IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Oleg Kontchaev
May 2, 2022

Outline

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

Executive Summary

- Summary of methodologies

Data Collection with API and WebScraping

Data Wrangling

Exploratory Data Analysis with SQL and Pandas

Visualization of Exploratory Data Analysis with Matplotlib

Interactive Visual Analytics and Dashboard

Predictive Analysis with Machine Learning (Classification)

- Summary of all results

Exploratory Data Analysis results

Predictive Analysis results

Introduction

- Project background and context

SpaceX is able to launch "Falcon 9" rockets with a cost of 62 million dollars while its competitors launches for 165 million dollars. A great deal of this difference is SpaceX can reuse the first stage of the launch. For this reason, if we can determine the result of landing status of first stage, we can determine the cost of a launch as well.

- Problems you want to find answers

In this project , our objective is to train machine learning models with publicly available data so that we can determine whether the first stage will be reusable or not. Thus estimate the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected by using 2 methods: SpaceX API and web scraping(Wikipedia)

- Perform data wrangling

  One hot encoding and dropping of irrelevant columns was performed

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

  The data was split into training and test sets. Four different classification models were built. Then, the accuracy of each model was evaluated.

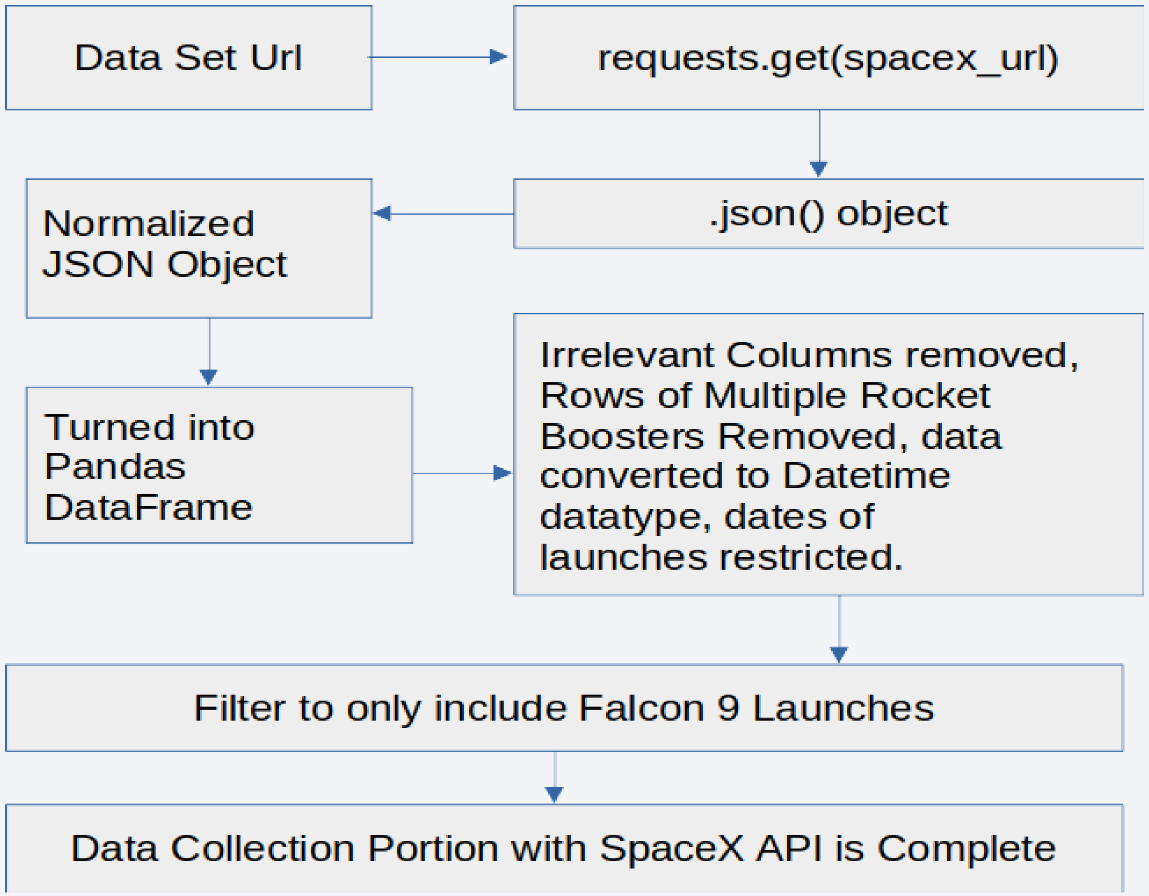
Data Collection

- **The data collection included the following steps:**
- Get request to the SpaceX API
- Decoded the response content using .json() function call and turn it into a pandas dataframe with
- .json_normalize()
- Cleaned the data of missing values
- Web scraping from Wikipedia for Falcon 9 launch data using BeautifulSoup
- Take the HTML table and convert it to a pandas dataframe for analysis

Data Collection – SpaceX API

- Key phrases and flowchart is to the right

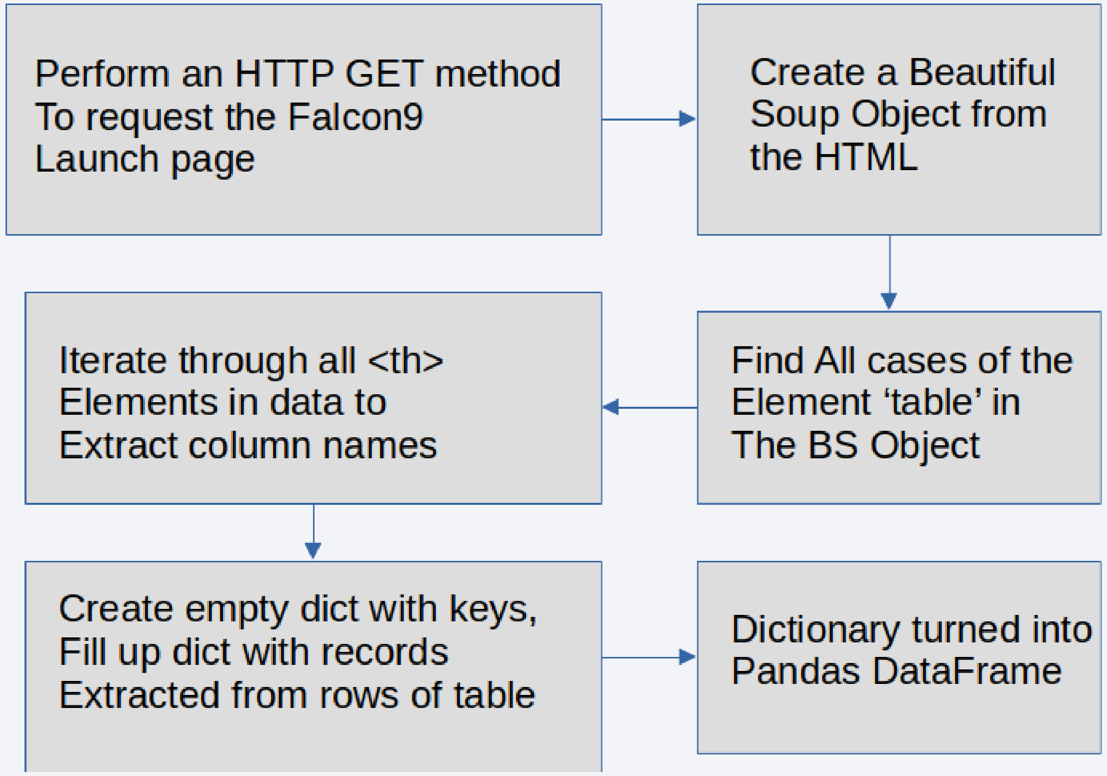
- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/c273fa6720a0cf16ee206d6274d85edb14eb2acb/jupyter-labs-spacex-data-collection-api%20(2).ipynb



Data Collection - Scraping

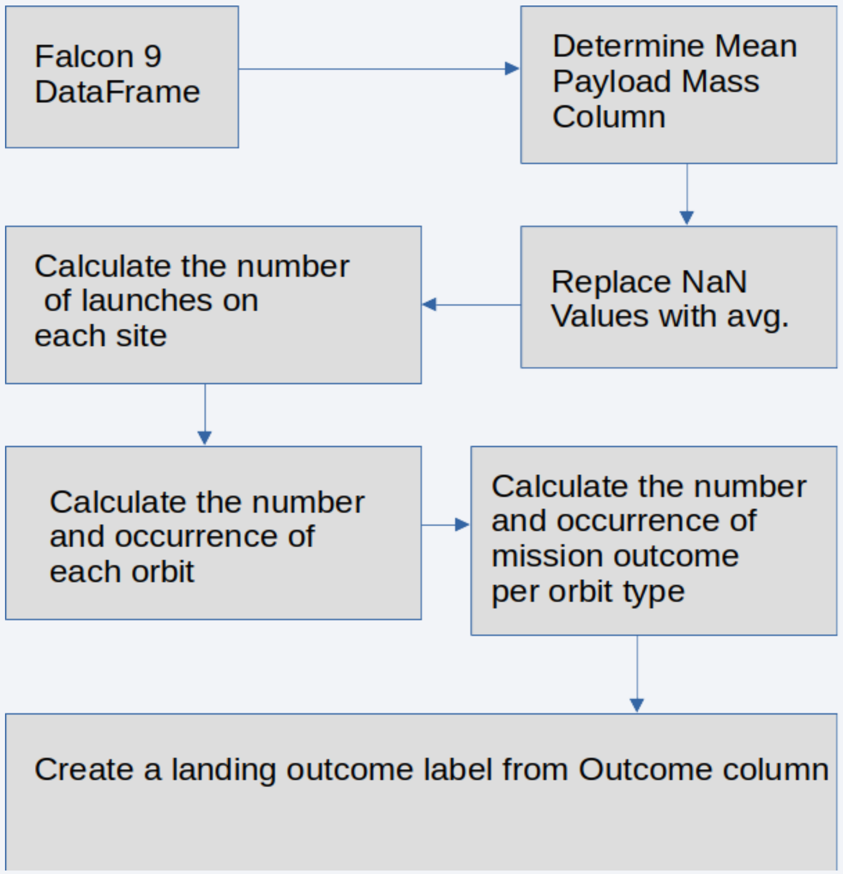
- Webscraping flowchart is to the right

- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- Performed Exploratory Data Analysis, determined the training labels, summarized the number of launches at each site, and occurrence of each orbit, then created landing outcome labels from outcome column
- Data Wrangling flowchart is to the right
- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



EDA with Data Visualization

- Graphs being drawn:

Scatter Graphs: Flight Number VS. Payload Mass, Flight Number VS. Launch Site, Payload VS. Launch Site, Orbit VS. Flight Number, Payload VS. Orbit Type, Orbit VS. Payload Mass

Bar Graph: Mean vs. Orbit

Line Graph: Success Rate vs. Year

- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

- SQL queries results
  - Displaying unique launch sites in the space mission
  - Displaying initial 5 records in the data where the launch sites begin with string ‘CCA’
  - Calculating total booster payload launched by NASA (CRS)
  - Showing the average payload mass carried by booster F9 v1.1
  - Finding the date of the first successful ground pad landing outcome
  - Listing booster names with payload between 4000 –6000 kg and landing outcome a success in drone ships
  - Displaying total number of successful and failed rocket launch outcomes
  - Listing booster names that carried highest payload mass
  - Displaying launch site names and booster versions that produced fail drone ship landing outcomes in 2015
  - Listing count of each landing outcome (e.g., Failure (drone ship)) in descending order between dates 4/6/2010 and 20/3/2017
- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql(1).ipynb

Build an Interactive Map with Folium

- Map objects created and added to a folium map
  site_map.add_child(circle) was used to add circles
  site_map.add_child(marker) was used to add markers
  site_map.add_child(lines) was used to add lines
- Those objects were added to pinpoint locations of launch sites, calculate distances, mark the outcome of launches for each site on the folium map
- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:

A pie chart and a scatter chart were added

- Charts were added to show:

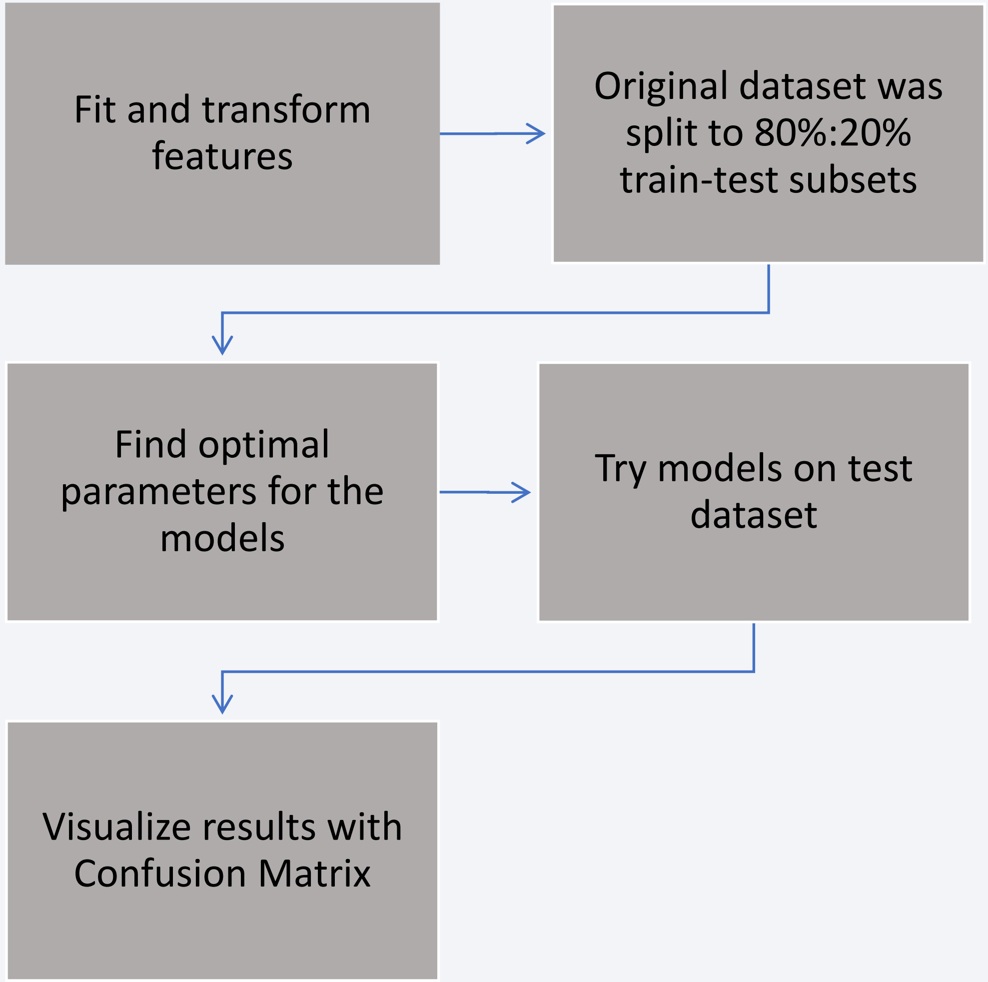
pie chart - the total successful launches count for all sites

scatter chart - the correlation between payload and launch success

- The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/labs%20spacex%20project%20-%20dashboard%20-%20Plotly%20Dash.py

Predictive Analysis (Classification)

The data were loaded, transformed , and split into training and testing sets. Then different machine learning models were built and tuned. The accuracy was used as the metric for all models. Found optimal parameters for the models. Finally, the best performing model was determined.



The link to GitHub: https://github.com/okontchaev/Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

- Exploratory data analysis results:
    - CCAFS LC-40, has a success rate of 60 %, KSC LC-39A and VAFB SLC 4E have a success rate of 77%
    - ES-L1, SSO, HEO, and GEO has the highest success rate of 1.0 = 100%
    - With Geo and MEO, greater success over time. Others see no difference over time
    - Polar, LEO, and ISS Orbits, when launching heavy payloads, have a greater success rate

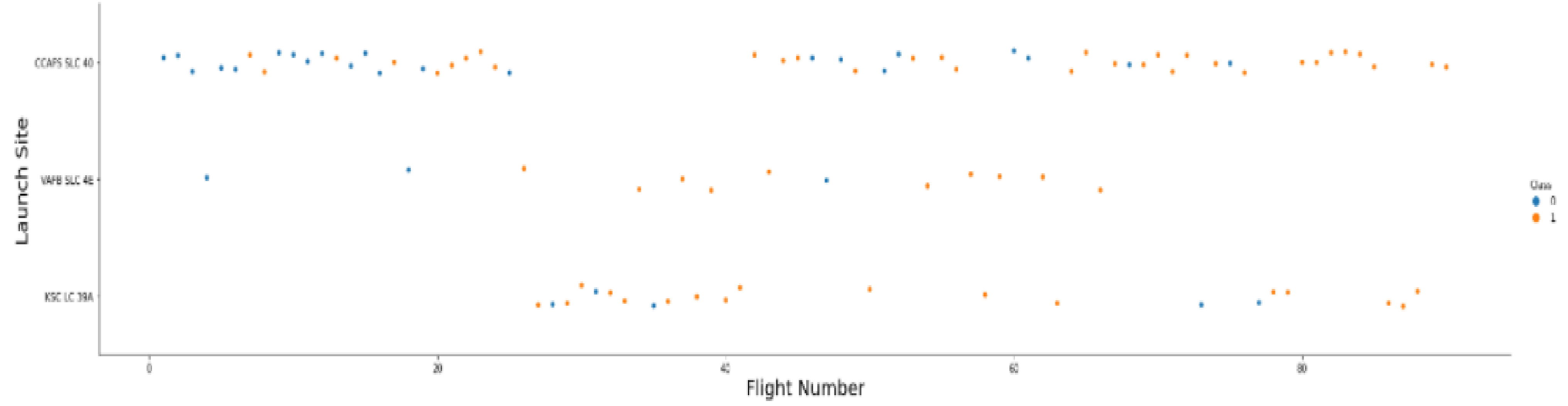
- Predictive analysis results: Decision Tree Classifier had the highest accuracy of 87.5%

Section 2

Insights drawn from EDA

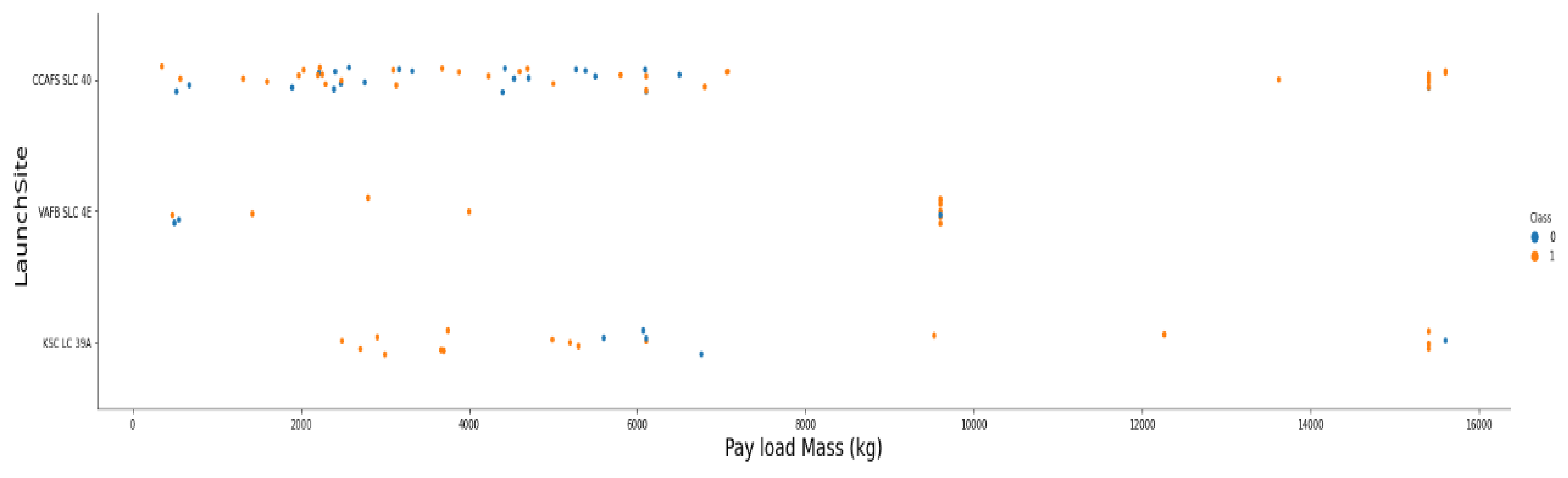
Flight Number vs. Launch Site

The chart shows a distribution of flights among three different launch sites. Color shows success vs failure. As we can see, the failure percentage is declining over time(Flight Number) for all sites



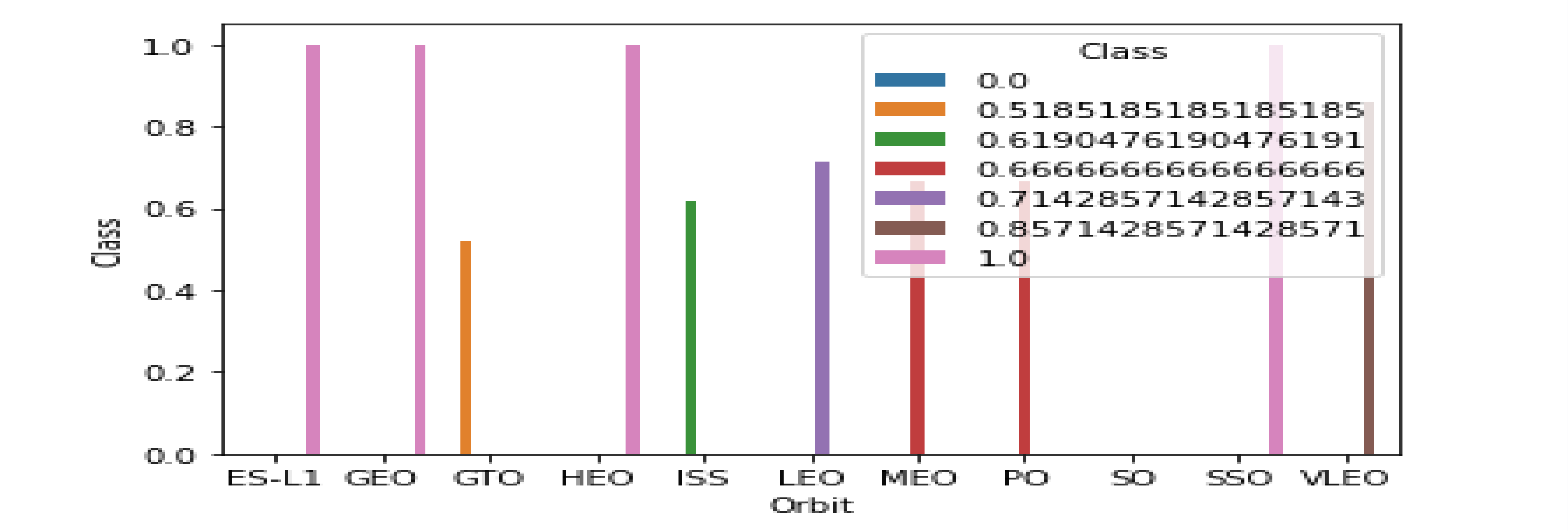
Payload vs. Launch Site

As we can see on Pay load vs Launch Site chart, the greater mass correlates with a success of the flight



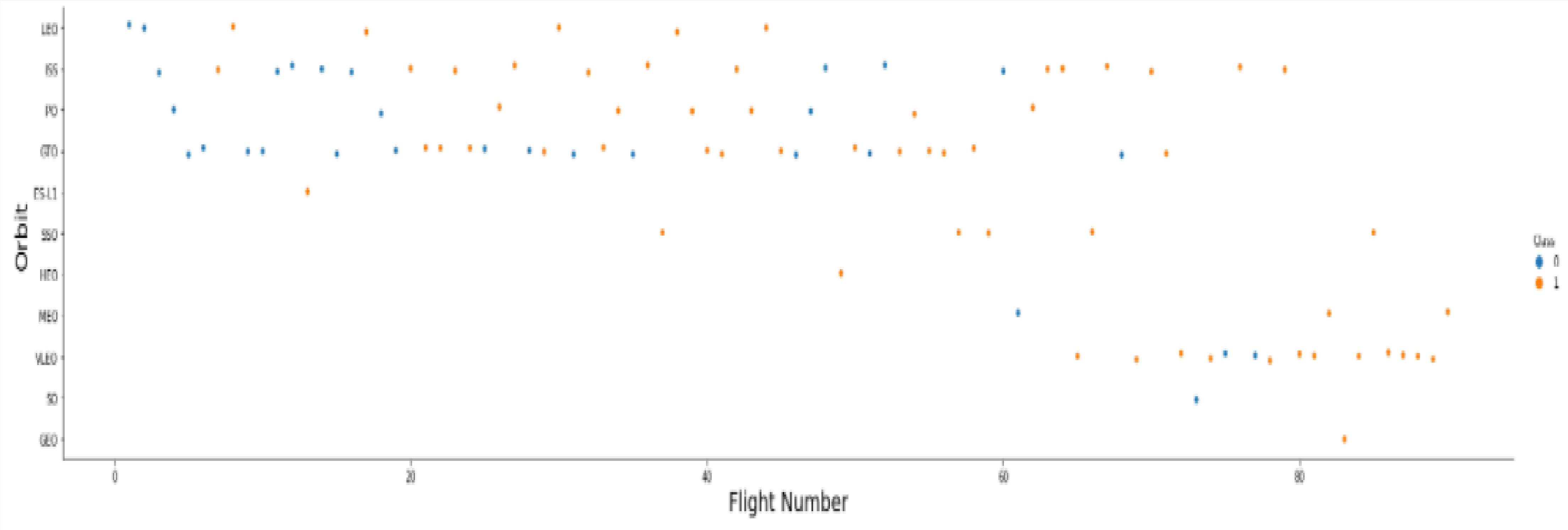
Success Rate vs. Orbit Type

From this plot, we can see that ES-L1, GEO, HEO, SSO have the most success rate



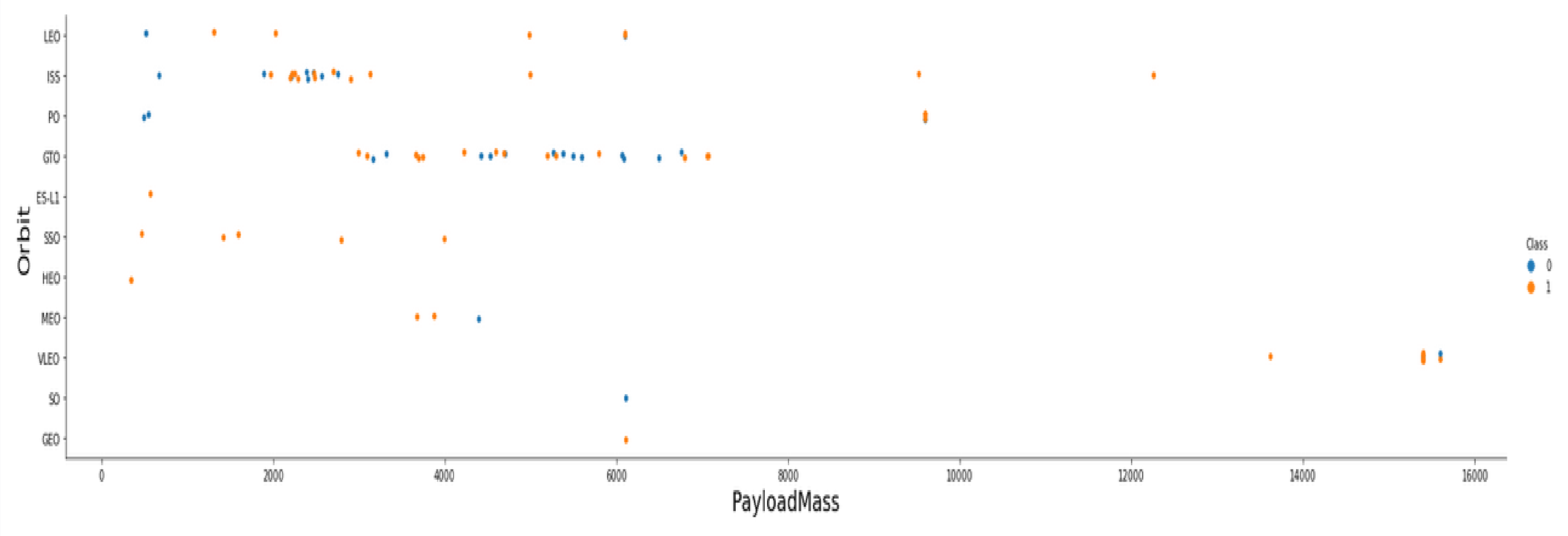
Flight Number vs. Orbit Type

This Flight Number vs. Orbit type plot shows that SSO is “all the time” successful orbit, for most of the other orbits, success rate improves over time(Flight Number)



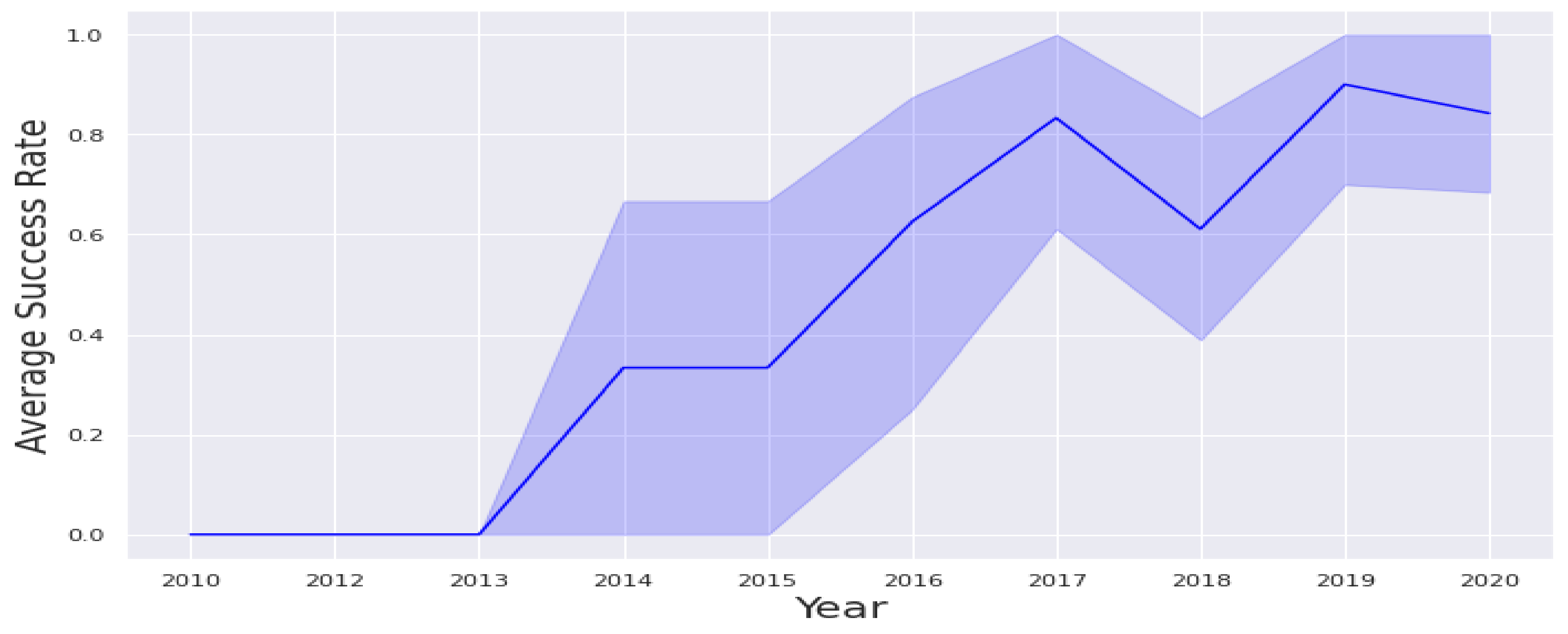
Payload vs. Orbit Type

Among the busiest launch sites, heavy load improves success percentage for LEO and SSO orbits and does not have any effect on SSO(100% success) and GTO(50/50 success/failure) orbits



Launch Success Yearly Trend

The success rate is improving since 2013, as shown on this chart



;

All Launch Site Names

- The names of the unique launch sites: CAAFS LC-40, CAAFS SLC-40, KSC LC-39A, and VAFB SLC-4E
- Query result:

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

- A short explanation:

This query %sql select distinct(LAUNCH_Site) from SPACEXTBL; **returns unique names only**

Launch Site Names Begin with 'CCA'

- Query: **%sql** select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
- Result with a short explanation: **limit** 5 returns only 5 records from **SPACEXTBL like** 'CCA%' implies that the Launch Site Name must start with CCA.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA: 45596
- Query: **%sql** select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)';

- Short explanation here:

Function sum calculates the total in the column PAYLOAD_MASS__KG_,

while where filters final result to apply to customer NASA (CRS) only

%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION LIKE '%F9 v1.1';

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1: 2928.4
- Query: **%sql** select avg(PAYLOAD_MASS_KG_) from SPACEXTBL where BOOSTER_VERSION LIKE '%F9 v1.1';
- Short explanation:

Function avg calculates the average in the column PAYLOAD_MASS_KG_,

while where and LIKE filters final result to apply to booster versions starting with ‘F9 v1.1’ (F9 version

First Successful Ground Landing Date

Query: **%sql SELECT MIN**(date) **AS** first_successful_landing_outcome
FROM SPACEXTBL
WHERE LANDING__OUTCOME **LIKE** '%Success (ground pad)%';

- The result with a short explanation:

2015-12-22

Function MIN calculates the minimum date, while WHERE and LIKE filters final result to apply to successful ground landings only

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: **%sql SELECT** BOOSTER_VERSION, LANDING__OUTCOME, payload_mass__kg_ **FROM** SPACEXTBL **WHERE** LANDING__OUTCOME **LIKE** '%Success (drone ship)%' **AND** payload_mass__kg_ **>** 4000 **AND** payload_mass__kg_ **<** 6000;
- The result with a short explanation:

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

SELECT function gets only the booster versions, WHERE and LIKE applied to successful landing outcomes on drone ships only, AND is used to specify an additional condition(payload being between 4000 kg and 6000 kg range)

Total Number of Successful and Failure Mission Outcomes

- Query: **%sql SELECT COUNT**(MISSION_OUTCOME) **FROM** SPACEXTBL **WHERE** MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = Failure(in flight)';
- The result: 100

Boosters Carried Maximum Payload

- Query: **%sql** select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- The result:

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

Query: **%sql SELECT** LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, **from** SPACEXTBL **WHERE** (LANDING__OUTCOME **=** 'Failure (drone ship)') **AND** (**LIKE** '%2015%');

- The result with a short explanation:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- Only 2 missions failed in 2015.

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

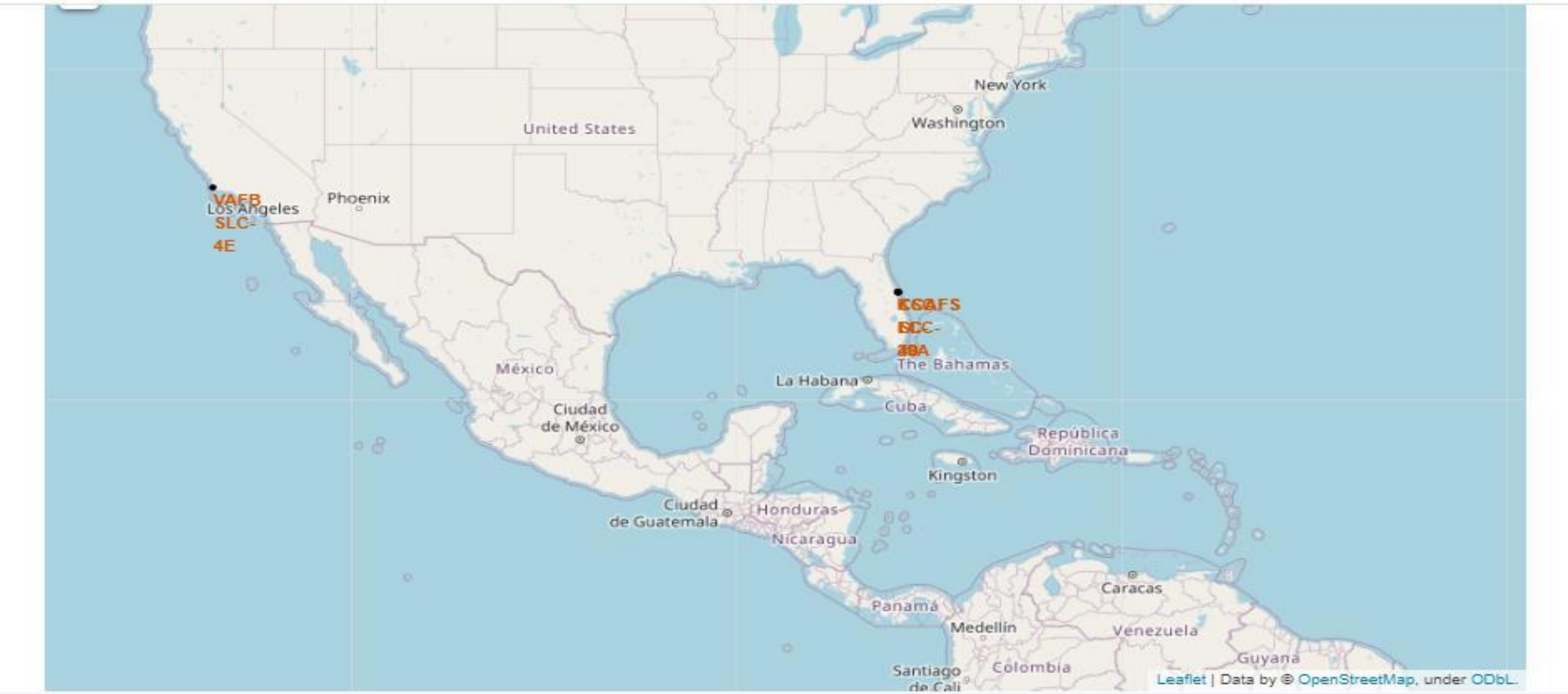
- Query: **%sql** SELECT Landing_Outcome, Count(*) AS Count_Outcomes FROM SPACEXTBL
- WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count_Outcomes DESC;
- The result:

number	landing__outcome	ranking
10	No attempt	1
5	Failure (drone ship)	2
5	Success (drone ship)	2
3	Controlled (ocean)	4
3	Success (ground pad)	4
2	Failure (parachute)	6
2	Uncontrolled (ocean)	6
1	Precluded (drone ship)	8

Section 3

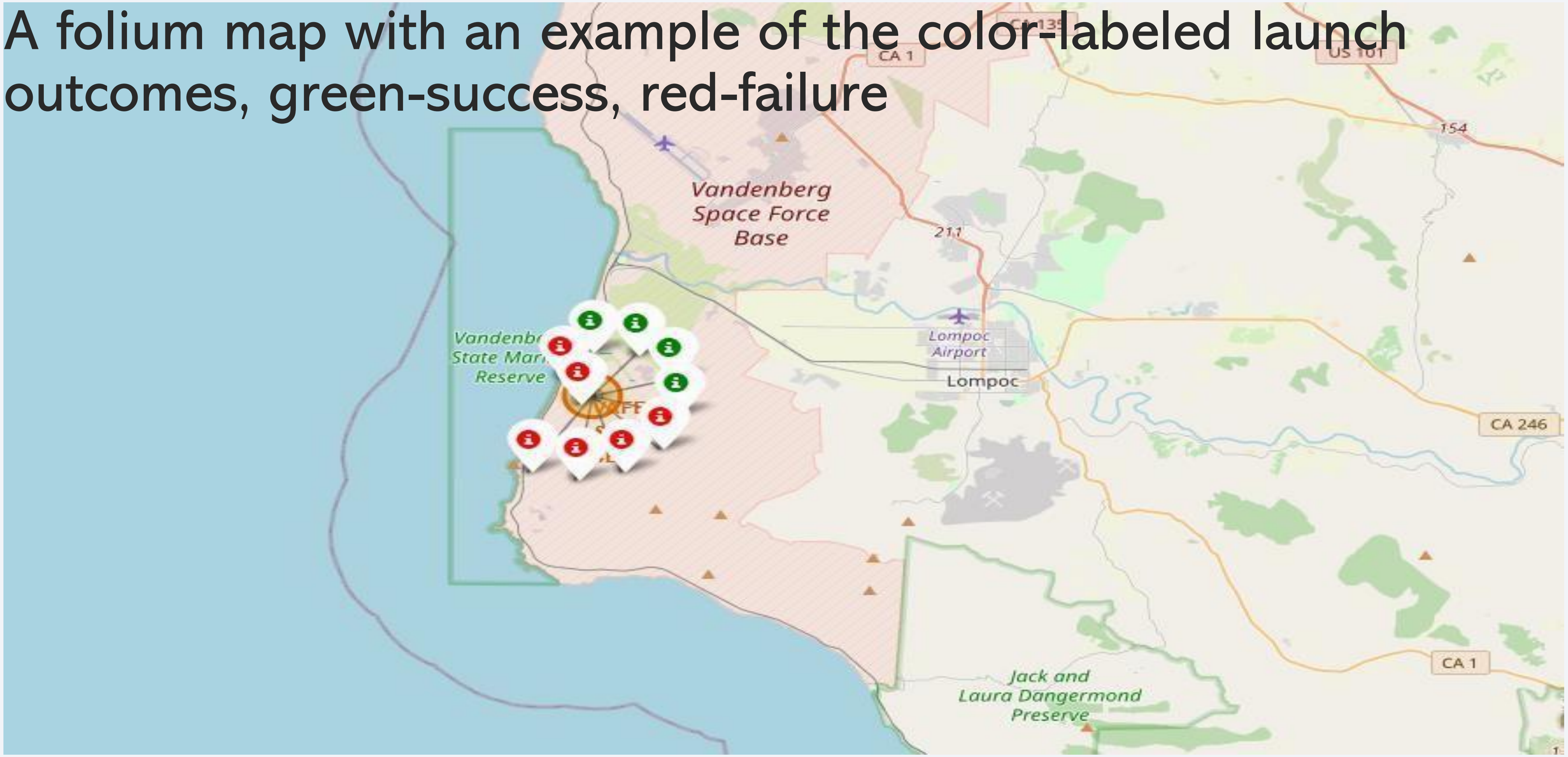
Launch Sites Proximities Analysis

Folium map shows all launch sites' location markers on a global map



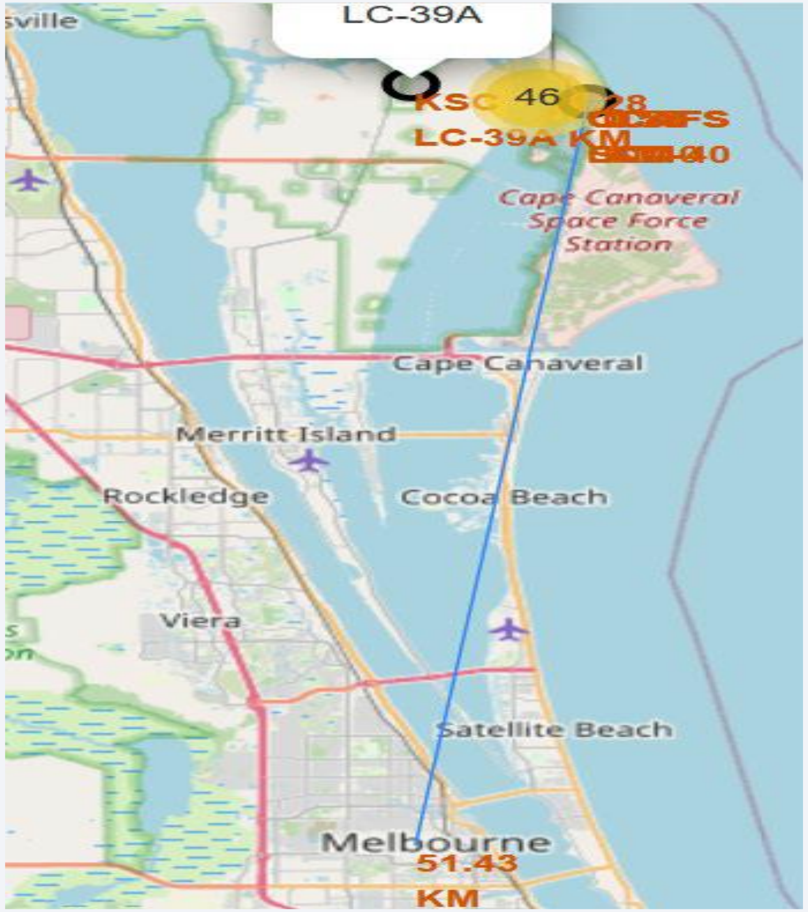
Folium Map: color-labeled launch outcomes

A folium map with an example of the color-labeled launch outcomes, green-success, red-failure



Launch Site Distances to its Proximities

51.43 KM from the nearest International Airport

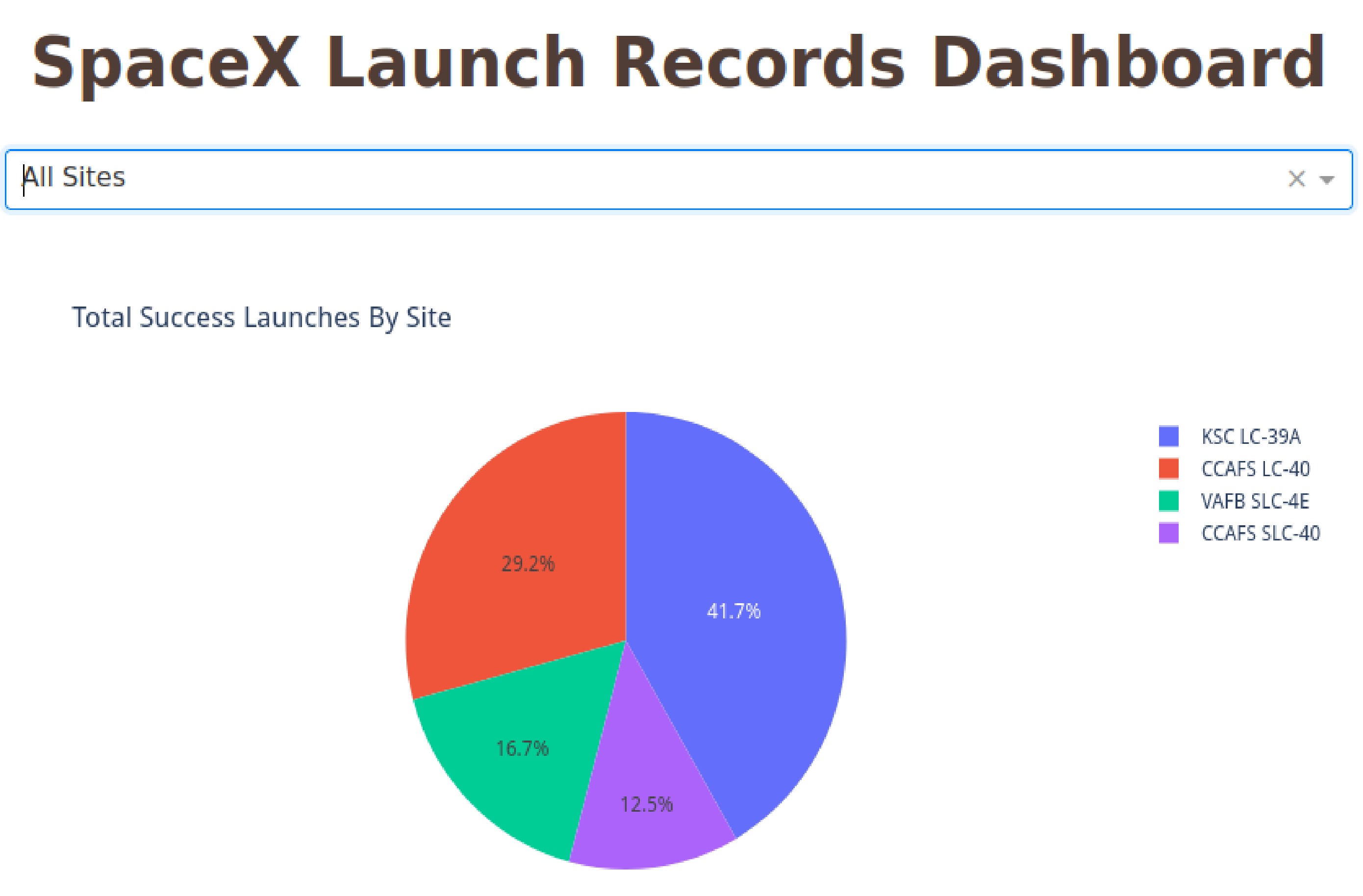


Section 4

Build a Dashboard with Plotly Dash

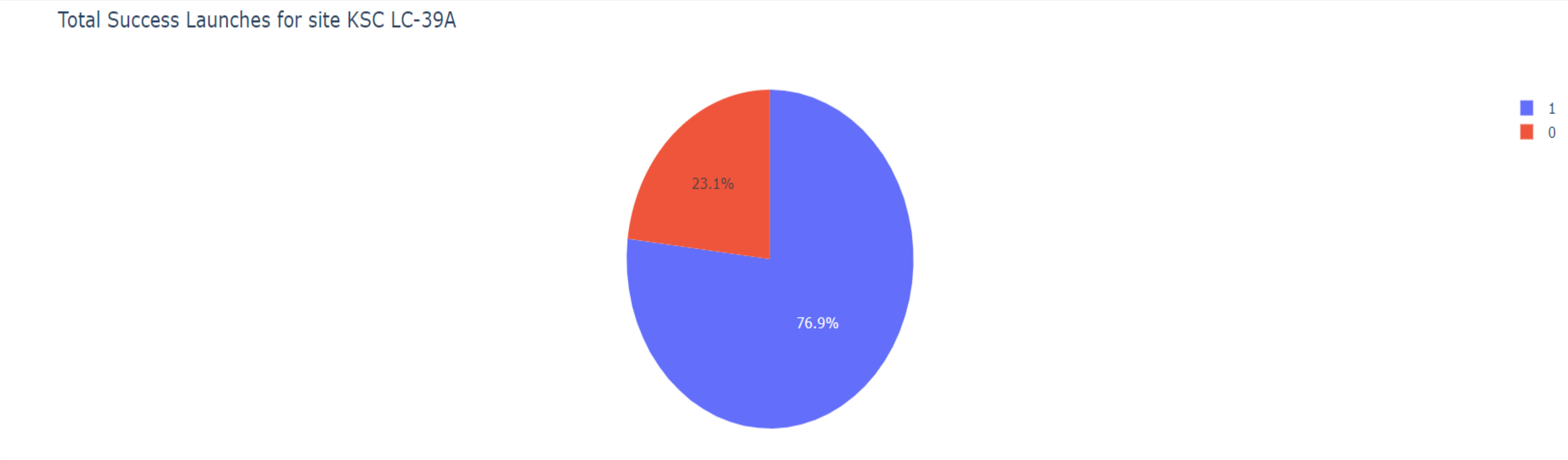
Pie Chart of Successful Landings For All Launch Sites

KSC LC-39A (Florida) has the highest number of successful launches

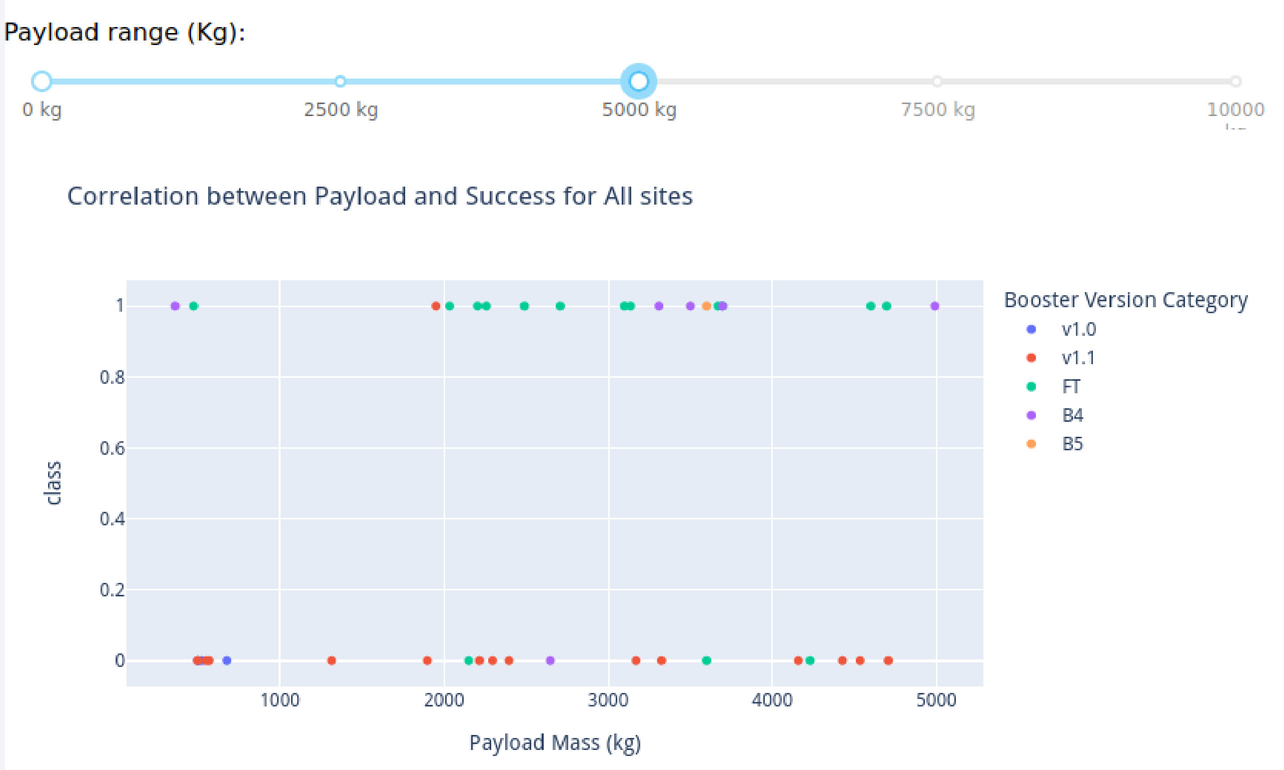


Pie chart for the launch site with highest launch success ratio

Pie chart for the launch site with highest launch success ratio is KSC, 76.9% of all launches were successful!



Payload vs. Launch Outcome for All Sites



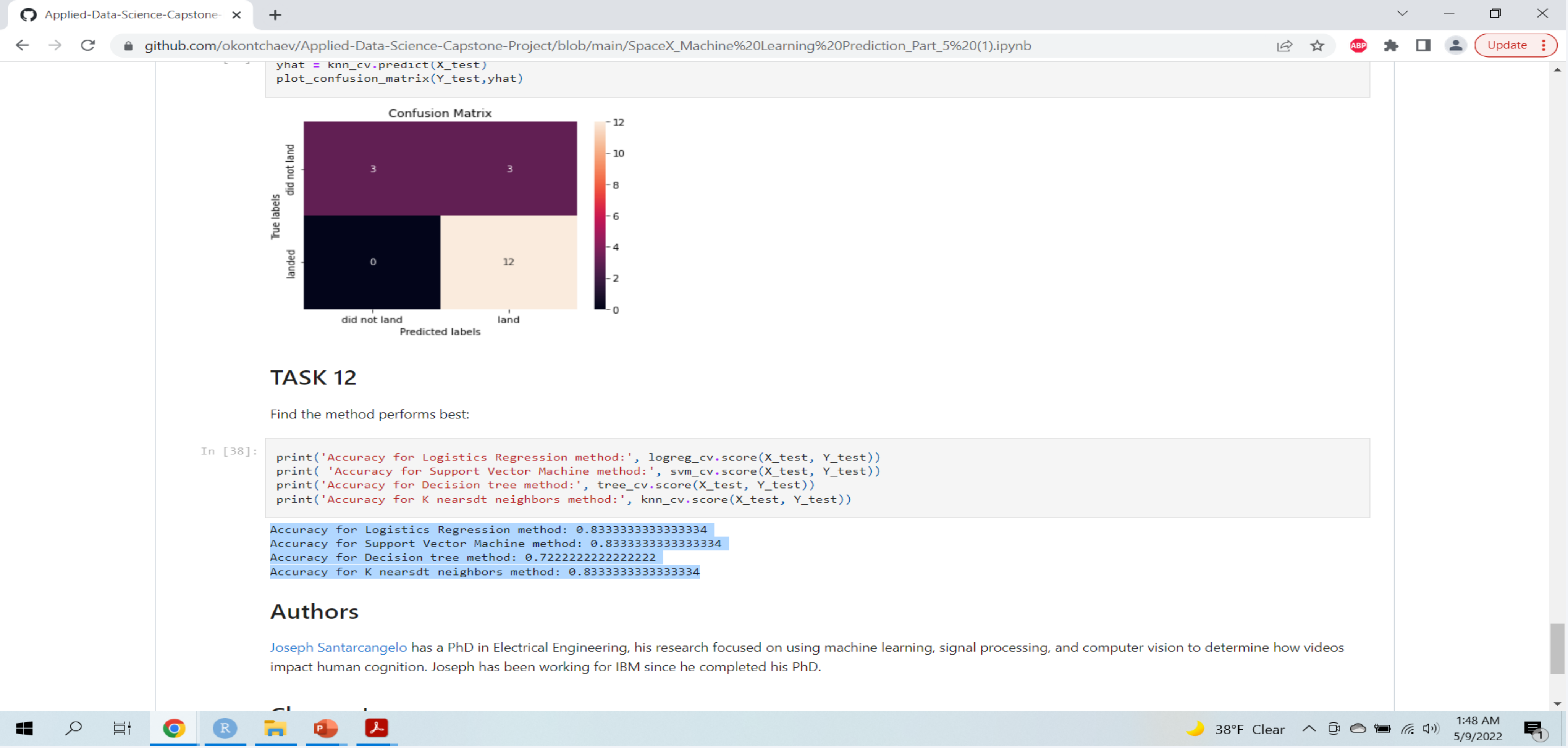
The plots presented above, clearly show that lower payloads result in higher success rates and the FT booster version has the largest success rate amongst all other boosters.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The machine model that has the lowest classification accuracy is the Decision Tree Classifier at 72.22 %, all other models reached the accuracy of 83.33%



Confusion Matrix

This is KNN Confusion Matrix. Along with SVM and LR it performed better then DT model.

Conclusions

- Launch success rate started to increase in 2013 till 2020.
- The success rate is generally increasing over time
- Launch site KSC LC-39A has the highest success ratio and the most successful launches
- Orbit types GEO, HEO, SSO and ES-L1 had the best success rate
- The Decision Tree Classifier has shown the less effective Machine Learning Algorithm in this project. SVM, Logistic Regression and KNNs has shown equal results

Appendix

- You are welcome to visit my GitHub repo at:

https://github.com/okontchaev

Thank you!