IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The data analysis process included several methodologies :
  - Data was gathered through web scraping and the SpaceX API.
  - Exploratory Data Analysis (EDA) was performed, which involved data cleaning, visualization, and interactive visual analysis.
  - Machine learning models were then used to make predictions.
  - Summary of results
  - Public data sources provided valuable information for the analysis.
  - EDA revealed the most significant features for predicting the success of launches.
  - Machine learning models helped identify the key characteristics influencing the likelihood of success, leveraging the entire dataset effectively.

Introduction

- The goal is to assess whether the new company, Space Y, can effectively compete with SpaceX.
- Key points of evaluation include:
- Desirable answers:
  - By predicting the likelihood of successful first-stage rocket landings, we can better estimate the overall costs of launches.
  - Identifying the best geographic locations for launches, based on factors like cost-efficiency, success rates, and environmental conditions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data regarding SpaceX was gathered from two key sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Data Wrangling:
  - The collected data was refined and enhanced by creating a landing outcome label, which was derived after summarizing and analyzing the available features
- Exploratory Data Analysis (EDA):EDA was conducted using both data visualization techniques and SQL queries to uncover insights and patterns in the data.

Methodology

Executive Summary

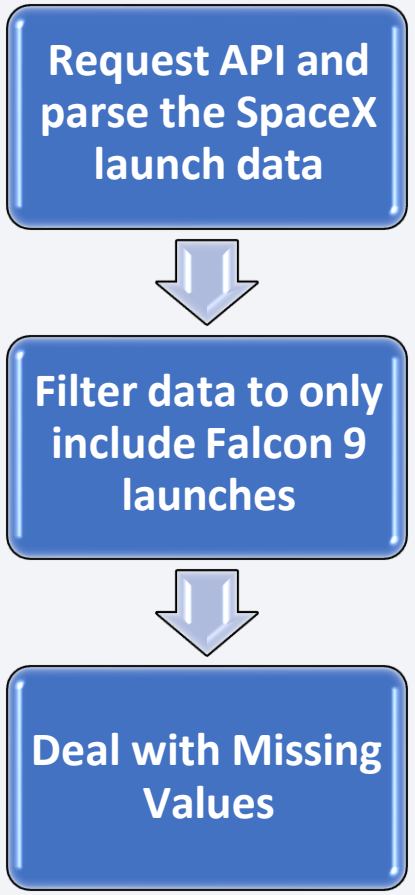
- Interactive Visual Analytics: Interactive visualizations were created using tools such as Folium for mapping and Plotly Dash for dashboards to analyze the data visually and interactively
- Predictive Analysis: Predictive modeling was performed using classification models.The collected data was normalized and split into training and test datasets. Four different classification models were applied, and their accuracy was evaluated using various combinations of parameters to determine the best-performing model.

Data Collection

- The datasets were gathered from two sources:SpaceX API: https://api.spacexdata.com/v4/rockets/
- Wikipedia: List of Falcon 9 and Falcon Heavy launches, using web scraping techniques.

Data Collection - SpaceX API

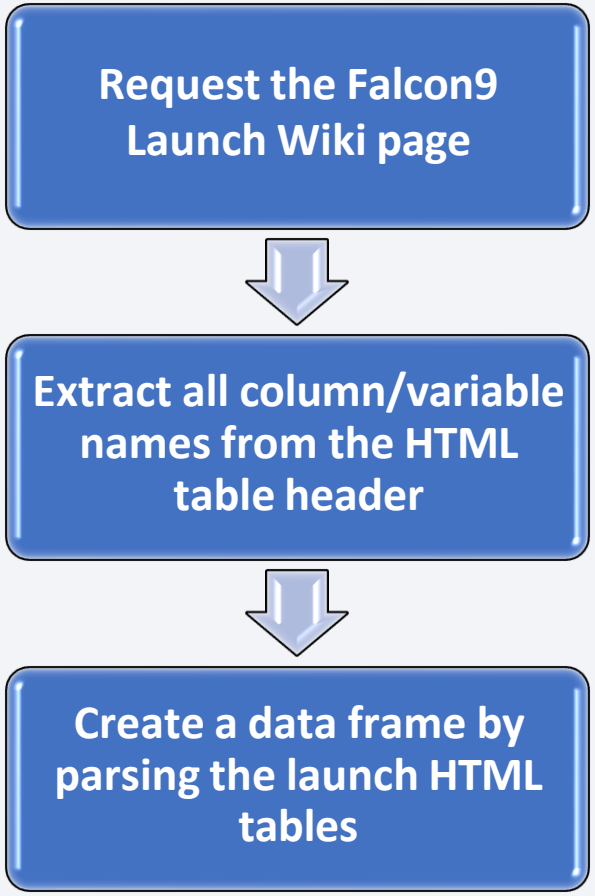
- SpaceX provides a public API that allows easy access to data, which can then be utilized for analysis.The API was integrated following the outlined flowchart, and the retrieved data was subsequently stored for further use.



- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Data on SpaceX launches can also be accessed from Wikipedia
- The data was downloaded following the process described in the flowchart and then stored for future analysis.
- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



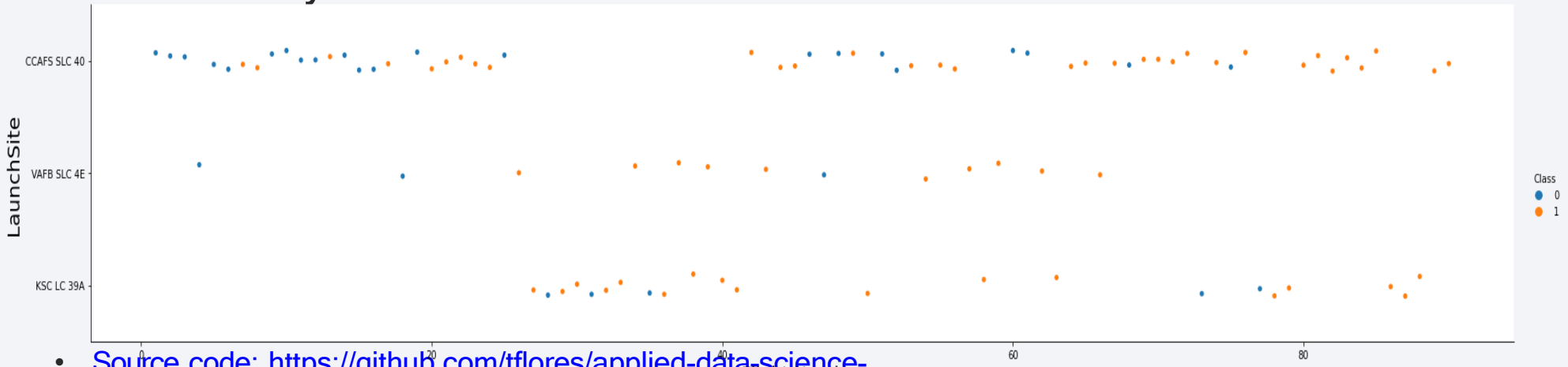
Data Wrangling

- Exploratory Data Analysis (EDA) was conducted on the dataset at the outset.
- This involved calculating the number of launches per site, the frequency of each orbit, and the mission outcomes categorized by orbit type..
- As a final step, a landing outcome label was generated based on the data in the Outcome column.

- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- To analyze the data, scatterplots and barplots were utilized to visualize the relationships between pairs of features, specifically:
- Payload Mass vs. Flight
- NumberLaunch Site vs.
- Flight NumberLaunch Site vs. Payload MassOrbit vs. Flight NumberPayload vs. Orbit



- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The following SQL queries were executed to extract insights from the dataset::
  - Retrieve the names of the unique launch sites involved in the space missions;
  - Identify the top 5 launch sites whose names start with the string 'CCA.';
  - Calculate the total payload mass carried by boosters launched by NASA (CRS).;
  - Determine the average payload mass carried by the booster version F9 v1.1.;
  - Find the date when the first successful landing outcome on a ground pad was achieved.;
  - List the names of boosters that succeeded in landing on a drone ship with a payload mass between 4,000 and 6,000 kg;
  - Count the total number of successful and failed mission outcomes ;
  - Identify the names of booster versions that have carried the maximum payload mass.;
  - Extract the failed landing outcomes on drone ships, along with their booster versions and launch site names for the year 2015; and
  - Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates June 4, 2010, and March 20, 2017..

- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/EDA.ipynb

Build an Interactive Map with Folium

Folium Maps utilized various visual elements for enhanced data representation:

- **Markers**: Represent specific points, such as launch sites.
- **Circles**: Highlight areas surrounding specific coordinates, such as the NASA Johnson Space Center.
- **Marker Clusters**: Group events at each coordinate, illustrating the number of launches occurring at a launch site.
- **Lines**: Indicate distances between two coordinates, providing a clear visual of spatial relationships.

- Source code: https://github.com/tflores/applied-data-science-capstone/blob/d232d76932163635b072952f121a8d70286e0d84/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

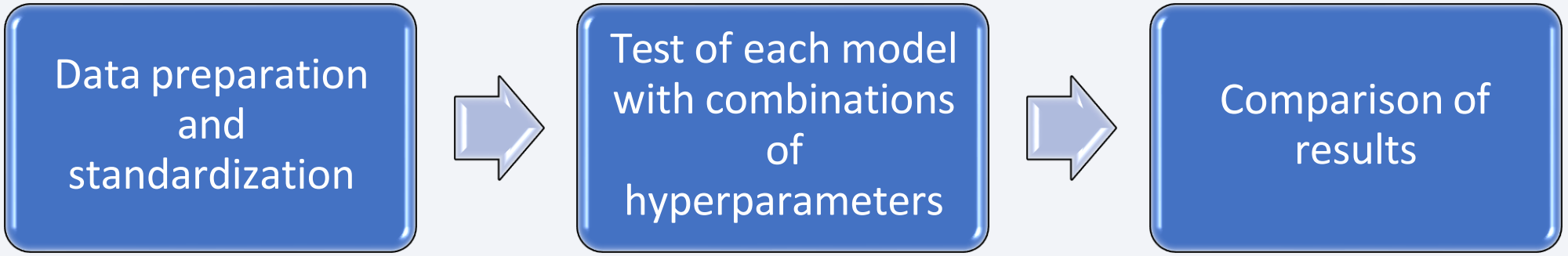
Build a Dashboard with Plotly Dash

The following graphs and plots were employed to visualize the data:
•**Percentage of Launches by Site**
•**Payload Range**
This combination facilitated a quick analysis of the relationship between payloads and launch sites, aiding in the identification of the most suitable locations for launches based on payload capacities.

- Source code: https://github.com/tflores/applied-data-science-capstone/blob/d232d76932163635b072952f121a8d70286e0d84/spacex_dash_app.py

Predictive Analysis (Classification)

Four classification models were compared in the analysis:
- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Decision Tree**
- **K-Nearest Neighbors (KNN)**



- Source code: https://github.com/tflores/applied-data-science-capstone/blob/d232d76932163635b072952f121a8d70286e0d84/Machine%20Learning%20Prediction.ipynb

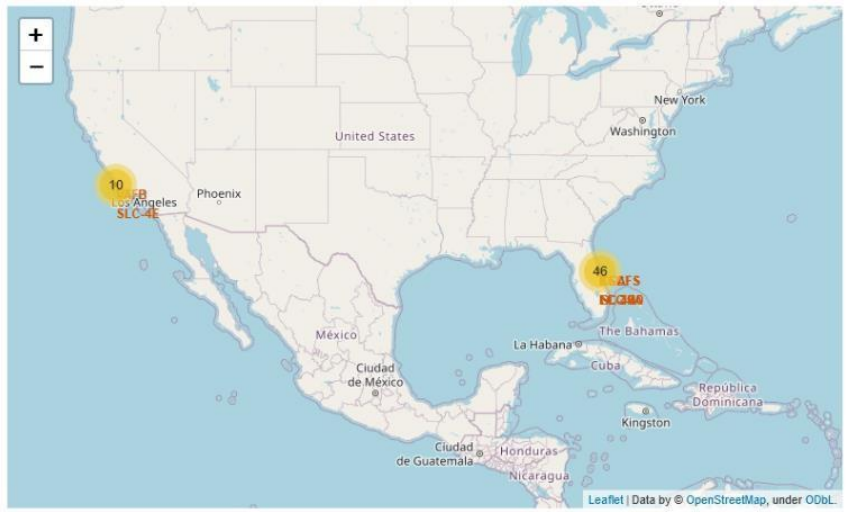
Results

Exploratory Data Analysis Results:

- SpaceX operates four different launch sites.
- The initial launches were conducted for SpaceX and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The first successful landing outcome occurred in 2015, five years after the first launch.
- Many Falcon 9 booster versions successfully landed on drone ships with payloads exceeding the average.
- Nearly 100% of mission outcomes were successful.
- Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015.
- The number of successful landing outcomes improved over the years.

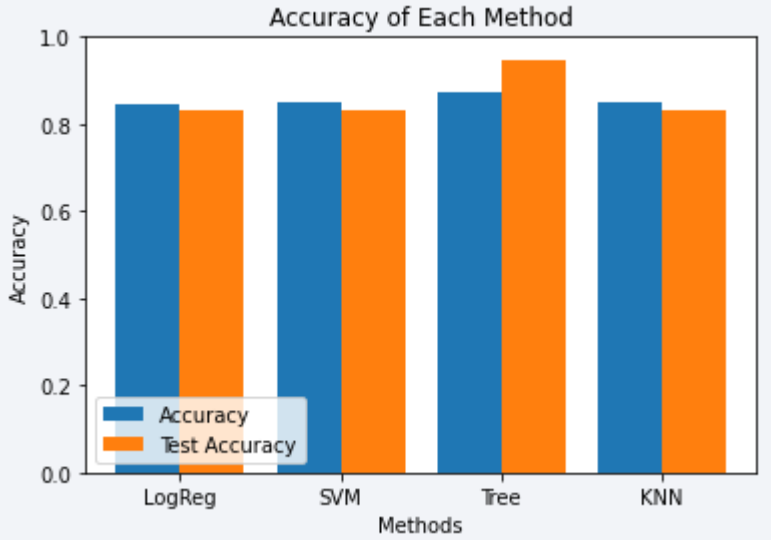
Results

- Through interactive analytics, it was identified that launch sites are typically located in safe areas, often near the sea, and are supported by robust logistical infrastructure. Additionally, most launches occur at east coast launch sites.



Results

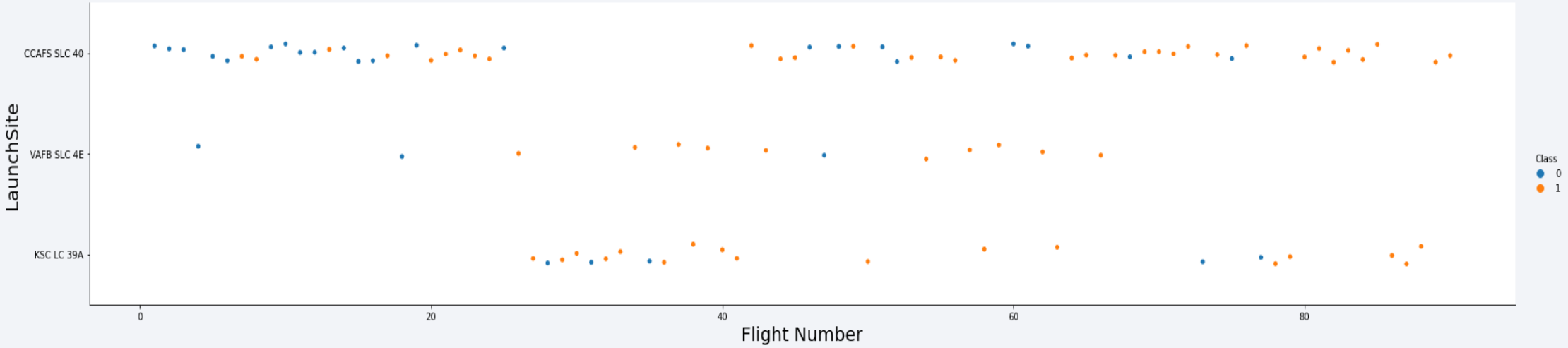
- The predictive analysis revealed that the **Decision Tree Classifier** is the most effective model for predicting successful landings, achieving an accuracy of over 87% and a test data accuracy exceeding 94%.



Section 2

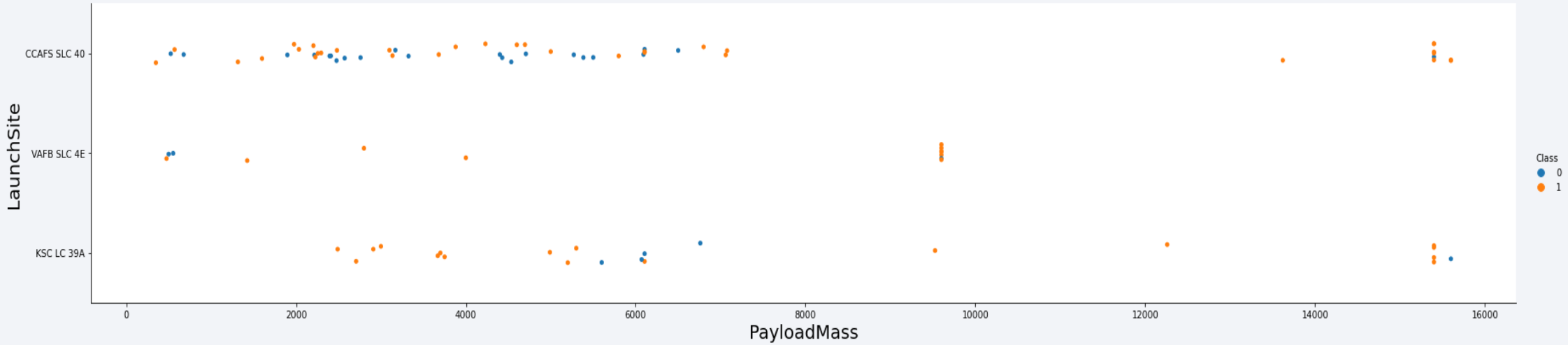
Insights drawn from EDA

Flight Number vs. Launch Site



- Based on the plot, it is evident that the best launch site currently is **CCAF5 SLC 40**, which has seen the highest number of successful recent launches.
  In second place is **VAFB SLC 4E**, followed by **KSC LC 39A** in third.
  Additionally, the overall success rate has shown improvement over time.

Payload vs. Launch Site



- Payloads exceeding 9,000 kg (approximately the weight of a school bus) exhibit an excellent success rate.
  However, payloads over 12,000 kg appear to be feasible only at the **CCAFS SLC 40** and **KSC LC 39A** launch sites.

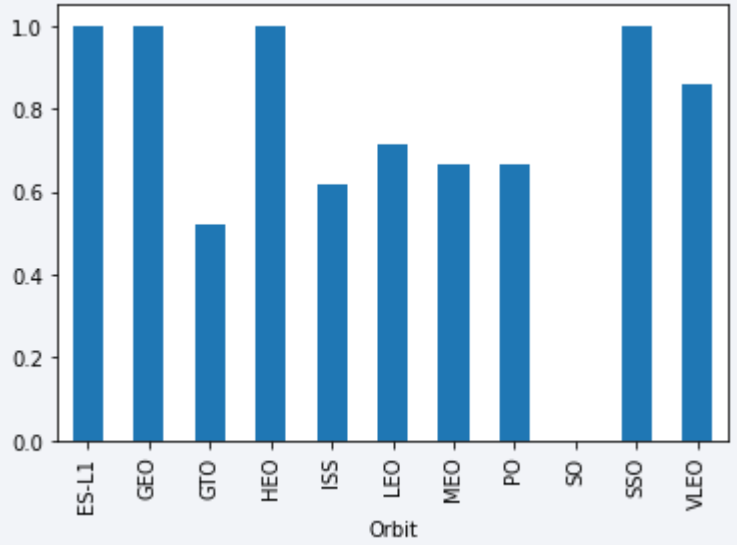
Success Rate vs. Orbit Type

The highest success rates are observed for the following orbits:

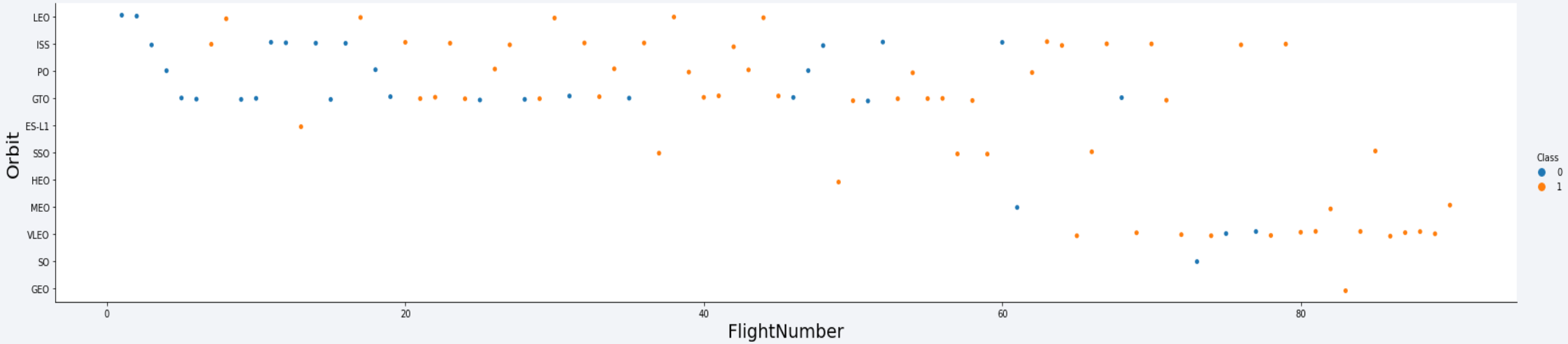
- **ES-L1**
- **GEO**
- **HEO**
- **SSO**

These are closely followed by:

- **VLEO** (above 80%)
- **LFO** (above 70%)

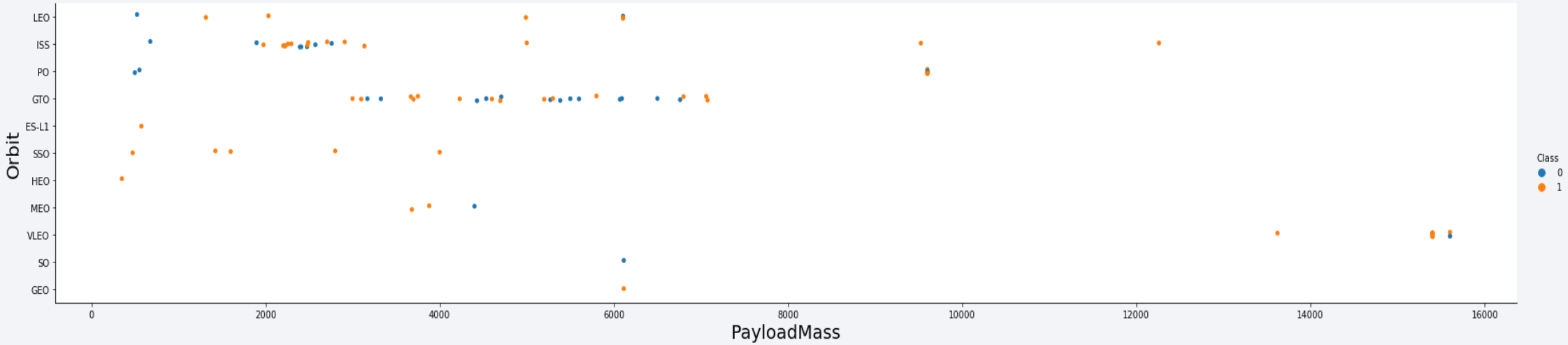


Flight Number vs. Orbit Type



- It seems that the success rate has improved over time for all orbits.
  Additionally, the **VLEO** orbit appears to be a promising business opportunity, given its recent increase in launch frequency.

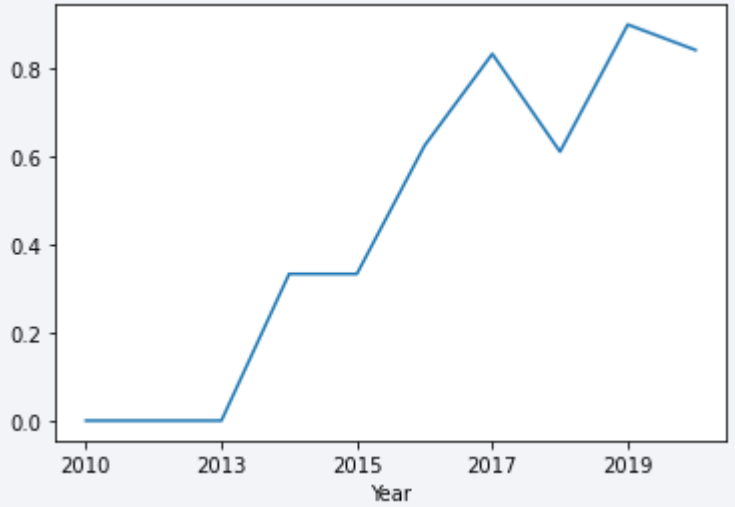
Payload vs. Orbit Type



- It appears that there is no correlation between payload and success rate for the **GTO** orbit.
  The **ISS** orbit demonstrates the widest range of payloads while maintaining a good success rate.
  Additionally, there have been relatively few launches to the **SO** and **GEO** orbits.

Launch Success Yearly Trend

The success rate began to rise in 2013 and continued to improve until 2020. The initial three years appear to have been a period of adjustments and technological enhancements.



All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- The total payload was calculated by summing all payloads with codes that include 'CRS,' which corresponds to NASA missions.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- By filtering the data for successful landing outcomes on a ground pad and retrieving the minimum date value, it is possible to identify that the first occurrence happened on **December 22, 2015**.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping the mission outcomes and counting the records for each group resulted in the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

These are the boosters that have carried the maximum payload mass recorded in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

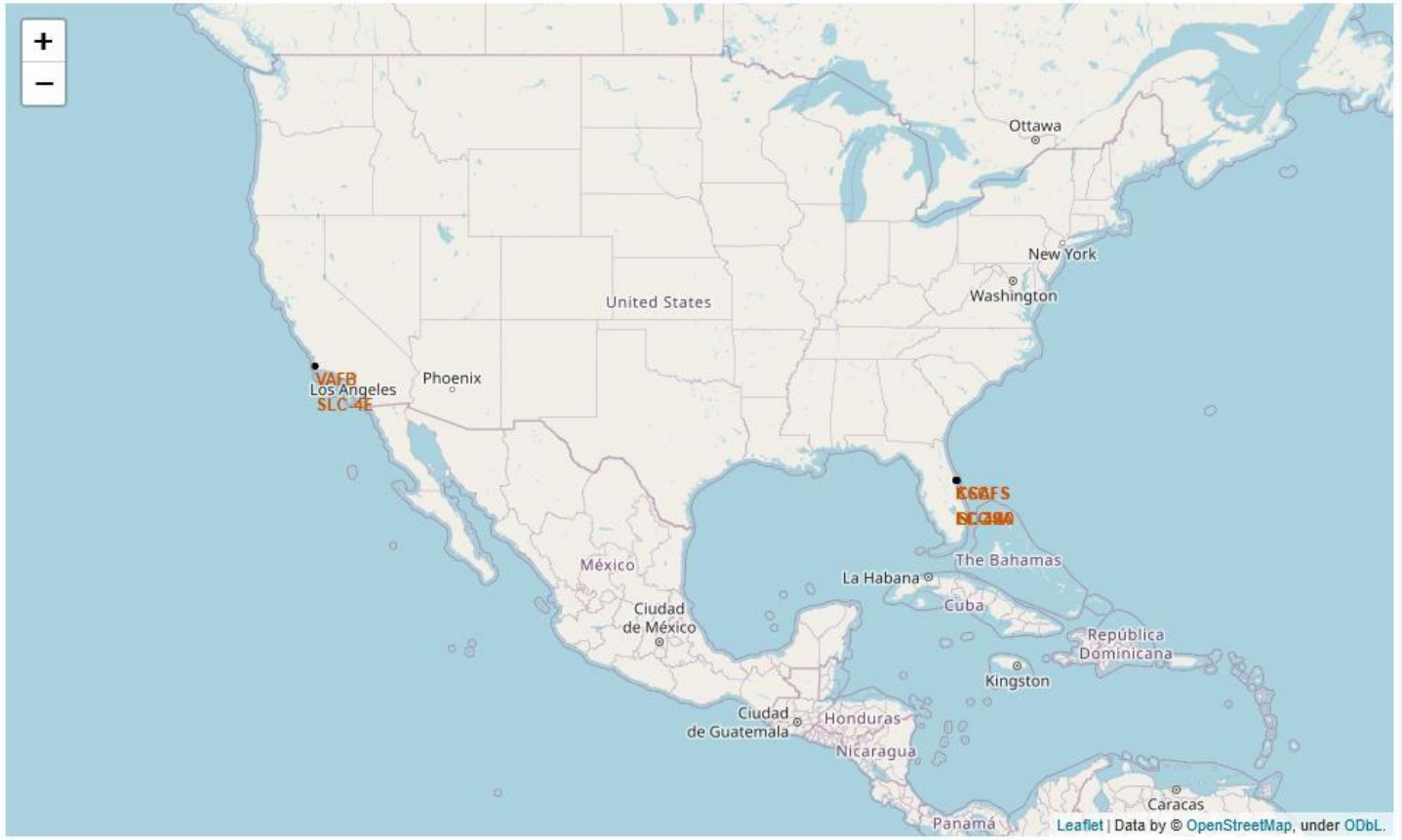
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

This data insight highlights the importance of considering "No attempt" outcomes in the analysis

Section 4

Launch Sites Proximities Analysis

All launch sites



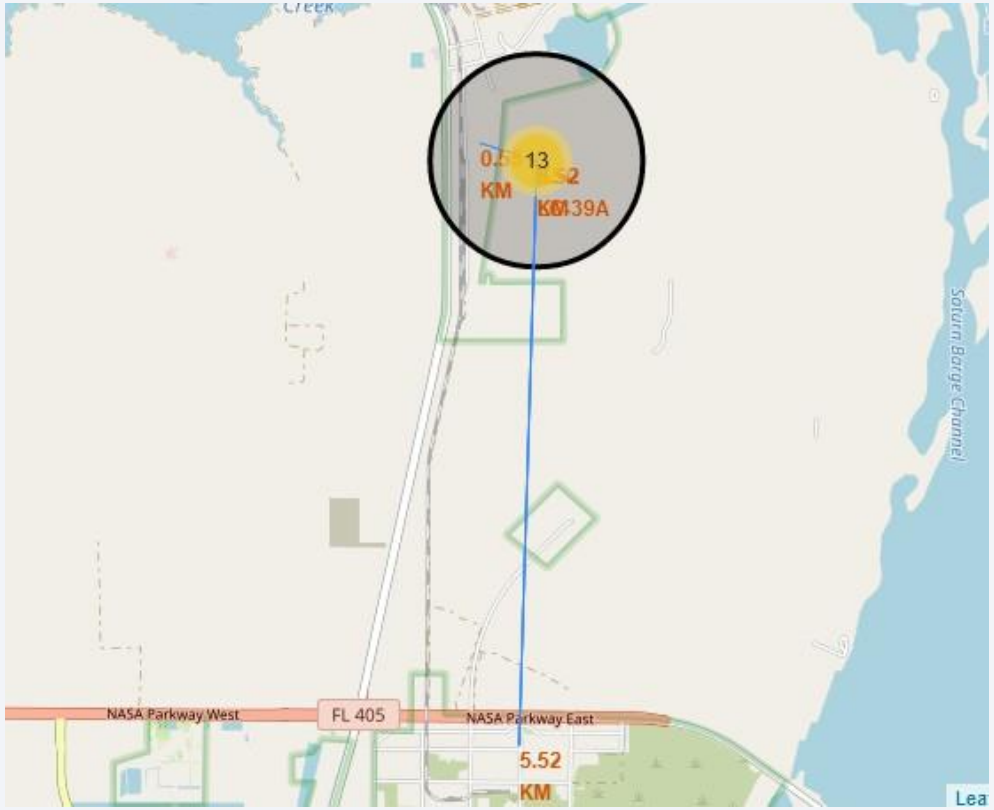
- Launch sites are typically located near the sea, likely for safety reasons, while still being in proximity to roads and railways for logistical support

Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

- Green markers represent successful outcomes, while red markers indicate failures.

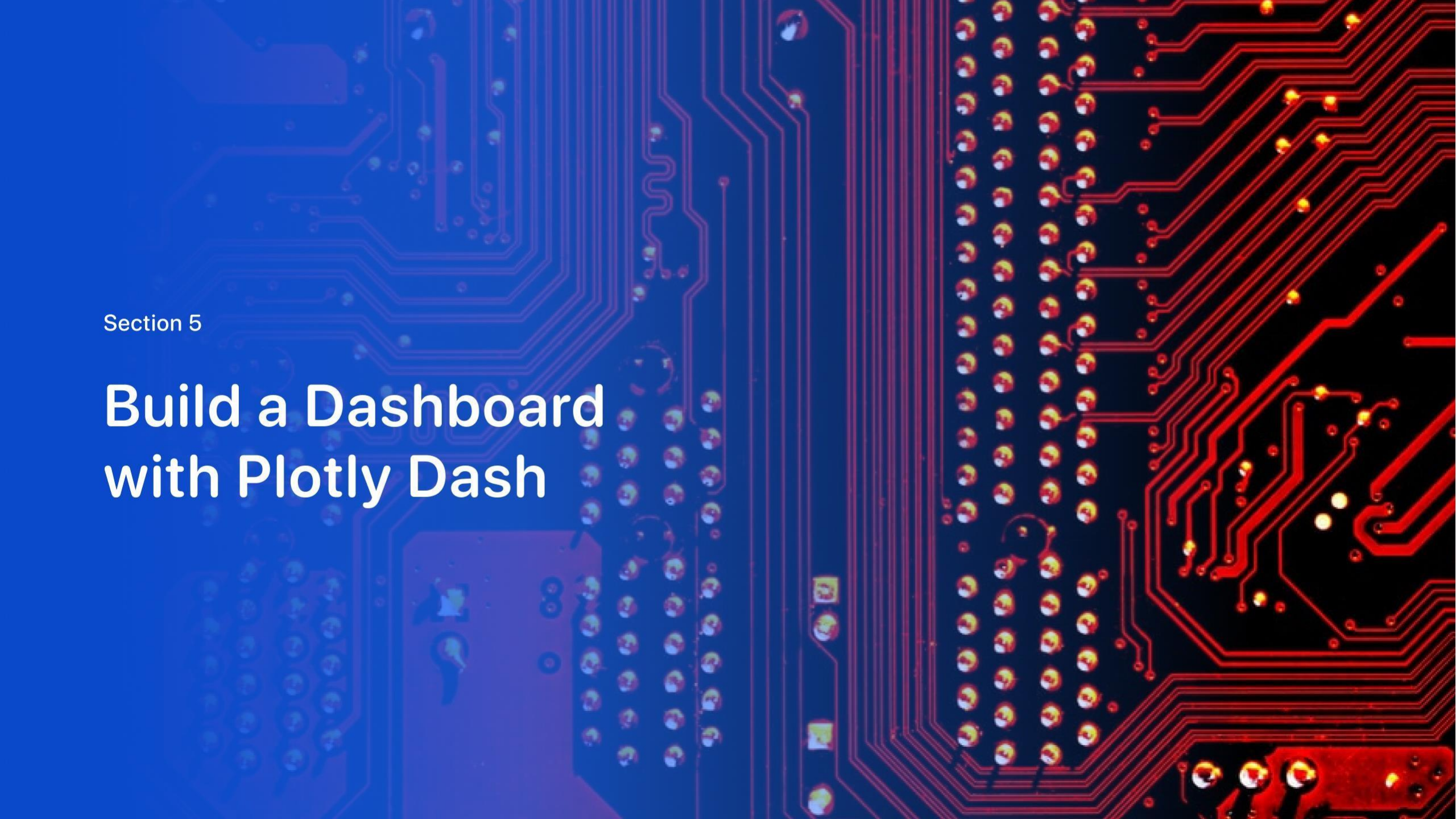
Logistics and Safety



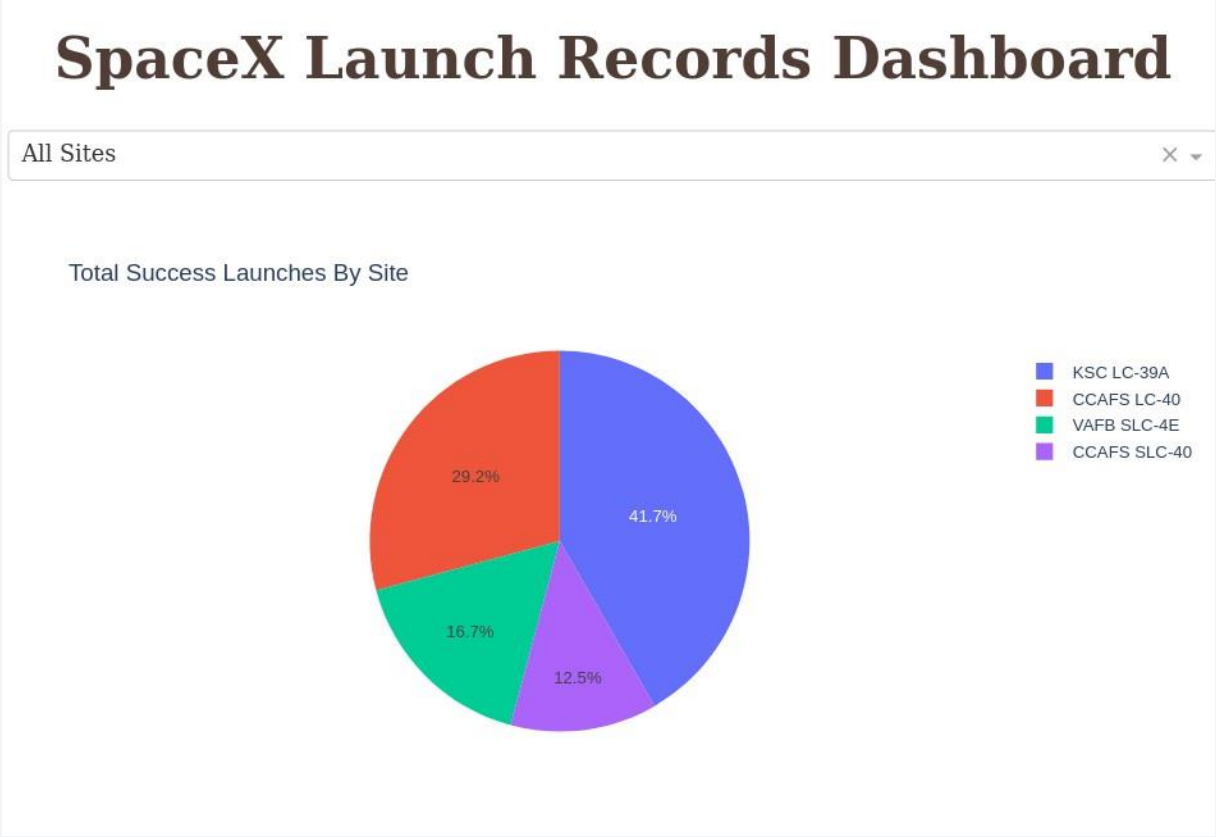
- The launch site **KSC LC-39A** has favorable logistical features, as it is located near both a railroad and road while being relatively distant from populated areas.

Section 5

Build a Dashboard with Plotly Dash

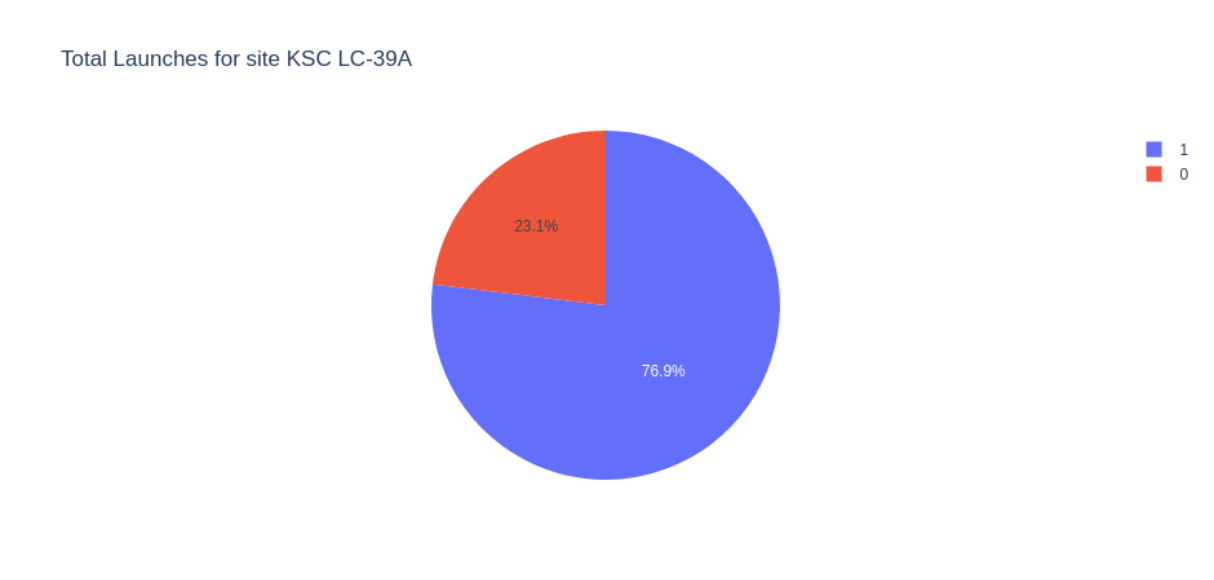


Successful Launches by Site



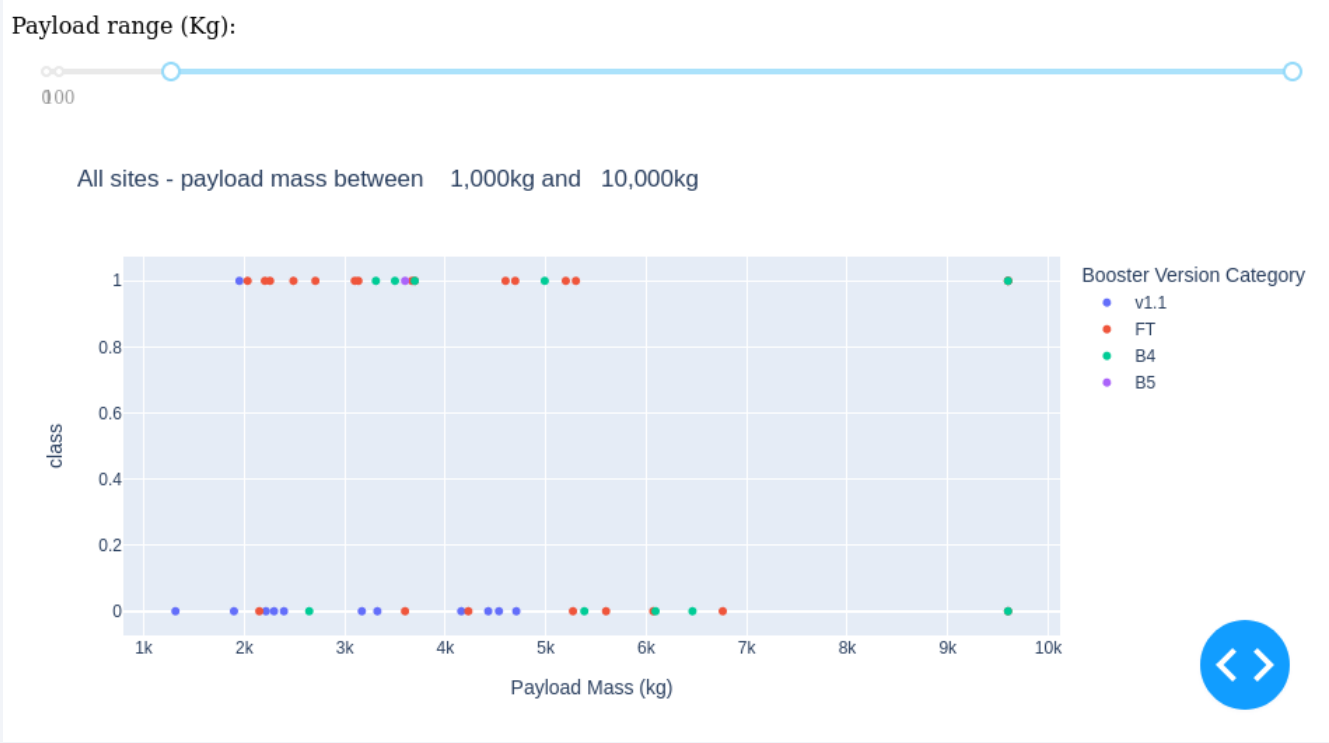
- The launch site appears to be a critical factor influencing the success of missions.

Launch Success Ratio for KSC LC-39A



- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.

Payload vs. Launch Outcome

- There’s not enough data to estimate risk of launches over 7,000kg

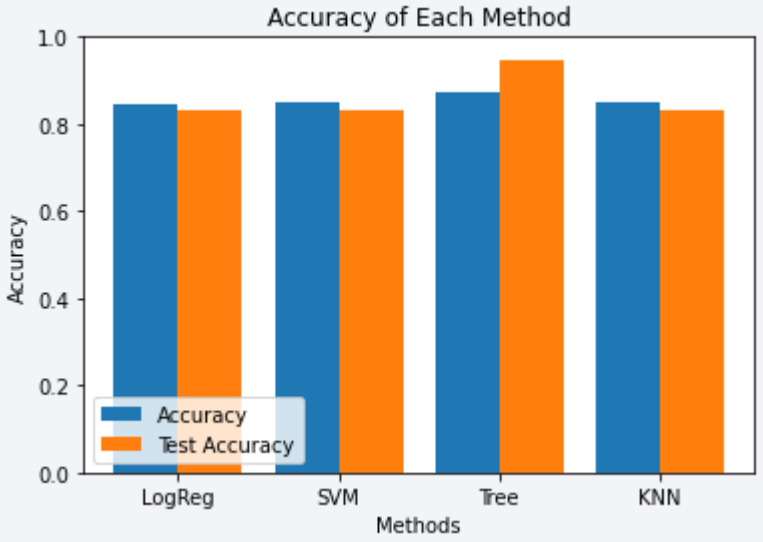
Section 6

Predictive Analysis (Classification)

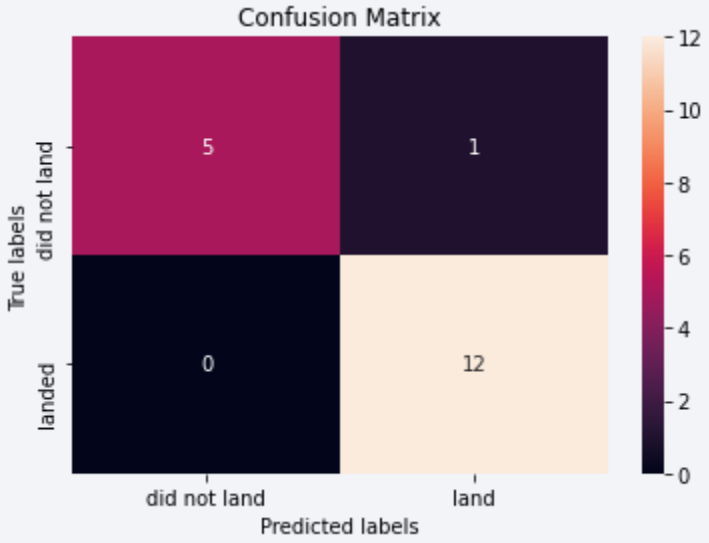
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Various data sources were examined, leading to more precise conclusions throughout the process.
- The optimal launch location is KSC LC-39A.
- Launches exceeding 7,000 kg tend to carry lower risks.
- While the majority of mission results are positive, the rate of successful landings appears to improve over time due to advancements in processes and rocket technology.
- A Decision Tree Classifier can be employed to forecast successful landings and enhance profitability.

Thank you!