IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- This project analyzes SpaceX Falcon 9 launch data to understand the factors influencing first-stage landing success.
- The workflow includes data collection (API + web scraping), data wrangling, exploratory data analysis (visual + SQL), interactive analytics (Folium + Plotly Dash), and predictive modeling.
- Key findings show that launch site, booster version, payload mass, and orbit type strongly correlate with landing success.
- A Decision Tree model achieved the best performance, making it suitable for estimating future mission outcomes

Introduction

- SpaceX significantly reduces launch costs by recovering and reusing Falcon 9 first-stage boosters. Understanding what drives successful landings is crucial for improving reliability and reducing mission cost.
- This project aims to answer the central question:
  Can we predict whether a Falcon 9 first stage will land successfully based on historical launch data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected launch records using SpaceX REST API and web scraping.
- Perform data wrangling
  - Cleaned, transformed, and harmonized raw launch data for analysis.
- Conducted using Python visualizations and SQL queries to identify trends and key factors.
- Created geospatial visualizations with Folium and a fully interactive dashboard with Plotly Dash.
- Perform predictive analysis using classification models
  - Built, tuned, and evaluated multiple ML models (Logistic Regression, SVM, KNN, Decision Tree).

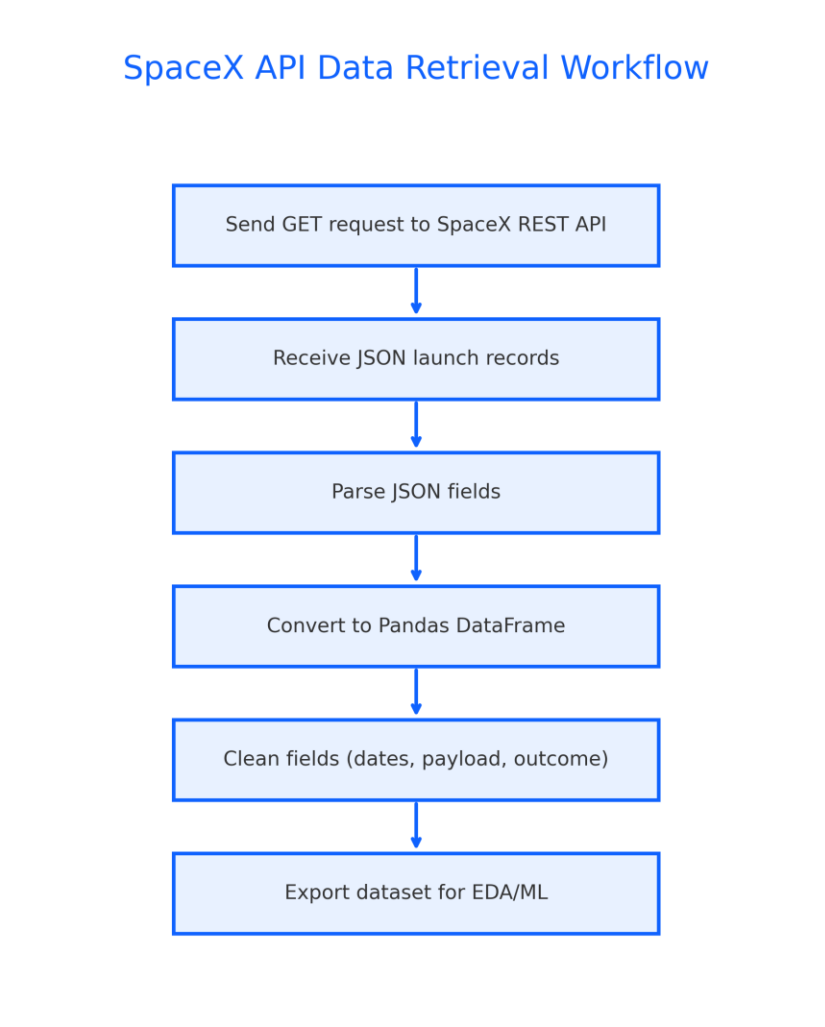
Data Collection

- Collected SpaceX Falcon 9 launch data from two primary sources:
- Public SpaceX REST API
- Web scraping of launch records from HTML tables
- Extracted key attributes such as payload mass, orbit type, launch site, booster version, and landing outcome.
- Combined datasets into a consolidated analytical DataFrame using Python.

Data Collection – SpaceX API

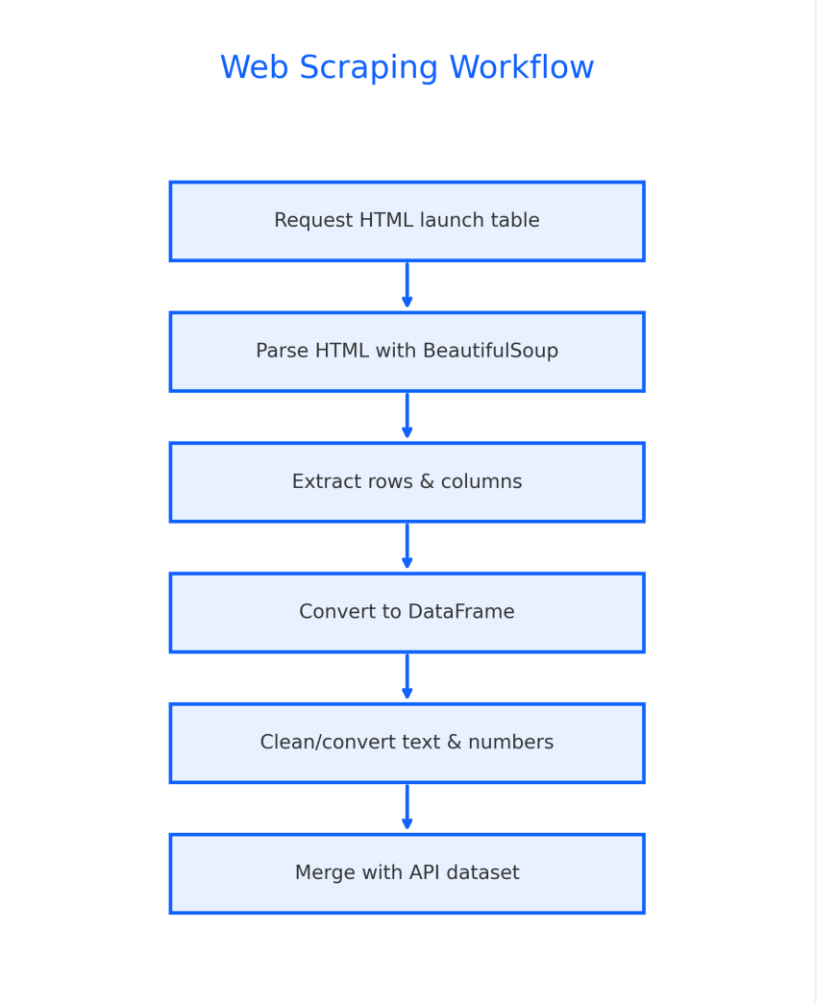
- Data Collection – SpaceX API
  - Retrieved launch metadata via SpaceX REST API using requests and JSON parsing.
  - Extracted fields: launch site, orbit, flight number, payload mass, booster version, and landing outcome.
  - Transformed API output into structured tables for further analysis.

- https://github.com/GyorfiCsaba/AppliedDataScienceCapstone



Data Collection - Scraping

- Data Collection – Scraping
  - Scraped launch information from a public SpaceX launch table using requests + BeautifulSoup.
  - Extracted complementary attributes not available in the API, such as detailed payload information and mission identifiers.
  - Merged scraped results with API dataset to resolve missing values and improve data completeness.
- https://github.com/GyorfiCsaba/AppliedDataScienceCapstone



Data Wrangling

- **Data Wrangling**
  - Cleaned and standardized raw SpaceX launch data.
  - Converted date formats, corrected inconsistent text fields, and handled missing payload values.
  - Constructed the target variable (class: 1 = success, 0 = failure).
  - Engineered additional features:
    - **Booster Version Category**
    - **Simplified Orbit Classification**
    - One-hot encoded categorical variables.
  - Prepared the final dataset for EDA, dashboarding, and machine learning.
- **GitHub URL:** https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

EDA with Data Visualization

- **EDA with Data Visualization**
- Several exploratory charts were created to understand relationships between launch features and landing outcomes.
  - **Scatter plots** were used to compare Flight Number, Payload Mass, and Orbit Type against Launch Site, revealing patterns in success rates.
  - **Bar charts** highlighted the success rate distribution across different orbit types.
  - **Line charts** showed trends in landing success rates across years.
  - **Correlation heatmaps** helped identify the most influential variables for predicting landing success.
- GitHub URL: https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

EDA with SQL

- **EDA with SQL**
- SQL queries were used to validate trends observed in the Python visual analysis:
  - Retrieved **unique launch sites**.
  - Filtered records using pattern matching (e.g., sites starting with CCA).
  - Computed **total payload mass** for specific customers (e.g., NASA).
  - Calculated **average payload mass** for selected booster versions.
  - Identified the **first successful ground landing date**.
  - Queried **success vs failure counts** across all missions.
  - Found boosters carrying the **maximum payload**.
  - Ranked landing outcomes across a defined date range (2010–2017).
    https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

Build an Interactive Map with Folium

- **Build an Interactive Map with Folium**
- Created an interactive Folium map to visualize:
  - **Launch site locations** using markers.
  - **Launch outcomes** using color-coded icons (green = success, red = failure).
  - **Distance calculations** from each launch site to nearby features such as coastline, roads, and railways.
- These map objects help visually analyze geographical influence on landing outcomes.
- Folium notebook: https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

Build a Dashboard with Plotly Dash

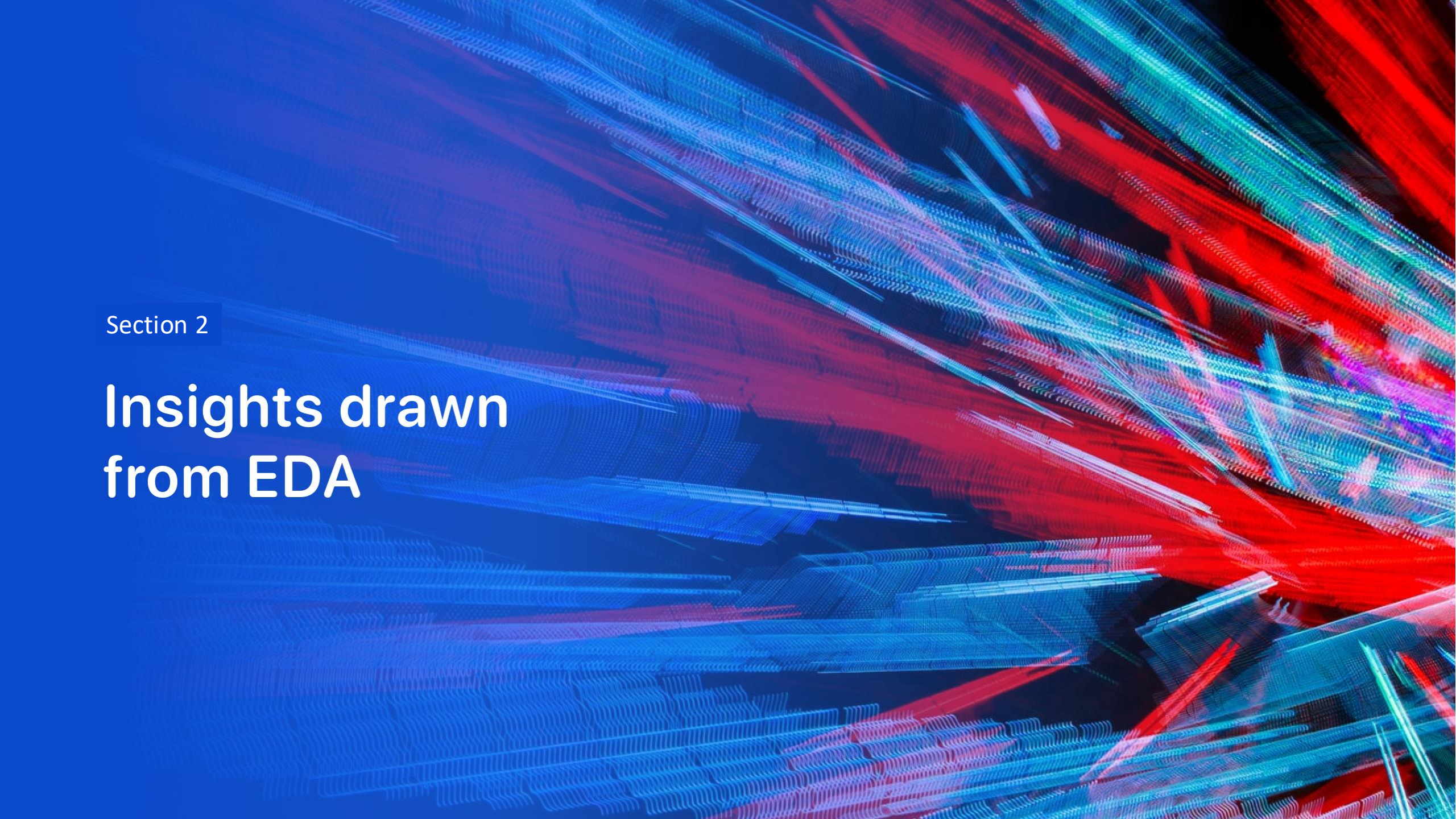
- **Build a Dashboard with Plotly Dash**
- Added interactive components:
  - **Launch Site Dropdown** to filter results by site or view all sites.
  - **Payload Range Slider** to dynamically update mission results by selected payload interval.
- Plots included in the dashboard:
  - **Pie chart** showing success counts (overall or per site).
  - **Scatter plot** comparing Payload vs. Landing Outcome with Booster Version color coding.
- These interactions let users explore patterns and derive insights from mission data.
- Dash notebook: https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

Predictive Analysis (Classification)

- **Predictive Analysis (Classification)**
- Built several machine learning classification models to predict first-stage landing success.
- Steps performed:
  - Preprocessed dataset (scaling, encoding).
  - Split data into training and test sets.
  - Trained Logistic Regression, SVM, KNN, and Decision Tree models.
  - Tuned hyperparameters using GridSearchCV.
  - Selected best-performing model based on cross-validation accuracy.
- **Best model:** Decision Tree
  - Cross-validation accuracy ≈ **0.88**
  - Test set accuracy ≈ **0.72**
- ML notebook: https://github.com/GyorfiCsaba/AppliedDataScienceCapstone

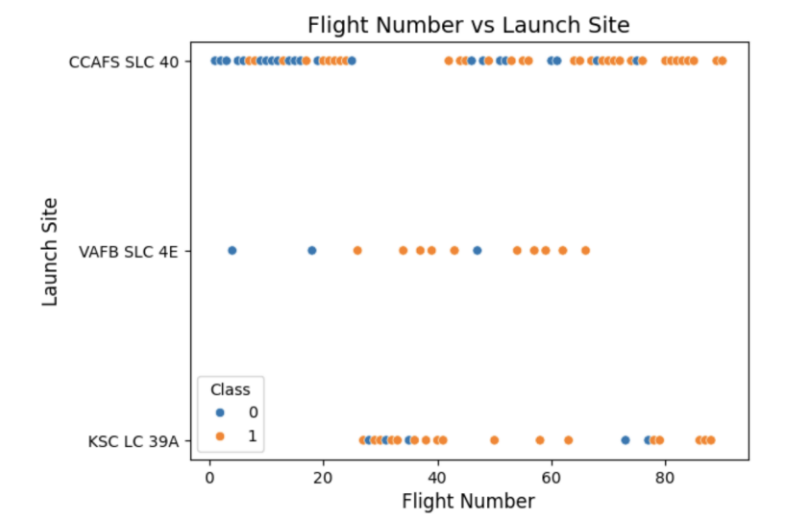
Results

- Exploratory Data Analysis:
  - Identified relationships between payload, orbit type, booster category, and landing outcomes.
- Interactive analytics (screenshots to be added):
  - Folium map results (launch site geography).
  - Dash dashboard output (success distribution, payload patterns).
- Predictive Analysis:
  - Evaluated multiple ML models.
  - Selected Decision Tree as the most accurate and interpretable classifier.


Section 2
Insights drawn from EDA

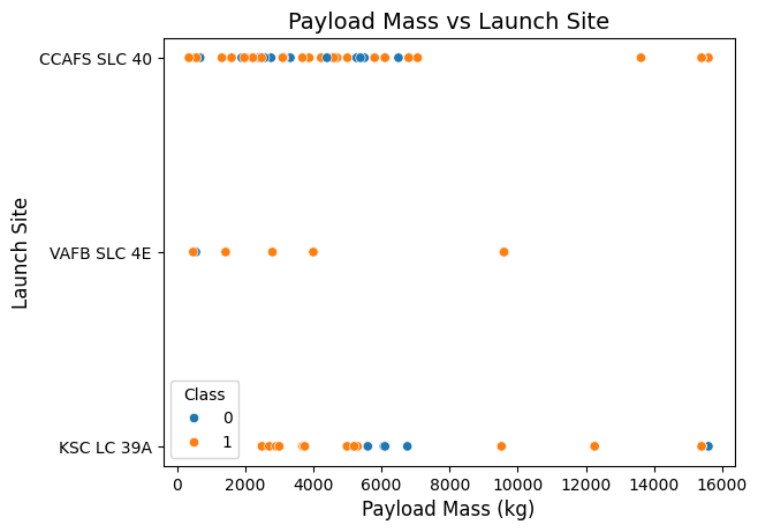
Flight Number vs. Launch Site

- Flight Number vs. Launch Site
  - This scatter plot visualizes how launch experience (Flight Number) varies across SpaceX launch sites.
  - Higher flight numbers generally correspond to later missions, which tend to show improved landing success due to engineering advancements.
  - Differences across sites reflect how frequently each site is used and how mission types vary.



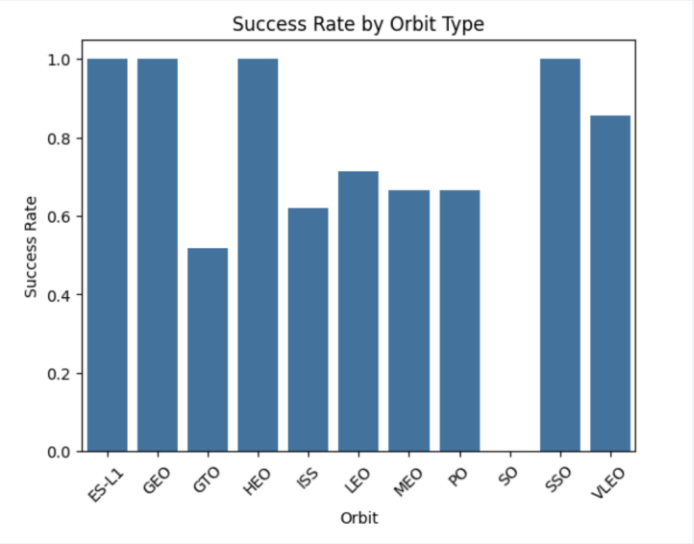
Payload vs. Launch Site

- This plot compares Payload Mass (kg) across launch sites.
- Some sites handle significantly heavier payloads, which impacts mission complexity and landing outcomes.
- Notable clusters indicate preferred launch sites for different payload categorie



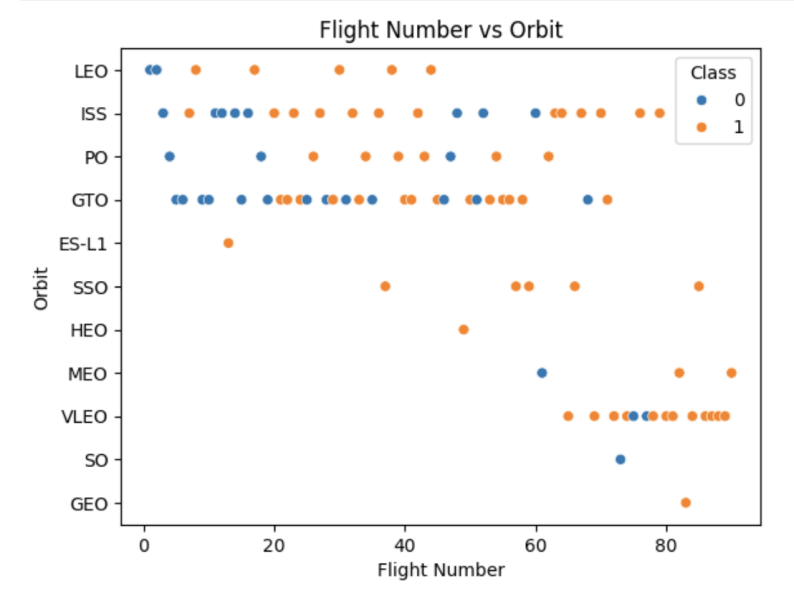
Success Rate vs. Orbit Type

- A bar chart shows how landing success probability varies by orbit type.
- Lower-energy orbits (LEO) typically have higher success rates, while high-energy orbits like GTO show lower success.
- Understanding orbit-related performance helps refine mission risk assessments.



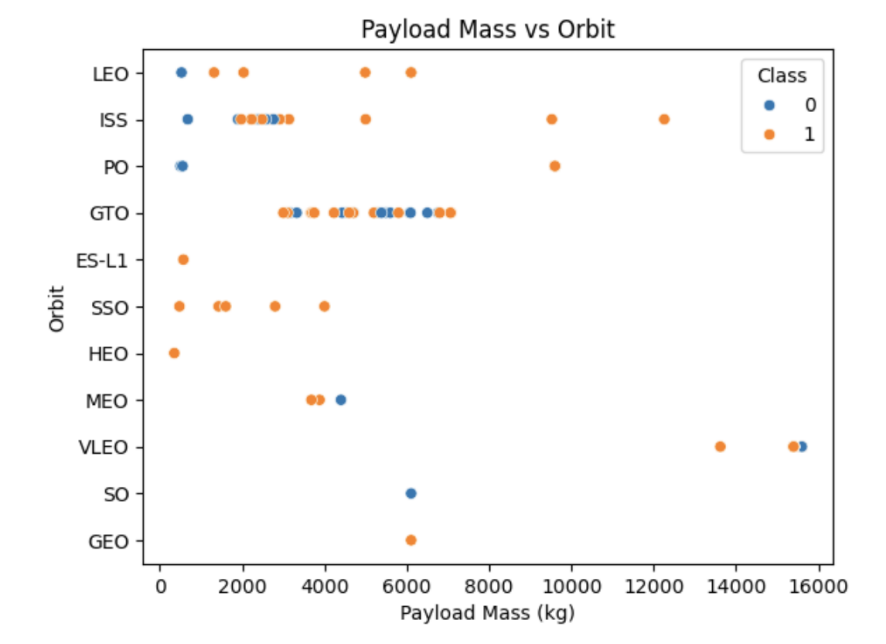
Flight Number vs. Orbit Type

- This scatter plot compares Flight Number with Orbit Type, helping to identify how mission profiles evolved over time.
- Higher flight numbers represent more recent missions, where improvements in booster design and operations are reflected in better landing outcomes.
- The distribution across orbit types shows which orbits became more common as SpaceX expanded its launch capabilities.



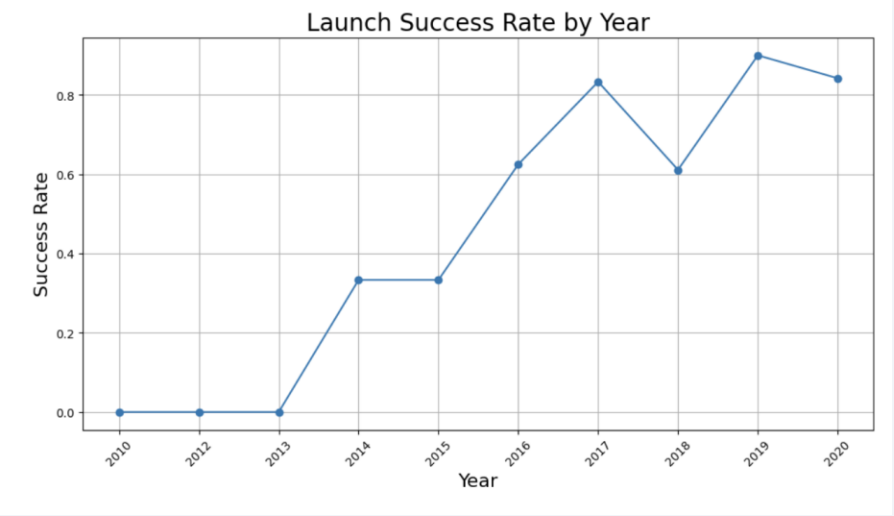
Payload vs. Orbit Type

- This plot illustrates how payload mass (kg) varies across orbits such as LEO, MEO, GEO, GTO, and SSO.
- Heavier payloads tend to be associated with more energy-intensive orbits, influencing mission difficulty and landing success.
- Visual clustering helps identify which orbits typically carry higher or lower payload categories.



Launch Success Yearly Trend

- A line chart was created to show the average landing success rate per year.
- The upward trend reflects SpaceX's rapid technological improvements in booster reusability.
- Early years show inconsistent outcomes; later years exhibit strong, stable success rates, aligning with the introduction of improved booster versions like F9 FT and Block 5.



All Launch Site Names

- SQL query retrieved all unique launch site names from the dataset.
- This confirms the presence of the four official SpaceX sites used during the analyzed timeframe:
  - KSC LC-39A
  - CCAFS LC-40
  - VAFB SLC-4E
  - CCAFS SLC-41 (depending on dataset version)
- These names form the basis for site-level analysis in later slides.

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

```
 * 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 query used pattern matching (LIKE 'CCA%') to filter launch sites whose names begin with "CCA".
- This returns Cape Canaveral–related pads, e.g.:
- CCAFS LC-40
- CCAFS SLC-41
- This demonstrates use of SQL string filtering in exploratory analysis.

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

```
 * 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

Total Payload Mass

- omputed the total payload mass carried by boosters for NASA missions.
- This helps quantify NASA's utilization of Falcon 9 platform and total delivered mass to orbit.
- The result represents cumulative payload over all NASA-related launches in the dataset.

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS total_payload
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

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

total_payload
45596

Average Payload Mass by F9 v1.1

- SQL query calculated the average payload mass carried by booster version Falcon 9 v1.1.
- This provides insight into the performance envelope of earlier Falcon 9 versions.
- Comparing averages across booster types helps understand engineering improvements over time.

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

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

avg_payload
2928.4

First Successful Ground Landing Date

- Retrieved the earliest date where the landing outcome was Success (ground pad).
- This marks a historical milestone for SpaceX, as successful ground landings enabled rapid refurbishment and reuse.
- The date confirms when reusability first became operationally reliable.

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

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

first_success_ground_pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- QL query filtered missions that:
- Landed successfully on a **drone ship**, and
- Carried **payload mass between 4000 and 6000 kg**.
- These missions demonstrate booster performance under high-payload, offshore recovery profiles.
- Result includes the booster IDs that met these criteria.

```
%%sql
SELECT DISTINCT "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.
```

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

Total Number of Successful and Failure Mission Outcomes

- SQL aggregation counted total successful vs failed landing outcomes across all missions.
- Provides a clear summary of Falcon 9 landing reliability.
- The ratio between successes and failures shows substantial improvement over the years.

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

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

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

Boosters Carried Maximum Payload

- SQL query was used to identify the booster(s) associated with the highest payload mass.
- This result highlights the missions that pushed Falcon 9's lift capacity to its limits.
- Knowing which boosters handled the maximum payload helps understand performance differences between older and newer booster versions.

```sql
%%sql
SELECT DISTINCT "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

2015 Launch Records

- Retrieved launch records from year 2015 where the landing outcome was a failure on drone ship.
- Query outputs include:
  - Booster version
  - Launch site
  - Landing outcome
- These failures represent the early phase of SpaceX's ocean-based recovery attempts, before the technique became more reliable.
- here

```
%%sql
SELECT substr("Date", 6, 2) AS month,
       "Landing_Outcome",
       "Booster_Version",
       "Launch_Site"
FROM SPACEXTABLE
WHERE substr("Date", 1, 4) = '2015'
  AND "Landing_Outcome" LIKE 'Failure (drone ship)%';
```

```
 * 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

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

- SQL query ranked landing outcomes in descending order based on frequency within the selected date range.
- Shows how often each outcome occurred, such as:
  - Success (ground pad)
  - Success (drone ship)
  - Failure (drone ship)
  - Failure (ground pad)
- This ranking helps visualize SpaceX's rapid improvements in booster recovery during this period.

```
%%sql
SELECT "Landing_Outcome",
       COUNT(*) AS outcome_count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_count DESC;
```

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

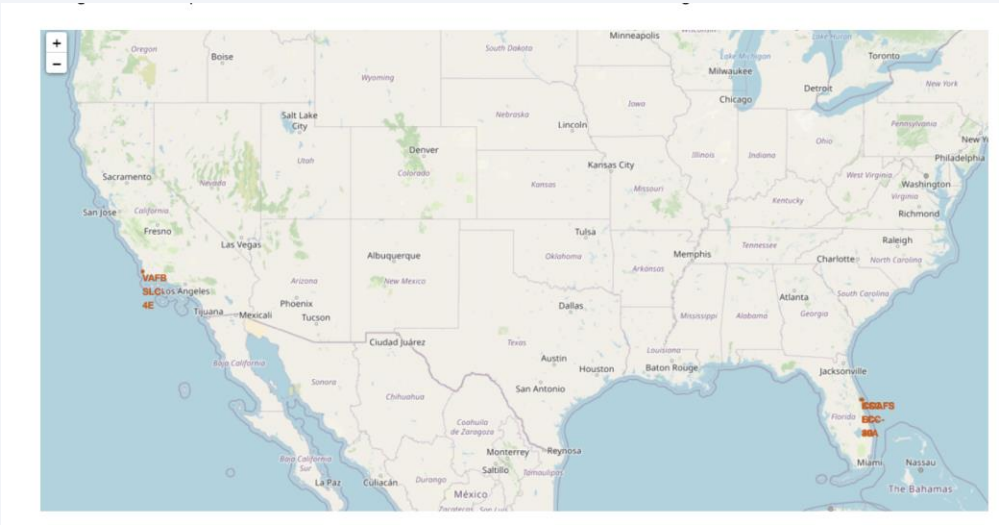
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Global Launch Site Map

- This Folium map displays **all SpaceX launch sites worldwide**, marked with interactive location pins.
- It provides a geographical overview of where Falcon 9 missions originate.
- Marker tooltips include the site name and coordinates, allowing users to explore spatial context easily.



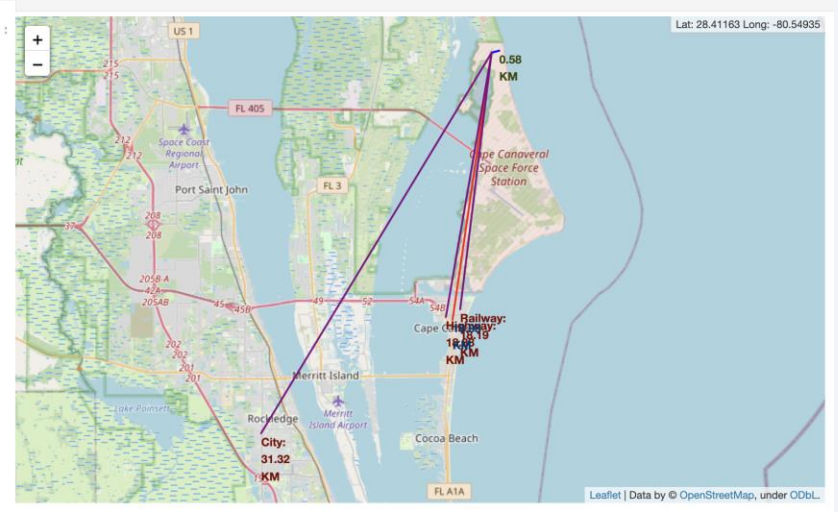
Launch Outcomes by Site (Color-Coded Markers)

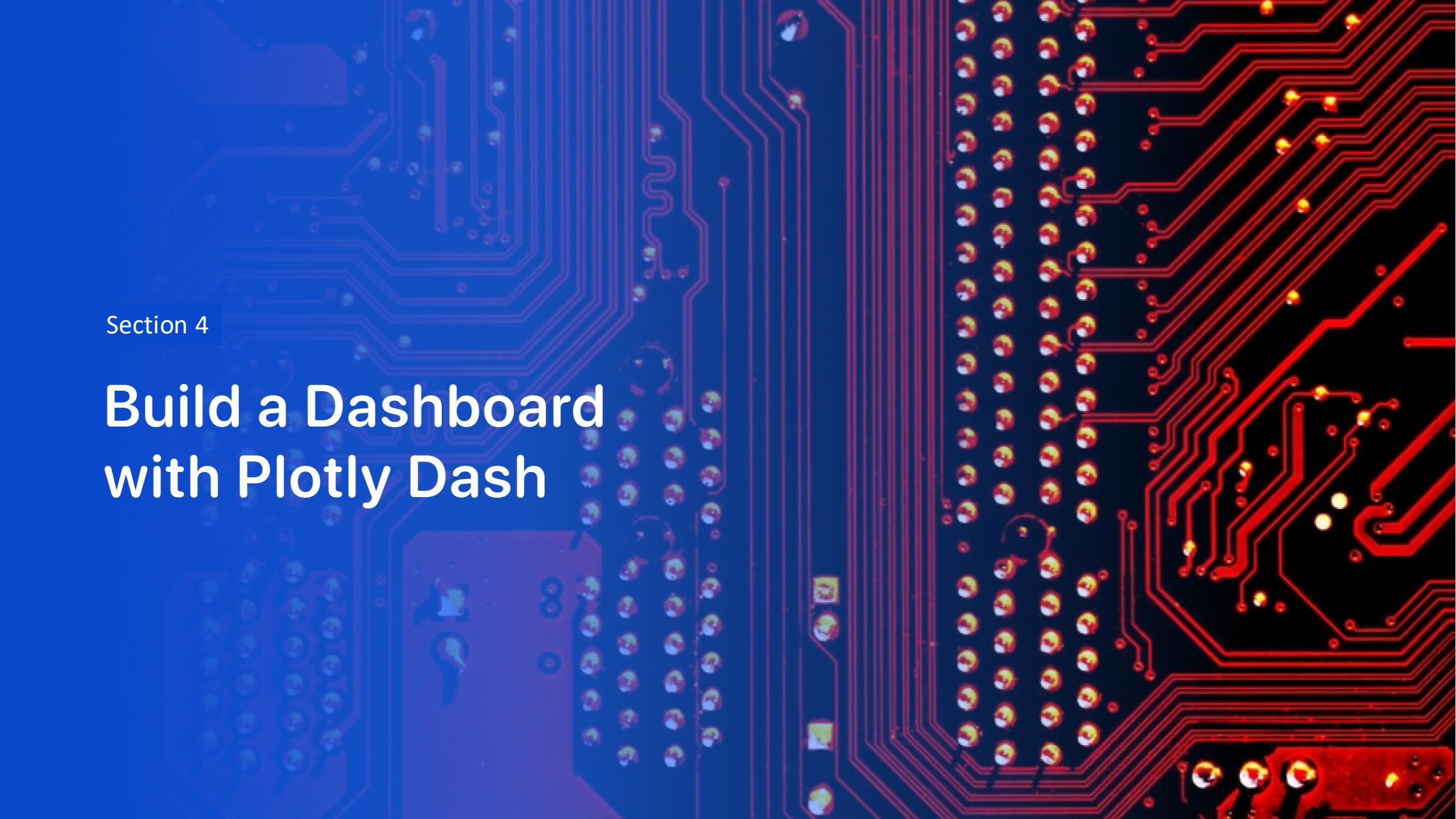
- This Folium map visualizes landing outcomes using **distinct colors**:
  - Green markers → Successful landings
  - Red markers → Failed landings
- The spatial pattern reveals which launch sites have the highest success density.
- This helps identify operational differences across locations.

- Explain the important elements and findings on the screenshot

Proximity Analysis of Launch Site Infrastructure

- This screenshot focuses on a selected launch site and shows:
- Nearby coastline
- Highways, railroads, or logistic routes
- Calculated distances from the site to the closest infrastructure
- This analysis is useful to understand safety margins, logistics, and site selection rationale.




Section 4
Build a Dashboard with Plotly Dash

Launch Success Counts Across All Sites

- Pie chart shows the total number of successful launches for each launch site when “All Sites” is selected.
- This gives an immediate understanding of which sites contribute most to overall mission success.
- Differences in slice size reflect varying launch frequencies and performance history.

Success Ratio for Top-Performing Launch Site

When a specific launch site is selected in the dashboard, the pie chart updates to show its success vs failure ratio.

This view highlights which site has the highest success rate, typically KSC LC-39A for this dataset.

This breakdown helps compare operational reliability across sites.

Payload vs. Launch Outcome Scatter (All Sites)

- This scatter plot visualizes Payload Mass (kg) versus Launch Outcome (class) for all launch sites.
- The range slider allows selecting different payload intervals, revealing success patterns across payload categories.
- Key insights from this chart typically show:
- Mid-range payloads (≈2000–5000 kg) exhibit the highest success rates.
- Extremely low or high payloads often correlate with more failures.
- Booster version categories (colored groups) highlight which generations perform best.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- A bar chart was generated to compare the accuracy of all trained classification models:
- Logistic Regression
- SVM
- K-Nearest Neighbors
- **Decision Tree (best performer)**
- The Decision Tree achieved the **highest cross-validation accuracy (~0.88)**.
- On the test set, the Decision Tree model reached an accuracy of **~0.72**, showing strong generalization for this dataset.

- Find which model has the highest classification accuracy

Confusion Matrix

- The confusion matrix of the **Decision Tree** model shows its performance on the test set:
- **True Positives (TP):** Correctly predicted successful landings
- **True Negatives (TN):** Correctly predicted failed landings
- **False Positives (FP):** Predicted success where actual outcome was failure
- **False Negatives (FN):** Predicted failure where actual outcome was success
- Interpretation:
- The model correctly identifies most successful landings.
- Misclassifications occur mostly in borderline cases with unusual payloads or rare orbits.

Conclusions

- Successfully built a complete end-to-end data science workflow using real SpaceX launch data.
- EDA revealed key factors influencing landing success:
- Launch site
- Booster version category
- Payload mass
- Orbit type
- Interactive analytics (Folium + Dash) provided deeper insights into spatial and operational patterns.
- Machine learning models were developed and evaluated, with the Decision Tree emerging as the best performer.
- Predictive modeling demonstrates how data-driven tools can support launch planning, cost estimation, and risk assessment.

Appendix

- dditional resources and artifacts from this project may include:
  - Python code snippets
  - SQL queries used in EDA
  - DataFrames, tables, and extended outputs
  - Extra diagnostic plots
  - Notebook references and intermediate results
- Full project repository:
- https://github.com/GyorfiCsaba/AppliedDataScienceCapstonet

Thank you!