Winning Space Race with Data Science

Feng Liu
6/9/2025

Outline

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

Executive Summary

- Objective: Predict the landing success of SpaceX Falcon 9 first stages using machine learning.
- Data Sources:
  - SpaceX API (launch data)
  - Wikipedia (historical launch records)
- Methodology:
  - Data wrangling and EDA using Pandas, SQL, Folium, Plotly
  - Predictive modeling with Logistic Regression, SVM, Decision Tree
- Key Findings:
  - Higher flight numbers, moderate payloads, and specific orbits increase landing success.
  - Decision Tree model achieved the highest accuracy (88%).
- Business Value:
  - Enables cost estimation and risk assessment for competitive launch bids.

Introduction

- Project background and context
  - Space X advertises Falcon 9 rocket launches with a much smaller cost comparing to other providers due to the reuse of its first stage.
  - The success of the first stage can determine the cost of a launch.
  - This information can be used for an alternate company to bid against space X for a rocket launch.

- Problems you want to find answers
  - Create a machine learning pipeline to predict if the first stage will land.

Section 1

Methodology

Methodology

Executive Summary

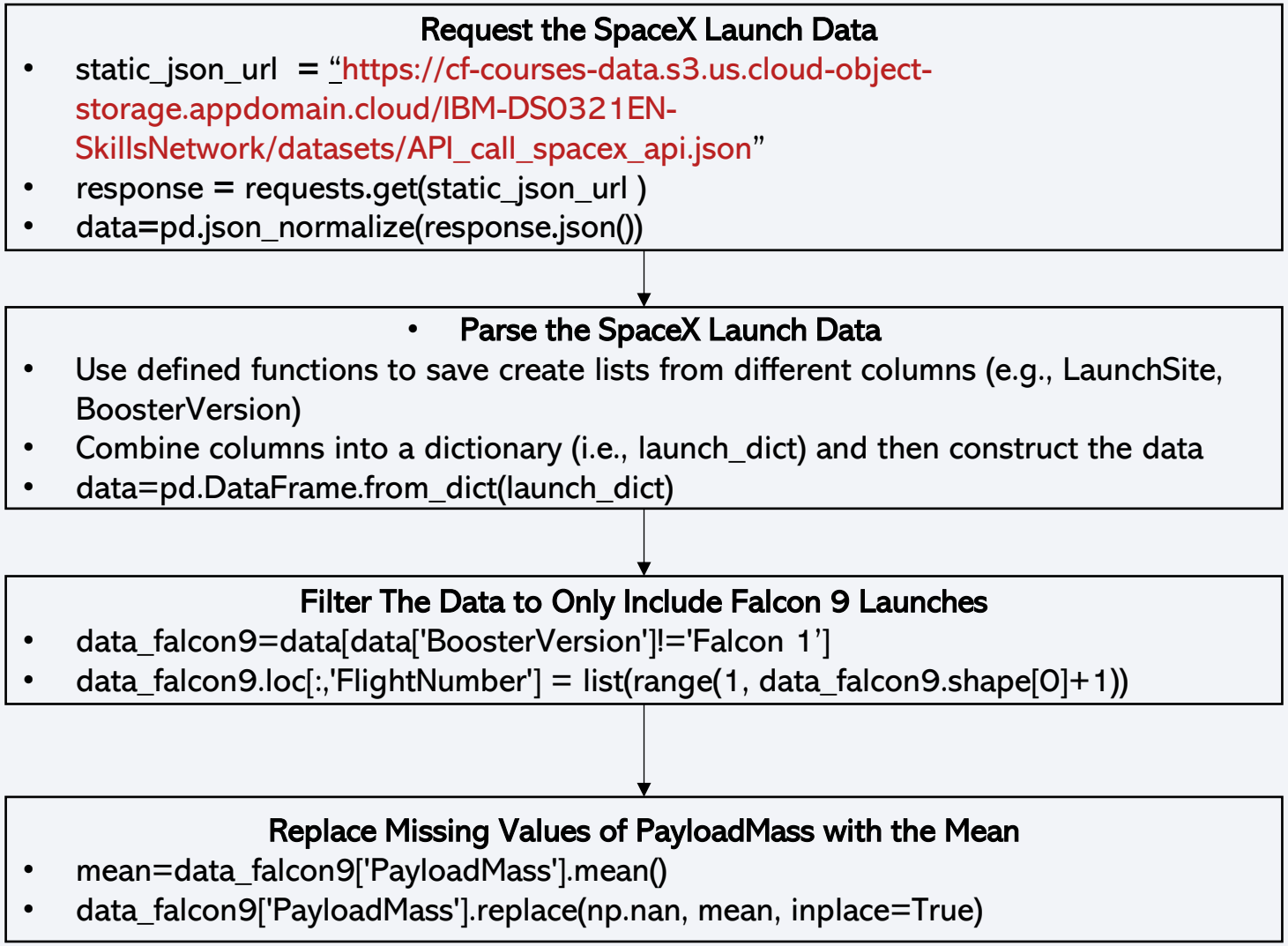
- Data collection methodology:
  - Requested and parsed from SpaceX API
  - Scraped from a Wikipedia page
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Requested and parsed from SpaceX API
- Scraped from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”

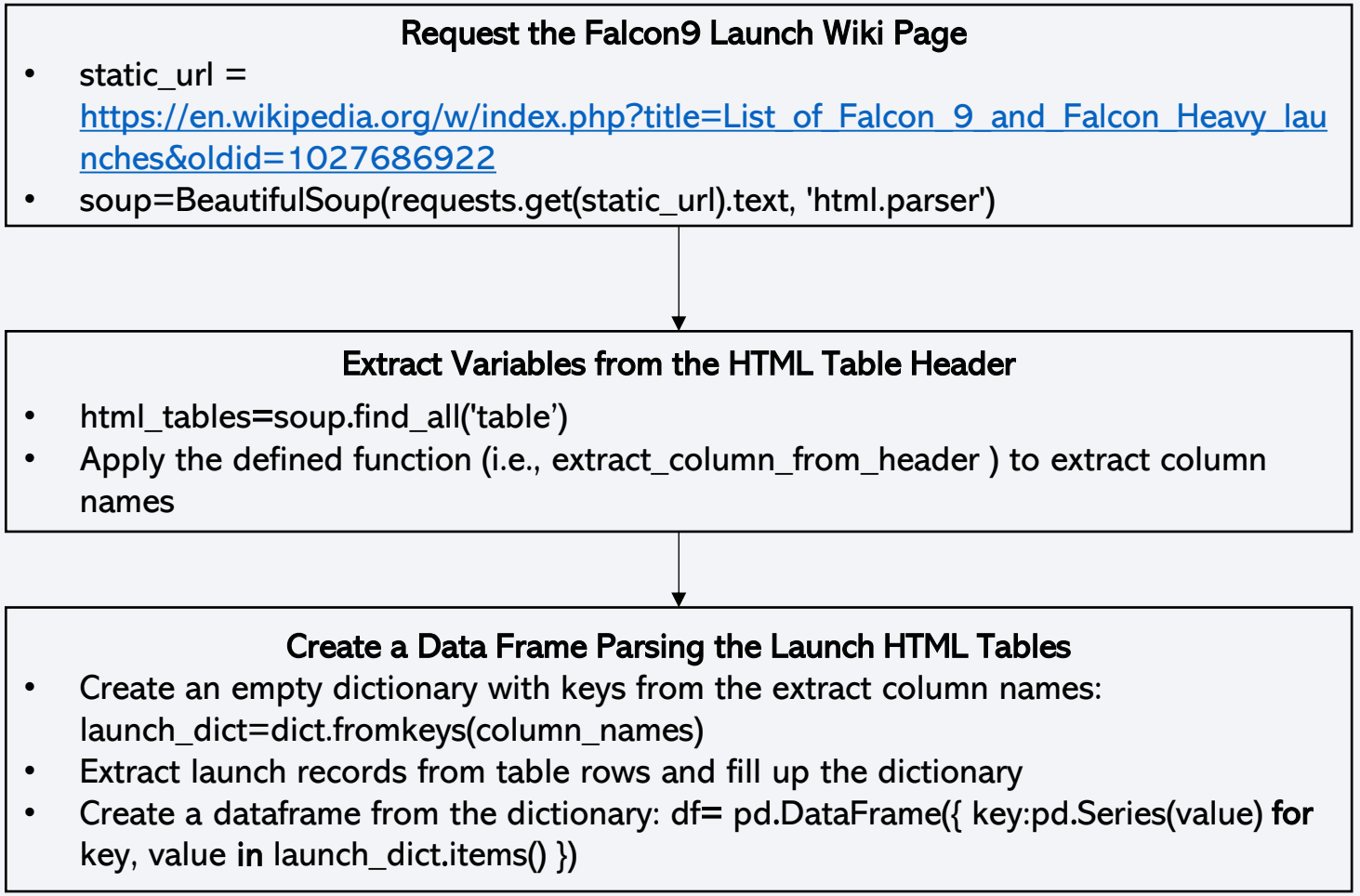
Data Collection – SpaceX API

- GitHub URL of the notebook



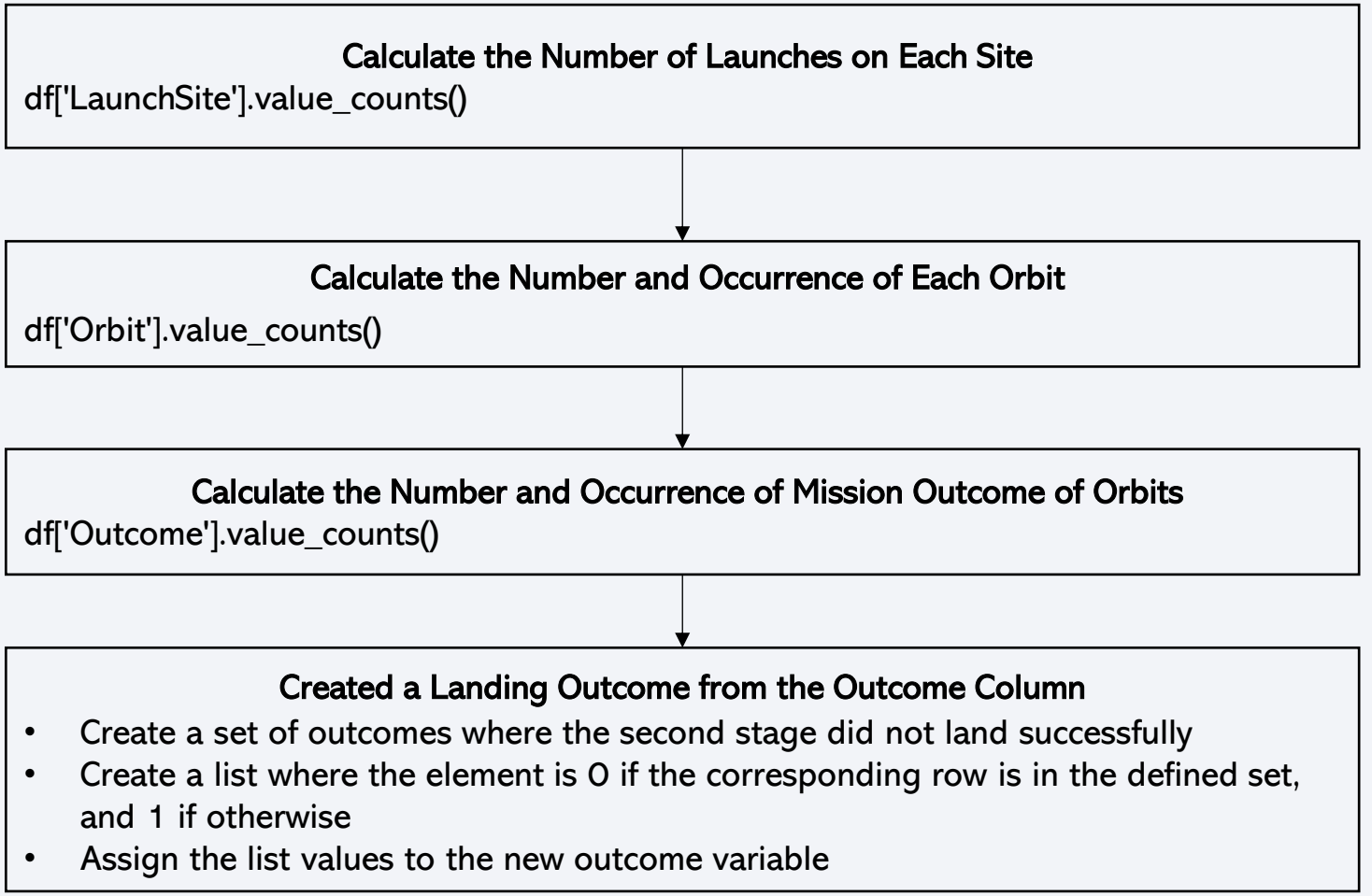
Data Collection - Scraping

- GitHub URL of the notebook



Data Wrangling

- GitHub URL of the notebook



EDA with Data Visualization

- Scatter plot of the following two variables overlaying the outcome to visualize the relationship between the two variables and how they affect the outcome
  - FlightNumber and PayloadMass
  - FlightNumber and LaunchSite
  - PayloadMass and LaunchSite
  - FlightNumber and Orbit
  - PayloadMass and Orbit
- Bar chart of SuccesRate of each Orbit to visualize the relationship between SuccessRate and Orbit type
- Line chart to visualize the success yearly trend
- GitHub URL of the notebook

EDA with SQL

- The name of each unique launch site
  - %sql SELECT DISTINCt(Launch_Site) from SPACEXTABLE
- Five records where launch sites begin with the string ‘CCA’
  - %sql SELECT Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
- The total payload mass carried by boosters launched by NASA (CRS)
  - %sql SELECT sum(PAYLOAD_MASS__KG_) as Total_Payload_NASA_CRS from SPACEXTABLE where Customer='NASA (CRS)'
- Average payload mass carried by booster version F9 v1.1
  - %sql SELECT round(avg(PAYLOAD_MASS__KG_),2) as Mean_Payload_Mass from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
- The date when the first successful landing outcome in ground pad achieved
  - %sql SELECT min(Date) as Date_1stSuccGrdLand from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'

EDA with SQL - Cont.

- The name of boosters with success in drone ship landing and payload mass 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
- The total number of successful and failure mission outcomes
  - %sql SELECT Mission_Outcome, count(*) as Number from SPACEXTABLE group by Mission_Outcome
- All booster versions with the maximum payload mass
  - %sql SELECT Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) as Max_Payload_Mass from SPACEXTABLE)

EDA with SQL - Cont.

- The records with the month, failure landing outcome in drone ship, booster version, launch site for the months in year 2015
  - %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 = 'Failure (drone ship)'
- The count of landing outcomes between the 06/04/2010 and 03/20/2017 ranked in descending order
  - %sql SELECT Landing_Outcome, count(*) as Number from SPACEXTABLE where Date >= '2010-06-04' and Date <= '2017-03-20' group by Landing_Outcome order by LandingOutcomesNum desc
- GitHub URL of the notebook

Build an Interactive Map with Folium

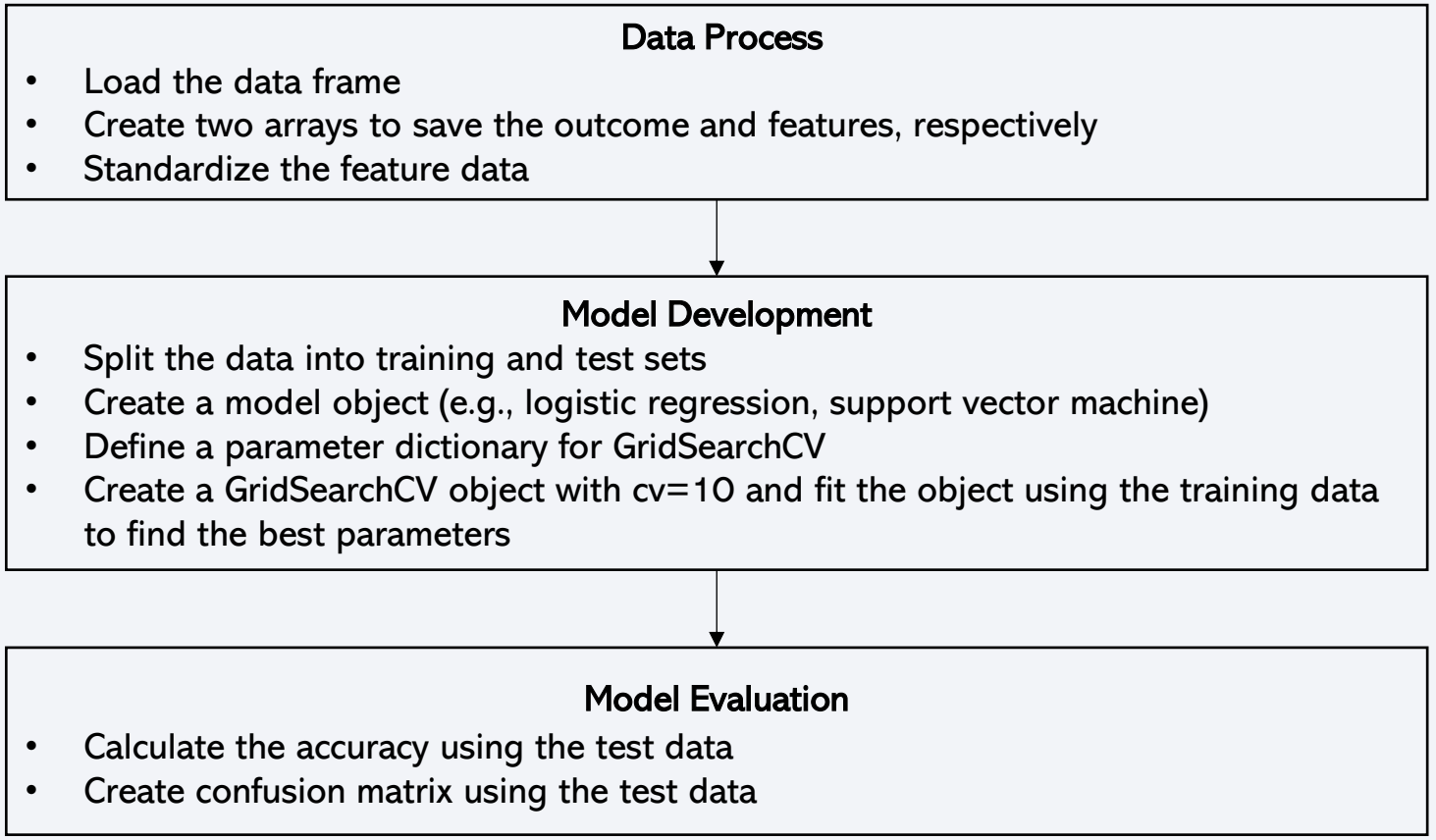
- Added a circle and marker for each launch site to the map to show where these sites are
- Added a marker for each launch site to show their launch records
- Added a line between the launch site CCAFS SLC-40 and the closest coastline and highway to show whether the site is close to coastline or highway.
- GitHub URL of the notebook

Build a Dashboard with Plotly Dash

- Added a dropdown menu to allow users to select a launch site
- Added a pie chart to show either the distribution of success count across all sites (when “All Sites” selected) or the success rate for a selected site
- Added a range slider to allow users to select a payload range
- Added a scatter plot to show the relationship between payload and the launch outcome and whether the relationship varies across different payload ranges selected using the range slider
- GitHub URL of the notebook

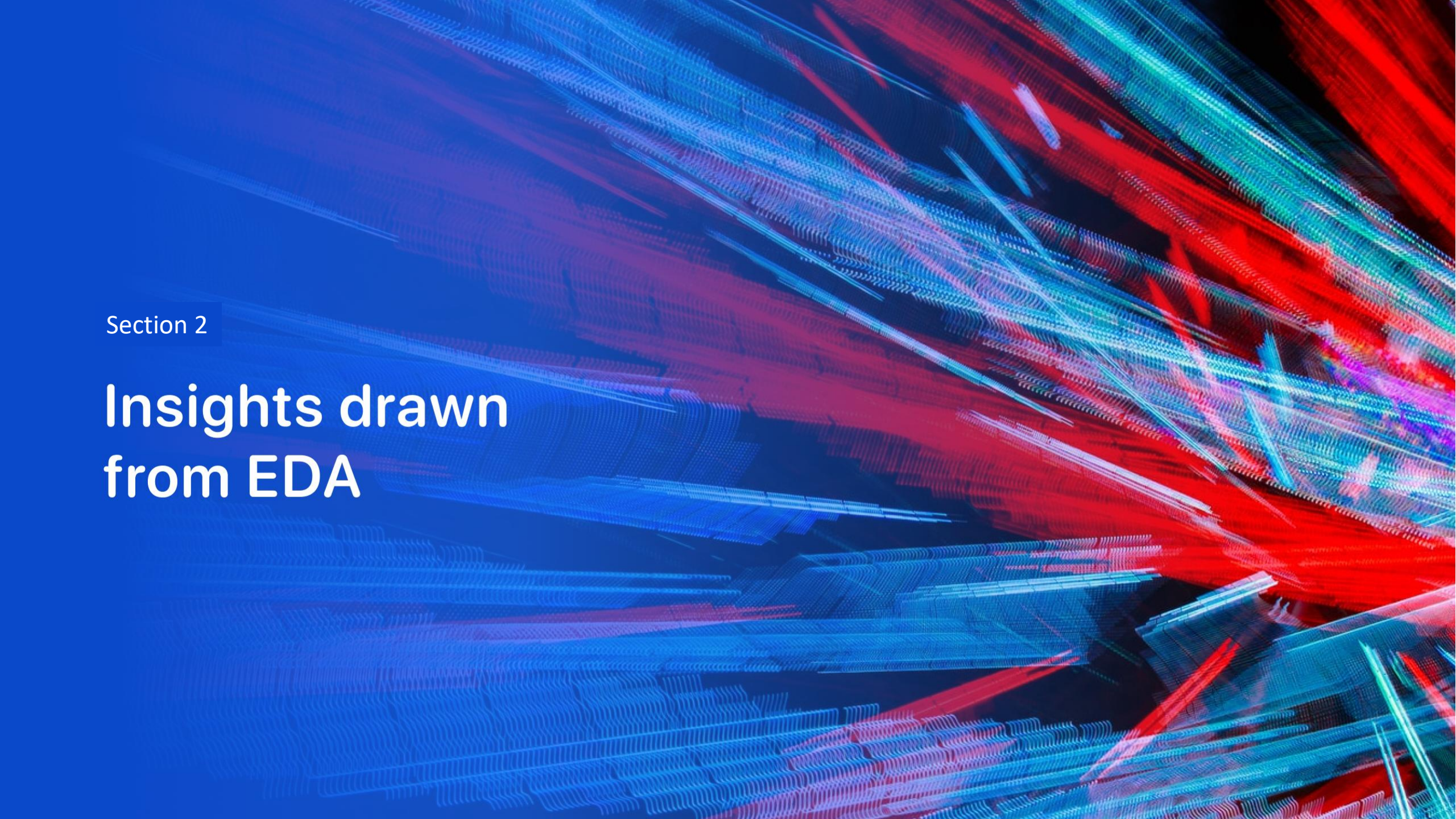
Predictive Analysis (Classification)

- GitHub URL of the notebook



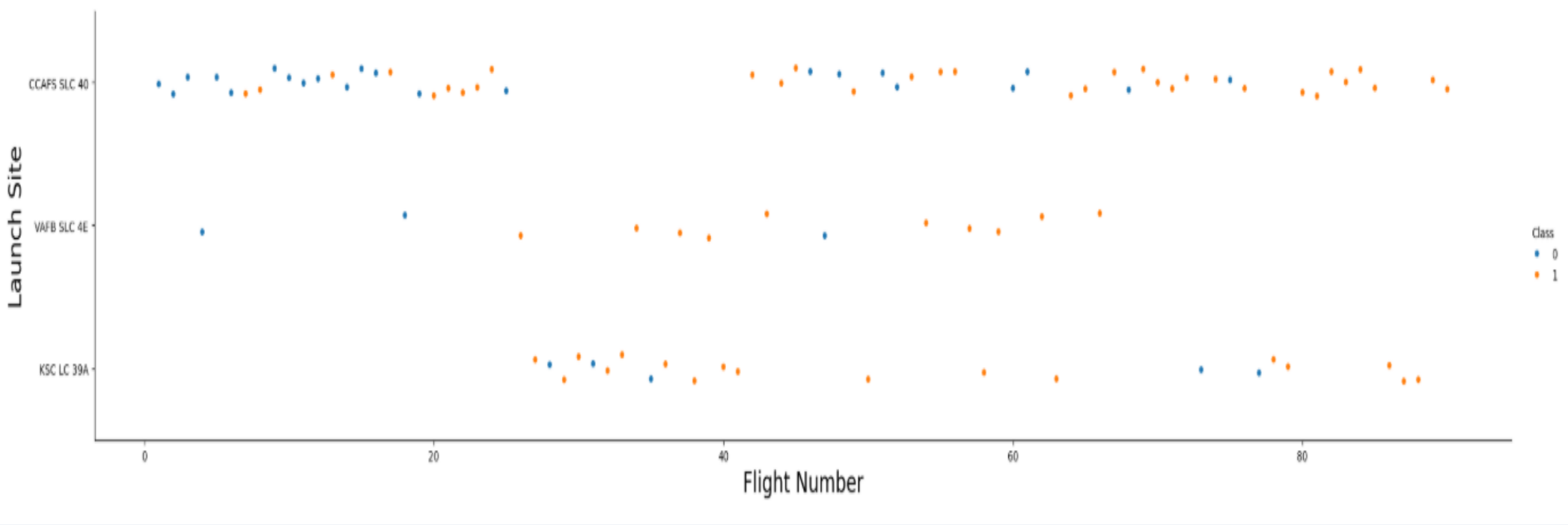
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

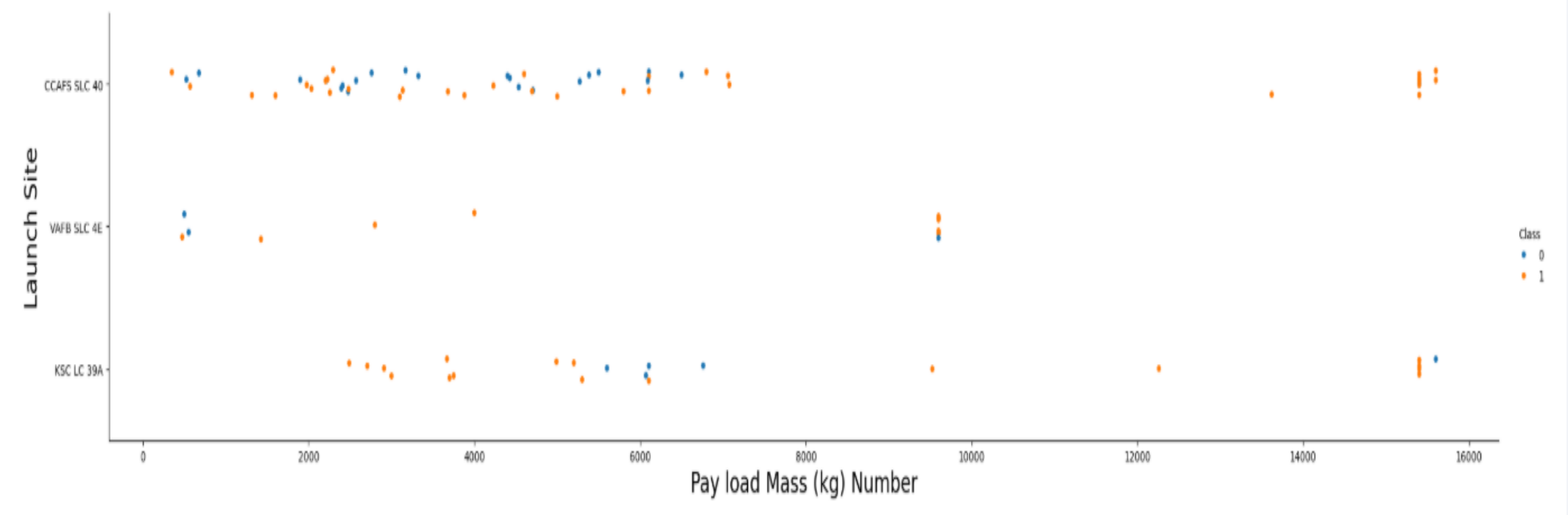
- As the flight number increases, the first stage is more likely to land successfully
- The pattern applies to all launch sites



Class: 1: success, 0: failure

Payload vs. Launch Site

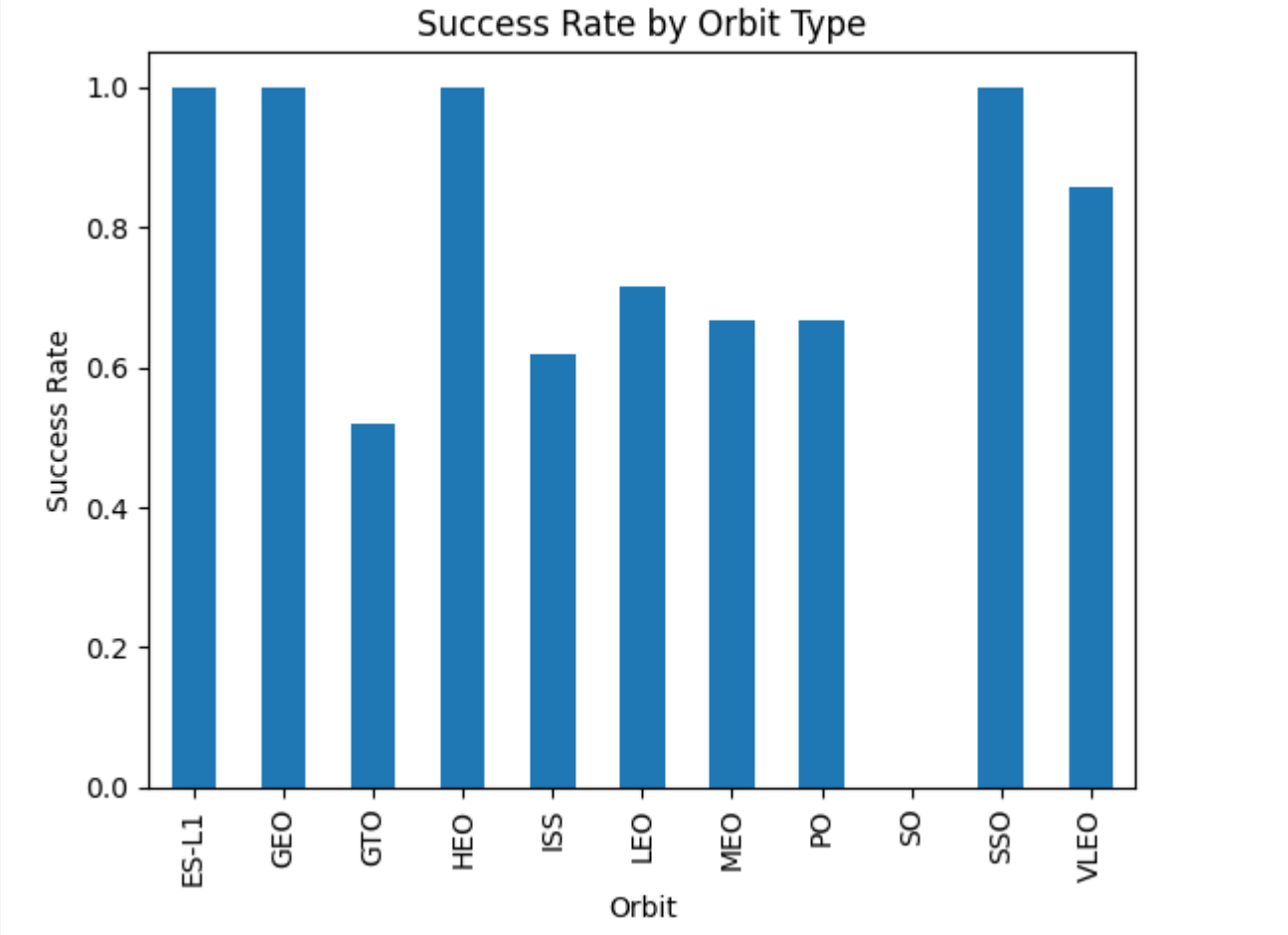
- For VAFB-SLC, no rocket was launched with payload mass more than 10000
- For KSC LC, all rockets were launched with payload mass more than 2000



Class: 1: success, 0: failure

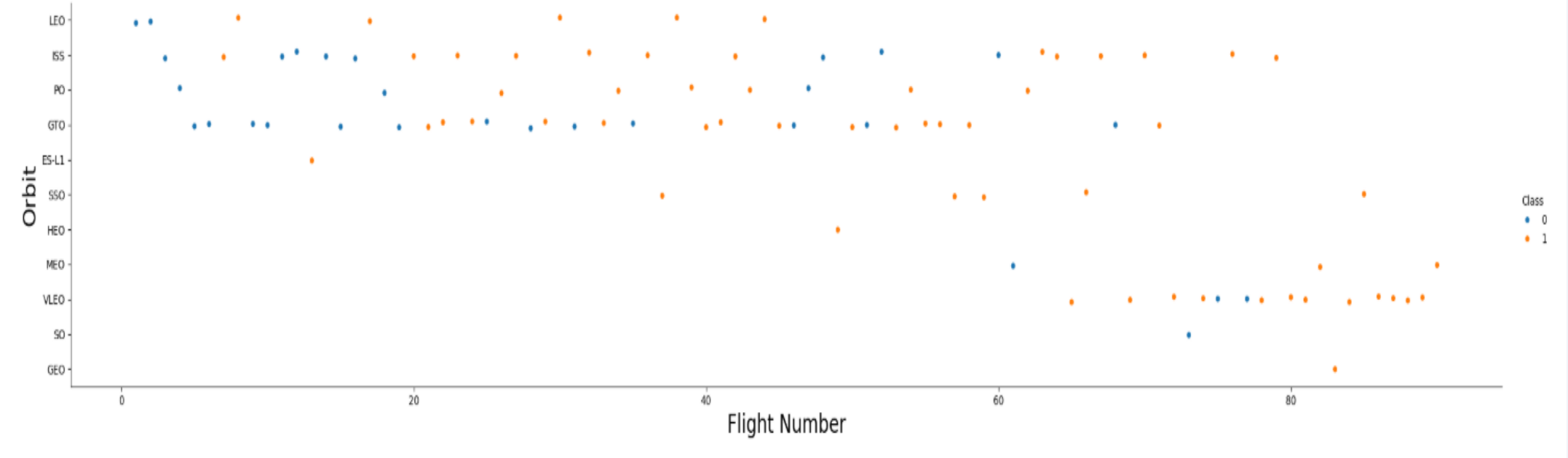
Success Rate vs. Orbit Type

- All sites except GTO have success rate above 60%
- The following sites have 100% success rate: ES-L1, GEO, HEO, SSO



Flight Number vs. Orbit Type

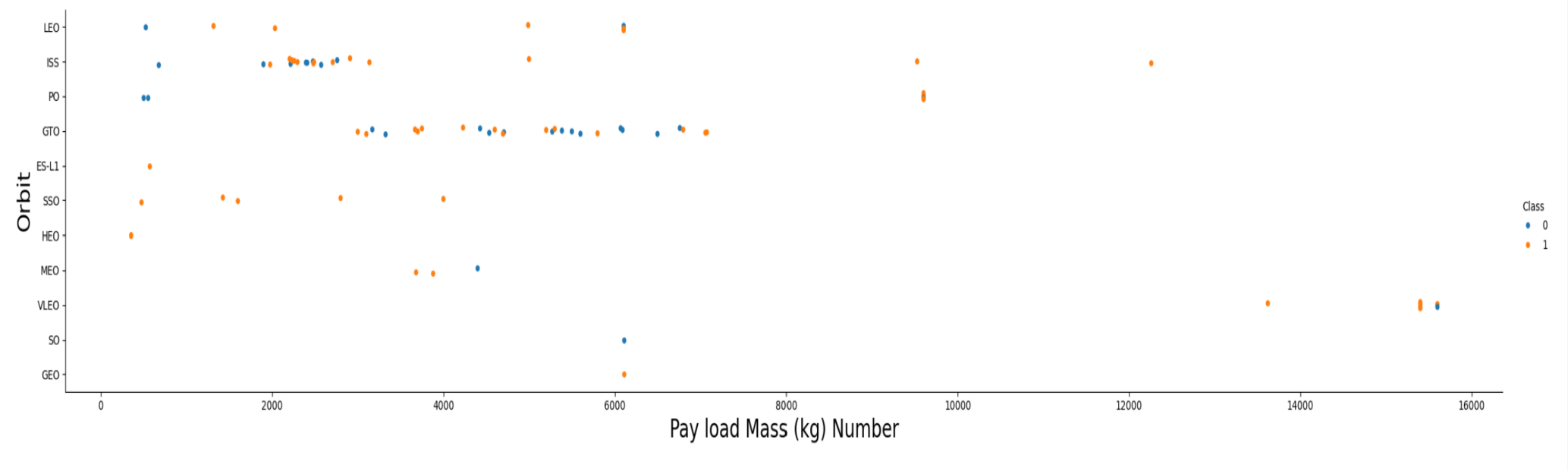
- As the flight number increases, the first stage is more likely to land successfully in LEO orbit
- No clear pattern was identified for other orbits



Class: 1: success, 0: failure

Payload vs. Orbit Type

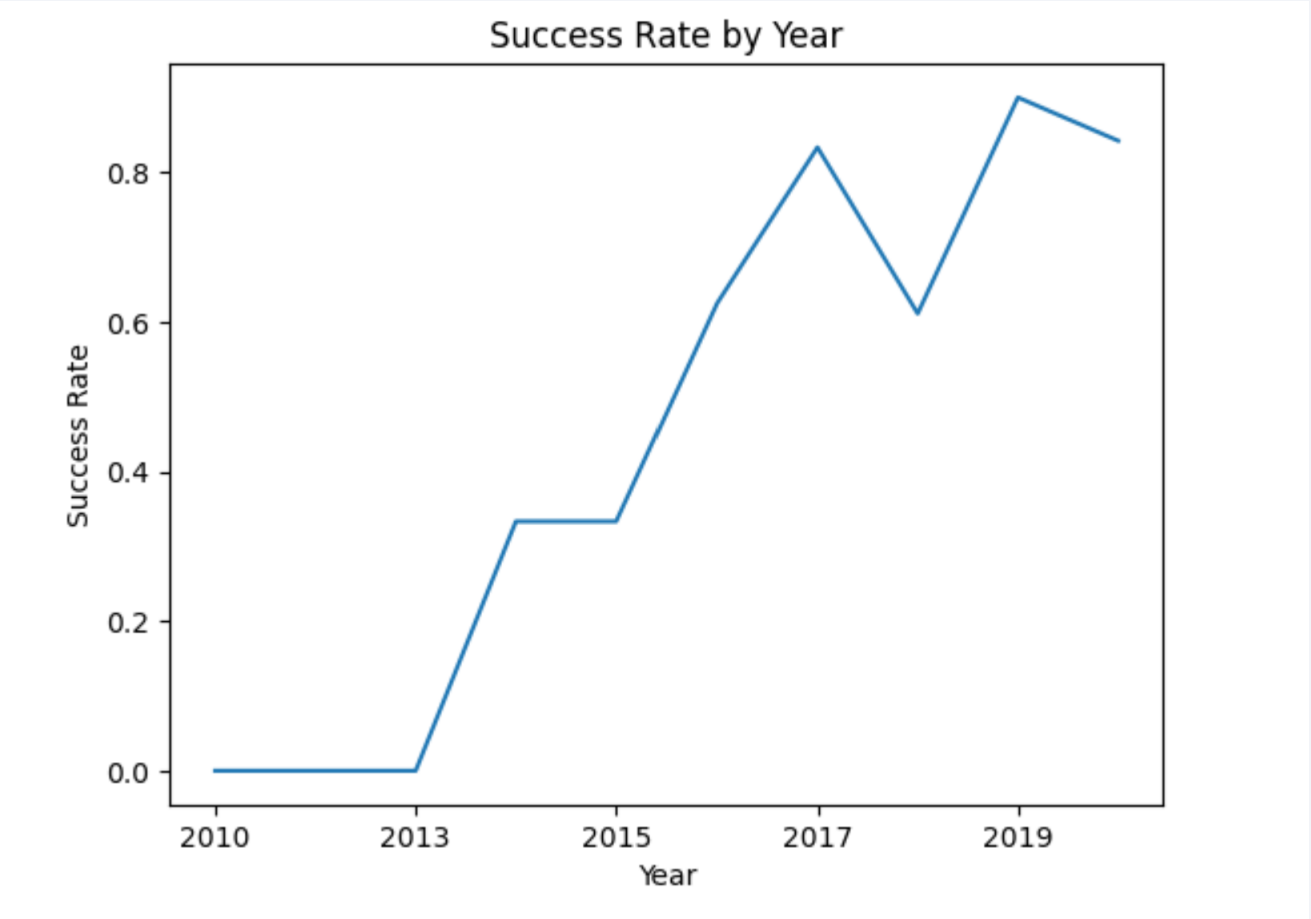
- With heavy payloads, there are more successful landings for PO, LEO and ISS
- No clear pattern was identified for other orbits



Class: 1: success, 0: failure

Launch Success Yearly Trend

- Success rate has increased since 2013 till 2020



All Launch Site Names

- There are four launch sites as shown in the image blow generated using the following query
  - %sql SELECT DISTINCT(Launch_Site) from SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

- The 5 records as shown in the image below were generated using the following query
  - %sql SELECT Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' limit 5

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The total payload carried by boosters from NASA (CRS), 45596, as shown in the image below, was generated using the following query
  - %sql SELECT sum(PAYLOAD_MASS__KG_) as Total_Payload_NASA_CRS from SPACEXTABLE where Customer='NASA (CRS)'

Total_Payload_NASA_CRS
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1, 2534.67, as shown in the image below, was generated using the following query
  - %sql SELECT round(avg(PAYLOAD_MASS__KG_),2) as Mean_Payload_Mass from SPACEXTABLE where Booster_Version like 'F9 v1.1%'

Mean_Payload_Mass
2534.67

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad, 12/22/2015, as shown in the image below, was generated using the following query
  - %sql SELECT min(Date) as Date_1stSuccGrdLand from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'

Date_1stSuccGrdLand
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are four boosters as shown in the image blow generated using the following query
  - %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

- There are total 100 successful and 1 failure mission outcomes, as shown in the image blow, generated using the following query
  - %sql SELECT Mission_Outcome, count(*) as Number from SPACEXTABLE group by Mission_Outcome

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

Boosters Carried Maximum Payload

- There were 12 boosters carrying the maximum payload mass as shown in the image on the right, generated using the following query
  - %sql SELECT Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) as Max_Payload_Mass 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

- There were 2 failed landing_outcomes in drone ship in 2015 as shown in the image below, generated using the following query
  - %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 = '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

- The count of landing outcomes was ranked from high to low, as shown in the image on the right, generated using the following query
  - %sql SELECT Landing_Outcome, count(*) as Number from SPACEXTABLE where Date >= '2010-06-04' and Date <= '2017-03-20' group by Landing_Outcome order by Number desc
- No attempt had the highest count (n=10), followed by Success (drone ship) and Failure (drone ship), each with 5, and Precluded (drone ship) had the lowest count (n=1)

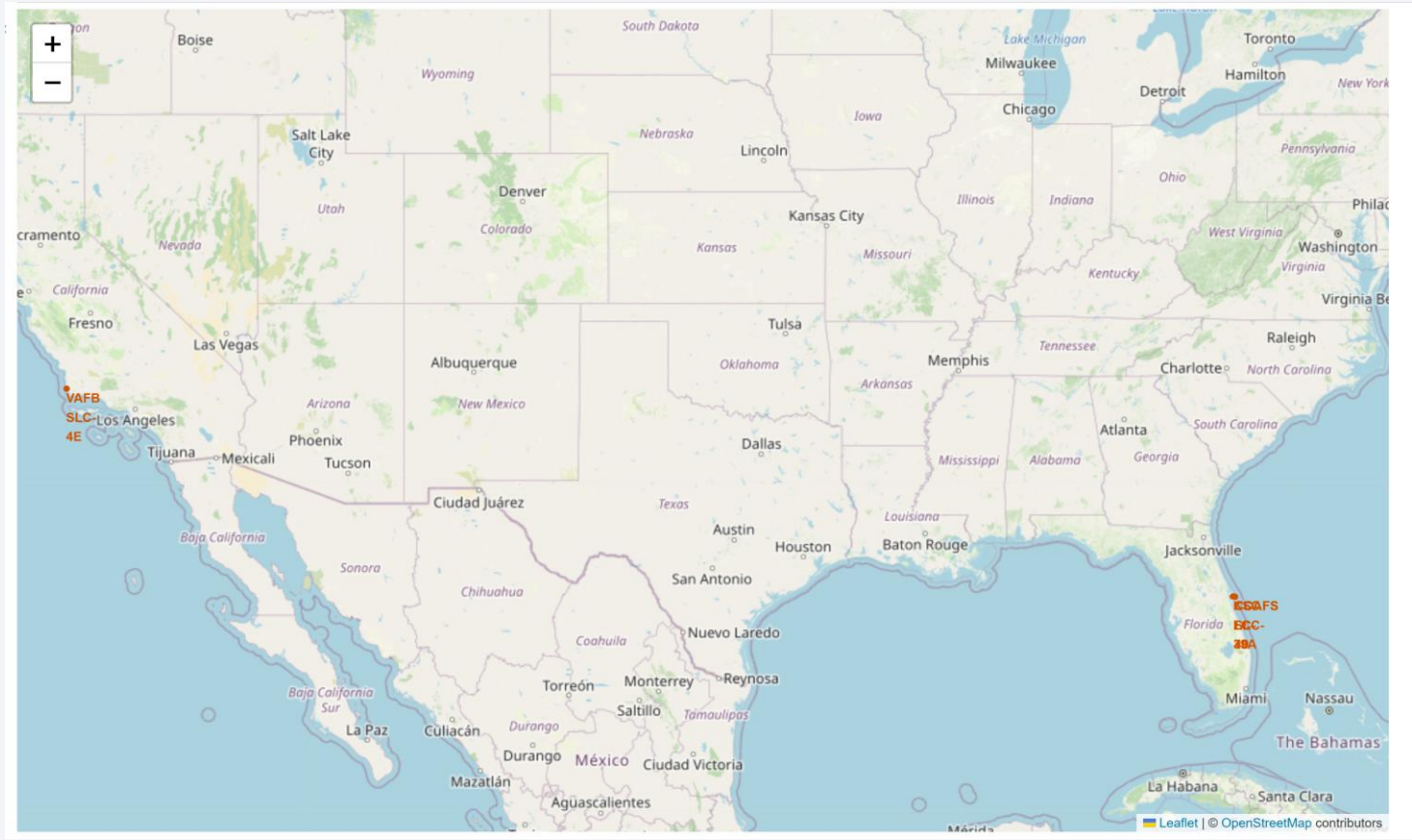
Landing_Outcome	Number
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Site Location

- The map with a circle and marker for each launch site is shown on the right with the circle indicating the location and the marker showing the name for each site
- All launch sites are located near coastlines

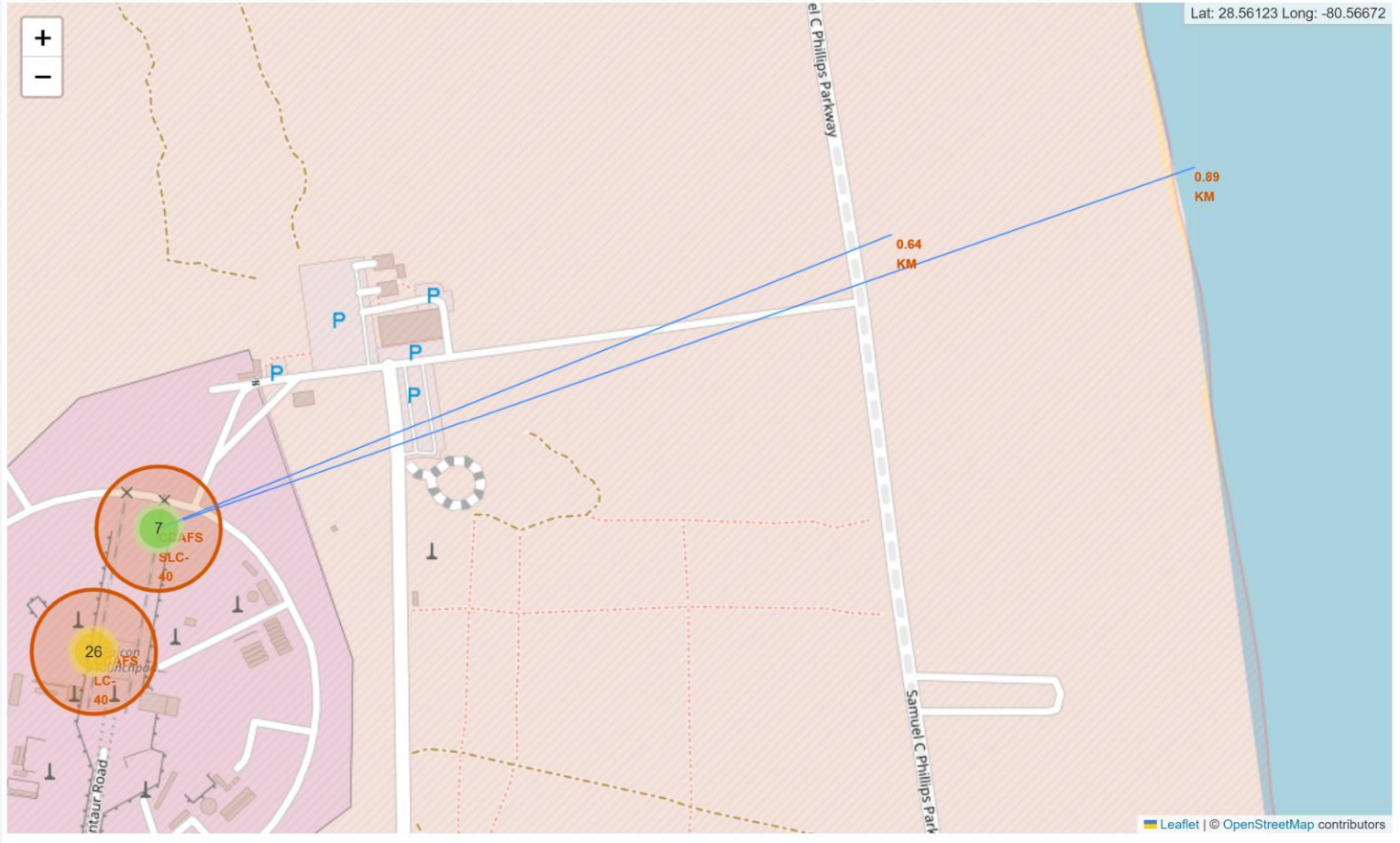


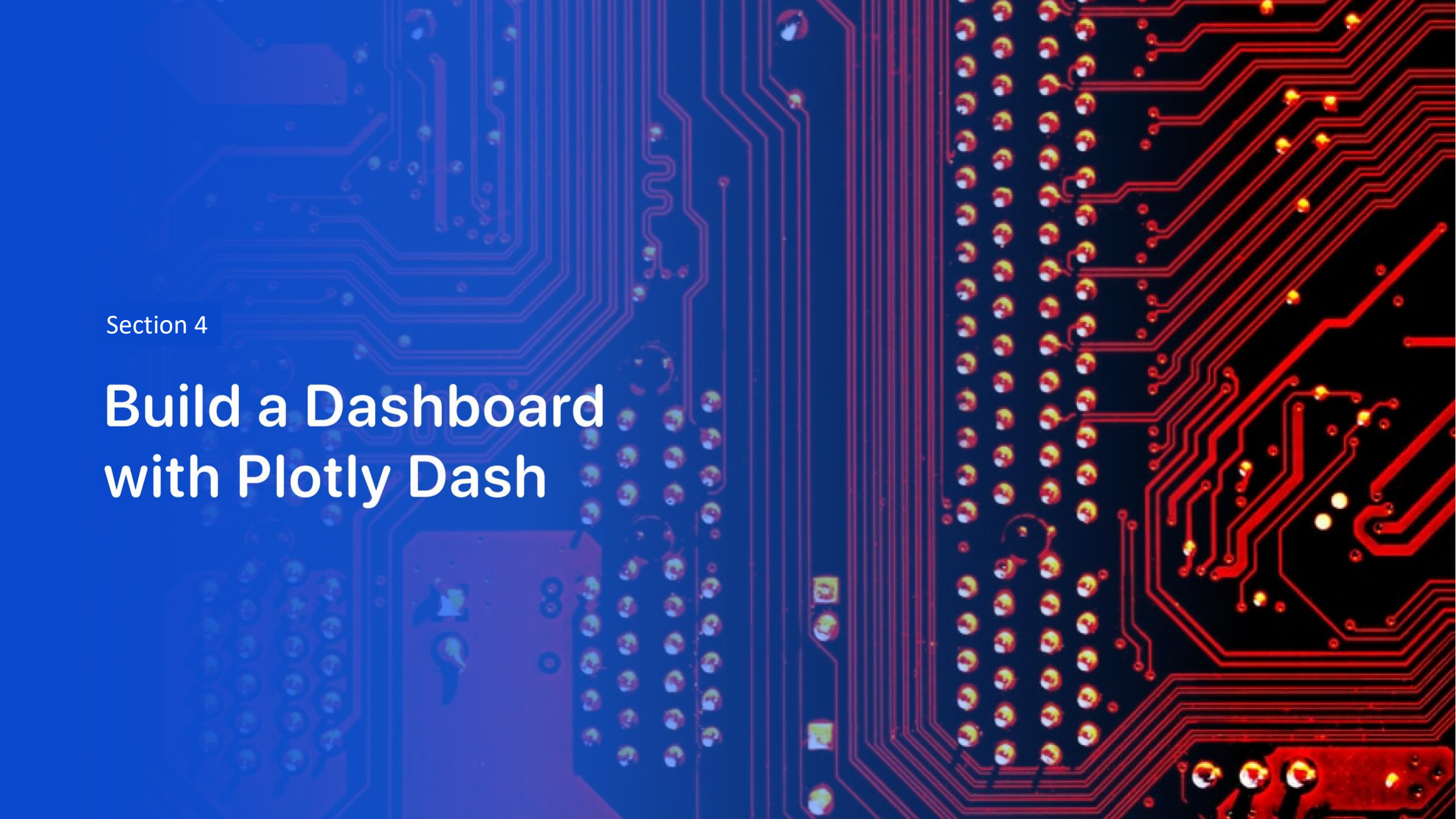
Site Launch Outcomes

- There were total 56 launches in the dataset and 46 of them took place in the sites located in the east coastline and 10 took place in the site located in the west coastline

Site Distance to Proximity

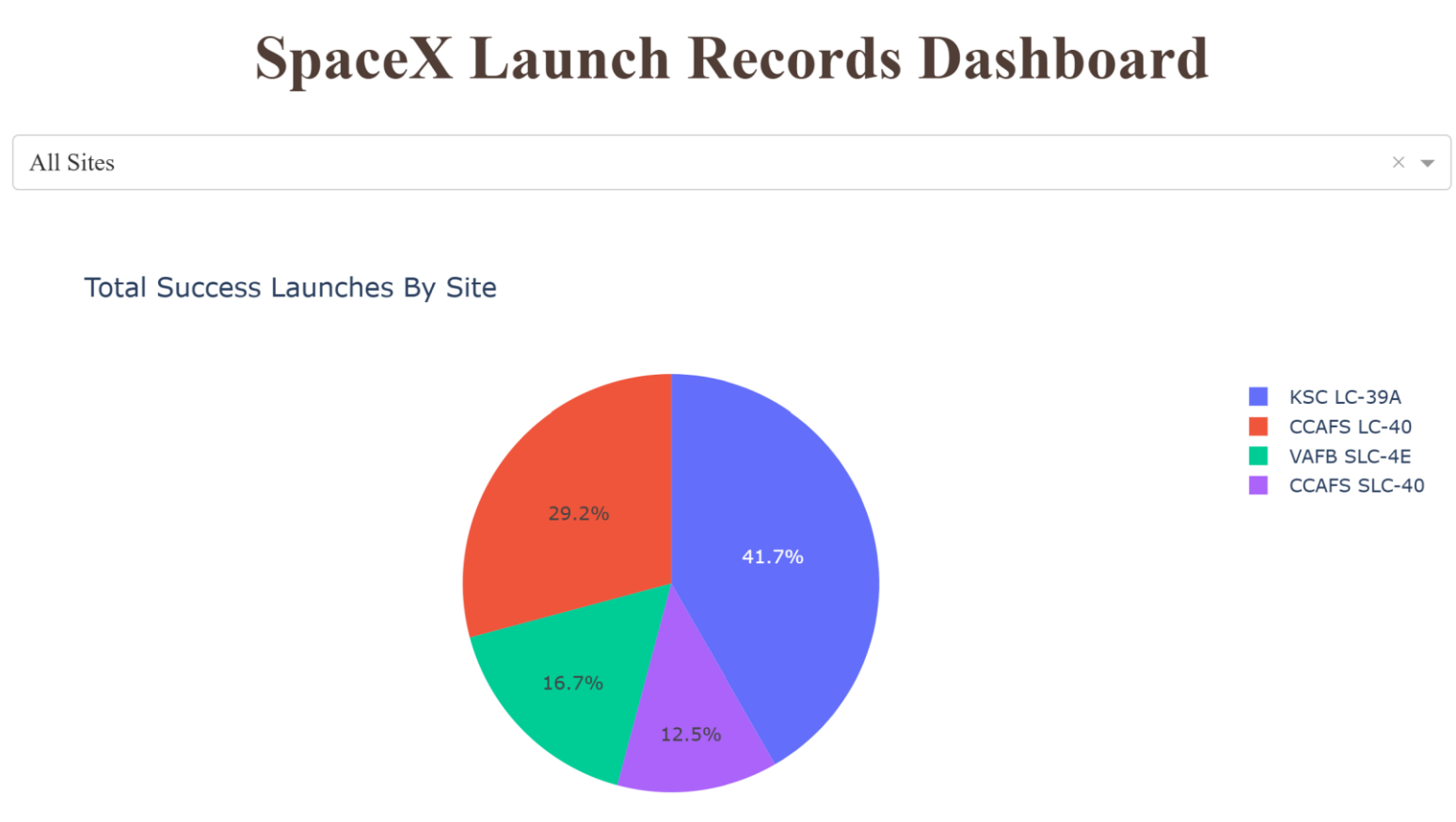
- The map shows the launch site CCAFS SLC-40 is close to a highway and the coastline (all <1KM)




Section 4
Build a Dashboard with Plotly Dash

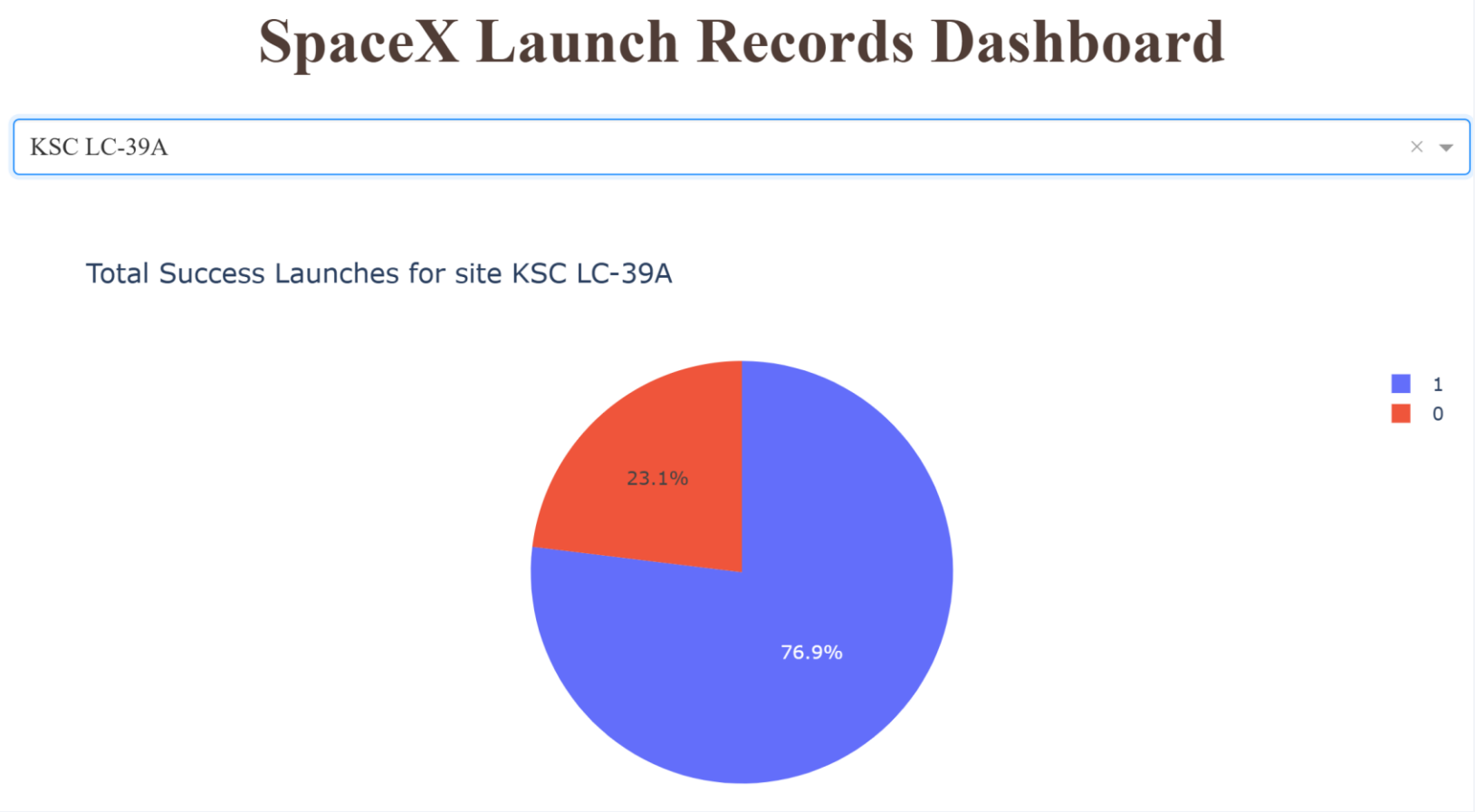
Pie Chart – Success Count

- More than 40% of the success launches took place in KSC LC, about 30% took place in CCAFS LC, 17% took place in VAFB SLS and 13% took place in CCAFS SLC



Pie Chart – Site with Highest Success Rate

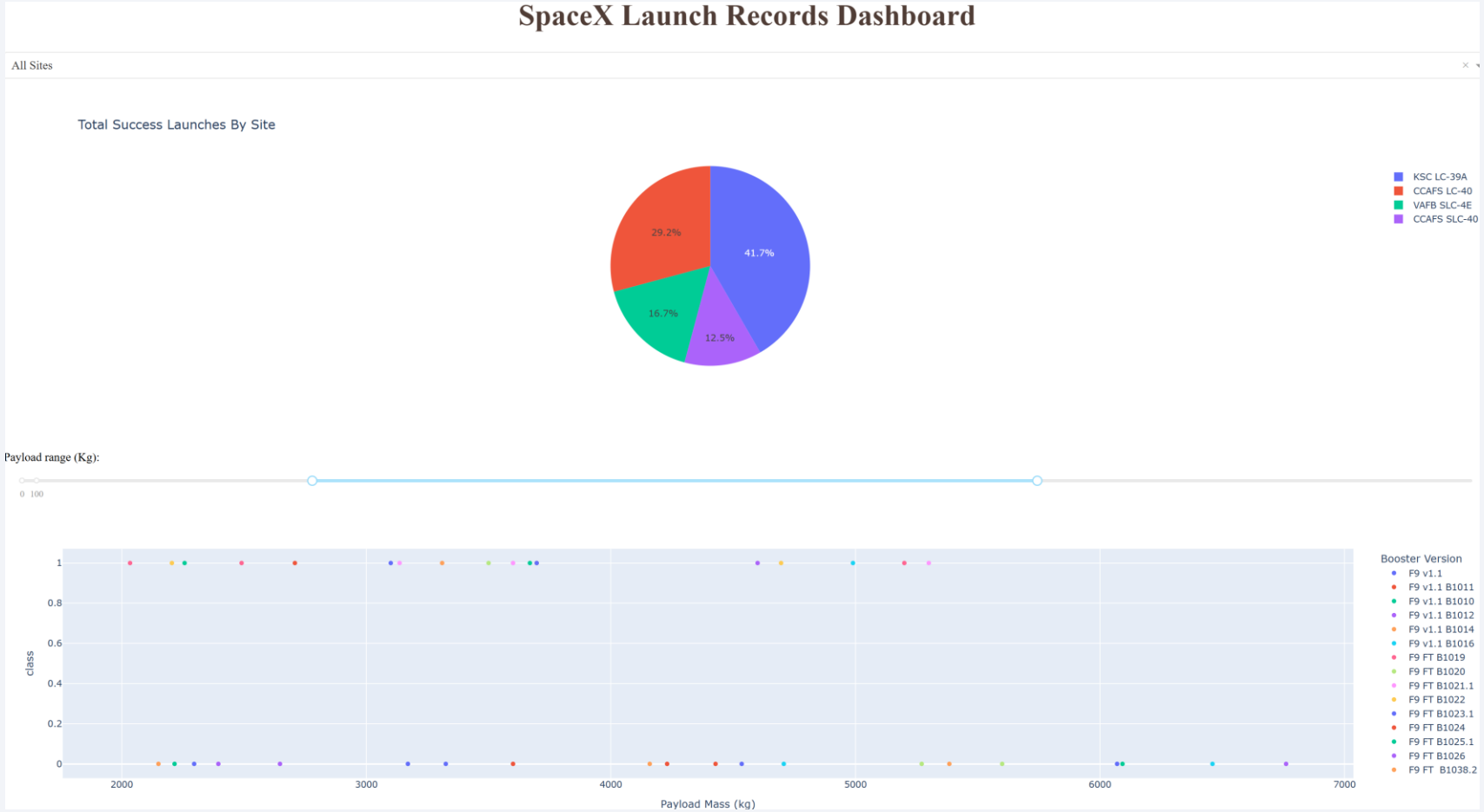
- KSC LC-39A has the highest success rate of 77%



1: success
0: failure

Scatter Plot – Playload vs. Launch Outcome

- Success rate varies by booster version and payload range
- For example, in the payload range between 2000 and 7000, the booster version F9 V1.1 had the highest success rate

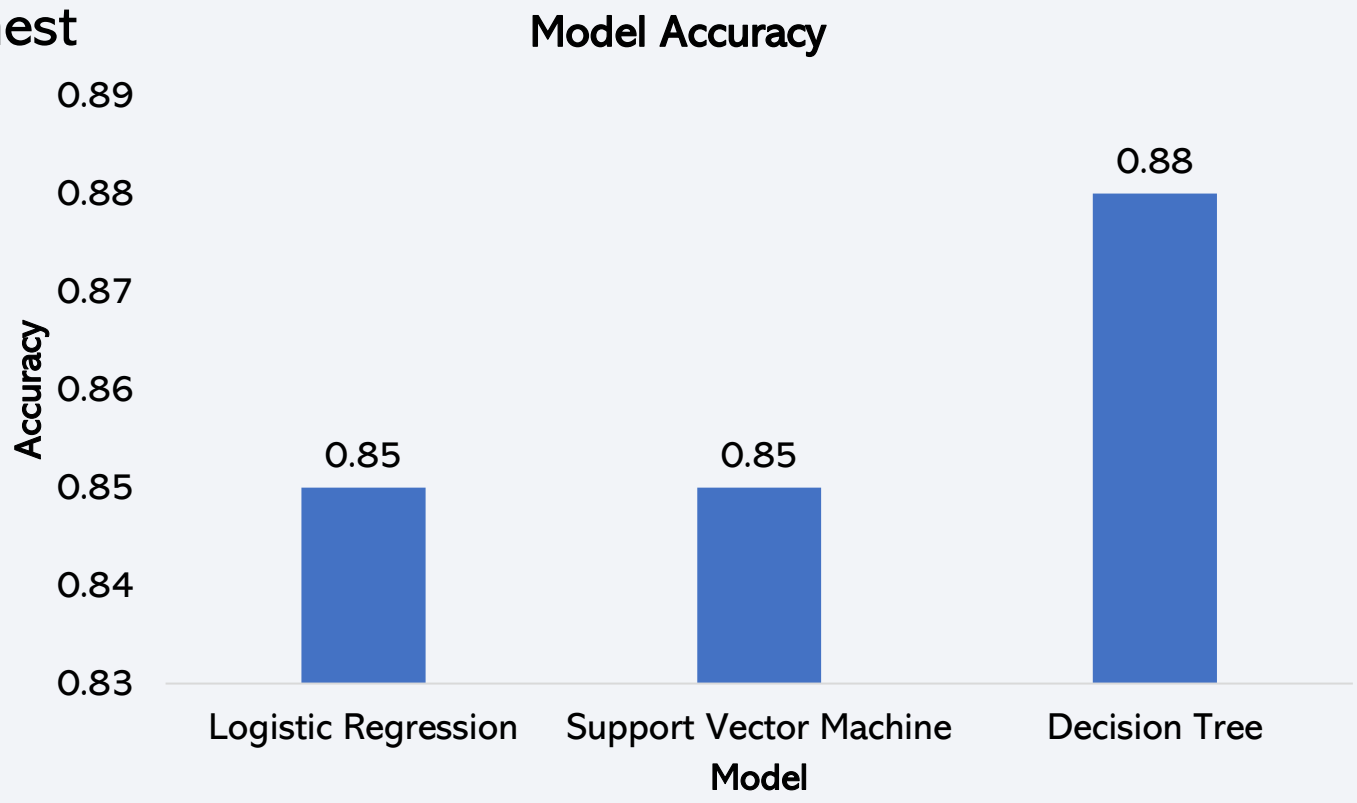


Section 5

Predictive Analysis (Classification)

Classification Accuracy

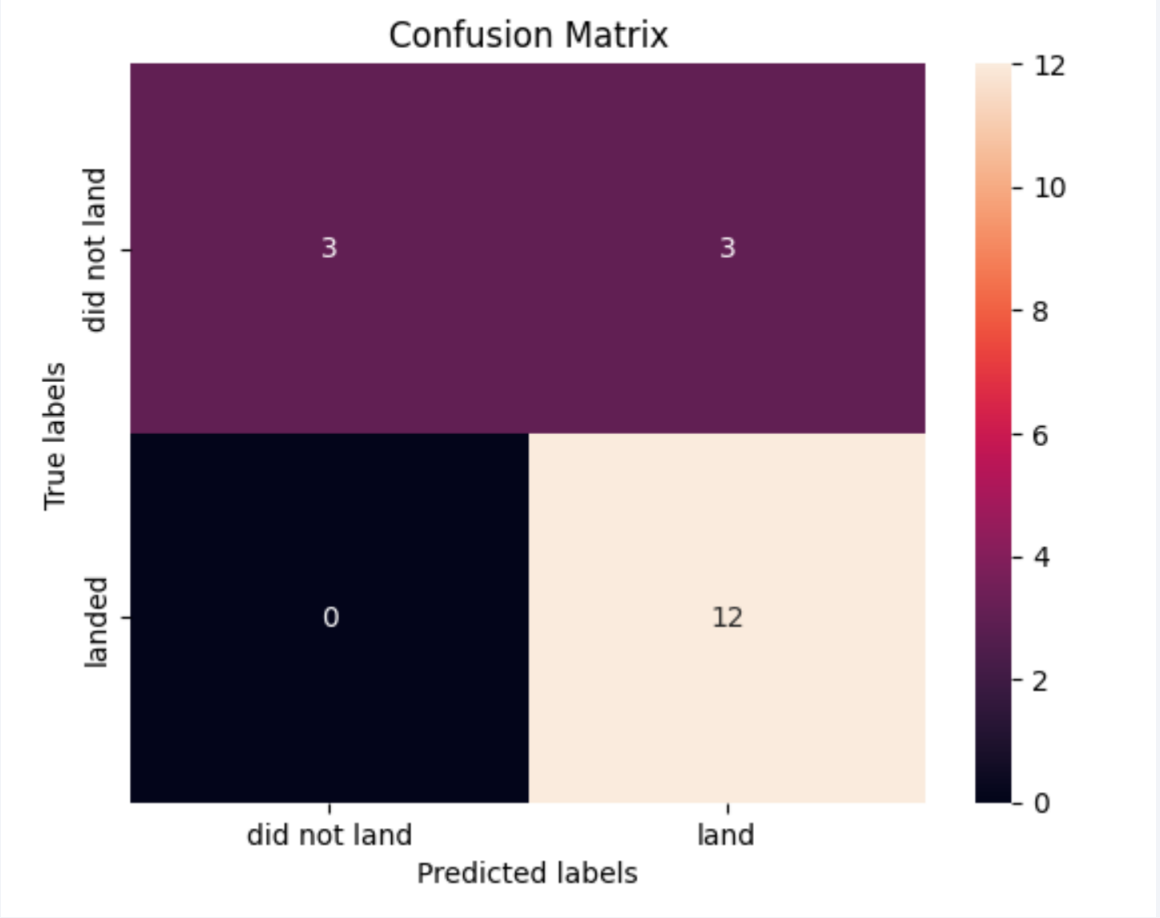
- The decision tree model has the highest classification accuracy



Confusion Matrix

- The following are different statistics in the confusion matrix of the decision tree model on the test data
  - True Positive (TP): 12
  - True Negative (TN): 3
  - False Positive (FP): 3
  - False Negative (FN): 0
  - Precision: TP/(TP+FP)=12/(12+3)=0.8
  - Reall: TP/(TP+FN)=12/(12+0)=1

Confusion Matrix of The Decision Tree Model On Test Data



Conclusions

- As the flight number increases, the first stage of launch is more likely to be successful
- The launch outcome is related to booster version, payload range and orbit type, and the success rate has increased since 2013
- All launch sites are located near coastlines with most launches taking place in the sites located near the east coastline
- The decision tree model has the highest accuracy among the three models tested

Thank you!