IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gerardo Avelar

October 2023

Outline

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

Executive Summary

We utilized several Machine Learning algorithms to **find the best predictor of successful launches** using historical data of more than 80 features (characteristics) of previous launches of this nature.

K-nearest-neighbors provided the highest precision and recall among all other classifier algorithms utilized. We will outline results of all algorithms during this presentation.

Introduction

In order for SpaceY to become a relevan player in the rocket launching market, we need to have a solid analytical background that enables us to maximize the probability of successful launches.

This presentation will provide in-depth details of different Machine Learning methodologies that aim to help us answer the question:

What is the probability that a launch will be successful?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We utilized publicly available data from SpaceX API and webscraped data from Wikipedia
- Perform data wrangling
  - Standard Scaling was applied to the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tested several hyperparameters on different ML Classifier Algorithms such as Decision Trees, Support Vector Machines, Logistic Regressions and others.

Data Collection

- Data sets were collected from:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used

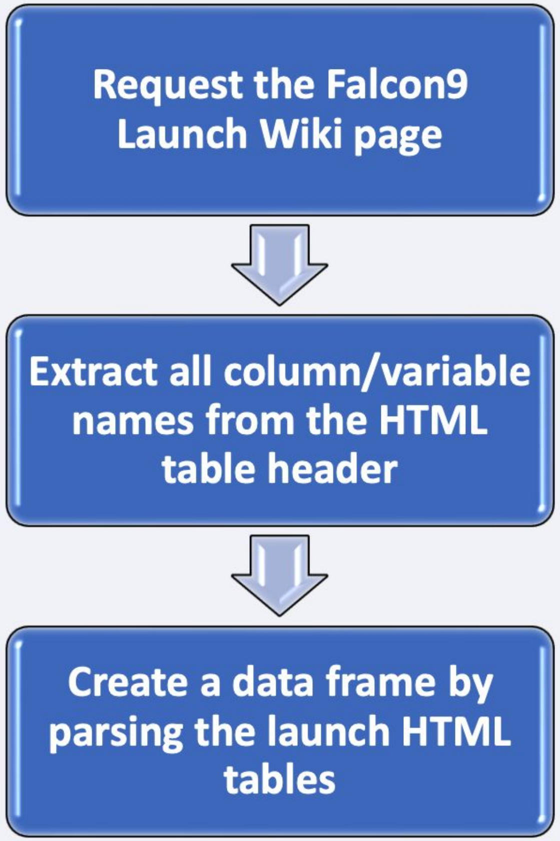
- This API was used according to the flowchart beside and then data is persisted.

Notebook source

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia

- Data are downloaded from Wikipedia according to the flowchart and then persisted.



[Notebook source]

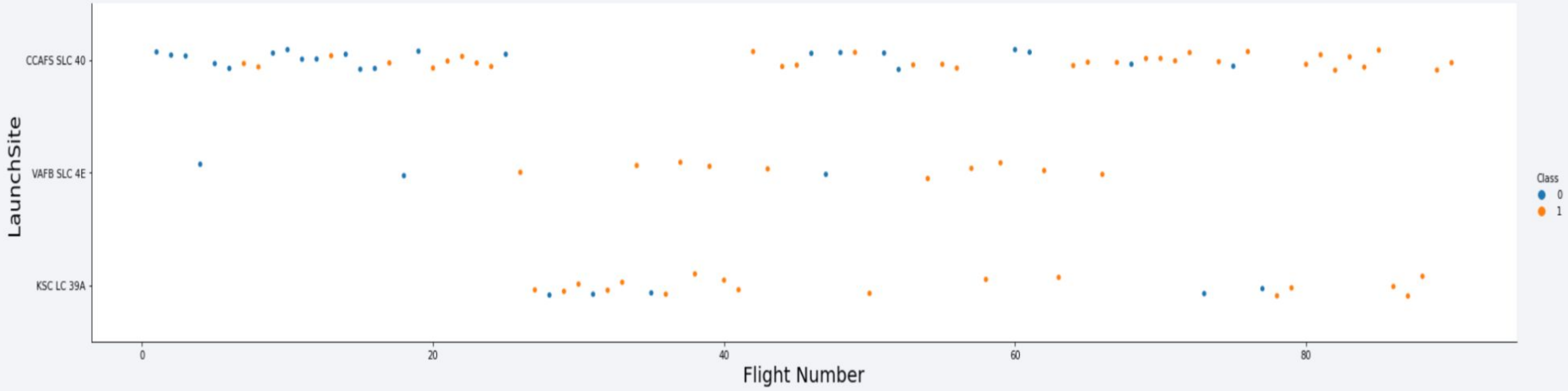
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.

Notebook source

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
    - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



[Notebook source]

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin 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 when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (groundpad)) between the date 2010-06-04 and 2017-03-20.

[Notebook source]()

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.

[Notebook source]

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 allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

[Source Code](#)

Predictive Analysis (Classification)

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

Notebook source

Results

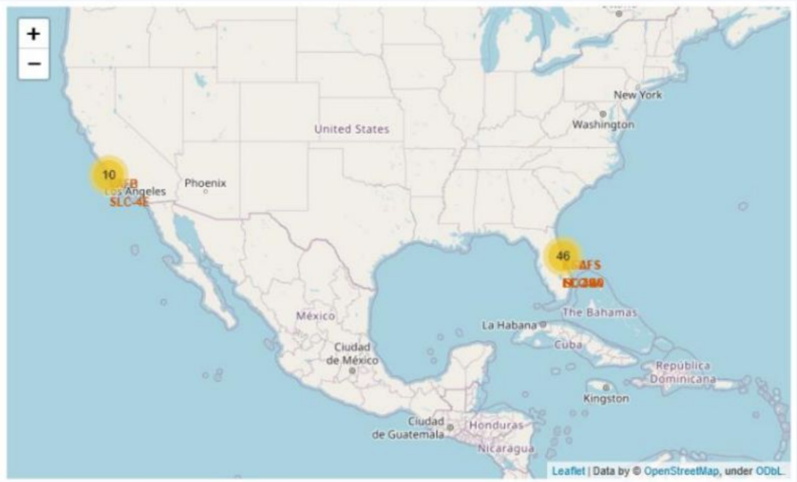
Exploratory data analysis results:

- 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

Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.

Most launches happens at east coast launch sites



Results

Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.

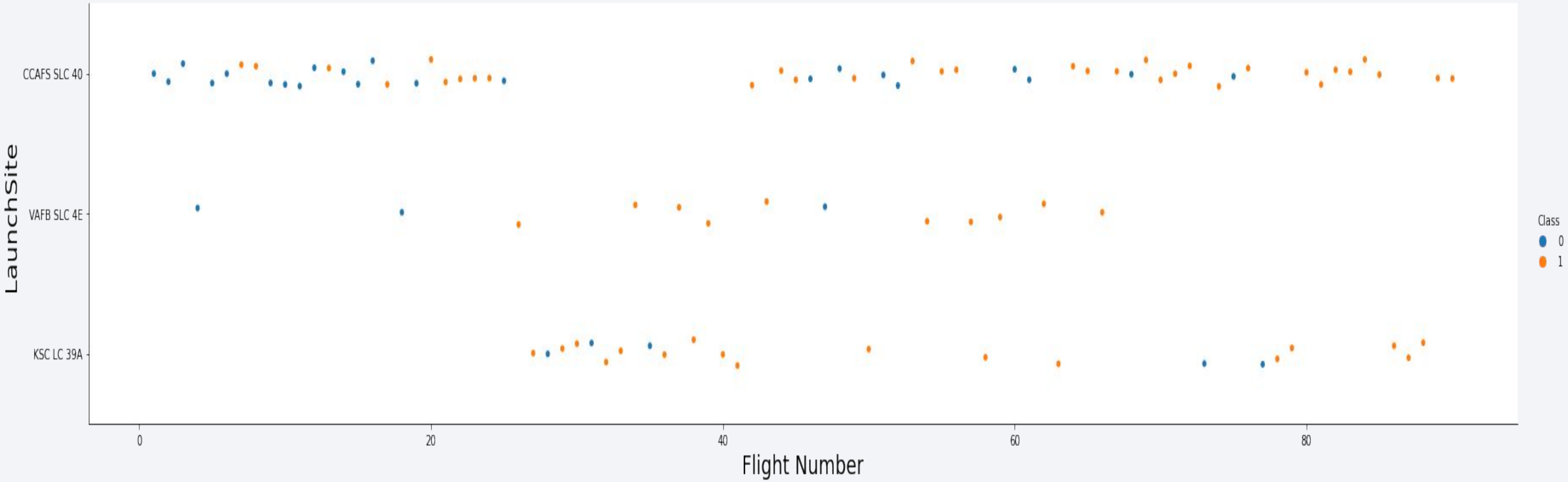
Most launches happens at east coast launch sites

Section
2

Insights drawn from EDA

Flight Number vs. Launch Site

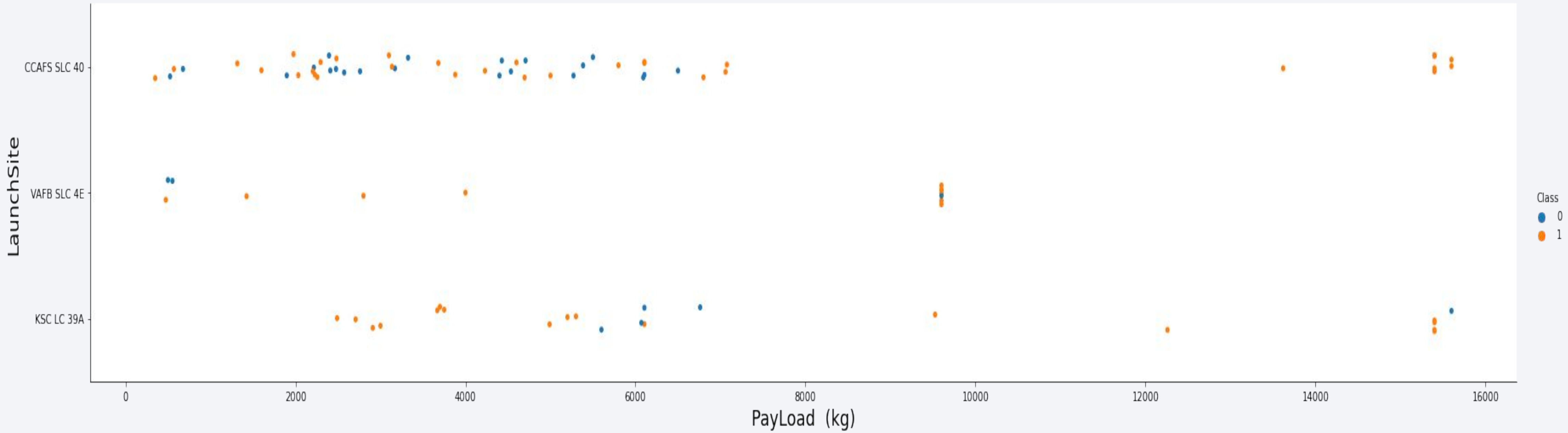
Visualize the relationship between Flight Number and Launch Site
We see that different launch sites have different success rates. But as we increase the number of flights the success rate increase.



Payload vs. Launch Site

Visualize the relationship between Payload and Launch Site

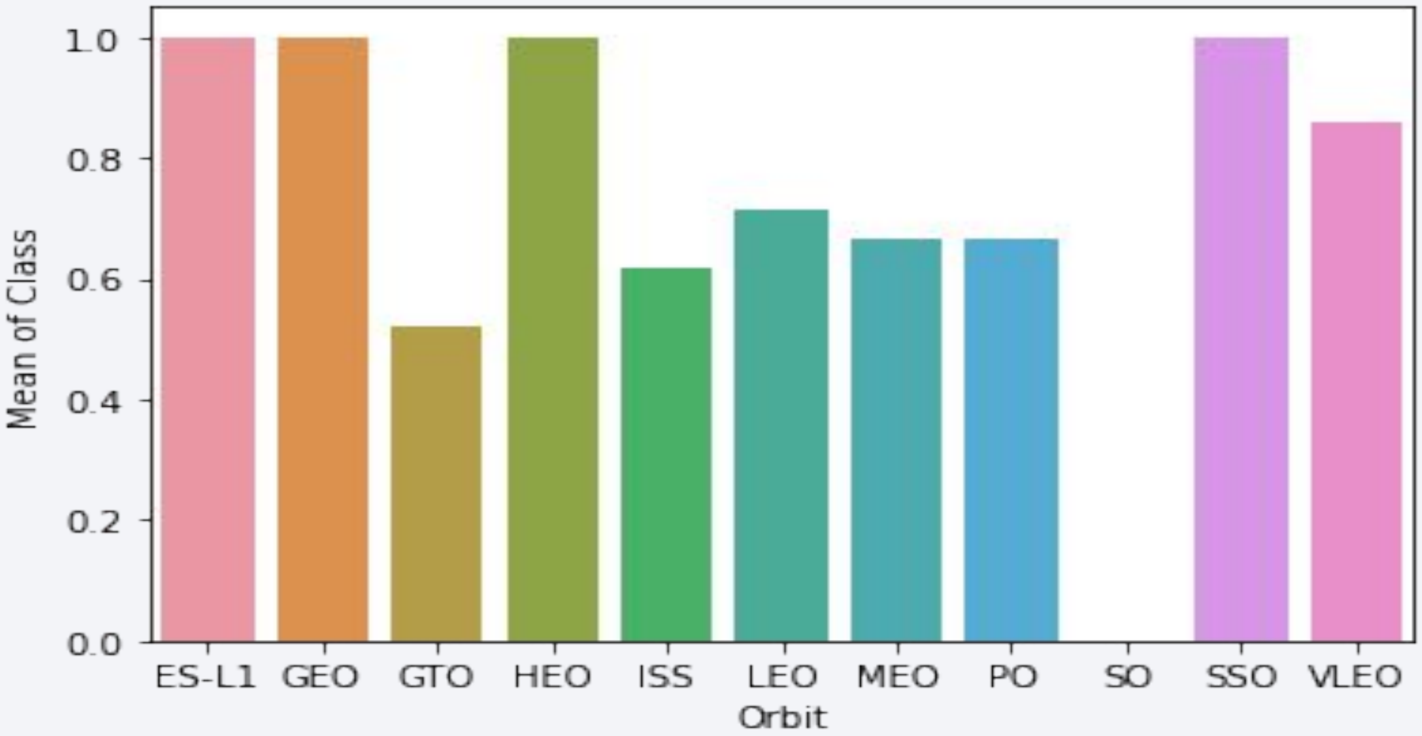
As well, if increase the number of we Pay Load Mass (kg) the success rate increase.



Success Rate vvss.. Orbit Type

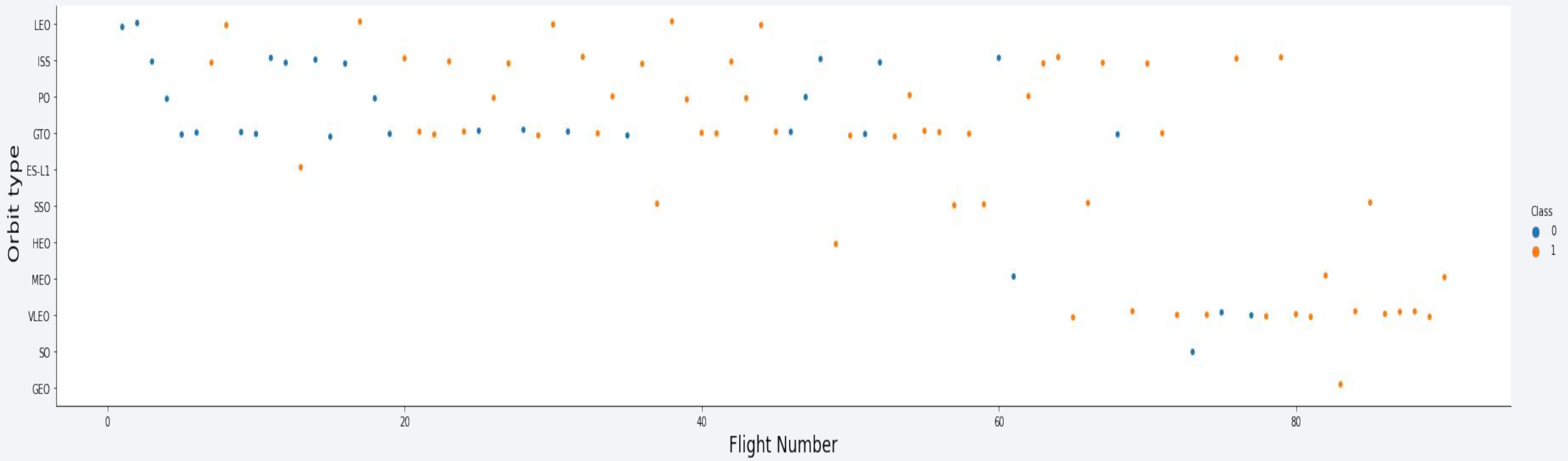
Visualize the relationship between success rate of each orbit type

- AS we can see ES-L1 , GEO, HEO and SSO have a success rates 100%.



Flight Number vs.. Orbit Type

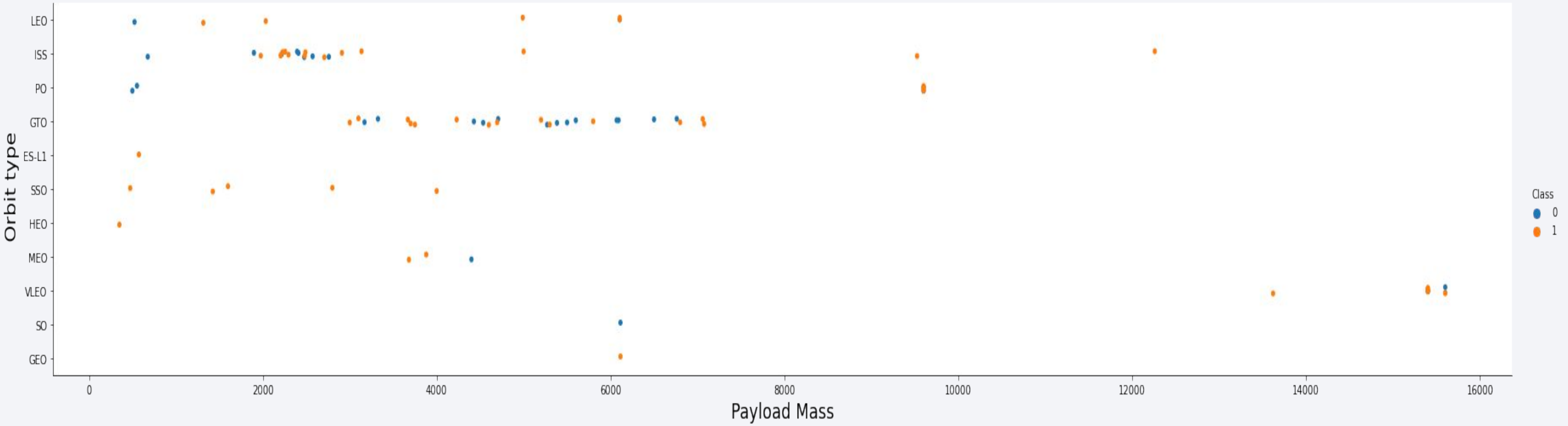
Visualize the relationship between FlightNumber and Orbit type



Payload vs. Orbit Type

Visualize the relationship between Payload and Orbit type

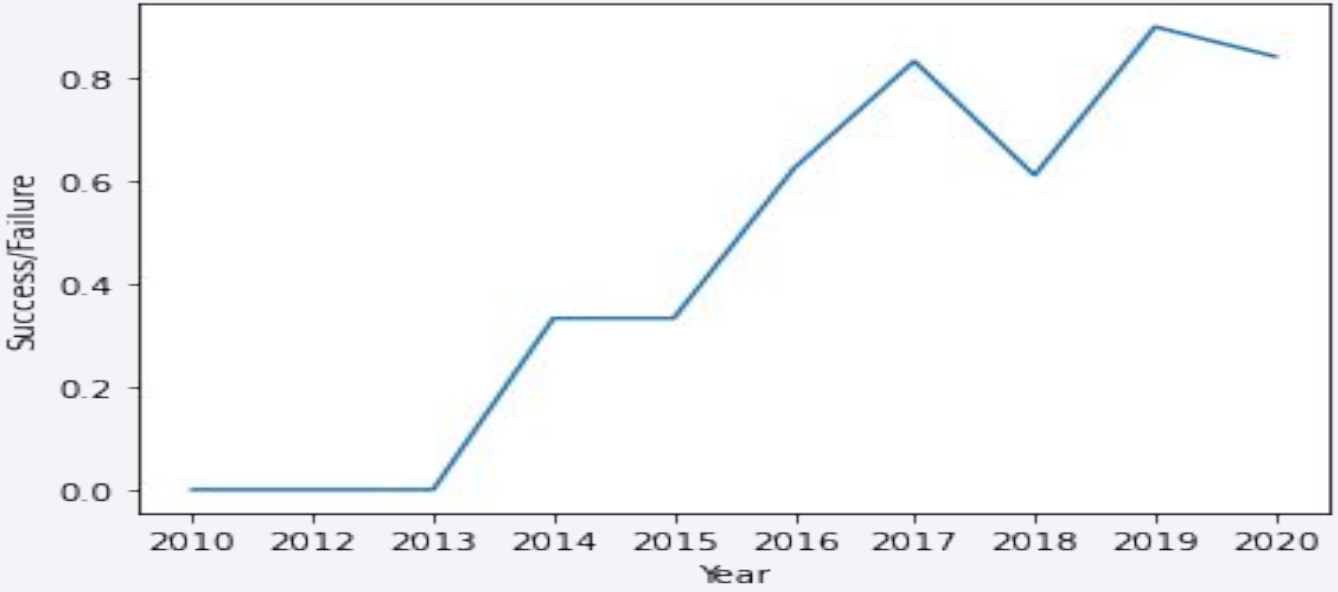
There is a connection between ISS and payload in the range of 2000 to 3000. Also between GTE and Payload at 4000 to 8000.



Launch Success Yearly Trend

Visualize the launch success yearly trend

We can note in the figure that the launch sites begin to increase in success rate from 2013 to approximately 2018, decreases slightly, and then returns to increase with the passage of the year



All Launch Site Names

- %sql SELECT DISTINCT(launch_site) FROM Spacex

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

Launch Site Names Begin with "CCA"

- %sql SELECT * FROM Spacex WHERE launch_site LIKE'CCA%' LIMIT 5

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- %sql SELECT SUM(payload_mass kg_) FROM Spacex WHERE customer='NASA (CRS)'

1
45596

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(payload_mass kg_) FROM Spacex WHERE booster_version='F9 v1.1'

1
2928

First Successful Ground Landing Date

- %sql SELECT MIN(DATE) FROM Spacex WHERE landing outcome='Success (ground pad)'

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT booster_version FROM Spacex WHERE landing outcome='Success (drone ship)'
  AND payload_mass kg_ 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

- %sql SELECT COUNT(MISSION_OUTCOME) AS missionoutcomes FROM Spacex WHERE mission_outcome LIKE 'Success%'

missionoutcomes
100

- %sql SELECT COUNT(MISSION_OUTCOME) AS missionoutcomes FROM Spacex WHERE mission_outcome LIKE 'Failure

missionoutcomes
1

Boosters Carried Maximum Payload

- %sql SELECT booster_version AS Maxboosterversion FROM Spacex WHERE payload_mass kg_=(SELECT
MAX(payload_mass kg_) FROM Spacex)

maxboosterversion
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

- %sql SELECT landing   outcome,booster_version,launch_site,DATE FROM Spacex
  WHERE
  landing   outcome='Failure (drone ship)' AND EXTRACT(YEAR FROM
  DATE)='2015'

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-10-01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

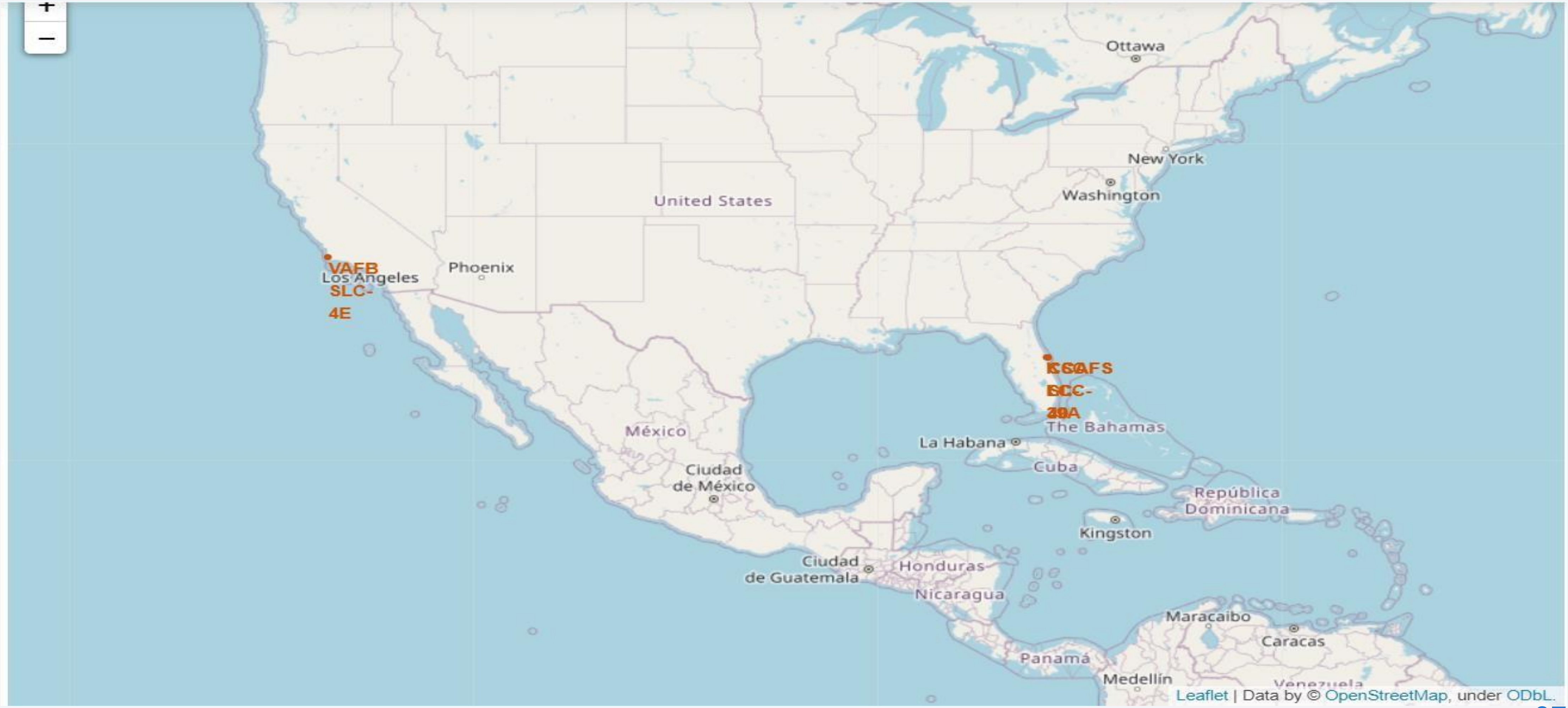
- %sql SELECT landing outcome , COUNT(landing outcome) FROM Spacex WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing outcome ORDER BY COUNT(landing outcome) DESC

landing__outcome	2
No attempt	10
Success (drone ship)	6
Failure (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

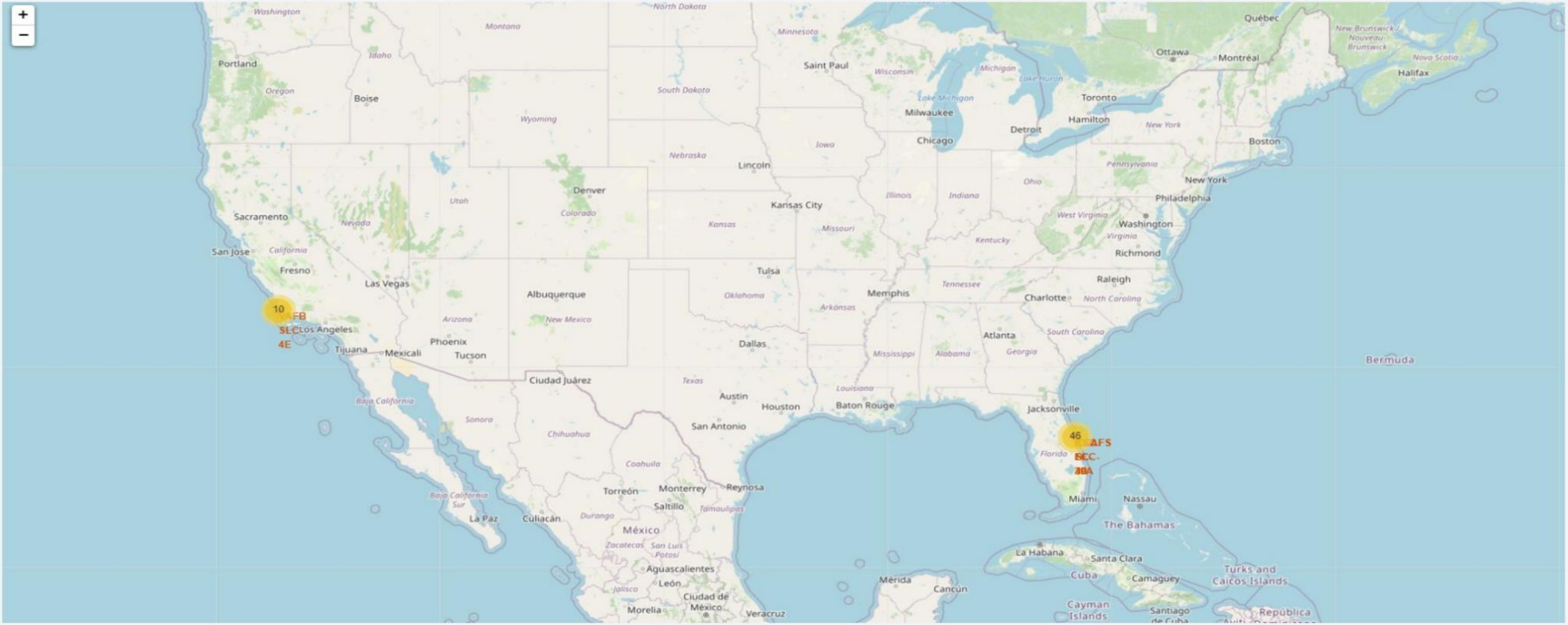
Section
3

Launch Sites Proximities Analysis

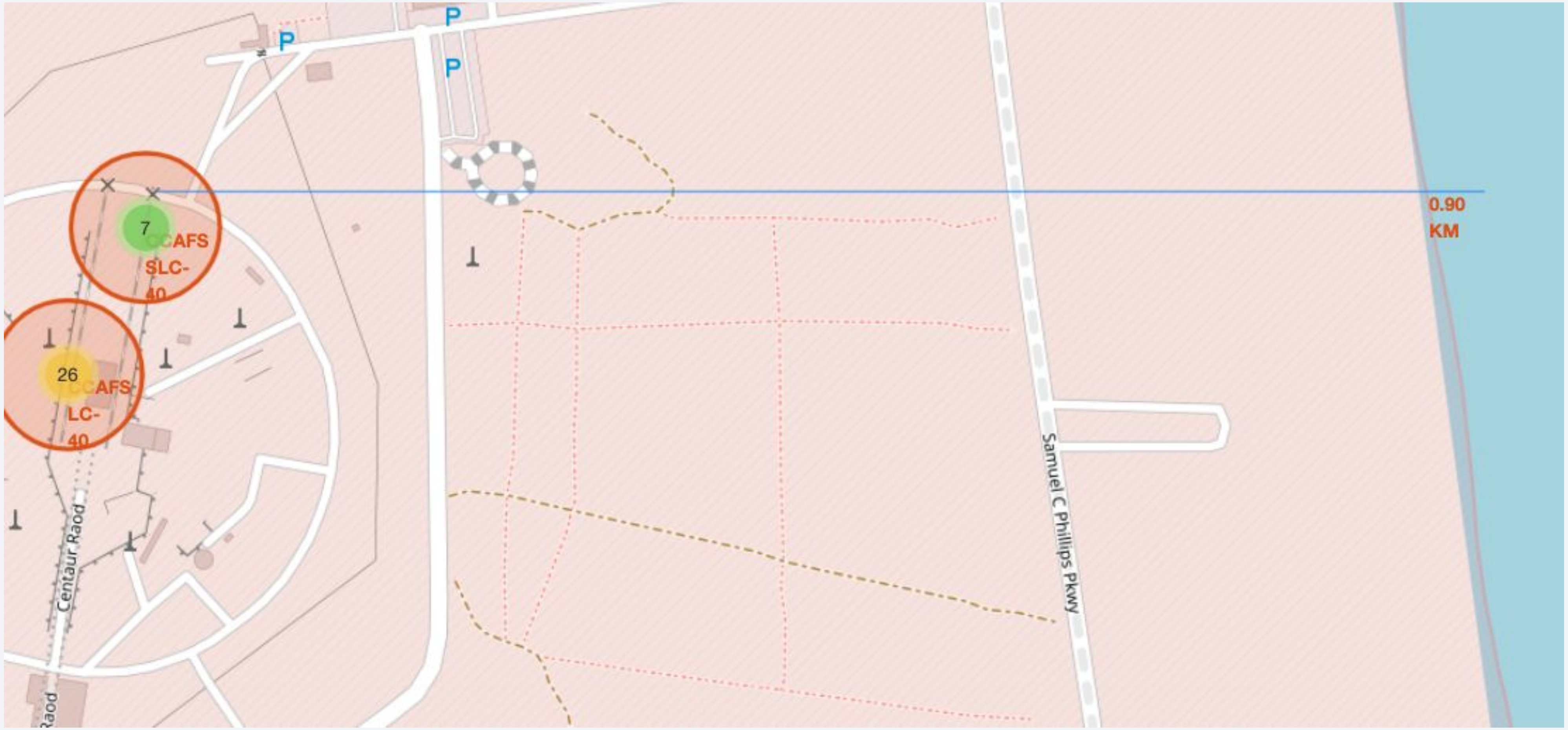
All launch sites on the site map



Launch sites by using Marker Cluster



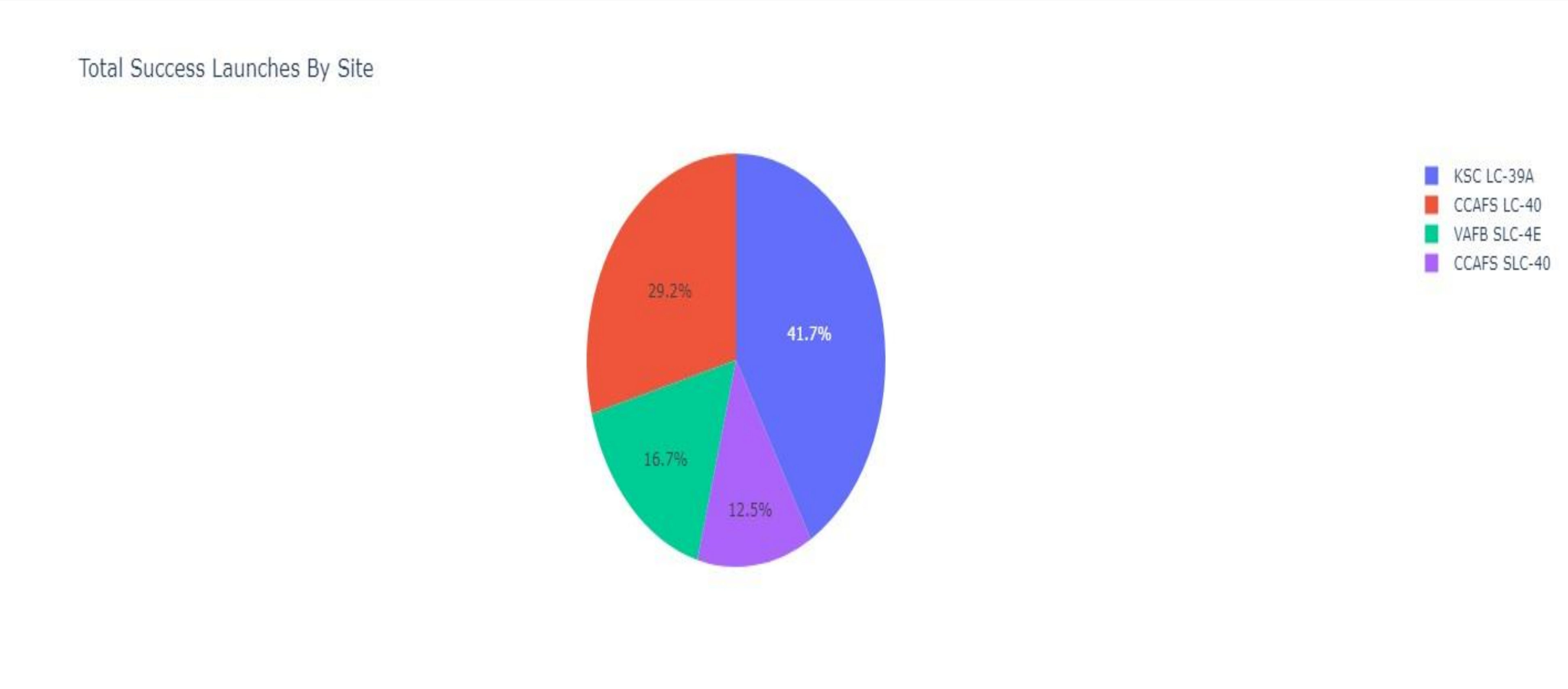
Marker distance between the coastline point and the launch site



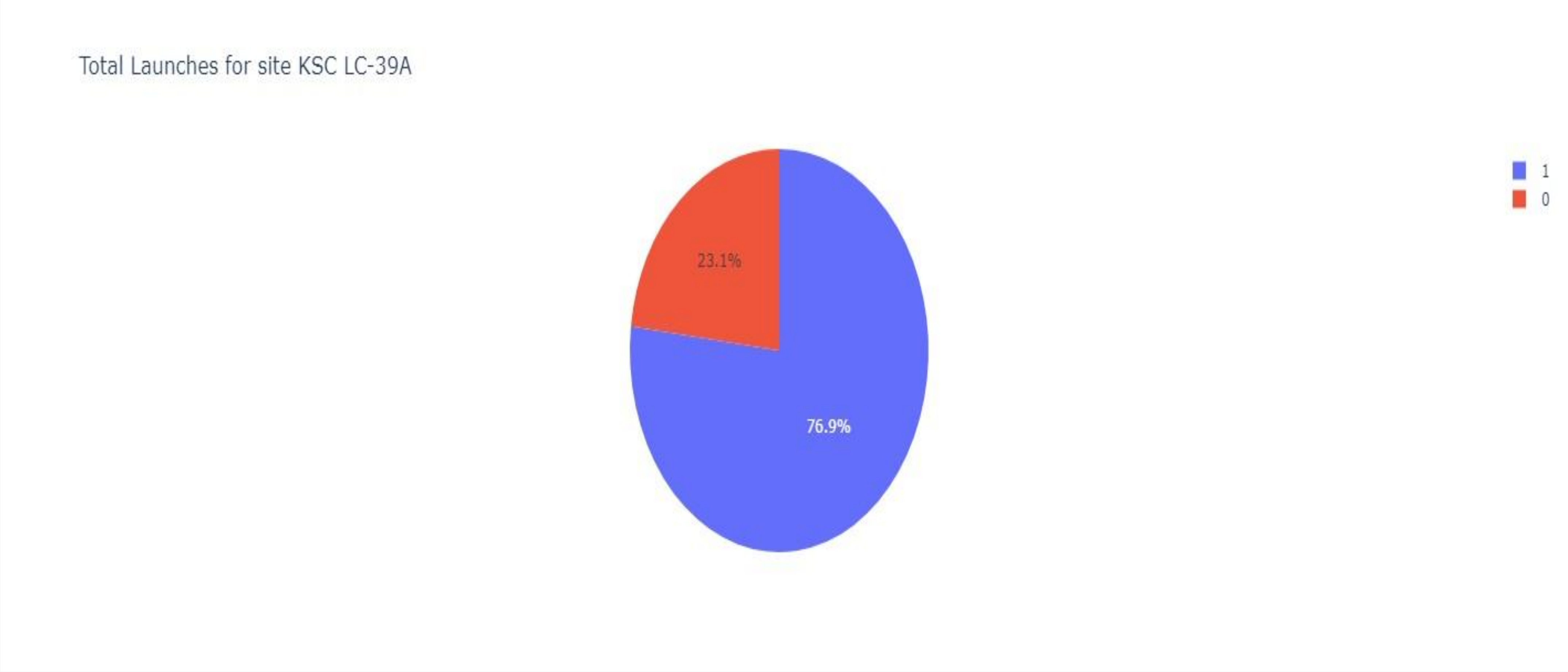
Section
4

Build a Dashboard with Plotly Dash

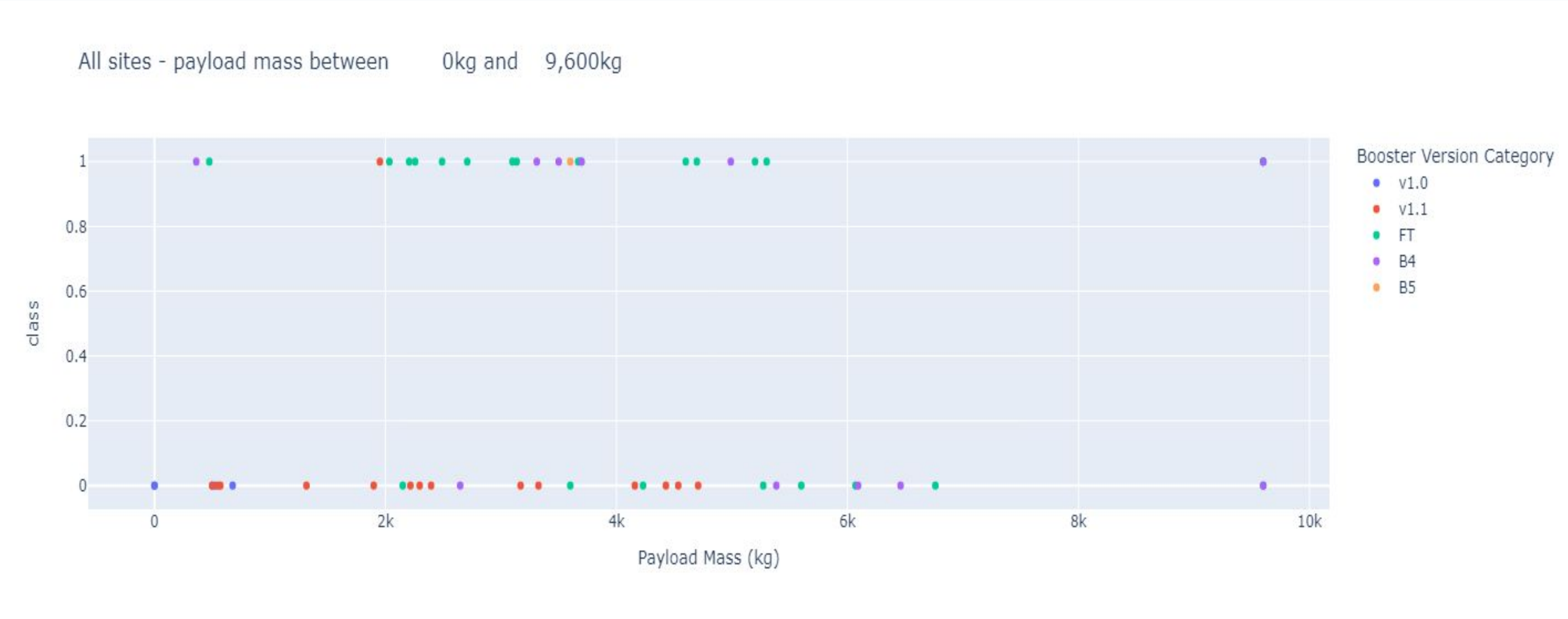
Total Success Launches



Launches for Hight site Score



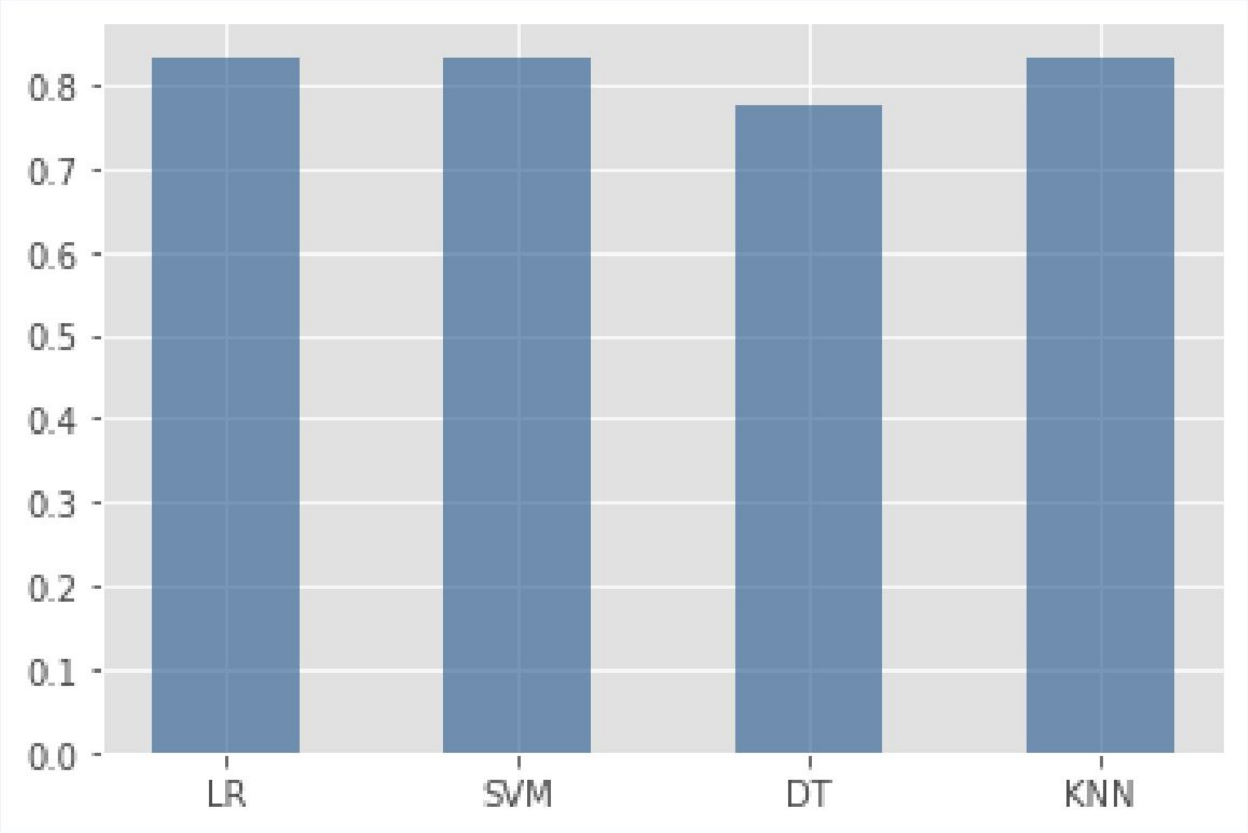
Payload vs. All Launch



Section
5

Predictive Analysis (Classification)

Classification Accuracy



As we can see Decision Tree has the highest accuracy with almost 0.89, then comes the remaining models with almost the same accuracy of 0.84.

Confusion Matrix

Measure	Derivations	Result
Precision	PPV = TP / (TP + FP)	0.67
Accuracy	ACC = (TP + TN) / (P + N)	0.89
F1 Score	F1 = 2TP / (2TP + FP + FN)	0.80

Conclusions

- After analysis data the CCAFS SLC-40 site and KSC LC-39A site are has most
  successful launches from all the sites.
- Orbit GEO,HEO,SSOES L1 has the best Success Rate.
- The payload of O kg to 5000 kg was more diverse than 6000 kg to 10000
- The Decision Tree model is the best in terms of prediction accuracy for this dataset.

References

- Confusion Matrix : https://onlineconfusionmatrix.com/
- Matplotlib - Bar Plot : https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm
- https://towardsdatascience.com/7-points-to-create-better-histograms-with-seaborn-5fb542763169
- https://www.researchgate.net/figure/Bar-chart-showing-the-performance-evaluation-in-our-data-loading- tests_fig4_268150621

Thank you!