IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SAMIRA GHORBANZADEH
2023-08-15

Outline

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

Executive Summary

- Summary of methodologies
- The following methodologies were used to analyze data:
- Data Collection using web scraping and SpaceX API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
- Machine Learning Prediction.
- Summary of all results
- It was possible to collected 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 by 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.

- 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

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources: • Space X API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (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
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

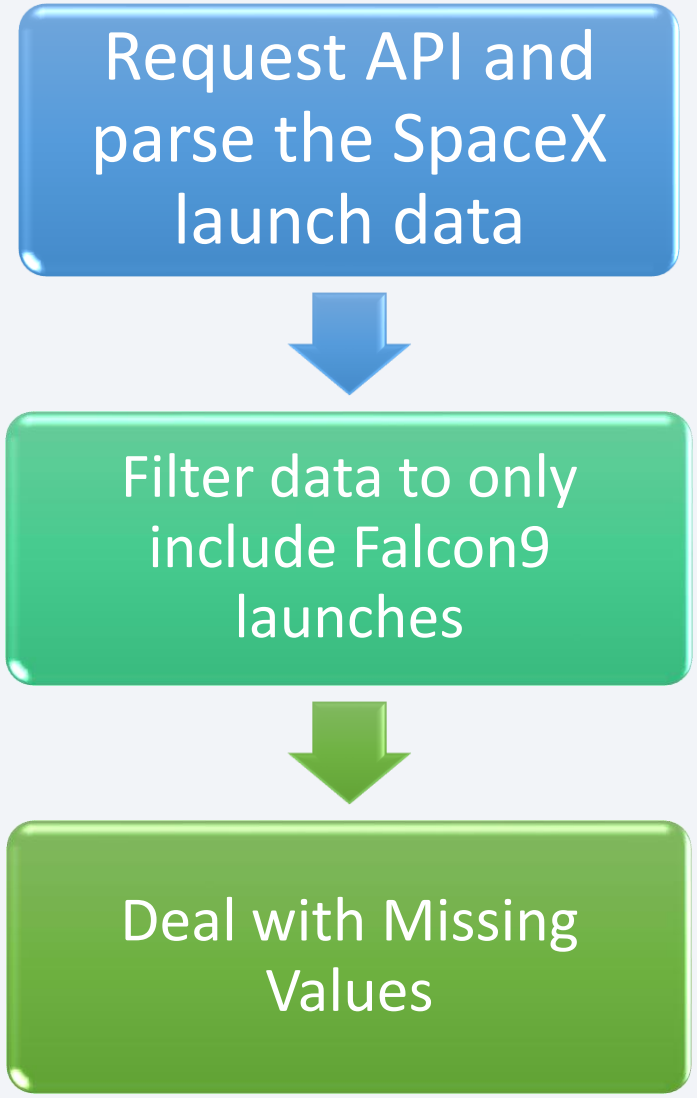
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

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 technics.

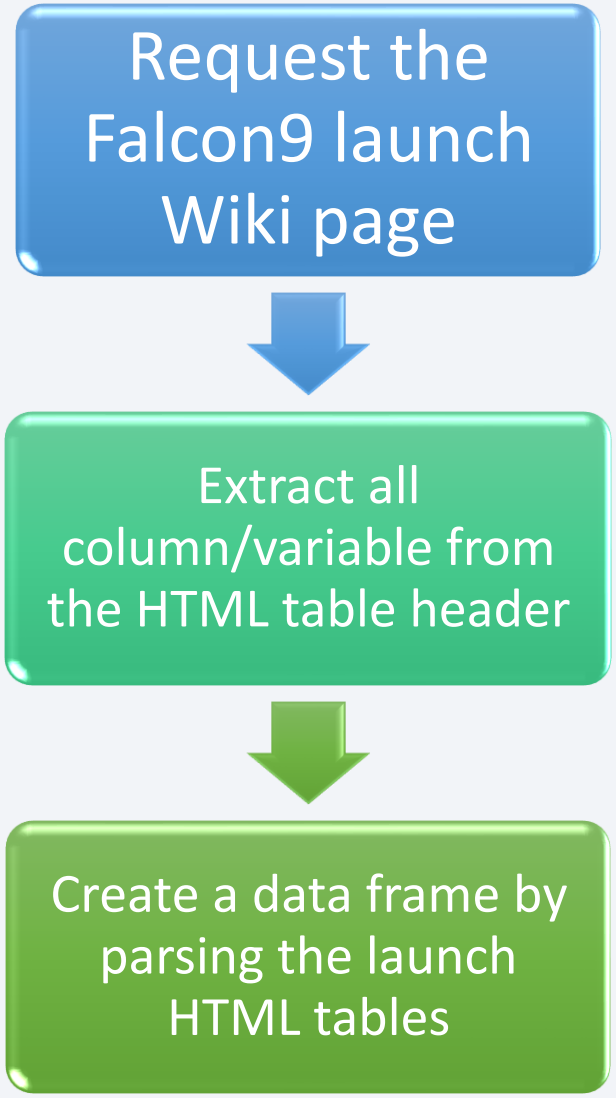
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week1/jupyter_labs_spacex_data_collection_api.ipynb



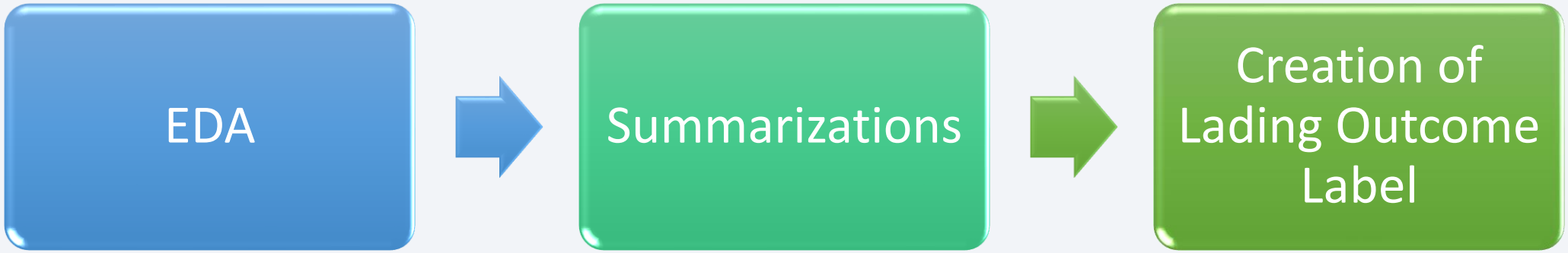
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week1/jupyter_labs_webscraping.ipynb



Data Wrangling

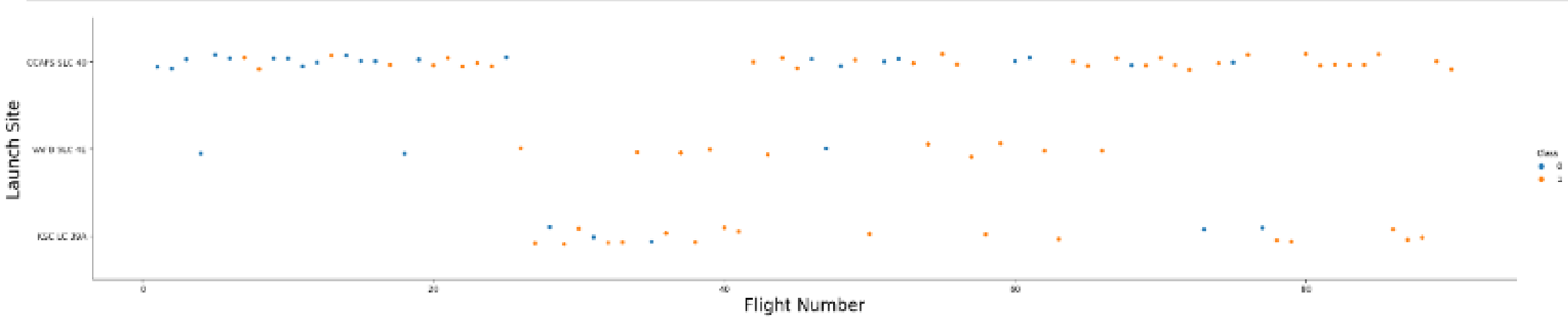
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week1/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week2/jupyter_labs_eda_dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
- Names of the 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;
- Date when the first successful landing outcome in ground pad was achieved;
- 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;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week2/jupyter_labs_eda_sql_coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates

- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week3/lab_jupyter_launch_site_location.ipynb

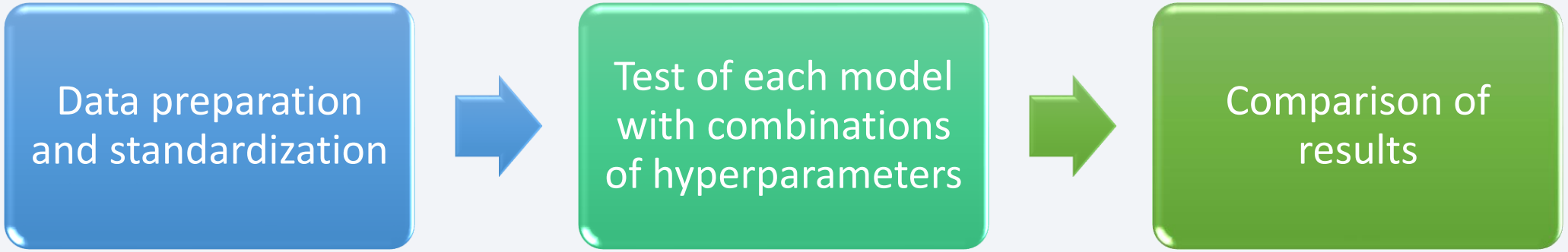
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week3/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models are compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



- Source code: https://github.com/Samira492/DataScienceCapstone-spacex/blob/main/week4/Machine%20Learning%20Prediction.ipynb

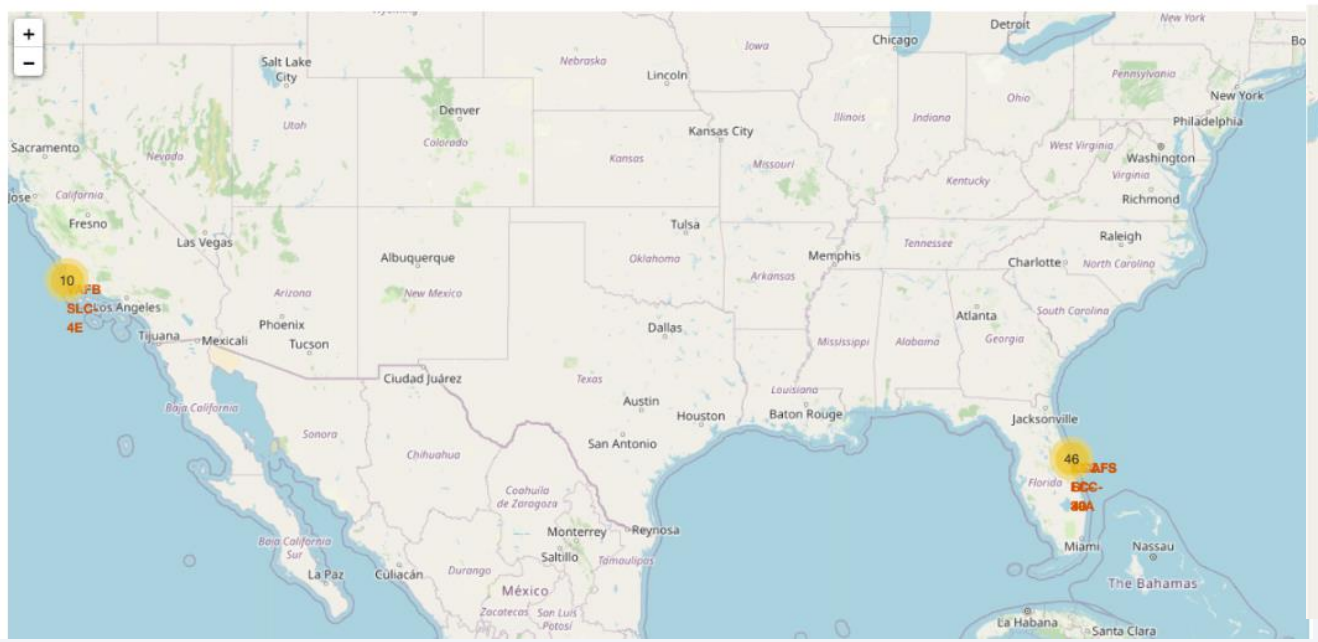
Results

- Exploratory data analysis results
  - Space X uses four 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

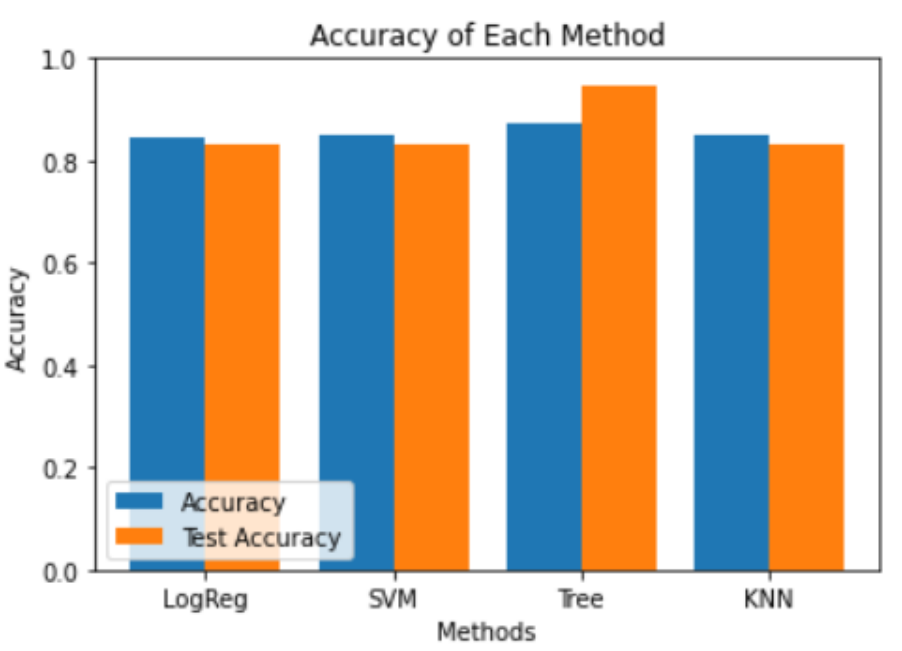
- Using interactive analytics was possible to identify that launch sites used to be in safety places, near sea, for example and have a good logistic infrastructure around.

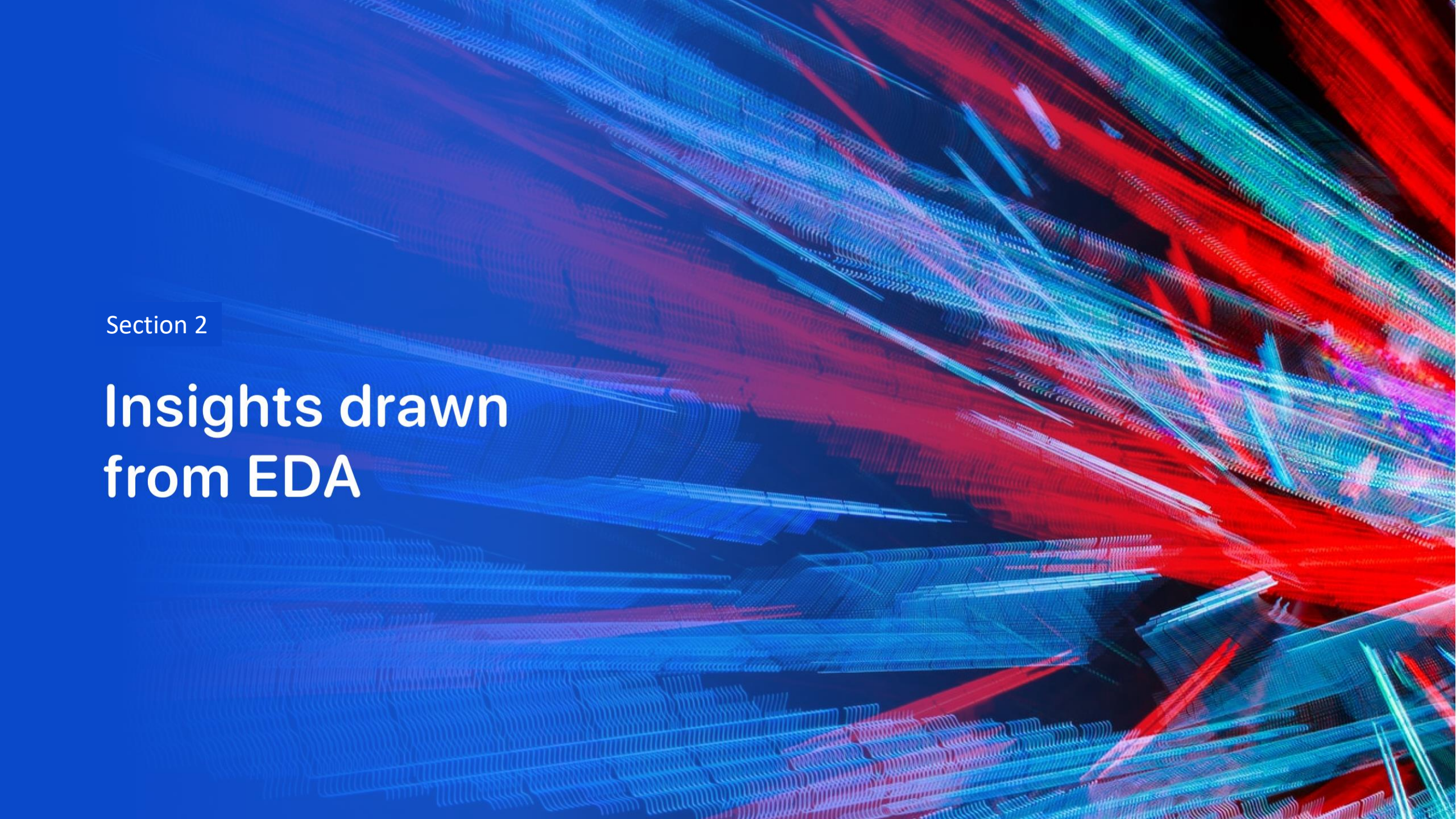
- Most launches occurs at east cost launch sites.



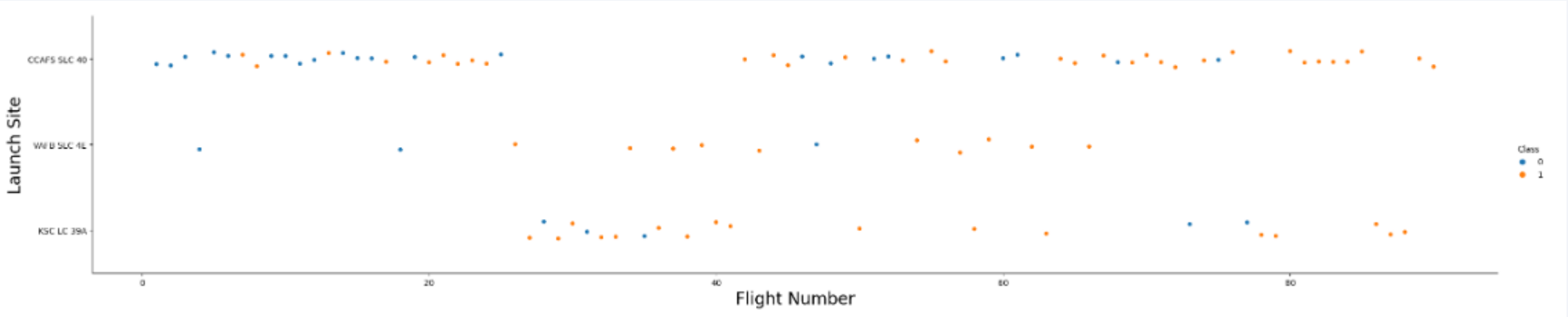
Results

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



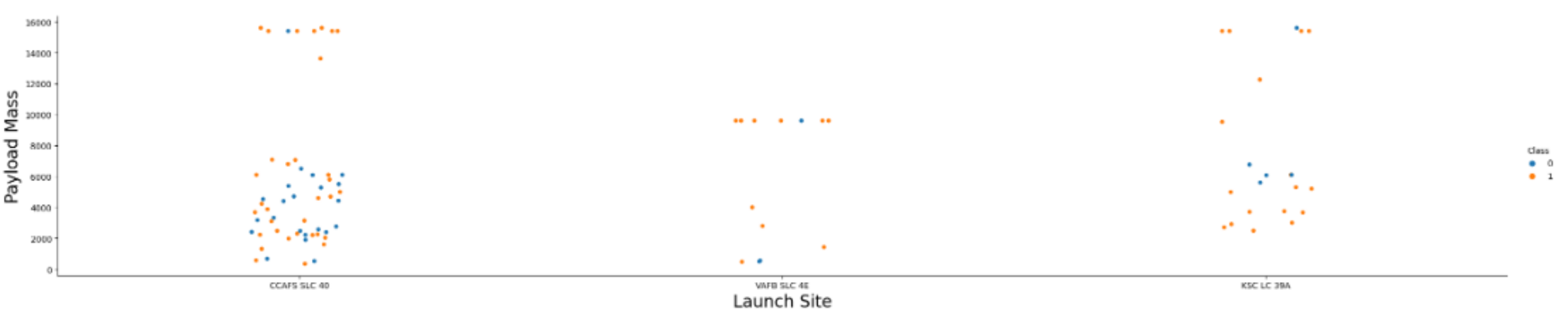

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The above plot shows 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.

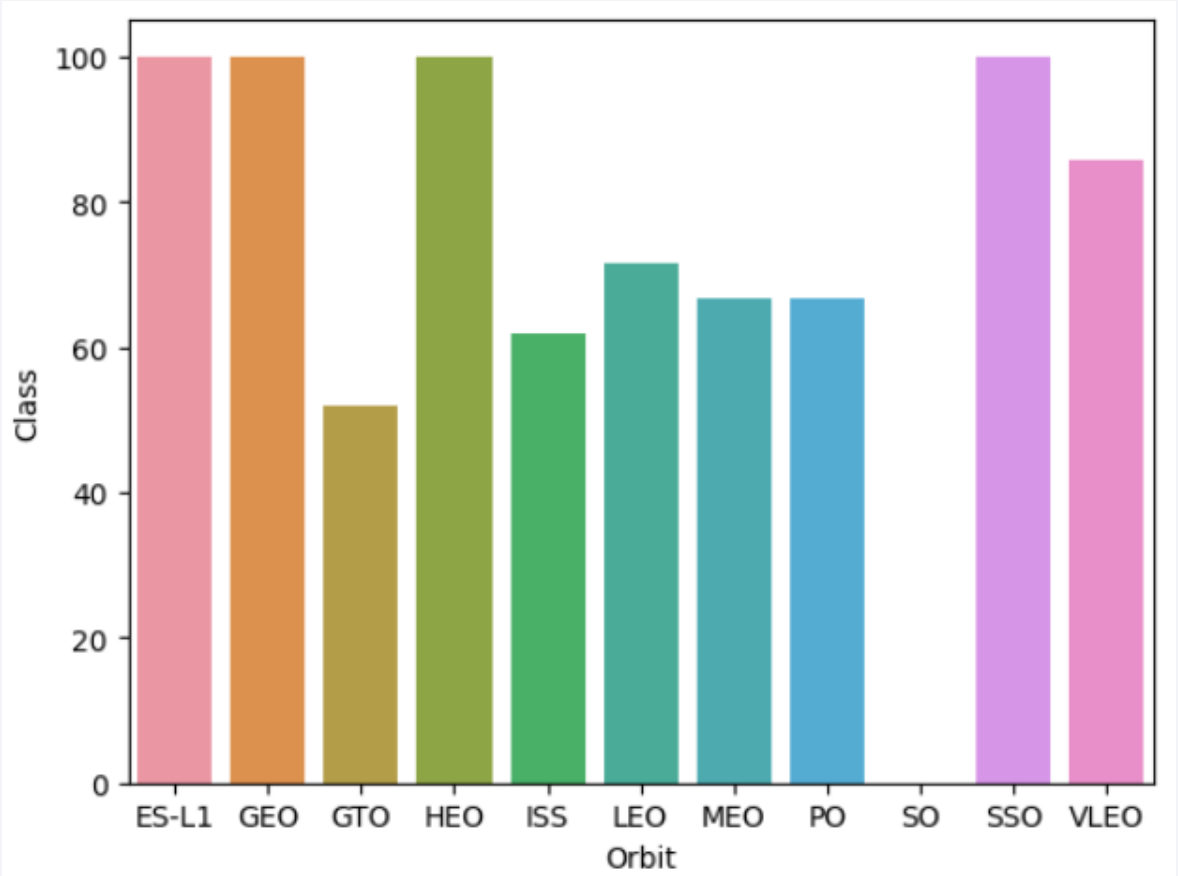
Payload vs. Launch Site



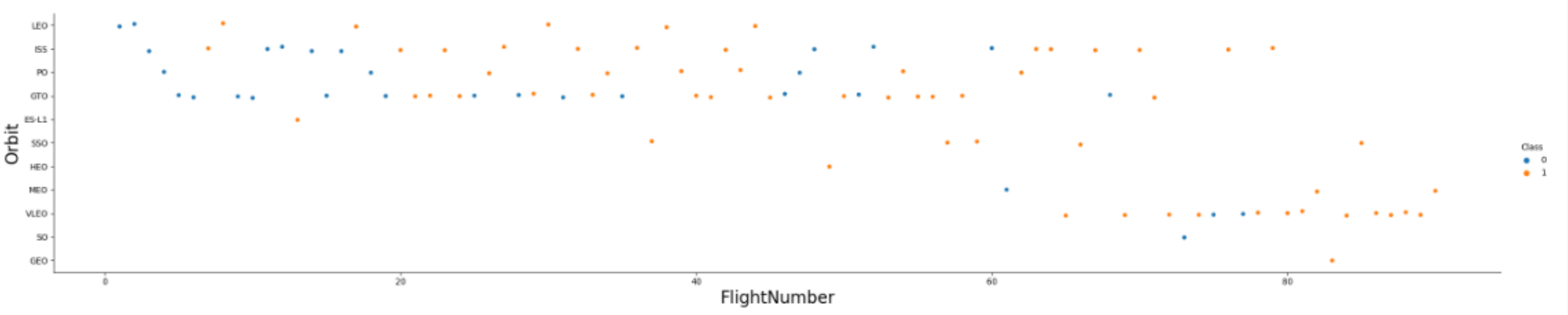
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The highest success rates occurs in orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO
- And this is followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

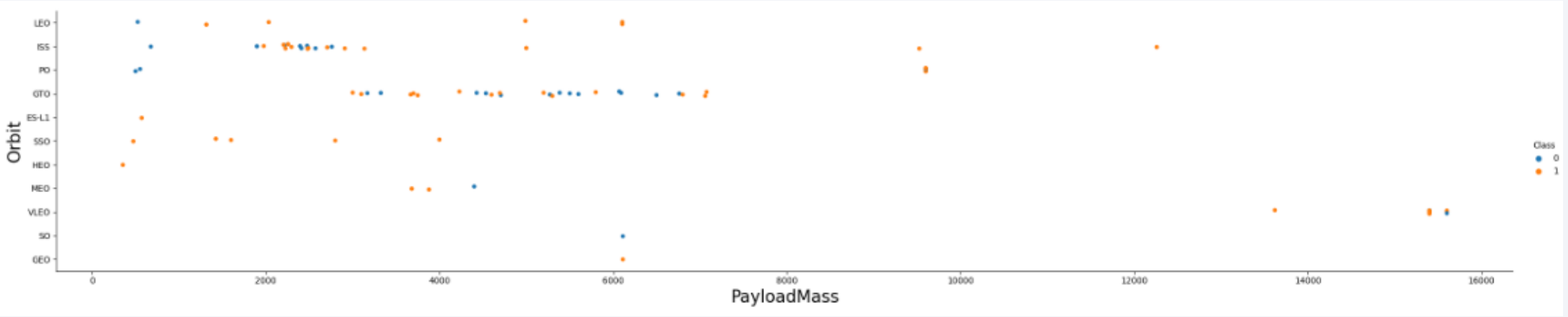


Flight Number vs. Orbit Type



- It seems that success rate improved over time to all orbits;
- The results show an increase in the frequency of VLEO orbit.

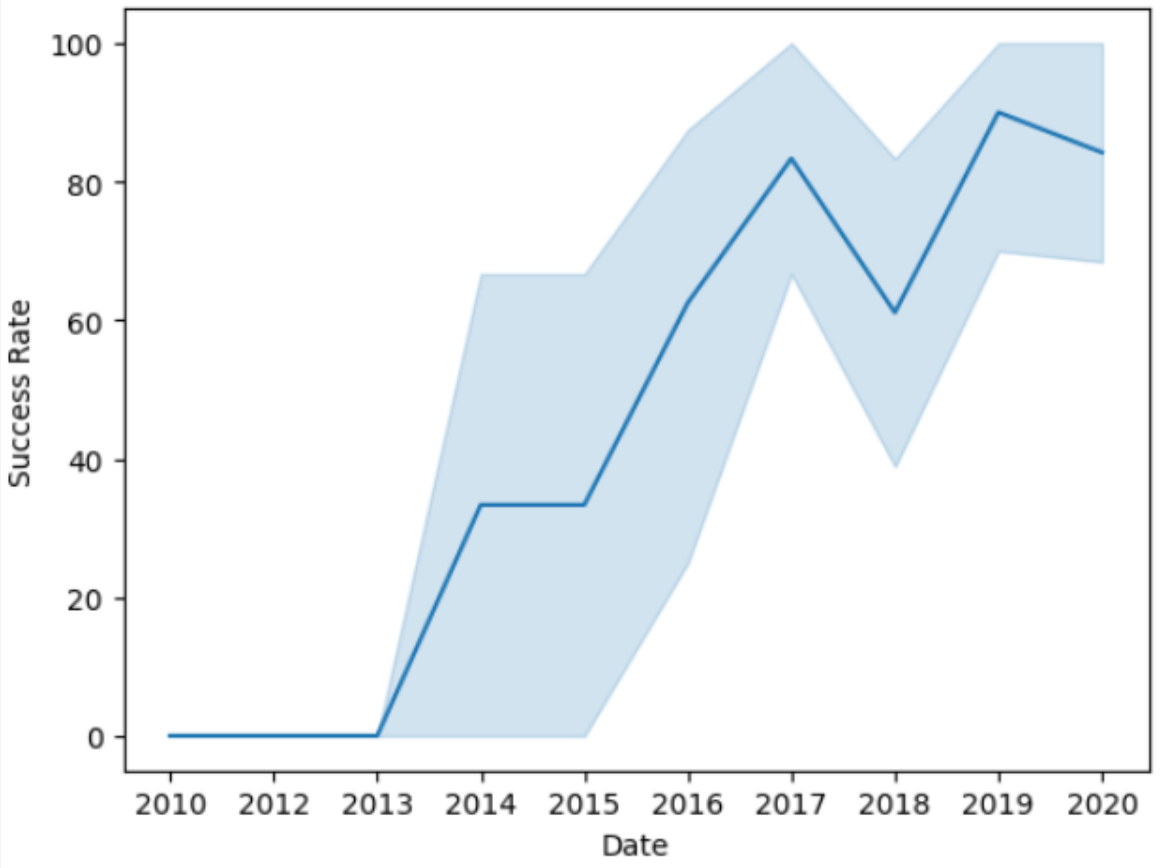
Payload vs. Orbit Type



- It seems that there is almost no relation between payload and success rate to orbit GTO;
- Few launches can be seen for the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing from 2013 and after a few fluctuations between 2017 and 2018 hit the pic to around 90% in 2019 before slight decreasing;
- It remained constant after 2010 in the following three years, maybe because of limitted technology in this period.



All Launch Site Names

- There are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- This obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- Launch sites beginning with `CCA` are:

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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- These five records representing samples of Cape Canaveral launches

Total Payload Mass

- Total payload mass carried by boosters from NASA:

Total Payload (kg)
111.268

- This figure calculated 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:

Avg Payload (kg)
2.928

- By filtering data with 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:

Min Date
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- After selecting distinct booster versions according to the filters above, these four are the results.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group resulted in the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

- The above list shows which boosters have carried the maximum payload mass registered in the dataset.

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- These tow boosters are the only occurrences.

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:

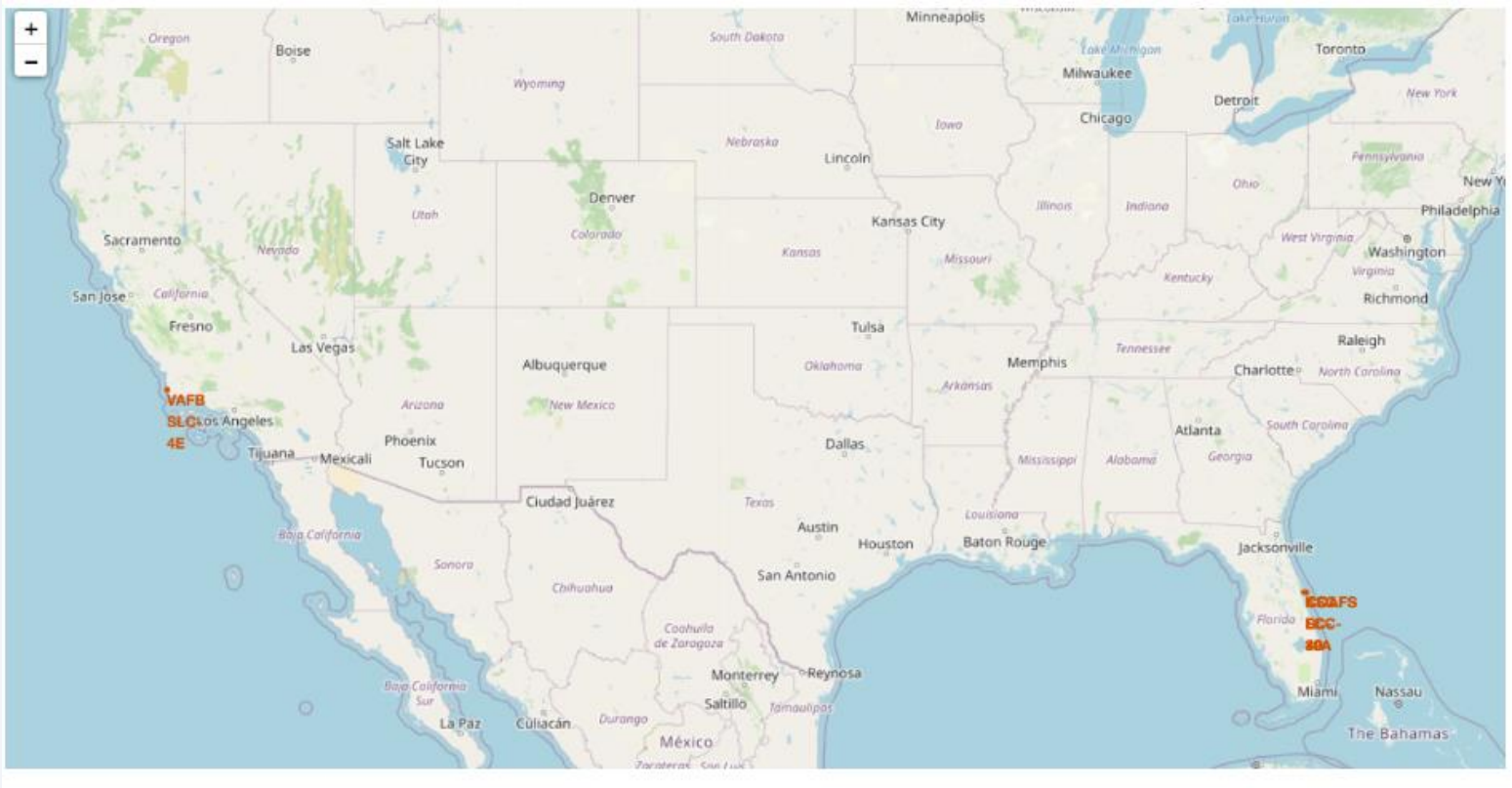
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- These results show that “No attempt” must be taken in account.

Section 3

Launch Sites Proximities Analysis

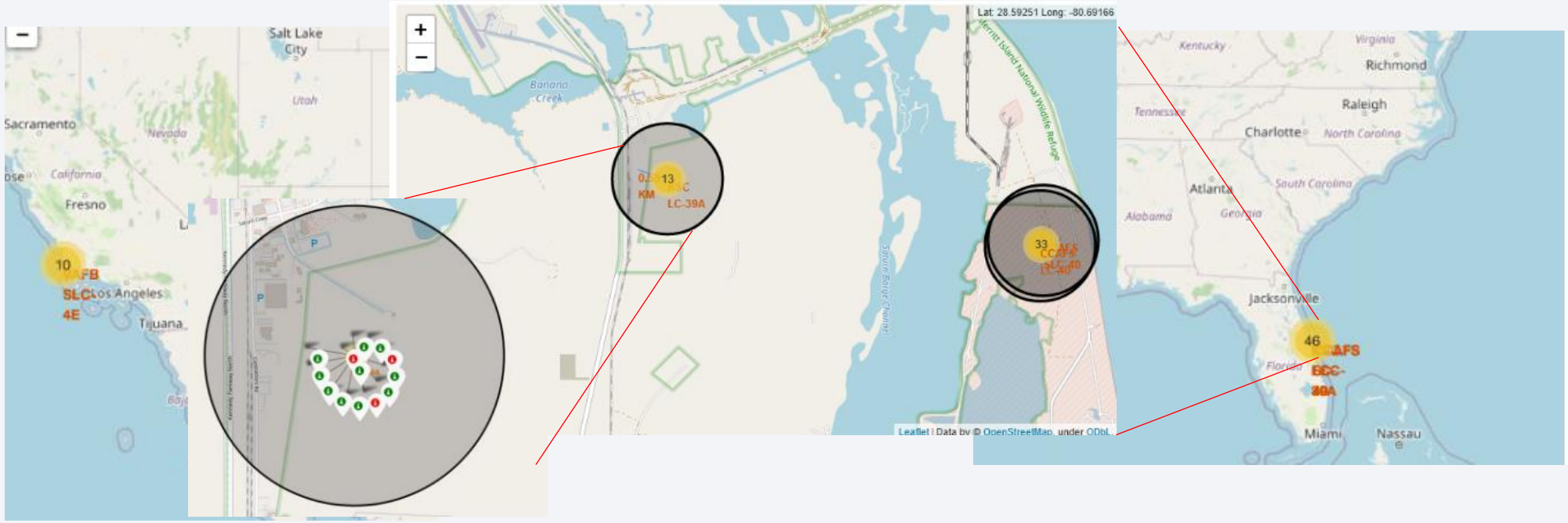
All launch sites



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

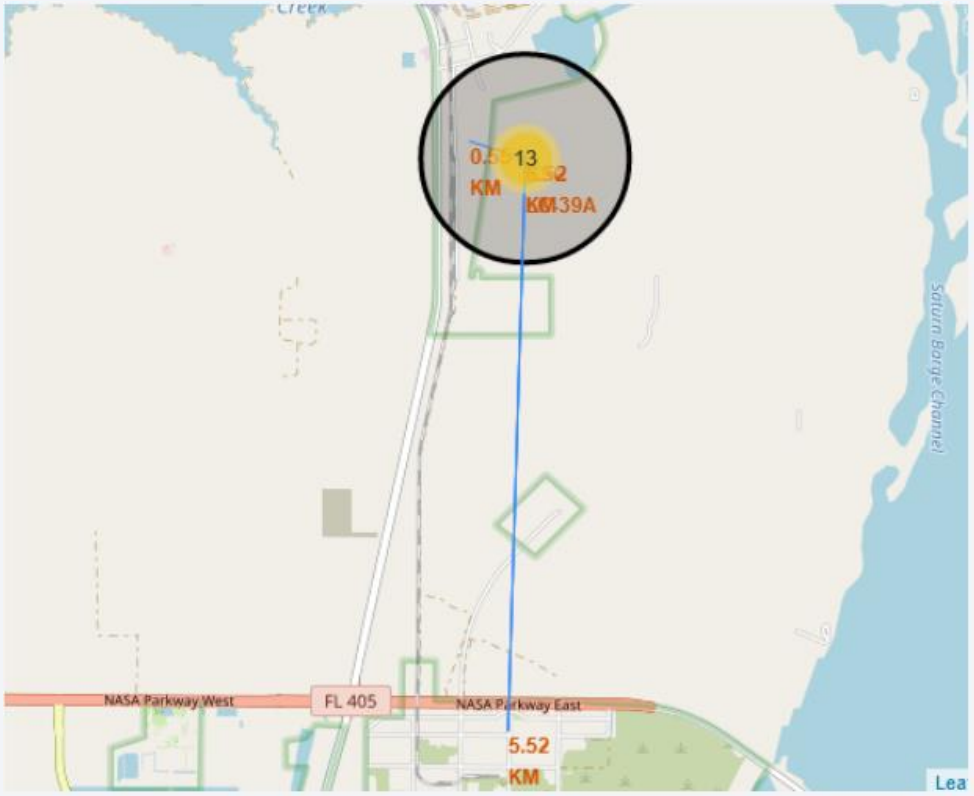
Launch Outcomes by Site

- Cases for KSC LC-39A launch site launch outcomes

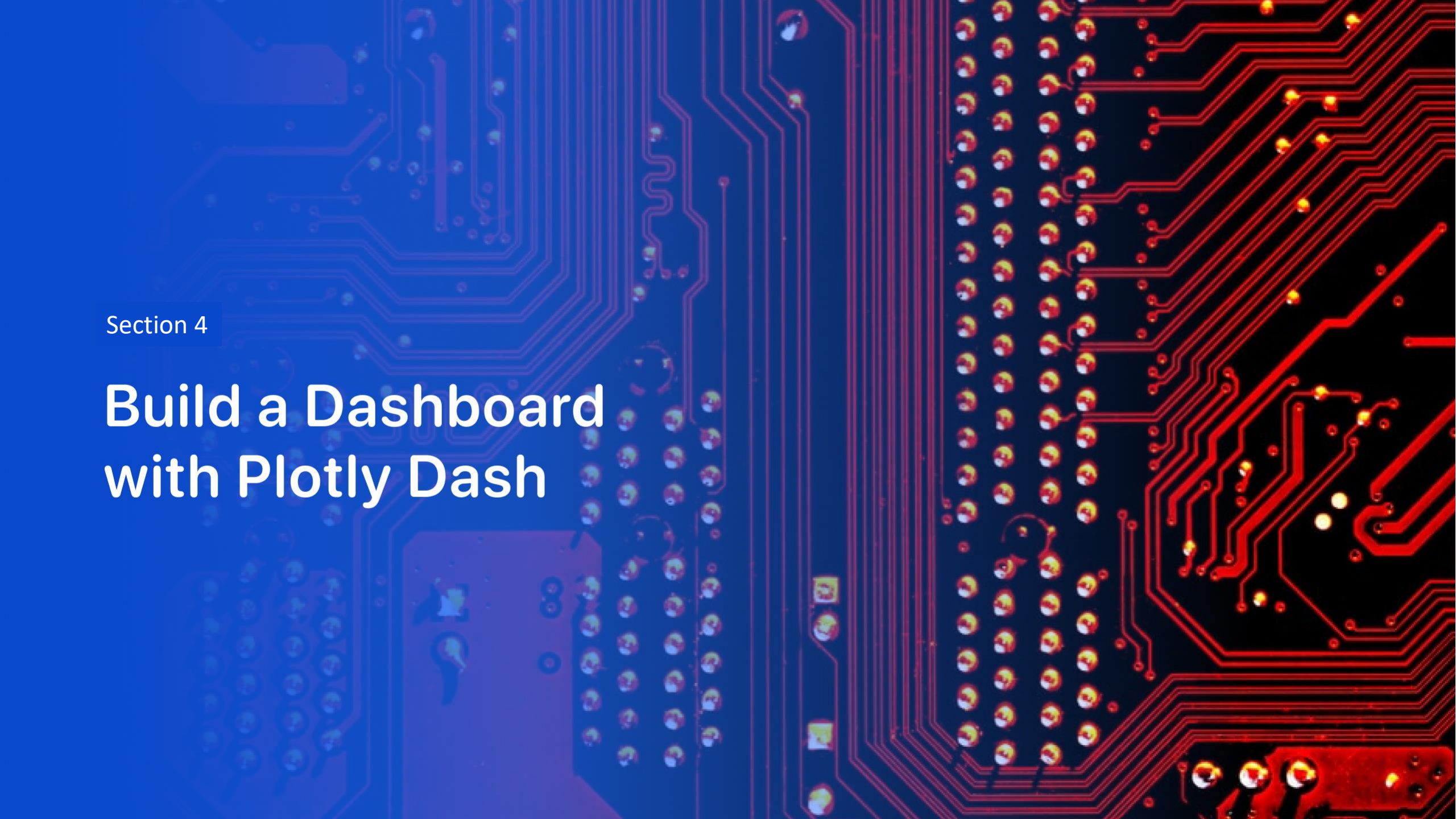


- Green points indicate successful and red ones indicate failure.

Logistics and Safety

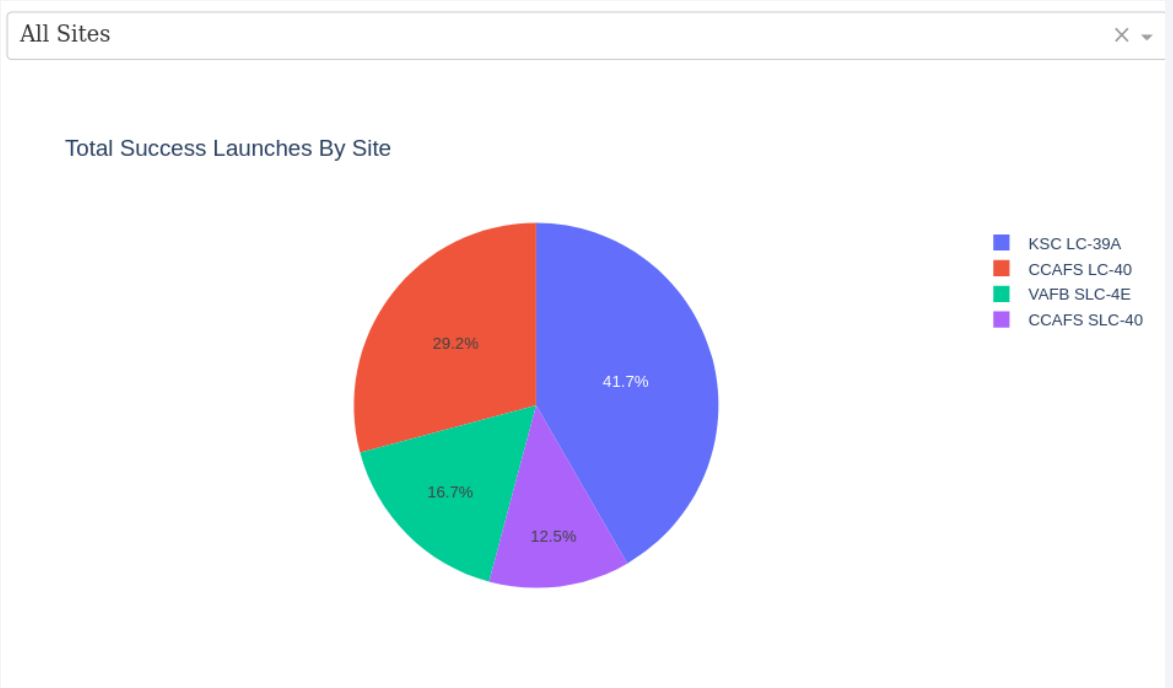


- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from populated areas.


Section 4
Build a Dashboard with Plotly Dash

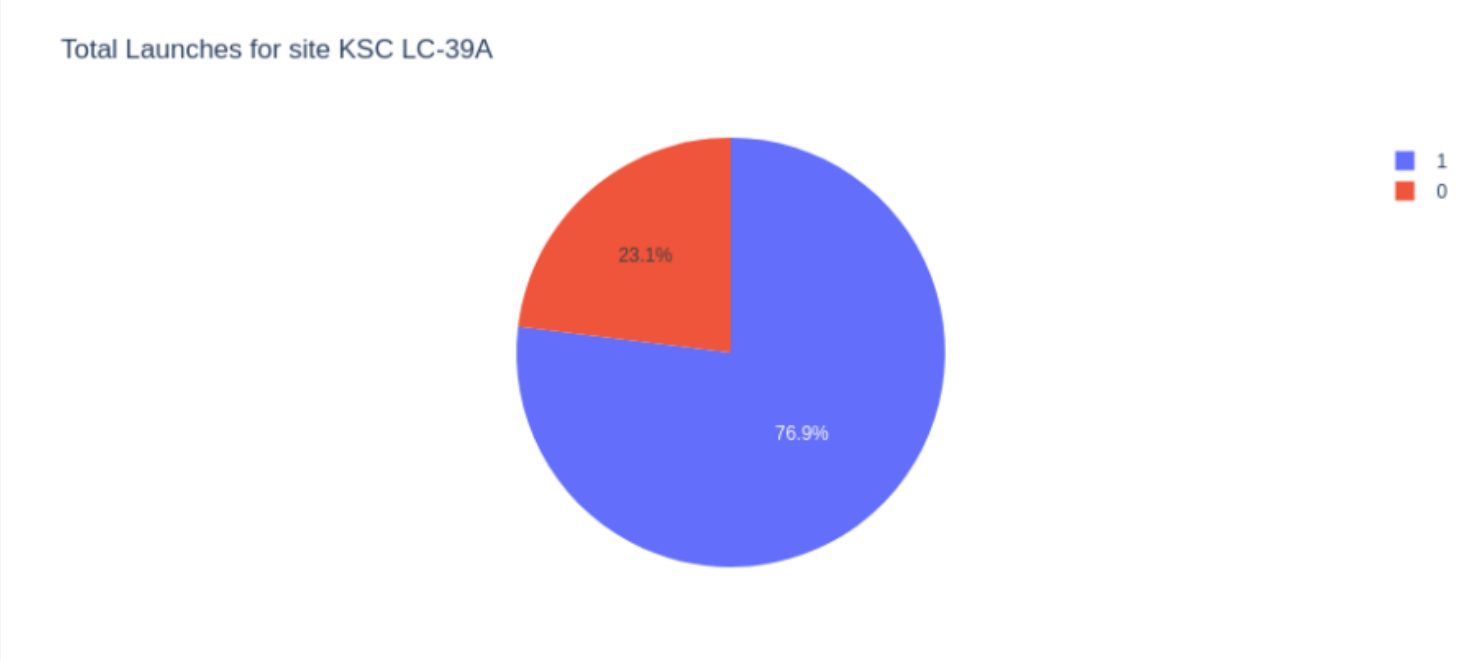
Successful Launches by Site

- Apparently, the location from where launches occur is a very important factor of success of missions.

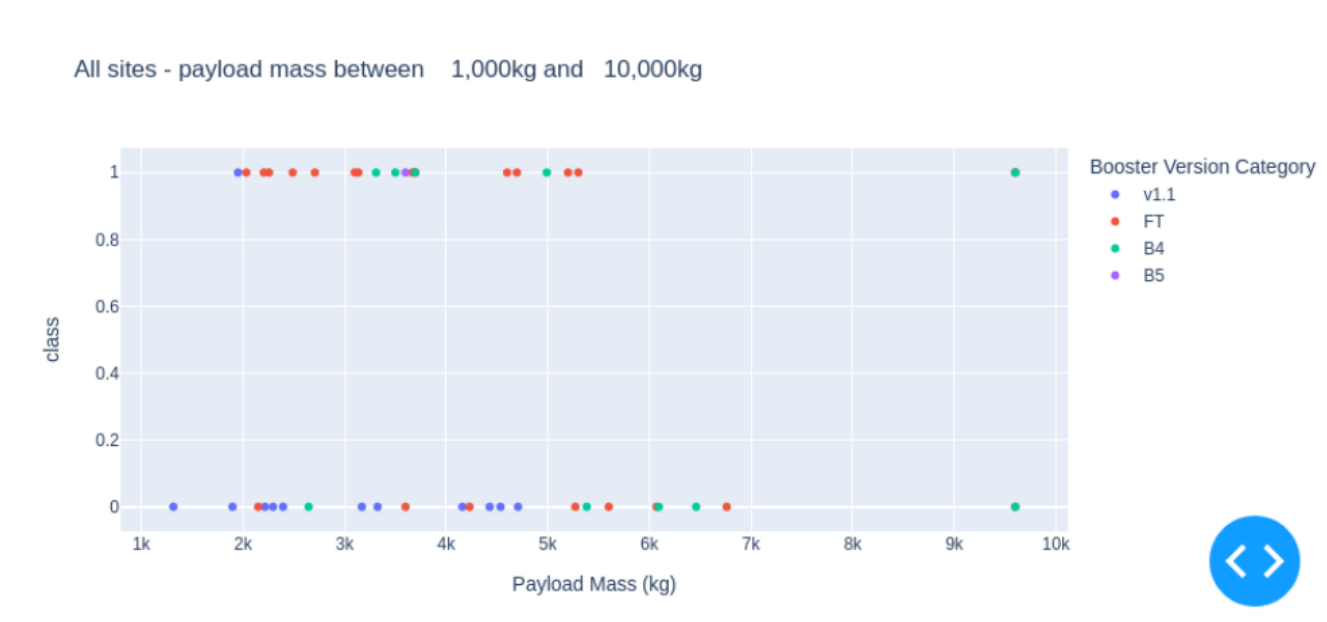


Launch Success Ratio for KSC LC-39A

- Around 77% of launches are successful in this site.



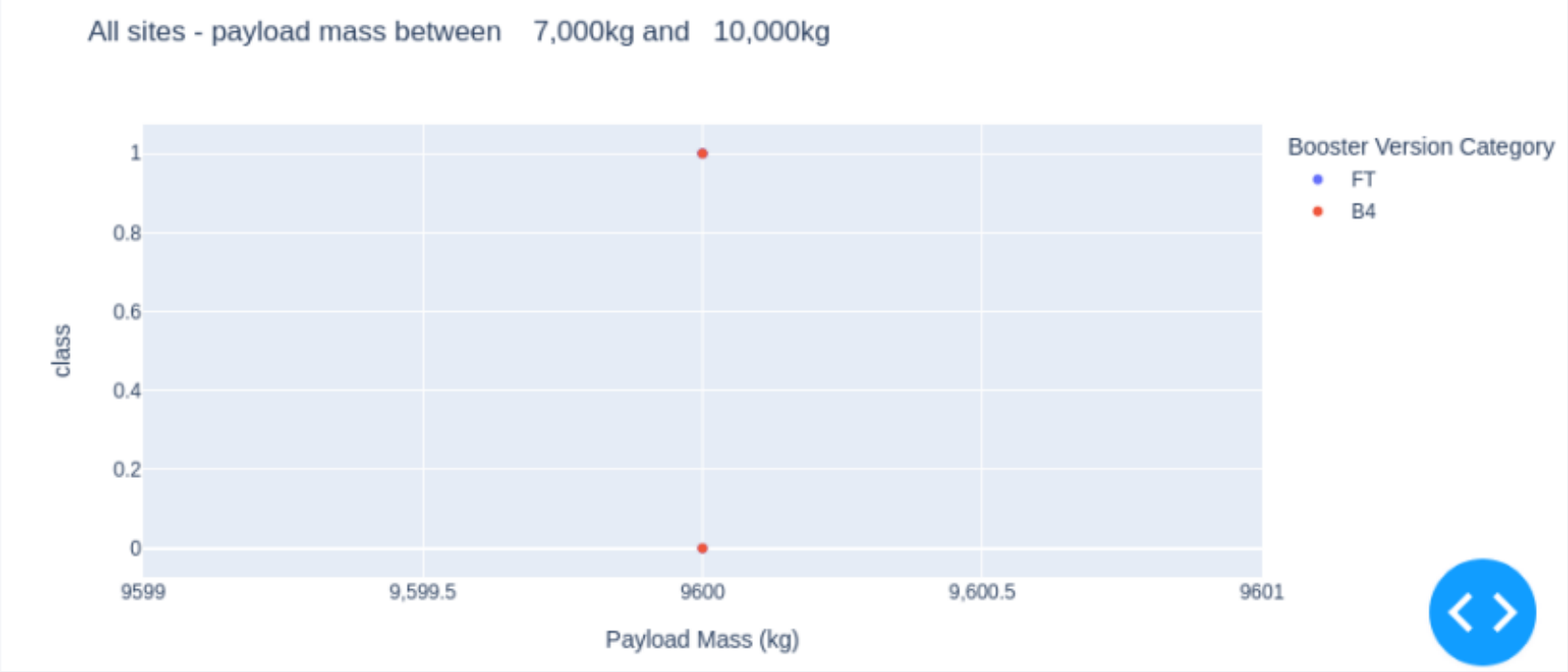
Payload vs. Launch Outcome



- The most successful combination are payloads under 6,000kg and FT boosters.

Payload vs. Launch Outcome

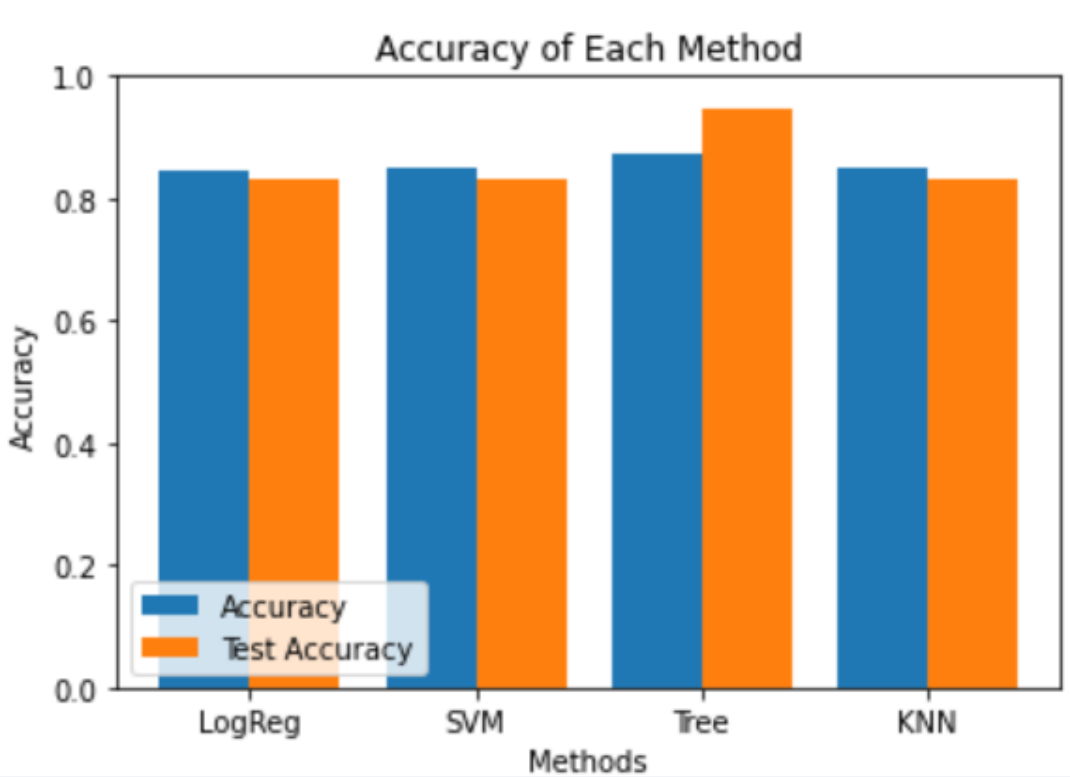
- It would be difficult to estimate risk of launches over 7,000kg because of limited available data in this range.



Section 5

Predictive Analysis (Classification)

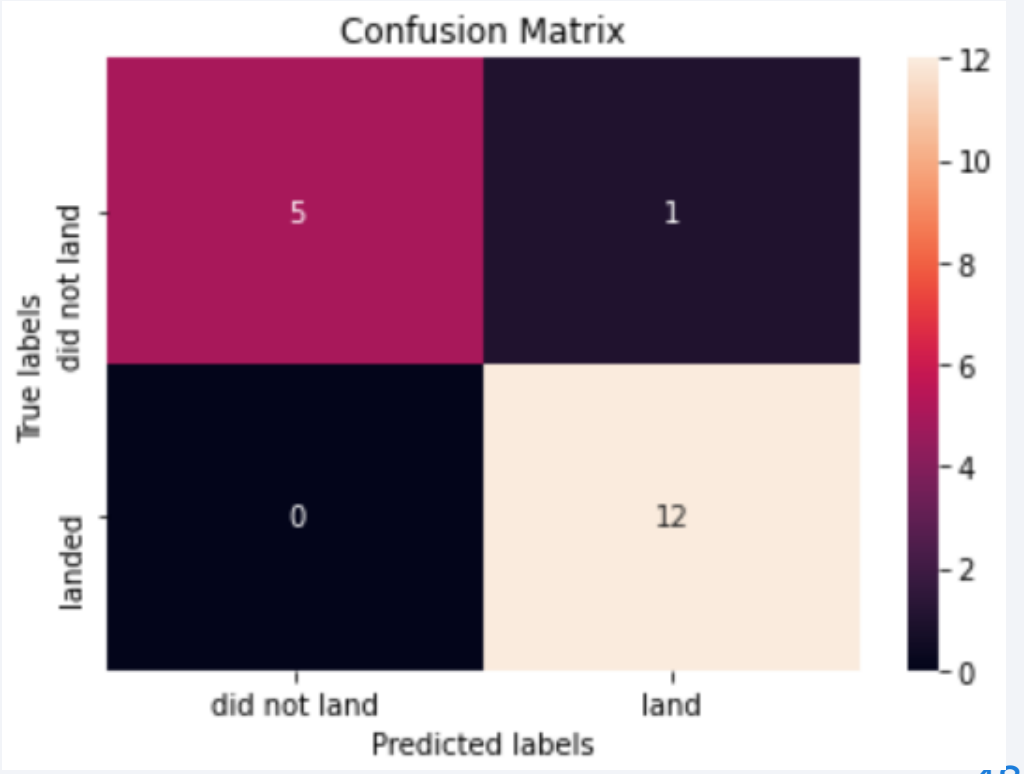
Classification Accuracy



- The bar chart represents four classification models were tested, and their accuracies;
- Decision Tree Classifier, showed the highest classification with accuracies over 87%.

Confusion Matrix

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



Conclusions

- In this study different data sources analyzed, refining conclusions along the process;
- The results showed that the best launch site is KSC LC-39A among others;
- There is less risk in Launches above 7,000kg;
- Because of the evolution of processes and rockets, most of mission outcomes are successful, and successful landing outcomes seem to improve over time.
- In prediction of successful landings, Decision Tree Classifier can be used with the highest accuracy, leading to making the most profits.

Appendix

- It is important to set a value to np.random.seed variable, to improve the model tests.

- As folium didn’t show maps on Github, I took screenshots myself.

Thank you!