IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christabel Ekeocha
August, 2025

Table of Contents

Executive Summary

- SpaceX is revolutionizing space travel by making it accessible due to relatively inexpensive launches. SpaceY is a new commercial rocket launch provider who wants to compete against SpaceX
- This project uses SpaceX API and web scraping to gather rocket launch data and predict whether first stages will land and be reused, cutting costs.
  - Data wrangling and cleaning was performed in Python.
  - Exploratory Data Analysis conducted with visualization and SQL.
  - Interactive analysis was developed with Folium and Plotly Dash.
  - Predictive analysis applied using classification models (Decision tree, KNN, SVM, Logistic regression).
- Given mission parameters such as payload mass and desired orbit, the models deployed in this report were able to predict the first stage rocket booster landing successfully with an accuracy level up to 87.8%.

GitHub: https://github.com/christabelek/Capstone

Introduction

- SpaceX has transformed the aerospace industry with reusable rockets and if our company, SpaceY can model landing outcomes, we can make further impact in the industry with even lower cost launches.

- The rocket ‘first stage is the the which is the most expensive part to build. SpaceX cuts costs by reusing this stage, and so finding variables that guarantee first stage recovery is a direct path to reducing expenditure.

Problem: Can we predict successful landings based on payload, orbit, and other variables?

Goal: Apply data science techniques to analyze launch data and build models that can predict outcomes and therefore prospective costs.

Section 1

Methodology

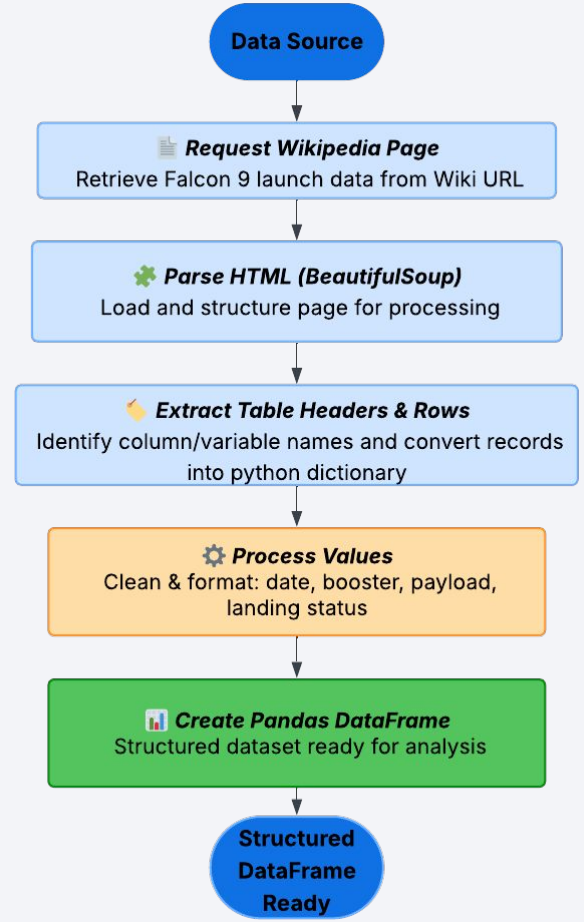
Methodology

Outline

- Data collection methodology:
  - SpaceX REST API calls
  - Web scraping Wikipedia launch records
- Perform data wrangling
  - Cleaning nulls, standardizing column names
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Fitted and analysed data Logistic Regression, SVM, Decision Tree and KNN models

Methodology - Data Collection with Web Scraping

- Falcon 9 launch records were scraped from Wikipedia page '[List of Falcon 9 and Falcon Heavy Launches]'
  - Requested the Falcon 9 Launch Wiki page from its Wikipedia URL
  - Used BeautifulSoup object functionality to effectively parse the script
  - Extracted all column/variable names from the HTML table header using defined function
  - Further parsed the HTML scripted table, loading values into a python dictionary
    - Used IBM provided functions to process values like date, booster version, landing status mass from web scraped table
  - Converted the dictionary into a Pandas dataframe for further analysis



[Github URL]

Methodology - Data Wrangling

- The data was explored and processed using appropriate python libraries in preparation for analysis

  - Created categorical variables for outcome success/failure

    - The training label: 'Class' was created based on the values in column, ‘Landing Outcome

Class = 0; first stage booster did not land successfully.
Values in ‘Landing Outcome’:

- None None; not attempted
- None ASDS; unable to be attempted due to launch failure
- False ASDS; *successfully landed to a drone ship*
- False Ocean; *successfully landed to the ocean*
- False RTLS; *unsuccessfully landed to a ground pad*

Class = 1; first stage booster landed successfully.
Values in ‘Landing Outcome’:

- True ASDS; *successfully landed on a drone ship*
- True RTLS; *successfully landed to a ground pad*
- True Ocean; *successfully landed in the ocean*

Github URL

Methodology - EDA with Data Visualization

- Used Matplotlib and Seaborn visualization libraries to plot and investigate various relationships:
  - Generated scatter plots to Visualize the relationship between
    - Flight Number and Launch Site,
    - Payload and Launch Site,
    - Flight Number and Orbit type,
    - Payload and Orbit type.
  - Used Bar chart to Visualize the relationship between success rate of each orbit type.
  - Line plot to Visualize the launch success yearly trend.

Github URL

Methodology - EDA with SQL

- SQL queries were run on an IBM DB2 instance to explore the data set further and identify the following:
  - **Unique launch sites** – Show all distinct launch site names in the missions.
  - **Launch sites starting with 'CCA'** – Display 5 records of sites whose names begin with "CCA".
  - **Total NASA (CRS) payload mass** – Calculate the sum of payload mass carried by NASA (CRS) boosters.
  - **Average payload for F9 v1.1** – Compute the mean payload mass for booster version *F9 v1.1*.
  - **First successful ground pad landing** – Find the earliest date a landing was successful on a ground pad.
  - **Boosters with specific payloads on drone ship** – List booster versions with successful drone ship landings carrying payloads between 4000 and 6000 kg.
  - **Mission outcomes count** – Count how many missions ended in success vs. failure.
  - **Boosters with maximum payload** – Identify booster versions that carried the heaviest payload mass.

Github URL

Methodology - Building an Interactive Map with Folium

- A map was created through Folium library to visualize data with geographical context
  - **Added launch site markers** to visualize spread and clusters
  - **Color-coded success/failure** to see whether certain areas were prone to be more or less successful
  - **Measured distances to highways/coastlines** to further conceptualize placement methodology

Github URL

Methodology - Building a Dashboard with Plotly Dash

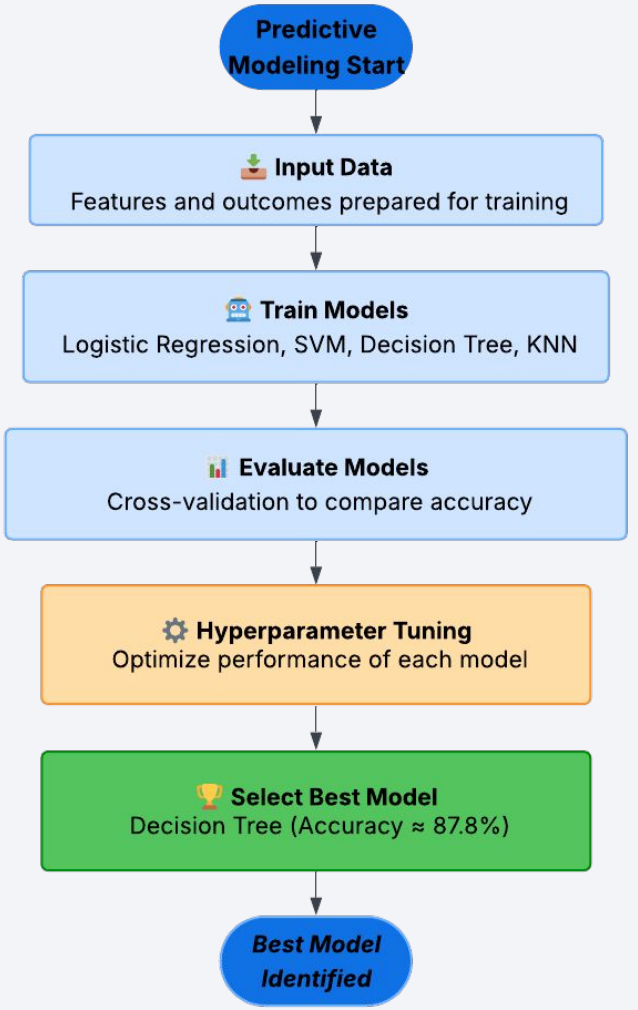
- Developed a **Plotly dashboard** for interactive data visualization
  - **Pie chart** – Launch success counts by site
  - **Scatter plot** – Payload vs. outcome with adjustable slider
  - **Interactive filters** – Payload range and booster version

- Enables quick exploration of distributions under different conditions (e.g., varying payloads or boosters)

Github URL

Methodology - Predictive Analysis (Classification)

- **Model Development** – Trained Logistic Regression, SVM, Decision Tree, and KNN
- **Evaluation** – Applied cross-validation to compare performance
- **Improvement** – Performed hyperparameter tuning for optimization
- **Best Model** – Decision Tree selected (Accuracy ≈ 87.8%)



[Github URL]

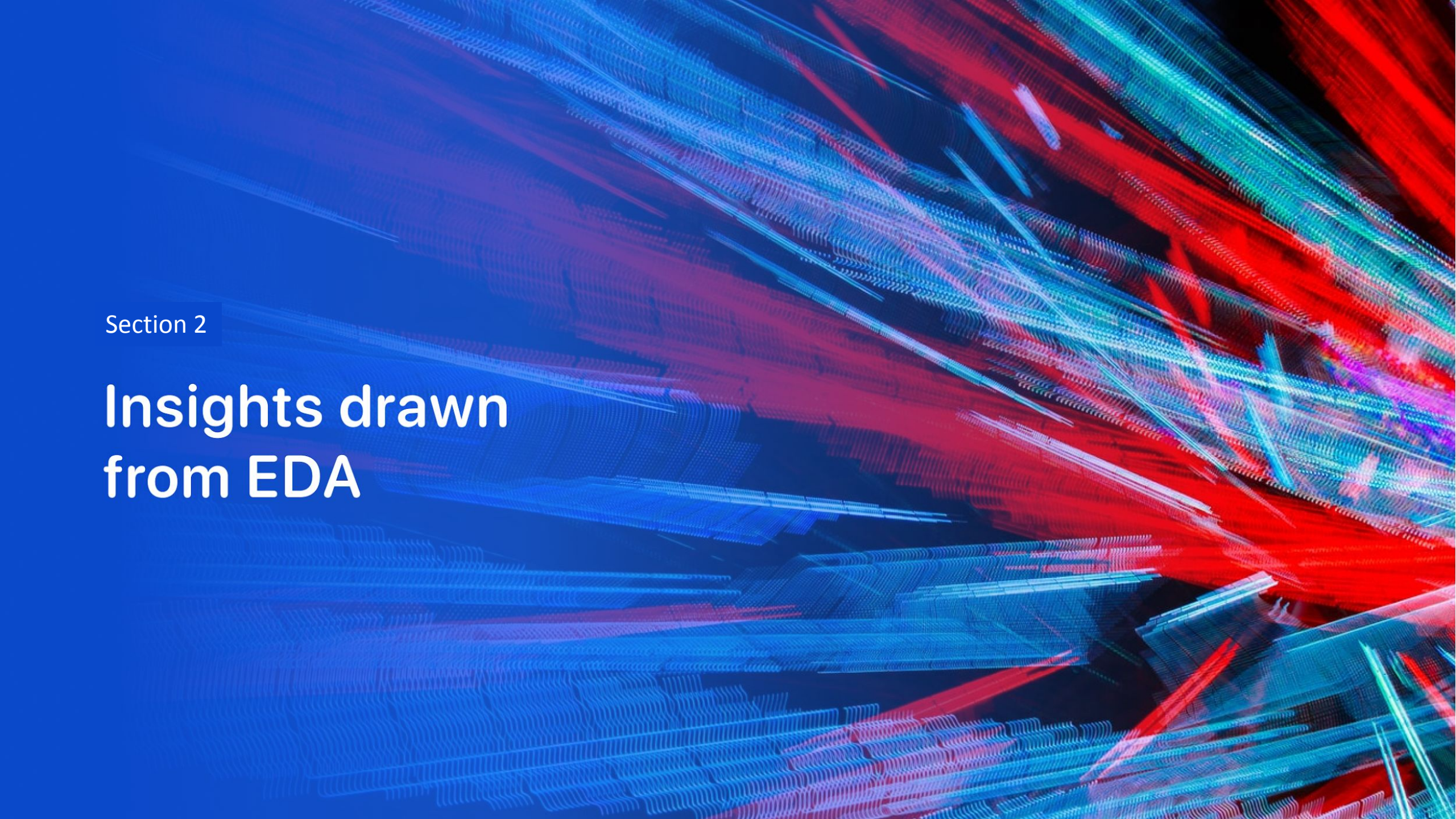
Results

Outline

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

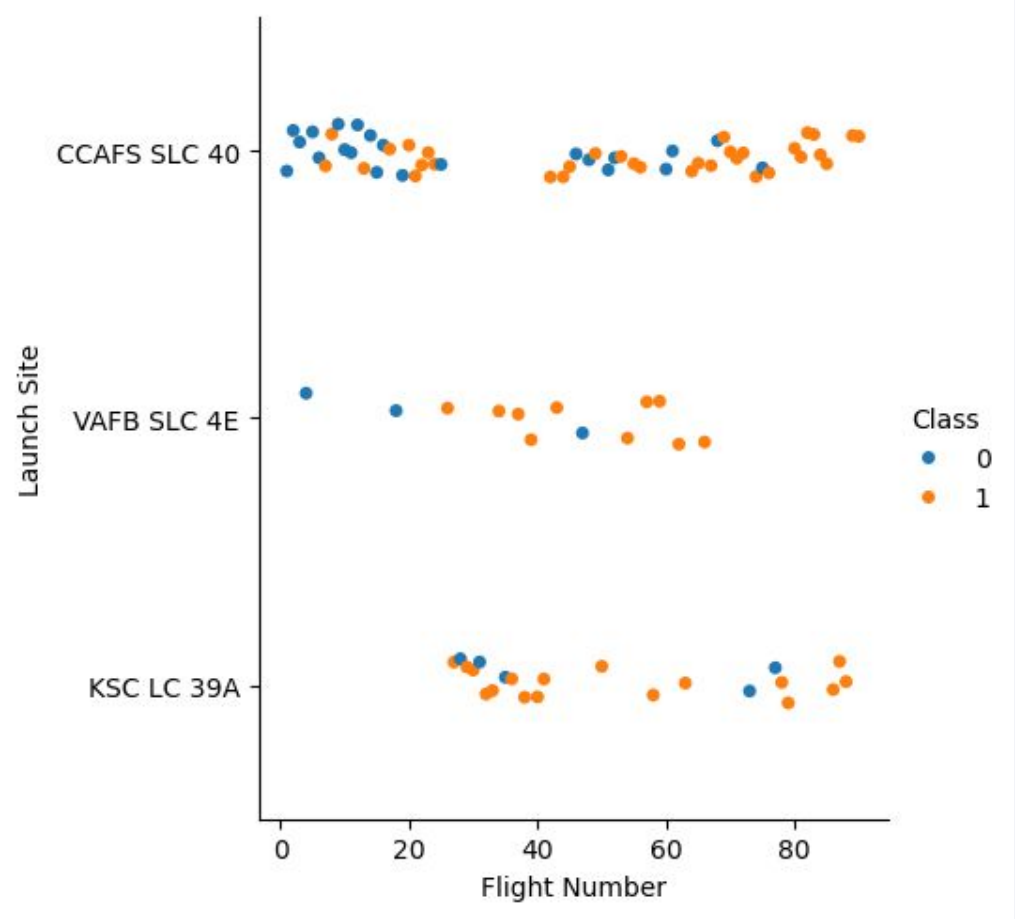
Section 2

Insights drawn from EDA



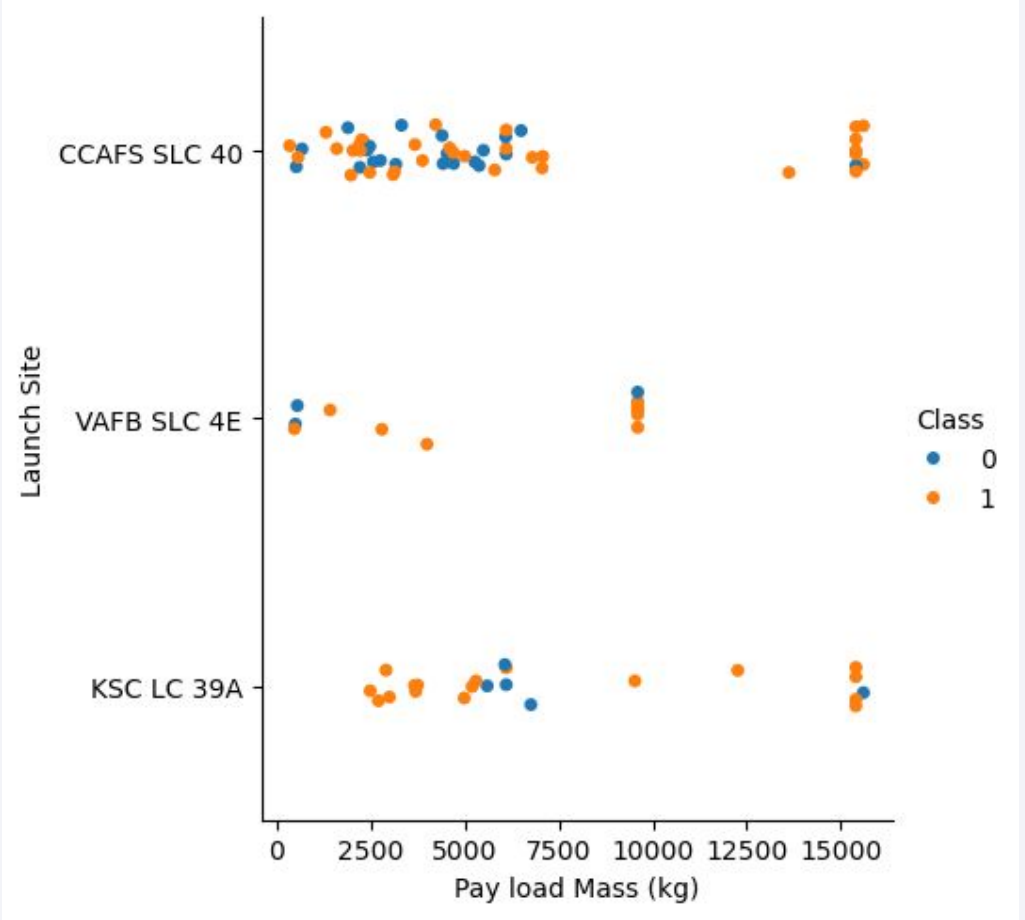
Flight Number vs. Launch Site

- Each point represents a **single launch**
- Multiple sites are shown on the y-axis → allows comparison across locations
- Distribution reveals **how many flights occurred at each site**
  - Most launches occurred at Cape Canaveral (CCAFS SLC 40)
  - Vandenberg (VAFB SLC 4E) was used sparingly for missions
  - Kennedy Space Center (KSC LC 39A) became active in later flights
- Overall mission success improved over time across all sites.



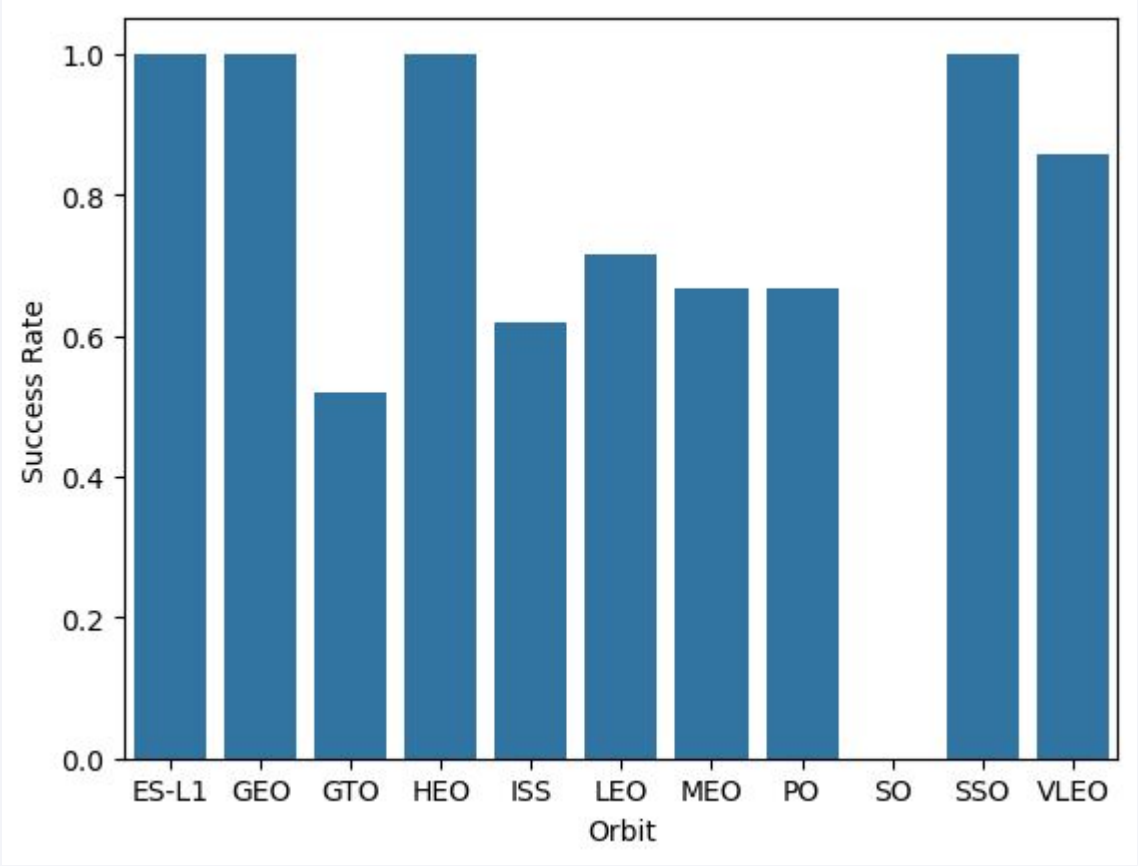
Payload vs. Launch Site

- Each point represents a **single launch with its payload mass**
- Allows comparison of **payload capacity handled by different sites**
- Clusters reveal which sites typically handle **heavier vs. lighter payloads**
  - We find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)
  - Cape Canaveral (SLC 40) manages the broadest range of payloads,
  - Kennedy (LC 39A) is key for very heavy missions, Vandenberg (SLC 4E) supports mid-weight specialized launches.
- Success rates improve with higher payload missions.



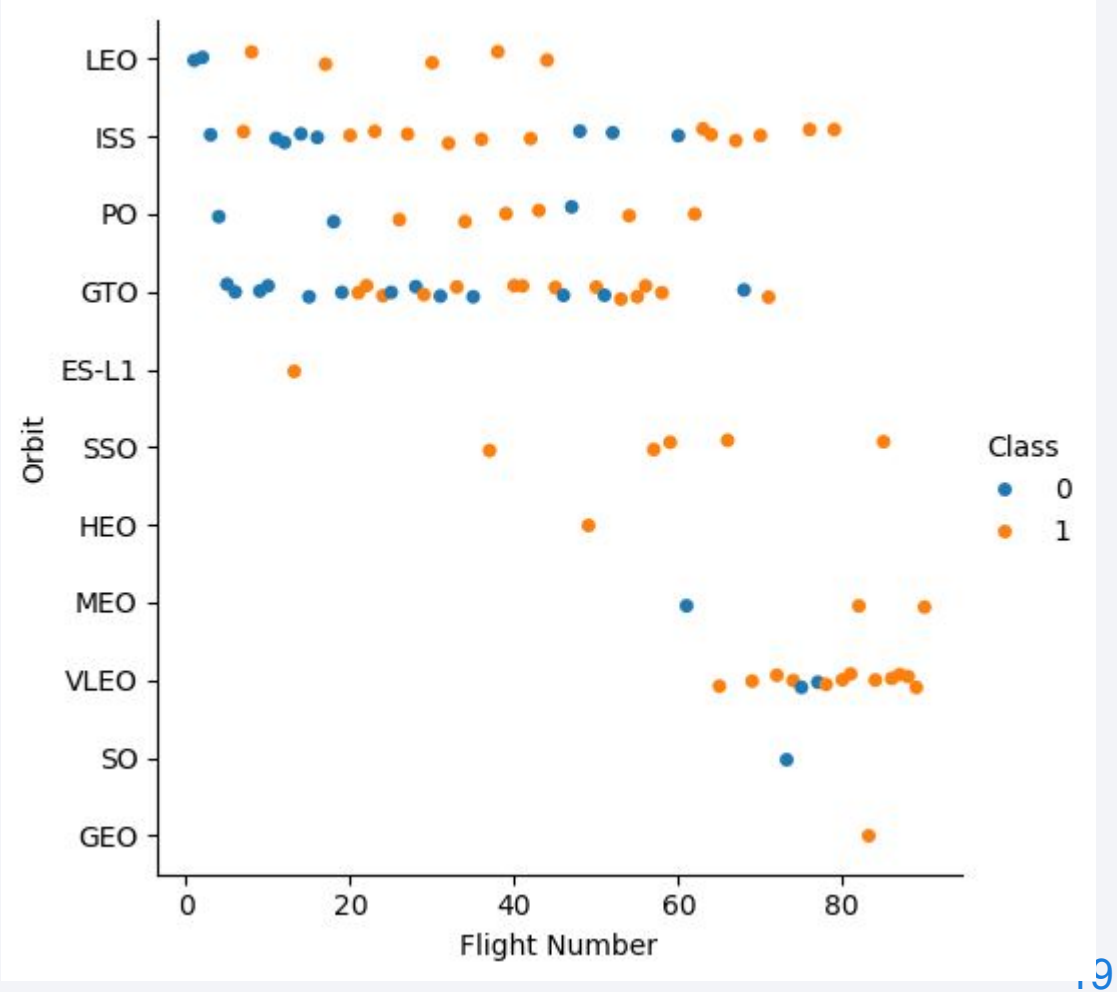
Success Rate vs. Orbit Type

- Each bar shows the **proportion of successful launches** for that orbit.
- Easier to compare **reliability across orbit types** at a glance.
- Some orbits have **consistently higher success rates**, suggesting greater operational maturity.
  - **Highest success rates** – GEO, ES-L1, HEO, and SSO missions achieved nearly 100% reliability.
  - **Moderate performance** – LEO, MEO, PO, VLEO, and ISS had mixed outcomes (≈65–85%).
  - **Lowest success rate** – GTO missions were least reliable (~50%), highlighting their complexity.



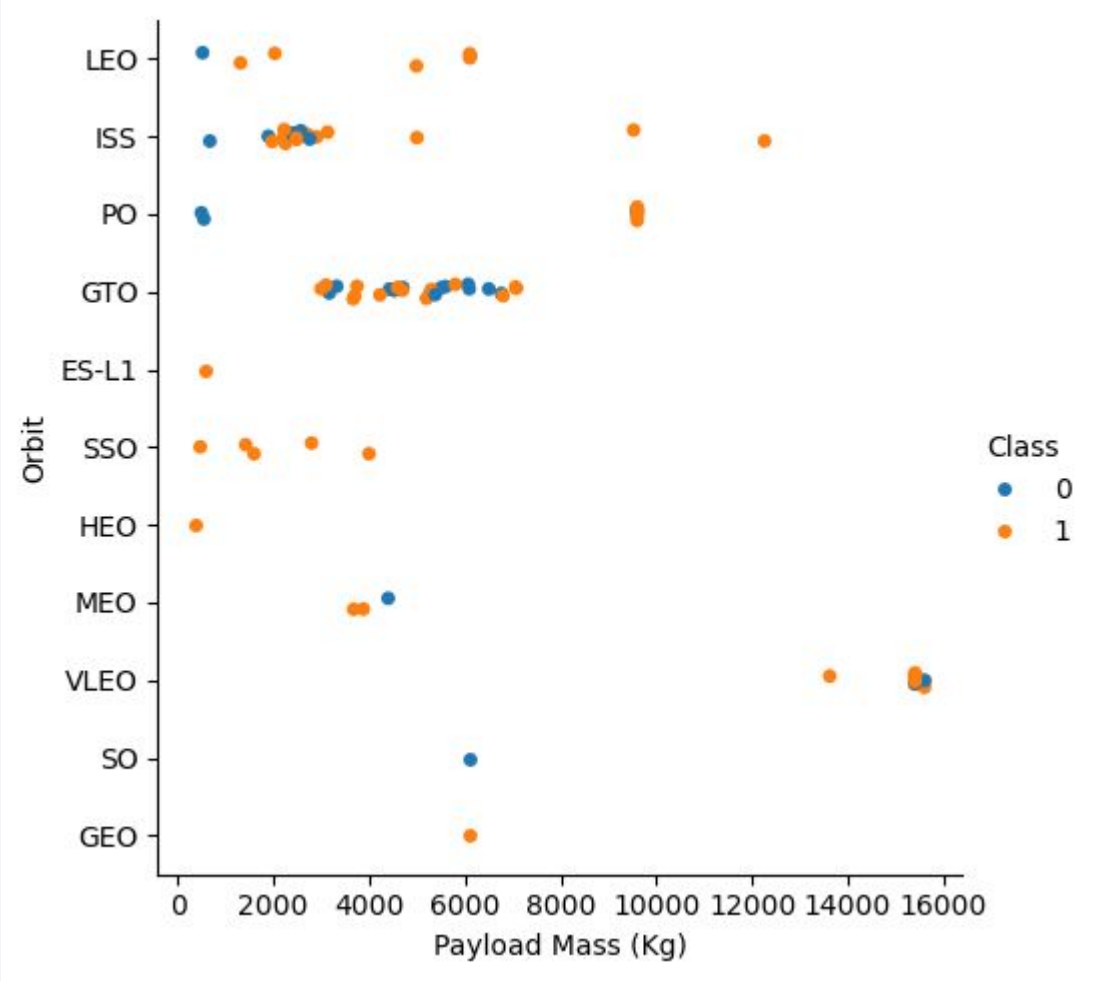
Flight Number vs. Orbit Type

- Each point represents a **single launch**.
- Shows how **different orbit types** were used across flight history.
- Distribution reveals **when new orbit types were introduced**, assuming flight numbers are assigned temporally
  - Newer orbit types (HEO, VLEO, MEO) appear later, with overall success improving over time.
- LEO, ISS, and PO dominate launch history as the most frequent mission orbits.
- GTO missions show the **highest rate of failure** compared to other orbits.



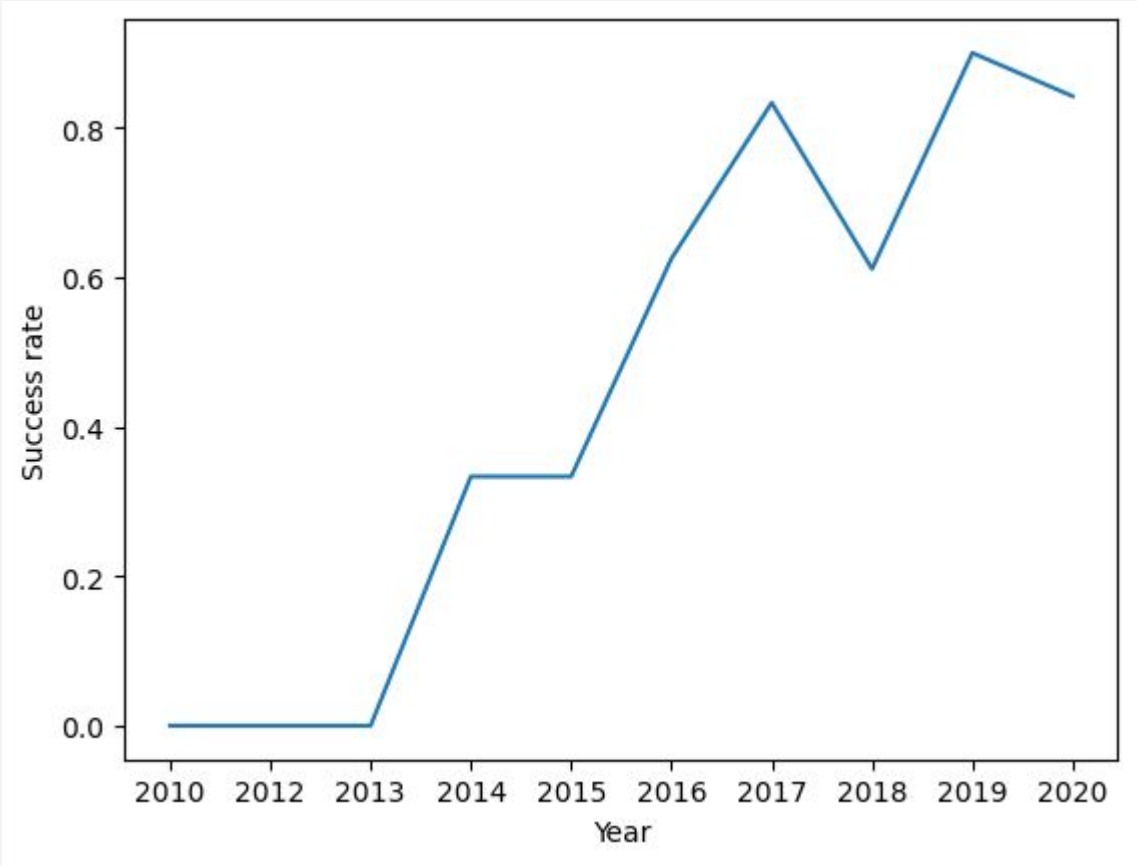
Payload vs. Orbit Type

- **Most launches** target LEO, ISS, and GTO, covering a wide payload range.
- **Heavy payload missions (≥10,000 kg)** were mostly successful, especially in LEO and ISS.
- **Failures** occurred more often in the **low-to-mid payload range**, notably in GTO and LEO.



Launch Success Yearly Trend

- Success rates were **near zero before 2014**, reflecting early testing.
- **Steady improvement** after 2014, with rates exceeding 80% by 2017.
- Despite minor dips, overall trend shows **consistent operational maturity**.



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

sum(PAYLOAD_MASS__KG_)
619967

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

avg(PAYLOAD_MASS__KG_)
6138.287128712871

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

```
%sql select count(Mission_Outcome) from SPACEXTABLE where Mission_Outcome like 'Success%'
```

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

count(Mission_Outcome)
100

```
%sql select count(Mission_Outcome) from SPACEXTABLE where Mission_Outcome like 'Failure%'
```

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

count(Mission_Outcome)
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) betw and 2017-03-20, in descending order

- Present your query result wi re

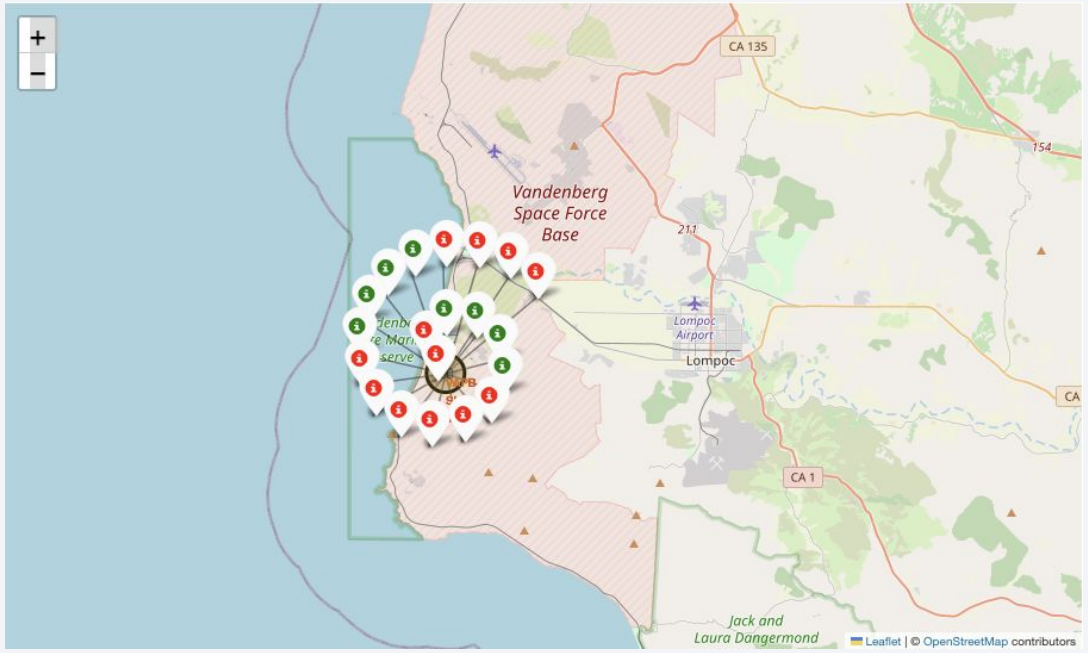
Landing_Outcome	LandingOutcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Global Map of SpaceX Launch Sites

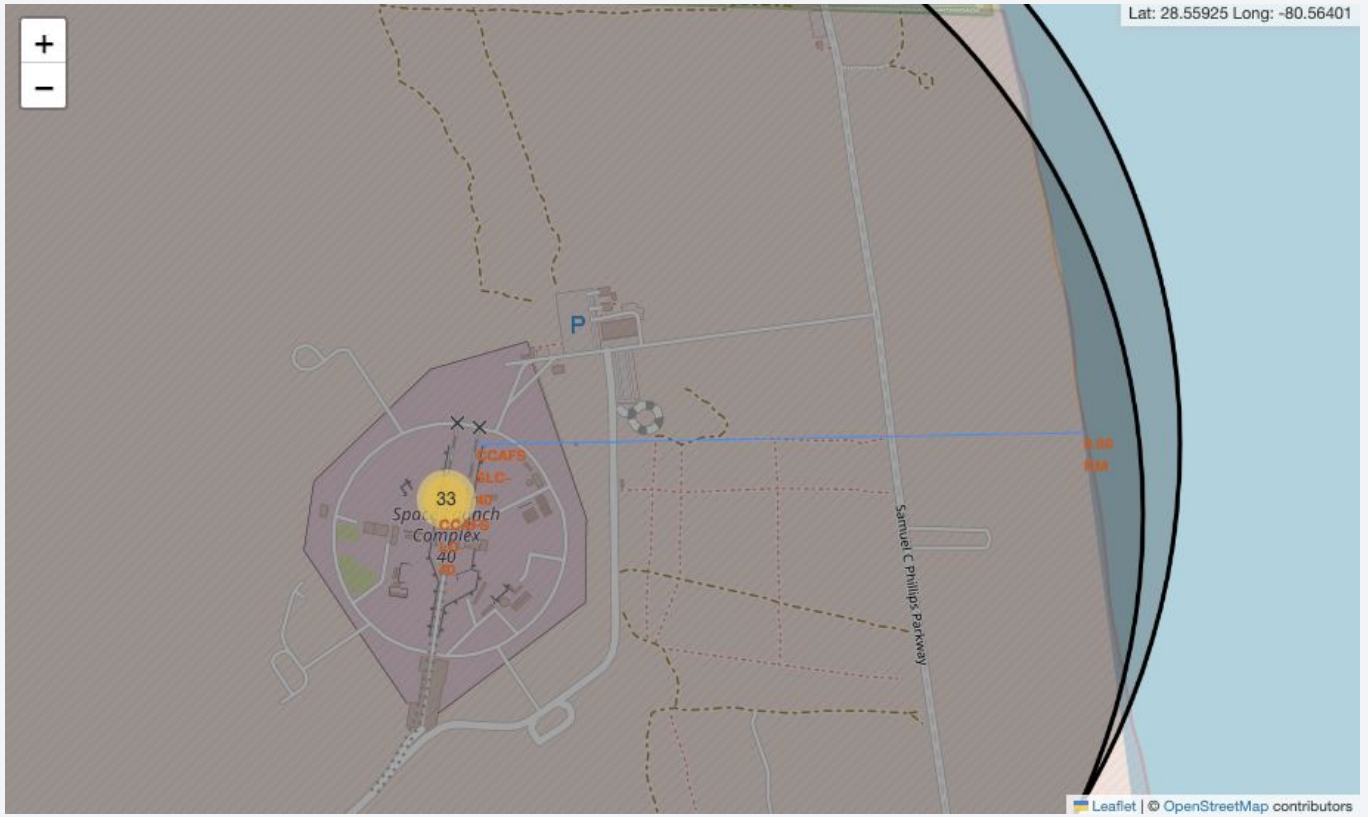
- The map displays **all SpaceX launch sites worldwide**, visualized using folium with interactive markers.
- Launch sites are concentrated in the **United States**, primarily along the **East Coast (Florida)** and the **West Coast (California)**.
- Florida hosts the majority of launches (e.g., **KSC LC-39A** and **CCAFS LC-40**), making it SpaceX's central hub for missions.
- California's **VAFB SLC-4E** supports polar orbit launches, highlighting its role in specific mission profiles.
- The clustering of markers emphasizes that **SpaceX operations are US-centric**, with no active international launch sites currently represented.

Mapping Launch Success and Failure

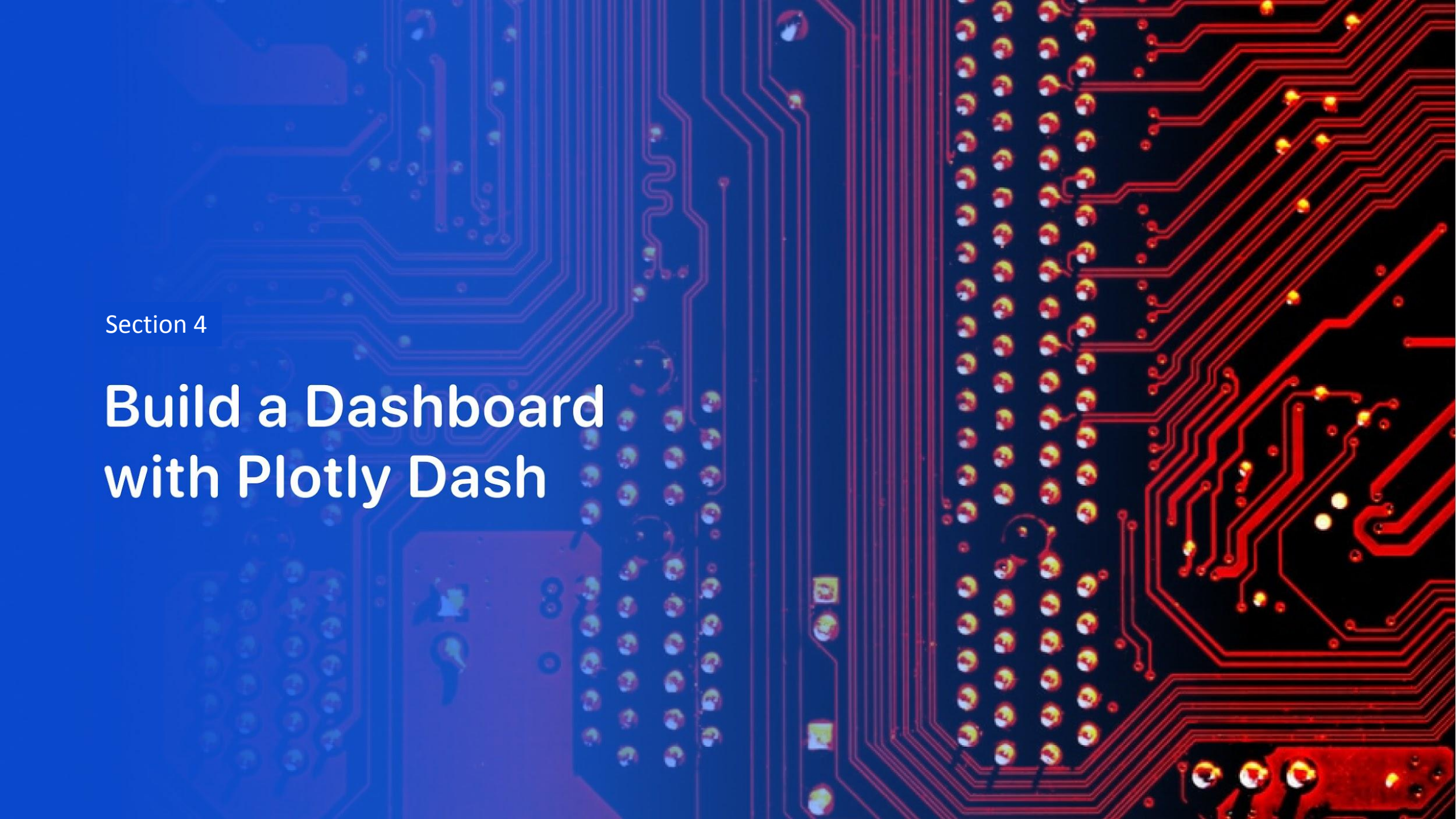


- The folium map visualizes **launch outcomes at Vandenberg Space Force Base**, with markers color-coded:
  - **Green markers** represent successful launches.
  - **Red markers** represent failed launches.
- The spiral clustering technique makes overlapping launch events easier to distinguish.
- The visualization highlights that while **successful launches (green) are present**, there is also a **notable number of failures (red)**, offering quick insight into reliability patterns at this site.
- This method allows an **intuitive geographic + outcome-based view** of launch history, making it easier to compare performance across sites.

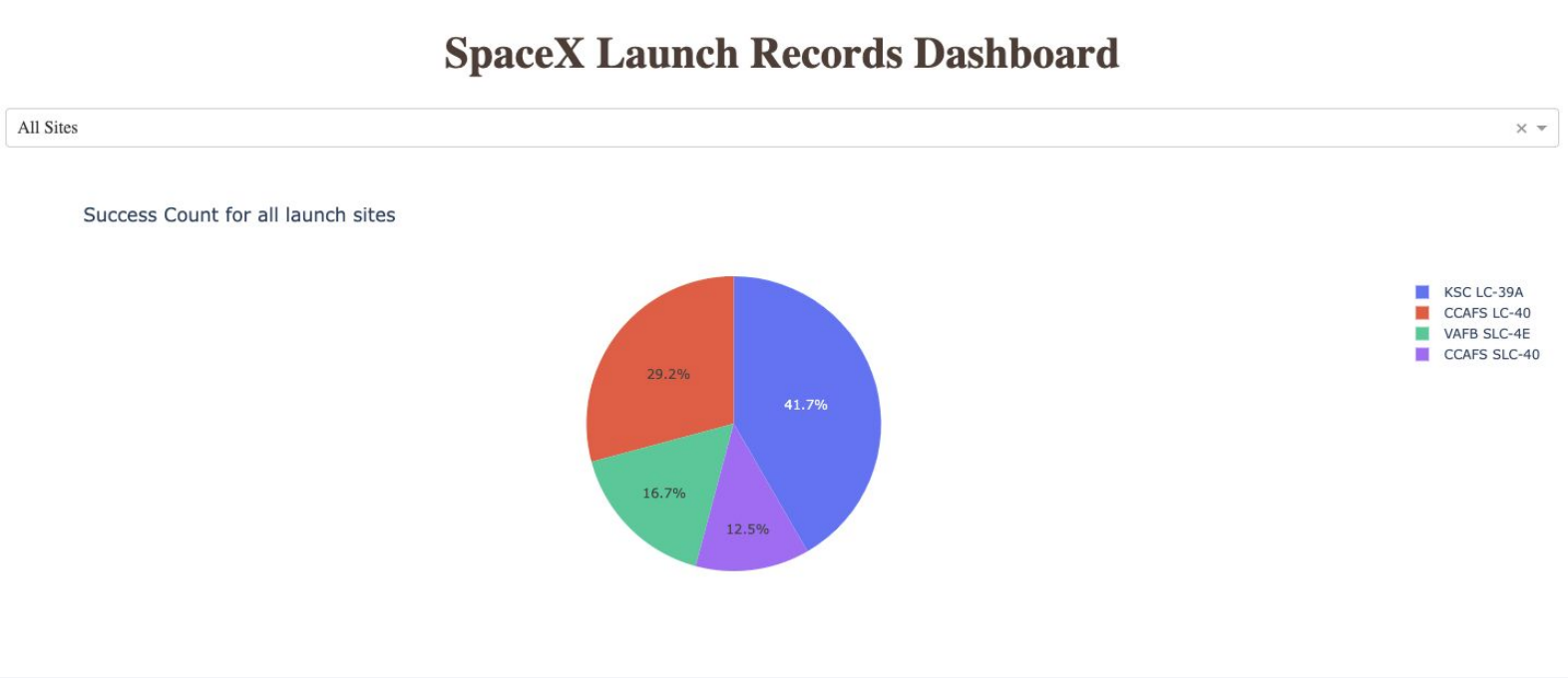
Launch Site Proximity Analysis



- The folium map zooms in on **Cape Canaveral's SLC-40 launch site**.
- Proximity lines and distance markers illustrate how close the launch site is to key infrastructure:
  - **Railway** (used for transporting heavy equipment and rocket components)
  - **Highway access** (supporting logistical operations and personnel transport)
  - **Coastline** (important for safety, as launch crashes in water bodies pose least risk to life and infrastructure)
- The **distance overlays** quantify these proximities, emphasizing how site location is optimized for **logistics, accessibility, and safety**. This spatial analysis helps explain why Cape Canaveral is a strategic hub for launches — combining coastal access with strong transport connectivity.

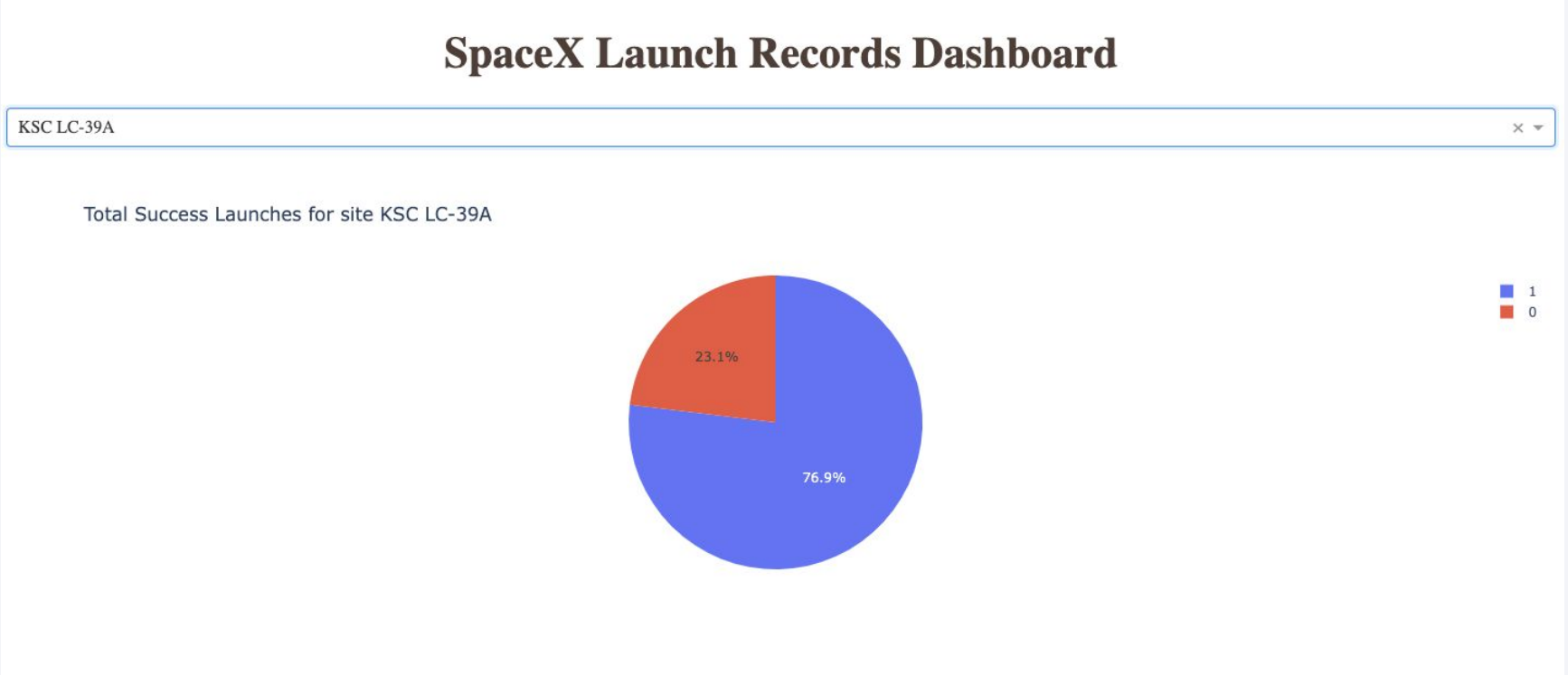

Section 4
Build a Dashboard with Plotly Dash

Launch Success Distribution Across All Sites



- The pie chart shows the **share of successful launches** at each SpaceX launch site.
- **KSC LC-39A** accounts for the largest portion (**41.7%** of all successes).
- **CCAFS LC-40** follows with **29.2%**, showing it as another major contributor.
- **VAFB SLC-4E** and **CCAFS SLC-40** contribute smaller shares (**16.7%** and **12.5%**, respectively).
- This distribution highlights that launch successes are not evenly spread but **dominated by a few high-performing sites**, with KSC LC-39A leading.

Launch Site with Highest Success Ratio



- This pie chart illustrates the distribution of successful vs. failed launches for the site with the **highest overall success ratio**.
- The **blue segment (76.9%)** represents successful launches, while the **red segment (23.1%)** shows failed attempts.
- This indicates that the selected launch site (e.g., **KSC LC-39A**) has consistently been the most reliable in SpaceX's portfolio, with over **three-quarters of launches achieving success**.

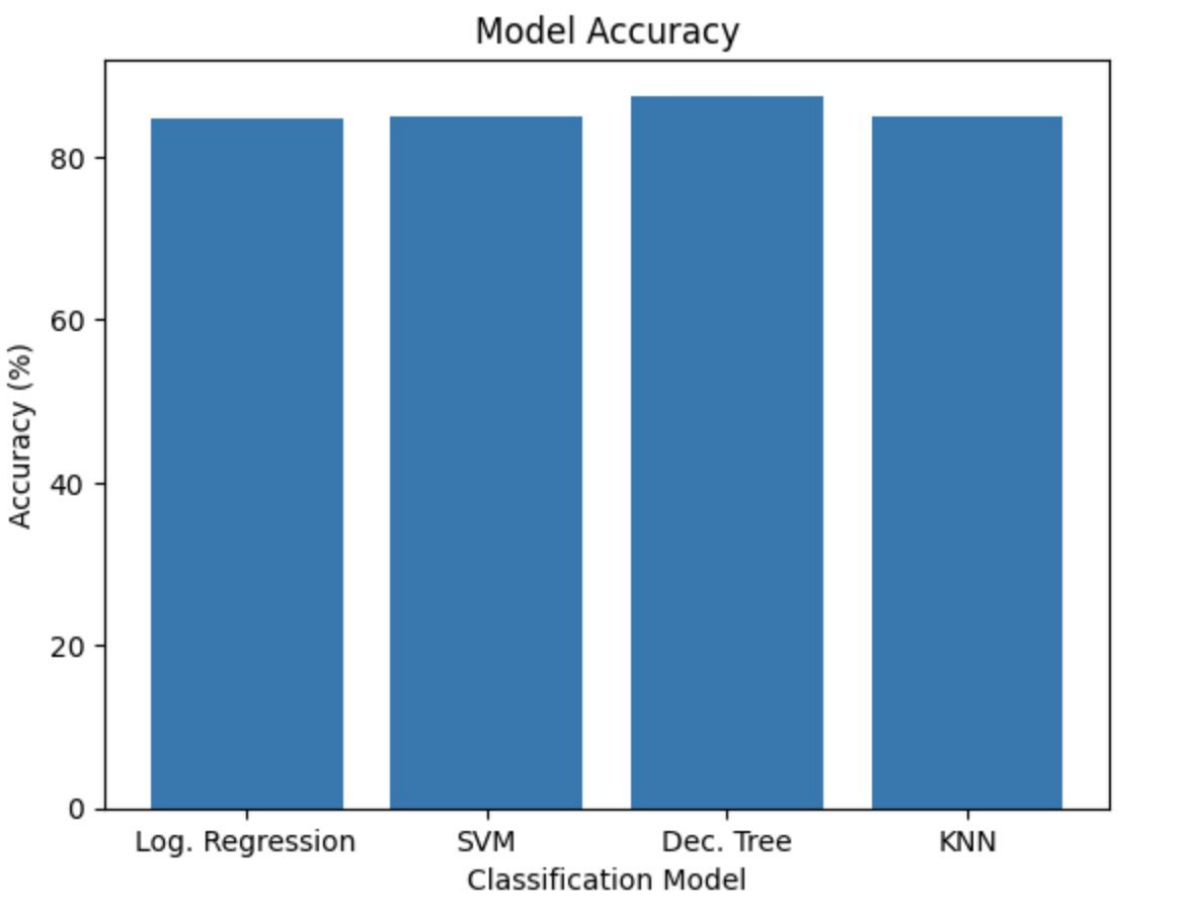
Impact of Payload Mass on Launch Outcome

The Plotly dashboard allows for manipulation of the the payload range with the slider above the charts

Section 5

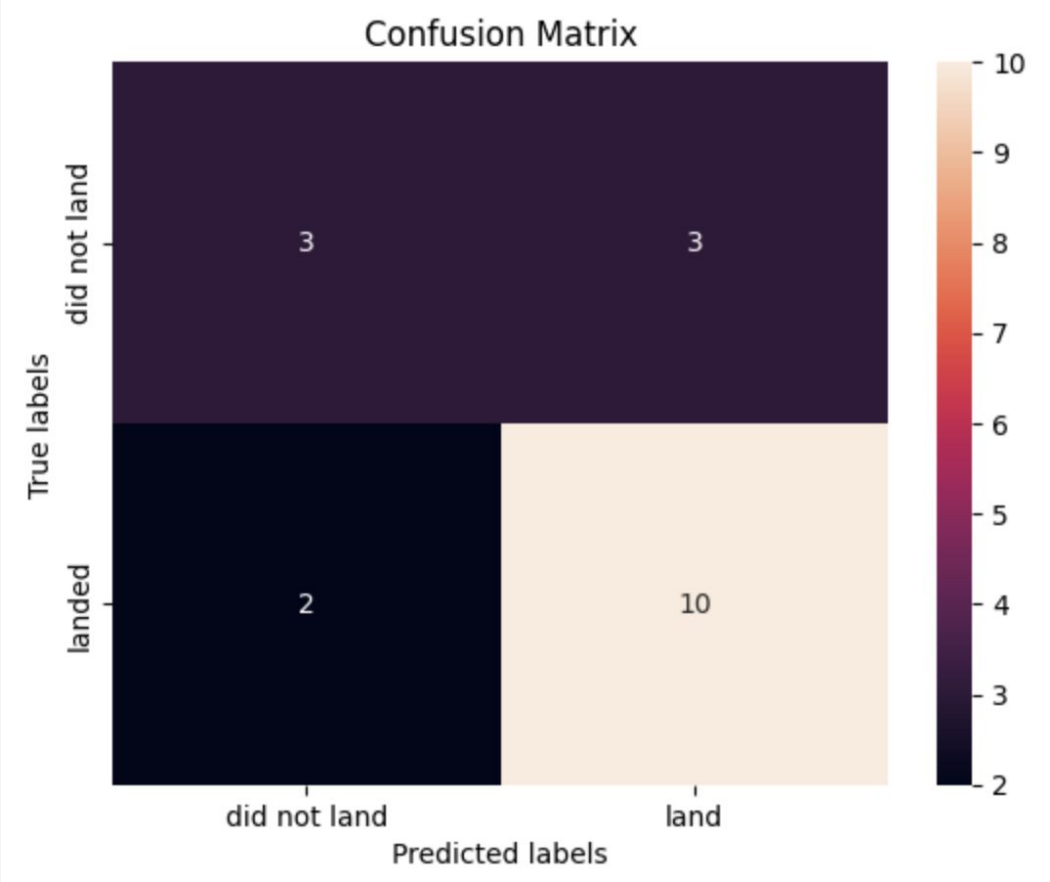
Predictive Analysis (Classification)

Classification Accuracy



The Decision tree model was found to have the highest accuracy (87.8%) amongst the models tested

Confusion Matrix



This decision matrix obtained from running the model on test data reveals the model found **10 True Positives, 2 False Positives, 3 True Negatives and 3 False Negatives**

Conclusions

Industry Context

- SpaceX has lowered launch costs through reusable rockets, setting the benchmark for competitors like SpaceY.
- Understanding launch outcomes and success drivers is critical for new entrants in the market.

Data Exploration & Visualization

- Cleaned and integrated SpaceX launch data using **Python, SQL, and API/web scraping**.
- **Exploratory Data Analysis** revealed clear trends:
  - Higher payloads and certain booster categories correlated with lower landing success.
  - Specific launch sites (e.g., **KSC LC-39A**) had the highest success ratios.
- **Interactive Dashboards (Plotly Dash & Folium)** enabled intuitive exploration of:
  - Payload vs. launch success scatter plots.
  - Success distribution across all launch sites.
  - Geographic site clustering and proximity to transport and safety infrastructure.

Predictive Modeling

- Applied **classification models** (Decision Tree, KNN, SVM, Logistic Regression) to predict landing success
- Achieved **up to 87.8% prediction accuracy**, demonstrating feasibility of outcome forecasting.
- Payload mass, orbit type, and booster version category emerged as **key predictive features**.

Key Insights

- SpaceX's operational success is tied to both **engineering design (reusable boosters)** and **site/infrastructure advantages** (proximity to coastline and transport).
- **Data-driven models** can guide competitors like SpaceY in selecting mission parameters that maximize landing success probability.
- Visual dashboards and mapping tools provide actionable transparency for stakeholders, investors, and operations planners.

Appendix

- All code is detailed in Jupyter Notebooks hosted at:
  - https://github.com/christabelek/Capstone

- Acknowledgments
  - Thank you to Joseph Santarcangelo, Yan Luo and Azim Hirjani for creating the course and materials.

Thank you!