Winning Space Race with Data Science

Louie Wang

Mar 30, 2022

Outline

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

Executive Summary

- Methodologies

  The report is trying to analyze SpaceX's Falcon 9 launching records in order to give our company, SpaceY, a more competitive niche in the industry.

  The report will start from Data Collection stage, followed by Data Wrangling and EDA using both visualization and SQL. Then several interactive visualizations, including dashboards, are introduced. Finally it will conclude based on a few machine learning algorithms.

- Summary of all results

  Based on the results from machine learning algorithms, the best method is Decision Tree.

Introduction

- Project Background

  The commercial space age is already here and there are many existing companies trying to make space travel affordable for everyone. SpaceX is among one of the most successful ones by reducing its costs of each spacecraft launch. In particular, the project is focusing on the data recorded for Falcon 9 from SpaceX. What makes Falcon 9 special is that they are able to recover the first stage of a launch, which substatially reduce the cost of each launching mission. But SpaceX does not always recover the first stage successfully.

- Problems

  This project is trying to determine if SpaceX will reuse the first stage after each launching mission by using the publicly available information and machine learning models.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX REST API, as well as Webscraping from Wikipedia

- Perform data wrangling

  Create one-hot encoding column for landing outcomes

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

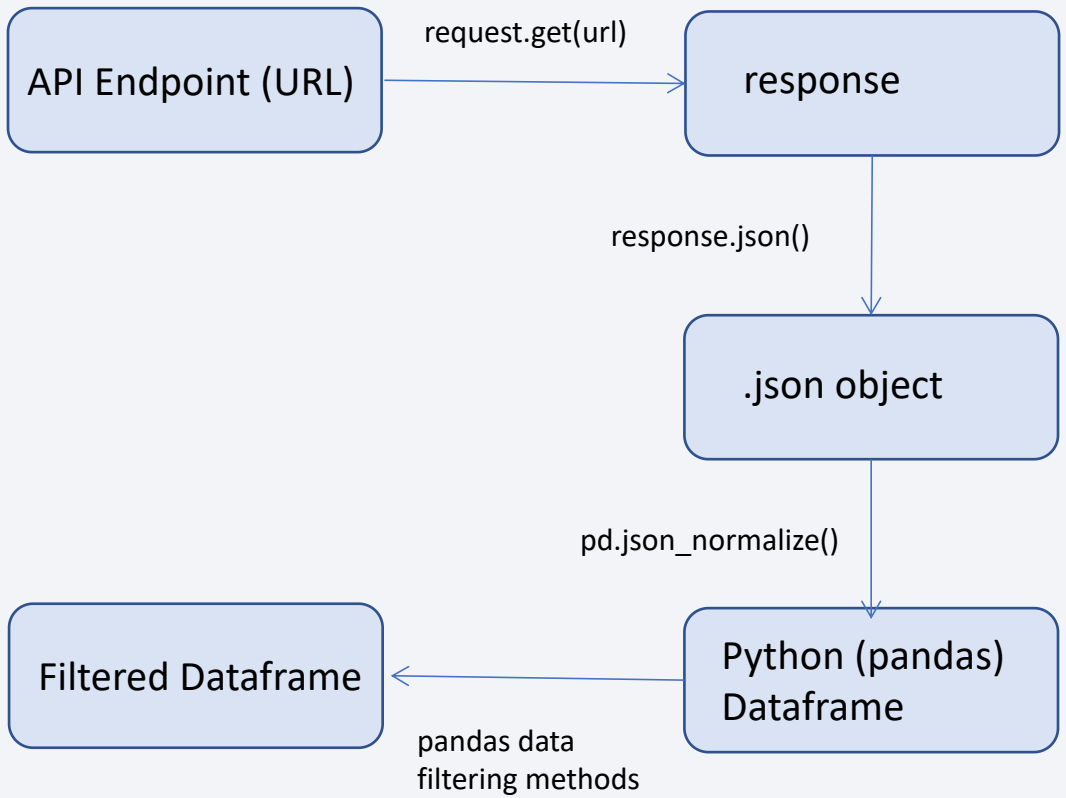
  Use cross validation for models and confusion matrices for evaluations

Data Collection

- **SpaceX API**
- **Scraping**

Data Collection – SpaceX API

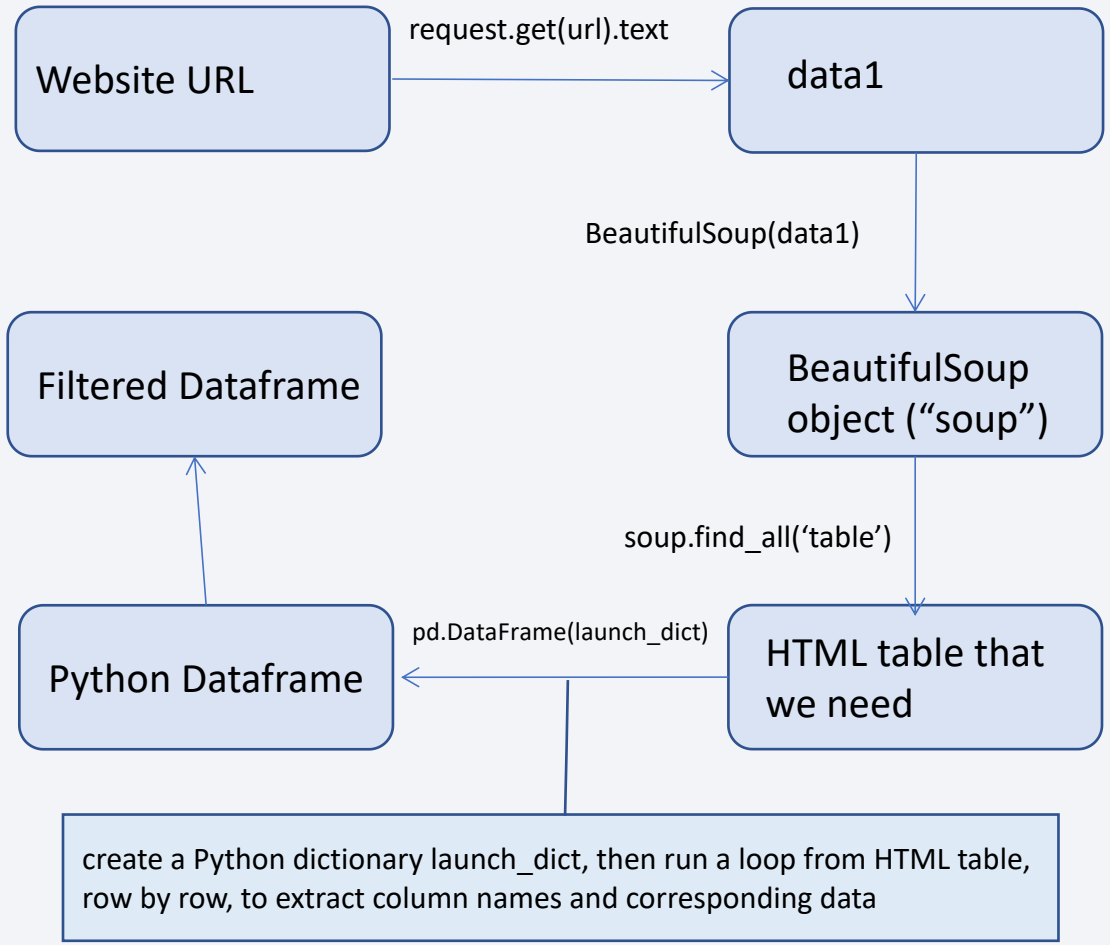
- Get data from an API (‘response’) by making an HTML request to the provided URL;
- Convert the response into a .json object, then use a Python pandas method to convert it into a pandas Dataframe;
- Lastsly, use pandas and self-defined functions to extract relevant information from the original data for our usage.



https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/Data%20Collection%20API.ipynb

Data Collection - Scraping

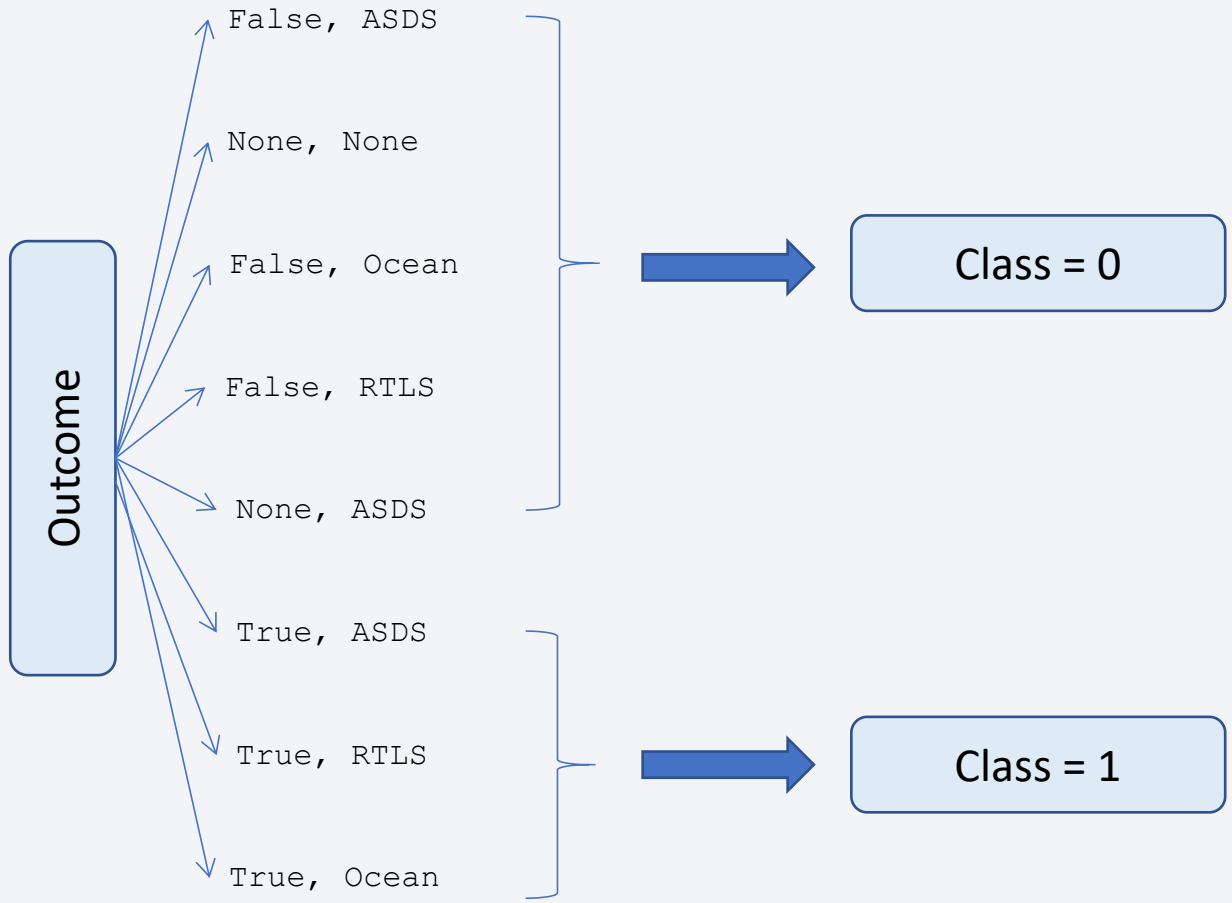
- Similar to API, we start with a URL and we access the information by using request;
- Create a BeautifulSoup object from the obtained information;
- Use “find_all” method to find all tables in the HTML page and pick the relevant one;
- Iterrate through each <th> to get the column names, which will later be used as column names in the pandas dataframe;
- Create a python dictionary with the column names, then iterrate through the HTML script to scrape data (<tr> - <td>);
- Lastly, filter the dataframe



https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/Data%20Collection%20with%20Webscraping.ipynb

Data Wrangling

- Data Wrangling involves two main sections.
- First, we checked for distinct values in the following columns in order to have a better understanding of the dataset: **LaunchSite**, **Orbit**, and **Outcome**.
- Second, we converted Outcome results into a binary column Class with 1 being “successfully landed” and 0 otherwise.



https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/EDA.ipynb

EDA with Data Visualization

We plotted the following visualization (more details to follow in later slides)

- Flight Number vs. Payload
- Flight number vs. Launch Sites
- Payload vs. Launch Sites
- Success rate vs. Orbit types
- Launch success rate over years
- Flight number vs. Orbit types
- Payload vs. Orbit types

https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/EDA%20with%20Visualization.ipynb

EDA with SQL

Tasks we performed:

- Unique launch sites in the mission
- 5 records where launch sites begin with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when first successful landing outcome in ground pad was achieved
- Names of boosters which have success in drone ship and have payload mass between 4000 kg and 6000 kg
- Total number of successful and failed mission outcomes
- Names of booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names in 2015
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20

https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Map objects created and added to the folium map

- Circle objects for each launch site with its name
- Marker cluster for each site (green for a success launch and red for a failuer)

  Success and failure are determined based on the ‘class’ column (1 for success, 0 for failure)

  Cluster is used to avoid overlapping of each individual launching point
- Mouse position showing the current latitude and longitude

  This ensures us to keep track of the coordinates of a point we are interested in measuring
- Distance marker for a launch site and a proximity point
- A polyline between a launch site and the selected proximity point

https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

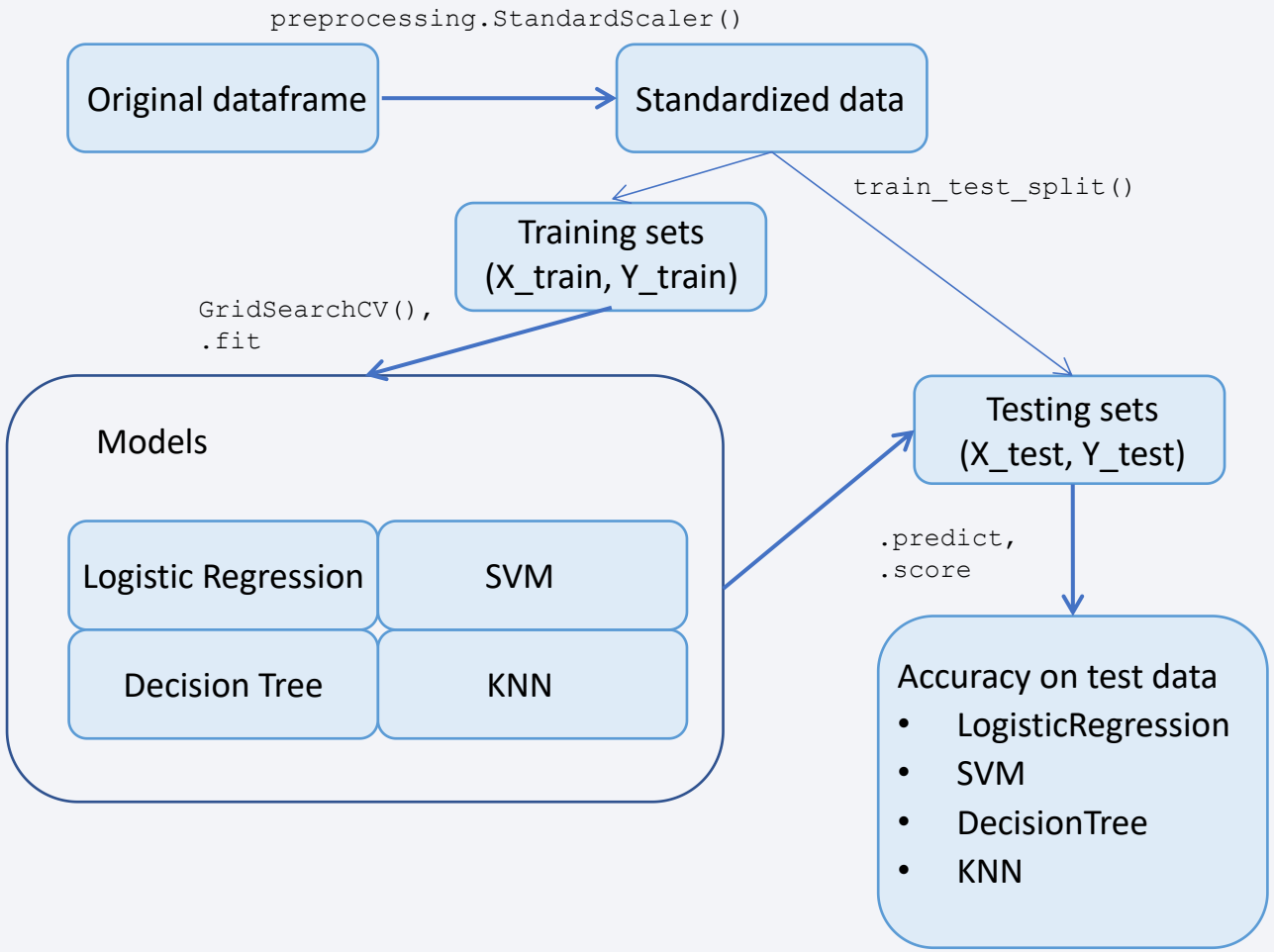
Plots/graphs and interactions added to a dashboard

- A dropdown menu for Launch Site selection
  - This menu enable the user to narrow down the selection
- A pie chart showing launch data for each launch site
  - The pie chart visualizes the proportions of successful and failed launches
- A payload range slider
  - The slider helps narrow down the selection by selecting interested payload range
- Scatter plot showing correlation betweem payload and launch sites
  - The scatter plot visualizes the correlation

https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- Standardize and normalize the data first. Then split the data into testing and training subsets.
- Models built: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K Nearest Neighbors (KNN)
- Use cross validation for each model to find the parameter(s) that can lead to the best result in training data
- Use testing data to get the accuracy for each model and determine the best one



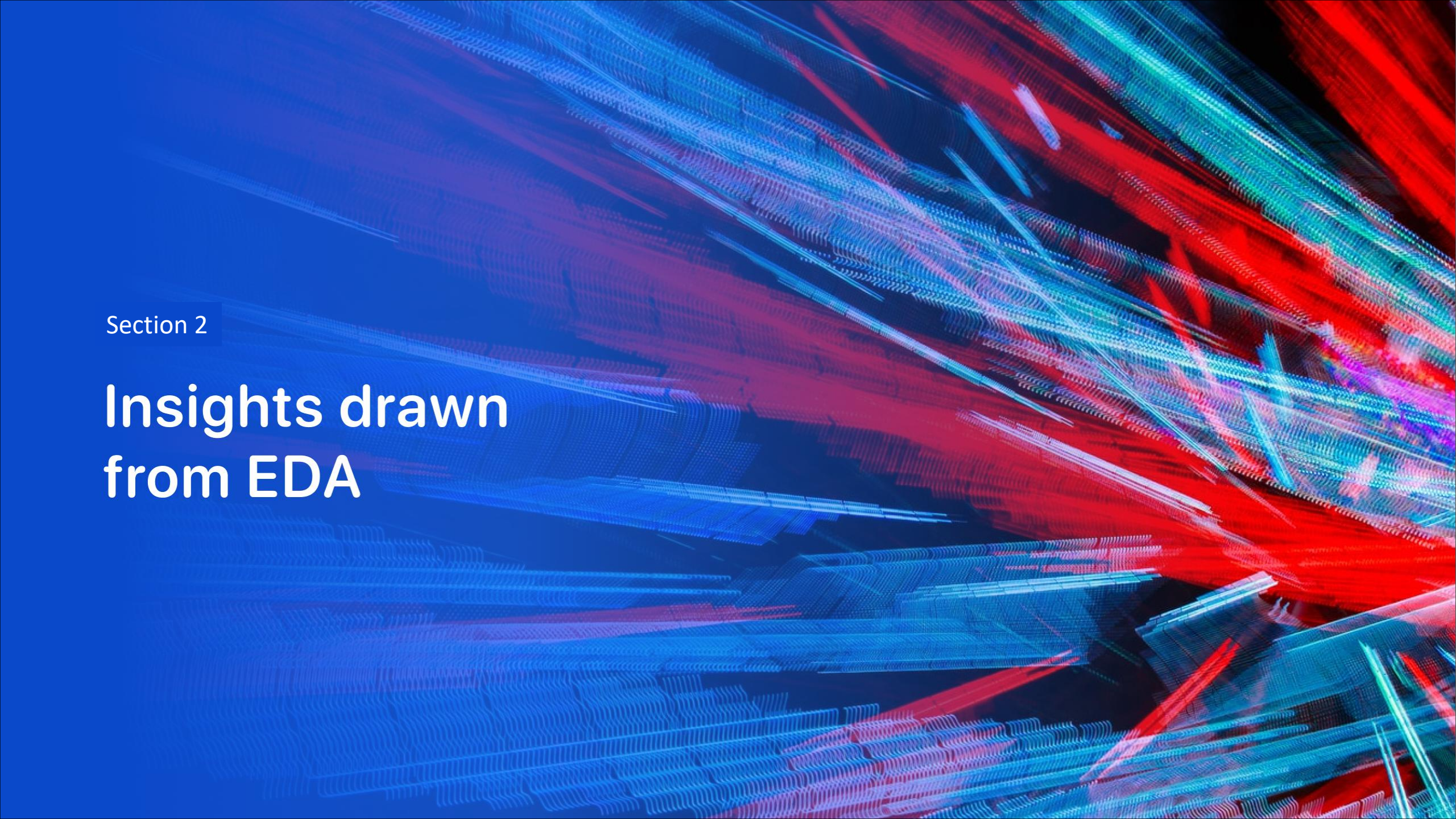
https://github.com/lsjzw5/coursera_IBM_Capstone/blob/6b552ceefc7765a97efb376459432051bf15bcdd/Machine%20Learning%20Prediction.ipynb

Results

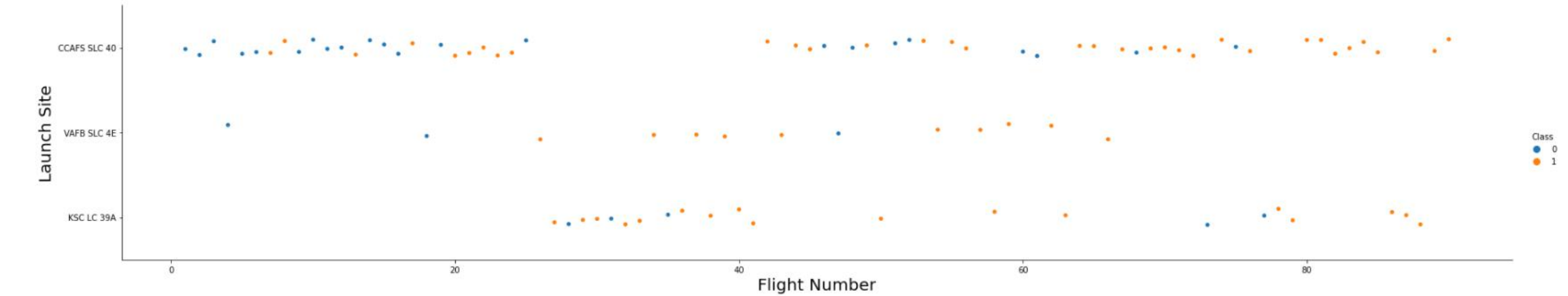
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



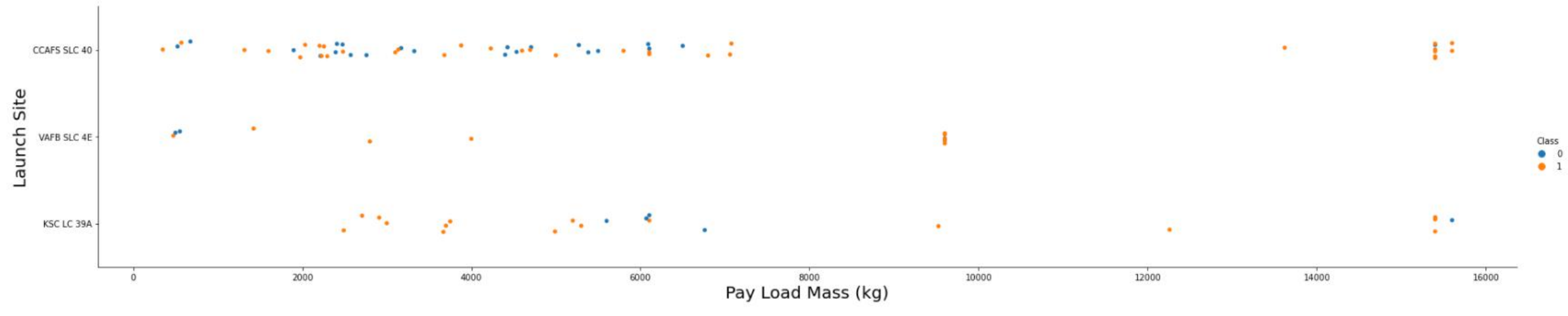
Flight Number vs. Launch Site



This plot shows that site CCAFS SLC 40 hosts the most launching missions and among the first 30 launches, most of them were not successful.

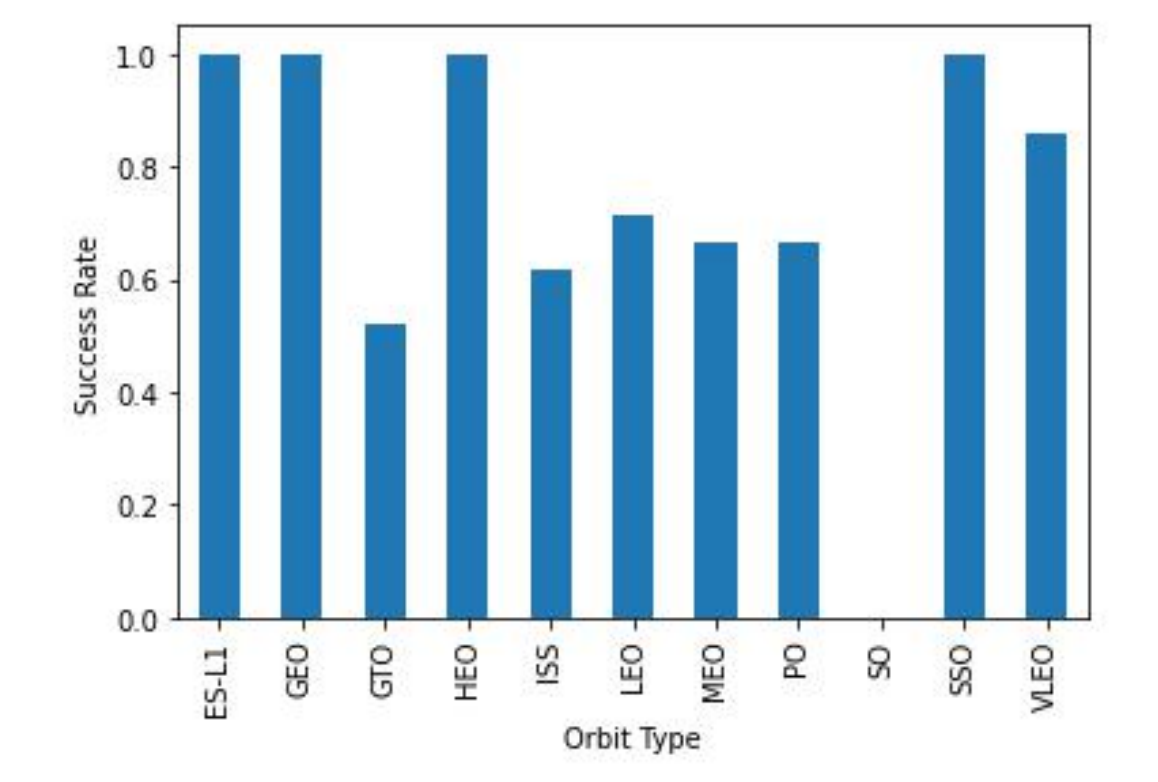
And among all three launch sites, VAFB SLC 4E seems to have the highest successful rate.

Payload vs. Launch Site



It is noticable that at VAFB SLC launch sites, there are no rockets launched for heavy payload mass (greater than 10000 kg).

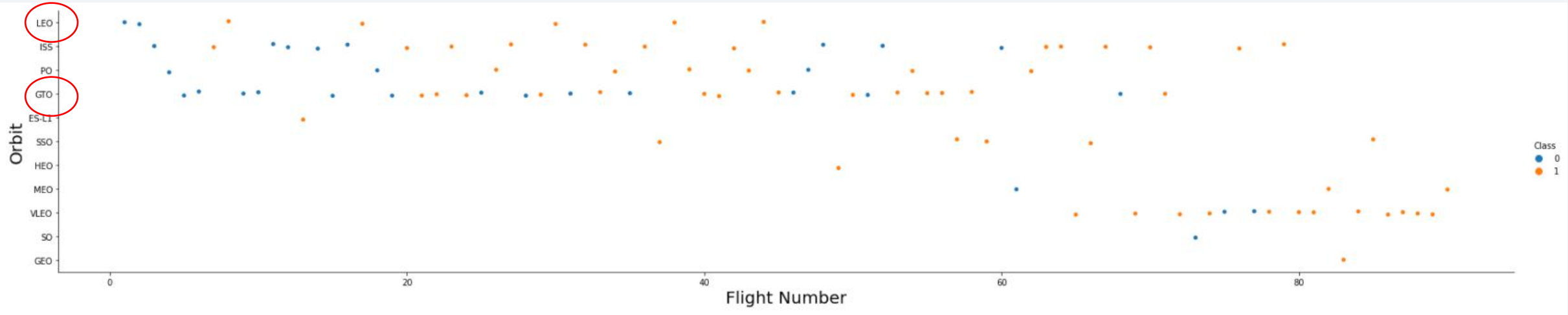
Success Rate vs. Orbit Type



There are several orbits that have the highest success rates: ES-L1, GEO, HEO, and SSO.

The lowest success rate is at the SO orbit.

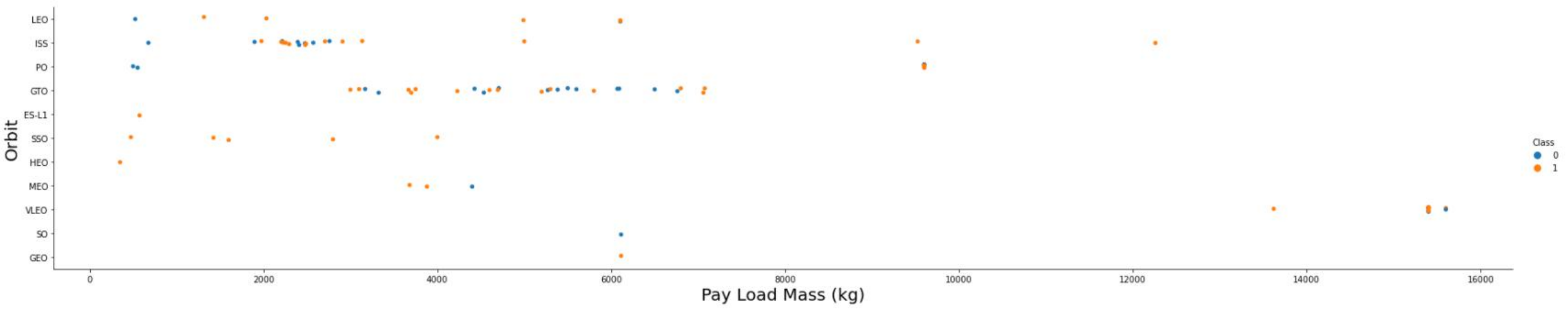
Flight Number vs. Orbit Type



In the LEO (Low Earth Orbit), the success rate appears to be related to the number of flights;
On the other hand, there seems to be no relationship between flight number and success rates in the GTO (Geostationary Transfer Orbit).

Plus, launches to SO and GEO are relatively recent (after at least 70 flights)

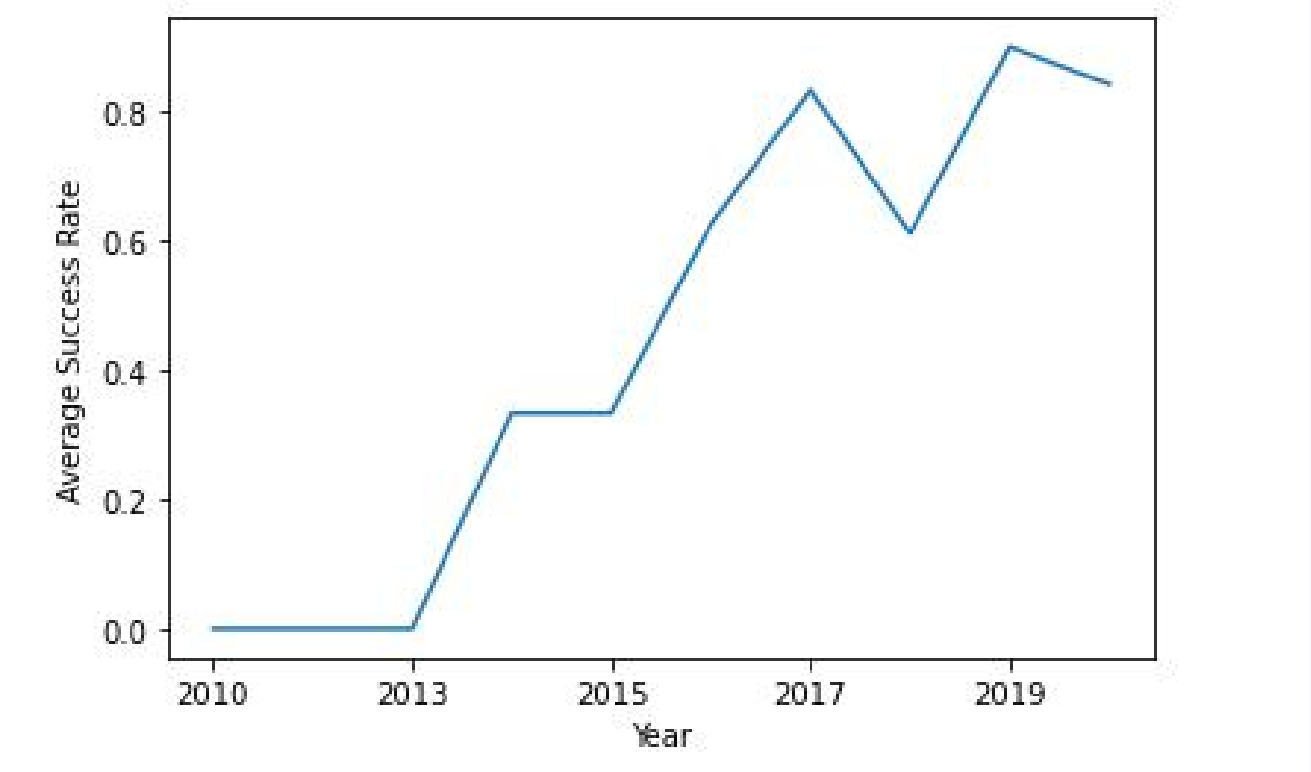
Payload vs. Orbit Type



With heavy payloads (>10000 kg), the success rate are more for LEO and ISS.

But for GTO, we cannot distinguish this relationship well.

Launch Success Yearly Trend



The success rate has been generally increasing since 2013, with a drop in 2018 and 2020.

All Launch Site Names

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

The four launch sites in all space missions are:

- 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

The results should all have launch sites named either “CCAFS LC-40” or “CCAFS SLC-40”

Total Payload Mass

total_payload__mass_kg_
45596

This is the total payload mass carried by boosters launched by NASA (CAS)

Average Payload Mass by F9 v1.1

average_payload_mass__kg_
2928

This is only the average payload mass by booster version F9 v1.1

First Successful Ground Landing Date

This is the date of the first successful landing outcome in **ground pad**.

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

These are the booster versions that have success in drone ship landings.

And their payload masses are between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

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

Among all missions, only 1 was a failure. It is worth noting that this table displays the outcomes for MISSION, not LANDING. The landing outcomes are more diversed and have a lower success rate in general.

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

These are the booster versions that have carried the maximum payload mass.

2015 Launch Records

Landing _Outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

These are the only two records of failed landing outcomes in year 2015, both launched from CCAFS LC-40.

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

Landing _Outcome	num_landing_outcome
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

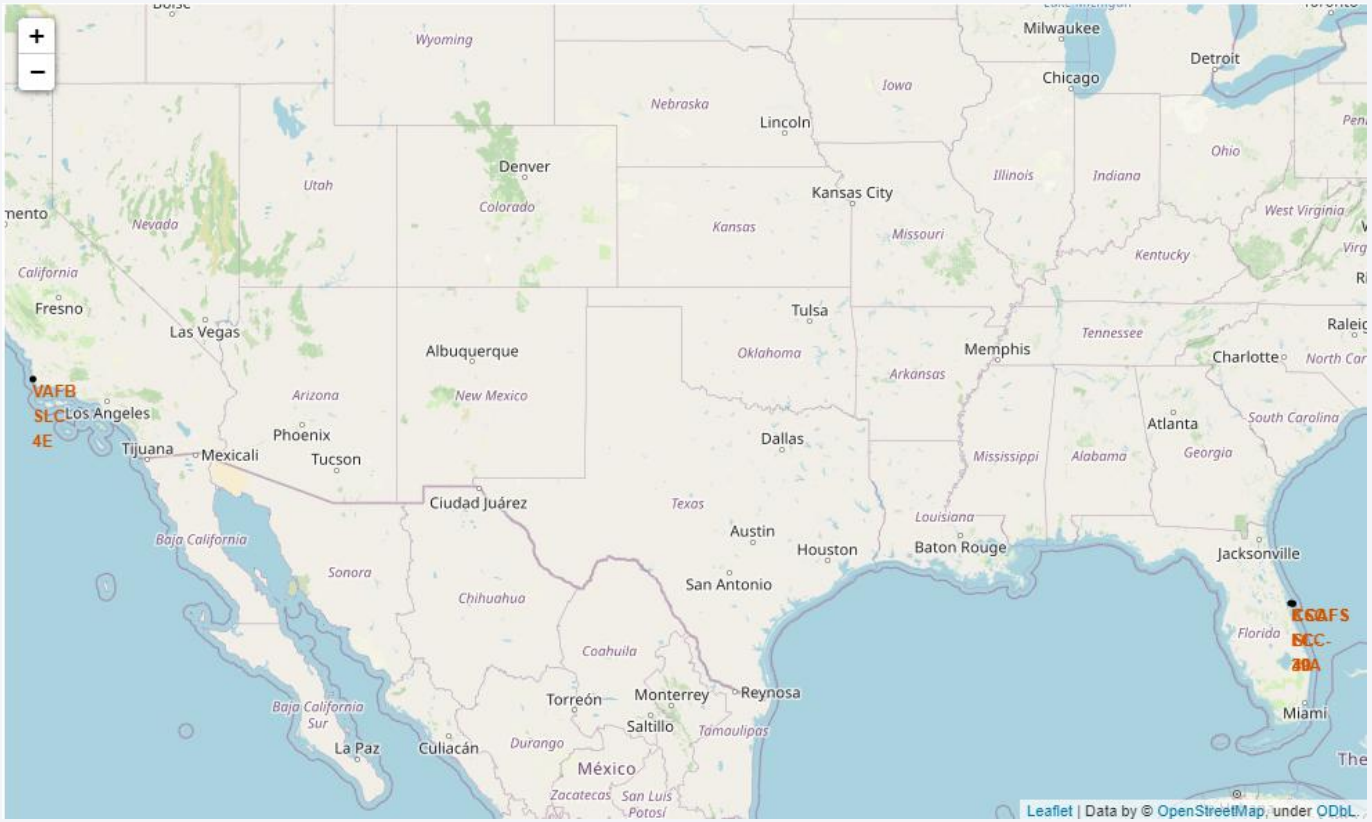
The table displays the ranking of all landing outcomes within the selected date range.

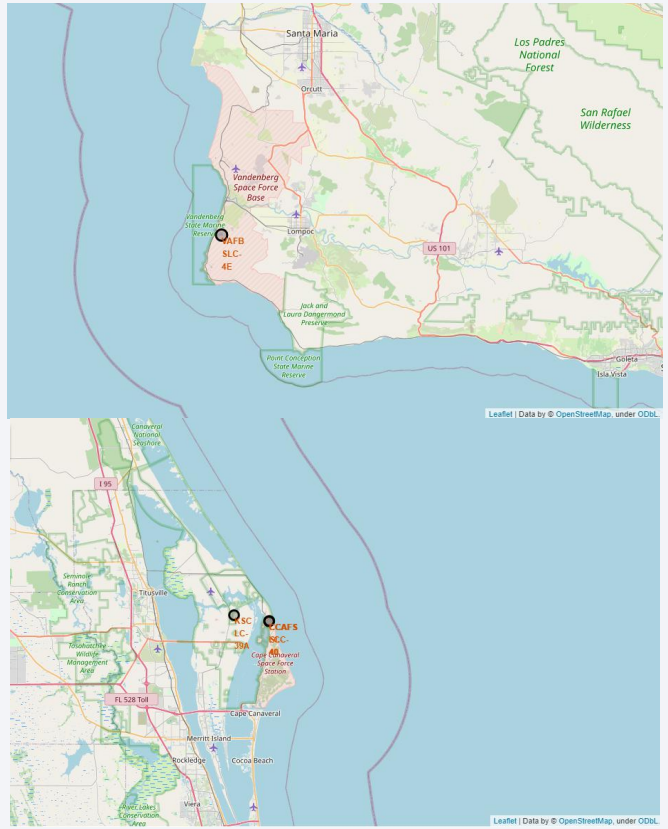
The number one landing outcome is “No attempt”.

None of the parachute landing options went successful during these time range.


Section 3
Launch Sites
Proximities Analysis

Folium Map - All Launch Sites Location



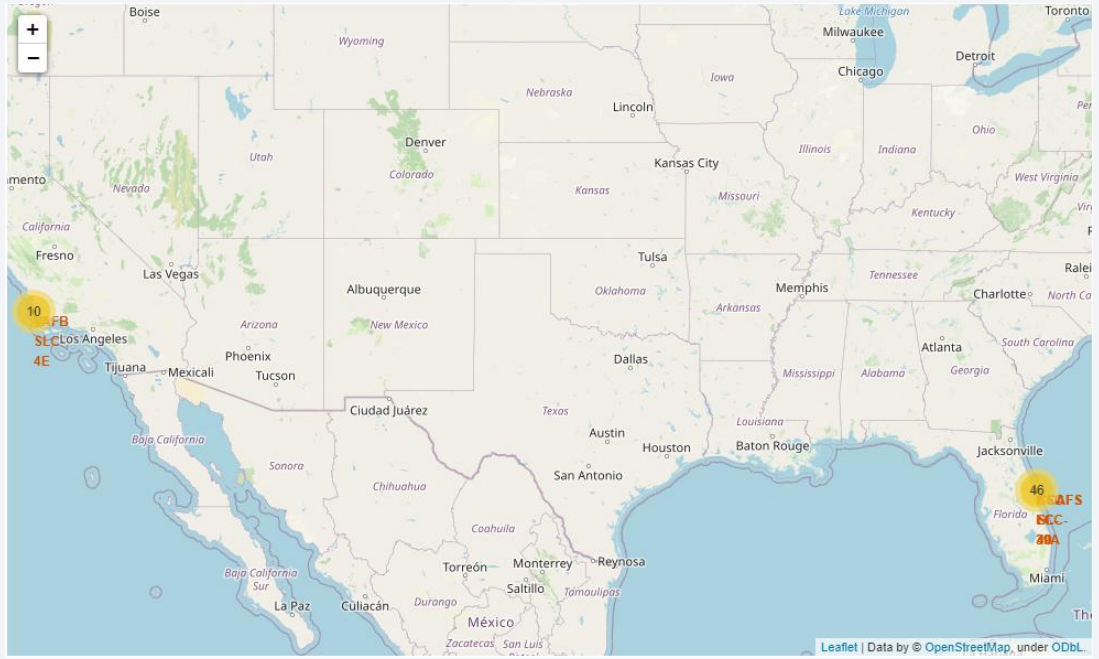


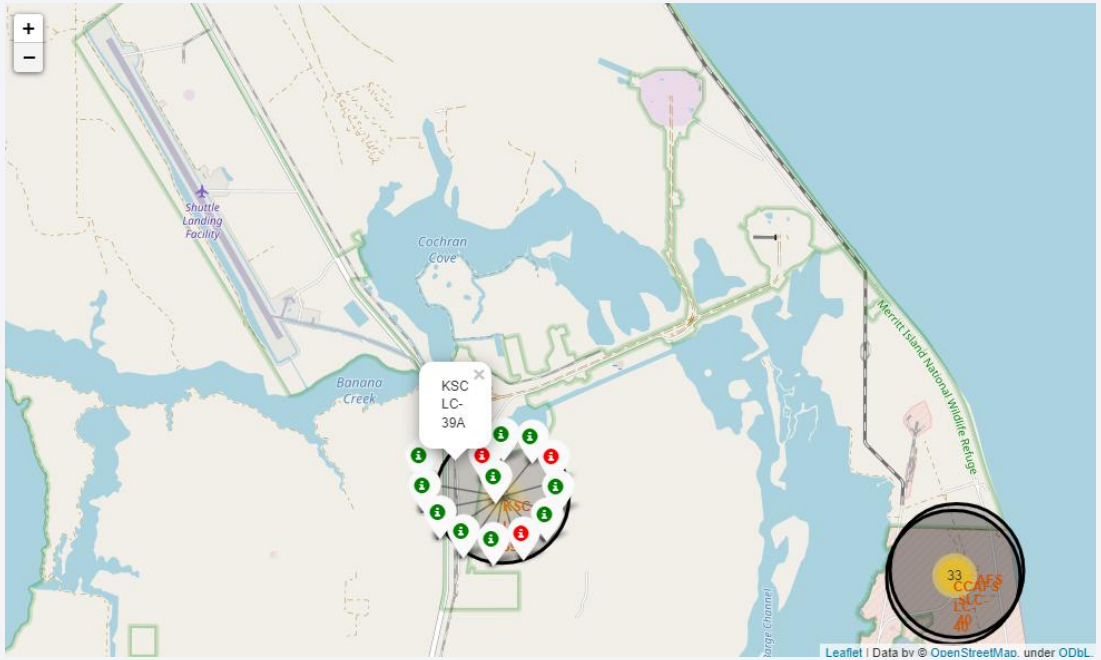
All launch sites are located in either CA or FL states.

They are close to the equator line and are all very close to coastlines.

Sites CCAFS LC-40 and CCAFS SLC-40 are in almost identical locations.

Folium Map - Success/Failed Launches for each Site

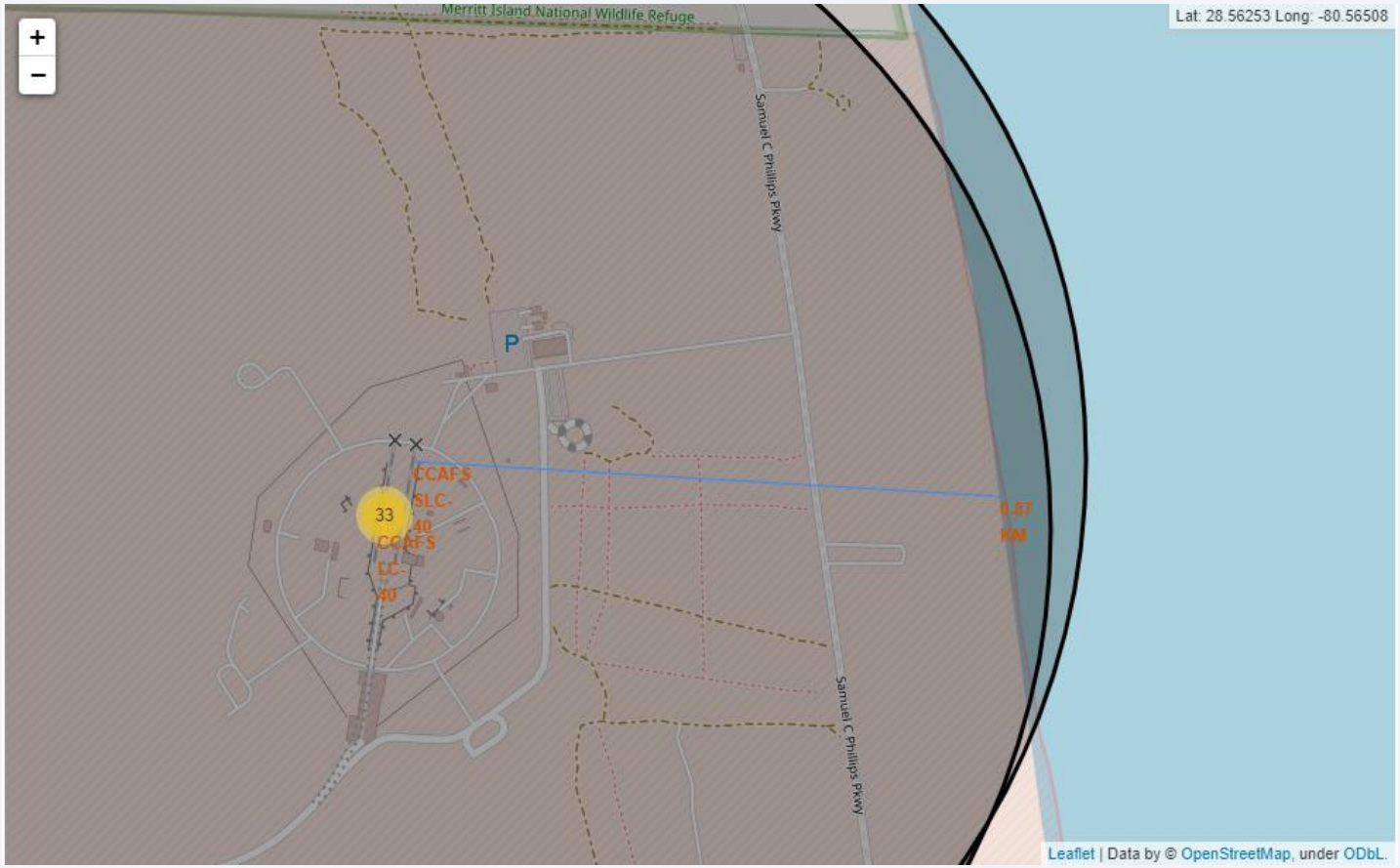




As shown in the left map, all sites are now marked with a number, indicating how many launches are recorded.

After zooming in a paarticular site and click on the number, as shown on the right sample map, a cluster of data is shown to display the ‘class’ of each launch, with green being a successful launch.

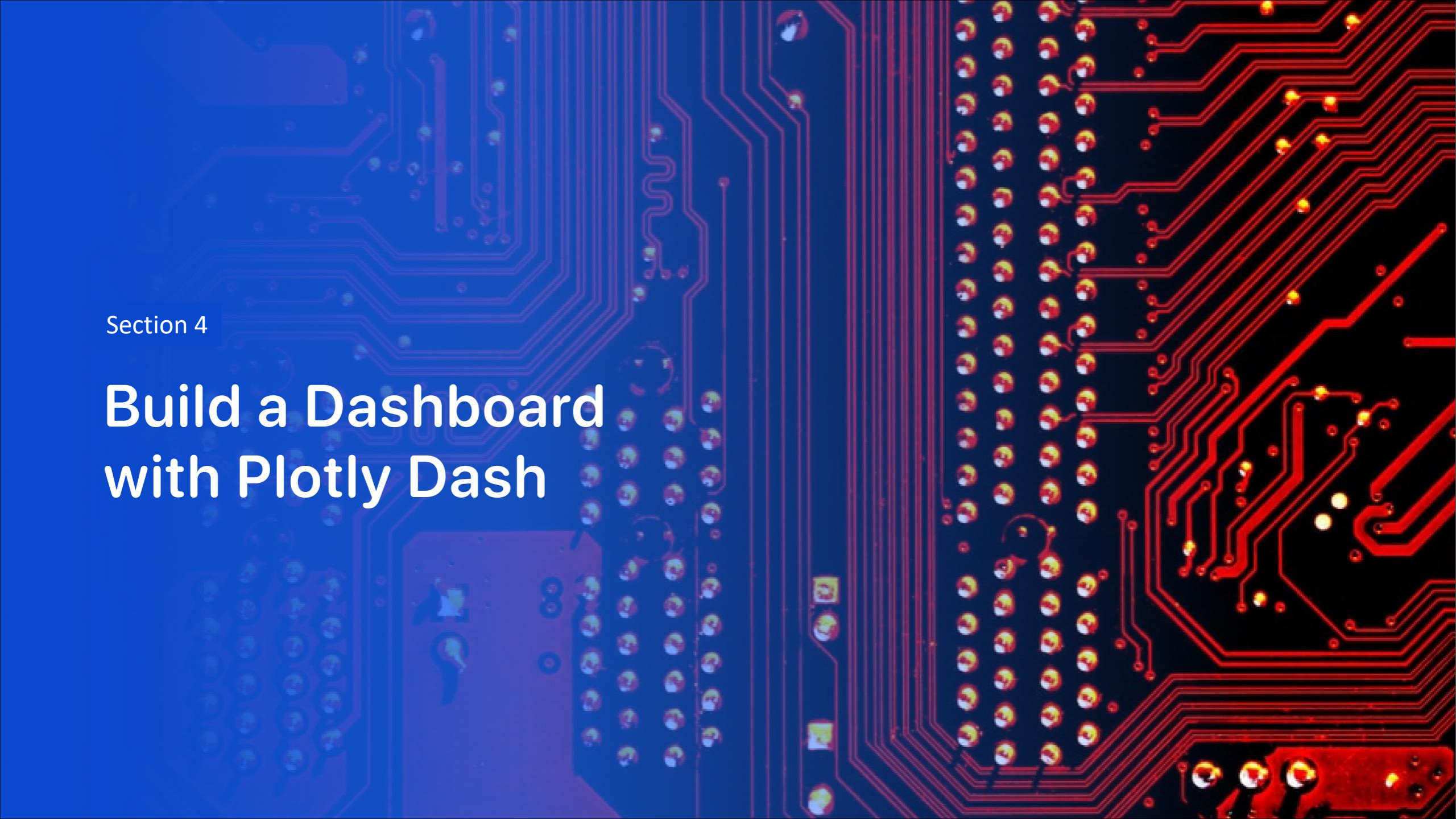
Folium Map - Distance between a Launch Site to its Proximities



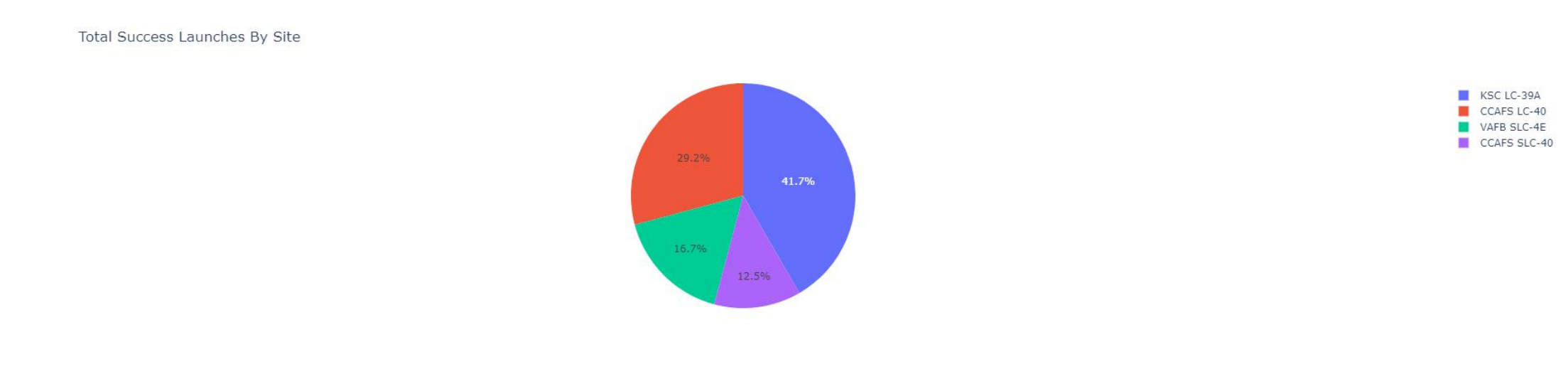
The map displays the distance between a launch site to one of its proximities (it can either be a railroad, a street, or in our case, a coastline).

Both the distance and a line drawn between the two locations are shown.

In this map, the distance between site CCAFS SLC-40 to its closest coastline is approximately 0.87 km.

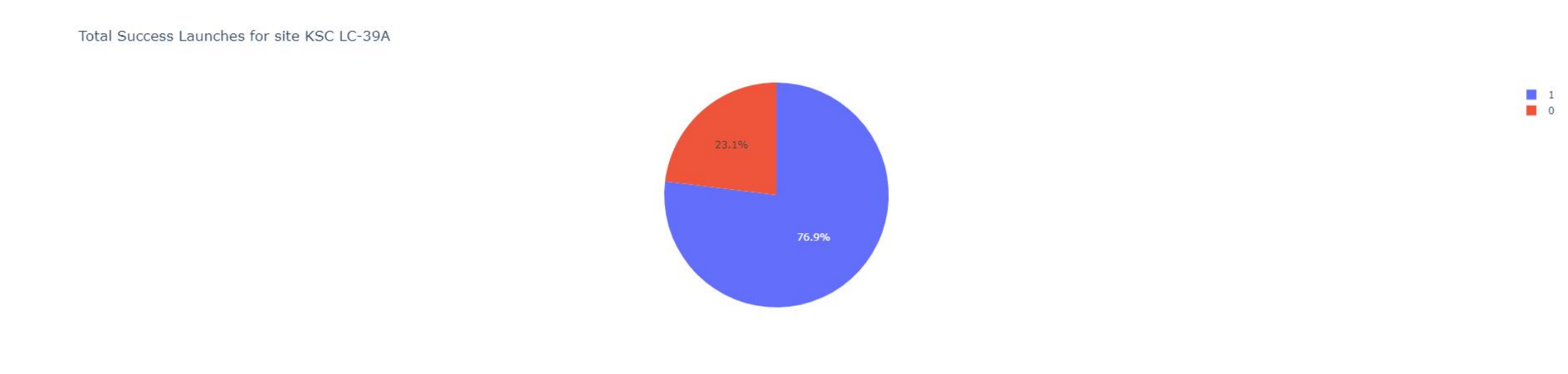

Section 4
Build a Dashboard with Plotly Dash

Dashboard - Launch Success for ALL Sites



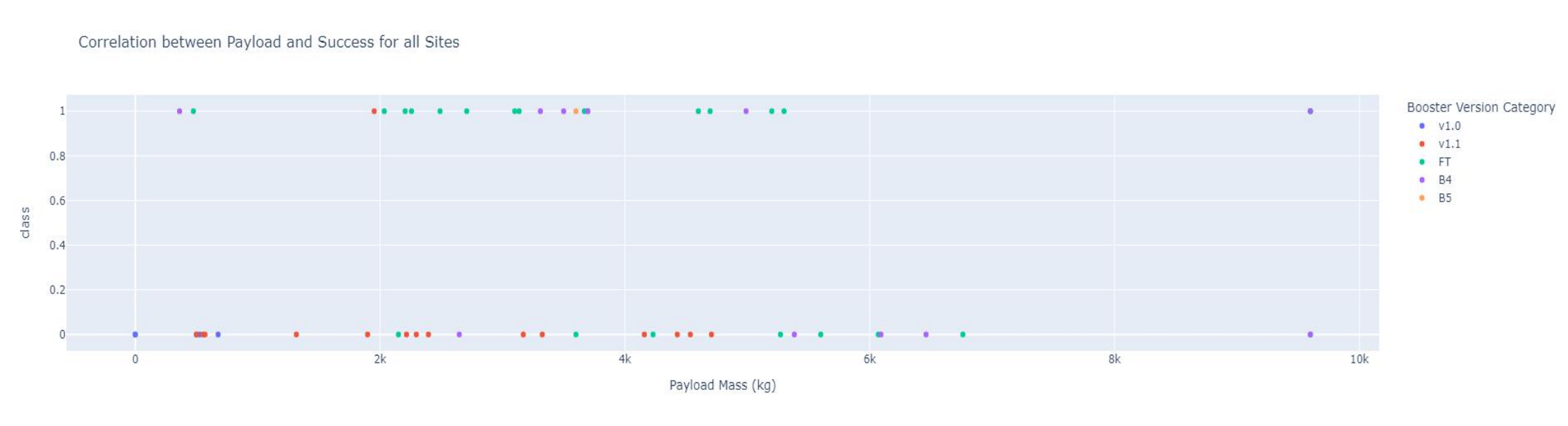
Site KSC LC-39A has the most success launching.
Site CCAFS SLC-40 has the fewest success launching.

Dashboard - Site with the Highest Success Ratio



At the site with the highest success ratio, we see that about 80% of the lauches were successful.

Dashboard - Payload vs. Launch Outcome, ALL Sites



The x-axis is payload mass in kg, y-axis is class (launch outcomes), and the colors are based on booster versions.

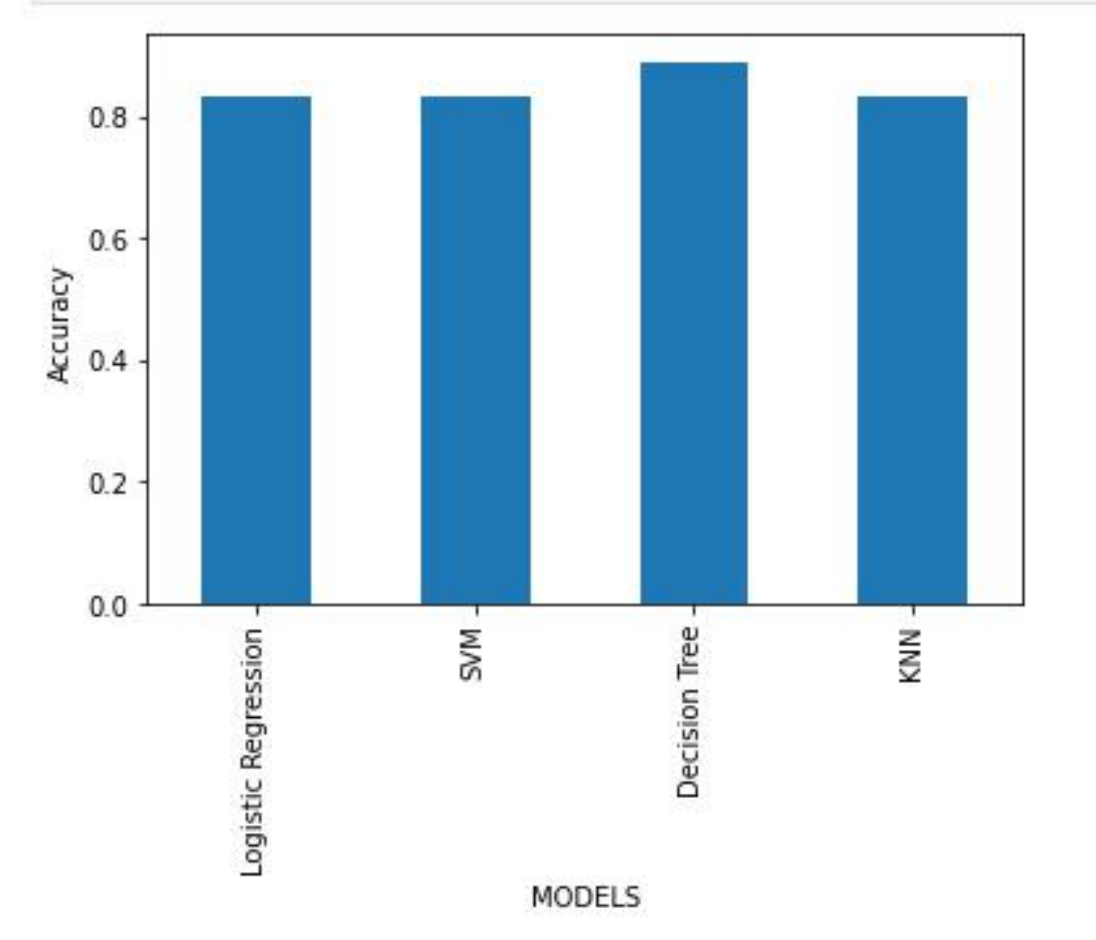
We see that booster version FT has a very high success rate as most of its points are in class = 1;

On the other hand, booster version v1.1 has a very low success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

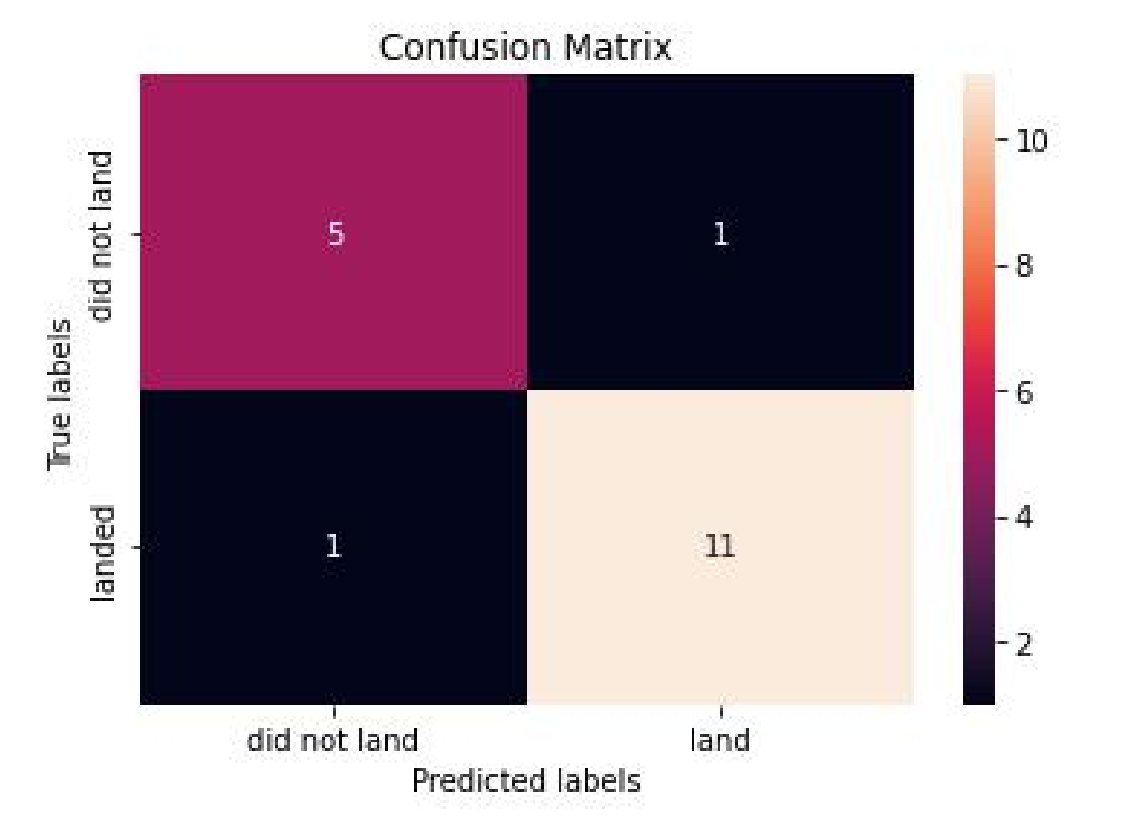


For this particular train-test split and cross validation, the best classification model is the **Decision Tree**, with the accuracy over 0.85.

However, all other models performed relatively well, too. All are having an accuracy over 0.8.

Such result may not be reproducible. It may vary based on the split of data and the cross validation process each time.

Confusion Matrix



The better a model is, the more accurately it predicts each class/label.

In this model, it correctly predicts 5 out of 6 cases where the rocket “did not land”. And it correctly predicts 11 out of 12 cases where the rocket “landed”.

Conclusions

- Most launch sites are located near coastlines and are near the Equator line.
- The Decision Tree model gives the most accurate predictions on our data. But with the data we have, all four machine learning algorithms can perform well.
- There are many factors that can affect the result of a launching mission, including but not limited to: booster versions, the orbit types, payload mass, etc.

Appendix

- Data collection urls: SpaceX API, Scraping.
- Github Repo for all codes: https://github.com/lsjzw5/coursera_IBM_Capstone
- Data used for Plotly Dashboard can be found here.

Thank you!