IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Bilal
24th November 2021

Outline

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

Executive Summary

- The project consisted of multiple steps where the data collection process started the pipeline for building a sustainable machine learning model to predict the launch outcome of a SpaceX Rocket.
- SpaceX API was used to gather most of data, however some of the data was collected using web scrapping to complete the dataset.
- Once data was collection was finished data wrangling was done to shape data into the best form for the machine learning algorithm.
- EDA was then done using both SQL and Visualization to gather a complete structure of the data.
- Once the data relationships were identified a machine learning model was trained on the data.

Introduction

Project background and context

SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

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?

Methodology

Executive Summary

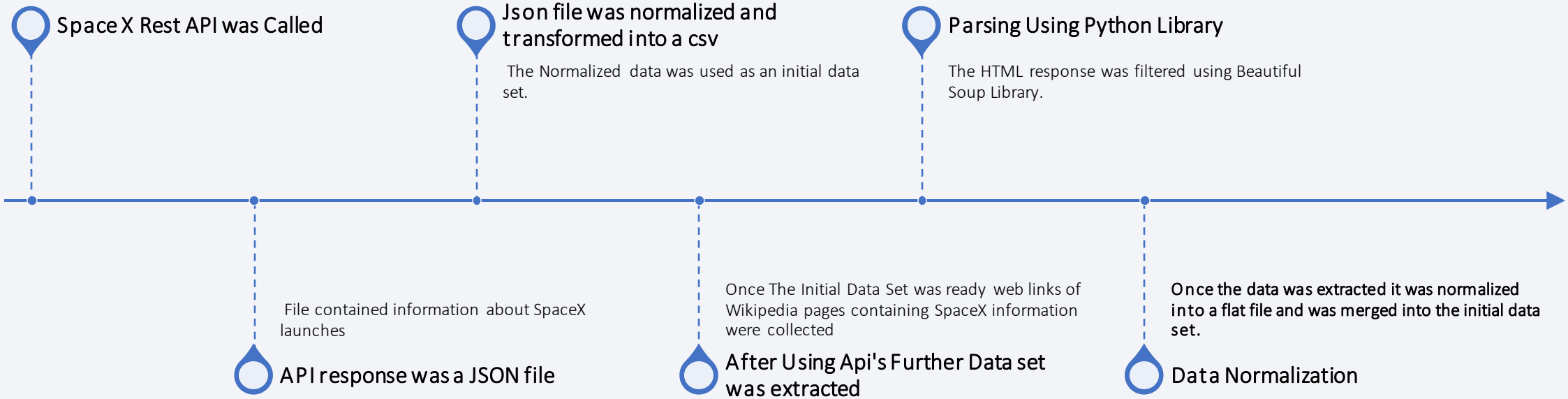
- Data collection methodology:
  - Data was collected using SpaceX Public API's however to complete the dataset some of the data was collected through web scraping of Wikipedia pages.
- Perform data wrangling
  - Data was Filtered and missing values were dealt with.
  - Binary Classification of Landing outcomes were one hot encoded.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Multiple predictive classification algorithms were used to ensure that the selected machine learning model contains the best possible results.

Section 1

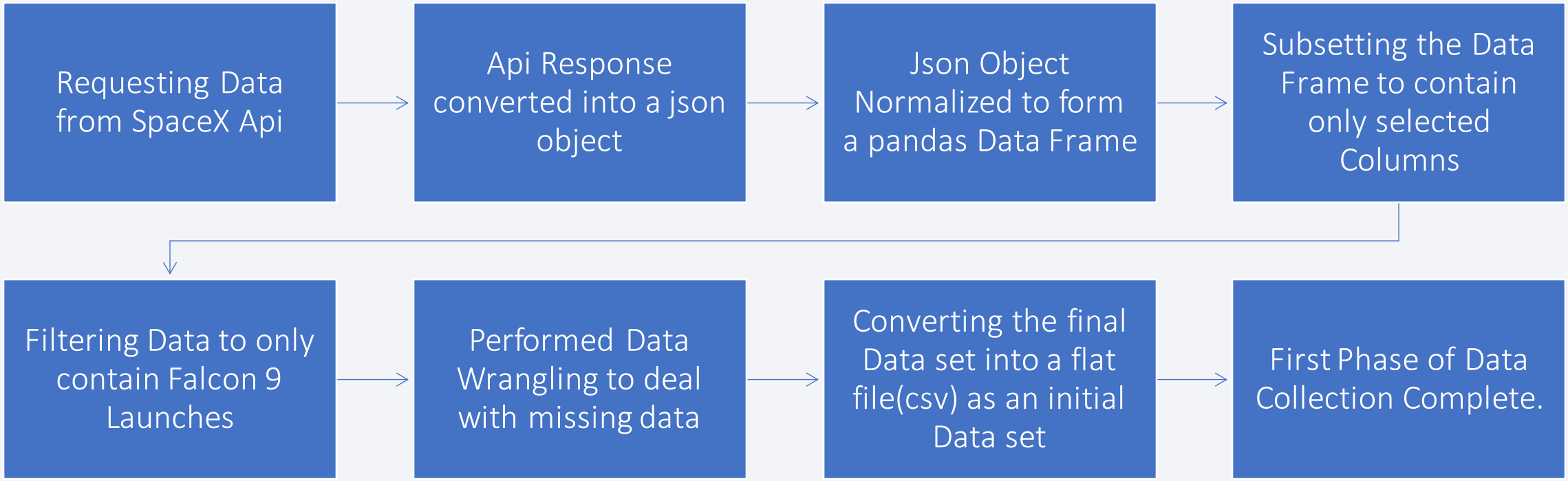
Methodology

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 page.
- Both data collection methods were done to ensure that complete information about the launches are present in our dataset for a detailed analysis.

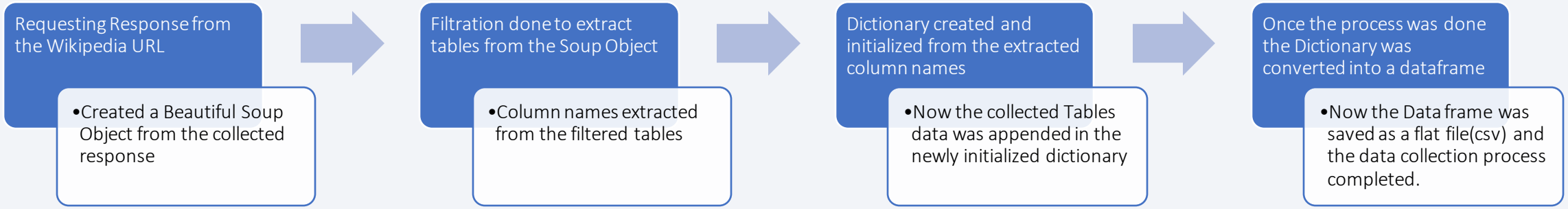


Data Collection – SpaceX API



GitHub URL: Data Collection API Lab

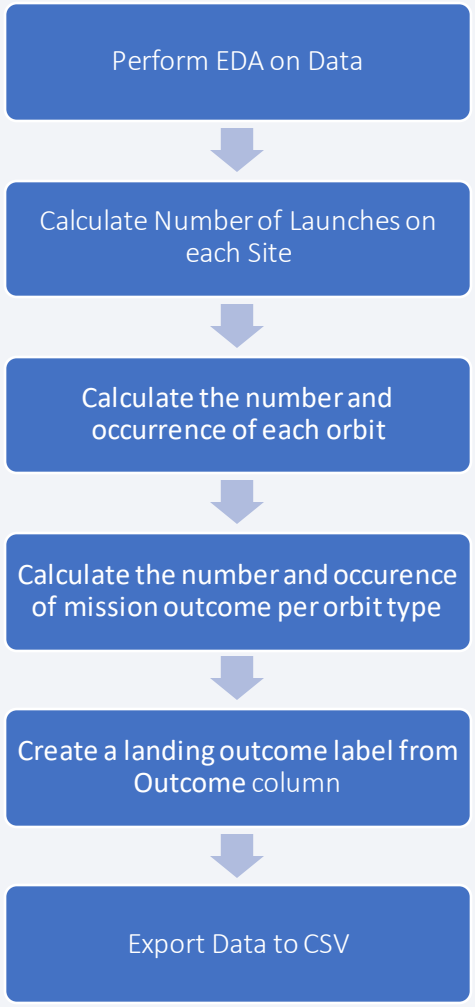
Data Collection - Scraping



GitHub URL: Data Collection Web Scrapping Lab

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully.
- Sometimes a landing was attempted but failed due to an accident. For example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship.
- This step mainly converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



GitHub URL: Data Wrangling Lab

EDA with Data Visualization

- Categorical Scatter Plots determine how much one variable is affected by another and were used to visualize relationship b/w multiple variables and landing outcomes
  - Flight Number vs. Payload Mass, Launch Site and Orbit Type
  - Payload Mass vs. Launch Site and Orbit Type
- A bar chart compares two groups of data and was plotted to visualize relationship between success rate of each orbit type.
- A line plot can be used to see trends/sequence and was used to visualize success rate yearly trends.

GitHub URL: EDA with Visualization Lab

EDA with SQL

- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string ‘CCA'.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster versions which have carried the maximum payload mass.
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.
- Ranking 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.

GitHub URL: EDA with SQL Lab

Build an Interactive Map with Folium

- Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center were plotted using its coordinates.
- Markers with Circle, Popup Label and Text Label of all Launch Sites were plotted using their coordinates to show their geographical locations and proximity to Equator and coasts.
- Each Launch Site had colored markers to identify successful and failed launches.
- Straight lines were plotted to mark distance between the launch site and its proximities for example railway, highway, closest city and coastline.

GitHub URL: Data Visualization with Folium

Build a Dashboard with Plotly Dash

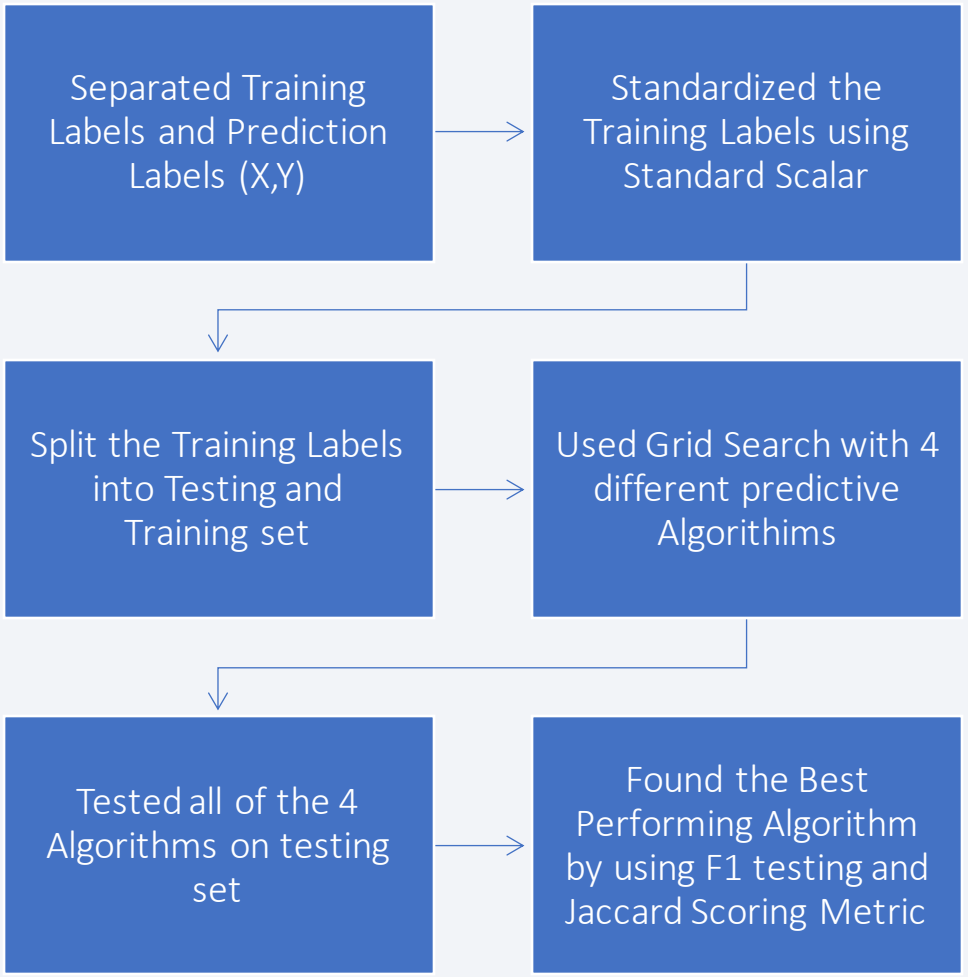
- A dropdown containing all launch sites was added.
- A pie chart was plotted which plotted the ratio of successful and unsuccessful launches of the selected site from the above-mentioned dropdown.
- A slider was introduced to select the Payload Range of the launches
- A scatter plot was added to visualize the relation between the Payload and Launch Outcome.

GitHub URL: Interactive Dashboard with Plotly Dash

Predictive Analysis (Classification)

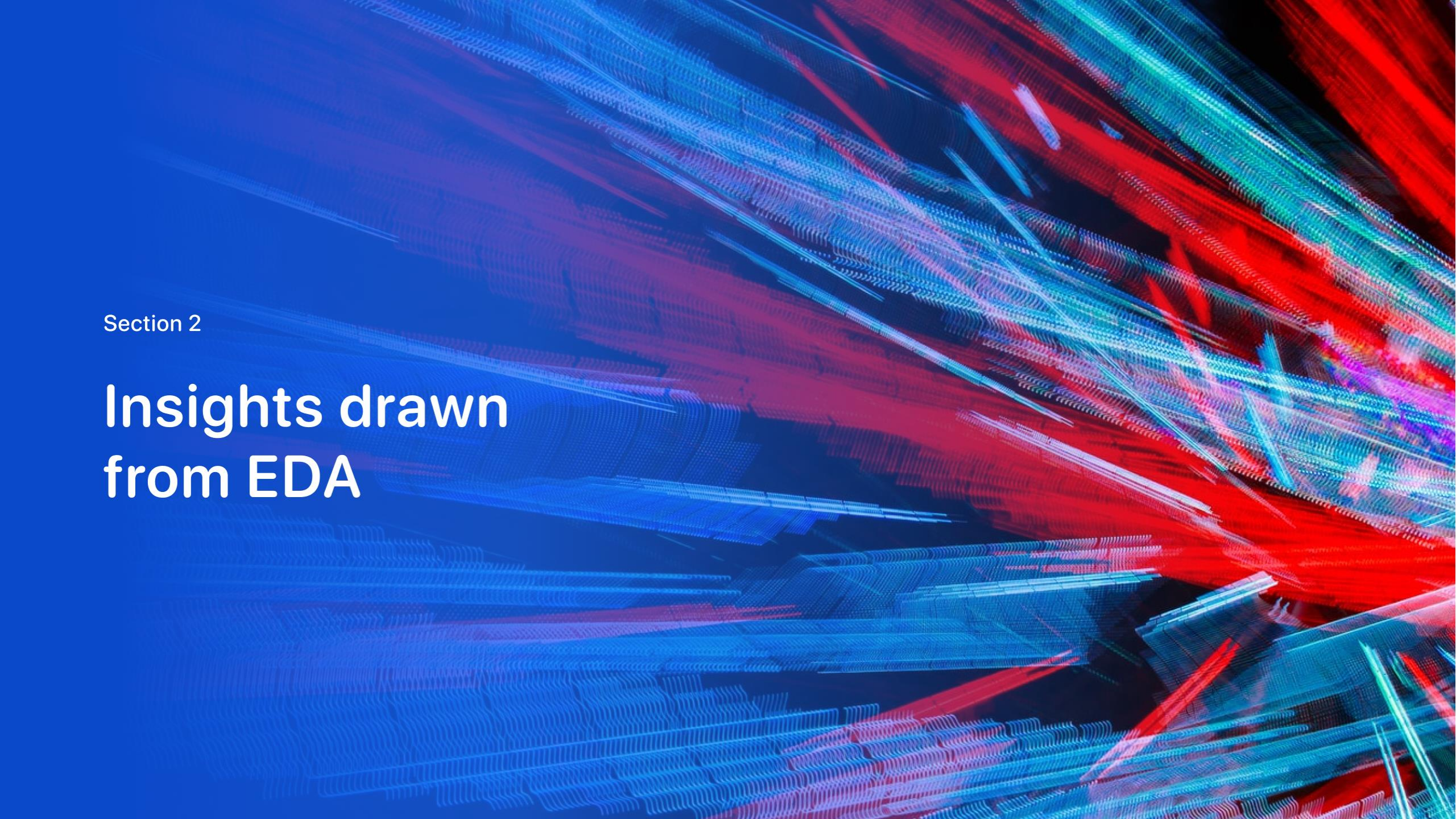
- Separated the Data
- Standardized the Data
- Split for training and testing
- Used GridSearchCV for finding Optimal Parameters
- Tested accuracy of the best model using Testing Data
- 4 different type of models were trained:
  - Logistic Regression Model
  - Support Vector Machine Model
  - Decision Tree Model
  - K-Neighbors Model

GitHub URL: Prediction using Machine Learning



Results

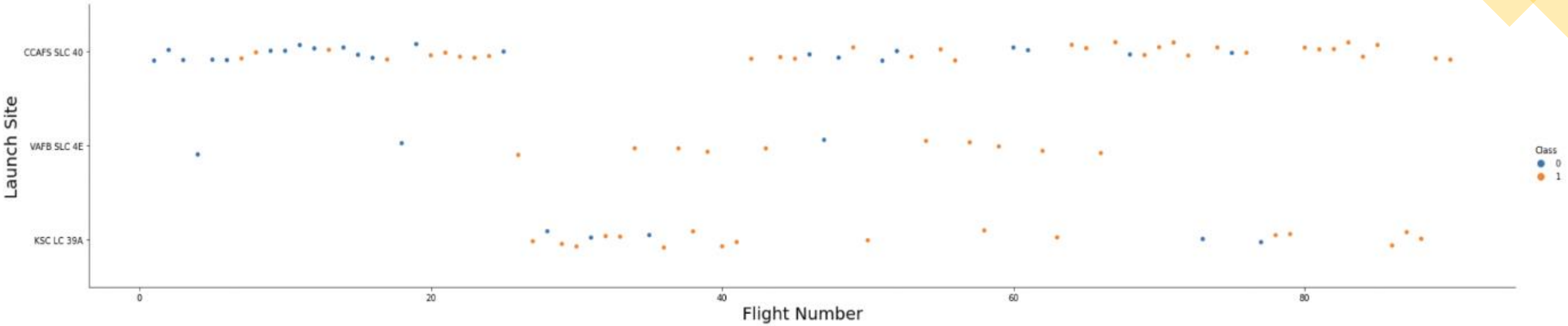
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

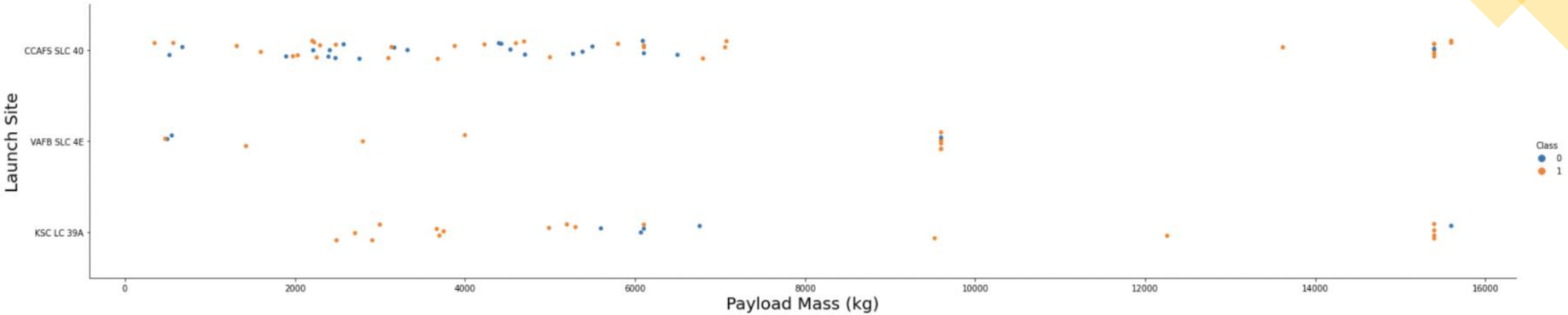
Insights drawn from EDA

Flight Number vs. Launch Site



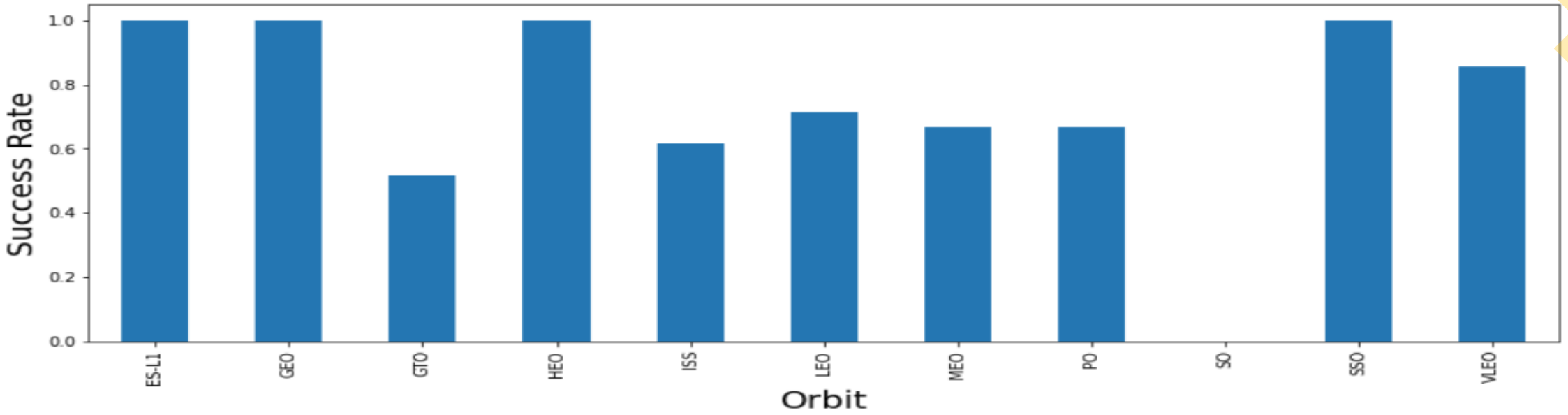
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that rate of success increases with the number of flights.

Payload vs. Launch Site



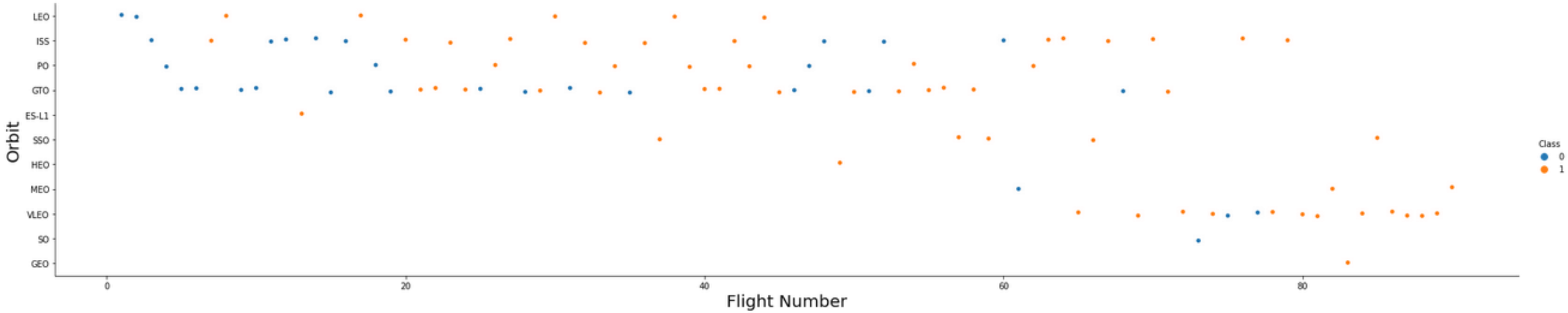
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type



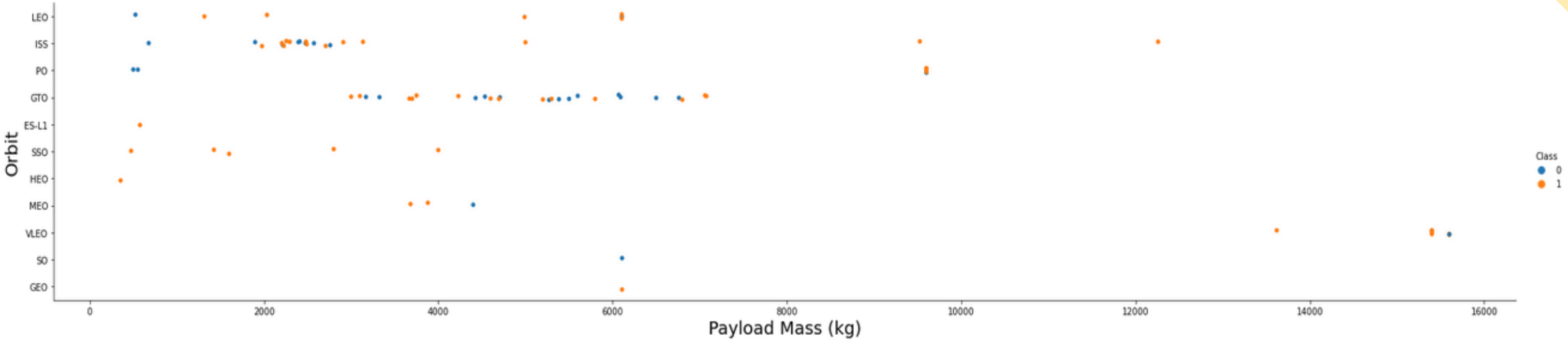
- ES-L1, GEO, HEO, SSO are orbits with 100% success rate.
- SO is the only orbit with 0% success rate.
- GTO, ISS, LEO, MEO, PO are orbits with success rate between 50% and 85%.

Flight Number vs. Orbit Type



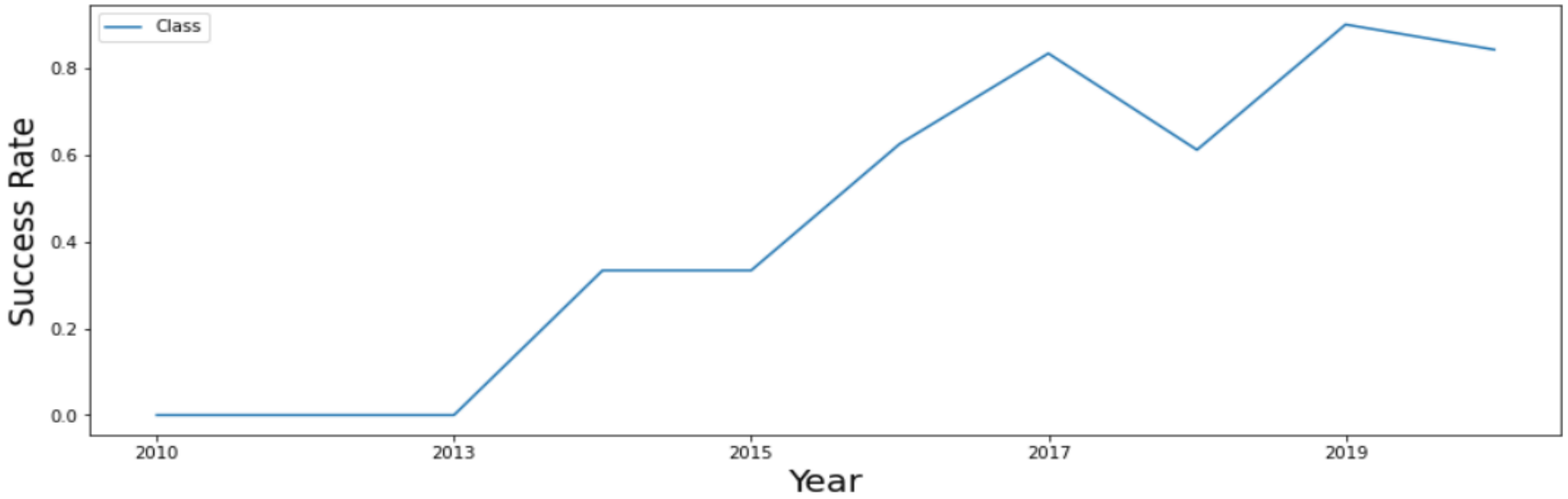
- This confirms that there is no linear relationship between number of flights and success ratio.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



- Success rate has been increasing from 2013 till 2019 (excluding 2018).
- From 2010 to 2013 the rate of success was 0%.

All Launch Site Names

- The names of the unique launch sites used by Falcon 9.

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

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- 5 records where launch sites begin with `CCA`

Total Payload Mass

- Calculated the total payload carried by boosters from NASA

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

booster_version	pmass
F9 v1.1	2928
F9 v1.1 B1003	500
F9 v1.1 B1010	2216
F9 v1.1 B1011	4428
F9 v1.1 B1012	2395
F9 v1.1 B1013	570
F9 v1.1 B1014	4159
F9 v1.1 B1015	1898
F9 v1.1 B1016	4707
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

First Successful Ground Landing Date

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

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

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

- Calculating total number of successful and failure mission outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

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

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

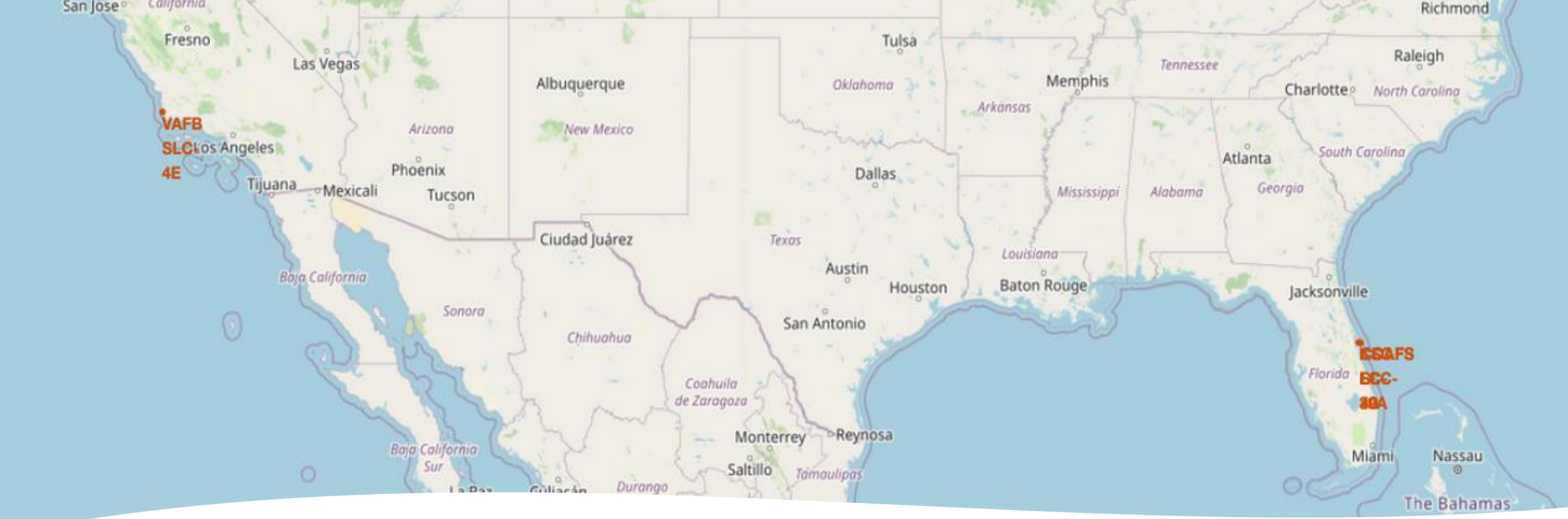
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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
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

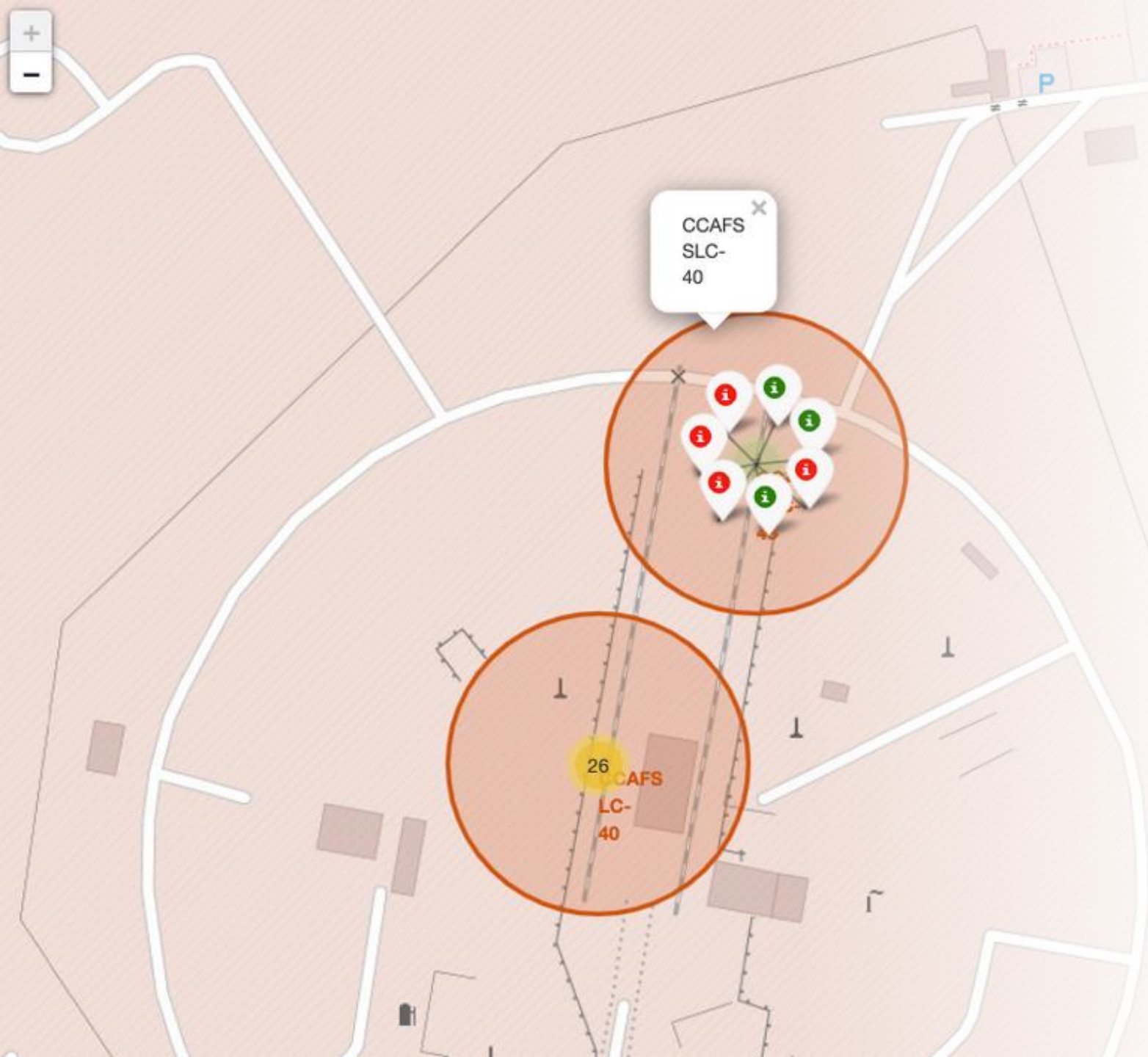
Section 4

Launch Sites Proximities Analysis



Launch Sites Marking

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.
- Three of the four launch sites are at the same place and can be seen on the lower right corner on the map above.

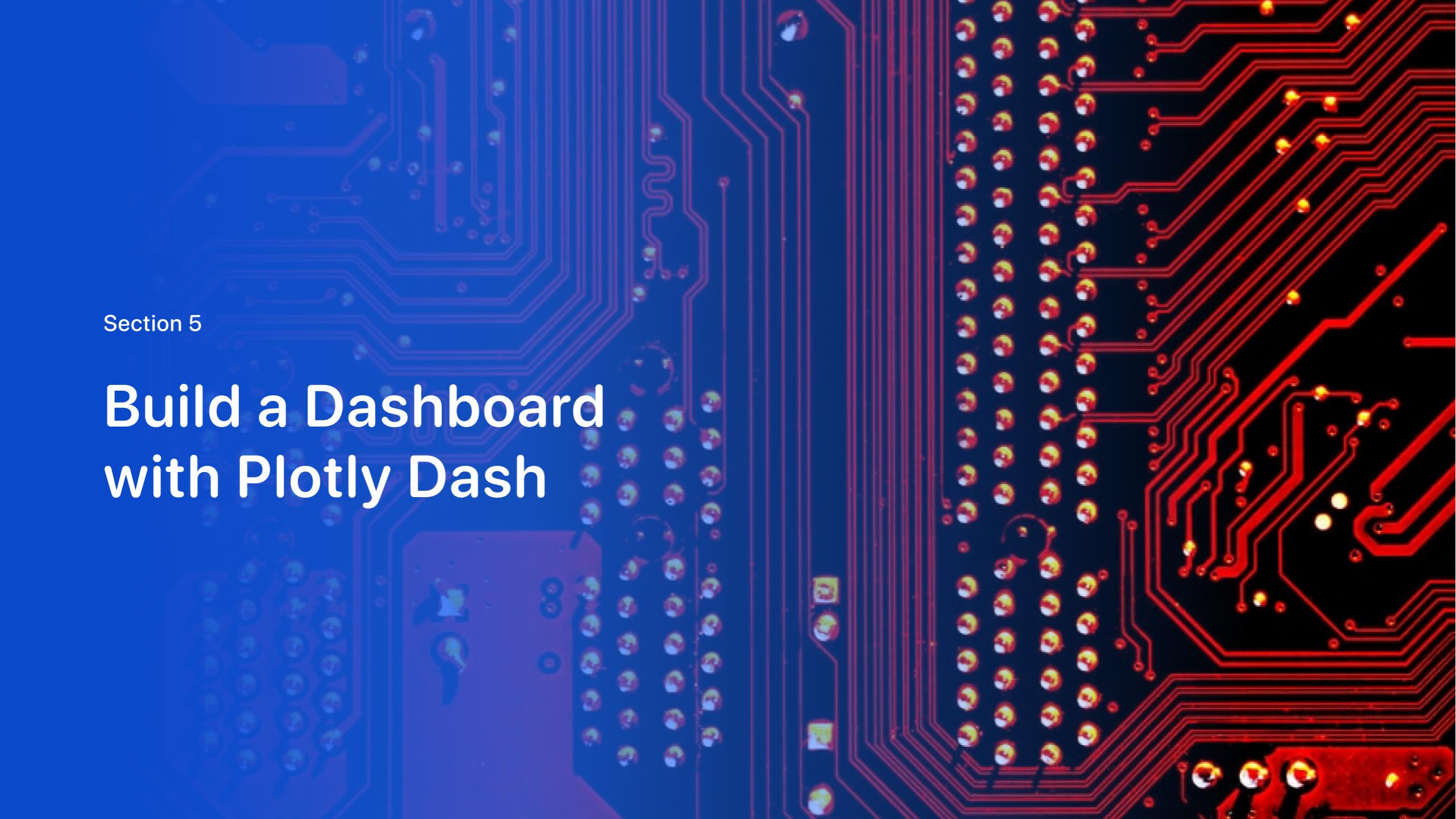


Color Coded Launches

- Color-labeled markers make it easier to identify which launch sites have relatively high success rates.
  - Green Markers visualize successful launches
  - Red Markers visualize Unsuccessful launches
- Interactive map makes it easier for the user to visualize the data inside a map

Distance Marking to Launch Sites Proximities

- Every launches sites and its proximities are marked with a line which states the distance between them.
- Launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed on the map.
- Are launch sites close to its proximities was the main question answered through the interactive map above.

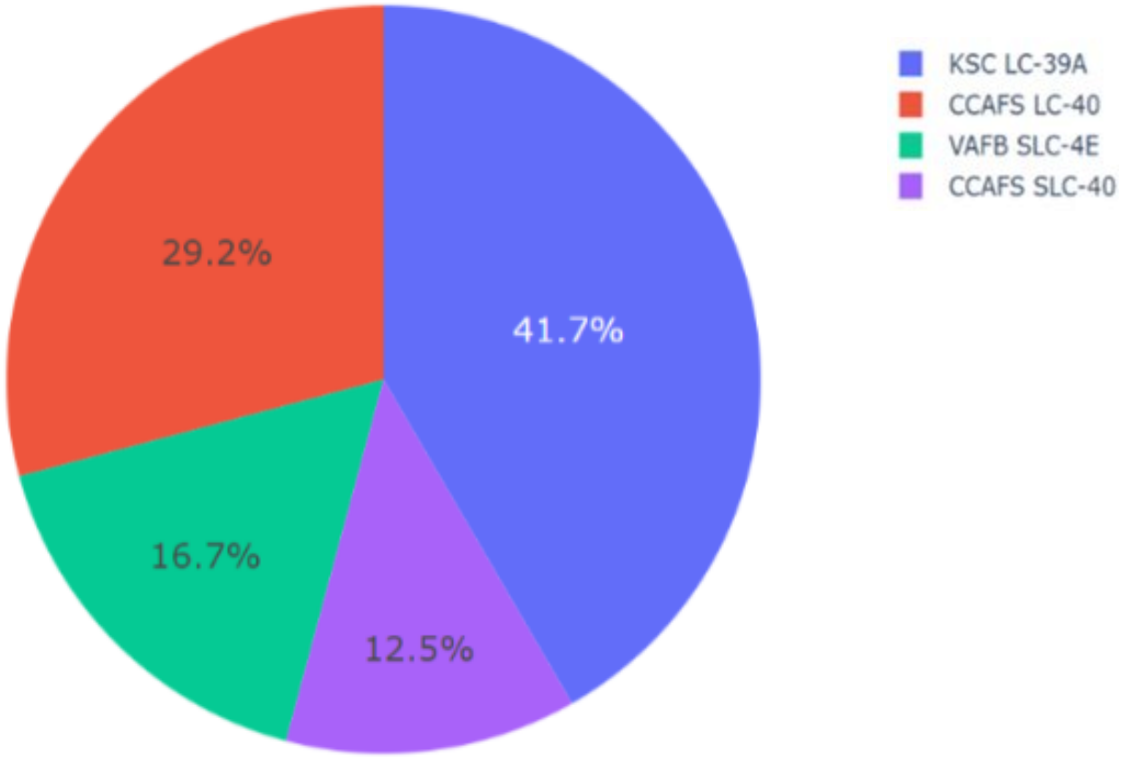


Section 5

Build a Dashboard with Plotly Dash

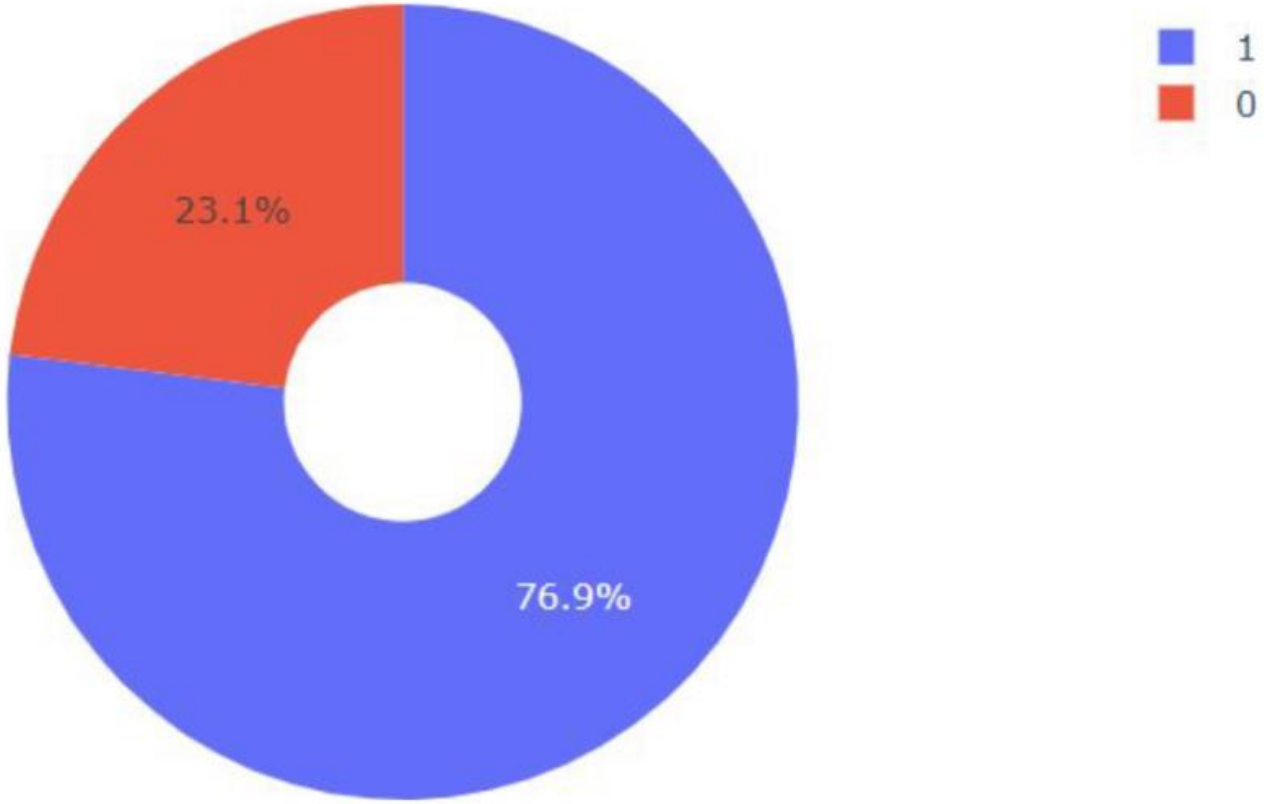
Successful Landing outcomes by Site

- 41.7% was the highest success ratio, achieved by the KSC LC-39A Launch Site.
- 12.5% was the lowest success ratio, achieved by CCAFS SLC-40
- Success ratio of each site individually was also displayed if the site was selected from the plotted dropdown menu



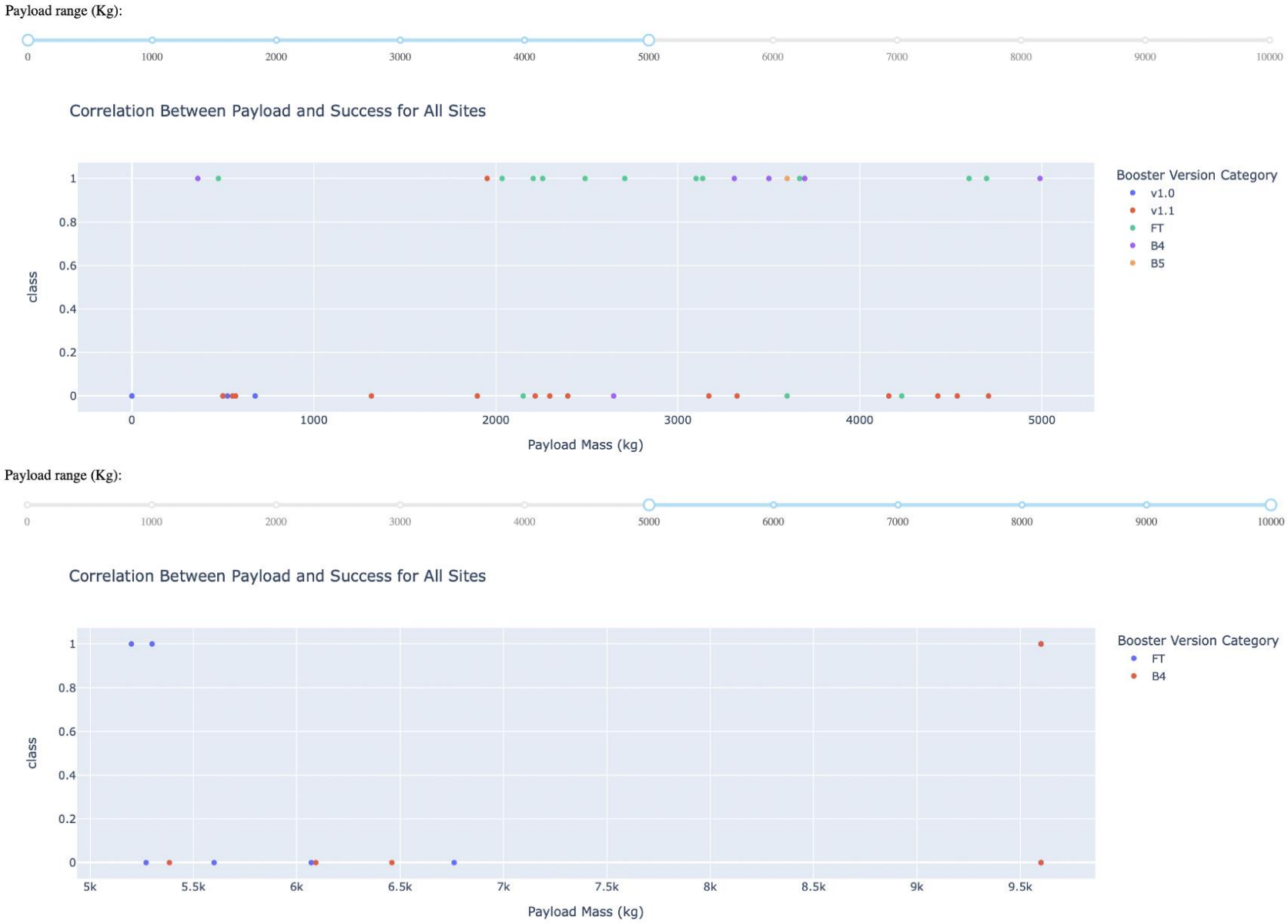
Most Successful Launch Site

- 76.9% was the success rate achieved by the KSC LC-39A Launch Site.
- 23.1% was the failure rate achieved by KSC LC-39A Launch Site.



Payload Mass vs. Launch Outcome

- Most successes were achieved when payload was between 2000 and 5500 kg.
- However, there is no relationship between the payload mass and launching outcome of the site.



Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Based on Testing Accuracy all models performed exactly same. One reason for this is the low amount of testing data.
- However, Training Accuracy of all models were not so similar, and a model can be chosen from the scoring charts.
- Decision Tree Model performed the best on training data, hence it can be chosen as the best performing model.

Testing Accuracy

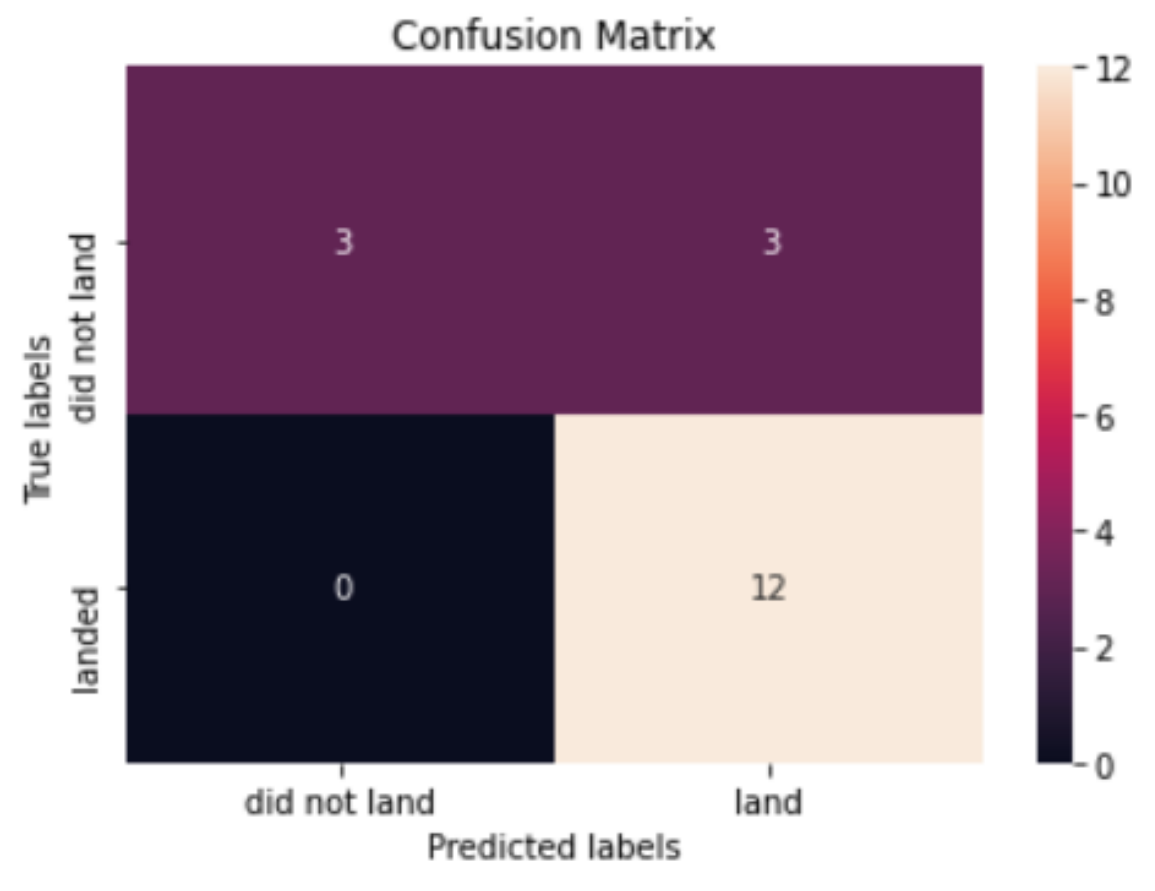
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Training Accuracy

	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

Confusion Matrix

- Examining the confusion matrix, we see that all models had the same results.
- We see that the major problem is false positives, as 3 of the unsuccessful launches were predicted to be successful.



Conclusions

- Decision Tree Algorithm works best for the given dataset.
- Although there is no relation between a successful launch and payload mass, most of the better results were from launches with low payload.
- There is a positive linear relationship between the increasing year and success ratio.
- Launch Site KSC LC-39A had the highest amount of success
- ES-L1, GEO, HEO and SSO had 100% success ratio.
- All launch sites were close to the line of equator and the coast line.

Appendix

- Github URL: Link to the Course Labs
- Github URL: Link to All Course Labs of Professional Certificate
- Special Thanks to all course instructors for their wonderful efforts into creating an amazing experience for all students. (Instructors list)

Thank you!