IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pouchias Konstantinos
20/07/2022

Outline

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

Executive Summary

The data analysis techniques employed during this project were as follows:

- Data collection using the SpaceX API and webscraping
- Data wrangling.
- Exploratory Data Analysis (EDA) including data visualization, SQL queries and interactive visual analytics, as well as building a dashboard.
- Machine learning model development and evaluation.

The results of the analysis were the following:

- Getting useful insights about the data and the correlation between different features and the target variable.
- Finding out which Machine Learning model is better for predicting the launch outcome.

Introduction

Taking the role of a data scientist working for a rocket company named SPACE Y, that wants to compete with SPACE X, we need to answer the following questions, :

- Will the first stage land successfully so that can be reused?

- What is the price of each launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  1. Request of an API
  2. Web scraping

- Perform data wrangling

  - Performing Exploratory Data Analysis (EDA) to find patterns in the data and adding a new column named "Class" to the dataframe, based on the outcome of a launch in binary values, 0 for failure and 1 for success

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
  - First the collected data were normalized before they got split in training and testing data sets. Then we used 4 different classification methods and after we determined the best combination of parameters for each method, we made a prediction of the landing outcome of the first stage.

Data Collection

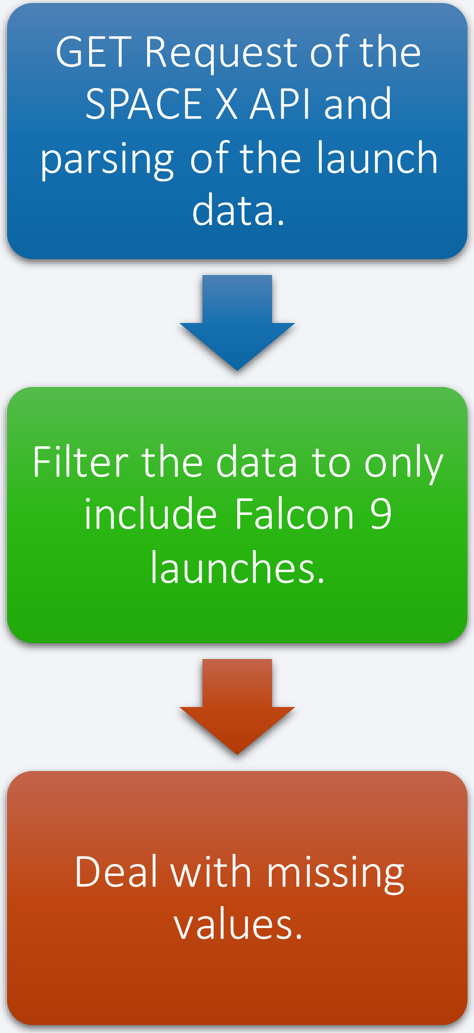
- The data sets of the problem was collected using two different ways.

1. Utilizing the **SPACE X REST API** (https://api.spacexdata.com/v4/launches/past)

2. From the wikipedia page using web scraping techniques. (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

- During the data collection process we utilized the free SpaceX REST API.
- The procedure we followed is depicted in the adjacent flowchart.
- GitHub URL: https://github.com/kpouchias/Applied-Data-Sciene-Capstone/blob/master/Week%201%20Data%20Collection%20API.ipynb



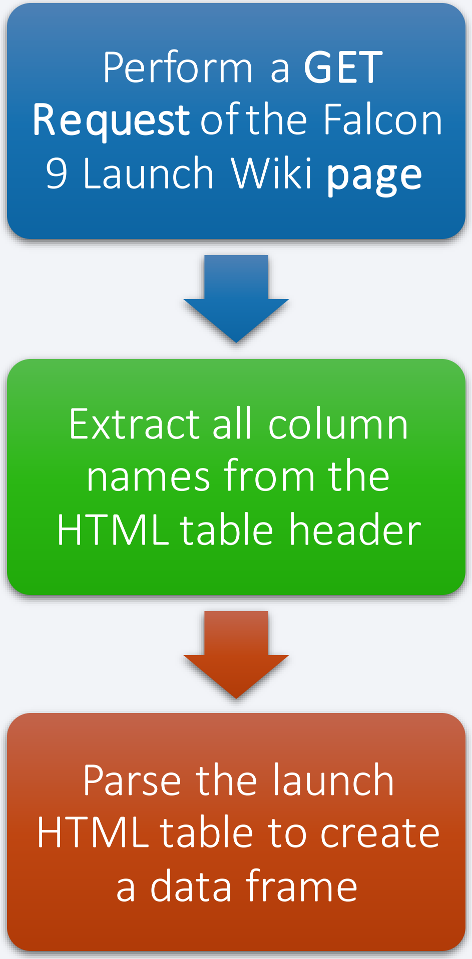
GET Request of the SPACE X API and parsing of the launch data.

Filter the data to only include Falcon 9 launches.

Deal with missing values.

Data Collection - Scraping

- Some web scraping techniques were used in order to collect the necessary data from the corresponding wikipedia page.
- The adjacent flowchart contains the 3 step process we followed.
- GitHub URL:https://github.com/kpouchias/Applied-Data-Sciene-Capstone/blob/d6b3ea8f50ccadf834f437e18259983f363cf4ab/Week%201:%20Data%20Collection%20with%20Web%20scraping.ipynb



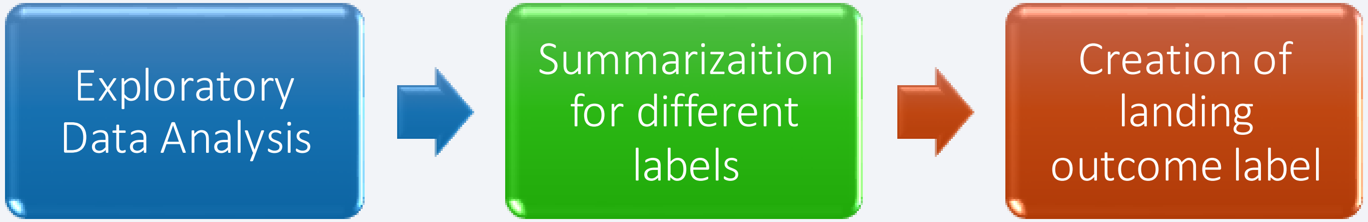
Perform a **GET Request** of the Falcon 9 Launch Wiki **page**

Extract all column names from the HTML table header

Parse the launch HTML table to create a data frame

Data Wrangling

- The dataset was first subjected to some exploratory data analysis (EDA), followed by summaries of launches per site, occurrences of each orbit, and occurrence of mission outcome per orbit type.
- Finally, the Outcome column was used to generate the landing outcome label.


Exploratory Data Analysis

Summarizaition for different labels

Creation of landing outcome label

- GitHub URL: https://github.com/kpouchias/Coursera_Capstone_Project/blob/master/Week%201:%20Data%20Wrangling.ipynb

EDA with Data Visualization

- On this section of our analysis the main objective was to find out, if and how different pairs of variables can affect the landing outcome. To do that we used different charts, in particular scatter, bar and line charts. The variable pairs we used were:

1) Payload Mass vs Flight Number (scatter plot)
2) Flight number vs Launch Site (scatter plot)
3) Launch Site vs Payload Mass (scatter plot)
4) Flight Number vs Orbit Type (scatter plot)
5) Payload Mass vs Orbit Type (scatter plot)
6) Success Rate vs Orbit Type (bar chart)
7) Average Success Rate vs Year (line chart)

- GitHub URL: https://github.com/kpouchias/Coursera_Capstone_Project/blob/112d41d9d6511d465290c3297c090da1d86646ad/Week%202:%20EDA%20with%20Visualization.ipynb

EDA with SQL

The SQL queries that were executed are the following:

- *Names of unique launch sites in the space mission*
- *The first 5 records where launch site starts with 'CCA'*
- *The total payload mass carried by boosters launched by NASA (CRS)*
- *The average payload mass carried by booster version F9 v1.1*
- *The date when the first successful landing outcome in ground pad was achieved*
- *The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
- *The total number of successful and failure mission outcomes*
- *The names of the booster versions which have carried the maximum payload mass*
- *The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015*
- *Rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order*
- GitHub URL: https://github.com/kpouchias/Coursera_Capstone_Project/blob/112d41d9d6511d465290c3297c090da1d86646ad/Week%202:%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium Maps were utilized with markers, circles, lines, and marker clusters.
- Markers represent points, such as launch sites
- Circles represent highlighted areas around specific coordinates
- Marker clusters represent groups of events in each coordinate, such as launches at a launch site
- Lines represent distances between two coordinates.

- GitHub URL: https://github.com/kpouchias/Coursera_Capstone_Project/blob/112d41d9d6511d465290c3297c090da1d86646ad/Week%203:%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

To visualize data, the following graphs and plots were used:

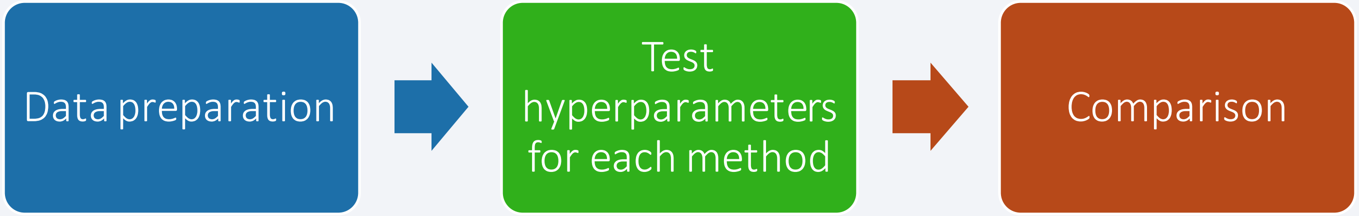
- Percentage of launches per site
- Payload range

This combination allowed for a quick analysis of the relationship between payloads and launch sites, assisting in determining where the ideal spot to launch is based on payloads.

- GitHub URL: https://github.com/kpouchias/Coursera_Capstone_Project/blob/112d41d9d6511d465290c3297c090da1d86646ad/Interactive_Dashboard.py

Predictive Analysis (Classification)

- After the preparation and the standardization, the data were split in training and testing set. Then we used the following classification methods: **Logistic Regression**, **Support Vector Machine (SVM)**, **Classification Tree**, **K-Nearest Neighbors.**
- First we found the best hyperparameters using the training set for every method and then we compared the accuracy of each method using the test set



- GitHub URL:
  https://github.com/kpouchias/Coursera_Capstone_Project/blob/112d41d9d6511d465290c3297c090da1d86646ad/Week%204:%20Machine%20Learning%20Prediction.ipynb

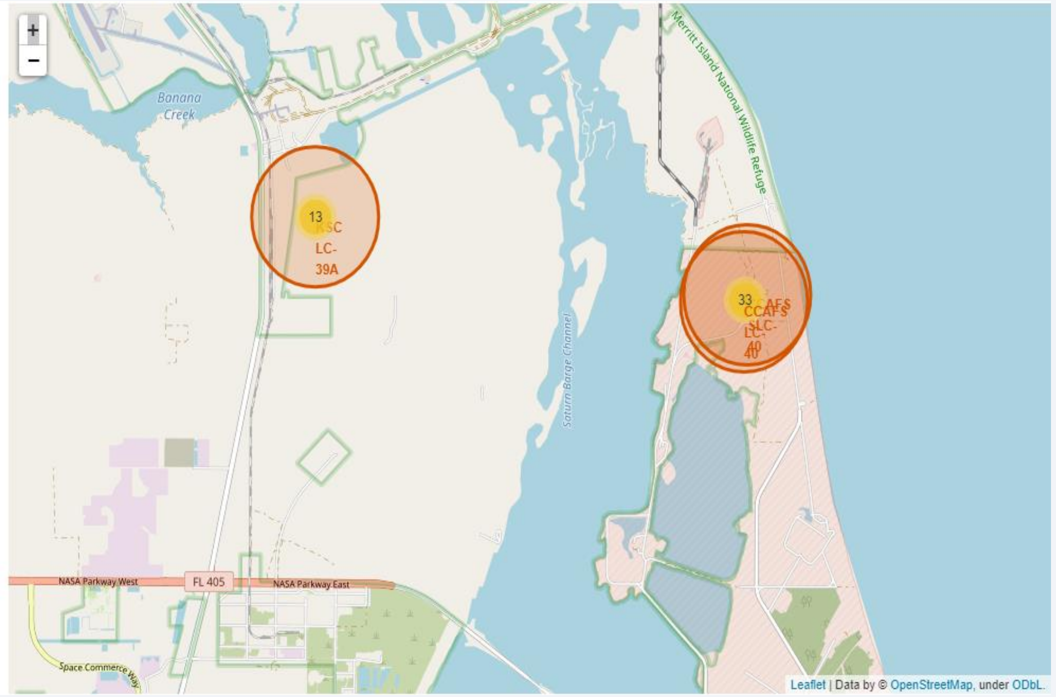
Results

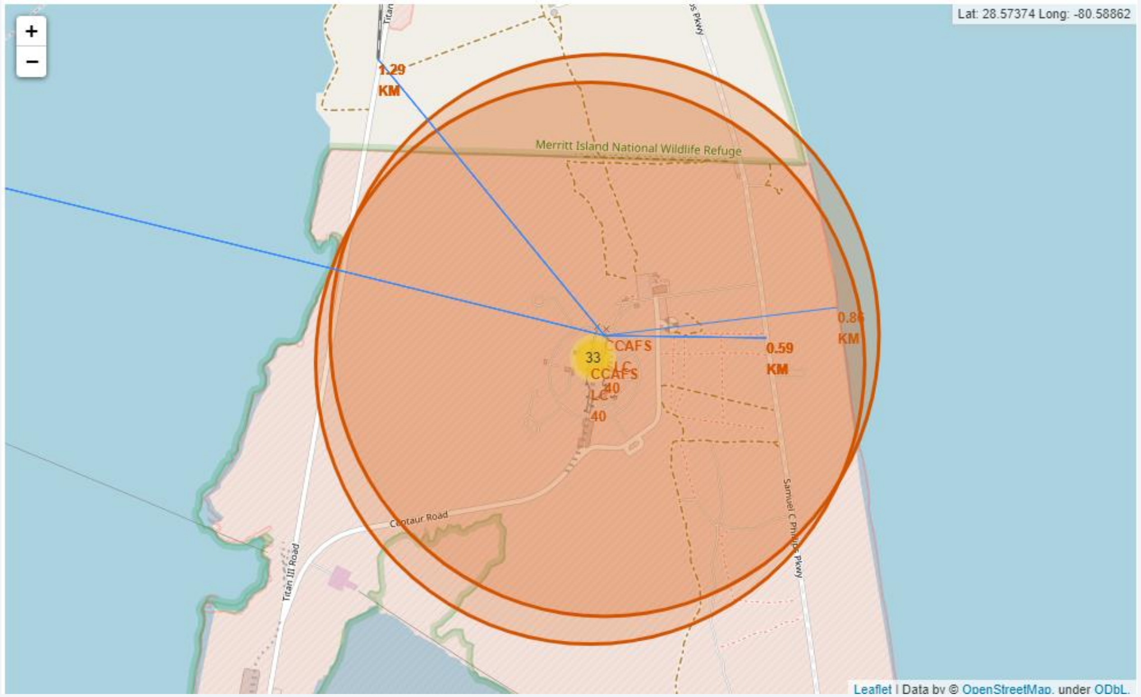
❑Exploratory data analysis results:

- 4 separate launch sites were used by SpaceX.
- CCAFS SLC-40 is the first launch site to be utilized as well as the one where the most launches occurred.
- The first successful landing outcome in ground pad was achieved in 12/22/2015, almost 5 years after the first launch took place.
- The success rate of landing outcomes increased over time.
- Nearly all mission outcomes were successful.
- The average Payload Mass carried by booster version F9 v1.1 was 2,928 kg.
- For Payload Mass over 8,000kg the success rate of landing outcome is almost 100%.
- The Orbit types with the highest landing outcome success rate were ES-L1, GEO, HEO, SSO.

Results

- It was feasible to determine, using interactive visual analytics, that launch sites used to be in secure locations, such as close to the sea, away from inhabited areas and had a robust logistic infrastructure surrounding them.

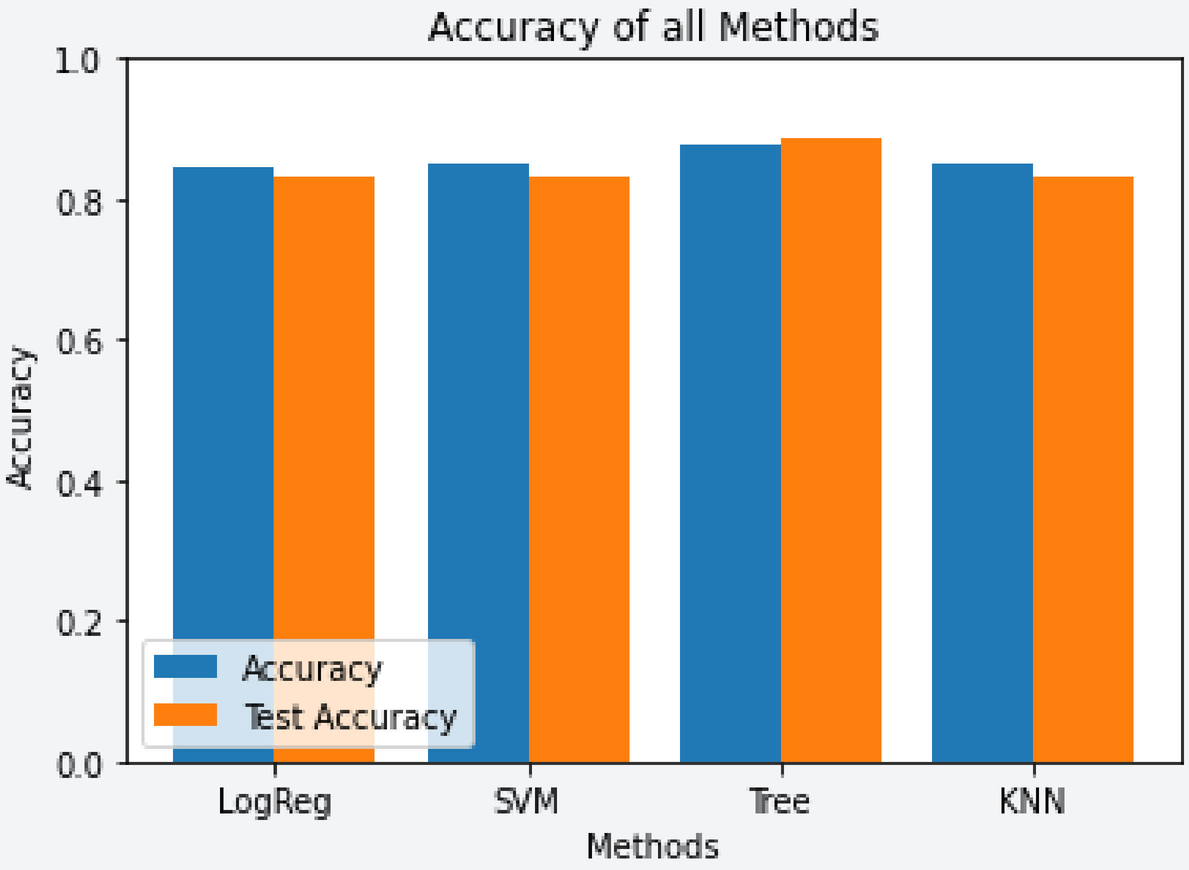




+
−
Lat: 28.57374 Long: -80.58862
1.29 KM
Merritt Island National Wildlife Refuge
0.86 KM
0.59 KM
33
CCAFS LC-40
CCAFS LC-40
Centaur Road
Titan III Road
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

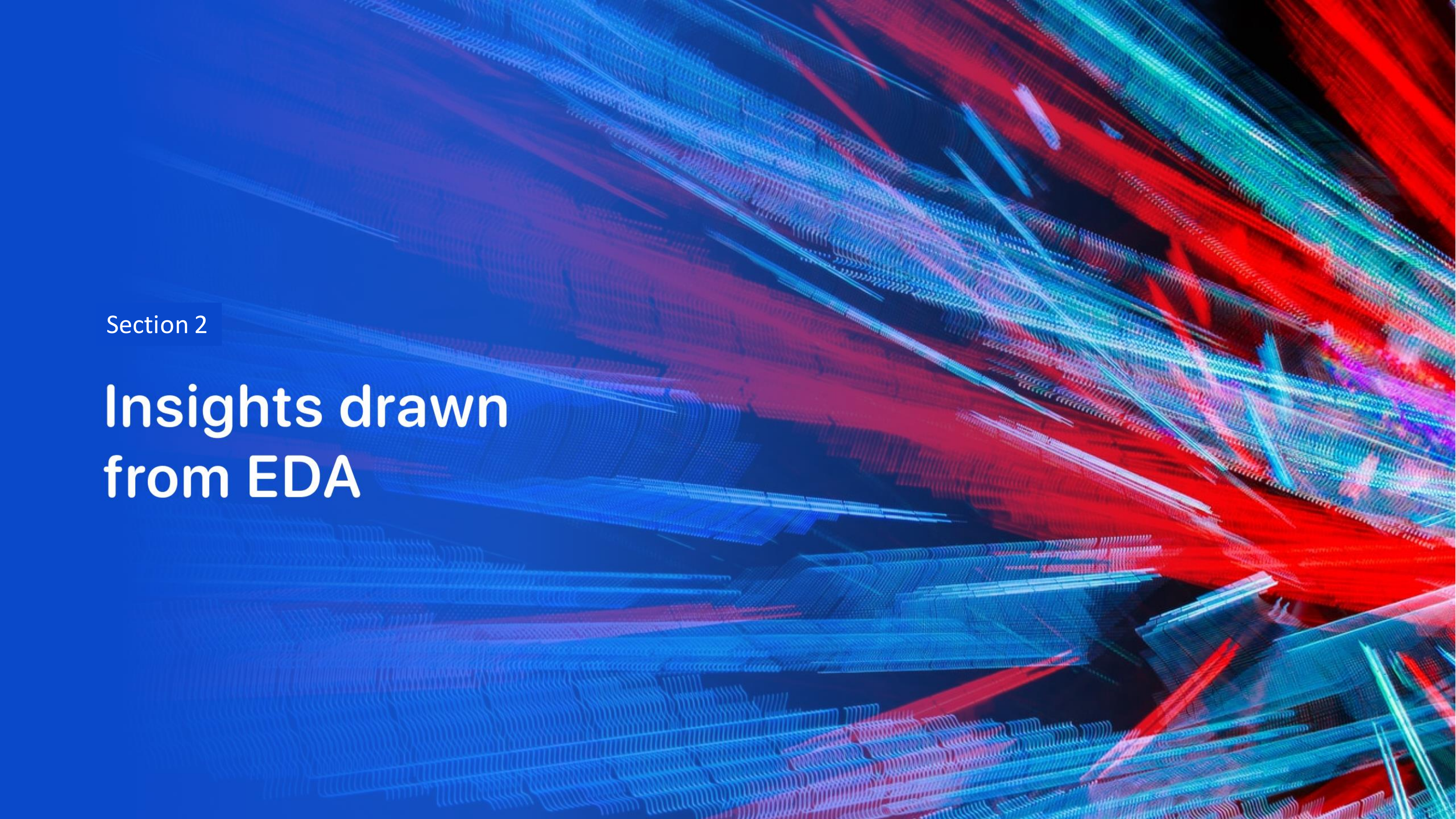
Results

- According to the predictive analysis, Decision Tree classifier is the best model to predict landing outcomes, achieving the highest accuracy scores both on the training as well as the testing data.

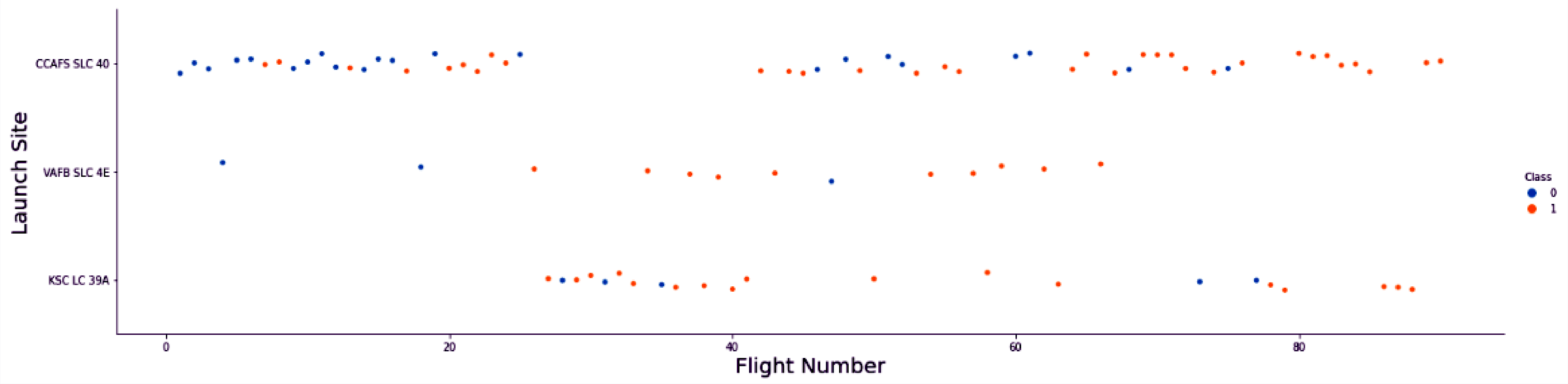


Section 2

Insights drawn from EDA



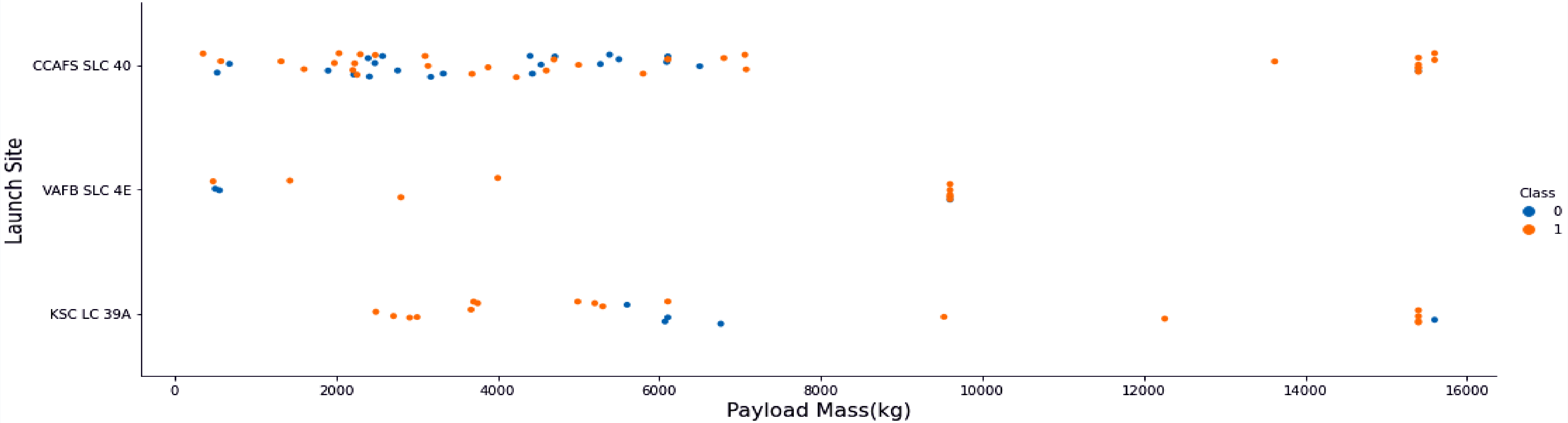
Flight Number vs. Launch Site



- CCAFS SLC- 40 is the Launch Site with most launches

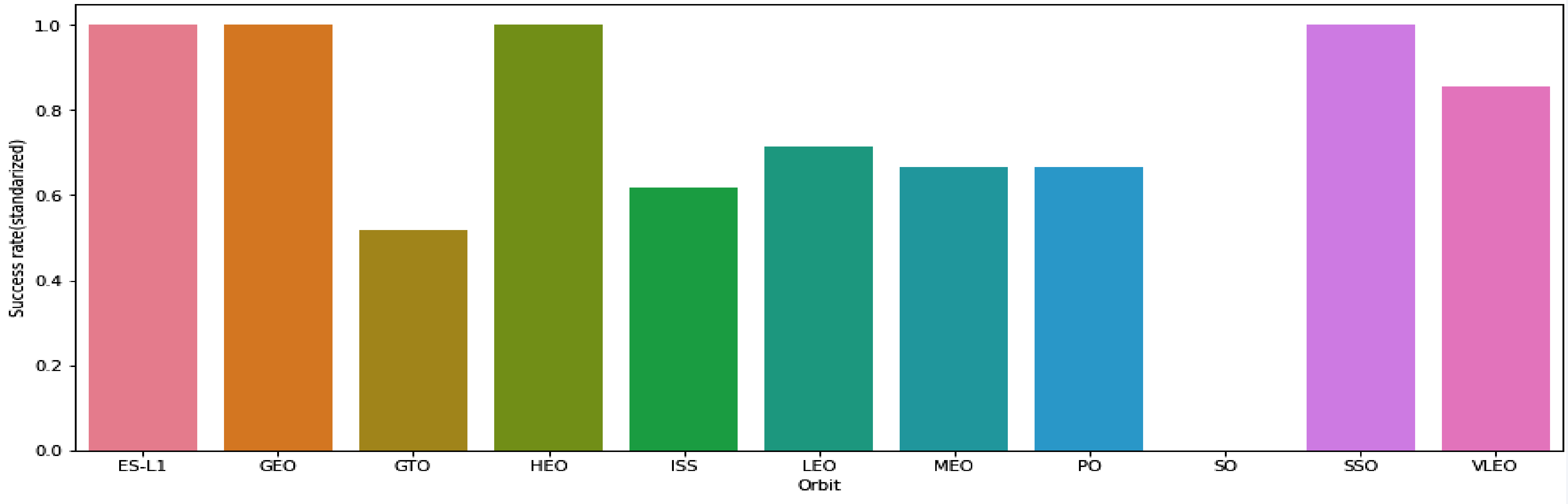
- The success rate overall increased over time

Payload vs. Launch Site



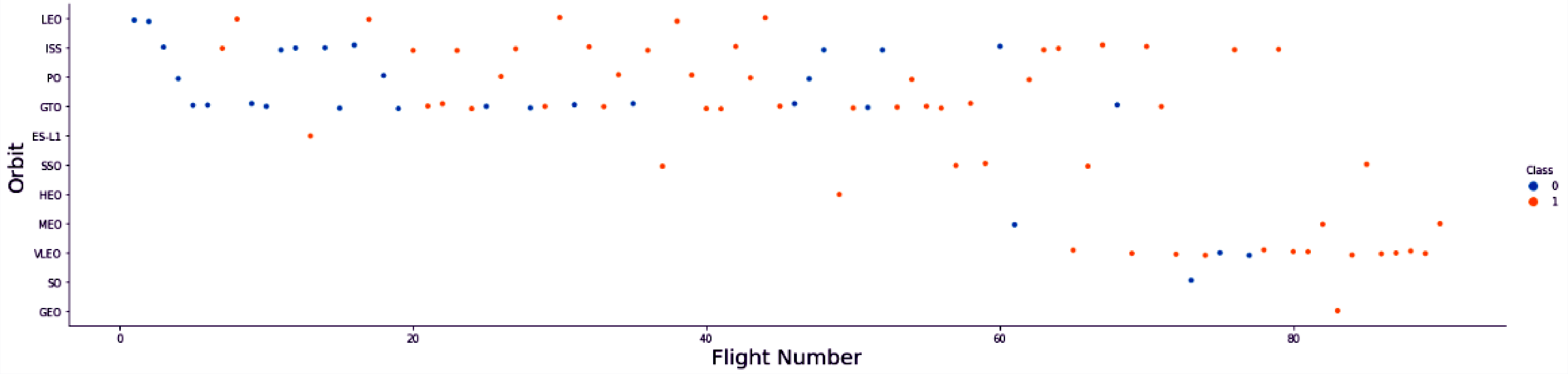
- In the majority of launches the payload mass was under 8000 kg.
- The success rate for payload mass greater than 8000 kg is nearly 100%.
- The launch site VAFB SLC-4E was used only for payload mass less than 10000 kg.

Success Rate vs. Orbit Type



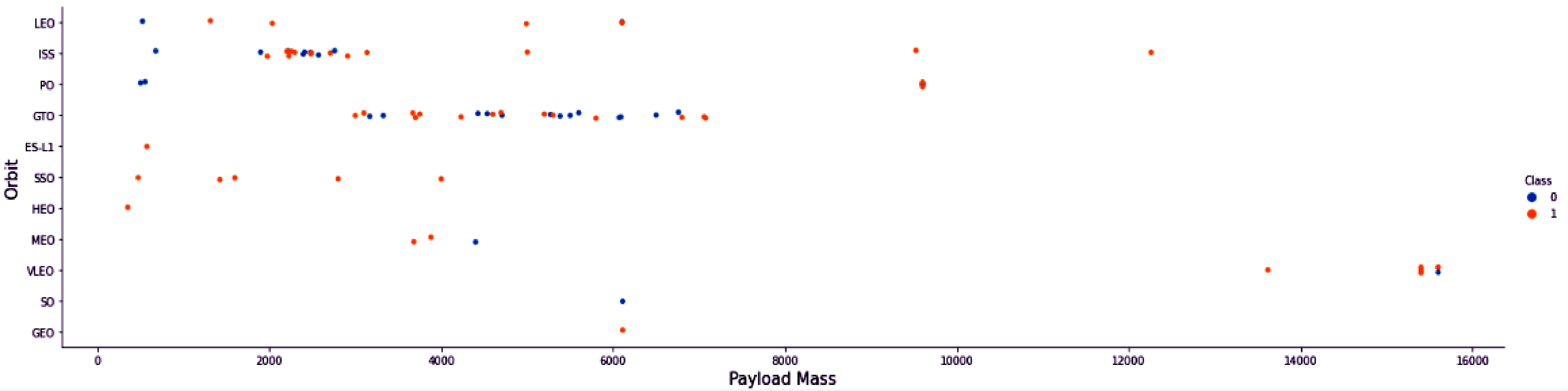
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate, but those Orbits were used only once.
- VLEO is the 4th better orbit according to the success rate but since it is also the 3rd most used orbit type we can presume that it is probably the most successful one.

Flight Number vs. Orbit Type



- On the first launches the most common orbits were LEO, ISS, PO and GTO.
- On the more recent launches, the most popular orbit type is the VLEO, with a very good success rate.

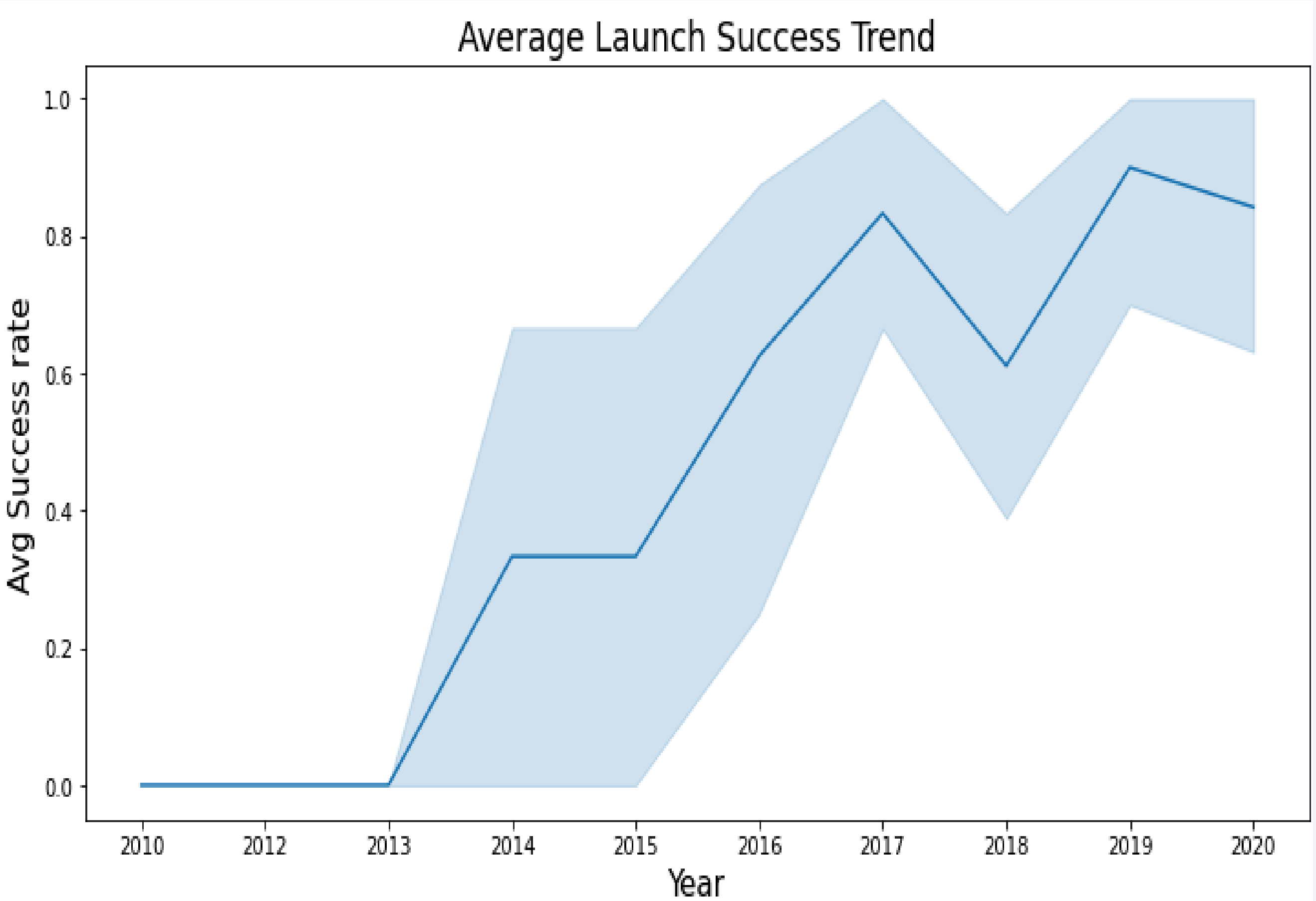
Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for Polar, LEO and ISS
- For the GTO orbit there is no clear pattern between payload mass and successful or unsuccessful landings.

Launch Success Yearly Trend

- The first 3 years of launches (2010-2013) seems to have been a period of adjustments.
- From 2013 till 2020 the success rate almost constantly increased, except of the period 2017-2018 in which there was a slight decrease.



All Launch Site Names

- Searching for distinct values in the column "launch_site" we obtained the following four unique Launch Sites.

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

Launch Site Names Begin with 'CCA'

- Five records where the launch site is on the Cape Canaveral area.

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

Total Payload Mass

- To calculate the total payload mass that were carried by NASA boosters, it was necessary to sum the payload mass of every booster whose payload had 'CRS' on it.

TOTAL PAYLOAD MASS(KG)
111268

Average Payload Mass by F9 v1.1

- Calculating the average payload mass carried by booster version F9 v1.1 by filtering the data according to the booster version in use.

AVERAGE PAYLOAD MASS(kg)
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was obtained by choosing the minimum value from all dates where a successful ground pad landing has occurred.

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- By filtering first all the boosters which have successfully landed on drone ship and then taking only those who had payload mass greater than 4000 but less than 6000, we get the following four booster versions.

BOOSTER VERSION
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- After grouping the data based on the mission outcome and the counting the records for each group we get the 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

- By identifying, first, the maximum payload mass and then searching which boosters have carried payload mass equal to the maximum, we get a list of all the 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
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

LANDING OUTCCOME	BOOSTER VERSION	LAINCH 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

- Ranking 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.

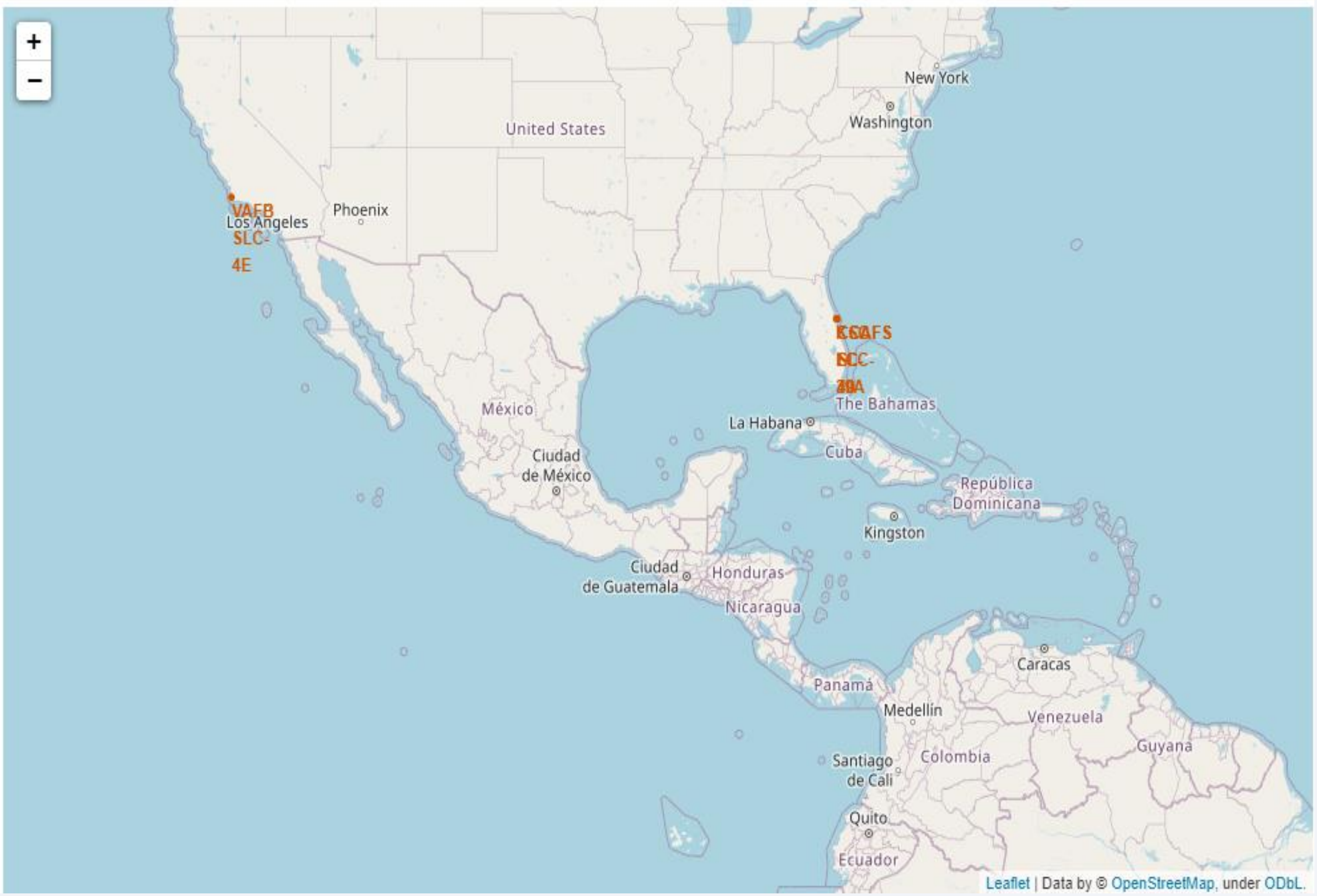
LANDING OUTCOME	COUNT OF LANDING OUTCOME
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

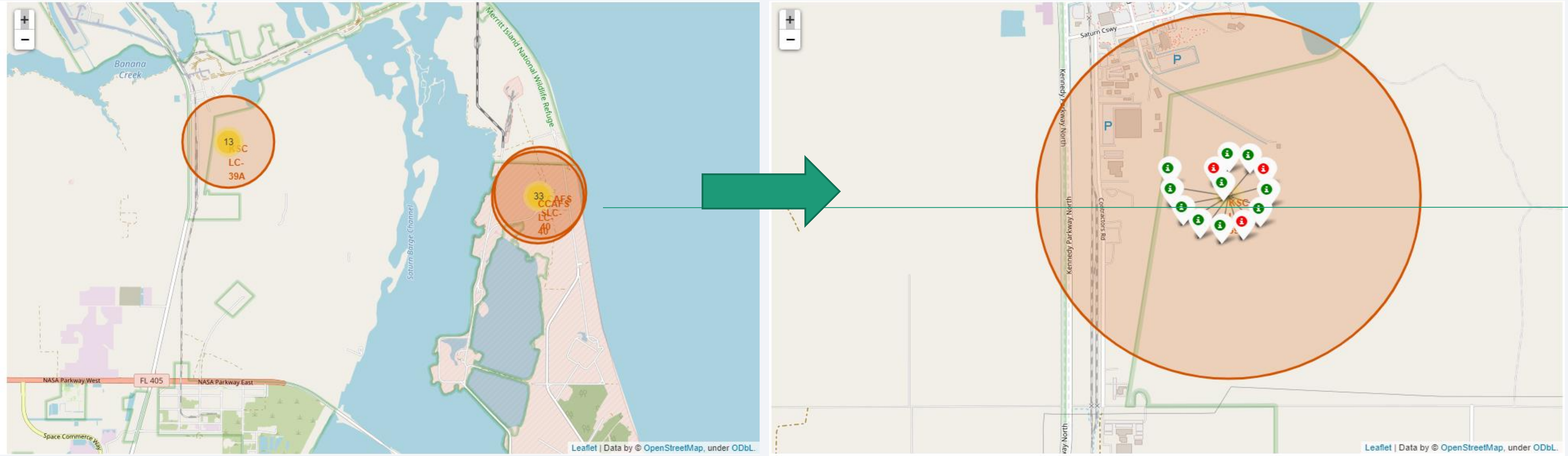
Launch Sites Locations

- The launch sites are very close to the equator.
- All the sites are in a very clear proximity to the coast.



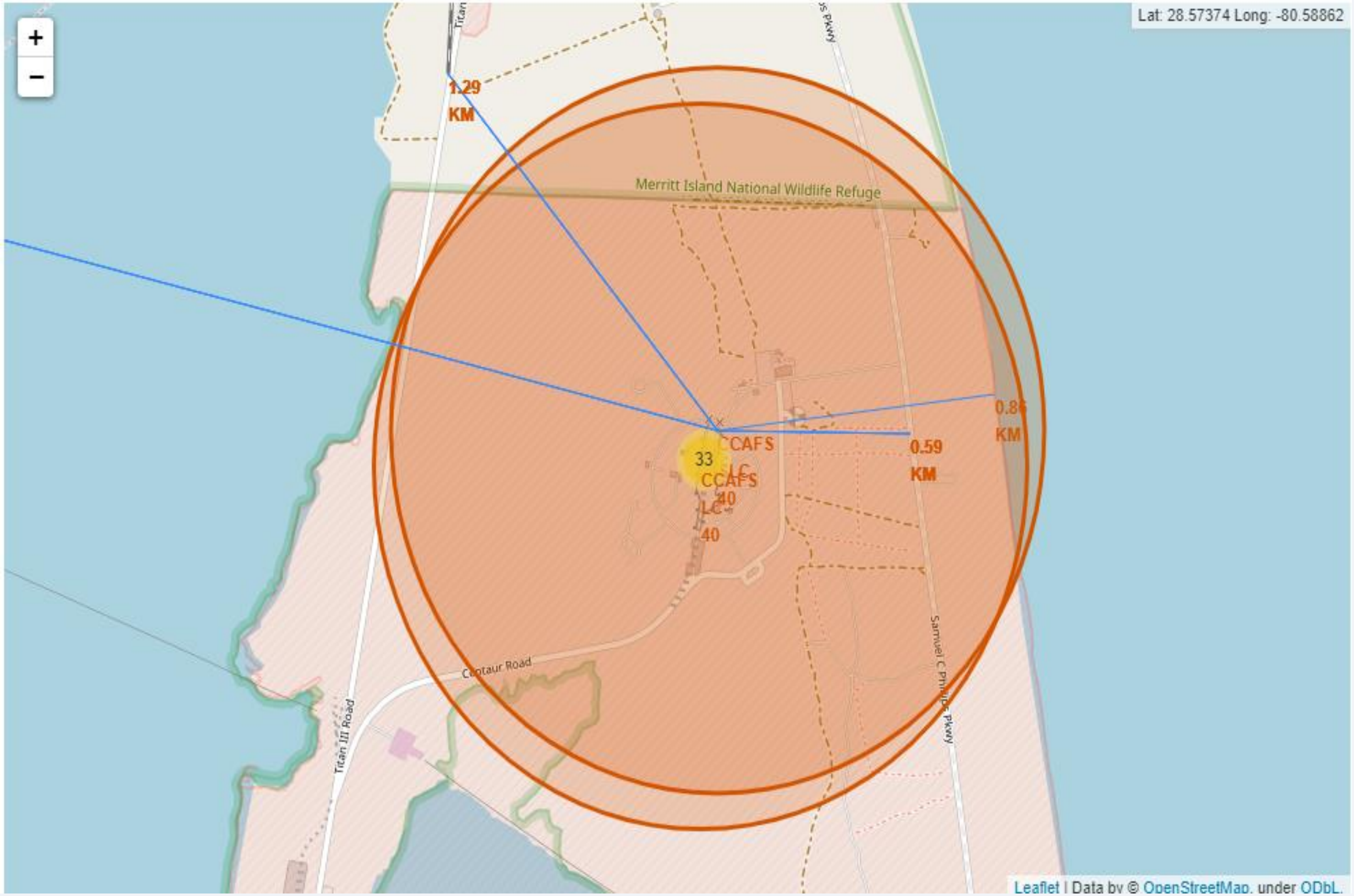
Launch Outcomes by site

- On the left image there are clusters of markers that visualize the launch outcome for every launch site.
- On the right image are depicted the markers on the launch site KSC LC-39A.
- Green markers indicate successful landing outcome while red indicate failure.



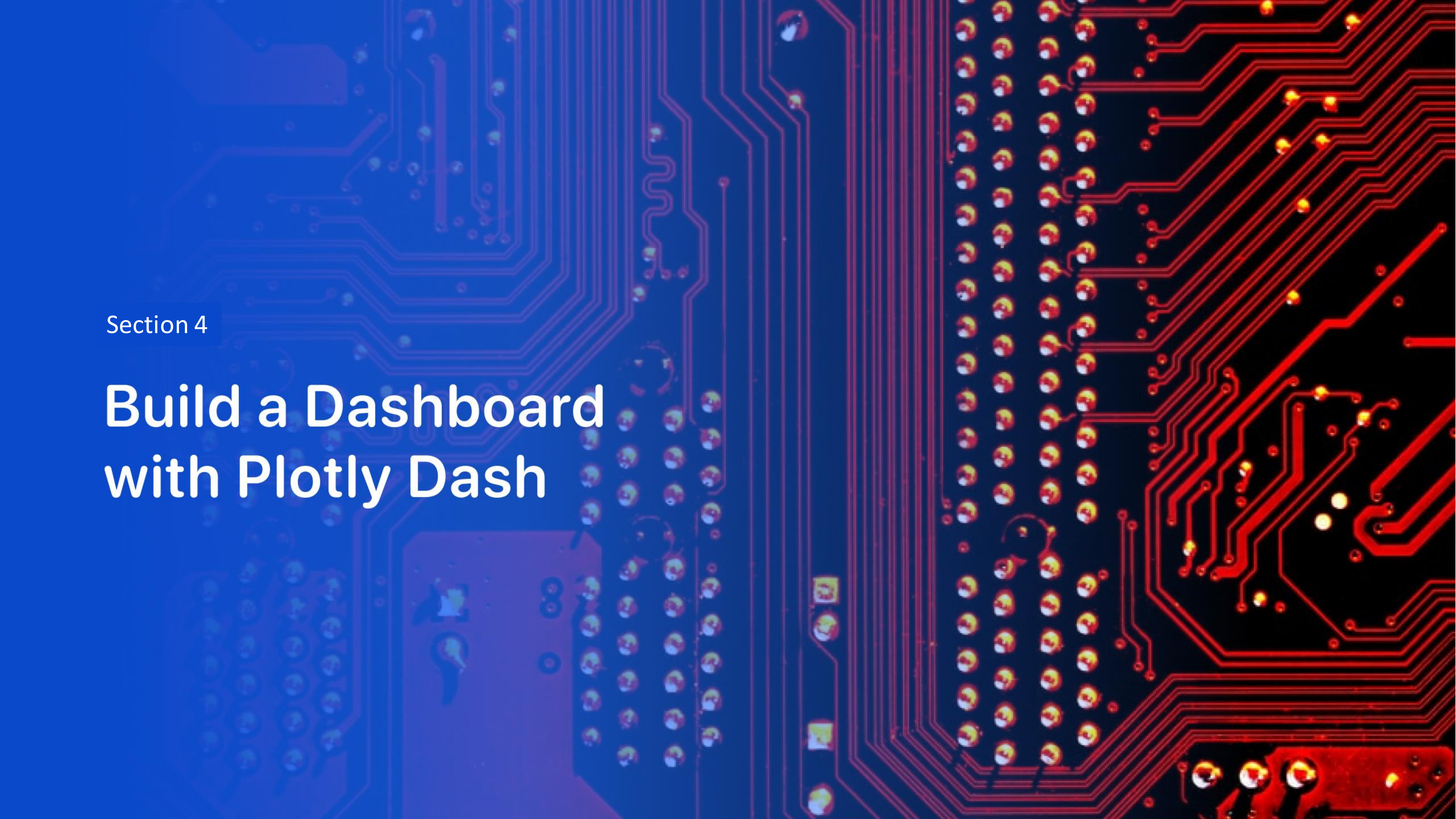
Relative position of launch sites

- Launch site CCAFS SLC-40 with distances calculated to its closest proximities is depicted in the adjacent image.
- The launch sites position seems to be strategically chosen so it can be as close as possible to every major transportation network such as railways, highways and coastline while as far as possible from residential areas.


+
−
Lat: 28.57374 Long: -80.58862
1.29
KM
Merritt Island National Wildlife Refuge
0.86
KM
0.59
KM
33
CCAFS
LC
CCAFS
40
LC
40
Centaur Road
Titan III Road
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

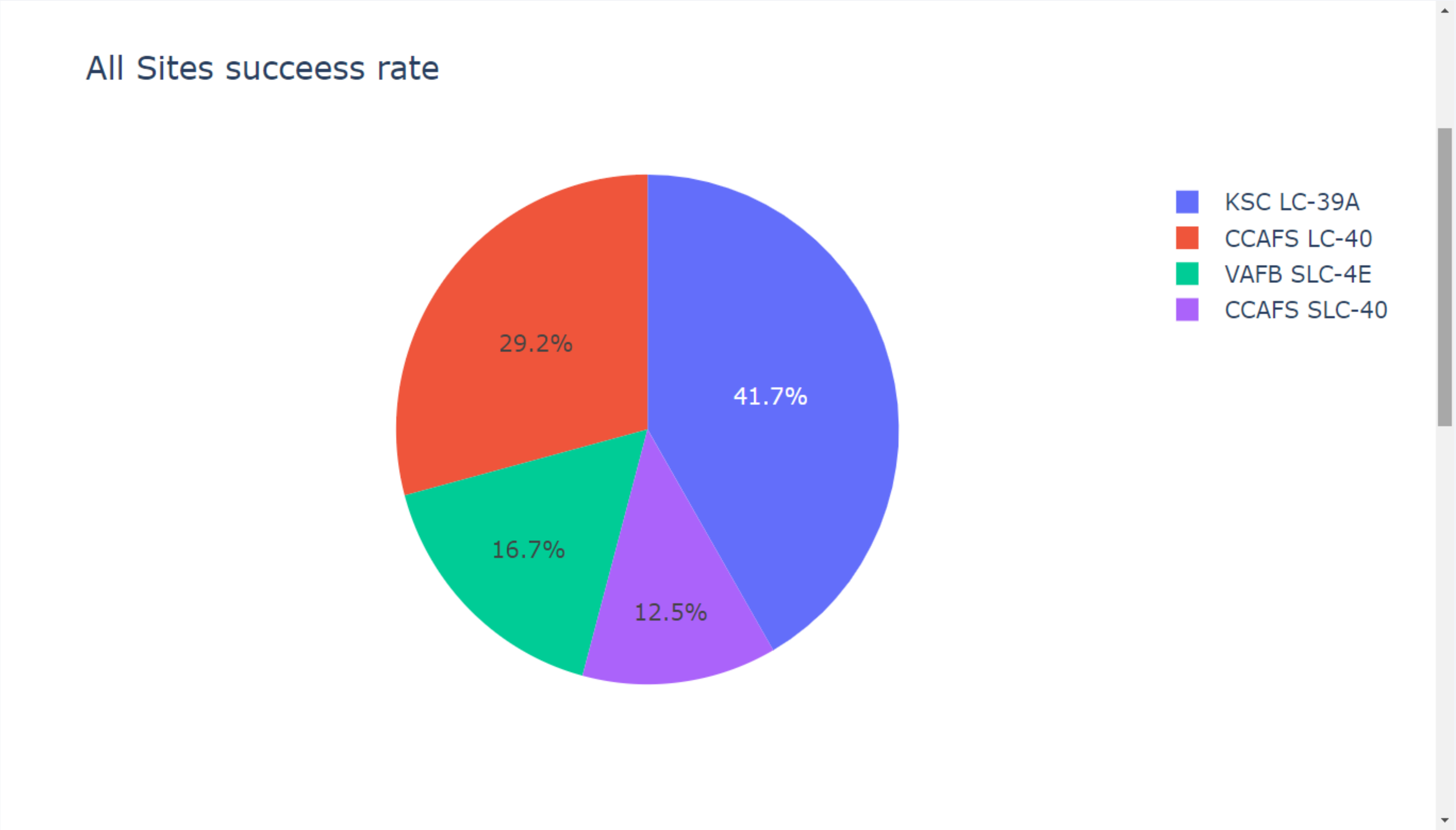
Section 4

Build a Dashboard with Plotly Dash



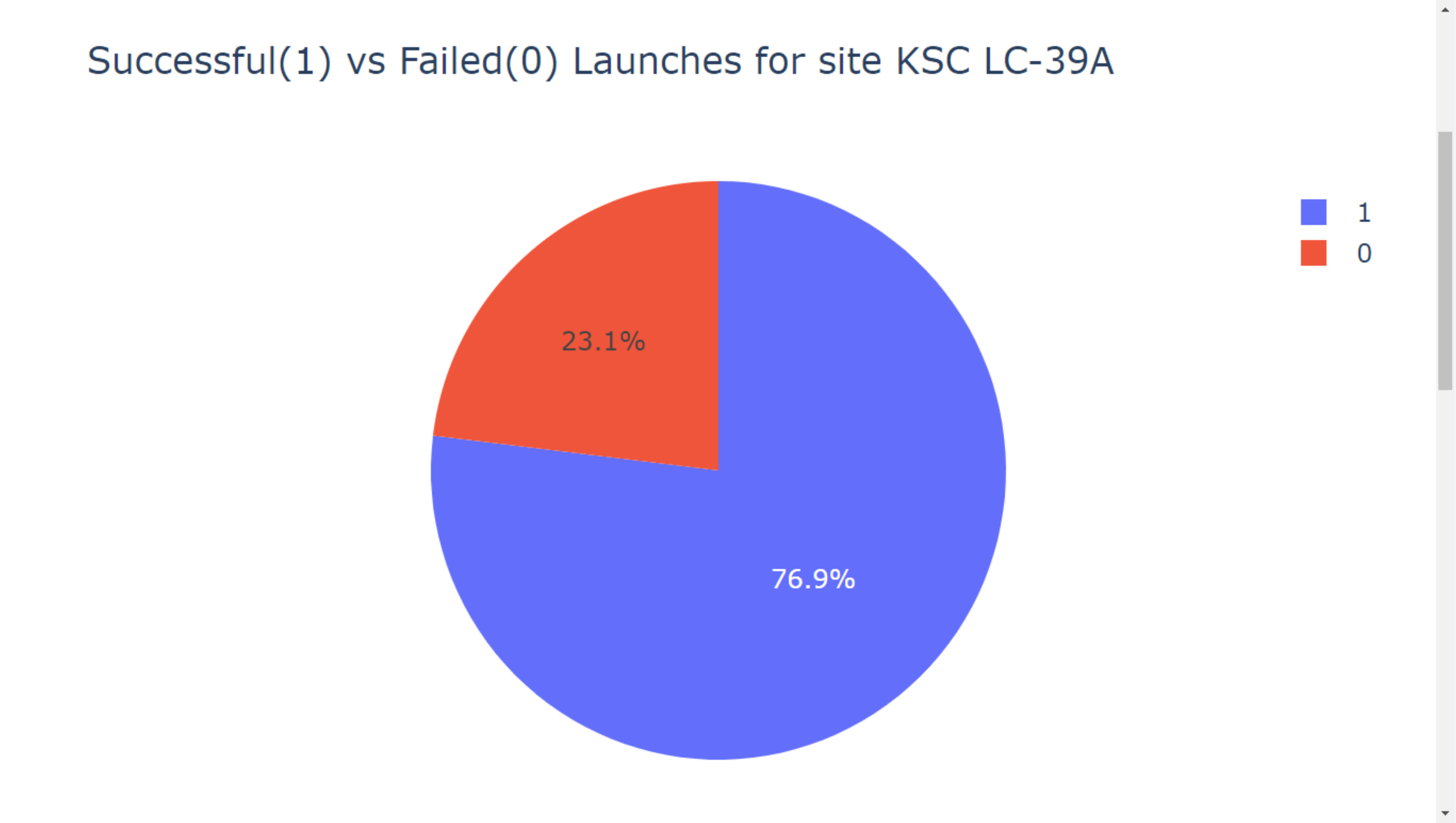
Success rate of launch sites

- The landing outcome of a mission can vary depending on the site from which it was launched.



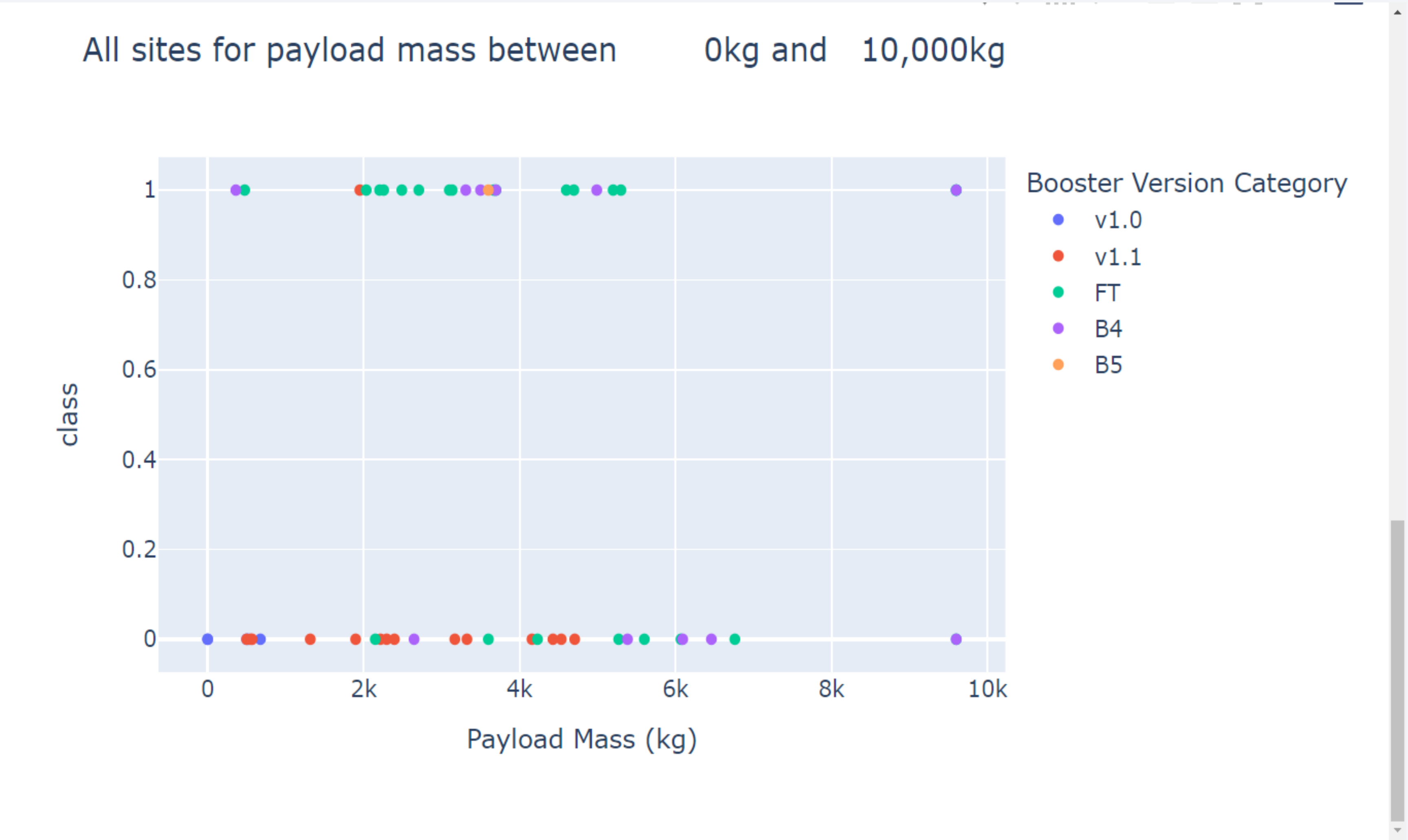
Most successful launch site

- Launch site KSC LC-39A is the site with the highest success rate.



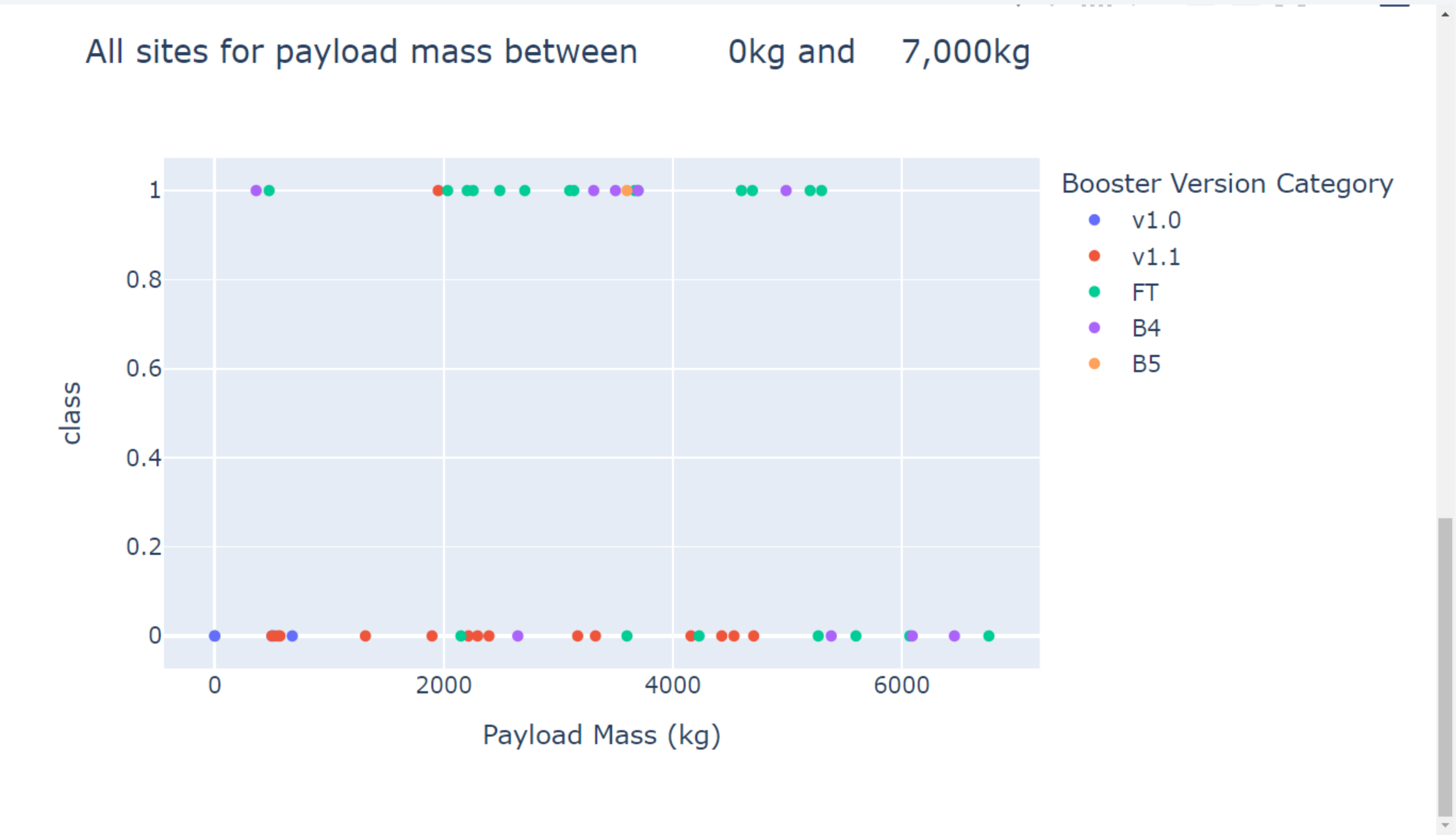
Payload vs Launch Outcome

- The majority of launches had a Payload Mass approximately between 0 and 7,000kg.



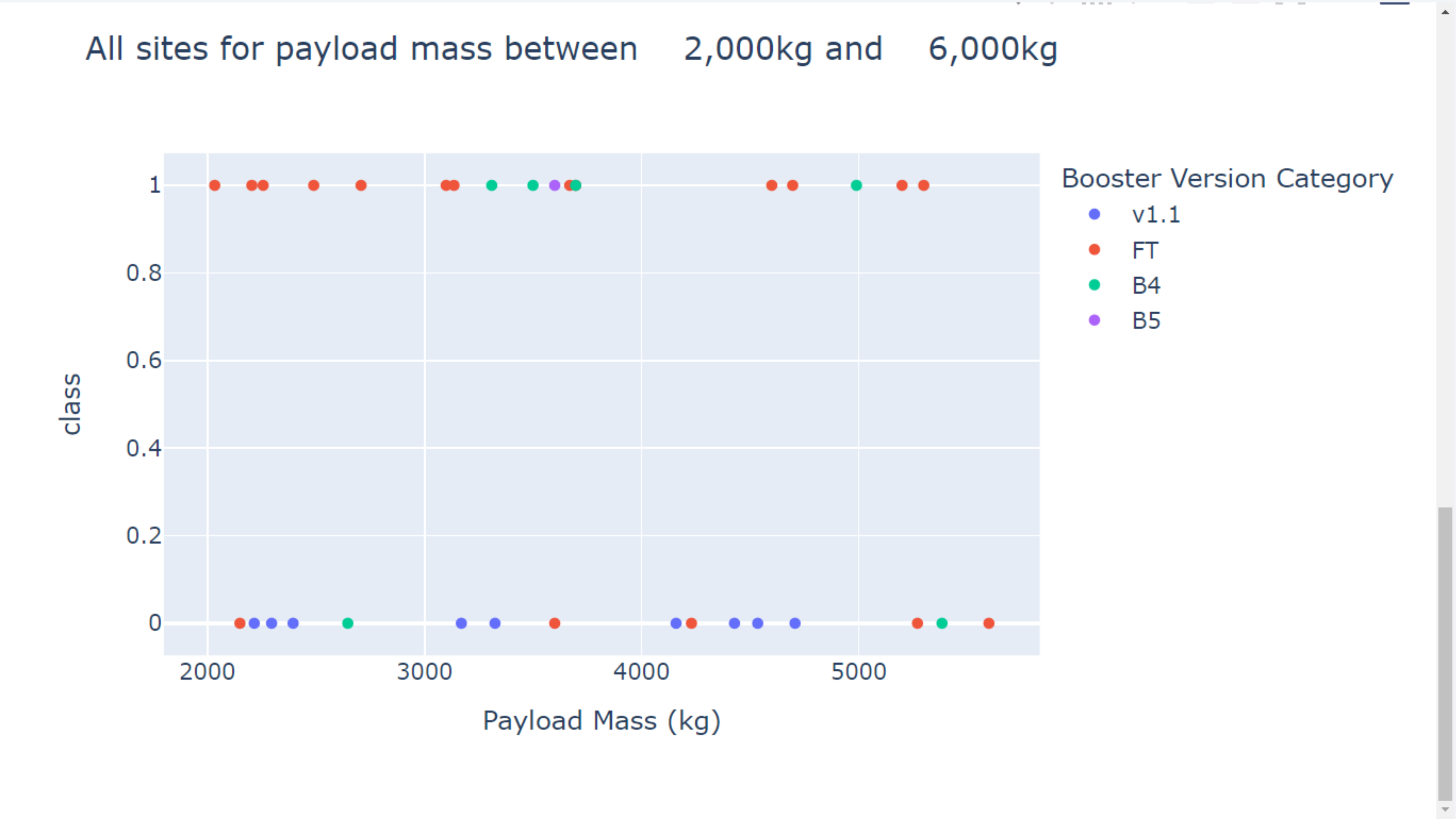
Payload vs Launch Outcome

- The majority of the successful lau nches seem to occur for Payload Mass approximately be tween 2,000 and 6,000kg.



Payload vs Launch Outcome

- The Booster Version with most successful landings for a payload mass between 2,000 and 6,000 kg is the Falcon 9 FT (Full Thrust).

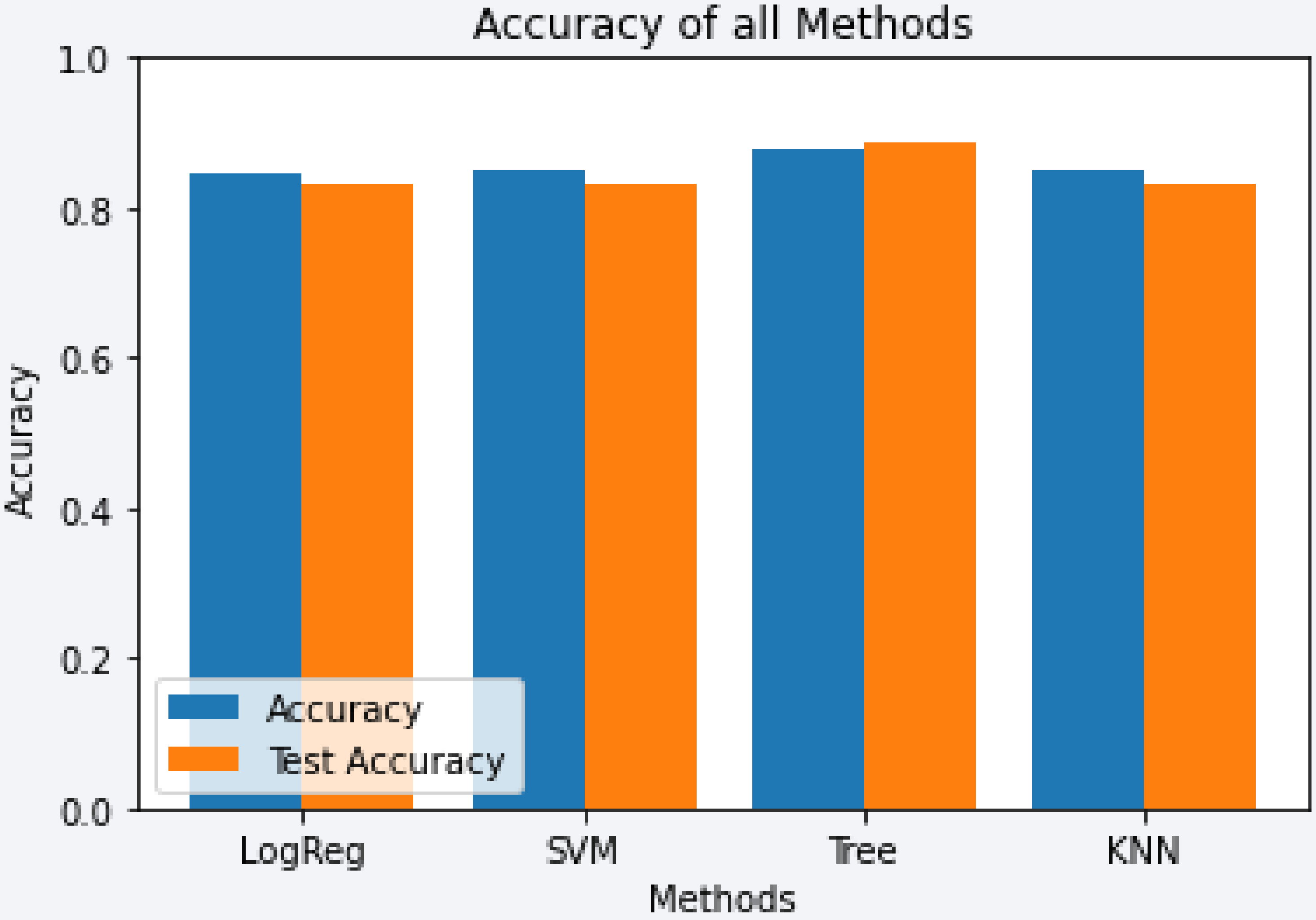


Section 5

Predictive Analysis (Classification)

Classification Accuracy

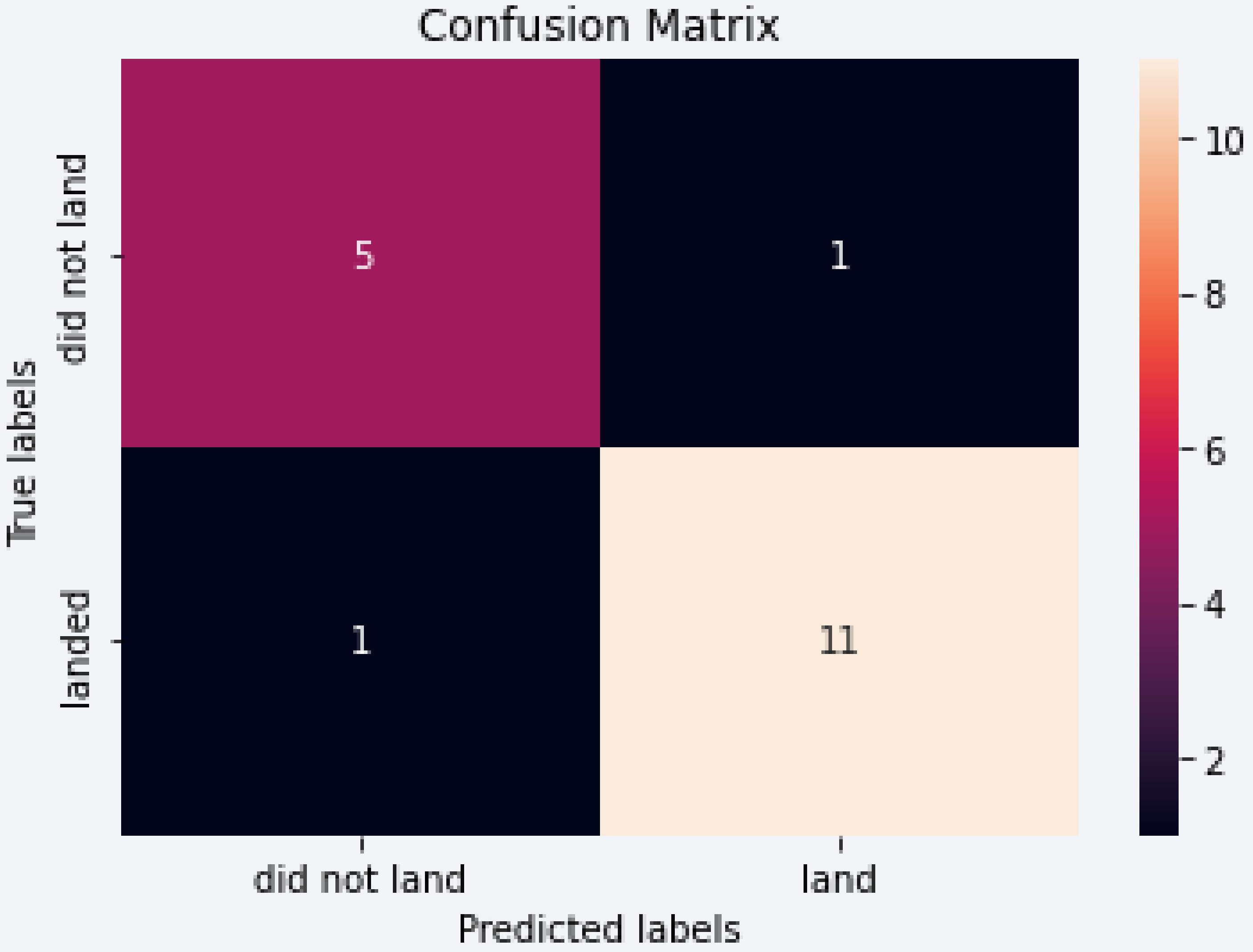
- The blue bar represents the accuracy score of each method on the training data while the orange bar the accuracy on the test data.
- From the bar chart we can conclude that the Classification Tree method has the best performance both on the training as well as on the test data.



Confusion Matrix

- There is only one False Positive and one False Negative

- The ability to predict True Positives is remarkable.



Conclusions

- CCAFS SLC-40 is the most used launch site.
- Percentage-wise KSC LC-39A is the most successful launch site.
- Launch sites CCAFS SLC-40 and KSC LC-39A can be used for missions where the Payload Mass is over 10,000kg.
- Missions where the Payload Mass is over 8,000kg succeed to land the first stage almost 100% of the time.
- Over time the success rate of landings has increased significantly
- Orbit type VLEO(Very Low Earth Orbits) it the most probable to result in a successful landing.
- Most accurate way to predict the landing outcome of the first stage is using the Decision Tree Classifier.

Thank you!