IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eduardo Emmanuel Rodríguez López
04/14/2024

Outline

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

Executive Summary

Summary of Methodologies.

- The main objective of this project is to predict whether the first stage of Space X's Falcon 9 will land successfully. If the first stage lands successfully, it can be reused on new launches, which reduces SpaceX's costs. Thus, in this project, a number of methodologies were applied in order to address this goal. These were:
    - Data Collection with the SpaceX API and Web Scraping
    - Data Wrangling
    - Exploratory Data Analysis (EDA): SQL queries and Data Visualization.
    - Interactive Visual Analysis, and
    - Predictive Analytics (Classification).

Executive Summary

Summary of Results.

- A number of Machine Learning models were tested (Logistic Regression, SVM, Decision Tree and KNN), with very good results, but very similar to each other for the test set. Given the above, just to find more significant differences, the complete dataset was used, with the SVM standing out as the best model for predicting whether the first stage landed successfully.

Introduction

- Space Y is a company that wants to compete against Space X. For this it wants to know the costs that Space X has in its rocket launches. These are greatly influenced by the fact of whether its first stage landed successfully or not, because if it did, it can be reused in new launches, and thus drastically reduce costs.

- Therefore, it is desired to know:
  - The impact that some variables have on the successful landing of the fisrt stage (and hence its cost).
  - The best launch sites.
  - Which of them has the highest first stage landing percentage.

Section 1

Methodology

Methodology

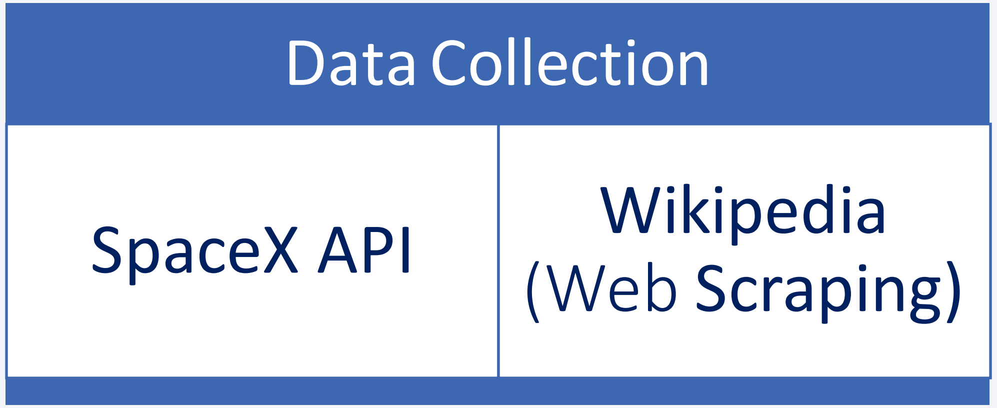
Executive Summary

- Data collection methodology:
  - The data were collected using two sources, SpaceX API, and Wikipedia (using Web Scraping)
- Perform data wrangling
  - The variables in the dataset were reviewed, for example its type or the amount of null data per column. The number of launches per site was examined and, finally, a landing outcome label was created from the Outcome column.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Finally, with the data processed up to this point, four different Machine Learning models were trained and tested: Logistic Regression, SVM, Decision Tree and KNN.

Data Collection

Data were collected mainly from two sources:

- SpaceX API (unofficial: GitHub - r-spacex/SpaceX-API: :rocket: Open Source REST API for SpaceX launch, rocket, core, capsule, starlink, launchpad, and landing pad data.).
- Wikipedia's List of Falcon 9 and Falcon Heavy launches page, using Web Scrapping. (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)



Data Collection

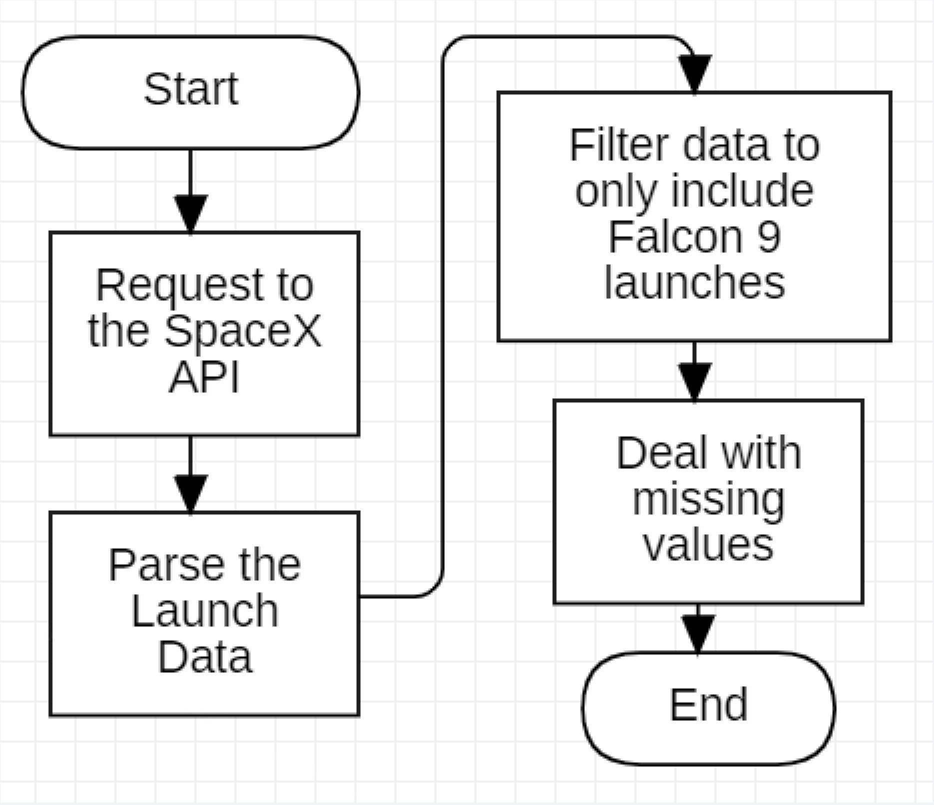
SpaceX API

Wikipedia
(Web Scraping)

Data Collection – SpaceX API

- The SpaceX API (unofficial) was used to obtain relevant launch data. The process that was followed is represented in the flowchart on the right.

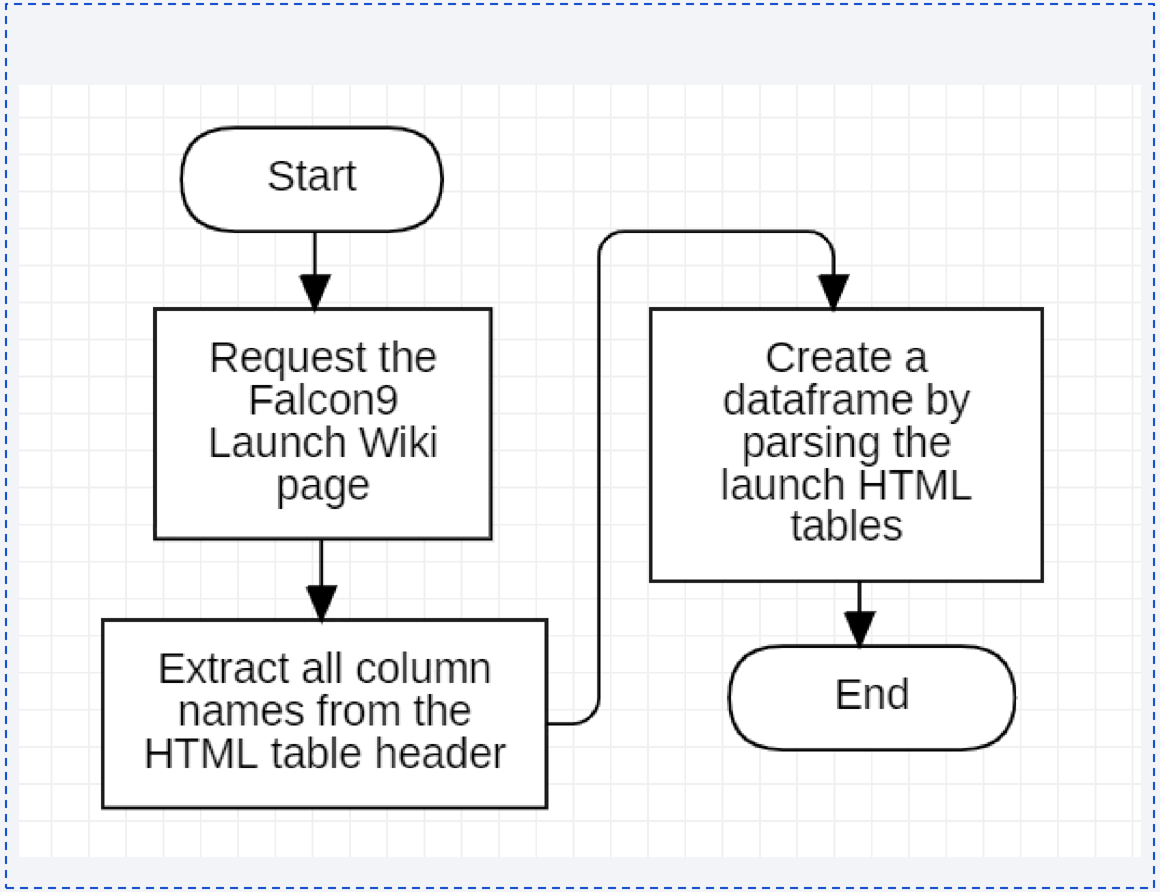
- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb


Start
Request to the SpaceX API
Parse the Launch Data
Filter data to only include Falcon 9 launches
Deal with missing values
End

Data Collection - Scraping

- Web Scraping was applied on the Wikipedia page of the Falcon 9 and Falcon Heavy launches Records. Please refer to the flowchart on the right.

- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Start

Request the Falcon9 Launch Wiki page

Extract all column names from the HTML table header

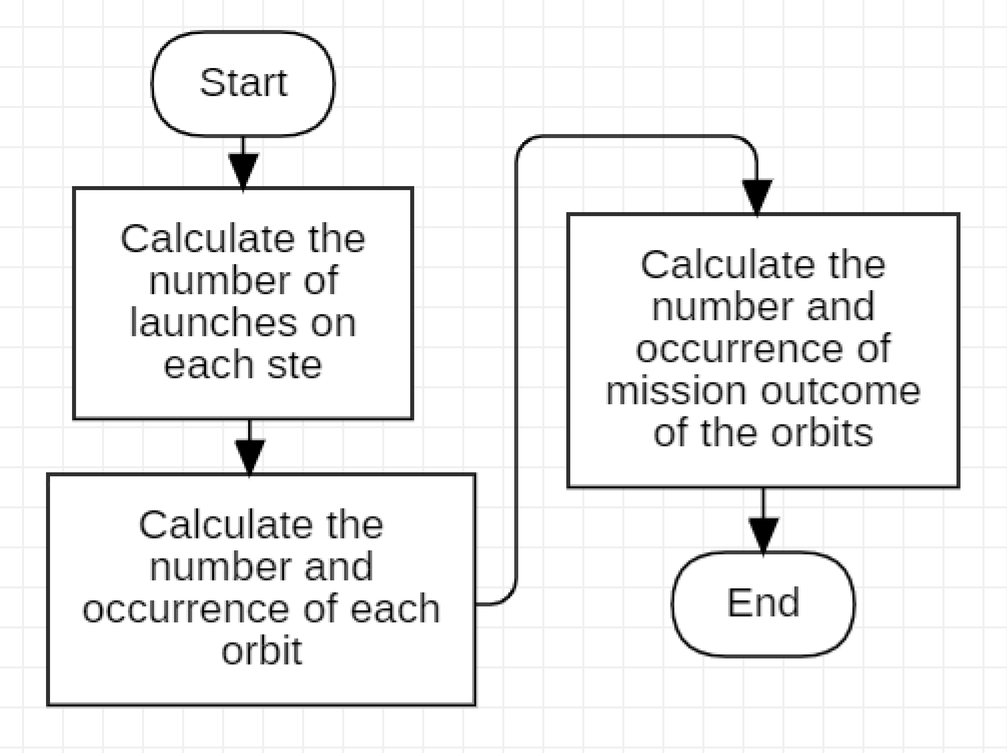
Create a dataframe by parsing the launch HTML tables

End

Data Wrangling

- During this phase, a basic EDA was performed and training labels were determined. Refer to the flow chart on the right.

- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



Start

Calculate the number of launches on each ste

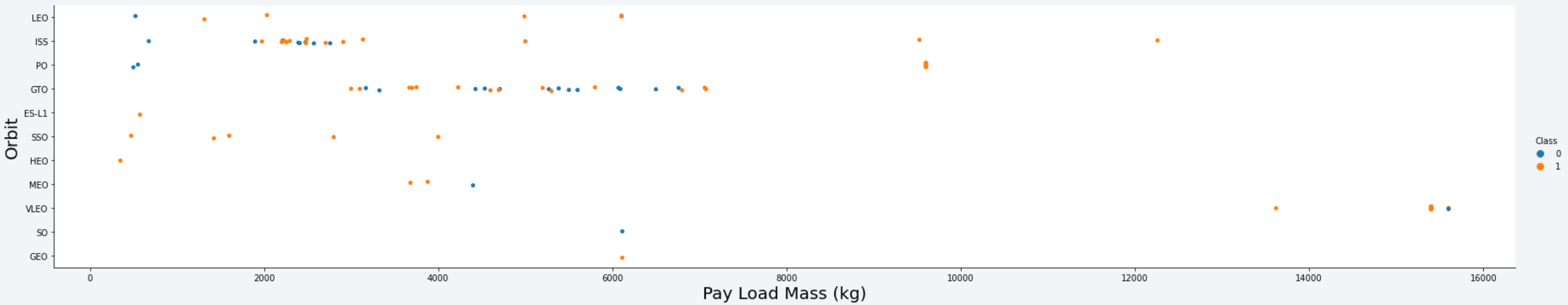
Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome of the orbits

End

EDA with Data Visualization

- At this stage, scatter and bar charts were used mainly because they are very useful to show the relationships between a pair of variables. As an example, a scatter plot was used to visualize the relationship between Payload and Orbit.



- A line graph was also used to show the relationship between class and years.
- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were made:
  - Get the names of the unique launch sites in the space mission.
  - Get the 5 records where launch sites begin with 'CCA'.
  - Get the total payload mass carried by booster launched by NASA (CRS).
  - Get the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass between 4000 – 6000kg.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass.
  - List the records of failure landing outcomes in drone ship (including months, booster versions and launch site) that were made in 2015.
  - Rank the count of landing outcomes between 2010-06-04 - 2017-03-20.
- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In this phase Folium was used to perform an interactive geographic analysis of the data. For example, to know the launch sites.
- Markers were used to mark the launch sites (as well as other places of interest, such as NASA Jonshon Space Center).
- The circles are used to highlight these markers and the clusters help to group several markers that are close to each other.
- On the other hand, lines help to draw lines between places of interest, for example, from a launch site to the nearest coast.

- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

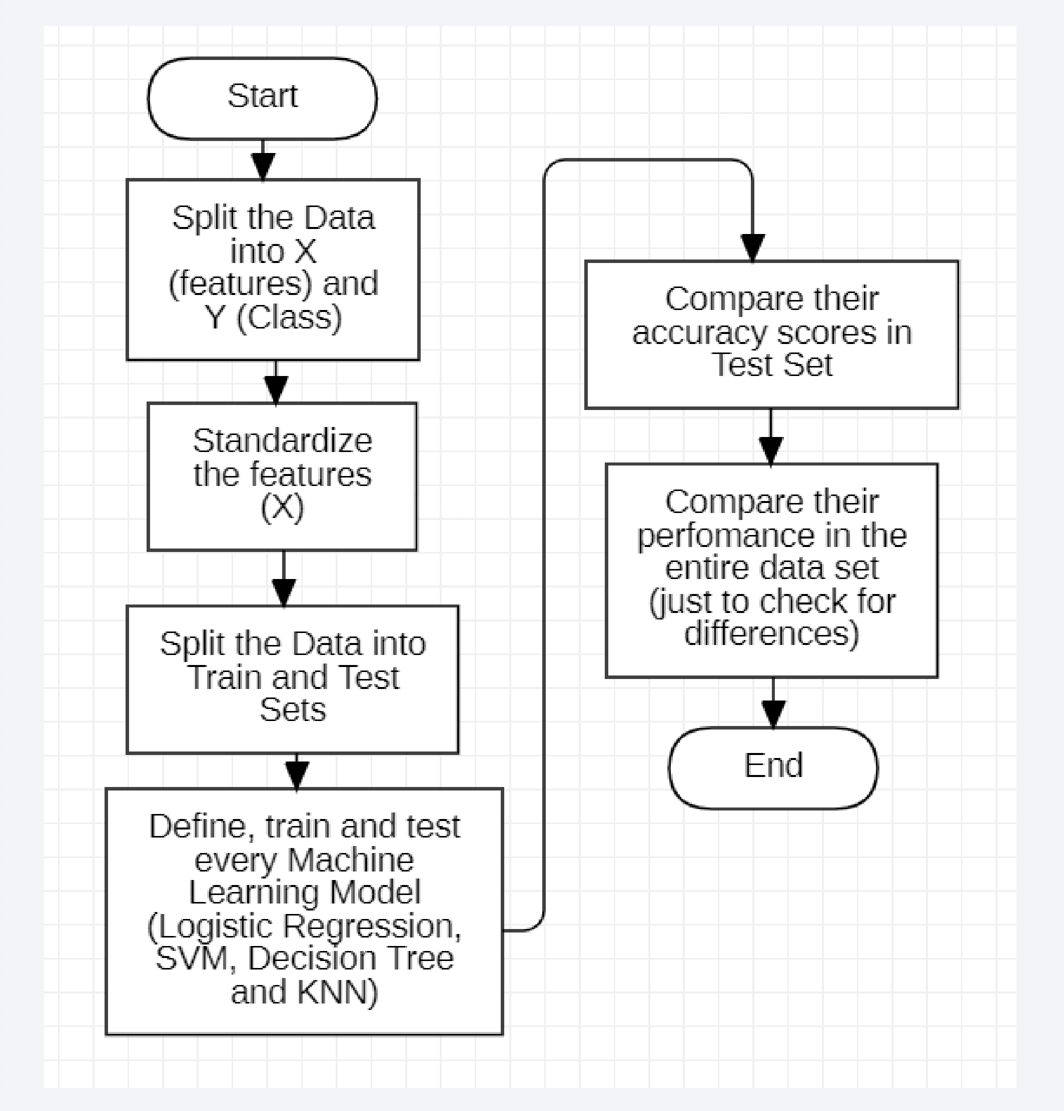
- An interactive web application was developed with Dash to visualize the data.
- Pie charts were used to represent the percentage of successful launches for each site, and
- Scatter plots were used to show the correlation between payload mass (kg) and successful launches for each site.
- In this way it is possible to identify which sites are most suitable for successful launches.

- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four different Machine Learning models (Logistic Regression, SVM, Decision Tree and KNN) were trained and tested. Very similar results were obtained with each model (which will be discussed later). See the diagram on the right for more details.

- GitHub URL: https://github.com/Edy-Blau/Applied_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Start

Split the Data into X (features) and Y (Class)

Standardize the features (X)

Split the Data into Train and Test Sets

Define, train and test every Machine Learning Model (Logistic Regression, SVM, Decision Tree and KNN)

Compare their accuracy scores in Test Set

Compare their perfomance in the entire data set (just to check for differences)

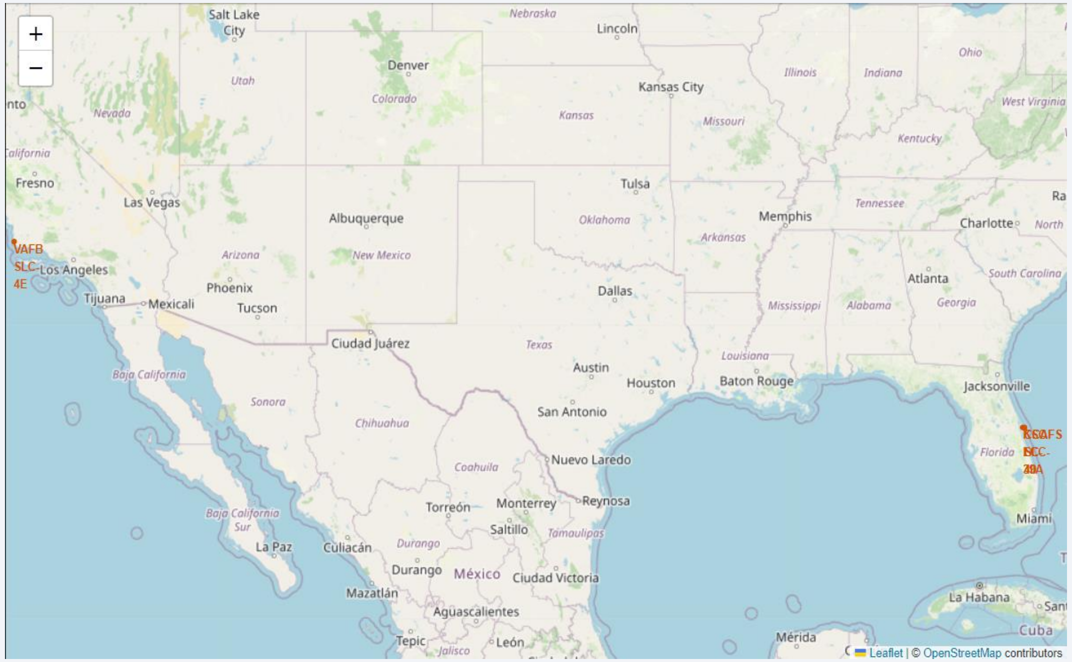
End

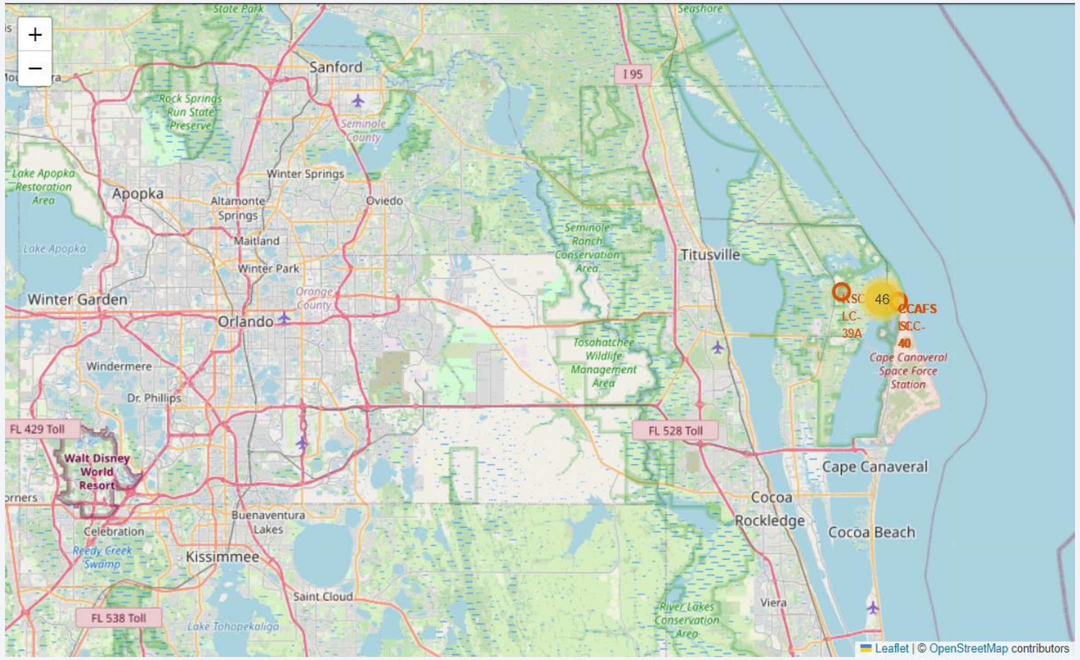
Results

- Exploratory data analysis results
  - The total payload mass carried by boosters launched by NASA was 45,596Kg.
  - The average payload mass carried by booster version F9 v1.1 was 2,928.4Kg.
  - The first successful landing outcome in ground as achieved on 2015-12-22.
  - Boosters F9 FT versions B1022, B1026, B1021.2, and B1031.2 had successful landing outcomes in drone ships even when their payload mass was greater than the average (4000 - 6000 Kg).
  - Almost all launch mission outcomes were successful.
  - Many Boosters versions carried the maximum payload mass.
  - Two versions of the Booster F9 v1.1 (B1012 and B1015) failed to land on drone ship in January and April 2015, respectively.
  - During 2010-06-04 to 2017-03-20, the most significant outcomes were equal number of successful and unsuccessful landings on drone ship (5), 3 successful landings on ground pad and 10 not attempted.

Results

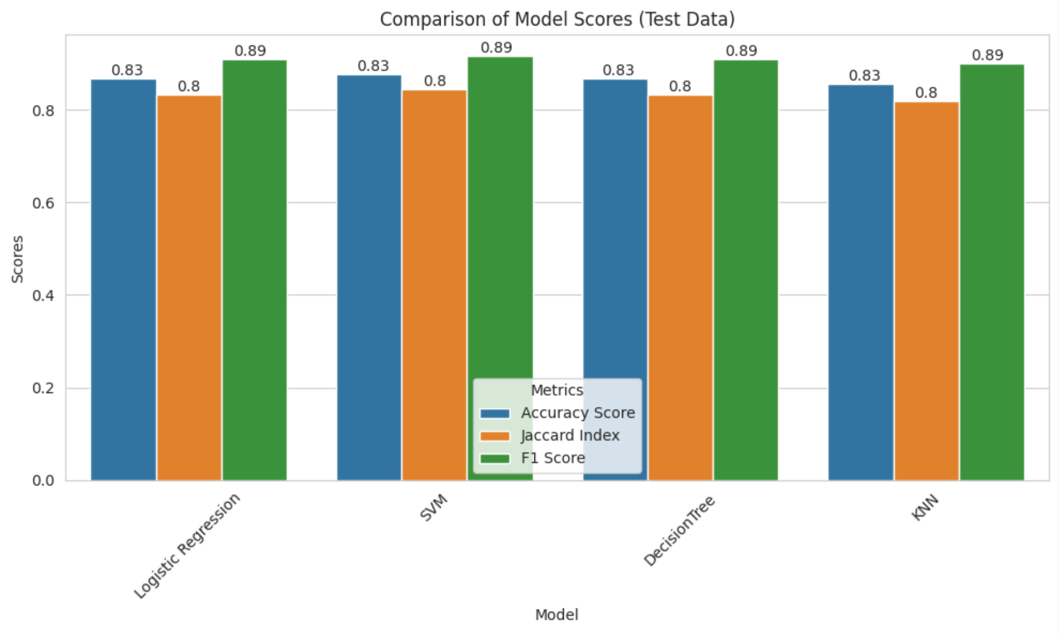
- Interactive analytics demo in screenshots
- The launch stations are close to the coasts and relatively far from large urban centers (a few kilometers), most likely for safety reasons.

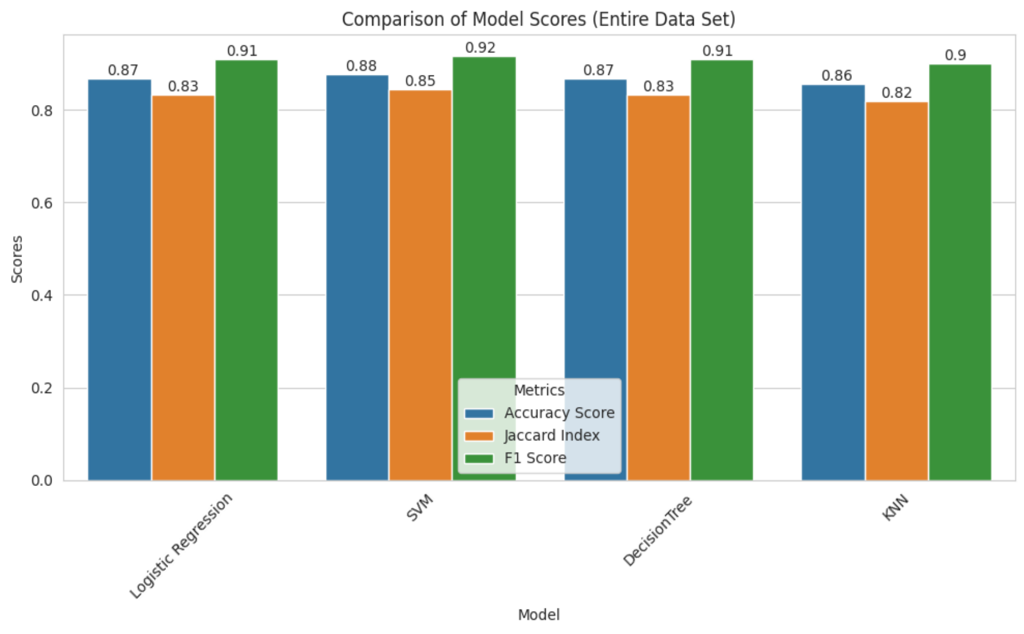



+
−
KSC LC-39A
46
CCAFS SLC-40
Cape Canaveral Space Force Station
Orlando
Titusville
Cape Canaveral
Leaflet | © OpenStreetMap contributors

Results

- Predictive analysis results
- All four models performed equally well when evaluated on the test set (left bar chart). However, in order to highlight the possible differences, the metrics were also calculated on the entire data set (right bar chart), being the SVM the one that performed the best.

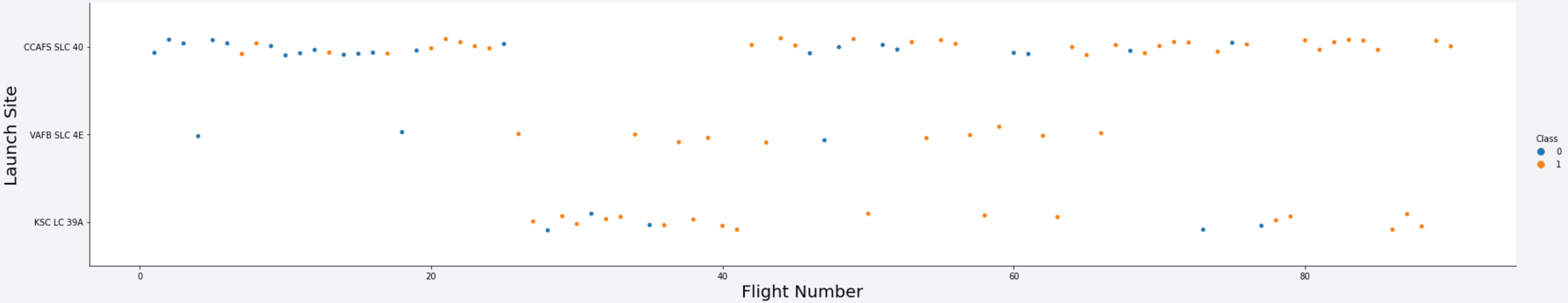




Section 2

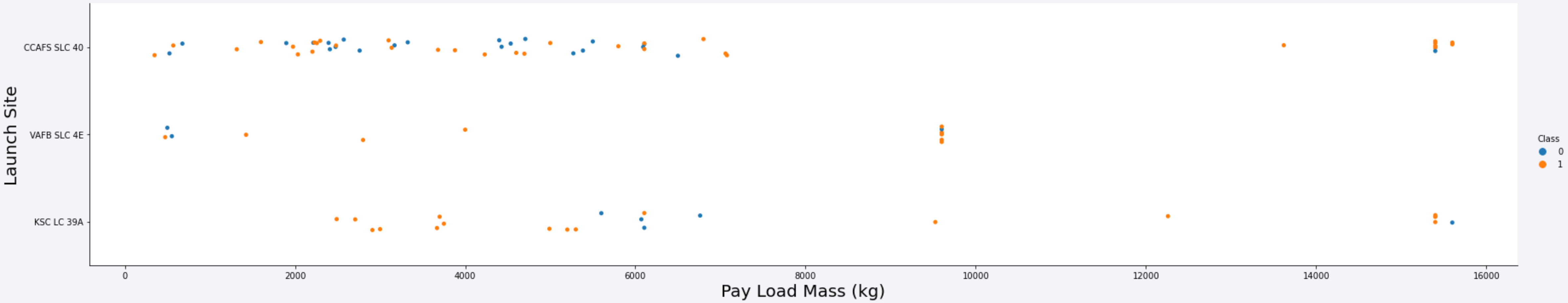
Insights drawn from EDA

Flight Number vs. Launch Site



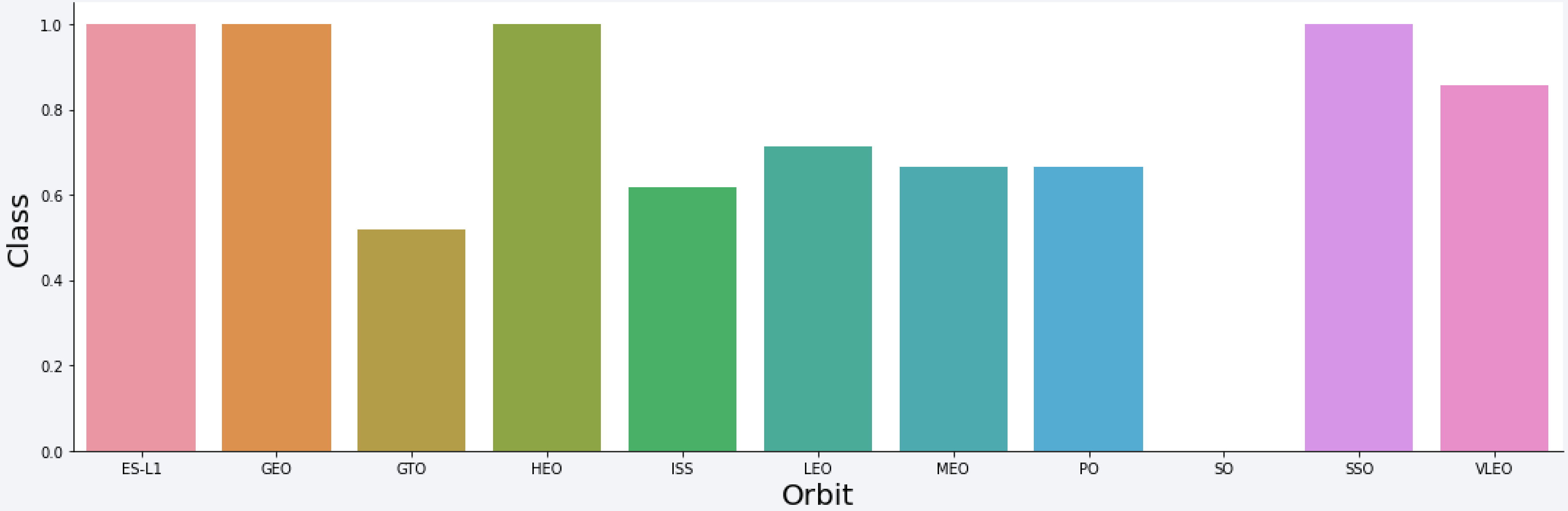
- In general, as the launches progressed, the probability of a successful landing increased. In addition, the site with the highest number of successful landings was CCAFS SLC 40.

Payload vs. Launch Site



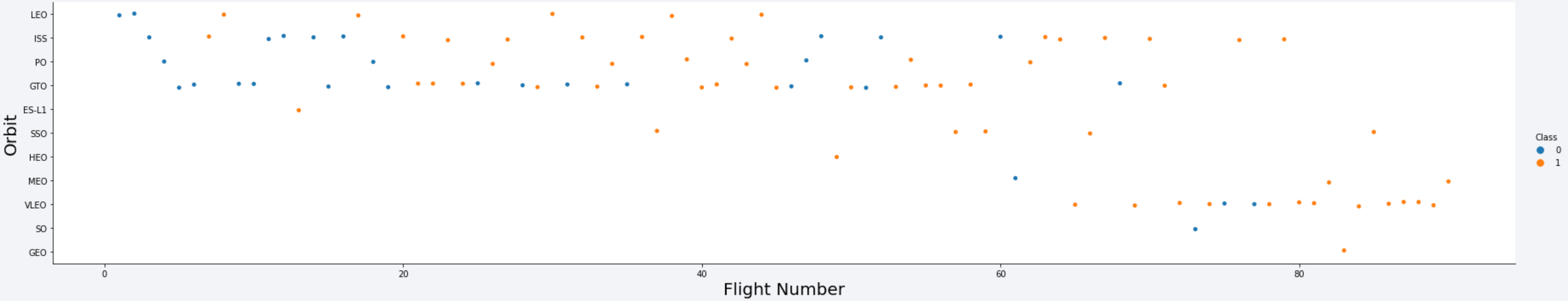
- Launches with payload mass greater than 12,000 kg were only performed in CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type



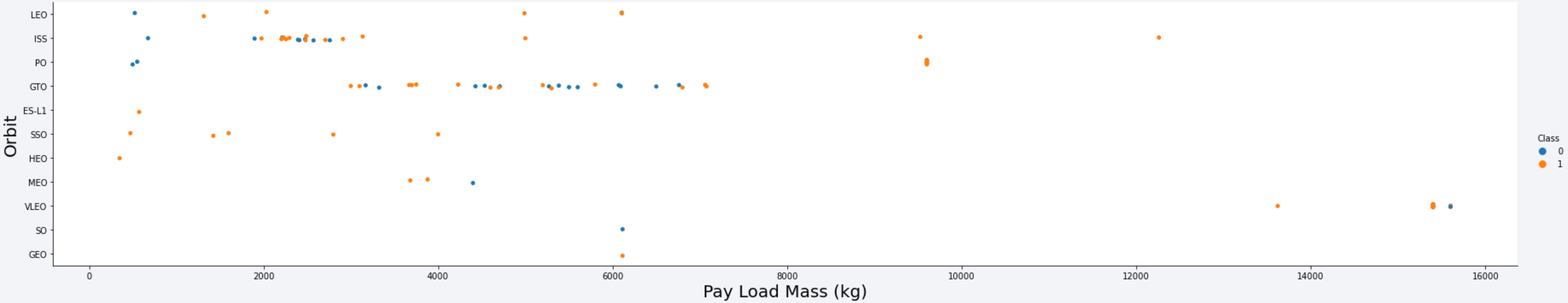
- Orbits with the highest success rate are ES-L1, GEO, HEO and SSO.

Flight Number vs. Orbit Type



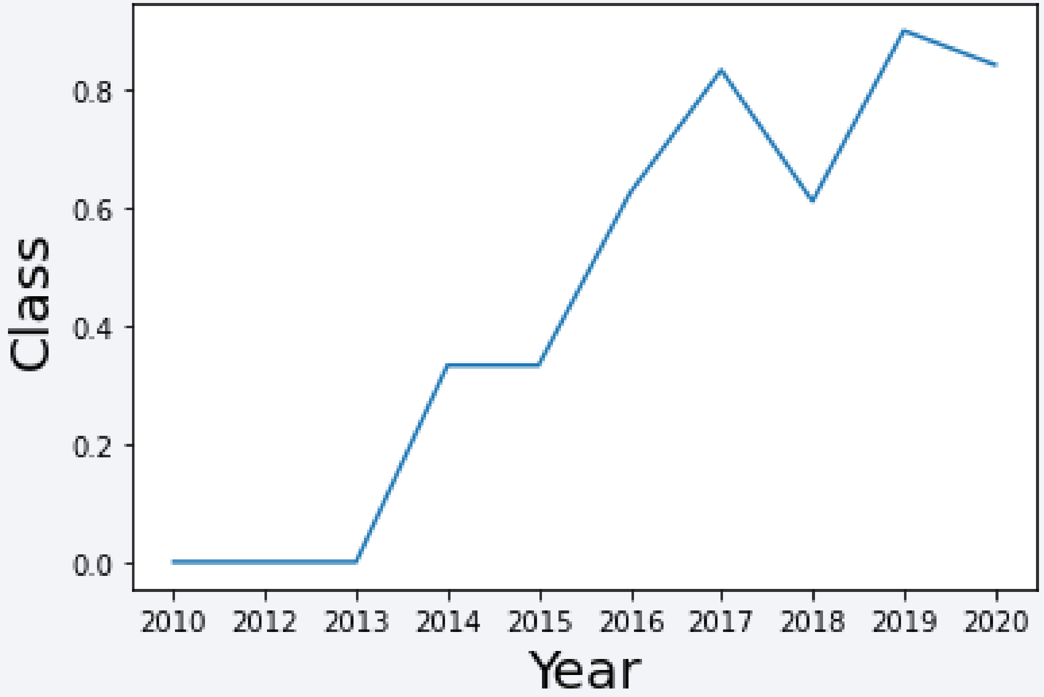
- From flight number 65 onwards (approximately) most of the flights were made in the VLEO orbit.

Payload vs. Orbit Type



- Launches with payloads below 8000 kg were performed in all orbits, but those with higher payloads were permormed only in ISS, PO and VLEO.

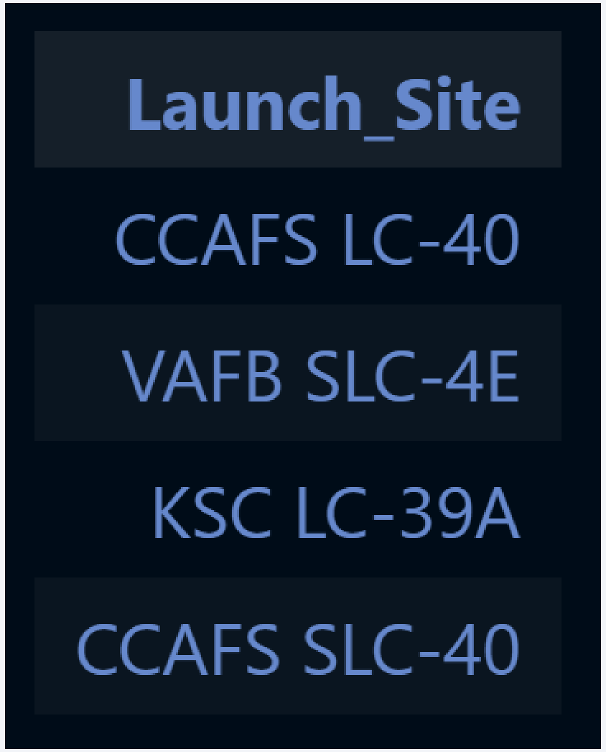
Launch Success Yearly Trend



- The percentage of successful launches starts from 2013, with a small drop in 2018 and an eventual recovery. In 2020 a small drop is seen again (probably due to the beginning of the COVID-19 pandemic).

All Launch Site Names

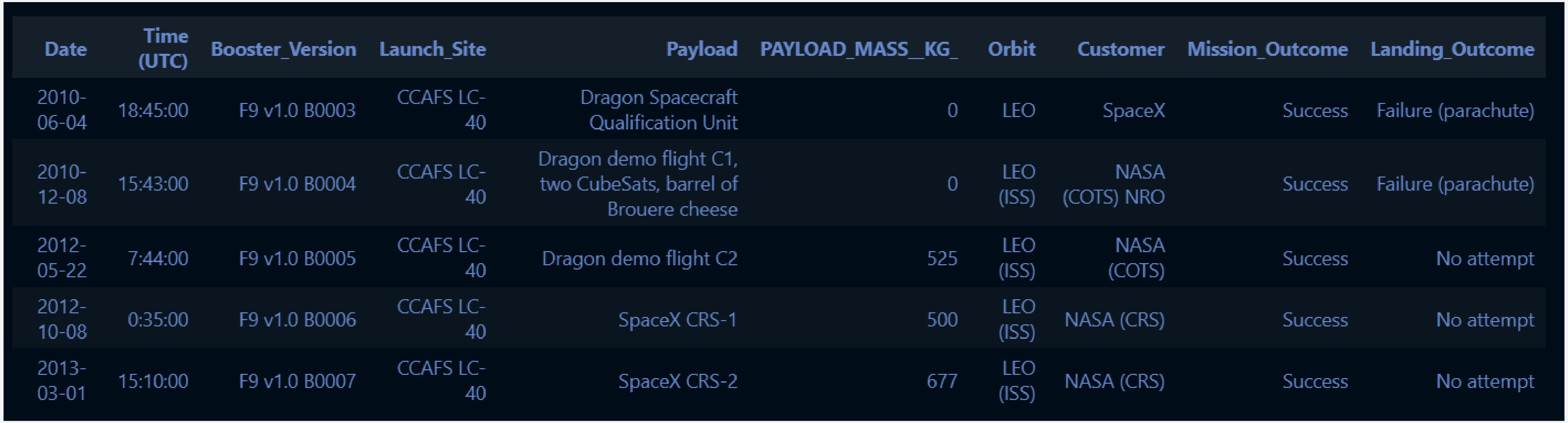
- According to the query, there are four launch sites.



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

Launch Site Names Begin with 'CCA'

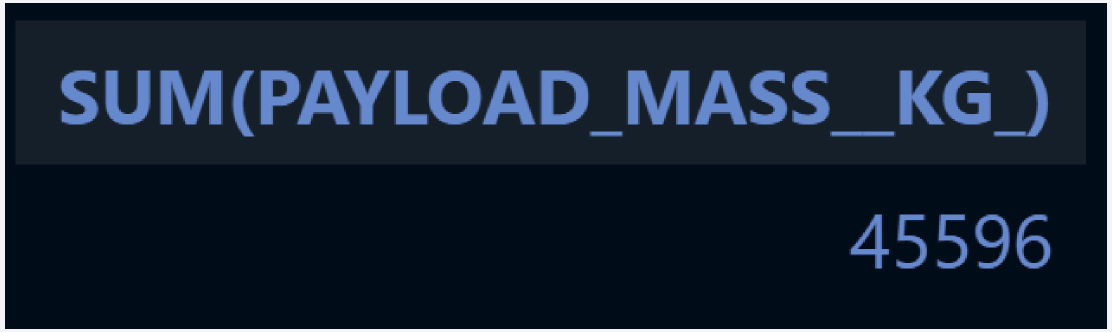
- 5 records where launch sites begin with `CCA`.
- All of them were launched between 2010 – 2013.



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

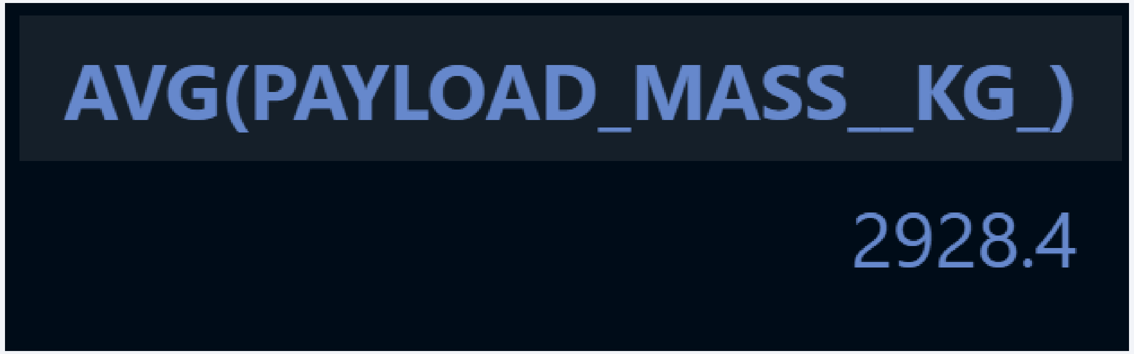
- The total payload carried by boosters from NASA.
- It was calculated using 'NASA (CRS)' as Customer.



SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

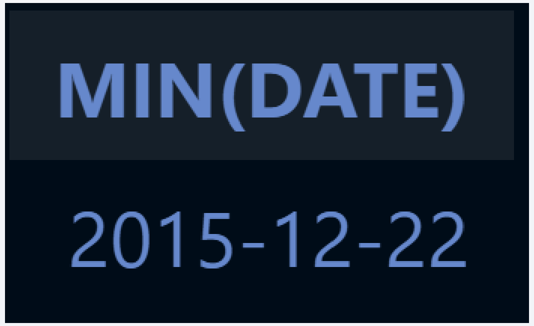
- The average payload mass carried by booster version F9 v1.1
- The average payload mass for this booster is almost 3,000Kg.



AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

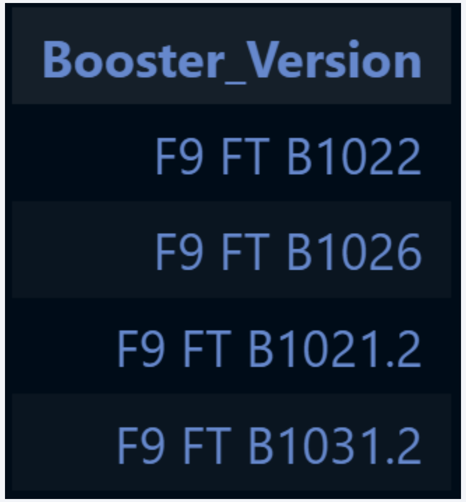
- The date of the first successful landing outcome on ground pad. It was December/22/2015.
- In order to find this date, first it is necessary to know the exact Landing Outcome, which is 'Success (ground pad)'.



MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

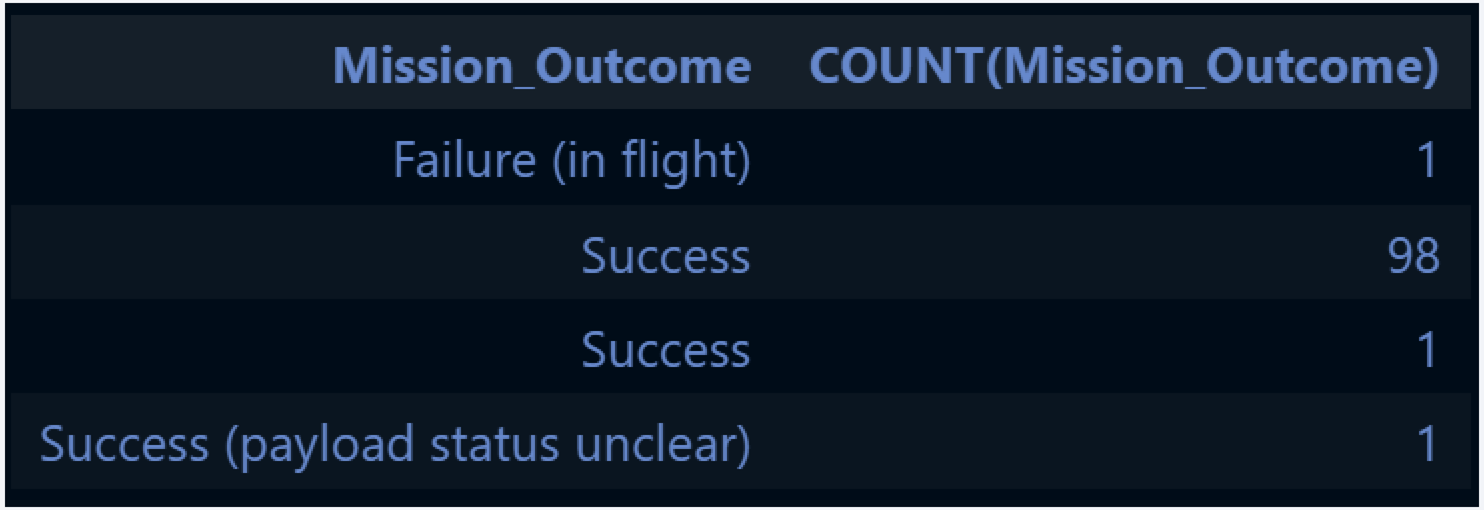
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- These are boosters F9 FT B1022, F9 FT B1026, B1021.2 & B1031.2.



Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

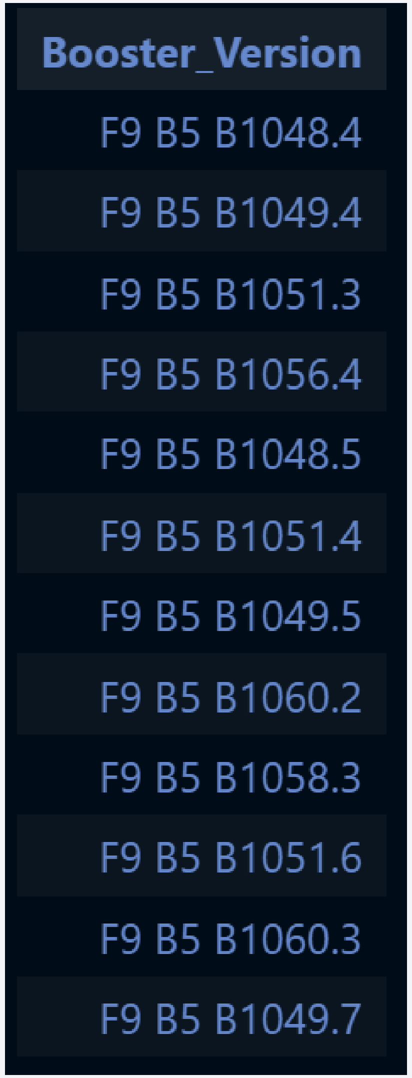
- The total number of successful and failure mission outcomes.
- Almost all the mission were successful.



Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

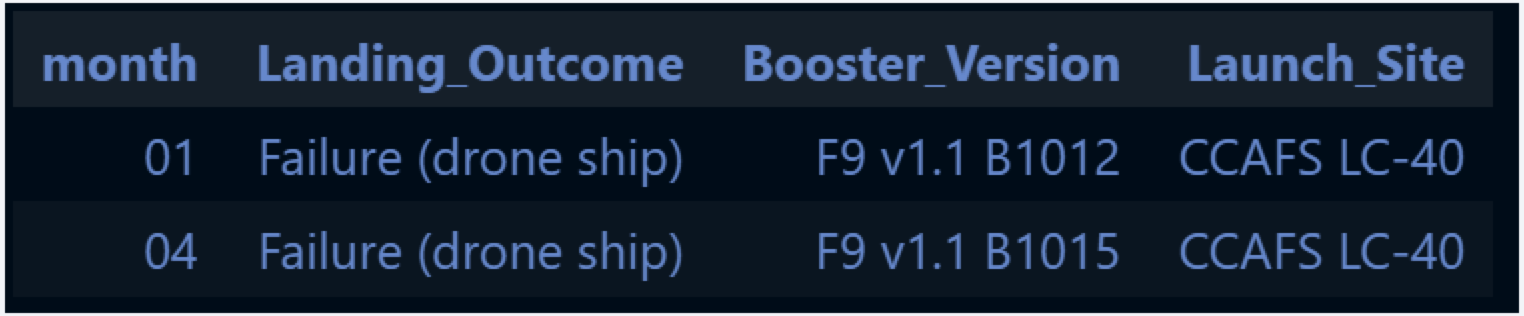
- These are the names of the boosters which have carried the maximum payload mass.



Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

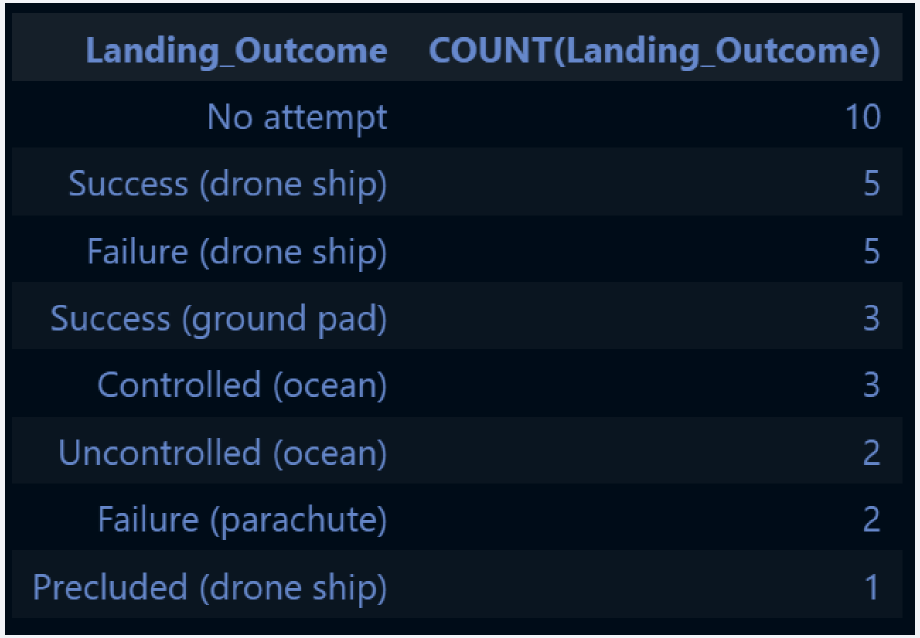
- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- They occurred in January and April.



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

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

- Ranking of landing outcomes between 2010-06-04 and 2017-03-20, in descending order.
- They were 10 Landing Outcomes with "No Attempt".



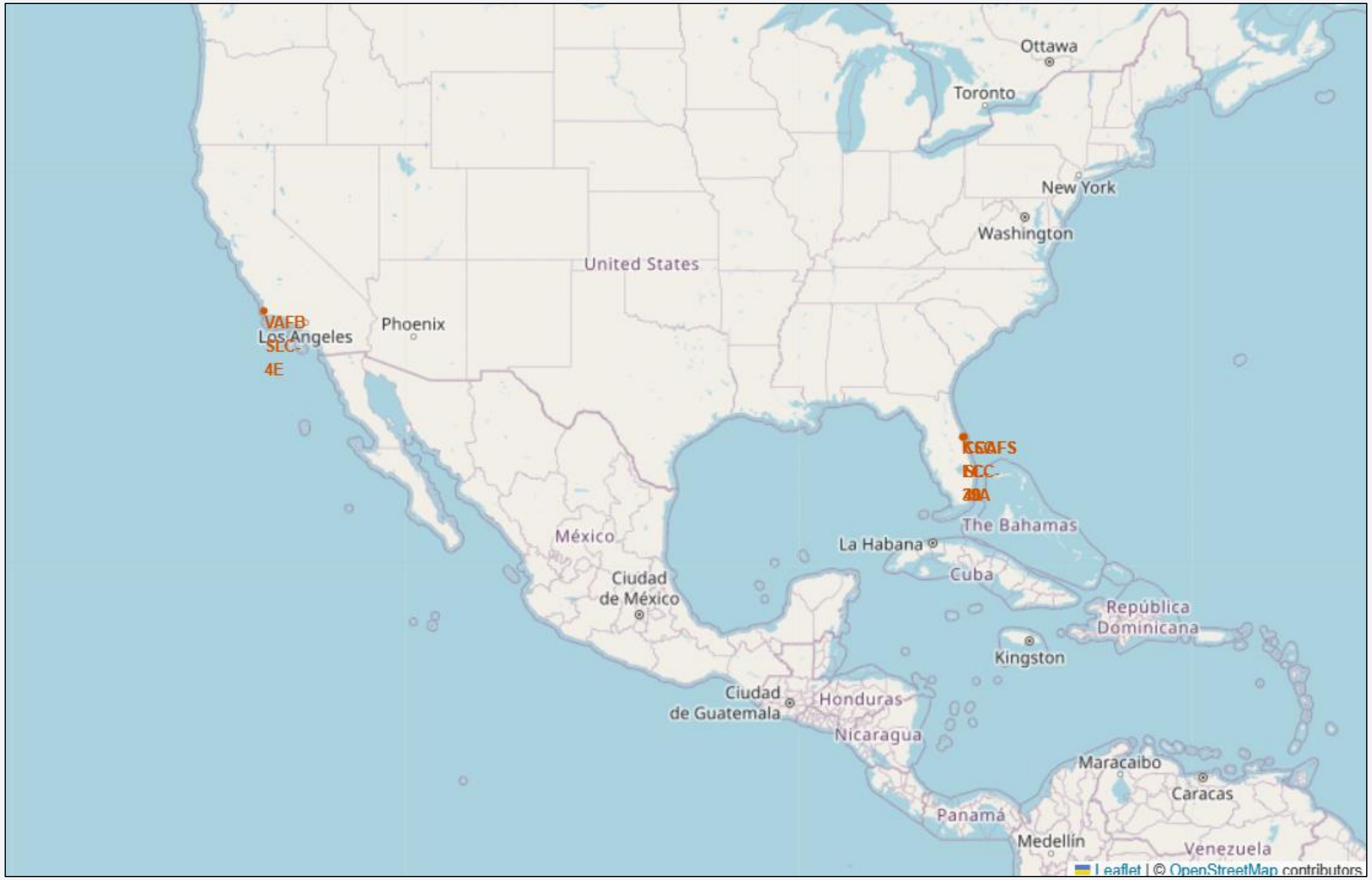
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

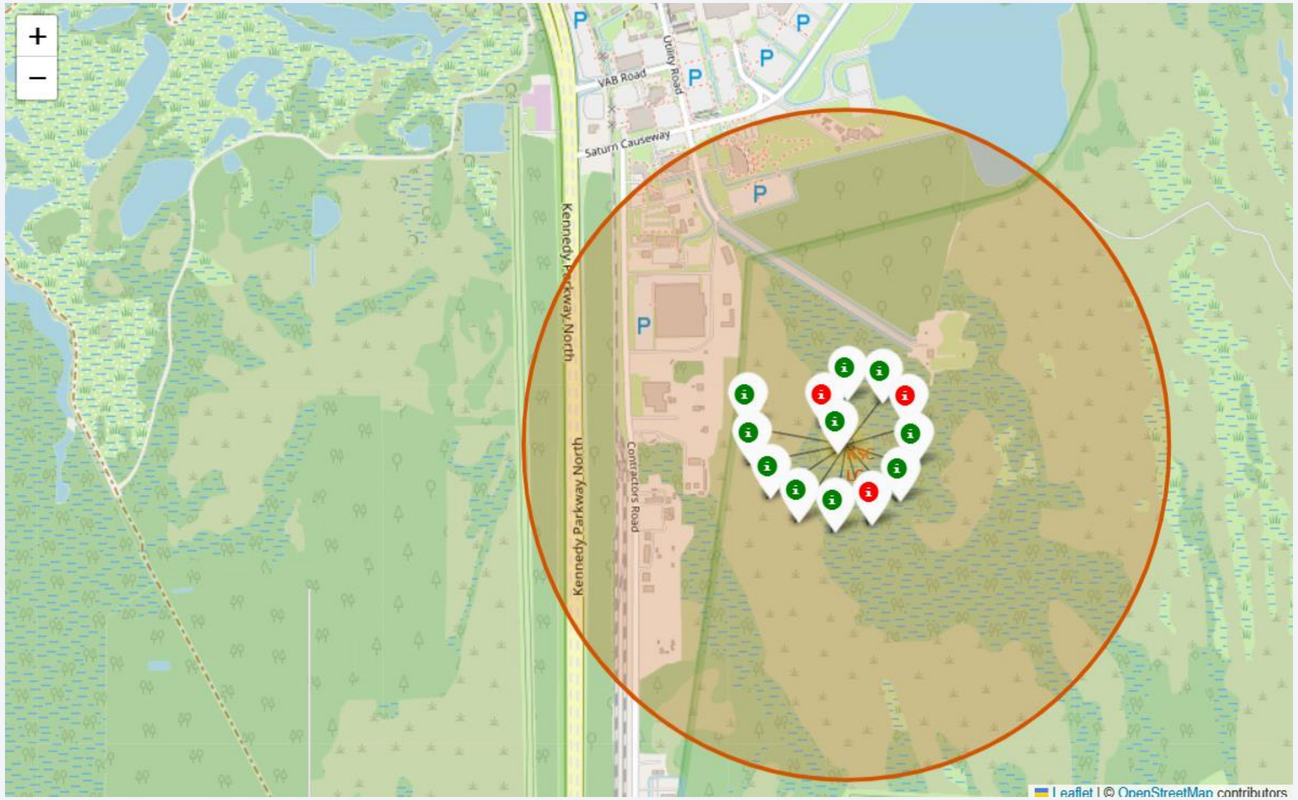
Launch Sites

- The Launch sites are near to the coast in California and Florida.



Launch Outcomes

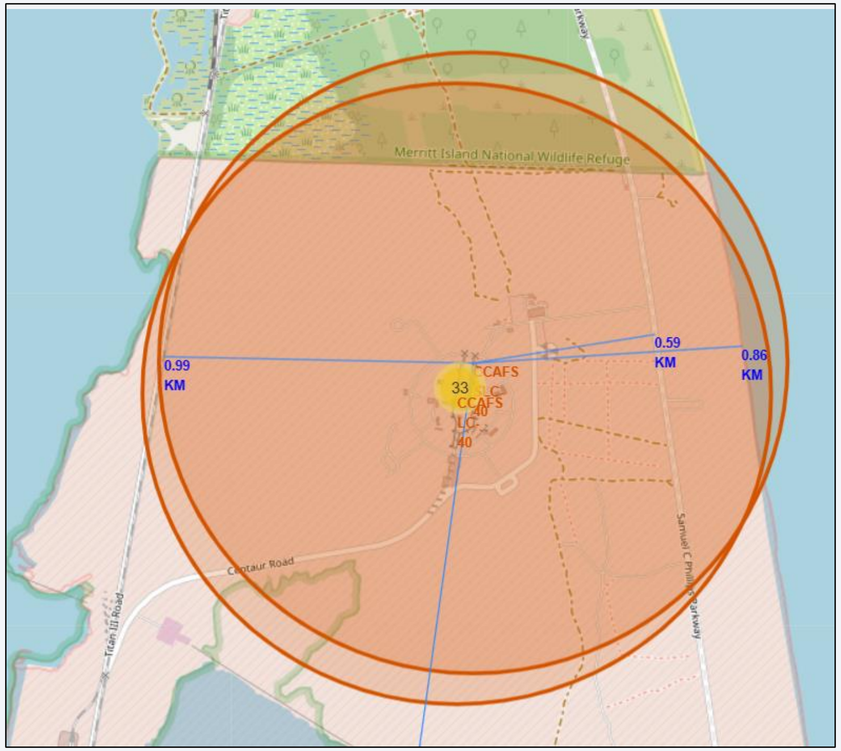
- Successful launch outcomes are represented in green, whereas failed launch outcomes are in red. These representations are from KSC LC-39A Launch Site.

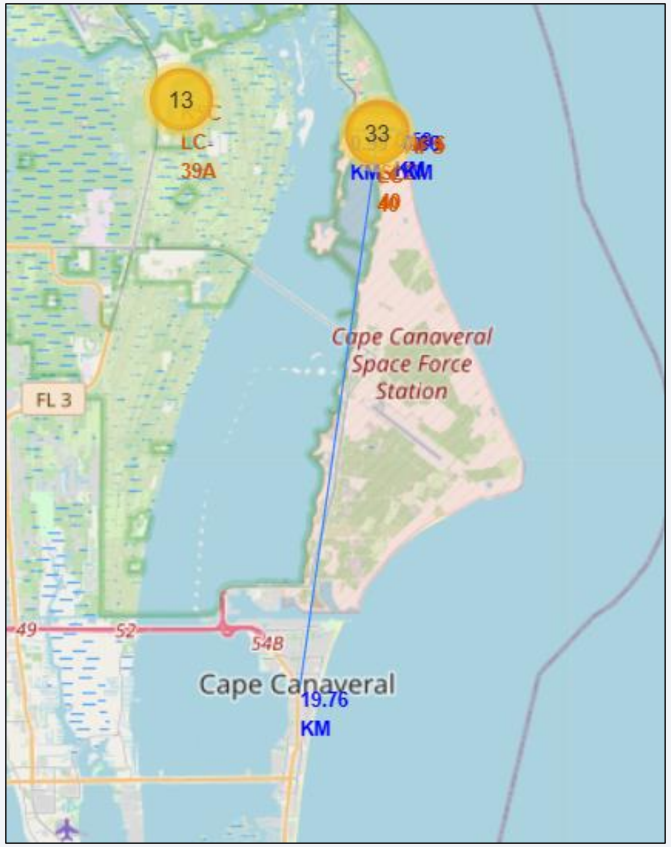


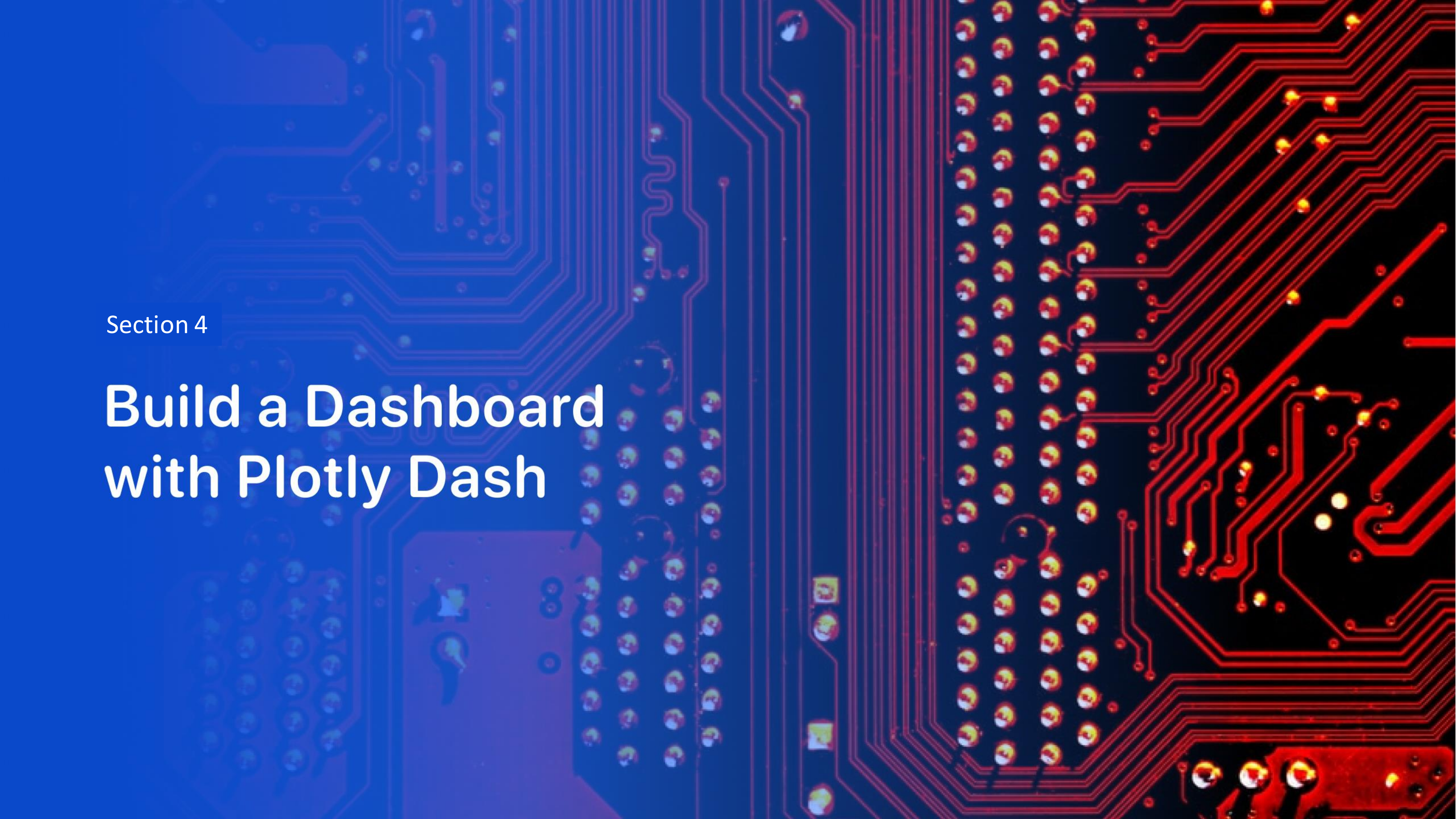
Launch Sites and their proximities

- The following distances are from the Launch Site CCAFS SLC-40:
  - Coastline: 0.86Km
  - Titan III Road (Railway): 0.99Km
  - Samuel C. Phillips Parkway (Highway): 0.59
  - Cape Canaveral (City): 19.76 Km.

- The distances to these points may be mostly due to logistical and security reasons.


Merritt Island National Wildlife Refuge
0.99 KM
0.59 KM
0.86 KM
33
CCAFS
LC-40
Centaur Road
Titan III Road
Samuel C Phillips Parkway



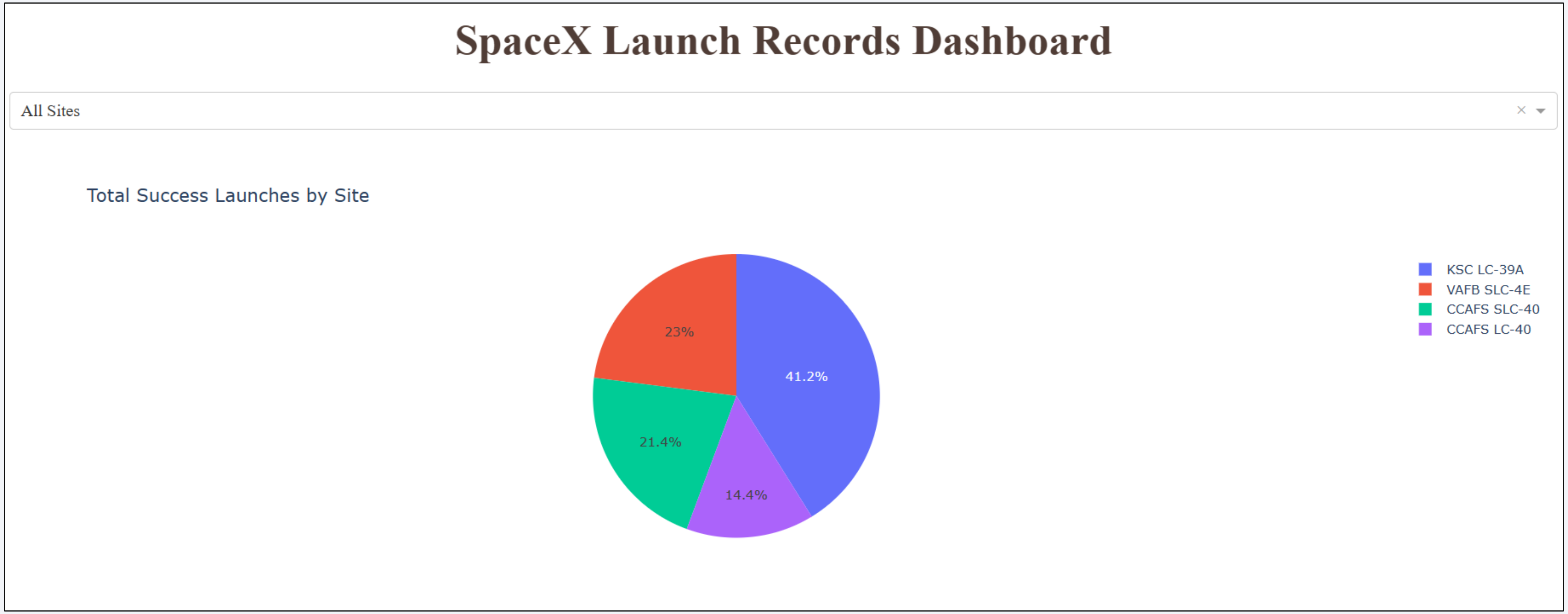


Section 4

Build a Dashboard with Plotly Dash

Total success launches by site

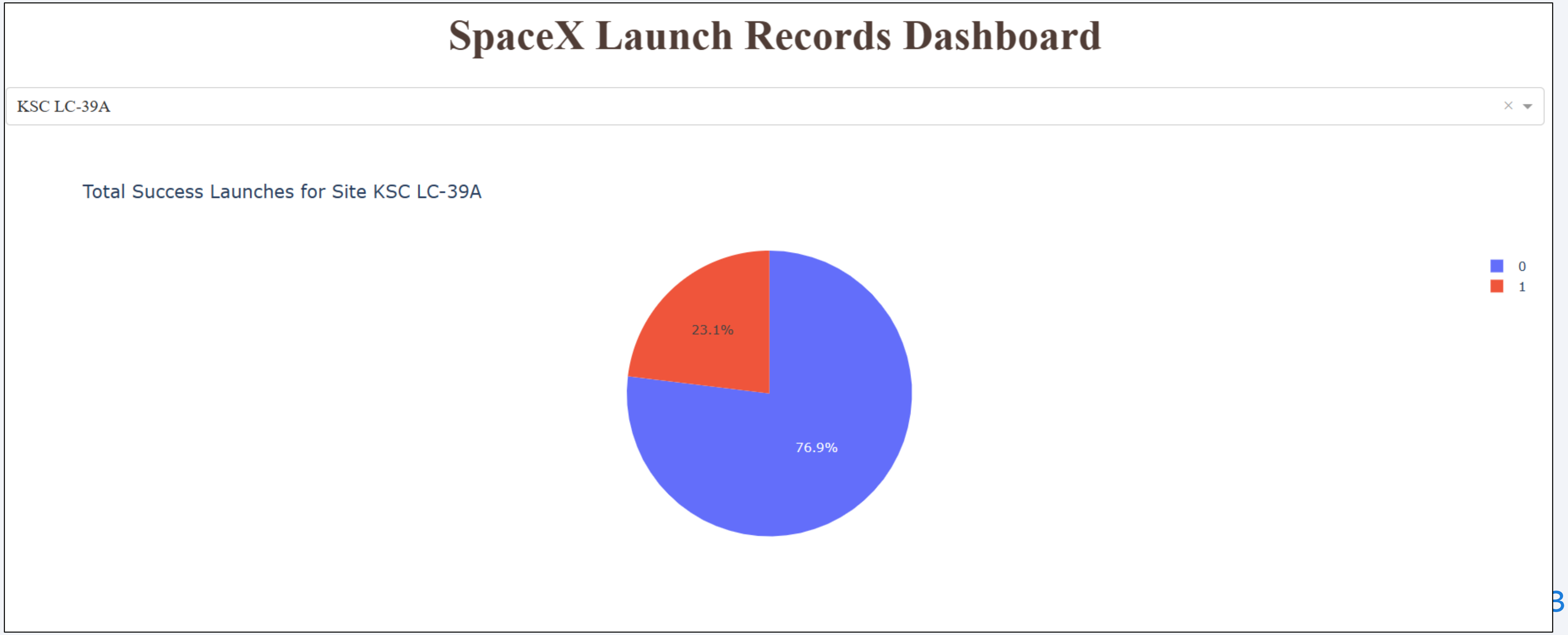
- The launch site is very important to determine the successful of a launch. In this case, for example, the most effective launch site is KSC LC-39A with 41.2% of the total success launches.



SpaceX Launch Records Dashboard

KSC LC-39A The most effective launch site

- KSC LC-39A has the highest launch success ratio.

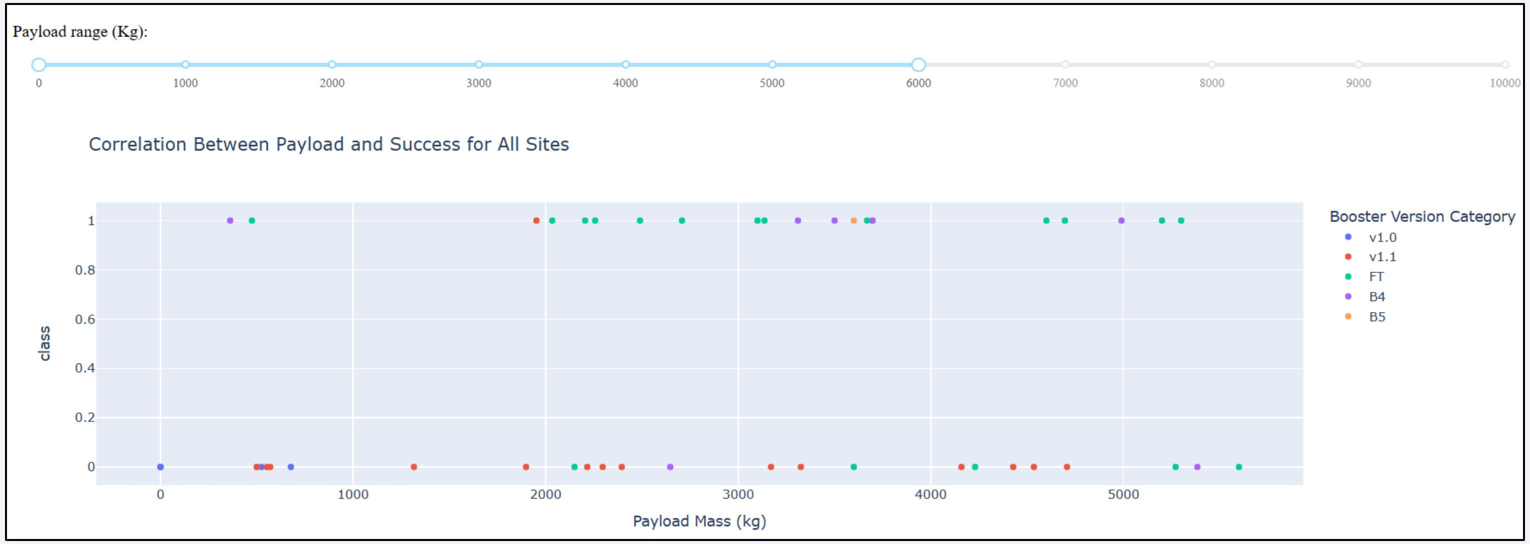


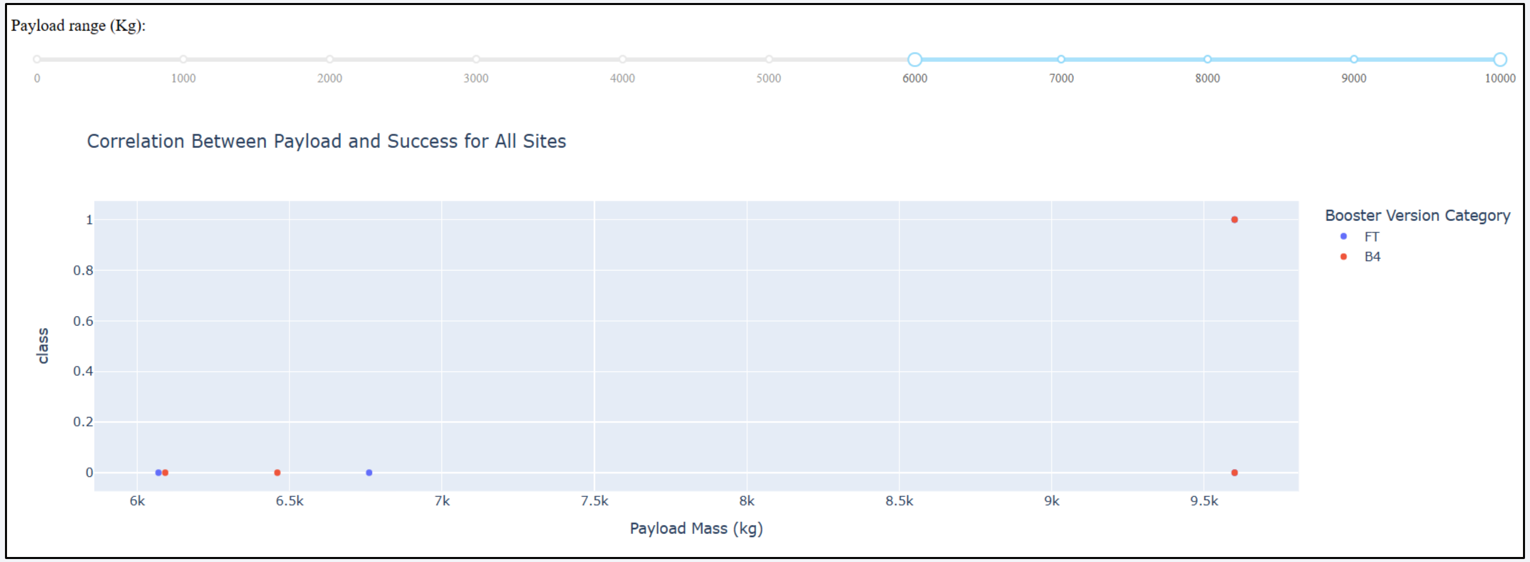
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

Payload vs Launch Outcome





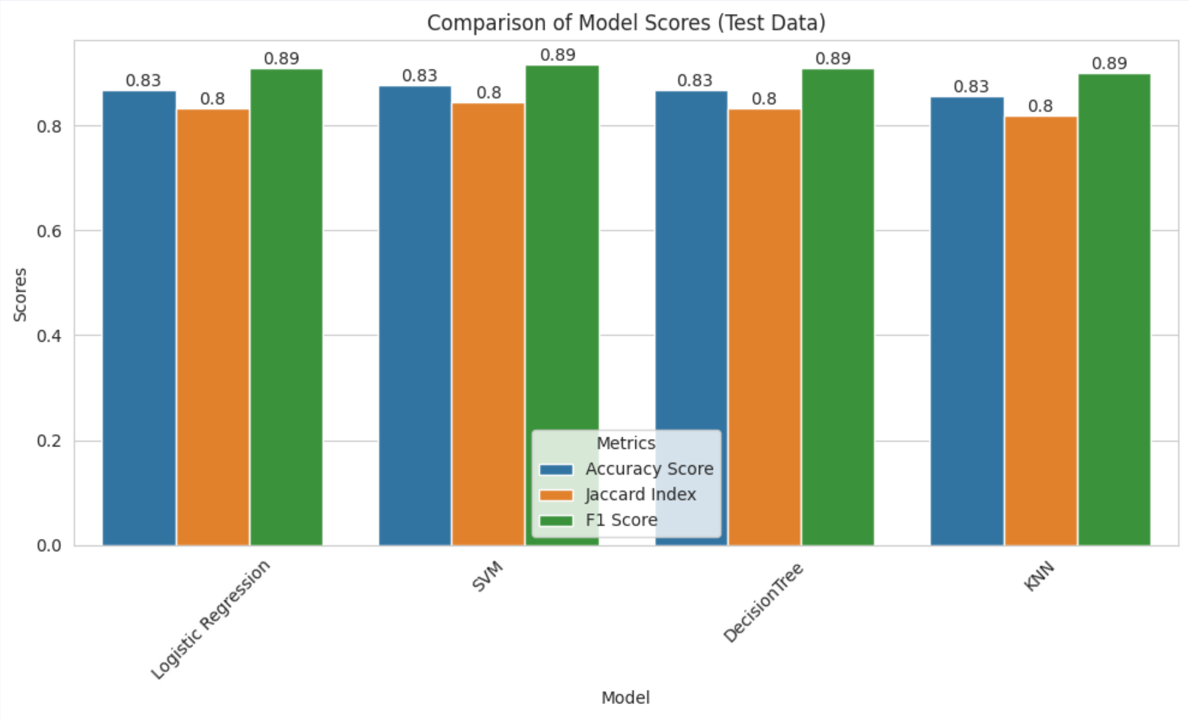
- For payloads less than 6000 kg, there are successful and failed launches, being the most effective booster the FT.
- Meanwhile, for loads greater than 6000Kg, very few launches were performed, being only one successful, with the B4 booster.

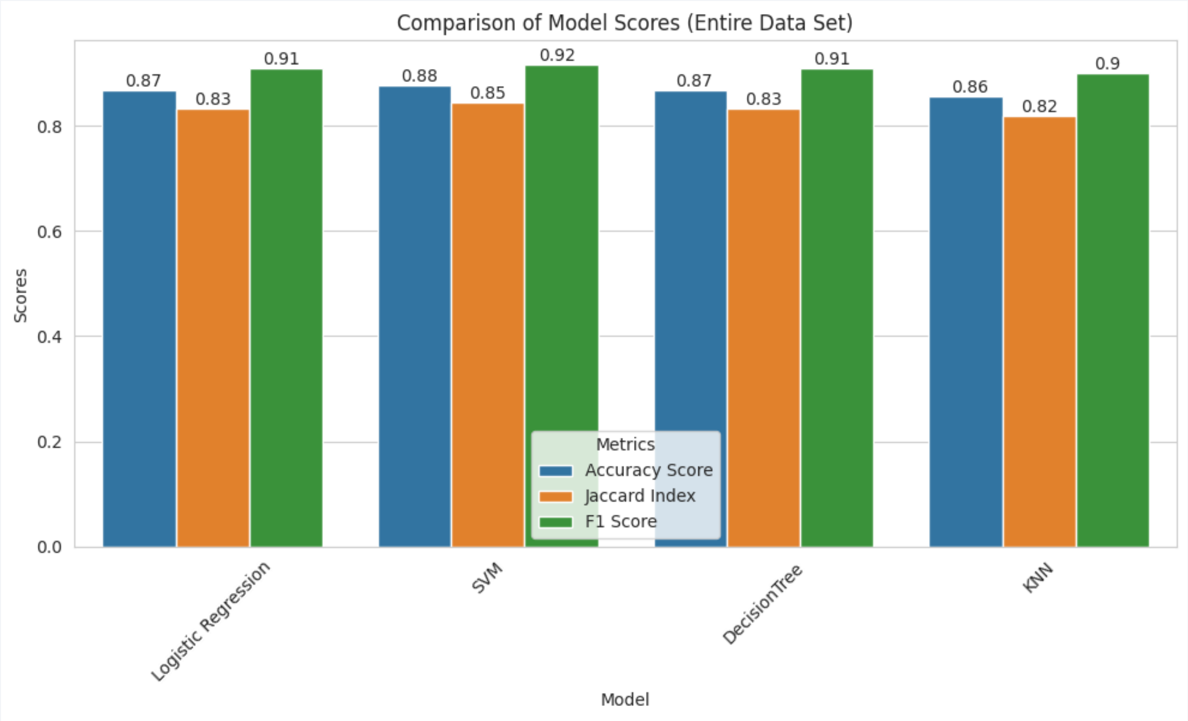
Section 5

Predictive Analysis (Classification)

Classification Accuracy

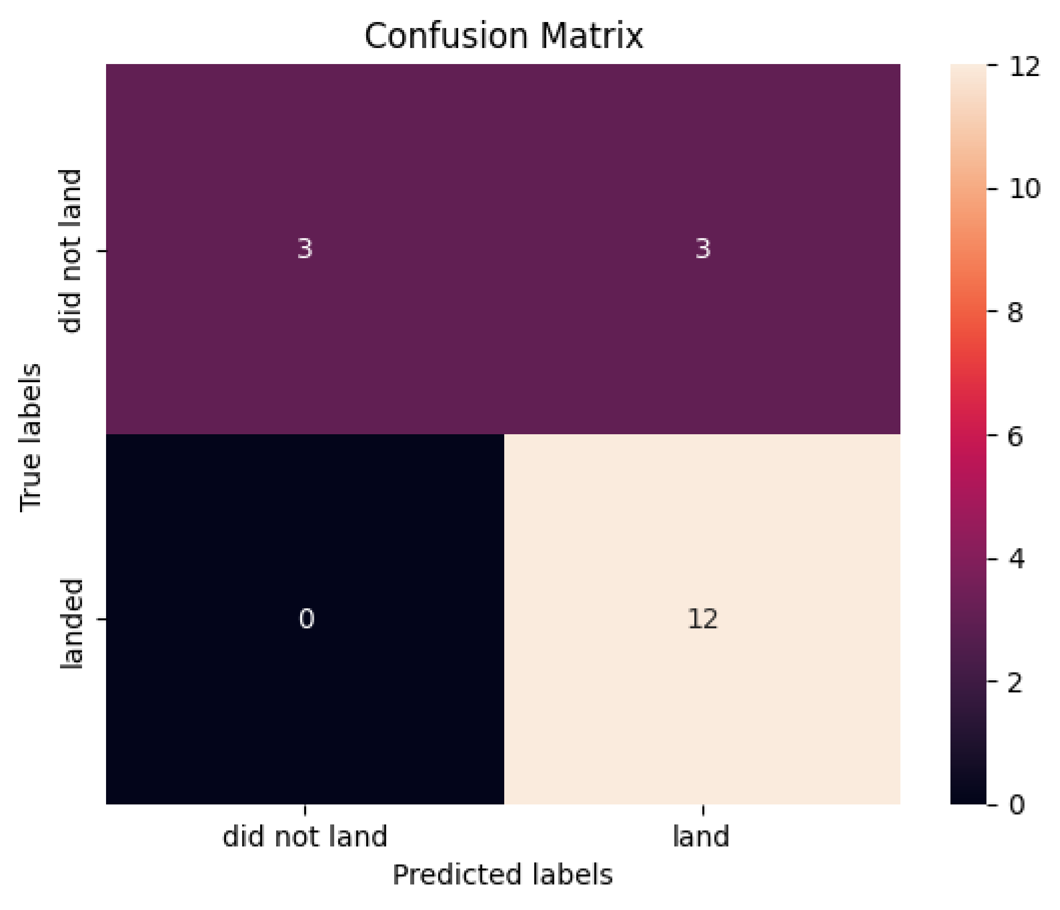
- The performance scores (accuracy, Jaccard index and F1-Score) were the same for the four models on the test set. For this reason these metrics were also calculated on the entire data set (just to check for differences). In this case, the best model was the SVM.

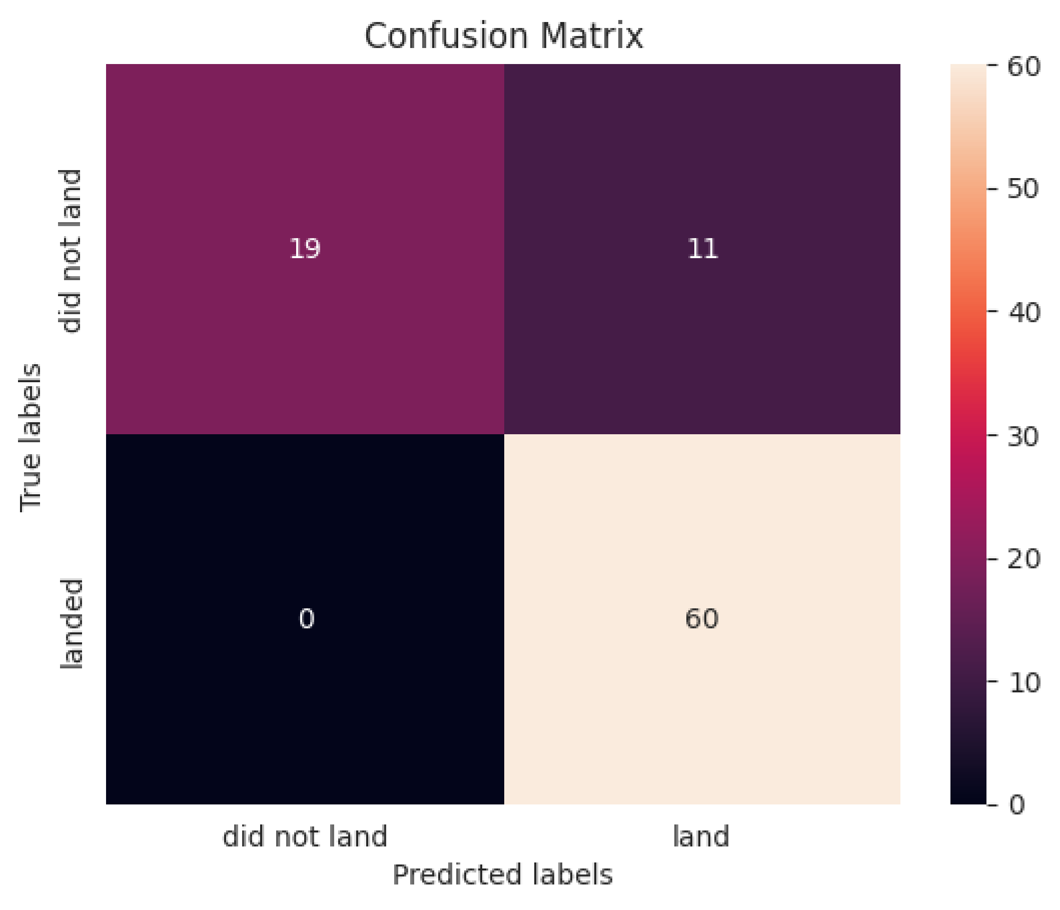




Confusion Matrix

- The confusion matrices for the SVM on the test set (left) and the complete dataset (right) are shown below.



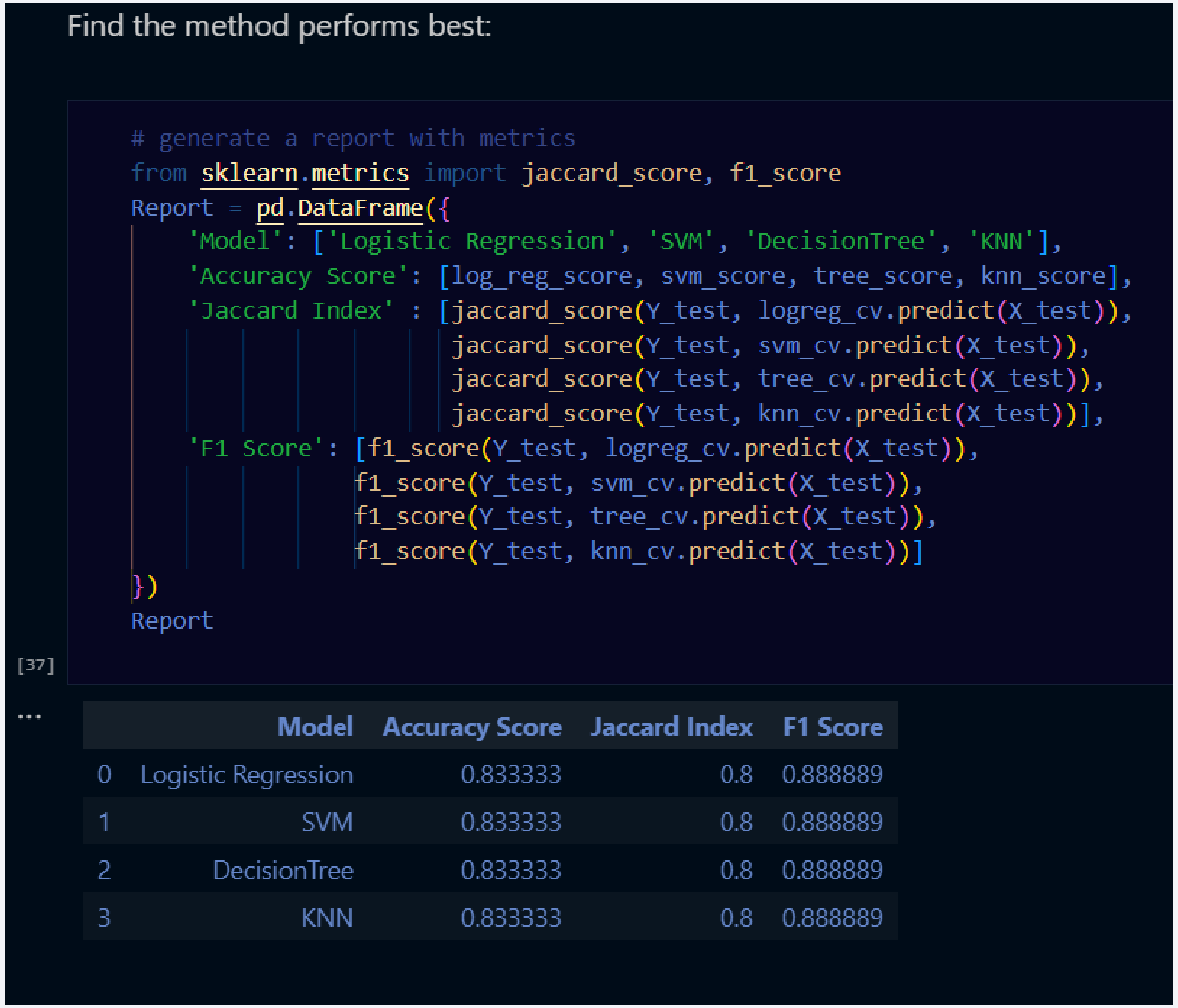


Conclusions

- Although almost 100% of the missions were successful, this does not guarantee that the landing outcome will also be successful. Fortunately, as the years go by and technology advances, each year experiences a greater effectiveness.
- All launch sites are located near the coast. Mainly for logistical and safety reasons.
- The best launch site is KSC LC-39A, located in Florida.
- The best Booster version was FT.
- Among the most important variables that influence the success of a launch outcome are the launch site, the payload mass and the orbit type.
- All the models performed equal on the test set (according to Accuracy, Jaccard Index and F1-Score), but if we consider the entire dataset, the best model is SVM.
- Based on the data, it can be possible to predict if the First Stage of a rocket will land successfully or not. But in order t improve the results, the Machine Learning models can be optimized

Appendix

- Below are some important screenshots of code snippets that may be of interest.

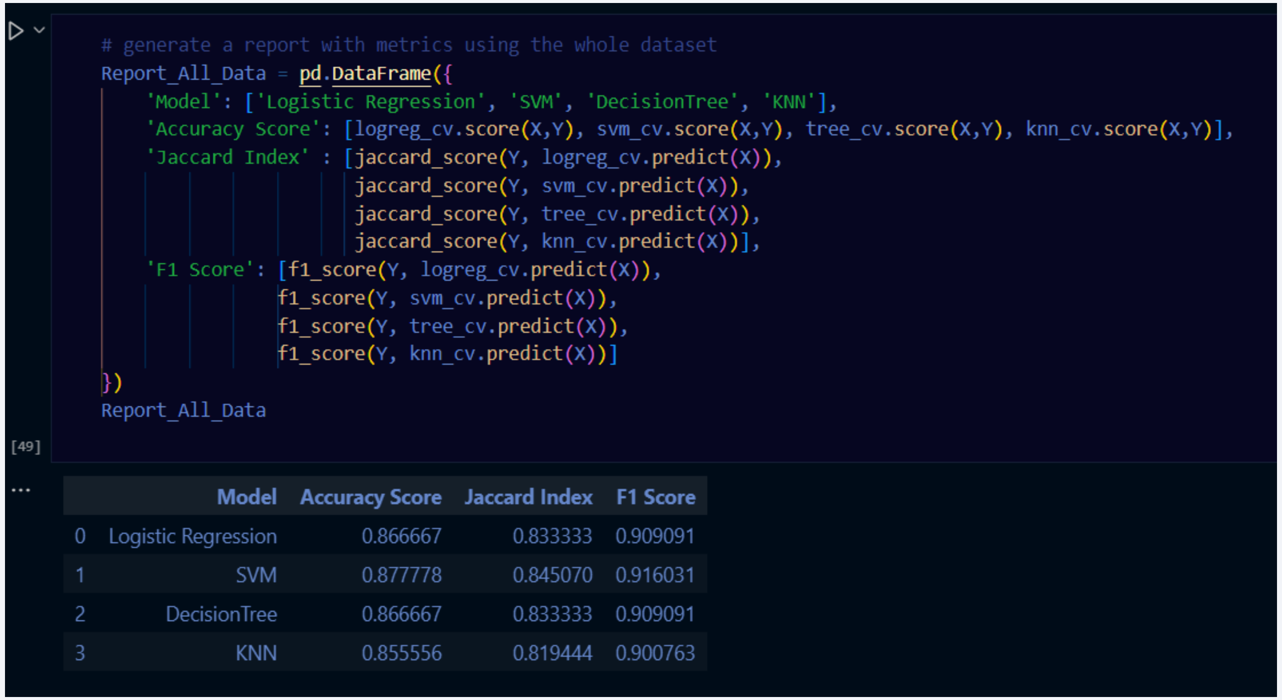


Find the method performs best:

```python
# generate a report with metrics
from sklearn.metrics import jaccard_score, f1_score
Report = pd.DataFrame({
    'Model': ['Logistic Regression', 'SVM', 'DecisionTree', 'KNN'],
    'Accuracy Score': [log_reg_score, svm_score, tree_score, knn_score],
    'Jaccard Index' : [jaccard_score(Y_test, logreg_cv.predict(X_test)),
                       jaccard_score(Y_test, svm_cv.predict(X_test)),
                       jaccard_score(Y_test, tree_cv.predict(X_test)),
                       jaccard_score(Y_test, knn_cv.predict(X_test))],
    'F1 Score': [f1_score(Y_test, logreg_cv.predict(X_test)),
                f1_score(Y_test, svm_cv.predict(X_test)),
                f1_score(Y_test, tree_cv.predict(X_test)),
                f1_score(Y_test, knn_cv.predict(X_test))]
})
Report
```

[37]

	Model	Accuracy Score	Jaccard Index	F1 Score
0	Logistic Regression	0.833333	0.8	0.888889
1	SVM	0.833333	0.8	0.888889
2	DecisionTree	0.833333	0.8	0.888889
3	KNN	0.833333	0.8	0.888889



```python
# generate a report with metrics using the whole dataset
Report_All_Data = pd.DataFrame({
    'Model': ['Logistic Regression', 'SVM', 'DecisionTree', 'KNN'],
    'Accuracy Score': [logreg_cv.score(X,Y), svm_cv.score(X,Y), tree_cv.score(X,Y), knn_cv.score(X,Y)],
    'Jaccard Index' : [jaccard_score(Y, logreg_cv.predict(X)),
                       jaccard_score(Y, svm_cv.predict(X)),
                       jaccard_score(Y, tree_cv.predict(X)),
                       jaccard_score(Y, knn_cv.predict(X))],
    'F1 Score': [f1_score(Y, logreg_cv.predict(X)),
                f1_score(Y, svm_cv.predict(X)),
                f1_score(Y, tree_cv.predict(X)),
                f1_score(Y, knn_cv.predict(X))]
})
Report_All_Data
```

[49]

	Model	Accuracy Score	Jaccard Index	F1 Score
0	Logistic Regression	0.866667	0.833333	0.909091
1	SVM	0.877778	0.845070	0.916031
2	DecisionTree	0.866667	0.833333	0.909091
3	KNN	0.855556	0.819444	0.900763

Thank you!