IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jessica Mariño
15-06-2022

Outline

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

Executive Summary

In this report we present the different methodologies applied to find the key points by which SpaceX succeeds and then predict from this data the possibility of competing against it.

This report is summarized in the next points:

- Data collection and preparation performing data wrangling.
- Identify important parameters and tendences applying EDA and SQL
- Visualization of variables using Folium and Plotly Dash
- Perform predictive analysis using classification models and find the best model.

Introduction

- Project background and context
  - **SpaceX** advertises Falcon 9 rocket launches 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
- Problem
  - An **alternate company** wants to bid against SpaceX for a rocket launch
  - We need to predict if the Falcon 9 first stage will land successfully to determine the cost of launch for the alternate company

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Request to the SpaceX API and Clean the requested data
- Perform data wrangling
  - Convert booster landed outcomes to Training Labels: 1 successful 0 unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class, standardize the data, split into training data and test data
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

- Using Request and Pandas python libraries

Request the SpaceX launch data
- GET requests

Clean data
- Convert data into a dataframe
- Take a subset of data
- Remove rows with multiple cores and payloads
- Restrict dates of the launches

Extract information
- use GET functions

Filter data
- Keep only Falcon 9 launches

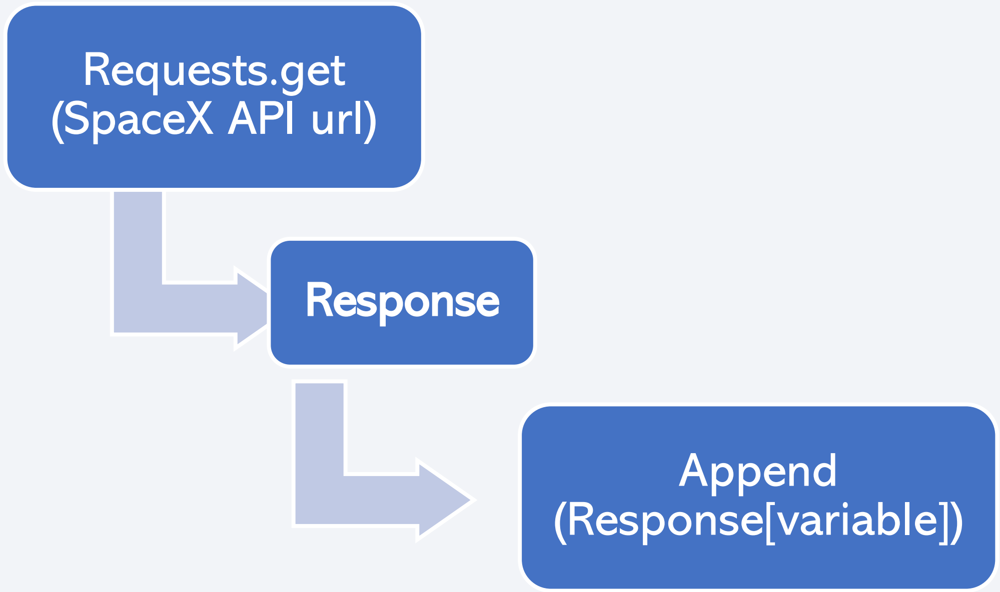
Deal with Missing Values
- Replace NAN Payload mass values with the mean value

Data Collection – SpaceX API

- Data Collection:

• Date • Booster Version • Payload Mass, and • Orbit • Launch Site • Outcome • Flights • GridFins	• Reused • Legs • Landing Pad • Block • Reused Count • Serial • Longitude • Latitude

- GET function structure (use to obtain Data):



- GitHub URL of the completed SpaceX API calls notebook: https://github.com/JessicaMarino/dataScience_repo/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia page
  - Use Requests and beautifulsoup4 python libraries

Request the Falcon9 Launch Wiki page from its URL

- data = requests.get(url).text
- *create a BeautifulSoup object from a response text content*

Extract all column names from the HTML table header

- Find table of the actual launch records
- Extract column names

Create a data frame by parsing the launch HTML tables

- Create an empty dictionary with keys from the extracted column names
- Extract each table
- Get table elements
- Use functions to scrap table
- Append key cells into the dictionary
- Create a dataframe from the dictionary

- GitHub URL of the completed web scraping notebook: https://github.com/JessicaMarino/dataScience_repo/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

- Apply EDA to find patterns in the data and determine training labels.

Calculate the number of launches on each site

- df["LaunchSite"].value_counts()

Calculate the number and occurrence of each orbit

- df["Orbit"].value_counts()

Calculate the number and occurence of mission outcome per orbit type

- landing_outcomes = df["Outcome"].value_counts()
- create a set of bad outcomes

Create a landing_class label from Outcome column

- Class = 0: bad outcome
- Class = 1: success outcome

Determine the success rate

- df["Class"].mean()

- GitHub URL of your completed data wrangling notebooks: https://github.com/JessicaMarino/dataScience_repo/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Perform EDA and Feature Engineering using Pandas, Matplotlib, and seaborn

Visualize relations:

- Payload Mass vs. Flight Number and overlay the outcome of the launch
- Launch Site vs. Flight Number
- Launch Site vs. Payload
- Success rate of each orbit
- Orbit type vs. Flight Number
- Orbit type vs. Payload
- launch success yearly trend

Select features to predict in the module.

Create dummy variables to categorical columns

- get_dummies()

Cast all numeric columns to float64

- .astype('float64').dtypes

- GitHub URL of completed EDA with data visualization notebook: https://github.com/JessicaMarino/dataScience_repo/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

1. Load dataset and store in a Db2 database

2. Connect to the database

   - Load SQL extension and establish a connection

3. Write and execute SQL queries to display information
   - names of the unique launch sites
   - launch sites begin with the string 'KSC'
   - total payload mass carried by boosters launched by NASA (CRS)
   - average payload mass carried by booster version F9 v1.1
   - Total number of successful and failure mission outcomes
   1. Date when the first successful landing outcome in drone ship was achieved
   2. Boosters that have success in ground pad and have payload mass in a certain range
   3. Booster versions that have carried the maximum payload mass.
   4. Records of successful landing outcomes in ground pad , for months in a year
   5. Rank the count of successful landing_outcomes between a delimited time.

- GitHub URL of completed EDA with SQL notebook: https://github.com/JessicaMarino/dataScience_repo/blob/master/jupyter-labs-eda-sql-edx.ipynb

Build an Interactive Map with Folium

1. Create a Folium.Map with initial center location in NASA Johnson Space Center at Houston, Texas
2. Mark all launch sites on a map
   - folium.Circle → add a highlighted circle area on coordinates
   - folium.Marker → create an icon showing their names as a popup labels on coordinates
3. Mark the success/failed launches for each site on the map
   - icon_color property → add green markers for success launches and red markers for failed launches
   - MarkerCluster() → simplify a map containing many markers having the same coordinate.
4. Calculate the distances between a launch site to its proximities based on Lat and Long values
   - MousePosition → get the coordinate (Lat, Long) for a mouse over on the map
   - folium.Marker → create marks at closest city, railway, and highway on the map
5. Draw a line between the marker to the launch site
   - folium.PolyLine → draw a line between the marker to the launch site
   - icon property → Display the calculated distance between coastline point and launch site

GitHub URL of completed interactive map with Folium map:
https://github.com/JessicaMarino/dataScience_repo/blob/master/lab_jupyter_launch_site_location.ipynb

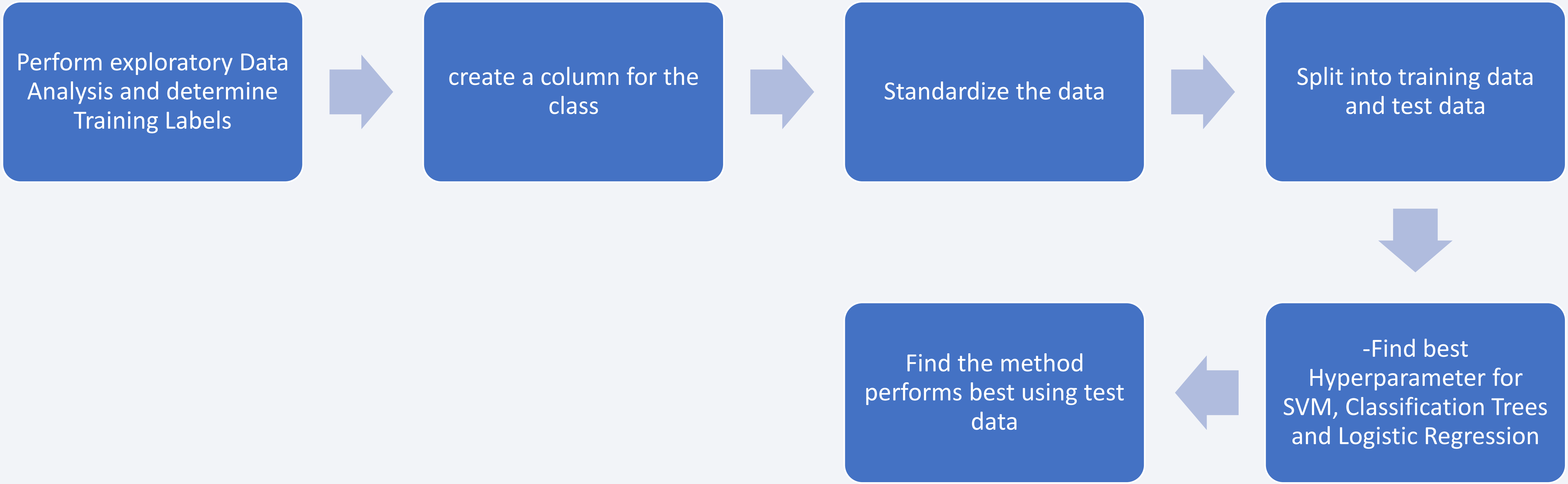
Build a Dashboard with Plotly Dash

The dash application "app" contains an app layout with the following elements:

1. Launch Site Drop-down Input Component
   - to select different launch sites
2. Callback function to render success-pie-chart based on selected site dropdown
   - to visualize launch success counts.
3. Range Slider to Select Payload
   - to select different payload range and identify some visual patterns
4. Callback function to render the success-payload-scatter-chart scatter plot
   - to observe how payload may be correlated with mission outcomes for selected site(s).

GitHub URL of completed Plotly Dash lab:
https://github.com/JessicaMarino/dataScience_repo/blob/master/spacex_dash_app2.py

Predictive Analysis (Classification)



- GitHub URL of completed predictive analysis lab: https://github.com/JessicaMarino/dataScience_repo/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

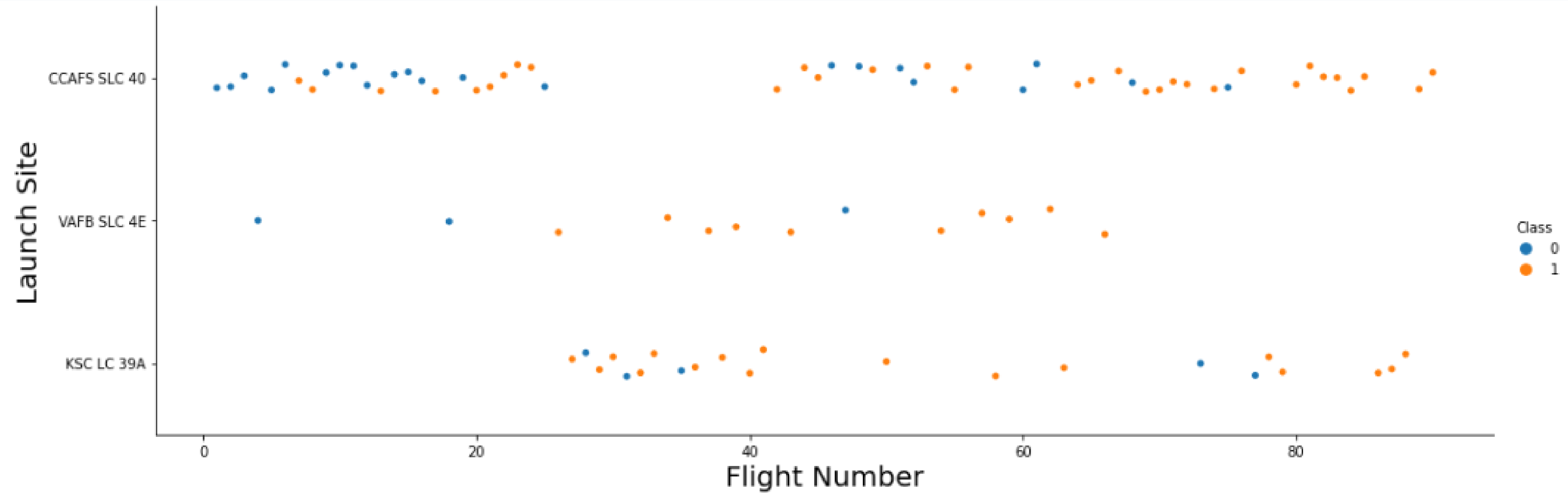
Results

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

Section 2

Insights drawn from EDA

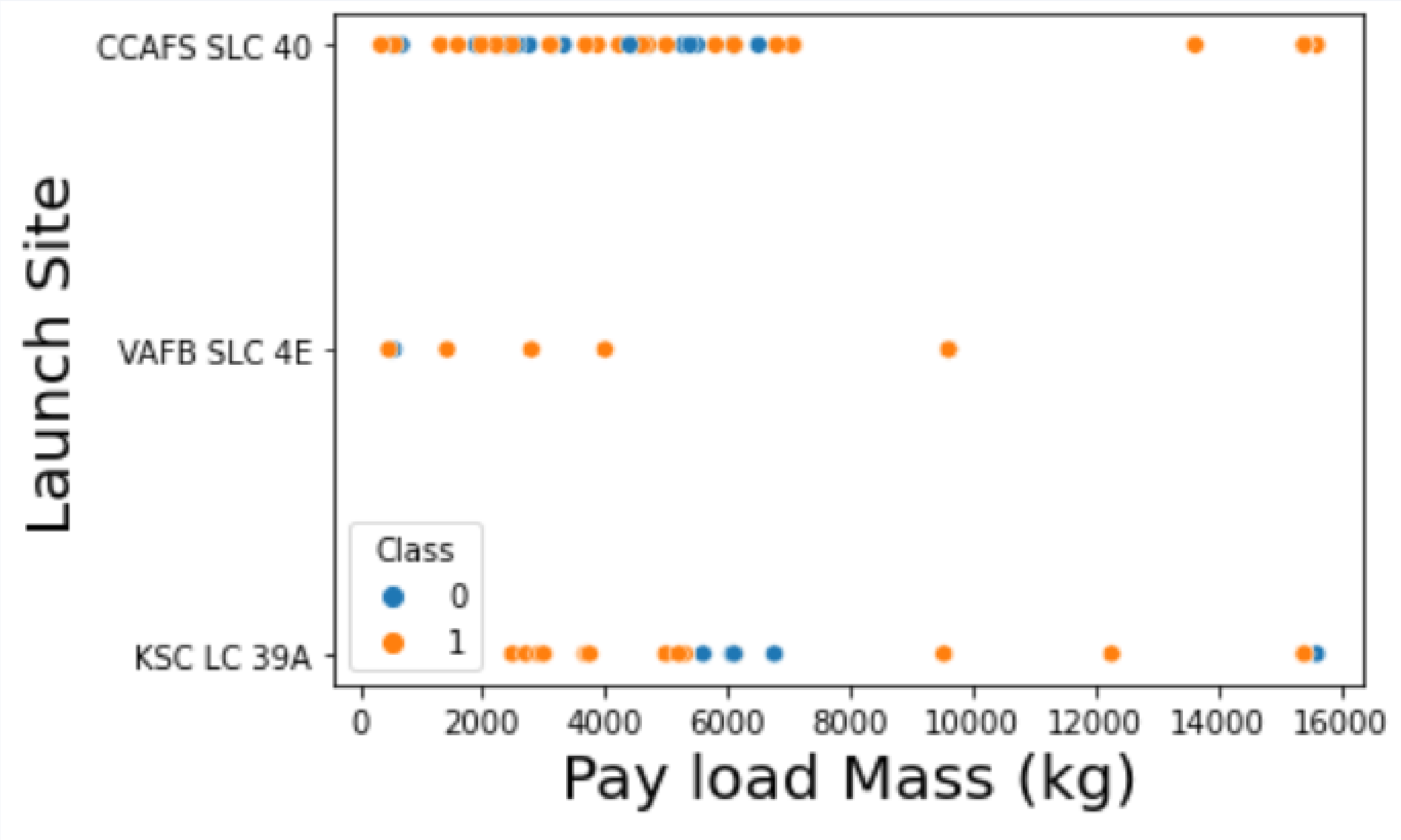
Flight Number vs. Launch Site



- We see that as the flight number increases, the first stage is more likely to land successfully for the three Launch Site.

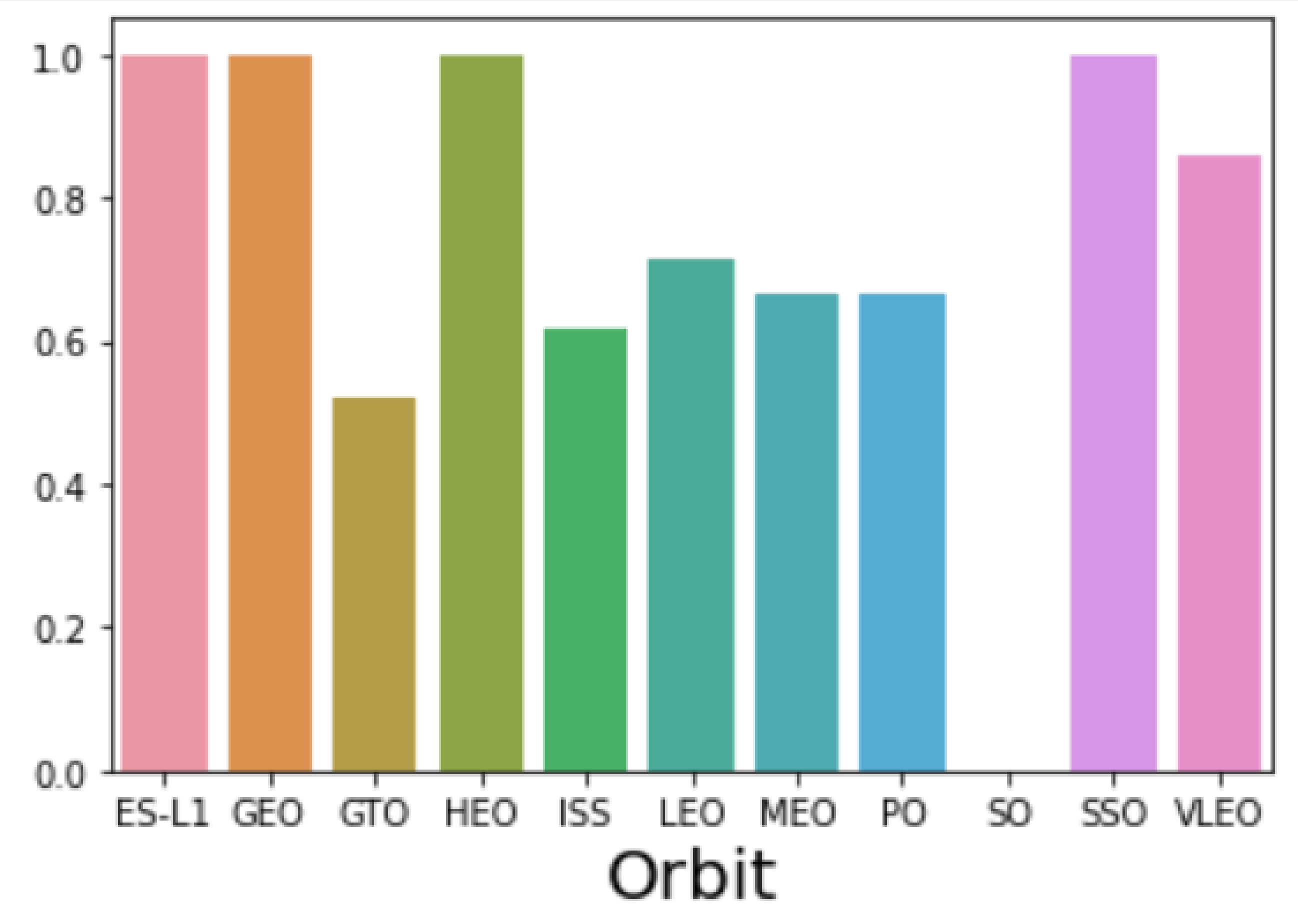
Payload vs. Launch Site

- In VAFB-SLC 4E launch site, no rockets launched for heavy payload mass (greater than 10000).
- In general, heavy payload mass is very seldom carried.



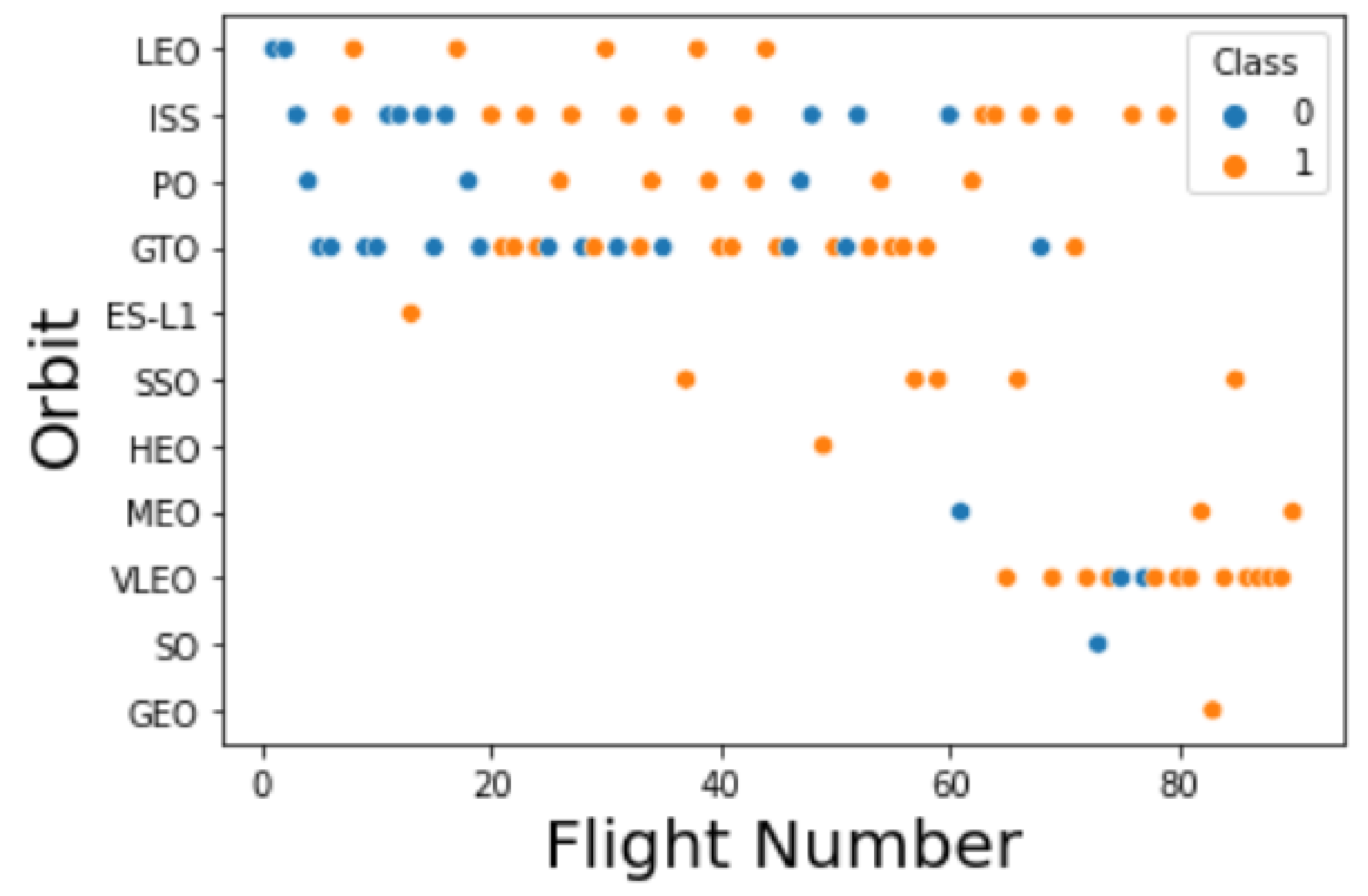
Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, and SSO orbits have high success rate.



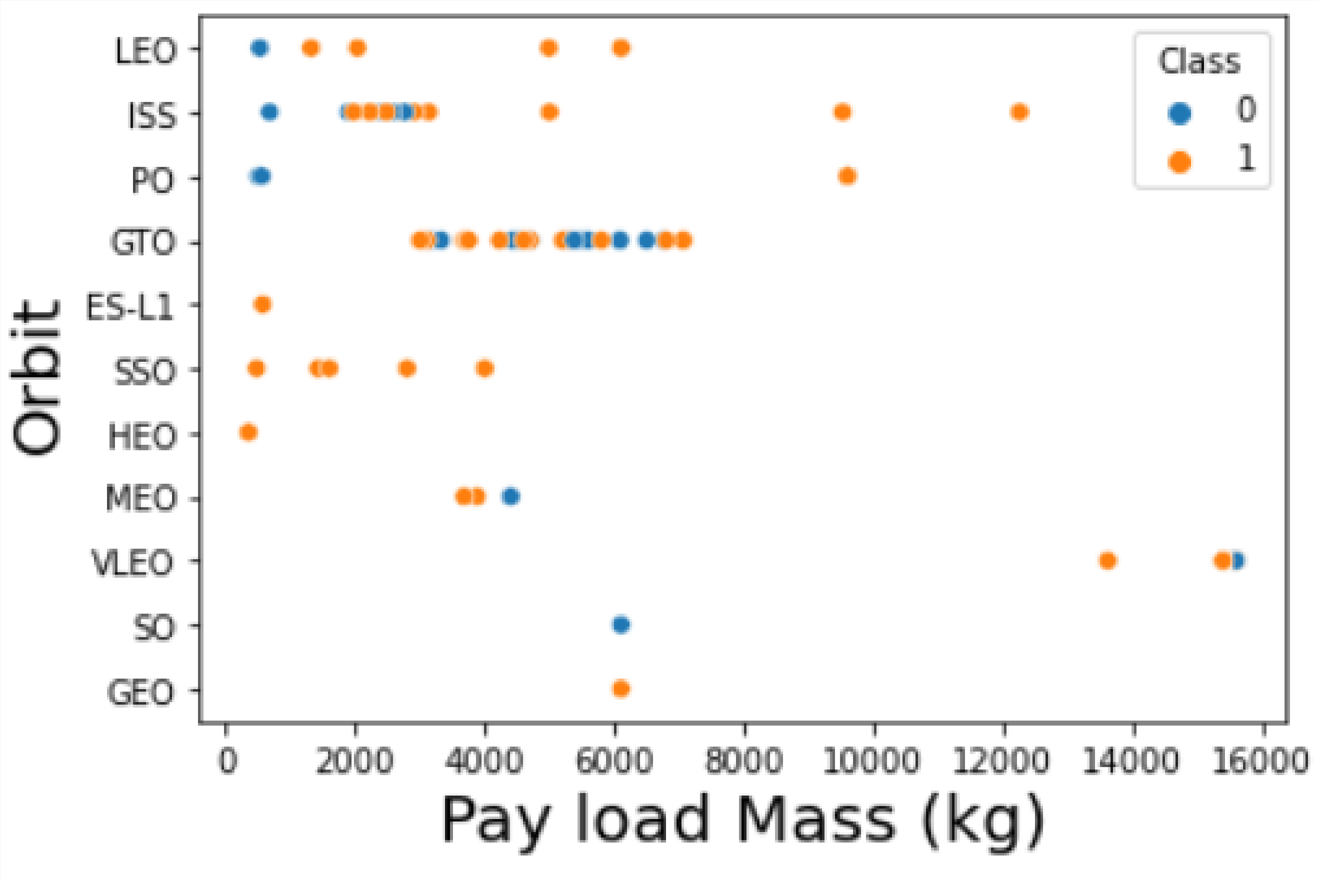
Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



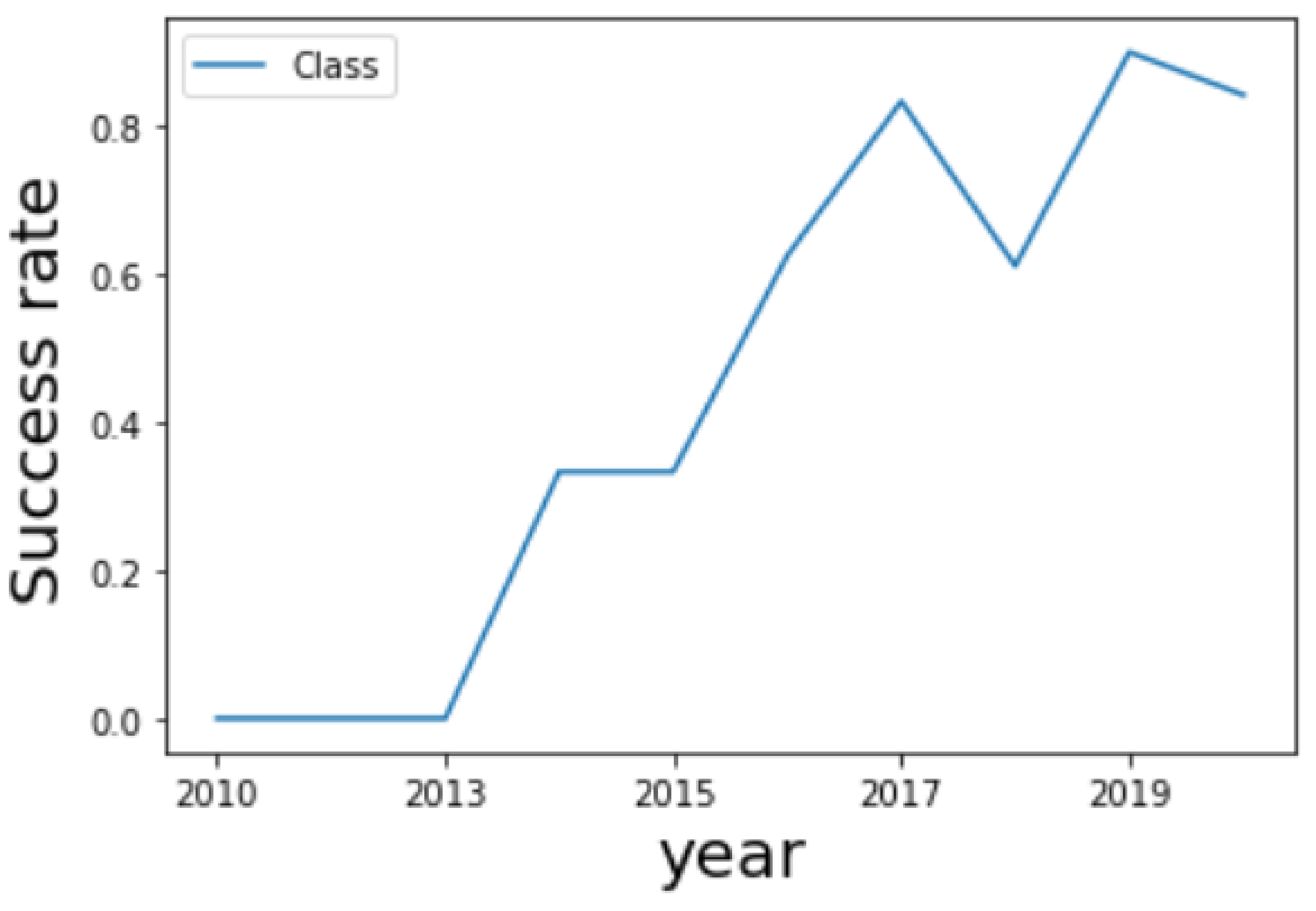
Payload vs. Orbit Type

- ISS orbit has more positive landing possibilities for heavy payloads.
- LEO and SSO are suitable for positive landing carrying a light payload mass.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

There are four launch Sites:

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

Launch Site Names Begin with 'KSC'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Most of the Landing outcomes for the launch site KSC LC-39A were successful

Total Payload Mass

- Query:

```
%%sql
select SUM(PAYLOAD_MASS__KG_)
from SPACEXTBL
where CUSTOMER LIKE 'NASA (CRS)';
```

- Result:

Total payload carried by boosters from NASA = 45 596 kg

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 = 2 928 kg

We can conclude that the booster version F9 v1.1 carries light payload mass.

First Successful Ground Landing Date

The First successful Ground Landing Date was 2016-04-08

After the first successful Ground Landing in 2016, the successful rate is not less than 60%.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Version

- F9 FT B1032.1
- F9 B4 B1040.1
- F9 B4 B1043.1

Only three Booster versions successfully land with Payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Mission Outcomes

- Failure (in flight) 1
- Success 99
- Success (payload status unclear) 1

Almost any mission outcomes successful.

Boosters Carried Maximum Payload

Booster version:	
• F9 B5 B1048.4	• F9 B5 B1049.5
• F9 B5 B1049.4	• F9 B5 B1060.2
• F9 B5 B1051.3	• F9 B5 B1058.3
• F9 B5 B1056.4	• F9 B5 B1051.6
• F9 B5 B1048.5	• F9 B5 B1060.3
• F9 B5 B1051.4	• F9 B5 B1049.7

We can see that the Booster versions F9 B5 B1xx can carry heavy Payloads

2015 Launch Records

Month	landing outcome	booster version	launch site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

In 2015, all the landing outcomes were successful, and almost all of them took place in KSC LC-39A launch site. The booster versions were F9 FT B1xx and F9 B4 B1xx.

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

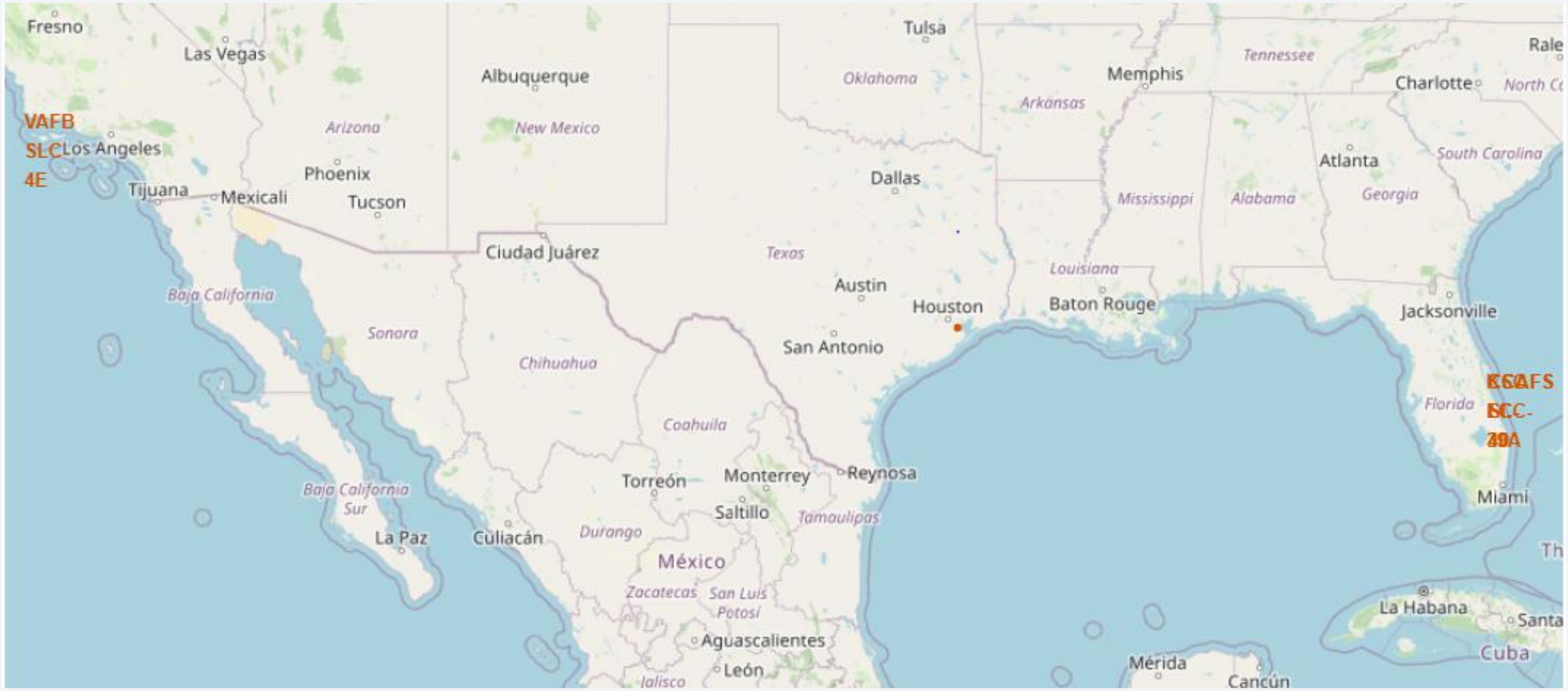
Landing outcome	
Success (ground pad)	3
Success (drone ship)	5

There are two types of landing ground pad and drone ship. The second has more possibilities of success.

Section 3

Launch Sites Proximities Analysis

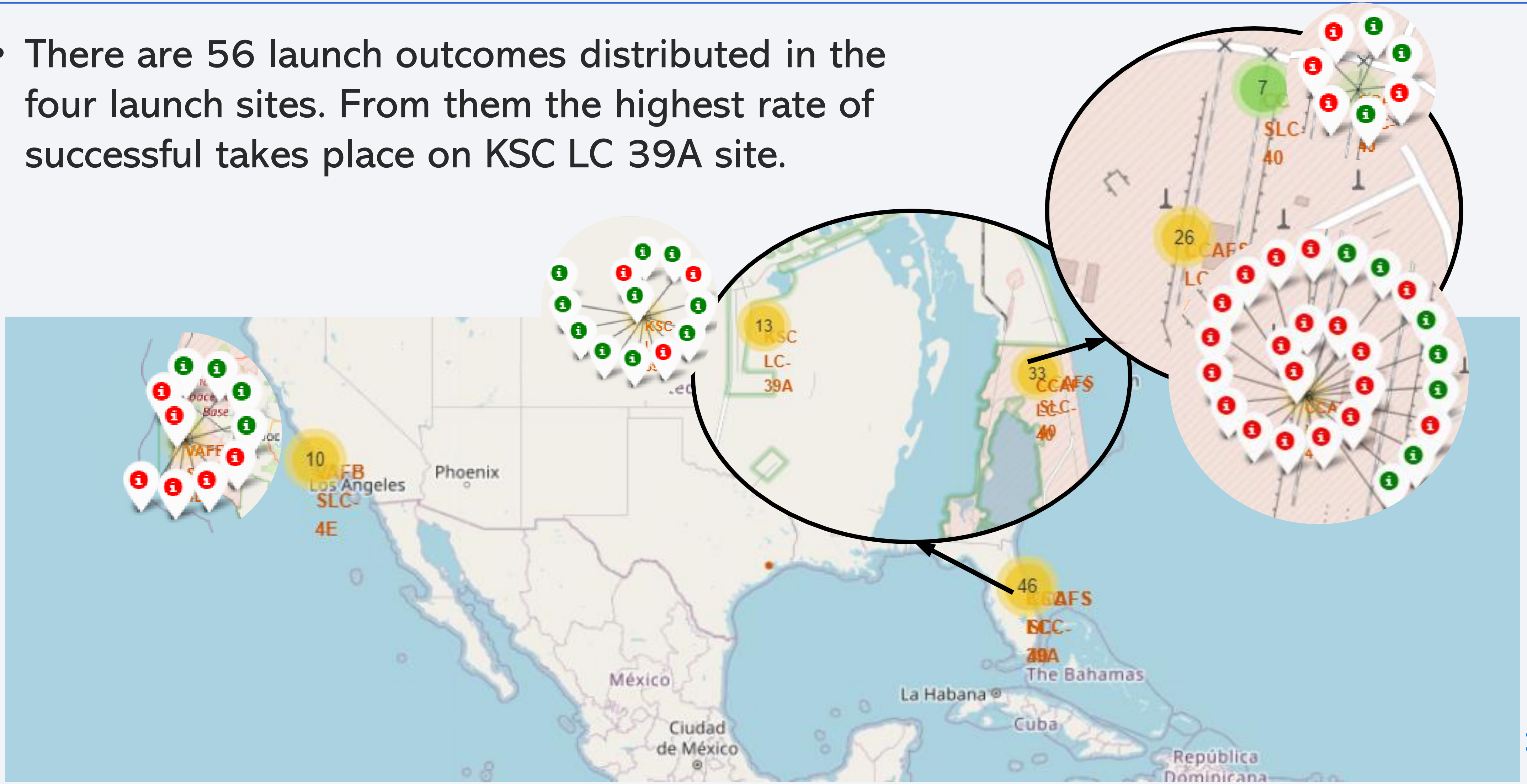
Launch site's location on global map



- The four launch sites are in very close proximity to the coast. Three of them are in Florida and one in Los Angeles
- All launch sites are in proximity to the Equator line.

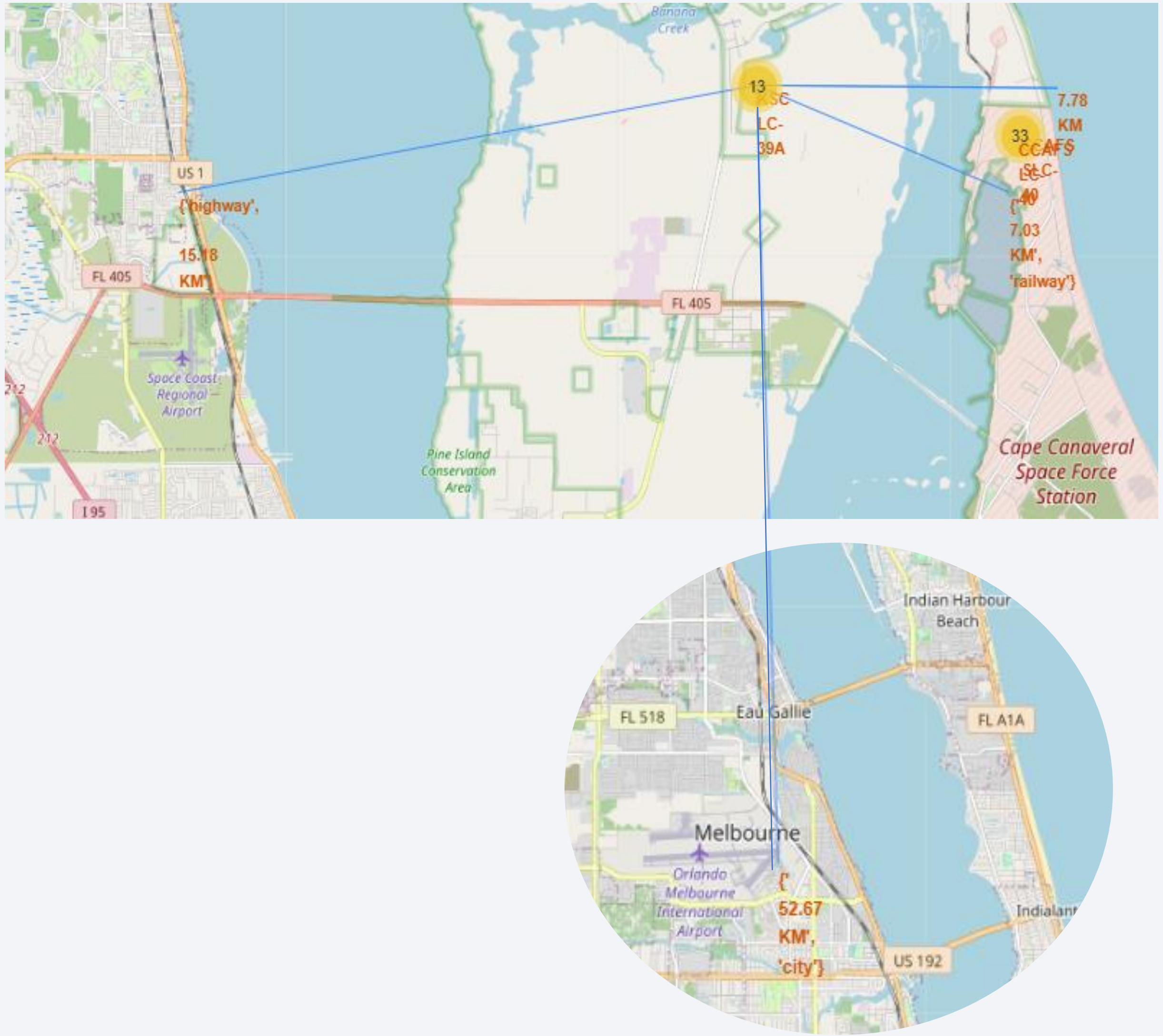
Launch outcomes on the map

- There are 56 launch outcomes distributed in the four launch sites. From them the highest rate of successful takes place on KSC LC 39A site.



KSC LC 39A and its proximities

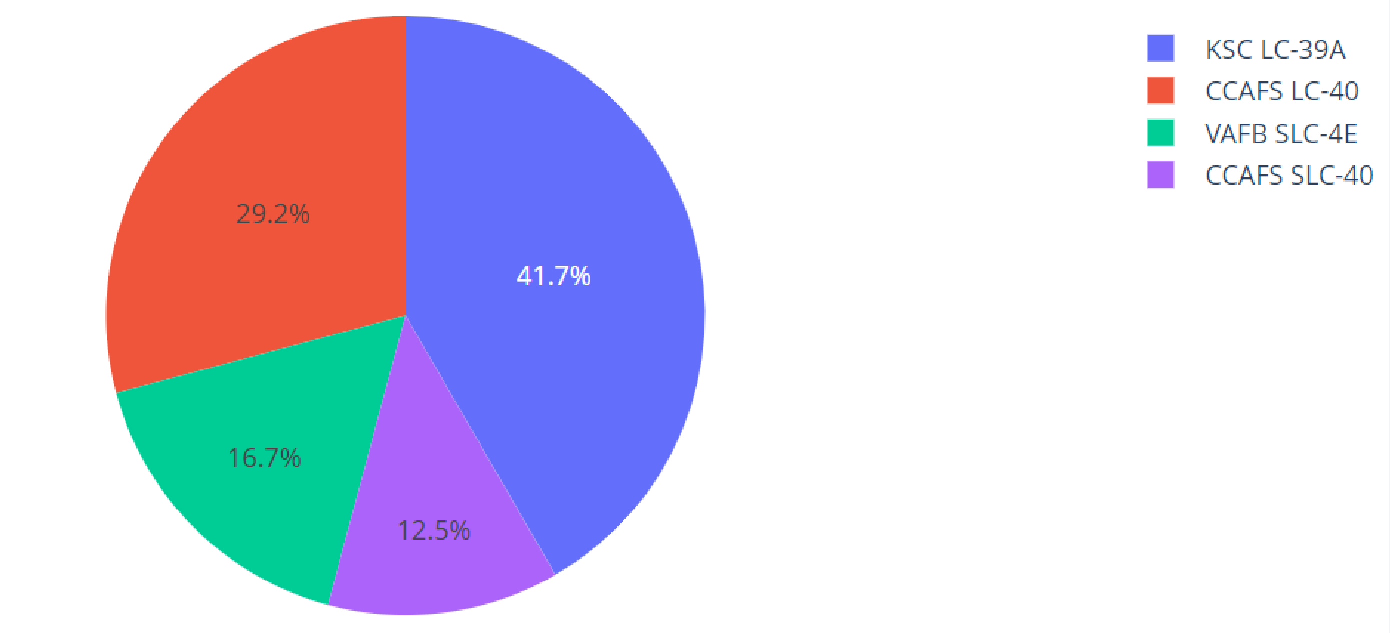
- The Launch site KSC LC 39A is very close to the coastline and the transportation routes.
- Distance to proximities
    - To Atlantic Coastline: 7.78km
    - To highway: 15.78km
    - To railway: 7.03km
    - To city Melbourne: 52.67km



Section 4

Build a Dashboard with Plotly Dash

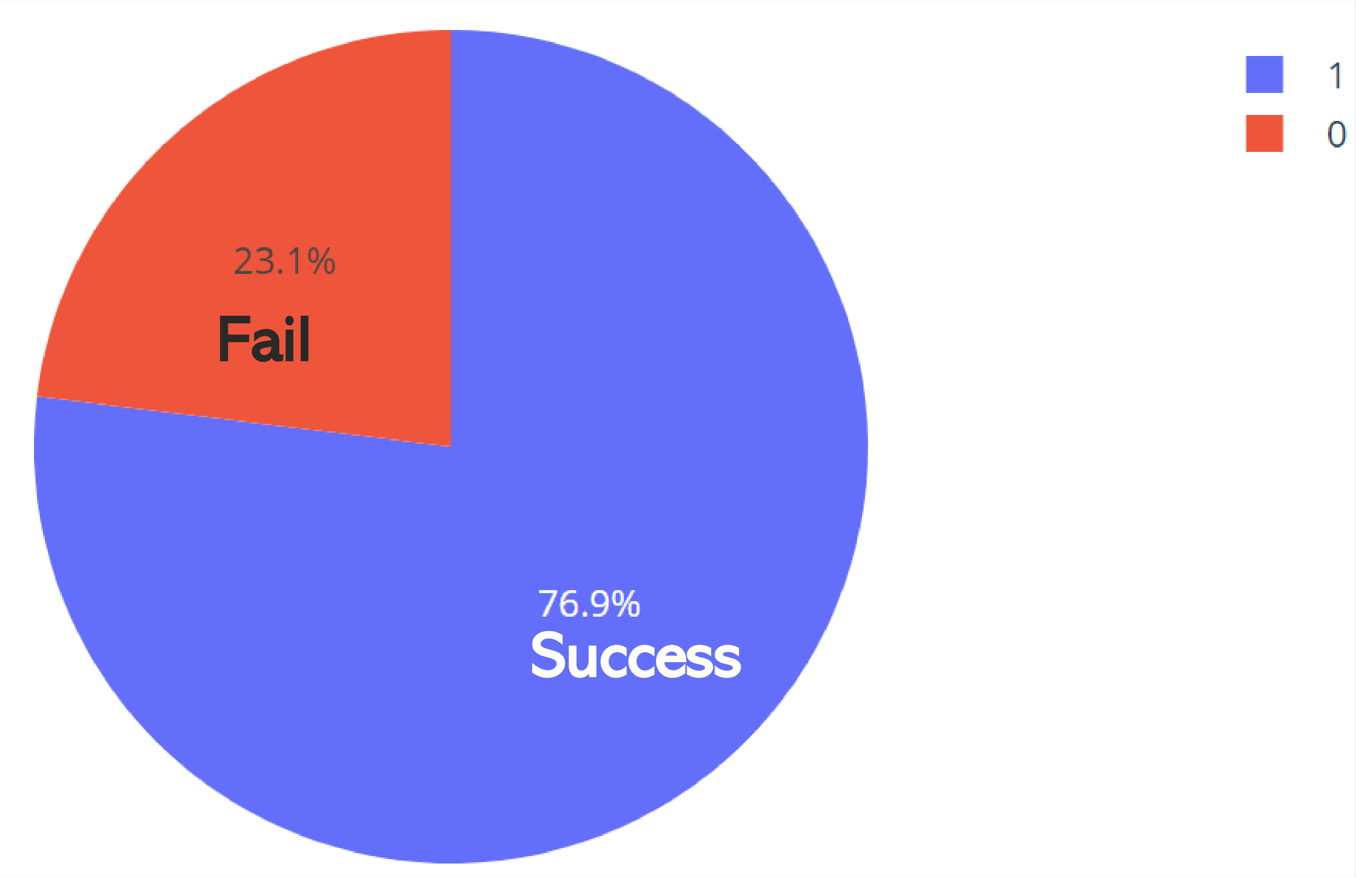
Total success launches by site



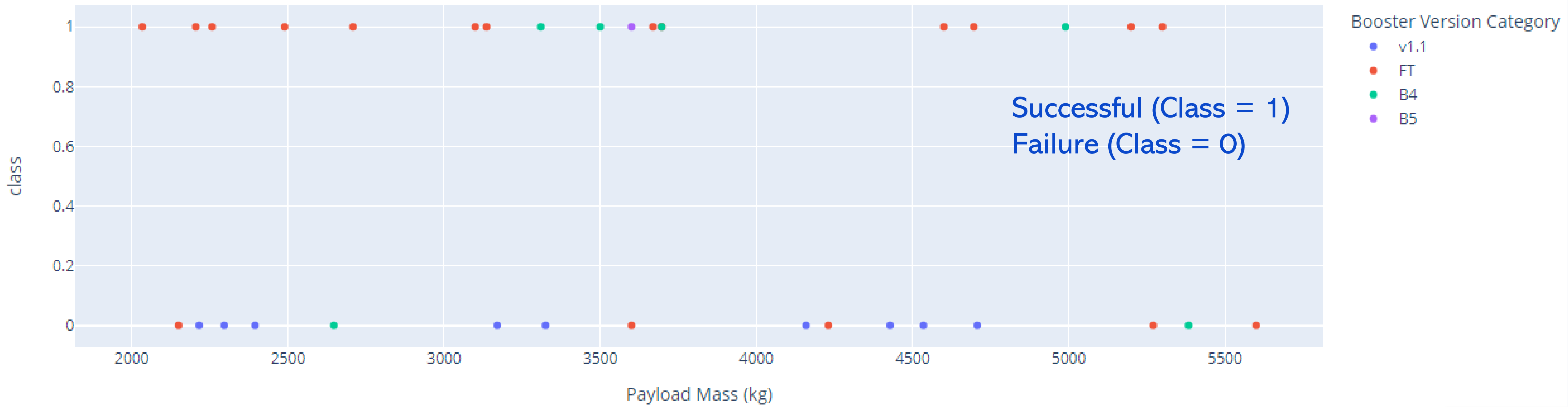
- The site with highest success is KSC LC-39A with almost 42% of the total success launches.
- The second success launches site is CCAFS LC-40 with 29%.
- The two places are near to Atlantic coast.

Total success launches in KSC LC-39A

- The KSC sites. LC-39A launch site has the highest number of success launches of the four analyzed sites.
- We can observe that for each 10 launches in this site, almost 8 are successful.



Payload Mass range with largest success rate



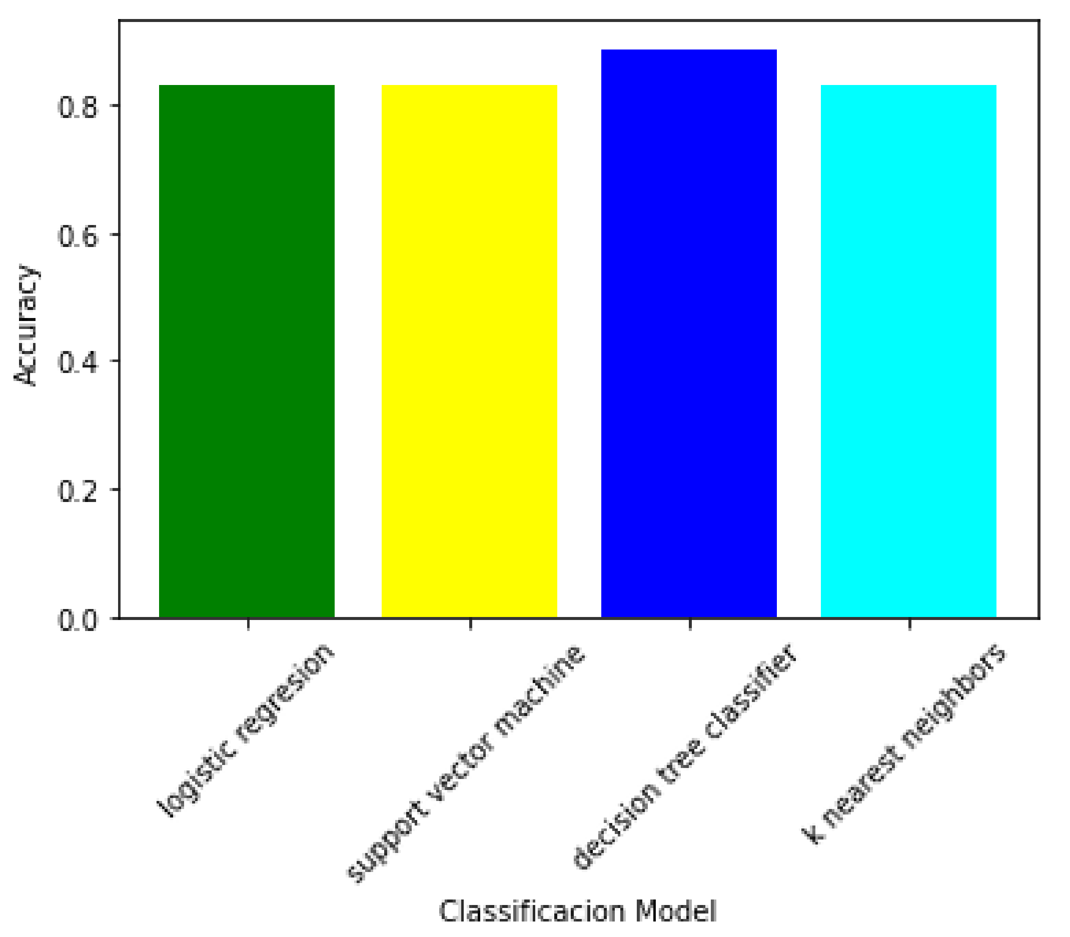
- The payloads from 2 000kg to 6 000kg is the range with largest success rate.
- In this range, the Booster version category FT is the more successful.

Section 5

Predictive Analysis (Classification)

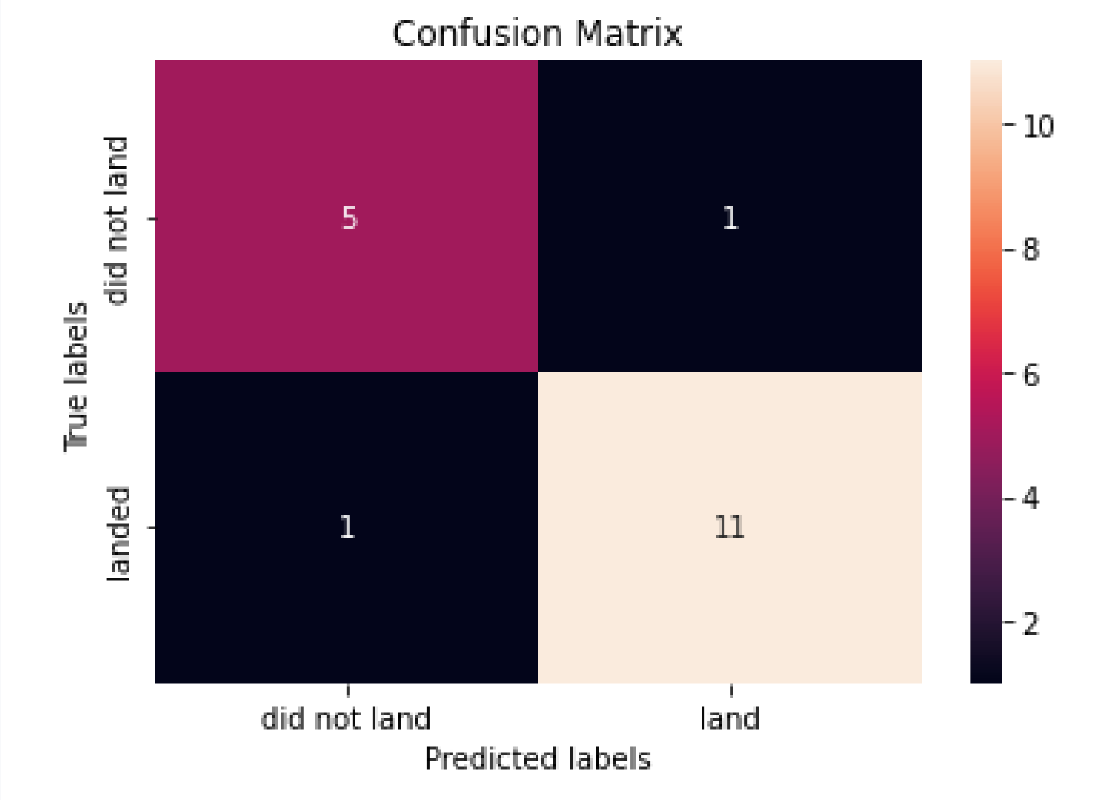
Classification Accuracy

- The four classification models have accuracy over 0.8.
- The model with the highest classification accuracy is the decision tree classifier model with 0.88.



Confusion Matrix of decision tree classifier

- The total number of positive landing outcomes are 12
- The total number of negative landing outcomes are 6
- The corrected predicted values are 16 from 18 total outcomes.
- The model makes only one error to predict positive landing as negative landing and vice versa. Therefore, the accuracy of the model is almost 89%



Conclusions

The **alternate company** has good possibilities to compete in the market against SpaceX since in the success rate of recover the first stage is positively increasing. In 2020 was 80%.

The following recommendation can be taken account in order to increase the opportunities:

- Offer the service for payload mass in the range from 2000kg to 6000kg since they are high demanded and there are a good success rate of recovering the first stage.
- Use booster versions of the series F9 B4 B1XX for medium payloads, and the orbit SSO.
- Increase the flight numbers, for more than 40 it is more likely to recover the first stage.
- Landing preferably in the KSC LC-39A site, since has the highest number of success launches.
- Use the decision tree classifier model to predict the launches outcomes since the accuracy of the model is around 89%.

Thank you!