IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Haripriya Prakash
20/11/2025

Outline

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

Executive Summary

- **Summary of methodologies**
  - Collected data via SpaceX API and web scraping (Wikipedia).
  - Performed data wrangling and feature engineering.
  - Conducted exploratory data analysis using Python visualization and SQL queries.
  - Built interactive dashboards with Plotly Dash and Folium maps.
  - Developed and compared 4 classification models (Logistic Regression, KNN, SVM, Decision Tree)
- **Summary of all results**
  - Dataset: 90 Falcon 9 launches (2010-2020)
  - Best Model: Decision Tree with 88.89% accuracy
  - Key Finding: Payload mass, orbit type, and launch site are critical predictors
  - Business Impact: Enables competitive cost estimation ($62M vs $165M)

Introduction

- **Project background and context**
  - SpaceX disrupts space industry with reusable rocket technology
  - Falcon 9 launches advertised at $62 million
  - Competitor launches cost upward of $165 million
  - Cost savings primarily from first-stage reusability
- **Problems you want to find answers**
  - Can we predict if the Falcon 9 first stage will land successfully?
  - What factors influence landing success?
  - How can this help competitors estimate launch costs?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - SpaceX REST API calls for launch data
  - Web scraping from Wikipedia for historical records
- **Perform data wrangling**
  - Created binary classification labels
  - Handled missing values
  - Feature engineering with one-hot encoding

Methodology (Cont)

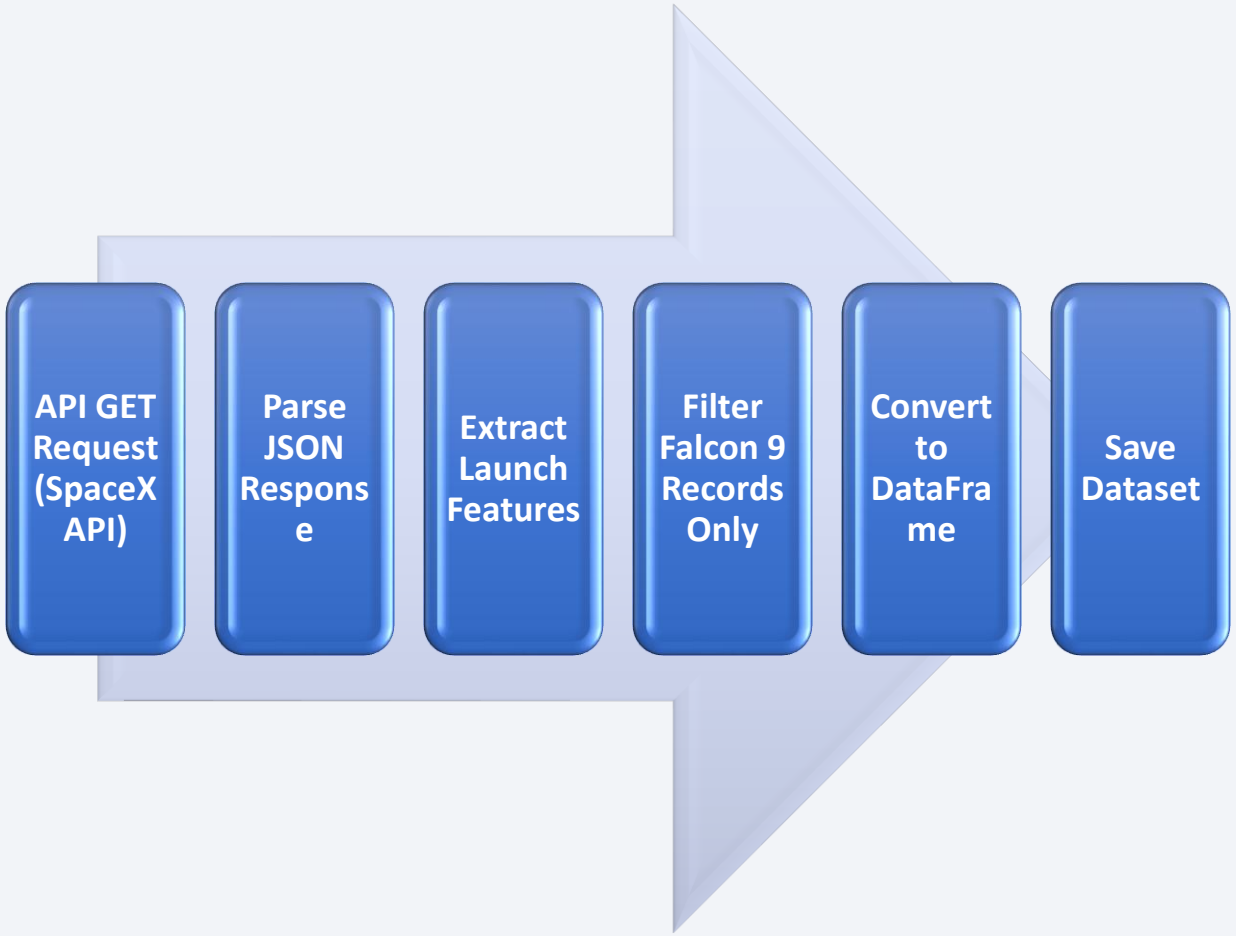
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - Visualization using matplotlib, seaborn
  - SQL queries for statistical analysis
- **Perform interactive visual analytics using Folium and Plotly Dash**
  - Visualization using matplotlib, seaborn
  - SQL queries for statistical analysis
- **Perform predictive analysis using classification models**
  - Built 4 classification models
  - Hyperparameter tuning with GridSearchCV
  - Model evaluation and comparison

Data Collection

- Collected SpaceX Falcon 9 launch data from two sources:
  - Source 1: SpaceX REST API for programmatic data extraction
  - Source 2: Wikipedia web scraping for historical launch records
- Structured data into pandas DataFrame for analysis
- Validated consistency between both data sources

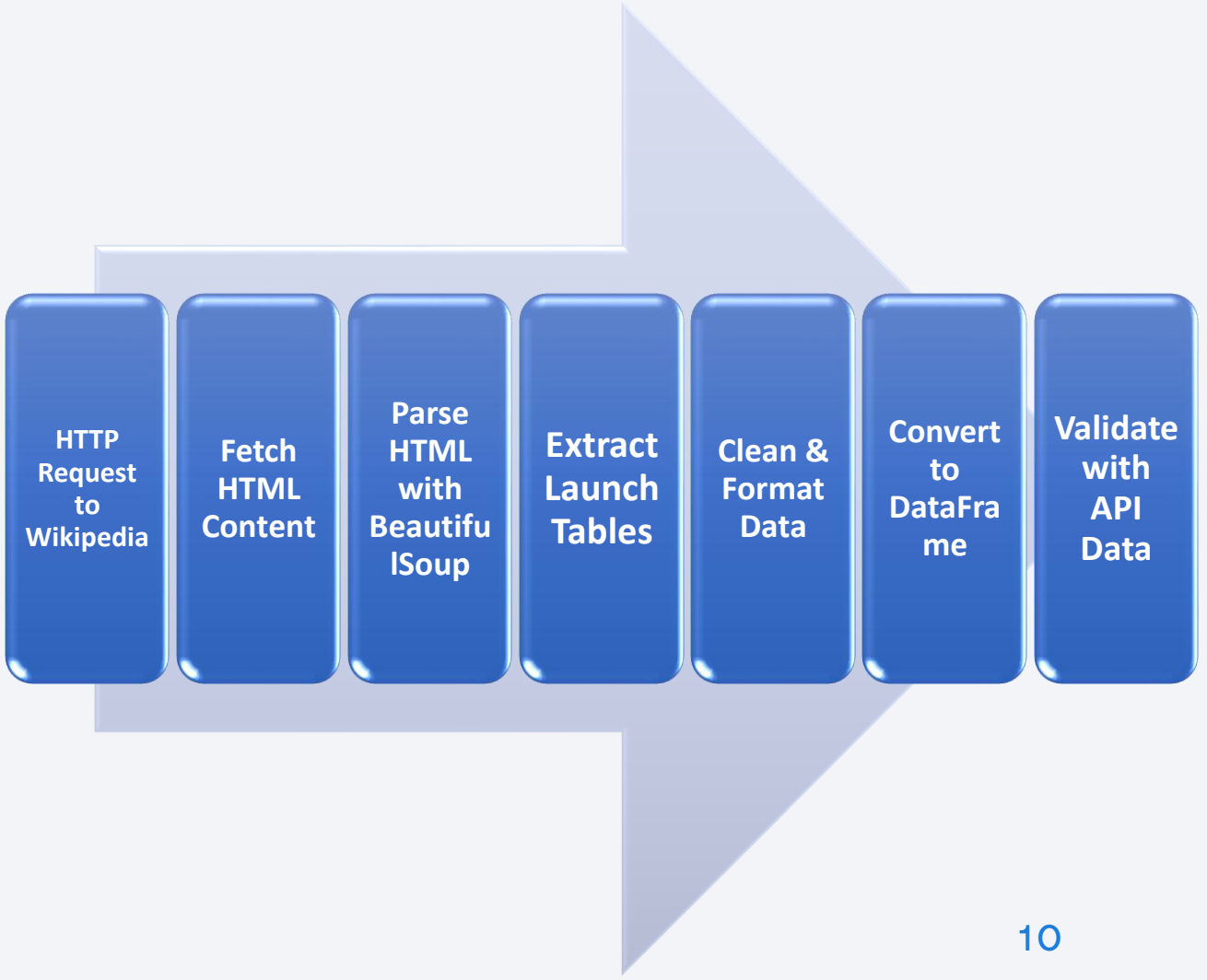
Data Collection – SpaceX API

- Made GET requests to SpaceX public REST API
- Endpoint: https://api.spacexdata.com/v4/launches/past
- Extracted features: flight number, date, rocket type, payload mass, orbit, launch site, landing outcome, grid fins, legs
- Filtered dataset for Falcon 9 launches only
- Converted JSON response to pandas DataFrame
- Saved as structured dataset for machine learning
- **GitHub URL:** AppliedDataScienceCapstone-ProjectFalcon9/1.Data-collection-API.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9



Data Collection - Scraping

- Scraped Falcon 9 launch records from Wikipedia
- URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Used Python requests library for HTTP calls
- Parsed HTML tables with BeautifulSoup
- Extracted columns: flight number, launch date, version, site, payload, orbit, customer, outcome
- Cross-validated with API data for accuracy
- **GitHub URL:** AppliedDataScienceCapstone-ProjectFalcon9/2.Data collection-webscraping.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9



Data Wrangling

- Filtered dataset to Falcon 9 launches only (removed Falcon 1 and Falcon Heavy)
- Created binary classification target variable (Class):
  - Class = 1: Successful landing (True ASDS, True RTLS, True Ocean)
  - Class = 0: Unsuccessful landing (False outcomes, None)
- Handled missing values in PayloadMass column (filled with mean: 6,104.96 kg)
- Performed feature engineering with one-hot encoding:

  - Orbit types (GTO, LEO, ISS, PO, SSO, etc.)

  - Launch sites (CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)

  - Boolean features (GridFins, Reused, Legs)
- Standardized feature scaling for machine learning models
- Final dataset: 90 launches with 18+ features ready for analysis

GitHub URL: AppliedDataScienceCapstone-ProjectFalcon9/3.Data wrangling.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

EDA with Data Visualization

- Charts plotted included:
  - **Scatter plots** for Flight Number vs Launch Site, Payload Mass vs Launch Site
  - **Bar charts** for Success Rate by Orbit Type and
  - **Line charts** for yearly trends
- These visualizations helped uncover patterns in launch success, the impact of payload and site, and improvements over time.
- The chosen charts made it easy to analyze key features like orbit, payload, and launch site against landing success.

- **GitHub URL:** AppliedDataScienceCapstone-ProjectFalcon9/5.EDA-Data Visualization.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

EDA with SQL

- Queried unique launch sites (3 facilities identified-CCAFS, KSC, VAFB)
- Filtered launches by site name patterns
- Calculated total payload mass by customer (NASA CRS)
- Analyzed average payload by booster version F9 v1.1
- Identified first successful landing date (Dec 22, 2015)
- Examined payload ranges for successful drone ship landings
- Computed mission outcome statistics (67% success rate)
- Found maximum payload missions and their outcomes
- Analyzed failed landings by year for learning trends
- Ranked landing outcomes over time (2010-2017)
- **GitHub URL**: AppliedDataScienceCapstone-ProjectFalcon9/4.EDA-SQL.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

Build an Interactive Map with Folium

- Summary of map objects:
  - Markers: Show the geo location for each launch site
  - Circles: Indicate specific event or site with color for success/failure
  - Clusters: Group markers to give an overview when there are many points
  - Lines: Show distances between launch sites and nearby infrastructure
- Purpose of these objects:
  - Visualize where launches happen
  - Explore if launches cluster geographically
  - Analyze proximity to key infrastructure (e.g., coast, road, rail)
  - Distinguish visually between successful and failed landings

- **GitHub URL**:AppliedDataScienceCapstone-ProjectFalcon9/6.Interactive Visual Analytics with Folium.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

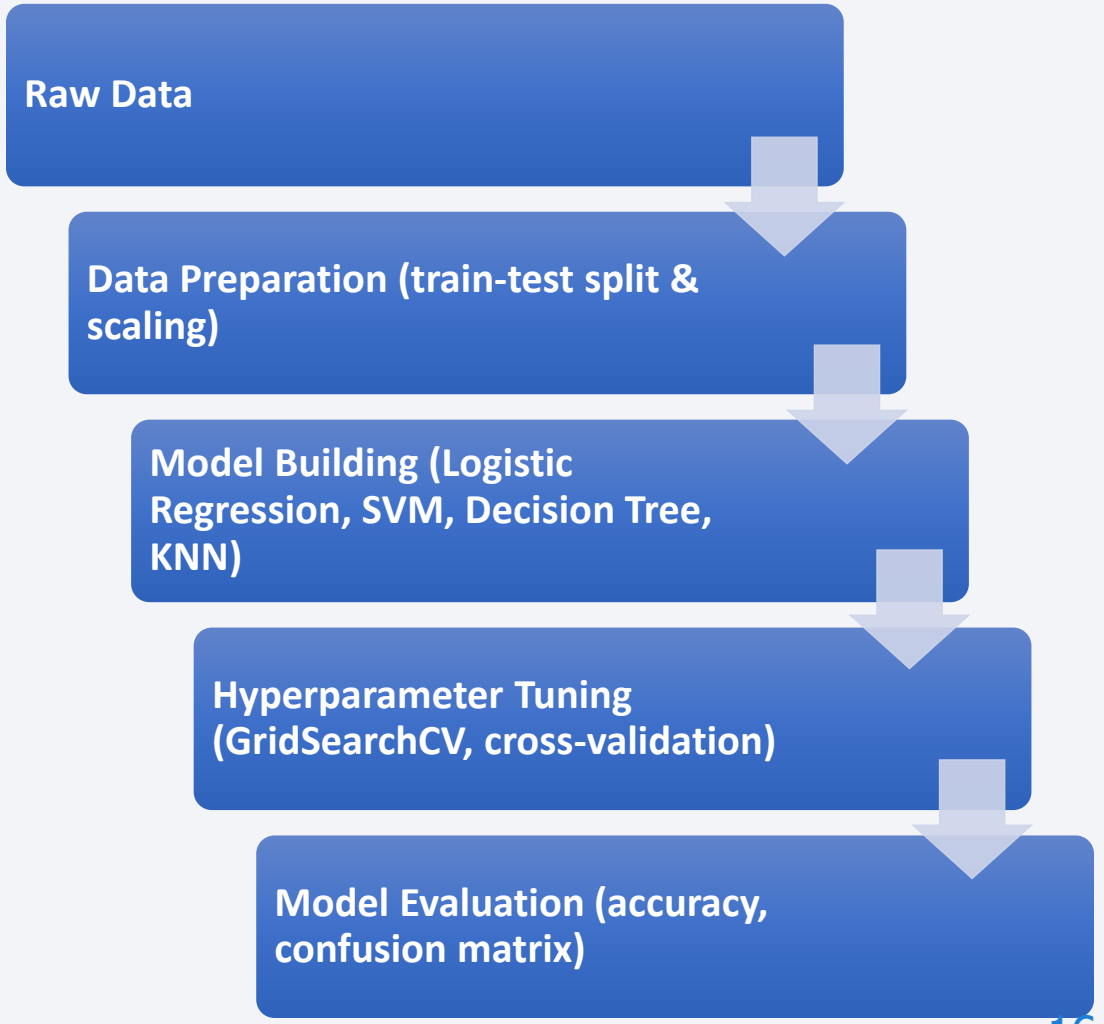
Build a Dashboard with Plotly Dash

- **Plots/graphs and interactions added**:
  - Pie chart showing the percentage of successful vs. failed launches for the selected launch site
  - Scatter plot of payload mass vs. launch outcome, colored by booster version category
  - Dropdown menu to select launch site (All Sites, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
  - Payload range slider to interactively filter launches by payload mass
- Why these plots and interactions were added:
  - The pie chart gives a clear view of the success rate for each launch site
  - The scatter plot helps analyze how payload mass and booster version relate to landing success
  - The dropdown lets users compare performance across different launch sites
  - The slider allows users to focus on specific payload ranges and see how success changes

- **GitHub URL:** AppliedDataScienceCapstone-ProjectFalcon9/7.spacex-dash-app.py at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

Predictive Analysis (Classification)

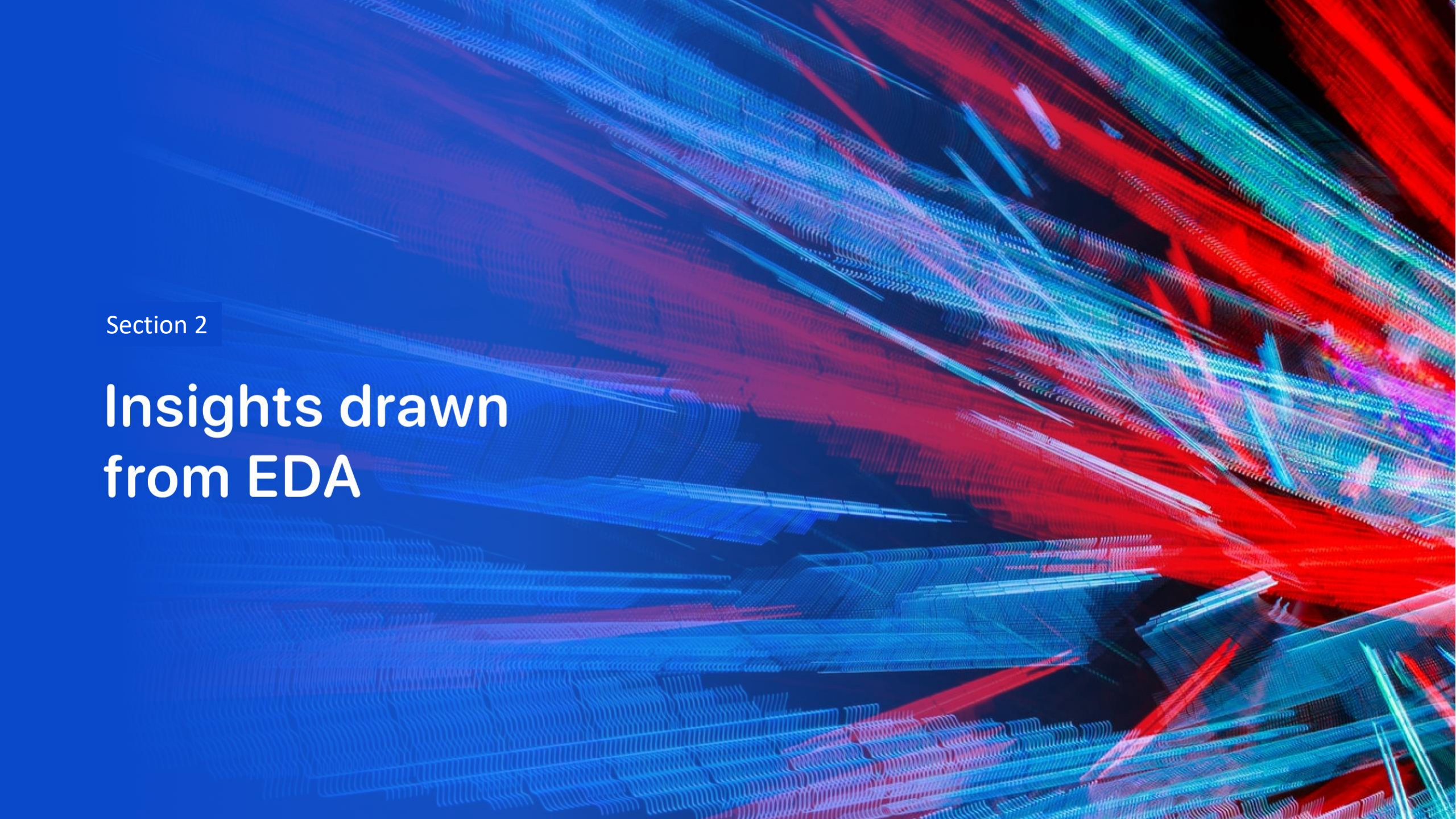
- Split the dataset into train and test sets
- Standardized the feature values
- Built and trained four classifiers:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- Used GridSearchCV with 10-fold cross-validation for best parameters on all models
- Measured model accuracy on test data
- Plotted confusion matrices for classification performance
- Model Comparison:
  - Decision Tree had the highest test accuracy: 0.89
  - Logistic Regression, SVM, and KNN: 0.83
- Best Performing Model:Decision Tree Classifier (Test accuracy: 0.89)



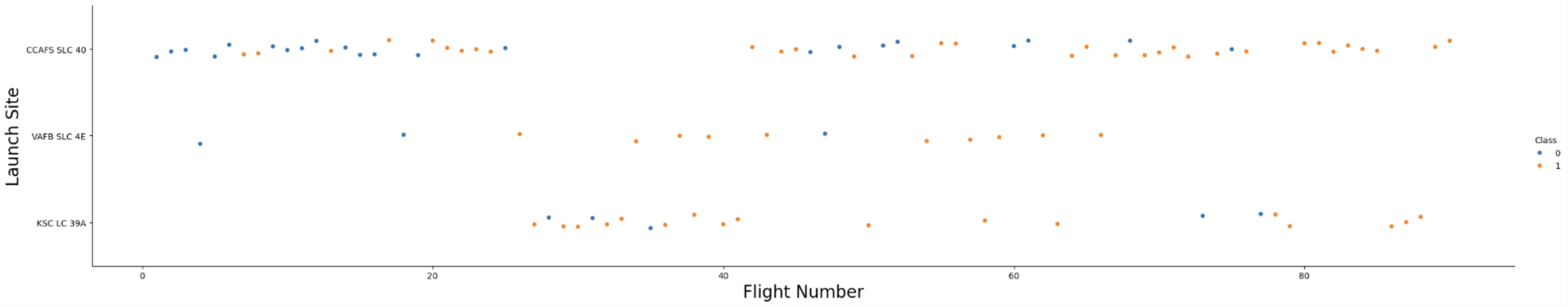
GitHub URL:AppliedDataScienceCapstone-ProjectFalcon9/8.SpaceX-Machine-Learning-Prediction.ipynb at main · haripriya-prakash/AppliedDataScienceCapstone-ProjectFalcon9

Results

- EDA Findings:
  - Success rate improved from 0% (2015) to 90% (2020)
  - KSC LC-39A site has highest success rate (~75%)
  - Optimal payload: 2,000-8,000 kg
  - LEO/ISS orbits: 85% success | GTO orbits: 40% success
- Interactive Analytics:
  - Folium maps revealed infrastructure proximity advantage for KSC
  - Dashboard enabled dynamic exploration of payload-site-success patterns
- Machine Learning Results:
  - Best Model: Decision Tree (89% accuracy)
  - Other models: 83% accuracy (Logistic Reg, SVM, KNN)
  - Key factors: Payload mass, launch site, orbit type, flight experience

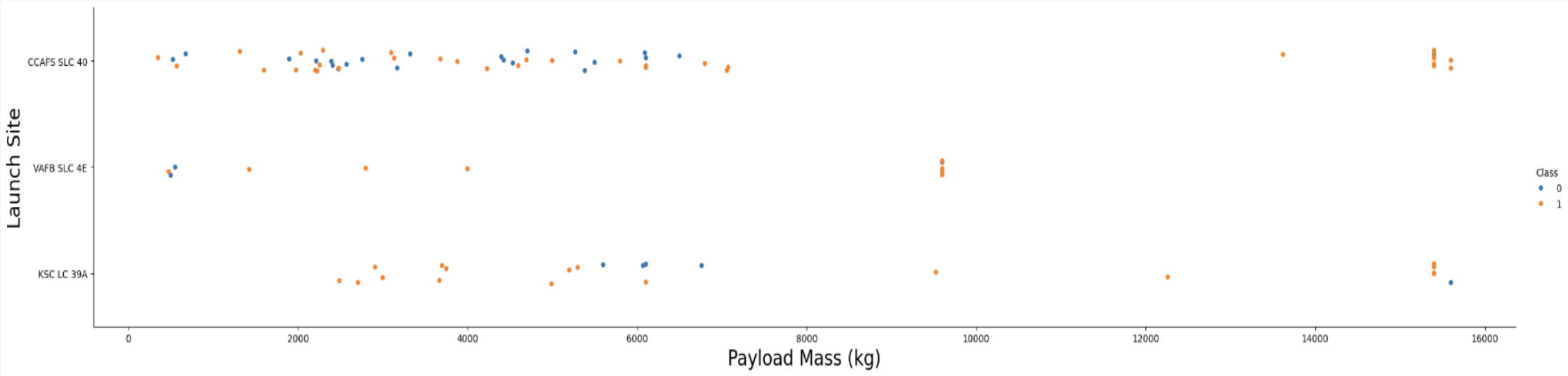

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Scatter plot shows each launch by site (y-axis) and flight number (x-axis)
- Color-coded: Blue (0) = Failed landing | Orange (1) = Successful landing
- Early flights (low flight numbers) show more failures (blue dots)
- Later flights show increased success (more orange dots)
- KSC LC-39A and CCAFS SLC-40 show higher success rates in later missions
- Insight: Success rate improved over time as SpaceX gained experience

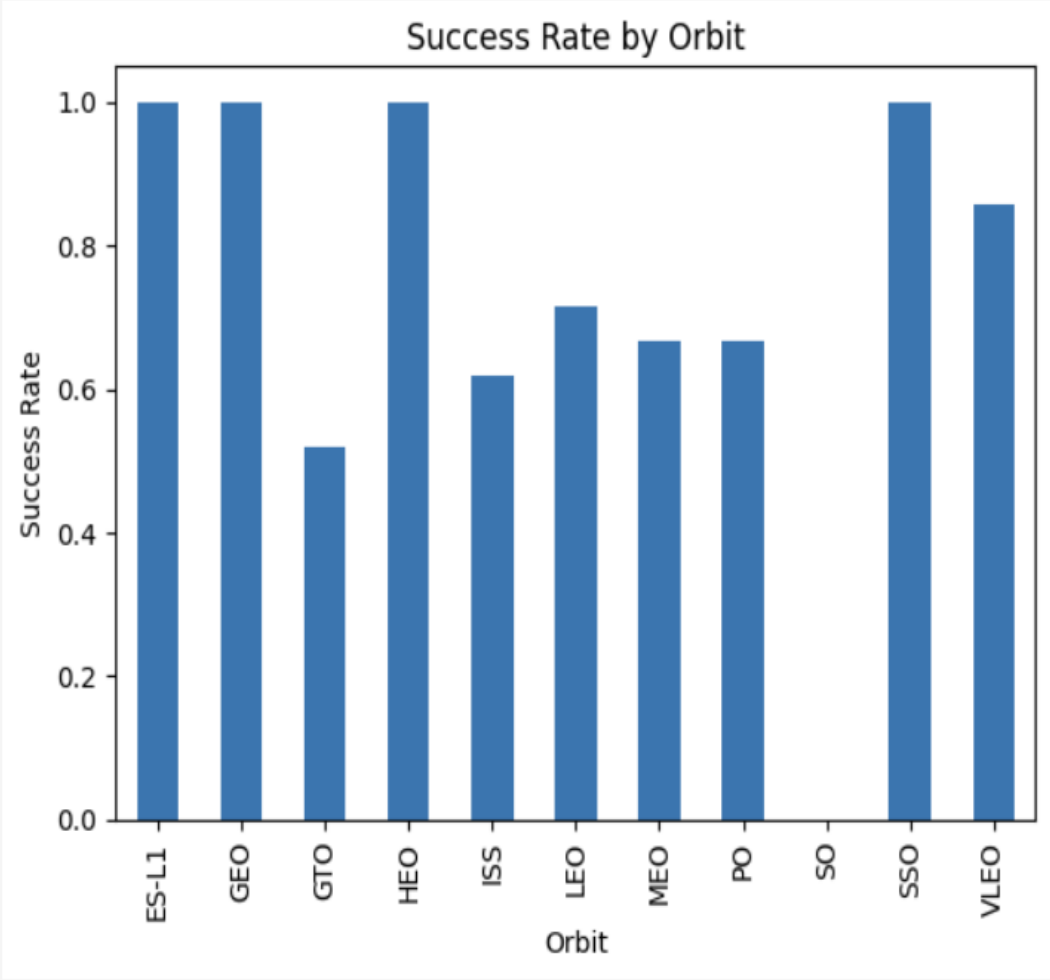
Payload vs. Launch Site



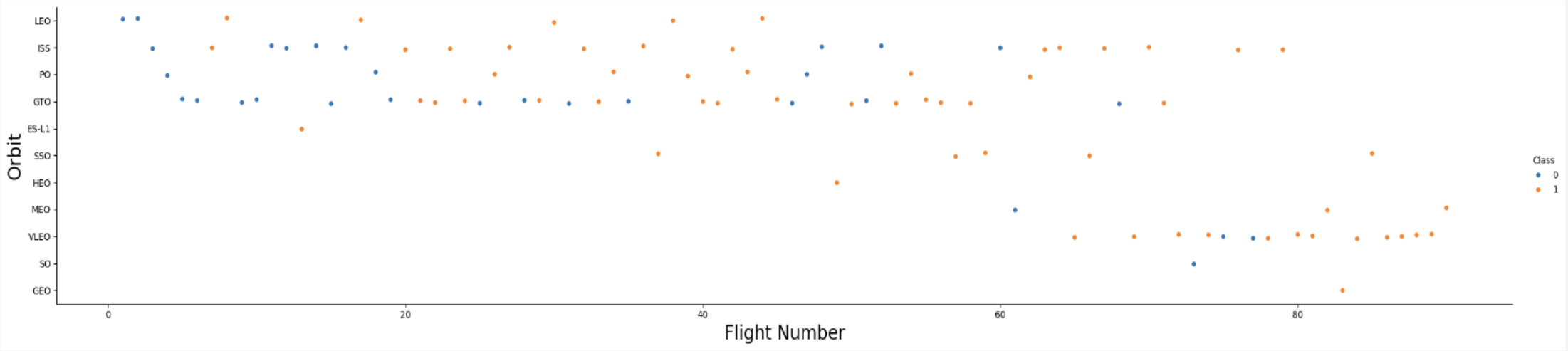
- Each point: one Falcon 9 launch (y = launch site, x = payload mass in kg)
- Color: Blue (0) = failed landing, Orange (1) = successful landing
- CCAFS SLC-40: Wide payload range, mix of results
- KSC LC-39A: Mostly moderate payloads, many successes
- VAFB SLC-4E: Fewer launches, moderate to low payloads
- Low payloads (<4000 kg): higher success at all sites
- Very high payloads (>10,000 kg): often successful, but fewer attempts
- Insight: Moderate payloads and KSC LC-39A give best results

Success Rate vs. Orbit Type

- Bar chart shows landing success rate for each orbit type.
- Highest success rate: ES-L1, GEO, HEO, SSO (mostly close to 1.0).
- LEO, VLEO, ISS orbits: also high success, but some failures.
- GTO orbit: lowest success rate (~0.5), due to high payload and mission difficulty.
- Some rare orbits (SSO, HEO) have 100% success but few launches.
- Key Insight: Missions to LEO/SSO/ISS are most reliable; GTO and challenging orbits have more frequent failures.

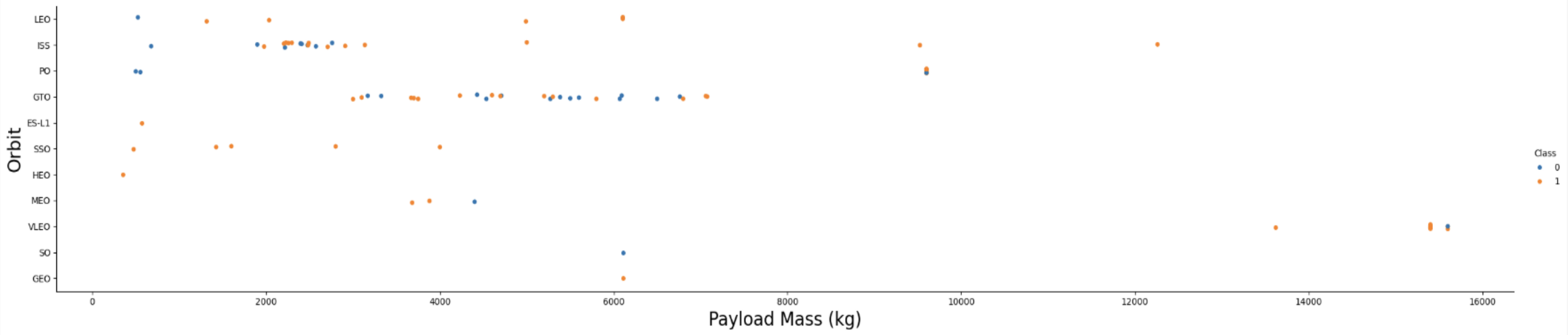


Flight Number vs. Orbit Type



- Each point: one launch (x = flight number, y = orbit type)
- Color: Blue (0) = failed landing, Orange (1) = successful landing
- Early flights (low flight numbers) mostly LEO, ISS, GTO orbits
- More orbit types (diversification) as flight numbers increase
- Success rate improves in later flights for most orbits (more orange on right)
- Some orbits (LEO, ISS, SSO) have higher concentration of successes

Payload vs. Orbit Type



- Each point: one launch (x = payload mass in kg, y = orbit type)
- Color: Blue (0) = failed landing, Orange (1) = successful landing
- Most successful landings (orange) for moderate payloads (2,000–8,000 kg), mainly LEO and ISS orbits
- GTO and other high-energy orbits: more failures (blue), often heavier payloads
- ES-L1, SSO, and some rare orbits: fewer launches, typically successful
- Insight: Moderate payload with LEO/ISS orbits gives best success; GTO orbits and very heavy payloads are more challenging

Launch Success Yearly Trend

- Line chart shows average landing success rate by year (2010–2020)
- Success rate was 0% until 2014, then rose sharply from 2015 onward
- Major jump after first successful landing in Dec 2015
- Steady improvements, reaching above 80% by 2019–2020
- Key Insight: Experience and technology upgrades led to dramatic rise in launch success over time.

All Launch Site Names

```
In [12]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

- There are four unique launch site names in the SpaceX data. Most are located at Cape Canaveral, with one each at Kennedy Space Center and Vandenberg Air Force Base.

Launch Site Names 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	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

- All five records are launches from Cape Canaveral sites beginning with 'CCA', reflecting the high launch frequency from this region.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [14]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[14]: SUM(PAYLOAD_MASS__KG_)
         45596
```

- NASA Commercial Resupply Services (CRS) missions have delivered a combined total of 45,596 kg of payload mass to the International Space Station using SpaceX boosters. This highlights SpaceX's role in supporting ISS cargo missions for NASA.

Average Payload Mass by F9 v1.1

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

In [15]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Out[15]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass carried by Falcon 9 v1.1 booster missions was about 2928 kg, indicating its moderate-lift launch capability compared to later versions.

First Successful Ground Landing Date

```
In [16]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
Out[16]: MIN(Date)
         2015-12-22
```

- The first successful Falcon 9 booster landing on a ground pad occurred on December 22, 2015. This historic milestone marked the beginning of routine booster recovery for SpaceX.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [17]:

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

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

Out[17]:

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

- These Falcon 9 Full Thrust boosters successfully landed on drone ships while carrying payloads between 4000 kg and 6000 kg. This demonstrates the reliability of these booster versions for medium-heavy satellite launches.

Total Number of Successful and Failure Mission Outcomes

In [18]:
```
%%sql
SELECT Mission_Outcome, COUNT(*) AS OutcomeCount
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

Out[18]:

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

- The vast majority of SpaceX missions in the dataset were successful (98 out of 101). Only a single failure (in flight) and one “success” with unclear payload status were recorded, reflecting high overall mission reliability.

Boosters Carried Maximum Payload

- Multiple Falcon 9 Block 5 boosters have carried the maximum recorded payload mass of 15,600 kg. This highlights the strong heavy-lift capability and reusability of the Block 5 version.

List all the booster_versions that have carried the maximum payload mass, using a subquery suitable aggregate function.

In [19]:
```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Out[19]:

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

2015 Launch Records

In [20]:
```
%%sql
SELECT substr(Date,6,2) AS Month,
       Landing_Outcome,
       Booster_Version,
       Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date,1,4) = '2015';
```

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

Out[20]:

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

- In 2015, there were two failed attempts to land Falcon 9 boosters on drone ships, both using v1.1 boosters from Cape Canaveral SLC-40. These early failures paved the way for later successful landings as SpaceX improved technology and techniques.

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

- Between 2010 and early 2017, most missions did not attempt a landing. Success and failure counts for drone ship landings were equal, while only a few achieved success on ground pads. The variety of outcomes reflects SpaceX's experimentation and rapid progress with landing technology in this period.

the date 2010-06-04 and 2017-03-20, in descending order.

In [21]:
```
%%sql
SELECT Landing_Outcome, COUNT(*) AS OutcomeCount
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY OutcomeCount DESC;
```

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

Out[21]:

Landing_Outcome	OutcomeCount
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

Section 3

Launch Sites Proximities Analysis

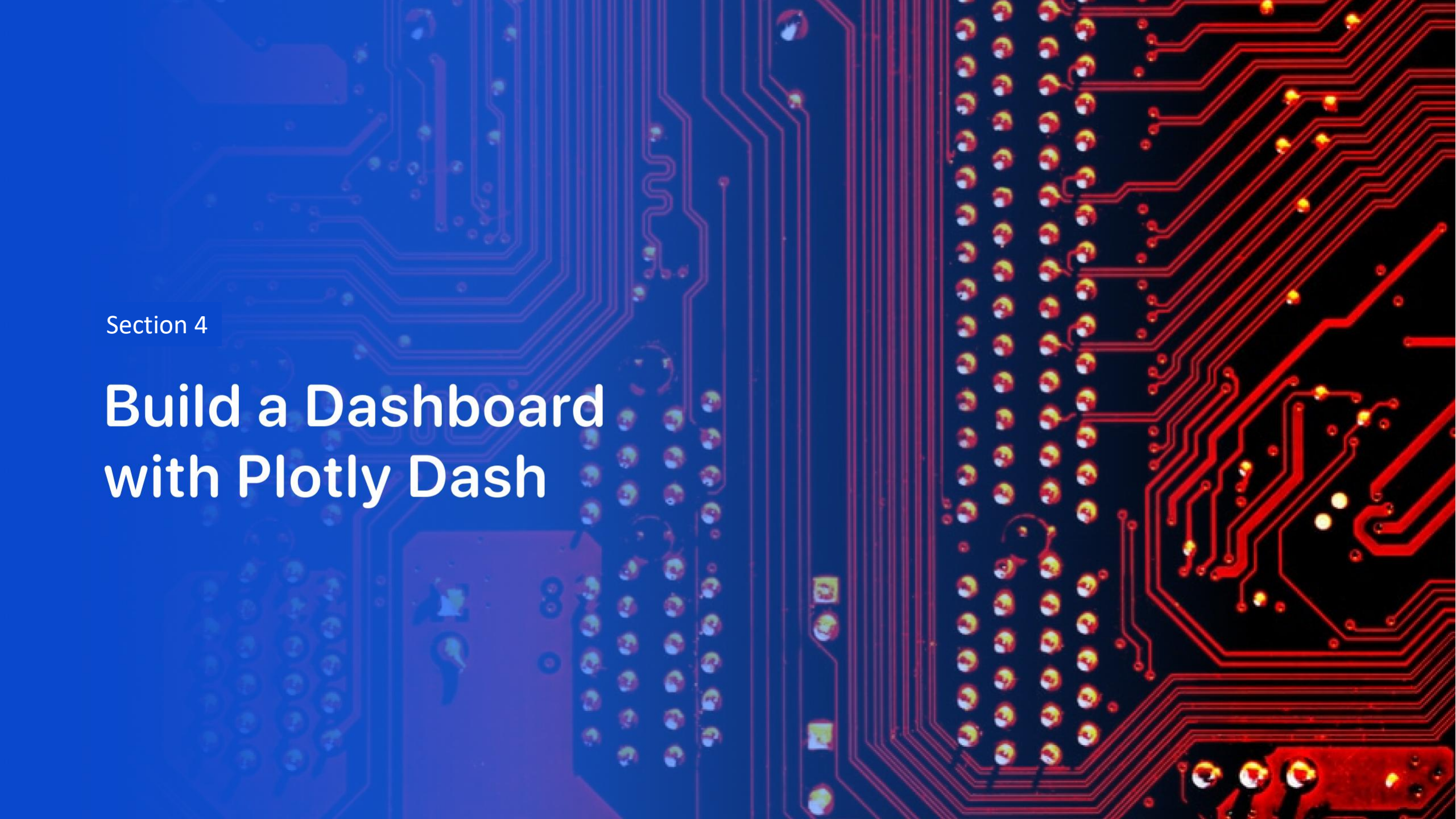
All Launch Sites: Location Markers

Color-Coded Markers for Launch Site Outcomes

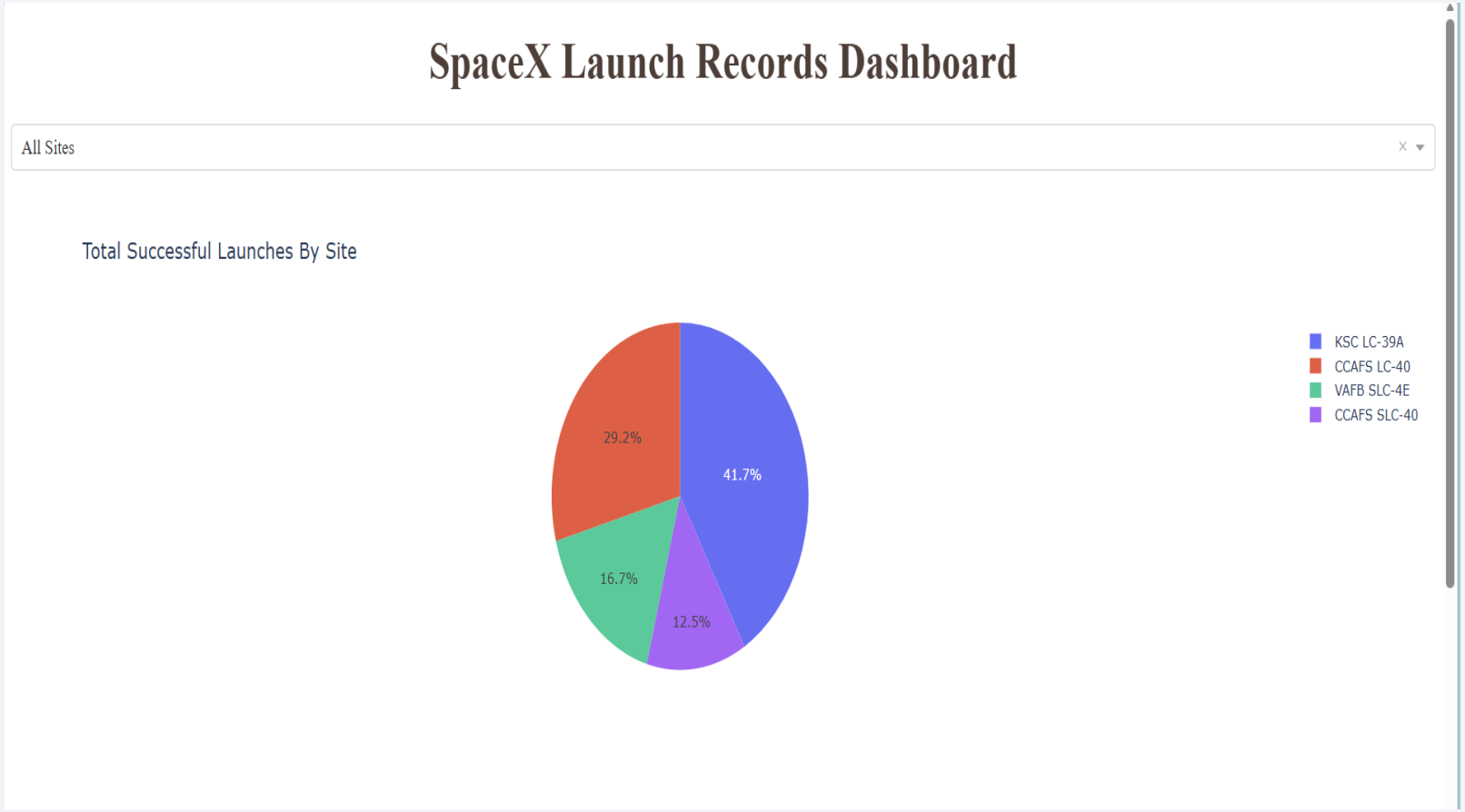
- The folium map shows SpaceX launch sites. Each marker's color indicates outcome: green for successful landings, red for failures. Marker clusters summarize results at busy sites, making it easy to see performance at a glance.

Launch Site Proximity to Coastline

- A blue line connects the launch pad and the nearest coastline.
- The distance (0.59 km) is shown using Geopy's calculation.
- This helps visualize how close the launch site is to water for safety and planning.

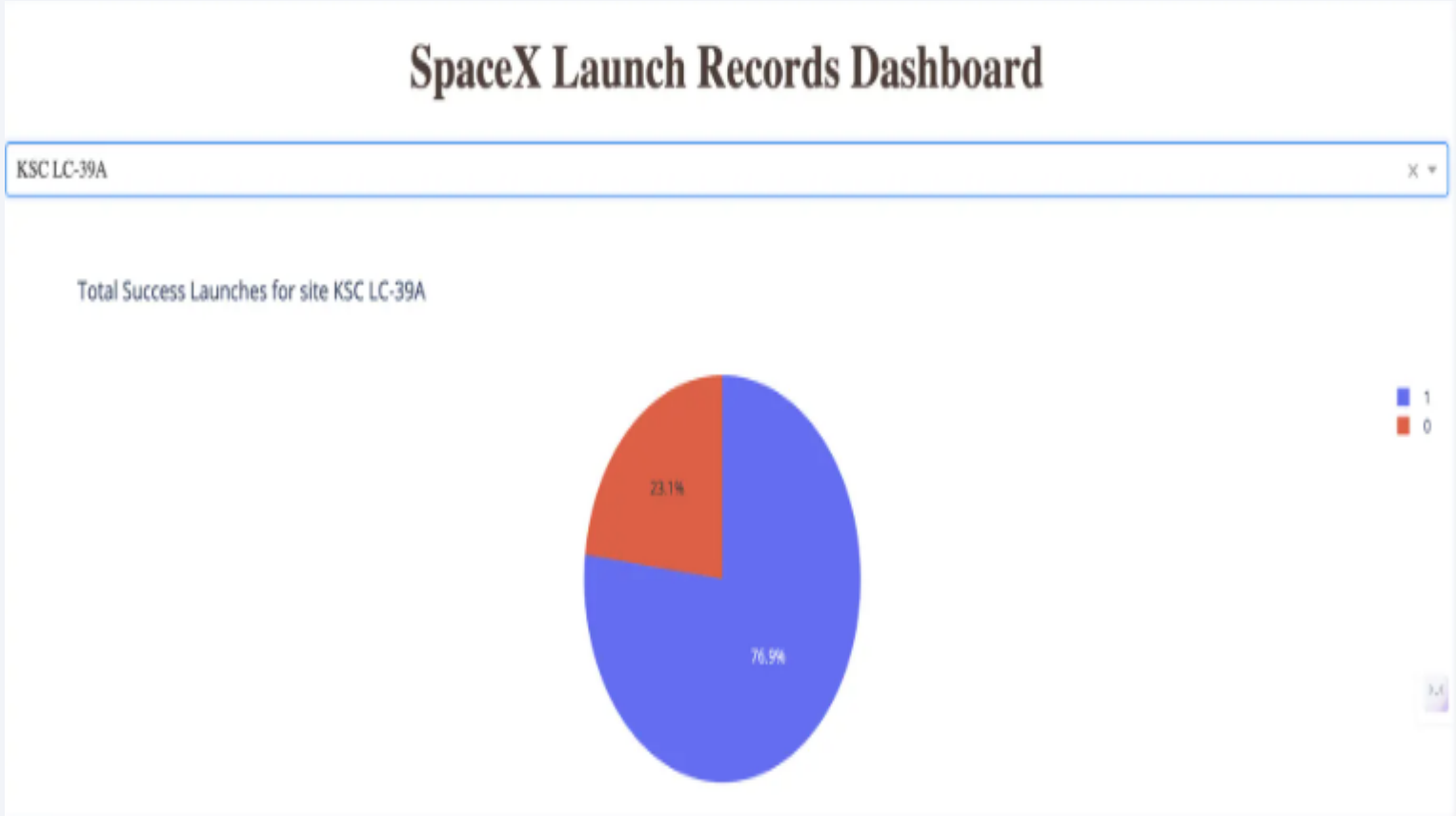

Section 4
Build a Dashboard with Plotly Dash

Launch Success Distribution by All Site



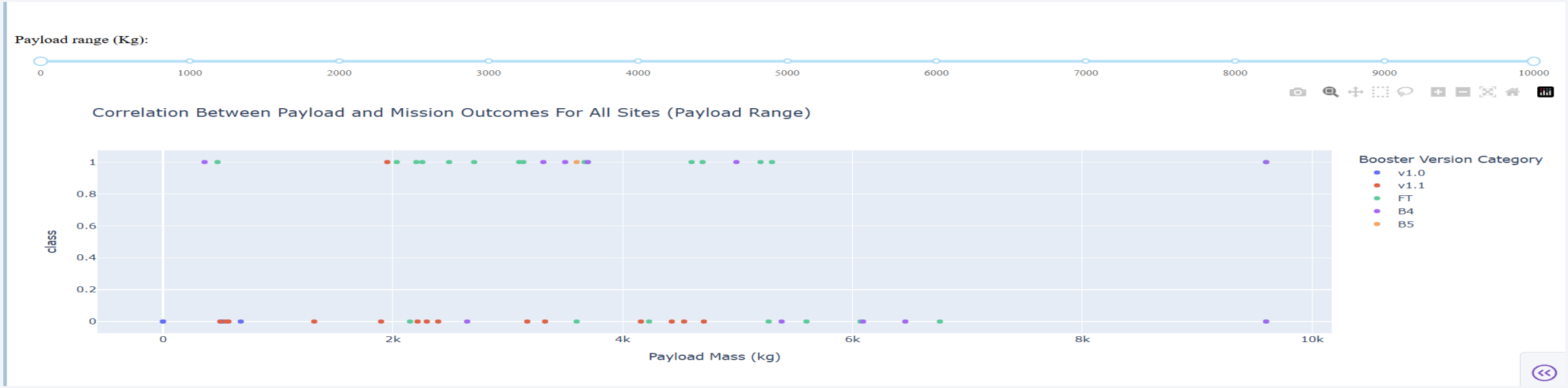
- Pie chart shows the percentage of successful launches for each site.
- KSC LC-39A has the largest share

Launch Success Ratio at KSC LC-39A (Highest Performance Site)



- The pie chart shows the proportion of successful vs. unsuccessful launches at KSC LC-39A.
- KSC LC-39A has the highest launch success ratio, with about 76.9% of launches marked as successful.
- The color legend helps easily identify success (blue) and non-success (red).

Payload Mass vs. Launch Outcome by Booster Version



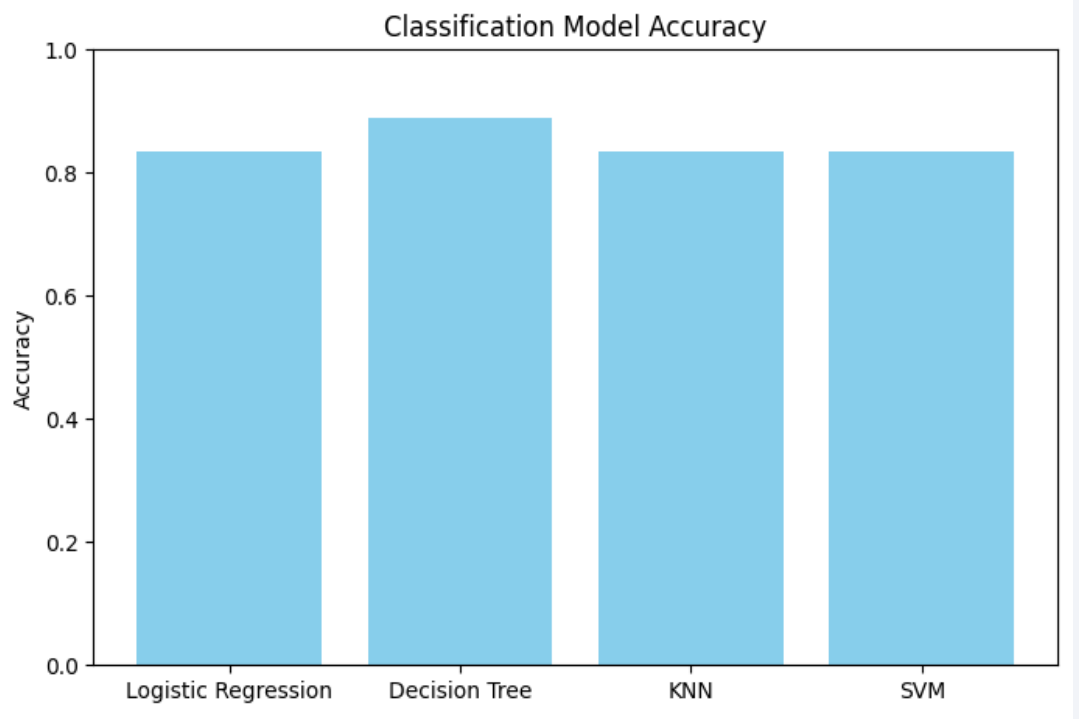
- Scatter plot shows how payload mass relates to launch success.
- Each dot is a launch, colored by booster type.
- Filter outcomes by payload with the slider.
- Modern boosters succeed more, especially at lower/mid payloads.
- Older boosters and heavy payloads have more failures.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

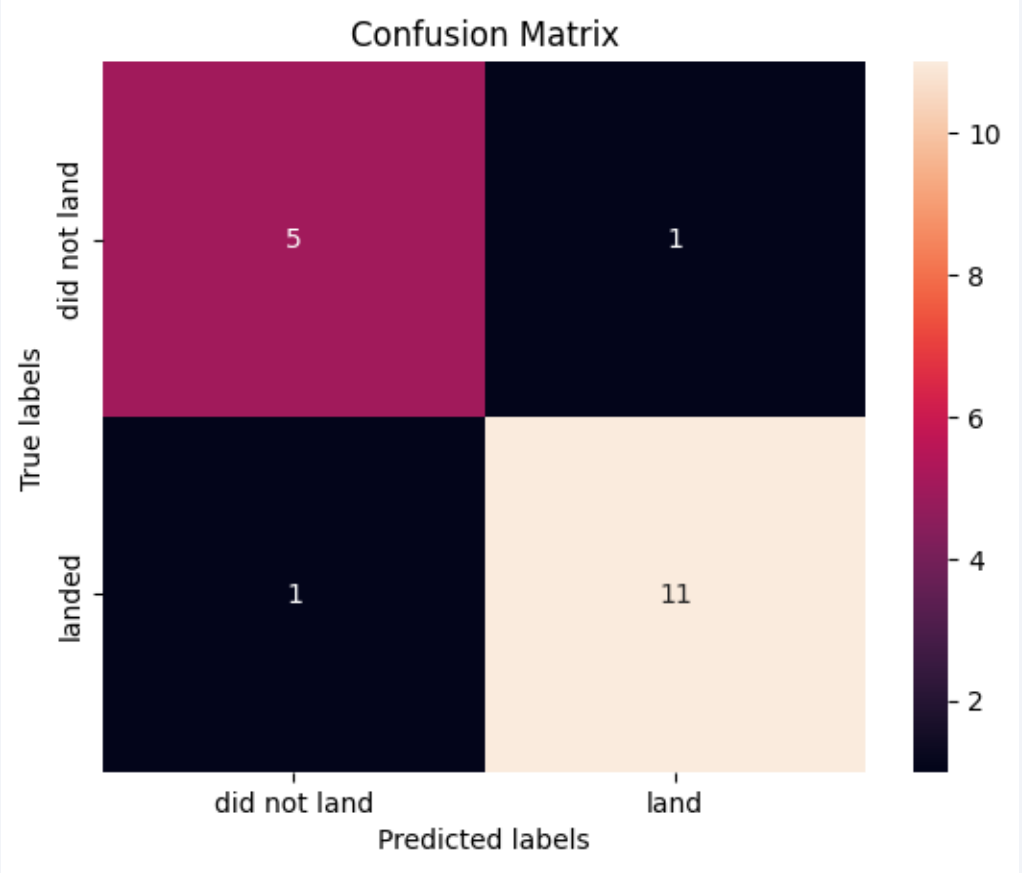
- The bar chart shows the classification accuracy for each built model.
- Each bar represents a different model (Logistic Regression, Decision Tree, KNN, SVM).
- The Decision Tree model has the highest classification accuracy among all.




```
--- Model Performance Comparison (Test Set) ---
                Model  Test Accuracy
        Decision Tree       0.888889
  Logistic Regression       0.833333
                  SVM       0.833333
                  KNN       0.833333
```

Confusion Matrix

- Confusion Matrix of the Best Model (**Decision Tree**):
- The matrix shows how well the model predicts land vs. did not land outcomes.
- Diagonal values: Correct predictions (5 “did not land”, 11 “landed”)
- Off-diagonal values: Errors (1 instance each type)
- Interpretation: The model is accurate, with very few misclassifications.



Conclusions

- Exploratory analysis revealed that launch success rates improve over time, with modern booster versions (FT, B4, B5) showing significantly higher reliability.
- KSC LC-39A demonstrated the highest launch success rate among all sites, making it the most reliable launch location.
- Payload mass and booster type are strong predictors of landing success, with moderate payloads and newer boosters achieving better outcomes.
- Machine learning models (Logistic Regression, Decision Tree, KNN, SVM) accurately predicted launch outcomes, with the Decision Tree model achieving the highest classification accuracy.
- The confusion matrix confirmed the Decision Tree model's strong performance with minimal misclassifications, validating its predictive reliability.
- These insights enable data-driven decision-making for optimizing launch site selection, payload planning, and mission risk assessment.

Appendix

- Data Sources
  - SpaceX REST API (launch records and mission details)
  - Web scraping from SpaceX Wikipedia page (historical launch data)
- Tools & Technologies
  - Programming Language: Python 3.x
  - Data Processing: Pandas, NumPy
  - Data Visualization: Matplotlib, Seaborn, Plotly, Folium
  - Interactive Dashboard: Dash by Plotly
  - Machine Learning: Scikit-learn (Logistic Regression, Decision Tree, KNN, SVM)
  - SQL Analysis: SQLite
- Key Notebooks
  1. Data Collection (API & Web Scraping)
  2. Data Wrangling
  3. Exploratory Data Analysis (EDA) - SQL
  4. EDA - Data Visualization
  5. Interactive Visual Analytics with Folium
  6. Interactive Dashboard Development
  7. Machine Learning Prediction

Thank you!