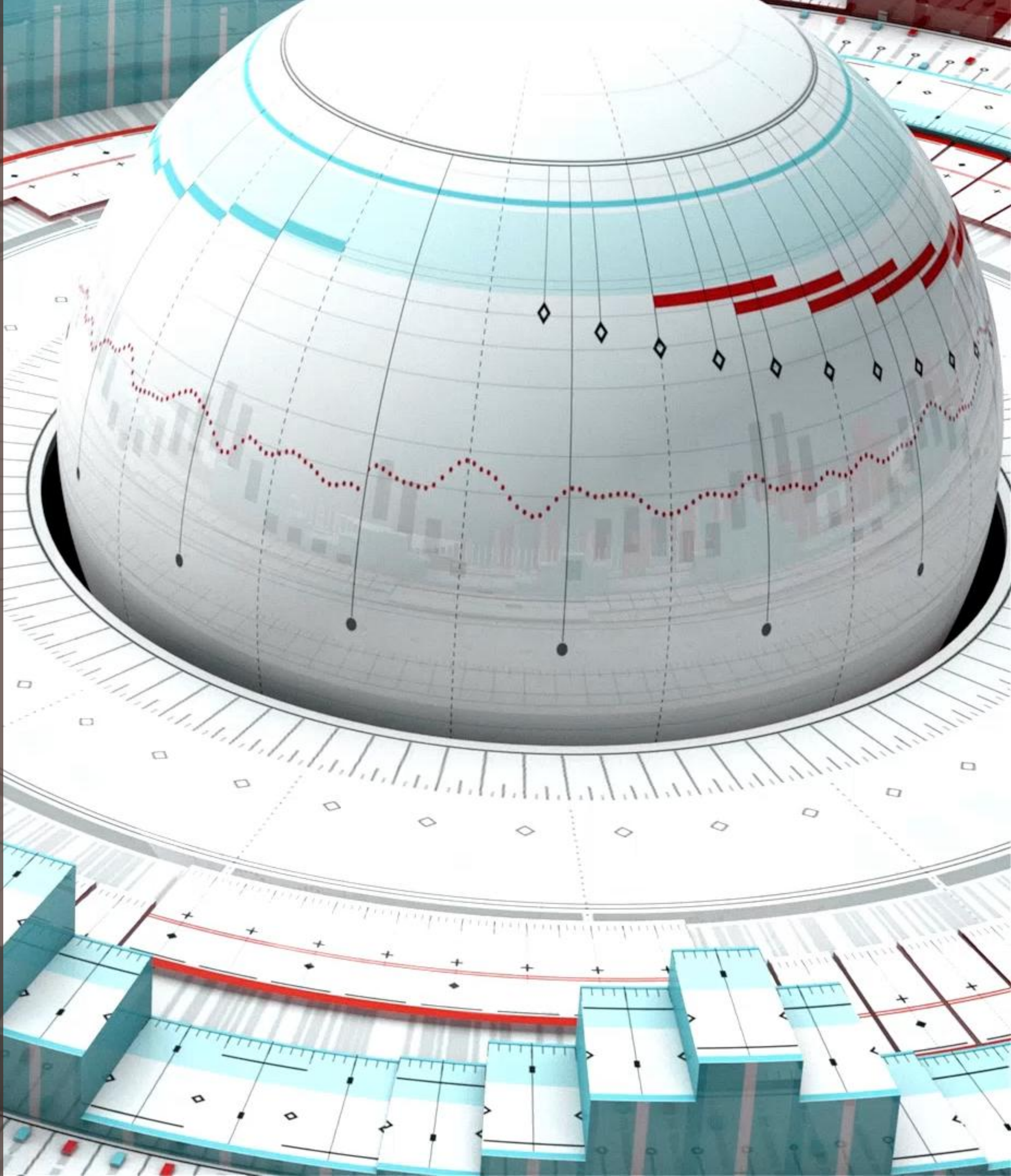
SpaceX Data Analysis



Outline

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	acecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
3	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
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
6	2013-09-29	16:00:00	F9 v1.1 B1003	VAFB SLC-4E	CASSIOPE	500	Polar LEO	MDA	Success	Uncontrolled (ocean)
7	2013-12-03	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt
8	2014-01-06	22:06:00	F9 v1.1	CCAFS LC-40	Thaicom 6	3325	GTO	Thaicom	Success	No attempt
9	2014-04-18	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
10	2014-07-14	15:15:00	F9 v1.1	CCAFS LC-40	Orbcomm-OG2 satellites	1316	LEO	Orbcomm	Success	Controlled (ocean)
11	2014-08-05	8:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8	4535	GTO	AsiaSat	Success	No attempt
12	2014-09-07	5:00:00	F9 v1.1 B1011	CCAFS LC-40	AsiaSat 6	4428	GTO	AsiaSat	Success	No attempt
13	2014-09-21	5:52:00	F9 v1.1 B1010	CCAFS LC-40	SpaceX CRS-4	2216	LEO (ISS)	NASA (CRS)	Success	Uncontrolled (ocean)
14	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
15	2015-02-11	23:03:00	F9 v1.1 B1013	CCAFS LC-40	DSCOVR	570	HEO	S. Air Force NASA NOAA	Success	Controlled (ocean)
16	2015-03-02	3:50:00	F9 v1.1 B1014	CCAFS LC-40	S-3A Eutelsat 115 West B	4159	GTO	ABS Eutelsat	Success	No attempt
17	2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
18	2015-04-27	23:03:00	F9 v1.1 B1016	CCAFS LC-40	Turkmen 52 / MonacoSAT	4707	GTO	n National Space Agency	Success	No attempt
19	2015-06-28	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)
20	2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
21	2016-01-17	18:42:00	F9 v1.1 B1017	VAFB SLC-4E	Jason-3	553	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
22	2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
23	2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
24	2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
25	2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
26	2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	S-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
27	2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
28	2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
29	2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
30	2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
31	2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
32	2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
33	2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
34	2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt
35	2017-06-03	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
36	2017-06-23	19:10:00	F9 FT B1029.2	KSC LC-39A	BulgariaSat-1	3669	GTO	Bulsatcom	Success	Success (drone ship)
37	2017-06-25	20:25:00	F9 FT B1036.1	VAFB SLC-4E	Iridium NEXT 2	9600	LEO	Iridium Communications	Success	Success (drone ship)
38	2017-07-05	23:38:00	F9 FT B1037	KSC LC-39A	Intelsat 35e	6761	GTO	Intelsat	Success	No attempt
39	2017-08-14	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

Executive Summary

Methodology Summary

- → Throughout this project a lot of data from various sources has come in to tell the story about the data from SpaceX.
- → The methodologies used come from both descriptive analytics which use data from real records documented by SpaceX depicting the relationships from launches, payloads, orbits, customers, outcomes and much more which say how the different variables are linked. This used great data visualisation techniques as well as exploratory data analytics to see the outcome of the data.
- → Another methodology which was employed was predictive analytics using modern methods such as k-nearest neighbours, support vector machines as well as decision trees to train models to see how data for future missions could turn out on a basic level to communicate the efficacy of the missions.

Results Summary

- → The analysis of SpaceX launch data demonstrates the feasibility of predicting successful rocket landings. By classifying key attributes into classes, valuable insights into launch performance and booster landings can be gained such as using the KSC LC-39A location as it has the highest success ratio. The SpaceX launch data yielded key insights, supported by clear numbers. Logistic Regression, SVM, Decision Tree, and KNN models were evaluated, with the Decision Tree achieving the highest accuracy at 87.5%, followed by SVM and KNN at 84.8%, and Logistic Regression at 84.6%. A scatter plot of payload mass versus launch outcome showed key trends across all sites, with specific payload ranges impacting landing success. Additionally, a pie chart comparison revealed that the site with the highest success ratio accounted for a significant portion of successful landings, underscoring its reliability. These results can enable better decision-making for future missions by optimising launch site selection and payload strategies, aligning with SpaceX's goals for increasing reusability and cost-efficiency.

Introduction

Project Background

This project uses data from SpaceX launches to predict the success of the Falcon 9 rocket's first stage landing. SpaceX has transformed space travel with its focus on reusability, significantly lowering costs and enhancing operational efficiency. By leveraging the SpaceX REST API, we gather historical data, including flight numbers, launch dates, booster versions, payload masses, orbit types, launch sites and landing outcomes. This dataset provides insights into the factors affecting landing success, with improvements noted since 2013. Our goal is to apply statistical and machine learning techniques to develop predictive models that assess landing success, thereby aiding decision-making for future missions and supporting SpaceX's commitment to innovation.

Context

SpaceX has enhanced its rocket landing capabilities, establishing itself as a leader in the aerospace industry. Its focus on reusability reflects a significant shift in how space missions are conducted. Understanding the factors that predict successful landings is crucial for optimising future missions and resource allocation. As space missions become increasingly complex, analysing historical launch data is essential. This project uses exploratory data analysis and machine learning to show insights to guide strategic decisions on launch site selection, payload considerations and mission planning. By accurately predicting the Falcon 9's first stage landing success, we can support SpaceX's mission to reduce costs and improve reliability in space travel.

Introduction – Problems to address

Data Wrangling and Cleaning:

How can we effectively preprocess the raw data obtained from the SpaceX API and web scraping to create a clean, usable dataset?

What strategies can we implement to handle missing or NULL values in key attributes, such as **PayloadMass**?

Feature Engineering:

Which attributes in the dataset are most strongly correlated with successful landingsand how can we derive additional features to improve our predictive models?

How can we determine the impact of various launch sites on landing success ratesand what other factors (e.g., payload mass, booster version) should be considered?

Model Development and Evaluation:

Which machine learning algorithms (e.g., Logistic Regression, Support Vector Machines, Decision Trees, K-Nearest neighbours) yield the best accuracy in predicting landing outcomes?

How can we optimise hyperparameters for these algorithms to enhance their performance?

Interactive Visual Analytics:

How can we utilise interactive visualisations to present findings effectively to stakeholders and facilitate deeper insights into launch patterns and success rates?

What insights can be gained from mapping launch sites and analysing their proximities in relation to landing outcomes?

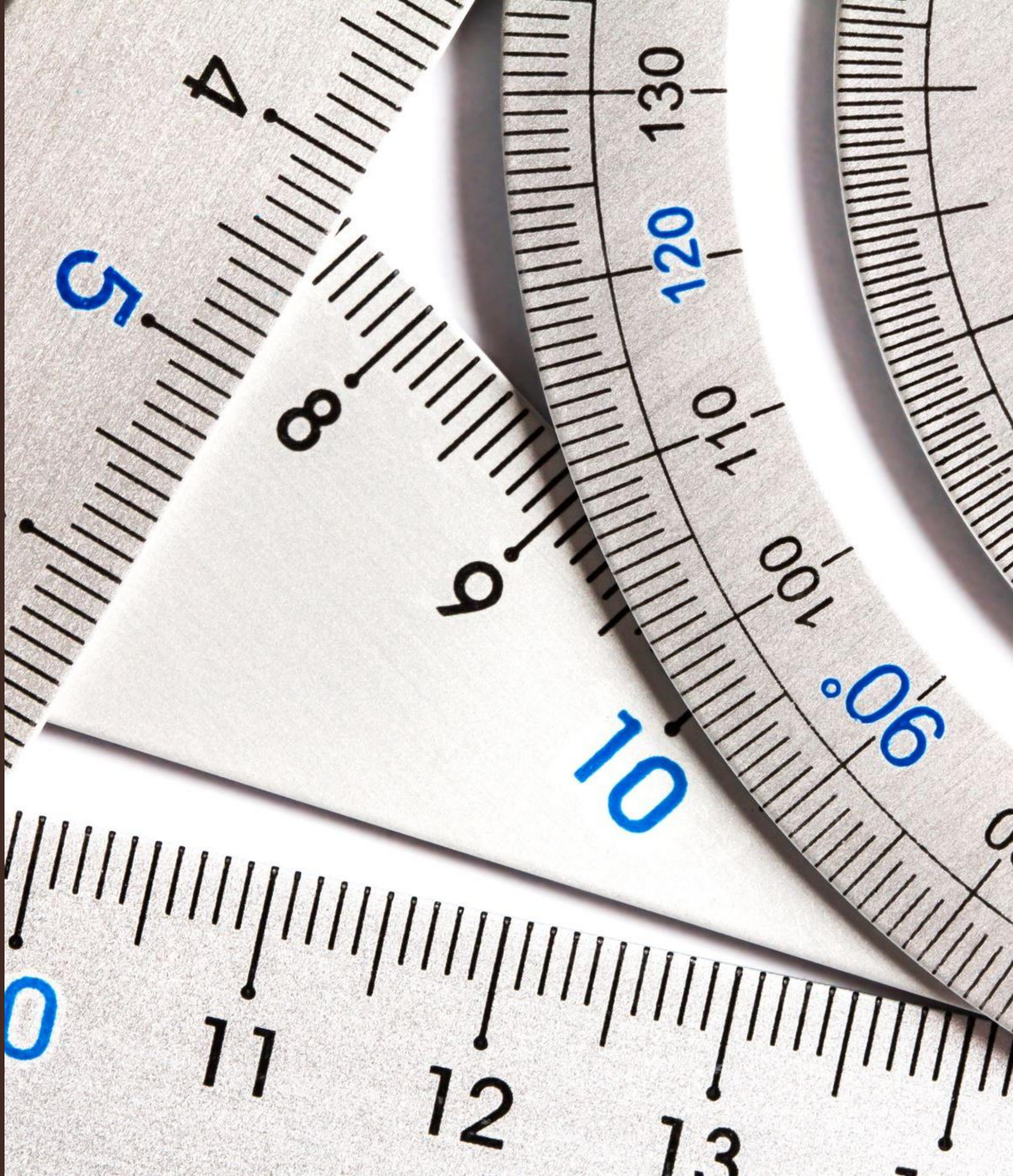
Predictive Accuracy and Insights:

What is the overall predictive accuracy of our final modeland how can the results be interpreted through metrics such as confusion matrices?

How can these insights inform future SpaceX missions and the strategic decisions surrounding rocket landings?

Section 1

Methodology



Methodology – Executive Summary

Data Collection Methodology
Data was collected through the **SpaceX REST API**, targeting the endpoint for past launches. This API provided information such as flight numbers, launch dates, booster versions, payload masses, orbit types, launch sites and landing outcomes. Additionally, web scraping was used to extract supplementary launch data from Wikipedia using BeautifulSoup.

Data Wrangling
The data underwent cleaning to address missing values (e.g., using the mean for missing **Payload Mass** values) and filtering to focus on Falcon 9 launches. Data normalisation using json_normalise ensured a flat structure for easier analysis, while additional API calls gathered specific details like booster or payload information.

Data Processing
The data was standardised to ensure consistency across features. Categorical variables, such as launch sites and orbit types, were converted to numerical values through one-hot encoding, making them suitable for inclusion in machine learning models.

Exploratory Data Analysis (EDA)
EDA was conducted through SQL queries and visualisations using Matplotlib and Seaborn. Key metrics, such as launch site success rates and payload mass effects on landing outcomes, were explored to uncover patterns and correlations in the data.

Interactive Visual Analytics
We used Folium to create an interactive map visualising launch sites and proximities and Plotly Dash to build a dashboard with sliders. These tools allowed real-time exploration of factors affecting landing success, such as payload mass and launch site.

Predictive Analysis Using Classification Models
Machine learning models, including Logistic Regression, Support Vector Machines, Decision Trees and K-Nearest neighbours, were used to predict the success of Falcon 9 landings. Hyperparameter tuning via Grid Search was conducted to improve model performance and evaluation was done using metrics such as accuracy and confusion matrices.

Building, Tuning and Evaluating Models
A pipeline was built to split the data into training and testing sets. Models were trained on the training set and hyperparameter tuning was applied using Grid Search. The best-performing model was evaluated using accuracy and other metrics on the test set, helping to predict future landing successes with high accuracy.

Data Collection

API Data Collection

SpaceX REST API: The primary dataset was collected which provides detailed information about launches. The API endpoint was used to gather historical launch data.

Requesting Data: Using the Python requests library, a GET request was made to the API endpoint.

Response Format: The data returned was in JSON format, containing a list of JSON objects representing individual launches.

Normalisation: The structured JSON data was normalised using the json_normalise function, transforming it into a flat, table-like format suitable for analysis.

BeautifulSoup: For additional information, such as specific payload details and historical records, web scraping was done using the Python package BeautifulSoup.

Target Source: Launch-related tables from relevant Wikipedia pages were identified as supplementary sources.

HTML Parsing: The HTML tables were parsed, and their data was converted into a Pandas Dataframe for further use.

Multiple API Endpoints: When certain attributes, such booster details, were missing from the primary dataset so further API requests were made to gather additional data.

Merging: Data from the API and web scraping was merged into a consolidated Dataframe.

Filtering and Cleanup: Launches related to Falcon 1 were filtered out, focusing solely on Falcon 9 for analysis consistency.

Data Collection – SpaceX API

[SpaceX API Calls Notebook]

1. API Endpoint Targeting was used for retrieving past launch data and GET request using the Python requests library to fetch data.

2. A GET request was made to the SpaceX API to retrieve the data. The response from the API was received in JSON format. The .json() method was used to parse the response into a Python-readable JSON object.

3. Data Normalisation: The structured JSON data was transformed into a flat, tabular structure from Pandas. The columns like Flight Number, Launch Date, Booster Version, Payload Mass, Orbit Type, Launch Site, and Landing Outcome were extracted from the response.

4. Additional API calls were made to endpoints to gather further details such as booster specs. The supplementary data was merged with the primary launch data using unique identifiers like Rocket ID.

5. Non-Falcon 9 launches were filtered out and missing values were addressed (calculating the mean for missing payload masses). The final clean dataset was stored in a Pandas Dataframe for analysis.

Data Collection – Scraping

Webscraping Notebook

1. Wikipedia pages containing Falcon 9 launch records and associated tables. The tables with valuable information, such as launch dates and payloads were identified for scraping.

2. The requests library was used to send an HTTP GET request to the target URL. Then HTML content of the web page was fetched and stored for further processing.

3. The HTML content was parsed using BeautifulSoup to target HTML table elements using methods like find() and find_all() to capture the relevant data.

4. The extracted table data was converted into a Pandas Dataframe for easy manipulation and integration with other datasets. The columns were then renamed for clarity and removing unwanted rows.

5. Any missing values in the scraped data were filled using appropriate methods, the mean.

6. The cleaned data from the web scraping process was merged with the API data to create a comprehensive dataset.

Data Wrangling

[Data Wrangling Notebook]

EDA with Data Visualisation

[Data Visualisation Notebook]

The charts used during the data visualisation stage were scatter plots, bar charts, line plots and the use of feature engineering afterwards to prepare the data for machine learning models, we created dummy variables from categorical columns. This transformation allows algorithms to better interpret the categorical data, enhancing model performance and accuracy in predictions

For each type of relationship there was a rationale for each chart selection based on their advantages and insights you can gain from looking at the data. The reasons for these are:

→ **Trend Identification:** Scatter plots were particularly useful for identifying relationships and trends over time, which is critical in assessing the evolution of SpaceX's capabilities and mission strategies.

→ **Comparative Analysis**: The bar chart was selected for its effectiveness in comparing categorical data, allowing for a straightforward understanding of success rates across different orbit types.

→ **Temporal Insights**: The line plot was essential for visualising changes over time, providing insights into operational improvements and helping to set benchmarks for future performance.

→ **Model Readiness**: The creation of dummy variables was a necessary step in preparing the data for machine learning, ensuring that the model could effectively learn from both categorical and numerical features.

EDA with SQL

SQL Notebook

Unique Launch Sites: To display the names of the unique launch sites used in SpaceX missions.

Launch Sites Starting with 'CCA':To display 5 records of launch sites that begin with the string 'CCA'.

Total Payload Mass for NASA (CRS): To calculate the total payload mass carried by all the boosters launched by NASA.

Average Payload Mass for F9 v1.1: To find the average payload mass carried by the booster version F9 v1.1.

Date of First Successful Ground Pad Landing: To list the date when the first successful landing outcome on a ground pad was achieved.

Boosters with Success in Drone Ship and Specific Payload Mass: To list the names of boosters that successfully landed on a drone ship and had a payload mass between 4000 and 6000.

Total Successful and Failed Mission Outcomes: To list the total number of successful and failed mission outcomes.

Boosters with Maximum Payload Mass: To list the names of the booster versions that have carried the maximum payload mass.

Monthly Records of Failed Drone Ship Landings (2015): To display records of month names, failure landing outcomes on drone ships, booster versions, and launch sites for the months in the year 2015.

Ranking of Landing Outcomes Between Dates: To rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the specified dates in descending order.

Build an Interactive Map with Folium

Launch Site Markers: Added markers at the specific launch site coordinates to visualise the exact locations of SpaceX launch sites. These provide an easy reference for identifying where launches take place geographically.

Circles Representing Proximity to Launch Site: Added circles around each launch site to represent proximity or range. This helps visualise the surrounding area and identify nearby infrastructure or features that influence launch decisions.

Lines Representing Distance Between Locations: Added lines connecting launch sites to important points such as the ocean or drone ship locations. This helps to analyse proximity to landing zones and how far boosters need to travel for recovery.

Popup Information on Markers: Added popups to each marker with detailed information such as the launch site name. This provides users with additional context when interacting with the map.

These map objects were chosen to provide a clear, interactive way to explore geographic factors related to SpaceX launch sites and mission success. Markers and circles help to identify key locations, and their proximities and lines allow for a better understanding of distances and logistical considerations. The interactive popups offer immediate insight without needing to cross-reference the data.

Folium Notebook

Build a Dashboard with Plotly Dash

Pie Chart (Launch Success Count for All Sites): Displays the total number of successful launches across all SpaceX launch sites, giving a clear picture of which sites have had the most success. Users can interact with the chart to focus on specific sites or see the overall distribution of success rates across multiple launch sites.

Pie Chart (Highest Launch Success Ratio by Site): Highlights the launch site with the highest success ratio. This allows users to identify which site has consistently performed the best in terms of successful launches. Clicking on the pie segments provides further information about specific launch sites and their contribution to the overall success.

Plot (Payload vs. Launch Outcome for All Sites): Shows the relationship between payload mass and the launch outcome across all sites. This helps visualise how payload mass impacts launch success. This includes a range slider to adjust the range of payload mass displayed, allowing users to analyse outcomes for specific mass ranges.

These charts and interactions were designed to provide a detailed and interactive overview of SpaceX's launch performance. The pie charts focus on success counts and success ratios, giving stakeholders a clear visual of which sites have performed best. The scatter plot helps analyse the impact of payload mass on launch outcomes, offering insights into potential factors that influence launch success. The interactions allow users to explore different aspects of the data, making the dashboard more dynamic and insightful for decision-making.

Dash Web App

Predictive Analysis (Classification)

Predictive Analysis Notebook

→ 1. Building the Classification Model

The process began by preprocessing the data. Numerical features were scaled to ensure consistency, and categorical features like orbit types and launch sites were converted into numeric form using dummy variables. Missing data was handled through manipulation. After preprocessing, the data was split into training and testing sets, 80/20, to train the model and evaluate its performance on unseen data.

→ 2. Model Selection: Several models were chosen for evaluation, including:

Logistic Regression for binary classification

Support Vector Machine (SVM) to find the best hyperplane separating the outcomes

Decision Tree Classifier for rule-based classification

K-Nearest neighbours (KNN) for proximity-based classification.

→ 3. Evaluating the model

Models were assessed using accuracy and F1-score. The accuracy measured overall performance, while precision and recall helped evaluate how well the model handled false positives and negatives. A confusion matrix was used to better visualise classification errors for successful and unsuccessful landings.

→ 4. Model Improvement

To improve performance, Grid Search was used for hyperparameter tuning, optimising parameters like C for Logistic Regression or max depth for Decision Tree. k-fold cross-validation was applied to validate performance across different subsets of the training data, reducing overfitting risks.

→ 5. Best Performing Model

After tuning, SVM with optimised hyperparameters emerged as the best model, offering the highest accuracy and balanced metrics. The final model was tested on the unseen data to confirm its reliability in predicting Falcon 9 landing outcomes.

Results

1. Exploratory Data Analysis Results

Launch Sites Analysis: A **bar chart** showed that certain launch sites, such as **KSC LC-39A** and **CCAFS SLC-40**, had significantly higher numbers of launches compared to others. This suggests these sites are more frequently used by SpaceX.

Payload Mass Analysis: A **scatter plot** of **Payload Mass vs. Flight Number** revealed that missions with higher payload masses were more likely to be successful, indicating a potential correlation between payload and mission outcome.

Success Rate by Orbit Type: A **bar chart** displayed that **LEO (Low Earth Orbit)** missions had the highest success rate, indicating the orbit type as an important predictor for landing outcomes.

Success Rate Over Time: A **line chart** demonstrated that SpaceX's mission success rate improved significantly over time, particularly after 2015, reflecting advancements in landing technology.

2. Interactive Analytics Demo in Screenshots

Launch Success Count for All Sites: This was represented in a **pie chart** showing the distribution of successful launches across various sites. By interacting with the chart, users could soon in on specific launch sites and examine their contribution to overall success.

Launch Site with Highest Success Ratio: A **pie chart** displayed the launch site with the highest success ratio. Users could select different sites using a dropdown menu for quick comparison.

Payload vs. Launch Outcome Scatter Plot: A **scatter plot** showed how payload mass influenced launch outcomes. A **range slider** was provided, allowing users to filter the data based on different payload ranges, which revealed that success rates varied across different payload masses.

3. Predictive Analysis Results

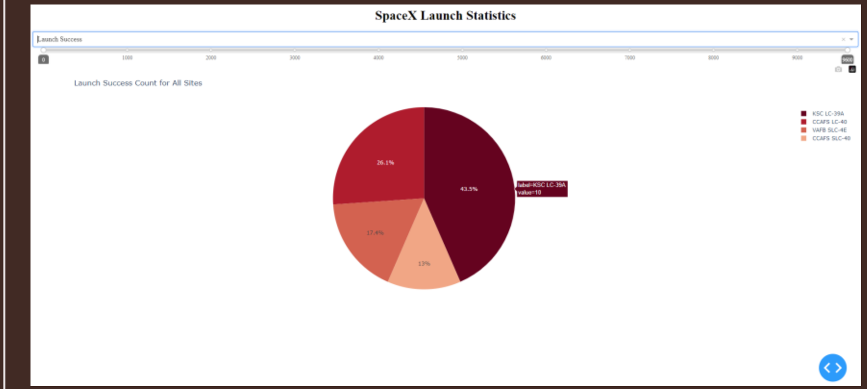
Logistic Regression: Achieved **moderate accuracy** but was less effective in handling complex relationships between features.

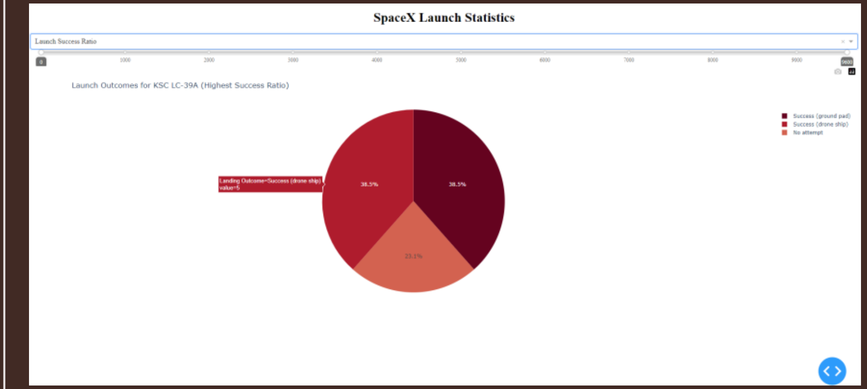
Support Vector Machine (SVM): Provided **strong performance** with a good balance between precision and recall, especially after **hyperparameter tuning** through **Grid Search**.

Decision Tree Classifier: Showed **decent accuracy** but was prone to overfitting even despite tuning.

K-Nearest Neighbours (KNN): Performance was **reasonable**, but it struggled with larger datasets and didn't generalise as well as SVM.

Best Model: After tuning, the **SVM model** achieved the highest accuracy and best overall performance, making it the best model for predicting Falcon 9 first-stage landing outcomes.



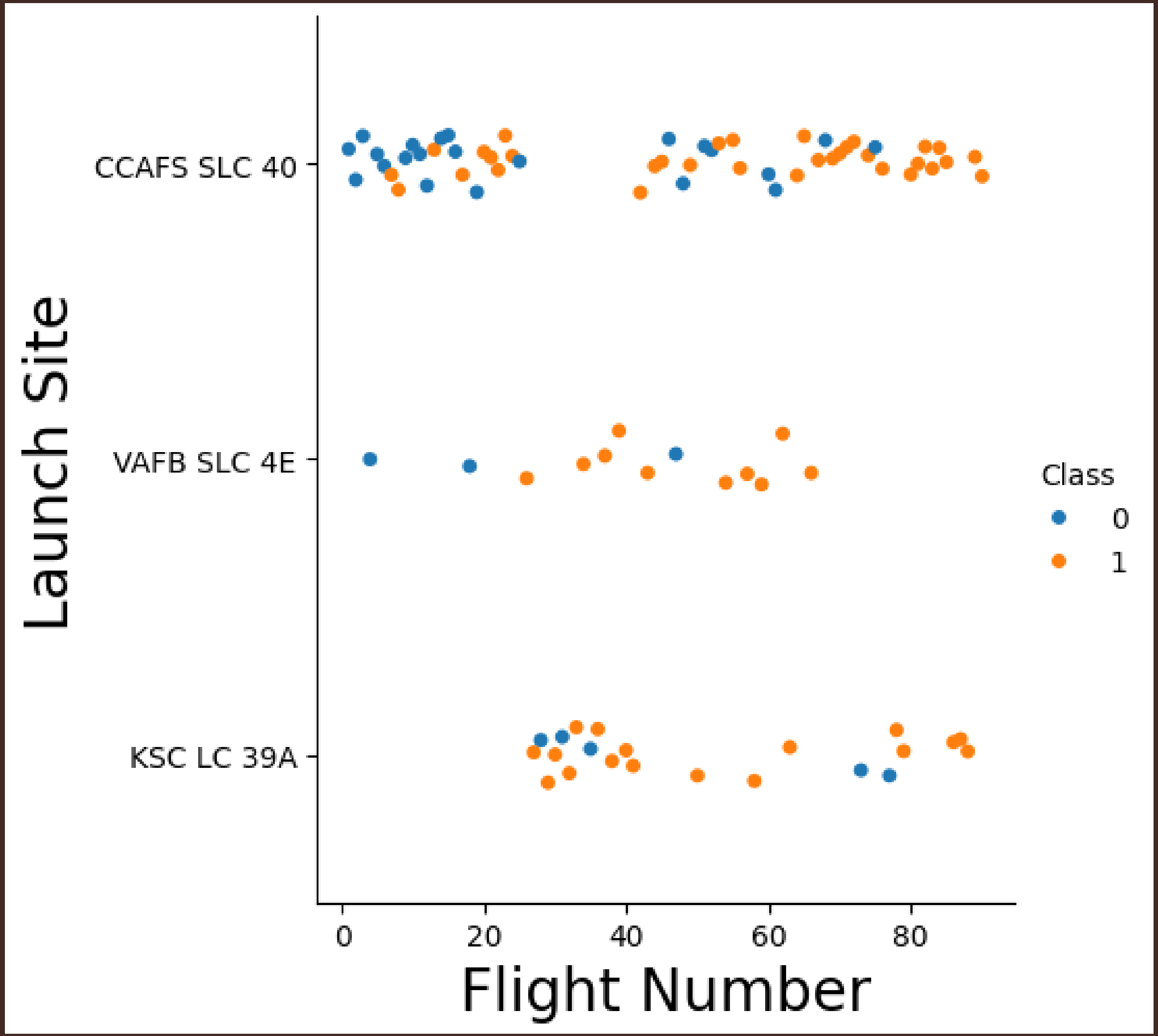


Section 2 Insights from EDA

Flight Number vs. Launch Site

CCAFS SLC 40 has the highest number of launches, with early flights (0-20) mostly showing unsuccessful outcomes, while later flights (40-90) indicate a shift towards successful landings, especially in flight clusters 40-50 and 70-80. This suggests improved success rates or procedural changes over time.

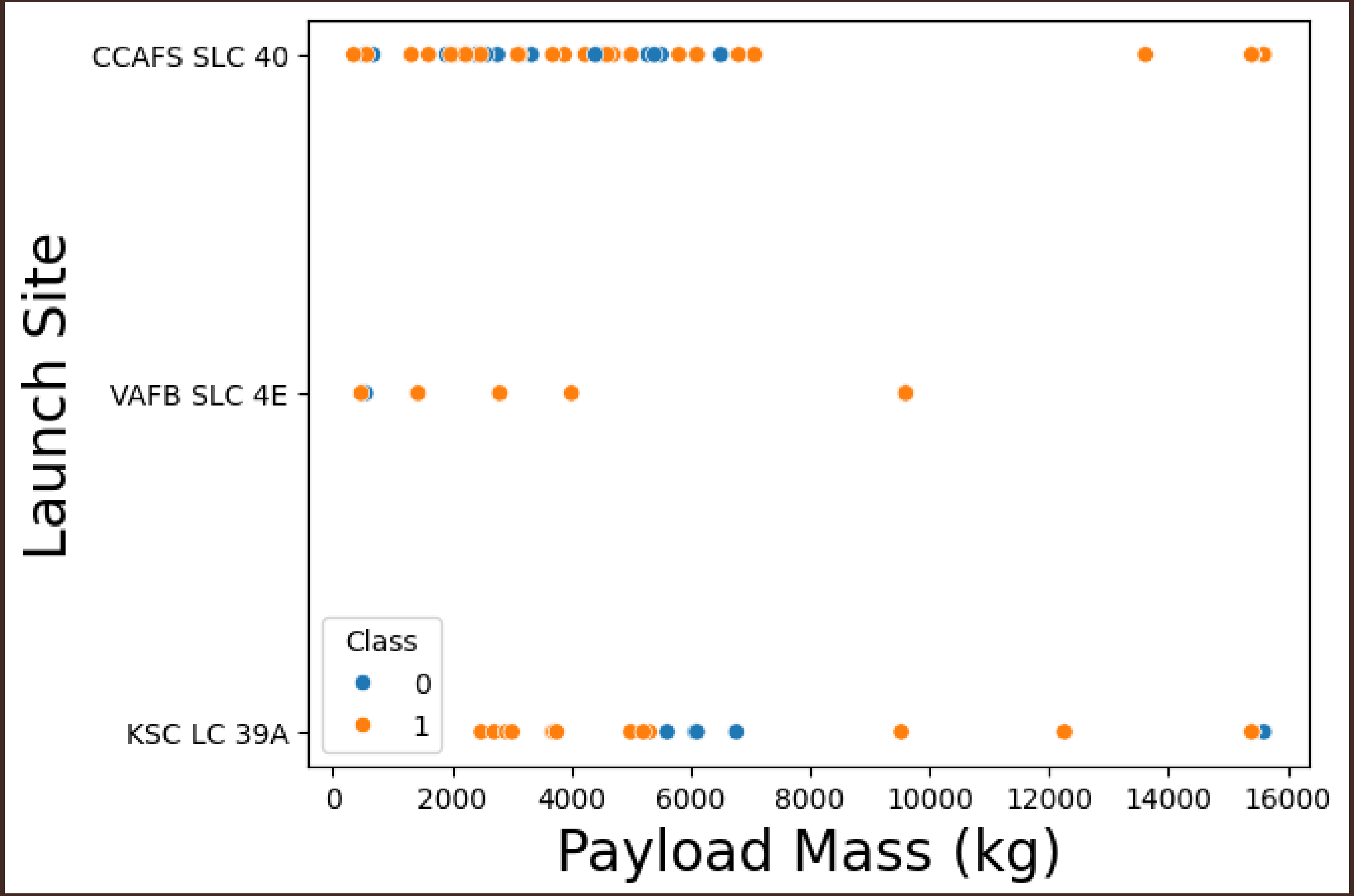
VAFB SLC 4E has fewer launches (10-12), with an even mix of successful and unsuccessful outcomes that seem to alternate rather than cluster, indicating varied mission outcomes. KSC LC 39A has two main clusters of launches around flight numbers 30-40 and 80-90, with fewer launches in between. It shows a slight predominance of successful landings in later flights. These patterns highlight operational and success trends across the sites, with CCAFS SLC 40 being the most active and improving over time.



Payload vs. Launch Site

CCAFS SLC 40 site has the highest number of launches, with low to medium payload mass launches mostly showing successful outcomes but launches do become unsuccessful as payload gets to 6000kg, however the site has attempted very large payloads and was successful. This suggests they could take projects with large payloads as it shows success but more data is required.

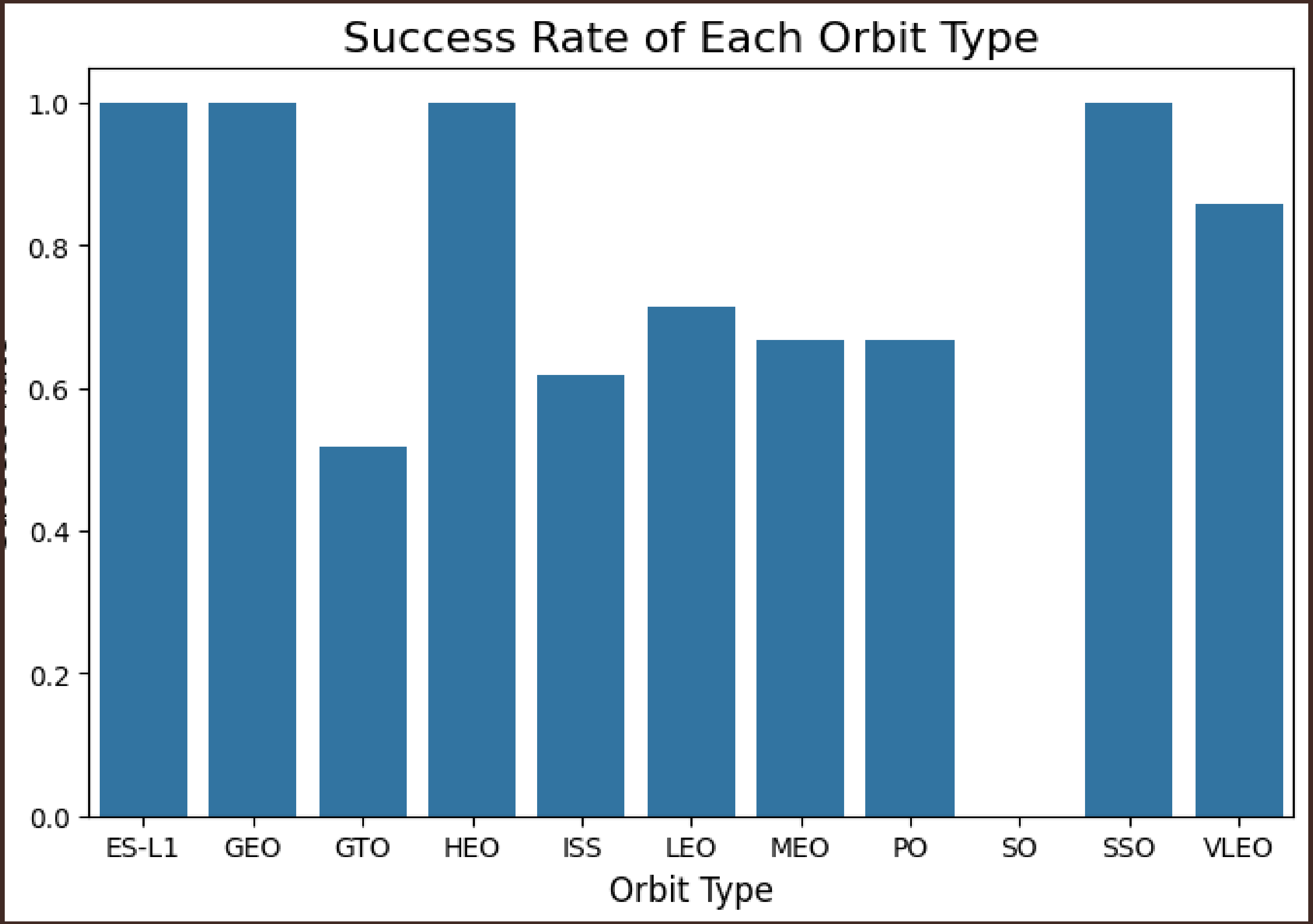
VAFB SLC 4E has fewer launches, with an even mix of payload sises but all were successful which suggests the site is good for varied projects but once again more data is required. KSC LC 39A has many successful payload launches however failed launches around 6000kg which suggests it should not those paylaods. It shows a slight predominance of successful landings in greater payloads however there is not enough data to make a conclusion. These patterns highlight operational and success trends of payloads across the sites, with CCAFS SLC 40 being the most active and successful.



Success Rate vs. Orbit Type

The bar chart reveals the success rates of different orbit types, with values ranging from 0% to 100%. Orbit types like ES-L1, GEO, HEO, and SSO boast a perfect 100% success rate, showing high reliability. In contrast, GTO has a lower success rate of around 60%, while ISS, LEO, MEO, and PO display moderate success rates between 70% and 80%. VLEO demonstrates strong performance with a success rate close to 90%.

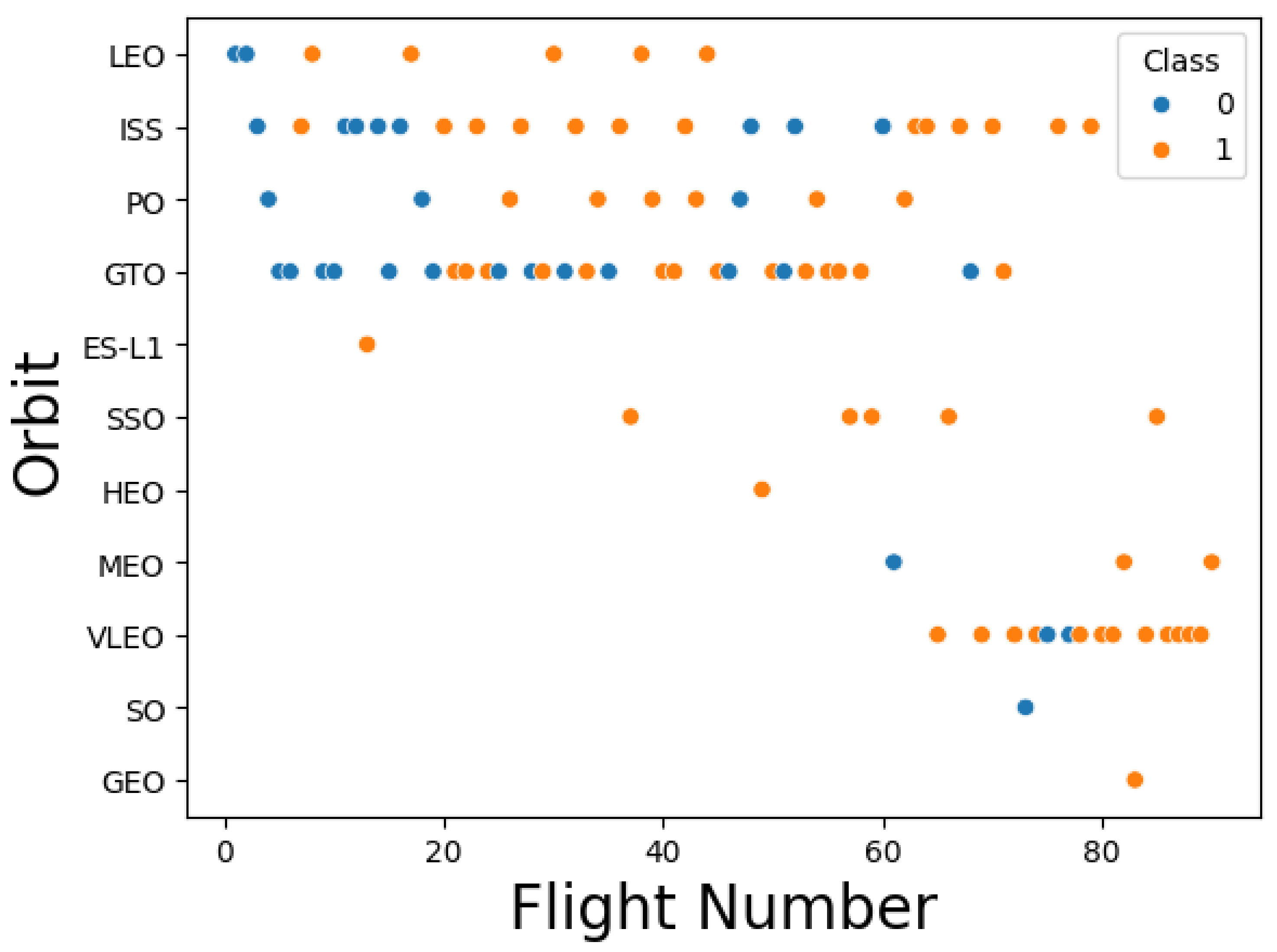
On the other hand, the SO orbit stands out with a 0% success rate, indicating significant challenges or a lack of successful missions. Overall, orbits like ES-L1 and GEO are consistently reliable, while GTO and ISS show more variability. This insight can help assess the difficulty or reliability of space missions based on their target orbit.



Flight Number vs. Orbit Type

The scatter plot reveals how the success of different payload orbits have changed over time. Orbit types like LEO, ISS, PO and GTO have been the primary orbit however not successful showing how we had to understand how the orbits worked fully for successful missions in later flights. In contrast, ES-L1, SSO, HEO, MEO, SO and GEO have much less launches which does not tell us a lot but the fact they can become useful as technology advances as some are successful.

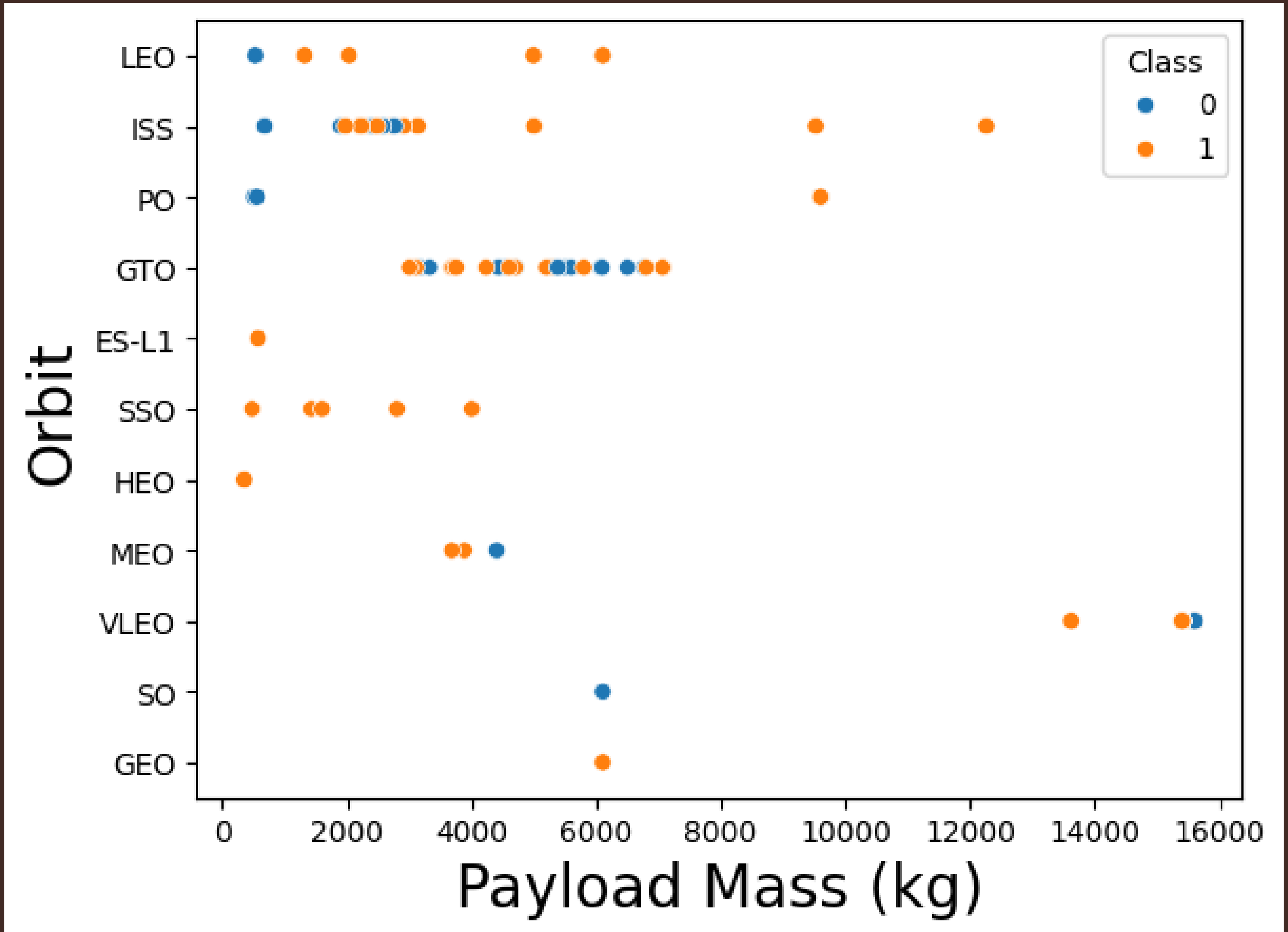
Finally, VLEO - which started in later flights - has a huge success ratio and is primarily used in later flight missions. Overall, orbits like ISS, GTO and PO are reliable now due to the amount of data and trials we have done on them, while SO and GEO are less so due to minimal flights. This insight can help assess the difficulty or reliability of space orbits in future missions and how much data we have on the specific orbit.



Payload vs. Orbit Type

The scatter plot reveals how the success of different payload masses and orbits used are related. Orbit types like ISS and GTO have been used for many payload masses but have varying success. This contrasts with SSO and LEO who have great success in small payloads showing how these orbits are useful for those specific payloads. Then payloads, ES-L1, PO, HEO, MEO, SO and GEO have much less launches, but which are successful which does not tell us a lot but the fact they can become useful as technology advances as some are successful.

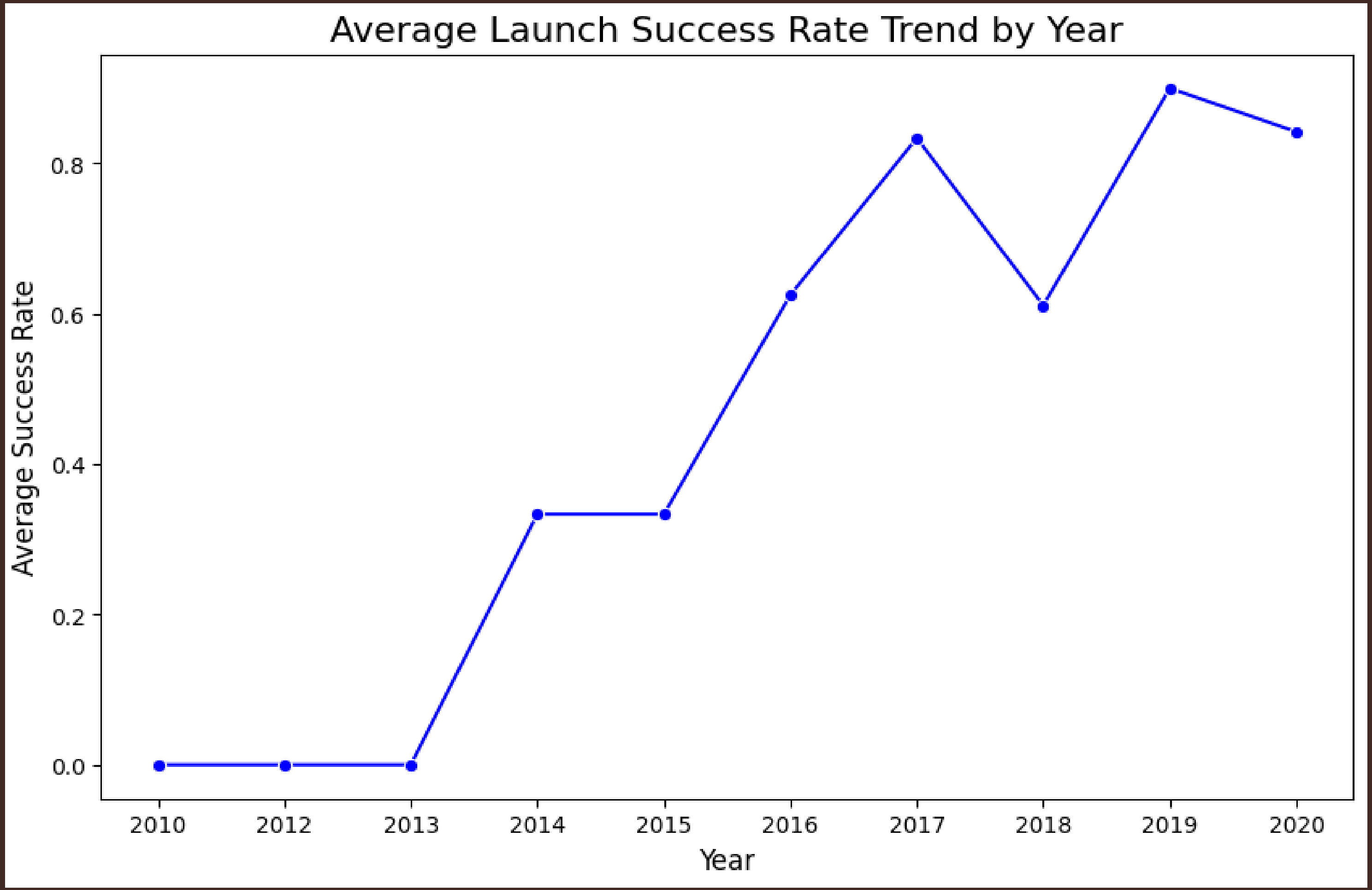
Finally, VLEO, is quite successful in very heavy payloads which suggests a very useful purpose for them. Overall, this insight can help assess the difficulty or reliability of different payloads in specific orbits and suggests which orbits in future missions has the highest chance of working based on the payload mass.



Launch Success Yearly Trend

The graph shows the trend of average launch success rates from 2010 to 2020, highlighting a general upward trajectory. This suggests that space launch reliability has improved over the decade. Notably, there was a significant jump in success rates from 2013, indicating advancements in technology and procedures. The highest success rate was reached in 2019, though a slight decline occurred in 2020.

Despite the overall improvement, the success rate fluctuates slightly between years, likely due to technical challenges, weather conditions, or mission complexities. However, the positive trend reflects progress in space launch technology, enhanced safety measures, and growing expertise in the field. This tells us the way missions are being executed is near perfect and we should continue in the same way.



All Launch Site Names

```
%sql select distinct "Launch_Site" from SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

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

- The SQL query is designed to retrieve a list of unique launch sites from the SpaceX dataset. By using the DISTINCT keyword, the query ensures that only unique values are returned, eliminating any duplicate entries from the Launch site column. The query operates on the SpaceX table, which contains data about SpaceX launches. This query is useful for identifying all distinct launch sites that have been used in SpaceX missions, providing insights into the diversity of locations from which launches have occurred.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The SQL query retrieves up to five records from the SpaceX table where the Launch site starts with the string 'CCA'. The LIKE operator is used to match launch site names that begin with 'CCA', which could correspond to sites like CCAFS. The LIMIT 5 clause restricts the output to just five records, making the query efficient by only showing a small subset of relevant results. This query helps quickly explore data related to specific launch sites that match the 'CCA' pattern.

Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" == 'NASA (CRS)';

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

sum(PAYLOAD_MASS__KG_)
45596

- The SQL query calculates the total payload mass, in kilograms, carried by rockets for NASA’s CRS missions. The SUM function adds up all the values from the payload mass column where the customer is identified as 'NASA (CRS)'. This query is useful for determining the total mass of cargo delivered by SpaceX to support NASA's CRS missions, which are crucial for resupplying the ISS. It provides insights into NASA’s payload operations with SpaceX.

Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" == 'F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- The SQL query calculates the average payload mass, in kilograms, carried by SpaceX's F9 v1.1 rockets. The AVG function is used to compute the mean value of the payload mass from the payload mass column, specifically for launches using the F9 v1.1 booster version. This query helps assess the typical payload capacity of the F9 v1.1, offering insights into how much weight this particular booster version typically transported during its missions.

First Successful Ground Landing Date

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" == 'Success (ground pad)';

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

min(Date)
2015-12-22

- The SQL query retrieves the earliest successful landing date of a Falcon 9 rocket on a ground pad. The MIN function is used to find the minimum value from the date column, which corresponds to the first instance of a successful landing classified as 'Success (ground pad)' in the landing outcome column. This query provides insight into when SpaceX first achieved a successful ground pad landing, an important milestone in its reusability efforts.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" == 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 and 6000;
```

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

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

- The SQL query retrieves the booster versions that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg. By filtering the results using the landing outcome as 'Success (drone ship)' and restricting the payload mass to this specific range, the query identifies which booster versions were able to handle such payloads during successful drone ship landings. This information helps assess booster performance under specific conditions.

Total Number of Successful and Failure Mission Outcomes

```
%sql select count("Mission_Outcome") from SPACEXTABLE;

 * sqlite:///my_data1.db
Done.

count(Mission_Outcome)
                   101
```

- The SQL query counts the total number of mission outcomes recorded in the SpaceX table. It provides the overall number of missions present in the dataset, regardless of whether the outcomes were successes or failures. This query is useful for understanding the sise of the dataset and how many missions have been logged for analysis.

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
```

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

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

- The SQL query retrieves the booster version associated with the mission that carried the maximum payload mass. It first finds the largest payload mass in the SpaceX table using the inner query and then matches that payload to the specific booster version in the outer query. This query helps identify which booster was used for the heaviest payload, offering insights into the capabilities of different booster versions.

2015 Launch Records

```python
query = """
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'
    END AS "Month_Name",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) == '2015'
  AND "Landing_Outcome" == 'Failure (drone ship)'
"""

# Execute the query
cur.execute(query)

# Fetch and print the results
results = cur.fetchall()
for row in results:
    print(row)
```

```
('January', 'Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40')
('April', 'Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40')
```

- The provided SQL query extracts specific information from the SpaceX table about rocket launches that failed while landing on a drone ship in the year 2015. It uses a CASE statement to convert the month portion of the Date field, which is formatted into its corresponding month name. The SELECT statement retrieves four columns and the WHERE clause filters the results to include only those entries from the year 2015 with a landing outcome of 'Failure (drone ship)'. After executing the query, the results are fetched and printed, allowing the user to see a list of failed launches along with their associated month, booster version, and launch site.
- This query is useful for analysing trends in landing failures over time and understanding the performance of specific boosters and launch sites.

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

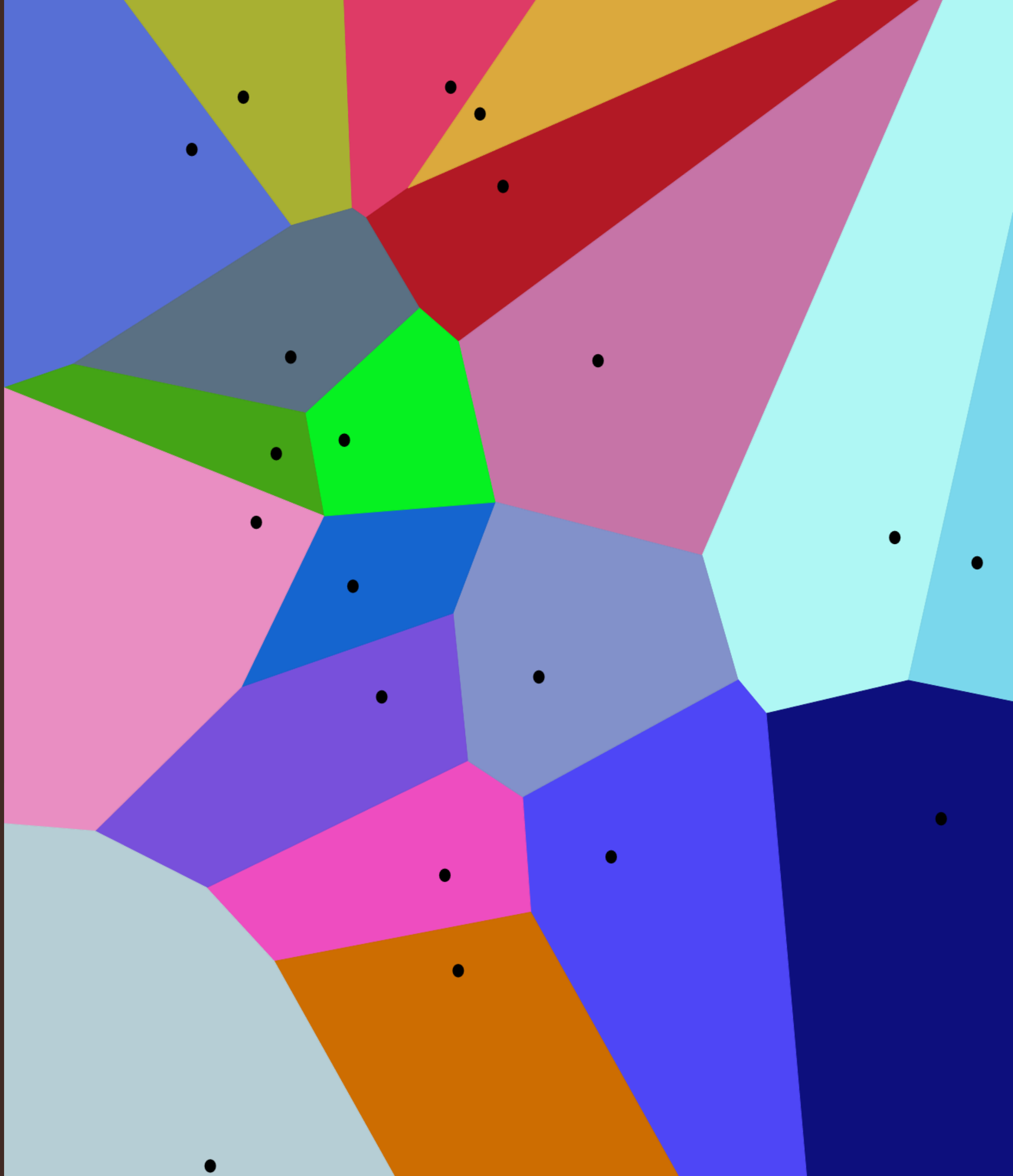
```
%sql select "Landing_Outcome", count("Landing_Outcome") as "Outcome_Count" from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' gro
```

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

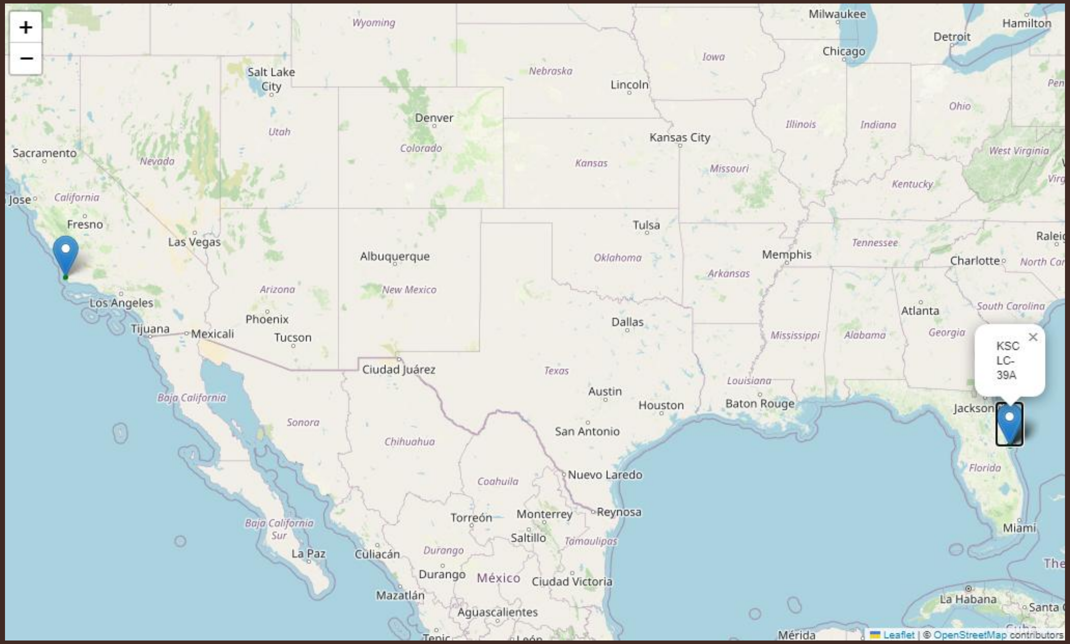
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This SQL query retrieves and counts the occurrences of each landing outcome from the SpaceX table for rocket launches that occurred between June 4, 2010, and March 20, 2017. It selects the landing outcome column and uses the count() function to calculate how many times each outcome appears, labelling this count as outcome count. The WHERE clause filters the records to only include those within the specified date range. The GROUP BY clause organises the results by each unique landing outcome. Finally, the ORDER BY clause sorts the results in descending order based on the outcome count, providing insight into the success and failure rates of rocket landings during that period.

Section 3
Launch Site Proximity Analysis



Launch sites on the map



- ❖ Launch sites are vital for space exploration and research, influencing mission planning by determining optimal trajectories and launch windows based on factors like Earth's rotation and target orbits. Their locations are needed for tracking and communication with spacecraft, ensuring mission success through effective coordination.

- ❖ In environmental monitoring, launch sites impact satellite orbits, which are essential for tracking Earth's climate and natural disasters. This knowledge helps calibrate satellite data for accurate environmental measurements.

- ❖ From a military perspective, monitoring launch sites is critical for early warning systems that detect missile launches and inform defence strategies.

- ❖ In scientific research, strategic site selection based on proximity to astronomical phenomena enabling valuable investigations.

Overall, markers on maps indicating launch site locations are indispensable for mission planning, environmental monitoring, military applications, and scientific research.

Successful and failed launches for each site on the map

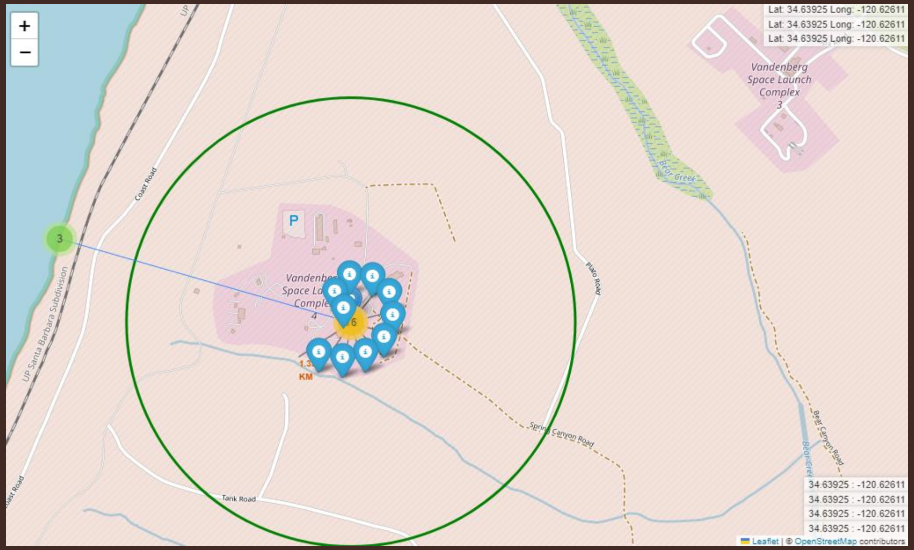
- Colour-labelled launch outcomes on maps offer significant visual information that serves multiple purposes. They enable quick assessments of mission success by using different colours to represent various outcomes, such as successful launches, failures, or partial successes

- This colour-coding facilitates an easy evaluation of the overall success rates of different launch sites. Furthermore, analysing the colour distribution can reveal trends and patterns, such as failed launches, which could indicate areas that require further investigation.

- Clusters of failed launches can highlight potential issues related to specific launch sites, weather conditions, or operational procedures providing insights for enhancing safety and reliability. The colour-labelled outcomes also allow for effective comparisons of performance helping to identify the most dependable options for future missions.

In summary, colour-labelled launch outcome maps are essential tools for understanding launch history, identifying performance patterns, and promoting safety in the field of space exploration.

Distances between a launch site and its proximities

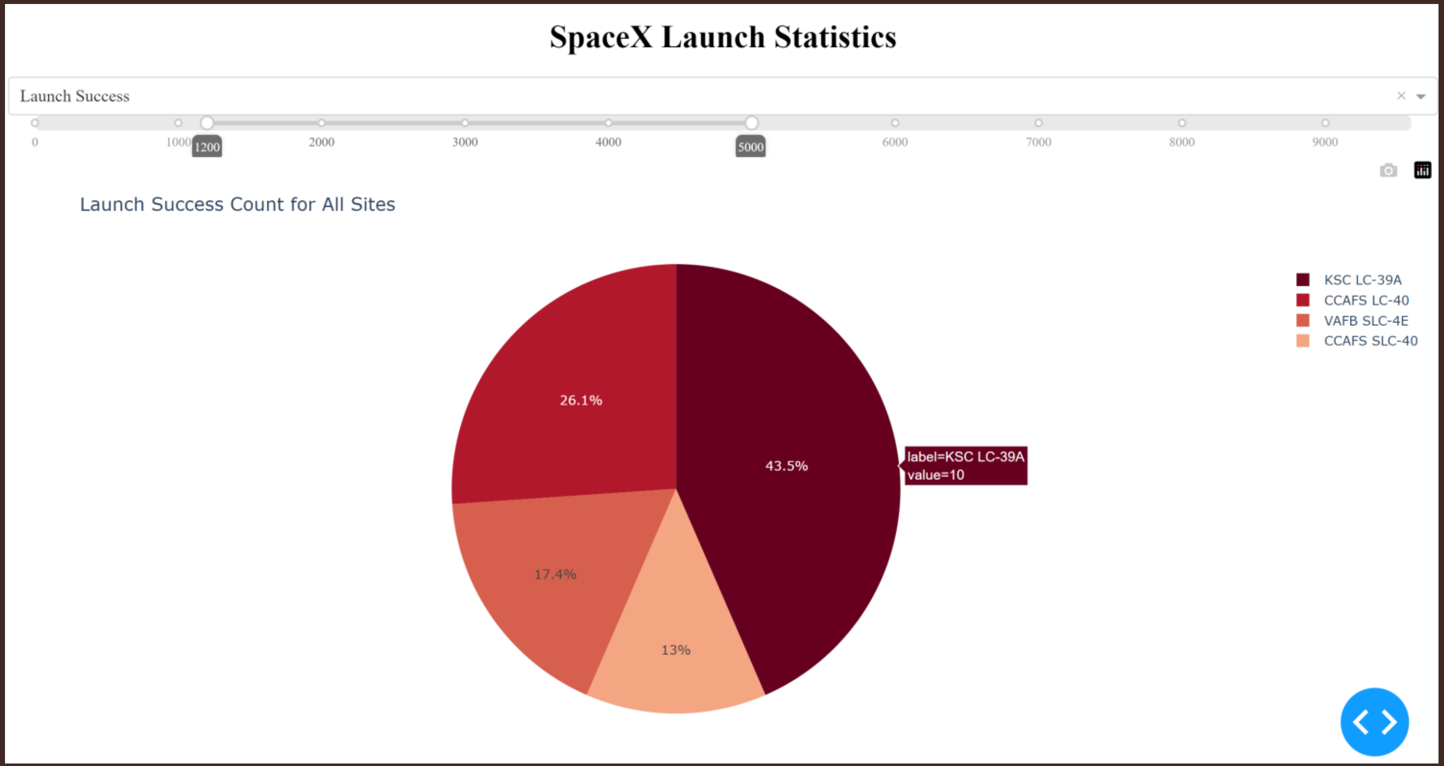


- ❖ The proximity of a launch site to infrastructure such as railways, highways, and coastlines plays a crucial role in its operational efficiency and effectiveness.
- ❖ Access to transportation networks enables the efficient movement of materials and simplifying post-launch cargo shipments. In emergencies, quick access to railways and highways allows for rapid deployment of response teams and equipment.
- ❖ The presence of a launch site can stimulate economic growth in the surrounding area, encouraging infrastructure development and attracting industries related to aerospace and logistics.
- ❖ However, there are environmental considerations to keep in mind, especially for coastal launch sites that can use marine transportation to minimise the impact of land-based problems. While proximity to populated areas may raise concerns about noise and pollution from launch operations, effective planning and mitigation measures can stop these issues.
- ❖ Well-developed transportation infrastructure enhances security by allowing controlled access to the launch site and improving the response capability of security if there are emergencies.

In conclusion, careful consideration of the proximity to transportation infrastructure is vital when selecting and developing launch sites, as it affects logistics, emergency readiness, economic development, environmental impact, and security.

Section 4
Dashboard with Plotly Dash

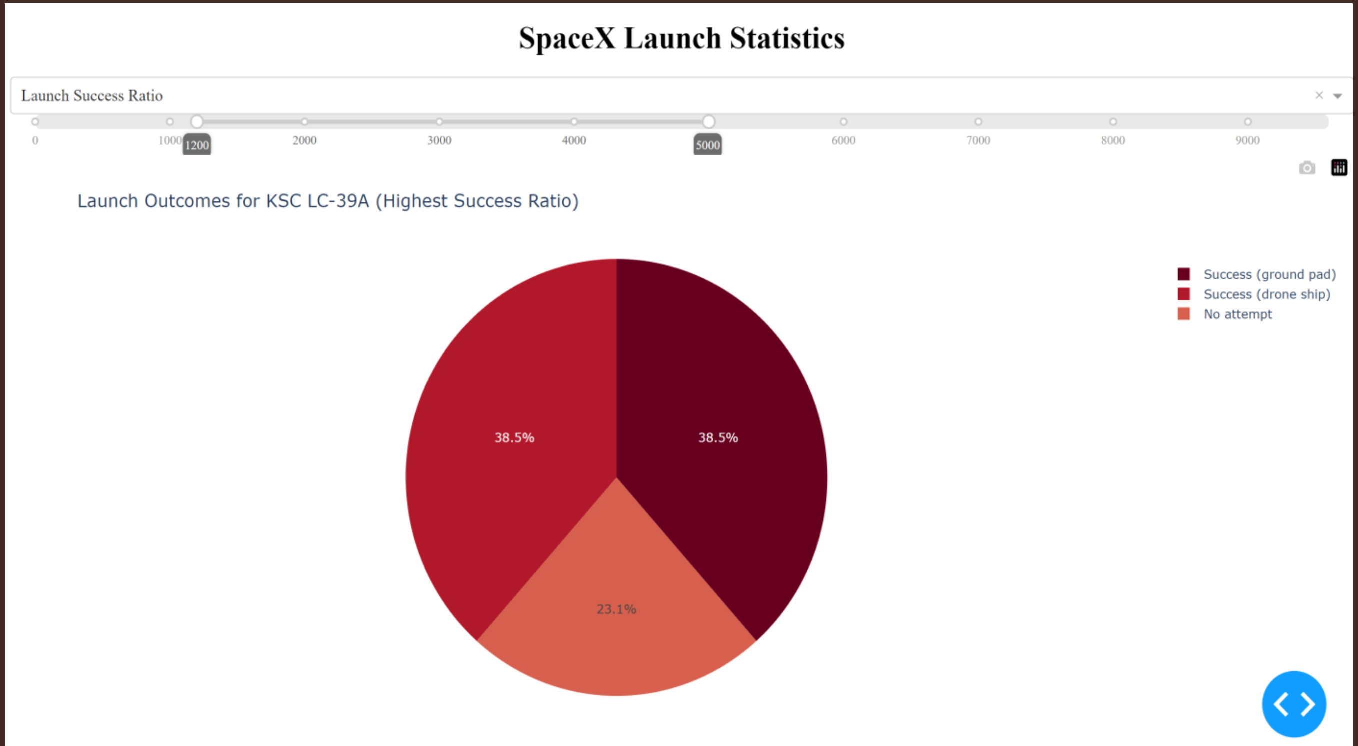
Launch success count for all sites



- ❖ The pie chart showing the launch success count for all sites offers a clear visual representation of their performance in terms of success rates. It allows a quick identification on which launch sites are performing well and which are not, facilitating comparison and strategic planning for future missions. In this case the KSC LC-39A is the most successful site, it can all be seen with percentage success.
- ❖ Additionally, the chart highlights trends, such as improvements in success rates at specific sites, guiding resource allocation and investment decisions.

Overall, the insights gained from the pie chart are crucial for making decisions and improving the efficacy and effectiveness of future launches.

Launch success ratio

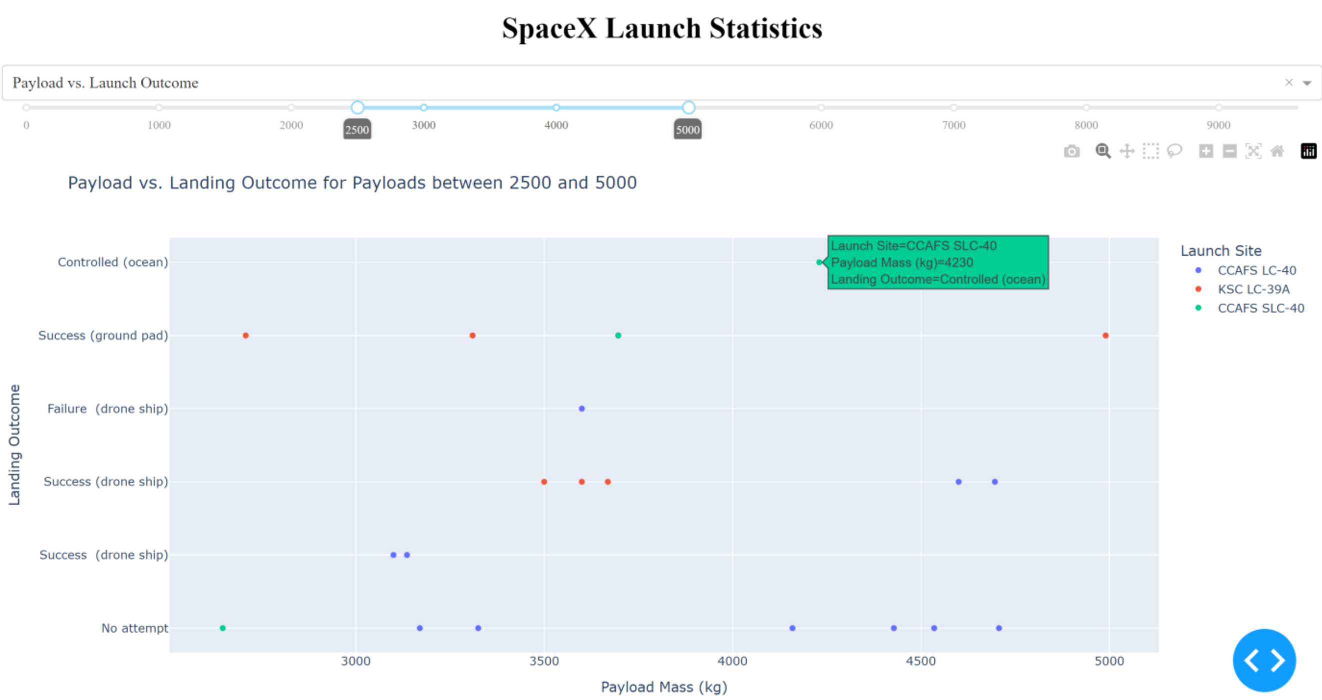


- The pie chart showing the highest launch success ratio is essential for understanding which launch location has consistently achieved successful missions and by how much even if the data changes in the future. By visually representing the success rates it allows stakeholders to easily identify the most reliable site for strategic decisions regarding future missions.

- The findings from this chart not only highlight the top-performing launch site but also suggest potential factors which we can analyse, if needed, contributing to its success. This information is valuable for organisations looking to optimise their launch strategies.

Ultimately, the chart serves as a key tool for enhancing mission planning and improving overall launch reliability in the space industry.

Payload vs. Launch Outcome for all sites, with different payload masses

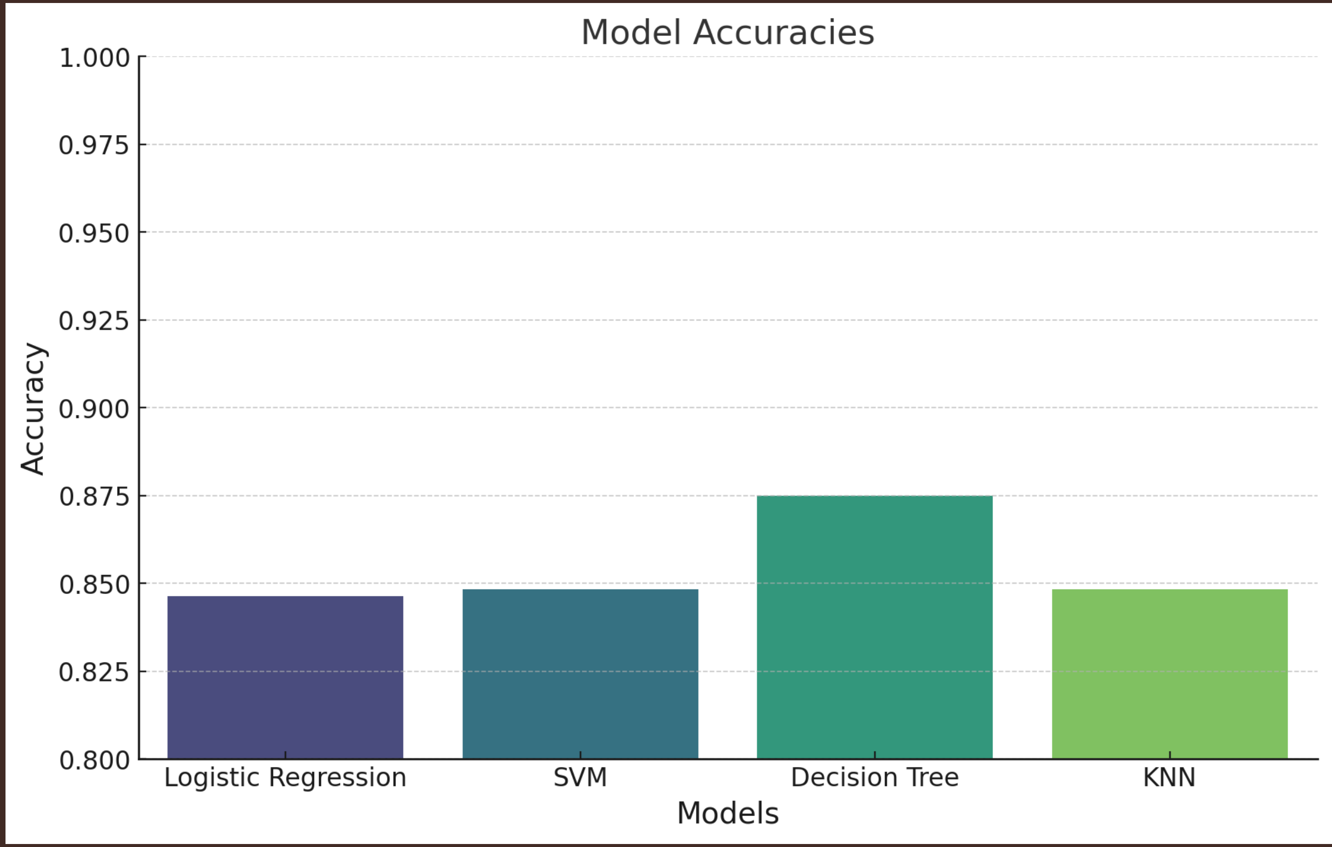


- The Payload vs. Launch Outcome scatter plot is necessary for analysing the relationship between the payload mass of various missions and their outcomes across different launch sites. By allowing users to filter payload sises this visualisation gives insights on how payload weight influences the success of launches. This helps in selecting optimal payload capacities to improve launch success rates.
- The findings from this scatter plot reveal patterns that may indicate thresholds or optimal ranges for payloads that correlate with successful launches. For instance, if a concentration of successful outcomes is observed within a certain payload range it suggests that staying within this range may enhance mission reliability.
- Conversely, clusters of failures at specific payloads could highlight the challenges/limitations during launches.

This information very useful for making decisions regarding payload specifications and launch strategies to maximise success rates in future missions.

Section 5
Predictive Analysis (Classification)

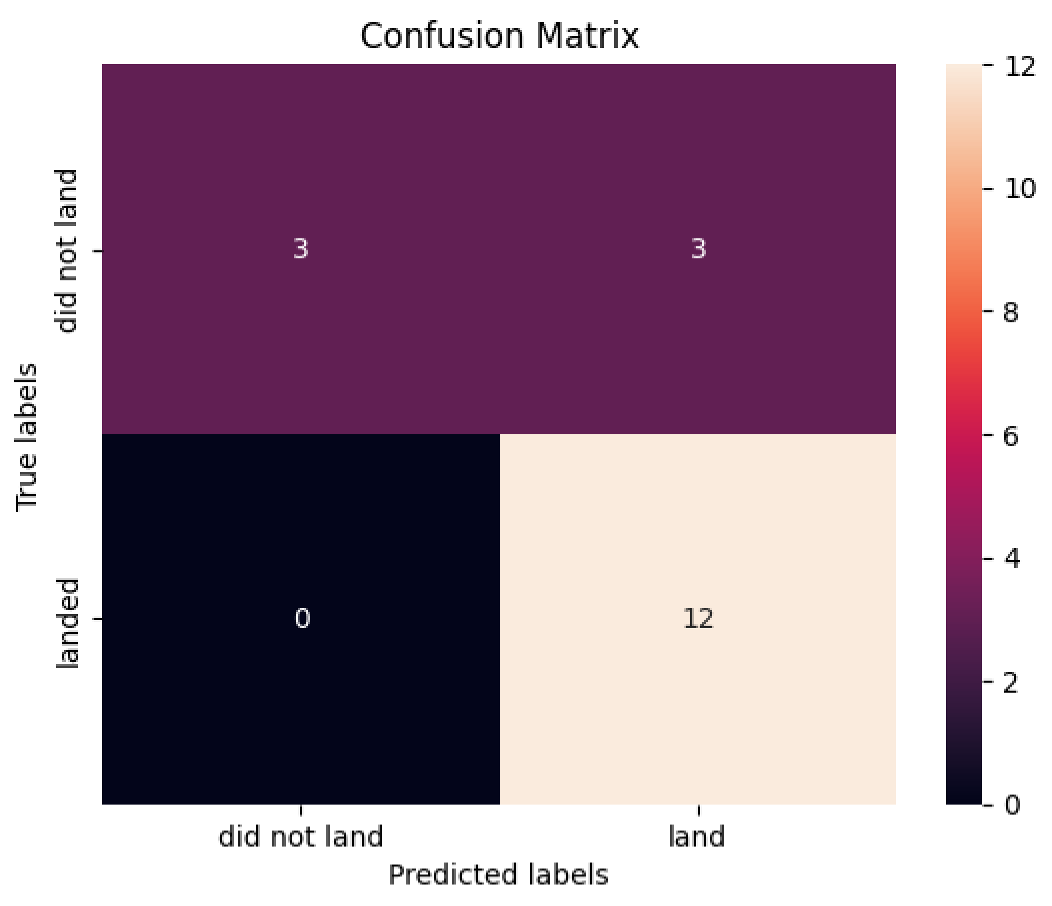
Classification Accuracy



- The bar plot visualises the accuracies of four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbours. The Decision Tree model stands out with the highest accuracy at 87.5%, outperforming the others at 84.6.

- The plot highlights the effectiveness of the Decision Tree algorithm in this analysis and shows for this specific type of data it is the best one to use. Although not all the possible algorithms were used, this fairly reudamentary algorithm is very useful in getting a quick and accurate result to predict the future missions. This can only get better as we provide more data

Confusion Matrix



- In the confusion matrix, there are 12 true positives, meaning the model correctly predicted that the rocket landed when it indeed did land. This means the model effectively identified instances of the positive class.
- However, there are 3 false positives, where the model incorrectly predicted landed when the true label was not landed. This highlights an issue with false positives, suggesting that the model may be over-predicting the positive class. This means more data is needed so this clarification can be made better by the algorithm.
- Overall, while the logistic regression model shows some capacity for distinguishing between classes, addressing the false positives is crucial for improving its accuracy and reliability.

Conclusion

This report has explored the core aspects of SpaceX's mission to improve the success of Falcon 9 rocket launches, particularly focusing on the first-stage landing. The discussion emphasises SpaceX's strategic focus on reusability, which reduces costs and boosts efficiency in space travel. By using historical data from the SpaceX REST API, the project analysed factors affecting landing success. This helps optimise future missions and supports resource allocation in a highly cost-sensitive and technically demanding field like space exploration.

The use of SQL queries to extract meaningful insights from the SpaceX data to show how critical data analysis is for mission planning. Specific queries were used to retrieve information on launch sites, payload masses etc to shed light on key performance indicators that influence launch success. Visualisations like pie charts and scatter plots helped to identify trends such as the relationship between payload masses and landing outcomes. These allowed decision-makers to quickly assess the performance of launch sites and strategies which is essential for improving safety and reliability in the future.

Machine learning models are also applied to predict the success of Falcon 9 landings based on various mission parameters. By evaluating algorithms like Logistic Regression, SVM, Decision Trees, and K-Nearest Neighbours we cam accurately forecast mission success for better risk management. Moreover, proximity to transportation networks, infrastructure development are considered in the strategic planning of launch sites. Data-driven decisions ensure that SpaceX continues to optimise emergency responses, while also addressing concerns such as noise and environmental impacts around launch sites.

In conclusion, this analysis demonstrates how SpaceX leverages historical data and machine learning to predict and enhance the success of future missions. By focusing on variables like payload mass, launch conditions etc. SpaceX can improve mission planning and ensure that its reusable rocket technology continues to lead the aerospace industry. This data-driven approach reduces costs and enhances the safety and efficiency of future space missions, to push the boundaries of space exploration and achieve future goals like interplanetary travel.

Appendix

Full GitHub Link - Location of all necessary files which is needed in this project. Consult the README.md for more information.

Historical SpaceX Data Used - Location of historical SpaceX data used in this project.

Analysed Wiki Article – The link to the Wiki in which data from tables was gathered before doing Exploratory data analysis.


THANK YOU