IBM Developer
SKILLS NETWORK

Data Science Capstone Using SpaceX Dataset

Outline

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

Outline

Summary of methodologies

- · **Data Collection**: Utilized APIs and Web Scraping for comprehensive data acquisition.
- · **Data Wrangling**: Cleaned and transformed raw data to ensure its suitability for analysis.
- · **Exploratory Data Analysis**: Conducted in-depth analysis using SQL and visualization techniques.
- · **Visualization and Interaction**: Developed interactive maps using Folium and created dashboards using Plotly Dash for dynamic data exploration.
- · **Machine Learning Predictions**: Implemented machine learning models for classification tasks.

Summary of all results

· Exploratory Data Analysis helped us achieve an enhanced understanding of **parameter relationships**.
· Visualizations, improved **comprehension of parameter effects**.
· Machine Learning, established an efficient **predictive model** with a minimum **accuracy of 83%.**

Intro

· Project background and context

Collaborating with SpaceY to analyze publicly accessible data from SpaceX, focusing on the performance of Falcon 9 stage 1 rockets. Understanding the success and failure of landing these rockets is crucial as they significantly impact launch expenses. The ability to reuse them has the potential to alter the competitive landscape in the rocket launch industry.

· Problems to find answers

· Identify the factors, circumstances and parameters that influence the successful landing of stage 1 rockets after each deployment.

· Forecast the outcome (success or failure) of a new rocket landing based on the collected parameters.

· Evaluate the accuracy of the predictions made using the aforementioned parameters.

- We are using a SPACEX data set to explore their data
- § Data Collection/Data Wrangling

· data collection and data wrangling methodology related slides (1 pt)

· EDA and interactive visual analytics methodology related slides (3 pts)

· Predictive analysis methodology related slides (1 pt)

· EDA with visualization results slides (6 pts)

· EDA with SQL results slides (10 pts)

· Interactive map with Folium results slides (3 pts)

· Plotly Dash dashboard results slides (3 pts)

· Predictive analysis (classification) results slides (6 pts)

· Conclusion slide (1 pts)

· Applied your creativity to improve the presentation beyond the template (1 pts)

METHODOLOGY

Methodology

- Data collection methodology:

o Data collection involved leveraging SpaceX's public APIs and implementing web scraping techniques.

- Perform data wrangling

o The data underwent one-hot encoding to enhance its suitability for utilization in learning algorithms.

- Perform exploratory data analysis (EDA) using visualization and SQL

o Uncover novel data patterns through the application of SQL and visualization techniques.

- Perform interactive visual analytics using Folium and Plotly Dash

o Interactive methods in this analysis involved the utilization of Plotly Dash for dashboards and Folium maps.

- Perform predictive analysis using classification models

o Various machine learning algorithms were assessed to determine the most effective method.

Data Collection

Using SpaceX API calls

- Collecting Data using SpaceX API calls
- Normalize Data using JSON functions
- Preprocess columns and rows
- Loading Dictionary and converting to DataFrame Object
- Filter for Falcon 9 data

Using Web Scraping

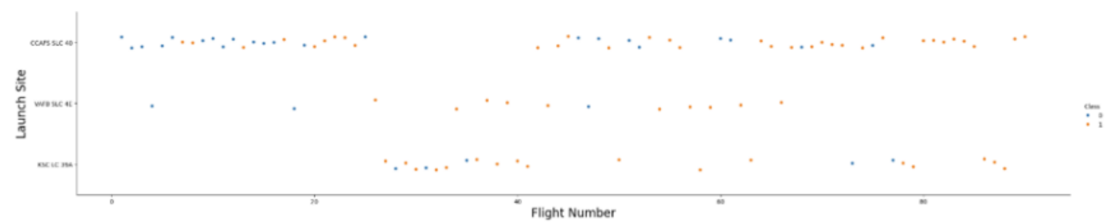
- Response from Webpage
- Create BeautifulSoup Object
- Find Relevant Tables & Columns
- Create dictionary object
- Convert to DataFrame object

Data Wrangling

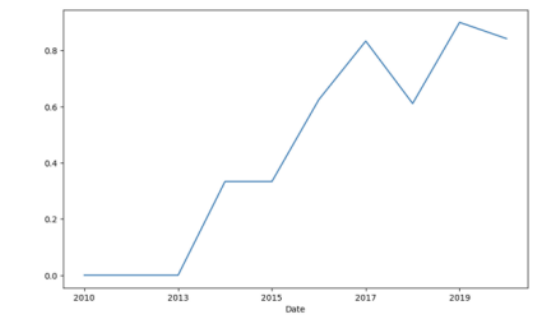
- Collecting Data using SpaceX API calls
- Reading the Data
- Checking Launch Sites and Orbit Counts
- Identifying Bad Outcomes
- Categorizing success and failure of landing into new column class

•EDA with Visualization

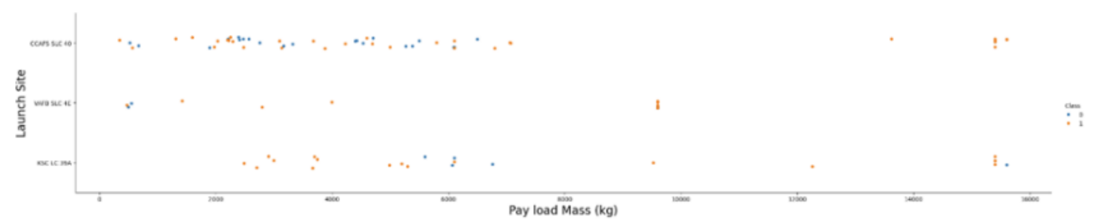
Relationship between Flight Number, Launch Site and landing success/failure class.



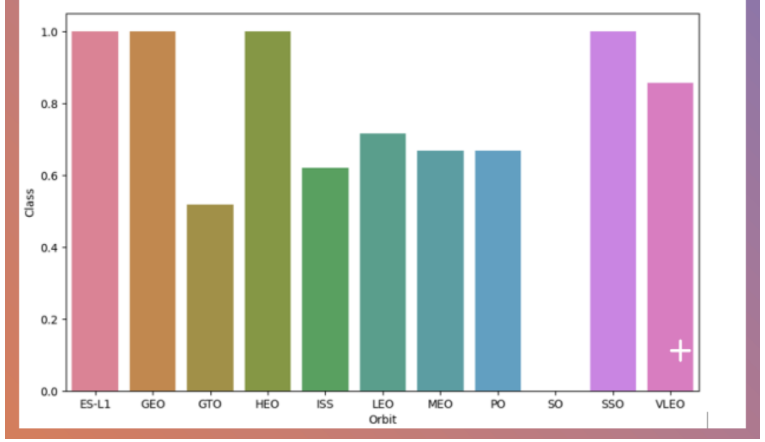
Launch success yearly trend.



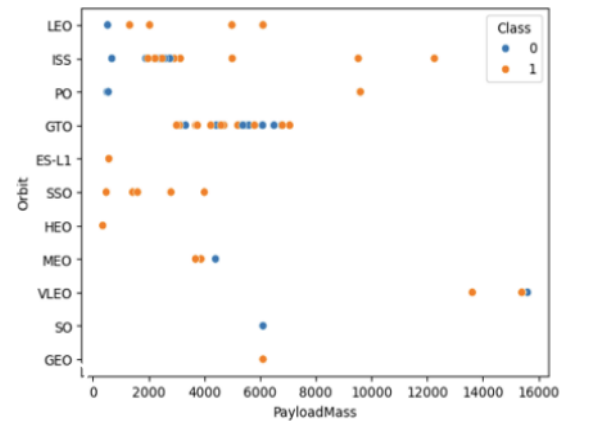
Relationship between Payload, Launch Site and landing success/failure class.



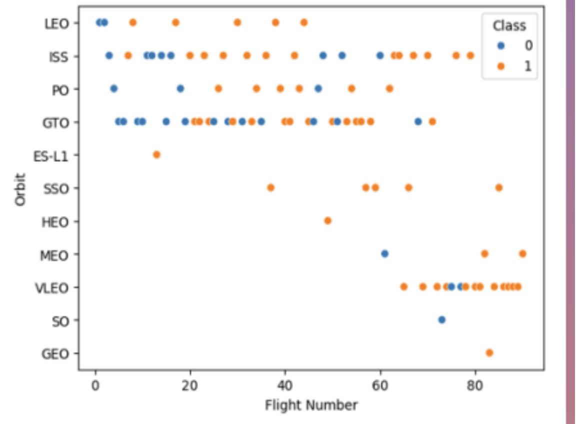
Relationship between Success rate and Orbit type



Relationship between Payload and Orbit Type



Relationship between Flight Number and Orbit Type



EDA with SQL Results

- Display the names of the unique launch sites in the space mission
- 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.
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months 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.

• Built an Interactive Map with Folium

- · Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map, using folium.Circle and folium.Marker
- · Assigned the feature launch outcomes (failure or success) to class 0 and 1. i.e., 0 for failure, and 1 for success, using the colorlabeled Marker Cluster , we identified which launch sites have relatively high success rate.
- · Calculated the distances between a launch site to its proximities. We answered some question for instance:
  - · Are launch sites near railways, highways and coastlines.
  - · Do launch sites keep certain distance away from cities.

Built a Dashboard with Plotly Dash

- · Built an interactive dashboard with Plotly Dash
- · Plotted pie charts showing the total successful launches by All sites or selected sites (selected from dropdown).
- · Plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version ranges (with dynamic payload range slider).

Predictive Analysis (Classification)

1

Building the model

- Create array Y for output class
- Standardize the data
- Split into training and testing set
- Build GridSeachCV model and fit the data

2

Evaluating the model

- Calculating the accuracies
- Calculating the confusion matrix
- Plot the results

3

Finding the optimal model

- Find the best hyperparameters for the model
- Find the best model with highest accuracy on testing data
- Confirm the optimal model

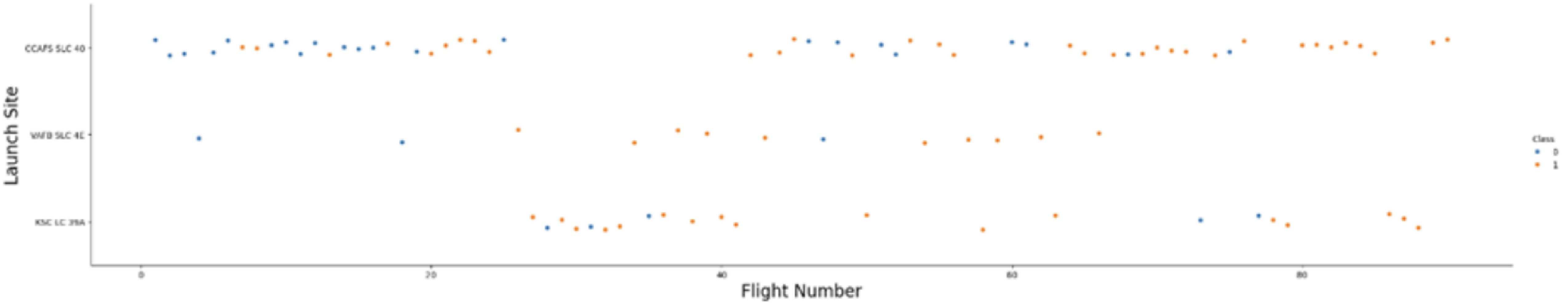
INSIGHTS DRAWN FROM EDA

Results

- The SVM, KNN and Logistic regression models are the best in terms of prediction accuracy
- · Low weighted payloads perform better than the heavier payloads
- · The success rates of SpaceX launches have significantly improved over the years
- · KSC LC 39A had the most successful launches from all the sites
- · Orbits GEO, HEO, SSO, ES L1 have the best success rates

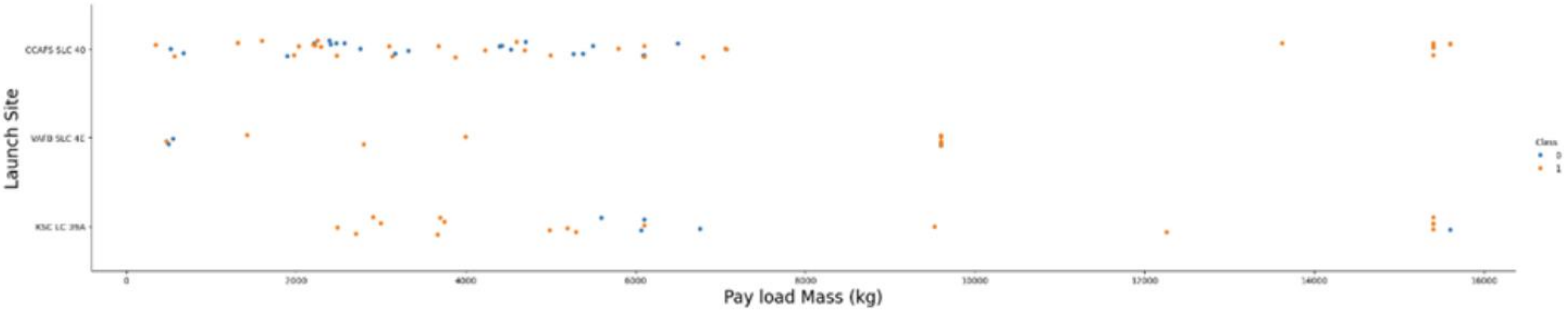
Flight Number v. Launch Site

- · Launches from the site of CCAFS SLC 40 are significantly higher than launches from other sites
- · With the increase in flight number, the success rates also increases



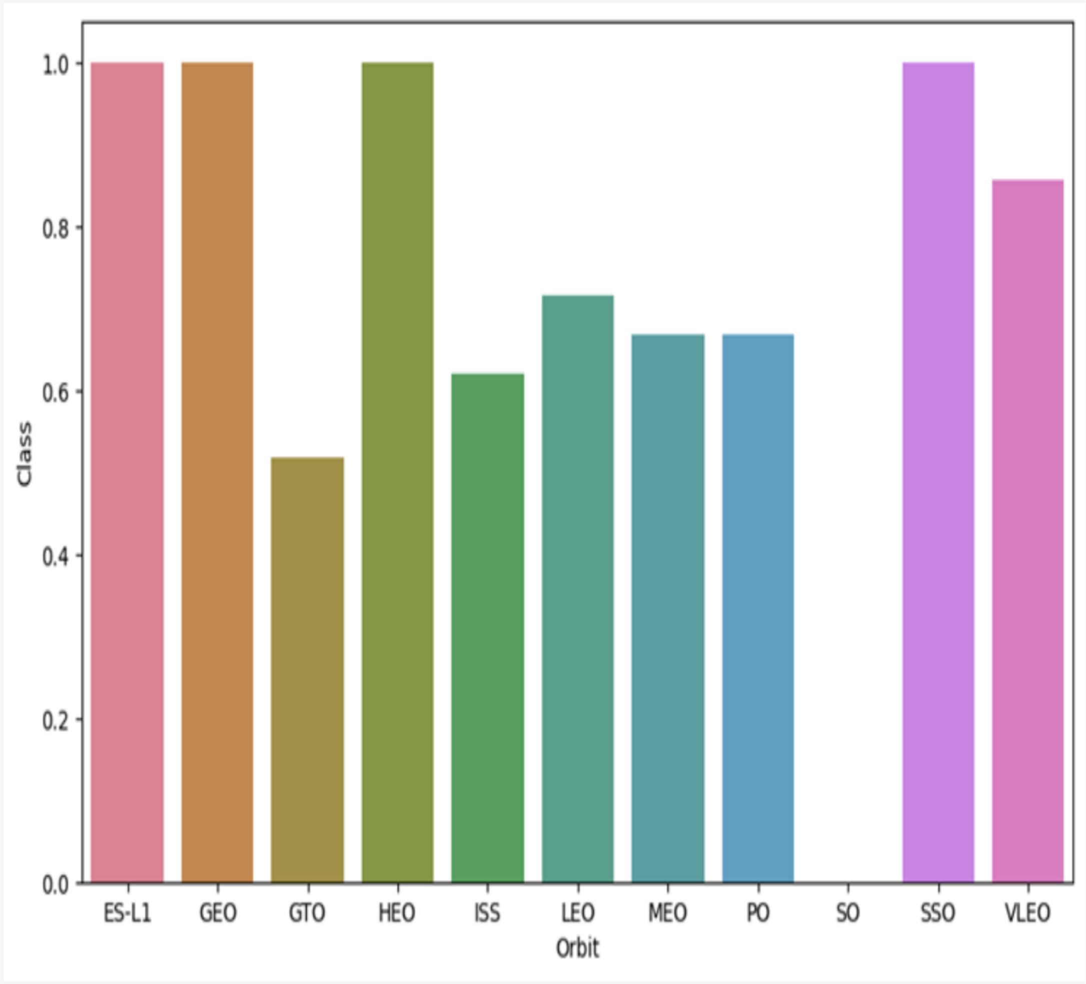
Payload v. Launch Site

- · Majority of Pay loads with lower Mass range have been launched from CCAFS SLC 40



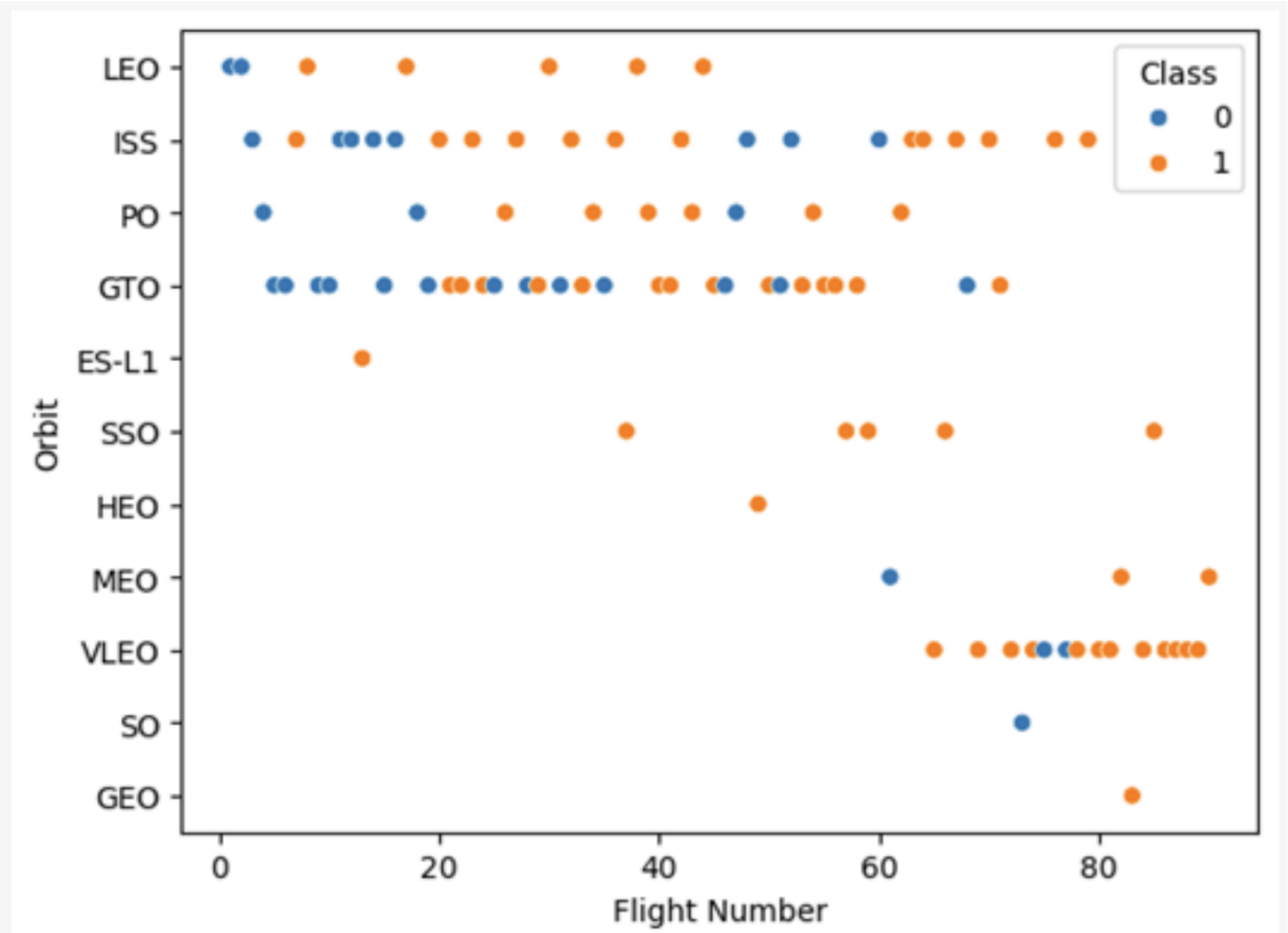
Success Rate v. Orbit Type

- · Following orbits have high success rate (100%)
- · ES-L1
- · GEO
- · HEO
- · SSO
- · SO has the least success (0%)



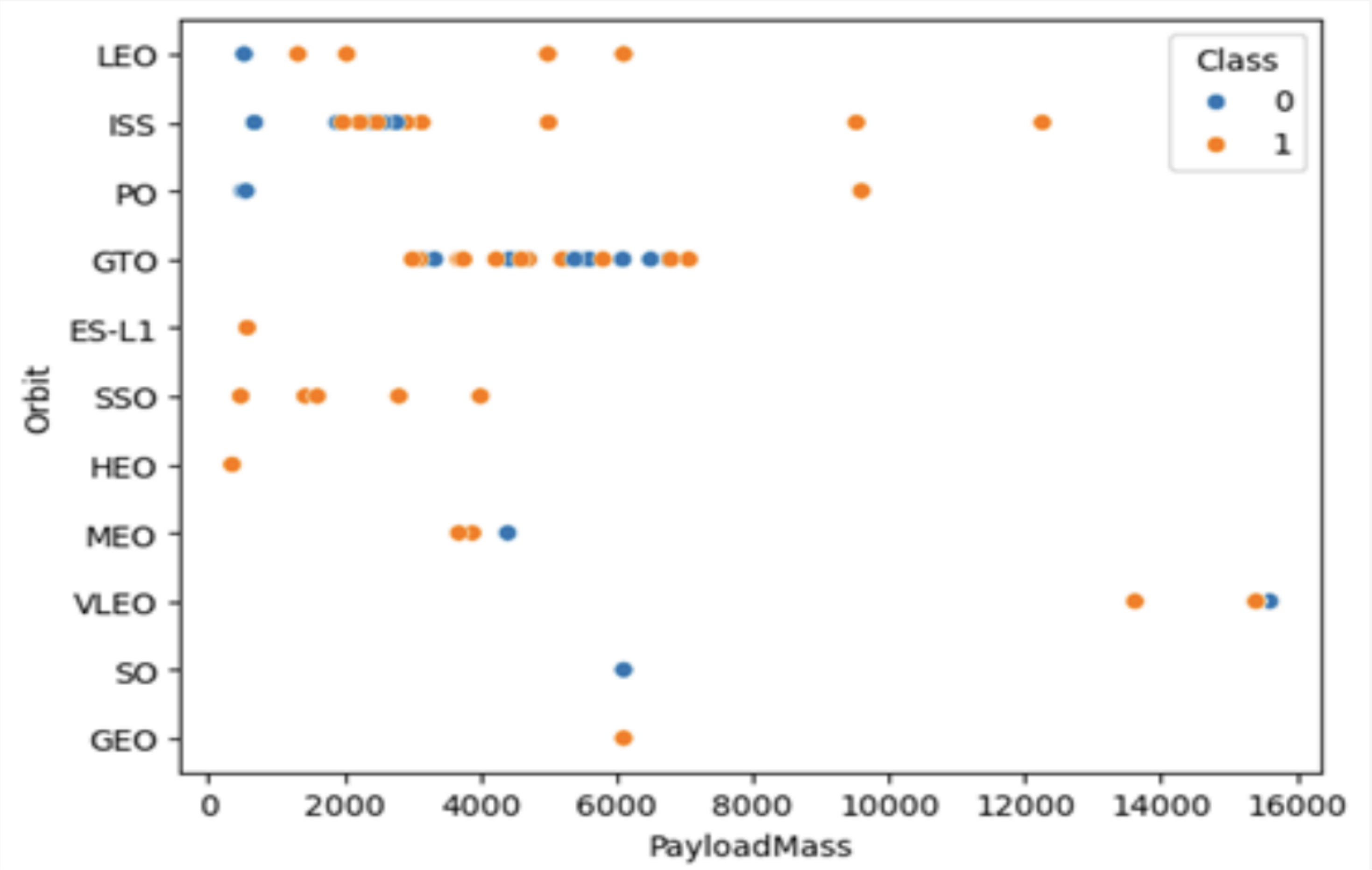
Flight Number v. Orbit Type

- · A trend can be observed where the latest launches have been shifted to VLEO orbit with good success rates



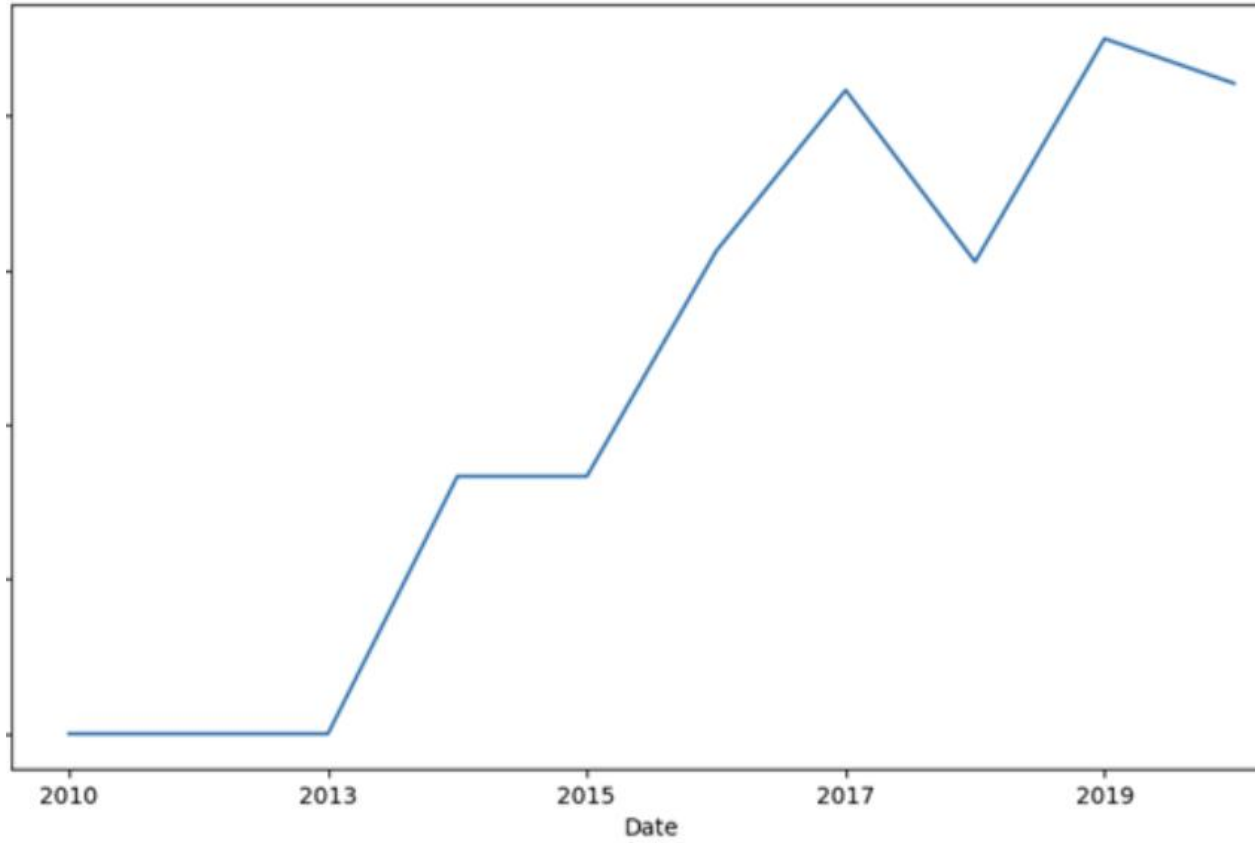
Payload v. Orbit Type

- · Co-relation between ISS orbit and payloads between range of 2000 and 4000 kgs.
- · Also a relationship between the GTO orbit and payload range of 3000-7000 kgs.



Launch Success Yearly Trend

- · Success rates are significantly increasing over the years, even with the dip and failures around 2018



All Launch Site Names

- Using DISTINCT keyword for listing unique names

```
[8]: %sql select distinct "Launch_Site" from SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

[8]:

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

Launch Site Names Begin with 'CCA'

- Used keyword LIMIT to display only 5 records

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' LIMIT 5
```

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

- Use the keyword SUM.

```
%sql select sum(PAYLOAD_MASS__KG_) as "Total_Payload_Mass" from SPACEXTABLE where "Customer"='NASA (CRS)';
```

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- Used the keyword AVG

```
%sql select avg(PAYLOAD_MASS__KG_) as "Average_Payload" from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%'
```

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

Average_Payload
2534.6666666666665

First Successful Ground Landing Date

- Used the keyword MIN on the date column with the appropriate condition to find the required parameter

```
%sql select min(Date) from SPACEXTABLE where "Landing_Outcome"="Success (ground pad)"
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Displaying list the names of boosters which have successfully landed on drone ship and had payload mass in between 4000 and 6000

```
%%sql
select distinct Booster_Version from SPACEXTABLE where "Landing_Outcome"="Success (drone ship)" and
PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * 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

· Used the query with GROUP BY feature · Results show 10 Failures and 61 Successful landing

```
%%sql
select sum(C) as "Total", Landing_Outcome from
    (select count(*) as C,SUBSTR(Landing_Outcome,1,7) as Landing_Outcome
        from SPACEXTABLE
        where (SUBSTR(Landing_Outcome,1,7) like ('Success%') or
                SUBSTR(Landing_Outcome,1,7) like ('Failure%'))
        group by Landing_Outcome)
    group by Landing_Outcome
```

* sqlite:///my_data1.db

Total	Landing_Outcome
10	Failure
61	Success

Boosters Carried Maximum Payload

- Using Sub-query to get the distinct list of booster version meeting the payload conditions

```
%%sql

select distinct Booster_Version from SPACEXTABLE
    where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE) order by 1
```

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

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

2015 Launch Records

- Matched the 2015 records using SUBSTR function in the where conditions

```
%%sql
select REPLACE(REPLACE(substr(Date,6,2),'01','January'),'04','April') as Month_Names,
        Landing_Outcome, Booster_Version, Launch_Site
        from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome='Failure (drone ship)'
```

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

Month_Names	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- · Using the RANK function on the descending order of count function used to count number of landing outcomes in the given date range.

```
%%sql
select RANK () OVER (
    ORDER BY count(*) desc) Rank, count(*) Count, Landing_Outcome from SPACEXTABLE
    where Date between '2010-06-04' and '2017-03-20'
    group by Landing_Outcome order by 1
```

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

Rank	Count	Landing_Outcome
1	10	No attempt
2	5	Success (drone ship)
2	5	Failure (drone ship)
4	3	Success (ground pad)
4	3	Controlled (ocean)
6	2	Uncontrolled (ocean)
6	2	Failure (parachute)
8	1	Precluded (drone ship)

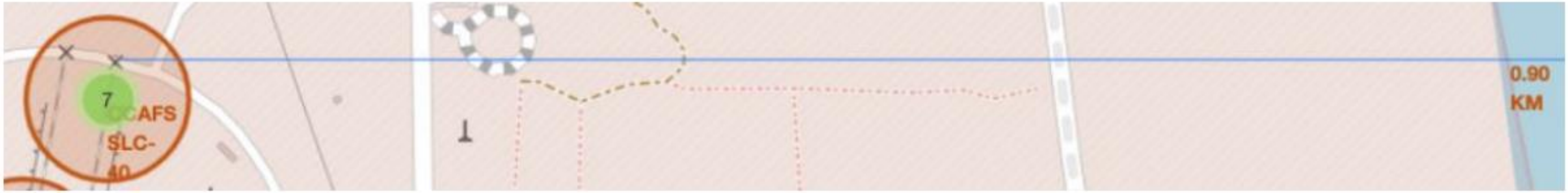
LAUNCH SITES PROXIMITIES ANALYSIS

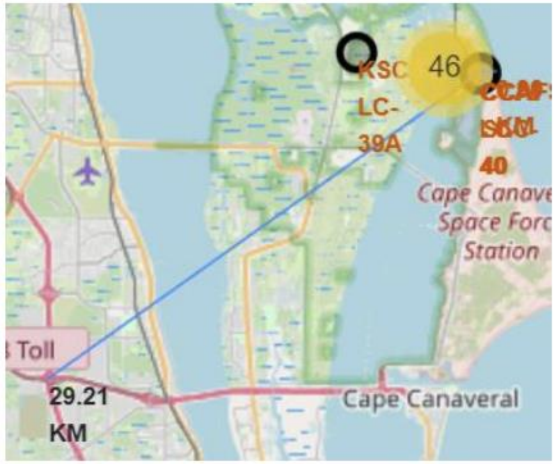
Launch Sites for SpaceX

- All SpaceX rockets are launched from the coastal lines of Florida and California in the United States of America

Distances between a launch sites to its proximities

- All infrastructures like, railway line, highway, coastal line and nearest city are in close proximities

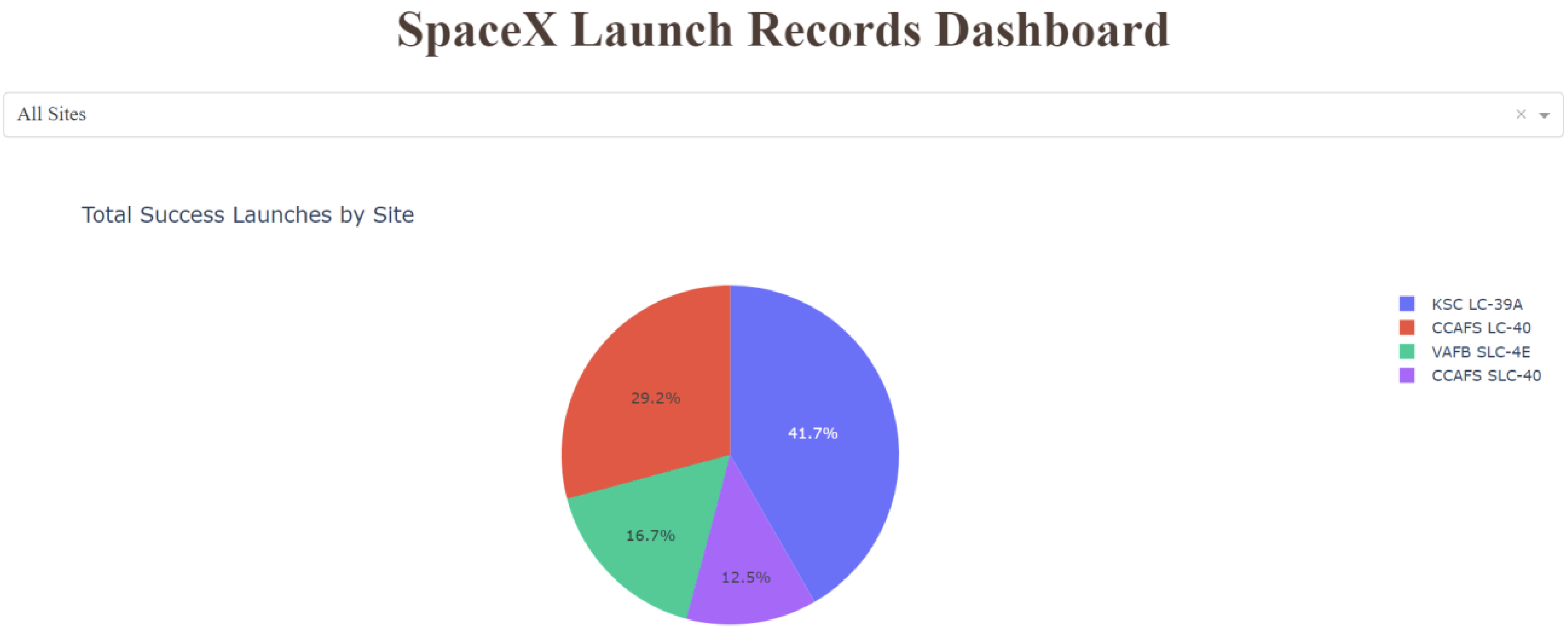




BUILD A DASHBOARD WITH PLOTLY DASH

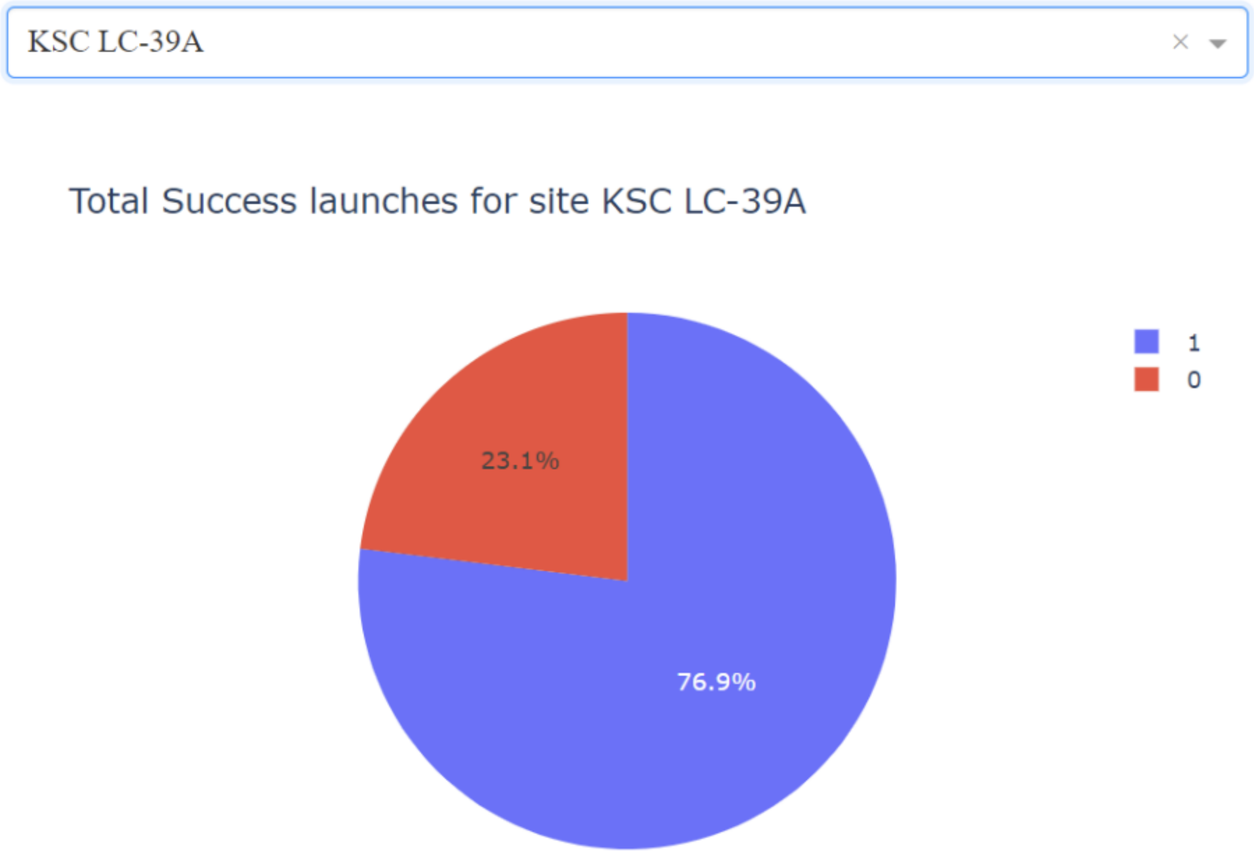
Total success Launches by all sites

- · KSC LC-39A has had the most successful launches in comparison to all the launch sites.



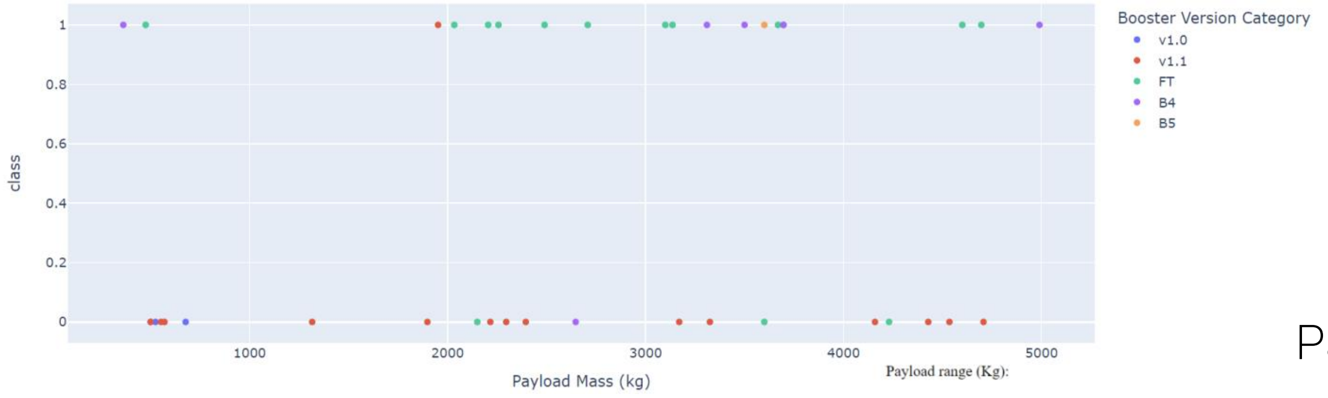
Success Rate by Site (KSC LC-39A)

- · KSC LC-39A has a 76.9% success rate as a launch site

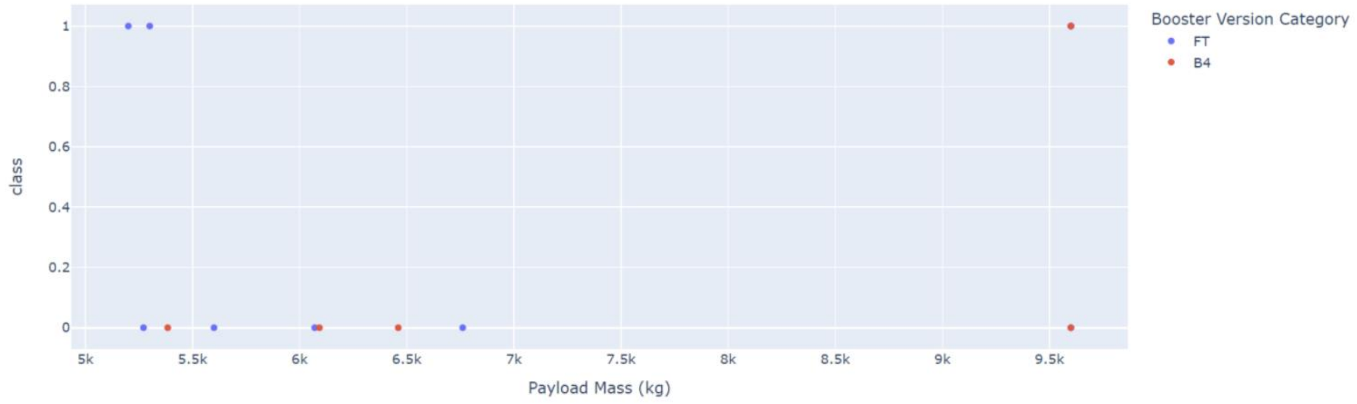


Success Rates on Variable Payload using Dash Rangeslider

Payload (0-5000kg) range for all sites



Payload (5000-10000kg) range for all sites



PREDICTIVE ANALYSIS (CLASSIFICATION)

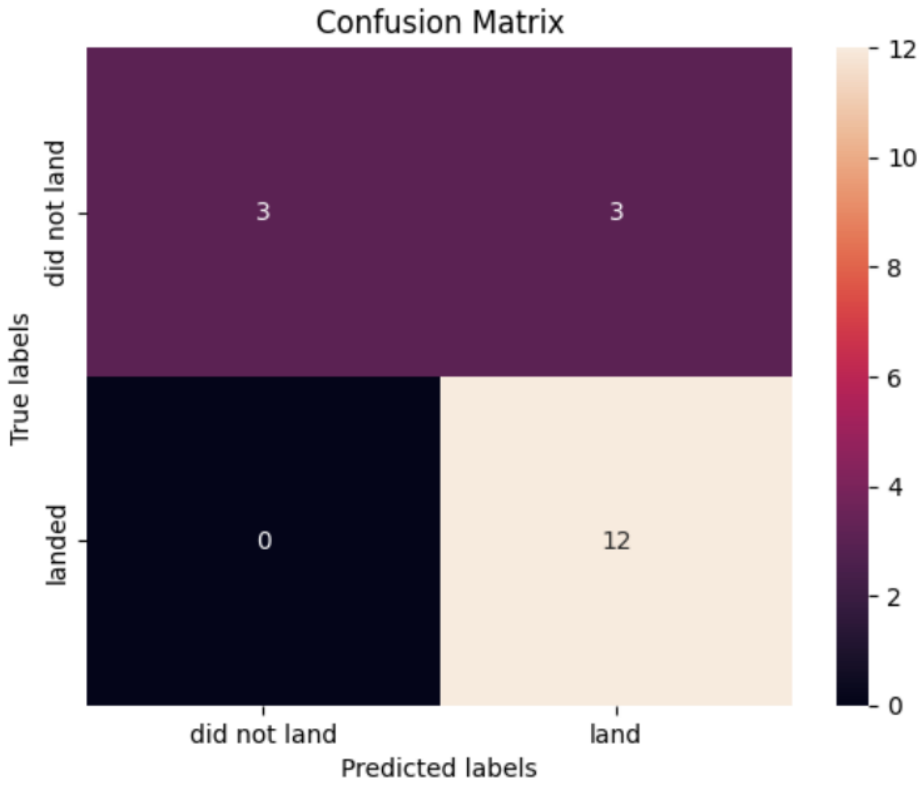
Classification Accuracy

```
print("Logistic Regression:",logreg_cv.score(X_test,Y_test))
print("SVM:",svm_cv.score(X_test,Y_test))
print("Decision Tree Classifier:",tree_cv.score(X_test,Y_test))
print("K-Nearest Neighbours:",knn_cv.score(X_test,Y_test))
```

```
Logistic Regression: 0.8333333333333334
SVM: 0.8333333333333334
Decision Tree Classifier: 0.8333333333333334
K-Nearest Neighbours: 0.8333333333333334
```

Confusion Matrix of KNN Model

Matrix shows that inaccuracies in the model result in displaying False negatives, but never result in False positives



Conclusions

- In summary:
- · The launch success rate exhibited a consistent increase from 2013 to 2020.
- · Lighter payloads demonstrated higher success rates compared to heavier payloads.
- · Orbits such as ES-L1, GEO, HEO, SSO, and VLEO achieved the highest success rates.
- · KSC LC-39A emerged as the site with the highest number of successful launches.
- · The K-Nearest Neighbor algorithm proved to be the most effective for this specific task.

☺ Thank you!