IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shailendra Kumar
4/30/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection, Data Wrangling, Data Analysis, Data Visualization, Dashboards and Predictive Analytics
- Summary of all results
  - Predict with accuracy of 83%
  - Generated Dashboard to find insights visually

Introduction

- Project background and context
  - We predicted if the Falcon 9 first stage will land successfully. 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.
    I
- Problems you want to find answers
  - We wanted to determine the price of each launch. We did this by gathering information about Space X and creating dashboards. We also determined if SpaceX will reuse the first stage. We trained a machine learning model and used public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches updated on 9th June 2021
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- We web scraped to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches updated on 9th June 2021

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Web scraping done for Falcon 9 launch records with BeautifulSoup
- Extracted a Falcon 9 launch records HTML table from Wikipedia
- Parsed the table and converted it into a Pandas data frame
    - Requested the Falcon9 Launch Wiki page from its URL
    - Extracted all column/variable names from the HTML table header
    - Created a data frame by parsing the launch HTML tables

Data Collection – SpaceX API

GitHub URL

https://github.com/shail-70/AppliedDataScienceCapstone/blob/02b15a862e53aada7a0234a5553b6f83786cd0cf/jupyter-labs-spacex-data-collection-api.ipynb

Flowchart:

- Made Request to the SpaceX API
- Cleaned the requested data
- Requested and parsed the SpaceX launch data using the GET request
- Filtered the dataframe to only include Falcon 9 launches
- Data Wrangling done
    - Dealt with Missing Values

Data Collection - Scraping

- GitHub URL

https://github.com/shail-70/AppliedDataScienceCapstone/blob/02b15a862e53aada7a0234a5553b6f83786cd0cf/jupyter-labs-webscraping%20(1).ipynb

Flowchart:

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame
- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Flowchart:
- Performed exploratory Data Analysis and determine Training Labels
  - Exploratory Data Analysis
    - Calculated the number of launches on each site
    - Calculated the number and occurrence of each orbit
    - Calculated the number and occurence of mission outcome per orbit type
  - Determine Training Labels
    - Create a landing outcome label from Outcome column
- GitHub URL:

https://github.com/shail-70/AppliedDataScienceCapstone/blob/02b15a862e53aada7a0234a5553b6f83786cd0cf/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Summary of charts plotted
  - plot out the FlightNumber vs. PayloadMass - Categorical Plot(Catplot)
  - Visualize the relationship between Flight Number and Launch Site (catplot)
  - Visualize the relationship between Payload and Launch Site (catplot)
  - Visualize the relationship between success rate of each orbit type (barplot)
  - Visualize the relationship between FlightNumber and Orbit type (Catplot)
  - Visualize the relationship between Payload and Orbit type (Catplot)
  - Visualize the launch success yearly trend (lineplot)
- Github URL:

https://github.com/shail-70/AppliedDataScienceCapstone/blob/02b15a862e53aada7a0234a5553b6f83786cd0cf/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Summary of the SQL queries performed to understand the Spacex DataSet
  - Loaded the dataset into the corresponding table in a Db2 database
  - Displayed the names of the unique launch sites in the space mission
  - Displayed 5 records where launch sites begin with the string 'CCA'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Listed the date when the first succesful landing outcome in ground pad was acheived.
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes
  - Listed the names of the booster_versions which have carried the maximum payload mass.
  - Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site
  - Ranked the count of successful landing_outcomes
- GitHub URL:

https://github.com/shail-70/AppliedDataScienceCapstone/blob/02b15a862e53aada7a0234a5553b6f83786cd0cf/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summary of map objects created and added to a folium map
  - Folium Circle to add a highlighted circle area with a text label on a specific coordinate
  - Folium.Marker for each launch site on the site map
  - Marker clusters to simplify a map containing many markers having the same coordinate.
  - MousePosition on the map to get coordinate for a mouse over a point on the map
  - Draw a PolyLine between a launch site to the selected coastline point
- We added those objects to
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- GitHub URL:

https://github.com/shail-70/AppliedDataScienceCapstone/blob/2fdb82234f35b369dee928c4953a476f019bfadb/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summary of plots/graphs and interactions we have added to dashboard
  - Added a Launch Site Drop-down Input Component
  - Added a callback function to render success-pie-chart based on selected site dropdown
  - Add a Range Slider to Select Payload
  - Add a callback function to render the success-payload-scatter-chart scatter plot

- We added those plots and interactions to:
  - Build a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- GitHub: https://github.com/shail-70/AppliedDataScienceCapstone/blob/4e4ad5b5a68b2f5b38626edad3e9d90025f19146/spacex_dash_app.py

Predictive Analysis (Classification)

- Summary of how we built, evaluated, improved, and found the best performing classification model
  - Performed exploratory Data Analysis and determined Training Labels
  - created a column for the class
  - Standardized the data
  - Split into training data and test data
  - -Found best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Found the method performs best using test data
- You need present your model development process using key phrases and flowchart
  - Loaded the dataframe
  - Created a NumPy array from the column Class
  - Standardized the data
  - split the data X and Y into training and test data
  - Created logistic regression, support vector machine , decision tree classifier, k nearest neighbors models then create a GridSearchCV object
  - Calculated the accuracy on the test data
  - Found the method performs best
- GitHub URL:
  - https://github.com/shail-70/AppliedDataScienceCapstone/blob/199310dc9a3e0bc4f5721dc95bc520ea13cda176/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

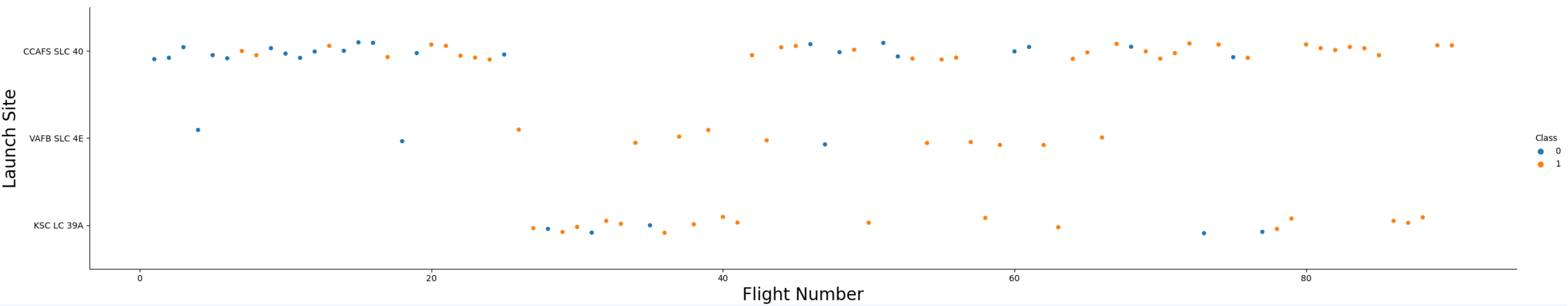
Results

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

Section 2

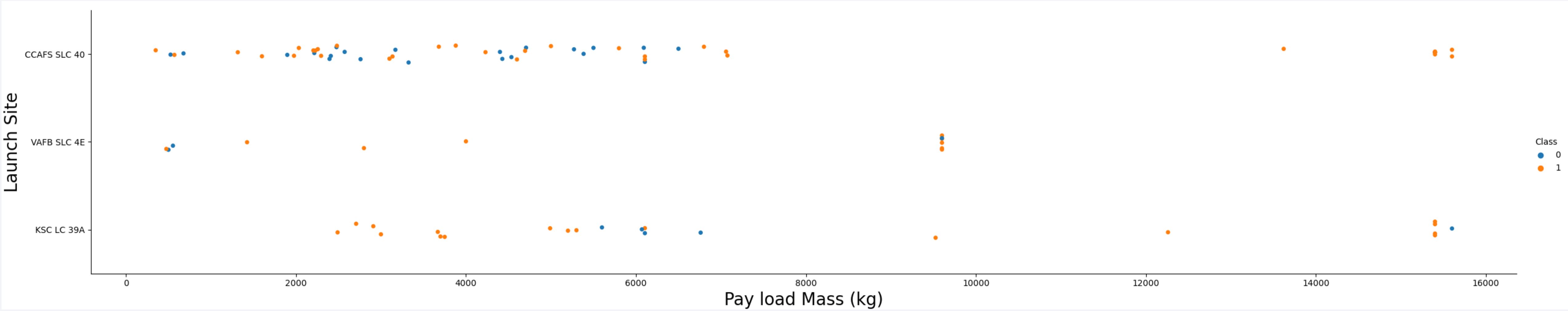
Insights drawn from EDA

Flight Number vs. Launch Site



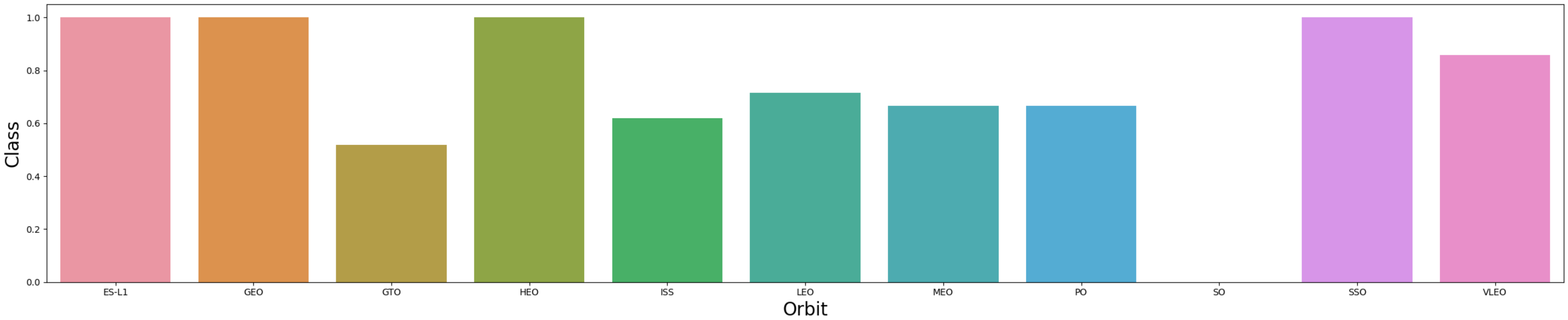
- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site



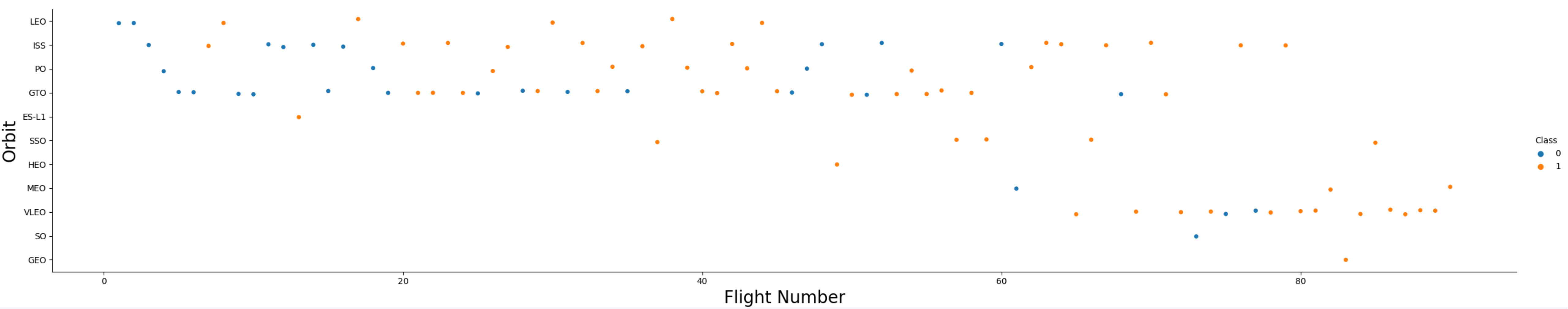
- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type



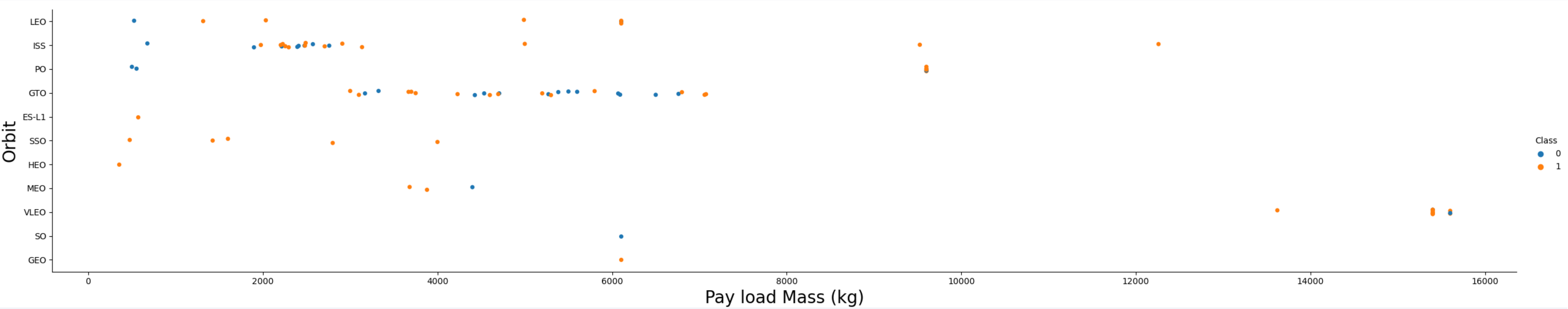
- Visualise relationship between success rate and orbit type

Flight Number vs. Orbit Type



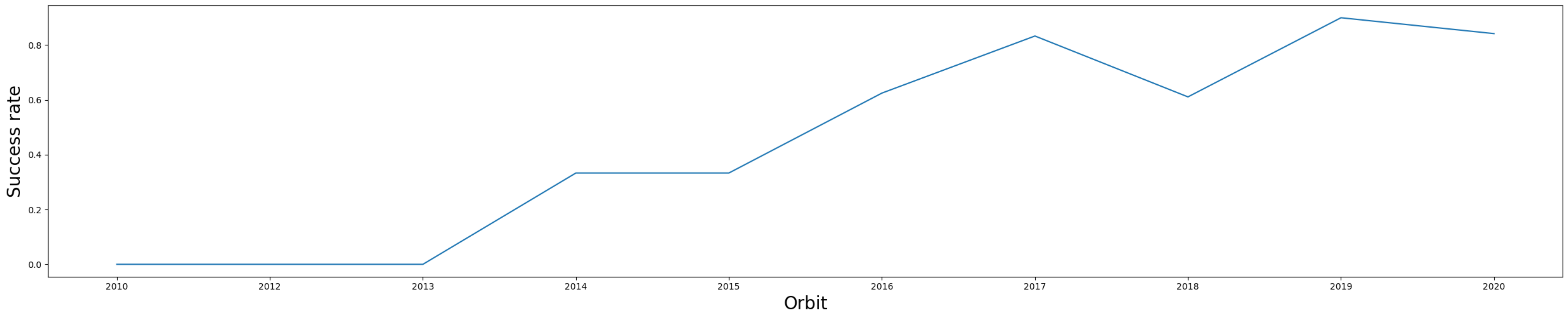
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

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) are both there here.

Launch Success Yearly Trend



- sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

- 'CCAFS LC-40', 'VAFB SLC-4E', 'KSC LC-39A', 'CCAFS SLC-40'

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - The launch site name of the 5 records is CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - 45596 Kg

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - 2928.4 Kg

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - 22nd December 2015

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
  - 'F9 FT B1022','F9 FT B1026','F9 FT B1021.2','F9 FT B1031.2'

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
    - 100 Successful
    - 1 Failure

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - '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

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

  - 'Jan', '2015', 'Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40'
  - 'Apr', '2015', 'Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40'
  - 'Jun', '2015', 'Precluded (drone ship)', 'F9 v1.1 B1018', 'CCAFS LC-40'

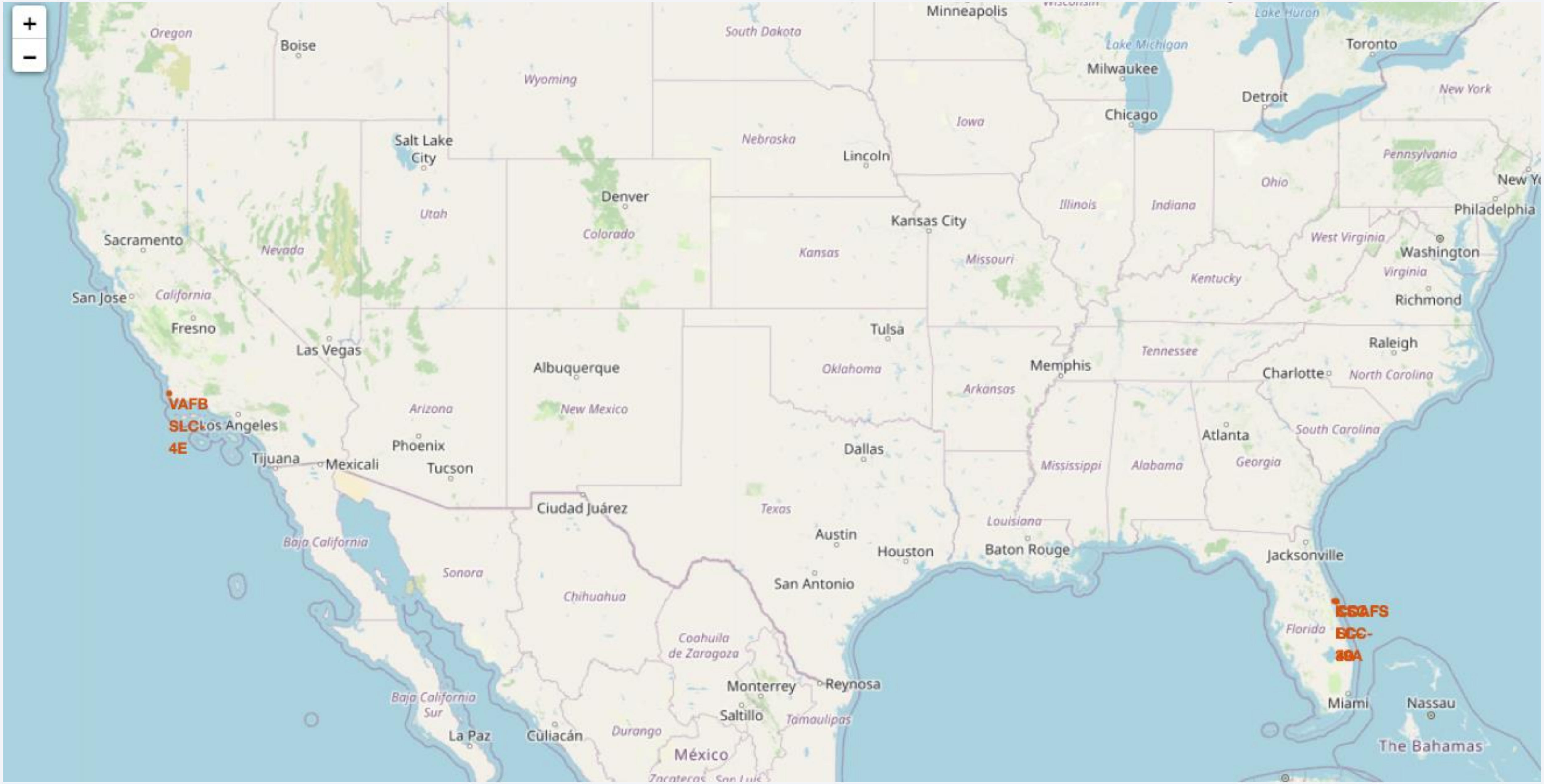
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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
  - '19-02-2017'
  - '14-01-2017'
  - '14-08-2016'
  - '18-07-2016'
  - '27-05-2016'
  - '06-05-2016'
  - '08-04-2016'
  - '22-12-2015'

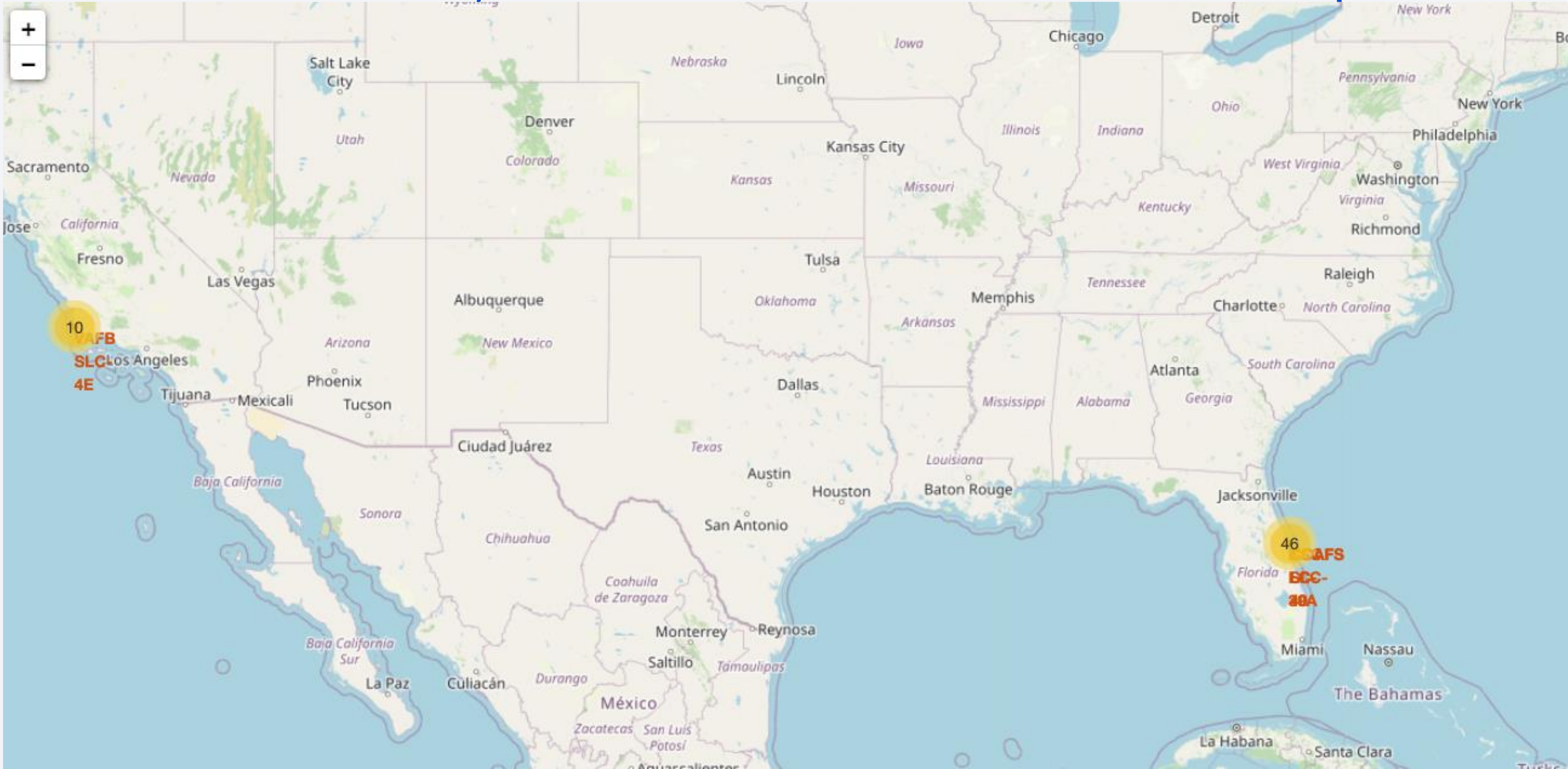
Section 3

Launch Sites Proximities Analysis

The generated map with marked launch sites



Mark the success/failed launches for each site on the map



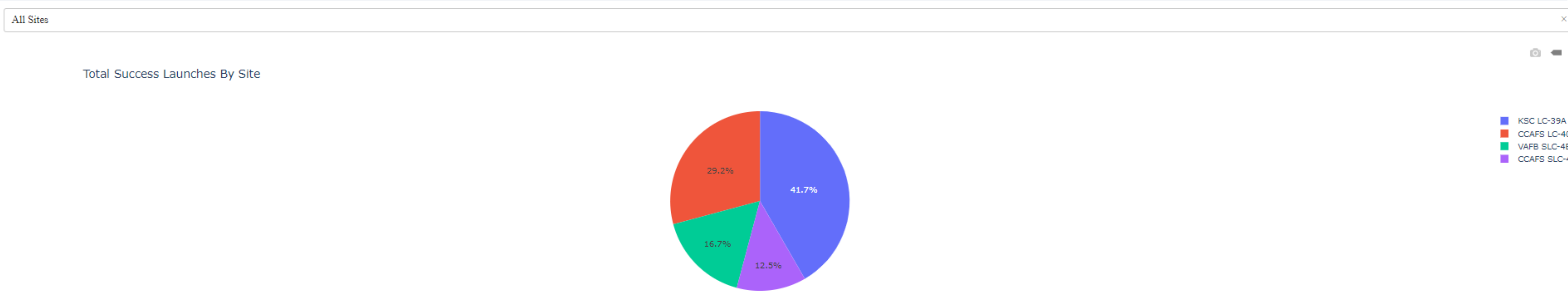
Selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

Section 4

Build a Dashboard with Plotly Dash

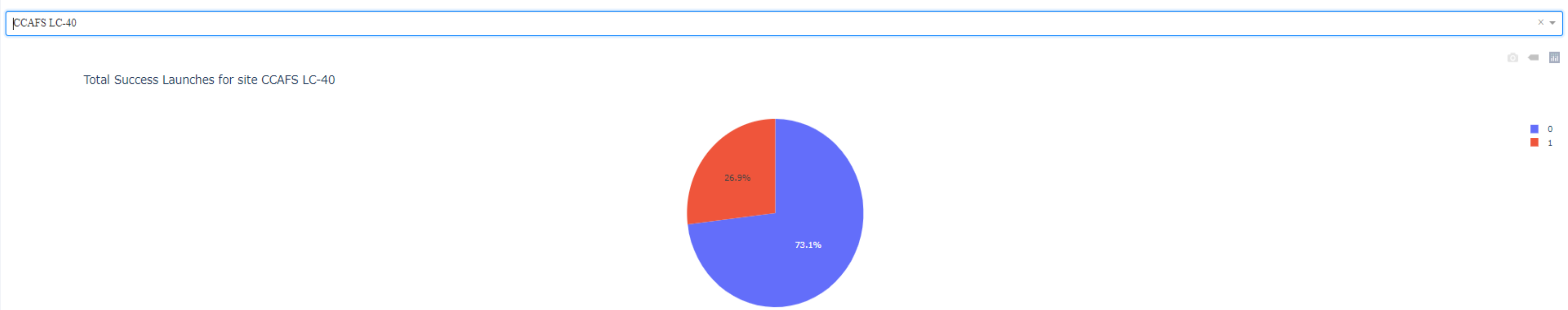
Launch success count for all sites

- Explain the important elements and findings on the screenshot



Total Success Launches

- Replace <Dashboard screenshot 2> title with an appropriate title



Correlation between Palyload and Succuss

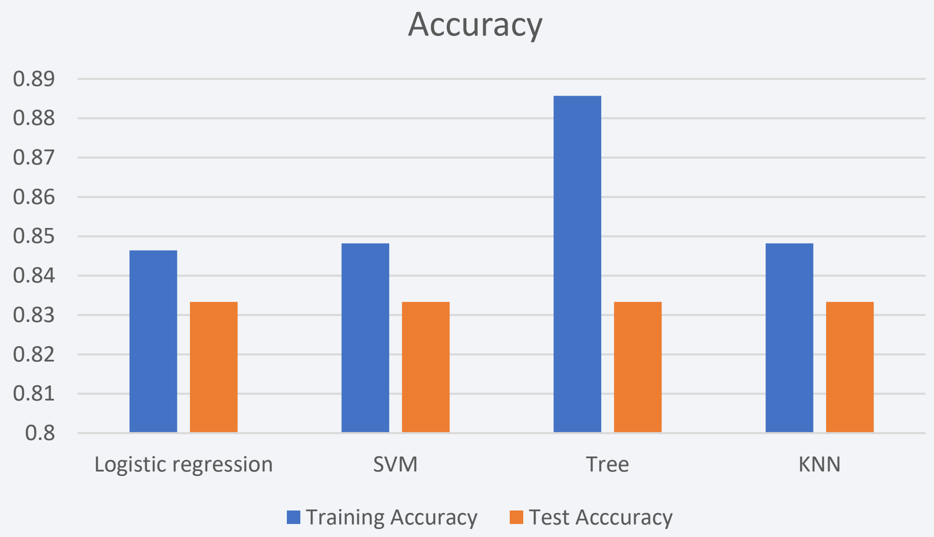
- Replace <Dashboard screenshot 3> title with an appropriate title

Section 5

Predictive Analysis (Classification)

Classification Accuracy

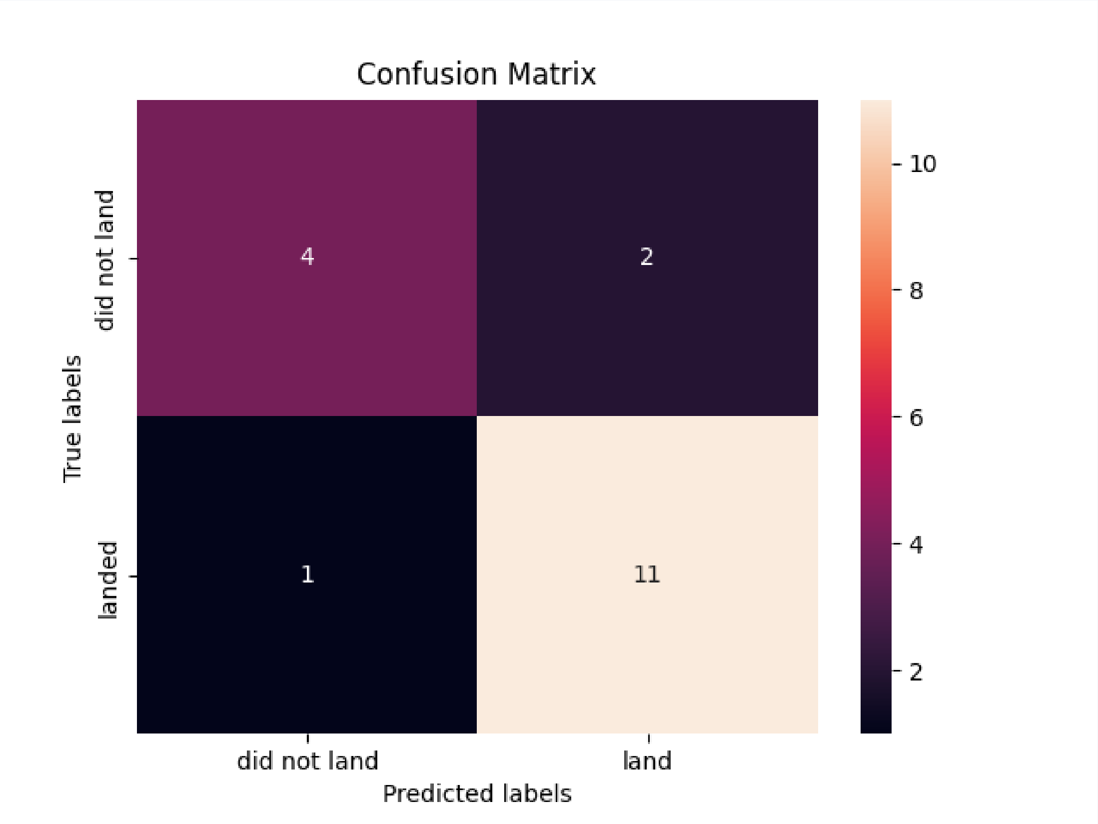
- Visualize the built model accuracy for all built classification models, in a bar chart



- Find which model has the highest classification accuracy
  - Tree model had highest training accuracy and similar test accuracy

Confusion Matrix

- confusion matrix of the best performing model (tree model). It had best training accuracy and similar test accuracy



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%.
- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)
- Success rate since 2013 kept increasing till 2020
- The first successful landing outcome on ground pad was on 22nd December 2015
- Generated Dashboard to analyze data visually
- Accuracy of predictive model is 83%

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!