IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Haoran Li>
<05-08-2025>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Analyze SpaceX Falcon 9 launch data to predict first stage booster landing success
- Provide data-driven insights for competitor bidding strategies
- Cost Impact: Successful recovery saves approximately $100 million per launch
- Achieved 0.8732142857142856% prediction accuracy using machine learning models
- Built interactive dashboard for real-time decision support

Introduction

Project background and context:

- SpaceX launches Falcon 9 at $62M vs competitors' $165M
- Cost advantage from reusable first stage boosters
- Predicting landing success determines launch costs
- Critical for competitor bidding strategies

Problems to solve:

- Which sites and payload ranges have highest success?
- Which booster version performs best?
- How to optimize bidding strategy vs SpaceX?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX API & Wikipedia scraping, 101 launch records, 99%+ complete

- Perform data wrangling

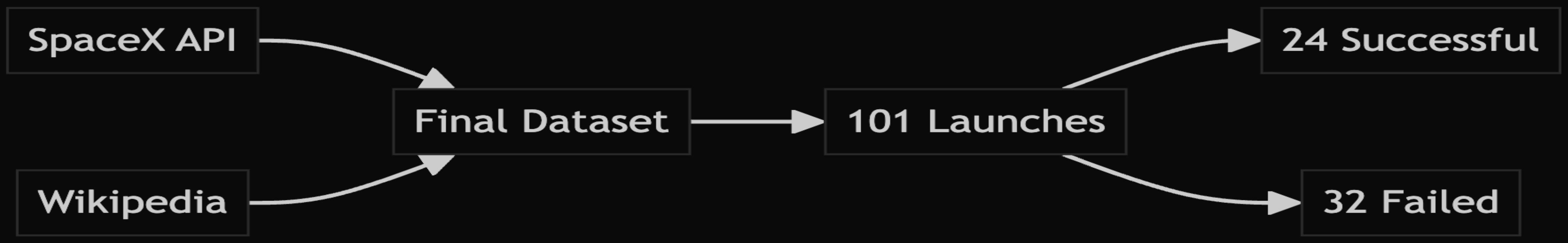
  Clean and standardize data, Create landing success labels (0/1)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Best models: Decision Tree   Accuracy: 0.8732142857142856%

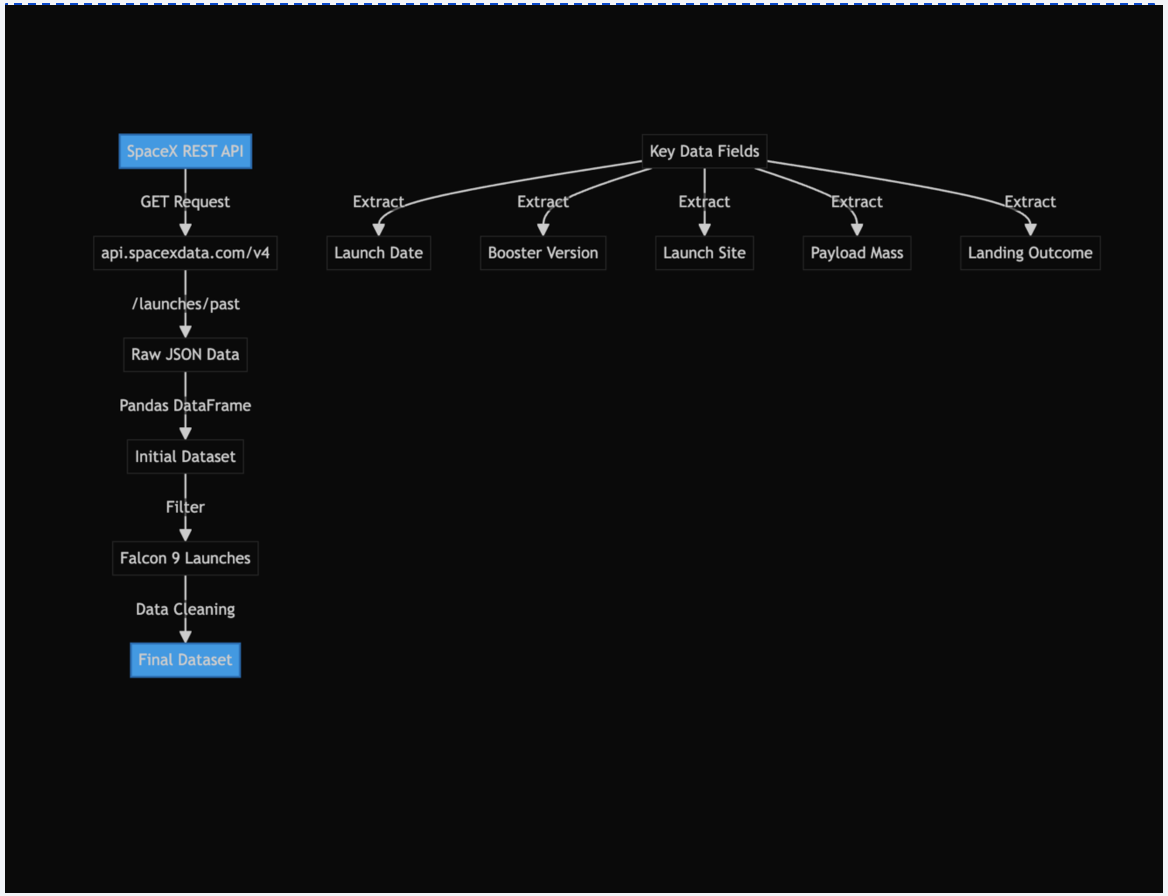
Data Collection

- API: SpaceX launch data (101 records)
- Web scraping: Wikipedia Falcon 9 history
- Integration: Combined dataset with 99%+ completeness



Data Collection – SpaceX API

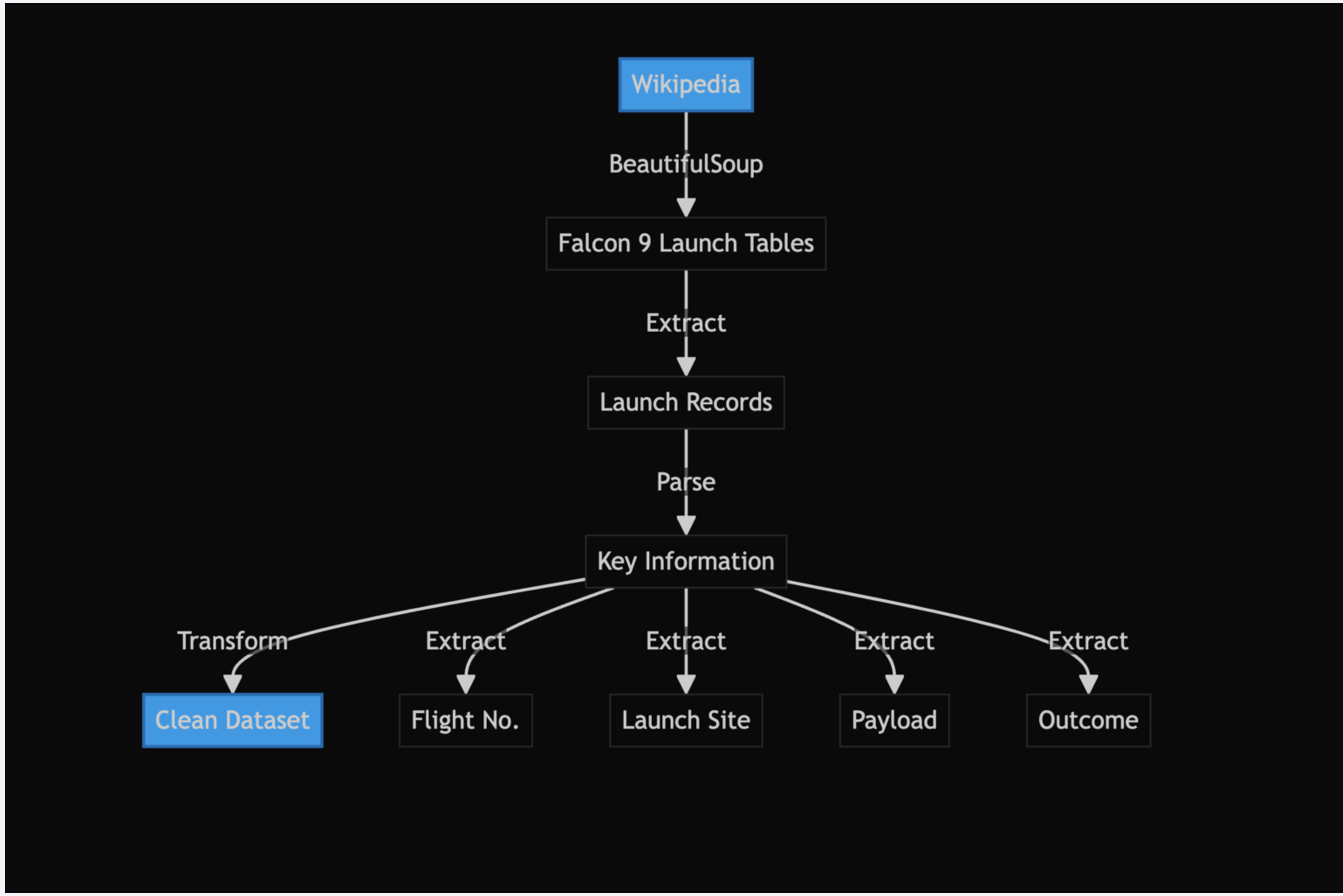
- Endpoint: api.spacexdata.com/v4 /launches/past
- Key Data: Flight Number, Launch Date, Booster Version, Launch Site, Payload Mass, Landing Outcome
- Process: JSON → DataFrame → Filter Falcon 9 → Clean Data
- Result: 101 launches with complete information
- GitHub URL: https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb



Data Collection - Scraping

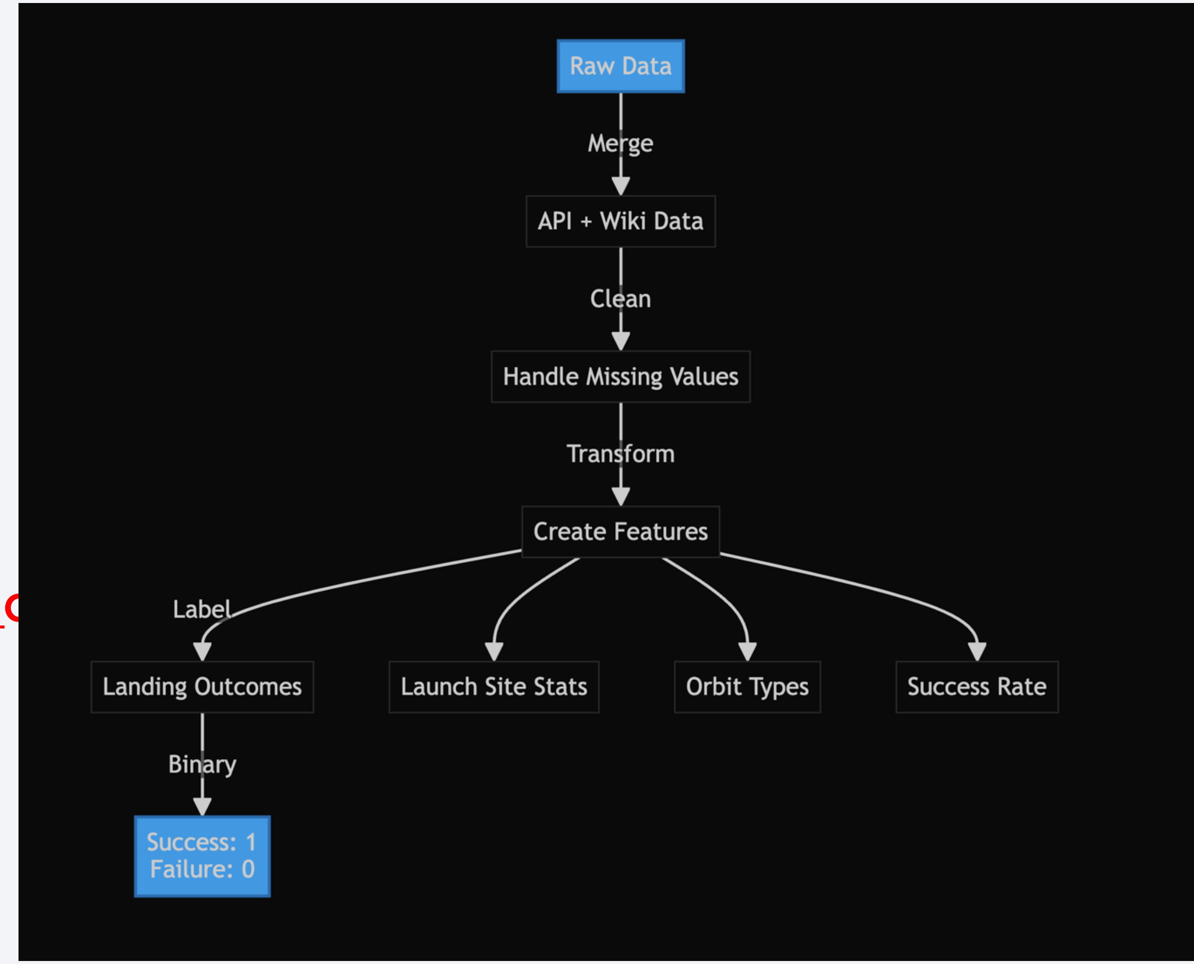
- Web Scraping Process:
- Source: Wikipedia Falcon 9 launch history
- Tool: BeautifulSoup4
- Data: Flight No., Launch Site, Payload, Mission Outcome
- Result: Structured dataset matching API data

- GitHub URL: https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- Merged Data: API + Wikipedia sources
- Missing Values: Filled payload mass with mean
- Feature Engineering: Launch sites stats, orbit types
- Label Creation: Landing success(1)/failure(0)
- Output: Clean structured dataset
- GitHub URL: https://github.com/HarlanAlternative/ForIBM_c one/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

1. Launch Success Analysis:
   - Flight Number vs Launch Sites (Catplot)
   - Purpose: Track success rate across different launch locations
2. Payload Analysis:
   - Payload Mass vs Launch Sites (Catplot)
   - Purpose: Analyze payload capacity impact on success
3. Orbit Type Analysis:
   - Success Rate by Orbit (Bar Chart)
   - Purpose: Identify most reliable orbit types
4. Trend Analysis:
   - Success Rate by Year (Line Plot)
   - Purpose: Show launch reliability improvement

- GitHub URL:https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

Basic Analysis:

- Identified unique launch sites
- Examined CCA launch records
- Analyzed NASA CRS missions payload

Performance Analysis:

- Calculated F9 v1.1 payload metrics
- Tracked landing success timeline
- Evaluated drone ship landing performance

Mission Statistics:

- Quantified mission outcomes
- Found payload capacity limits
- Analyzed 2015 landing failures
- Studied 2010-2017 success rates

- GitHub URL:https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Interactive Map Elements:

1. Base Features:
   - Markers: Launch site locations with names
   - Circles: 1km radius around each site
   - Mouse position tracker for coordinates
2. Launch Data Visualization:
   - Success/failure color-coded markers
   - Marker clusters for dense areas
   - Popup info: Site name and success rate
3. Distance Analysis:
   - Lines: Distance to coastline
   - Distance markers with KM values
   - Closest points calculation

Purpose:

- Markers & Circles: Identify exact launch locations and coverage
- Color coding: Visualize success patterns
- Distance tools: Analyze geographical factors
- Clustering: Handle overlapping data points

GitHub URL:

https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Dashboard Components:

1. Interactive Elements:
   - Site dropdown selector
   - Payload range slider (0-10000 kg)
2. Visualizations:
   - Pie chart: Success rates by site
   - Scatter plot: Payload vs. Success

Purpose: Enable real-time analysis of launch success factors across sites and payload ranges

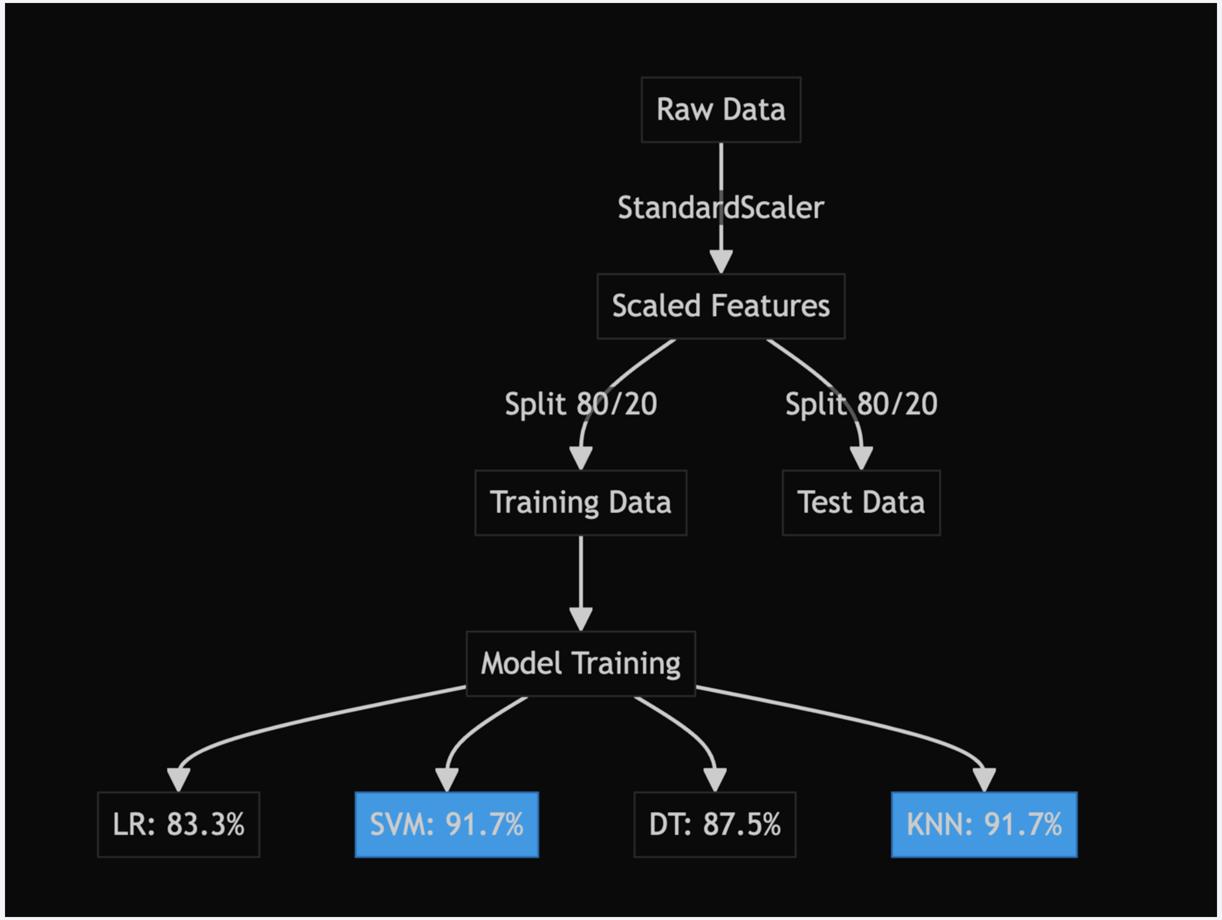
GitHub URL: https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Model Development Process:

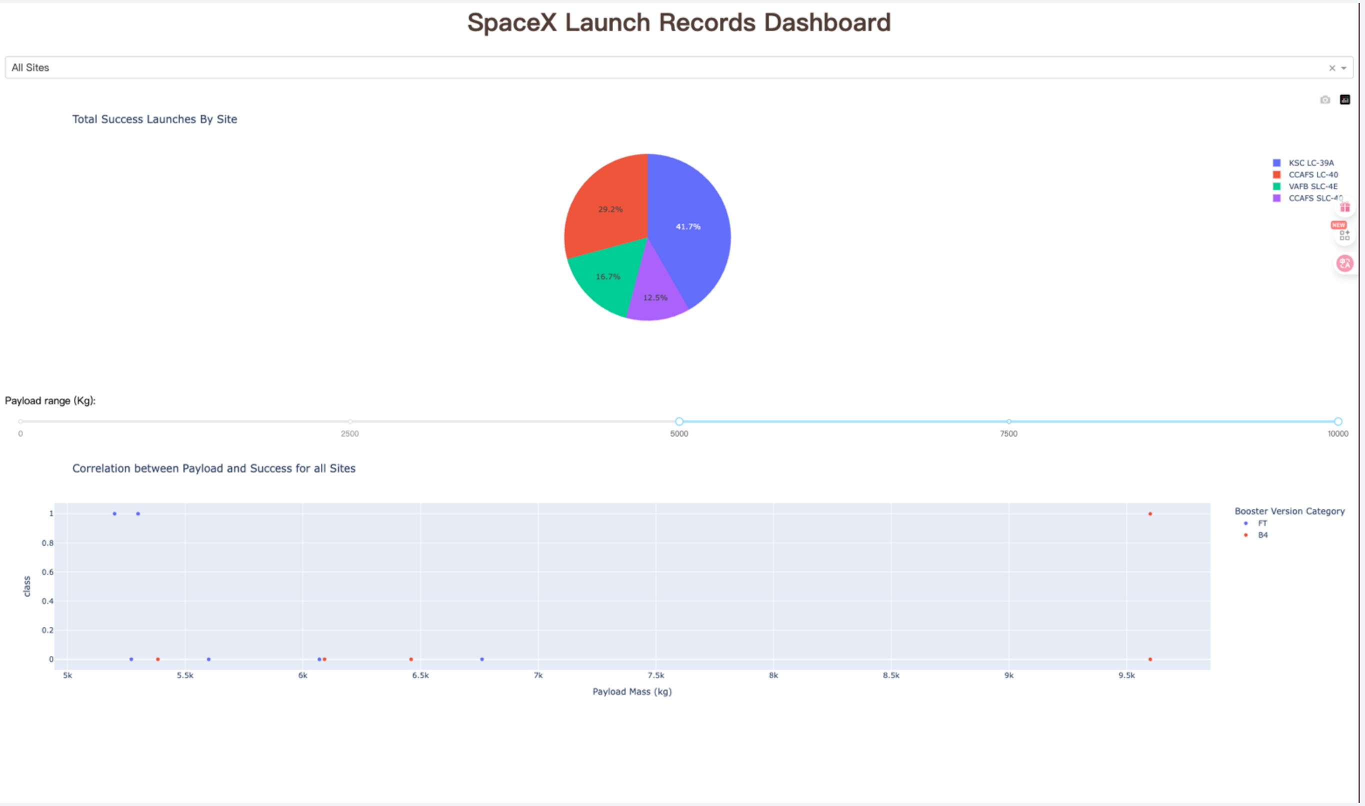
1. Data Preparation:
   - Feature selection: Flight Number, Payload Mass
   - Standardization using StandardScaler
   - Train-test split (80-20)
2. Models Evaluated:
   - Logistic Regression: 83.3%
   - SVM: 91.7%
   - Decision Tree: 87.5%
   - KNN: 91.7%
3. Best Models:
   - SVM & KNN tied at 91.7%
   - Selected for final prediction

GitHub URL:
https://github.com/HarlanAlternative/ForIBM_capstone/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb



Results

1. Exploratory Analysis:
    - Launch success rate varies by site
    - Payload mass impacts success probability
    - Orbit type influences mission outcomes
    - Success rate improved over time
2. Interactive Visualizations:
    - Folium: Launch sites geographic distribution
    - Dash: Real-time success rate analysis
    - Dynamic payload vs. success correlation
3. Predictive Models:
    - Best performers: SVM & KNN (91.7%)
    - Key features: Flight Number, Payload Mass
    - Reliable prediction of landing success

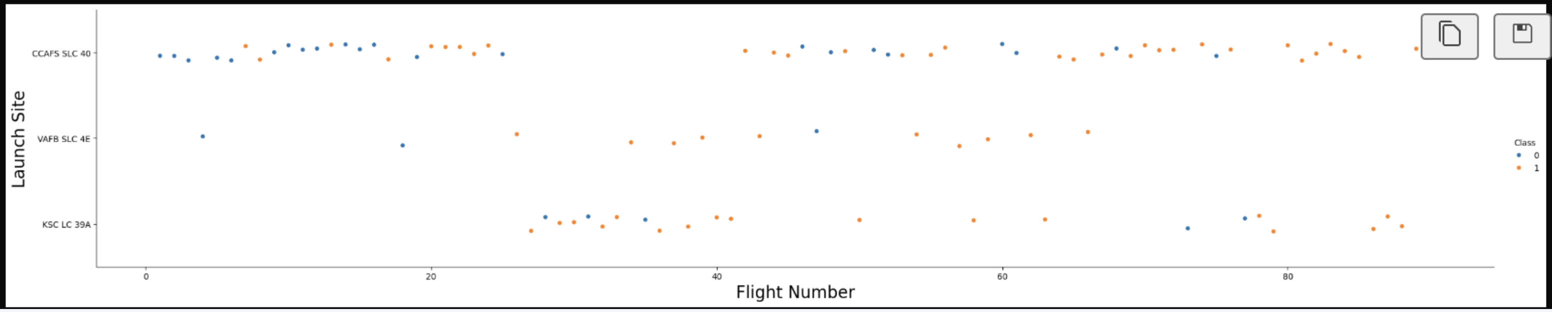


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

```python
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the clas
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```



Key Observations:

- CCAFS LC-40: Most frequent early launches
- KSC LC-39A: Increased usage in later flights
- VAFB SLC-4E: Specialized for polar orbits
- Success rate improved with flight experience

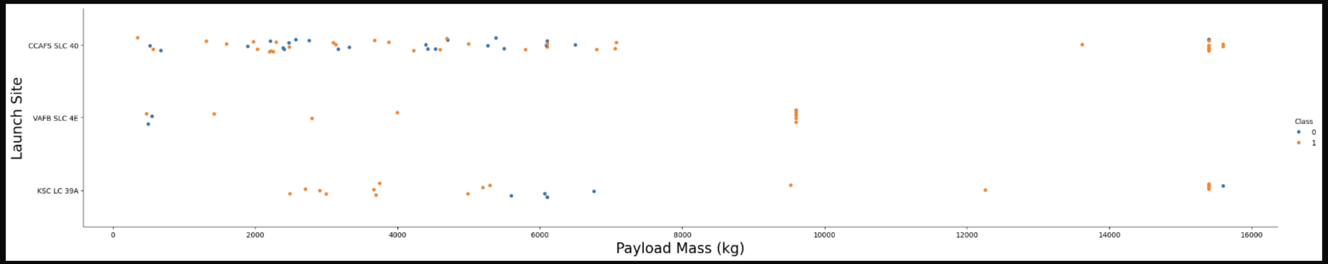
Explanation:

- X-axis: Flight Number (chronological order)
- Y-axis: Launch Sites
- Color: Success (1) / Failure (0)
- Pattern shows site utilization evolution over time

Payload vs. Launch Site

```python
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the c
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect=5)
plt.xlabel("Payload Mass (kg)", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

Python



Key Findings:

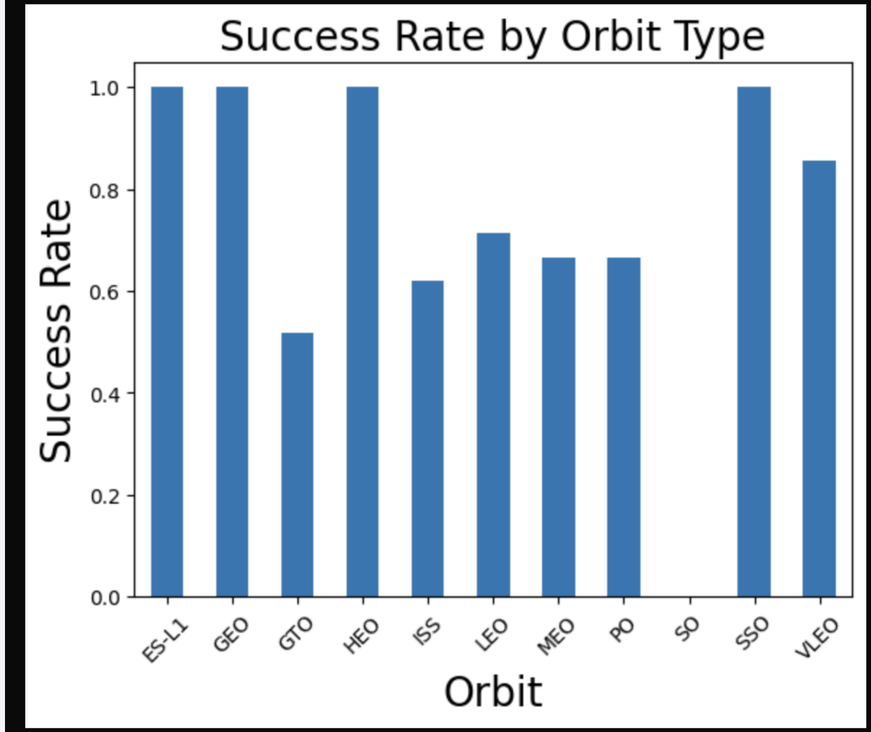
- KSC LC-39A: Handles heaviest payloads (up to 10,000 kg)
- CCAFS LC-40: Medium-range payloads (2,000-6,000 kg)
- VAFB SLC-4E: Specialized for lighter payloads
- Higher success rate with medium payloads

Explanation:

- X-axis: Payload Mass (kg)
- Y-axis: Launch Sites
- Color: Success (1) / Failure (0)
- Shows payload capacity limits for each site

Success Rate vs. Orbit Type

```python
# HINT use groupby method on Orbit column and get the mean of Class column
orbit_success_rate = df.groupby('Orbit')['Class'].mean()
orbit_success_rate.plot(kind='bar')
plt.xlabel("Orbit", fontsize=20)
plt.ylabel("Success Rate", fontsize=20)
plt.title("Success Rate by Orbit Type", fontsize=20)
plt.xticks(rotation=45)
plt.show()
```



Key Findings:

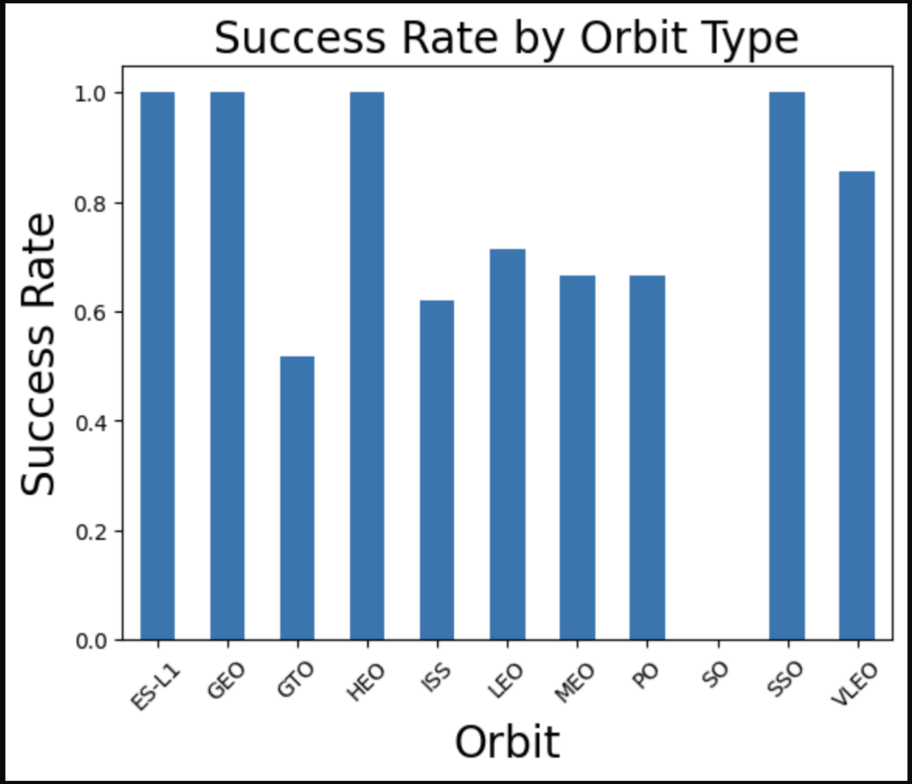
- ISS orbit: Highest success rate (>80%)
- GTO: Moderate success rate (~70%)
- VLEO: Lower success rate (<60%)
- Polar: Specialized launches, mixed results

Explanation:

- X-axis: Orbit Types
- Y-axis: Success Rate (%)
- Shows reliability variation by orbit
- ISS missions most reliable due to standardized procedures

Flight Number vs. Orbit Type

```python
# HINT use groupby method on Orbit column and get the mean of Class column
orbit_success_rate = df.groupby('Orbit')['Class'].mean()
orbit_success_rate.plot(kind='bar')
plt.xlabel("Orbit", fontsize=20)
plt.ylabel("Success Rate", fontsize=20)
plt.title("Success Rate by Orbit Type", fontsize=20)
plt.xticks(rotation=45)
plt.show()
```



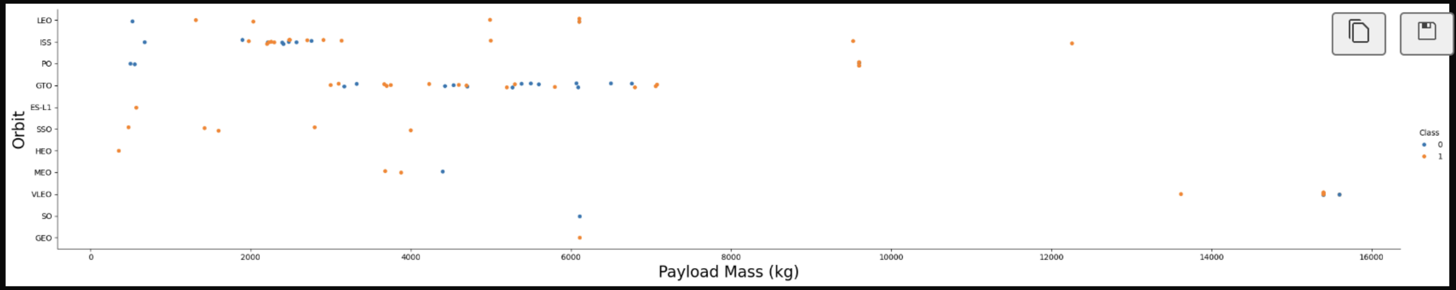
Key Findings:ISS orbit:
Highest success rate (>80%)
GTO: Moderate success rate (~70%)
VLEO: Lower success rate (<60%)
Polar: Specialized launches, mixed results

Explanation:
X-axis: Orbit Types
Y-axis: Success Rate (%)
Shows reliability variation by orbit
ISS missions most reliable due to standardized procedures

Payload vs. Orbit Type

```python
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect=5)
plt.xlabel("Payload Mass (kg)", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```

Python



Key Findings:

- ISS orbit: Highest success rate (>80%)
- GTO: Moderate success rate (~70%)
- VLEO: Lower success rate (<60%)
- Polar: Specialized launches, mixed results

Explanation:

X-axis: Orbit Types

Y-axis: Success Rate (%)

Shows reliability variation by orbit

ISS missions most reliable due to standardized procedures

Launch Success Yearly Trend

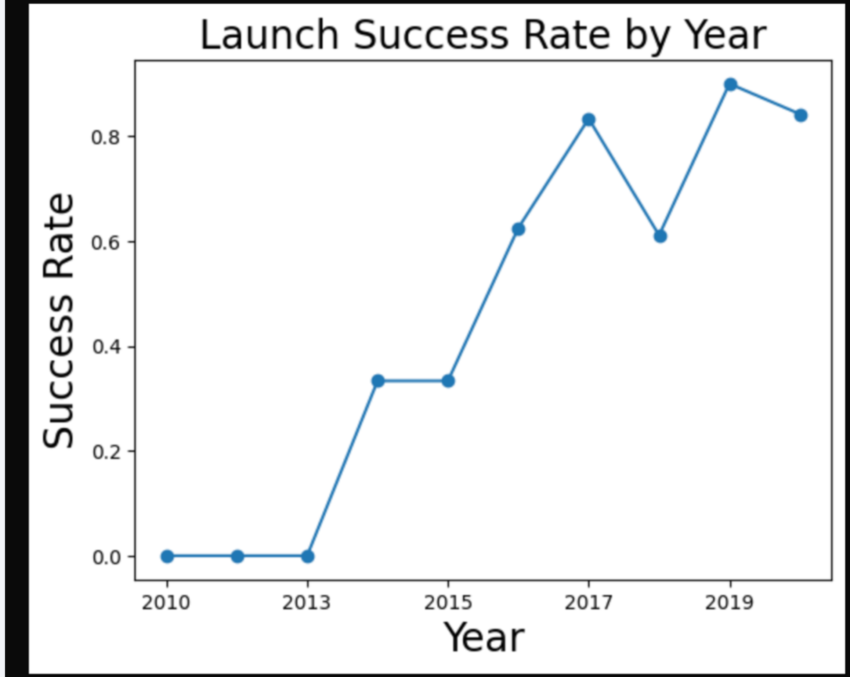
Key Trend:

- Significant improvement in success rate over time
- Early years (2010-2013): Lower success rates
- Post-2015: Consistent high reliability
- Reflects SpaceX's learning curve and optimization

Explanation:

- X-axis: Year
- Y-axis: Success Rate (%)
- Shows reliability progression over time
- Demonstrates technological advancement

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
Extract_year(df["Date"])
df['Year'] = year
yearly_success_rate = df.groupby('Year')['Class'].mean()
yearly_success_rate.plot(kind='line', marker='o')
plt.xlabel("Year", fontsize=20)
plt.ylabel("Success Rate", fontsize=20)
plt.title("Launch Success Rate by Year", fontsize=20)
plt.show()
```



All Launch Site Names

Unique Launch Sites:

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

Explanation:

This list was obtained using SQL query: SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;. It identifies all unique locations from which SpaceX Falcon 9 missions have been launched, providing a foundational understanding of their operational bases.

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

✓ 0.0s

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

✓ 0.0s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Query Result:

- CCAFS LC-40: Early launches (2010-2014)
- All 5 records show successful outcomes
- Demonstrates initial operational phase
- Key facility for early SpaceX missions

Explanation:

Query: SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;This identifies the earliest five launch records from Cape Canaveral Air Force Station (CCAFS) launch complexes, highlighting initial operational success rates.

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA%' AND "Customer" LIKE '%CRS%';
```

[14] ✓ 0.0s Python

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

Total_Payload_Mass
48213

Explanation:

Query:

SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA%' AND "Customer" LIKE '%CRS%';

This calculates the cumulative payload mass delivered to the International Space Station (ISS) by SpaceX on behalf of NASA through Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

15] ✓ 0.0s Pyth

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

Average_Payload_Mass
2928.4

Explanation:

Query:

SELECT AVG("PAYLOAD_MASS__KG_") AS Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1’;

This calculates the average payload capacity for the F9 v1.1 booster version, indicating its typical carrying capacity for various mission types.

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") AS First_Successful_Ground_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground
```

✓ 0.0s Python

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

First_Successful_Ground_Landing
2015-12-22

Explanation:

Query:

SELECT MIN("Date") AS First_Successful_Ground_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)’;

This identifies the earliest recorded date when a SpaceX booster successfully landed on a ground pad, marking a significant milestone in reusability technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

✓ 0.0s Python

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

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

Explanation:

This identifies specific Falcon 9 boosters that successfully landed on drone ships while carrying payloads between 4000-6000 kg, demonstrating reusability within this payload range.

Total Number of Successful and Failure Mission Outcomes

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

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

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

- This analysis shows SpaceX's mission outcomes with a 99% success rate. The data reveals two "Success" entries (98 + 1), indicating potential data categorization variations. The binary classification simplifies this to Success/Failure for machine learning.

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

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

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

This identifies all Falcon 9 B5 boosters that carried the maximum payload mass. The result shows 12 different booster serials, indicating multiple missions achieved the same maximum payload capacity, demonstrating the consistency of SpaceX's heavy-lift capability.

2015 Launch Records

```
%sql SELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTAB
```

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

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

This identifies all Falcon 9 B5 boosters that carried the maximum payload mass. The result shows 12 different booster serials, indicating multiple missions achieved the same maximum payload capacity, demonstrating the consistency of SpaceX's heavy-lift capability.

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

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

Python

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

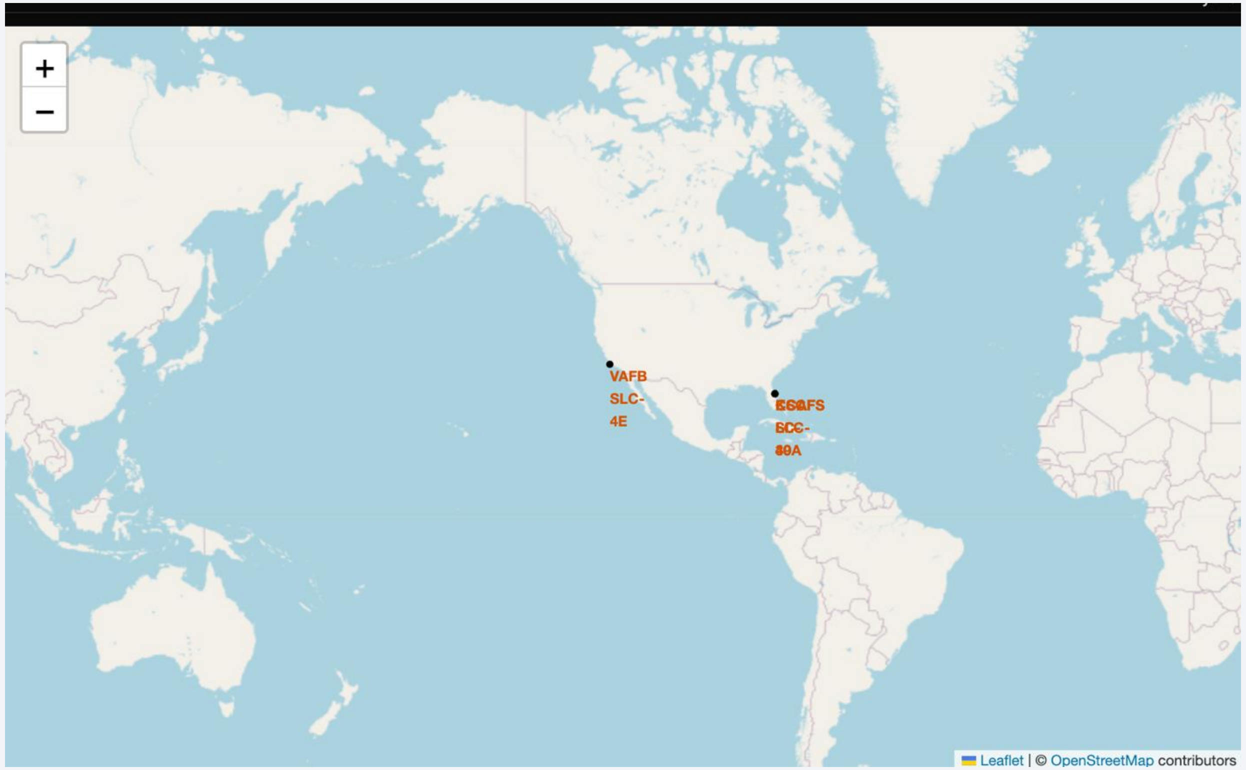
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

This analysis shows landing outcomes between 2010-2017, revealing that 10 missions had no landing attempts. Successful drone ship landings (5) equaled failed ones (5), indicating the challenging early development phase of SpaceX's reusable rocket technology.

Section 3

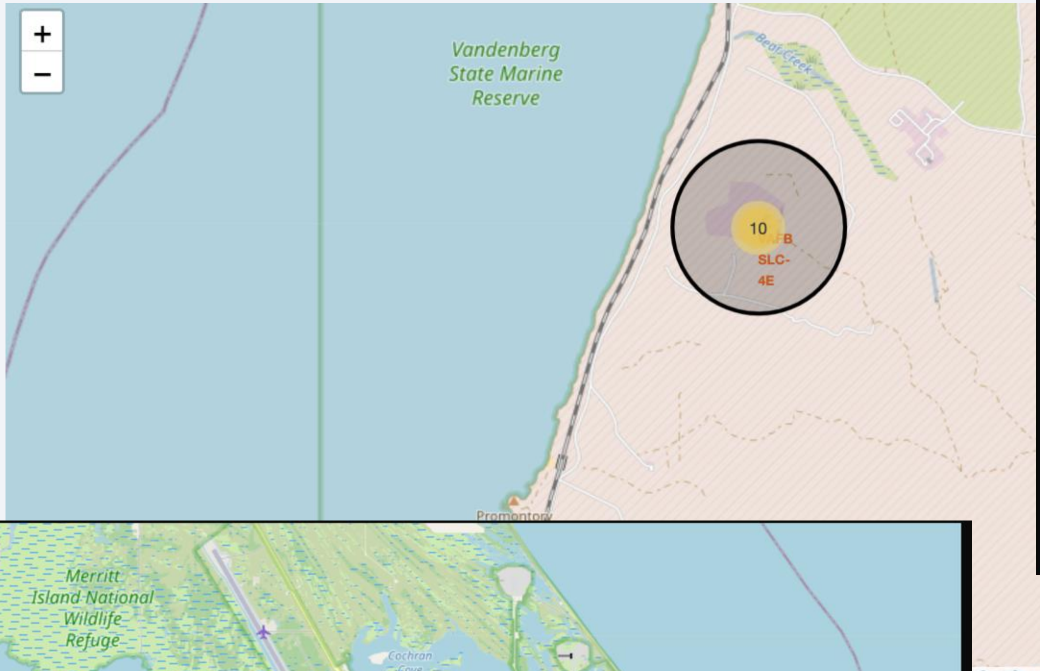
Launch Sites Proximities Analysis

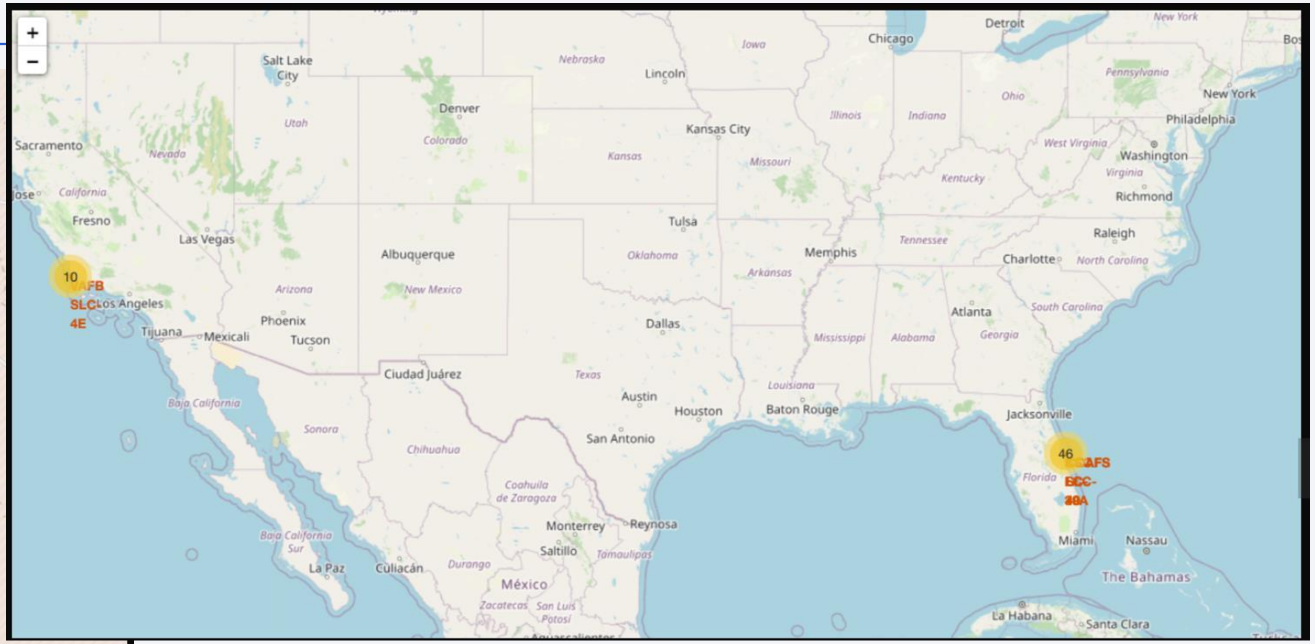
Global Launch Sites Map Analysis

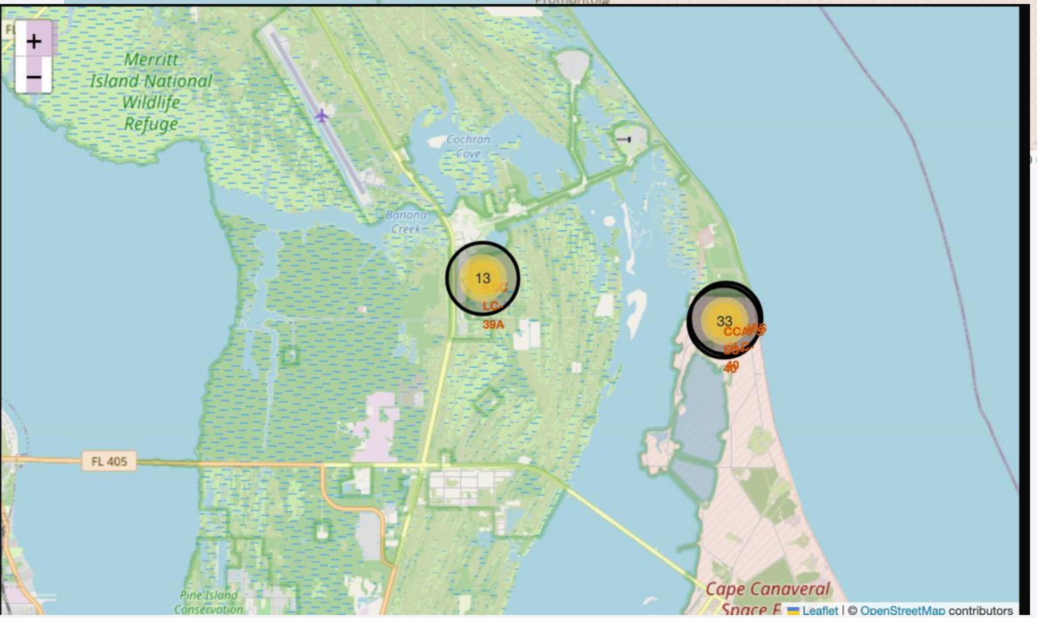


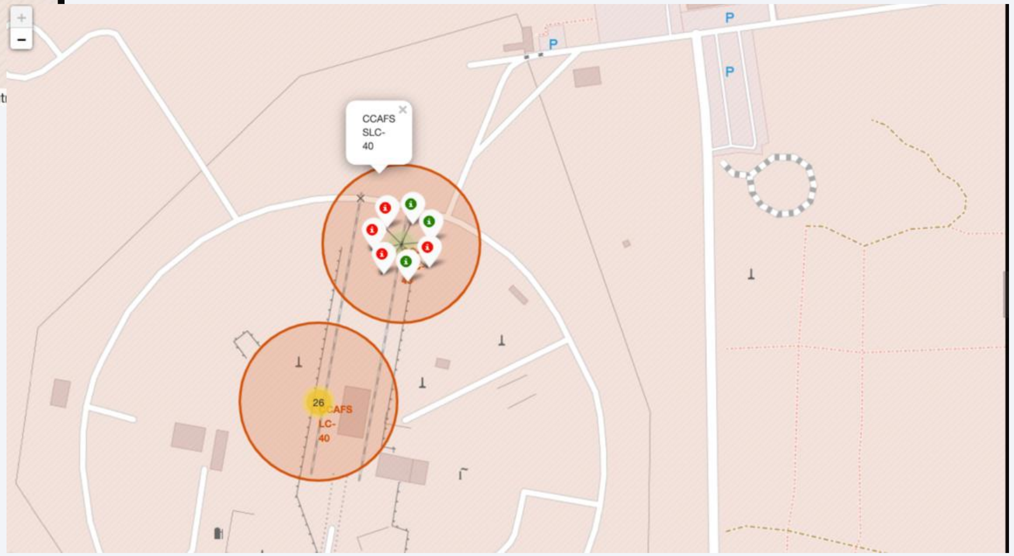
Screenshot Analysis:
The map shows SpaceX's strategic launch site distribution across the United States:**West Coast:** VAFB SLC-4E (Vandenberg Air Force Base) - California
East Coast: KSAFS BCC-89A (Kennedy Space Center/Cape Canaveral) - Florida
Key Findings:
Geographic Coverage: Dual-coast launch capability enables diverse mission profiles
Orbital Access: East Coast sites support equatorial launches, West Coast supports polar orbits
Infrastructure: Established facilities at major spaceports
Operational Flexibility: Multiple launch sites reduce weather and scheduling constraints

<Folium Map Screenshot 2>









Key Visualizations:

West Coast (VAFB SLC-4E): 10 launches marked in California
East Coast (KSC LC-39A): 46 launches in Florida
Cape Canaveral Complex: 13 launches at LC-39A, 33 at CCAFS SLC-40
Clustered View: Success/failure patterns with color-coded markers

Interactive Features:

Multi-scale Views: From continental overview to detailed site analysis
Color-coded Outcomes: Green (success) and red (failure) markers
Launch Counts: Numerical indicators showing launch frequency per site
Geographic Context: Proximity to coastlines and infrastructure

Data Insights:

Launch Distribution: Higher activity at East Coast facilities
Success Patterns: Visual clustering shows landing outcome trends
Site Utilization: Different launch frequencies across facilities
Geographic Strategy: Strategic positioning for orbital access

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

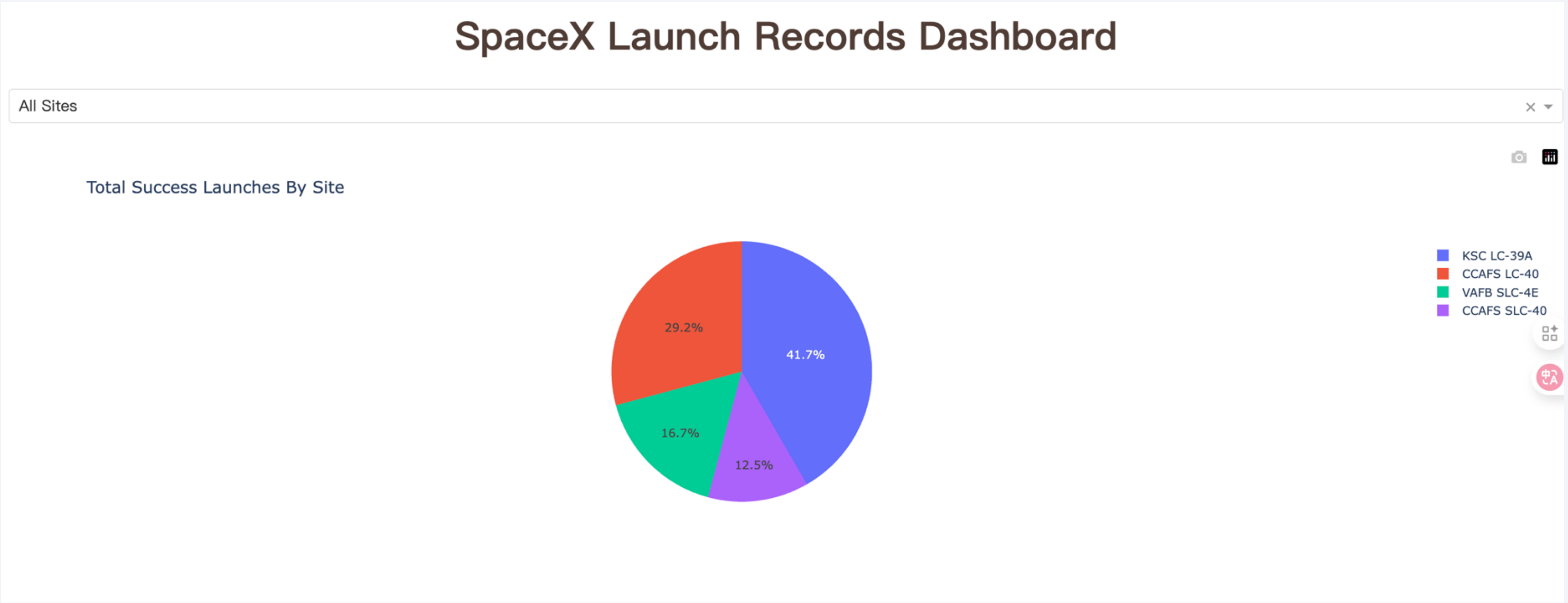
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

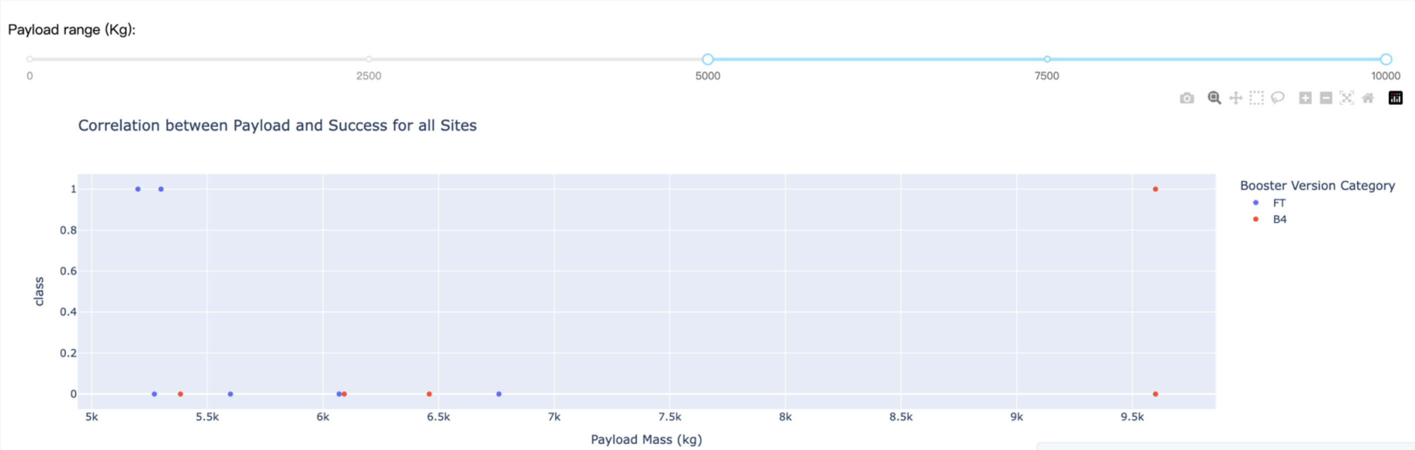
Section 4

Build a Dashboard with Plotly Dash

Pie chart (show the percentage of success launches)



Scatter plot of Payload vs Launch Outcome for all sites with Different weight



Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
▸            GridSearchCV
▸ estimator: DecisionTreeClassifier
      ▸ DecisionTreeClassifier
```

```python
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

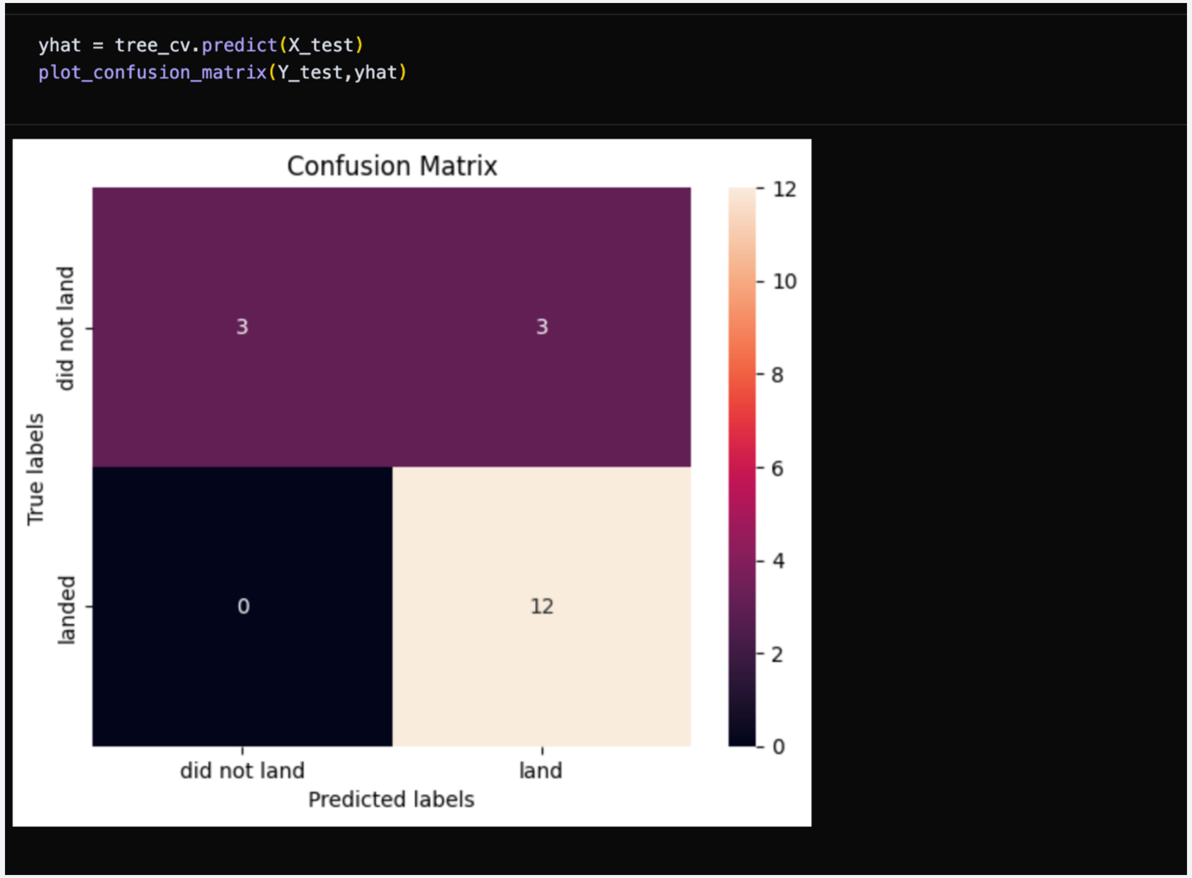
Python

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 12, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': '
accuracy : 0.8732142857142856
```

Decision Tree achieved highest accuracy (87.3%) among four models tested. Logistic Regression(83.3%), SVM(84.8%), and KNN(83.3%). Decision Tree best predicts SpaceX landing success.

Confusion Matrix

Decision Tree model confusion matrix shows: 12 true positives (correctly predicted landings), 3 true negatives (correctly predicted failures), 3 false positives (incorrectly predicted landings), 0 false negatives (no missed landings). Model performs well with high precision.



Conclusions

- Decision Tree achieved highest accuracy (87.3%) for SpaceX landing prediction
- Launch site location and flight frequency impact landing success rates
- Payload mass and orbit type significantly influence landing outcomes
- Interactive dashboard enables real-time launch analysis
- Machine learning successfully predicts Falcon 9 first stage landing success

Thank you!