IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nafees
May 2025

Outline

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

Executive Summary

Summary of methodologies

This project aimed to predict the success of Falcon 9 first stage landings using machine learning. It involved data collection from APIs and web scraping, data cleaning, visualization, and modeling.

Summary of all results

Data Insights:

- **Identified key factors influencing the successful landing of the Falcon 9 first stage.**
- **Visualized geographical patterns and success rates.**

Model Performance:

- **Decision Tree: 94.44% accuracy.**
- **SVM and K-Nearest Neighbors: 83.33% accuracy.**

Key Findings:

- **Launch site and payload mass play a significant role in landing success.**
- **The Decision Tree model proved to be the most effective.**

Introduction

Project background and context

SpaceX reduces launch costs by reusing rocket stages. Our goal is to identify factors influencing successful landings and develop accurate prediction models. By accurately predicting landing success, we can estimate launch costs and provide valuable insights for companies.

Problems you want to find answers

- What factors impact the successful landing of the Falcon 9 first stage?
- How can machine learning models accurately predict the outcome of the landing?
- Which machine learning model is most effective at predicting landing success?

Section 1

Methodology

Methodology

Data collection methodology:

- Launch data was sourced from the SpaceX API and Wikipedia. APIs provided technical details; web scraping retrieved historical launch records.

Perform data wrangling

- Cleaned and merged datasets.
- Missing values handled.
- New features engineered.
- Standardization ensured for analysis.

Perform exploratory data analysis (EDA) using visualization and SQL

- EDA was performed using Matplotlib and Seaborn to uncover trends. SQL queries were used for additional insights. Charts included histograms, bar plots, and heatmaps.

Methodology

Perform interactive visual analytics using Folium and Plotly Dash

- Folium maps illustrated launch sites and outcomes.
- Plotly Dash enabled dynamic exploration of launch success based on payload and location.

Perform predictive analysis using classification models

- Built classifiers: Logistic Regression, SVM, KNN, and Decision Trees.
- Used GridSearchCV for hyperparameter tuning and cross-validation for evaluation.

Data Collection

- Data collection process involved a combination of API requests from SpaceX RESTAPI and Web Scraping data from a table in SpaceX's Wikipedia entry.
- We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Flow for Data Collection

RestApi:

Web Scraping:

Data Collection – SpaceX API

Initiate API Request

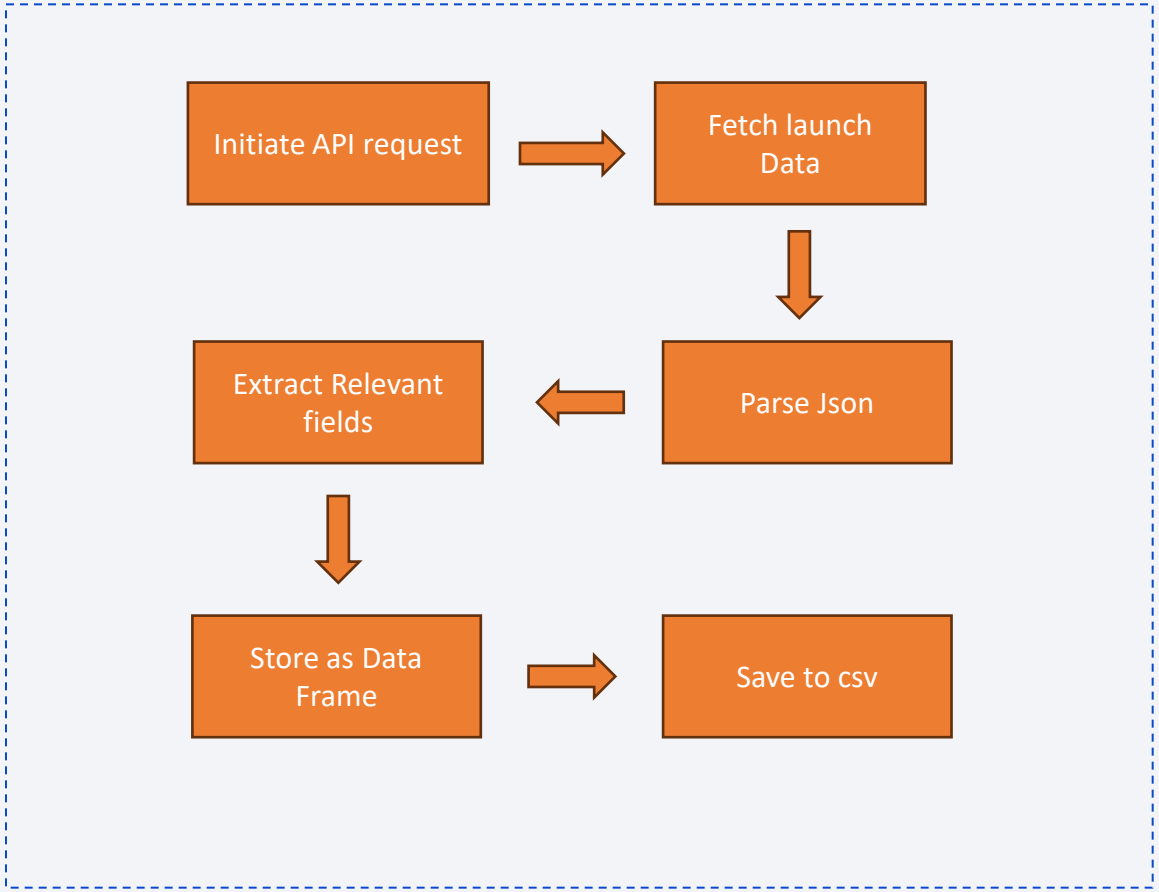
- Use Python's requests library to connect to the SpaceX API.

Parse API Response

- Convert API response from JSON to a Python dictionary.
- Extract relevant fields: launch date, launch site, payload mass, rocket type, outcome.

Store Data as Data Frame

- Save extracted data into a pandas DataFrame.
- Store the DataFrame locally for further processing.



Data Collection - Scraping

Initiate Web Scraping

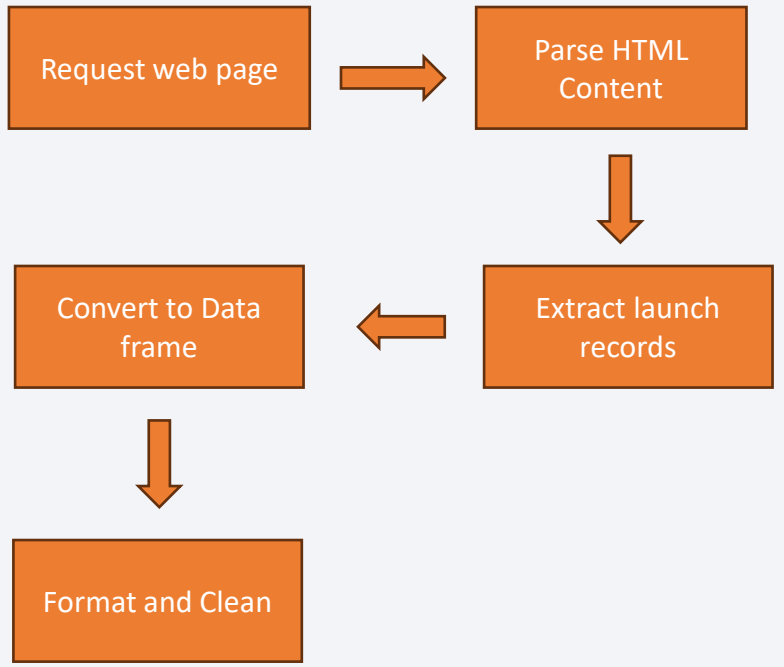
- Use Python's requests library to fetch the HTML content of the Wikipedia page.

Parse HTML Content

- Use BeautifulSoup to parse the HTML content.
- Extract the HTML table containing Falcon 9 launch records.

Convert to Data Frame

- Convert the extracted HTML table into a pandas Data Frame.
- Clean and format the Data Frame, ensuring data consistency.



Data Wrangling

Data wrangling involves cleaning, transforming, and organizing raw data into a structured format suitable for analysis.

Data Cleaning

- Identify and fill or remove missing values in the dataset.
- Use appropriate techniques or drop rows/columns with excessive missing data.

Data Transformation

- Convert data types to appropriate formats.
- Standardize text.
- Create new features from existing data.
- Normalize/scale numerical features to ensure consistency

Data Wrangling

Data Integration

- Merge datasets collected from different sources (API, web scraping) into a single dataset.
- Ensure consistent column names and data formats across datasets.

EDA with Data Visualization

Charts Plotted:

- **Histograms:**
  - Used to visualize the distribution of numerical variables such as launch success rates, payload mass, and flight number.
  - Helps in understanding the spread and central tendency of the data, identifying outliers, and assessing data skewness.
- **Bar Charts:**
  - Used to compare categorical variables such as launch outcomes (success/failure) across different categories like launch sites or rocket types.
  - Provides a clear comparison of frequencies or proportions within categorical data, highlighting patterns or trends.
- **Line Charts:**
  - Used to track trends over time, such as the success rate of Falcon 9 launches across different years.
  - Reveals patterns and helps in understanding performance trends or changes over specific periods.

EDA with Data Visualization

- Scatter Plots:
  - Used to explore relationships between two numerical variables, such as payload mass vs. launch success.
  - Identifies correlations or dependencies between variables, visualizing how one variable changes concerning another.

EDA with SQL

Aggregate Queries:

- Calculated the overall total number of launches.
- Counted both successful and failed launches.
- Determined success rates based on launch site and rocket type.

Filtering Queries:

- Narrowed down the data to focus on specific launch outcomes (success or failure).
- Applied filters to extract launches based on criteria such as launch date or rocket configuration.

EDA with SQL

Sorting Queries:

- Organized the data to highlight trends or identify outliers.
- Ordered launches by date or success rate for further analysis.

Subqueries:

- Used nested queries to calculate derived metrics, such as average payload mass per launch site.
- Employed subqueries to conduct in-depth analysis within larger datasets.

Build an Interactive Map with Folium

Map Objects Created:

Markers:

- Positioned markers to represent launch sites on the map.
- Each marker corresponds to a specific geographic location where SpaceX launches took place.

Circles:

- Added circles around launch sites to show proximity zones.
- Circles help illustrate the areas surrounding launch sites that may impact decision-making.

Lines:

- Drew lines to link launch sites with nearby zones or other relevant locations.
- Lines offer spatial context and highlight connections between key points related to the launches.

Build a Dashboard with Plotly Dash

Plots/Graphs Added:

Success Pie Chart:

- Shows the breakdown of successful versus failed launches.
- Helps visualize the overall success rate and identify performance trends.

Success-Payload Scatter Plot:

- Displays the correlation between payload mass and launch success.
- Provides insights into how payload mass impacts mission outcomes.

Interactions Added:

Launch Site Dropdown:

- Allows users to select specific launch sites for detailed analysis.
- Enhances filtering and focused exploration based on geographical locations.

Range Slider for Payload:

- Lets users dynamically adjust the payload mass range for more tailored exploration.

Predictive Analysis (Classification)

Data Preprocessing:

- Standardized features to ensure all variables contribute equally to the model.
- Split the data into training and test sets for validation purposes.

Model Selection:

- Applied various classification algorithms, including SVM, Decision Trees, and K-Nearest Neighbors (KNN).
- Chose algorithms suited for binary classification tasks based on the project's needs.

Hyperparameter Tuning:

- Utilized GridSearchCV to systematically find the optimal hyperparameters.
- Fine-tuned parameters like C (SVM), max_depth (Decision Trees), and n_neighbors (KNN).

Predictive Analysis (Classification)

Model Evaluation:

- Assessed models using cross-validation to ensure robustness and generalization.
- Used metrics such as accuracy, precision, recall, and F1-score to evaluate performance.

Improvement Iterations:

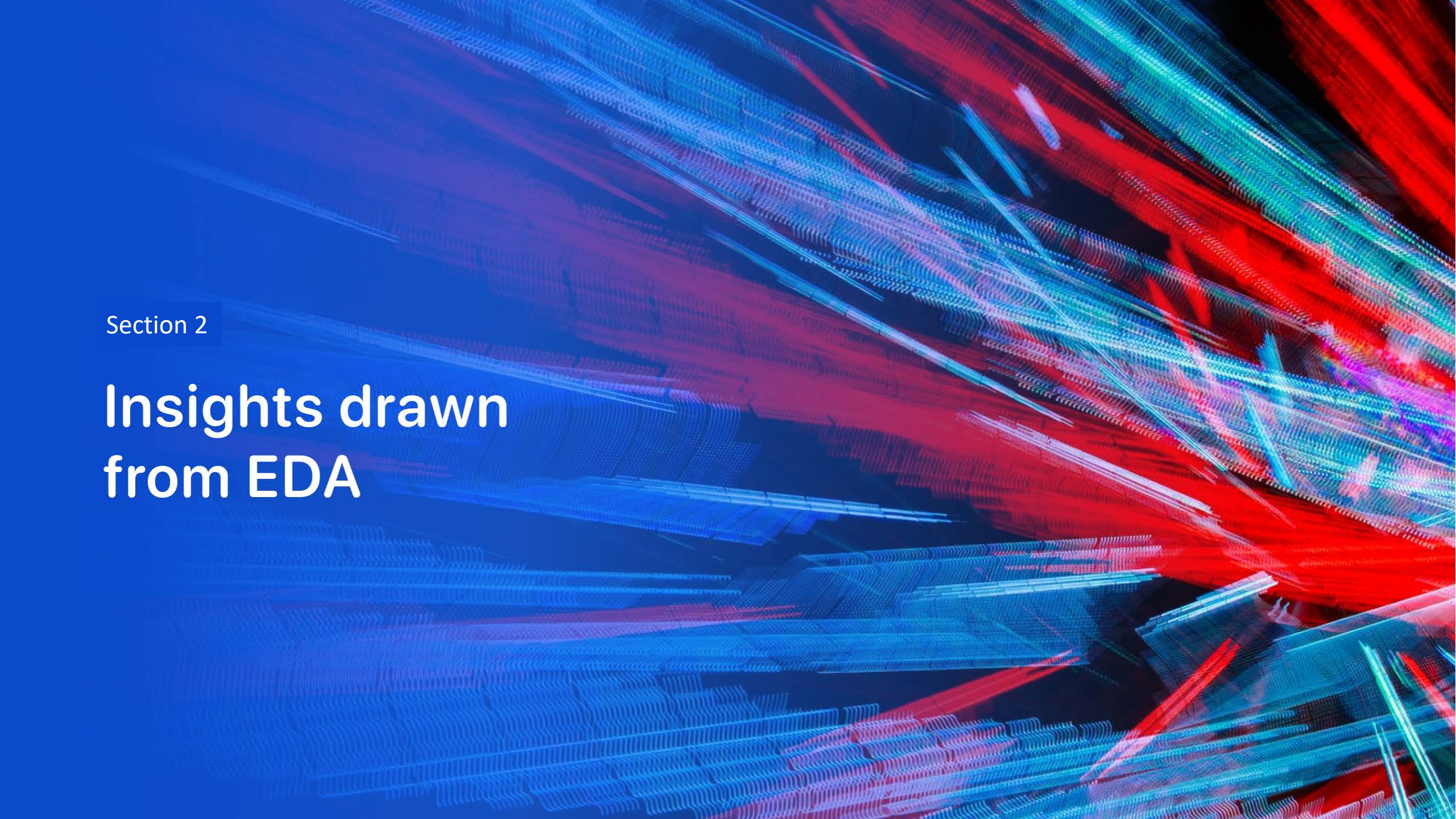
- Made iterative adjustments to models based on insights from validation results.
- Fine-tuned hyperparameters to enhance predictive accuracy and reliability.

Selection of Best Performing Model:

- Selected the model with the highest accuracy on the test set as the top performer.

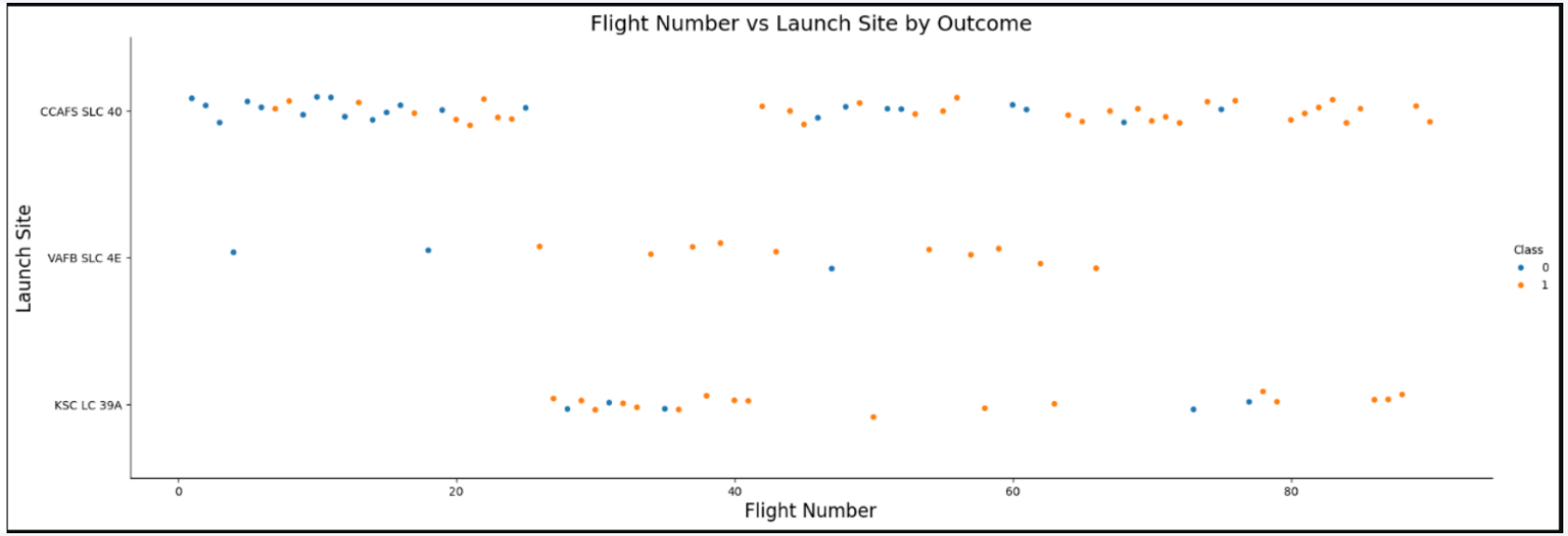
Section 2

Insights drawn from EDA



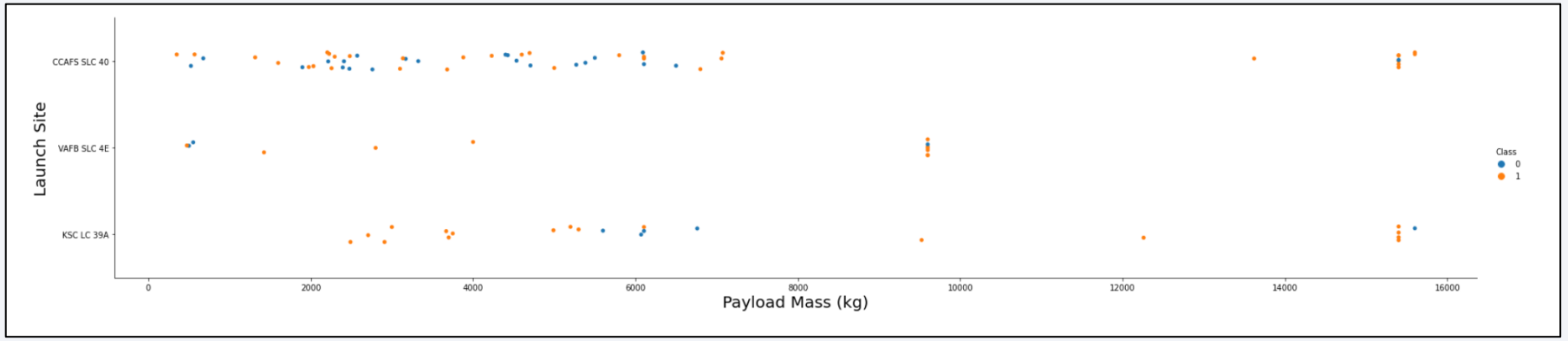
Flight Number vs. Launch Site

- The distribution indicating that factors other than the launch site itself may influence the landing success
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.



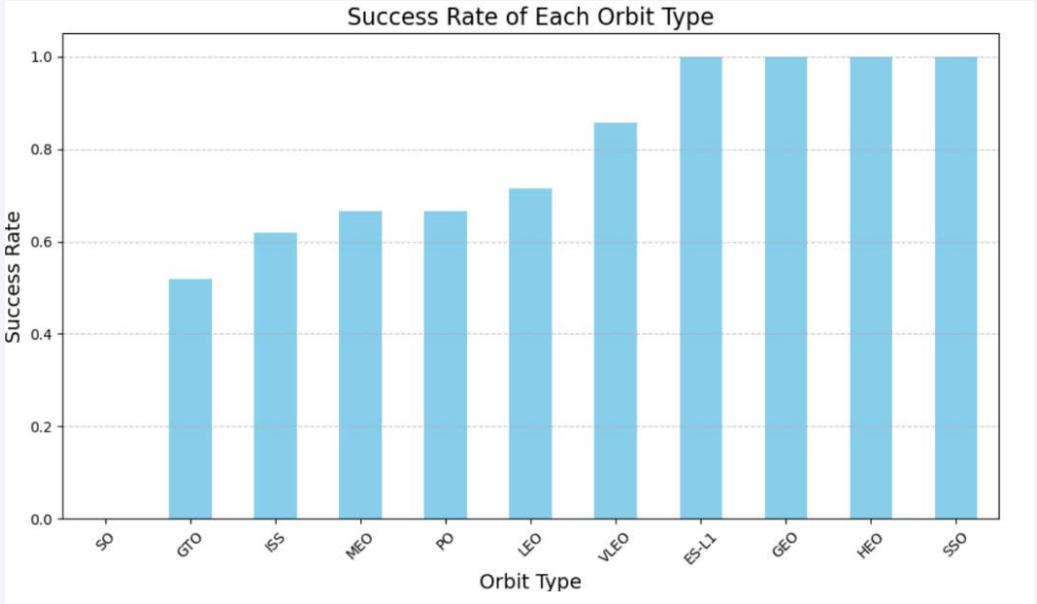
Payload vs. Launch Site

- Most launches from the CCAFS SLC 40 site handle payloads below 10,000 kg, while the VAFB SLC 4E and KSC LC 39A sites have a wider range of payload masses
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



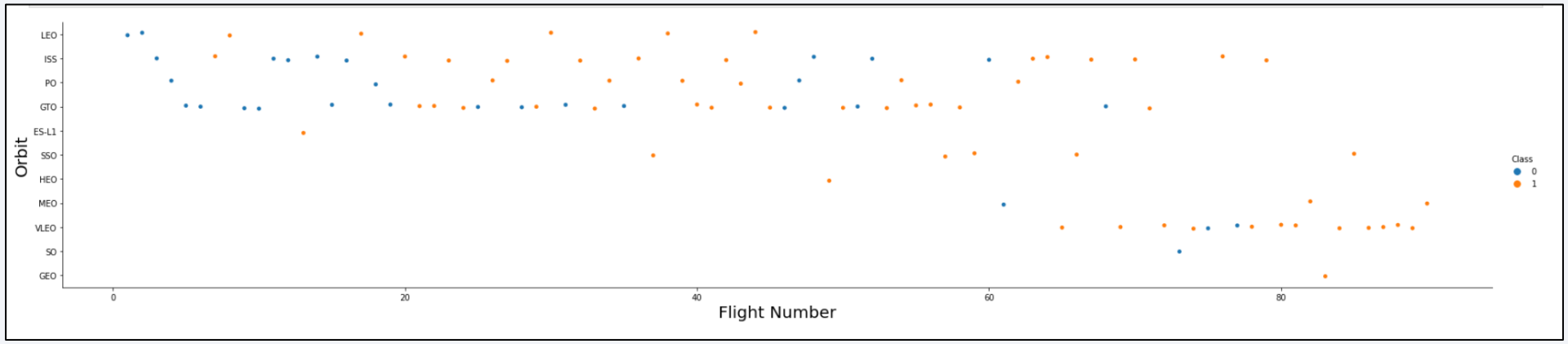
Success Rate vs. Orbit Type

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



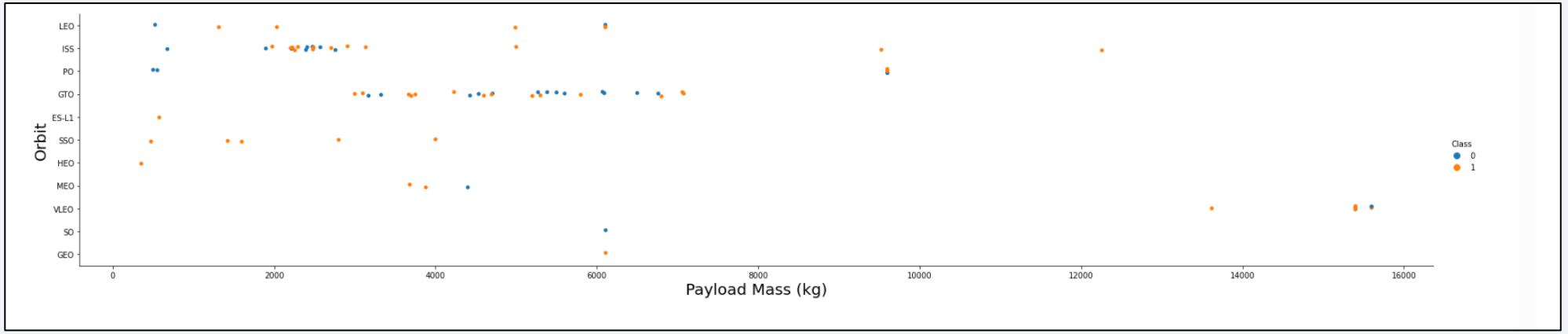
Flight Number vs. Orbit Type

- The success rate of launches improves with higher flight numbers
- Mix Outcomes with Orbit Specific Performance.



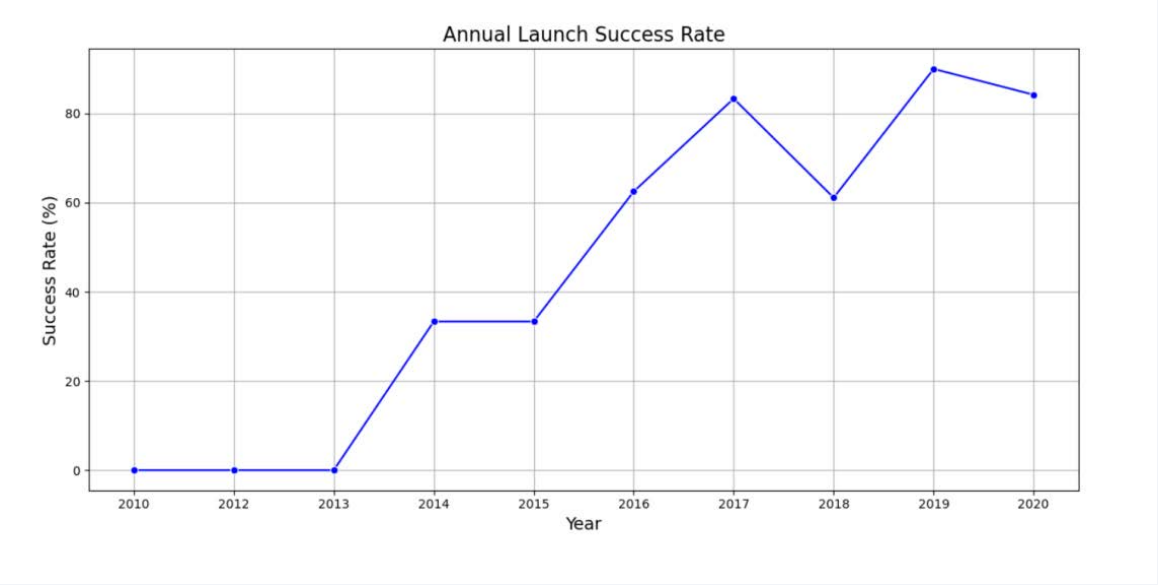
Payload vs. Orbit Type

- For payloads less than 6000 kg Successful landings are more frequent across all orbit types.
- For GTO orbit with payload greater than 5000, show more failures.
- Less Samples for payloads greater than 6000



Launch Success Yearly Trend

- Success rate kept increasing since 2013.
- This indicates increasing reliability and success in launches over the years despite slight fall in 2018



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

[12]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

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

[12]:

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

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

[13]:
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_O
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[14]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

[14]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[15]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

[15]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[17]:
```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

[17]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[18]:
```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ b
```

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

[18]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[19]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

[19]:

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

Boosters Carried Maximum Payload

Task 8

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
[21]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTA
```

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

[21]:

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

2015 Launch Records

```
        ELSE 'Unknown'
    END AS "Month_Name",
    "Mission_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTABLE
WHERE
    substr("Date", 0, 5) = '2015';
```

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

[16]:

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

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

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

[27]:
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

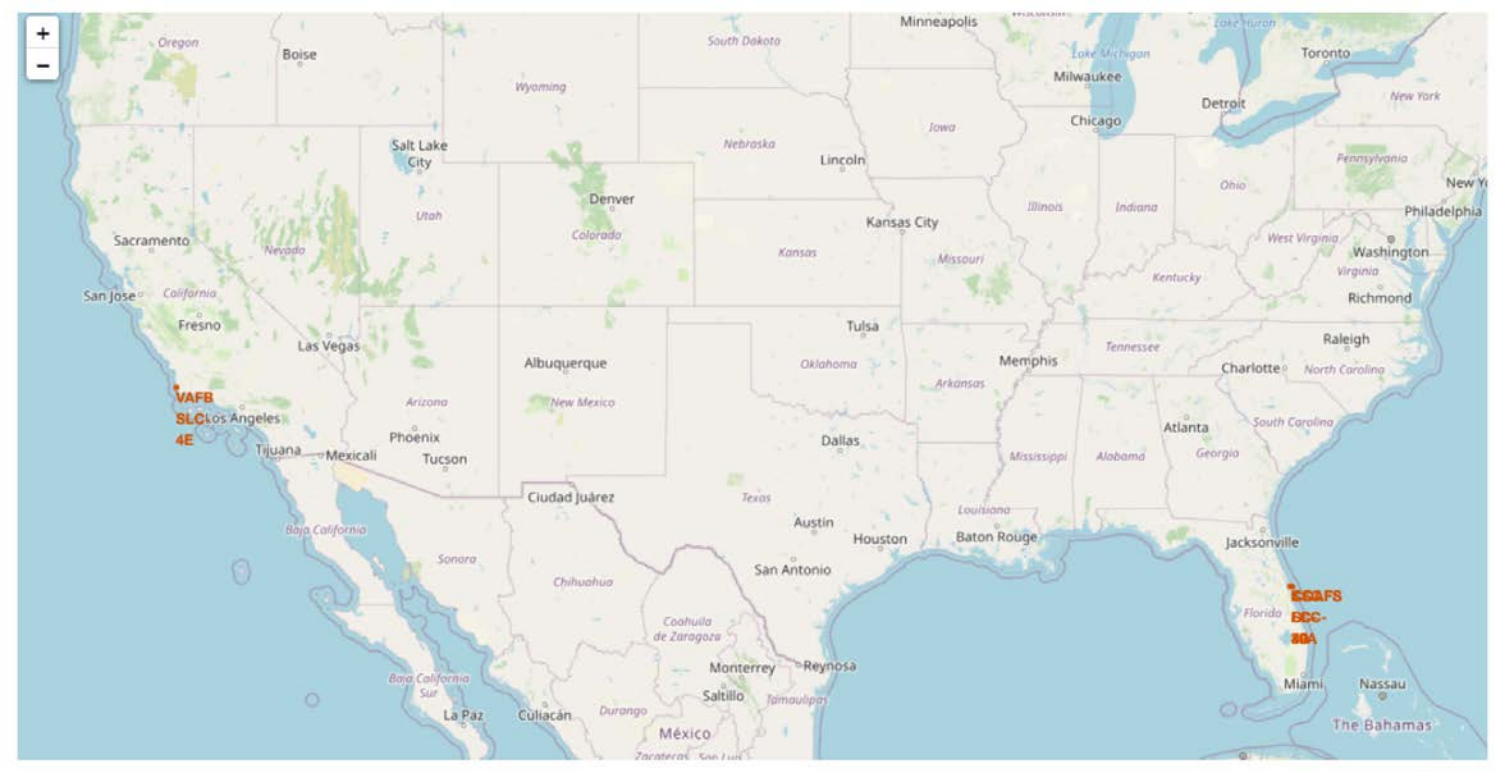
[27]:

Landing_Outcome	count_outcomes
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

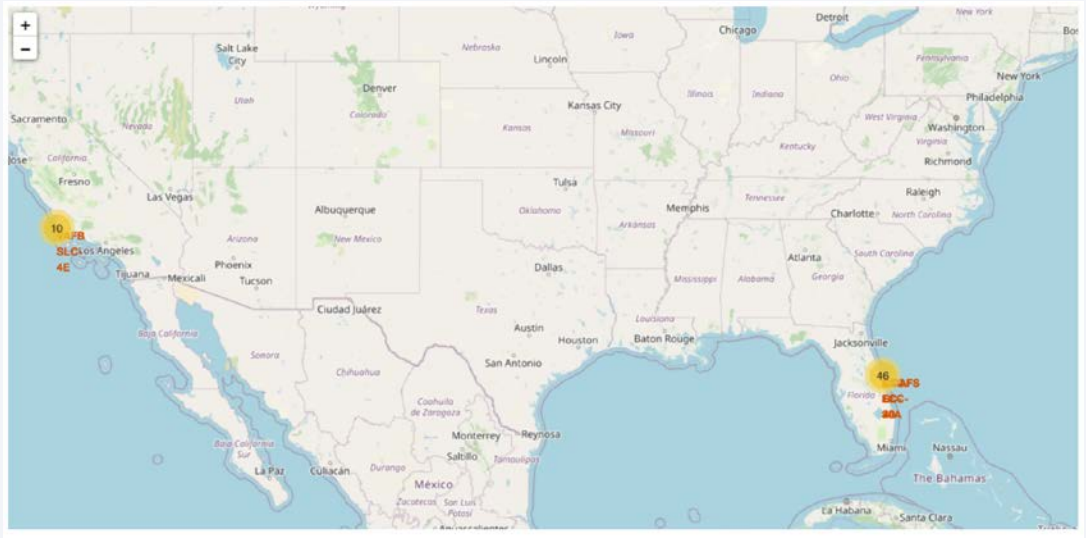
Mark all launch sites on a map

- Most of Launch sites considered are in proximity to the Equator line
- All launch sites considered are in very close proximity to the coast



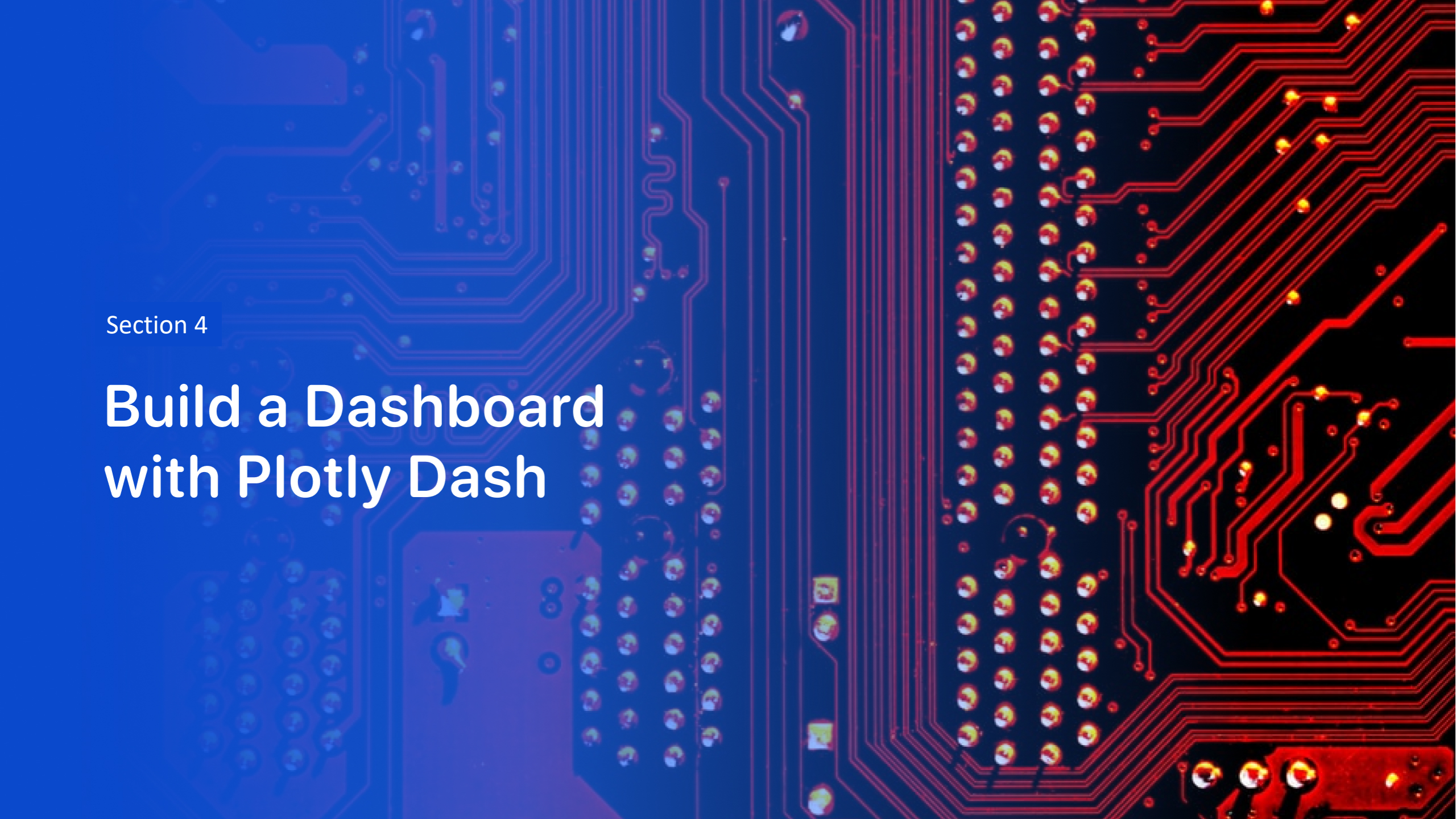
Mark the success/failed launches for each site on the map

- Success shows with Green Markers and Failure shows with red markers.
- Clustering helps manage a large number of markers more efficiently and reveals patterns that might be overlooked in a less structured plot. By analyzing marker colors and the information in popups, we can gain a better understanding of the characteristics and distribution of SpaceX launches.



Calculate the distances between a launch site to its proximities

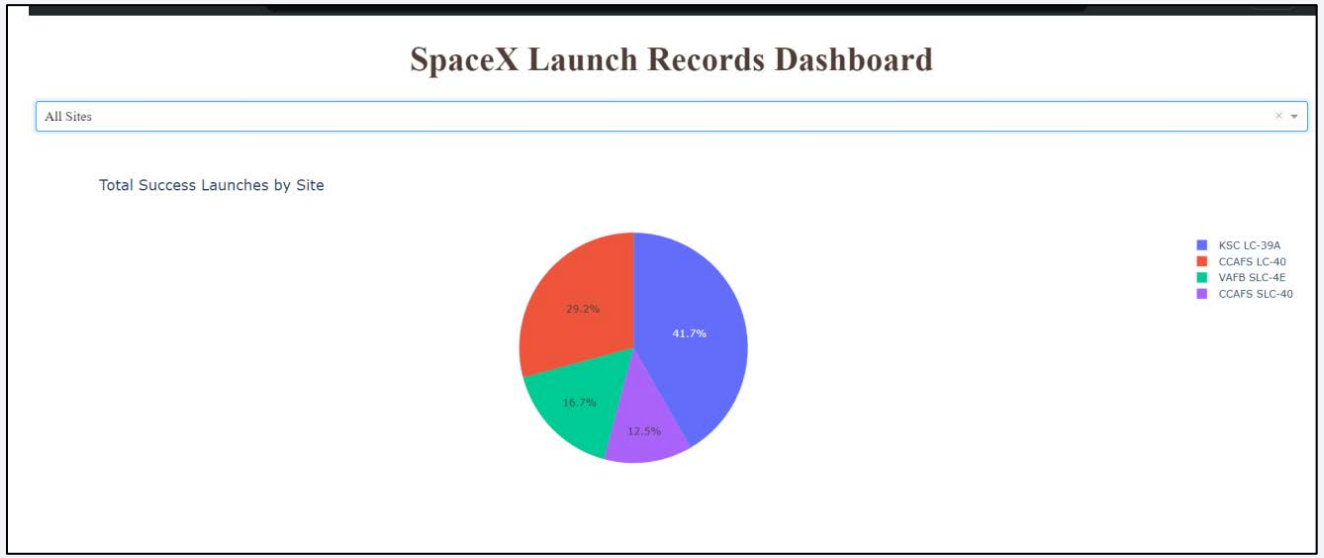
- From the visual analysis of the launch site, we can clearly see that it is:
  - Relatively close to highway
  - Relatively close to coastline


Section 4
Build a Dashboard with Plotly Dash

Launch Success Records Dashboard

Success Rates

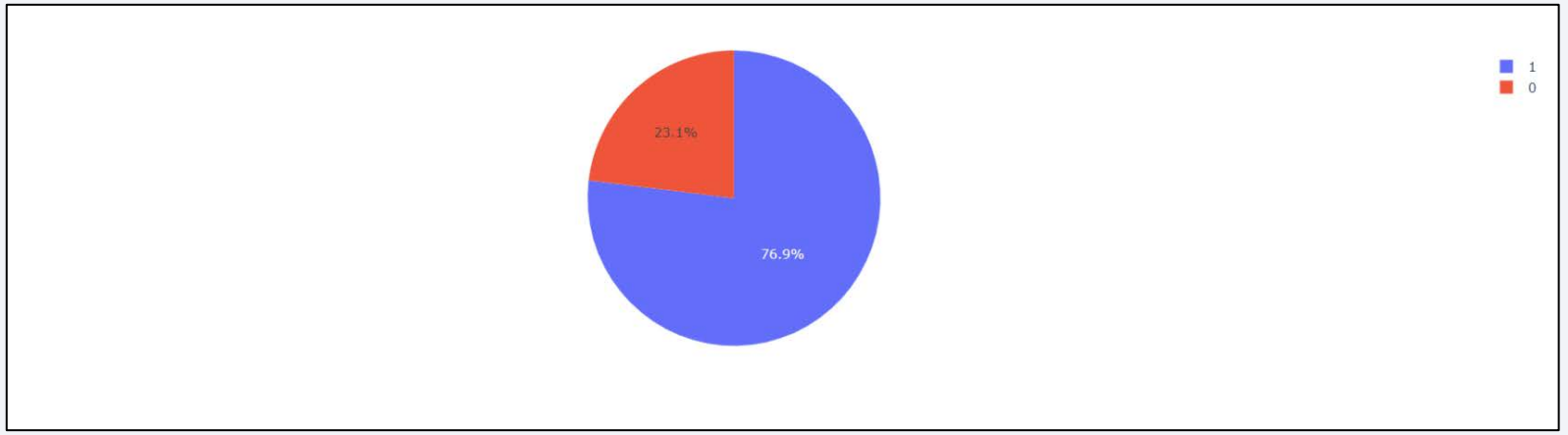
- CCAFS LC-40: 29.2%
- CCAFS SLC-40: 12.5%
- VAFB SLC-4E: 16.7%
- KSC LC-39A: 41.7%



Highest launch Success ratio Dashboard

KSC LC-39A:

- Class 1 (Successful Launches): 76.9%
- Class 0 (Unsuccessful Launches): 23.1%



Payload vs. Launch Outcome scatter plot

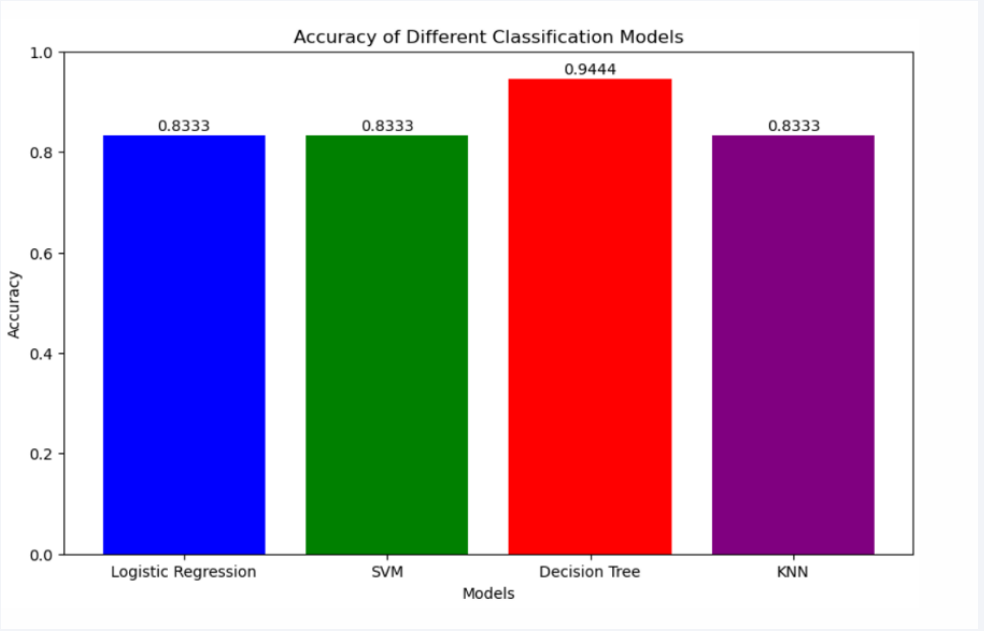
- Booster version “FT” has a high success rate across various payload masses.
- Booster version “v1.0” has fewer launches and may require further analysis to understand its performance.

Section 5

Predictive Analysis (Classification)

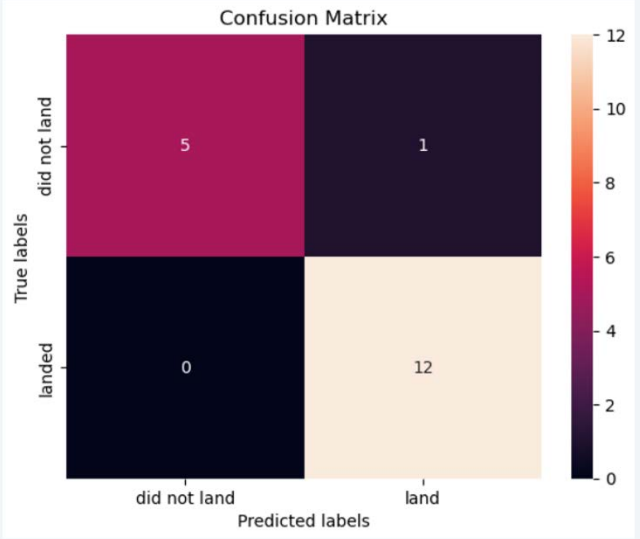
Classification Accuracy

- The Decision Tree model has the highest classification accuracy.
- Decision Tree has Accuracy: 94%



Confusion Matrix

- The Decision Tree Predictor achieved a high accuracy with a score of 94.44%, with a significant number of true positives and true negatives.
- With no false negatives indicates that the model reliably predicts successful landings



Conclusions

- The "KSC LC-39A" launch site boasts the highest success rate, with 41% of launches being successful.
- The Decision Tree model achieved the highest classification accuracy.
- The majority of launch sites are located near the Equator, with all sites being very close to the coast.
- Launch success rates have improved over the years.
- Orbits such as ES-L1, GEO, HEO, and SSO all have a 100% success rate.

Appendix

- For Reference Jupyter Notebooks: https://github.com/NafeesDev1/AppliedDataScience

Thank you!