Winning Space Race with Data Science

Christopher Prior
21/11/2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection via SpaceX REST API and web scraping
- Data wrangling and preprocessing
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models (Logistic Regression, SVM, Decision Tree, KNN)

Summary of all results

- Successfully collected and processed SpaceX launch data

- Identified key factors affecting landing success (Flight Number, Payload Mass, Orbit Type, Launch Site)

- Built and evaluated multiple classification models

- Achieved best model accuracy - [Will be determined after running the models]

Introduction

Project background and context

- SpaceX advertises Falcon 9 rocket launches at $62 million vs. competitors at $165+ million
- Cost savings come from reusing the first stage
- Predicting first stage landing success helps determine launch cost
- This information is valuable for competitive bidding

Problems you want to find answers to

- Can we predict if the Falcon 9 first stage will land successfully?
- What factors influence landing success?
- Which machine learning model performs best for this prediction?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology -

- Collected data from SpaceX REST API
- Web scraped additional launch records from Wikipedia

Perform data wrangling

- Data Quality Assessment - Identified and calculated percentage of missing values (e.g., LandingPad had 28.89% missing values)
- Data Type Identification - Classified columns as numerical (FlightNumber, PayloadMass, Flights, etc.) or categorical (BoosterVersion, Orbit, LaunchSite, etc.)

Perform exploratory data analysis (EDA) using visualization and SQL

- Analyzed relationships between variables using visualizations
- Performed SQL queries to extract insights
- Identified patterns in launch success rates

Perform interactive visual analytics using Folium and Plotly Dash

- Created interactive maps showing launch sites
- Built dashboard with interactive visualizations

Perform predictive analysis using classification models

- Built and tuned multiple classification models

- Evaluated model performance using cross-validation

- Selected best performing model based on test accuracy

Data Collection

SpaceX REST API

Collected launch data including rocket information, payloads, launchpads, and core landing outcomes from Space X API Rest Endpoints

Web Scraping -

Extracted historical launch records from Wikipedia page "List of Falcon 9 and Falcon Heavy launches“, using beautiful soup to parse the html

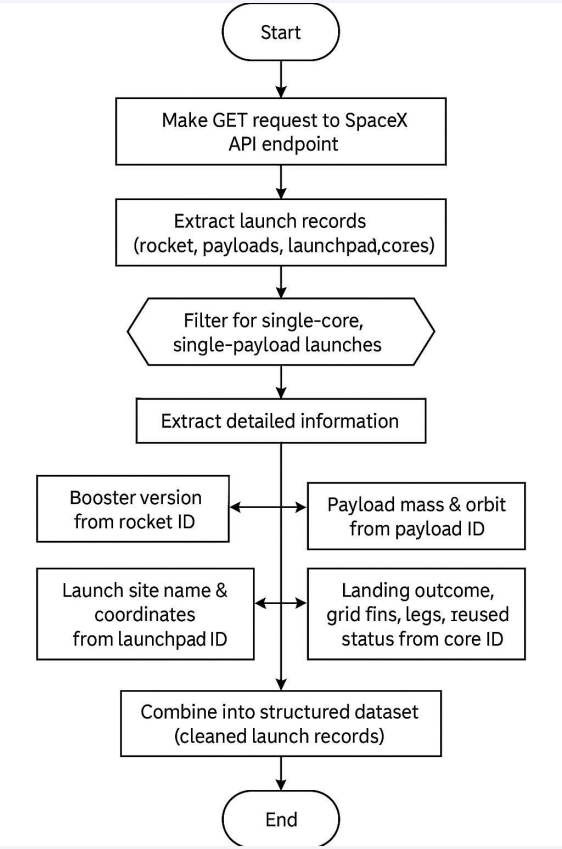
Data Sources -

Rocket information (booster version, block, serial)

- Payload data (mass, orbit type)
- Launch site information (name, coordinates)
- Landing outcomes (success/failure, landing type)

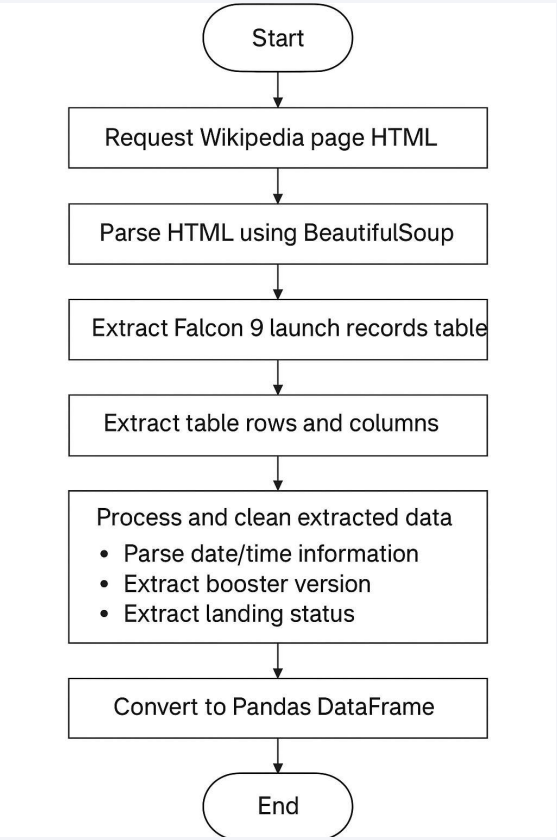
Data Collection – SpaceX API

https - //github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%201/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

- https - //github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%202/jupyter-labs-webscraping.ipynb



Data Wrangling

Data Cleaning

- Removed rows with missing critical data
- Filtered for single-core, single-payload launches
- Converted date formats to datetime
- Handled missing values

Data Transformation

- Extracted nested JSON data from API responses
- Created binary classification column (Class - 1 for success, 0 for failure)
- Standardized categorical variables
- Created dummy variables for categorical features

Feature Engineering

- Selected relevant features - FlightNumber, PayloadMass, Orbit, LaunchSite, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial
- Encoded categorical variables
- Standardized numerical features

https -//github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%203/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Charts plotted and why**
- 1. Flight Number vs. Payload Mass - To identify relationship between launch experience and payload capacity
- 2. Flight Number vs. Launch Site - To see if success rate improves with experience at each site
- 3. Payload Mass vs. Launch Site - To understand payload capacity differences across sites
- 4. Success Rate by Orbit Type (Bar Chart) - To identify which orbits have highest success rates
- 5. Flight Number vs. Orbit Type - To see if experience affects success differently by orbit
- 6. Payload Mass vs. Orbit Type - To understand payload constraints by orbit
- 7. Yearly Success Trend (Line Chart) - To observe improvement in success rate over time
- **Key Findings**
- Success rate increases with flight number (more experience = better success)
- VAFB-SLC has no heavy payload launches (>10,000 kg)
- LEO orbit shows relationship between flight number and success
- Success rate has been increasing since 2013

https -//github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%204/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

SQL Queries Performed

- 1. Unique Launch Sites - `SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE`
- 2. Launch Sites Beginning with 'CCA' - `SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5`
- 3. Total Payload Mass for NASA (CRS) - `SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'`
- 4. Average Payload Mass for F9 v1.1 - `SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'`
- 5. First Successful Ground Landing - `SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'`
- 6. Successful Drone Ship with Payload 4000-6000 - `SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000`
- 7. Total Success/Failure Count - `SELECT "Mission_Outcome", COUNT(*) as Count FROM SPACEXTABLE GROUP BY "Mission_Outcome"`
- 8. Maximum Payload Boosters - `SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)`
- 9. 2015 Failure Records - `SELECT substr("Date", 6, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr("Date", 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)%'`
- 10. Ranked Landing Outcomes (2010-2017) - `SELECT "Landing_Outcome", COUNT(*) as Count FROM SPACEXTABLE WHERE "Date" >= '2010-06-04' AND "Date" <= '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC`

https://github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%205/edadataviz.ipynb

Build an Interactive Map with Folium

Map objects created

- Launch Site Markers: Added markers for all launch sites (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
- Color-coded Markers: Different colours for successful vs. failed landings
- Circle Markers: Radius proportional to payload mass
- Distance Calculations: Calculated distances to nearby features (railway, highway, coastline)

Why added

- - Visualize geographic distribution of launch sites
- - Identify patterns in success rates by location
- - Understand proximity to infrastructure
- - Provide interactive exploration of launch data

- https://github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%206/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plots/graphs and interactions

- Pie Chart: Launch success count for all sites
- Pie Chart: Launch site with highest success ratio
- Scatter Plot: Payload vs. Launch Outcome with range slider
- Interactive Filters: Launch site dropdown, payload range slider, booster version selector

Why added

- - Enable interactive exploration of data
- - Allow users to filter by different criteria
- - Visualize relationships between variables
- - Identify patterns in success rates by payload range and booster version

https://github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%207/spacex-dash-app.py

Predictive Analysis (Classification)

Model Development Process

1. Data Preparation:

- Standardized features using StandardScaler
- Split data: 80% training, 20% testing
- Created binary target variable (Class: 1=success, 0=failure)

2. Model Building:

- Logistic Regression with GridSearchCV (C: [0.01, 0.1, 1], penalty: l2, solver: lbfgs)
- Support Vector Machine with GridSearchCV (kernel: linear/rbf/poly/sigmoid, C: logspace(-3,3,5), gamma: logspace(-3,3,5))
- Decision Tree with GridSearchCV (criterion: gini/entropy, max_depth: 2-18, min_samples_leaf: 1-4, min_samples_split: 2-10)
- K-Nearest Neighbors with GridSearchCV (n_neighbors: 1-10, algorithm: auto/ball_tree/kd_tree/brute, p: 1-2)

3. Model Evaluation:

- Used 10-fold cross-validation for hyperparameter tuning
- Evaluated on test set using accuracy score
- Generated confusion matrices for each model

4. Best Model Selection:

- Compared test accuracies across all models
- Selected model with highest test accuracy

- https://github.com/ChrisPrior96/IBM-Data-Science/blob/main/Applied%20Data%20Science%20Capstone/Lab%208/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory data analysis results:

- Success rate increases with flight number (experience matters)
- Payload mass affects success rate
- Orbit type influences landing success
- Launch site location impacts success
- Success rate has improved over time (2013-2020)

Predictive analysis results

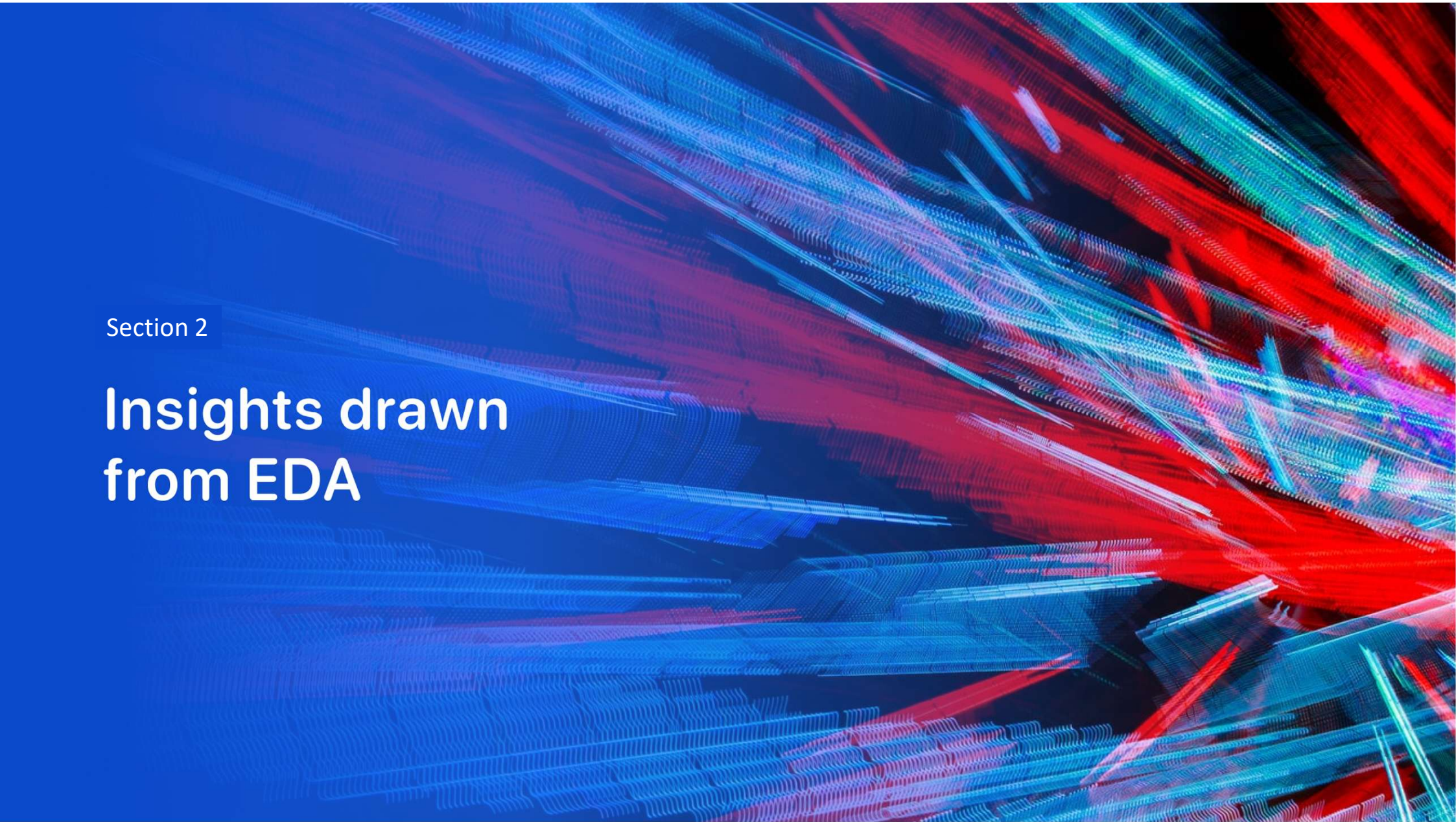
Logistic Regression Test Accuracy: 0.8333333333333334

SVM Test Accuracy: 0.8333333333333334

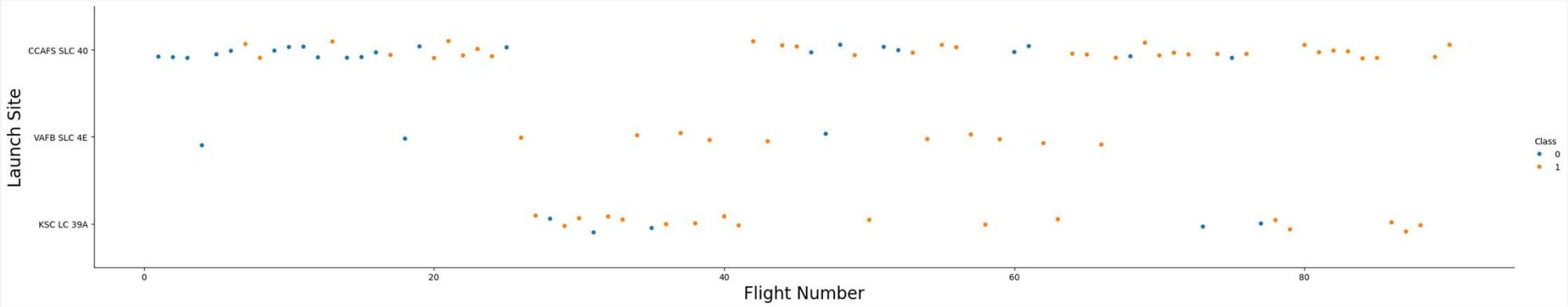
Decision Tree Test Accuracy: 0.8333333333333334

KNN Test Accuracy: 0.8333333333333334

Best Method: Logistic Regression with accuracy: 0.8333333333333334

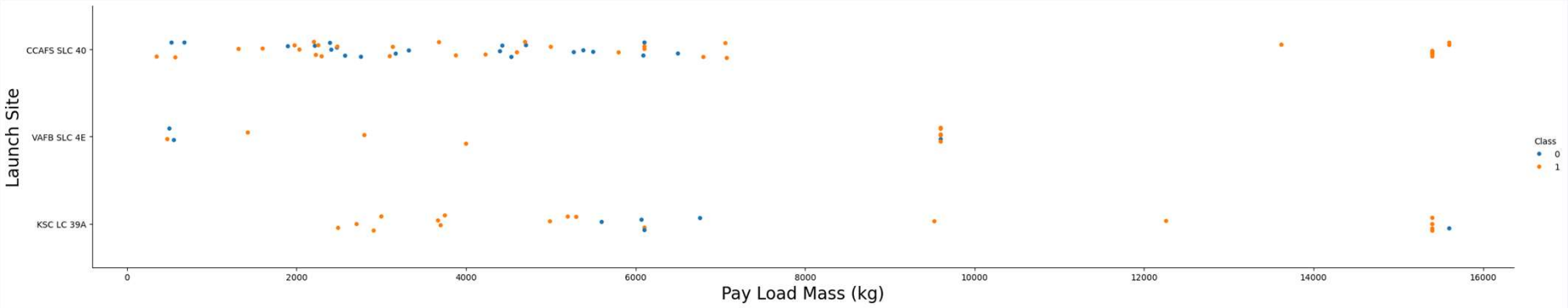

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Success rate improves with flight number, especially at certain launch sites

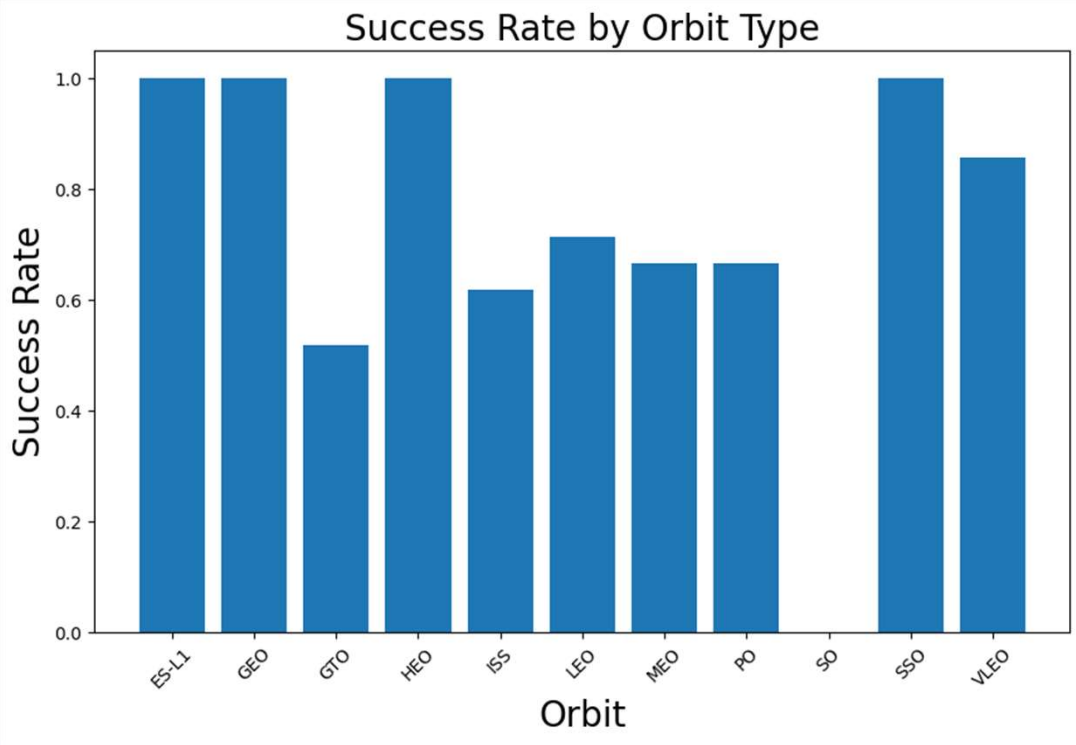
Payload vs. Launch Site



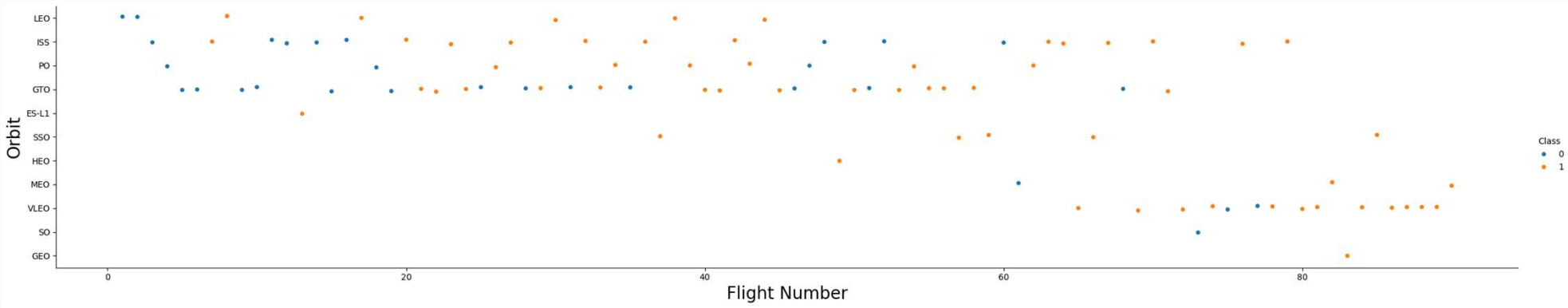
VAFB-SLC has no heavy payload launches; payload affects success differently by site

Success Rate vs. Orbit Type

- Highest Success Rate
- ES-L1
- GEO
- HEO
- SSO

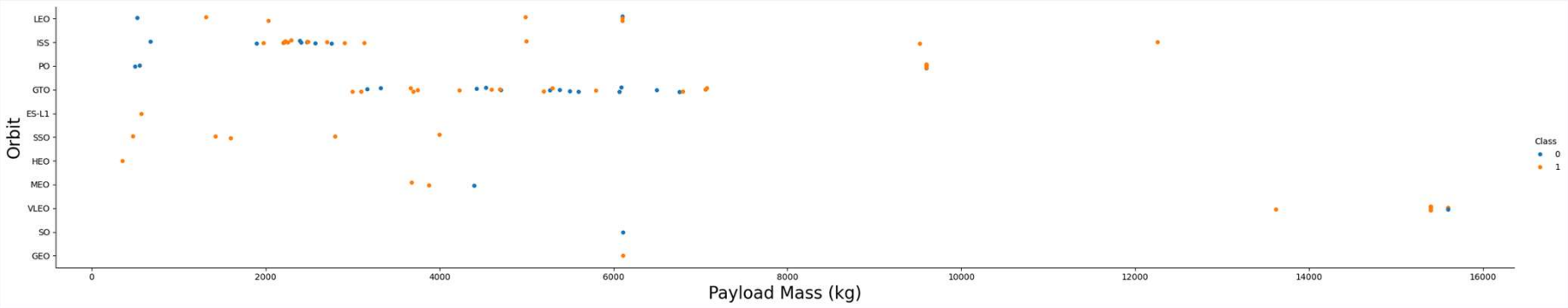


Flight Number vs. Orbit Type



- EO Shows no real relationship between flight number and success; GTO shows no clear relationship

Payload vs. Orbit Type



- Different orbits have different payload constraints and success patterns

Launch Success Yearly Trend

- Success rate has been increasing since 2013, reaching peak in 2020

All Launch Site Names

Query: `SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE`

Result: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

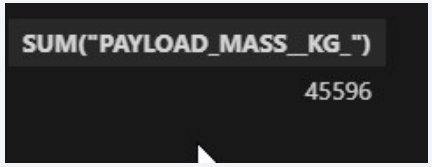
Launch Site Names Begin with 'CCA'

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

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

- %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'



Average Payload Mass by F9 v1.1

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

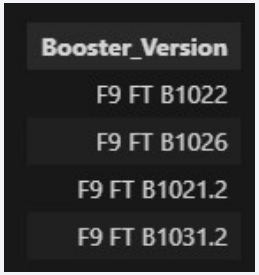
AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

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

Successful Drone Ship Landing with Payload between 4000 and 6000

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



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

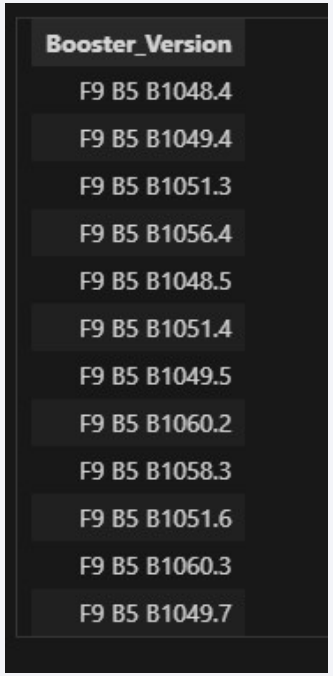
Total Number of Successful and Failure Mission Outcomes

%sql SELECT "Mission_Outcome", COUNT(*) *as* Count FROM SPACEXTABLE GROUP BY "Mission_Outcome"

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

Boosters Carried Maximum Payload

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



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

%sql SELECT substr("Date", 6, 2) *as* Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr("Date", 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)%'

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 SELECT "Landing_Outcome", COUNT(*) as Count FROM SPACEXTABLE WHERE "Date" >= '2010-06-04' AND "Date" <= '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC
```

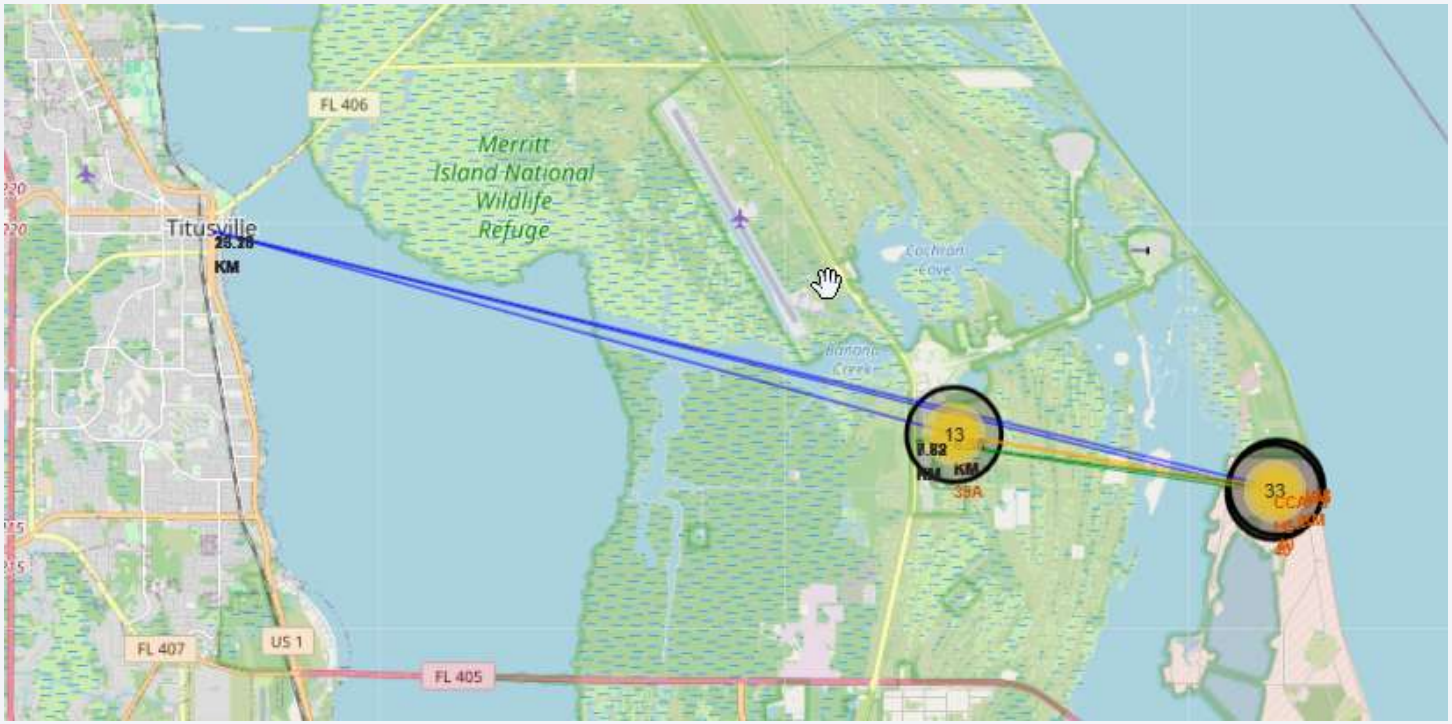
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

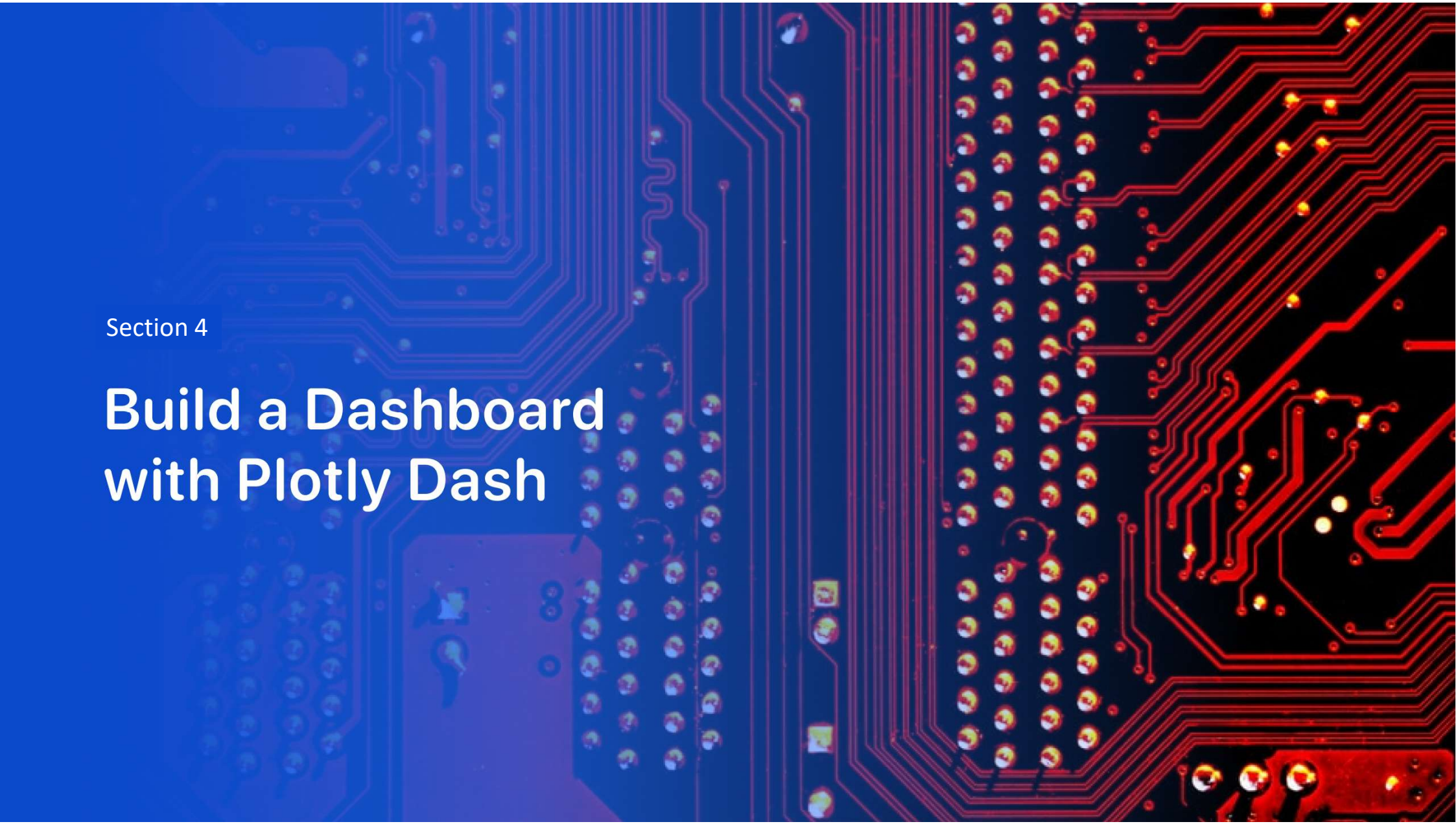

Section 3
Launch Sites
Proximities Analysis

All Launch Sites

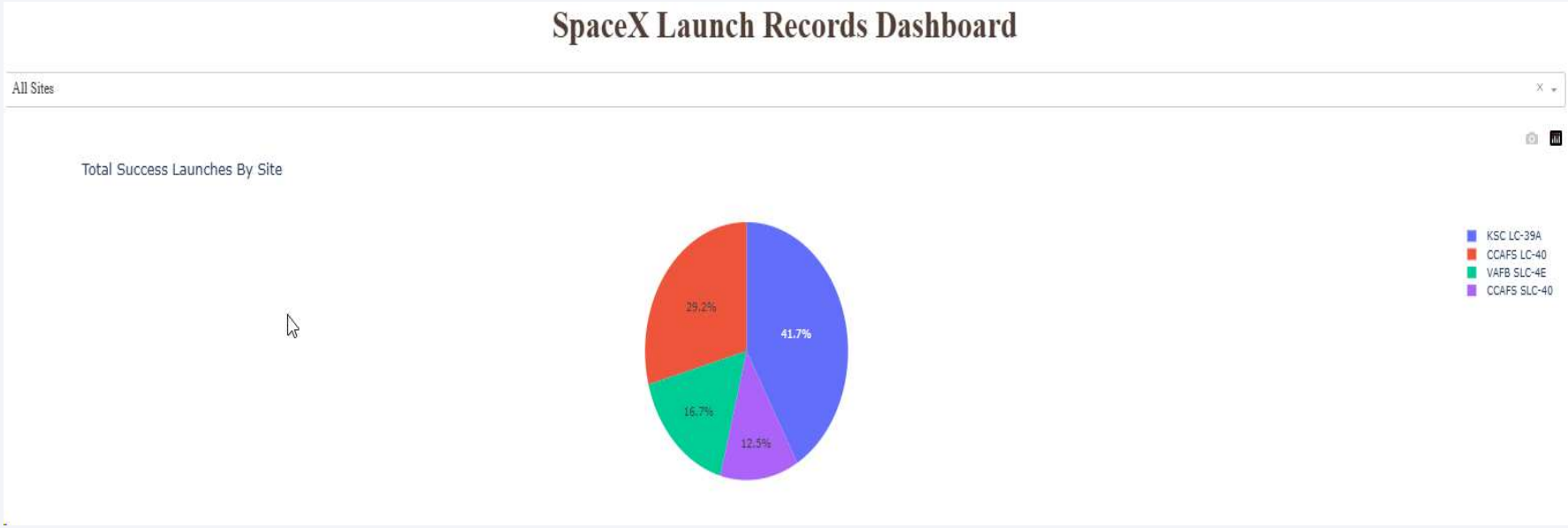
Colour coded Launch Site Map

Distances between a launch site to its proximities



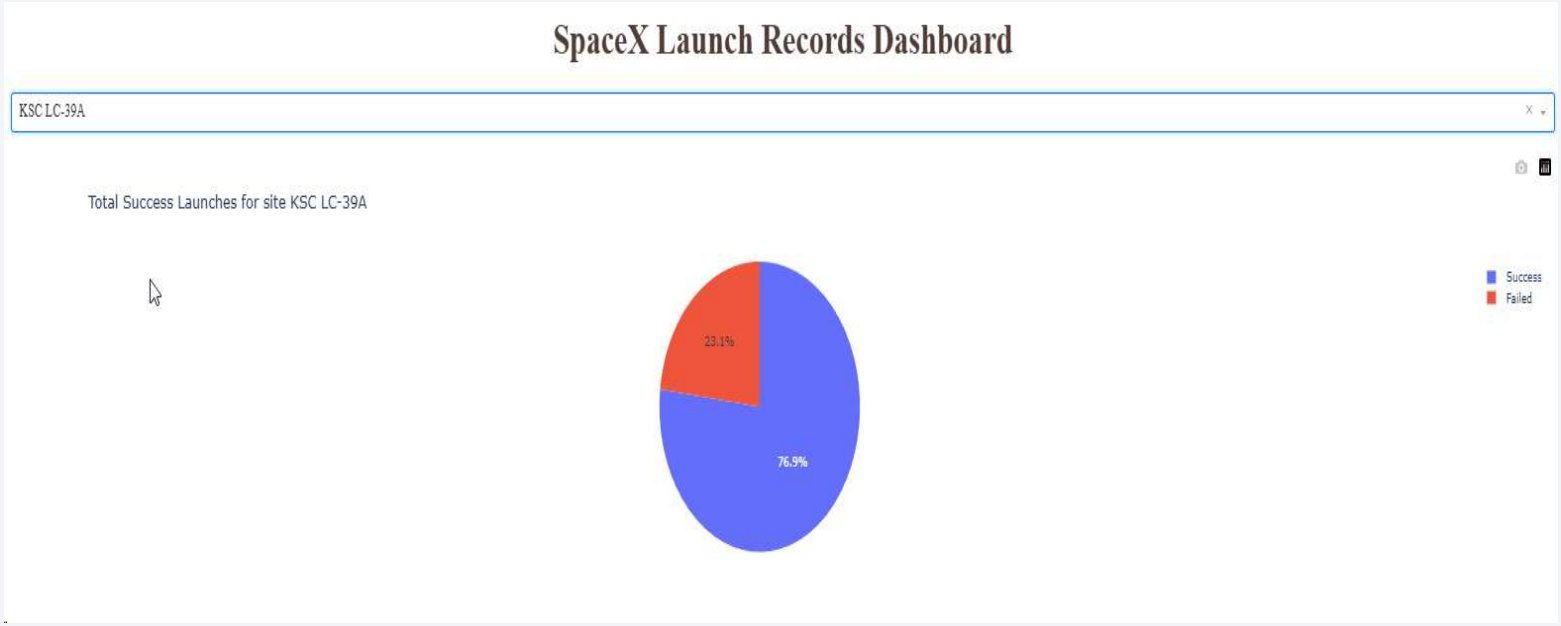

Section 4
Build a Dashboard with Plotly Dash

Total Success by Site

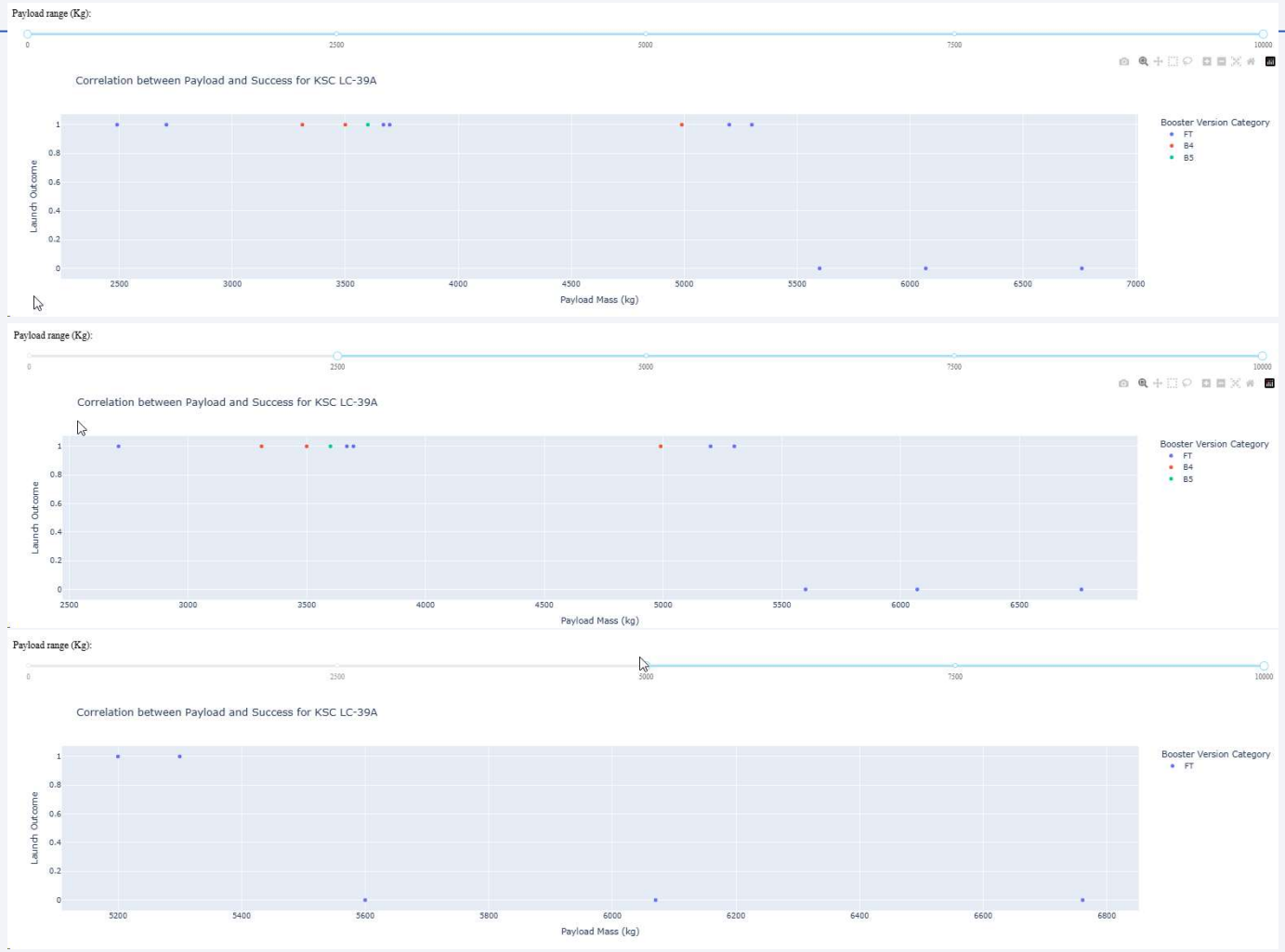


- KSC LC-39A is the most successfull

Total Successful Launches for site KSC LC-39A



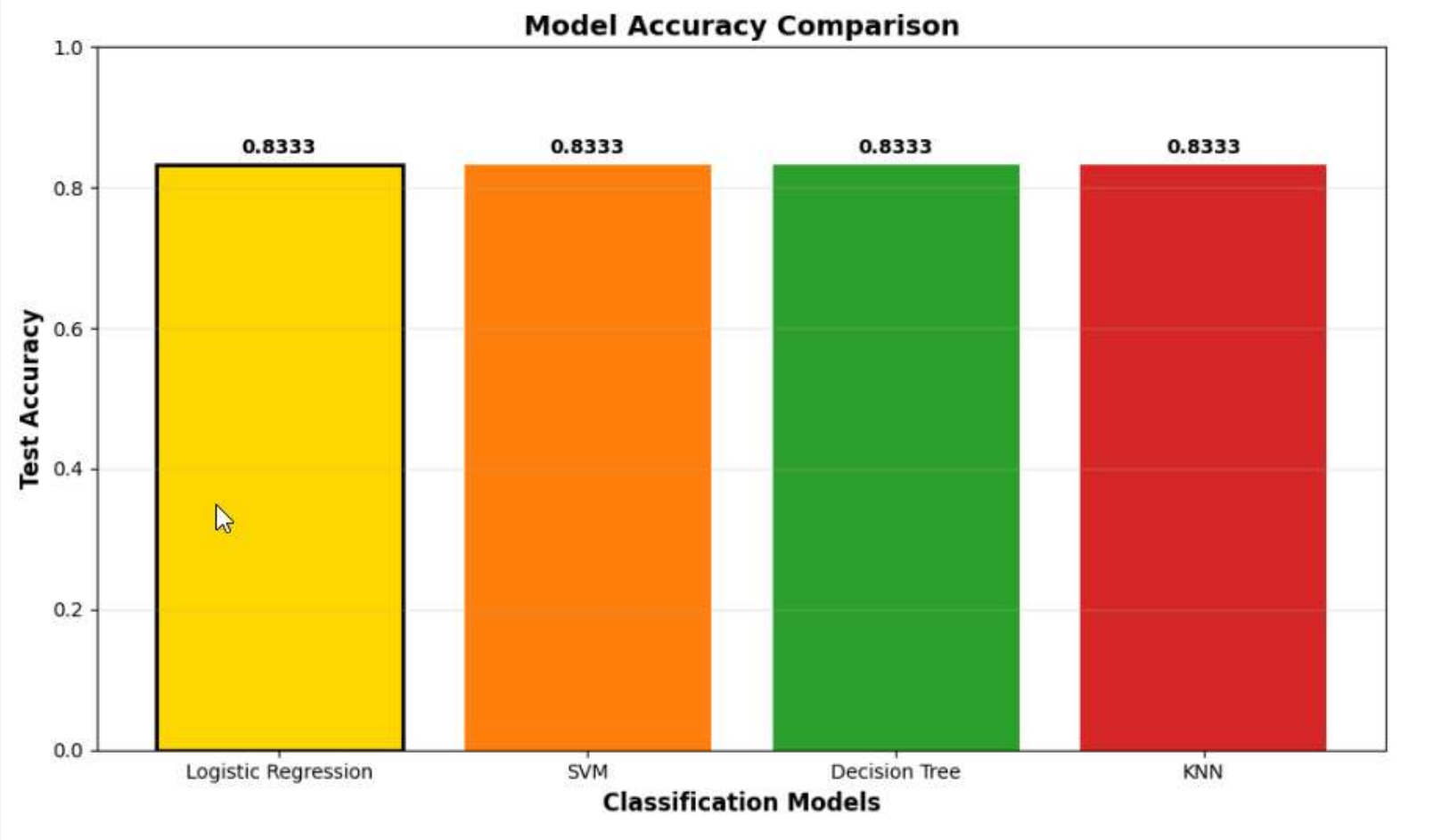
Correlation between payload and Success for KSC-LC-39A



Section 5

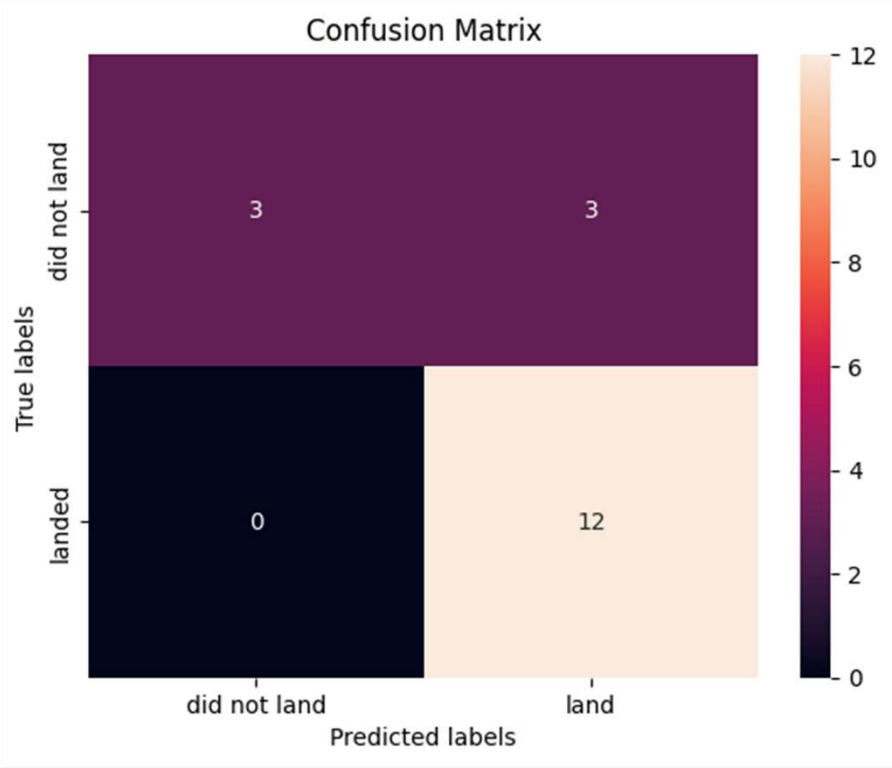
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- True Positives: Correctly predicted successful landings
- False Positives: Incorrectly predicted successful landings
- True Negatives: Correctly predicted failed landings
- False Negatives: Incorrectly predicted failed landings



Conclusions

- Flight number (experience) is a key factor in landing success - success rate increases with more launches
- Payload mass and orbit type significantly influence landing outcomes
- Launch site location affects success rates, with some sites showing better performance
- Machine learning models can effectively predict landing success, with Logistic Regression
- achieving 83.4% accuracy

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!