IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bomin Kim
May 5, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using APIs and Web Scraping
  - Data Wangling Python Libraries
  - Exploratory Data Analysis (EDA) using SQL and Data Visualization
  - Interactive Visual Analytics with Folium and Dashboard with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- As data scientists working for SpaceY that would like to compete with SpaceX, we have decided to analyze their rocket launches using the data provided. The key differentiation of SpaceX is that they reduce the total price of launches by reusing the first stage of the launch.
- With the information we have regarding the rocket launches, can we determine if SpaceX will reuse the first stage of the launch? In other words, can we predict the success of the launch and determine the key factors for the successful launch?

Section 1

Methodology

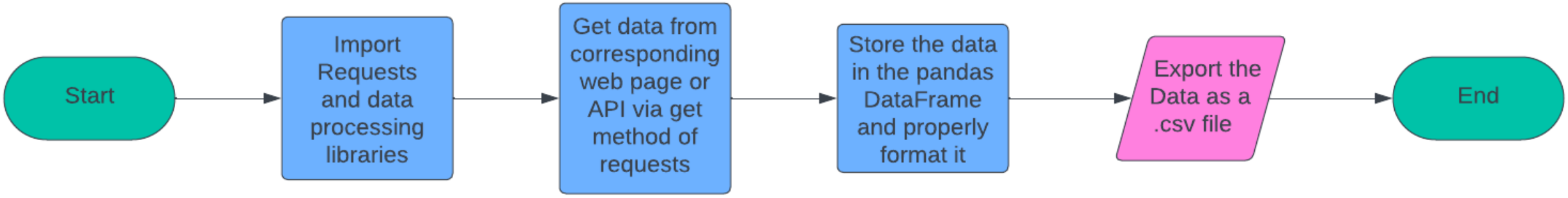
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected with Data Collection API and Web Scraping
- Perform data wrangling
  - Checked the value counts of significant features and converted the landing outcome labels into numerical values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - With the significant features, performed data pre-processing and fit the data to different classification models with different hyperparameters to find the best model. Evaluation was done by confusion matrix.

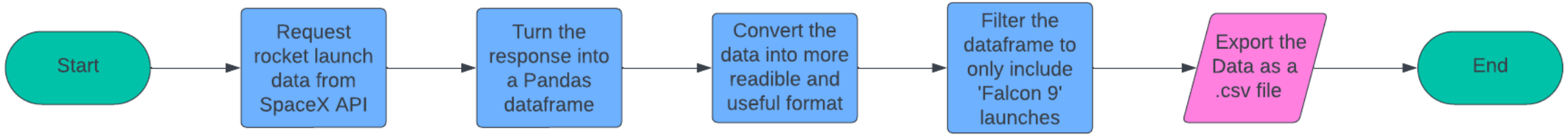
Data Collection

- Data is collected through the request to the SpaceX API.
- Data Wrangling and Formatting was partially done as well.



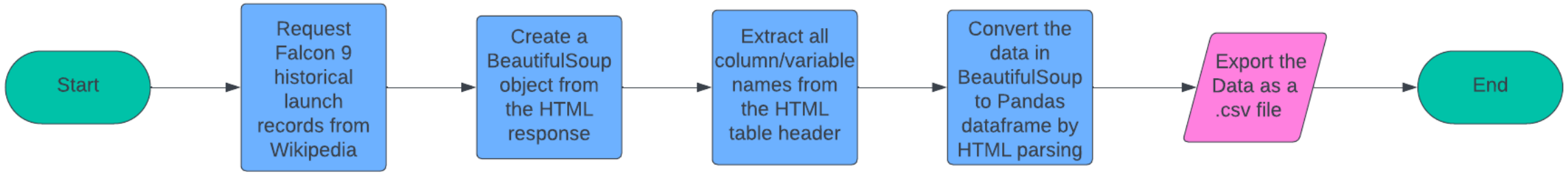
Data Collection – SpaceX API

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



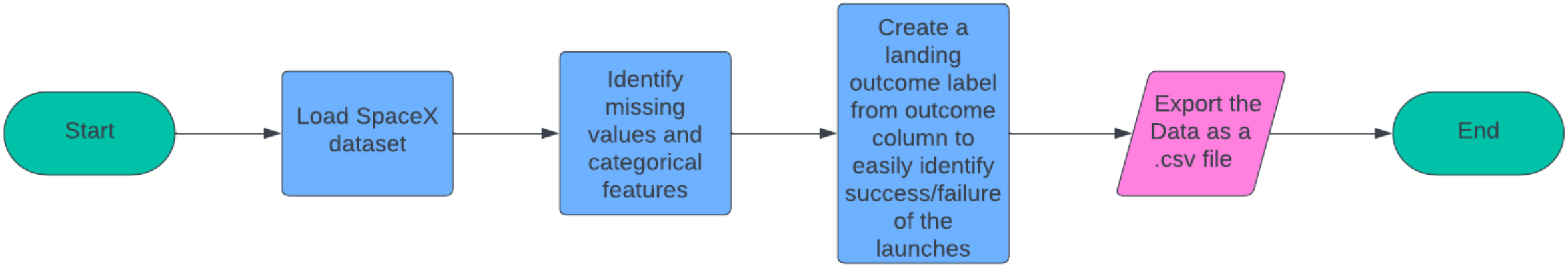
Data Collection - Scraping

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/edadataviz.ipynb
- Plotted charts:
  - FlightNumber vs. LaunchSite (scatter plot): Explore the relation between flight numbers, launch sites, and success
  - Payload vs. LaunchSite (scatter plot): Explore the relation between payload, launch sites, and success
  - Orbit vs. Success Rate (bar chart): Explore the relation between the orbit and success rates
  - FlightNumber vs. Orbit (scatter plot): Explore the relation between flight numbers, orbit, and success
  - Payload vs. Orbit (scatter plot): Explore the relation between payload, orbit, and success
  - Launch Success Yearly Trend (line chart): Explore the relation between years and success rates

EDA with SQL

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- SQL Queries:
  - Display the names of the unique launch sites in the space mission
    SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
  - Display 5 records where launch sites begin with the string 'CCA'
    SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
  - Display the total payload mass carried by boosters launched by NASA (CRS)
    SELECT sum("PAYLOAD_MASS__KG_") AS "Total_Payload" FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)"
  - Display average payload mass carried by booster version F9 v1.1
    SELECT avg("PAYLOAD_MASS__KG_") AS "Average_Payload" FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"

EDA with SQL

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- SQL Queries:
  - List the date when the first successful landing outcome in ground pad was achieved
    SELECT min("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)"
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)
  - List the total number of successful and failure mission outcomes
    SELECT "fcount" AS failures, "scount" AS successes FROM (SELECT count("Mission_Outcome") AS "fcount" FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "Failure%") AS f, (SELECT count("Mission_Outcome") AS "scount" FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "Success%") AS s

EDA with SQL

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- SQL Queries:
  - List the names of the booster_versions which have carried the maximum payload mass
    SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
  - List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for months in year 2015
    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)"
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
    SELECT "Landing_Outcome", count("Landing_Outcome") AS "Count" FROM SPACEXTABLE WHERE "Date" BETWEEN "2010-06-04" AND "2027-03-20" GROUP BY "Landing_Outcome" ORDER BY count("Landing_Outcome")DESC

Build an Interactive Map with Folium

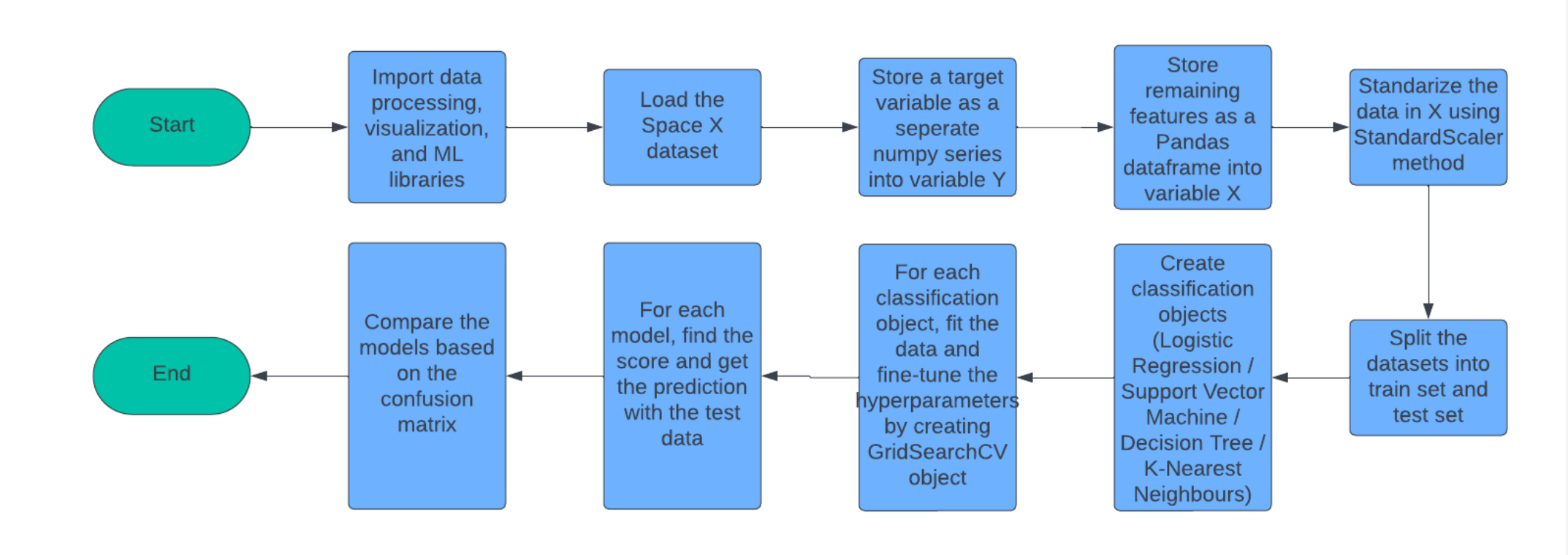
- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/lab_jupyter_launch_site_location.ipynb
- Map Objects:
  - Launch Site Info:
    - Markers: indicate each site on the map, display the names of the launch sites, and show the number of launches
    - Circles: indicate the approximate area of the launch sites
  - Launch Results:
    - Markers: indicate whether the launch succeeded or failed on each launch site
  - Launch Site Proximities:
    - Lines: visualize the distance from each launch site to the close railway, highway, coastline, and city

Build a Dashboard with Plotly Dash

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/spacex_dash_app.py
- Dropdown: lets a user to choose a targeted launch site or all launch sites to view its dashboard
- Pie Chart: displays a success count of each launch site if all launch sites option is chosen, and a success count and a failure count if one launch site option is chosen
- Range Slider: lets a user set the range of interest of a payload mass
- Scatter Plot: displays a relation between success and payload mass

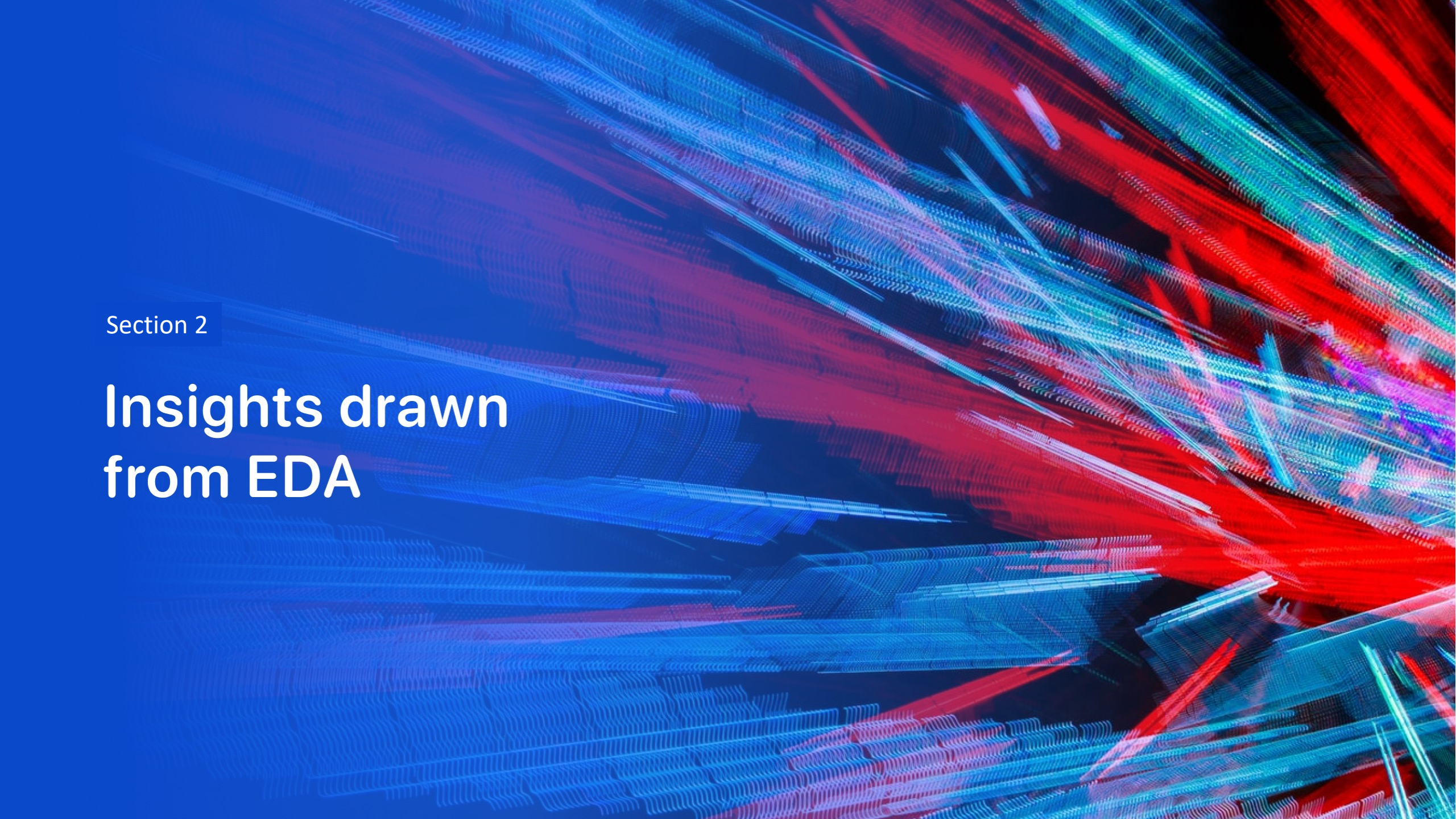
Predictive Analysis (Classification)

- https://github.com/Bomin75/IBM-Data-Science-Professional-Certificate/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



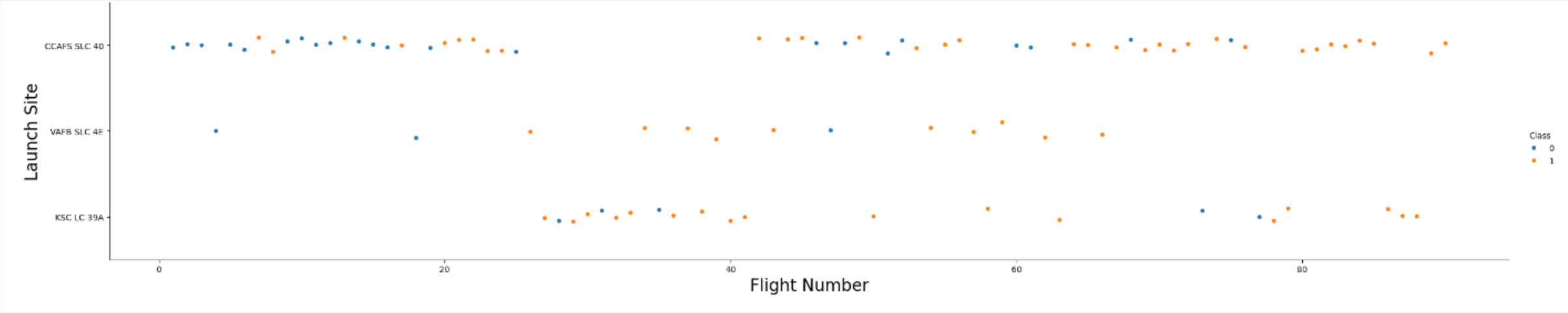
Results

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


Section 2
Insights drawn from EDA

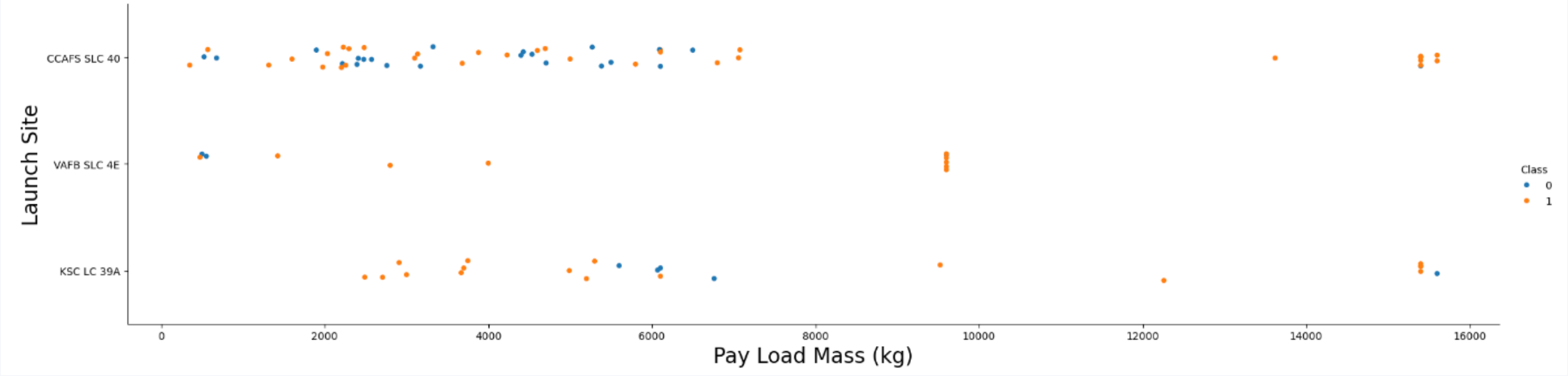
Flight Number vs. Launch Site

- Observation: as flight number increases, there are more successful launch cases from launch site CCAFS SLC 40



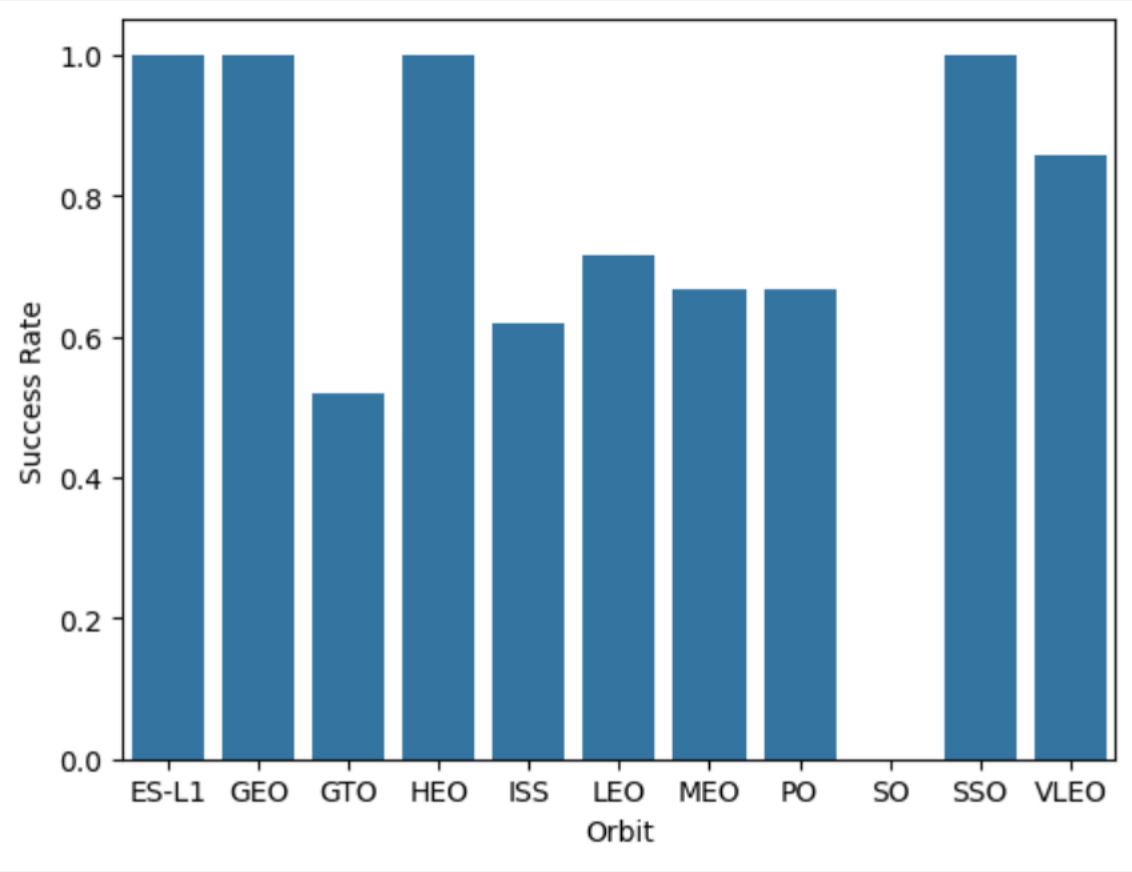
Payload vs. Launch Site

- Observation: the VAFB-SLC launch site has no rockets launched for heavy-payload-mass (greater than 10000)



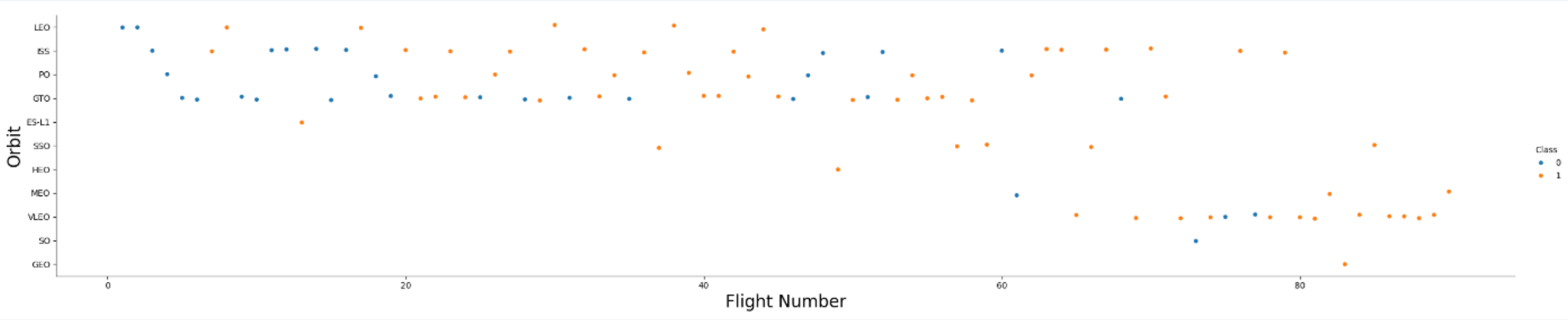
Success Rate vs. Orbit Type

- Observation: orbits ES-L1, GEO, HEO, and SSO have the highest success rate followed by orbit VLEO



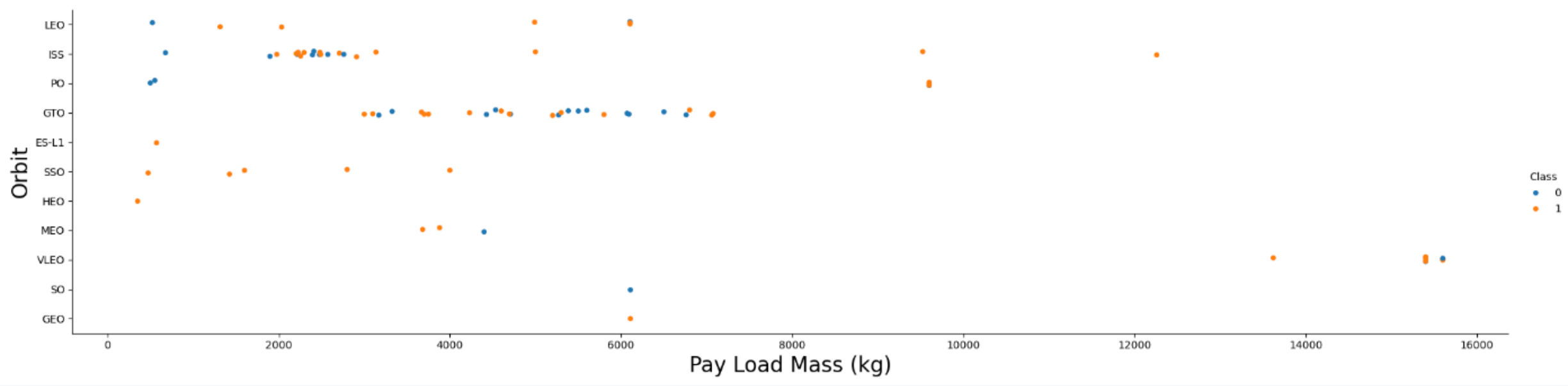
Flight Number vs. Orbit Type

- Observation: the success of the launches in the LEO orbit is related to the flight number while the success in the GTO orbit has no significant relationship with the flight number



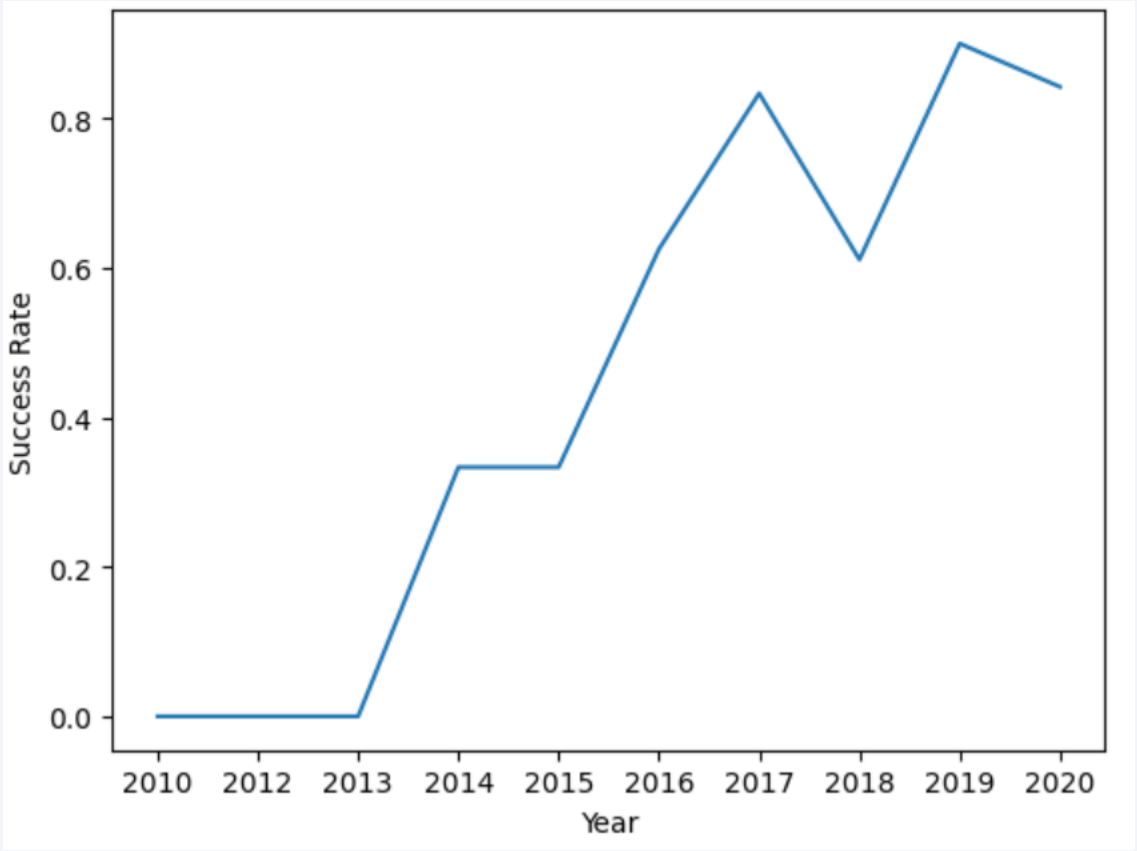
Payload vs. Orbit Type

- Observation: the heavier the payload mass is, the greater the success rate becomes for the orbits Polar, LEO, and ISS



Launch Success Yearly Trend

- Observation: the success rate tended to grow since 2013



All Launch Site Names

- Find the names of the unique launch sites

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total_Payload
45596

Average Payload Mass by F9 v1.1

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

Average_Payload
2928.4

First Successful Ground Landing Date

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

min("Date")
2015-12-22

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

failures	successes
1	100

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

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

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	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

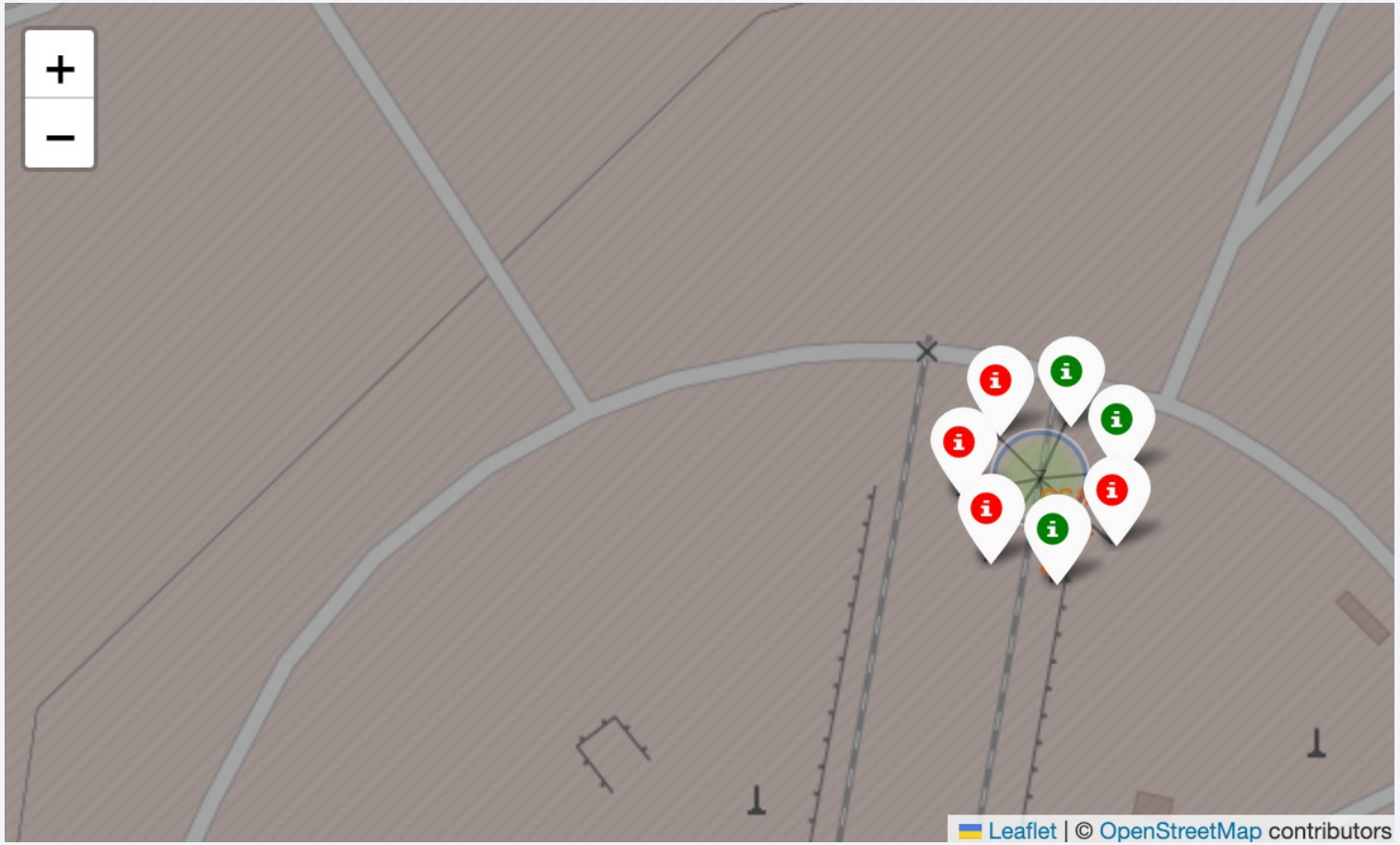

Section 3
Launch Sites
Proximities Analysis

Launch Site Info

- Observation: the launch sites are close to the coastline

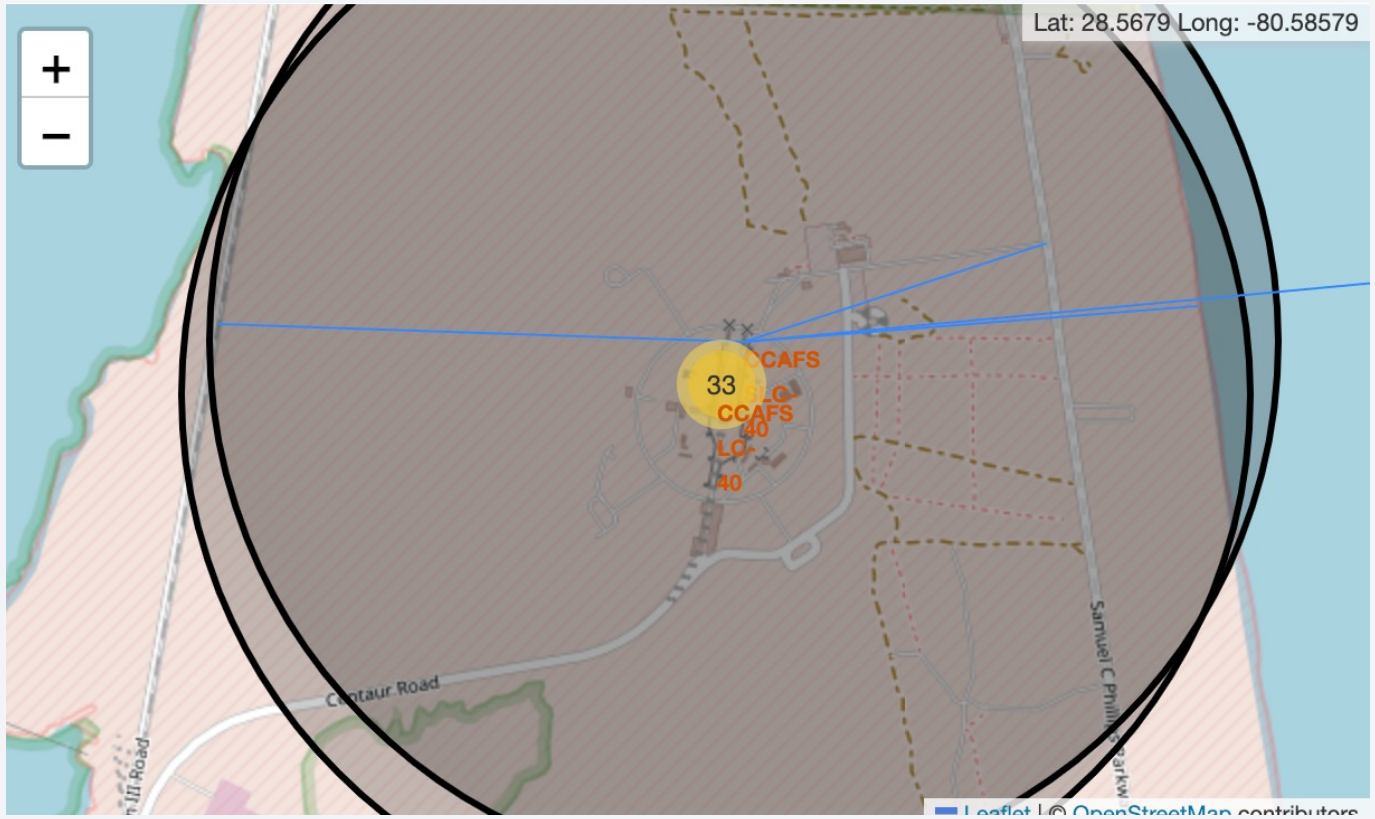
Launch Results

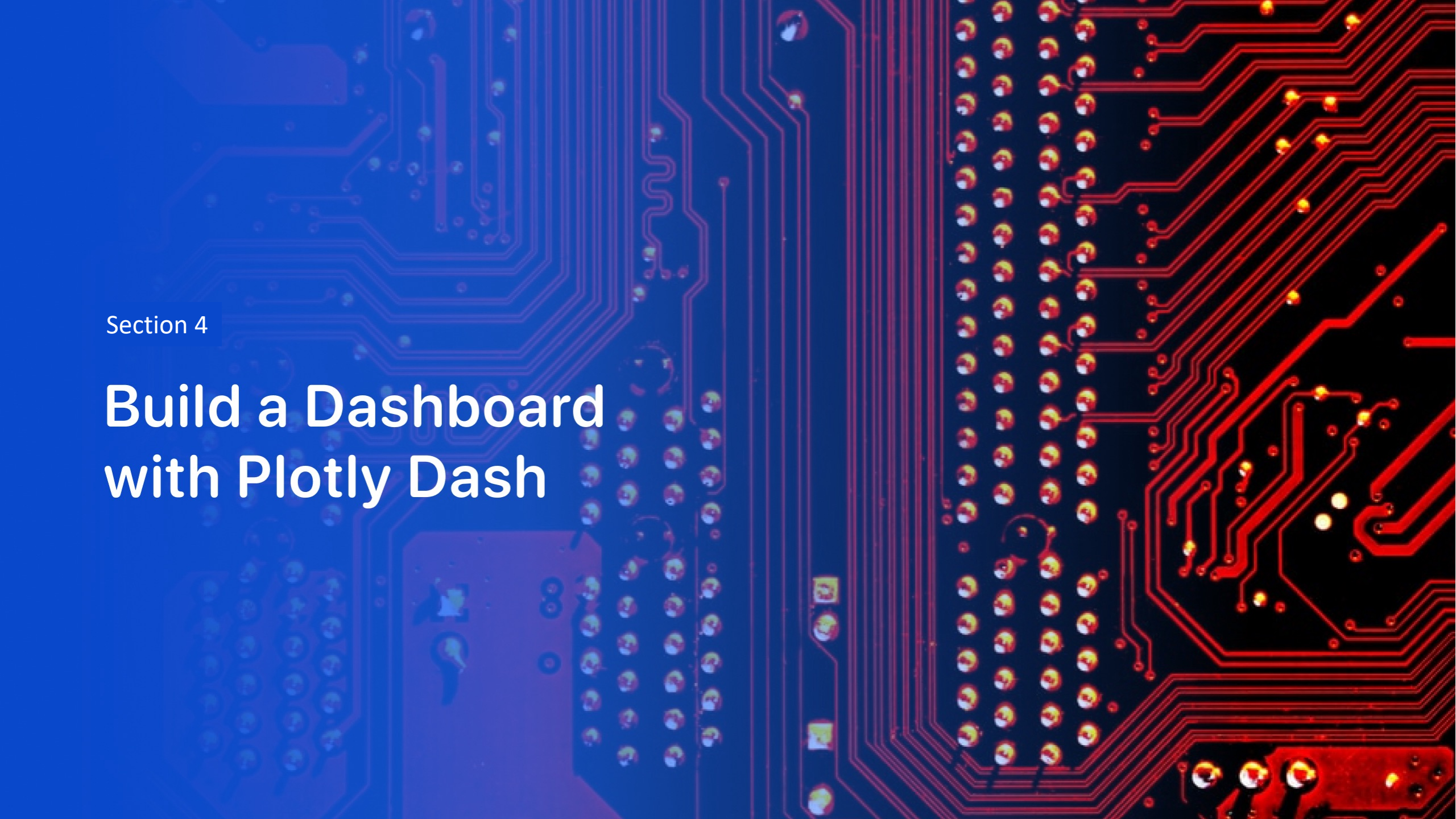
- Observation: launch site KSC LC-39A has the highest success rate



Launch Site Proximities (CCAFS SLC-40)

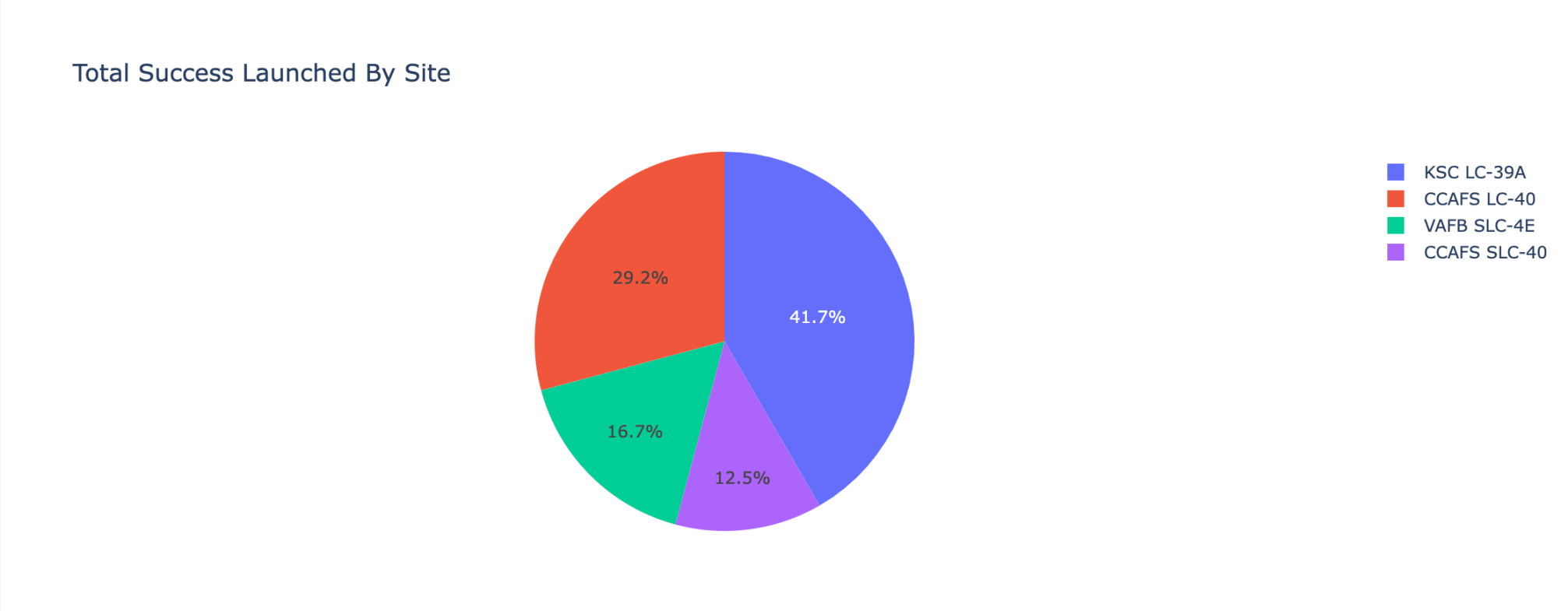
- Observation: launch site CCAFS SLC-40 is close to railways, highways, and coastline




Section 4
Build a Dashboard with Plotly Dash

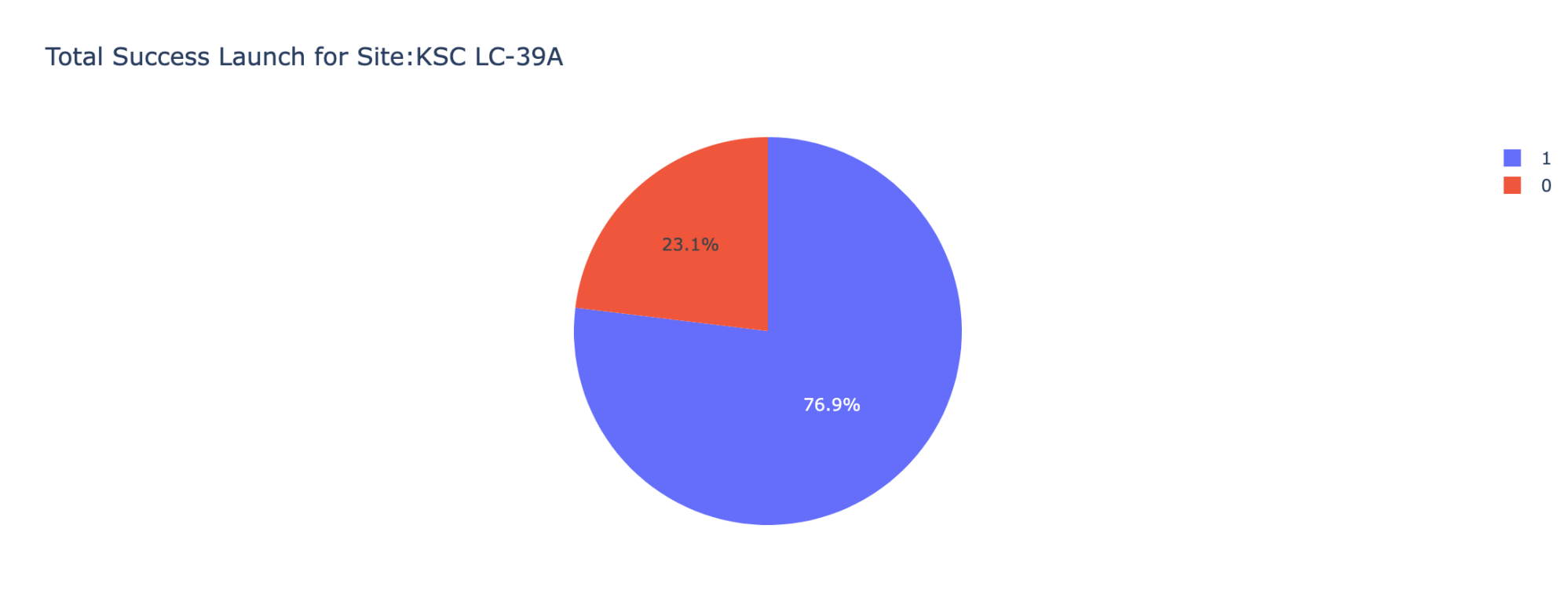
Total Success Launched By Site

- Observation: site KSC LC-39A has the highest success count



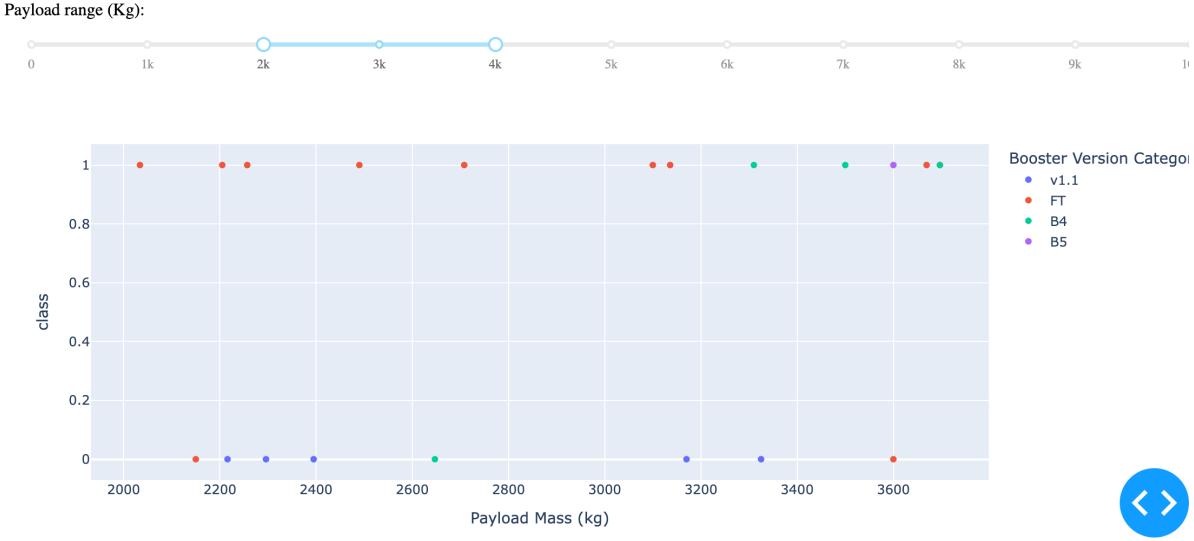
Total Success Launch for Site KSC LC-39A

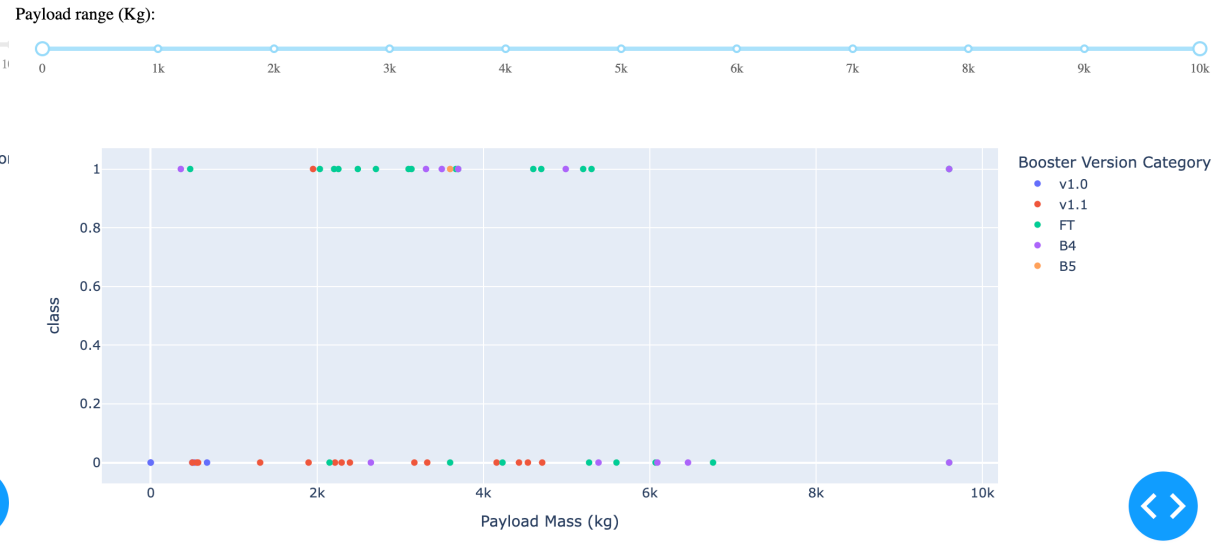
- Observation: KSC LC-39A has a success rate of 76.9%



Payload vs. Launch Outcome for all sites by Ranges

- Observation: payload range from 2k to 4k and the booster version FT have the highest success rate



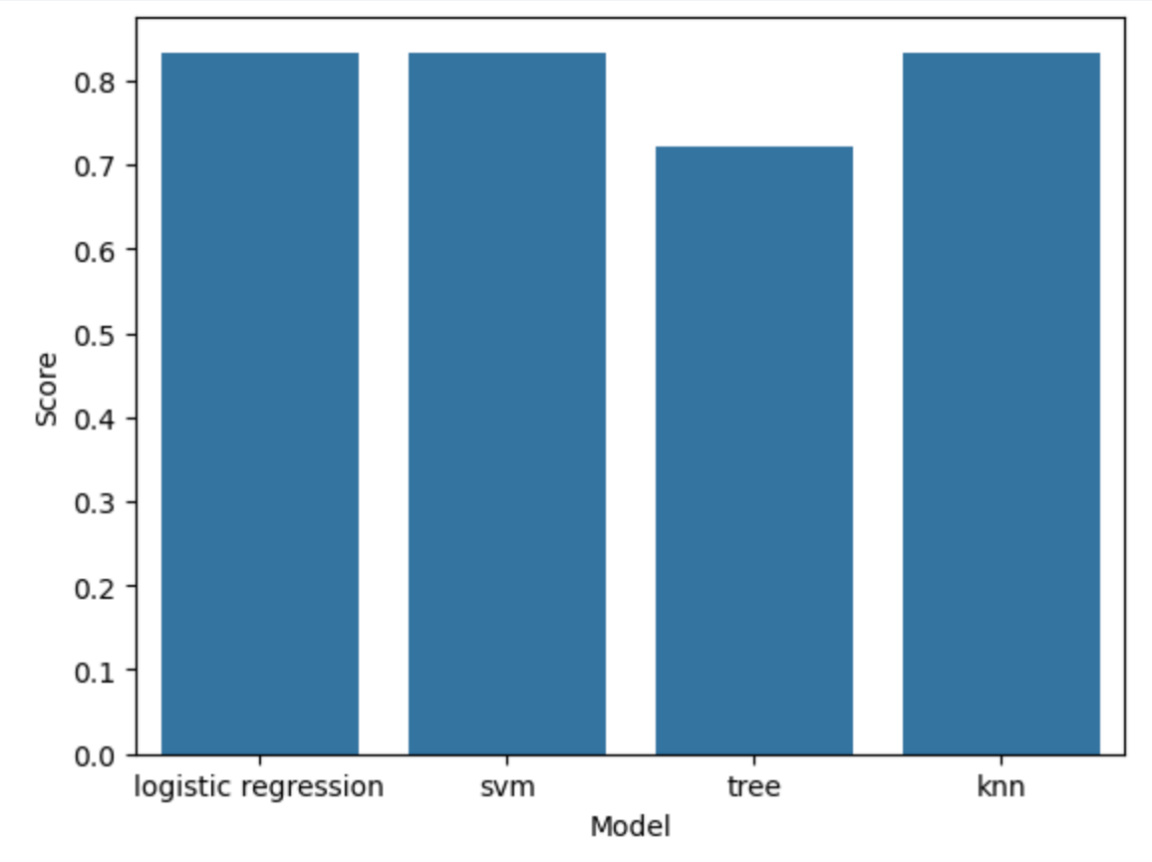


Section 5

Predictive Analysis (Classification)

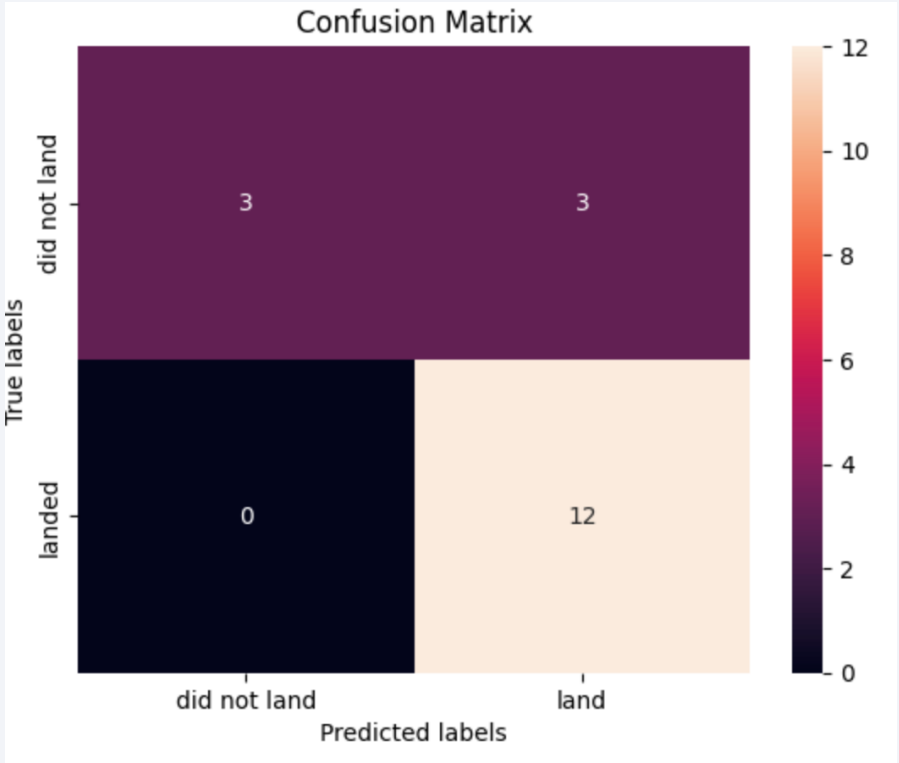
Classification Accuracy

- Logistic Regression, SVM, and KNN have the same accuracy on the test data and are the highest.



Confusion Matrix

- Logistic regression, SVM, and KNN are the models with the highest accuracy and have the same confusion matrix
- The major problem is false positives



Conclusions

- Overall, as flight number increases, the number of successful launch cases also rises
- The launch site with the highest success rate and the highest success count is KSC LC-39A
- The booster version with the highest success rate is FT, on the range of 2k to 4k, in particular
- The orbits ES-L1, GEO, HEO, and SSO have the highest success rate followed by the orbit VLEO
- With the provided data from SpaceX, the success of the launch can be predicted with the accuracy of ~83.34%

Thank you!