Winning Space Race with Data Science

Hamza Elhaj
June 5, 2022

Outline

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

Executive Summary

- Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards
- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- Findings of this project was shown that machine learning techniques has a high abilities to predict the outcomes of rocket landing with high success rate.
- Decision tree model shows the best prediction performance among the other trained models with training accuracy of 87.5% and testing accuracy of 94.4%.

Introduction

- **Project background and context:**

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

- **Problems you want to find answers:**

In this case, 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. Therefore, this project aims to determine whether a rocket will land on earth successfully or not.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - The source information of this project was extracted from the Wikipedia website using web scrapping technique. The dataset used in this project was generated by making a get request to the SpaceX API.
- **Perform data wrangling**
  - The dataset was first preprocessed by calculating the number of launches on each site. We then calculated the number and occurrence of each orbit, by orbit type. Finally, we created a landing outcome label from the Outcome column.

Methodology Executive Summary (Continued)

- **Exploratory data analysis (EDA) using visualization and SQL**
  - The information in the dataset was explored using some data visualization techniques to explore the hidden patterns from the data. The dataset was also uploaded to the database system of ibm-db2 to preserve and maintain the data and extract the needed information at the needed time.
  - Perform EDA using data visualization and SQL.
- **Interactive visual analytics using Folium and Plotly Dash**
  - Interactive visualizations were created using Folium and Plotly Dash libraries.
- **Predictive analysis using classification models**
  - Four supervised machine learning techniques were used to build predictive models to predict the outcomes of rocket landing.

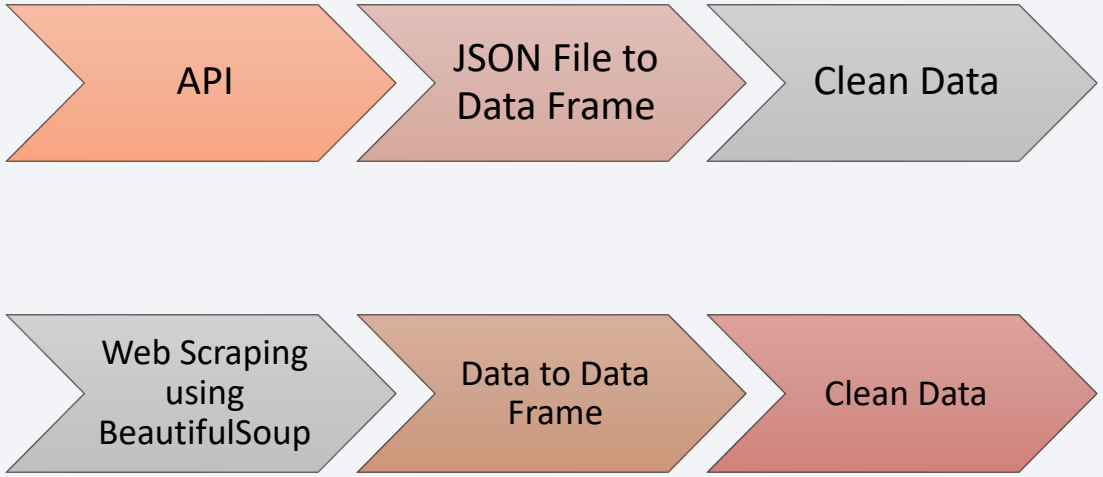
Data Collection

- Data collection process involved a combination of API requests from SpaceX public API and web scraping data from a table in Space X's Wikipedia entry.
- **Space X API Data Columns:**

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

- **Wikipedia Webscrape Data Columns:**

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.



Data Collection (API Request)

- **Data was collected as follow:**
  - In this study, we managed to collect the Data with an API.
  - Data that are gathered from the API, specifically the **SpaceX REST API**, provided us with information about launches, including the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - We performed a get request using the requests library to obtain the launch data, which we will use to get the data from the API. This result can be viewed by calling the .json() method. The response will be in the form of a JSON, specifically a list of JSON objects.
  - To convert this JSON to a data frame, we used json_normalize function. Then we parsed the data from those tables and converted them into a Pandas data frame for further visualization and analysis.

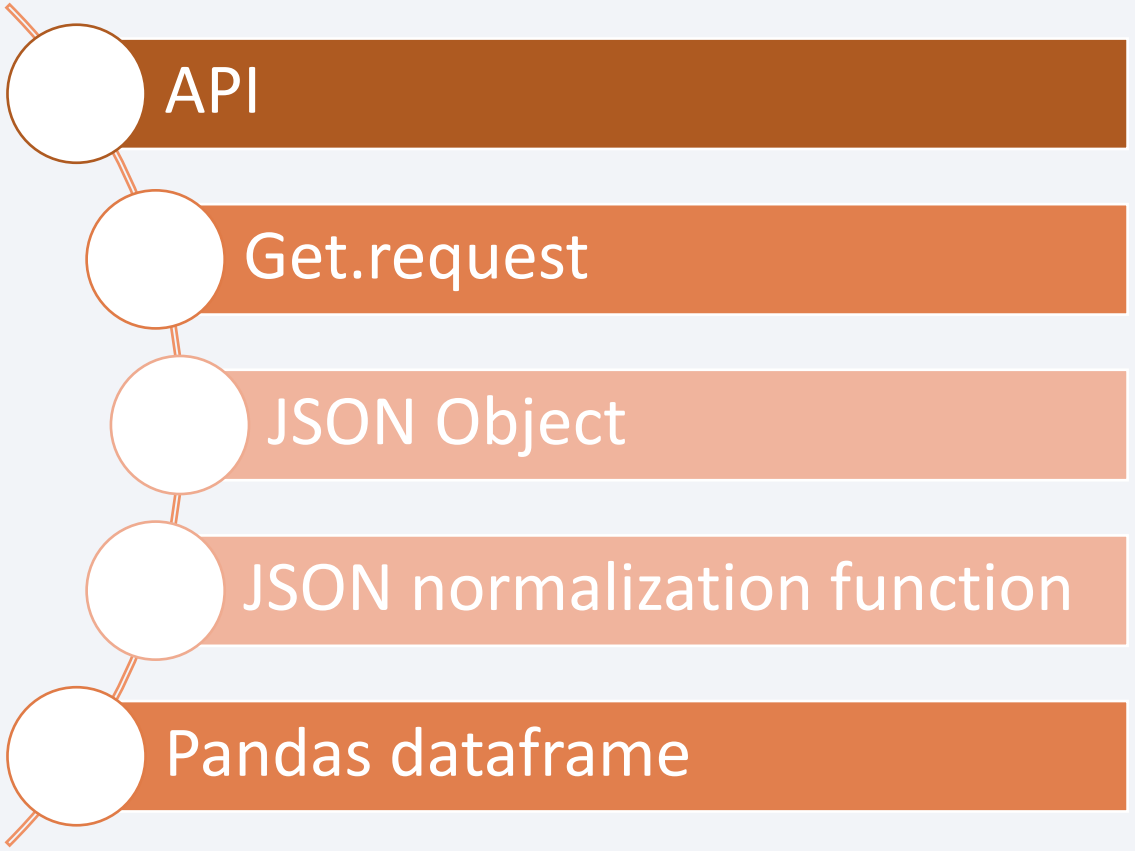
Data Collection (Web Scraping)

- **Data was collected as follow:**
  - Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages.
  - In this study, we also collected some information about the Falcon 9 launch from Wikipedia website.
  - First, we performed an HTTP get method to request the Falcon9 Launch HTML page, as an HTTP response.
  - Then we created a BeautifulSoup object from the HTML response.
  - We extracted all column/variable names from the HTML table header, since we want to collect all relevant column names from the HTML table header. Then, we need to iterate through the <th> elements and apply the provided extract_column_from_header() to extract column name one by one.
  - Finally, we created a data frame by parsing the launch HTML tables.

Data Collection – SpaceX API

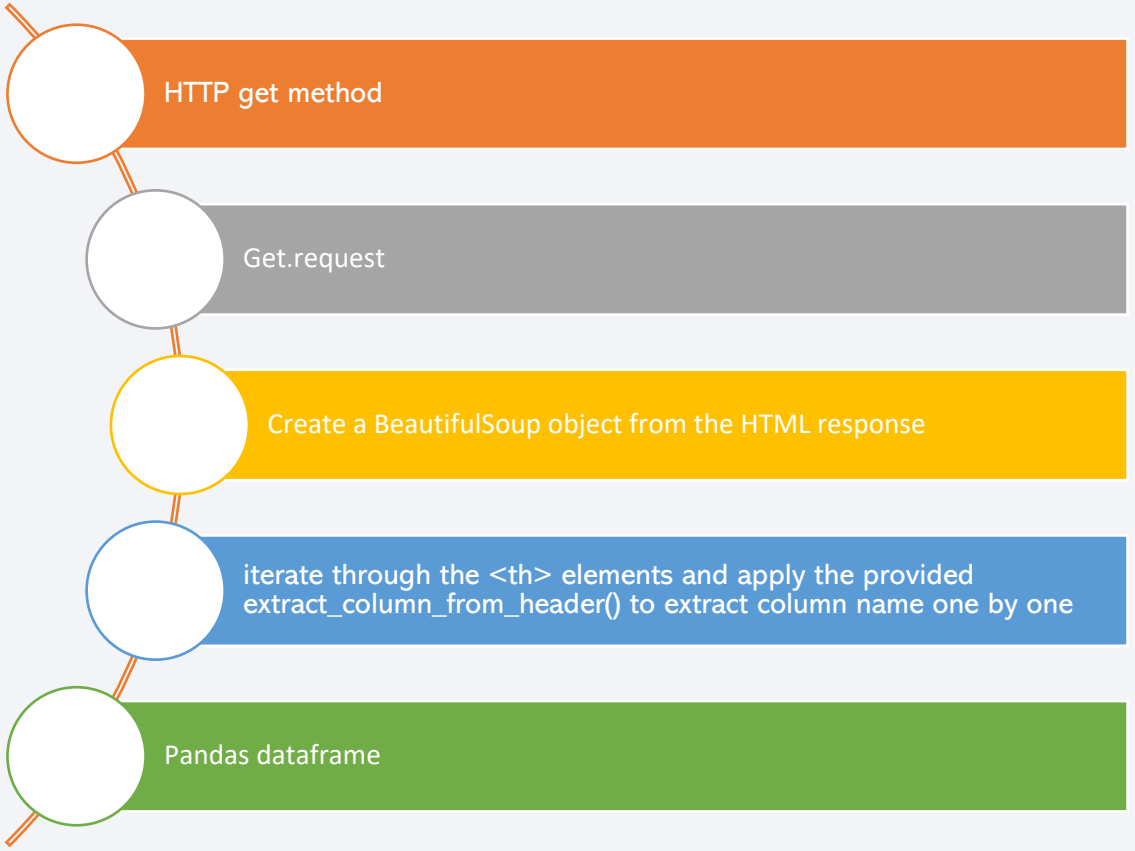
- The following figure describes the process of data collection in four main steps to collect the data.
- For more information, please find the below link to my GitHub Repository:

GitHub URL



Data Collection - Scraping

- The following figure describes the process of data collection in web scraping process.
- For more information, please find the below link to my GitHub Repository:
- GitHub URL



Data Wrangling

- In this study, 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. The main task from data wrangling is to convert these outcomes into training dataset for the next stage.
- Data wrangling was involved checking the messing data from each attribute. We then grouped features based on their success rate of landing. After that, features encoding were applied to convert categorical features into numerical features. One-hot encoding technique was used for this purpose. Finally, we created a final dataset to be used for model training. Data wrangling steps can be found below.

- For more information, please find the below link to my GitHub Repository: GitHub URL

EDA with Data Visualization

- Multiple charts were constructed to extract insights from the datasets. We created two scatter plots to find the relationship between flight number and both pay load mass and lunch site. Another bar plot were constructed between the rocket orbit and the mean of class column to see the success rate for each orbit. We then created another scatter plot to explore if there is any relationship between the Flight Number and Orbit type. Another scatter plot was constructed from Payload vs. Orbit scatter to reveal the relationship between Payload and Orbit type. Finally, we created a line plot to understand the launch success yearly trend for the period of 10 years.
- For more information, please find the below link to my GitHub Repository:

GitHub URL

EDA with SQL

- Exploratory data analysis was also performed using SQL language. Queries used in this study was as follow:
  - Loaded data set into IBM DB2 Database.
  - Queried using SQL Python integration.
  - Queries were made to get a better understanding of the dataset.
  - Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes.
- For more information, please find the below link to my GitHub Repository:

  GitHub URL

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- For more information, please find the below link to my GitHub Repository:

  GitHub URL

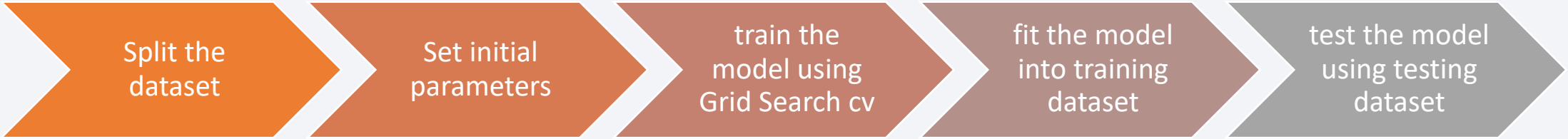
Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show the distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual sites and payload mass on a slider between0 and10000kg.
- The pie chart is used to visualize the launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.
- For more information, please find the below link to my GitHub Repository:

GitHub URL

Predictive Analysis (Classification)

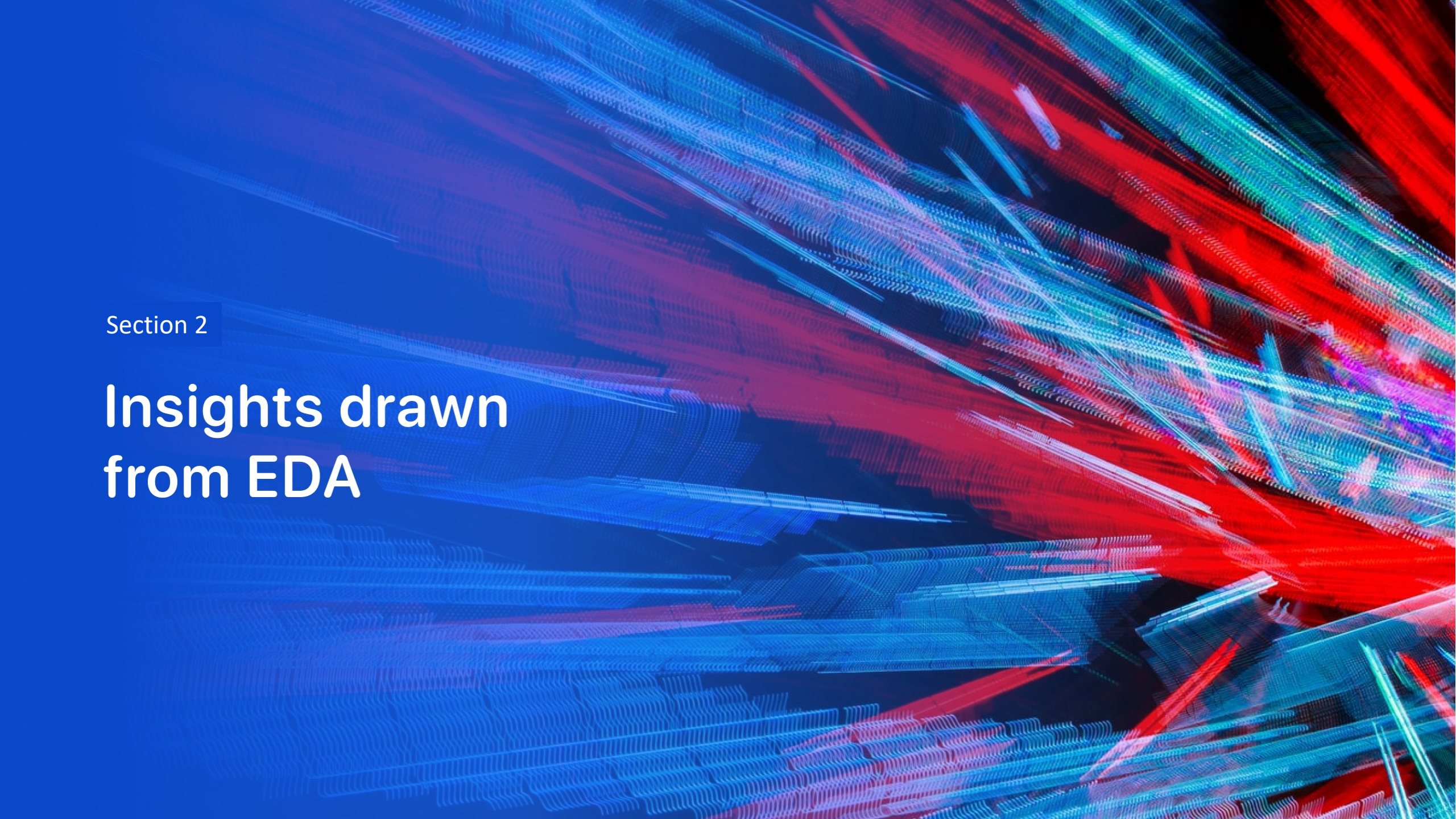
- After preparing the data for analysis, we created FOUR main SUPERVISED machine learning (ML) model techniques to predict the success lunch of the next rocket.
- The first step before training the model is splitting the dataset into training and testing data using train_test_split technique.
- We then set an initial parameters for each classifier. We created each model and then trained them using Grid Search Cross validation. Grid search technique is an effective technique that can be used to search for the best hyperparameters that can provide the best prediction outcomes of the model and provide parameters that deliver the optimal performance. This technique also help to reduce the bias-variance error through finding the trade-off border between both error types. We also applied 10-fold cross validation technique to divide the data during the training in order to reduce the bias in the model.
- We then developed 4 ML models and trained them on the training set and tested their performance on out-of-sample data using the testing dataset. The model development process can be found below

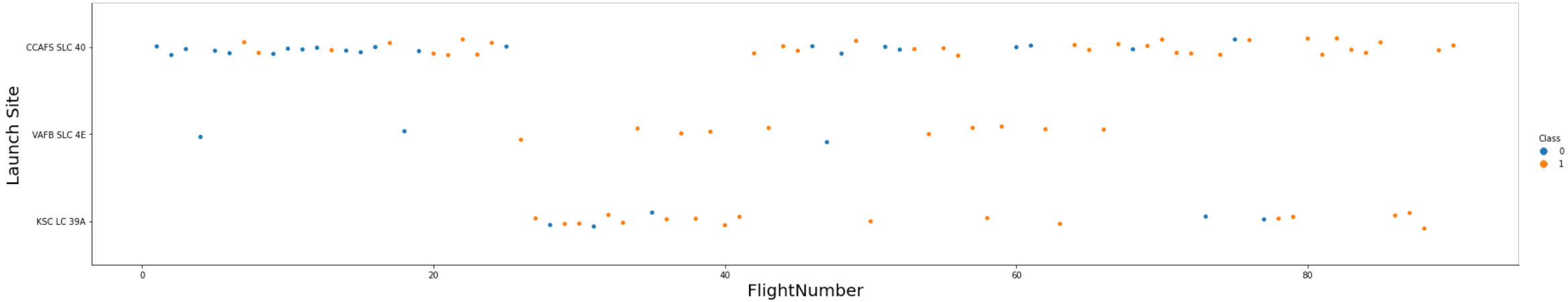


- The GitHub URL of the completed SpaceX scraping calls notebook can be found here: GitHub URL

Results

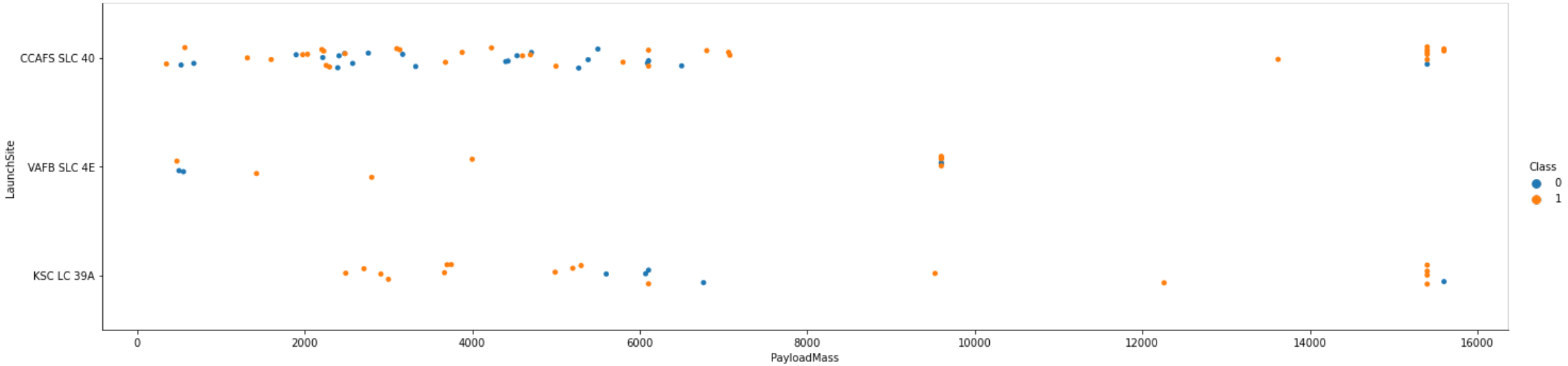
- Findings of this project was shown that machine learning techniques has a high abilities to predict the outcomes of rocket landing with high success rate.
- Decision tree model shows the best prediction performance among the other trained models with training accuracy of 87.5% and testing accuracy of 94.4%.


Section 2
Insights drawn from EDA



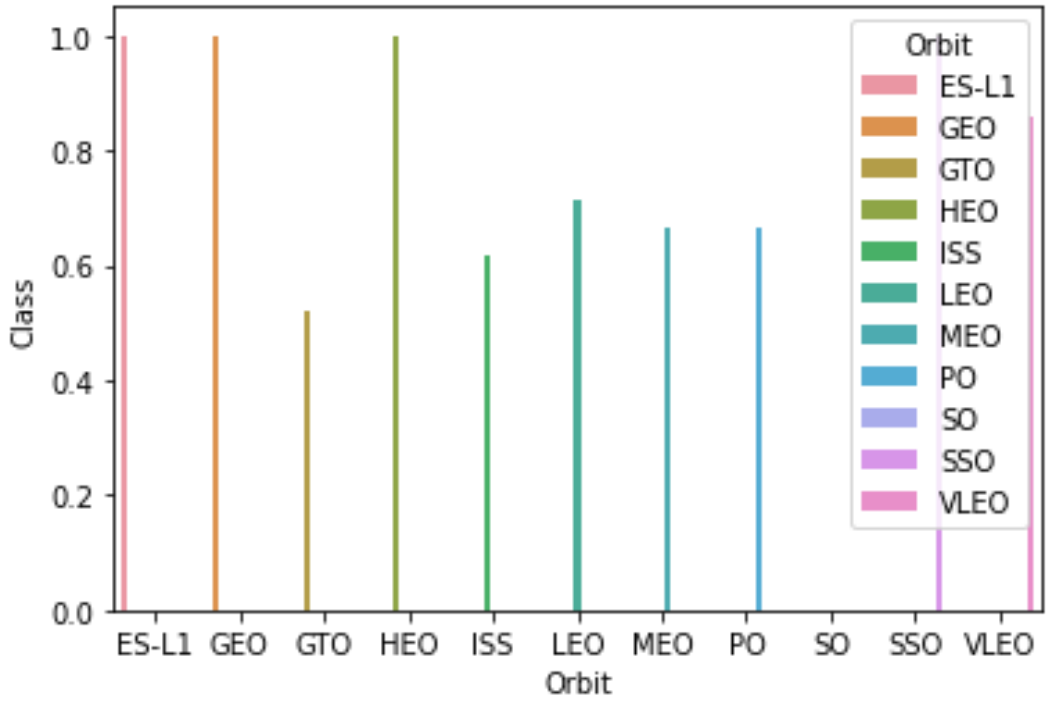
Flight Number vs. Launch Site

- It can be concluded from this graph that the recent flights shows a better success rate at each of the launching sites.



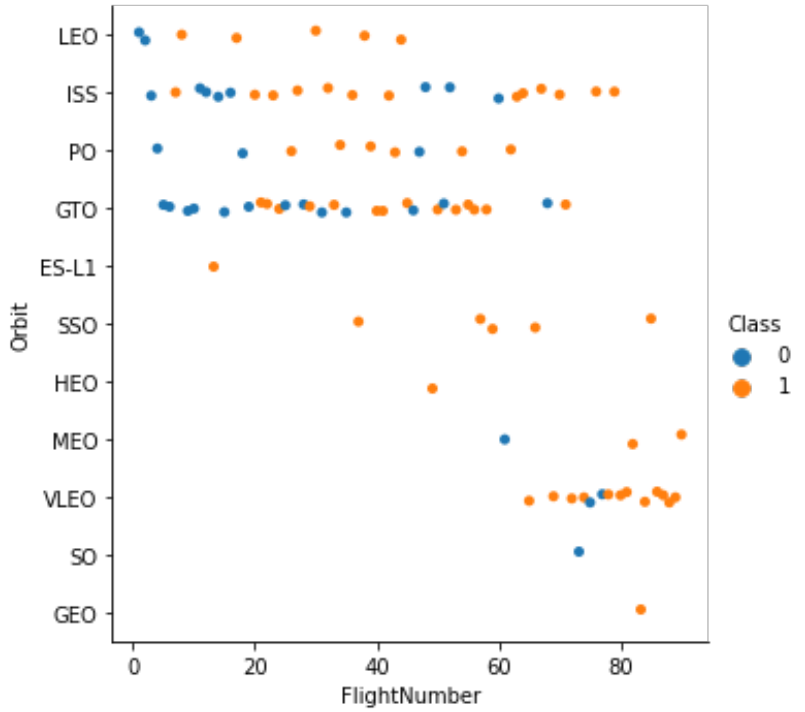
Payload vs. Launch Site

- It can be seen from the above figure that the launching site of CCAFS SLC 40 has the least success rate of landing among the other sites.



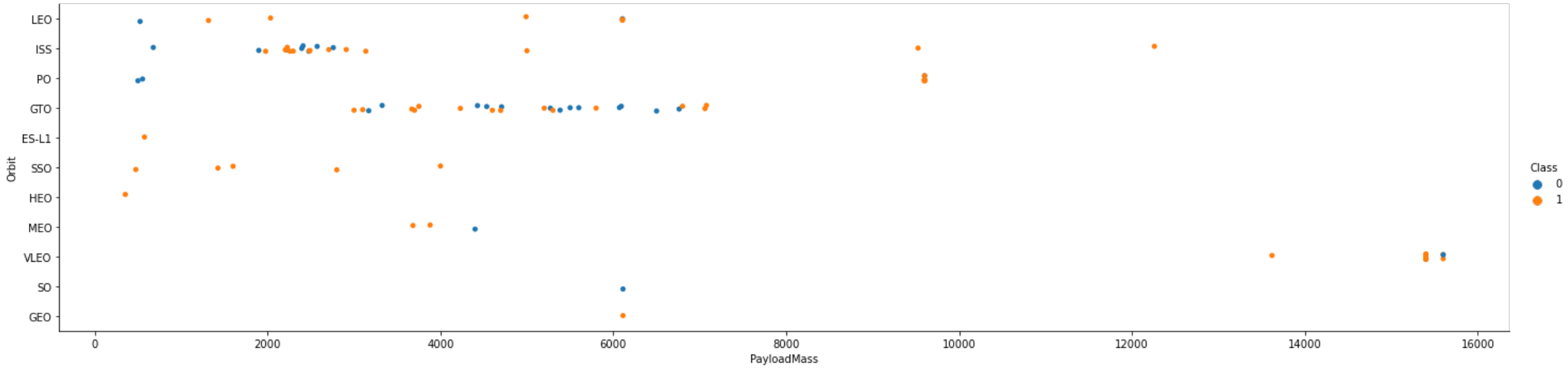
Success Rate vs. Orbit Type

- The figure above shows that the orbit SO has the lowest success rate among other orbits.
- While many orbits has a 100% success rate of landing such as ES-L1, GEO, HEO, and SSO.



Flight Number vs. Orbit Type

- It can be concluded from this figure that the latest flights has more success rate of landing than at the beginning time of flights.

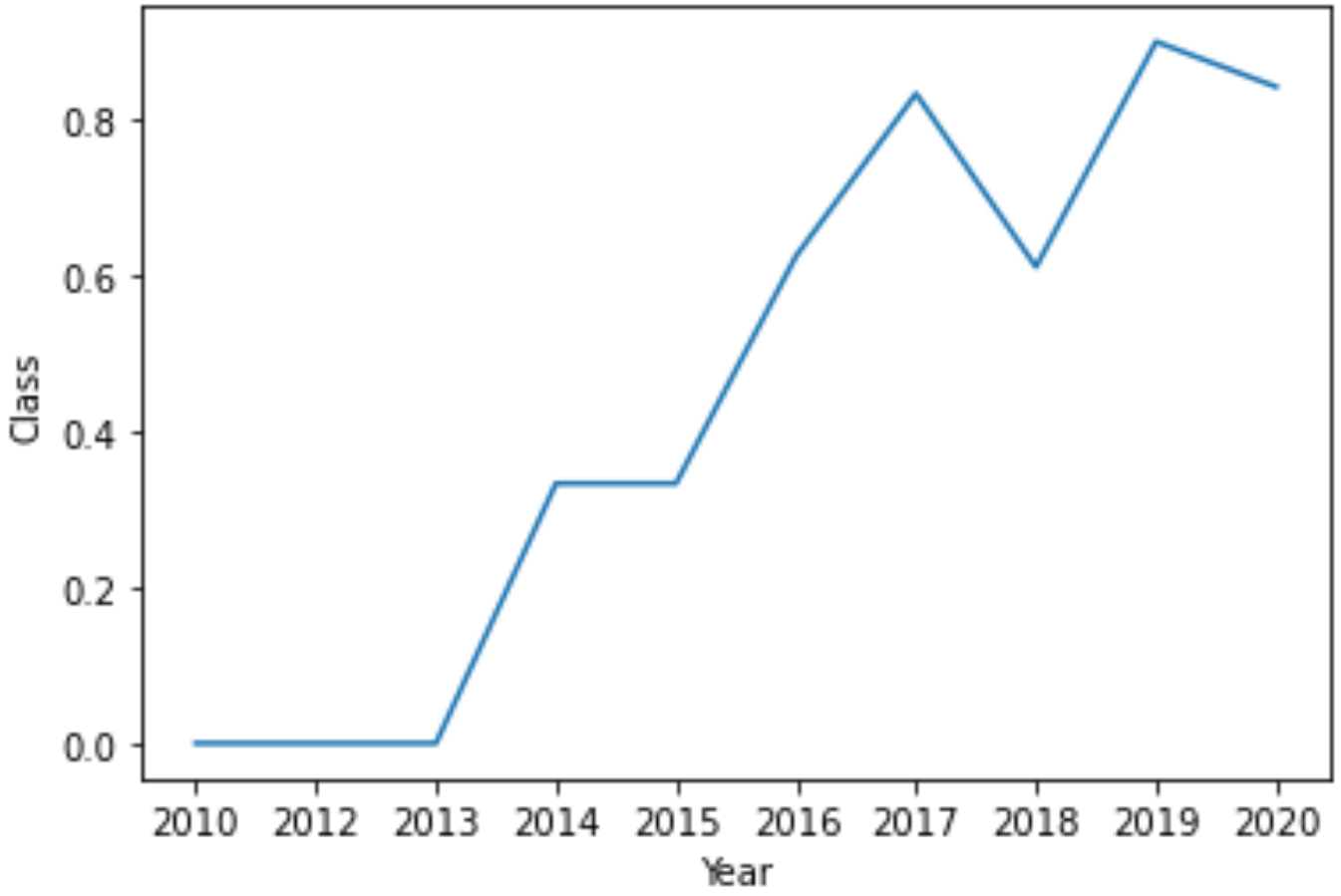


Payload vs. Orbit Type

- We can conclude from this figure that the success rate has a positive relationship with the pay load mass.
- The more the pay load mass, the better the success rate of different orbits.
- SSO, HEO, ES-L1, GEO orbits has the highest success rate.
- Some orbits such as VLEO has a better chance of a successful landing when the pay load mass is high, in despite of others like HEO that has a low rate of pay load mass.

Launch Success Yearly Trend

- We can see that since 2013, the success rate of rocket landing start to increase.
- The best success rate was achieved in 2019 and the score was 87%.



Result set 1

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

All Launch Site Names

Launch Site Names Begin with 'CCA'

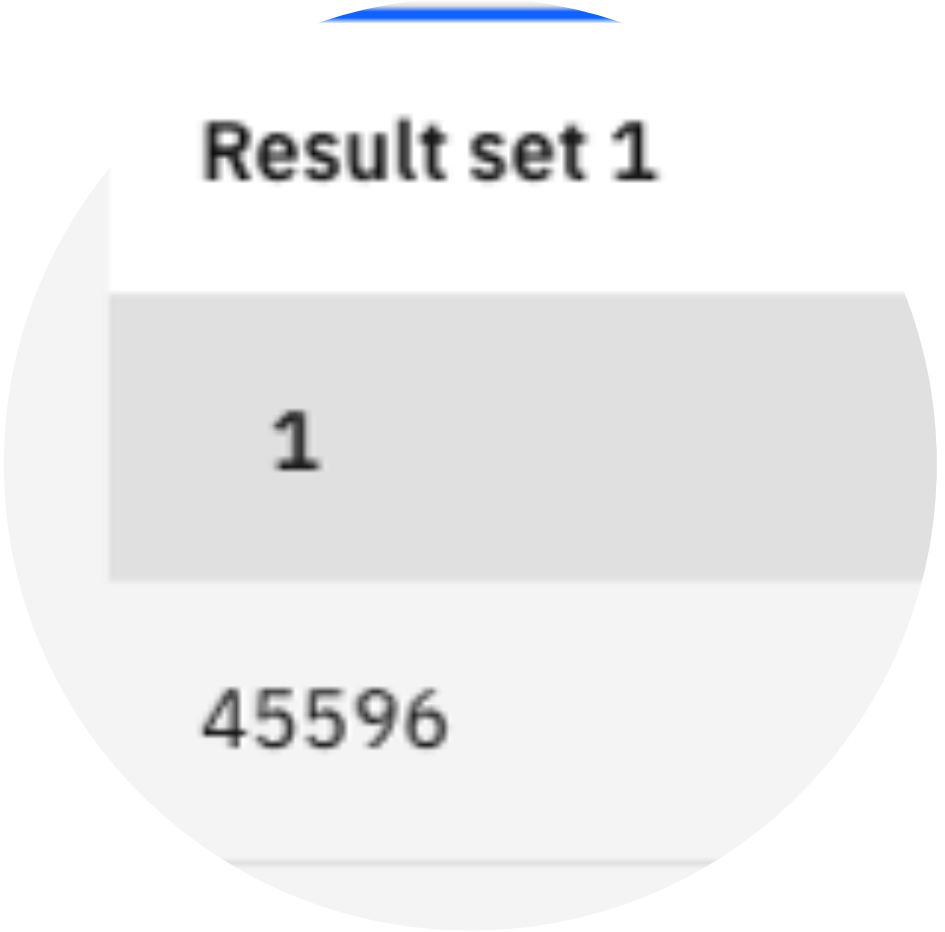
- The results are presented in the following graph.

Result set 1

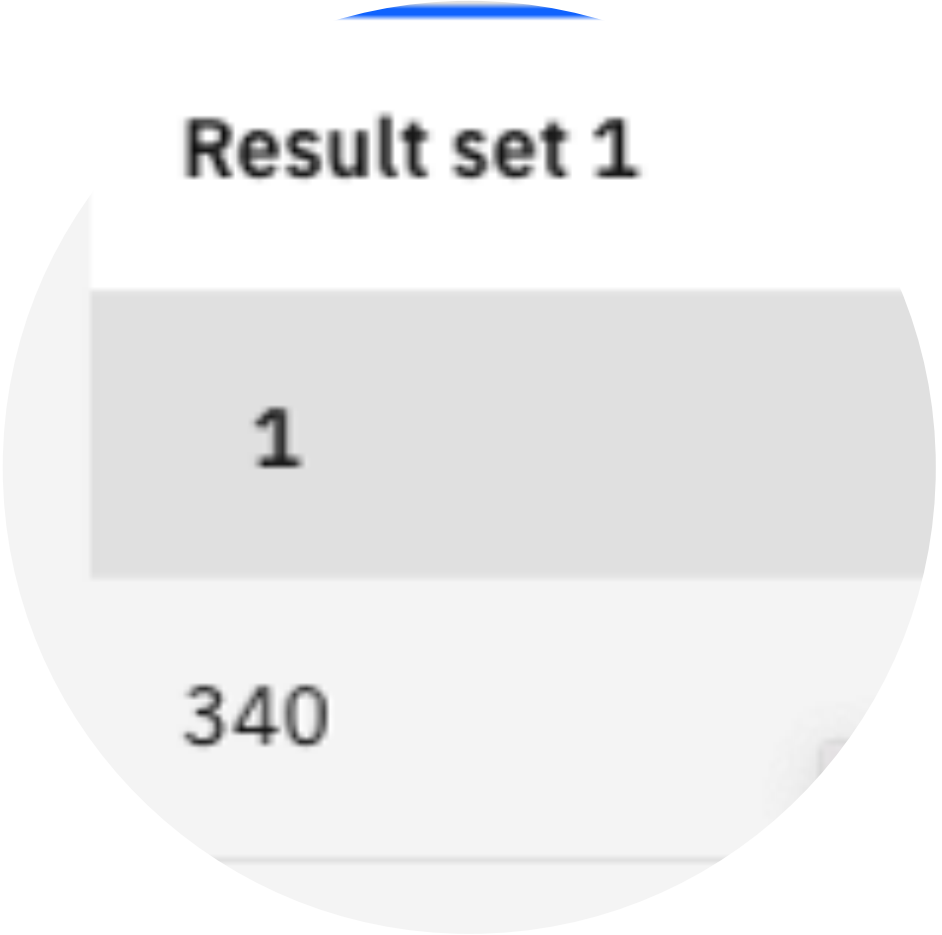
LAUNCH_SITE
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- It is estimated to be about 45,596



Average Payload Mass by F9 v1.1 was found to be 340



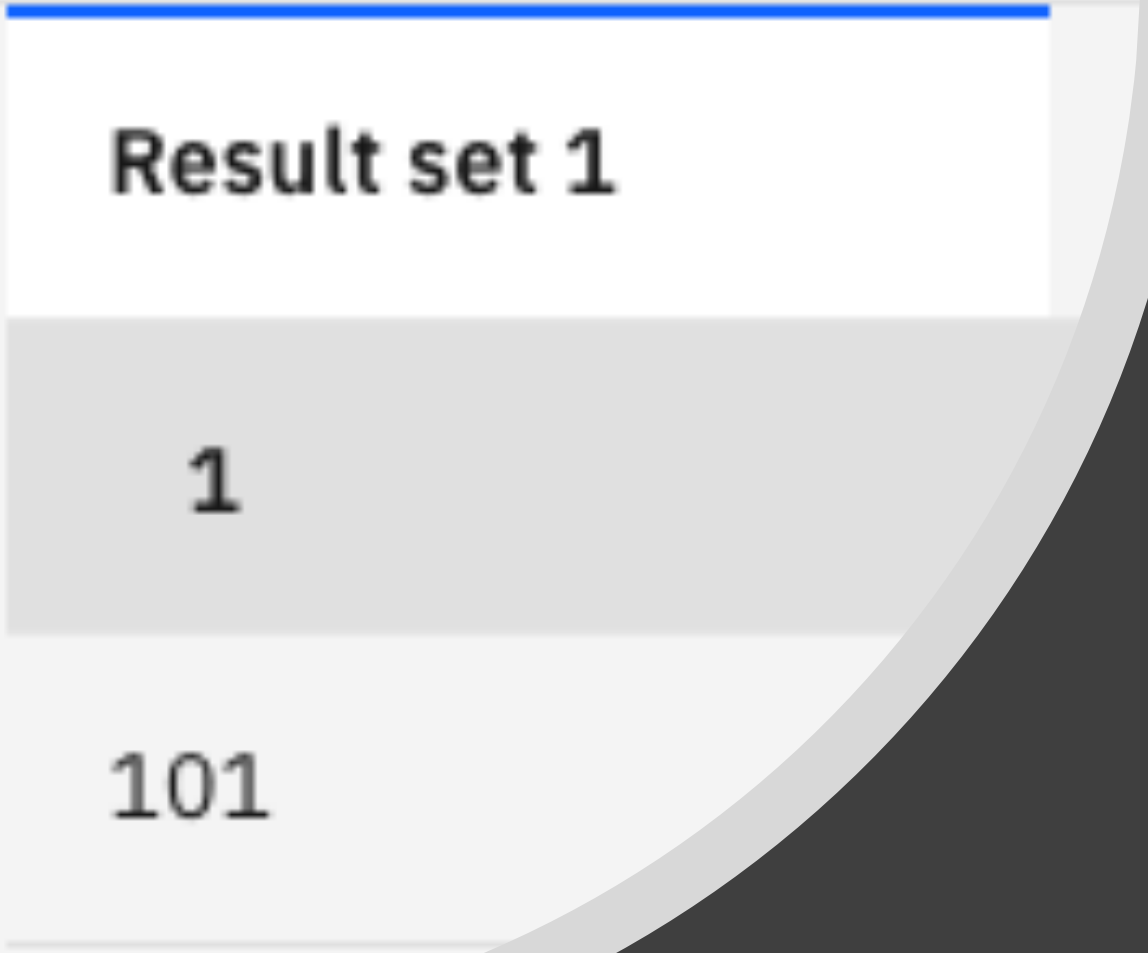
First Successful Ground Landing Date was in Dec 22, 2015

Min Date
2015-12-22

Result set 1

BOOSTER_VERSION
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2

Successful Drone Ship Landing with Payload between 4000 and 6000 are presented in the following graph.



Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed mission was 101 trials.

Boosters Carried Maximum Payload

- 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
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 many failed landing outcomes all occur at the launching site of CCAFS LC 40 & VAFB SLC-4E with various booster versions.

Result set 1

Find

LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
Failure (parachute)	F9 v1.0 B0003	CCAFS LC-40
Failure (parachute)	F9 v1.0 B0004	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E

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

- We can conclude that the landing outcomes of the rocket has more successful rate than failure.

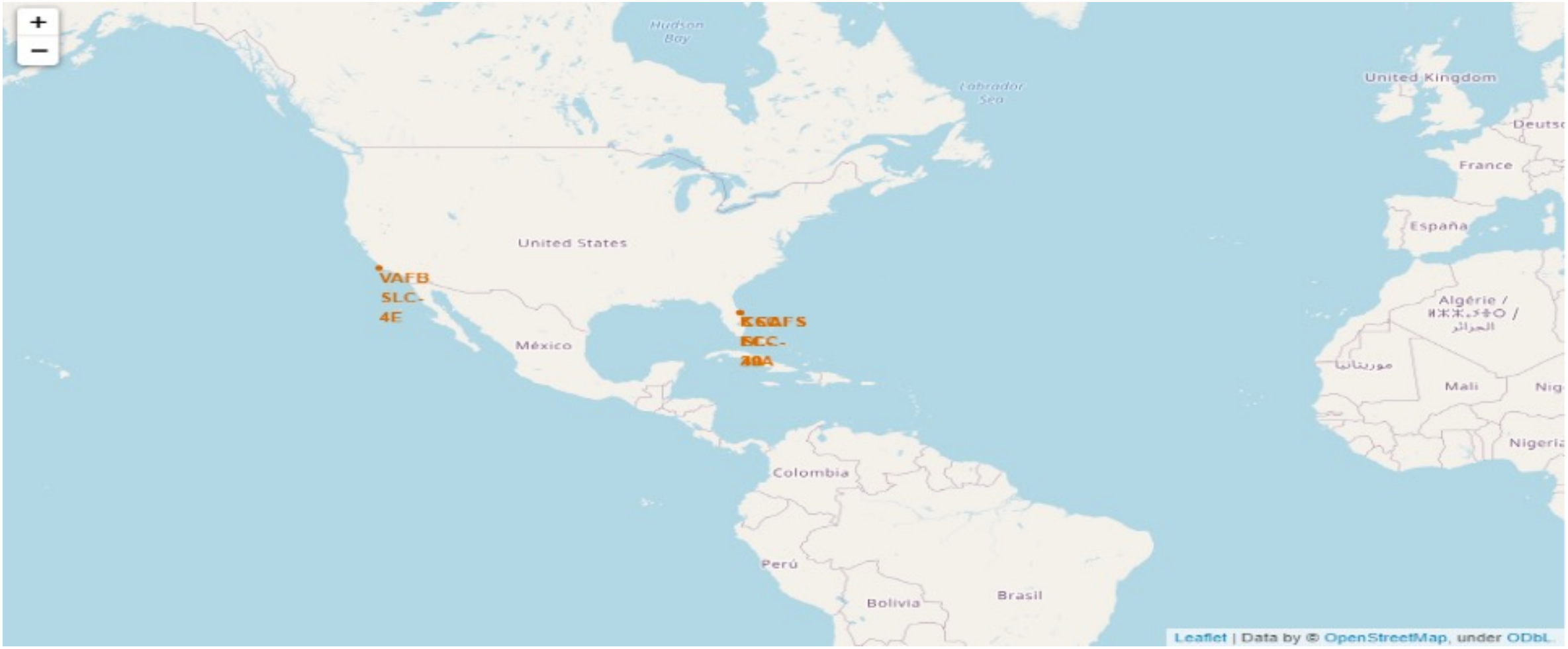
Result set 1

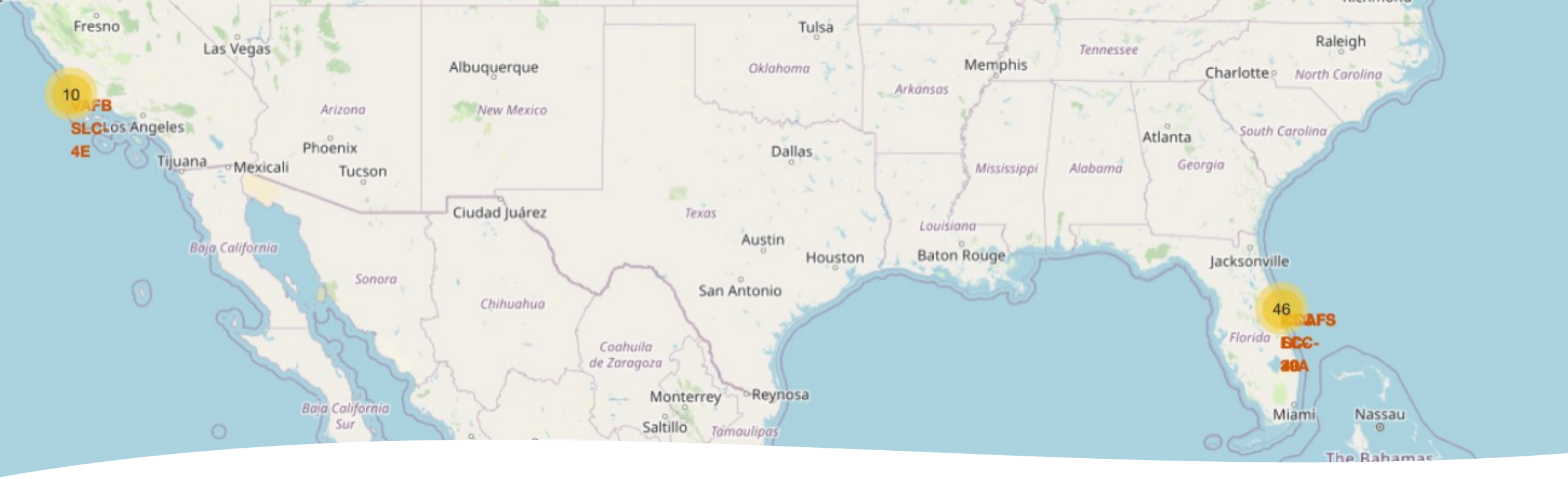
LANDING__OUTCOME	2
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5


Section 3
Launch Sites
Proximities Analysis

Nasa JSC Site

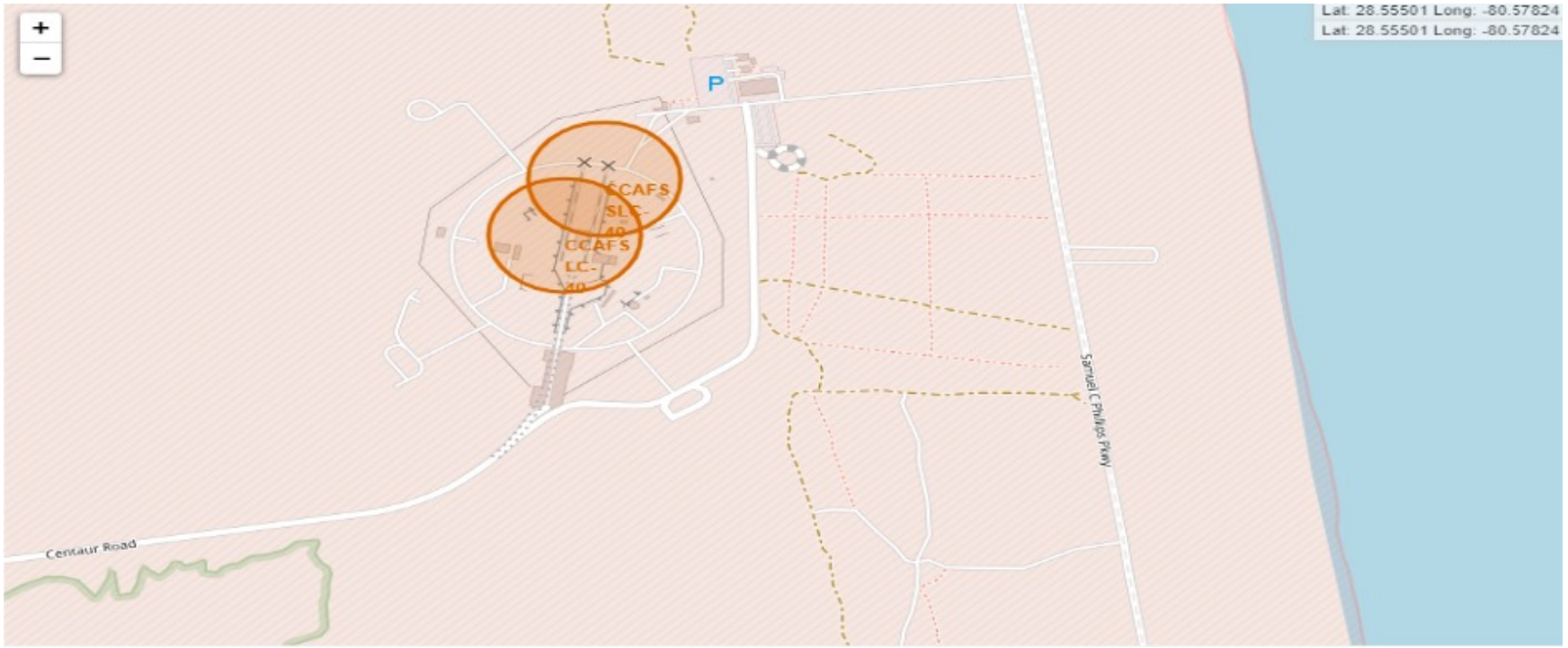
- This figure represents the location of Nasa JSC site





All launch sites on the map

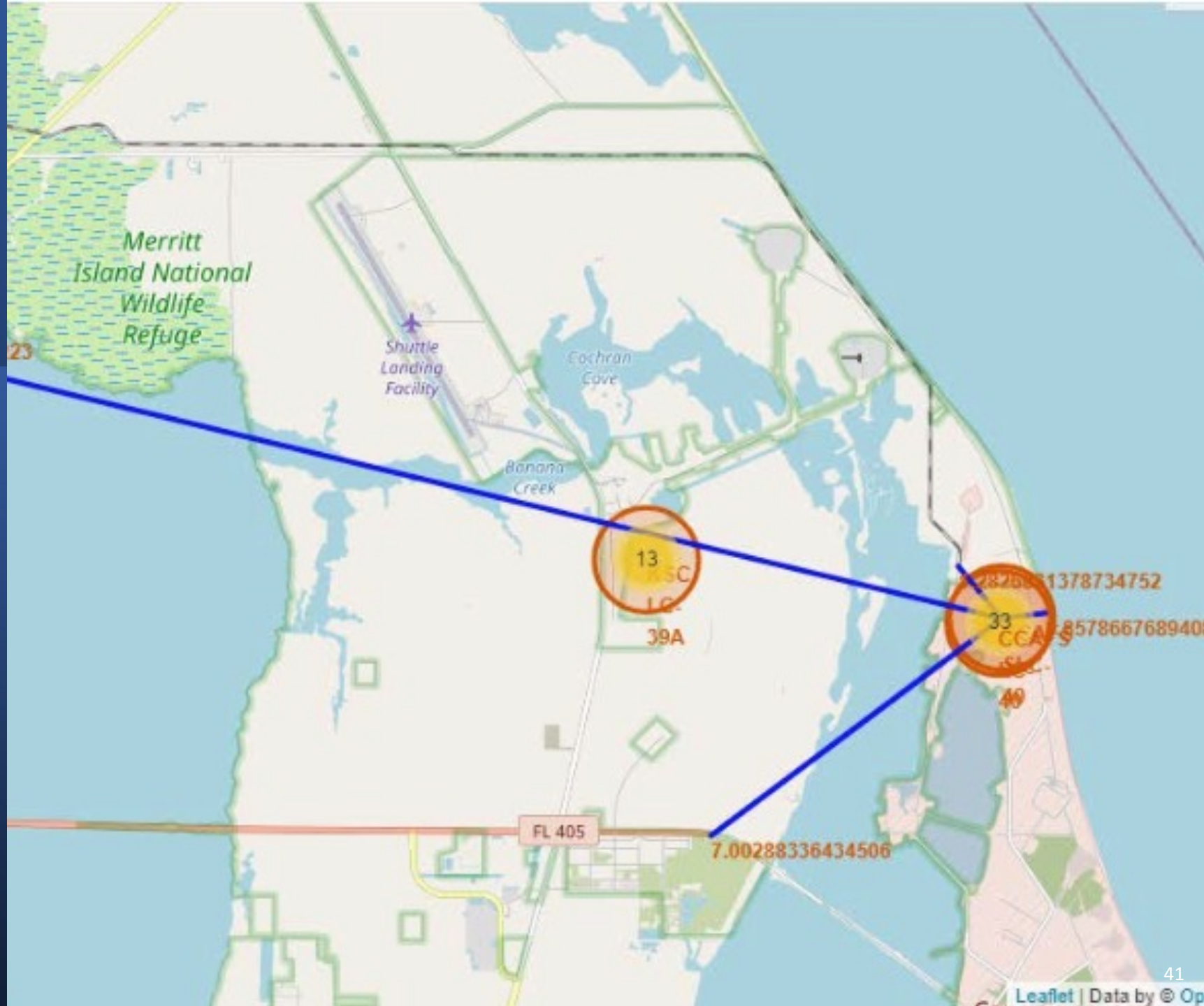
- This figure shows each rocket launch site.
- The figure shows that most of the launch locations are coastal.
- The launch only happens in one of the four launch sites, which means many launch records will have the exact same coordinate. Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.



Color-coded Launch Sites

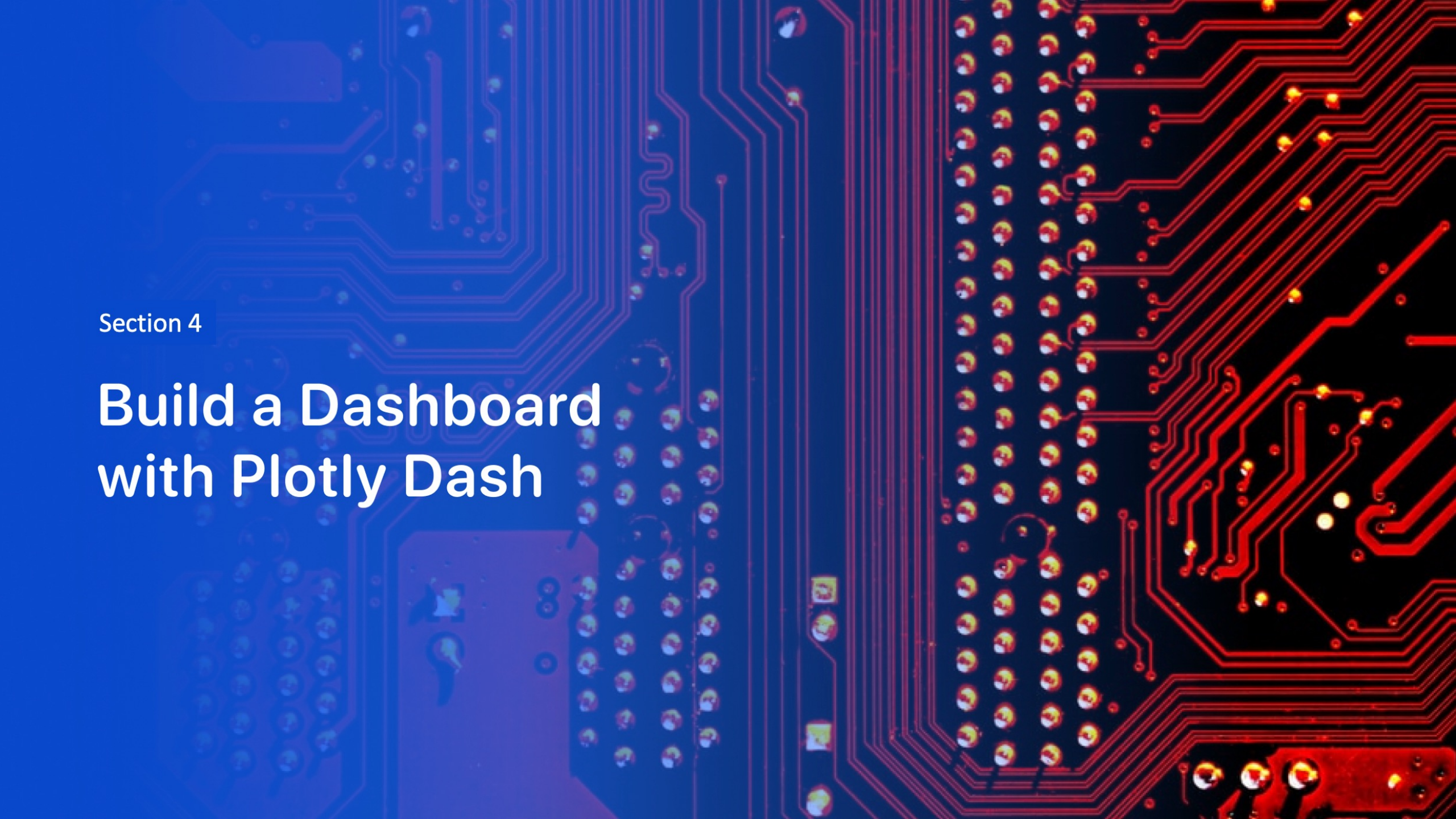
- In this figure, we can explore one of the launch sites and its proximity to any railway, highway, coastline, etc. This is an important issue since we need to make sure that these sites are far away from the residential areas in case of landing failure of the rocket will not cause any threats to human lives and properties.

The distances between CCAFS SLC-40 to its proximities



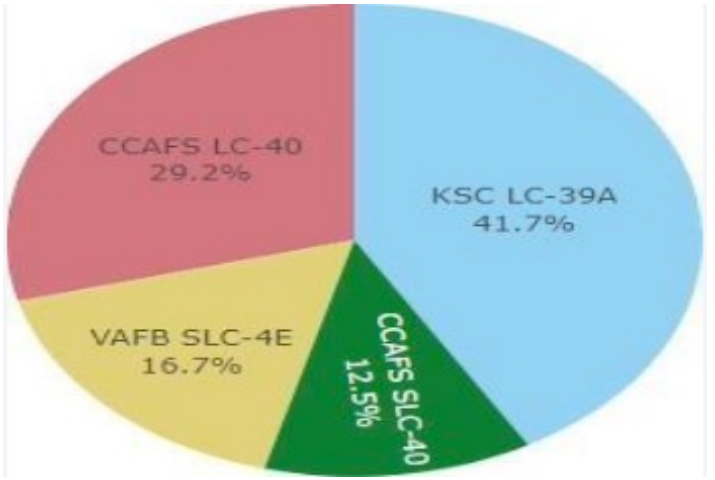
Section 4

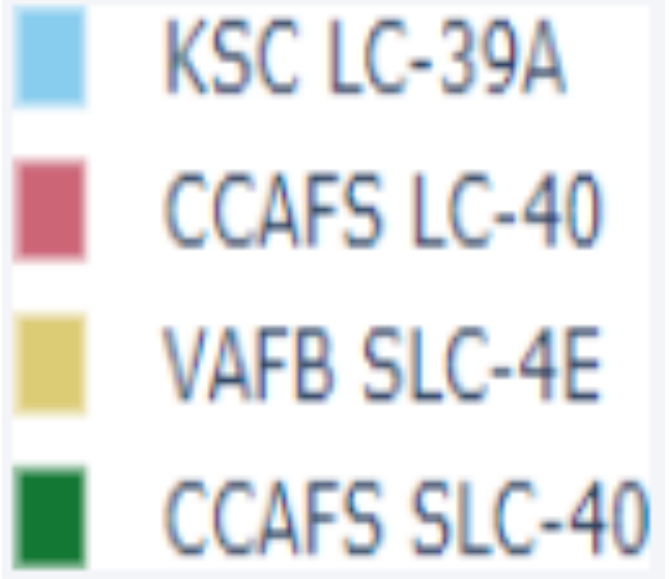
Build a Dashboard with Plotly Dash



Successful Launches

- The place from where launches are done seems to be a very important factor of success of missions.

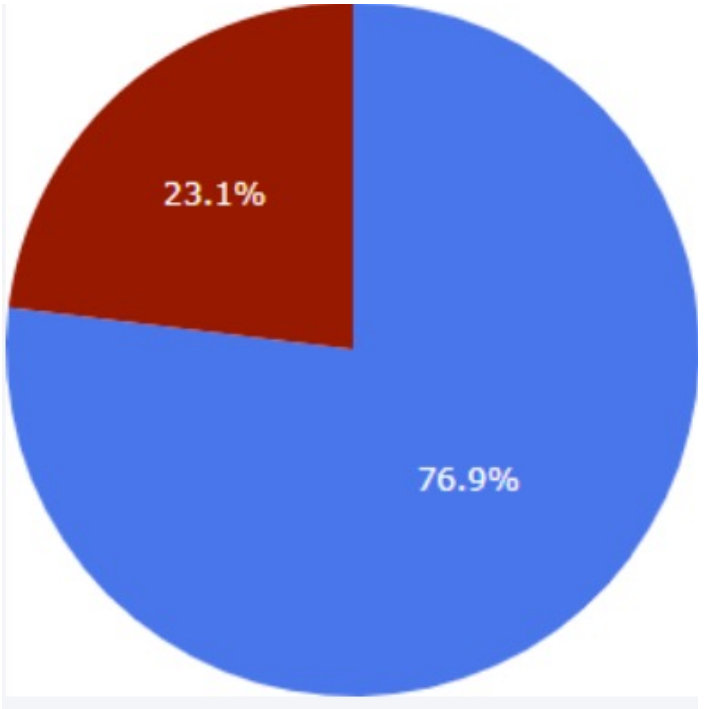




Successful Launches Ratio

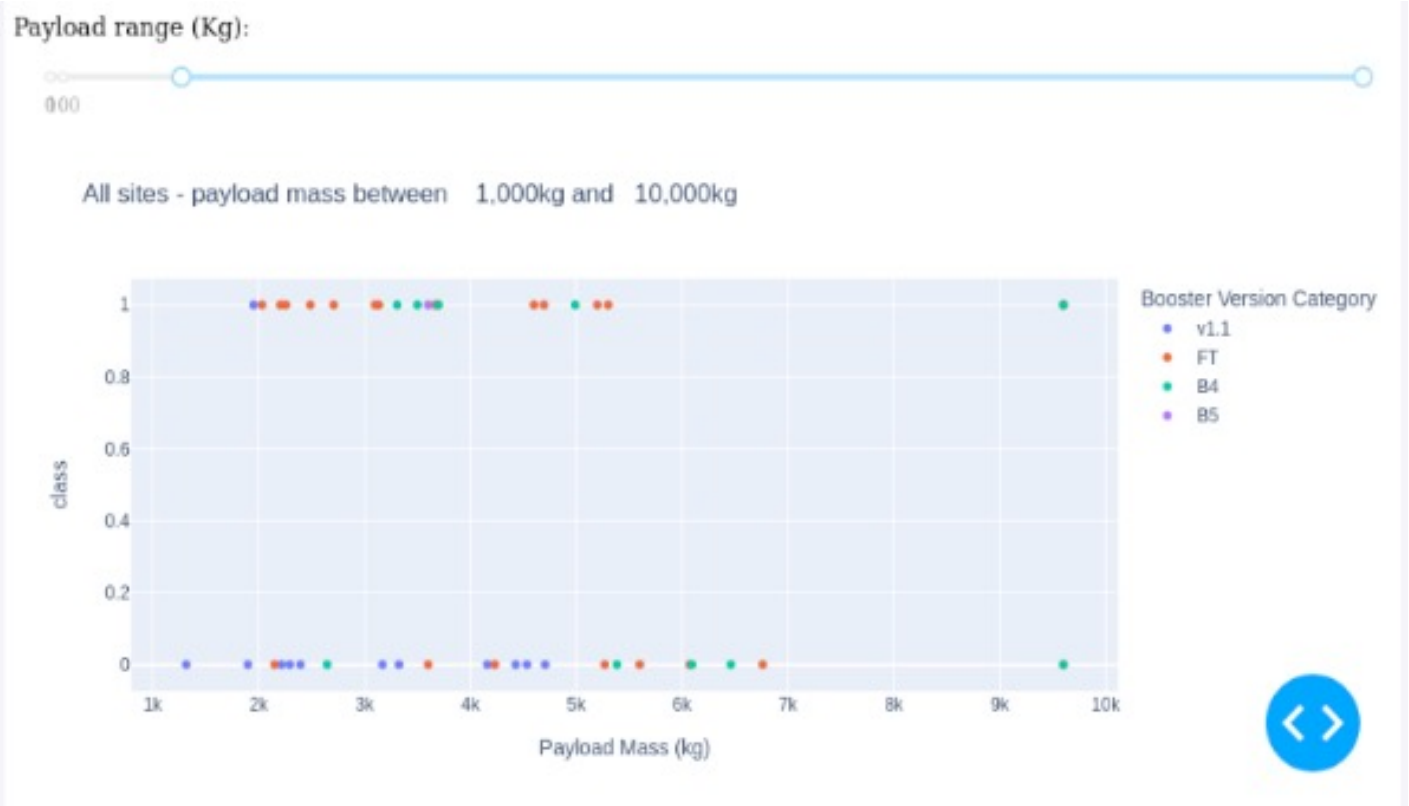
- 76.9% of launches are successful in this site.

KSC LC-39A Success Rate (blue=success)



Payload vs Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.

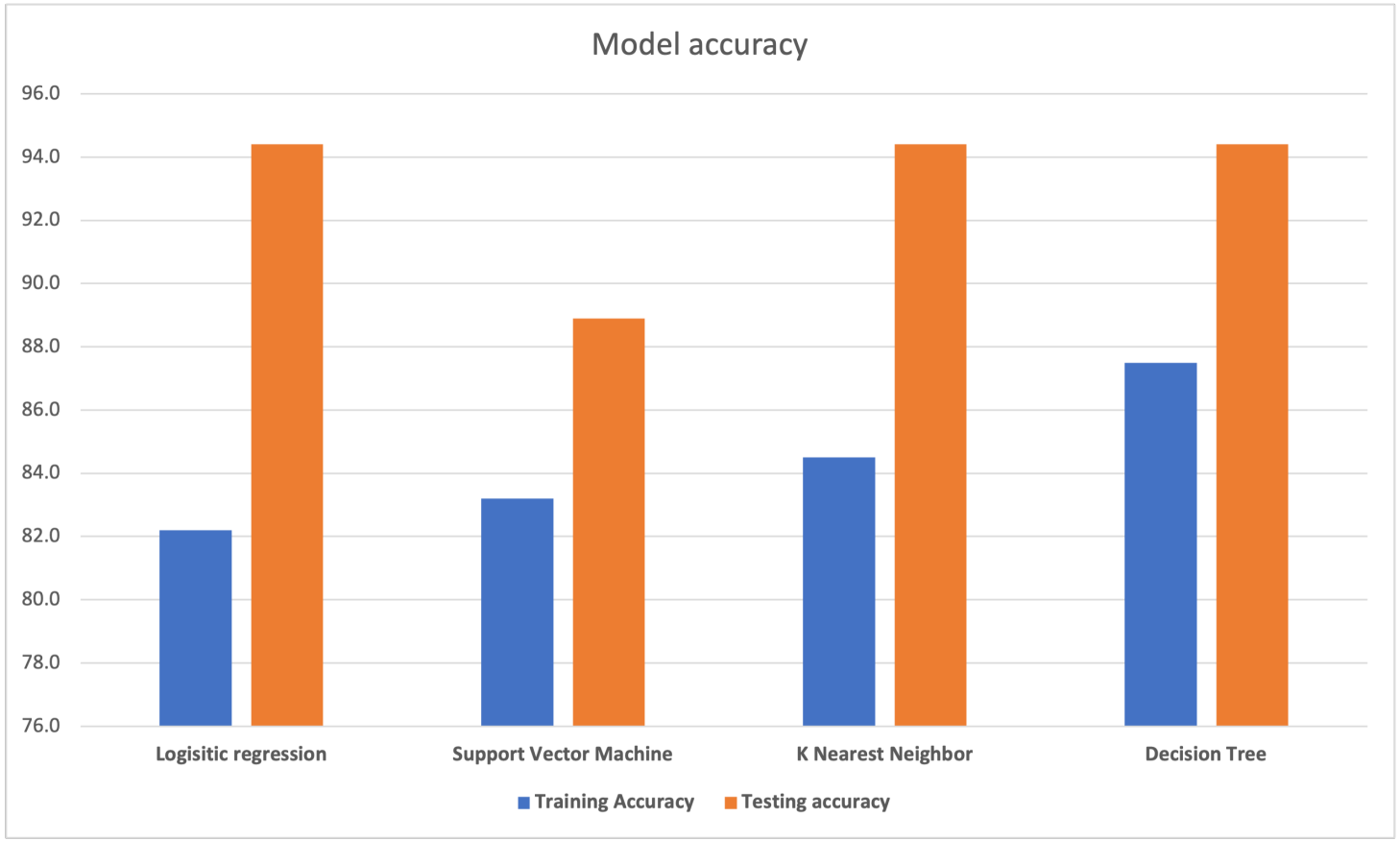


Section 5

Predictive Analysis (Classification)

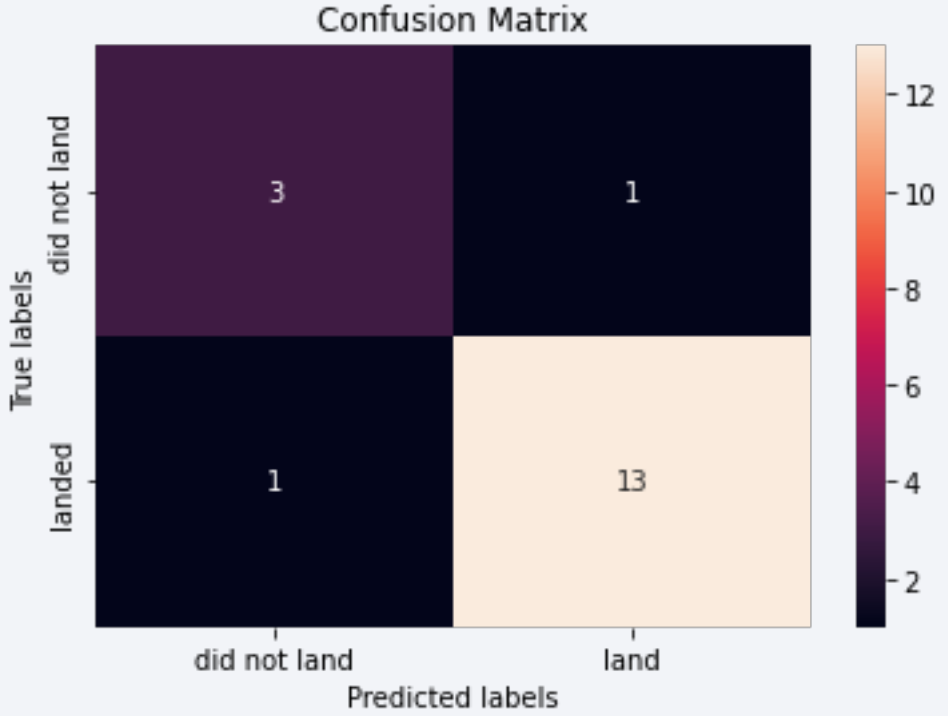
Classification Accuracy

- It can be seen here that Decision tree model has the best training accuracy among the trained models with best testing accuracy shared with LR & KNN.
- Support vector machine has the poorest performance in predicting the rocket landing.



Confusion Matrix

- Decision tree model was predicted 16 trials of landing successfully among 18 trails, including predicting 13 of successful landing and 3 did not landed successfully.
- DT model also mis-class two samples from the training set, one was landed successfully but predicted otherwise and one did not land successfully but predicted otherwise.



Conclusions

- It can be concluded from this project that:
  - Launches above 7,000kg are less risky.
  - The best launch site is KSC LC-39A.
  - Machine learning techniques are robust techniques that can be used to predict whether a rocket will land successfully or not.
  - Data visualization is an important technique to explore the hidden patterns from the data that can be used in decision making.
  - Web scraping can help extract information that is hard to be obtained as csv files or other structured data.
  - Model hyperparameter tuning is an essential task that can help reduce the error of prediction in models and find the best parameters that delivers the optimal performance.

References Links

- Hands-on Lab : String Patterns, Sorting and Grouping
- Hands-on Lab: Built-in functions
- Hands-on Lab : Sub-queries and Nested SELECT Statements
- Hands-on Tutorial: Accessing Databases with SQL magic
- Hands-on Lab: Analyzing a real World Data Set
- GitHub Repo URL

Thank you!