IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Massi-Nissa Abboud
Wed. 19 oct 2022

Outline

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

Executive Summary

- This study aims to analyze the landing attempts of the first stage of Space X's Falcon 9, in order to predict if the landing will succeed based on several informations such as launch site, weight, etc…
- We firstly collected data from multiple sources and cleaned it .
- After this, we performed exploratory data analysis using SQL and visualizing tools like matplotlib to extract some useful insights from the data.
- We then build an interactive dashboard in which we can visualize these insights in an interactive way.
- We finally build several ML models in order to predict if the first stage of Falcon 9 will land successfully.

Executive Summary

Some useful insights we found when analizing our data:

- CCAFS SLC 40 launch site have lowest success rate (60%) but a higher number of attempts.
- No rockets are launched for heavypayload mass(greater than 10000) from the VAFB-SLC launchsite.
- Rockets which are launched on ES-L1, GEO, HEO,SSO,VLEO orbits have a success rate higher than 0.8.
- For heavy payloads, the successful landing rate are more for Polar,LEO and ISS.
- In the LEO orbit, the Success appears related to the number of flights.
- Sucess rate kept increasing since 2013.
- Launch sites are close to the coast, have near railways, far from highways and cities.

Introduction

Project background and context:

In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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 by our company SpaceY to bid against SpaceX for a rocket launch.

Problems for which we want to find answers:

- What are the main characteristics of the first stage of a rocket?
- Which parameters influence the choice of the launch site?
- What are the geographical features shared by the launch sites?
- Define if a first stage will land successfully or not.

Section 1

Methodology

Methodology

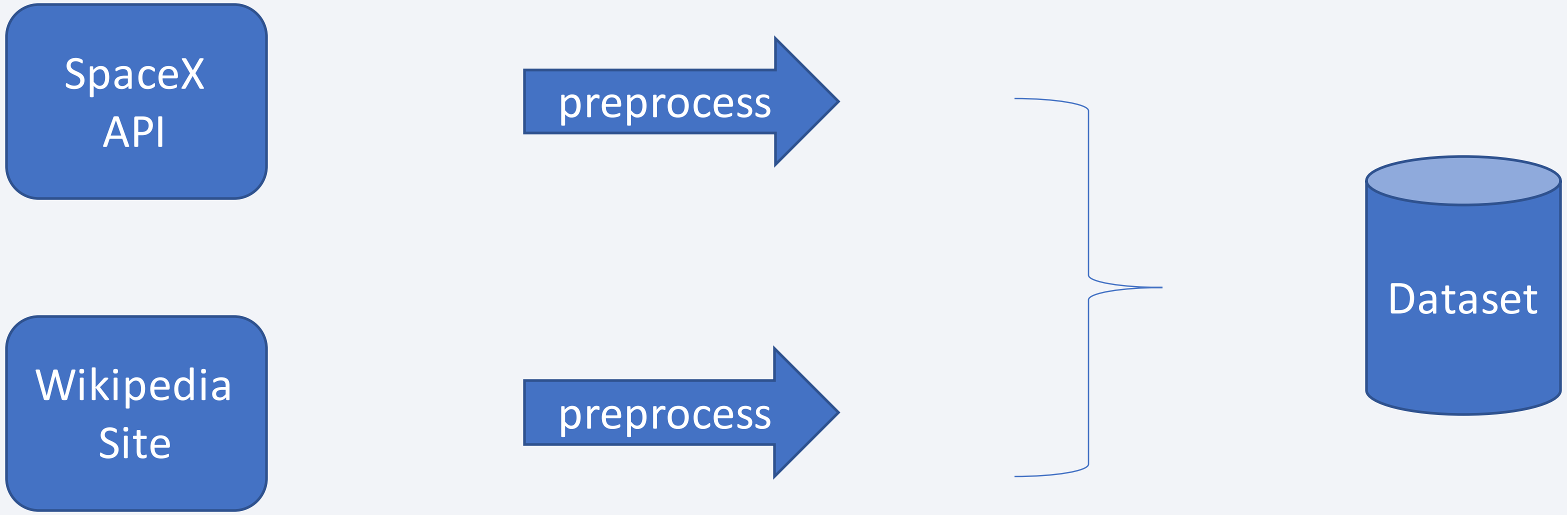
Executive Summary

- Data collection methodology:
  - Data was collected from two main sources : the SpaceX API , data scraping from the Wikipedia site.
- Perform data wrangling
  - Clean the data and preprocess it. Adding a "class" column which represents the success or failure of a landing.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models
  - Train different models and extract for each one of them the best parameters.
  - Test the accuracy of each model and determine the one which performs the best

Data Collection

Data was collected from two main sources :

- SpaceX API
- Web scraping from Wikipedia

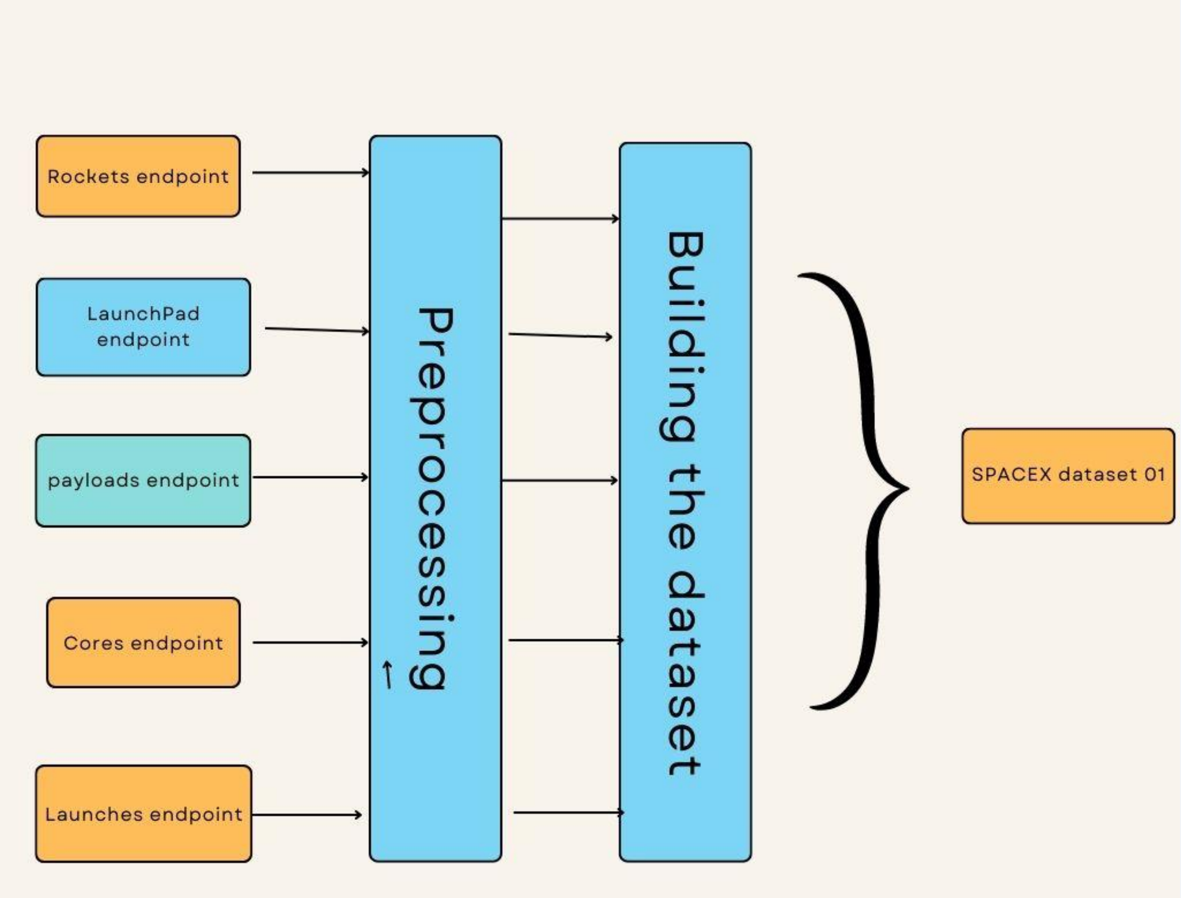


Data Collection – SpaceX API

Data are collected from the SpaceX API using different get requests to the following endpoints:

- https://api.spacexdata.com/v4/rockets/
- https://api.spacexdata.com/v4/launchpads/
- https://api.spacexdata.com/v4/payloads/
- https://api.spacexdata.com/v4/cores/
- https://api.spacexdata.com/v4/launches/past/

You can check the data collection through this link : https://github.com/MassixXx/Capstone-Project-Space-X/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

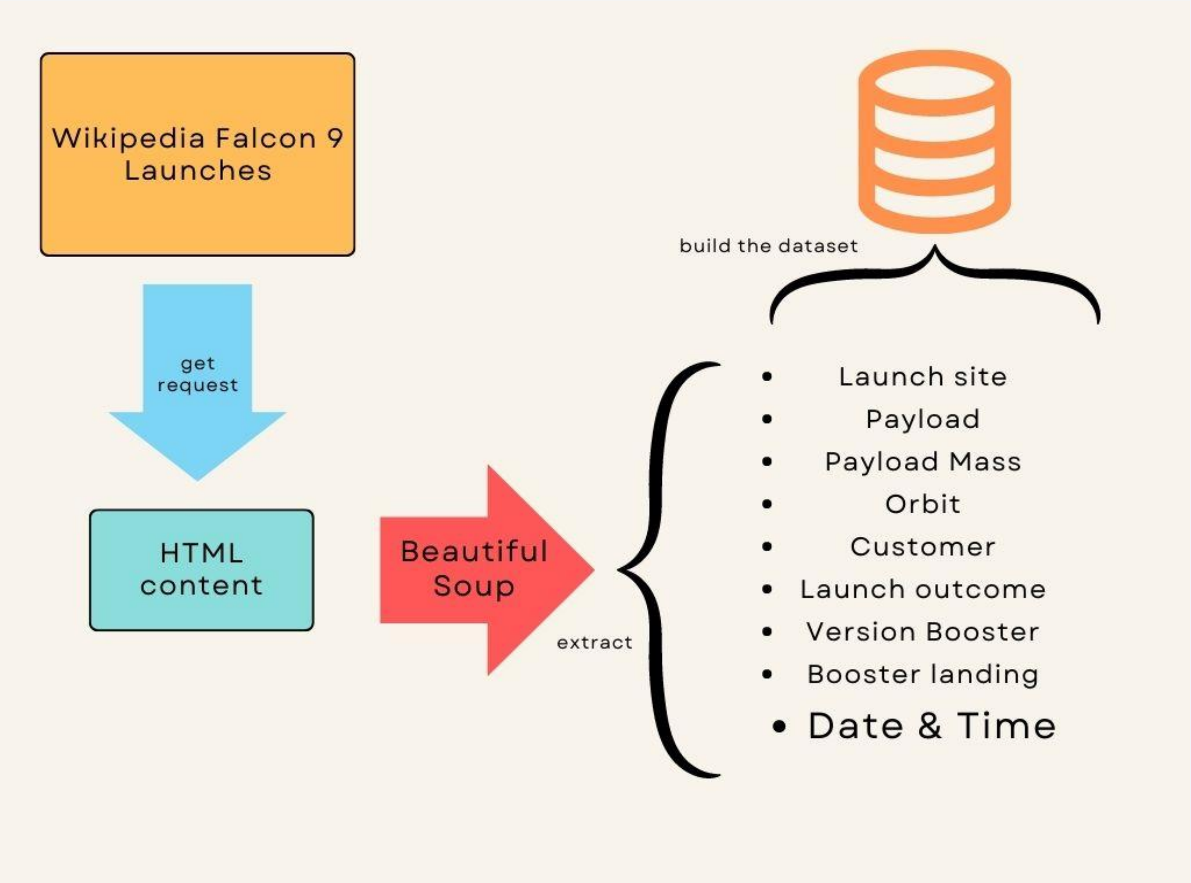


Data Collection - Scraping

Data has been collected using web scraping through this steps:

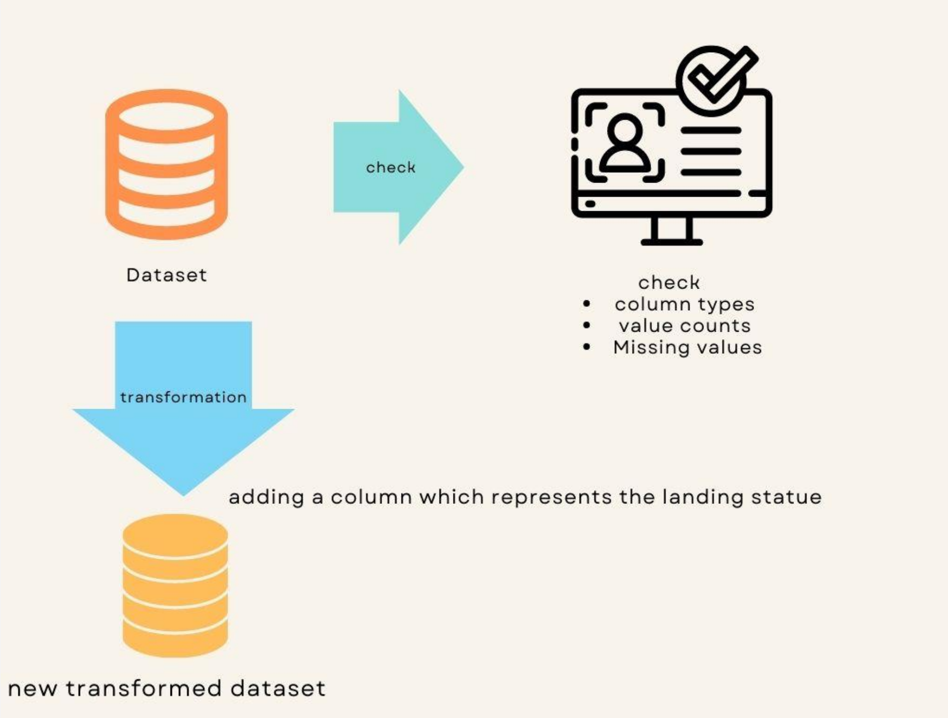
- Perform a request to this link and retrieve the HTML page content.
- Extract all column/variable names
- Create a data frame by parsing the launch HTML tables

Follow this link for more information.



Data Wrangling

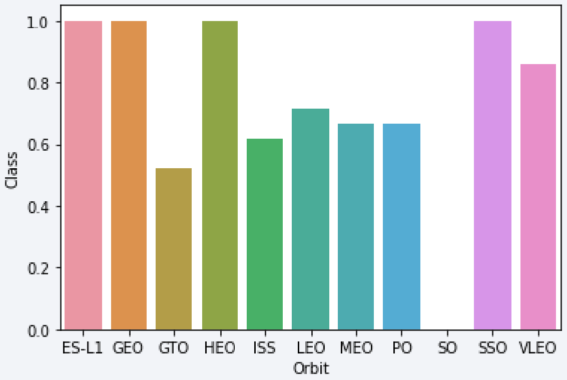
- Check data types
- Check value counts and missing values
- Adding a new column "class" which takes 1 if the landing success and 0 otherwise.

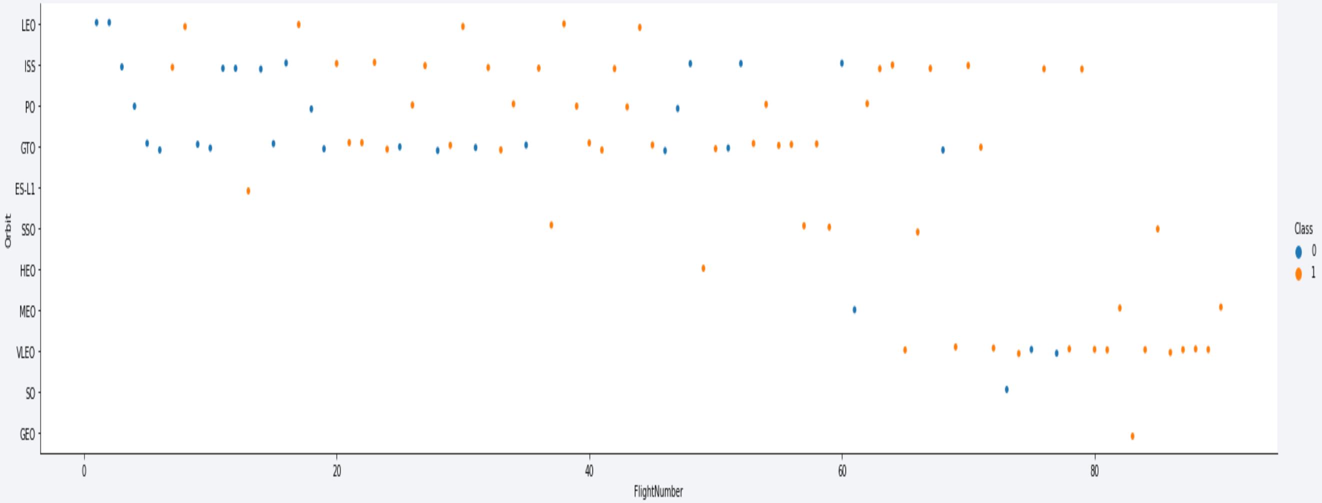


More details can be found at this github repo.

EDA with Data Visualization

- Visualizing tools that we used :
  - Scatter Plot : Visualize the relation between flies parameters.
  - Bar chart : To see the relationship between success rate of each orbit type.
  - More infos can be found in our github repo.





EDA with SQL

Perform different SQL queries in order to:

- *Display the names of the unique launch sites in the space mission*
- *Display 5 records where launch sites begin with the string 'CCA'*
- *Display the total payload mass carried by boosters launched by NASA (CRS)*
- *Display average payload mass carried by booster version F9 v1.1*
- *List the date when the first successful landing outcome in ground pad was acheived*
- *List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
- *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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*
- *Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order*

Follow this link to the [github repository](#) for more details.

Build an Interactive Map with Folium

Map objects used in the map creation :

- Circle : draw a circle around launch sites in order to locate them.
- Marker : Stacked on the circle, display the name of each launch site.
- Marker Cluster : An object that cluster markers that contains colored icons, when green ones represents the successful landings and the red ones the failed landings. Each launch site own a marker cluster.
- Lines : Line which represents the distance between a launch site and the nearest coast/railway/highway/city.

You can find the file in the github repository .

Build a Dashboard with Plotly Dash

Plotly components of the dashboard :

- A dropdown menu : to select a specific launch site or all of them.
- A pie chart : which represents the repartition of successful landing among the launch sites (when all sites are specified in the dropdown menu) or the Success/Failure repartition of the landings in a particular launch site (when only one launch site is selected).
- A range Slider : to specify the range of payload mass of the launches that we want to visualize in the scatter plot.
- A scatter plot : which represents the repartition of successful and failed launches by payload mass in one or all of the launch sites.

The script is available in the github repository .

Predictive Analysis (Classification)

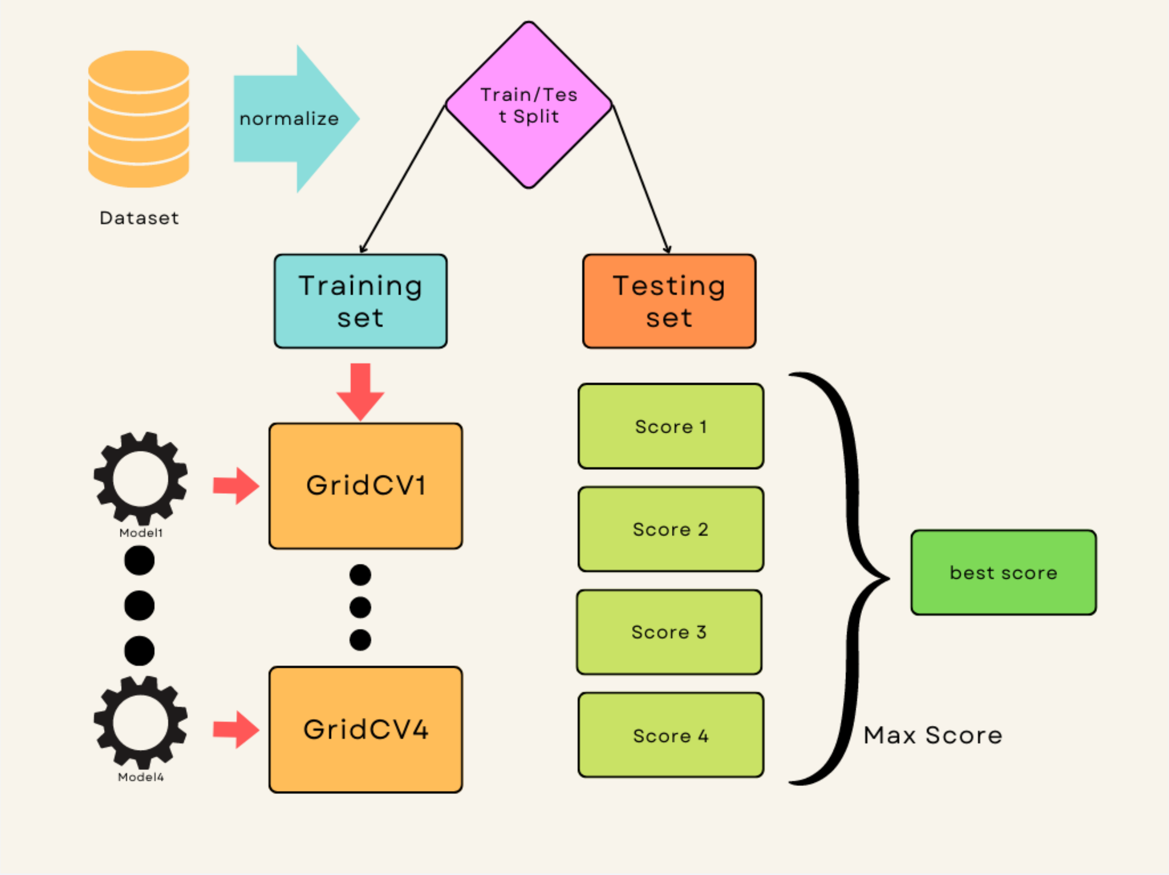
Data Classification Pipeline :

- Standardize the data
- Train/Test Split
- For each model of this list (logistic regression, SVM, KNearestNeighbor, Decision tree classifier):
  - Build a GridCV model which will contain the model and fit it with the best parameters using cross validation to increase its precision.
  - Test with the data in the test set and get the score.
  - Plot the confusion matrix of each model
- Determine the model with highest accuracy.

This work can be found in the github repository

Predictive Analysis (Classification)

Data Classification Flowchart



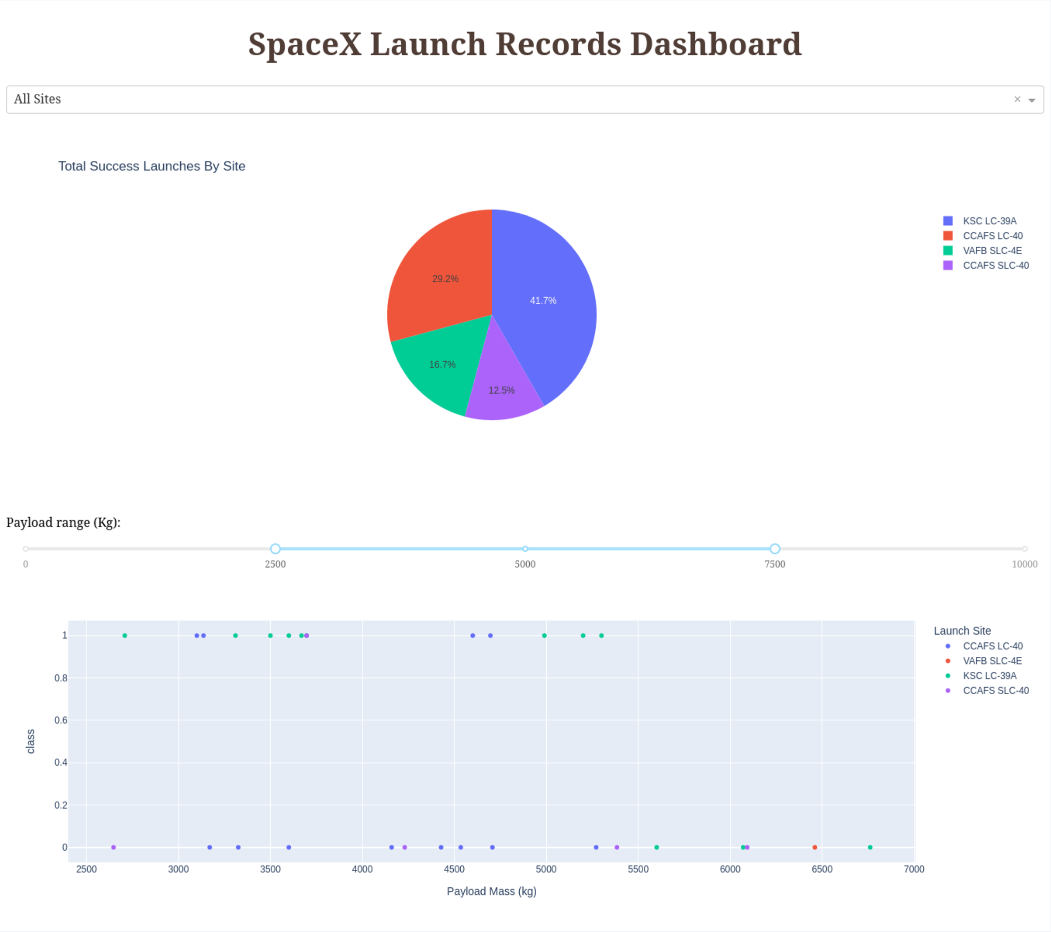
Results

Exploratory data analysis results

- CCAFS SLC 40 launch site have lowest success rate (60%) but a higher number of attempts.
- No rockets
  are launched for heavypayload mass(greater than 10000) from the VAFB-SLC launchsite.
- Rockets which are launched on ES-L1, GEO, HEO,SSO,VLEO orbits have a success rate higher than 0.8.
- For heavy payloads, the successful landing rate are more for Polar,LEO and ISS.
- In the LEO orbit, the Success appears related to the number of flights.
- Sucess rate kept increasing since 2013.
- Launch sites are close to the coast, have near railways, far from highways and cities.

Results

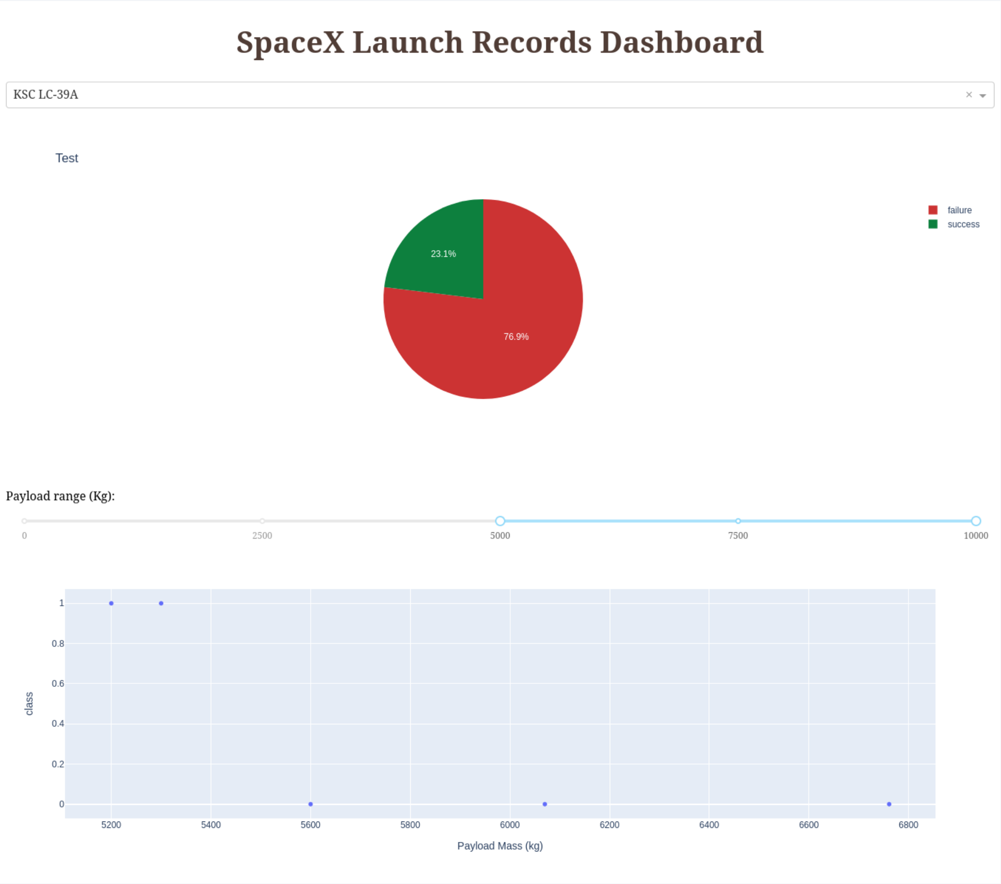
Interactive analytics demo



SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):



SpaceX Launch Records Dashboard

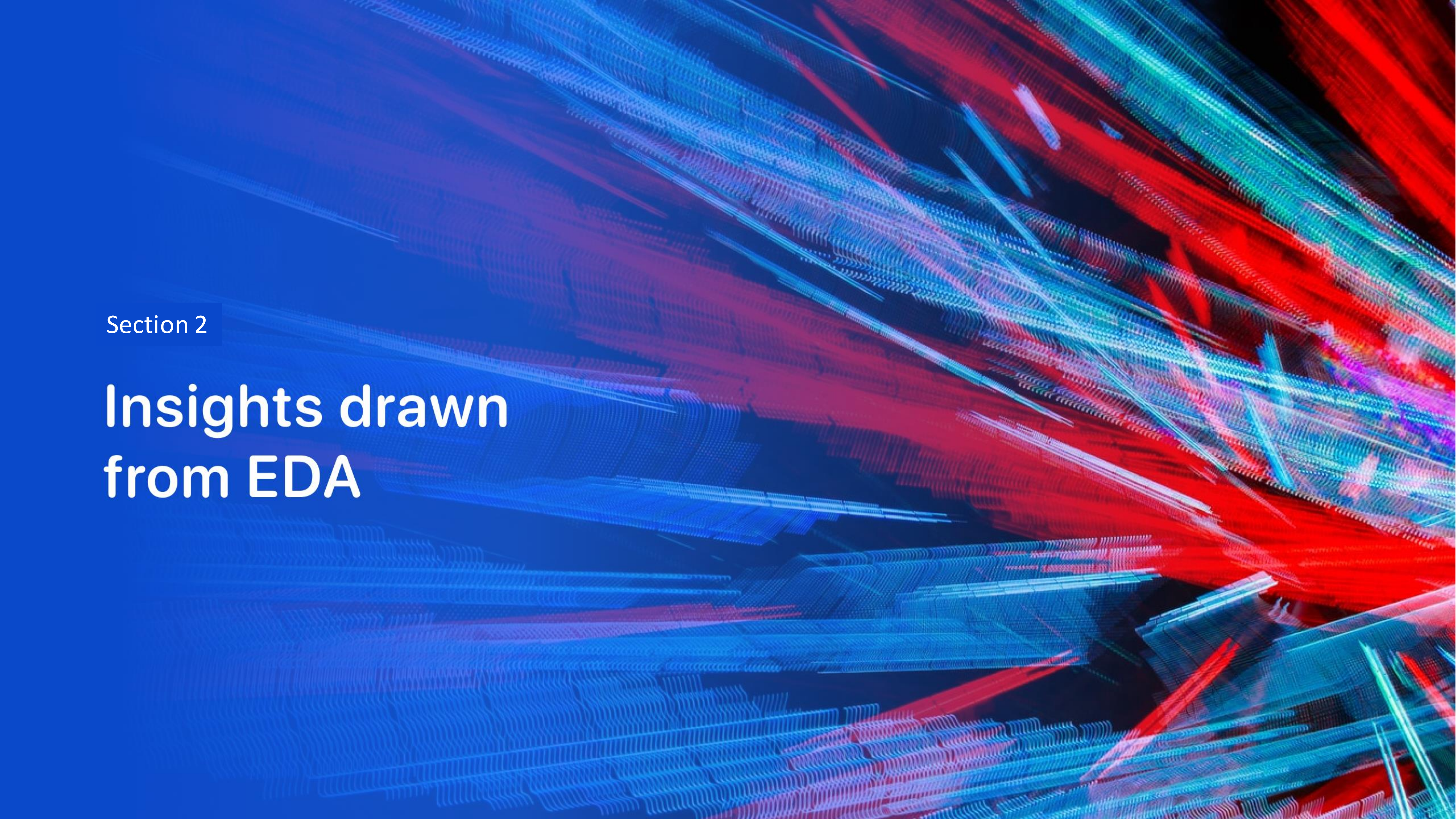
KSC LC-39A

Payload range (Kg):

Results

Predictive analysis results

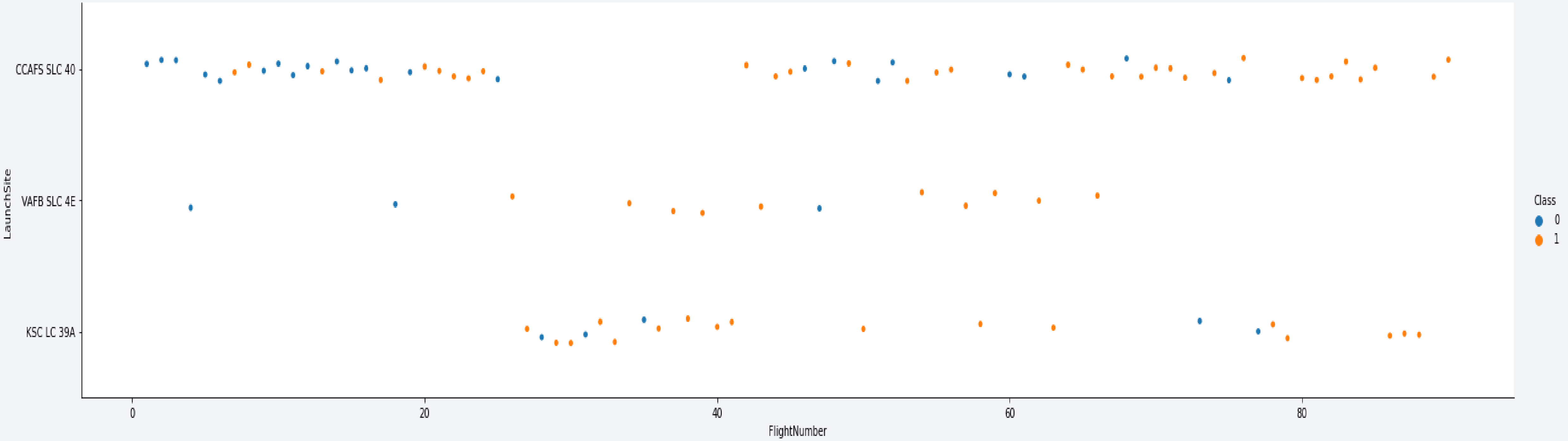
Model	Accuracy
Logistic regression	0.83
SVM	0.78
Decision Tree	0.78
K Nearest Neighbour	0.83



Section 2

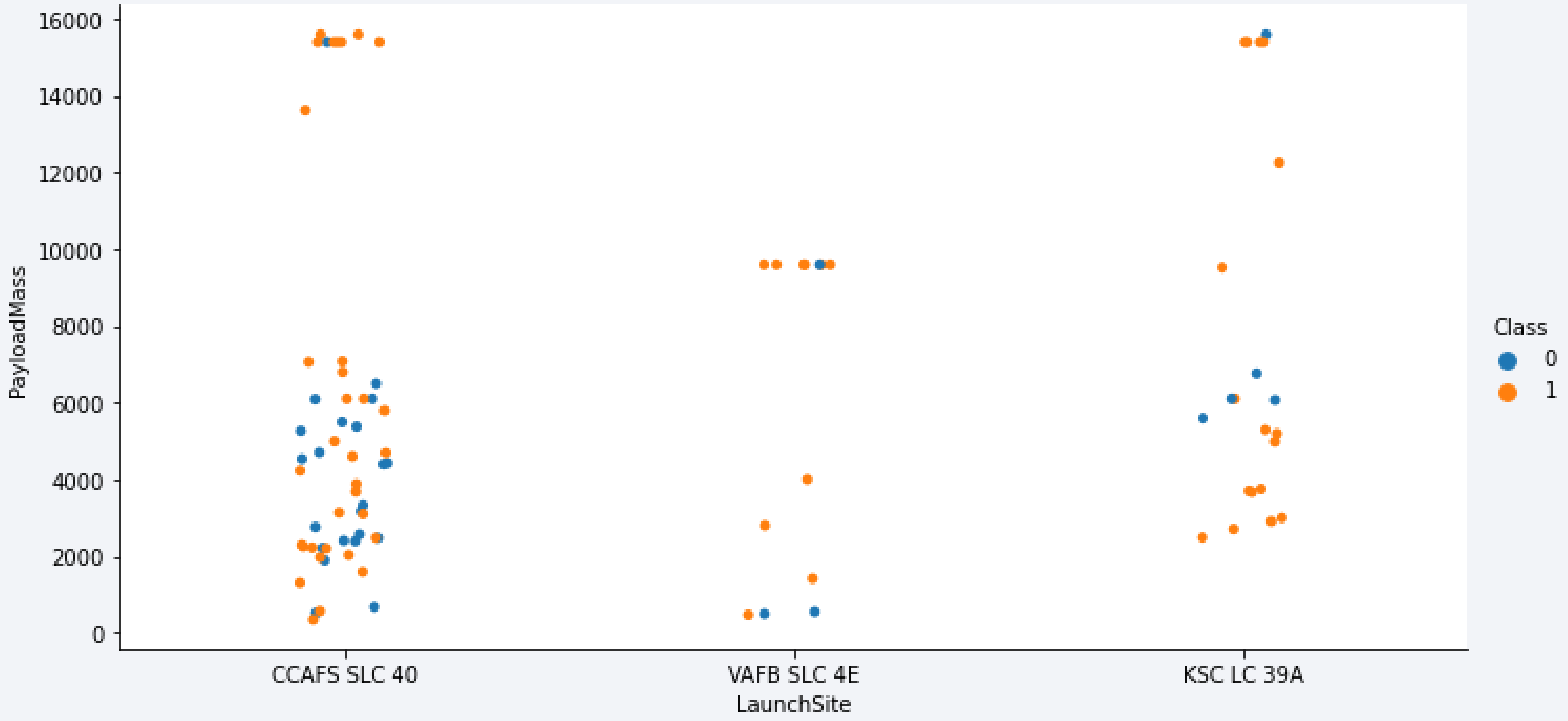
Insights drawn from EDA

Flight Number vs. Launch Site



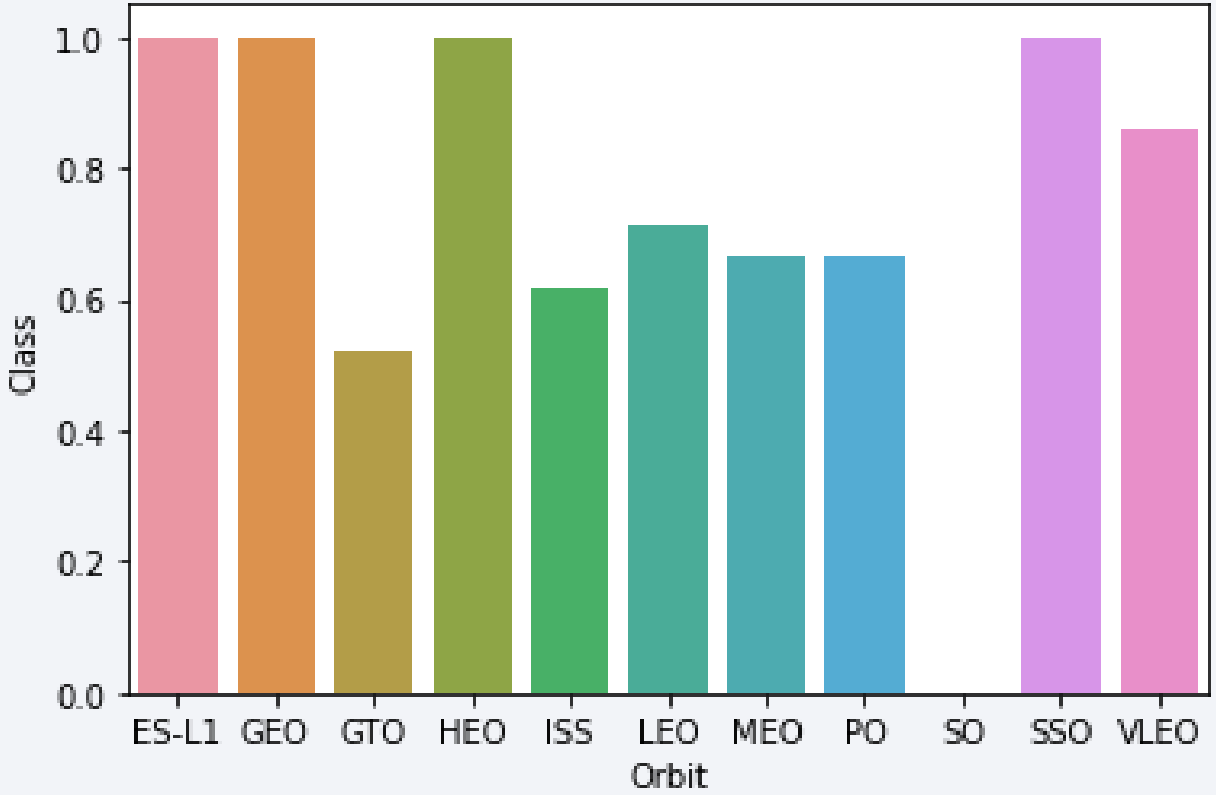
- CCAFS SLC 40 was the first launch site and the most used until today.
- The exploitation of the KSC LC 39A launch station begun after the 27th launch site.
- The VAFB SLC 4E station is the least used but has the highest rate of success landing.

Payload vs. Launch Site



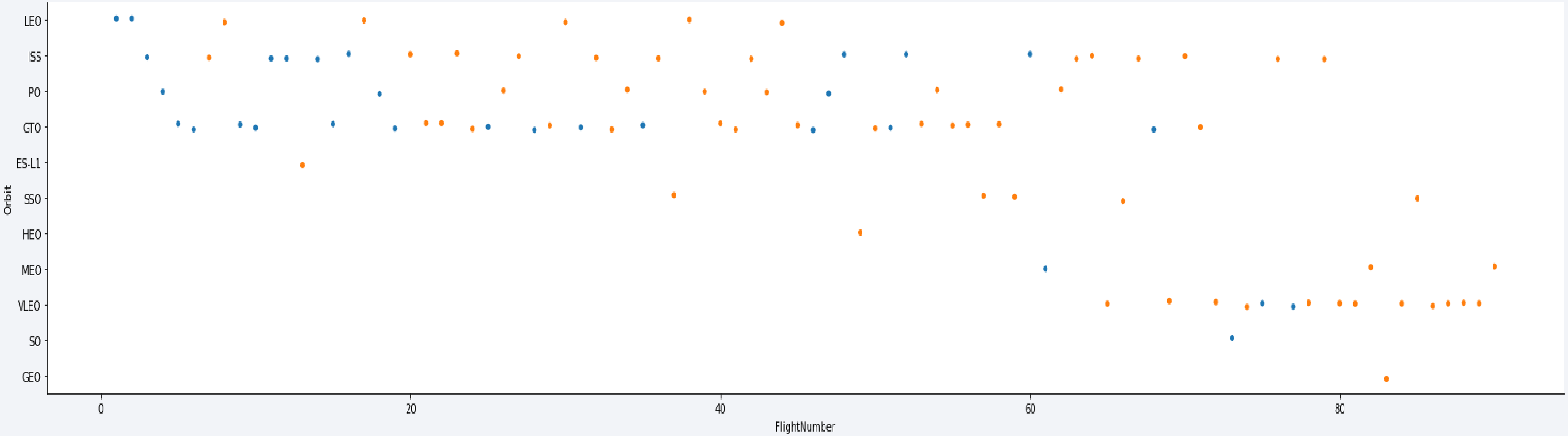
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- Most of the payloads weigh less than 8000 kg.
- The heavypayloads have a high sucess rate.

Success Rate vs. Orbit Type



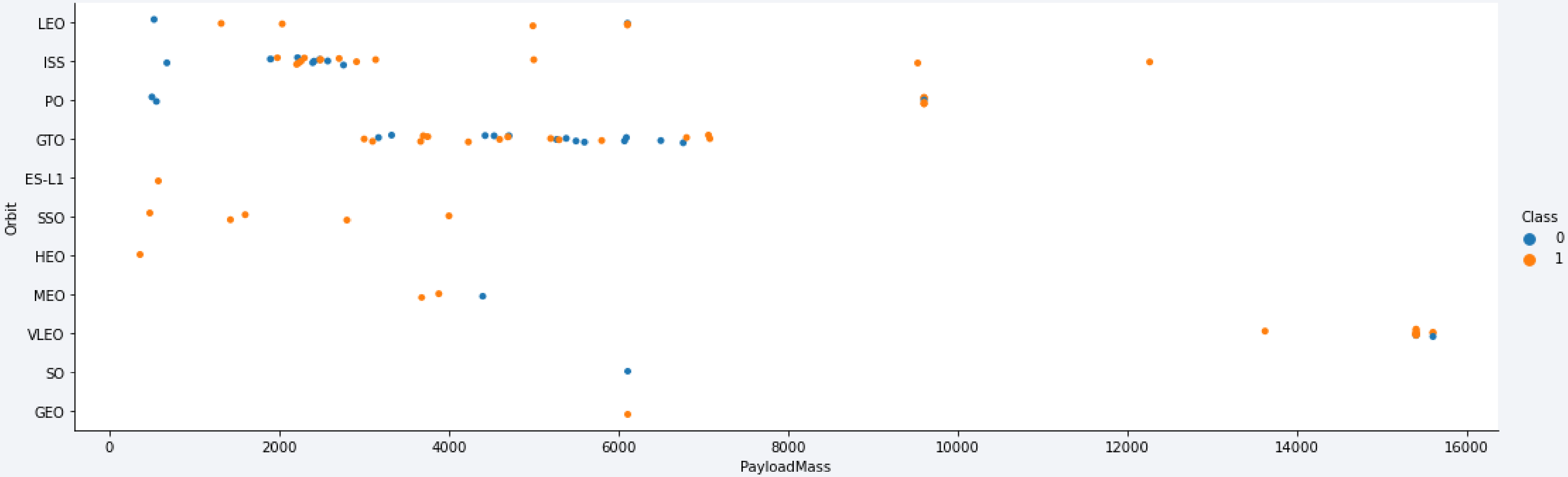
- The orbits ES-L1,GEO,HEO,SSO have a perfect success rate of 1.0.
- VLEO orbit has a good success rate near 0.8.
- Other orbits have a success rate between 0.5 and 0.7.
- Very few data of the SSO orbit.

Flight Number vs. Orbit Type



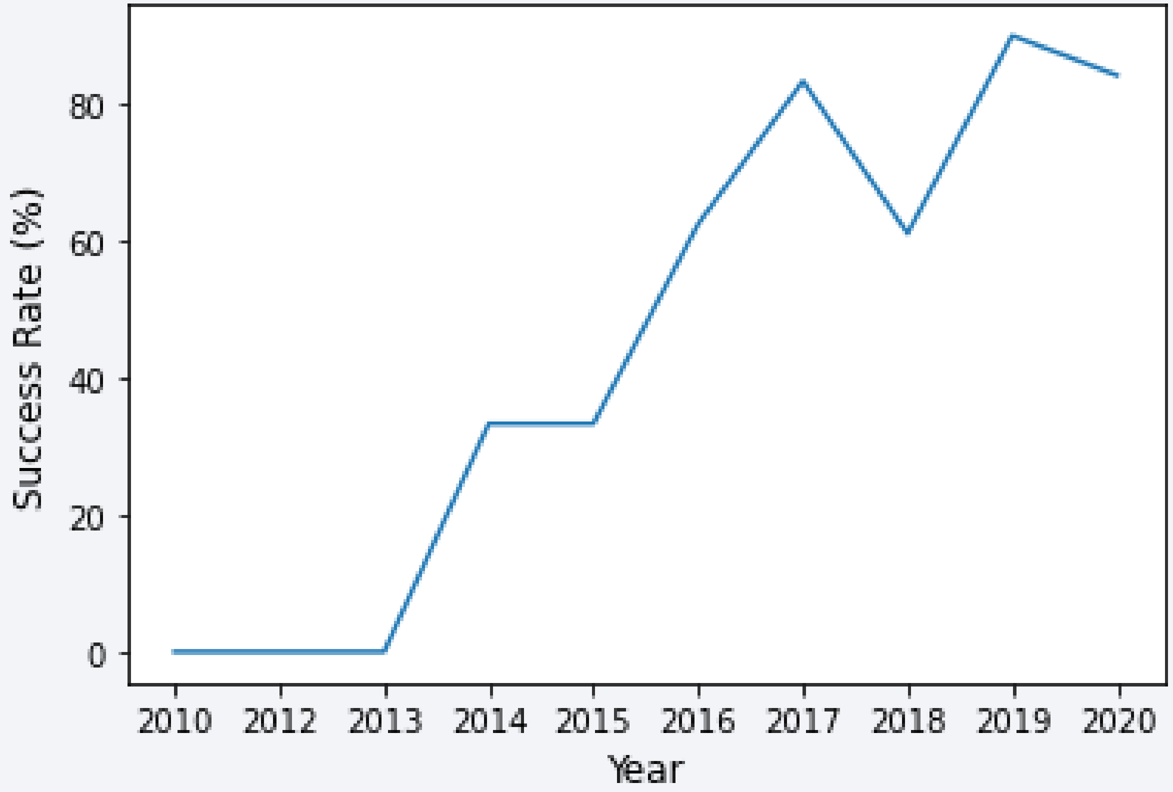
- Most of the launches were made on the GTO and ISS orbits.
- From the 65th launch, most of launches are made on VLEO and the number of launches on GTO decrease.

Payload vs. Orbit Type



- Most of the rockets that have payload mass between 3000 and 8000 are launched on the GEO site.
- Most of the rockets that have a payload mass between 2000 and 3000 are launched on the GTO site

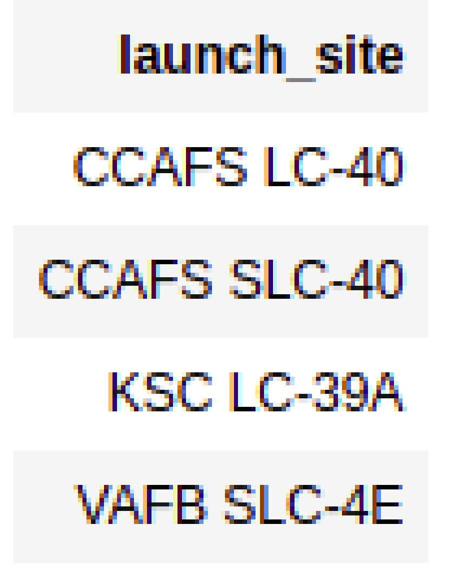
Launch Success Yearly Trend



- Success rate since 2013 kept increasing till 2020.
- Since 2016, the success rate is higher than 60%.

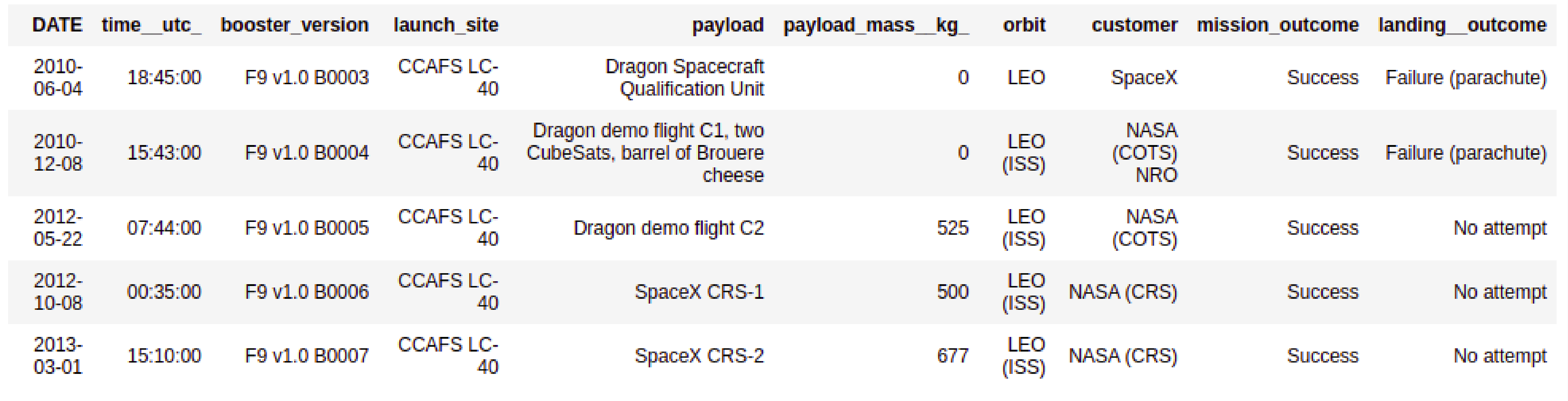
All Launch Site Names

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



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

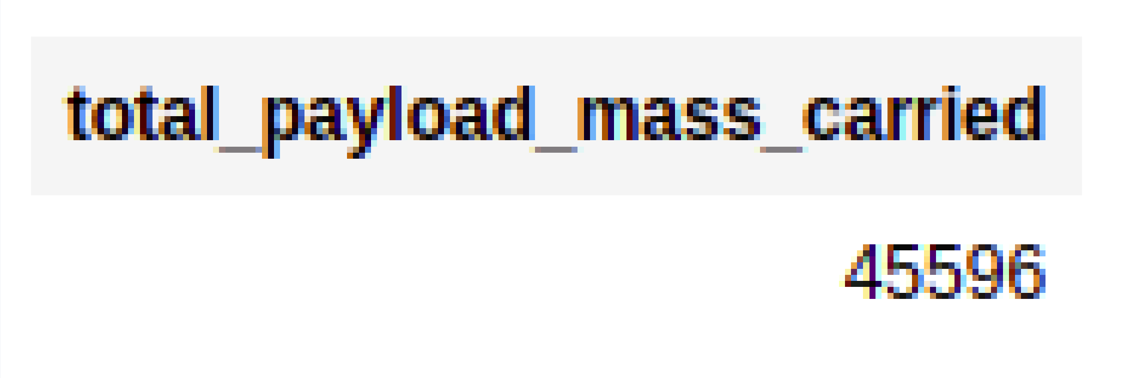
Launch Site Names Begin with 'CCA'



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	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

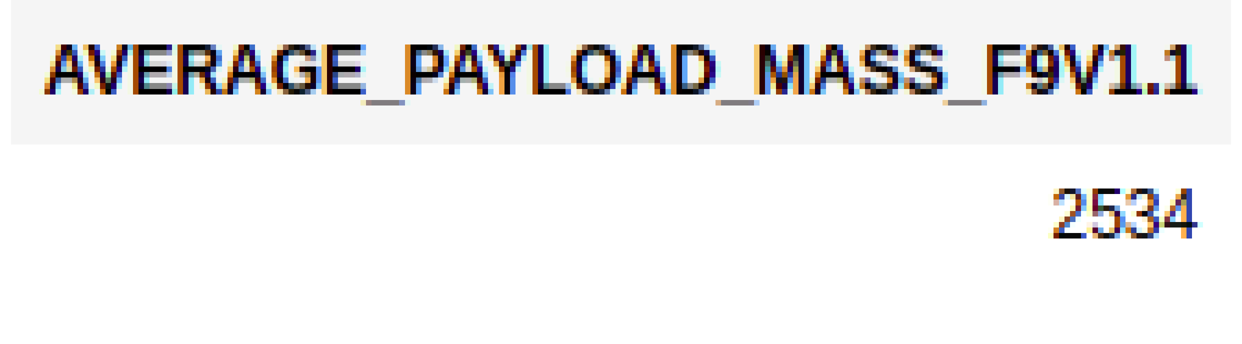
- CCAFS LC-40
- CCAFS SLC-40

Total Payload Mass



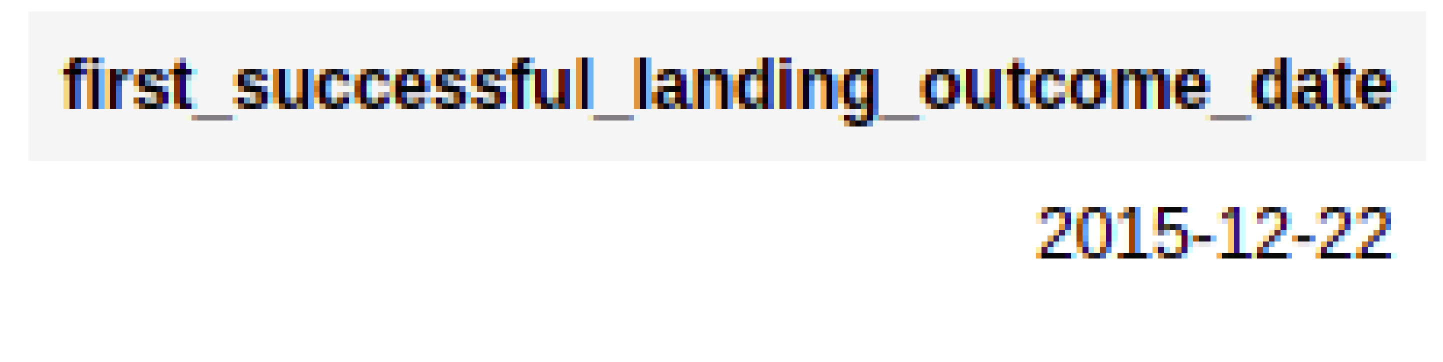
total_payload_mass_carried
45596

Average Payload Mass by F9 v1.1



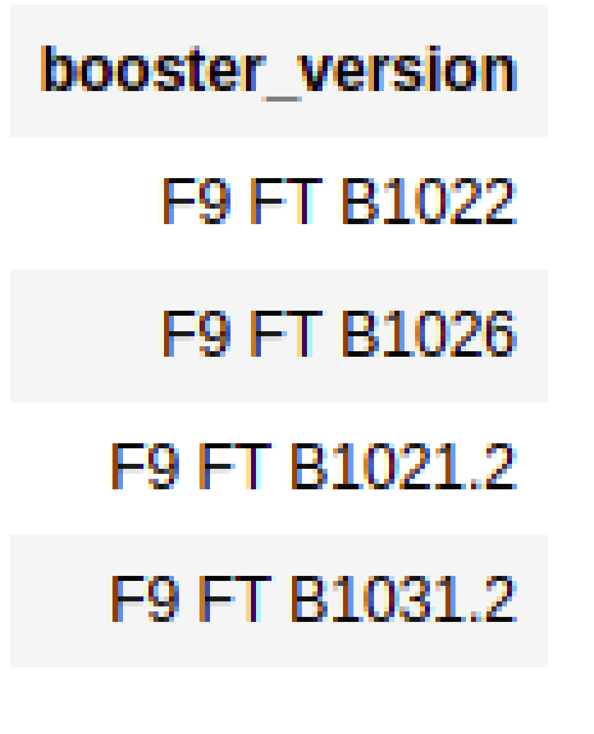
AVERAGE_PAYLOAD_MASS_F9V1.1
2534

First Successful Ground Landing Date



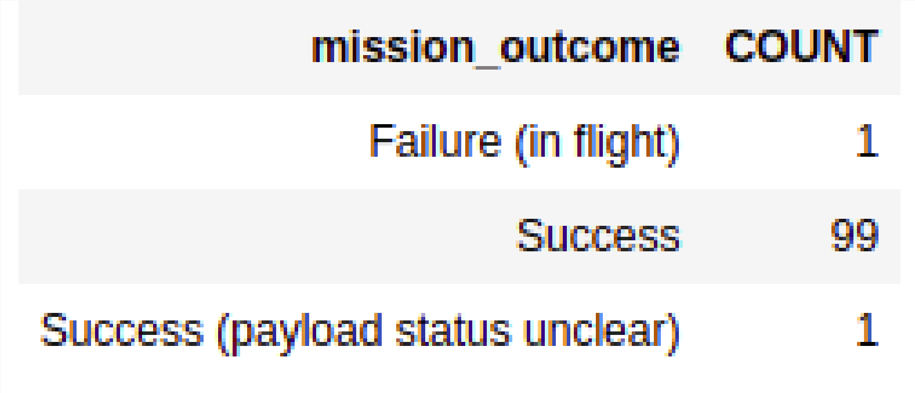
first_successful_landing_outcome_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



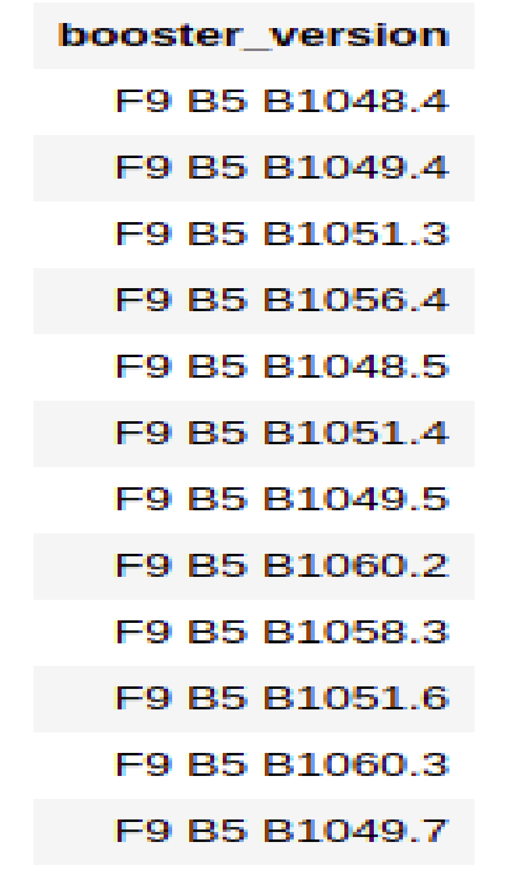
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



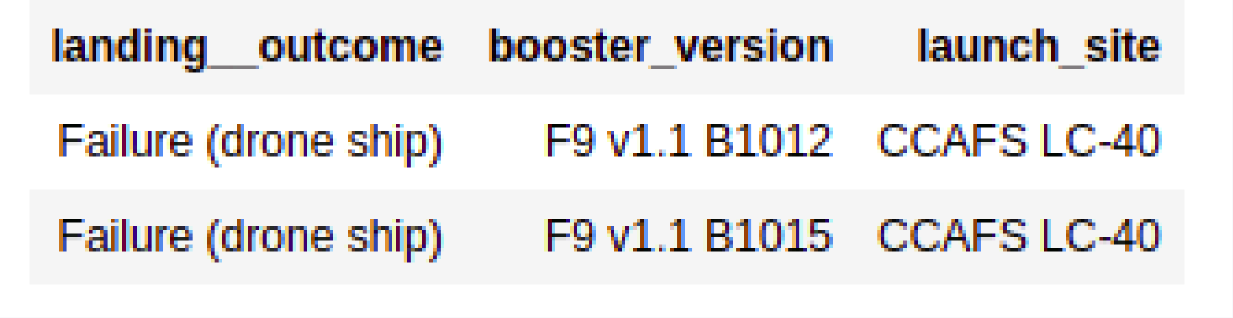
mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload



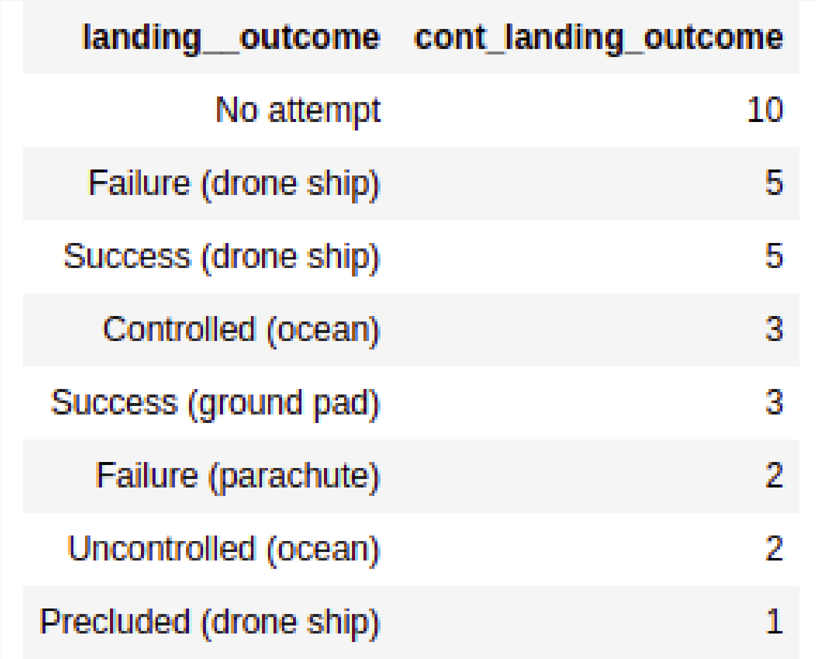
booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records



landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

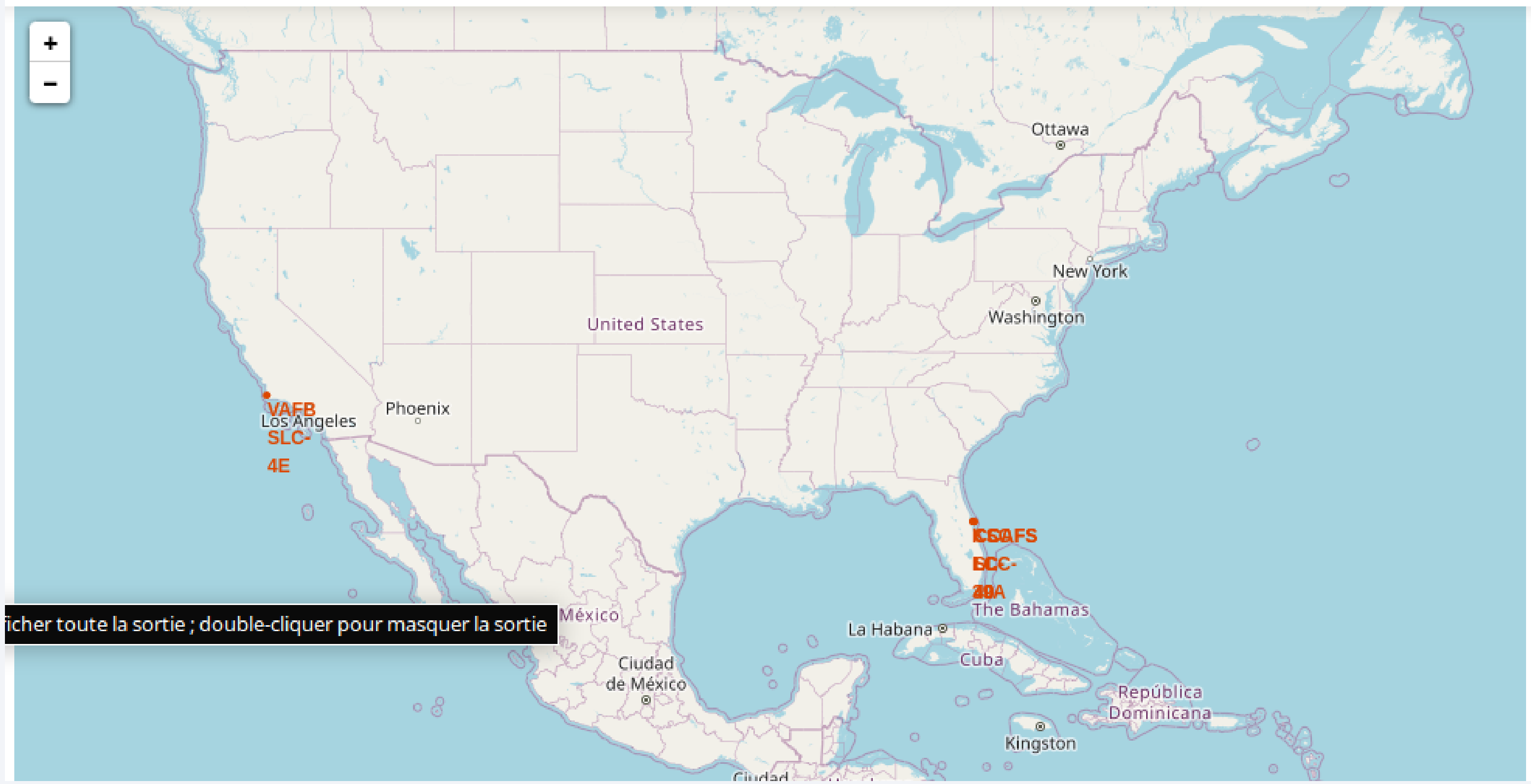


landing__outcome	cont_landing_outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

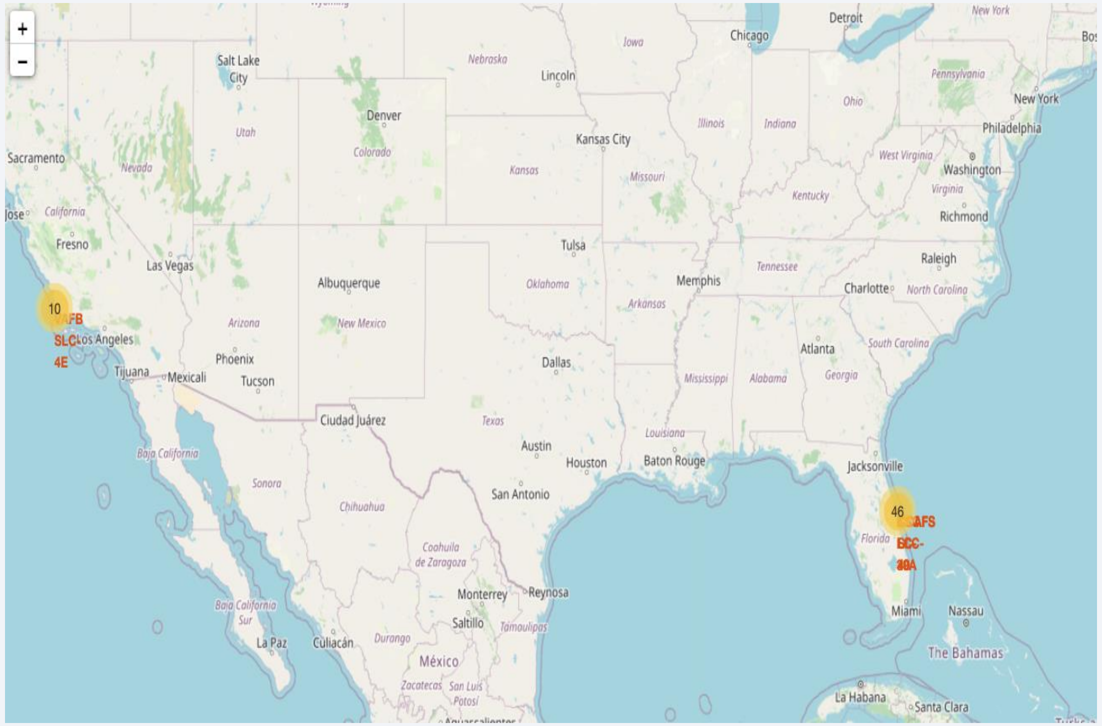
Launch Sites Proximities Analysis

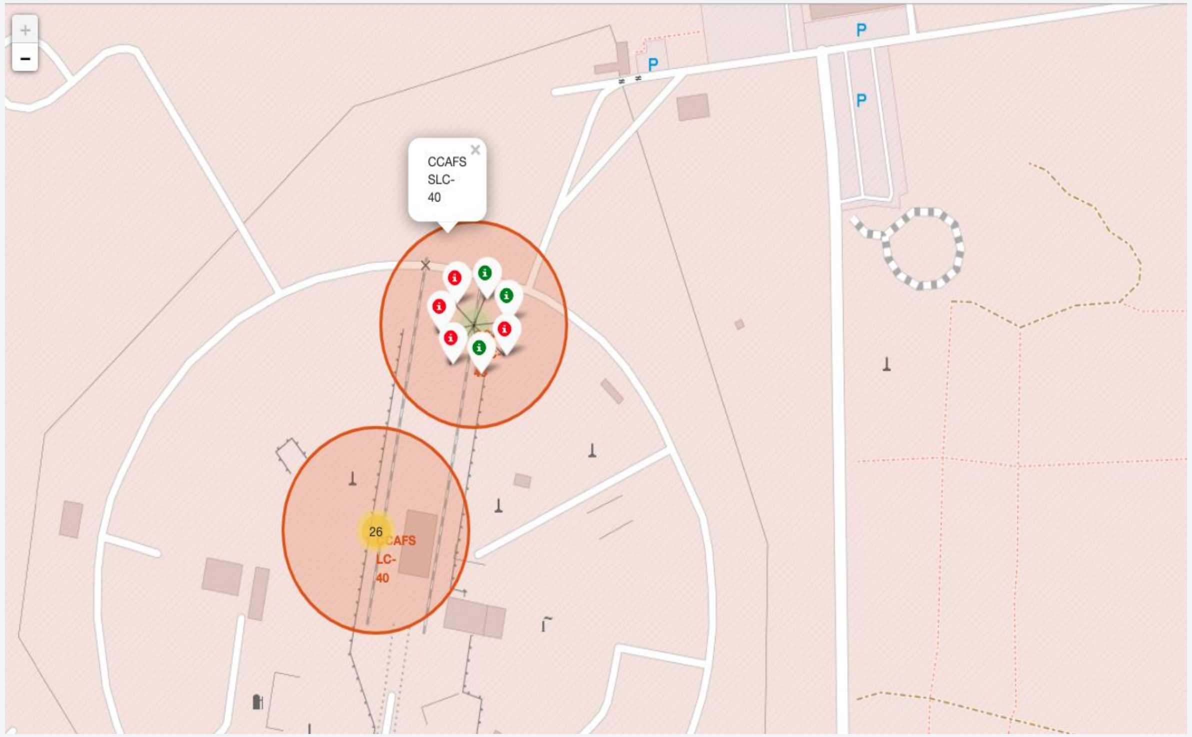
Launch Sites on the Map



This map shows the repartition of the launch sites in the U.S map

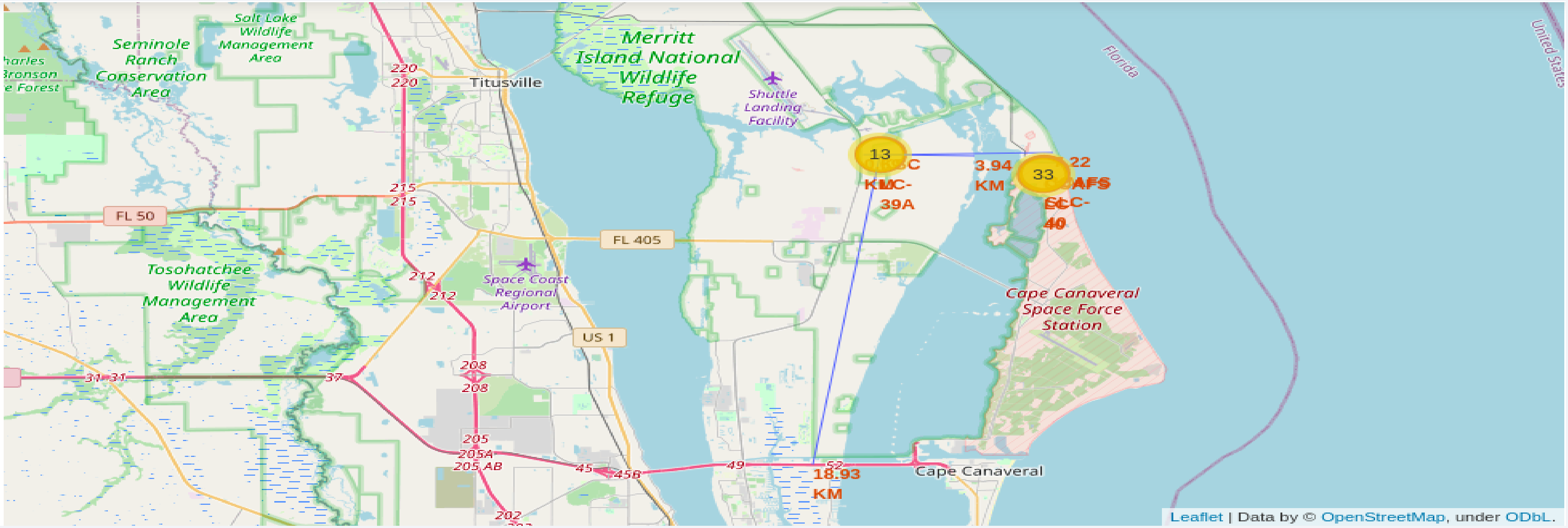
Success/Failed Launches For each Site on the Map



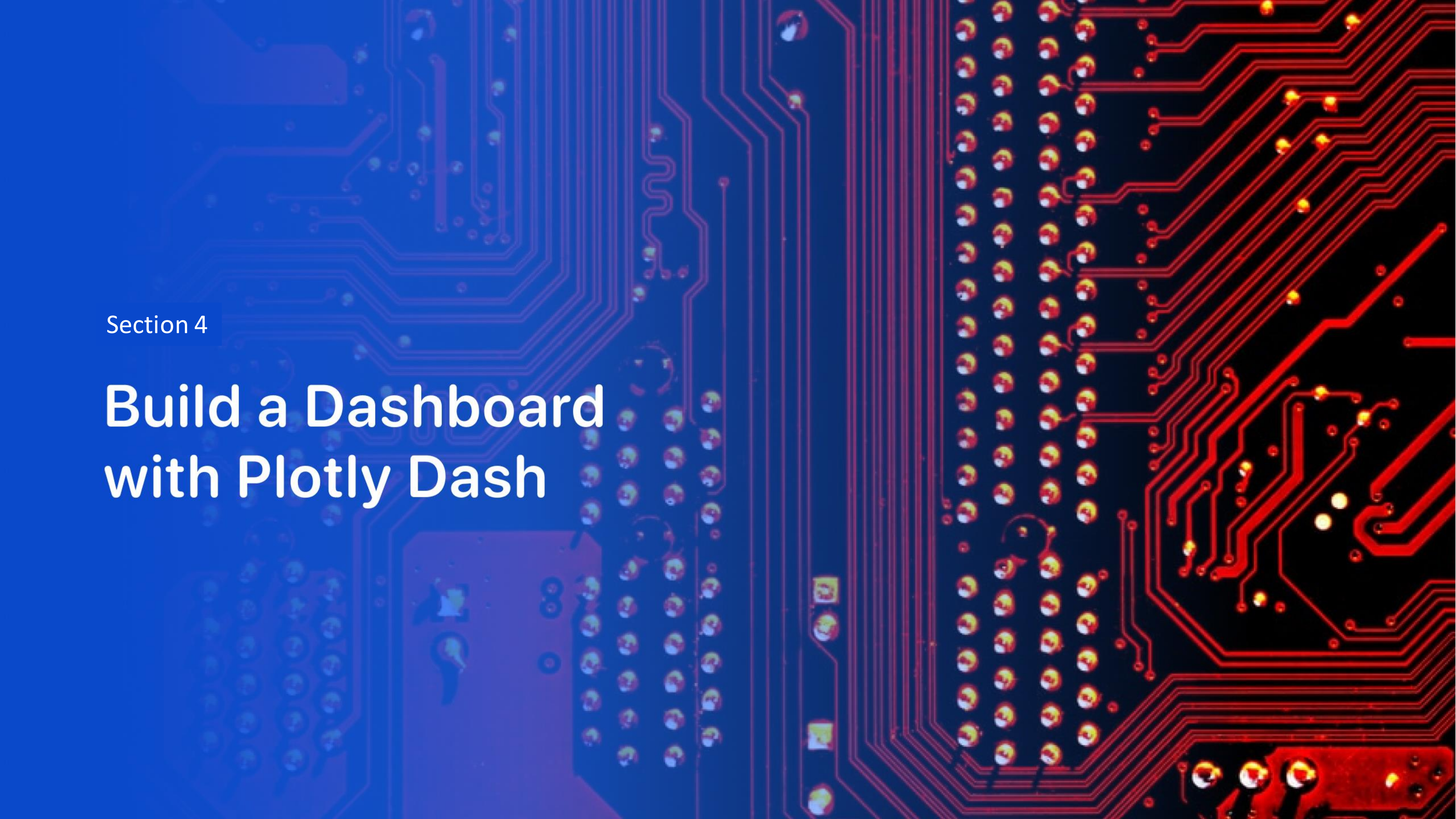


In this map, we can visualize success and failure launches for each launch Site

Distances between a Launch Site to its Proximities


Salt Lake Wildlife Management Area
Seminole Ranch Conservation Area
harles Bronson e Forest
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
Titusville
220
215
FL 50
FL 405
Tosohatchee Wildlife Management Area
212
Space Coast Regional Airport
US 1
31
37
208
205
205A
205 AB
45
456
49
202
13
C
KMC-
39A
3.94
KM
33
22
AFS
SLC-
40
Florida
United States
Cape Canaveral Space Force Station
18.93
KM
52
Cape Canaveral
Leaflet | Data by © OpenStreetMap, under ODbL.

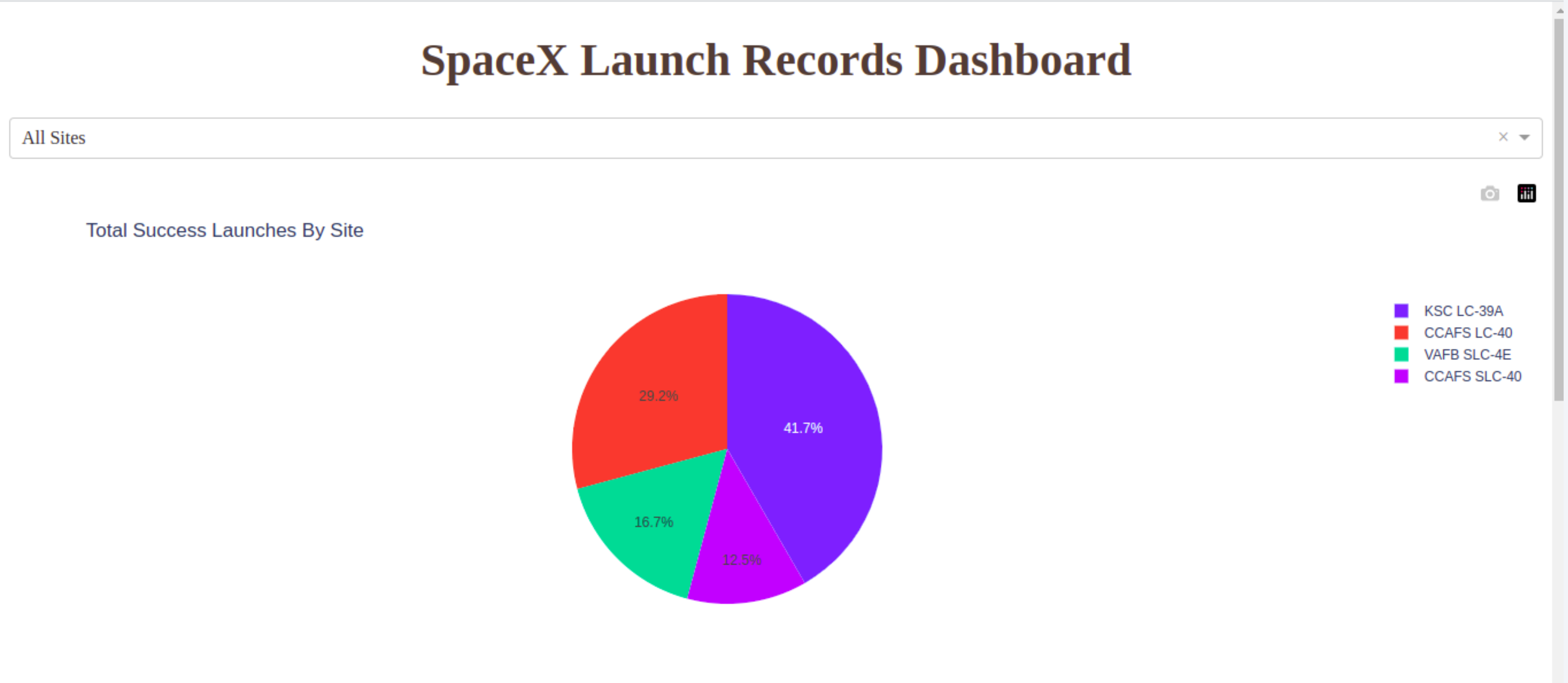
- Launch sites are closer to coastline in order to minimize damages in cases of crashes of the first stage.
- Launch sies are close to highways in order to transporting materials into the station.
- Launch sites are far from cities and highways in order to avoid human damages when the landing fails.



Section 4

Build a Dashboard with Plotly Dash

Total Success Launches By Site

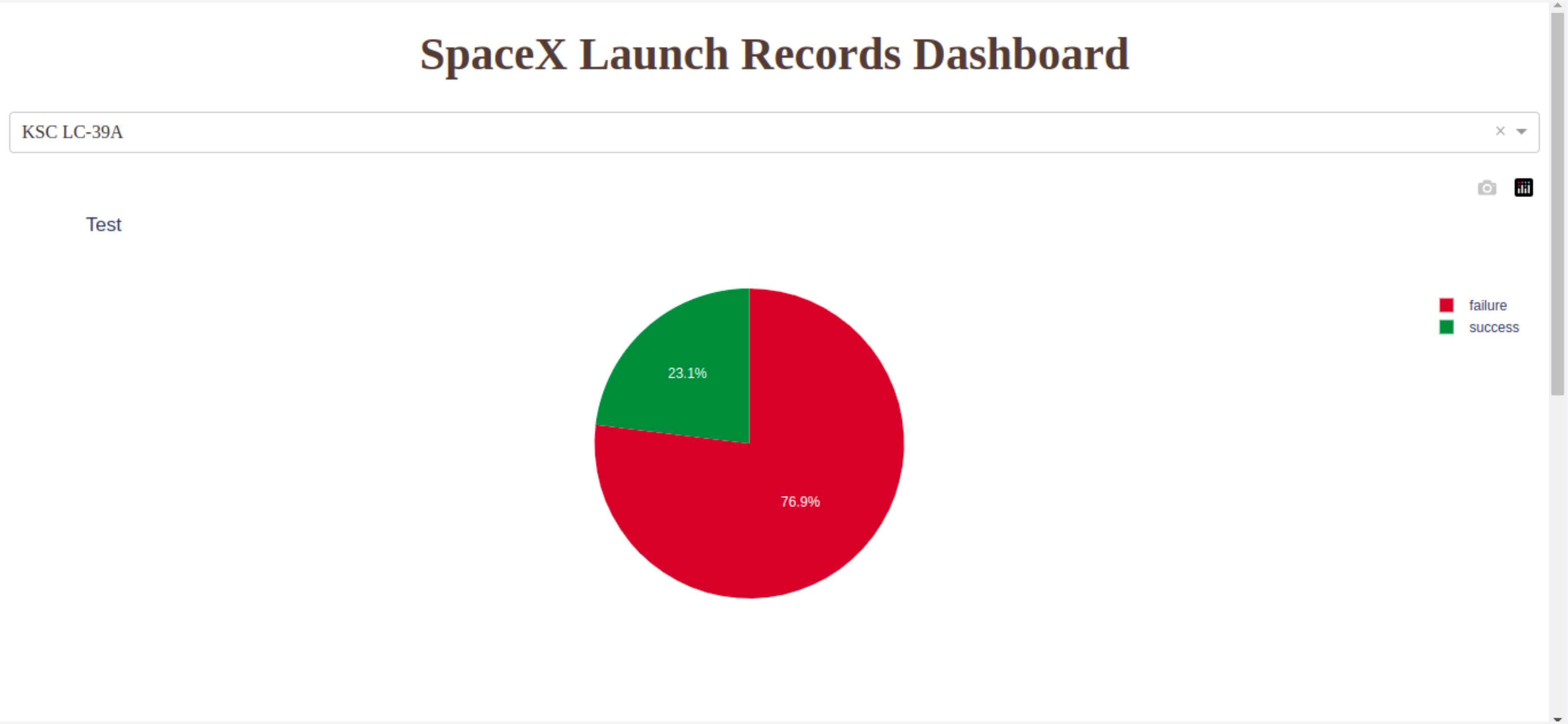


SpaceX Launch Records Dashboard

All Sites

- We notice that the majority of success launches are made on the KSC LC-39A

Success Ratio of the KSC LC-39A Launch Site

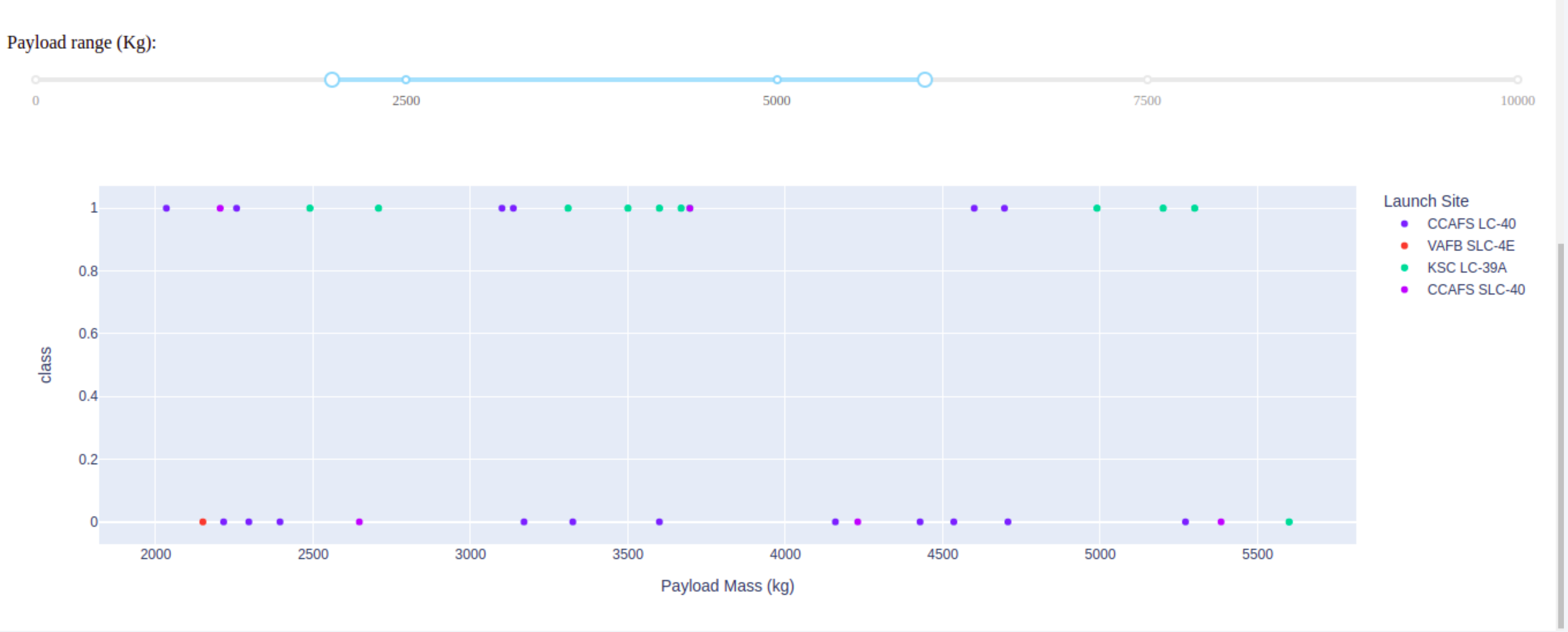


SpaceX Launch Records Dashboard

KSC LC-39A

- We notice that about 77% of the launches made on this site succeed

Distribution of Successful Launches By Payload Mass

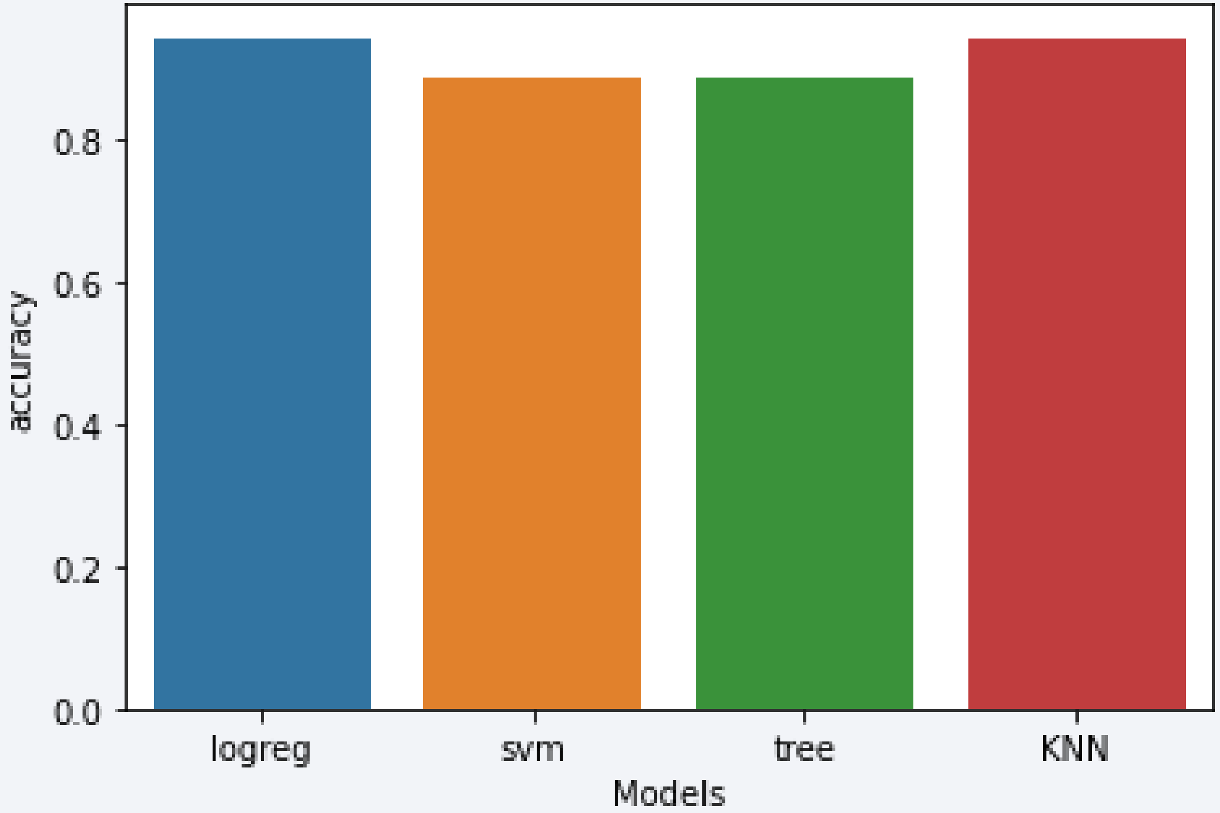


- We notice that the range where success launches rate is higher is the range 2000 - 5000.

Section 5

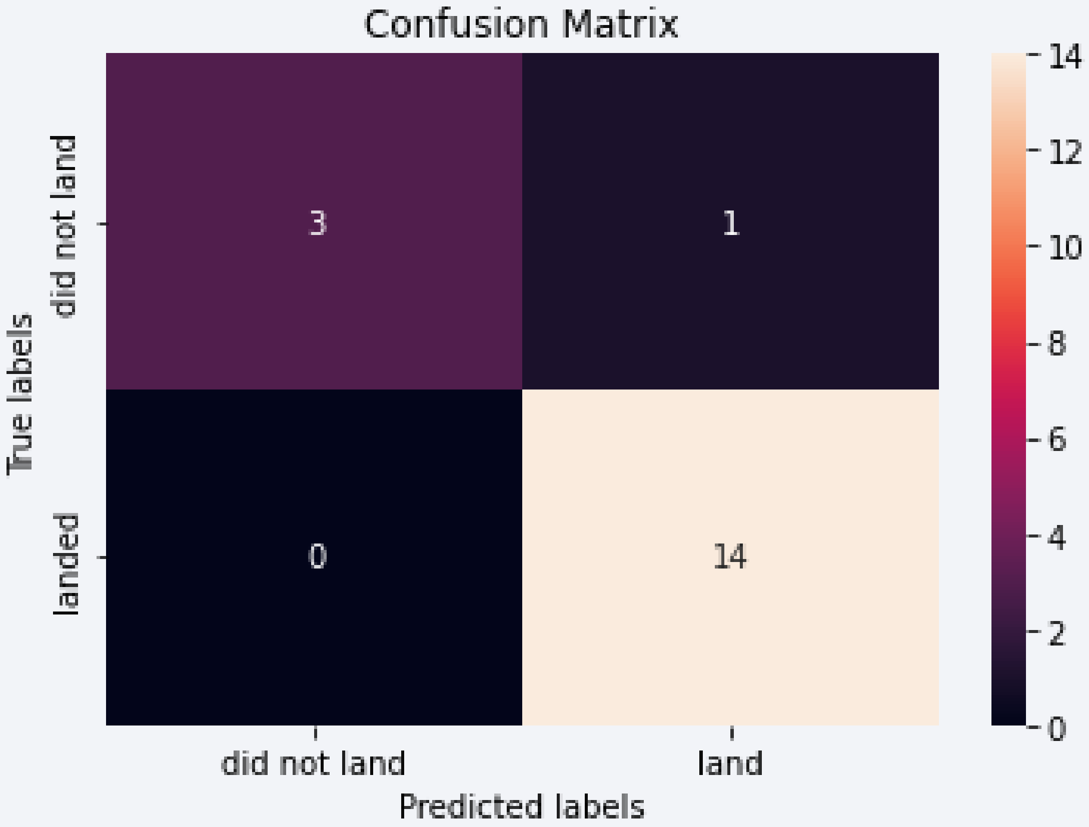
Predictive Analysis (Classification)

Classification Accuracy



- The model that perform better are logreg and KNN with an accuracy of 8.33.
- The other models have also good scores (0.78)

Confusion Matrix



- We see that the model performs well detecting the launches that did not land (3/3).
- But it lacks some accuracy when predicting the launches that will land.
- This problem may be fixed with preparing a bigger training set with more data of failure value.

Conclusions

- CCAFS SLC 40 was the first used launch site and the most used until today.
- For the VAFB-SLC launchsite there are no
  rockets launched for heavypayload mass(greater than 10000).
- The orbits ES-L1,GEO,HEO,SSO have a perfect success rate of 1.0.
- Most of the launches were made on the GTO and ISS orbits, but for the last 30 launches, the most used is VLEO
- Success rate since 2013 kept increasing till 2020.
- Launch sites are close to coastlines and railways and far from highways and cities.
- majority of success launches are made on the KSC LC-39A (approximatively 77% of success rate)
- Our best model (SVM) has an accuracy rate of 0.83 which is very good and can be higher with a bigger dataset.

Appendix

- A great thanks to the IBM Skills Network team which made this amazing certificate possible and for all the authors.

Thank you!