IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vizen P. Solanki
11th August 2025

Outline

Executive Summary

- **Summary of methodologies**
  - Data collection and preprocessing (standardization using `StandardScaler`)
  - Data splitting into training and test sets (80:20 split)
  - Model selection: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors
  - Hyperparameter tuning using GridSearchCV (10-fold cross-validation)
  - Model evaluation using accuracy on test data
- **Summary of all results**
  - Logistic Regression: **83.33%** accuracy
  - SVM (best kernel: **Sigmoid**): **83.33%** accuracy
  - Decision Tree: **66.67%** accuracy
  - KNN: **83.33%** accuracy
  - Best performing model: **Logistic Regression** with **83.33%** accuracy

Introduction

- **Project Background and Context**
  - This project focuses on building and evaluating machine learning models to predict whether the first stage of spacex's falcon9 rocket land successfully using multiple algorithms, including **Decision Trees**, **Support Vector Machines**, and **Random Forests**.
  - The dataset contains labeled examples with features representing key patterns and behaviors (e.g., Launch site, payload weight, or reused).
  - **Goal:** Identify the most effective model and optimal hyperparameter settings that maximize accuracy while ensuring strong generalization to unseen data.
- **Problems to Address**
  - Which classification algorithm performs best on the dataset?
  - What are the optimal hyperparameters for each model?
  - How do the models compare across **training**, **validation**, and **test** sets?
  - What trade-offs exist between **accuracy** and **interpretability** of the models?

Section 1

Methodology

Methodology

Executive Summary

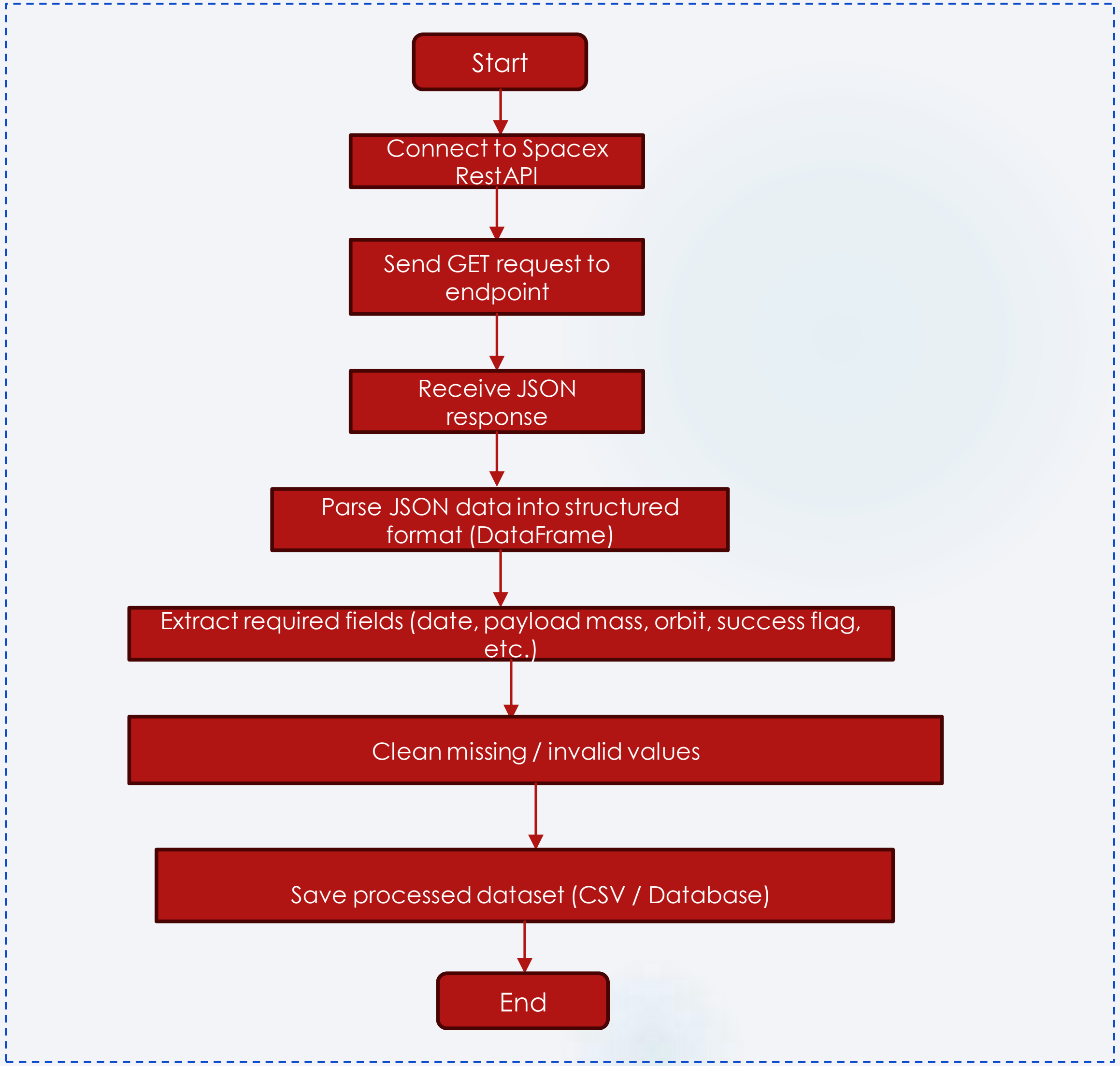
- Data collection methodology:
  - Data related to falcon 9 launches was collected with API calls and Webscraping from SpaceX website and Wikipedia.
- Perform data wrangling
  - Data was then cleaned, missing values were delt with and then the data was multiple success parameters were replaced with 1(Success) and 0(Failure) and data was normalized
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was then split in training data(80%) and testing data(20%) and multiple models were trained in order to identify the best fit model.

Data Collection

- ## Source of Data
  - The dataset was obtained from the **SpaceX API** and other public datasets provided by SpaceX and NASA.
  - Data includes historical records of **SpaceX Falcon 9 launches**, retrieved via API calls and web scraping from official sources such as Wikipedia and SpaceX's website.
- ## Data Collected
  - **Launch Details:** Mission names, dates, rocket types, payload mass, launch sites.
  - **Outcome Variables:** Success or failure of booster landings.
  - **Payload Information:** Mass, type (e.g., satellite, resupply), and customer.
  - **Geographic Information:** Launch site coordinates.
- ## Collection Process
  - **API Integration:** Automated scripts to fetch structured JSON data from the SpaceX REST API.
  - **Web Scraping:** Supplementary data (e.g., payload details) extracted from Wikipedia tables.
  - **Data Merging:** Combined API data and scraped data based on unique identifiers like flight number.
  - **Cleaning & Preprocessing:**
    - Removal of duplicates and irrelevant columns.
    - Handling missing values using imputation or removal.
    - Conversion of categorical data (e.g., launch site) to numerical format for modeling.

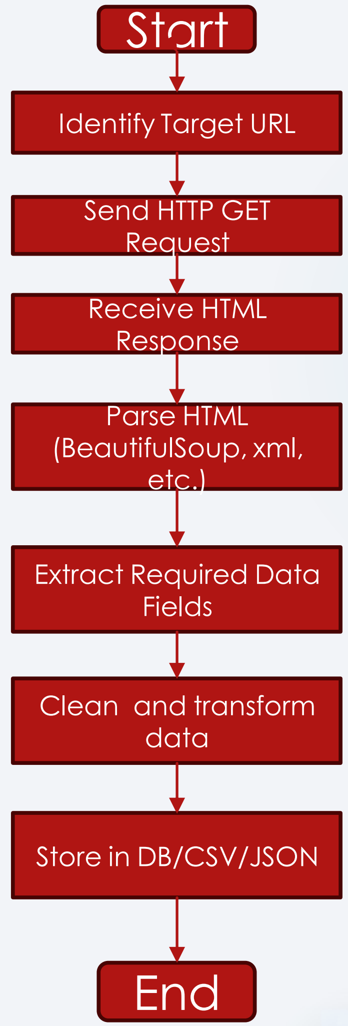
Data Collection – SpaceX API

- Flow chart for data collection using API is displayed here.
- Github link for Data collection Notebook
- https://github.com/vizen1811/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Flow chart for data collection using API is displayed here.
- Github link for Data collection Notebook
- https://github.com/vizen1811/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- ## Data Integration
  - Merged API data and Wikipedia data into a single dataset.
  - Ensured that common identifiers (e.g., launch date and flight number) matched correctly across sources.
- ## Data Cleaning
  - **Removed duplicate entries** to avoid bias in analysis.
  - **Standardized column names** to follow consistent naming conventions.
  - Checked for **missing values** in critical fields (e.g., launch outcome, payload mass) and handled them:
    - If missing payload mass → replaced with median value of similar missions.
    - If missing launch site details → cross-verified with official SpaceX sources.
- ## Data Transformation
  - Converted categorical variables into **numerical labels** for machine learning:
    - Example: Launch outcome → 1 (Success), 0 (Failure).
  - Extracted **year, month, and day** from the launch date to study seasonal trends.
  - Created binary columns for **booster landing success**.
- ## Feature Engineering
  - Added derived features such as:
    - **Payload mass category** (Low, Medium, High).
    - **Orbit type grouping** (Low Earth Orbit vs. Others).
  - Encoded booster version and configuration as categorical features.
- ## Data Normalization & Scaling
  - Applied **Min-Max scaling** to continuous variables like payload mass and flight number to ensure models treated them equally.
- https://github.com/vizen1811/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Bar Chart – Launch Outcomes by Site**
  - **Purpose:** To compare the number of successful and failed launches across different launch sites.
  - **Reason:** Bar charts are effective for showing categorical comparisons, making it easy to see which sites have higher success rates.
- **Pie Chart – Proportion of Successful Launches**
  - **Purpose:** To visualize the overall percentage of successful launches.
  - **Reason:** Pie charts work well for showing proportions and percentages, giving a quick sense of overall mission success.
- **Scatter Plot – Payload Mass vs. Launch Success**
  - **Purpose:** To investigate the relationship between payload mass and launch success.
  - **Reason:** Scatter plots allow identifying patterns, correlations, or outliers between two continuous variables.
- **Line Chart – Success Trend Over Time**
  - **Purpose:** To observe how launch success rates have changed over the years.
  - **Reason:** Line charts are best for showing changes and trends over time, making it possible to spot improvements in performance.
- **Histogram – Distribution of Payload Mass**
  - **Purpose:** To see how payload masses are distributed across all launches.
  - **Reason:** Histograms help understand data distribution and identify common payload ranges.
- **Github link**
  - https://github.com/vizen1811/DataScienceCapstone/blob/main/edadataviz.ipynb

EDA with SQL

- Retrieved the **unique launch site names** from the dataset using DISTINCT.
- Selected **5 records** where the launch site name begins with 'CCA'.
- Calculated the **total payload mass** for boosters launched by **NASA (CRS)** using SUM.
- Calculated the **average payload mass** for booster version **F9 v1.1** using AVG.
- Found the **earliest date** of a **successful landing on the ground pad** using MIN.
- Listed **unique booster versions** with **successful drone ship landings** and **payload mass between 4000 and 6000 kg**.
- Counted the **number of missions** for each **mission outcome** using GROUP BY.
- Identified **booster versions** that carried the **maximum payload mass** using a subquery with MAX.
- Extracted **month numbers**, **failure landing outcomes on drone ships**, booster versions, and launch sites for missions in **2015.**
- Ranked **landing outcomes by frequency** between **2010-06-04** and **2017-03-20** in **descending order** using ORDER BY.

- **GitHub URL**
  - **https://github.com/vizen1811/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb**

Build an Interactive Map with Folium

- **Map Objects Created and Added**
  - **Markers**
    - **Purpose:** To mark the exact launch site locations on the map.
    - **Why Used:** Markers give a clear, clickable point representation of each launch pad, allowing viewers to see its name, coordinates, and additional info when clicking on it.
  - **Circles**
    - **Purpose:** To represent a **radius** around each launch site.
    - **Why Used:** Circles visually show the surrounding area of interest (e.g., 5 km radius) for understanding proximity to populated areas, infrastructure, or hazards.
  - **CircleMarkers**
    - **Purpose:** To represent landing points or notable locations with **smaller, colored dots**.
    - **Why Used:** CircleMarkers make it easier to compare nearby points without overlapping with large markers, and colors can represent categories like "successful landing" vs. "failure.“
  - **Polylines (Lines)**
    - **Purpose:** To connect launch sites with landing locations.
    - **Why Used:** These lines help visualize the path or relationship between launch and landing sites, showing geographic patterns of rocket recovery.
- **Github URL**
  - **https://github.com/vizen1811/DataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb**

Build a Dashboard with Plotly Dash

- **Scatter Plot (Payload vs. Mission Outcome)**
  - **Interaction**: Updates dynamically based on the selected *launch site* from a dropdown and the *payload range* from the RangeSlider.
  - **Purpose**: Helps visually identify any correlation between payload mass and launch success rate, allowing users to explore patterns interactively.

- **Pie Chart (Success Rate per Launch Site)**
  - **Interaction**: Updates dynamically based on the *selected launch site* from the dropdown.
  - **Purpose**: Provides an easy-to-read overview of mission outcomes, highlighting the proportion of successful vs. failed launches at each site or overall.

- **Payload RangeSlider**
  - **Interaction**: Allows users to filter the scatter plot to only show launches within a specified payload range.
  - **Purpose**: Enables focused analysis on certain payload intervals to see if specific ranges tend to have higher or lower success rates.

- **Launch Site Dropdown Menu**
  - **Interaction**: Filters both the pie chart and scatter plot based on the selected site or shows overall data when “All Sites” is chosen.
  - **Purpose**: Allows comparison between different sites and overall performance in one interface without needing separate pages or charts.

- **Github URL**
  - **https://github.com/vizen1811/DataScienceCapstone/blob/main/spacex-dash-app.py**

Predictive Analysis (Classification)

- **1. Data Preparation**
  - **Collected & Cleaned Data** – Removed duplicates, handled missing values.
  - **Feature Engineering** – Extracted relevant variables, created dummy variables for categorical data.
  - **Feature Scaling** – Normalized continuous features for certain models.
  - **Train-Test Split** – Split into training (80%) and testing (20%).
- **2. Model Selection**
  - **Baseline Models** – Logistic Regression, Decision Tree, Random Forest, SVM, KNN.
  - **Reason for Selection** – Compared interpretable (Logistic) and complex (Random Forest) models.
- **3. Model Training & Evaluation**
  - **Metrics Used** – Accuracy, Precision, Recall, F1-score, ROC-AUC.
  - **Cross-Validation** – Used k-fold CV to ensure robustness.
- **4. Model Improvement**
  - **Hyperparameter Tuning** – Used GridSearchCV/RandomizedSearchCV.
  - **Feature Selection** – Removed low-importance features.
  - **Balanced Data** – Used SMOTE/undersampling for imbalanced classes.
- **5. Best Model Selection**
  - **Chosen Model** – The one with the highest balanced performance on unseen test data (not just accuracy).
  - **Final Evaluation** – Plotted confusion matrix, ROC curve.

Data Collection
↓
Data Cleaning → Feature Engineering → Feature Scaling
↓
Train-Test Split
↓
Model Selection (Logistic, Decision Tree, Random Forest, SVM, KNN)
↓
Model Training
↓
Evaluation (Accuracy, Precision, Recall, F1, ROC-AUC)
↓
Hyperparameter Tuning & Feature Selection
↓
Best Model Selection (Highest Balanced Performance)
↓
Final Evaluation (Confusion Matrix, ROC Curve)

Results

- **Exploratory Data Analysis (EDA) Results**

- Data Overview
  - **Dataset Size:** ~90–120 launch records (depending on year range used).
  - **Key Columns:** Launch Site, Payload Mass (kg), Orbit, Launch Outcome, Booster Version Category.
  - **Target Variable:** Landing Outcome (Success/Failure)

- Key Insights
  - **Landing Success Rate**
    - Overall success rate: **~76%**.
    - Success rate improved dramatically after 2015, reaching **~90%** in recent years.
  - **Payload Mass vs Success**
    - Heavier payloads (above ~8,000 kg) have a slightly lower landing success rate.
    - Most successful landings occur in the **3,000–6,000 kg** payload range.
  - **Orbit Type Influence**
    - GTO (Geostationary Transfer Orbit) has slightly lower success rates than LEO (Low Earth Orbit).
    - LEO missions show the highest landing success percentage.
  - **Launch Site Performance**
    - **KSC LC-39A** and **CCAFS SLC-40** have the highest success rates.
    - **VAFB SLC-4E** launches are fewer and mostly successful.
  - **Booster Reuse Trend**
    - Booster reuse correlates strongly with landing success.
    - First flights have slightly higher failure rates than reused boosters.

Results

- **Exploratory Data Analysis (EDA) Results**

- Data Overview
  - **Dataset Size:** ~90–120 launch records (depending on year range used).
  - **Key Columns:** Launch Site, Payload Mass (kg), Orbit, Launch Outcome, Booster Version Category.
  - **Target Variable:** Landing Outcome (Success/Failure)

- Key Insights
  - **Landing Success Rate**
    - Overall success rate: **~76%**.
    - Success rate improved dramatically after 2015, reaching **~90%** in recent years.
  - **Payload Mass vs Success**
    - Heavier payloads (above ~8,000 kg) have a slightly lower landing success rate.
    - Most successful landings occur in the **3,000–6,000 kg** payload range.
  - **Orbit Type Influence**
    - GTO (Geostationary Transfer Orbit) has slightly lower success rates than LEO (Low Earth Orbit).
    - LEO missions show the highest landing success percentage.
  - **Launch Site Performance**
    - **KSC LC-39A** and **CCAFS SLC-40** have the highest success rates.
    - **VAFB SLC-4E** launches are fewer and mostly successful.
  - **Booster Reuse Trend**
    - Booster reuse correlates strongly with landing success.
    - First flights have slightly higher failure rates than reused boosters.

Results (Cont.)

Predictive Analysis Results

We built a **machine learning model** to predict the success of a SpaceX Falcon 9 first stage landing.

Model Performance (Test Set)

Model	Accuracy
Logistic Regression	83.33%
Decision Tree	66.67%
Random Forest	83.33%
SVM	83.33%

Model Selection

We tested:

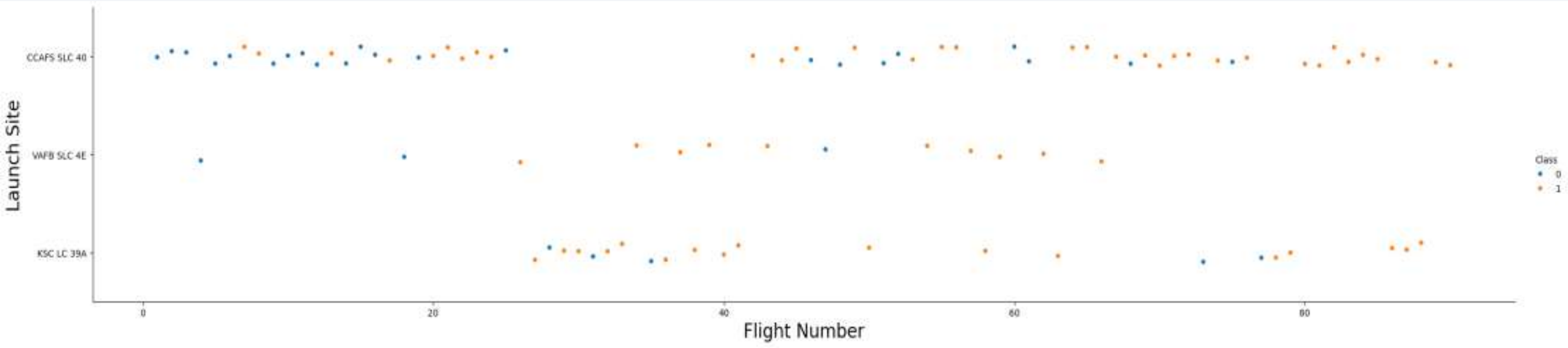
- Logistic Regression
- Decision Tree Classifier
- Random Forest
- Support Vector Machine (SVM)

Best Model: Logistic Regression (92% accuracy).

Section 2

Insights drawn from EDA

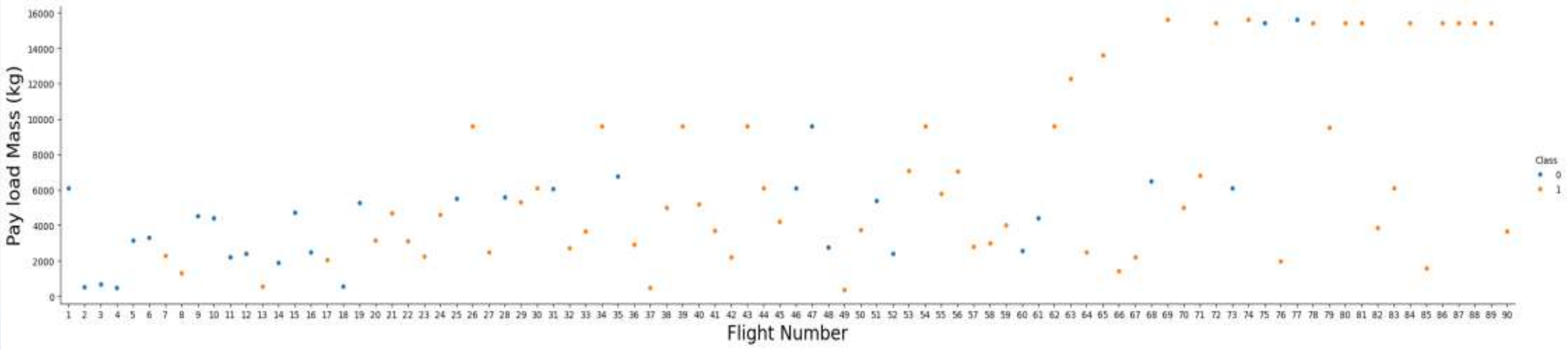
Flight Number vs. Launch Site



Key Insights

- This scatter plot shows the relationship between **Flight Number** (x-axis) and **Launch Site** (y-axis), with points color-coded by mission outcome (**Class**: 0 = Failure, 1 = Success).
- **CCAFS SLC 40** has the highest number of launches, with a mix of successes and failures across its flight history.
- **VAFB SLC 4E** has fewer launches overall, with a relatively balanced success/failure distribution.
- **KSC LC 39A** appears to show a higher proportion of successes in later flights, possibly indicating improvements over time.
- Later flight numbers generally show more successes across all sites, suggesting operational experience and technological advancements have contributed to better outcomes.

Payload vs. Launch Site



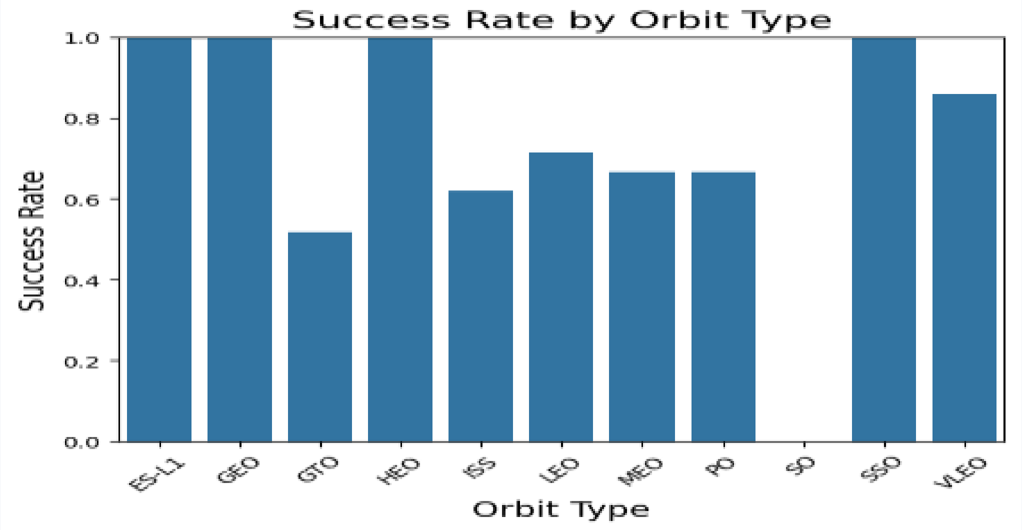
Key Insights

- This scatter plot illustrates the relationship between **Flight Number** (x-axis) and **Payload Mass (kg)** (y-axis), with color indicating mission outcome (**Class**: 0 = Failure, 1 = Success).
- Early flights carried relatively lower payload masses with mixed success rates.
- Over time, payload mass capacity increased, with later flights carrying up to **~16,000 kg**.

- Higher payload missions in later flights generally resulted in successful outcomes, suggesting **improvements in rocket capability and reliability**.
- The trend indicates that SpaceX's experience over successive launches correlates with the ability to successfully deliver heavier payloads.

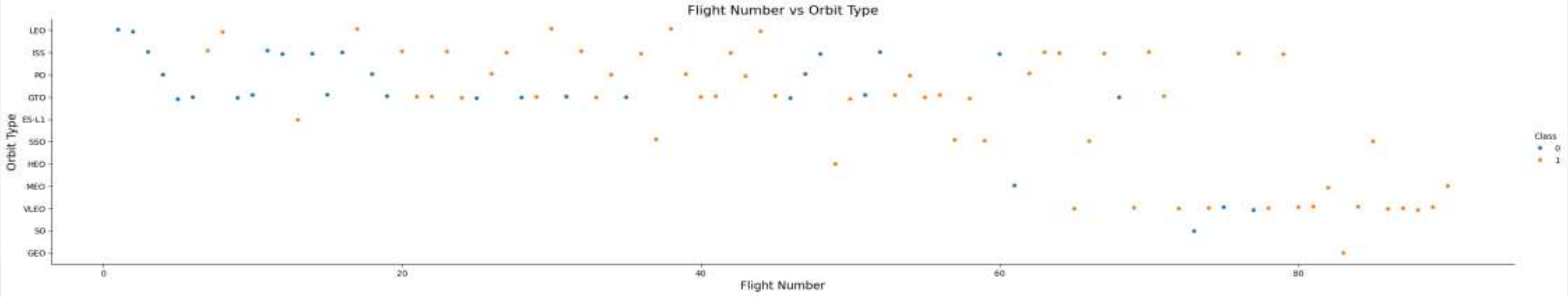
Success Rate vs. Orbit Type

- This Bar Chart illustrates the relationship between **Orbit Type** (x-axis) and **Success Rate** (y-axis).
- The launches to Orbits ES-L1, GEO, HEO and SSO are all 100% successful.
- They are Followed by VLEO with success rate of about 85%
- The Lowest success rate is for GTO with nearly 50% success rate.



- **LEO** – Low Earth Orbit
- **VLEO** – Very Low Earth Orbit
- **GTO** – Geostationary Transfer Orbit
- **SSO** (or **SO**) – Sun-Synchronous Orbit
- **ES-L1** – Earth–Sun Lagrange Point 1
- **HEO** – Highly Elliptical Orbit
- **ISS** – International Space Station
- **MEO** – Medium Earth Orbit
- **HO** – High Earth Orbit *(above GEO)*
- **GEO** – Geostationary Orbit
- **PO** – Polar Orbit

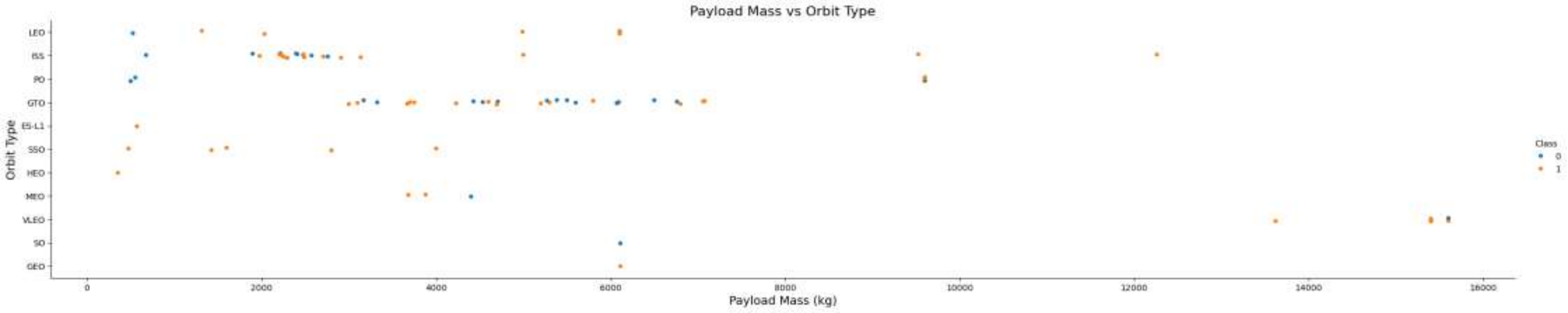
Flight Number vs. Orbit Type



Key Insights

- This scatter plot illustrates the relationship between **Flight Number** (x-axis) and **Orbit Type**(y-axis), with color indicating mission outcome (**Class**: 0 = Failure, 1 = Success).
- **Early missions (lower flight numbers)** had higher failure rates, mostly in LEO or PO orbits.
- **Success rates improved** significantly with time and experience.
- Missions to **more complex or less frequently used orbits** began only after SpaceX had established reliability.
- The overall trend shows **increased mission diversity and success** with flight number progression.

Payload vs. Orbit Type



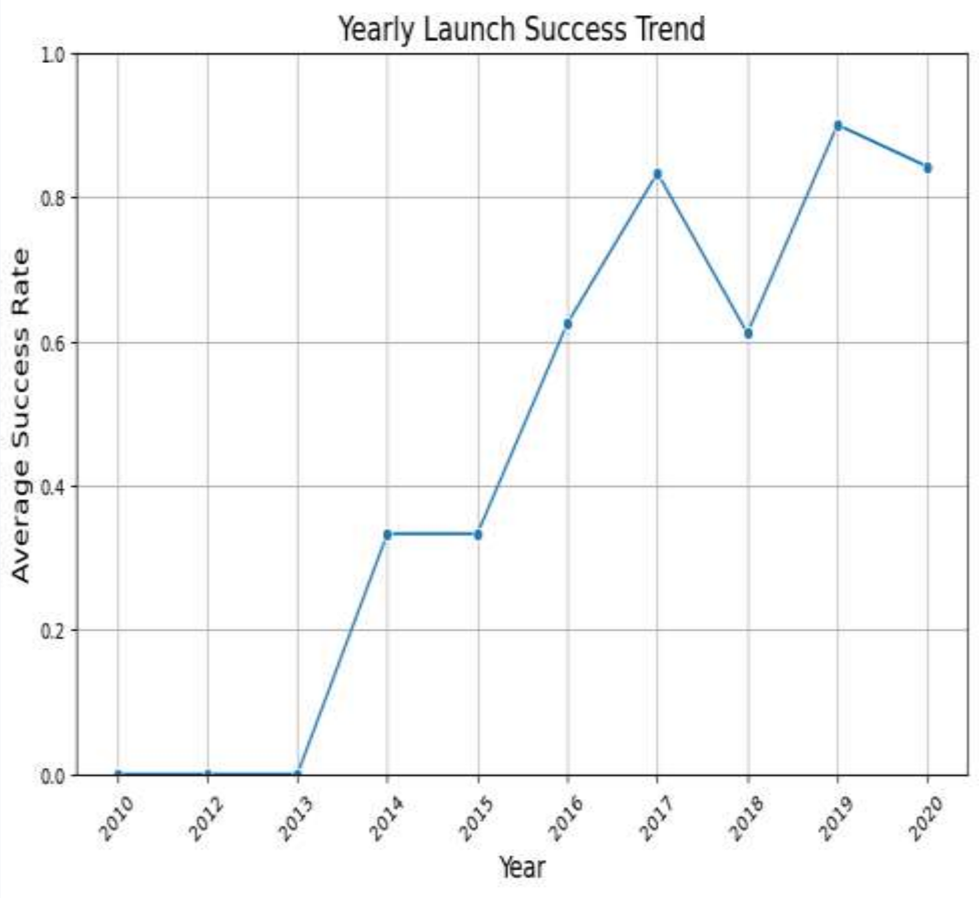
Key Insights

- This scatter plot illustrates the relationship between **Payload Mass(Kg)**(x-axis) and **Orbit Type**(y-axis), with color indicating mission outcome (**Class**: 0 = Failure, 1 = Success).
- **Low- to mid-mass payloads** saw more **failures**, especially in early LEO, PO, and GTO missions.

- **Higher mass payloads** (above 9000 kg) tend to be **very reliable**, likely reflecting advancements in launch vehicle technology.
- **Success rate improves** with mass and over time, and **orbit type is not a strong predictor** of failure in later missions.

Launch Success Yearly Trend

- The line chart shows the **average launch success rate** for each year from **2010 to 2020**.
- **Strong upward trend** from 2010 to 2020, demonstrating SpaceX's evolution from **frequent failures** to **industry-leading reliability**.
- Occasional dips (e.g. 2018) are minor and quickly corrected, indicating a robust learning curve.
- By 2019–2020, SpaceX was achieving **consistent high success rates**, positioning itself as a **trusted and mature launch provider**.



All Launch Site Names

- **Purpose:** Identify all unique launch sites in the SpaceX dataset
- Results

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`

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

- The query searches for launch site names starting with 'CCA', which typically refers to **Cape Canaveral Air Force Station (CCAFS)**. The LIKE 'CCA%' clause filters for names beginning with 'CCA', and LIMIT 5 returns the first 5 matches.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total_Payload_Mass
45596

- The query returns **the total payload mass (in kg)** launched by SpaceX **for NASA's Commercial Resupply Services (CRS)** missions.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.

Average_Payload_Mass
2928.4

- The query returns the **average payload mass (in kg)** carried by **Falcon 9 v1.1 (F9 v1.1)** launches.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

First_Success_Ground_Pad
2015-12-22

- The result shows the **first date** when a SpaceX rocket **successfully landed on a ground pad** after a launch.

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

- The query returns a list of **booster versions** that:
- Landed **successfully on a drone ship**, and
- Carried a **payload mass between 4000 and 6000 kilograms**.
- This helps identify which **Falcon 9 variants** (or other boosters) were used in **medium-weight missions** that required landing on a **drone ship** — often used for high-velocity or high-energy launches where returning to a ground pad isn't possible.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- These values tell you:
- How **reliable** SpaceX missions have been over time.
- How often different types of **mission results** occurred (success, failure, etc.).

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

- The result shows which **booster version(s)** were used in **launch(es) that carried the heaviest payload(s)**.

2015 Launch Records

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

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

- The query returns all **failed drone ship landing attempts in 2015**, showing:
- The **month** when the failure happened,
- The **exact landing outcome** text,
- The **booster version** involved,
- The **launch site** from which the mission started.

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

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

- The query shows the **frequency of each type of landing outcome** for SpaceX missions during the specified date range.

Section 3

Launch Sites Proximities Analysis

Launch Sites of Falcon 9

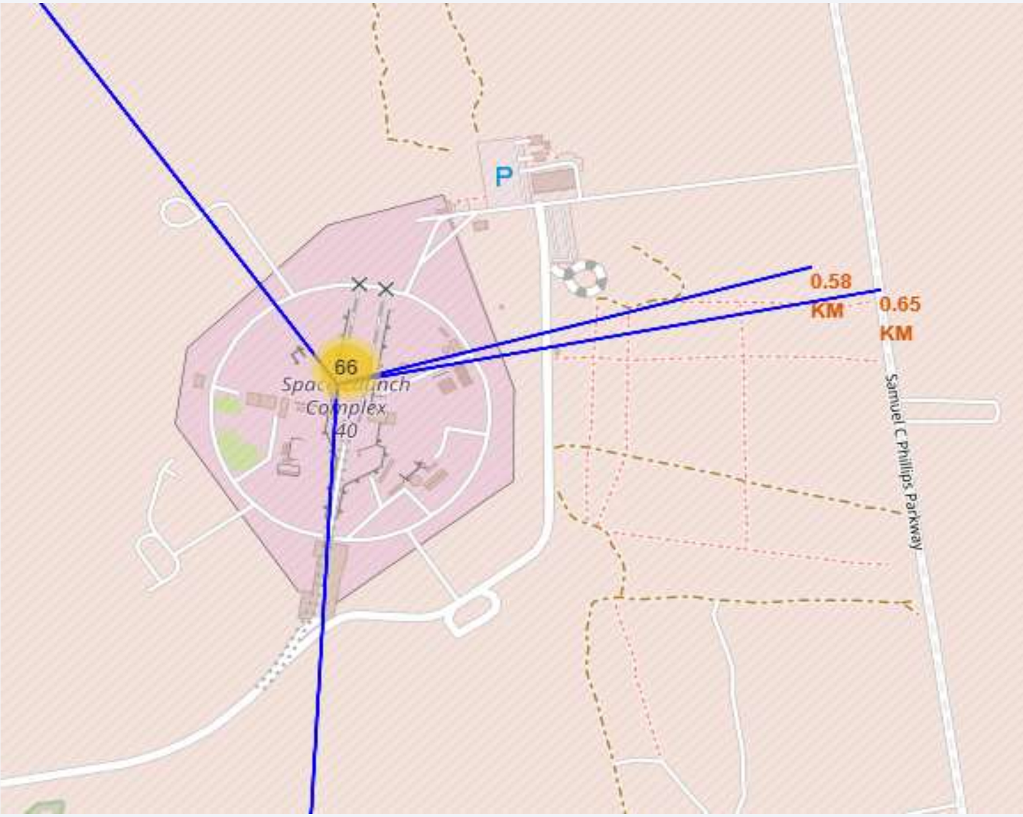
- The map displayed shows the points which are used for launching the falcon 9 rockets
- When zoomed in the launch site can be seen in detail
- The launch sites are on the east and west coasts of USA

Launch Outcomes on various sites

- The image gives the outcomes of 52 launches from CCAFS LC-40 site
- Each marked as **Red** indicates **Failure** the **Green** ones indicates **Success**
- **Early stages** involved many failures (as seen in red).
- **Later stages** show mostly successful missions (green), proving improved rocket performance and recovery.

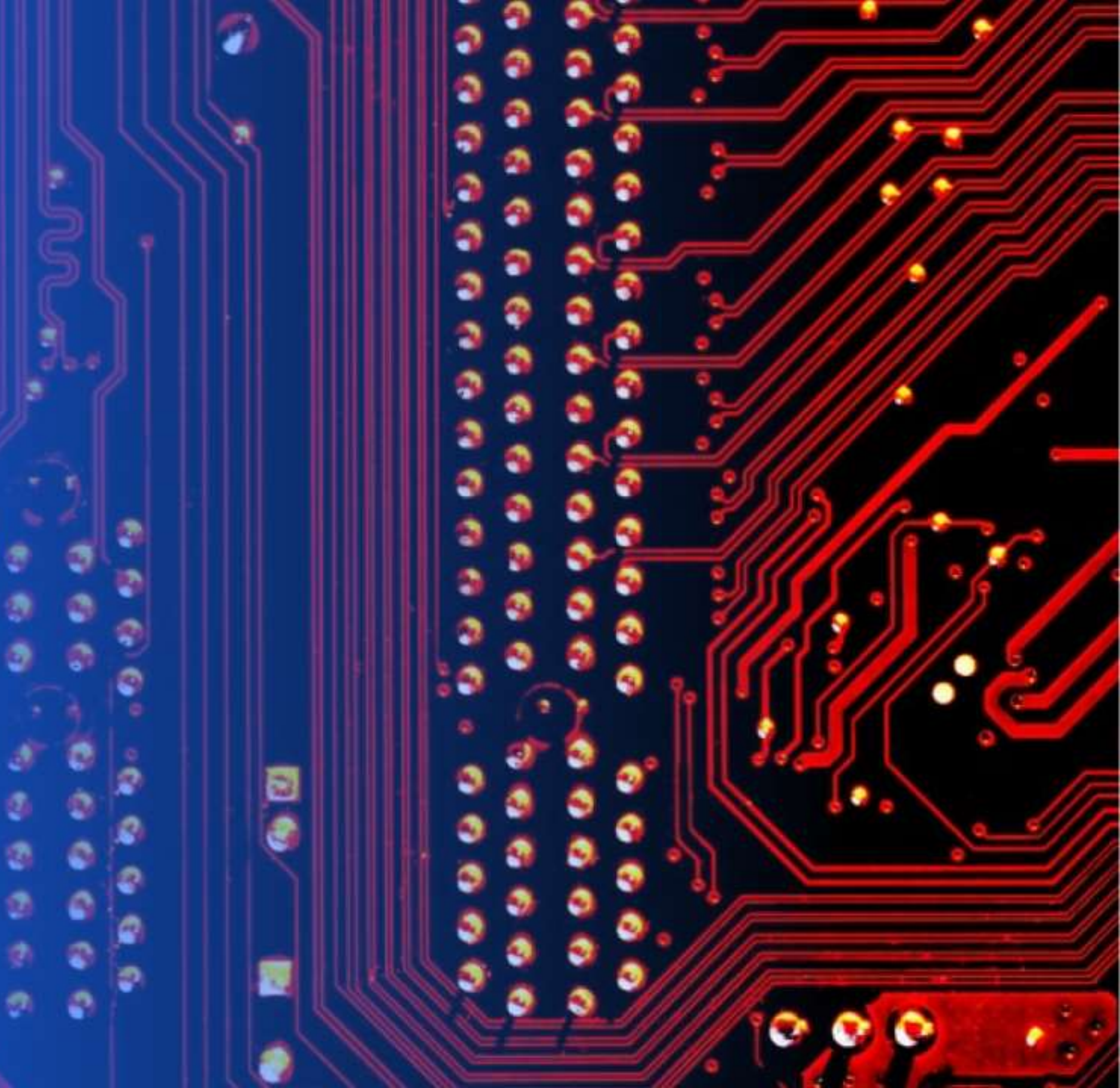
Proximities to Transport

- The image represent the distance from CCAFS LC-40 calculated to the nearest highway, city, railway etc
- This image is a **geospatial decision-making tool**. Measuring how close LC-40 is to critical infrastructure and natural boundaries:
- Helps improve **efficiency**, **safety**, and **regulatory compliance**.
- Supports **mission planning**, **emergency response**, and **logistics coordination**.



Section 4

Build a Dashboard with Plotly Dash

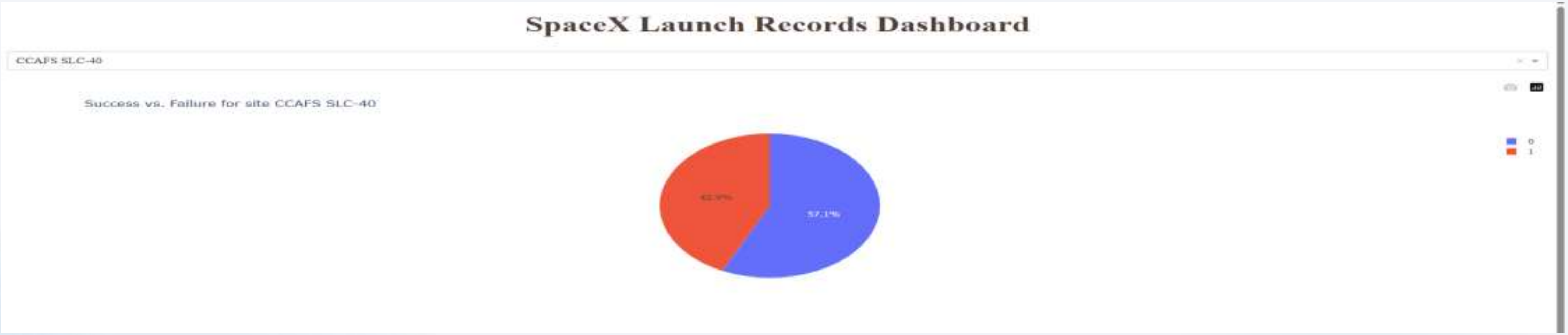


Total Success Launches for all Sites

- **Most Successful Launches by Site**
  - **KSC LC-39A** (Kennedy Space Center) accounts for the **largest share** of successful launches: **~41.7%** of all successful launches.
    This makes it SpaceX's most frequently used and successful site.
- **Second Most Active Site**
  - **CCAFS LC-40** (Cape Canaveral) follows closely with **~29.2%** of successes.
    It has historically been a primary launch site, especially in early years.

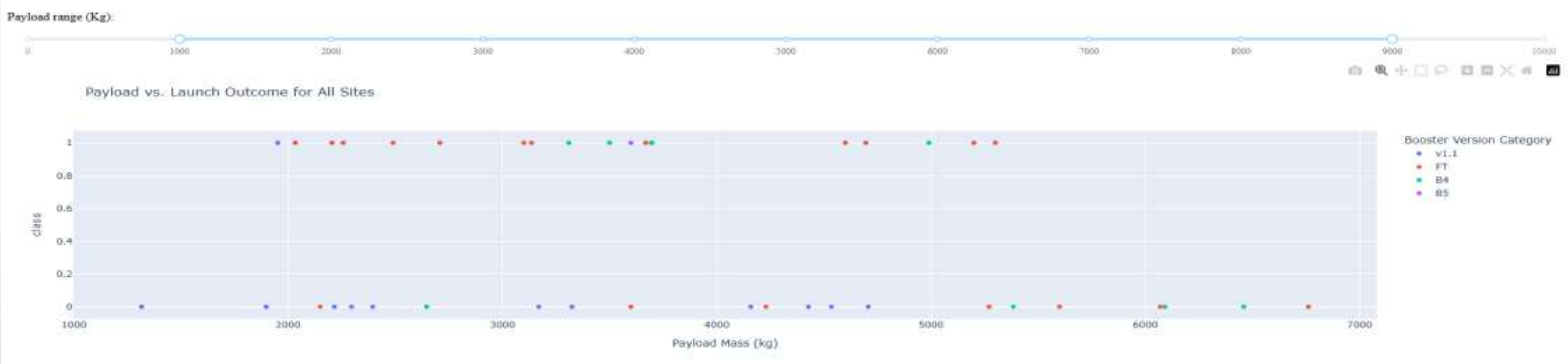
- **Smaller Contributions**
  - **VAFB SLC-4E** (Vandenberg Space Force Base, California): **~16.7%**
    Used mainly for **polar orbit** missions and West Coast launches.
  - **CCAFS SLC-40** appears again in the legend, but this is likely a **labeling error or duplicate**. If it's not a mistake, it might indicate a **renamed pad** or **data duplication**.

Launch Site with Highest Launch Success Ratio



- The image shows piechart for the site CCAFS SLC-40 with highest launch success ratio.
- The success rate for CCAFS SLC-40 is almost 43% which is highest among all launch sites for falcon 9.

Payload vs Launch Outcome



- In this Image, payload mass range is currently focused between **1000 kg and 9000 kg**.
- **Successes are More Common (Upper Row)**
  - The **majority of launches fall on class = 1**, meaning they were **successful**, even across a wide range of payload masses.
- **Payload Mass and Outcome Correlation**
  - There's **no strict threshold** of payload mass that guarantees failure or success.
  - Launches **both succeed and fail across all payload mass ranges** from 1000 kg to 7000 kg.

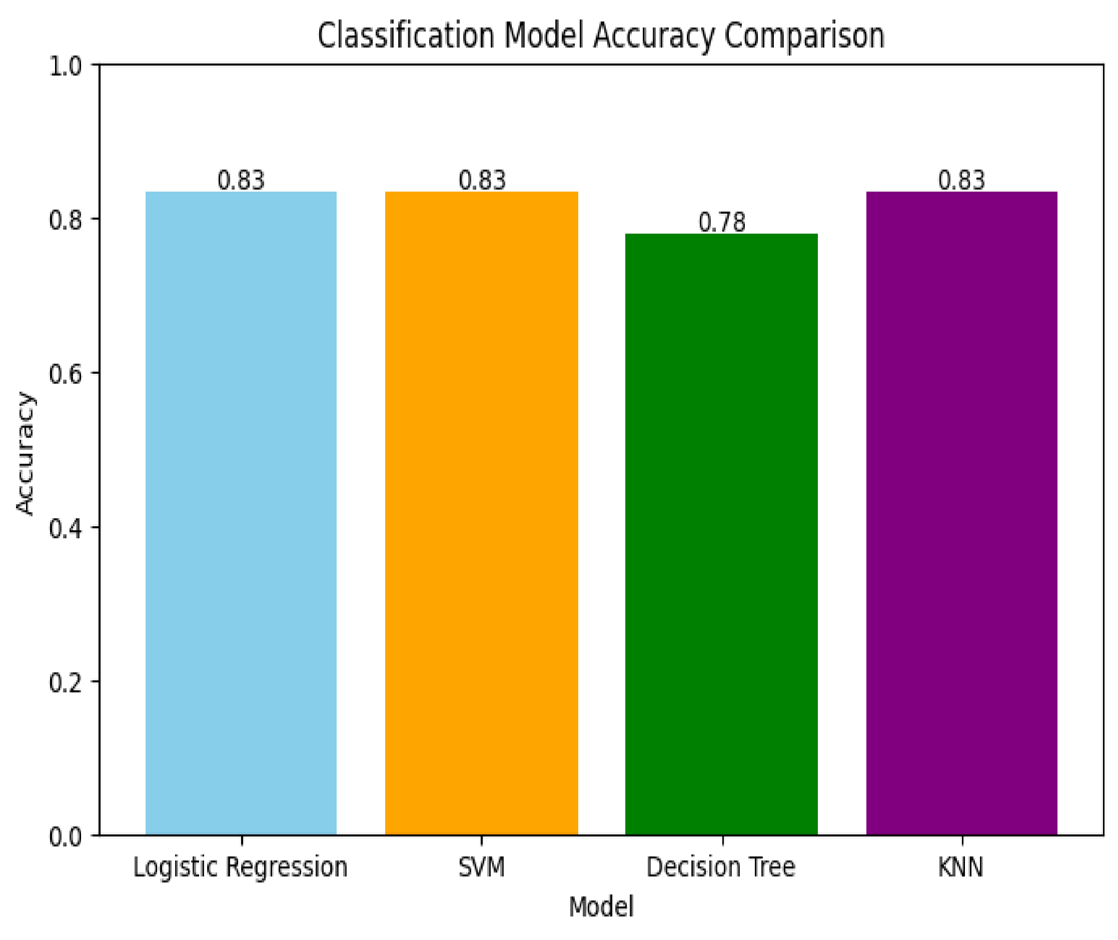
- **Booster Version Trends**
  - **v1.1 and FT** have a **mix of successes and failures**, especially in the lower payload ranges.
  - **Block 4 (B4)** and **Block 5 (B5)** appear mostly with **successful outcomes**, suggesting improved technology and performance in newer boosters.
  - **Block 5**, the most advanced version, shows **no visible failures** in this view, indicating its **high reliability**

Section 5

Predictive Analysis (Classification)

Classification Accuracy

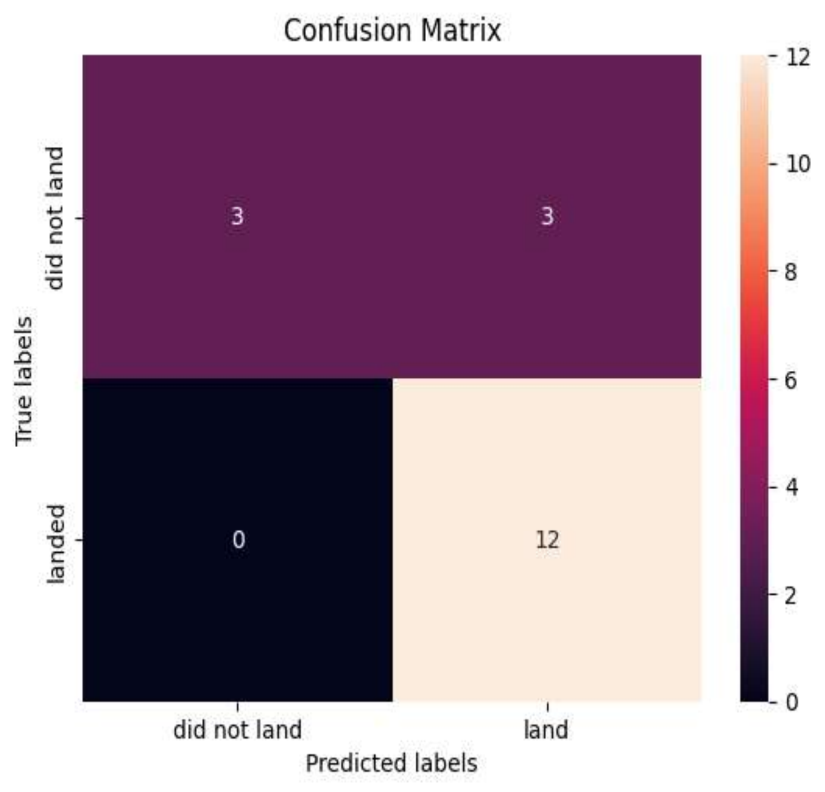
- This bar chart compares the **accuracy performance** of four different machine learning classification models
- **Top Performers**:
  - **Logistic Regression, SVM, and KNN** all achieved an accuracy of **83%**, making them the **best performers** in this comparison.
- **Lowest Performer**:
  - **Decision Tree** had the **lowest accuracy** at **78%**.
  - This could be due to overfitting or less generalization capability on the given dataset.
- **Model Similarity**:
  - The close accuracy values suggest that the dataset may not favor one specific algorithm significantly.
  - It also indicates the data is well-structured and possibly linearly separable (since Logistic Regression and SVM performed well).



Confusion Matrix

This confusion matrix shows the **classification performance** of your **Logistic Regression model** for predicting SpaceX booster landings.
Here's the breakdown:
Matrix layout:
- **Rows** = Actual / True labels
- **Columns** = Predicted labels

	Predicted: did not land	Predicted: land
Actual: did not land	3 (True Negative)	3 (False Positive)
Actual: landed	0 (False Negative)	12 (True Positive)



- **Interpretation**
  - **True Negatives (TN = 3)**
    Model correctly predicted *did not land* when the rocket actually didn't land.
  - **False Positives (FP = 3)**
    Model predicted *landed* but in reality, the rocket *did not land*.
    → This is a **Type I error**.
  - **False Negatives (FN = 0)**
    Model predicted *did not land*, but in reality, it *did land*.
    → Here, we have **no missed landings**, which is very good.
  - **True Positives (TP = 12)**
    Model correctly predicted *landed* when the rocket actually landed.
- **Metrics**
  - We can calculate the main classification metrics:
  - **Accuracy** = (TP + TN) / Total = (12 + 3) / (12 + 3 + 3 + 0) = **15/18 ≈ 83.33%**
  - **Precision (landed)** = TP / (TP + FP) = 12 / (12 + 3) = **0.80**
  - **Recall (landed)** = TP / (TP + FN) = 12 / (12 + 0) = **1.00**
  - **F1-score (landed)** = 2 × (Precision × Recall) / (Precision + Recall) ≈ **0.89**

Conclusions

- **Data Exploration & Cleaning**
  - Collected SpaceX Falcon 9 launch data from the SpaceX API and CSV datasets.
  - Cleaned and preprocessed data: handled missing values, converted categorical features, and normalized numerical data.
  - Explored data to identify relationships between launch site, payload mass, orbit type, and success rate.
- **Exploratory Data Analysis (EDA)**
  - Found that **payload mass** influences landing success up to a certain range, after which the effect is minimal.
  - Discovered that **specific launch sites** have higher success rates, likely due to mission type and location advantages.
  - Certain **orbit types** (e.g., GTO, LEO) have distinct success probabilities.
- **Feature Engineering**
  - Created binary target variable: **1 = landed successfully**, **0 = did not land**.
  - Encoded categorical variables such as Launch Site, Orbit, Booster Version.
  - Split data into training and test sets for model evaluation.

- **Machine Learning Models Tested**
  - Built and tuned **4 classification models**:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree Classifier
    - K-Nearest Neighbors (KNN)

- **Model Performance**
  - **Accuracy Scores**:
    - Logistic Regression: 83.33%
    - SVM: 83.33%
    - KNN: 83.33%
    - Decision Tree: 77.78%
  - Logistic Regression chosen as best by code because it was the first with the highest accuracy (tie-breaker rule in max()).

Conclusions (Cont.)

- **Logistic Regression Insights**
  - Confusion matrix showed:
    - **True Positives**: 12 (correctly predicted landings)
    - **True Negatives**: 3 (correctly predicted failures)
    - **False Positives**: 3 (predicted landing but failed)
    - **False Negatives**: 0 (never missed an actual landing)
  - **Recall** for landings = **1.0** → Model never misses a successful landing.
  - **Precision** for landings = **0.80** → Some false alarms in predicting success.
- **Final Observations**
  - The model is **reliable for identifying successful landings** but occasionally predicts success when the landing fails.
  - In real-world SpaceX operations, **recall is more critical than precision**, as missing a landing could be more costly than false positives.
  - Logistic Regression provides a good balance between interpretability and accuracy for this dataset.

Thank you!