IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Narendra Nath Rit
22 September, 2022

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection through API(mainly SpaceX API)
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis(EDA) with SQL
  - Exploratory Data Analysis(EDA) with Data Visualization
  - Interactive Visual Analytics with Folium
  - Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results:
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context:

  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, much of the 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 if an alternate company (depicted here as SpaceY) wants to bid against SpaceX for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers:
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features (like the place for landing etc.) that determine the success rate of a successful landing.
  - What operating conditions (like the payload etc.) needs to be in place to ensure a successful landing program?
  - The total cost estimation for launches by predicting the successful landings of the first stage of the rockets.

Section 1

Methodology

Methodology

Executive Summary:

- Data collection methodology:
    - Data was collected using two sources:
        - SpaceX API (https://api.spacexdata.com/v4/rockets/)
        - Web Scraping from Wikipedia (https://en.Wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling:
    - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features.
    - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL.

Methodology

Executive Summary:

- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models.
  - Building the classification models using the data that was collected after normalizing and dividing the data into training and testing data sets.
  - Tuning the classification models using the train and test data sets and further doing the fine tuning by adjusting the data set after evaluation process.
  - Evaluating the classification models for better accuracy and fine tuning then using those models for prediction.

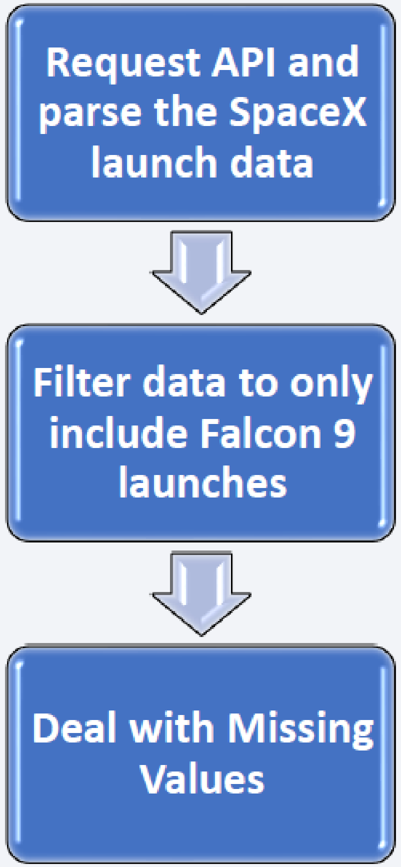
Data Collection

- The data was collected using various methods:
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with Python module, BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- We used the get request to the SpaceX API (publicly available) to collect and clean the requested data and did some basic data wrangling and formatting as mentioned in the flowchart beside.

- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/1.%20Data%20Collection%20API.ipynb



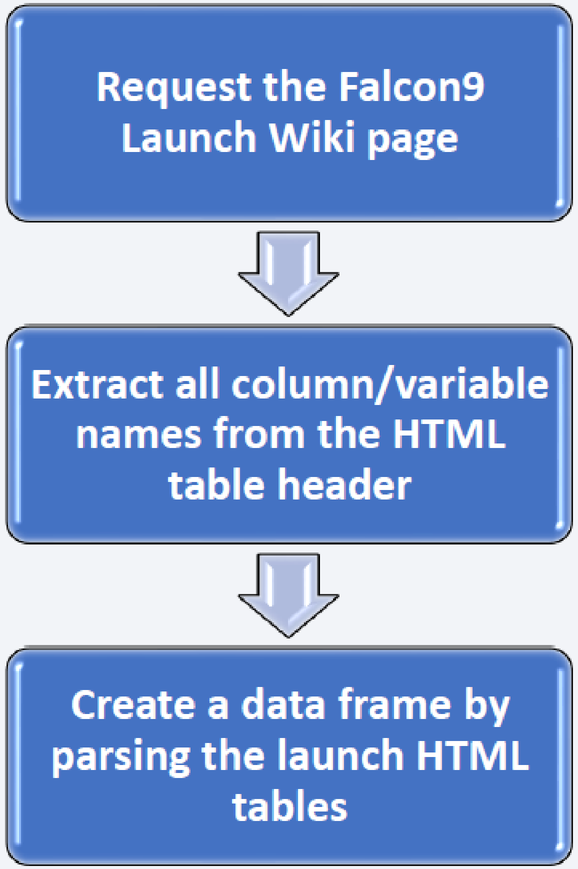
Request API and parse the SpaceX launch data

Filter data to only include Falcon 9 launches

Deal with Missing Values

Data Collection - Scraping

- We applied web scrapping to web scrap the Falcon 9 launch records with BeautifulSoup. Then we parsed the table and converted it into a pandas dataframe as per the flowchart beside.
- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/2.%20Data%20collection%20with%20Web%20Scrapping.ipynb



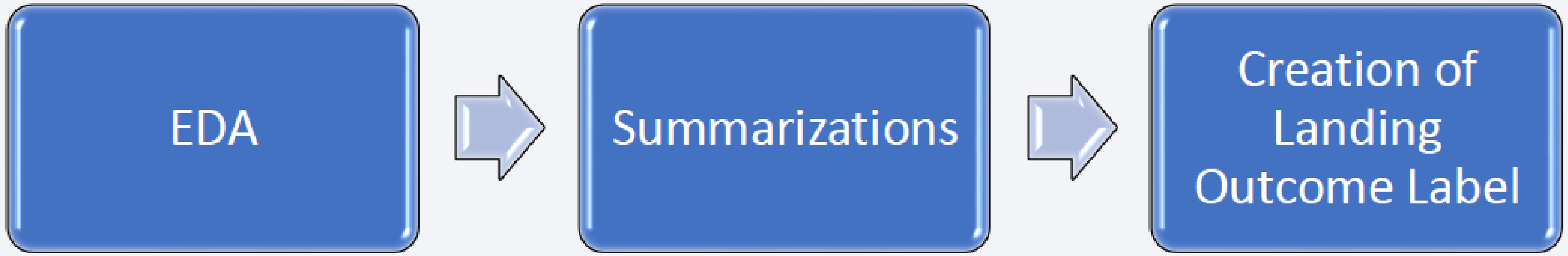
Request the Falcon9 Launch Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits.
- We created landing outcome label from outcome column as per the flowchart below and exported the results to csv.
- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/3.%20Data%20Wrangling.ipynb



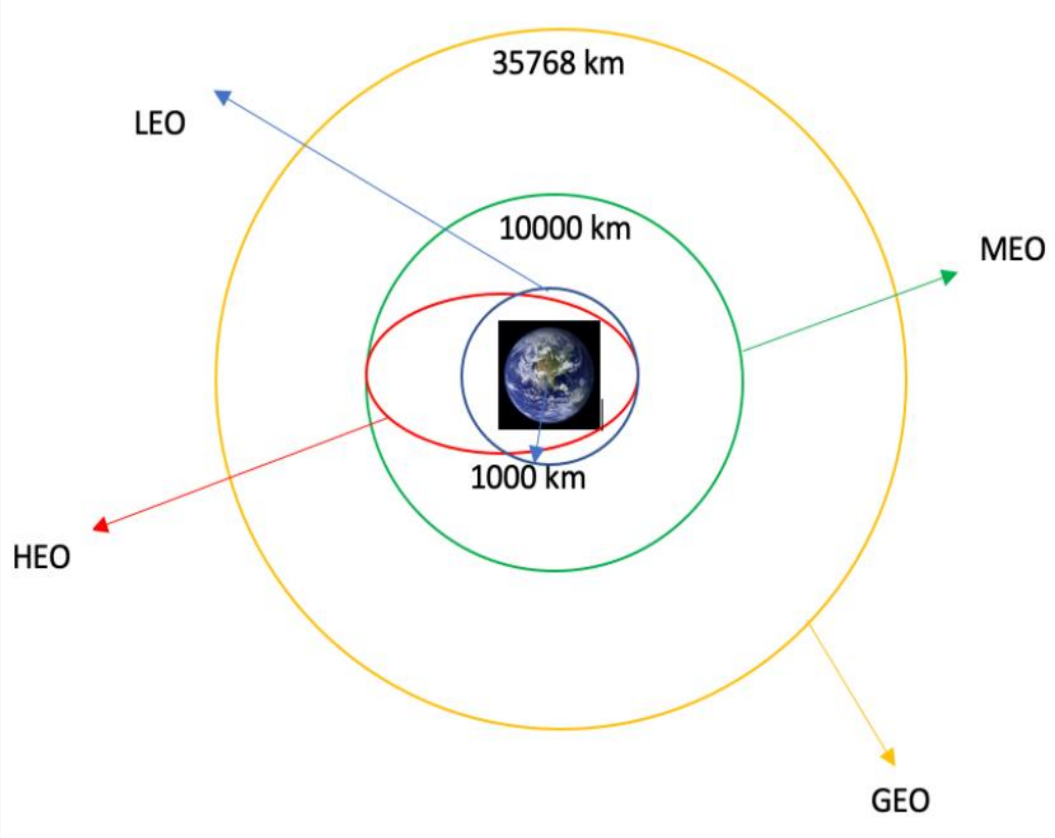
EDA

Summarizations

Creation of Landing Outcome Label

Data Wrangling

- **LEO**: It is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less.
- **HEO:** It is a highly elliptical orbit with high eccentricity, usually referring to one around Earth above the altitude of geosynchronous orbit (35,786 km or 22,236 mi).
- **MEO:** It is a geocentric orbit ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi).
- **GEO:** It is a circular geosynchronous orbit 35,786 kilometers (22,236 miles) above Earth's equator and following the direction of Earth's rotation.



EDA with SQL

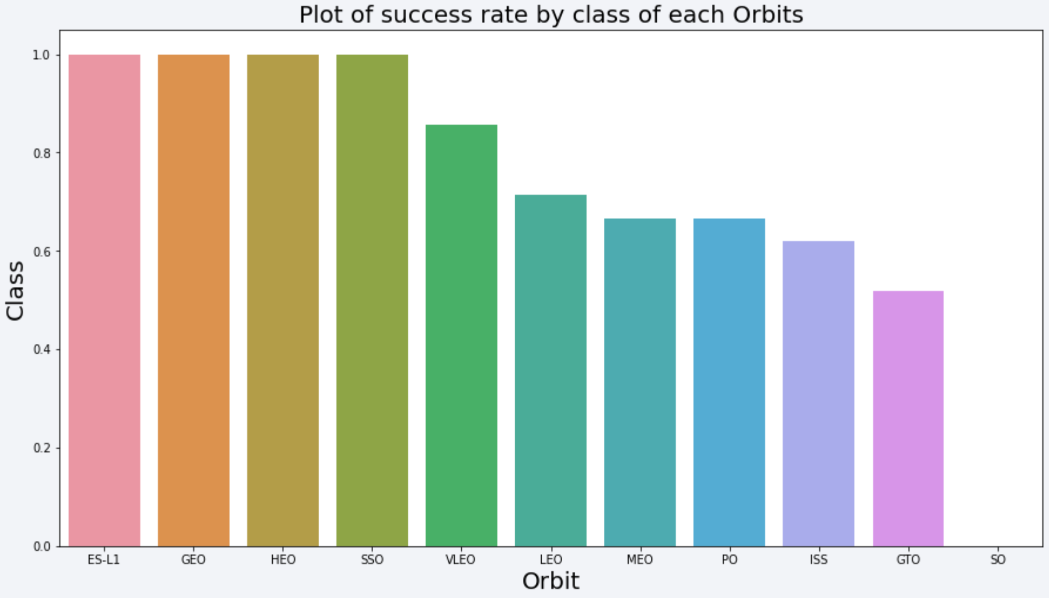
- We loaded the SpaceX dataset using SQLite3 module without leaving the jupyter notebook.

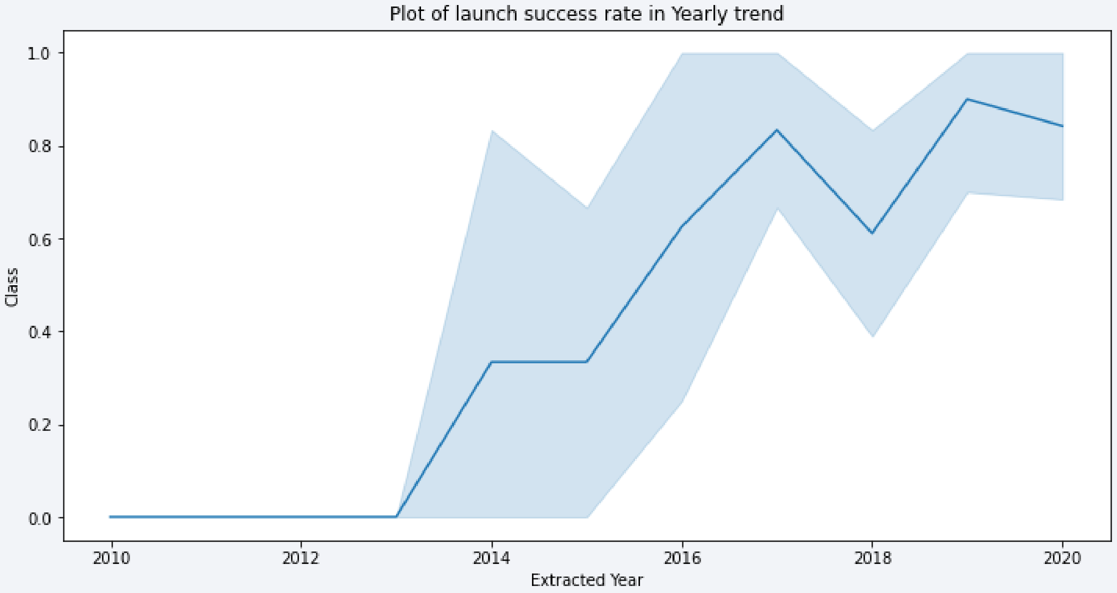
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - The names of unique launch sites in the space mission,
  - Top 5 launch sites whose name begin with the string 'CCA',
  - Total payload mass carried by boosters launched by NASA (CRS),
  - Average payload mass carried by booster version F9 v1.1,
  - The date when the first successful landing outcome in ground pad was achieved,
  - Names of the boosters which have success in drone ship having payload mass between 4000 to 6000 kg
  - The 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 sites names for in year 2015,
  - The rank of the count of landing outcomes (such as failure(drone ship) or success (ground pad)) between the date 2010-06-04 to 2017-03-20.

- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/4.%20EDA%20with%20SQL.ipynb

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.





- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/5.%20EDA%20with%20Data%20Visualization.ipynb

Build an Interactive Map with Folium

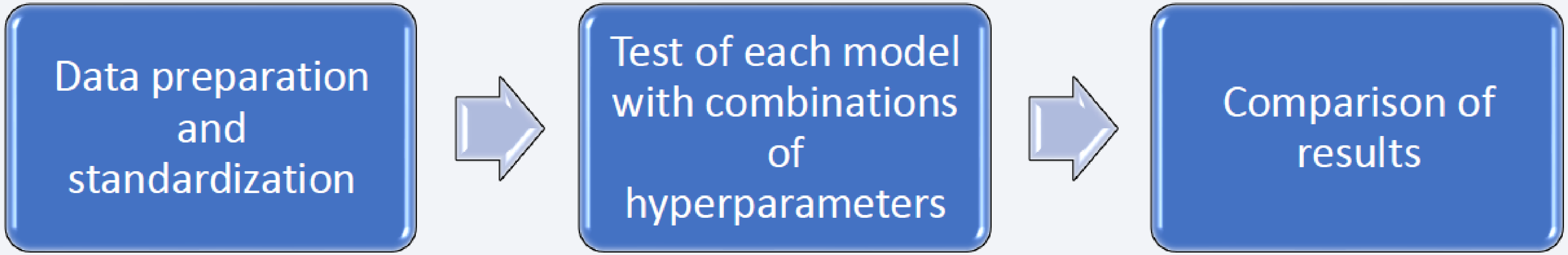
- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the Folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
  - Are launch sites near railways, highways and coastlines?
  - Do launch sites keep certain distance away from cities?
- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/6.%20Interactive%20Visual%20Analytics%20with%20Folium%20Lab.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash.
- We plotted pie charts showing the total launches by a certain sites.
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The link to the notebook is:
  https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/7.%20SpaceX%20Launch%20Records%20Dashboard%20App.py

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing. Then we built different machine learning models and tune different hyperparameters using GridSearchCV. We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- After comparing four classification models (Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest neighbors) we found the best performing classification model.
- The link to the notebook is: https://github.com/nnrit7/Applied_Data_Science_Capstone/blob/main/8.%20Machine%20Learning%20Prediction.ipynb


Data preparation and standardization

Test of each model with combinations of hyperparameters

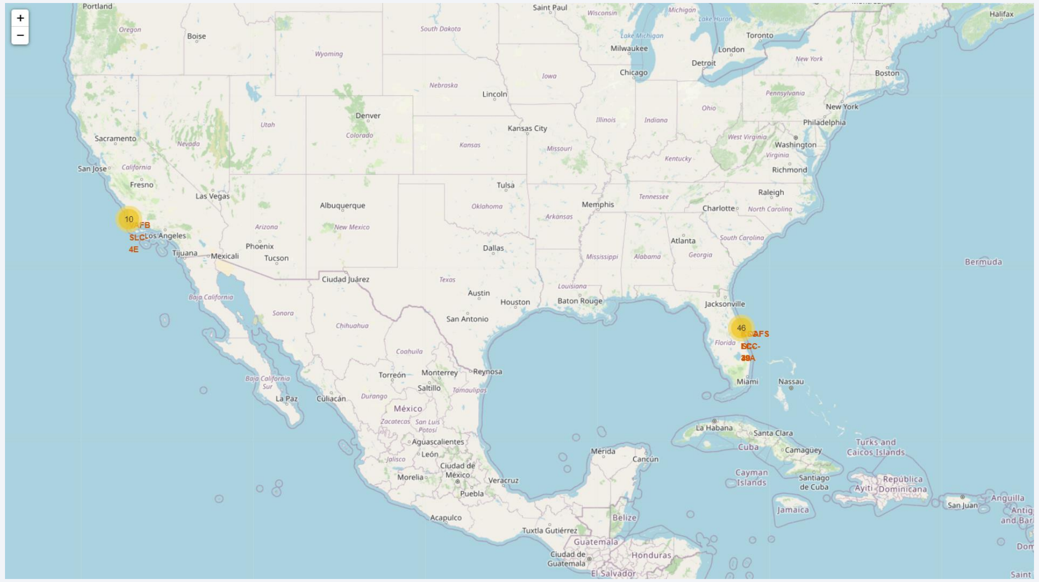
Comparison of results

Results

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites;
  - The first launches were done to SpaceX 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.

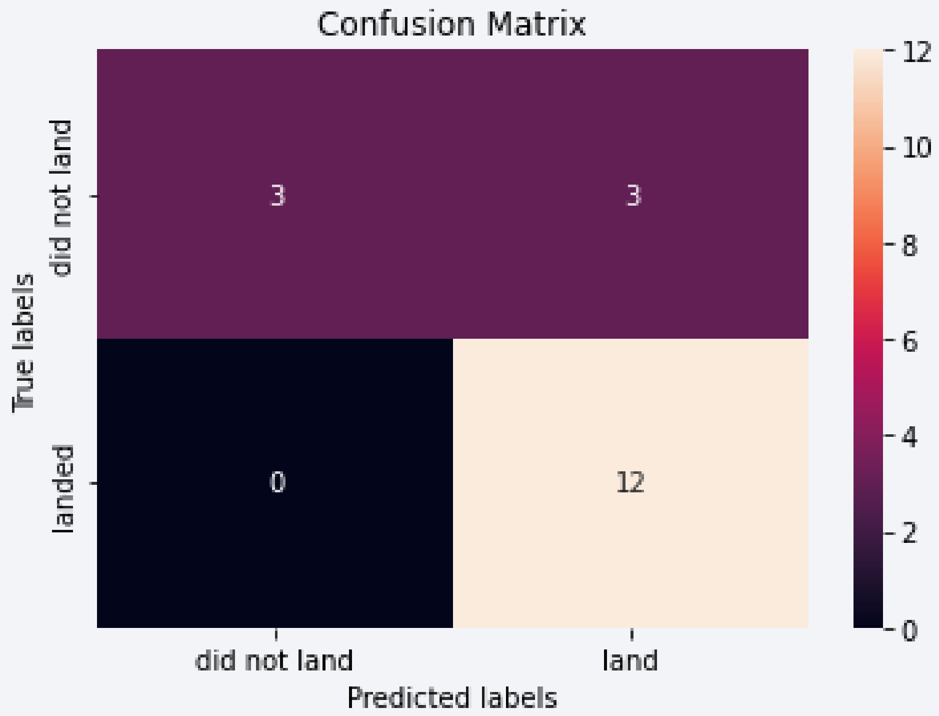
Results

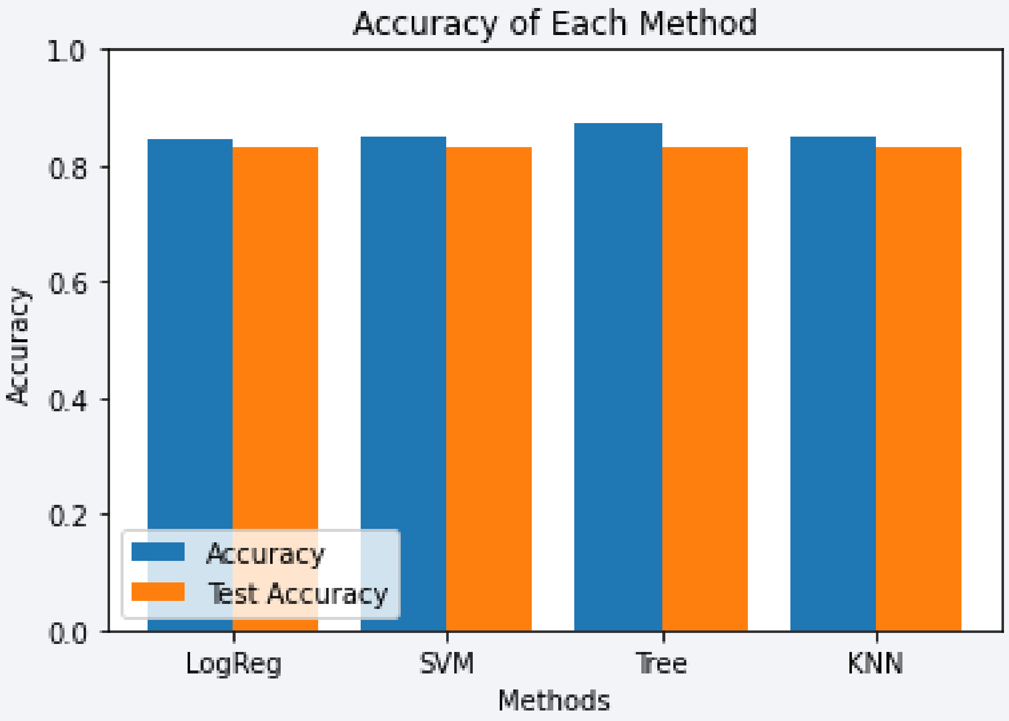
- Interactive data analysis results:
  - Launch sites are far away from localities but highway and railway is present for communication.
  - Launch sites are use to be in safety places like near sea and have a good logistic infrastructure around.
  - Most launches happens at east coast launch sites.

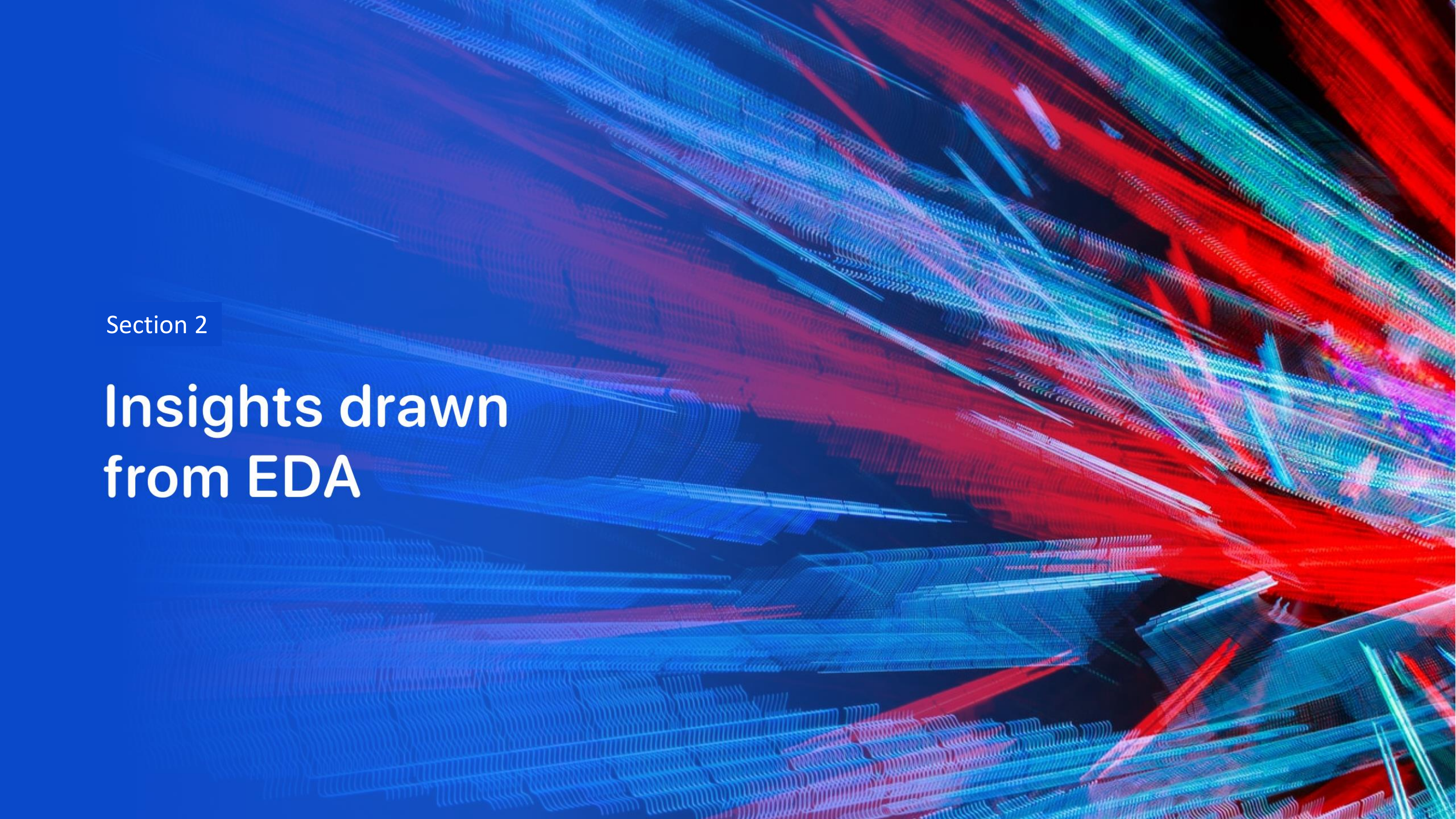


Results

- Predictive data analysis results:
    - Amongst all models that we tested Decision Tree Classifier is came out as best model to predict successful landings having accuracy over 87% and accuracy for test over 83%.





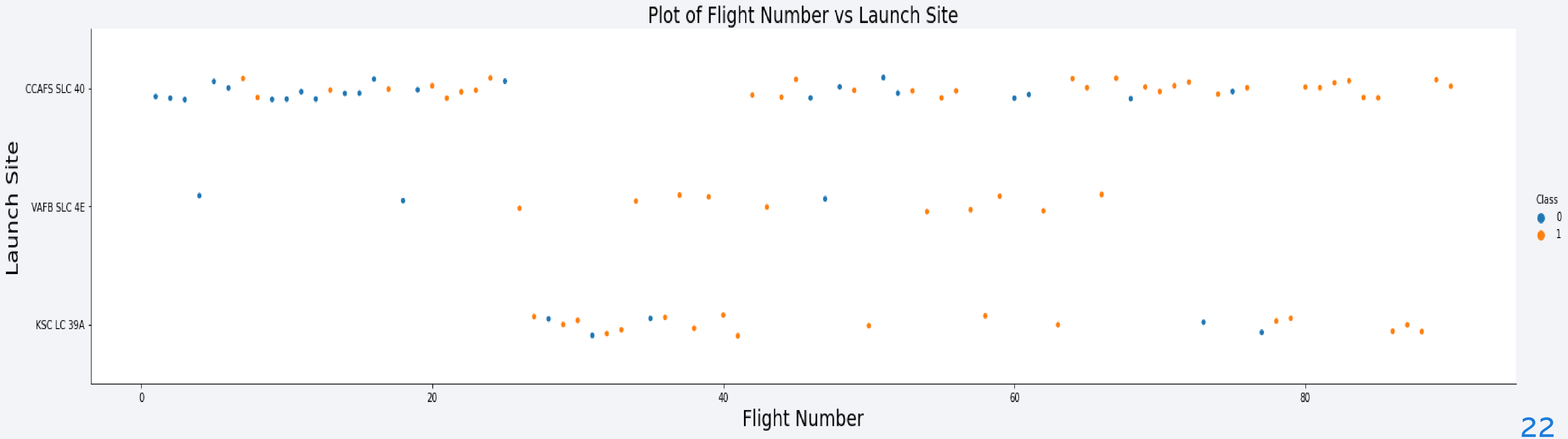


Section 2

Insights drawn from EDA

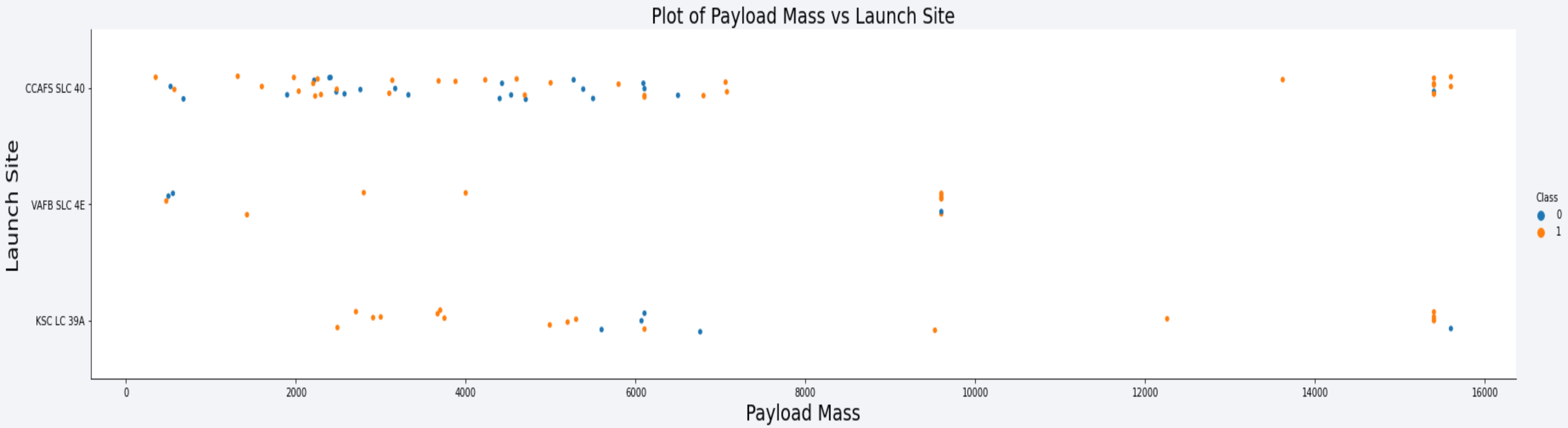
Flight Number vs. Launch Site

- According to the plot below, it’s possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
- In second place VAFB SLC 4E and third place KSC LC 39A.
- It’s also possible to see that the general success rate improved over time.



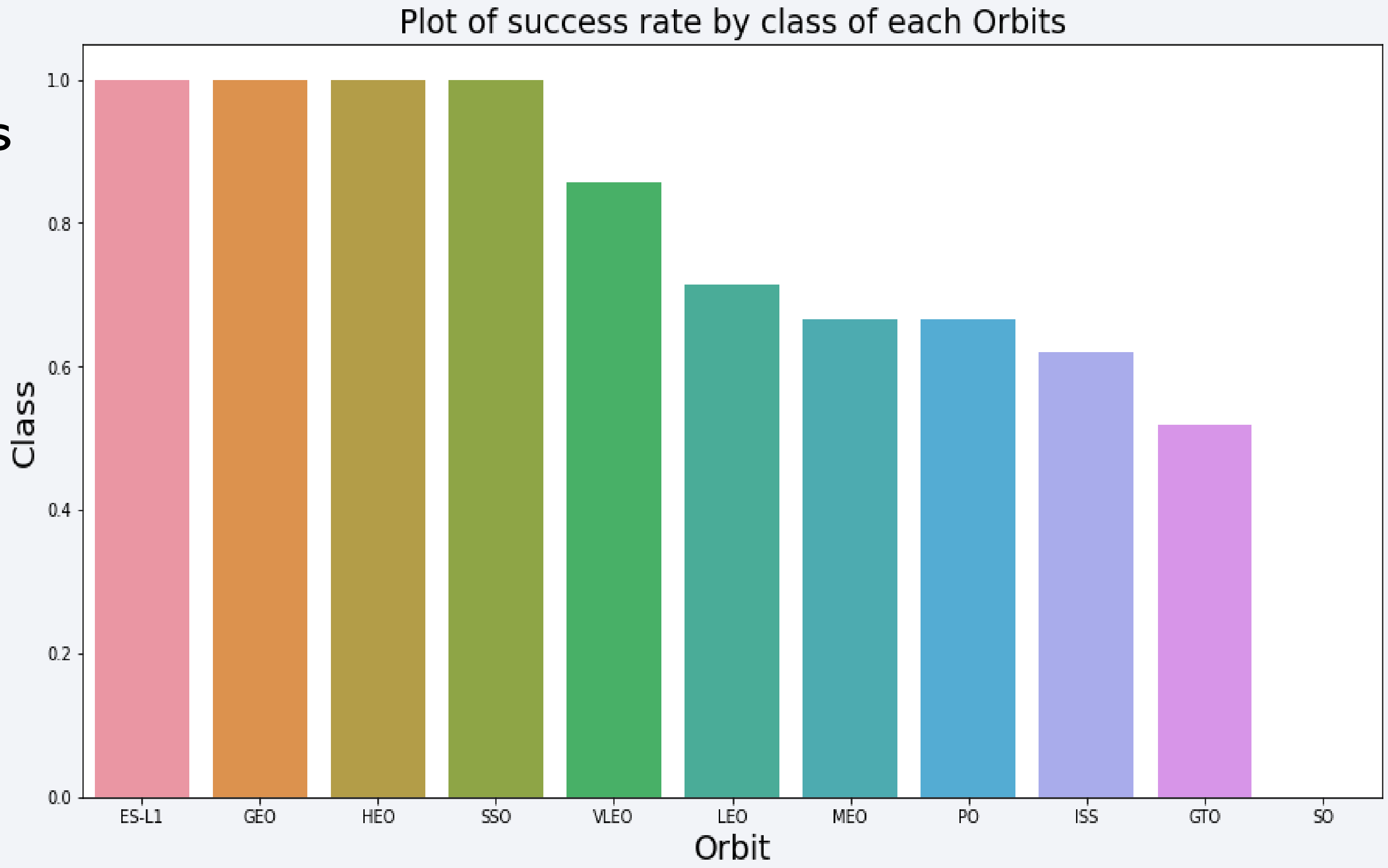
Payload vs. Launch Site

- According to the plot below, payloads over 9,000kg (about the weight of a school bus) have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



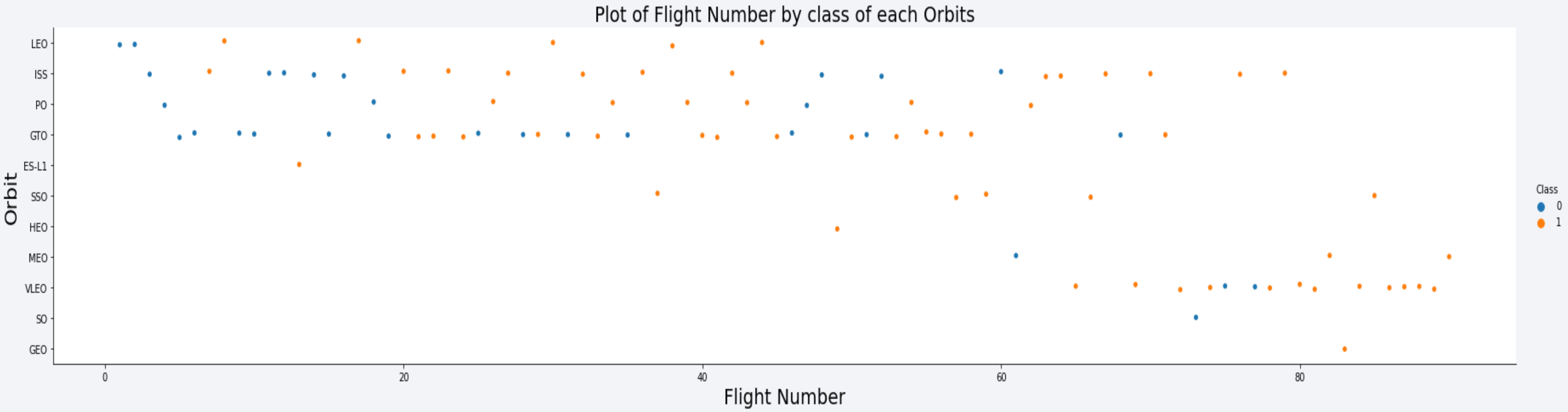
Success Rate vs. Orbit Type

- According to the plot beside, the biggest success rates happens to orbits as:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by orbits as:
  - VLEO (above 80%)
  - LFO (above 70%)



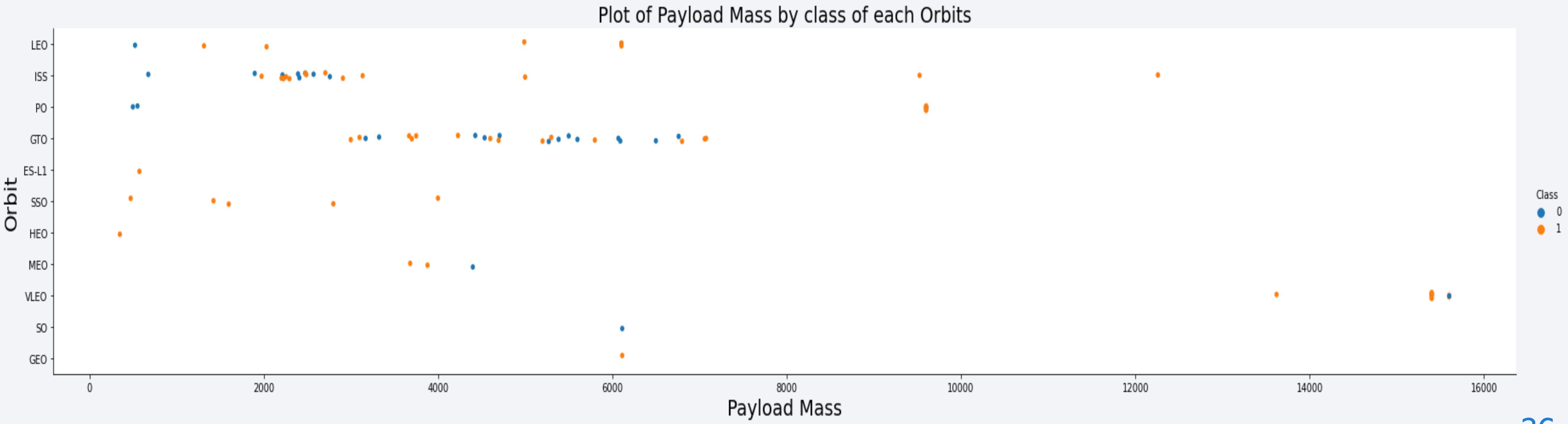
Flight Number vs. Orbit Type

- Apparently from to the plot below, success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.



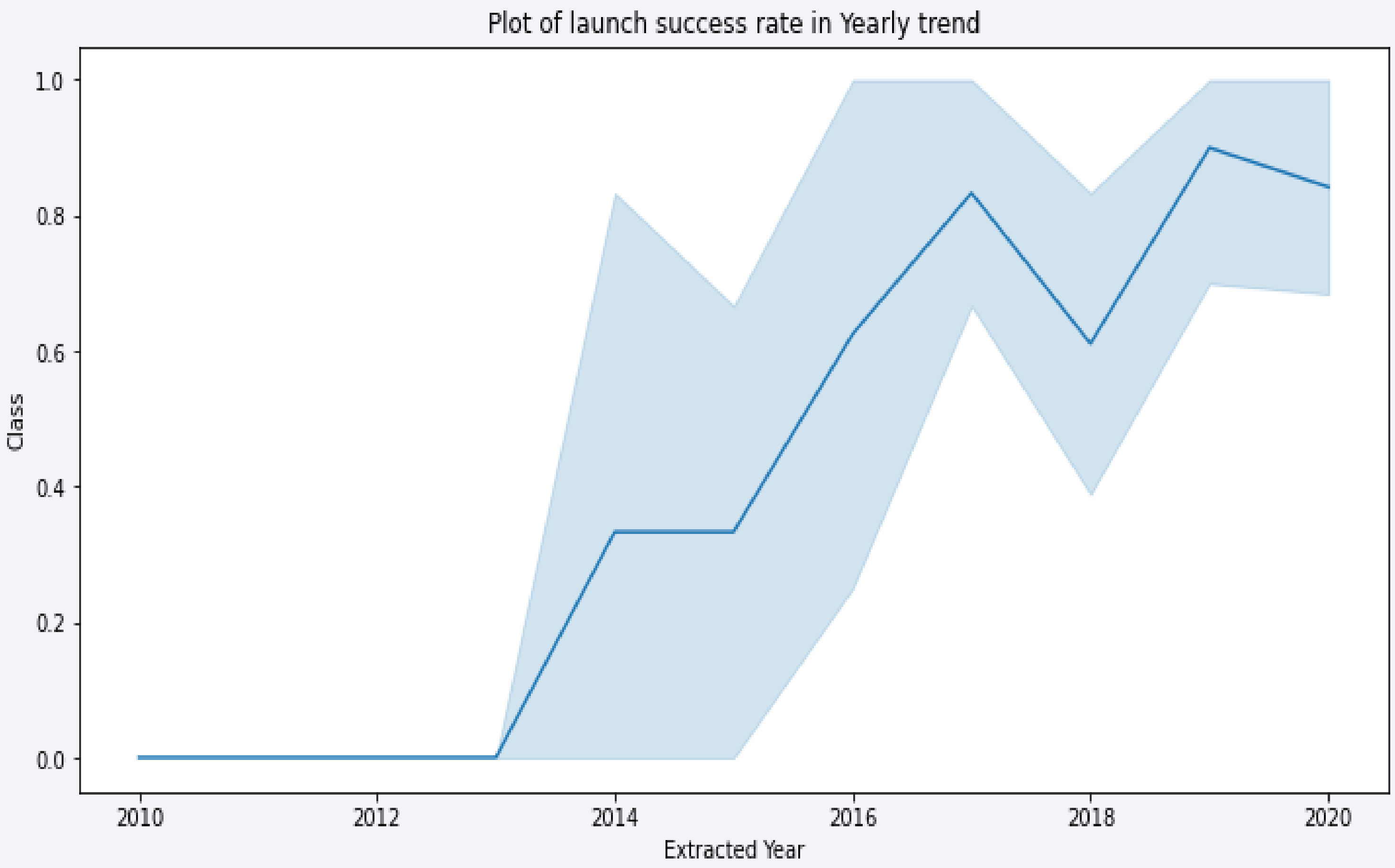
Payload vs. Orbit Type

- Apparently from to the plot below, there is 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.



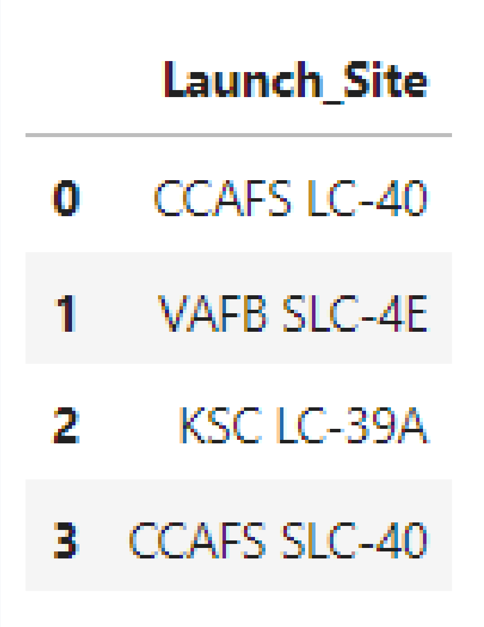
Launch Success Yearly Trend

- According to the plot beside, success rate started increasing from 2013 and kept until 2020.
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:

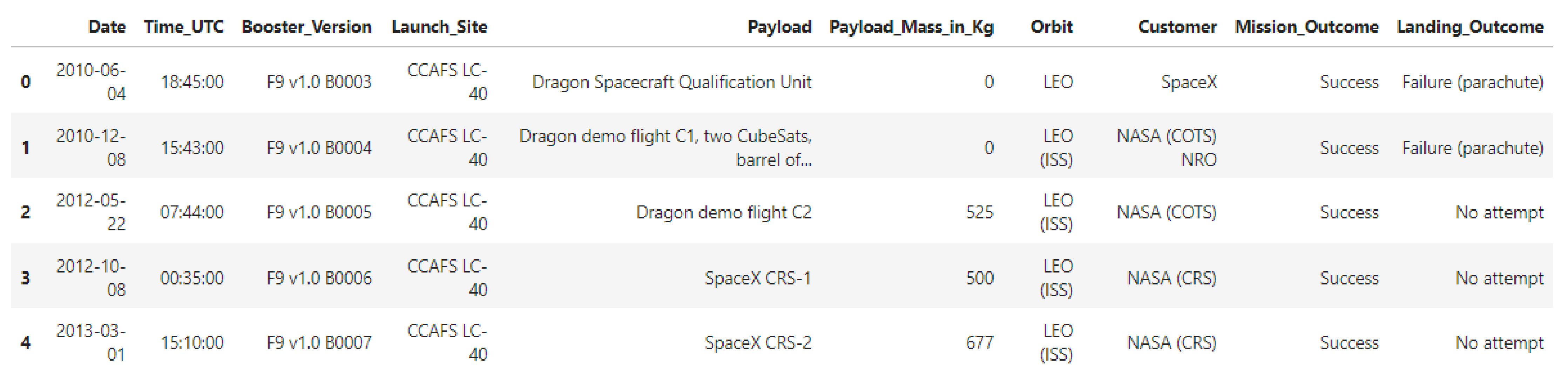


	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

- They are obtained by selecting unique occurrences of “Launch_Site” values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA’:

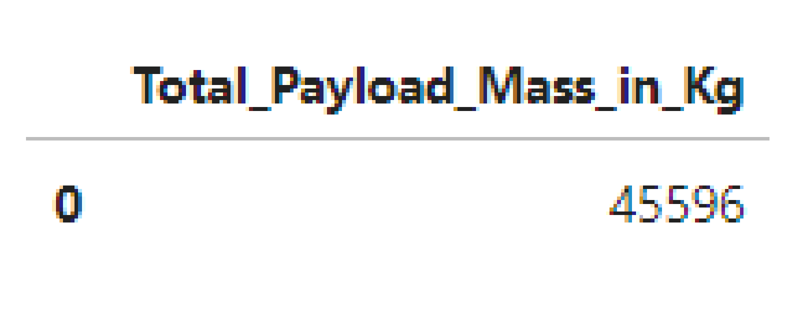


	Date	Time_UTC	Booster_Version	Launch_Site	Payload	Payload_Mass_in_Kg	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of…	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

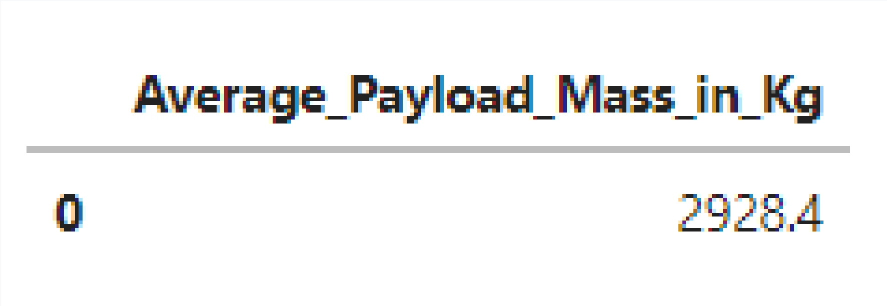


	Total_Payload_Mass_in_Kg
0	45596

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

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

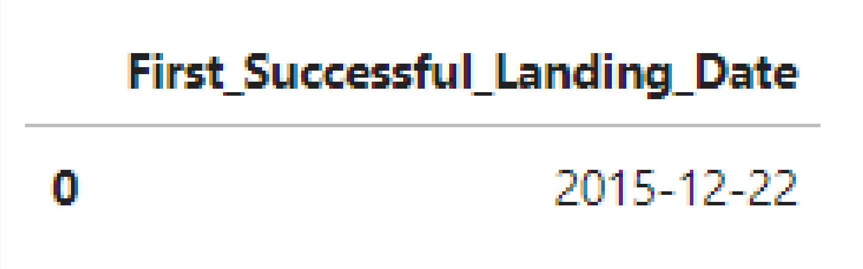


	Average_Payload_Mass_in_Kg
0	2928.4

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

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

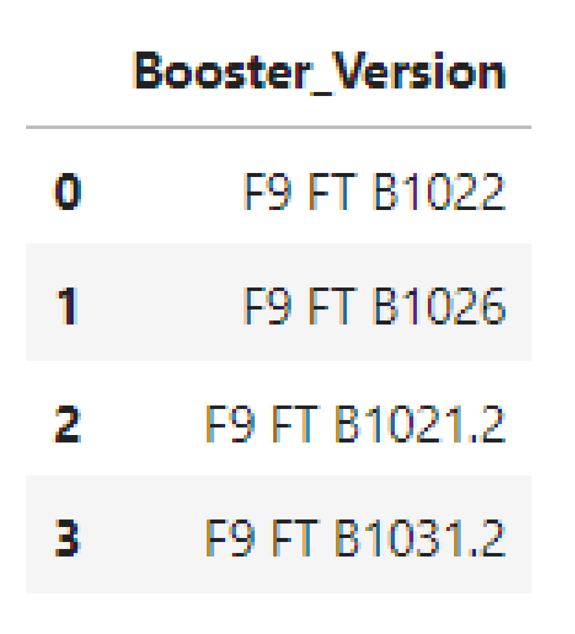


	First_Successful_Landing_Date
0	2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

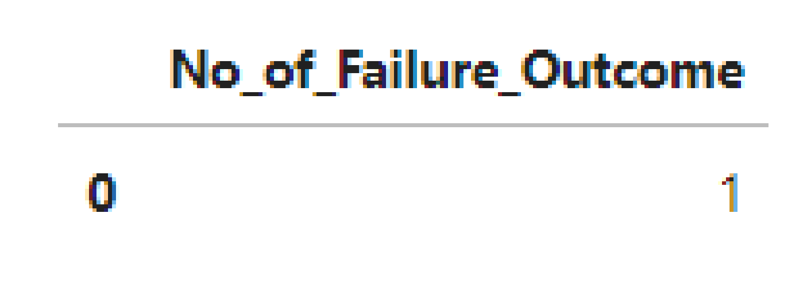


	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:



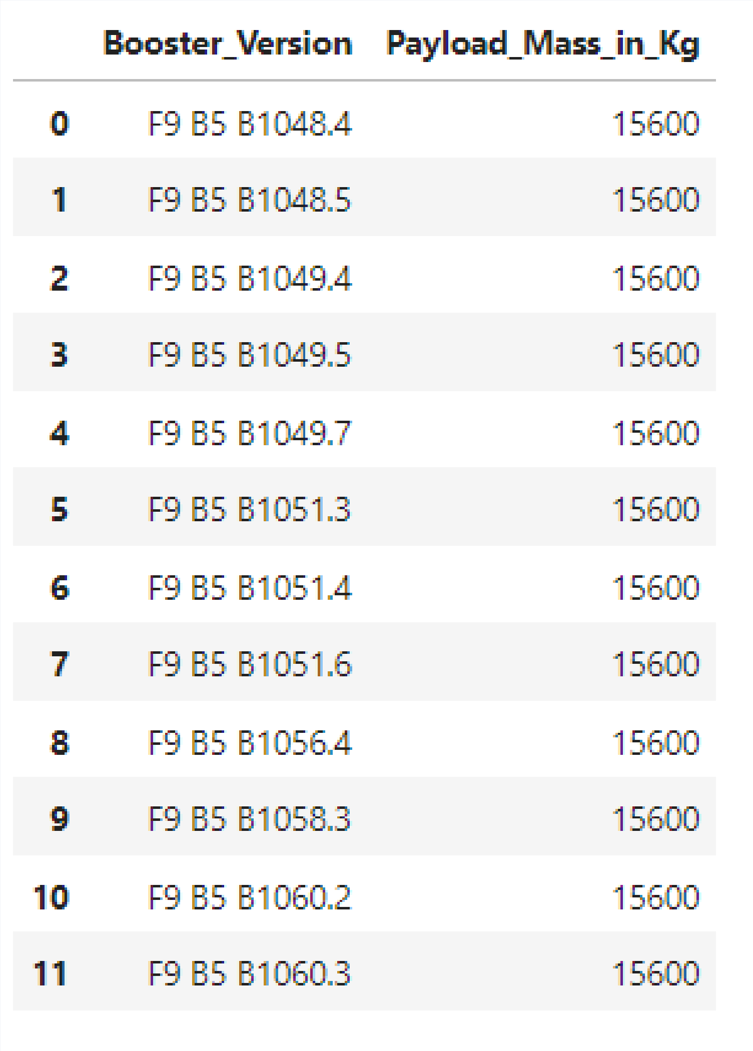
	No_of_Failure_Outcome
0	1

- 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:

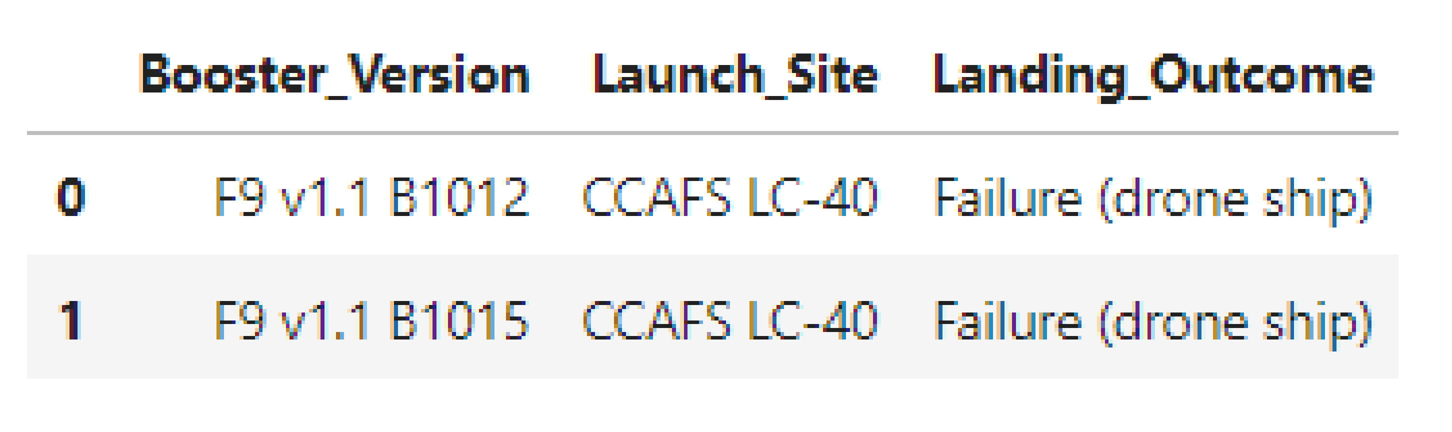
- These are the boosters which have carried the maximum payload mass registered in the dataset.



	Booster_Version	Payload_Mass_in_Kg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

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



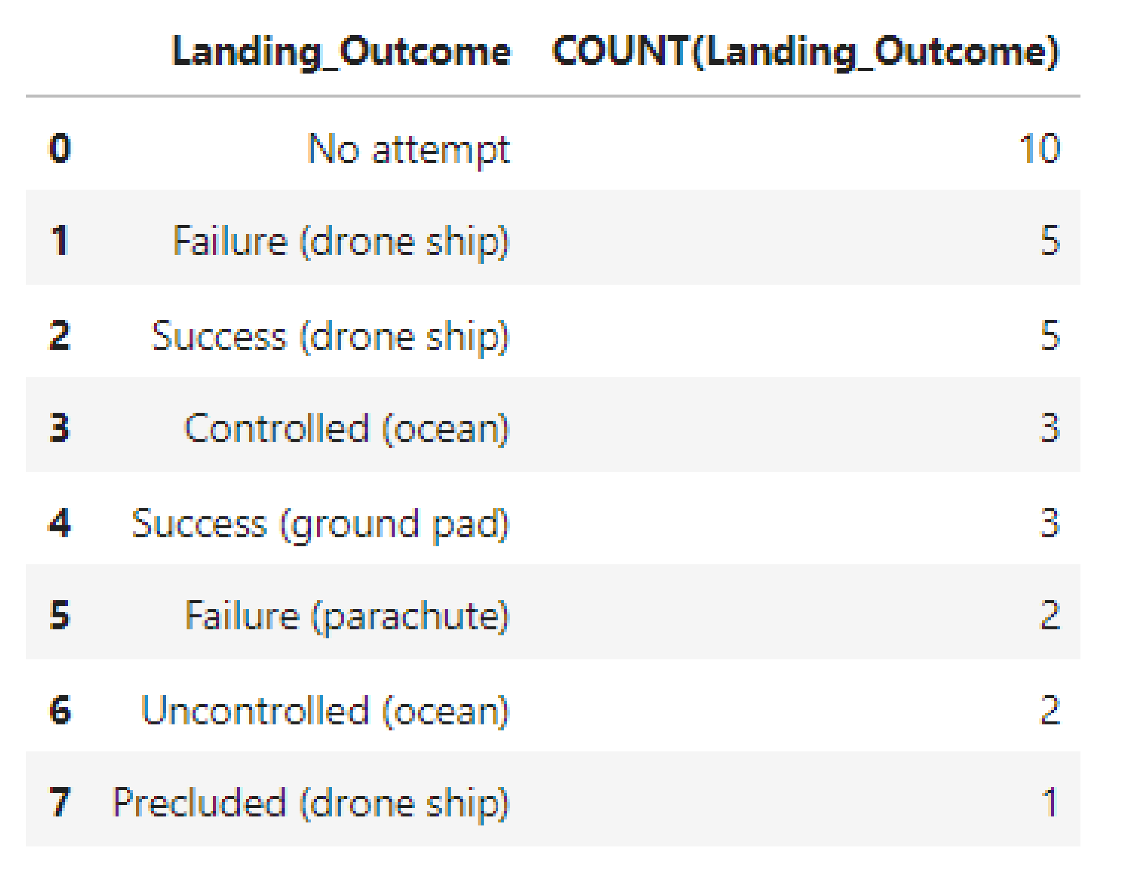
	Booster_Version	Launch_Site	Landing_Outcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

- This view of data alerts us that “No attempt” must be taken in account.



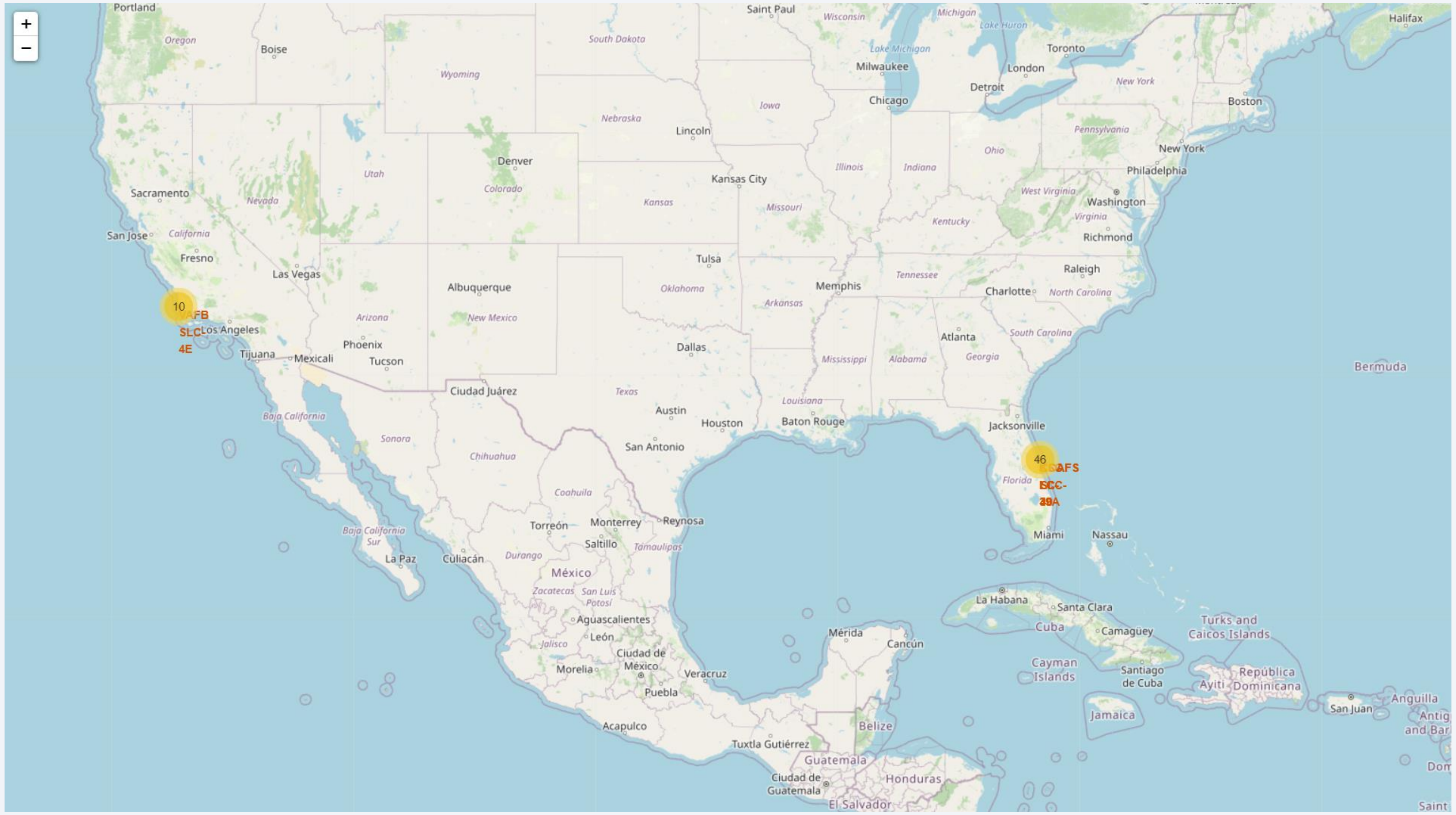
	Landing_Outcome	COUNT(Landing_Outcome)
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Failure (parachute)	2
6	Uncontrolled (ocean)	2
7	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

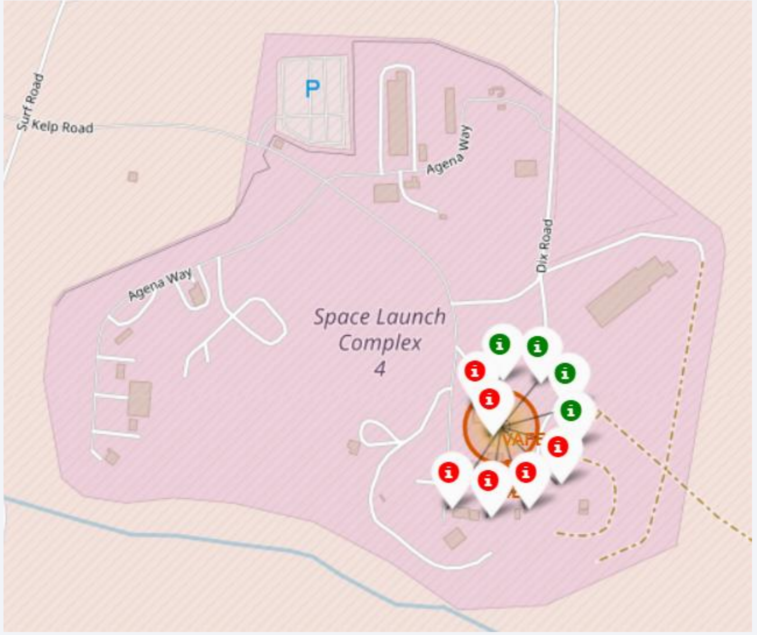
All Launch Sites Global Map Markers

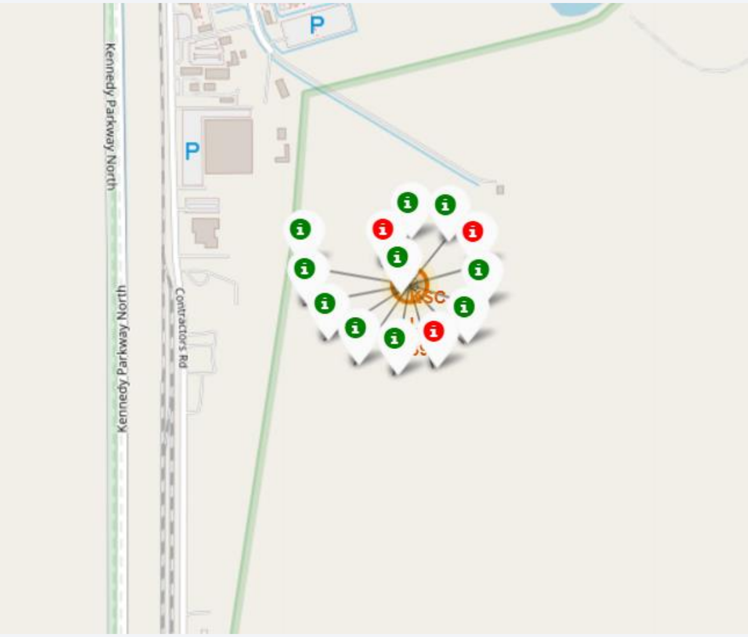
- Launch sites are near to sea and far from locality having good logistic infrastructure and communication by highway and railway.

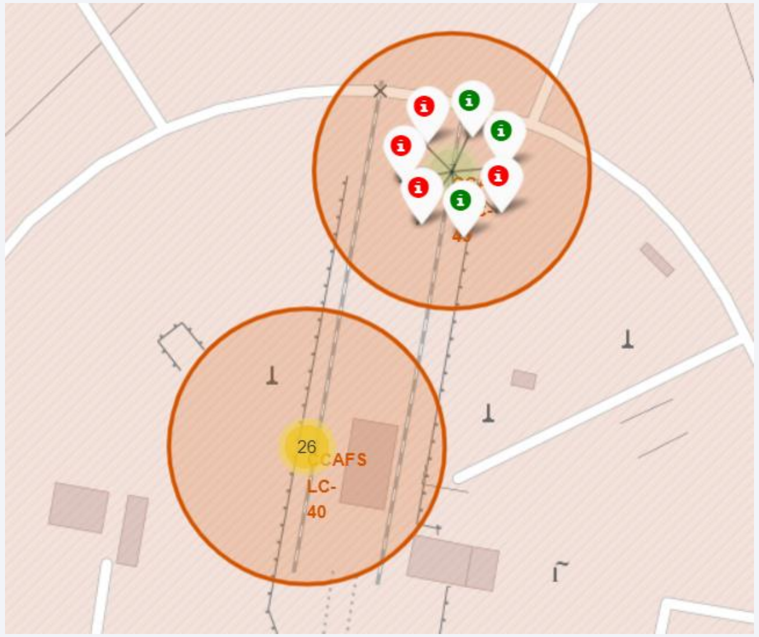


Launch sites Showing by Markers with Color Labels

- Launch site outcomes for VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 (left to right).
- Green markers indicate successful and red ones indicate failure.

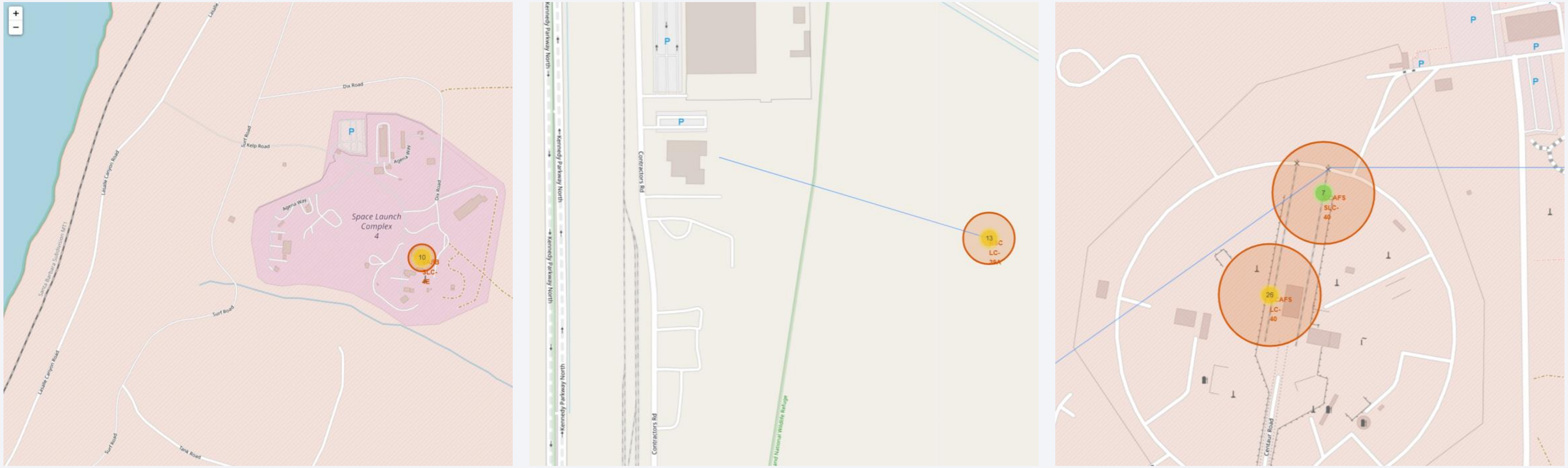


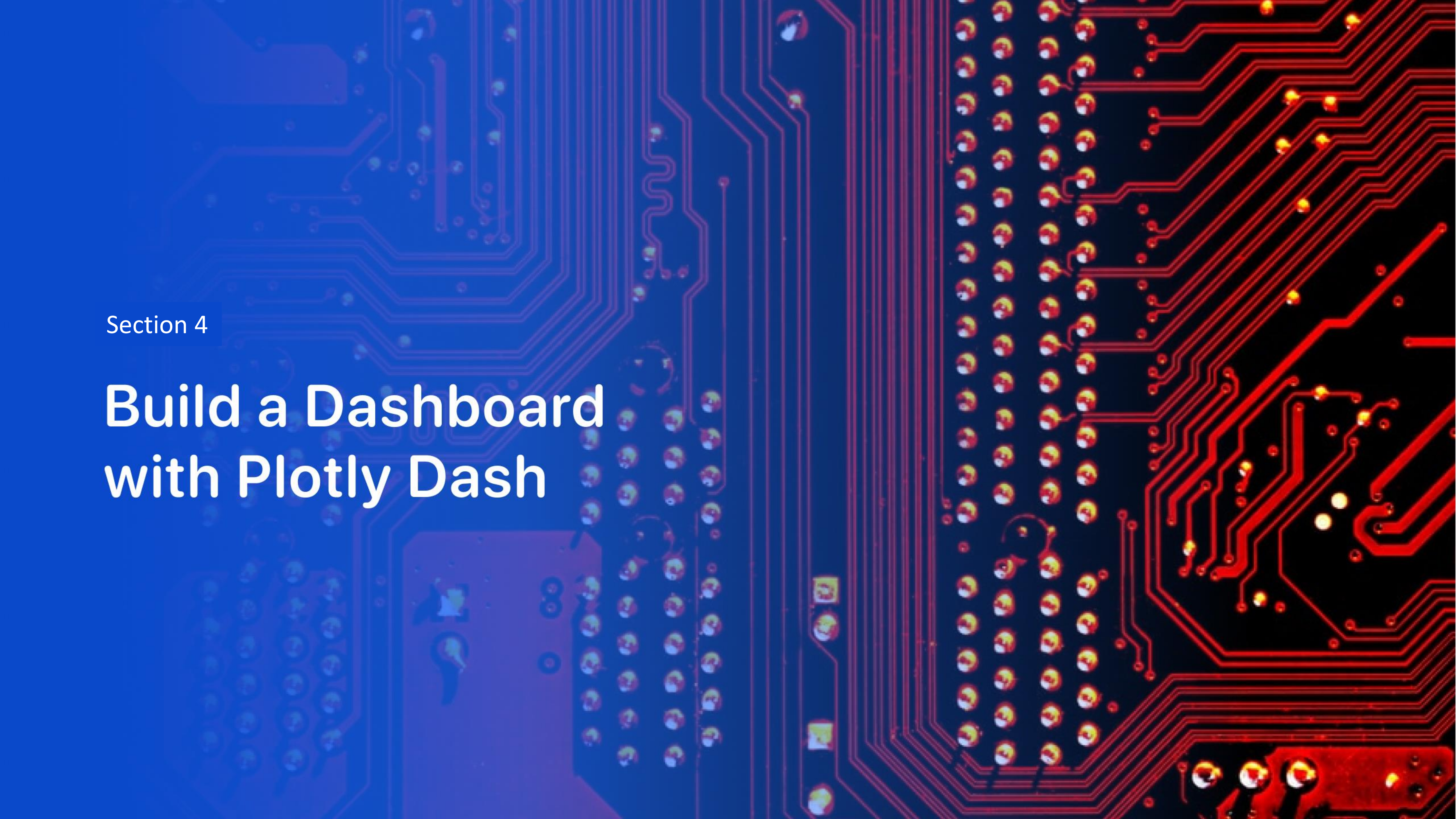




Launch Sites

- Launch sites (VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 (left to right)) have good logistics aspects along with railways and highways nearby but not very close. Launch sites are also far away from localities and close to coastline.



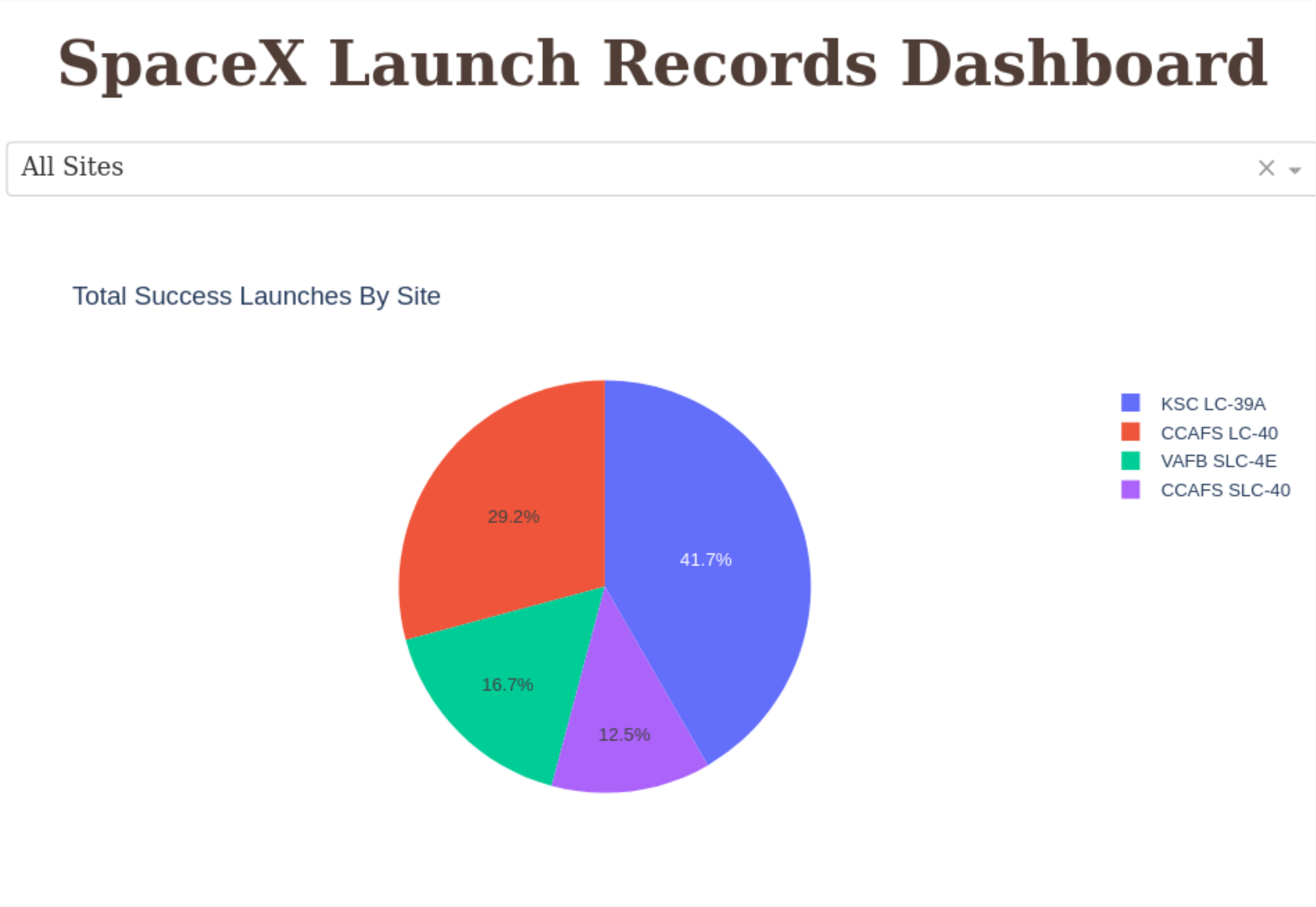


Section 4

Build a Dashboard with Plotly Dash

Pie Chart Showing Successful Launches by Site

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

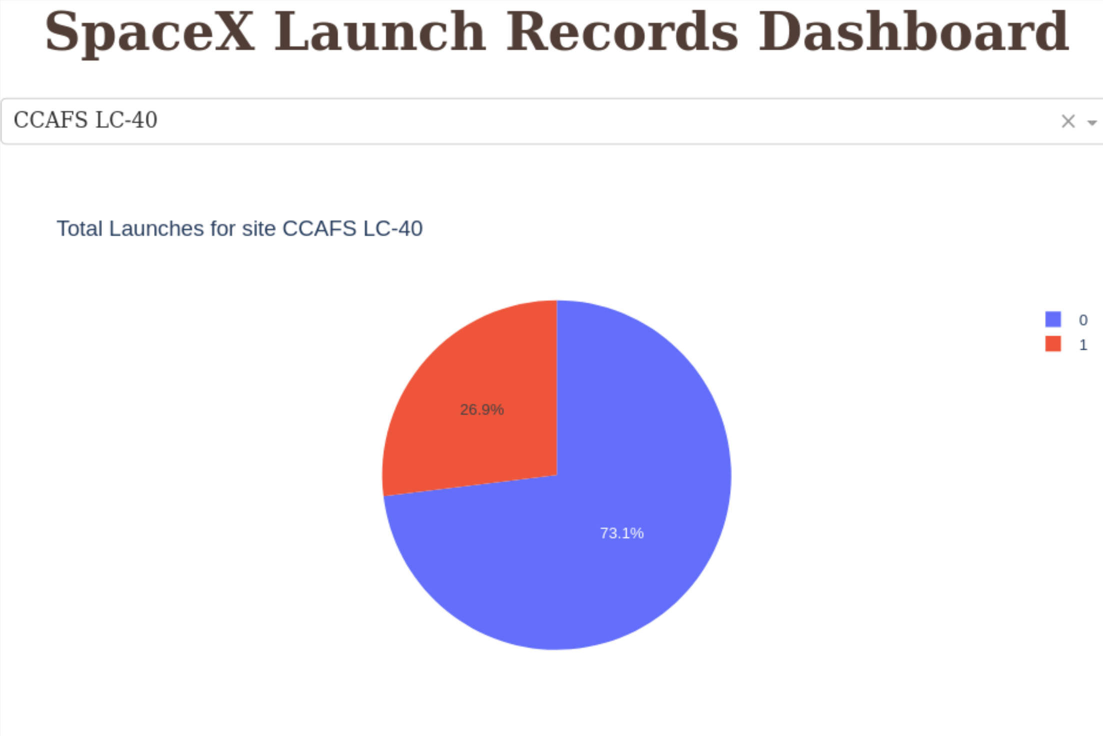


SpaceX Launch Records Dashboard

All Sites

Pie Chart Showing Launch Success Ratio for Launching Sites

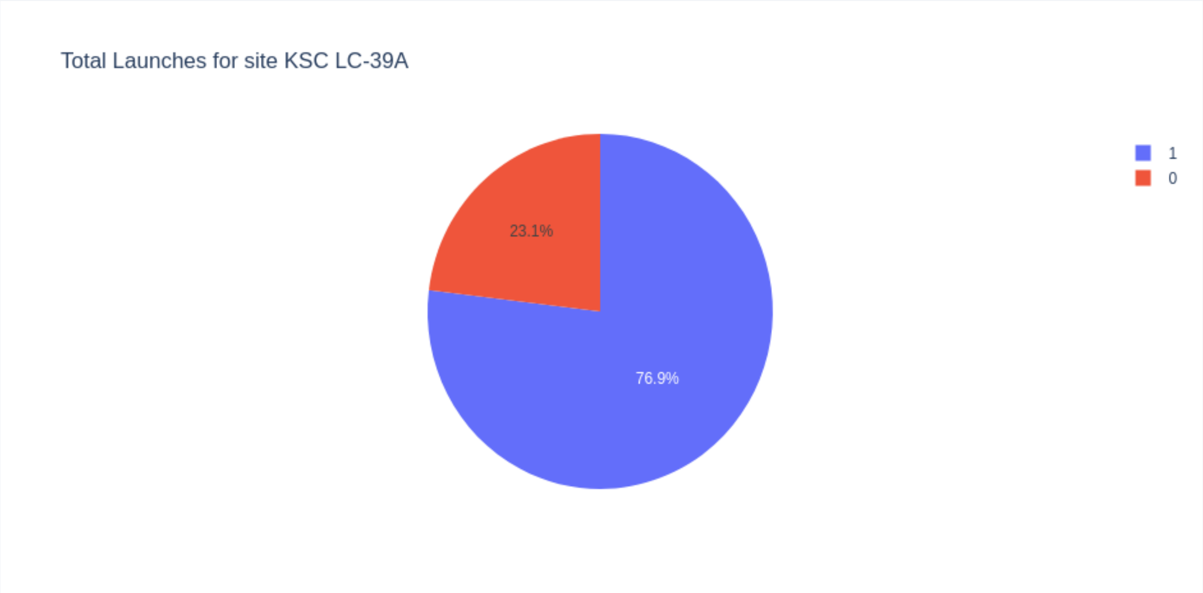
- 73.1% and 76.9% of launches are successful for CCAFS LC-40 and KSC LC-39A launching sites.



SpaceX Launch Records Dashboard

CCAFS LC-40 × ▾

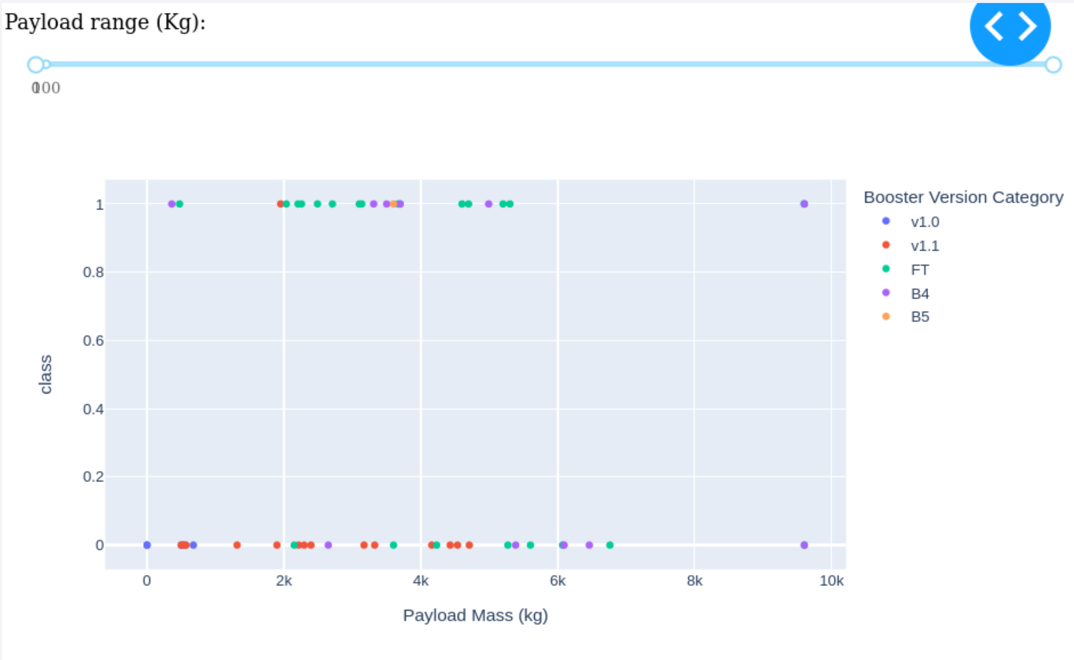
Total Launches for site CCAFS LC-40

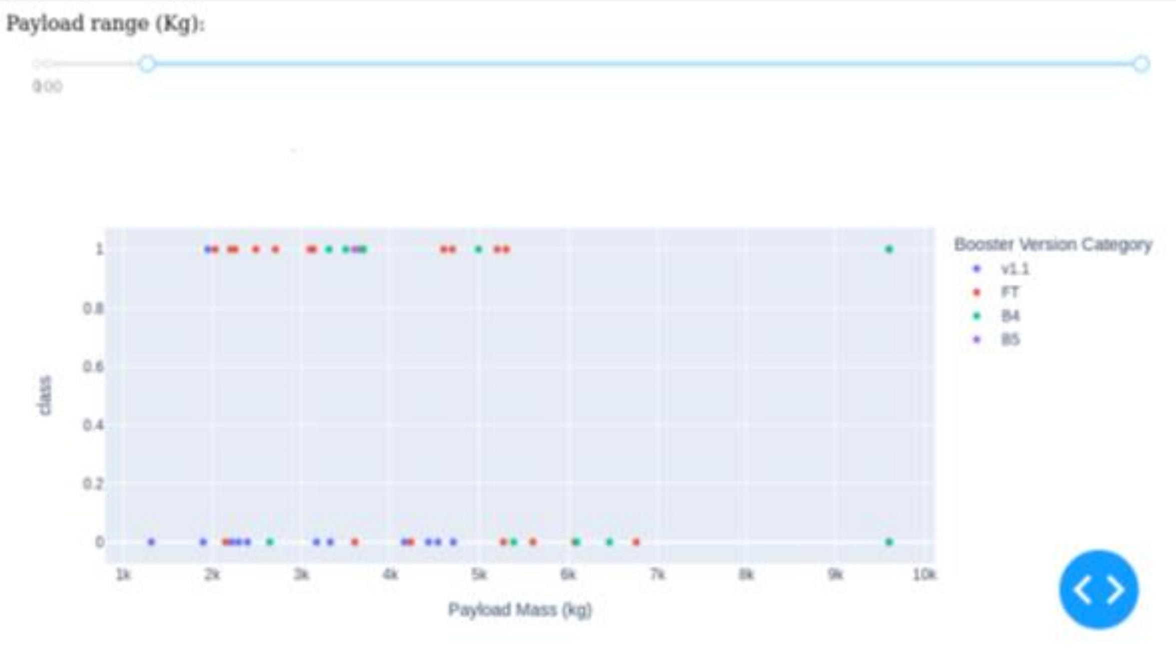


Total Launches for site KSC LC-39A

Graph for Payload vs Launch Outcome

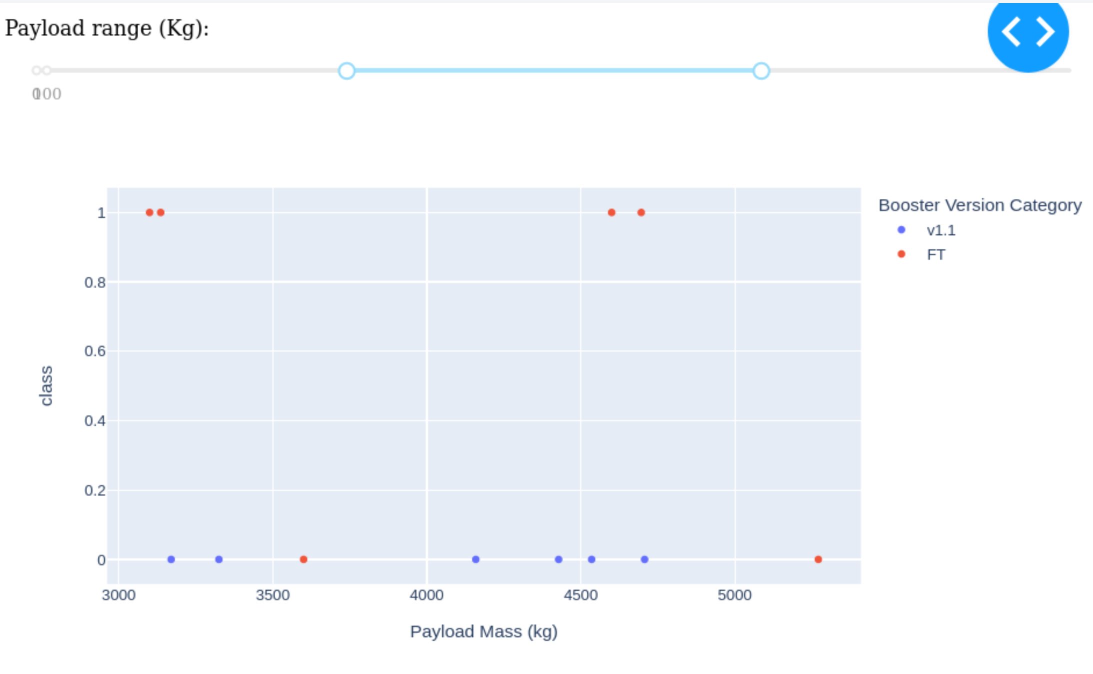
- Most of the launches are done by v1.1 and FT boosters.
- Payloads under 6000 kg and FT boosters are the most successful combination.

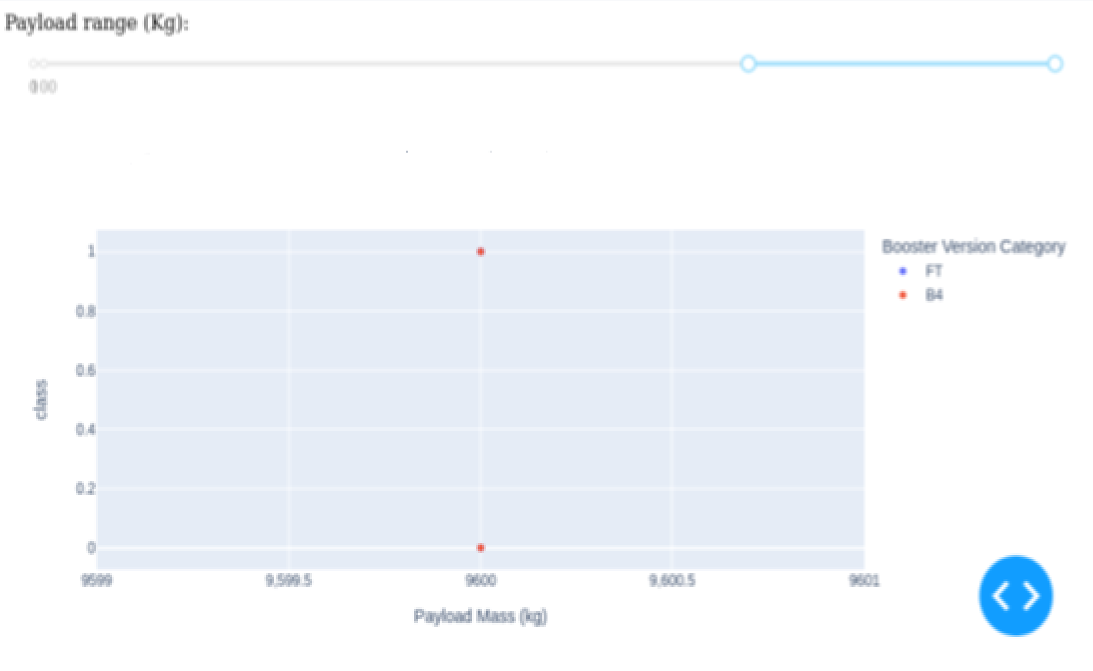




Graph for Payload vs Launch Outcome

- Success rate for FT booster version with payload between 4000 to 7000 kg is 100%.
- There's not enough data to estimate risk of launches over 7,000 kg.





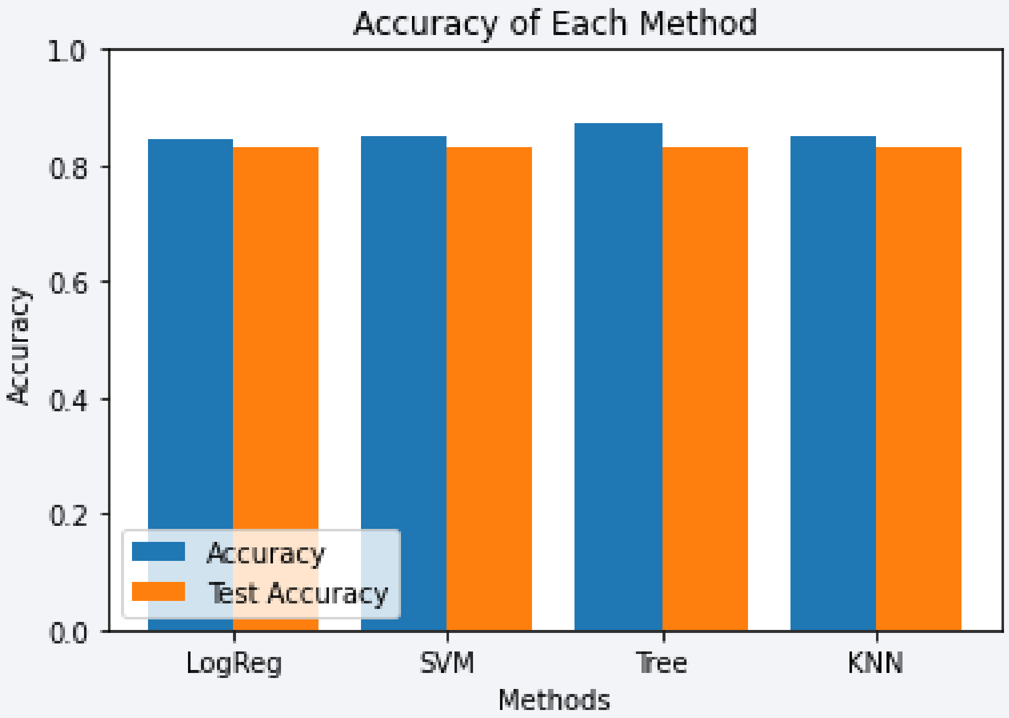
Section 5

Predictive Analysis (Classification)

Classification Accuracy

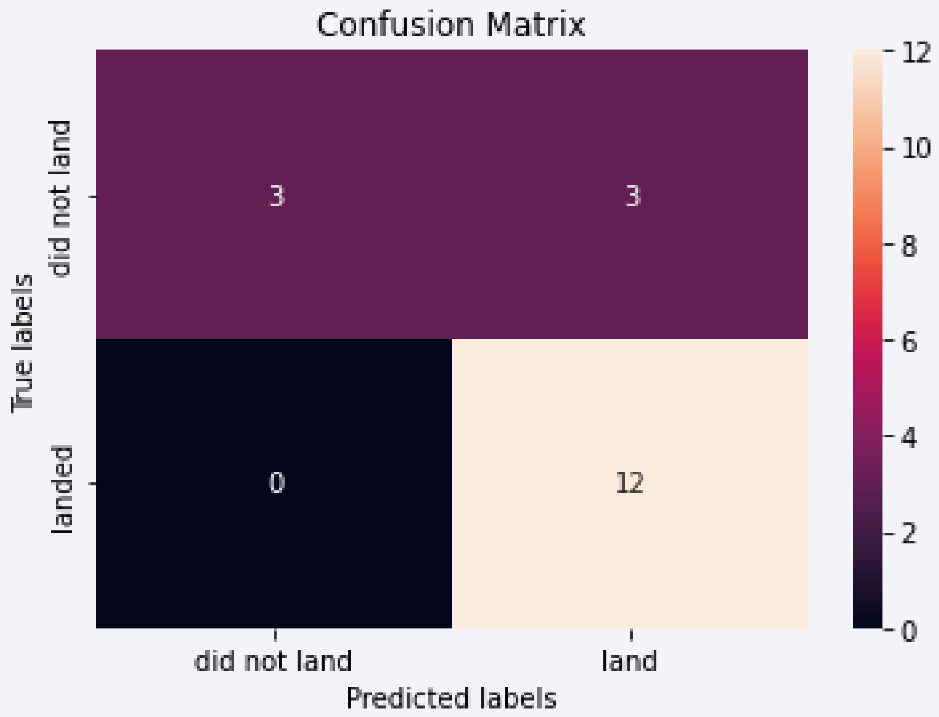
- Four classification models were tested, and their accuracies are plotted beside.

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets. Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- Launches above 7,000kg are less risky.
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- Google Colaboratory were used for analysis for its user friendliness and simplicity.
- All ipynb and py files were uploaded in Github repository.
- Folium and Plotly dash didn’t show maps on Github repository, instead screenshots were uploaded.
- All csv files, graphs, charts, plots, flowchart images are uploaded in Github repository.

Thank you!