IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tetteh Kwadjo
25/05/2023

Outline

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

Executive Summary

- **Summary of methodologies**

In this capstone project we are looking to predict using various machine learning algorithms if the SpaceX Falcon 9 first stage booster rockets will land successfully.

Successful landing and recovery depends on many factors for example orbit, payload mass, booster versions, launch sites etc.

Methods and Steps to take

- Data collection via Api, web scrapping and Data wrangling
- Exploratory Data Analysis with Data Visualization and Exploratory Data Analysis with SQL
- Interactive Map with Folium and Dashboard with Plotly Dash
- Predictive analysis or classification

- **Summary of all results**

- Exploratory Data Analysis results
- Interactive maps and dashboard
- Predictive analysis results

Introduction

- **Project background and context**

Space X 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 Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. Using Data, public information and machine learning models, we are going to predict the success or failure of the first stage

- **Problems you want to find answers**
- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- What conditions allow for the highest success rate
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data and dropping unnecessary columns
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

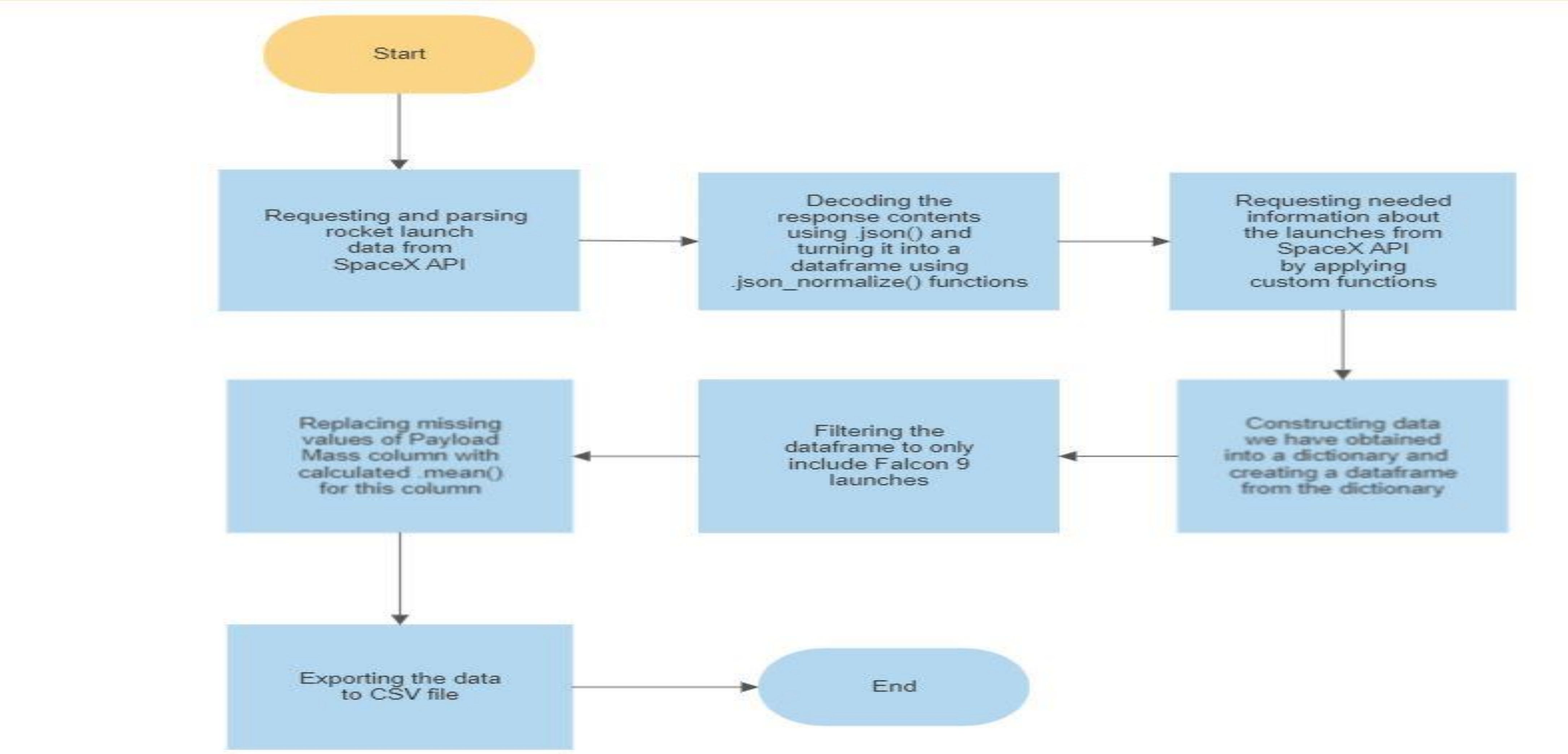
Data was collected through a process involving a combination of API requests from SpaceX REST API i.e. https://api.spacexdata.com/v4/rockets/ and Web Scraping data from a table in SpaceX's Wikipedia entry i.e. https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 .

We Use these data collection methods in order to get more complete

information about the launches for a more detailed analysis.

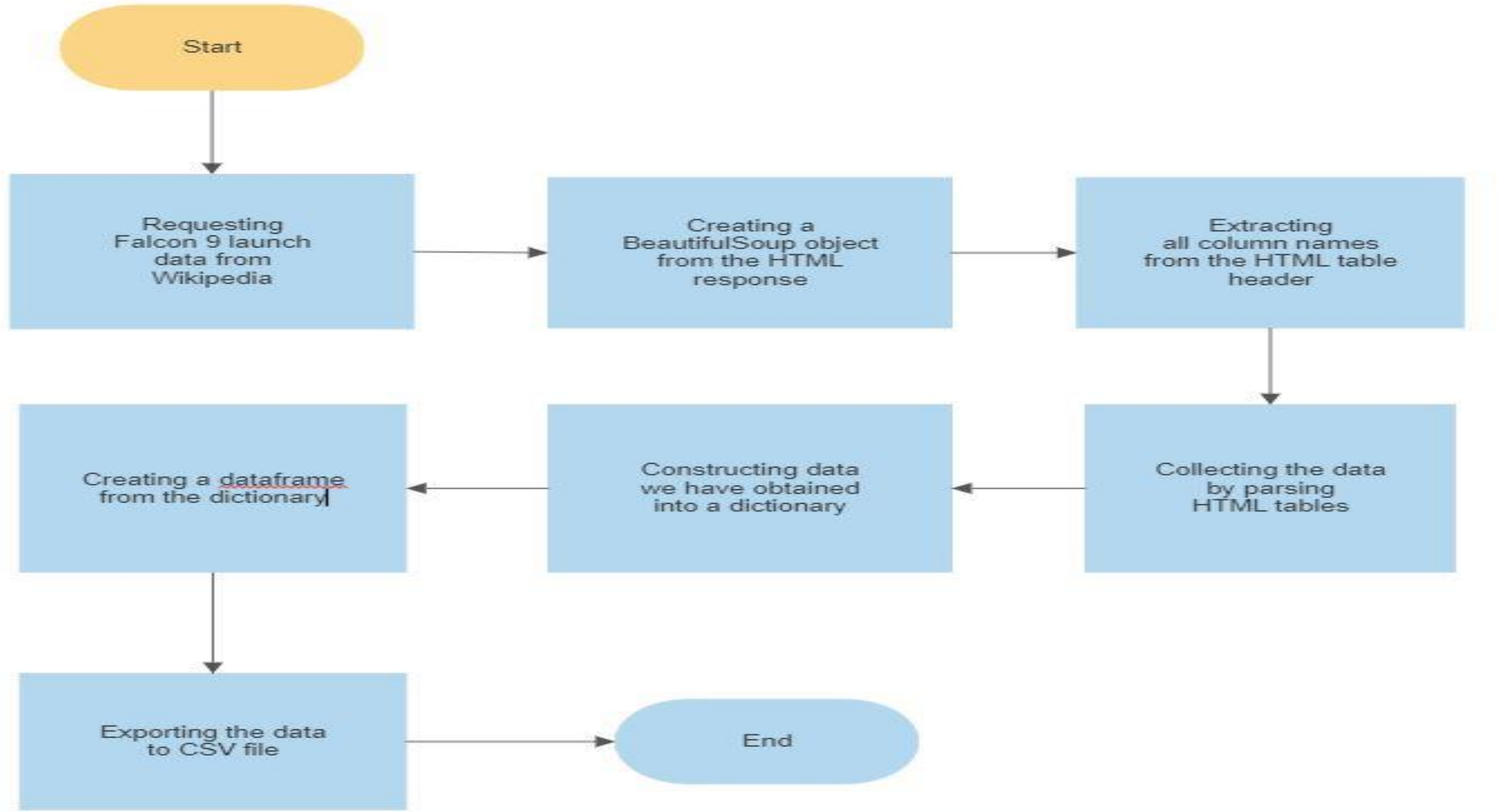
- Data Columns are obtained by using SpaceX REST API:FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



[Github Data collection API Link]

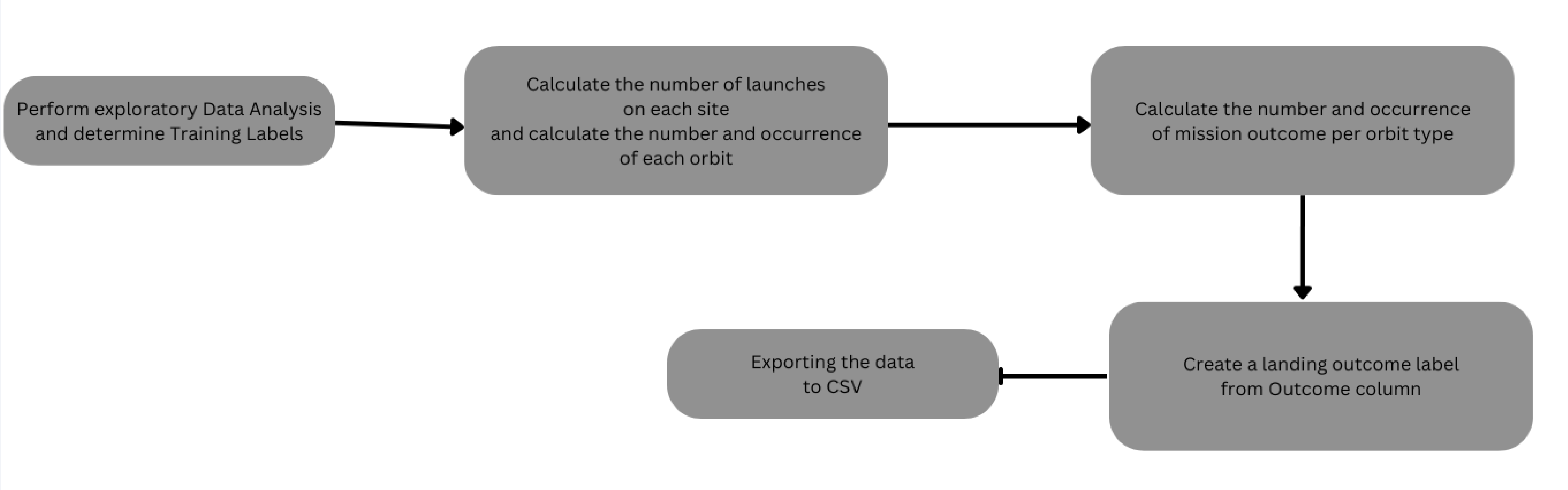
Data Collection - Scraping



[Github data collection and web scrapping link]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident, True Ocean, True RTLS, True ASDS means the mission has been successful. False Ocean, False RTLS, False ASDS means the mission was a failure. We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure. [Github wrangling Link]



EDA with Data Visualization

- **Charts that were plotted:**
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type

- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Success rate vs Orbit type

- Line charts show trends in data over time (time series)

Success rate vs year

EDA with Data Visualization Link

EDA with SQL

Performed the following SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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
- EDA with SQL Link

Build an Interactive Map with Folium

- **Markers of all Launch Sites:**

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

-Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

- **Coloured Markers of the launch outcomes for each Launch Site:**

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- **Distances between a Launch Site to its proximities:**

- 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
- Interactive Folium Map

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart 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.

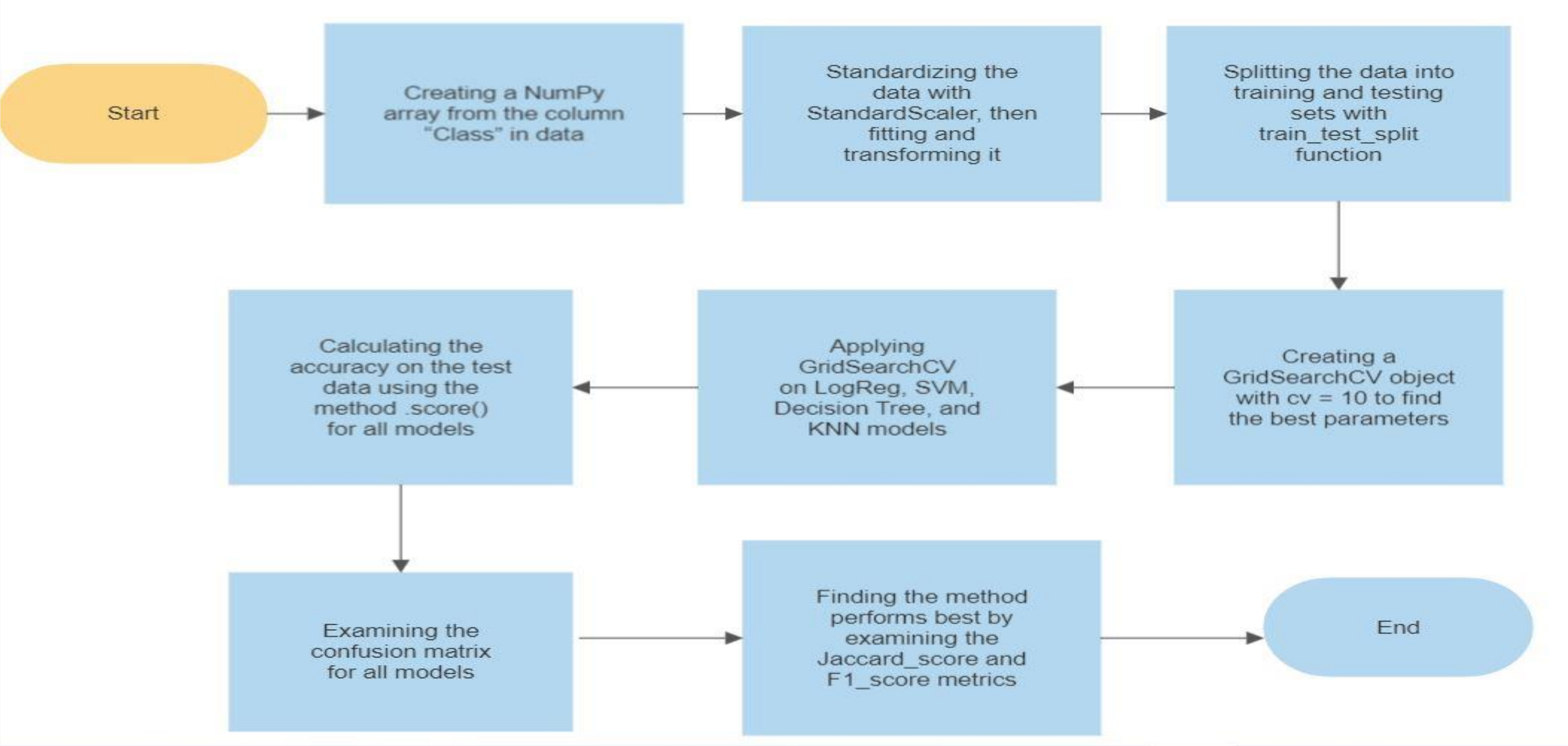
Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.
- Plotly Dash App

Predictive Analysis (Classification)



[Machine Learning Prediction Link]

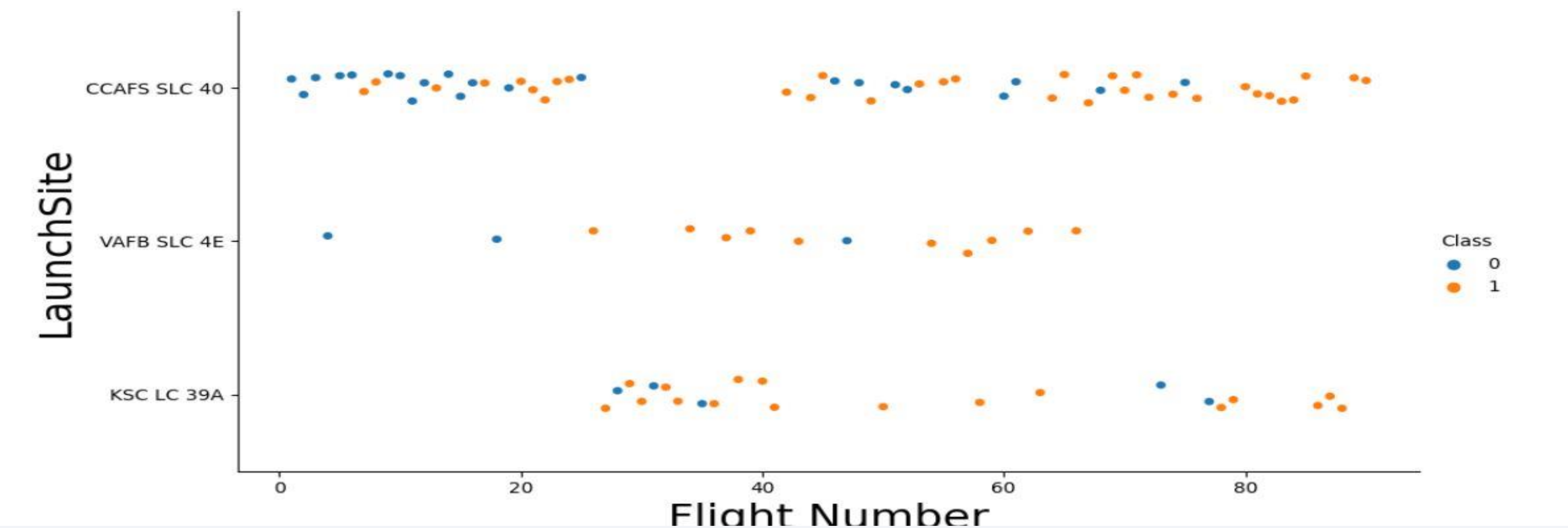
Results

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

Section 2

Insights drawn from EDA

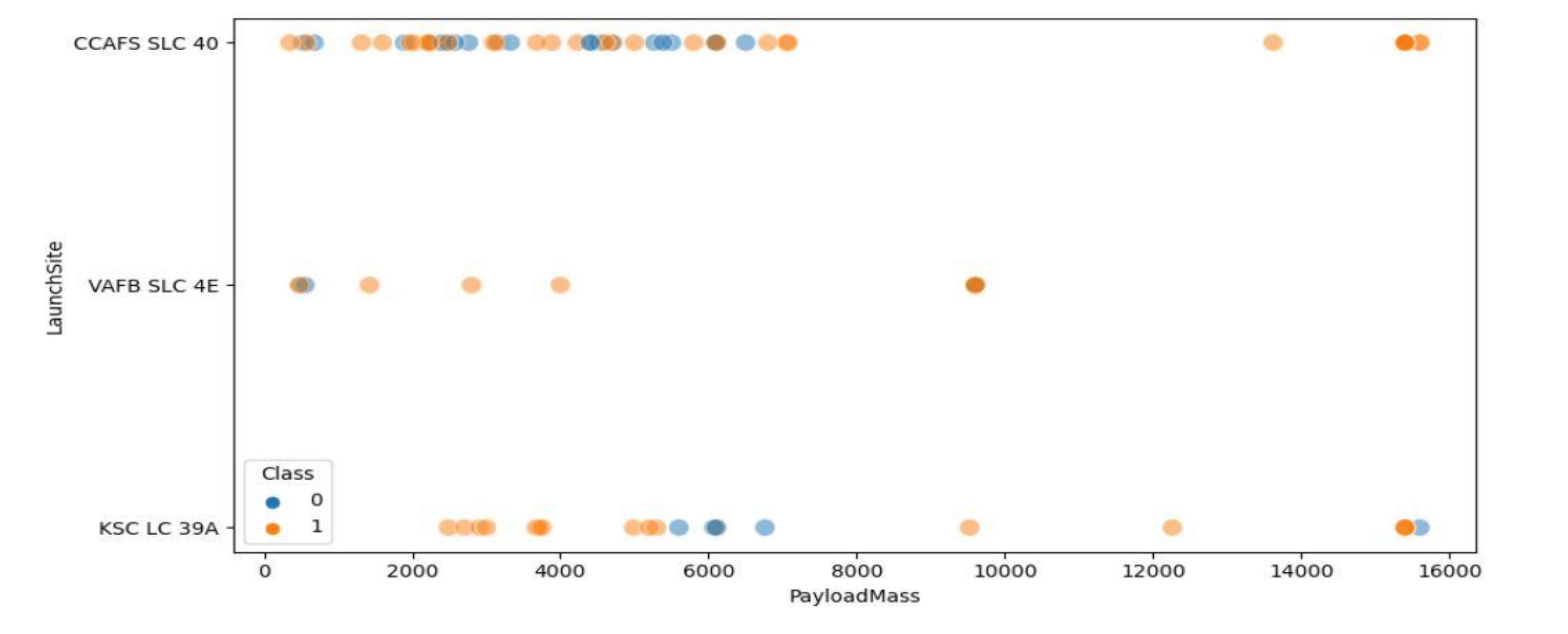
Flight Number vs. Launch Site



Explanation:

- It can be assumed that each new launch has a higher rate of success.
- We observe that, for each site, the success rate is increasing with flight number

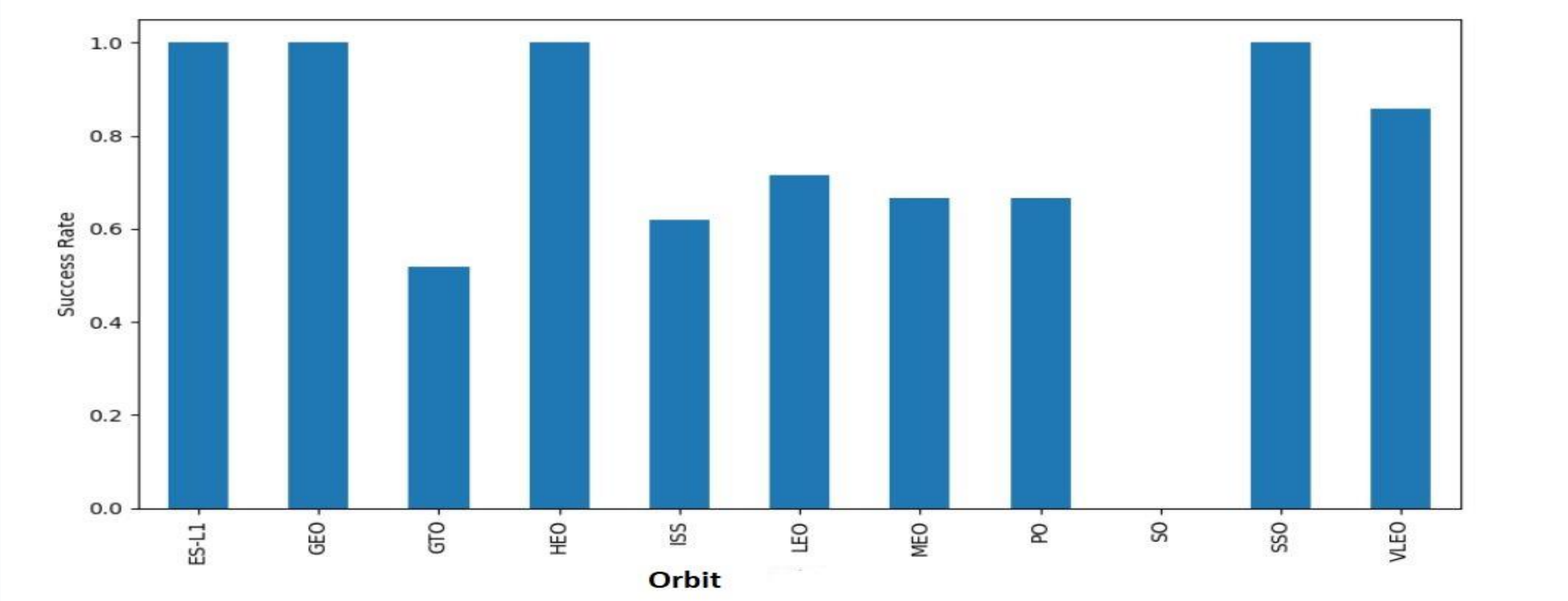
Payload vs. Launch Site



Explanation:

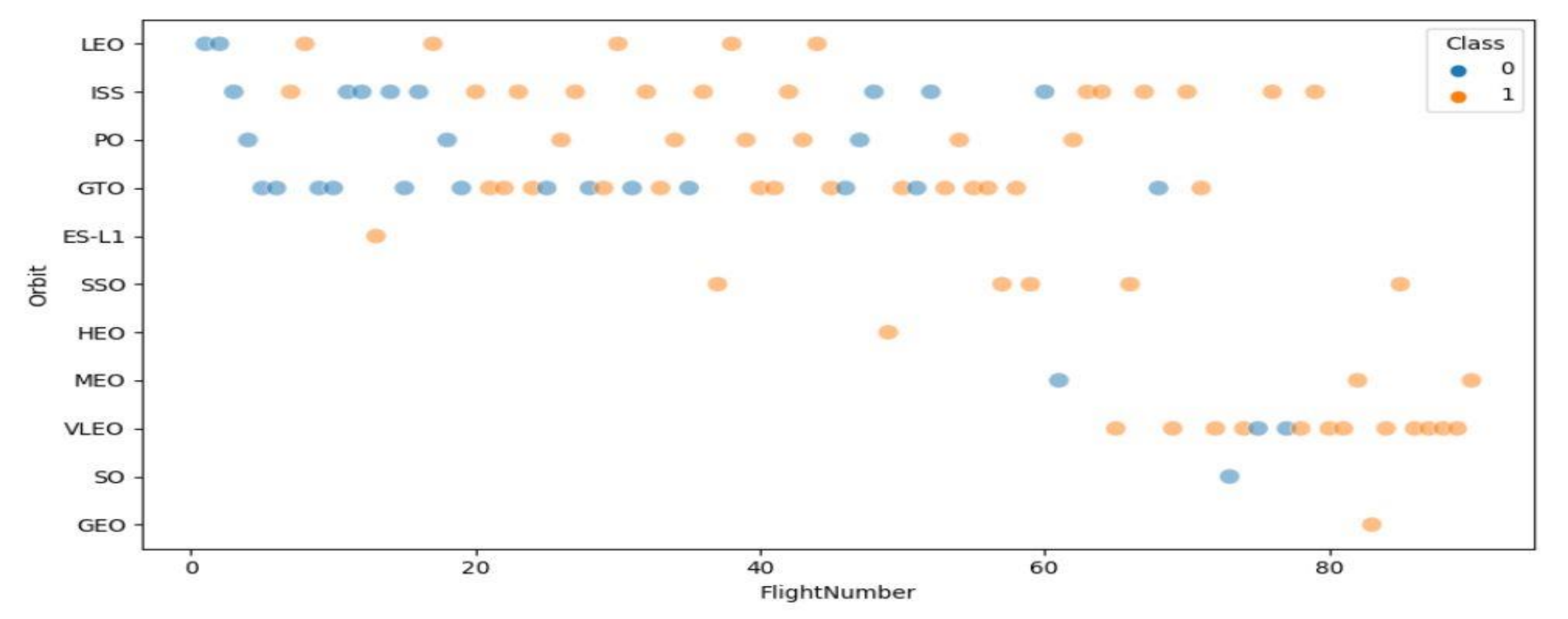
- For all launch sites the higher the payload mass, the higher the success rate.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg

Success Rate vs. Orbit Type



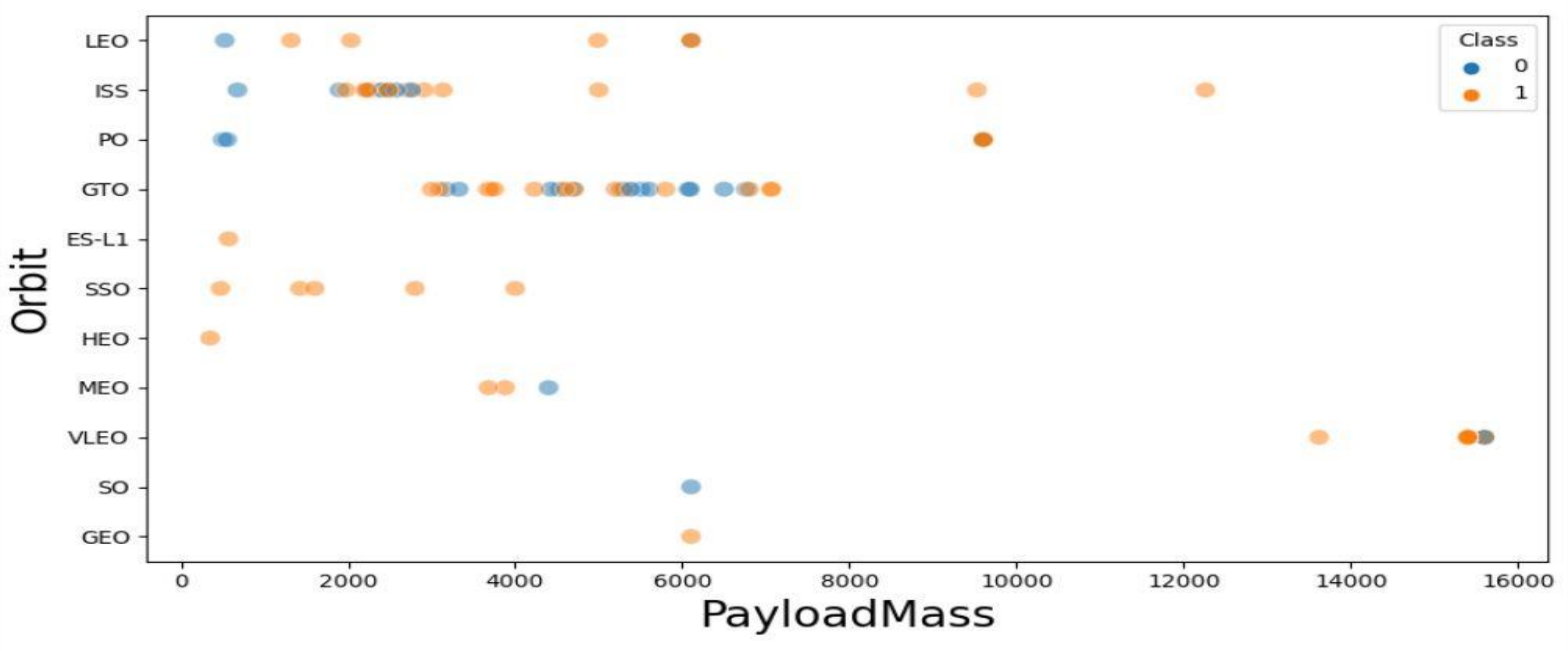
- Orbits with 0% success rate: SO
- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 85%:GTO, ISS, LEO, MEO, PO

Flight Number vs. Orbit Type



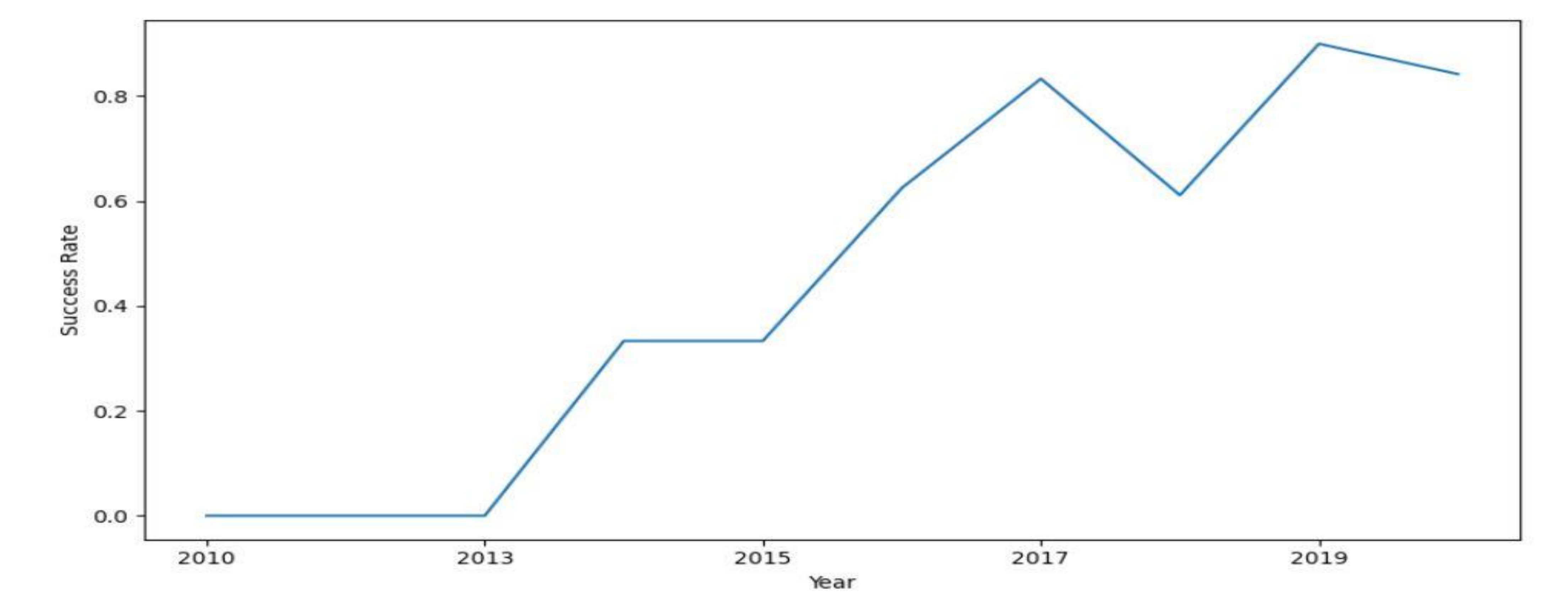
- We notice that the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights. But we can suppose that the high success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches or other orbits.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



- The success rate has been increasing since 2013

All Launch Site Names

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

In [7]:

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Out[7]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- The use of DISTINCT in the query allows us to remove duplicate LAUNCH_SITE and produce unique launch site names

Launch Site Names Begin with 'CCA'

In [10]:
```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

- The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA.

Total Payload Mass

```
In [16]:
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayLoadMassNASA_CRS
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';

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

Out[16]:

TotalPayLoadMassNASA_CRS
45596.0

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

Average Payload Mass by F9 v1.1

In [7]:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as AVG
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Out[7]:

AVG
2534.6666666666665

- Using the avg() function and where clause to display the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
In [22]: %%sql
         SELECT MIN(Date) as earliest_date
         FROM SPACEXTBL
         WHERE Landing_Outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

Out[22]:
```

earliest_date
01/08/2018

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [26]: %%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;

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

Out[26]:

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

- This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

In [29]:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS Count
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Out[29]:

Mission_Outcome	Count
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Listing the total number of successful and failure mission outcomes by utilizing the group by clause

Boosters Carried Maximum Payload

In [30]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

Out[30]:

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

- Using a subquery to find the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

In [38]:

```
%%sql
SELECT substr(Date, 4, 2) as Month ,LANDING_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE
FROM SPACEXTBL
where substr(Date,7,4)='2015' AND LANDING_OUTCOME = 'Failure (drone ship)' ;
```

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

Out[38]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- SQLite doesn’t support monthname() function so we use Substr function to process date in order to get month or year. Substr(DATE, 4, 2) shows month and Substr(DATE,7, 4) shows year, we use this to query month, booster version and launch site where landing was unsuccessful and landing date took place in 2015

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

In [8]:

```
%%sql
SELECT landing_outcome, COUNT(landing_outcome) AS count
FROM spacextbl
WHERE date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY landing_outcome
ORDER BY count DESC
```

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

Out[8]:

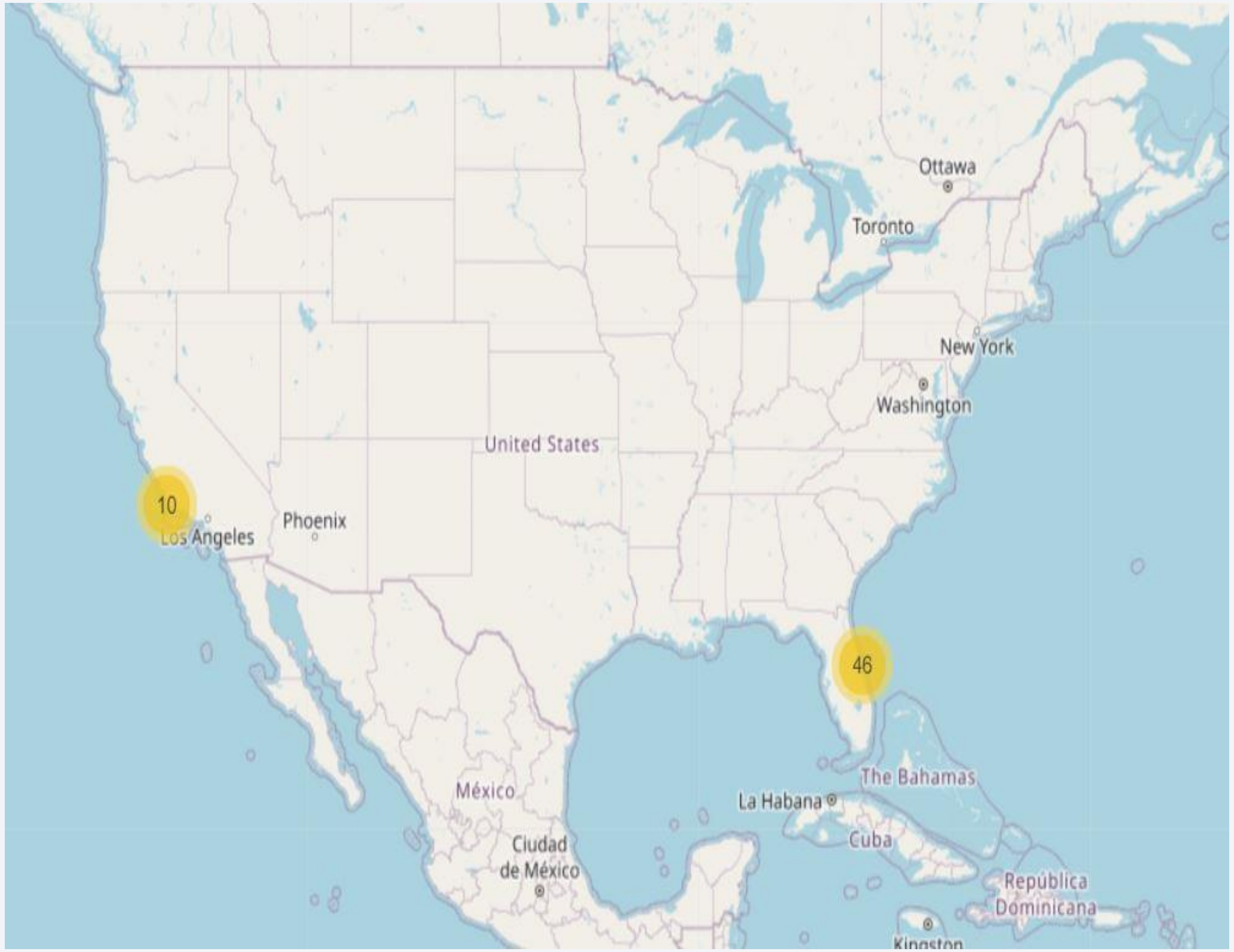
Landing_Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- Ranking 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

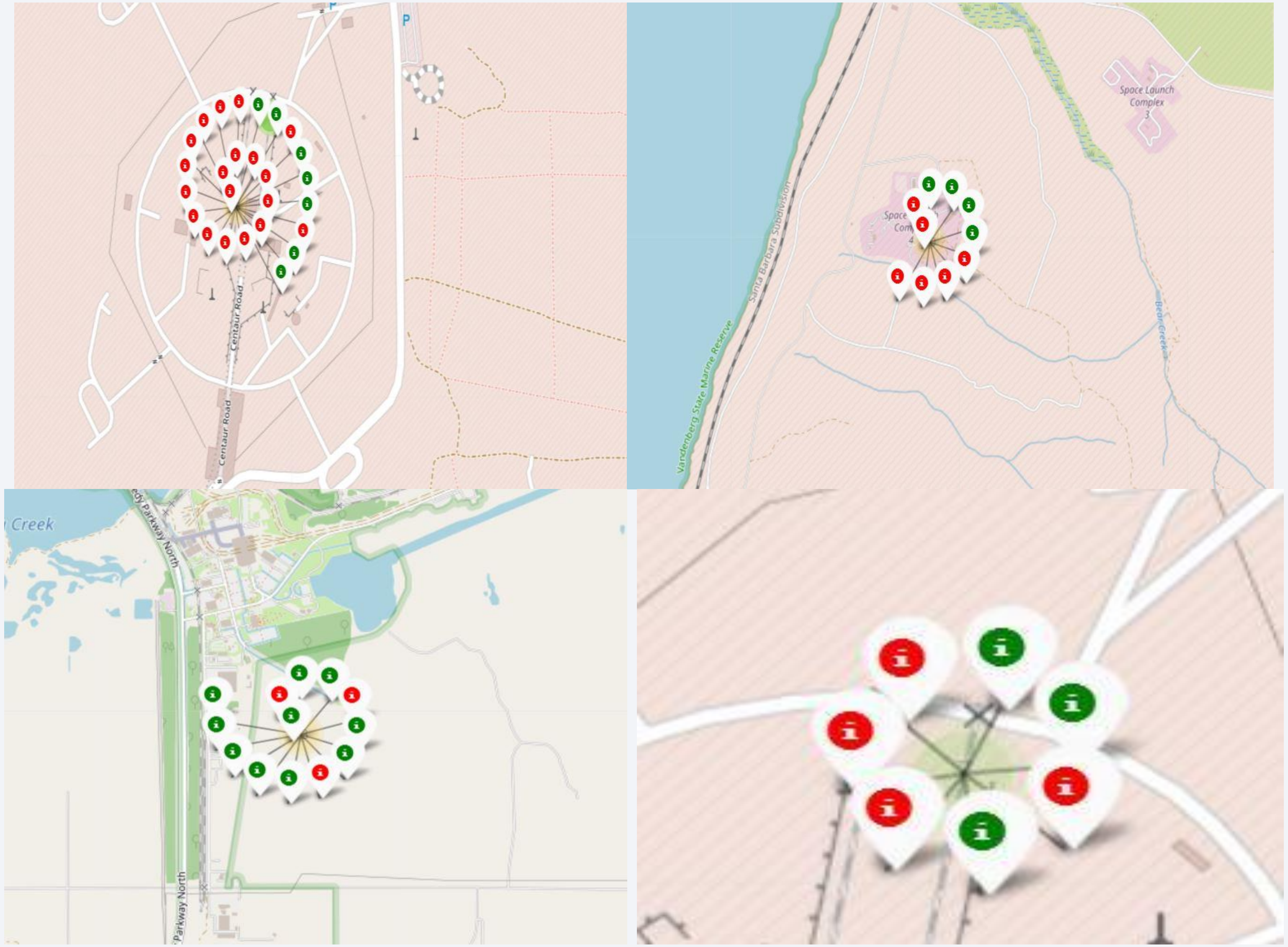
Launch Sites Proximities Analysis

Launch Sites in U.S.A



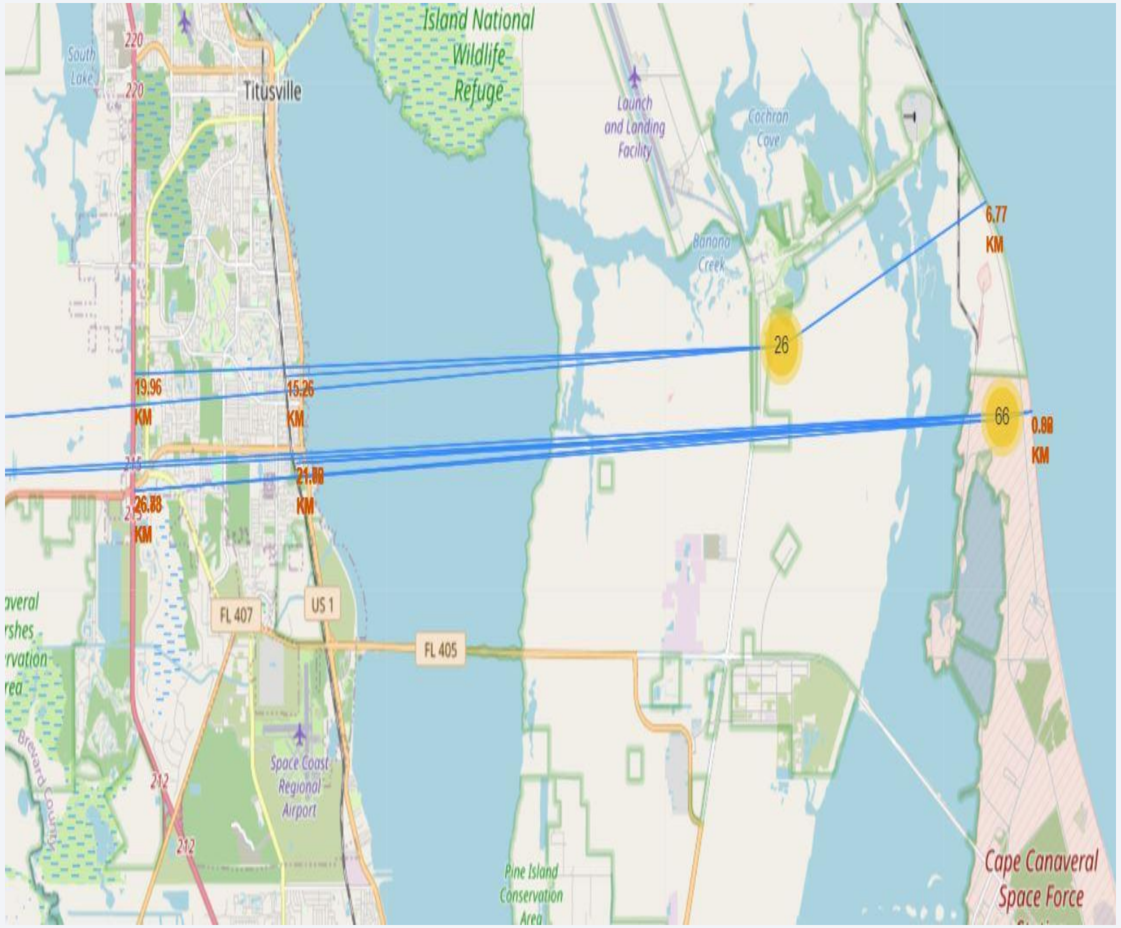
- All launch sites are in very close proximity to the coast, also all of Launch sites are in proximity to the Equator line to save fuel when launching

Colored Folium map for Launch outcomes



- Green marker represents successful launches. Red marker represents unsuccessful launches.
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- Launch Site KSC LC-39A has a very high Success Rate.

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



- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- relative close to railway (15.23 km)

- relative close to highway (20.28 km)

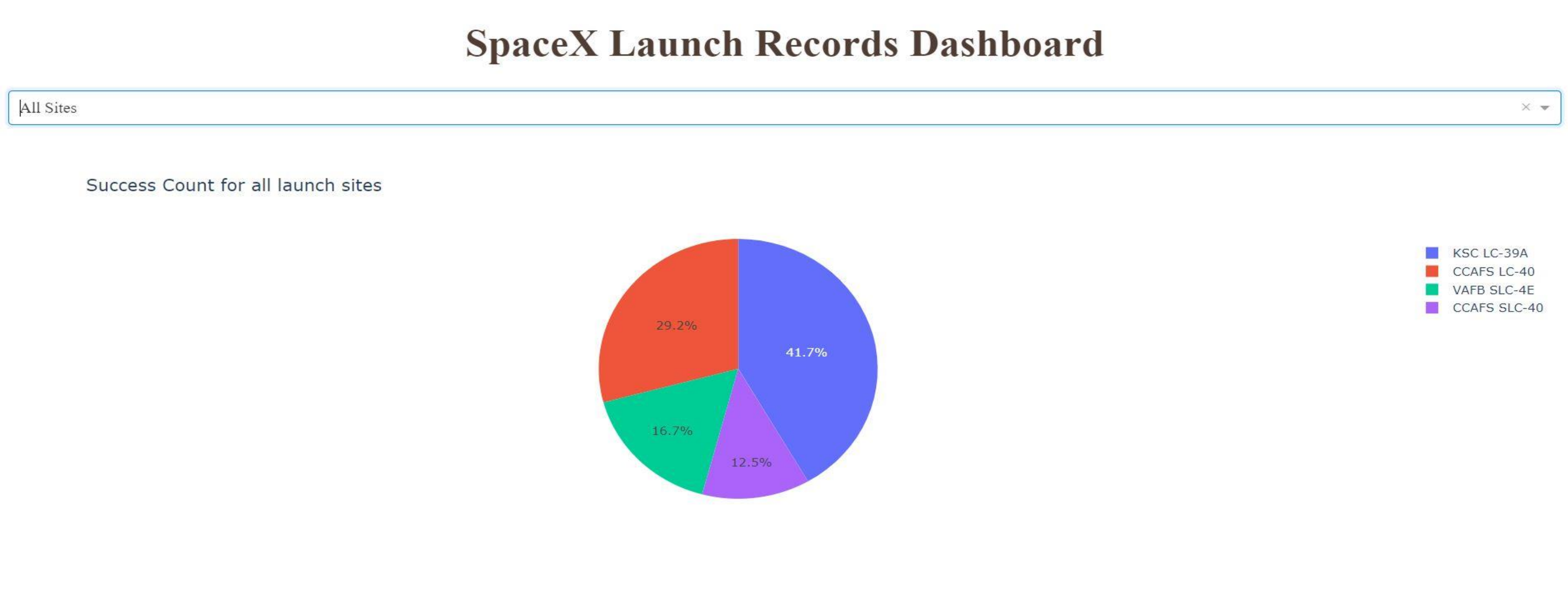
- relative close to coastline (14.99 km)

- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover much distances in a few seconds. It could be potentially dangerous to populated areas.

Section 4

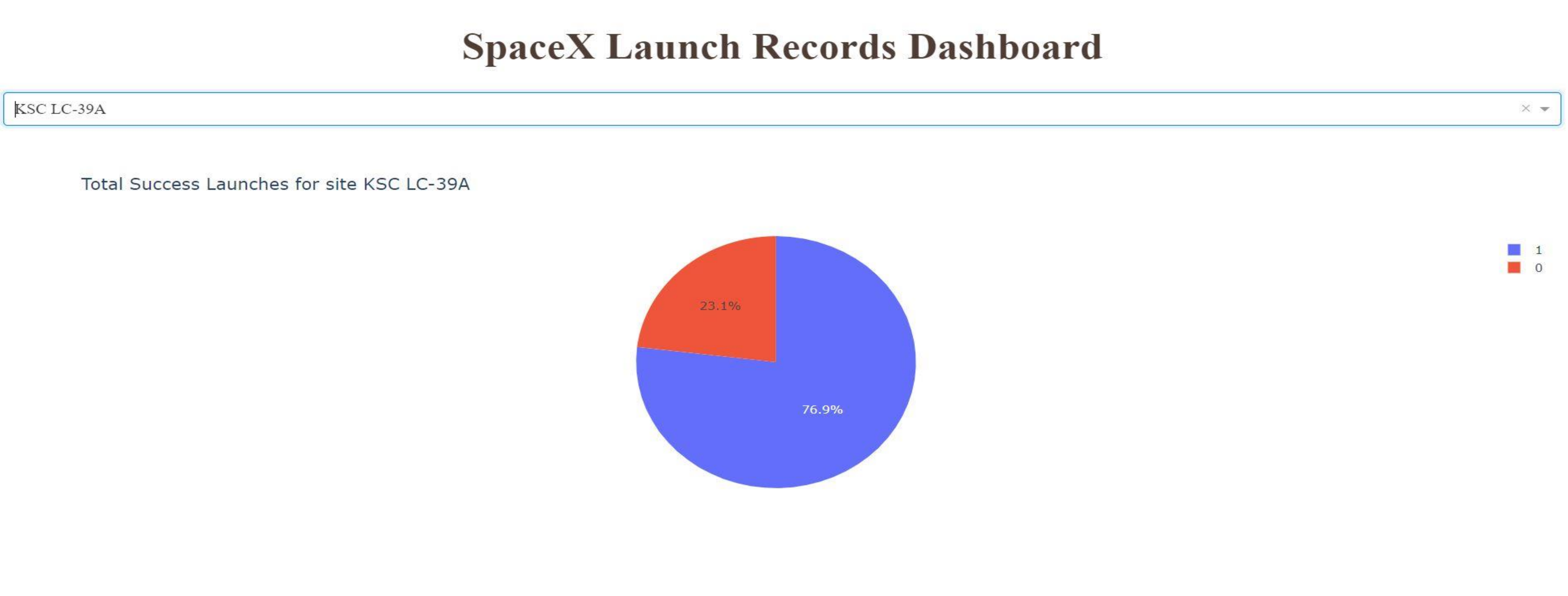
Build a Dashboard with Plotly Dash

Total successful launches for all sites



- KSC LC-39A has the most success rate of all launch sites.

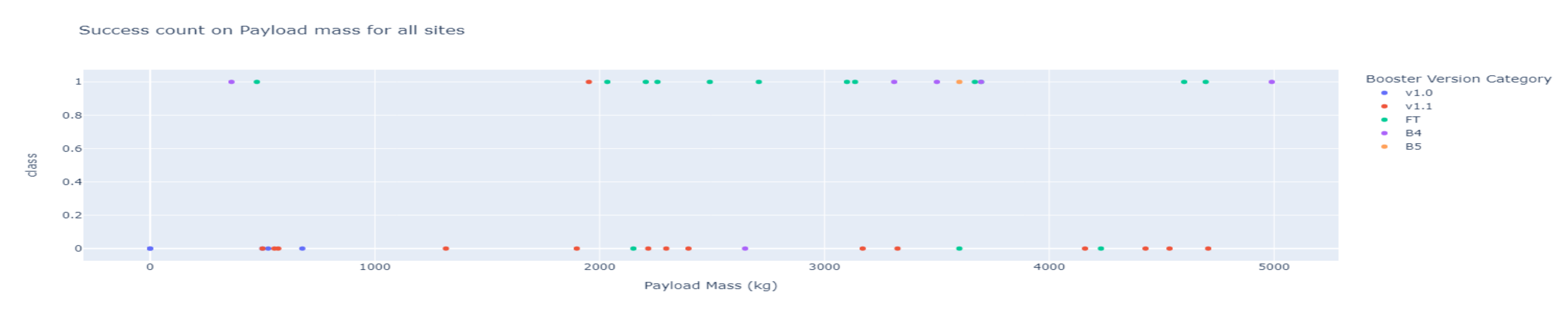
Launch site with highest launch success ratio



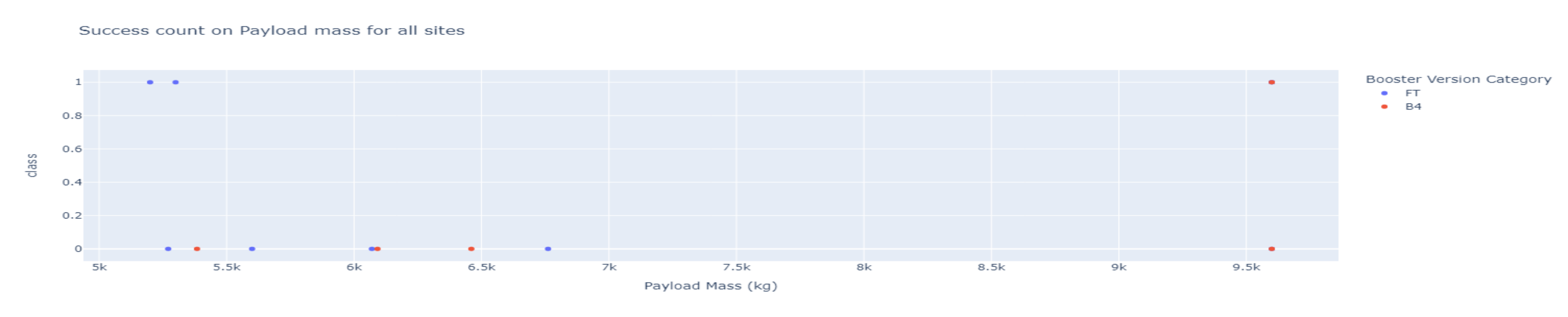
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and 3 failed landings.

Payload Mass vs. Launch Outcome for all sites

- **payload 0kg to 5000kg.**



- **Payload 5000kg to 10000kg**



- Payloads between 2000 and 5500 kg have the highest success rate.

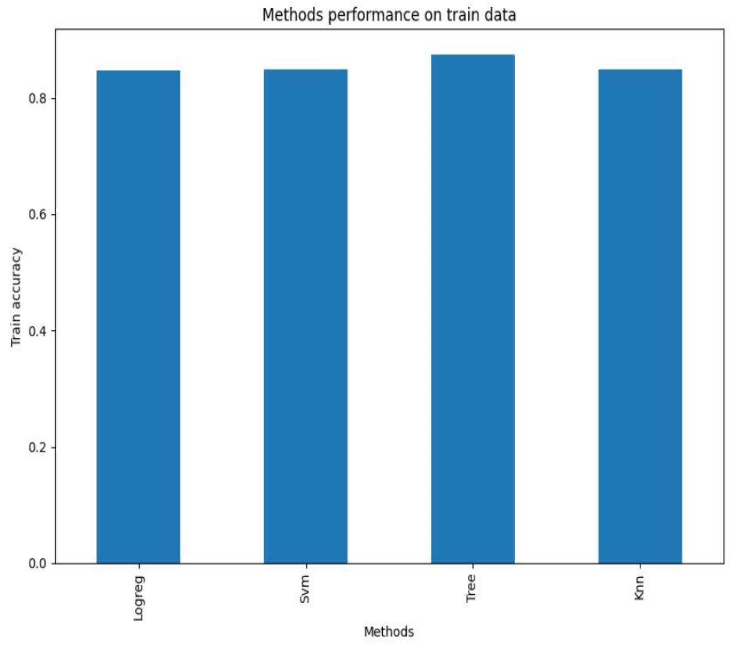
Section 5

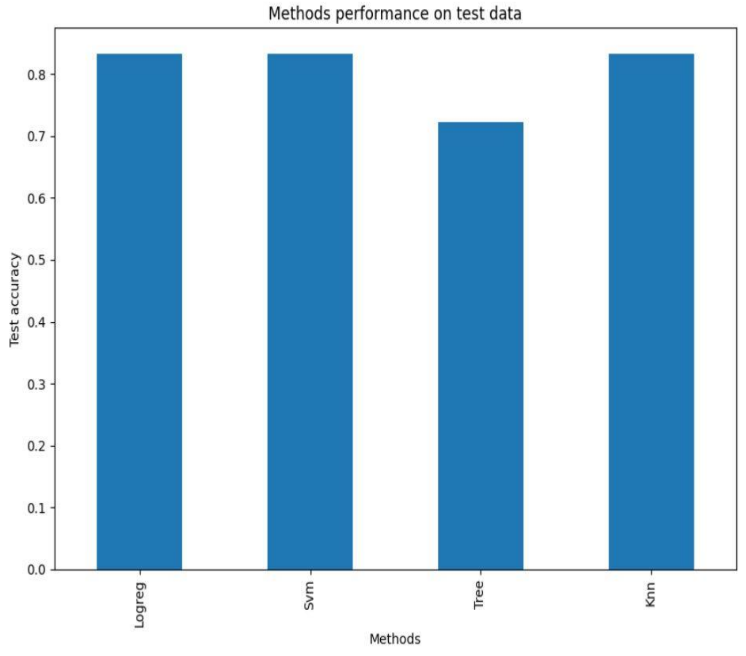
Predictive Analysis (Classification)

Classification Accuracy

[34]:

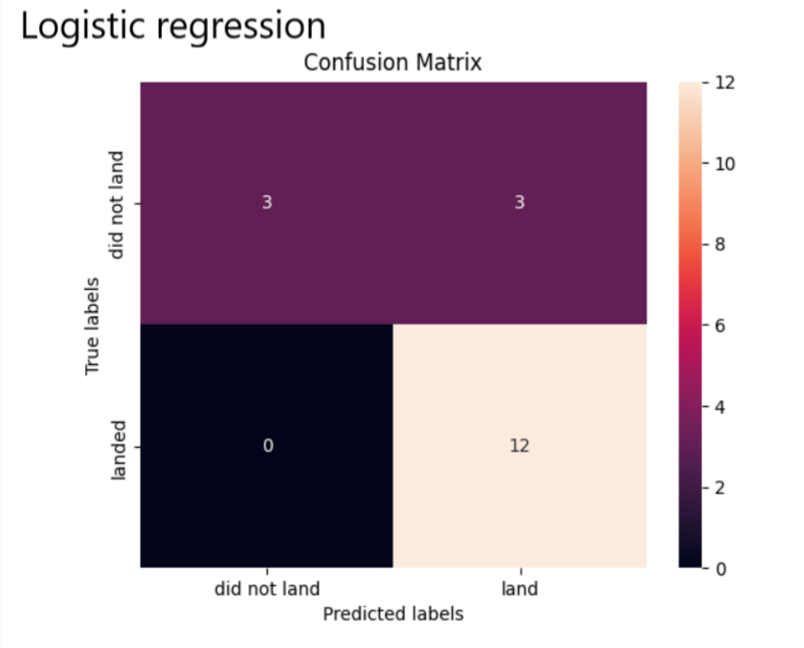
	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.875000	0.722222
Knn	0.848214	0.833333

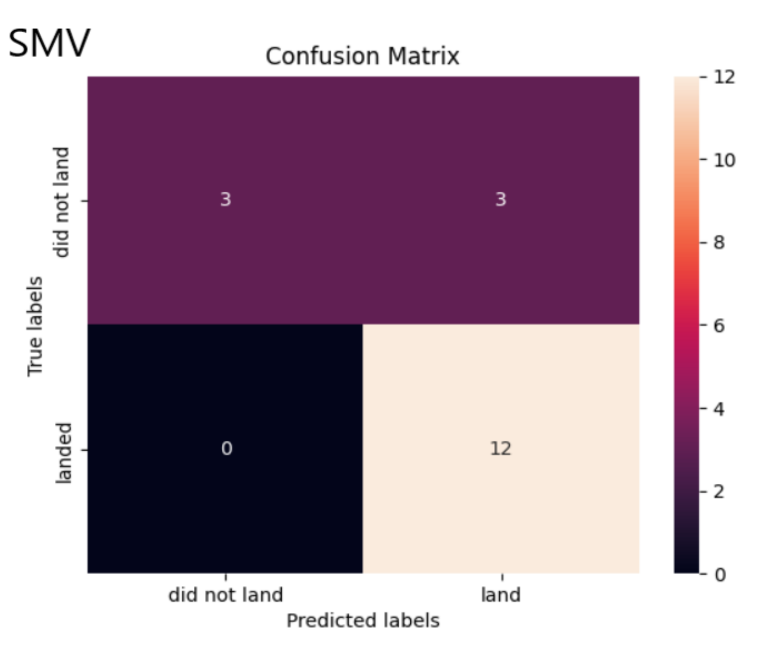


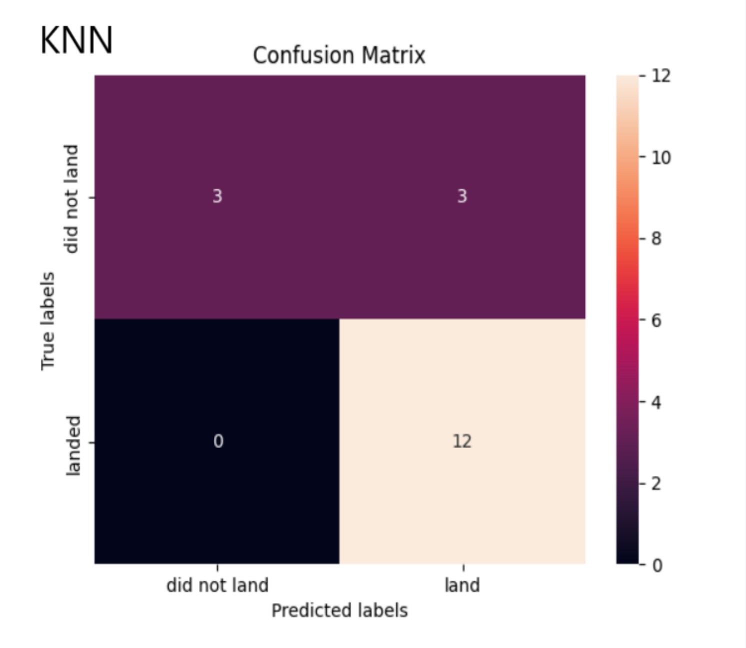


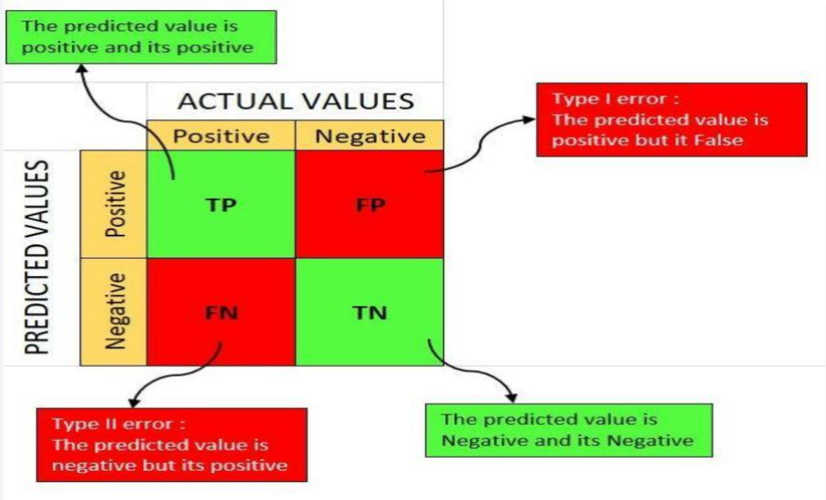
- Logistic Regression (0.8333333333333334), SVM (0.8333333333333334), and KNN (0.8333333333333334) all performed equally well and have the highest accuracy, while Decision Tree (0.7222222222222222) cannot identify true negatives.

Confusion Matrix









- Test accuracy are equal for the three models, the confusion matrices are also identical. The main problem of these models are false positives.

Conclusions

- The success or failure of a mission can be explained by several factors or conditions such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge and experience between launches that allowed launches go from a launch failure to a success.
- The orbits with the highest success rate are ES-L1, GEO, HEO and SSO they have a 100% success rate
- The success rate of launches increases over the years.
- For this dataset, we can choose Logistic Regression, SVM or KNN as the best models for predicting launch outcome they have the same test accuracy and similar train accuracy.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- KSC LC-39A has the highest success rate of the launches from all the sites.

Thank you!