Winning Space Race with Data Science

Jesús Antonio Cortés Asencio
February 19 2023

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scrapping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Folium
- Machine Learning Predictions

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive analytics result

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX. One reason SpaceX can do this is the rocket launches are relatively inexpensive, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems we want to find answers:
  - What factors determine if the rocket will land successfully
  - The interaction amongst various features that predict the rate of successful landing
  - If we can predict the success of landing in the first stage of launching, then we can determine the cost of launch. **This information can be used if an alternate company wants to bid against SpaceX for a rocket launch**

Section 1

Methodology

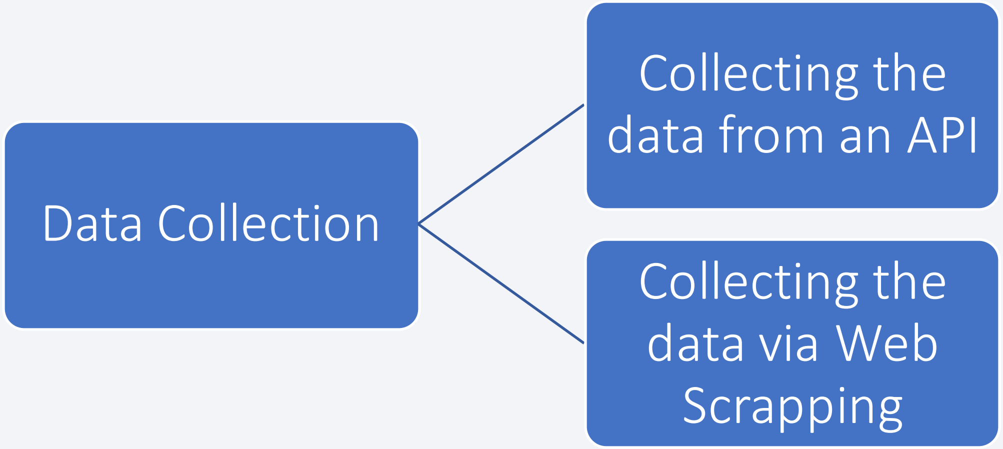
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scrapping 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 was done via two methods: Data gathered from an API, specifically the SpaceX REST API. This gave us information about launches and rockets used, payload delivered, launch specifications, landing specifications and landing outcomes of the Falcon 9 rocket.
- The second method was web scrapping some HTML tables that contained valuable Falcon 9 launch records



Data Collection – SpaceX API

- We performed a get request to the SpaceX API. Basic data wrangling and formating was done: request to the SpaceX API. Clean the requested data

- GitHub URL of the completed SpaceX API calls notebook API Data Collection

Request and parse the SpaceX launch data using the GET request

↓

Filter the dataframe to include only Falcon 9 launches

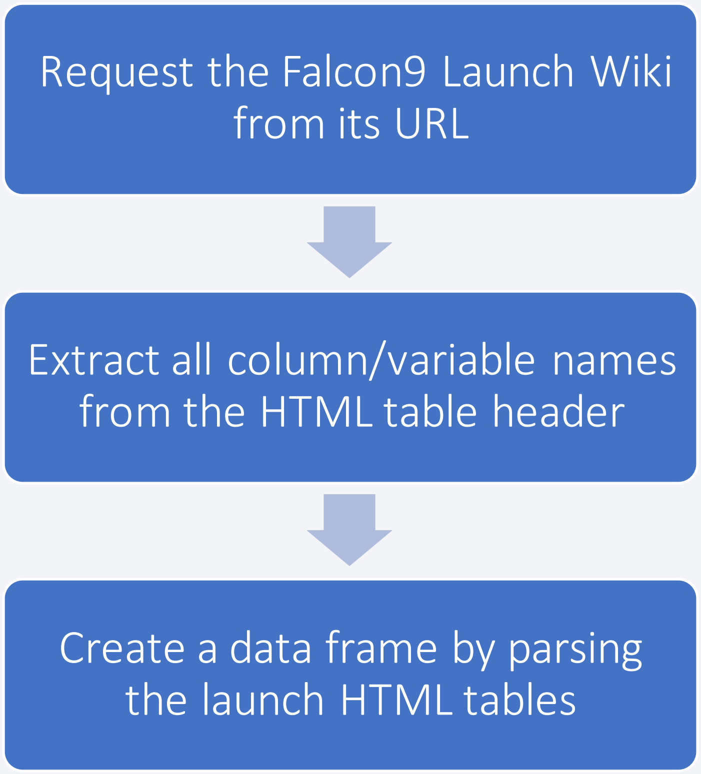
↓

Dealt with missing values

Data Collection - Scraping

- We web scrapped Falcon 9 launch records with BeautifulSoup: We extracted a Falcon 9 launch records HTML table from Wikipedia. Parsed the table and convert it into Pandas data frame

- GitHub URL of the completed web scraping notebook [Data Collection Web Scrapping]()



Data Wrangling

- The column Outcome indicates if the first stage successfully landed. There are 8 of them. True ASDS means the booster successfully landed to a drone ship. False ASDS means the mission outcome was unsuccessfully. We converted landing outcomes to Classes 0 or 1, where 0 means a bad outcome and 1 a good outcome.

Calculate the number of launches on each site → Calculate the number and occurence of each orbit → Calculate the number and occurence of mission outcome per orbit type → Create a landing outcome label from Outcome column

- **GitHub URL of the completed web scraping notebook** [Data Wrangling]()

EDA with Data Visualization

- We performed exploratory Data analysis and feature engineering using Pandas and Matplotlib. Specifically we plotted the charts:
  - Flight number vs Launch Site (scattered plot)
  - Pay load mass vs Launch Site (scattered plot)
  - Success rate of each orbit (bar plot)
  - Flight number vs Orbit (scattered plot)
  - Pay load mass vs Orbit (scattered plot)
  - Launch success yearly trend (line plot)
- GitHub URL of your completed EDA with data visualization notebook EDA with Data Visualization

EDA with SQL

- We delved deeper into the data set using SQL analysis, and we found:
  - The unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - The total Payload mass carried by boosters launched by NASA (CRS)
  - The average Payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - The names of the booster versions which have carried the maximum payload mass
- GitHub URL of your completed EDA with SQL notebook SQL EDA

Build an Interactive Map with Folium

- We analyzed launch site and success rates using Folium. We created a map and marked all launch sites with a highlighted circle. Then we marked the success/failed launches for each site on the map with a MarkerCluster object. Finally, we calculated the distance between launch site to its proximities.
- We created the circles and markers on the launch sites to visualize this locations by pinning them on a map. And then we plot distance lines to the proximities of launch sites to seek relations of success rates to proximities to railways, highways, coastlines, cities.
- GitHub URL of completed interactive map with Folium Folium Analysis

Build a Dashboard with Plotly Dash

- We created a Dashboard using Plotly Dash. We add the option to visualize in a pie plot the Total Success Launches by Site and scatter plots of success count on Payload mass for all sites, in this case we also add a slider to select the Payload range (kg).
- We added the pie chart for an easy visualization of successful launches per site and we added the scatter plots to see the correlation between payload and launch success.
- The GitHub URL of the completed Plotly Dash Dashboard

Predictive Analysis (Classification)

- We built a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully. This included: preprocessing, standarized data and train-test-split data. We trained the model to perform Grid Search, which allowed us to find the hyperparameters that allow a given algorithm to perform best.
- The GitHub URL of completed predictive analysis lab Predicitve Analysis

Machine Learning Models tested:

Logistic Regression	Support Vector machines	Decision Tree Classifier	K-nearest neighbors

Results

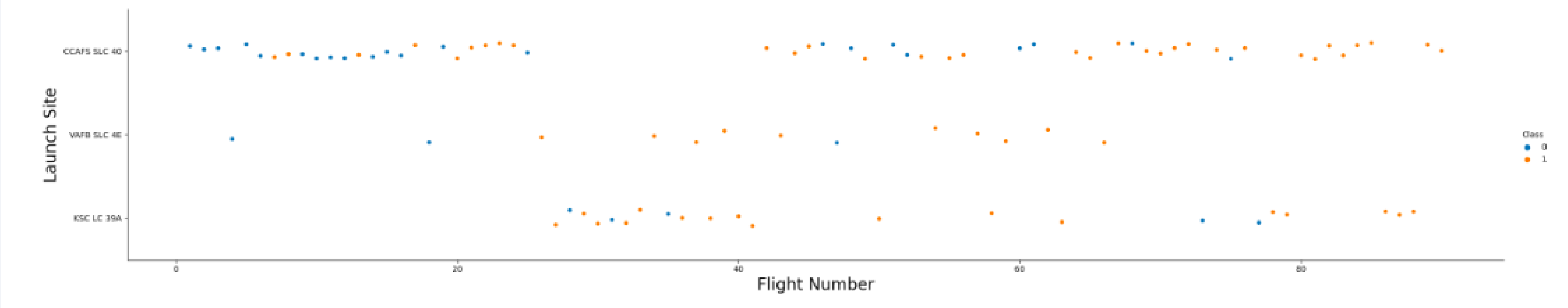
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

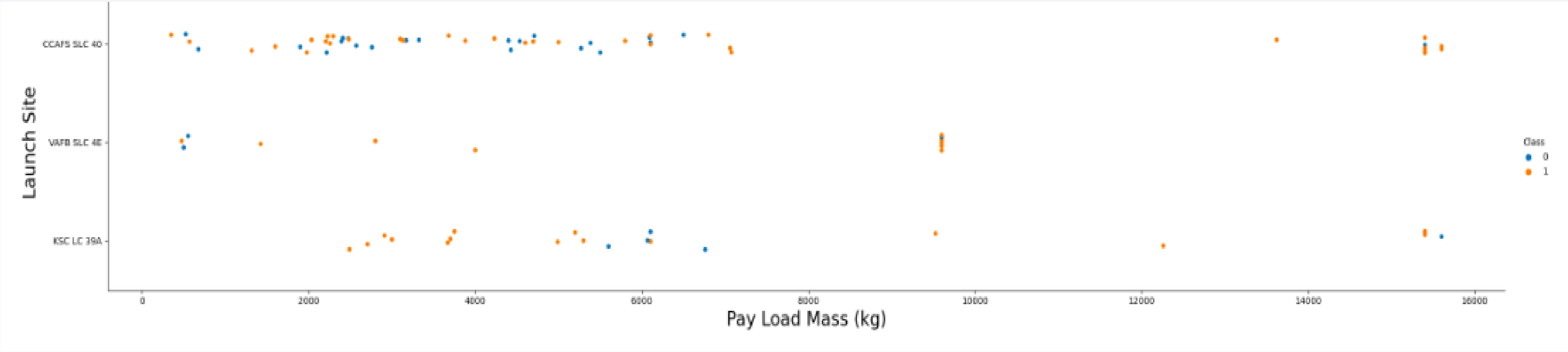
Flight Number vs. Launch Site

- We can see that as Flight Number increases the success of all Launch Sites also increase



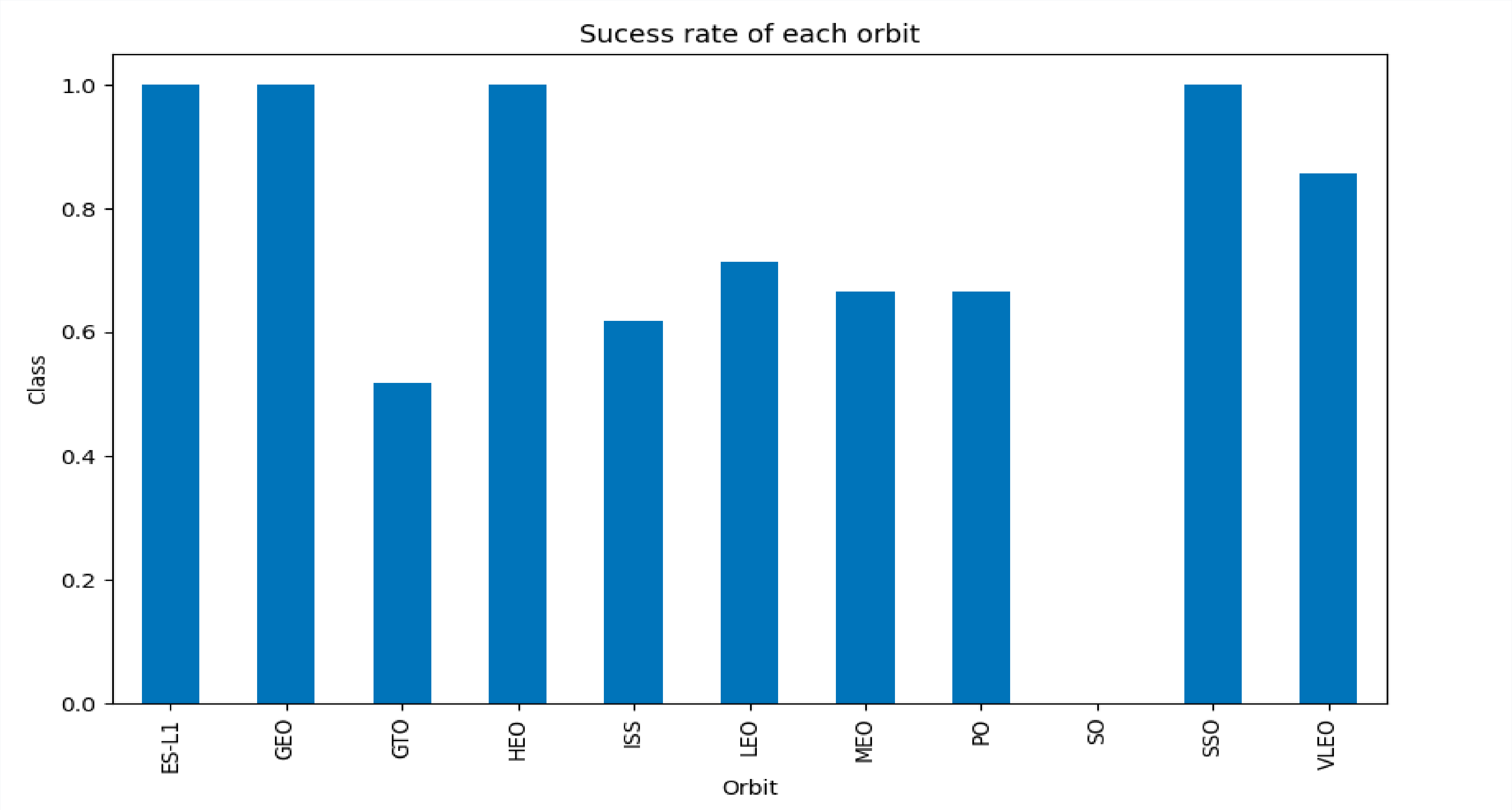
Payload vs. Launch Site

- We find for the VAFB-SLC Launch site there are no rockets launched for heavy payload mass >1000. Also, for the CCAFS SLC 40 Launch Site we find a high rate of success launches for heavy payload mass.



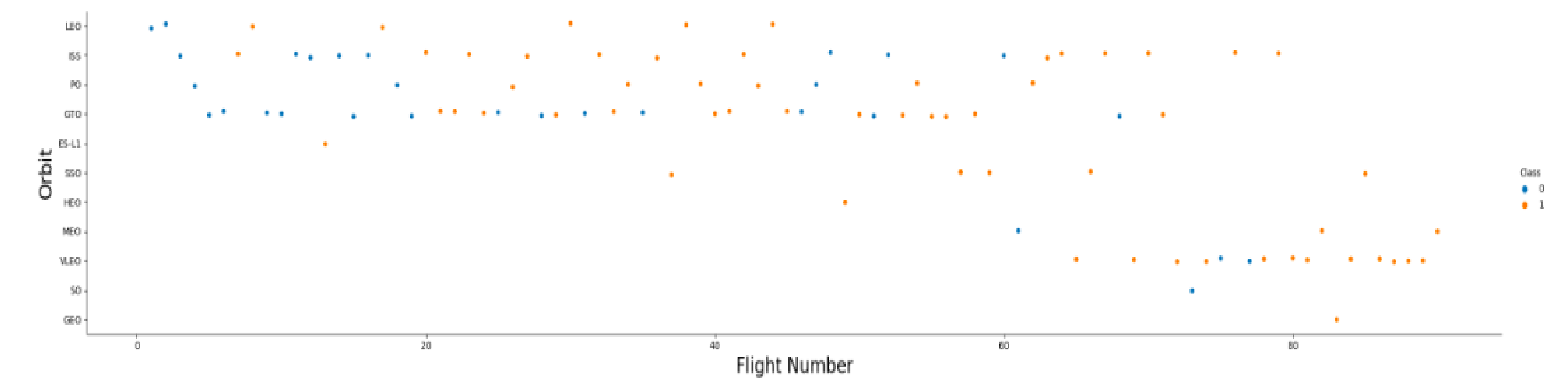
Success Rate vs. Orbit Type

- Clearly, we see the most successful rate on orbits: SSO, HEO, GEO and ES-L1. In contrast, we found the least successful rate in orbit GTO.



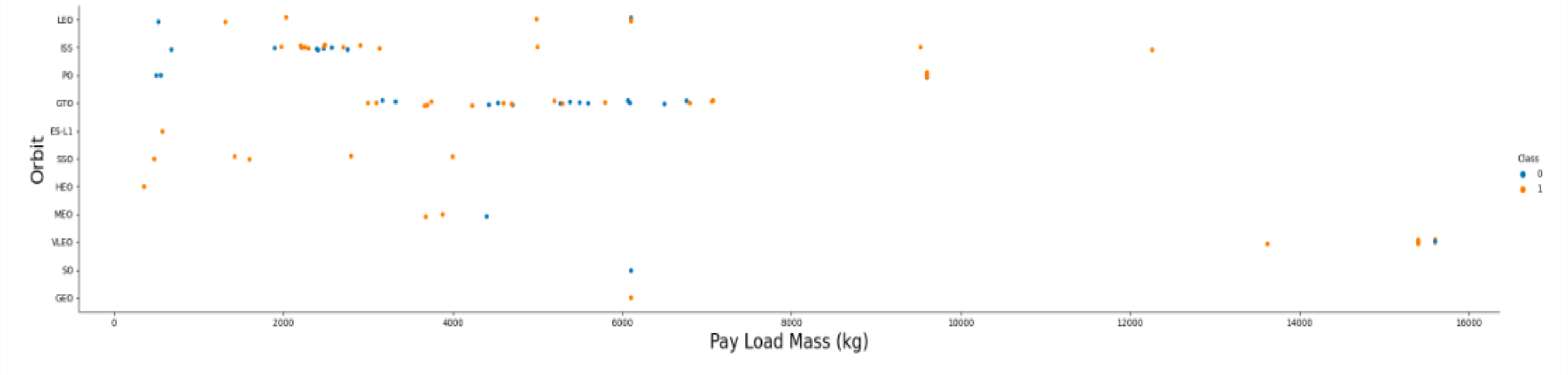
Flight Number vs. Orbit Type

- We see that in the LEO orbit the success appears related to the numbers of flights.



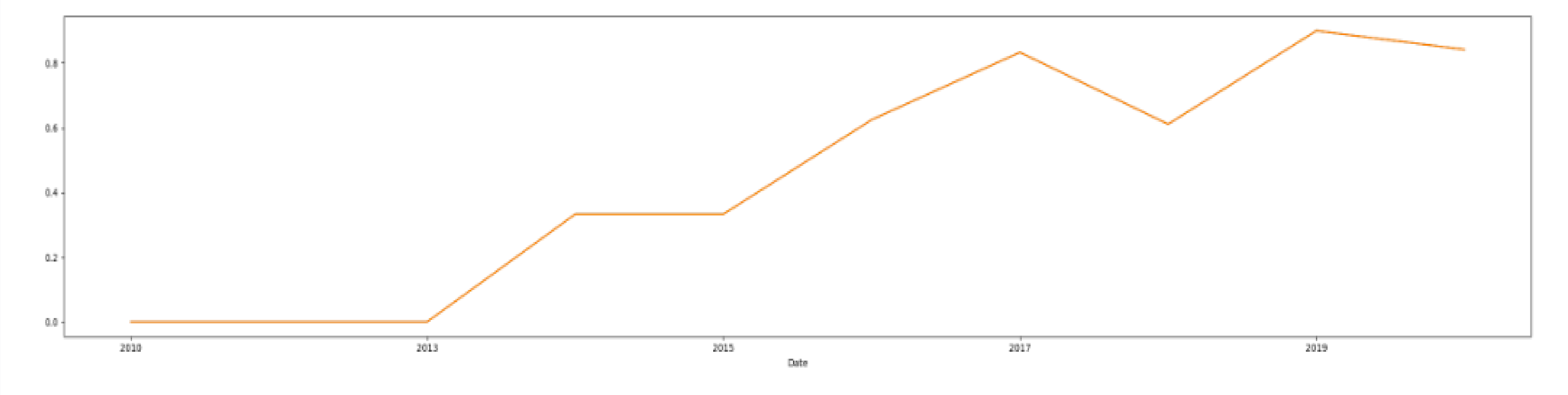
Payload vs. Orbit Type

- With heavy payloads the successful landing are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing are both present.



Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing until 2020



All Launch Site Names

- We identified all the unique names of the Launch sites, there are 4 different ones :

```
In [12]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

- Here we display 5 records whose Launch Sites begin with "CCA"

```
3]:  %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5

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

3]:

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

Total Payload Mass

- We found the total sum of the Payload Mass = 45596 kg

```
In [18]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = "NASA (CRS)"

 * sqlite:///my_data1.db
Done.

Out[18]:  SUM(PAYLOAD_MASS__KG_)
                           45596
```

Average Payload Mass by F9 v1.1

- We calculated the Average of the Payload Mass of the Falcon 9 v1.1

```
In [20]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version= "F9 v1.1"

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

Out[20]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We display the date of the first successful Ground Landing: 01-05-2017

In [18]:
```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome"= "Success (ground pad)";
```

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

Out[18]:

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- We displayed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [20]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome"= "Success (drone ship)" AND PAYLOAD_MASS__KG_<6000 AND PAYLOAD_MASS__I

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

Out[20]:

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

Total Number of Successful and Failure Mission Outcomes

- We found the number of failure mission outcomes is 1, meanwhile the total number of successful outcomes is 99

In [6]:
```
%sql SELECT COUNT(*) AS "List of FAILURE and SUCCESS mission outcomes" FROM SPACEXTBL WHERE Mission_Outcome LIKE "%%Success%%" UNION SELEC
```

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

Out[6]:

List of FAILURE and SUCCESS mission outcomes
1
100

Boosters Carried Maximum Payload

- We display the names of the booster which have carried the maximum payload mass

```
In [8]: %sql SELECT Booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =  (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL );

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

Out[8]:

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

- We display the list of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
In [6]: %sql SELECT substr(Date,4,2), "Landing _Outcome", Booster_version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _

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

Out[6]:

substr(Date,4,2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- We ranked 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

In [9]:
```
%sql SELECT Date, "Landing _Outcome" FROM (SELECT Date, "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome"= "Success") WHERE DATE
```

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

Out[9]:

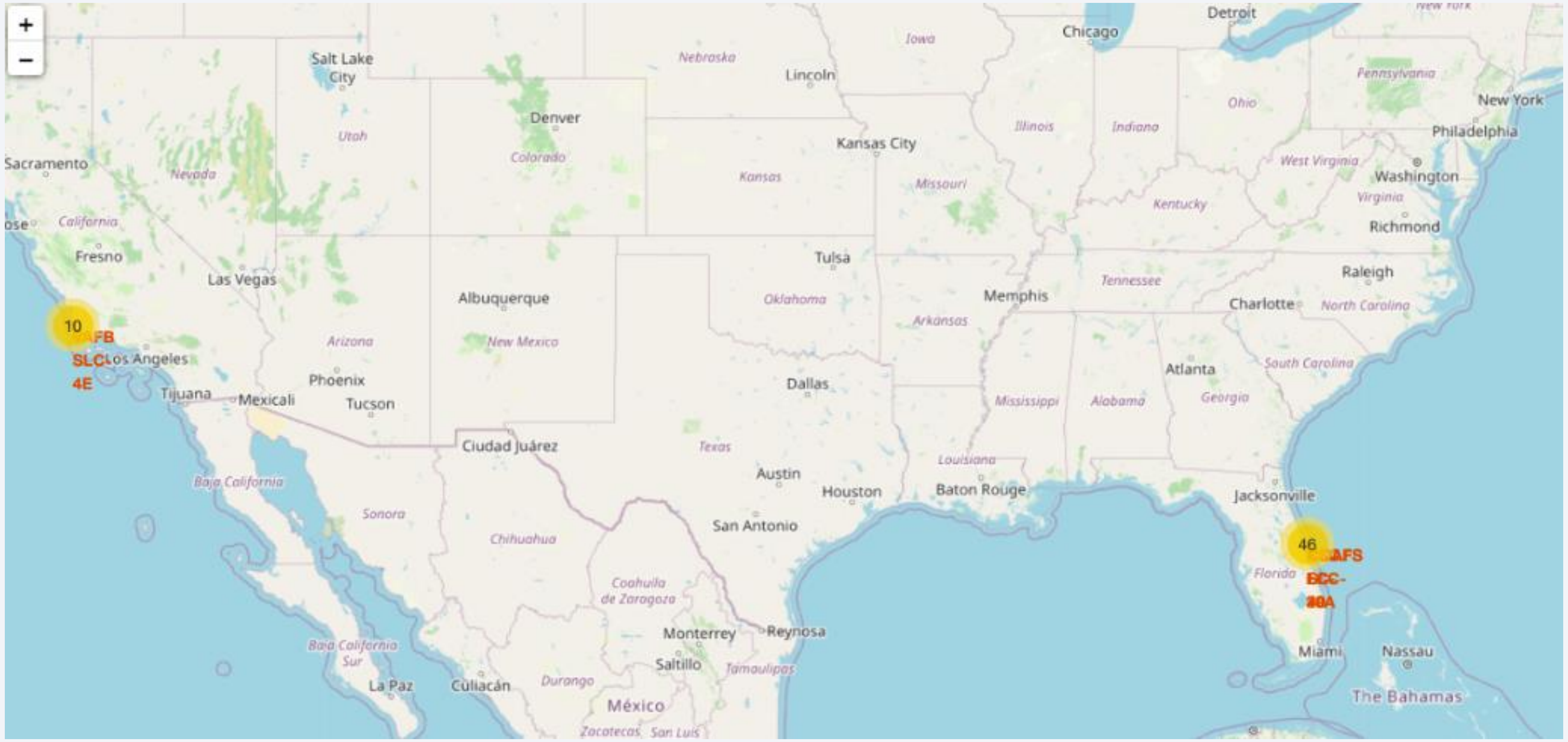
Date	Landing _Outcome
07-01-2020	Success
07-03-2020	Success
04-06-2020	Success
13-06-2020	Success
07-08-2020	Success
18-08-2020	Success
06-10-2020	Success
18-10-2020	Success
05-11-2020	Success
16-11-2020	Success
06-12-2020	Success
11-01-2019	Success
12-06-2019	Success
11-11-2019	Success
05-12-2019	Success
17-12-2019	Success
07-08-2018	Success
10-09-2018	Success
08-10-2018	Success
15-11-2018	Success

Section 3

Launch Sites Proximities Analysis

Folium Map with all Launch Sites locations

We created a Folium map where all Launch sites locations are highlighted with a yellow circle. Moreover, we display the number of launches per launch site

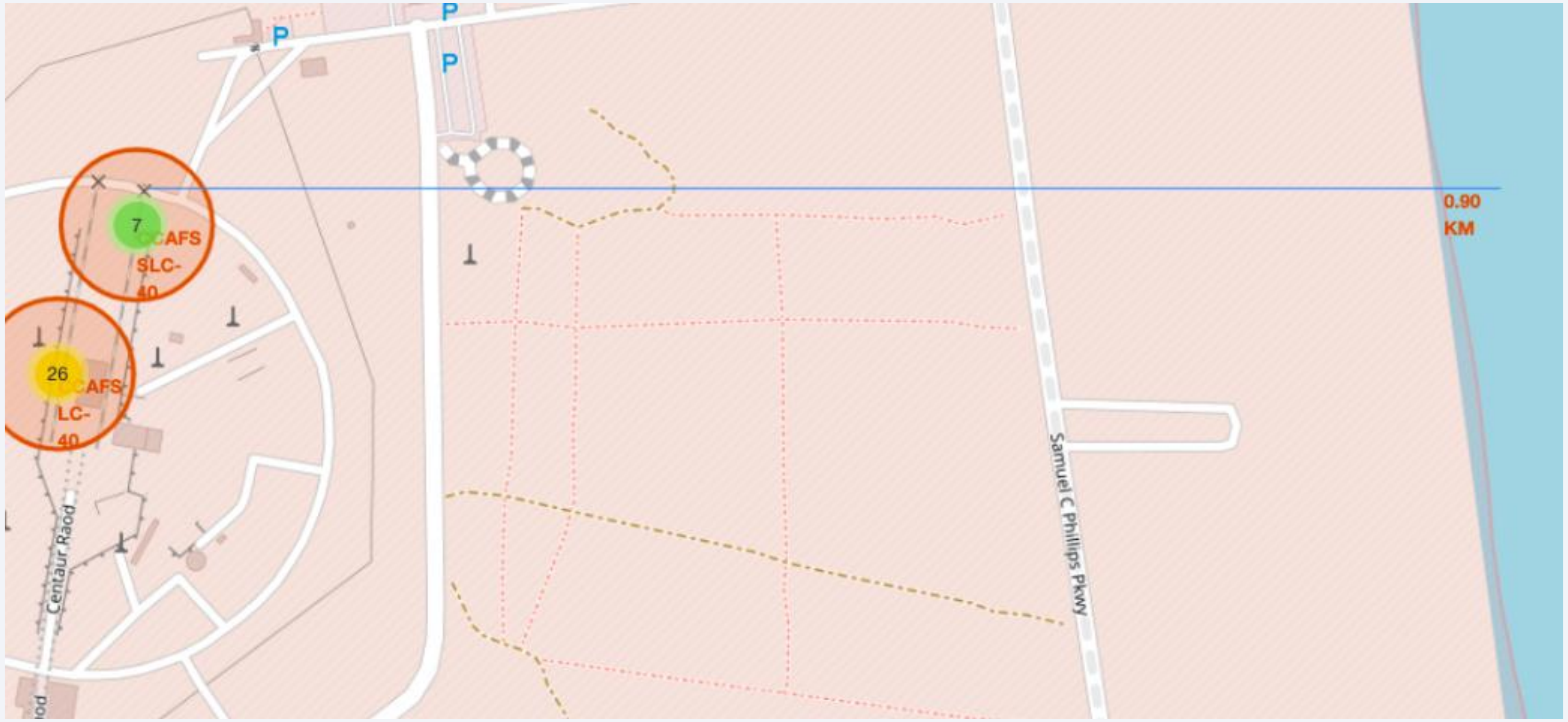


Folium map with success/failed launches for each site

For each launch site, we can vizualize with green marks the success launches and with red marks the failed ones.

Folium Map with Launch Site proximities

Moreover, we added a function to calculate and display de distance from Launch Sites to coast lines, highways and railways

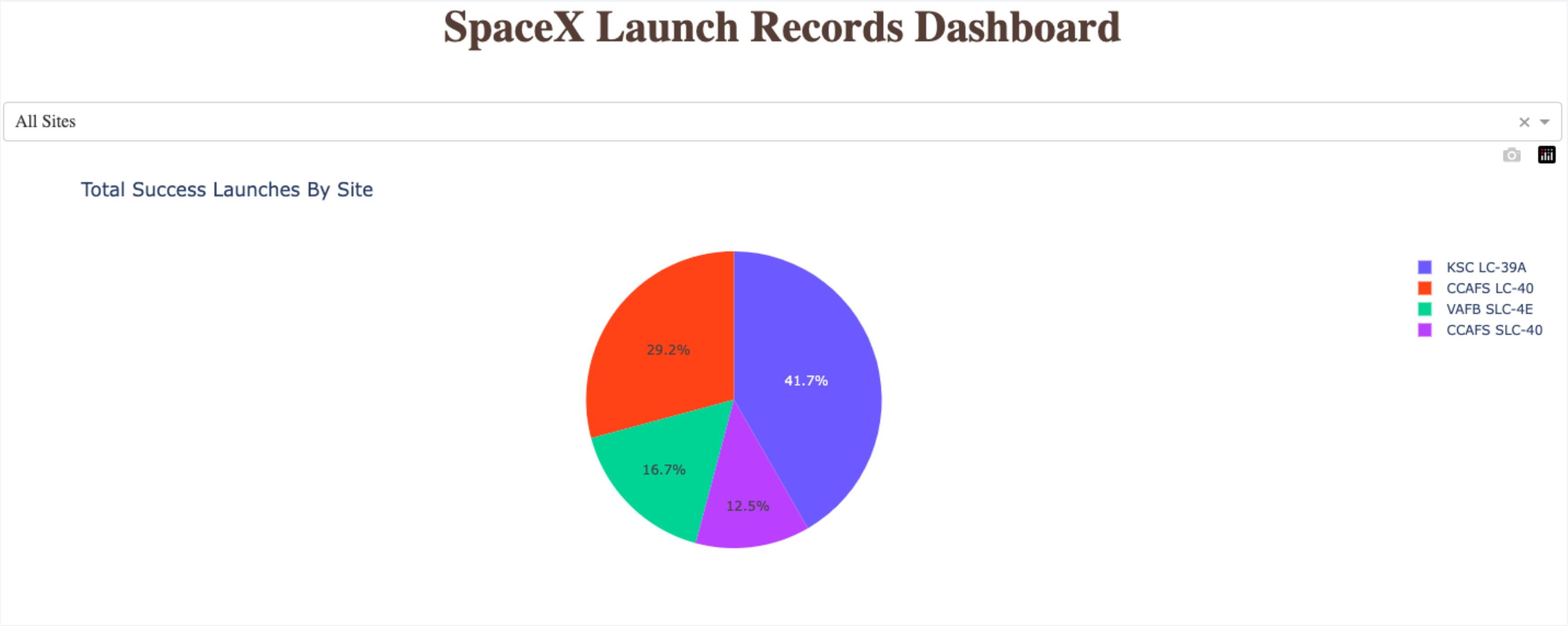


Section 4

Build a Dashboard with Plotly Dash

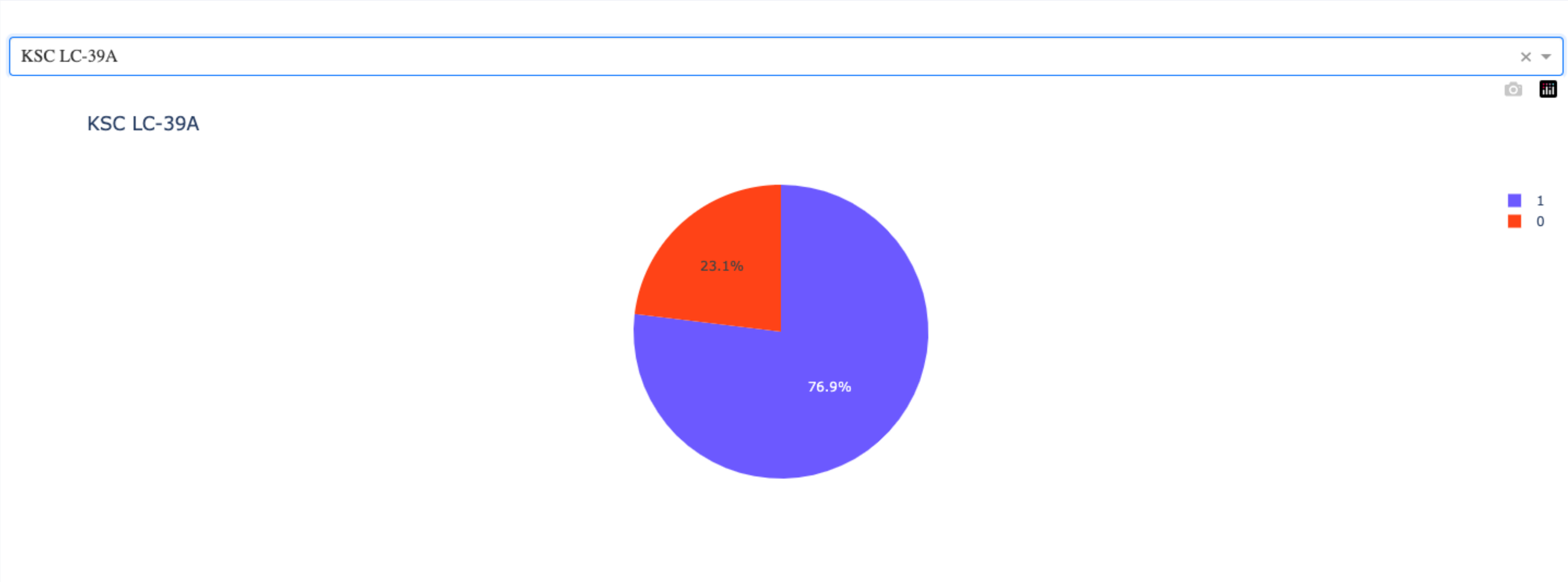
Dashboard Total success count for all sites

We visualize that the Launch Site KSC LC-39A has the most successful launches of all sites.
In contrast, the Launch Site CCAFS SLC-40 has the least successful launches of all sites



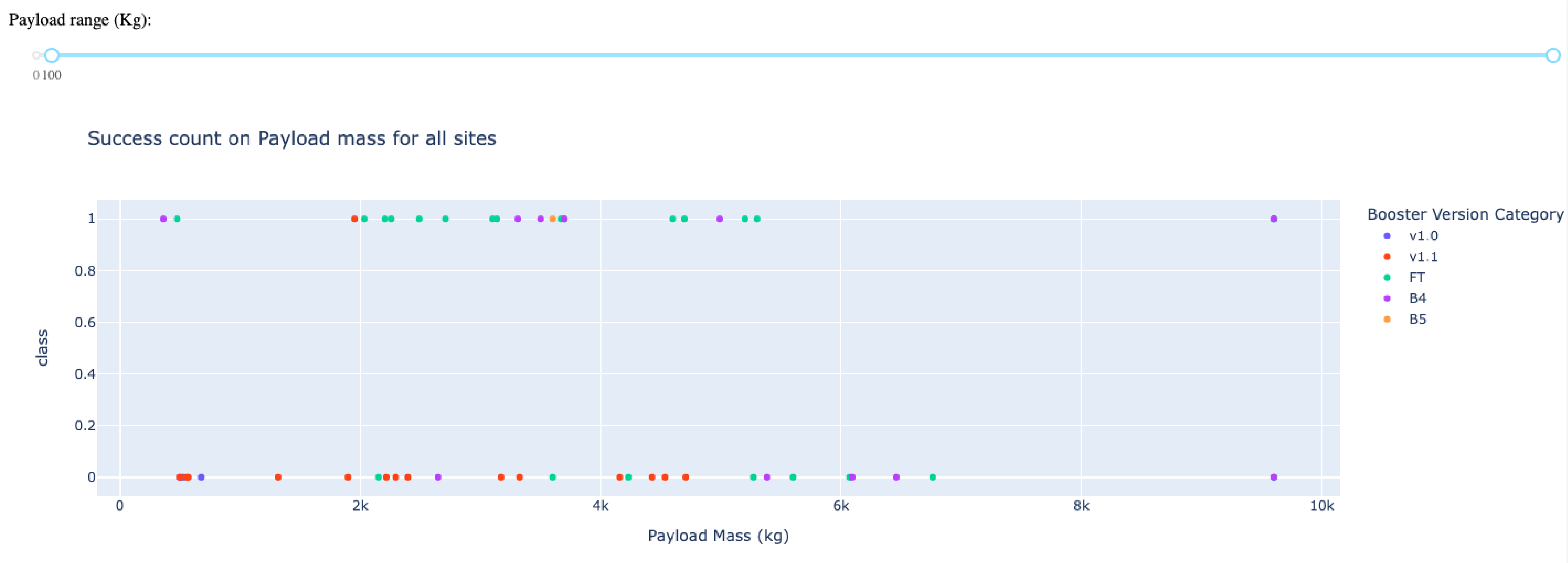
Dashboard Pie Chart of launch site **KSC LC-39**

- We display the success ratio of the KSC LC-39 Launch Site. We see almost 77% of the launches result in a successful launch.



Dashboard Payload vs Launch Outcome

We found that the most successful payload range is between 2k-6k. Also, the most successful booster version is FT.

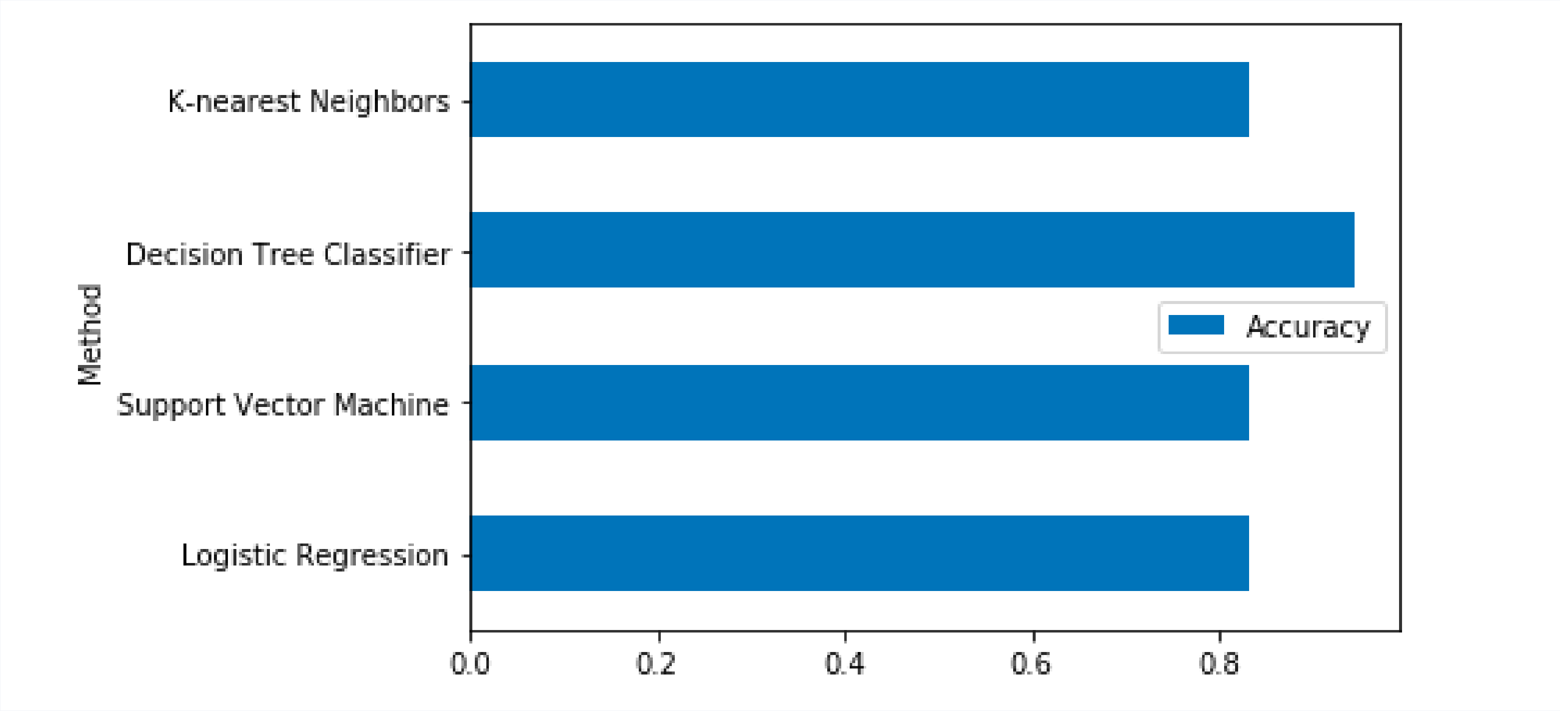


Section 5

Predictive Analysis (Classification)

Classification Accuracy

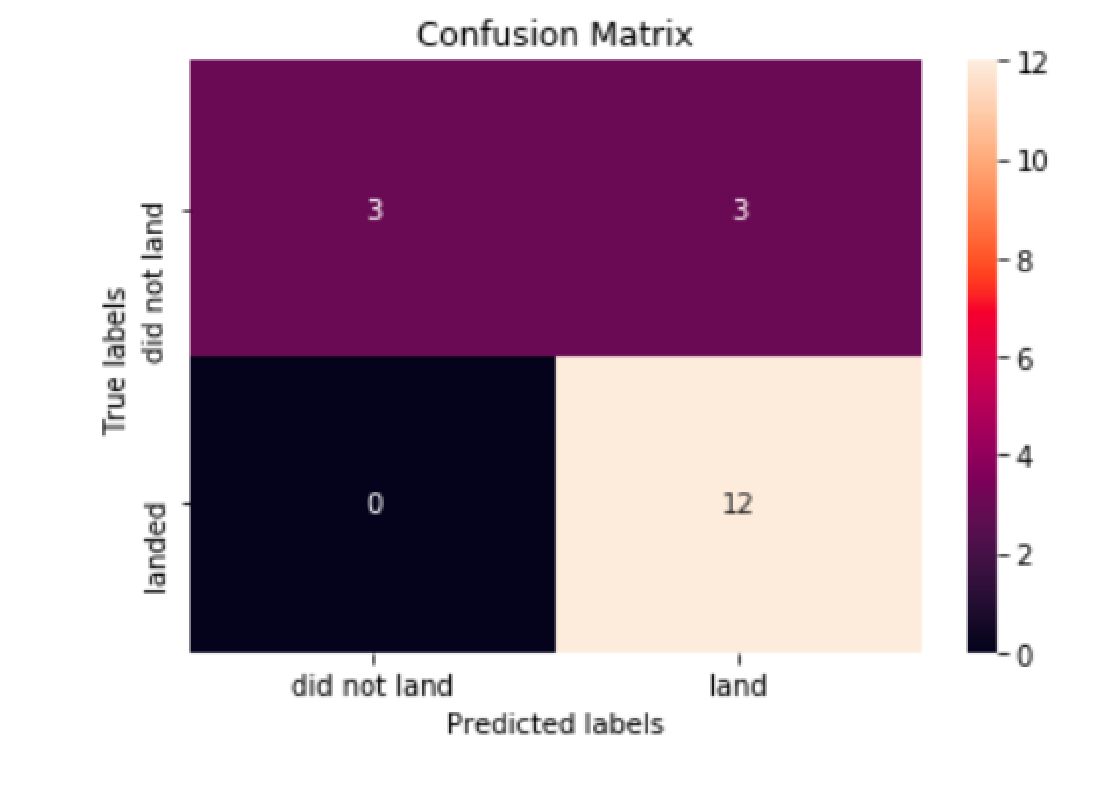
- We visualize the built model accuracy for all built classification models:



Confusion Matrix

- We show the confusion matrix of the best performing model:

- In the model, we find of 18 predicitions: 12 were correctly predicted in the land class, 3 were correctly pedicted in the did not land calss. Also, we have 3 false positivies



Conclusions

- SpaceX has launched rockets in 4 different locations
- The first success landing on ground pad was in 2015. Five years later than the first launch
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015: F9 v.1.1 B1012 and F9 v1.1. B1015
- Since 2013 the number of landing outcomes kept increasing
- Using Folium Maps, was possible to identify that launch sites use to be in safety places, near sea.
- Predictive Analysis showed that a Decision Tree Classifier algorithm is the best model to predict successful landings, having an accuracy over 87%.

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets are found in the GitHub project Falcon 9 Project

Thank you!