IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

María Rodríguez
19/09/2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection through API
  - Data collection using web scraping
  - Data wrangling
  - Exploratory data analysis with SQL
  - Exploratory data analysis using data visualization
  - Interactive visual analytics with Folium
  - Machine Learning prediction
- **Summary of all results**
  - Exploratory data analysis results
  - Screenshots from interactive analysis
  - Predictive analytics results

Introduction

- **Project background and context**

  While other providers rocket launch cost upward of 165 million dollars each, Falcon 9 advertises their cost in an average of 62 million dollars, thanks in a big part due to Space X can reuse the first stage. In the other hand, to make this happen the first stage should land successfully. If we can determine if the first stage will land, we can determine the cost of a launch and with this information an alternate company could bid against Space X for a rocket launch.

  The goal of this project is to create a Machine Learning pipeline to predict if the first stage will land successfully, this way we can predict estimate cost of a successful launch.

- **Problems you want to find answers**

  - Identify factors which determine the successfully land of the rocket.
  - Interaction amongst various features that determine the success rate of a successful landing.
  - Operating conditions required to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX Rest API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features.
- 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 collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems.

In this dataset this process was made using REST API and web scrapping from Wikipedia.

REST API: Started using get request method. Once information was obtained, response content was decoded as Json and converted into a pandas data frame using json_normalize(). After that, data was cleaned, looked for missing values and corrected if needed.

Web scraping: BeautifulSoup was used to extract launch records as HTML table, parse and convert it to a pandas data frame for an exhaustive analysis.

Data Collection – SpaceX API

Completed notebook:

SpaceX data collection API

get request using API for SpaceX data

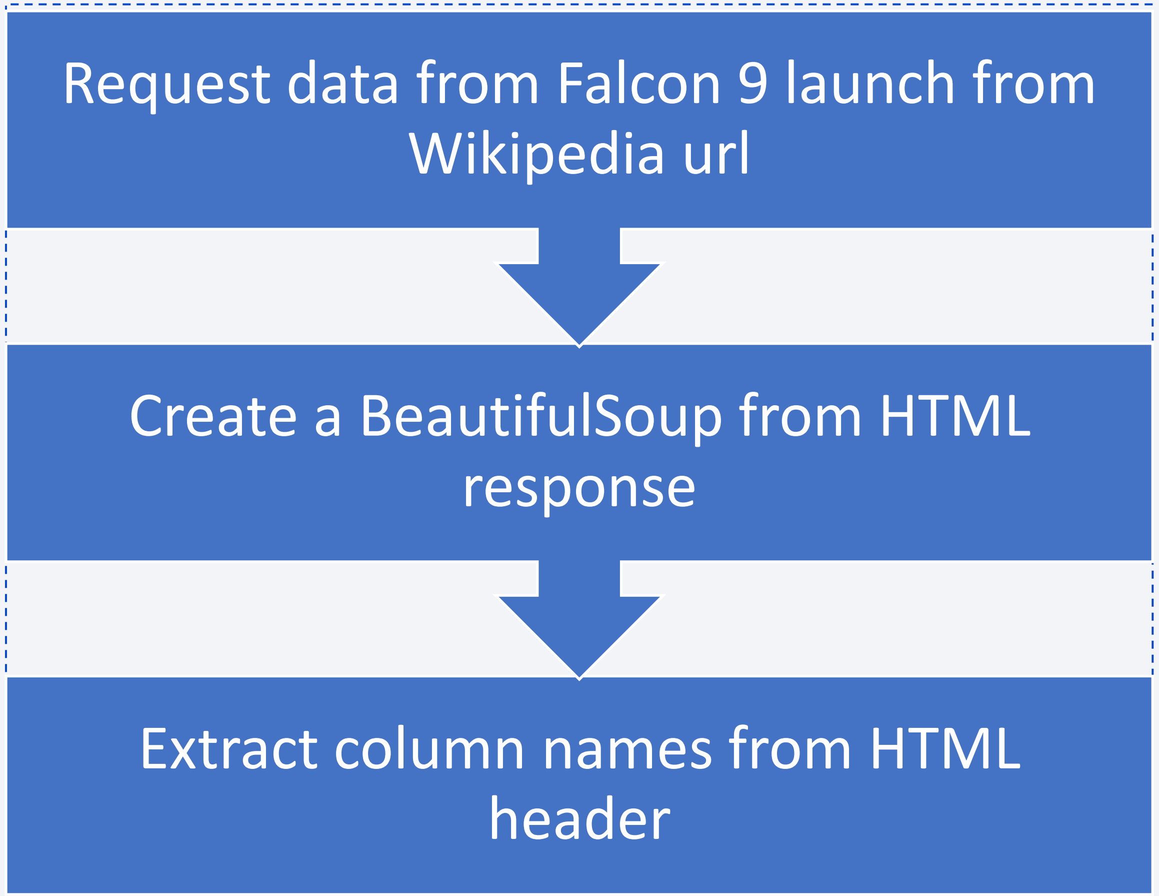
json_normalize and pandas to convert json result into a dataframe

Data cleaning and correcting missing values

Data Collection - Scraping

Completed notebook:

[SpaceX data collection using webscraping]



Data Wrangling

Data wrangling: is the process of transforming and mapping data from aw into another format with the intent of making it more appropriate and valuable for a variety of purposes such as an easy access and Exploratory Data Analysis (EDA).

Process: First, calculate number of launches for every site and the number and occurrence of mission outcome per orbit type.

Create a landing outcome label from outcome column to make it easier for further analysis, visualization and Machine Learning process. Last but not least, export the result to a CSV file.

Completed notebook: SpaceX data wrangling process

EDA with Data Visualization

First a scatter graph show us the relationship between for the following attributes:

- Payload vs Flight number.
- Flight number vs Launch site.
- Payload vs Launch site.
- Flight number vs Orbit type.
- Payload vs Orbit type.

Scatterplots help showing attributes' dependency on each other. It's simple to see factors affecting the most to the success of the landing outcomes.

EDA with Data Visualization

After analyze the information from scatter plot and visualize possible relationships, more visualization tools are needed for further analysis, such as bar and line plots graphs.

Bar graphs are useful to see relationships between attributes, in our case is used to determine which orbit have the highest probability of success.

Line graphs are used to show trends of the attribute over time.

Feature Engineering will help to predict success by creating dummy variables to categorical columns.

Completed notebook: SpaceX EDA with data visualization

EDA with SQL

We performed different queries using SQL to understand better information obtained, e.g.:

- **Displaying:**

Names of launch sites.

Total payload mass carried by booster launched by NASA (CRS).

Average payload mass carried by Booster version F9 v1.1.

- **Listing:**

Date when first successful landing outcome in ground pad was achieved.

Names of boosters which have success with payload mass over 4000 and less that 6000.

Total number of successful and failure missions

Completed notebook: SpaceX EDA with SQL

Build an Interactive Map with Folium

An interactive map was created with Folium for better visualize where successful and failed attempts took place and where launch sites are placed.

Latitude and longitude coordinates for each launch site were taken and a circle marker added with the name of it.

Failure attempts were assigned to red color while successful were assigned to green color using MarkerCluster().

To better understand where launch sites are place distance from railways, highways, coastlines and nearby cities were calculated and provided in the interactive map.

Completed notebook: SpaceX Interactive Folium map

Build a Dashboard with Plotly Dash

An interactive dashboard was created using Plotly dash so if information is updated so does the dashboard.

In this dashboard we can find a pie chart with total launches in which we could select an specific location if needed.

A scatter graph is in the dashboard too, which shows the relationship between outcome and payload mass in kilograms, where we can also select between the different booster versions.

Completed dashboard: SpaceX Dashboard

Predictive Analysis (Classification)

Building

- Load dataset into Numpy and Pandas.
- Transform data and split it for training and test datasets.
- Review performance of each type of ML to decide which one to use.
- Set parameters and algorithms to GridSearch and fit into the dataset.

Evaluating

Review accuracy for each model.

Get tuned hyperparameters for each type of algorithms.

Plot a confusión matrix.

Improving

Use Feature Engineering and algorithm tuning.

Select

Model with the best accuracy score is the best performing model to be used.

Completed predictive analysis: SpaceX Predictive Analysis

Results

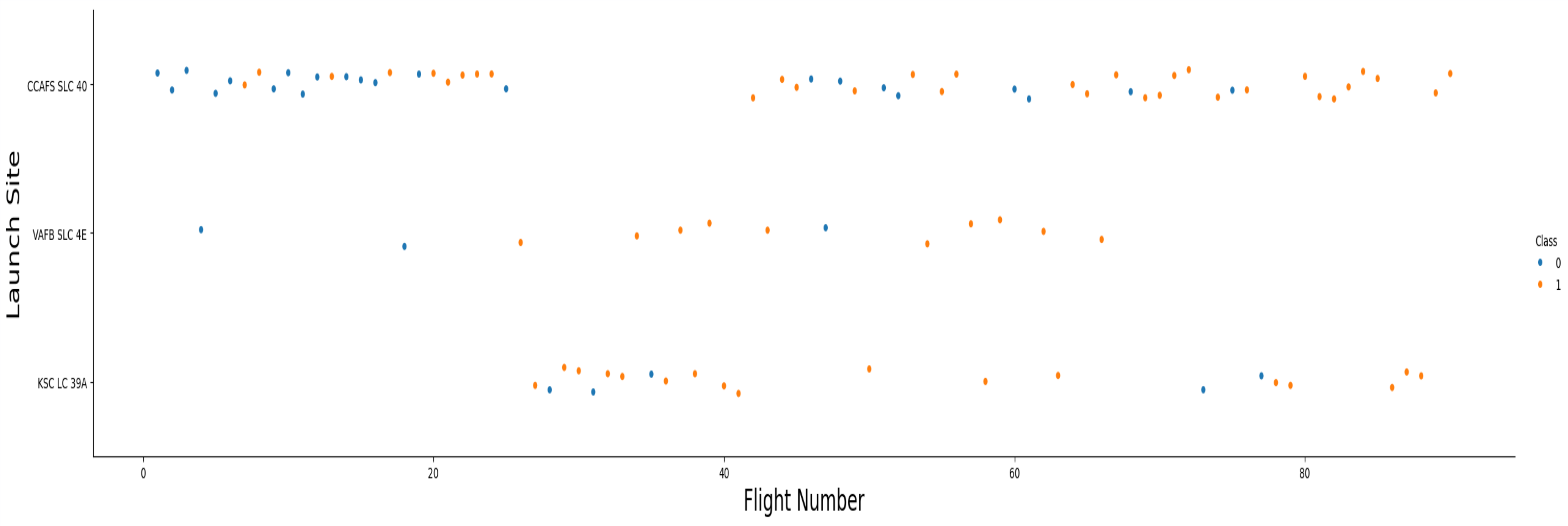
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

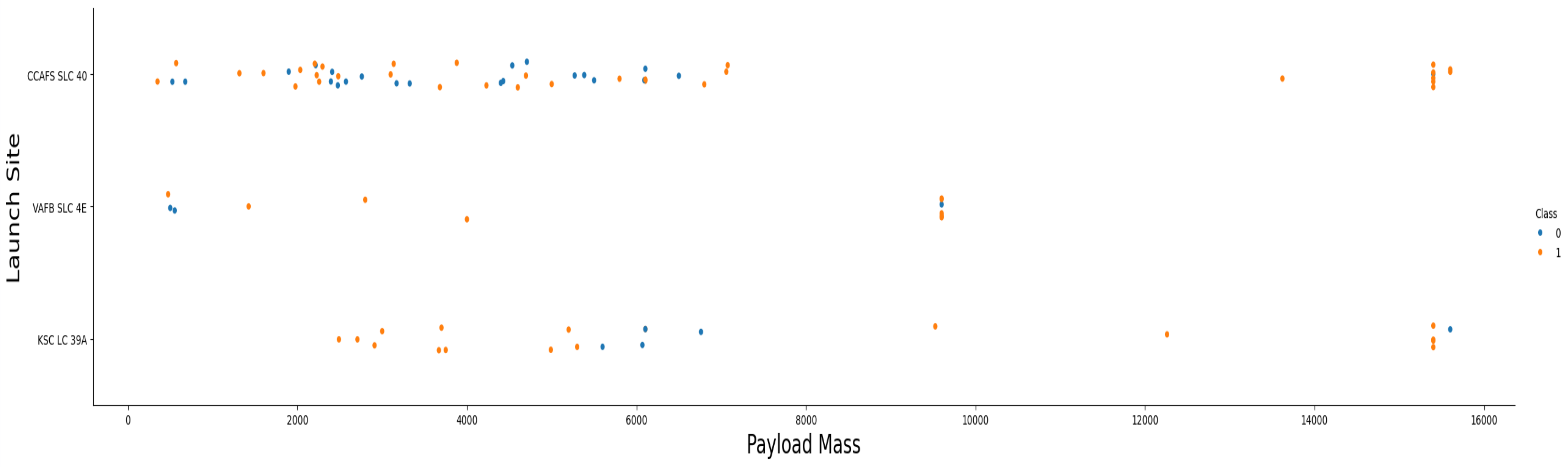
Flight Number vs. Launch Site

As we can see below, there are more launches from CCAPS and less from VAFB but the amount of successful launches are higher in the second one and as the flight number increases, the first stage is more likely to land successfully.



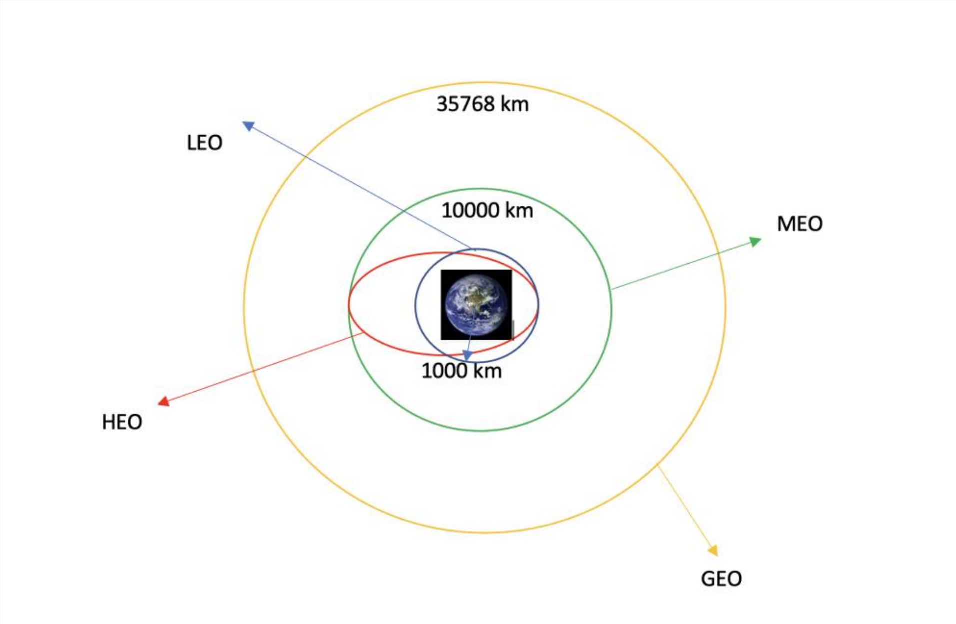
Payload vs. Launch Site

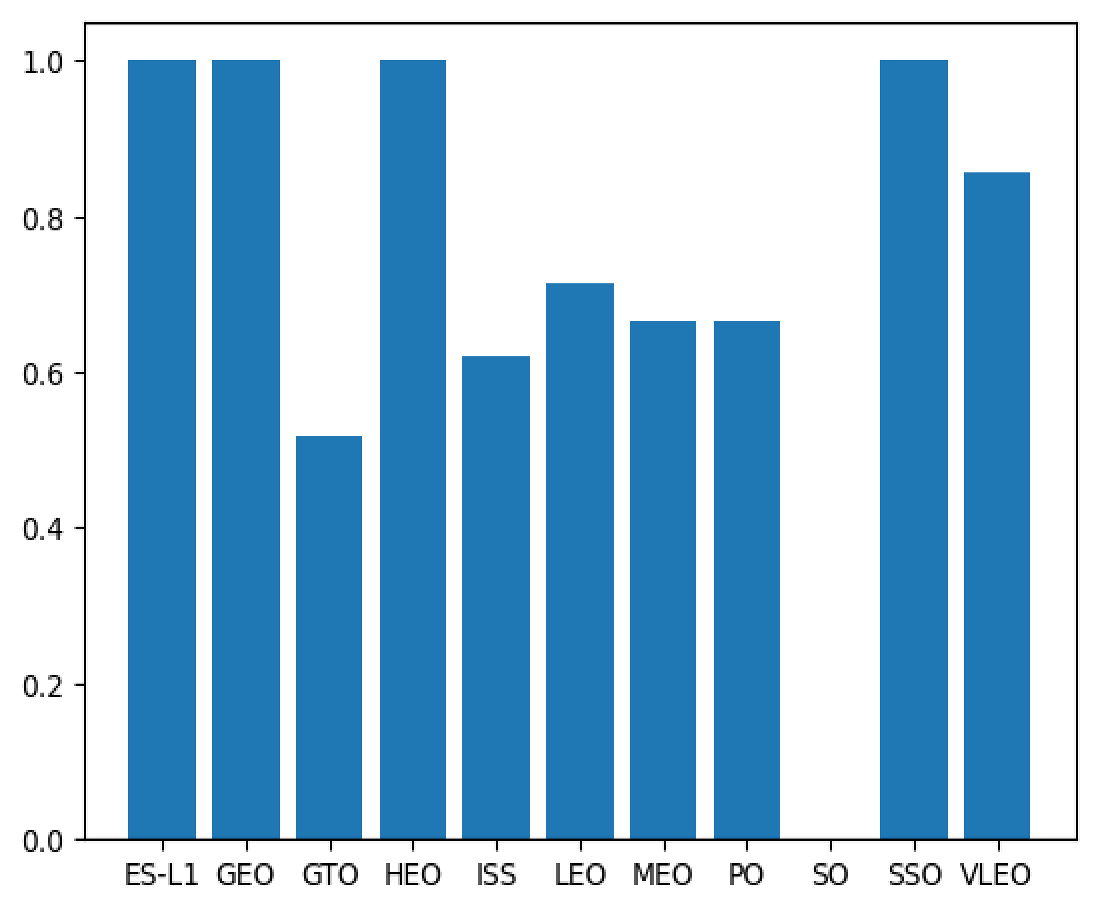
Now if we observe Payload Mass vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10.000).



Success Rate vs. Orbit Type

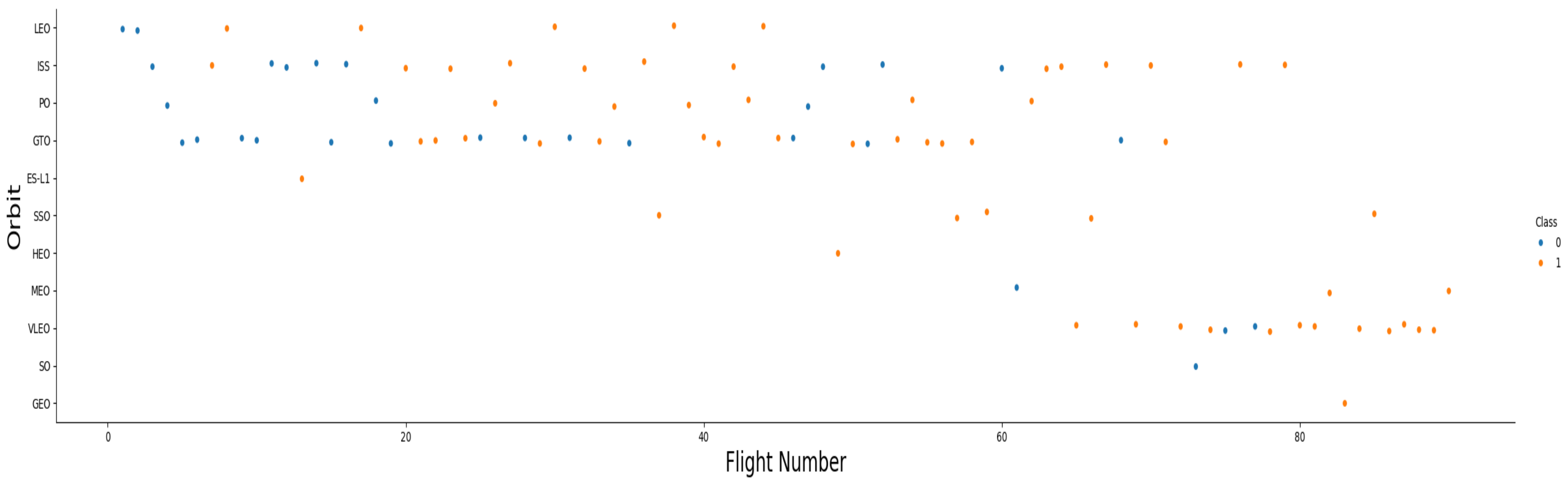
The orbits with higher success rate are ES-L1, GEO, HEO and SSO, the lowest success rate is for SO, meanwhile the rest have and average success rate.





Flight Number vs. Orbit Type

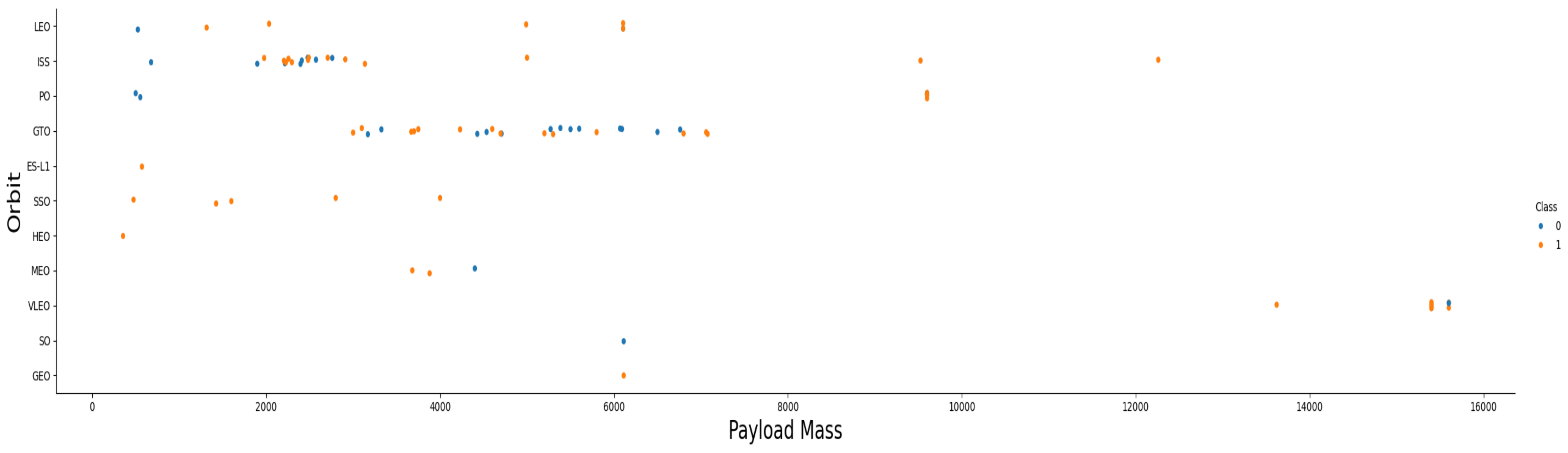
In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

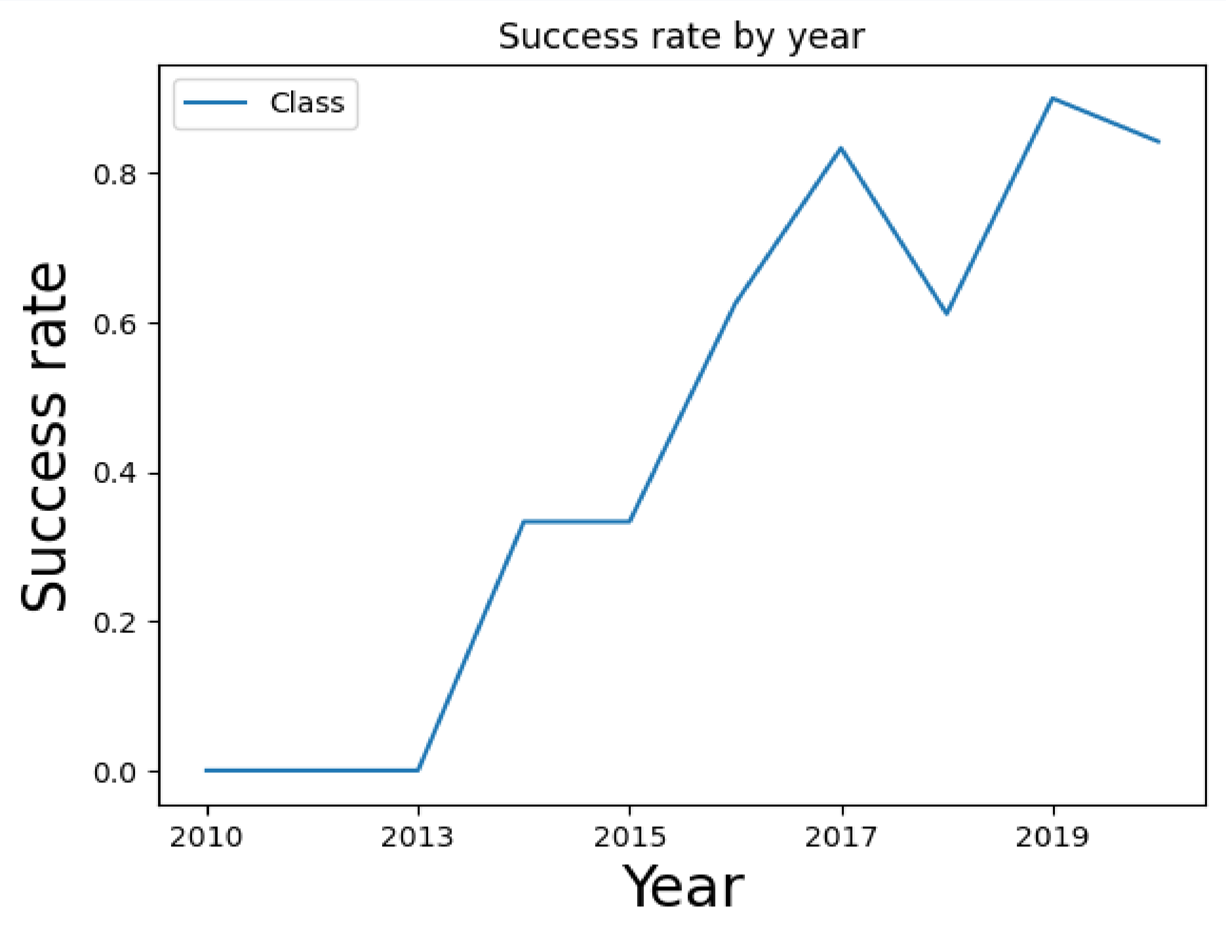
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

Success rate since 2013 kept increasing till 2020.



All Launch Site Names

Using SQL we can find the unique names from launch sites as showed 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

With SELECT DISTINCT sql returns unique names from variable ‘Launch site’ and FROM we are selecting ‘Spacextbl’ which is the table where information is contained.

Launch Site Names Begin with 'CCA'

If we want to select 5 records where launch sites begin with ‘CCA’ the command LIKE is used with % for missing values and LIMIT with the number of results wanted.

```
%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
2010-06-04	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

If we want to calculate the total payload carried by boosters from NASA, we need the function SUM which sums the column PAYLOAD MASS KG and with the WHERE clause we filter results for NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_)  FROM SPACEXTBL WHERE CUSTOMER="NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

To calculate the average payload mass carried by booster version F9 v1.1 we use function AVG which means average after SELECT and filter results for version like with WHERE clause, including LIKE in case some version are written in a different way.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_)  FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE "F9 v1.1%";
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

First successful landing on ground pad was on July 22th in 2018 according with SQL, to get this information we SELECT the older date with MIN(DATE) filtering with WHERE clause the LANDING OUTCOME in which dates are

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME="Success";
```

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

MIN(DATE)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 we SELECT column called Booster Version FROM table called SPACEXTBL filtering with the WHERE clause by Payload mass AND Success (drone ship)

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) AND (Landing_Outcome="Success (drone ship)");
```

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

If we want to calculate the total number of successful and failure mission outcomes we need to SELECT column Mission Outcome, COUNT the quantity of that Mission Outcome and GROUP BY the different Mission Outcome we found in our table.

```
%sql SELECT Mission_Outcome, COUNT (Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

Here are the list of the names of the booster which have carried the maximum payload mass, to get this information we need a subquery looking for maximum payload mass.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below:

```
%sql SELECT substr(Date,6,2) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome F
FROM SPACEXTBL WHERE Landing_Outcome="Failure (drone ship)" AND substr(Date,0,5)="2015";
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Ranking shows Failure (drone ship) or Success (ground pad) in descending order:

```
%sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY Landing_Outcome HAVING Landing_Outcome= 'Success (ground pad)' OR Landing_Outcome='Failure (drone ship)'
    ORDER BY Landing_Outcome DESC;
```

Landing_Outcome	COUNT(*)
Success (ground pad)	3
Failure (drone ship)	5

Section 3

Launch Sites Proximities Analysis

Launch sites for SpaceX in USA

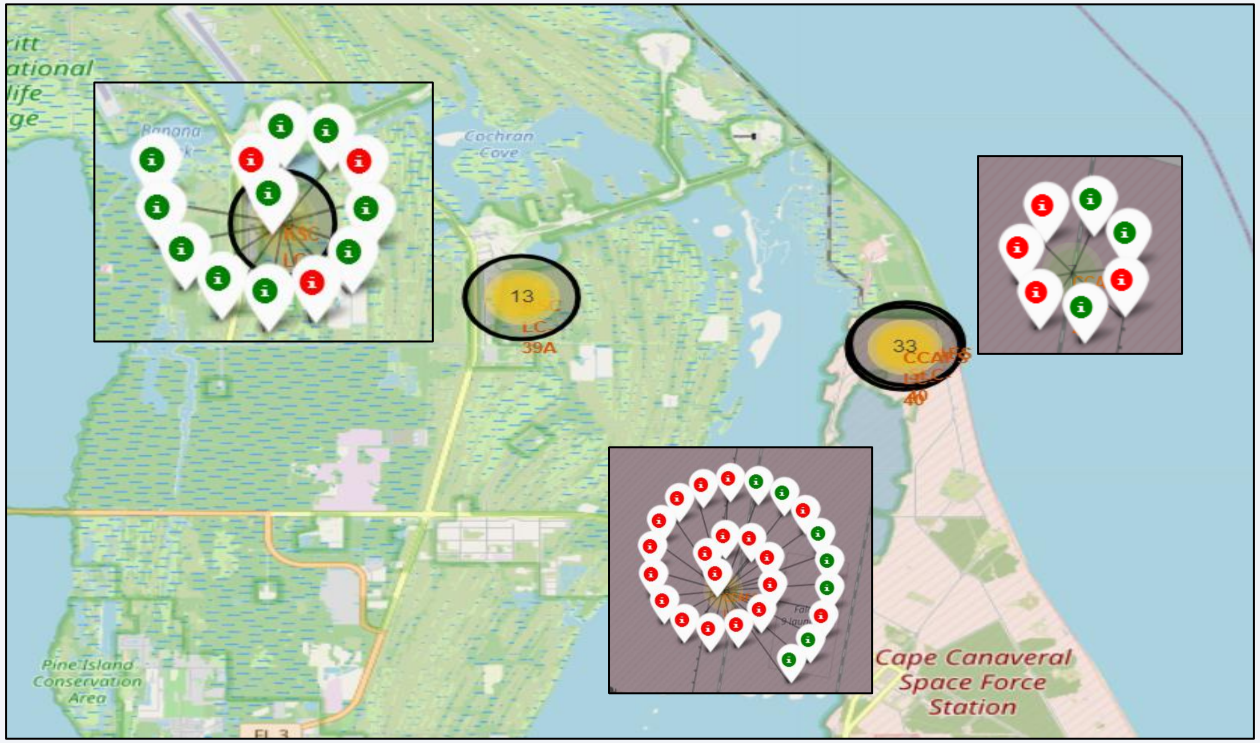
As we can see launch sites are place close to the coast within USA.

Launch sites with color labels

Green markers shows successful launches while red ones shows failed ones.

California

Florida



Launch Site distance to key places

Launch sites are close to railways as transport is cheaper an easy by train in case they need materials from other states they are close to highways too as is better for people who work there and materials which can't be transported by train. ·

Launch sites are close to coastline for safety reasons but aren't close to cities to avoid risks.

Section 4

Build a Dashboard with Plotly Dash

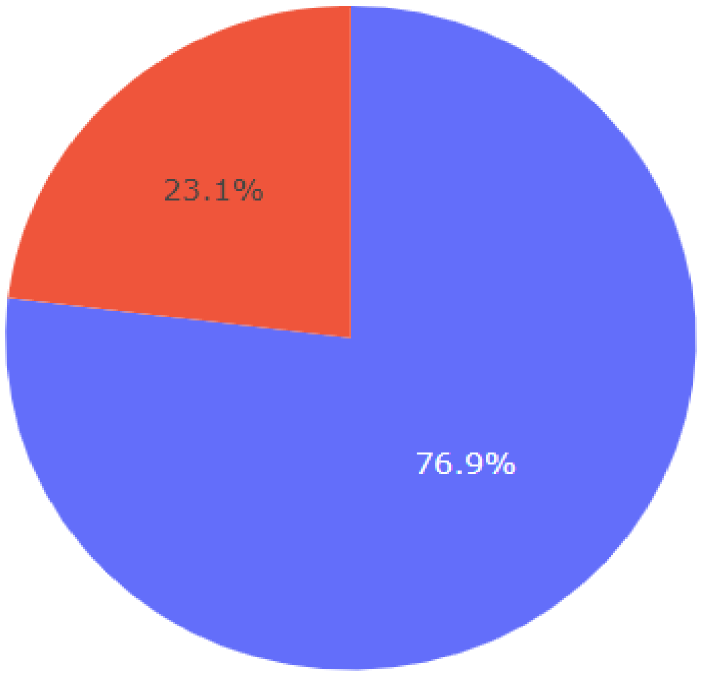
Pie chart: Success by each launch site

Most successful launches are from KSC LC 39A

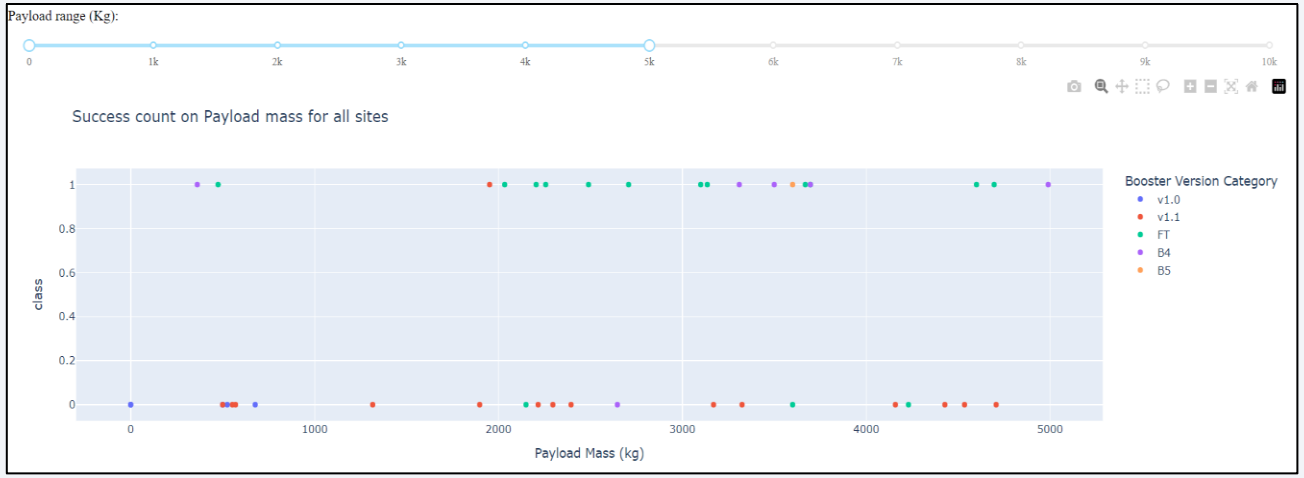
Pie chart from KSC LC 39A

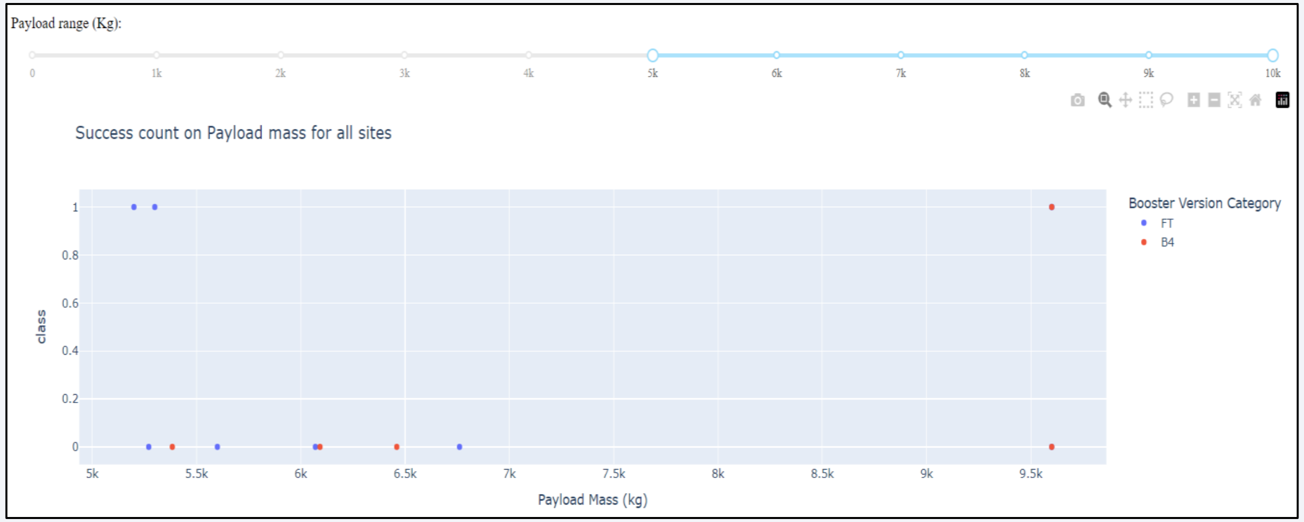
Total Success Launches for site KSC LC-39A

KSC LC 39A has the higher successful rate with a 76,9% of success versus 23,1% of failed attempts.



Payload vs. Launch Outcome scatter plot for all sites





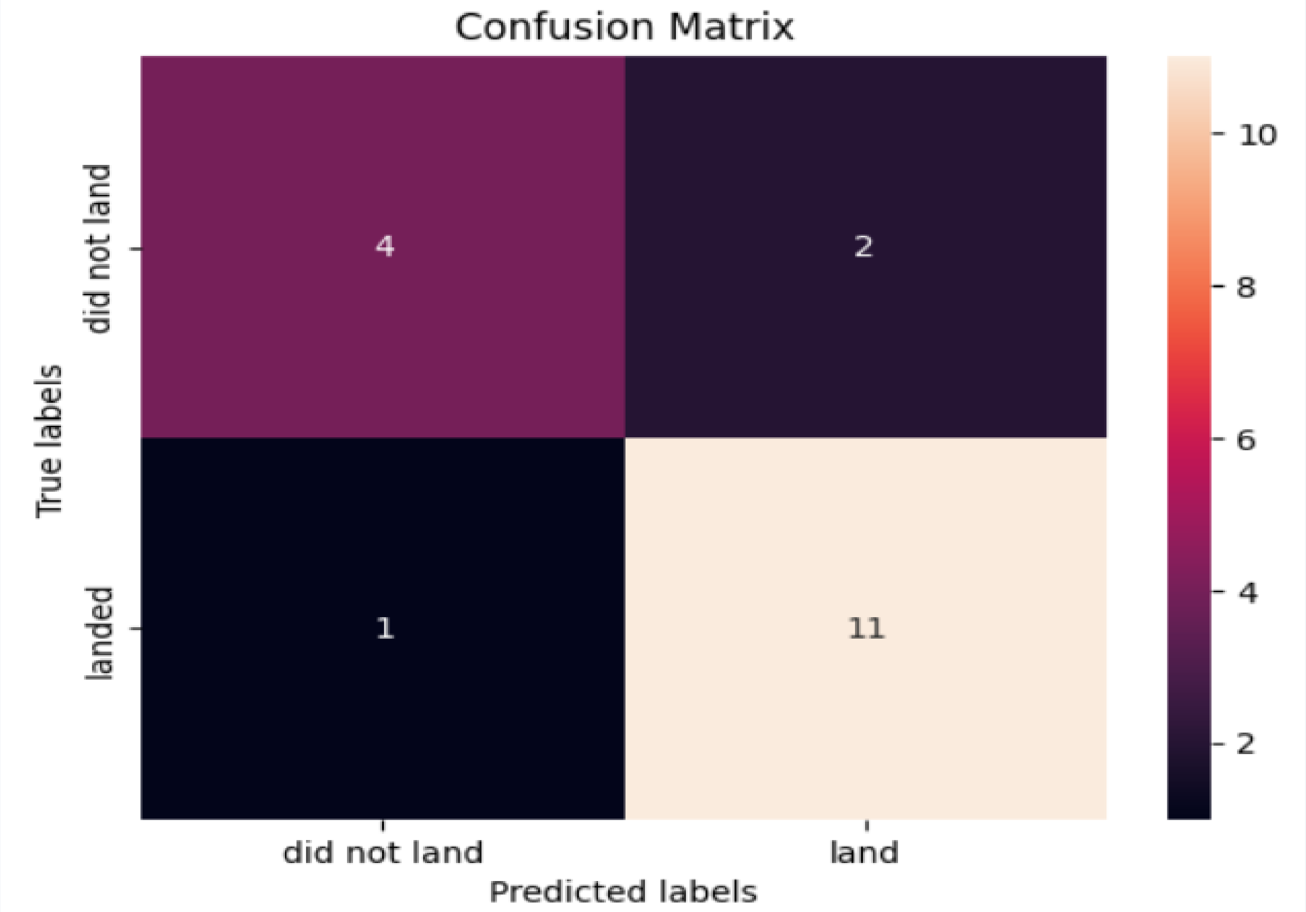
Success rates for payloads over 5000kg are lower than for smaller payloads.

Section 5

Predictive Analysis (Classification)

Confusion Matrix

Confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes better than other methods.



Conclusions

- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm in this case.

Thank you!