IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bhavishya Vudatha

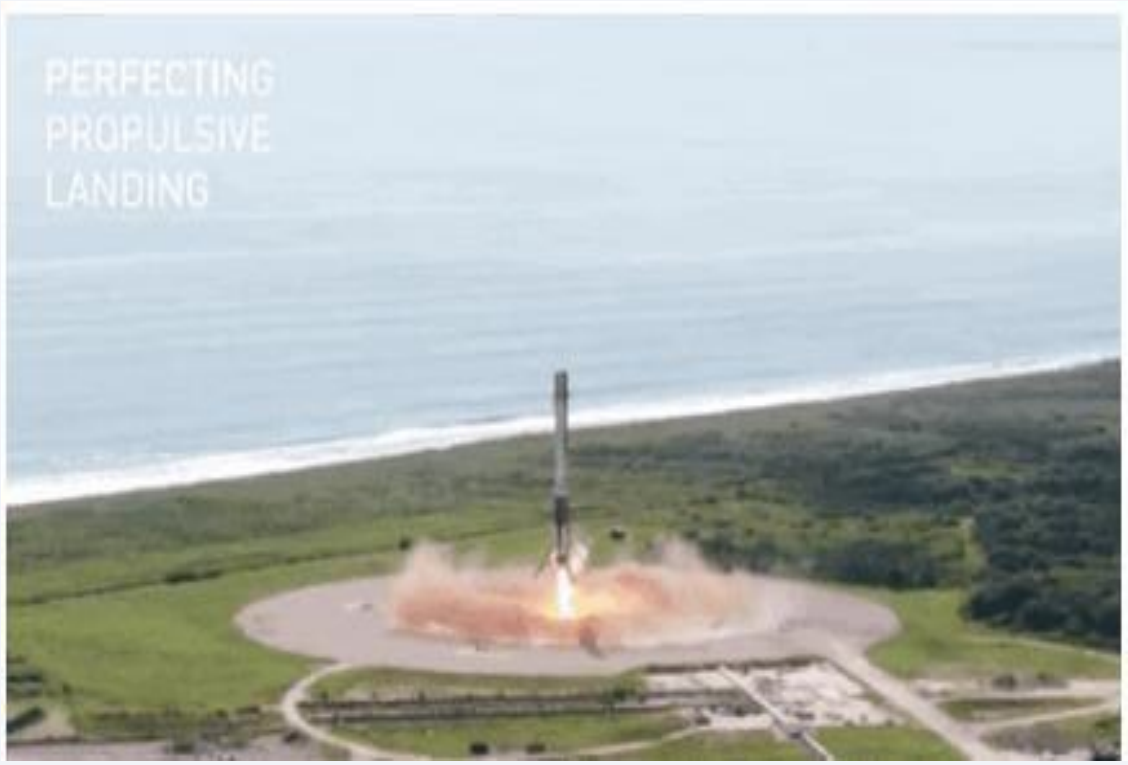
Outline

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

Executive Summary

- Summary of methodologies
  - SpaceX Data Collection using SpaceX API
  - SpaceX Data Collection with Web Scraping
  - SpaceX Data Wrangling
  - SpaceX Exploratory Data Analysis using SQL
  - Space-X EDA DataViz Using Pandas and Matplotlib
  - Space-X Launch Sites Analysis with Folium-Interactive Visual Analytics and Ploty Dash
  - SpaceX Machine Learning Landing Prediction

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.

- Problems you want to find answers

  In this capstone, we will predict if the Falcon 9 first stage will land successfully using data from Falcon 9 rocket launches advertised on its website.

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Exploratory data analysis (EDA) using SQL
- Visualization using Python pandas and Matplotlib
- Visual analytics using Folium and Plotly Dash
- Predictive analysis using various classification models

Data Collection

- Description of how SpaceX Falcon9 data was collected.
  - Data was first collected using the SpaceX API (a RESTful API) by making a GET request. The SpaceX launch data was requested and parsed using this GET request, then the response content was decoded as a JSON result, which was subsequently converted into a Pandas DataFrame.
  - After converting the data into a DataFrame, it was observed that each feature consisted of IDs instead of actual data. To resolve this, a series of additional API calls were made to extract the detailed data from various endpoints, replacing the IDs with the corresponding data in the DataFrame.
  - Also performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches of the launch records are stored in a HTML. Using BeautifulSoup and request Libraries, I extract the Falcon 9 launch HTML table records from the Wikipedia page, Parsed the table and converted it into a Pandas data frame.

Data Collection- API

- Data collected using SpaceX API (a RESTful API) by making a get request to the SpaceX API then requested and parsed the SpaceX launch data using the GET request and decoded the response content as a Json result which was then converted into a Pandas data frame

- Git hib URL - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - WebScraping

- Performed web scraping to collect Falcon 9 historical launch records from a Wikipedia using BeautifulSoup and request, to extract the Falcon 9 launch records from HTML table of the Wikipedia page, then created a data frame by parsing the launch HTML.

- After obtaining and creating a Pandas DF from the collected data, data was filtered using the ***Booster Version*** column to only keep the Falcon 9 launches, then dealt with the missing data values in the ***LandingPad*** and ***PayloadMass*** columns. For the ***PayloadMass*** , missing data values were replaced using mean value of column.
- Also performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
- Github URL - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- After obtaining and creating a Pandas DF from the collected data, data was filtered using the ***BoosterVersion*** column to only keep the Falcon 9 launches, then dealt with the missing data values in the ***LandingPad*** and ***PayloadMass*** columns. For the ***PayloadMass*** , missing data values were replaced using mean value of column.
- Also performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determined the label for the data.
- Git Hub URL - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with SQL

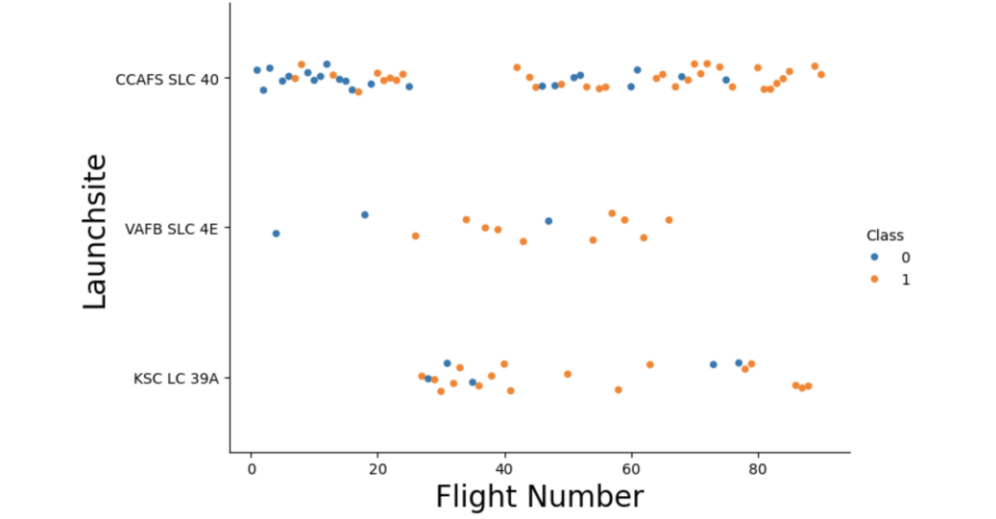
- Performed data Analysis using SQL
  - Load the dataset into the corresponding table in a Db2 database.
  - Analyzing data with various SQL Queries

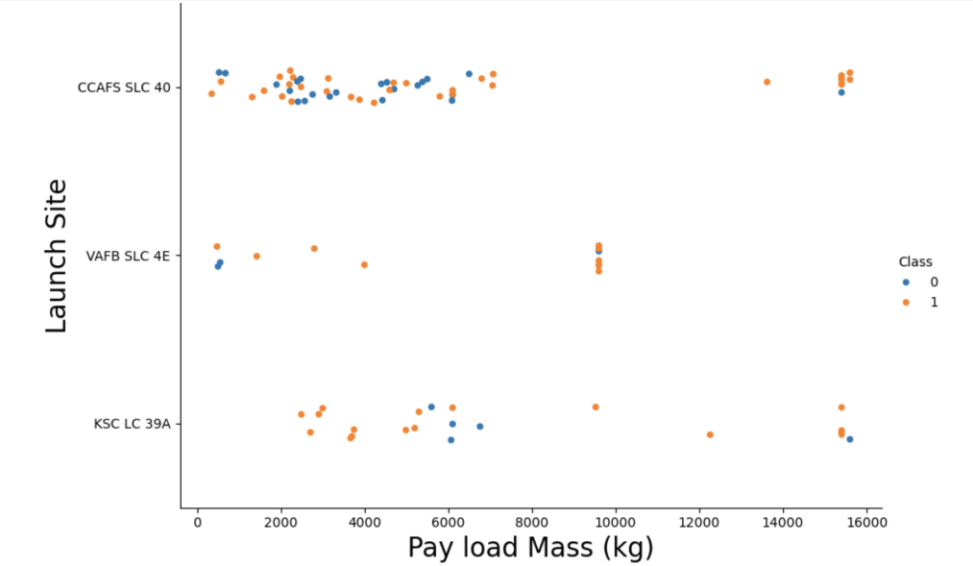
- Git Hub URL - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

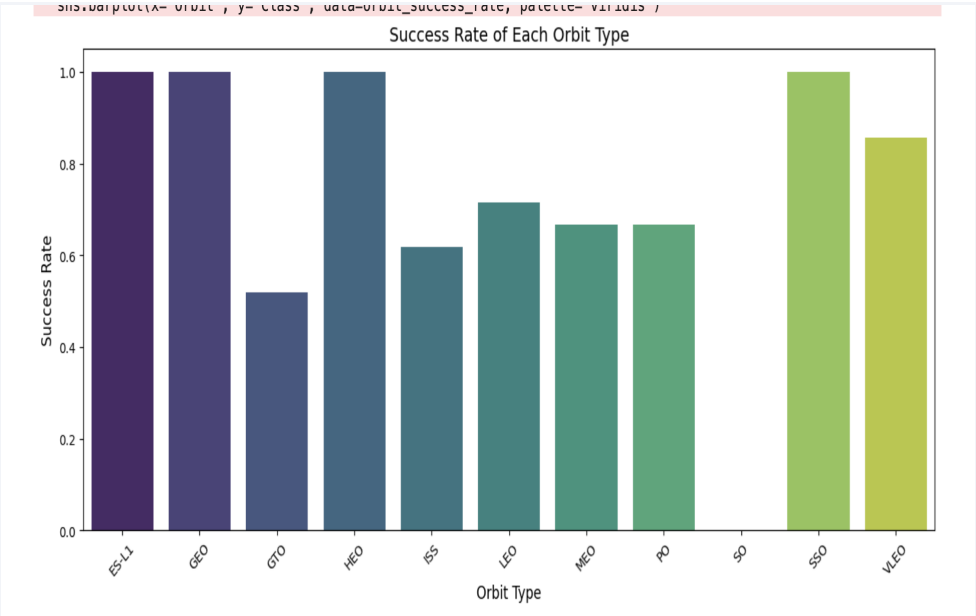
EDA with Data Visualization

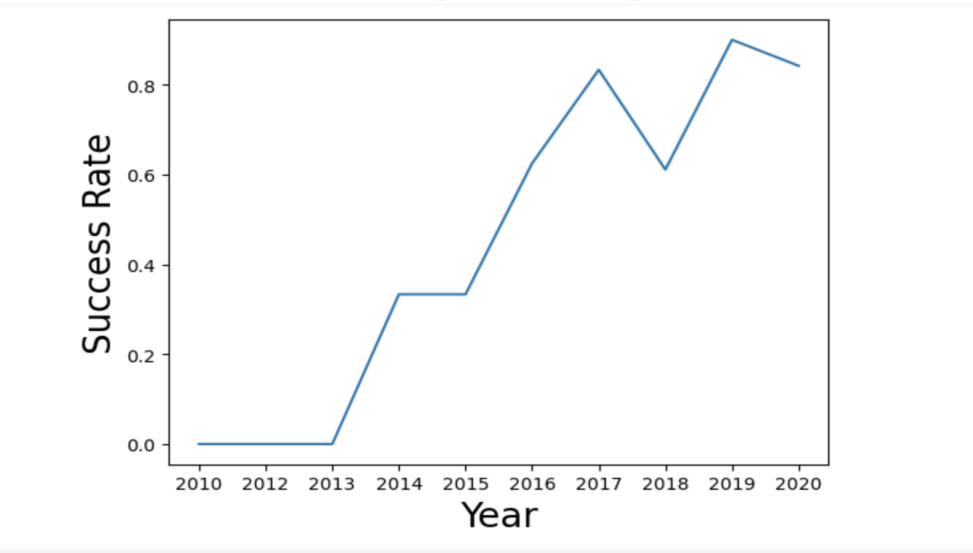
- Performed data Analysis and Feature Engineering using Pandas and Matplotlib.i.e.
  - Exploratory Data Analysis
  - Preparing Data Feature Engineering
- Used scatter plots to Visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.
- Used Bar chart to Visualize the relationship between success rate of each orbit type
- Line plot to Visualize the launch success yearly trend.
- Git Hub - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/Data%20Visulization%20using%20pandas%20and%20matplotlib.ipynb

EDA with Data Visualization (Plots Cont.…)









Interactive Map Using Folium

- Created folium map to marked all the launch sites, and created map objects such as markers, circles, lines to mark the success or failure of launches for each launch site.
- Created a launch set outcomes (failure=0 or success=1).
- Marked the distance between coast and Launch site.
- Marked the distance between Railways and Launch site.
- Marked the distance between City and Launch site.
- Git Hub - https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/lab_jupyter_launch_site_location.ipynb

Dashboard using Plotly Dash

- Built an interactive dashboard application with Plotly dash by:

  -Adding a Launch Site Drop-down Input Component

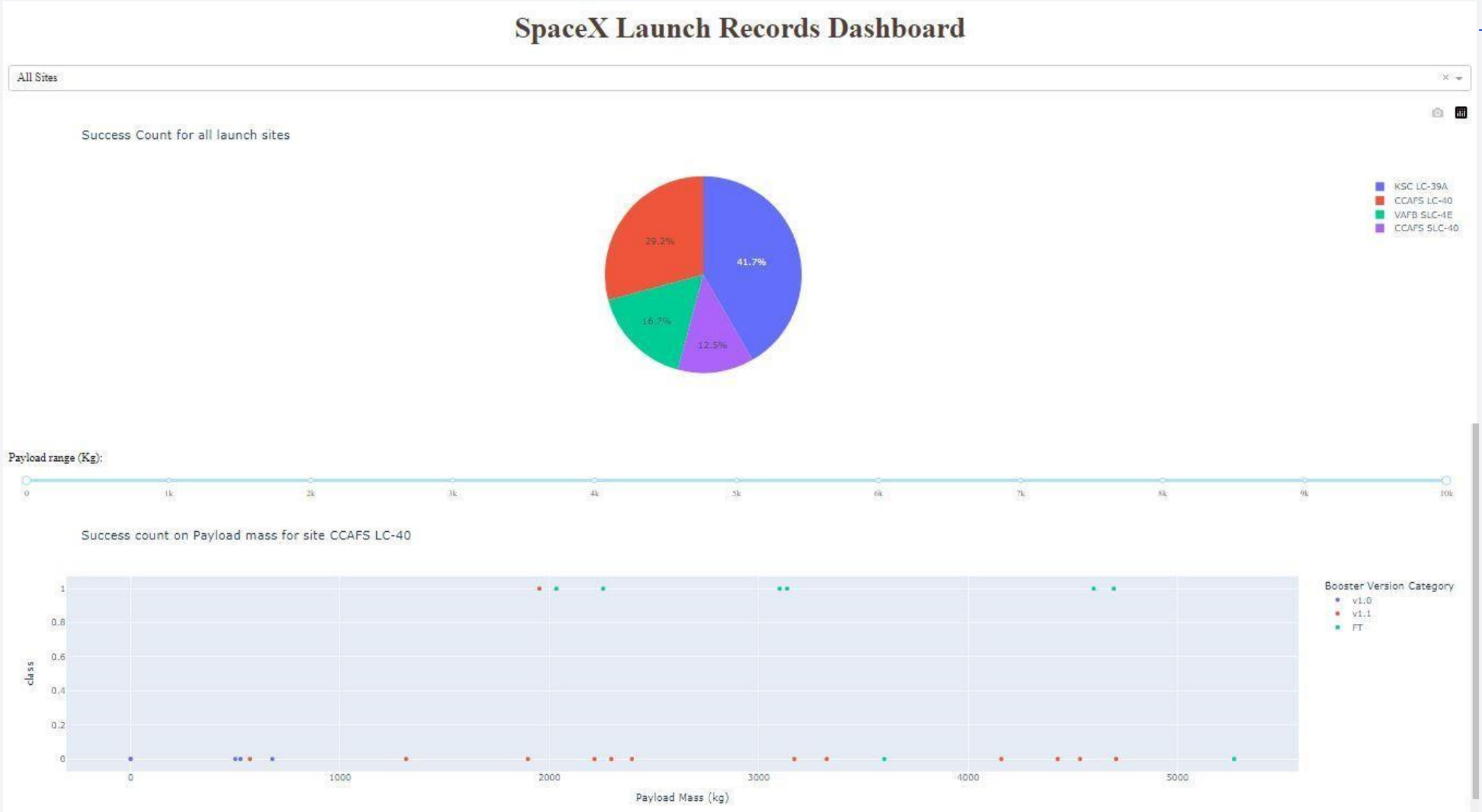
  -Adding a callback function to render success-pie-chart based on selected site dropdown

  -Adding a Range Slider to Select Payload

  -Adding a callback function to render the success-payload-scatter-chart scatter plot

  - Git Hub Url- https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/spacex_dash_app.py

SpaceX Dash App



Predictive Analysis (Classification)

- Summary of Building, Evaluating, Improving, and Identifying the Best Performing Classification Model
- After loading the data into a Pandas Data Frame, I began by performing exploratory data analysis and determining the training labels. This process involved:
  - Creating a NumPy array from the "Class" column in the dataset by applying the to_numpy() method and assigning it to the variable Y as the outcome variable.
  - Standardizing the feature dataset (X) using the preprocessing. StandardScaler() function from Scikit-learn to ensure the features have a mean of 0 and a standard deviation of 1.
  - Splitting the data into training and testing sets using the train_test_split function from sklearn.model_selection, with the test_size parameter set to 0.2 and random_state set to 2.

Predictive Analysis (Classification)

- In order to find the best ML model/ method, I have used various classification models such as SVM, Classification Trees, k nearest neighbors and Logistic Regression.

- GithubURL-https://github.com/bhavi40/SpaceX-Falcon-9/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Predictive Analysis (Classification)

- The table below shows the test data accuracy score for each of the methods comparing them to show which performed best using the test data between SVM, Classification Trees, k nearest neighbors and Logistic Regression;

Out[68]:

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

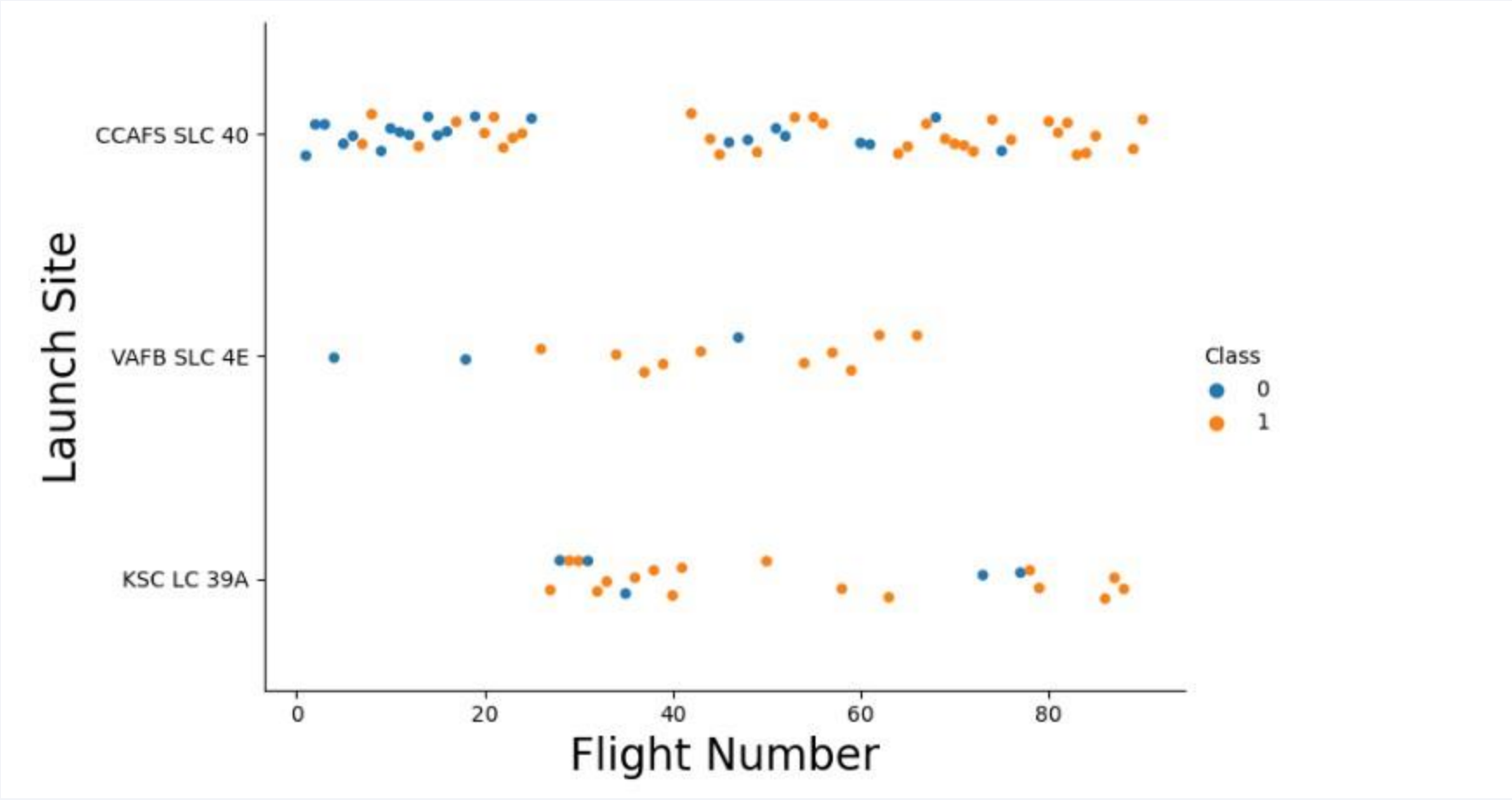
Results

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

Insights Drawn From EDA

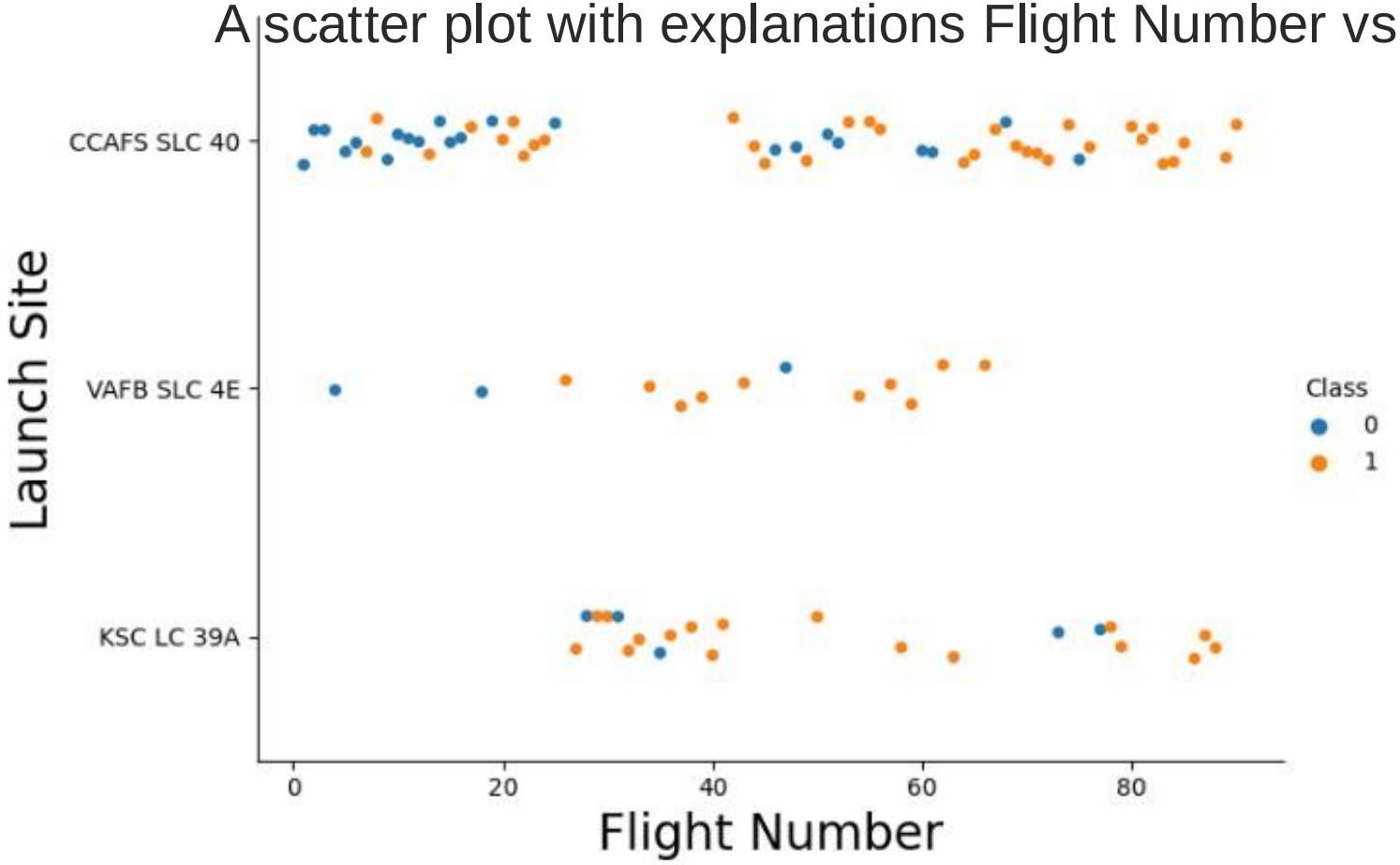
Flight Number vs. Launch Site

A scatter plot of Flight Number vs. Launch Site



Flight Number vs. Launch Site with explanations

A scatter plot with explanations Flight Number vs. Launch Site

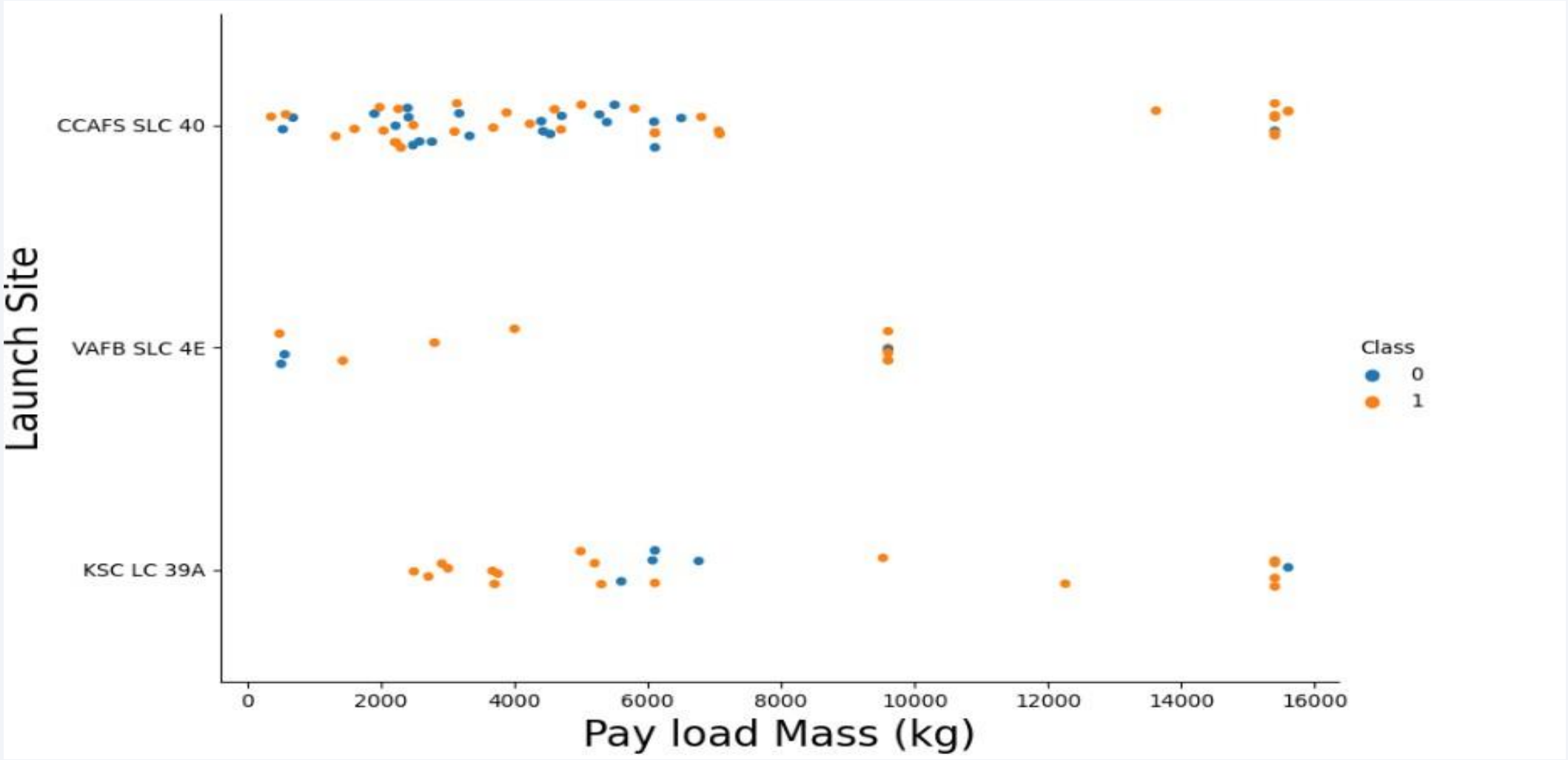


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

We can deduce that, as the flight number increases in each of the 3 launcg sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight.

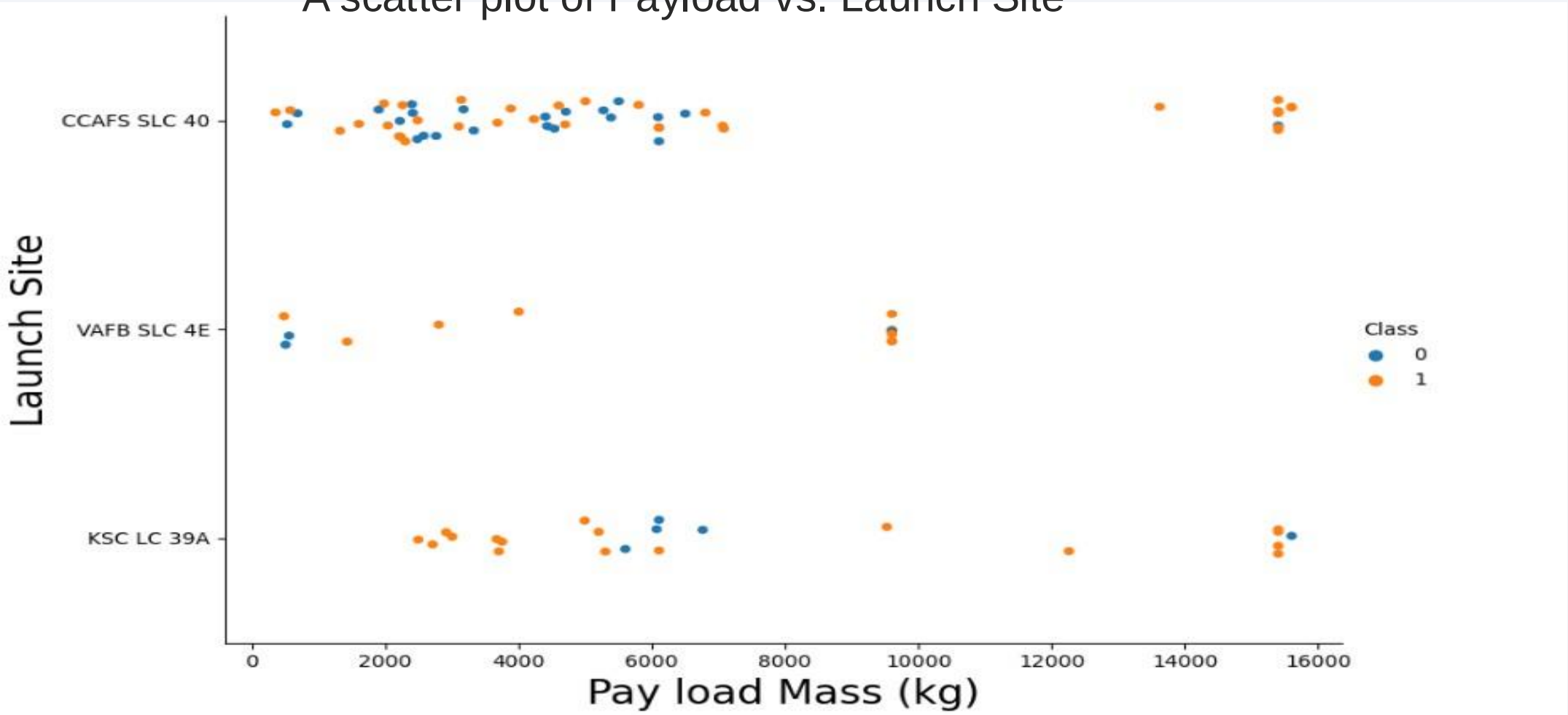
Payload vs. Launch Site

A scatter plot of Payload vs. Launch Site



Payload vs. Launch Site with explanations

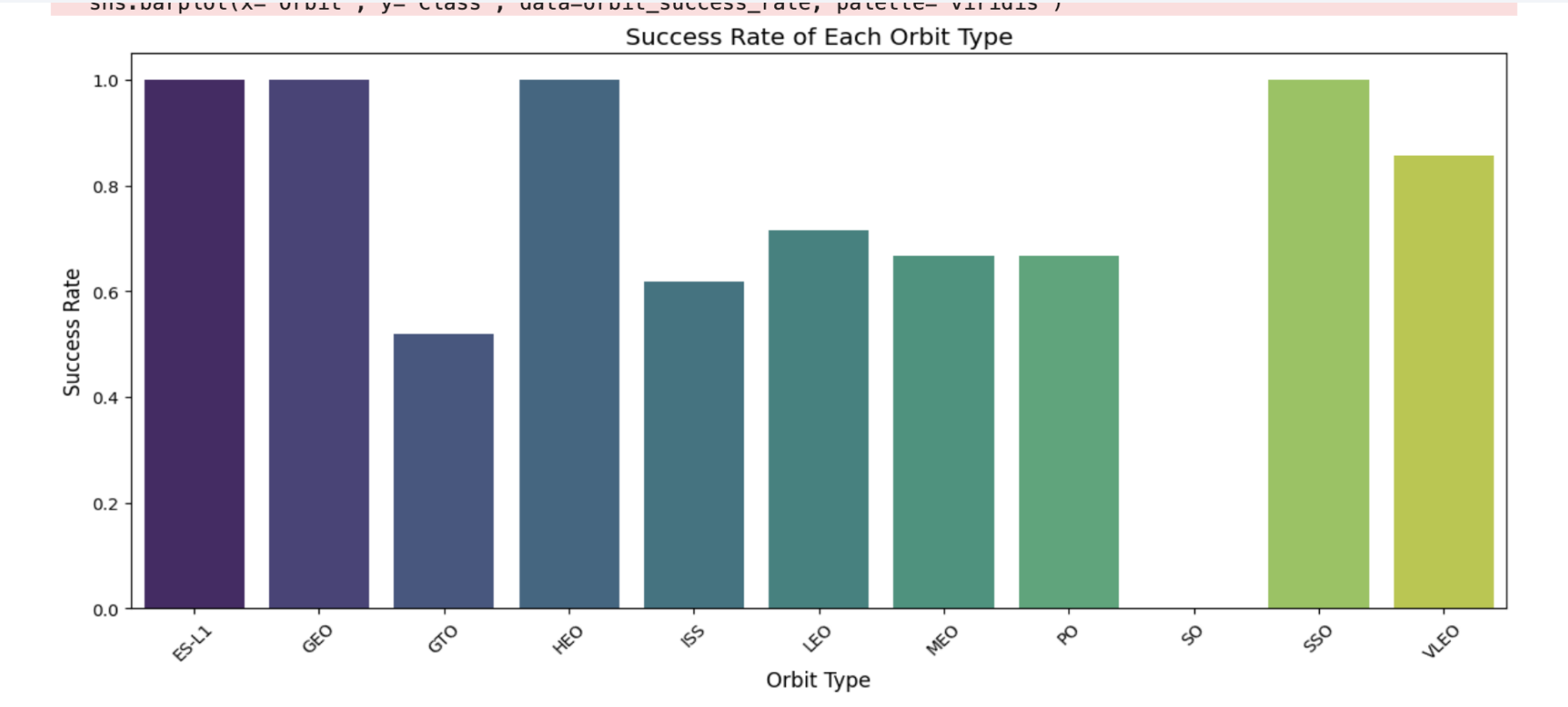
A scatter plot of Payload vs. Launch Site



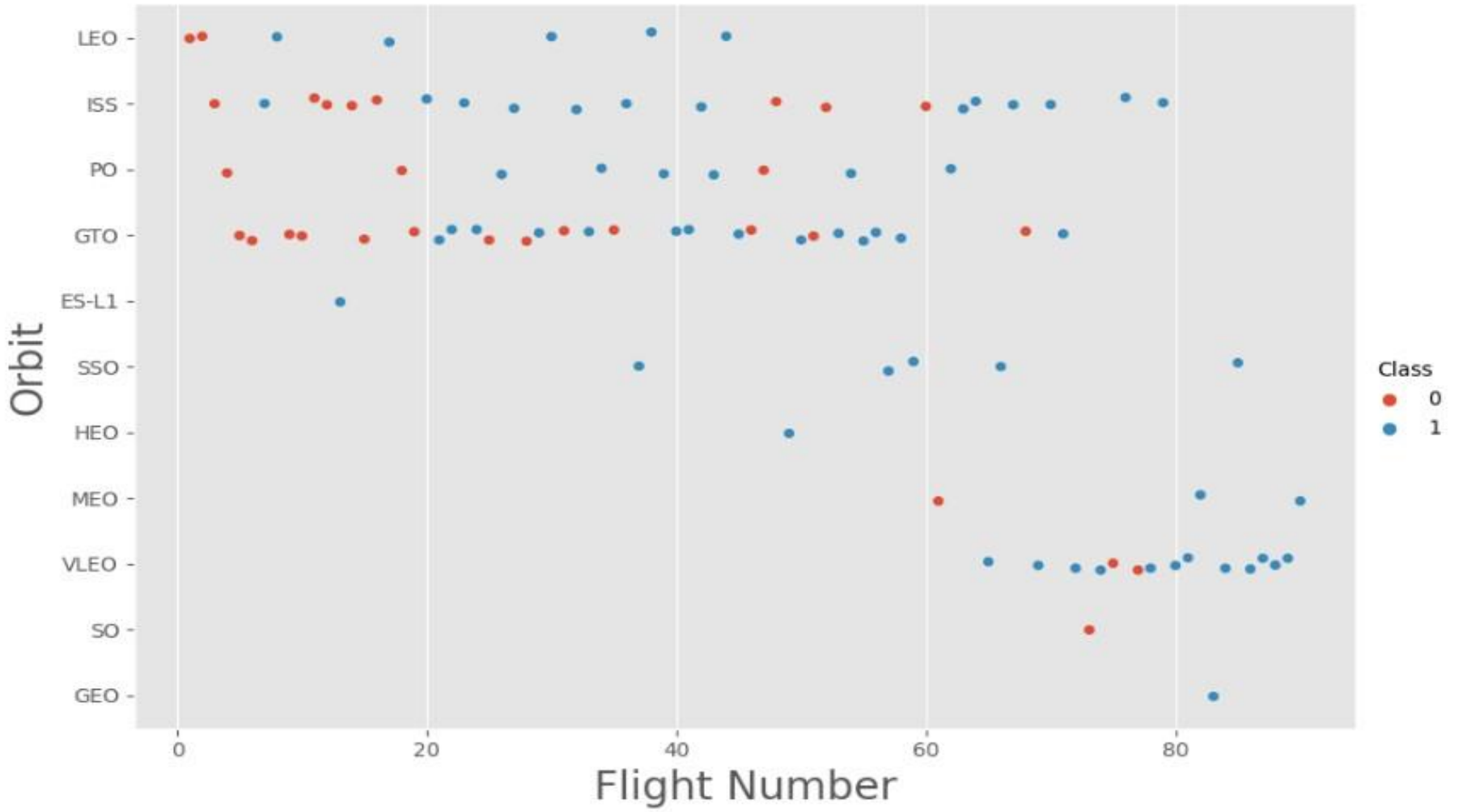
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type



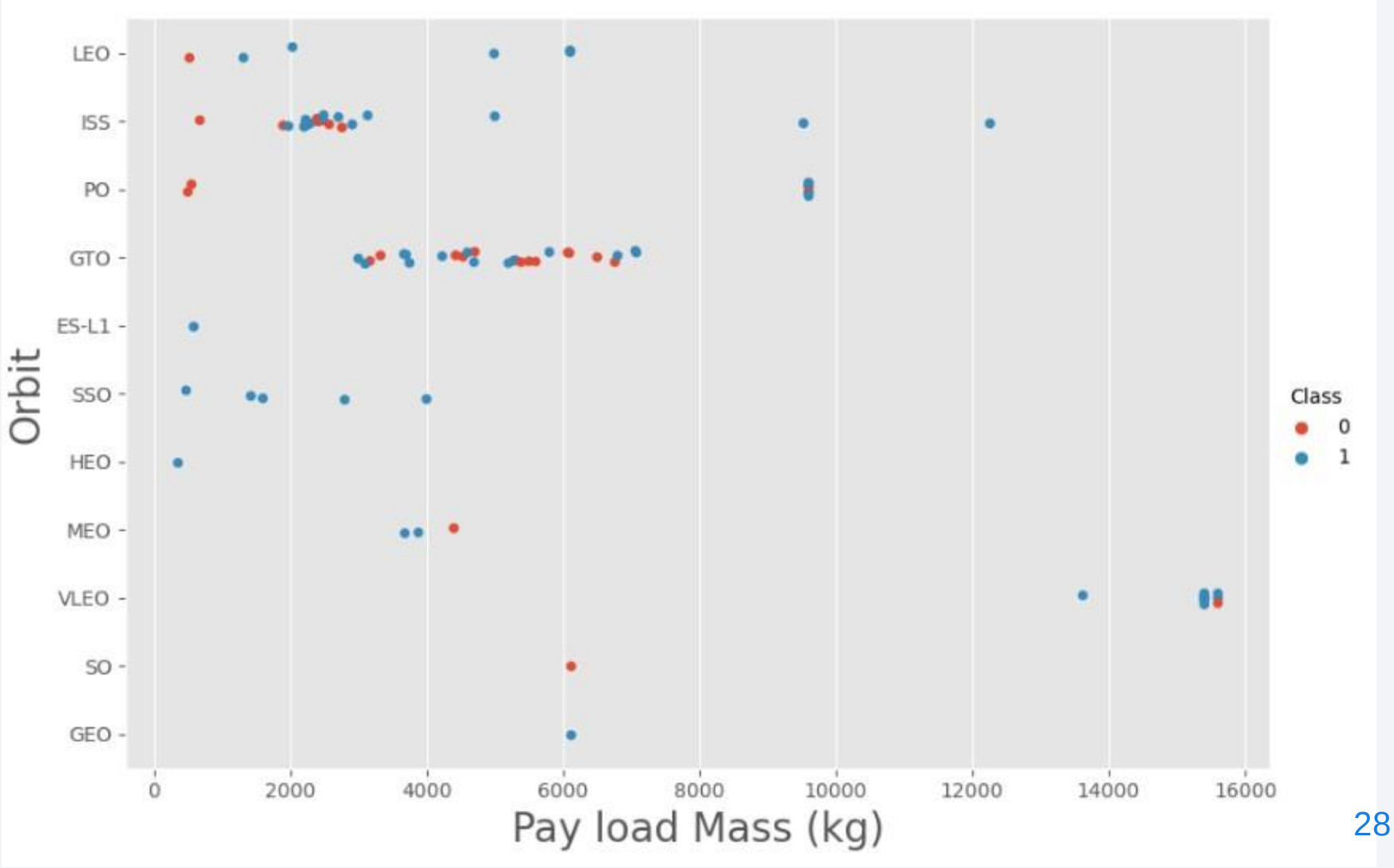
Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

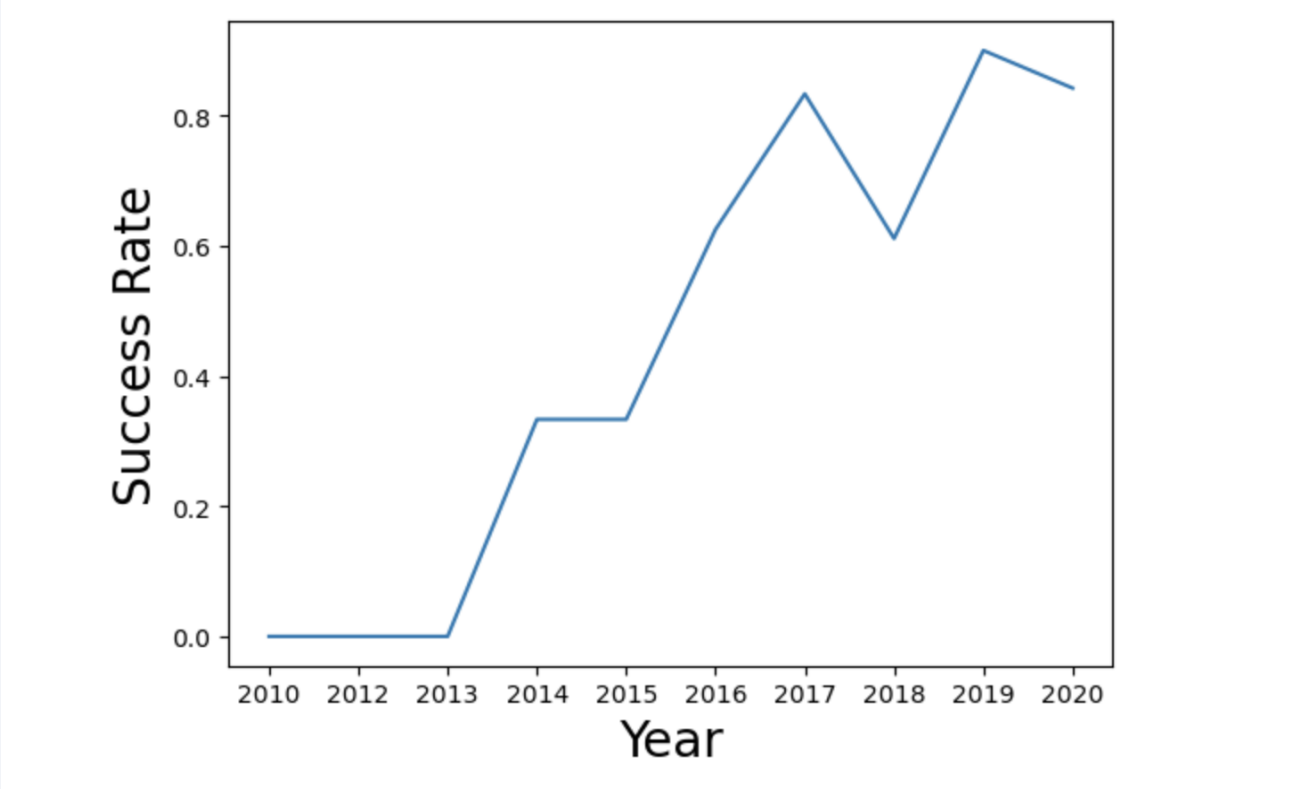
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) both have near equal chances.



Launch Success Yearly Trend

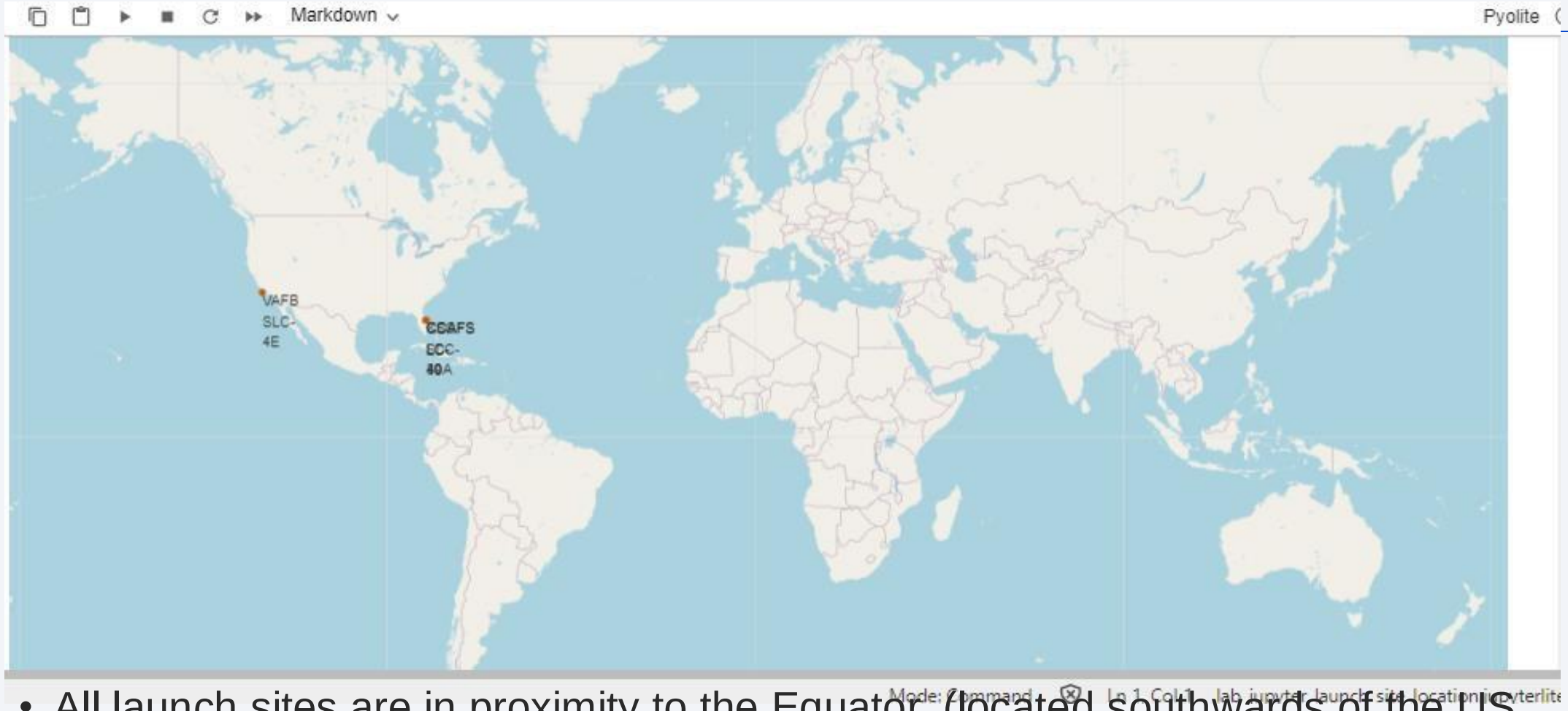
success rate kept going up till 2020



Section 3

Launch Sites Proximities Analysis

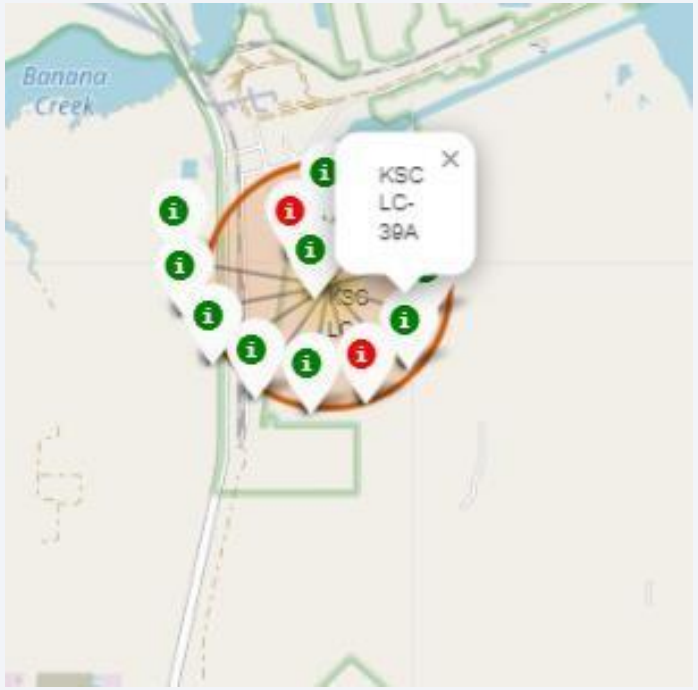
Markers of all launch sites on global map



- All launch sites are in proximity to the Equator, (located southwards of the US map). Also all the laumch sites are in very close proximity to the coast.

Launch outcomes for each site on the map With Color Markers

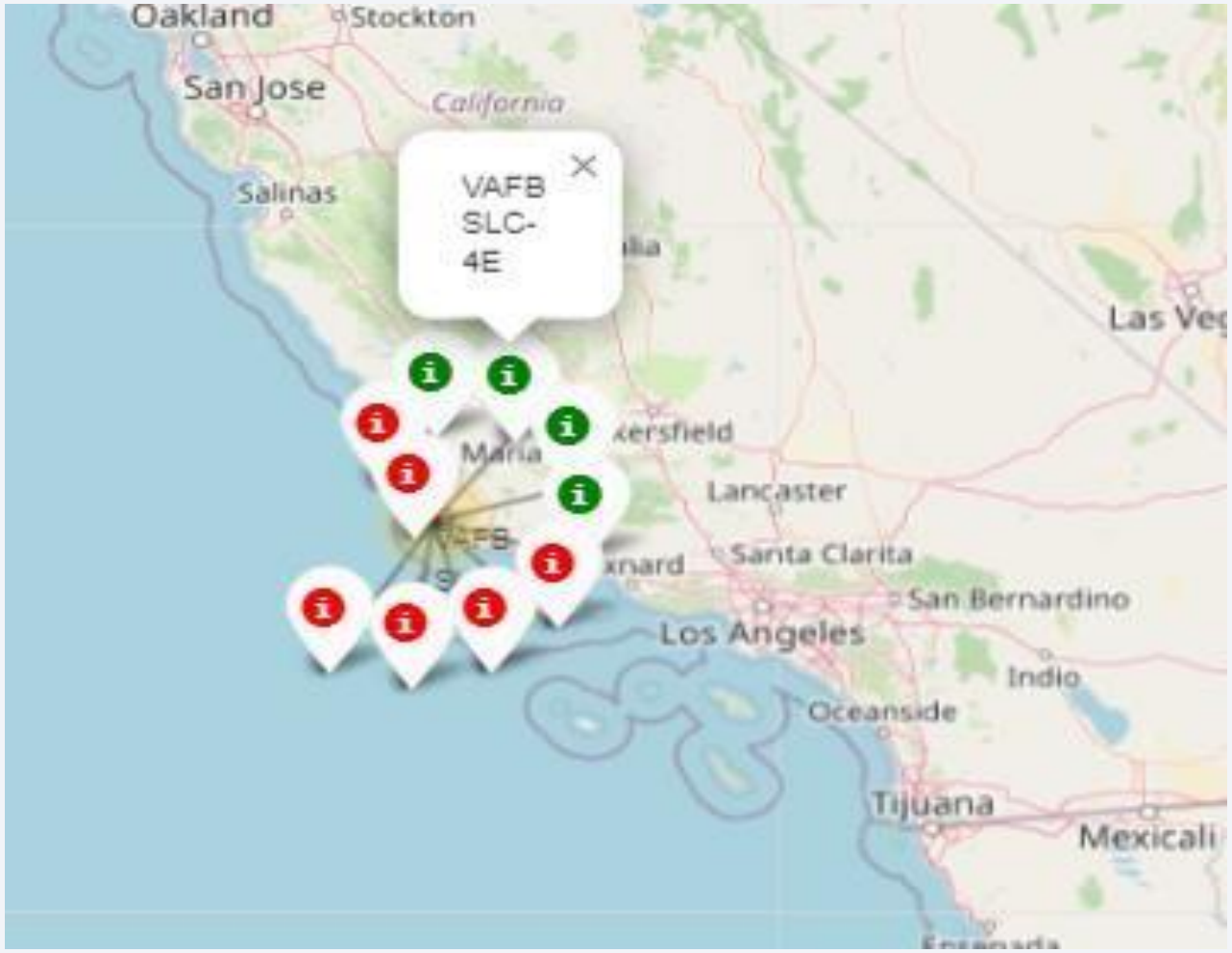
Florida Sites



- In the Eastern coast (Florida) Launch site KSC LC-39A has relatively high success rates compared to CCAFS SLC-40 & CCAFS LC-40.

Launch outcomes for each site on the map With Color Markers

West Coast/ Carlifonia

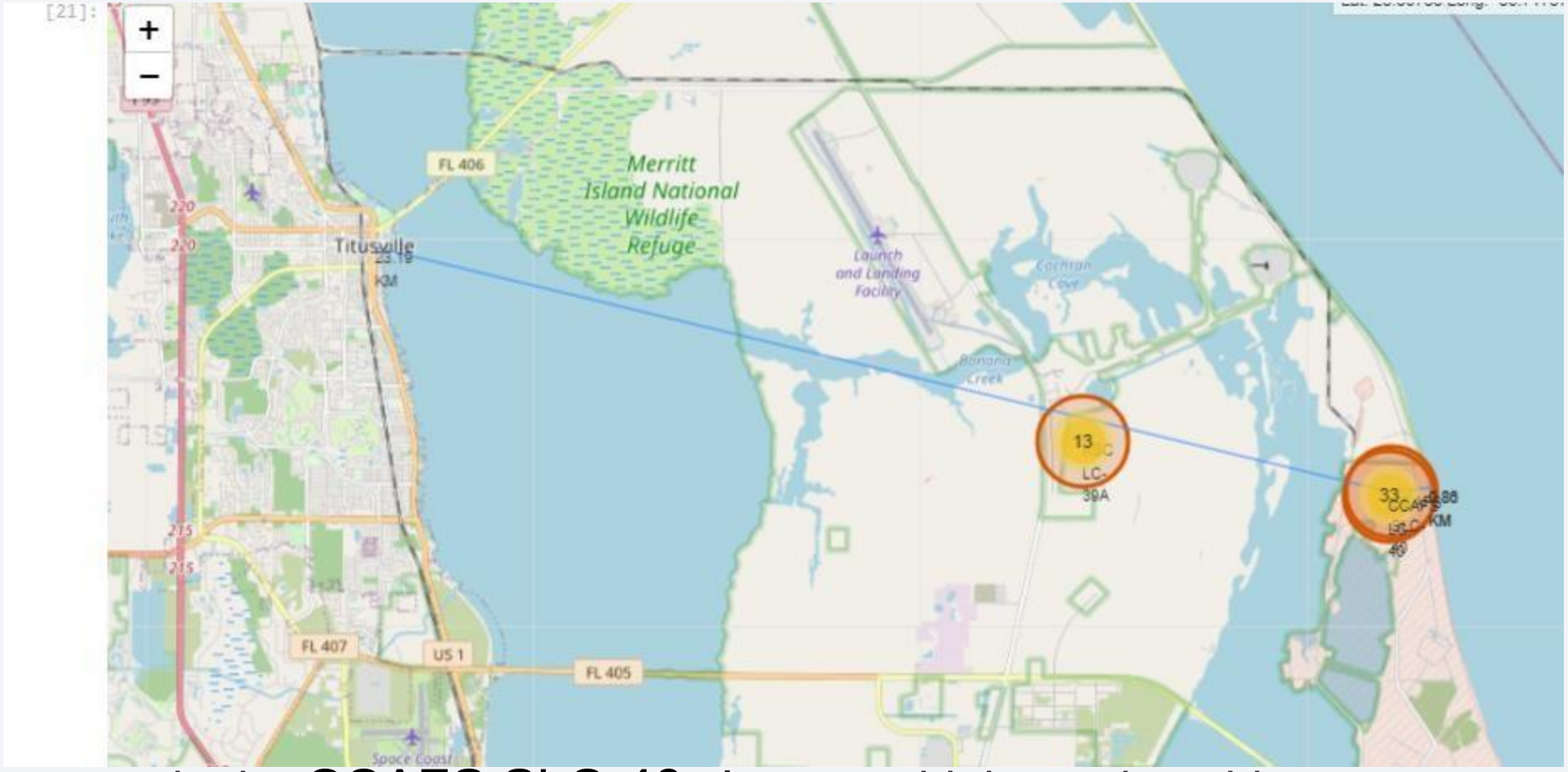


- In the West Coast (Californai) Launch site VAFB SLC-4E has relatively lower success rates 4/10 compared to KSC LC-39A launch site in the Eastern Coast of Florida.

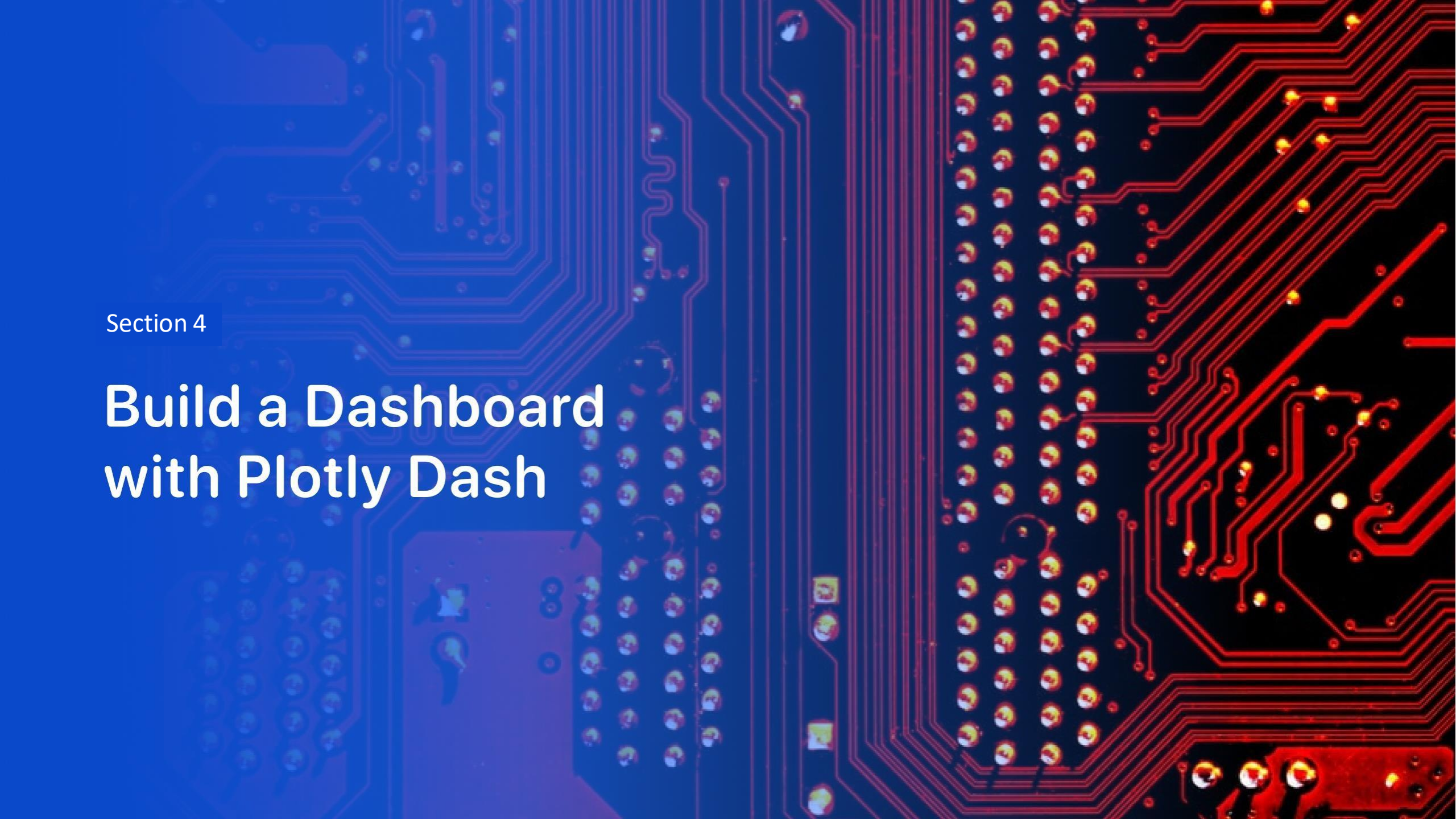
Distances between a launch site to its proximities

- Launch site CCAFS SLC-40 proximity to coastline is 0.86km

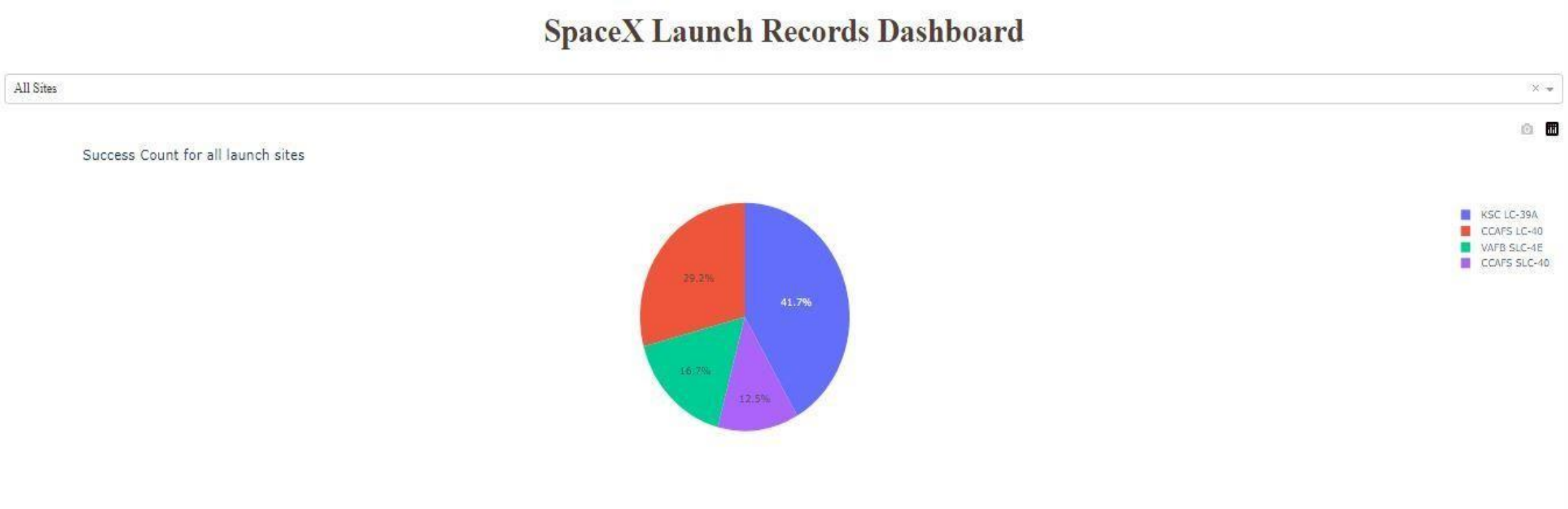
Distances between a launch site to its proximities



- Launch site CCAFS SLC-40 closest to highway (Washington Avenue) is 23.19km

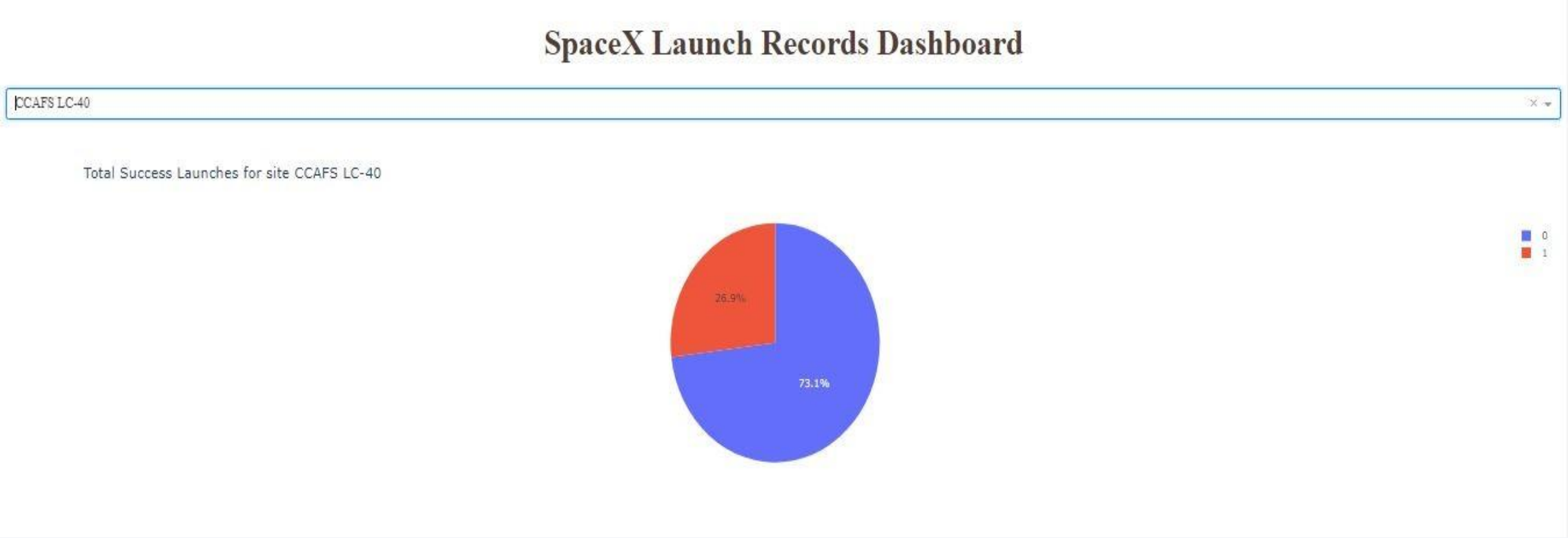

Section 4
Build a Dashboard with Plotly Dash

Pie-Chart for launch success count for all sites



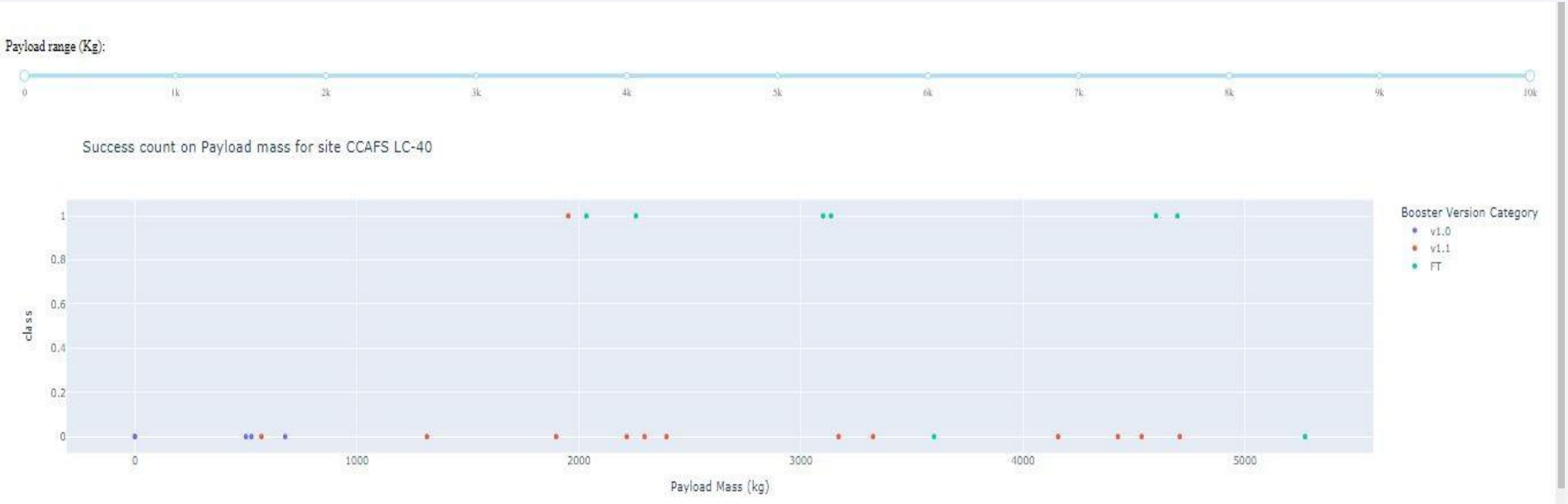
- Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%

Pie chart for the launch site with 2nd highest launch success ratio



- Launch site CCAFS LC-40 had the 2nd highest success ratio of 73% success against 27% failed launches

Payload vs. Launch Outcome scatter plot for all sites



- For Launch site CCAFS LC-40 the booster version FT has the largest success rate from a payload mass of >2000kg

Predictive Analysis

Classification Models Accuracy

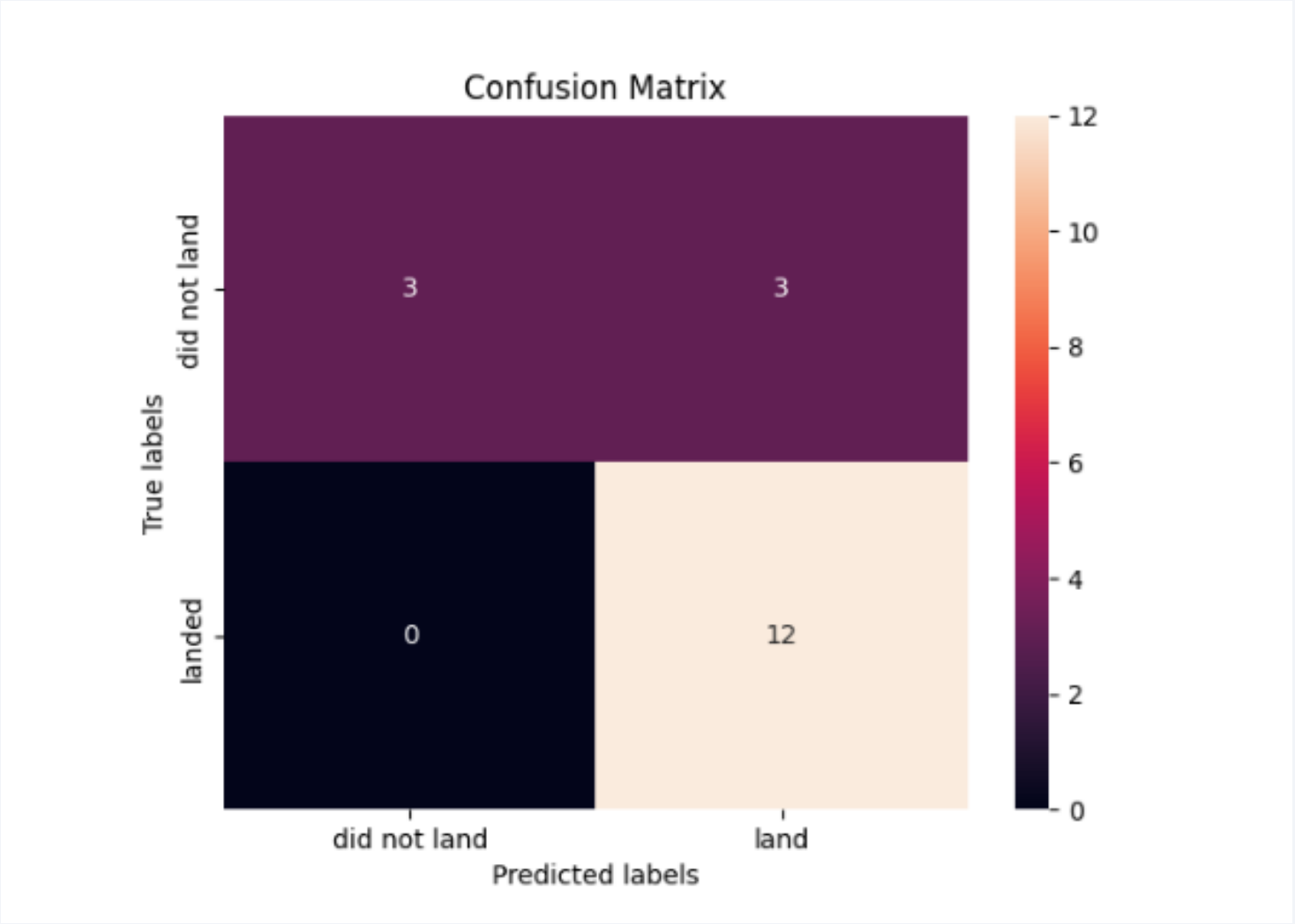
Out[68]:

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

All the methods perform equally on the test data: i.e. They all have the same accuracy of 0.833333 on the test Data

Confusion Matrix

- All the 4 classification model had the same confusion matrixes and were able equally distinguish between the different classes. The major problem is false positives for all the models.



Conclusions

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- We can deduce that, as the flight number increases in each of the 3 launcg sites, so does the success rate. For instance, the success rate for the VAFB SLC 4E launch site is 100% after the Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rates after 80th flight
- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Conclusions Cont….

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here
- Anf finally the sucess rate since 2013 kept increasing till 2020.


Thank You