IBM DATA SCIENCE CAPSTONE - SPACEX PROJECT

APARNA V

Date : 10-01-2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- **Methodologies**
  - Data Collection and Data Wrangling
  - Exploratory Data Analysis
  - Visual analytics and Dashboard
  - Machine Learning Prediction

- **Results**
  - EDA results
  - Interactive analytics and data visualizations
  - Predictive analysis results

INTRODUCTION

- **Project background and Context**
  - The SpaceX spacecraft manufacturing company launches its Falcon 9 rockets with a cost of 62 million dollars, as advertised by the company on its website. Meanwhile, all the other providers perform the same with a cost of around 165 million dollars each. The reduced cost of the SpaceX rocket launches is due to the reuse of the first stage. Determining whether the first stage will land successfully or not will help any other company in bidding against SpaceX.
- **What needs to be solved?**
  - Which factors influence the successful landing of the first stage?
  - What is the effect of the relationship of each rocket variables on the outcome?
  - What measures does SpaceX carry out in order to achieve the best landing success rate.

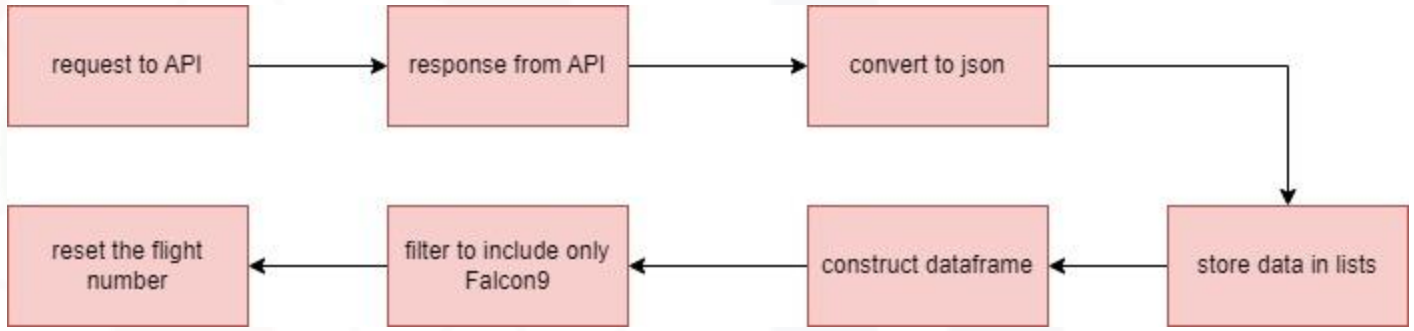
METHODOLOGY

- **Data Collection**
  - Using SpaceX REST API
  - Web scrapping from Wikipedia
- **Data wrangling**
  - One hot encoding data fields
  - Dropping irrelevant columns
- **EDA and Visualization**
  - EDA using SQL
  - Finding patterns between data using scatter plots and bar graphs
- **Interactive visual analytics**
  - Using Folium
  - Interactive dashboard using Plotly dash
- **Predictive analysis**
  - Create and evaluate classification models
  - Find the best model

METHODOLOGY

Data Collection via API

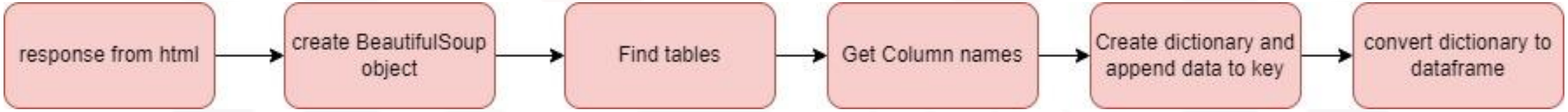
```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```



SpaceX Falcon9 launch data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	BoosterVersion2
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857	True
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857	True
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857	True
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093	True
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857	True

Data Collection-Web scrapping




```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response)
```

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables=soup.find_all('table')
html_tables
```

```
df=pd.DataFrame.from_dict(launch_dict, orient='index')
df=df.transpose()
df.head()
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	F9 v1.0B0003.1	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Success\n	4 June 2010	18:45
1	1	F9 v1.0B0004.1	Dragon	0	LEO	NASA	Success	F9 v1.0B0003.1	Success	4 June 2010	18:45
2	1	F9 v1.0B0005.1	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0003.1	Success	4 June 2010	18:45
3	1	F9 v1.0B0006.1	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0004.1	Success\n	4 June 2010	18:45
4	1	F9 v1.0B0007.1	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0005.1	Success\n	4 June 2010	15:43

Data Wrangling

- Number of launches on each site

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

- Number and occurrence of each orbit

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

- Create a landing outcome label from outcome column

```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

EDA Data Visualization

- Scatter plots
  - Flight number vs. Payload mass
  - Flight number vs. Launch Site
  - Payload vs. Launch Site
  - Flight number vs. Orbit type
  - Payload vs. Orbit type
- Bar graph
  - Plot of success rate by class of each orbits
- Line plot
  - Plot of launch success yearly trend

EDA with SQL

- 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 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 successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

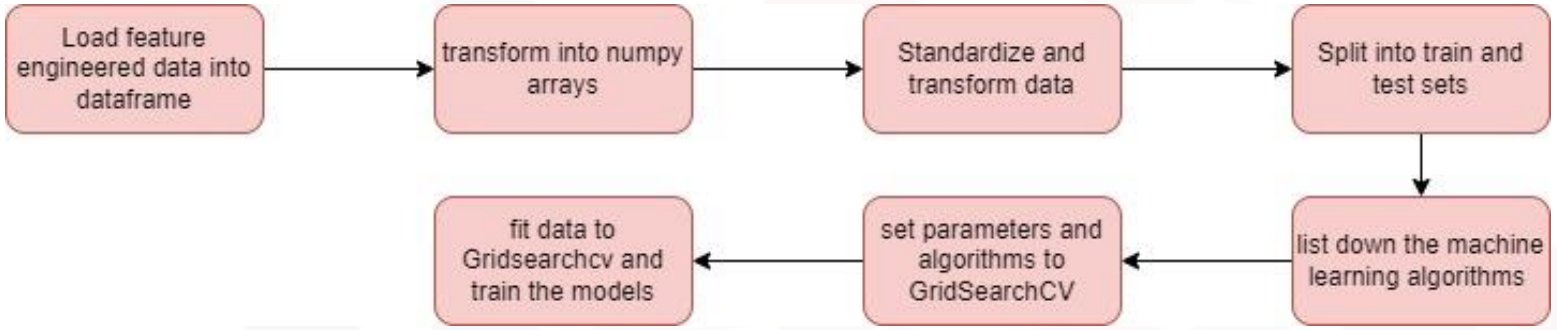
Interactive map with Folium

- Mark all launch sites on the map
- Mark the success/failed launches for each site on the map
- Calculate the distance between a launch site to its proximities

Interactive dashboard with Plotly

- Pie chart showing the percentage of success in relation to launch site
- Scatter graph showing correlation between payload and success for all sites or by certain launch sites.

Machine Learning prediction



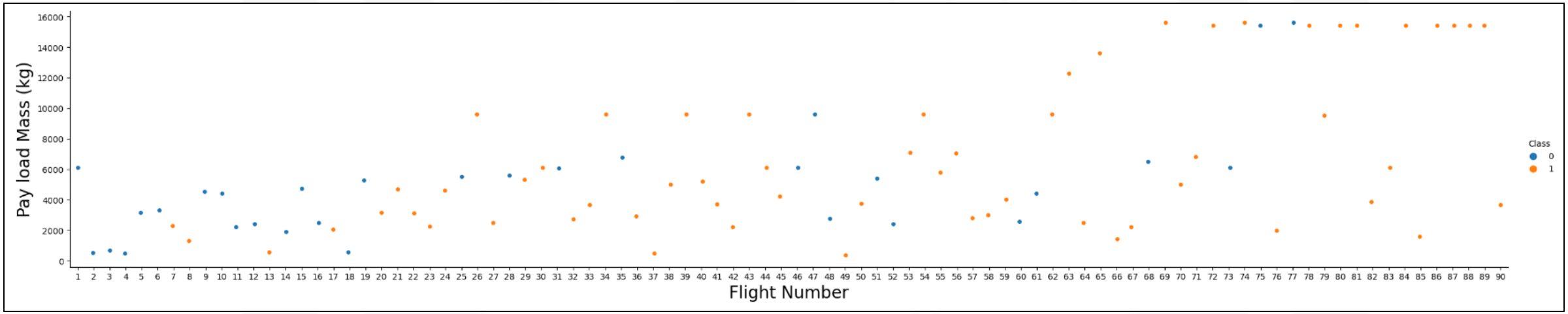
- Model Evaluation
  - Check accuracy for each model
  - Find the best parameters for each model
  - Plot confusion matrix

- Best model
  - The model with the best accuracy is the best model

RESULTS

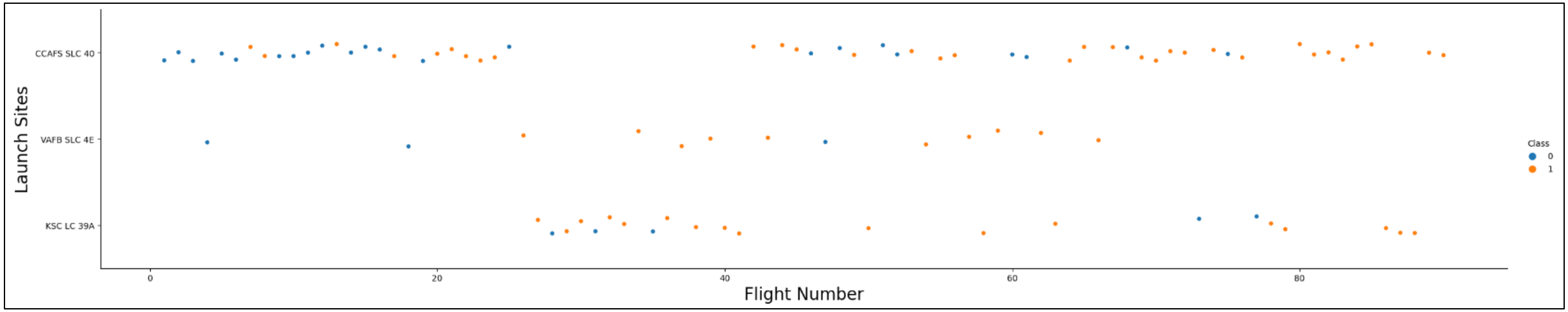
EDA with Visualization

- Flight Number vs. Payload mass



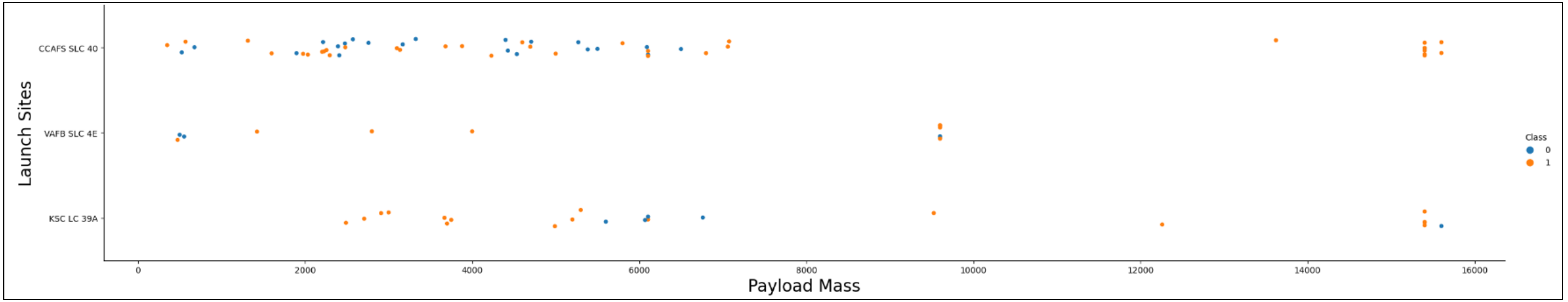
The more massive the payload, the less likely the first stage will return.

Flight number vs. Launch site



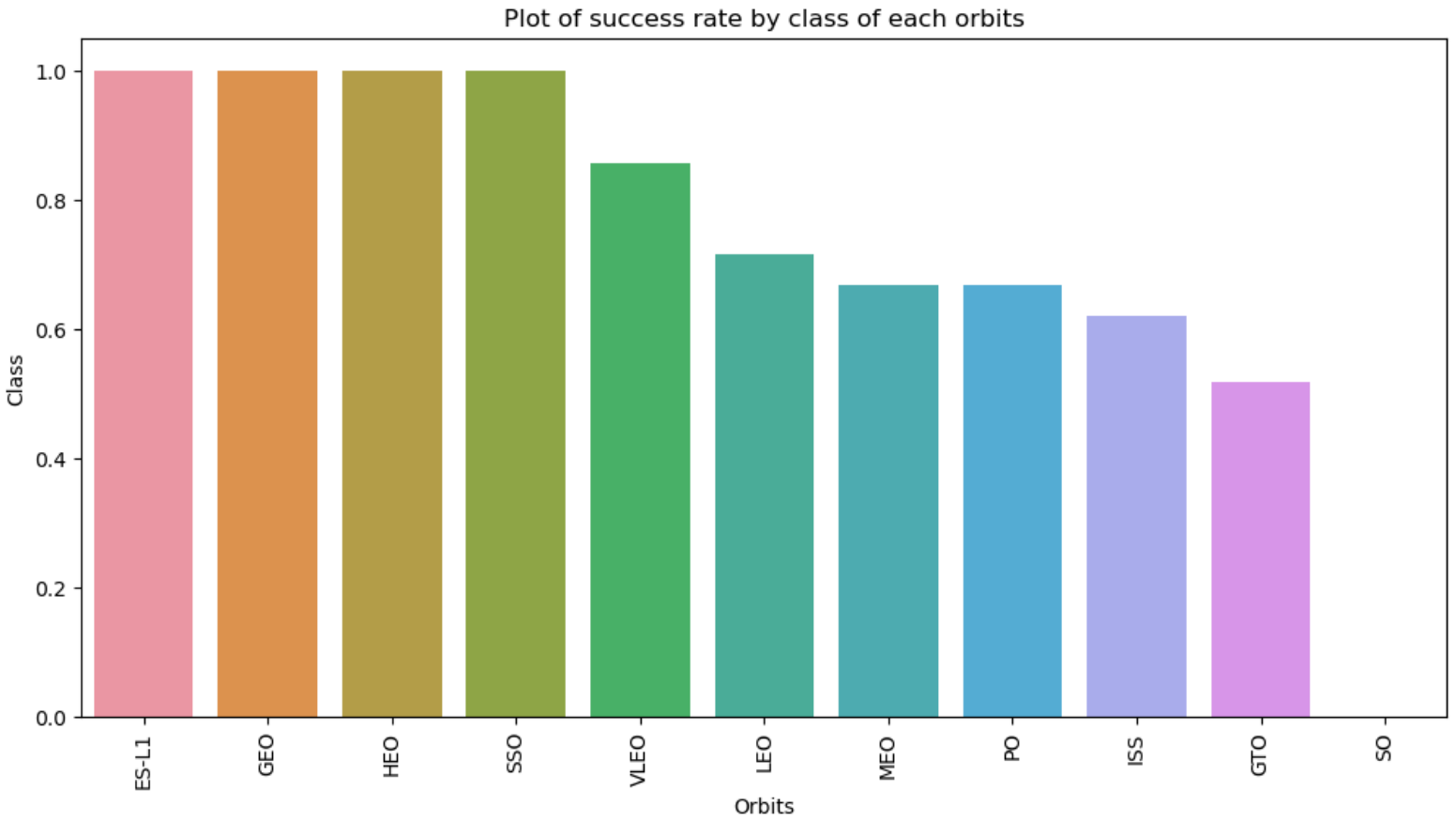
The success rate increases with higher flight numbers

Payload mass vs. Launch site



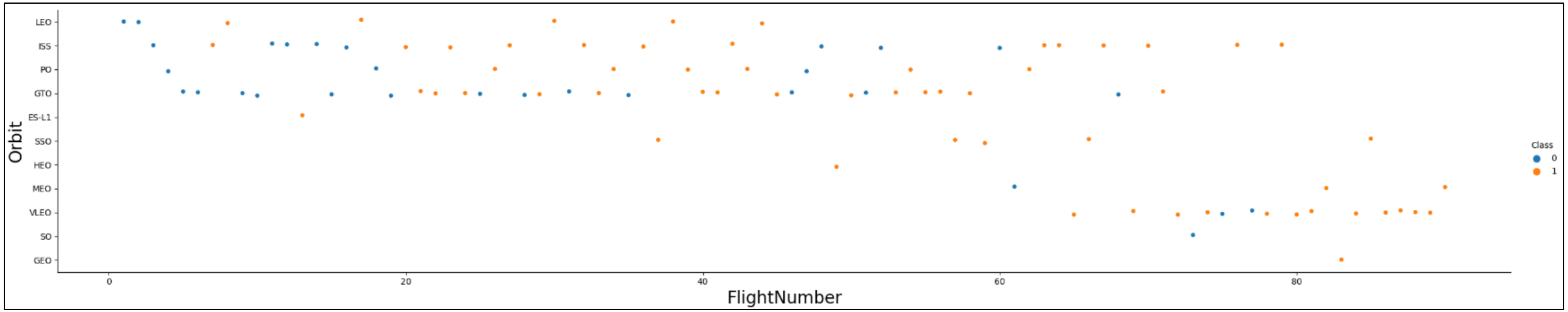
The greater the payload mass, higher the success rate.

Plot of success rate by class of each orbits



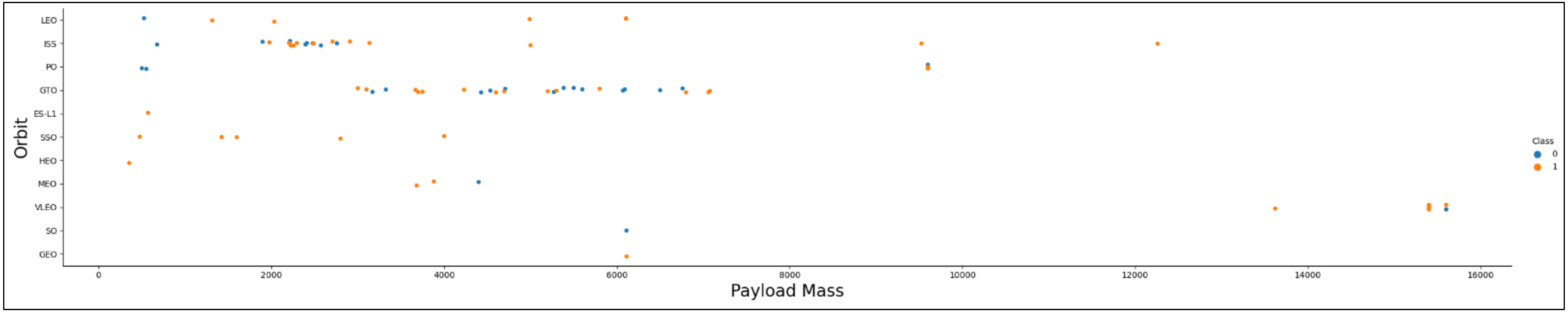
ES-L1, GEO, HEO and SSO have the highest success rates.

Flight number vs. Orbit



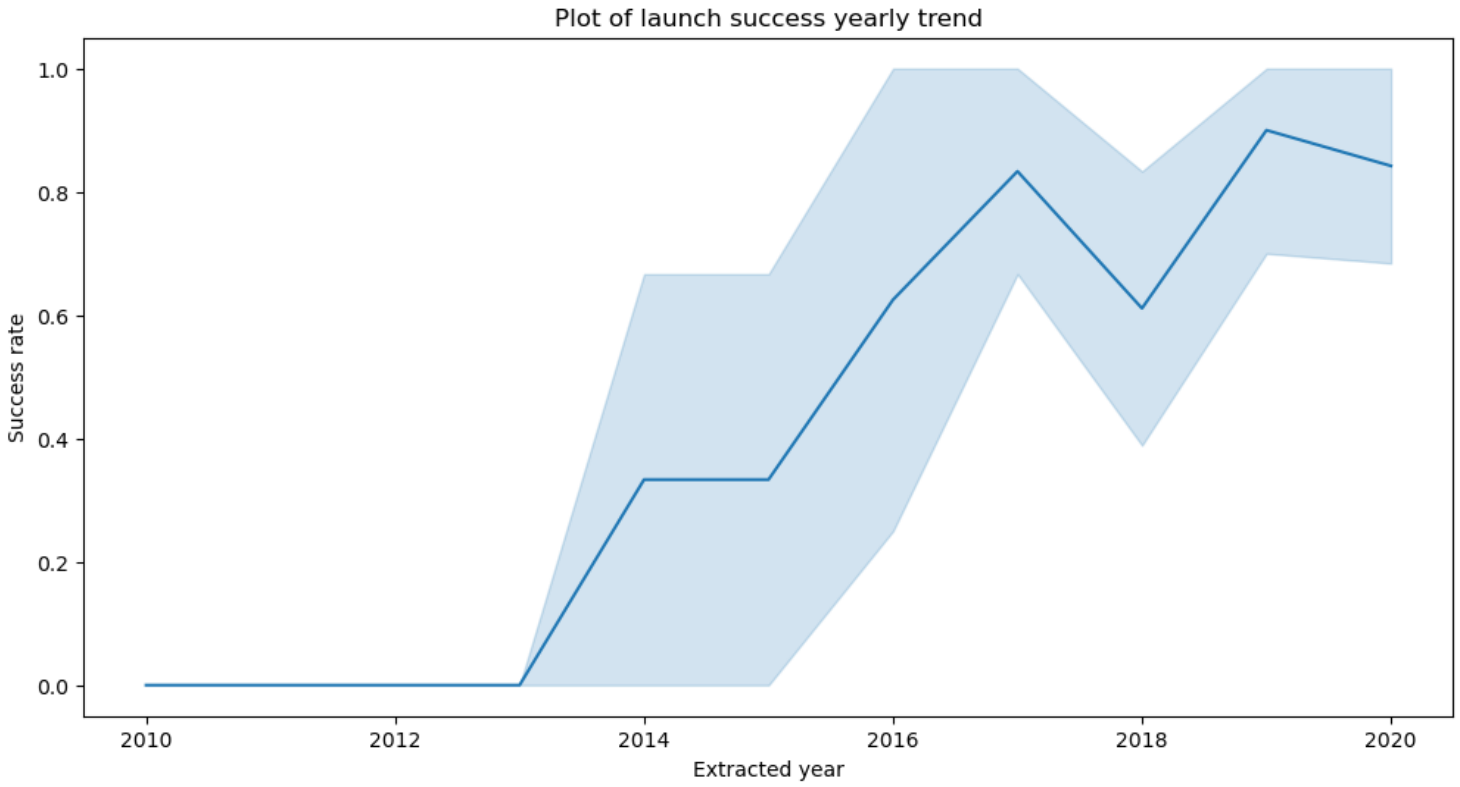
For LEO, the success rate increases with the flight number.
There is no significant relationship for any other orbit, with the flight number.

Payload mass vs. orbit



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch success yearly trend



The success rate kept increasing relatively from 2013 till 2020, though there is a dip at 2017 and 2019.

EDA SQL results

- All launch sites

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

```
%sql SELECT DISTINCT Launch_Site as "LaunchSites" FROM SPACEXTBL;
```

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

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

- Payload mass launched by NASA(CRS)

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Launch site names beginning with 'CCA'

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

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

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

Average payload mass carried by booster version F9 v1.1

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

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

Date when first successful landing outcome in ground pad was achieved

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE SPACEXTBL.'Landing _Outcome'=='Success (ground pad)';
```

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

MIN(Date)
01-05-2017

Names of boosters which have success in payload mass 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 SPACEXTBL.'Landing _Outcome' = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 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 failed mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "SUCCESS COUNT" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

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

SUCCESS COUNT
100

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "FAILURE COUNT" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%';
```

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

FAILURE COUNT
1

```
%sql SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%' OR MISSION_OUTCOME LIKE 'Failure%';
```

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

COUNT(MISSION_OUTCOME)
101

Names of the booster versions which have carried the maximum payload mass.

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

Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

Date	Landing _Outcome	Booster_Version	Launch_Site
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

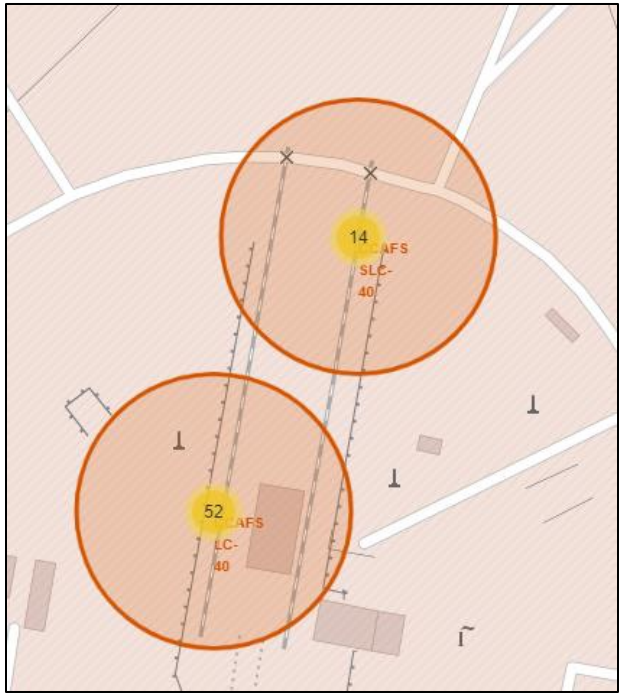
Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

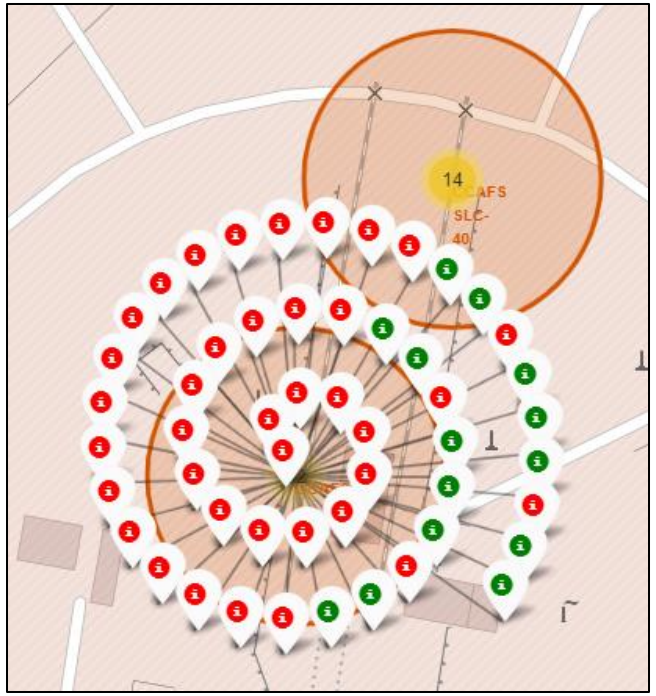
Landing _Outcome	TOTAL COUNT
Success	20
Success (drone ship)	8
Success (ground pad)	6

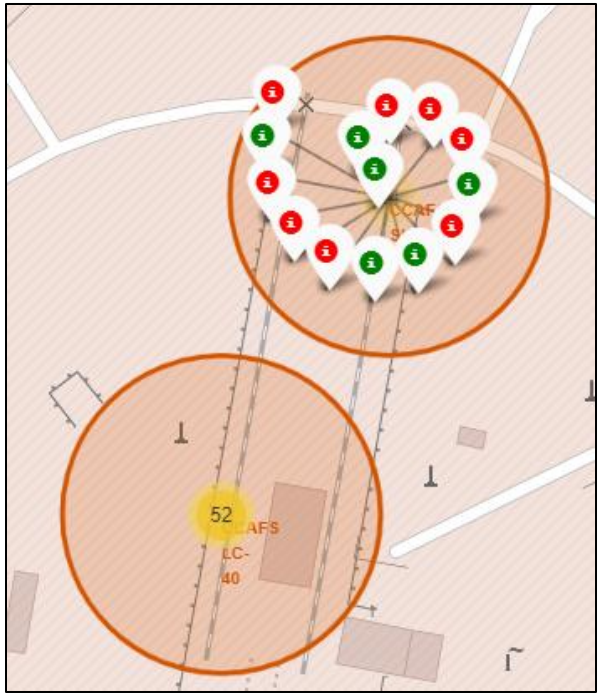
Interactive map with Folium

- Global map markers
- The SpaceX launch sites are located on the coasts of USA-Florida and California.

Color markers

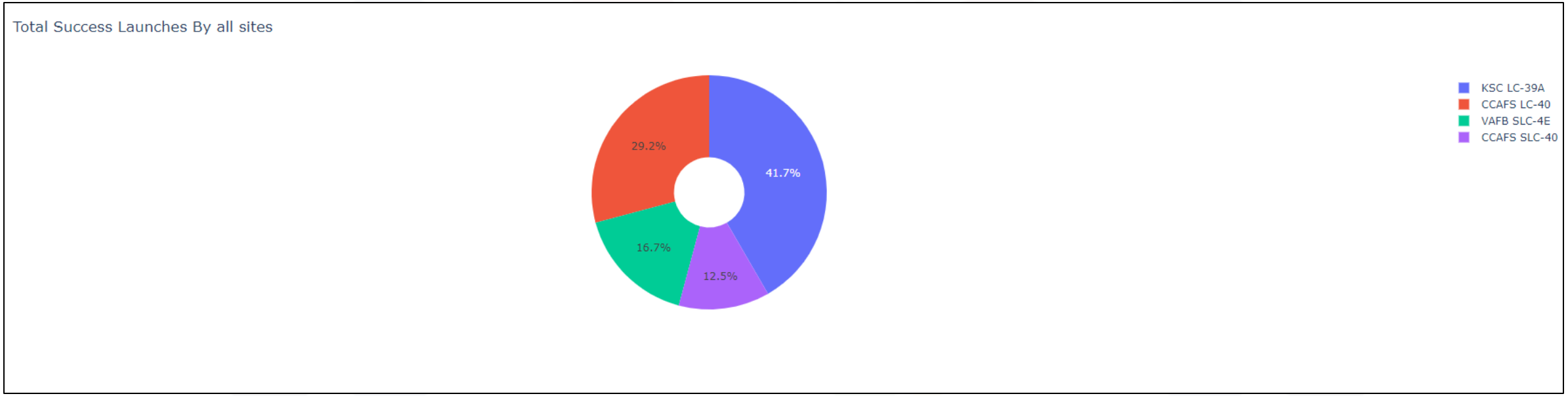






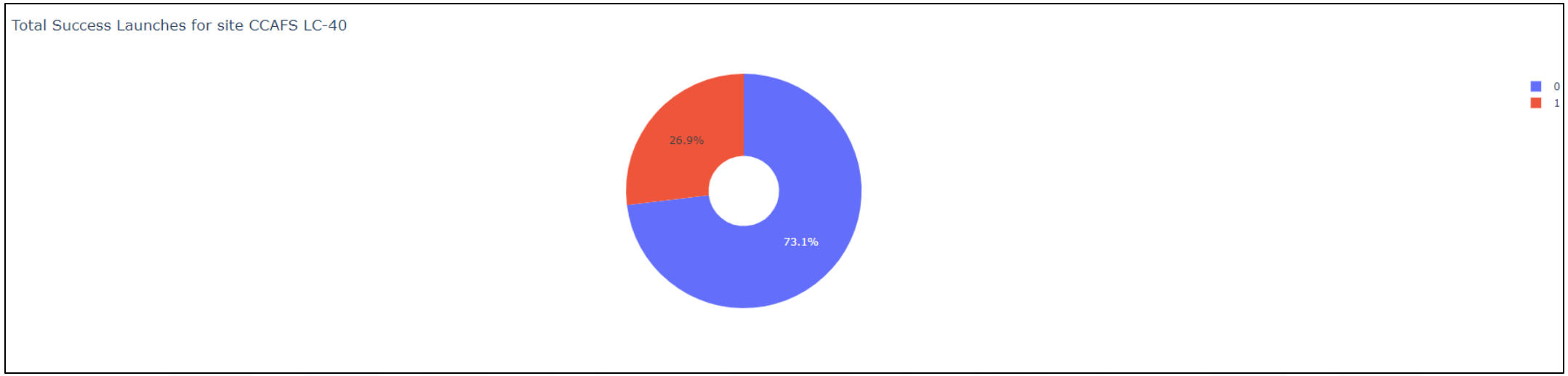
The green markers indicate successful landings and the red markers indicate failed landings.

Interactive Dashboard with Plotly

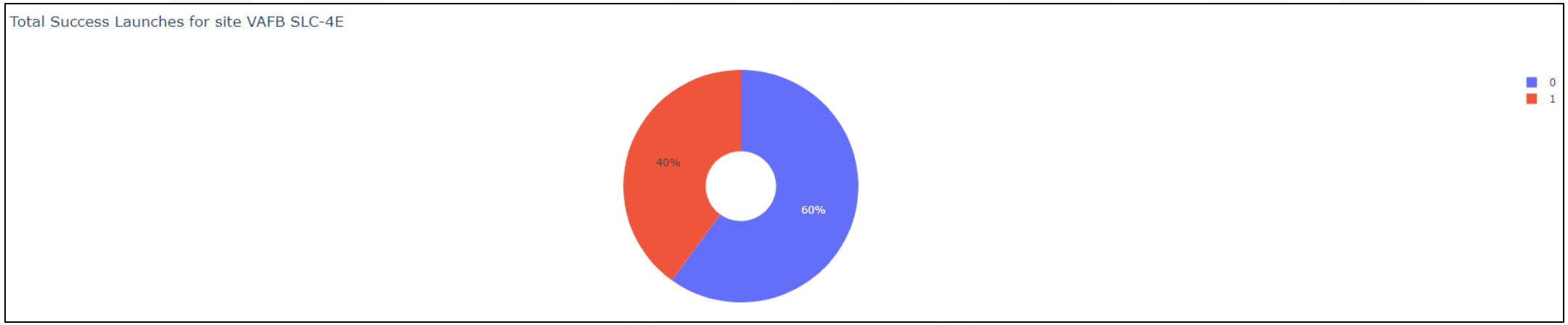


Pie chart showing the successful launch percentage by each launch site
The launch site KSC LC-39A has the highest number of successful launches.

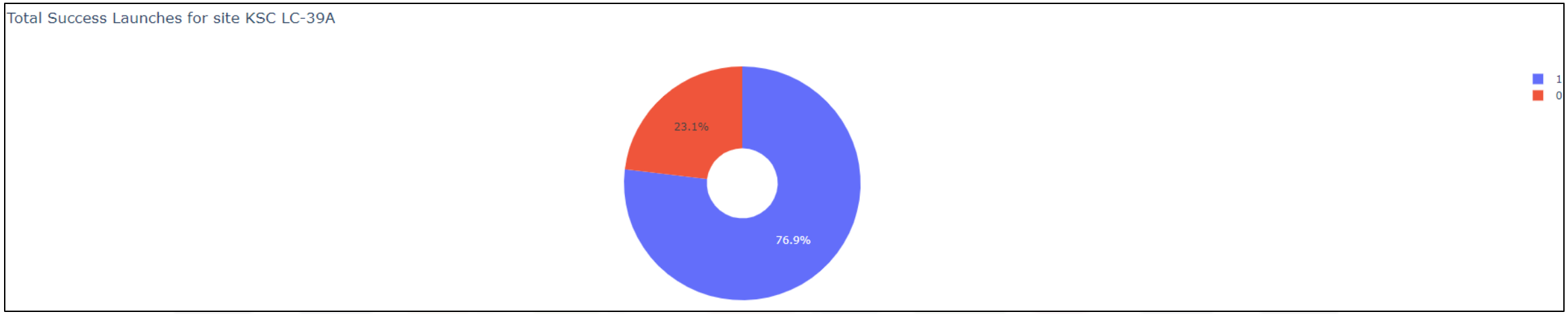
Success launch ratio for CCAFS LC-40



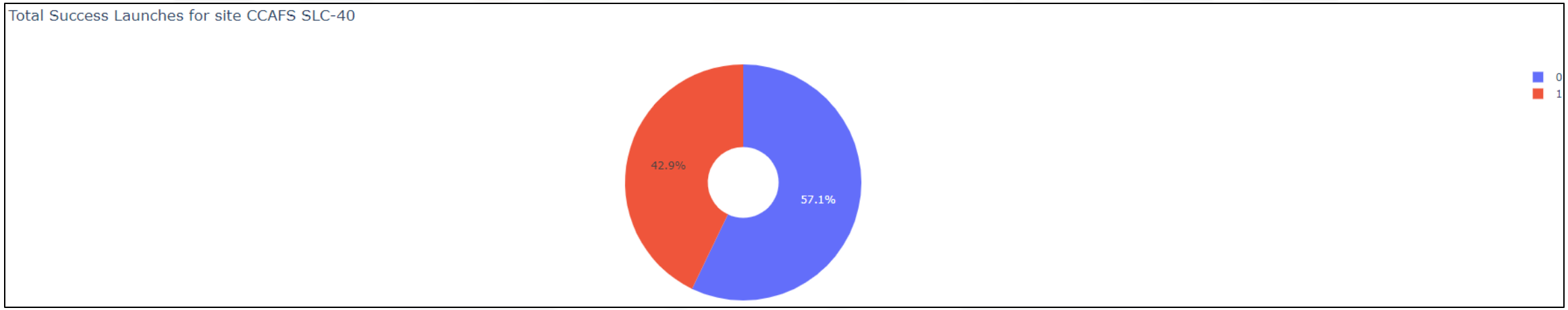
Success launch ratio for CCAFS LC-40



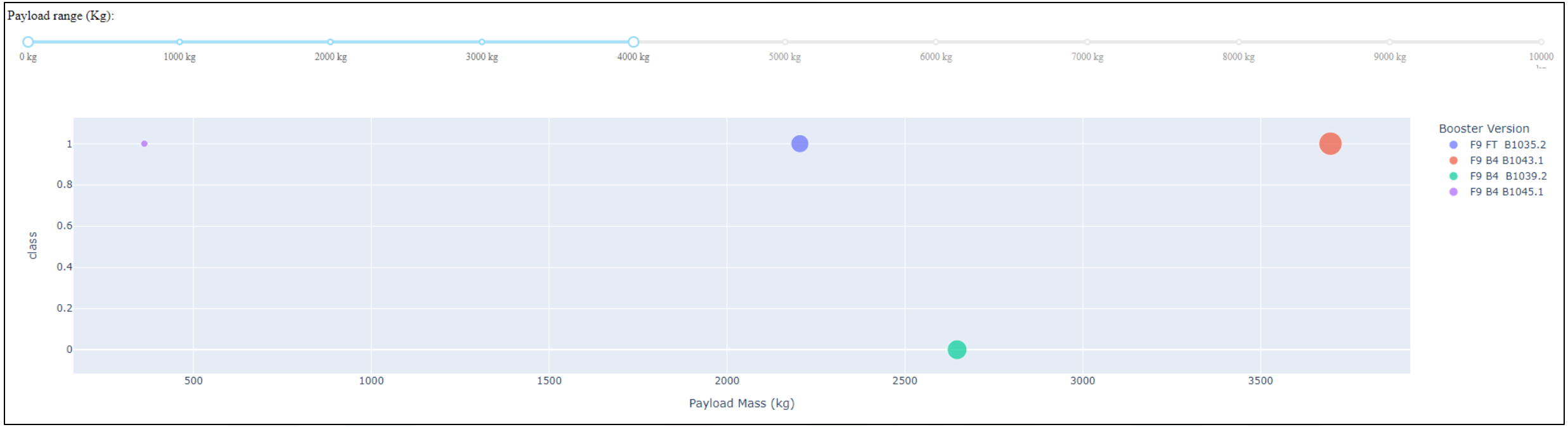
Success launch ratio for KSC LC-39A



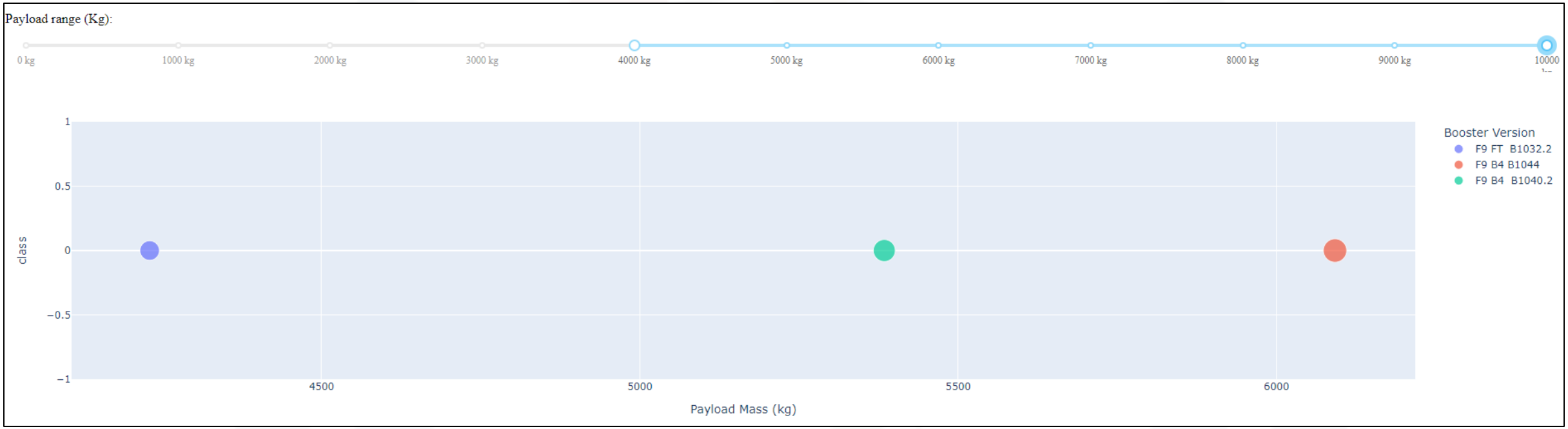
Success launch ratio for CCAFS SLC-40



Scatter Plots-Payload vs. Launch Outcome



Low payload range : 0kg to 4000kg



High payload range : 4000kg-10000kg

Predictive Analysis-Classification

- Algorithms used : KNN, Decision Tree, SVM, Logistic Regression
- Accuracy obtained for each model:

S.No.	Model	Accuracy
1	KNN	0.84821
2	Decision Tree	0.89107
3	SVM	0.84821
4	Logistic Regression	0.84642

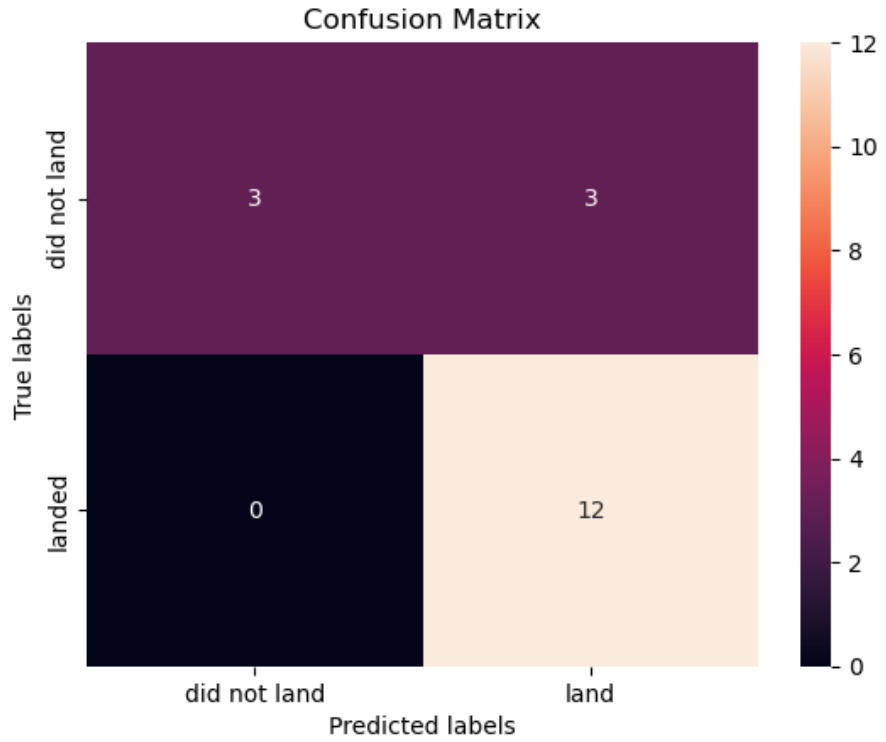
- Thus, Decision tree is the best model as it has the highest accuracy.

- The best parameters for the tree were found :

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
accuracy : 0.8910714285714285
```

- The accuracy of the model on the test data was found to be 83.33%
- The confusion matrix for the decision tree classification:




CONCLUSION

- Orbits GEO, HEO, SSO, ES-L1 has the best success rates
- The launch site KSC LC-39A has the most successful launches
- From the line plot, it is evident that the success rate increases with increase in time
- The low weighted payloads has higher success rate than high weighted payloads
- The Decision tree model is the best algorithm for this modelling with an accuracy of 83.33%.