IBM

SPACE X FALCON 9 LANDING PREDICTION

Presentation by Huda Basit

OUTLINE

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

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGIES

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

• SUMMARY OF ALL RESULTS

- Exploratory Data Analysis results
- Interactive analytics demo
- Predictive analysis results

INTRODUCTION

Project Background and Context:

SpaceX stands out as the foremost player in the commercial space industry, revolutionizing space travel with its cost-effective approach. Notably, the company offers Falcon 9 rocket launches at a significantly lower price point of 62 million dollars compared to other providers, whose costs can soar to 165 million dollars per launch. A substantial portion of these savings stems from SpaceX's groundbreaking ability to reuse the first stage of its rockets. Hence, predicting the success of first stage landings is pivotal in determining the overall cost of a launch. Leveraging publicly available data and employing machine learning models, our objective is to forecast whether SpaceX will successfully reuse the first stage.

Questions to Be Addressed:

1. How do variables such as payload mass, launch site, number of flights, and orbits influence the likelihood of a successful first stage landing?
2. Is there a discernible trend indicating an increase in the rate of successful landings over the years?
3. Which algorithm yields the most optimal performance for binary classification in this scenario?

METHODOLOGY

Data Collection Approach:

- Utilizing the SpaceX Rest API
- Employing Web Scraping techniques from Wikipedia
- Conducted Data Wrangling
- Filtering and refining the dataset
- Addressing missing values appropriately
- Employing One Hot Encoding to facilitate binary classification
- Conducted Exploratory Data Analysis (EDA) utilizing visualization tools and SQL
- Conducted Interactive Visual Analytics utilizing Folium and Plotly Dash
- Conducted Predictive Analysis employing classification models
- Developing, refining, and assessing classification models to ascertain optimal outcomes

DATA COLLECTION API

Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

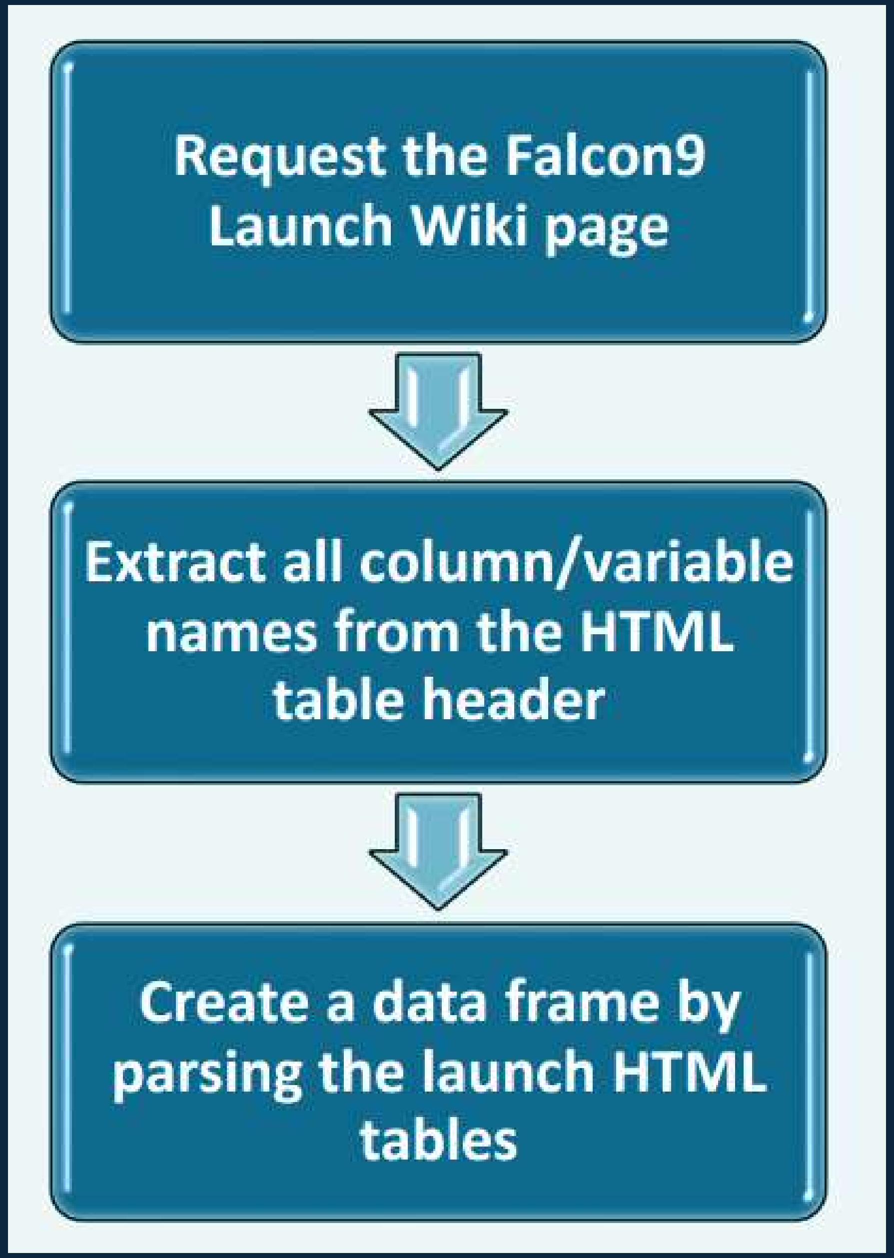
SpaceX offers a public API from where data can be obtained and then used; this API was used according to the flowchart beside and then data is persisted.

Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/data-collection-api.ipynb

DATA COLLECTION SCRAPING

Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

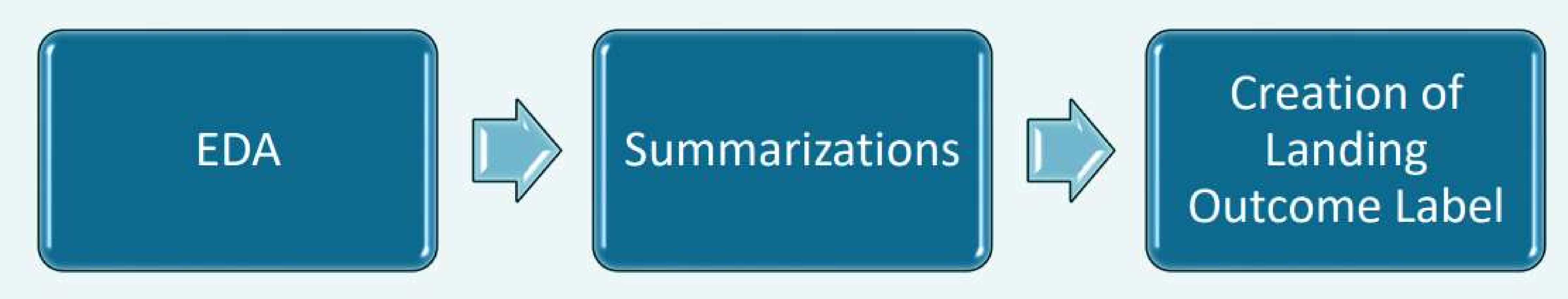
Data from SpaceX launches can also be obtained from Wikipedia; Data are downloaded from Wikipedia according to the flowchart and then persisted.



Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/webscraping.ipynb

DATA WRANGLING

Within the dataset, various scenarios arise where the booster fails to land successfully. On certain occasions, landing attempts may fail due to accidents. For instance, "True Ocean" denotes a successful landing in a specific oceanic region, while "False Ocean" signifies an unsuccessful landing in the same region. Similarly, "True RTLS" indicates a successful landing on a ground pad, whereas "False RTLS" denotes an unsuccessful attempt. "True ASDS" signifies a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful landing on such a vessel. These outcomes are predominantly converted into Training Labels, where "1" denotes a successful booster landing and "0" signifies an unsuccessful attempt.



Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/Data%20wrangling.ipynb

EDA WITH DATA VISUALIZATION

Visualizations were generated for the following comparisons:

Flight Number versus Payload Mass
Flight Number versus Launch Site
Payload Mass versus Launch Site
Orbit Type versus Success Rate
Flight Number versus Orbit Type
Payload Mass versus Orbit Type
Success Rate Yearly Trend

Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/DataVizEDA.ipynb

EDA WITH SQL

Performed 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

Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/eda-sql.ipynb

INTERACTIVE MAP WITH FOLIUM

Markers of all Launch Sites

Coloured Markers of the launch outcomes for each Launch Site

Distances between a Launch Site to its proximities

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between two coordinates.

Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/launch_site_location.ipynb

DASHBOARD WITH PLOTLY DASH

The following graphs and plots were used to visualize data

- Percentage of launches by site
- Payload range

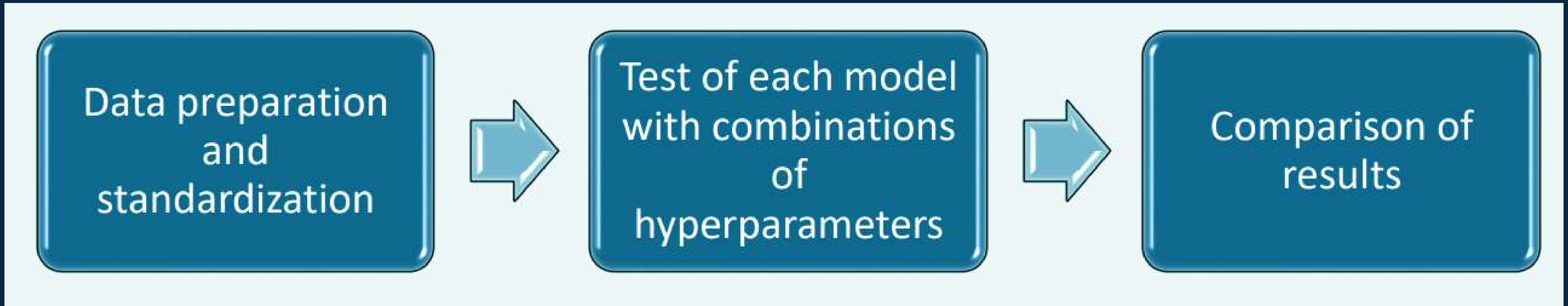
This combination helped analyze the relation between payloads and launch sites, leading to identify where is best place to launch according to payloads.

Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/spacex_dash_app.py

PREDICTIVE ANALYSIS

Four classification models were compared:

- Logistic regression
- Support vector machine
- Decision tree
- K nearest neighbors



Source Code: https://github.com/Hudabasit/SpaceX_Falcon9_Landing_Prediction/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

RESULTS

Exploratory data analysis findings reveal the following insights:

- SpaceX operates from four distinct launch sites
- Initial launches were conducted for SpaceX itself and NASA
- The average payload of the F9 v1.1 booster is 2,928 kg
- The first successful landing occurred in 2015, five years after the initial launch
- Many Falcon 9 booster versions successfully landed on drone ships, particularly those carrying payloads above the average
- Nearly 100% of mission outcomes were successful
- In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships
- The success rate of landing outcomes improved over the years

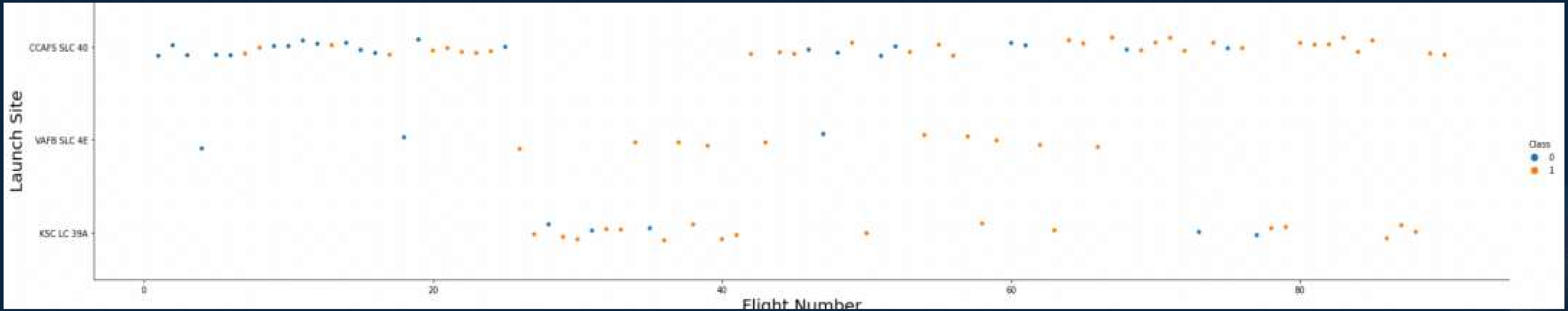
RESULTS

Through interactive analytics, it was discernible that launch sites are strategically located in secure areas, often in proximity to the sea, ensuring safety measures, and boasting robust logistical infrastructure. The majority of launches occur at launch sites along the east coast.

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings.

INSIGHTS FROM EDA WITH VISUALIZATION

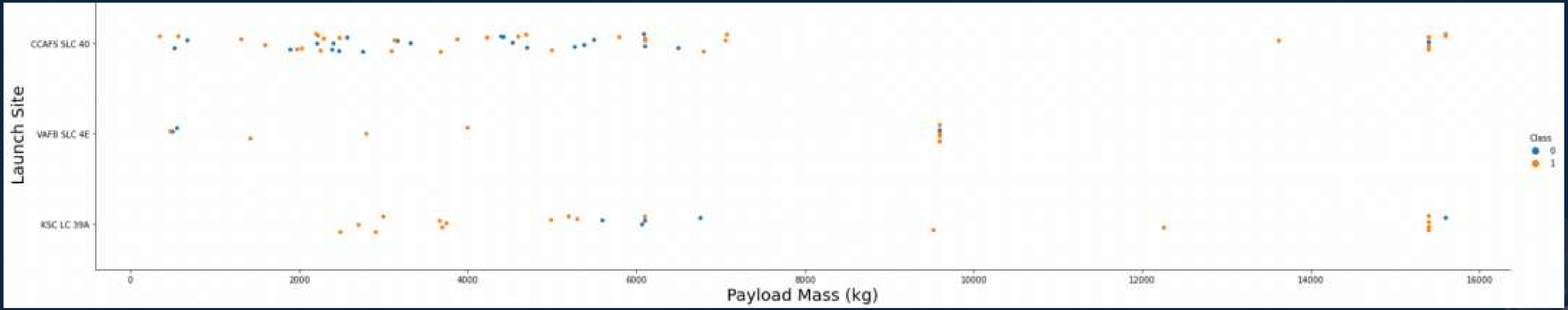
FLIGHT NUMBER VS. LAUNCH SITE



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.

INSIGHTS FROM EDA WITH VISUALIZATION

PAYLOAD VS. LAUNCH SITE

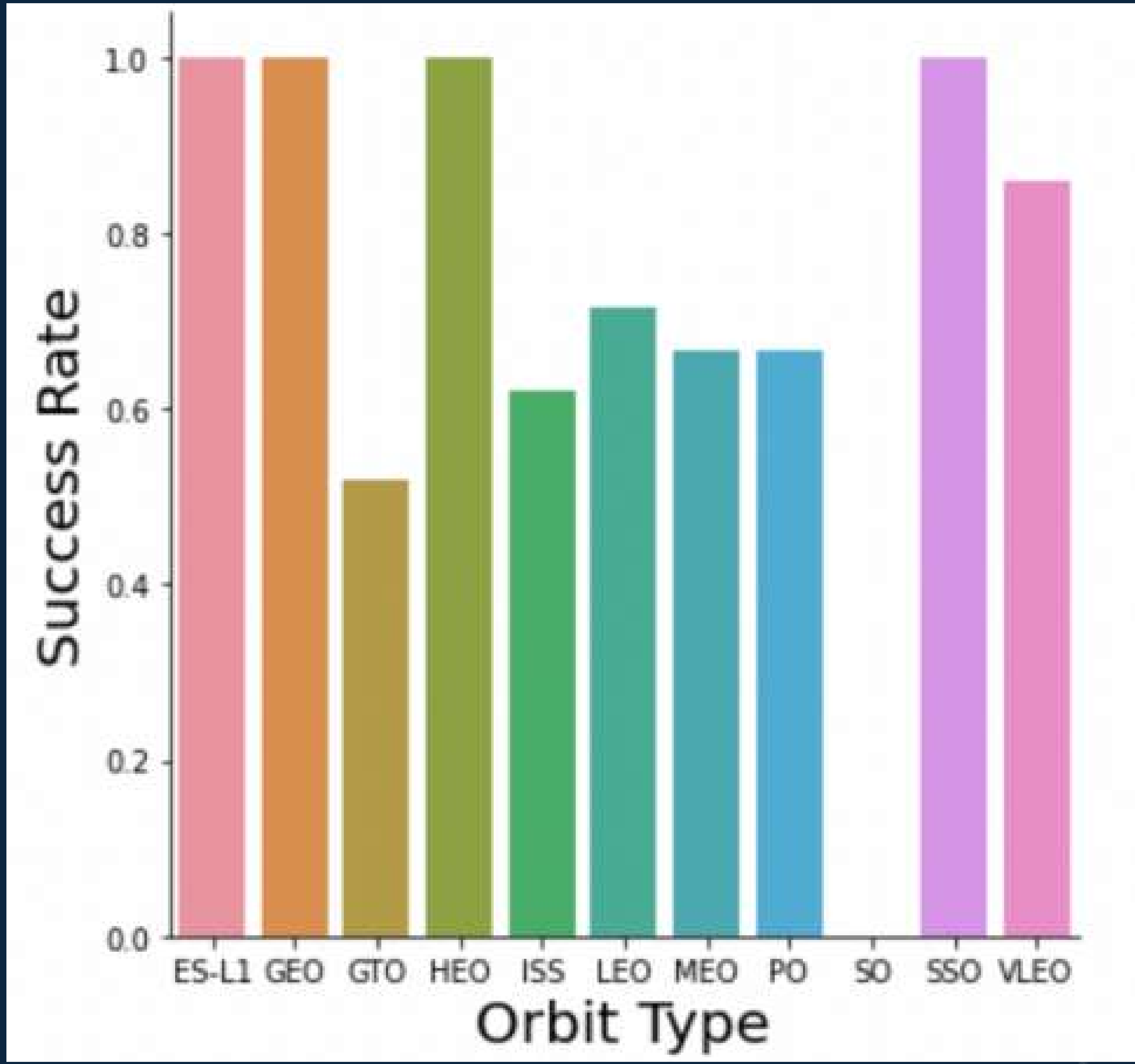


For every launch site the higher the payload mass, the higher the sccess rate.
Most of the launches with payload mass over 7000 kg were successful.
KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

INSIGHTS FROM EDA WITH VISUALIZATION

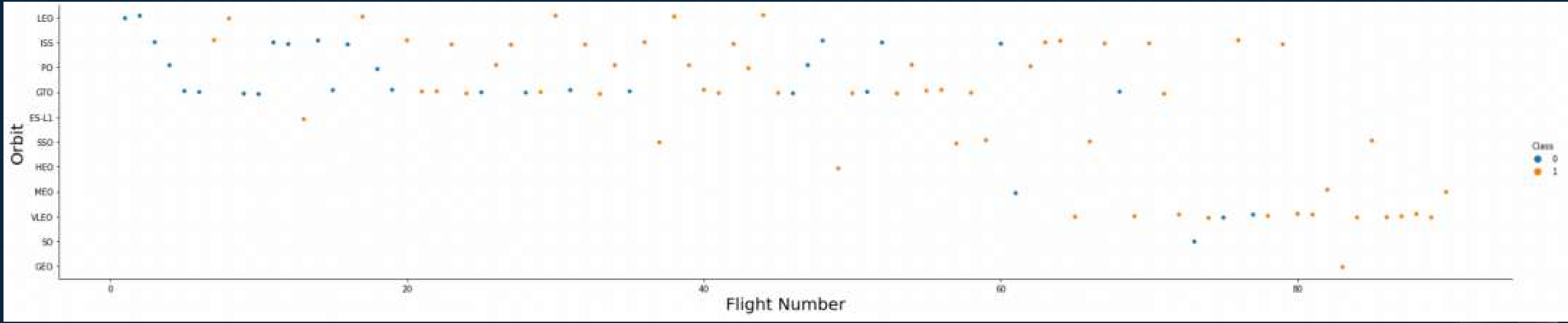
SUCCESS RATE VS. ORBIT TYPE

ORBITS WITH 100% SUCCESS RATE:
– ES-L1, GEO, HEO, SSO
ORBITS WITH 0% SUCCESS RATE:
– SO
ORBITS WITH SUCCESS RATE BETWEEN 50% AND 85%:
– GTO, ISS, LEO, MEO, PO



INSIGHTS FROM EDA WITH VISUALIZATION

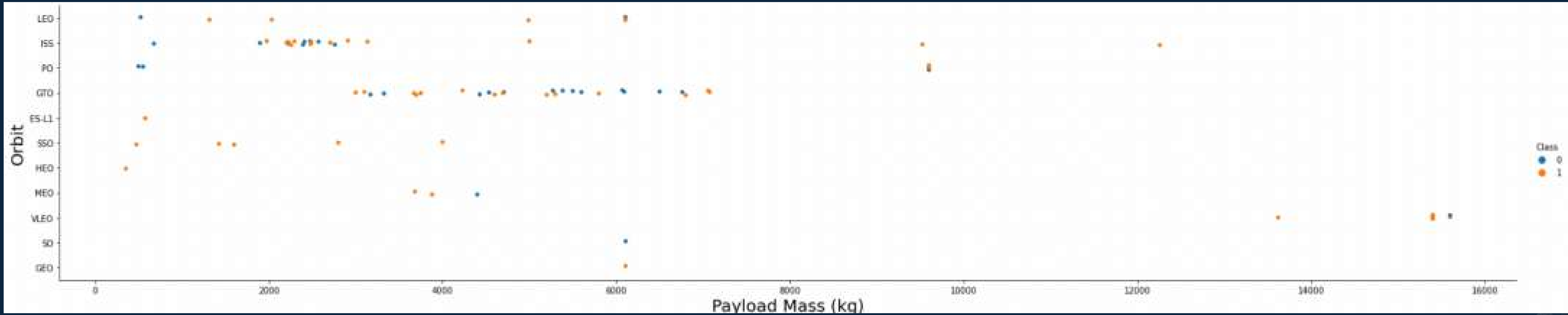
FLIGHT NUMBER VS. ORBIT TYPE



In the Low Earth Orbit (LEO), success seems to be correlated with the number of flights. Conversely, there appears to be no discernible relationship between flight number and success when in the Geostationary Transfer Orbit (GTO).

INSIGHTS FROM EDA WITH VISUALIZATION

PAYLOAD MASS VS. ORBIT TYPE

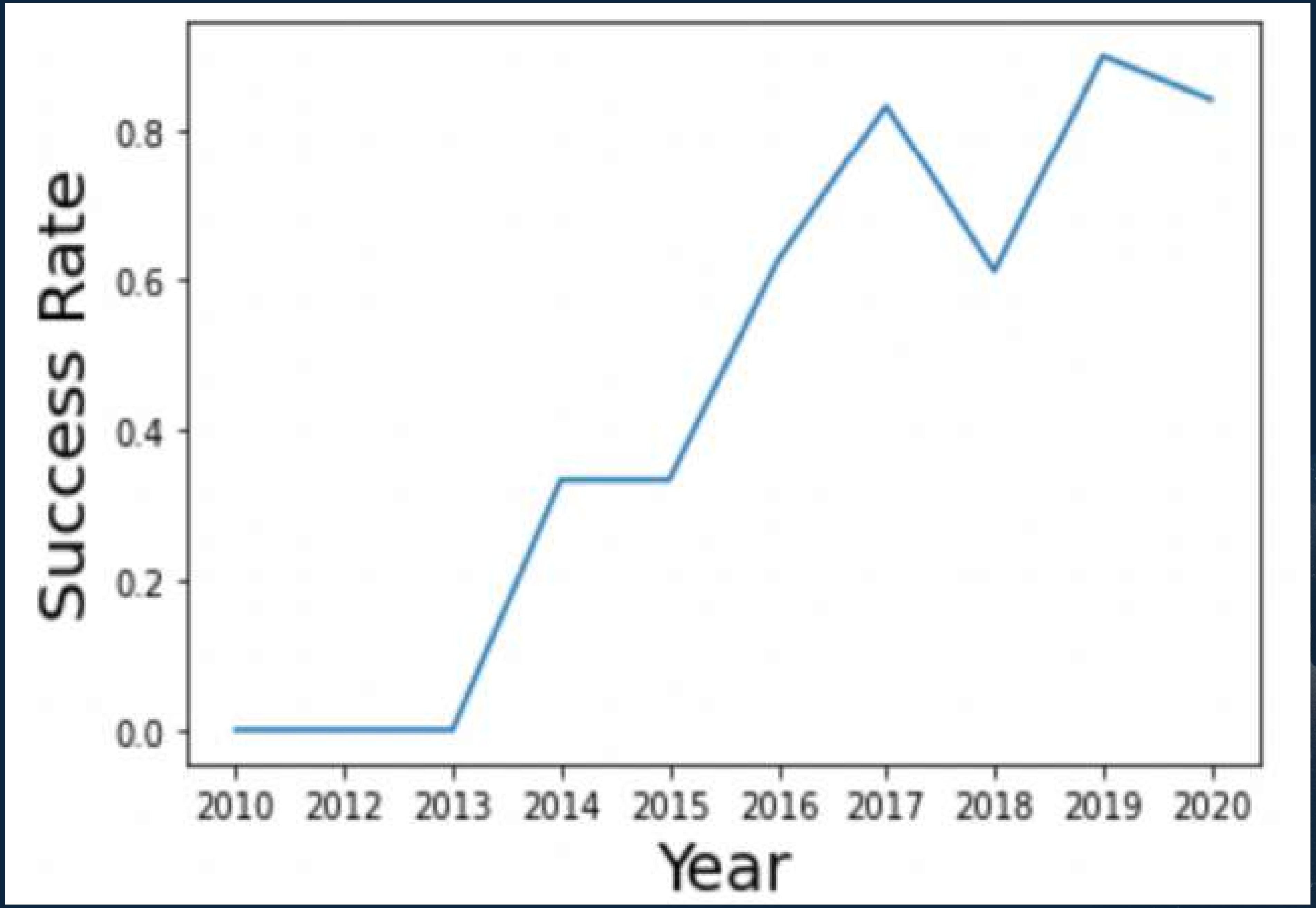


Large payloads exert a detrimental effect on Geostationary Transfer Orbits (GTO) while demonstrating a positive impact on Geostationary Orbits (GEO) and Polar Low Earth Orbits (LEO) such as those used for the International Space Station (ISS).

INSIGHTS FROM EDA WITH VISUALIZATION

LAUNCH SUCCESS YEARLY TREND

The success rate has been consistently rising from 2013 to 2020.



INSIGHTS FROM EDA WITH SQL

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

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

INSIGHTS FROM EDA WITH SQL

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

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

INSIGHTS FROM EDA WITH SQL

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

total_payload_mass
45596

INSIGHTS FROM EDA WITH SQL

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

total_payload_mass
45596

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

average_payload_mass
2534

INSIGHTS FROM EDA WITH SQL

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

first_successful_landing
2015-12-22

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

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

INSIGHTS FROM EDA WITH SQL

Listing the total number of successful and failure mission outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

INSIGHTS FROM EDA WITH SQL

Listing the 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

INSIGHTS FROM EDA WITH SQL

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

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

INSIGHTS FROM EDA WITH SQL

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.

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

INTERACTIVE MAP WITH FOLIUM

Most launch sites are situated near the Equator line, where the land moves at a faster pace compared to other locations on Earth's surface. Objects at the equator are already moving at a speed of 1670 km/hour due to the Earth's rotation. When a spacecraft is launched from the equator, it inherits this speed and continues to orbit the Earth at the same velocity due to inertia. Launching rockets towards the ocean from these sites helps minimize the risk of debris falling or exploding near populated areas, as all launch sites are located very close to the coast.

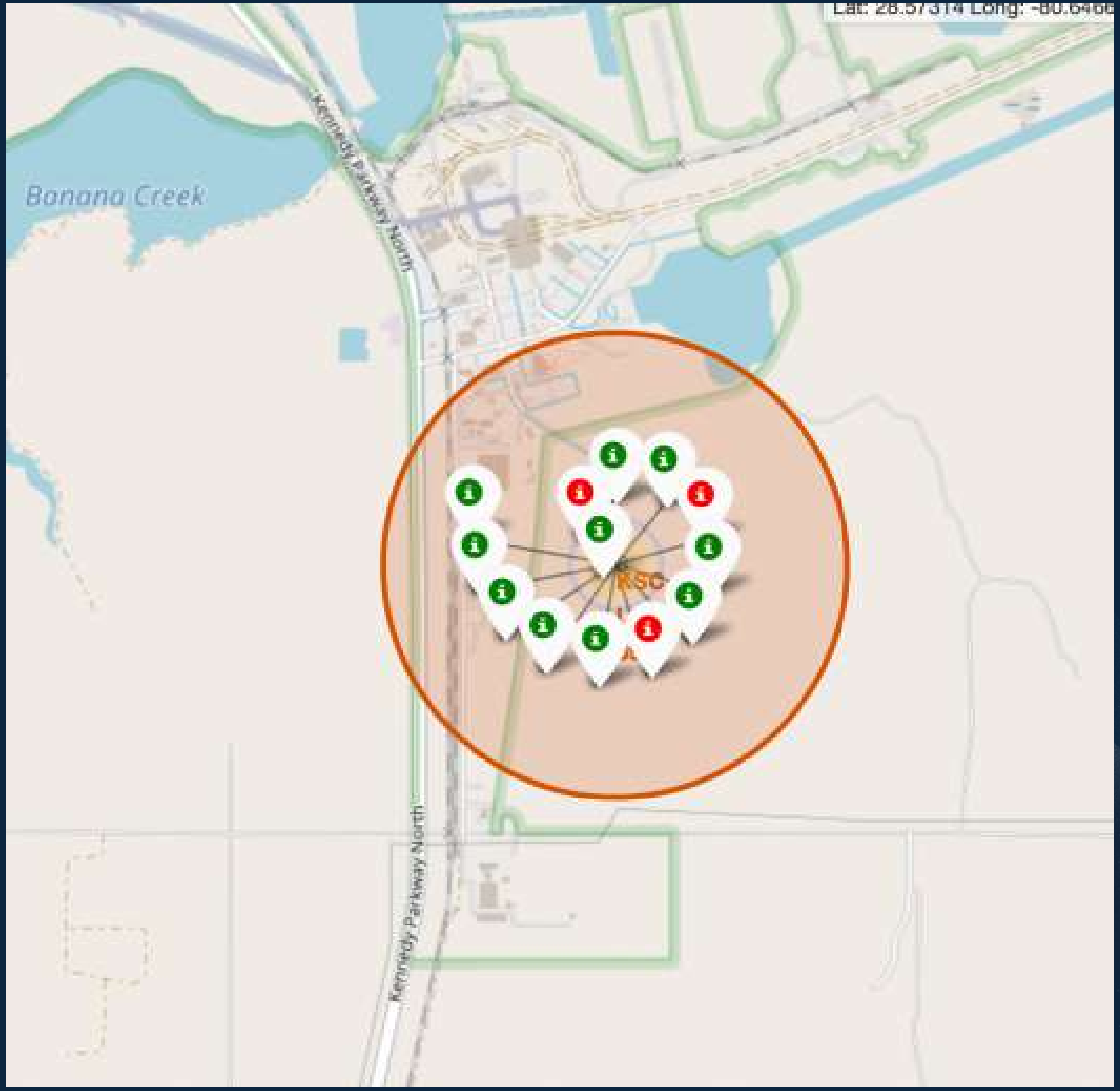
INTERACTIVE MAP WITH FOLIUM

The color-coded markers should help us quickly identify the launch sites with comparatively high success rates

Green Marker = Successful Launch

Red Marker = Failed Launch

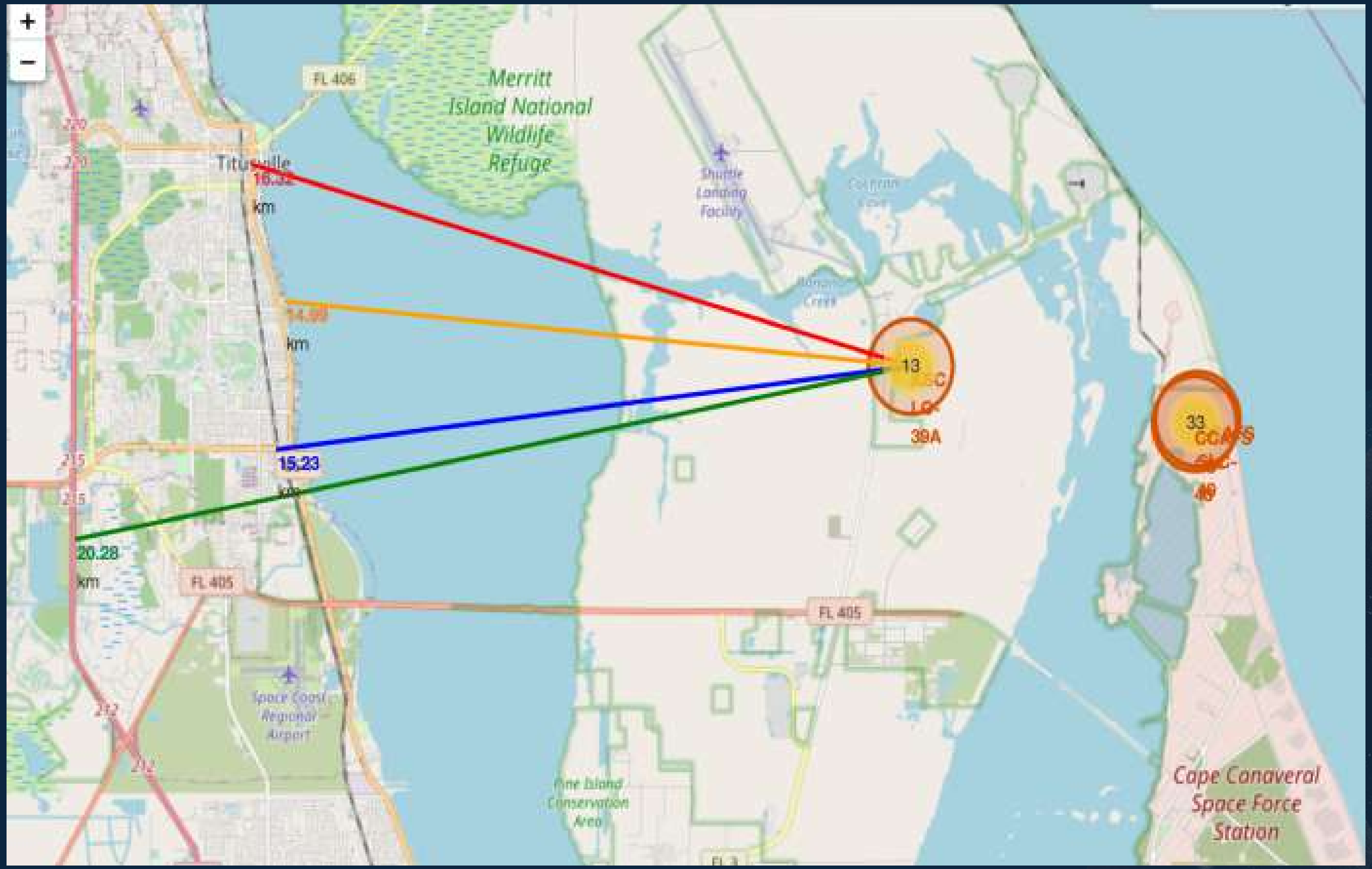
Launch Site KSC LC-39A has a very high Success Rate.



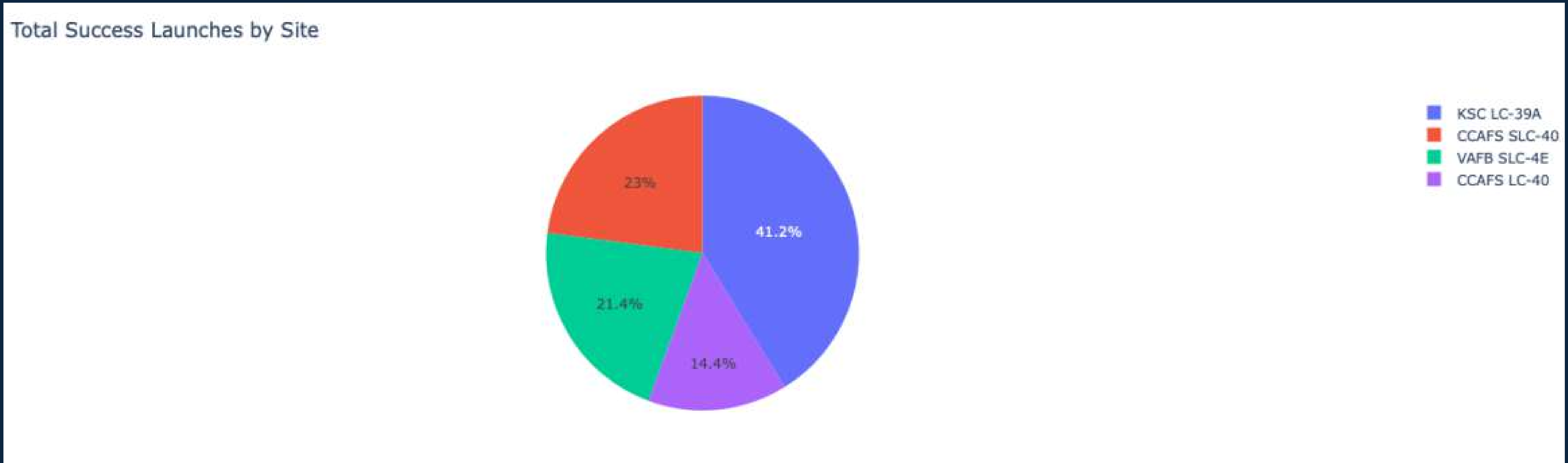
INTERACTIVE MAP WITH FOLIUM

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 distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



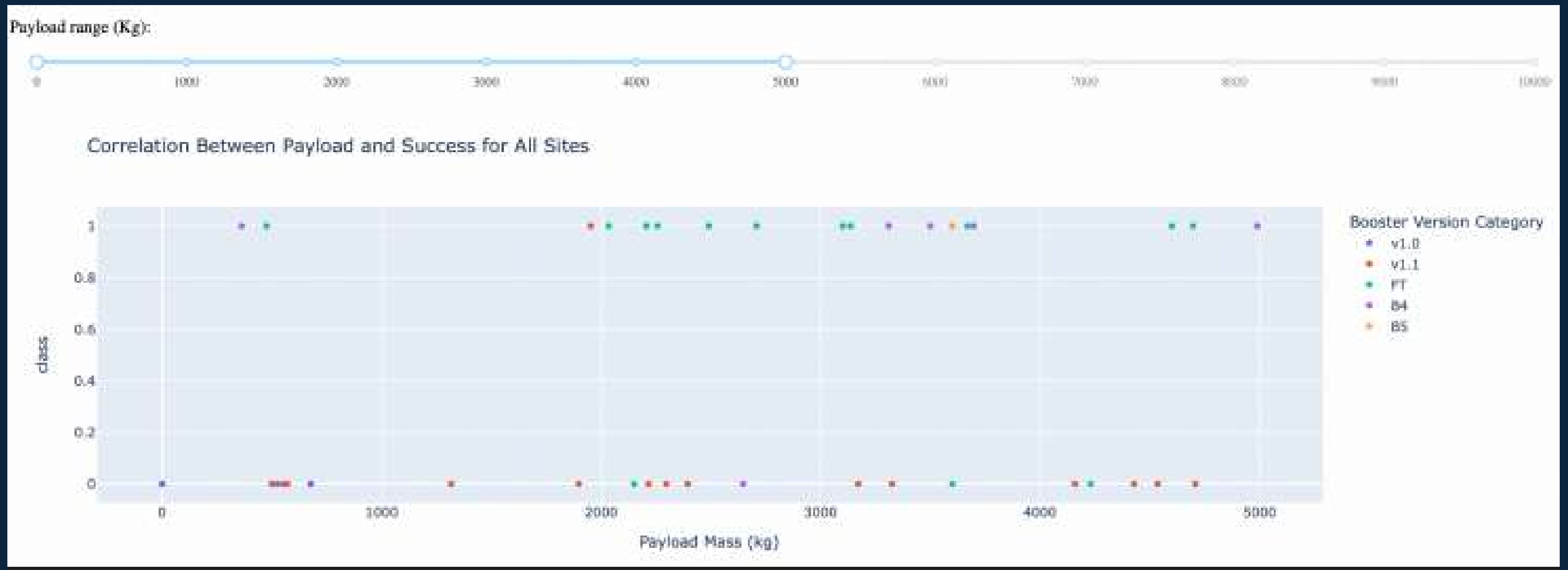
DASHBOARD WITH PLOTLY DASH

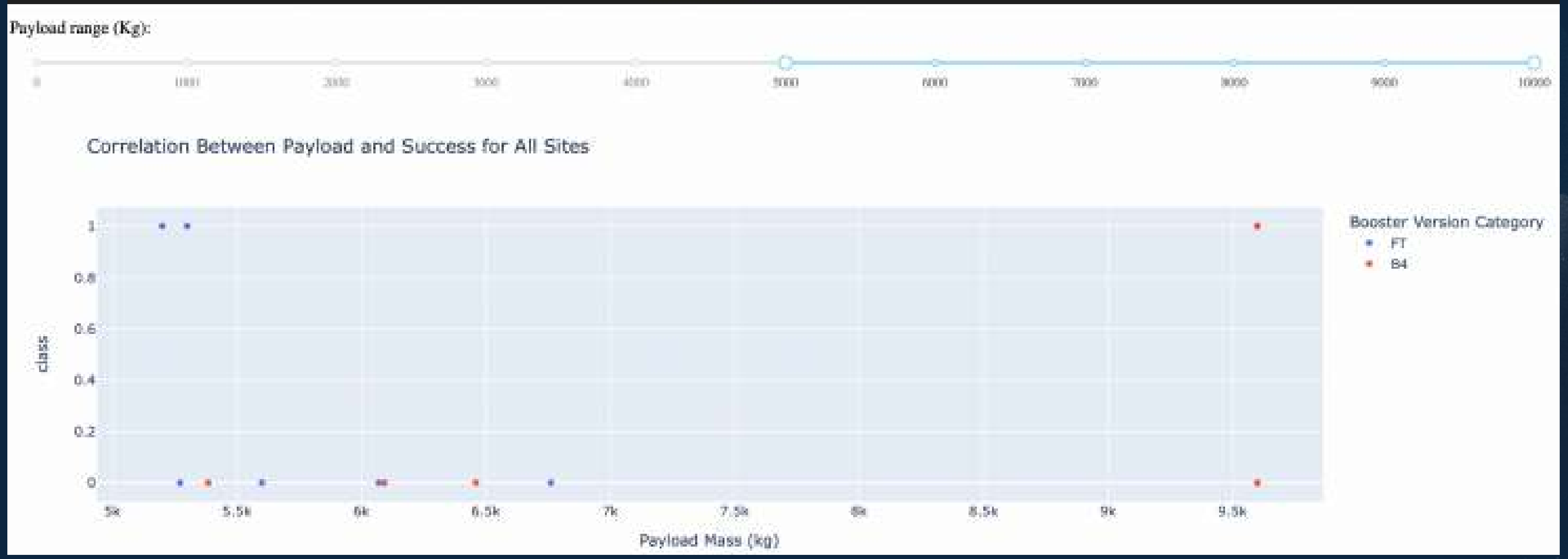


The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

DASHBOARD WITH PLOTLY DASH

The charts show that payloads between 2000 and 5500 kg have the highest success rate.





PREDICTIVE ANALYSIS CLASSIFICATION

Scores and Accuracy of the Test Set

	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

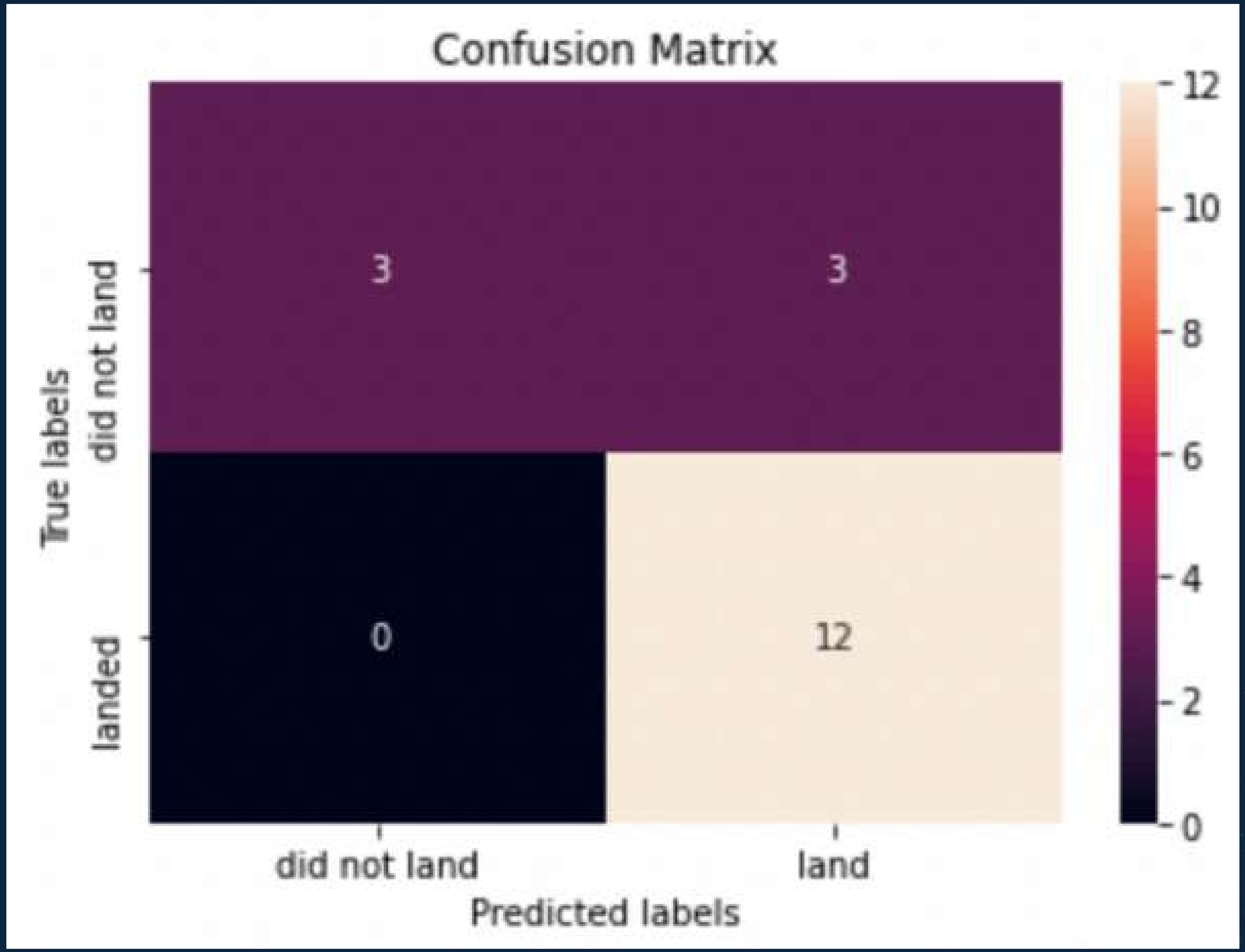
Scores and Accuracy of the Entire Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

PREDICTIVE ANALYSIS CLASSIFICATION

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



CONLUSION

- Decision Tree Model is the best algorithm for this dataset. Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

APPENDIX

Scatter plots illustrate the relationship between variables, providing insights that may be utilized in machine learning models. Bar charts facilitate comparisons between discrete categories, highlighting the relationship between the categories being compared and a measured value. Line charts depict trends in data over time, allowing for the analysis of time series data.

IBM

THANK YOU

Presentation by Huda Basit