IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yanika Dontong
08/03/2022

Outline

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

Executive Summary

- Data collection methodology
  - Collect data through SpaceX API and scrap from Wiki
- Data Wrangling
  - Modify columns and deal with missing values
- EDA using visualization and SQL
- Interactive visual analytics on Folium and Plotly Dash
- Prediction analysis using classification model
  - Develop Logistic regression, SVM, Decision Tree and KNN models
- Summary of all results
  - A later launch at KSC LC-39A, has a light weight has the most success rate
  - The best predict models have 83.33% accuracy

Introduction

- SpaceX has a lower cost on rocket launch than others companies
- SpaceX can recover and reuse Stage One of rocket (which is large and expensive) on Falcon 9
- The Recovery of Stage One is not always success
- So, we will training the machine models to predict whether the Stage One can be recovery or not

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected by using SpaceX API and scraping on Wikipedia
- Perform data wrangling
  - Data is transformed by one-hot encoder on categorial 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
  - Split data set into train and test data
  - Develop various models and using Grid Search CV on train data to find best parameter
  - Assess accuracy score of each models using best parameter on test data

Data Collection

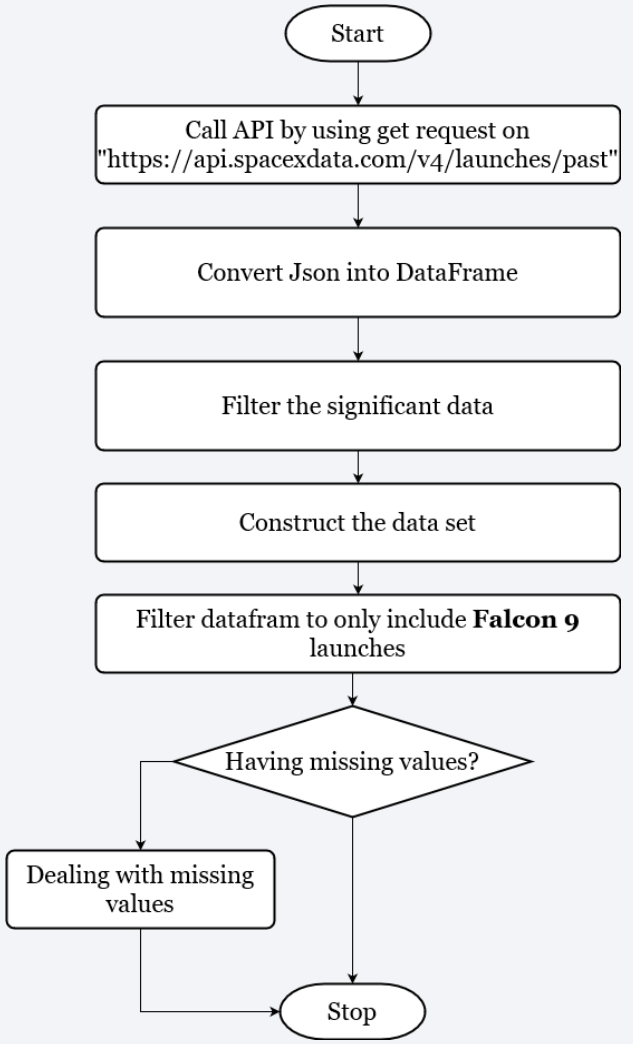
Using 2 Methods to collect data

- SpaceX API
  - Get request and parse the data to Json
  - Filter the features and Convert to Data frame
  - Clean data by dealing with missing values
- Web Scraping on Wikipedia
  - Request the Falcon9 Launch Wiki page from its URL
  - Extract data from HTML table header
  - Parse HTML table to data frame

Data Collection – SpaceX API

- This flowchart demonstrated how to collect data using SpaceX API

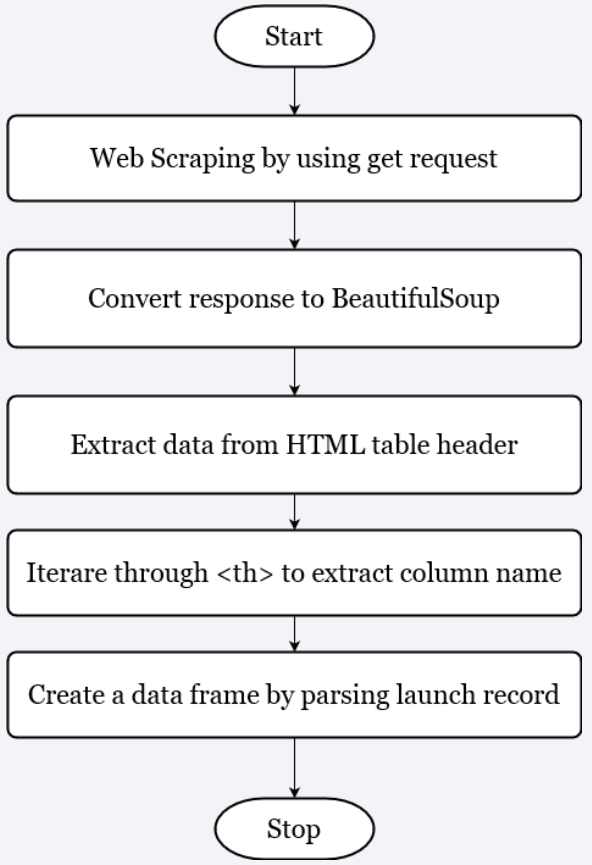
- Notebook link: Data Collection using SpaceX API



Data Collection - Scraping

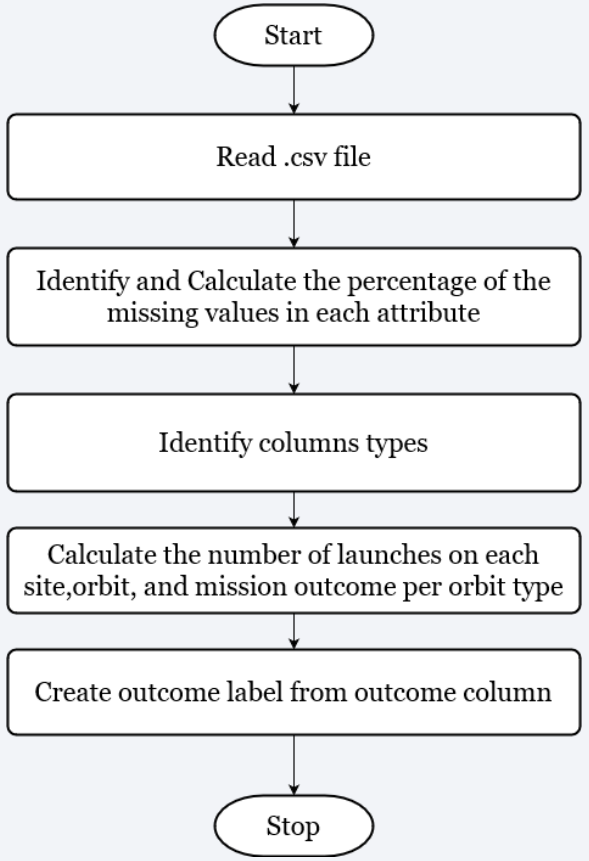
- This flowchart demonstrated how to collect data by scraping on the Falcon9 Launch Wiki page

- Notebook link: Data Collection by Scraping on Wiki



Data Wrangling

- The process:
  - Perform exploratory data analysis and determined the percentage of missing values in each value
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create outcome label
- Notebook link: Data Wrangling



EDA with Data Visualization

- Visualize the relationship between flight number and launch Site, payload and launch site, flight number and orbit type, the payload and orbit type using scatter plot because we want to know if these two feature have a significant correlation to success rate
- Visualize the relationship between success rate of each orbit type using bar plot because we want to know which orbits have high success rate
- Visualize the launch success yearly trend using line plot because we want to know the trend of success rate
- Notebook link: EDA with Data Visualization

EDA with SQL

- Connect to IBM DB2 using Python and wrote queries to find these insight
  - Unique launch site in space mission
  - Launch sites begin with "CCA"
  - Total payload mass by boosters launched by NASA (CRS)
  - 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
  - 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 the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Notebook link: EDA with SQL

Build an Interactive Map with Folium

- Created marker to point out the launch sites
- Add circle to launch sites to see the proximity area in 1 km.
- Add marker cluster to each launch sites to mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities then add the line and the label to show the distance
- Notebook link: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Plot the pie chart to demonstrate the percentage of success launch for all sites and see the percentage between failure and success landing in specific launch site
- Plot the scatter plot to demonstrate the Correlation between Payload and Success of all sites and also for the individual launch site by adjusted payload range on slider
- Python file link: Plotly Dash

Predictive Analysis (Classification)

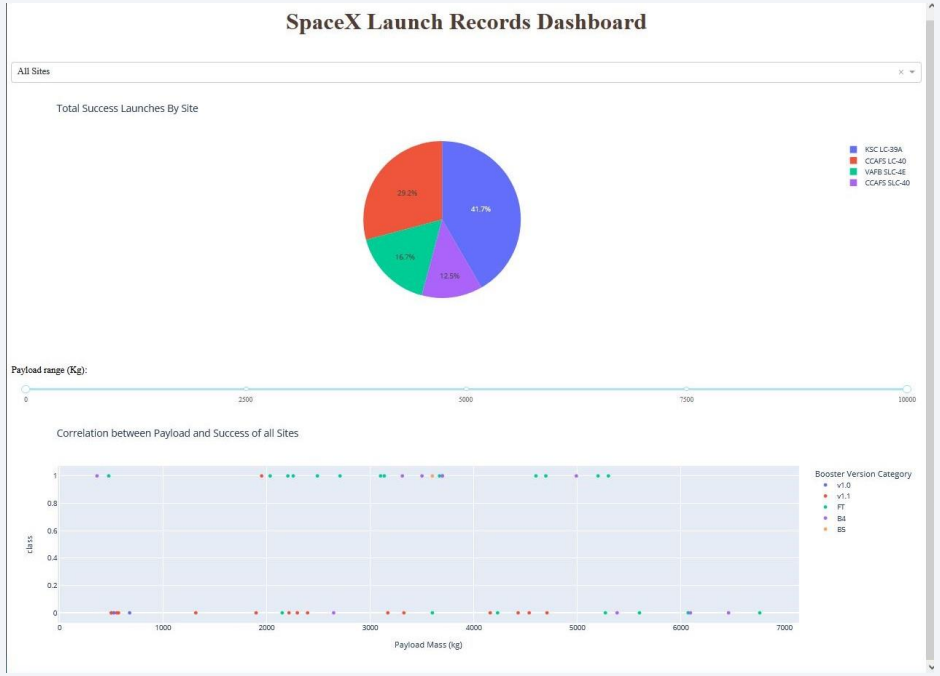
- Transform data and Split train and test set to validation data and training data
- Build Logistic Regression, Support Vector Machine, Decision tree and KNN models and using Grid Search CV to find the best parameters for each model
- Use Accuracy score to metric the model for both validation data and training data
- Found the best model
- Notebook link: Predictive Analysis

Results

- Exploratory data analysis results
  - SpaceX has an increasing of success rate in launching rocket since 2013 to 2020
  - The most massive of payload mass, the less likely the first stage will return
  - Heavy payload had more success rate on landing in Polar orbit type
  - CCAFS SLC 40 has launched most rocket
  - The best way to recover the part is by using drone ship

Results(con.)

- Interactive analytics demo in screenshots



- Predictive analysis results

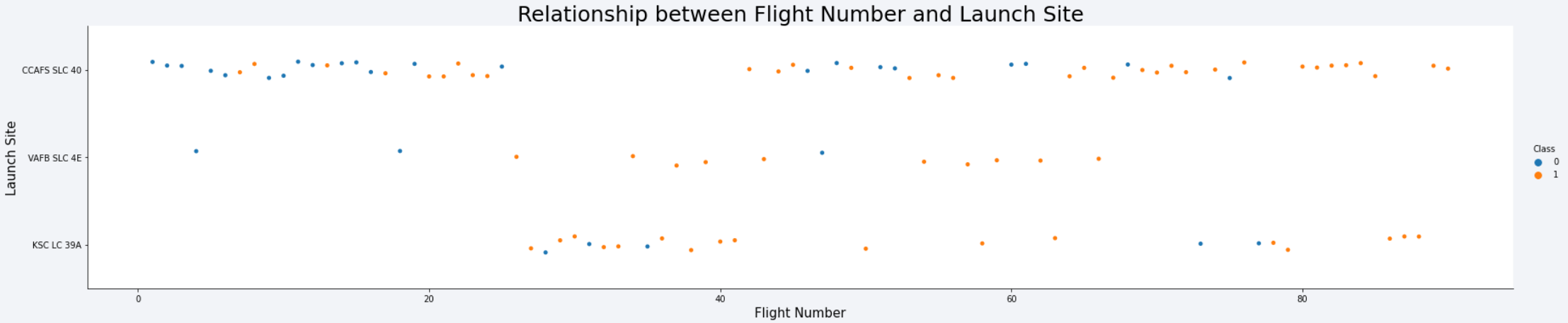
  Best models have prediction score which is 83.33% and they are logistic regression, SVM, and KNN

Section 2

Insights drawn from EDA

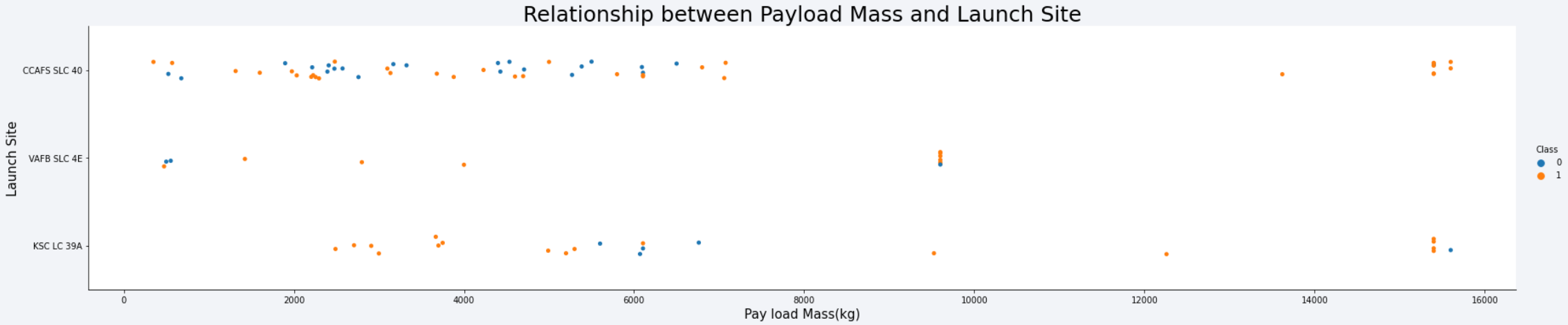
Flight Number vs. Launch Site

- Successful rate has grown over Flight Number and **CCAFS SLC 40** is the popular launch site and **KSC LC 39A** have the most percentage of success rate



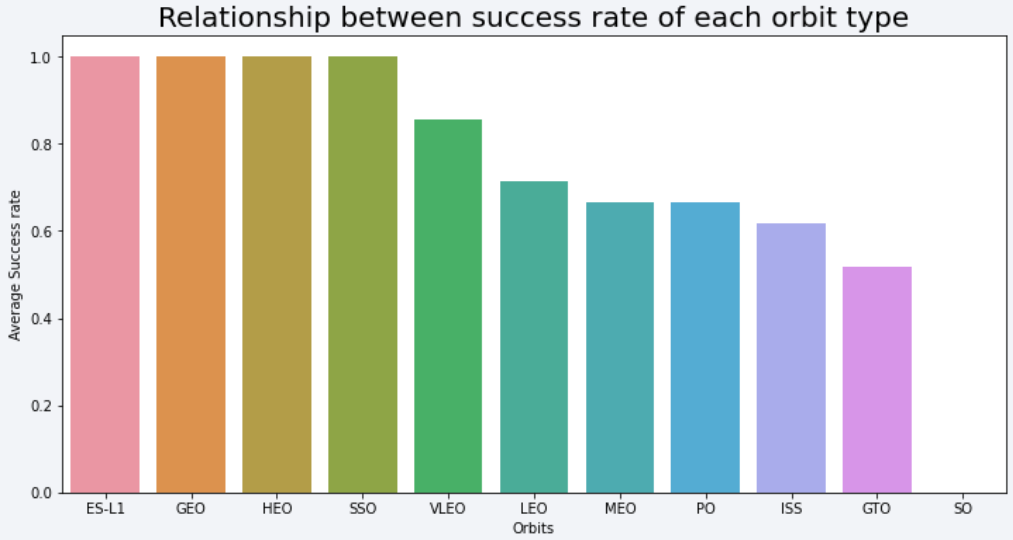
Payload vs. Launch Site

- Small Payloads (<6000kg) seem to has more success rate



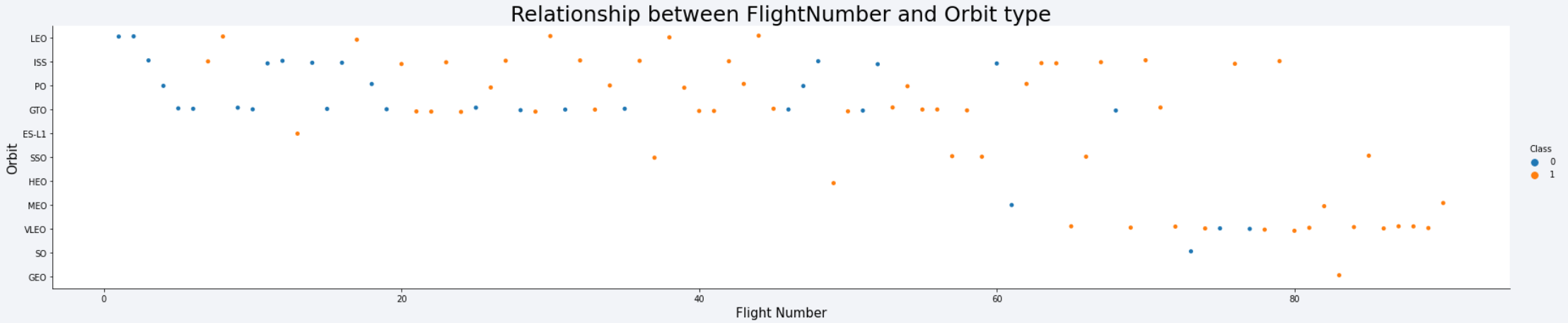
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO , and SSO orbits has the most success rate



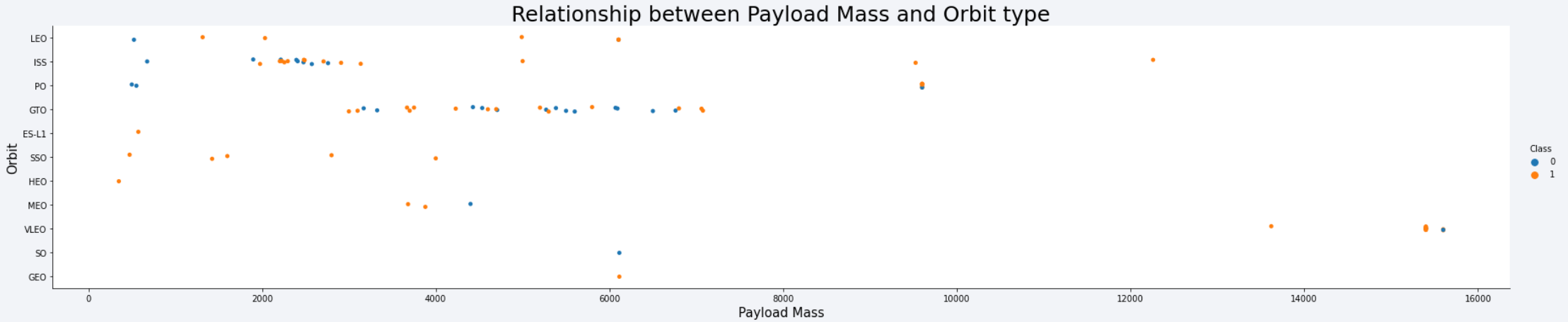
Flight Number vs. Orbit Type

- ES-L1, GEO, HEO have launch only once so the reliable can be skew
- LEO, ISS, SSO , and VLEO seem to have high success rate



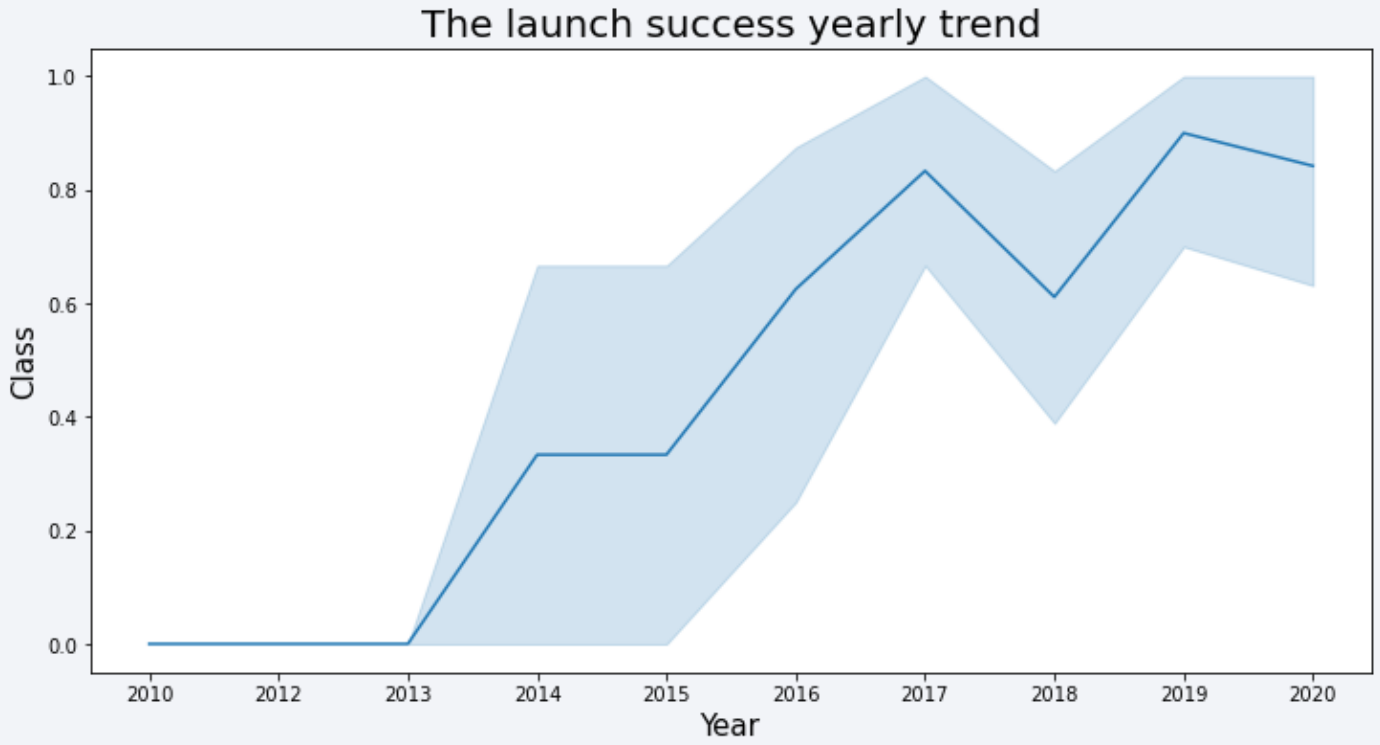
Payload vs. Orbit Type

- SSO have more success rate maybe because of having light weight



Launch Success Yearly Trend

- The success rate is increasing in between 2013 to 2017



All Launch Site Names

- There are 4 launch sites

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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

Total Payload Mass

- the total payload mass carried by boosters launched by NASA (CRS)

booster_version	total_payload_mass
F9 B4 B1039.2	2647
F9 B4 B1039.1	3310
F9 B4 B1045.2	2697
F9 B5 B1056.2	2268
F9 B5 B1058.4	2972
F9 B5 B1059.2	1977
F9 B5B1050	2500
F9 B5B1056.1	2495
F9 FT B1035.2	2205
F9 FT B1021.1	3136
F9 FT B1025.1	2257
F9 FT B1031.1	2490
F9 FT B1035.1	2708
F9 v1.0 B0006	500
F9 v1.0 B0007	677
F9 v1.1	2296
F9 v1.1 B1010	2216
F9 v1.1 B1012	2395
F9 v1.1 B1015	1898
F9 v1.1 B1018	1952

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 which is 2928 kg

booster_version	total_payload_mass
F9 v1.1	2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 22 July 2018

DATE	landing__outcome
2018-07-22	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 versions of booster that success on drone ship recovery and has a mass between 4,000 and 6,000

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- There are 100 of success mission and 1 of failed mission

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

Boosters Carried Maximum Payload

- List of the names of the booster which have carried the maximum payload mass which is 15,600 kg

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- There are 2 rockets have the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- Rank of 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_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

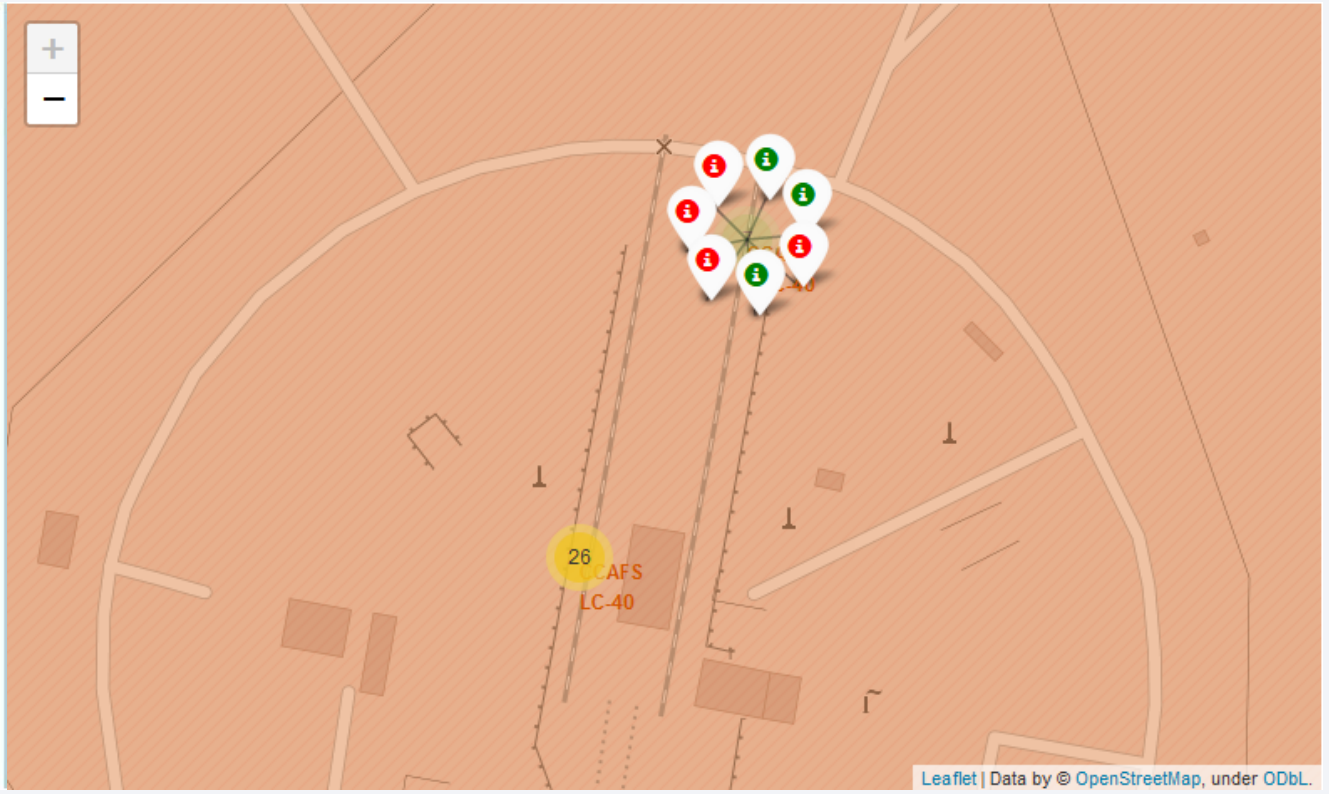

Section 3
Launch Sites
Proximities Analysis

Launch Sites Locations

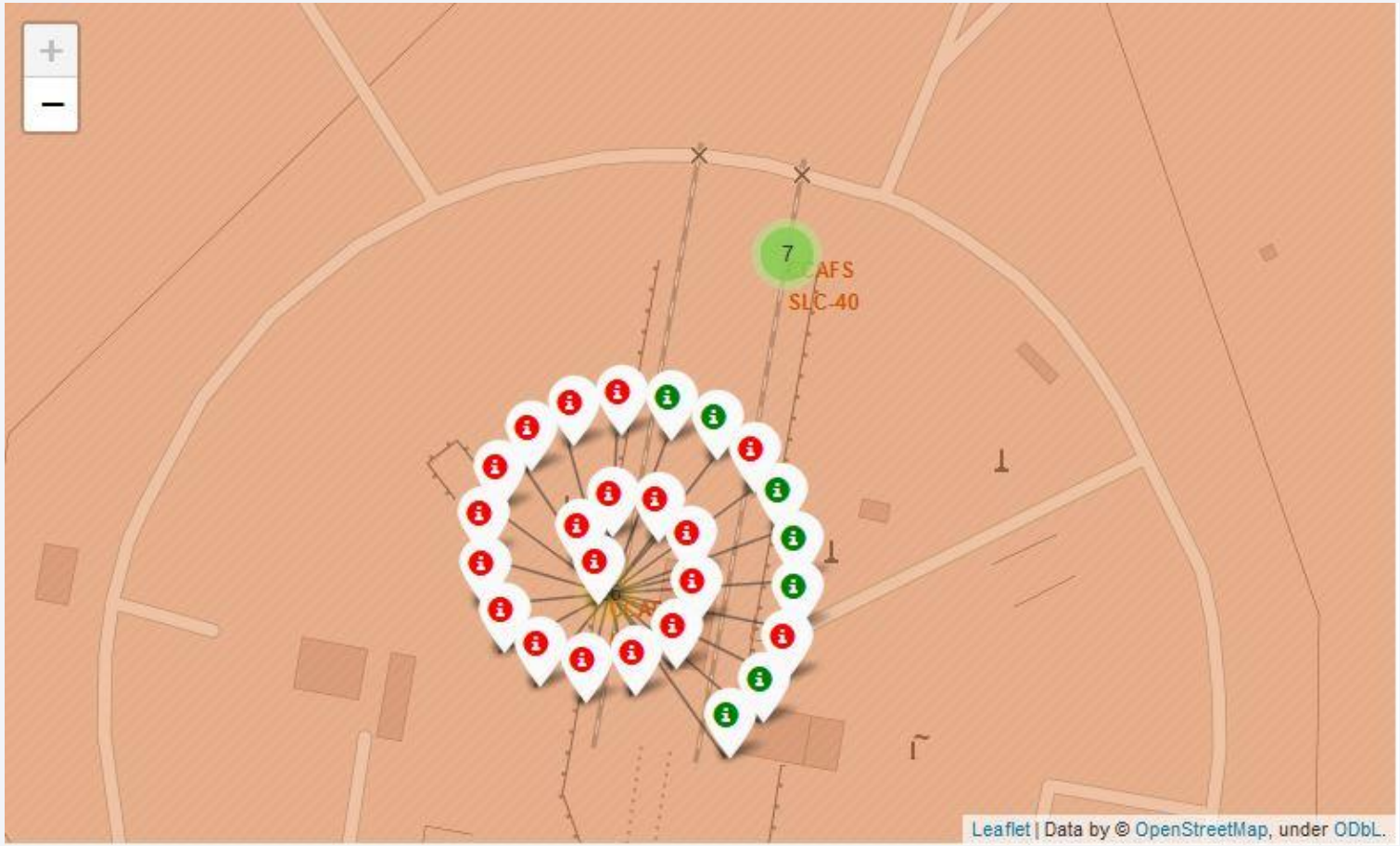
- SpaceX has 1 launch site in California, and the others 3 launch sites in Florida

CCAFS SLC-40 Recovery Result

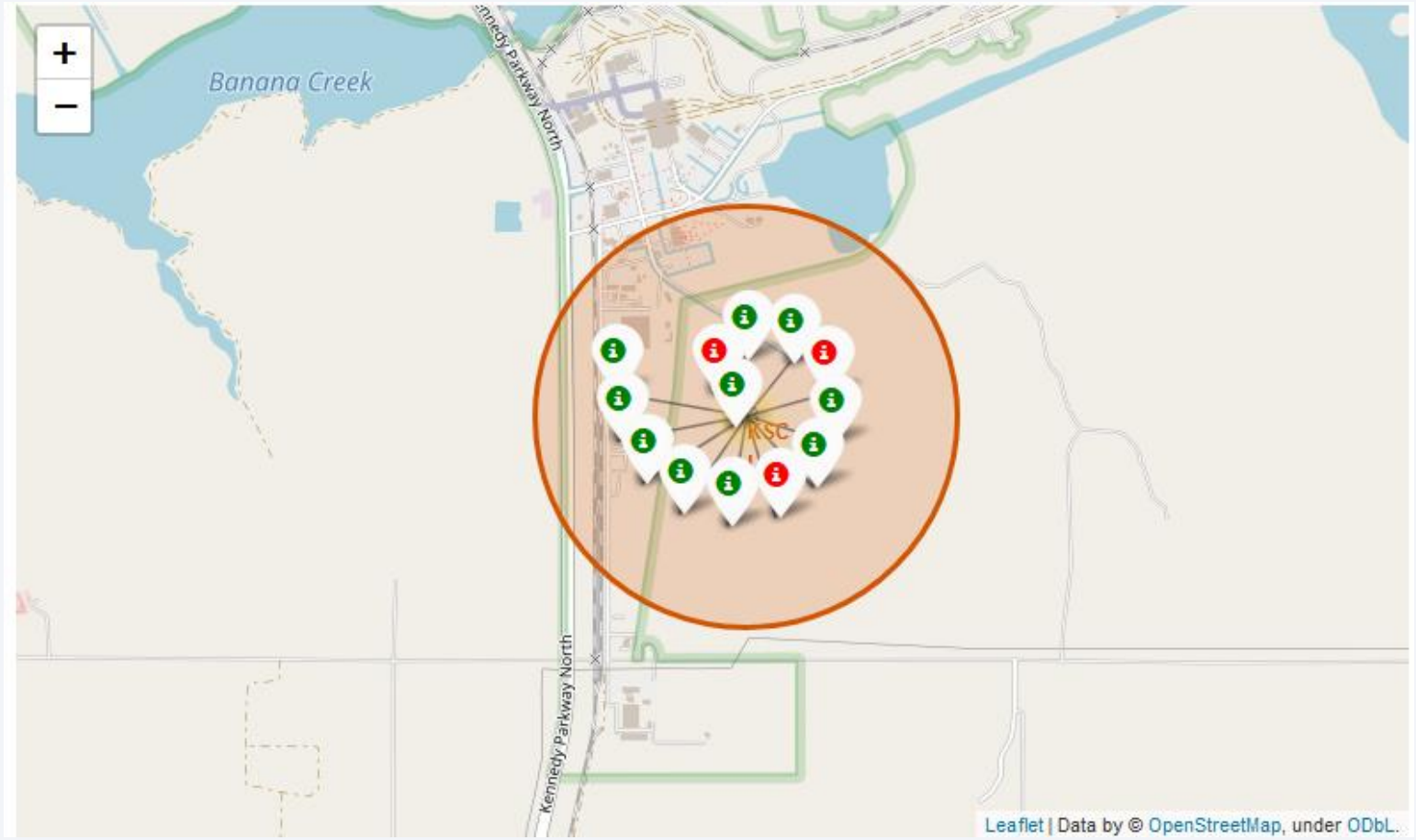
- Green marker represented successfully recovery of Falcon9 stage one
- Red marker represented successfully recovery of Falcon9 stage one
- Marker with 26 and has yellow color represented the number of all recovery that happen in the site



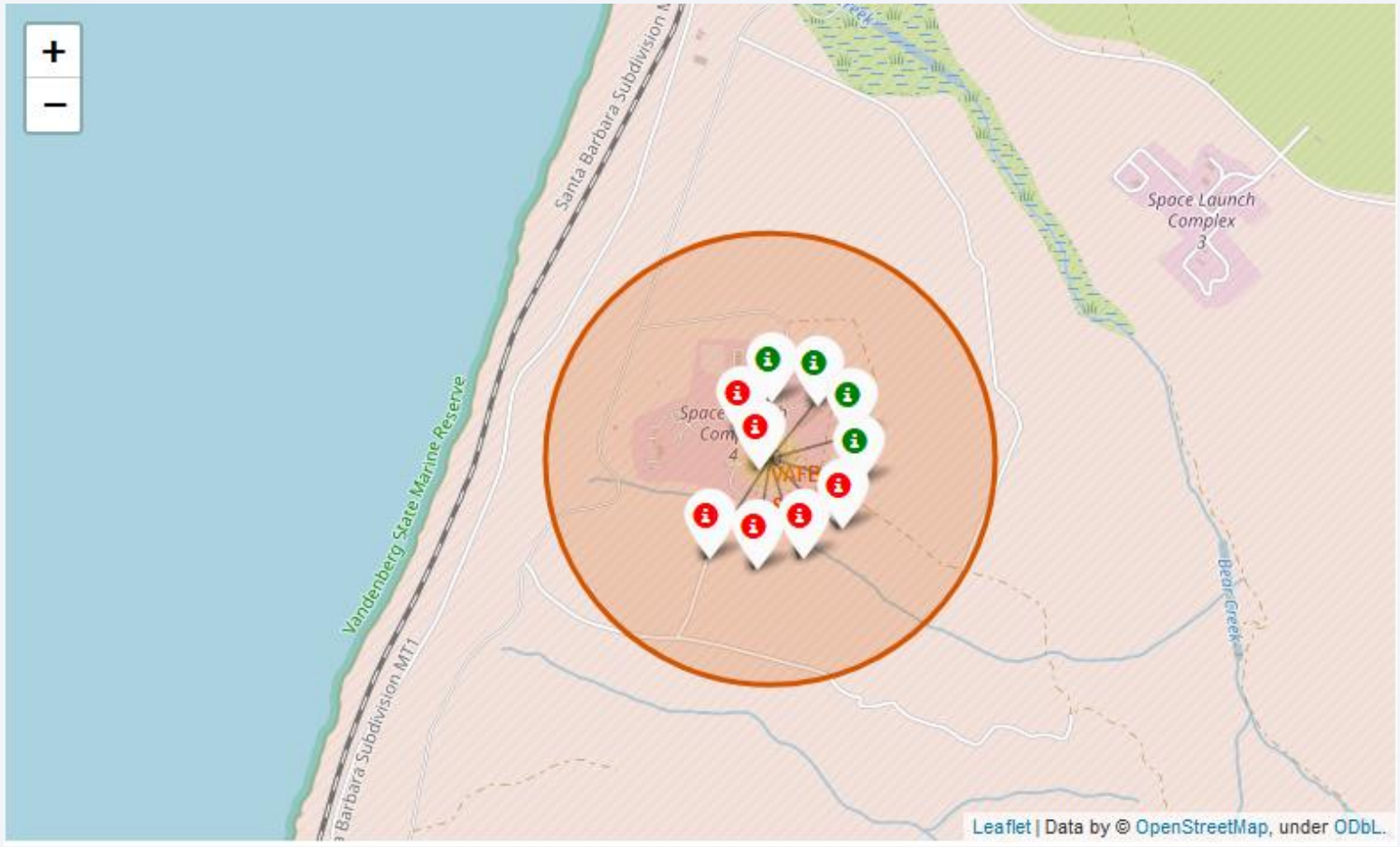
CCAFS LC-40 Recovery Result



KSC LC-39A Recovery Result

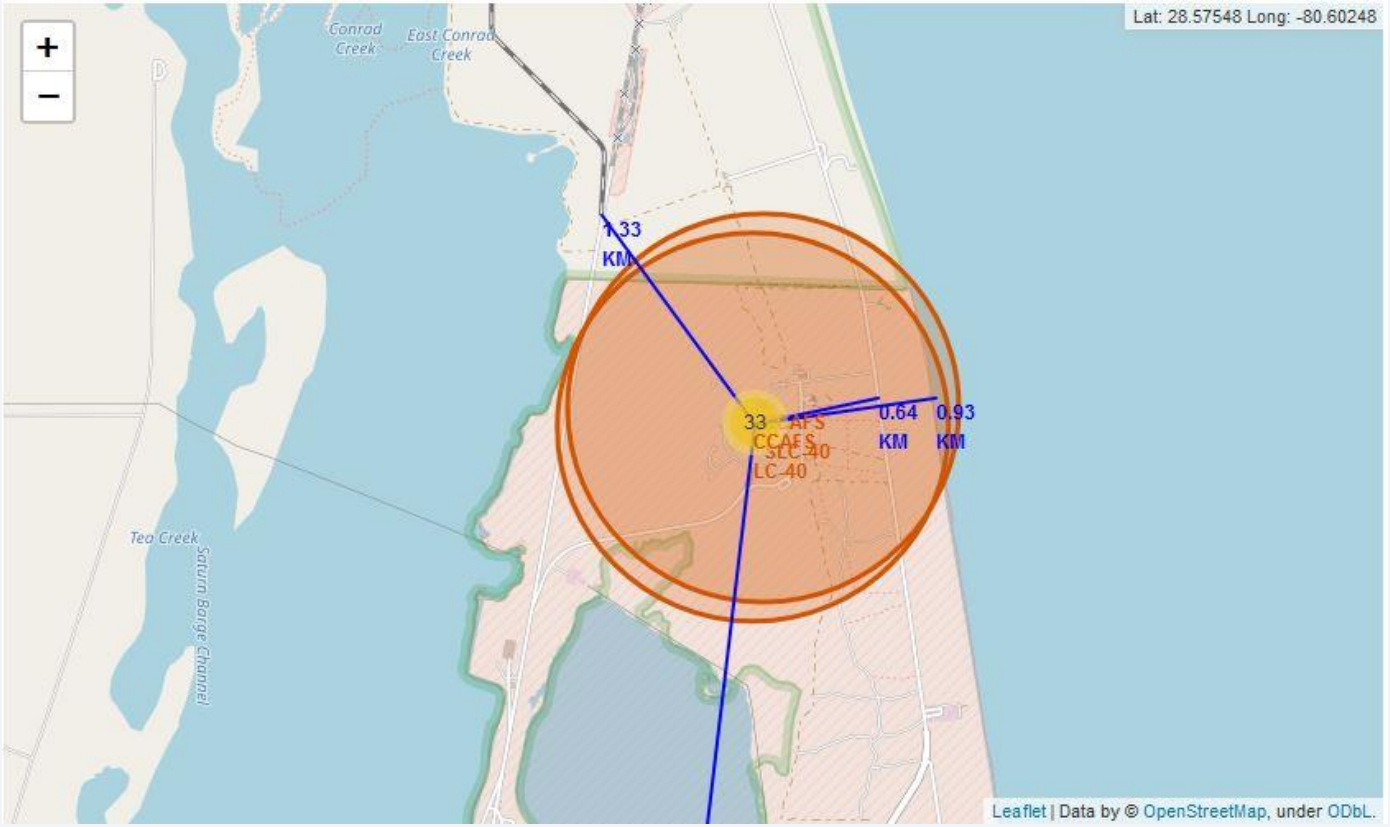


VAFB SLC-4E Recovery Result



Nearby CCAFS SLC-40 & CCAFS LC-40

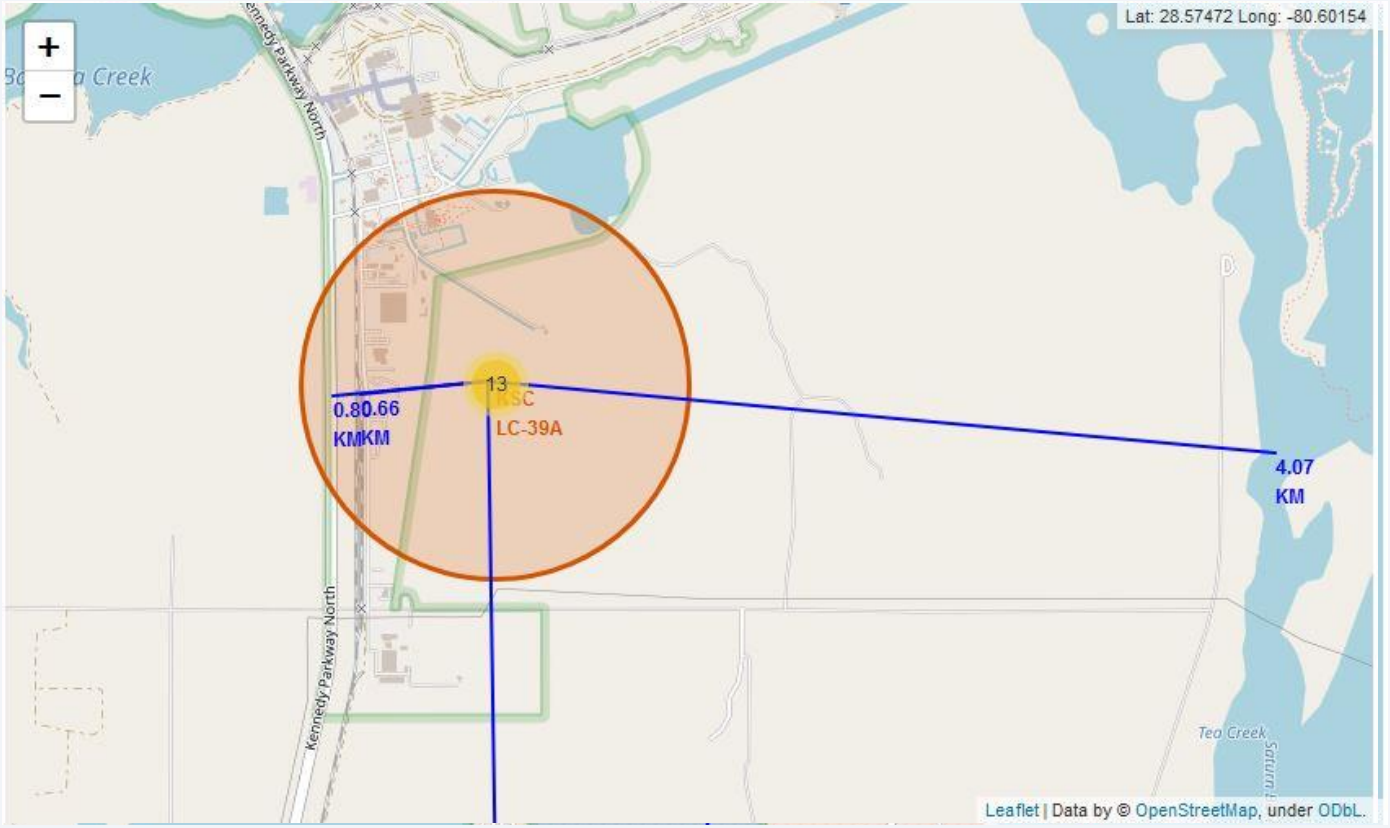
- The line represented the distance to nearest coastline, city, railway and highway
- These two sites are close to (<5km) railway, highway and coastline



Nearby KSC LC-39A

- The line represented the distance to nearest coastline, city, railway and highway

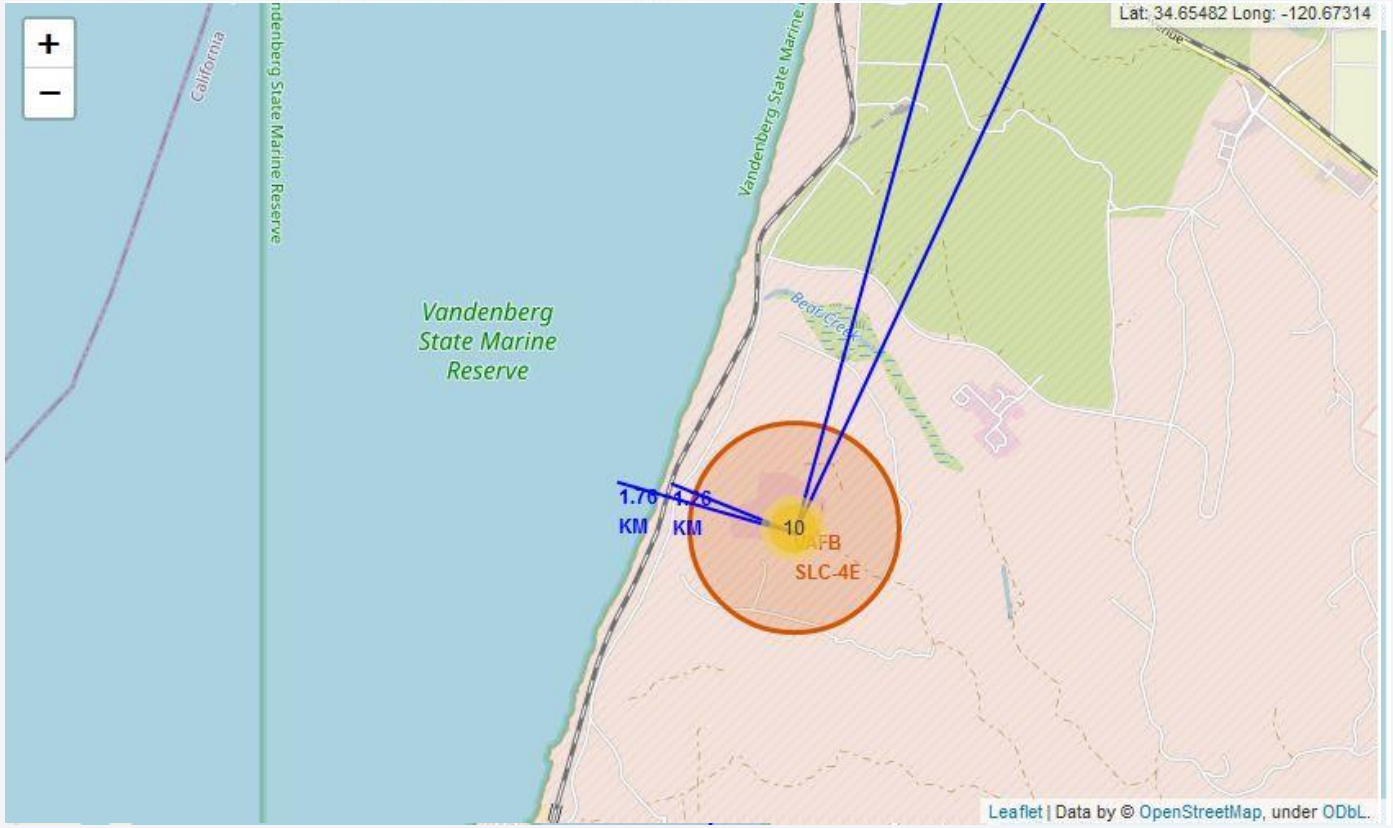
- This site is close to (<5km) railway, highway and coastline

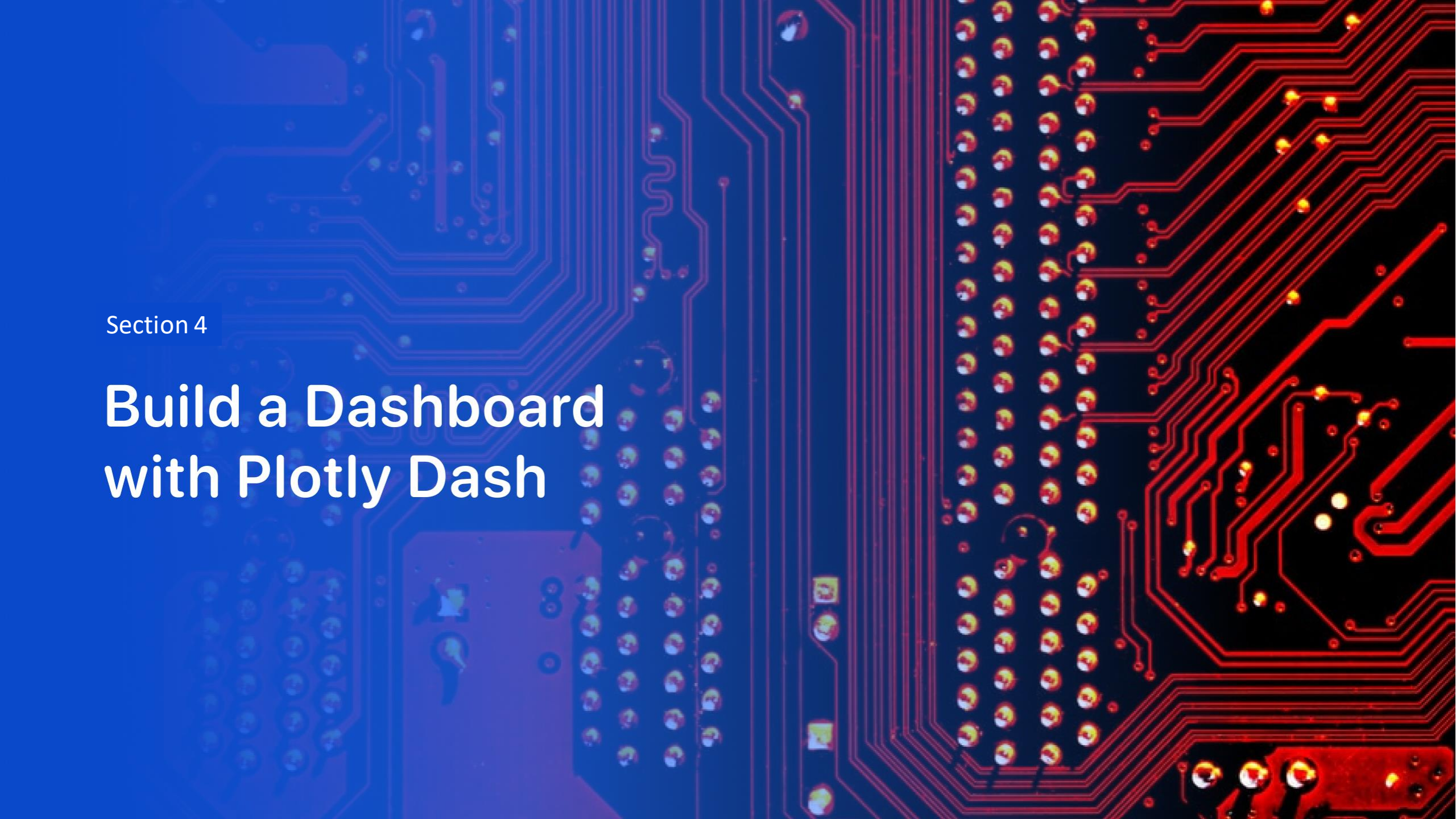


Nearby KSC LC-39A

- The line represented the distance to nearest coastline, city, railway and highway

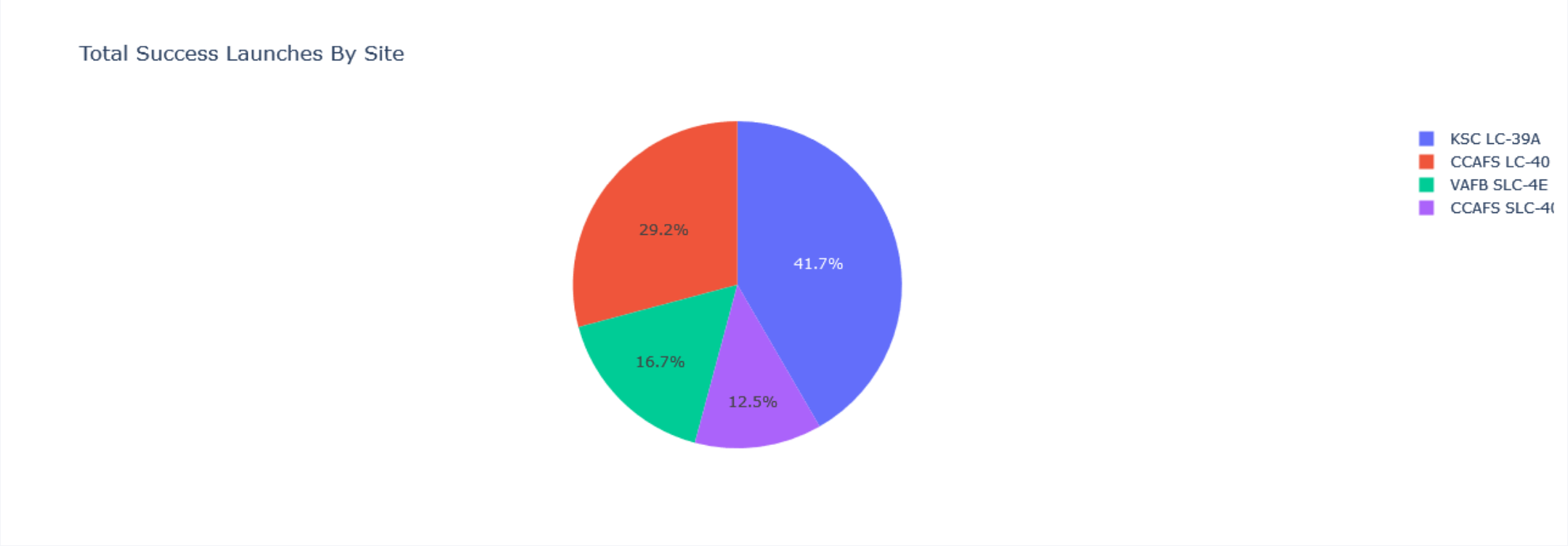
- This site is close to (<5km) railway and coastline




Section 4
Build a Dashboard with Plotly Dash

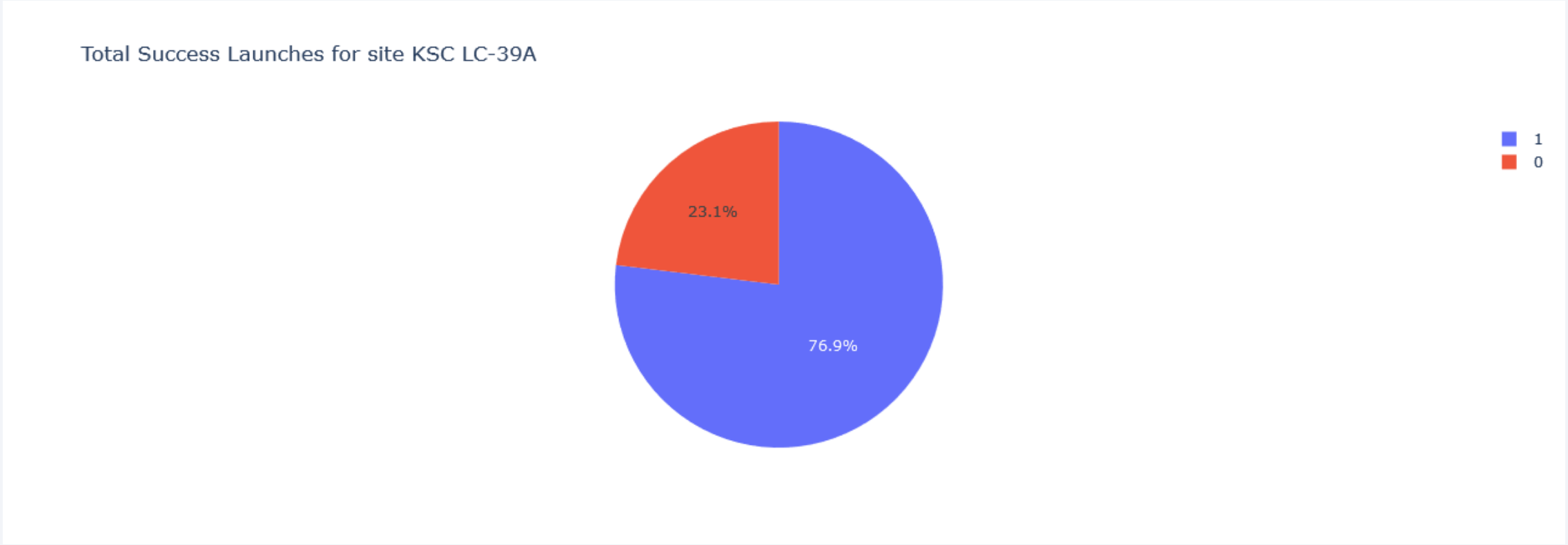
Pie Chart of Success Rate in all Sites

- KSC LC-39A has the most success rate among the others



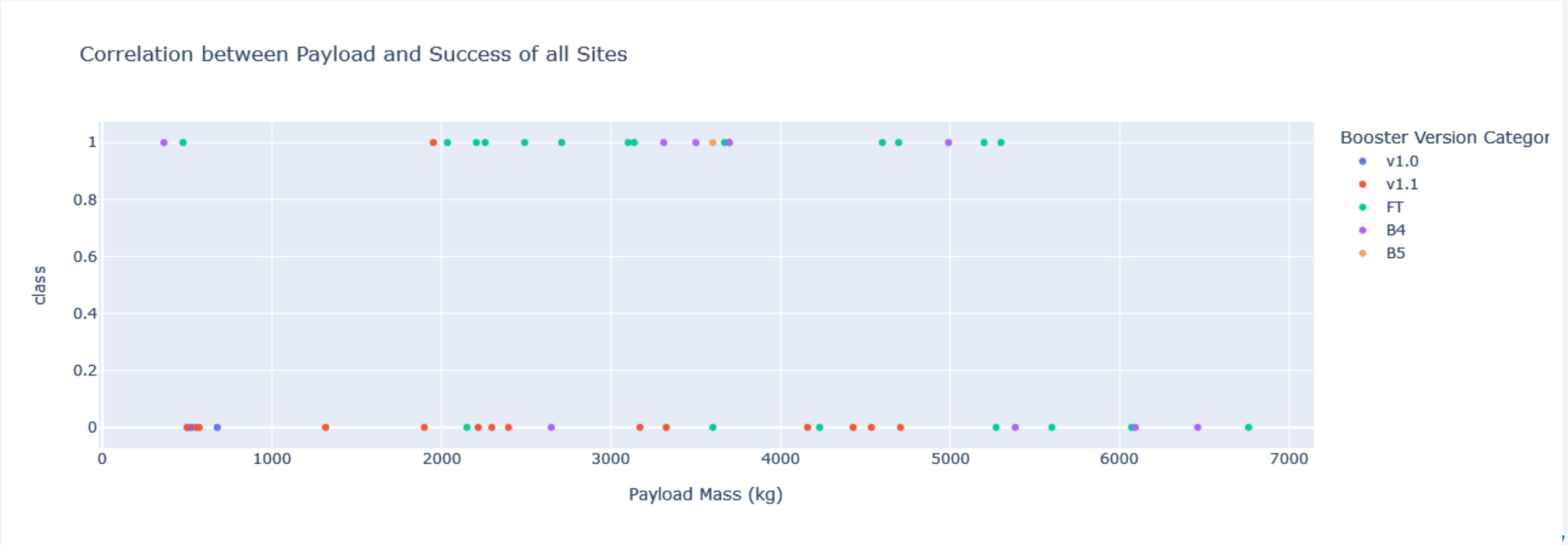
Most Successful Launch Site by Proportion

- KSC LC-39A has the most success rate which is 76.9%



Recovery result vs Payload Mass in each Booster version

- FT is the most booster version that success
- The success recovery has a payload mass between 2,000 and 4,000 kg

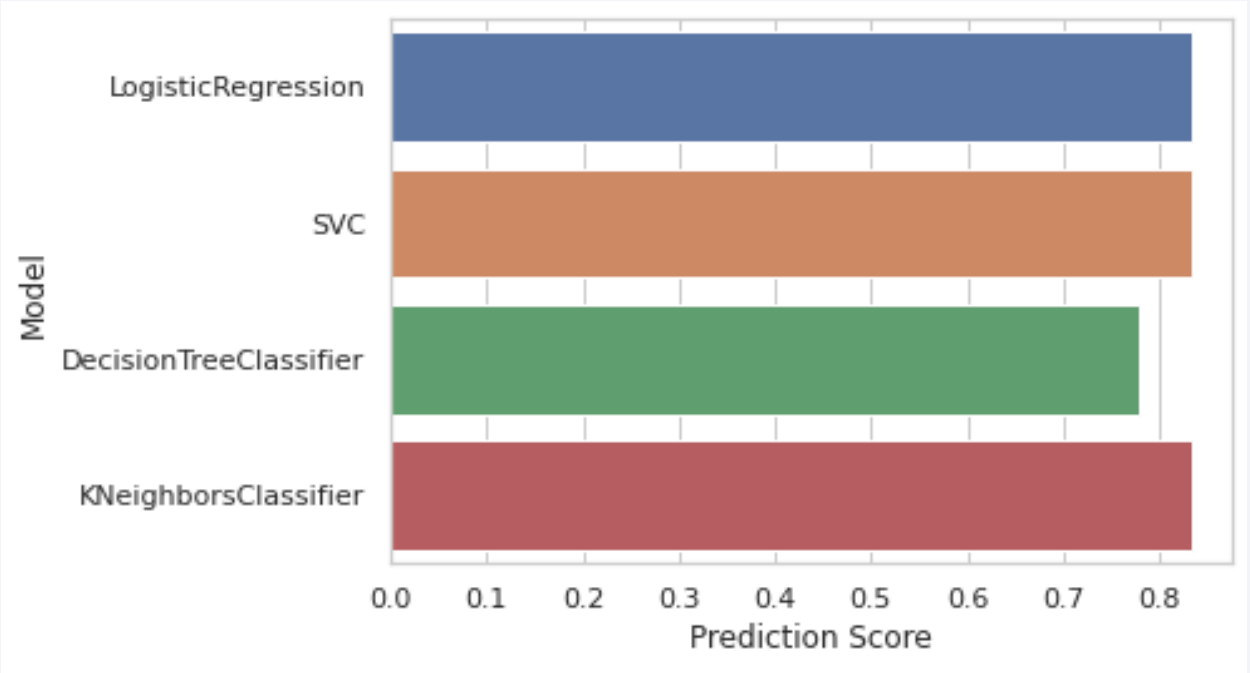


Section 5

Predictive Analysis (Classification)

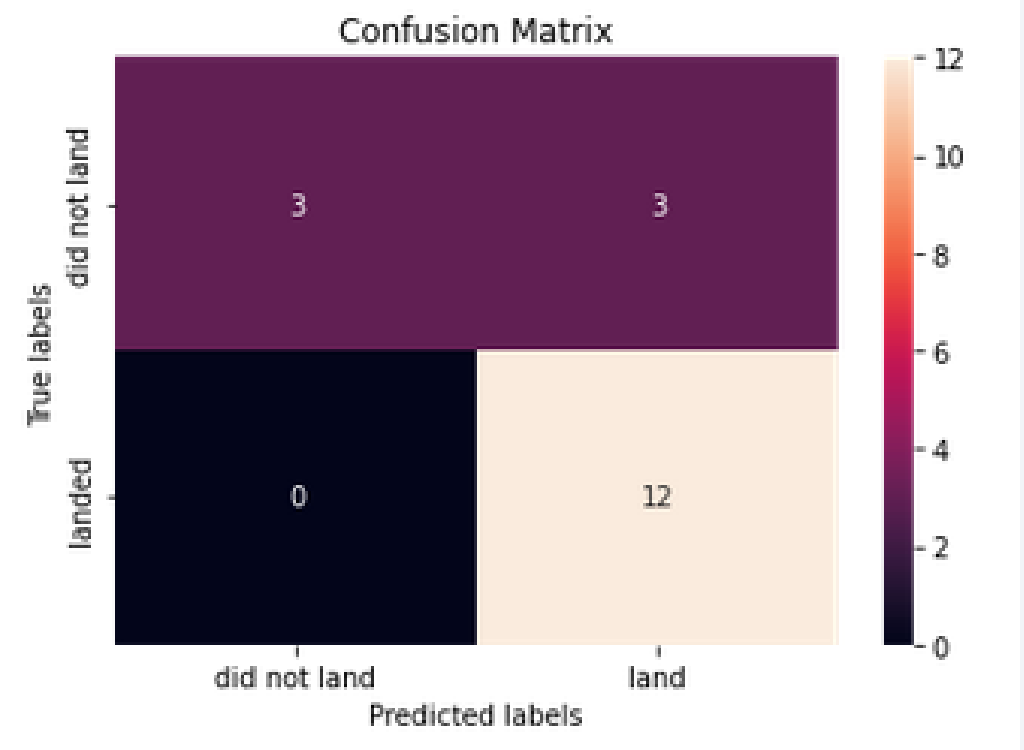
Classification Accuracy

- The bar plot show the accuracy of model. As can be seen, most model have equal accuracy score which is logistic regression, SVM, and KNN have a score 83.33% but for Decision tree is 77.8% which is the lowest



Confusion Matrix

- The models perform the same result which have wrong prediction in 3 landings that predict the they were land but actually didn't land



Conclusions

- SpaceX's recovery is more likely to be success when:
  - Launch in after year 2016
  - Has a light weight (2000-4000kg)
  - Launch from KSC LC-39A

- The best models have the accuracy score 83.33% at predicting the landing

Thank you!