SpaceY – Predicting the reuse of the rocket's first stage

Nebojša Dinčić

May 31, 2025.

EXECUTIVE SUMMARY

- This research examines various factors that influences safe landing of a Falcon 9 rocket and its reuse. This enables SpaceY to predict launch price and be competitive to SpaceX.
- Methodologies used:
  - Data collecting (SpaceX REST-API data and web scrapping from Wikipedia)
  - Data wrangling (data cleaning and creating success/fail variable)
  - Exploratory data analysis using Pandas and SQL
  - Visualization of relationships among features
  - Geo-visualization of launch sites and launch outcomes
  - Creating and evaluating four machine learning models for predicting outcome of landing
- Conclusion: all four ML models perform equaly well, enabling SpaceY to predict whether the rocket can be reused, which will decrease the cost of launching.

TABLE OF CONTENT

INTRODUCTION

- **SpaceY** is new aspiring space technology company. They want to perform launches using Falcon 9 rockets.
- The first stage of a Falcon 9 rocket may be reused if it landed successfully. This reuse makes launches significantly cheaper.
- SpaceX is a leader in the domain, so SpaceY stuff should learn from their public data.
- Questions:
  - **Explore what parameters influences reusing the rocket's first stage.**
  - **Make a machine learning model that predicts the landing outcome a Falcon 9 rocket.**

METHODOLOGY

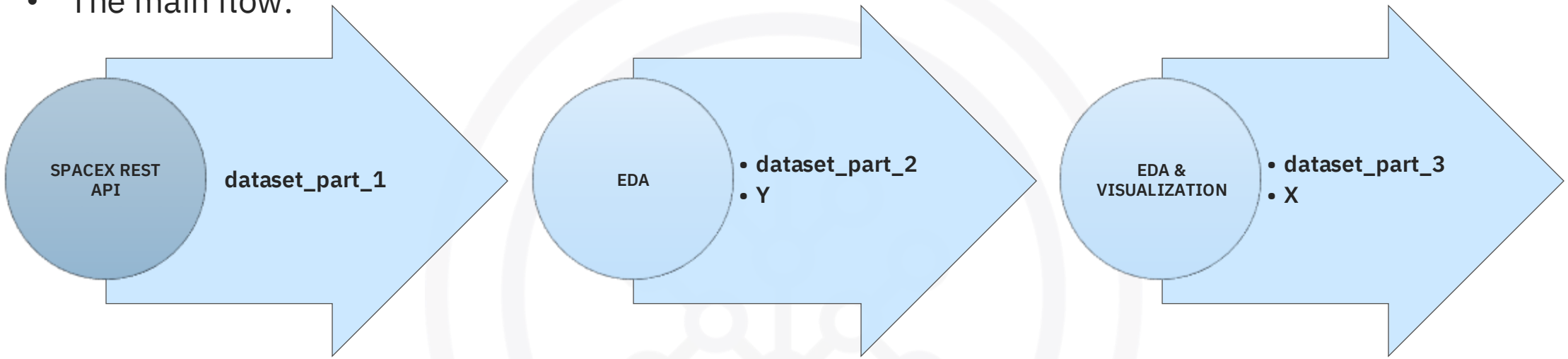
- Data sources:
  - SpaceX REST-API
  - Falcon 9 historical launches webpage (Wikipedia)
  - SpaceX.csv that will be read as a database
  - space_launch_geo.csv
- Programming languages and libraries
  - SQL
  - Python (Pandas, Numpy, Matplotlib, Seaborn, BeautifulSoup, Plotly, Dash, Scikit-Learn)
- Data acquisition:
  - Using SpaceX REST-API we obtained data as a series of JSON objects, then transform them into flat table (dataframe)
  - Using BeautifulSoup we perform web scrapping from the Falcon 9 Wikipedia webpage. Then we parse the launch records HTML table and obtain Pandas dataframe
  - During the data wrangling process, we again use SpaceX REST-API to decode some values

Data Collection

- Data collection is obtained combining web scrapping data from the Falcon 9 historical launches Wikipedia webpage and using SpaceX REST-API requests.
- More precisely, using BautifulSoup from the webpage we extracted the third table with the column names: 'Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'
- Then we parsed the HTML table and extracted data in the form of dictionary, created the dataframe, and saved it as a CSV file spacex_web_scraped.csv

RESULTS

- The main flow:



- Auxilliary flow: spacex_web_scrapped, SpaceX.csv (SQL), spacex_launch_geo (Folium), spacex_launch_dash (Plotly Dash)

SpaceX REST-API

- Using SpaceX REST-API, we obtain the data regarding the SpaceX launchings. Then we decode the response content as a JSON using .json(), and transform it into Pandas dataframe using .json_normalize().
- Resulting dataframe has a lot of ID-type data in its columns ('rocket', 'payloads', 'launchpad'). We dropped all columns except: 'rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc'.
- Knowing the rocket ID, we use the function getBusterVersion to obtain the booster name (as the 'BoosterVersion' column). Knowing the launchpad ID, we use the function getLaunchSite to find the launch site informations (as 'LaunchSite', 'Longitude' and 'Latitude' columns). Knowing the payloads ID, we use the function getPayloadData to find the payload informations (as 'PayloadMass' and 'Orbit' columns). Finally, applying the function getCoreData to the column 'cores' that contains a dictionary, we obtain various useful informations regarding cores and landing (as columns 'Block', 'ReusedCount', 'Serial', 'Outcome', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad').
- From the column 'date_utc' we extract the date (as the column 'date'). We also restrict the date to the period before November 13, 2020.

SpaceX REST-API - Data Wrangling

- Since we are interested only in Falcon 9 rockets with one core and one payload, we drop all rows where several cores or more than one payload appear. For convenience, we also reset the values in 'FlightNumber' columns.
- Examining the dataframe for missing values as a result gives 5 missing 'PayloadMass' column values and 26 missing 'LandingPad' values. We replace former by the mean payload mass, while keeping the latter to be None (which means that a landing pad was not used).
- As a result, we have the dataframe with 90 rows and 17 columns.
- The dataframe saved as dataset_part_1.csv file.

Web Scrapping

- Using BeautifulSoup, we scrapped the content from the "List of Falcon 9 and Falcon Heavy launches" Wikipage (updated on June 9th 2021)
- Using the function extract_column_from_header() we extracted the column names from the third HTML table ('Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome').
- Then we created launch_dict dictionary and populated it by parsing the HTML table. We used the following functions: date_time, booster_version, landing_status, get_mass. Each of them act on a HTML table cell and return: date and time, booster version, landing status, mass, respectively.
- We replaced 'Date and time ()' key by 'Date' and 'Time', and also added some more keys: 'Version Booster' and 'Booster landing'.
- We convert this dictionary to Pandas dataframe with 121 rows and 11 columns: 'Flight No.', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Version Booster', 'Booster landing', 'Date', 'Time', 'Launch Site'
- The dataframe saved as spacex_web_scraped.csv file.

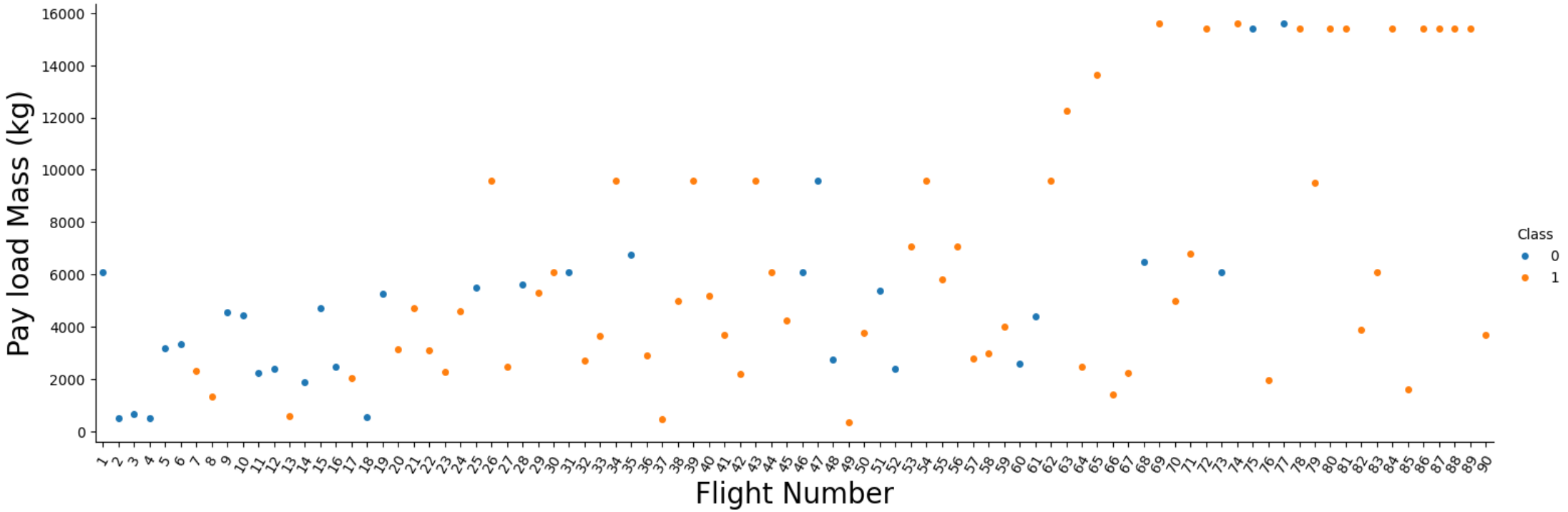
Exploratory Data Analysis - Pandas

- We performed EDA on the dataset_part_1.csv file and obtained the following conclusions:
  - Three Space X launch facilities: Cape Canaveral Air Force Space (CCAFS SLC 40), Vandenberg Air Force Base (VAFB SLC 4E), Kennedy Space Center (KSC LC 39A)
  - Launches distribution: more than 60% launches from CCAFS, 26% from KSC, 14% from VAFB
  - Orbit distribution: 30% launches to GEO (geosynchronous orbit), about 23% to ISS (international space station), about 16% to VLEO (very low Earth orbit)
  - Outcome distribution: **2/3 of all landings were successful** (more detailed, 45.6% successfully landed on a drone ship, 15.6% on a ground pad and 5.6% to a specific region in the ocean).
- **We added the binary column 'Class' that contains the value 1 if the landing was successful and 0 otherwise.**
- Such dataframe with 90 rows and 18 columns was saved as dataset_part_2.csv file.

Exploratory Data Analysis and Visualization

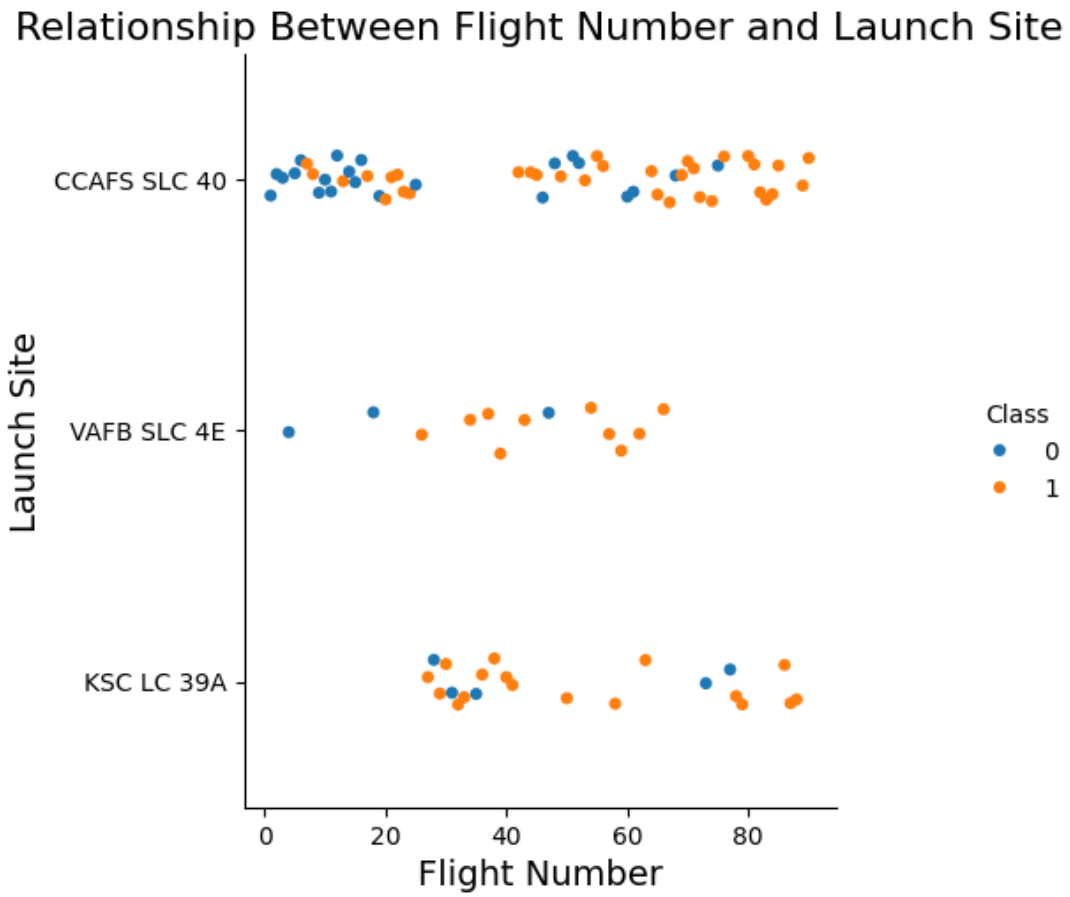
- We performed EDA on the dataset_part_2.csv file. We examined visually relationship between columns 'FlightNumber', 'PayloadMass', 'LaunchSite', 'Orbit' and 'Class' to prepare for features engineering.
- We often considered Seaborn scatterplots with two columns from the said set and 'Class' as hue.
- In order to prepare for machine learning phase, we casted all numerical columns to float64 datatype, and created dummy variables for four category columns ('Orbits', 'LaunchSite', 'LandingPad' and 'Serial') using OneHotEncoder.
- Resulting dataframe has 80 columns, and we saved it as dataset_part_3.csv file.

Payload Mass vs Flight Number



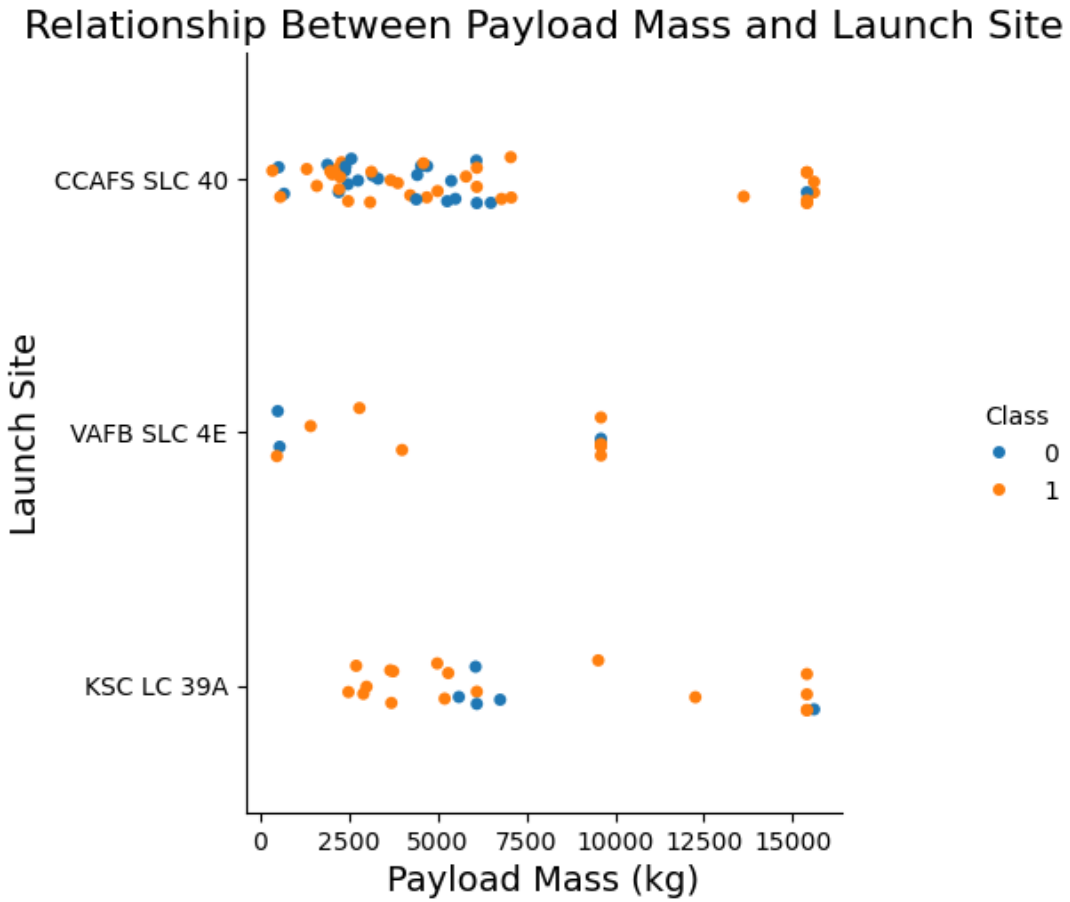
- As the Flight Number increases, the rocket's first stage more likely lands successfully
- With increasing Flight Number, the successful landing increases even for heavier payloads

Launch Site vs Flight Number



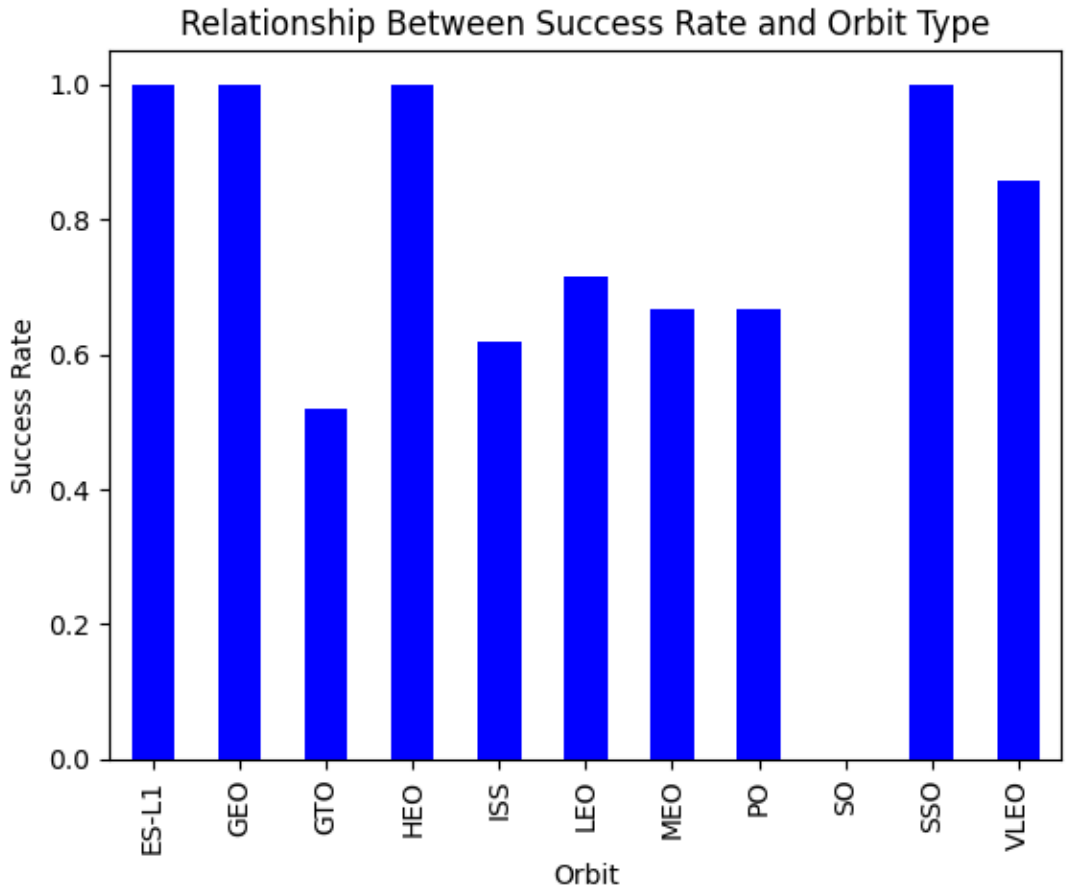
- CCAFS SLC 40 has the most launches, VAFB SLC 4E the least
- Successful landing becomes more likely on all three launch sites with increasing of flight number

Launch Site vs Payload Mass



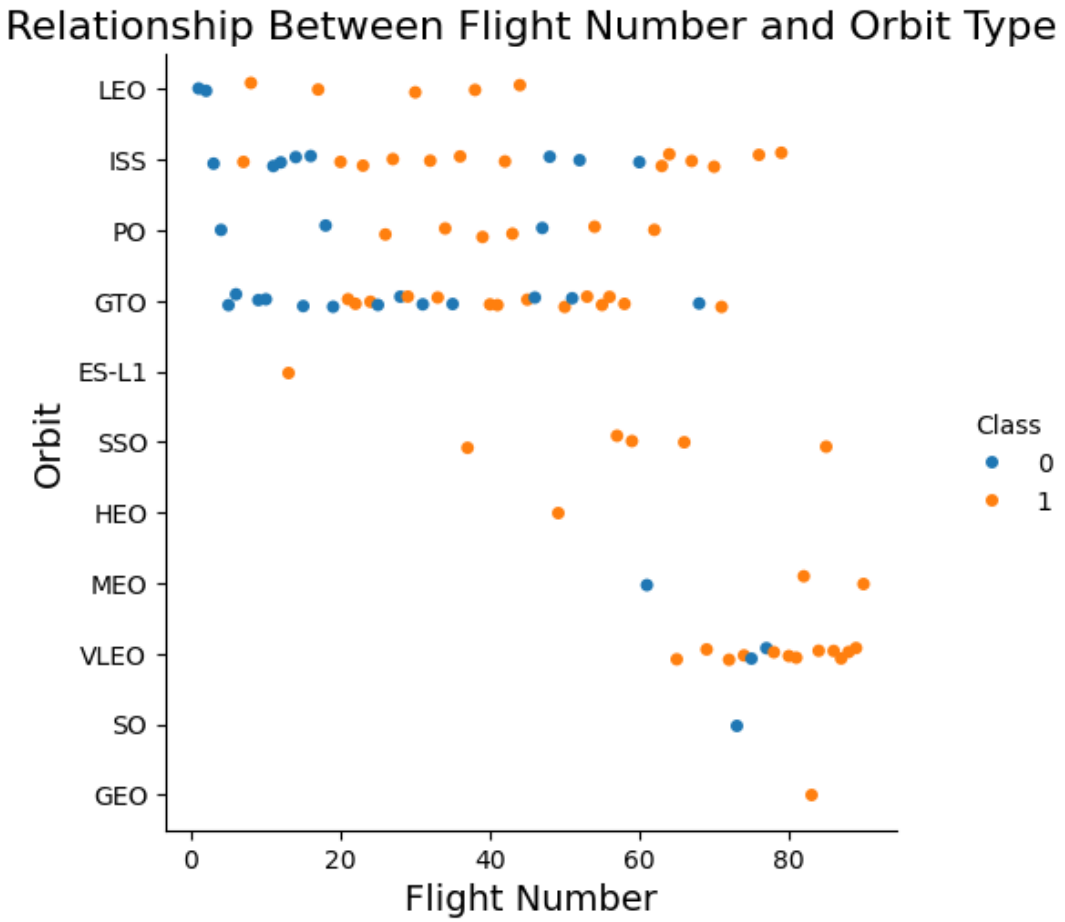
- No launches with payload greater than 10 tons from VAFB SLC 4E
- The relationship between the payload mass and the successful landing is rather complex, especially on CCAFS SLC 40 launch site.

Success Rate vs Orbit Type



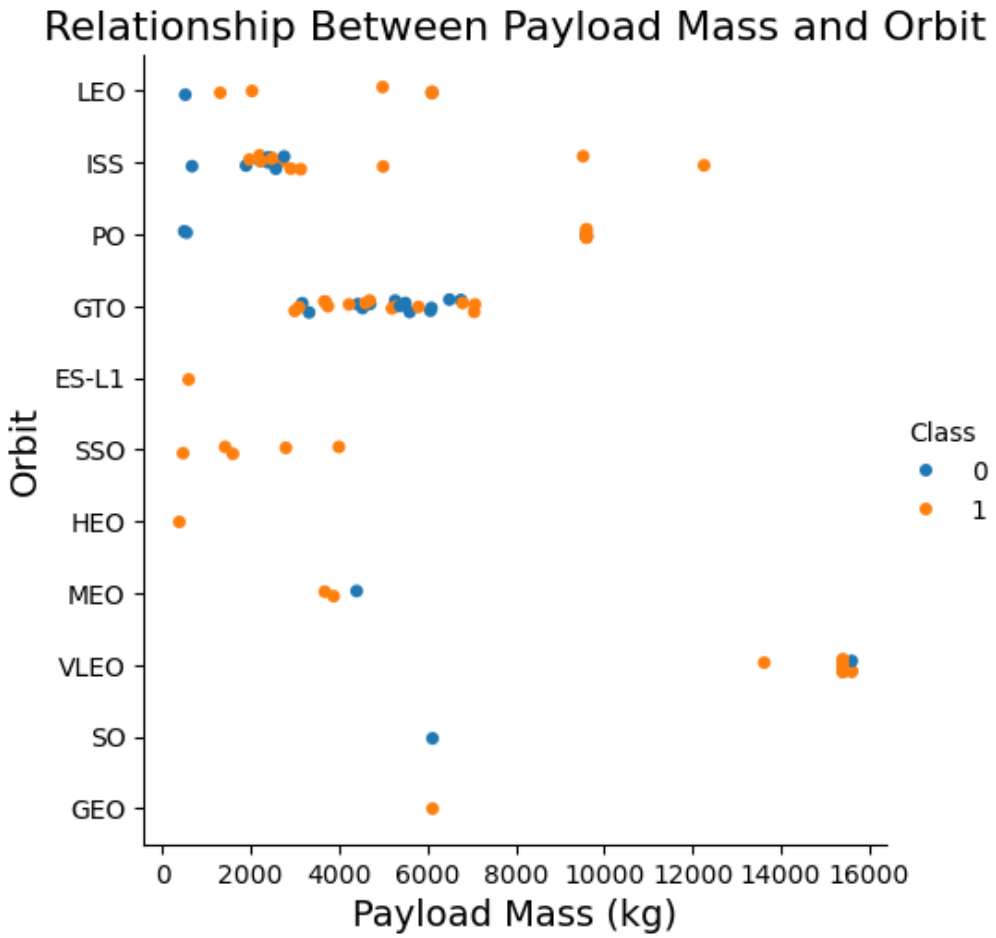
- ES-L1, GEO, HEO, SSO orbits are completely successful
- SO orbit is completely unsuccessful
- The success rate of remaining 6 orbits is between 50% and 85%

Orbit Type vs Flight Number



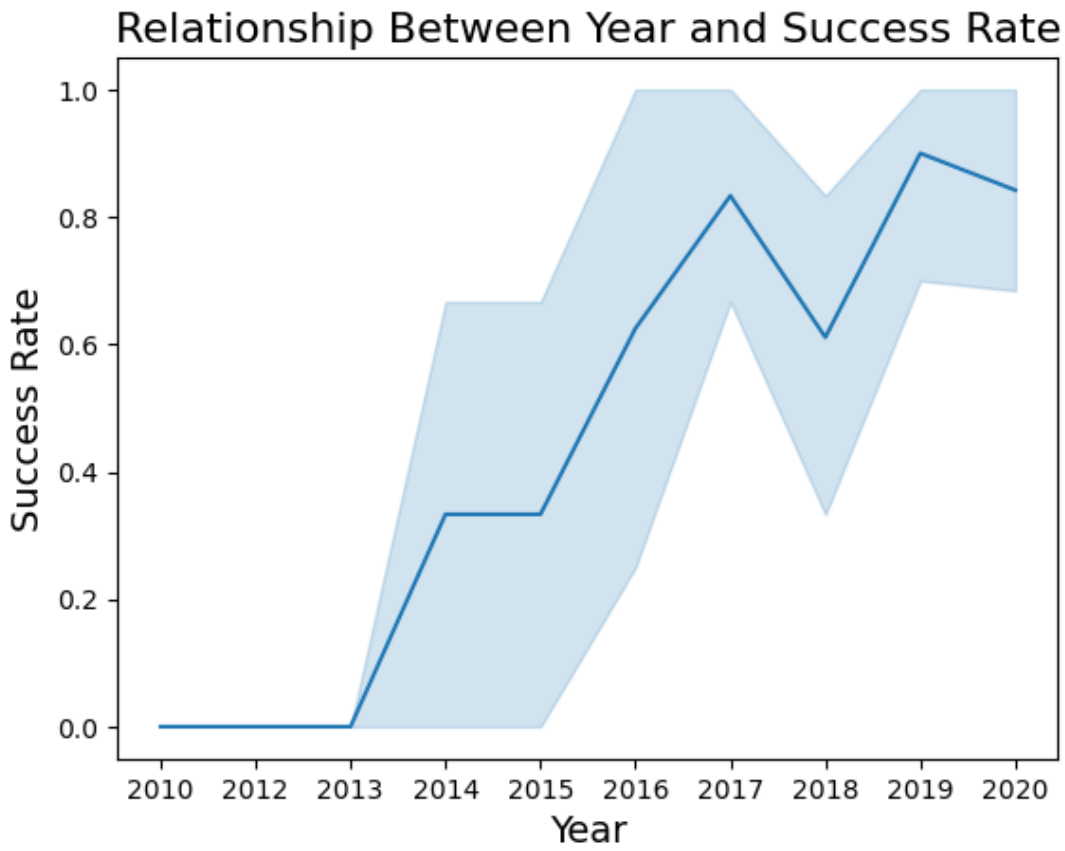
- In LEO orbit, success seems to be related with number of flights, but for GTO orbit there is no apparent relation
- SSO orbit appears to be 100% successful. Orbits ES-L1, HEO and GEO are also 100% successful, but they have only one launch per orbit.
- On the other hand, SO has only one launch and it failed.

Orbit Type vs Payload Mass



- With heavy payloads the successful landing or positive landing rate are more for PO,LEO and ISS.
- However, for GTO orbit, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Success Rate vs Year



- Since 2013, success rate increases, except for a small decrease in 2018.

Exploratory Data Analysis - SQL

- We performed EDA on the Spacex dataset using the SQL. The dataset was provided as Spacex.csv file, read using Pandas and converted to SQL table.
- Executing SQL queries we learned various facts:
  - Four launching sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
  - Average payload mass carried by booster Falcon 9 v1.1 is almost 3 tons
  - Almost 45.6 tons total payload mass was carried by boosters launched by NASA (CRS)
  - The first successful landing in a ground pad was achieved on December 22nd, 2015.
  - There were four boosters (B1022, B1026, B1021.2, B1031.2) that successfully landed in a drone ship and carried payload between 4 and 6 tons
  - Among 101 mission, 100 of them had successful landing.

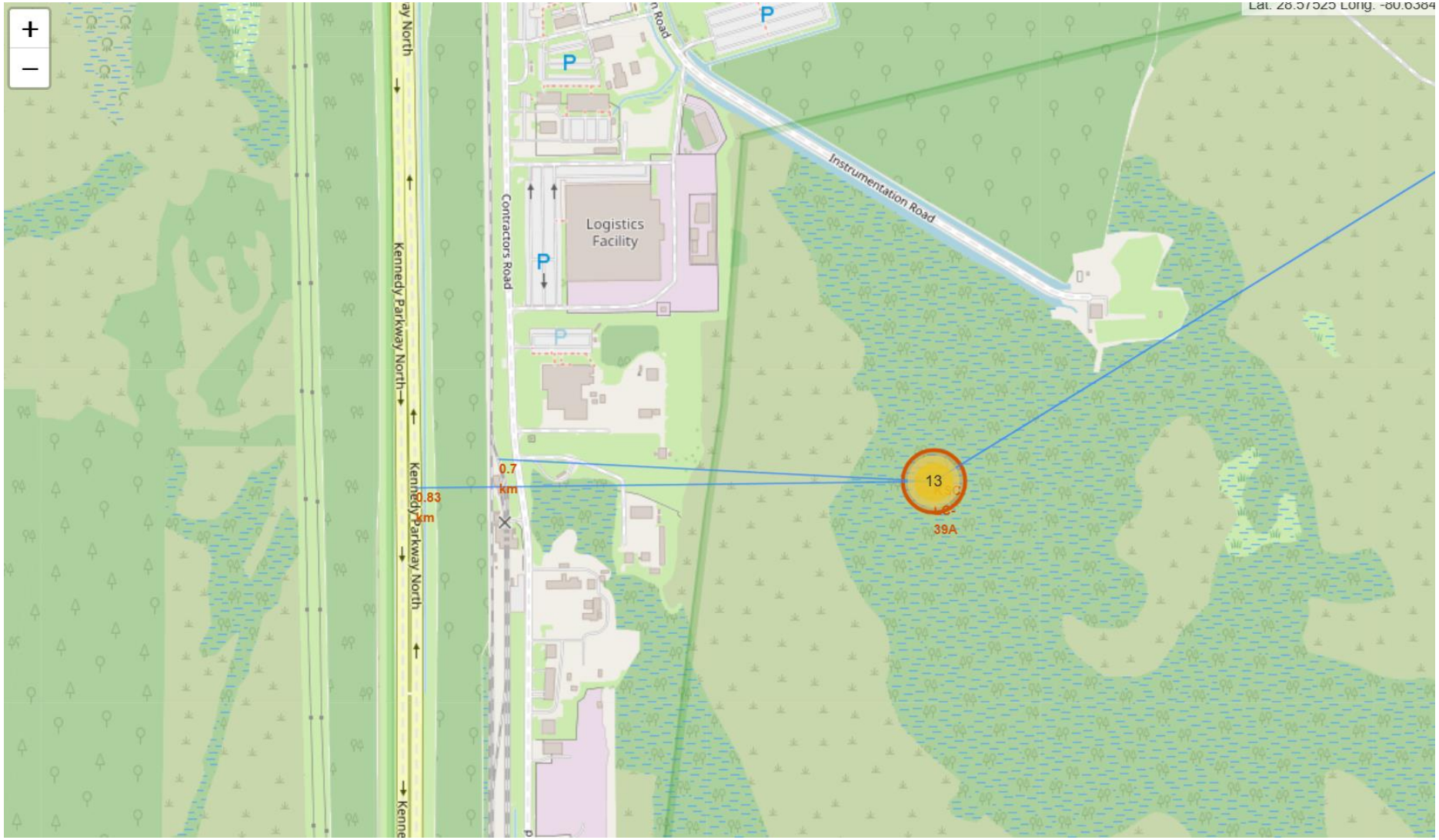
Geography of Launches

- We use the spacex_launch_geo.csv file, which among other informations contains the longitude and latitude of each launch site. The file contains 56 rows.
- Using Folium, we visualize each launch site as well as all landings on the map.
- We also examine the distance between launch sites and coastlines, highways, railways, and cities.
  - It is evident that launch sites are in proximity to railways and highways (often within a kilometer), probably due to transporting equipment. They are close to coasts, probably because if a launch is unsuccessful, it is far less dangerous to crash onto the ocean than onto a city.
  - The launch sites are away from cities because of safety, noise, and pollution.
  - Note also that all launch sites are near the Equator, because they use natural boosts due to the rotation of the Earth.

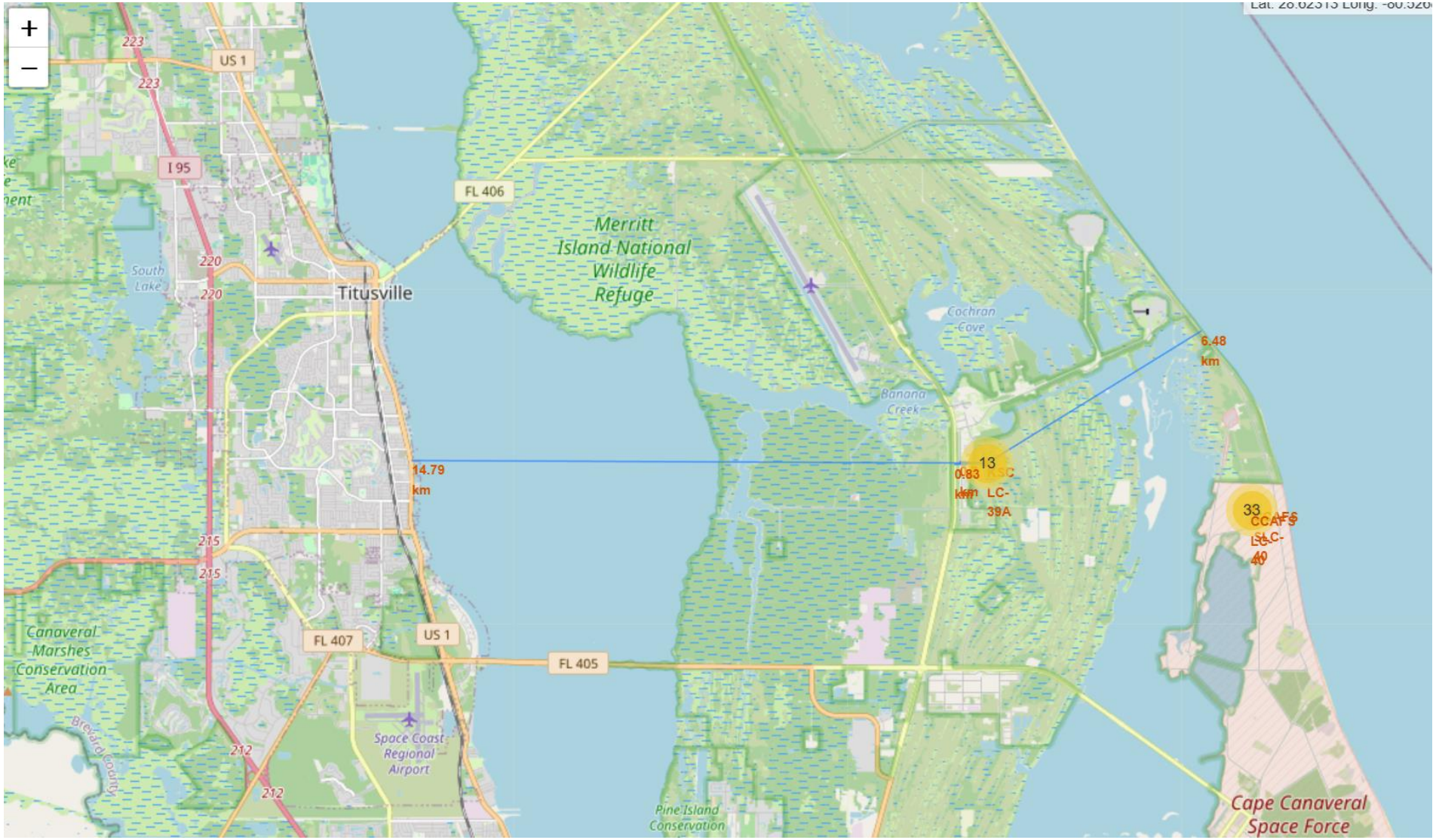
All four launch sites

All 56 launches

KSC LC-39A had 13 launches



KSC LC-39A highway and railway distances

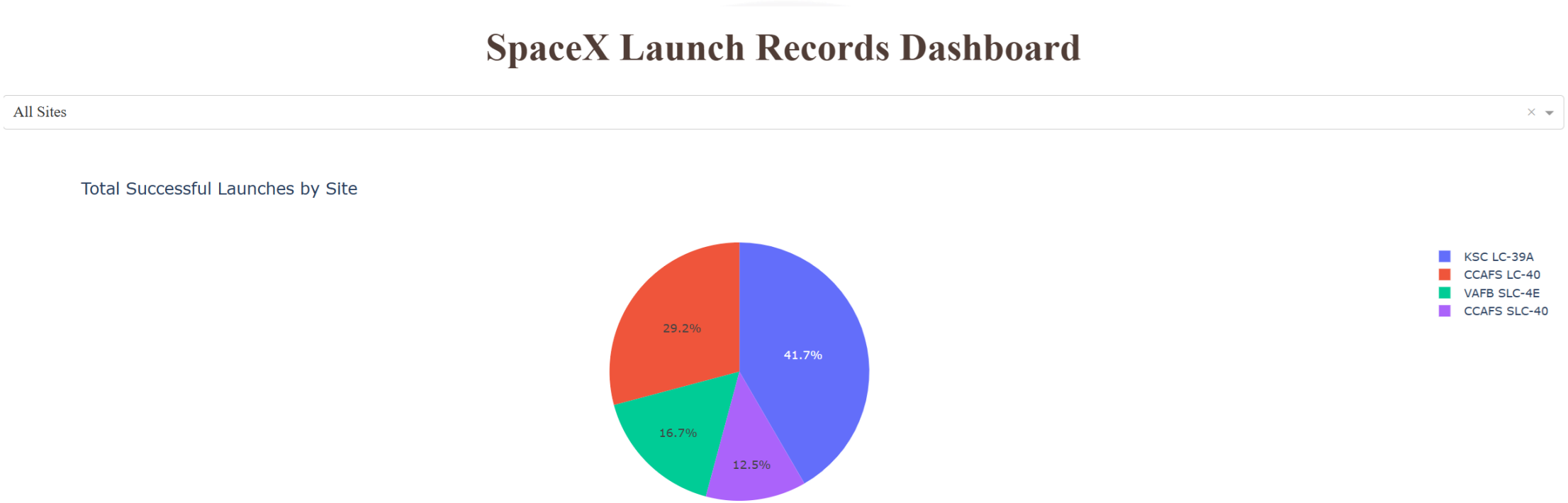


KSC LC-39A coastline and city distance

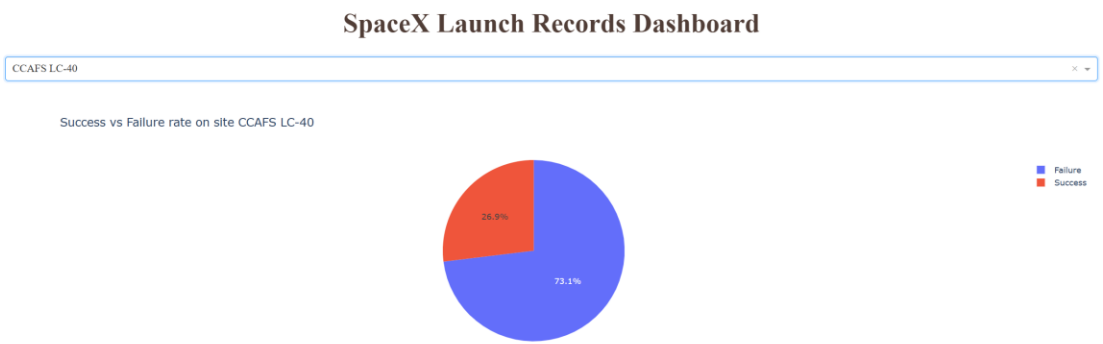
Plotly Dashboard

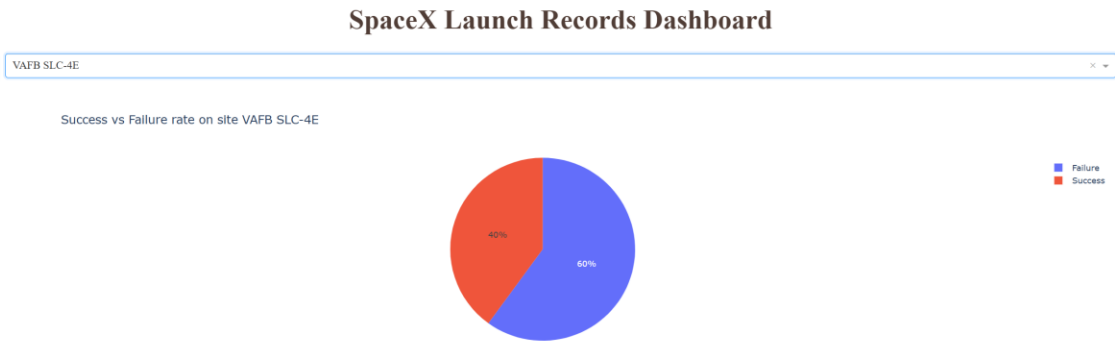
- The input file is spacex_launch_dash.csv
- Using Plotly Dash, we created the interactive dashboard that shows the pie chart with total successful launches for each of the four launch sites or all cumulatively (as chosen from the drop-down menu).
- Depending on the chosen launch site and the total payload (range is selected using two pointers), the scatterplot that shows booster version as well as the success is shown.

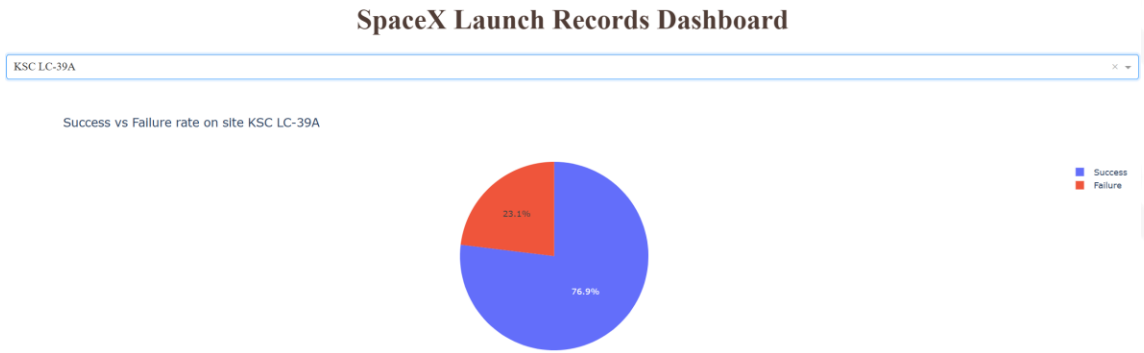
Plotly Dashboard

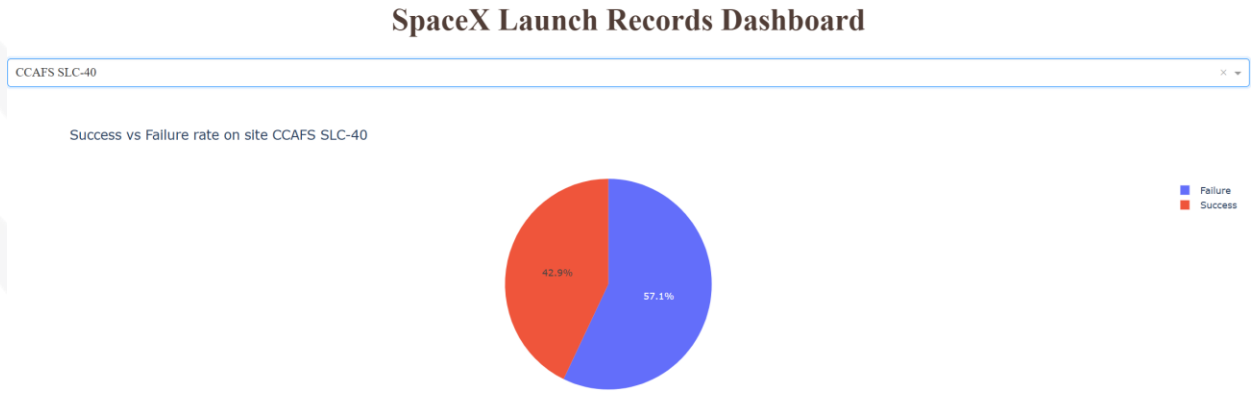


The most successful launch site is KSC LC-39A



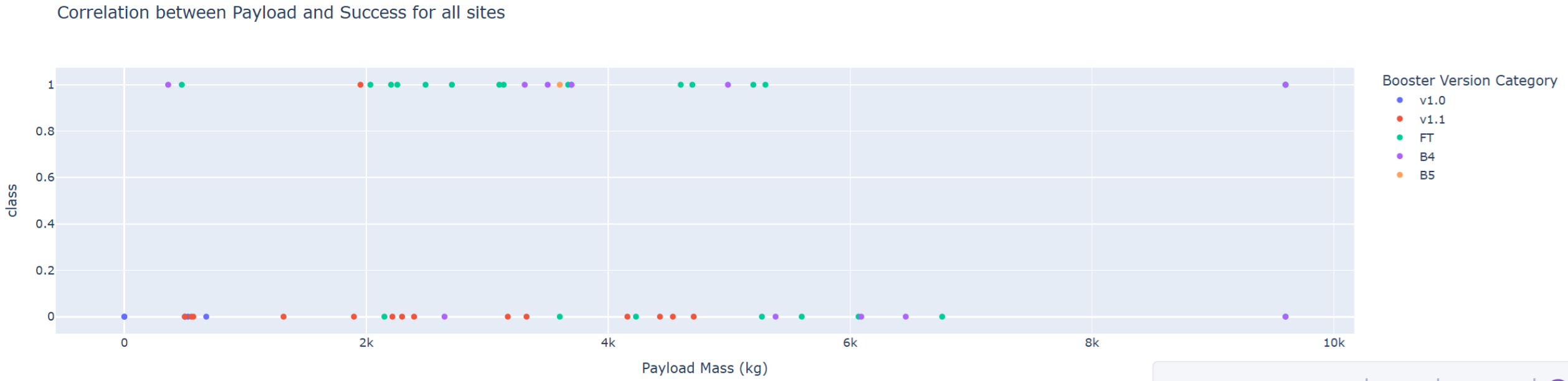




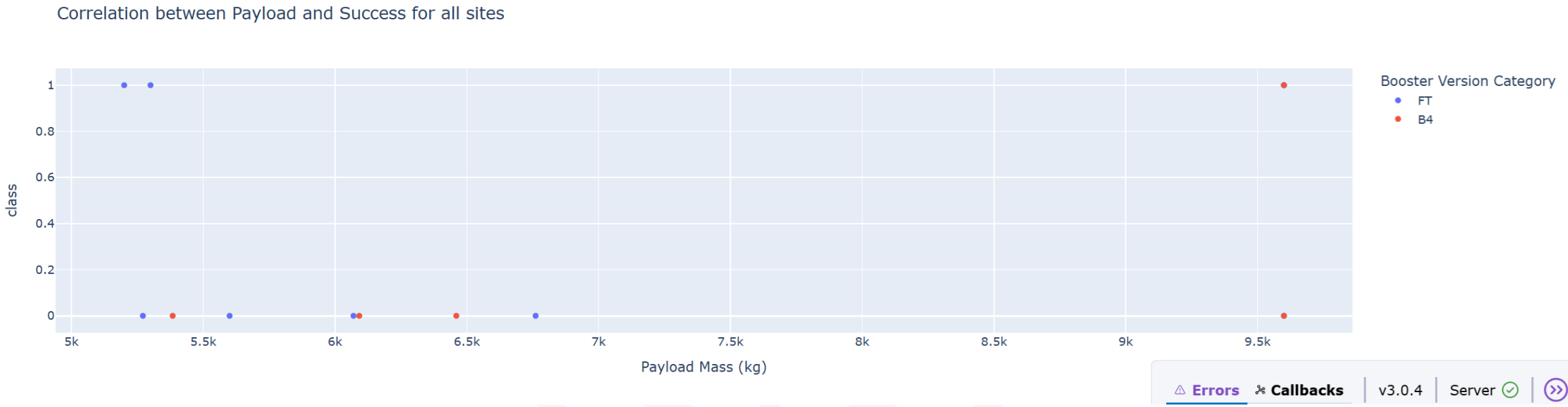


The most successful launch site is KSC LC-39A

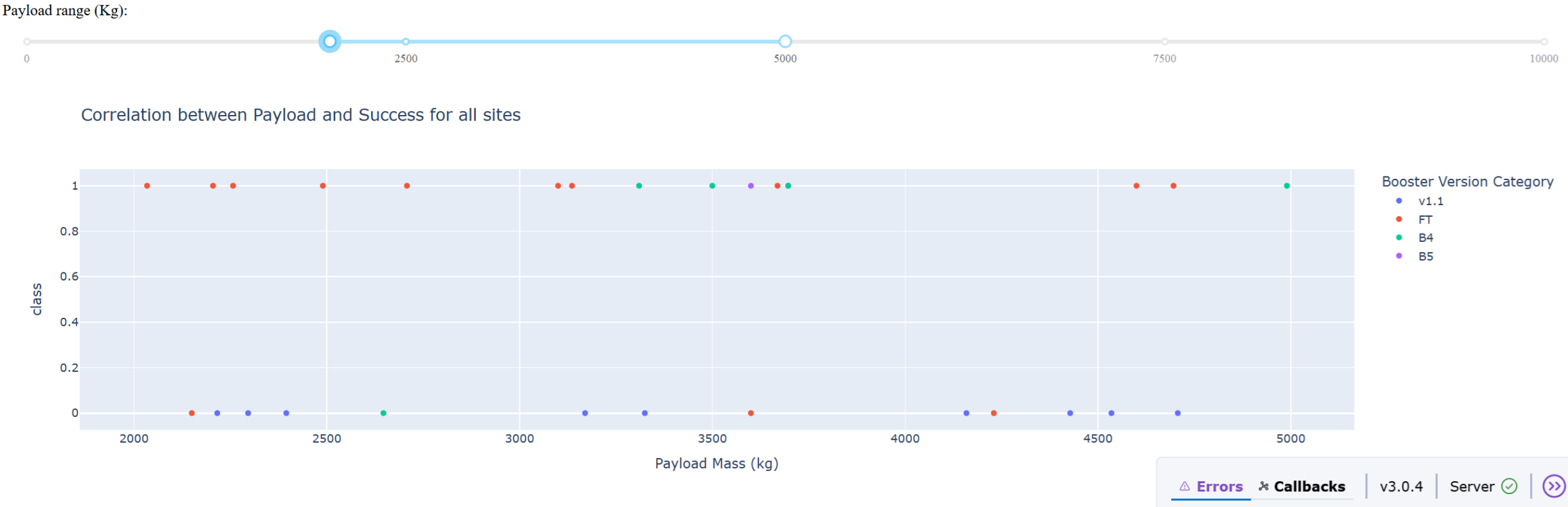
Payload range (Kg):



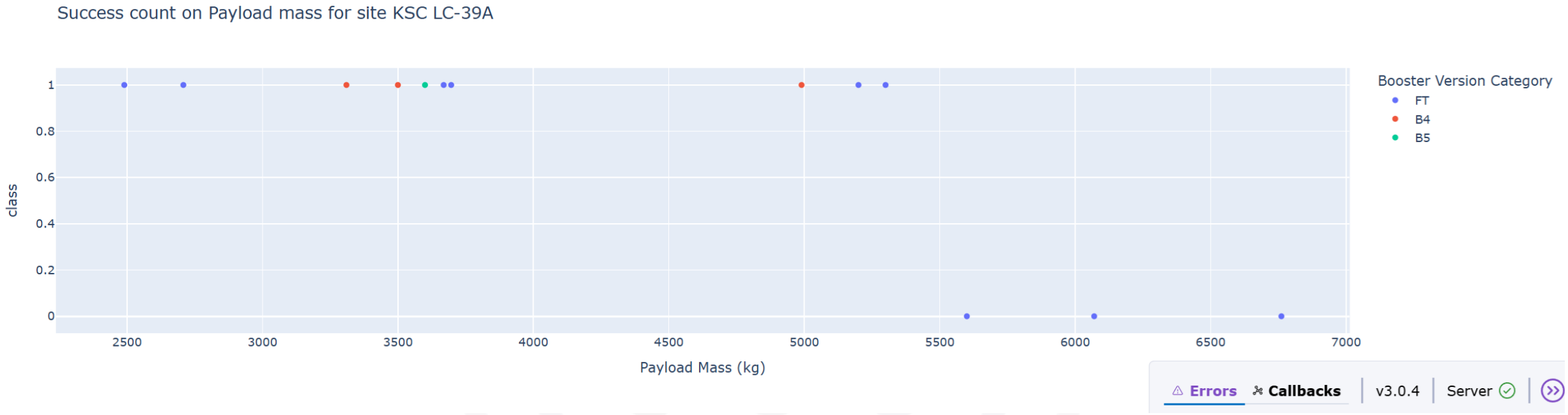
Payload range (Kg):



Correlation between Payload and Success when Payload is heavy (5-10 tons)



Correlation between Payload and Success when Payload is moderate (2-5 tons)



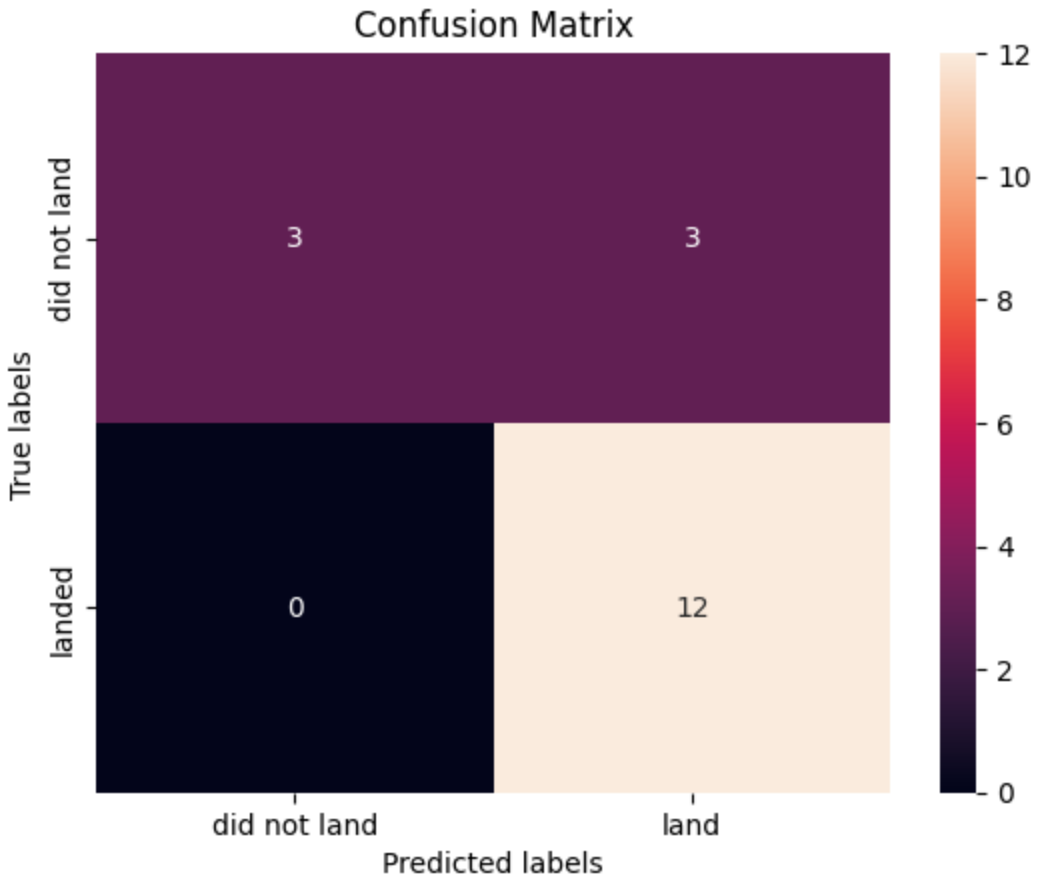
- Correlation between Payload and Success for KSC LC-39A
- High success for light and moderate weights (<5500 kg)
- High failure for heavy weights

Machine Learning Methods Applied

- We read the file dataset_part_2.csv into the dataframe data, and dataset_part_3.csv into the features dataframe X (90 rows and 83 columns)
- From the dataframe data we extract the target column 'Class' as Y
- After pretprocessing, we split the data X and Y into train and test parts (test size 0.2)
- We used the following methods
  - Logistic Regression
  - Support Vector Machine
  - Decission Tree Classifier
  - K-Nearest Neighbor
- Using GridSearchCV with cv=10, we found optimal parameters for all four methods.

Machine Learning Methods - Evaluation

- Used scikit-learn library.
- All four machine learning methods considered (Logistic Regression, Support Vector Machine, Decission Tree Classifier, K-Nearest Neighbor) perform the same.
- The accuracy for each of them is 83.33%
- They all share the confusion matrix
- Confusion matrix outputs
  - **12 True Positive**
  - 3 True Negative
  - **3 False Positive**
- Precision, Recall and F1 score for all four methods are also the same, and they are: 0.8, 1 and 0.89, respectively.



DISCUSSION

- The most successful launch site is KSC LC-39A (Kennedy Space Center Launch Complex 39A), especially for payload less than 5.5 tons.
- All launch sites are near coasts, railways and highways, but away from cities.
- All four machine learning models perform equaly. Caveat: the test data is small.

CONCLUSION

- We examined the factors that influence the outcome of landing the Falcon 9 rocket. If the landing is successful, we may reuse its first stage, and make launching cheaper.
- We built and evaluated four machine learning models for predicting the outcome of the landing.
- Suggested launch site would be KSC LC-39A.

APPENDIX

- Title page picture from https://wall.alphacoders.com/big.php?i=1180322
- The complete Jupyter notebooks can be found on github https://github.com/NDinca/SpaceY