IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adamu Aminu Habib
29th Nov, 2023

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection using Web Scraping and a Get Request to the SpaceX API;
  - Exploratory Data Analysis (EDA) with SQL in a Db2 database.
  - Exploratory Data Analysis (EDA) and Feature Engineering, including data wrangling, data visualization, Interactive
    visual analytics with Folium Lab and Dashboard with Ploty Dash
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collect valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - The best hyperparameters were {'criterion': 'gini', 'max_depth’: 12, ‘max_features’: ‘auto’, 'min_samples_leaf': 2, 'min_samples_split': 2}, ‘min_samples’: 10, ‘splitter’: ‘random’.
  - Resulting in an accuracy of 0.89.

Introduction

Project Background and Context: With its Falcon 9 rocket launches, SpaceX has upended the space sector for a cost of 62 million dollars, while competitors charge upwards of 165 million dollars. SpaceX's revolutionary ability to reuse its rockets' first stage is a major element in this cost savings. This ground-breaking method has transformed space flight and raised the necessity for anticipatory analysis of the pivotal first-stage landing event.

Problems you want to find answers to: Analyzing the Falcon 9 first-stage landings for predictability.Recognizing the relationship between total launch expenses and successful landings.giving rival businesses useful information so they can make well-informed bids against SpaceX.

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
  - I refined the SpaceX dataset through comprehensive data wrangling, involving steps such as handling missing data, cleaning, feature engineering, and creating a landing outcome label. This meticulous process ensures data integrity and prepares it for in-depth analysis and interpretation.
  - 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
    - The process included selecting suitable algorithms, fine-tuning parameters, and assessing model performance using metrics such as accuracy, precision, recall, and F1 score. Data collected up to this point were normalized, divided into training and test datasets, and evaluated using four distinct classification models. The accuracy of each model was assessed using various parameter combinations.

Data Collection

- Data sets were collected from Space X API using a Get Request (https://api.spacexdata.com/v4/rockets/) and

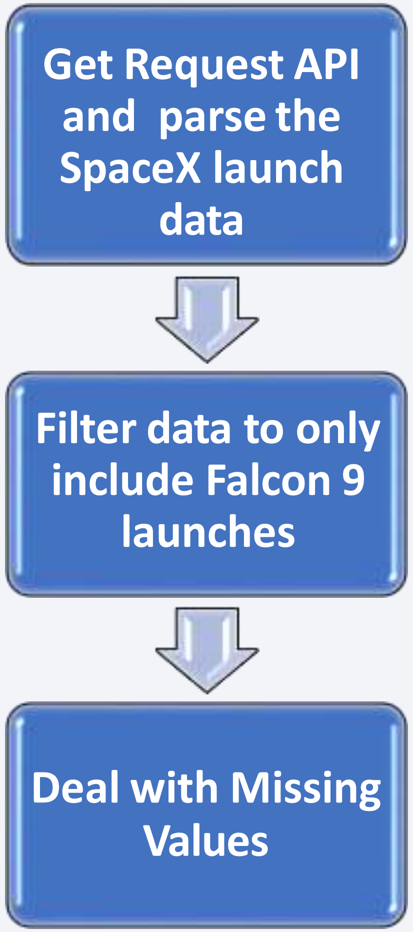
- from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technique.

Data Collection - SpaceX API

- SpaceX provides a public API from which data can be obtained and subsequently utilized;

- The API, in accordance with the accompanying flowchart, was employed to acquire data, which was then persisted.

- Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/Data%20Collection%20API.ipynb



Get Request API and parse the SpaceX launch data

Filter data to only include Falcon 9 launches

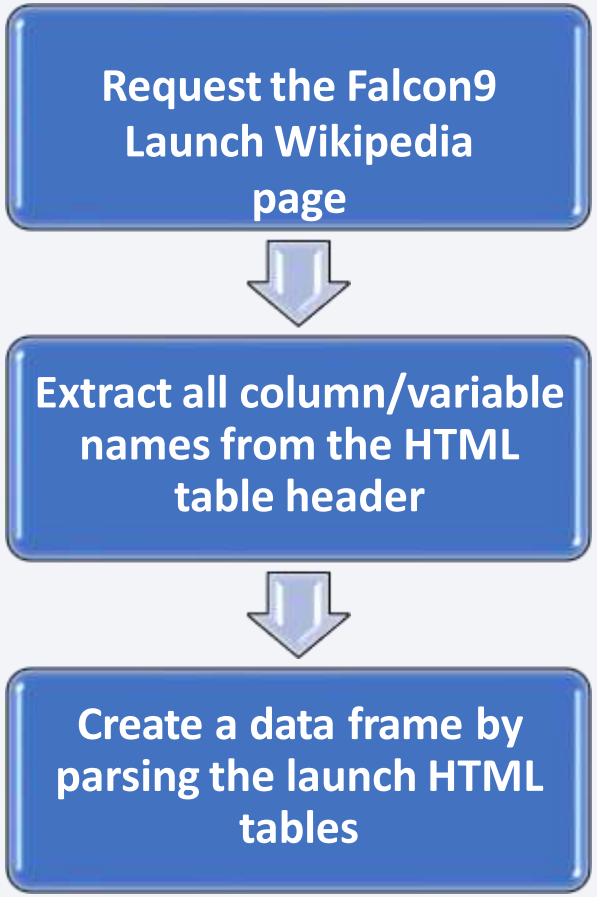
Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be sourced from Wikipedia;

- Following the outlined flowchart, data is downloaded from Wikipedia and subsequently persisted.

- Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/Data%20Collection%20with%20Web%20Scraping.ipynb



Request the Falcon9 Launch Wikipedia page

Extract all column/variable names from the HTML table header

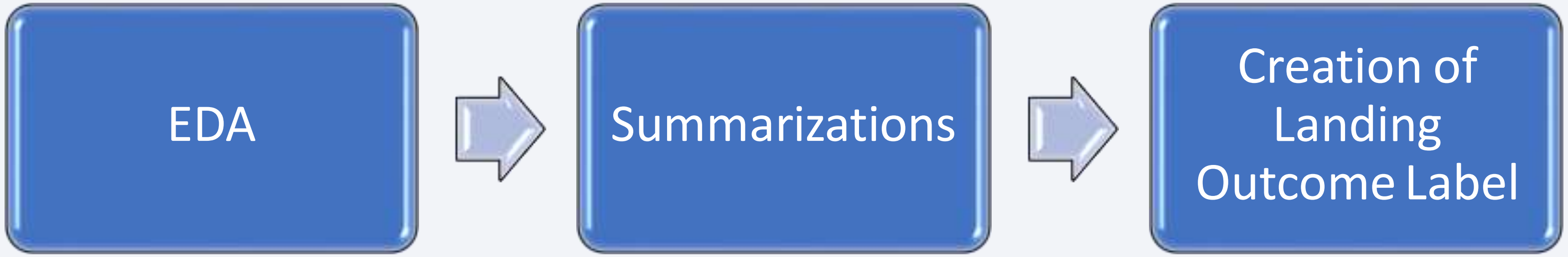
Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially, the dataset underwent Exploratory Data Analysis (EDA).

- Subsequently, calculations were made for summaries of launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type.

- Finally, the landing outcome label was generated based on the data in the Outcome column.

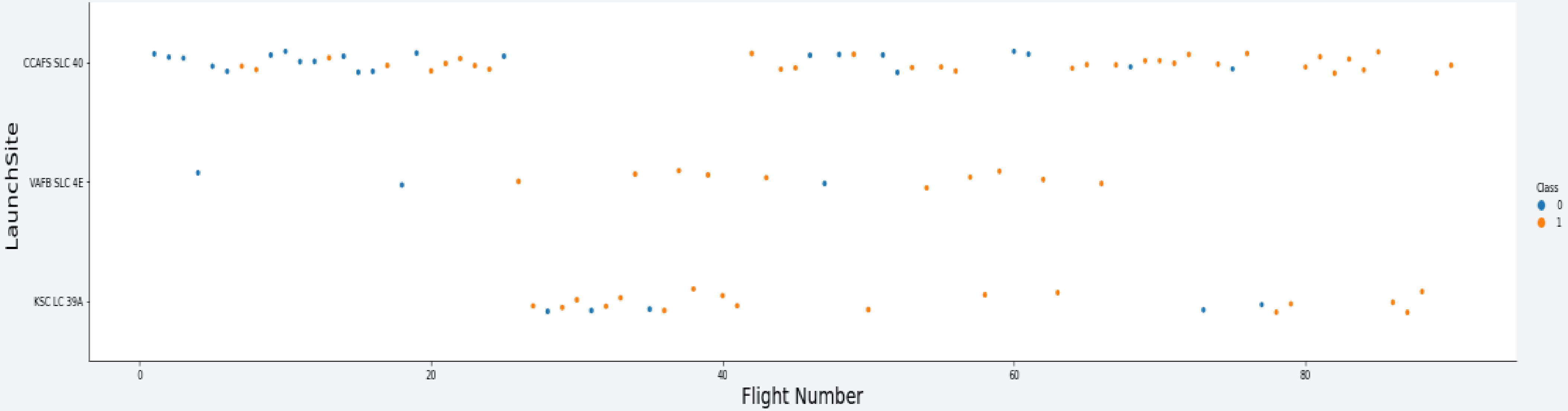


- Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/Data%20wrangling%20-%20Spacex.ipynb

EDA with Data Visualization

- Exploring the data involved utilizing scatterplots and barplots to visualize relationships between pairs 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/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/EDA%20with%20Visualization.ipynb

EDA with SQL

The following SQL queries were executed:

- Names of the unique launch sites in the space mission.
- Top 5 launch sites with names beginning 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 of the first successful landing outcome on a ground pad.
- Names of boosters with success in drone ship and payload mass between 4000 and 6000 kg.
- Total number of successful and failed mission outcomes.
- Names of booster versions carrying the maximum payload mass.
- Failed landing outcomes in a drone ship, including booster versions and launch site names in the year 2015.
- Rank of the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20.

Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/EDA%20with%20SQL.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.
  - Lines are used to indicate distances between two coordinates.

- Source code:

https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

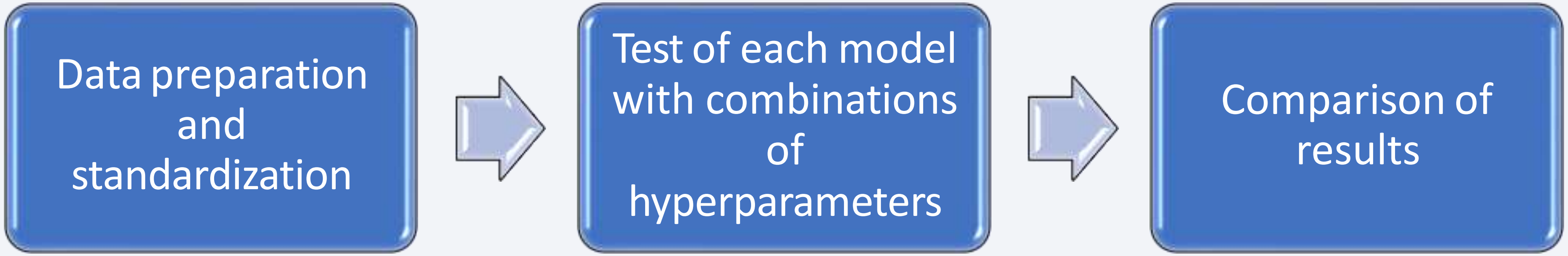
- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range

- This combination facilitated a swift analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch locations based on payload considerations.

- Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/spacex_dash_app.py

Predictive Analysis (Classification)

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



Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

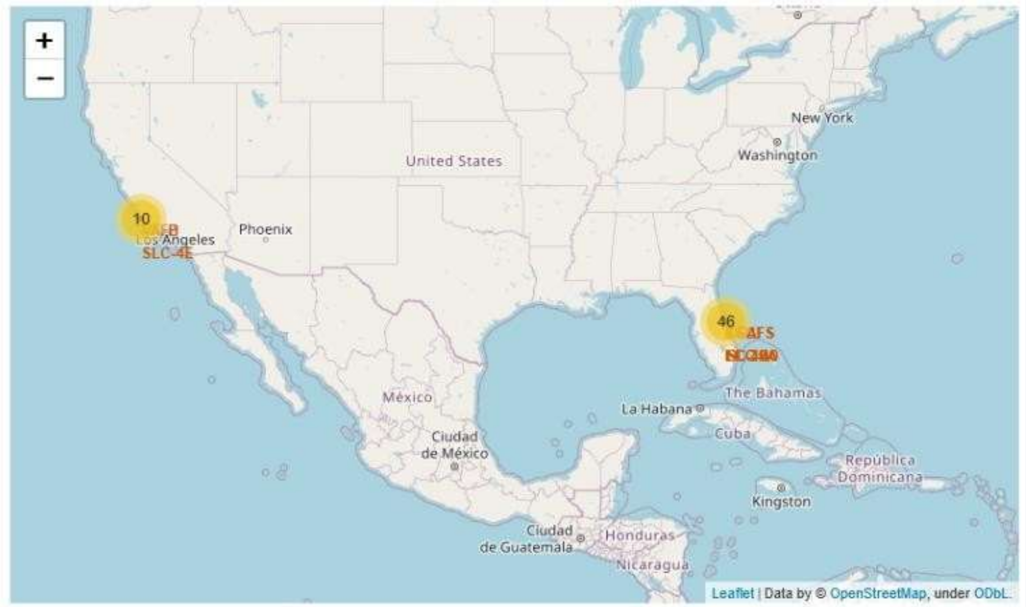
- Source code: https://github.com/HabibAnalyticsPro/My_Applied_Data_Science_Capstone_Projects/blob/master/applied-data-science-capstone-master/Machine_Learning_Prediction_Part_5_SpaceX.ipynb

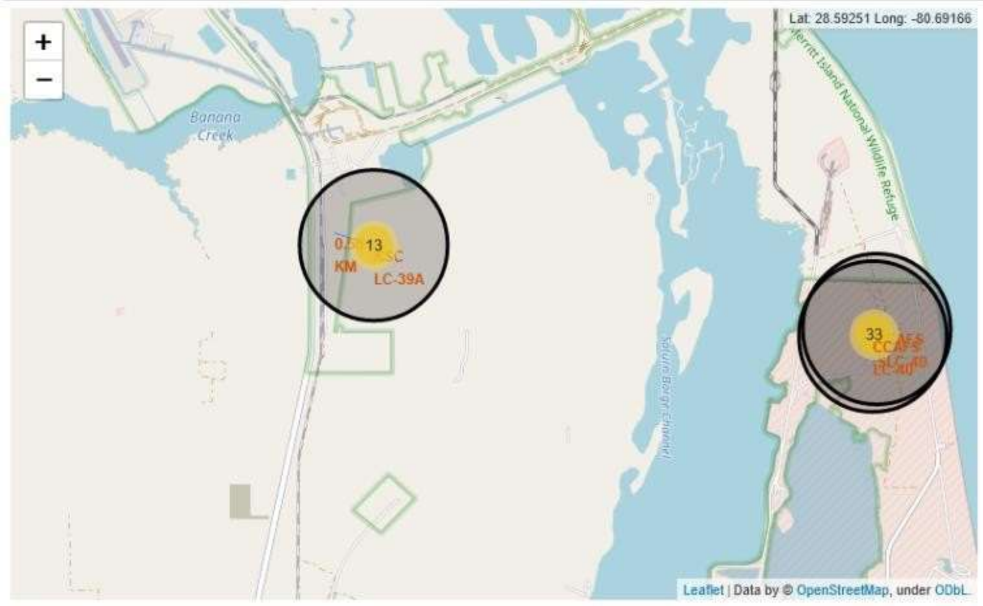
Results

- 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

- By employing interactive analytics, it was feasible to discern that launch sites tend to be located in secure areas, often in proximity to the sea, ensuring safety, and boasting robust logistic infrastructure. Additionally, a majority of launches occurs at launch sites situated on the east coast.

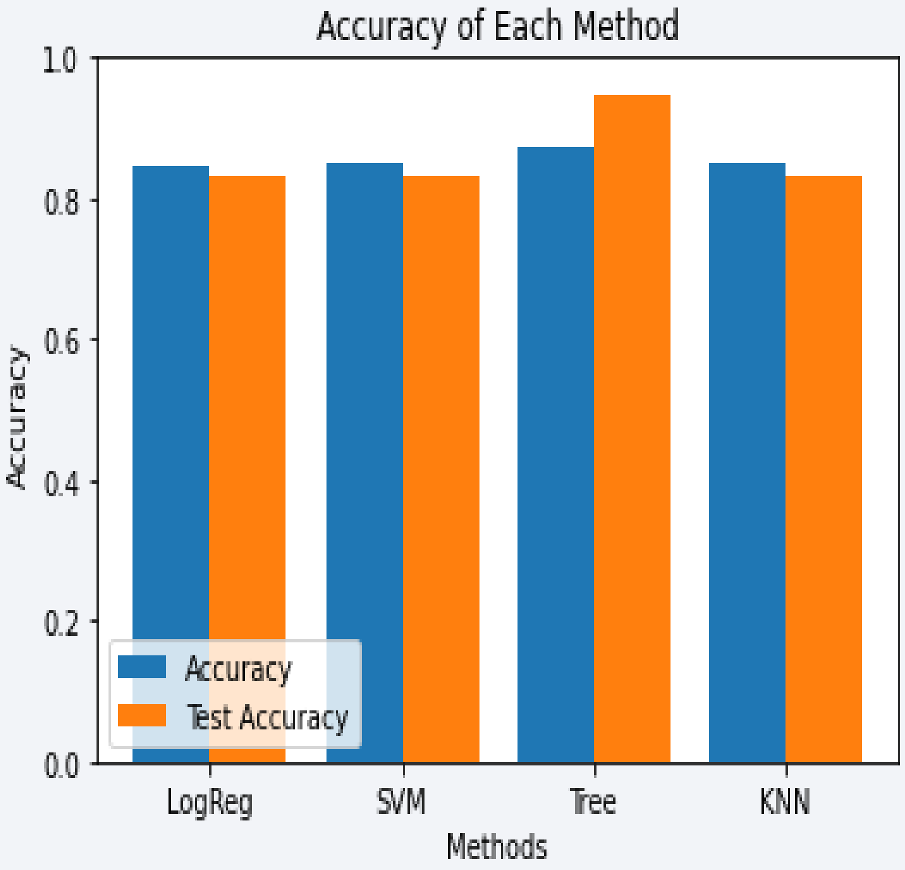




+
−
Lat: 28.59251 Long: -80.69166
Banana Creek
Merritt Island National Wildlife Refuge
13
KM
LC-39A
33
CCAFS
Leaflet | Data by © OpenStreetMap, under ODbL.

Results

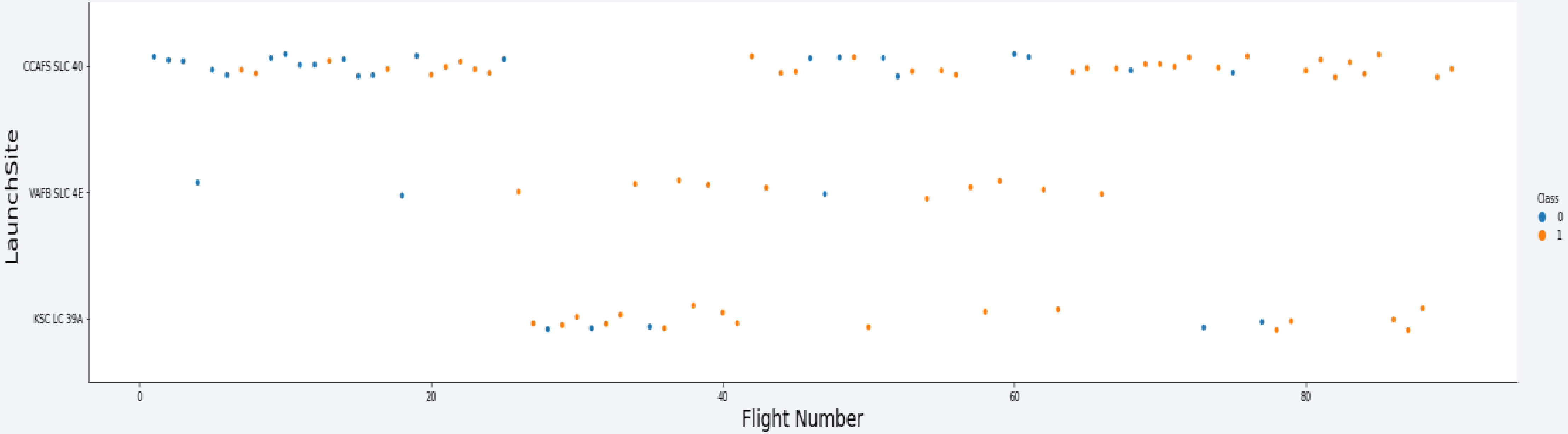
- Predictive analysis revealed that the Decision Tree Classifier emerged as the most effective model for predicting successful landings, boasting an accuracy exceeding 87%, with a test data accuracy surpassing 94%.



Section 2

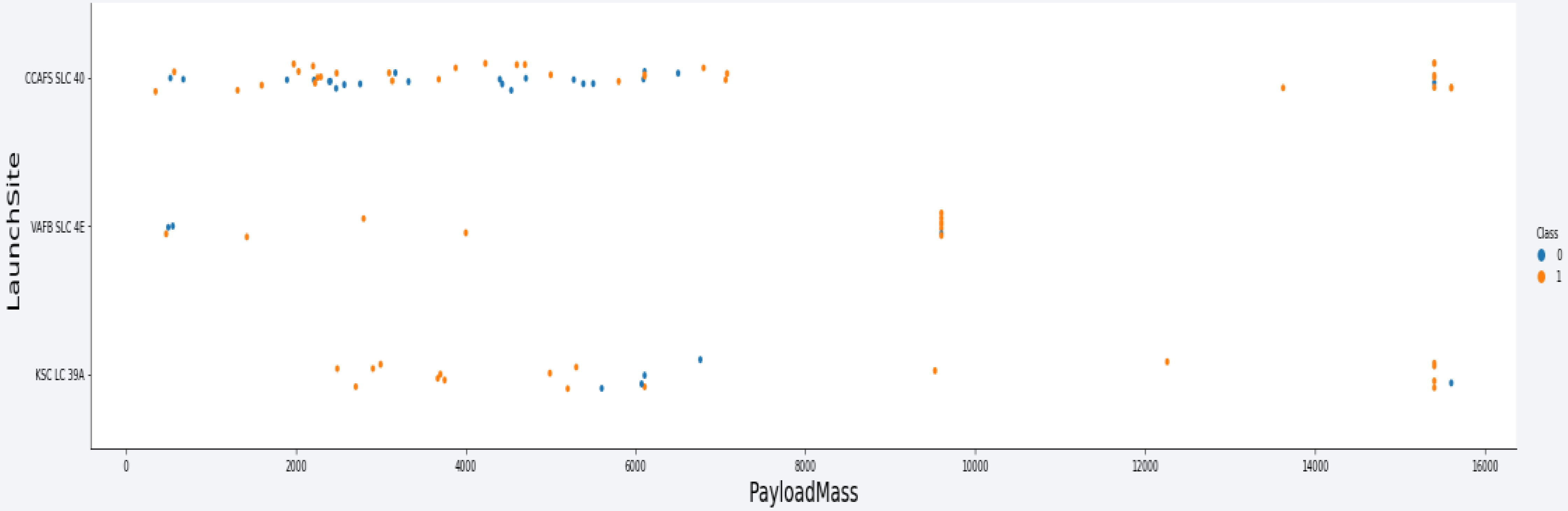
Insights drawn from EDA

Flight Number vs. Launch Site



- Based on the depicted plot, it's evident that the current optimal launch site is CCAF5 SLC 40, showcasing a high success rate in recent launches.
- Following closely, VAFB SLC 4E holds the second position, and KSC LC 39A secures the third spot.
- Additionally, the plot illustrates an overall improvement in the success rate over time.

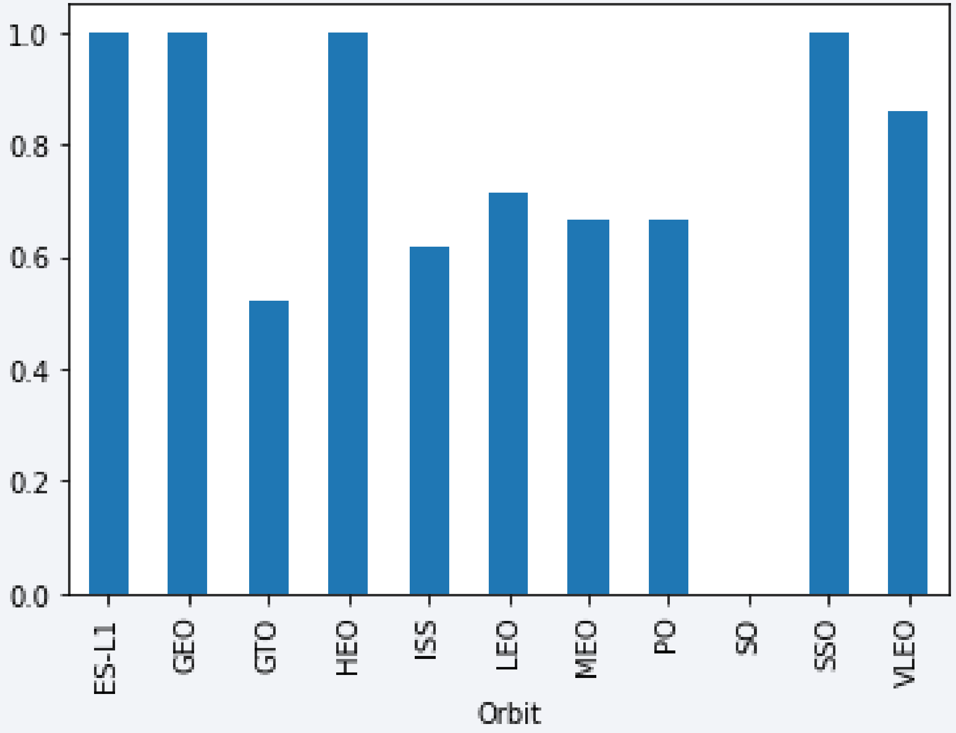
Payload vs. Launch Site



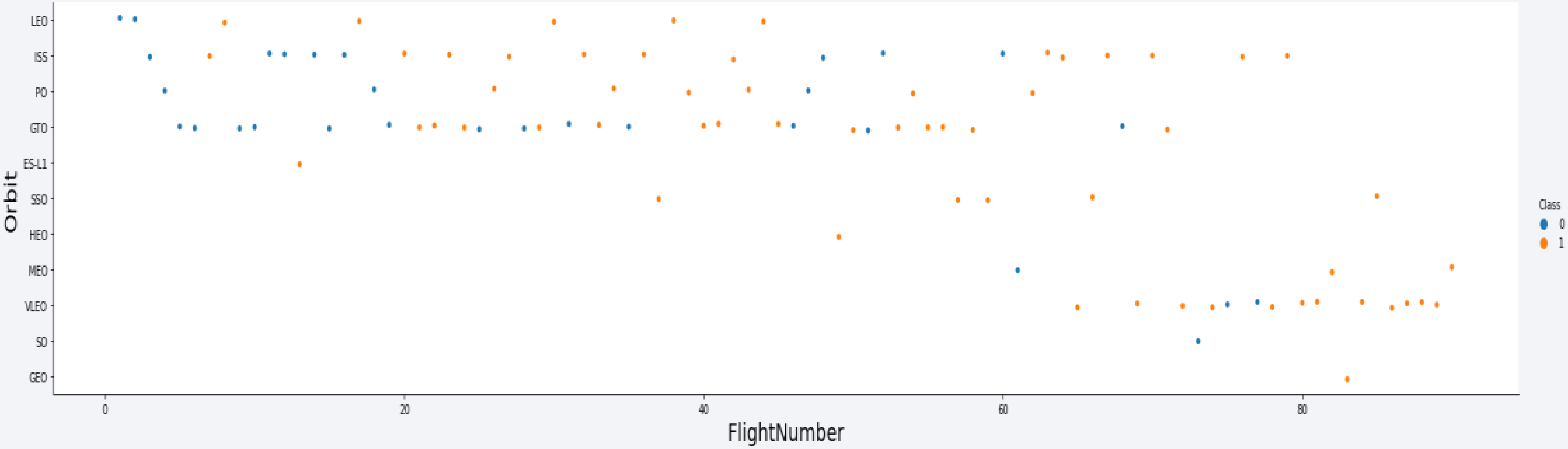
- Payloads exceeding 9,000kg, approximately the weight of a school bus, exhibit an outstanding success rate. Notably, payloads surpassing 12,000kg appear viable primarily at CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (above 80%), and
  - LFO (above 70%).

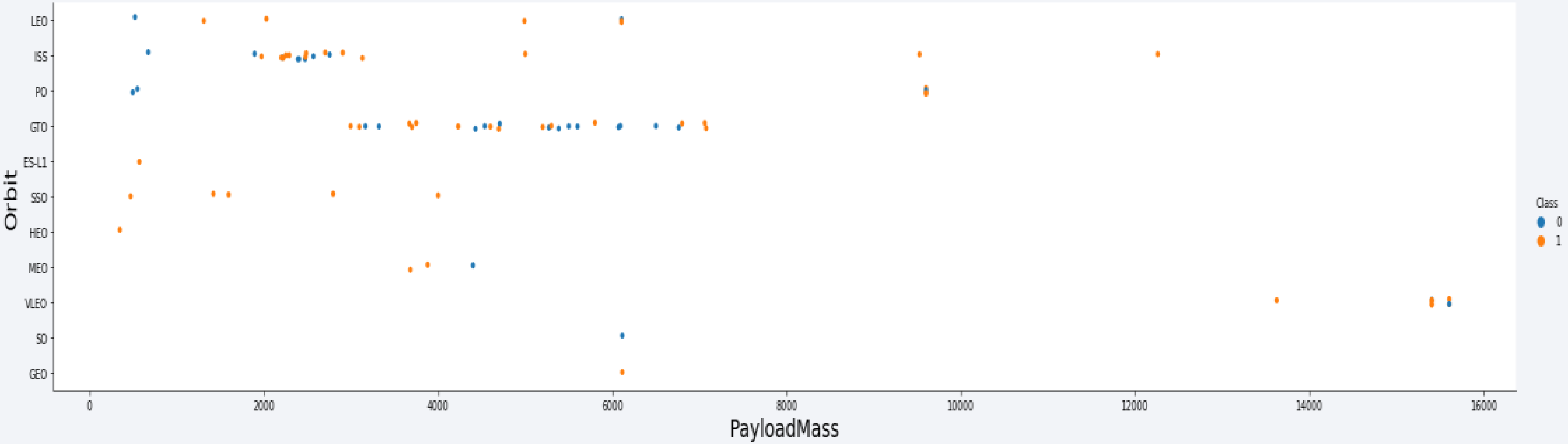


Flight Number vs. Orbit Type



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

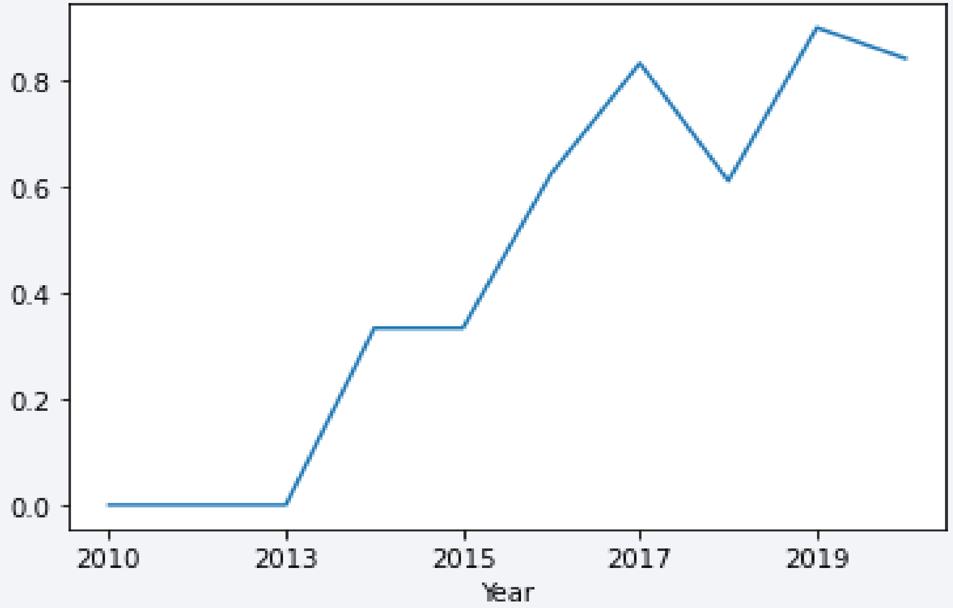
Payload vs. Orbit Type



- Apparently, 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

- The success rate initiated an upward trend in 2013 and continued until 2020.
- The initial three years appear to have served as a period of adjustments and technological improvements, leading to the subsequent increase in success rates.



All Launch Site Names

- According to data, there are four launch sites:

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

- 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 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

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

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- The total payload, as calculated above, is obtained by summing all payloads whose codes contain 'CRS,' corresponding to NASA.

Average Payload Mass by F9 v1.1

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

Avg Payload (kg)
2.928

- 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:

Min Date
2015-12-22

- By filtering the data for successful landing outcomes on the ground pad and extracting the minimum date value, it is possible to pinpoint the first occurrence, which took place 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

- 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:

Mission Outcome	Occurrences
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

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

- These are the boosters which 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

- The list above has the only two 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:

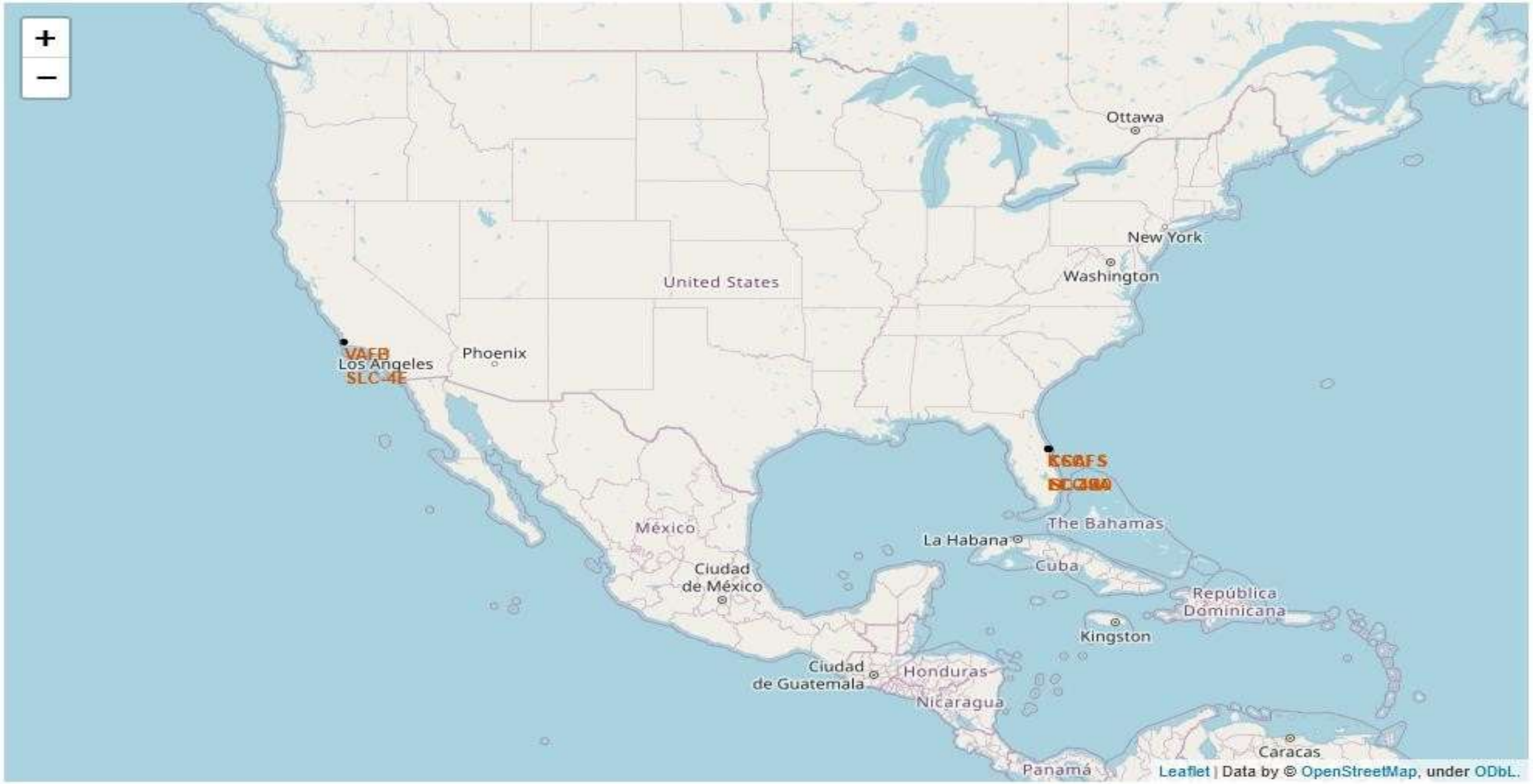
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- This data perspective emphasizes the significance of considering "No attempt" as a relevant category in our analysis.

Section 4

Launch Sites Proximities Analysis

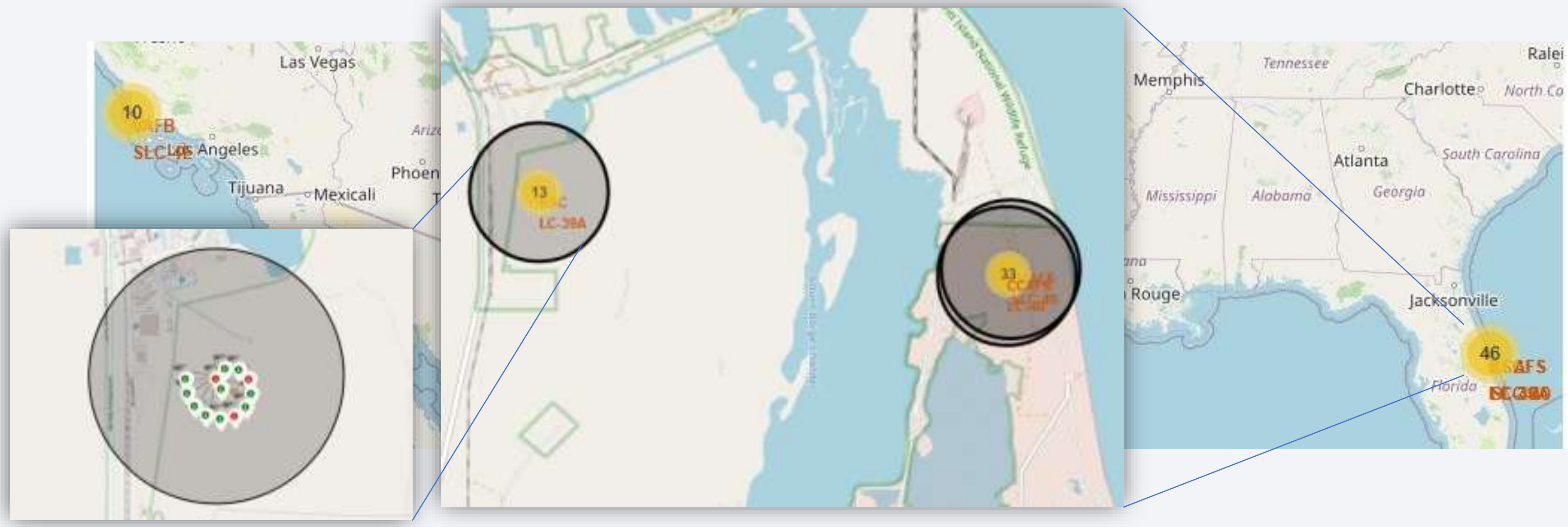
All launch sites



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

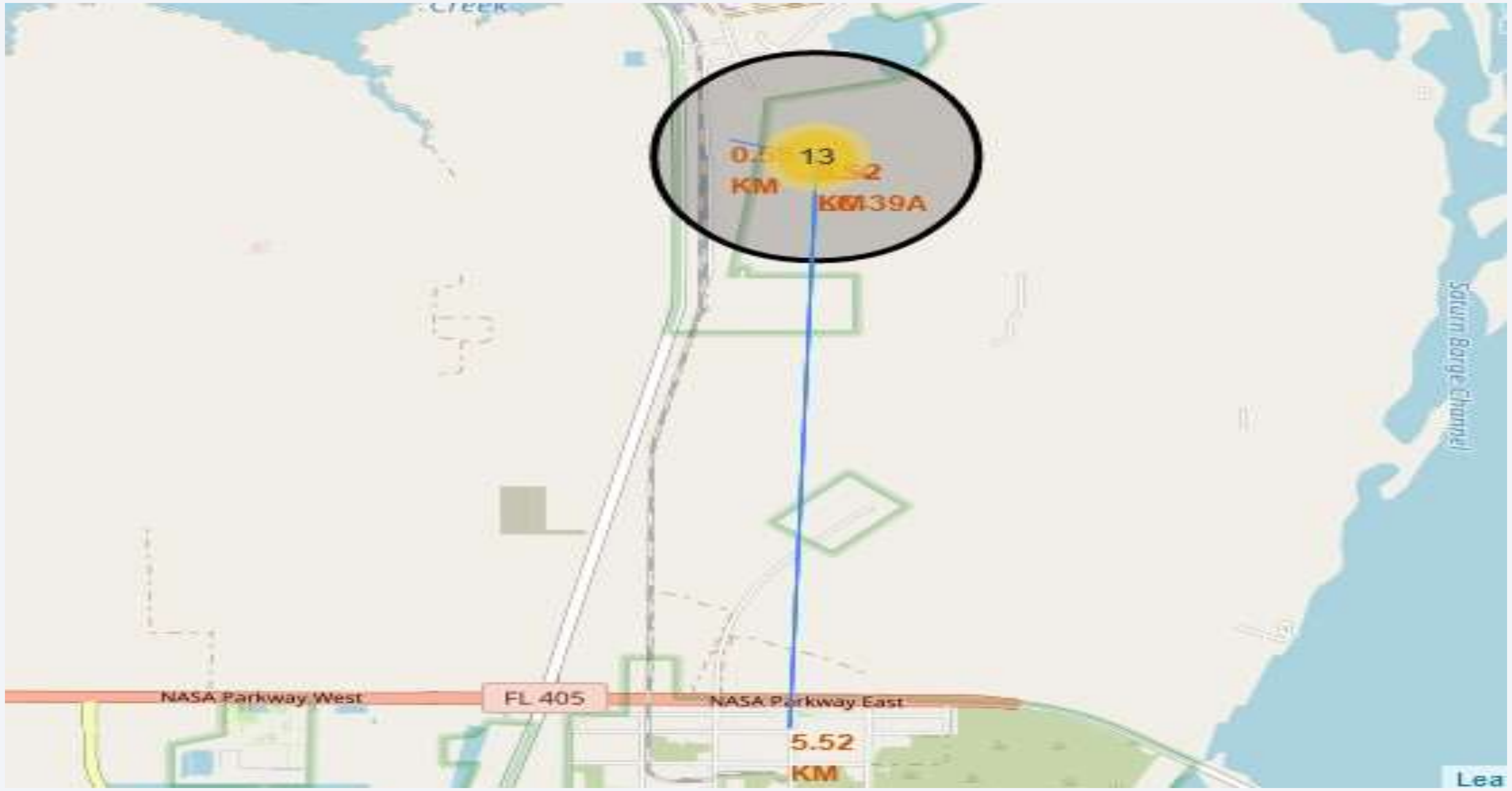
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes



- Green markers indicate successful and red ones indicate failure.

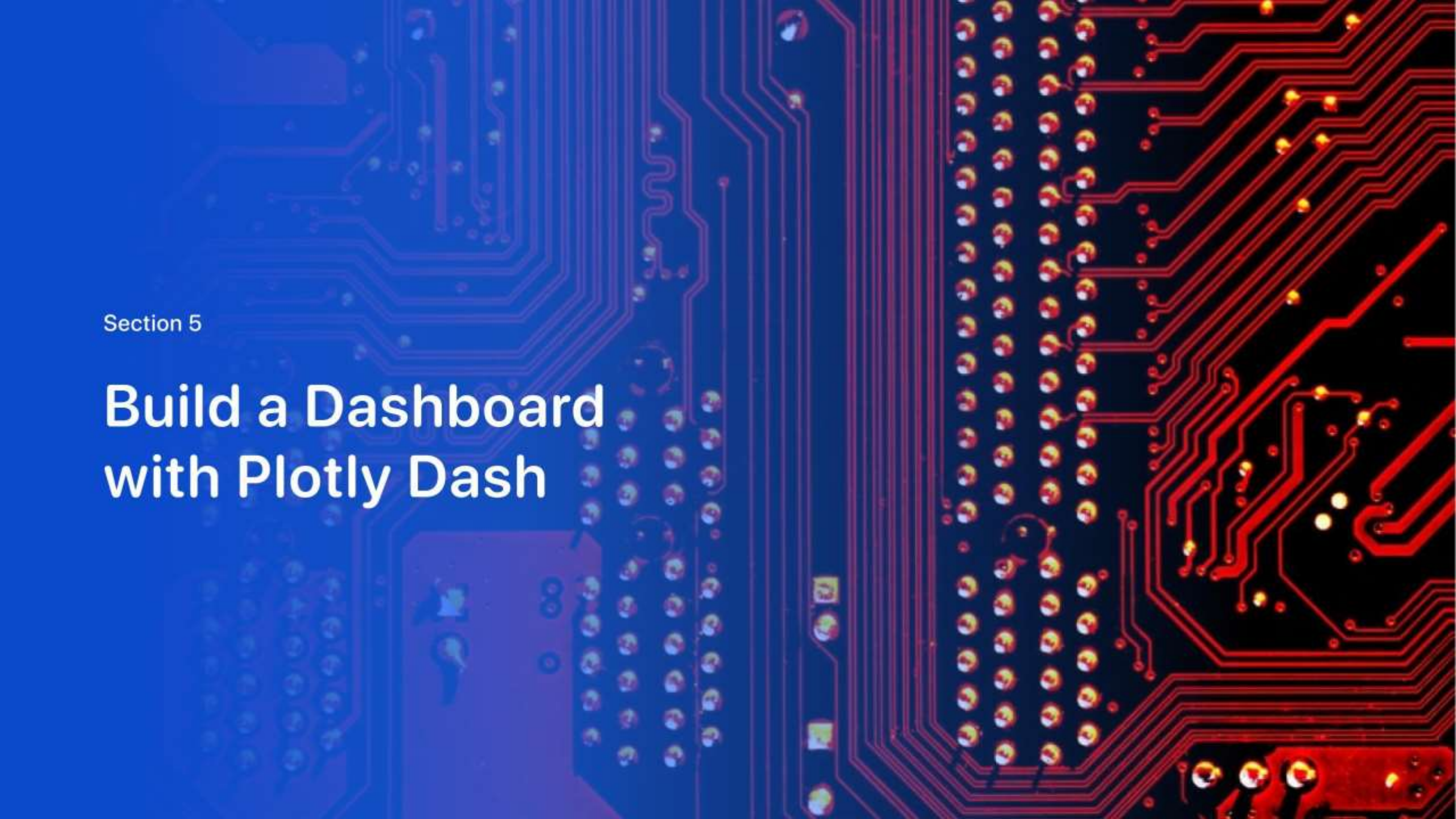
Logistics and Safety



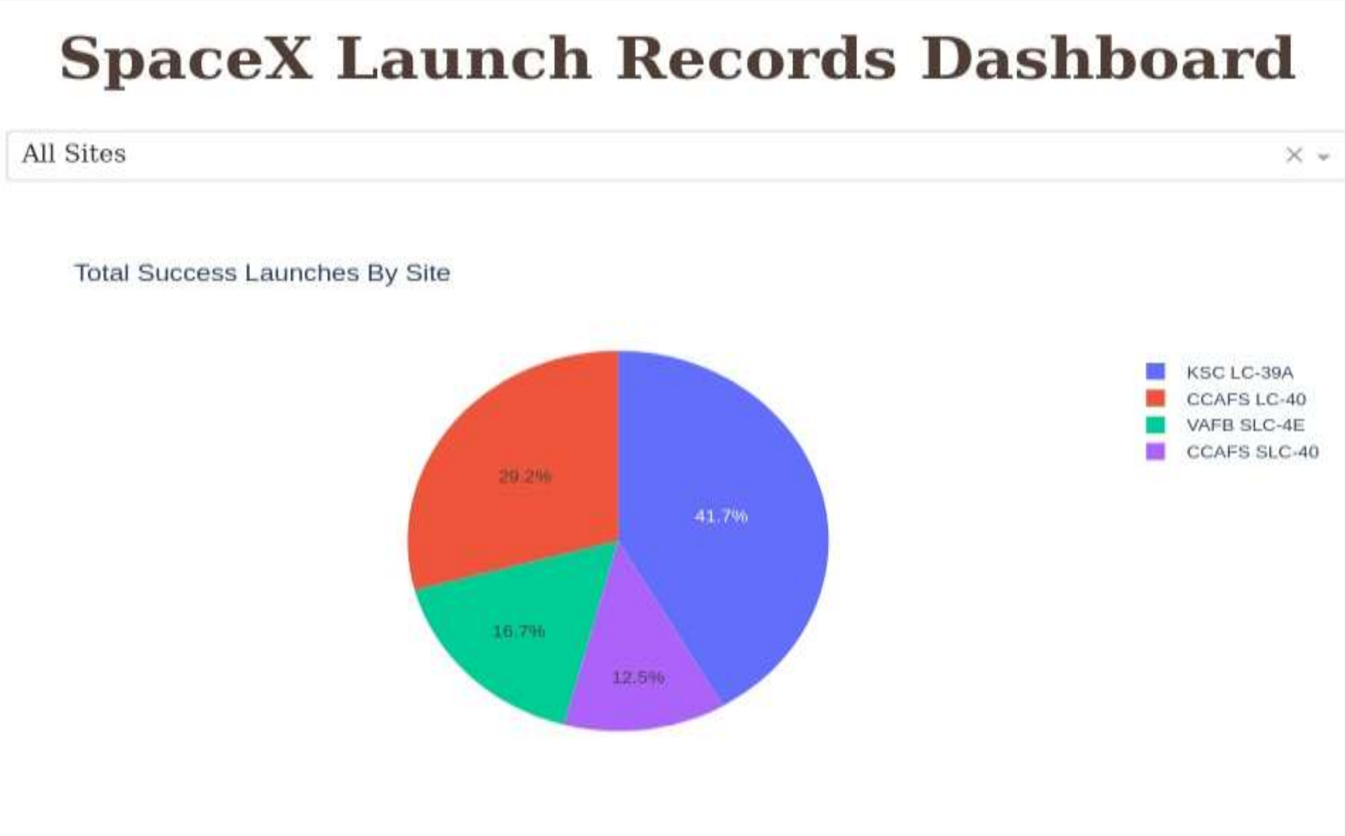
- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 5

Build a Dashboard with Plotly Dash



Successful Launches by Site

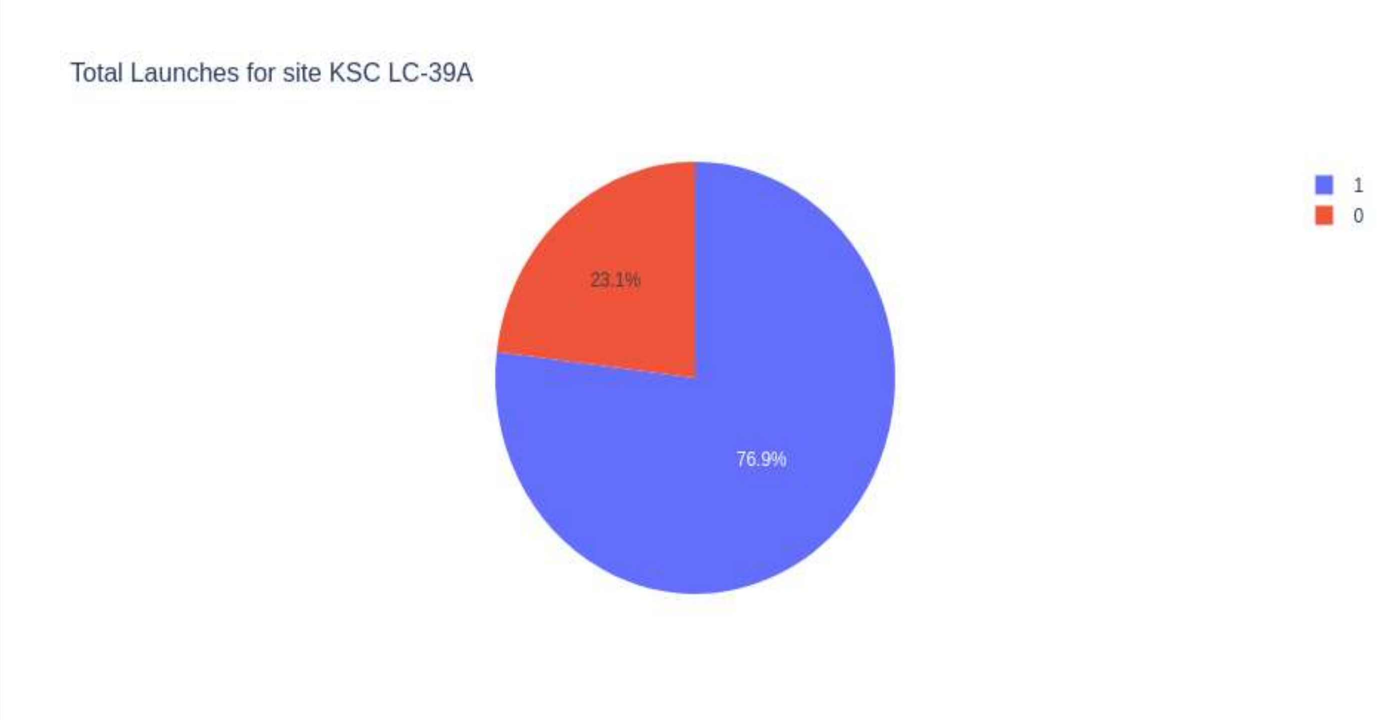


SpaceX Launch Records Dashboard

All Sites

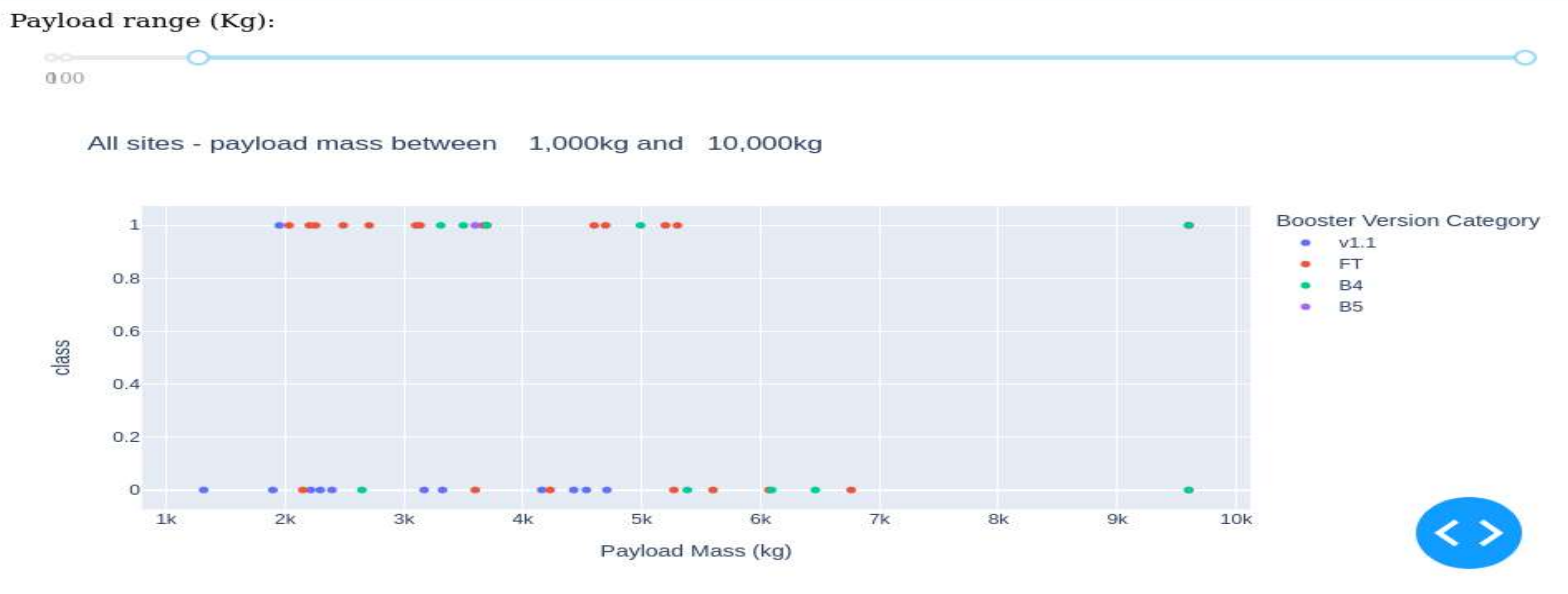
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



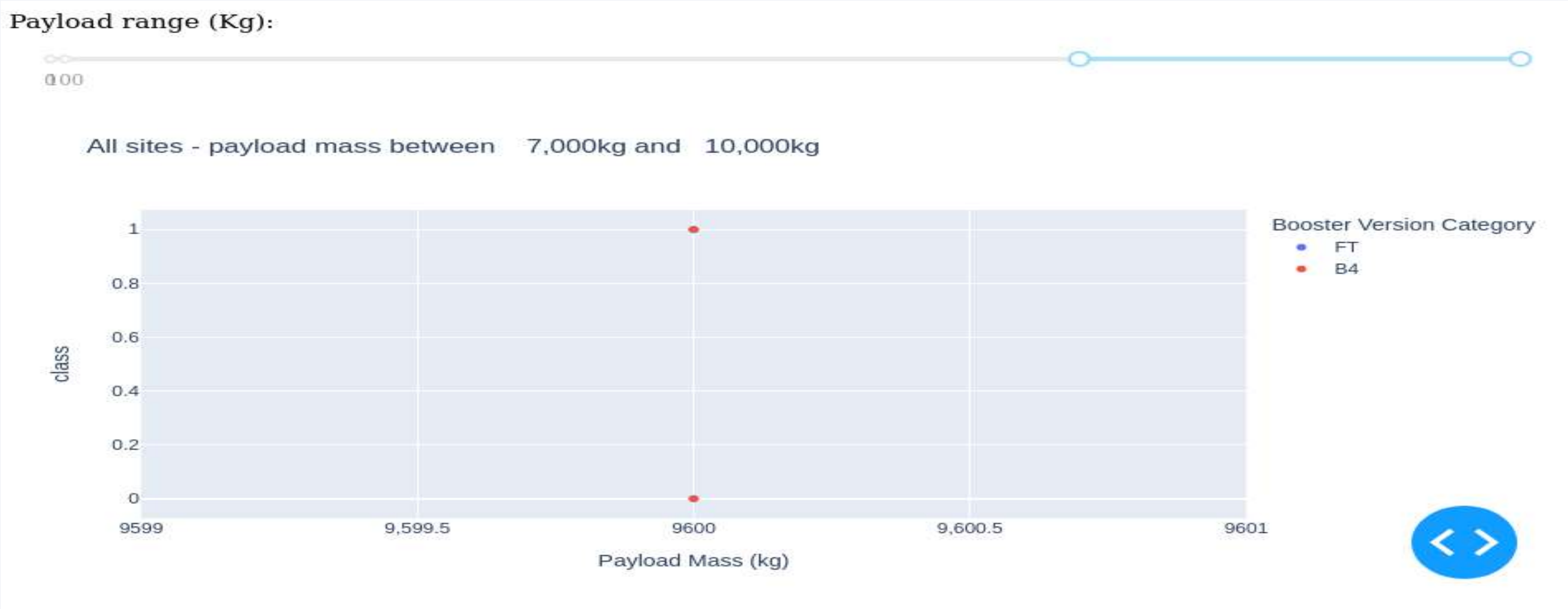
- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



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

Payload vs. Launch Outcome



Payload range (Kg):

All sites - payload mass between 7,000kg and 10,000kg

- There’s not enough data to estimate risk of launches over 7,000kg

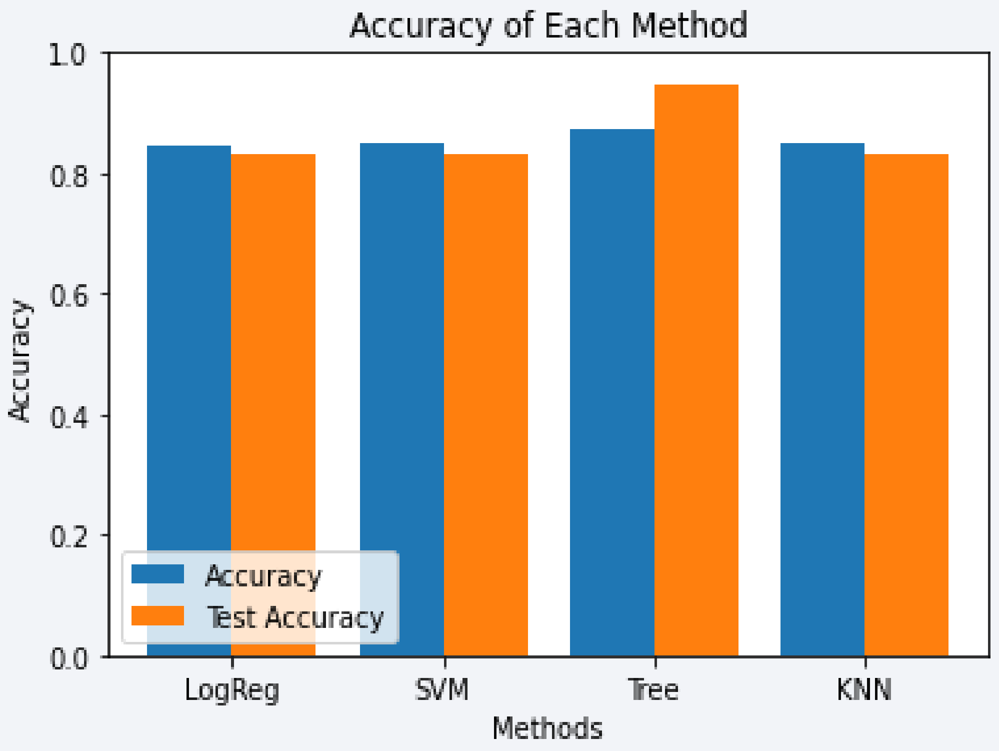
Section 6

Predictive Analysis (Classification)

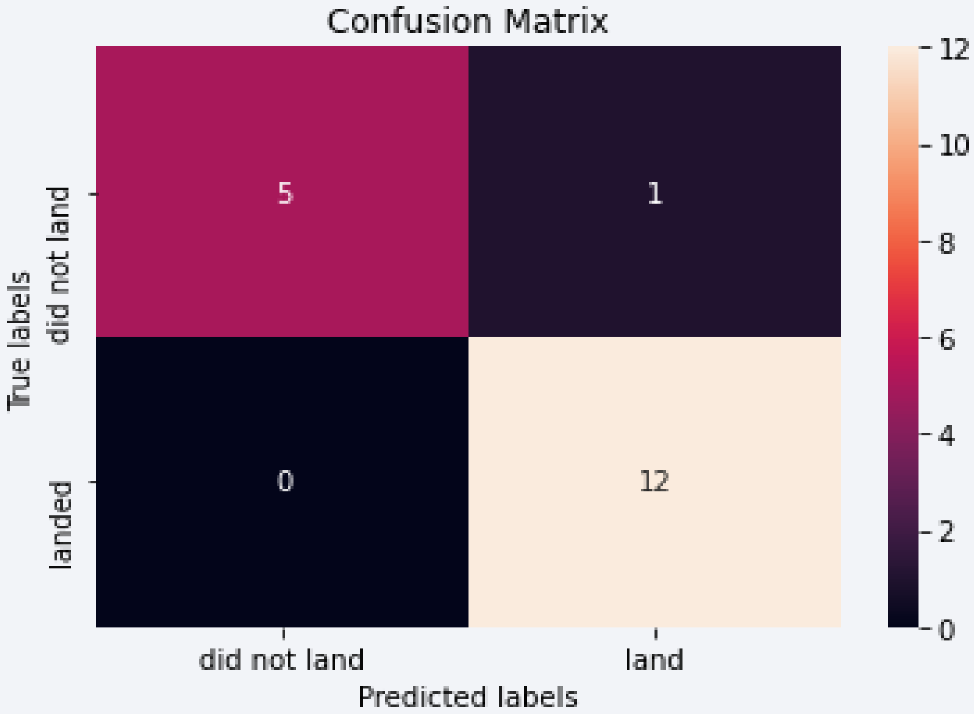
Classification Accuracy

- 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 of Decision Tree Classifier



- 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

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix 1

- **Data Source:**
  - SpaceX API: Acquired real-time and historical data directly from the SpaceX API, ensuring accuracy and timeliness.
  - Wikipedia (Web Scraping): Extracted supplementary information from Wikipedia through web scraping techniques, enriching the dataset.

- **Technical Details:**
  - Data Cleaning: Implemented thorough data cleaning procedures to handle missing values, inconsistencies, and outliers.
  - Feature Engineering: Generated new features and modified existing ones to enhance the predictive power of the dataset.
  - Normalization and Scaling: Ensured uniformity by applying normalization and scaling techniques to numerical features.
  - Model Selection: Find best Hyperparameter for SVM, Classification Trees and Logistic Regression.
  - Evaluation Metrics: Utilized accuracy, precision, recall, and F1 score for comprehensive model performance assessment.

Appendix 2

Model Parameters:

Logistic Regression:

Parameter 1: [GridSearchCV(cv=10, estimator=LogisticRegression() param_grid={'C': [0.01, 0.1, 1], 'penalty': ['l2'],
'solver': ['lbfgs']})]

K Nearest Neighbor:

Parameter 1: [GridSearchCV(cv=10, estimator=KNeighborsClassifier() param_grid={'algorithm': ['auto', 'ball_tree',
'kd_tree', 'brute’], 'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'p': [1, 2]})]

Support Vector Machine:

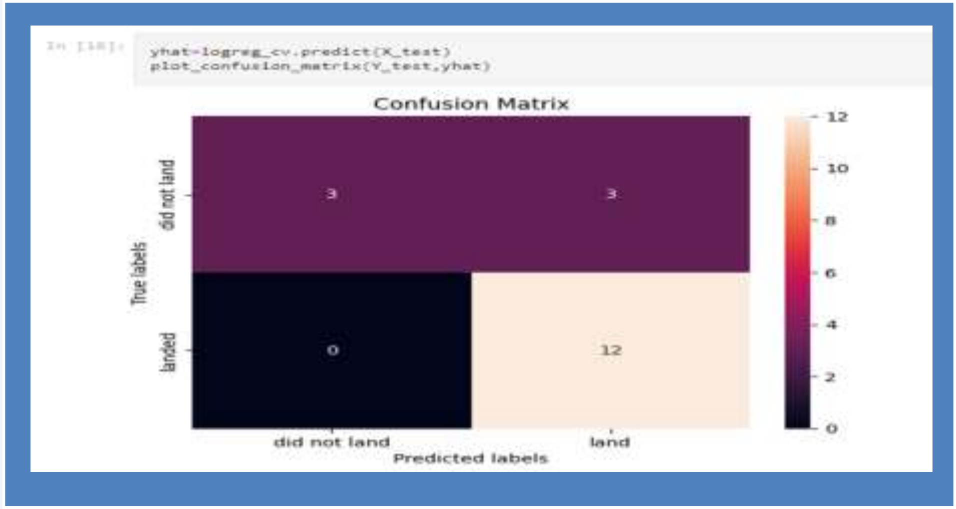
Parameter 1: [GridSearchCV(cv=10, estimator=SVC(),
param_grid={'C': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
1.00000000e+03]),
'gamma': array([1.00000000e-03, 3.16227766e-02, 1.00000000e+00, 3.16227766e+01,
1.00000000e+03]),
'kernel': ('linear', 'rbf', 'poly', 'rbf', 'sigmoid')})]

Decision Tree:

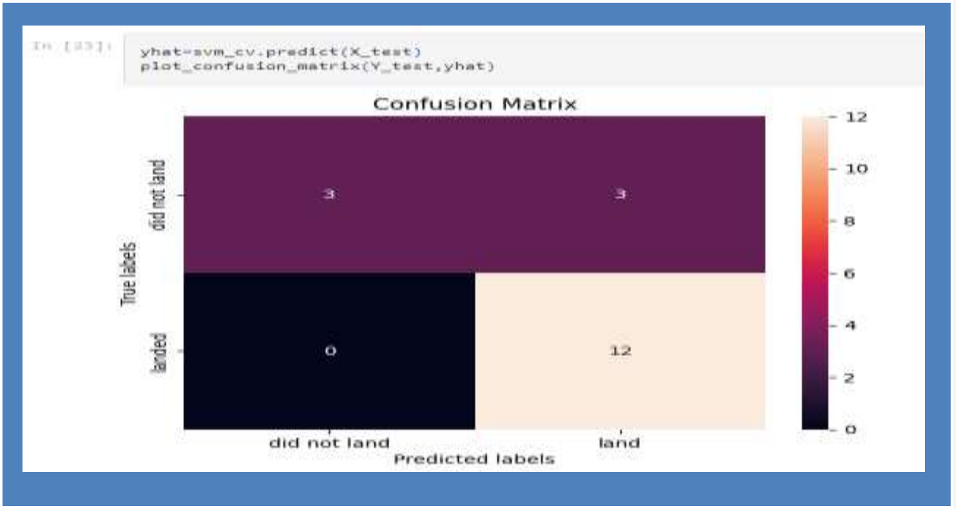
Parameter 1: [GridSearchCV(cv=10, estimator=DecisionTreeClassifier(), param_grid={'criterion': ['gini', 'entropy'],
'max_depth': [2, 4, 6, 8, 10, 12, 14, 16, 18],
'max_features': ['auto', 'sqrt'],
'min_samples_leaf': [1, 2, 4],
'min_samples_split': [2, 5, 10],
'splitter': ['best', 'random']})]

Appendix 3

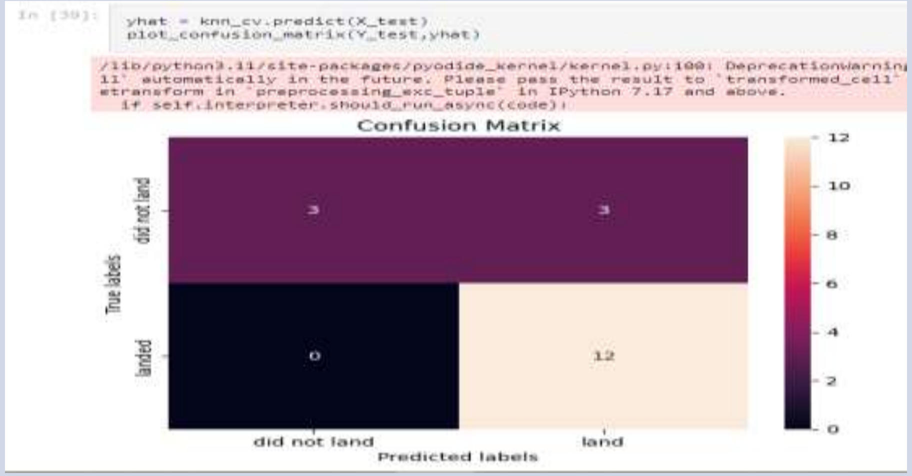
Additional Visuals:



```
In [18]: yhat=logreg_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```

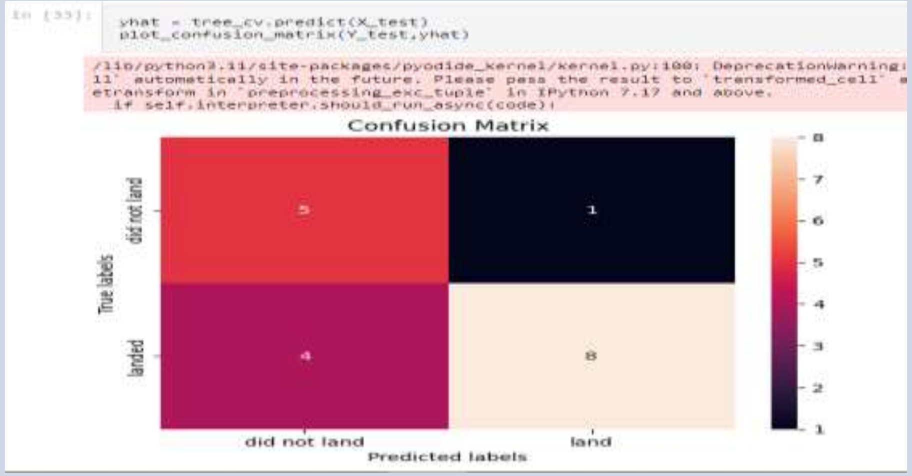


```
In [23]: yhat=svm_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



```
In [30]: yhat = knn_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)

/lib/python3.11/site-packages/pyodide_kernel/kernel.py:100: DeprecationWarning
11` automatically in the future. Please pass the result to `transformed_cell`
etransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  if self.interpreter.should_run_async(code):
```



```
In [33]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)

/lib/python3.11/site-packages/pyodide_kernel/kernel.py:100: DeprecationWarning:
11` automatically in the future. Please pass the result to `transformed_cell` a
etransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  if self.interpreter.should_run_async(code):
```

Thank you!