Data Science Capstone project

Sina Baghdadi

August ,2021

SpaceX Falcon9 first stage Landing Prediction

Overview

- Collecting data with REST API and web scraping
- Data wrangling with python
- Data visualization in Python
- Exploratory data analysis
- Build geographic maps with Folium
- Build a Dashboard with Plotly Dash
- Model Building and predictive analysis
- Results

Introduction

SpaceX designs, manufactures and launches the world's most advanced rockets and spacecraft. 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. In each launch we have two stages, the first stage does most of the work and is quite large and expensive than the second stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. In this project, I will predict if the Falcon 9 first stage will land successfully.

Methodology

Data collection

In this project, I used the SpaceX REST API to collect data .This API give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

Also to collect Falcon 9 historical launch records, I used the BeautifulSoup library of Python for web scraping from the Wikipedia page.

Data collection – SpaceX API

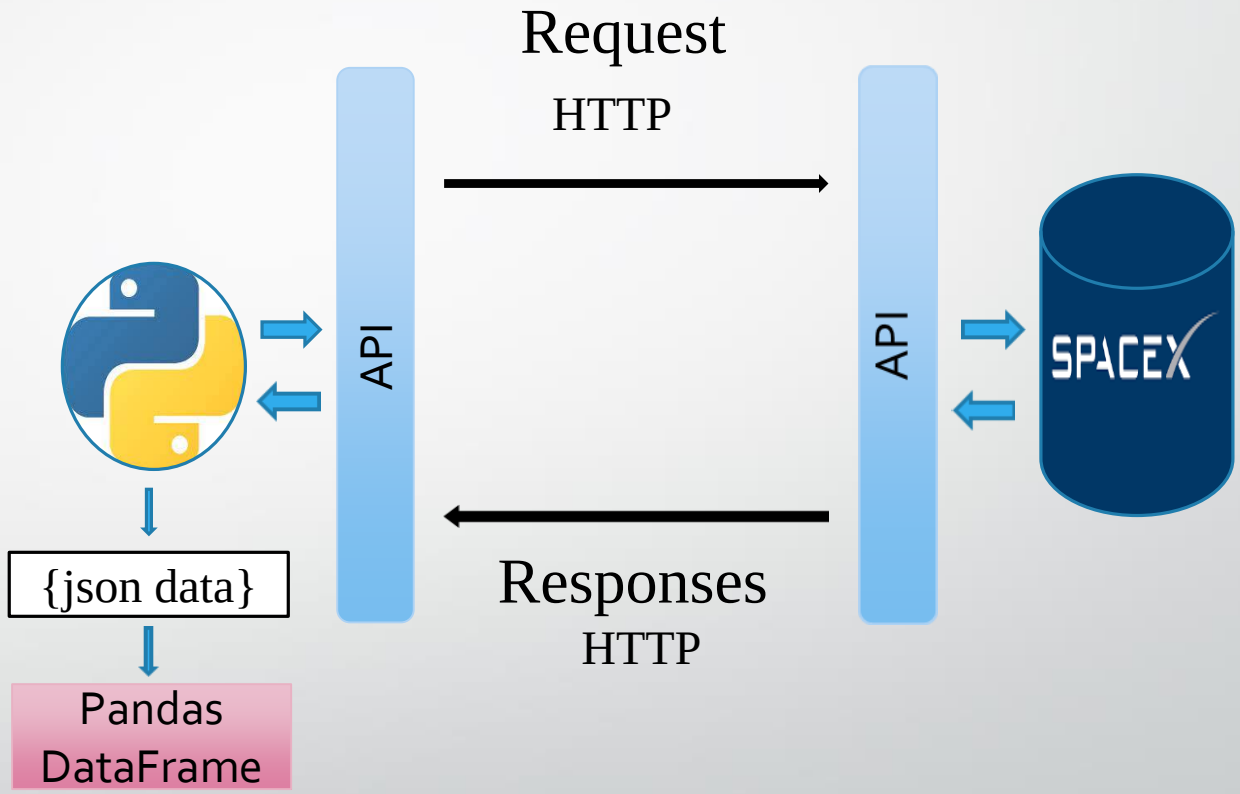
SpaceX API :

https://api.spacexdata.com/v4/launches/past

- Request to the SpaceX API
- Clean the requested data

GitHub URL of:

Code link



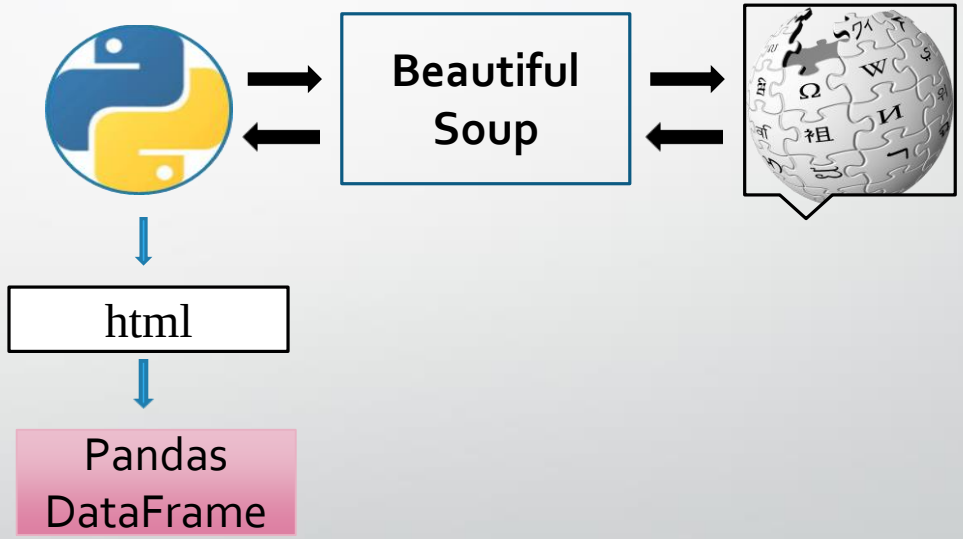
Data collection – Web scraping

Scraping Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

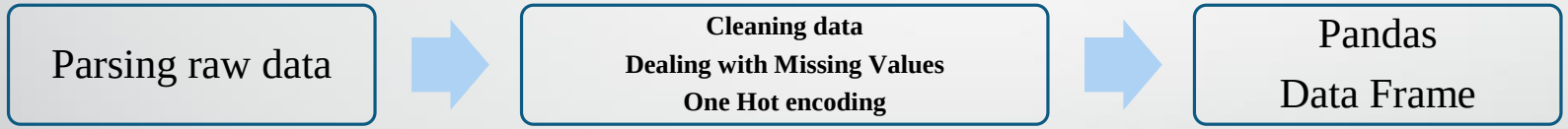
GitHub URL :

Code link



Data wrangling

After collecting raw data, we need to parsing the html table and json data to create the data frame for Falcon 9 launch records.



GitHub URL of :

Code link

EDA with data visualization

In this stage, for understanding how would variables can affect the launch outcome we need to plot charts and check for any existing patterns and relationships between variables

- Scatter plot for Flight Number vs. Payload Mass and overlay the outcome of the launch
- Scatter plot for Flight Number vs. Launch Site and overlay the outcome of the launch
- Visualize the relationship between Payload and Launch Site
- Bar plot for relationship between success rate of each orbit type
- Visualize the relationship between Flight Number and Orbit type
- Scatter plot for Payload and Orbit type
- Plotting the launch success yearly trend

GitHub URL of :

Code link

EDA with SQL

Explore the database to find the answers to these queries:

- Names of the unique launch sites in the space mission
- List of the 5 records where launch sites begin with the string 'CCA’
- The 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 outcome in ground pad
- List of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List of the total number of successful and failure mission outcomes
- List of the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL:
Code link

Build an interactive map with Folium

The launch success rate may depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

So we find the launch sites on the map and mark them, also we need to mark the success/failed launches for each site and calculate the distances between a launch site to its proximities.

We add this objects to Folium map :

- Launch site
- Mark of success/failed launches for each site
- Marker clusters for the same coordinates of launch records.
- Calculate the distances between a launch site to its proximities and mark them as polyline objects

GitHub URL :

Code link

Build a Dashboard with Plotly Dash

We will use Python Plotly Dash package for building a interactive dashboard application for explore and manipulate data in an interactive and real-time way.

With interactive visual analytics, we could find visual patterns faster and more effectively.

- Interactive pie chart for finding which site has the highest success rate
- Interactive scatter chart for payload mass vs outcome of the launch overlay the booster version

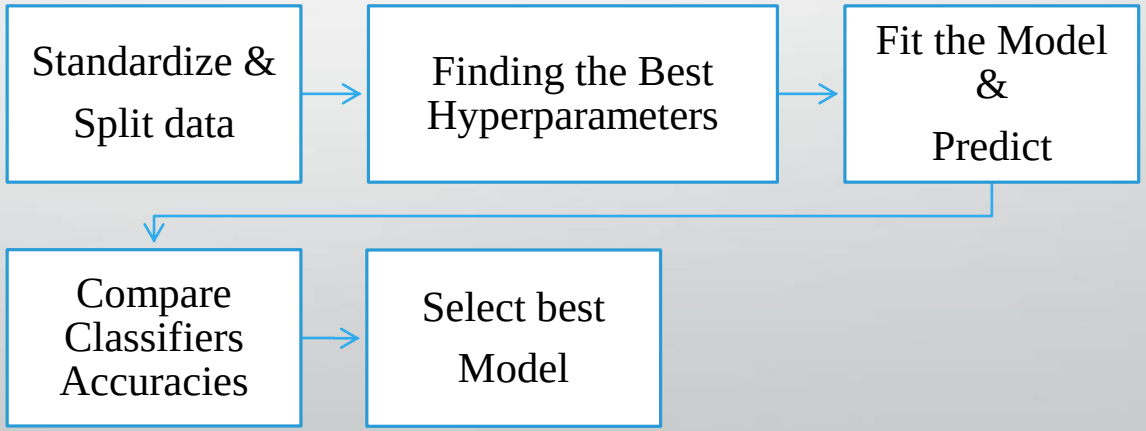
GitHub URL :
Code link

Predictive analysis (Classification)

After EDA, we select the important features that will be used in success prediction in the module. After that we need to Standardize the data and split it into training and testing data, also training data is divided into validation data, a second set is used for training data; then the models are trained and hyperparameters are selected using the function Grid SearchCV.

We will use these machine learning classifiers to find the best classification method for deploying the model and predictive analysis :

- Logistic Regression
- Support Vector Machine
- Decision Tree
- KNN



GitHub link :

Code link

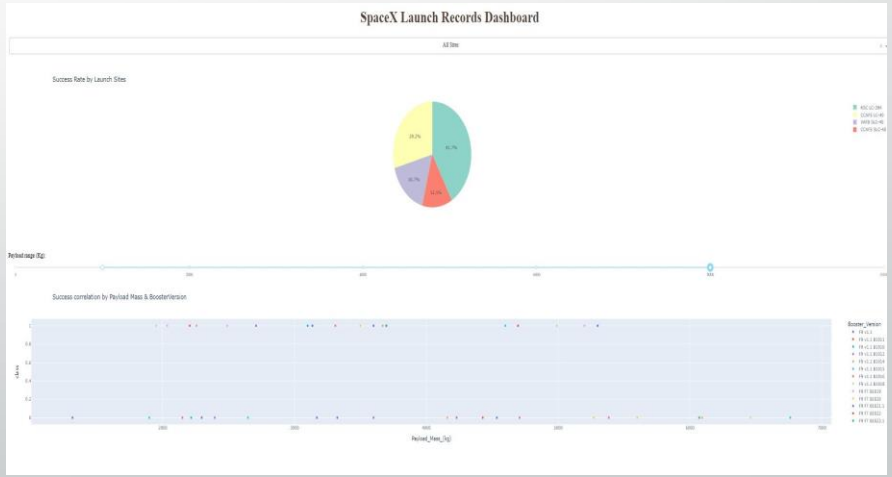
Results

After EDA, I found :

- different launch sites have different success rates
- As the number of flights increases, the success rate increases, especially for the CCAFS SLC 40 site
- As the payload mass increases, the success rate increases, especially for the CCAFS SLC 40 site
- GTO orbits have The least success rate
- Heavy payloads have a negative influence on GTO orbits and positive on LEO and ISS orbits.
- In general, the trend line of success rate, except for 2018, has been upward from 2013 to 2020

Best classifier :

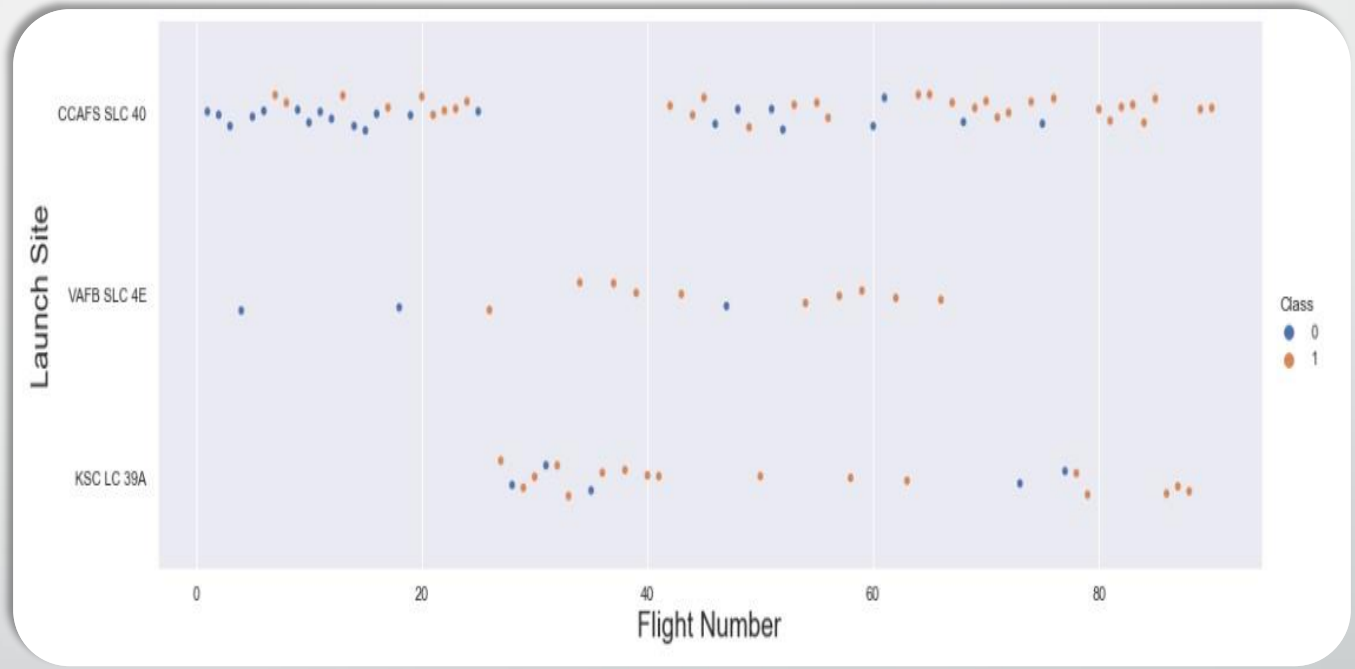
After finding the best hyperparameters for all of the algorithms and fit them on data, I found Decision Tree has the best performance and accuracy with the lowest FP errors compared to the other classifiers.



EDA with Visualization

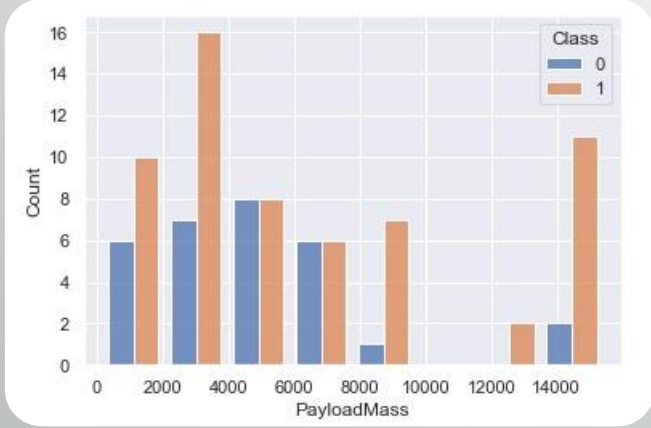
Flight Number vs. Launch Site

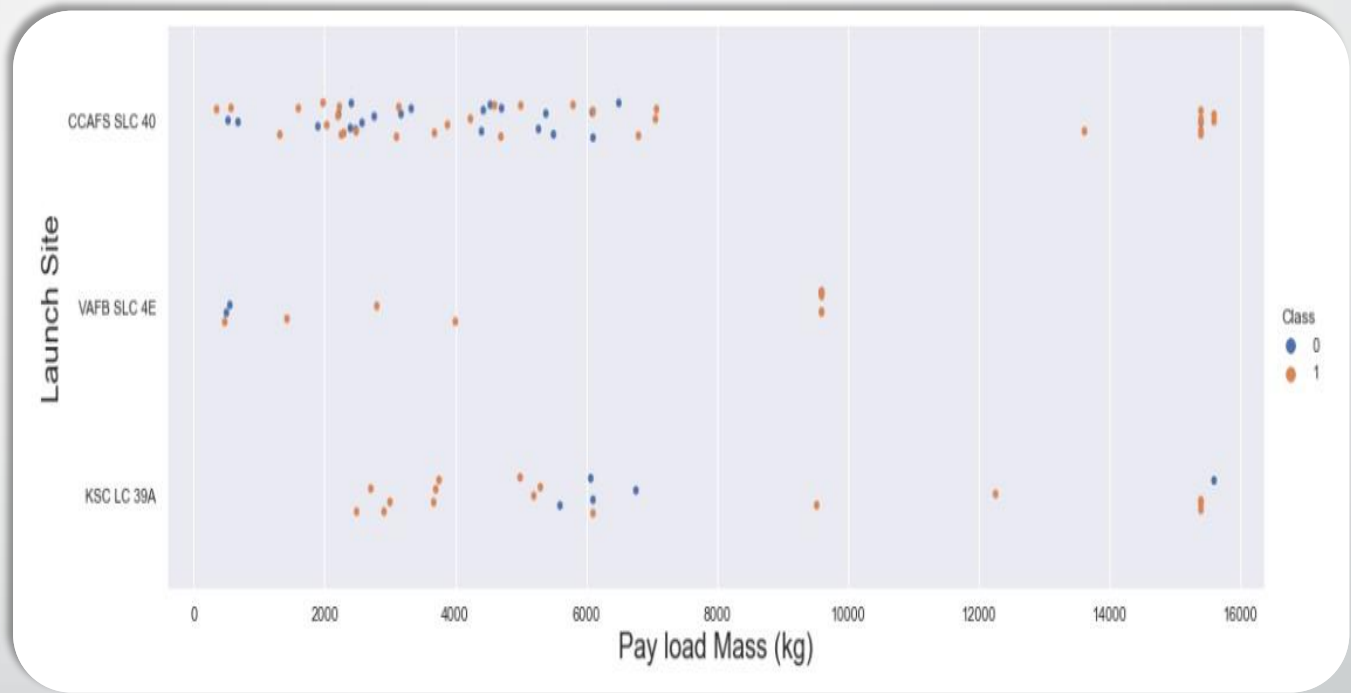
- Different launch sites have different success rates
- As the number of flights increases, wee see the success rate increases for the CCAFS SLC 40 site



Payload vs. Launch Site

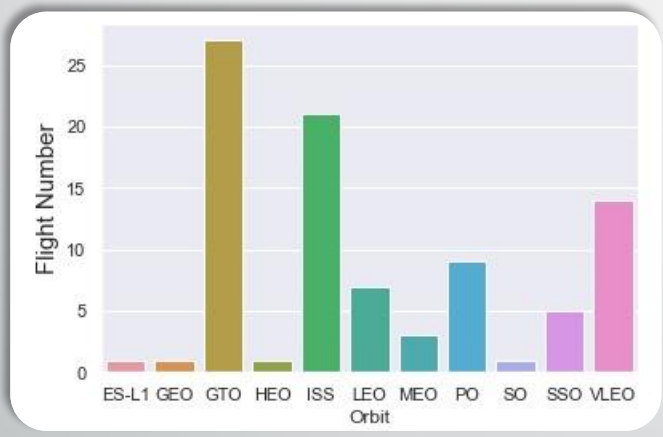
- As the payload mass increases, the success rate increases, for the CCAFS SLC 40 site
- Highest failure rate have occurred between 2,000 and 6,000 kg

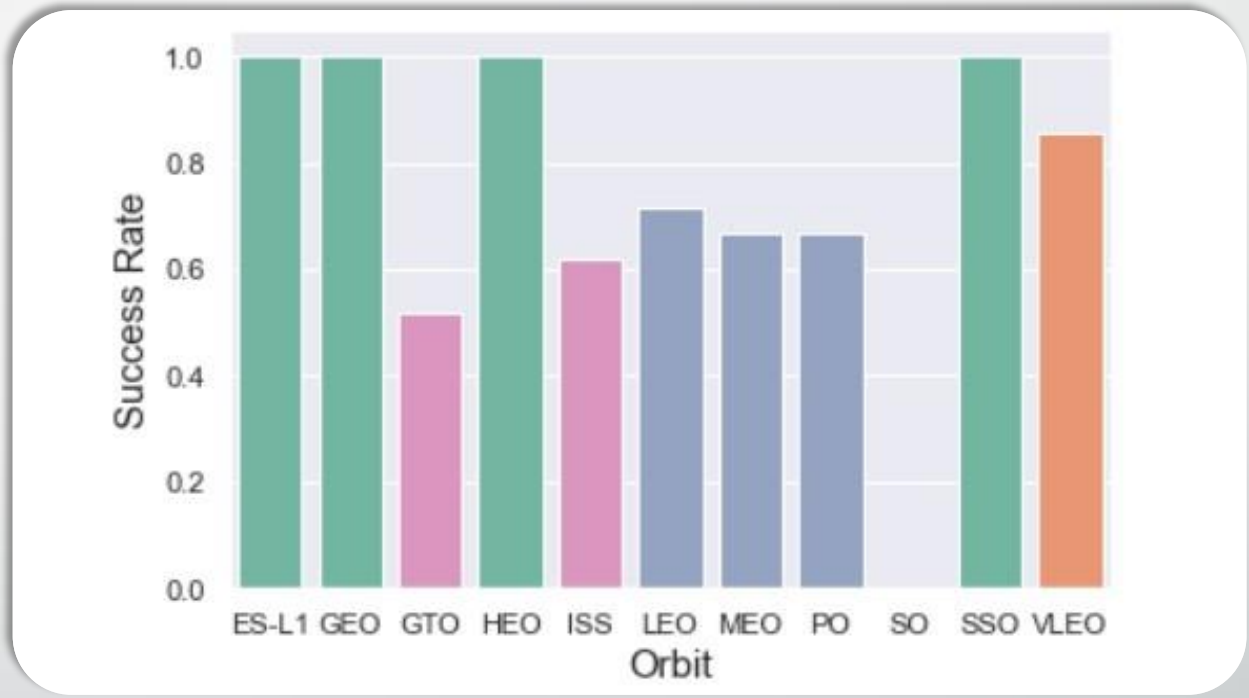




Success rate vs. Orbit type

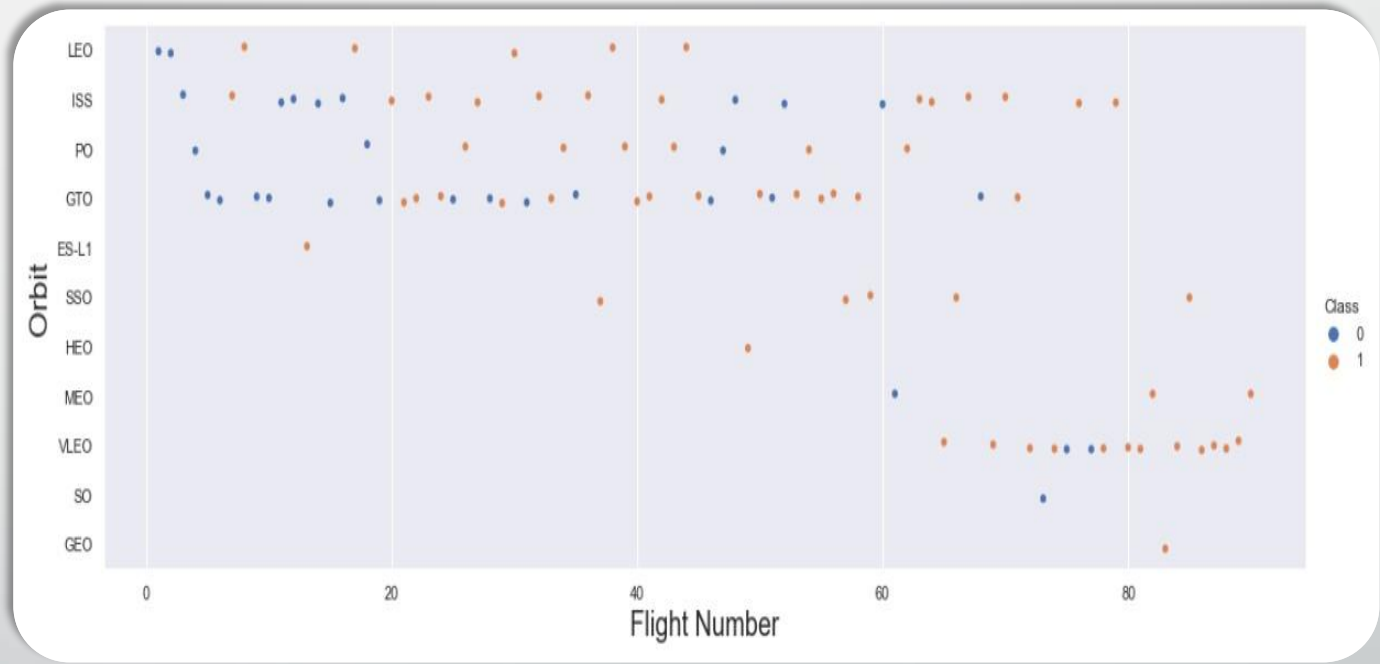
- GTO orbits have The highest flight number and lowest success rate





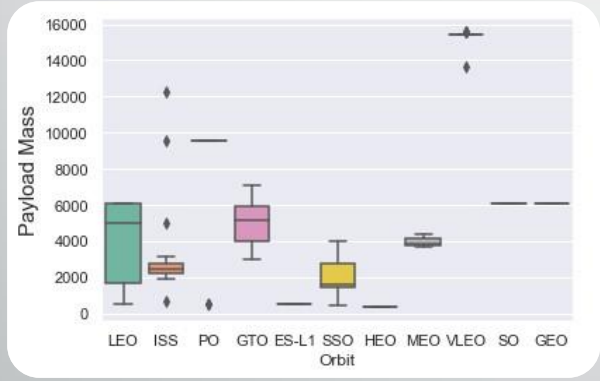
Flight Number vs. Orbit type

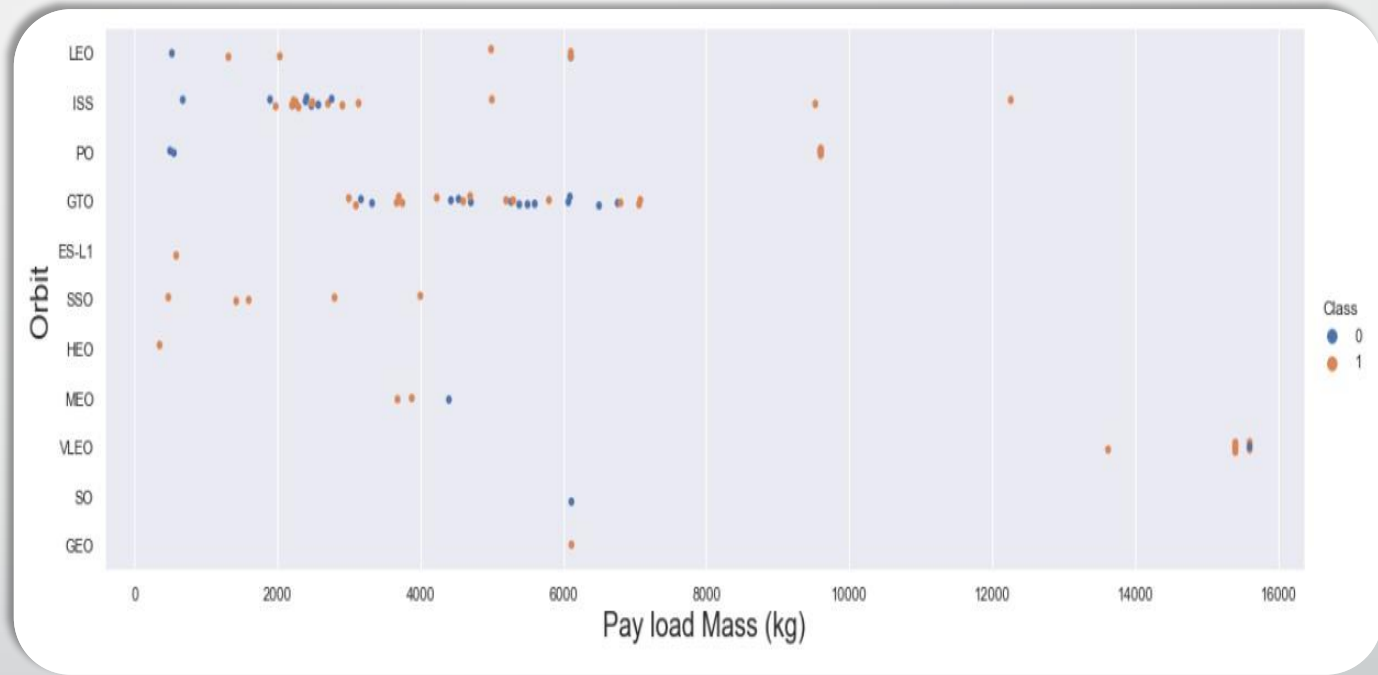
As the number of flights increases, the success rates increases for LEO, MEO and VLEO orbits



Payload vs. Orbit type

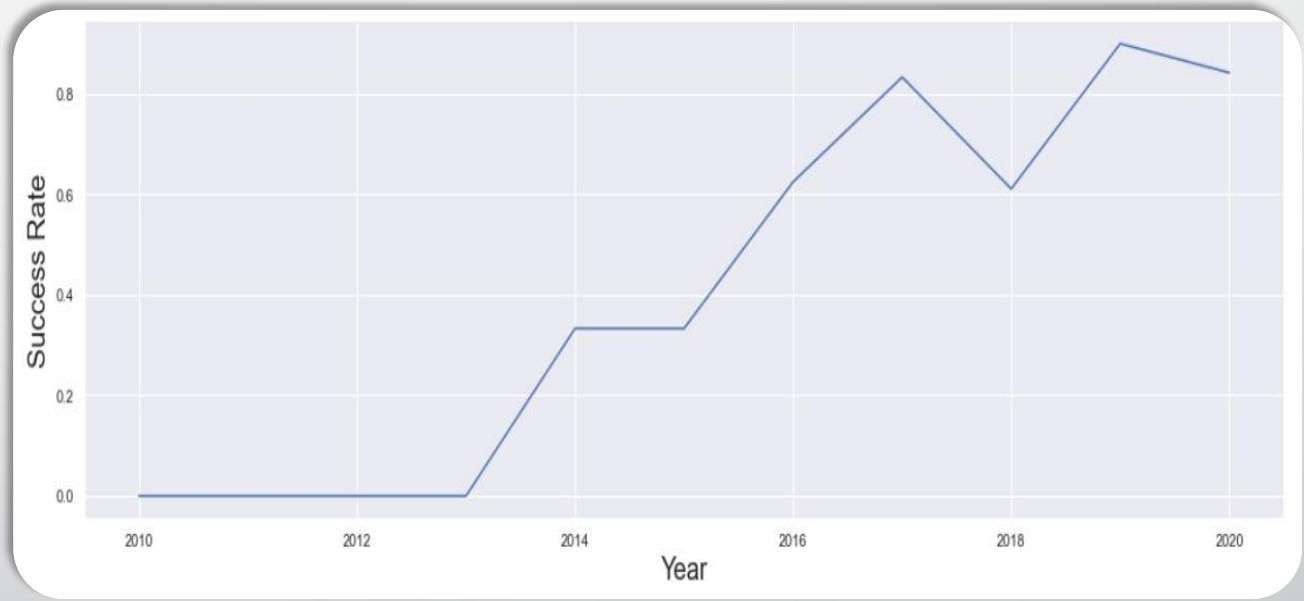
Heavy payloads have a negative influence on GTO orbits and positive on LEO and ISS orbits.





Launch success yearly trend

Success rate trend lines, except for 2018, has been upward from 2013 to 2020



EDA with SQL

All launch site names

Find the names of the unique launch sites

```
%sql select distinct(LAUNCH_SITE) from Spacex;
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Launch site names begin with `CCA`

Find all launch sites begin with `CCA`

```
%sql select * from Spacex where LAUNCH_SITE  like 'CCA%' limit 5;
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Calculate the total payload carried by boosters from NASA

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

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.
```

total_payload_mass
45596

Average payload mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_)as average_payload_mass from spacex where booster_version = 'F9 v1.1';
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.
```

average_payload_mass
2928

First successful ground landing date

Find the date when the first successful landing outcome in ground pad

```
%sql select DATE, time__utc_ from spacex where \
(landing__outcome = 'Success (ground pad)' )  order by DATE limit 1;
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

DATE	time__utc_
2015-12-22	01:29:00

Successful drone ship landing with payload between 4000 and 6000

List the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

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

Total number of successful and failure mission outcomes

Calculate the total number of successful and failure mission outcomes

```
%sql select count(*) as total_number from spacex where \
(mission_outcome like 'Success%') or (mission_outcome like 'Failure%');
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

total_number
101

Boosters carried maximum payload

List the names of the booster which have carried the maximum payload mass

```
%sql select distinct(booster_version) from spacex \
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex);
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
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

2015 launch records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%sql select monthname(DATE) as Month, booster_version, launch_site, landing__outcome \
from spacex where (landing__outcome = 'Failure (drone ship)' )and year(DATE)=2015  ;
```

```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain
Done.
```

MONTH	booster_version	launch_site	landing__outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank success count between 2010-06-04 and 2017-03-20

Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

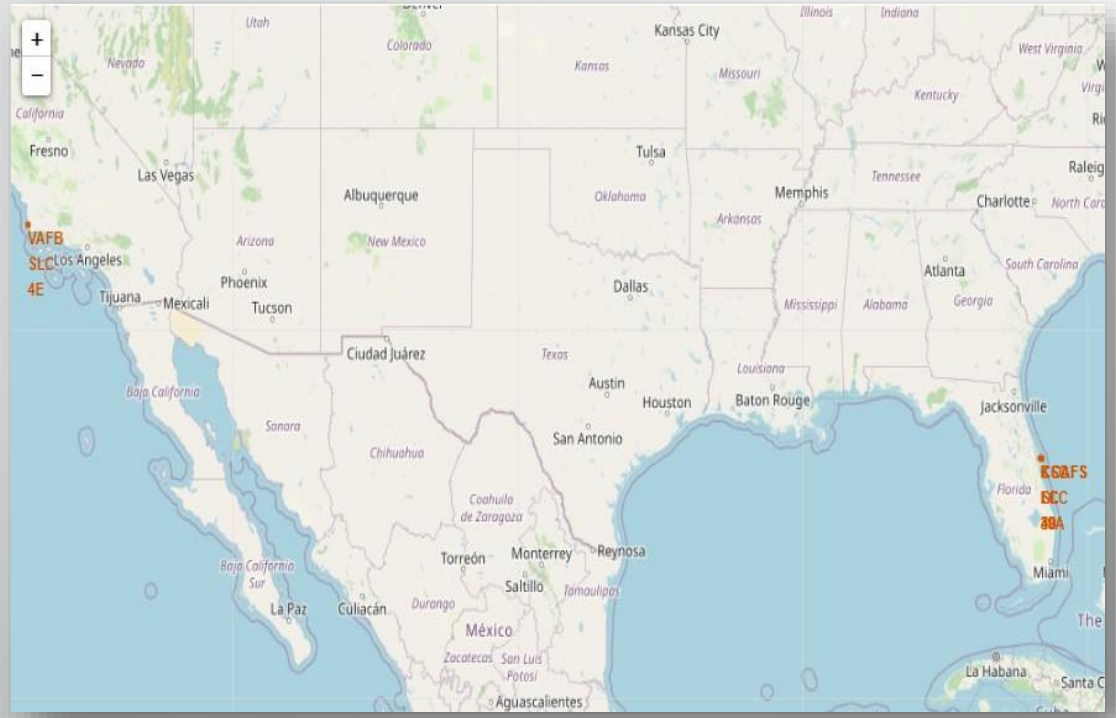
```
%sql select DATE as successful_date from spacex where (landing__outcome like '%Success%' )\
and DATE in (select DATE from spacex where DATE between '2010-06-04' and '2017-03-20') order by DATE desc;
```

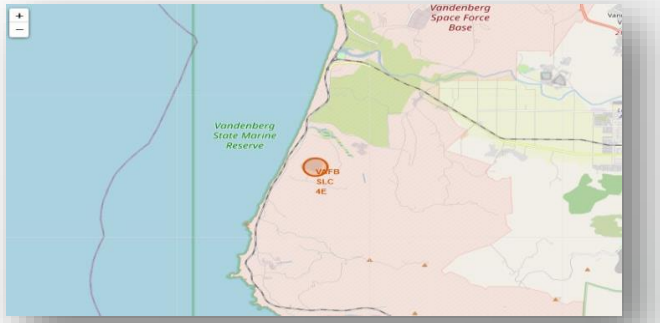
```
 * ibm_db_sa://gvr91967:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.
```

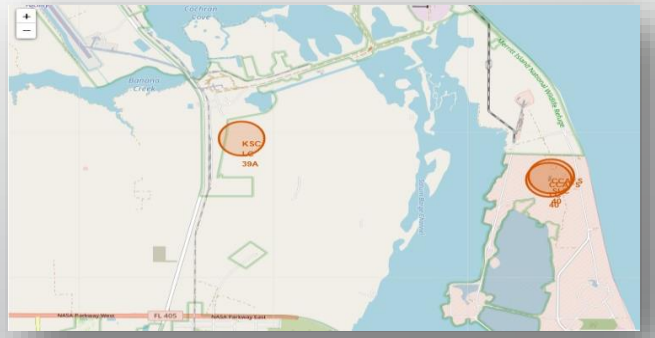
successful_date
2017-02-19
2017-01-14
2016-08-14
2016-07-18
2016-05-27
2016-05-06
2016-04-08
2015-12-22

Interactive map with Folium

Launch Sites Locations

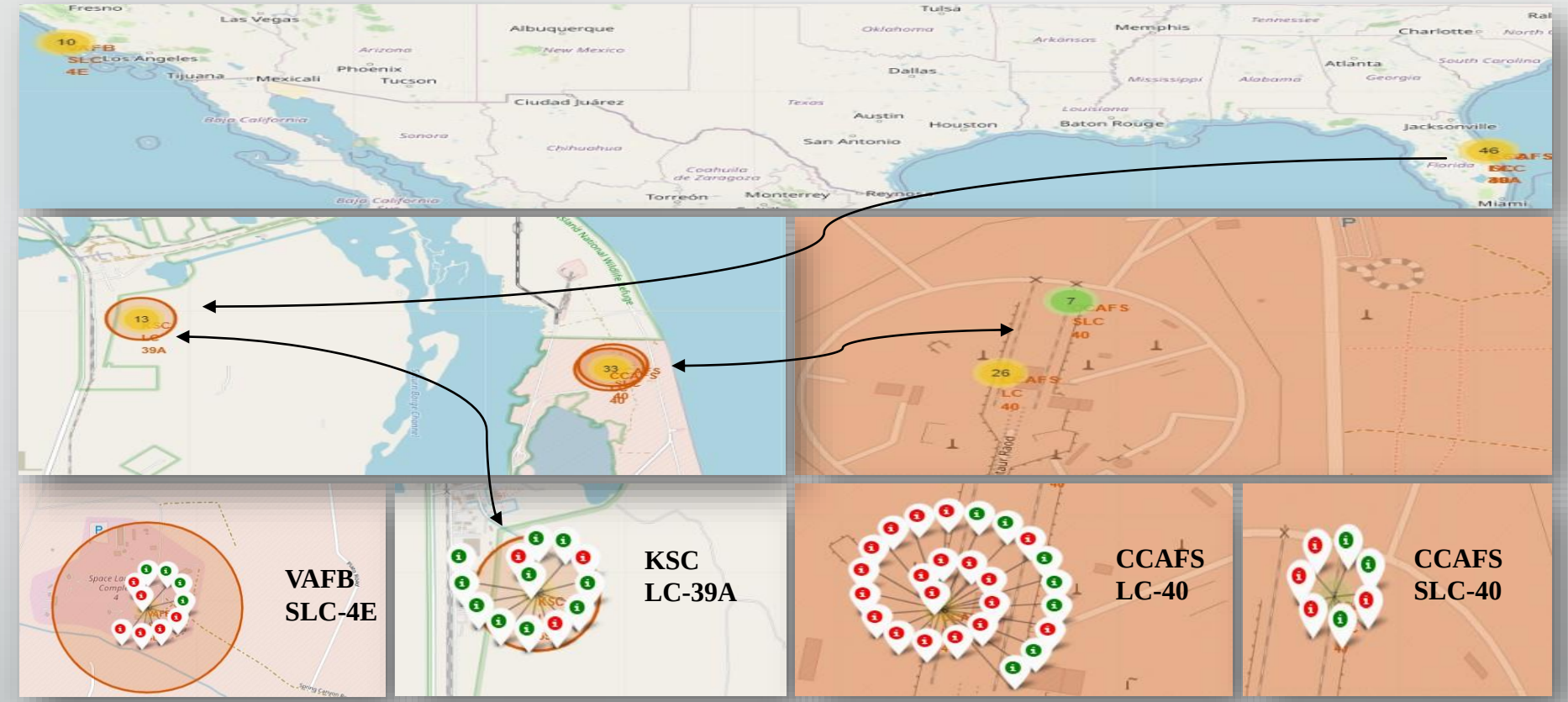






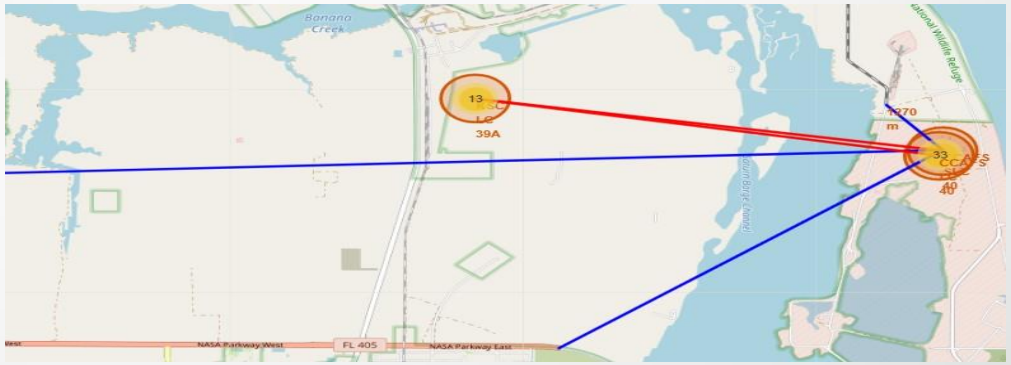
We see all of launch sites are very close proximity to the coast

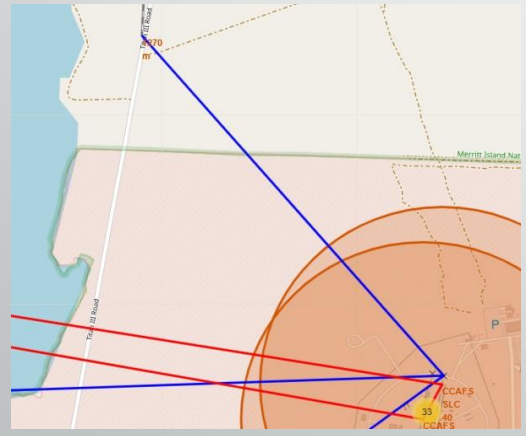
Geo Map of Launch Outcomes



- Success Outcomes : Green mark
- Failed Outcomes :Red mark
- KSC LC-39A launch site has the highest success rate

Distances between a launch site to its proximities





For CCAFS SLC-40
Distance from
Orlando : 78.4 km
Railway : 1.27 km

```
#distance_railway = calculate_distance(lat1, lon1, lat2, lon2)
x = [28.57205, -80.58528]
CCAFS_rw = [28.56341, -80.57678]

distance_railway = calculate_distance(CCAFS_rw[0], CCAFS_rw[1], x[0], x[1])
distance_railway
```

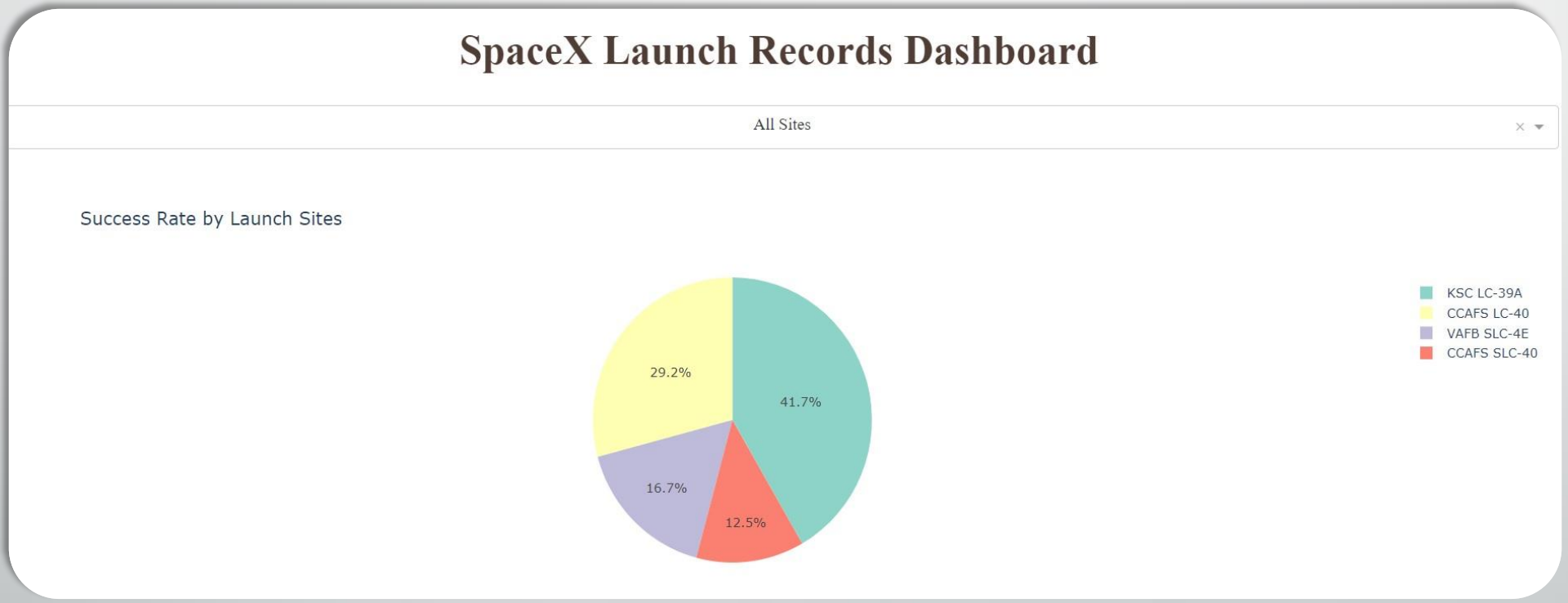
```
1.2700576811493218 km
```

```
# Create a marker with distance to a closest city, coastline, highway, etc.
# Draw a line between the marker to the launch site
Orlando = [28.5384, -81.3789]
N_HW = [28.52561, -80.63435]
LC = [28.562302, -80.577356]
SLC = [28.563197, -80.576820]
KSC = [28.573255, -80.646895]

P1 = (CCAFS_rw, Orlando)
P2 = (CCAFS_rw, N_HW)
P3 = (LC, SLC)
P4 = (LC, KSC)
P5 = (SLC, KSC)

dist_Orlando = calculate_distance(CCAFS_rw[0], CCAFS_rw[1], Orlando[0], Orlando[1])
dist_Orlando
```

```
78.41913203972042 km
```

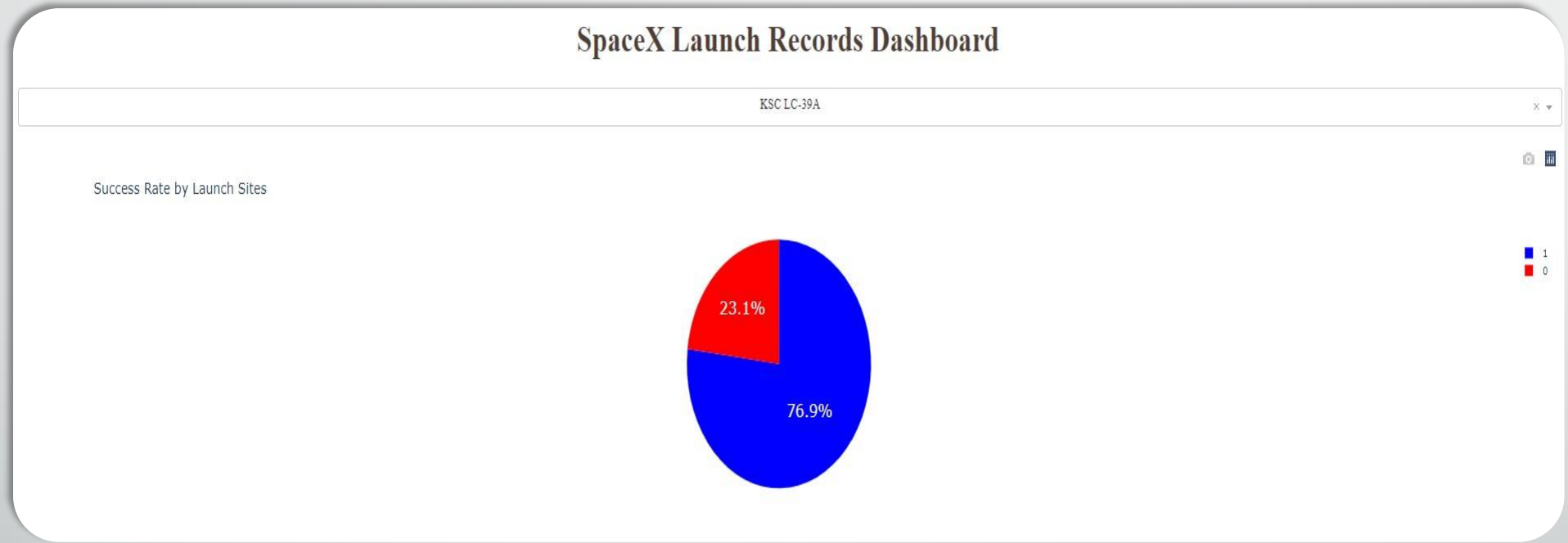
Build a Dashboard with Plotly Dash

Piechart of success rate by launch Sites



CCAFS SLC-40 launch site has the lowest success rate

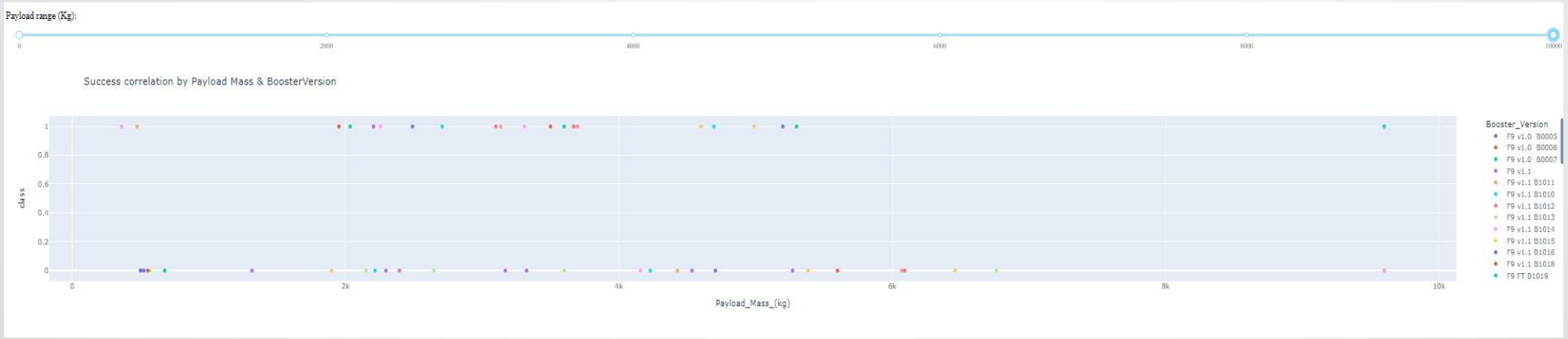
Launch site with the highest success rate

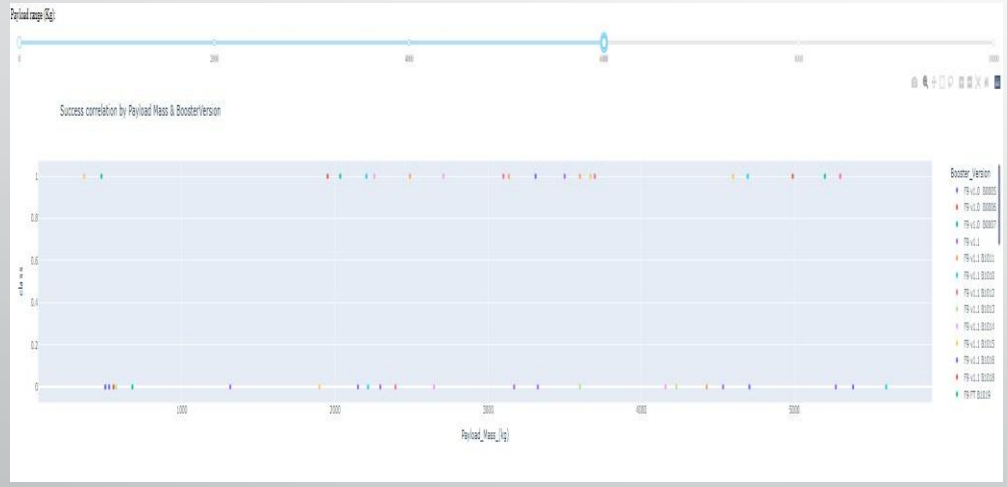


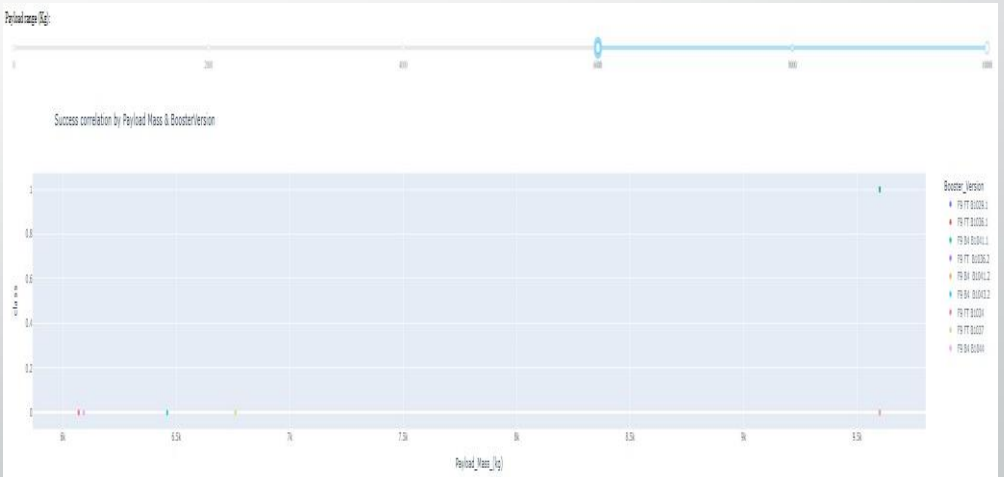
KSC LC-39A launch site has the highest success rate

Interactive scatter plot

Payload vs. Launch Outcome





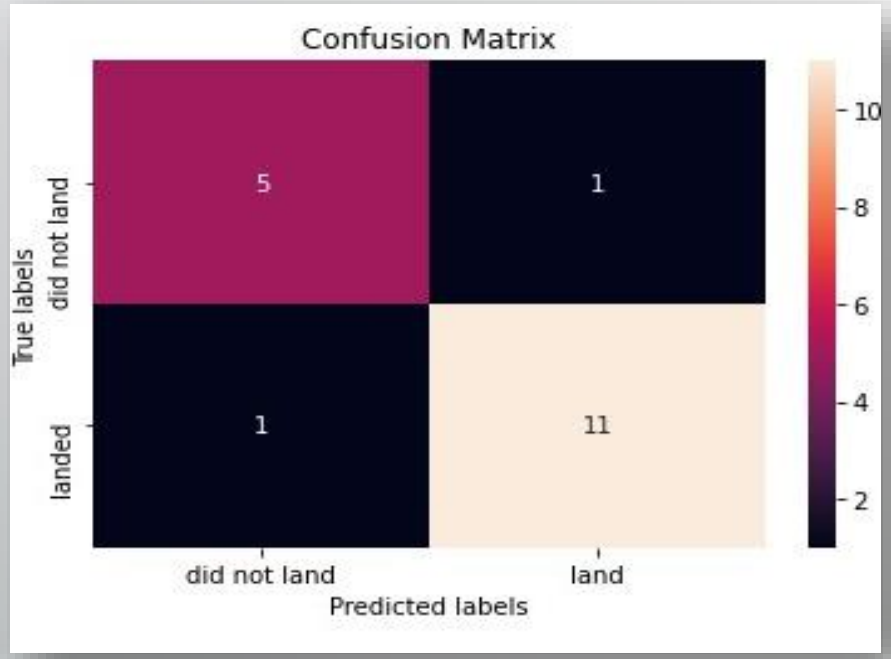


Most of successful launch have Payloads between 2,000 and 6,000

For Payloads between 6,000 and 10,000, only one launch has been successful

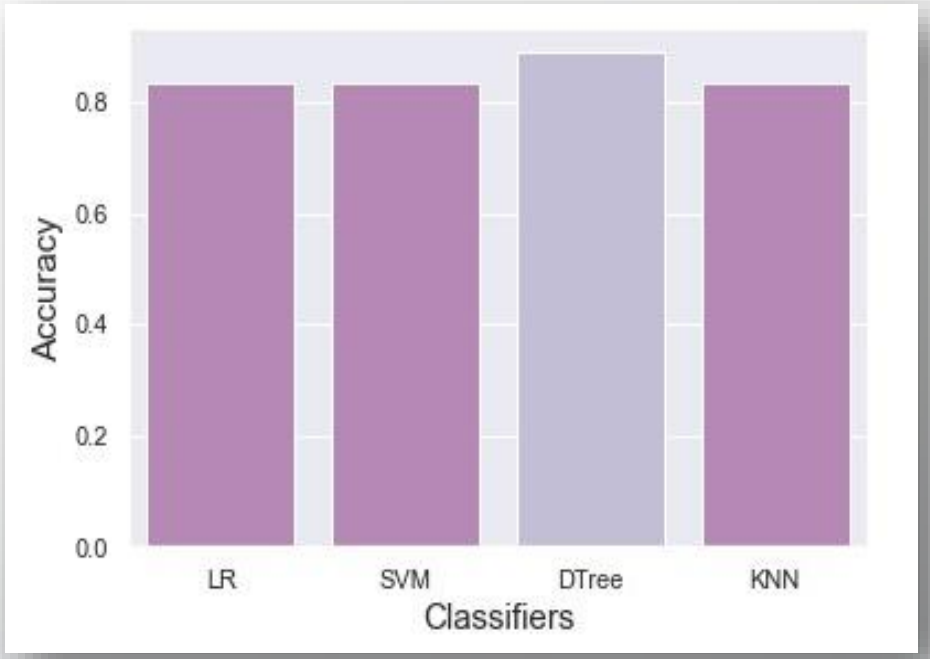
Predictive analysis (Classification)

Confusion Matrix



Decision Tree

Classification Accuracy



Decision Tree has the best performance and accuracy with the lowest FP errors compared to the other classifiers.

CONCLUSION

- We can confidently choose the decision tree as the best classifier, but it is better to test with other classifiers such as neural network to be sure.
- Highest failure rate have occurred between 2,000 and 6,000 kg for both launch sites CCAFS SLC 40 and KSC LC-39A , but in VAFB SLC-4E, no launch has been done in this range, which must be further investigated.
- To increase success rate we need to select CCAFS SLC 40 site only for payloads masses between 8000 and 16000 kg.
- The biggest failures occurred in GTO orbits launches outcome with pay load masses in the range 3000 and 7000 kg , so to increase our success rates we have to choose higher payloads masses for the our launches of this circuit.