IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sam
2025-10-02

Outline

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

Executive Summary

Data & Methods:

- Collected data via SpaceX API and web scraping
- Created success/fail classification variable
- Explored payload, launch site, flight trends with visualization
- Analyzed launch metrics (payload statistics, success counts) using SQL
- Examined geographic impact on launch outcomes
- Built predictive models: Logistic Regression, SVM, Decision Tree, KNN

Key Insights:

- Launch success has improved over time
- KSC LC-39A launch site leads in success rate
- Orbits ES-L1, GEO, HEO, & SSO have 100% success
- Launch sites cluster near equator and coastlines
- Decision Tree model slightly outperforms others

Introduction

SpaceX leads the space industry by making space travel more affordable and accessible.

Accomplishments:

- Delivering cargo and crew to the ISS
- Launching satellite internet constellation
- Conducting manned space missions
- Key innovation: Reusing the Falcon 9 rocket's first stage, reducing launch costs to $62M vs. competitors' $165M+
- Predicting if first stage will land successfully helps estimate launch cost
- Approach: Use public data and machine learning to predict first-stage landing success

Exploration Focus

- Impact of payload mass, launch site, flight count, and orbits on landing success
- Trends in landing success rates over time
- Identify best predictive model for landing outcome (binary classification)

Section 1

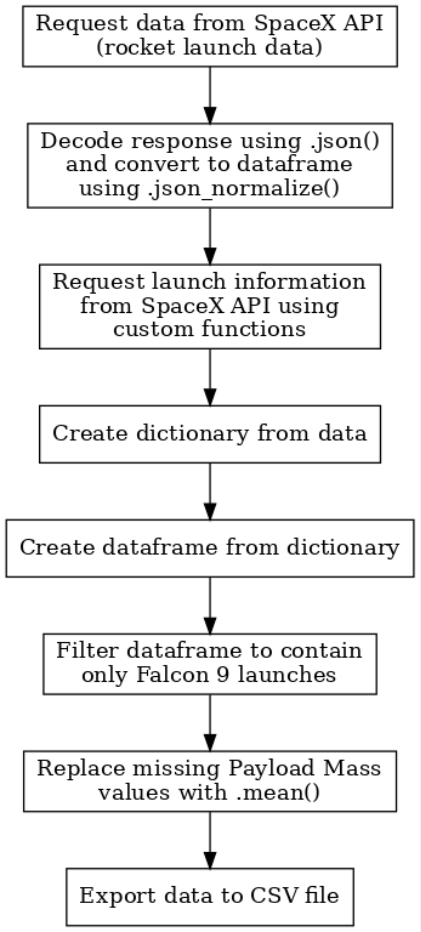
Methodology

Methodology

- Collected data using SpaceX REST API and web scraping
- Prepared data by filtering, handling missing values, and applying one-hot encoding
- Explored data through exploratory data analysis (EDA) with SQL and visualization
- Created interactive visualizations using Folium and Plotly Dash
- Built and tuned classification models to predict rocket landing outcomes

Data Collection API

- Request rocket launch data from SpaceX REST API
- Decode JSON response and convert to dataframe with .json_normalize()
- Use custom functions to retrieve launch information
- Convert data to dictionary, then create dataframe
- Filter for Falcon 9 launches only
- Handle missing Payload Mass values by replacing with mean
- Export cleaned data to CSV file



Data Collection Web Scraping

- Request raw HTML from Wikipedia Falcon 9 launch list page
- Parse HTML with BeautifulSoup to create soup object
- Extract column headers from table header tags
- Parse launch data from HTML tables into a structured dictionary
- Convert dictionary to pandas DataFrame
- Export cleaned DataFrame to CSV file for analysis

Data Wrangling

- Extracted Falcon 9 launch data from Wikipedia table using BeautifulSoup
- Parsed HTML to identify column headers and collect data from tables
- Converted data into a structured dictionary and then to pandas DataFrame
- Filtered for Falcon 9 launches only
- Replaced missing Payload Mass values with the column mean
- Exported the cleaned dataset to CSV for further analysis

EDA with Data Visualization

Visualized relationships using scatter plots:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Payload Mass vs. Orbit Type
- Used scatter plots to identify potential predictive variables
- Employed bar charts to compare discrete categories and their measured values
- These visual analyses guide feature selection for machine learning models

EDA with SQL

- Data Queries & Analysis Overview
- Displayed unique launch sites
- Extracted 5 records with launch site names beginning with "CCA"
- Total payload mass for NASA (CRS) booster launches calculated
- Listed:
- Date of first successful ground pad landing
- Boosters with drone ship success & payload mass between 4,000–6,000 kg
- Total count of successful vs. failed missions
- Failed drone ship landings in 2015 with booster versions and launch sites
- Landing outcomes count between June 2010 and March 2017 (descending)

Build an Interactive Map with Folium

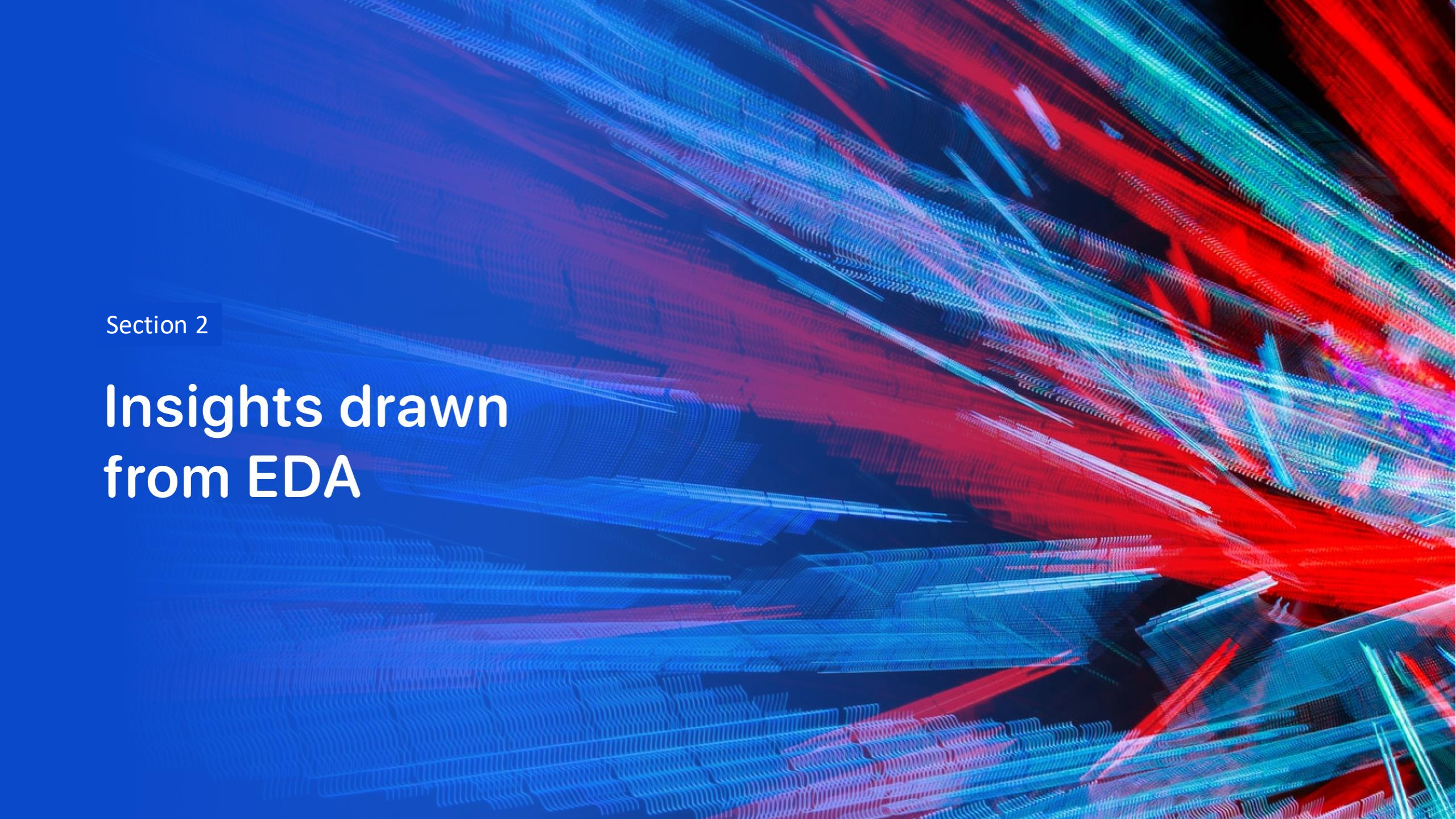
- Added blue circle marker at NASA Johnson Space Center (Lat: 29.5622, Lon: -95.0908) with popup label
- Placed red circle markers at all launch sites with latitude/longitude coordinates and popup labels
- Colored markers show launch outcomes: green for successful, red for unsuccessful launches
- Included colored lines depicting distances from launch site CCAFS SLC-40 to nearby coastline, railway, highway, and city

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites
- • Allow user to select all launch sites or a certain launch site
- Pie Chart Showing Successful Launches
- • Allow user to see successful and unsuccessful launches as a percent of the total
- Slider of Payload Mass Range
- • Allow user to select payload mass range
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
- • Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score and Accuracy


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

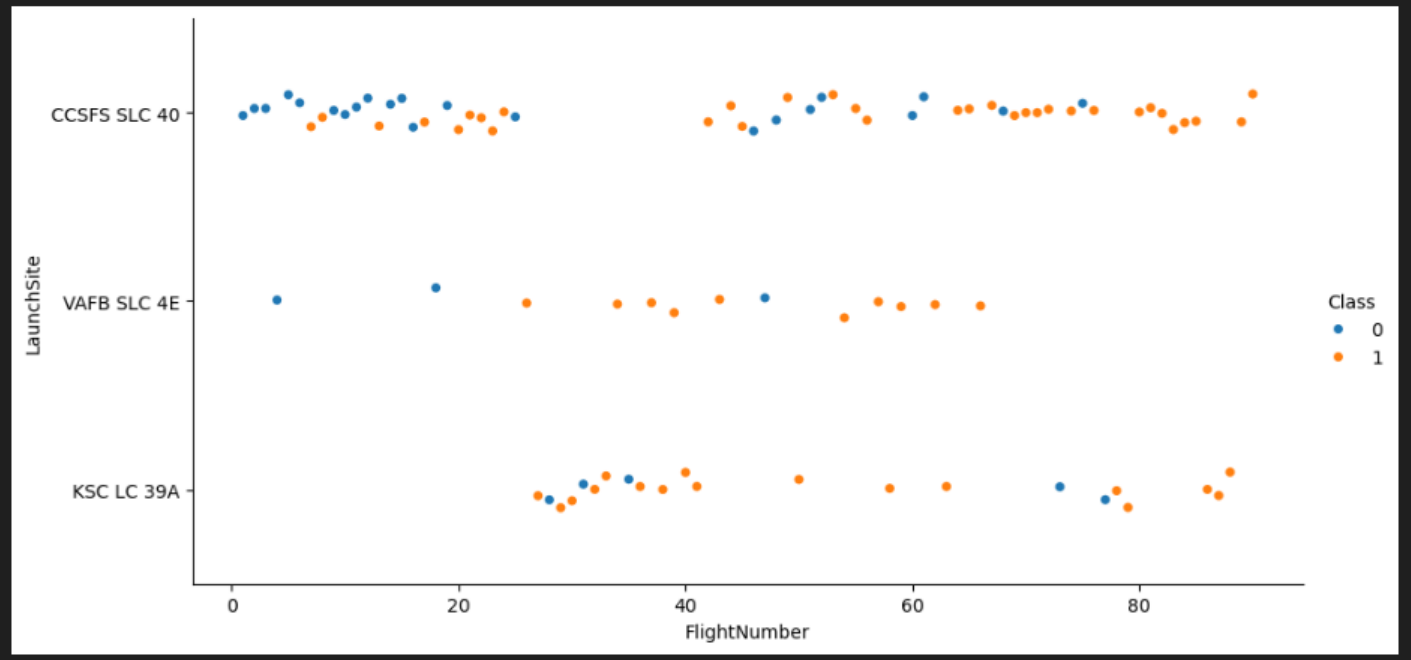
Analysis of Flight Number and Launch Sites on Launch Outcome

- The majority of launches occur from CCSFS SLC-40, showing both early unsuccessful attempts and later consistent improvements in landing success.
- KSC LC-39A launches appear in later flight numbers and generally demonstrate higher landing success rates, reflecting its use in more recent missions.
- VAFB SLC-4E has fewer launches overall but shows a mix of outcomes, with successes becoming more common as flight experience accumulates.

Overall, the analysis suggests that while launch site influences mission allocation, success rates improve over time across all sites, emphasizing operational maturity rather than location-specific performance differences.

```
[71]  1  sns.catplot(data=df_api, x='FlightNumber', y='LaunchSite', hue='Class', aspect=2)
      ✓ 0.1s                                                                    Python
```

```
<seaborn.axisgrid.FacetGrid at 0x7f9c89874ec0>
```



Payload vs. Launch Site

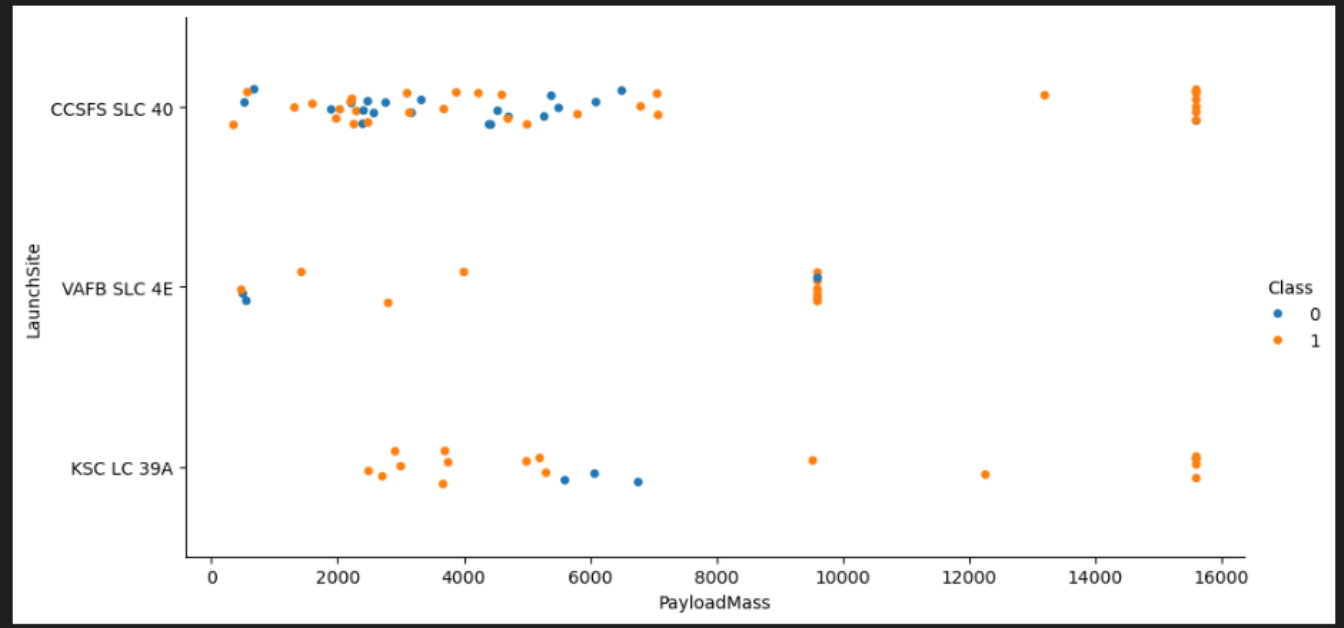
Analysis of Payload Mass and Launch Sites on Launch Outcome

- The relationship between **PayloadMass** and **LaunchSite** with mission outcome (Class) is visualized using a categorical scatterplot.
- **CCSFS SLC-40** shows the largest number of launches across a wide range of payload masses, with successful outcomes (orange) dominating at both moderate and very high payload levels.
- **KSC LC-39A** handles several high-payload missions (approaching 16,000 kg) with mostly successful outcomes, highlighting its role in heavy-lift launches.
- **VAFB SLC-4E** shows fewer launches and mostly moderate payloads, with mixed success, suggesting more limited but specialized use.
- Overall, higher payload missions tend to be concentrated at CCSFS SLC-40 and KSC LC-39A, with improved success rates over time, while VAFB SLC-4E plays a smaller role in heavy-lift missions.

```
1  sns.catplot(data=df_api, x='PayloadMass', y='LaunchSite', hue='Class', aspect=2)
```

✓ 0.1s Python

```
<seaborn.axisgrid.FacetGrid at 0x7f9c8957ff80>
```



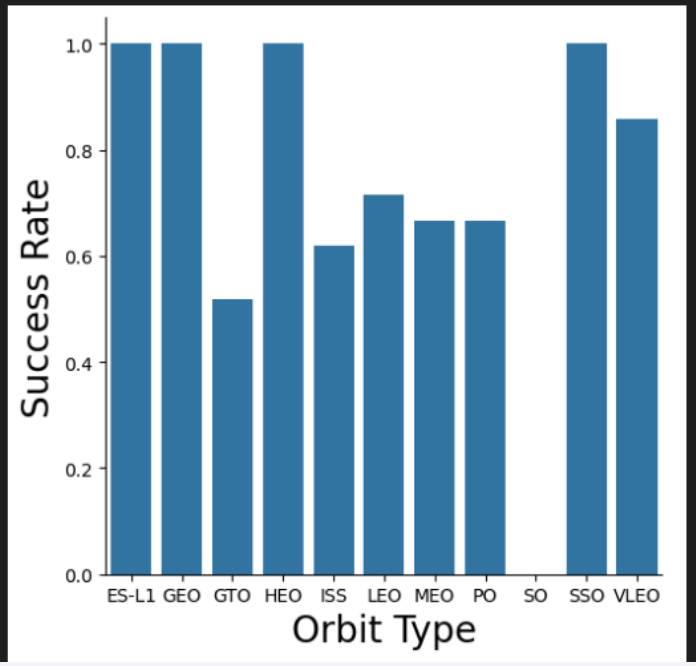
Success Rate vs. Orbit Type

Analysis of Orbit Type on Launch Success Rate

- The plot shows the relationship between orbital destination (Orbit Type) and the probability of mission success (Success Rate).
- Missions targeting **ES-L1, GEO, HEO, and SSO** achieve a **100% success rate**, indicating consistent reliability for these orbits.
- **GTO (Geostationary Transfer Orbit)** has the lowest success rate, reflecting the higher energy requirements and complexity of achieving this orbit.
- Medium-demand orbits such as **LEO, MEO, PO, and ISS** show moderate success rates (60–70%), suggesting that while common, these missions still carry operational challenges.
- Overall, the analysis highlights that **orbit type influences mission success**, with highly demanding transfer orbits like GTO posing the greatest challenge, whereas specialized or stable orbits exhibit stronger reliability.

```
# HINT use groupby method on Orbit column and get the mean of Class column
sns.catplot(x= 'Orbit', y = 'Class', data = df_api.groupby('Orbit')['Class'].mean().reset_index(), kind = 'bar')
plt.xlabel('Orbit Type',fontsize=20)
plt.ylabel('Success Rate',fontsize=20)
plt.show()
```

74] ✓ 0.1s Python

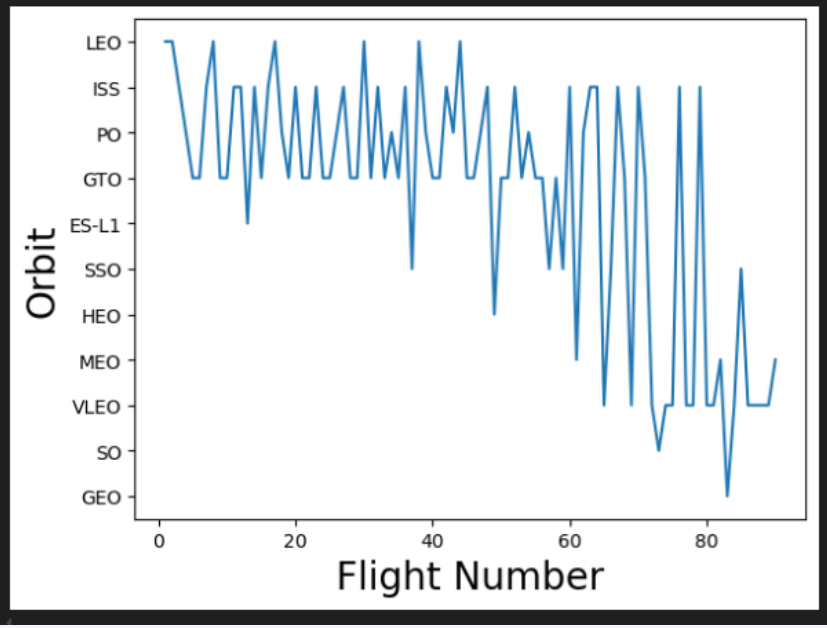


Flight Number vs. Orbit Type

- The plot illustrates the evolution of orbital destinations (Orbit) over sequential launches (FlightNumber).
- Early missions are dominated by **LEO** and **ISS** launches, reflecting SpaceX's focus on low Earth orbit and cargo delivery missions during initial operational phases.
- As flight numbers increase, a wider variety of orbits (e.g., **GTO, SSO, VLEO, GEO, ES-L1**) appear, highlighting the company's diversification into more complex and higher-energy missions.
- The increasing spread of orbit types in later launches indicates SpaceX's transition from primarily resupply missions to broader commercial and scientific payload deployment.
- This trend underscores the **expansion of operational capability** over time, with later missions targeting more challenging orbits that demonstrate technical maturity and mission versatility.

```
1  sns.lineplot(y="Orbit", x="FlightNumber",data=df_api)
2  plt.xlabel("Flight Number",fontsize=20)
3  plt.ylabel("Orbit",fontsize=20)
4  plt.show()
```

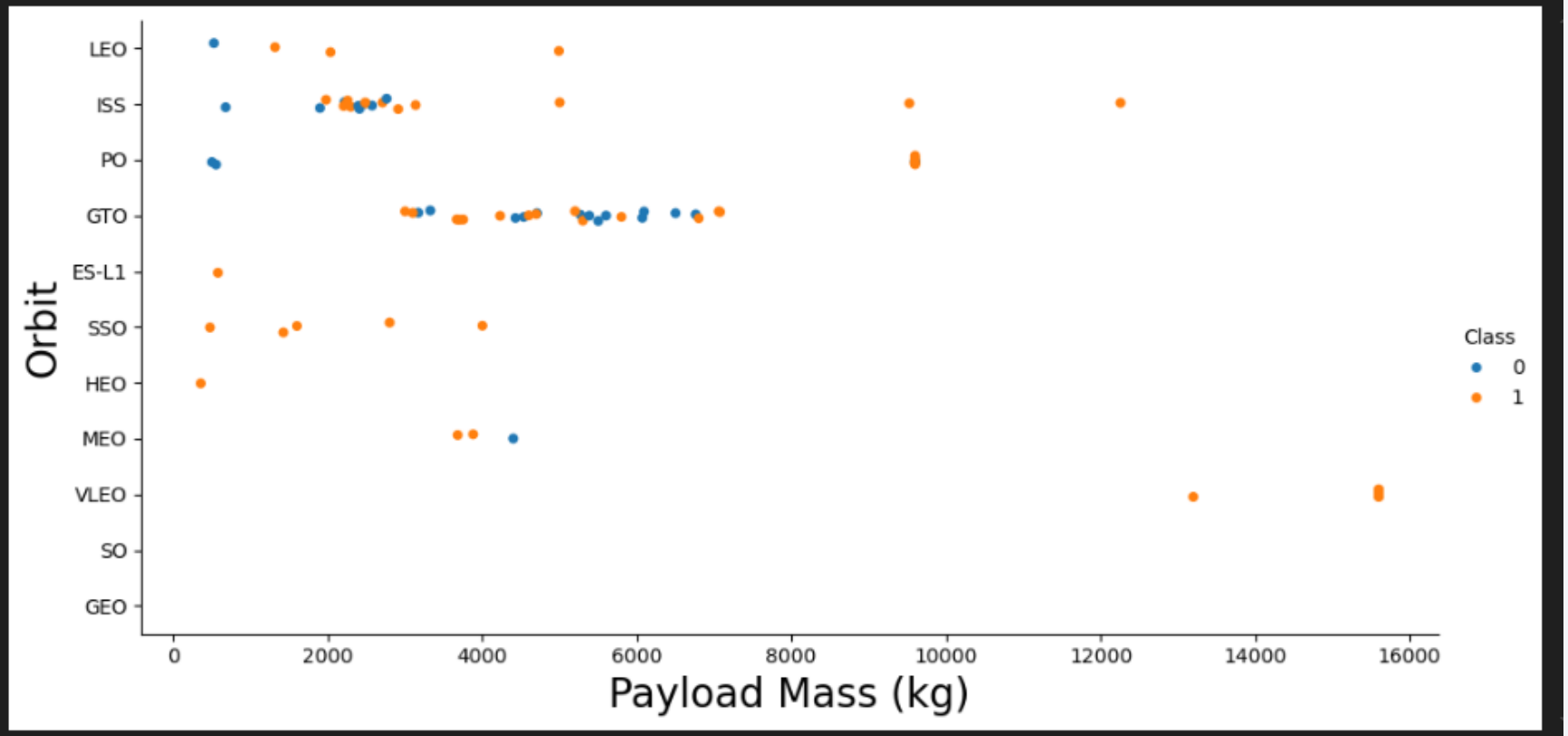
0.0s Python



Payload vs. Orbit Type

```
1
2  # 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
3  sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df_api,aspect=2)
4  plt.xlabel("Payload Mass (kg)",fontsize=20)
5  plt.ylabel("Orbit",fontsize=20)
6  plt.show()
```

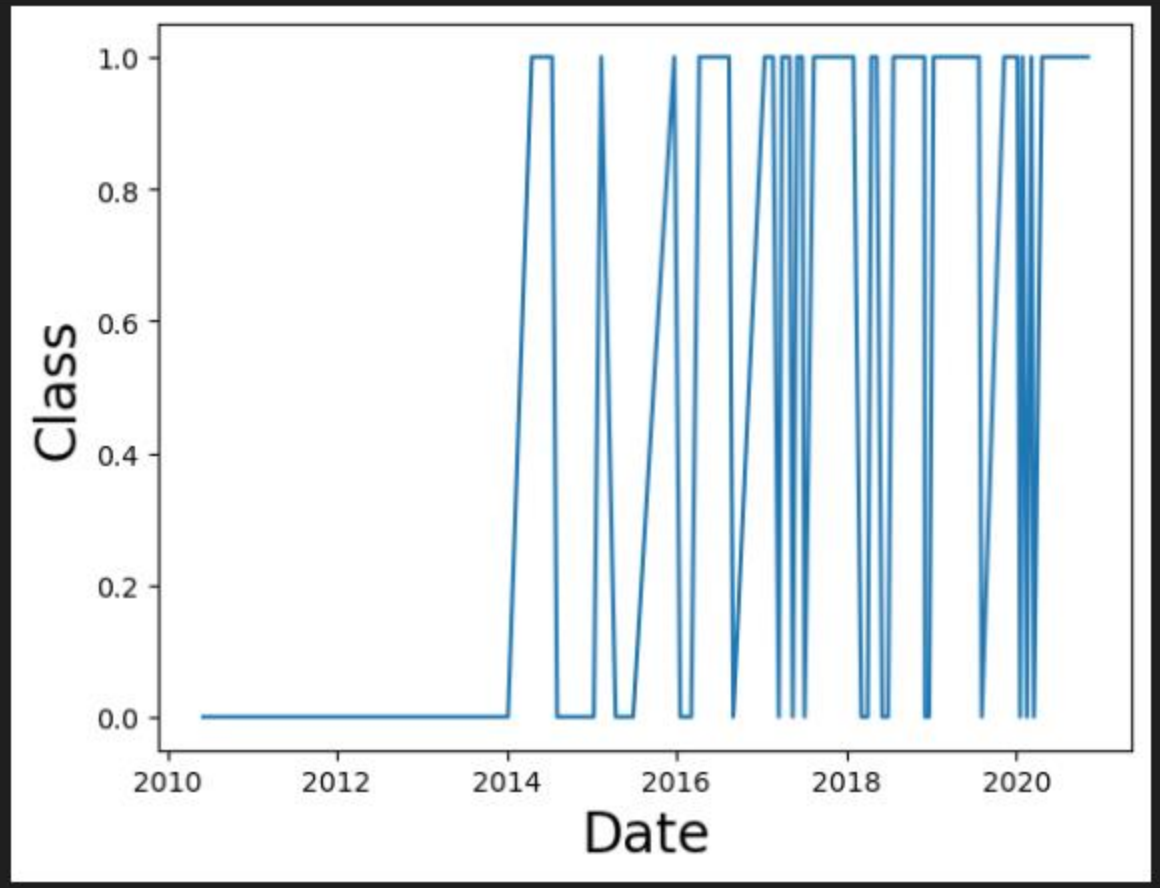
✓ 0.2s

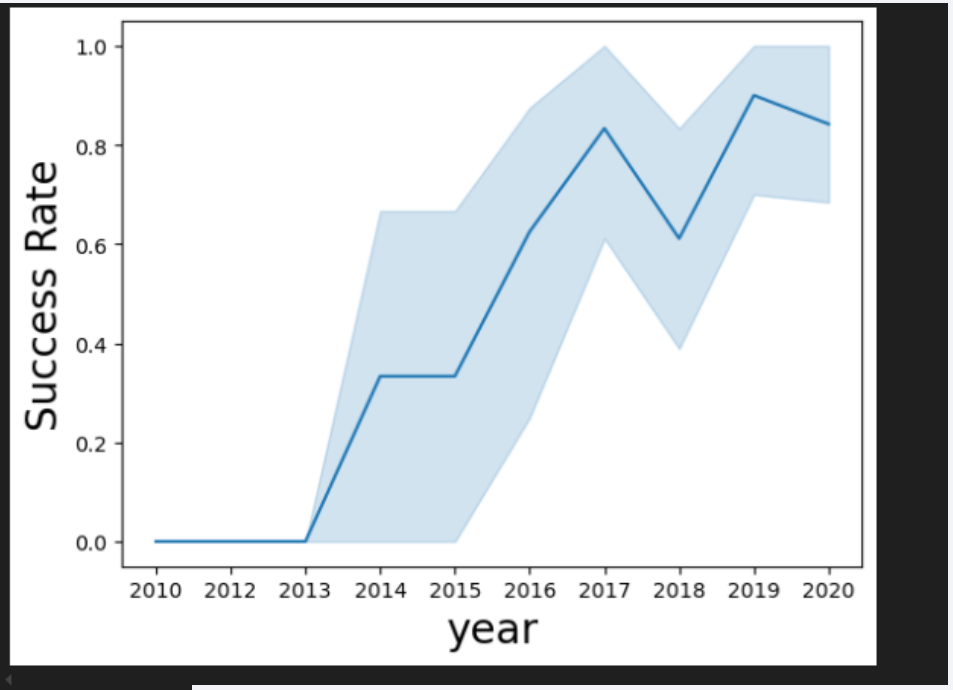


Launch Success Yearly Trend

```
### TASK 6: Visualize the launch success yearly trend
sns.lineplot(y="Class", x="Date", data=df_api)
plt.xlabel("Date",fontsize=20)
plt.ylabel("Class",fontsize=20)
plt.show()
```

✓ 0.0s





All Launch Site Names

Launch Site Names

- CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40)
- CCAFS SLC-40 (Space Launch Complex 40, renamed from LC-40)
- KSC LC-39A (Kennedy Space Center Launch Complex 39A)
- VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4E)

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Launching close to the equator provides a natural boost from Earth's rotation
- Rockets gain additional velocity, reducing fuel and booster needs
- Equatorial advantage is especially valuable for prograde orbits like geostationary transfer orbit (GTO)
- Launch sites like Cape Canaveral and KSC LC-39A at ~28.5°N benefit but a site nearer the equator would receive higher boosts

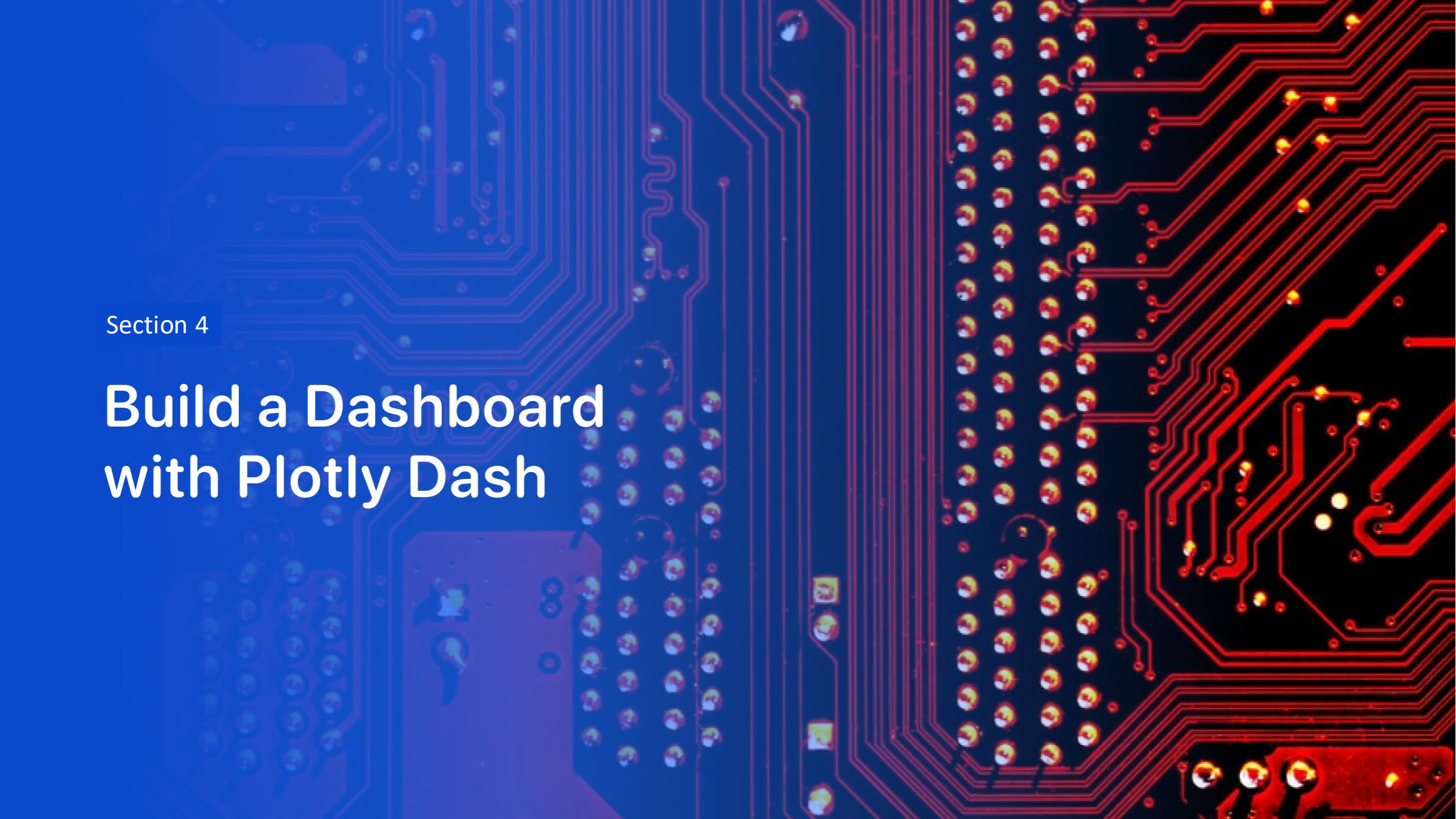
<Folium Map Screenshot 2>

- Green markers: Successful launches
- Red markers: Unsuccessful launches
- CCAFS SLC-40 launch site: 3 successful launches out of 7
- Success rate of 42.9%

<Folium Map Screenshot 3>

Distances from CCAFS SLC-40 Launch Site

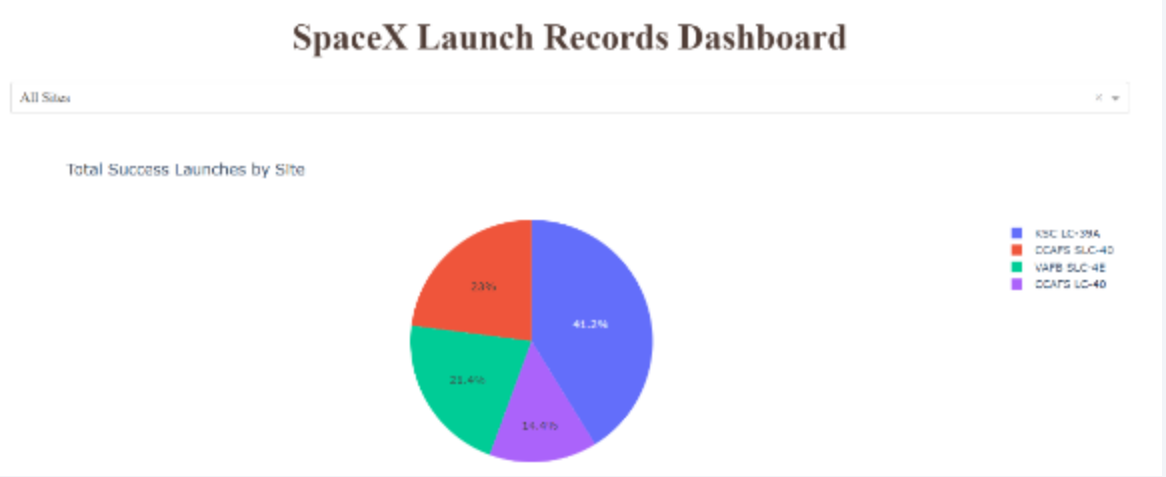
- 0.86 km to nearest coastline
- 21.96 km to nearest railway
- 23.23 km to nearest city
- 26.88 km to nearest highway


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

Launch Success by Site

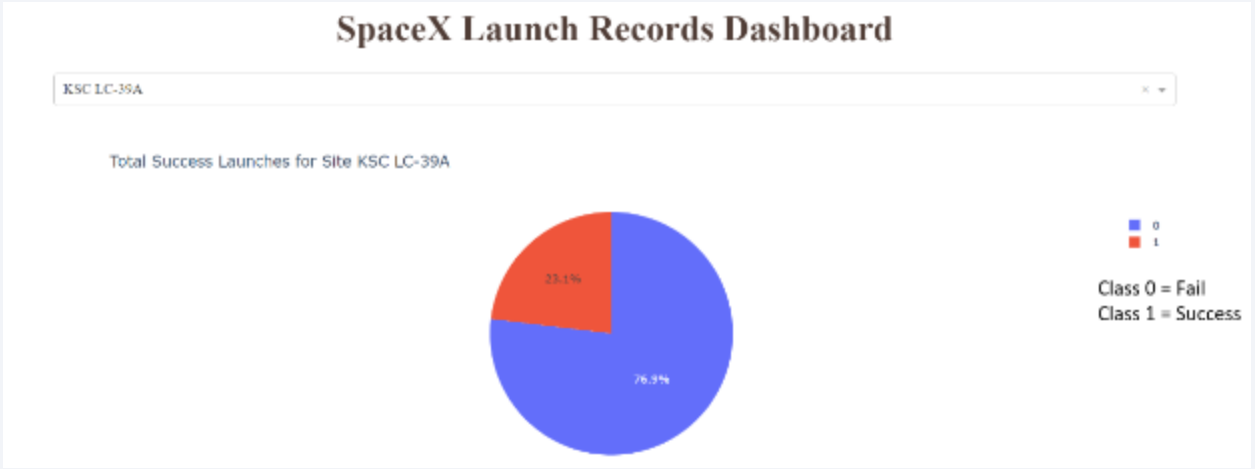
- KSC LC-39A leads with the highest percentage of successful launches
- Success rate at KSC LC-39A: 41.2% of total successful launches among all sites



<Dashboard Screenshot 2>

Launch Success at KSC LC-39A

- Highest success rate among launch sites: 76.9%
- 10 successful launches and 3 failed launches recorded



<Dashboard Screenshot 3>

Payload Mass and Launch Success

- Payloads between 2,000 kg and 5,000 kg exhibit the highest success rates
- Success indicated by 1, failure by 0 in the outcome variable
- This range likely represents optimal payload mass for reliable booster performance

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model Accuracy and Evaluation

- All models showed similar performance and accuracy, likely due to the small dataset size
- Decision Tree model slightly outperformed others based on best_score_
- best_score_ represents the average score across all cross-validation folds for the best parameter combination
- This metric helps evaluate model performance during hyperparameter tuning

```python
from sklearn.metrics import jaccard_score, f1_score

# Examining the scores from Test sets
jaccard_scores = [
                  jaccard_score(Y_test, logreg_yhat, average='binary'),
                  jaccard_score(Y_test, svm_yhat, average='binary'),
                  jaccard_score(Y_test, tree_yhat, average='binary'),
                  jaccard_score(Y_test, knn_yhat, average='binary'),
                 ]

f1_scores = [
             f1_score(Y_test, logreg_yhat, average='binary'),
             f1_score(Y_test, svm_yhat, average='binary'),
             f1_score(Y_test, tree_yhat, average='binary'),
             f1_score(Y_test, knn_yhat, average='binary'),
            ]

accuracy = [logreg_score, svm_cv_score, tree_cv_score, knn_cv_score]

scores_test = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]))
scores_test
```

✓ 1.0s Open 'scores_test' in Data Wrangler

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

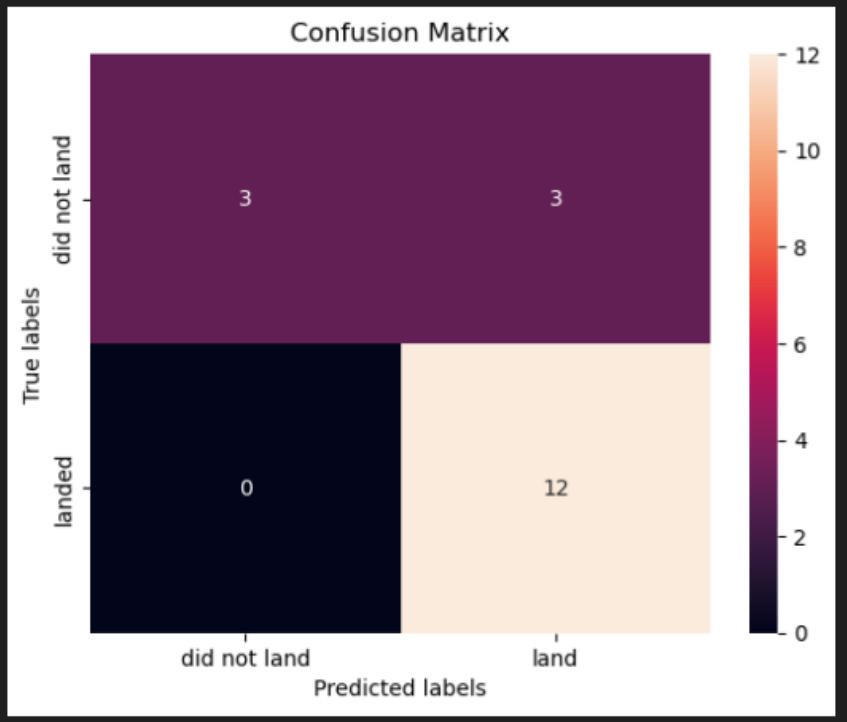
Confusion Matrix

Performance Summary: Confusion Matrix

- A confusion matrix summarizes a classification model's performance by comparing predicted vs. actual outcomes
- All models produced identical confusion matrices in this study
- Presence of false positives (Type I errors) indicates some incorrect positive predictions
- False positives can negatively impact model reliability and need to be minimized

```
parameters = {
    'criterion': ['gini', 'entropy'],
    'splitter': ['best', 'random'],
    'max_depth': [2*n for n in range(1, 10)],
    'max_features': ['sqrt', 'log2', None],  # Correct, no 'auto'
    'min_samples_leaf': [1, 2, 4],
    'min_samples_split': [2, 5, 10]
}
tree = DecisionTreeClassifier()
tree_cv = GridSearchCV(estimator=tree, cv=10, param_grid=parameters).fit(X_train, Y_train)

print("Tuned hyperparameters: (best parameters)", tree_cv.best_params_)
print(f"Model Trainning Accuracy: {tree_cv.best_score_*100:.2f}%")

print(f"Model Testing Accuracy: {tree_cv.score(X_test,Y_test)*100:.2f}%")

tree_pred = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test, tree_pred)

✓ 10.8s
```

```
Tuned hyperparameters: (best parameters) {'criterion': 'entropy', 'max_depth': 6, 'max_features':
Model Trainning Accuracy: 88.93%
Model Testing Accuracy: 83.33%
```



Conclusions

- Model Performance: All models performed similarly, with the Decision Tree slightly outperforming others
- Equator Advantage: Most launch sites are near the equator, gaining a natural boost from Earth's rotation, reducing fuel and booster needs
- Proximity to Coast: All launch sites are close to the coast, facilitating safer launch trajectories over water
- Launch Success: Success rates have increased over time, reflecting improvements in rocket technology and operations
- KSC LC-39A: Highest success rate among launch sites; 100% success for launches under 5,500 kg payload
- Orbits: ES-L1, GEO, HEO, and SSO orbital launches have 100% success rate
- Payload Mass: Higher payload mass correlates with higher success rates across launch sites

Thank you!