IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fabian Pedreros
Jan 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Web scraping and API use to data collection.
  - Exploratory Data Analysis (EDA) to understand and visualize the collected data.
  - Supervised Machine Learning algorithms to classify the results of a landing (success or not)
- Summary of all results
  - It is possible to create a Machine Learning model that helps to predict if a SpaceX landing is successful or not.

Introduction

- Project background and context
  - SpaceX is a US aerospace manufacturing and space transportation services company, is at the forefront of the industry because its launch costs are the lowest in the market, in large part due to the fact that it reuses the first state. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. For the project are taking in account the Falcon 9 rockets.
- Problems you want to find answers
  - What attributes of a landing can be used to predict if is successful or not?
  - What is Machine Learning model that can predict with the major precision if a landing is successful or not?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collection of the Falcon 9 landings is done using two sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Public data from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - Summarized data analysis is performed to understand the distributions of landings by orbit and launch sites, as well as the creation of the dependent variable 'Class', which identifies whether a landing was successful (1) or not (0).

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Normalization and data encoding was need it due to the nature of the variables.
  - Training and test datasets have been created for the independent variables that explain whether a spacecraft lands or not.
  - Three different classification algorithms (SVM, Classification Trees and Logistic Regression) are trained.
  - Grid Search is used to establish which are the best hyper parameters that give the highest accuracy for each model.
  - The selection of the best model is done by selecting the highest accuracy..

Data Collection

The data sets were collected using two different sources and techniques:

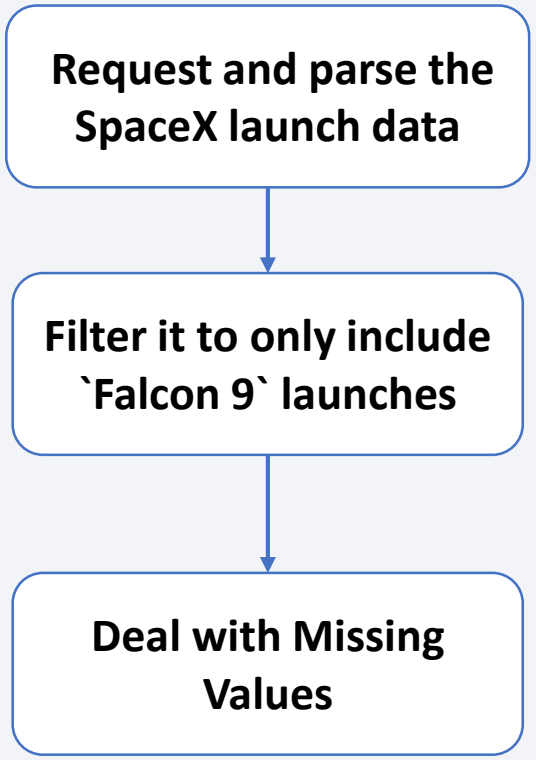
- API request. For SpaceX data launch using a service provided for the enterprise.

- Web Scraping. To request public data available in the web, where we can find information about the rockets.

Data Collection – SpaceX API

- It is executed a data extraction using the API public services for SpaceX. This data is requested initially in Json format and then transformed and filtered into a DataFrame.

Code:
https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

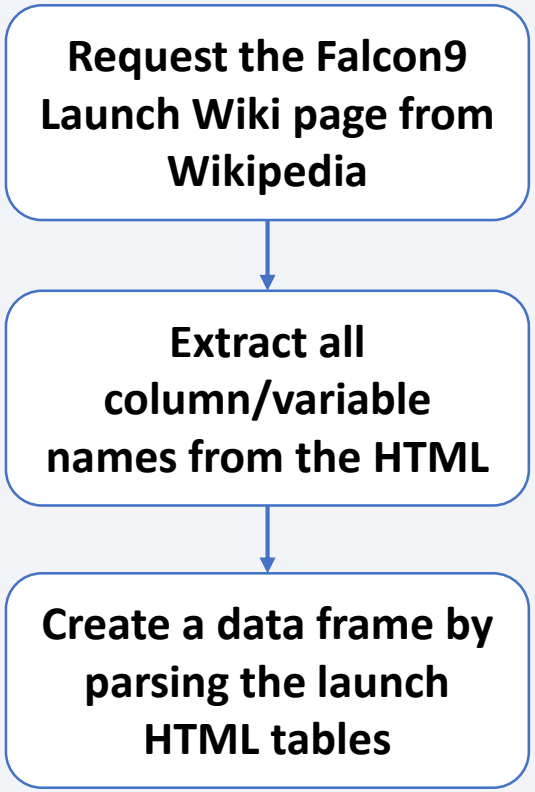


Data Collection - Scraping

- Web scrap Falcon 9 launch records is done using `BeautifulSoup`, extracting the Falcon 9 launch from the HTML Table and then convert it to a data frame.

Code:

https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

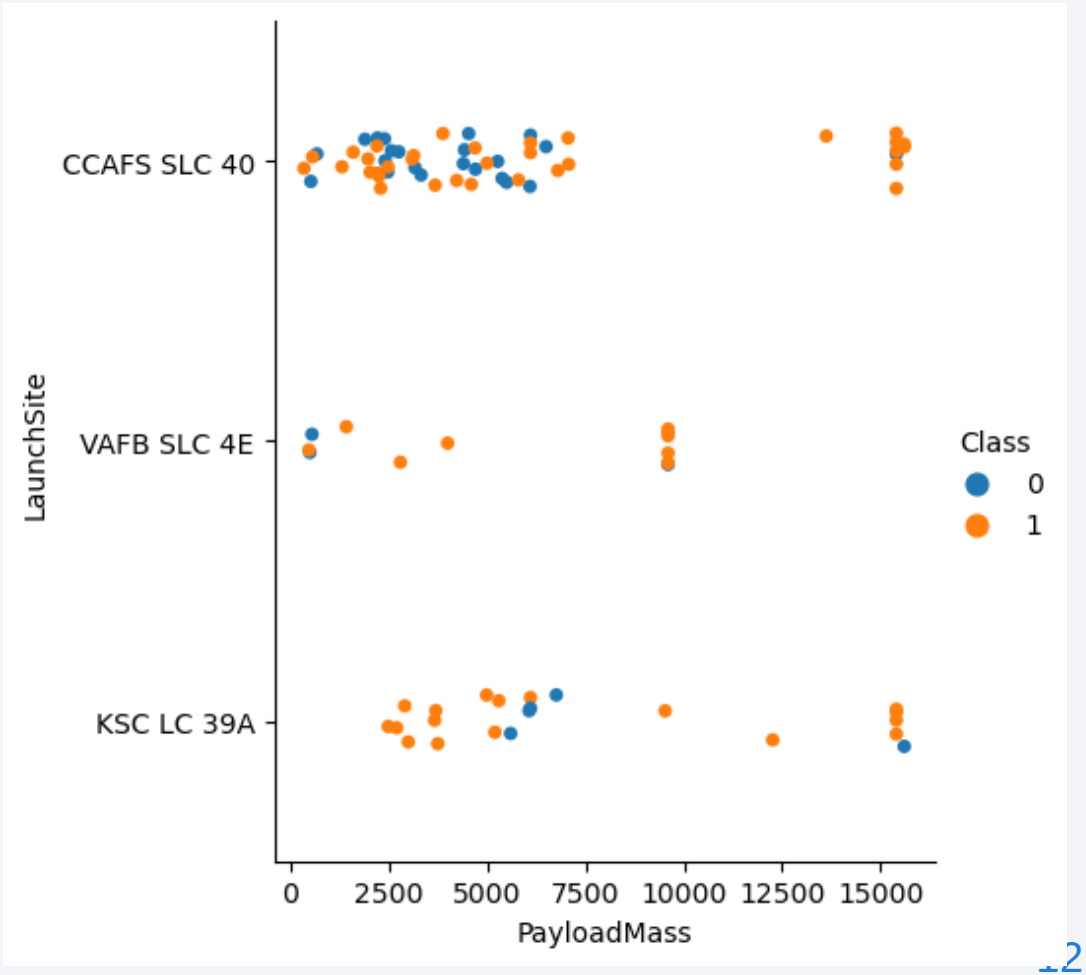
In order to understand the data and find relevant attributes than can explain the landing success, is done a EDA to explore relationships via visualization as:

- Payload mass vs flight number
- Site number vs Launch site
- Payload mass vs Launch site
- Orbit success
- Flight number vs Orbit Type
- Orbit vs Flight number

Code:

https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Launch site vs Payload mass success



EDA with SQL

The SQL queries that were performed to explore the data were:

- 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 achieved.
- 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
  Code: https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Object such as markers, and circles were added to a folium map in order to show graphically and geographically the sites and results from the rockets were launched. Lines were used to show and measure the distance between a point of interest and the launch sites.
- This objects are relevant to gain information about the locations, and examine if a relation with the sites environment can conduct to a successful landing.

Code: https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

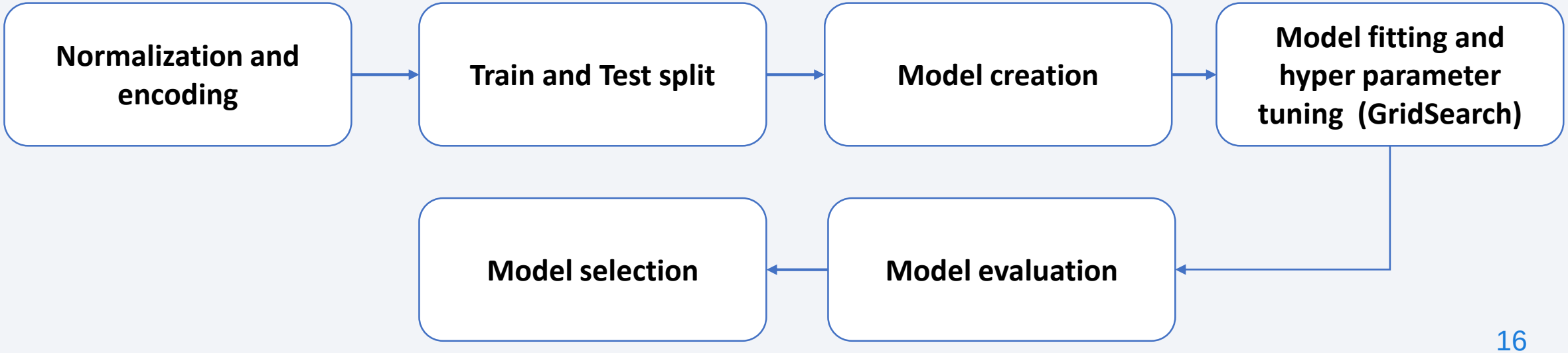
- The plots and graphs added to the dashboard were:
  - Pie chart for the launches by site.
  - Scatterplot with Payload Mass vs Class.
- With interactions like a list where you can select the launch to use in the pie chart and a slide for the Payload range to visualize in the scatterplot.
- These graphs helps to understand how the success is distributed in the different locations, and how the payload range can be an attribute that explains the landing success.

Code: https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/spacex_dash_app2.py

Predictive Analysis (Classification)

- The process of creating a supervised classification model to provide which landing is satisfactory included separating the data into training and test subsets, normalizing and coding the values, as well as training three different models (SVC, Tree classification, logistic regression) and evaluating them using accuracy metric.

Code: https://github.com/FabianPedreros/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



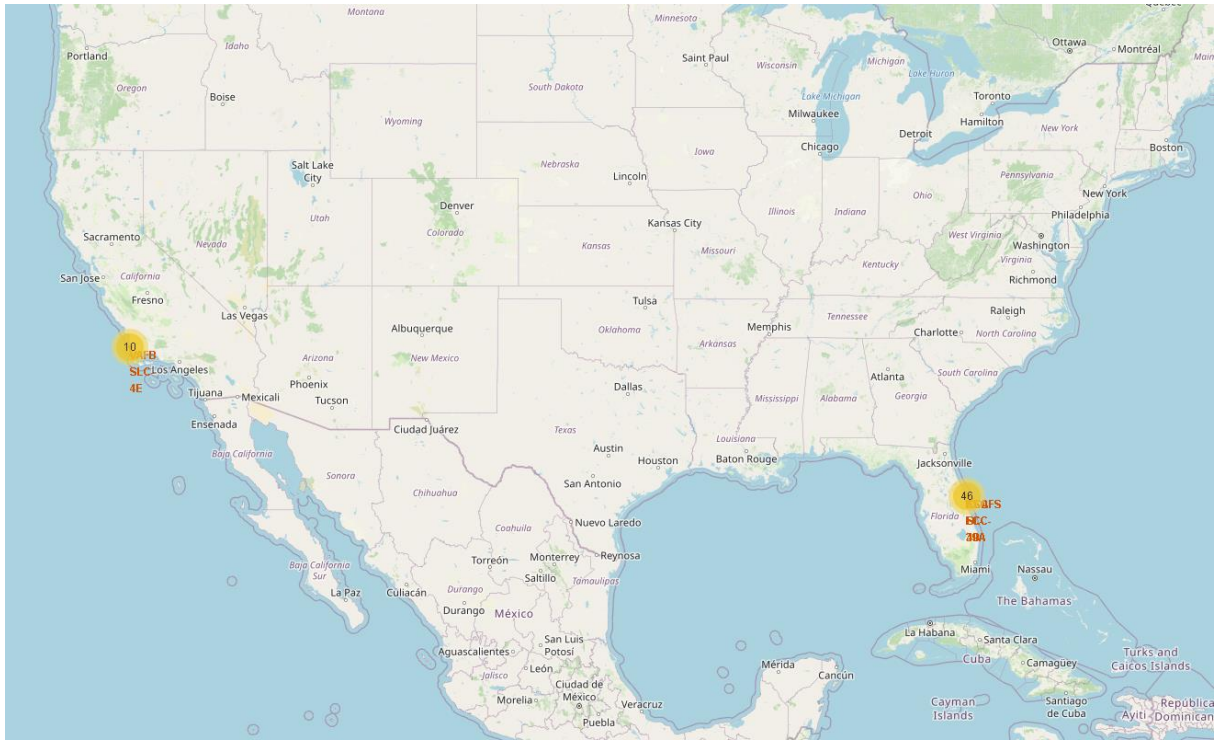
Results

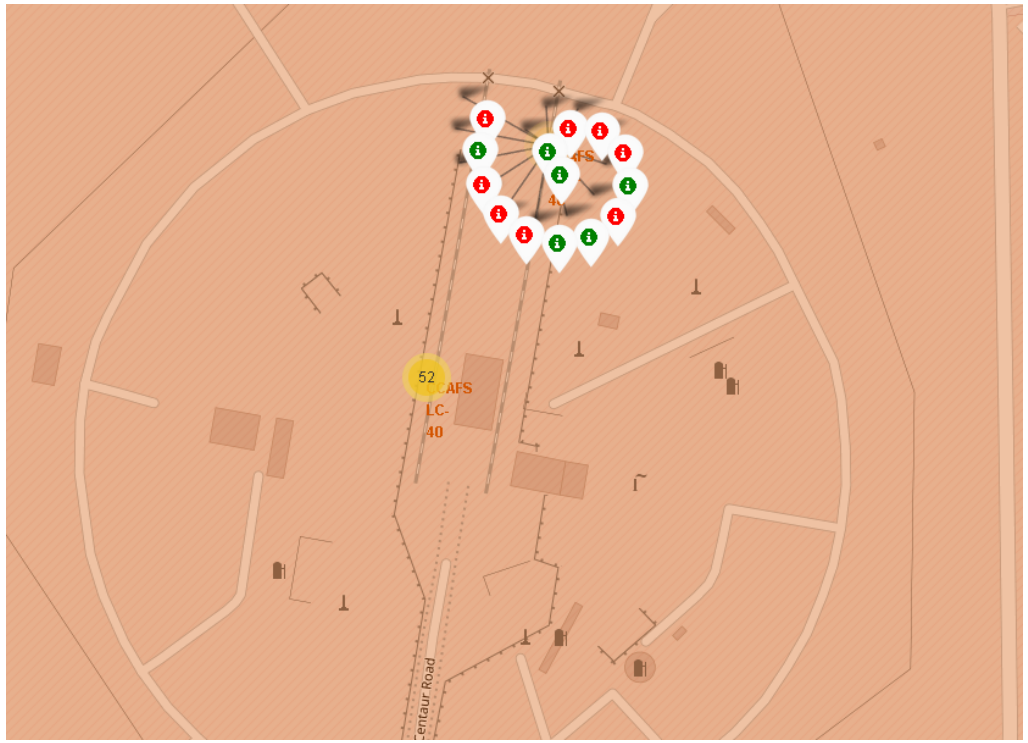
Exploratory data analysis results:

- As SpaceX launches more launches, the more successful they become.
- The site with the highest success rate is the called KSC LC 39A with almost 77% success rate.
- The VAFB SLC 4E site has a high success rate, however it is the site with the fewest launches of the three sites evaluated.
- There are orbits with a high success rate such as SSO (100%) and VLEO (greater than 80%) which have also had more than five launches.
- The orbit with the worst landing success rate is GTO, with 55%.The orbit together with PayloadMass seem to have a correlation with landing success.

Results

Interactive analytics demo in screenshots

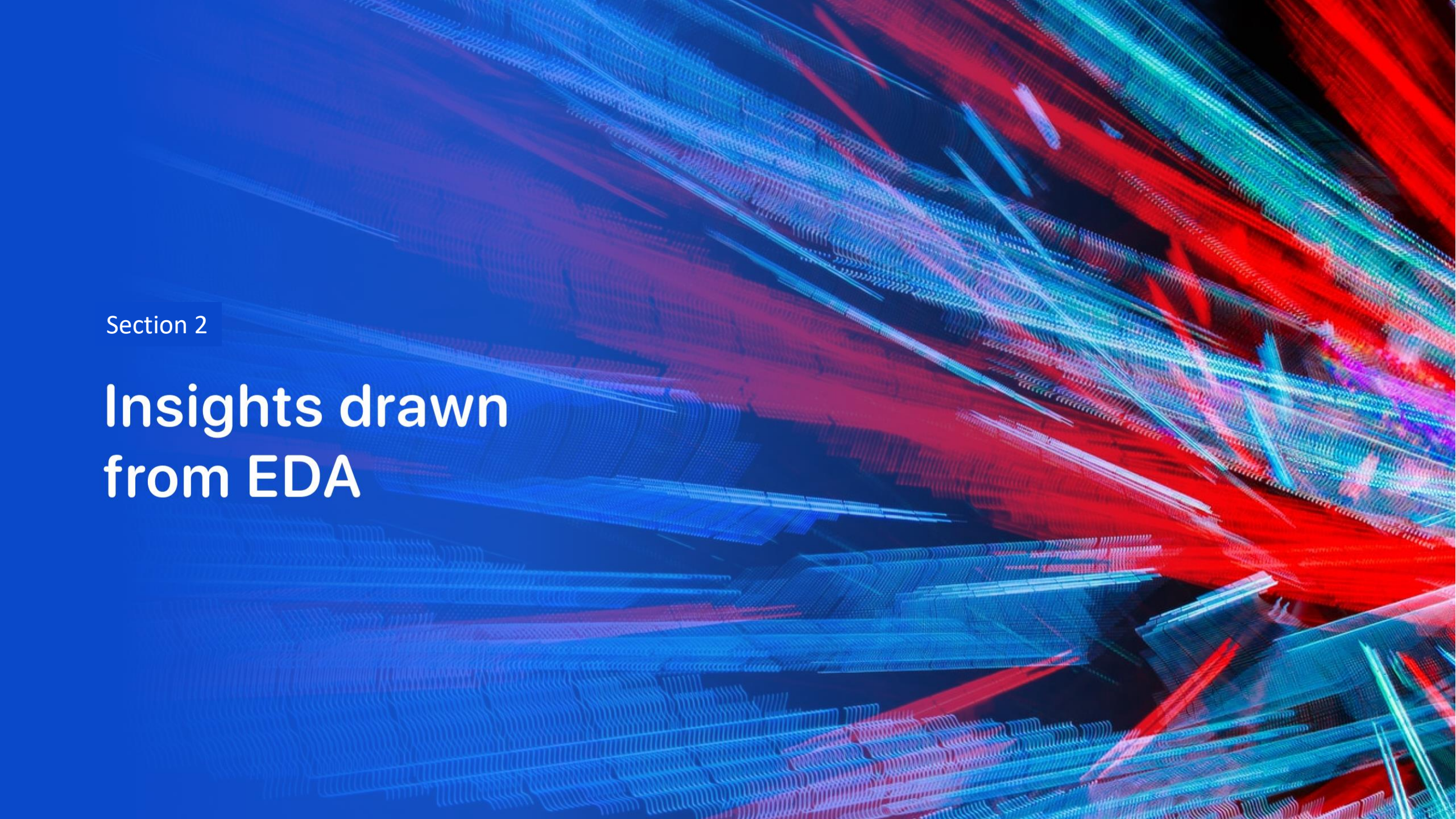




Results

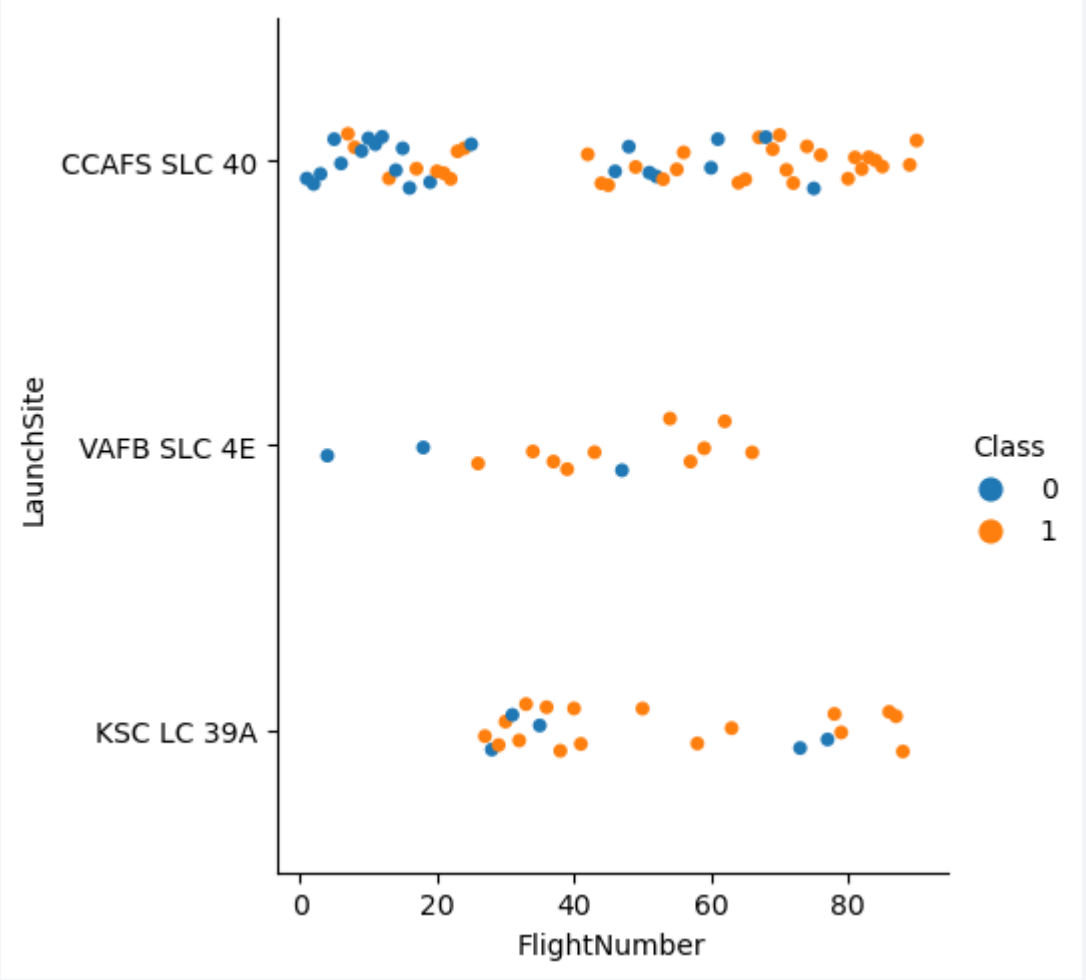
Predictive analysis results

- The best performing supervised classification machine learning algorithm in terms of accuracy was found to be the decision tree, with a training accuracy of 89% and test accuracy of 83%.
- The best performing decision tree model was the hyper-parameter model equal to: 'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'.


Section 2
Insights drawn from EDA

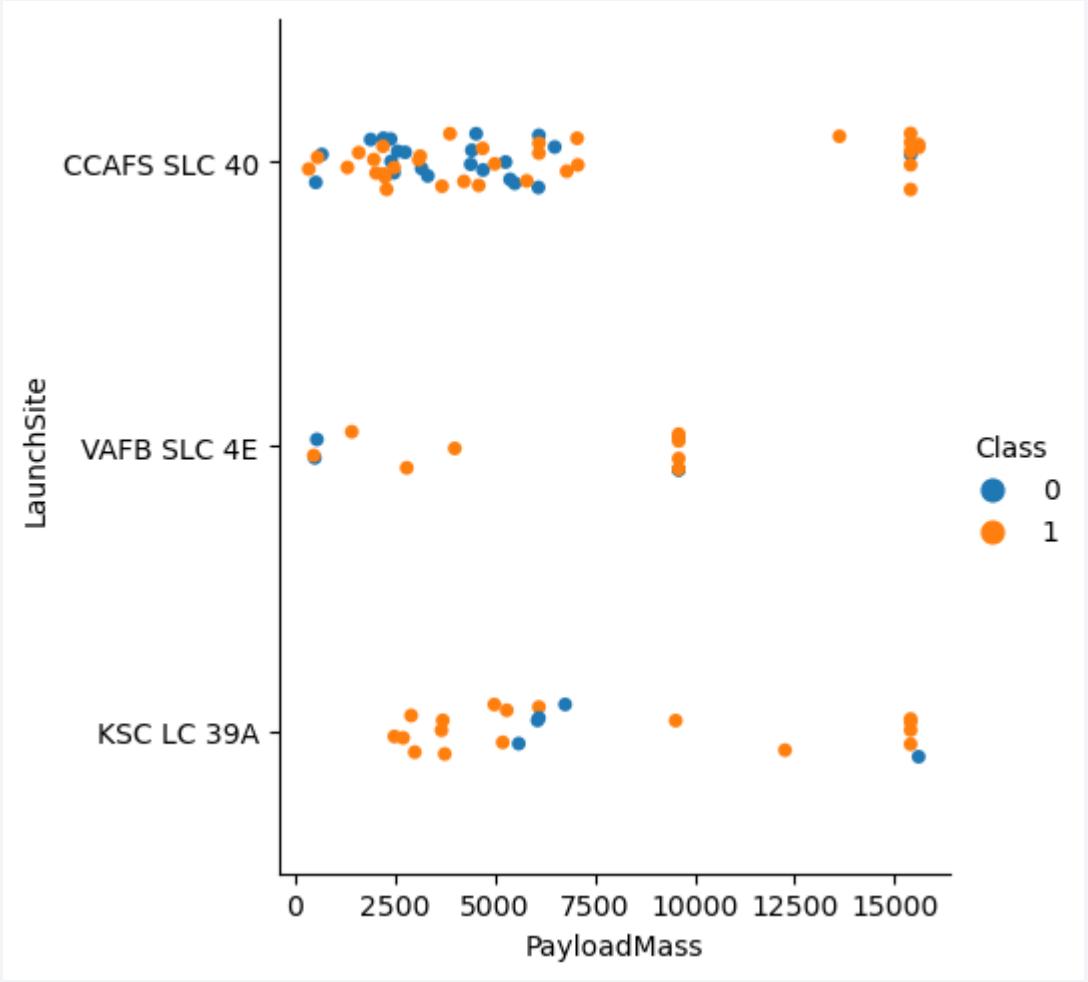
Flight Number vs. Launch Site

- The launch site with the highest accuracy is KSC LC 39A.
- The launch site with the worst performance is the CCAFS SLC 40.
- Over the time the landing success has been improved.



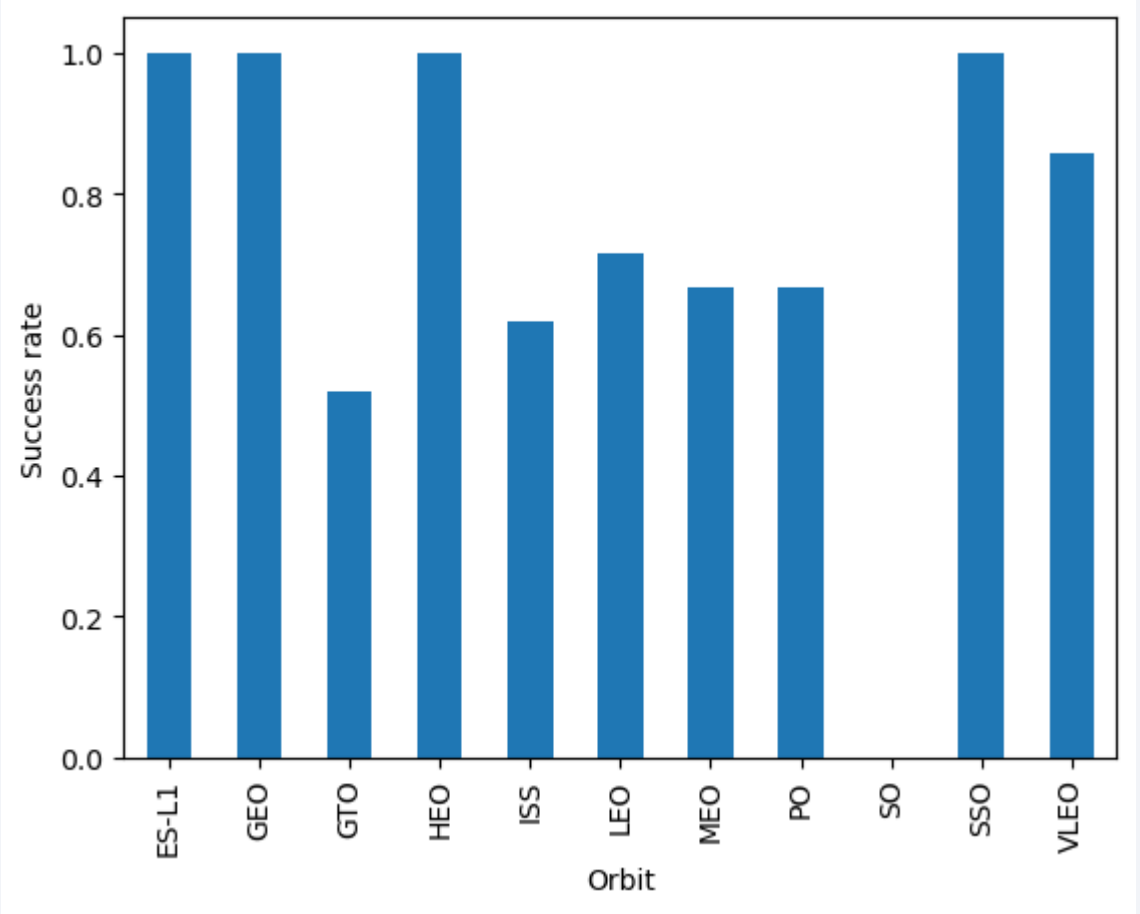
Payload vs. Launch Site

- This scatterplot show us how the first state rockets with a mass below the 7500 kg in the CCAFS site had a very low success. Meanwhile in the other two sites the flights with the same mass were successful.
- The landings with greater mass of 7500 to were more successful.



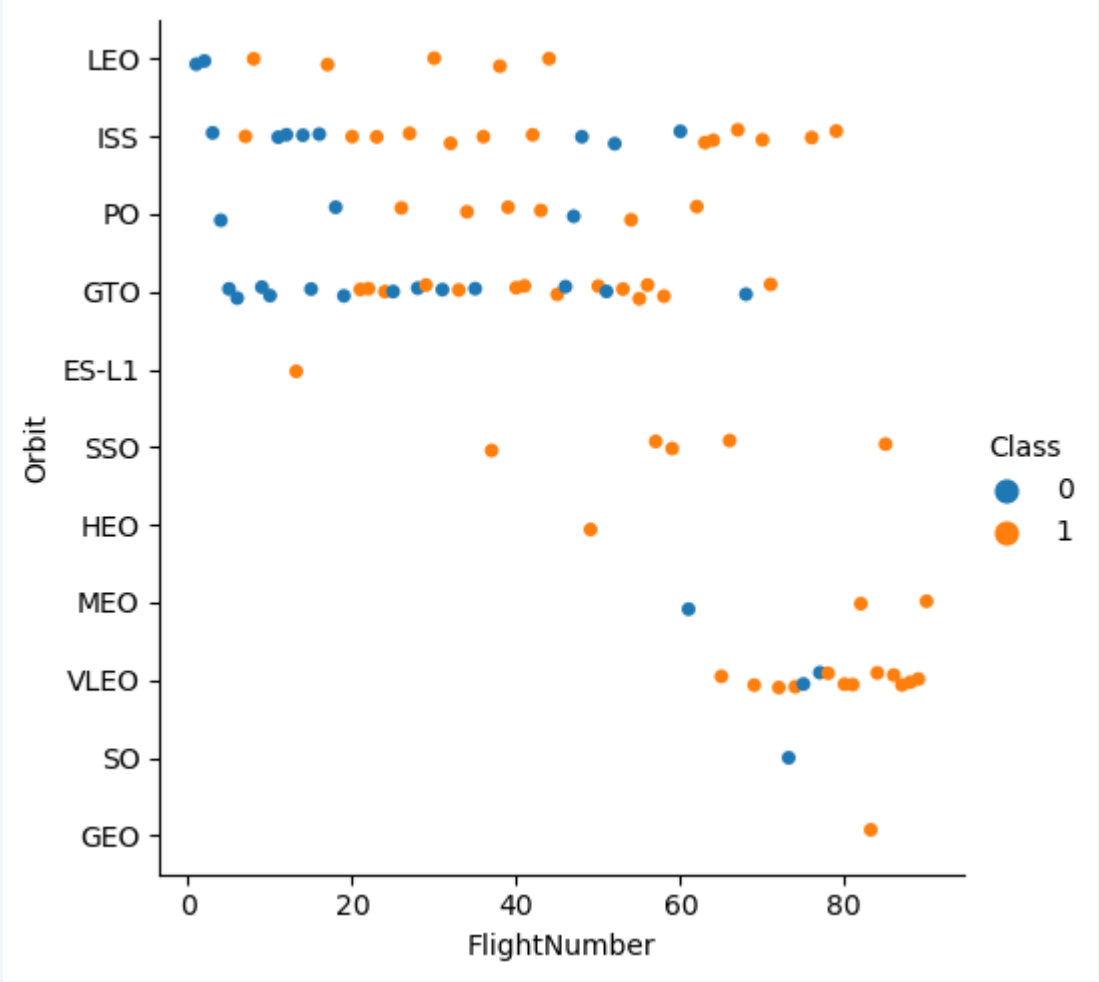
Success Rate vs. Orbit Type

- The orbits with the highest successful rate were ES-L1, GEO, HEO and SSO with 100%
- But should be considered the number of flights to that orbits.
- VLEO seem to be a great accuracy and GTO the worst. Not taking into account the extreme points.



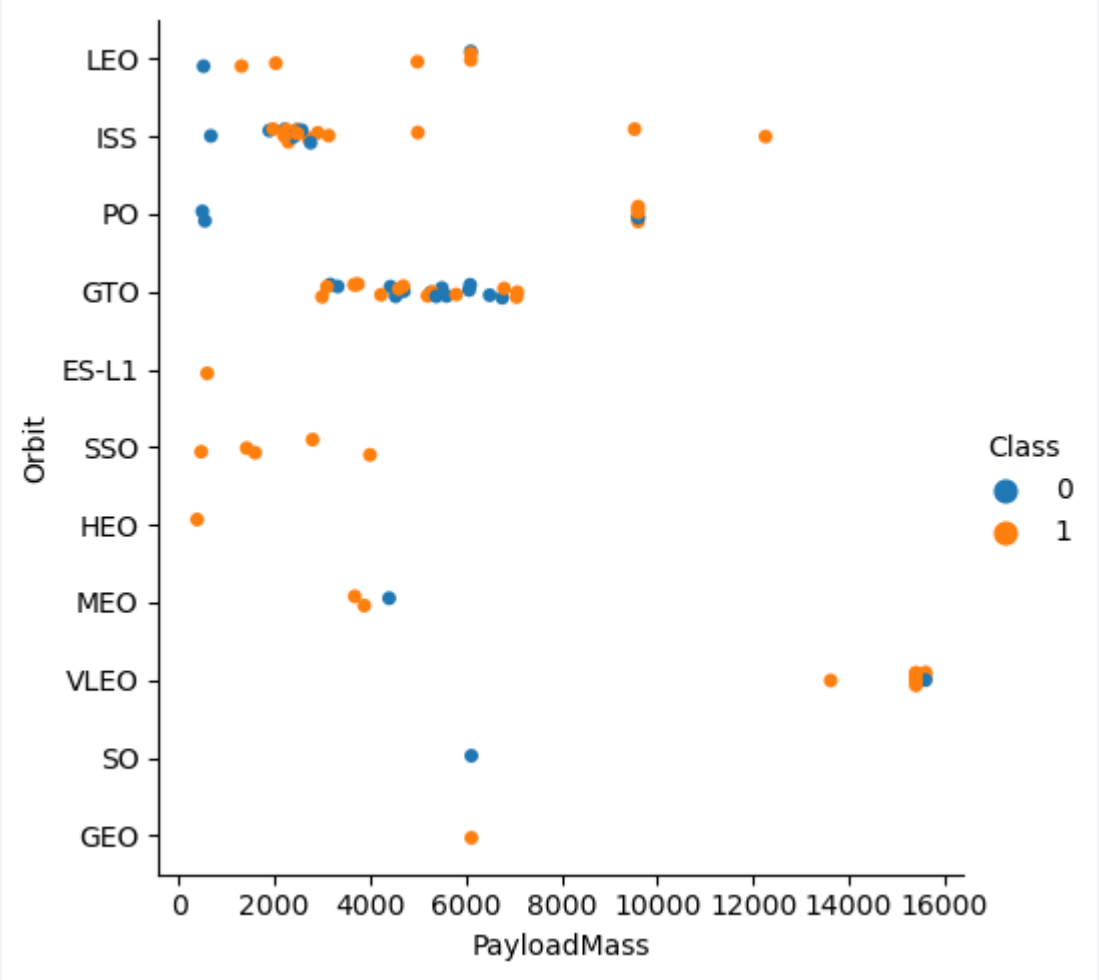
Flight Number vs. Orbit Type

- The GEO, HEO, and ES orbits had a high success rate due to the few flights done to that orbits.
- The most successful orbit is the SSO.
- LEO had improved the success over the flights.



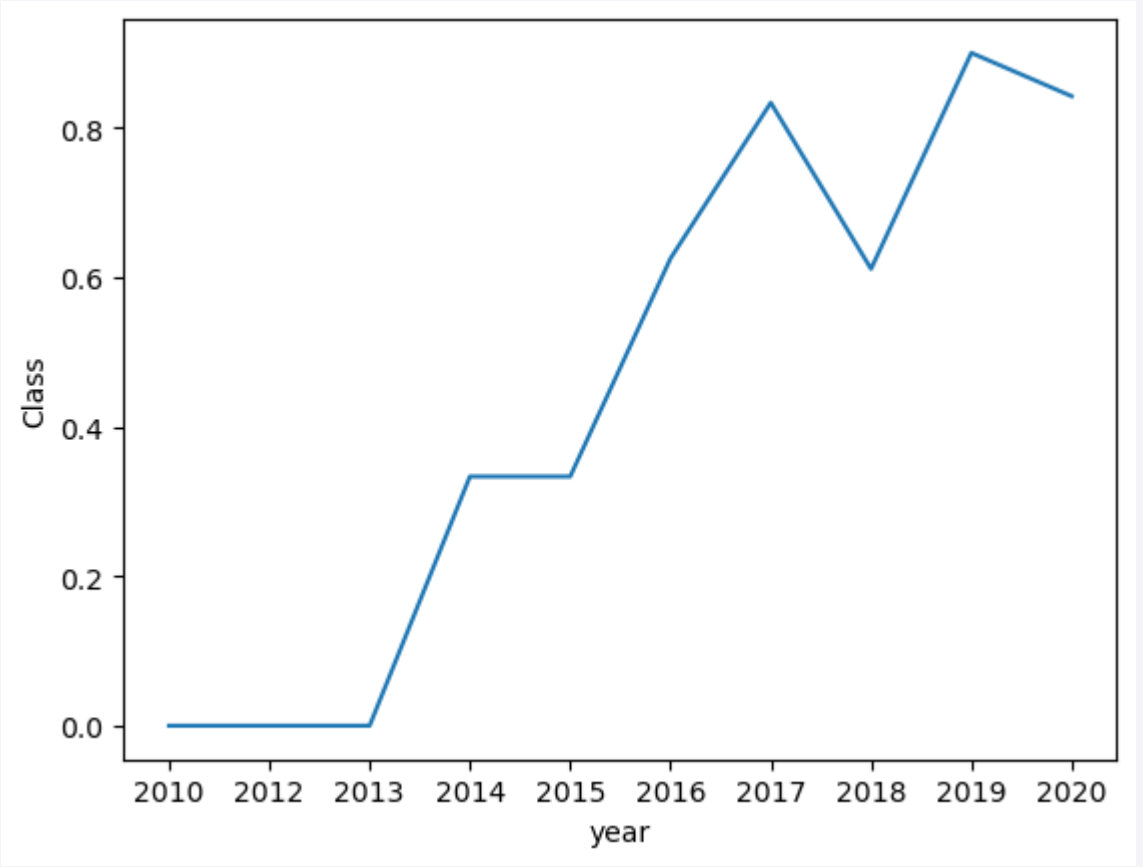
Payload vs. Orbit Type

- The landings with a mass greater that 7500 had a greater success rate.
- The SSO orbit has a great success rate.



Launch Success Yearly Trend

- In the scatterplot we can see how the launch success has increased over the years, with a drop in 2018.



All Launch Site Names

- Query the unique launch sites in the space mission.

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

Launch Site Names Begin with 'CCA'

- Query all the data from the table filtering the Launch sites that begin with ‘CCA’ and limiting the results to the first five rows

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_ _KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Group the Payload mass data using a sum and filtering the customer to be the NASA.

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Group the payload mass by the Booster version, filtering the results by the booster version that contains F9 V1.1

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1 B1003	500.0
F9 v1.1 B1010	2216.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1012	2395.0
F9 v1.1 B1013	570.0
F9 v1.1 B1014	4159.0
F9 v1.1 B1015	1898.0
F9 v1.1 B1016	4707.0
F9 v1.1 B1017	553.0
F9 v1.1 B1018	1952.0

First Successful Ground Landing Date

- Query the dates that were a ‘Success (ground pad)’ order ascending by the date and just taking the first row.

Date
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query the booster version where the landing outcome is equal to ‘Success (drone ship)’ and the payload mass is between 4000 a 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- Listing the possible landing outcomes

Landing _Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attemp

Boosters Carried Maximum Payload

- Find the maximum payload mass and query the booster version that have that 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

- Select the landing month, the landing outcome, booster version and launch site where the landing outcome are equal to ‘Failure (drone ship)’ and the year is 2015.

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

- Showing the count of rows where the Landing outcome were a success and the date was between 4-06-2010 and 20-03-2017, ordered in descending order.

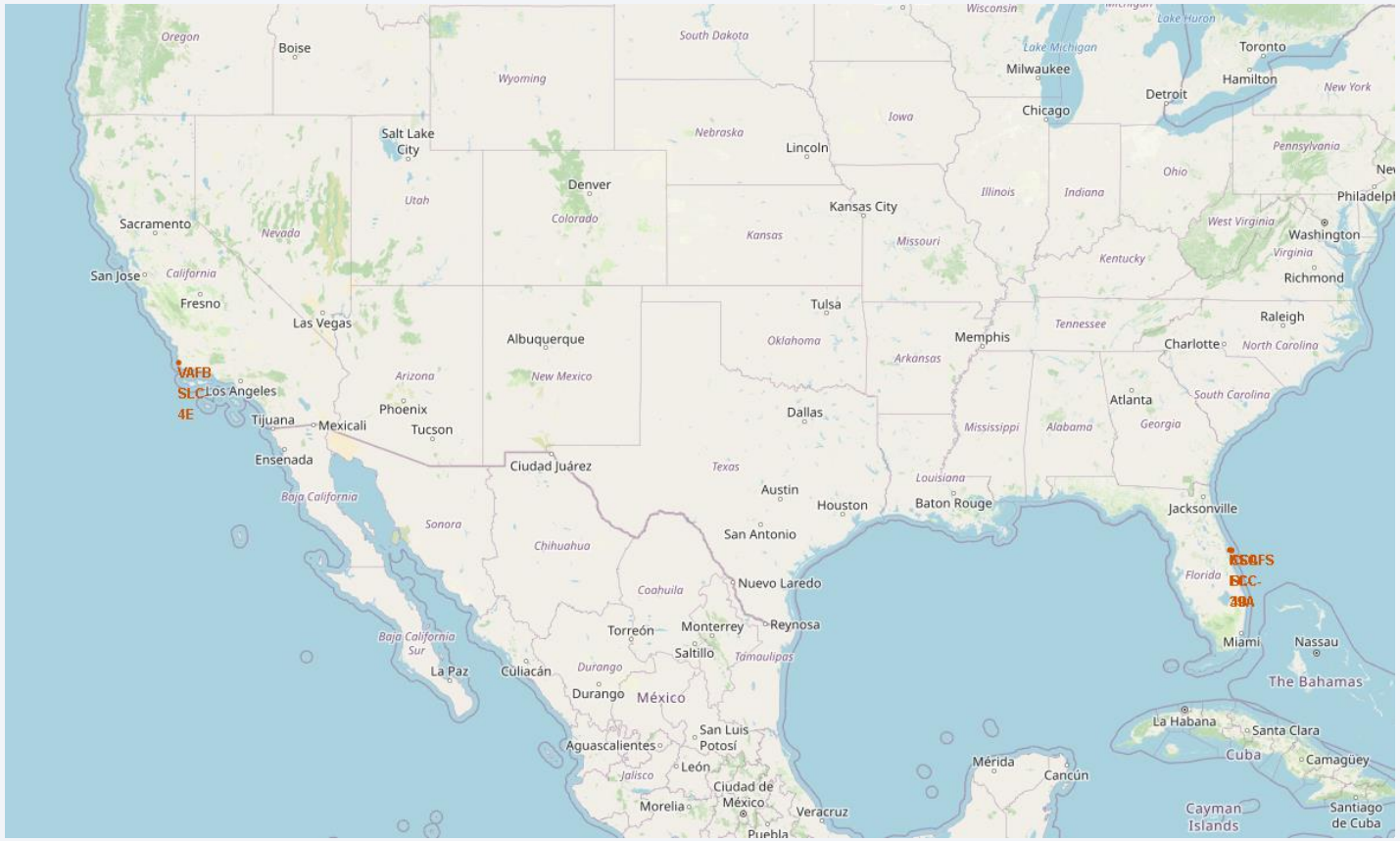
Landing _Outcome	COUNT
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

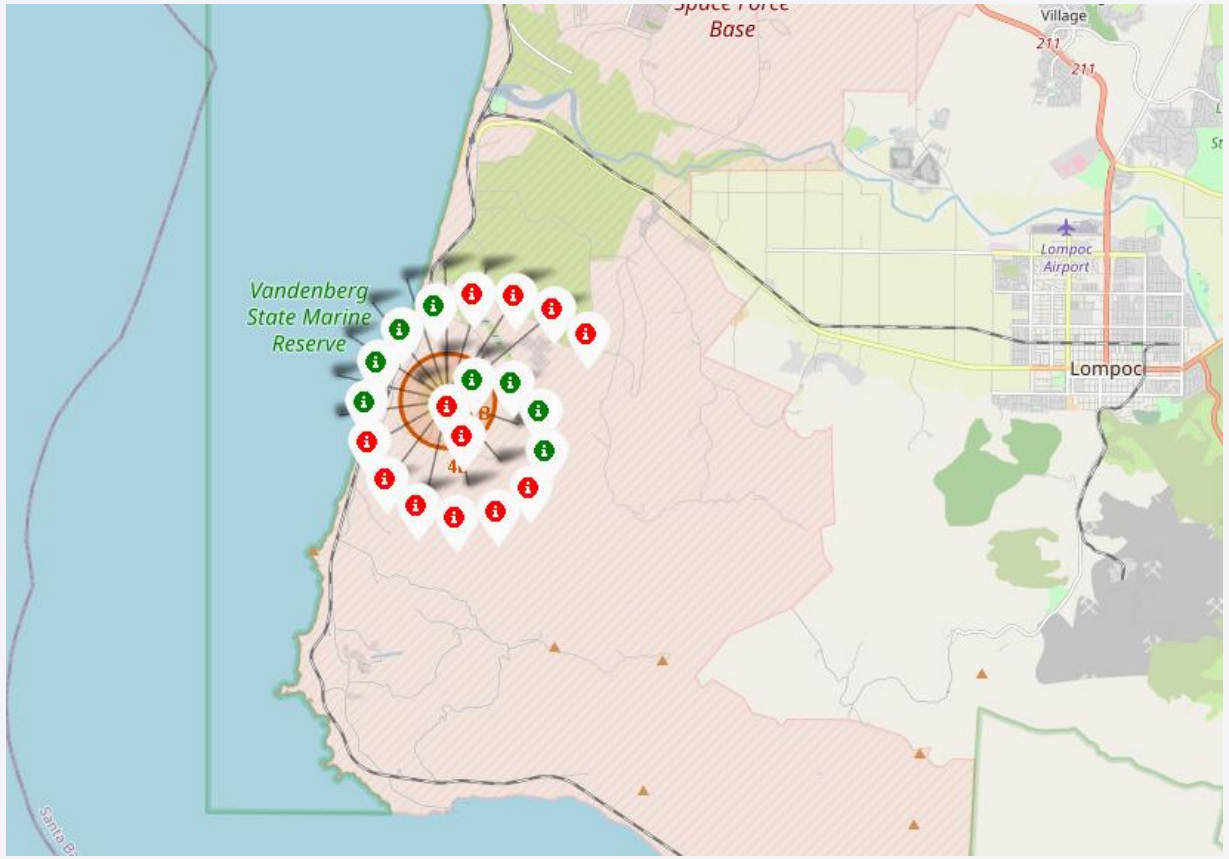
All launch sites

- The map shows where Falcon 9 launches have taken place. There are three different launch places but two global cluster, one in the Florida and the other near Los Angeles. This are near the ocean.



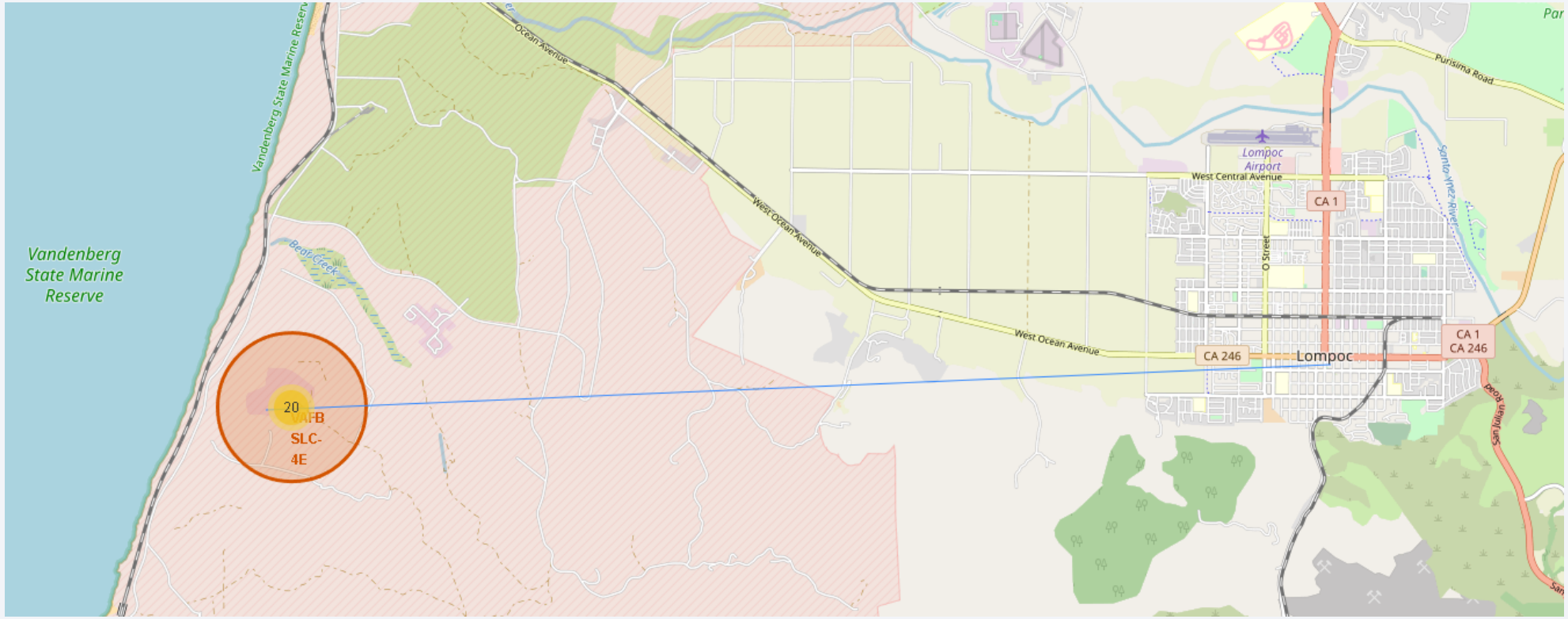
Launch outcomes

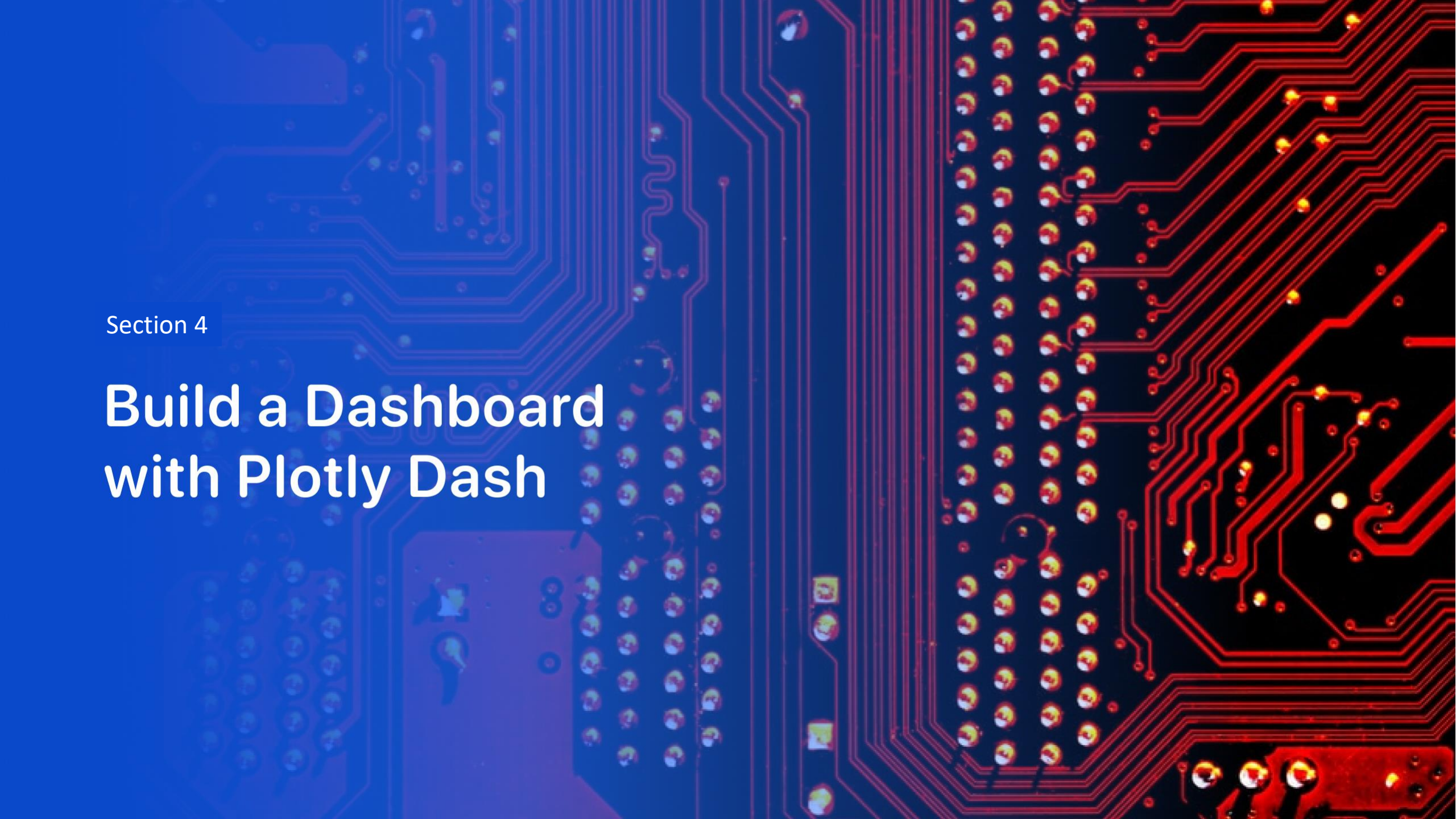
- In the map we can see in each location the distribution of results, in red the Failed ones and in green the Successful ones.



<Folium Map Screenshot 3>

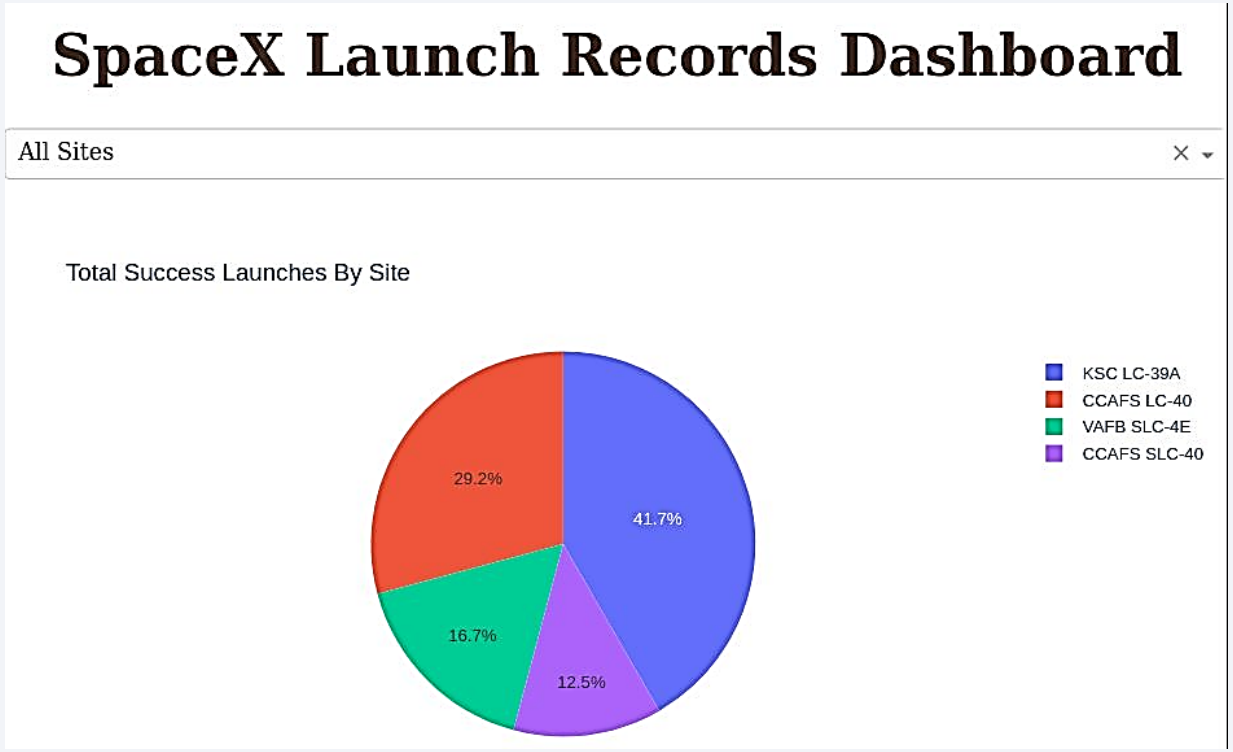
- The launch sites are far away from cities where the failed landings can be dangerous.




Section 4
Build a Dashboard with Plotly Dash

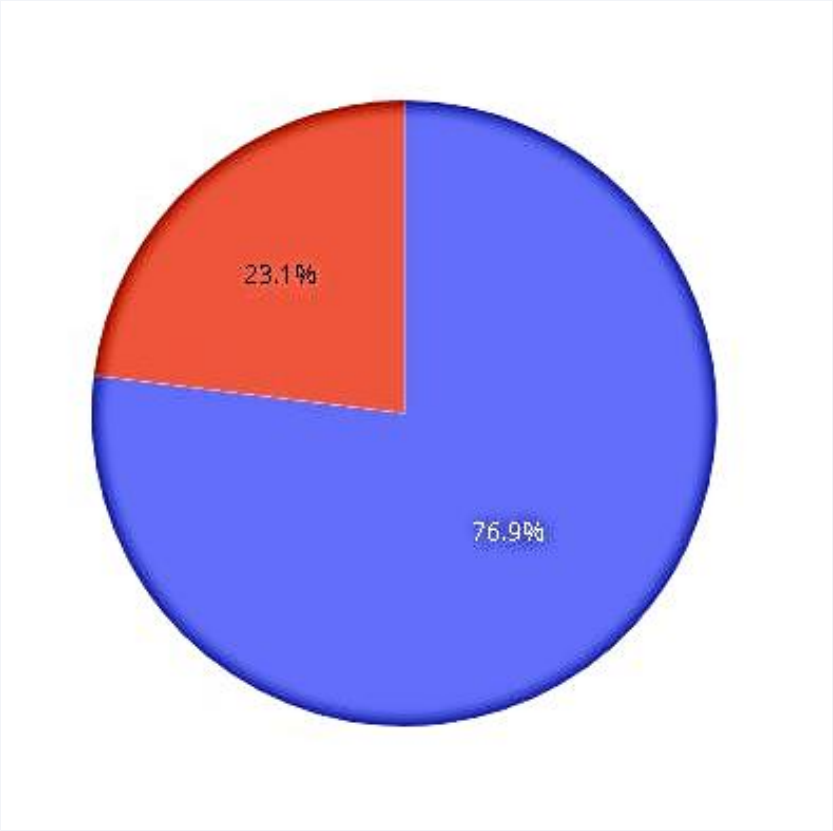
Success rate for all sites

- In the pie chart we can see how the major part of landing successes are from KSC LC-40



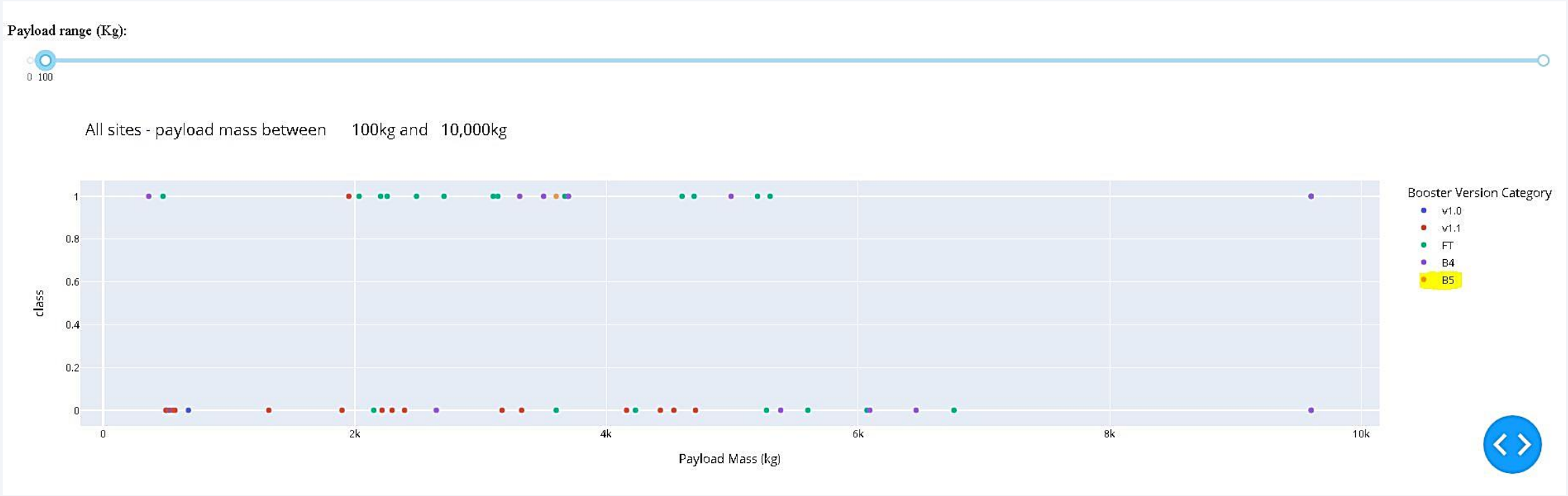
Site highest success ratio

- In the pie char we can see how the site launch with the highest success ratio is the KSC LC-39A with almost 70% of success.



<Dashboard Screenshot 3>

- Here we can see the flights class for every booster version and the relation with the payload mass, the booster version with the highest success rate is the B5.

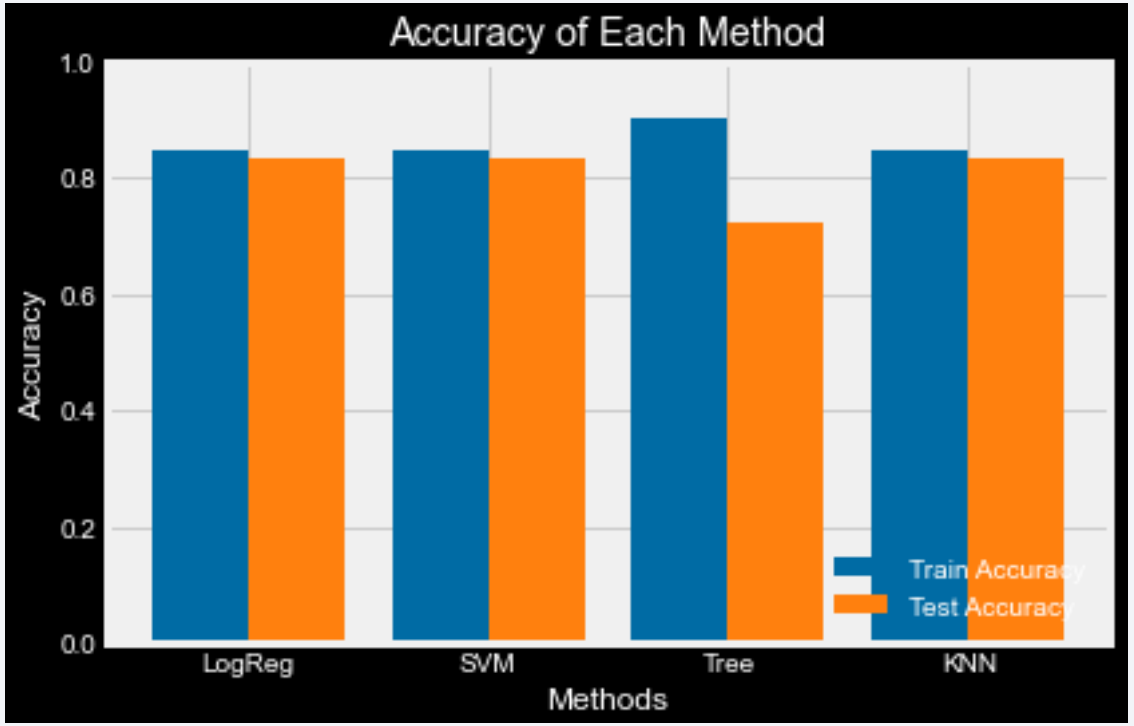


Section 5

Predictive Analysis (Classification)

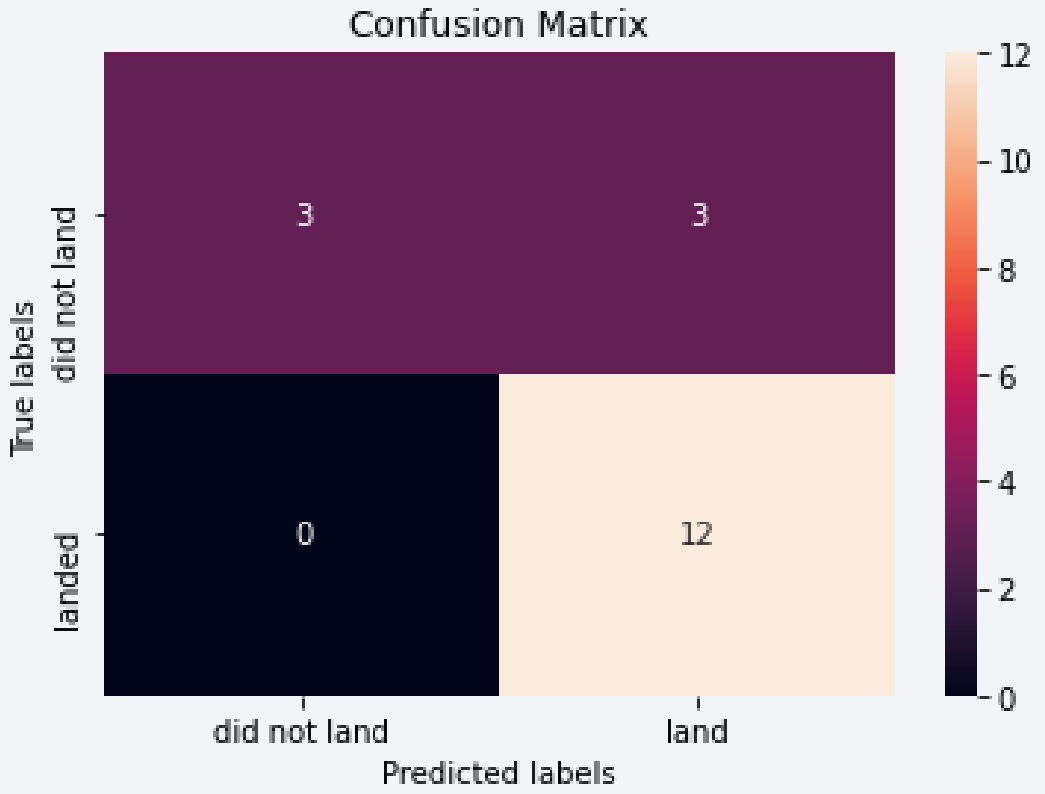
Classification Accuracy

- In the bar chart we can see that the train accuracy is the highest for the tree model but the best test accuracy is for equal for lineal the models.



Confusion Matrix

- The lineal models had the same precision, with some error for the predicted landing that in the reality did not land. In this cases we had false positives.



Conclusions

- The site launch seems to have a correlation with a successful landing, in this case the SC LC-39A site has the highest success rate.
- The payload mass seems to have a positive correlation with a successful landing, over the 7500 kg the success rate is high.
- It is possible to predict the success or not of a landing using a supervised machine leaning model, SVC, Logistic Regression and KNN had the same precision in the test data.

Appendix

- All the project with the different assets are in the GitHub repository:

  https://github.com/FabianPedreros/IBM-Data-Science-Capstone

Thank you!