IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tamir Chong
6/1/2022

Outline

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

Executive Summary

- Collection of data from the SpaceX API only including Falcon 9 Launches, web scraping to collect Falcon 9 historical launch records from a Wikipedia.
- Some SQL analysis was performed, folium maps and a dashboard were created to have a better understanding of the data
- 4 classification models were trained and compared to select the best one
- A Decision Tree model would be the best option to predict the first stage landing with variables like payload and launch site

Introduction

- SpaceX is an American space manufacturer, a provider of space transportation services, and a communications corporation. SpaceX has the goal of reducing space transportation costs to enable the colonization of Mars.
- SpaceX originally intended to follow its light Falcon 1 launch vehicle with an intermediate capacity vehicle, the Falcon 5. The company instead decided in 2005 to proceed with the development of the Falcon 9, a reusable heavier lift vehicle.
- The company 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.

- 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

Section 1

Methodology

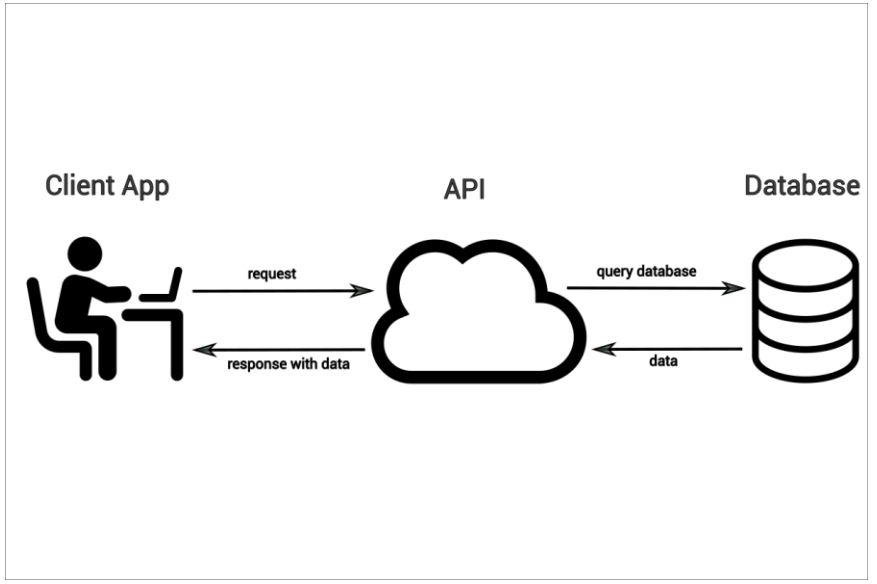
Methodology

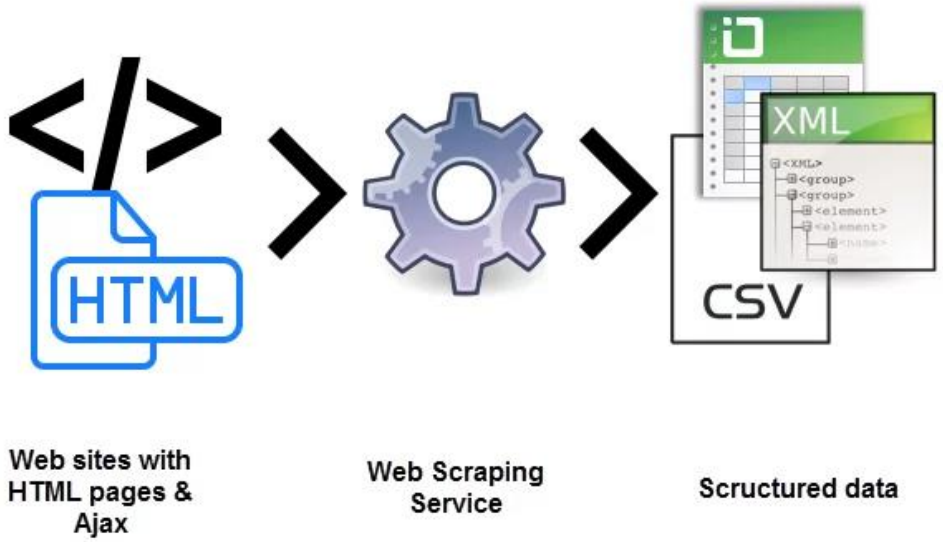
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

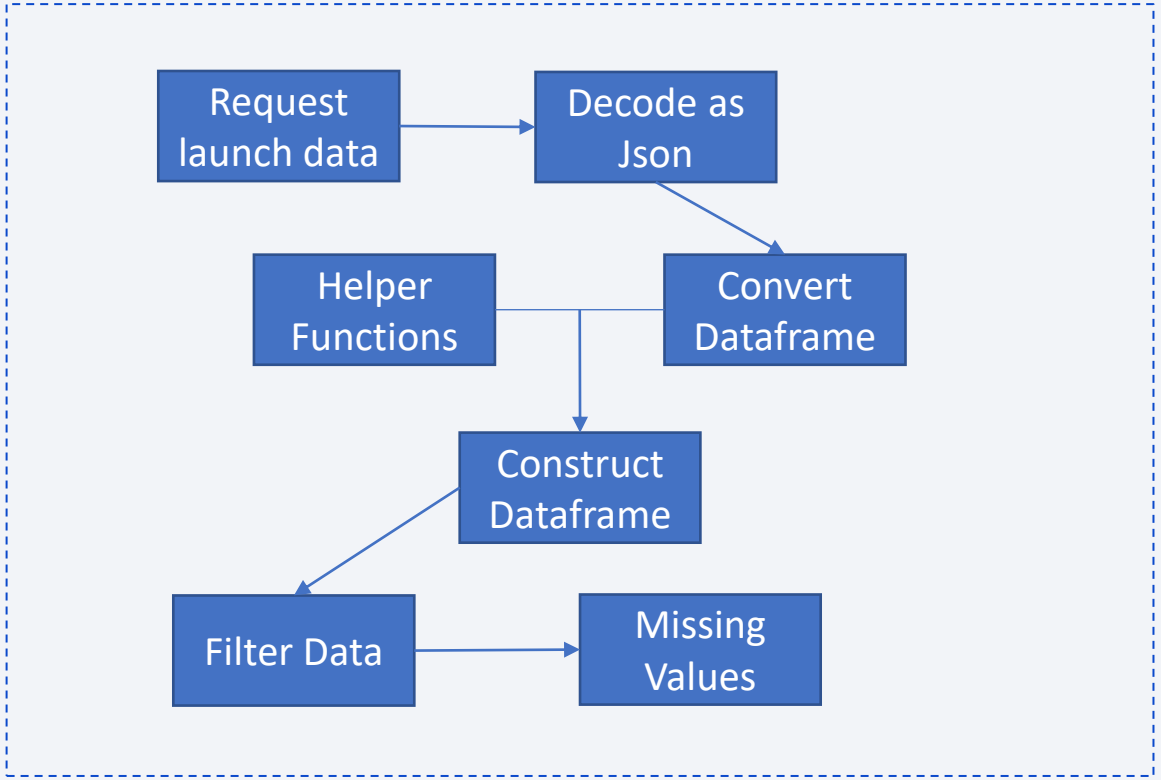
- Collecting data from the SpaceX API and making sure the data is in correct format to create a dataframe that only includes Falcon 9 Launches.
- Performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches





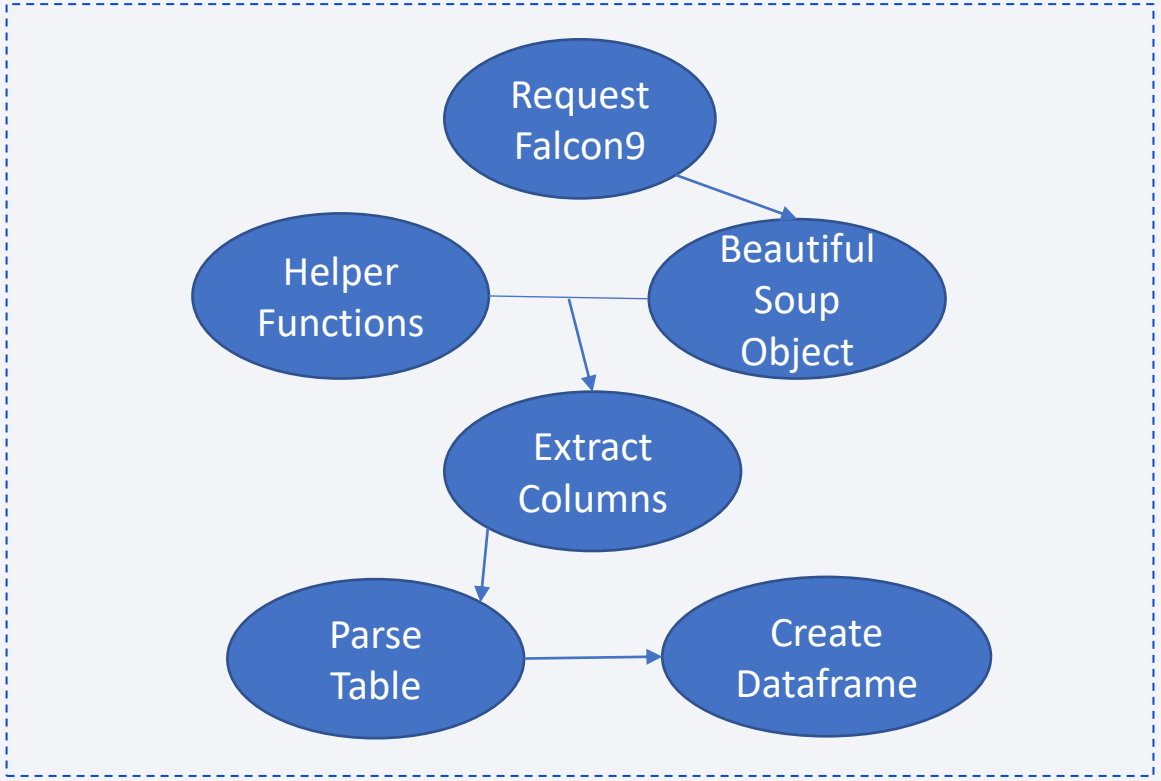
Data Collection – SpaceX API

- SpaceX URL:
  - https://api.spacexdata.com/v4/launches/past
- GitHub URL:
  - https://github.com/geofias/Applied-Data-Science-Capstone/blob/02efddf397990ddc7953987bcc3ec31859b93ae9/jupyter-labs-spacex-data-collection-api.ipynb

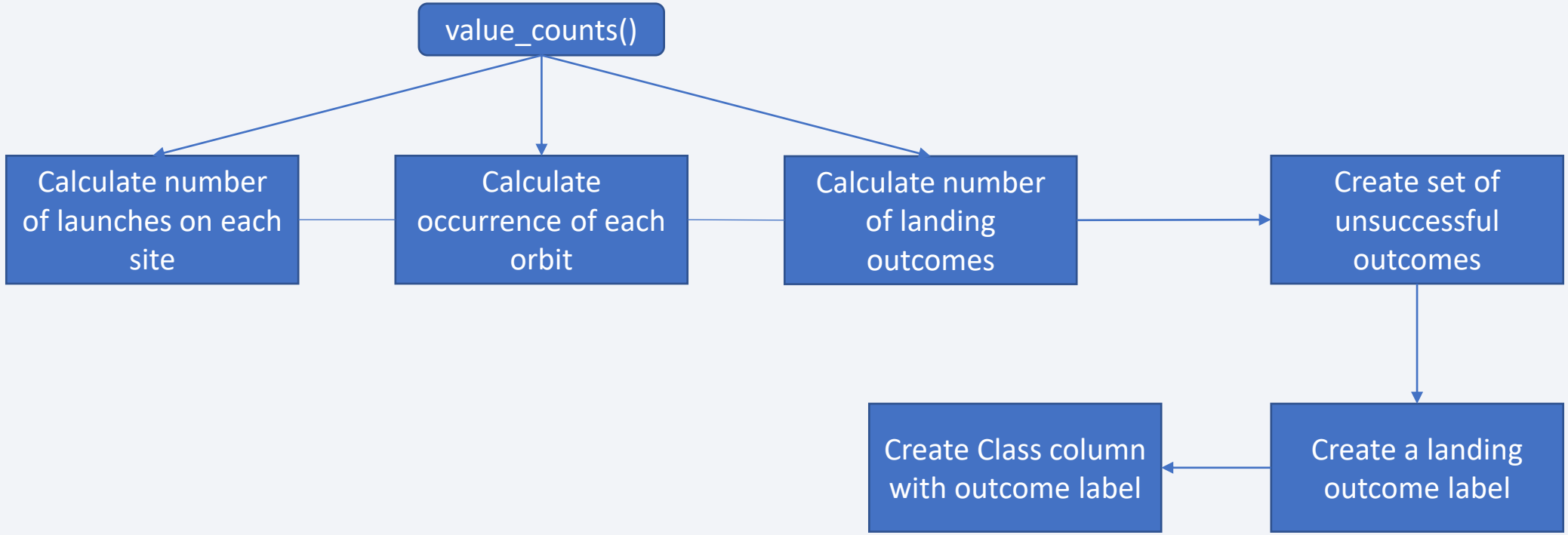


Data Collection - Scraping

- Wikipedia URL of the List of Falcon 9 and Falcon Heavy launches:
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- GitHub URL:
  - https://github.com/geofias/Applied-Data-Science-Capstone/blob/15ebe90d5c647f154d69fbeb1694a03ddc4424aa/jupyter-labs-webscraping.ipynb



Data Wrangling



- GitHub URL:
  - https://github.com/geofias/Applied-Data-Science-Capstone/blob/15ebe90d5c647f154d69fbeb1694a03ddc4424aa/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **FlightNumber vs PayloadMass (Scatter Plot):** Visualize the relationship between the two variables and see how would they affect the launch outcome.
- **FlightNumber vs LaunchSite (Scatter Plot):** Visualize the relationship between Flight Number and Launch Site.
- **Payload vs Launch Site (Scatter Plot):** Observe if there is any relationship between Launch Sites and their Payload mass.
- **Orbit vs Class (Bar Chart):** Check if there are any relationship between Success Rate and Orbit type.
- **FlightNumber vs Orbit (Scatter Plot):** For each orbit, I want to see if there is any relationship between FlightNumber and Orbit type.
- **Payload vs Orbit (Scatter Plot):** Reveal the relationship between Payload and Orbit type.

GitHub URL: https://github.com/geofias/Applied-Data-Science-Capstone/blob/15ebe90d5c647f154d69fbeb1694a03ddc4424aa/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015
- Ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL: https://github.com/geofias/Applied-Data-Science-Capstone/blob/15ebe90d5c647f154d69fbeb1694a03ddc4424aa/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Explain why you added those objects
- Circle at NASA Johnson Space Center's coordinate to represent the site on the map
- Marker to create an icon as a text label showing the NASA Johnson Space Center's name
- Circles representing different launch sites on the map
- Markers with the name of each launch site as icons on the map
- Maker Clusters to represent success/failed launches for each site on the map
- Lines showing the distance between launch sites to their proximities

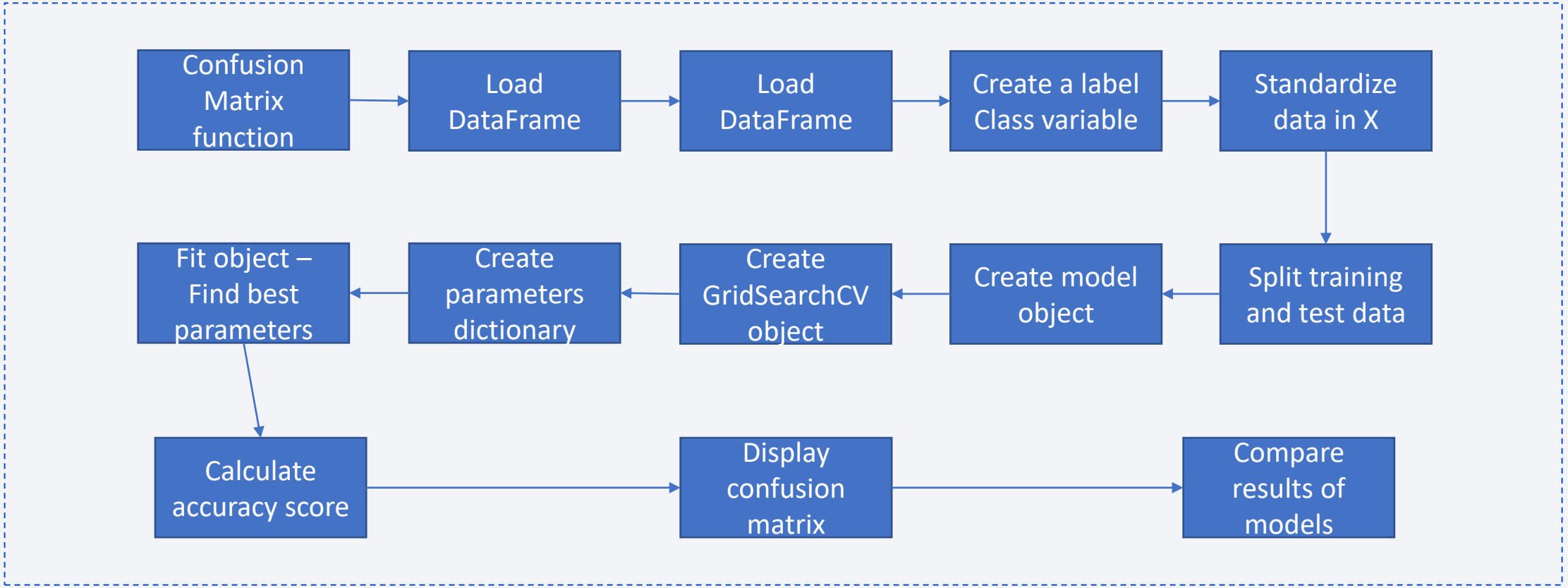
GitHub URL: https://github.com/geofias/Applied-Data-Science-Capstone/blob/15ebe90d5c647f154d69fbeb1694a03ddc4424aa/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart to display the total success launches by all sites and the rate success/failure if a specific site is selected
- Scatter Plot showing the relationship between Class and Payload Mass variables were can try different ranges of payload mass and it's all classified by the Booster Version

GitHub URL: https://github.com/geofias/Applied-Data-Science-Capstone/blob/117b6059e0a9c4b4de1e6a19798c3d3c3087aabc/Plotly-Dash-dashboard.ipynb

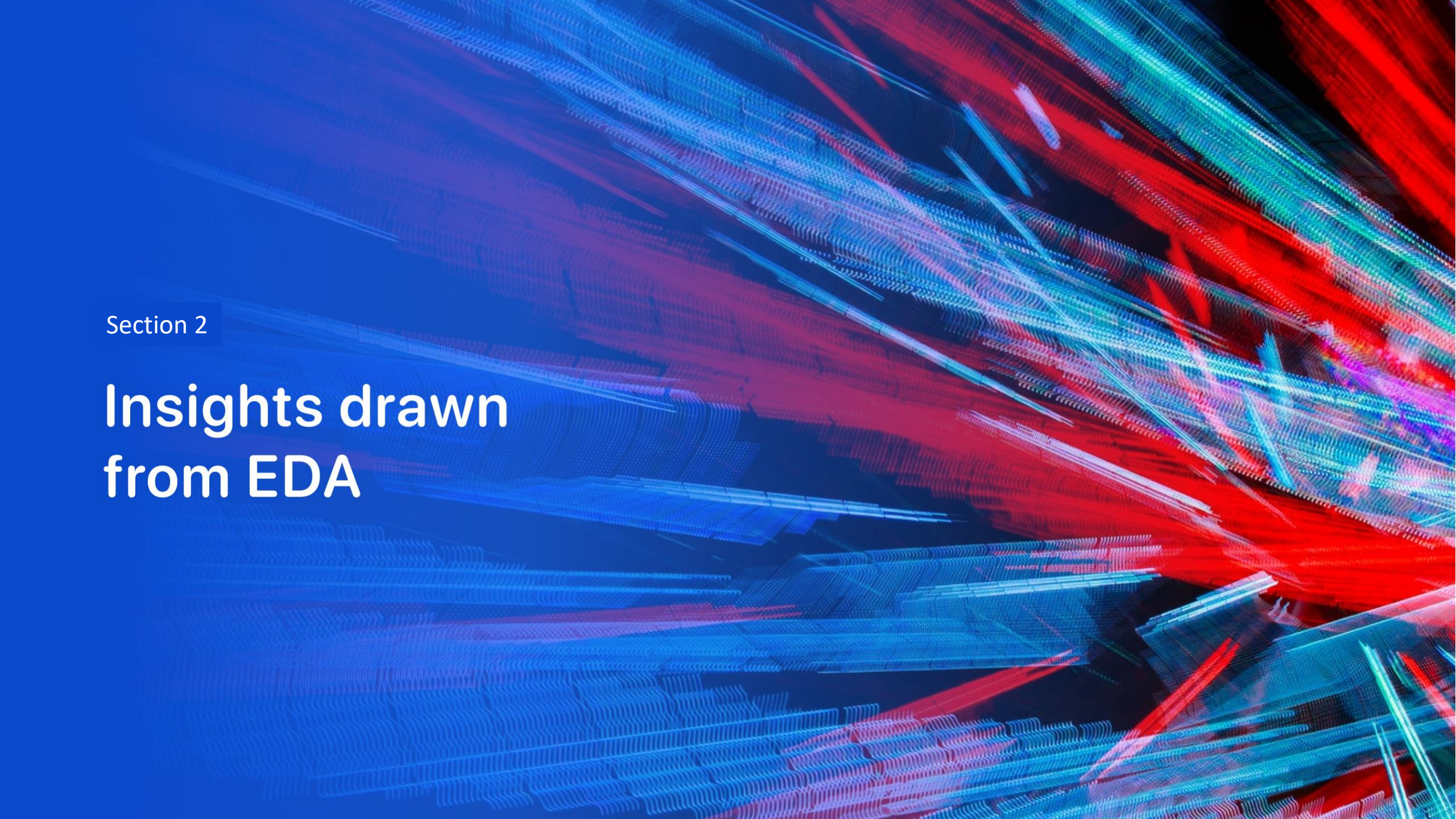
Predictive Analysis (Classification)



GitHub URL: https://github.com/geofias/Applied-Data-Science-Capstone/blob/117b6059e0a9c4b4de1e6a19798c3d3c3087aabc/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

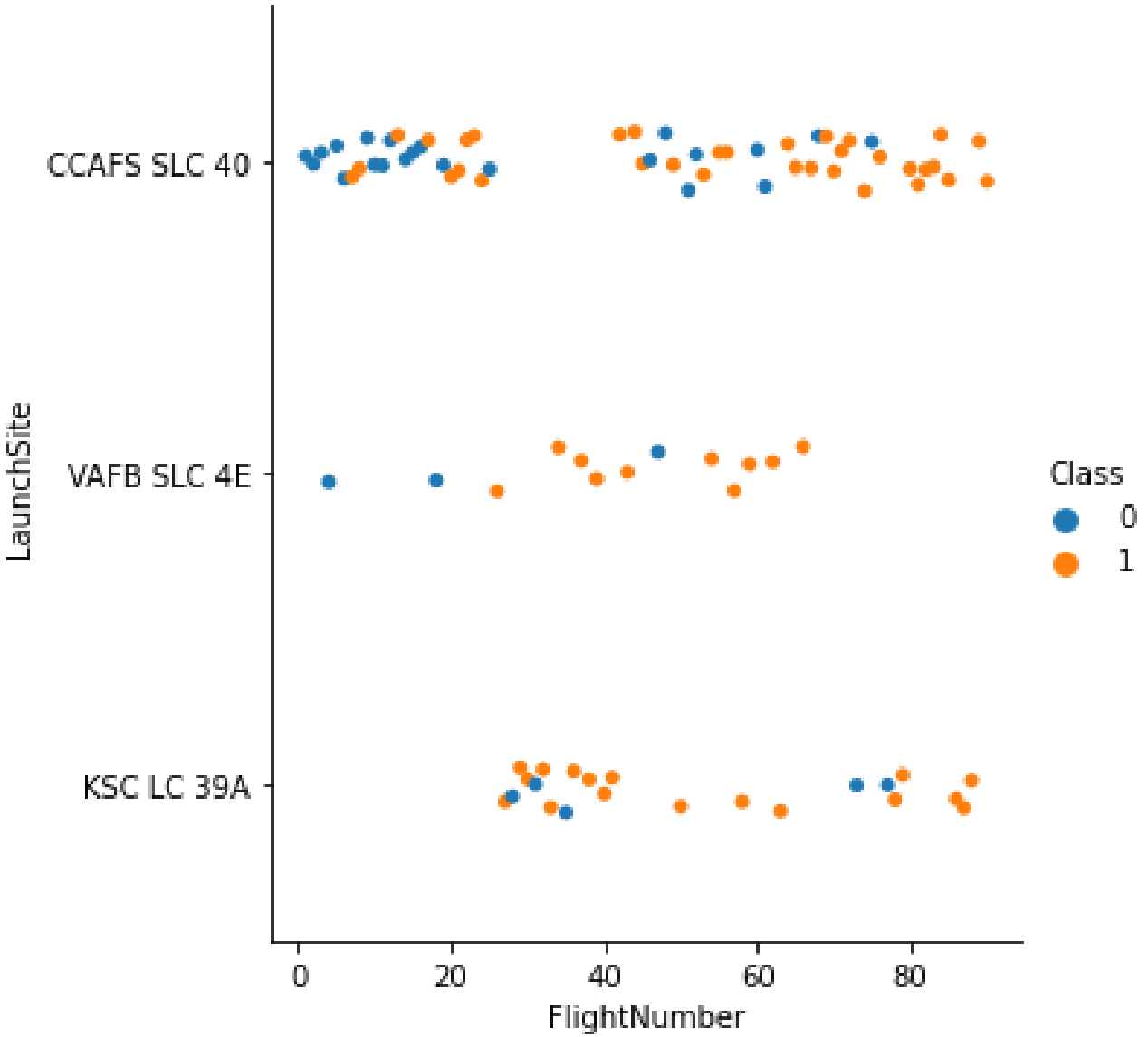
Results

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


Section 2
Insights drawn from EDA

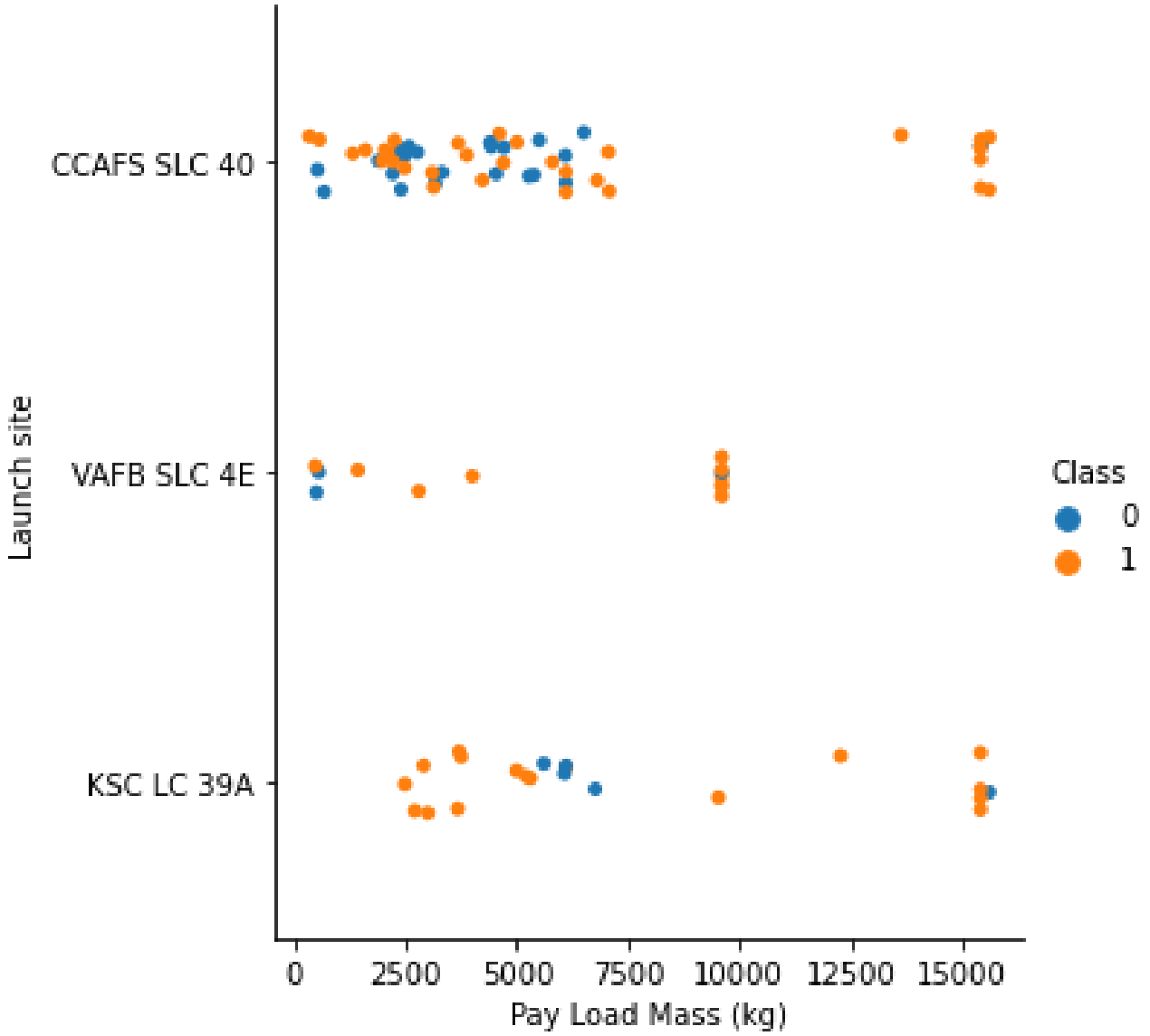
Flight Number vs. Launch Site

- As the flight number increases, the first stage is more likely to land successfully being CCAFS SLC 40 the launch site with more successful landings as flight number increases



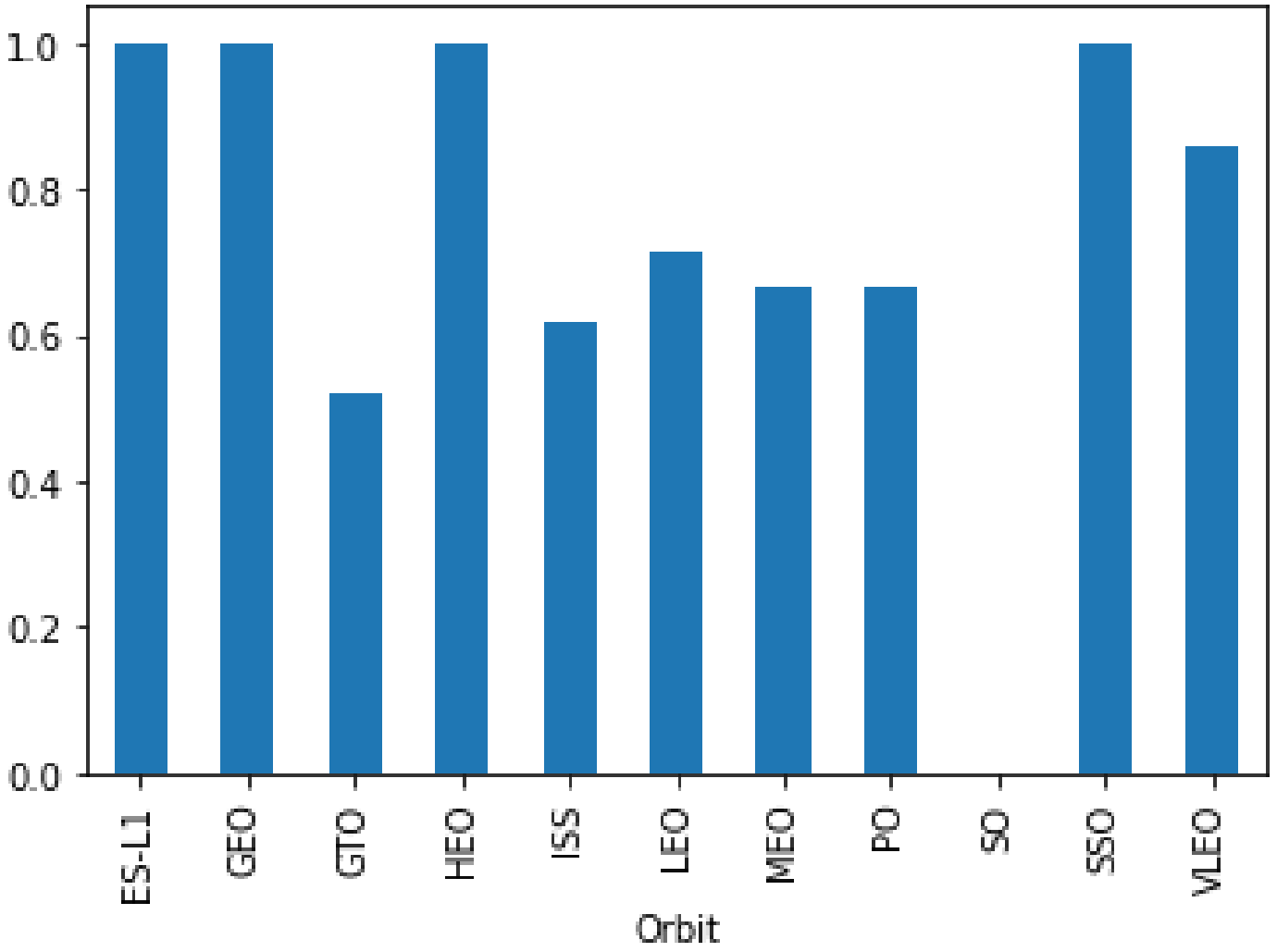
Payload vs. Launch Site

- For the VAFB-SLC there are no rockets launched for heavypayload mass (greater than 10000)
- For CCAFS SLC 40 and KSC LC 39A the first stage is more likely to land successfully with a payload mas greater than 15000



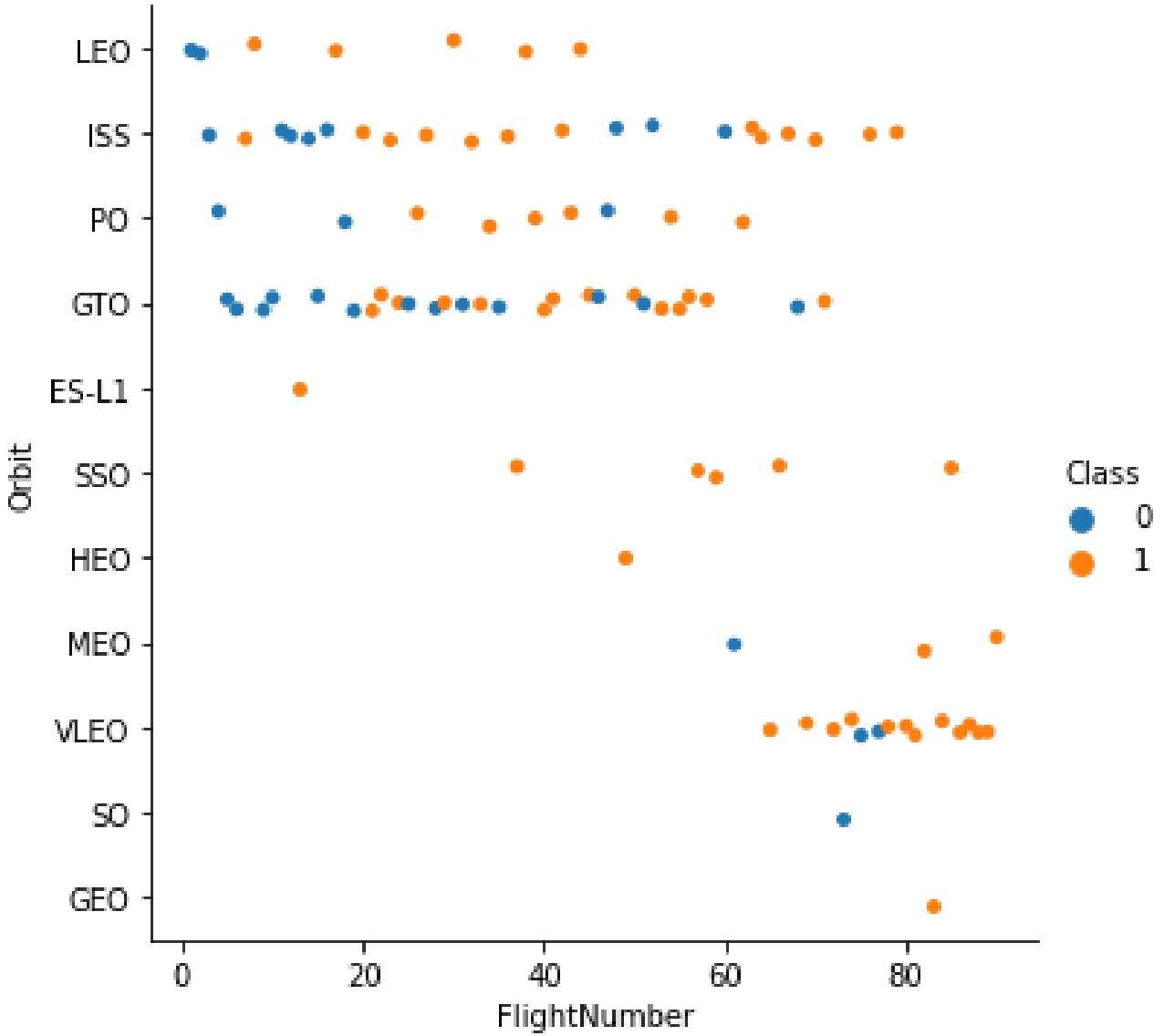
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits have the greatest success rate of 1
- VLEO has a very high success rate of 0.85 and so



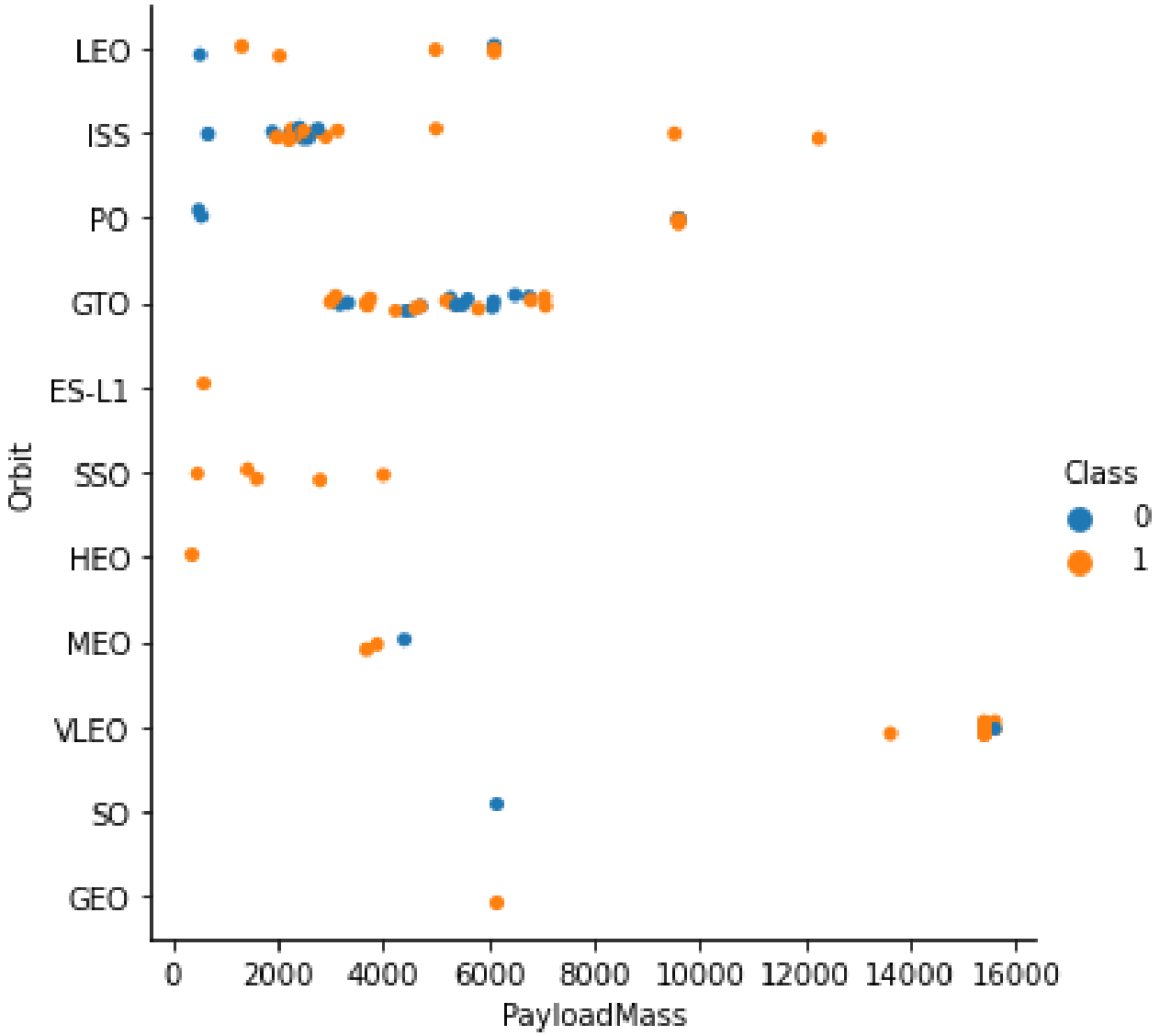
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit



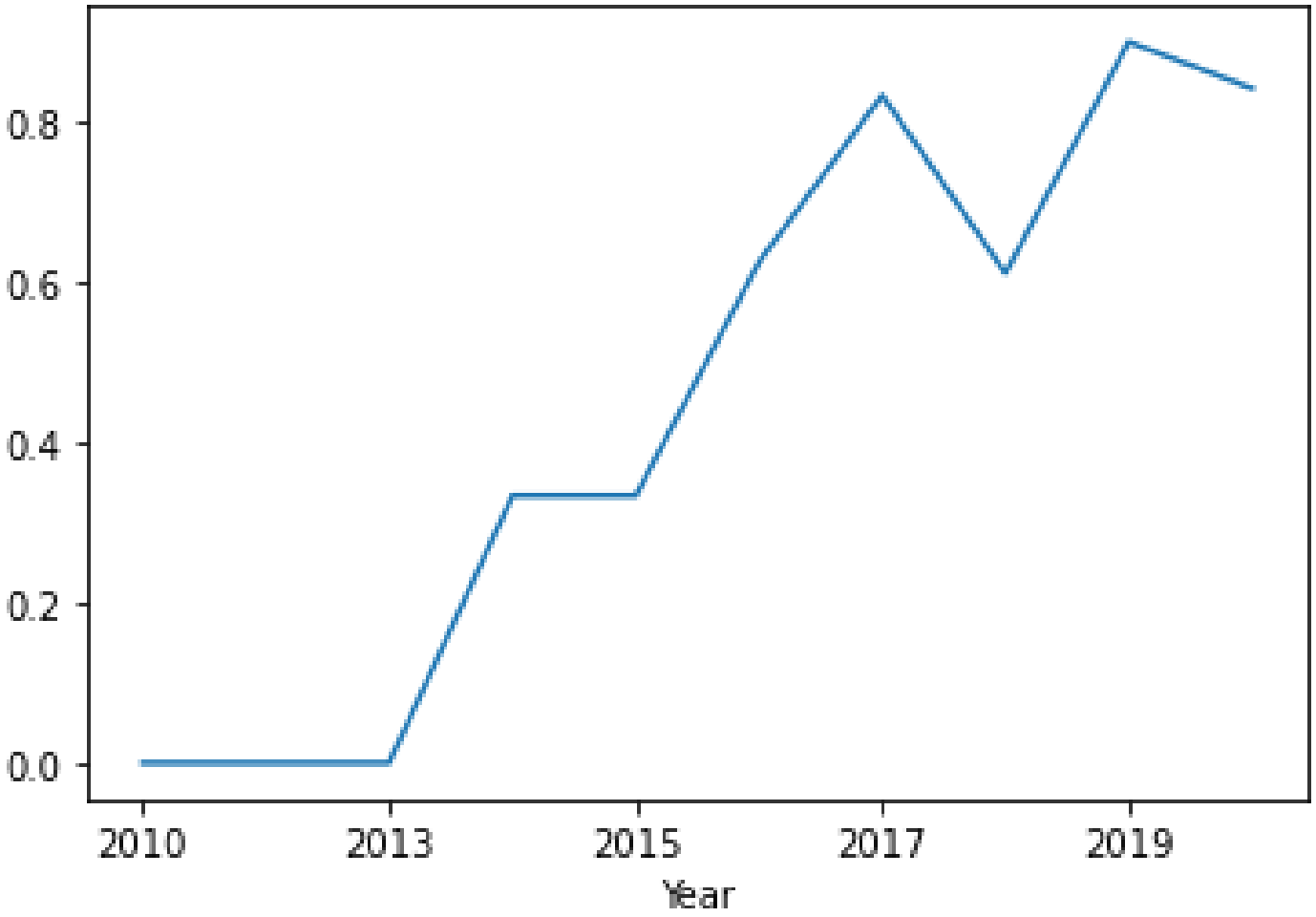
Payload vs. Orbit Type

- With heavy payloads the successful landing rate are more for PO, LEO and ISS
- For GTO we cannot distinguish this well as both positive landing rate and negative landing are both there



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020
- There seem to be a drop between 2017 and 2020 but the trend is positive



All Launch Site Names

Names of the unique launch sites:

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

Used DISTINCT and then selected the column LAUCNH_SITE from SPACEXTBL

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with “CCA”:

- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40
- CCAFS LC-40

Selected the LAUNCH_SITE column from SPACEXTBL then apply were for launch sites that begin with “CCA” and finally limited the result to 5

Total Payload Mass

Total payload carried by boosters from NASA = 45596 kg

Did a sum on the payload_mass_kg column were the customer was equal to “NASA (CRS)”

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 = 2928 kg

Used the AVG function on the payload_mass_kg column were the booster version was equal to “F9 v1.1”

First Successful Ground Landing Date

Date of the first successful landing outcome on ground pad:

- 2015-12-22

Used the MIN function on the DATE column were landing outcome is “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Selected the column booster_version from SPACEXTBL were the landing outcomes where equal to “Success (drone ship)” and payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes:

- Failure (in flight): 1
- Success: 99
- Success (payload status unclear): 1

Selected the mission_outcome column, used the COUNT() function to count number of successful and failure mission outcome while grouping the data by the mission_outcome column

Boosters Carried Maximum Payload

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

- Selected the booster_version column from SPACEXTBL were payload is at its maximum (used a subquery)

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Selected the columns landing__outcome, booster_version, launch_site from SPACEXTBL were landing outcome is “Failure (drone ship)” and date is in year “2015”

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

Selected the landing__outcome column and applied the COUNT function to count the landing outcomes between the date 2010-06-04 and 2017-03-20, grouped by the landing outcome and ordered by the counted values in descending order

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

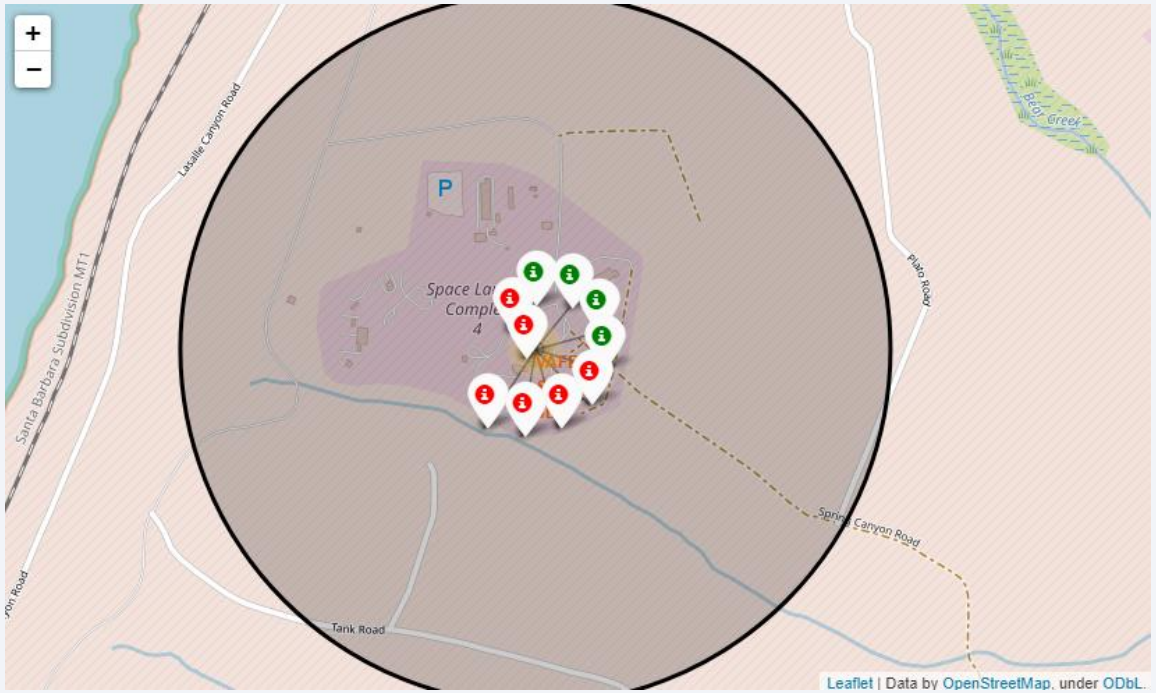
Section 3

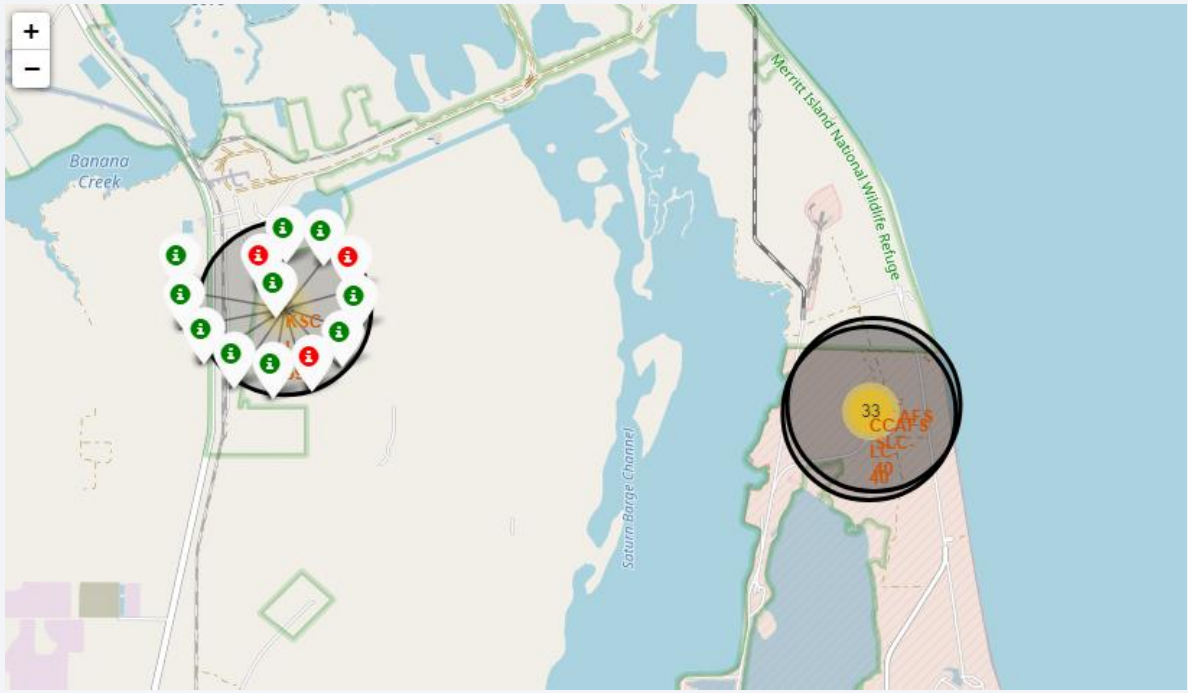
Launch Sites Proximities Analysis

Launch Site Locations

- All launch sites are close to the coasts
- They are all far from the Equator line
- KSC LC-39A, CCAFS LC-40, CCAFS SLC-40 are very close to each other

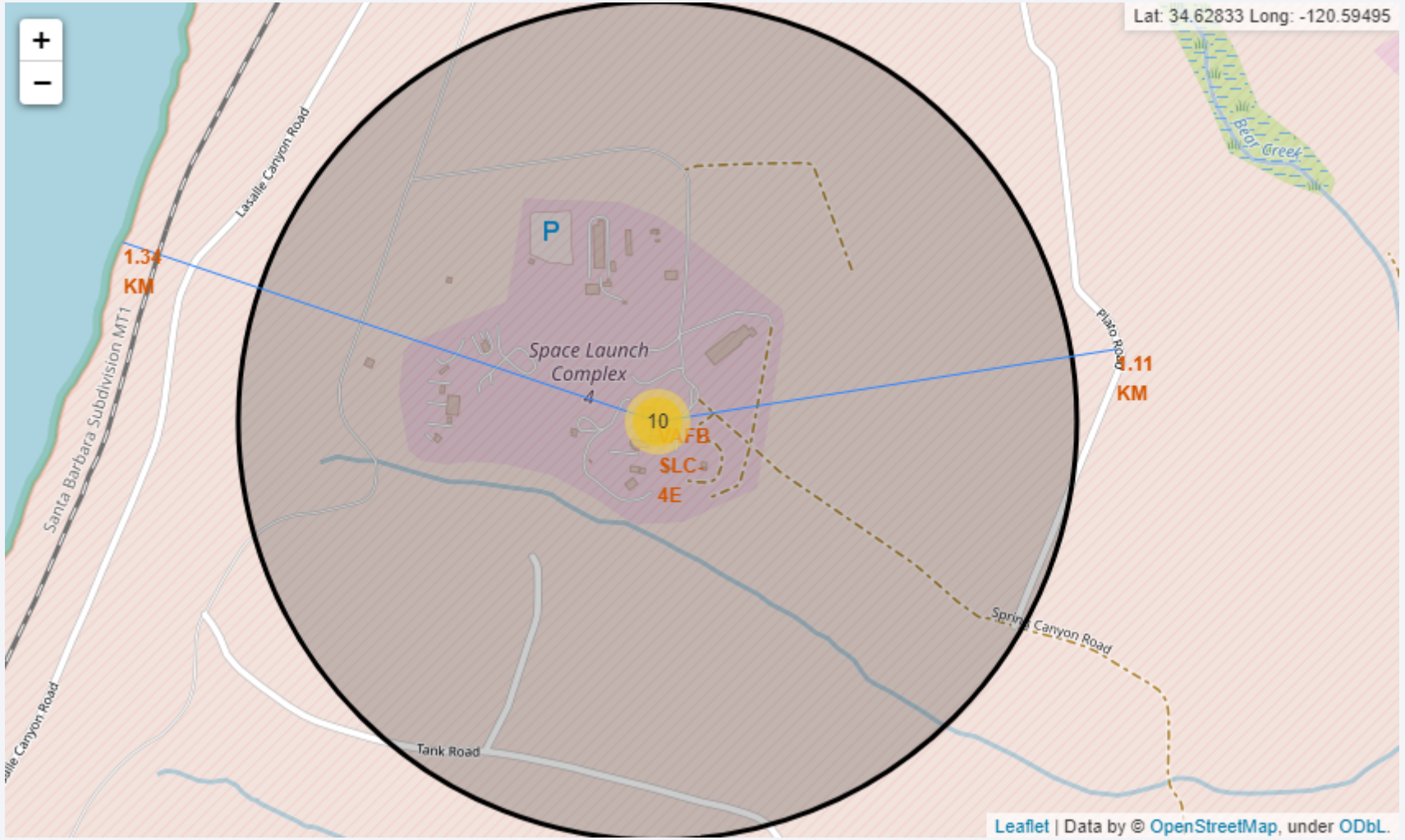
SpaceX Launch Outcomes



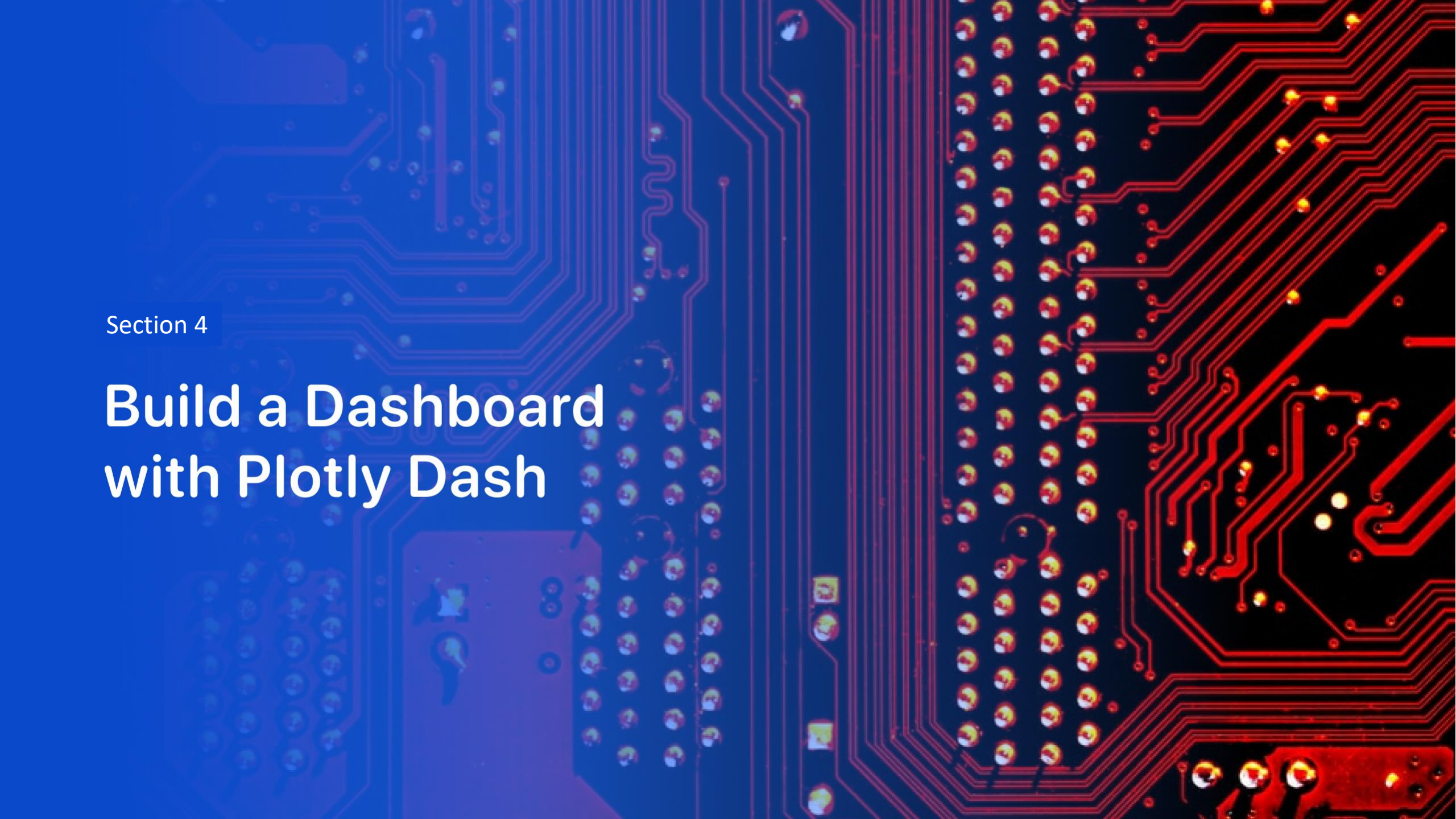


- The KSC LC-39A lunch site has the higher success rate

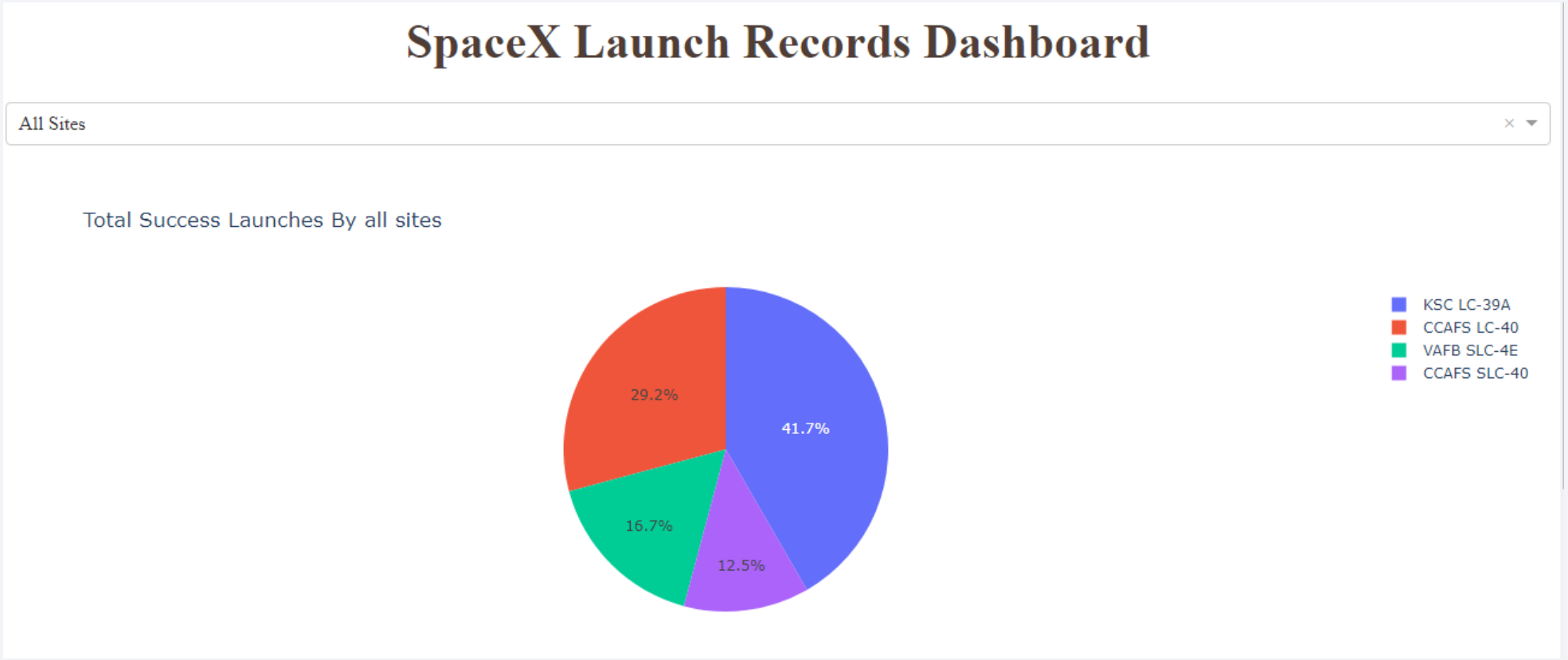
Proximities of KSC LC-39A launch site



- The launch site is close to highways
- The launch site is close to the coastline
- This launch is very far from cities

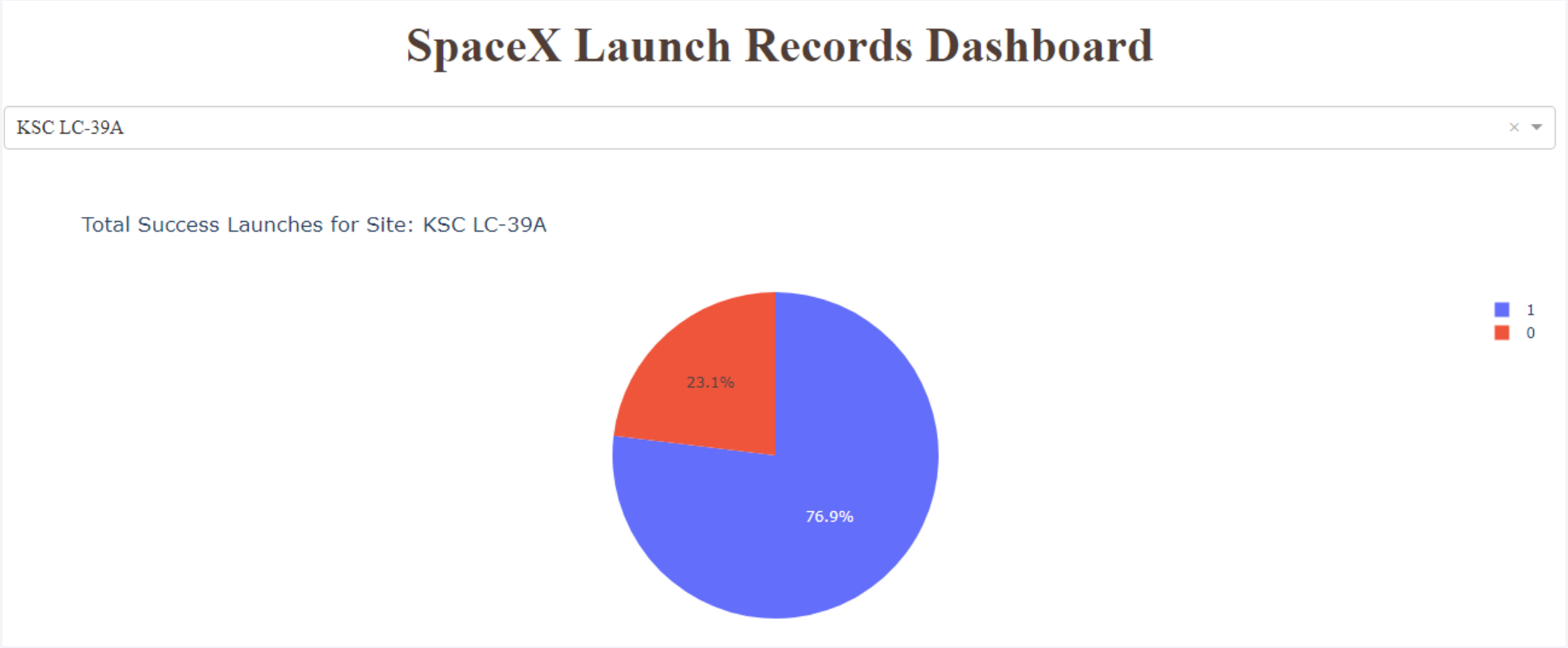

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



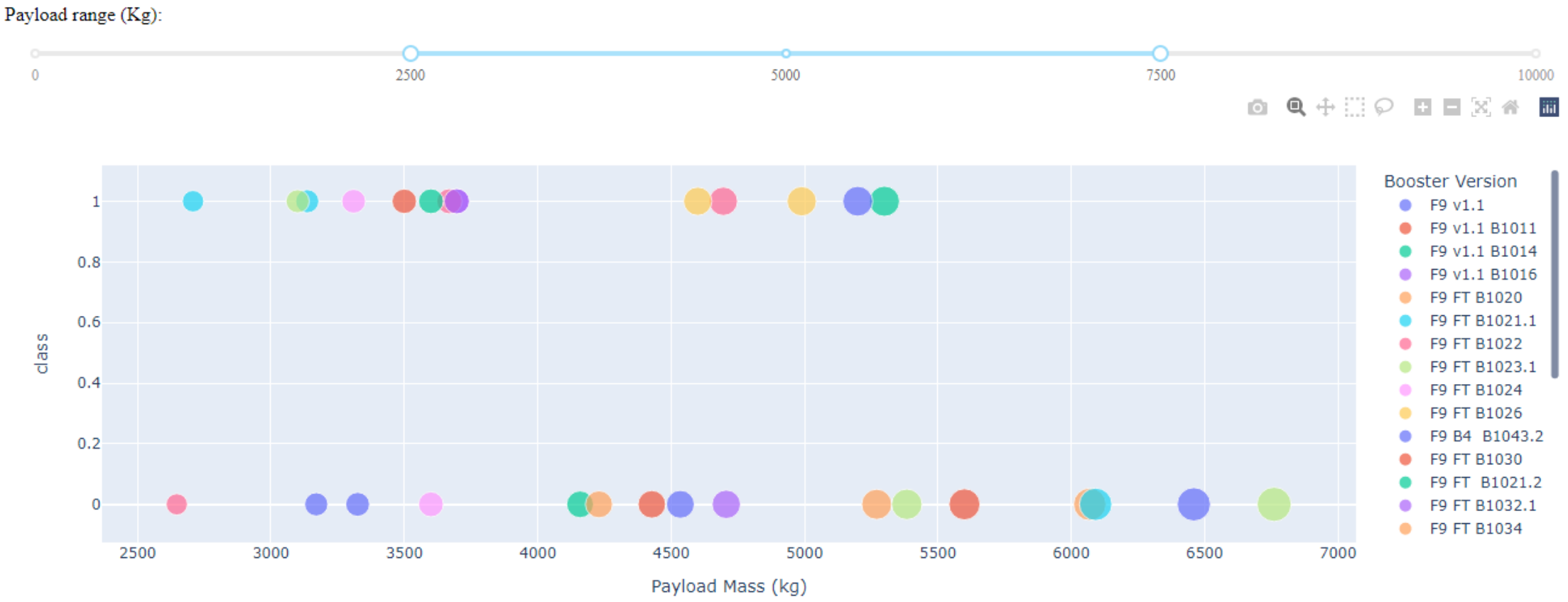
- The KSC LC-39A launch site has the highest success rate
- The CCAFS SLC-40 launch site has the lower success rate

KSC LC-39A Launch Site Outcome



- 76.9% of the launch site outcome successfully landed
- Only 23.1% of the outcomes were failures

Payload vs. Launch Outcome for all sites



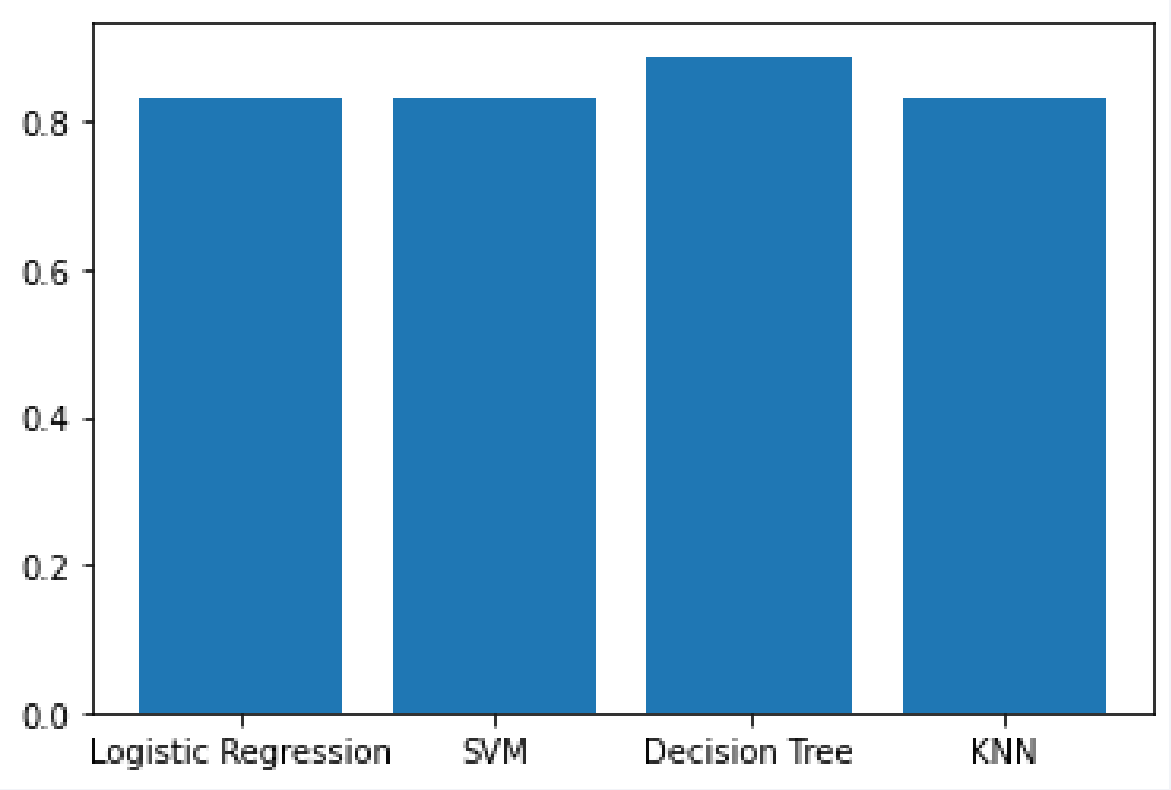
- The (2500-5000) payload range have the largest success range
- The greater the payload the lowest the success rate

Section 5

Predictive Analysis (Classification)

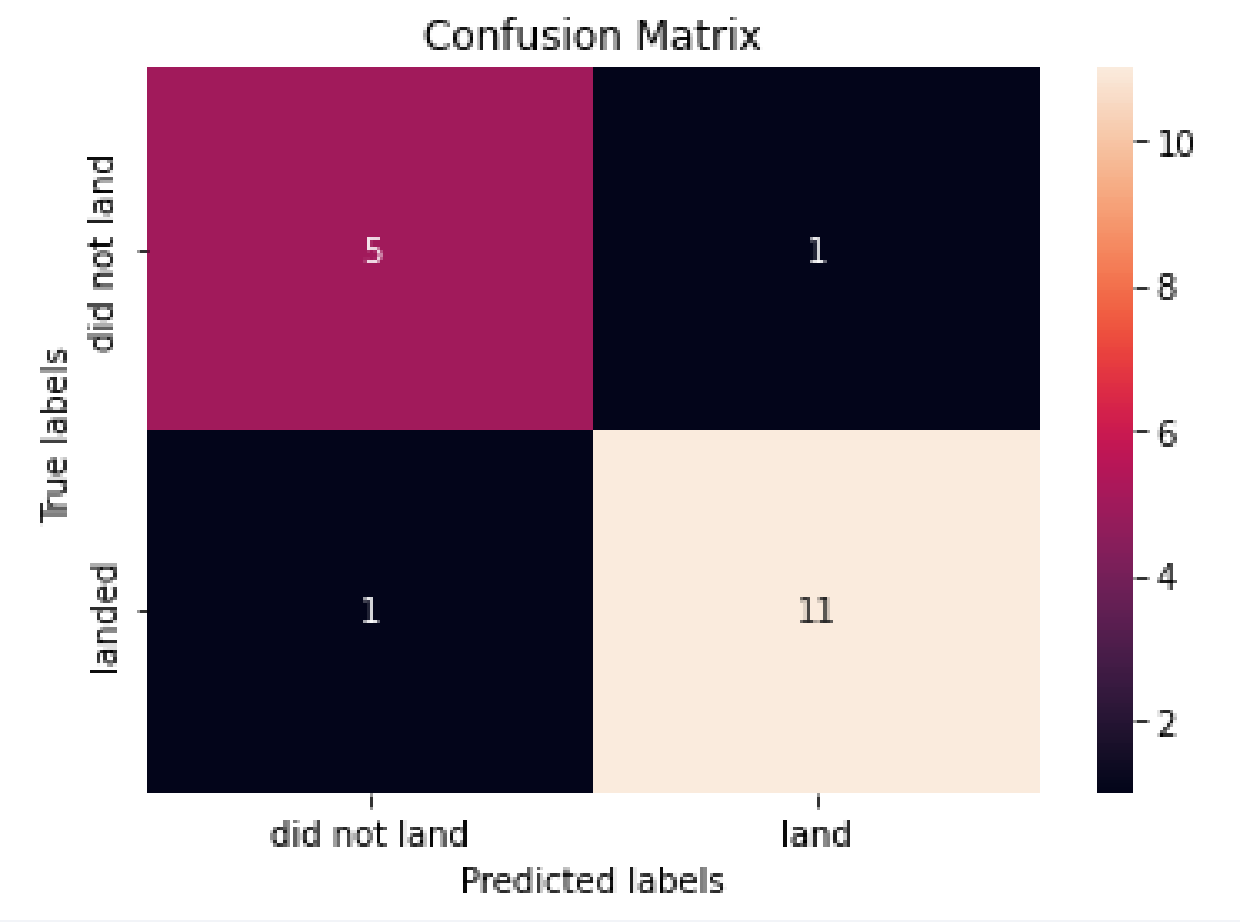
Classification Accuracy

- The Decision Tree model has the highest accuracy



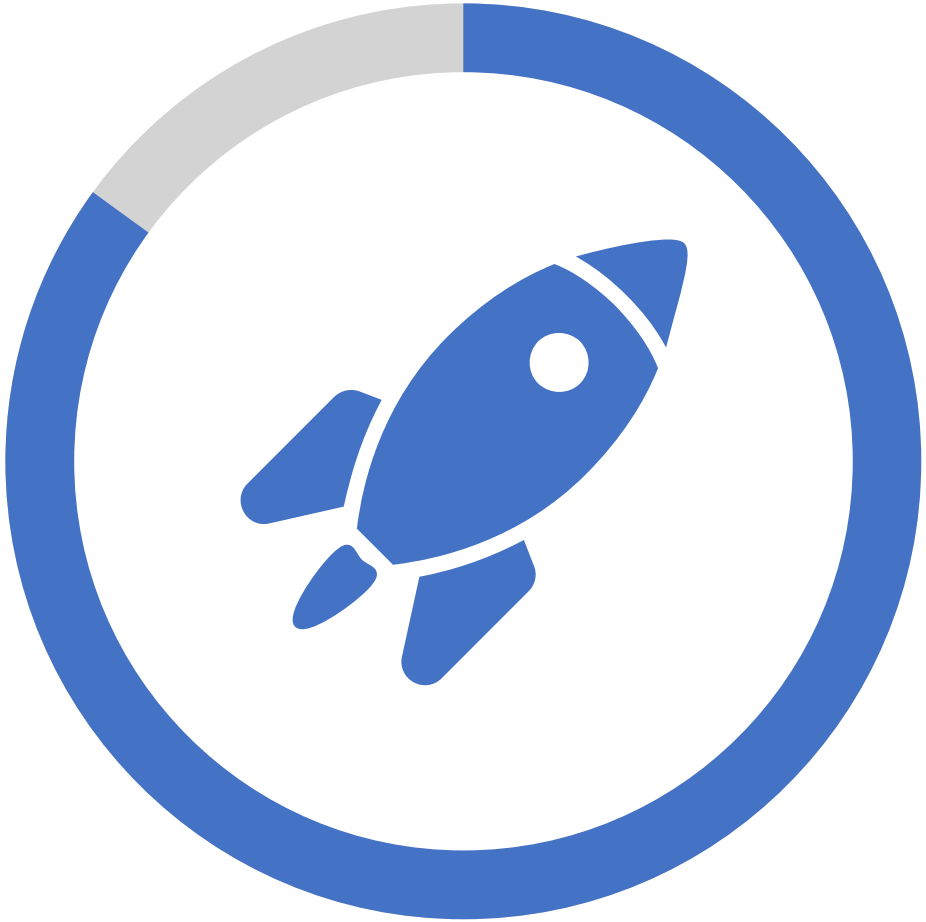
Decision Tree Confusion Matrix

- There is 11 True Positives, that means that there is 11 successful landings
- It shows 5 True Negatives, so there are 5 failure landings

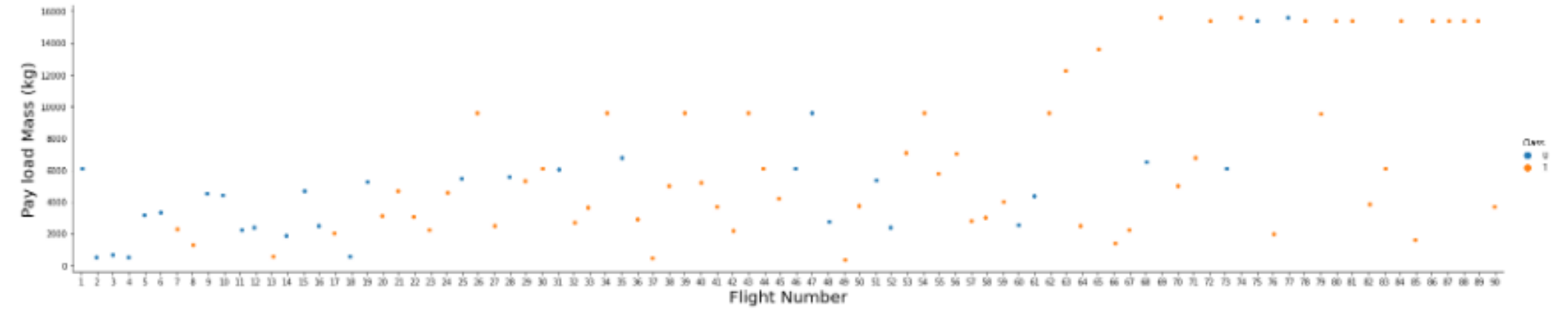


Conclusions

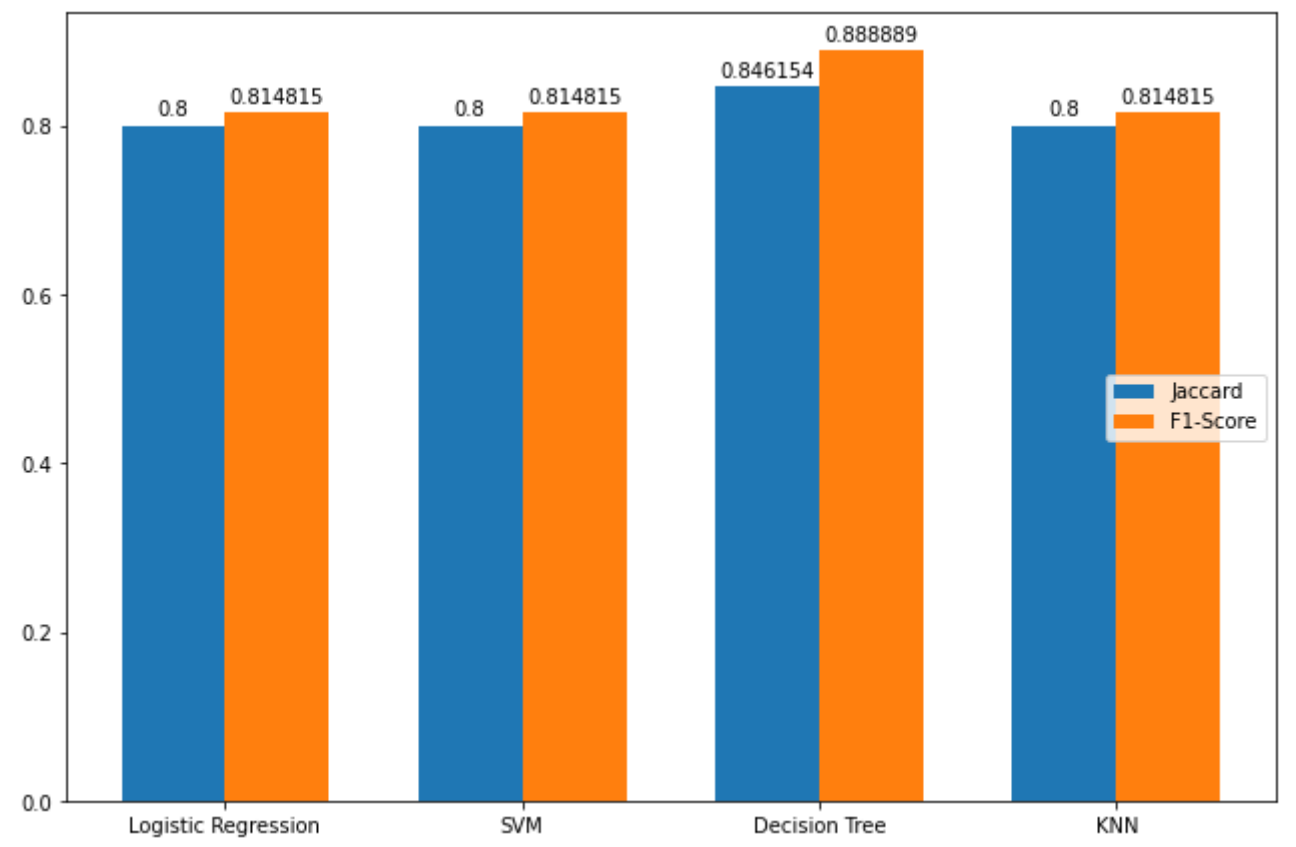
- The KSC LC-39A Launch Site is the site with the highest successful launches
- The payload mass is a determining variable to determine the success of the first stage landing
- VLEO orbit has one of the best results with an 85% success rate
- The best model to predict the first stage landing would be a decision tree



Appendix



Appendix


0.8
0.814815
0.8
0.814815
0.846154
0.888889
0.8
0.814815
0.0
0.2
0.4
0.6
0.8
Jaccard
F1-Score
Logistic Regression
SVM
Decision Tree
KNN

Thank you!