IBM

Applied Data Science Capstone

Alvin Kariri

9th August 2025

Outline

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

EXECUTIVE SUMMARY

- In this capstone project the goal is to develop a machine learning model to predict the likelihood of a SpaceX Falcon 9 rocket landing successfully using historical launch data.
- The main steps in the project involved:
  - Data collection through the SpaceX API followed by wrangling and processing.
  - Exploratory Data Analysis (EDA)
  - Interactive data visualizations and querying the data through a database in order to make gainful insights into the data.
  - Deploying an interactive dashboard using dash app and plotly.
  - Model training and evaluation
- The visualizations show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.
- One of the major insights gained was the success rate of various booster versions for the rocket.
- From the model training and evaluation, the decision tree classifier was concluded to be the best machine learning algorithm for predictions

INTRODUCTION

- This capstone project focuses on predicting whether the Falcon 9 first stage will land successfully—a critical factor in SpaceX’s cost advantage. While competitors charge $165 million per launch, SpaceX offers launches at $62 million, primarily due to first-stage reusability. By analyzing landing success, we can:
  - **Estimate Launch Costs**: Accurately predict expenses for SpaceX and competitors.
  - **Optimize Mission Planning**: Identify high-probability landing conditions to reduce risk.
  - **Support Competitive Bidding**: Provide data-driven insights for rivals bidding against SpaceX.
- While some "unsuccessful" landings are intentional (e.g., controlled ocean splashdowns for testing), our model distinguishes these from true failures. Using machine learning, we analyze key launch features—payload mass, orbit type, launch site, and booster version—to answer the core question:
- **"Given a Falcon 9’s launch parameters, will its first stage land successfully?"**
- This tool not only predicts outcomes but also uncovers hidden patterns in SpaceX’s landing success, offering strategic insights for the space industry.

Section 1

Methodology

METHODOLOGY

- The overall methodology includes:
  1. Data collection, wrangling, and formatting, using:
     - SpaceX API
     - Web scraping (Beautiful Soup)
  2. Exploratory data analysis (EDA), using:
     - Pandas and NumPy
     - SQLite3
  3. Data visualization and dashboard deployment using:
     - Matplotlib and Seaborn
     - Folium
     - Dash
  4. Machine learning prediction, using the techniques:
     - Logistic Regression
     - Support Vector Machine (SVM)
     - Decision Tree Classifier
     - K-nearest Neighbors (KNN)

METHODOLOGY

Data collection

- To build a predictive model for Falcon 9 landing success, data was collected from two primary sources:
  1. SpaceX API
     - Source: [SpaceX API]
     - Method:
       - Called /launches, /launch sites, /cores, and /payloads endpoints.
       - Filtered for Falcon 9 missions using rocket ID falcon9.
     - Key Data:
       - Booster version, core reuse count, and precise landing status.
  2. Web Scraping (Wikipedia)
     - Source: [List of Falcon 9 launches]
     - Method:
       - Used BeautifulSoup to parse HTML tables.
       - Extracted: *Launch date, payload, launch site, and landing outcome*.
- Integration:
  - Merged Wikipedia (historical context) with API (technical details) using pandas.
  - Final dataset: 178 launches (2012–2023) with 12+ features (e.g., payload_mass_kg, landing_success).
- Tools: Python (requests, BeautifulSoup, pandas), Jupyter Notebook.

METHODOLOGY

1 Data collection (SpaceX API call)

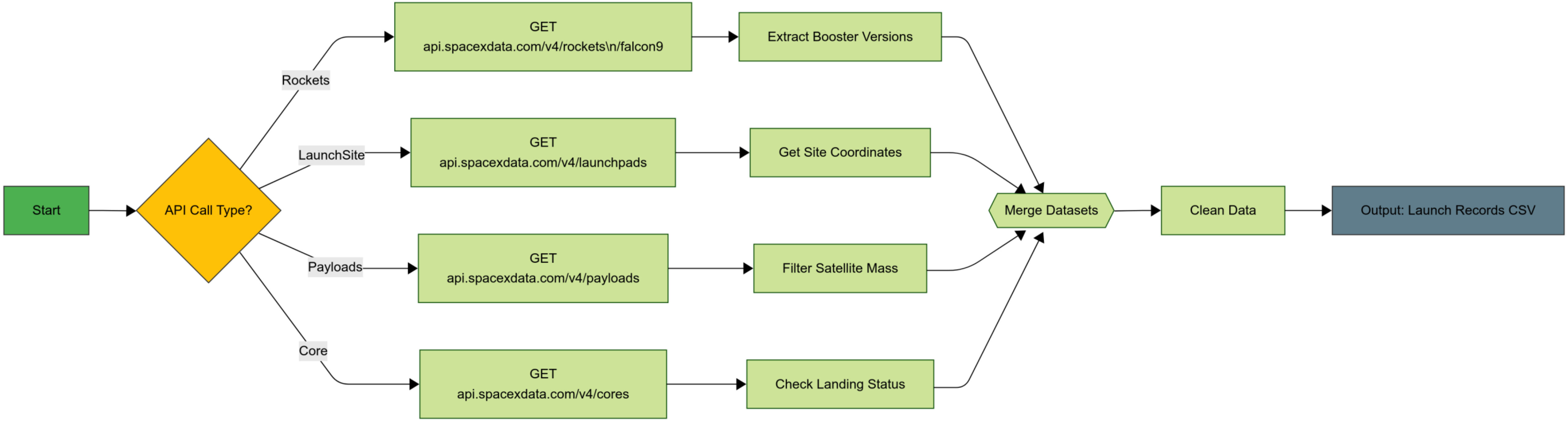
1. SpaceX API
    - The static rest API used is SpaceX API.
    - Four functions were used to append the API URL in order to obtain the individual datasets for: rocket(booster versions), cores, launch sites and payload mass.
    - The individual datasets are merged into one launch records csv.
    - The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
    - In the data frame there were missing values in the payload mass column which were replaced by the mean value of the column. The data frame is later loaded into a csv for later use.
    - We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
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
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
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
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
		2013-				CCSFS SLC	None										

METHODOLOGY

1 Data collection (SpaceX API call)

- <u>SpaceX API Flowchart</u>:



GitHub URL: [GitHub API Repo Link]

METHODOLOGY

1 Data collection (Web Scraping)

2. Web scraping
   - The data is scraped from List of Falcon 9 Launches.
   - The website contains only the data about Falcon 9 launches.
   - We use the get request to obtain the html page in text form then use Beautiful soup to extract the launch data by parsing through the html elements.
   - We then use pandas to create a data frame by parsing the launch html tables.
   - We end up with 121 rows or instances and 11 columns or features.
   - After this we convert the data frame into a csv to be used later on.
   - The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

METHODOLOGY

1 Data collection (Web scraping)

- Web scraping Flowchart:

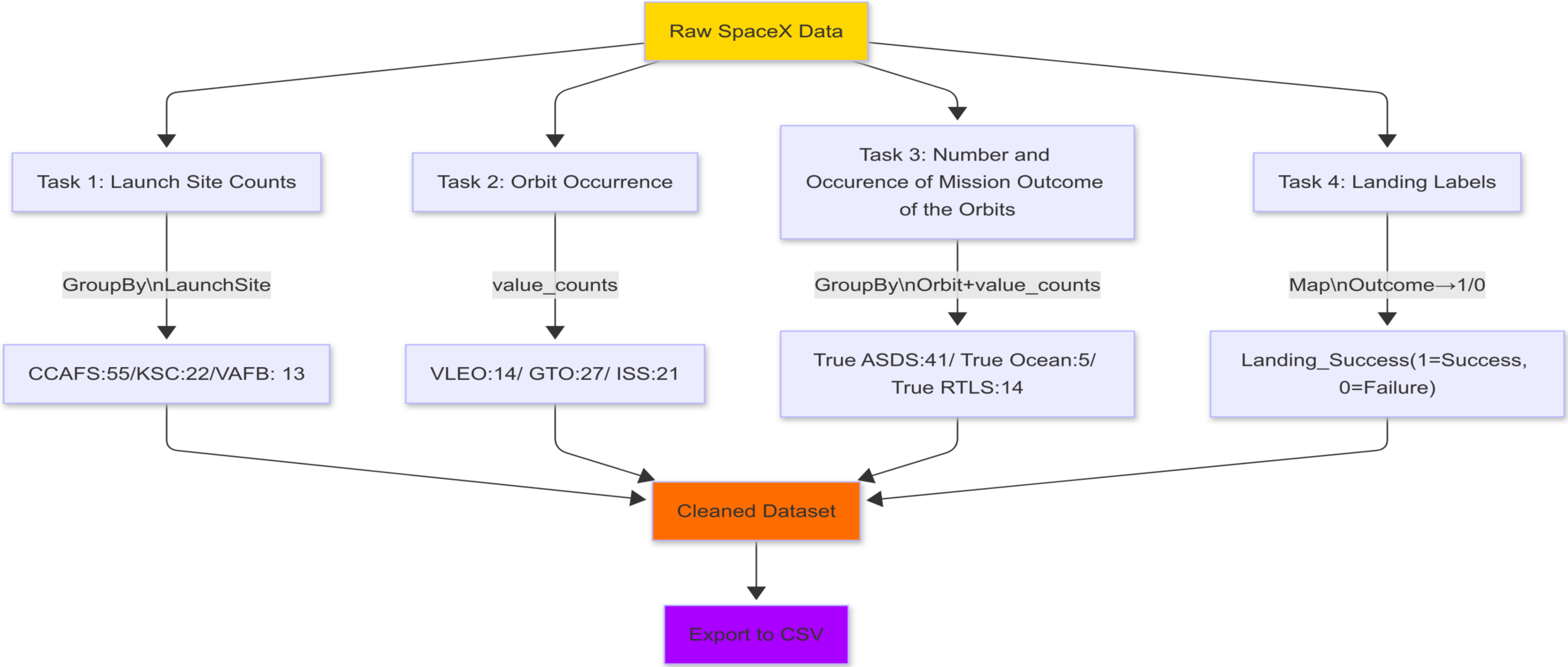
GitHub URL: GitHub Web Scraping Repo Link

METHODOLOGY

1 Data wrangling

- In the data wrangling stage, we load the csv created from the data frame that was created using launch data from the SpaceX API call. The data has 90 rows and 17 columns.
- We first start by identifying numerical and categorical columns. This info will be useful later on.
- After this we determine the number of unique launch sites which are found to be three. This info helps us determine the number of launches per site. Each launch is aims to a dedicated orbit. These sites are:
  - Complex 40 **VAFB SLC 4E**
  - Vandenberg Air Force Base Space Launch Complex 4E **(SLC-4E)**
  - Kennedy Space Center Launch Complex 39A **KSC LC 39A**
- Orbit distribution is determined next by mapping the frequency of each orbit. There are 11 orbit types.
- Thereafter we determine the frequency of landing outcomes. Of the seven possible outcomes 3 belong to the successful category while the remaining indicate failure/non landing.
- Using this information, we create an extra column in the data frame called ‘Class. The column uses binary classification to determine whether landing was successful or not with ‘0’ representing a failed landing and 1 a successful landing. The dataset is loaded into a csv and is now ready for EDA and visualization.

1 Data wrangling Flowchart



GitHub URL: [GitHub Data Wrangling Repo Link]

METHODOLOGY

2 Exploratory Data Analysis (EDA) with SQLite3

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
    - The number of launches on each launch site
    - The number of occurrence of each orbit
    - The number and occurrence of each mission outcome
- SQL
  - The data is queried using SQLite3 to answer several questions about the data such as:
    - The names of the unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1

GitHub URL: GitHub SQL Repo Link

METHODOLOGY

3 Data Visualization with Matplotlib and Seaborn

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type

GitHub URL: GitHub EDA Data Visualization Repo Link

METHODOLOGY

3 Data Visualization with Folium

- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps. This is achieved through the use of folium objects. These objects include:
    1. Markers
       - These are map objects used to pinpoint specific locations on a site map (e.g., launch sites).
       - They are added in order to:
         - Highlight key sites (e.g., **KSC LC-39A** for high-success launches).
         - Provide clickable popups with details (successful launch sites).
    2. Circles
       - These are radius-based zones around points of interest.
       - They are added in order to:
         - Visualize landing hazard areas
         - Show payload mass impact
    3. Lines
       - These are polylines used to connect two different points of coordinates on a map.
       - They are added in order to show the distance between the launch site and its proximities.
    4. GeoJSON layers
       - These are vector overlays.
       - They are used in order to:
         - To display regional success rate (color-coded by launch site)
         - Integrate external geospatial data

- The Folium library is used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

GitHub URL: [GitHub Folium Maps Repo Link]()

METHODOLOGY

3 Data Visualization with Dash by plotly

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - The dropdown allows the user to choose the specific launch site whose data you want to visualize.
  - The range slider is added in order to customize the payload mass ranges.
  - This allows scatterplots for different payload masses and the mission outcome to be plotted.
  - This allows us to gain insights on the relationship between and payload mass and the mission outcome.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site.
    - The correlation between payload mass and mission outcome (success or failure) for each launch site.

GitHub URL: GitHub Plotly Dash App Repo Link

METHODOLOGY

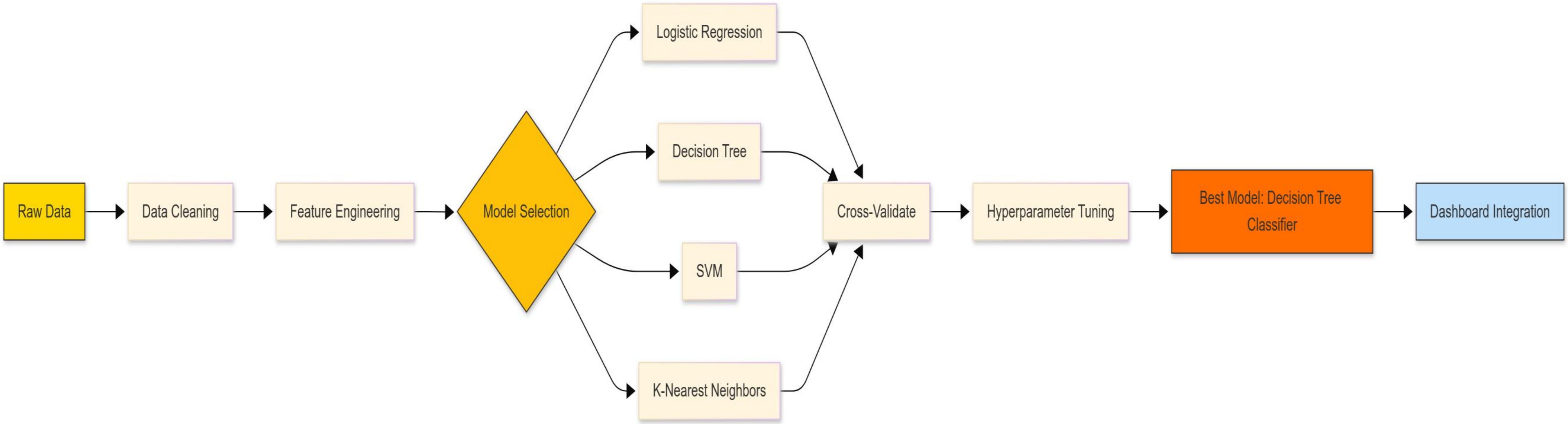
4 Machine Learning Prediction

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Use one-hot encoding to turn categorical data into numerical data
  - Standardizing the data
  - Splitting the data into training and test data.
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Test model accuracy on the testing set
  - Find the best combination of hyperparameters for each model using grid search and folds
  - Evaluate the models based on their accuracy scores and confusion matrix

METHODOLOGY

4 Machine Learning Prediction

- Machine Learning Flowchart:



GitHub URL: GitHub Machine Learning Prediction Repo Link

RESULTS

- The results are split into 5 sections:
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis

1. <u>Exploratory Data Analysis with Visualization</u>
   - From the analysis we were able to determine:
     - Flight Number vs. Payload Mass: Show trends in success rates over time and payload constraints.
     - Launch Site Analysis: Revealed site-specific success patterns.
     - Temporal Improvements: Demonstrated SpaceX's progress in landing reliability over successive missions.

2. <u>SQL (EDA with SQLite3)</u>
   - The analysis done using SQL was able to determine the following key insights:
     - Identify unique launch sites
     - The average payload mass for booster version F9 v1.1 which was determined to be 2,534 kg, suggesting its suitability for medium-weight payloads.
     - Success Rate: 98 out of 101 missions were successful (Success), with only 1 in-flight failure.
     - Max Payload Carriers: Boosters like F9 B5 B1048.4 and F9 B5 B1049.4 carried the heaviest payloads (~16,000 kg).

RESULTS

3. Folium
   - The folium results enable us to determine the location of launch sites on the map with a tremendous degree of accuracy.
   - It also enables us to create geolocation tags for mission launches so we can visualize where successful/failed launches occurred.
   - Folium also allows us to determine the proximity of launch sites to various points of interest.
4. Dash
   - Dash by plotly allows us to create customizable dashboard that allows anyone with no programming language to create visualizations of the launch data ranging from graphs and pie charts to scatterplots and bar charts.
   - In our case we used dash to create a dashboard that can visualize the percentage of successful/failed missions grouped by launch sites and also create a scatterplot that shows the correlation between payload mass, booster versions and success rate.
   - The range of the payload mass is controlled by a slider that has a rang of 0 – 10,000 kg which more than satisfies all our case scenarios.
5. Predictive analysis (Machine learning models)
   - Using sci-kit learn along with pandas we can import modules for machine learning packages.
   - The data was split into training and test data and fed into the ML algorithms. The ML models picked for training were: k-nearest neighbors, SVMs, logistic regression and decision tree classifiers.
   - Various parameters were fitted for each individual model. This is thanks to GridSearch which allows for that.
   - After training the models we are able to get the best parameters for each model. Afterwards test data was used on the models and the accuracy scores were obtained. This predictions and test data were used to generate confusion matrices and give us insight into which model is best for prediction.

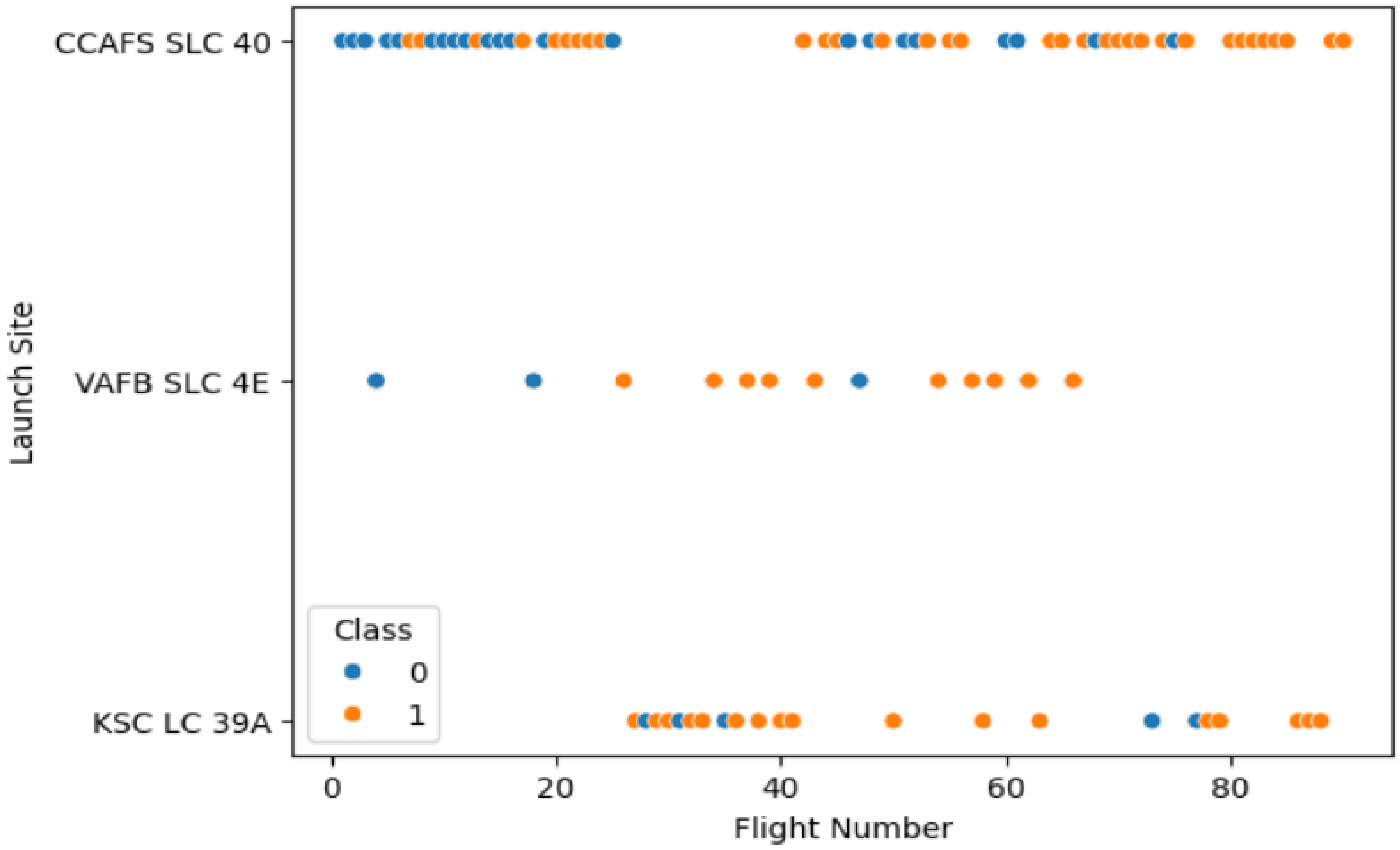
Section 2

Insights drawn from EDA

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between flight number and launch site:

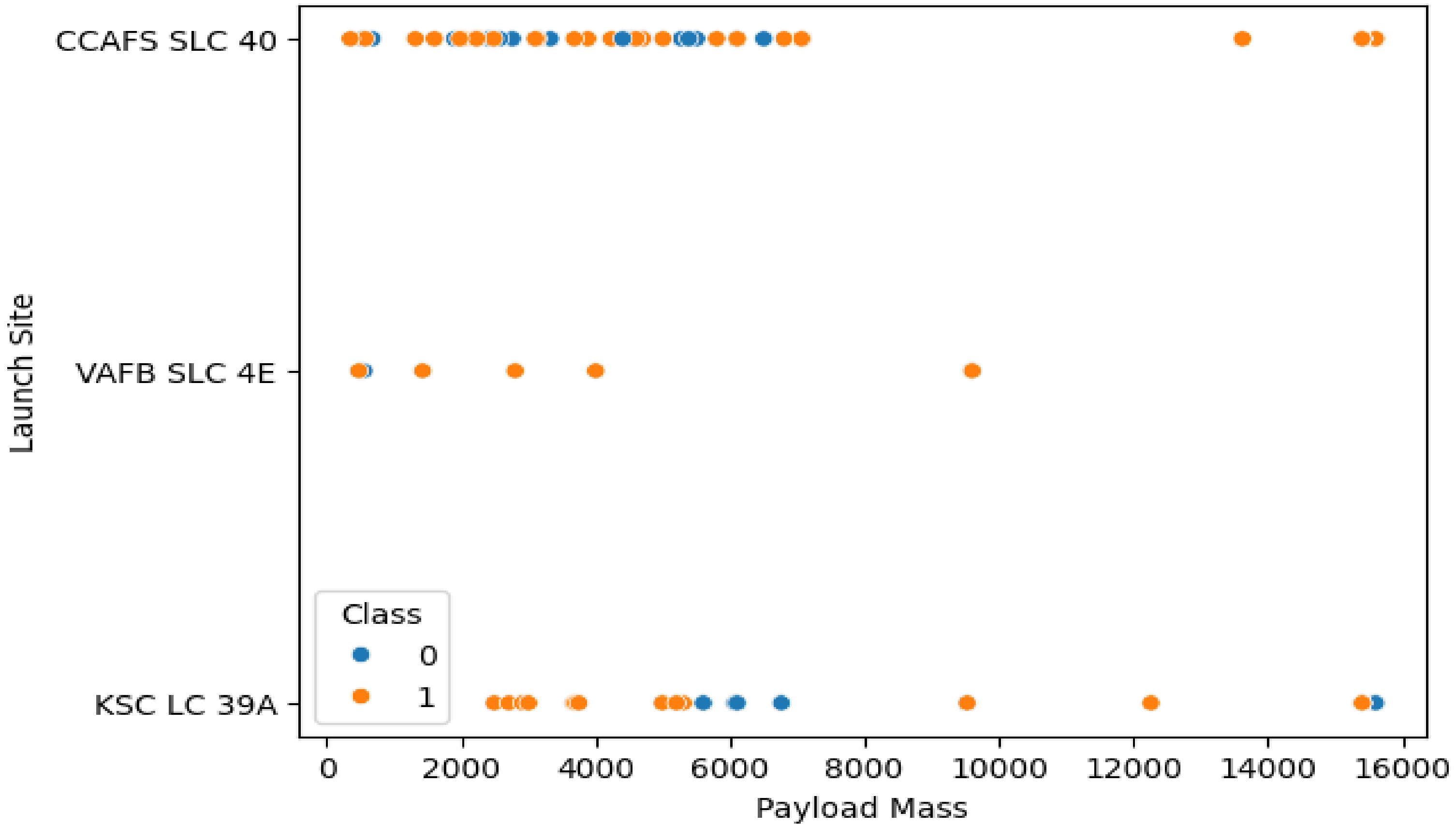


- From the scatterplot we can see that with an increase in flights there was a corresponding increase in successful mission outcomes.
- This may be due to an improvement in iterative design of the rockets over time.
- Launch site CCAFS SLC 40 has the greatest number of flights indicating its consistency and reliability as a launch site.
- VAFB SLC 4E has the fewest number of flights which might indicate its underlying unreliability as a consistent launch site.

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and launch site:

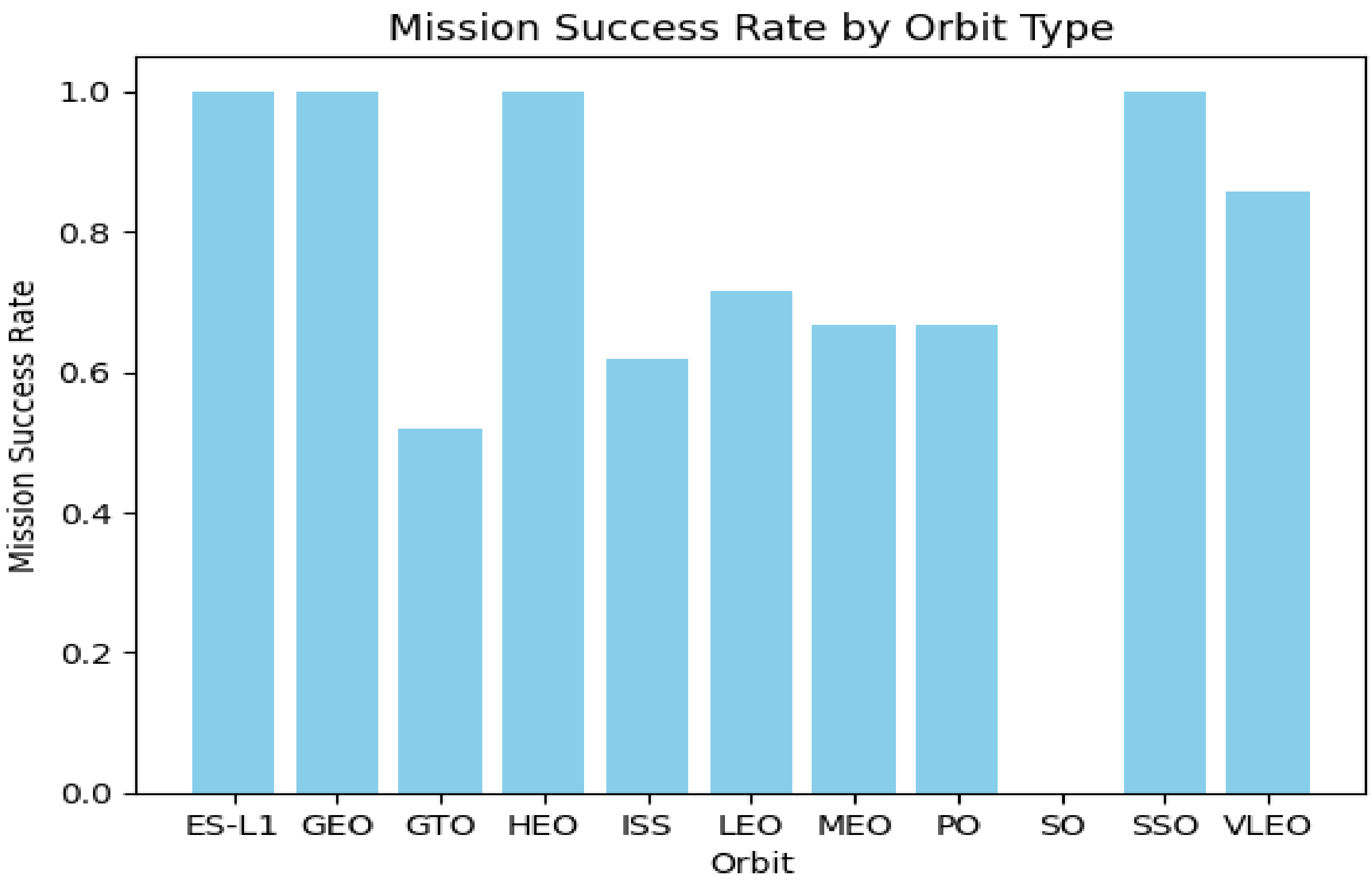


- From the scatterplot we can see a higher payload mass correlates with increased risk (more Class 0 outcomes), especially at CCAFS SLC 40.
- Lighter payloads are consistently successful across all sites.
- Site selection may depend on payload mass, with CCAFS SLC 40 handling heavy/large missions and others optimized for lighter loads.

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between mission success rate and orbit type:

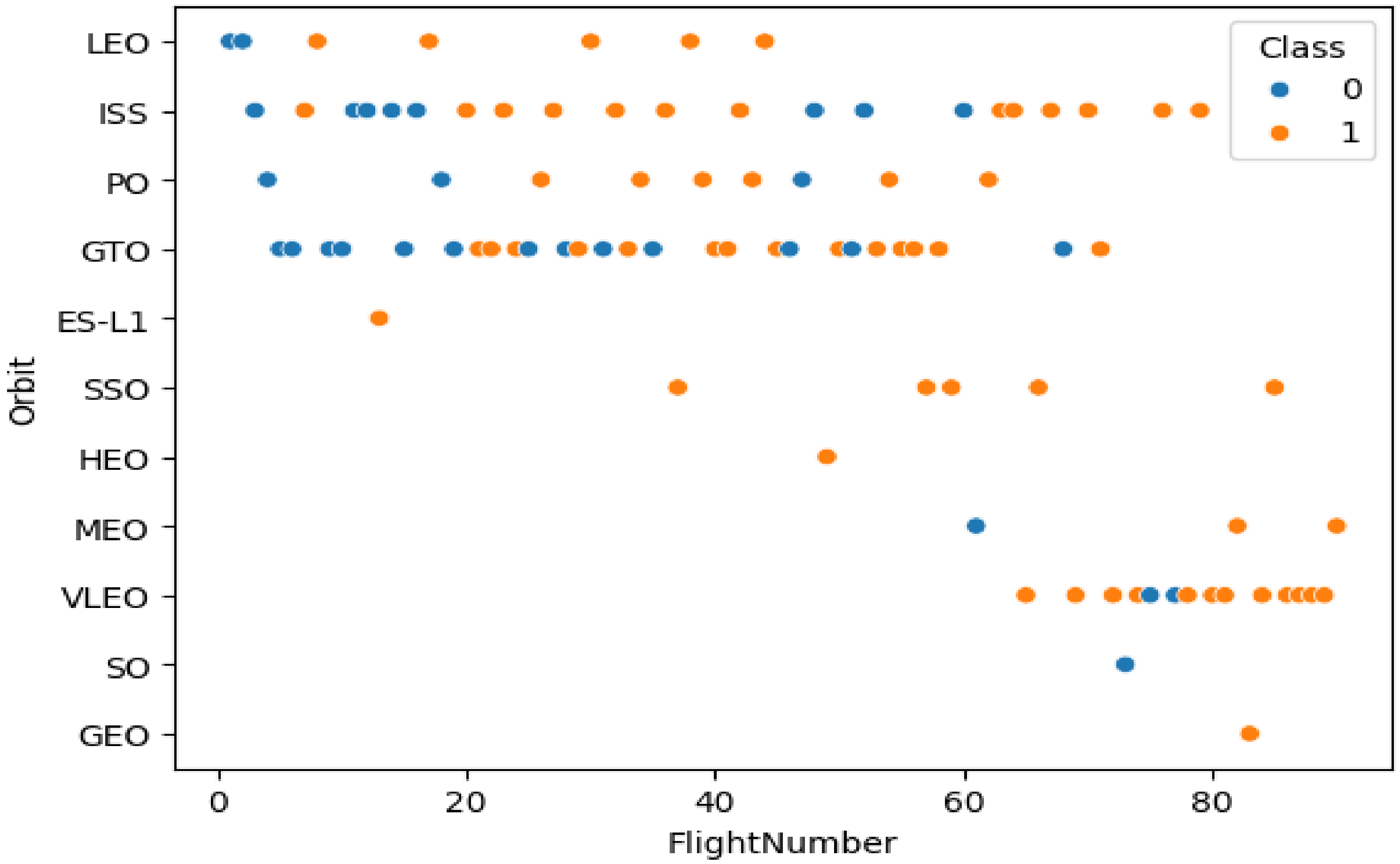


- Established orbits (LEO, GEO, GTO) have high success rates, reflecting operational maturity.
- Specialized/complex orbits (ES-L1, HEO, VLEO) face higher failure risks, suggesting technical hurdles or less optimized launch systems.
- Mission planning could prioritize proven orbits for critical payloads, while newer orbits may require further testing.

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between flight number and orbit type:

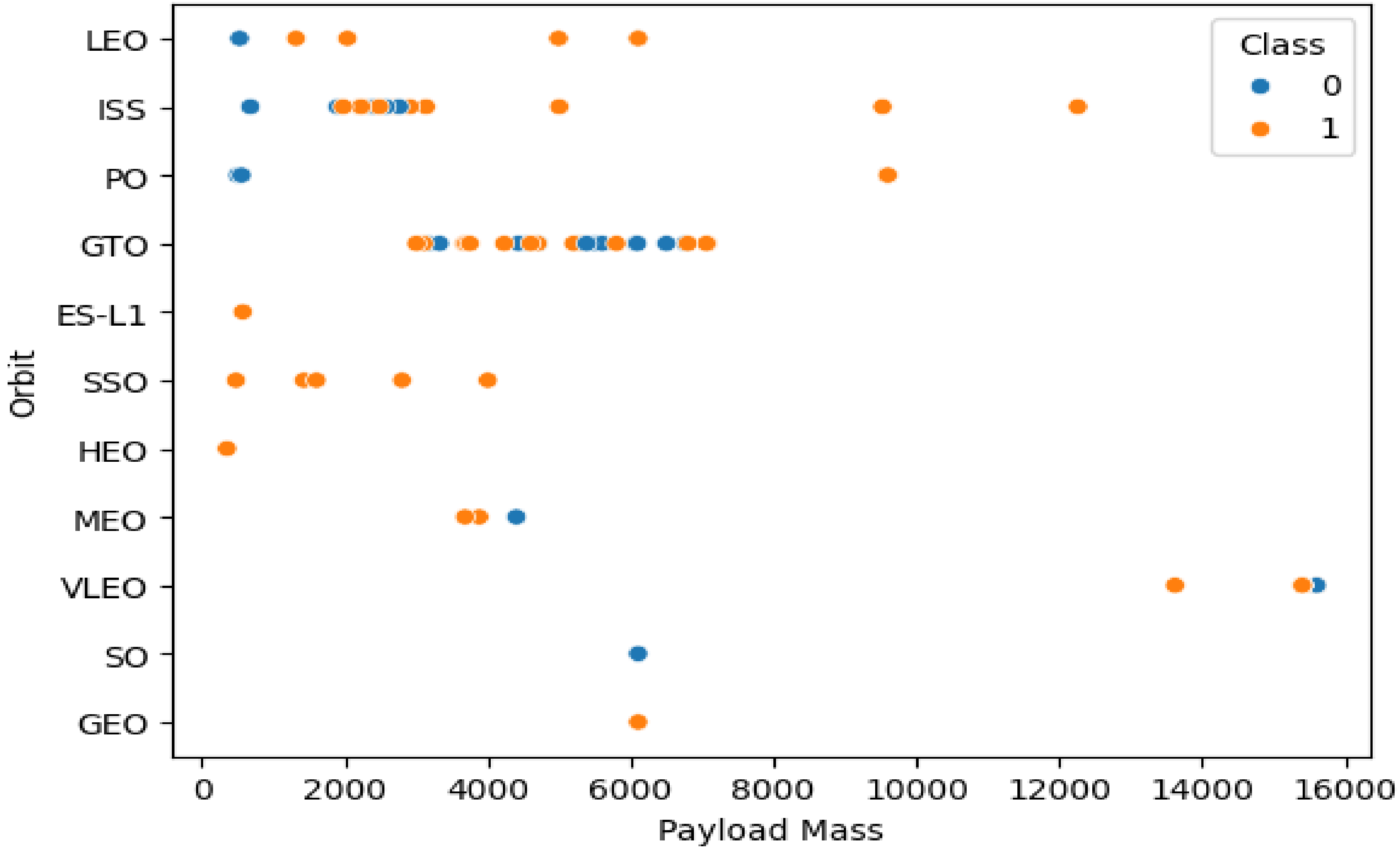


- **LEO** and **GEO** dominate mission history, while newer/niche orbits emerge as flight numbers increase.
- Flight number progression may correlate with technological advancement (e.g., later missions tackle complex orbits).
- Gaps in certain orbits (e.g., **SO**, **HEO**) suggest limited use cases or higher risk.
- Early flights (lower numbers, left side) may focus on **GTO, GEO, or LEO**, indicating foundational missions.
- Later flights (higher numbers, right side) show diversification into **SSO, PO, MEO**, suggesting expanded capabilities.

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and orbit type:

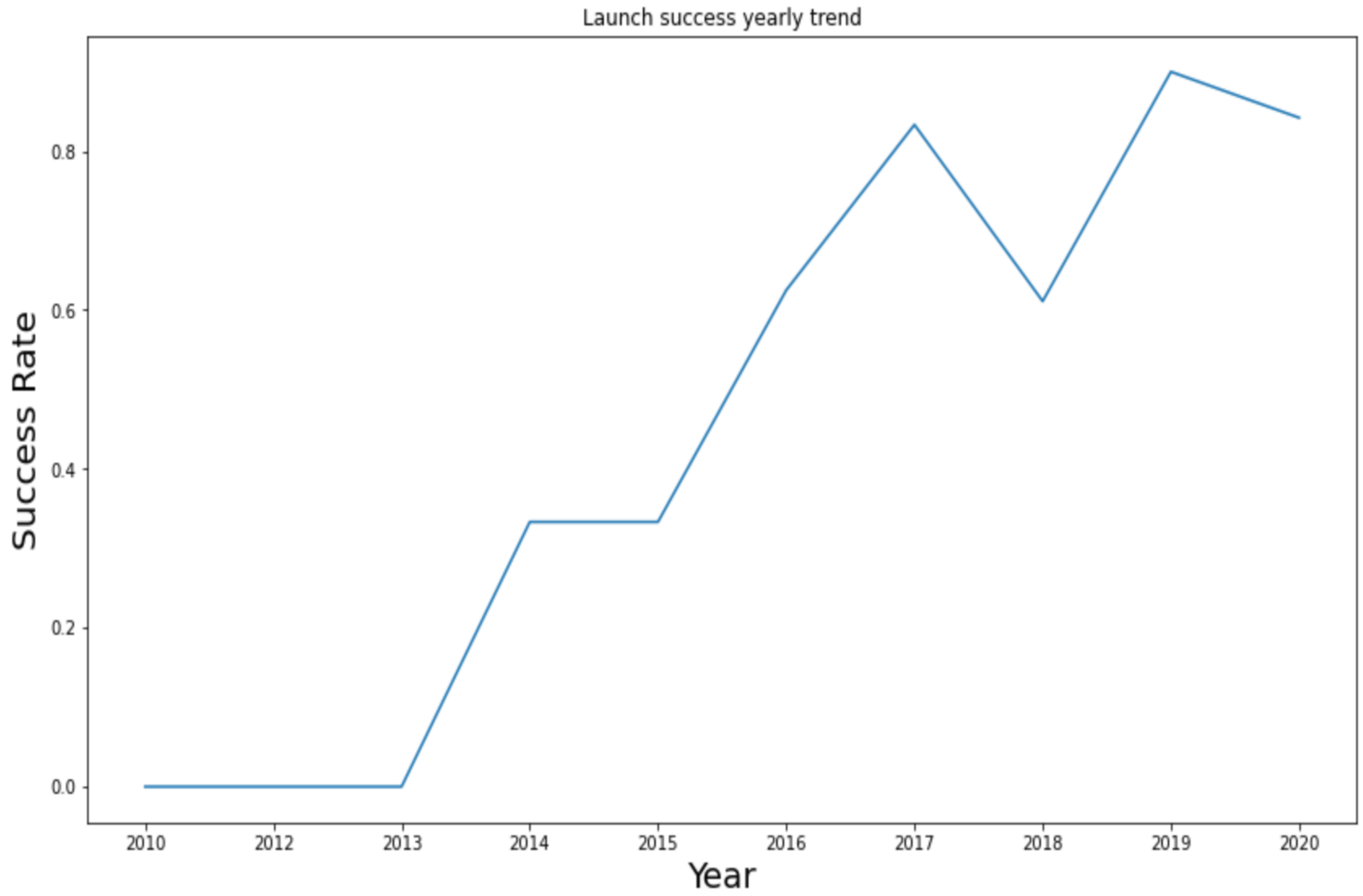


- Orbits like LEO (Low Earth Orbit), ISS, GTO (Geostationary Transfer Orbit), and SSO (Sun-Synchronous Orbit) have more data points, indicating frequent usage.
- Some orbits (e.g., SO, VLEO, HEO) have very few data points, indicating rarer missions.
- Most payloads cluster below 8000 kg, though some values go up to ~16,000 kg.
- GEO and GTO orbits tend to have heavier payloads.
- Lighter payloads are more common in SSO, LEO, and PO missions.

RESULTS

1 Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and orbit type:



- **Early years (pre-2014):** 0% success rate (no launches or all failures).
- **2014–2015:** Modest improvement (~33% success).
- **2016:** Surge to >60% (reliability improvements).
- **2017:** Peak at ~85% (strong performance).
- **2018:** Drop to ~60% (setbacks).
- **2019:** Best year (~95% success).
- **2020:** Slight decline (~85%), still high.
- **Trend:** Clear progress post-2013, reflecting maturing launch systems.

RESULTS

2 SQL (EDA with SQL)

- The list contains four distinct launch sites:
  - **CCAFS LC-40** and **CCAFS SLC-40** are separate designations at Cape Canaveral (likely different pads or historical naming conventions).
  - **KSC LC-39A** (Kennedy Space Center) and **VAFB SLC-4E** (Vandenberg Space Force Base) are independent sites.
- The names of the unique launch sites in the space mission:

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

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

RESULTS

2 SQL (EDA with SQL)

- The query filtered records where Launch_Site starts with 'CCA' (all from Cape Canaveral, Florida).
- All 5 missions were successful, early SpaceX flights (2010–2013) using Falcon 9 v1.0.
- Payloads include Dragon spacecraft demos and NASA CRS missions to the ISS.
- Note: Two early flights had parachute landing failures, while others made no landing attempt.

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

RESULTS

2 SQL (EDA with SQL)

- The total payload mass carried by boosters launched by NASA (CRS):

Total payload mass by NASA (CRS)
45596

- The result may be useful for assessing overall launch capacity or mission volume.
- The query result below shows the average payload mass carried by booster version F9 v1.1:

Average payload mass by Booster Version F9 v1.1
2928

- Helps assess performance capabilities of F9 v1.1 compared to other variants.
- The date when the first successful landing outcome in ground pad was achieved occurred on December 22, 2015.

Date of first successful landing outcome in ground pad
2015-12-22

- This date marks a **historic milestone** in SpaceX's reusable rocket technology.
- It demonstrates the first proven capability to recover and potentially reuse orbital-class boosters.

RESULTS

2 SQL (EDA with SQL)

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- The query list shows Falcon 9 booster versions that met these criteria:
  - Successfully landed on a drone ship (autonomous ocean platform).
  - Carried payloads between 4,000–6,000 kg (mid-range mass missions).
- Key Observations:
  - Includes multiple variants:
    - Early models: F9 v1.1 (e.g., B1011, B1016).
    - Upgraded versions: F9 FT (e.g., B1021.2, B1032.1).
    - Block 4/5 iterations (e.g., B1040.1, B1051.2), showing reusability progress.
  - Payload range (4k–6k kg) suggests missions like:
    - ISS resupply (e.g., CRS flights).
    - Mid-sized satellite deployments (e.g., GTO Comsats).
- Significance:
  - Demonstrates reliable drone ship landings for moderately heavy payloads.
  - Highlights SpaceX's iterative booster improvements (v1.1 → Block 5).

RESULTS

2 SQL (EDA with SQL)

- The query result below shows the total number of successful and failure mission outcomes:

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

- **98.0% Success Rate** (100/101 missions) excluding the unclear payload case.
- **99.0% Success Rate** if counting all "Success" labels (101/102).
- The single **in-flight failure** suggests:
  - A rare but critical anomaly (e.g., propulsion/structural issue).
  - High reliability overall, with failures being outliers.
- Notes**:**
  - The **"payload status unclear"** case may require investigation (partial success?).
  - The duplicate "Success" entry should be audited for data consistency.

RESULTS

2 SQL (EDA with SQL)

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

- The query result show the list of names of booster versions which have carried the maximum payload mass.
- Key observations made are:
  1. All Boosters Are Block 5 Variants
     - Indicates SpaceX's most advanced and capable Falcon 9 iteration (e.g., B1048.4, B1051.6).
     - Higher-numbered suffixes (e.g., .4, .6) suggest multiple reuses of the same booster.
  2. Payload Capacity Insight
     - These boosters were selected for missions requiring maximum mass-to-orbit performance.
     - Likely used for:
       - Heavy GTO communications satellites (~5,500–8,000 kg).
       - Crew Dragon or cargo resupply missions with max capacity.
  3. Reusability Demonstrated
     - Boosters like B1049.7 (flown 7 times) show high reuse counts, proving reliability even under heavy payload demands.

RESULTS

2 SQL (EDA with SQL)

- The query result below shows list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- January 2015: F9 v1.1 B1012 failed to land on a drone ship after launching from CCAFS LC-40.
- April 2015: F9 v1.1 B1015 also failed during a drone ship landing attempt from CCAFS LC-40.
- Both failures occurred early in SpaceX’s **reusability program**, when drone ship landings were experimental.
- These were part of **CRS (Cargo Resupply)** missions to the ISS, where landing attempts were secondary to primary payload delivery.
- Highlights the learning curve in SpaceX’s reusable rocket development.
- All failures were with Falcon 9 v1.1, an early iteration later upgraded for reliability.
- By late 2015, SpaceX achieved its first successful drone ship landing (December 2015).

RESULTS

2 SQL (EDA with SQL)

- The query result shows the frequency of landing outcomes between 2010 and 2017:

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

- **Most frequent outcomes**:
  - "No attempt" (10x): Early missions (2010–2013) often skipped landings to prioritize payload delivery.
  - Drone Ship Success/Failure (5x each): Reflects experimental phase of ocean landings.
  - Ground Pad Success (5x): Includes historic first landings at LZ-1 (e.g., December 2015).
- **Less Frequent Outcomes**:
  - Controlled Ocean (3x): Intentional water landings (e.g., expendable missions).
  - Uncontrolled Ocean (2x)/Parachute Failure (2x): Early recovery attempts (2010–2012).
  - Precluded Drone Ship (1x): Aborted landing due to technical issues.
- **Trends & Implications**:
  - 2010–2013: Focus on mission success, not recovery ("No attempt" dominates).
  - 2014–2017: Rise in landing attempts, with 50% drone ship success rate (5/10 attempts succeeded).
  - Progress: By 2017, ground pad success became routine, paving the way for Block 5 reusability.

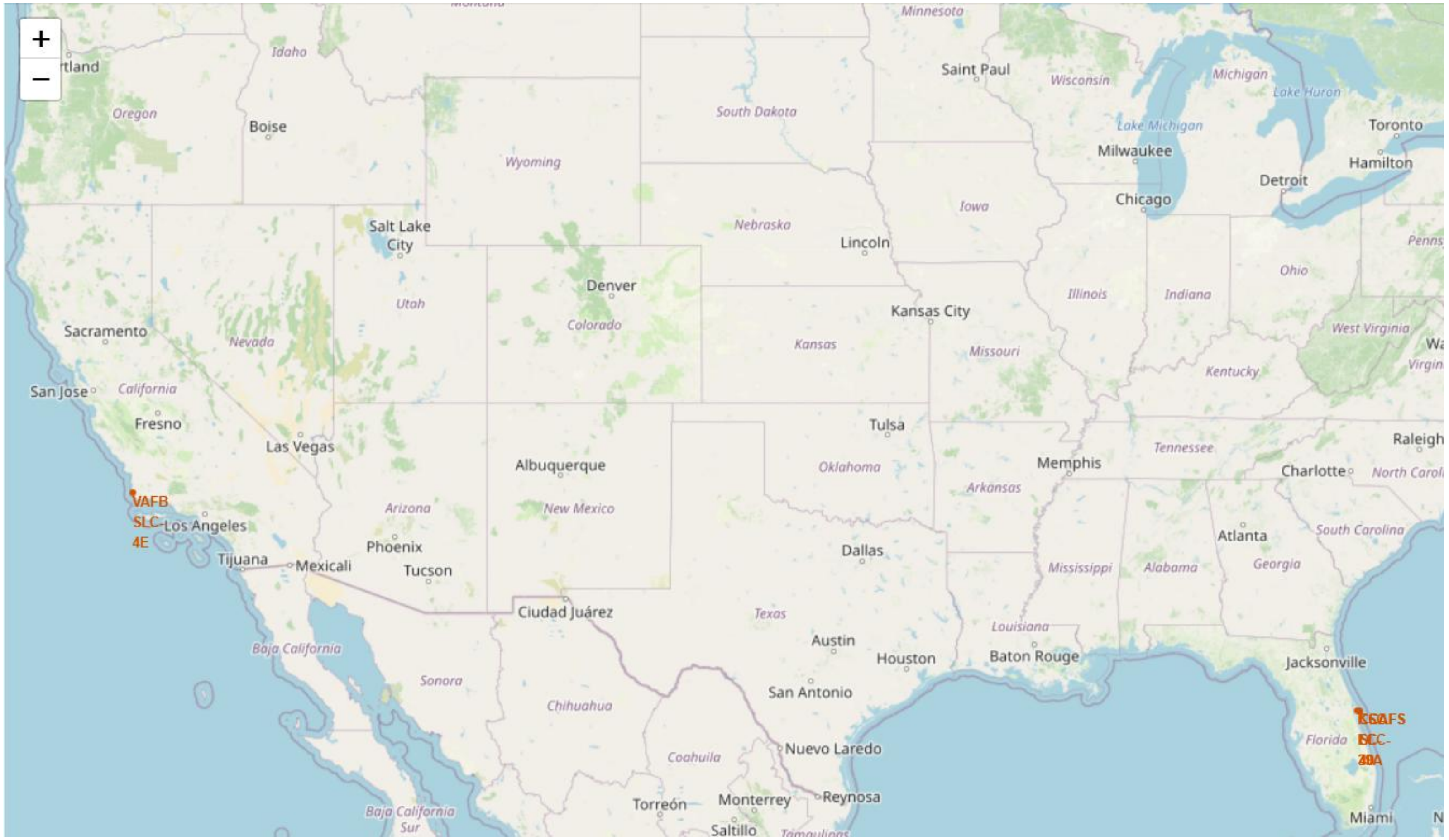
Section 3

Launch Sites Proximities Analysis

RESULTS

3 Folium

- All launch sites on map:

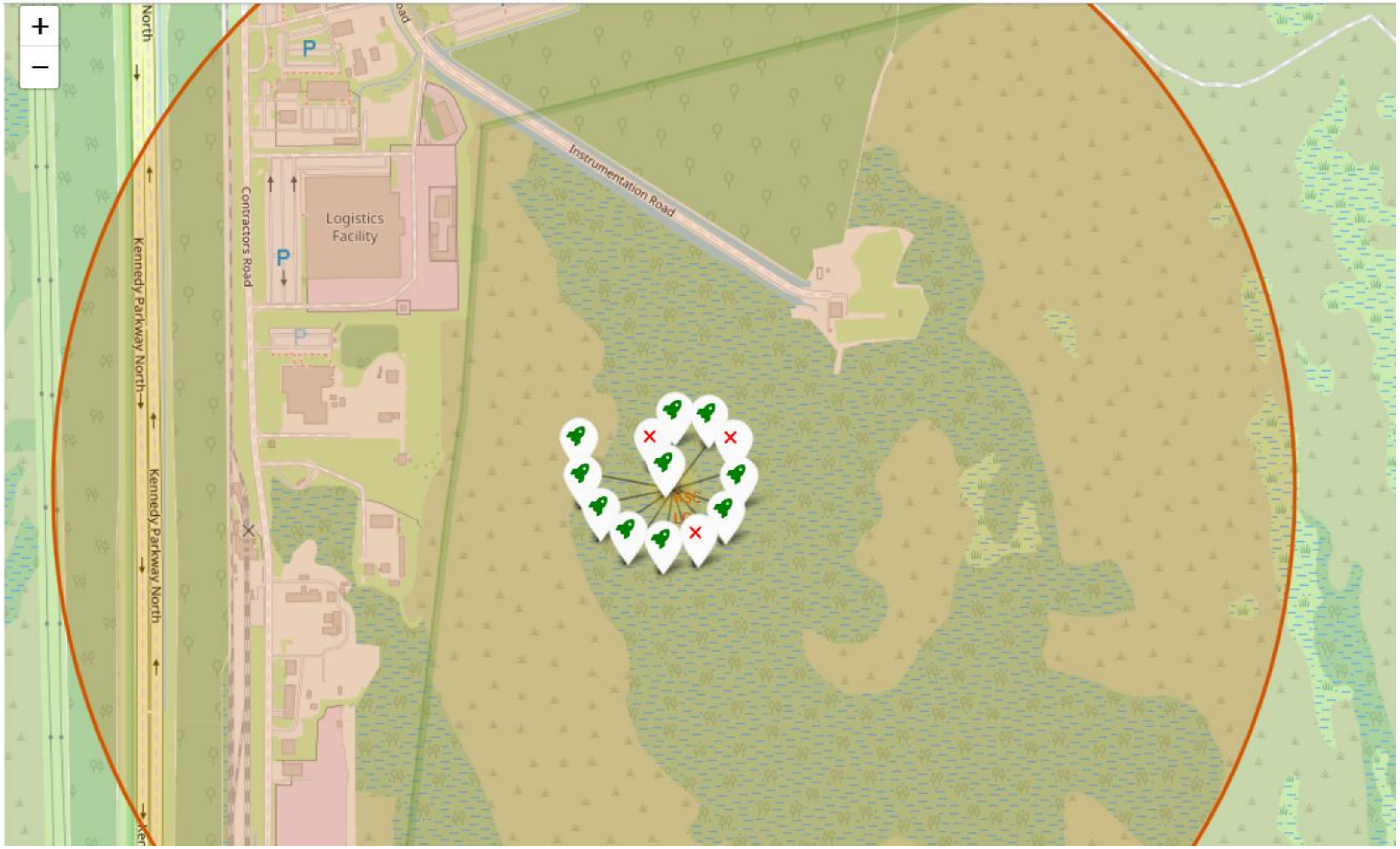


- The folium map image shows the location of the launch sites on a map of the United States.
- The three clusters of labels in orange represent the launch sites being:
  - **VAFB** – Vandenberg Air Force Base (now Vandenberg Space Force Base), a major launch site for polar orbit missions.
  - **CCSFS** – Cape Canaveral Space Force Station.
  - **KSC-** Kennedy Space Center
- The base layer is a Leaflet-style map with light terrain and city names.

RESULTS

3 Folium

- Color-labeled launch outcomes on the map:

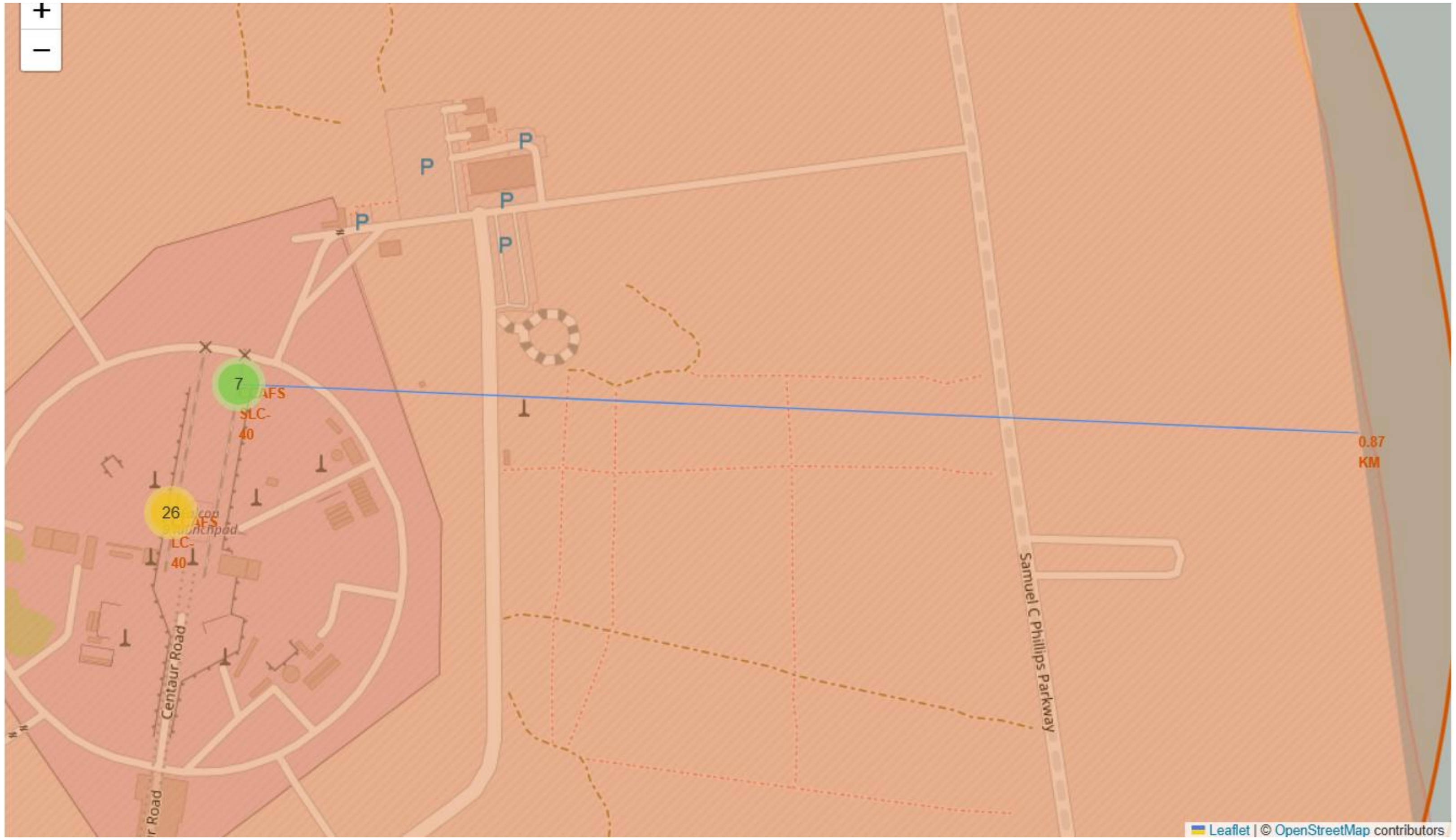


- The folium map image shows the outcome of launches at Kennedy Space Centre.
- The exact launch complex is the KSC SLC-39A.
- It’s famously known been used for Apollo, Shuttle, and now SpaceX launches.
- The color-coded tags indicate launch outcome with the green rocket tags signifying successful launches and red ‘X’ tags signifying failed launches.
- Out of a total 13 launches 10 have been successful with only 3 having failed.
- This represents a 76.92% success rate which is quite high but still can be improved on.

RESULTS

3 Folium

- Distance between a launch site to its proximities (coastline):

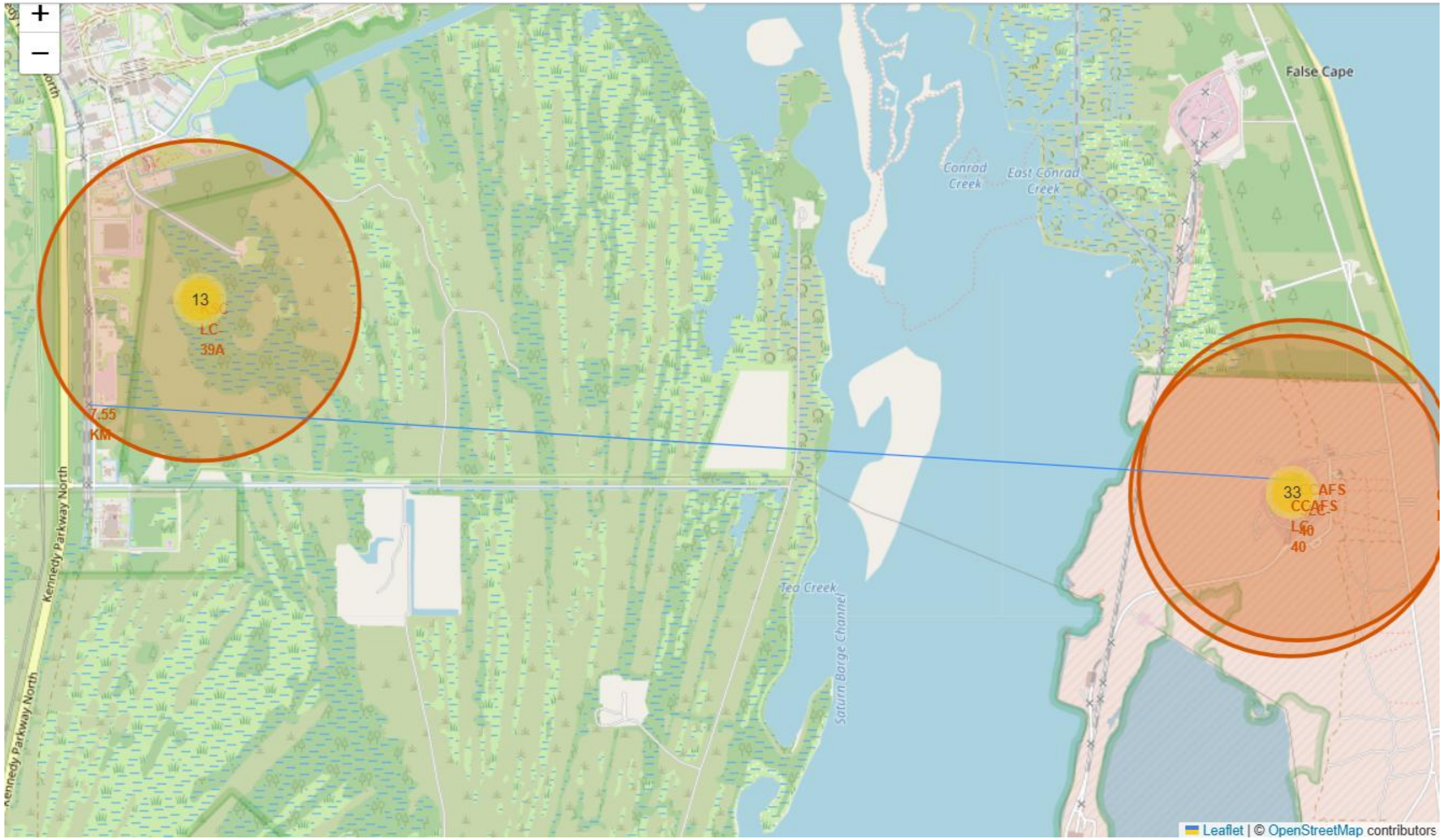


- The folium map image shows the distance between CCAFS SLC-40 (Cape Canaveral) to the Florida coastline which is 0.87 km (870 metres).

RESULTS

3 Folium

- Distance between a launch site to its proximities (railway):

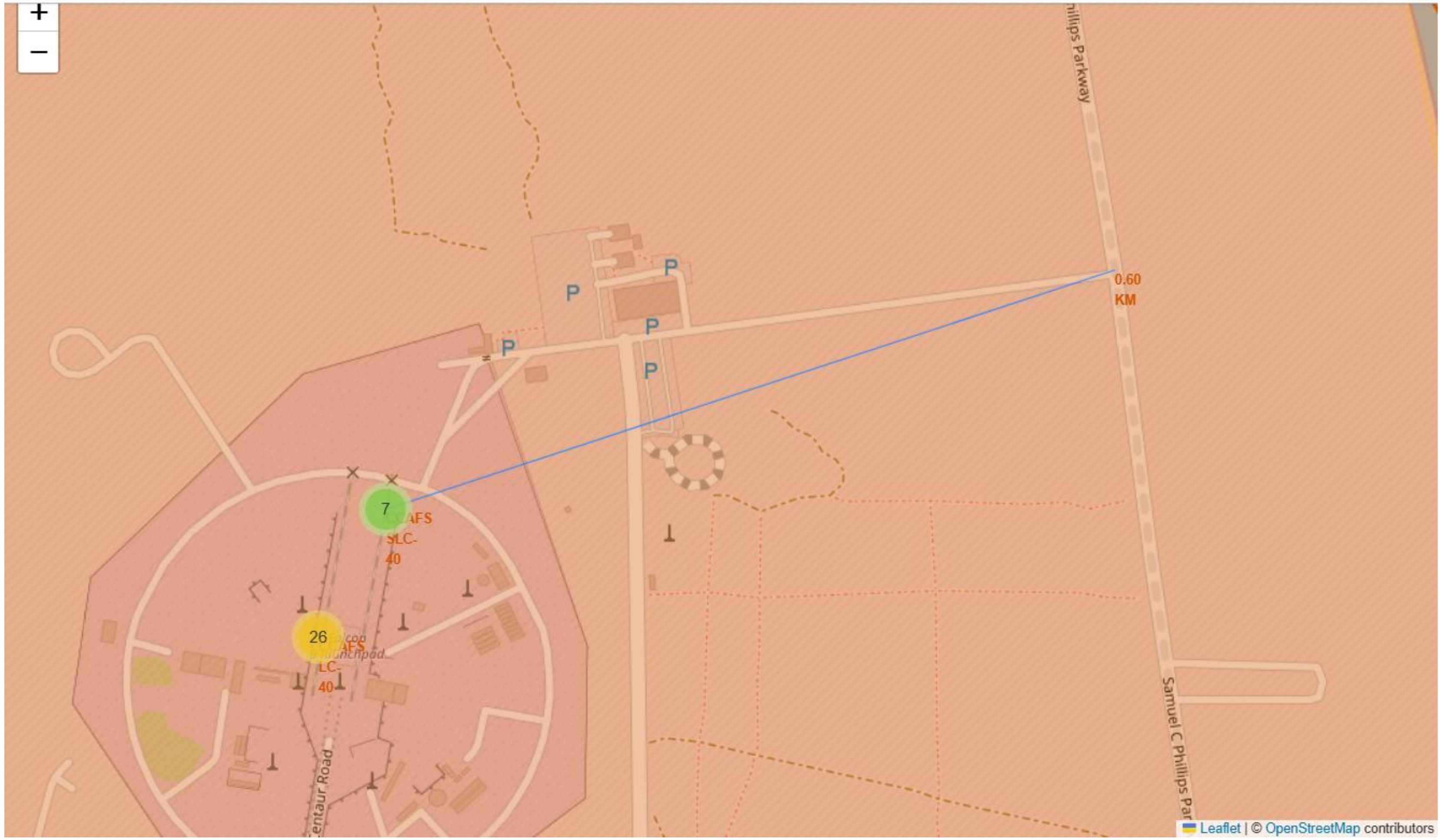


- The folium map image shows the distance between CCAFS SLC-40 (Cape Canaveral) to the nearest railway line (Kennedy Parkway North line) which is 7.55 km
- The distance is significant and could indicate an underlying want for launches to not be too close to civilian infrastructure.

RESULTS

3 Folium

- Distance between a launch site to its proximities (road):

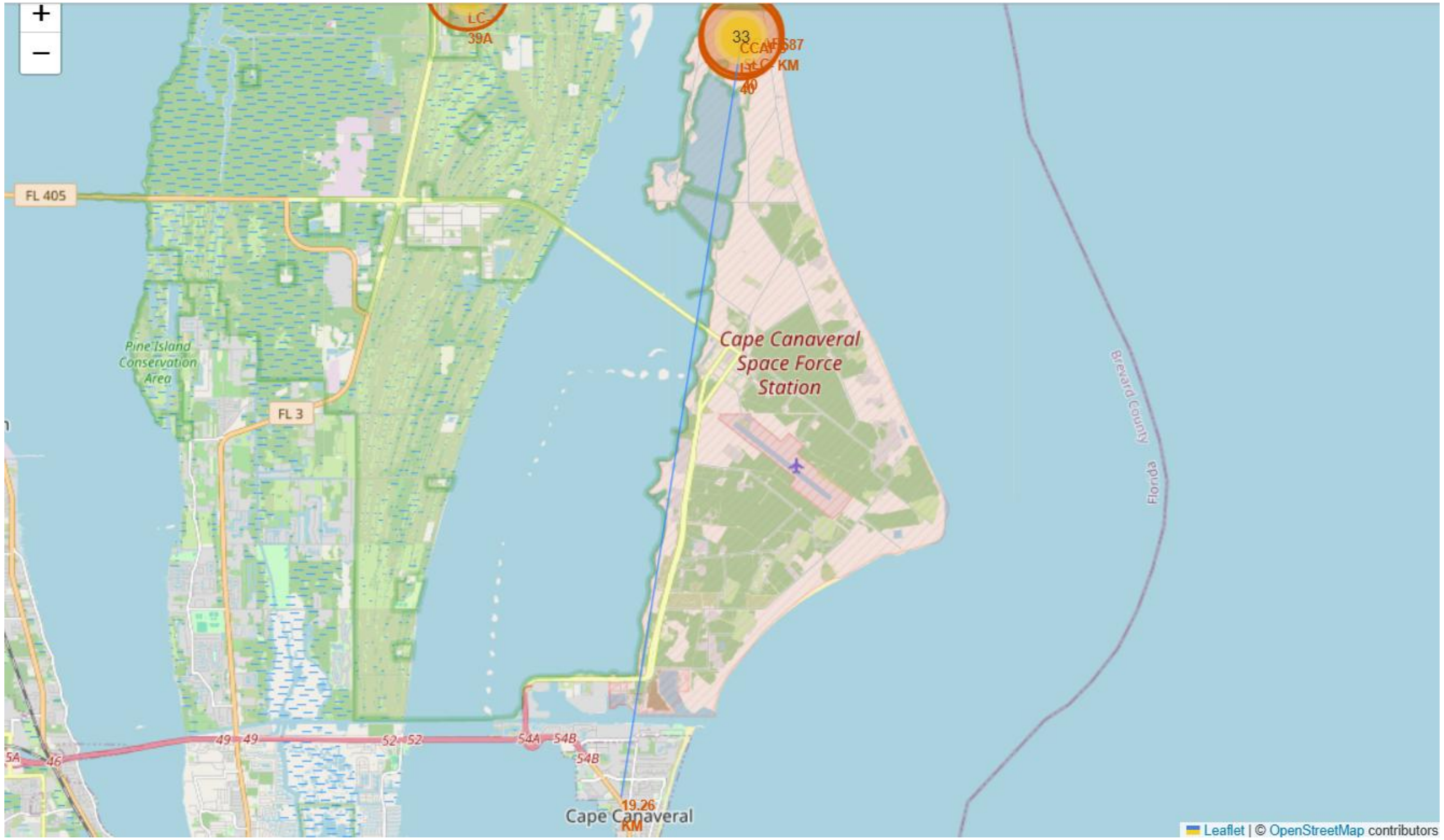


- The folium map image shows the distance between CCAFS SLC-40 (Cape Canaveral) to Samuel C Phillips Parkway road which is 0.60 km (600 metres).
- This road may be a service/access road used by the air base hence it doesn't receive civilian traffic and is restricted from the general public.

RESULTS

3 Folium

- Distance between a launch site to its proximities (city):



- The folium map image shows the distance between CCAFS SLC-40 (Cape Canaveral) to Cape Canaveral city. The distance is 19.26 km
- The sizeable distance between the launch site and the city may be due to the hazardous nature of rocket launches hence it's best to be as far as possible from a city so as not to increase the risk of danger or disruption to civilian life in the city.

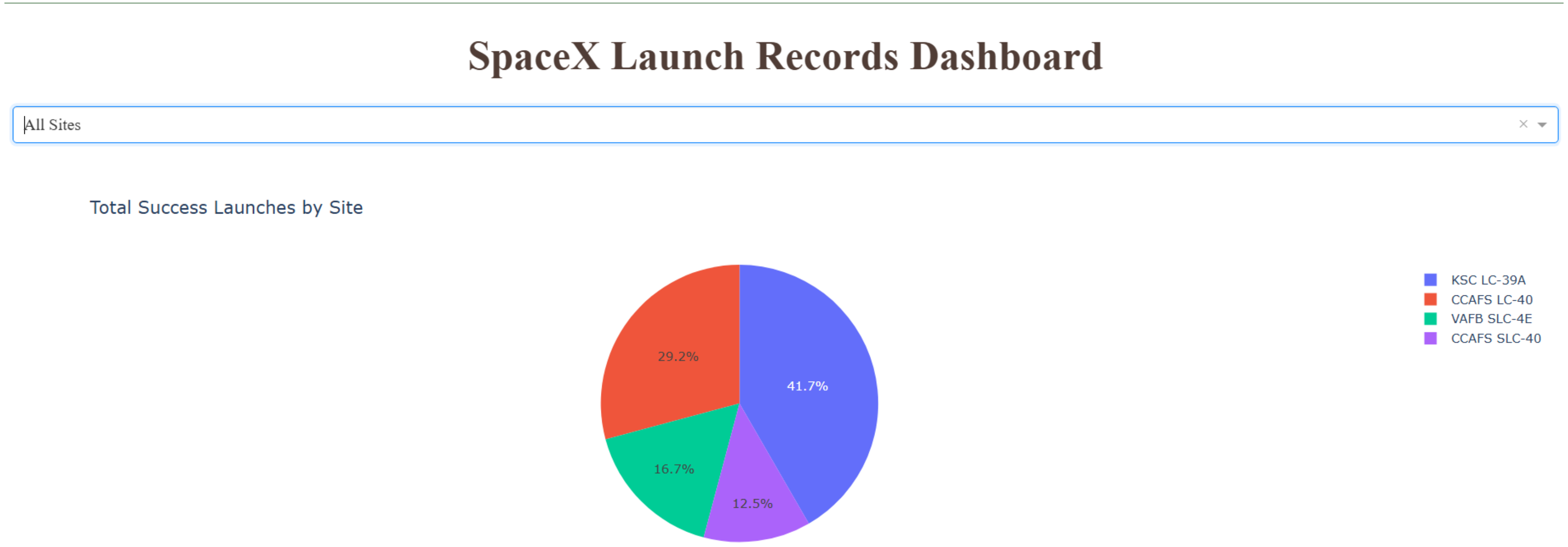
Section 4

Build a Dashboard with Plotly Dash

RESULTS

4 Dash

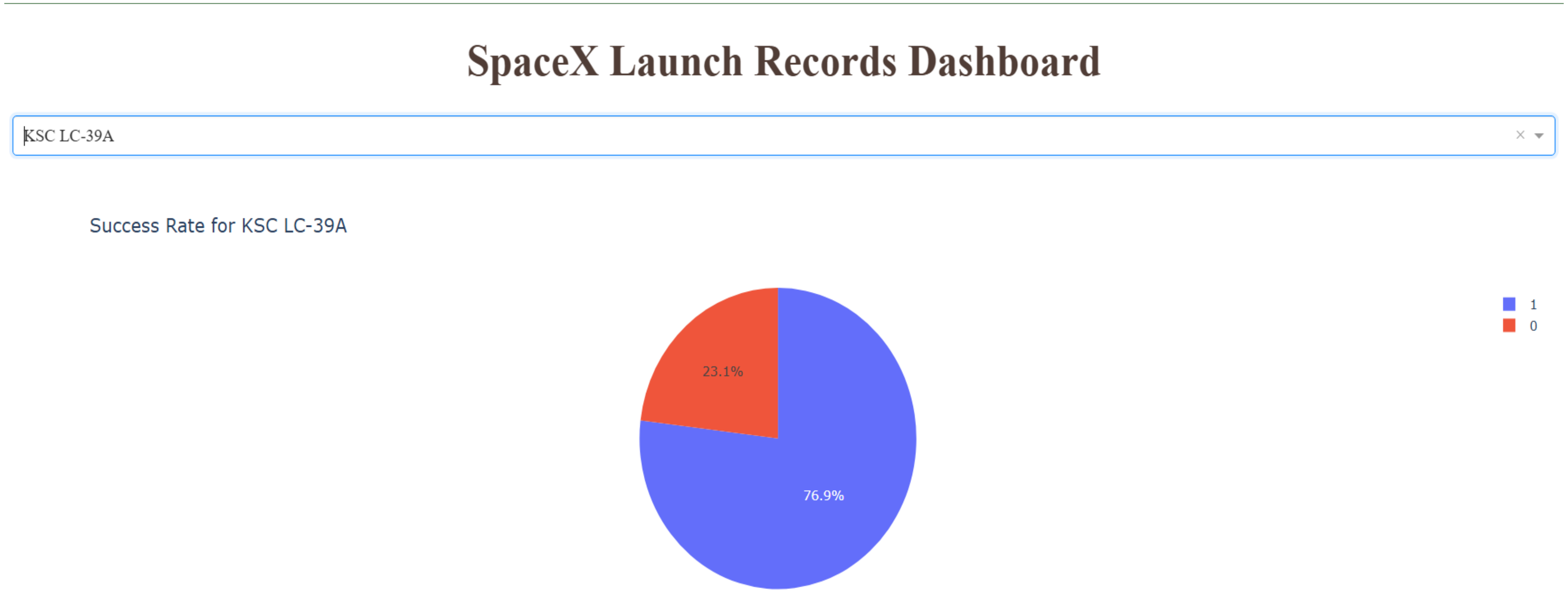
- The image below shows a pie chart displaying total success launches by site.
- It covers all launches across all sites. It focuses on successful launches (mission outcome).
- Percentages are relative to total successes, not absolute launch counts. KSC LC-39A has 41.7% of successful launches (highest share).
- Discrepancy between CCAFS sites suggests SLC-40 was less utilized than LC-40.



RESULTS

4 Dash

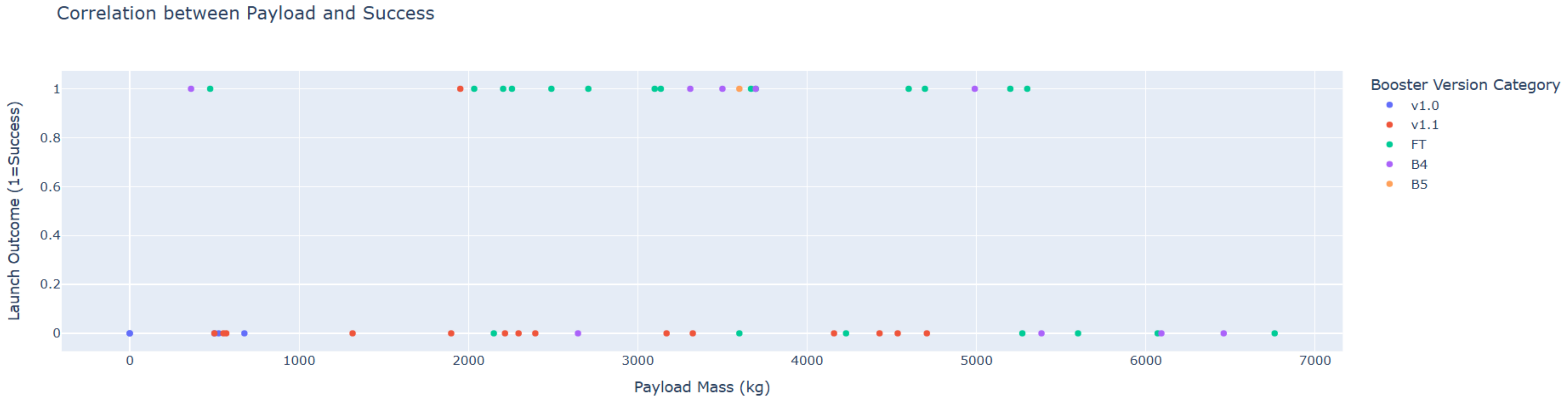
- The image below analyzes launch outcomes specifically for Kennedy Space Center's LC-39A, SpaceX's primary pad for high-profile missions. 0 signifies failure while 1 signifies success.
- The vast majority of launches from this pad succeeded with a 76.9% success rate. A 23.1% failure rate represents a minority of missions that encountered issues.
- From the analysis we can conclude KSC LC-39A maintains the strongest success rate of all sites.



RESULTS

4 Dash

- The image below shows a scatterplot displaying payload vs. success rate correlation
- The plot compares mission success rate by various booster versions across a wide range of payload masses.
- Lower Payloads (1,000–3,000 kg): Likely show highest success rates across all booster versions (less stress on systems).
- Mid-Range (4,000–5,000 kg): Success rates may dip slightly, especially for older boosters (V1.0/V1.1).
- B4/B5 boosters seem to have high success rates across all payload ranges. They also maintain a high success rate for heavy payloads indicating an optimization for said loads. For heavy payloads they are essential.
- There seems to be no failures below 2,000 kg: Light payloads are consistently safe.



Section 5

Predictive Analysis (Classification)

RESULTS

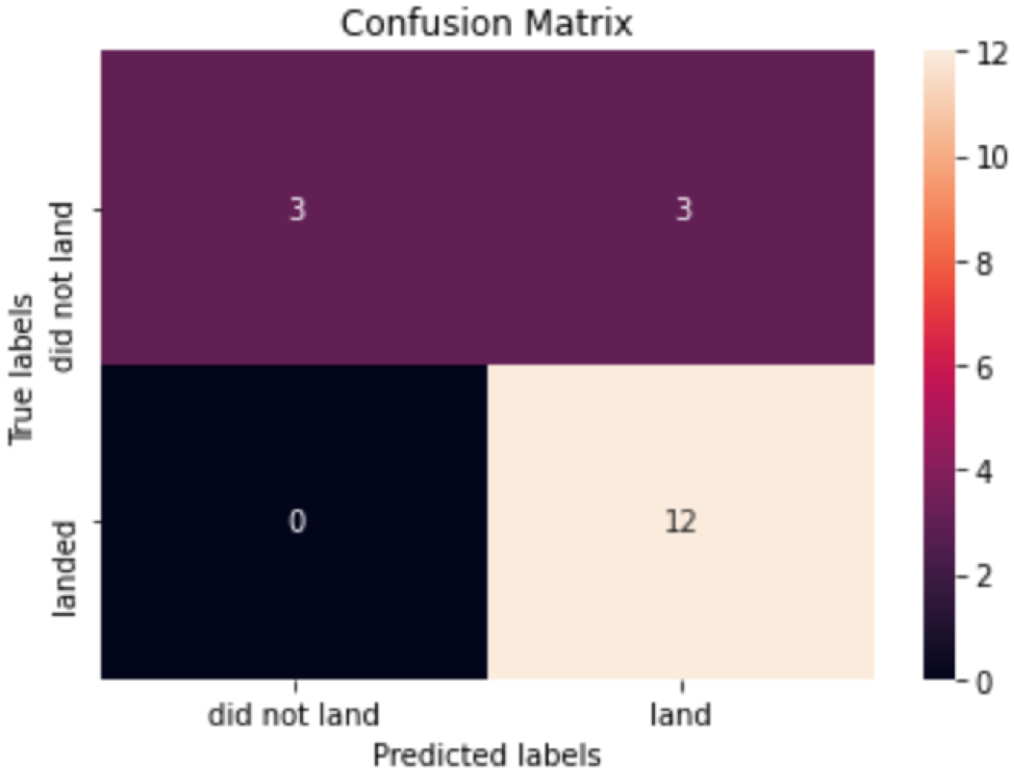
5 Predictive Analysis

- During the modelling stage 4 models were tested:
  - **Logistic Regression**
  - **Support Vector Machines (SVMs)**
  - **Decision Tree Classifier**
  - **K-nearest Neighbours**
- Modelling took place using cross-validation and fold techniques, with varying parameters tested. Their respective cross-validation scores (CV scores), accuracy scores and best parameters were recorded in order to determine the best model.
- Across the 4 models their accuracy scores were exactly identical when the models were tested on the test set and this was further backed by their confusion matrices whereby upon plotting, the number of recorded true positives, true negatives, false positives and false negatives were equal across aboard.
- Given this conundrum the GridSearch CV scores were the next best metric used to determine the best model. K-nearest neighbours and SVM had the same CV score of 0.848214
- The decision tree classifier had the best CV score with 0.9054 hence was determined to be the best model for making predictions.
- A bar chart was also plotted to compare the accuracy and CV scores of all 4 models in order to get a better grasp of the marginal difference between the models' scores.
- The respective confusion matrices for each model are in the accompanying slides along with the bar chart.

RESULTS

5 Predictive Analysis

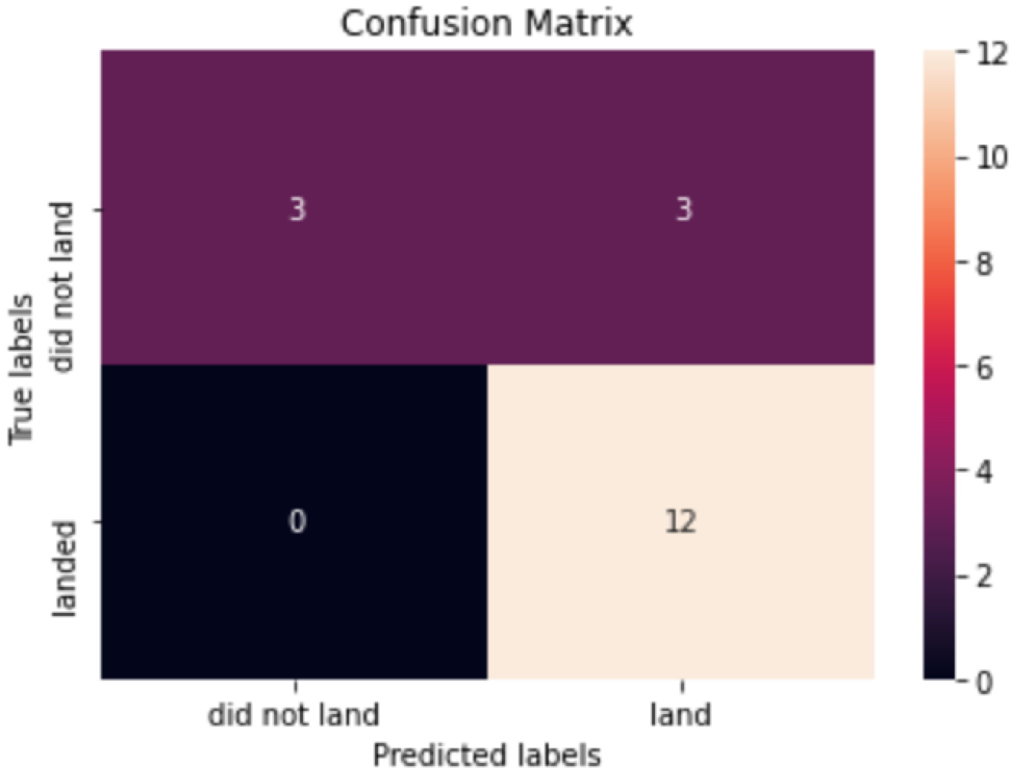
- <u>Logistic regression</u>
  - GridSearch CV best score: 0.846429
  - Accuracy score on test set: 0.833333
  - Confusion matrix:



RESULTS

5 Predictive Analysis

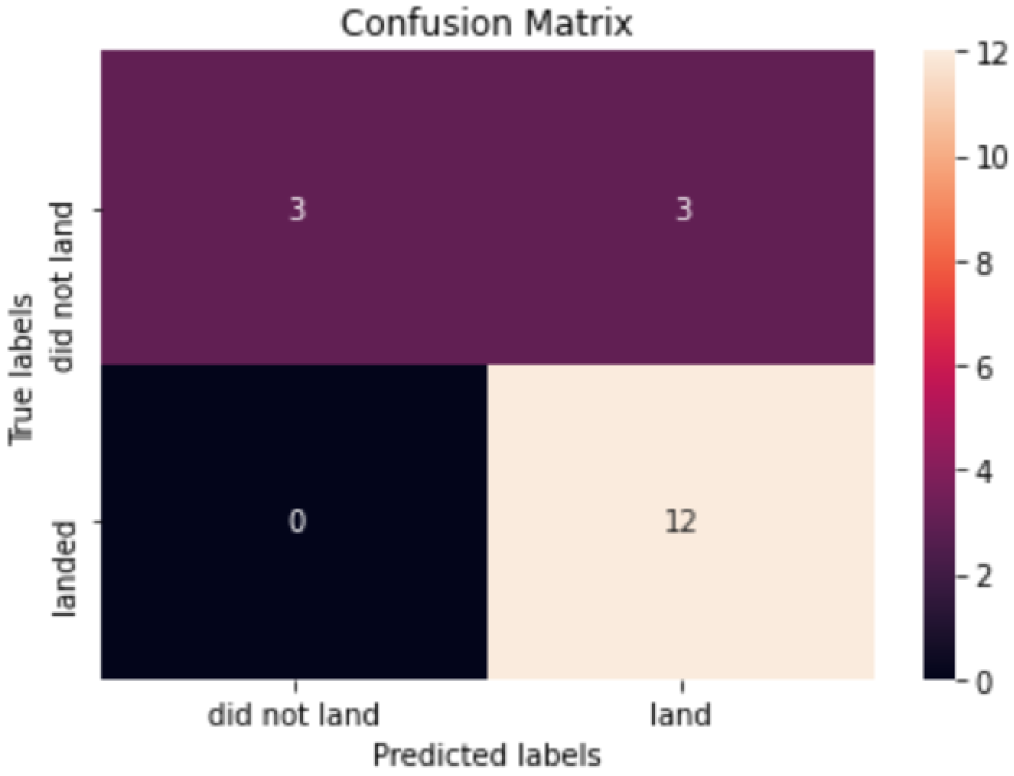
- Support Vector Machine (SVM)
  - GridSearch CV best score: 0.848214
  - Accuracy score on test set: 0.8333333
  - Confusion matrix:



RESULTS

5 Predictive Analysis

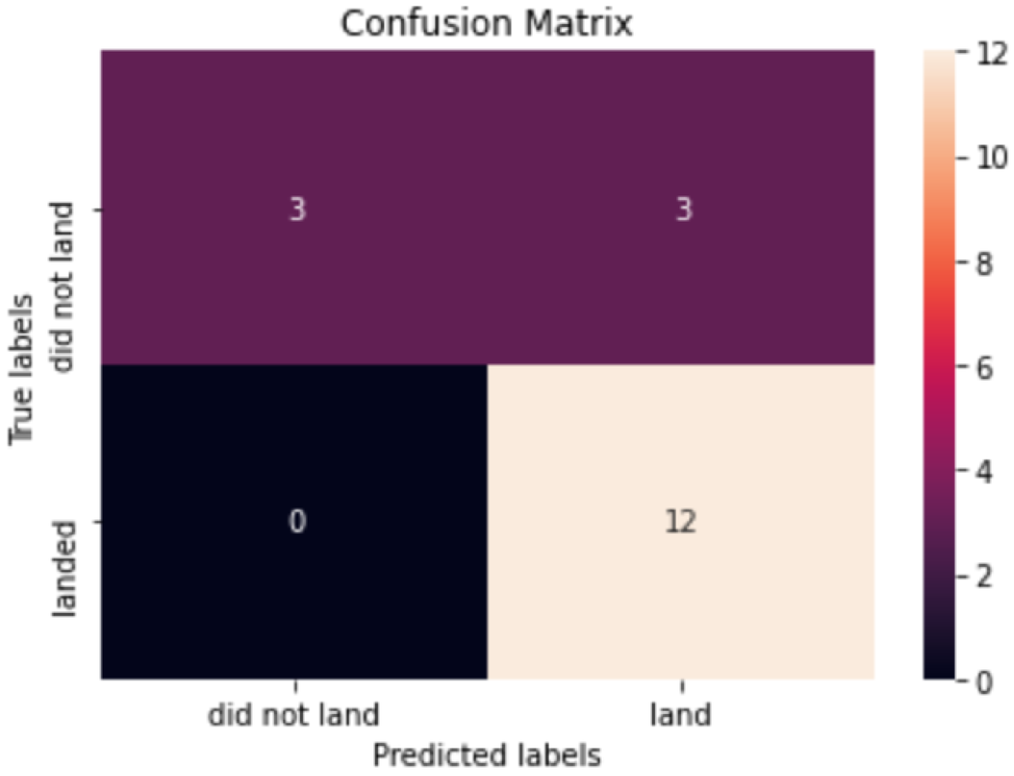
- <u>Decision Tree Classifier</u>
  - GridSearch CV best score: 0.875000
  - Accuracy score on test set: 0.833333
  - Confusion matrix:



RESULTS

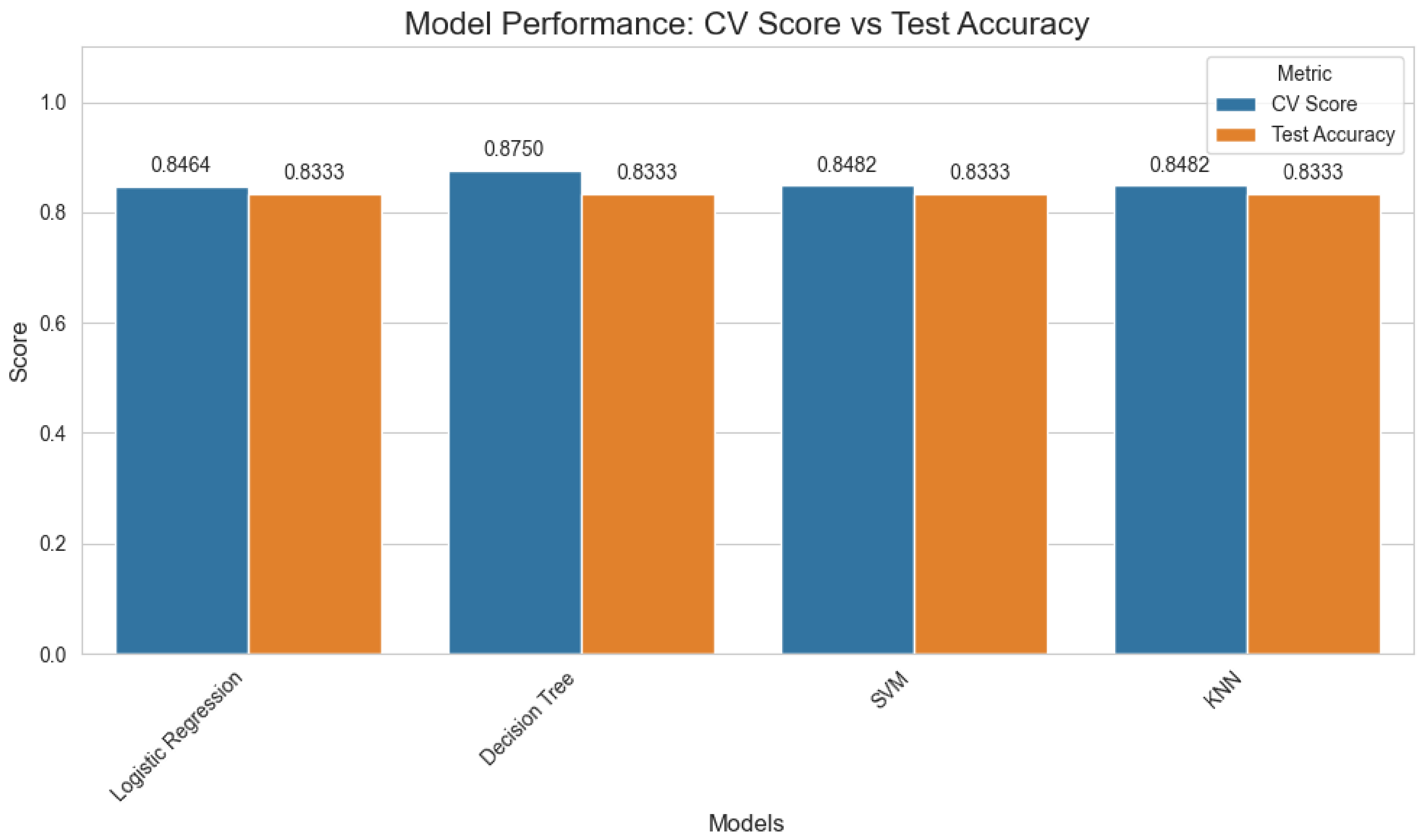
5 Predictive Analysis

- K-Nearest Neighbors
  - GridSearch CV best score: 0.848214
  - Accuracy score on test set: 0.833333
  - Confusion matrix:



RESULTS

5 Classification Accuracy



DISCUSSION

- From the overall analysis and results we can see that some features may have a correlation with the mission outcome in several ways. Some of the features that standout include booster versions and payload mass.
- We've also gained key insights into launch sites that have the highest success rate with KSC LC-39A being the standout.
- Block 5 boosters have been shown to demonstrate the highest reliability especially for heavier payloads (6,000 -7,000 kg).
- Lighter payloads (<4,000 kg) have shown to have near-perfect success rates across all booster versions.
- Drone ship landings improved from 50% success (2015) to near-perfect today.
- First successful ground pad landing occurred on **December 22, 2015**, marking a reusability milestone.
- Overall success rate exceeded **98%** in recent years, with failures concentrated in early missions or complex orbits.
- All in all, SpaceX's iterative design improvements and site optimizations have revolutionized launch reliability, enabling cost-effective access to space across diverse mission profiles.
- This is best evidenced in the launch yearly trend graph which shows improvement by the year.

CONCLUSION

- The goal of this project, was to try to develop a model that could predict if the first stage of a given Falcon 9 launch will land or not in order to determine the cost of a launch.
- The data was retrieved and complied through multiple API calls. It underwent cleaning and wrangling to prepare it for exploratory data analysis (EDA).
- The data was visualized through various graphs and plots and queried through a database so as to gain keen insights.
- It was established that different variables and how they interacted with one another greatly affected the outcome of a launch. Each feature of a Falcon 9 launch, such as its payload mass or orbit type, was found out to affect the mission outcome in a certain way.
- As per the insight gathered from the data, KSC LC-39A is the most optimal launch site for use.
- The data underwent feature engineering and preprocessing in order to prepare it to be fed into the machine learning algorithms.
- This was project was found to be a binary classification problem.
- Several machine learning algorithms were employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by the Decision Tree algorithm performed the best among the 4 machine learning algorithms employed.

Thank you!