IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chi Sum, TSE
14/7/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
- Data Collection: The methodology involved gathering relevant data from various sources such as databases, APIs, or online repositories. This step ensured a comprehensive and diverse dataset for analysis.
- Data Wrangling: Data wrangling encompassed cleaning, transforming, and preprocessing the collected data to ensure its quality and compatibility for analysis. This step included handling missing values, dealing with outliers, and performing necessary data transformations.
- Exploratory Data Analysis with Data Visualization: Exploratory Data Analysis (EDA) was conducted to gain insights, discover patterns, and identify relationships within the dataset. Data visualization techniques were employed to present the findings visually, making it easier to understand and communicate the information effectively.
- Exploratory Data Analysis with SQL: In addition to traditional EDA techniques, SQL queries were used to explore the dataset further. SQL allowed for aggregating data, extracting specific information, and performing complex operations on databases, providing an additional analytical perspective.
- Building an Interactive Map with Folium: The methodology involved using the Folium library in Python to create interactive maps. Folium enabled the overlaying of data onto maps and the visualization of geospatial information. This step allowed for a spatial analysis and the representation of data in a geographic context.
- Building a Dashboard with Plotly Dash: A web-based dashboard was developed using Plotly Dash. This methodology facilitated the creation of interactive and user-friendly dashboards to showcase visualizations, metrics, and key insights derived from the data analysis. Dashboards provided an intuitive way to explore and present information.
- Predictive Analysis (Classification): Predictive analysis involved applying machine learning algorithms, specifically classification techniques, to make predictions or classify data based on training data patterns. This step aimed to identify patterns and relationships that could be used to predict or classify future instances accurately.

Executive Summary (con’t)

- Summary of all results
- Data Collection: Valuable data was successfully collected from public sources, ensuring a comprehensive dataset for analysis.
- Exploratory Data Analysis (EDA): Through EDA, important features were identified that have a significant impact on predicting the success of launchings. This analysis provided insights into the characteristics that contribute to successful outcomes.
- Machine Learning Prediction: Machine learning models were employed to predict the key characteristics that drive successful opportunities. By utilizing the collected data, the best-performing model was determined, which successfully identified the important factors that lead to favorable outcomes.

Introduction

- Background
- SpaceX, a leading company in the commercial space industry, has revolutionized space travel by making it more affordable. They achieve significant cost savings by reusing the first stage of their Falcon 9 rockets. Predicting the success of the first stage landing is crucial for determining the launch cost. In this project, we aim to use machine learning models and public information to predict whether SpaceX will reuse the first stage.
- Questions to be Answered:
- Impact of Variables: How do payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Success Rate Over Time: Does the rate of successful landings increase over the years?
- Best Classification Algorithm: What is the most effective algorithm for binary classification in predicting the reuse of the first stage?

Section 1

Methodology

Methodology

Executive Summary

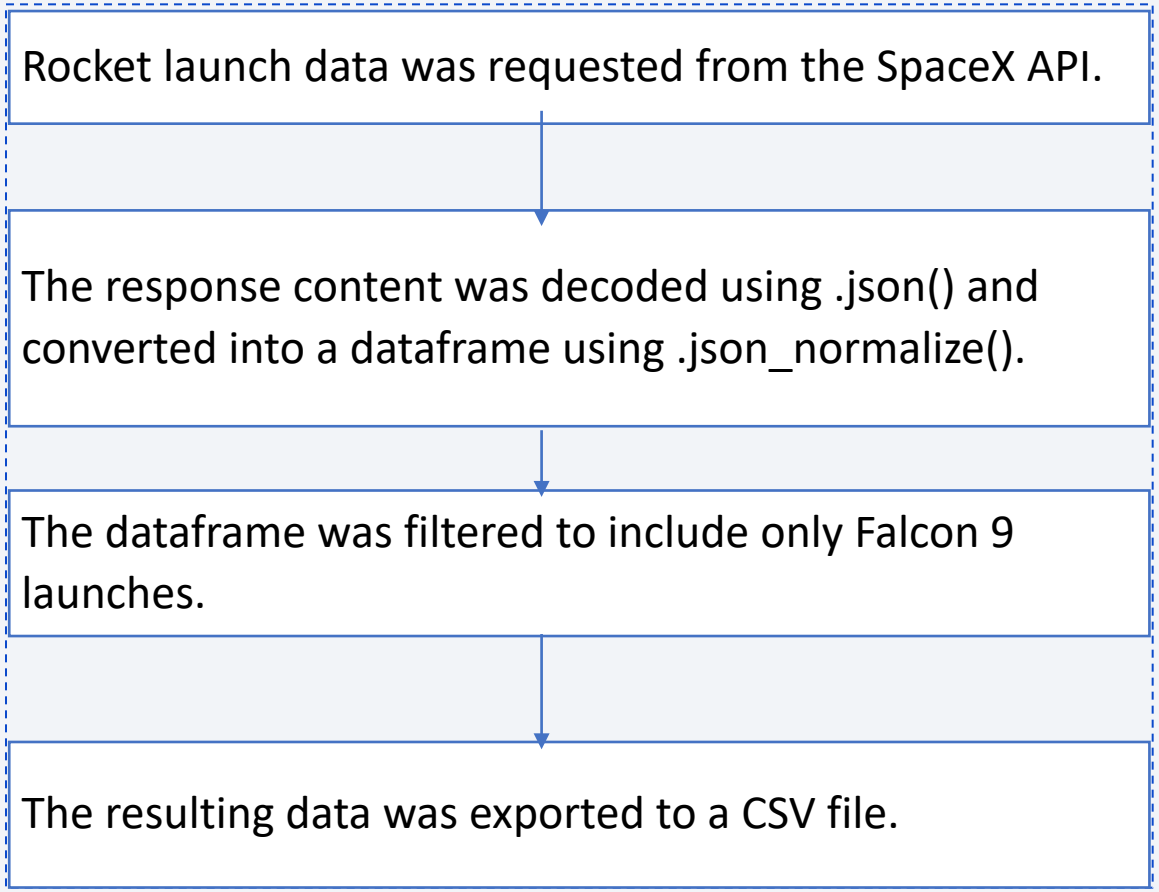
- Data collection methodology:
  - Space X API
  - Webscraping from Wikipedia
- Perform data wrangling
  - the collected data was filtered, missing values were handled, and One Hot Encoding was applied to prepare the data for binary classification, while also enriching the dataset by creating a landing outcome label based on summarized and analyzed features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data was normalized, split into training and test datasets, and evaluated using four different classification models with varying parameter combinations, aiming to build, tune, and evaluate the models to achieve the best results in terms of accuracy.

Data Collection

- Data collection involved a combination of API requests from SpaceX REST API and web scraping from SpaceX's Wikipedia entry.
- Both methods were used to ensure comprehensive information about launches for detailed analysis.
- The SpaceX REST API provided the following data columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Web scraping from Wikipedia provided the following data columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

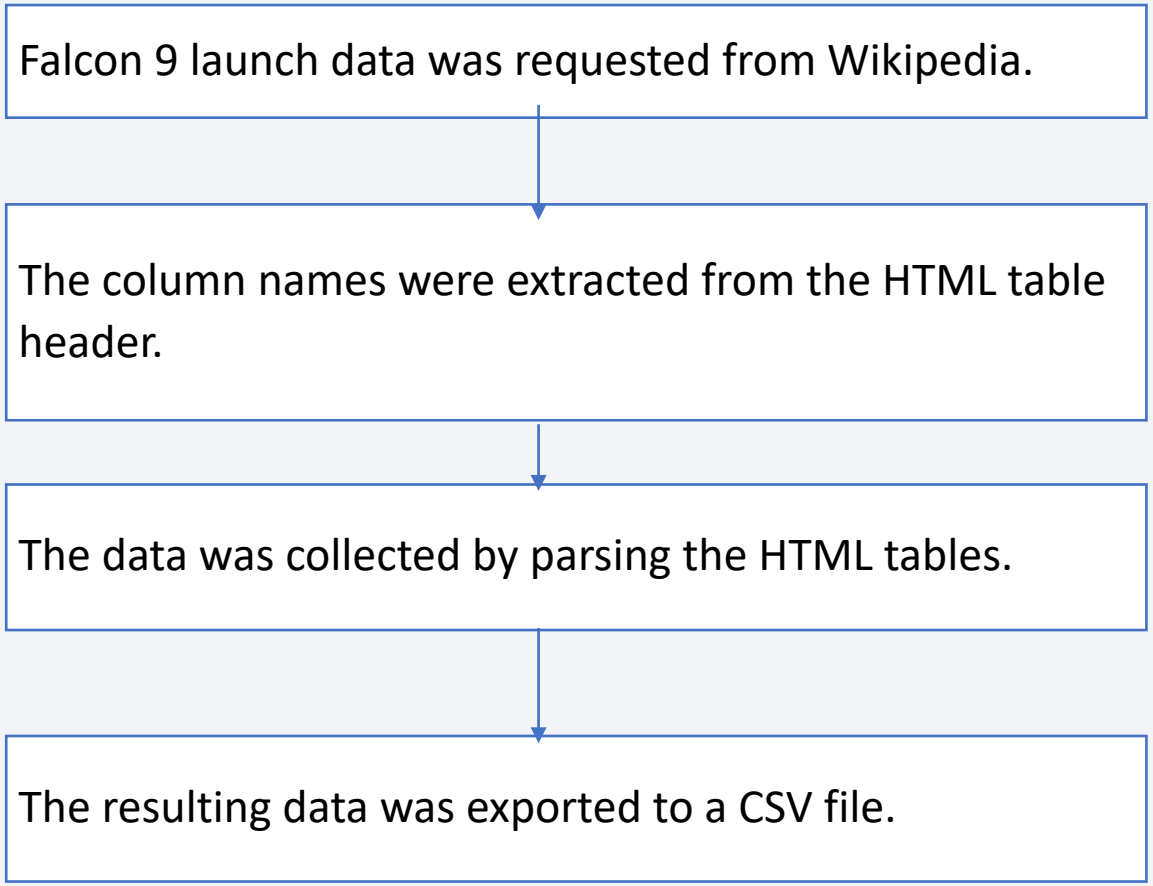
- SpaceX provides a public API that was utilized to obtain data.
- The data was obtained by following the flowchart provided and then persisted for further use.



- GitHub URL : Data Collection API

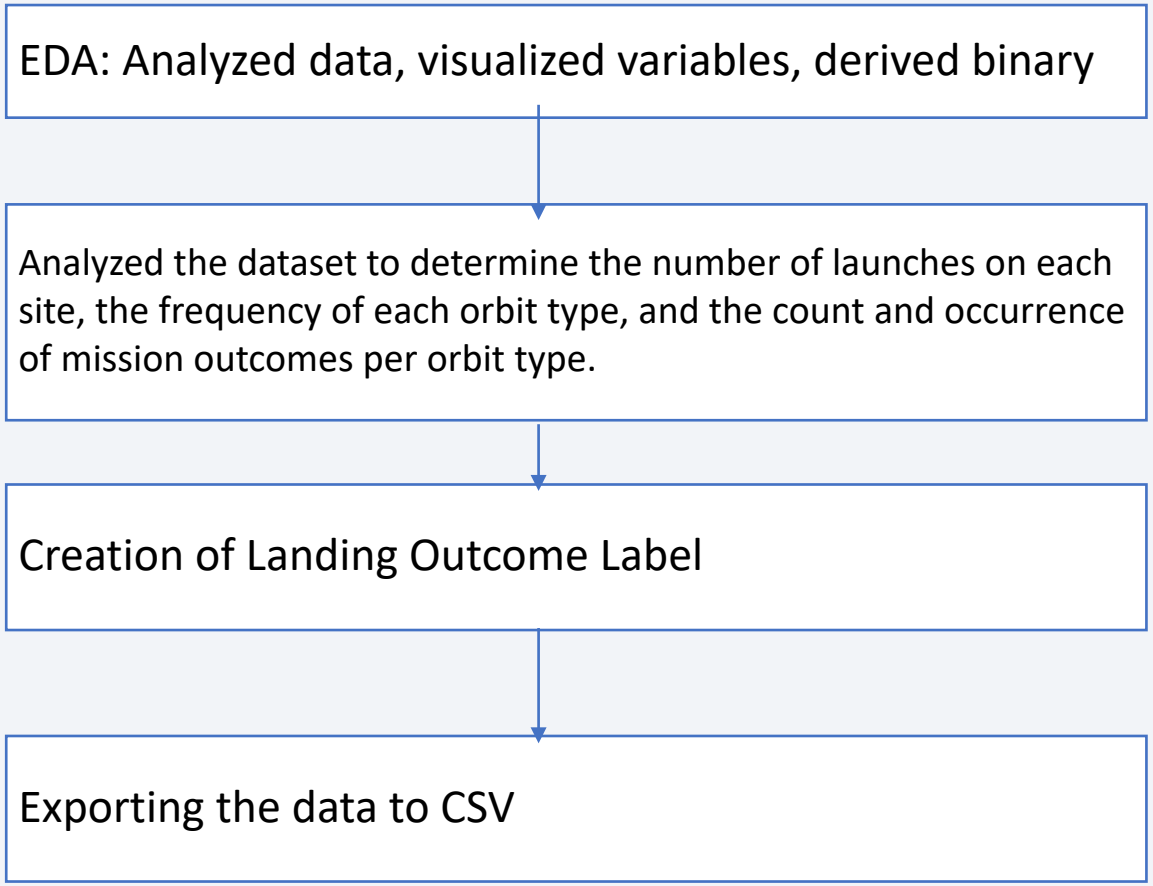
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia.
- The data is downloaded from Wikipedia based on the provided flowchart.
- Once downloaded, the data is saved for further use.
- GitHub URL: Data Collection with Web Scraping



Data Wrangling

- In the dataset, different cases exist where the booster did not land successfully.
- Various outcomes indicate the landing attempts, such as True Ocean (successful landing in the ocean), False Ocean (unsuccessful landing in the ocean), True RTLS (successful landing on a ground pad), False RTLS (unsuccessful landing on a ground pad), True ASDS (successful landing on a drone ship), and False ASDS (unsuccessful landing on a drone ship).
- These outcomes are converted into training labels, where "1" represents a successful booster landing and "0" represents an unsuccessful landing.
- GitHub URL: Data Wrangling



EDA with Data Visualization

- Plotted scatter plots to examine relationships between variables, such as Flight Number vs. Payload Mass, Flight Number vs. Launch Site, and Payload Mass vs. Launch Site, to potentially utilize them in a machine learning model.
- Created bar charts to compare categories, including Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, and Payload Mass vs. Orbit Type, to identify relationships between specific categories and measured values.
- Utilized line charts to visualize the yearly trend of the success rate, providing insights into the performance over time and identifying any patterns or trends.

EDA with SQL

- Extracted the names of unique launch sites in the space mission.
- Retrieved the top 5 launch sites with names starting with 'CCA'.
- Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Determined the average payload mass carried by booster version F9 v1.1.
- Identified the date of the first successful landing outcome on a ground pad.
- Retrieved the names of boosters with successful landings on a drone ship and payload mass between 4000 and 6000 kg.
- Calculated the total number of successful and failed mission outcomes.
- Identified the booster versions that carried the maximum payload mass.
- Retrieved failed landing outcomes on a drone ship, along with their booster versions and launch site names for the year 2015.
- Determined the rank of landing outcome counts (Failure - drone ship or Success - ground pad) between the dates 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

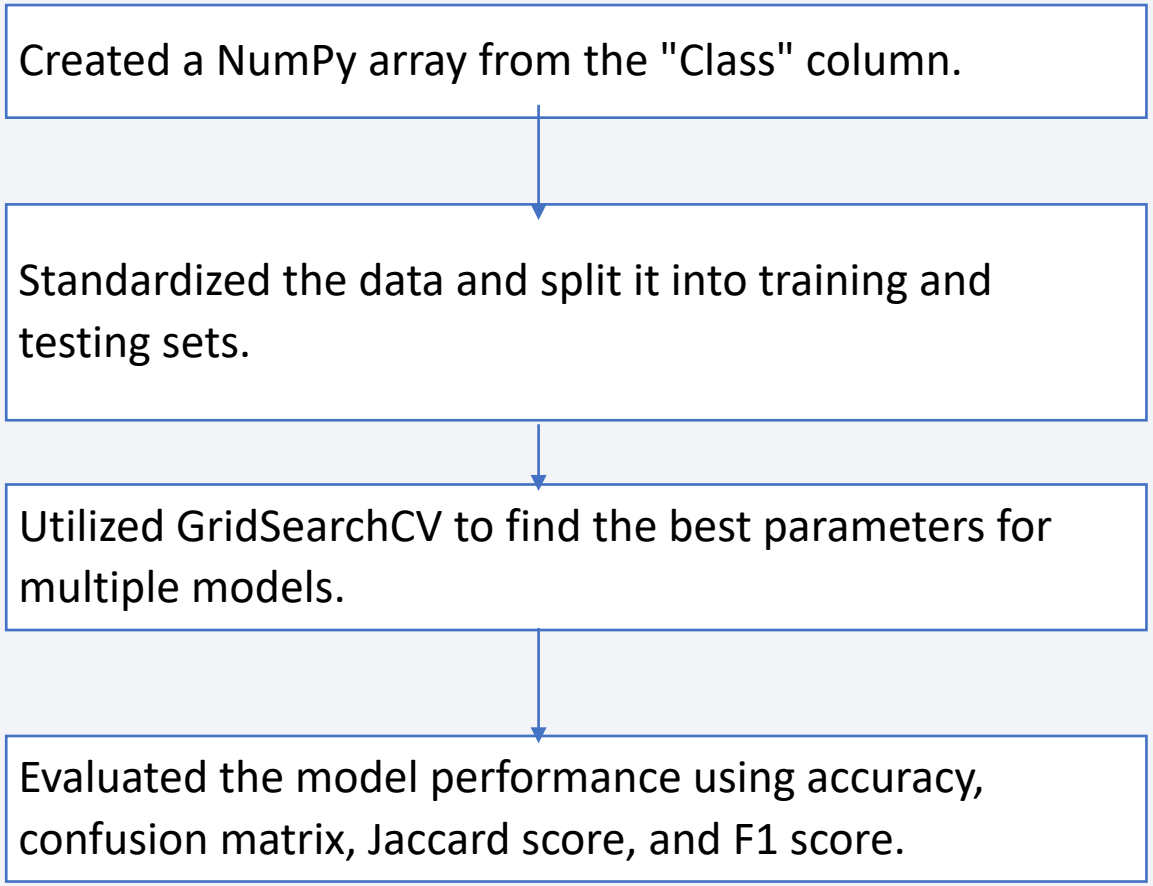
- Markers, Circles, Lines, and Marker Clusters:
- Utilized Markers to indicate points such as launch sites. Implemented Circles to highlight areas around specific coordinates, such as NASA Johnson Space Center.
- Employed Marker Clusters to group events in each coordinate, such as launches at a specific launch site.
- Incorporated Lines to indicate distances between two coordinates, such as the distance between a launch site and its proximities.
- Markers of Launch Sites:
- Added Markers with Circle, Popup Label, and Text Label for all launch sites, showcasing their geographical locations and proximity to the Equator and coasts.
- Coloured Markers of Launch Outcomes: Utilized coloured Markers (Green for success, Red for failure) with Marker Clusters to indicate the success rates of launches at each launch site.
- Distances between Launch Sites and Proximities:
- Added coloured Lines to visualize distances between a launch site (e.g., KSC LC-39A) and its proximities, including the Railway, Highway, Coastline, and Closest City.

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List: Implemented a dropdown list to facilitate Launch Site selection.
- Pie Chart for Success Launches: Utilized a pie chart to display the total count of successful launches for all sites, and if a specific Launch Site was chosen, showed the success vs. failed counts for that site.
- Slider for Payload Mass Range: Added a slider component to enable the selection of a Payload mass range. Scatter Chart for Payload Mass vs. Success Rate by Booster Versions: Incorporated a scatter chart to visualize the relationship between Payload mass and Launch Success rate for different Booster Versions.
- Additional Visualization Techniques: Utilized a percentage chart to showcase the distribution of launches by site. Employed a payload range visualization to identify the relationship between payloads and launch sites, aiding in determining the optimal launch site based on payload requirements.

Predictive Analysis (Classification)

- Four classification models were compared
- logistic regression
- support vector machine
- decision tree
- k nearest neighbours.



Results

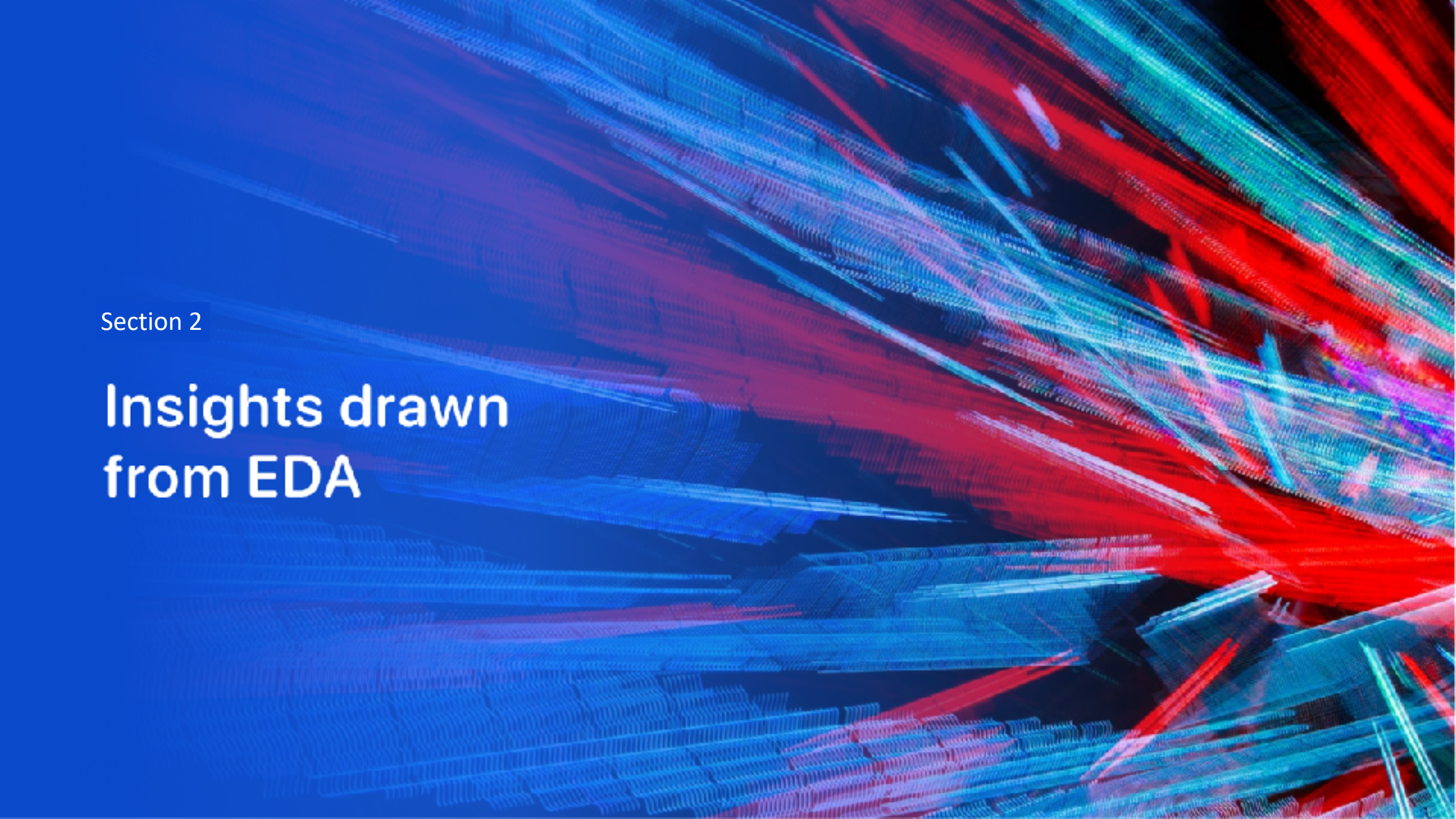
- SpaceX operates four different launch sites for its missions.
- Initially, the first launches were conducted by SpaceX itself and NASA.
- The average payload of the F9 v1.1 booster is approximately 2,928 kg.
- The first successful landing outcome occurred in 2015, five years after the initial launch.
- Many Falcon 9 booster versions achieved successful landings on drone ships, particularly with payloads above the average.
- Nearly 100% of mission outcomes were successful. In 2015, two booster versions (F9 v1.1 B1012 and F9 v1.1 B1015) failed at landing on drone ships.
- The number of successful landing outcomes improved over the years, indicating progress and improvements in landing capabilities.

Results

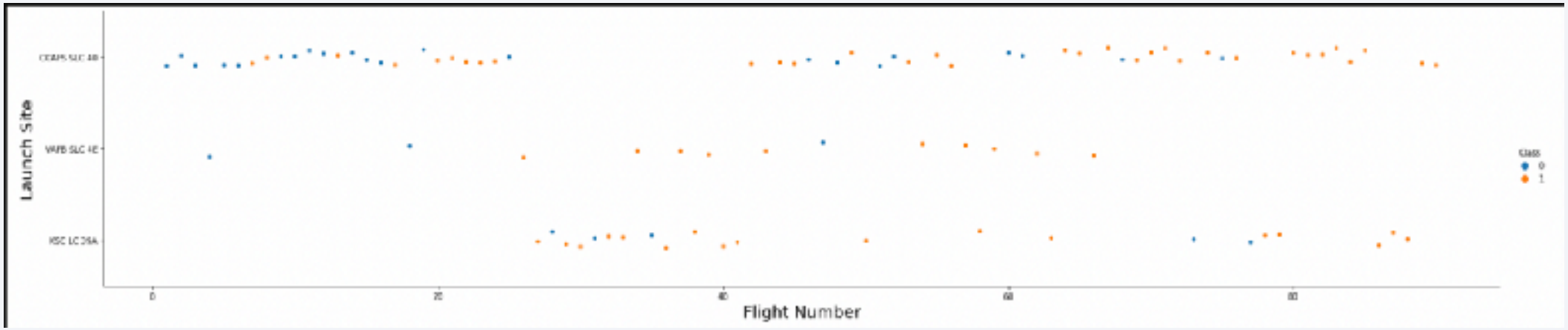
- Most of the launches took place at the East Coast.

Results

- Decision Tree Classifier proved to be the best model which predicted successful landings, having accuracy over 87% and accuracy for test data over 94%.

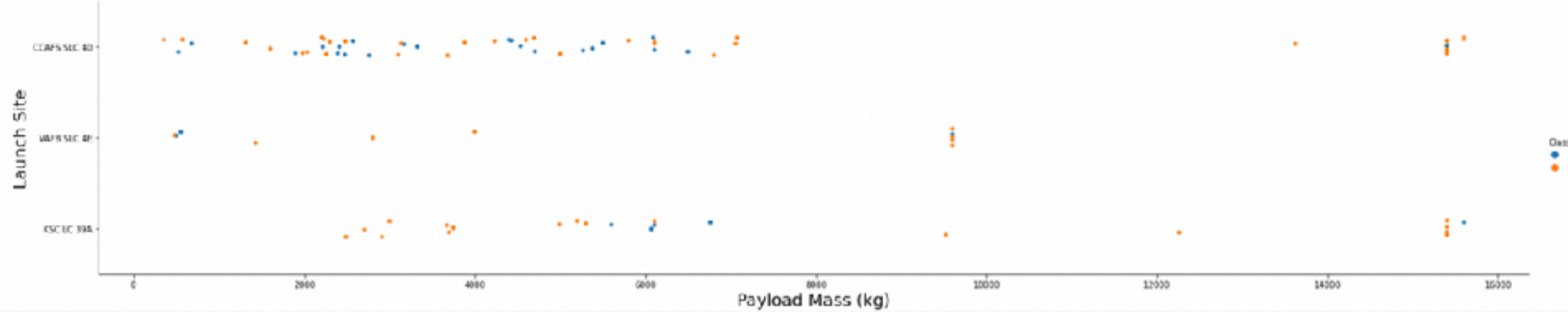

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- The best launch site in terms of recent success rate is CCAF5 SLC 40, accounting for approximately half of all launches.
- The second and third best launch sites are VAFB SLC 4E and KSC LC 39A, respectively.
- The general success rate of launches has improved over time, with earlier flights experiencing more failures and later flights having higher success rates.

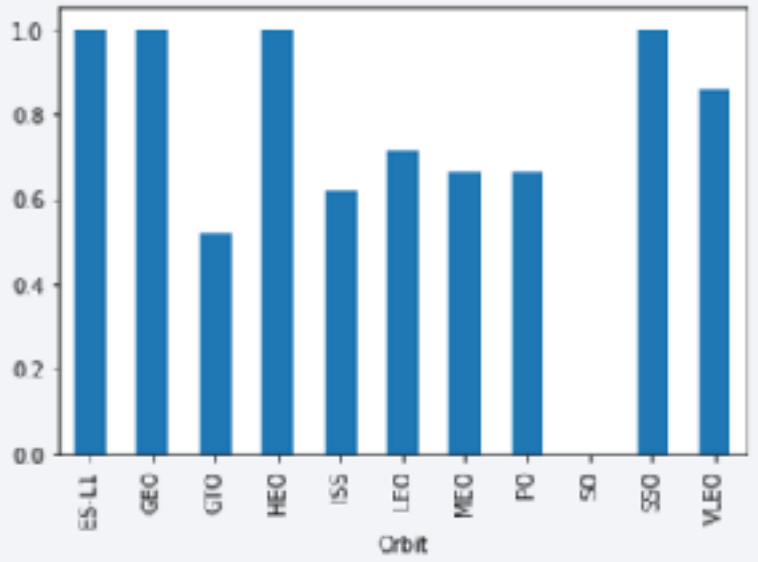
Payload vs. Launch Site



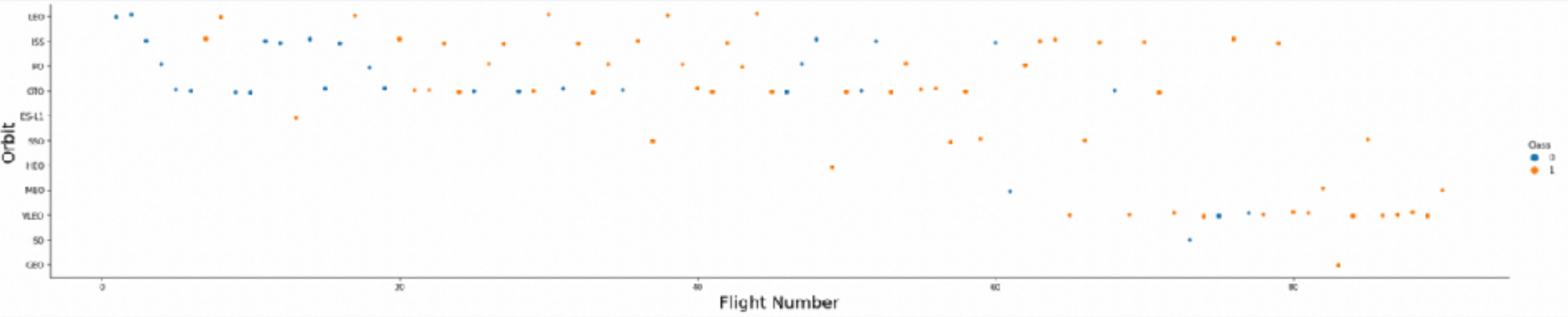
- Higher payload mass is correlated with a higher success rate for every launch site.
- Launches with payload mass over 7000 kg have a higher success rate.
- Payloads over 9000 kg have an excellent success rate, comparable to the weight of a school bus.
- Payloads over 12,000 kg appear to be possible only at CCAFS SLC 40 and KSC LC 39A launch sites. KSC LC 39A shows a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type

- 100% success rate: ES-L1, GEO, HEO, SSO
- Success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO
- 0% success rate: - SO

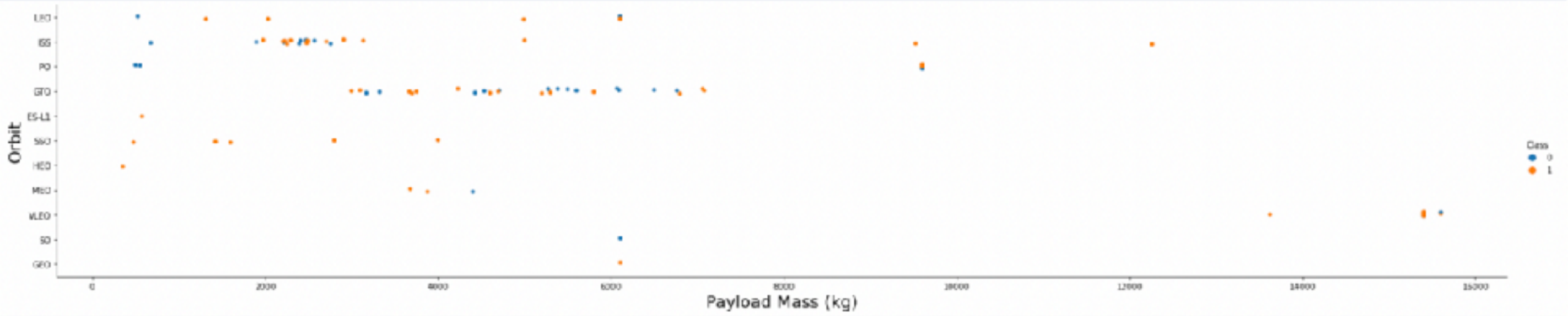


Flight Number vs. Orbit Type



- In the LEO orbit, the success rate appears to be related to the number of flights, suggesting that more experienced flights have a higher success rate.
- However, there seems to be no relationship between flight number and success rate in the GTO orbit.
- Overall, the success rate has improved over time for all orbits, indicating advancements and improvements in launch capabilities.
- The frequency of launches in the VLEO orbit has recently increased, indicating a potential new business opportunity in this orbit.

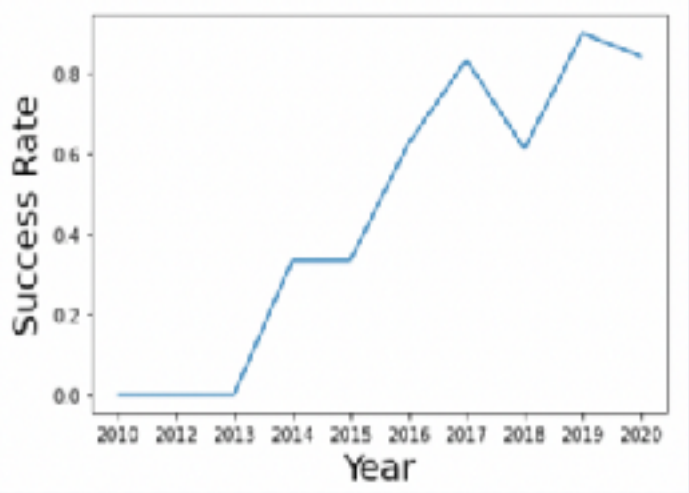
Payload vs. Orbit Type



- Payload size has a negative impact on GTO orbits but a positive influence on GTO and Polar LEO orbits, while success rates show no clear correlation with payload in GTO orbits, the ISS orbit demonstrates a wide payload range and a high success rate, and the SO and GEO orbits have fewer launches.

Launch Success Yearly Trend

- Since 2013, the success rate kept increasing.



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

total_payload_mass
45596

Average Payload Mass by F9 v1.1

average_payload_mass
2534

First Successful Ground Landing Date

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

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

Boosters Carried Maximum Payload

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

2015 Launch Records

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

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

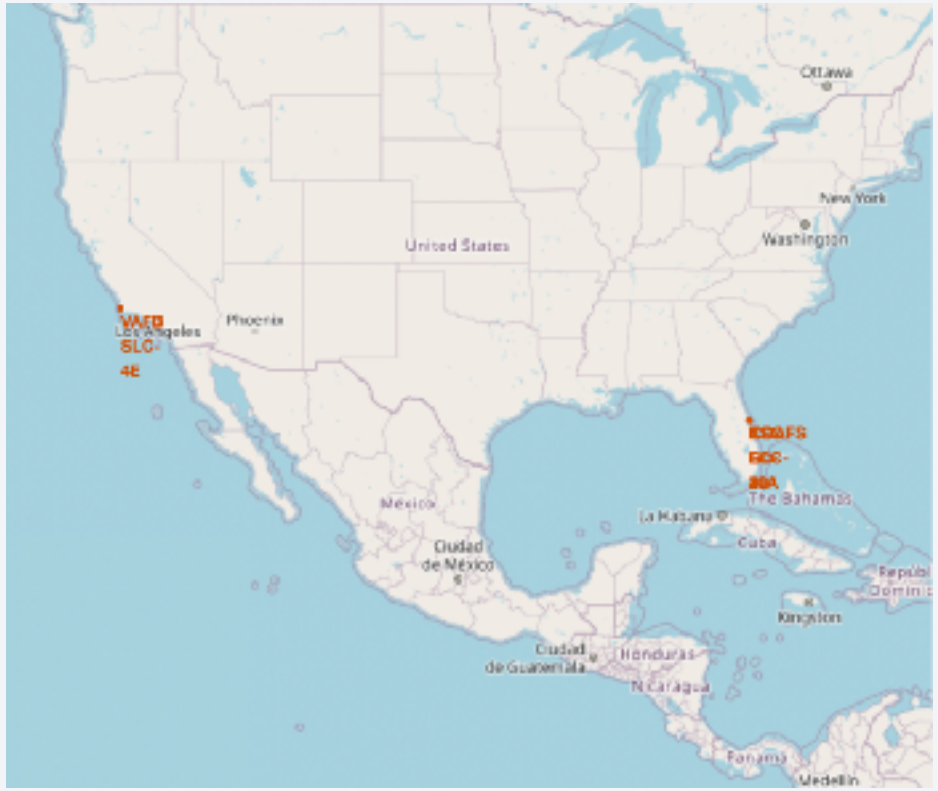
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

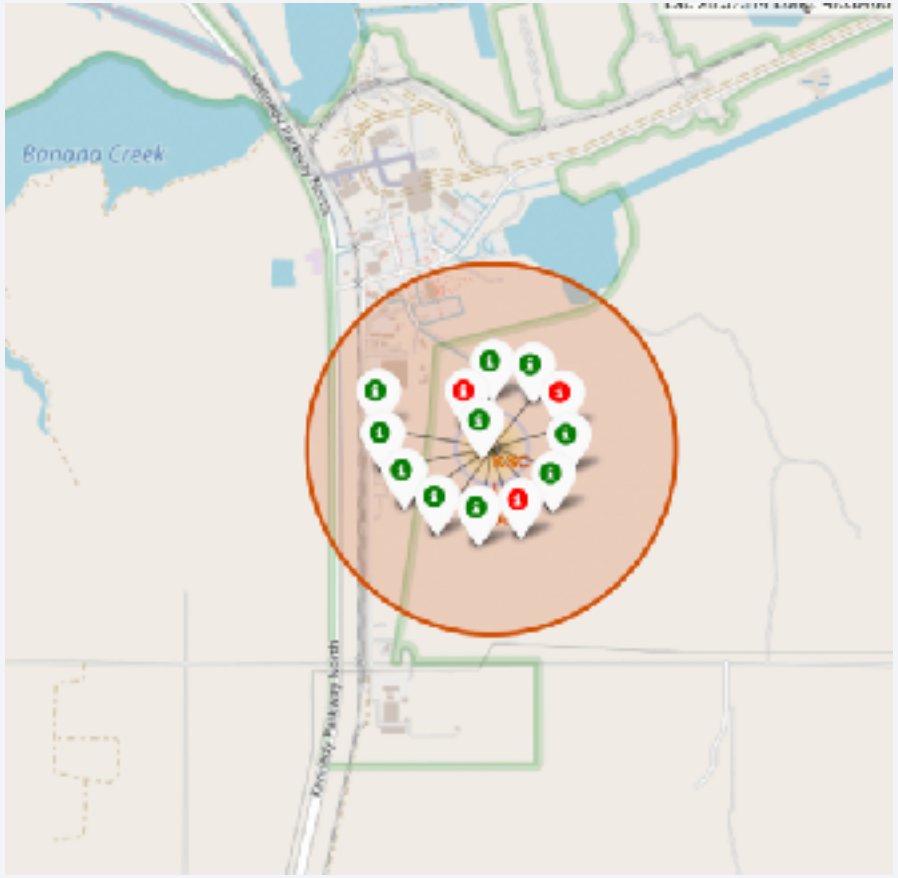
All launch sites’ locations

- All launch sites are situated near coastlines



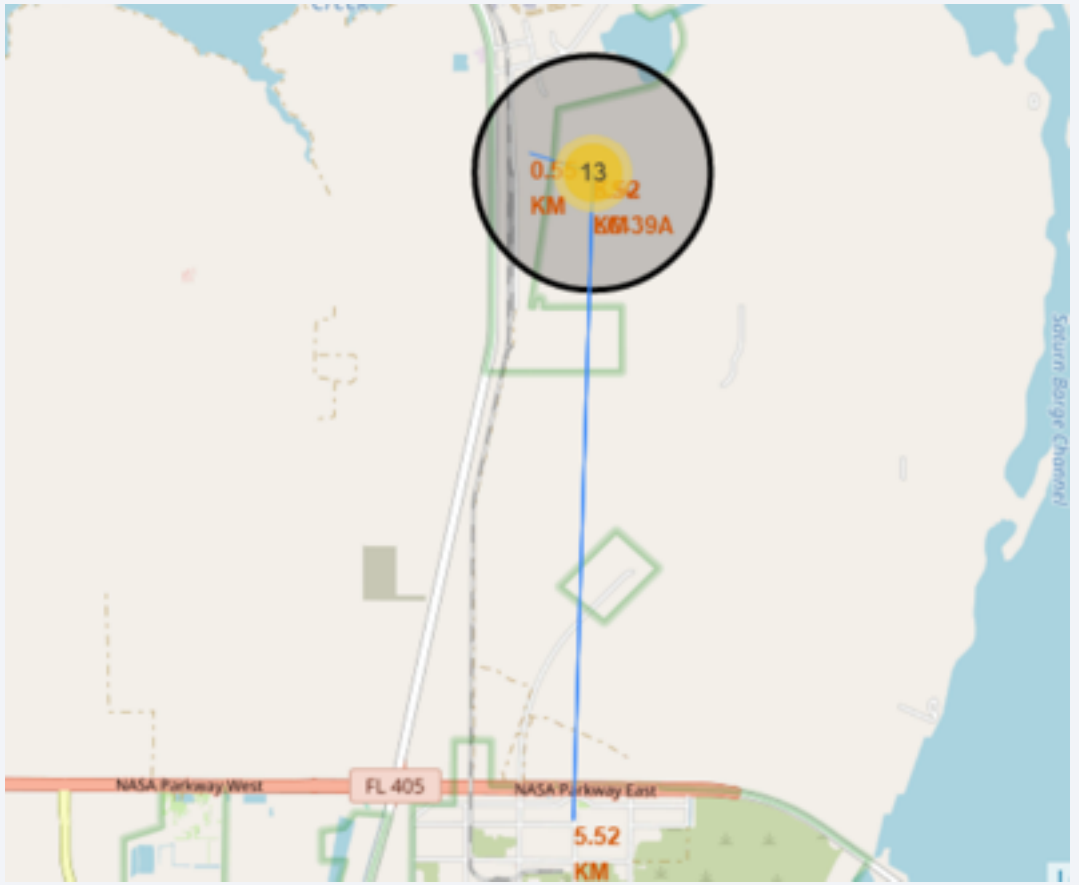
Site with highest success rate

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



Feature of Launch Site KSC LC-39A

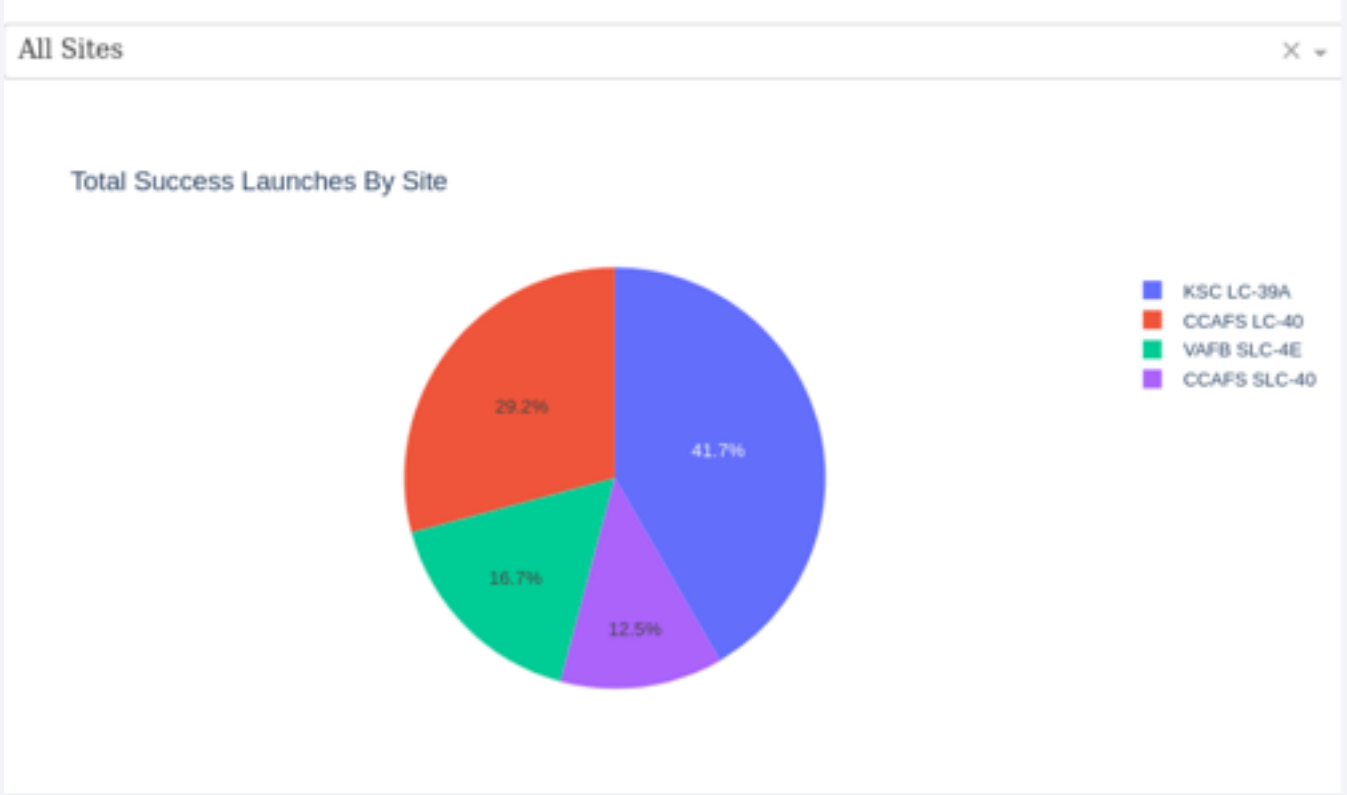
- good logistics aspects, being close to railroads, highway and coastline



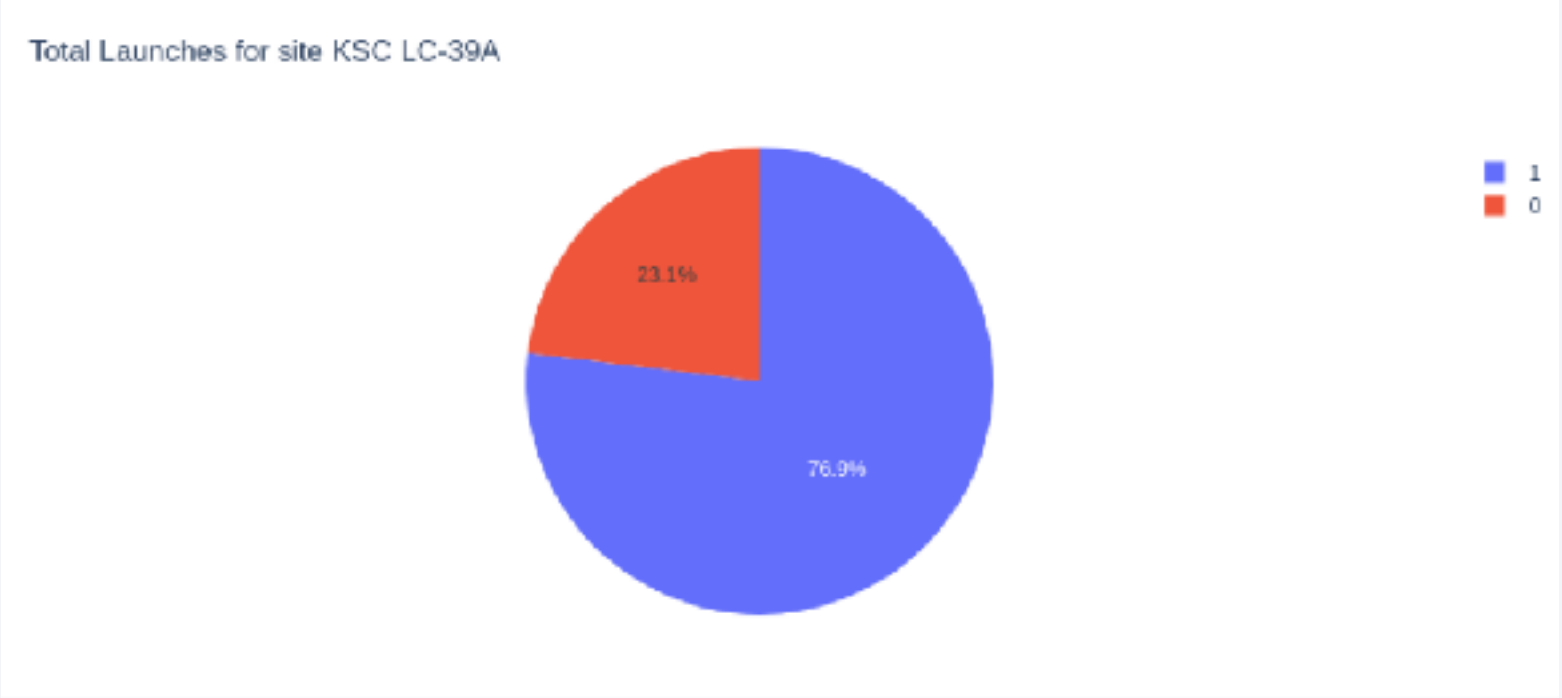
Section 4

Build a Dashboard with Plotly Dash

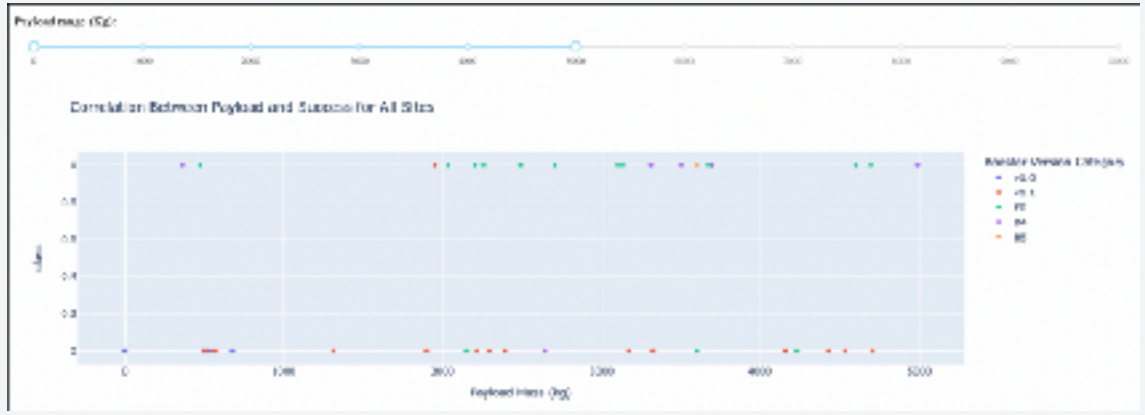
Success rate by launch site

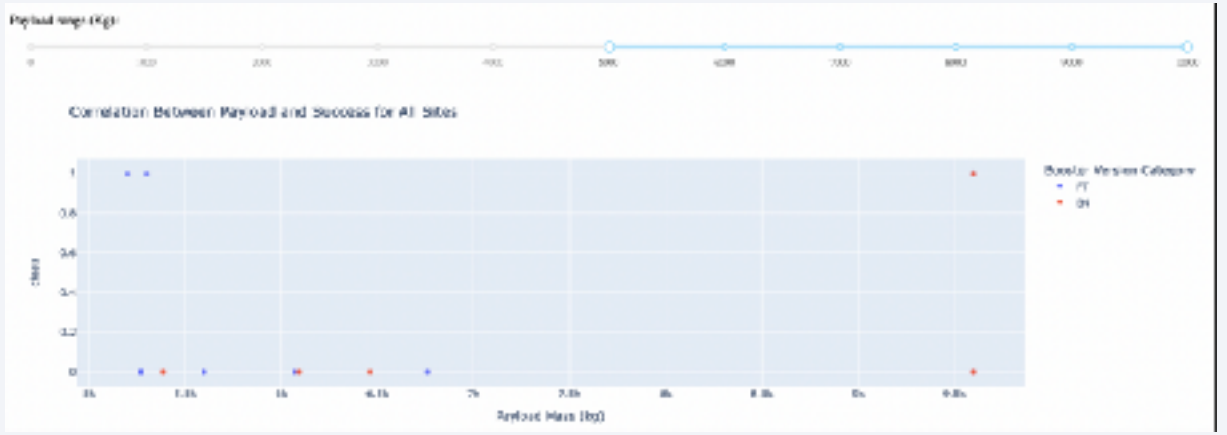


Launch site with highest launch success rate



Payload vs. Launch Outcome by site





Section 5

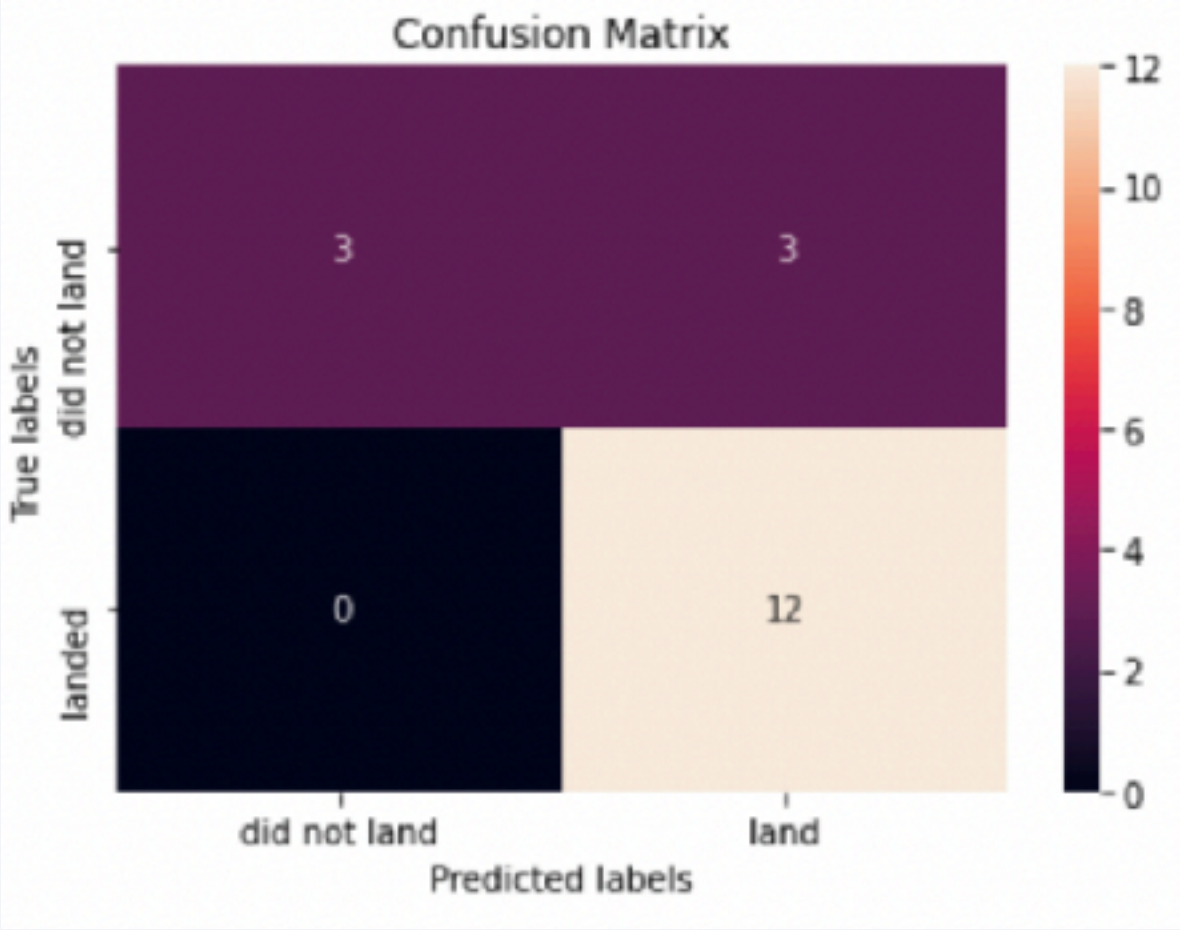
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree Classifier has the highest classification accuracy which is over 87%.

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

Confusion Matrix



Conclusions

- The Decision Tree model is determined to be the best algorithm for this dataset.
- Launches with lower payload masses tend to have better success rates compared to launches with larger payload masses.
- The majority of launch sites are located near the Equator and in close proximity to the coast.
- The success rate of launches has been observed to increase over the years.
- KSC LC-39A is identified as the launch site with the highest success rate. Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.
- The analysis of different data sources led to refined conclusions throughout the process.
- Launches with payloads above 7,000kg are considered to be less risky.

Thank you!