IBM Developer
SKILLS NETWORK

The SpaceX Falcon 9 Success: **Exploratory Data Analysis** and **Landing Prediction**

Sahana Ashok

Wednesday, October 16, 2024

Agenda

Executive Summary (Abstract)

1. Introduction
2. Problem Statement
3. Methodology
4. Results
5. Insights from EDA
6. Launch Site Proximities Analysis
7. Launch Data Reporting and Dashboarding
8. Predictive Analysis
9. Conclusion

References

Appendix

Executive Summary

The objective of this study is to **predict the success of Falcon 9's first-stage landings**, which plays a crucial role in reducing the cost of rocket launches through reusability. **SpaceX charges approximately $62 million per launch** compared to other providers whose costs exceed $165 million. By accurately predicting whether the first stage will land successfully, alternative launch providers can **better estimate their competitive bids and optimize launch planning**. The dataset used for this analysis was collected through SpaceX API requests and web scraping of historical launch data from Wikipedia. After merging, preprocessing, and wrangling, the data was prepared for exploratory analysis. Relationships between various features, including launch site location, payload mass, orbit type, and landing outcomes, were studied. Preliminary trends indicated a correlation between launch site proximity to coastlines and landing success rates. Using SQL for data manipulation and visualization tools for trend discovery, we explored success rates across multiple factors. For predictive modeling, the data was standardized, and the features were engineered to improve accuracy. Binary labels were assigned to outcomes, and three supervised machine learning classifier models were trained and optimized through hyperparameter tuning after splitting into train-test sets. The models were evaluated based on accuracy. **Logistic Regression**, **SVM**, and **KNN** classifier models performed with the same accuracy of 83%. However, a common challenge across models was the presence of False Positives. This study provides actionable **insights into launch success patterns** and highlights areas for future improvements in predictive modeling. It also offers a foundational framework for competitive analysis, helping potential providers **optimize bids and assess factors for building launch sites**.

Part A

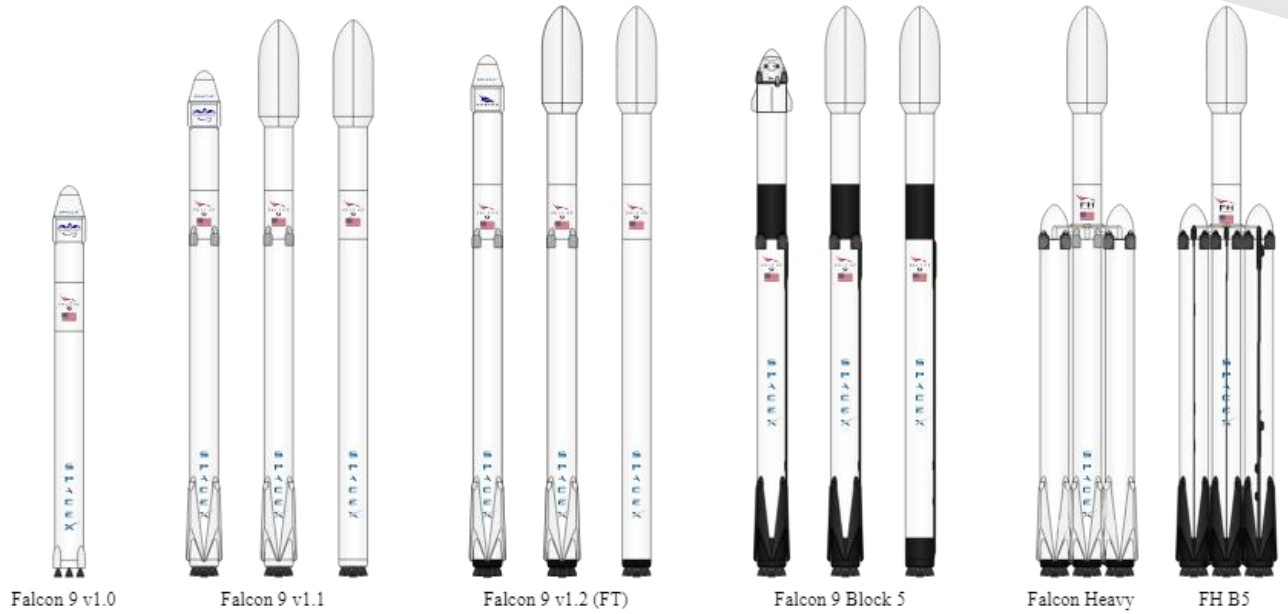
1. Introduction
2. Problem Statement
3. Methodology
4. Results

1. Introduction

The cost of rocket launches is a critical factor for organizations planning satellite deployments, space missions, or competitive bids in the aerospace market. Predicting whether the first stage will be reused allows for more accurate cost estimation, helping potential competitors or partners in optimizing their bids and launch strategies.

SpaceX Falcon 9

- SpaceX has disrupted the industry by offering launches at approximately $62 million per launch, compared to competitors charging over $165 million.
- A key component of this cost advantage is reusability of the Falcon 9 first stage, which significantly reduces operational expenses by reserving some fuel to land the 1st stage rocket booster.
- Proclamations indicate a 1st stage Falcon 9 booster costs at least $15 million without R&D cost recoupment or profit margin.



2. Problem Statement (Objective)

SpaceY is a new commercial rocket launch provider. As a Data Scientist employed by SpaceY, I aim to determine the price of each launch by gathering information on SpaceX launches and creating Dashboards for reporting. Based on Predictive Modelling, I aim to determine whether SpaceX will reuse the 1st stage.

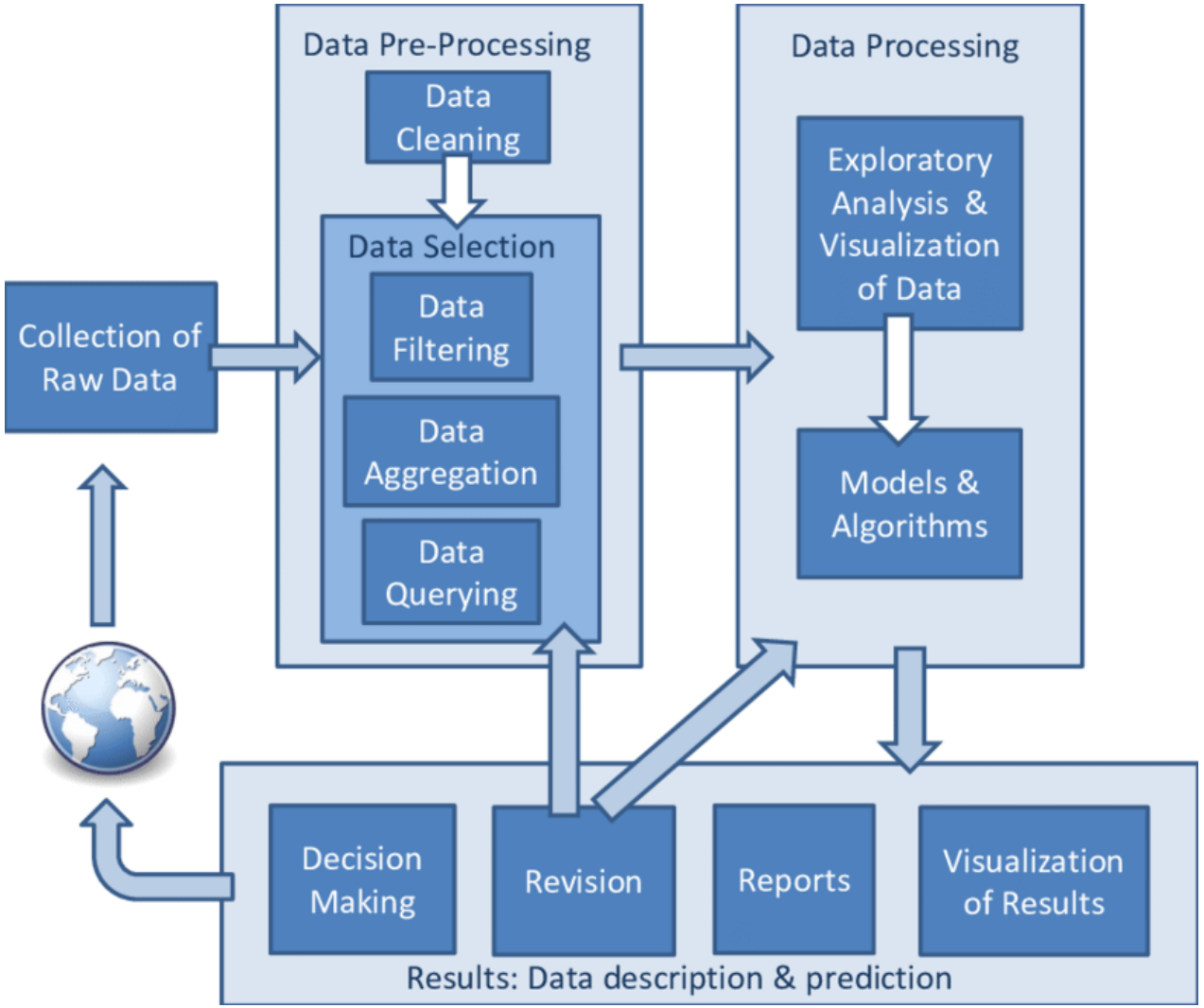
The key objectives include:

1. Data Collection and Wrangling of SpaceX launch information by
   - Web Scraping from Wiki pages.
   - Parsing and Handling missing data.
   - Obtaining an overview of the data.
2. Exploratory Data Analysis by
   - Feature Engineering of landing outcomes.
   - Observing landing trends and statistics.
   - Analyzing the attribute patterns.
   - Calculating correlation between attributes.
3. Visual Analytics of launch data insights by
   - Marking locations and proximities of launches.
   - Discovering map patterns.
   - Analyzing optimal launch site choices.
4. Predictive Analysis and Evaluation of models by
   - Target Prediction of whether Falcon 9 will land, given the mission parameters.
   - Determining the model with best accuracy and hyperparameter values.
   - Determining landing status via Confusion Matrices for each model.

Using the 1st stage landing predictions as a proxy for the cost of a launch, SpaceY will be able to make more informed bids against SpaceX.

1. Data Collection
2. Data Wrangling
3. Exploratory Data Analysis (EDA)
4. Data Visualization
5. Model Development, Predictive Analysis, and Model Evaluation
6. Reporting of Results to stakeholders

4. Methodology



Proposed System

DATA ANALYSIS

1. Data Collection
2. Data Cleaning
3. Data Description and Overview
4. Target Trends
5. Data Visualization

TARGET PREDICTION

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree Classifier
4. K-Nearest Neighbour
5. Metric Analysis

ROADMAP

Below demonstrates the phases of Data Analysis and methods performed in each stage.

Data Collection

Procurement of launch data was performed using Open Source SpaceX REST API and by means of Web Scraping Falcon 9 Wiki.

SPACEX API

- Open Source REST API that contains launch data including information like rocket used, payload delivered, landing outcomes, launch specifications, etc.
- Past launch information is collected via requests sent to endpoint https://api.spacexdata.com/v4/launches/past.

STEPS PERFORMED

1. Requested SpaceX launch data by sending a GET request to endpoint.
2. Parsed the JSON response into a Pandas DataFrame.
3. Obtained an overview on available data and its attributes.
4. Filtered the DataFrame to include only Falcon 9 launches.
5. Handled missing values by replacing PayloadMass by its mean.

SpaceX API Implementation Notebook

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/1-jupyter_labs_spacex_data_collection_api.ipynb

WEB SCRAPING

- Wikipedia page titled *List of Falcon 9 and Falcon Heavy launches* containing Falcon 9 launch data subject to web scraping using BeautifulSoup package, which parsed table HTML tags into DataFrame. Falcon 9 launch dataset was limited to launches before December 7, 2020.
- Functions for specific launch IDs used to fetch further information, include getBoosterVersion(), getLaunchSite(), getPayloadData(), and getCoreData().

STEPS PERFORMED

1. GET request sent to Falcon 9 Wiki URL.
2. BeautifulSoup object created, and HTML table details extracted from response.
3. Column names parsed from table header fetched and Dictionary created with these keys to store data.
4. Data from table details parsed and appended to Dictionary keys as values.
5. Dictionary converted to DataFrame, and saved as CSV.

Web Scraping of Falcon9 Wiki Page Notebook

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/2-jupyter_labs_webscraping.ipynb

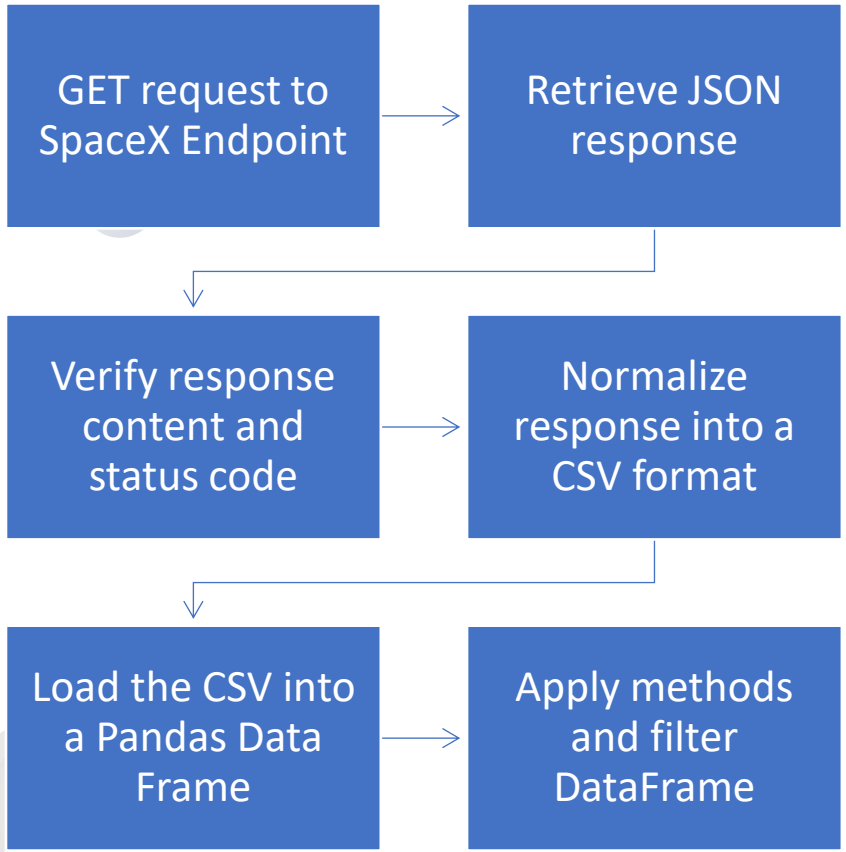
Falcon 9 launch dataset was limited to launches before December 7, 2020. Below is the Wiki table that was parsed.

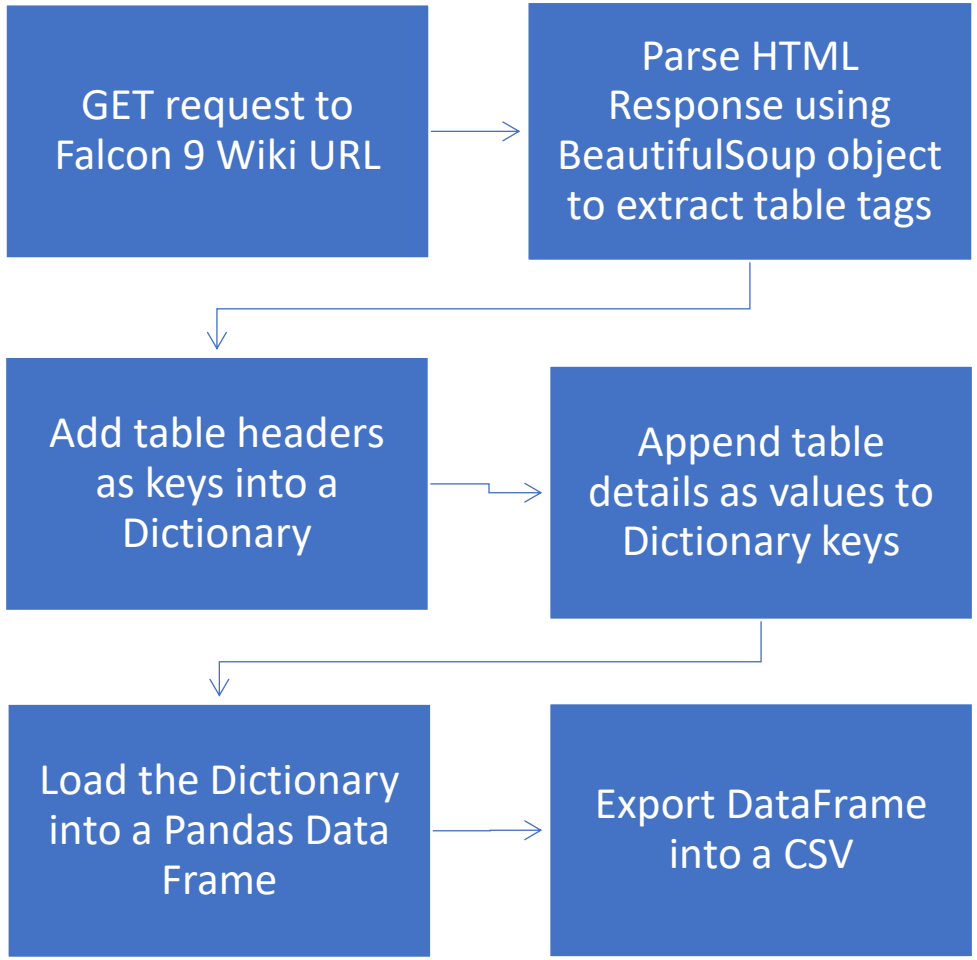
2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

SPACEX API RESPONSE VS WEB SCRAPING





Data Wrangling

The *Outcome* column indicates whether the 1st stage successfully landed. It is a set of 8 values, and is the Target variable, that must be converted into a categorical binary value 0/1 by One-Hot encoding. Site and Orbit based counts was calculated based on outcome, before encoding was performed.

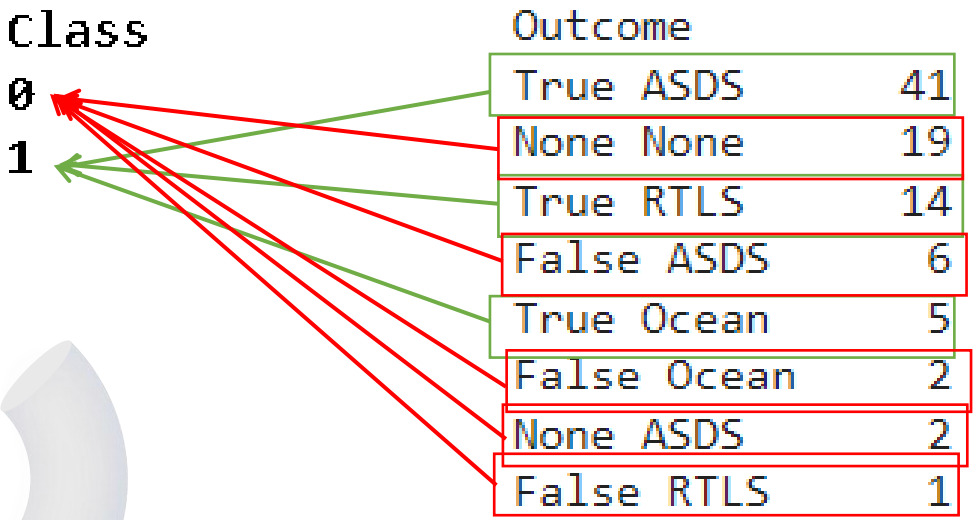
STEPS PERFORMED

1. Identified and calculated the percentage of missing values of each attribute.
    - Null values in *Payload* mass replaced by mean.
    - *LandingPad* Null values unaltered, as these denoted no usage of landing pads.
2. Identified numerical and categorical columns.
3. Counts of below calculated based on *Outcome*:
    - Number of launches in each site.
    - Number and occurrence of each orbit.
    - Number and occurrence of mission outcome of the orbits.
4. Created a landing outcome label *Class* from *Outcome* column by One Hot Encoding.
5. Determined mean success rate of landing.
6. Exported prepared data as CSV.

ONE HOT ENCODING CRITERIA

Class = 1: 1st Stage booster landed **successfully**.

- True ASDS (Drone ship landing succeeded)
- True RTLS (Ground pad landing succeeded)
- True Ocean (Ocean Landing succeeded)



Class = 0: 1st Stage booster **failed** to land successfully.

- None None (Not attempted)
- False ASDS (Drone ship landing failed)
- False Ocean (Ocean landing failed)
- None ASDS (Unable to have been attempted due to launch failure)
- False RTLS (Ground pad landing failed)

Data Wrangling Implementation Notebook

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/3-labs_jupyter_spacex_Data_wrangling.ipynb

Exploratory Data Analysis

EDA by SQL Query Execution

SQL QUERIES EXECUTED

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/3-labs_jupyter_spacex_Data_wrangling.ipynb

1. Displaying the names of the **unique launch sites** in the space mission.
2. Displaying 5 records where **launch sites** begin with the string '**CCA**'.
3. Displaying the **total payload mass** carried by **boosters launched by NASA (CRS).**
4. Displaying **average payload mass** carried by **booster version F9 v1.1**.
5. Listing the **date** when the **successful landing outcome in ground pad** was achieved.
6. Listing the names of the **boosters** which have **success in drone ship** and **payload mass in range (4000,6000).**
7. Listing the **total number of successful and failure mission outcomes**.
8. Listing the names of the **booster_versions** which have carried the **maximum payload mass**, using subquery.
9. Listing the records which will display **the month names**, **failure landing_outcomes in drone ship**, **booster versions, launch site** for the months **in year 2015**.
10. Ranking the **count of landing_outcomes** between the date 2010-06-04 and 2017-03-20 in descending order.

DATA VISULATIZATION

The data set is read into the Pandas DataFrame. Using Matplotlib and Seaborn visualization libraries, the below relationships were plotted and analyzed for trends, patterns, and insights.

PLOTS

1. Effect of continuous launch attempts for each payload mass, on outcome. (**FlightNumber vs PayloadMass**)*
2. Effect of continuous launch attempts for each launch site, on outcome. **(FlightNumber vs LaunchSite**)*
3. Relationship between payload mass and launch site, based on outcome**. (PayloadMass vs LaunchSite**)*
4. Bar chart representing success rates based on each **Orbit** type, to aggregate counts.
5. Effect of continuous launch attempts for each orbit type, based on outcome. (**FlightNumber vs Orbit**)*
6. Relationship between payload mass and orbit, based on outcome. (**PayloadMass vs Orbit**)*
7. Line chart representing average **yearly launch success trends** from 2010-2020, to plot time series.

*Hue of *Class* overlayed since categorical impact also required for analysis. Scatter plots catplot() made.

Some attributes can determine Stage 1 reuse, like launch sites and timeline trends. These attributes have been combined for feature engineering. The below features were selected to be used in success prediction.

- FlightNumber
- PayloadMass
- Orbit
- LaunchSite
- Flights
- GridFins
- Reused
- Legs
- LandingPad
- Block
- ReusedCount
- Serial

FEATURE ENGINEERING

1. Applied get_dummies() to categorical feature columns Orbits, LaunchSite, LandingPad, and Serial.
2. Applied OneHotEncoder to the feature attributes.
3. Displayed the results using the method head().
4. Casted the entire DataFrame to variable type float64.
5. Exported the Dataframe to CSV.

Plotting and Feature Engineering

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/5-jupyter_labs_eda_dataviz.ipynb

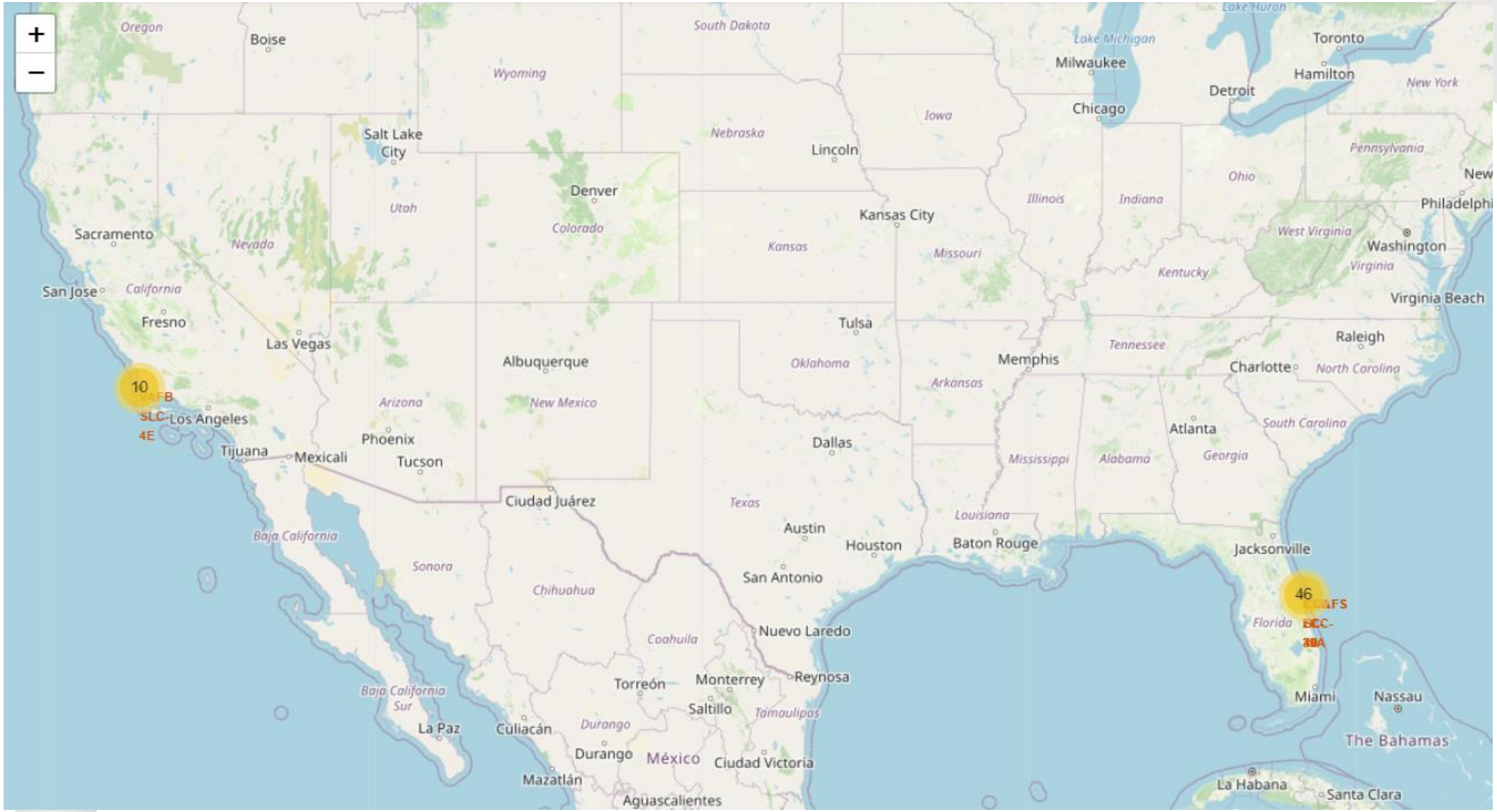
Interactive Map using Folium

Launch success rate may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors. These can be discovered by analyzing existing launch site locations.

MARKING LAUNCH SITES AND PROXIMITIES ON MAP

spacex_launch_geo.csv is an augmented dataset with latitude and longitude added for each site. It is fetched into a DataFrame and plotted over a world map using Folium package.

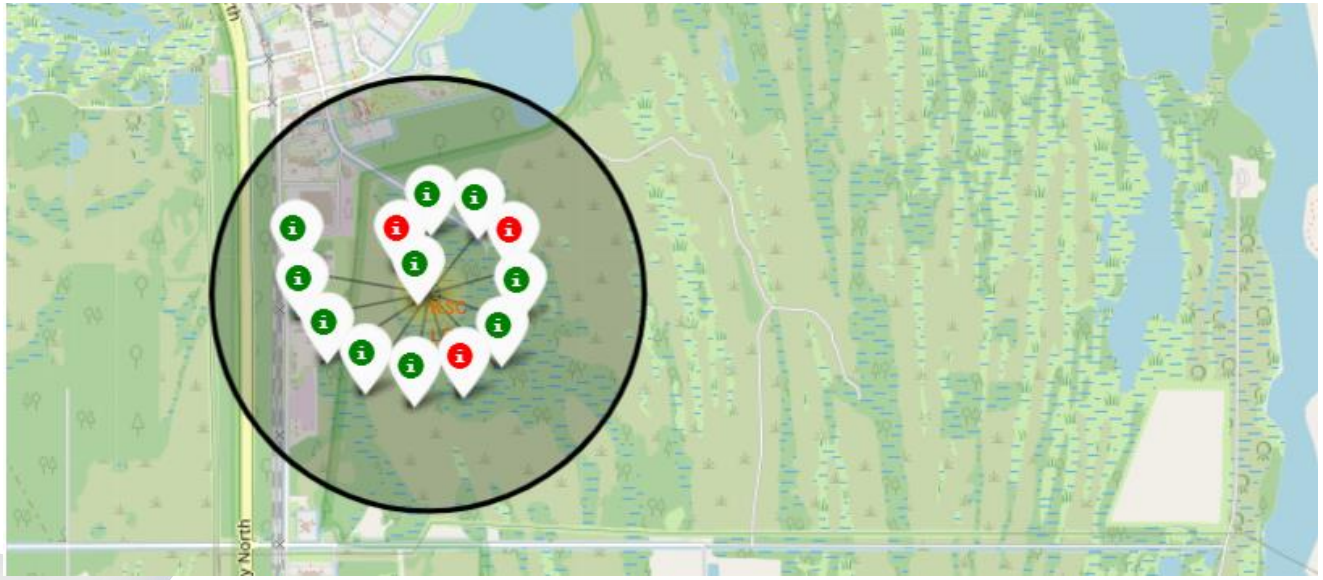
1. Folium.Map object created with initial center location to be NASA Johnson Space Center at Houston, Texas.
2. Folium.Circle is used to add a highlighted circle area with a text label for each coordinate.
3. Folium.Marker used over each launch site to format the name of each site.
4. Launch site proximities analysed from Equator and coast.


Oregon
Boise
South Dakota
Wyoming
Lake Michigan
Milwaukee
Toronto
Hamilton
Detroit
New York
Chicago
Iowa
Nebraska
Salt Lake City
Lincoln
Pennsylvania
Ohio
Denver
Utah
Illinois
Indiana
Kansas City
Colorado
Sacramento
Nevada
Kansas
Missouri
West Virginia
Washington
Kentucky
Virginia
San Jose
California
Virginia Beach
Fresno
Tulsa
Raleigh
Las Vegas
Tennessee
Albuquerque
Oklahoma
Memphis
Charlotte
North Carolina
Arkansas
10
AFB
SLC
4E
Los Angeles
Arizona
New Mexico
Atlanta
South Carolina
Phoenix
Dallas
Tijuana
Mexicali
Tucson
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Louisiana
Austin
Houston
Baton Rouge
Jacksonville
Baja California
Sonora
San Antonio
Chihuahua
46
AFS
Florida
Nuevo Laredo
Coahuila
Torreón
Monterrey
Reynosa
Miami
Nassau
Baja California Sur
Saltillo
Tamaulipas
La Paz
Culiacán
Durango
The Bahamas
Durango
México
Ciudad Victoria
Mazatlán
La Habana
Santa Clara
Aguascalientes

MARKING LAUNCH OUTCOMES FOR EACH SITE

Launch outcomes were added to each site, to discover which sites have high success rates. Since many launch records will have the exact same coordinate, Marker clusters have been used to portray a map containing many markers having the same coordinate.

1. Marker colors based on *Class* value initialized for each site and added to DataFrame.
2. For each launch record, folium.Marker is created with icon color as provided marker color.
3. Each folium.Marker is added to folium.MarkerCluster variable after initializing it.



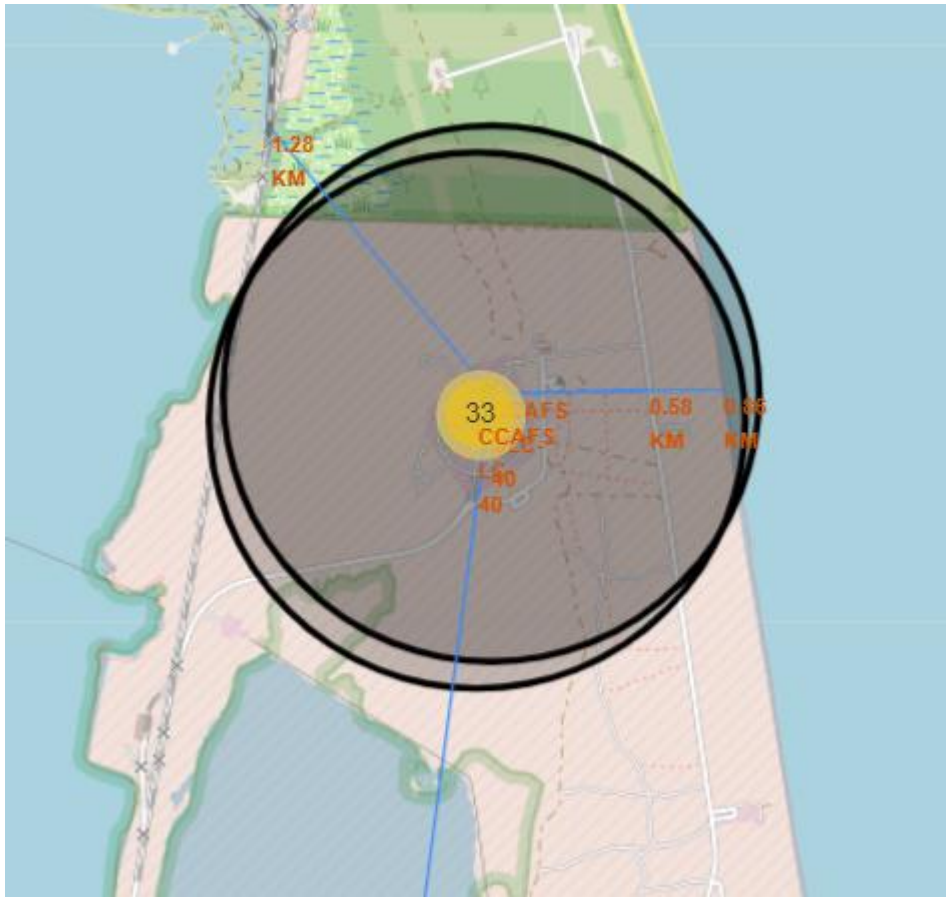
[Folium Launch Site Proximity Mapping]

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/6-lab_jupyter_launch_site_location.ipynb

LAUNCH SITE PROXIMITY DISTANCE CALCULATION

The data set is read into the Pandas DataFrame. Using Matplotlib and Seaborn visualization libraries, the below relationships were plotted and analyzed for trends, patterns, and insights.

1. Added folium.MousePositions to fetch coordinates of railways, highway, coastline, etc.
2. Calculated the distance between each coordinate and the launch sites by Haversine formula.
3. Created folium.Markers to display the distance value to each coordinate.
4. Used the folium.PolyLine to plot a line between the coordinates and launch site coordinate.
5. Observed proximity finding and reasoned their values with respect to railways, highways, coastlines, and cities.



HAVERSINE FORMULA

Haversine formula is used to calculate the great-circle distance between two points on the surface of a sphere given their latitudes and longitudes. This formula accounts for the curvature of the Earth, making it more accurate than simple Euclidean distance when working with geographical coordinates.

Given 2 coordinates on earth (lat1,lon1) and (lat2,lon2), the Haversine formula is written as:

$$a = \sin^2\left(\frac{\Delta \text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta \text{lon}}{2}\right)$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R \cdot c$$

Where $\Delta\text{lat} = \text{lat2} - \text{lat1}$

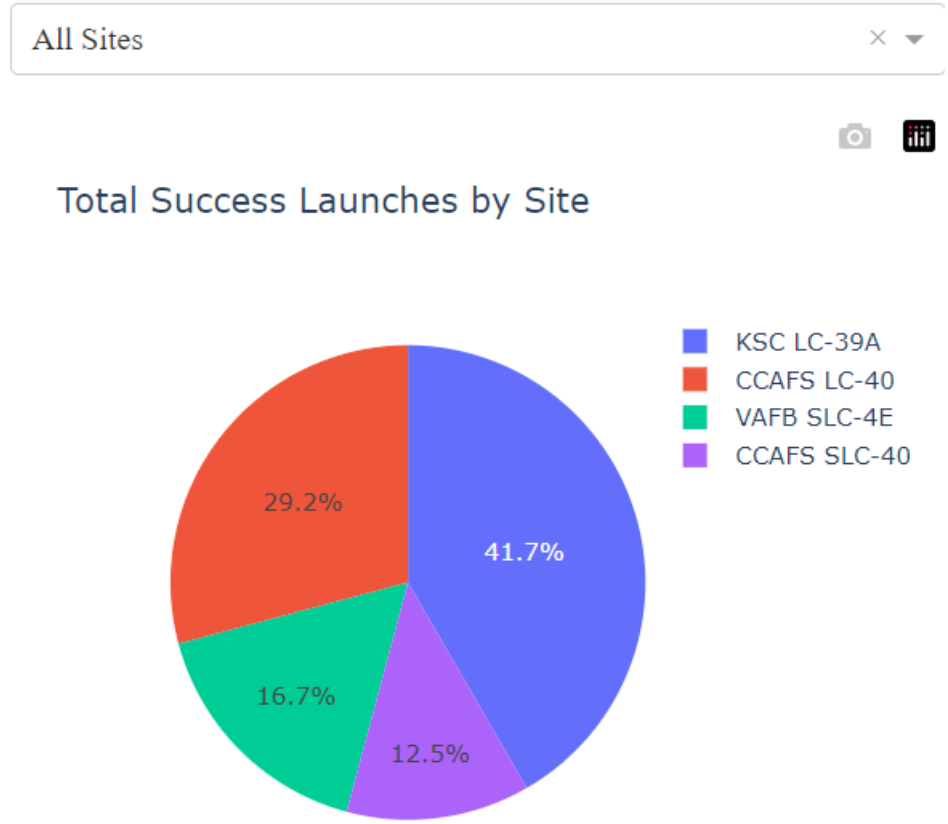
$\Delta\text{lon} = \text{lon2} - \text{lon1}$

R is the Earth's radius (mean radius ~6371 km)

d is the distance between the two points along the surface

Interactive Dashboard using Plotly

Used Python interactive dashboarding library Plotly Dash to allow stakeholders to explore and manipulate launch record data in an interactive and real-time manner as a part of reporting. The dashboard summarizes launch site based outcomes. This has been done by adding Dropdowns, Pie charts, Sliders, and Plots by rendering callback functions for each input, to display the dropdown and plots.



REPORTING

1. Site based success rates for all launches.
2. Success rates per launch site, with site selected in dropdown.
3. Scatter plot of outcome based on Payload mass, categorized by Booster version category, with payload as slider input.

These details were used to analyze success rates based on launch sites and payload.

Below is the success rate for all sites, for a payload mass between 3k to 7k.

Launch Success Rate Dashboarding

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/7_spacex_dash_app.py

Predictive Analysis (Classification)

Binary classification is performed over the *Class* attribute, that determines whether the 1st stage Falcon 9 landing was successful or not. The data set is preprocessed and split into testing and training sets. These are subject to Grid Search. The data is fed into 4 classifiers, of which the best accuracy is chosen by Cross Validation with multiple hyperparameter values. The landing states predictions are reviewed through a Confusion matrix for each model.

LIBRARIES IMPORTED

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Scikit-learn

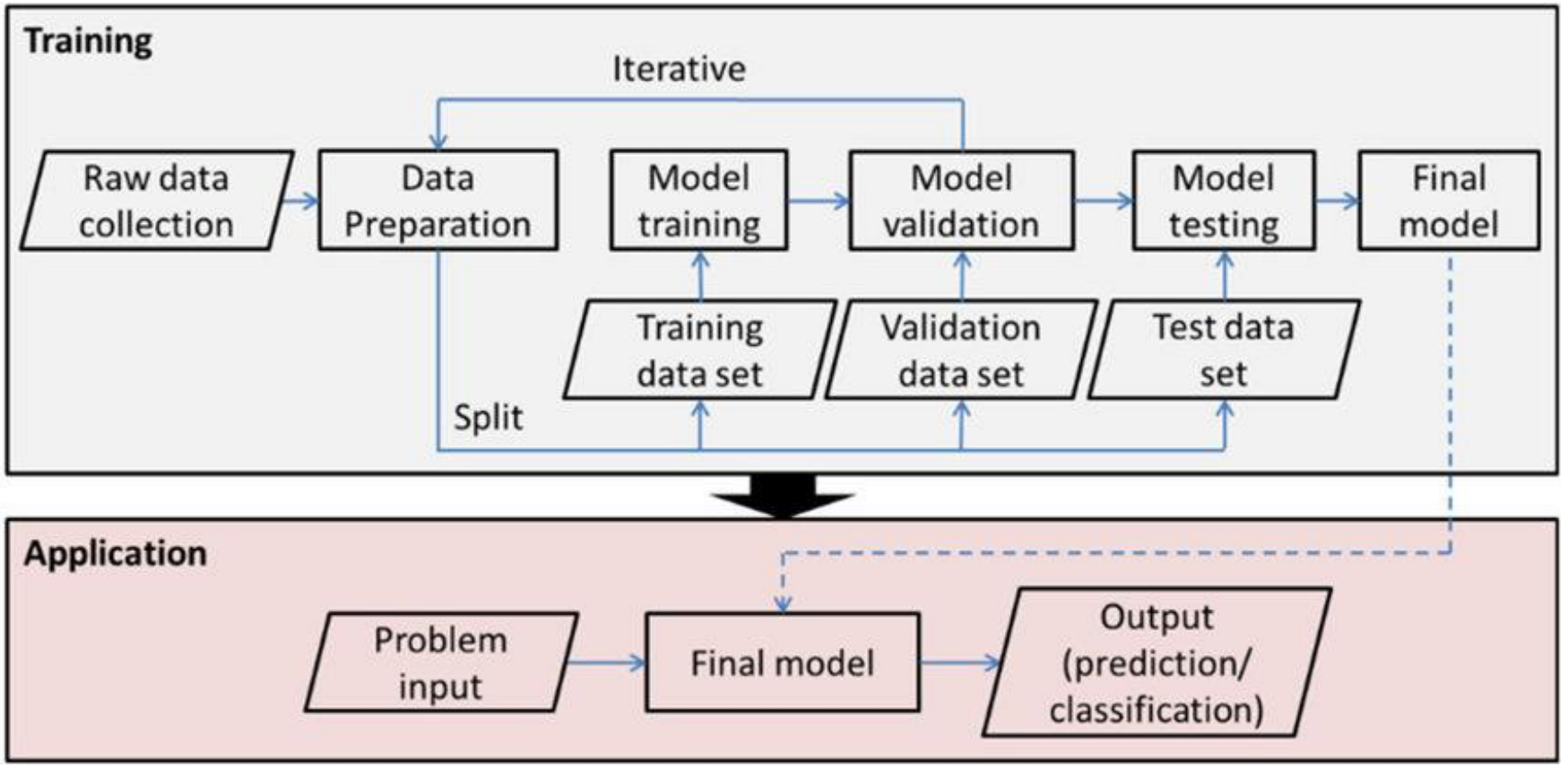
SUPERVISED CLASSIFICATION MODELS USED

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree Classifier
4. K Nearest Neighbors Classifier

Launch Success Classification Predictive Analysis

https://github.com/Lochana24/SpaceX-Falcon-9-First-Stage-Landing-Analysis/blob/main/8-SpaceX_Machine_Learning_Prediction_Part_5_jupyterlite.ipynb

In supervised classification, an ML model learns to map input features X to discrete class labels y by minimizing a specific loss function (e.g., cross-entropy for probabilistic models). During training, the model iteratively adjusts its parameters based on labeled data to reduce the discrepancy between predicted and actual labels. The model is the evaluated on unseen test data to assess its performance through metrics such as accuracy, precision, recall, and the F1-score.



STEPS PERFORMED

1. Imported libraries and respective Classifiers.
2. Defined function to create confusion matrix given the real target, and predicted target values.
3. Loaded the DataFrame from prepared CSV after data wrangling as the Training Set.
4. Stored our target label *Class,* into a Numpy array as the Test Set.
5. Standardized the Training Set data using StandardScaler.
6. Split the data into training data and test data, with test_size=0.2 and Random_state=2.
7. Fit the training data to the 4 classifier models.
8. Selected the tuned hyperparameters for each model, by performing a cross-validated grid-search using GridSearchCV with cv=10.
9. Predicted and Evaluated accuracy of each model using test data.
10. Plotted confusion matrices for each model to calculate the False Positives, False Negatives, True Positives, and True Negatives.

4. Results and Observations

- EDA Observations and Patterns
- Launch Site Proximities Observations
- Launch Data Report Insights
- Predictive Analysis Results

Wednesday, Oct 16, 2024

EDA OBSERVATIONS AND PATTERNS

1. As the **flight number increases**, the likelihood of **successful landings improves**, indicating potential benefits from experience and iterative design.

2. Payload mass plays a significant role**: heavier payloads reduce the chances of a successful landing**, with CCAFS SLC 40 showing higher success rates for such missions, while VAFB SLC 4E is less favored for heavy payloads. The **maximum payload recorded was 15,600 kg**, associated with the F9 B5 booster. **KSC LC 39A** demonstrates higher success rates for **lighter payloads** (below 6,000 kg).

3. In terms of orbital destinations, **100% success rates** were achieved for **ES-L1, GEO, HEO, and SSO**. Only the **SO orbit** experienced 100% failed launches.

4. Success tren**ds over time show consistent improvement**. Between **2013 and 2017**, the success rates **increased or remained stable**, with 2014 exhibiting no growth in success rate. However, the **rates declined briefly in 2018** but have shown an **upward trend since** then. This suggests ongoing improvements in operations, potentially leading to enhanced reliability over time.

LAUNCH SITE PROXIMITIES OBSERVATIONS

1. Launch sites are strategically positioned **near the equator** to optimize performance and safety. Proximity to the equator provides additional rotational velocity, reducing fuel requirements and improving efficiency by minimizing inclination adjustments.

2. They are **near coastlines** as well. Coastal regions, being remote and sparsely populated, offer **increased trajectory options** and **safe recovery** in case of failures, along with reduced noise pollution. There is also **lower risk of debris impact on land**, and **easier recovery from the ocean**.

3. The **KSC** launch site shows the **highest success rate**, with **76.9% of missions succeeding**, whereas **CCAFS** and **VAFB** have recorded **more failures**.

4. Launch sites are also **close to railways, highways, and coastlines**, ensuring seamless transportation of **materials, freight, and human resources**. Railways support the movement of heavy machinery and assembly components, while highways facilitate easier logistics for staff and smaller equipment.

5. Their situation **far from cities** minimizes **noise interference**, ensures **safety** from potential debris impacts, and aligns with **emergency protocols** for launches.

LAUNCH DATA REPORT INSIGHTS

1. The **KSC LC-39** launch site leads in performance with the **highest successful launches** (41.7%) and the **largest success rate** of **76.9%**. In contrast, **CCAFS SLC-4E** has the **lowest success rate**, achieving only **12.5%**.

2. The payload range of **2,000 to 4,000 kg shows the highest success**, followed by the **4,000 to 6,000 kg** range. Payloads **exceeding 6,000 kg** exhibit the **lowest success rates**.

3. Booster version **FT** has the **most successful launches**, with **B4** following closely, while **v1.1** records the fewest successful launches.

PREDICTIVE ANALYSIS RESULTS

1. **LR, SVM, and KNN models** performed favorably, correctly predicting all 12 successful landings with 0 False Negatives (Type II error). All models showed occurrence of False Positives (Type I error).

2. Above 3 models achieved the **highest accuracy of 83.33%.** The Decision Tree model had lowest accuracy of 78.2%.

Part B

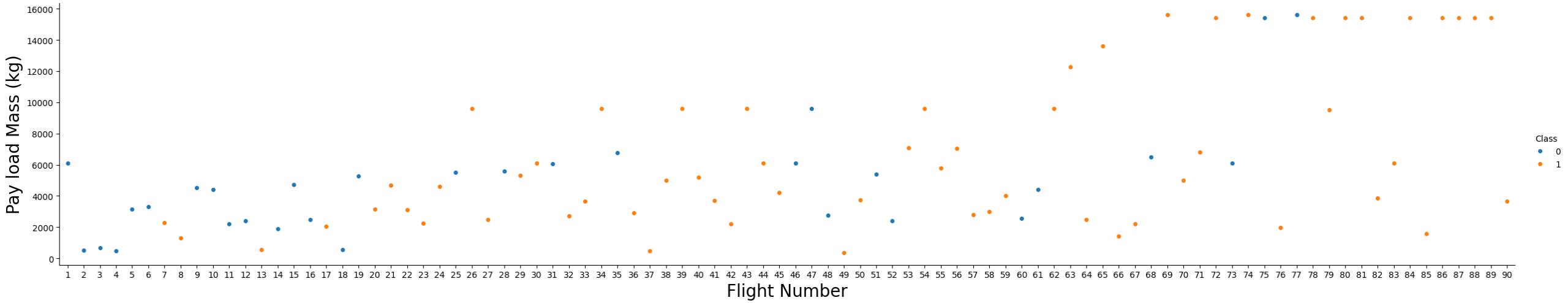
5. Insights from EDA
6. Launch Site Proximities Analysis
7. Launch Data Reporting and Dashboarding Insights
8. Predictive Analysis
9. Conclusion

5. Insights from EDA

FLIGHT NUMBER VS PAYLOAD MASS

As the count of flights increase, success rate more likely increases. The success rate is proportional to the Flight number.

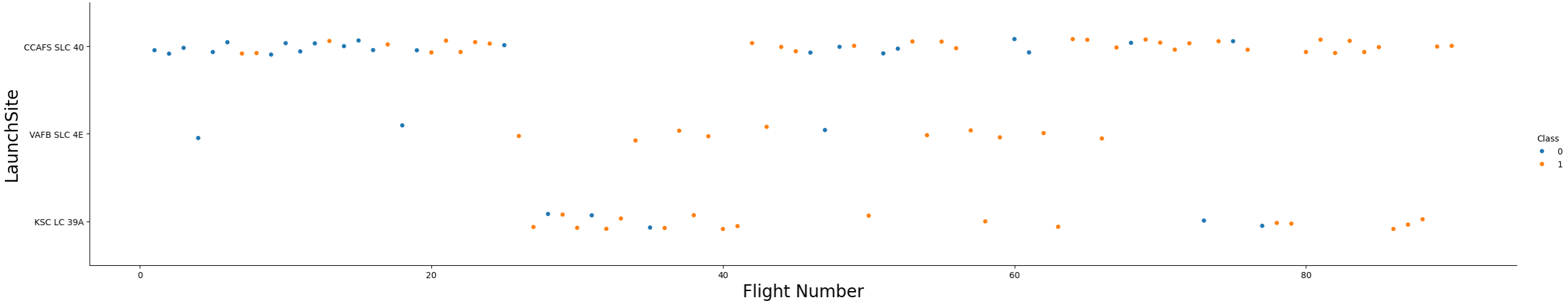
Payload mass on the other hand, has lower success rates when more massive. With more weight, it is less likely the 1st stage will land. It negatively correlates with a successful outcome.



FLIGHT NUMBER VS LAUNCH SITES

CCAFS SLC 40 seems to be where most of the early 1st stage missions took place. The success rates have increased with flight count. So, CCAFS SLC 40 seems to have the highest early failures of 1st stage landing.

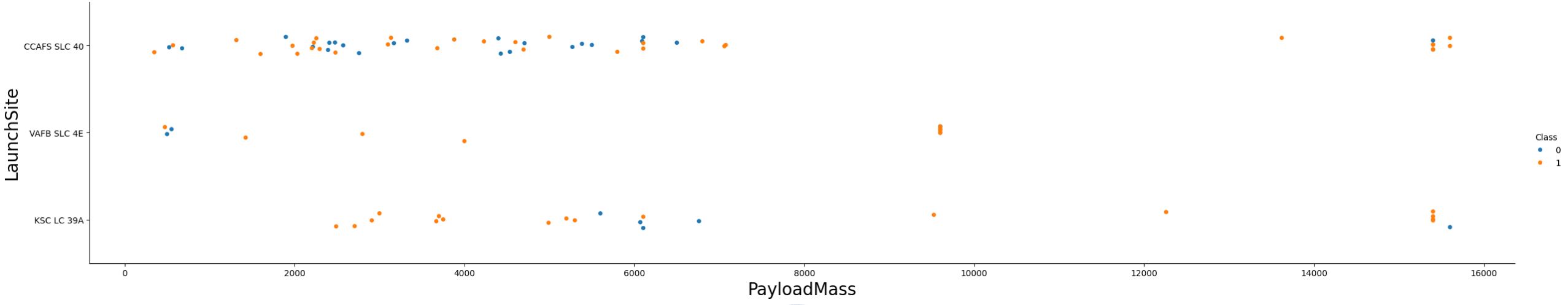
The first successes of VAFB SLC 4E and KSC LC 39A have been after a flight number of 20, unlike that for CCAFS SLC 40.



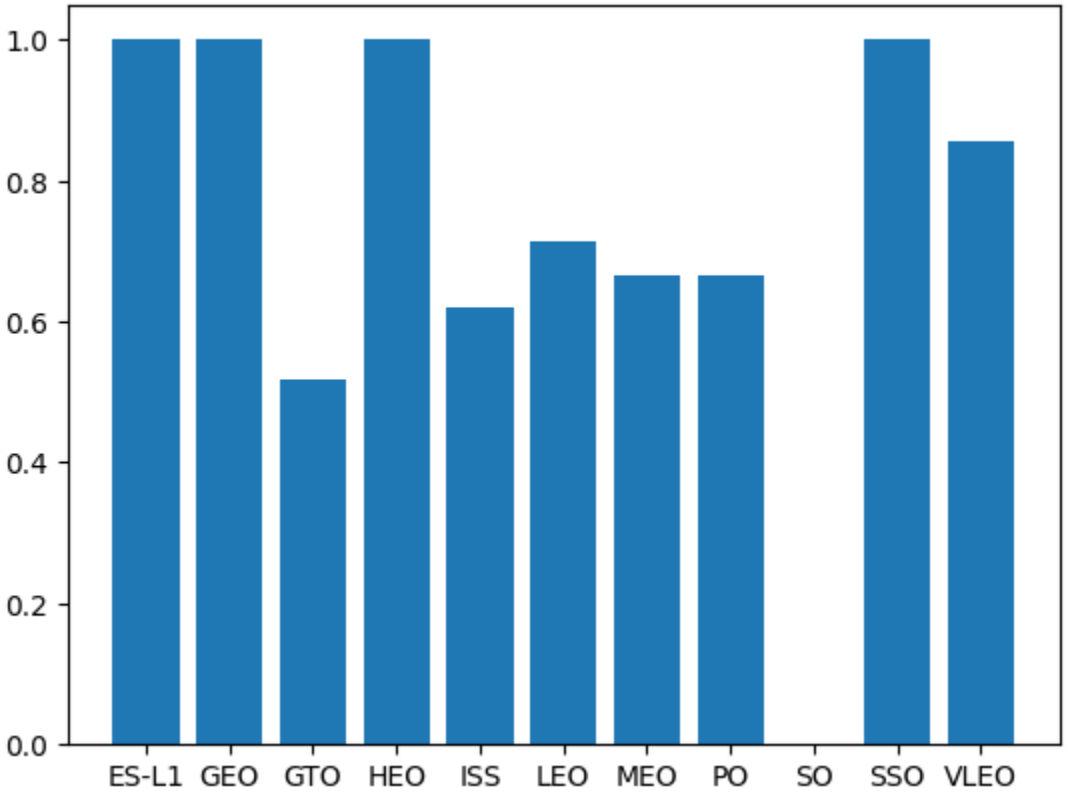
PAYLOAD VS LAUNCH SITES

CCAFS SLC 40 has higher success rates for heavier payloads. Both CCAFS SLC 40 and KSC LC 39A seem to be favored for heavy payloads. VAFB SLC 4E seems to be unfavored for heavier payloads.

KSC LC 39A has higher success rates than other launch sites for lighter payloads, that are below 6000.



ORBIT WISE SUCCESS RATE



100% success rates are observed for orbits:

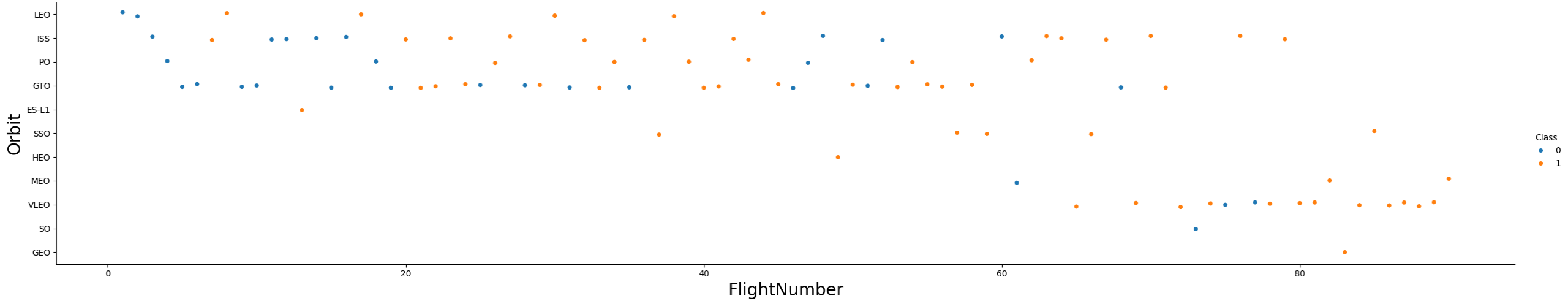
- Sun-Earth Lagrange Point 1 (ES-L1)
- Geostationary Equatorial Orbit (GEO)
- Highly Elliptical Orbit (HEO)
- Sun-Synchronous Orbit (SSO)

All orbits except for SO have had successful launches.

FLIGHT NUMBER VS ORBIT TYPE

In the LEO orbit, success rate seems to increase with the number of flights. Other orbits seem to have no specific correlation with the flight number otherwise.

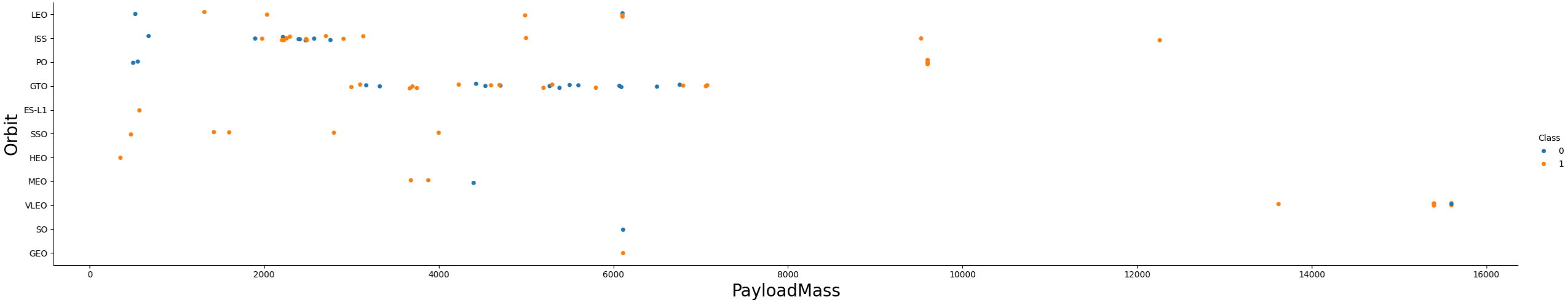
However, the flight count is positively correlated to the success rate.



PAYLOAD MASS VS ORBIT TYPE

Heavier payloads have higher success rates for Polar, LEO, and ISS. Lighter payloads have good success rates for SSO, HEO, and MEO.

Majority of launches have been attempted into orbits GTO, PA, ISS, and LEO. Heavy payloads has low modes.



LAUNCH SUCCESS YEARLY TREND (2010-2020)

Percentage increase of success remain non-negative between 2013 to 2017. It was stable in 2014.

Success rates declined in 2018, post which it has increased.

NAMES OF THE UNIQUE LAUNCH SITES

Unduplicated entries of Launch_Site entries were fetched from SPACEXTABLE. There are 4 launch sites present in the data set.

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

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

LAUNCH SITES THAT BEGIN WITH 'CCA'

5 records whose Launch_Stite begin with 'CCA' were fetched from SPACEXTABLE. LIKE criteria provided to start with CCA as CCA%. Delimiter of 5 entries included using LIMIT constraint.

```
%sql select * from SPACEXTABLE 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
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 BY NASA BOOSTERS

Aggregate SUM() was applied to payload_mass__kg_ over records from SPACEXTABLE whose customer is NASA (CRS), to get total payload mass in Kg.

```
[ ] %sql select sum(payload_mass__kg_) from SPACEXTABLE WHERE customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
sum(payload_mass__kg_)
45596
```

AVERAGE PAYLOAD MASS BY BOOSTER F9 V1.1

Aggregate AVG() was applied to payload_mass__kg_ over records from SPACEXTABLE whose booster_version is F9 v1.1, to get average payload mass in Kg.

```
[ ] %sql select avg(payload_mass__kg_) from SPACEXTABLE WHERE booster_version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
avg(payload_mass__kg_)
2928.4
```

DATE OF FIRST SUCCESSFUL GROUND PAD LANDING

Aggregate MIN() applied to DATE over records of SPACEXTABLE that fulfil criteria where landing_outcome is successful in ground pad, using WHERE clause.

```
[ ] %sql select min(DATE) from SPACEXTABLE WHERE landing_outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22

BOOSTERS WITH DRONE SHIP SUCCESS AND PAYLOAD BETWEEN 4000 AND 6000

Booster version is selected from SPACEXTABLE records that fulfil criteria where landing_outcome is successful in ground pad and payload mass is between 4000 kg and 6000 kg, using WHERE clause.

```
[ ] %sql select 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

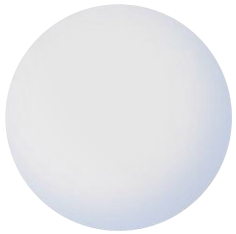
TOTAL COUNTS OF SUCCESSFUL AND FAILED MISSION OUTCOMES

Aggregate COUNT() was applied to mission_outcome over records of SPACEXTABLE to get counts of each landing outcome. This was categorized by GROUP BY clause for each mission outcome.

```
[ ] %sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

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

Mission_Outcome	count(mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1



BOOSTER VERSIONS WHICH HAVE CARRIED MAXIMUM PAYLOAD MASS

Aggregate MAX() over payload_mass__kg_ was applied for records in SPACEXTABLE to return the maximum value of payload mass. This formed the subquery. The maximum payload is 15600.

Booster version and corresponding payload mass is selected from SPACEXTABLE records that fulfil criteria where payload_mass__kg_ is equal to the returned maximum payload mass 15600, using WHERE clause.

```
[ ] %sql select booster_version, payload_mass__kg_ from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 LAUNCH RECORDS

Month, booster version, launch site, and month is selected from SPACEXTABLE records that fulfil criteria where landing_outcome is a failure for drone ship, for the year 2015 using WHERE clause.

The year is matched using the SUBSTR() query that returns the substring of characters from the Date attribute. Month is also selected this way.

```
[ ] %sql select  substr(Date,6,2) as Month, booster_version, launch_site from SPACEXTABLE where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

Aggregate COUNT() is used to rank the landing outcomes from SPACEXTABLE records that fulfil criteria where the date is between 2010-06-04 and 2017-03-20 using WHERE clause.

This is categorized by the landing_outcome values using GROUP BY clause. The results are ordered by descending counts using the ORDER BY clause.

```
[ ] %sql select count(landing_outcome), landing_outcome from SPACEXTABLE where DATE between '2010-06-04' and '2017-03-20' group by landing_outcome order by count(landing_outcome) desc
```

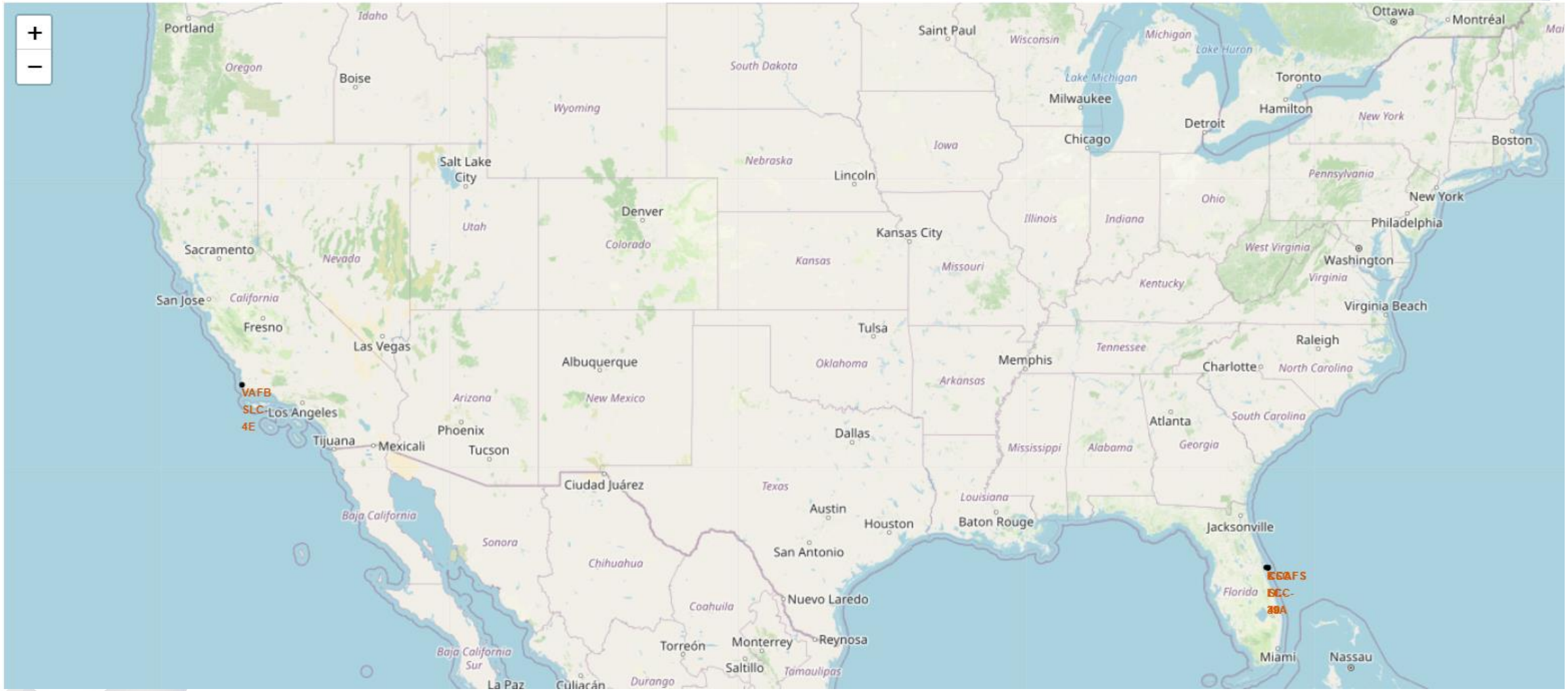
```
 * sqlite:///my_data1.db
Done.
```

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

6. Launch Sites Proximities Analysis

Wednesday, Oct 16, 2024

ALL LAUNCH SITES



ELEMENTS

Map is created with an initial center location. **Circles** are added to highlight an area around each site given their coordinates. Launch sites are labelled by **Markers** with their names. Each Marker is added to the Map.

QUESTIONS

Are all launch sites in proximity to the Equator line?

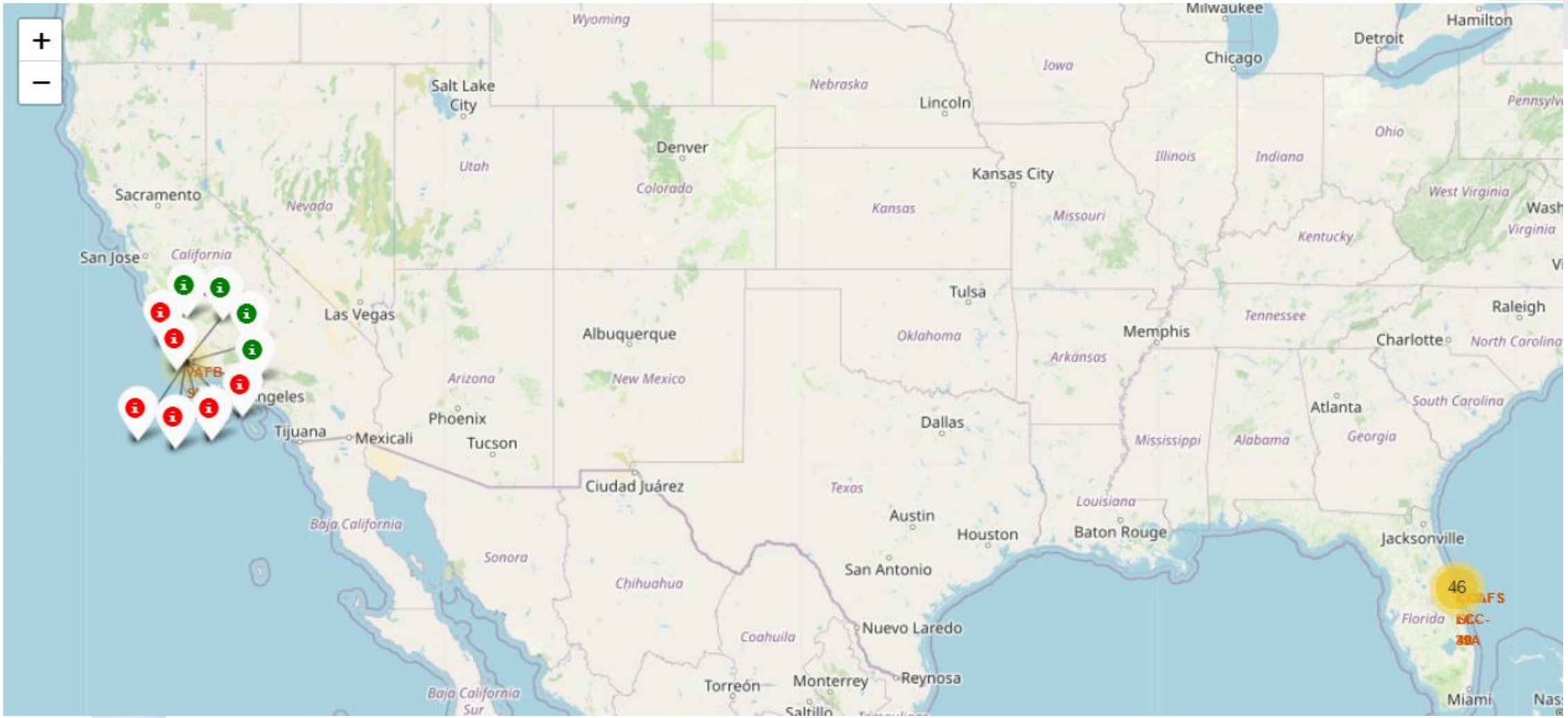
Are all launch sites in very close proximity to the coast?

FINDINGS

Yes, the launch sites are found to be in proximity to the Equator line and coast.

1. Earth's rotational velocity, being fastest at the equator can supplement to extra speed, reducing the fuel load and cost required. This location leads to minimal inclination changes, leading to better fuel efficiency.
2. Coastal regions are comparatively remote and safe to carry launch operations. Being remote, trajectory variations are increased, recovery from damage are easier, and noise pollution is reduced.

LAUNCH OUTCOMES PER SITE

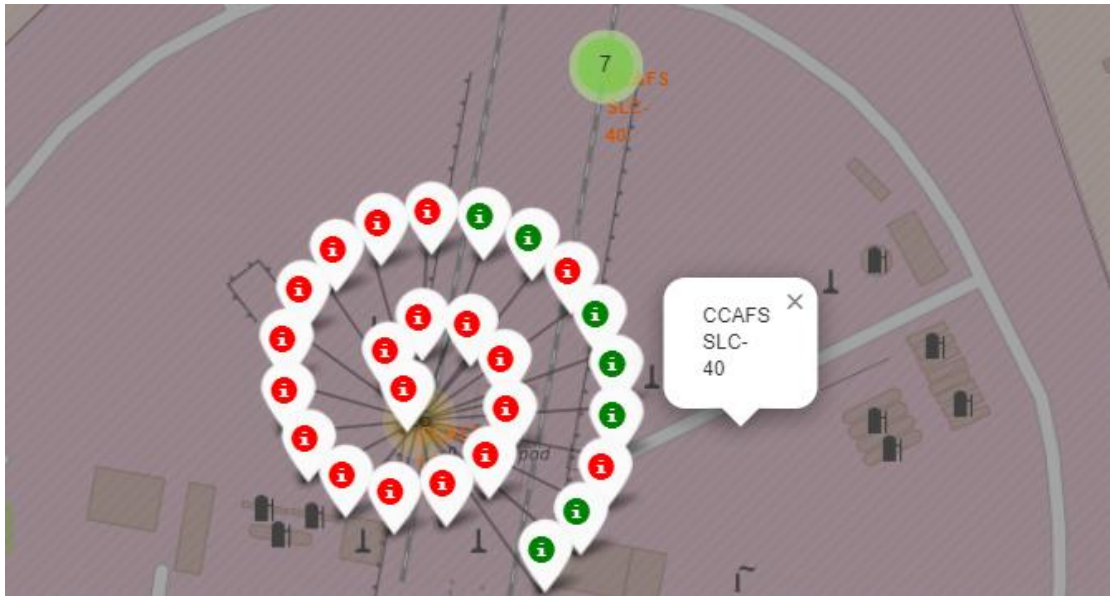


ELEMENTS

MarkerCluster used to contain multiple **Markers** having the same coordinate, for each site. A successful launch is represented by a Green Marker, while a failed launch is represented by a Red Marker. The color is represented in the Icon Color. Each Marker is added to the MarkerCluster of the Map.

FINDINGS

- KSC has majority successes of 76.9%.
- Majority of the launches for CCAFS and VAFB have been failures.



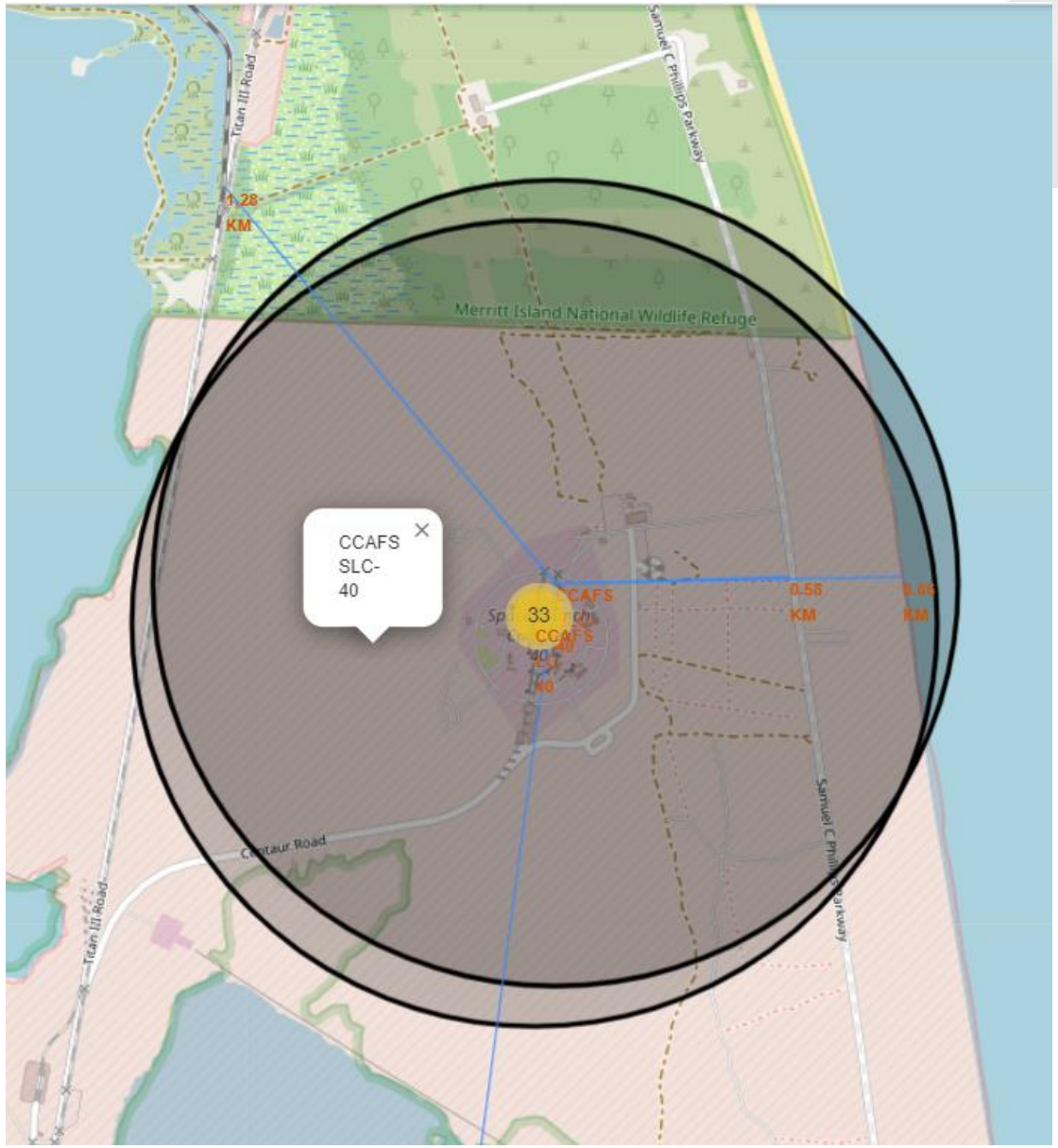
LAUNCH SITE DISTANCE TO PROXIMITIES

ELEMENTS

MousePosition is used to get coordinate for a mouse over any point. This is added to the Map. This is used to fetch the coordinates for proximities of nearest railways, highways, and coastlines.

The coordinates are used to calculate the distance from the launch site by Haversine formula. **Marker** is used to label the distance between each proximity. These are added to the Map.

PolyLine is drawn from the launch site to each proximity on providing their coordinates. These are added to the Map.



Launch sites are in close proximity to railways, highways, and coastline. They are maintained at a certain distance from cities.

QUESTIONS

Are launch sites in close proximity to railways?

Yes. Railways can be used to carry freight load, machinery, and raw materials required for assembly, product development, and launch. Transportation is made easier.

Are launch sites in close proximity to coastline?

Yes. Coastlines are remote and sparsely populated when compared to cities. This location is ideal due to the following few reasons:

- Wider range of trajectories.
- Safety from risk of falling debris into landmass.
- Reduced noise, airspace, and radio frequency interference with populated areas.
- Easier recovery from ocean.

.

QUESTIONS FINDINGS

Are launch sites in close proximity to highways?

Yes. Highway transportation of human resources and materials is allowed.

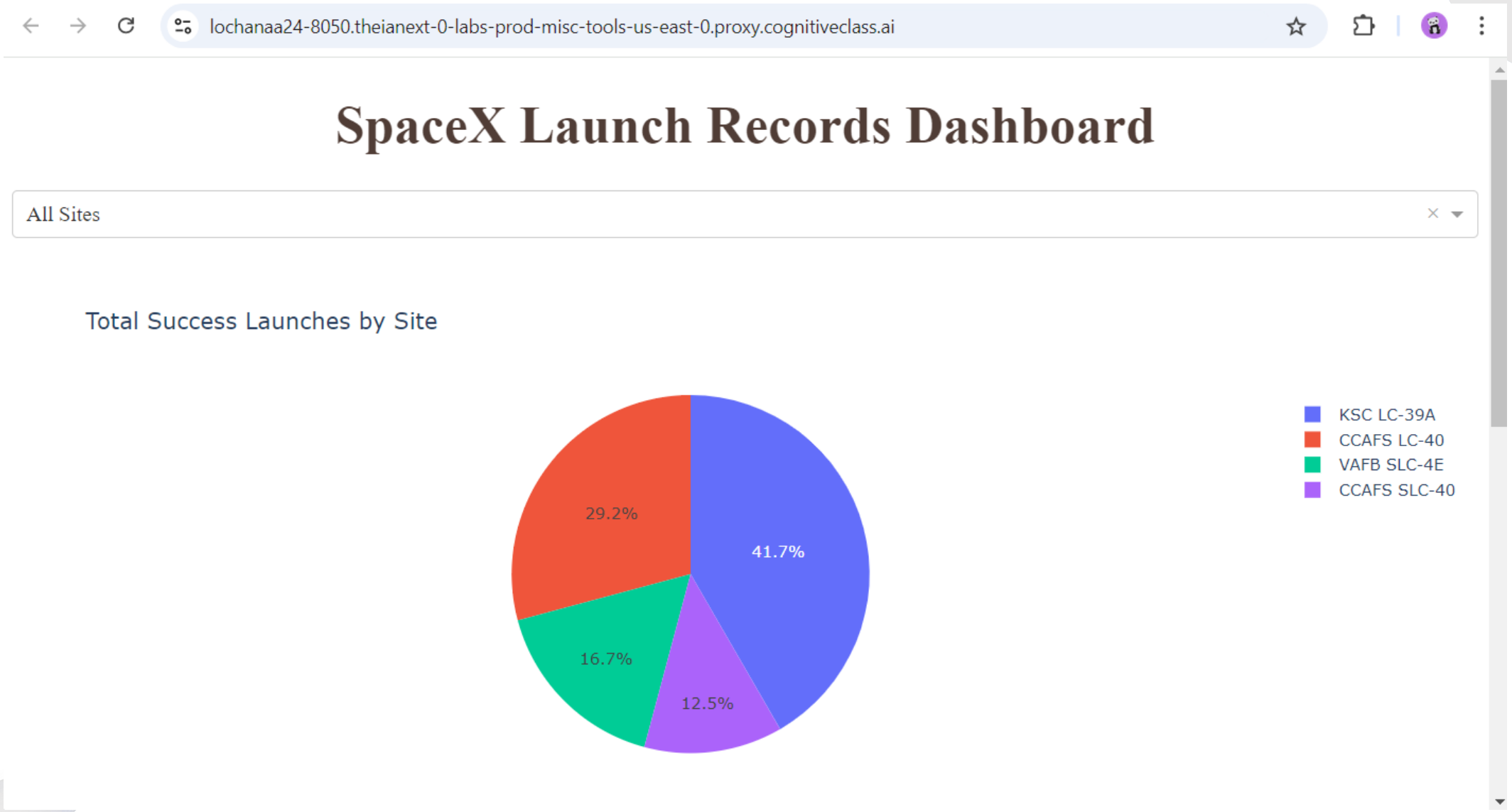
Do launch sites keep certain distance away from cities?

Yes. Cities are populated areas and are not remote. Launch sites are located away from cities due to following few reasons:

- Lower noise interference and pollution from cities.
- Safety from risk of falling debris into landmass.
- Launch safety protocols, in case of emergencies.

6. Launch Data Reporting and Dashboarding

SUCCESS PERCENTAGE FOR ALL SITES

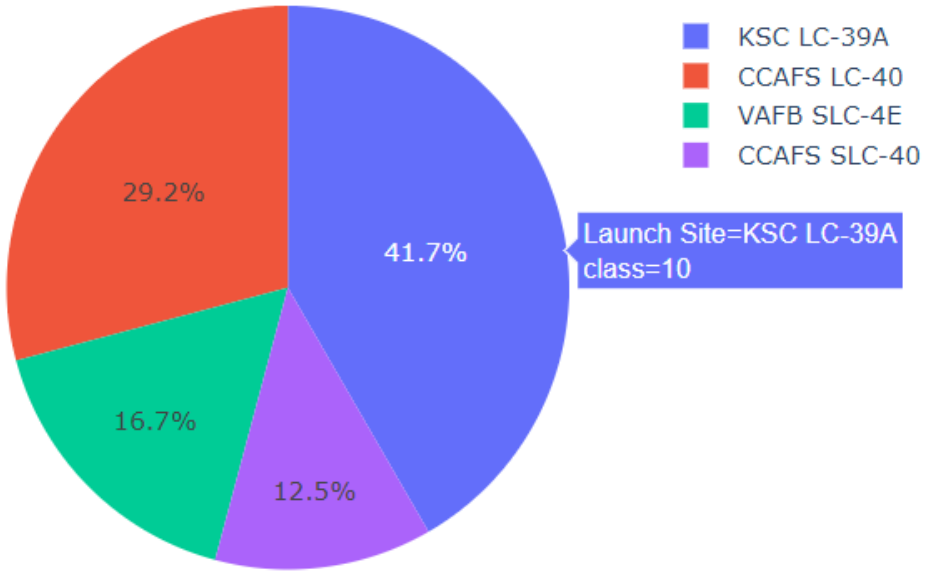


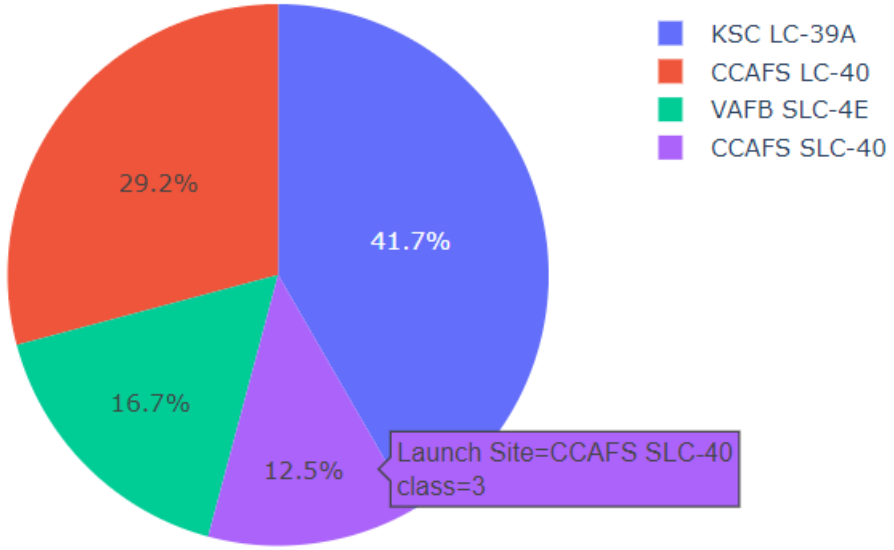
ELEMENTS

Dropdown is added to list available 4 launch sites and a combined selection of all sites. A **Pie Chart** that represents the success percentage of each site is displayed on selecting the *All Sites* option. Along with it is a **Legend** expressing the color mapped to each site.

FINDINGS

- KSC LC-39 has the maximum success rate, with a value of 41.7%. 10 successful launches were recorded.
- CCAFS SLC-4E has the minimum success rate, with a value of 12.5%. Only 3 successful launches were recorded.





SUCCESS PERCENTAGE FOR KSC LC-39A

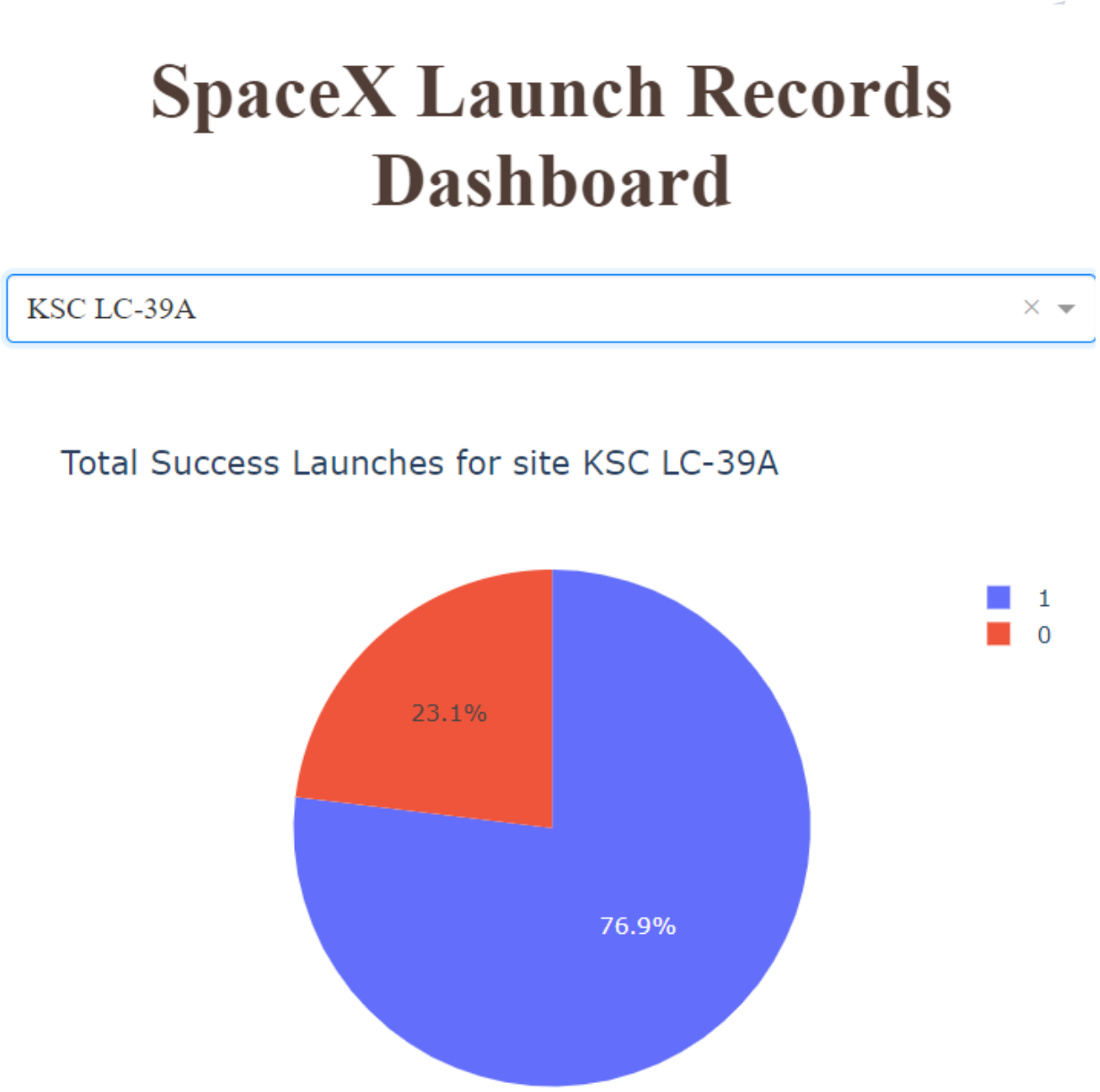
ELEMENTS

The launch site with highest success rate is KSC LC-39A.

A **Pie Chart** representing the success and failure percentages is displayed on selecting the *KSC LC-39A* option in the **Dropdown**. The **Legend** expresses Class=0 being a success and Class=1 being a failure.

FINDINGS

- The success rate of the site is 76.9%.
- The failure rate of the site is 23.1



PAYLOAD VS LAUNCH OUTCOME

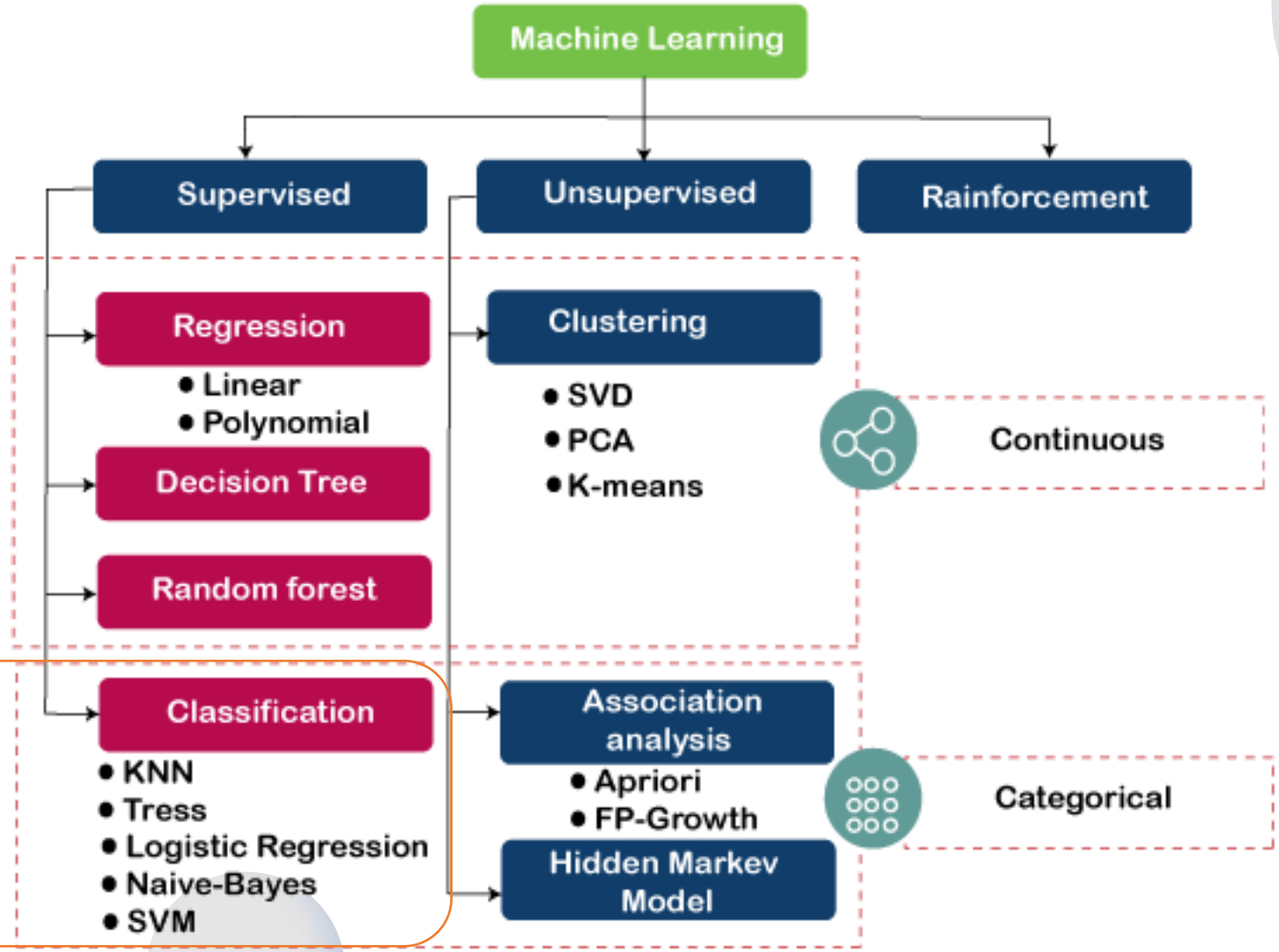
SCATTER PLOT

The payload range with the highest success launches is between 2,000 kg to 4,000 kg, followed by the payload range of 4,000 to 6,000 kg, as can be seen by the count of points. Payloads above 6,000 kg have the lowest success rates.

Booster version FT has the highest successful launches, followed by B4 with the second highest among all booster versions. Booster v1.1 has the lowest successful launch count.

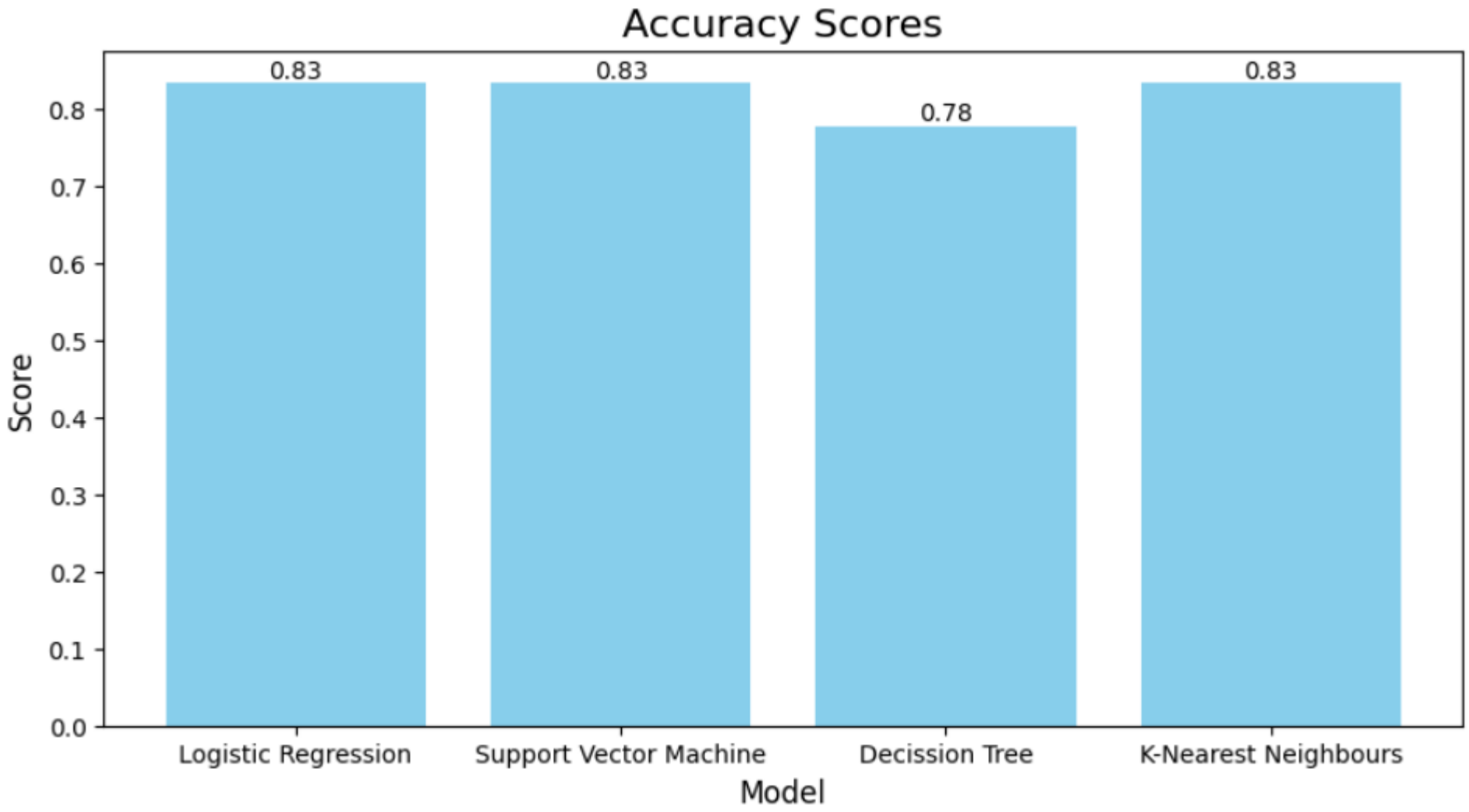
7. Predictive Analysis

Classification

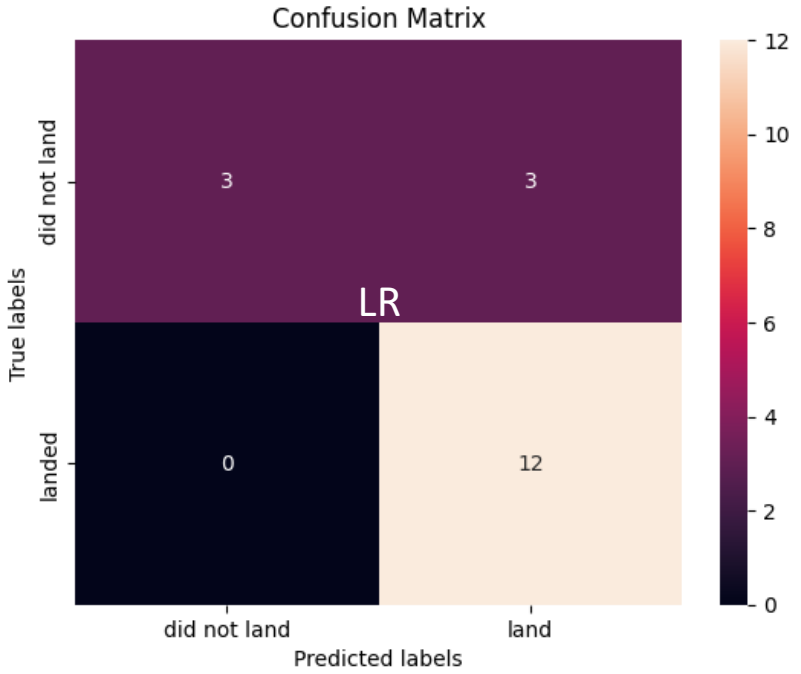


CLASSIFICATION ACCURACY (83.3%)

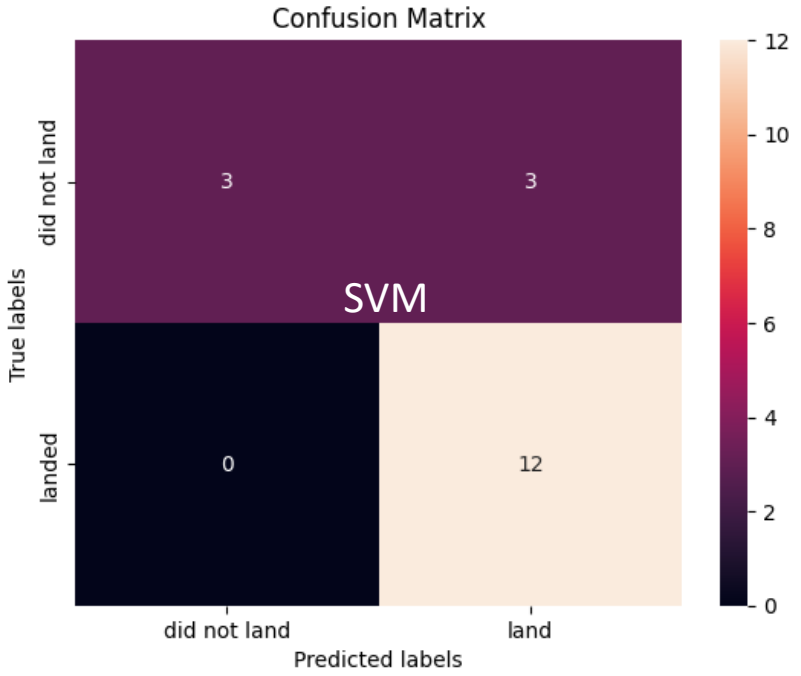
Logistic Regression, SVM, and KNN classifiers have the highest accuracy score of **0.83(bar).**

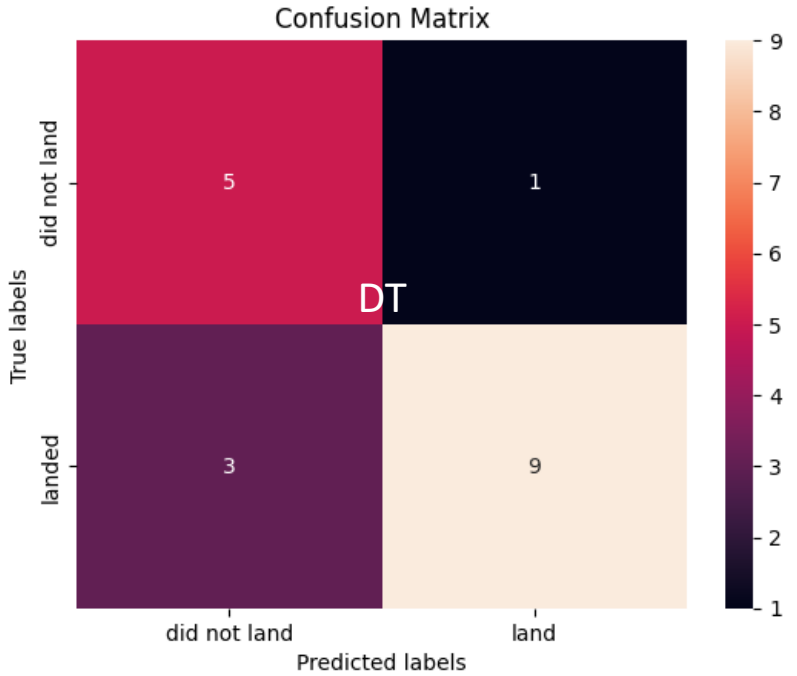


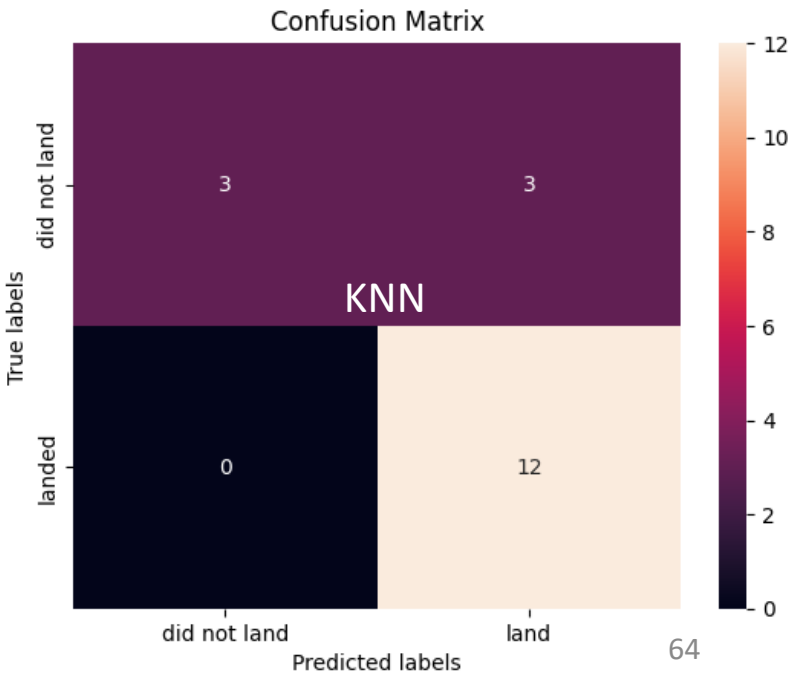
CONFUSION MATRIX



- LR, SVM, and KNN models favorable as their confusion matrix show that they predicted all 12 successful landing correctly, with 0 false negatives.
- The Decision Tree model predicted only 9 successful landing correctly and had 3 false positives.
- All LR, SVM, and KNN models have the same accuracy of 83.33% as portrayed previously, hence these are the best performing models.





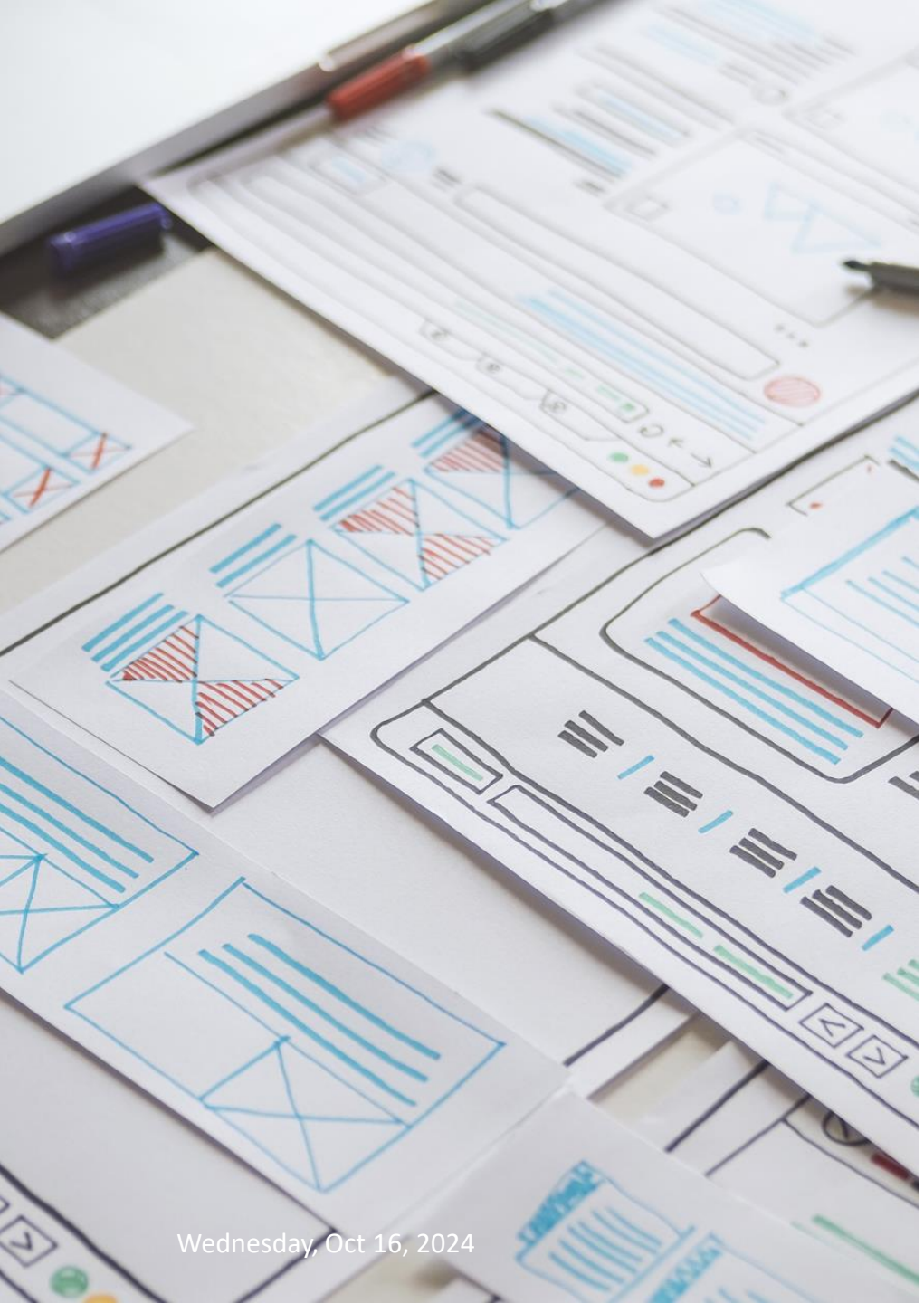


8. Conclusion

This study provides valuable insights into the factors influencing the success of Falcon 9 1ST stage landings and the potential for accurate landing predictions using machine learning models.

The observations revealed that **flight experience and payload mass significantly impact landing success**, with lighter payloads showing higher success rates. KSC LC-39A emerged as the most reliable launch site, particularly for lighter payloads, while VAFB SLC-4E performed poorly for heavier payloads. Over time, launch success rates have shown a positive trend, with occasional setbacks, indicating improvements in operational efficiency. Number of flights were positively correlated with success rates while payload mass was found to be inversely correlated with success rates, where payloads over 6,000 kg exhibited lower reliability. The proximity analysis highlighted the strategic importance of locating launch **sites near the equator and coastlines**, maximizing fuel efficiency and safety while ensuring easier recovery and minimizing environmental impact. Sites are also conveniently connected to **railways and highways** to facilitate logistics and transportation. As per data analysis methodology, prepared data was subject to supervised classification. **Logistic Regression, SVM,** and **KNN** models were the best performers, achieving **83.33% accuracy** and correctly predicting all the successful landings without false negatives. However, all models struggled with false positives, where failed landings were incorrectly predicted as successful, indicating room for further refinement.

Overall, this study demonstrates that data-driven analysis and classifier models can effectively predict launch success, aiding in launch planning and cost estimation. Future work could focus on refining predictive models to reduce false positives and incorporate additional features such as weather conditions and technical parameters to improve accuracy.



References

1. Wikimedia Foundation. List of falcon 9 and Falcon Heavy launches. Wikipedia. *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*
2. Postman. R/spacex API docs. r/SpaceX API Docs. *https://docs.spacexdata.com/*
3. Folium#. Folium - Folium 0.16.1.dev76+g2921126e documentation. *https://python-visualization.github.io/folium/latest/*
4. Dash documentation & user guide. Plotly. *https://dash.plotly.com/*
5. Coursera. (n.d.). Applied Data Science Capstone. Coursera. *https://www.coursera.org/learn/applied-data-science-capstone*

**as per instructions from Coursera Applied Data Science Captsone Report (Module 5)*

Appendix I: Data Set

Below are a few key attributes present in the data set.

List of Falcon 9 and Falcon Heavy Launches

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Column	Description
Flight No.	Unique identifier for each launch
Date and Time	Launch timestamp
Booster Version	Type of first stage booster
Launch Site	Location of the launch
Payload	Description of the payload
Payload Mass	Mass of the payload in kilograms
Orbit	Target orbit for the payload
Customer	Entity that owns the payload
Launch Outcome	Result of the launch (e.g., success, failure)
Booster Landing	Result of the first stage landing

Appendix II: Languages and Tools

Language: Python and SQL

Libraries:

1. BeautifulSoup → Web Scraping
2. SQLite3 → Database
3. Folium → Map Plot
4. Dash
5. Plotly

   Dashboarding
6. Numpy
7. Pandas
8. Seaborn
9. Scikit-Learn
10. Matplotlib

    EDA and Visualisation

Tool: Google Colaboratoy

‘ Let the data speak ’

Thank you
Sahana Ashok