IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Asma Alkhalaf
29-08-2023

Outline

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis Results
- Interactive Maps and Dashboard screenshots
- Results of Predictive Analysis

Introduction

Project background and context

The project aims to predict whether Falcon 9 first stage will land successfully. SpaceX is a successful company that advertises Falcon 9 rocket on its website with a launch cost of 62 million dollars. Thus, if we are able to determine if the first launch is successful, we will be able to determine the cost. It's an important information for other companies if they wish to compete with SpaceX for a rocket launch.

Problems you want to find answers

- What factors affect the launch of Falcon 9 and whether it will land successfully or not?
- What features determine the success rate of a landing?
- What are the best conditions to consider when trying to achieve a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SaceX REST API
  - Using Web scraping from Wikipedia
- Perform data wrangling
  - Filtering the data and dealing with missing values
  - Using One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected using the following methods

- SpaceX REST API and we obtained the following information
  - Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Wikipedia web scraping and we obtained the following information
  - Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

Requesting Rocket Launch data from SpaceX API

Converting the Response to a JSON File and turning it into a Datafarme

Requesting needed information by applying custom function

Creating a dictionary with our data

Creating a Dataframe

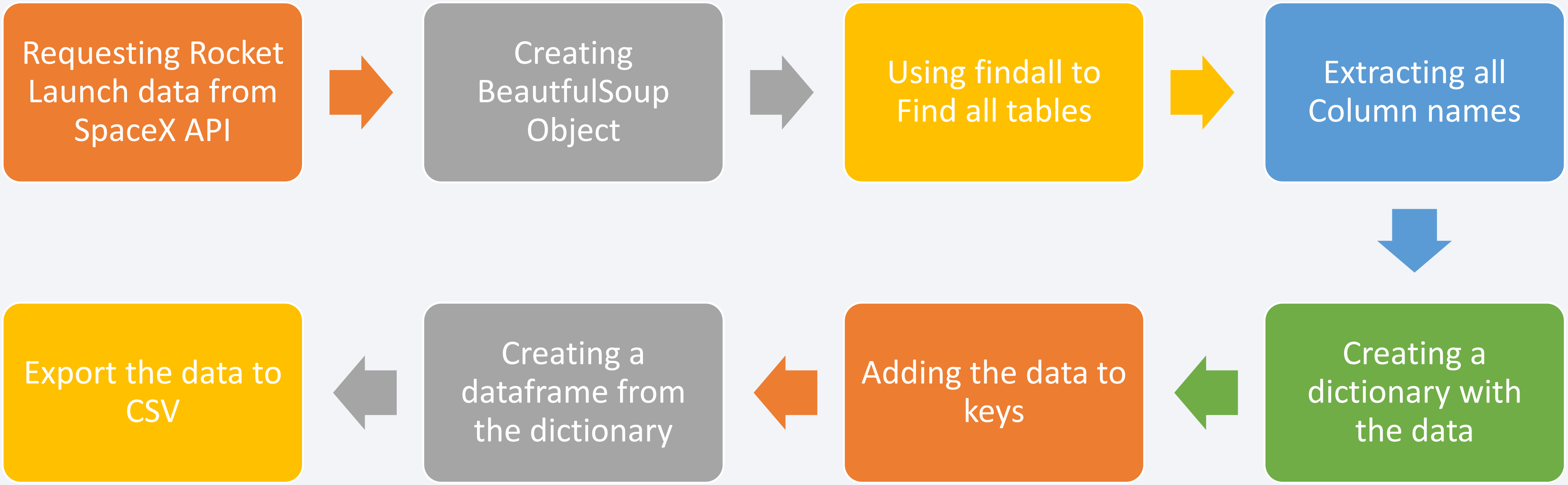
Filtering the dataframe to only include Falcon 9

Dealing with the missing values

Export the data to CSV

Data Collection SpaceX API - GitHub URL

Data Collection - Scraping

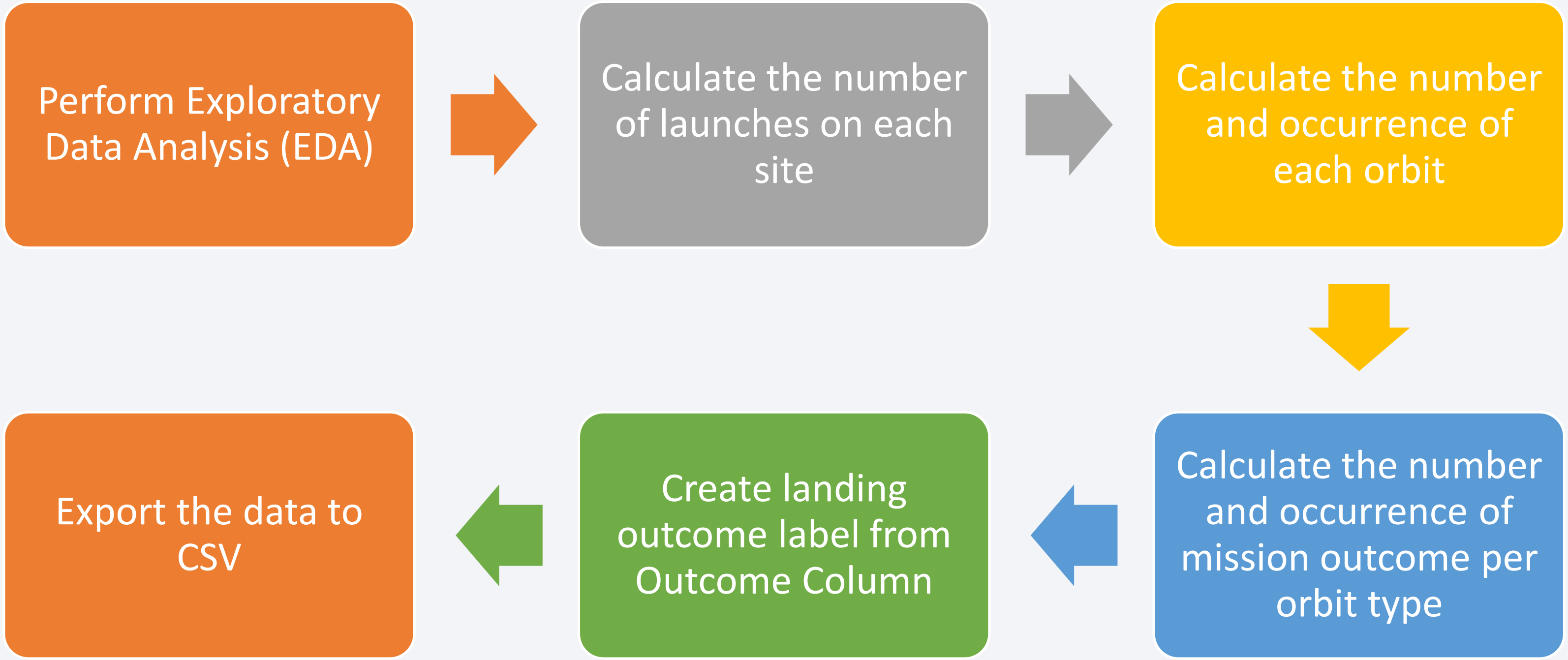


[Data Collection Web Scraping - GitHub URL]

Data Wrangling

In the dataset, there are several cases where the booster did not land successfully.

- True Ocean, True RTLS, True ASDS means the mission has been successful.
- False Ocean, False RTLS, False ASDS means the mission was a failure.



Data Wrangling - GitHub URL

EDA with Data Visualization

The charts that were plotted used are as follows,

1- Scatter Plots

We used this type of chart because it shows the relationships between variables. If a relationship exists, they can be used in a machine learning model. This relationship is called correlation.

2- Bar Charts

Bar charts are used to show the relationship between a numeric and categorical variables.

3- Line Charts

Line charts are used to show trends in data over a period of time.

[EDA with data Visualization - GitHub URL]

EDA with SQL

The SQL queries that were performed are as follows,

- Displaying the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List 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 04-06-2010 and 20-03-2017 in descending order.

[EDA with SQL - GitHub URL]

Build an Interactive Map with Folium

The following were the steps taken to build an interactive map with Folium:

- We first marked all launch sites and added Marker with Circle, Popup Label and lines in order to mark the success or failure of the launches of each site.
- We then assigned features for the outcomes of the launch to class 0 for failure and 1 for success.
- Added coloured Markers of success (**Green**) and failed (**Red**) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.
- We finally calculated the distances between a launch site to its proximities like Railway, highway and closest city.

Launch Site Location - GitHub URL

Build a Dashboard with Plotly Dash

The following were the steps taken to build a dashboard with Plotly:

- We first added a dropdown list to enable Launch Site selection.
- The pie chart was and used to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- We also added a Range slider to select a payload mass in a fixed range.
- We finally chose to add a scatter chart because it shows the relationship between two variables and in this case the correlation between Payload and Launch Success in particular.

Dashboard with Plotly - GitHub URL

Predictive Analysis (Classification)

The classification model was built in the following steps,

Data Preparation	Load the dataset	Normalize the data	Split data into training and test sets
Model Preparation	Selection of machine learning Algorithms	Set parameters for each algorithm GirdSearchCV	Training GridSearch Model with training dataset
Model Evaluation	Get best hyperparameters for each type of model	Compute accuracy for each model with test dataset	Plot Confusion Matrix
Model Comparison	Comparison of models according to their levels of accuracy	The model with the best accuracy is the one chosen	

[Predictive Analysis - GitHub URL]

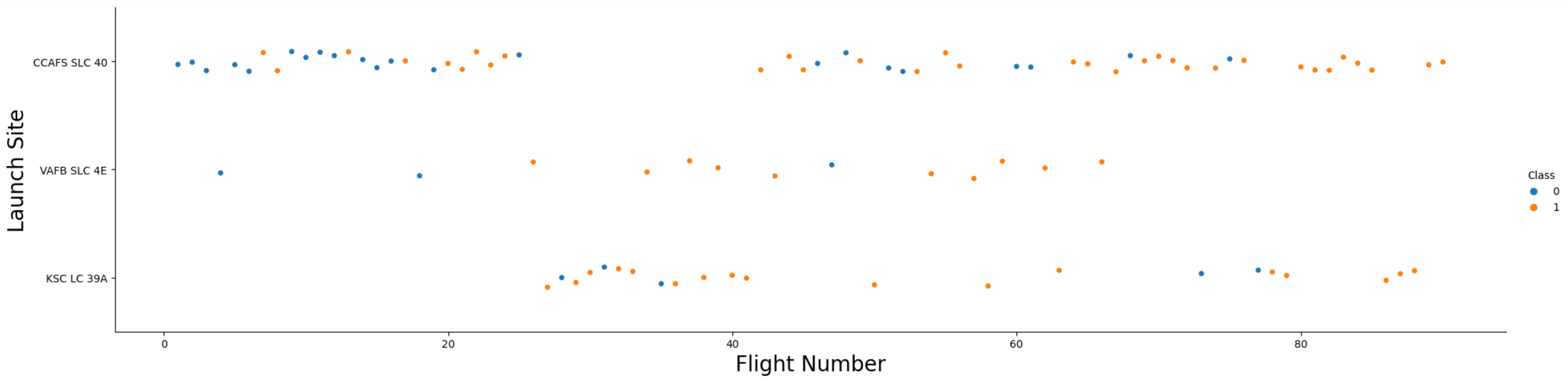
Results

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

Section 2

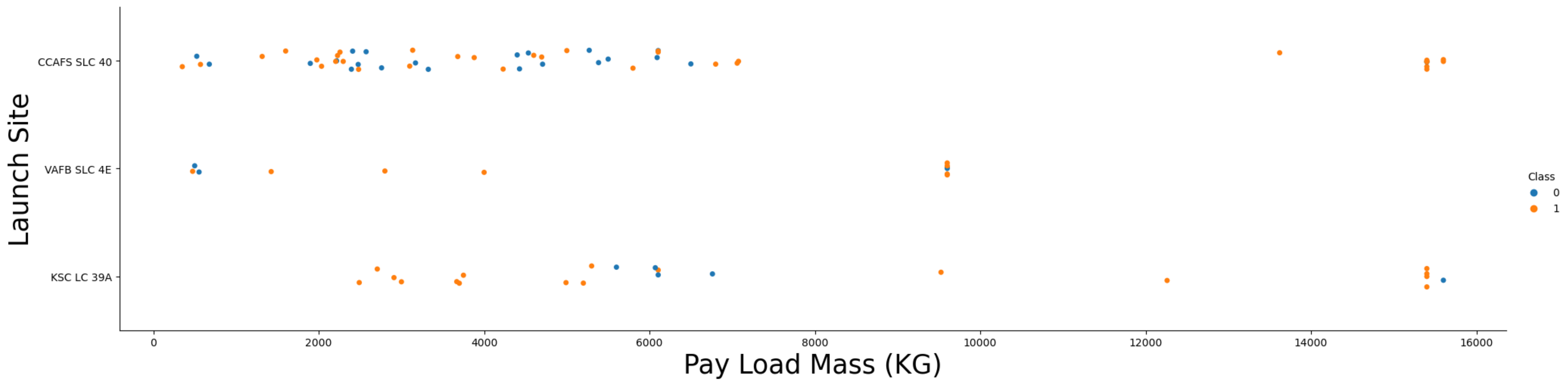
Insights drawn from EDA

Flight Number vs. Launch Site



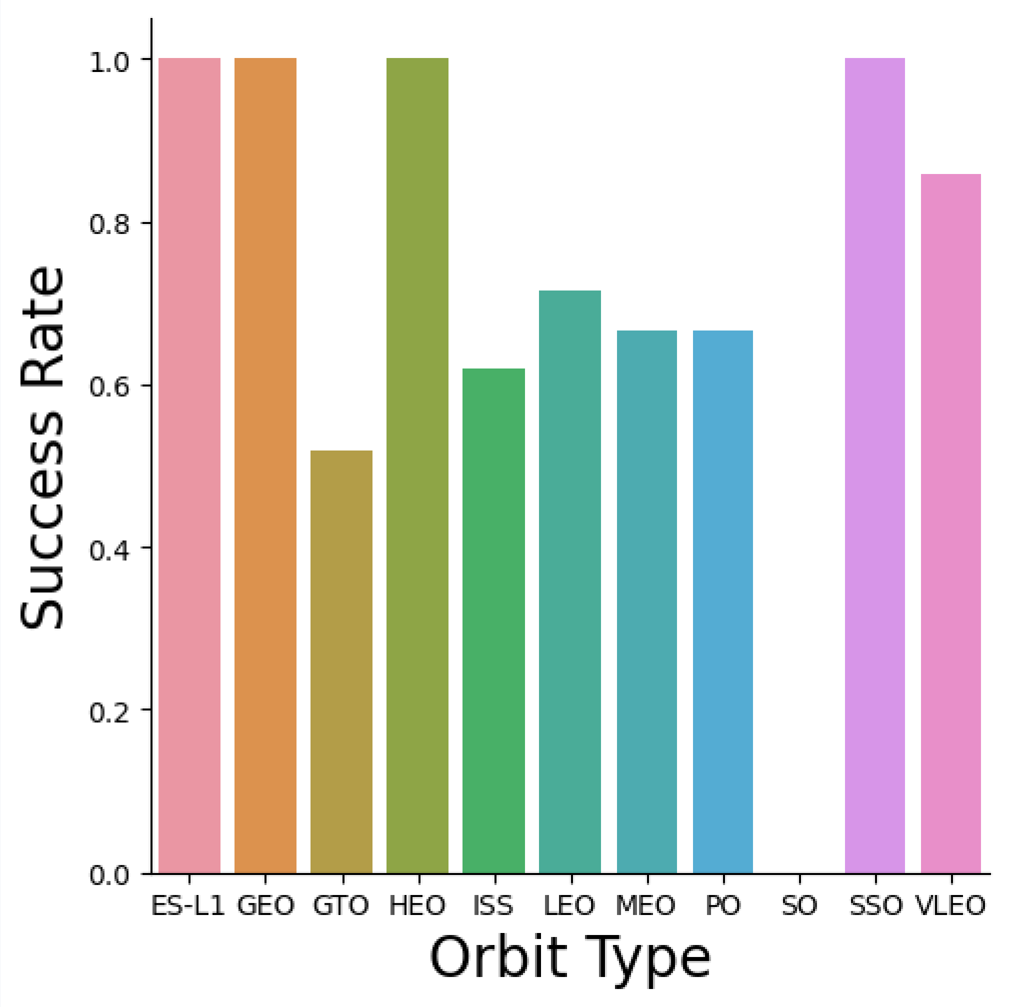
- From the plot, we can observe that the success rate for each flight increases with each launch

Payload vs. Launch Site



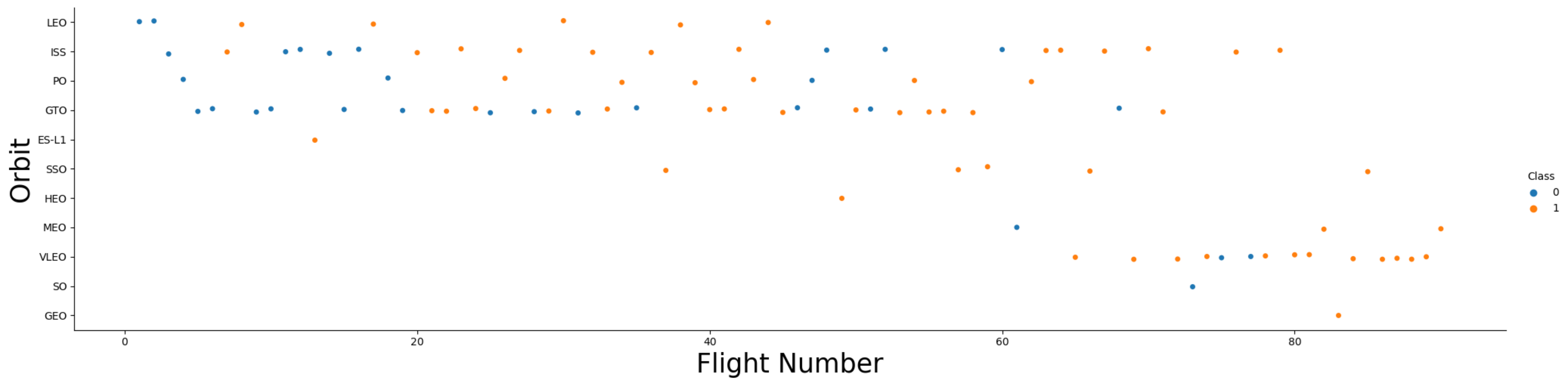
- For every launch site the greater the payload mass, the higher the success rate.
- Payloads over 9,000kg have a high success rate;

Success Rate vs. Orbit Type



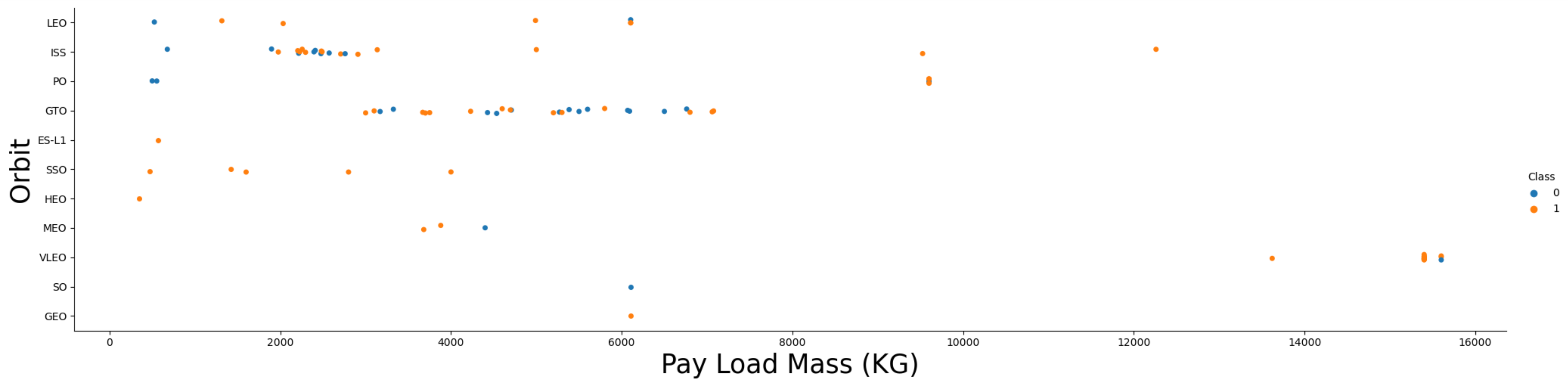
- Orbits with 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO
  - VLEO

Flight Number vs. Orbit Type



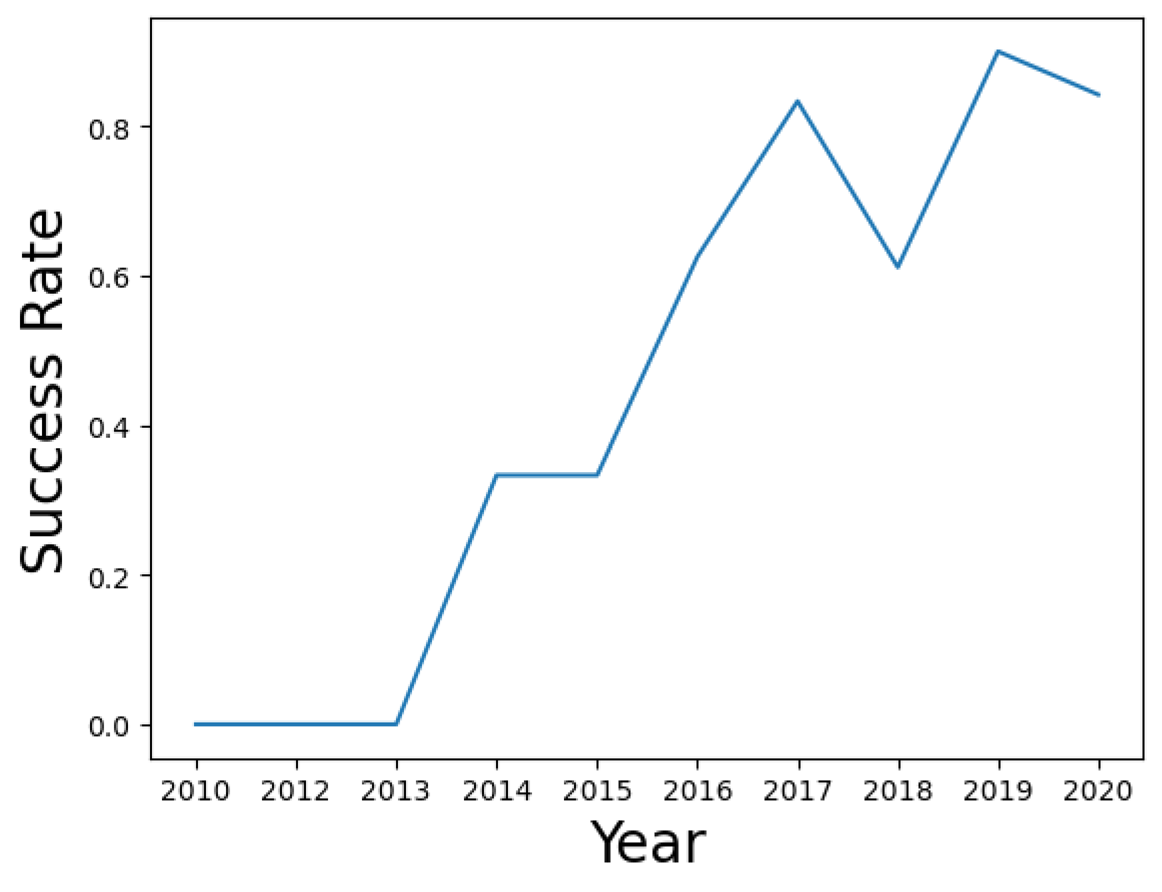
in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

The use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE.

Launch Site Names 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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

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

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

total_payload_mass
45596

Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) as AVG_PAYLOAD FROM SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

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

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_LAUNCH FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_SUCCESSFUL_LAUNCH
2015-12-22

Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BET
```

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

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

Listing 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 SPACEXTBL group by mission_outcome;
```

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

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

Displaying the total number of successful and failure mission outcome

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

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

2015 Launch Records

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND \
LANDING_OUTCOME = 'Failure (drone ship)';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015

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

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDIN
```

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

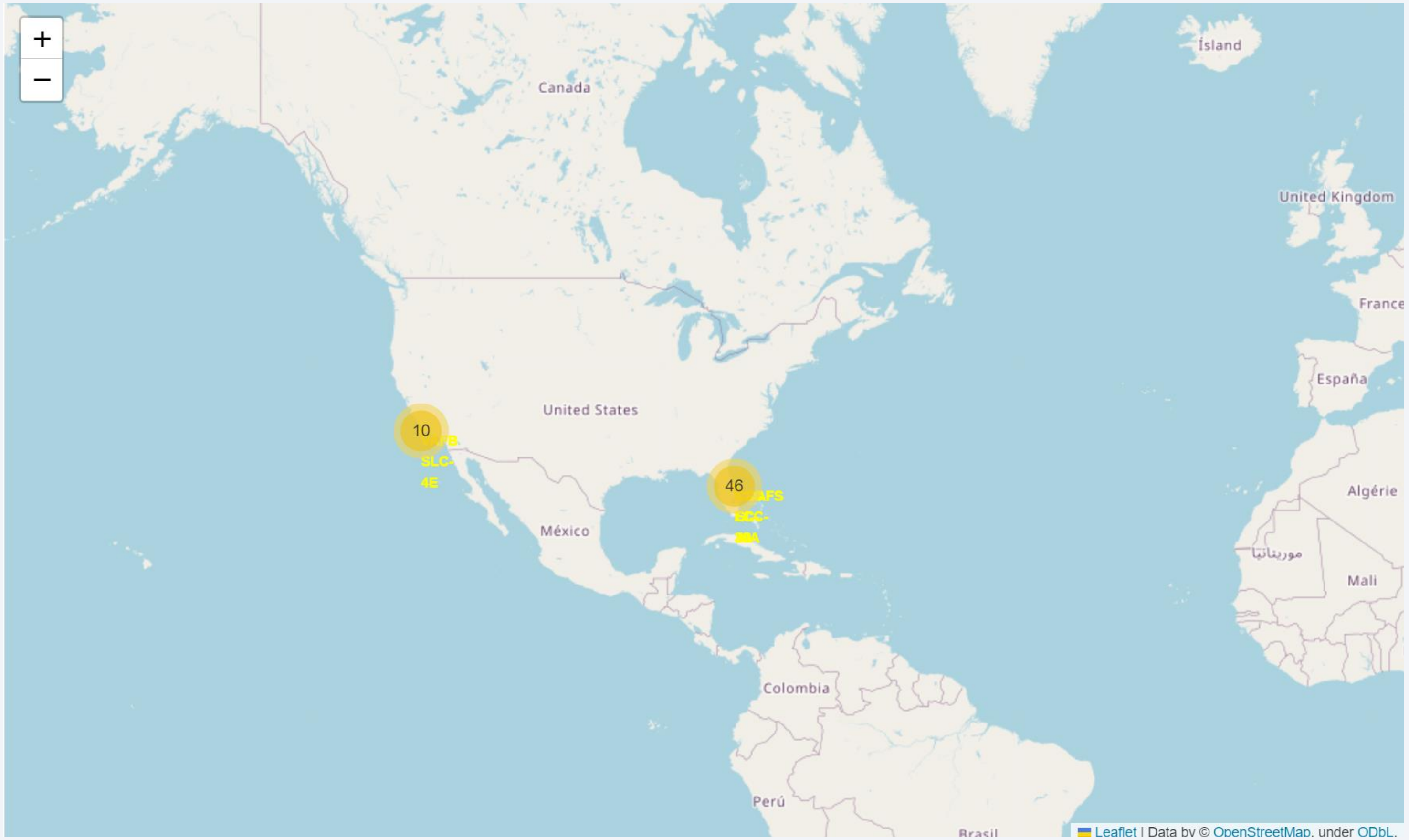
Landing_Outcome	QTY
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

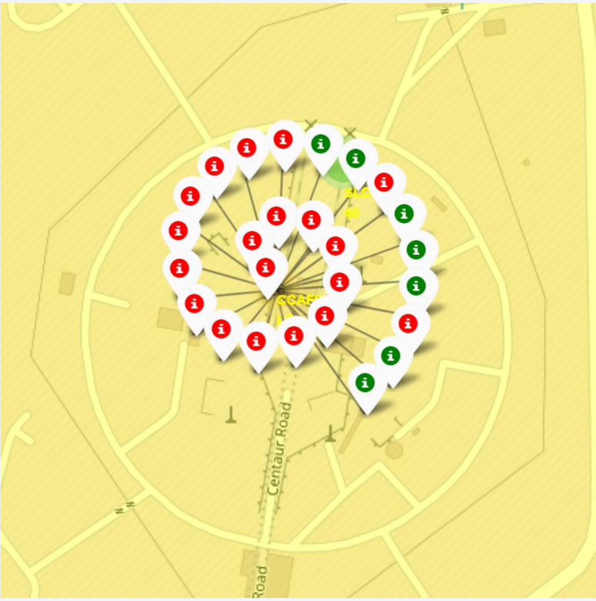
Launch Sites Proximities Analysis

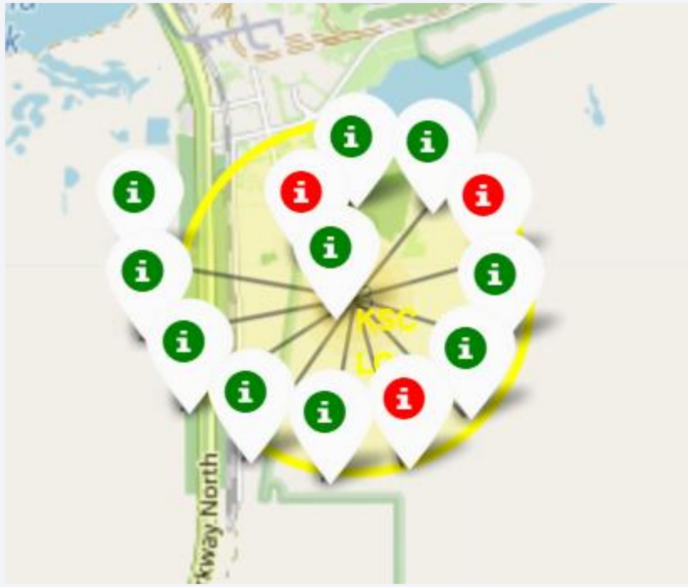
All launch sites' location markers on a global map



We can see that Space X launch sites are located on the coast of the United States

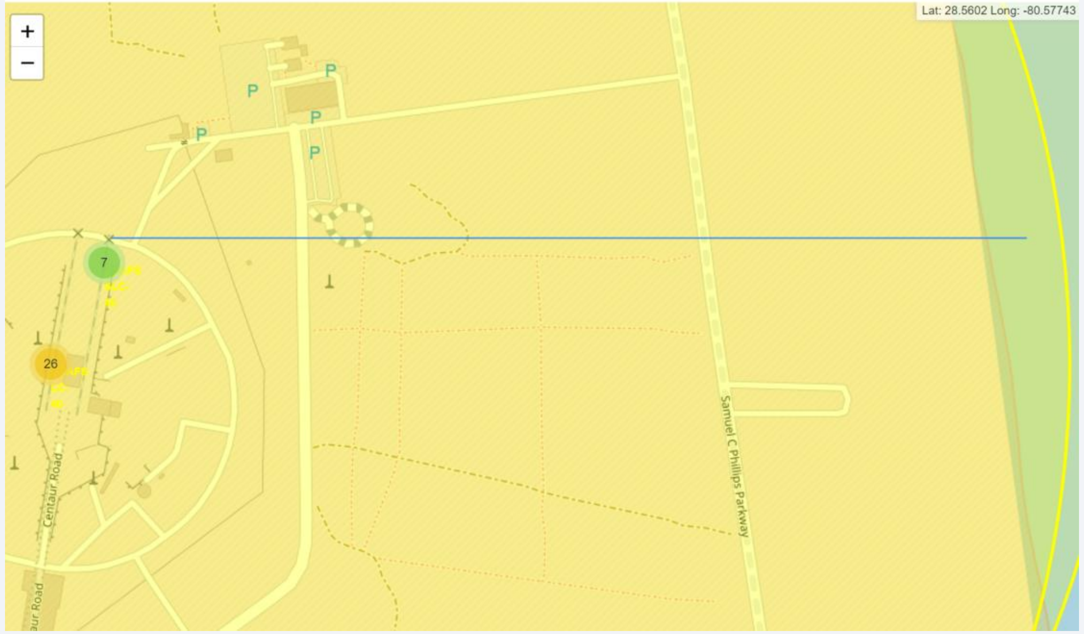
Colour labeled launch records on the map

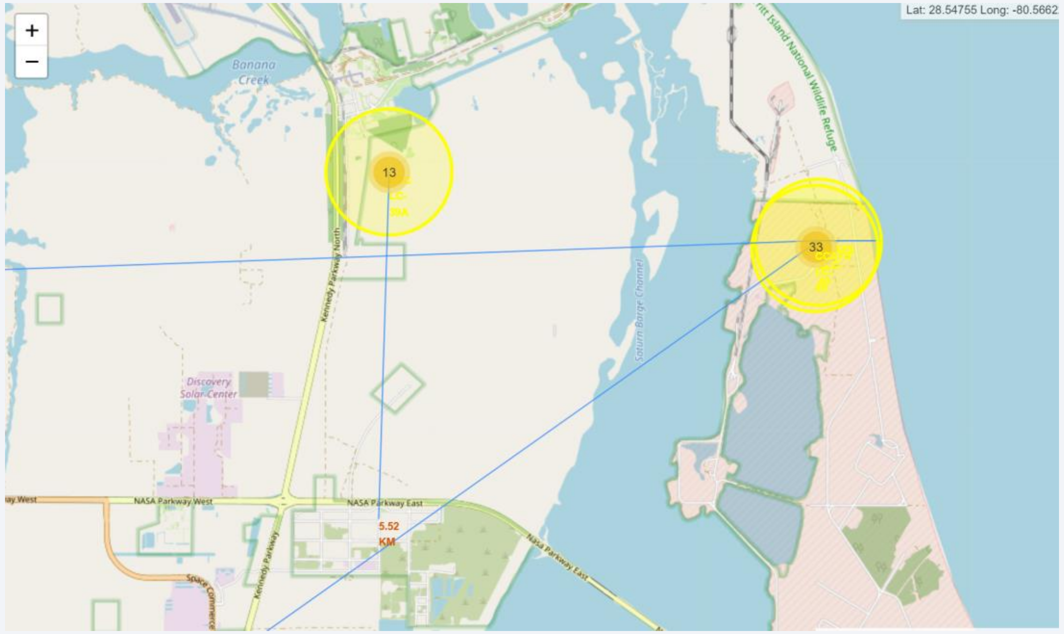




Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.

Distance from the launch site KSC LC-39A to its proximities



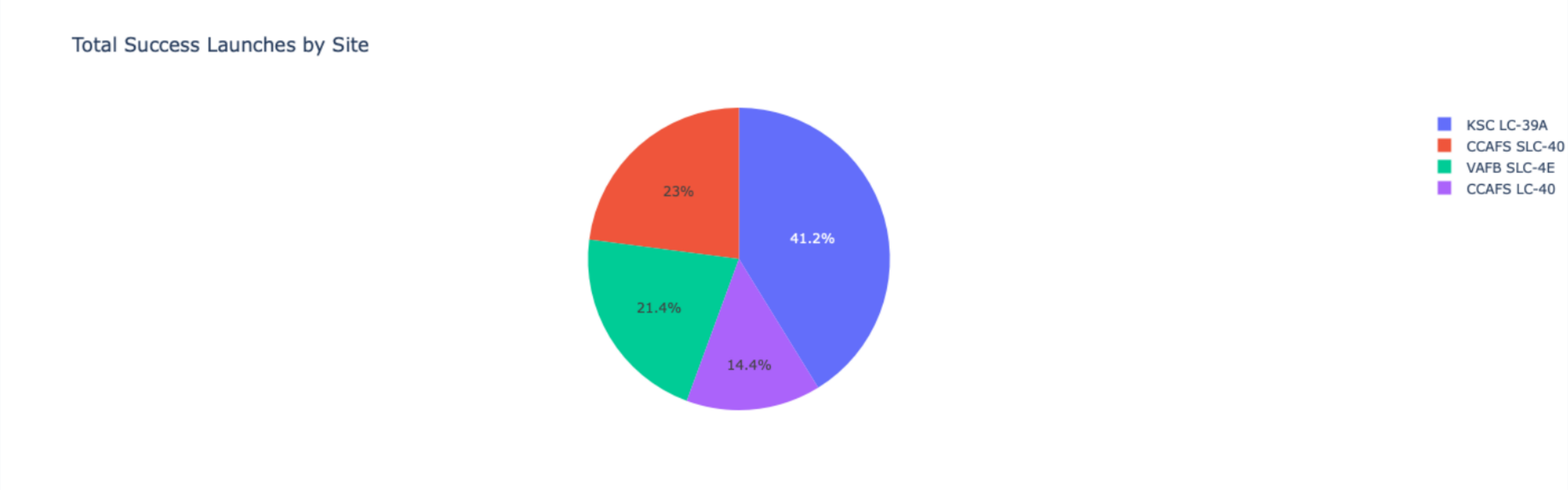


Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 4

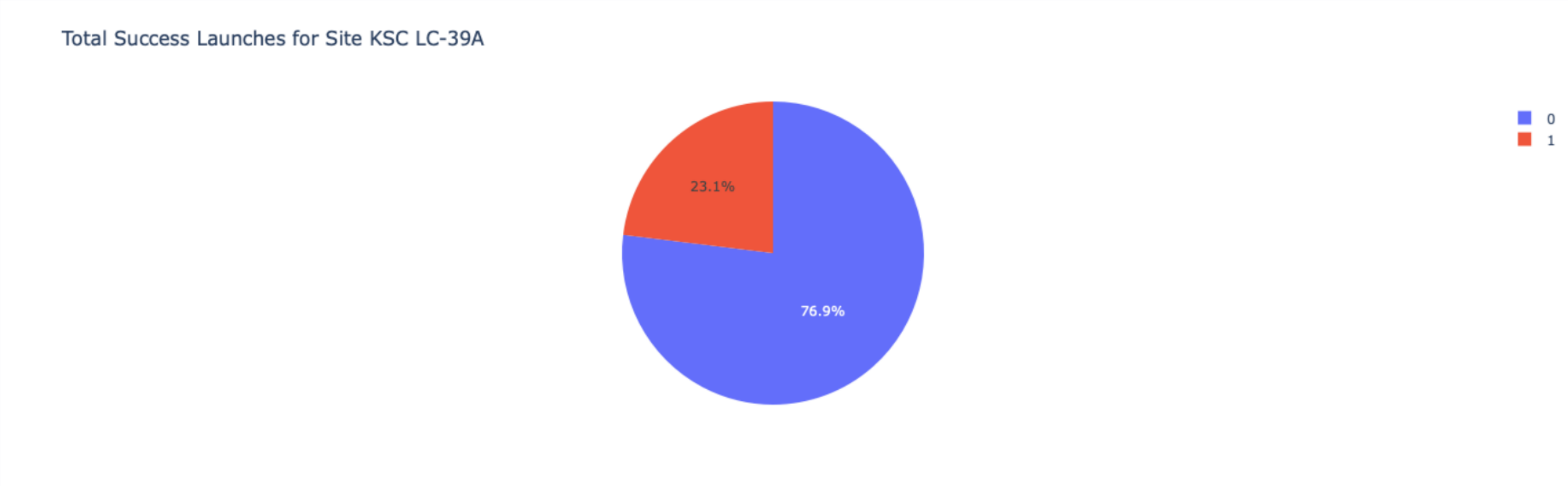
Build a Dashboard with Plotly Dash

Dashboard – Total success by Site



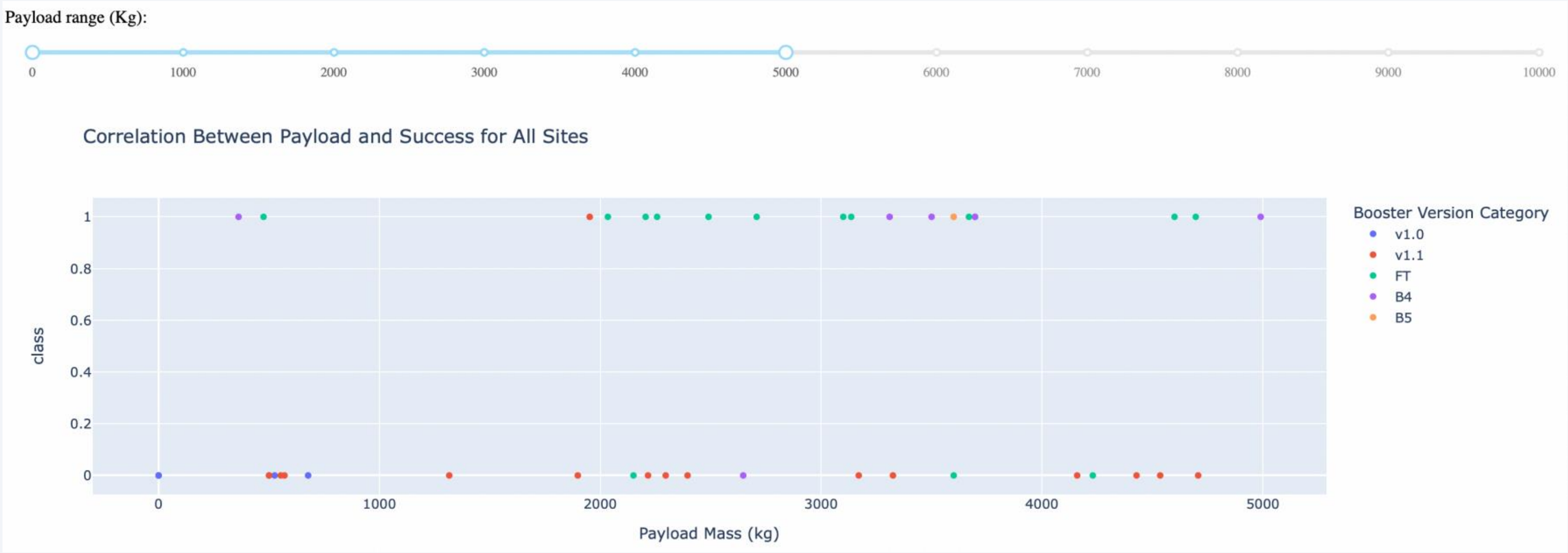
We can observe that KSC LC-39A has the best success rate of launches.

Dashboard – Total success launches for Site KSC LC-39A



KSC LC-39A has the highest launch success rate of (76.9%)

Dashboard – Payload mass vs Outcome for all sites with different payload mass selected



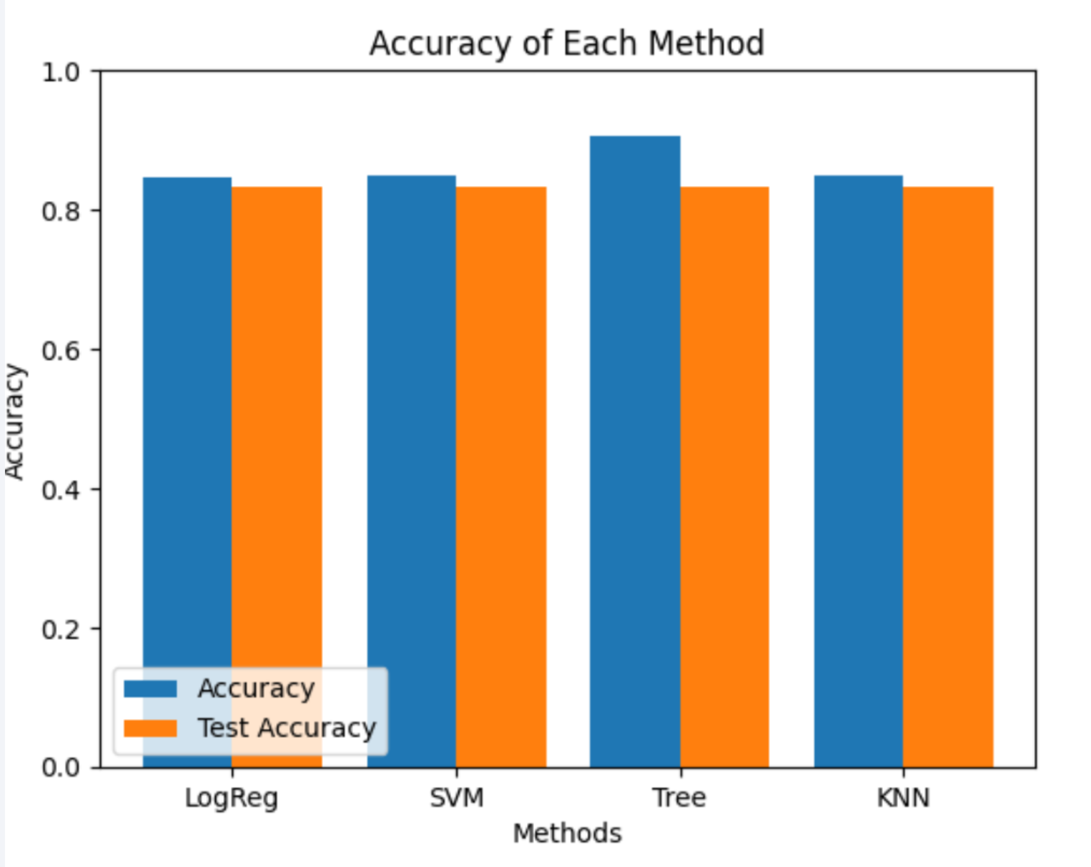
We can observe that low weighted payloads have a better success rate than the heavy weighted payloads

Section 5

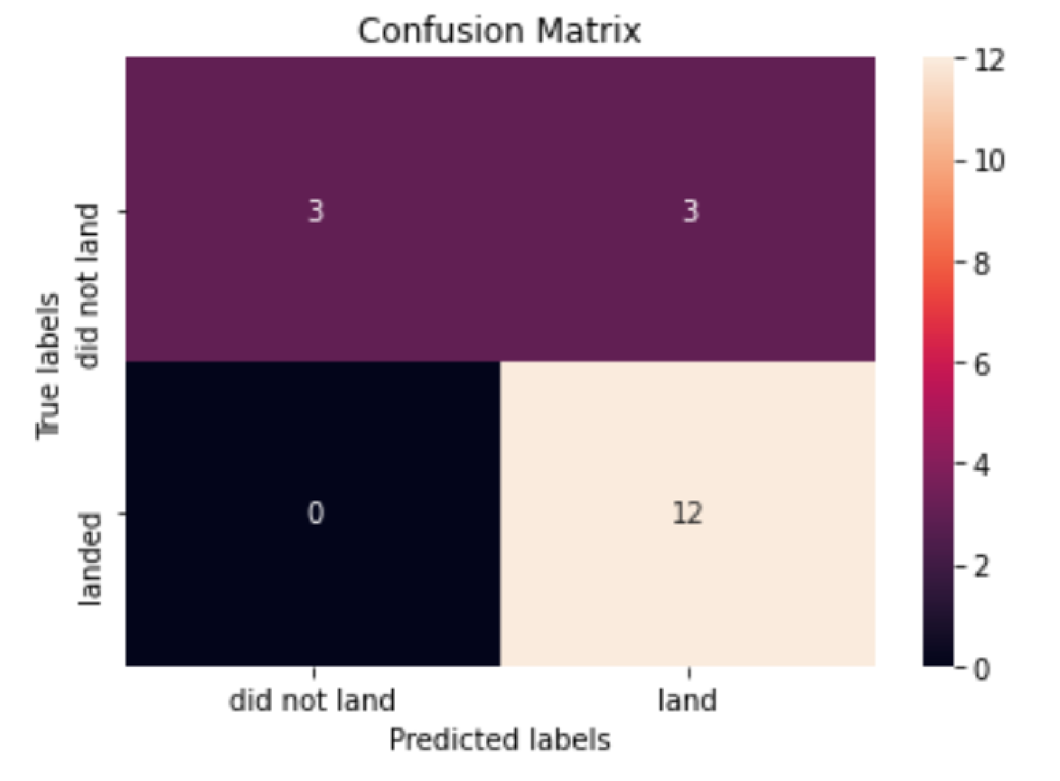
Predictive Analysis (Classification)

Classification Accuracy

We can conclude that the decision tree classifier is the model with the highest classification accuracy



Confusion Matrix



Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Launches above 7,000kg are less risky;
- Successful landing outcomes rate increase over the years.
- From all sites, KSC LC-39A has the highest success rate of the launches.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!