Winning Space Race with Data Science

<Hasan Roshan>
<June 11, 2024>

Outline

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

Executive Summary

- **Summary of methodologies:**
  - Data Collection:
    - API Calls
    - Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA)
    - EDA with Data Visualization
    - EDA with SQL
  - Interactive Map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- **Summary of all results:**
  - Over time, the success rate of the Falcon9 launches increased
  - Different methods are available to predict the success rate with acceptable accuracy. For instance:
    - Medium and heavy payloads have high success rates
    - If SpaceX is launching from KSC LC-39A site, the outcome is most likely a success.
    - Any of the four models developed here (LogReg, SVM, Decision Tree, KNN) can be used with similar accuracy (0.83333).

Introduction

- **Background and Context:**
  - The commercial space age is making space travel more affordable.
  - Companies like Virgin Galactic, Rocket Lab, Blue Origin, and especially SpaceX are leading the charge.
  - SpaceX stands out due to its reusable rockets, significantly reducing launch costs to $62 million compared to $165 million from other providers.
  - Their success includes missions to the ISS, the Starlink satellite internet project, and manned space missions.
  - The key to their cost efficiency is the reusable first stage of the Falcon 9 rocket.
- **Problems Statement:**
  - How to predict the landing success of SpaceX's Falcon 9 first stage?
  - Accurate predictions can help determine launch costs and provide valuable insights for other companies competing in the space launch market.

Section 1

Methodology

Methodology

Executive Summary

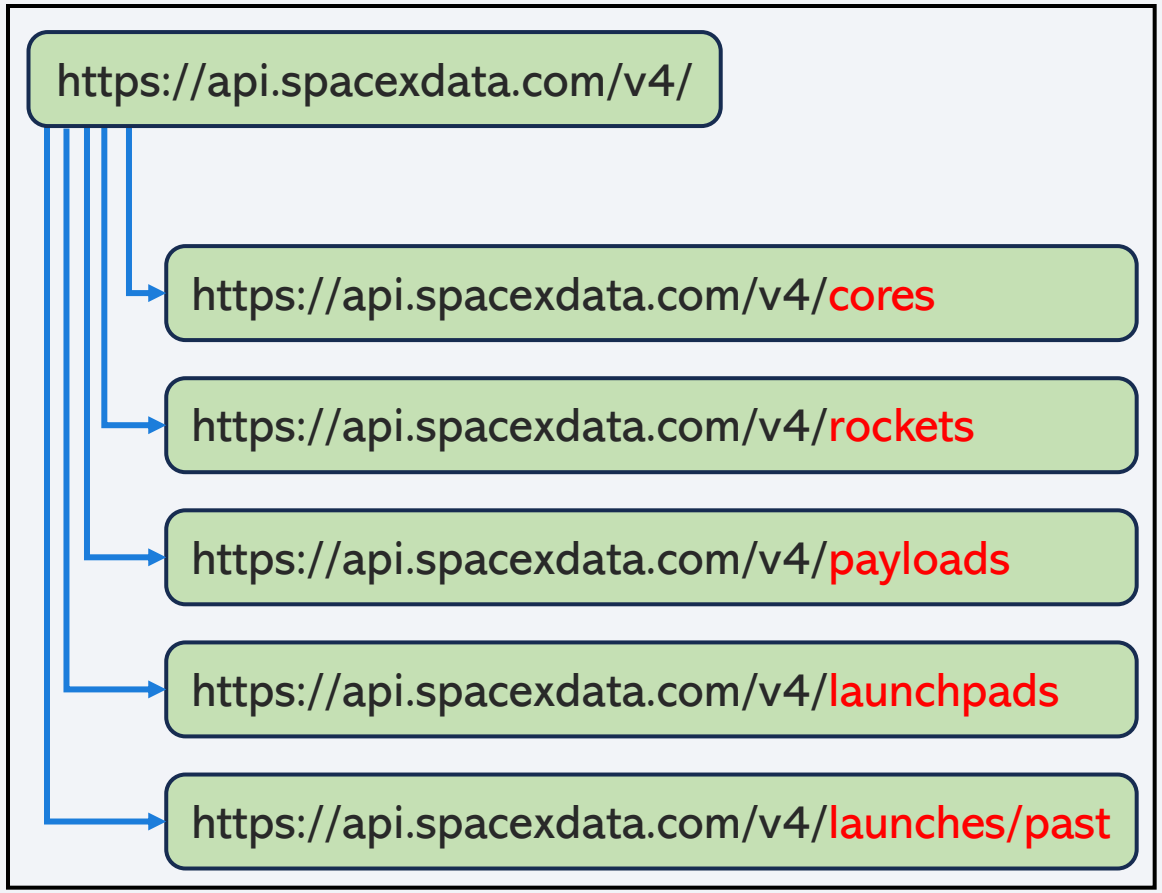
- Data collection methodology:
  - Two methods were used for data collection: (1) API Calls, (2) Web Scraping
- Perform data wrangling
  - Exploratory Data Analysis (EDA) was performed to create a ‘Class’ label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 3-stage process: Data preparation, finding best model-in-method, Methods Comparison

Data Collection

- Two methods were used for data collection:
  - API Calls: The primary data source is the SpaceX REST API, accessed via the endpoint api.spacexdata.com/v4/launches/past, providing comprehensive launch details; data is collected using Python's requests library and converted from JSON to a flat table using Panda's "json_normalize" method.
  - Web Scraping: Additional launch data is sourced from Wikipedia pages using the BeautifulSoup package to scrape HTML tables, which are then parsed and converted into Panda's data frames for clean, structured analysis and visualization.

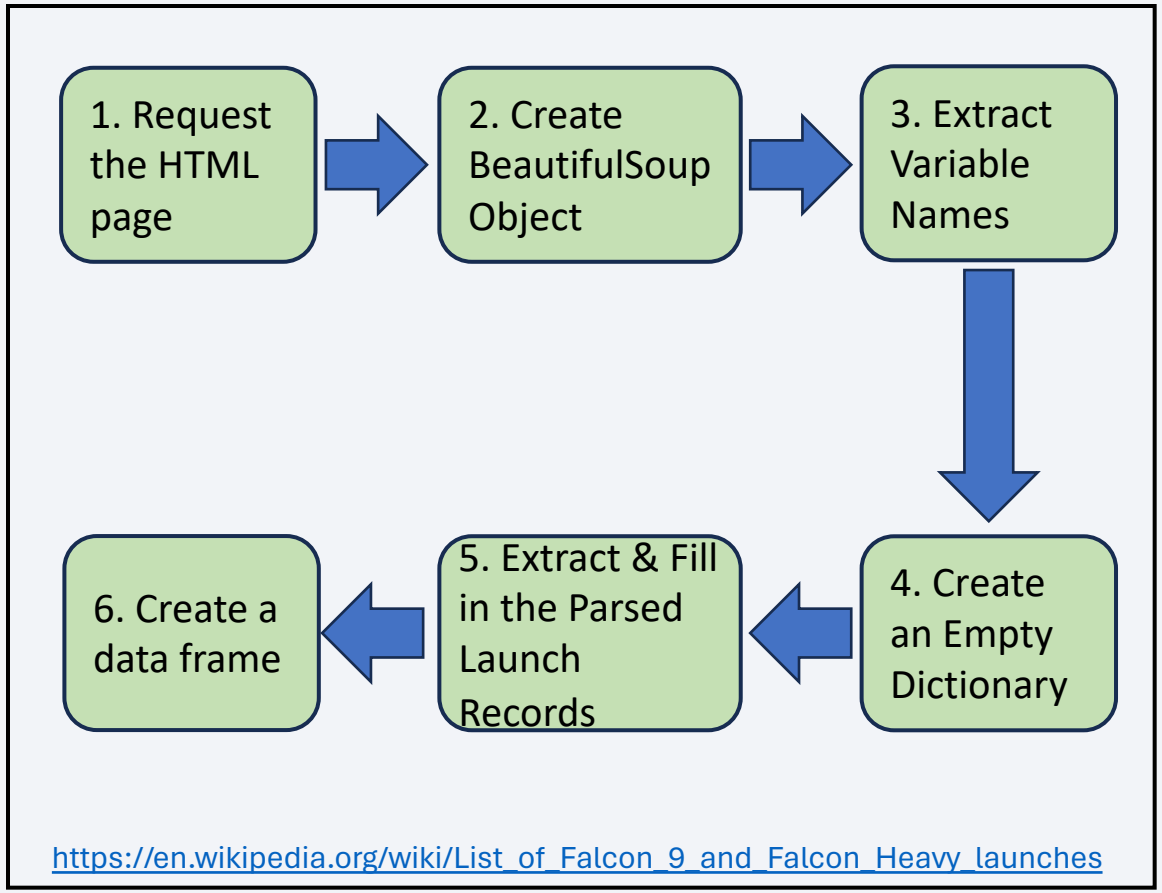
Data Collection – SpaceX API

- Import Libraries & Define Auxiliary Functions
- Request & parse the SpaceX launch data using the GET request:
  - Create global variables & apply Auxiliary functions
  - Construct a dataset from the dictionary of combined columns
- Filter the dataframe to only include Falcon 9 launches
- Data Wrangling: replace missing values using the .mean() method
- Libraries used: Pandas, NumPy, datetime.
- GitHub URL



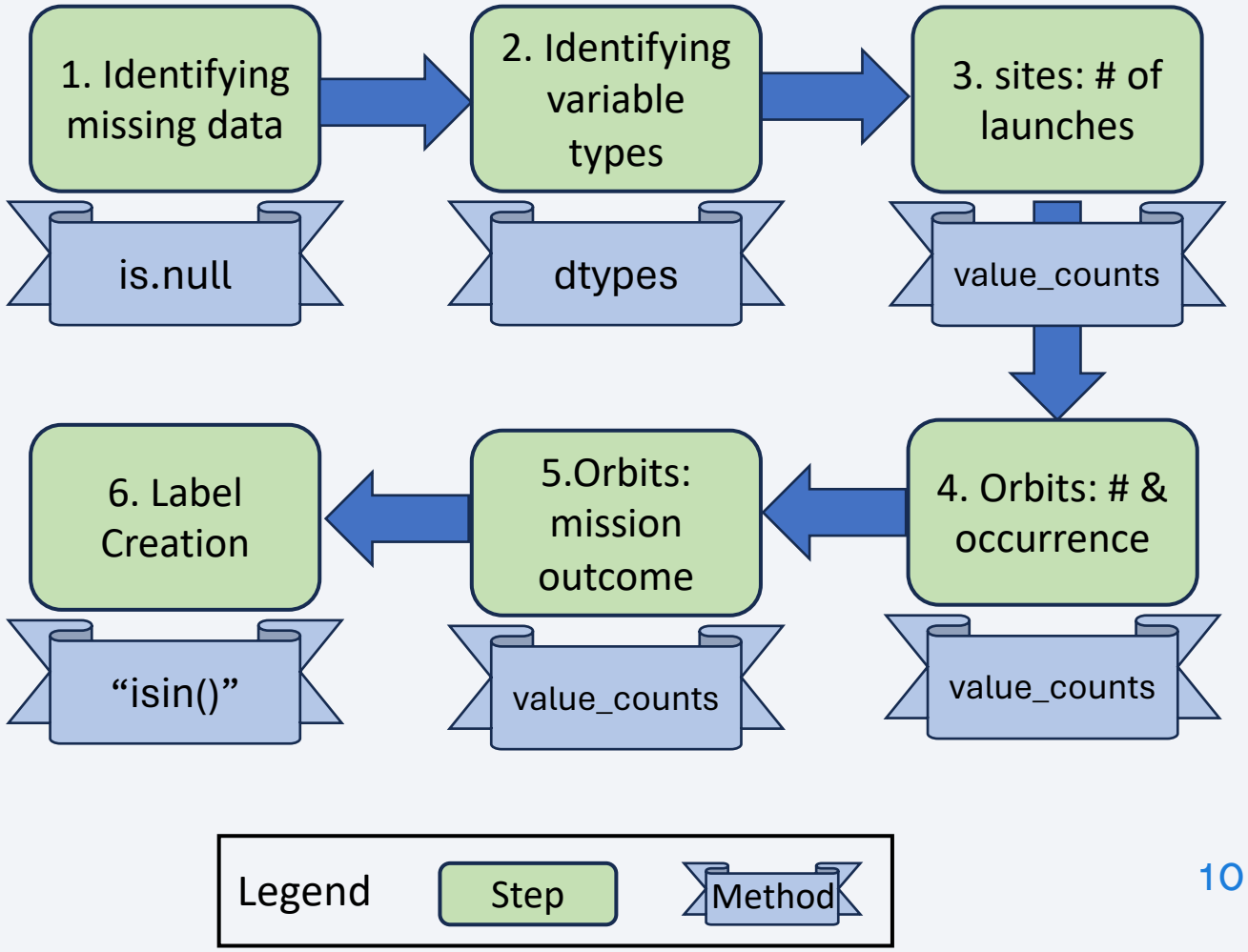
Data Collection - Scraping

- Web scraping consists of two main steps:
  - Extract Falcon 9 launch records with BeautifulSoup
  - Parse the table and convert it into a Pandas data frame:
- Source: Wikipedia
- Libraries used: Pandas, BeatifulSoup, request
- GitHub URL



Data Wrangling

- Exploratory Data Analysis (EDA) was performed to convert outcome of the booster's landing into a numerical, two-value label named 'Class' for training supervised models.
- The 'Class' label groups eight distinct outcomes into two groups, where `1` means the booster successfully landed, and `0` means it was unsuccessful.
- GitHub URL



EDA with Data Visualization

- Exploratory Data Analysis (EDA) was performed to extract insights about the relationship between different variables
- 3 types of data visualizations:
  - Scatter Plots: 5
  - Bar Chart: 1
  - Line Plot (Time-series): 1
- Libraries used: Pandas, NumPy, Seaborn
- GitHub URL

Type	Reason (example)
Scatter Plot	Effect of experience & weight on success rate
	Discover effect of location & experience on success rate
	If heavy payloads are more successful in specific location
	If achieving success took more time for specific orbits.
	If heavy payloads are more successful in a specific orbit
Bar Plot	Which orbit has the highest success rate?
Line Plot	Identify overall success trend

EDA with SQL

- 10 SQL Queries Performed in 4 groups:
  - Launch Site Analysis
  - Payload Mass Analysis
  - Landing Outcome Analysis
  - Booster Performance
  - Year-Specific Analysis
- Libraries Used: sqlalchemy, ipython-sql, csv, sqlite3, Pandas
- GitHub URL

Query Group	Example
Launch Site Analysis	Unique launch sites
Payload Mass Analysis	Total payload mass for NASA (CRS) boosters
Landing Outcome Analysis	Dates of first successful ground pad landings
Booster Performance	Booster versions with the maximum payload mass
Year-Specific Analysis	2015 landing outcomes, booster versions, and launch sites

Build an Interactive Map with Folium

- Steps Taken:
  - Marked NASA JSC & 3 launch sites on the map:
    - 4 markers & circles
    - Insight: Sites are close to coastlines
  - Marked the successful/failed launches:
    - 56 markers & circles
    - Insight: Visual comparison of sites
  - Marked & displayed distances between a launch site to its proximities:
    - 2 lines & distance markers
- Libraries Used:
  - Pandas, folium, MarkerCluster, MousePosition, DivIcon
- Additional data Source:
  - spacex_launch_geo.csv
- GitHub URL

Build a Dashboard with Plotly Dash

- Steps Taken:
  - Added a Drop-down Input Component to:
    - Interactively select a launch site
  - Added a callback function to:
    - Render a pie chart based on the dropdown input
    - Visualize the proportion of success/failure outcomes
  - Added a Range Slider to Select Payload:
    - Investigate the effect of different payload ranges on mission outcomes
- Libraries Used: Pandas, dash, plotly.express
- GitHub URL

Predictive Analysis (Classification)

Data Preparation

- Created a column for the class (Y)
- Standardized the data (X)
- Split data into training data and test data

Find the best Hyperparameters

- For each method (Logistic Regression, SVM, Classification Trees, and KNN):
  - Created a model object
  - Created a GridSearchCV object
  - Fitted the object to find the best parameters from the parameters' dictionary
  - Performed accuracy tests on the test data for the best model & stored the results

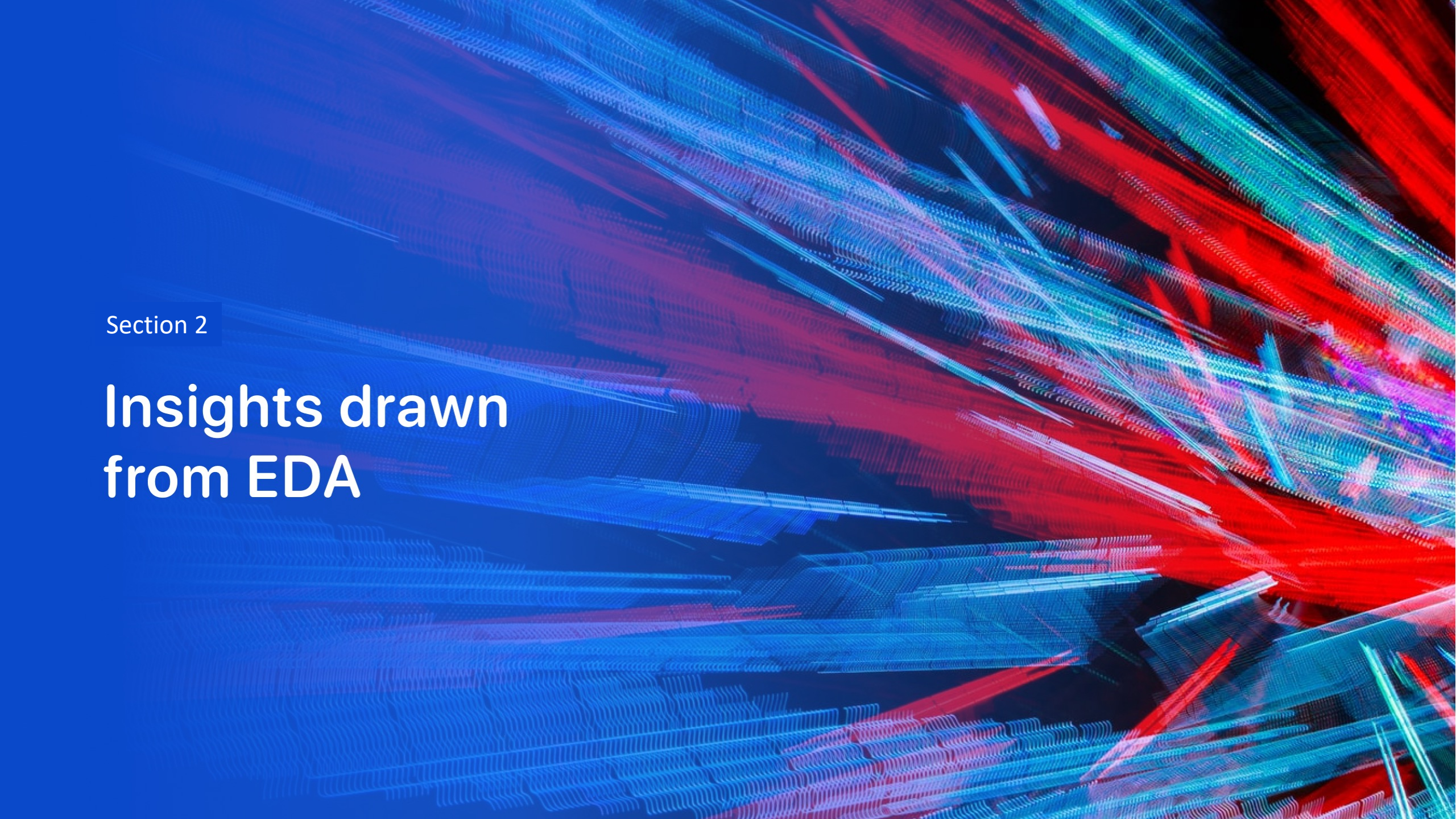
Find the best method

- Presented the performance of all best models in a tabular format
- Compared and identify the best model

- Libraries Used:
  - Pandas
  - NumPy
  - Seaborn
  - matplotlib.pyplot
  - sklearn
- GitHub URL

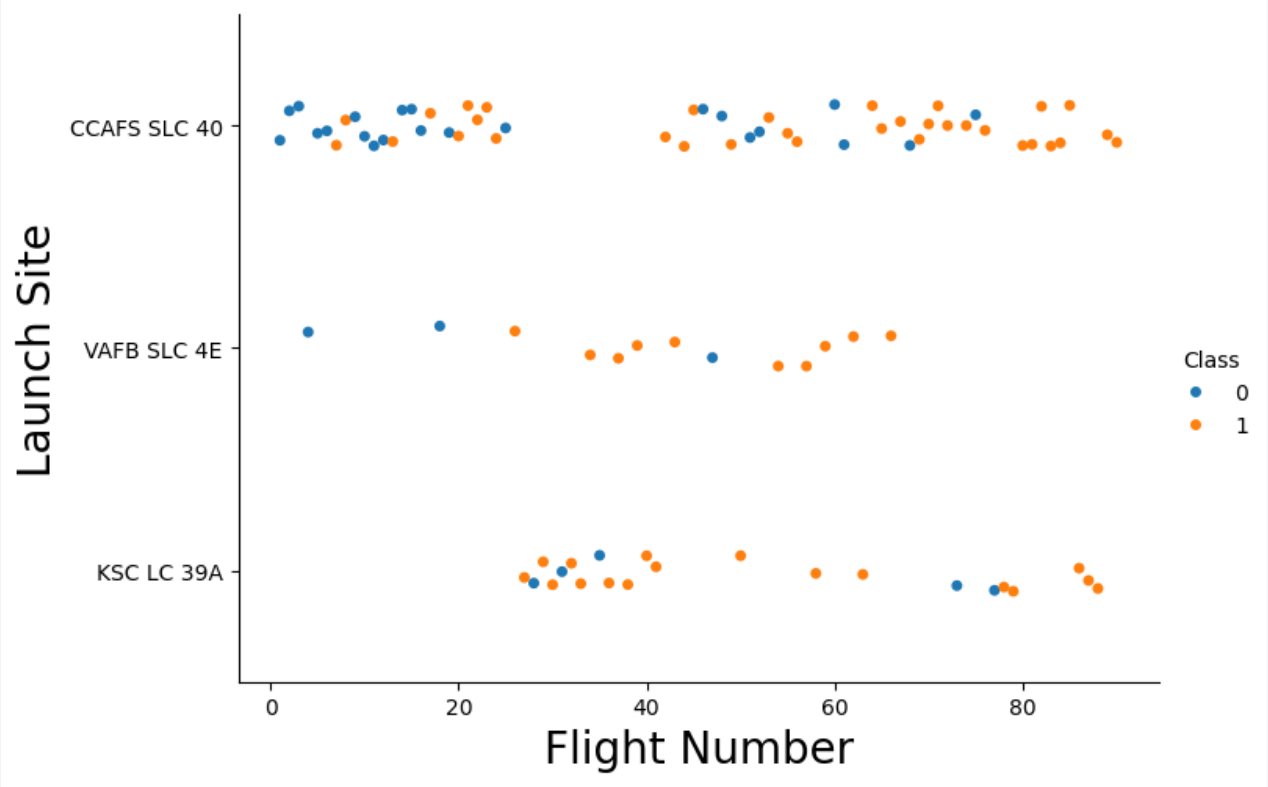
Results

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


Section 2
Insights drawn from EDA

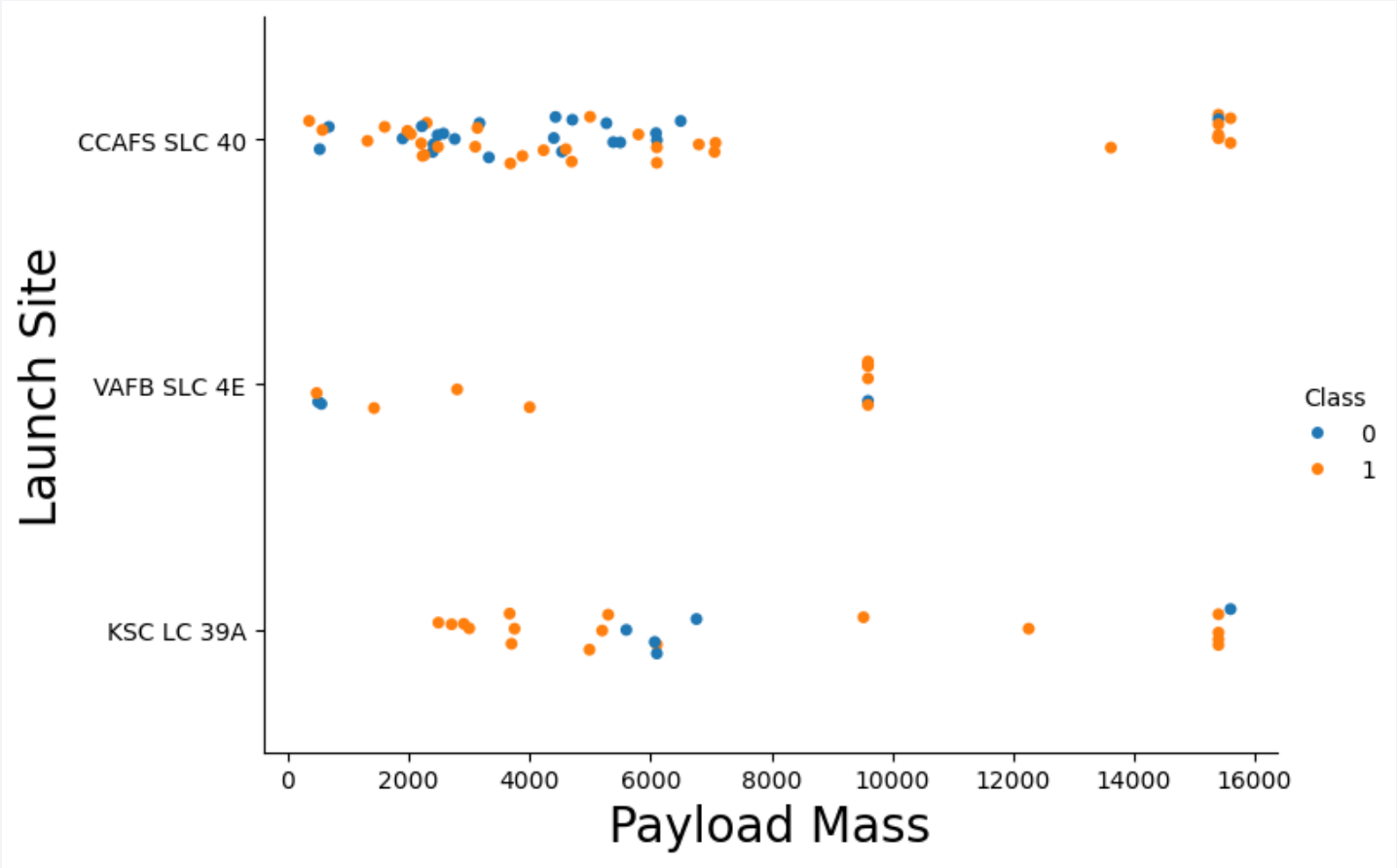
Flight Number vs. Launch Site

- Insights:
  - Majority of the first 20 launches were in one site
  - VAFB SLC site began with two consecutive failure, but ended up being the most successful site
  - Over time (increasing flight number), successful launches increases



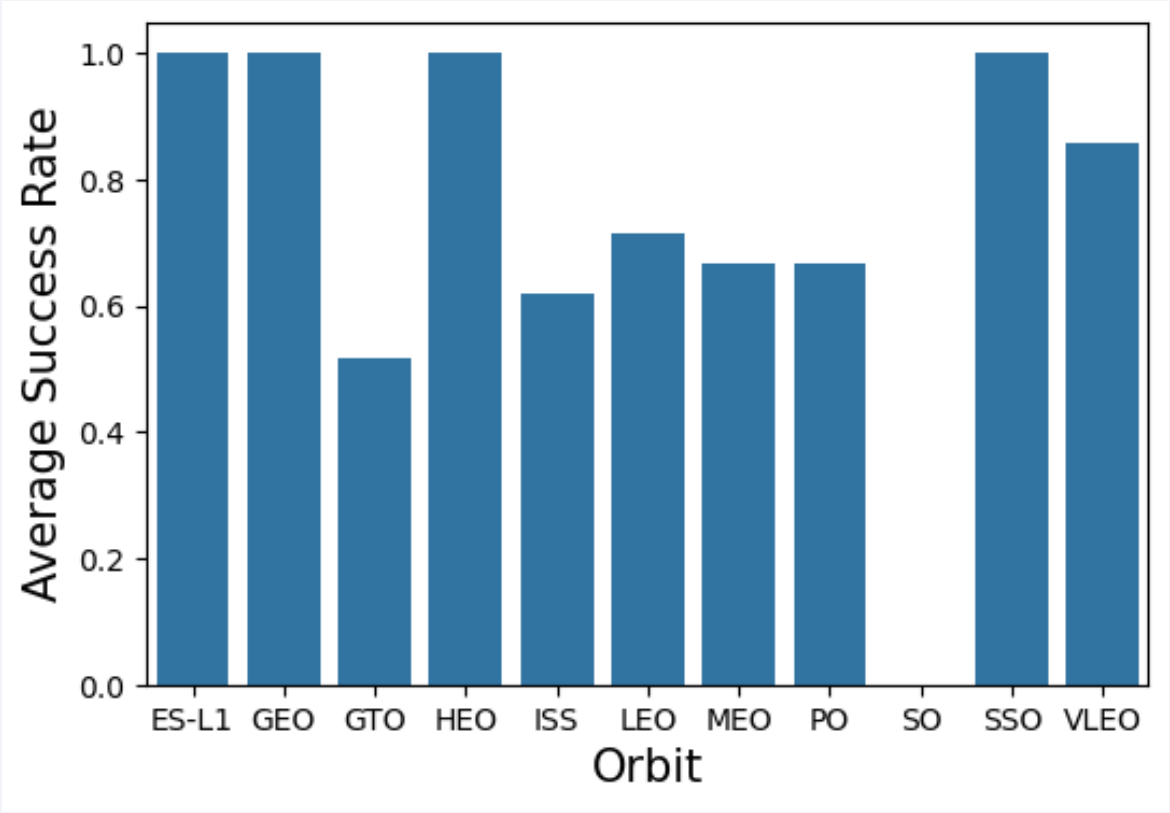
Payload vs. Launch Site

- Insights:
  - Most failures are from lighter payloads and in one site
  - Medium and heavy payloads have high success rates



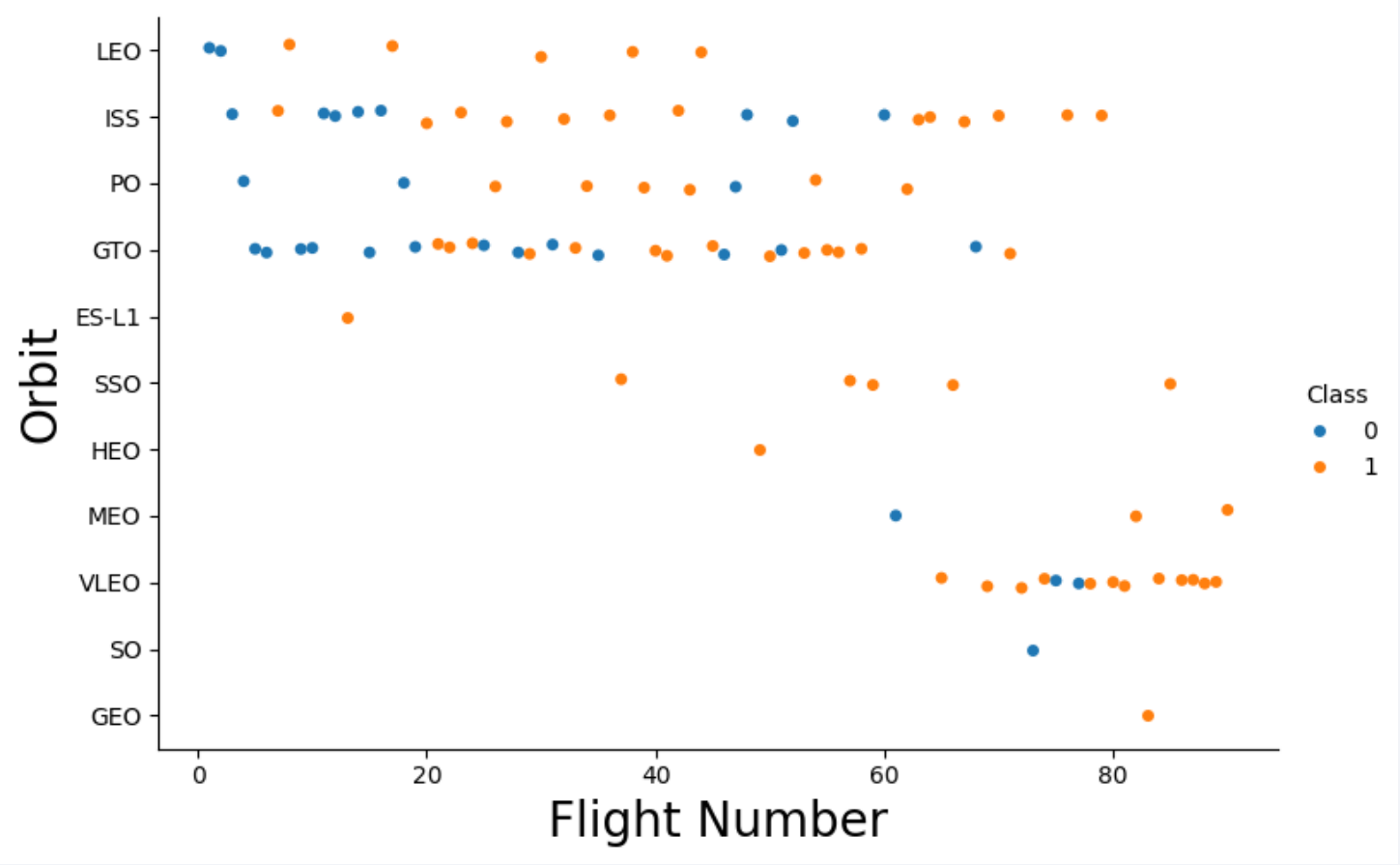
Success Rate vs. Orbit Type

- Insights:
  - The success rates of four orbits are perfect
  - All launches of Falcon9 to SO orbit were unsuccessful



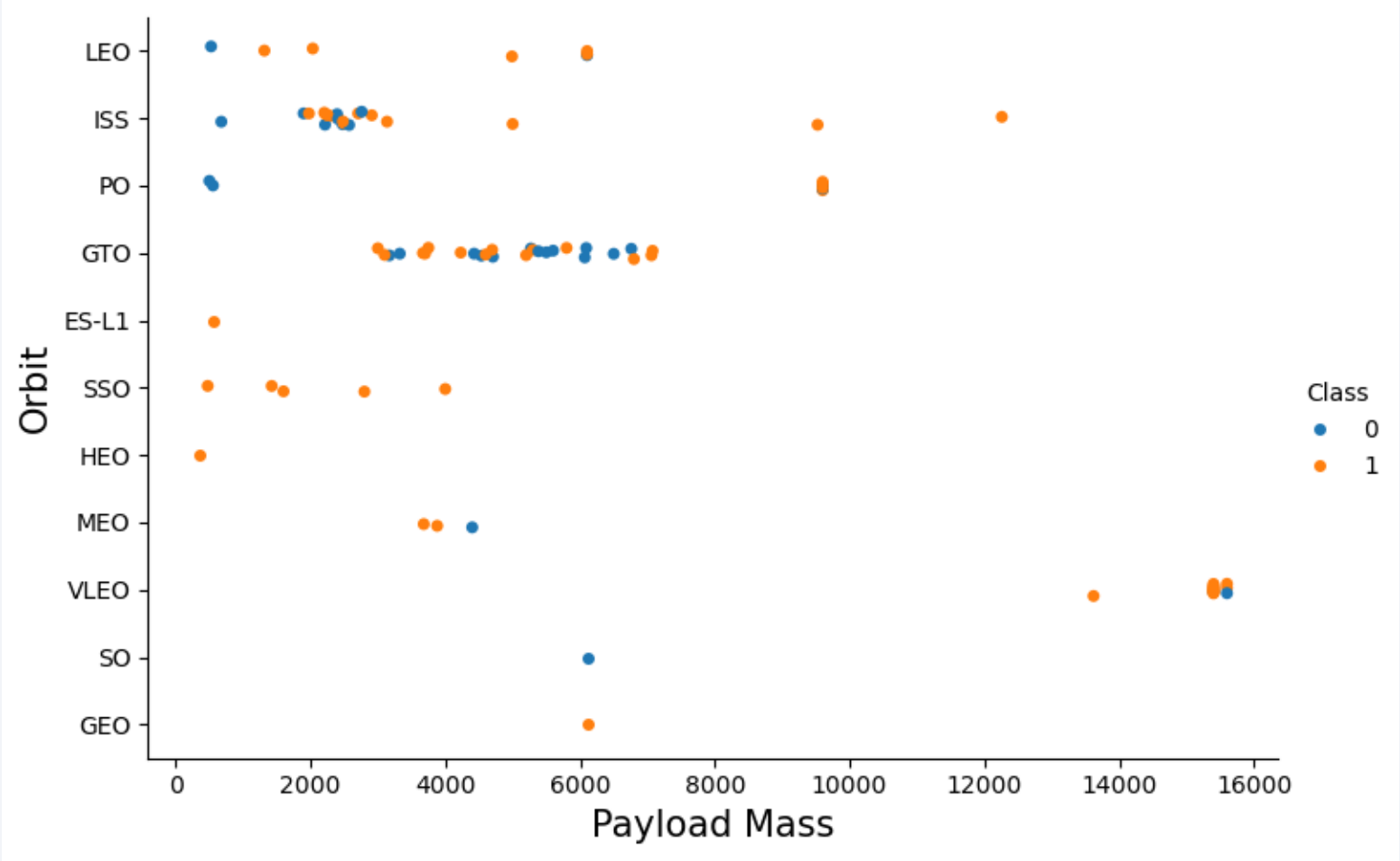
Flight Number vs. Orbit Type

- Insights:
  - Over time, successful launches increase for all orbits
  - Many orbits (like GEO, SO, VLEO) were later become selected



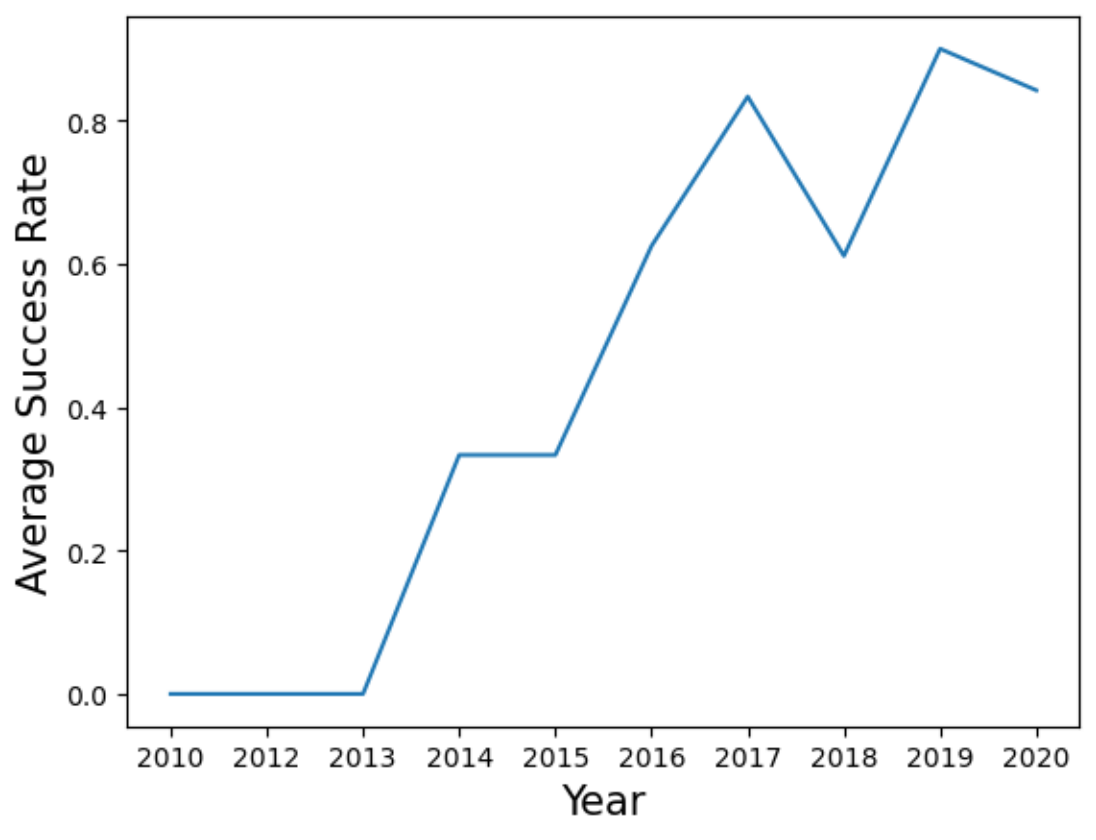
Payload vs. Orbit Type

- Insights:
  - We can predict the range of orbits based on the payload mass
  - Certain orbits (e.g., ES-L1, SSO, MEO) have exclusively had lower-range payload mass



Launch Success Yearly Trend

- The overall trend shows an increasing success rate over time
- The 2018 decline can be explained by further analysis of other charts



All Launch Site Names

- The snippet code and output on the right shows how to find the names of the unique launch sites

```
%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'

- The bellow snippet code shows 5 records where launch sites begin with `CCA`

```
%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

- The total payload carried by boosters from NASA

```
%%sql SELECT SUM("PAYLOAD_MASS__KG_")
FROM SPACEXTABLE WHERE "Customer" LIKE "NASA%";
```

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

SUM("PAYLOAD_MASS__KG_")
99980

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

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

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

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

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

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed mission outcomes

```
%%sql SELECT COUNT(*) FROM SPACEXTABLE
WHERE (("Landing_Outcome" LIKE "Success%")
       OR ("Landing_Outcome" LIKE "Failure%"));
```

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

COUNT(*)
71

Boosters Carried Maximum Payload

- The names of the boosters that have carried the maximum payload mass
- 12 distinct booster versions

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

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

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

2015 Launch Records

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

```
%%sql SELECT SUBSTR(Date,6,2), "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE
WHERE ((SUBSTR("Date",0,5)="2015") AND ("Landing_Outcome"= "Failure (drone ship)"));
```

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

SUBSTR(Date,6,2)	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

- Ranked count of landing outcomes between June 4, 2010, and March 20, 2017, in descending order

```
%%sql SELECT "Landing_Outcome",
COUNT("Landing_Outcome") as TOTAL FROM SPACEXTABLE
GROUP BY "Landing_Outcome"
HAVING ("Date" BETWEEN "2010-06-04" AND "2017-03-20")
ORDER BY TOTAL DESC;
```

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

Landing_Outcome	TOTAL
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Site

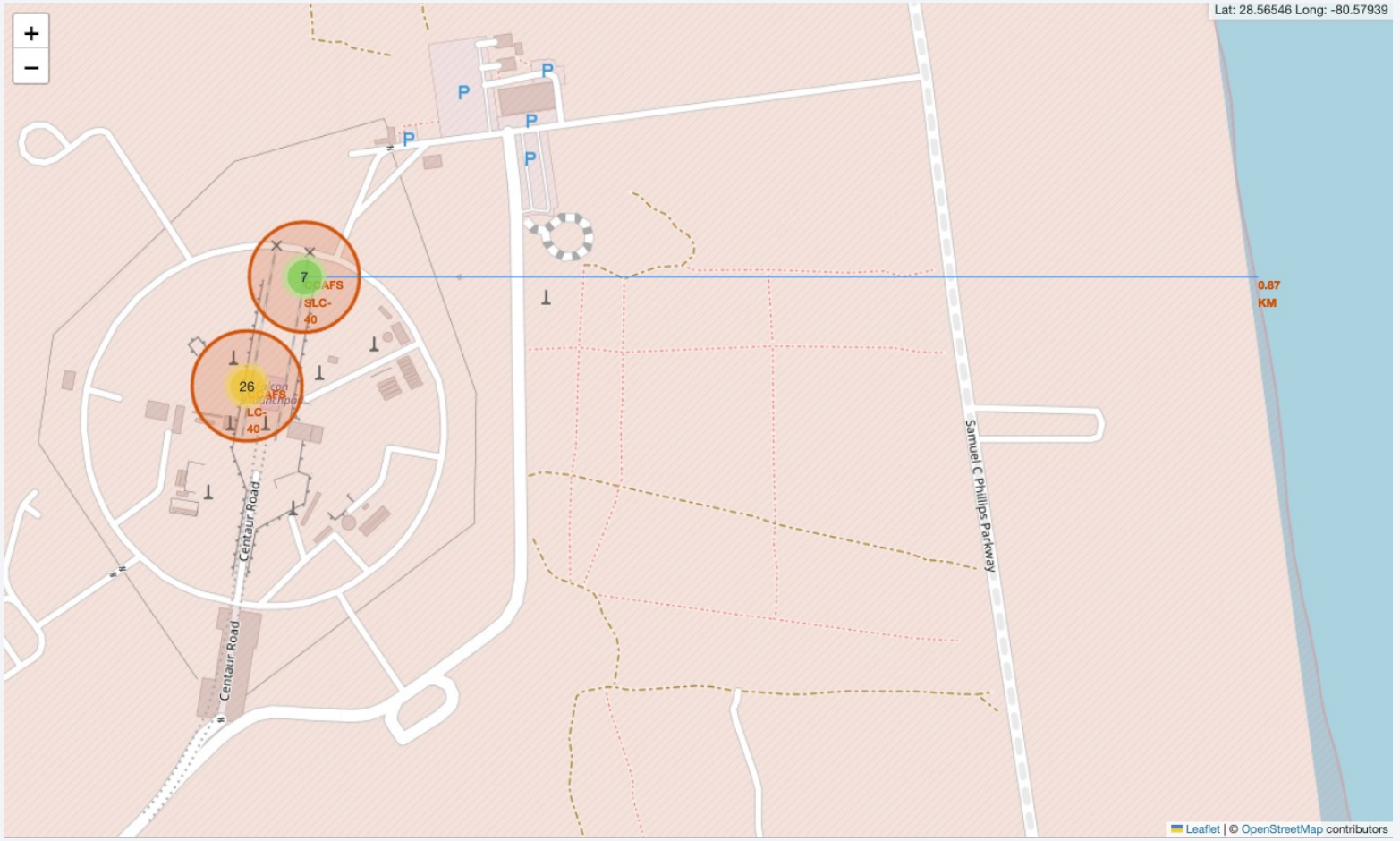
- Insights:
  - Launch sites are in proximity to coastlines
  - All are relatively far away from the equator

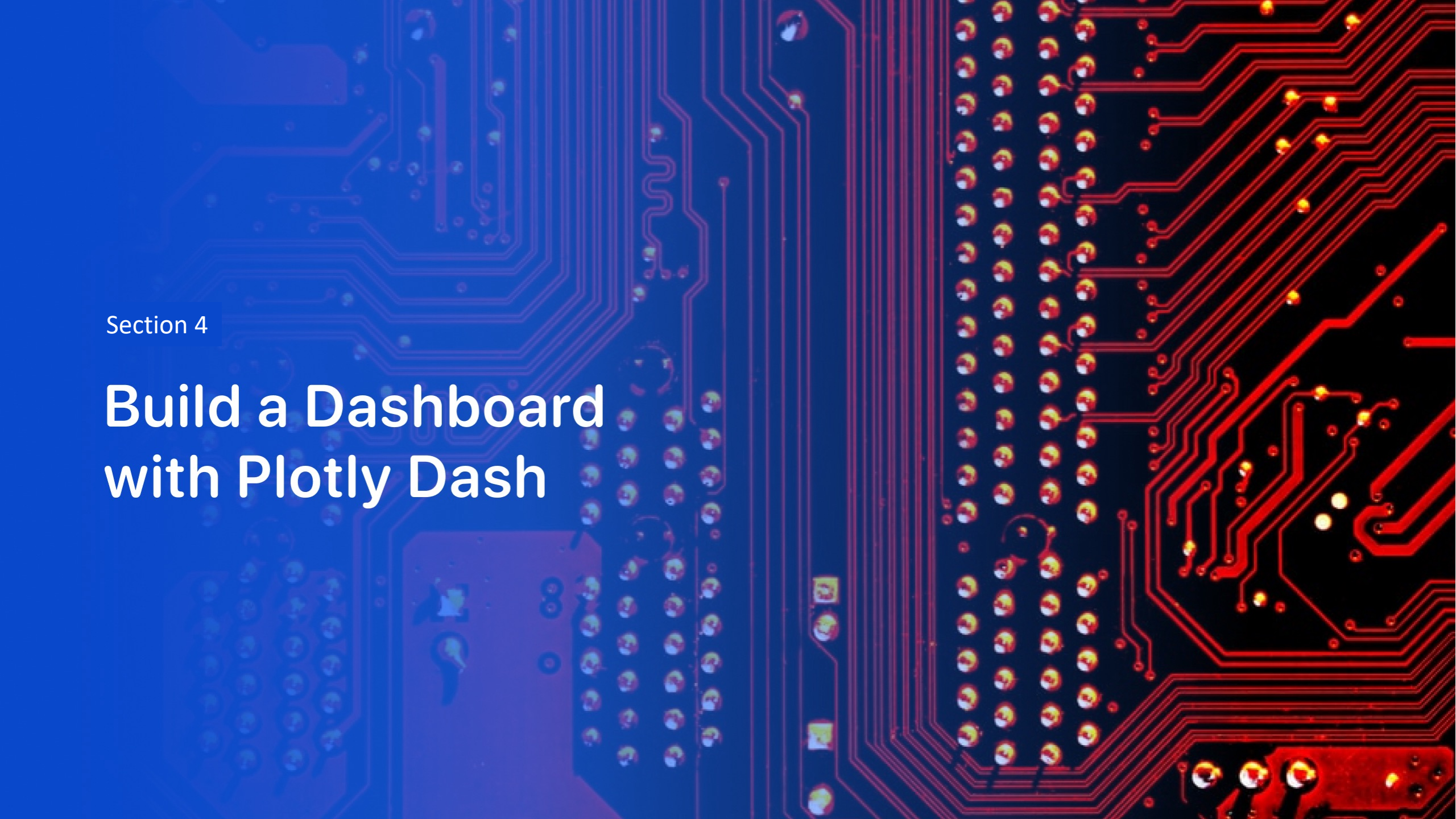
Visualizing Launch Outcomes on an Interactive Map

- KSC LC-39A is the launch site with the highest success rate
  - Green markers: success (10)
  - Red markers: failure (3)

Launch Site Proximities

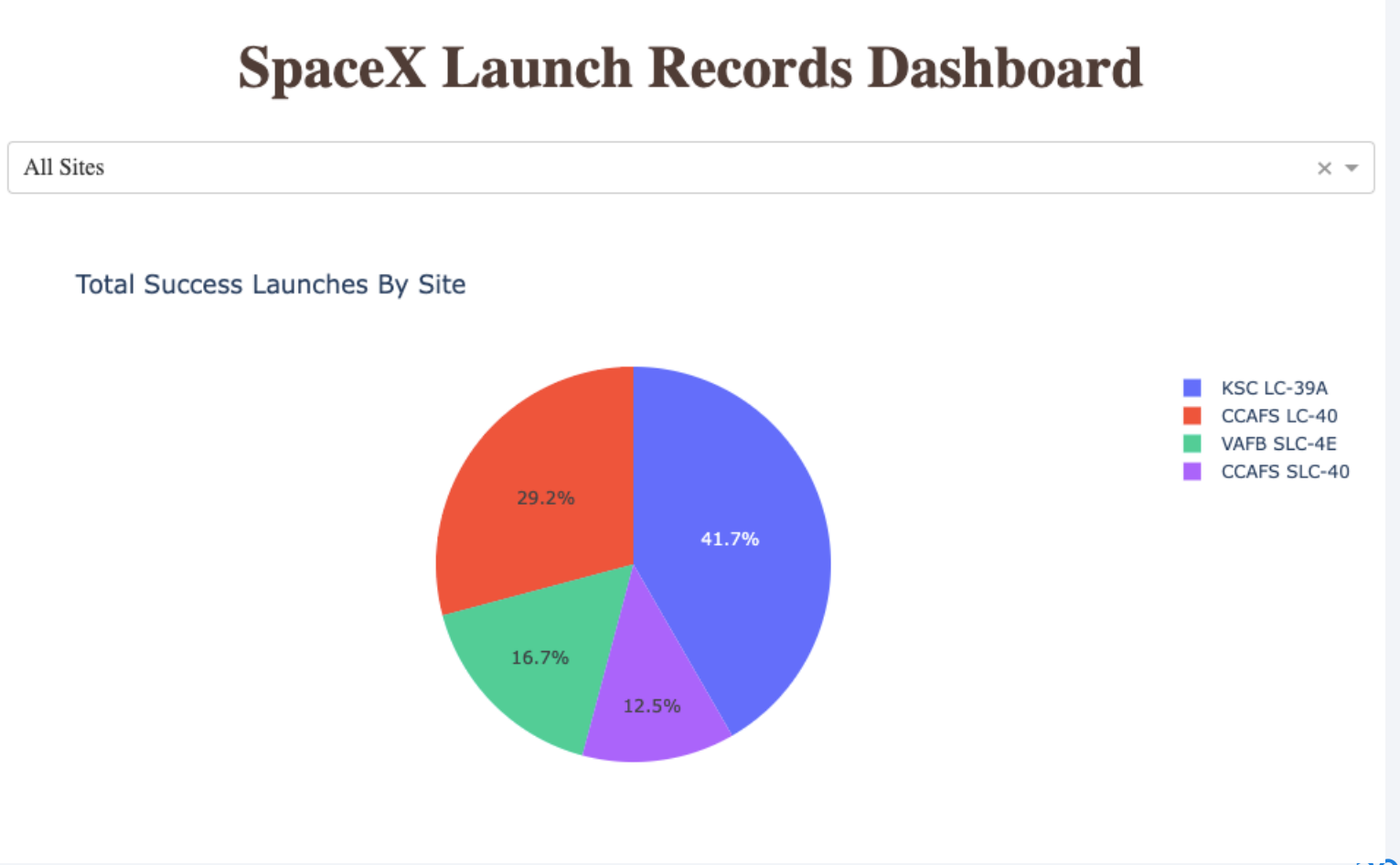
- Red circles: launch sites: Blue line: distance from the nearest coastline (0.87 Km)
- Insights:
  - Launch sites are in proximity to coastlines, roads, and railroads




Section 4
Build a Dashboard with Plotly Dash

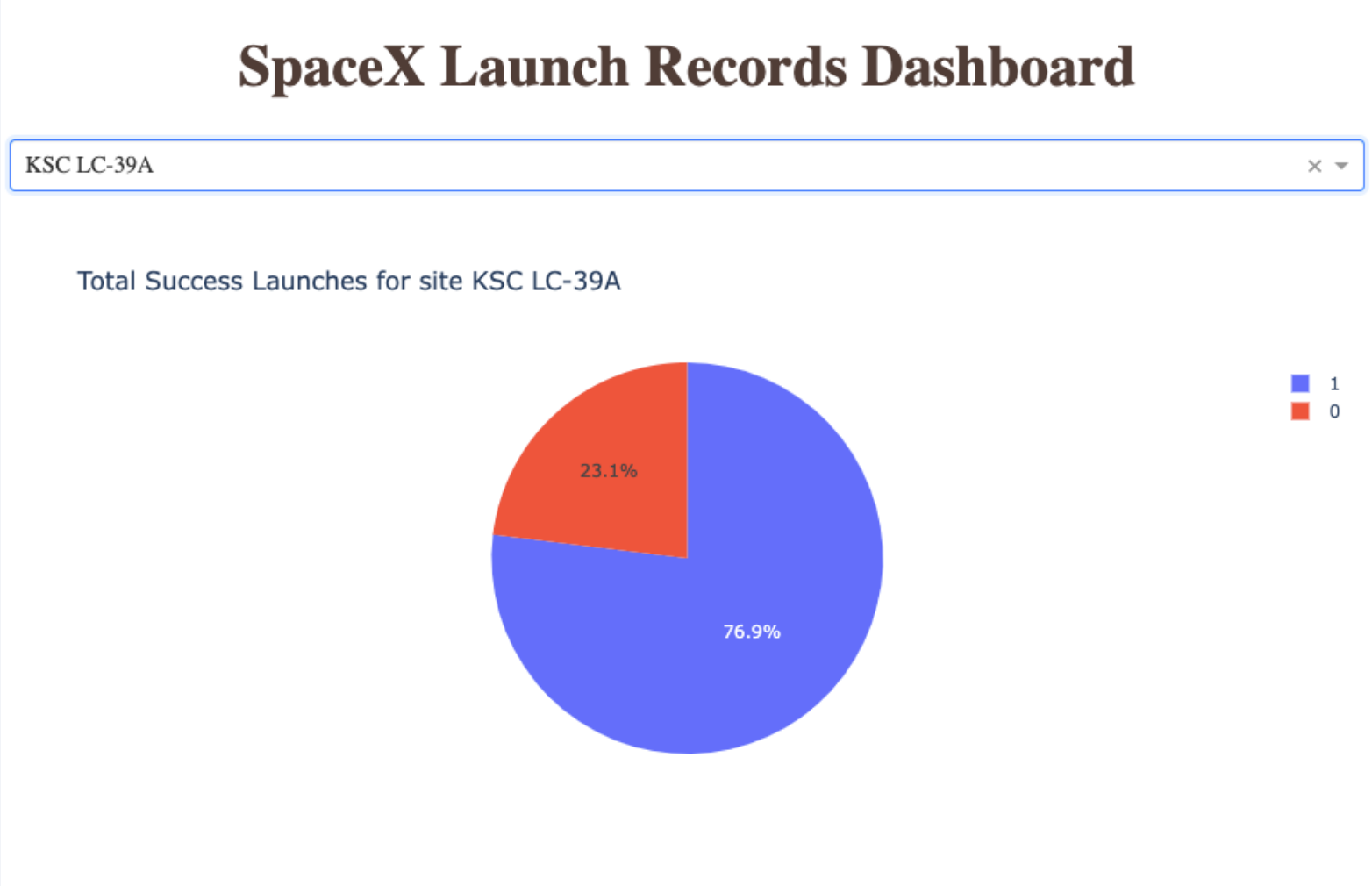
Launch Success Count for all Sites

- The pie chart represents the proportion of launch success count for each site
- Insights:
  - KSC LC-39A has the highest success
  - VAFB SLC-4E is the worst



Drop-Down Selection

- The drop-down menu enables the user to select the launch site
- Insights:
  - The launch site with the highest launch success ratio is KSC LC-39A



Range Slider

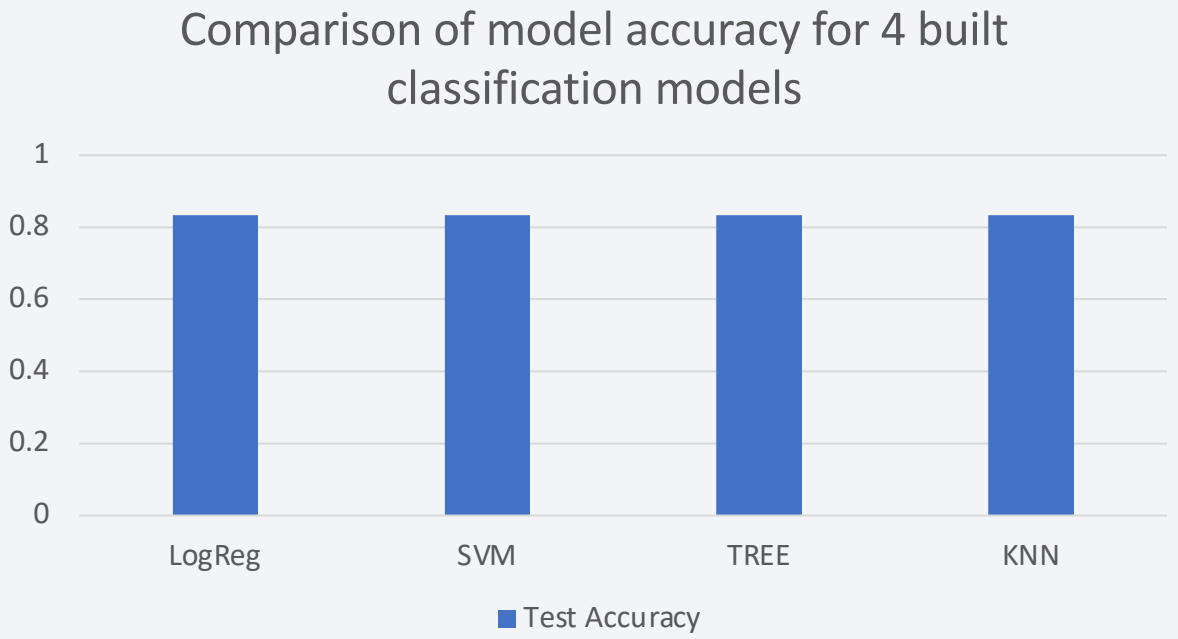
- Comparison of two different payload mass ranges
  - Left: payload range: 0-1000 Kg
  - Right: 4000-6000 Kg
- Insights:
  - FT booster version has a lower success rate in the narrow payload range
  - B4 has a higher rate in the narrow range

Section 5

Predictive Analysis (Classification)

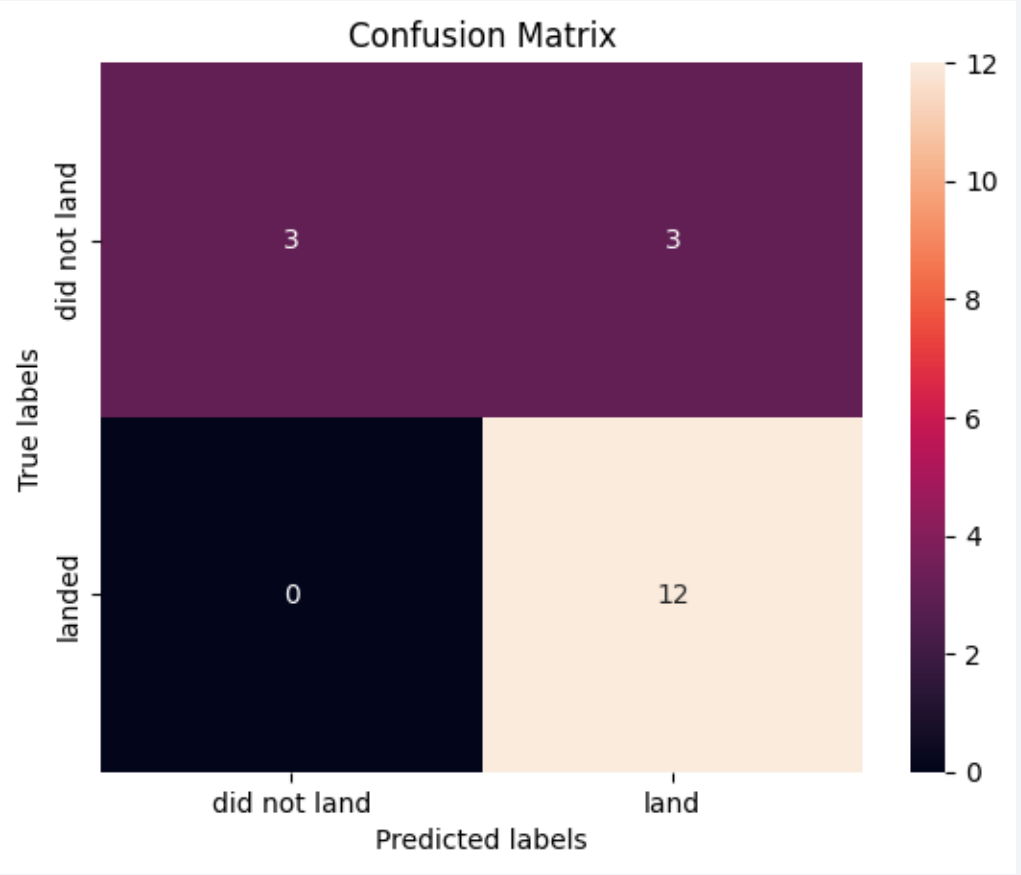
Classification Accuracy

- All models perform similarly
  - Accuracy = 0.833333
  - This could be due to the small size of the data



Confusion Matrix

- Confusion matrix for the KNN model
  - All models perform the same
  - All have similar confusion matrix



Conclusions

- Over time, the success rate of the Falcon9 launches increased
- Different methods are available to predict the success rate with acceptable accuracy
    - EDA provides insights into the relationship between variables like PayloadMass, LaunchSite, and Orbit that can be used to predict outcomes. For instance:
        - Medium and heavy payloads have high success rates
    - Interactive visualizations enable us to gain further insights. For instance:
        - Proximity to coastline, roads, railroad, etc, is important for a new company in site selection
        - If SpaceX is launching from the KSC LC-39A site, the outcome is most likely a success.
- Classification models enable us to provide accurate predictions about the outcome, enabling a rival company to compete with SpaceX.
    - Any of the four models developed here (LogReg, SVM, Decision Tree, KNN) can be used with similar accuracy (0.83333).

Appendix

- Libraries Used:
  - Pandas
  - NumPy
  - datetime
  - requests
  - BeautifulSoup
  - sqlalchemy
  - ipython-sql
  - csv
  - sqlite3
  - Seaborn
  - matplotlib.pyplot
  - folium
  - MarkerCluster
  - MousePosition
  - DivIcon
  - dash
  - plotly.express
  - sklearn (scikit-learn)

Thank you!