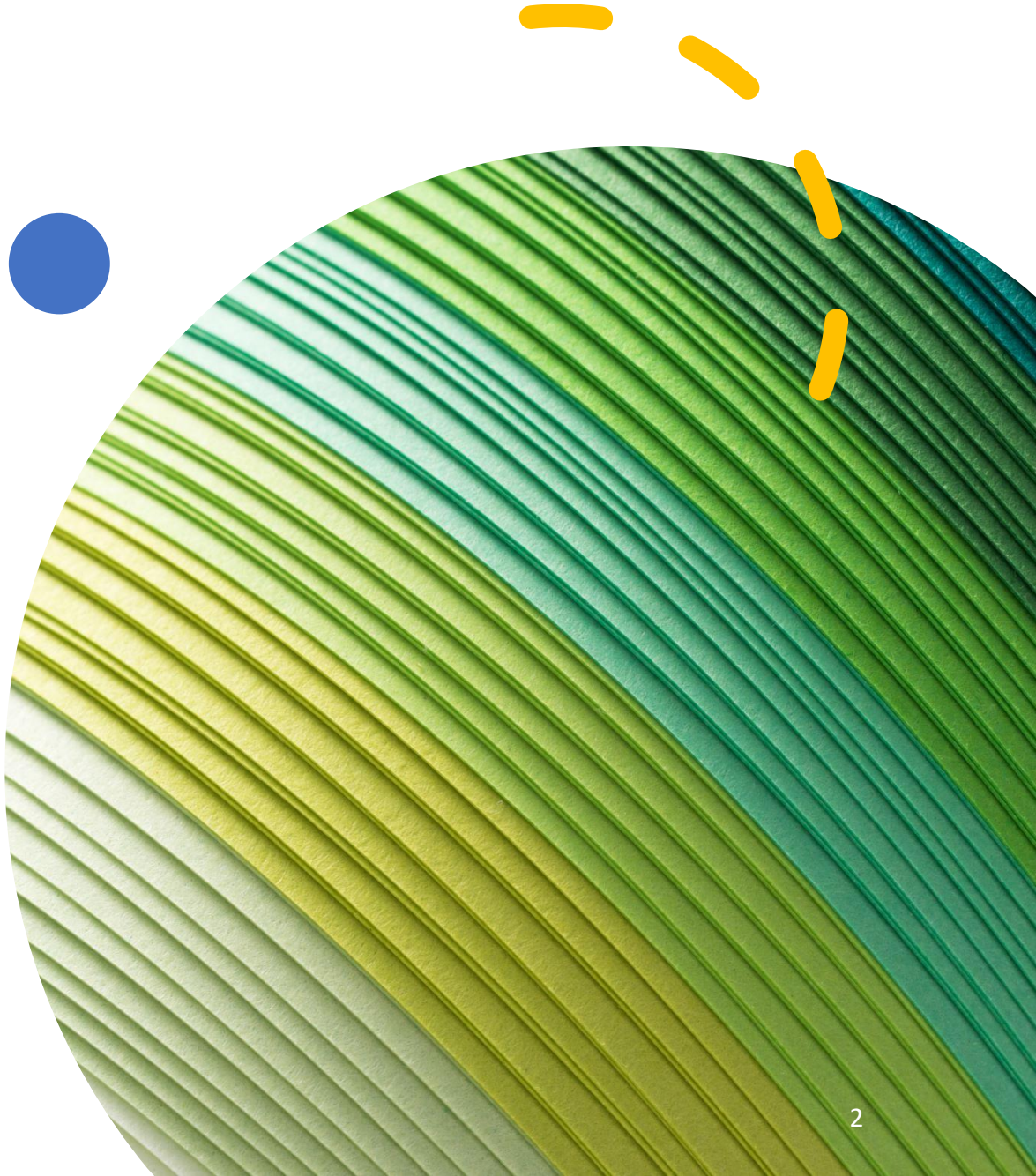
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alain Villesuzanne
05/08/2023

Outline

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



Executive Summary

- SpaceY, a burgeoning player in the commercial rocket launch sector, is positioning itself to challenge the dominance of SpaceX. With SpaceX offering launch services at a base cost of $62 million and a reusability feature, SpaceY aims to enhance its competitive edge by accurately predicting the success of first-stage rocket landings.
- Our study is a comprehensive endeavor that merges data from public SpaceX APIs and leverages web scraping techniques to extract information from SpaceX's Wikipedia page. The methodology encompasses a series of stages, including data collection, wrangling, exploratory analysis, and sophisticated visualization techniques. This culminated in the identification of crucial features and the meticulous fine-tuning of machine learning models.
- Our models, which encompass Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K Nearest Neighbors (KNN), consistently yield an impressive accuracy rate of approximately 83.33% (sometimes more for Decision Tree) in predicting the success of first-stage rocket landings. It's noteworthy that these models demonstrate a propensity to slightly over-predict success.
- In light of SpaceX's public disclosure of a $15 million cost for the first-stage Falcon 9 booster, our models offer a reliable means of predicting first-stage landing success with a commendable accuracy level of 83.3% or more. This predictive capability equips SpaceY with a powerful tool for strategic decision-making, enabling precise cost estimation and competitive bidding against SpaceX.
- It should be noted, however, that work to enrich the analysis data will be essential to maintain and make reliable the relevance of the estimates.

Introduction : Predicting Falcon 9 First Stage Landing for Cost Efficiency

- **SpaceX**, at the forefront of **innovation** in the **aerospace** industry, has consistently redefined the possibilities of **space travel** by making it more accessible and **cost-effective**. The company's remarkable achievements range from supplying the International Space Station to deploying a groundbreaking satellite constellation for global internet connectivity. A pivotal factor in SpaceX's ability to revolutionize space travel lies in its **unique approach** to rocket launches, exemplified by the **Falcon 9 rocket**.
- The **Falcon 9 rocket** stands out as a **pioneering solution**, boasting a relatively **low** launch **cost** of **$62 million** compared to the industry standard of upwards of $165 million per launch. This drastic cost differential can be attributed to SpaceX's ingenious strategy of **reusing the first stage** of the Falcon 9 rocket, an innovation that dramatically **reduces** production **expenses** and fuels a more sustainable space travel model.
- However, the **decision** to reuse the first stage hinges on a critical question: ***Will the first stage successfully land after launch?*** This question is not only pivotal for **operational success** but also for determining the **cost** of a launch. **Predicting** the **landing outcome** of the Falcon 9 first stage is the **focal point** of our analysis.
- By harnessing the power of **data science** and **machine learning**, this project aims to **forecast the probability of a successful landing** for the Falcon 9 first stage. The predictions derived from this analysis will serve for **estimating launch costs**. Such **insights** hold significant value for not only SpaceX but also **competing entities** that seek to bid against SpaceX for rocket launches.
- We will explore the **methodologies** employed to **collect** and **preprocess** data, delve into **exploratory data analysis** to uncover **trends** and **patterns**, and detail the construction of **predictive models** that offer actionable **insights** into the likelihood of successful first stage landings. Ultimately, our endeavor seeks to contribute to the continued transformation of space travel by fostering a deeper **understanding** of the **factors** that drive the **cost-effective** and groundbreaking launches epitomized by **SpaceX's Falcon 9 rocket**.

Section 1

Methodology

Methodology

1. Data Collection and Wrangling: The data **collection** process combines SpaceX's **REST API** and **web scraping** techniques to amass a comprehensive dataset. This includes **historical launch information** and crucial features. The **dataset** is then subjected to **wrangling processes**, such as one-hot encoding for **categorical** features, **filtering**, and handling **missing values**. The result is a **structured** and **cleaned** dataset ready for analysis.

2. Exploratory Data Analysis (EDA) and Visualization: EDA plays a pivotal role in understanding the data's **characteristics** and **relationships**. Through **SQL queries** and **interactive visualizations**, the analysis uncovers insights related to launch sites, payload masses, booster versions, mission outcomes, and more. **Matplotlib**, **Seaborn**, and visualization libraries like **Folium** and **Plotly Dash** facilitate the representation of **trends** and **patterns**, enhancing our understanding of the data.

3. Interactive Visual Analytics: Folium and **Plotly Dash** are employed to create **interactive maps** and **dashboards**, respectively. These tools enable stakeholders to explore and manipulate data **dynamically**. The interactive map provides **insights** into launch site **locations** and **proximities**, while the dashboard allows for **real-time exploration** of flight records, success rates, payload masses, and landing outcomes.

4. Predictive Analysis using Classification Models: The predictive analysis phase centers on building and evaluating **classification models** to forecast the landing outcomes of Falcon 9 first stages. Multiple models, including Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors Classifier, are trained and evaluated. **Hyperparameter** tuning using **GridSearchCV** optimizes model performance, ultimately allowing for informed predictions.

In essence, the methodologies seamlessly integrate various techniques from data collection to predictive analysis. The collective efforts enable a comprehensive **understanding** of the **factors** influencing first stage landings, culminating in **actionable insights** that can drive **cost efficiency** and competitive strategies in the dynamic space travel industry. This analysis stands as a testament to the **power** of **data science** and its potential to shape the future of space exploration.

Data Collection

- Data collection refers to the process of **gathering**, **acquiring**, and **procuring** raw information or data points from various sources.
- Data collection encompasses several key aspects:

- **Source Identification**: Data scientists need to identify the sources from which they will collect data. These sources can include databases, APIs, websites…

- **Data Gathering**: Once sources are identified, data is collected using appropriate methods, such as web scraping, API requests, data extraction from files, or manual entry. The goal is to retrieve the relevant information needed for analysis.

- **Data Validation and Cleaning**: Collected data may contain errors, inconsistencies, or missing values. Data validation involves checking for correctness and reliability, while data cleaning involves addressing issues such as duplicates, outliers, and missing values to ensure the data's quality.

- **Data Transformation**: Data may need to be transformed into a suitable format for analysis. This can involve tasks like reshaping data, converting data types, creating new features, and normalizing or standardizing variables.

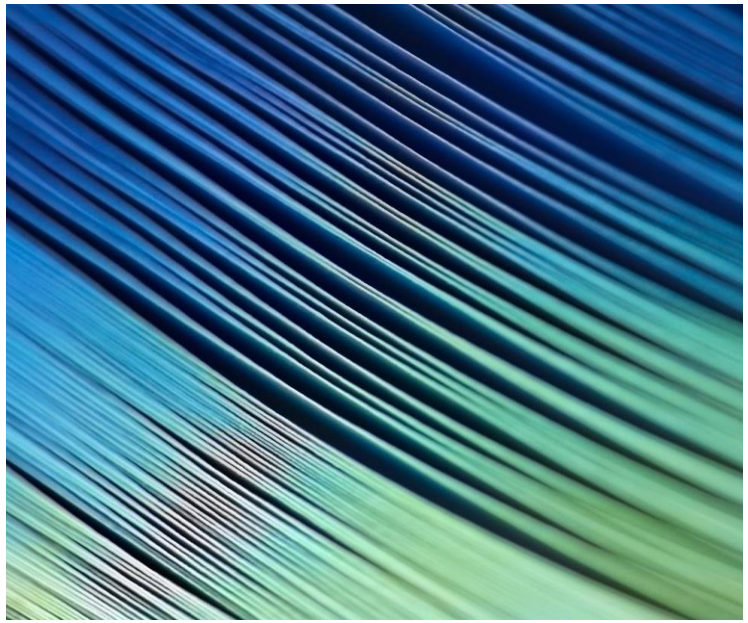
Data Collection (REST API)

Request	Request data from SpaceX API (rocket launch data)
Request	Request information about the launches from SpaceX API using custom functions
Decode	Decode response using . json () and convert to a dataframe using . json_normalize
Create	Create dictionary from the data
Create	Create dataframe from the dictionary
Filter	Filter dataframe to contain only Falcon 9 launches
Replace	Replace missing values of Payload Mass with calculated .mean()
Export	Export data to csv file

The information obtained by the API are :

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

SpaceX REST API URL



Data Collection (Web Scraping)

Request	Request data (Falcon 9 launch data) from Wikipedia
Create	Create BeautifulSoup object from HTML response
Extract	Extract column names from HTML table header
Collect	Collect data from parsing HTML tables
Create	Create dictionary from the data
Create	Create dataframe from the dictionary
Export	Export data to csv file

The information obtained by the webscrapping of Wikipedia are :

Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Wrangling

At this stage, an **initial Exploratory Data Analysis (EDA)** is carried out on certain data, along with the determination of the **data labels** that will be used for **supervised** model training.

1. **Analyze:**
    - **Calculate Launches for Each Site:** The dataset is analyzed to calculate the **number of launches** that occurred at each **launch site**. This information provides an overview of **launch activity** at different **locations**.
    - **Analyze Orbits:** The data is examined to calculate the **number** of occurrences for each **orbit type**. This analysis helps understand the distribution of missions based on their designated orbits.
    - **Evaluate Mission Outcomes:** The occurrences and outcomes of **missions** are assessed based on their **outcome types**. This step involves calculating the number and occurrence of **successful** and **unsuccessful** missions for each **outcome** type.
2. **Create Landing Outcome Label:** A new **categorical variable**, labeled ***Class***, is generated to represent the landing outcomes of the missions. This **label** is assigned based on the mission Outcome (*True*/*False*) and landing location (code) components of the ***Outcome*** column. **Successful landings** are represented by a value of **1**, and **unsuccessful landings** are represented by a value of **0**. Different scenarios such as ***True ASDS***, ***True RTLS***, and ***True Ocean*** indicate **successful** landings and are assigned the value **1**. Conversely, scenarios like ***False ASDS***, ***False RTLS***, and ***False Ocean*** signify **unsuccessful** landings and are assigned the value **0**.
3. **Export to CSV:** After the necessary data transformations and label creation, the processed data is exported to a CSV file. This organized dataset serves as the foundation for subsequent exploratory data analysis (EDA), visualization, and machine learning tasks.

EDA with Data Visualization

- This stage involves a comprehensive exploration of the dataset using various visualization techniques to uncover meaningful patterns, relationships, and trends. This phase leverages the power of visual representations to gain insights into the factors influencing the success of landing outcomes. The following steps were undertaken as part of this phase:

- **Data Loading and Preparation:** The first step was to load the dataset into a Pandas DataFrame, a tabular data structure that facilitates data manipulation and analysis. This provided a structured foundation for further visualization.

- **Scatter Plots:** Scatter plots were created to visualize the relationships between different attributes. Specifically, scatter plots were generated for Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Orbit vs. Payload Mass. Scatter plots are powerful tools for depicting correlations and dependencies between variables, helping to identify any trends or patterns.

- **Bar Graphs:** Bar graphs were employed to compare discrete categories and highlight relationships between specific categories and measured values. The success rate was plotted against Orbit Type, revealing insights into which orbits had the highest probability of success. Bar graphs are instrumental in showcasing categorical relationships in a clear and intuitive manner.

- **Line Graphs:** Line graphs were utilized to illustrate trends in data over time, specifically, the success rate over the years. This time series analysis provides a visualization of the launch success yearly trend, helping to identify any consistent patterns or changes in success rates over time.

- **Feature Engineering:** After gaining insights from scatter plots, bar graphs, and line graphs, feature engineering was performed to enhance the dataset for future predictive modeling. Dummy variables were created for categorical columns, which would be useful for predicting success outcomes in subsequent modules.

- By visualizing the relationships between different attributes, the study can uncover valuable insights that inform subsequent analysis, predictive modeling, and decision-making processes.

EDA with SQL

- This phase involves leveraging structured query language (SQL) to perform exploratory data analysis (EDA) on the loaded dataset within an IBM DB2 database. This phase aims to extract meaningful insights and patterns from the data. Several SQL queries are executed to uncover key information:

- **Unique Launch Sites:** Display the distinct launch site names, revealing the different locations from which Falcon 9 rockets are launched.
- **Filtered Launch Sites:** Show specific launch sites where their names start with 'CCA', aiding in isolating launches from particular sites.
- **Payload Analysis:** Calculate the total payload mass carried by boosters launched as part of NASA's CRS program, providing insights into payload distribution.
- **Booster Version Analysis:** Compute the average payload mass carried by the F9 v1.1 booster version, shedding light on payload trends.
- **First Successful Ground Pad Landing:** Identify the date of the initial successful landing on a ground pad, marking a significant milestone.
- **Drone Ship Landings with Payload Range:** List boosters successfully landing on drone ships with a payload mass between 4000 and 6000 kg, showcasing precise landing achievements.
- **Mission Outcome Count:** List the total count of successful and failed mission outcomes, offering an overview of success rates.
- **Max Payload Booster Versions:** Determine booster versions that have carried the maximum payload mass, providing insights into payload capabilities.
- **Failed Landings in 2015:** Detail failed landing outcomes on drone ships, including booster versions and launch sites, for each month in 2015, highlighting temporal patterns.
- **Ranking Landing Outcomes:** Rank the count of landing outcomes (failure or success) between specific dates, aiding in identifying periods of high and low success rates.

Build an Interactive Map with Folium

- This phase involves using the Folium library in Python to create an interactive map. This map visually represents launch sites, launch outcomes, and distances between sites and key locations. Key steps include:
- **Launch Sites and Markers:** Blue circles mark launch sites with popup labels. Red circles represent all sites with labels showing their names, aiding in site distribution understanding.
- **Launch Outcomes:** Markers indicate success (green) or failure (red), grouped with MarkerCluster. This highlights success rates across launch sites.
- **Distance Calculation:** Lines show launch site distances to landmarks like railways, highways, coastlines, and cities, offering insights into proximity relationships.
- **Insights and Analysis:** The interactive map helps interpret spatial patterns, such as launch success near specific features.
- **Geographical Context:** Combining markers, circles, lines, and clusters provides a comprehensive view, enabling to explore spatial aspects and draw informed conclusions.

Build a Dashboard with Plotly Dash

- The "Build a Dashboard with Plotly Dash" phase involves creating an interactive dashboard using the Plotly Dash library. This dashboard enables stakeholders to explore and manipulate data interactively and in real-time. The key steps of this phase are as follows:
- **1. Launch Sites Dropdown List:**

  - Set up a dropdown list to allow users to select a specific launch site or all launch sites.

- **2. Success Visualization with a Pie Chart:**

  - Created a pie chart displaying the percentage of successful and failed launches for the chosen launch site in the dropdown.

  - Used colors to differentiate between successful and failed launches.

- **3. Payload Mass Range Slider:**

  - Added a range slider to enable users to select a payload mass range.

- **4. Payload Mass vs. Success Rate by Booster Version with a Scatter Plot:**

  - Generated a scatter plot to show the correlation between payload mass and launch success rate for different booster versions.

  - Utilized colors to distinguish between various booster versions.

- This interactive dashboard allows users to select specific parameters such as the launch site and payload mass range, and instantly observe updated visualizations based on their selections. It provides an overview of success rates, relationships between payload mass and launch success, and facilitates informed decisions regarding optimal launch sites based on payload requirements.

Predictive Analysis (Classification)

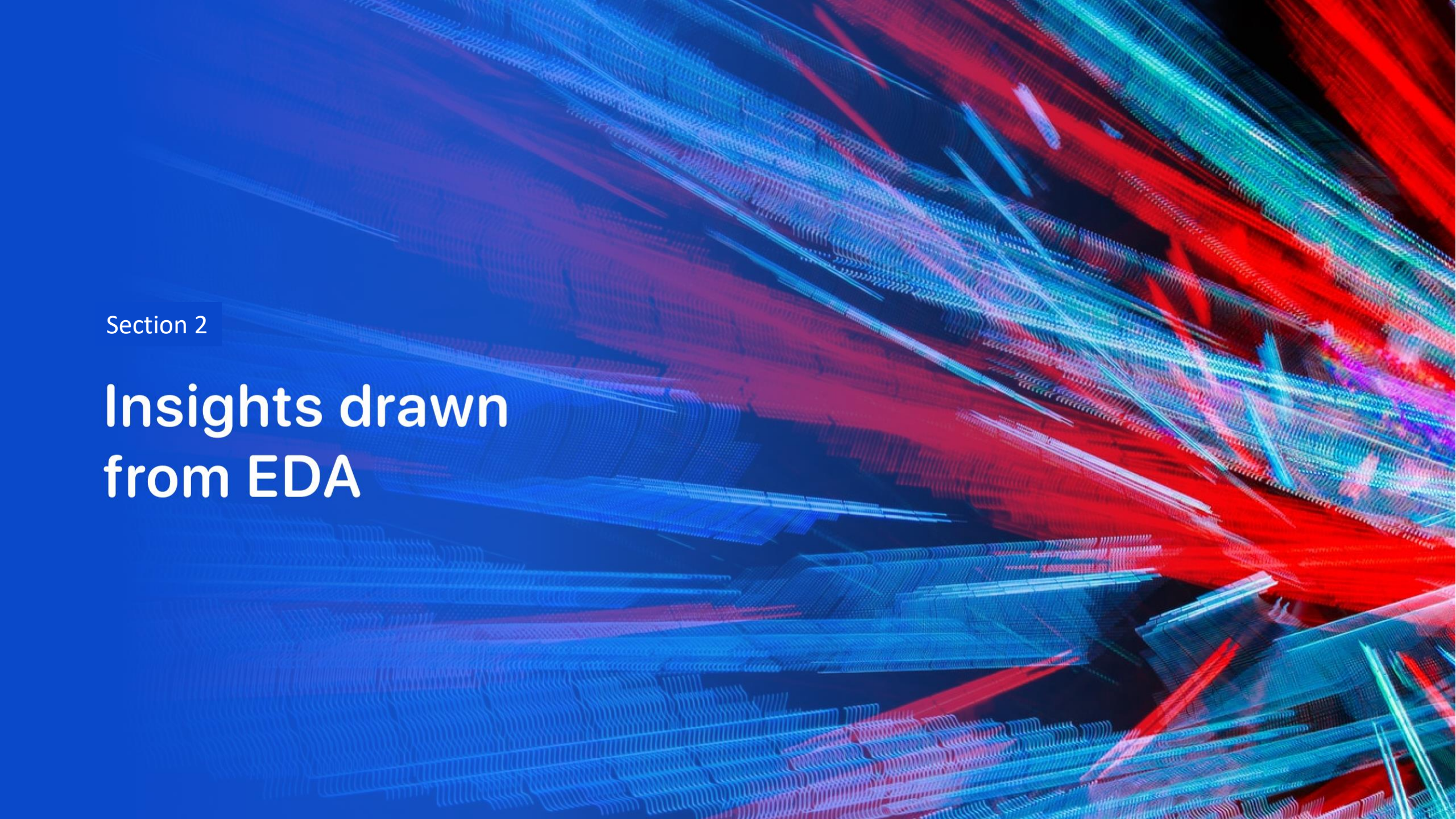
- The "Predictive Analysis (Classification)" phase in the study of SpaceX Falcon 9 launch data involves the following streamlined steps:
- **Data Preparation:**
  - Load the dataset into NumPy and Pandas.
  - Transform and standardize the data.
  - Split the data into training and test sets.
- **Model Selection and Training:**
  - Choose machine learning algorithms, such as Logistic Regression, SVM, Decision Tree, and KNN.
  - Set algorithm parameters using GridSearchCV.
  - Fit models to the training dataset.
- **Model Evaluation:**
  - Calculate accuracy scores for each model on the test data.
  - Evaluate confusion matrices to assess performance.
- **Model Comparison and Selection:**
  - Compare models based on accuracy, Jaccard score, and F1 score.
  - Choose the best-performing classification model.
- In this phase, predictive models are built and trained using the prepared data. The models are evaluated, compared, and the best model is selected based on its accuracy and other performance metrics. This process allows for informed decision-making in predicting launch outcomes for SpaceX Falcon 9 missions.

Results

EXPLORATORY DATA ANALYSIS RESULTS

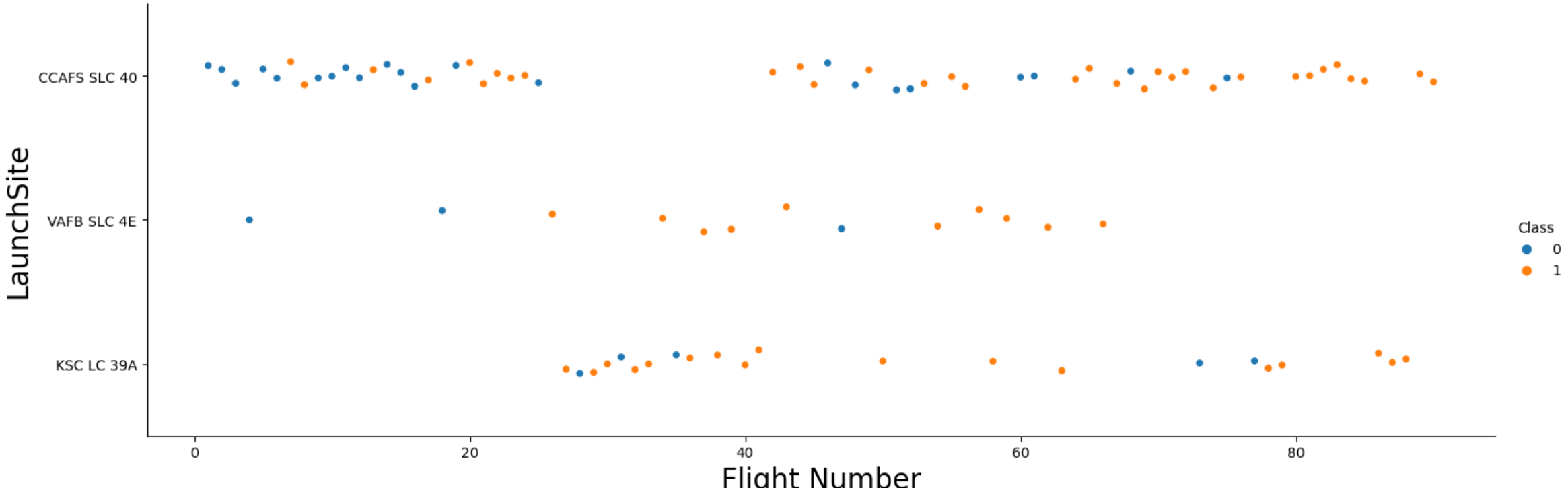
INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS


Section 2
Insights drawn from EDA

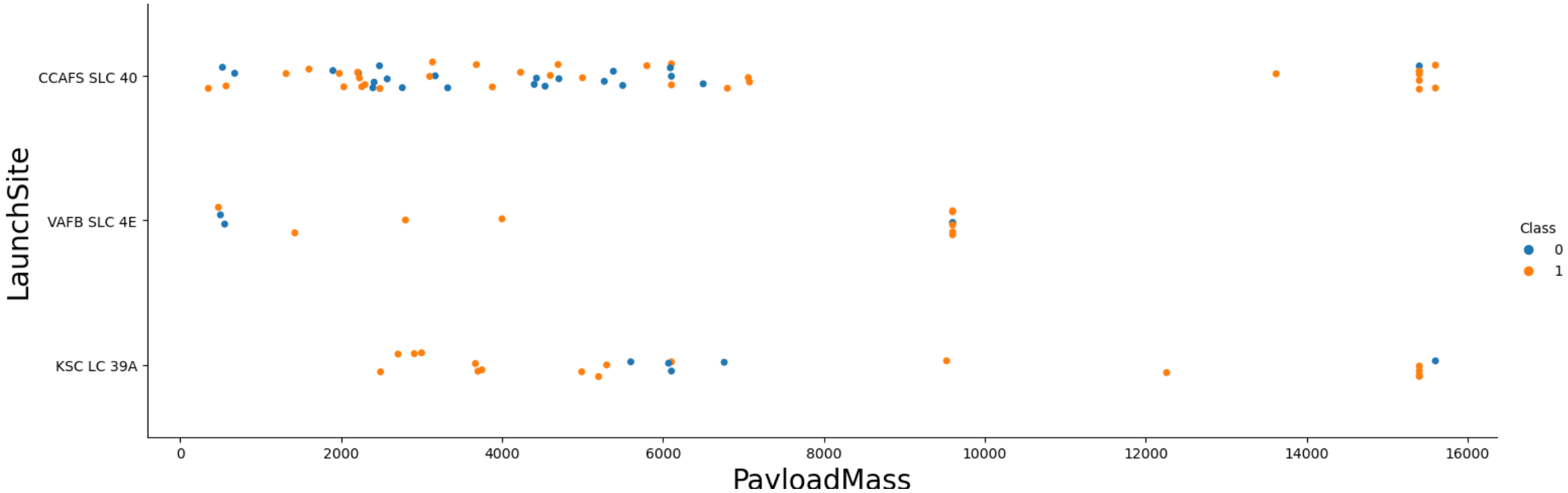
Flight Number vs. Launch Site

- The general trend is an increase in success rate with time and number of launches.
- The **CCAFS SLC-40** is the most used, with a success rate that really increases in recent launches, while this trend has been earlier for other sites.



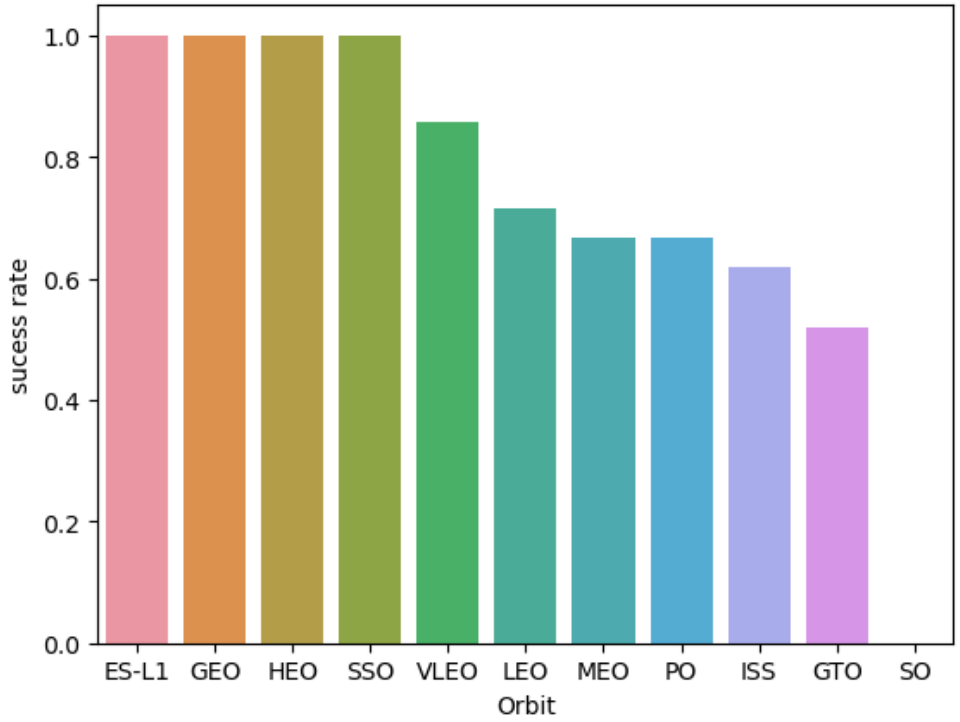
Payload vs. Launch Site

- Most launches have a payload < 7000 kg. Only CCAFS SLC 40 and KSC LC 39A perform payloads > 15000 kg.
- Paradoxically, the success rate increases with payload. But it seems to be more of a correlation than a causation. Indeed, the high payloads were sent during the last missions, by nature more reliable, The factor is rather temporal.



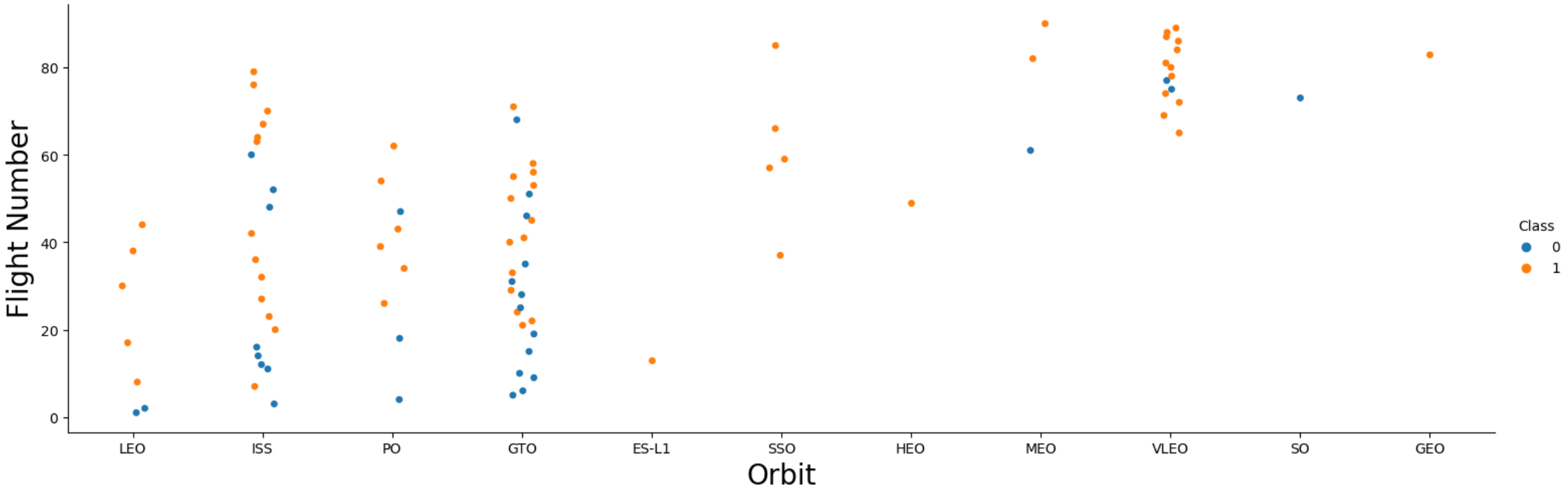
Success Rate vs. Orbit Type

Orbit ES-L1, GEO, HEO, SSO have the highest success rate (100%). VLEO orbit, has a rate > 85%



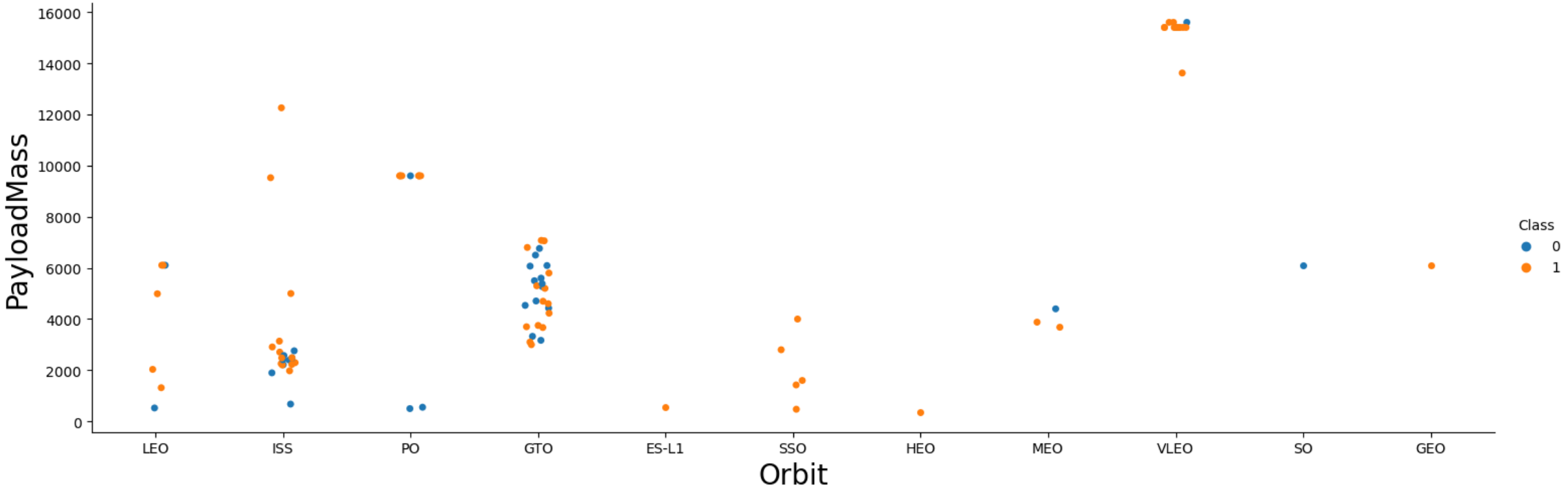
Flight Number vs. Orbit Type

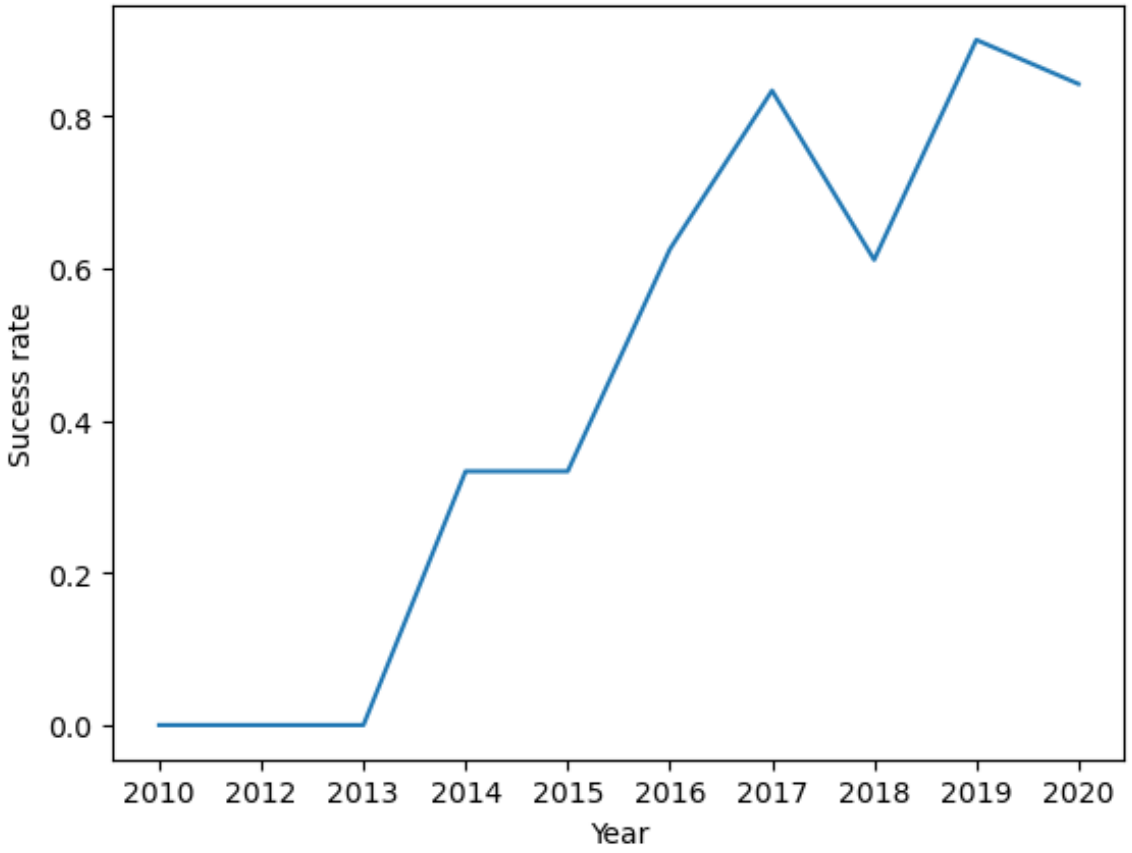
- Again, the success rate increases with time and the number of flights.
- However, the low number of launches in some orbits does not allow comparisons to be made on this point. It is noted that the VELEO orbit is the most frequent for recent missions.
- The ISS and GTO orbits have been the most regularly used since the beginning.



Payload vs. Orbit Type

- The orbits comprising the majority of launches (ISS and GTO) concern virtually non-overlapping payload ranges (2000-3500 and 3500-8000), indicating a certain specialization. Just like the VLEO orbit which seems reserved for the heaviest launches (> 15000).
- Other orbits have too few launches to provide meaningful data.
- As we have seen before, the correlation between the payload and the success rate is more temporal.





Launch Success Yearly Trend

- The trend is towards a high growth in the success rate since 2013.
- The decline in 2018 seems to be due to a lack of information (outcomes to "None") rather than failures.
- The level reached is now approaching 100%.

All Launch Site Names

This query retrieves unique launch site names from the SPACEXTABLE table. The DISTINCT keyword ensures that only distinct (non-duplicate) launch site names are returned. This query is helpful for obtaining a concise list of all the different launch sites used for SpaceX Falcon 9 launches.

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Launch Site Names Begin with 'CCA'

The query retrieves all data columns for the first 5 rows (**`LIMIT 5`**) of the SPACEXTABLE where Launch_Site starts with the letters 'CCA' (**`LIKE 'CCA%'`**), providing information on the first five SpaceX Falcon 9 launches that took place in the Cap Canaveral area.

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 query calculates the total (**SUM**) payload mass in kilograms for all SpaceX Falcon 9 launches where the customer is 'NASA (CRS)'. The result will be a single value representing the sum of payload masses for these launches.

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

The query calculates the average (**AVG**) payload mass in kilograms for all SpaceX Falcon 9 launches with booster versions starting (**LIKE**) with 'F9 v1.1'. The result will be a single value representing the average payload mass for these launches.

```
SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

Avg_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

This query retrieves the earliest date (**`MIN(Date)`**) of a successful ground pad landing (**`Landing_Outcome = 'Success (ground pad)'`**) from the SPACEXTABLE table containing SpaceX Falcon 9 launch data.

The result will show the first successful ground pad landing date.

```sql
SELECT MIN(Date) AS first_success
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The query retrieves a list of distinct booster versions for SpaceX Falcon 9 launches where the landing outcome is a successful landing on a drone ship (**`Success (drone ship)`**) and the payload mass is between 4000 and 6000 kilograms. The result will be a list of unique booster versions that meet these criteria.

```sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ > 4000
  AND PAYLOAD_MASS__KG_ < 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

The query retrieves a list of unique mission outcomes along with the count of occurrences for each (**GROUP BY**) outcome from the SPACEXTABLE table. This can provide valuable insights into the distribution of mission outcomes for SpaceX Falcon 9 launches, helping to analyze the success and failure rates of the missions.

```
SELECT Mission_Outcome, COUNT(*) AS Nb
FROM SPACEXTABLE
GROUP BY TRIM(Mission_Outcome);
```

Mission_Outcome	Nb
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The query retrieves the `Booster_Version` and `PAYLOAD_MASS__KG_` columns from the table where the payload mass matches the maximum (**MAX**) payload mass recorded in the table. This maximum is the result of a sub-query.

It helps identify the booster version(s) associated with the highest payload mass.

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

```
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

2015 Launch Records

```sql
SELECT
    CASE SUBSTR(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
        ELSE 'Invalid Month'
    END AS Month_2015,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(Date, 1, 4) = '2015'
    AND Landing_Outcome = 'Failure (drone ship)'
```

This query retrieves SpaceX Falcon 9 launch data for the year 2015 with a specific focus on failed landings on drone ships. It includes a calculated Month_2015 column displaying month names based on the month numbers from the Date column.

The result provides insights into landing outcomes, booster versions, and launch sites for failed drone ship landings specifically in the year 2015.

Month_2015	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

This SQL query retrieves the count of distinct landing outcomes for SpaceX Falcon 9 launches between June 4, 2010, and March 20, 2017.

The date parts are concatenated using the || operator and SUBSTR to form a single integer value representing the date.

The data is then grouped by landing outcome, the count of occurrences of each outcome is calculated, and the results are presented in descending order of the count (Nb).

This analysis helps understand the distribution of different landing outcomes during the specified period.

```
SELECT Landing_Outcome, COUNT(*) AS Nb
FROM SPACEXTABLE
WHERE CAST(
    SUBSTR(Date, 1, 4)
    || SUBSTR(Date, 6, 2)
    || SUBSTR(Date, 9, 2) AS INTEGER
) BETWEEN 20100604 AND 20170320
GROUP BY Landing_Outcome
ORDER BY Nb DESC;
```

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

Section 3

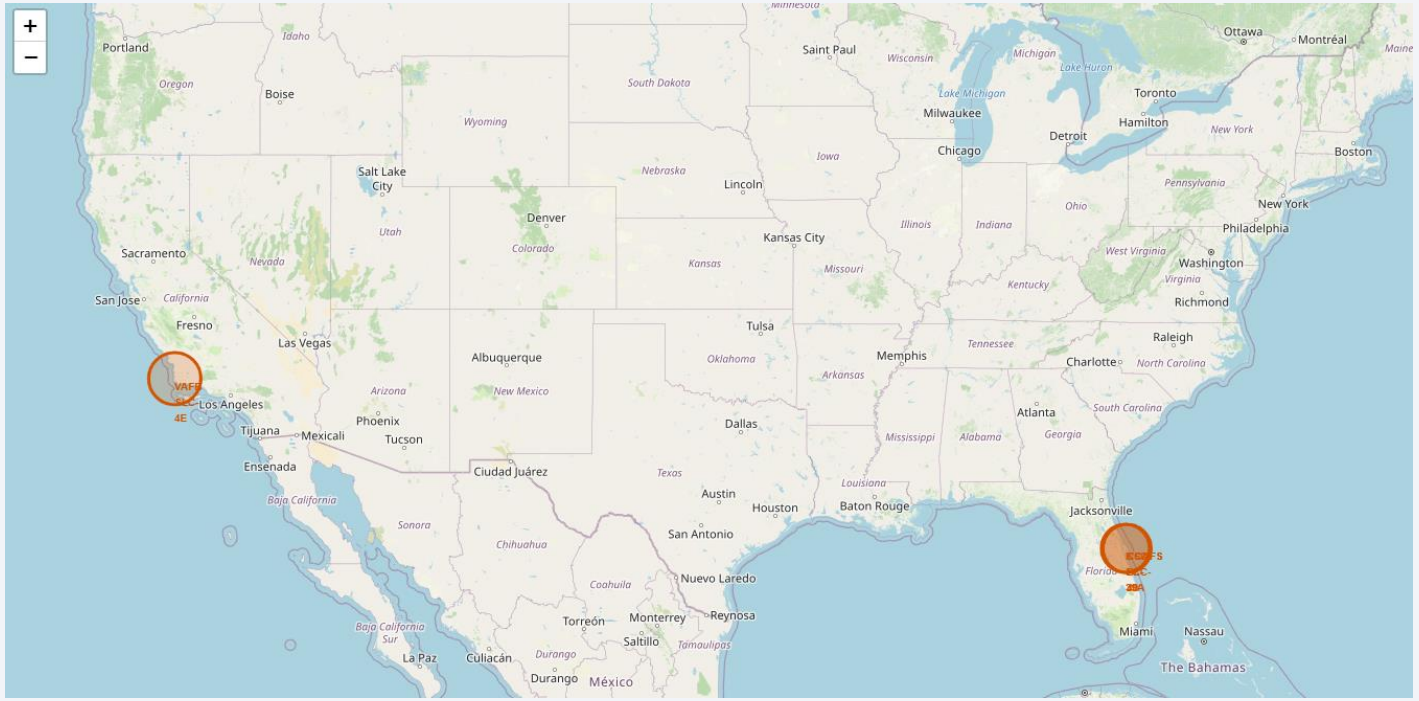
Launch Sites Proximities Analysis

SpaceX Falcon 9 launch sites location

The launch sites for SpaceX's Falcon 9 rocket are located at several key locations:

- Cap Canaveral Space Launch Complex (CCAFS) - Florida, USA
- Kennedy Space Center Launch Complex (KSC LC) - Florida, USA
- Vandenberg Space Force Base (VSFB) - California, USA

At this scale we can observe that launch sites for SpaceX's Falcon 9 rockets are strategically located near the equator and the sea for optimal mission performance and safety.



Near the Equator:

- Launching near the equator utilizes Earth's rotational speed, boosting launch velocity and payload capacity

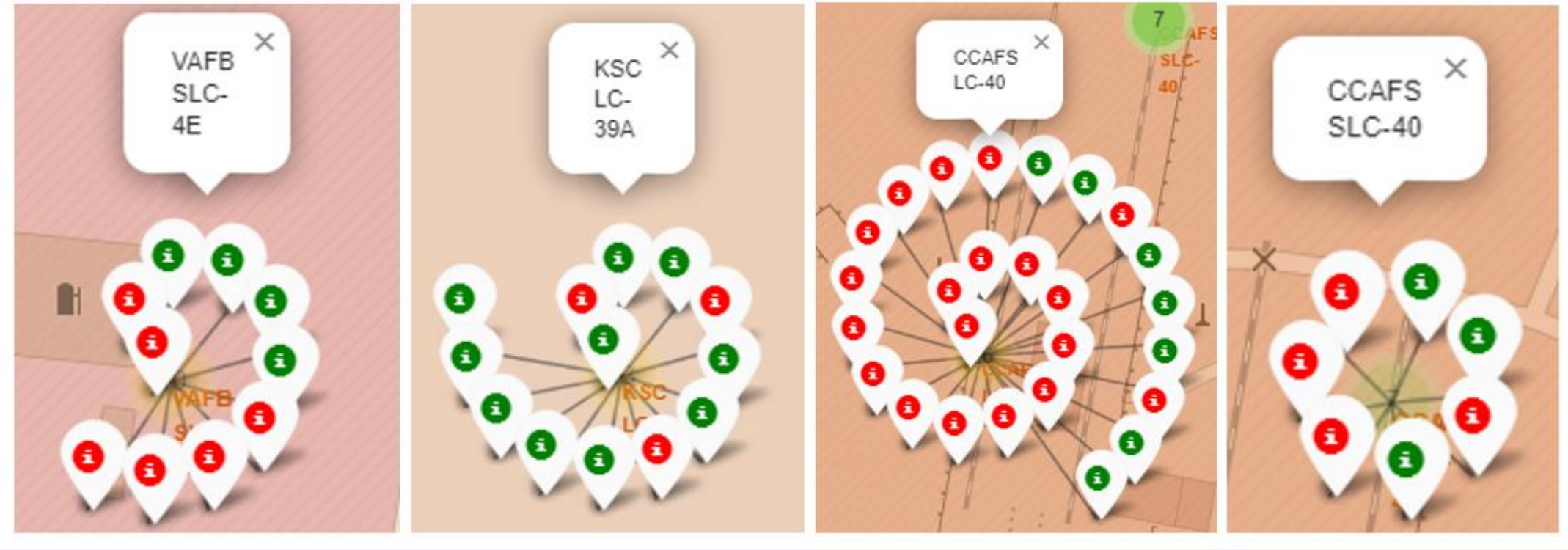
Near the Sea :

- Safe zones for rocket stages during different flight phases.
- Reusable boosters can land on drone ships, cutting costs.
- logistical advantages (transportation access).

SpaceX Falcon 9 launch sites outcomes

Outcomes: successful launches: **green** / failed launches: **red**

- **VABF SLC-4E**: 4/10 (success rate **40%**)
- **KSC LC-39A**: 10/13 (success rate **76.92%**) => best
- **CCAFS LC-40** : 26/7 (success rate **26.92%**)
- **CCAFS SLC-40**: 3/7 (success rate **42.86 %**)

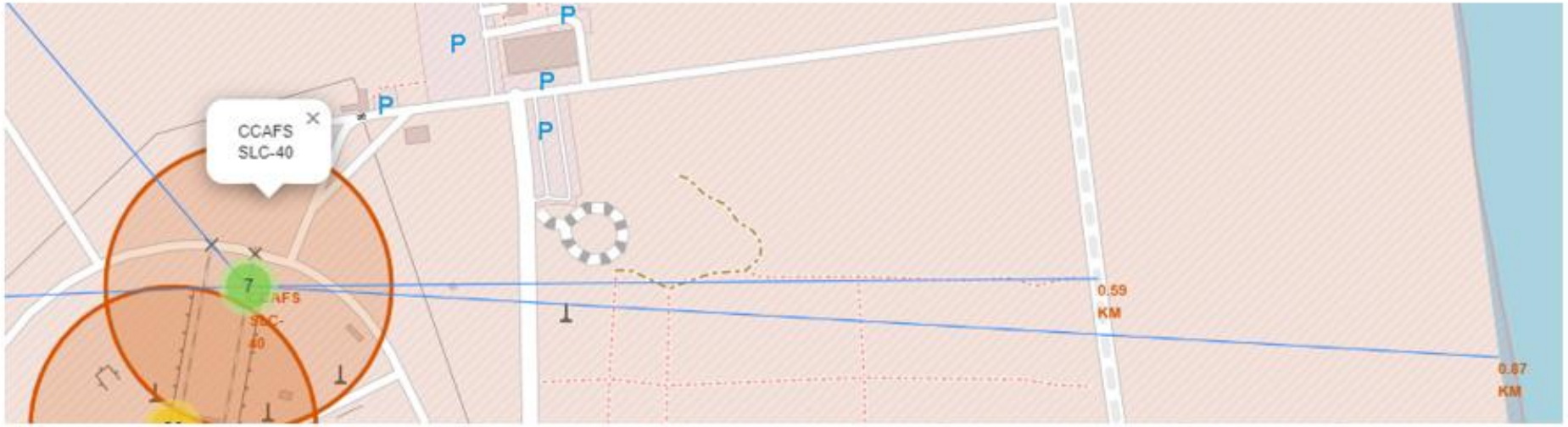


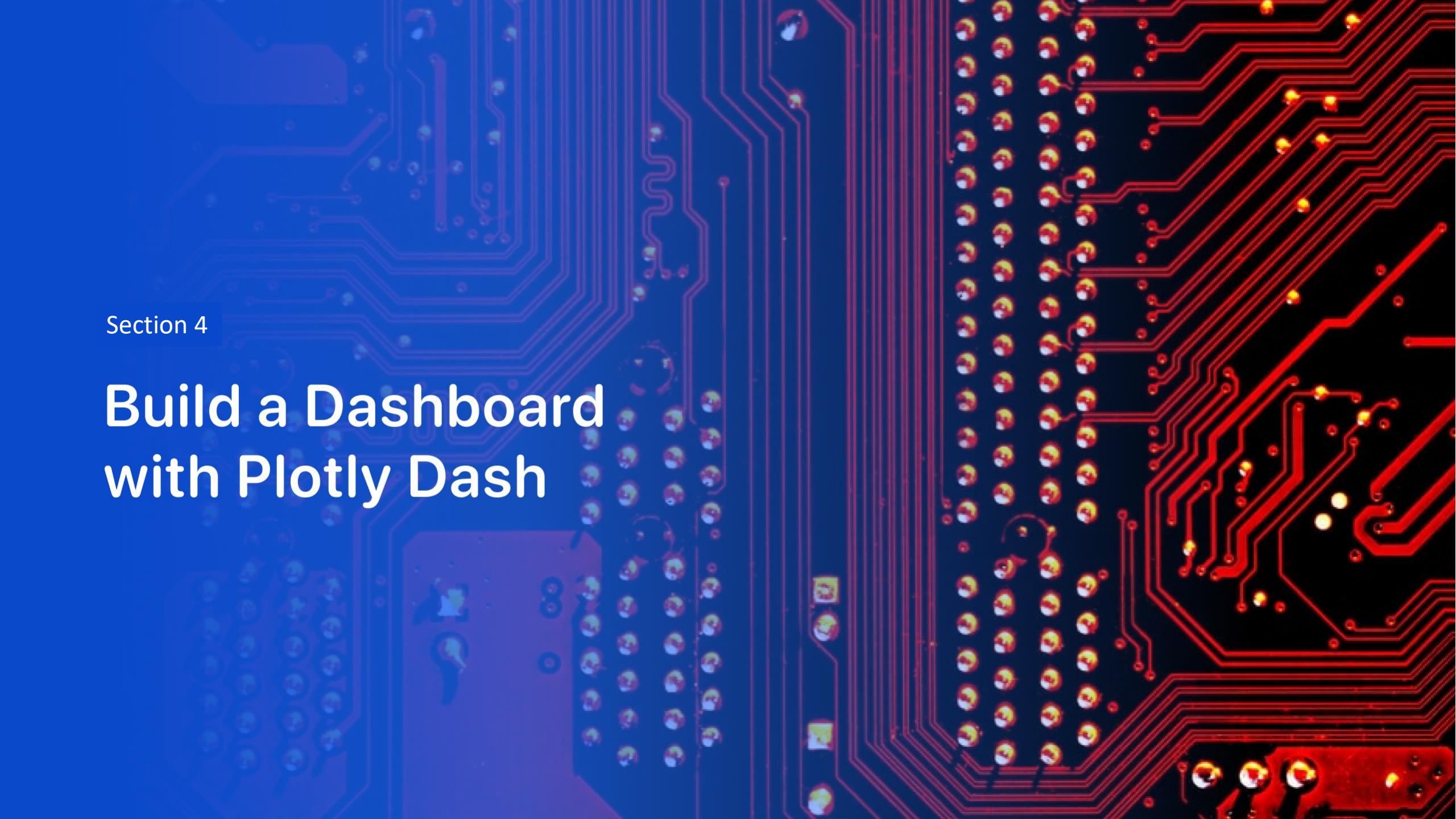
Distances between CCAFS SLC 40 and its proximities

railways (**1.28** km), **highways** (**0.59** km), **coastline** (**0.87** km), **cities** (**78,66** km)

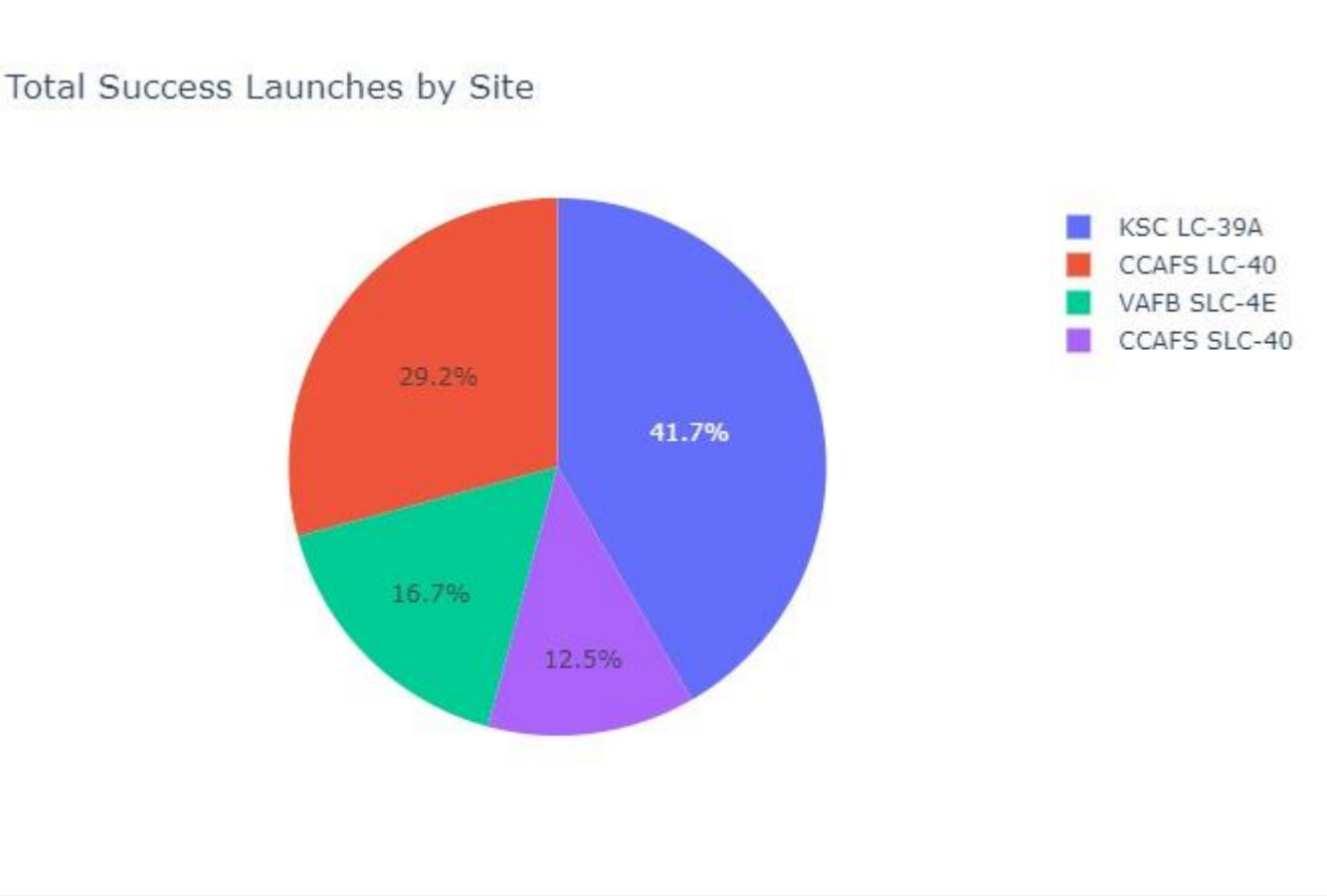
Locally, the distance between **CCAFS SLC-40** and nearby geographic landmarks reveals important siting considerations (shared by the other sites):

- **Safety Zones**: Launch sites are situated in areas with minimal population density and structures to ensure public safety during rocket launches and potential failures.
- **Accessibility**: Proximity to transportation networks, such as highways, railways, and ports, allows for efficient transportation of rocket components and fuel.




Section 4
Build a Dashboard with Plotly Dash

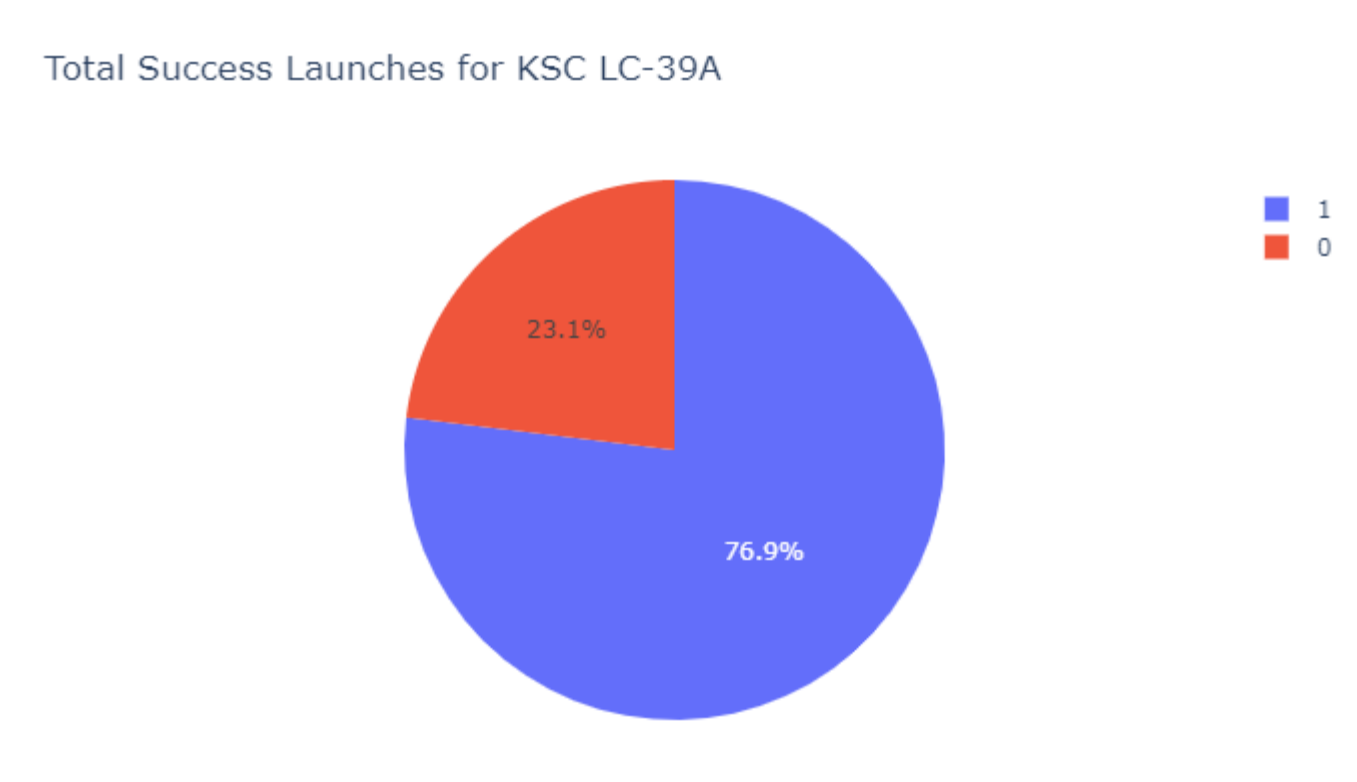
Launch success rate for all sites



KSC LC-39A was the site with the highest number of successful launches 41.7% of the total.

The following slide will show that this is not due to the overall number of launches from this site.

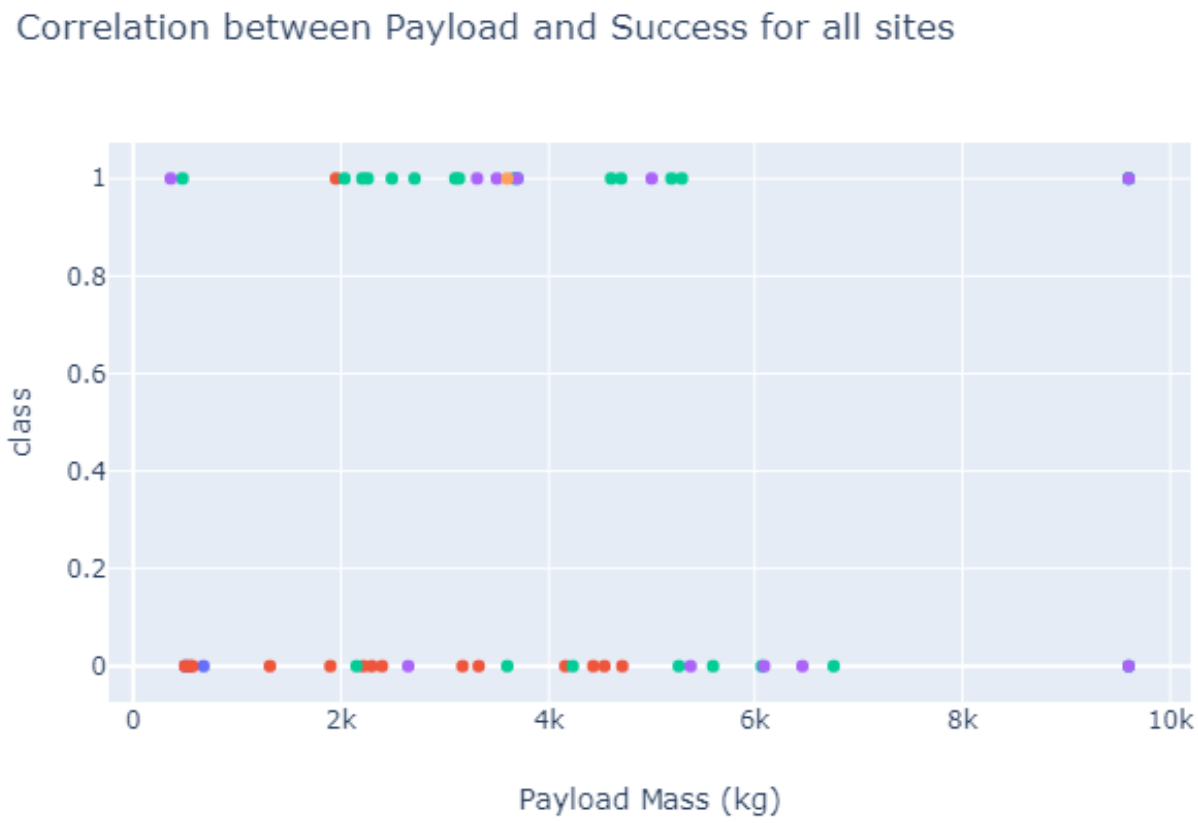
Success Rate for KSC LC-39A

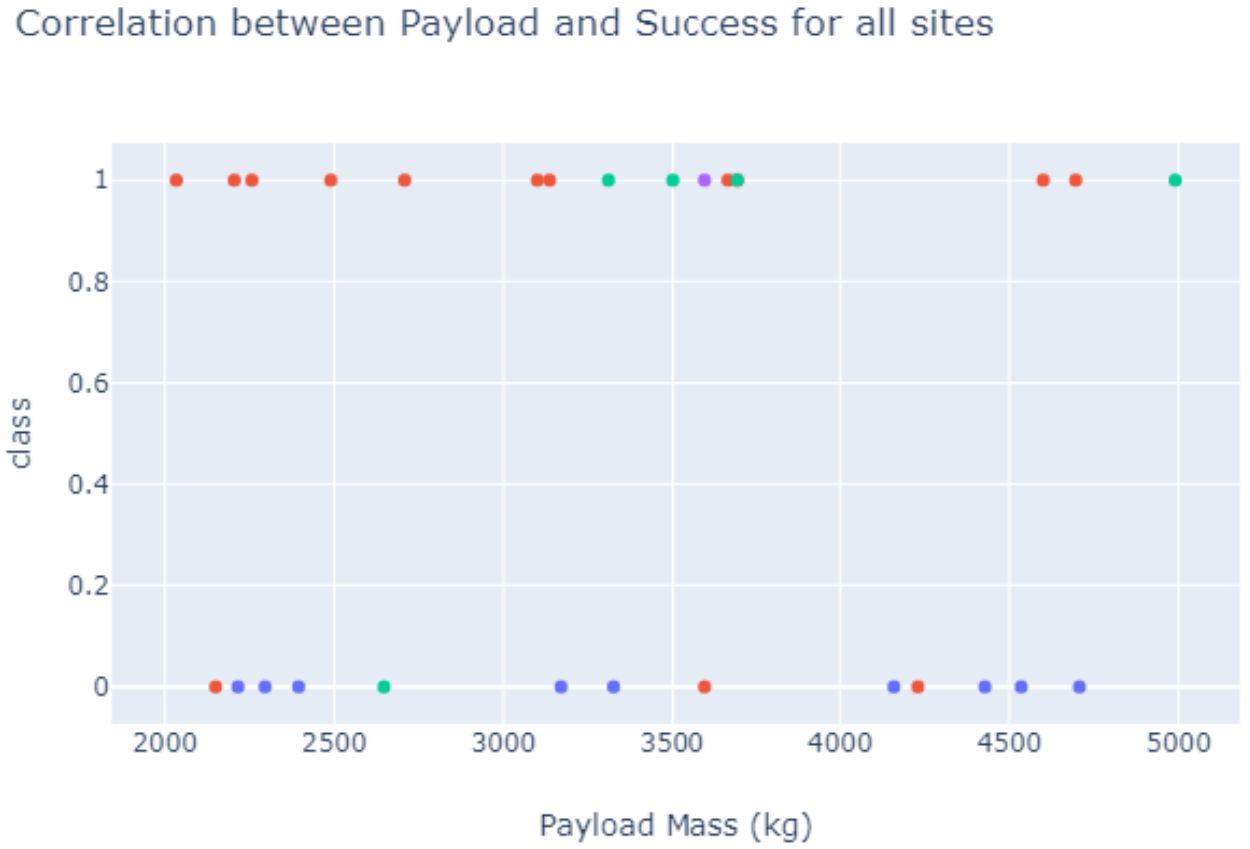


Here we have confirmation that the KSC LC-39A site has the highest number of successful launches, primarily because of a high success rate (76,9%).

Payload vs. Launch Outcome (all sites)

- Payloads <5000 kg have the best success rate.
- Overall, and particularly for payloads between 2000 and 5000 kg, the FT booster has the best success rate.



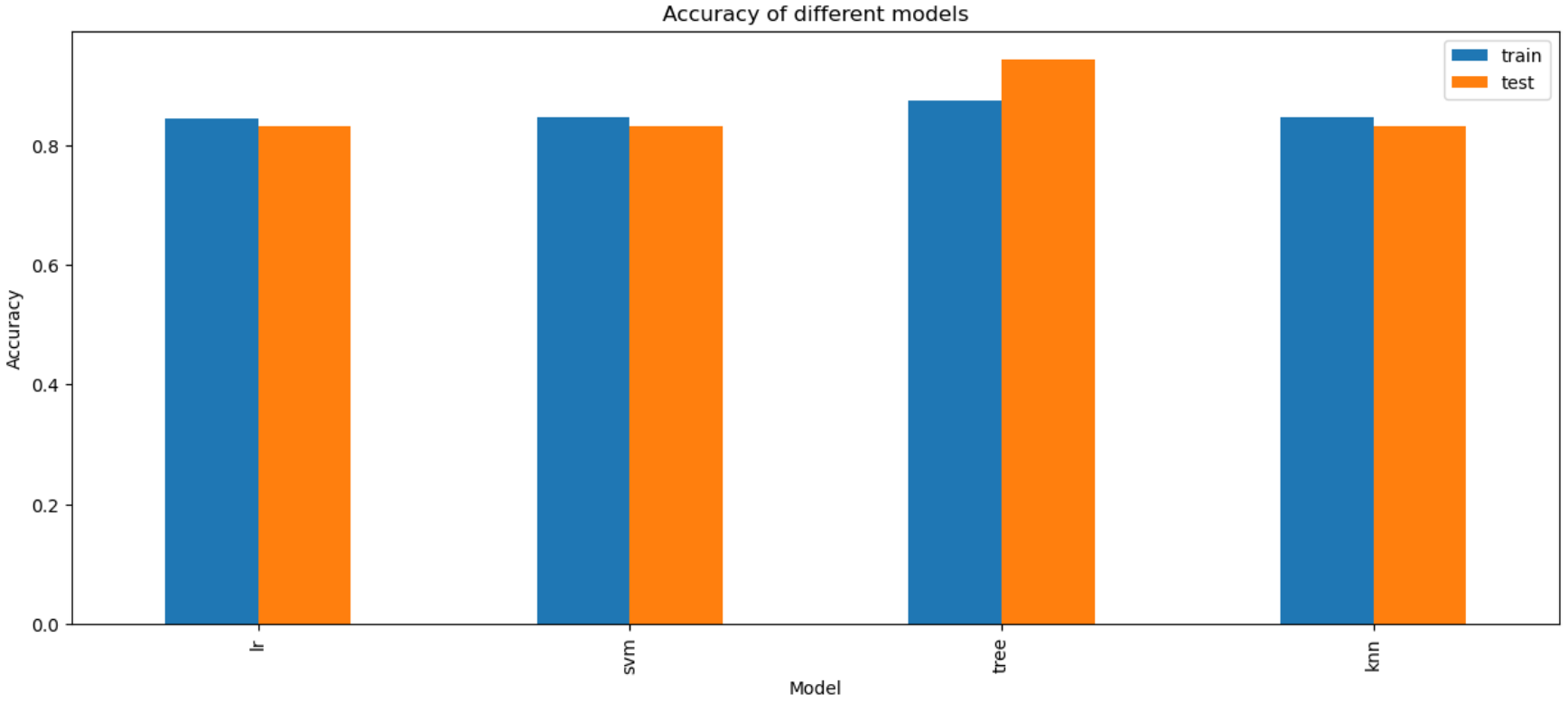


Section 5

Predictive Analysis (Classification)

Classification Accuracy

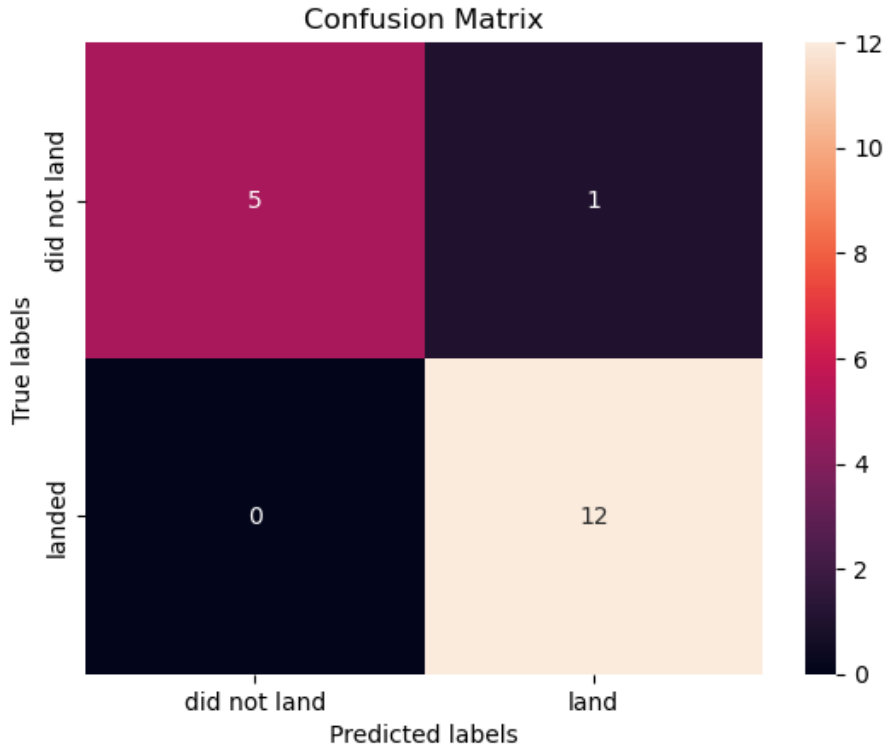
- The different models have similar classification accuracy.
- However, the decision tree performs slightly better with certain values of the **random_state** parameter (here the famous **42**).



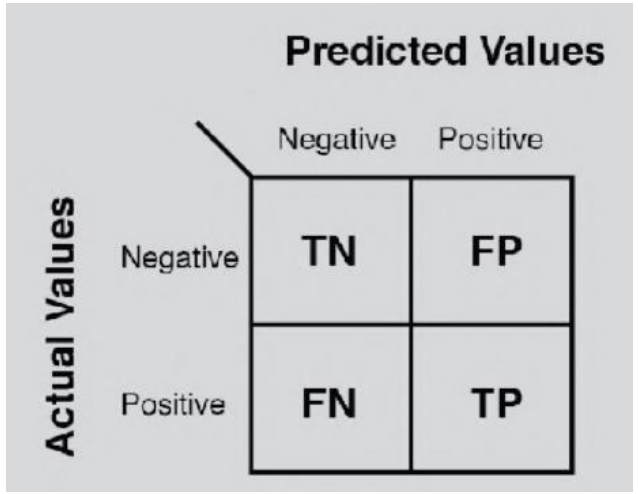
	LR	SVM	Tree	KNN
Jaccard	0.800000	0.800000	0.923077	0.800000
F1	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

	train	test
lr	0.846429	0.833333
svm	0.848214	0.833333
tree	0.885714	0.888889
knn	0.848214	0.833333

Confusion Matrix (decision tree)

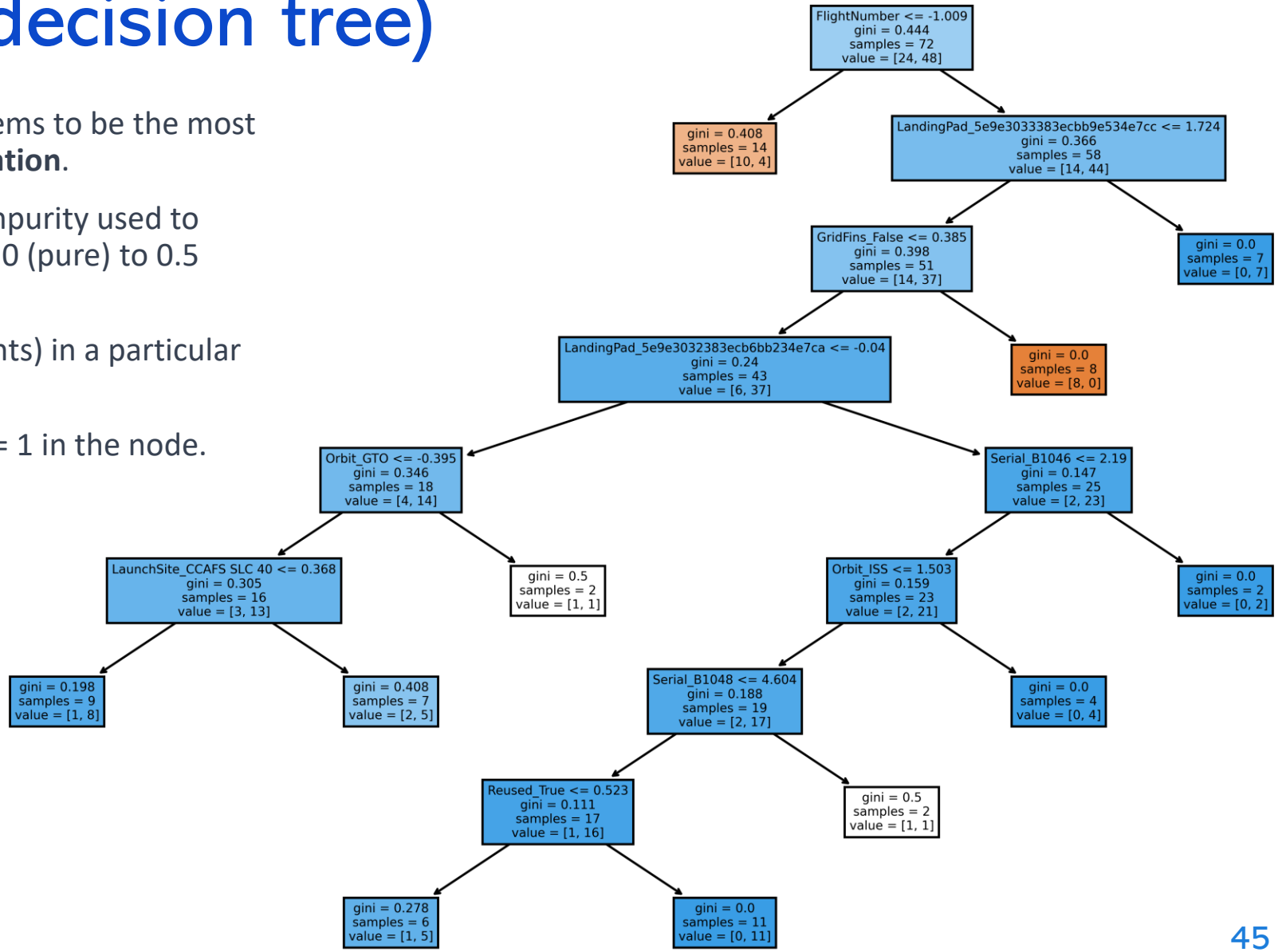


- **True Positives (TP):** positive instances that were correctly predicted by the model.
- **True Negatives (TN):** negative instances that were correctly predicted by the model.
- **False Positives (FP):** negative instances incorrectly classified as positive by the model.
- **False Negatives (FN):** positive instances incorrectly classified as negative by the model.
- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- The issue is the **false positives (FP)** : unsuccessful landings are classed as successful landings.



Visualization (decision tree)

- The **decision tree** being the one that seems to be the most efficient, here is its **graphical representation**.
- **gini** : Gini impurity, a measure of data impurity used to determine optimal splits. It ranges from 0 (pure) to 0.5 (mixed).
- **samples**: number of instances (data points) in a particular node.
- **value**: Split between class = 0 and class = 1 in the node.



Conclusions

- The analysis of SpaceX Falcon 9 launch data reveals key insights:

- Decision Tree Classifier is the best ML model with some parameters.
- A lower payload leads to higher success rates. The success of recent lauches with heavy payloads seems to be related to a temporality factor (overall growth in reliability)
- SpaceX's success rate improved since 2013.
- KSC LC-39A is the most successful launch site. It shares with other sites geographical characteristics related to logistics or security, the point directly impacting the success of the missions being the position close to the equator.
- Specific orbits like ES-L1, GEO, HEO, SSO are 100% successful, but the small number of some flights prevents reliable conclusions from being drawn.
- Interactive dashboard aids visualization.
- Predicting the probability of successful landings reduces costs.
- Needs: Expand the data set, refine the selection of essential attributes.

Appendix

Notebooks:

- Data Collection (API)
- Data Collection (webscraping)
- Data wrangling
- EDA with Visualization
- EDA with SQL
- Interactive Maps Analytics with Folium
- Interactive Dashboard
- Machine Learning Prediction

Thank you!