IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name : Shivasharanu
Date : July 14, 2025

Outline

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

Executive Summary

- **Project Objective**

  To build a predictive model that determines the likelihood of Falcon 9 rocket first-stage landing success, enabling cost-effective launch decisions.

- **Key Methodologies**
  - Data collection via SpaceX API + web scraping
  - Data cleaning & feature engineering (landing outcome, payload, site, etc.)
  - Exploratory Data Analysis (EDA)
  - Model building with Logistic Regression, SVM, Decision Tree, KNN
  - Model tuning using GridSearchCV with 10-fold cross-validation
  - Interactive visualizations using Plotly & Folium
- **Summary of all results**
  - Best Model: Decision Tree
  - Cross-Validation Accuracy: 88.57%
  - Key Influencing Features: Launch Site, Payload Mass, Orbit

Introduction

- Project Background & Context
  - SpaceX has revolutionized space transportation by making rocket reusability a core mission strategy.
  - The ability to successfully land the first-stage booster after launch is critical to reducing costs and ensuring sustainability.
  - Predicting the success of landings helps optimize planning, minimize risk, and improve operational efficiency.
- This project aims to answer the following:
  - Can we predict whether a Falcon 9 first-stage booster will land successfully?
  - What factors (e.g., launch site, payload, orbit type) most influence landing success?
  - Which machine learning model gives the most accurate predictions?
  - How can this model support cost-effective and risk-aware launch decision-making?

Section 1

Methodology

Methodology

Executive Summary

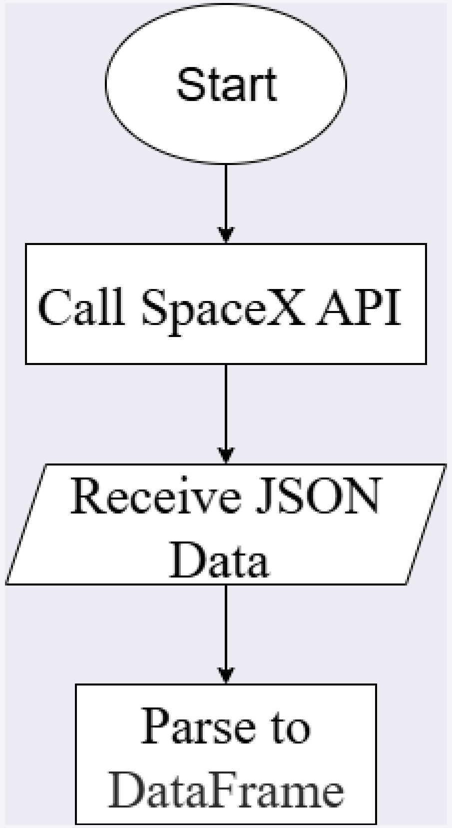
- Data collection methodology:
  - Launch data was collected using the SpaceX REST API and web scraping from Wikipedia to compile relevant features like payload, launch site, and landing outcome
- Perform data wrangling
  - The collected data was cleaned and structured by handling missing values, converting data types, and engineering new features for modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models—Logistic Regression, SVM, Decision Tree, and KNN—were trained to predict first-stage landing success.

Data Collection

- The dataset was collected using two primary sources:
- **SpaceX REST API** – to retrieve structured launch data (e.g., flight number, launch site, payload mass, orbit, landing success).
- **Wikipedia Web Scraping** – to extract additional launch information such as booster version and detailed launch outcomes.
- The collected raw data was then **merged**, **filtered**, and **stored in structured format** for further cleaning and analysis.

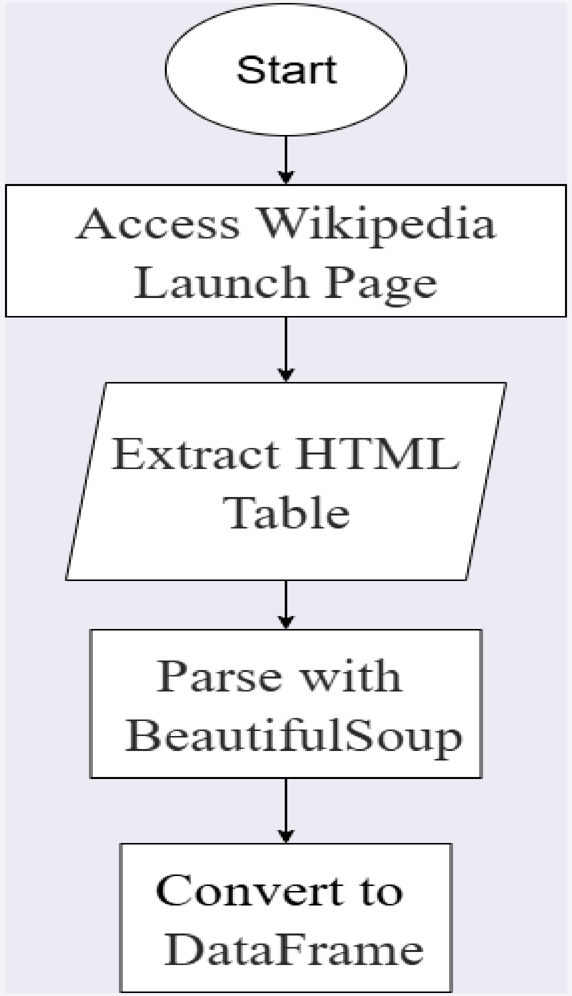
Data Collection – SpaceX API

- Data was collected by making **SpaceX REST API calls** to retrieve structured launch data in JSON format, which was then parsed into a DataFrame for further processing.
- GitHub URL of the completed SpaceX API calls notebook https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/jupyter-labs-spacex-data-collection-api-v2.ipynb



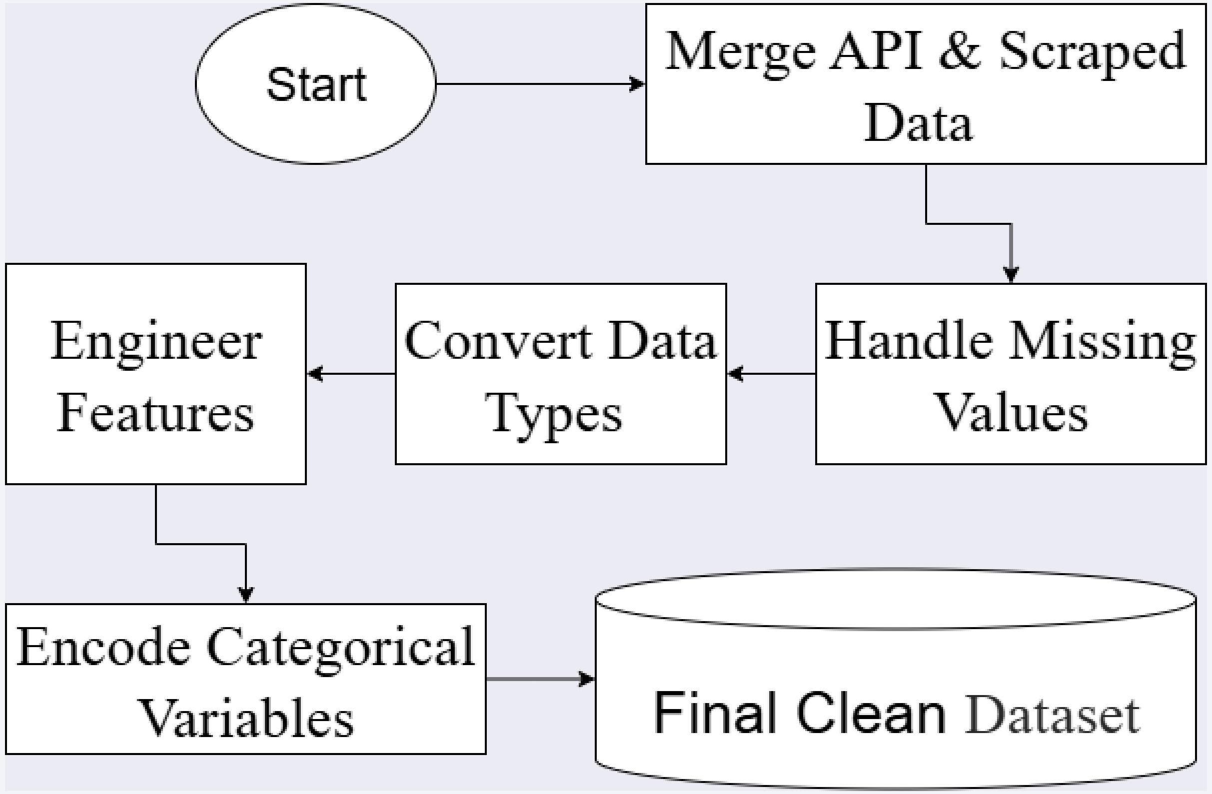
Data Collection - Scraping

- Web scraping was performed on the **Wikipedia Falcon 9 launch history page** to extract tabular launch data using BeautifulSoup, which was then parsed into a structured DataFrame.
- GitHub URL of the completed web scraping notebook https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/jupyter-labs-webscraping.ipynb



Data Wrangling

- The collected data was processed by cleaning null values, converting data types, engineering new features (e.g., landing outcome), and preparing it for analysis and modeling.
- GitHub URL of your completed data wrangling related notebooks https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

- Various charts such as **scatter plots, pie charts, bar graphs, and geospatial maps** were used to explore feature distributions, analyze relationships (e.g., payload vs. success), compare launch site performance, and visualize geographical launch patterns to gain insights into the factors influencing landing success.
- GitHub URL of your completed EDA with data visualization notebook https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- SELECT statements to retrieve key columns such as Launch Site, Payload Mass, and Mission Outcome.
- WHERE clauses to filter data based on conditions like Landing Outcome = 'Success' or specific Launch Site.
- GROUP BY queries to aggregate data by Launch Site and Orbit to count successes or failures.
- ORDER BY to sort launch records by payload mass, success rate, or launch date.
- COUNT() and AVG() functions to calculate total successful landings and average payload.
- LIMIT to display top N results (e.g., top 5 successful sites).

- GitHub URL of your completed EDA with SQL notebook https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We created multiple interactive Folium maps to visualize launch site locations, proximity to key infrastructure, and outcomes of SpaceX missions using markers, circles, lines, and popups for detailed spatial insights.
- **Markers**: Placed on launch sites and key locations (e.g., coastline, cities) to visually represent their coordinates and roles in analysis.
- **CircleMarkers**: Used to indicate launch sites with varying color or size to reflect success/failure status.
- **Lines (PolyLines)**: Drawn between launch sites and nearest features (e.g., coastline, railways, highways) to show proximity and analyze logistic relationships.
- **Popup Labels / Tooltips**: Added to markers to display detailed information like site name or distance.
- **Distance Markers (DivIcon)**: Used to display the calculated distances between launch sites and nearby features directly on the map.

- GitHub URL of your completed interactive map with Folium map https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

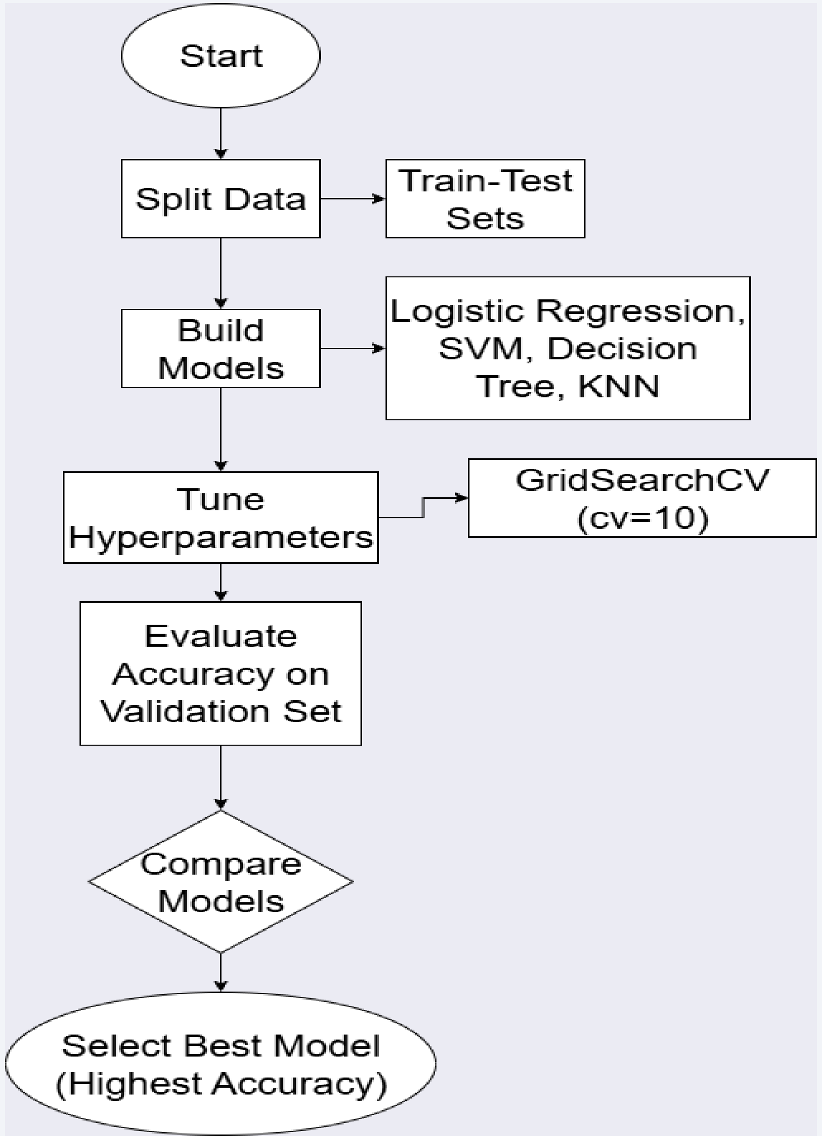
- The dashboard was built using **Plotly Dash**, incorporating interactive components like pie charts, scatter plots, dropdowns, and sliders to explore SpaceX launch data dynamically.
- Pie Chart: Shows overall success distribution across all launch sites or compares success vs. failure for a selected site — helps understand site-wise performance.
- Scatter Plot: Displays the relationship between payload mass and landing success — used to identify trends or thresholds affecting outcomes.
- Dropdown Menu: Allows users to select a specific launch site or view data for all sites — enables targeted analysis.
- Range Slider: Lets users filter payload mass range — helps explore how different payload sizes impact success rate.
- Dynamic Updates: All charts respond to user inputs (site selection, payload range) — makes the dashboard interactive and user-driven.

- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/spacex-dash-app.py

Predictive Analysis (Classification)

- We built multiple classification models (Logistic Regression, SVM, Decision Tree, KNN), tuned them using **GridSearchCV** with 10-fold cross-validation, evaluated them using accuracy scores, and selected the **best-performing model** based on validation performance.

- GitHub URL of your completed predictive analysis lab https://github.com/Shivasharanu-SN/Data-Science/blob/main/Applied%20data%20science%20project/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb



Results

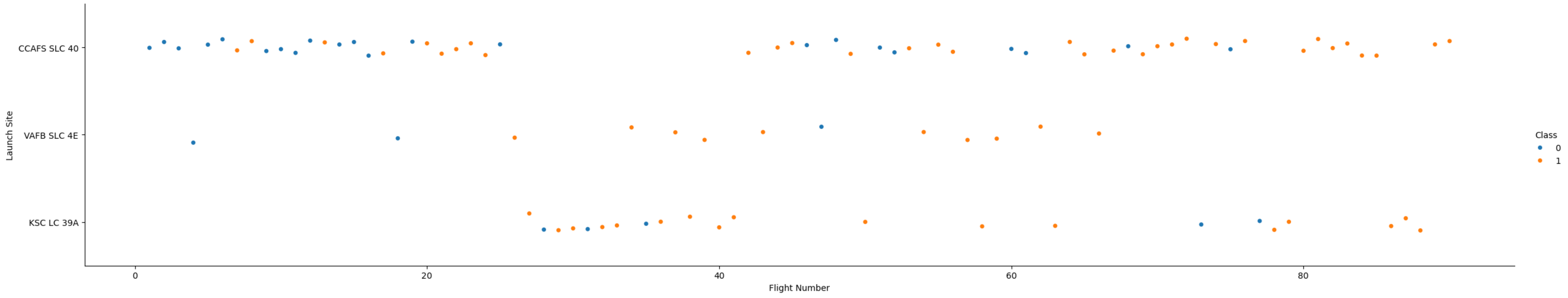
- Exploratory data analysis results
  - Launch success was found to vary significantly by launch site, orbit, and payload size.
  - Payloads within a moderate range showed higher success rates compared to extreme values.
- Interactive analytics
  - The dashboard allowed dynamic filtering by launch site and payload range.
  - Visual updates in pie and scatter plots provided real-time exploratory insights.
- Predictive analysis results
  - Among all models, Decision Tree performed best with the highest validation accuracy.
  - Launch Site, Orbit, and Payload Mass were key predictors of landing success.

Section 2

Insights drawn from EDA

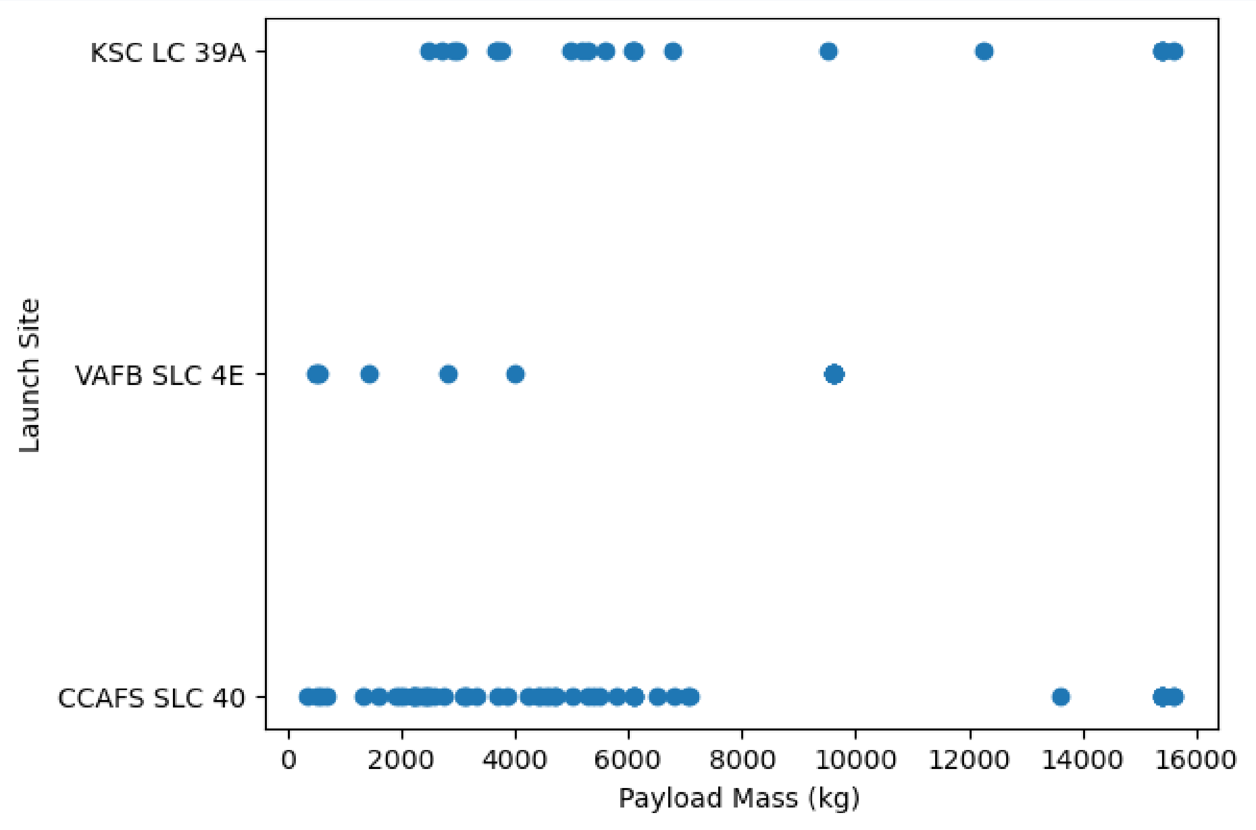
Flight Number vs. Launch Site

- The scatter plot of Flight Number vs. Launch Site shows how launches are distributed over time across different sites.
- It helps identify launch frequency patterns and site activity progression throughout SpaceX's mission history.



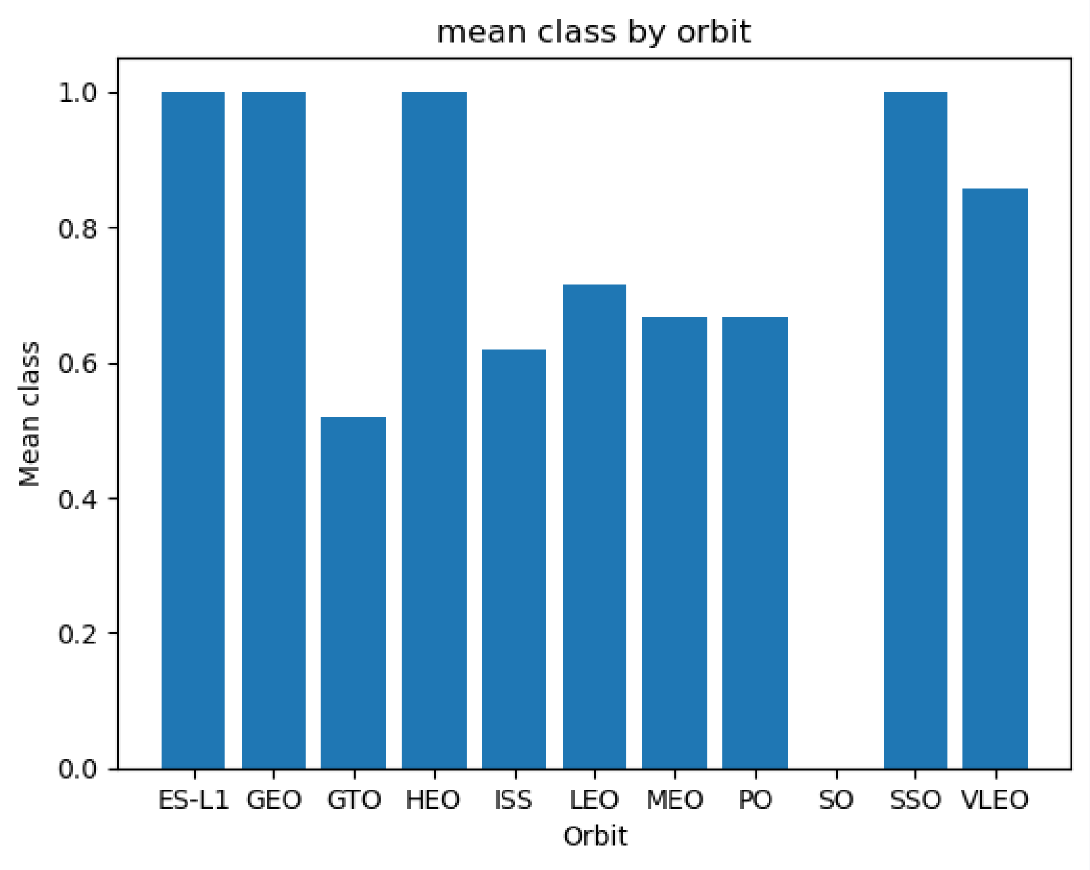
Payload vs. Launch Site

- The scatter plot of Payload Mass vs. Launch Site reveals how different sites handled varying payload sizes.
- It highlights that some launch sites are associated with heavier or more frequent payload missions, aiding in capacity analysis.



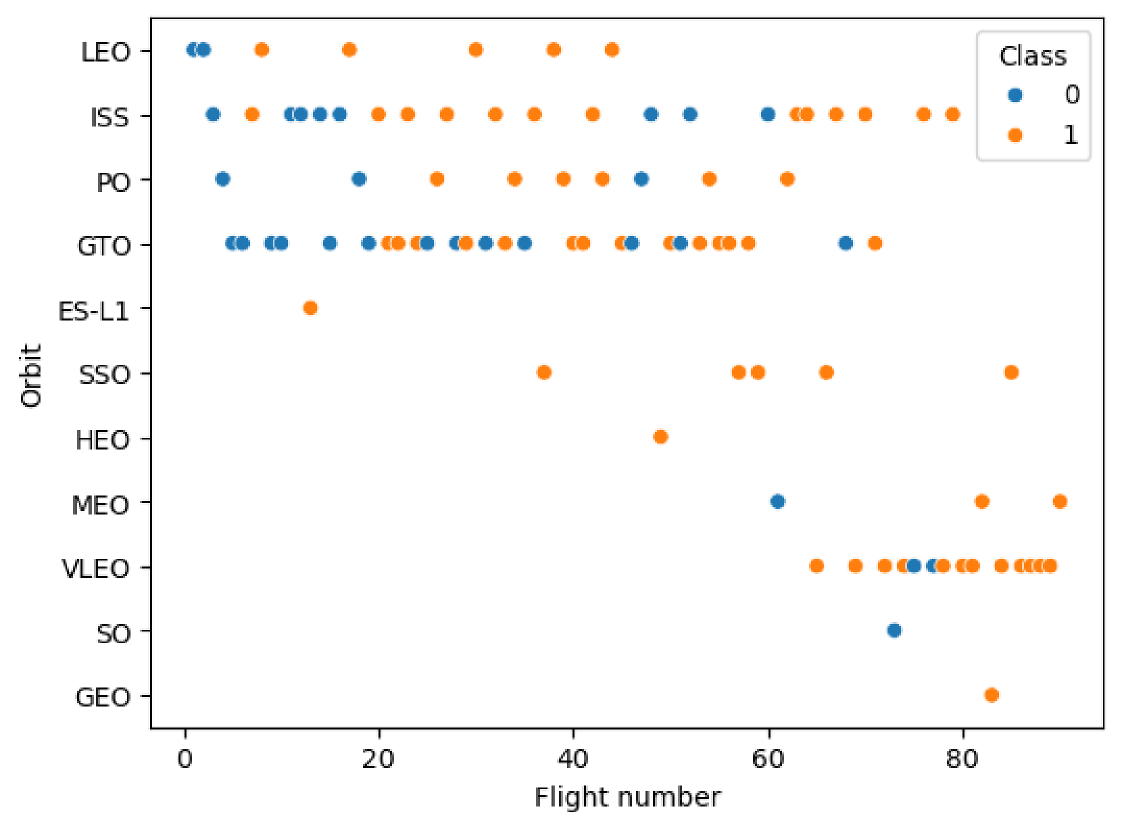
Success Rate vs. Orbit Type

- The bar chart displays the landing success rate for each orbit type, helping compare performance across mission profiles.
- It shows which orbits (e.g., GEO, SSO) are associated with higher success rates, supporting mission planning and risk analysis.



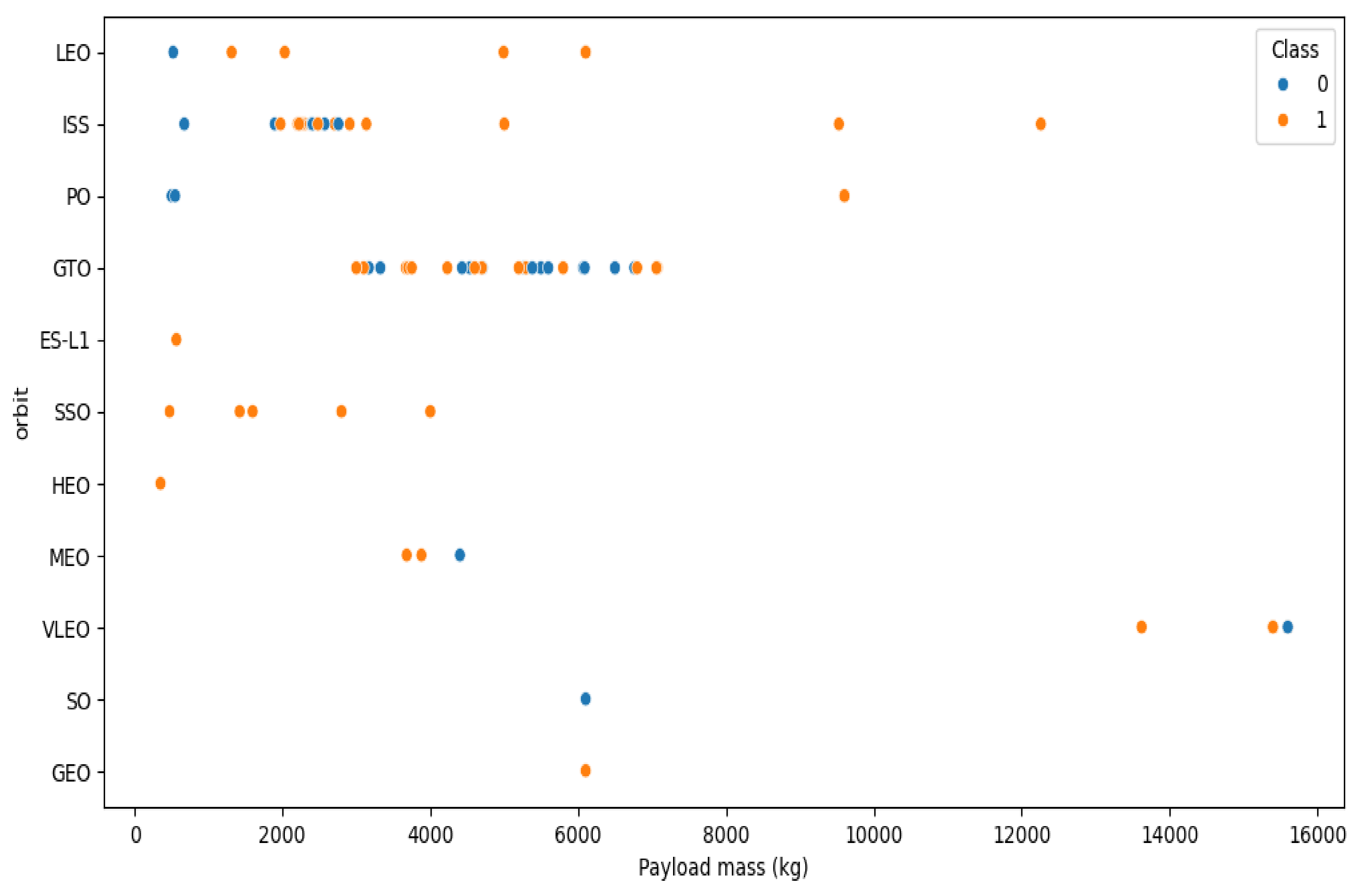
Flight Number vs. Orbit Type

- The scatter plot of Flight Number vs. Orbit Type illustrates how orbit selection has evolved over time across missions.
- It reveals trends such as increased use of certain orbits (e.g., VLEO or ISS) as SpaceX expanded its launch capabilities.



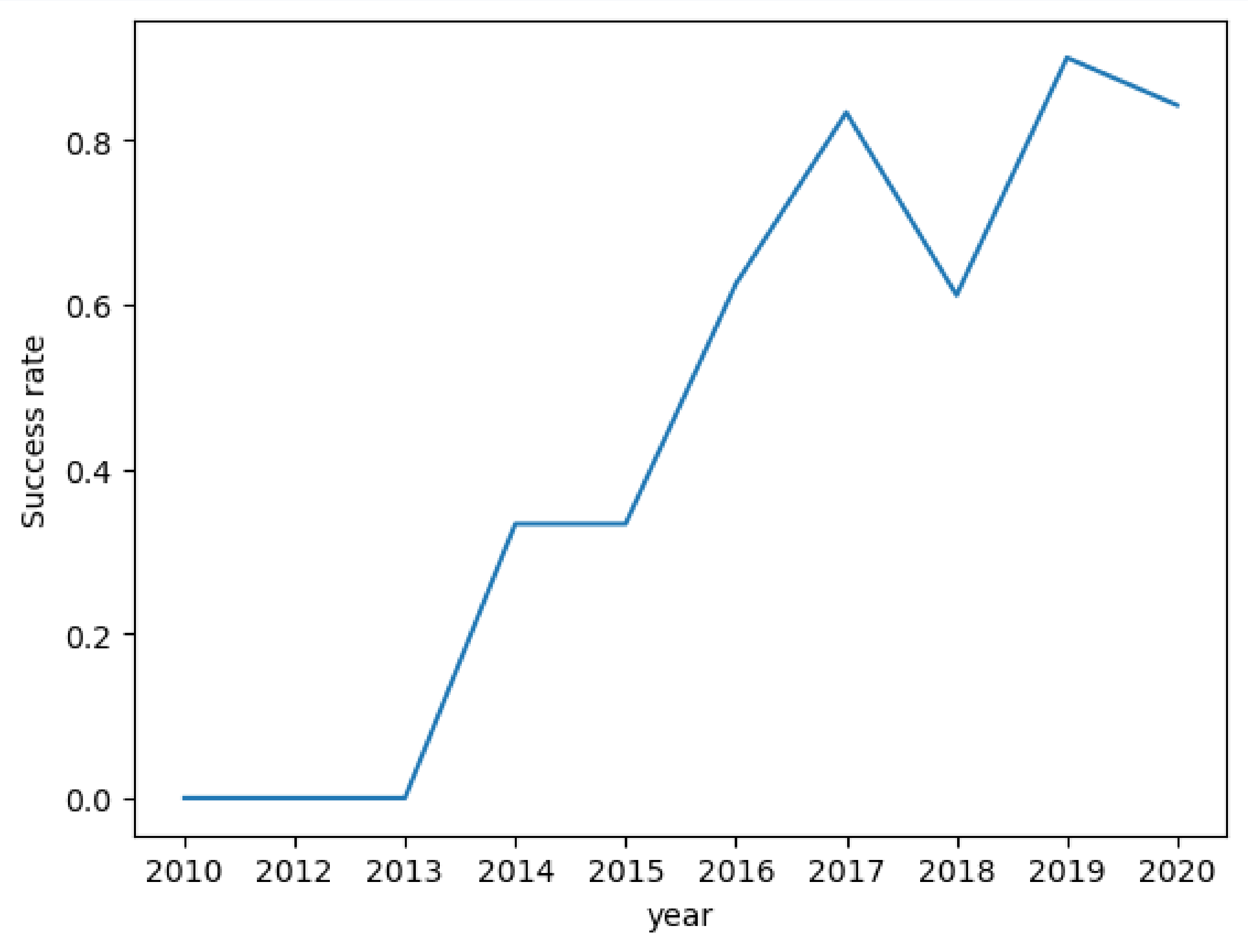
Payload vs. Orbit Type

- The scatter plot of Payload Mass vs. Orbit Type shows how different orbits are associated with varying payload capacities.
- It helps identify which orbits typically carry heavier or lighter payloads, providing insight into mission design constraints.



Launch Success Yearly Trend

- The line chart shows the average landing success rate per year, tracking SpaceX's performance over time.
- It highlights a clear upward trend, indicating consistent improvement in landing reliability year over year.



All Launch Site Names

- This query retrieves all distinct launch site names from the dataset, helping identify where SpaceX launches have taken place. These sites are used to analyze launch performance and frequency.

```
[10]: %sql select DISTINCT Launch_Site from spacextbl;

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

[10]:

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

Launch Site Names Begin with 'CCA'

[11]:
```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5;
```

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

[11]:

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

- This query filters the dataset to show the first five launch records where the launch site name starts with "CCA", which stands for Cape Canaveral Air Force Station. These sites are frequently used by SpaceX for missions to various orbits.

Total Payload Mass

```
[12]: %sql select sum(payload_mass__kg_) from spacextbl where customer like '%NASA%';

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

[12]:

sum(payload_mass__kg_)
107010

- This query calculates the total payload mass (in kg) carried by boosters for missions that involved NASA as a customer. The %NASA% pattern ensures that all entries mentioning NASA (e.g., "NASA (CRS)", "NASA/ULA") are included in the total.

Average Payload Mass by F9 v1.1

```
[13]: %sql select avg(payload_mass__kg_) from spacextbl where booster_version like 'F9 v1.1%';

 * sqlite:///my_data1.db
Done.

[13]:
```

avg(payload_mass__kg_)
2534.6666666666665

- This query calculates the average payload mass (in kg) carried specifically by the F9 v1.1 booster version. It helps assess the typical mission capacity of that particular Falcon 9 configuration.

First Successful Ground Landing Date

```
[14]: %sql select min(date) from spacextbl where landing_outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[14]:

min(date)
2015-12-22

- This query finds the earliest date on which SpaceX achieved a successful landing on a ground pad, showcasing a major milestone in their reusability goals. The MIN() function returns the first such occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[15]: %sql select booster_version from spacextbl where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000;

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

[15]:

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

- This query lists the distinct booster versions that successfully landed on drone ships and carried payloads between 4000 kg and 6000 kg. This highlights boosters capable of mid-weight missions with precise drone ship recovery.

Total Number of Successful and Failure Mission Outcomes

[16]:
```
%sql select mission_outcome, count(*) from spacextbl group by mission_outcome;
```

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

[16]:

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

- This query groups all missions by their outcome type and counts how many times each occurred. It provides a quick summary of SpaceX's overall success and failure rates across all launches.

Boosters Carried Maximum Payload

```
[17]: %sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl);

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

[17]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- This query identifies the booster version(s) that carried the heaviest payload ever recorded in the dataset. It uses a subquery to fetch the maximum payload and returns the corresponding booster(s).

2015 Launch Records

```
[18]: %%sql
select substr(date,6,2), landing_outcome, booster_version, launch_site
from spacextbl
where substr(date,1,4) == '2015' and landing_outcome = 'Failure (drone ship)';

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

[18]:

substr(date,6,2)	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

- This query filters for missions in the year 2015 that attempted to land on a drone ship but failed, showing their booster versions and launch sites. It highlights early challenges SpaceX faced in landing recovery.

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

[19]:
```
%%sql
select landing_outcome, count(*) as outcome_count
from spacextbl
where date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by outcome_count desc;
```

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

[19]:

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

- This query ranks the frequency of landing outcomes during the specified date range in descending order, helping identify which types of landings were most common or problematic during SpaceX's early landing development phase.

Section 3

Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites

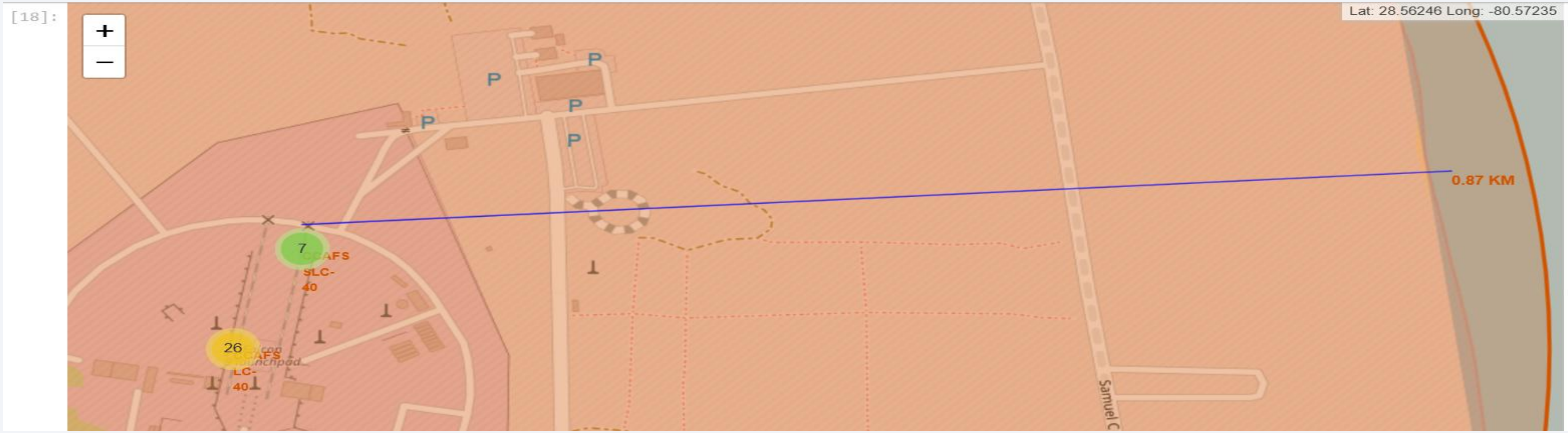
```
display(site_map)
```

- Location Markers: Each marker on the map represents a SpaceX launch site, placed using their real-world geographic coordinates.
- Popup Labels: Clicking on a marker reveals the launch site name, helping identify its location (e.g., CCAFS, VAFB, KSC).
- Geographic Spread: All launch sites are concentrated in the United States, indicating SpaceX's centralized operations.
- Visual Insight: This map provides a spatial understanding of SpaceX’s infrastructure distribution, essential for analyzing logistical and trajectory factors in launches.

Launch Outcomes Visualized by Color on Global Map

- Color-Coded Markers: Each marker’s color represents the launch outcome — for example, green for success and red for failure.
- Enhanced Context: This visual overlay provides immediate feedback on launch performance by location.
- Geographic Trends: You may observe that certain sites have higher success rates, while others show mixed outcomes, giving clues to operational maturity or complexity of missions.

Launch Site Proximity Analysis with Distance Markers

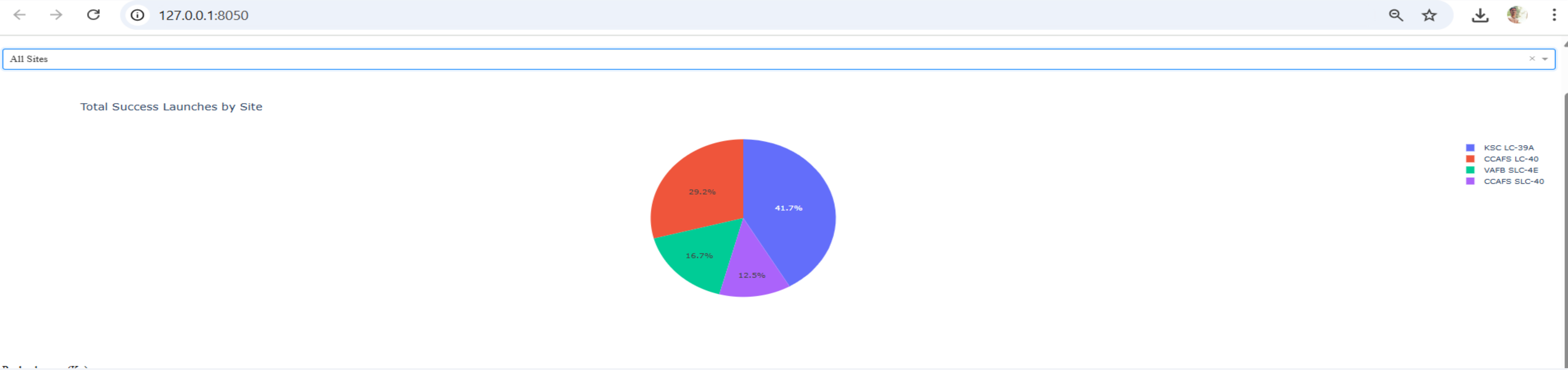


- Selected Launch Site: A highlighted marker indicates the launch site (e.g., CCAFS LC-40).
- Proximity Lines: Lines connect the launch site to nearby coastline, railway, and highway points.
- Distance Labels: Each line includes a label displaying the calculated distance (in KM) between the launch site and the respective feature.
- Insight: This visual helps assess the strategic accessibility of the launch site for logistics, recovery operations, and safety planning.

Section 4

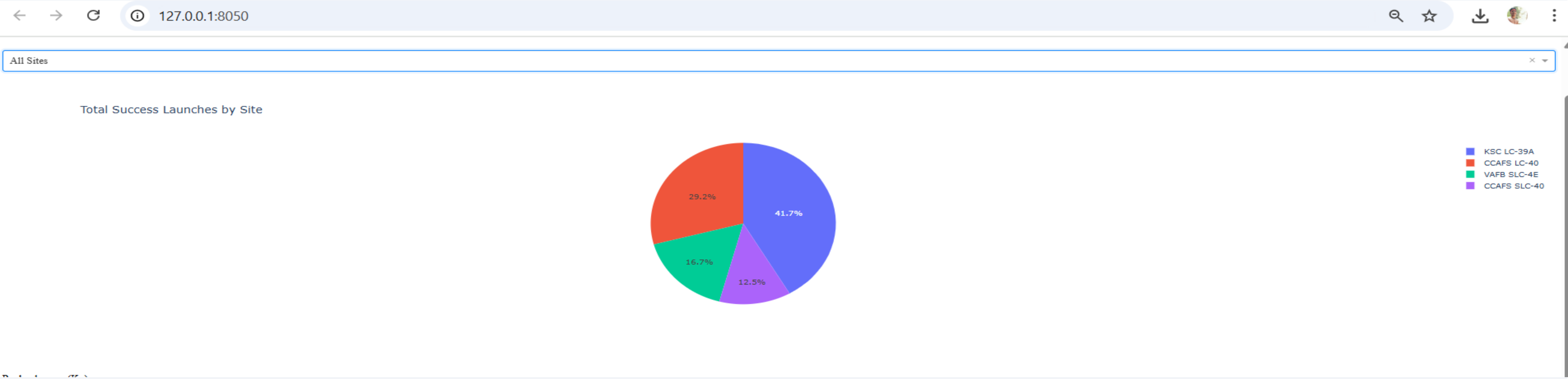
Build a Dashboard with Plotly Dash

Launch Success Count Across All Sites



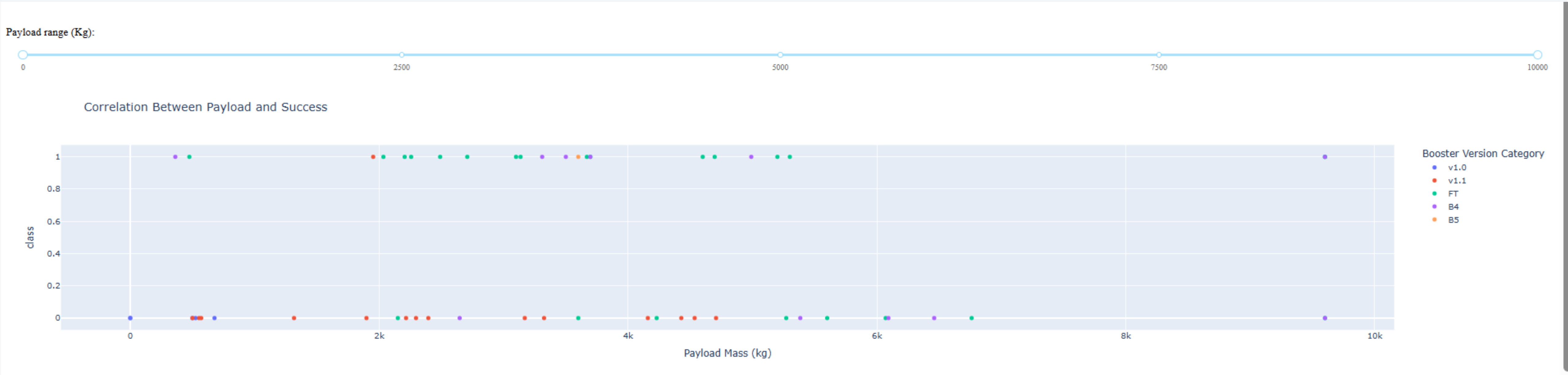
- Pie Chart Segments: Each slice represents a launch site and its proportion of successful launches.
- Comparative View: The chart highlights which sites have contributed most to successful missions, offering a quick glance at operational hotspots.
- Insight: Sites like KSC LC-39A or CCAFS LC-40 may show larger shares, indicating higher usage and reliability.

Success vs. Failure Breakdown for Top Performing Launch Site



- Focused Analysis: This pie chart zooms in on the launch site with the highest success ratio (e.g., KSC LC-39A).Segment
- Breakdown: It visually compares the count of successful vs. failed launches from that specific site.
- Insight: A large green (success) segment indicates strong operational performance, suggesting the site's maturity, infrastructure advantage, or mission-critical role.

Payload Mass vs. Launch Outcome Across All Sites

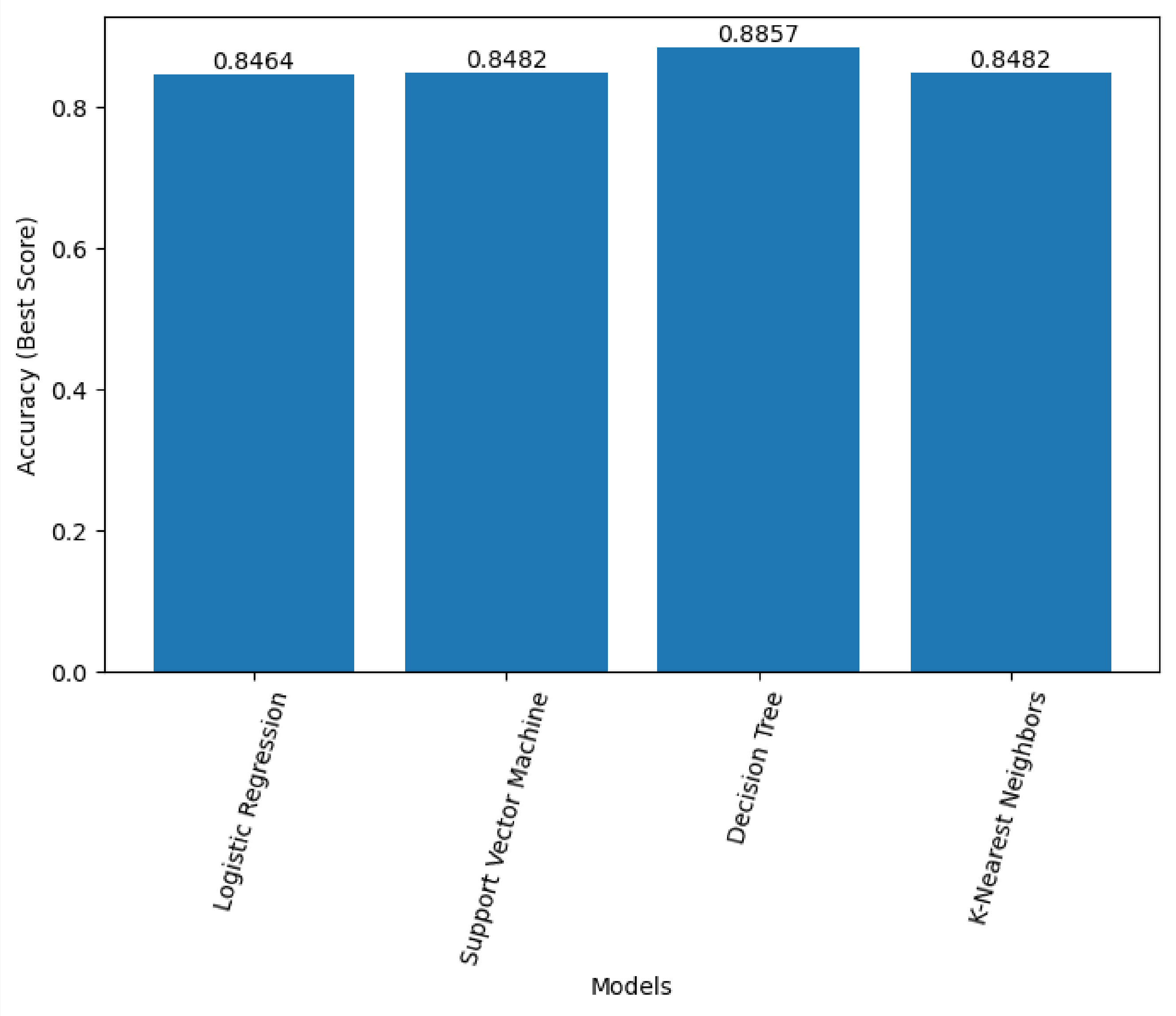


- Scatter Plot Dynamics: Each point represents a launch, plotted by payload mass and launch outcome, color-coded for success or failure.
- Range Slider Interaction: The payload range slider filters launches to explore trends in different payload categories (e.g., light vs. heavy missions).
- Insight: Success rates tend to be higher for payloads under 6000 kg. Boosters like F9 FT and F9 Block 5 show strong success clusters, indicating reliability for medium to heavy payloads.

Section 5

Predictive Analysis (Classification)

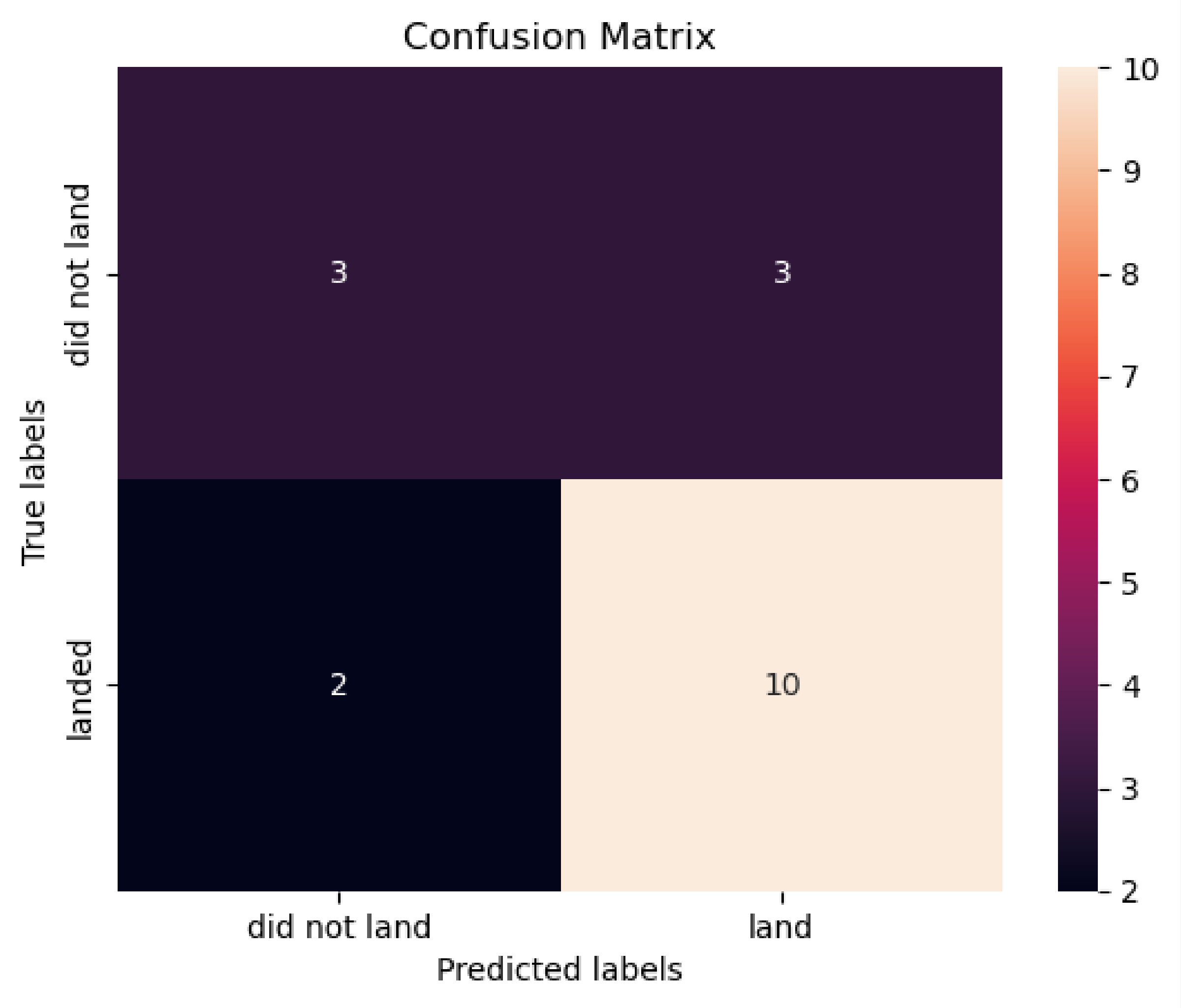
Classification Accuracy



- Bar Chart Representation: Each bar displays the classification accuracy of a model (e.g., Logistic Regression, SVM, Decision Tree, KNN).
- Clear Comparison: The visual makes it easy to compare the performance of different models on the validation/test set.
- Insight: The model with the highest classification accuracy Decision Tree demonstrates the best predictive performance in this SpaceX launch success classification task.

Confusion Matrix

- True Positives (TP): Cases where the model correctly predicted a successful launch.
- True Negatives (TN): Cases where the model correctly predicted a failure.
- False Positives (FP): Cases where the model predicted a success, but it was a failure.
- False Negatives (FN): Cases where the model predicted a failure, but it was a success.
- Insights: The confusion matrix provides a detailed evaluation of the model's performance beyond overall accuracy. The high number of true positives and true negatives in the matrix confirms that the model is reliable in both predicting launch successes and failures. Low false positives/negatives further suggest balanced prediction capability.



Conclusions

- SpaceX launch activity is concentrated at a few key U.S. sites, with varying success rates influenced by site infrastructure and mission complexity.
- Exploratory data analysis revealed patterns between payload mass, orbit type, and launch outcomes, aiding in identifying risk factors for launch failure.
- Interactive visualizations using Folium and Plotly Dash provided clear insights into spatial logistics and operational performance.
- Among the classification models built, the best performing model Decision Tree achieved the highest accuracy, making it suitable for predicting future mission outcomes.
- Data-driven decision-making is feasible for launch reliability analysis, showcasing how machine learning and visualization can support aerospace operations.

Appendix

GitHub link of project : https://github.com/Shivasharanu-SN/Data-Science/tree/main/Applied%20data%20science%20project

Thank you!