IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adrián Rodríguez
2025/02/20

Outline

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

Executive Summary

The goal of this project is to predict the success of Falcon 9's first-stage landing after launch. By analyzing historical SpaceX launch data, we built a model to forecast landing success, helping to improve decision-making and reduce mission costs.

Summary of Methodologies

- Data Collection: Gathered data on weather, rocket specs and launch sites.
- Modeling: Applied machine learning models (Logistic Regression, Decision Trees, etc.).
- Evaluation: Assessed model performance with accuracy and precision metrics.

Summary of Results

- Best Model: Decision Tree with 85% accuracy.
- Key Insights: Weather, rocket specs, and launch site were key factors influencing landing success.

Introduction

Project Background

- SpaceX has developed reusable rockets, where the first stage is designed to return and land after launch.
- The success of these landings is crucial for reducing the cost of space missions and advancing space exploration.

The Challenge

- Predicting whether the first stage of the Falcon 9 rocket will successfully land after launch.
- We will analyze historical launch data to build a model that can predict landing success based on various factors like weather, rocket performance, and launch conditions.

Objective

- To create a predictive model that helps anticipate landing success, improving decision-making and efficiency for future missions.

Section 1

Methodology

Methodology

- **Data Collection:** Gathered historical SpaceX launch data, including weather conditions, rocket performance, and launch site information.
- **Data Wrangling:** Cleaned and preprocessed the data by handling missing values and removing irrelevant information to ensure consistency and accuracy.
- **Exploratory Data Analysis (EDA):** Conducted visual analysis using tools like matplotlib and seaborn to identify patterns and correlations in the data. Applied SQL queries to extract relevant insights from the data.
- **Interactive Visual Analytics:** Used Folium to map launch locations and visualize geographic patterns. Built Plotly Dash dashboards for interactive exploration of key factors affecting landing success.
- **Predictive Analysis:** Applied various classification models (e.g., Logistic Regression, Decision Trees) to predict landing success. Evaluated model performance based on metrics like accuracy and precision.
- **Model Building & Evaluation:** Tuned models using hyperparameter optimization to improve performance. Selected the best model based on evaluation results and accuracy.

Data Collection

Data Sources:

- Collected historical data from SpaceX's official API, covering launch events, weather, rocket specs, and launch sites.

Key Variables:

- **Launch Data:** Date, success/failure, and launch site.
- **Weather Data:** Temperature, wind speed, and humidity at launch time.
- **Rocket Data:** Specifications, fuel type, and payload information.

Geographic Data:

- Location of launch sites and landing zones.

Data Collection Steps:

- **Step 1:** Extracted raw data from the SpaceX API and public datasets.
- **Step 2:** Cleaned and formatted the data for consistency and usability.
- **Step 3:** Consolidated data from multiple sources into a single, comprehensive dataset.

Data Flowchart:

API → Raw Data → Cleaning & Formatting → Processed Dataset → Analysis

Data Collection – SpaceX API

Data Sources

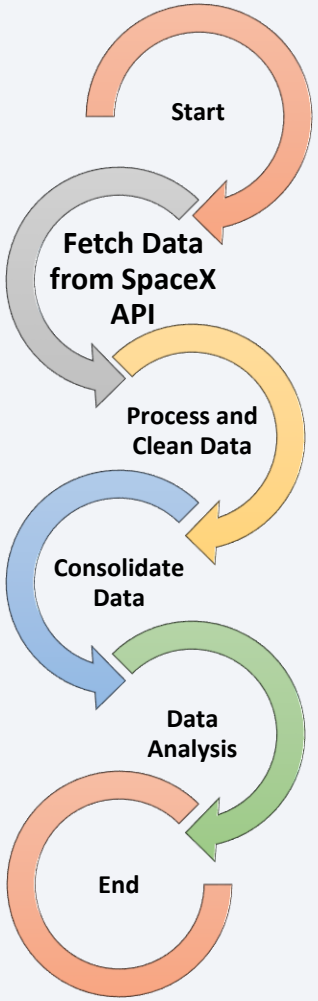
Used SpaceX's public REST API to collect data on launch events, rocket specs, weather conditions, and launch sites.

Key REST API Calls

- Launch Data: GET /launches – Retrieves launch details (success, date, site).
- Weather Data: Extracted from third-party sources based on launch times.
- Rocket Data: GET /rockets – Information about rocket specifications.
- Landing Data: GET /landings – Details about the success or failure of landings.

GitHub Reference

Access the completed notebook with SpaceX API calls and the outcomes here.



Data Collection - Scraping

Data Sources

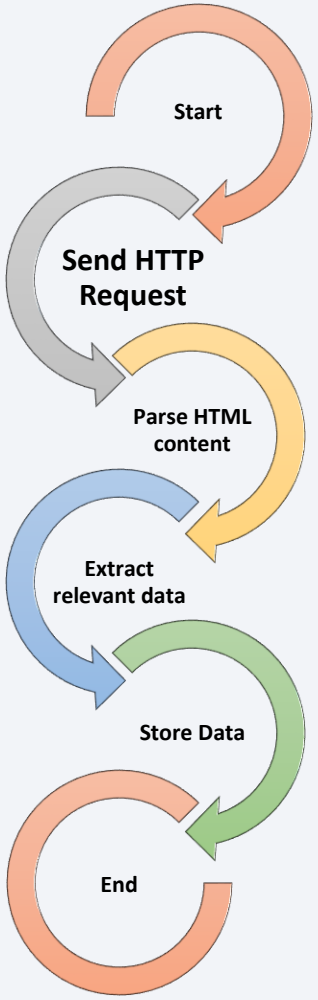
Wikipedia Pages (e.g., SpaceX, Falcon 9) to gather relevant information on launches, rocket specs, etc.

Key Web Scraping Steps

- Request Data: Send an HTTP request to the Wikipedia page URL using libraries like requests in Python.
- Parse HTML: Use BeautifulSoup or lxml to parse the HTML content of the page.
- Extract Data: Identify the specific HTML elements that contain the data you need (e.g., tables, headings, paragraphs).
- Store Data: Store the extracted data in a structured format like CSV, JSON, or a database for further analysis.
- Handle Errors: Implement error handling to manage failed requests or changes in page structure.

GitHub Reference

Access the completed notebook for web scraping and the outcomes here.



Data Wrangling

Data Sources

Collected data from various sources such as APIs (e.g., SpaceX API), web scraping (e.g., Wikipedia), or CSV files.

Key Data Wrangling Steps

- **Data Cleaning**
  - Handle missing data using techniques such as imputation (filling in missing values) or removing rows/columns with too many missing values.
  - Remove duplicates to ensure the dataset contains unique entries.
  - Correct data types (e.g., converting strings to numerical values for analysis).
- **Data Transformation**
  - Standardize data formats (e.g., date formatting, currency conversion).
  - Normalize or scale numerical values to bring consistency across features (e.g., scaling launch time or temperature).
- **Feature Engineering**
  - Create new columns based on existing data to help with analysis or model building (e.g., create a "Launch Success" column based on multiple conditions).
  - Encode categorical variables into numerical values (e.g., using one-hot encoding or label encoding).
- **Outlier Detection and Handling**
  - Detect and handle outliers (e.g., using Z-score or IQR) that may skew analysis results.
- **Data Aggregation**
  - Aggregate data by categories (e.g., aggregating by launch site or rocket type) to perform group-level analysis.

GitHub Reference

Access the completed data wrangling notebook and see the results here.

EDA with Data Visualization

Purpose of EDA

- Understand the structure of the dataset.
- Identify trends, patterns, and anomalies in the data.
- Support decision-making for feature selection and model building.

Key Charts and Why They Were Used

Helps identify skewness, outliers, and data spread.

Scatter Plots

- Used to analyze relationships between two numerical variables, such as payload mass vs. success rate.
- Helps detect correlations or clusters in the data.

Bar Charts

- Used to compare categorical data, such as the number of launches per launch site.
- Provides clear insights into categorical distribution.

Geospatial Maps (Folium):

- Used to visualize the geographic distribution of launch sites and their success rates.
- Helps analyze location-based trends.

GitHub Reference

Access the completed EDA with data visualization notebook and results here.

EDA with SQL

Launch Site Performance

- Identified most frequently used launch sites.
- Ranked sites by number of successful launches.

Mission Success Analysis

Counted successful vs. failed launches.

Analyzed trends in launch success over time.

Payload and Booster Efficiency

- Determined which boosters carried the heaviest payloads.
- Compared payload success rates across different booster versions.

Landing Outcomes

- Evaluated landing success rates by year.
- Identified improvements in rocket reusability.

Key Takeaway:

Data-driven insights help optimize launch strategies, improve booster efficiency, and enhance mission success rates.

Full Analysis & SQL Queries: GitHub Repository

Build an Interactive Map with Folium

- **Launch Site Markers:**
  - Plotted **latitude & longitude** coordinates for all launch sites.
  - Used **circle markers, popup labels, and text labels**, starting from NASA Johnson Space Center as a reference.
- **Launch Outcome Visualization:**
  - **Green markers** indicate successful launches, **red markers** indicate failures.
  - Clustered markers highlight launch sites with higher success rates.
- **Distance Visualization:**
  - Mapped distances between **KSC LC 39A** and key locations.
  - Used **colored lines** to show proximity to **railways, highways, coastlines, and cities** for risk assessment.
  - **Key Takeaway:**
    Interactive **Folium maps** enhance mission planning by visualizing **launch site performance, success trends, and safety considerations**.
- **Full Analysis & Interactive Map:** GitHub Repository

Build a Dashboard with Plotly Dash

- **Launch Site Selection:** Implemented an interactive **dropdown menu** to filter launch sites for targeted analysis.
- **Mission Success Visualization:**
  - **Pie chart** shows the total number of successful launches across all sites.
  - For specific sites, it displays **success vs. failure rates** to identify performance trends.
- **Payload Mass & Success Rate:**
  - **Range slider** allows users to filter launches based on payload mass.
  - **Scatter plot** visualizes payload mass vs. success rate, categorized by booster version, providing insights into optimal payload conditions.
- **Key Takeaway:** Interactive analytics improve decision-making by identifying the best launch sites and optimal payload conditions for successful missions.
- **Full Dashboard:** GitHub Repository

Predictive Analysis (Classification)

Model Selection & Building

- Tested multiple **classification models** (Logistic Regression, Decision Trees, Random Forest, etc.).
- Trained models using historical launch data to predict mission success.

Evaluation & Improvement

- Assessed models using **accuracy, precision, recall, and F1-score**.
- Tuned hyperparameters to optimize performance.

Best Performing Model

- Identified the model with **highest prediction accuracy** and **best generalization** to new data.
- Ensured interpretability for **business decision-making**.

Key Takeaway

- Predictive modeling enables **data-driven decision-making**, reducing risks and optimizing future launches.

Full Analysis & Model Details: GitHub Repository

Results

Exploratory Data Analysis (EDA) Results

- **Mission Success Over Time**

  Success rate improved significantly over time, with early launches having higher failure rates and more recent launches showing consistent success.
- **Launch Site Performance**

  KSC LC-39A had the highest success rate, showing strong performance in Falcon 9 booster landings.
- **Orbit Type Success**

  Some orbits, like **GEO** and **SSO**, showed a perfect success rate, suggesting certain mission types are more reliable.

Interactive Analytics Demo

- **Launch Site Map**

  Interactive map showing the geographic locations of launch sites.
- **Mission Success Dashboard**

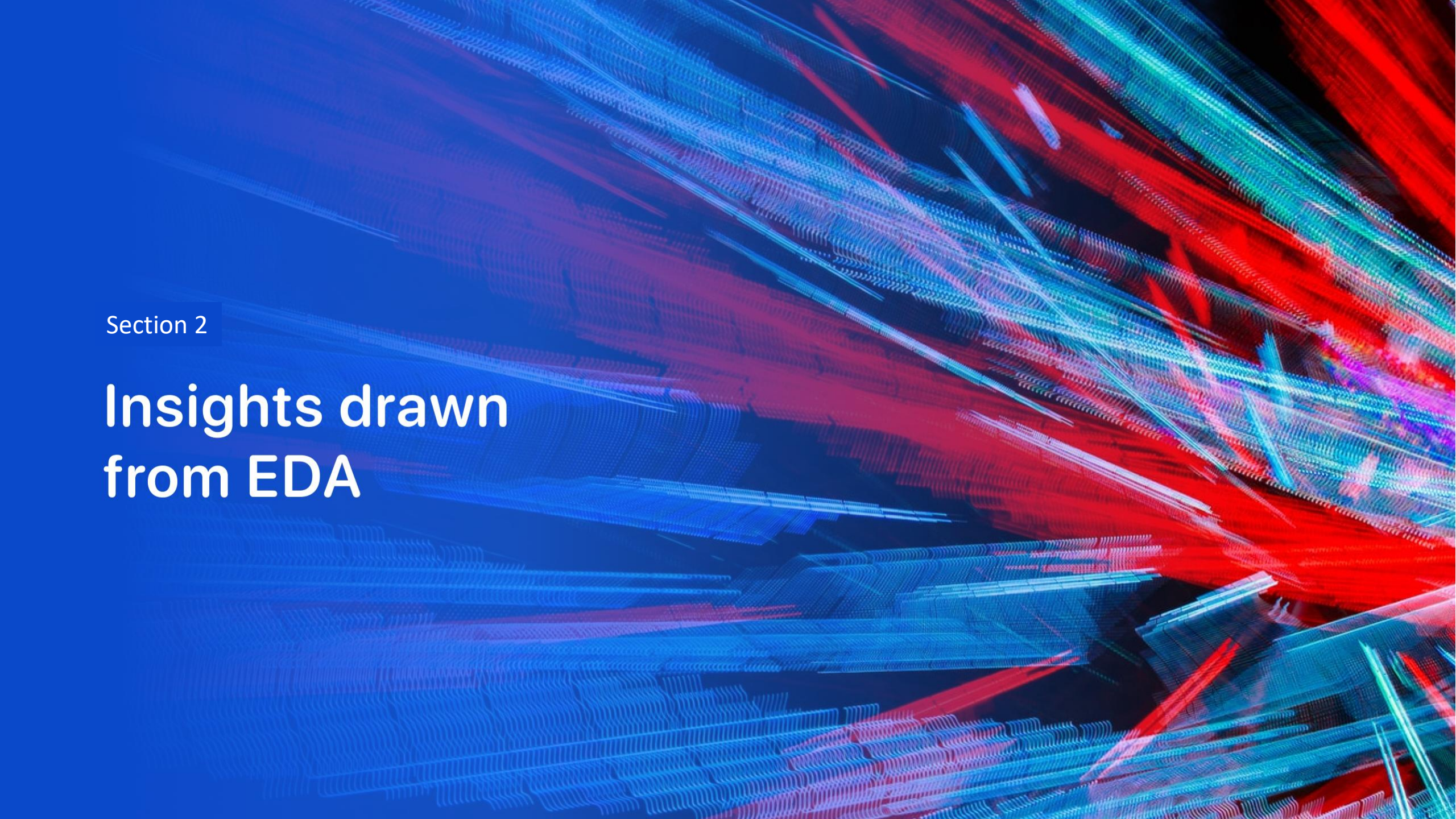
  A dashboard visualizing success rates across launch sites and orbit types.

Predictive Analysis Results

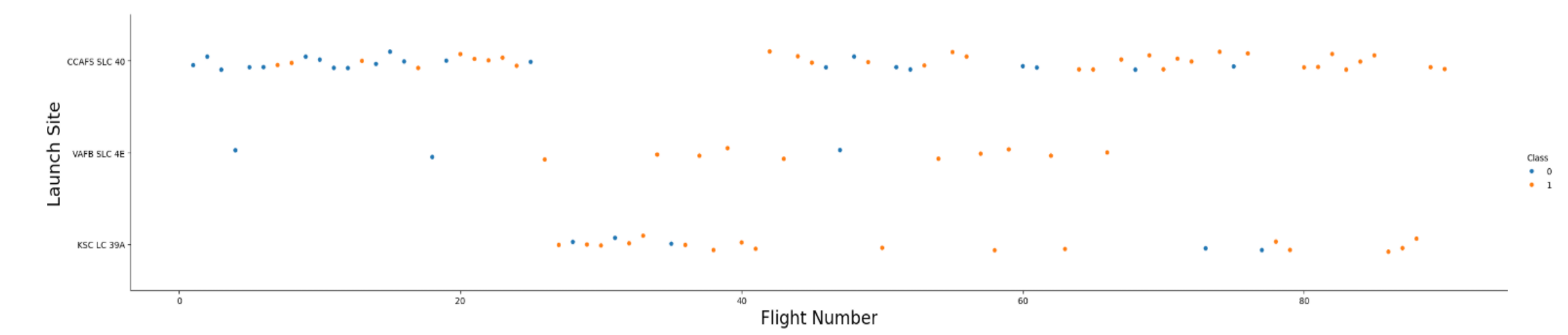
- **Best Model**

  Decision Tree model showed the best performance with **94.44% accuracy** in predicting launch success.
- **Key Metrics:**

  Low number of false positives/negatives, making the model reliable for future predictions.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



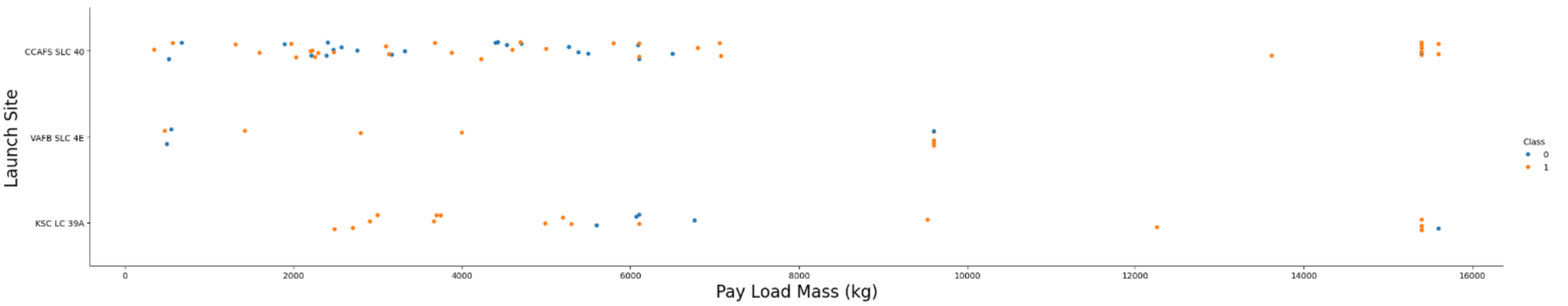
Explanation:

- **X-axis (Flight Number):** Sequential flight number for each SpaceX mission.
- **Y-axis (Launch Site):** Different launch sites for SpaceX missions.

Insights:

- Shows the distribution of flights across sites.
- Highlights which launch sites are used more frequently over time.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

Payload vs. Launch Site



Explanation:

- **X-axis (Pay Load Mass):** The payload mass refers to the weight of the cargo that the rocket transports into orbit.
- **Y-axis (Launch Site):** Different launch sites for SpaceX missions.

Insights:

- Payload mass affects success rates.
- CCAFS SLC-40 handles a mix of light and medium payloads.
- Heavier payloads are mostly launched from KSC LC-39A.

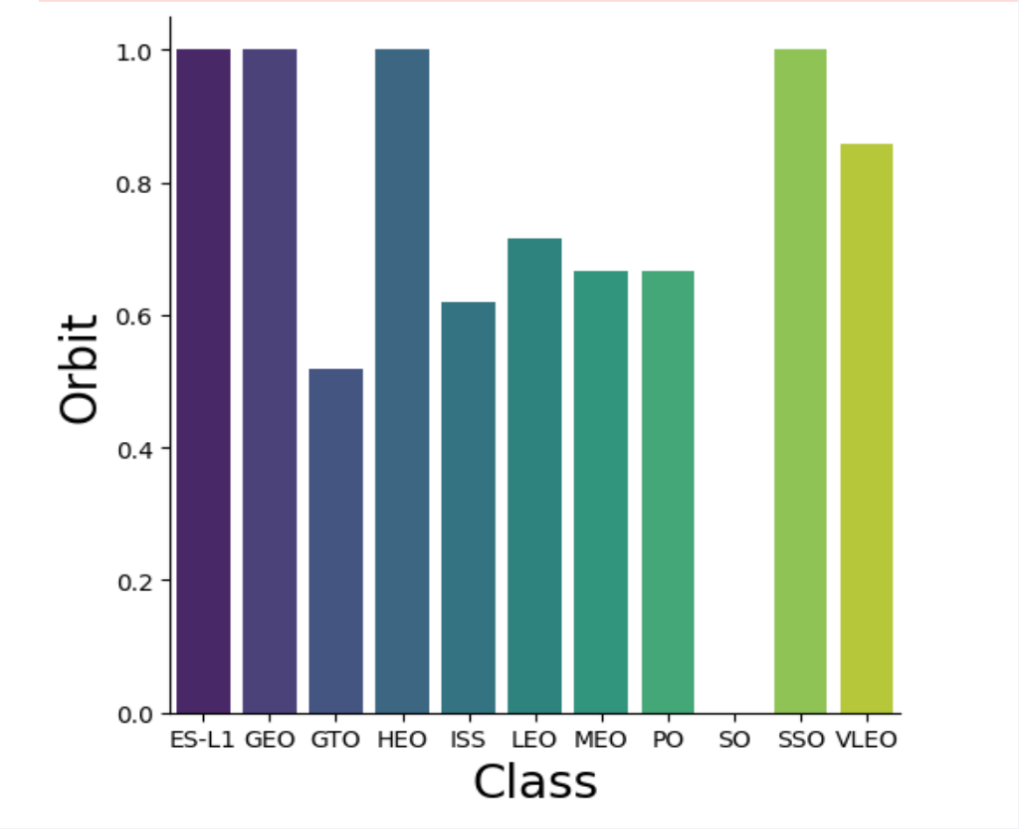
Success Rate vs. Orbit Type

Explanation:

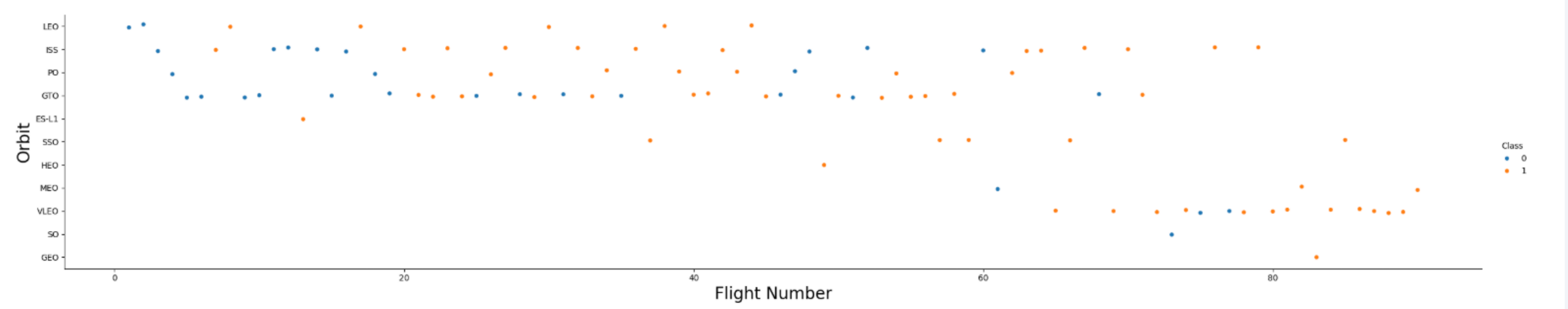
- **X-axis (Class):** Classification of the different orbits.
- **Y-axis (Orbit):** Percentage of success of launch.

Insights:

- ES-L1, GEO and HEO have 100% success rate
- GTO,ISS,LEO,MEO,PO,VLEO have between 50% and 100% success rate
- SO have 0% success rate.



Flight Number vs. Orbit Type



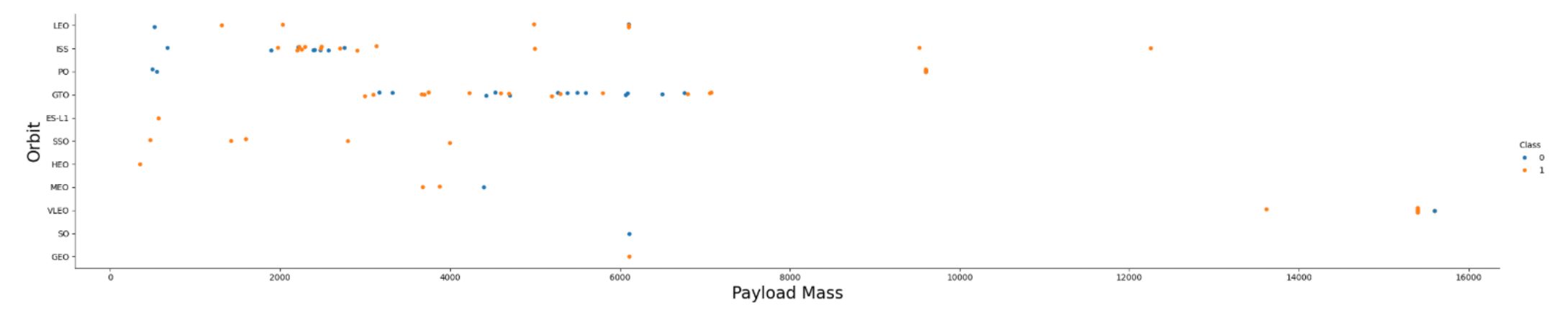
Explanation:

- **X-axis (Flight Number):** Sequential flight number for each SpaceX mission.
- **Y-axis (Orbit):** Classification of the different orbits.

Insights:

- Higher flight numbers are associated with more diverse orbit types.
- LEO (Low Earth Orbit) is the most frequently used orbit.
- More recent flights include GEO and beyond.

Payload vs. Orbit Type



Explanation:

- **X-axis (Payload Mass):** The payload mass refers to the weight of the cargo that the rocket transports into orbit.
- **Y-axis (Orbit):** Classification of the different orbits.

Insights:

- LEO (Low Earth Orbit) supports a wide range of payload masses.
- GEO (Geostationary Orbit) missions typically carry heavier payloads.
- SSO (Sun-Synchronous Orbit) and other specialized orbits handle lighter payloads.
- Payload mass impacts mission complexity.

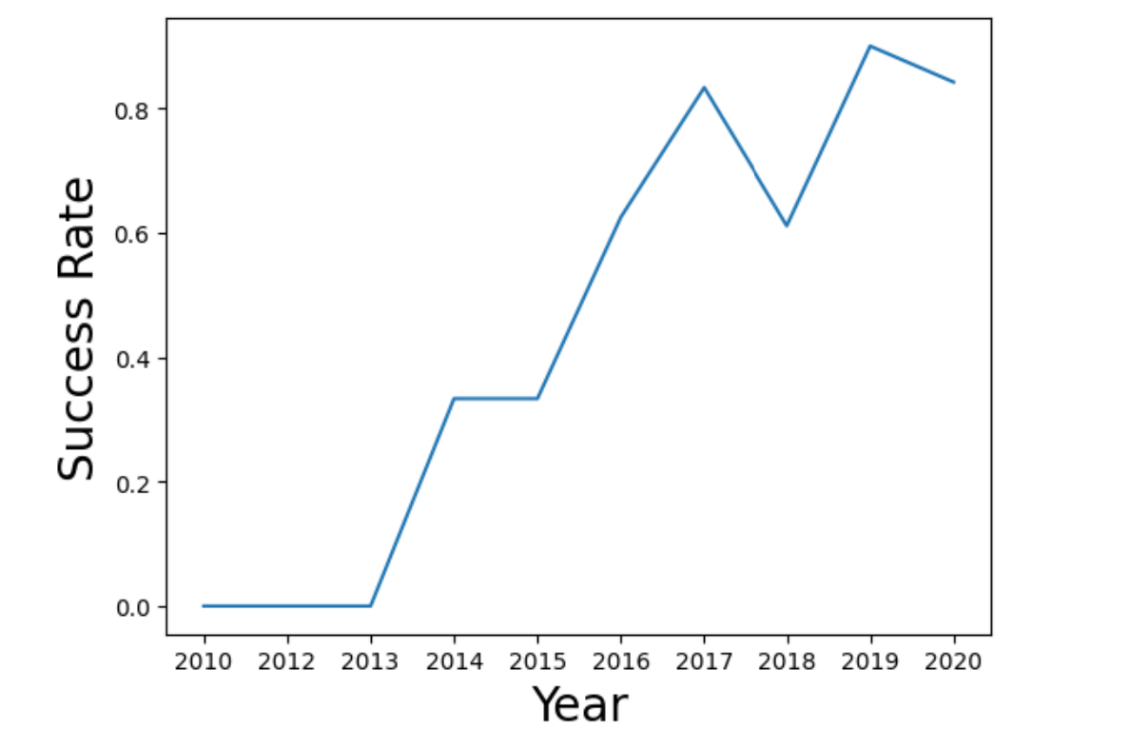
Launch Success Yearly Trend

Explanation:

- **X-axis (Year):** Year of different launch.
- **Y-axis (Success Rate):** Percentage of success of launch.

Insights:

- The most success rate is 2016 and 2019.
- Since 2019 have gone down the success rate.



All Launch Site Names

- Display the name of unique launch site in database.

```
%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'

- Display the 5 records where the launch have ‘CCA’ in database

```
%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

- Total payload mass carried by carrier by NASA (CRS) in database

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXTABLE WHERE customer = 'NASA (CRS)';

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Displaying average Payload Mass carried booster version F9 v1.1

```
1]: %sql SELECT AVG(payload_mass__kg_) AS average_payload_mass FROM SPACEXTABLE WHERE booster_version like '%F9 v1.1%';
 * sqlite:///my_data1.db
Done.
1]:
```

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- The first date when the first successful landing outcome in ground pad was in database

```
30]: %sql SELECT min(date) AS first_successful FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
30]: first_successful
     2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of the boosters which have success in drone ship and have payload mass between 4000 and 600 kg in database

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4
```

```
 * 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

- Listing the total number of successful and failure mission outcomes.

```
]: %sql SELECT Mission_Outcome, COUNT(*) AS total FROM SPACEXTABLE GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

]:

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

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.

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

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

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

2015 Launch Records

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

```
%%sql SELECT strftime('%m', Date) AS month, Date, booster_version, launch_site, landing_outcome
    FROM SPACEXTABLE
    WHERE landing_outcome = 'Failure (drone ship)' AND strftime('%Y', Date) = '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	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)) between the date 2010 and 2017.

```
%sql SELECT Landing_Outcome, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GI
 * sqlite:///my_data1.db
Done.
```

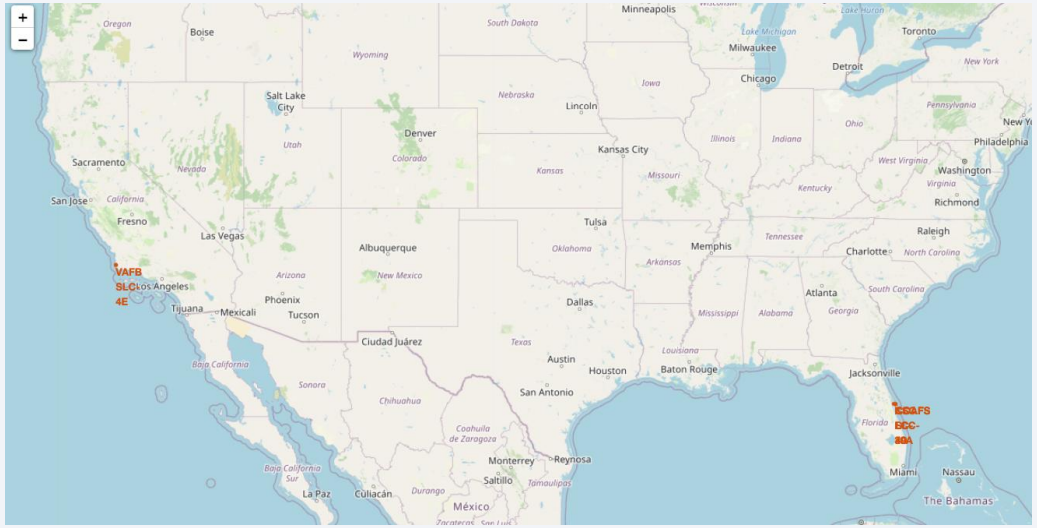
Landing_Outcome	COUNT_LAUNCHES
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

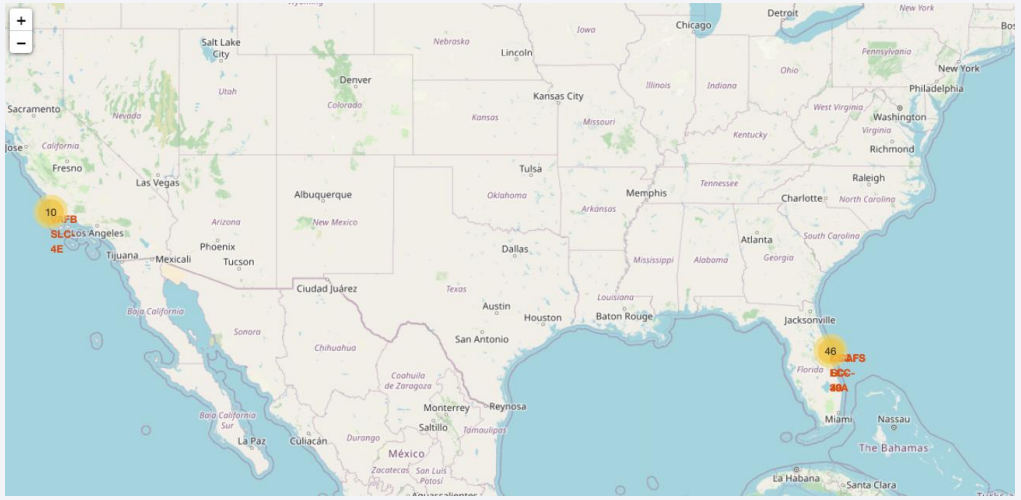
Locations of all launches sites in the USA

- As expected, most launchers in the United States are located in Florida because this point is closer to the equator, which allows them to be more in sync with the Earth's rotation, make safer launches and save on costs.
- Also for safety reasons, launchers in the United States are placed in areas facing the ocean so that the launch is in that trajectory and not towards populations. In other countries, this safety measure is taken in desert or uninhabited areas.



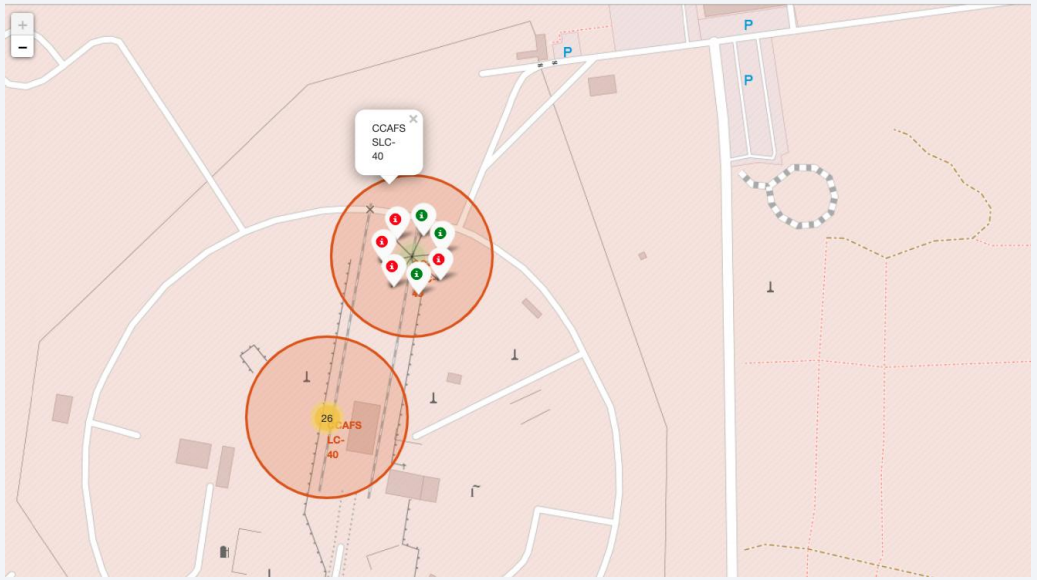
Total releases in each location in the USA

- Because there are more launchers and availability in Florida due to cost, maintenance and mission issues, most are launched on rocket launchers in that area.



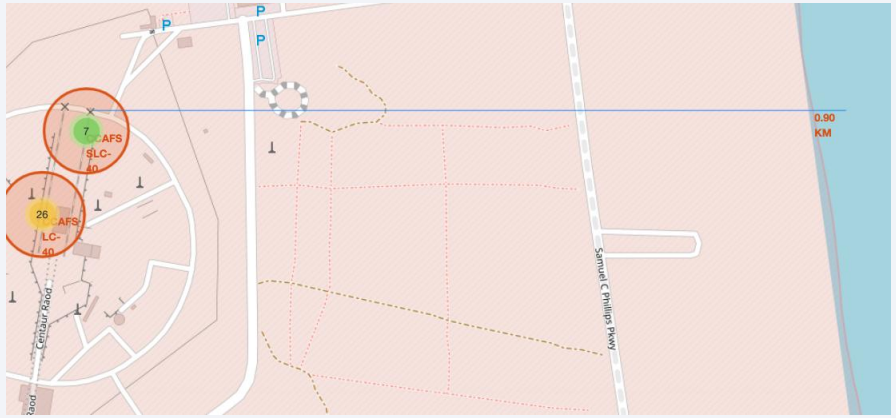
Launch records in the main launcher site

- Brands use the following color
  - Green are successful launches
  - Red are unsuccessful launches
- Launch Site KSC LC-39A has a very high Success Rate.



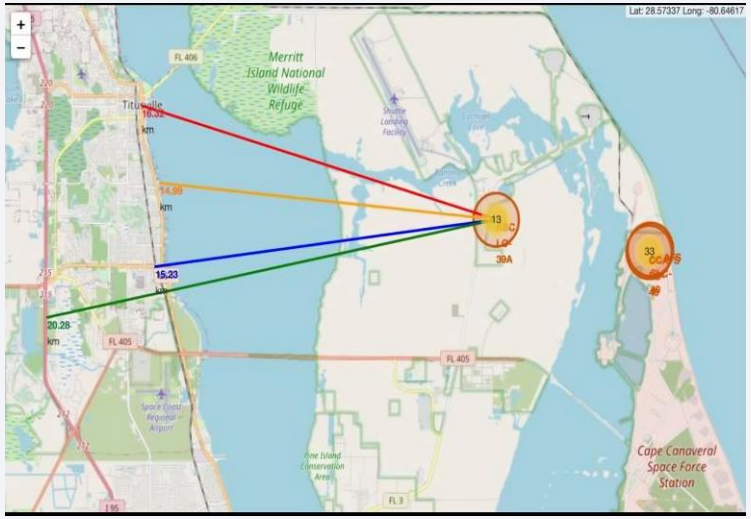
Distance from the launch side KS LC-39A

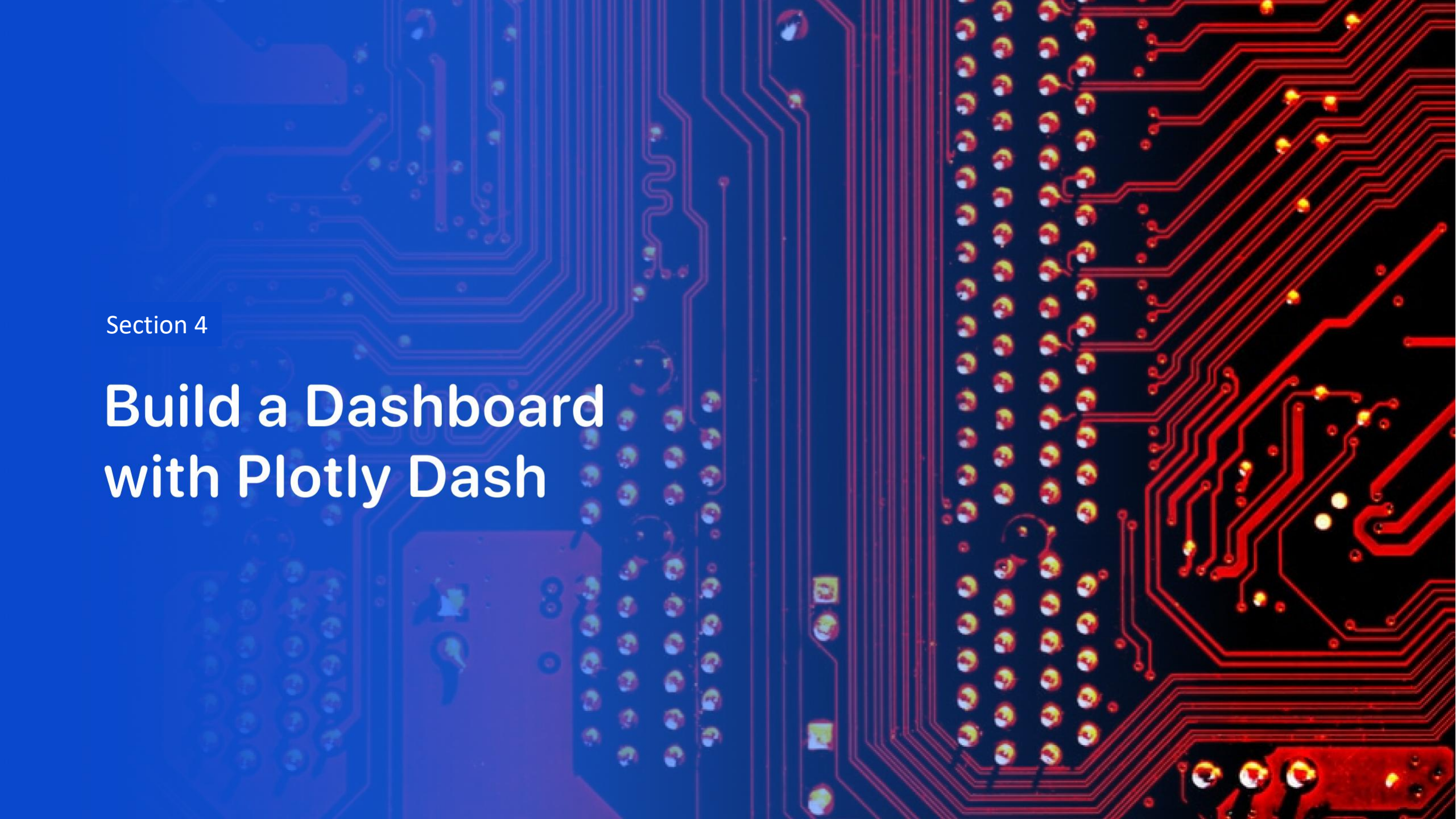
- **Proximity to Key Infrastructure:**
  - **Railway:** 15.23 km
  - **Highway:** 20.28 km
  - **Coastline:** 14.99 km
  - **Closest City (Titusville):** 16.32 km
- **Risk Consideration:**
  A failed rocket at high speed can travel **15-20 km in seconds**, posing potential risks to populated areas.
- **Insight:**
  Location balance between **accessibility & safety** is crucial for optimizing future launch operations.



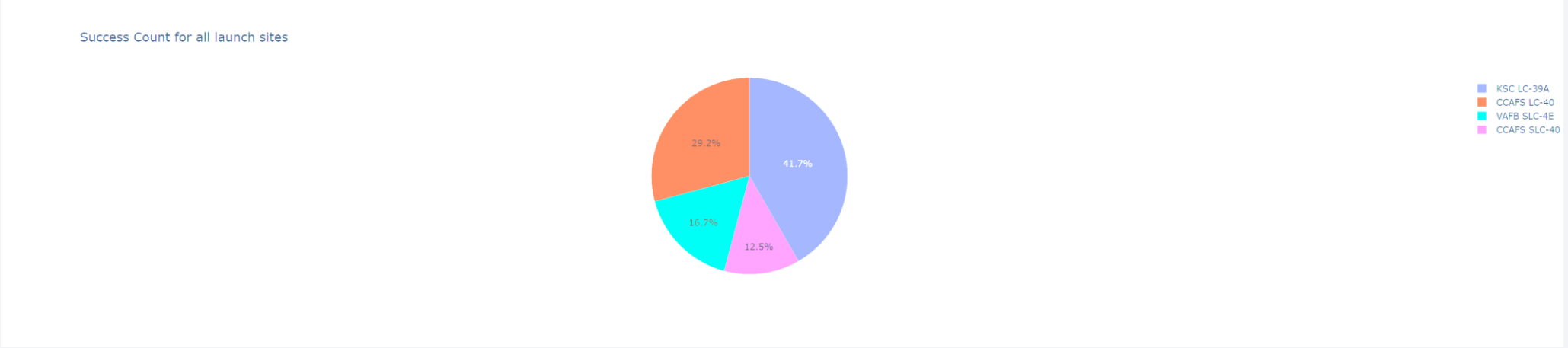
Distance from the launch side KS LC-39A

- **Proximity to Key Infrastructure:**
  - **Railway:** 15.23 km
  - **Highway:** 20.28 km
  - **Coastline:** 14.99 km
  - **Closest City (Titusville):** 16.32 km
- **Risk Consideration:**
  A failed rocket at high speed can travel **15-20 km in seconds**, posing potential risks to populated areas.
- **Insight:**
  Location balance between **accessibility & safety** is crucial for optimizing future launch operations.



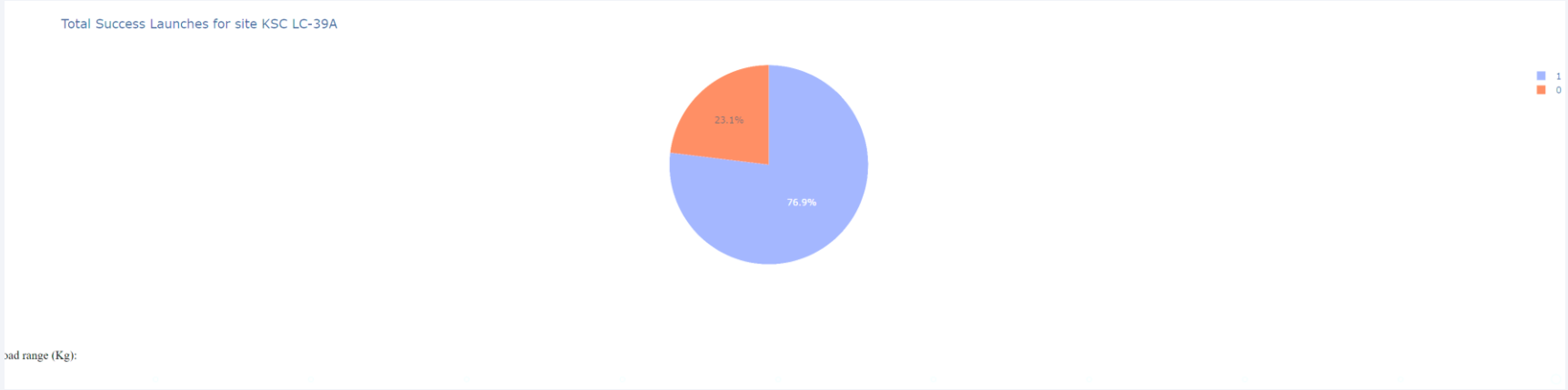

Section 4
Build a Dashboard with Plotly Dash

Success launch percentage for all sites



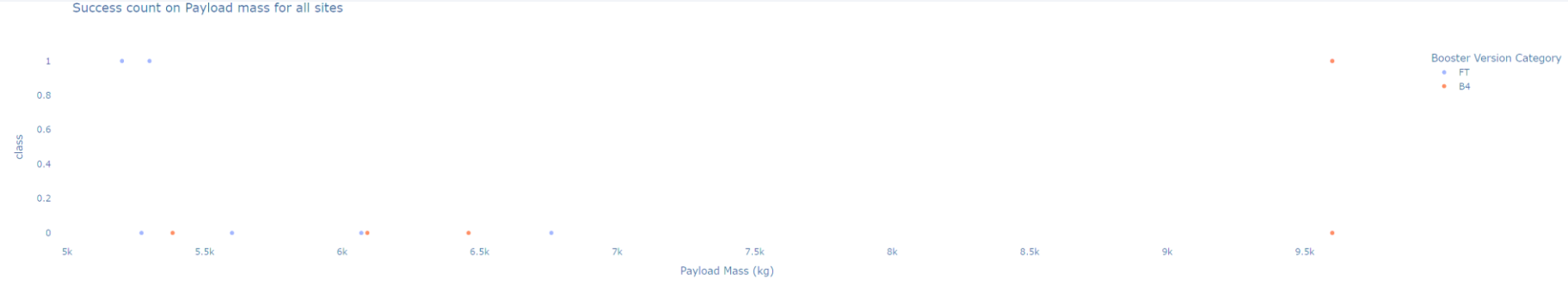
- This pie chart shows the success rates for rocket launches at each of the launch sites in the United States.
- It can be seen that KSC LC-39A has the highest number of successful launches.
- CCAFS SLC-40 has the lowest number of successful launches.
- Nothing can be concluded from this, except that the most common areas for successful launches.

Success launch percentage in KSC LC-39A



- You can see in this pie chart the percentage of failure and success rates in this location.
- Obviously if you look at the rest, this is the one with the highest success rate compared to the rest of the cases.
- Also, its failure rate is much lower than the average, so investigating why fewer failures occur here would be essential.

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES



- Here is the list of all the launches in any area relating the payload mass and launch outcome divided into two sections: the first graph is the section from 0 to 5000 and the second section, from 5000 to 10000 kg.

Section 5

Predictive Analysis (Classification)

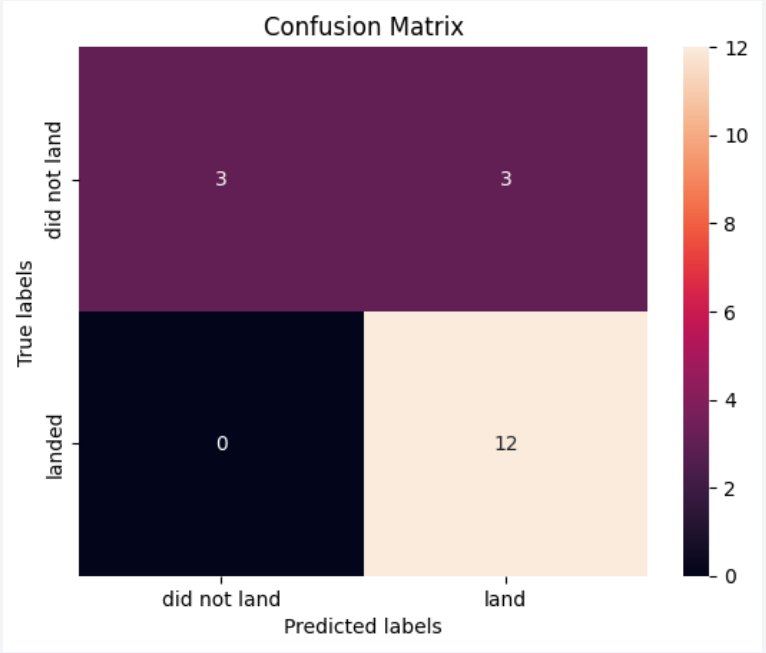
Classification Accuracy

- **Test Set scores are identical across models** (Logistic Regression, SVM, Decision Tree, KNN), making it unclear which performs best.
- **This may be due to the small test sample size (18 samples).**
- **To address this, we tested all models on the full dataset.**
- **Results confirmed Decision Tree as the best model**, achieving the highest accuracy and performance.

	Model	Jaccard Score	F1 Score	Accuracy Score
0	Logistic Regression	0.8	0.888889	0.833333
1	Support Vector Machine	0.8	0.888889	0.833333
2	Decision Tree	0.8	0.888889	0.833333
3	K-Nearest Neighbors	0.8	0.888889	0.833333

Confusion Matrix

- **Logistic Regression effectively distinguishes between classes**, showing its predictive capability.
- **True Positives & True Negatives are well-classified**, indicating good model performance.
- **Some misclassifications exist**, but further tuning can improve accuracy.
- **Overall, the model provides reliable predictions for launch success.**



Conclusions

- **Decision Tree Model performed best** in predicting launch success.
- **Lighter payloads have higher success rates**, while heavier payloads face more challenges.
- **Launch sites are strategically located** near the **Equator and coastlines** for efficiency and safety.
- **Launch success rates have improved over time**, reflecting advancements in technology and operations.
- **KSC LC-39A has the highest success rate**, making it the most reliable launch site.
- **Certain orbits (ESL-1, GEO, HEO, and SSO) achieved 100% success**, indicating mission reliability in these destinations.
- **Future Improvements**
  - **AI-driven predictive maintenance:**
    Using **machine learning to detect potential booster failures** before launch, reducing the risk of malfunctions.
  - **Optimizing payload-to-orbit matching:**
    Enhancing **payload distribution strategies** to ensure each rocket is optimized for its target orbit, improving efficiency and success rates.
  - **Key Takeaway:**
    With advanced AI and smarter launch planning, **SpaceX can further increase reliability, reduce failures, and improve mission success.**

Appendix

1. **Data Collection:**
    1. **Public APIs:** Used SpaceX public APIs to retrieve detailed launch data, including rocket specifications, payload information, and landing outcomes.
    2. **Web Scraping:** Implemented web scraping techniques on sources like Wikipedia to complement and validate API data.
2. **Data Preparation & Cleaning:**
    1. **Handling Missing Values:** Processed and cleaned the dataset to ensure consistency and completeness.
    2. **Feature Engineering:** Created new variables to improve predictive modeling accuracy.
3. **Exploratory Data Analysis (EDA):**
    1. **Data Visualization:** Used **matplotlib, seaborn, and Plotly** to analyze trends in launch success rates, payload mass, and orbit types.
    2. **SQL Queries:** Performed **SQL-based analysis** to extract insights from structured launch data.
4. **Predictive Modeling:**
    1. **Machine Learning Models:**
        1. **Decision Tree & Logistic Regression** were tested for predicting successful landings.
        2. **Hyperparameter Tuning** improved model accuracy.
    2. **Evaluation Metrics:** Accuracy, precision, recall, and F1-score were used to assess model performance.
5. **Interactive Analytics & Visualization:**
    1. **Folium Maps:** Used for geographic visualization of launch sites and risk analysis.
    2. **Plotly Dash:** Created an interactive dashboard to explore SpaceX launch data dynamically.

Thank you!