >7000 kg are highly likely successful in launching in all sites
- However, a clear relationship between Payload and Launch Site is yet to be determined.



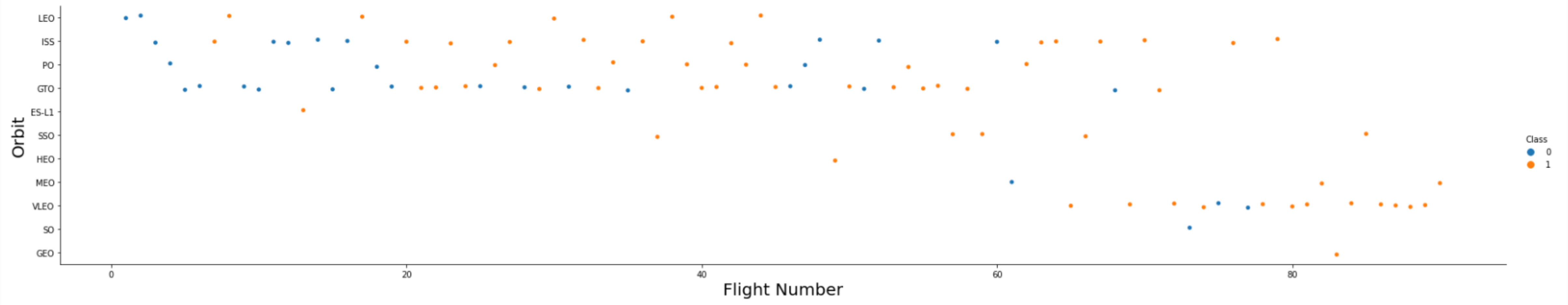
Success Rate vs. Orbit Type

- Top 4 orbits with highest success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- SO orbit did not show any rates of success
- However, as some of the highest success rate only launched once (e.g. GEO, HEO,ES-L1) more data is needed to observe the relationship between success rate vs orbit type



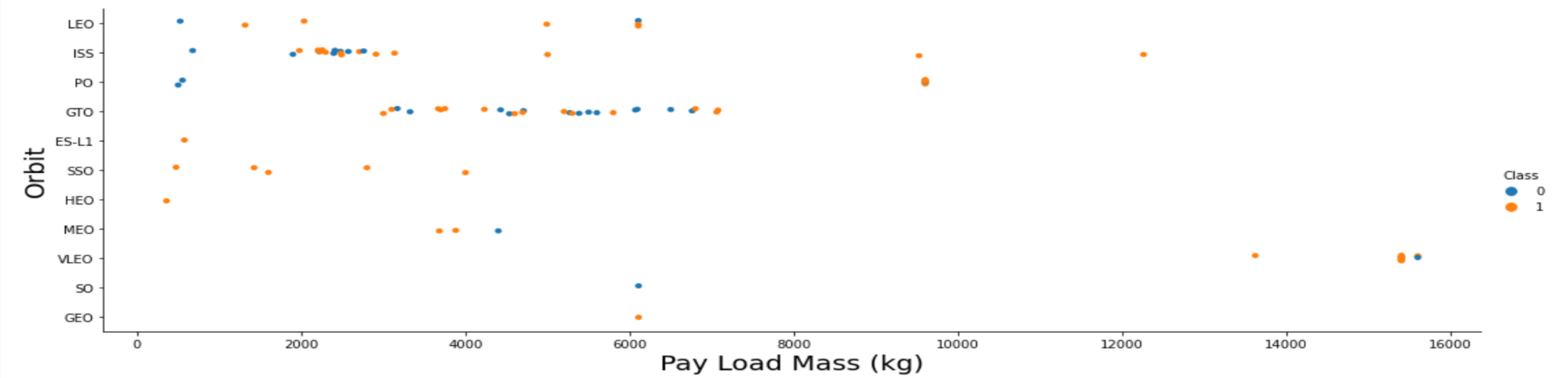
Flight Number vs. Orbit Type

- The more flight frequency on each orbit, the higher the success rate
- Flights to VLEO orbit seemed to have flight frequency above 40 with the higher success rate



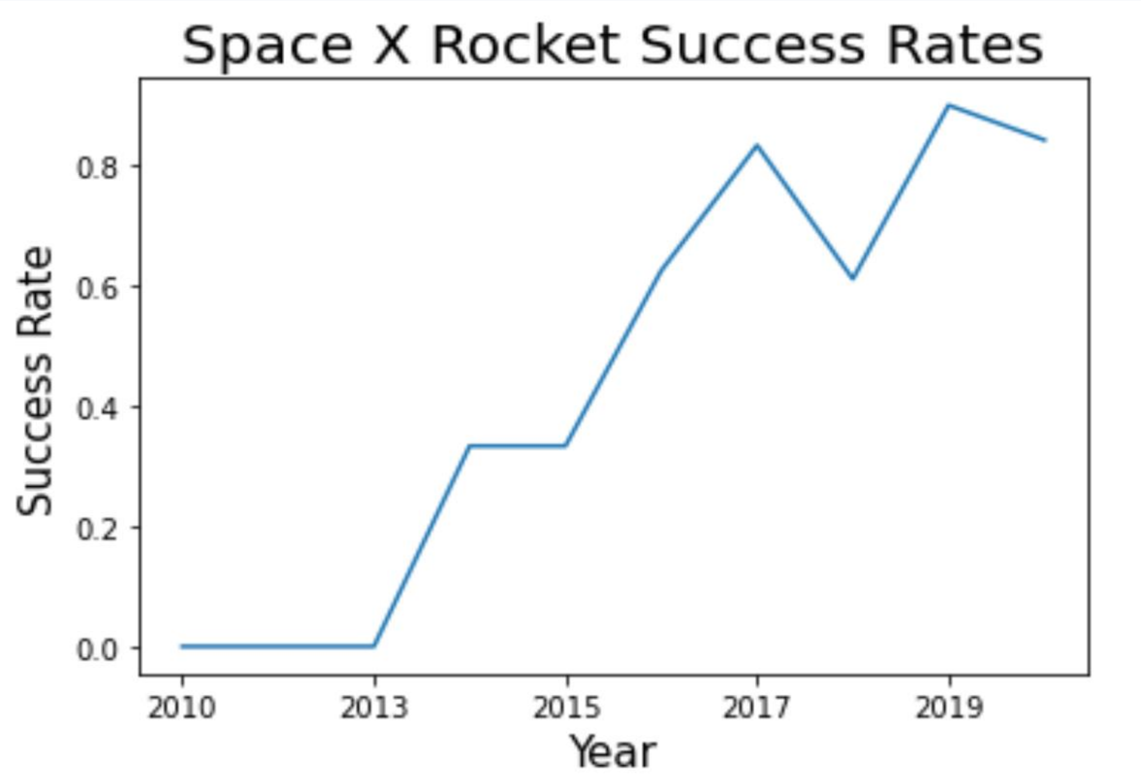
Payload vs. Orbit Type

- There is no significant relationship in success rate between Payload and to orbit GTO.
- Payload mass higher than 3000 kg showed success landing to orbit ISS
- Launch to orbit SSO showed the most landing success however the max payload launched was at 4000 kg
- More data is needed for orbits ES-L1, HEO, MEO, VLEO, SO and GEO to observe any trends



Launch Success Yearly Trend

- SpaceX success rates began to increase in the year 2013 until 2020 although there was a decrease in the rate in 2018.



All Launch Site Names

- There are 4 launch sites as below:

```
%sql select distinct(Launch_Site) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

- SQL query with **distinct()** function was used to obtain unique launch sites from the data as shown above.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The results above were obtained using the conditions **where**, **like** and **limit** as part of the query to obtain the result above.

Total Payload Mass

- The total Payload mass carried by boosters from NASA (CRS) was calculated as 45,596 kg.
- SQL **sum()** function for Payload mass was used and the **where** condition was included in the query to filter only customers from ‘NASA (CRS)’ as shown below:

```
%sql select sum(PAYLOAD_MASS__KG_) as "Total_Payload_Mass_KG" from SPACEXTBL where Customer == 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Total_Payload_Mass_KG
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928.4 kg as below.
- The result was obtained using SQL query
  - **avg** to calculate the average payload and
  - **where** query to filter the booster version

```
%sql select avg(PAYLOAD_MASS__KG_) as "Average_Payload_Mass_KG" from SPACEXTBL where Booster_Version == 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Average_Payload_Mass_KG
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was Dec 12th 2022.

```
%%sql

SELECT  min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2))  as "First_successful_GP_Landing" from SPACEXTBL
where "Landing _Outcome"="Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

First_successful_GP_Landing
20151222

- The result above was obtained using SQL **min()** query function for the minimum date and filtering the data by successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters that have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

```
%sql select Booster_Version from SPACEXTBL where [Landing _Outcome]='Success (drone ship)' and \
PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

- To achieve the results, SQL **where** clause was used to filter the boosters which successfully landed on drone ship and added the **and** condition to specify successful landing with payload mass between 4000 and 6000 kgs.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed mission outcomes are as below:

```
%sql select Mission_Outcome, Count(Mission_Outcome) as Count from SPACEXTBL group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The results above were achieved by applying **count()** function and grouping "Mission_Outcome".

Boosters Carried Maximum Payload

- The names of boosters that carried the maximum payload mass are as below:

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL) \
order by Booster_Version
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The results above were obtained using SQL subquery in the **where** clause and the **max()** function

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015:

```
%sql select substr(Date, 4, 2) as 'Month', substr(Date,7,4) as 'Year', Booster_Version, Launch_Site, [Landing _Outcome] \
from SPACEXTBL where substr(Date,7,4) = '2015' and [Landing _Outcome] = 'Failure (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Year	Booster_Version	Launch_Site	Landing _Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The results above were achieved by specifying the year as 2015, and using the **where** clause and the **and** condition to filter the failed landing outcomes in drone ship.
- Booster version and launch site were included in the summary

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

- Ranking of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order is as below:

```
%sql select Date, [Landing _Outcome], count ([Landing _Outcome]) as Landing_Outcome_Count from SPACEXTBL \
where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604' and '20170320' \
group by [Landing _Outcome] order by count([Landing _Outcome]) desc
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Landing _Outcome	Landing_Outcome_Count
22-05-2012	No attempt	10
08-04-2016	Success (drone ship)	5
10-01-2015	Failure (drone ship)	5
22-12-2015	Success (ground pad)	3
18-04-2014	Controlled (ocean)	3
29-09-2013	Uncontrolled (ocean)	2
04-06-2010	Failure (parachute)	2
28-06-2015	Precluded (drone ship)	1

- Results above were achieved by selecting the date, landing outcome, count() function to count the landing outcomes together with the where clause to filter the date of landing outcomes between 2010-06-04 and 2017-03-20.

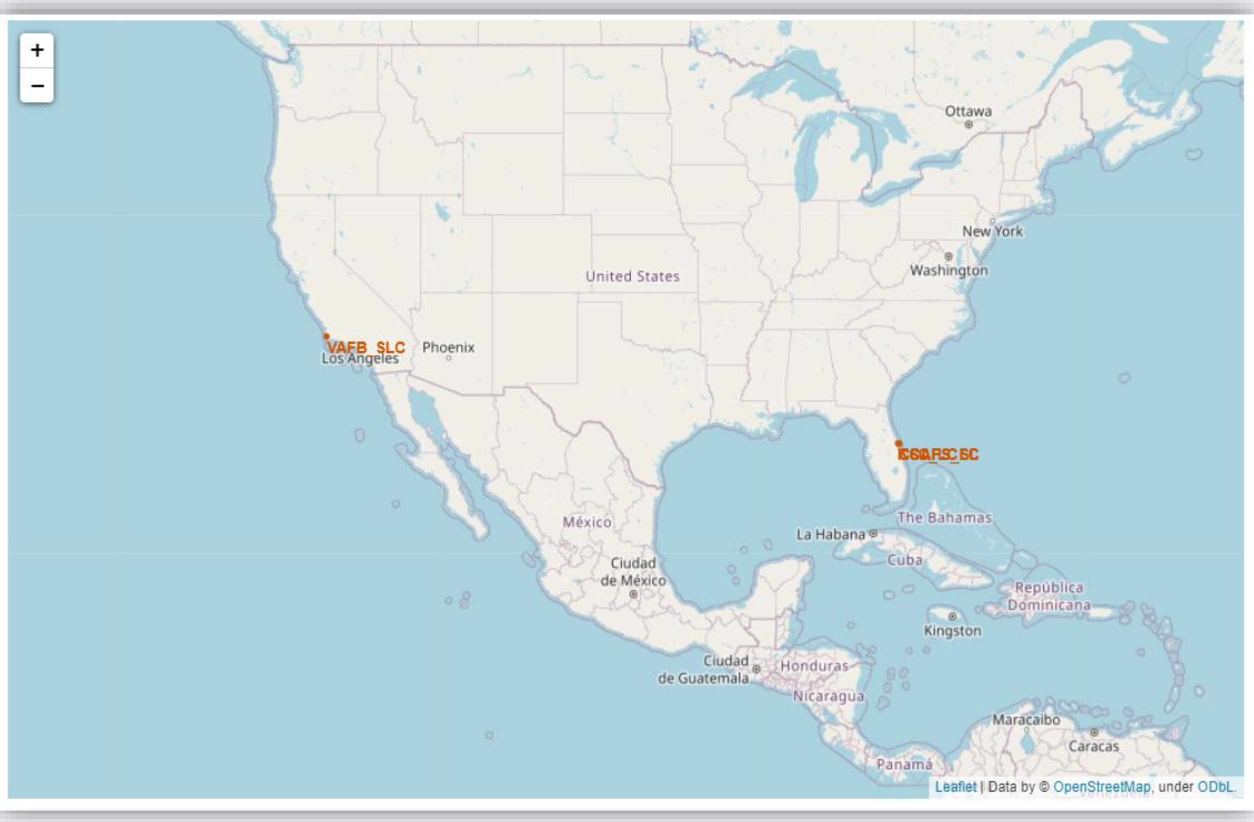
- Group by clause was included in the query to group the landing outcomes and Order by clause was used to list the grouped landing outcome in descending order.

Section 3

Launch Sites Proximities Analysis

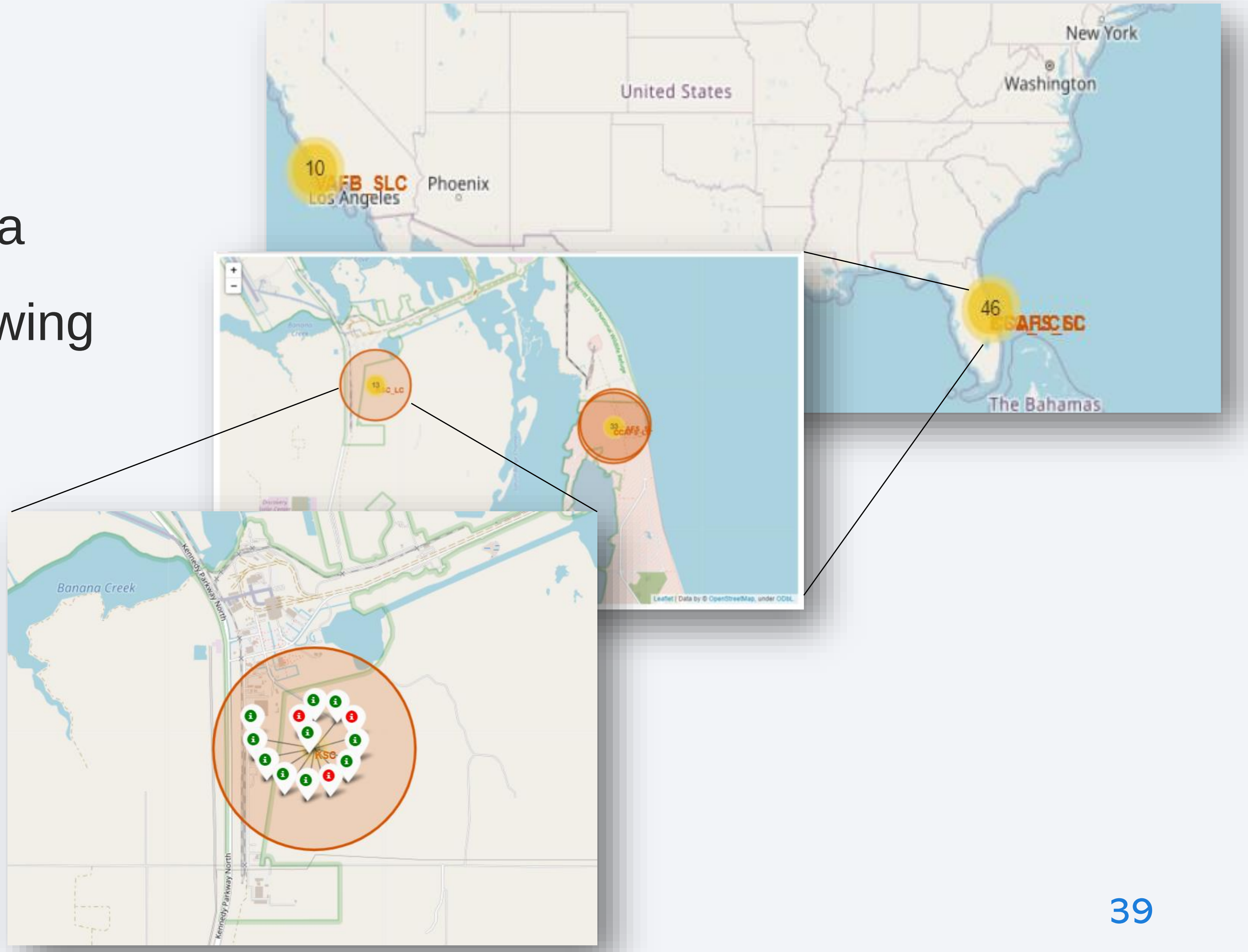
SpaceX Launch Sites

- The figure shows all SpaceX launch sites are located in the United States of America (USA).
- It is noted that the sites are located on the east coast (Florida, USA) and the west coast (California, USA).



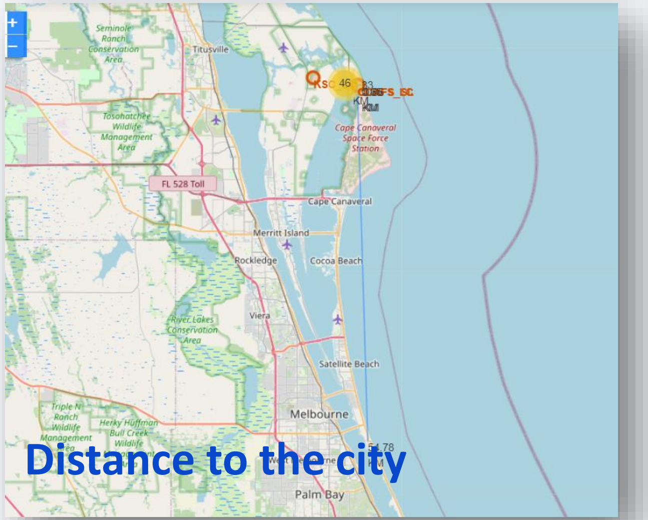
Launch sites' markers with color labels

- KSC LC-39A launch site in Florida
- Launch site with color labels showing outcomes of launch
- Color labels:
  - Green marker –Successful
  - Red marker - Failed



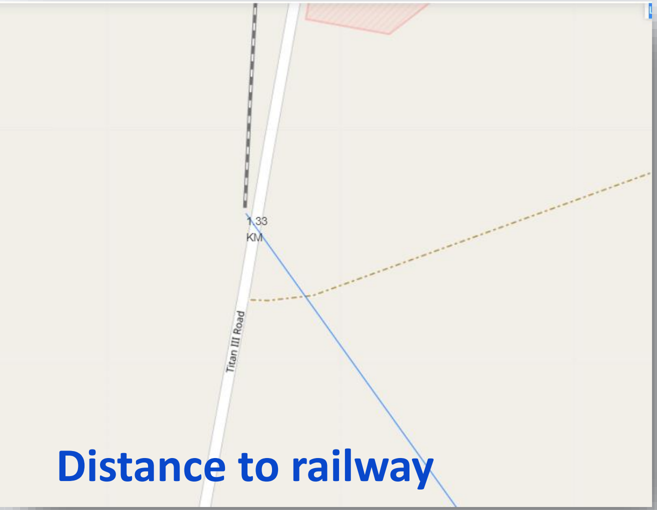
Launch Sites Distance to Closest Landmarks

- Both launch sites CCAFS are situated near highway and railway which indicates good logistic aspects
- However, both sides are relatively isolated from the nearest city and close to the coastline for launch purposes.



Distance to the city

Distance to the coast and the highway

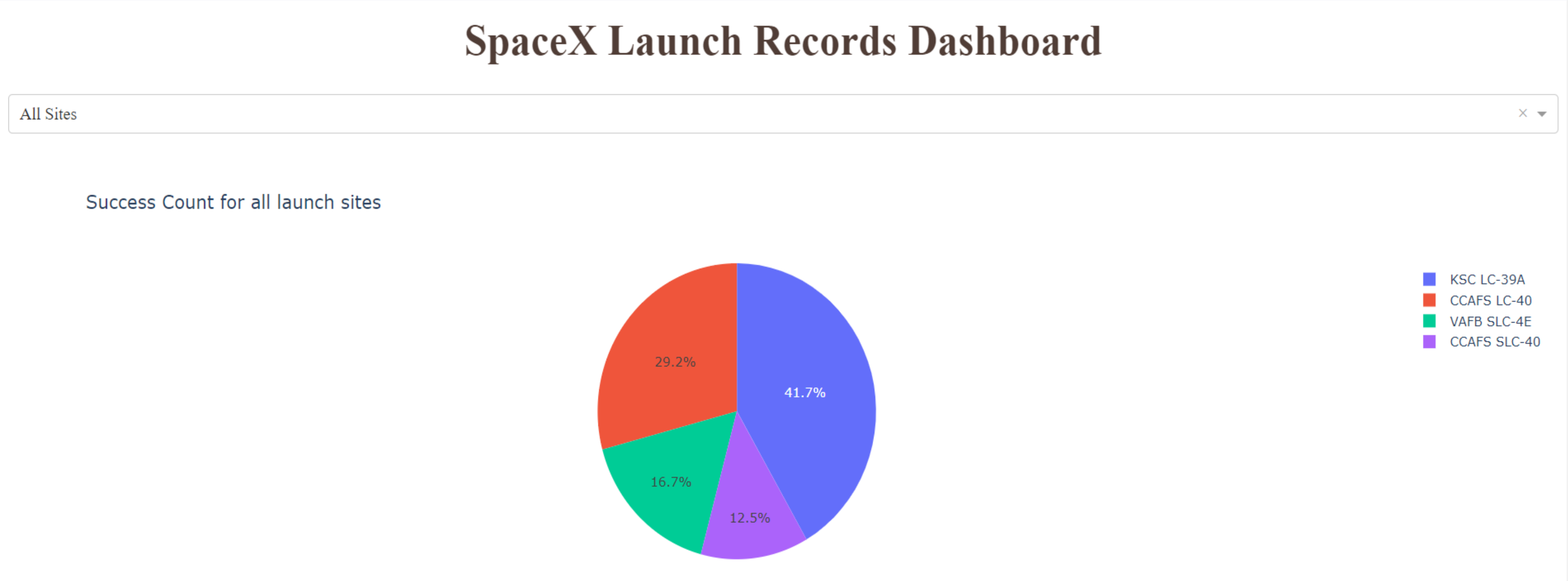


Distance to railway

Section 4

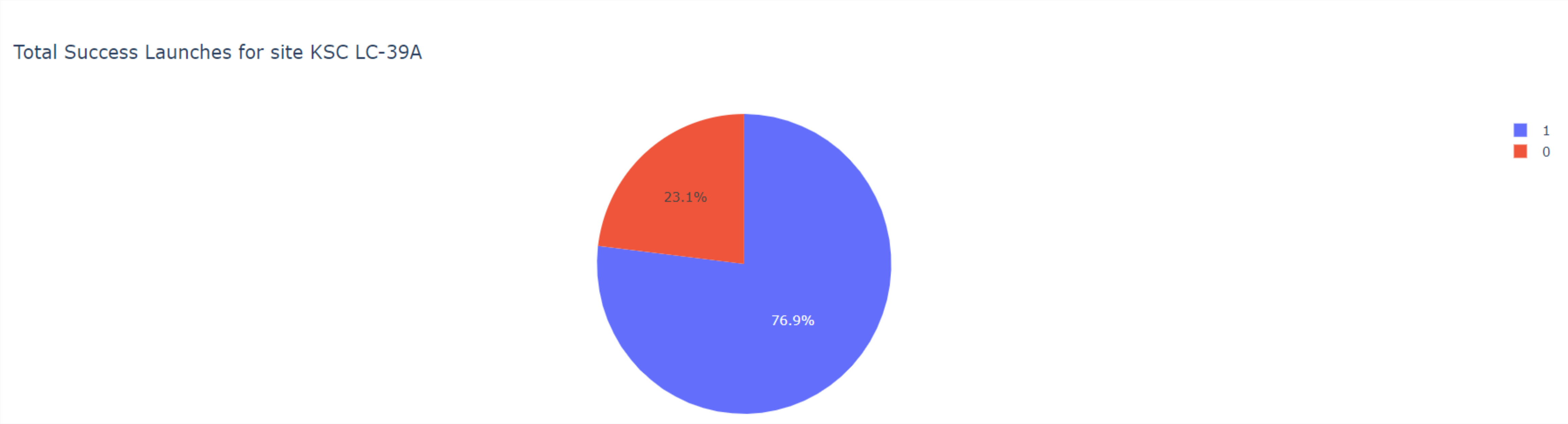
Build a Dashboard with Plotly Dash

Launch Success Counts (All sites)



- KSC-LC-39A had the most successful launches of all the sites with 41.7% success rate.
- CCAFS SLC-40 had the least success count of all the sites with 12.5% success rate

Launch Success Ratio for Site KSC LC-39A



- Site KSC LC-39A had proven 76.9% success rate while the remaining 23.1% was reported as failed launch.

Payload vs Launch Outcome with Different Boosters

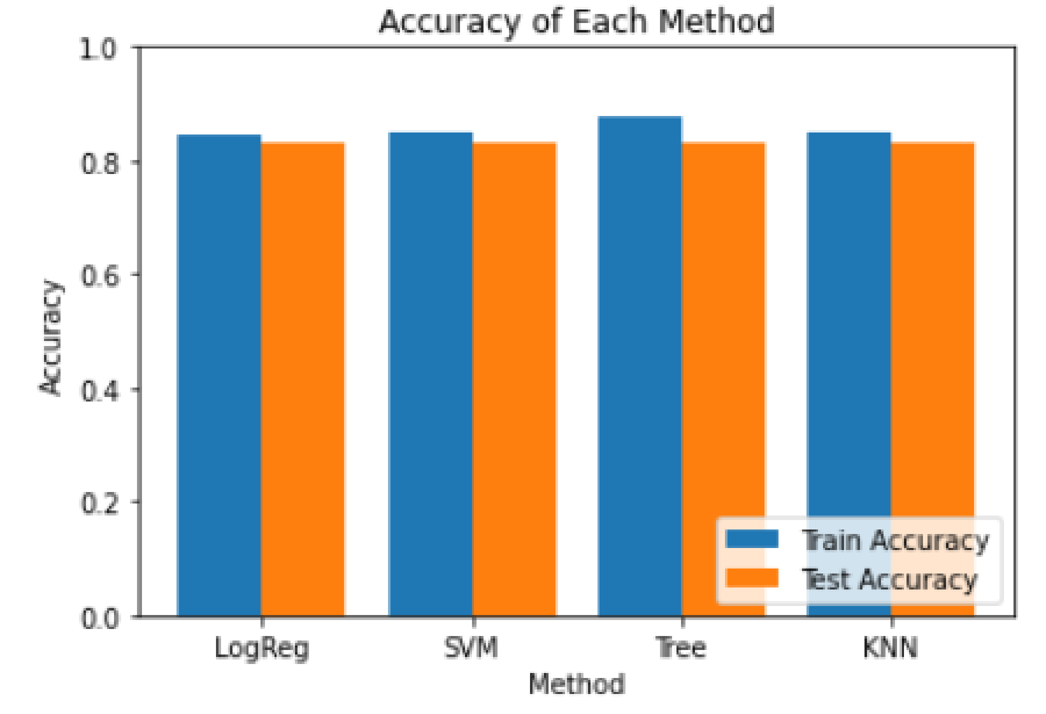
- Booster version FT with payloads less than 6000 kgs showed the most successful launch count compared to other boosters.
- More data is needed to evaluate launch outcome boosters with payloads heavier than 6000 kgs.

Section 5

Predictive Analysis (Classification)

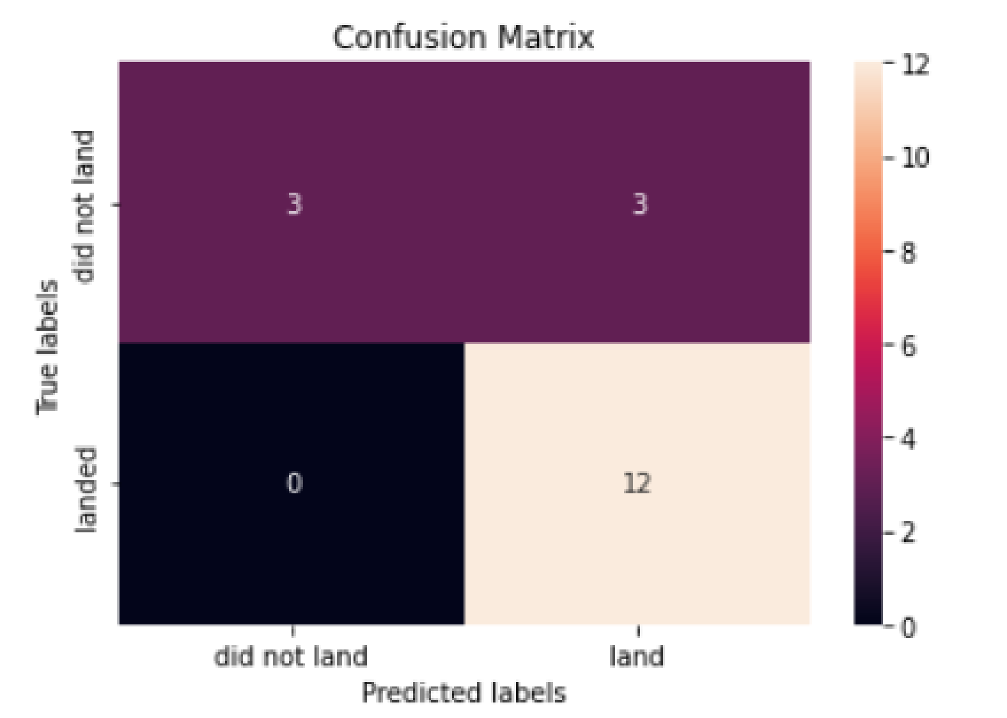
Classification Accuracy

- 4 classification models were evaluated:-
  1. Logistic Regression
  2. Support Vector Machine (SVM)
  3. Decision Tree
  4. K-Nearest Neighbors (KNN)
- The best model with the highest classification accuracy is the Decision Tree classifier with an accuracy of 87%.



Confusion Matrix

- The confusion matrix of the Decision Tree classifier model has no difference from the other models.
- Although it has the highest accuracy, false positives and false negatives are noted as shown in the figure.



Conclusions

- FT booster version with low payloads (<6000 kgs) has a high success landing outcome compared to other boosters although more data is needed to confirm this result.
- Success rate in rocket launches for SpaceX has increased since 2013 until 2020; this showed improvement in launch missions over time.
- Launch sites are located on the east and the west coast of the USA and farther away from cities. The best launch site is KSC LC-39A with 41.7% success rate.
- The Decision Tree classifier model has been tested as the best model for use in predicting launch outcomes for this project.

Thank you!