IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carlos Eduardo
15/08/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - EDA Results
  - Interactive analytics
  - Predictive analysis

Introduction

Project Background and Context

We live in an era where space travel is becoming 'affordable' for everyone. Among all the space enterprises, perhaps the most successful is SpaceX with great achievements such as:

- Sending spacecraft to the International Space Station
- Starlink, a satellite internet constellation providing satellite Internet access
- Sending manned missions to Space

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Main Questions to be answered

For this project, we are trying to ask some questions like:

- Can we predict the success of the first stage landing based on the data available?
- Does the geographic location of a launch site affects the success of the first stage landing?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection Methodology
  - SpaceX Public Rest API
  - Web Scrapping mission data from Wikipedia
- Perform data wrangling
  - Analyze and select relevant features for the model
  - One Hot encoding data fields
  - Cleaning null values and irrelevant columns
- Perform exploratory Data Analysis (EDA) using visualization and SQL
- Perform interactive visualization analytics using folium and Plotly Dash
- Perform predictive analysis using classification models
  - Test and train classification models such as: Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Decision Trees, evaluating each one for the best classifier.

Data Collection

Rest API

The first method used to collect data, was the SpaceX Rest API. It gives us information about the past missions, rocket launches, among others.

The API is available at: https://api.spacexdata.com/v4/

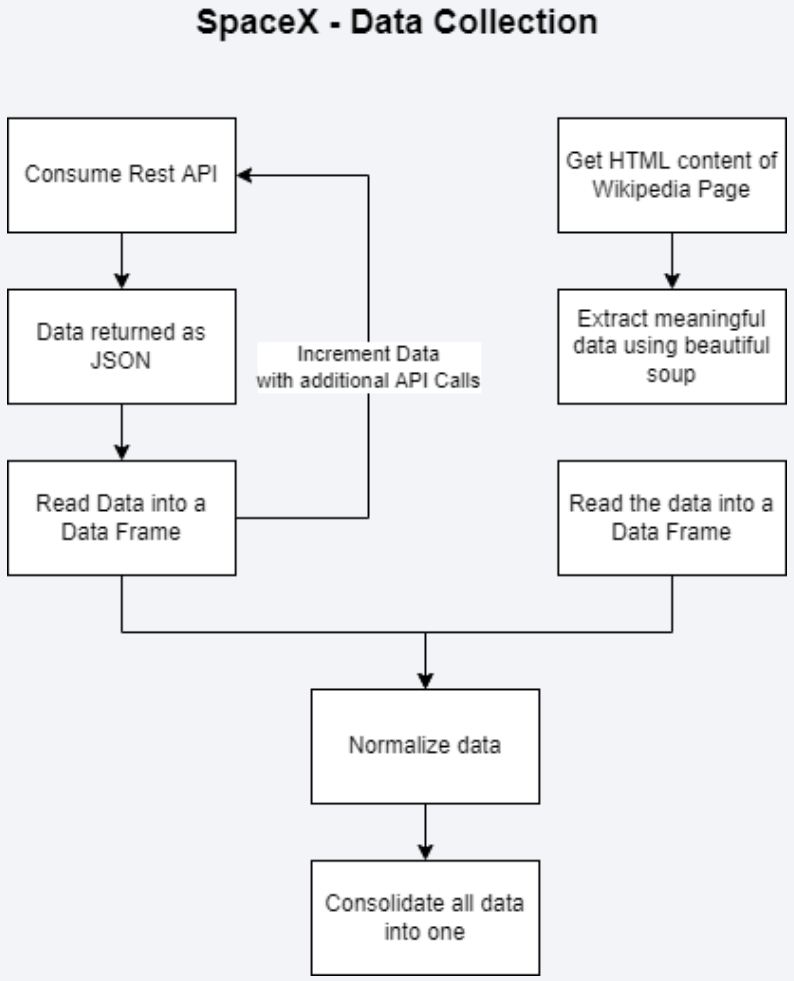
Web Scrapping

For complementing the API data, the Wikipedia Page was scrapped using Beautiful Soup. With that, we incremented our data with information on Falcon 9 and Falcon Heavy Launches records.

Data Collection

Flow Chart of Data Collection

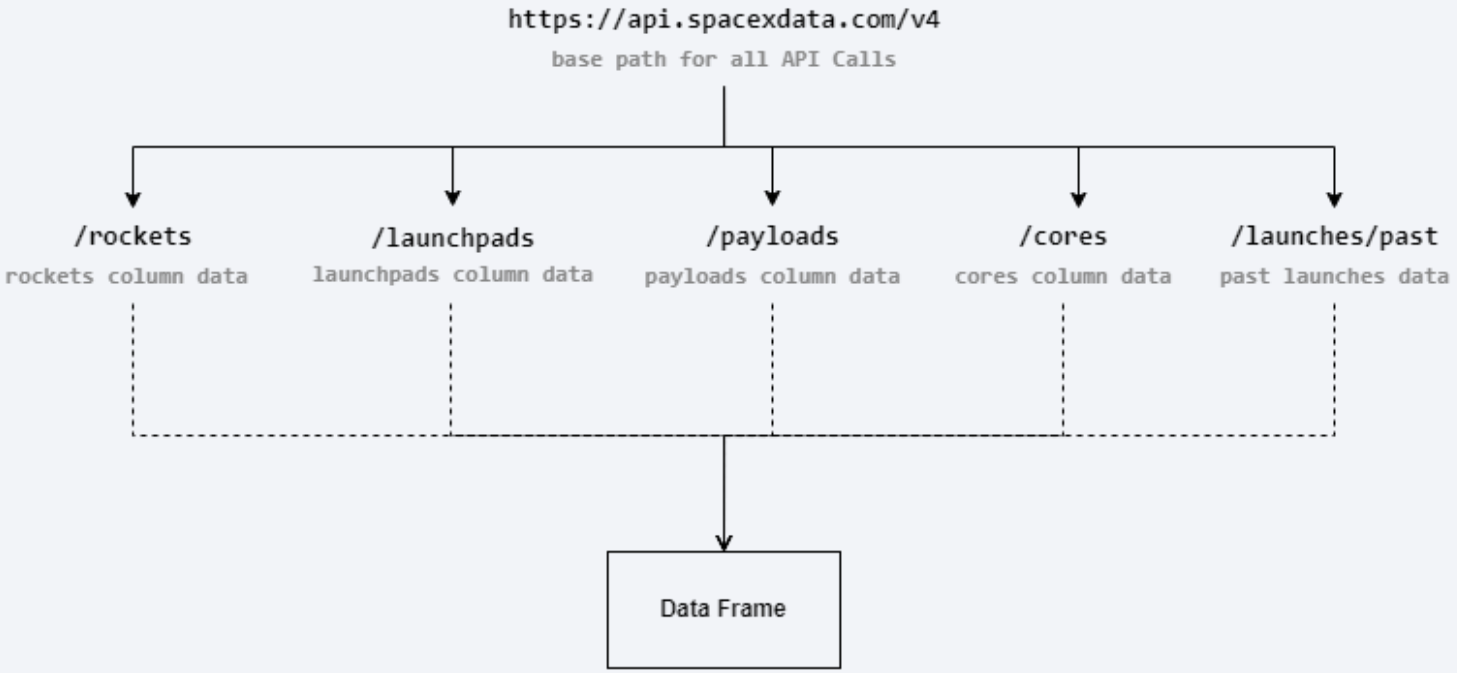
Code Reference

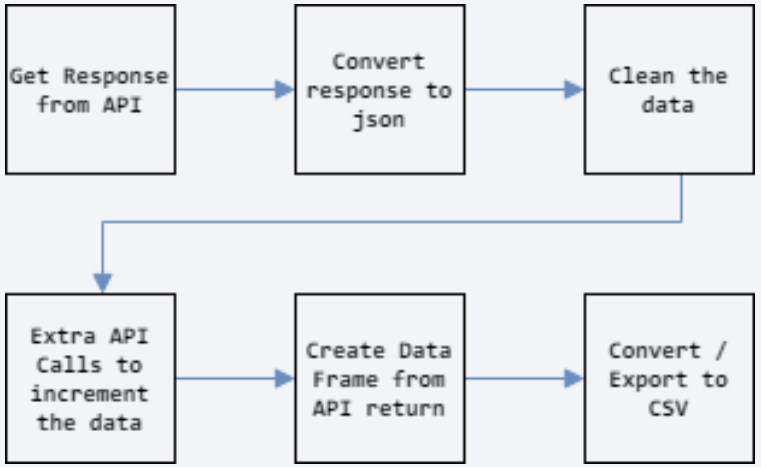


Data Collection – SpaceX API

API Calls made to the SpaceX Rest API

Code Reference

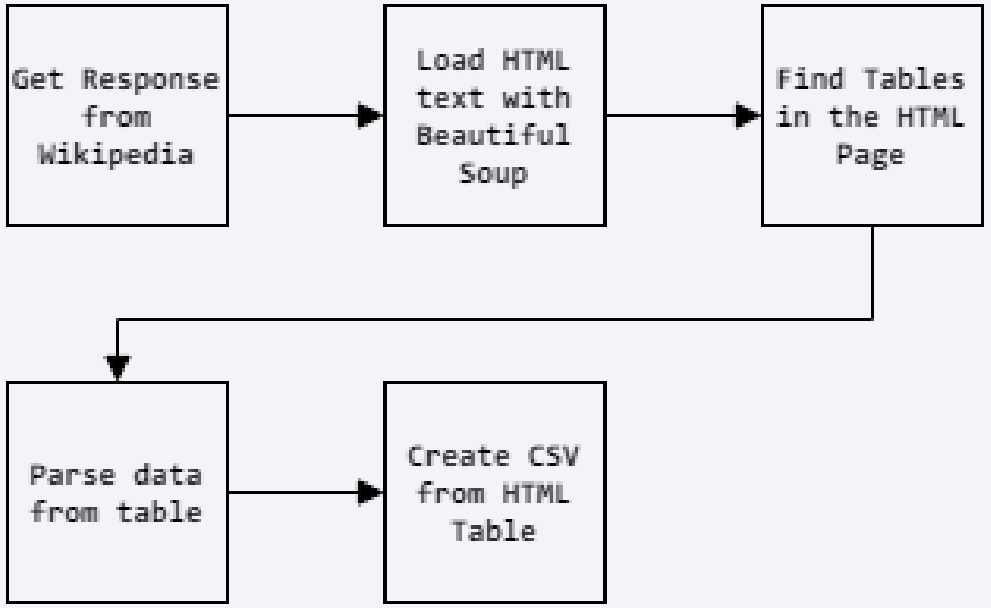




Data Collection - Scraping

- Web Scrapping from Wikipedia

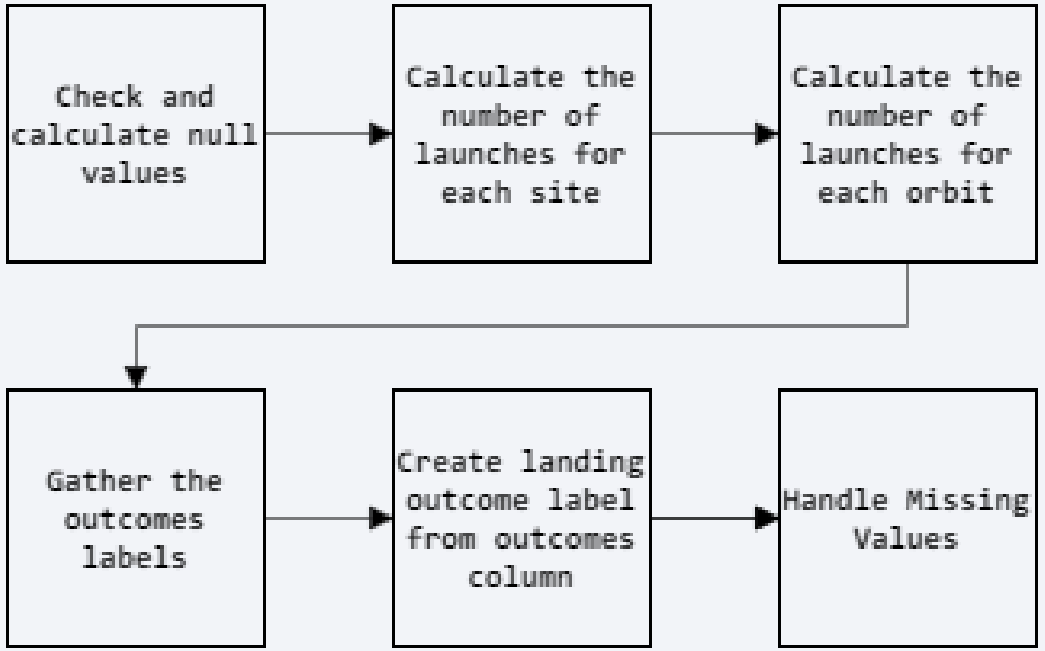
Code Reference



Data Wrangling

- Data Processing / Wrangling

Code Reference



EDA with Data Visualization

Categorical Graphs

- Flight Number x Payload Mass
- Flight Number x Launch Site

Bar Graphs

- Success Rate x Orbit

Scatter Graphs

- Flight Number x Payload Mass

Line Graphs

- Success Rate x Year

Code Reference

EDA with SQL

- SQL was used to extract come insights from our data:
  - Displaying the names of the unique lauch 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 which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Code Reference

Build an Interactive Map with Folium

- Folium objects added to our final map
  - Circle object added at NASA Johnson Space Center's coordinate
  - Circle object for each launch site
  - Marker Cluster used to group the success /failed markers in our map
  - Markers to show success and failed landing outcomes for each launch site
  - Markers and Lines to show the distance between launch sites and key locations like: closest railway, closest highway, closest city and coastlines.

Code Reference

Build a Dashboard with Plotly Dash

- Our dashboard is dynamic and can be used to analyze our data while changing variables. The main components added are:
  - **Charts**:
    - Pie Chart – Shows the total success and failures for one or all launch points
    - Scatter Chart – Shows the relationship between the Payload Mass Carried, and the success rate.
  - **Inputs:**
    - Dropdown – Used to allow the user to select a launch site or all launch sites to see the statistics
    - Ranger Slider – Used to select the range of the payload mass in a fixed range

Code Reference

Predictive Analysis (Classification)

- Data Preparation
  - Loading data
  - Normalize data
  - Split train / test data
- Model Preparation
  - Initialize Machine Learning Algorithms
  - Select parameters for each algorithm using GridSearchCV
  - Train the GridSearchCV model with the training data
  - Find the best parameters for each model
- Model Evaluation
  - Compute the accuracy of each algorithm with the best hyperparameters
  - Plot confusion matrix for evaluating results
- Model Comparison
  - Compare models based on the accuracy on a new test data
  - Examine the model with best accuracy

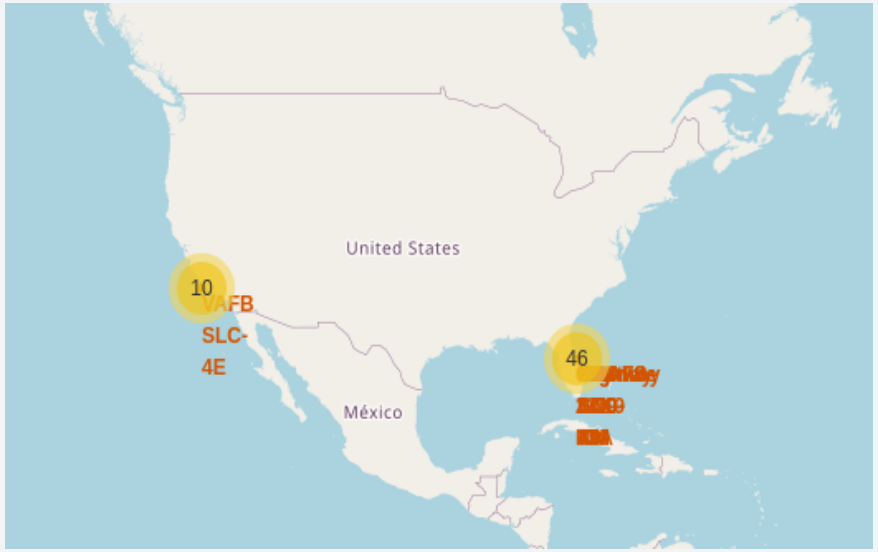
Code Reference

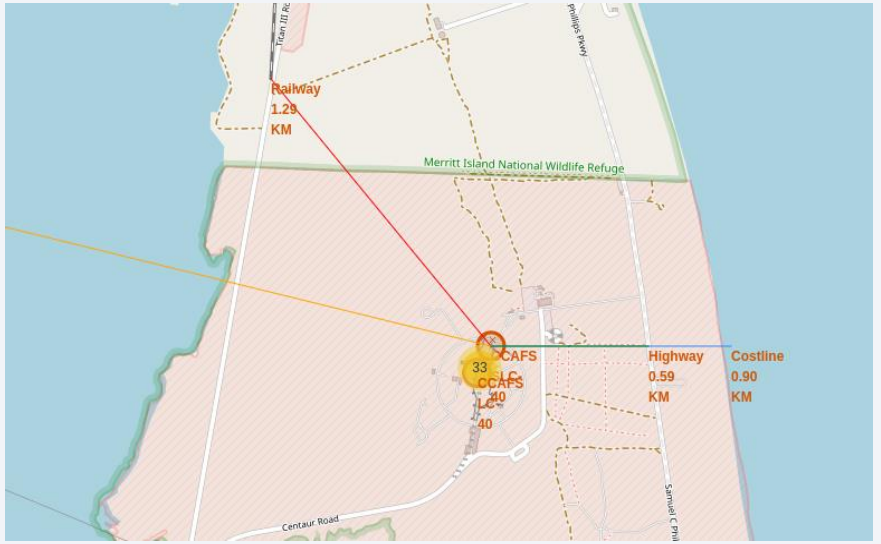
Results

- Exploratory data analysis results
  - Space X has 4 different launch site. Two of them are really close to each other.
  - The first launch recorded as successful was in 2015
  - The success rate for falcon 9 boosters is above the average.
  - Almost all of the missions were completed successfully
  - Booster with version beginning with F9 B5 have caried the most payload mass
  - As the years passes, the chance of a successful land outcome increases.

Results

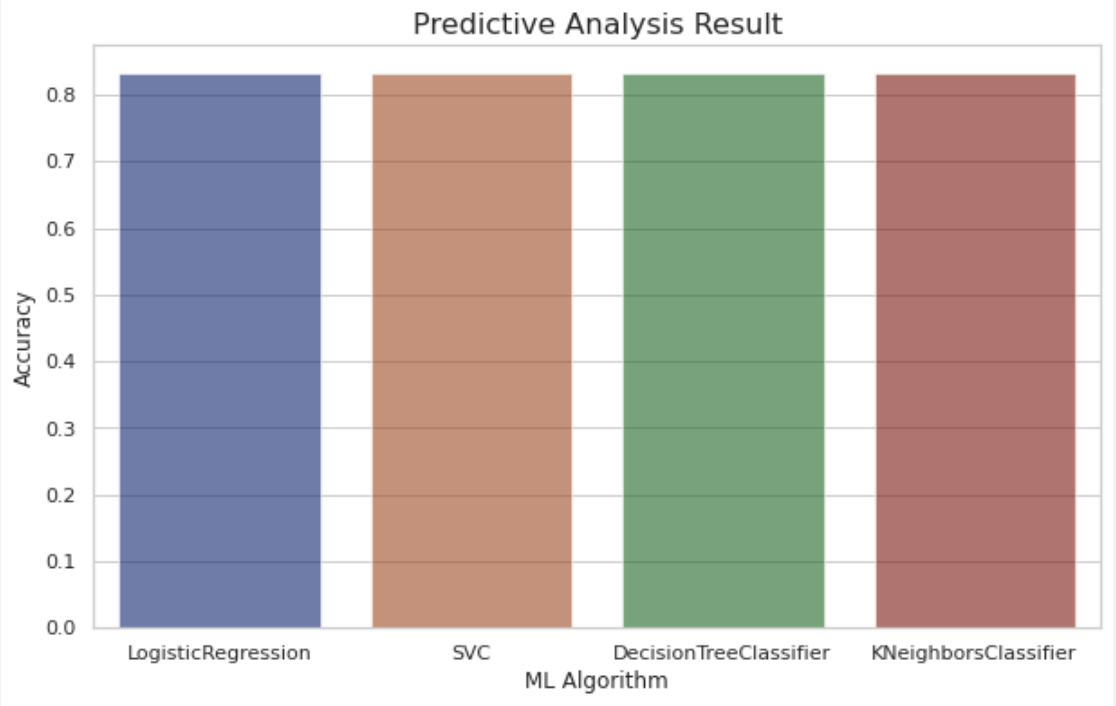
- With the use of interactive analytics on a map, we could find out that all launch sites are located in a strategical place, far from cities and highways, being close to coastlines, for safety.





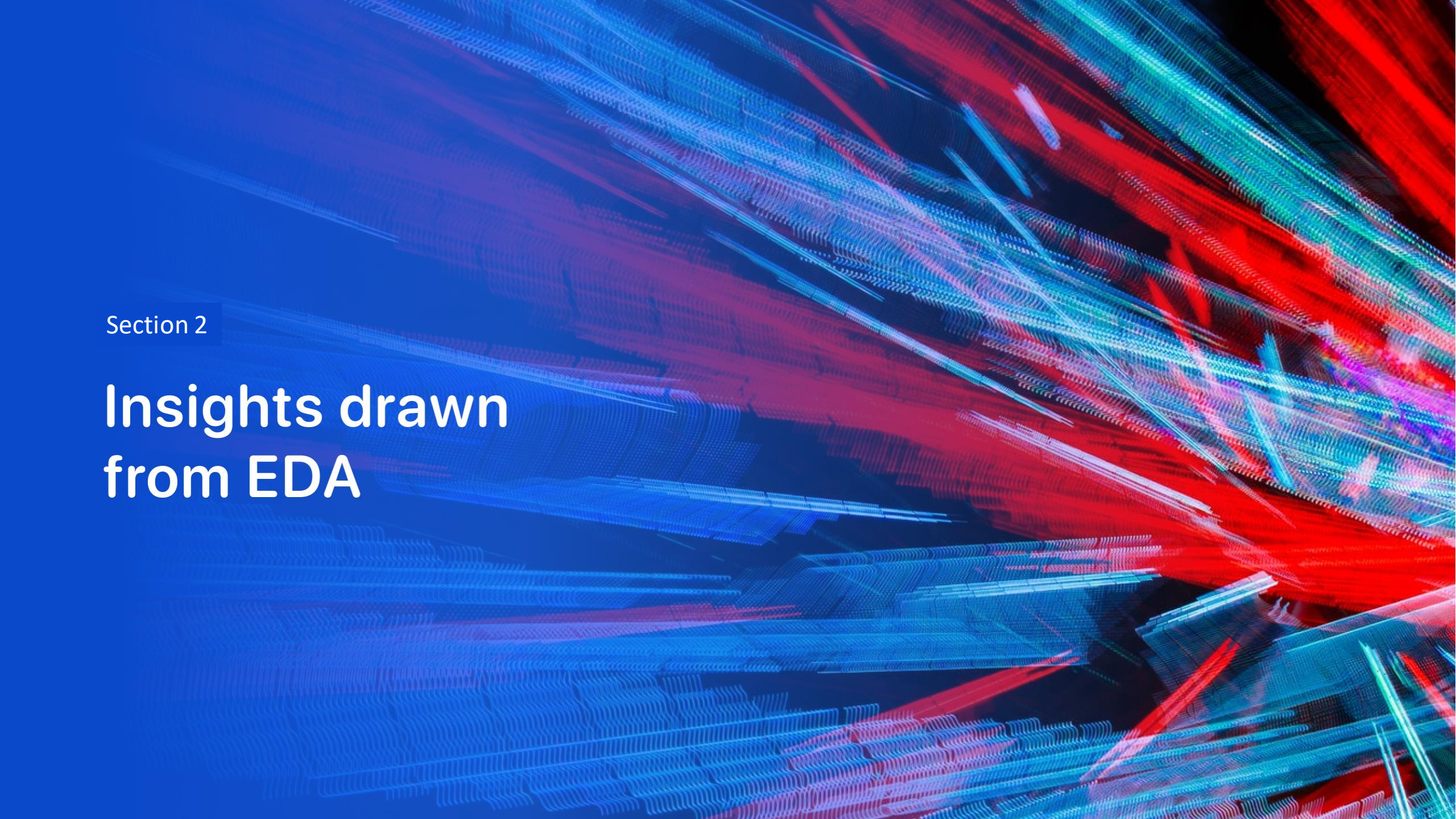
Results

- Predictive Analysis showed that all models had the same accuracy of 83.33%. That being any of the algorithms are well suited for our data.

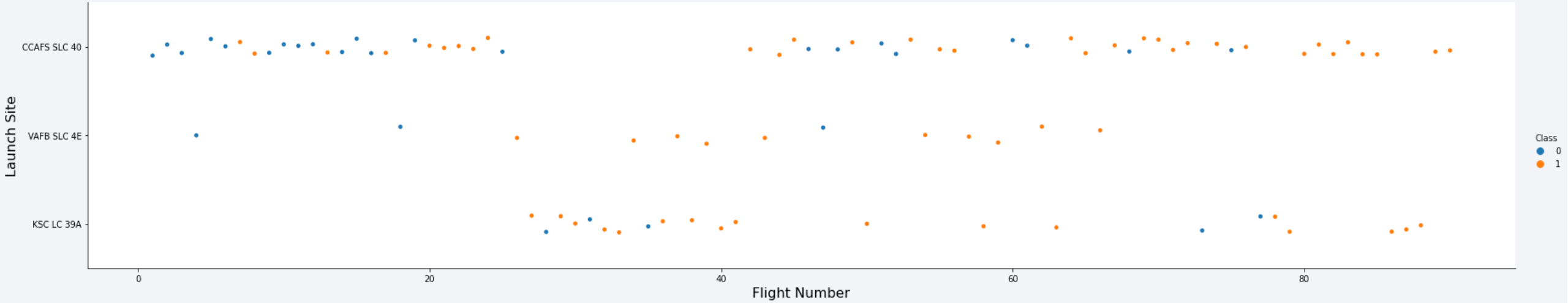


Section 2

Insights drawn from EDA

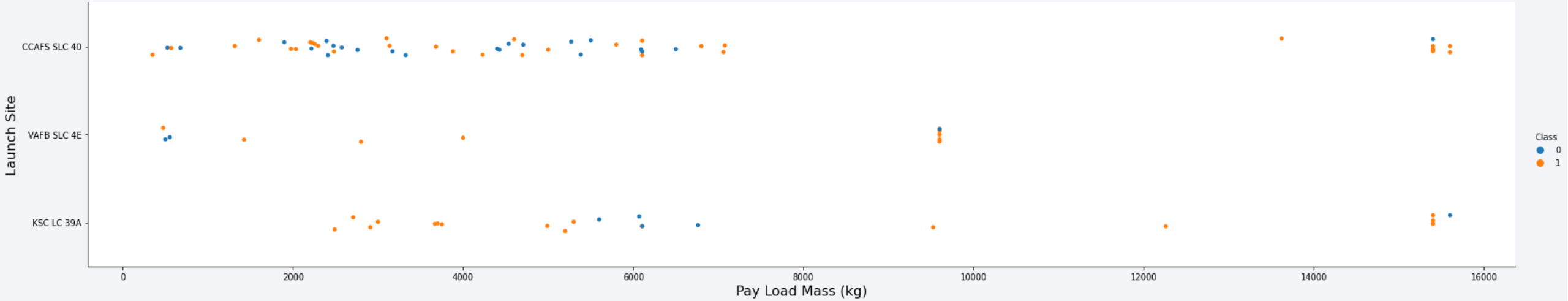


Flight Number vs. Launch Site



- We can see that, as the number of lights increases, the success rate there is a tendency that the success rate will increase as well.

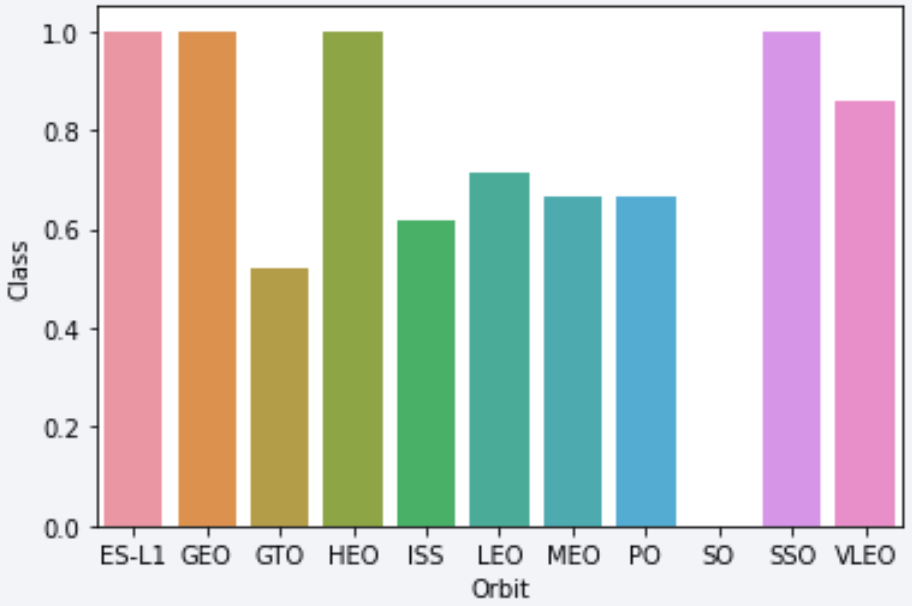
Payload vs. Launch Site



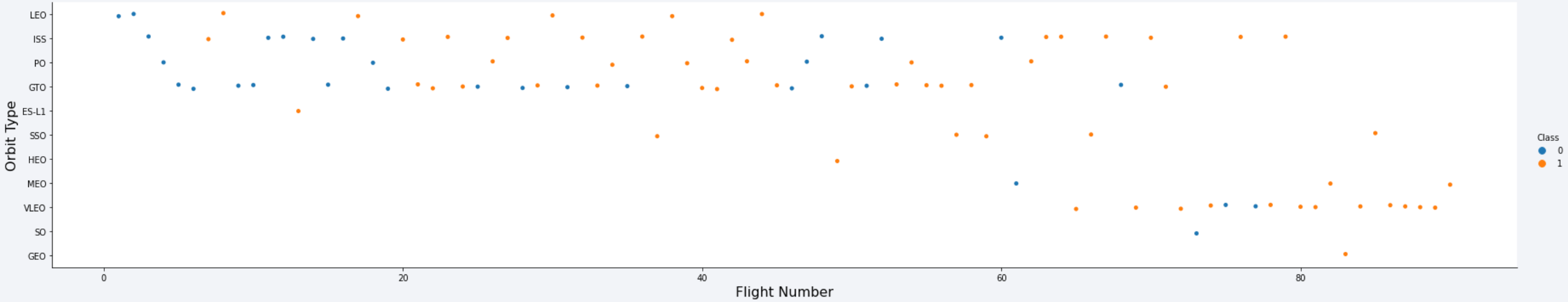
- CCAFS SLC 40 and KSC LC 39A are the launch sites that have caried the more heavier payloads, with a good success rate.
- VAFB SLC 4E did not carried any heavy payload mass(> 10000 KG), and is great for launches with payload mass up to 6000 KG.

Success Rate vs. Orbit Type

- Orbits with high success rate: (+80%)
  - ES-L1
  - GEO
  - HEO
  - SSO
  - VLEO

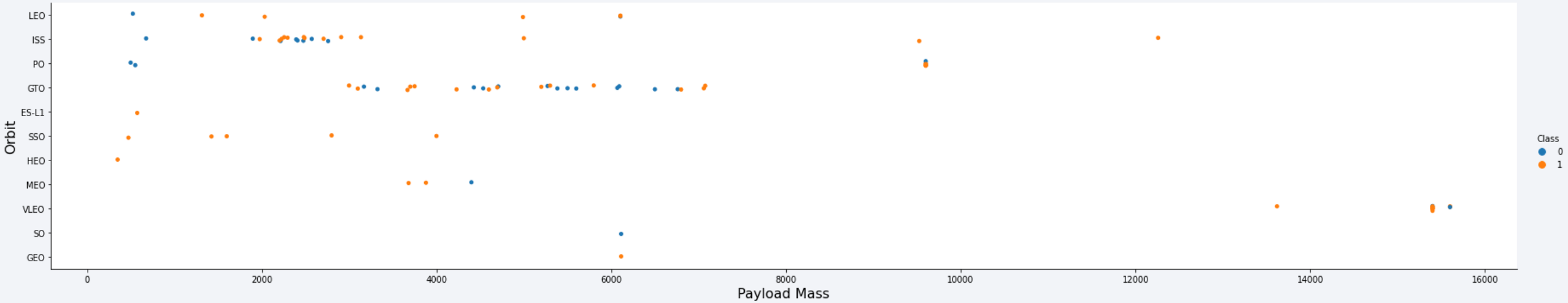


Flight Number vs. Orbit Type



- As the flight number increases, there is a tendency that the success rate for all orbits will increase as well.
- VLEO shows a great success rate, and can be a good opportunity and investment.

Payload vs. Orbit Type



- The graph shows us that Orbits like VLEO and ISS are the only ones that had a high payload mass.
- There is no apparent relationship between the success rate for the orbits based on the payload mass for all orbits.
- SO and GE have few launches.
- VLEO is the orbit with the heavier payloads.
- In general, there aren't many heavy payloads missions for all the orbits.

Launch Success Yearly Trend

- Since 2013, the success rate has drastically increased, and had its peak in 2019.
- There is a trend that the success rate will increase over the years

All Launch Site Names

- There are 4 launch sites in our data.

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

Launch Site Names Begin with 'CCA'

- The following are the first 5 launch sites with name beginning with 'CCA'

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

Total Payload Mass

- The total payload mass in KG carried by boosters launched by NASA:

total_kg
45596

Average Payload Mass by F9 v1.1

- The average payload mass in KG carried by booster F9 v1.1

total_kg
2928

First Successful Ground Landing Date

- Accordingly to our query result, the first successful landing outcome on a group pad was registered in December 22nd of 2015

date_
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- By querying our data for boosters versions with successful landings on drone ship where the payload mass was greater than 4000 KG, but less than 6000 KG, 4 booster versions were found:

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

Total Number of Successful and Failure Mission Outcomes

- With our data, we can see that almost all missions made, except for 1 have been successful, giving us a 99% mission success rate

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Boosters versions that carried the heavier payload seen in our data

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- In 2015, there were only 2 missions in which the landing of the first stage in drone ship have failed:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

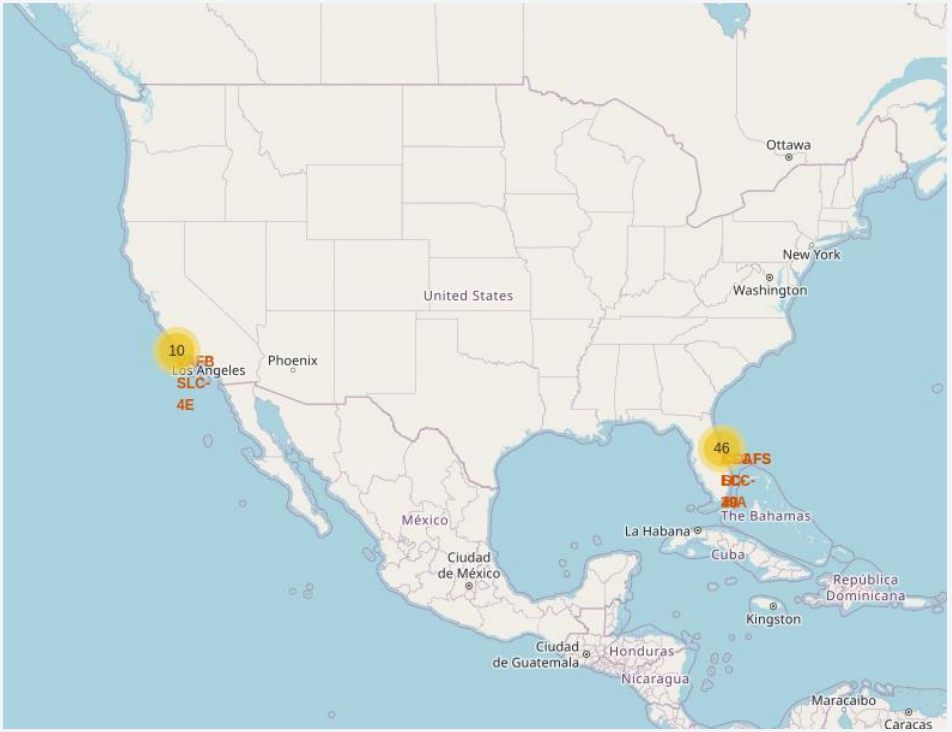
- Between the date 2010-06-04 and 2017-03-20, there were made 31 launches. More the 50% of them were either failed or not attempted (17)
- Most of the failures were attempted on drone ship

landing_outcome	total
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

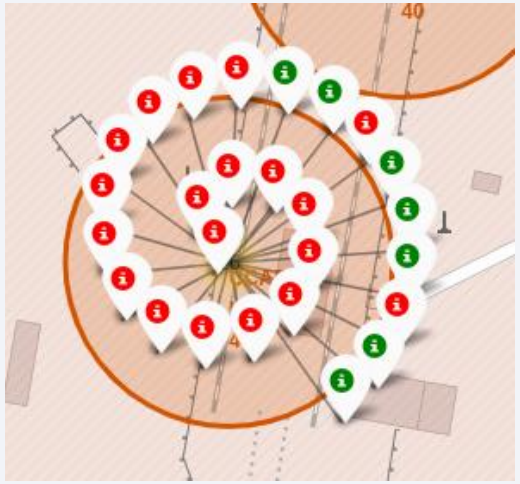
Launch Sites Location

- The following map shows us a global view of the location of all launch sites.
- We can see that they are all situated near south-east and south-west of the united states.
- All launch sites are near a coastline

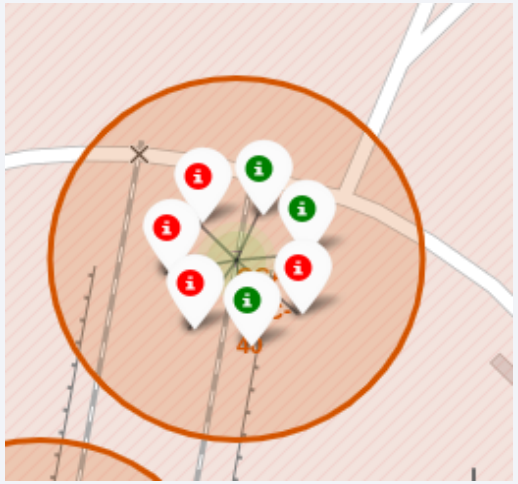


Success / Failure rate for Launch Sites

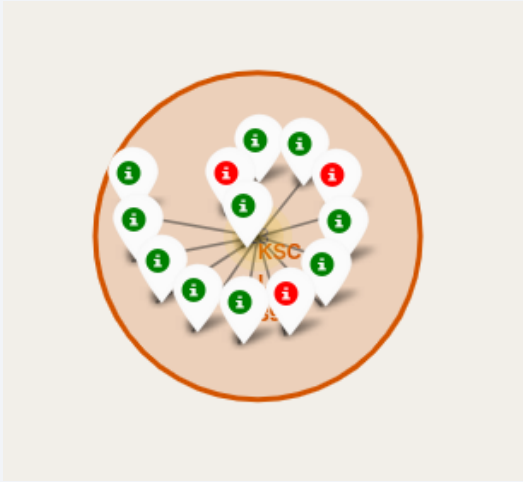
- The Green marker represents successful launches. Red marker represents unsuccessful launches.
- We note that KSC LC-39A has a higher launch success rate and CCAFS LC-40 has a lower success rate.



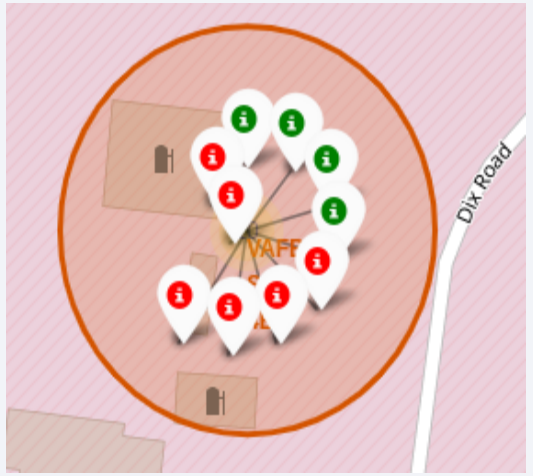
CCAFS LC-40



CCAFS SLC-40

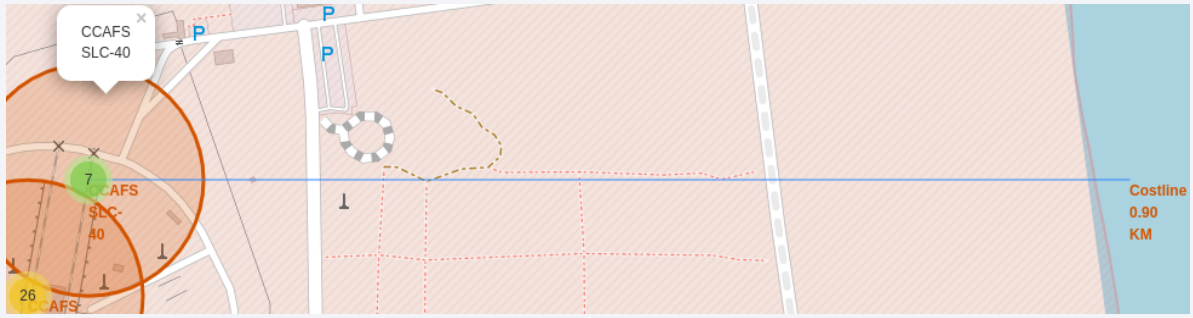


KSC LC-39A



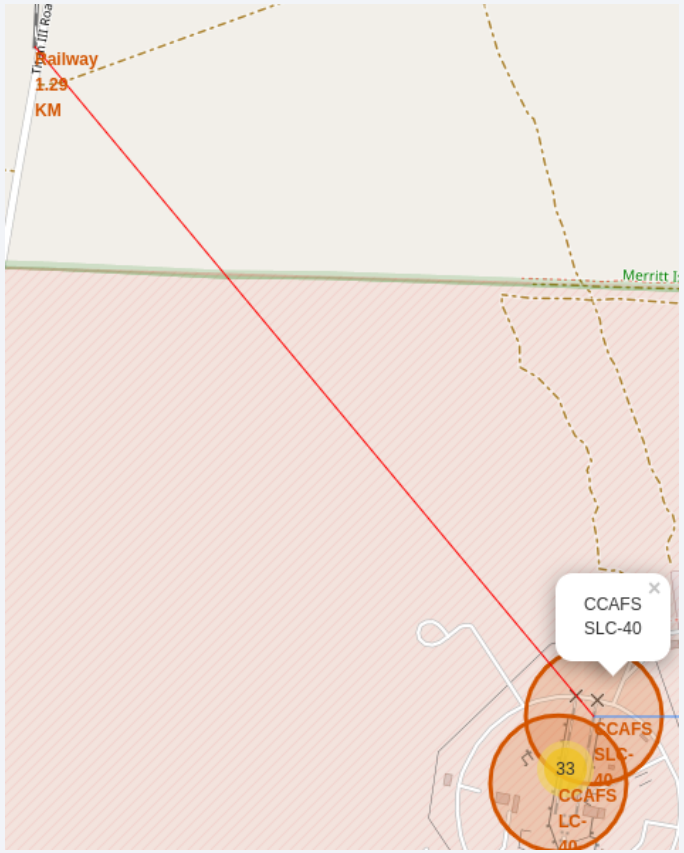
VAFB SLC-4E

Launch Sites Logistic Location



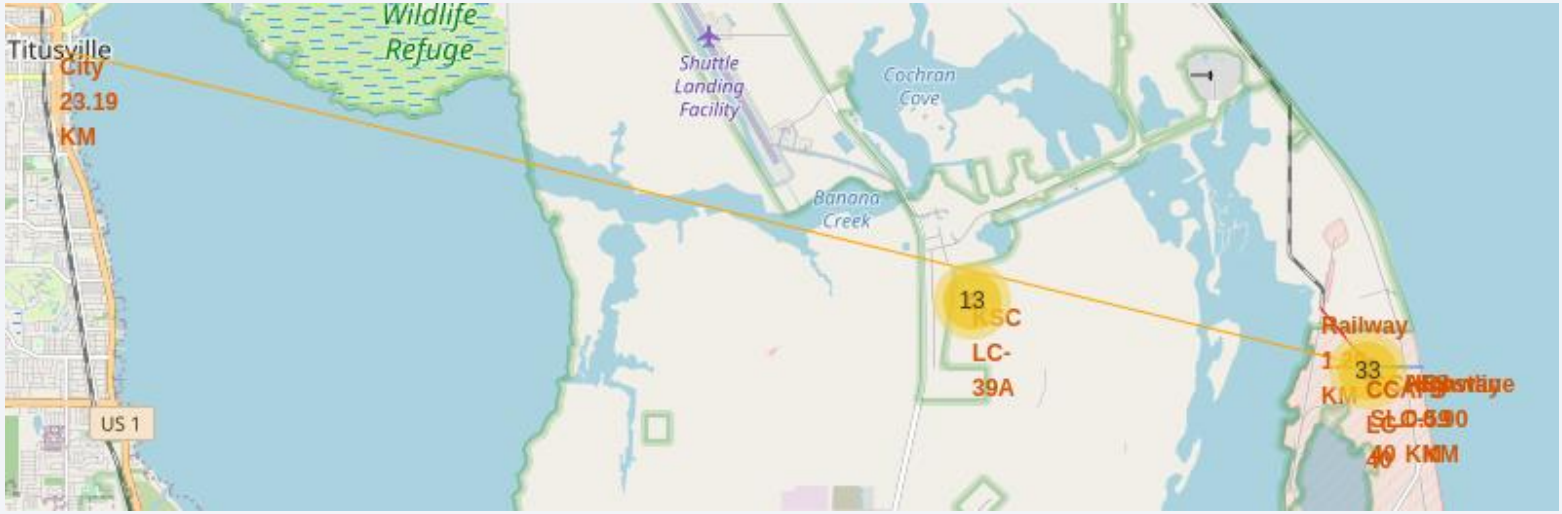
Launch sites tend to be close to coastlines

Launch sites tend to be close to highways

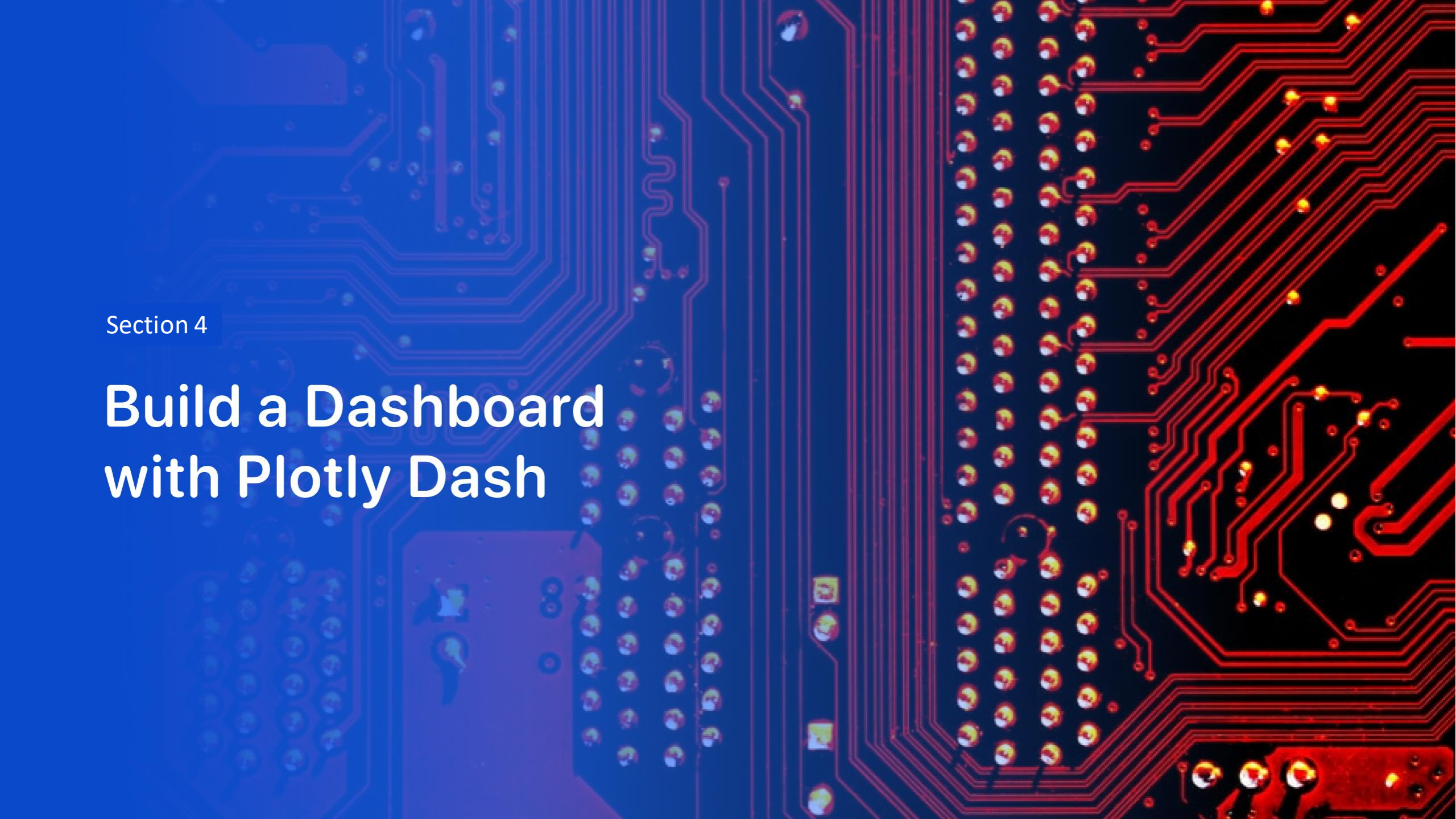


Launch sites seems to keep a certain distance of railways, but not much

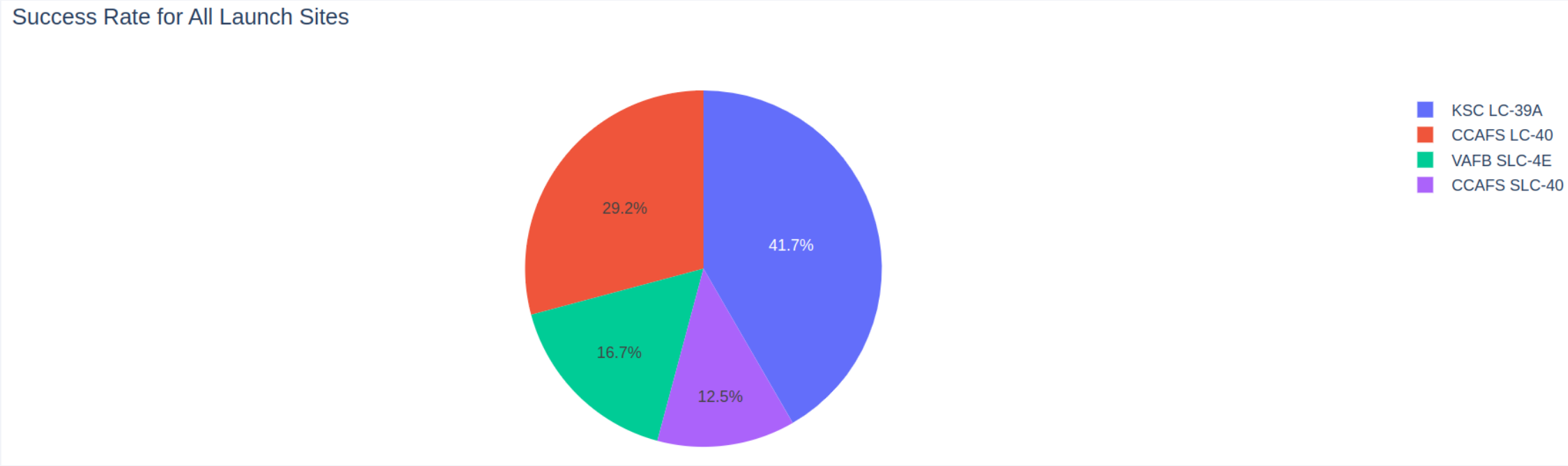
Launch Sites Logistic Location



Launch sites are far away from cities and populated areas

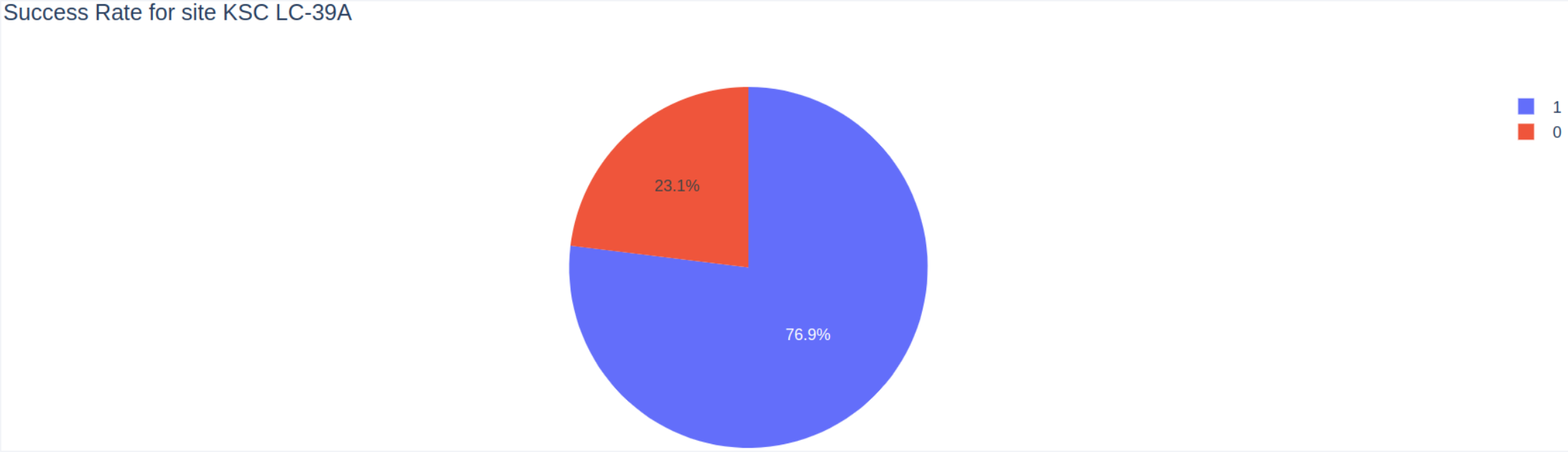

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Success Rate by Launch Site



- Launch Site KSC LC-39A shows to be the best Launch Site with 41.7% of success rate
- CCAFS SLC-40 appears to be the launch site with the least successful rate, with 12.5%

Dashboard - Launch Site with Highest Success



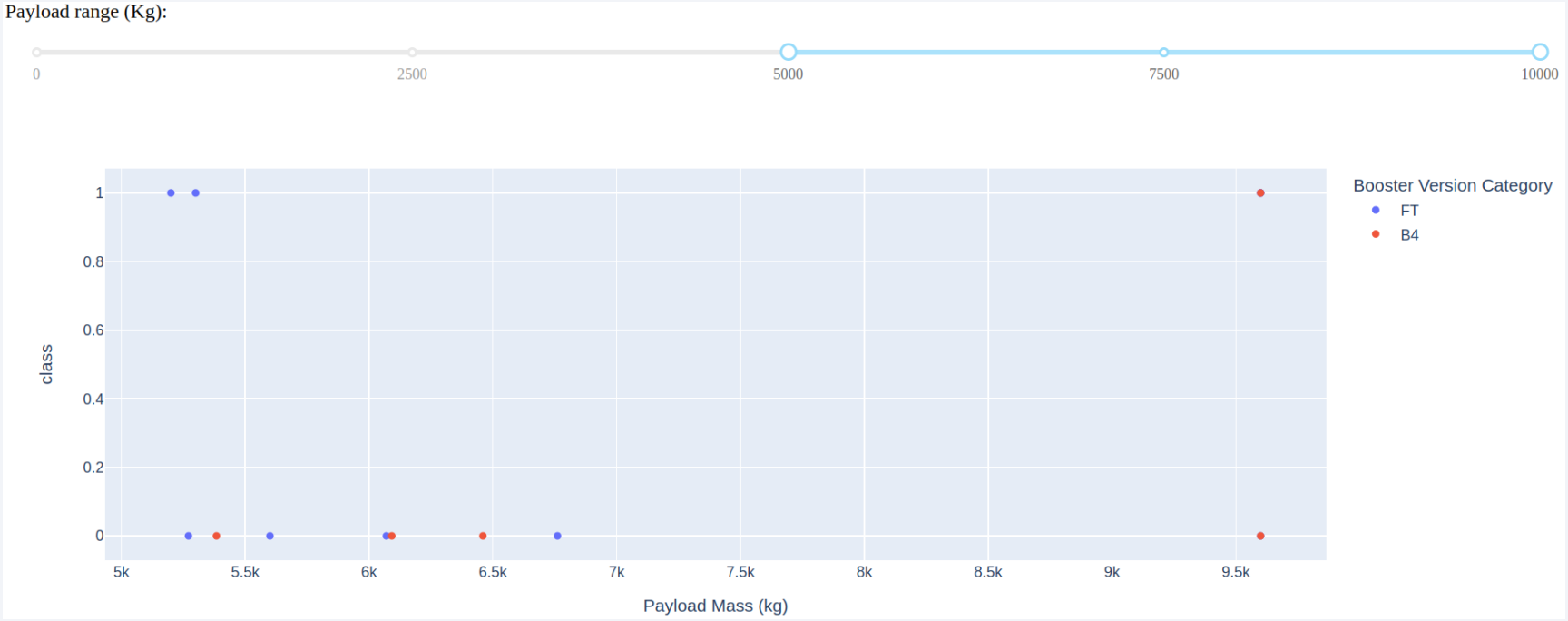
KSC LC-39A is the Launch Site with the higher success rate, with 76.9% success, and 23.1% failure rate.

Dashboard - Payload Mass vs Landing Outcome

Low Weighted Payload 0 - 5000

Dashboard - Payload Mass vs Landing Outcome

Heavy Weighted Payload 5000 - 10000



Dashboard - Payload Mass vs Landing Outcome

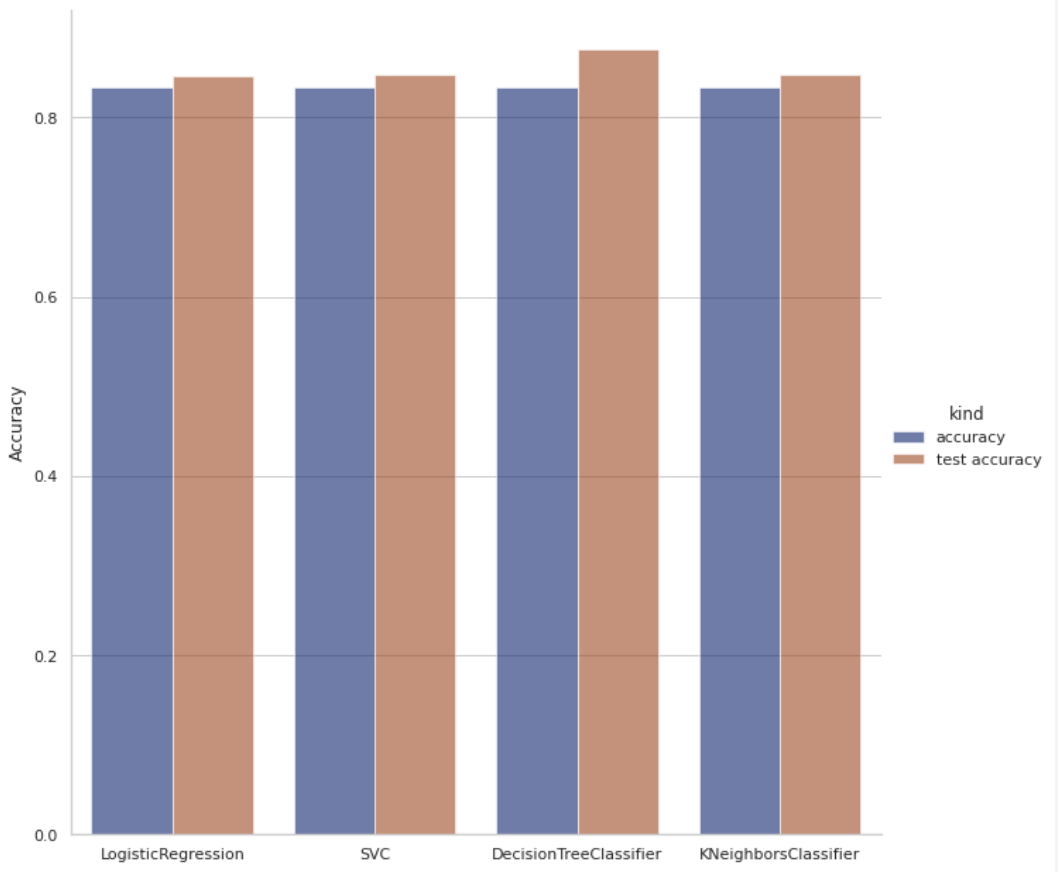
- The previous scatter plots shows us that there are fewer launches made with heavier payloads and, when they are made, the first stage of the launch usually fails, with a high failure rate
- For low weighted payload mass, we have much more launches, and a considerably higher success rate than the heavier payloads.
- Only 2 booster versions have caried a high payload mass: **FT** and **B4**

Section 5

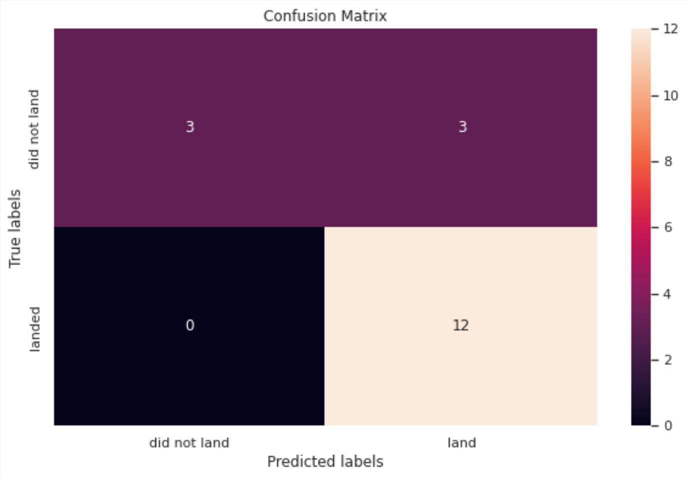
Predictive Analysis (Classification)

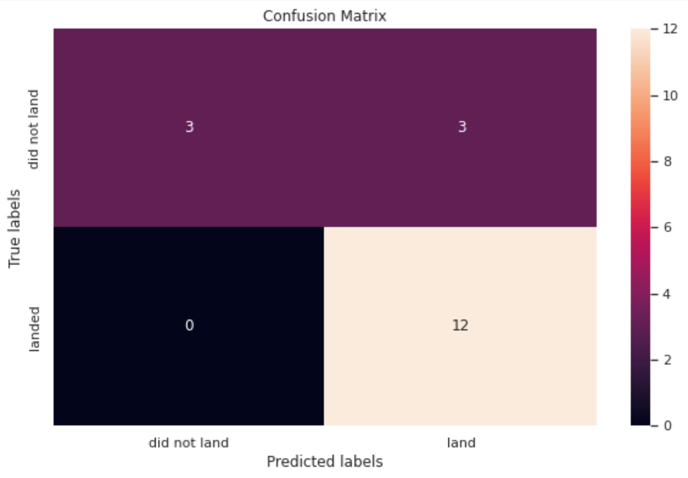
Classification Accuracy

- The accuracy for all the models were the same: 83.33%.
- Decision Trees has a higher accuracy on test data, making it the best suited model.

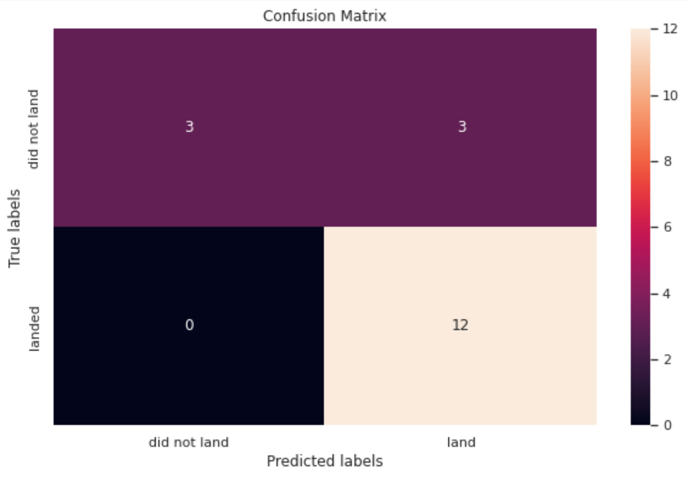


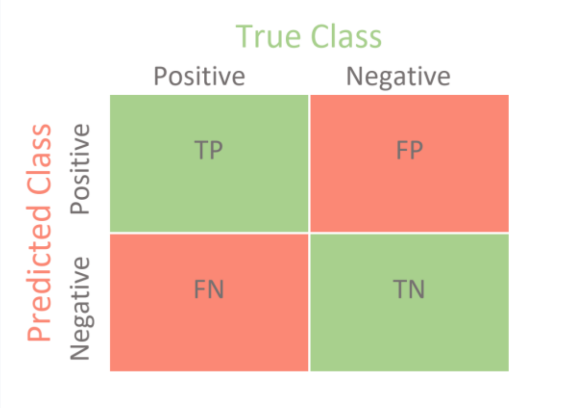
Confusion Matrix





- Since the accuracy of the models are equal, the confusion matrix for each one is identical.
- We can see the false positives are the main problem with our models.





Conclusions

- Different data sources were consumed in order to have a richer dataset.
- The success of a mission can be explained by several factors. That includes the launch site, payload mass, and number of previous launchers.
- Depending on the orbit the mission aiming, and the payload mass, since some orbits seems to only work with certain payload masses.
- All the tested algorithms shows an identical accuracy, showing that all of them are well suited for our problem.
- All the launch sites are strategically located, near coastlines and far away from cities and populated areas.
- There are not many launches made with really heavy payload mass.
- 99% of the missions were successful, and the success of the first stage landing is increasing over the years.

Appendix

- In order to maintain the accuracy of the models in a predictable state, I set `np.random.seed` to 0 in the beginning of the notebook.
- The full project is available on Github.

Thank you!