IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Faheem Ayub
Nov, 04th 2025

Outline

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

Executive Summary

- We will predict the successful landing of the SpaceX Falcon 9 first stage or otherwise, using machine learning classification algorithms
- Main steps of the project:
  - Data collection using SpaceX API and web scraping
  - Exploratory data analysis (EDA)), including data wrangling, data visualization and interactive visual analytics
  - Machine learning prediction
- Result Summary
  - It was possible to collect valuable data from public sources
  - EDA allowed to identify which features are the best to predict success of launchings
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity in the best way, using all collected data

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- We will predict 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
- 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 by Space Y for a rocket launch
- The main question: for a given set of features about a Falcon 9 rocket launch which include its **payload mass**, **orbit type**, **launch site**, and so on, will the first stage of the rocket land successfully
- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets
  - Where is the best place to make launches

Section 1

Methodology

Methodology

Our analysis comprised the following four stages:-

1. Data collection, wrangling, and formatting, using:
   - SpaceX API (https://api.spacexdata.com/v4/rockets/)
   - Web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
2. Perform data wrangling
   - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
3. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
4. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
5. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

Data Collection

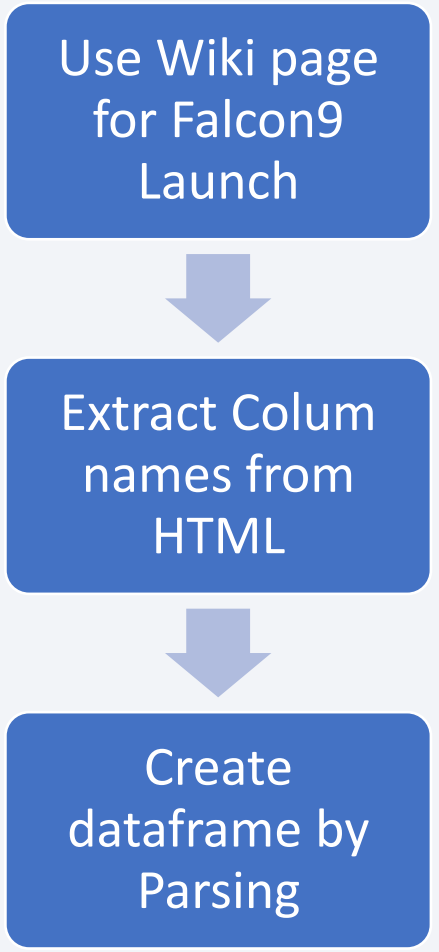
- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/ and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping techniques.

Data Collection – SpaceX API

- SpaceX API
  https://api.spacexdata.com/v4/rockets/
  - The API provides data about many types of rocket launches done by SpaceX; the data is therefore filtered to include only Falcon 9 launches
  - Every missing value in the data is replaced with the mean of the column that the missing value belongs to
  - We end up with 90 rows or instances and 17 columns or features. The picture shows first few rows of the data
- Source Code: https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

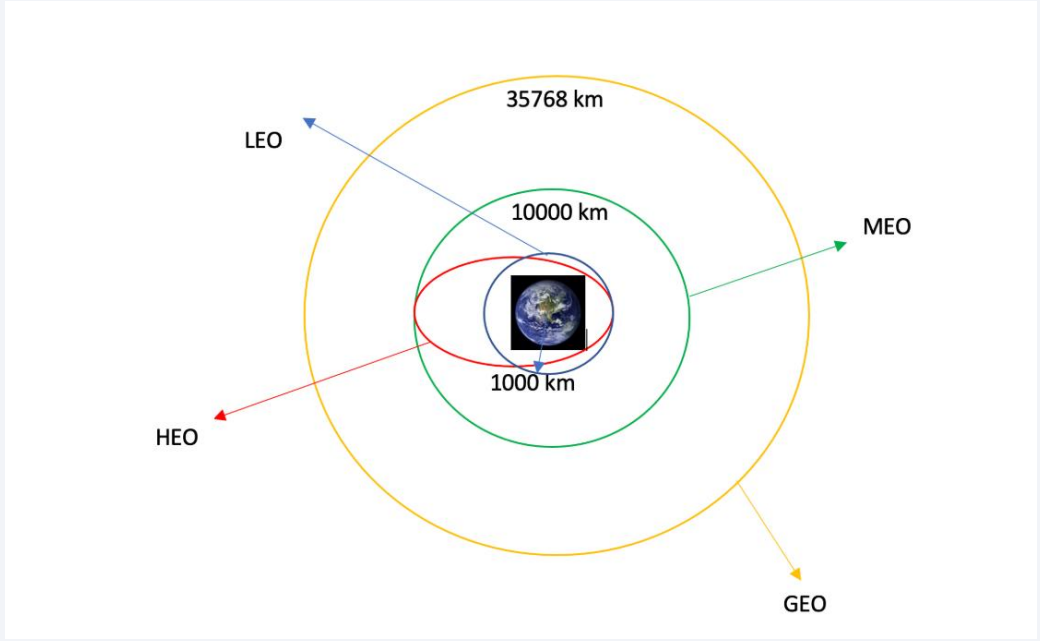
Data Collection - Scraping

- Web scraping was done using https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data
- Source Code: https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

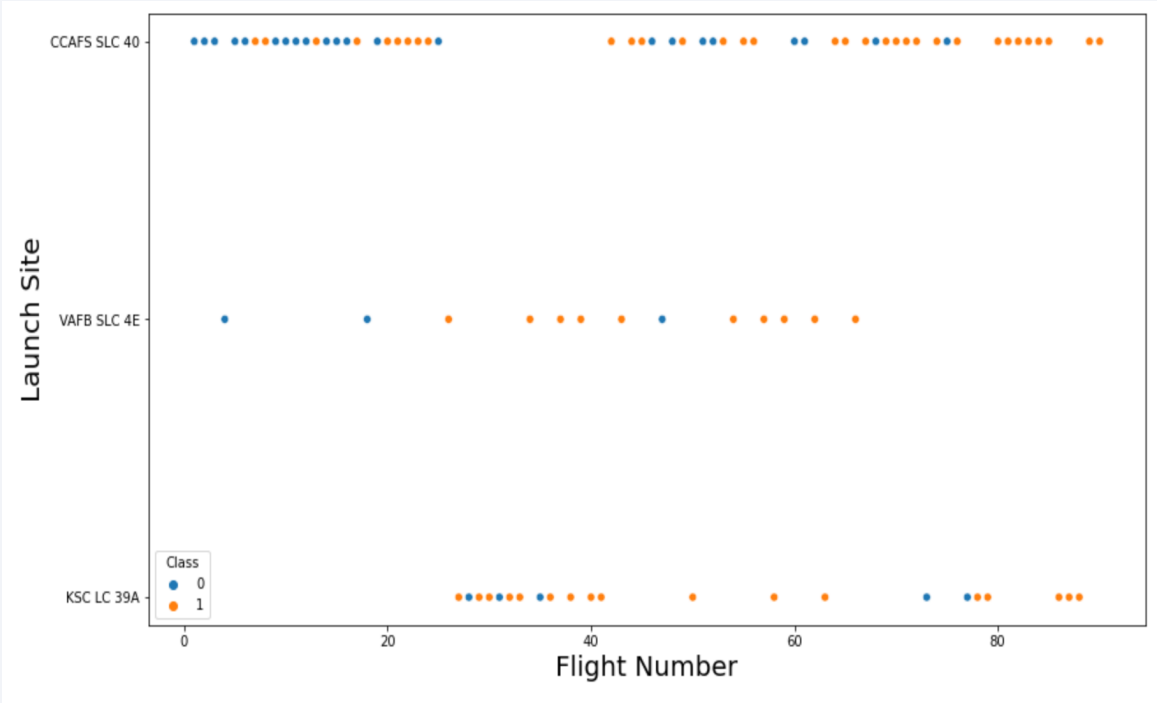
- Missing values and column type were identified
- Frequencies of occurrences of various launch sites and orbits were identified
- Mission outcome of each orbit was identified and label for outcome were assigned; 1 = success, 0 = failure
- Source Code: https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA → Summary → Landing outcome label

EDA with Data Visualization

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, Charts include:
    - Scatter: Flight number and payload vs launch site. Flight number and payload vs Orbit.
    - Bar: Success rate by Orbit
    - Line Plot: The number and occurrence of each mission outcome



- https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/edadataviz.ipynb

EDA with SQL

- The data is queried using SQL to answer several questions about the data such as:
  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
- https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Functions from the Folium libraries are used to visualize the data through interactive maps.
- The Folium library is used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway
- Various objects were added to know insight on:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?
- https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
- Using a pie chart and a scatterplot, the interactive site shows:
  - The total success launches from each launch site
  - The correlation between payload mass and mission outcome (success or failure) for each launch site
- These plots gave interactive information on successes on each launch site and correlation between payload mass and mission outcome of each launchsite
- https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/spacex_dash_app_fa.py

Predictive Analysis (Classification)

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Find the best combination of hyperparameters for each model
  - Evaluate the models based on their accuracy scores and confusion matrix

- https://github.com/proudfirst-debug/Data-Science-capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

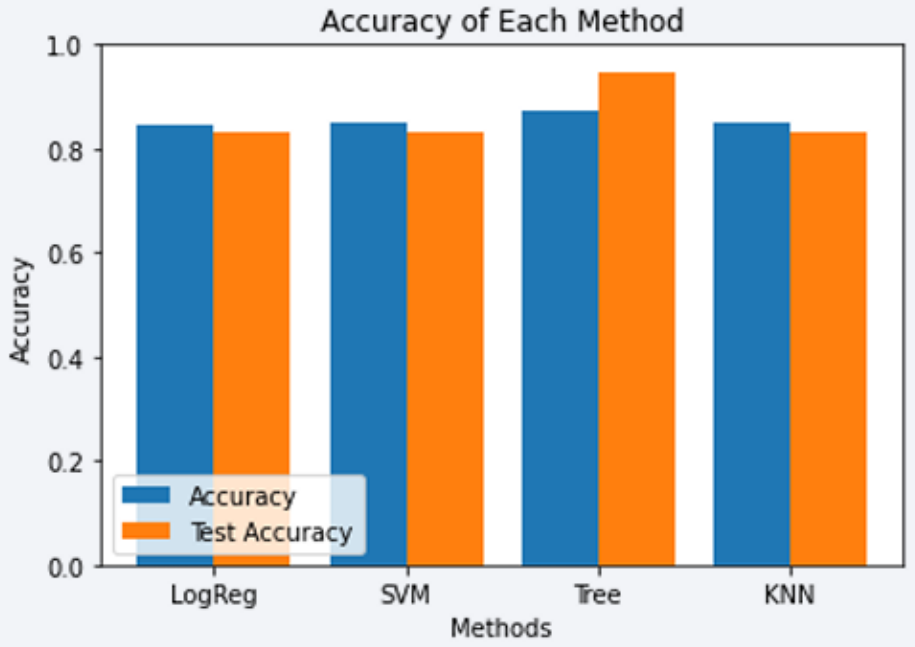
- Exploratory data analysis results:
- Space X uses 4 different launch sites
- The first launches were done to Space X itself and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 fiver year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average (Almost 100% of mission outcomes were successful)
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The number of landing outcomes became as better as years passed

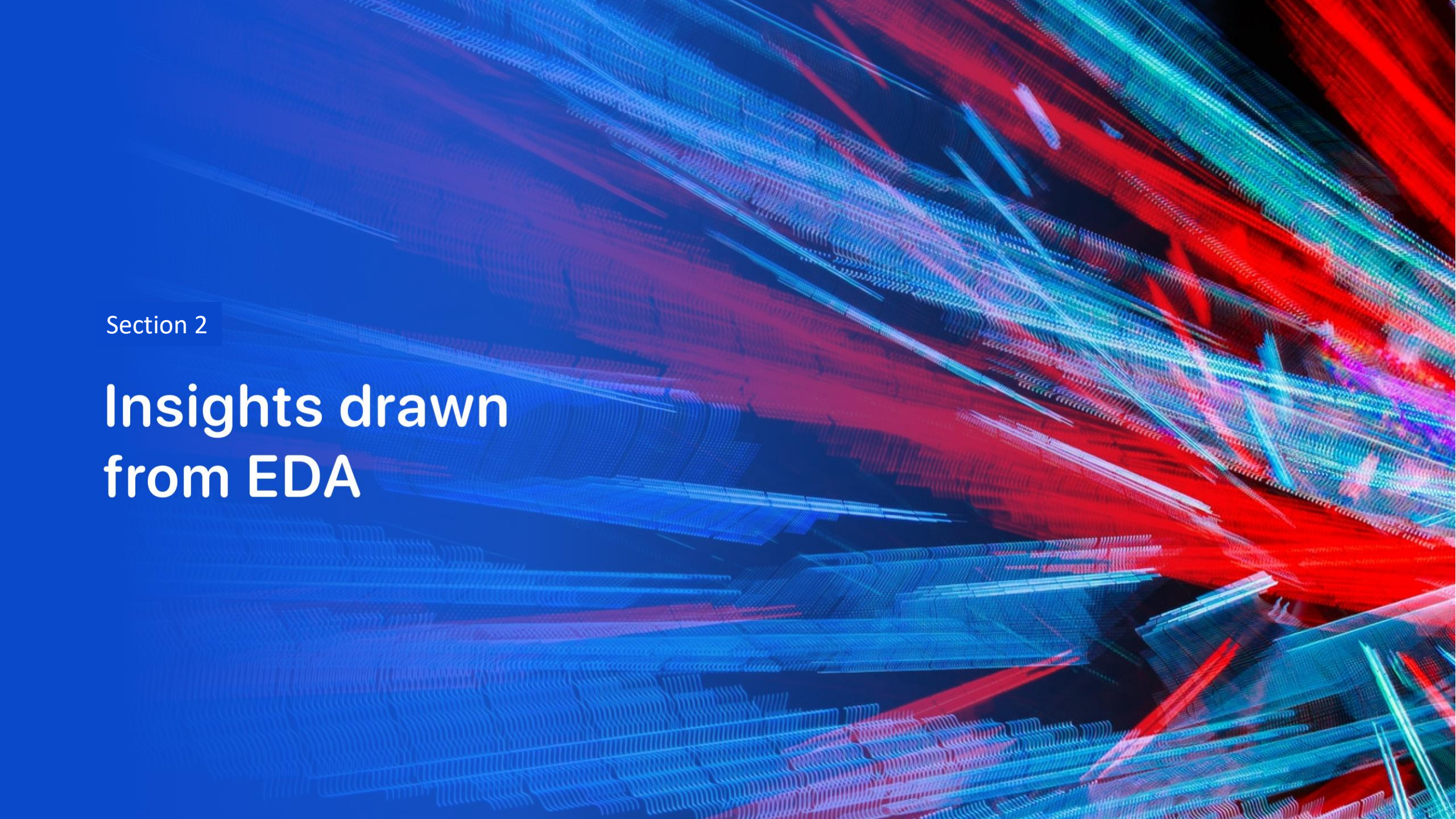
Results

- Using interactive analytics, it was possible to identify that launch sites used to be in safe places, near the sea, for example and have a good logistic infrastructure around
- Most launches happened at launch sites of East Coast

Results

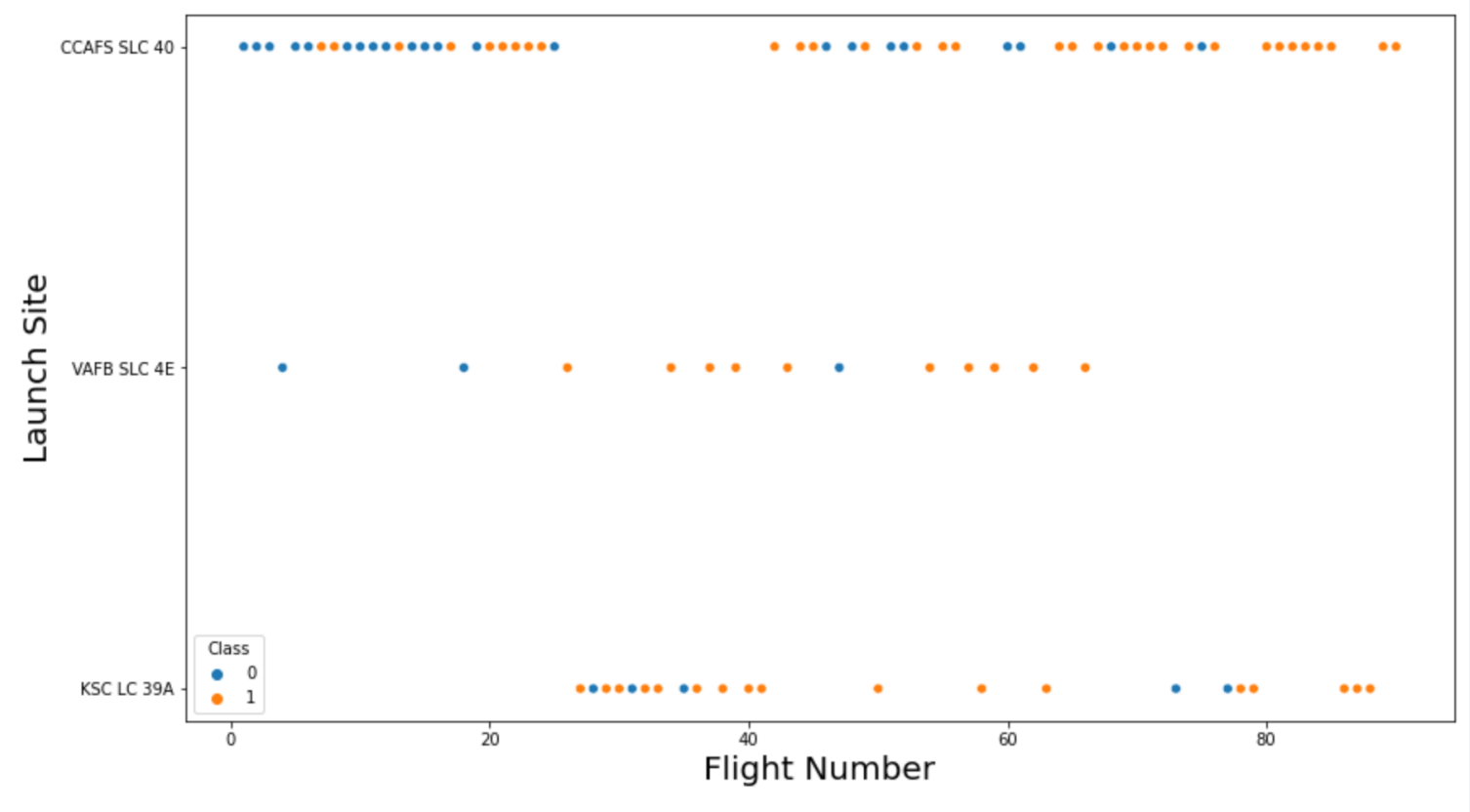
- Predictive Analysis showed that the Decision Tree Classifier is the best model to predict successful landings
- Accuracy over 87% and accuracy for test data over 94%




Section 2
Insights drawn from EDA

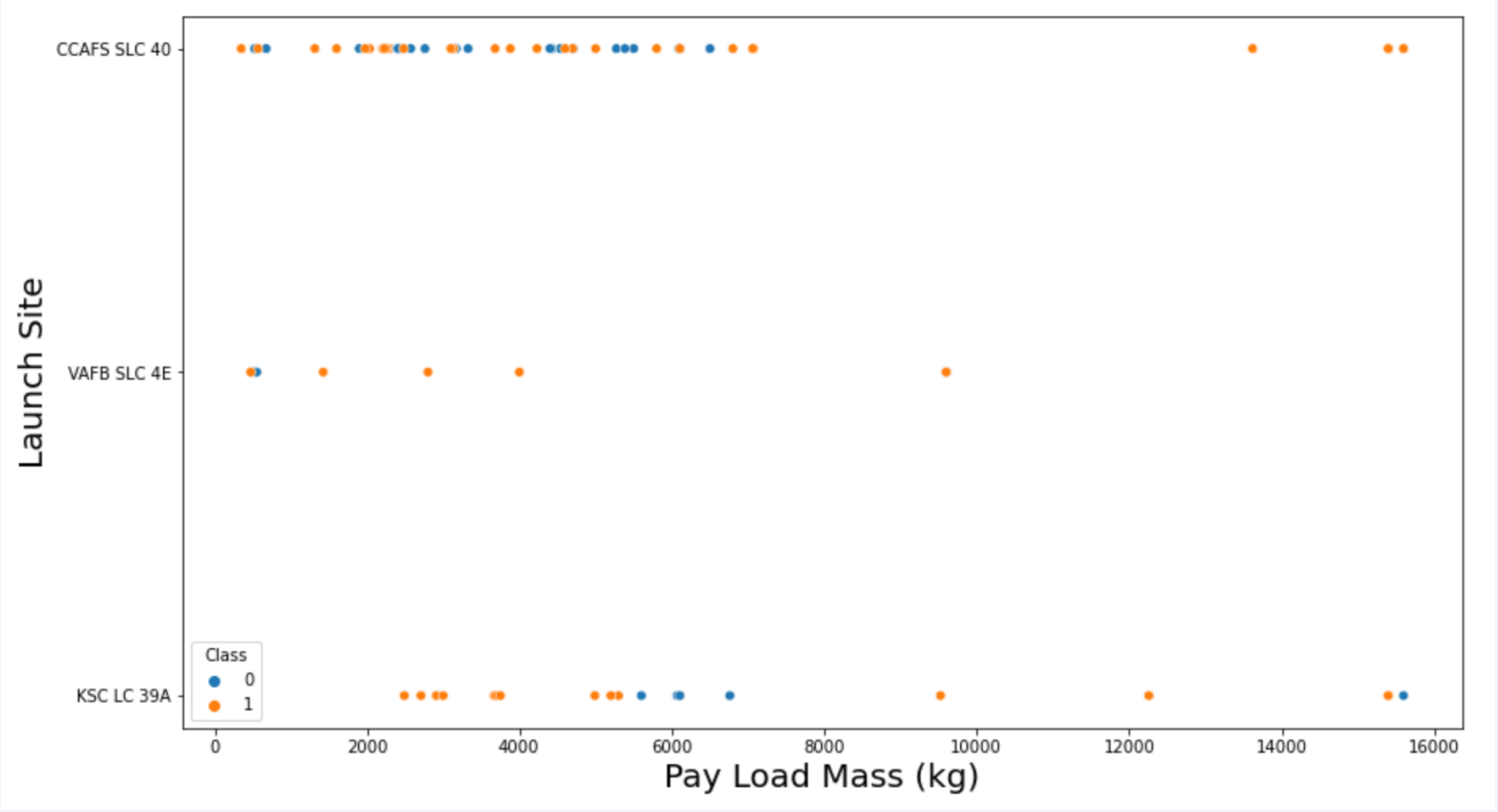
Flight Number vs. Launch Site

- Top best launch sites
  - CCAF5 SLC 40
  - VAFB SLC 4E
  - KSC LC 39A
- General success rate improved over time



Payload vs. Launch Site

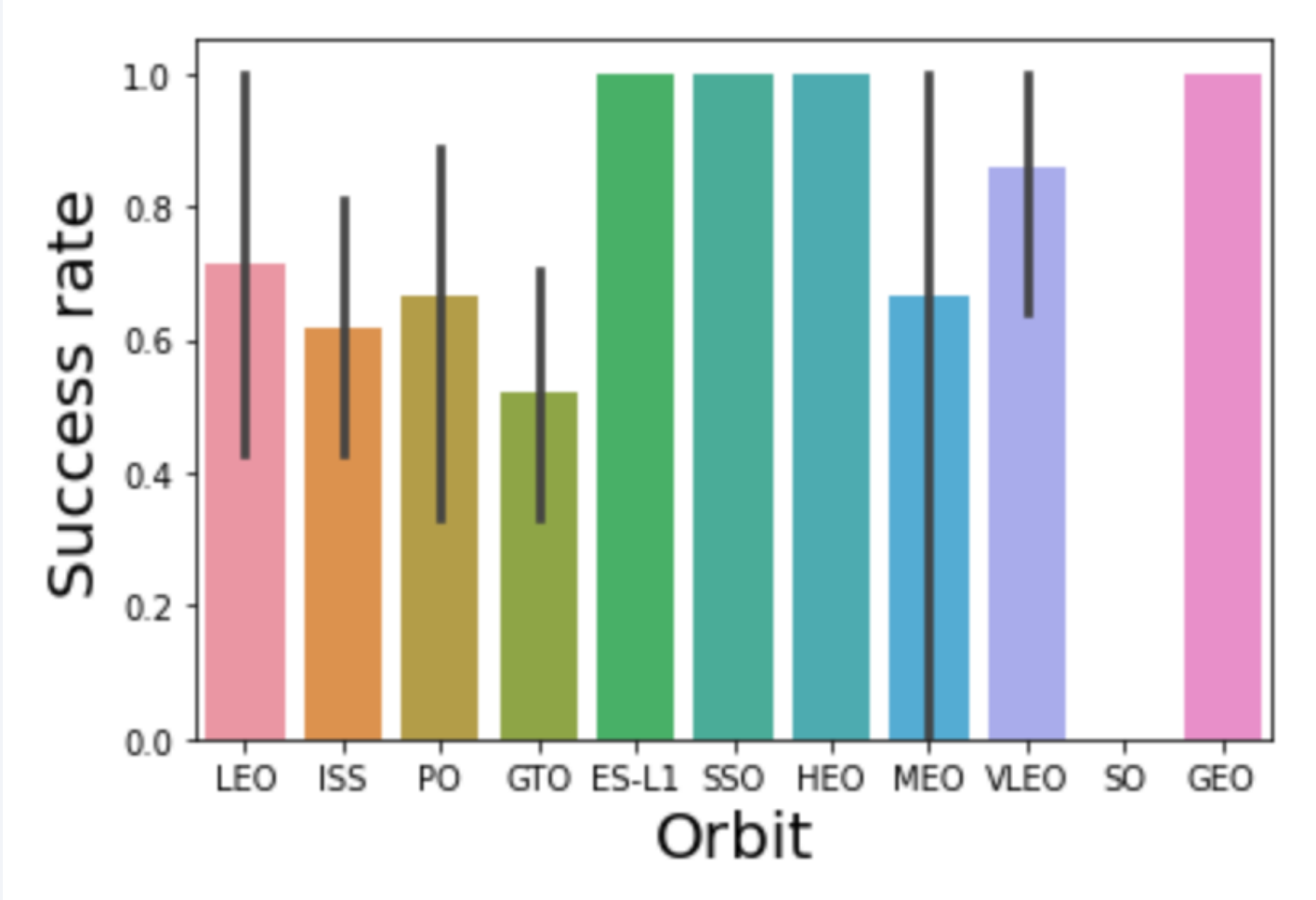
- Over 9,000kg payload have an excellent success rate
- Over 12,000kg payloads seem to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



Success Rate vs. Orbit Type

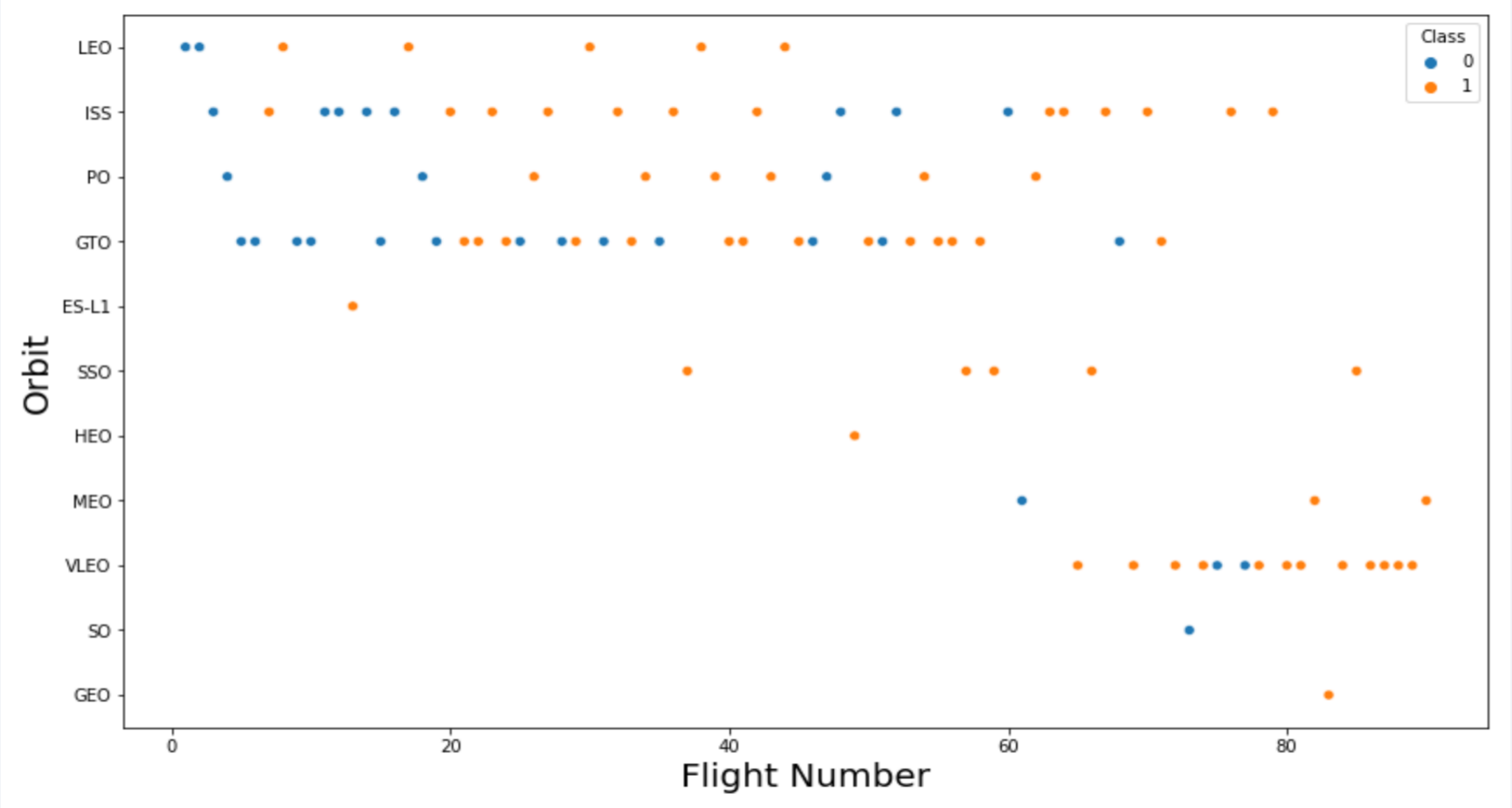
The highest success rate orbits are:

- ES-L1
- GEO
- HEO
- SSO



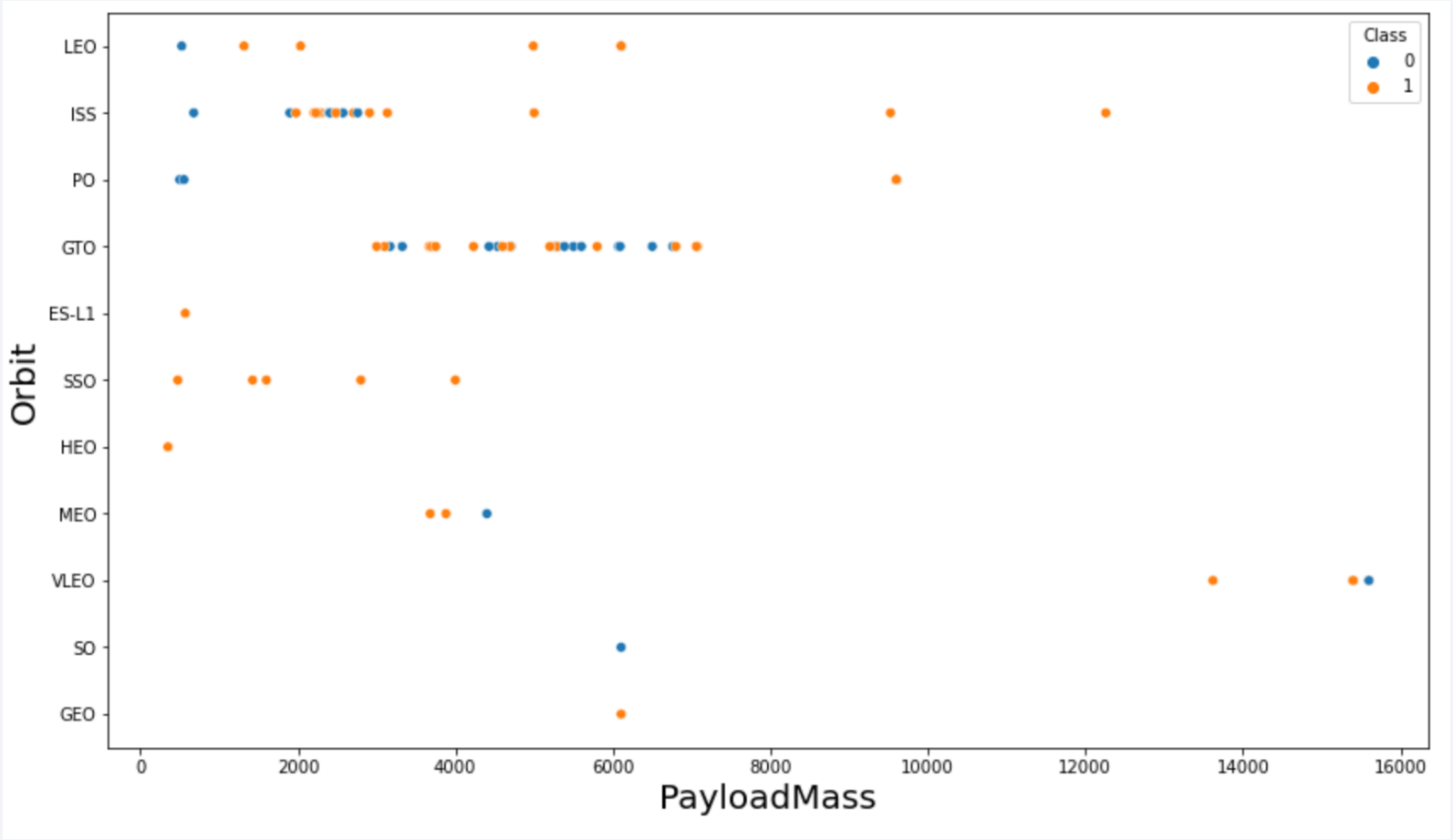
Flight Number vs. Orbit Type

- Success rate improved over time to all orbits
- VLEO orbit seems a new business opportunity (increase of its frequency)



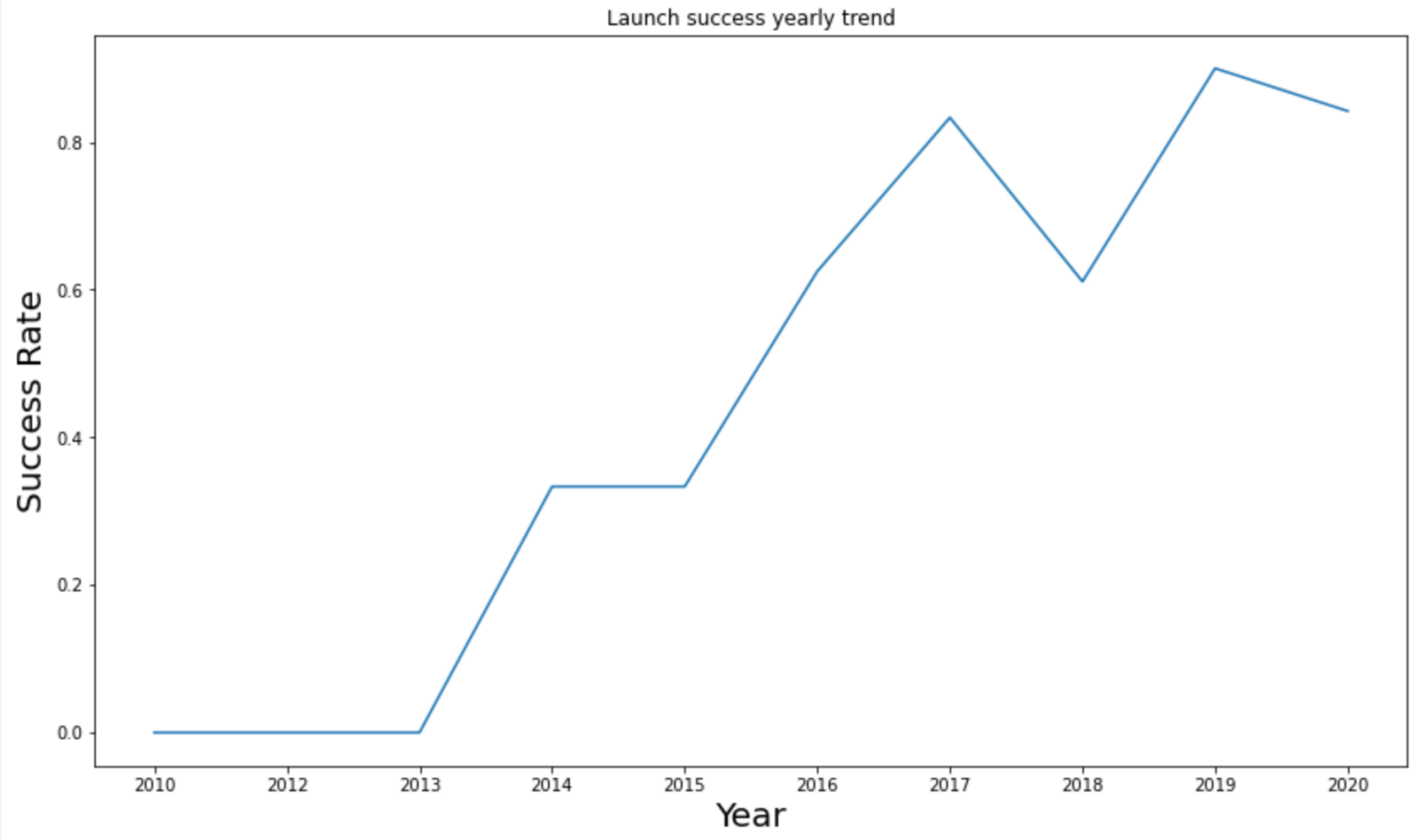
Payload vs. Orbit Type

- No relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO



Launch Success Yearly Trend

- First three years have no success
- Success rate started increasing in 2013 and continued until 2020



All Launch Site Names

- Four launch sites
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Otained by selecting unique occurrences of “launch_site” values

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Here we can see five samples of Cape Canaveral launches

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attempt

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg for booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

- By filtering data by successful landing outcome on ground pad and getting the minimum value for the date, it’s possible to identify the first occurrence: Min Date 2015-12-22

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four boosters have success in drone ship and have payload mass greater than 4000 but less than 6000, these are:
  - F9 FT B1021.2
  - F9 FT B1031.2
  - F9 FT B1022
  - F9 FT B1026

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes
  - Success 99
  - Success (payload status unclear) 1
  - Failure (in flight) 1
- Grouping mission outcomes and counting records for each group led us to the summary above

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass
- These are the boosters which have carried the maximum payload mass registered in the dataset

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

Booster Version
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

- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015
  - F9 v1.1 B1012 CCAFS LC-40
  - F9 v1.1 B1015 CCAFS LC-40
- The list above has only two occurrences

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017 03-20
- This view of data alerts us that “No attempt” must be taken in account

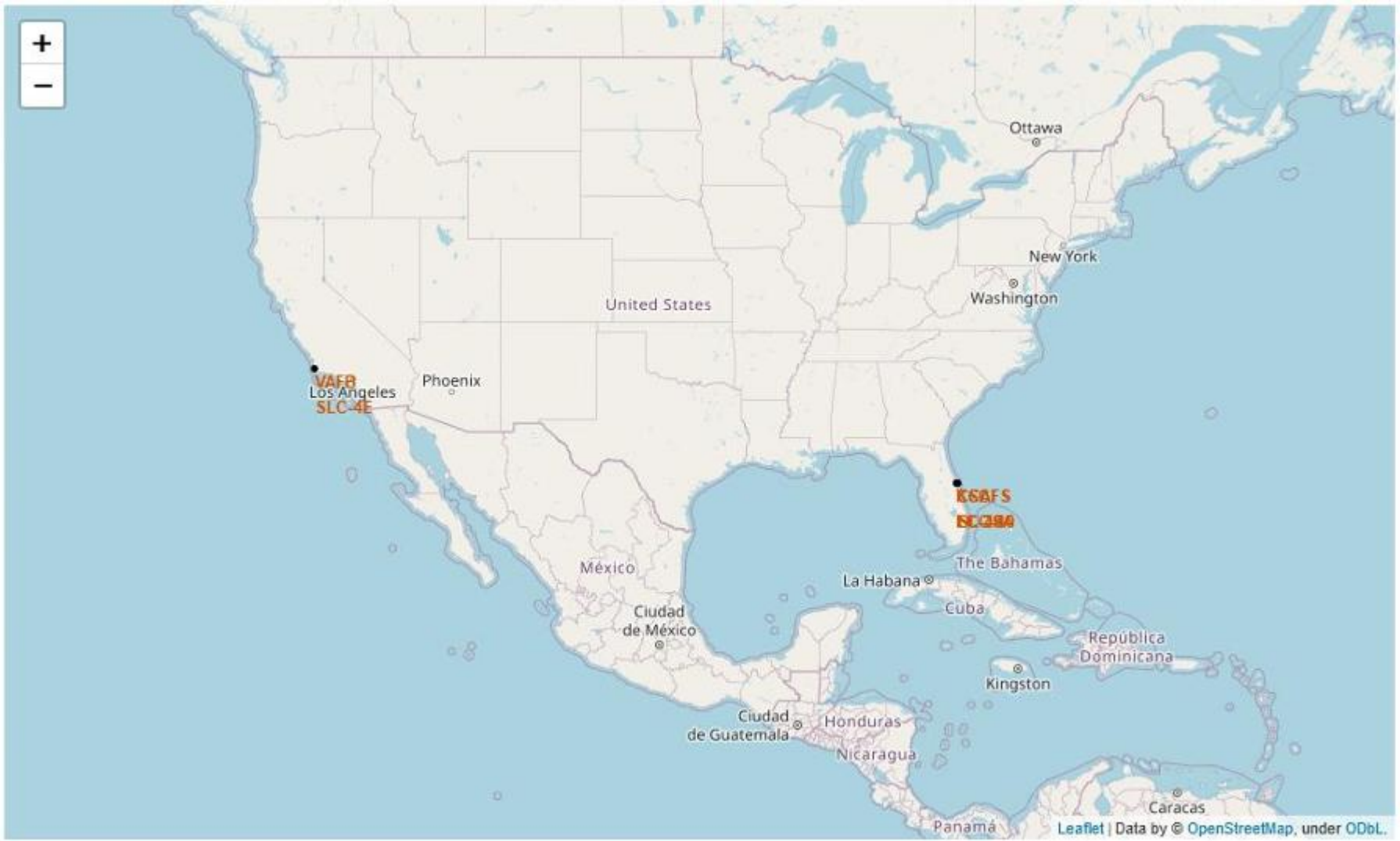
Landing Outcomes	Occurrences
No attempt	10
Failure - drone ship	5
Success - drone ship	5
Controlled – ocean	3
Success - Ground pad	3
Failure – parachute	2
Uncontrolled – ocean	2
Precluded – drone ship	1

Section 3

Launch Sites Proximities Analysis

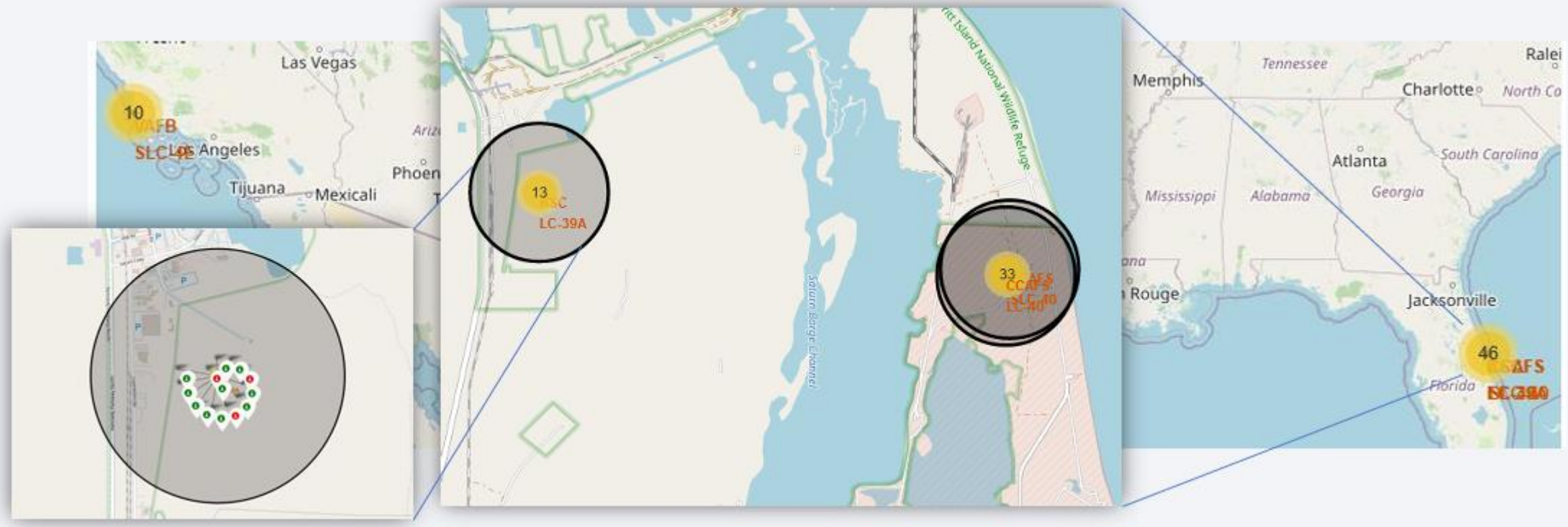
Launch Sites (All)

Launch sites are near sea, probably by safety, but not too far from roads and 38 railroads



<Folium Map Screenshot 2>

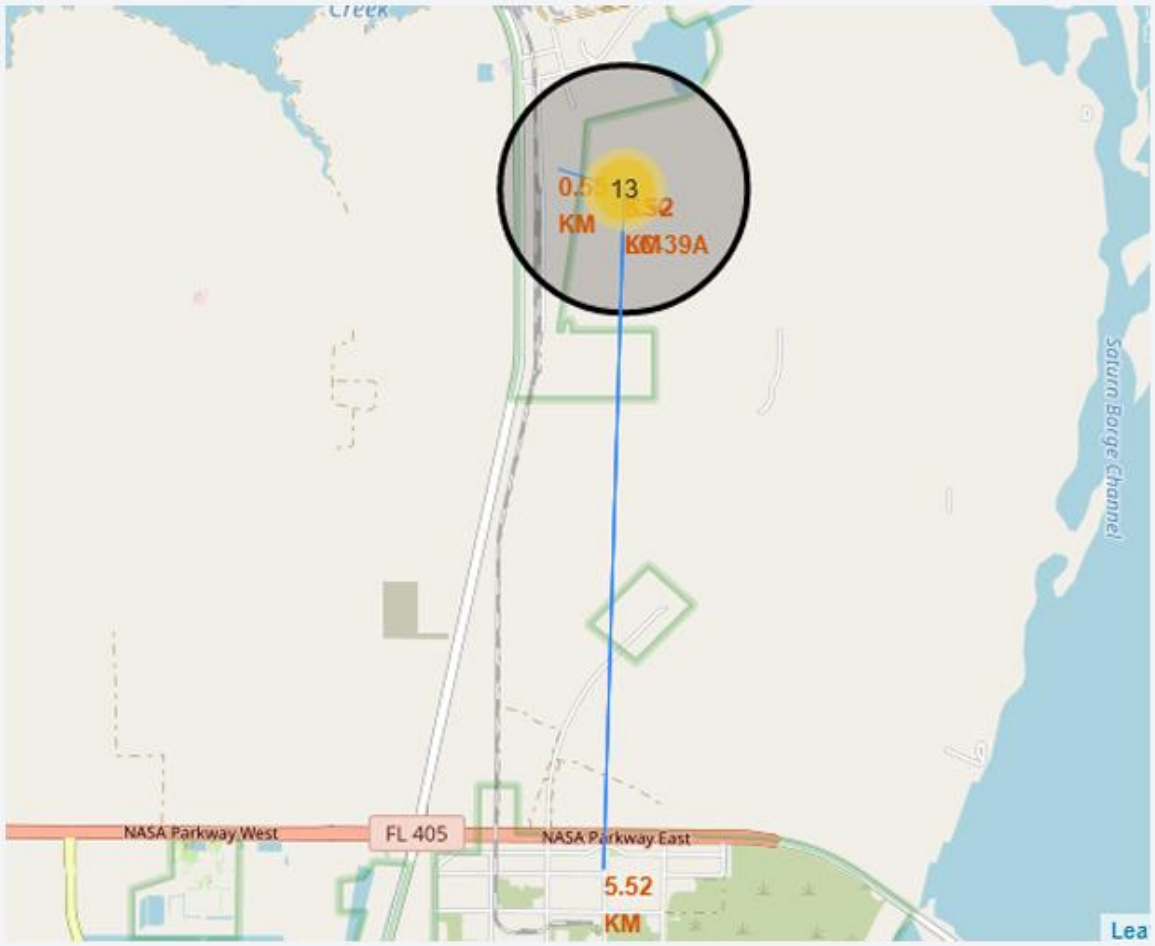
Green markers indicate successful and red ones indicate failure

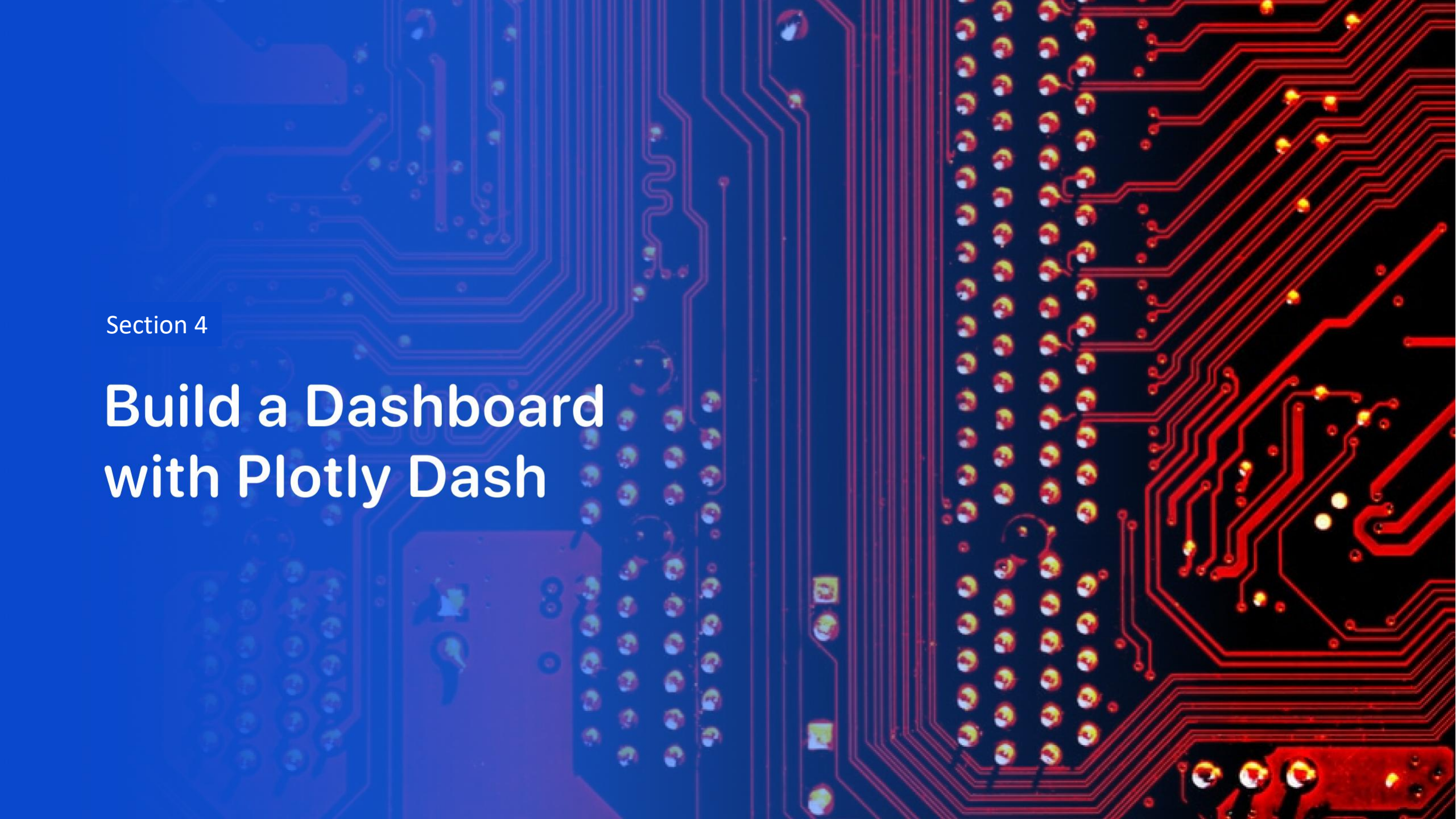


Green markers indicate successful and red ones indicate failure

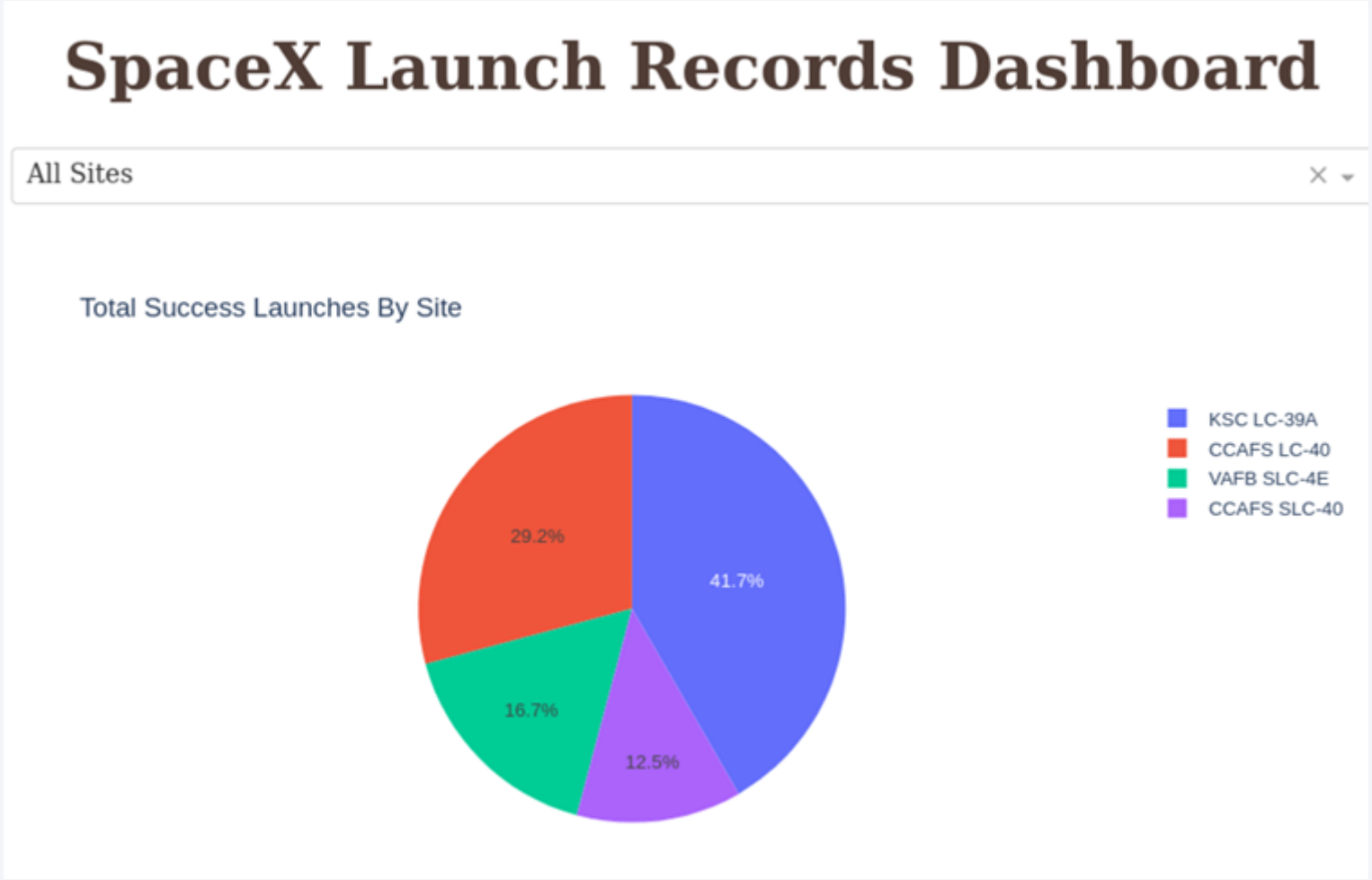
Safety

Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas




Section 4
Build a Dashboard with Plotly Dash

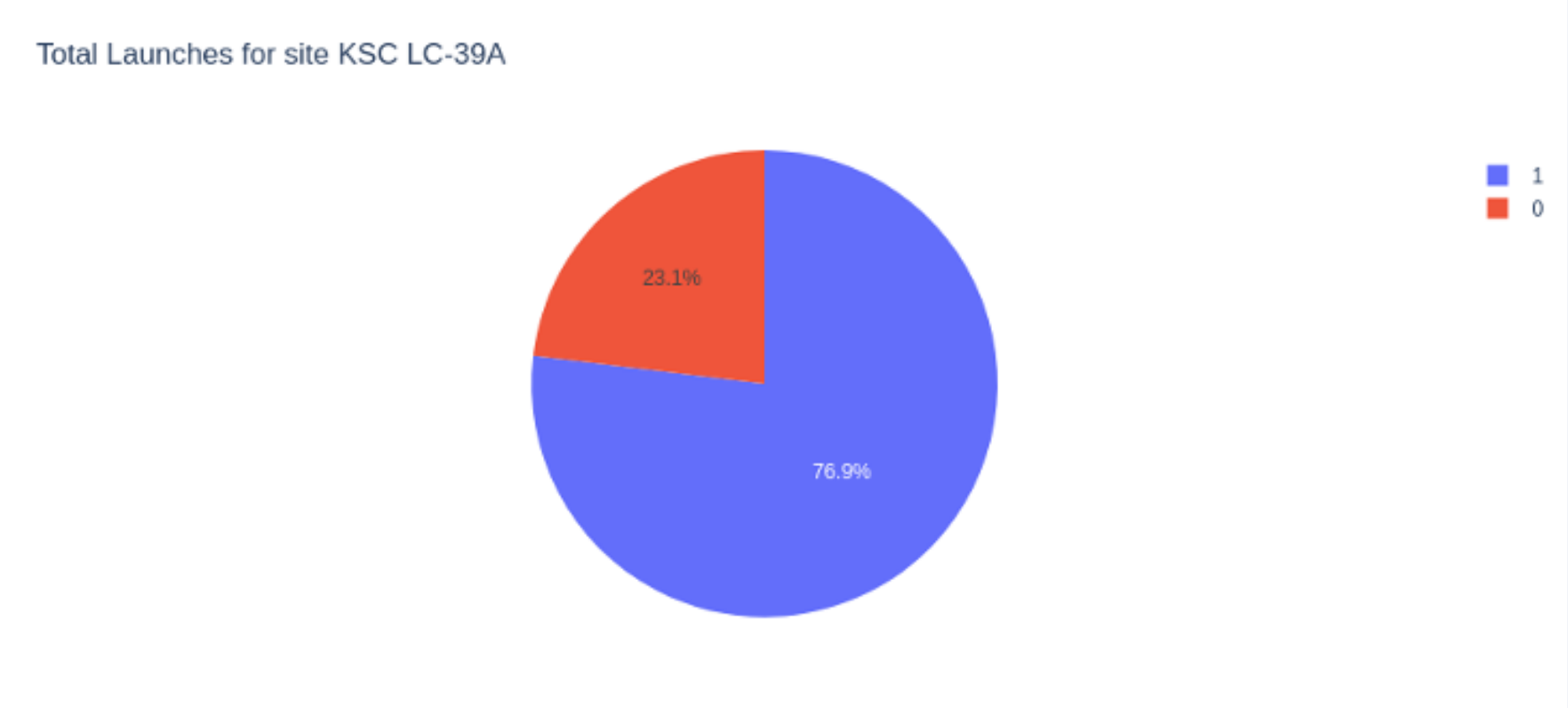
Successful Launch Sites



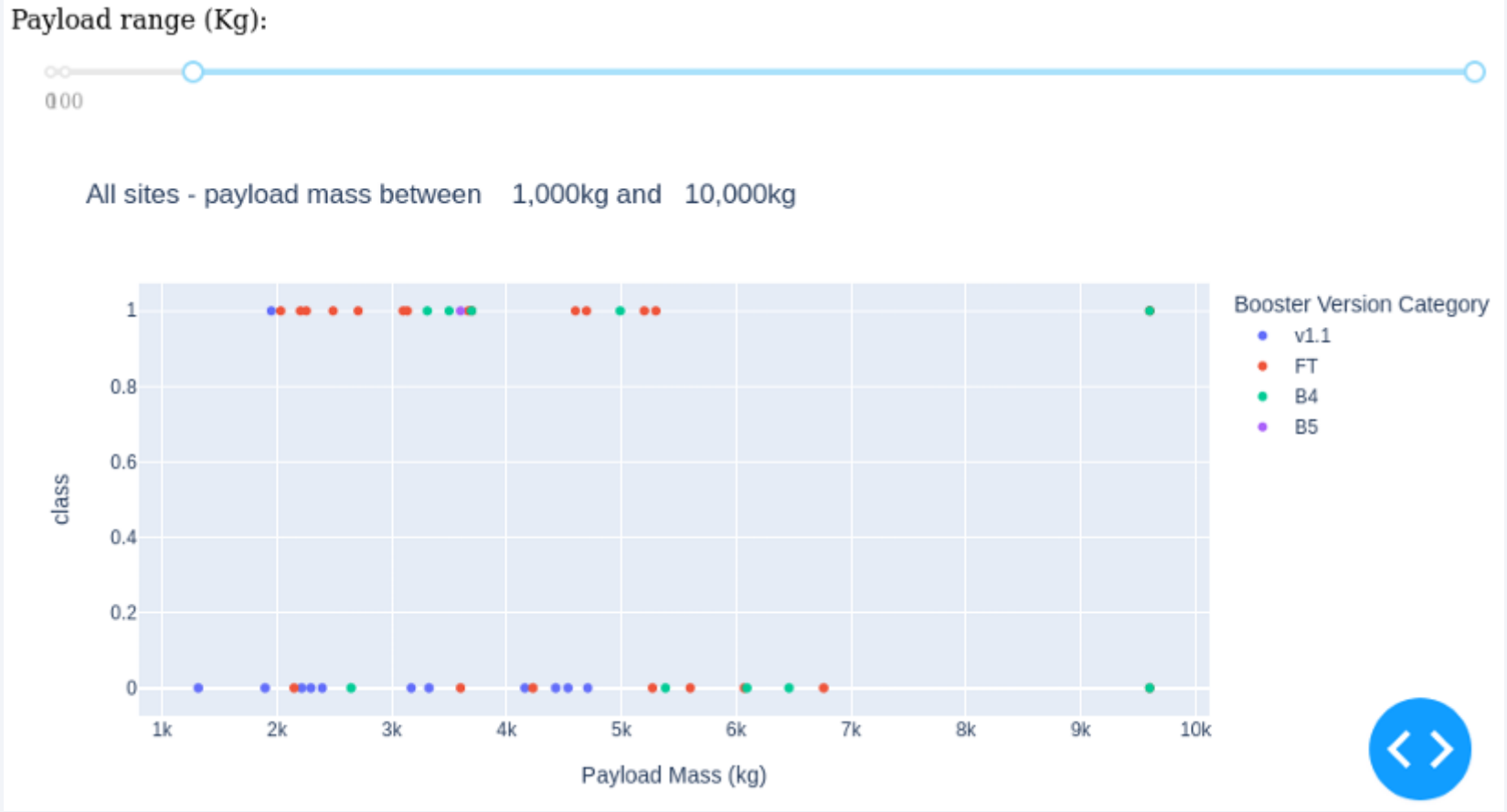
Launch site plays important role in success

KSC LC-39A – Launch Success

- 76.9% of launches are successful



Launch Outcome vs Payload



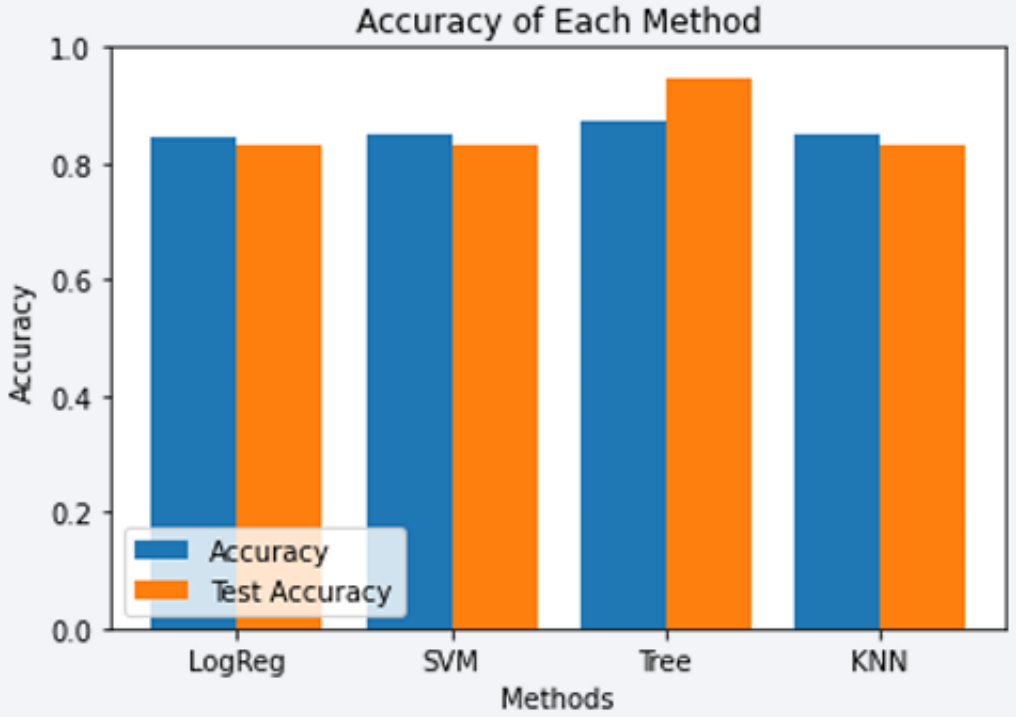
- The most successful combination: Payloads under 6,000kg and FT boosters

Section 5

Predictive Analysis (Classification)

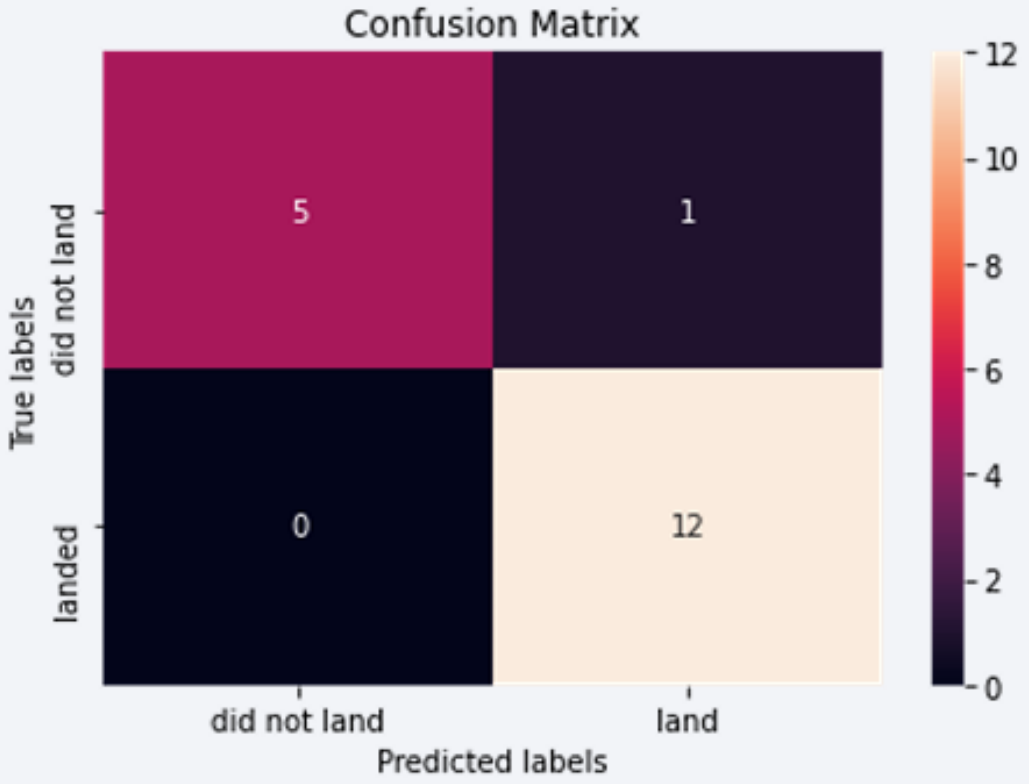
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted
- The model with the highest classification accuracy is the Decision Tree Classifier, which has an accuracy of over 87%



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative



Conclusions

- Different data sources were analyzed, refining conclusions along the process
- The best launch site is KSC LC-39A
- Successful landing outcomes seem to improve over time
- Decision Tree Classifier can be used to predict successful landings and increase profits

Appendix

- GitHub Repository of the complete project
- https://github.com/proudfirst-debug/Data-Science-capstone-project

Thank you!