IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sergei Lobia
05.03.2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

Summary of results

- Exploratory Data Analysis results
- Interactive analytics demo
- Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing? - Does the rate of successful landings increase over the years? - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

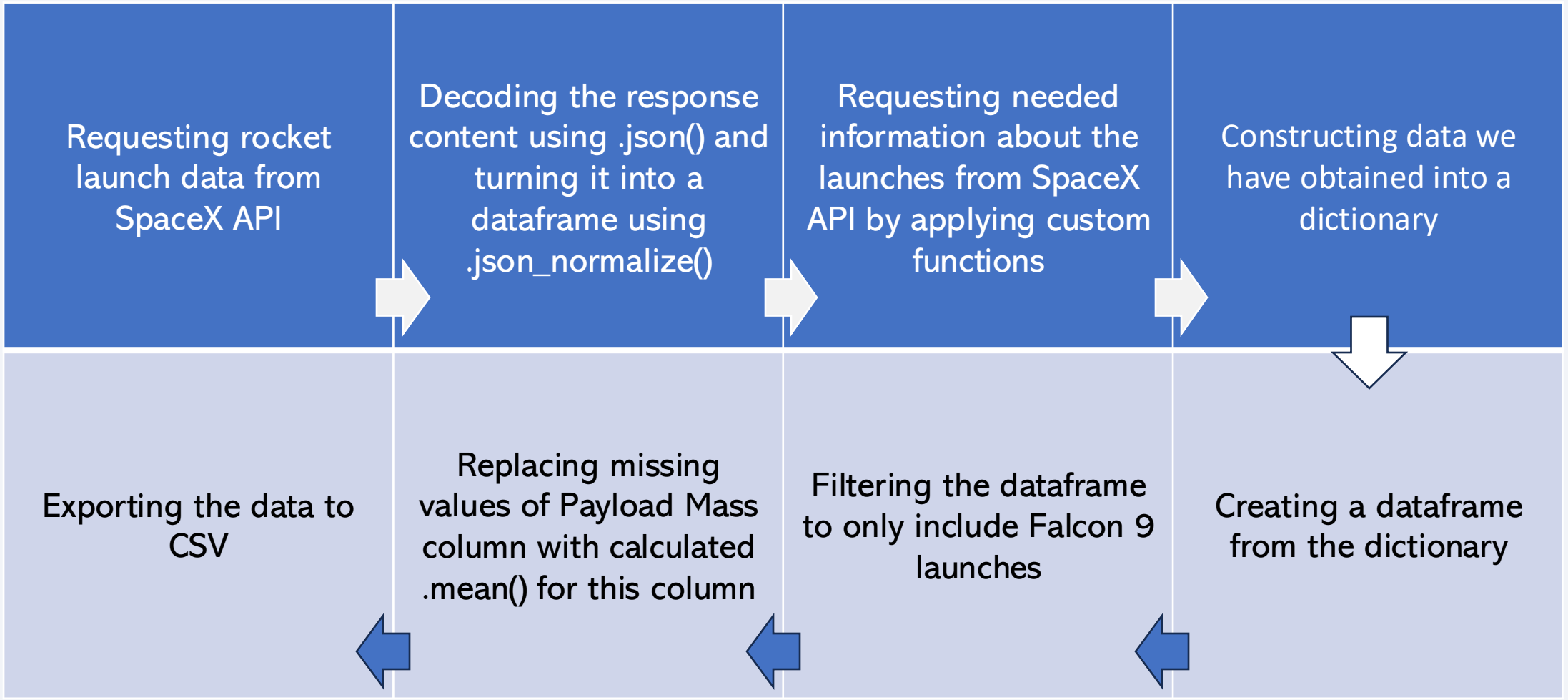
Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both these data collection methods in order get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

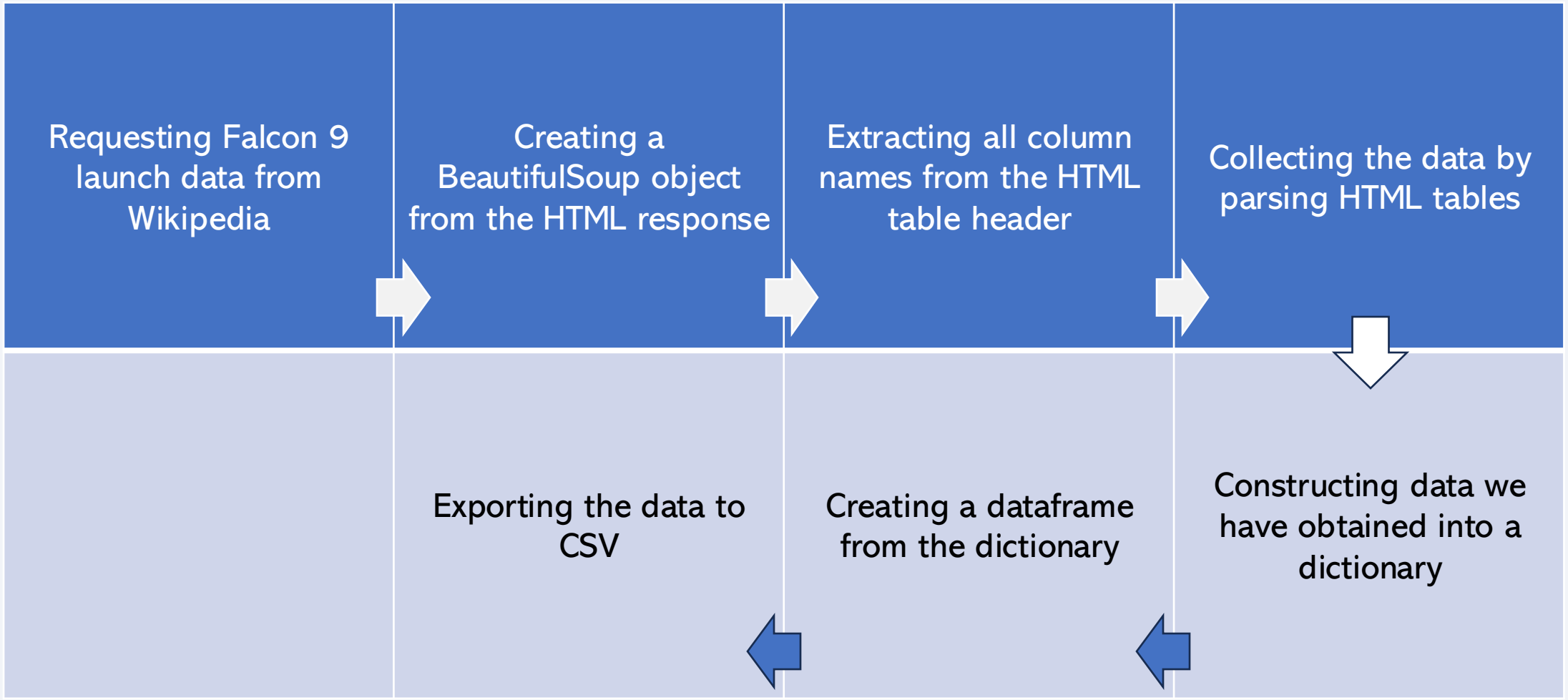
Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



URL: Data Collection API

Data Collection - Scraping



URL: Data collection with web scraping

Data Wrangling

- Performation of the exploratory Data Analysis and determine Training Labels
- Calculation of the number of launches on each site
- Calculation of the number and occurrence of each orbit
- Calculation of the number and occurrence of mission outcome per orbit type
- Landing outcome label from 'Outcome' column creation
- Exporting the data to CSV

URL: Data Wrangling

EDA with Data Visualization

Charts were plotted:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

Scatter plots show the relationship between variables.

Bar charts show comparisons among discrete categories.

Line charts show trends in data over time.

URL: EDA with Data Visualization

EDA with SQL

SQL queries are performed:

- Names of the unique launch sites
- Launch site names begin with `CCA`
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- First successful ground pad landing date
- Successful drone ship landing with payload between 4000 and 6000
- Total number of successful and failure mission outcomes
- List of boosters carried maximum payload
- Year 2015 records
- Rank success count between 2010-06-04 and 2017-03-20

URL: EDA with SQL

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

The launch success rate may depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

URL: Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash

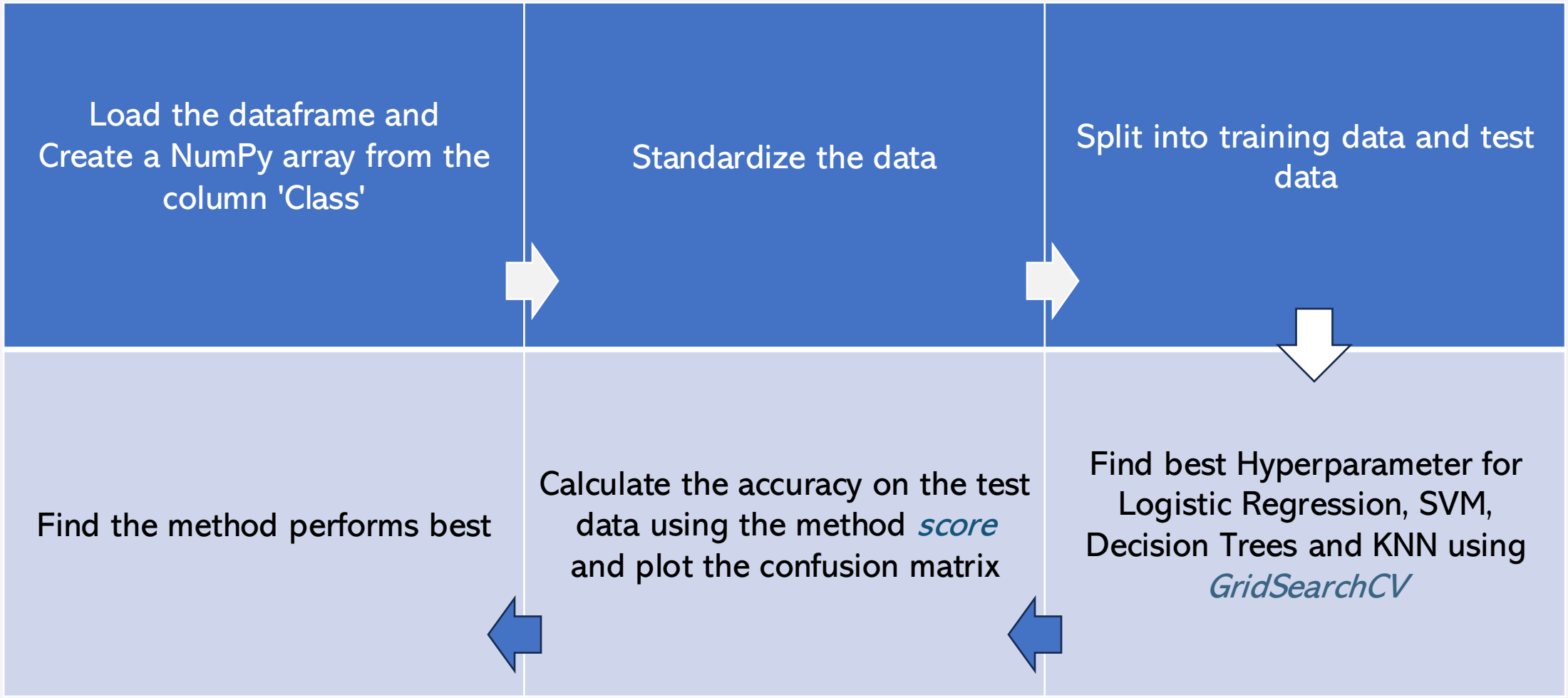
Plots/graphs and interactions on dashboard:

- Launch success count for all sites
- Launch site with highest launch success ratio
- Payload Mass vs. Launch Outcome for all sites

Visualization of the more successful launch site, also Payload Mass vs. Launch Outcome for all sites.

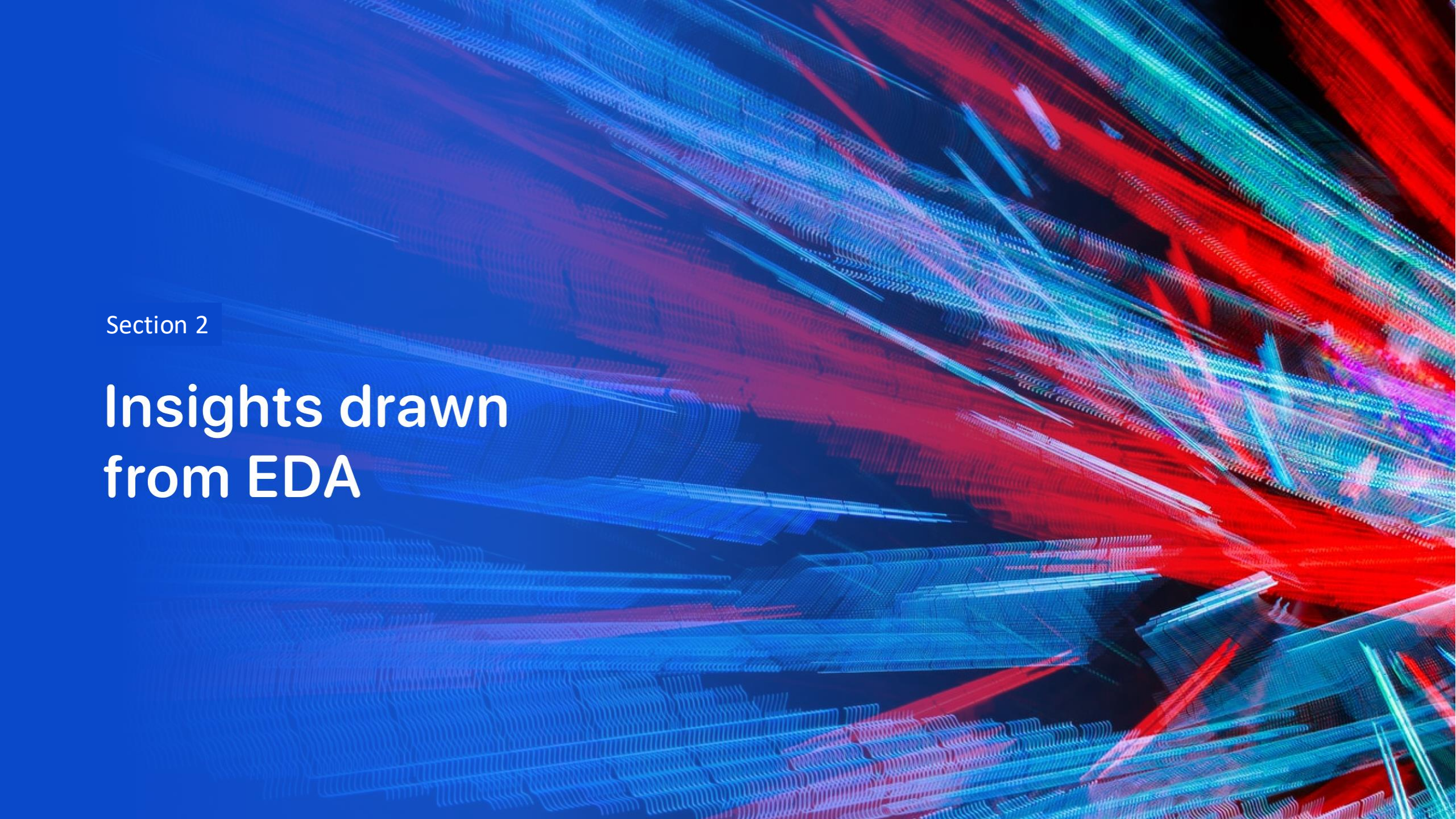
URL: Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

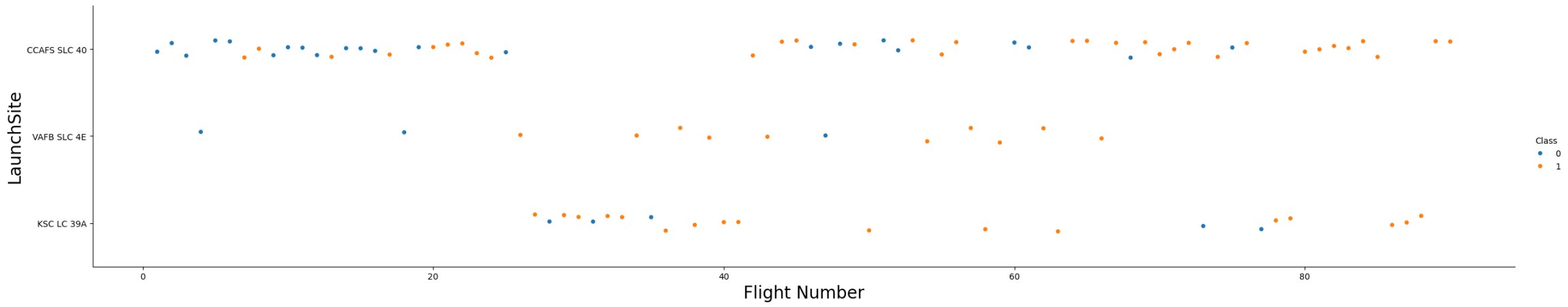


Results

- The success rate of launches is high
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Most of launch sites are in proximity to the Equator line
- KSC LC-39A has the highest success rate of the launches from all the sites
- Decision Trees Model is the best method for this dataset

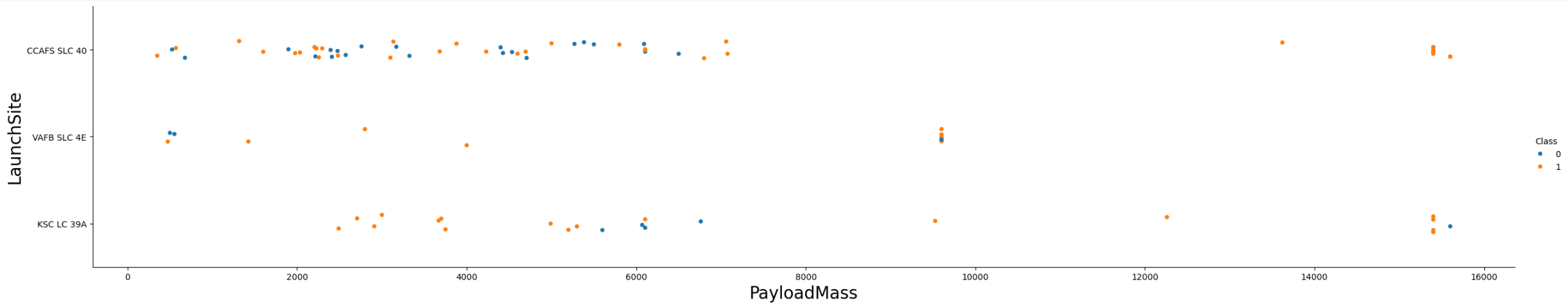

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



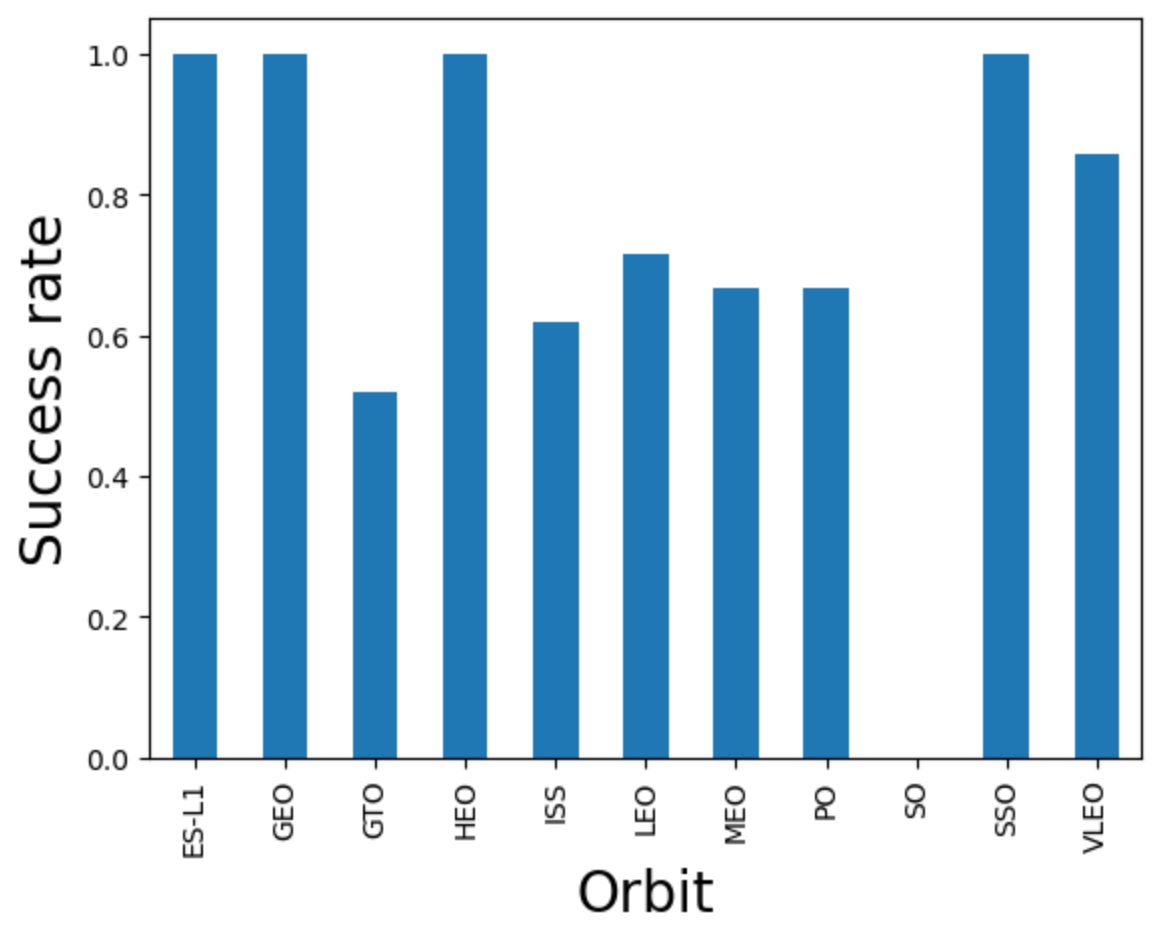
- The CCAFS SLC 40 launch site is more usable

Payload vs. Launch Site



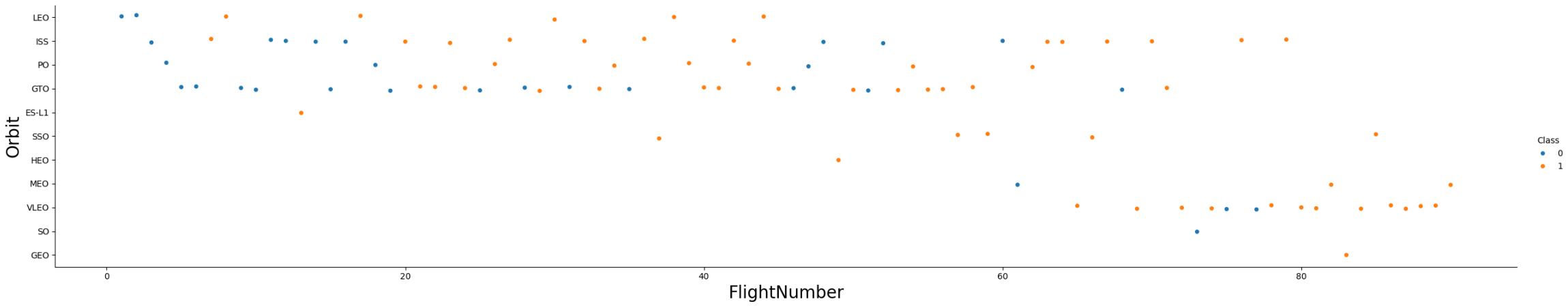
- The CCAFS SLC 40 launch site is more usable as for heavier as for lighter payloads
- The KSC LC 39A launch site used for heavier payloads vs VAFB SLC 4E

Success Rate vs. Orbit Type



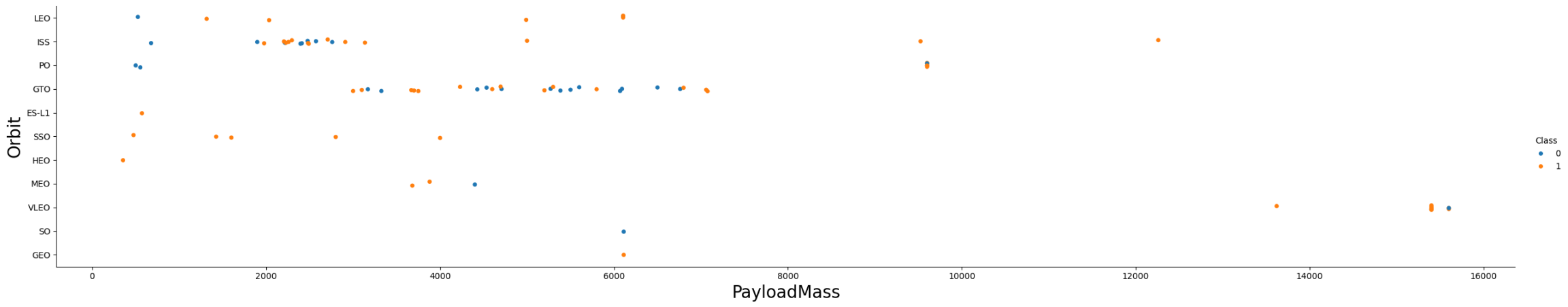
- ES-L1, GEO, HEO, SSO are orbits with 100% success rate
- GTO, ISS, LEO, MEO, PO are orbits with success rate between 50% and 85%
- SO only orbit with 0% success rate

Flight Number vs. Orbit Type



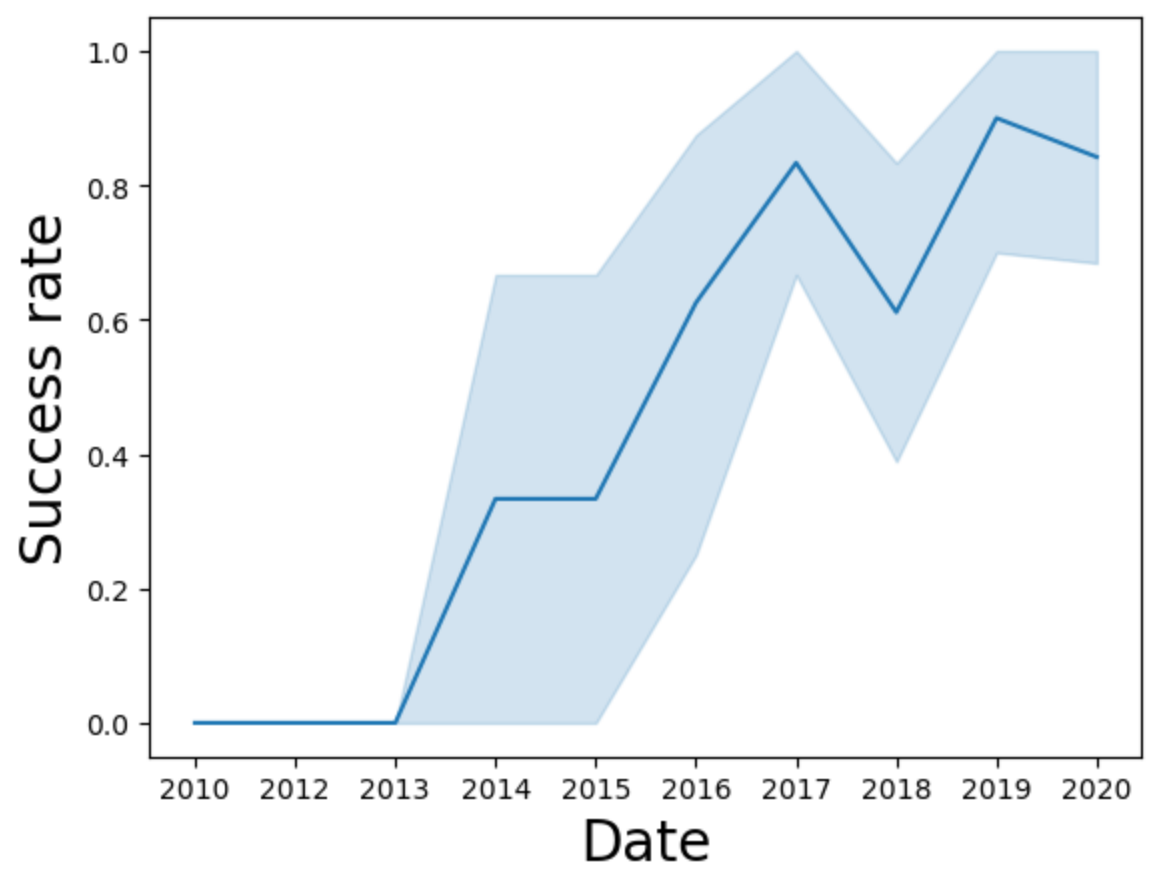
- VLEO is more successful orbit, also actively used last time
- The earliest flights all failed while the latest flights all succeeded

Payload vs. Orbit Type



- Heavy payloads have a negative effect on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch Success Yearly Trend



- The success rate since 2013 increases dramatically till 2020

All Launch Site Names

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

- Displaying the names of the unique launch sites

Launch Site Names 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

- Displaying of the launch sites names begin with 'CCA'

Total Payload Mass

Total_payload_mass
48213

- Displaying the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

Average_payload_mass
2534.6666666666665

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

First Successful Ground Landing Date

2015-12-22

- Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Listing the names of the boosters which have success in drone ship landing with payload between 4000 and 6000 kgs

Total Number of Successful and Failure Mission Outcomes

Failure (in flight)	1
Success	100

- Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

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

- Listing the names of the boosters which have carried the maximum payload mass

2015 Launch Records

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Listing the missions in year 2015

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

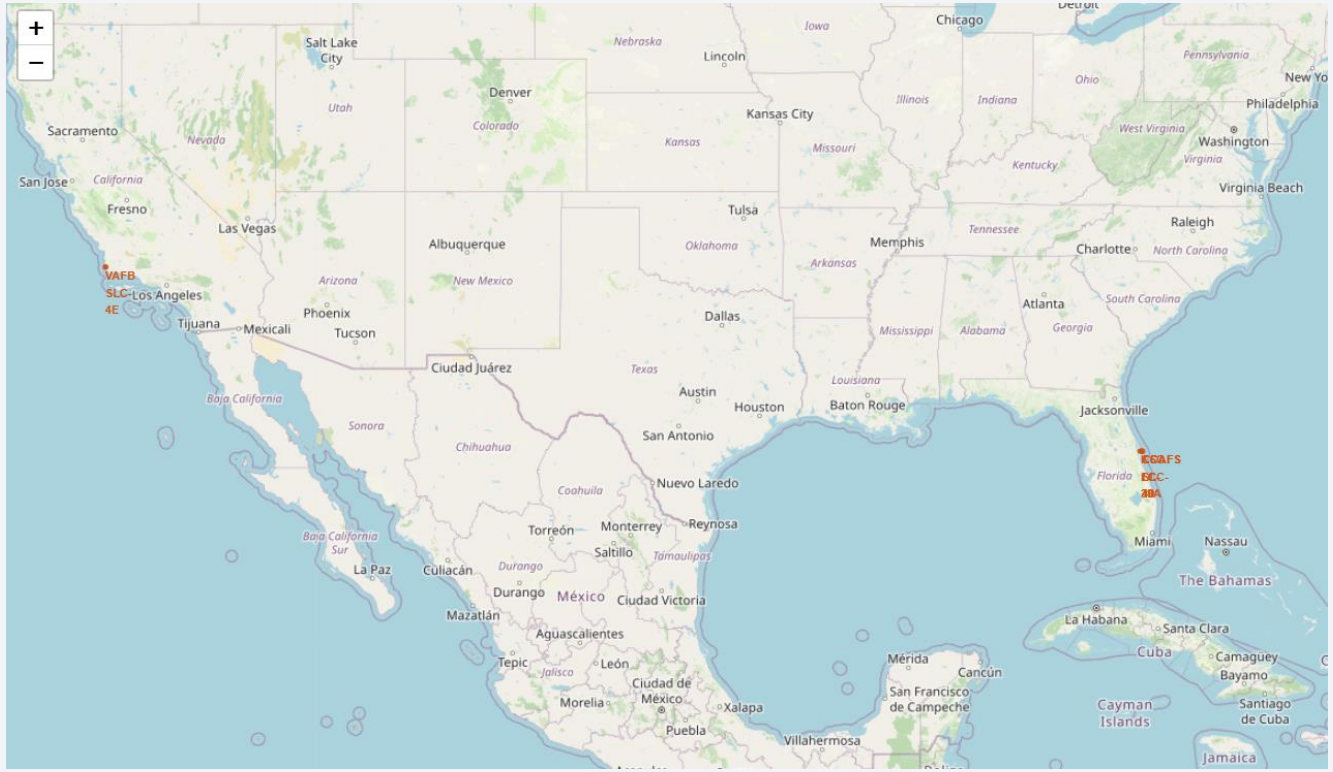
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

Launch sites on a map

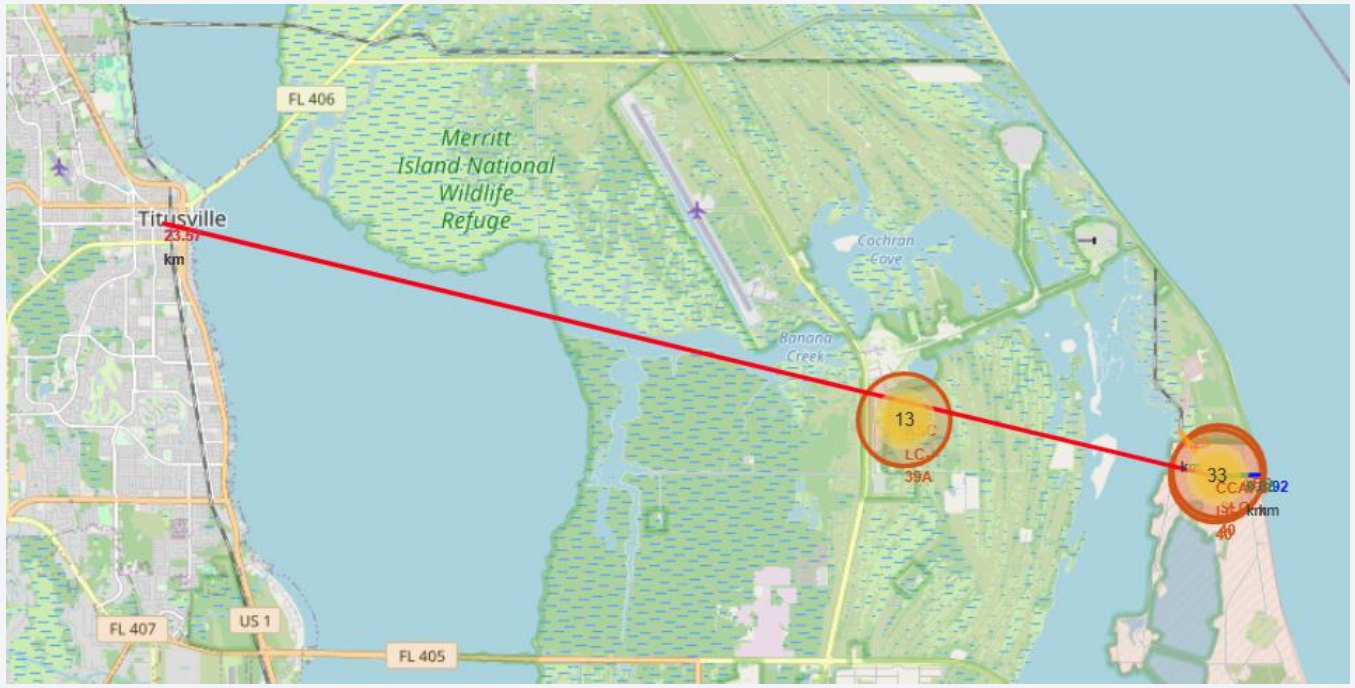


- Most of Launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast

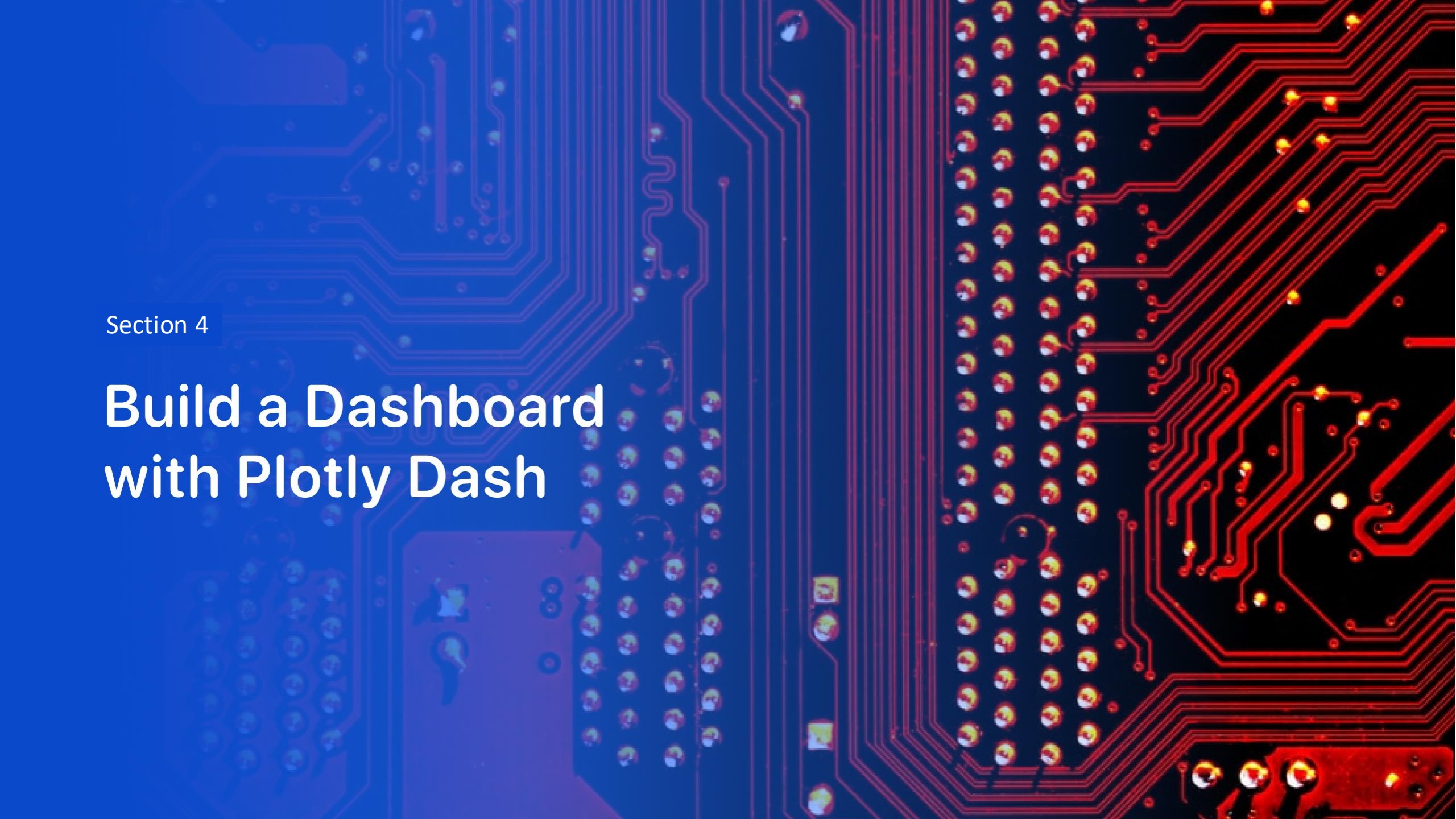
Success/failed launches for each site

- **Green Marker** = Successful Launch
- **Red Marker** = Failed Launch

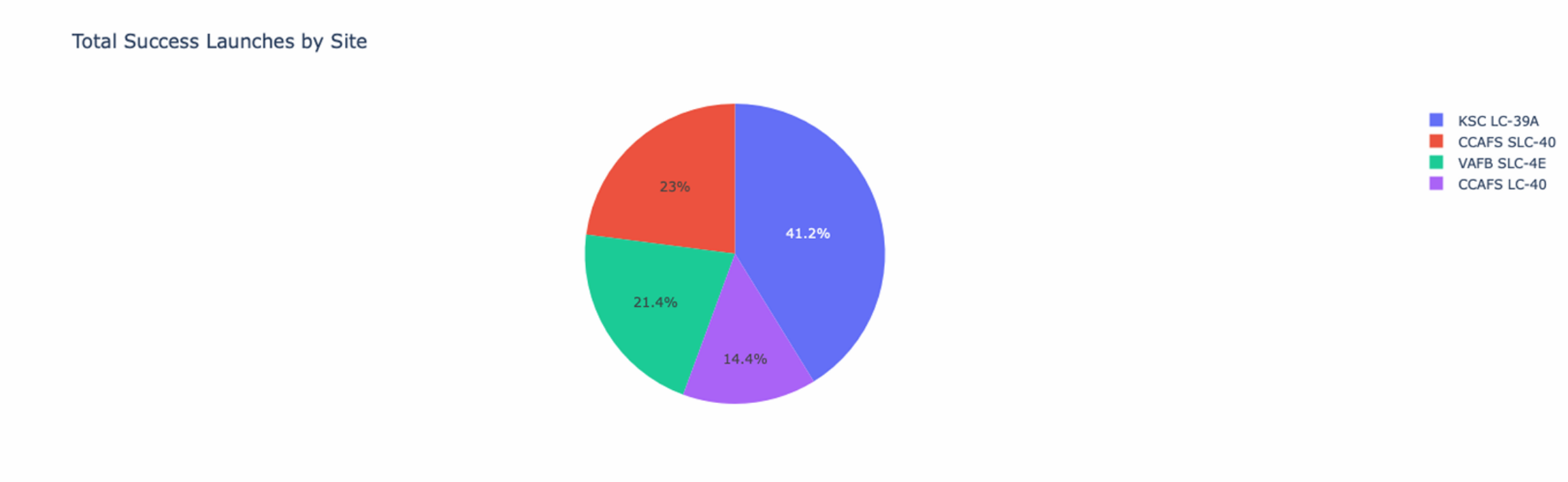
Distance from the launch site CCAF LC-40 to its proximities



- Close to coastline – 0,92 km
- Close to highway – 0,65 km
- Relatively close to city – 23,57 km

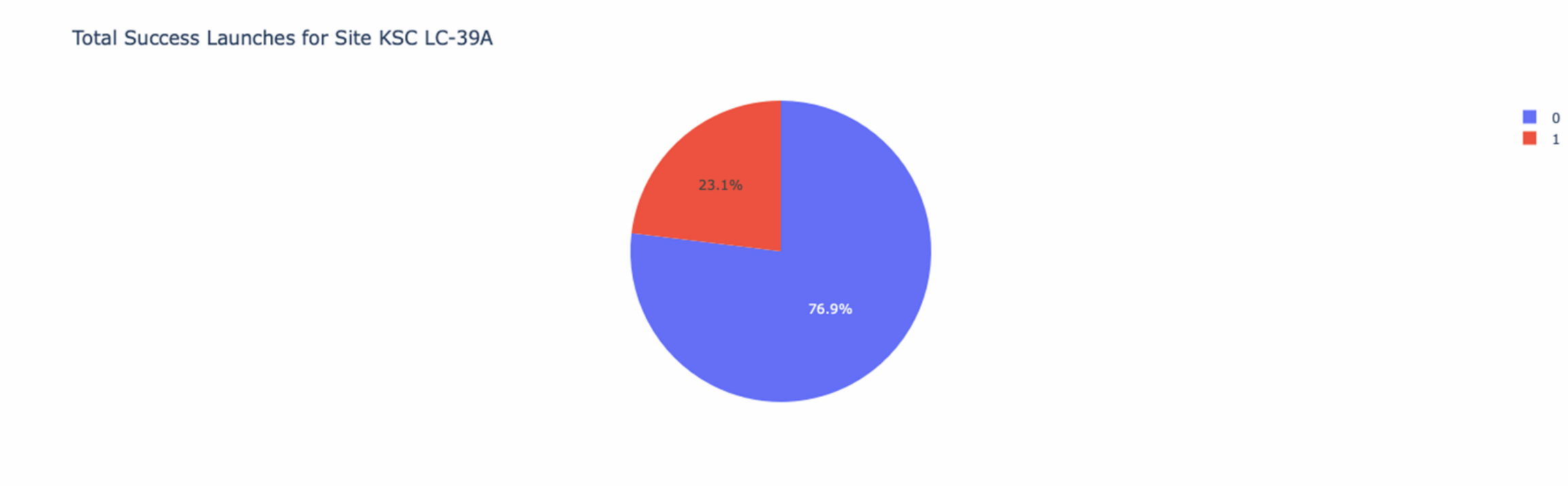

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



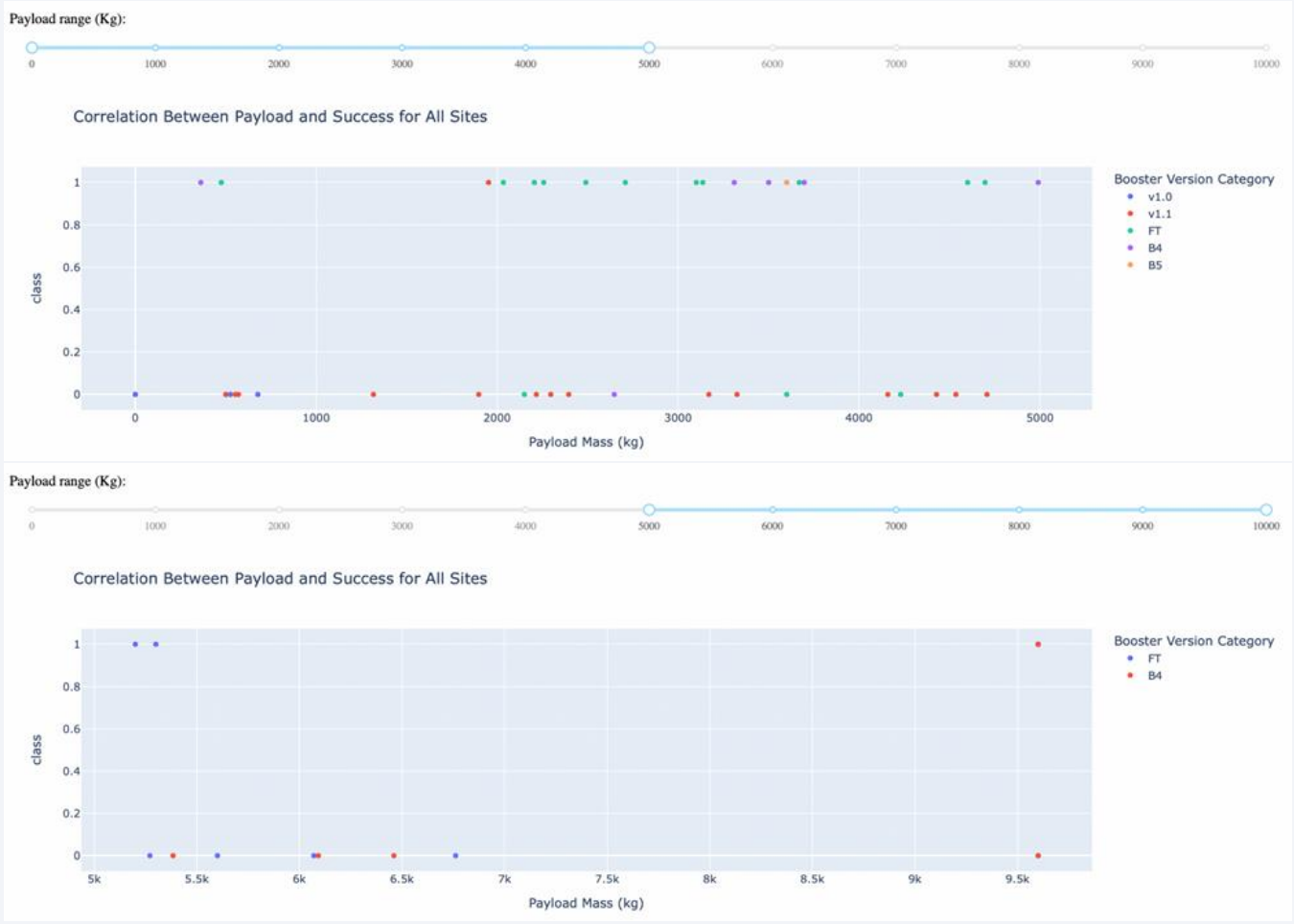
- KSC LC-39A has the most successful launches

Launch site with highest launch success ratio



- KSC LC-39A has the highest launch success rate

Payload Mass vs. Launch Outcome for all sites



- Payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores/Accuracy of the entire data set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

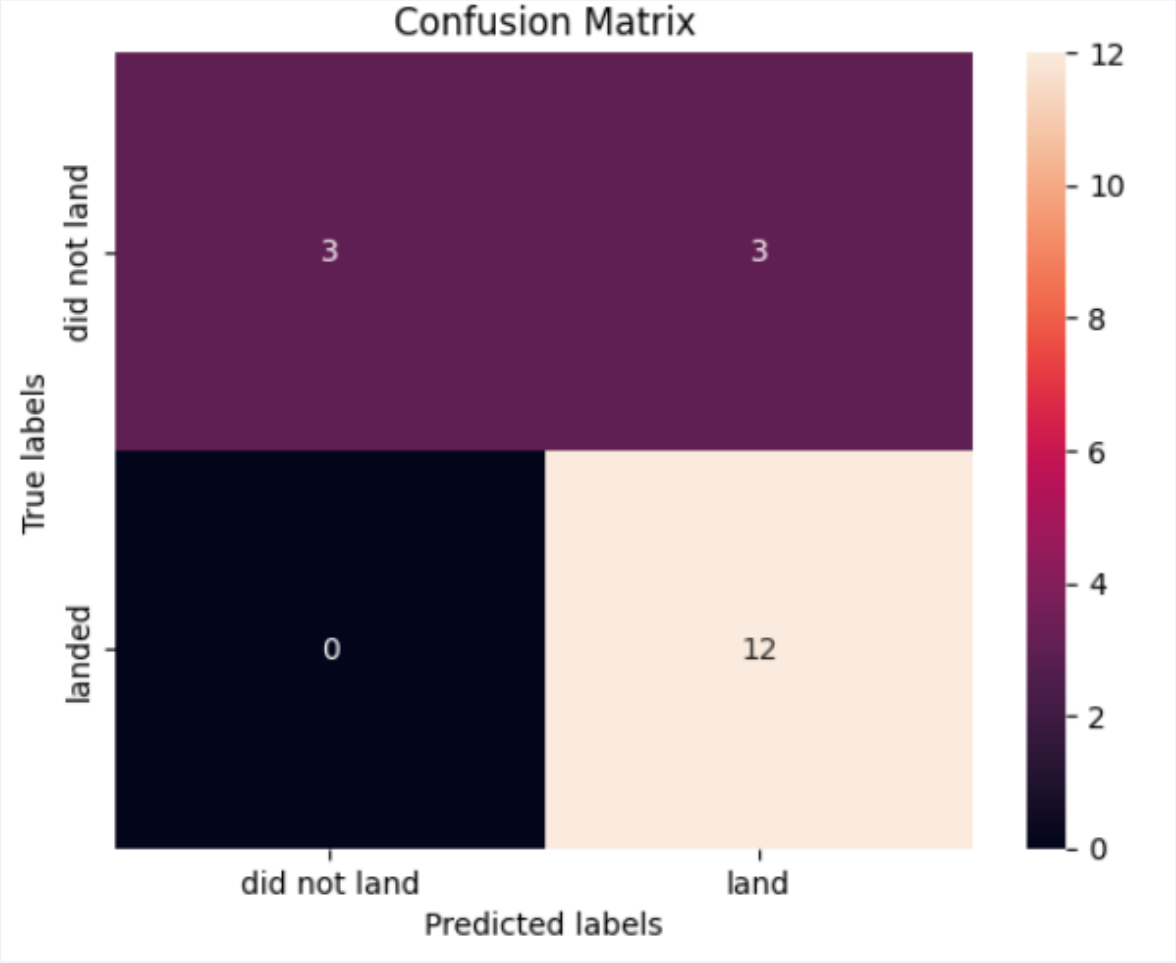
Score/Accuracy of the test data set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores/Accuracy of the test data set not shown which method performs best, may be due to the small test sample size,

Scores/Accuracy of the entire data set confirm that the best model is the Decision Tree Model

Confusion Matrix



- The main problem is false positives

Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass
- The success rate of launches increases over the years
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast
- KSC LC-39A has the highest success rate of the launches from all the sites
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Decision Tree Model is the best method for this dataset

Appendix

- GitHub
- Jupiter Lab
- IBM
- Coursera

Thank you!