Winning Space Race with Data Science

Subham Gaurav

24th August 2024

Outline

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

Executive Summary

Summary of methodologies

Data Collection using API

Data Collection using Web Scraping

Explanatory Analysis Using SQL

Explanatory Analysis Using Visualization

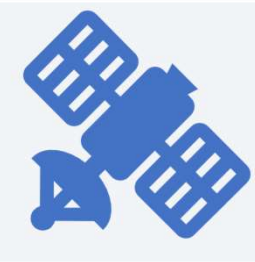
Summary of all results

Interactive Plotly Dashboard for Data Presentation

Predictive Analysis Result using Multiple Predictive Model

Introduction

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.



Aim to answer the following Questions

How do Multiple Variables such as Payload Mass , Orbit , Launch Site have Affected the respective Outcome of the Mission.

Whether the trend of Successful Operation have increased gradually over the year

Which Predictive Model should be employed for the most accurate predication.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collection process involved using two primary sources to gather comprehensive information about SpaceX launches:
    - SpaceX REST API: To fetch detailed technical and launch-specific information.
    - Wikipedia Web Scraping: To retrieve additional details not available through the API, such as customer information and payload details
- Perform data wrangling
  - The landing outcomes are being simplified into binary labels:
    - "1": Successful landing (True Ocean, True RTLS, True ASDS).
    - "0": Unsuccessful landing (False Ocean, False RTLS, False ASDS).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Training of machine learning models to predict future landing outcomes. By using a binary classification approach, we can leverage logistic regression, decision trees, or other classification algorithms to analyze patterns and factors that contribute to successful or failed landings.

Data Collection

Data was collected using both SpaceX's REST API and web scraping from Wikipedia. This approach ensured the completeness of the dataset, combining different data sources for more comprehensive analysis.

1. **SpaceX REST API**:

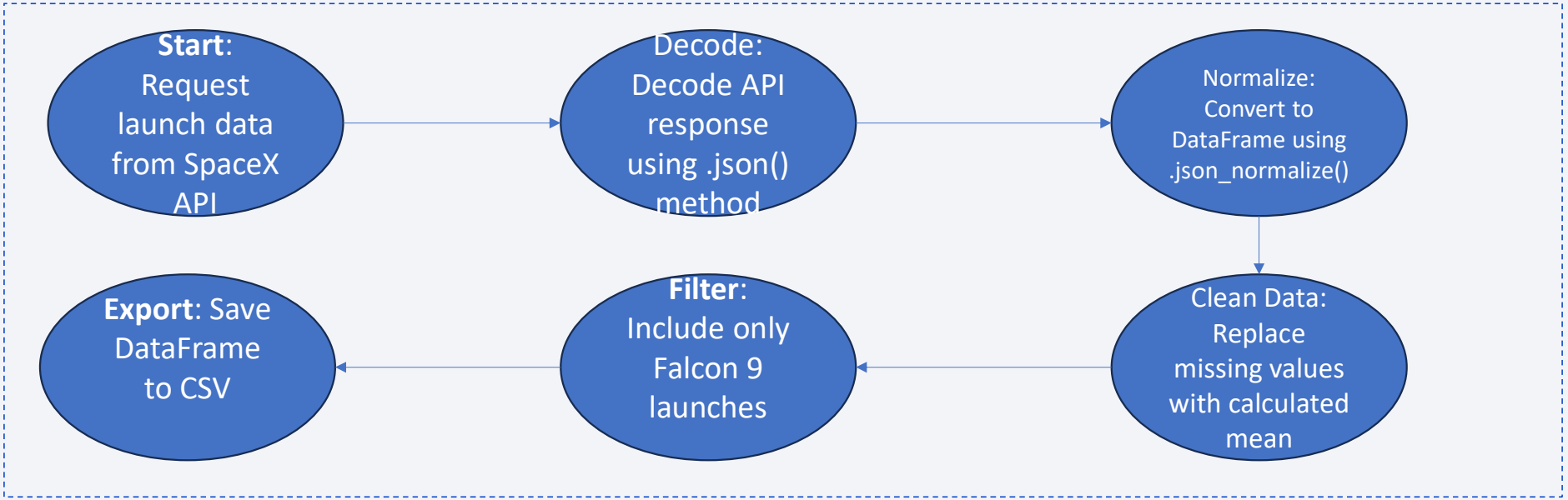
The SpaceX REST API provided detailed information on rocket launches, which was obtained through API requests and processed into a DataFrame using .json_normalize(). This data was then cleaned by replacing missing values in crucial fields, such as Payload Mass, with the mean value of the respective column. The analysis was focused solely on Falcon 9 launches, so other rocket types were filtered out. The final cleaned data was exported to a CSV file for further use.

2. **Web Scraping from Wikipedia**:

Web scraping was conducted on the SpaceX Wikipedia page to gather additional Falcon 9 launch data. This involved sending an HTML request to access the page and using BeautifulSoup to parse and extract data from the HTML tables. The column names were directly taken from the table headers, and the extracted data was structured into a dictionary before being converted into a DataFrame. This DataFrame was also exported to a CSV file, maintaining consistency in data storage across sources. The combination of these data collection methods provided a more robust dataset, enhancing the accuracy and reliability of the subsequent analysis.

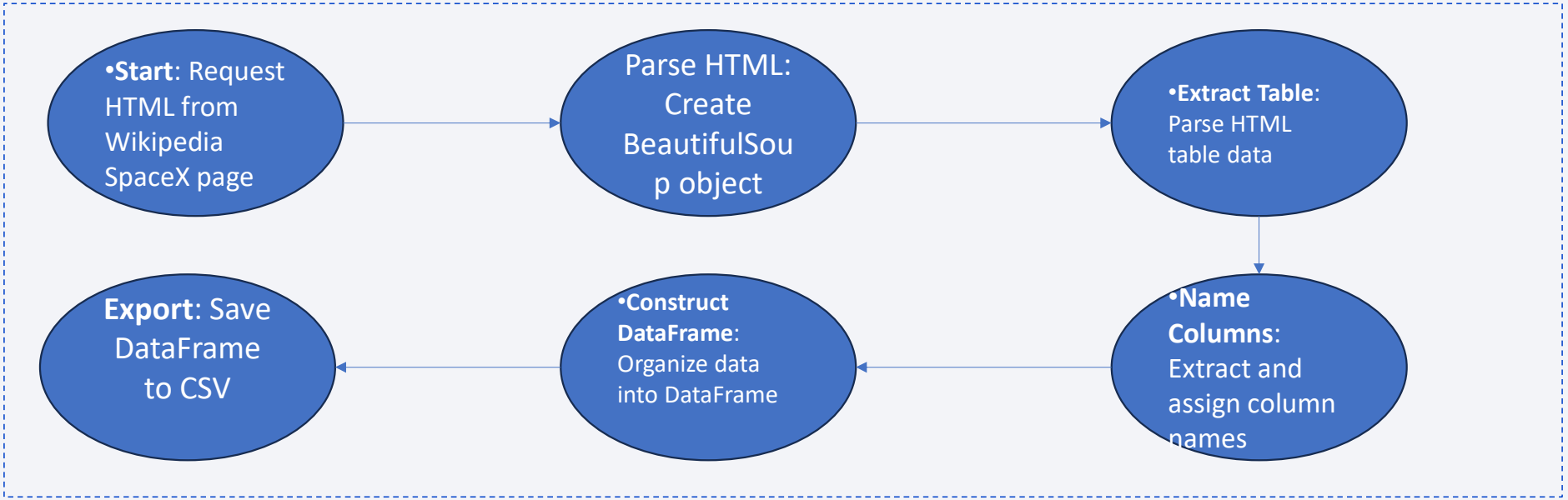
Data Collection – SpaceX API

- SPACE X API



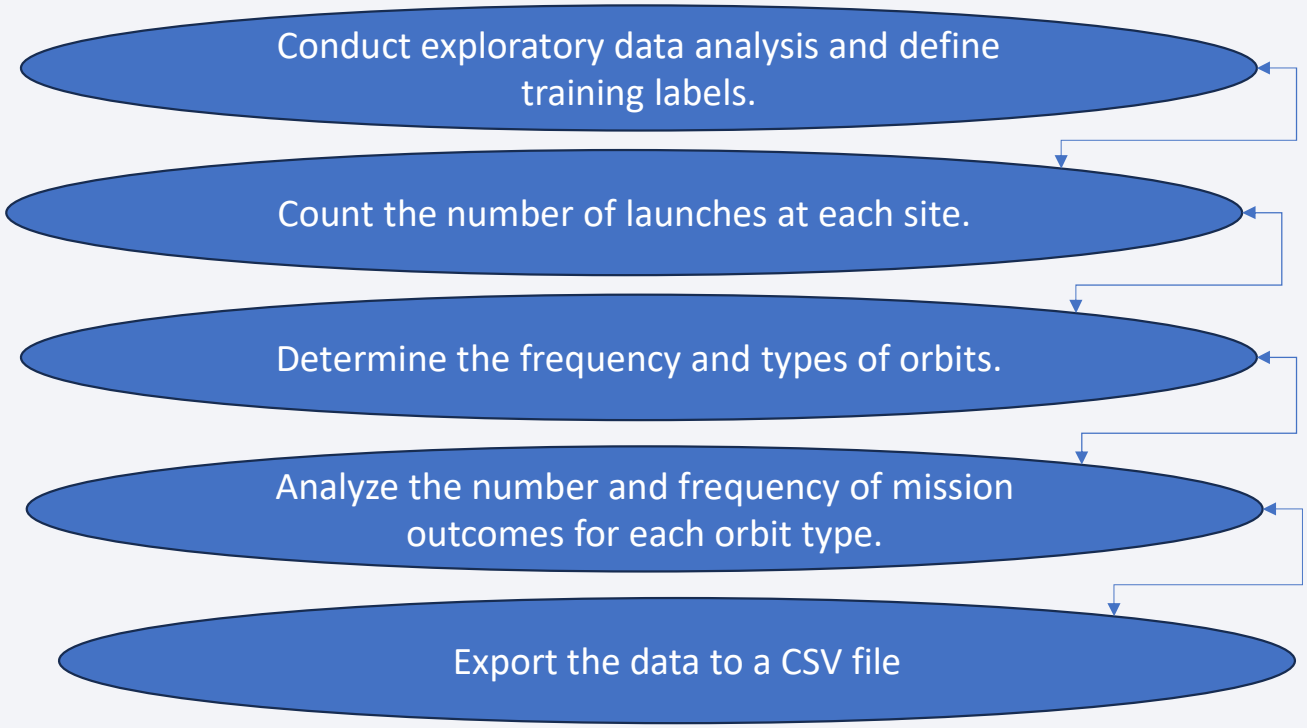
Data Collection - Scraping

- Web Scarping



Data Wrangling

- In the dataset, there are various scenarios where the booster failed to land successfully. For instance, "True Ocean" indicates a successful landing in a specific ocean region, while "False Ocean" signifies an unsuccessful landing in that region. Similarly, "True RTLS" denotes a successful landing on a ground pad, whereas "False RTLS" indicates an unsuccessful landing on a ground pad. "True ASDS" means a successful landing on a drone ship, while "False ASDS" represents an unsuccessful landing on a drone ship. For training purposes, we convert these outcomes into labels: "1" represents a successful landing, and "0" represents an unsuccessful landing.
- Data Wrangling



EDA with Data Visualization

1. Flight Number vs. Payload Mass **Purpose:** To observe if there is a trend or pattern between the flight number and payload mass. Scatter plots reveal relationships that might be useful for predictive modeling.
2. Flight Number vs. Launch Site **Purpose:** To analyze which launch sites are used for different flight numbers and their success rates. This helps in understanding the distribution of launches across sites.
3. Payload Mass vs. Launch Site **Purpose:** To examine the relationship between payload mass and the launch site. Bar charts provide insights into how payload mass affects success rates at different launch sites.
4. Orbit Type vs. Success Rate **Purpose:** To visualize the success rate for different orbit types. This helps identify which orbits have higher success rates and informs the analysis of orbit-related factors.
5. Flight Number vs. Orbit Type **Purpose:** To see if the number of flights impacts success rates for various orbits. Scatter plots can highlight any significant trends or outliers.
6. Payload Mass vs. Orbit Type **Purpose:** To explore how payload mass influences success rates across different orbit types. This chart helps understand the impact of payload weight on mission success.
7. Success Rate Yearly Trend **Purpose:** To track the success rate of launches over the years. Line charts are used to identify any trends or improvements in success rates over time.

- Data Visualization

EDA with SQL

1. **Display unique launch site names**: Retrieved all unique launch site names from the dataset.
2. **Records where launch sites begin with 'CCA'**: Showed 5 records for launch sites starting with 'CCA'.
3. **Total payload mass for NASA (CRS) launches**: Calculated the total payload mass for boosters launched by NASA under CRS.
4. **Average payload mass for booster version F9 v1.1**: Computed the average payload mass carried by the F9 v1.1 booster version.
5. **Date of first successful ground landing**: Listed the date when the first successful landing on a ground pad occurred.
6. **Successful drone ship landings with payload between 4000 and 6000**: Identified boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000 kg.
7. **Total successful and failed mission outcomes**: Counted the total number of successful and failed mission outcomes.
8. **Boosters with maximum payload mass**: Listed booster versions that carried the maximum payload mass.
9. **Failed drone ship landings in 2015**: Listed failed drone ship landings, including booster versions and launch site names for 2015.
10. **Ranking of landing outcomes between 2010-06-04 and 2017-03-20**: Ranked the count of landing outcomes (e.g., Failure on drone ship, Success on ground pad) in descending order between specified dates.

- SQL QUERIES

Build an Interactive Map with Folium

1. Markers with Circles and Labels

- **Description:** Added markers with circles and text labels for each launch site, including NASA Johnson Space Center and other key sites.
- **Purpose:** To visually indicate the geographical locations of launch sites on the map. The circles and labels help in identifying each site's position and provide contextual information.

2. Coloured Markers for Launch Outcomes

- **Description:** Used coloured markers to represent successful (green) and failed (red) launches.
- **Purpose:** To visually differentiate between successful and unsuccessful launches. This color-coding allows for quick assessment of launch outcomes and highlights patterns or concentrations of success or failure.

3. Coloured Lines to Show Distances

- **Description:** Added lines connecting the launch site (e.g., KSC LC-39A) to nearby features such as railways, highways, coastlines, and cities.
- **Purpose:** To illustrate the proximity of launch sites to critical infrastructure and geographical features. This visualization helps in understanding potential risks and the accessibility of launch sites.

- Folium Application

Build a Dashboard with Plotly Dash

1. Launch Sites Dropdown List

- **Description:** A dropdown menu allowing users to select a specific launch site.
- **Purpose:** To filter and view data specific to the selected launch site, making it easier to analyze performance and outcomes for individual sites.

2. Pie Chart for Success Launches

- **Description:** A pie chart displaying the proportion of successful launches versus failed launches for all sites or for the selected site.
- **Purpose:** To provide a visual representation of the success rates, allowing for a quick comparison of launch success across different sites or overall.

3. Slider for Payload Mass Range

- **Description:** A slider enabling users to select a range of payload masses.
- **Purpose:** To filter and analyze the impact of payload mass on launch success rates within the chosen range, facilitating detailed exploration of how payload size affects outcomes.

4. Scatter Chart of Payload Mass vs. Success Rate

- **Description:** A scatter plot showing the relationship between payload mass and success rates for different booster versions.
- **Purpose:** To visualize how payload mass correlates with launch success and identify any patterns or outliers related to different booster versions.

- Plotly Dashboard

Predictive Analysis (Classification)

1. Building the Models

Models Created: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

Process: Implemented each classification model using the training dataset. The models were initially built with default parameters to establish baseline performance.

2. Evaluating the Models

Metrics Used: Jaccard Score, F1 Score, Confusion Matrix, and accuracy.

Process: Each model's performance was evaluated using these metrics to assess their effectiveness in predicting successful and unsuccessful launches. The confusion matrix provided insights into false positives and false negatives.

3. Improving the Models

Techniques Used:

Standardization: Applied StandardScaler to standardize the data, ensuring all features contribute equally to the model.

Hyperparameter Tuning: Used GridSearchCV with cross-validation to find the optimal parameters for each model, enhancing their performance.

Process: Conducted grid search to explore different hyperparameter values and selected the best-performing configurations for each model.

4. Finding the Best Performing Model

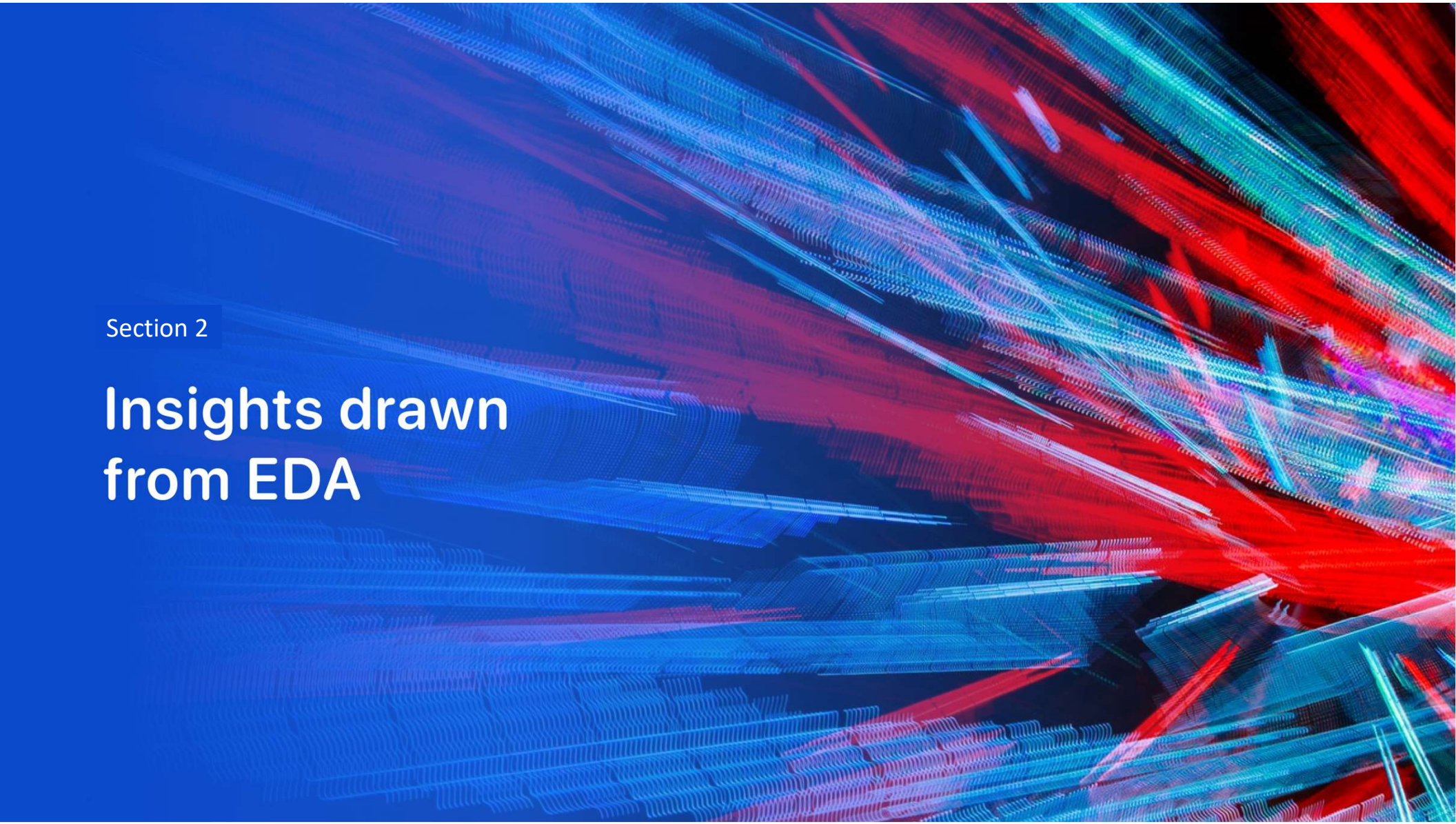
Comparison: Compared models based on accuracy scores, Jaccard Score, and F1 Score from both the test set and the entire dataset.

Results: Decision Tree emerged as the best model with higher accuracy and better performance metrics compared to others. It consistently showed strong results across various evaluations.

- Model Development

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

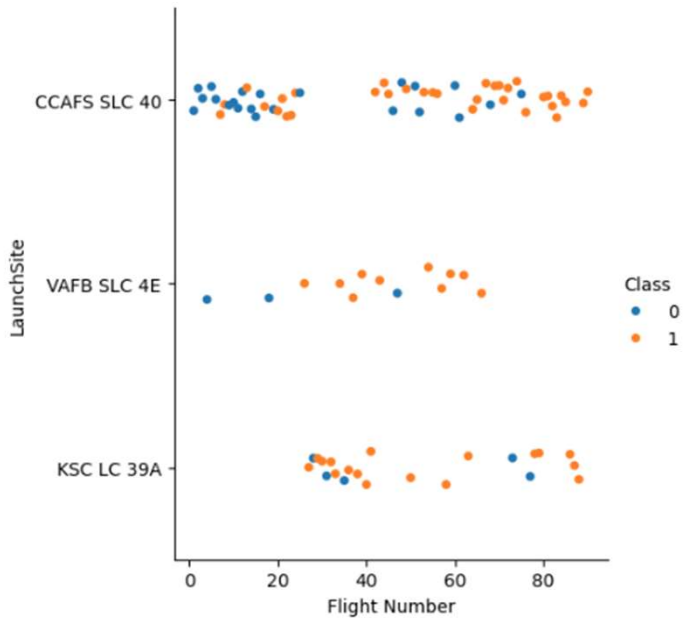


Section 2

Insights drawn from EDA

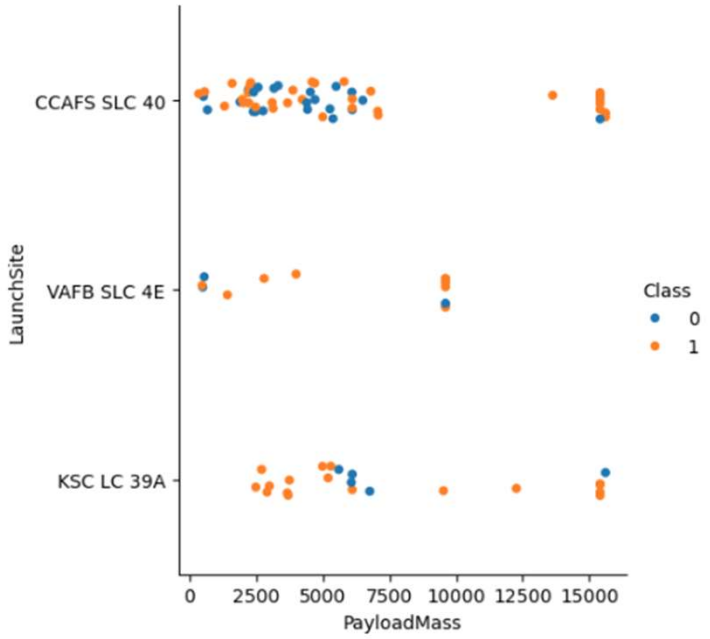
Flight Number vs. Launch Site

- Earlier flights had lower success rates compared to more recent ones. Launch sites like CCAFS SLC 40 had a high number of launches, while VAFB SLC 4E and KSC LC 39A exhibited higher success rates. Newer flights showed improved success rates.



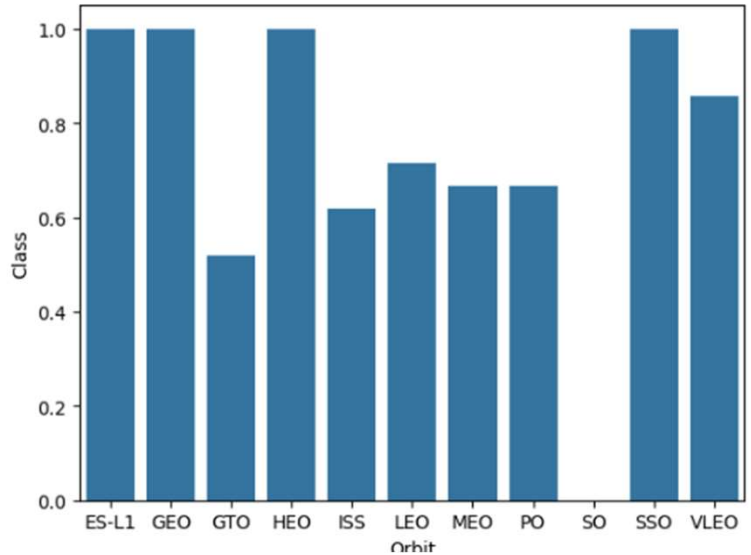
Payload vs. Launch Site

- Higher payload masses generally correlated with higher success rates. Most successful launches involved payloads over 7000 kg. KSC LC 39A showed a 100% success rate for payload masses under 5500 kg.



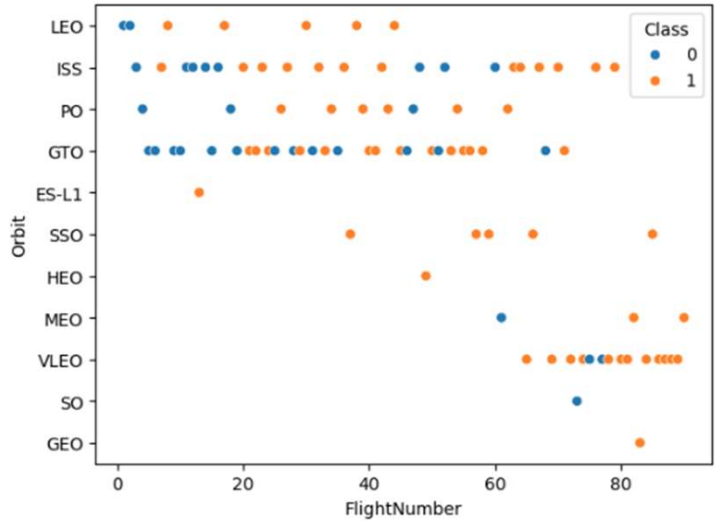
Success Rate vs. Orbit Type

- Orbits like ES-L1, GEO, HEO, and SSO had 100% success rates, while SO orbits had 0% success rate. Other orbits had mixed success rates between 50% and 85%.



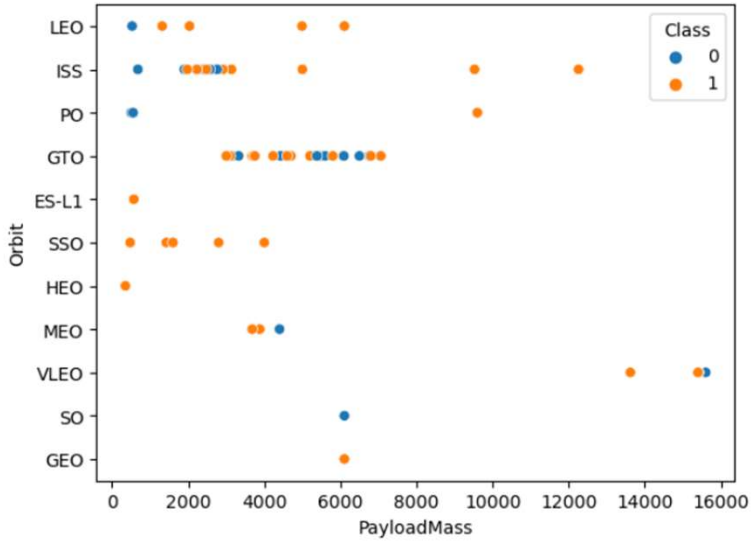
Flight Number vs. Orbit Type

- For LEO orbits, success rates appeared related to the number of flights. However, there was no clear relationship between the number of flights and success rates in GTO orbits.



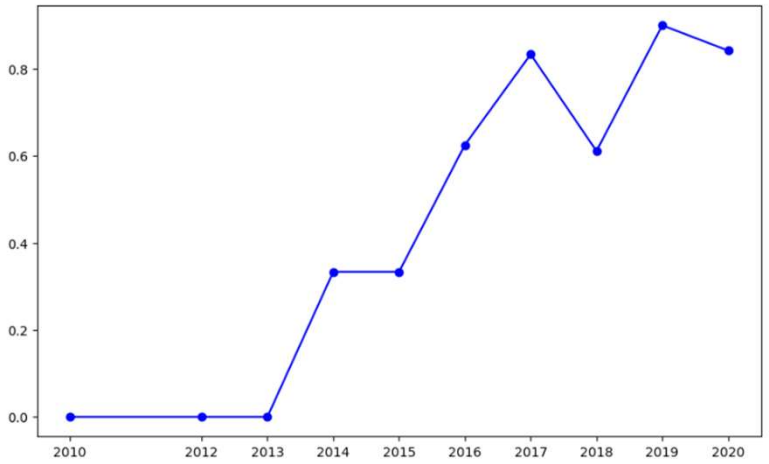
Payload vs. Orbit Type

- Heavy payloads negatively impacted success rates in GTO orbits but had a positive effect on ISS and Polar LEO orbits.



Launch Success Yearly Trend

- Success rates improved progressively from 2013 to 2020, indicating an upward trend in successful launches over the years.



All Launch Site Names

- %sql Select Distinct "Launch_Site" from SPACEXTABLE;

- DISTINCT Keyword: Ensures that each launch site name appears only once in the result, eliminating any duplicates.

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

Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5
- Limit Keyword : Ensures that only 5 Result will be shown.
- Like Keyword : Ensure that it matches the regex as mentioned

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Succes
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	Succes
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Succes
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Succes
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Succes

Total Payload Mass

- %sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE WHERE Customer like "%NASA%“
- This SQL query calculates the total payload mass (in kilograms) for rows in the "SPACEXTABLE" where the customer field contains the substring "NASA"

SUM("PAYLOAD_MASS__KG_")
107010

Average Payload Mass by F9 v1.1

- %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE WHERE Booster_Version like "%F9 v1.1%“
- This SQL query calculates the average payload mass (in kilograms) for rows in the "SPACEXTABLE" where the booster version contains the text "F9 v1.1"

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- %sql select MIN("Date") from SPACEXTABLE WHERE Landing_Outcome like "%Success (ground pad)%“
- This SQL query finds the earliest date from the "SPACEXTABLE" where the landing outcome includes the phrase "Success (ground pad).

MIN("Date")
2015-12-22

First Successful Ground Landing Date

- %sql select MIN("Date") from SPACEXTABLE WHERE Landing_Outcome like "%Success (ground pad)%“
- This SQL query finds the earliest date from the "SPACEXTABLE" where the landing outcome includes the phrase "Success (ground pad).

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT DISTINCT ("Booster_Version") from SPACEXTABLE WHERE Landing_Outcome like "%Success (drone ship)%" and PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
- This SQL query retrieves a list of unique booster versions from the "SPACEXTABLE" where the landing outcome includes "Success (drone ship)" and the payload mass is between 4,000 and 6,000 kilograms.

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

Total Number of Successful and Failure Mission Outcomes

- %sql Select "Mission_Outcome", COUNT("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome“
- This SQL query counts the number of occurrences of each unique mission outcome in the "SPACEXTABLE" and displays the mission outcome along with its count.

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

Boosters Carried Maximum Payload

- %sql Select DISTINCT("Booster_Version") from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
- This SQL query retrieves the unique booster version(s) from the "SPACEXTBL" table that corresponds to the maximum payload mass found in that table.

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

- %sql SELECT SUBSTR("Date", 6,2) as month,Booster_Version,Launch_Site from SPACEXTBL where "Landing_Outcome" like '%Failure (drone ship)%' and SUBSTR(Date,1,4)='2015’;
- This SQL query extracts the month from the "Date" column and selects the booster version and site for entries in the "SPACEXTBLlaunch" table where the landing outcome includes "Failure (drone ship)" and the year is 2015.

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- %sql select Landing_Outcome,COUNT(Landing_Outcome) from SPACEXTABLE where "Landing_Outcome" in ('Failure (drone ship)','Success (ground pad)') and Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;
- This SQL query counts the number of times each landing outcome ("Failure (drone ship)" or "Success (ground pad)") occurs in the "SPACEXTABLE" for entries with a date between June 4, 2010, and March 20, 2017. It groups the results by landing outcome and orders them in descending order of count.

Landing_Outcome	COUNT(Landing_Outcome)
Failure (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

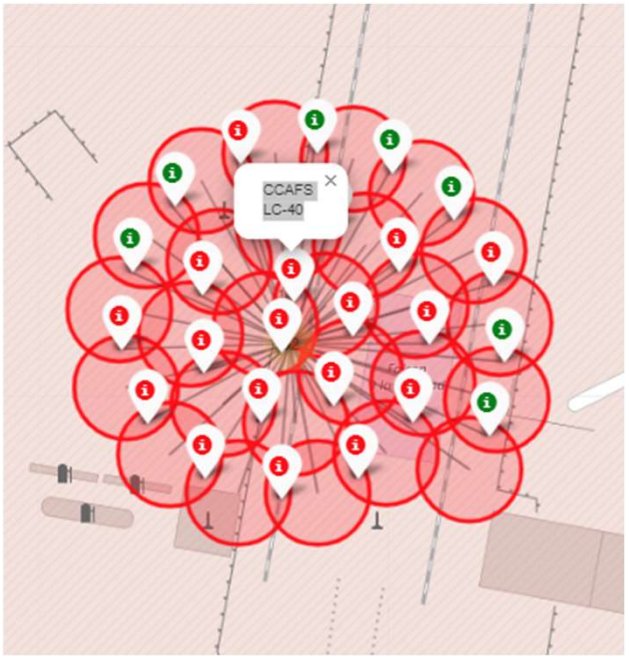
Global Map of Launch Sites with Location Markers

- **Distribution of Launch Sites:**
  - "The map shows the global distribution of SpaceX launch sites. We observe that most sites are located in regions close to the equator, such as Florida and California. This positioning is advantageous for launching satellites into orbit due to the higher rotational speed of the Earth at the equator."
- **Proximity to Coastlines:**
  - "All launch sites are strategically placed near coastlines. This minimizes the risk to populated areas in case of launch failures, as any debris or explosions would occur over the ocean."
- **Concentration of Launch Sites:**
  - "There is a notable concentration of launch sites in the United States, reflecting SpaceX's operational base and primary launch activities."

Colour-labeled launch records for CCAFS LC-40

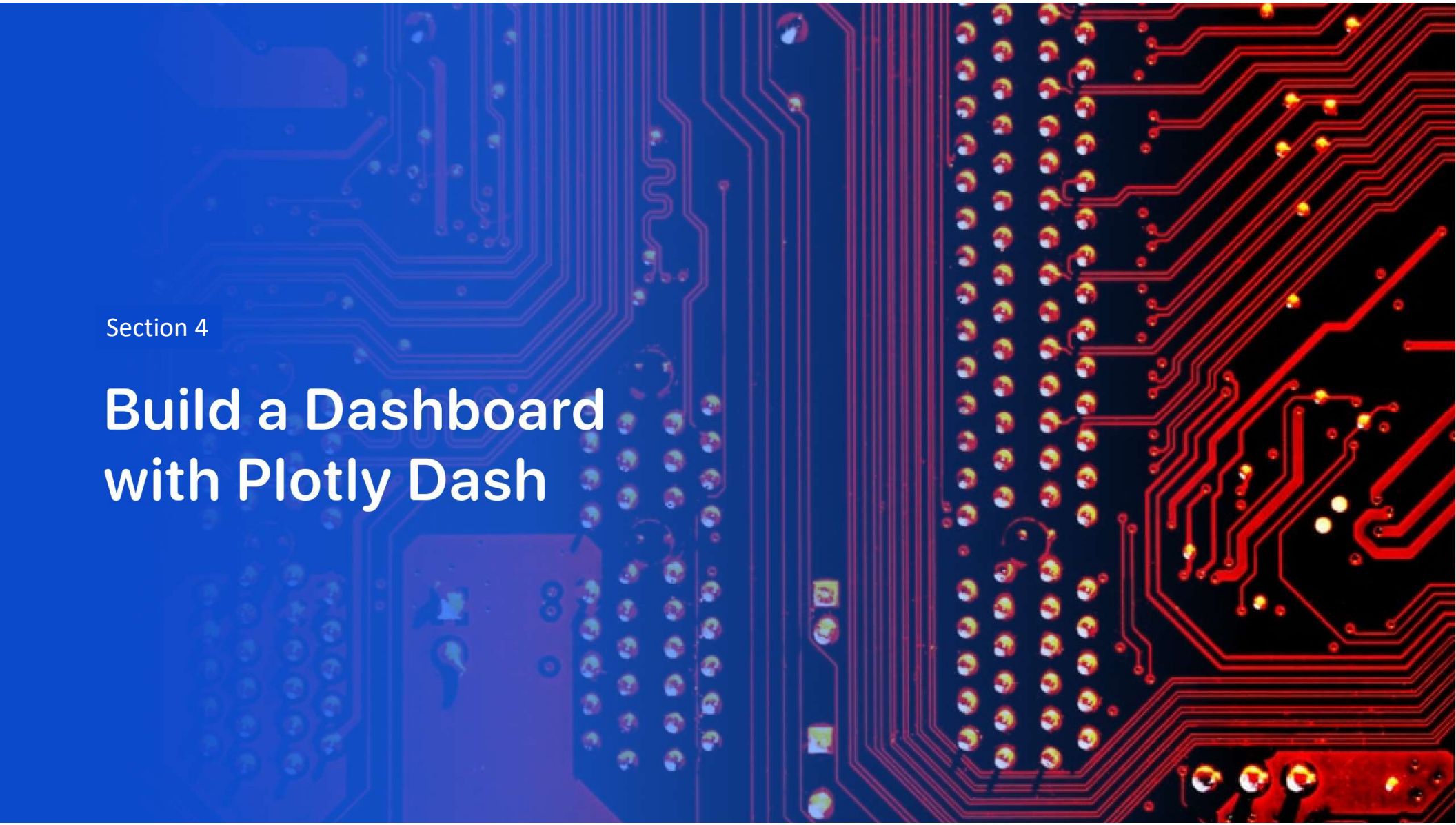
- By looking at the color-coded markers, we can quickly see which launch sites have higher success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch

- The CCAFS LC -40 launch site has a particularly high success rate.



Distance from the launch site CCAFS SLC-40

- Visual analysis of the CCAFS SLC-40 launch site reveals that it is:
  - Relatively close to a railway (0.98 km)
  - Relatively close to a highway (0.61 km)
  - Relatively close to the coastline (0.87 km)
  - Additionally, the launch site is near the city of Titusville (23.21 km).
  - A failed rocket, moving at high speed, could cover distances of 15-20 km in a matter of seconds, posing a potential danger to nearby populated areas.

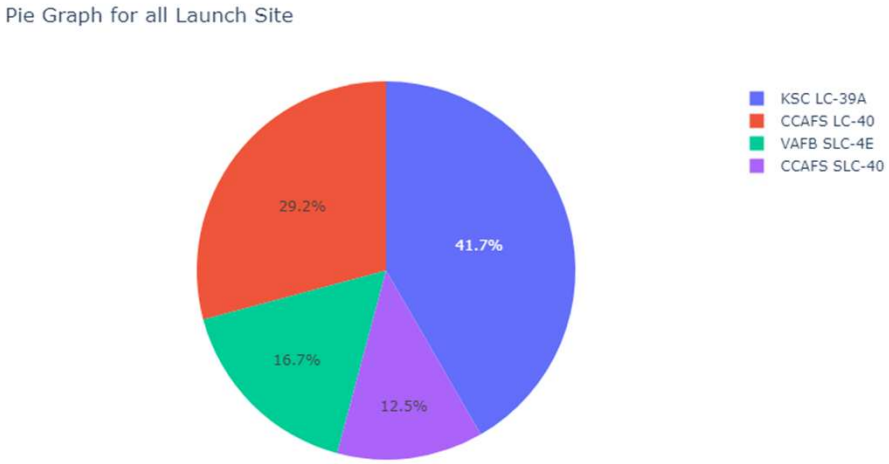


Section 4

Build a Dashboard with Plotly Dash

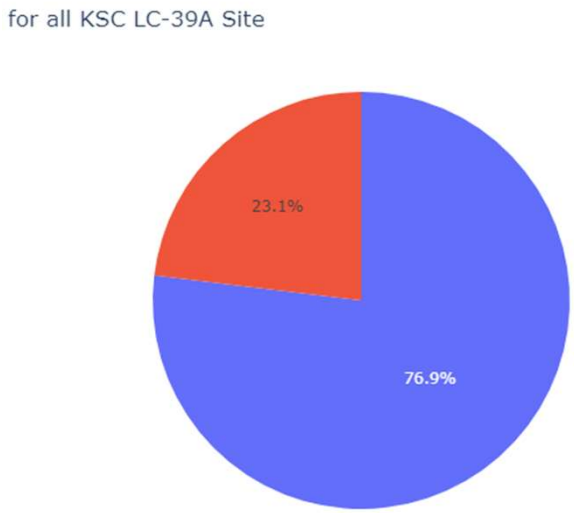
Launch success count for all sites

- The chart clearly indicates that KSC LC-39A has the highest number of successful launches among all the sites.
- While on the other hand ,Contribution by different Launch Site are as follows :
  - CCAFS LC-40 (29.2%)
  - VAFB SLC-4E (16.7%)
  - CCAFS SLC-40 (12.5%)



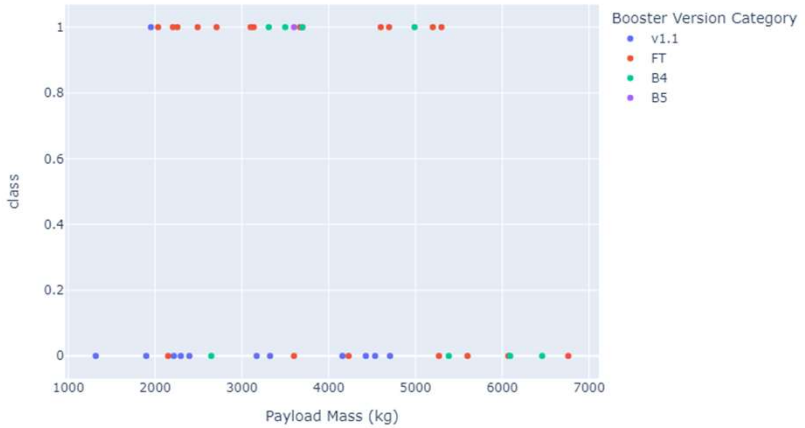
Launch site with highest launch success ratio

- The pie chart shows the launch success rate for the Kennedy Space Center Launch Complex 39A (KSC LC-39A).
- Key Information:
  - Highest Launch Success Rate: KSC LC-39A has the highest launch success rate among all launch sites, with a success rate of 76.9%.
  - Successful Launches: There have been 10 successful launches from this site.
  - Failed Landings: There have been 3 failed landings at this site.
  - Success vs. Failure: The blue section of the pie chart (76.9%) represents successful launches, while the red section (23.1%) represents failed landings.



Payload Mass vs. Launch Outcome for all sites

- The chart shows a scatter plot with various Booster Version Categories represented by different colors. The x-axis indicates the payload mass (in kilograms), and the y-axis represents the success or failure of the booster launches.
- Here's a general breakdown of the chart: X-axis (Payload Mass):
  - Represents the weight of the payload that each booster version attempted to carry.
  - Y-axis (Success/Failure): Represents whether the launch was successful or not.
  - Typically, success could be denoted at the top and failure at the bottom.
  - Color Coding: Different colors represent different booster versions. Each color-coded point represents an individual launch attempt, categorized by its booster version.



Section 5

Predictive Analysis (Classification)

Metrics on Test Set:

	Model	Accuracy	Jaccard Score	F1 Score
0	Logistic Regression	0.833333	0.800000	0.888889
1	Support Vector Machine	0.833333	0.800000	0.888889
2	Decision Tree	0.888889	0.846154	0.916667
3	K Nearest Neighbors	0.833333	0.800000	0.888889

Metrics on Entire Dataset:

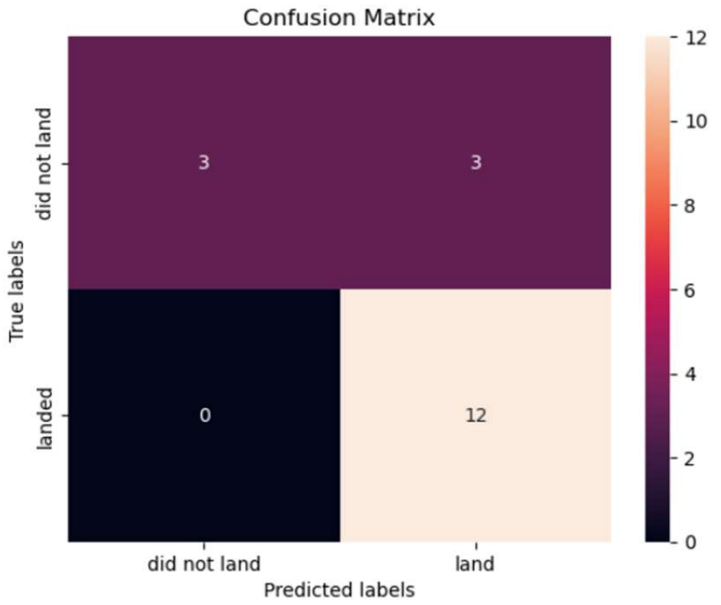
	Model	Accuracy	Jaccard Score	F1 Score
0	Logistic Regression	0.866667	0.833333	0.909091
1	Support Vector Machine	0.877778	0.845070	0.916031
2	Decision Tree	0.844444	0.784615	0.879310
3	K Nearest Neighbors	0.855556	0.819444	0.900763

Classification Accuracy

- Based on the Test Set scores, it's not possible to determine which method performs best.
- The identical Test Set scores might be due to the small test sample size of 18 samples. To address this, we evaluated all methods using the entire dataset.
- The results from the entire dataset confirm that the Support Vector Machine Model is the best. This model not only has higher scores but also the highest accuracy.

Confusion Matrix

- By examining the confusion matrix, we can see that SVM effectively distinguishes between different classes. However, the main issue with this model is the number of false positives.



Conclusions

- The Support Vector Model proves to be the most effective algorithm for this dataset.
- Launches with lower payload masses tend to have better outcomes compared to those with higher payload masses.
- Most launch sites are situated near the Equator and are all located close to the coast.
- The success rate of launches has improved over the years.
- KSC LC-39A boasts the highest success rate among all launch sites.
- Orbits such as ES-L1, GEO, HEO, and SSO exhibit a 100% success rate.

Appendix

Data Collection and Wrangling

```
import requests
import pandas as pd
from bs4 import BeautifulSoup
```

Exploratory Data Analysis (EDA) with Visualization

```
Import matplotlib.pyplot as plt
import seaborn as sns
```

Building Interactive Map with Folium

```
import folium
```

Building a Dashboard with Plotly Dash

```
import dash
import dash_core_components as dcc
import dash_html_components as html
import plotly.express as px
```

Predictive Analysis (Classification)

```
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, jaccard_score, f1_score
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
```

Data Preprocessing

```
from sklearn.preprocessing import OneHotEncoder
```

Thank you!