IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Takudzwa GUTU

15 Sep 2023

Outline

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

Executive Summary

The goal of this research is to study and analyze SpaceX Falcon 9 data, collected from various sources and employ Machine Learning models to predict the success of first stage landing.

Summary of methodologies

The research took on many methods to collect and analyze data, build and evaluate machine learning models, and make predictions. The methods and processes are listed below.

- Data collection through API and Web scraping.
- Data transformation and data wrangling.
- Exploratory data analysis with SQL.
- Data visualization to build maps using folium.
- Dashboarding with Plotly Dash.
- Machine learning model building.

Summary of all results

The results of this research will be from the visualisations, the data analysis and predictive model analysis.

Introduction

- Project background and context

The launch cost of space travel expeditions is a major decision factor for the participants in the industry. By reusing first stage capabilities, SpaceX may be ahead of it's competitors by lowering launch costs.

- Problems

Whether the first stage of SpaceX Falcon 9 will land successfully.

The effects of different variables such as launch site, booster version, payload mass on the landing outcomes.

Establishing correlations between the variables and success rate of launch.

Section 1

Methodology

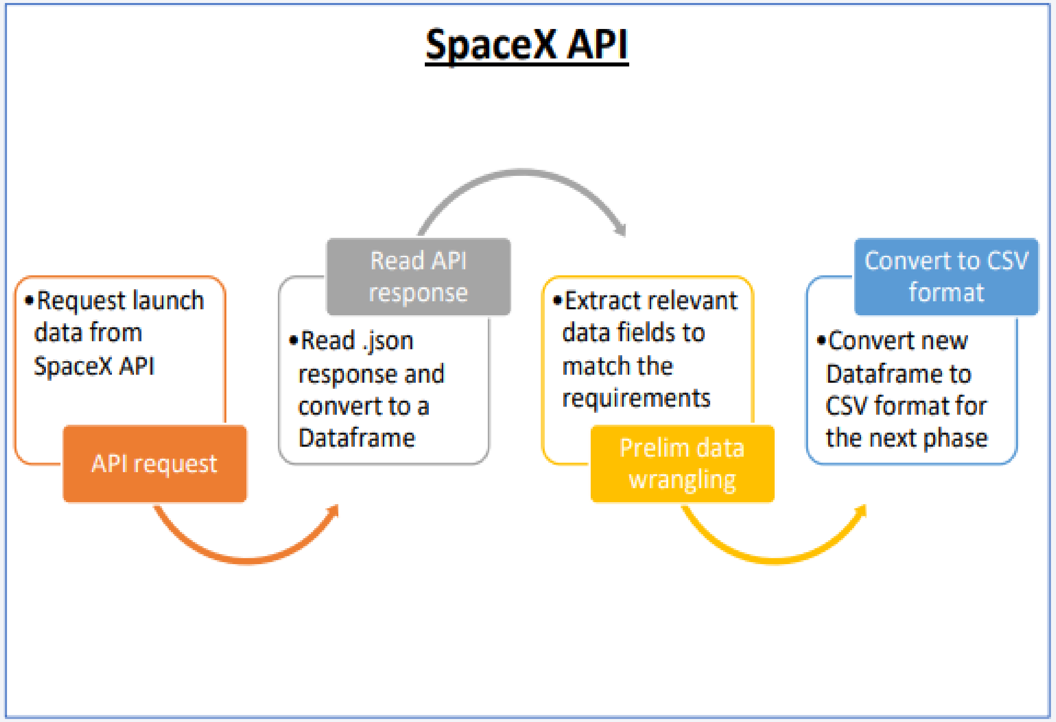
Methodology

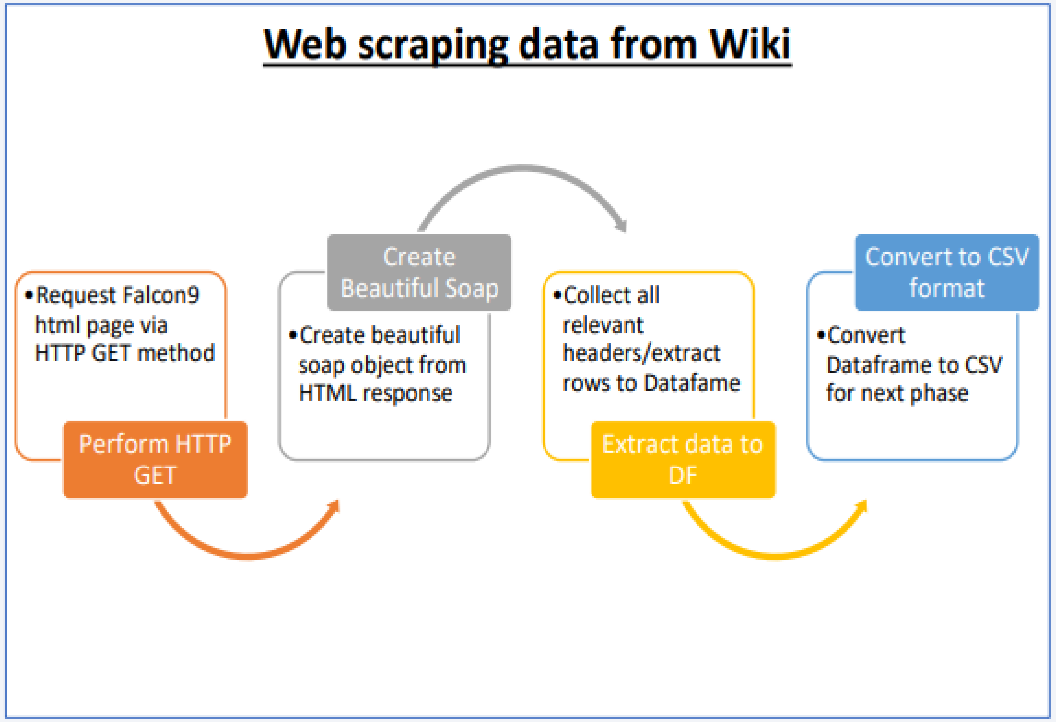
Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and by web scrapping Flacon 9 and Falcon Heavy launch records from Wikipedia.
- Perform data wrangling
  - Assigning labels for training the models by converting mission outcomes to labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Creating a “class” column.
  - standardize and transform the data.
  - Find a suitable classification algorithm.

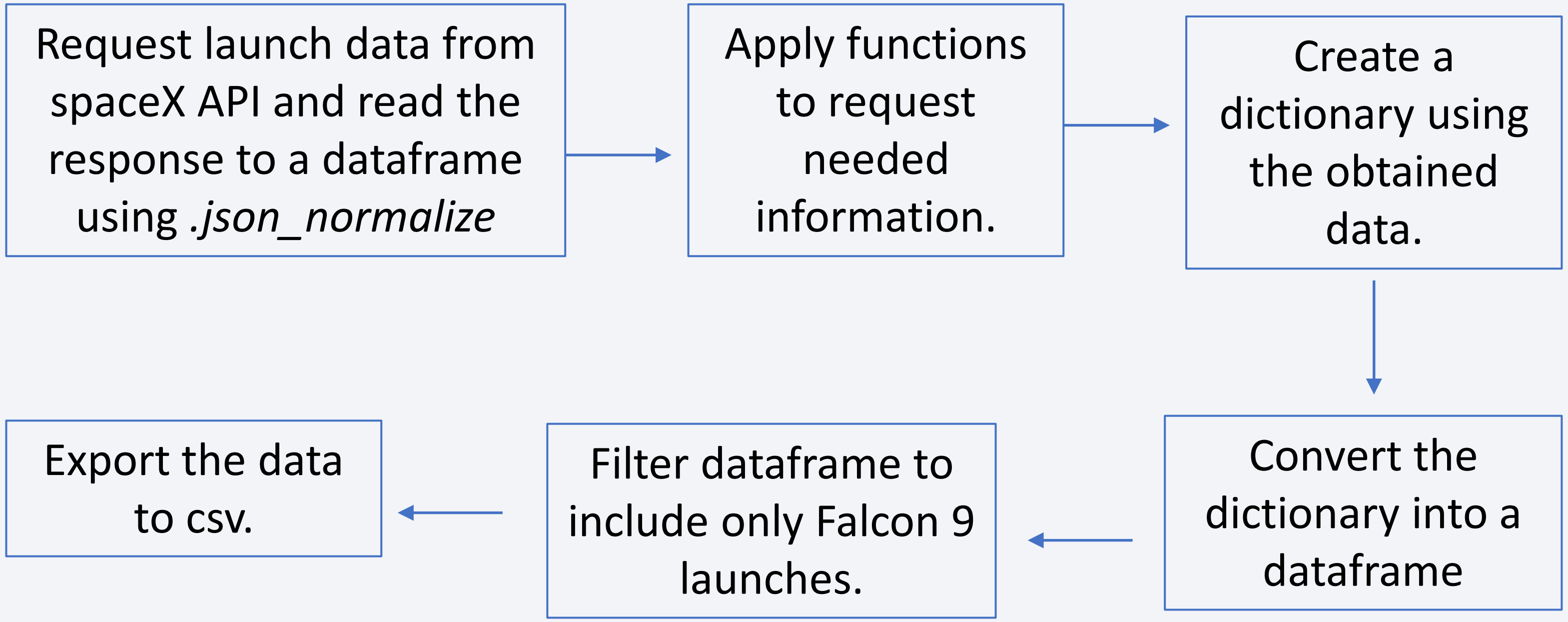
Data Collection

- Data was collected using SpaceX API and web scrapping relevant launch records from Wikipedia



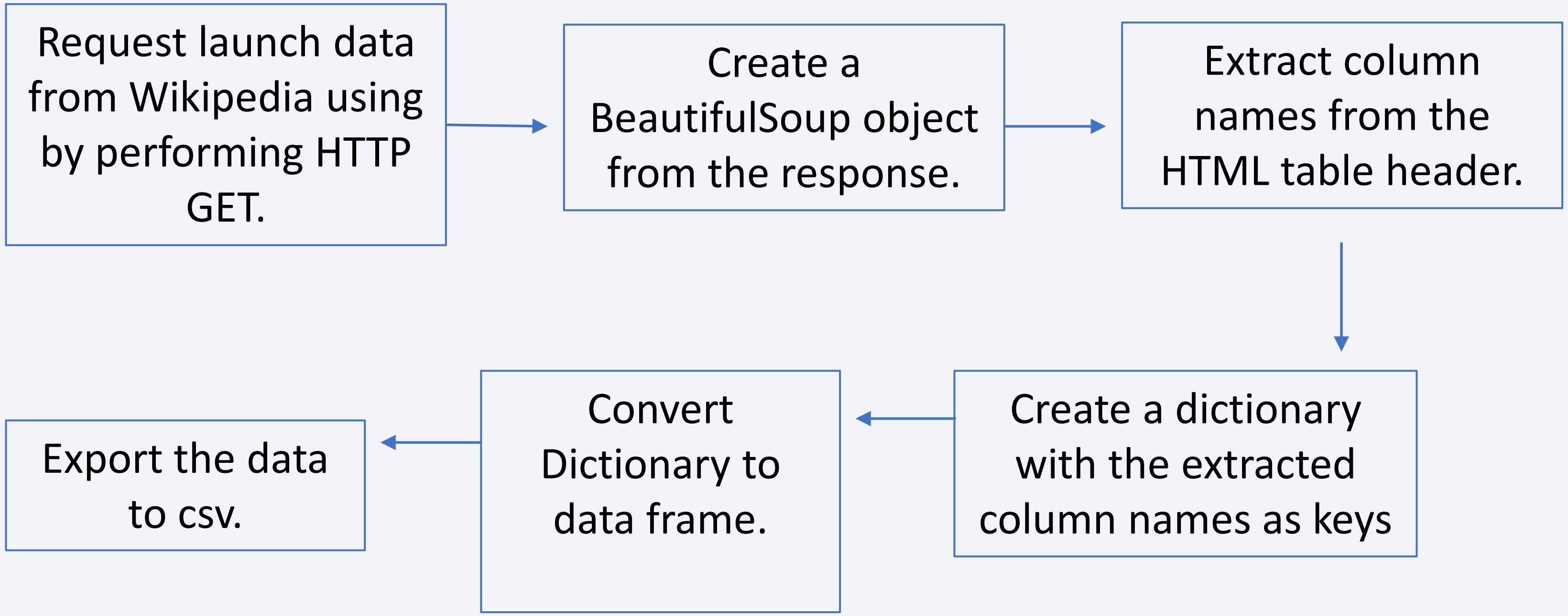


Data Collection – SpaceX API



GitHub: [Data Collection – SpaceX API]

Data Collection - Scraping



GitHub: [Data Collection – Scraping]

Data Wrangling

Exploratory data analysis was carried out to find patterns and help make labels for training the models.

For training, a value of 1 was given to each outcome where the booster successfully landed and 0 was assigned to each outcome where the booster did not land successfully.

Other labels that were assigned are:

- True Ocean: for a mission outcome which successfully landed in a specific region of the ocean.
- False Ocean: for a mission outcome which unsuccessfully landed in a specific region of the ocean
- False RTLS: for a mission outcome which unsuccessfully landed on a ground pad.
- True ASDS: for a mission outcome which successfully landed on a drone ship.
- False ASDS for a mission outcome which unsuccessfully landed on a drone ship.

GitHub: Data Wrangling

EDA with Data Visualization

Scatter plots show a relationship between two variables. The following charts were plotted as scatter plots: Flight number vs launch site, Payload vs Launch site, Flight number vs Orbit type, Payload vs Orbit type.

Bar charts are used to compare categorical variables. A bar chart of the success rate of each orbit type was plotted.

Line charts are used to show and track changes in data over a specified period of time. The year trend of the average launch success rate was shown through a line chart

GitHub: [EDA with Data Visualisation]

EDA with SQL

SQL queries were performed to do the following:

- Display 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. Use a subquery.
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- 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

GitHub: [EDA with SQL]

Build an Interactive Map with Folium

Map Objects that were created:

Markers of all launch sites using circles, pop up label and text label. This showed the launch sites on the map.

Coloured markers of the launch outcomes for each launch site. Success was shown in green and failures were shown in red.

Calculated distances between a launch site and it's proximities such as a highway, coastline using coloured lines for neater visualization.

GitHub: [Build an Interactive Map with Folium]

Build a Dashboard with Plotly Dash

The following plots and interactions were added to the dashboard:

A launch site dropdown list. This was to allow the option to choose a desired launch site.

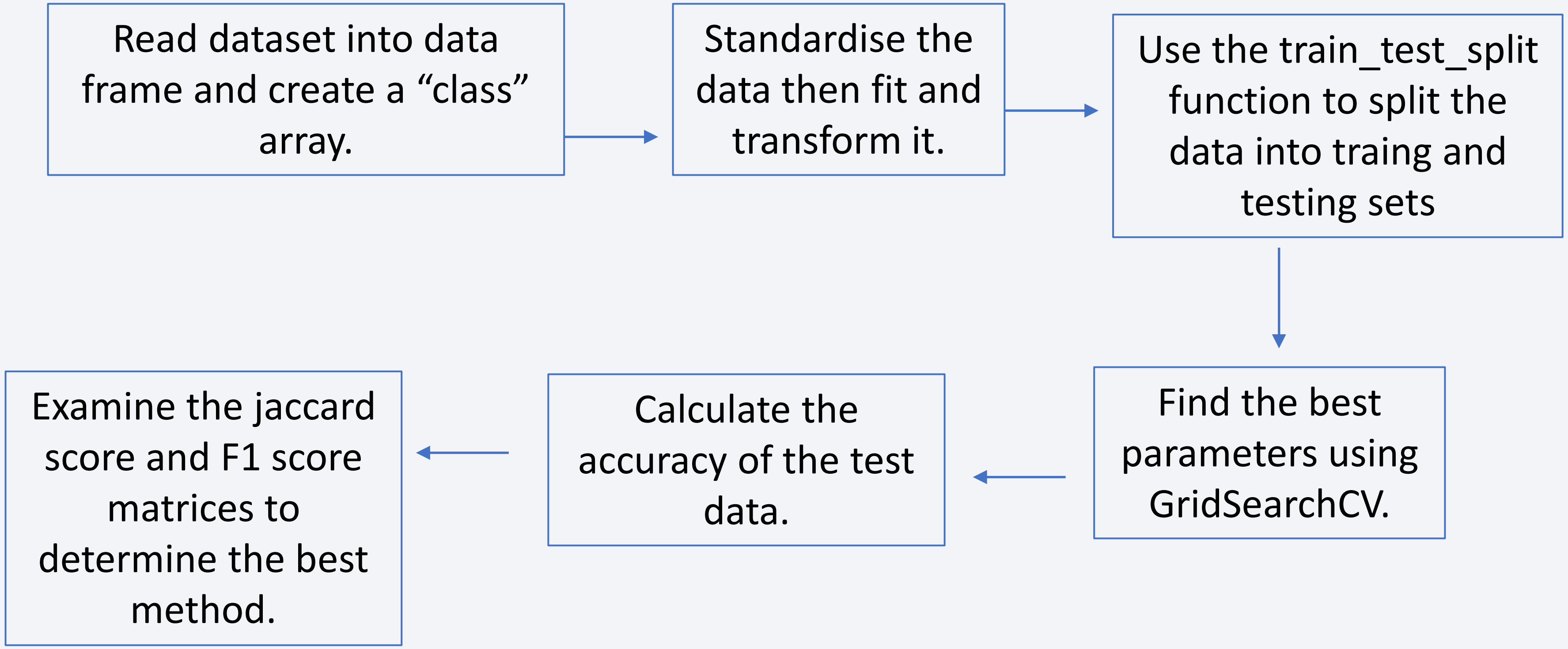
A pie chart to visualize the successful launches compared to failed counts of all launches.

A slider to select the payload range.

A scatter chart of payload mass vs success rate.

GitHub: [Build a Dashboard with Ploty Dash]

Predictive Analysis (Classification)



GitHub: [Predictive Analysis (Classification)]

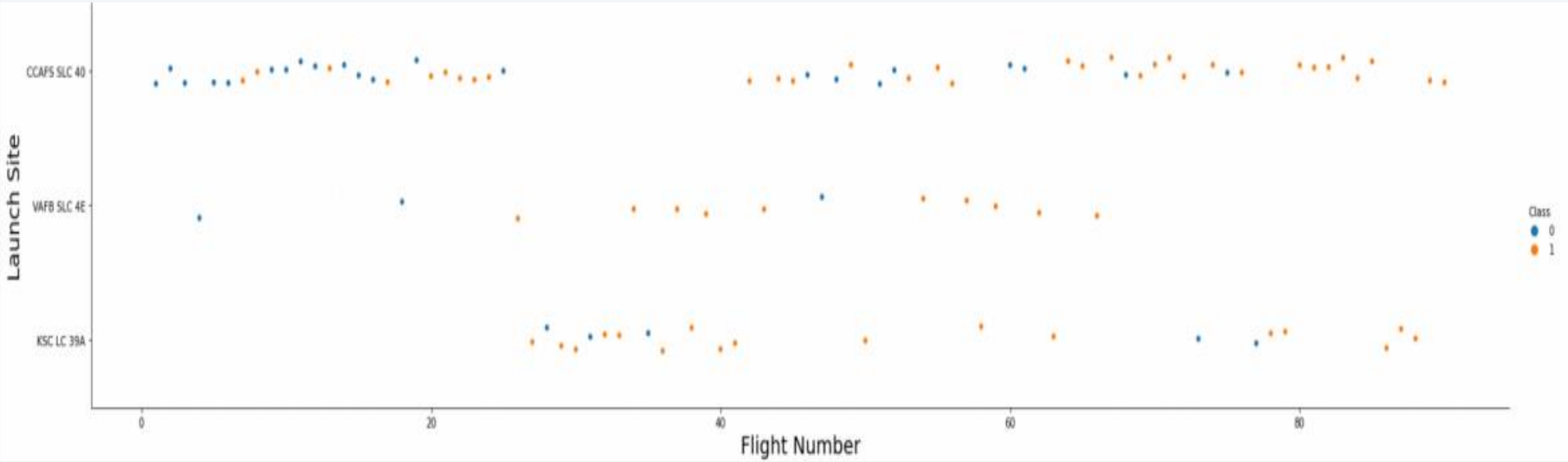
Results

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

Section 2

Insights drawn from EDA

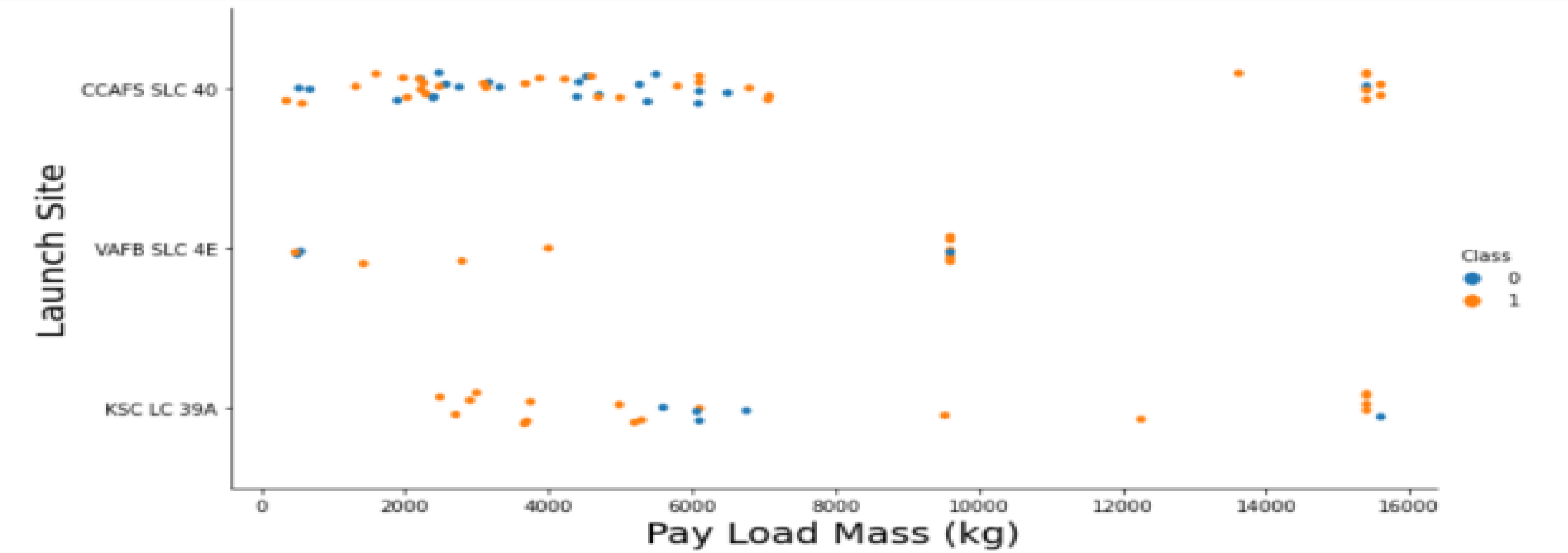
Flight Number vs. Launch Site



Class 0 represents unsuccessful flights. Class 1 represents successful fights.

The trend shown is that the success rate increases as the flight numbers increase.

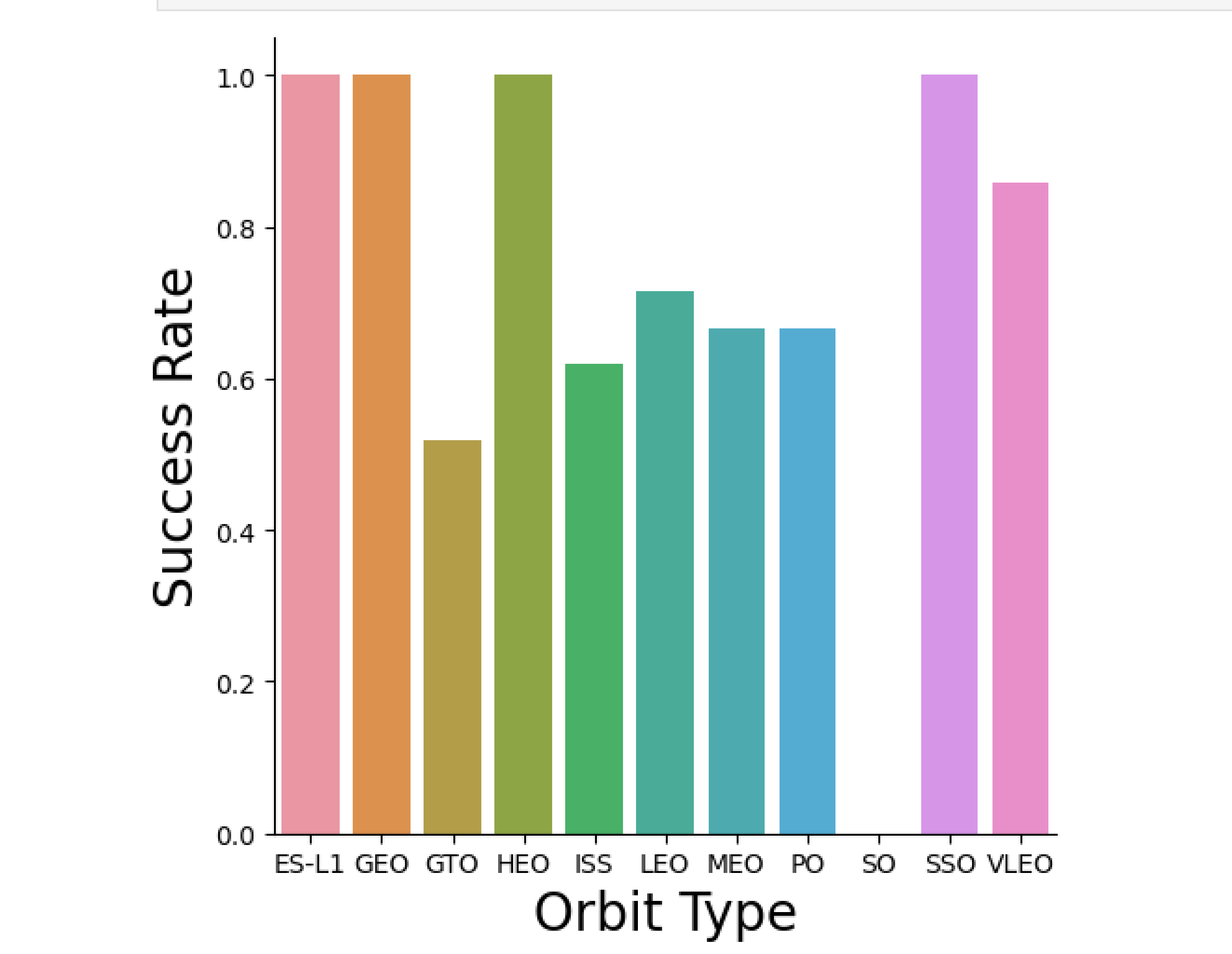
Payload vs. Launch Site



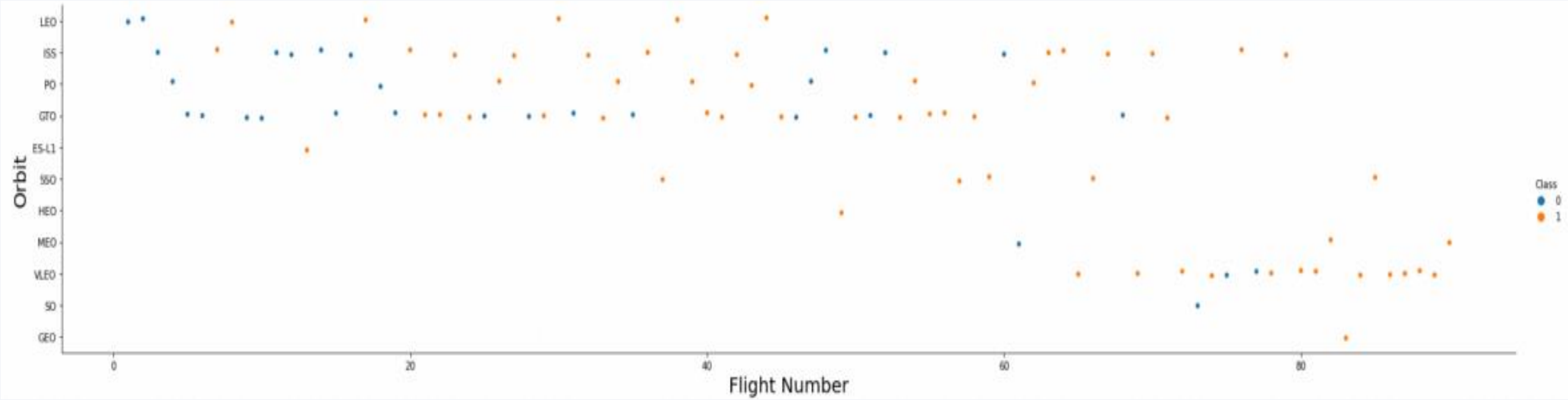
Class 0 represents unsuccessful flights. Class 1 represents successful fights.

KSC LC 39A flights with a payload mass lower than 5500kg were all successful.

Success Rate vs. Orbit Type



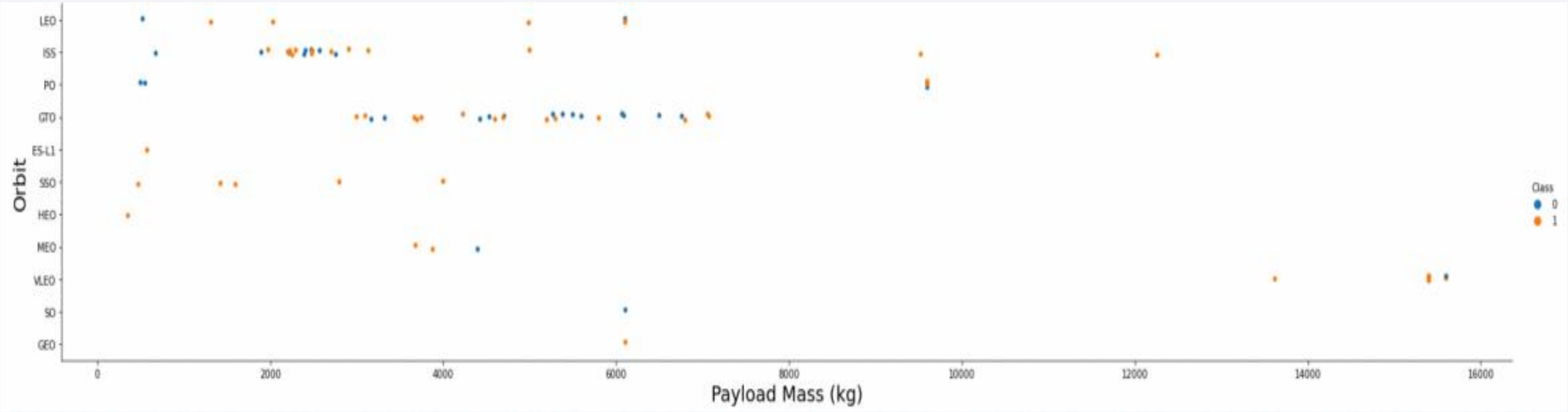
Flight Number vs. Orbit Type



Class 0 represents unsuccessful flights. Class 1 represents successful fights.

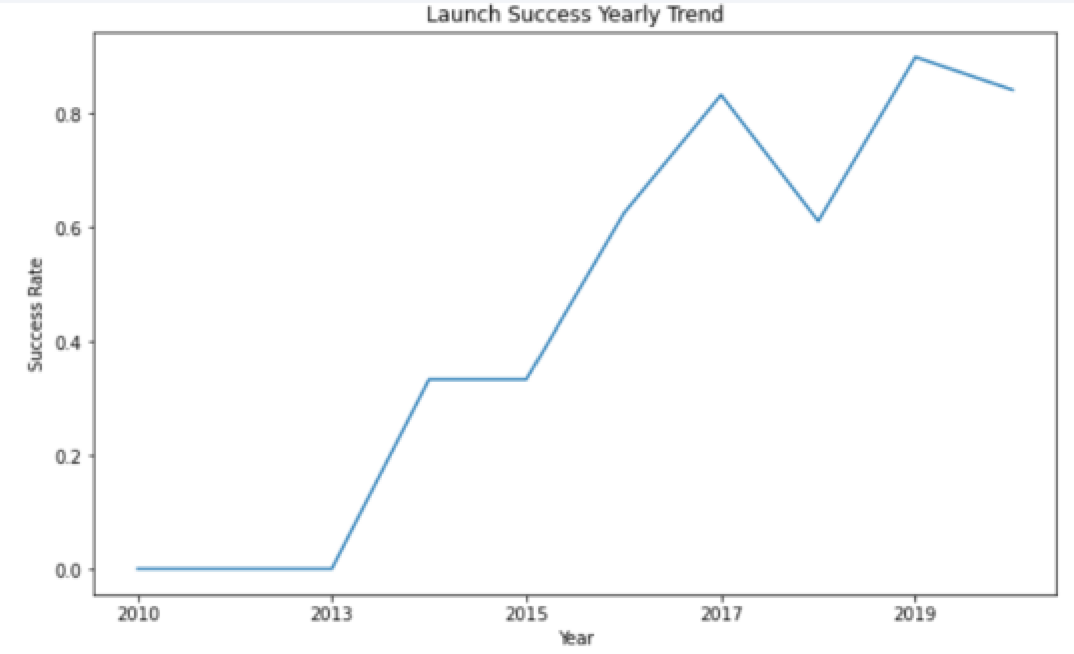
The trend in general shows that success rate increases with the number of flights, particularly with LEO orbit.

Payload vs. Orbit Type



Class 0 represents unsuccessful flights. Class 1 represents successful fights.

Launch Success Yearly Trend



The average yearly launch success rate is shown to be flate from 2010 to 2013 and between 2014 to 2015..

It began to increase in 2013 to 2017. a decrese is observed in from 2017 before a continuation of the uptrend in 2018

All Launch Site Names

Display the names of the unique launch sites in the space mission

[9]:

```
%%sql

select distinct Launch_Site from spacextbl
```

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

[9]:

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

The “distinct” function allows to select and list only the unique sites without repetition.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

[10]:
```
%%sql

select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
```

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

[10]:

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

- The “limit” enables to display only 5 names.

Total Payload Mass

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

```
[11]: %%sql

select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

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

[11]:

sum(PAYLOAD_MASS__KG_)
45596

The “sum” function sums up the selected data.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
[12]: %%sql

select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
```

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

[12]:

avg(PAYLOAD_MASS__KG_)
2928.4

The “avg” function calculates the average of the selected data

First Successful Ground Landing Date

```
[19]: %%sql

select min(Date) as min_date from spacextbl where Landing_Outcome = 'Success (ground pad)';

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

min_date
2015-12-22

The dates will be ordered hence the need for the "min" function. This function will select the lowest (earliest) date.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[ ]: %%sql

select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing__Outcome = 'Success (drone ship)');
```

There are 2 conditions specified after the “where” clause. The first is for mass greater than 4000, the second is for mass less than 6000. The use of the “and” operator returns data from when both conditions are met or true.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

[14]:
```
%%sql

select Mission_Outcome, count(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
```

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

[14]:

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

The “count” function allowed for counting of each instance of a mission outcome.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[20]:
```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl)
```

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

[20]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

[21]:
```
%%sql

select Landing_Outcome, Booster_Version, Launch_Site from spacextbl where Landing_Outcome = 'Failure (drone ship)'
```

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

[21]:

Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

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.

[22]:
```
%%sql

select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

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

[22]:

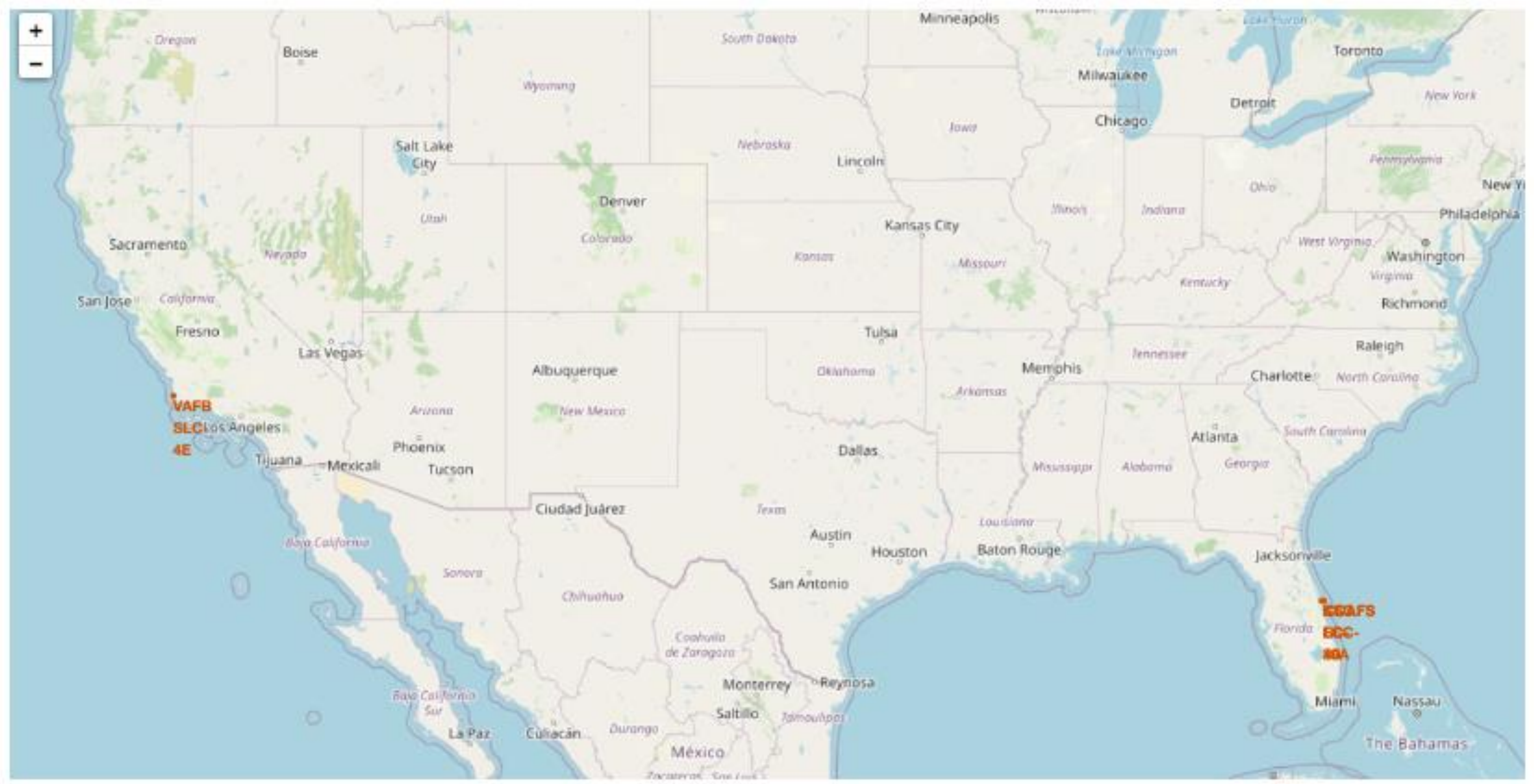
Landing_Outcome	LandingCounts
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Landing outcomes listed in descending order by making use of the “Order by” function

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

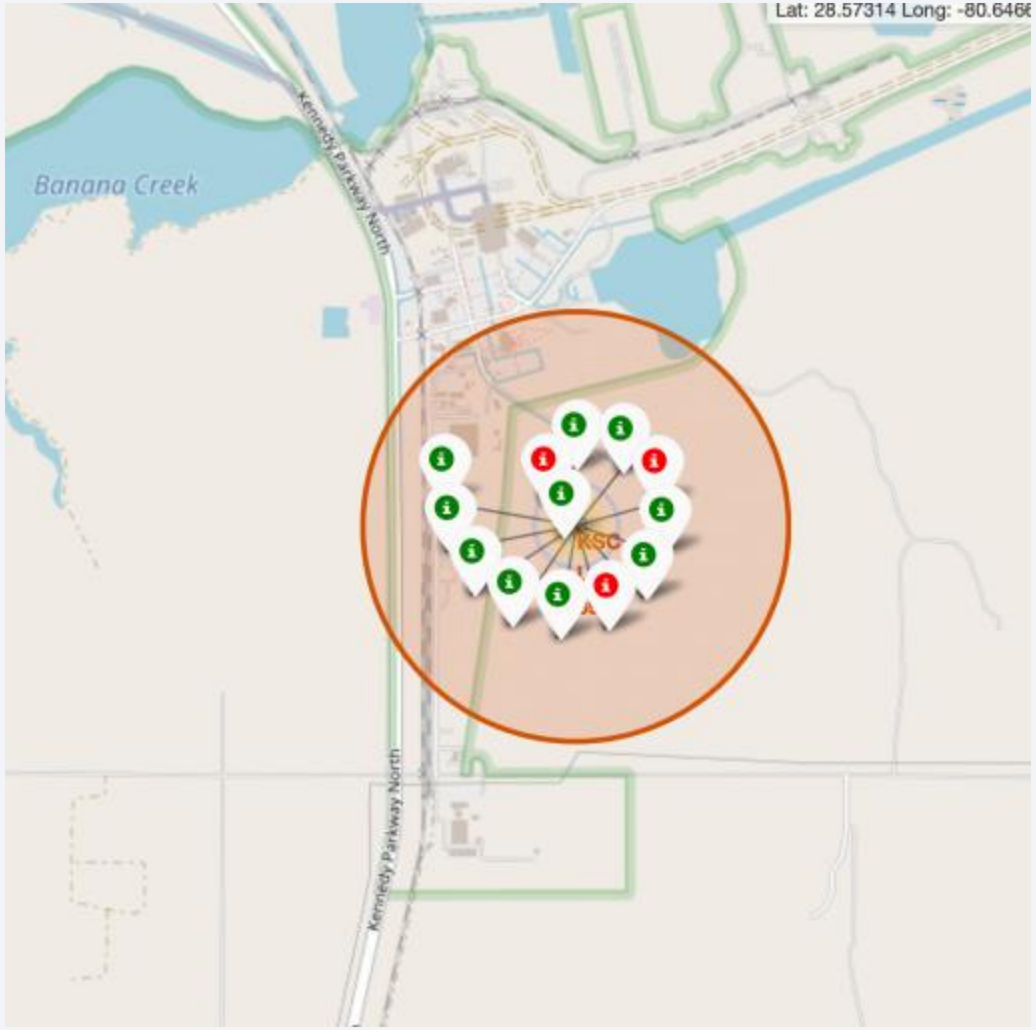


All launch site locations are near a coastline and close to the sea

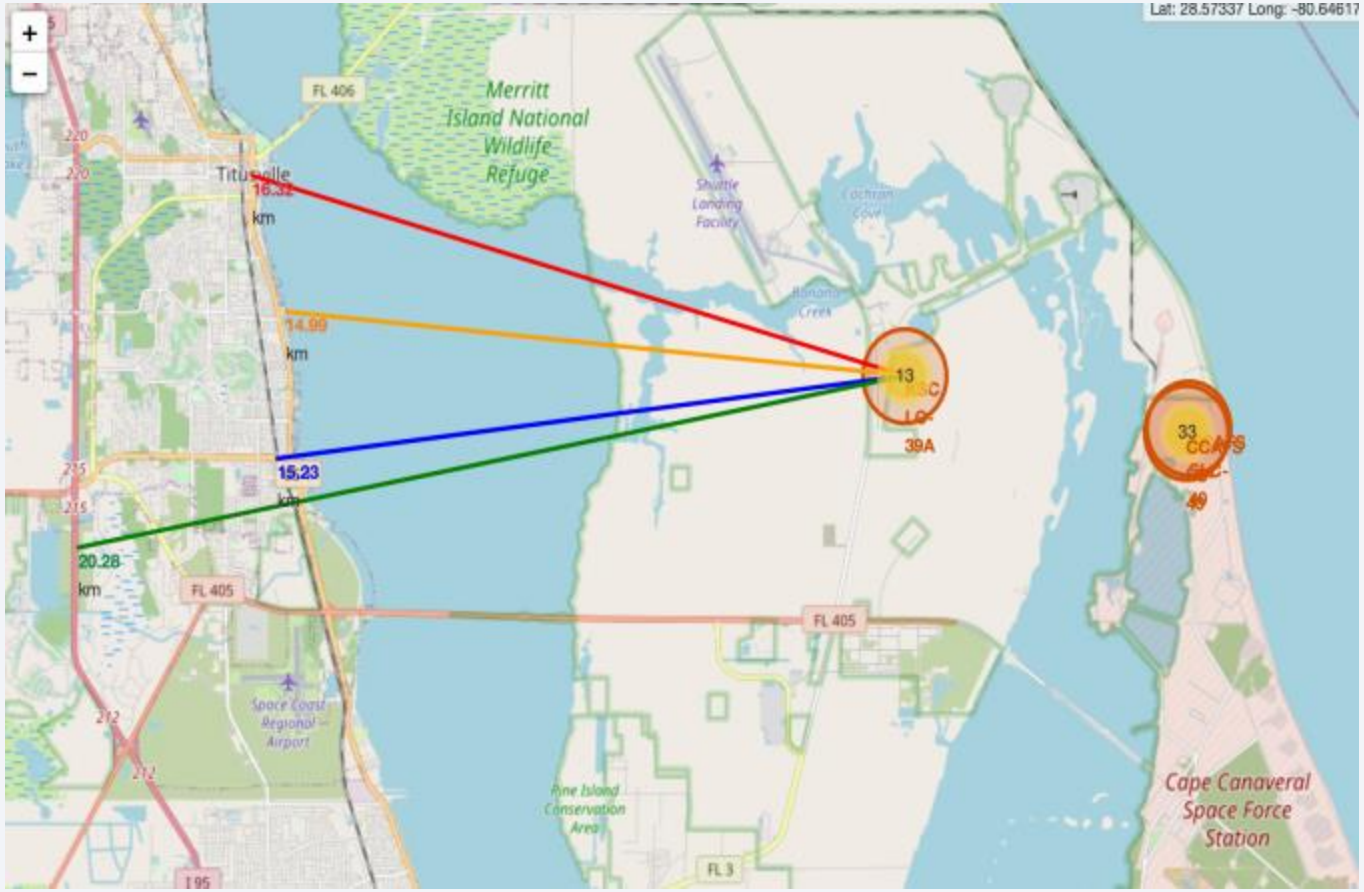
Launch Outcomes at site KSC LC -39 A

Green markers are for successful launches.

Red markers are for failures.



Distance from KSC LC- 39 A to proximities

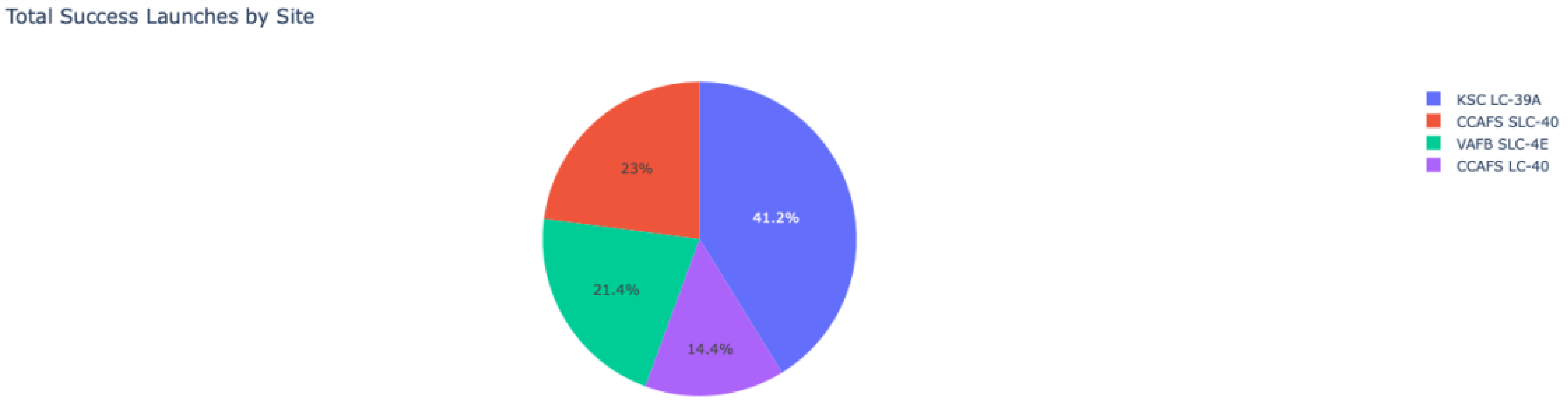


It can be seen that site KSC LC - 39 A is a15.23km from the railway

Section 4

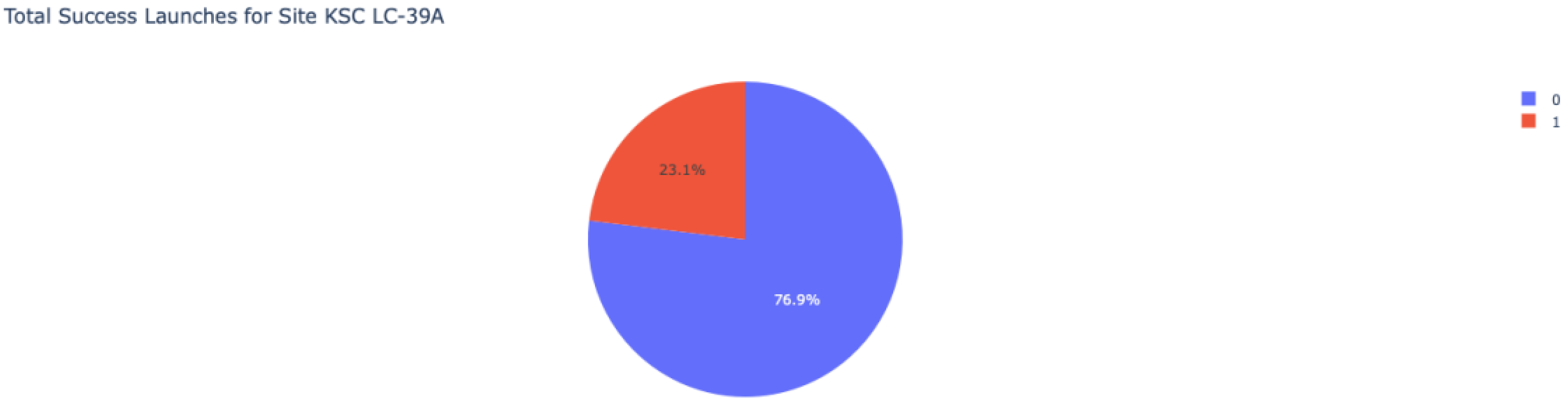
Build a Dashboard with Plotly Dash

Launch success count for all sites



Site KSC LC-39A has the most successful launches.

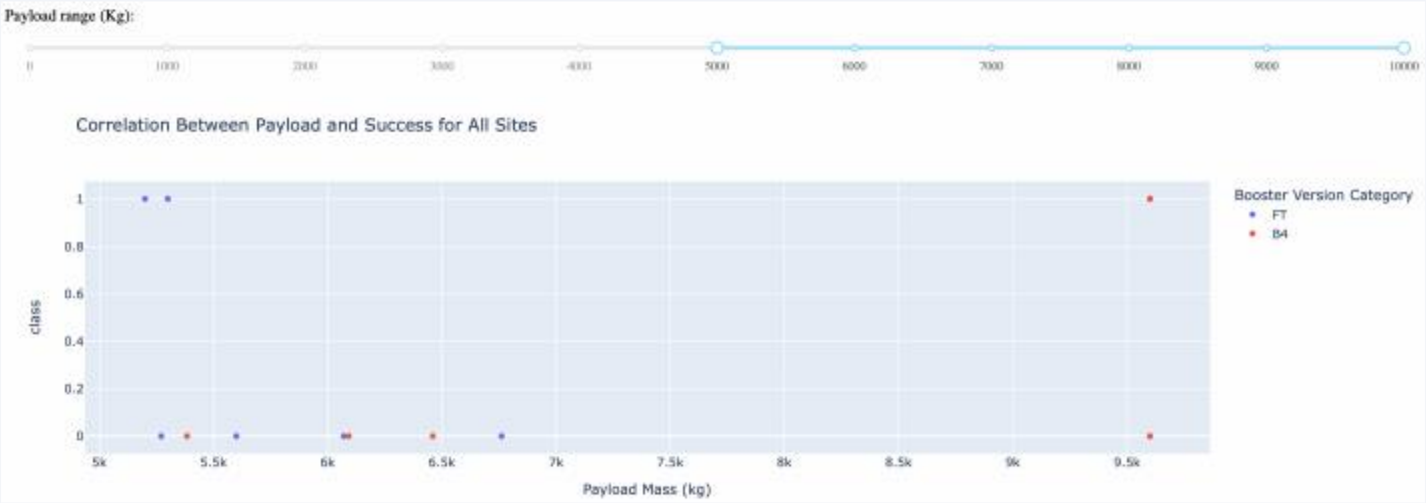
Launch site with highest launch success ratio

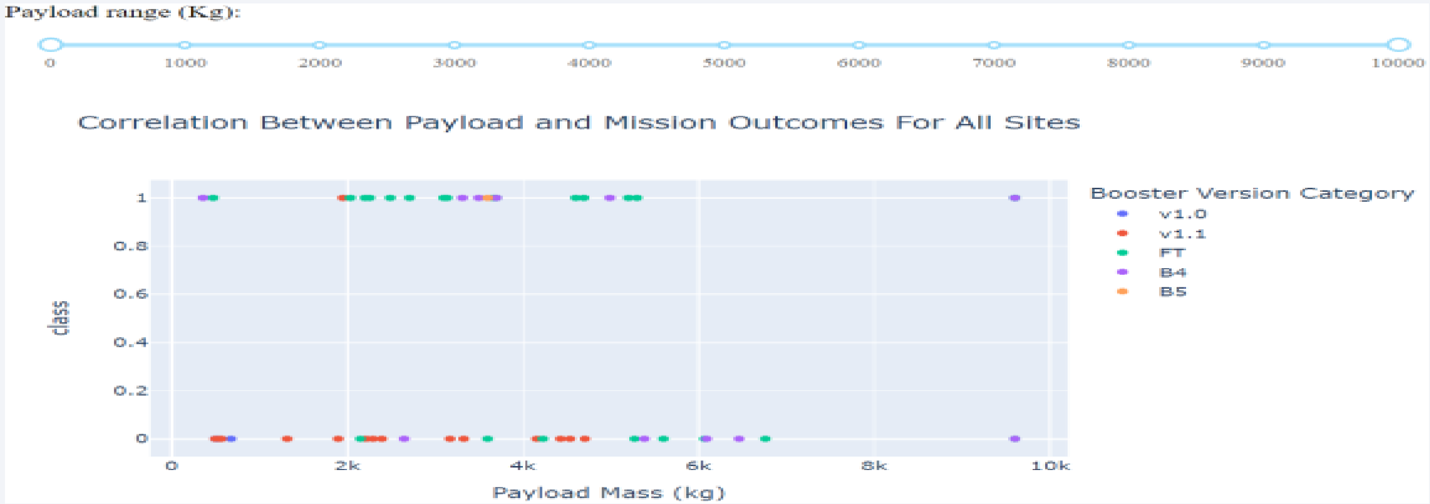


Site KSC LC -39 A has the highest launch success rate.

The ratio of successful launches is 76.9%

Payload vs. Launch Outcomes



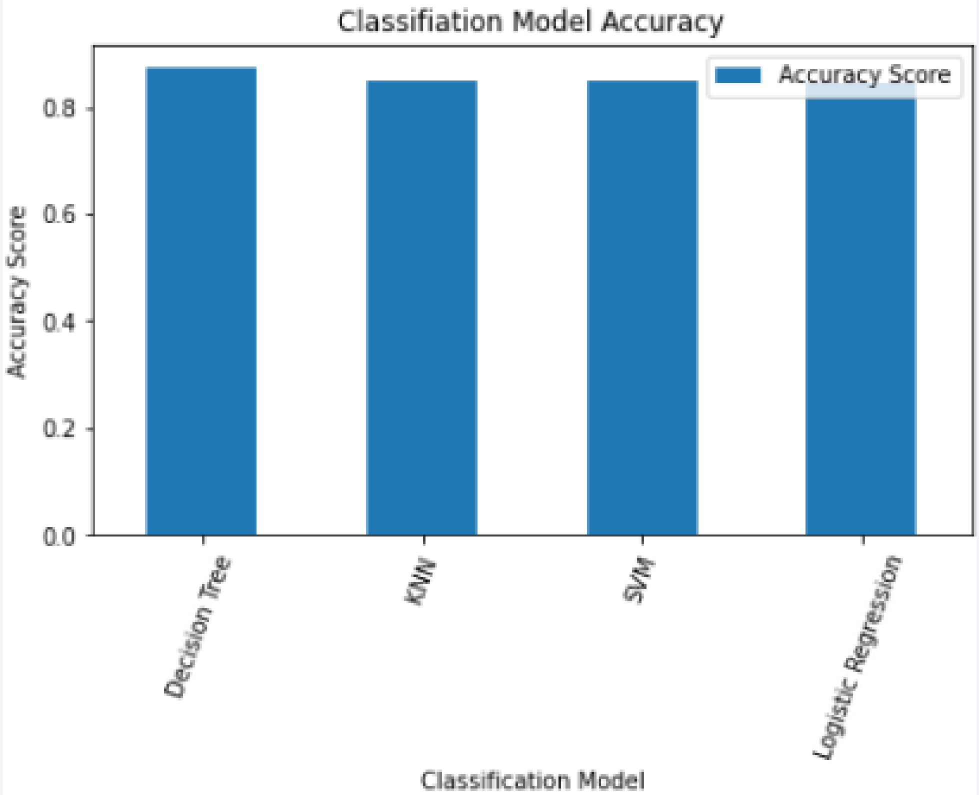


- Booster version FT had success at payloads of around 5000kg.

Section 5

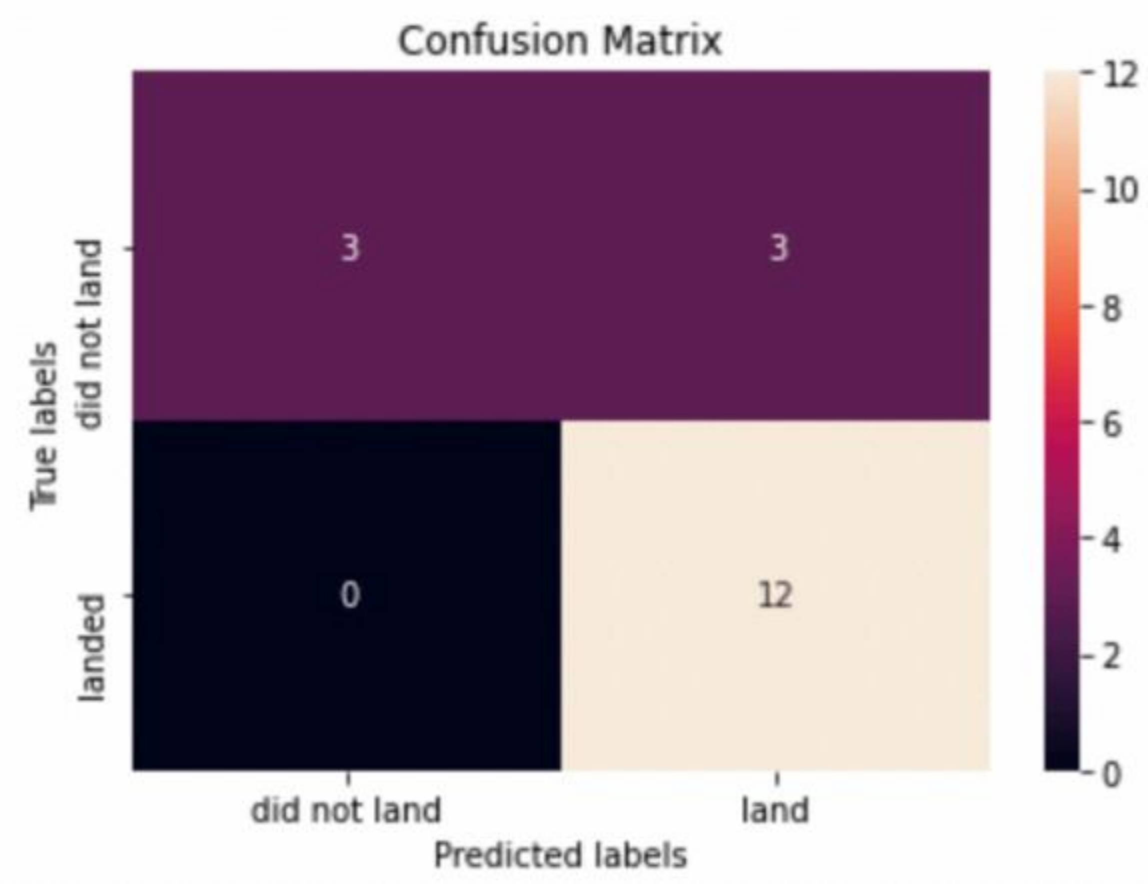
Predictive Analysis (Classification)

Classification Accuracy



Decision tree model has the highest classification accuracy.

Confusion Matrix



We see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- As the number of launches increases the numbers of successful landings also increases.
- Launches with a lower payload tend to be more successful than those with a higher payload.
- The average yearly success rate is increasing over the years.
- Site KSC LC -39A has the highest success rate.
- Decision tree model is the best algorithm for the data set

Appendix

- Relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that were created during this project are available in the sections that discuss them.

Thank you!