IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kateb zakaria
12/14/2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

- Project background and context
- Problems you want to find answers

Introduction

- Project background and context

Space Exploration Technologies Corp. (SpaceX) is an American spacecraft manufacturer, launcher, and a satellite communications corporation headquartered in Hawthorne, California. It was founded in 2002 by Elon Musk with the stated goal of reducing space transportation costs to enable the colonization of Mars. The company advertises Falcon9 rocket launches on its website, with a cost of 62 million dollars, other providers cost upward of 165 million dollars each.

Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond. Falcon 9 is the world's first orbital class reusable rocket. Reusability allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access.

- Problems you want to find answers

Introduction

Falcon 9
Overview

HEIGHT	70 m / 229.6 ft
DIAMETER	3.7 m / 12 ft
MASS	549,054 kg / 1,207,920 lb
PAYLOAD TO LEO	22,800 kg / 50,265 lb
PAYLOAD TO GTO	8,300 kg / 18,300 lb
PAYLOAD TO MARS	4,020 kg / 8,860 lb

Introduction

- Project background and context
- Problems you want to find answers

- Problems you want to find answers:

i. How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

ii. How do variables such as payload, launch site, number of flights and orbits influence the successful landing of the first stage?

iii. Is there an increase in successful landings over time?

iv. What is the best algorithm that can be used for binary classification in this case?

v. Which algorithm should be used for binary classification in this case?

vi. What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping using beautifulsoup library from Wikipedia.
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification

Methodology

Executive Summary

- 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 in order to obtain the best results

Data Collection

Data collection process involved a combination of API requests from SpaceX REST

API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete

information about the launches for a more detailed analysis.

Data Collection

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, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API (click the GitHub logo to go to the GitHub repo)

01 - Requesting rocket launch data from SpaceX API

02 - Check the content of the response and the response status code

03 - Decode the response content as a *json* using and turn it into a Pandas dataframe using the normalization method *.json_normalize()*

04 - Requesting needed information about the launches from SpaceX API by applying custom functions

05 - Constructing data we have obtained into a dictionary and convert it into a panda dataframe

06 - Filtering the dataframe to only include Falcon 9 launches

07 - Replacing missing values of Payload Mass column with calculated .mean() for this column

08 – Export the result into a CSV file

Data Collection – Scraping (click the GitHub logo to go to the GitHub repo)

01 -Request the Falcon9 Launch Wiki page from its URL

02 - Extract all column/variable names from the HTML table header

03 - Create a data frame by parsing the launch HTML tables

04 - Constructing data
we have obtained into a dictionary

05 - Creating a dataframe
from the dictionary

06 – Export the result into a CSV file

GitHub

Data Wrangling (click the GitHub logo to go to the GitHub repo)

01 - Perform exploratory Data Analysis and determine Training Labels

02 - Calculate the number of launches on each site

03 - Calculate the number and occurrence of each orbit & the number and occurrence of mission outcome per orbit type

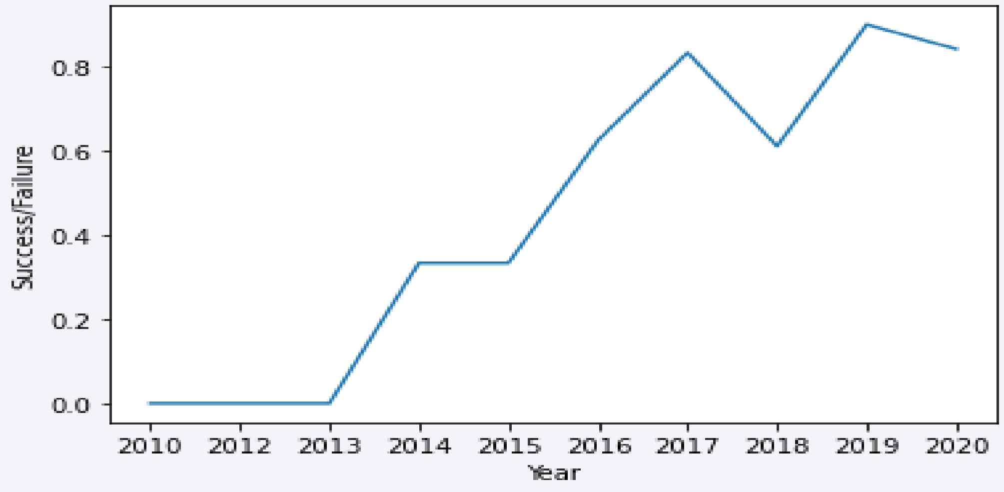
04 - Create a landing outcome label from Outcome column

05 Exporting the resulting data
to CSV file

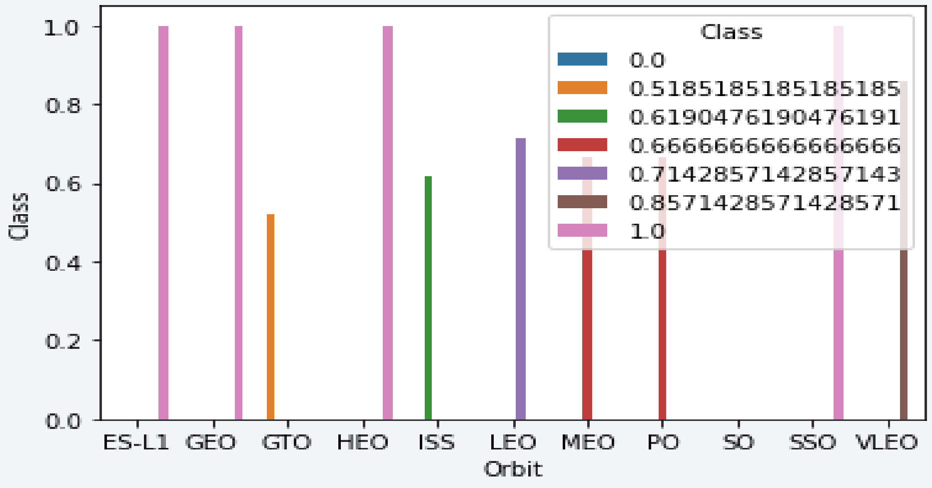
GitHub

EDA with Data Visualization (click the GitHub logo to go to the GitHub repo)

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.



the launch success yearly trend



bar chart for the success rate of each orbit

GitHub

EDA with SQL (click the GitHub logo to go to the GitHub repo)

- We loaded the SpaceX dataset into a Db2 database in IBMCloud.
- We applied EDA with SQL to get insights from the data, writing queries to find:
  - ➢ The names of unique launch sites in the space mission.
  - ➢ The total payload mass carried by boosters launched by NASA (CRS).
  - ➢ The average payload mass carried by booster version F9 v1.1.
  - ➢ The total number of successful and failure mission outcomes.
  - ➢ The failed landing outcomes in drone ship, their booster version and launch site names.

GitHub

Build an Interactive Map with Folium (click the GitHub logo to go to the GitHub repo)

- We marked all launch sites and added map objects like markers, circles, lines to be marked.
- successful or failed launches for each site on the folium map.
- We have assigned the results of the feature launch (failure or success) to class 0 and 1, that is, 0 for error and 1 for success.
- Using groups of colored markers, we determined which launch sites are relatively successful.
- We calculated the distances between a launch site and its vicinity

GitHub

Build a Dashboard with Plotly Dash (click the GitHub logo to go to the GitHub repo)

- We built an interactive dashboard with Plotly dash with the following:
- 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.

GitHub

Predictive Analysis (Classification) (click the GitHub logo to go to the GitHub repo)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using
- feature engineering and algorithm tuning

Predictive Analysis (Classification) (click the GitHub logo to go to the GitHub repo)

01 - loaded the data using numpy and pandas, transformed the data, split our data into training and testing

→ **02 -** Creating a GridSearchCV object with cv = 10 to find the best parameters

→ **03 –** Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

→ **04 -** Calculating the accuracy on the test data using the method .score() for all models

→ **05 -** Examining the confusion matrix for all models

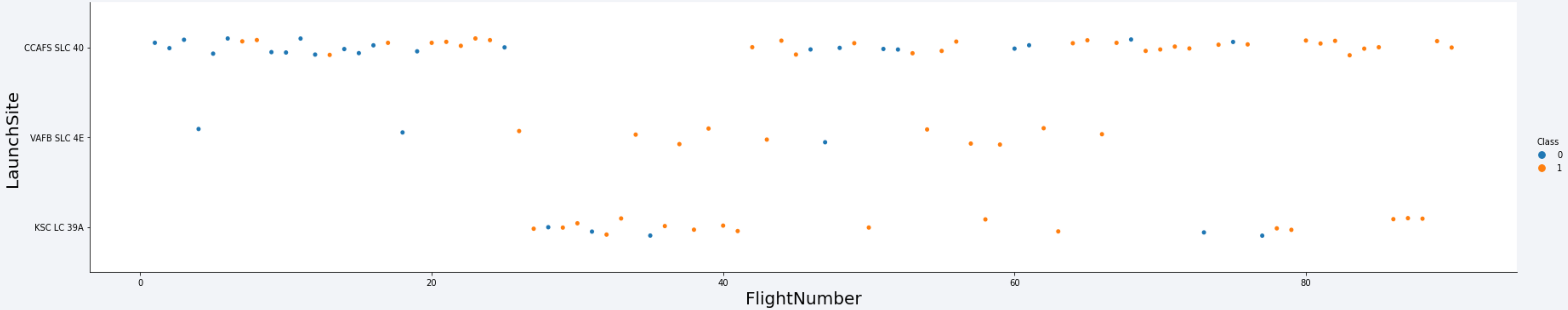
→ **06 -** Finding the method performs best by examining the Jaccard_score and F1_score metrics

GitHub

Section 2

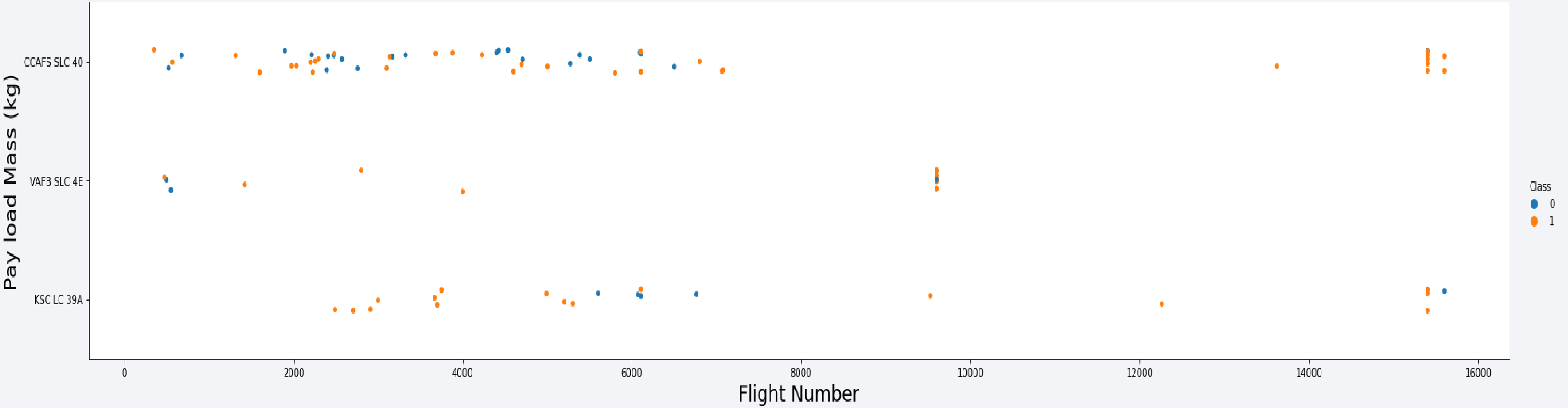
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



- For every launch site the higher the payload mass, the higher the success rate
- The grater the payload mass for lunch site CCAFS SLC 40 the higher the success rate
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

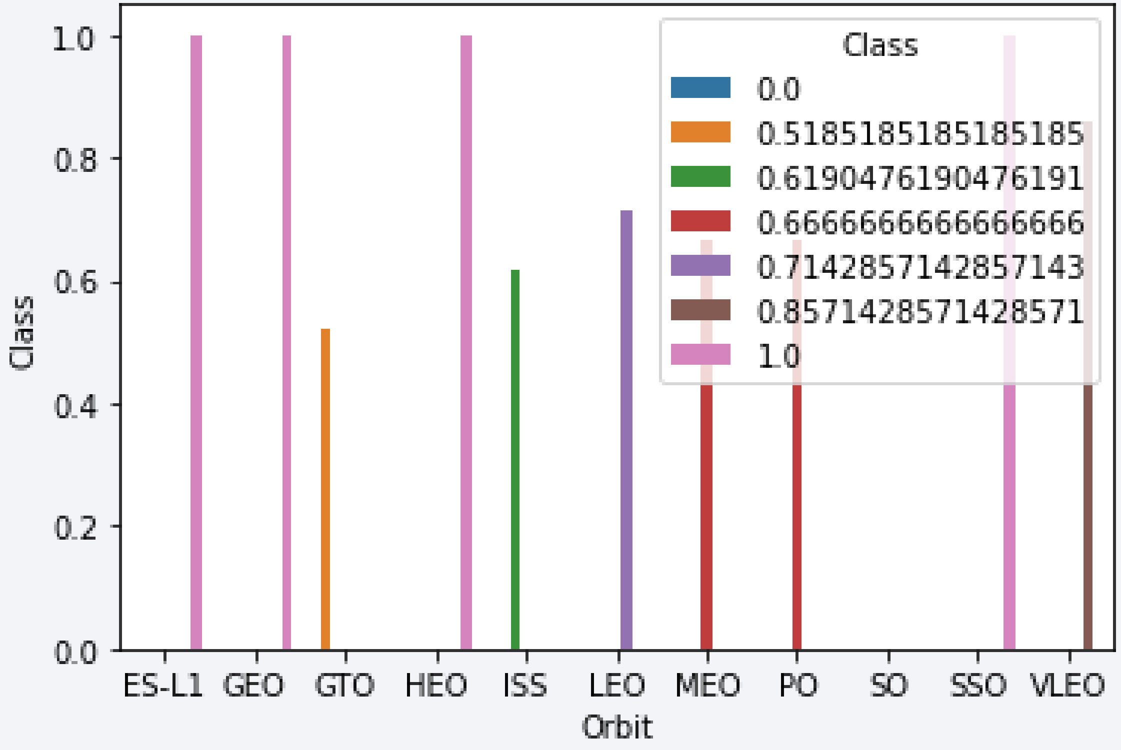
Success Rate vs. Orbit Type

- Orbits with 100% success rate:

  - ES-L1, GEO, HEO, SSO

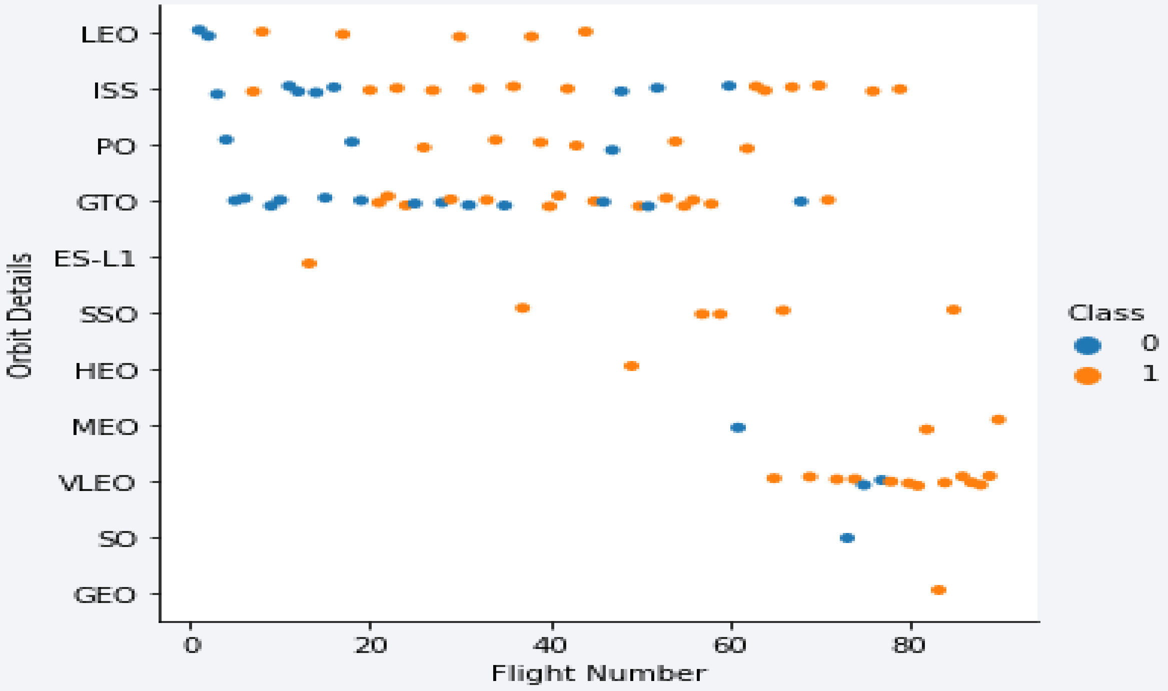
- Orbits with 0% success rate:

  - SO



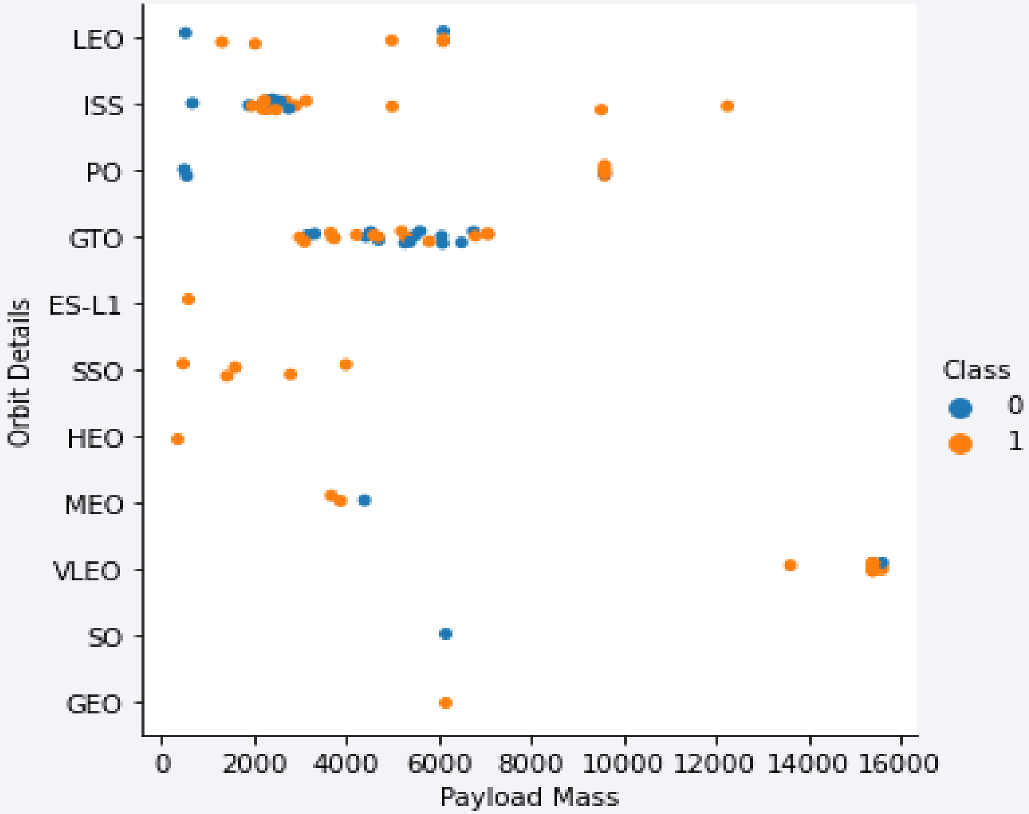
Flight Number vs. Orbit Type

- 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.



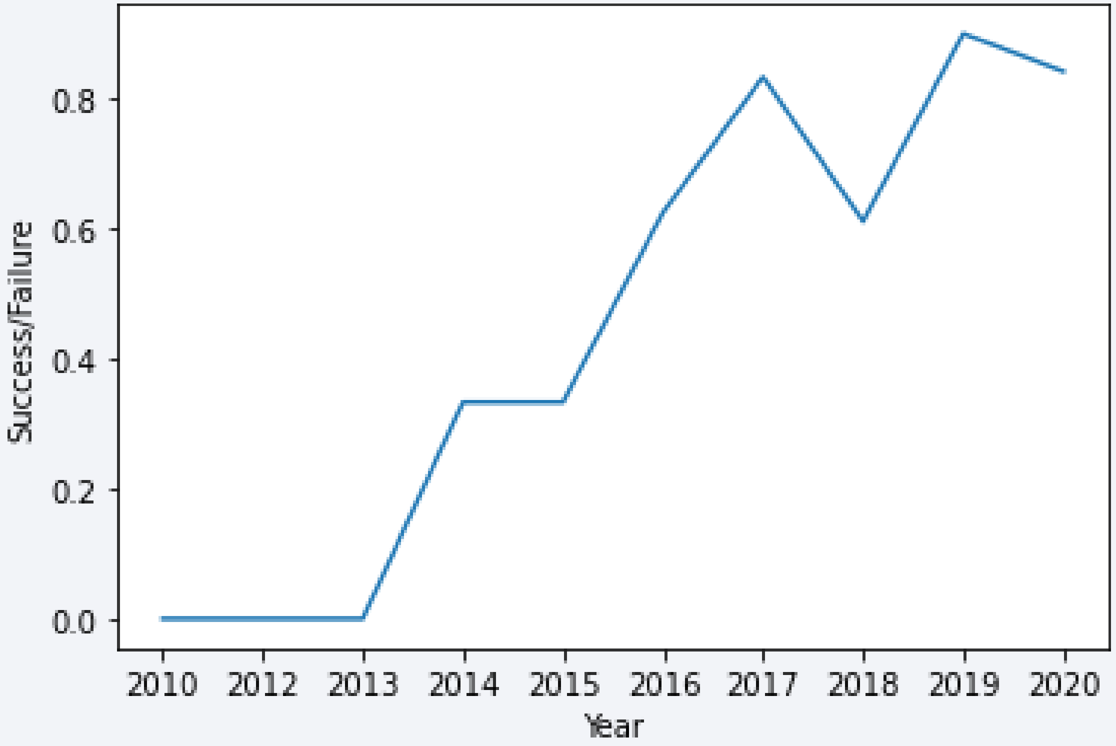
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020



All Launch Site Names

```
%sql select Unique(LAUNCH_SITE) from SPACEXX;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXX where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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)
2012-08-10	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
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

- We used the query above to display 5 records where launch sites begin with `CCA

Total Payload Mass

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

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

total_payload_mass
22007

- 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 average_payload_mass from SPACEXX where booster_version like '%F9 v1.1%';

 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

average_payload_mass
3226

- Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(DATE) as first_successful_landing from SPACEXX where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

first_successful_landing
2017-01-05

- 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 BOOSTER_VERSION from SPACEXX where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

booster_version
F9 FT B1022
F9 FT B1031.2

- Listing 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

```
%sql select mission_outcome, count(*) as total_number from SPACEXX group by mission_outcome;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

mission_outcome	total_number
Success	44
Success (payload status unclear)	1

- Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION from SPACEXX where payload_mass__kg_ = (select max(PAYLOAD_MASS__KG_) from SPACEXX);
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

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

2015 Launch Records

```
%%sql select monthname(DATE) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, landing__outcome from SPACEXX
     where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
October	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXX
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://crx89249:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

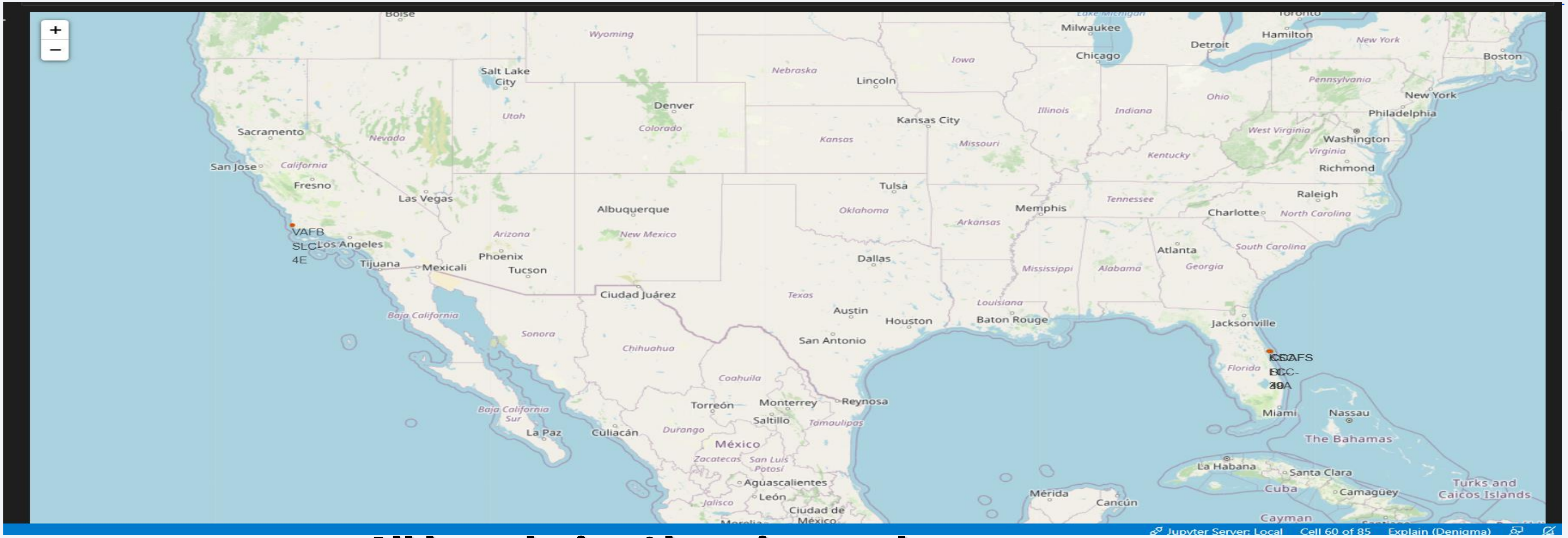
landing__outcome	count_outcomes
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

- 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

<Folium Map Screenshot 1>

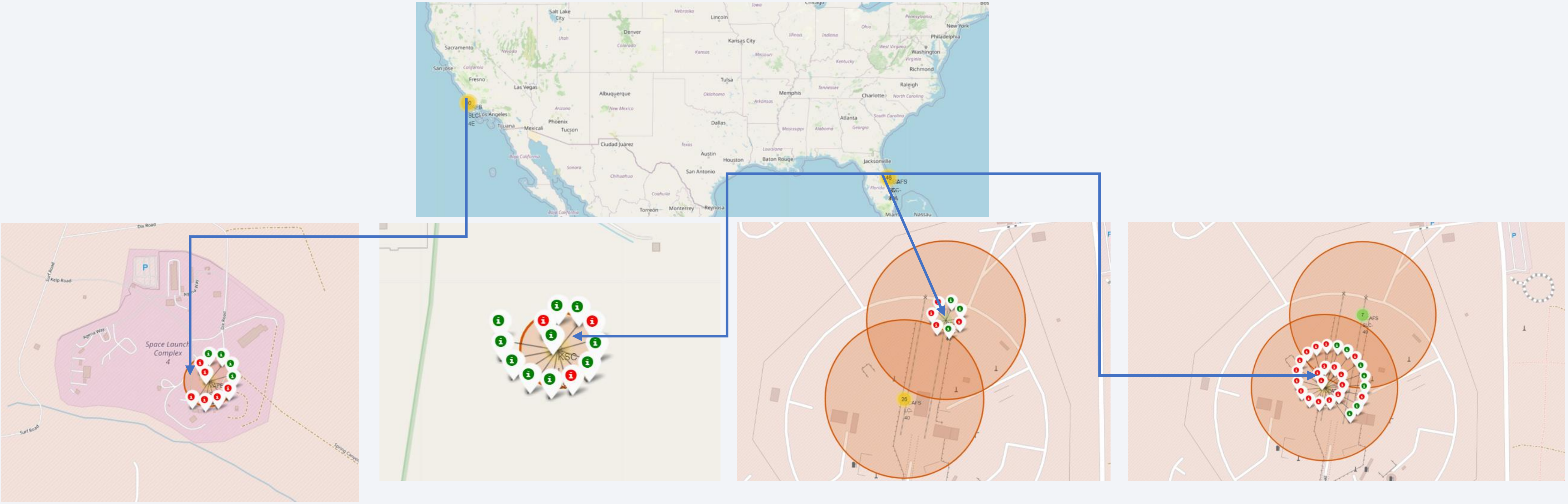


All launch sites' location markers on a map

- Most of Launch sites are in proximity to the Equator line
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

<Folium Map Screenshot 2>

Color-labeled launch records on the map

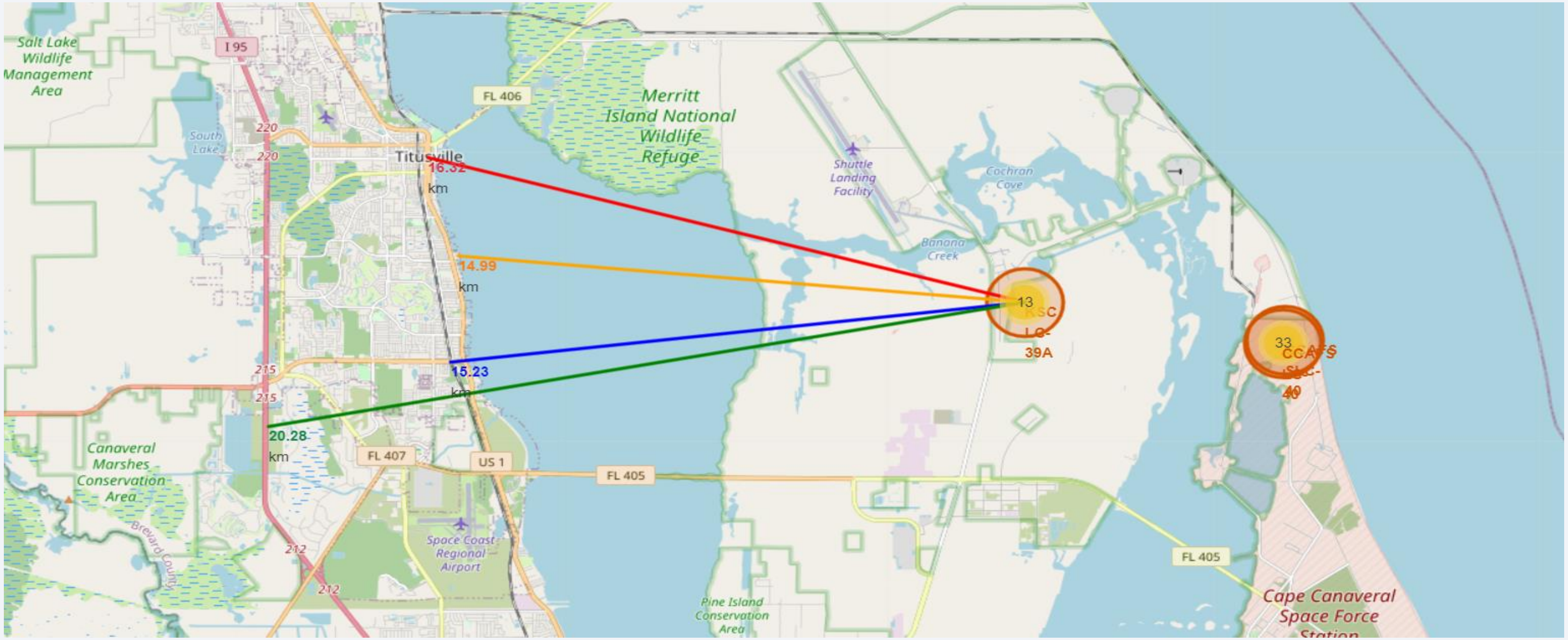


Green Marker = Successful Launch

Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate

<Folium Map Screenshot 3>

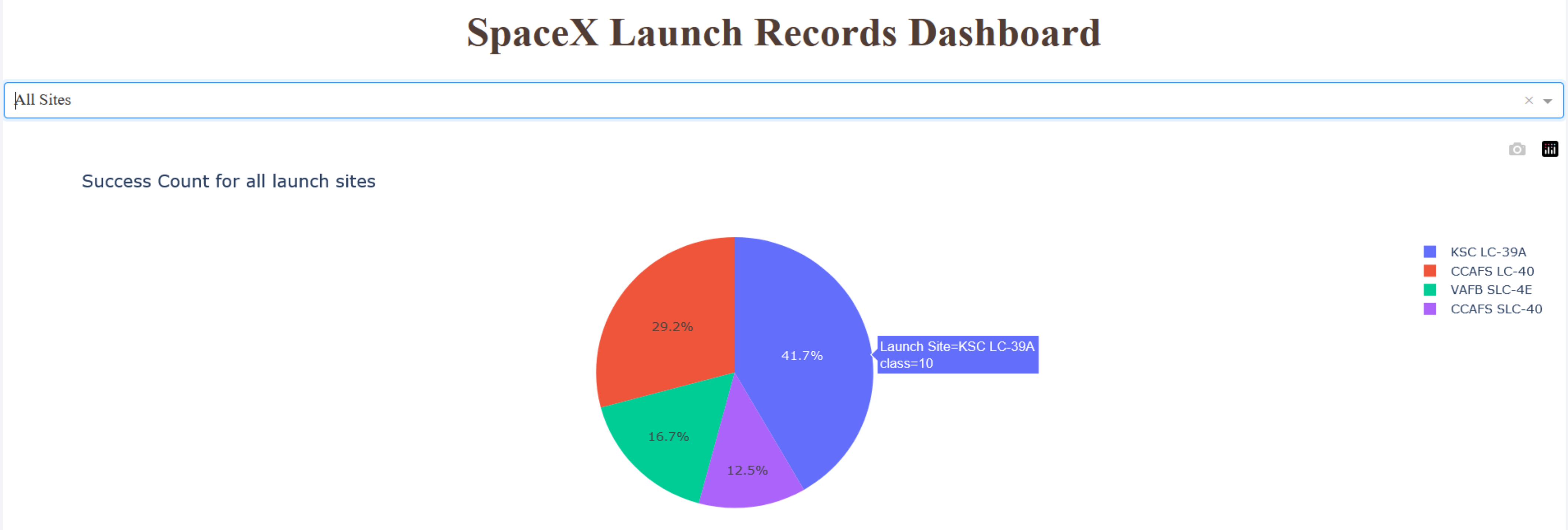


- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km)

Section 4

Build a Dashboard with Plotly Dash

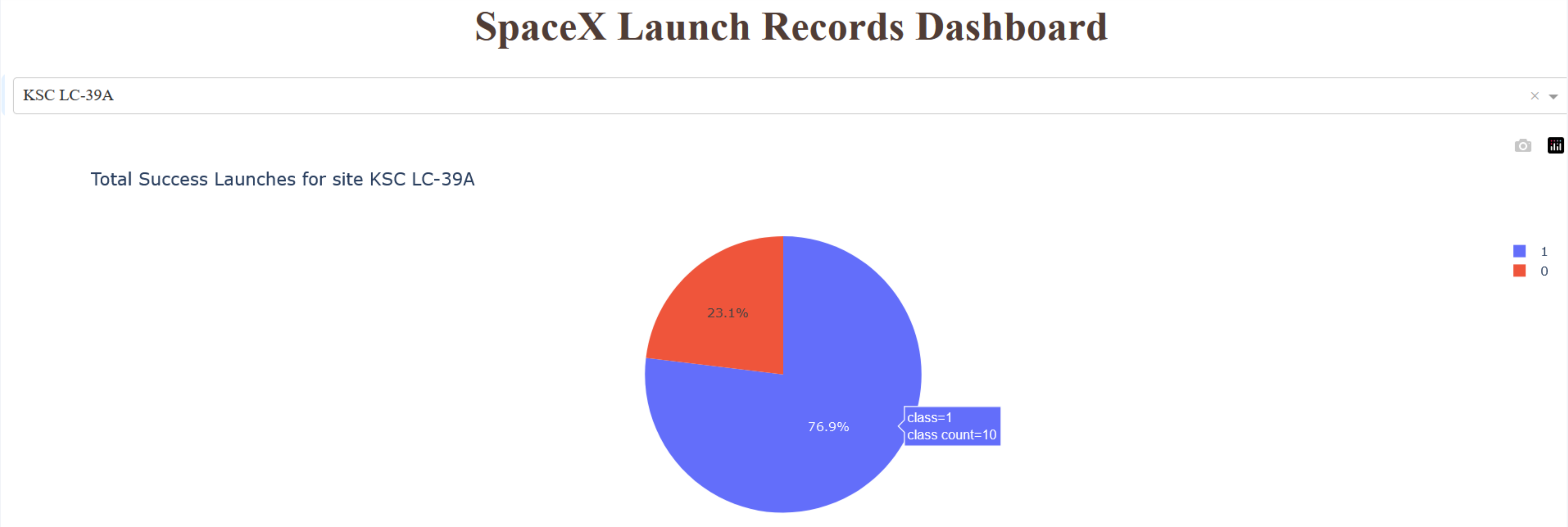
<Dashboard Screenshot 1>



Pie chart showing the success percentage achieved by each launch site

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

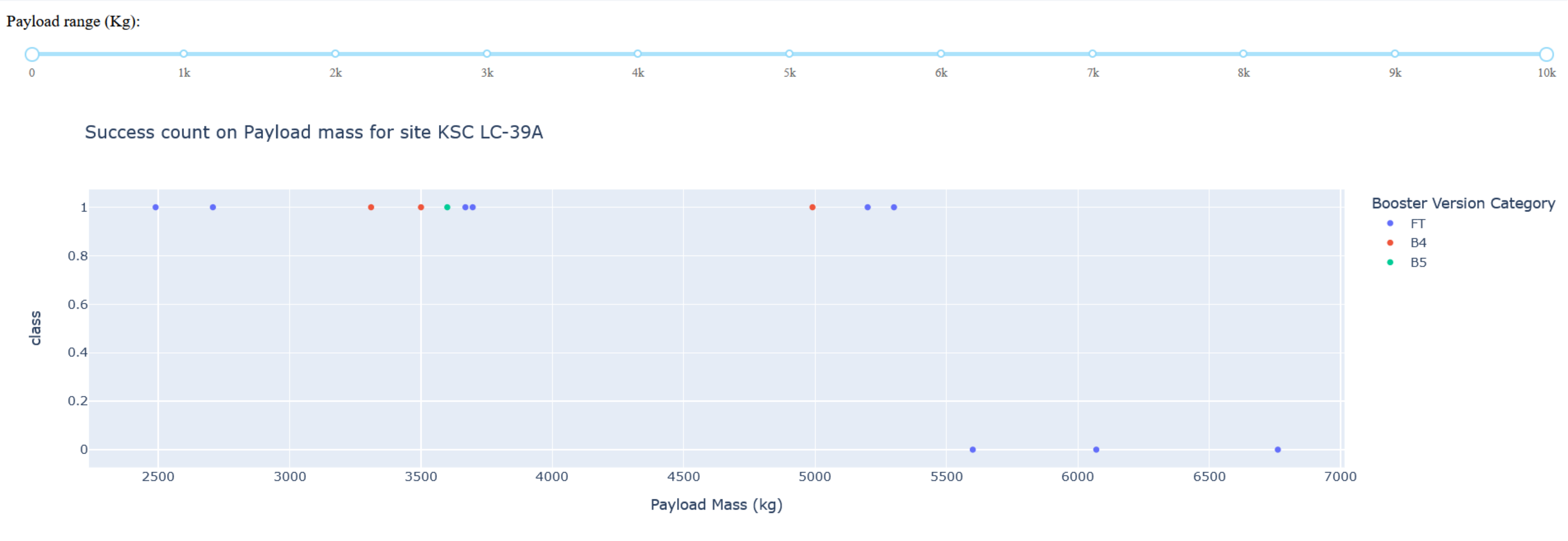
<Dashboard Screenshot 2>



Pie chart showing the total success lunches for the site KSC LC-39A

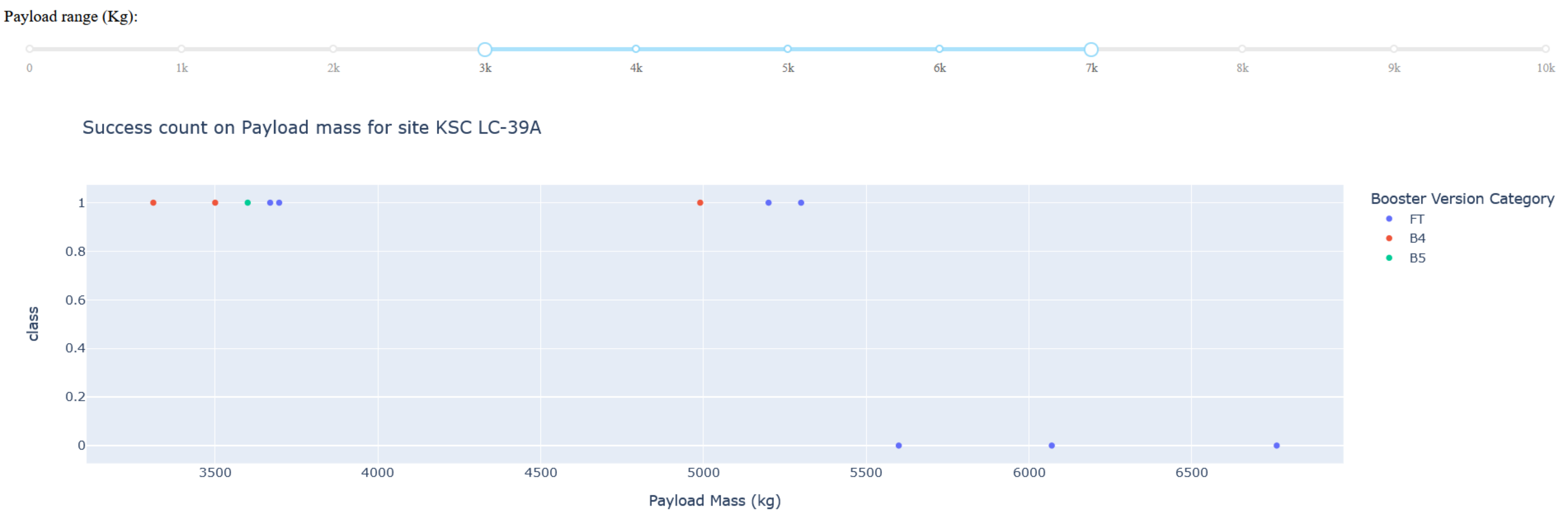
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

<Dashboard Screenshot 3>



Scatter plot of Payload vs Launch Outcome for all sites Payload 0 (kg) to 10K (kg)

<Dashboard Screenshot 3>



Scatter plot of Payload vs Launch Outcome for all sites Payload 3K (kg) to 7K (kg)

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Based on the scores of the Test Set, we can not confirm which method is the best.
Practically all these algorithms give the same result

Because there is no major difference between the results The scores from the dataset confirm that the best model is the decision tree model. Not only does this model have higher scores, it also has the highest accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.840580	0.819444
F1_Score	0.909091	0.916031	0.913386	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556

Scores and Accuracy of the Entire Data Set

Confusion Matrix

- The decision tree classifier confusion matrix demonstrates that the classifier can differentiate between different classes. The main problem has to do with false positives. the classifier reported that the landing was successfully completed.

Conclusions

We can conclude that:

- ❖The larger the flight amount at a launch site, the greater the success rate at a launch site.
- ❖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 for this task

Thank you!