IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adetola Adedoyin
09/12/2024

O utlin e

- Executive Summary
- Introduction
- Methodology
- Results
- Con c lus ion
- Ap pe n dix

Executive Summary

- General Overview
  - The project aims to predict the launch costs and reusability of SpaceX's Falcon 9 rocket by analyzing data on SpaceX.
  - Goals include determining launch prices, developing a machine learning model to predict the reuse of the rocket's first stage, and presenting key findings.

- Summary of methodologies
  - Data Collection involves gathering and preparing SpaceX launch data from the SpaceX REST API, focusing on filtering, cleaning, and enriching the dataset for predictive modeling. This includes transforming the data into a structured format, addressing missing values, and converting the landing outcome into a binary classification for analysis.
  - Data Wrangling reviews key attributes like Flight Number, Date, Booster version, Payload Mass, Orbit, Launch Site, and Outcome. The Outcome is crucial as it is transformed into a binary variable, with 0 for unsuccessful landings and 1 for successful ones, aiding in predictive analysis.

Executive Summary

- Summary of methodologies (continue):

  The text outlines a comprehensive analysis of Falcon 9's first-stage landings, which includes:

  - Exploratory Data Analysis: Identifying key attributes (e.g., launch sites, payload mass) that predict landing success using machine learning and one-hot encoding for categorical variables.
  - Interactive Visual Analytics: Utilizing Folium for launch site geolocation analysis and creating a Plotly Dash dashboard for real-time data exploration through interactive components.
  - Predictive Analysis: Developing a machine learning pipeline for landing success prediction, involving data standardization, train-test split, and evaluating multiple algorithms (Logistic Regression, SVM, Decision Tree, KNN) with Grid Search for hyperparameter optimization. Model performance is assessed with a confusion matrix to find the most accurate model.

Executive Summary

- Summary of all results
  - As flight numbers increase, first-stage landing success rates rise, with VAFB SLC-4E and KSC LC-39A leading in success. No heavy payloads launched from VAFB SLC, while CCAFS LC-40 has mixed results. Success is highest in ES-L1, GEO, HEO, and SSO orbits, with LEO success linked to flight numbers, but no correlation in GTO orbit. Heavy payloads improve success in Polar, LEO, and ISS orbits, whereas GTO shows varied outcomes. Success rates peaked until 2014, stabilized from 2014-2015, then showed slight declines in 2018 and 2020. The first successful ground landing was on 22-12-2015. Launch sites are near the equator for efficiency and away from cities for safety. KSC LC-39A has the highest success rate (41.7%) and most successful launches (76.9%). B5 and FT boosters excel with payloads over 4000 kg. Logistic Regression proved to be the most accurate model, with 12 true positives and 3 false positives.

In troduction

- Project Overview

- Space Industry Overview: Companies such as Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are transforming the landscape of space travel.
- Achievements of SpaceX: SpaceX has reached significant milestones, including sending spacecraft to the International Space Station and launching Starlink.
- Cost-Effectiveness: The Falcon 9 launch costs are considerably lower at $62 million due to its reusability, in contrast to competitors that charge around $165 million.
- Falcon 9 Details: The Falcon 9 features a two-stage design, with the first stage responsible for the majority of the launch and designed for reuse.
- Recovery: SpaceX’s approach to reusability lowers costs, distinguishing it from other launch service providers.

- Project Focus

- Objective: Determine the likelihood of a successful landing of SpaceX's Falcon 9 first stage, which impacts launch expenses.
- Mission Aim: Develop a model using data to forecast first-stage landings, aiding in the estimation of launch costs.

Section 1

Methodology

M et ho do lo gy

Executive Summary

- Data collection methodology:
  - Collecting and preparing SpaceX launch data from the SpaceX REST API in addition with web scraping techniques, focusing on filtering, cleaning, and enriching the dataset for predictive modeling. Transformation of the data into a structured format, handling missing values, and converting the landing outcome into a binary classification variable for analysis.
- Perform datawrangling
  - Reviewing key attributes such as Flight Number, Date, Booster version, Payload Mass, Orbit, Launch Site, and Outcome. The Outcome attribute is particularly important as it will be transformed into a binary classification variable, where 0 indicates an unsuccessful landing and 1 indicates a successful landing, facilitating predictive analysis.
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
  - Developing a machine learning pipeline to predict the success of Falcon 9's first stage landing, incorporating preprocessing steps for data standardization and the application of the train-test split technique to partition the dataset into training and testing subsets. Evaluation of several algorithms, including Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-Nearest Neighbors, with Grid Search employed to optimize hyperparameters, followed by model performance assessment using a confusion matrix to determine the model with the highest accuracy.

Data Collection

- API Usage: Data was collected from the SpaceX REST API, which provides detailed information about past launches, including rocket specifications, payloads, launch and landing details, and the outcome of landings.
- Data Retrieval: A GET request was made using Python’s requests library to fetch the data from the API. The response from the API was in the form of a JSON object, which contained a list of past launch records.
- Data Conversion: The JSON response was converted into a flat table using the json_normalize() function. This function helps to "normalize" the nested JSON structure into a Pandas DataFrame, making the data more accessible for analysis.
- Data Filtering: The dataset contained information for both Falcon 1 and Falcon 9 boosters. The data for Falcon 1 launches was filtered out to focus solely on Falcon 9 launches.
- Handling Missing Data: The dataset contained NULL values, particularly in the "PayloadMass" column. To deal with this, the mean value of "PayloadMass" was calculated and used to replace the NULL values. The "LandingPad" column, which had NULL values representing when no landing pad was used, was left unchanged for later handling with one-hot encoding.
- Data Enrichment: To supplement missing or incomplete data (e.g., rocket IDs), additional API requests were made targeting specific endpoints (such as /booster, /launchpad, /payload, and /core) to gather detailed information for each ID number, which was then used to enrich the dataset.

Data Collection – SpaceX API

- API Usage: Data was collected from the SpaceX REST API, which provides detailed information about past launches, including rocket specifications, payloads, launch and landing details, and the outcome of landings.
- Data Retrieval: A GET request was made using Python’s requests library to fetch the data from the API. The response from the API was in the form of a JSON object, which contained a list of past launch records.
- Data Conversion: The JSON response was converted into a flat table using the json_normalize() function. This function helps to "normalize" the nested JSON structure into a Pandas DataFrame, making the data more accessible for analysis.

•GitHub - Data Collection - API

Data Wrangling

- Data Wrangling Process:
- Attribute Identification: Key attributes included Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Serial, Grid Fins, Legs, Landing Pad, Reused Count, Longitude, and Latitude.
- Launch Site Analysis: Launch sites such as Vandenberg AFB, Kennedy Space Center, and CCAFS SLC-40 were categorized to analyze performance and outcomes.
- Orbit Classification: Orbits were classified:
  - LEO: Low Earth Orbit (up to 2,000 km altitude).
  - GTO: Geosynchronous Transfer Orbit (35,786 km altitude).
- Outcome Standardization:
  - Outcomes were binary:
    - 1: Successful landing (e.g., True ASDS on a drone ship).
    - 0: Unsuccessful landing (e.g., False ASDS).
  - A classification variable Y was created for analysis.
- Model Preparation: Data was structured for predictive modeling by standardizing outcomes into a binary format.
- GitHub - Data Wrangling

EDA with Data Visualization

- Exploratory Data Analysis (EDA): EDA was conducted to understand the structure and patterns within the Falcon 9 launch data, serving as a foundational step for further predictive modeling.
- Attribute Analysis: Key attributes, such as launch number, launch site, and payload mass, were evaluated to determine their correlation with the success of the first-stage landing.
- Success Rate Trends:
  - Historical data revealed that Falcon 9's landing success rate has improved since 2013, with distinct differences in success rates across launch sites.
  - For instance, CCAFS LC-40 has a 60% success rate, while KSC LC-39A and VAFB SLC 4E exhibit approximately 77% success rates. Combining features revealed additional insights, such as a 100% success rate at CCAFS LC-40 for payloads exceeding 10,000 kg.
- Feature Engineering:
  - Relevant attributes were combined and analyzed to uncover patterns that could predict successful landings.
  - Categorical variables, such as launch site, were converted into numerical formats using one-hot encoding to make them suitable for machine learning algorithms.
- GitHub –EDA (Data Visualization)

EDA withSQL

- SQL queries performed:
  - Display the names of unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string 'CCA'.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- GitHub – EDA (SQL)

Build an Interactive Map with Folium

- Folium: Used for geospatial analysis, marking launch sites, and analyzing proximities on interactive maps to identify patterns and optimal launch site locations.
- Provides real-time data exploration with zoom, pan, filter, search, and linking.
- Objectives:
    - Mark launch sites (clusters) on the map
    - Mark the successful (green)/ Unsuccessful (red) launches for each site
    - Calculate the proximity of a launch site to railway/ highway/ coastline/ cities

- GitHub - Folium

Build a Dashboard with Plotly Dash

- Purpose: Build an interactive dashboard for deeper insights into the SpaceX dataset, improving data analysis.
- Plotly Dash: Creates a dashboard with interactive elements (e.g., dropdown lists, range sliders) to manipulate pie and scatter charts.
- Outcome:
  - Plots:
    - Total Successful Lauches by Site: Pie Chart that shows the success rate based on the Launch Site
    - Payload vs Success for All Sites Scatter Plot: Binary Scatter Plot with a slider for payload mass that shows the Booster Version Category success considering the payload mass.
    - The plots assist to determine which Launch Sites are the most successful. In addition, they assist determining the most suitable booster based on the required payload.

- GitHub - Dash

Predictive Analysis (Classification)

- Objective: Develop a machine learning pipeline to predict successful landings of Falcon 9's first stage.
- Preprocessing: Standardize data for uniformity, binary values for successful/ unsuccessful launches.
- Train-Test Split: Divide data into training (80% of the dataset) and testing (20% of the) subsets for model validation.
- Model Training: Test multiple algorithms - Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors.
- Hyperparameter Tuning: Use Grid Search to identify the optimal hyperparameters for model performance.
- Evaluation: Assess model accuracy and output a confusion matrix for performance validation.
- Outcome: Identify the best predictive model with the highest accuracy for Falcon 9's first-stage landing success.

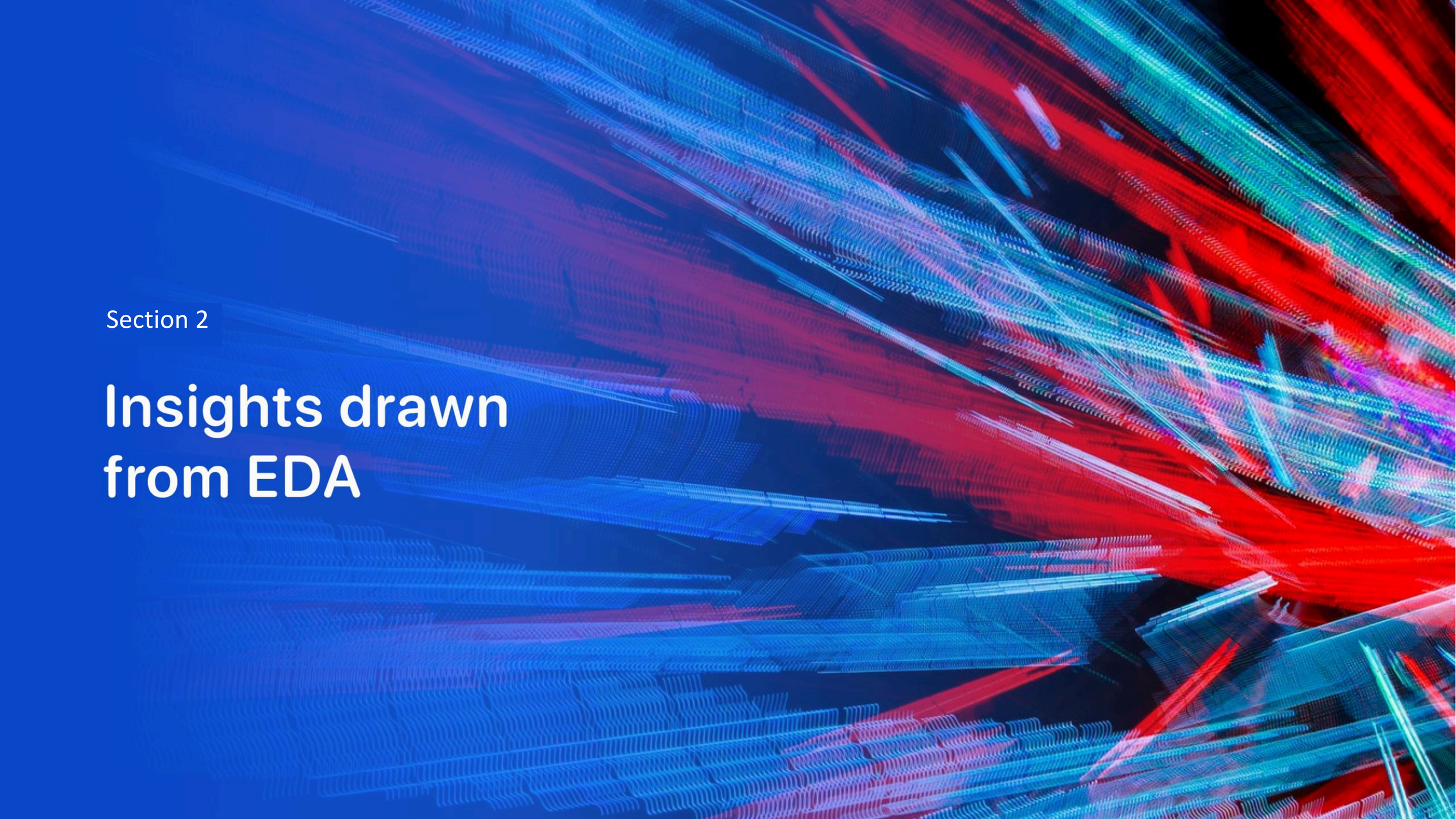
- GitHub – Predictive Analysis

Results

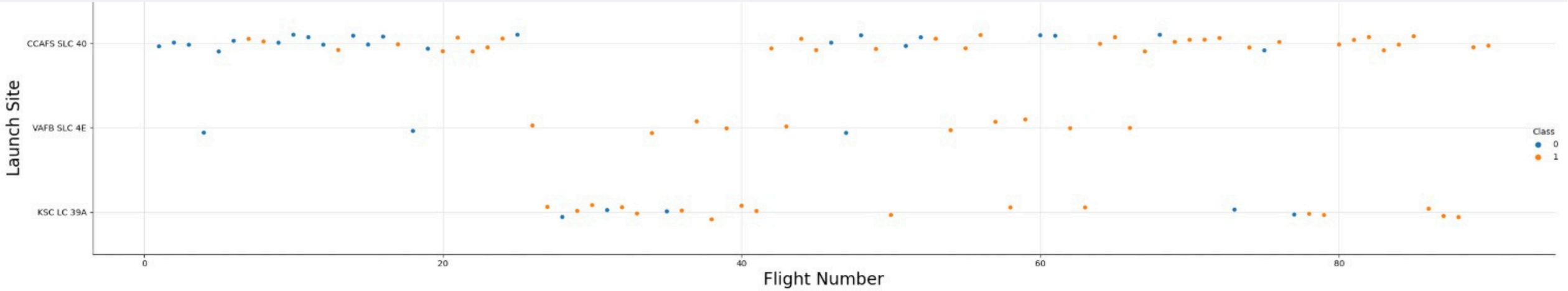
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



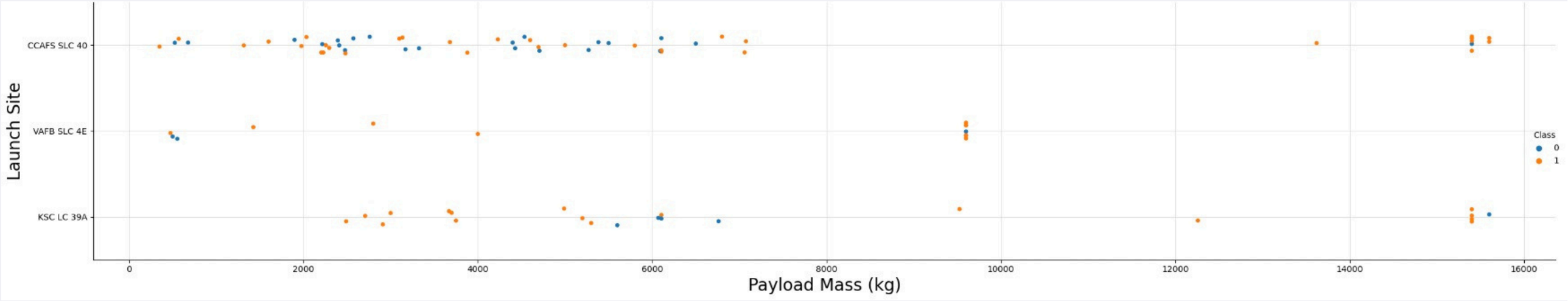
Flight Number vs. LaunchSite



Observations:

- As the flight number increases, the first stage is more likely to land successfully (orange dots).
- Most launches from VAFB SLC-4E and KSC LC-39A land successfully.

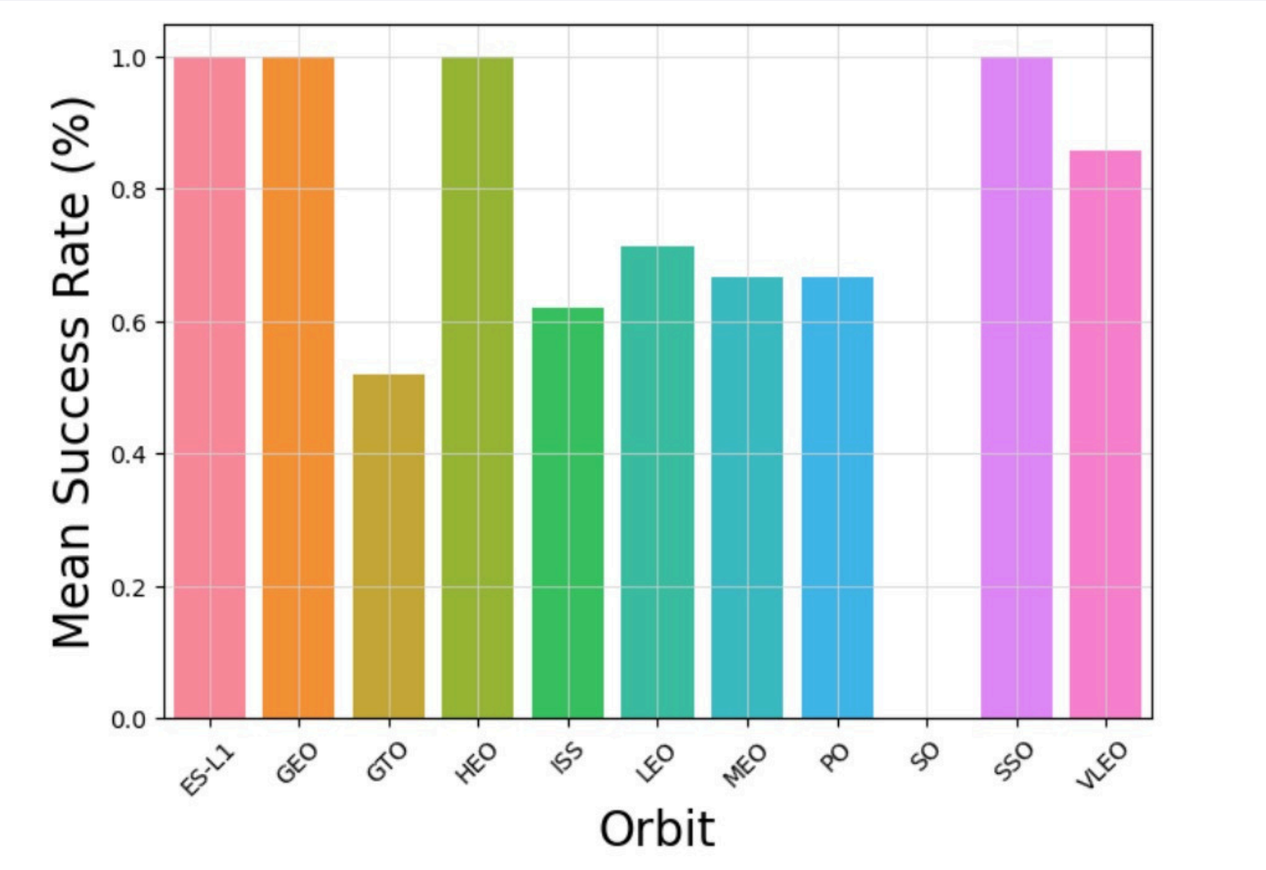
Payload vs. Launch Site



Observations:

- No heavy payload mass (greater than 10 000 kg) rockets were launched from the VAFB-SLC launch site.
- CCAFS LC-40 launch site has a mixed success in all range of payload mass rockets.

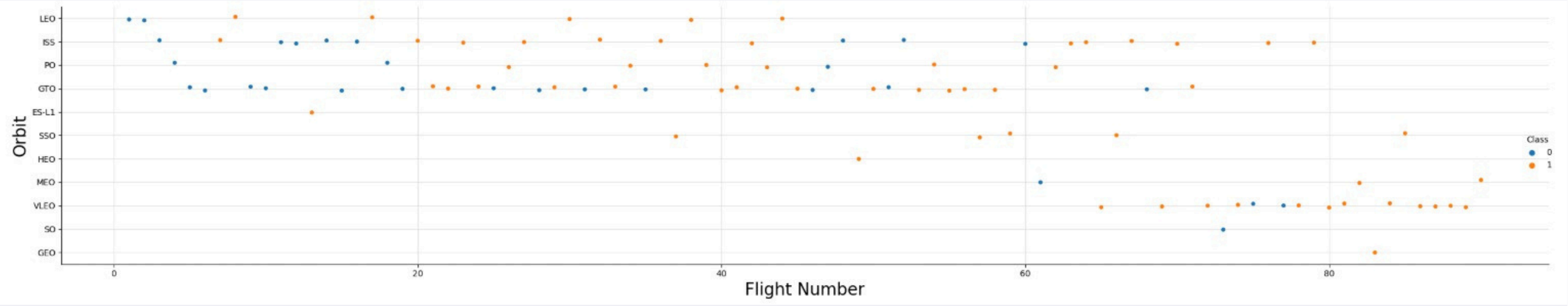
Success Rate vs. Orbit Type



Observations:

- The highest success rate was observed in the ES-L1, GEO, HEO and SSO orbit types.

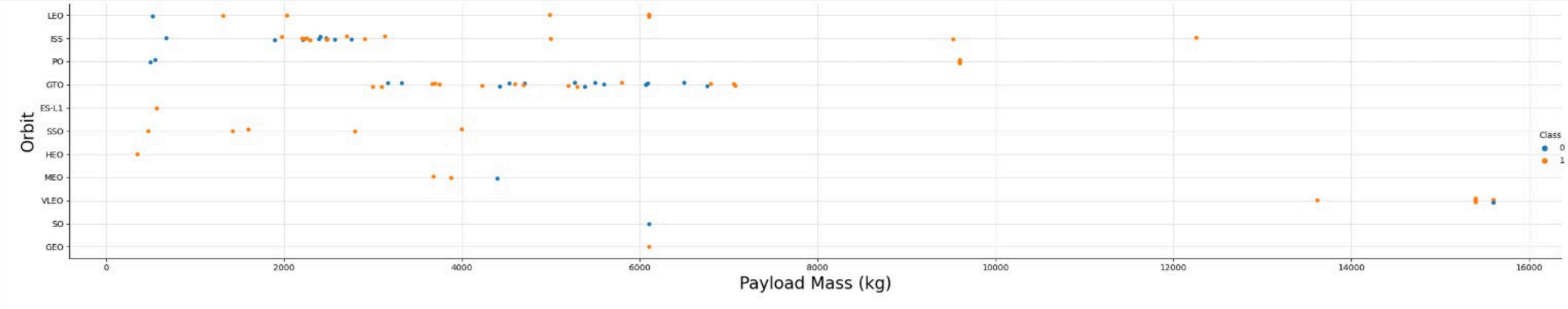
Flight Number vs. Orbit Type



Observations:

- In some Orbit Types (for exampleLow Earth Orbit (LEO)), the success rate appears to correlate with the number of flights.
- In contrast, no significant relationship is observed between the flight number and success rate in some other Orbit Types (for example the Geostationary Transfer Orbit (GTO)).

Payload vs. OrbitType



- Observations:
- For Polar, LEO, and ISS orbits, successful landings or positive landing rates are higher with heavy payloads.
- However, for GTO orbit, the distinction is less clear, as both successful and unsuccessful landing rates are observed.

Launch Success Yearly Trend

- Observations:
- The yearly trend line increases until 2014 and stabilizes from 2014 to 2015. Itincreases again onwards to 2019 with small declines being observed on 2018 and 2020

All Launch Site Names

- The launch sites were retrieved through a query that identifies their unique values in the dataset and are shown below.

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

Launch Site Names Begin with 'CCA'

- Below fiverecords where launch sites begin with `CCA` are presented. The results were retrieved through a query filtering the launch sites based on the characters on the string value ('CCA') and were limited to the first five.

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

- The total payload carried by boosters from NASA (CRS) is presented below. The result was calculated through a query that filters the 'Customer' columnfor the string value of 'CRS'(NASA) and then sums the payload mass.

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is presented below. The result was calculated through a query thatfilters the 'Booster Version' column to the string value of 'F9 v1.1' and then computes the average payload mass of all the filtered records.

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is shown below. The query filters the column of 'Landing Outcome' to the sting value of 'ground pad' and then computes the minimum value of the 'Date' column to show the earliest date where a successful landing on the ground padwas achieved.

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below are listed the names of the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. The query filters all the unique values on the column 'Booster Version' if the payload mass is between 4000 and 6000 kg.

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 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2

Total Number of Successful and FailureMission Outcomes

- The total counts of successful and unsuccessful mission outcomes are summarized below. The query calculates the number of entries in the 'Mission Outcome' column, categorizing them as successful or unsuccessful based on their respective mission status.

```
Successful Mission Outcomes:
 * sqlite:///my_data1.db
Done.
```

COUNT(*)
100

```
Unsuccessful Mission Outcomes:
 * sqlite:///my_data1.db
Done.
```

COUNT(*)
1

Boosters Carried MaximumPayload

- The names of the booster which have carried the maximum payload mass are listed on the right. The query retrieves the 'Booster Version' and 'Payload Mass' for the heaviest payload launched. The inner query finds the maximum payload mass, and the outer query selects the records matching this value.

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

- The failedlanding outcomes in drone ship, their booster versions, andlaunch site names for in year 2015 are listed below.The query extracts data from the dataset for 2015, filtering rows where 'Landing Outcome' indicates a drone ship failure (filtering the string value 'Failure (drone ship)'). It retrieves the mission's launch month, mission outcome, booster version, and launch site, using substring extraction for months while limiting the resultswithin the specified time range (2015).

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

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, are ranked below in descending order. The query retrieves the count of each 'Landing Outcome' from the dataset for the date range between June 4, 2010, and March 20, 2017. It groups the results by 'Landing Outcome' and sorts them in descending order based on the count of each outcome.

Landing_Outcome	COUNT(Landing_Outcome)
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

Section 3

Launch Sites Proximities Analysis

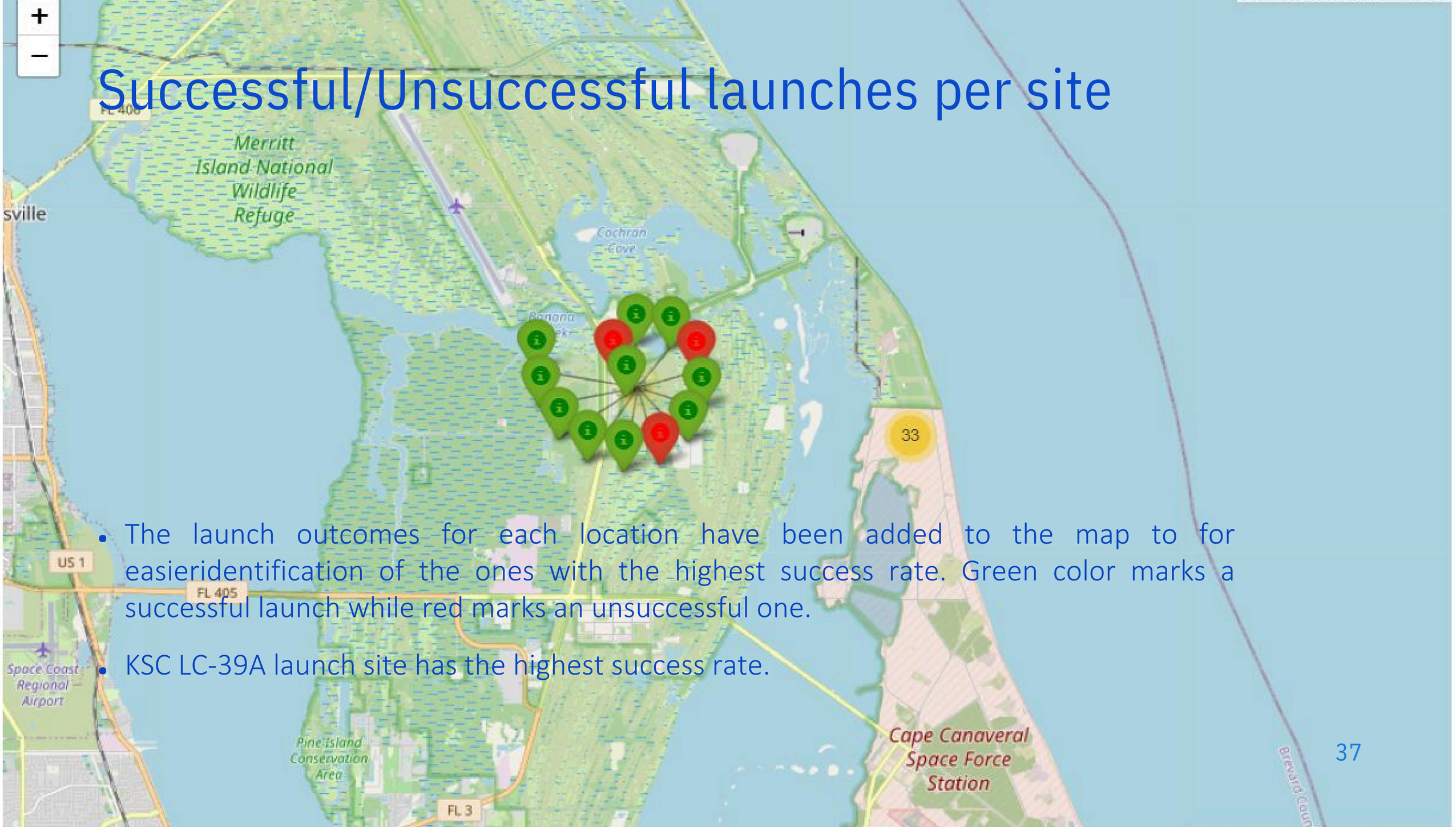
Launch Sites

- The location of the launch sites is shown on this map.
- The launch sites are in close proximity to the Equator line.

Launching from the near-equatorial locations provides a free speed boost from the Earth's rotation, which in turn saves fuel and helps rockets reach orbit more efficiently. With the Earth's rotation aiding their trajectory, rockets can also stay on course with fewer costly course corrections.

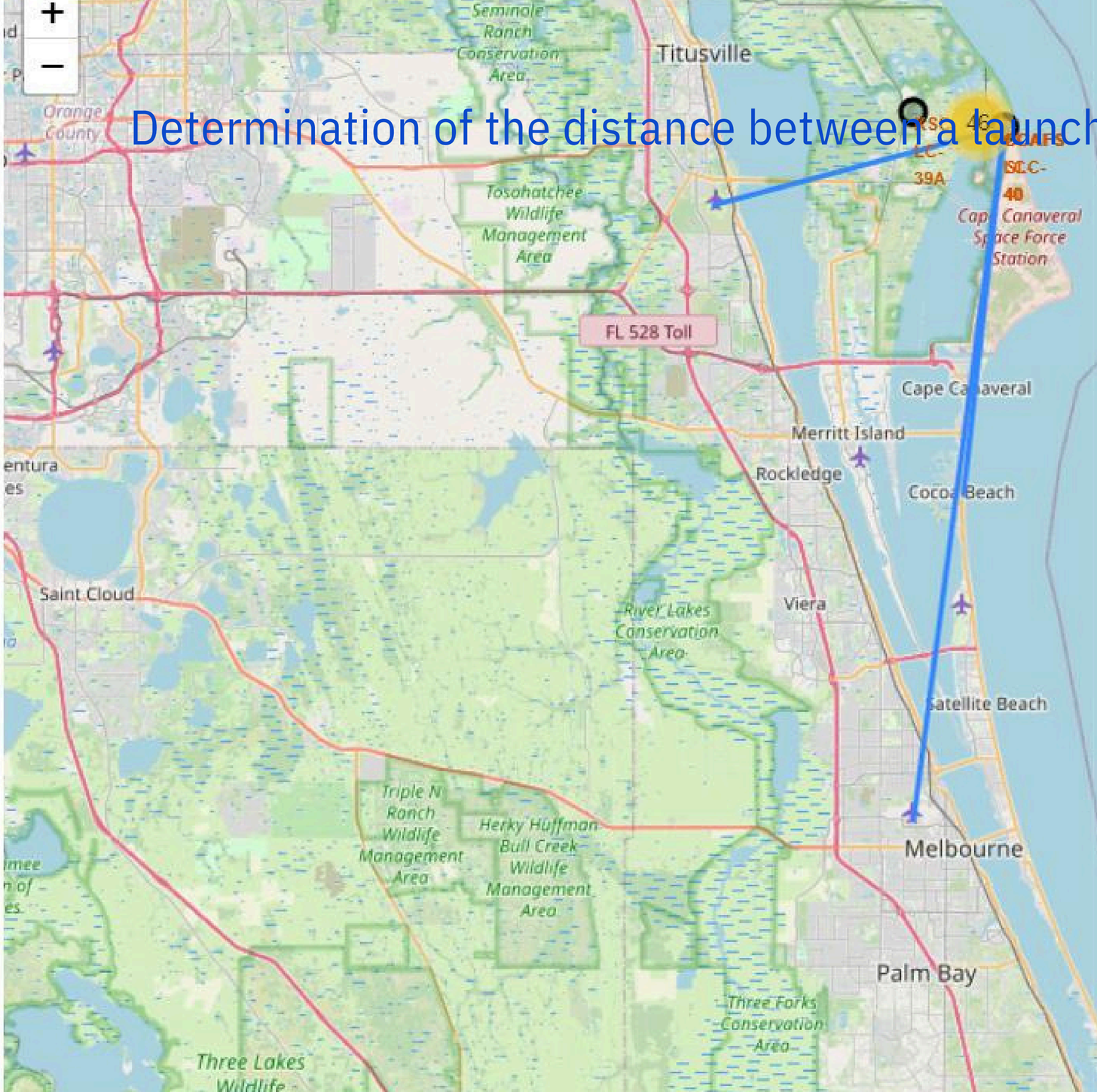
•Additionally, launch sites are located near the ocean because it provides a safe and efficient launching environment. It reduces the risk of damage and casualties in the event of a rocket failure, and also allows for a clear pathway downrange, minimizing the impact of rocket debris. Also, the ocean provides easy access to transportation facilities and a source of water for firefighting efforts.

Successful/Unsuccessful launches per site



- The launch outcomes for each location have been added to the map to for easieridentification of the ones with the highest success rate. Green color marks a successful launch while red marks an unsuccessful one.
- KSC LC-39A launch site has the highest success rate.

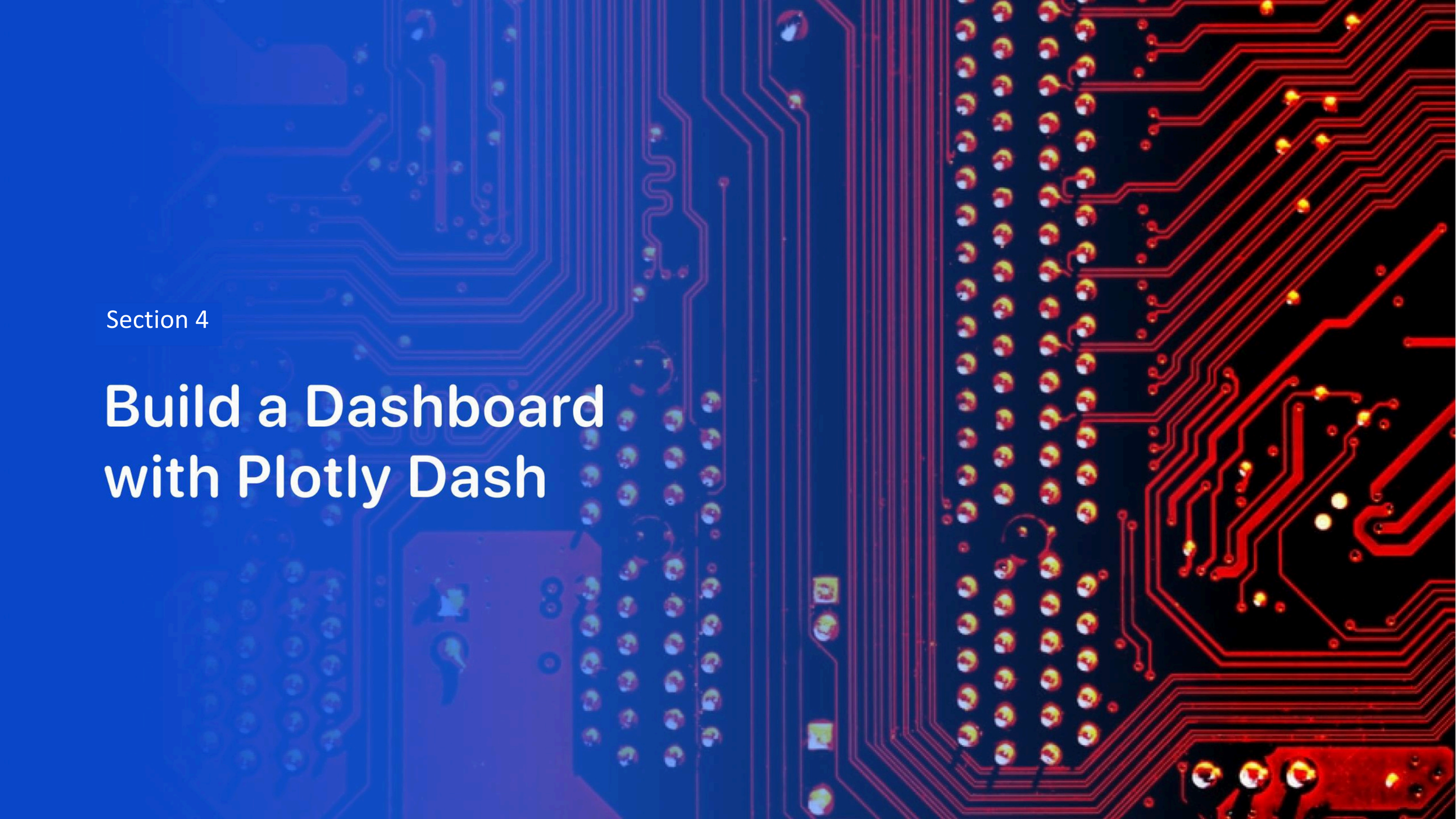
Determination of the distance between a launch site and its surroundings



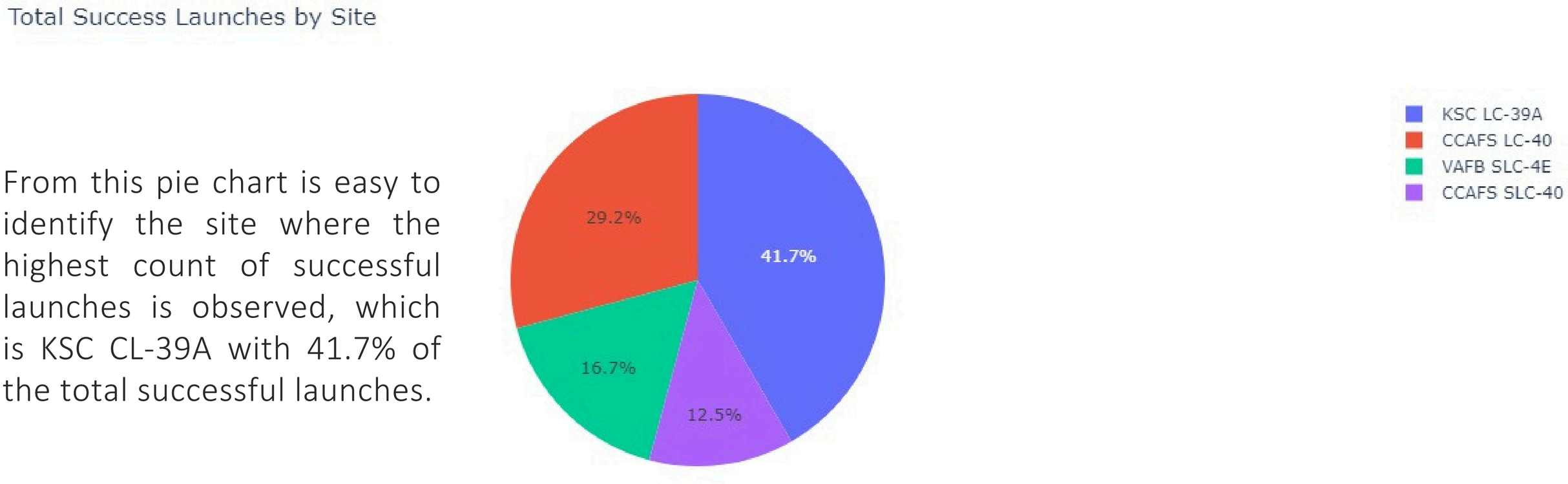
- Launch sites near railways ensure efficient transportation of heavy rocket components/ supplies.
- Highways provide logistical support for transporting materials, personnel, and supplies.
- Coastal locations reduces risks by enabling launches over unpopulated areas. Also, it reduces disturbance from the site's activities.
- Sites are distanced from cities to reduce risks and noise disturbances.

Section 4

Build a Dashboard with Plotly Dash



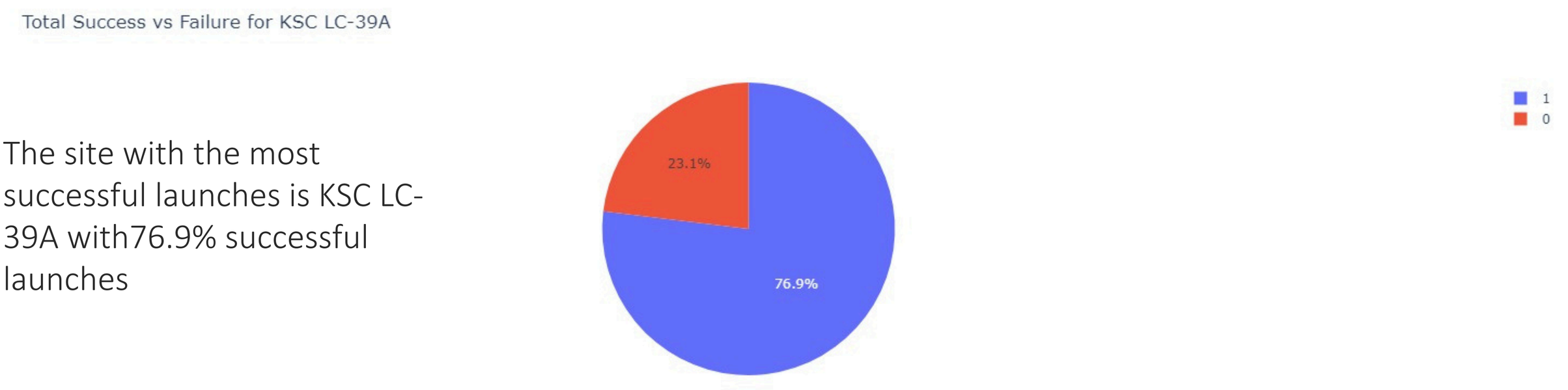
Launch success count for all sites



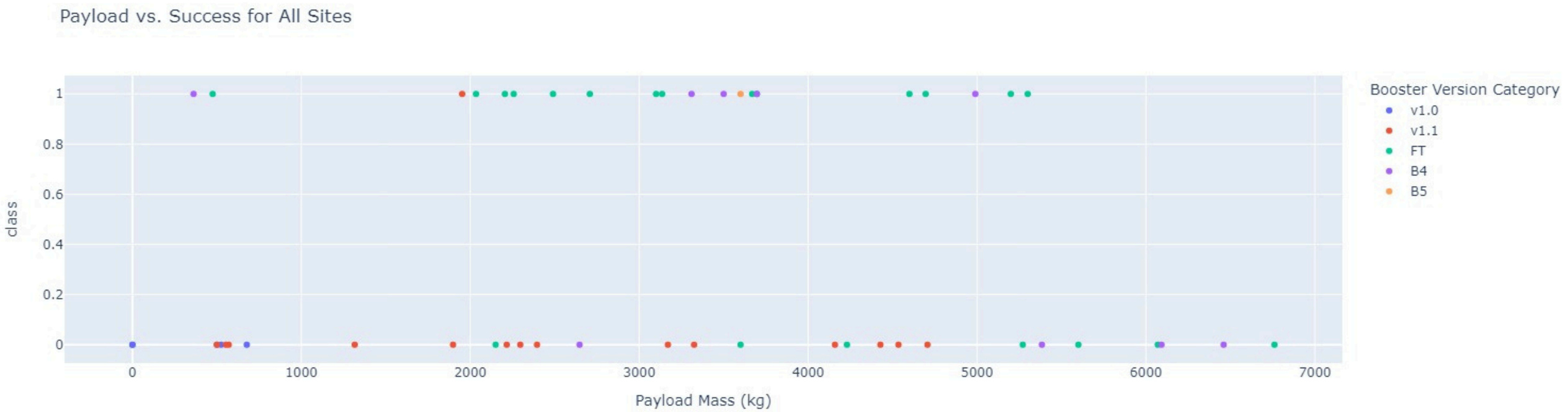
- From this pie chart is easy to identify the site where the highest count of successful launches is observed, which is KSC CL-39A with 41.7% of the total successful launches.

Site with the highest launch success ratio

- The site with the most successful launches is KSC LC-39A with76.9% successful launches



Payload vs. Launch Outcome per Launch Site

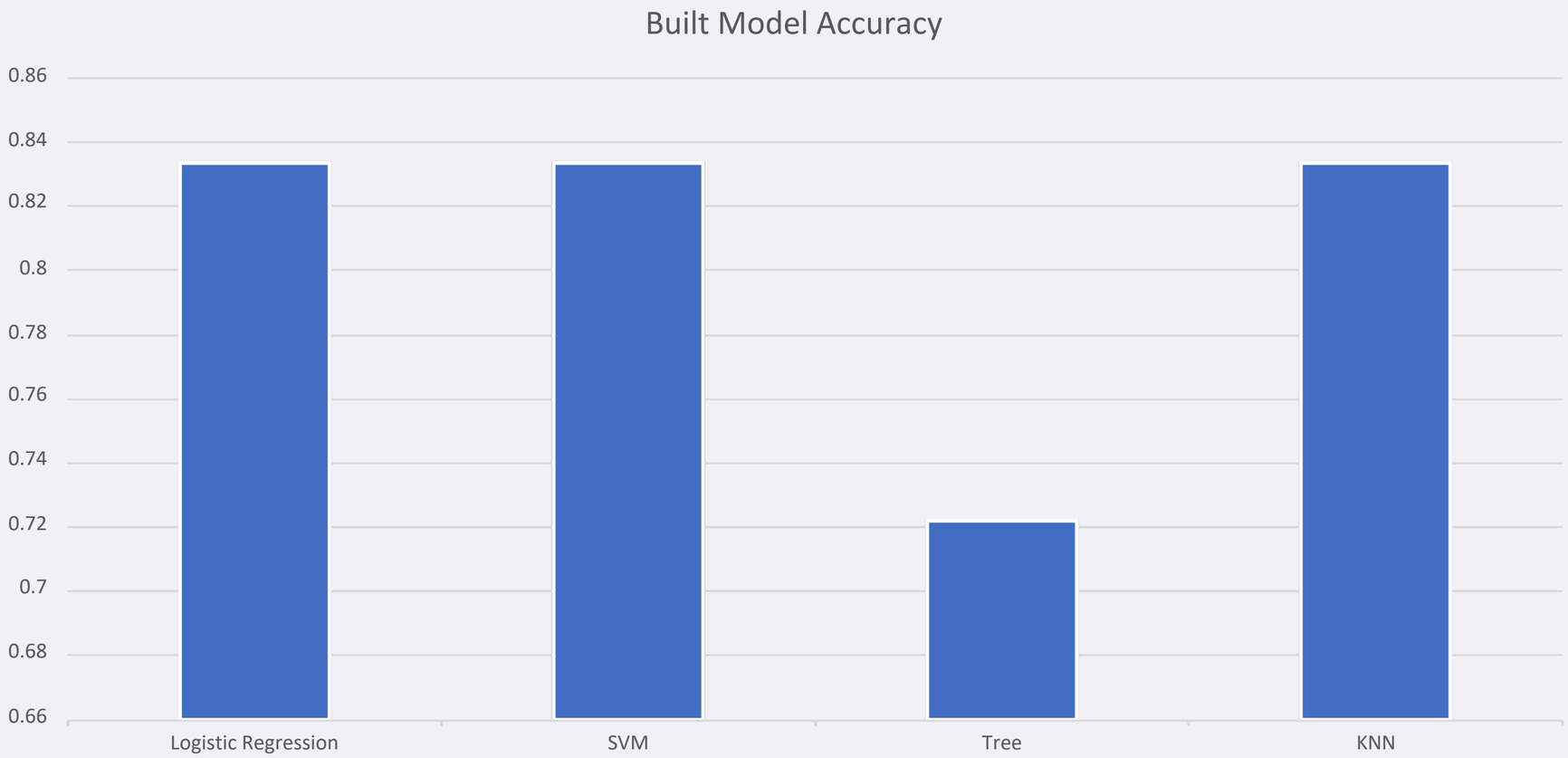


- B5 and FT boosters have shown the highest percentages of success.
- FT and B4 boosters have shown the highest percentage of success regarding the payload mass (over 4000 kg).

Section 5

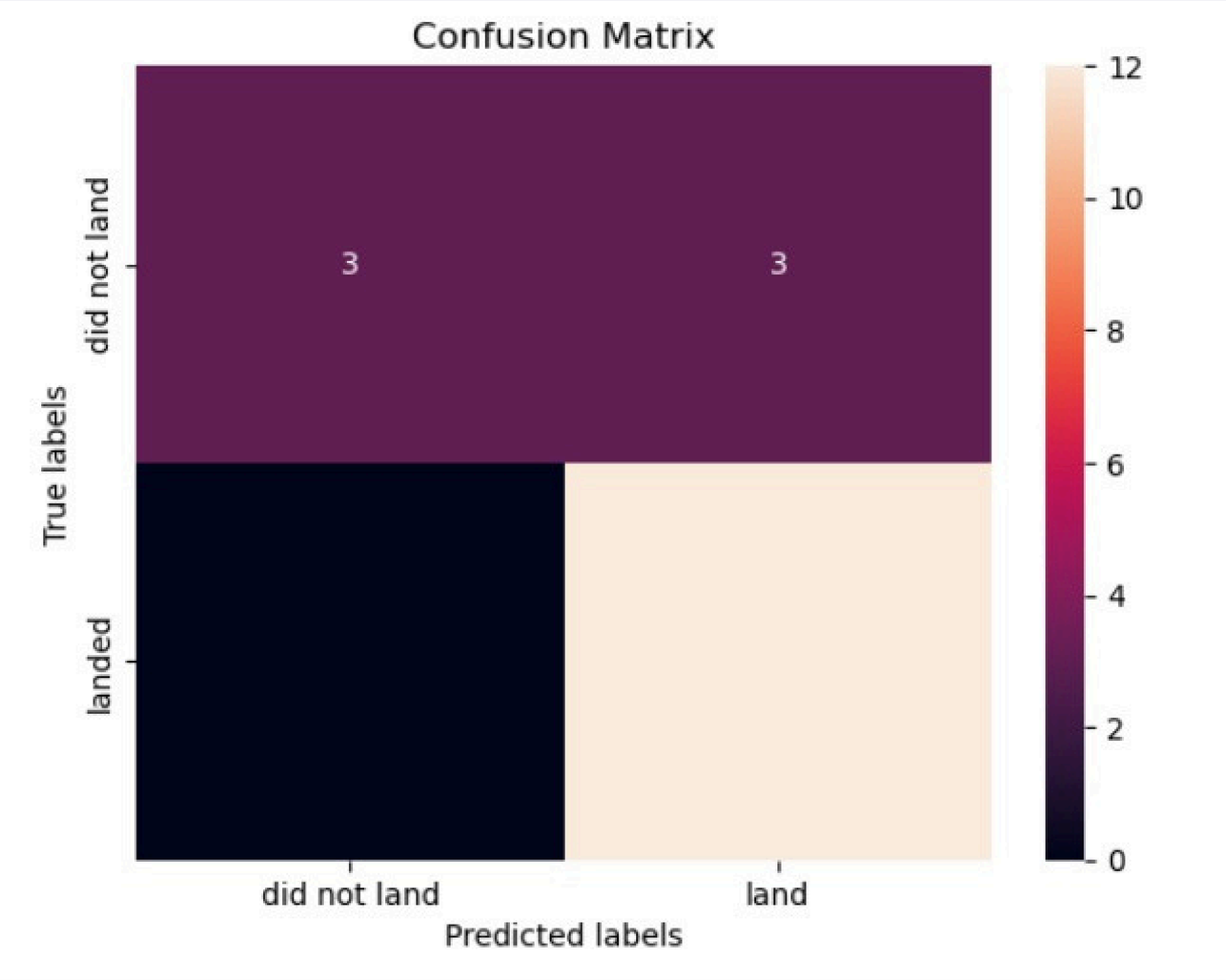
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression is the build model with the highest accuracy

Confusion Matrix



- The confusion matrix reveals the performance of the logistic regression model in distinguishing between classes. The model accurately predicts 12 instances where the rocket successfully landed (True Positive). However, it produces 3 False Positives, where the model incorrectly predicts a successful landing for rockets that did not land.

C onc lus io ns

- **Launch Sites and Success Rates**
- Most launches from VAFB SLC-4E and KSC LC-39A succeed.
- Heavy payloads (>10,000 kg) weren’t launched from VAFB SLC.
- CCAFS LC-40 shows mixed success across payload ranges.
- KSC LC-39A has the highest success rate (41.7% in pie chart, 76.9% overall).
- Launch outcomes are mapped: green for success, red for failure.
- **Payloads and Success Rates**
- Highest success rates: ES-L1, GEO, HEO, and SSO orbits.
- LEO success correlates with flight numbers.
- Heavy payloads succeed more in Polar, LEO, and ISS orbits.
- GTO success rates unclear with mixed results.
- FT and B4 boosters excel for payloads >4,000 kg.
- **Flight Number and Success**
- Higher flight numbers increase first-stage success.
  GTO shows no flight number-success correlation.

Conclusions (continue)

- Booster Performance
- B5 and FT boosters have the highest success rates.
- Yearly Trends
- Trends rise until 2014, stabilize in 2014-2015, and increase post-2015.
- First ground landing succeeded on 22-12-2015.
- Launch Site Proximities and Infrastructure
- Equatorial sites use Earth’s rotation for efficiency.
- Coastal sites reduce risks, manage debris, and offer firefighting resources.
- Railways enable heavy component transport; highways support logistics.
- Coastal and remote locations minimize risks and disturbances.
-
- Model Performance
- Logistic Regression achieves highest accuracy.
  Confusion matrix: 12 True Positives, 3 False Positives.

Appendix

- GitHub - Capstone Main Page
- Link to Dashboard

Thank you!