Winning Space Race with Data Science

Nabil T.
6 March 2024

Outline

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

Executive Summary

- Data were collected from two sources, the SpaceX API and web scrapings from the Wikipedia page of the Falcon 9 and Falcon Heavy launches list.
- After we have prepared the data from the API and web scraping, we performed some preliminary EDA and data wrangling
- We performed data analysis and the results are as follows:
  - The most effective models are Logistic Regression, Decision Tree, and KNN
  - Launch sites with higher flight amount tend to have better success rates.
  - Orbits like ES-L1, GEO, HEO, SSO have the highest success rates.

Introduction

- A key factor in SpaceX's success is their cost-effective launches. Their Falcon 9 rockets, significantly cheaper than competitors due to reusable first stages, cost around $62 million compared to the industry average of $165 million. Determining the cost of a launch therefore hinges on predicting whether the first stage will land successfully.
- This project aims to address this challenge for a new rocket company, Space Y, by training a machine learning model to predict reusability based on publicly available information, offering valuable insights for competitive pricing strategies.

Section 1

Methodology

Methodology

Executive Summary

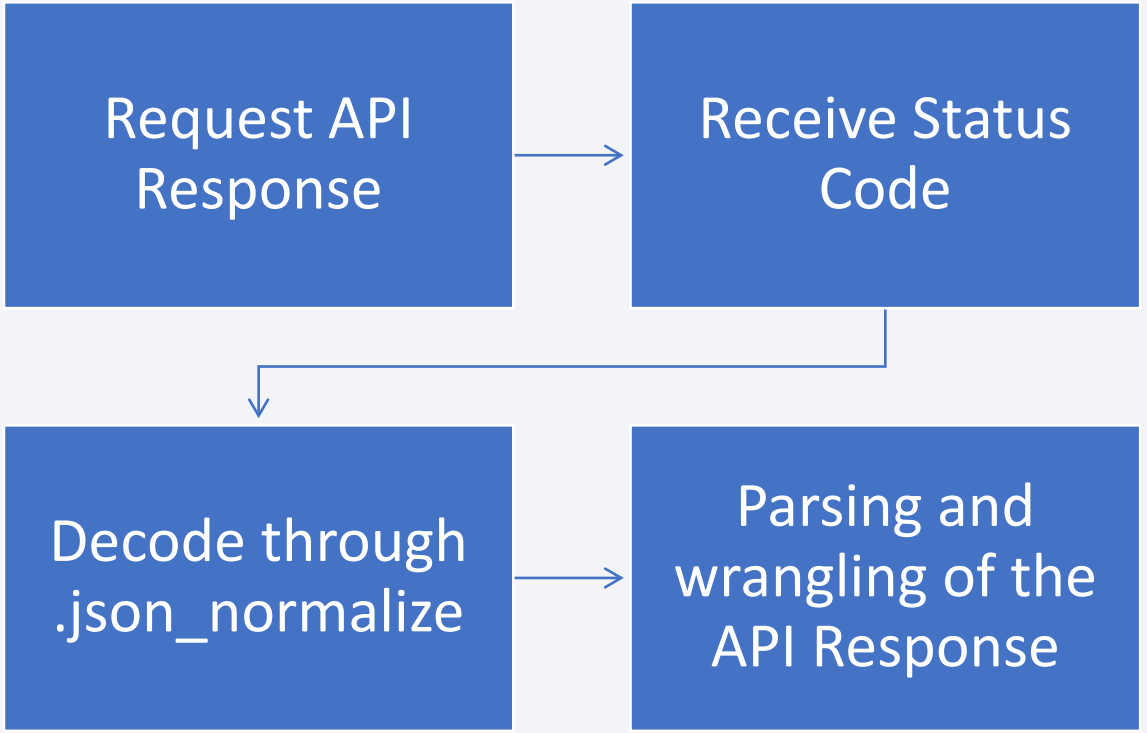
- Data collection methodology:
  - Collected from SpaceX API and web scraping
- Perform data wrangling
  - Using several data wrangling methods
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using 4 methods

Data Collection

- Data were collected from two sources, the SpaceX API and web scrapings from the Wikipedia page of the Falcon 9 and Falcon Heavy launches list.
- The following is the URL to the SpaceX API and the Wikipedia page:
  - https://api.spacexdata.com/v4/launches/past
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- On the next pages I will explain the high-level process.

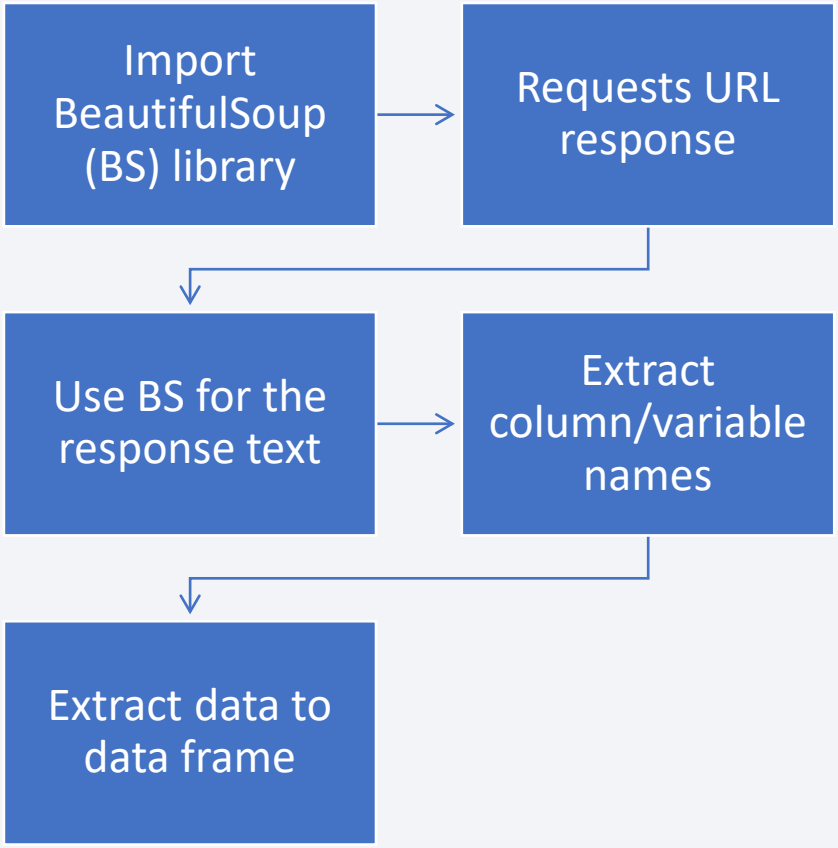
Data Collection – SpaceX API

- Prior to the API request, we defined several functions to help us later with the data parsing.
- We request the API response and conduct data parsing and wrangling.
- Github URL: IBM_DS_Capstone_Project/jupyter-labs-spacex-data-collection-api.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)



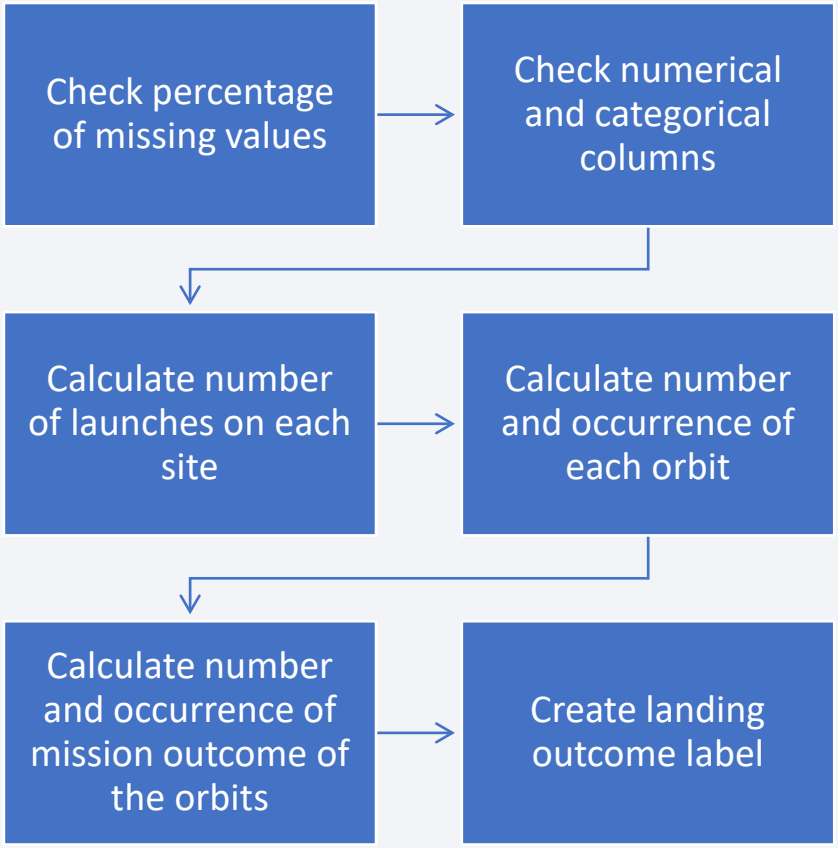
Data Collection - Scraping

- We use the BeautifulSoup library to do the web scraping
- Github URL: IBM_DS_Capstone_Project/jupyter-labs-webscraping.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)



Data Wrangling

- After we have prepared the data from the API and web scraping, we performed some preliminary EDA and data wrangling
- In the end, we prepare the outcome label from the Outcome column
- Github URL: [IBM_DS_Capstone_Project/labs-jupyter-spacex-Data wrangling.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)]



EDA with Data Visualization

- We used several charts to analyze the data, as follows:
  - **Scatter plot:** To see the relationship between two variables (e.g., FlightNumber vs Payload)
  - **Categorical scatter plot:** To see the relationship between a numberical and categorical variable (e.g., LaunchSite vs FlightNumber)
  - **Bar chart:** To see the magnitude of numerical amount of a categorical variable (e.g., success rate mean of each orbit)
  - **Line chart:** To see time trends (e.g., launch success yearly trend)
- Github URL: IBM_DS_Capstone_Project/jupyter-labs-eda-dataviz.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)

EDA with SQL

- We have performed SQL queries, as follows:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the records displays the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
  - 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.
- Github URL: IBM_DS_Capstone_Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)

Build an Interactive Map with Folium

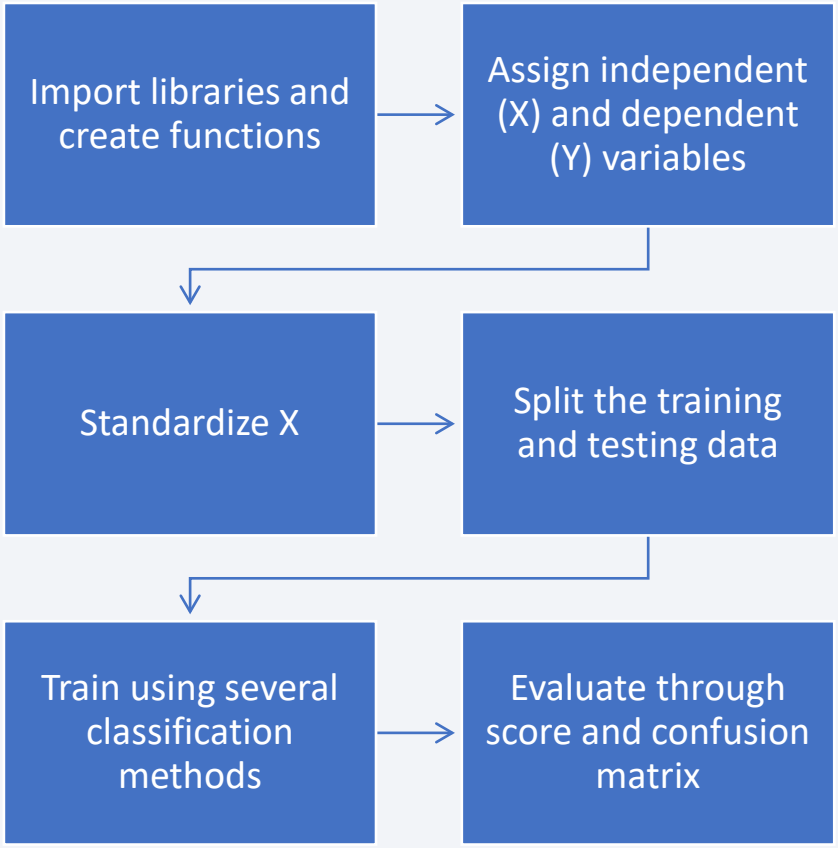
- We uses the following map objects in Folium:
  - **Circle:** To mark the launch sites
  - **Marker:** To add markers
  - **MarkerCluster:** To cluster multiple markers
  - **Lines:** To show distance between places
- Github URL: IBM_DS_Capstone_Project/lab_jupyter_launch_site_location.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)

Build a Dashboard with Plotly Dash

- On the Dash Application, we have added several plots:
  - **Pie chart:** To see the proportion of successful launches in launch sites
  - **Scatter plot:** To see the relationship between Payload and the Success Rate
- Furthermore, we also added dropdowns and sliders, as follows:
  - **Dropdown:** To filter between launch sites and updates the pie chart
  - **Slider:** To update the scatter plot payload range
- Github URL: IBM_DS_Capstone_Project/spacex_dash_app.py at main · ntegar/IBM_DS_Capstone_Project (github.com)

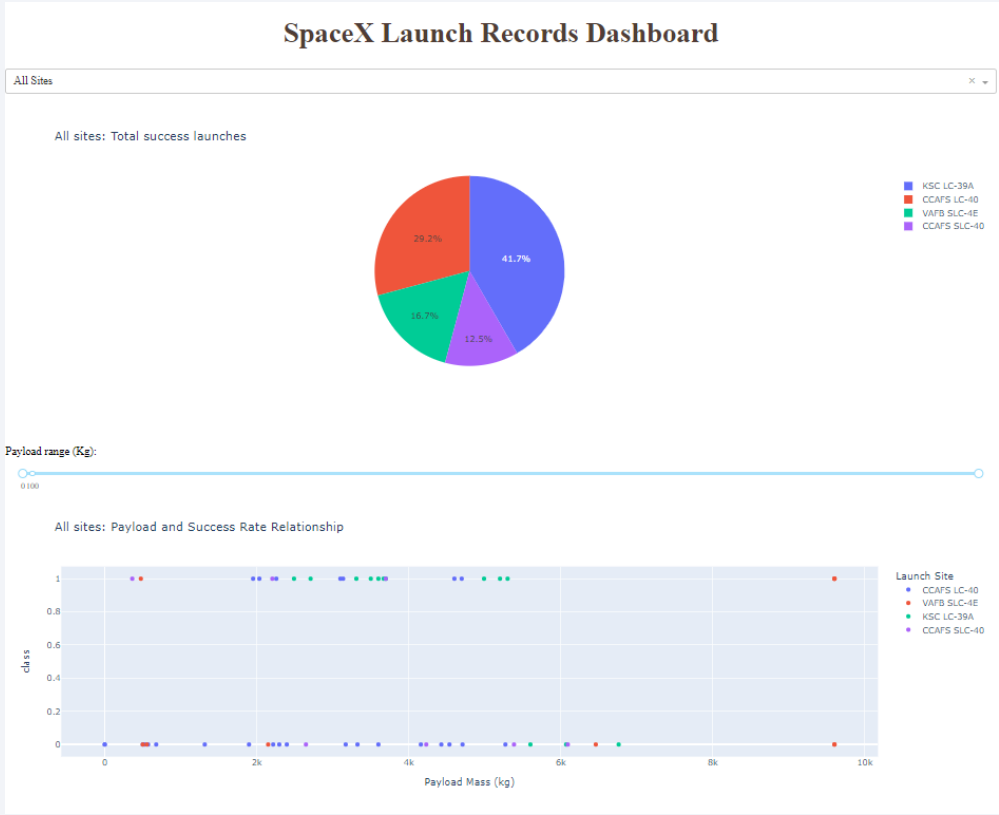
Predictive Analysis (Classification)

- Finally, we developed classification models using several methods:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Github URL:
  IBM_DS_Capstone_Project/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · ntegar/IBM_DS_Capstone_Project (github.com)



Results

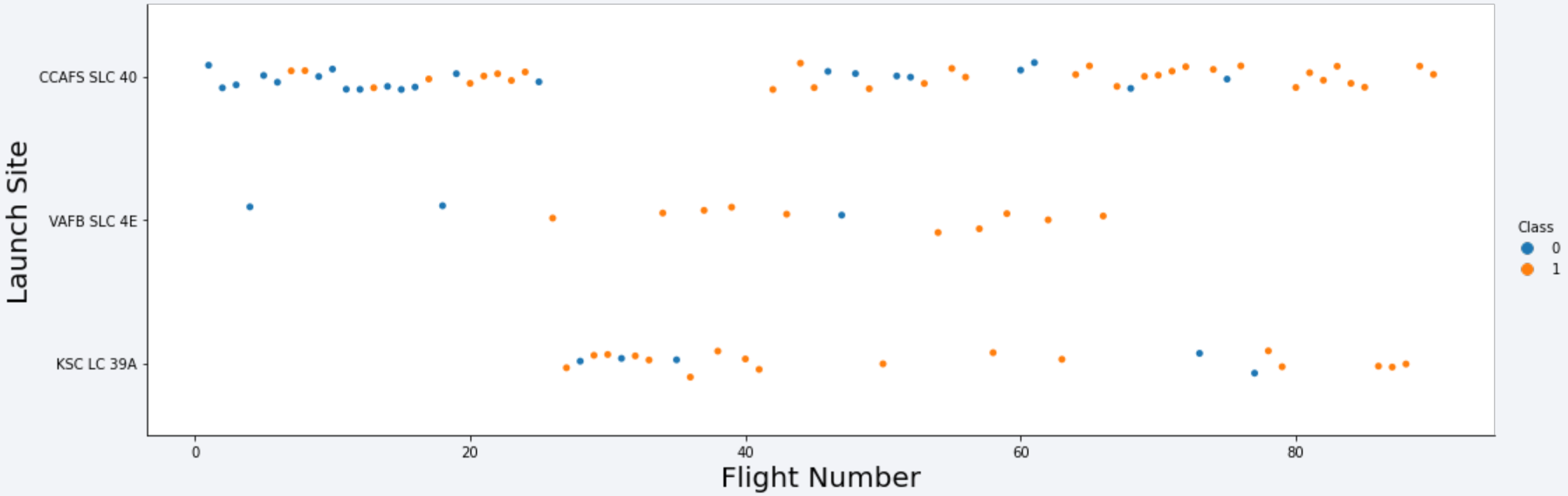
- EDA:
  - The higher the flight number, the success rate tend to be a little bit more successful
  - High payload mass tend to have higher success rate
  - ES-L1, GEO, HEO, SSO orbits have a perfect success rate
  - In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO, and ISS
  - The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.
- Predictive analysis
  - Logistic Regression, SVM, and KNN has similar accuracy
- Interactive analytics screenshot



Section 2

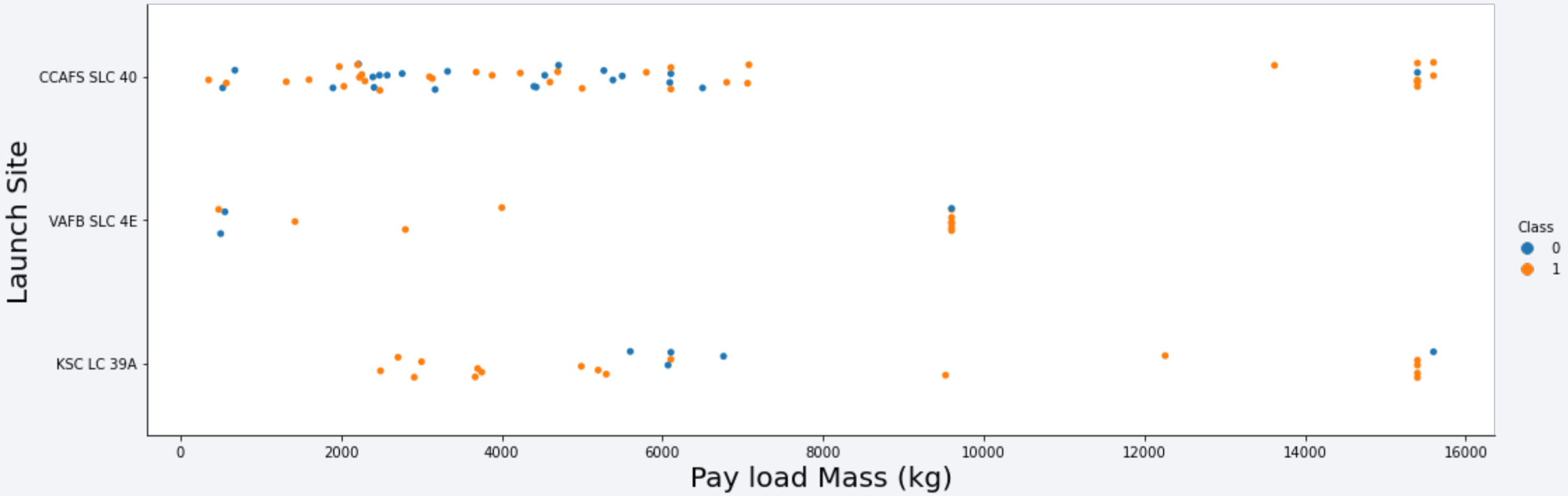
Insights drawn from EDA

Flight Number vs. Launch Site



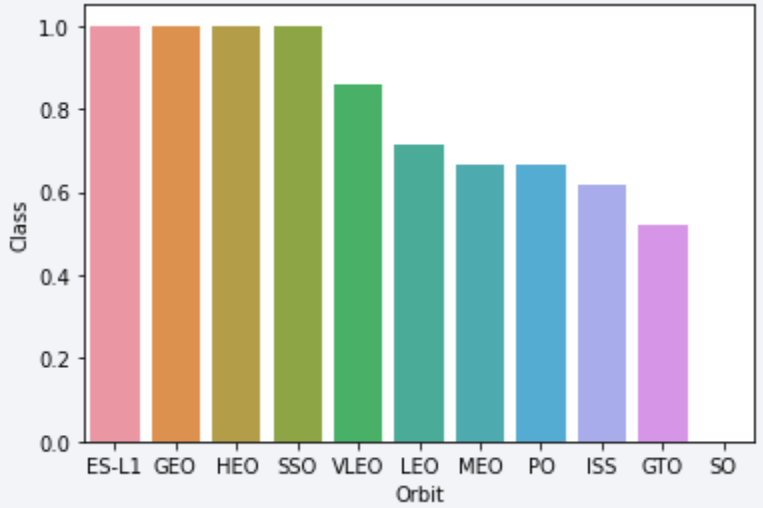
The higher the flight number, the success rate tend to be a little bit more successful

Payload vs. Launch Site



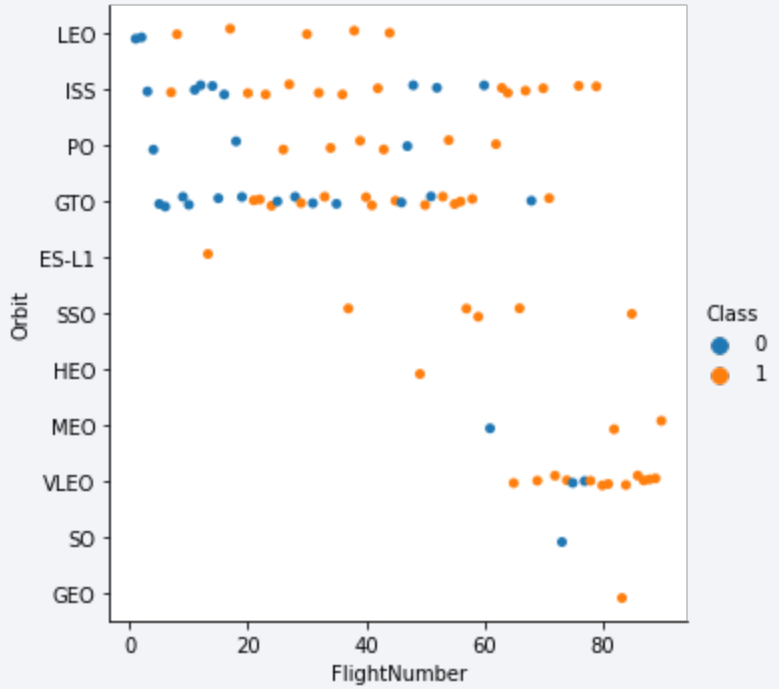
High payload mass tend to have higher success rate

Success Rate vs. Orbit Type



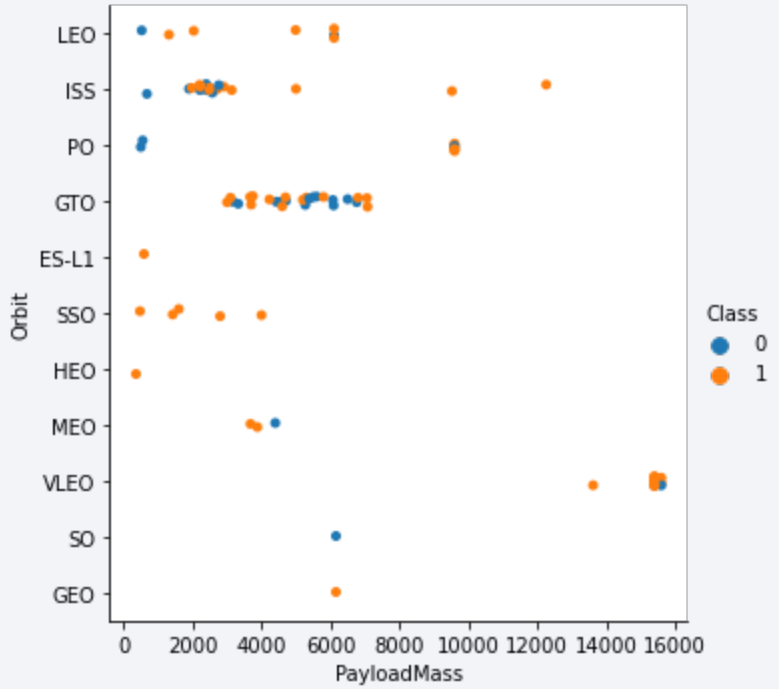
ES-L1, GEO, HEO, SSO orbits have a perfect success rate

Flight Number vs. Orbit Type



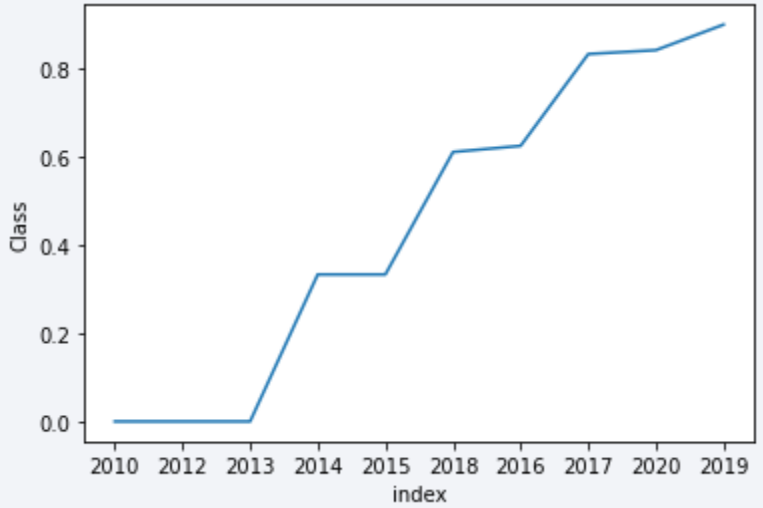
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

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) are both there here.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

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

There are 4 unique launch sites

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

These are 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

SUM(PAYLOAD_MASS_ _KG_)
45596

The total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

The average payload mass carried by booster version F9 v1.1 is ~2,434 kg

First Successful Ground Landing Date

The date when the first successful landing outcome in ground pad was achieved is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Above are the 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

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

The total number of successful and failure mission outcomes are as above

Boosters Carried Maximum Payload

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

On the side are names of the booster_versions which have carried the maximum payload mass.

2015 Launch Records

substr(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

Above are the records which display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

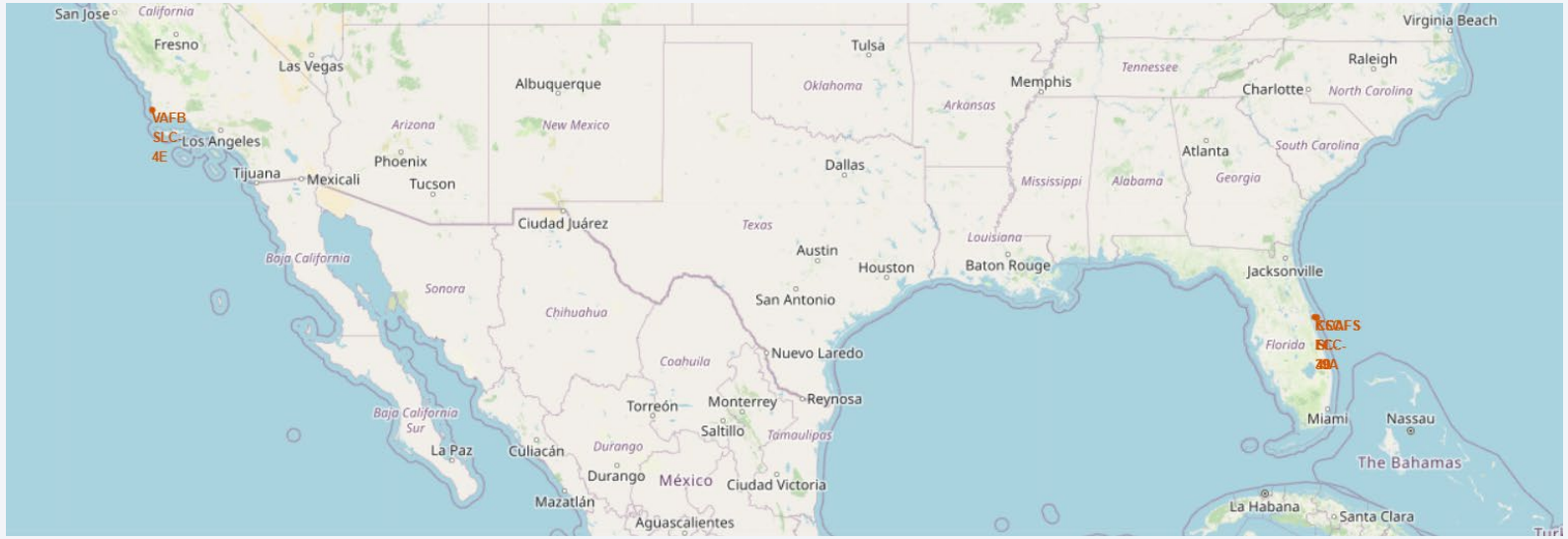
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

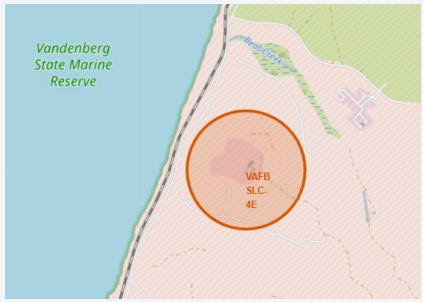
Landing_Outcome	COUNT(Landing_Outcome)
Failure (parachute)	31

Above is 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.


Section 3
Launch Sites
Proximities Analysis

Launch Sites Locations

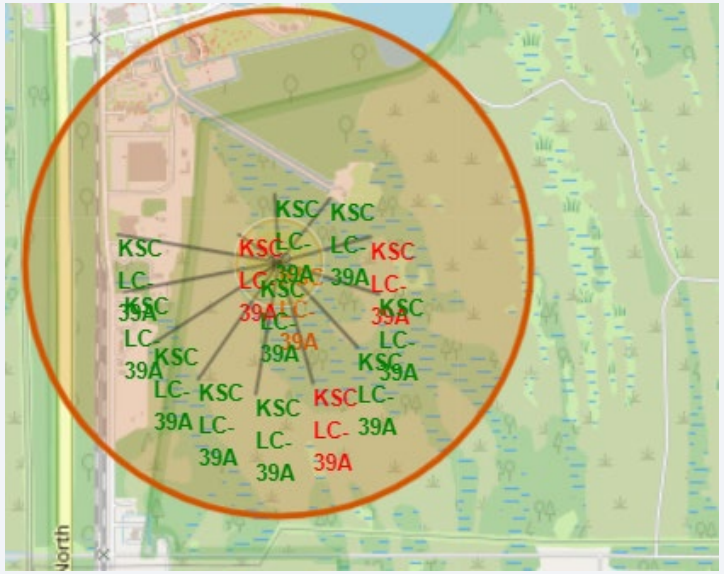


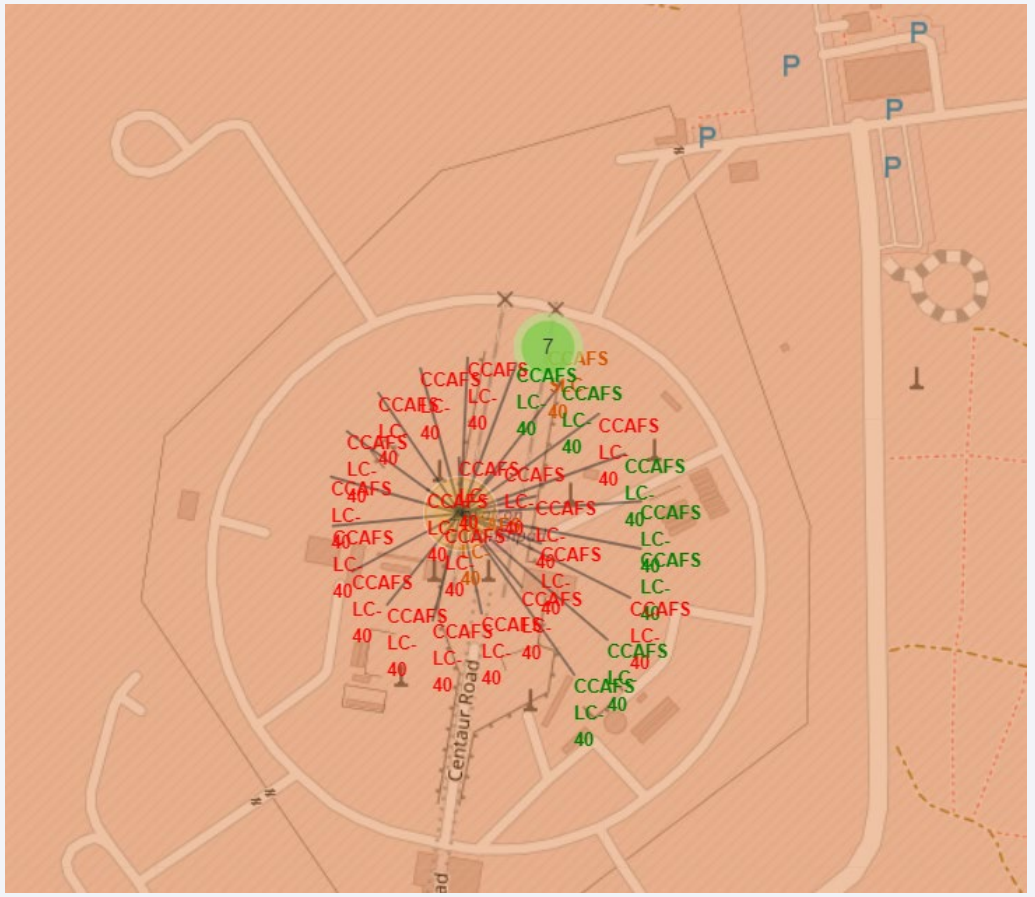


There are 4 launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E. Three of them are clustered nearby.

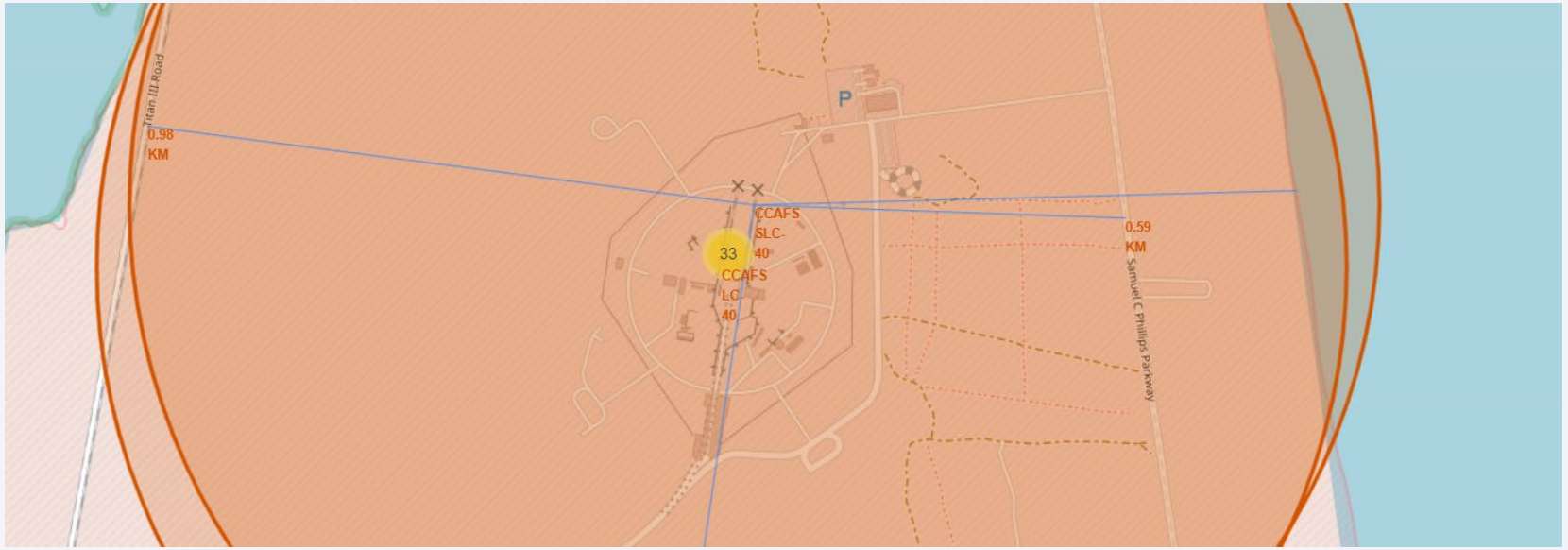
Launch Outcomes

Below are examples of the launch outcomes

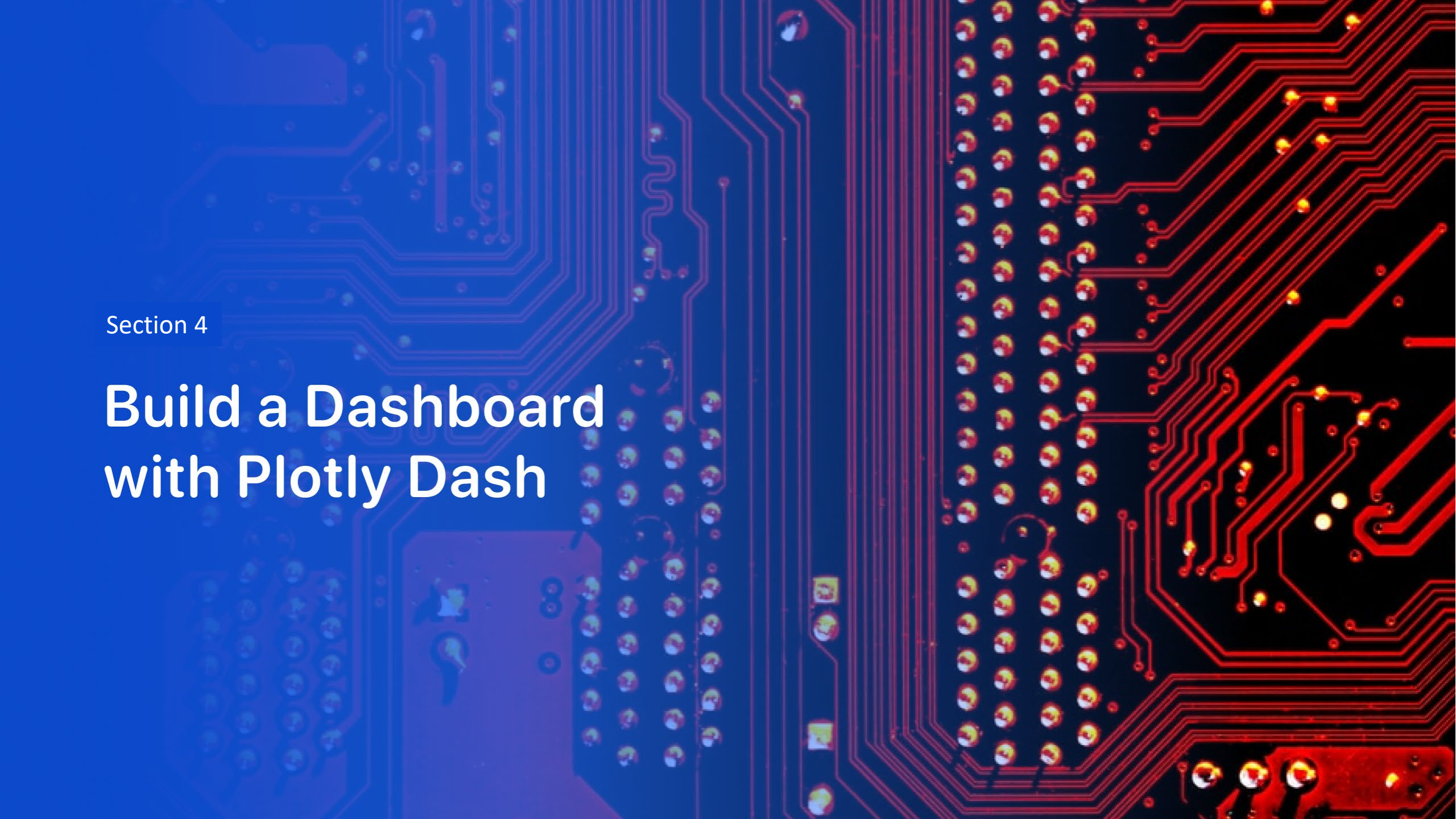




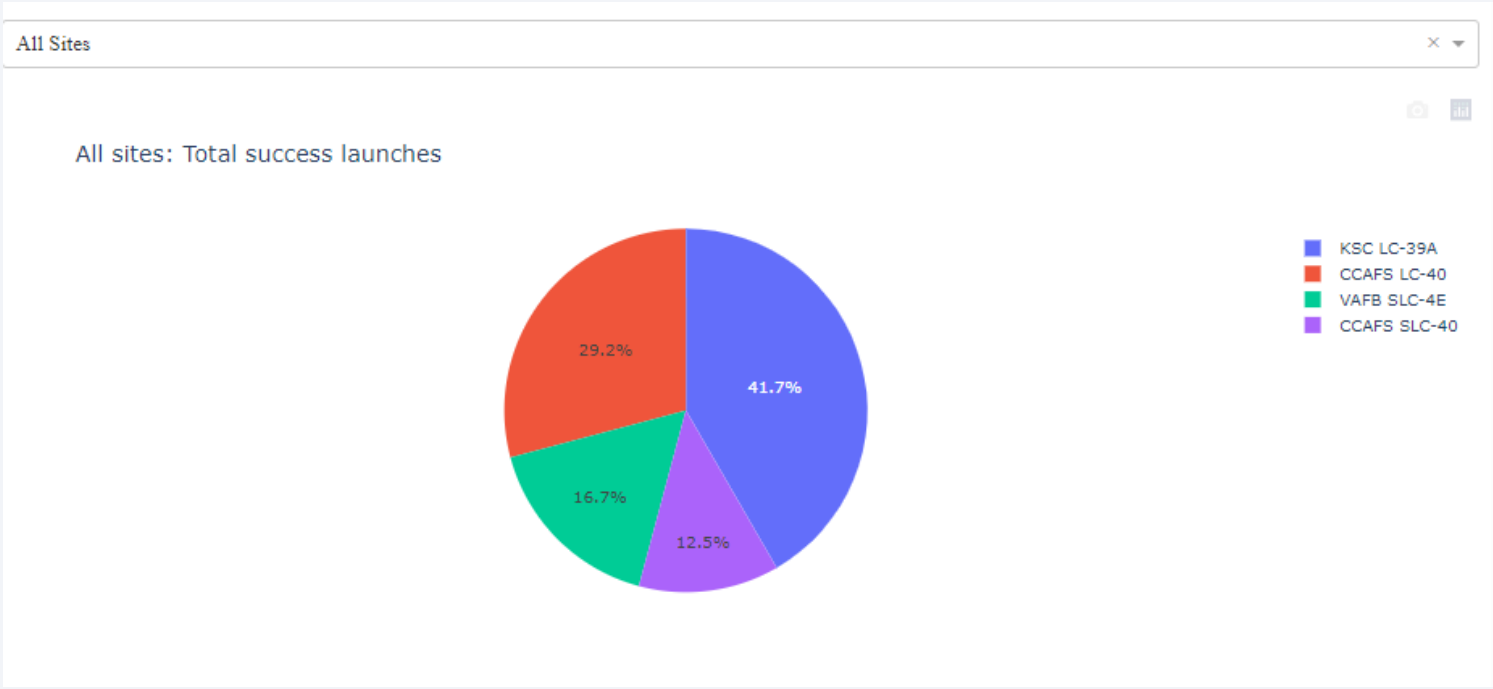
Launch Site Proximities



Above are the proximities from a launch site to the railway, highway, coastline

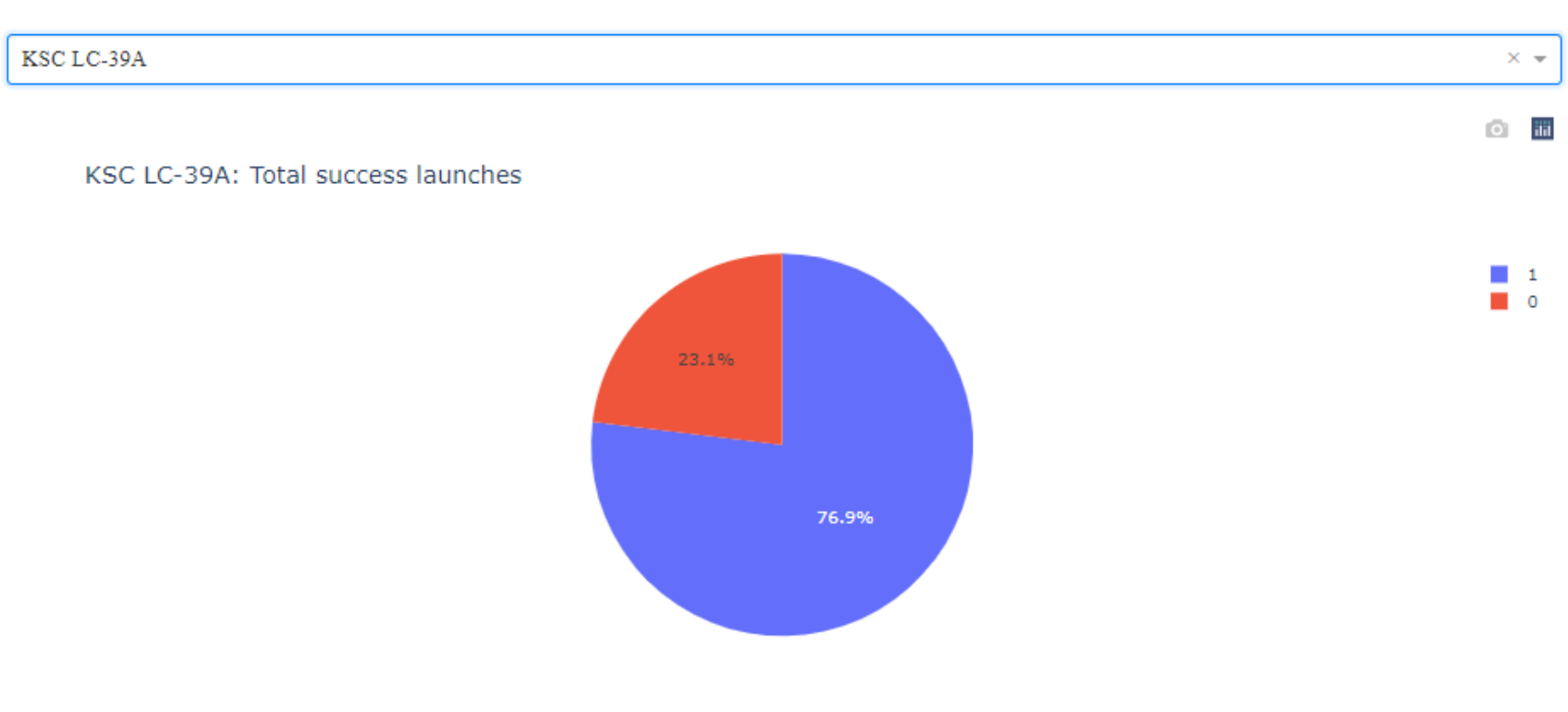

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



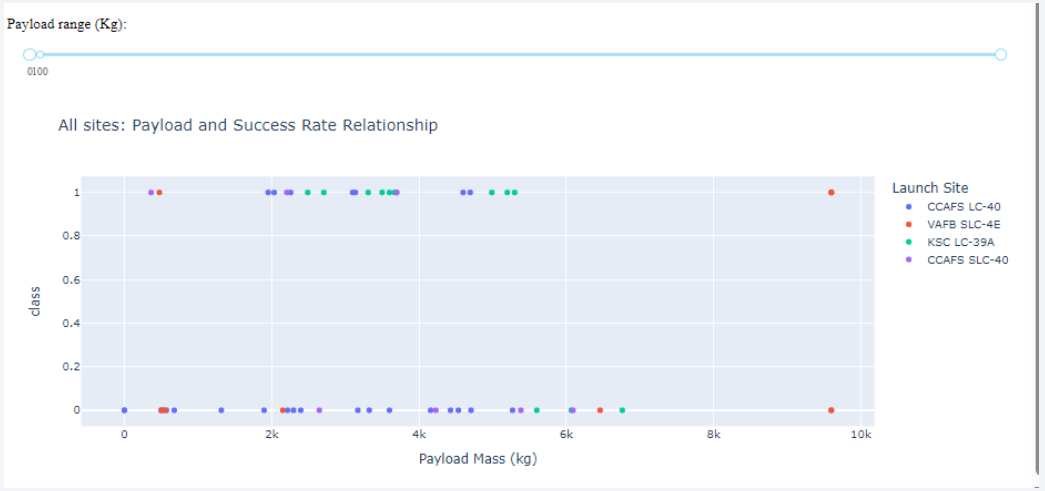
Above is a pie chart for the launch success count. The largest is KSC LC-39A and then CCAFS LC-40.

Launch Site with Highest Launch Success Ratio



KSC LC-39A has a 76.9% success rate

Payload VS Launch Outcome Scatter Plot



Different results from different payload

Section 5

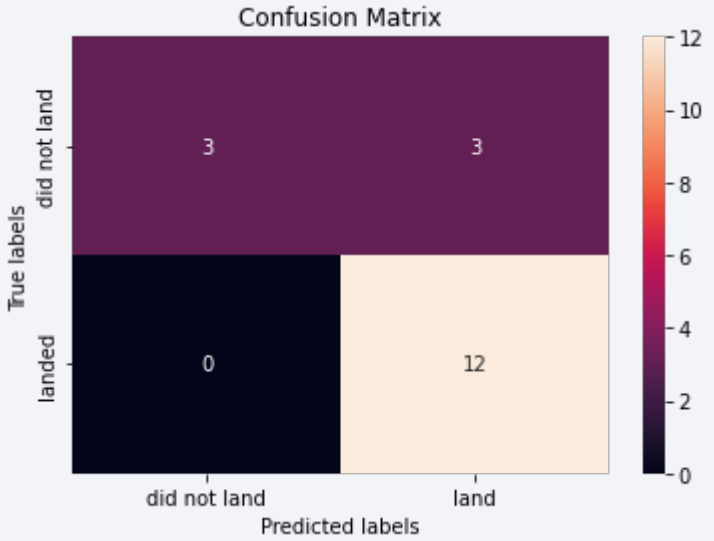
Predictive Analysis (Classification)

Classification Accuracy

- Logreg Score: 0.8333333333333334 Best Accuracy: 0.8464285714285713
- SVM Score: 0.8333333333333334 Best Accuracy: 0.8482142857142856
- Tree Score: 0.7777777777777778 Best Accuracy: 0.8892857142857145
- KNN Score: 0.8333333333333334 Best Accuracy: 0.8482142857142858

- The model with highest accuracy is Logistic Regression, Decision Tree, and KNN

Confusion Matrix



- Those three models have a quite good true positive rate. However there are 3 false positives.

Conclusions

- The most effective models are Logistic Regression, Decision Tree, and KNN
- Launch sites with higher flight amount tend to have better success rates.
- Orbits like ES-L1, GEO, HEO, SSO have the highest success rates.

Appendix

- N/A

Thank you!