IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anirudh Kunte
26th July 2024

Outline

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

Executive Summary

Summary of Methodologies - This research aims to identify the key factors that contribute to a successful rocket landing. The following methodologies were employed:

- Data Collection: Utilized SpaceX REST API and web scraping techniques to gather relevant data.
- Data Wrangling: Processed the data to create a variable indicating success or failure outcomes.
- Exploratory Data Analysis (EDA): Used data visualization techniques to examine factors such as payload, launch site, flight number, and yearly trends.
- SQL Analysis: Calculated key statistics such as total payload, payload ranges for successful launches, and the number of successful and failed launches.
- Geographical Analysis: Assessed launch site success rates and their proximity to geographical markers.
- Visualization: Mapped launch sites with the highest success rates and analyzed successful payload ranges.
- Predictive Modeling: Built models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) algorithms.

Results
Exploratory Data Analysis: • Identified KSC LC-39A as the launch site with the highest success rate. • Determined that orbits ES L1, GEO, HEO, and SSO have a 100% success rate. • Observed an improvement in launch success rates over time.
Geospatial visualization - Most launch sites are situated near the equator and all are in close proximity to the coast.
Predictive modeling - All predictive models showed similar performance on the test set, with the decision tree model slightly outperforming the others.

Introduction

Project background and context

- SpaceX, a pioneer in the aerospace industry, advertises the cost of a Falcon 9 rocket launch at 62 million dollars, which is significantly lower than the 165 million dollars charged by other providers. This substantial cost reduction is primarily due to SpaceX's innovative ability to reuse the rocket's first stage.
- The objective of this project was to predict the success of the first stage landing of the SpaceX Falcon 9 rocket. Accurate predictions would enable a competing startup to make more strategic and informed bids against SpaceX for rocket launches. By determining the likelihood of a successful first stage landing, it becomes possible to estimate the cost of a launch, providing crucial information for alternate companies bidding against SpaceX.

Problems to Address

- How do payload mass, launch site, number of flights, and orbits impact the success of the first-stage landing?
- What is the rate of successful landings over time?
- How do we effectively predict a successful landing (binary classification) using past data?

Section 1

Methodology

Methodology

Data Collection

Collected data using SpaceX REST API and web scraping techniques.

Data Wrangling

- Filtered data to retain only relevant records (e.g., Falcon 9 launches).
- Handled missing values by imputing or removing them.
- Applied one-hot encoding to categorical variables to prepare the data for analysis and modeling.

Exploratory Data Analysis (EDA)

- Explored the data using SQL queries to derive insights.
- Performed data visualization techniques to identify trends and patterns.

Data Visualization

- Created interactive visualizations using Folium and Plotly Dash.
- Mapped launch sites and analyzed their proximities and success rates.

Methodology

Predictive Modeling

- Built classification models to predict landing outcomes.
- Models used: Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN).
- Tuned and evaluated models to find the best-performing model and parameters.

Model Evaluation

- Evaluated model performance using accuracy, precision, recall, and F1-score.
- Selected the best model based on evaluation metrics and performance on the training and test set.

Data Collection

Data was collected from the following sources:

- SpaceX API: https://api.spacexdata.com/v4/launches
- Web scraping the Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

Import Libraries and Define Functions

- Imported required libraries.
- Defined helper functions for data extraction.

Request Data from SpaceX API

- Made GET requests to retrieve launch data from the SpaceX API: https://api.spacexdata.com/v4/launches
- Extracted relevant information from the API responses.

Data Cleaning and Formatting

- Filtered data to retain only Falcon 9 launches.
- Reset the FlightNumber column post-filtering.

Data Wrangling

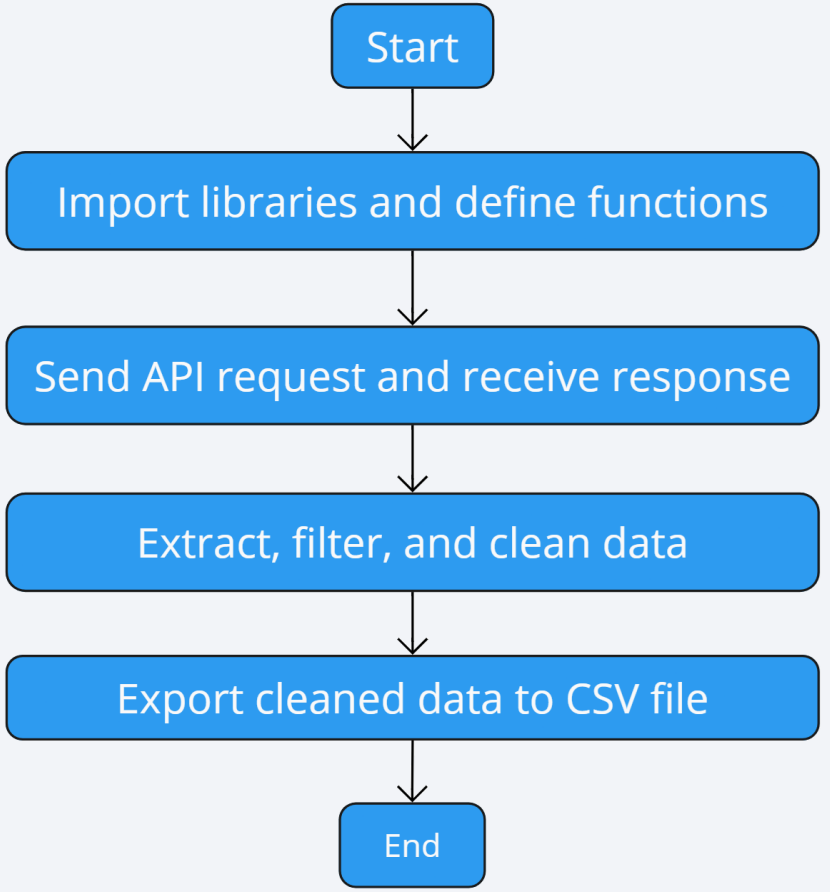
- Handled missing values in the dataset.
- Replaced np.nan values in PayloadMass with the mean.

Export Cleaned Data

- Exported the processed data to a CSV file for subsequent analysis.

Data Collection – SpaceX API

- API Request: Sent GET requests to the SpaceX API using response = requests.get('https://api.spacexdata.com/v4/launches')
- Response Handling: Received JSON responses containing launch details using response.json()
- Filtered the dataset to retain only Falcon 9 launches using Pandas
- Identified and handled missing values in the dataset using df.isnull() and df.fillna().
- GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX-Data-Collection-API.ipynb



Data Collection – Scraping

Import Libraries and Define Functions

- Imported necessary libraries.
- Defined helper functions for processing HTML tables.

Target URL Identification

Targeted the Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Fetch and parse HTML Content

- Used requests to retrieve the HTML content of the Wikipedia page.
- Utilized BeautifulSoup to parse the HTML content.

Extract Launch Records Table

Identified and extracted the HTML table containing Falcon 9 and Falcon Heavy launch records.

Convert to DataFrame

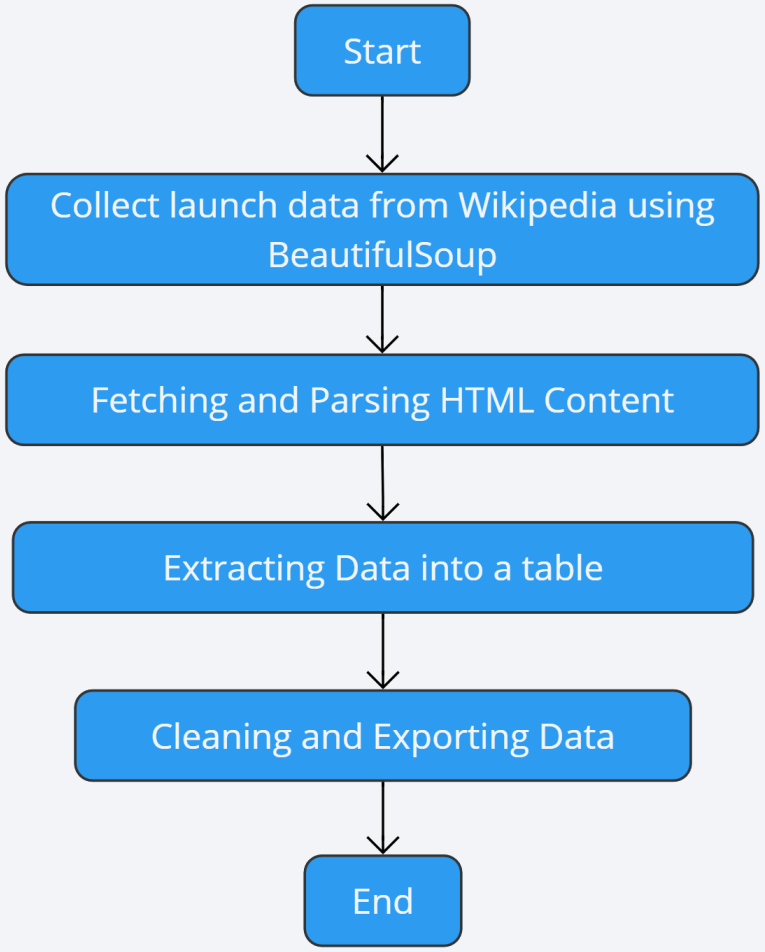
Parsed the HTML table and converted it into a Pandas DataFrame.

Clean and Export Data

- Handled annotations, missing values, and inconsistent formatting in the extracted data.
- Exported the cleaned data to a CSV file for subsequent analysis.

Data Collection - Scraping

- Fetching and Parsing HTML Content:
  - Retrieved HTML content using response = requests.get(url).
  - Parsed content with soup = BeautifulSoup(response.content, 'html.parser').
- Extracting Data and creating dataframe:
  - Extracted the HTML table containing launch records using soup.find_all()
  - Converted HTML table to DataFrame using pd.read_html()
- Cleaning and Exporting Data:
  - Handled missing values and formatting using df.replace() and df.dropna().
  - Exported cleaned data to CSV using df.to_csv()
- GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX-Webscraping.ipynb



Data Wrangling

Data Loading:
Loaded data into a Pandas DataFrame.

Data Cleaning:
Checked and handled missing values.

Feature Engineering:
Created new columns and transformed existing ones.
Example: df['PayloadMass_kg'] = df['PayloadMass'] * 1000

Landing Outcome Encoding:
Converted landing outcomes into binary values.
Example: landing_outcome_dict = {'True Ocean': 1, 'False Ocean': 0, 'True RTLS': 1, 'False RTLS': 0, 'True ASDS': 1, 'False ASDS': 0}
df['LandingOutcome'] = df['LandingOutcome'].map(landing_outcome_dict)

Export Cleaned Data:
Exported cleaned data to a new CSV file.

GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX-Data%20wrangling.ipynb

Landing Outcomes Data

True Ocean: Successful landing to a specific region of the ocean

False Ocean: Unsuccessful landing to a specific region of the ocean.

True RTLS: Successful landing on a ground pad.

False RTLS: Represented an unsuccessful landing on a ground pad.

True ASDS: Indicated a successful landing on a drone ship.

False ASDS: Unsuccessful landing on a drone ship.

Outcomes Conversion: Converted landing outcomes into binary values: 1 for all successful landings and 0 for all unsuccessful landings.

EDA with Data Visualization

Data Loading:
Loaded cleaned data into a Pandas DataFrame.

Exploratory Data Analysis (EDA):
Performed initial data exploration.

Data Visualization:
Scatterplots for viewing the relationship, Bar charts for showing comparison among discrete categories, and Trendline (time series) for change over time

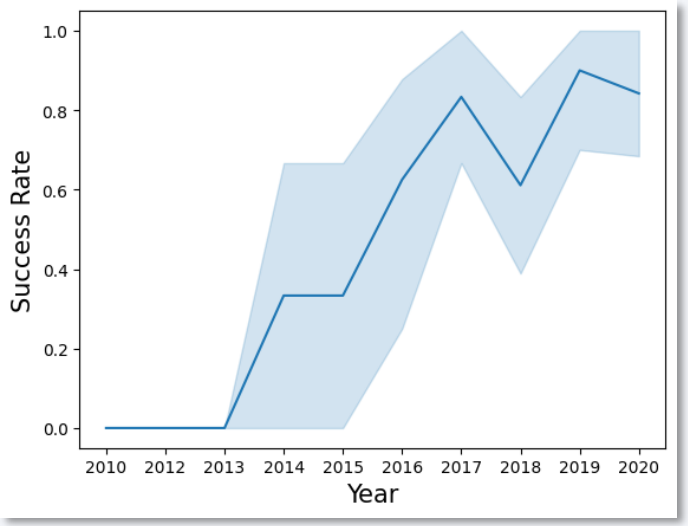
Key Findings:
Observed trends such as improved launch success rates over time and the impact of factors such as payload on launch success.

Export Visualizations:
Saved plots for reporting.

GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX-EDA-with-DataVisual.ipynb

Key Data Visualizations:

- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Success Rate vs. Orbit Type
- Flight Number vs. Orbit Type
- Payload vs. Orbit Type
- Yearly Trend of Launch Success



EDA with SQL

Key Insights from SQL

- Unique Launch Sites: Identified distinct launch sites.
- Launch Sites Starting with 'CCA': Filtered launch sites with specific criteria.
- Total Payload by NASA Boosters: Calculated total payload carried by NASA.
- Average Payload by Booster Version: Determined average payload for specific booster versions.
- First Successful Ground Landing Date: Found the earliest date of successful ground landings.
- Successful Drone Ship Landings: Identified successful drone ship landings with specific payload ranges.
- Total Successful and Failed Missions: Counted the number of successful and failed missions.
- Booster with Maximum Payload: Identified the booster carrying the maximum payload.
- Failed Landing Outcomes in 2015: Analyzed failed landings in 2015.
- Ranking Landing Outcomes: Ranked landing outcomes by frequency.

GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX-EDA-with-SQL.ipynb

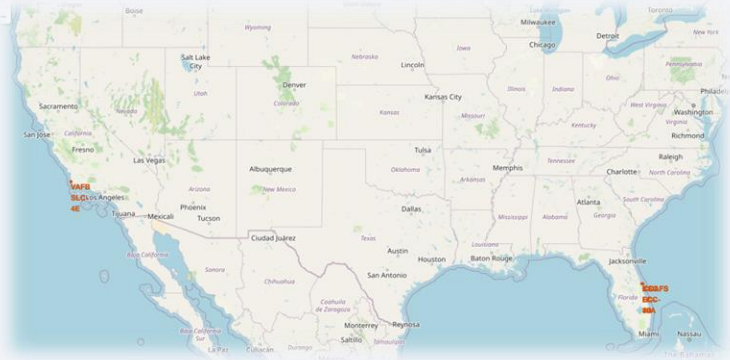
Build an Interactive Map with Folium

Key steps for Geospatial Visualization

Data Loading: Load cleaned data into a DataFrame.

Setup Folium Map: Initialize the map centered on launch sites.

Add Launch Sites Markers: Mark launch sites with relevant information



Add Circles for Launch Outcomes: Added circles to represent the success or failure of each launch. **Green** for successful and **Red** for unsuccessful launch.

Add Lines for Launch Paths: Added lines to visualize the paths from the launch sites to the landing outcomes**.**

GitHub URL to the notebook: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/SpaceX_locations_analysis_folium.ipynb

Build a Dashboard with Plotly Dash

Dropdown Menu:

Allows users to select different launch sites to filter the data.

Pie Chart:

Displays the total success launches by site.

Slider of Payload Mass Range

Allow user to select payload mass range

Scatter Plot:

Shows the correlation between payload and launch success.

Github URL to the .py file: https://github.com/anirudh-ak/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data Loading: Load cleaned data into a DataFrame.

Data standardization: Standardize the data with StandardScaler. Fit and transform the data.

Data Splitting: Split data into training and testing sets.

Model Building: Build and train classification models (Logistic Regression, SVM, Decision Tree, KNN).

Model Evaluation: Evaluate models using accuracy, precision, recall, and F1-score.

Hyperparameter Tuning: Optimize model performance using GridSearchCV.

Best Model Selection: Choose the best-performing model.

Results

Exploratory Data Analysis

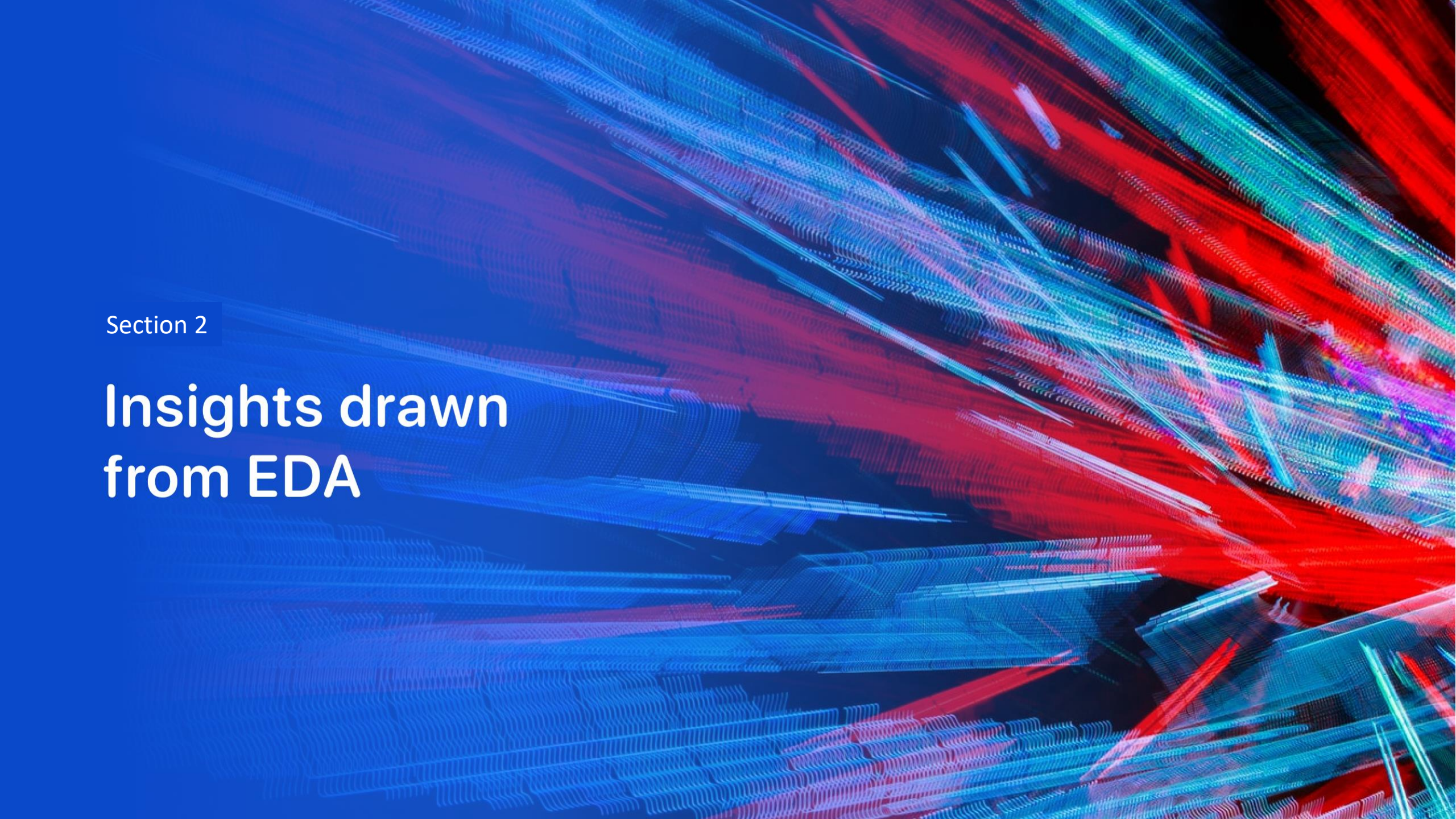
- Launch success has improved over time.
- KSC LC-39A has the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Visual Analytics

- Most launch sites are near the equator and close to the coast.
- Launch sites are strategically located far from cities, highways, and railways to avoid damage from failed launches, yet close enough to support launch activities.

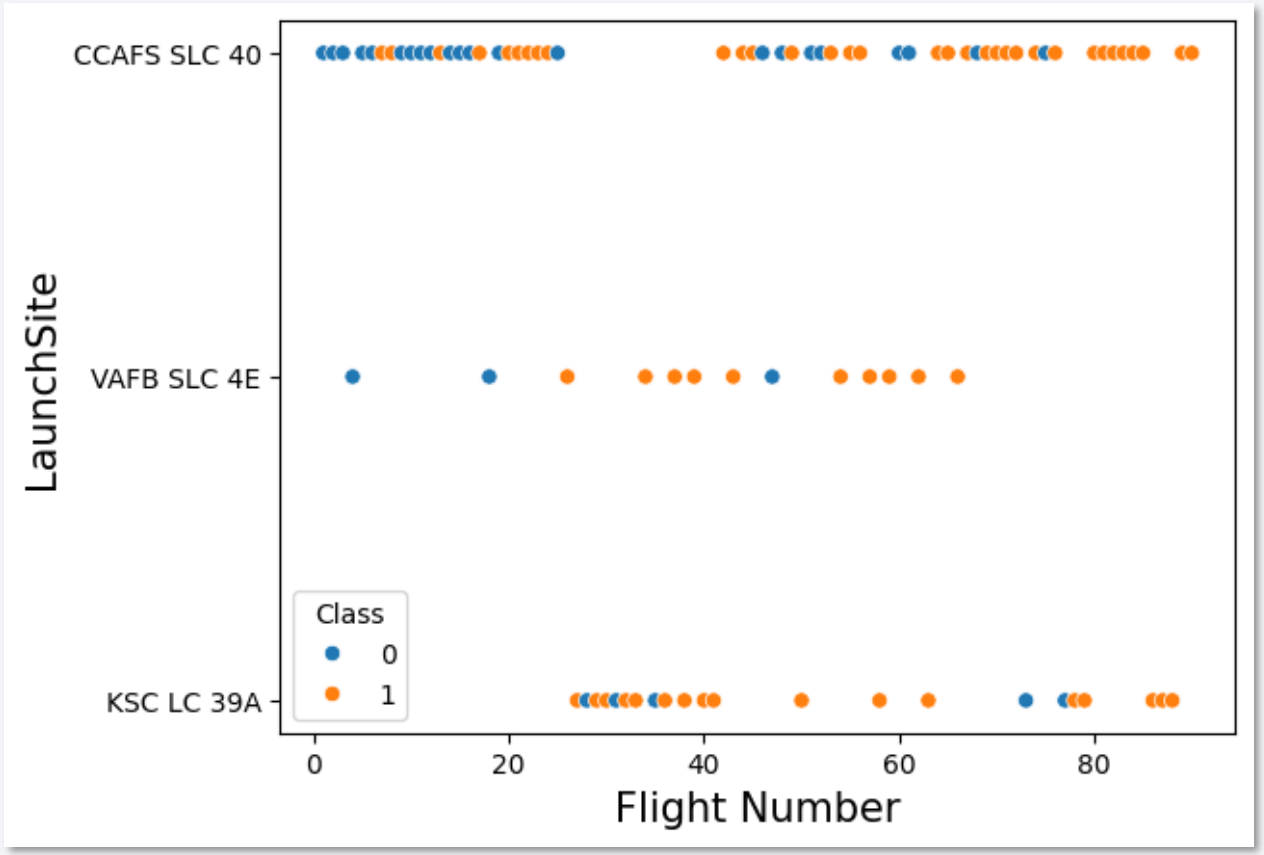
Predictive Analytics

It is difficult to say which model is best for the dataset, since all models except the decision tree have the same accuracies. However, the Decision Tree model has a higher training accuracy but still a low testing accuracy.


Section 2
Insights drawn from EDA

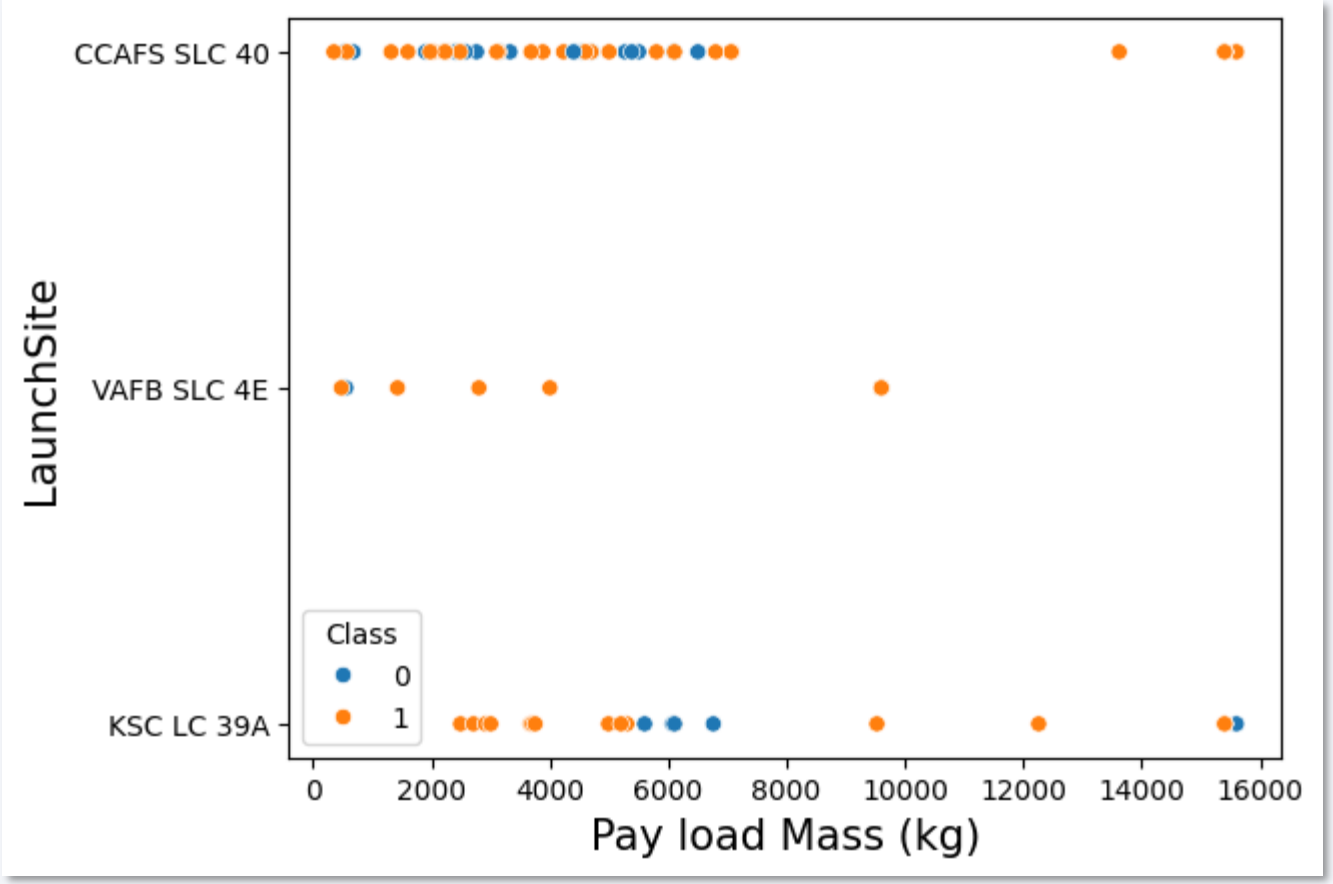
Flight Number vs. Launch Site

- CCAFS SLC 40 is the launch site with the highest number of successful launches.
- However, the sites VAFB SLC 4E and KSC LC 39A have a higher success rate.
- It is evident that newer flights have a higher rate of success



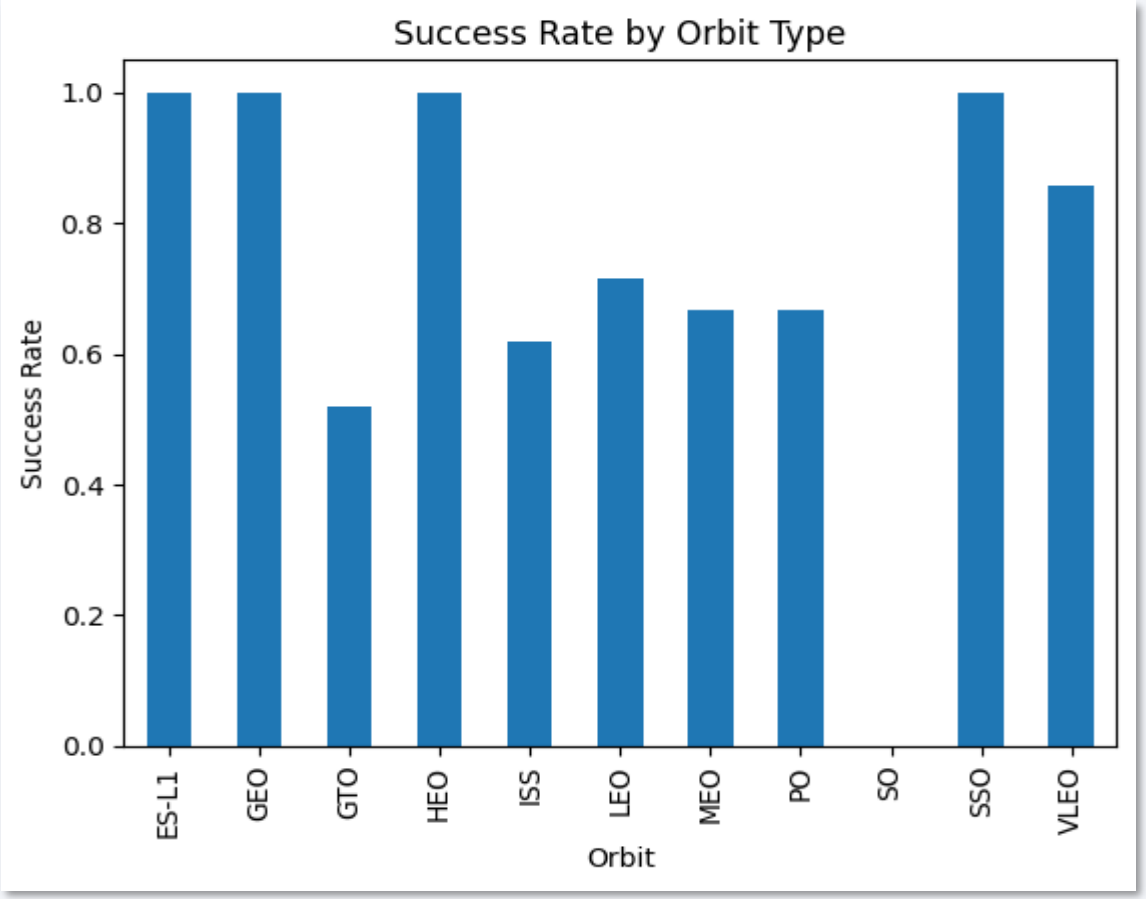
Payload vs. Launch Site

- There seems to be a higher success rate for launches with payloads under 5000 kg at CCAFS SLC 40.
- VAFB SLC 4E, despite fewer launches, shows a higher success rate
- KSC LC 39A has a balanced mix, indicating that this site handles a wide range of payloads with varying success.
- There are few data points for payloads above 10000 kg, indicating fewer launches with very high payloads, but these tend to be successful.



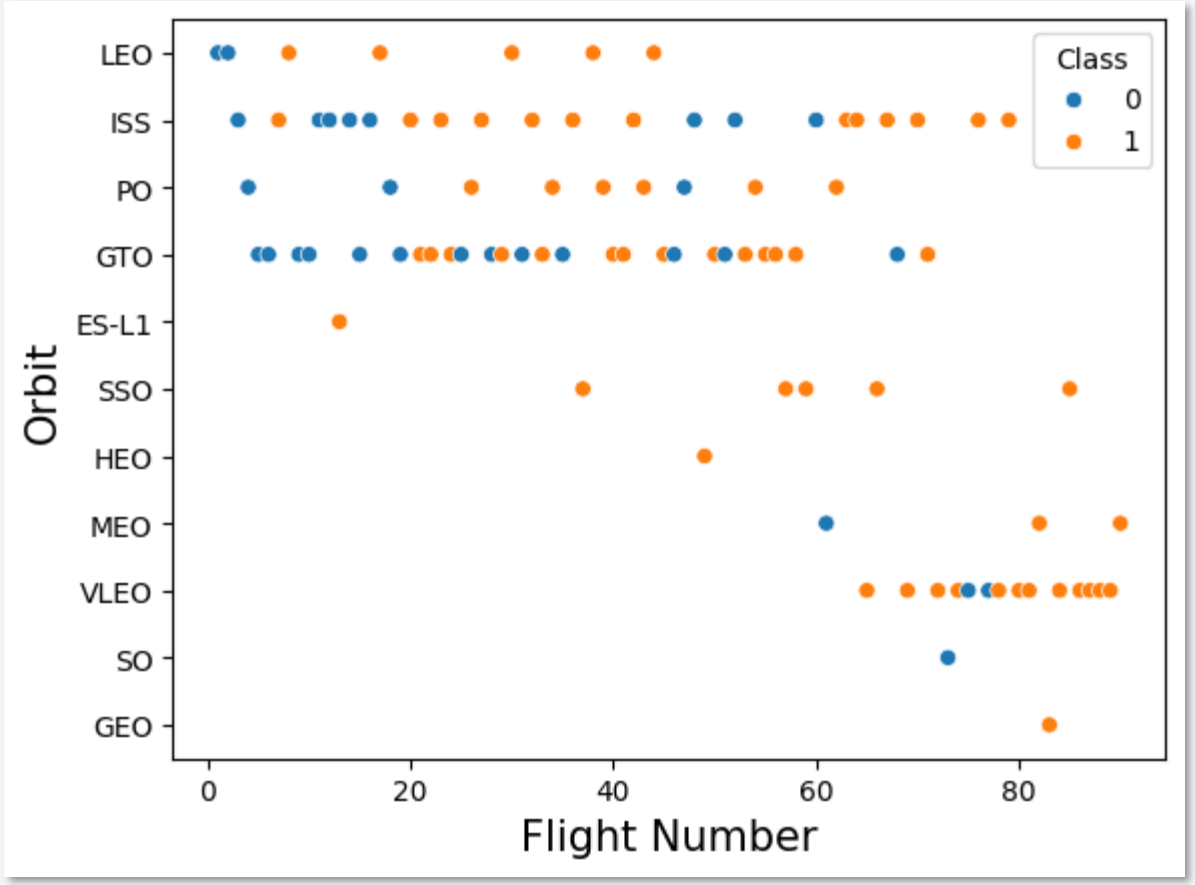
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate. These orbits consistently result in successful missions.
- Orbits ISS, LEO, MEO, PO, and VLEO have moderate success rates, ranging from around 60% to 80%.
- The orbit GTO has a significantly lower success rate compared to other orbits, indicating potential challenges or higher risks associated with missions to this orbit.



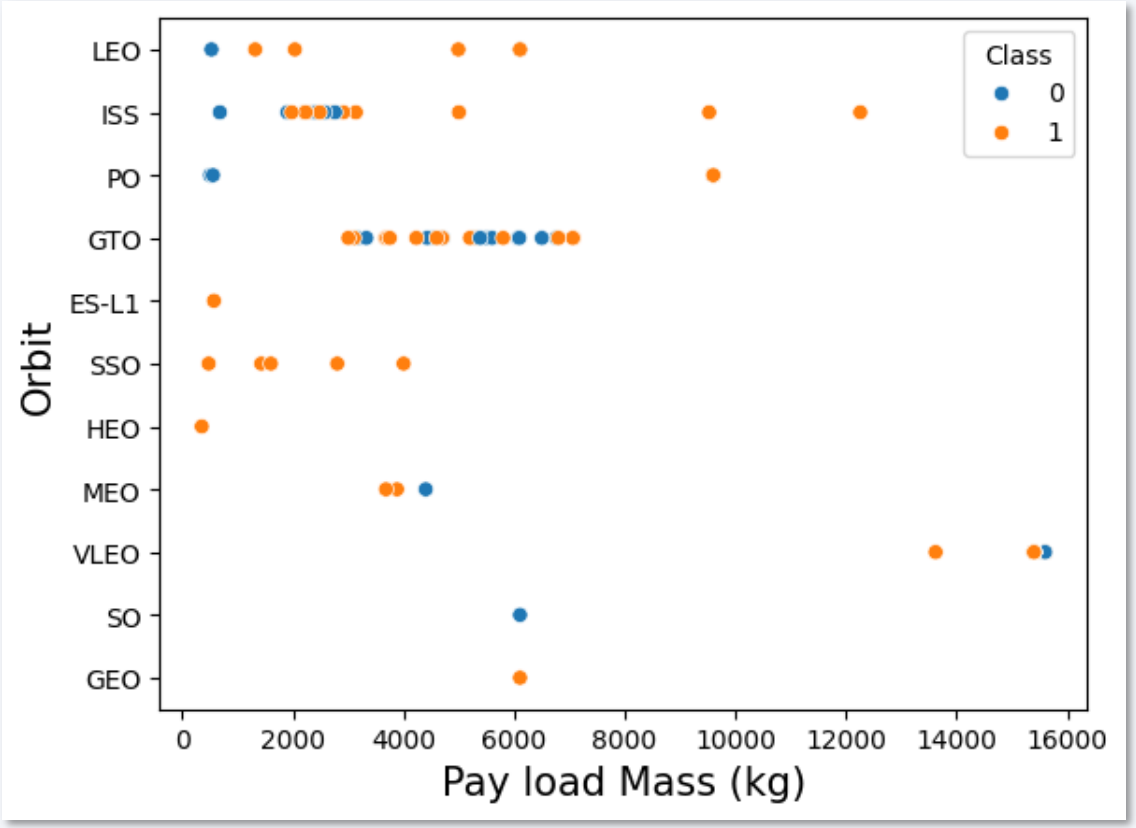
Flight Number vs. Orbit Type

- Later launches generally show more successful landings, especially noticeable in orbits like LEO and ISS.
- Orbits like GEO, HEO, and ES-L1 have higher success rates (more orange dots) even with fewer flights.
- GTO and PO show a mix of successes and failures across various flight numbers.
- Orbits such as SSO and VLEO show consistent success rates across different flight numbers, with fewer failures.
- LEO and ISS, which have a higher density of launches, display a mix of successes and failures, indicating variability in success rates.



Payload vs. Orbit Type

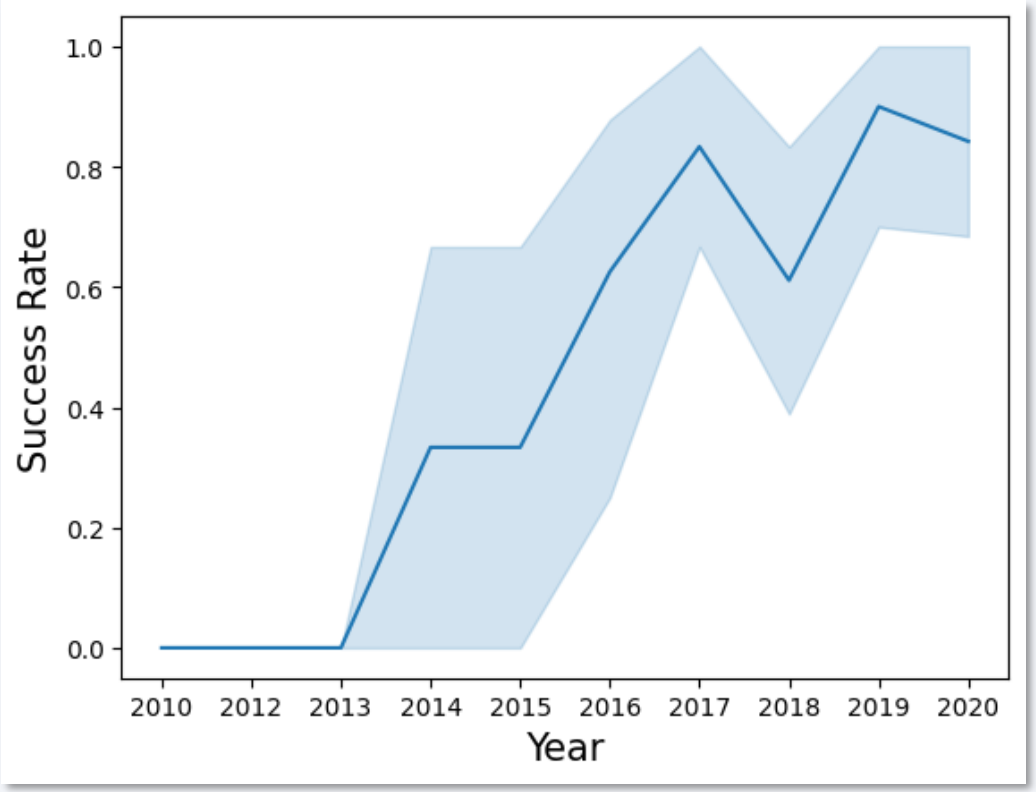
- Higher payload masses (above 6000 kg) tend to have higher success rates across multiple orbits, particularly noticeable in GTO and GEO orbits.
- Orbits such as LEO, ISS, and PO show a mix of successes and failures in the 0-6000 kg payload range.
- SSO, HEO, and ES-L1 orbits demonstrate high success rates across varying payload masses, indicating reliability regardless of payload size.
- GTO orbit shows mixed results across a wide range of payload masses, suggesting variability in success rates.



- There are a few outliers with extremely high payloads (above 14000 kg) that are successful, indicating that large payloads can still achieve successful landings under the right conditions.

Launch Success Yearly Trend

- Success rate was very low initially, with little to no successful launches.
- Success rate started to improve gradually, reaching around 40% by 2015.
- A sharp increase in success rate is observed after 2015, reaching nearly 80% by 2017.
- Indicates significant improvements in technology and operations during this period.
- Some fluctuations in the success rate during 2017-2019, but generally remained high.
- Success rate peaked around 90% in 2019 before dropping slightly. Indicates consistent and reliable performance in 2019-2020



All Launch Site Names

- There are 4 launch sites:

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

- Query: Selecting distinct on launch site column gives us this result

```
%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

Launch Site Names Begin with 'CCA'

- Using LIKE ‘CCA%’ in query we are able to find out launch sites starting 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

Total Payload Mass

- Calculating SUM on the payload mass and filtering on customer column as NASA (CRS) gives us the result of 45596 kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculating AVG on the payload mass column and filtering on the Booster version as F9 v1.1 gives us the result of 2928.4 kg

```
%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

First Successful Ground Landing Date

- Selecting MIN on the date column filtering on landing outcome as “Success (ground pad)” gives the result as follows:

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Selecting Booster version column filtered on landing outcome column as “Success (drone ship)” and on payload mass column between 4000 and 6000 gives us the answer as follows:

```
1 %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

Total Number of Successful and Failure Mission Outcomes

- Taking count of all mission outcomes using group by yields the following result:

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- Here we first need to find out the maximum value of the payload mass and then fetch the boosters that have carried that 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

We use a subquery to find the MAX of the payload column and then extract all the boosters that have carried that payload.

2015 Launch Records

- We take substrings of the date column to extract month and year, along with booster version and launch site column and filter on the year as 2015 and the landing outcome column as 'Failure (drone ship)' to get the following result:

```
%sql SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version,
Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND
substr(Date,0,5)='2015';
```

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

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

- We count the Landing outcome column after filtering on the date and grouping on Landing outcome column. Then we arrange the results in descending order:

```
%sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

Landing_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

Section 3

Launch Sites Proximities Analysis

All Launch sites

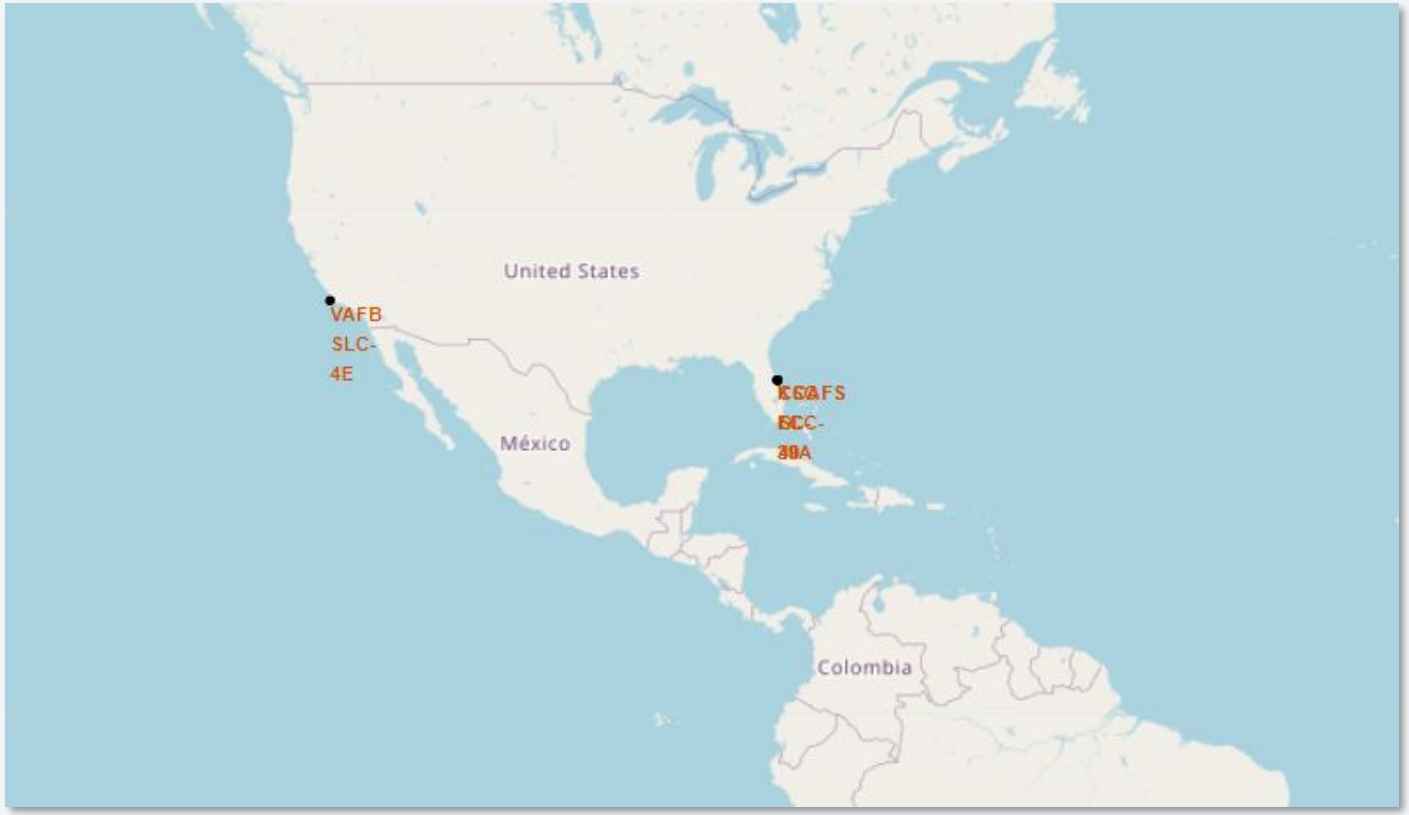
The launch sites are on the East and West coast of the United States.

They are closer to the sea and the equator

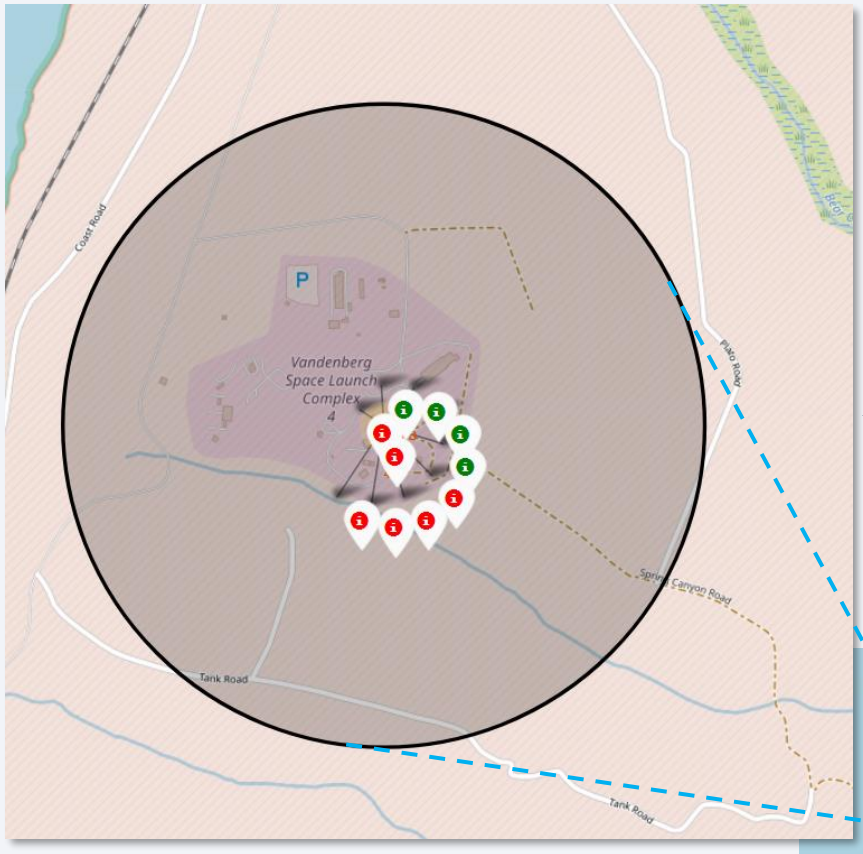
This makes sense because it is much easier to launch rockets from the equator because of:

- Ease of launching to equatorial orbit
- Additional natural boost due to the rotational speed of the earth; saves extra fuel and booster cost

The sites are also close to roads and railroads

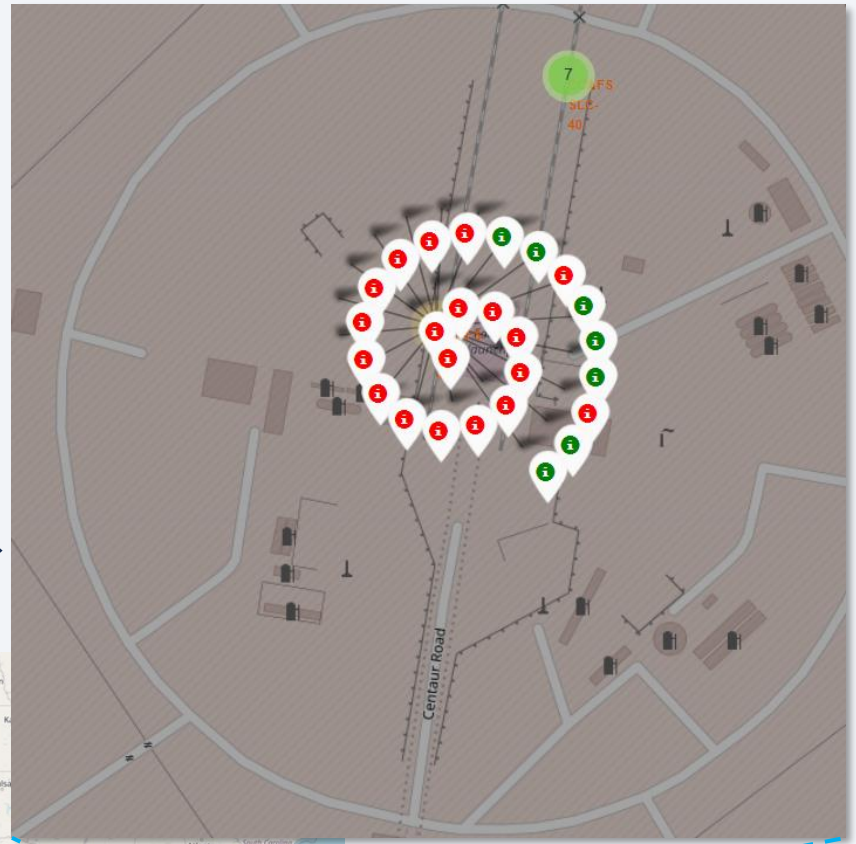


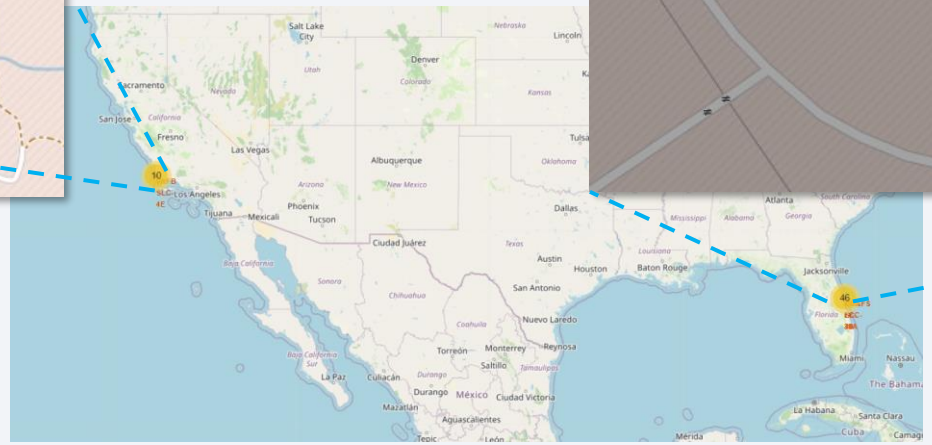
Launch sites with outcomes



VAFB SLC-4E

CCAFS LC-40





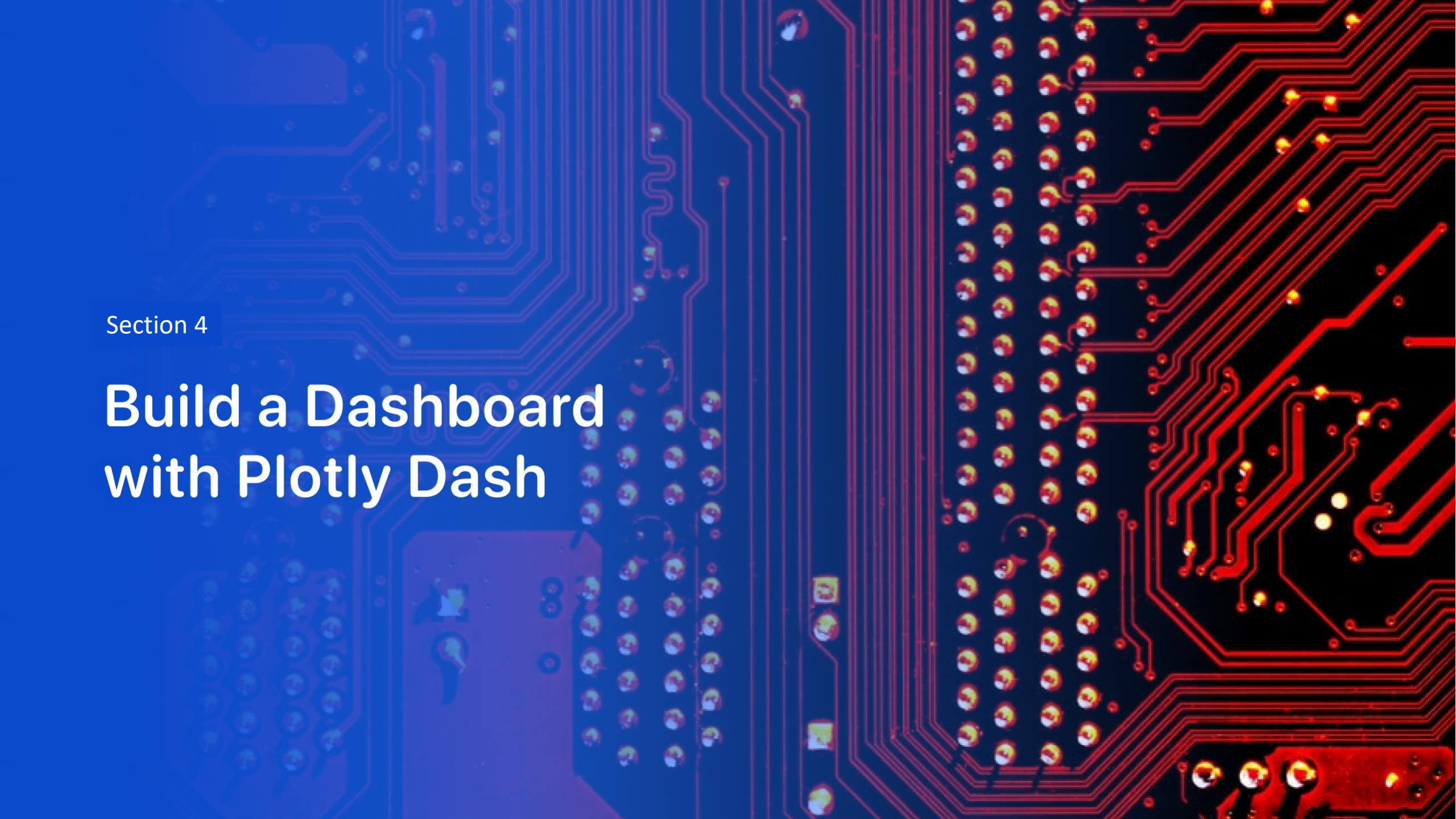
Green markers indicate successful launches and Red ones indicate failed ones

Distances of launch sites from landmarks

Distance of launch site from railway	15.23 km
Distance of launch site from highway	20.28 km
Distance of launch site from nearest coastline	6.33 km
Distance of launch site from city	16.32 km

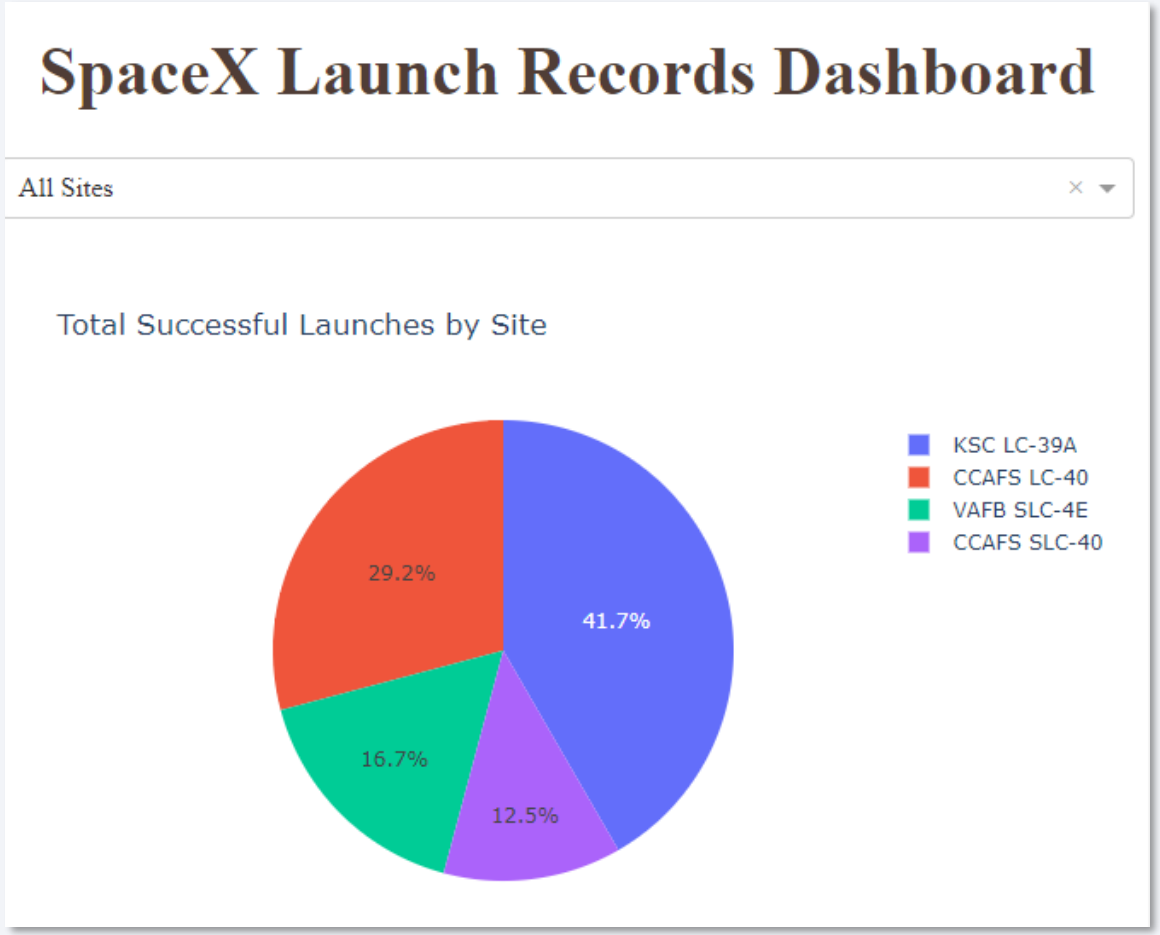
Close to coastline: To ensure that spent stages or failed launches drop along the launch path, avoiding populated areas to prevent accidents.

Away from transportation/Infrastructure and Cities: Sites need to be distant enough to avoid damage from failed launches but still close to roads, rails, and docks to facilitate transportation of people and materials for launch activities.


Section 4
Build a Dashboard with Plotly Dash

Total successful launches by site

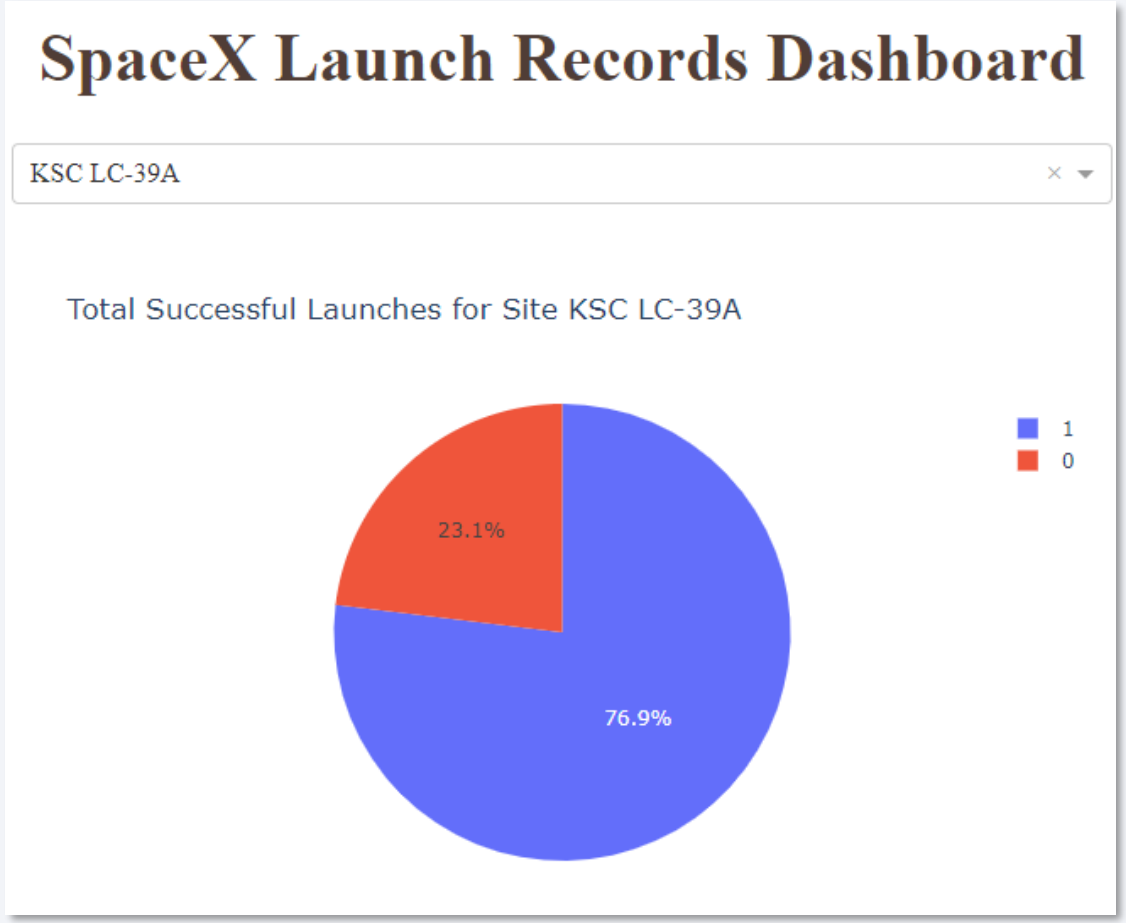
KSC LC-39A has the **most successful launches** amongst all launch sites (**41.2%**)



Launch Success (KSC LC-39A)

The **highest success rate** amongst all launch sites is exhibited by **KSC LC-39A** (**76.9%**)

10 successful launches and 3 failed launches



Payload mass and success of launch

Class 1: Success
Class 0: Failure

- Successful launches are predominantly concentrated in the mid-range payload masses (around 2000 to 6000 kg).
- Failures are more scattered across the lower payload masses, particularly below 2000 kg.
- FT and B5 booster versions show a higher frequency of successful launches across a broader range of payload masses.

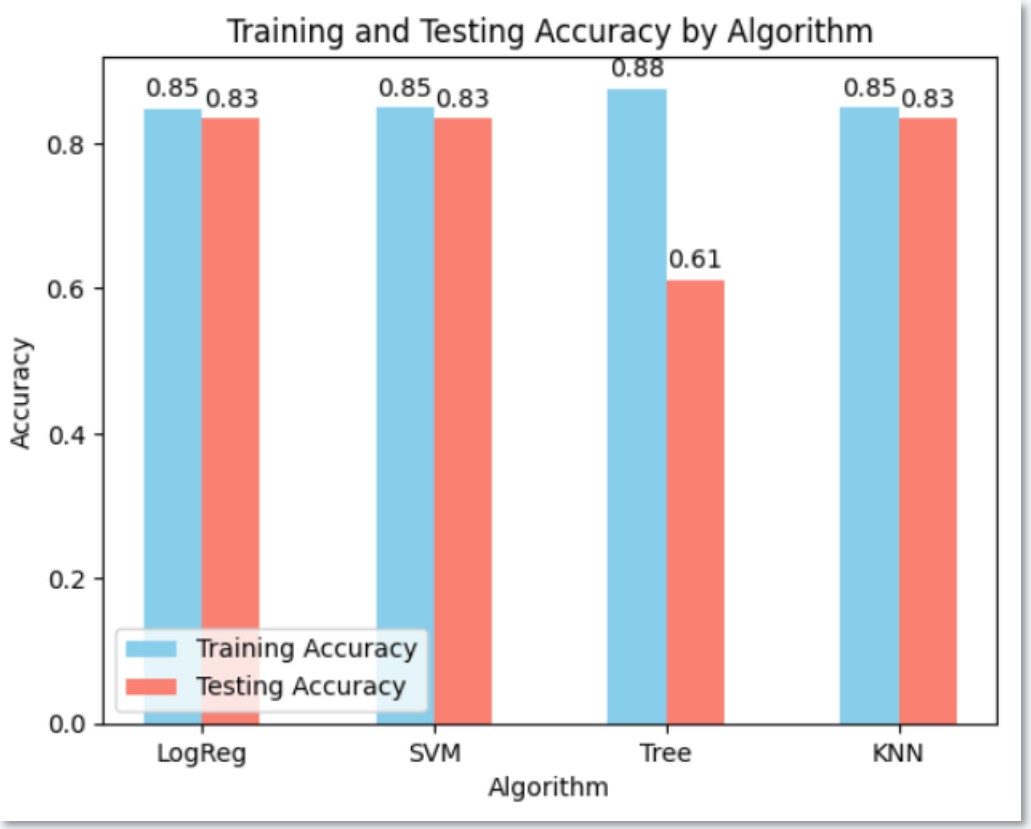
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Created a bar chart for comparison of the training and testing accuracy figures.
- All models have same training and testing accuracies except decision tree. It has a higher training accuracy and a lower testing accuracy

	LogReg	SVM	Tree	KNN
Accuracy	0.833333	0.833333	0.611111	0.833333
F1 Score	0.888889	0.888889	0.720000	0.888889
Precision	0.800000	0.800000	0.692308	0.800000
Recall	1.000000	1.000000	0.750000	1.000000



Confusion Matrix

Performance

High True Positives (12) and Zero False Negatives (0):

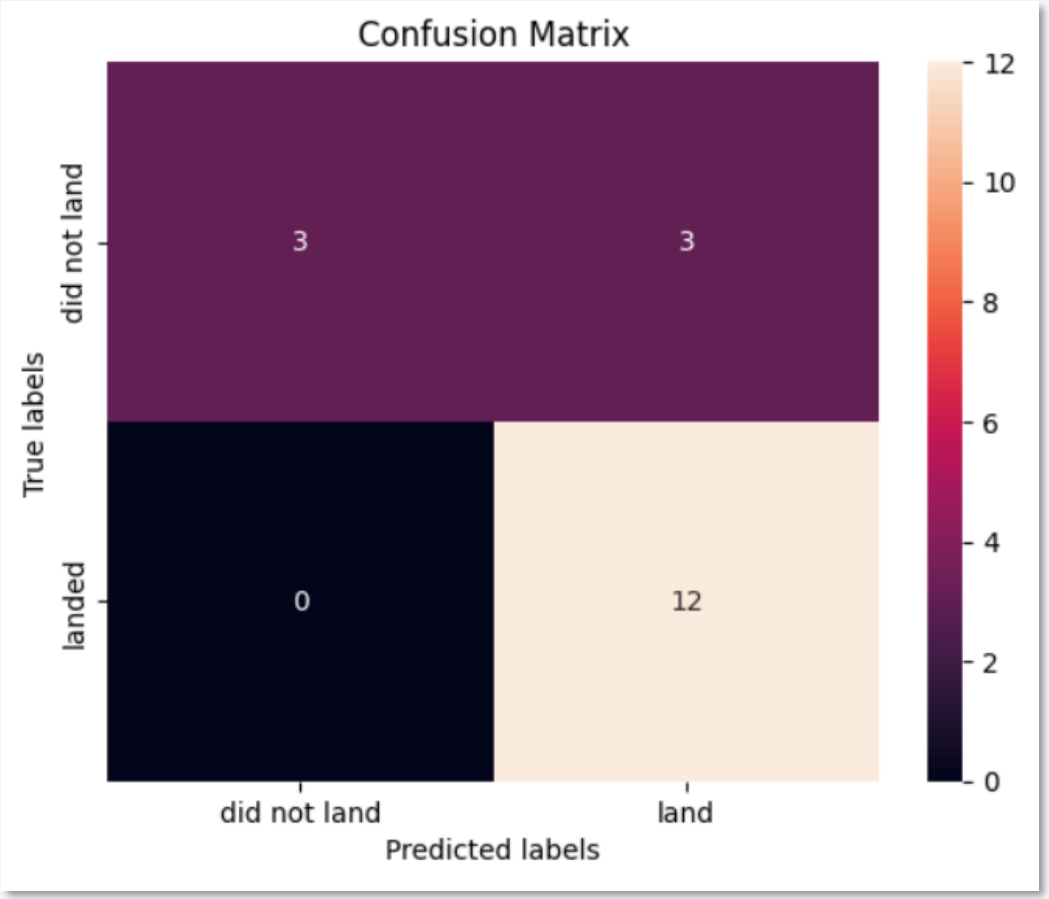
Indicates the model is highly effective at correctly predicting landings.

Presence of False Positives (3):

Indicates the model has some inaccuracies in predicting landings for launches that did not actually land.

Precision: 12/(12+3) = 0.8

Recall: 12/(12+0) = 1.0



Conclusions

Model Performance
All the models performed similarly on the train and test set with the decision tree performing slightly better on the training set and slightly worse on the test set

Launch Sites
- Equator: Launch sites are predominantly located near the equator, leveraging Earth's rotational speed for a natural propulsion boost, thereby reducing the need for extra fuel and boosters.
- Coast: All launch sites are situated along the coast.
- KSC LC-39A: KSC LC-39A boasts the highest success rate among all launch sites, achieving a 100% success rate for payloads under 5,500 kg.

Launch Success Trends
Launch success rates have improved consistently over time.

Orbits
High Success Orbits: Orbits such as ES-L1, GEO, HEO, and SSO have maintained a 100% success rate.

Payload Mass
Correlation with Success: Generally, across all launch sites, an increase in payload mass correlates with a higher success rate.

Appendix - Observations

- It has been observed that a larger dataset can help with better predictions as it will be able to capture more pattern related to the launch success.
- Ensemble models such as XGBoost were not used in this exercise. They perform way better on classification tasks and can be explored.
- Some techniques to gauge feature importance, such as SHAP can be used to further improve the accuracy.

Thank you!