IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akhilesh Vyas
10-May-2023

Outline

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

Executive Summary

- In the process of creating predictive models, SpaceX data is collected using python packages from the REST API (https://api.spacexdata.com/v4/) and transformed into pandas data frame. Pandas library is used to wrangle data, filling missing values and determining class labels from landing outcome feature. Exploratory data analysis (EDA) is performed to know relation between different data features using visualization plot such as scatter, bar, line etc. Moreover, SQL EDA helps us to explore more about data such distinct launch sites, maximum payload booster version, total success & failure launch mission. In the data analysis, it is found that first stage launch has more success rate after 2013 onwards. Further, ES-L1 , GEO , HEO and SSO orbit type have highest success rate and GTO has the lowest success. Interactive visual analytics is used to explore more about relation between payload mass and landing outcome. Heavy payload has low success rate in landing. Folium Map helps to determine different launch sites in the map and also to visualize success & failure rate. Predictive models (SVM, LR, Tree, KNN) are trained using GridSearchCV to predict success & failure of first stage launch using data features. Decision Tree model performs best to determine landing outcome as success and failure.

Introduction

- Satellites are launched using rockets where many stages are involved before delivery of the satellite to the final orbit. In this case, the first stage is very important and expensive. SpaceX company has a very low cost of launch compared to its competitors as it can reuse the rocket's first stage. The project uses predictive models to determine if the first stage of SpaceX will be successful or not. In addition, EDA methods is used to answer what is the relation between different features, and how these features impact success/failure landing outcome.

Section 1

Methodology

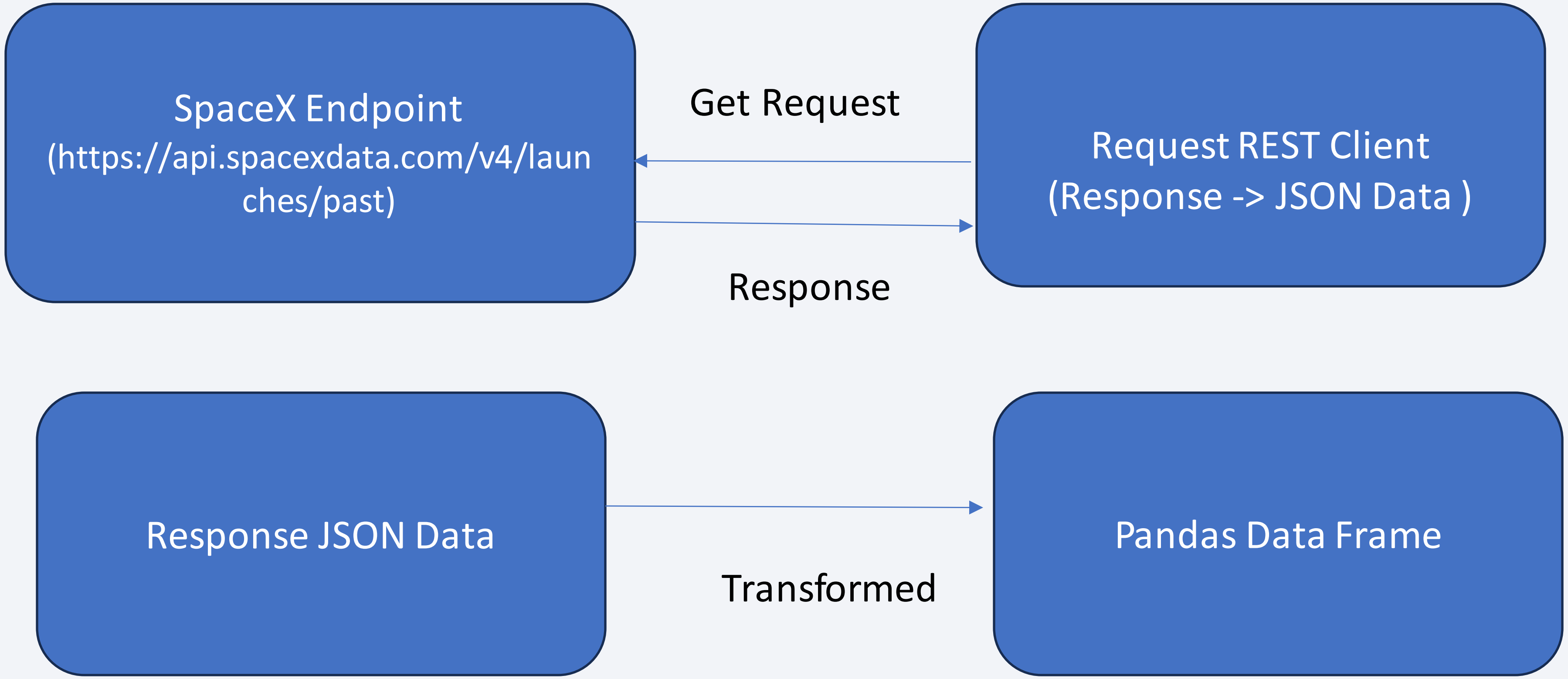
Methodology

Executive Summary

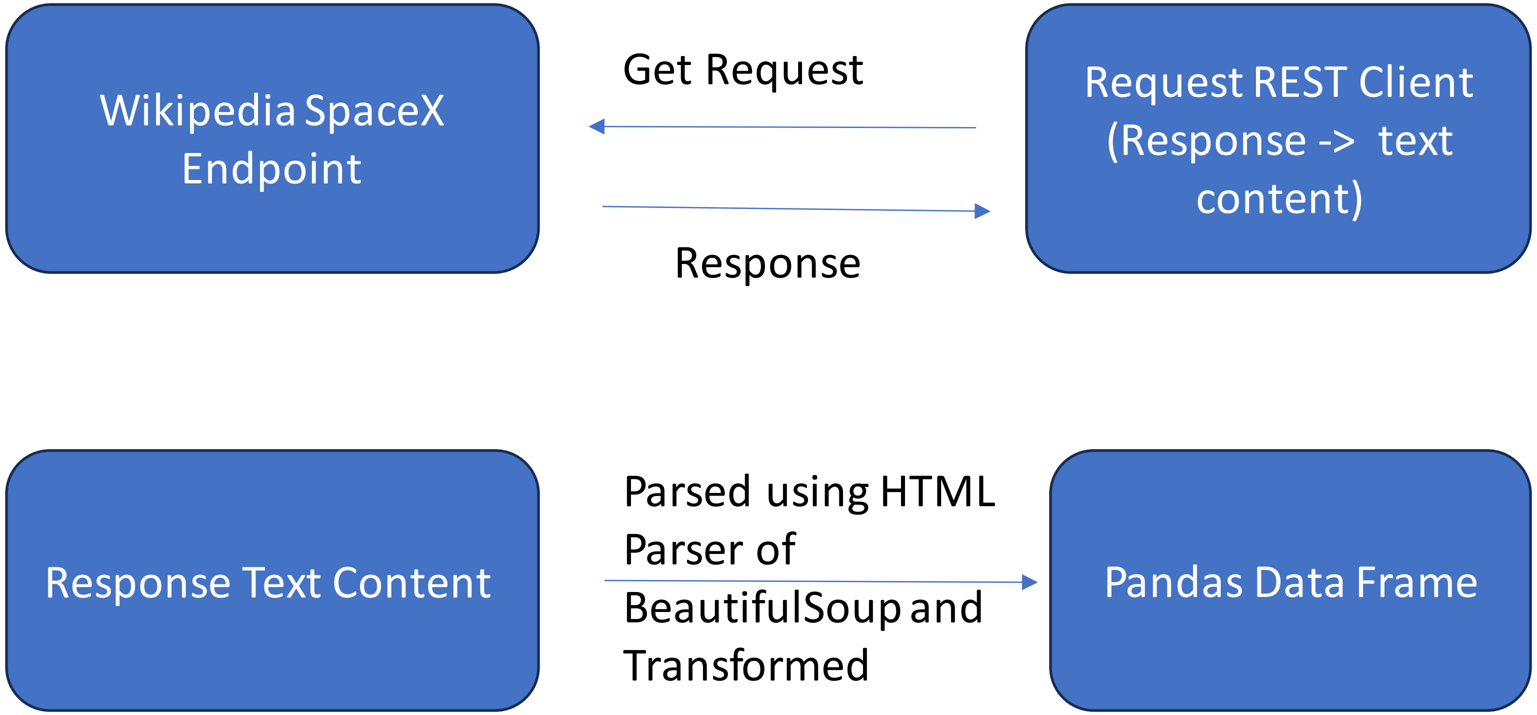
- Data collection methodology:
  - Python request package is used to collect data from SpaceX REST API (URL: https://api.spacexdata.com/v4/)
- Perform data wrangling
  - Python pandas is used to wrangle data and deal with null values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data is standardized and split into train and test data to train different classification models (Logistic Regression, SVM, Decision Tree and KNN) using Python Scikit-Sklearn library. Model are tuned using GridSearchCV and evaluated through confusion matrix.

Data Collection

From SpaceX Endpoint


SpaceX Endpoint (https://api.spacexdata.com/v4/launches/past)
Get Request
Request REST Client (Response -> JSON Data)
Response
Response JSON Data
Transformed
Pandas Data Frame

From Wikipedia URL Link



Wikipedia SpaceX Endpoint

Get Request

Response

Request REST Client (Response -> text content)

Response Text Content

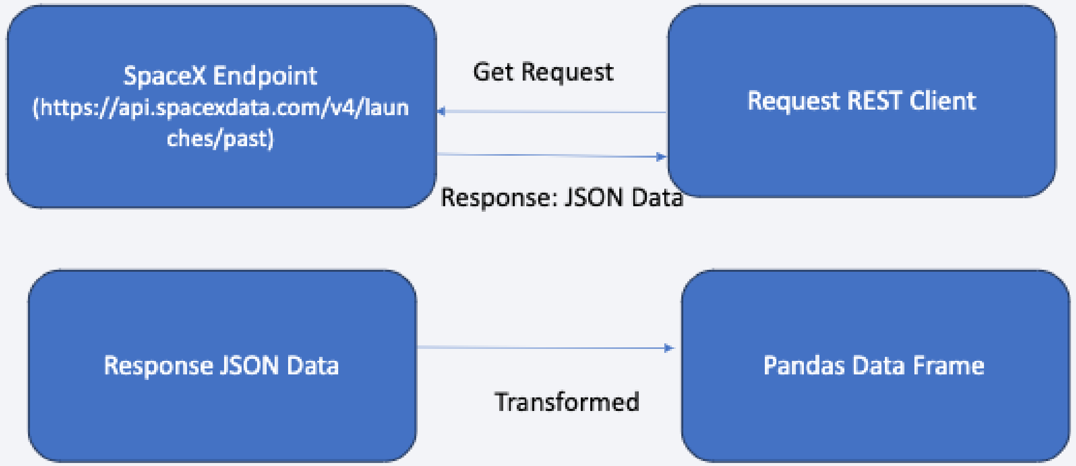
Parsed using HTML Parser of BeautifulSoup and Transformed

Pandas Data Frame

Data Collection – SpaceX API

- Data collection with SpaceX REST API
  - Collect data using Python library Request from SpaceX URL
  - Transform and convert JSON response data into pandas data frame

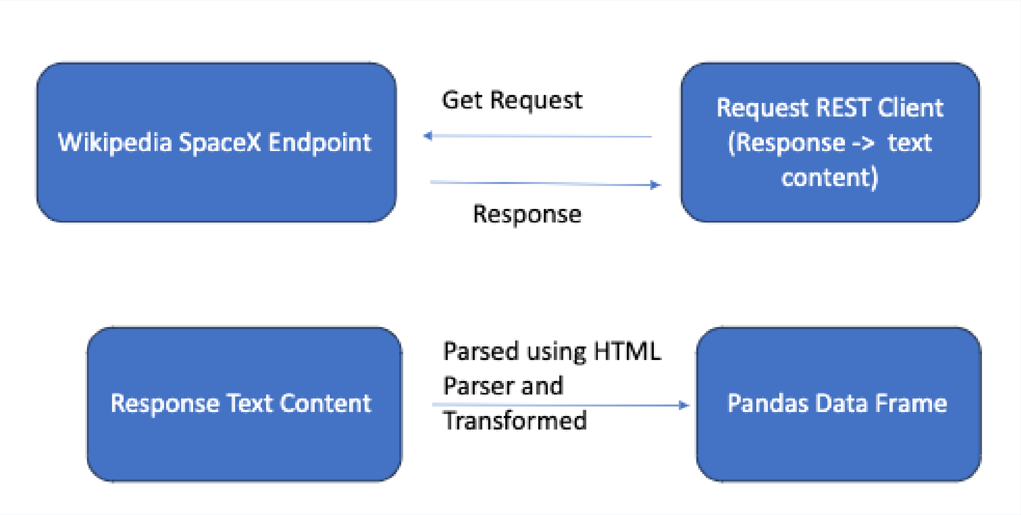
- GitHub URL



Data Collection - Scraping

- Data Collection from Wikipedia
    - Extract Falcon 9 launch record from HTML Table using Python Library BeautifulSoup
    - Parse Table and transform SpaceX table into Pandas data frame

- GitHub URL


Wikipedia SpaceX Endpoint
Get Request
Response
Request REST Client (Response -> text content)
Response Text Content
Parsed using HTML Parser and Transformed
Pandas Data Frame

Data Wrangling

- Methods
  - Pandas API is used to sample data and deal with null values
    - Replace PayloadMass feature null value using average value
  - GitHub URL

EDA with Data Visualization

The relationship between FlightNumber and Launch Site

- First stage launches success is increased with flight numbers from different launch sites. Launch site CCAFS LS-40 has low success rate in initial flights

The relationship between Payload and Launch Site

- VAFB-SLC has not launched any rockets having payload more than 10K. CCAFS-CLC 40 has more success rate for heavy payloads more than 12k

The relationship between success rate of each orbit type

- ES-L1 , GEO , HEO and SSO has orbit type have highest success rate and GTO has the lowest success rate.

The relationship between FlightNumber and Orbit type

- LEO orbit is having better success rate as flight number increases and GTO orbit has no relation with flight number in case of success rate

The relationship between Payload and Orbit type

- SSO has 100% success rate. LEO, Polar and ISS performs better with heavy payload.

EDA with Data Visualization

The relationship between Success rate and Years

- Initially, success rate is low till 2013 and then it increases with the years .

GitHub URL

EDA with SQL

- Select Unique Launch Sites
- Get first 5 Records having Launch site starting with string 'CCA'
- Total payload mass carried by customer NASA
- Average payload mass carried by booster version F9 v1.1
- Get first date of successful landing outcome in ground pad
- Get the names of the successful boosters having 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 having the month, 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.
- GitHub URL

Build an Interactive Map with Folium

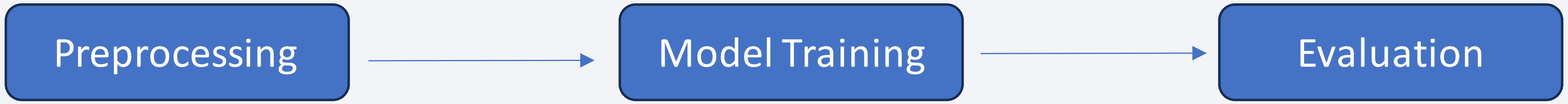
- Creation of map having launch sites with circle, marker, marker-cluster, mouse position and polyline
- Purpose of Map Objects
  - Map (To represent different launch sites in the map)
  - Circle (To draw for circle for launch sites coordinates)
  - Marker (To create leaflet marker with text for launch sites)
  - Marker-Cluster (To create markers for success and failure of launch sites)
  - Mouse-Position (To show mouse position coordinates in the map)
  - Polyline (To draw polyline overlays on map to show distance between coast line and launch site)
- GitHub URL - Note: If GitHub is not able to render map then use check pdf here.

Build a Dashboard with Plotly Dash

- Dashboard Items:
  - Title (To show title of the dashboard)
  - Dropdown Menu (To select different launch site)
  - Pie Chart (To show total success launches by site)
  - Payload Slider (To select payload range)
  - Scatter Plot (To show correlation between payload mass and success for site)
- GitHub URL

Predictive Analysis (Classification)

- Exploratory Data Analysis is performed and determined Training Labels for classification problem
- Class variable is used as target variable
- Data is standardized to create more generalized model
- Data is split into training data and test data
- GridSearchCV (CV=10) is utilized to find best Hyperparameter for KNN, SVM, Decision Trees and Logistic Regression models
- Confusion and accuracy matrix is used to select best mode
- GitHub URL



Preprocessing

Model Training

Evaluation

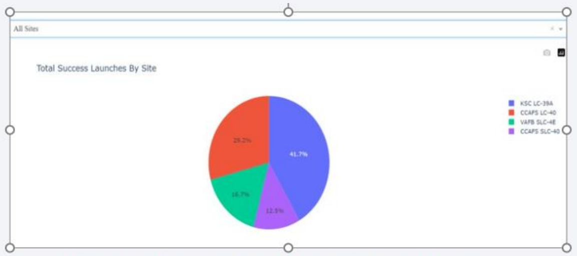
Standardized,
transformation ,
Label selection

GridSearchCV and
classification model
training

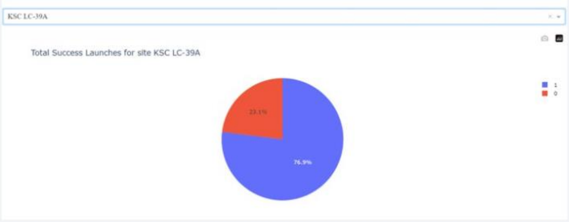
Accuracy and
Confusion Matrix

Results

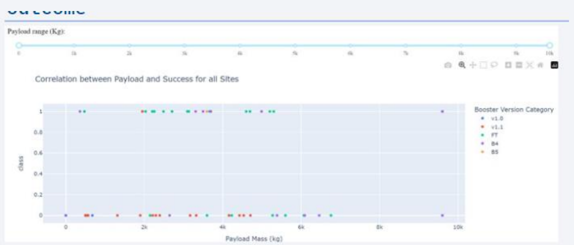
- Exploratory data analysis results
  - First stage launches success is increased with flight numbers from different launch sites
  - ES-L1 , GEO , HEO and SSO orbit types have highest success rate and GTO has the lowest success rate.
  - SSO has 100% success rate. LEO, Polar and ISS performs better with heavy payload.
  - Success rate trends positively after 2013.
- Interactive analytics demo screeshots



LSC LC-39A has maximum success launches ratio and CCAFS SLC has the lowest one

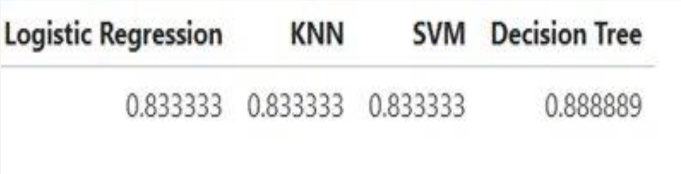


The Pie Chart shows that KSC LC-39A sites has more than 75% success. Class success=1 and failure=0

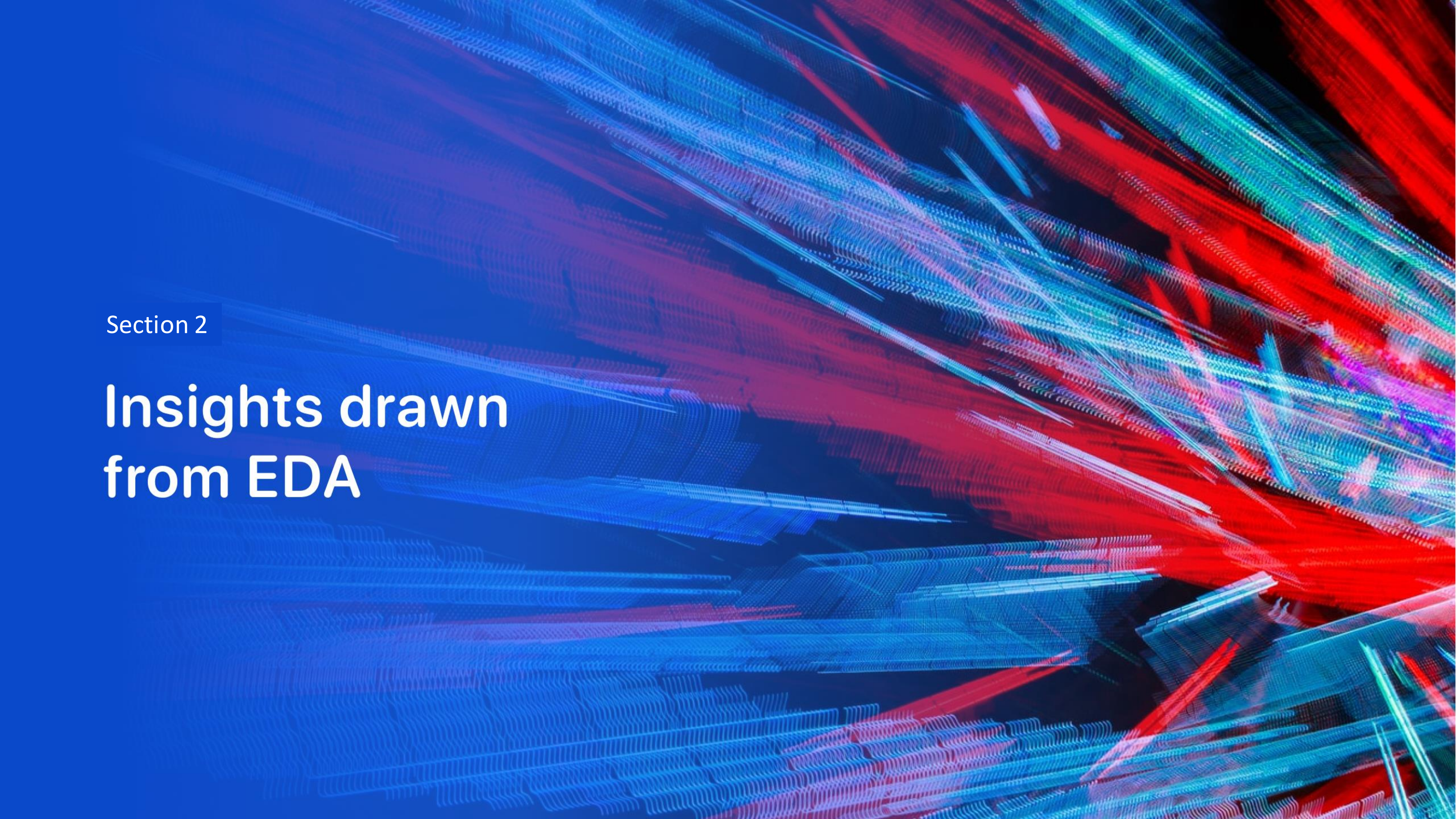


The Scatter plot shows that FT booster has very high success rate and payload mass more than 6k has very less success rate.

- Predictive analysis results –Decision tree model performs best to predictive first stage landing outcome.



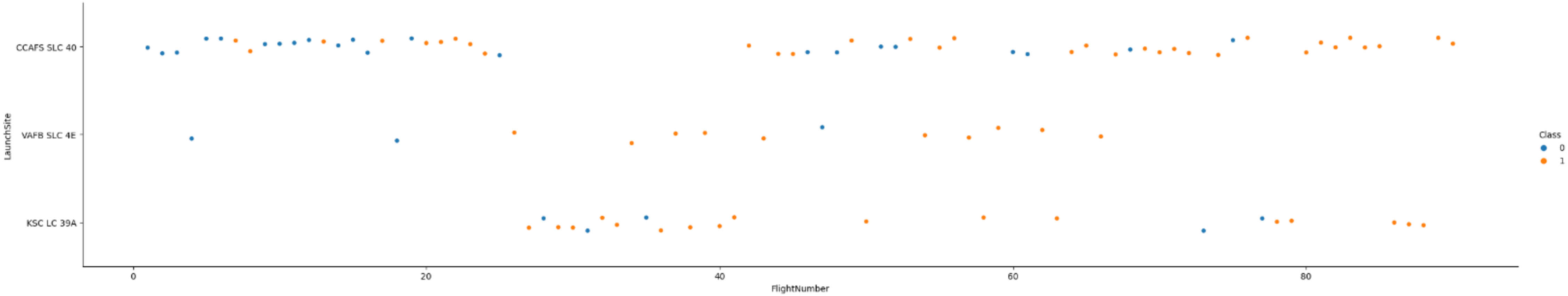
Logistic Regression	KNN	SVM	Decision Tree
0.833333	0.833333	0.833333	0.888889



Section 2

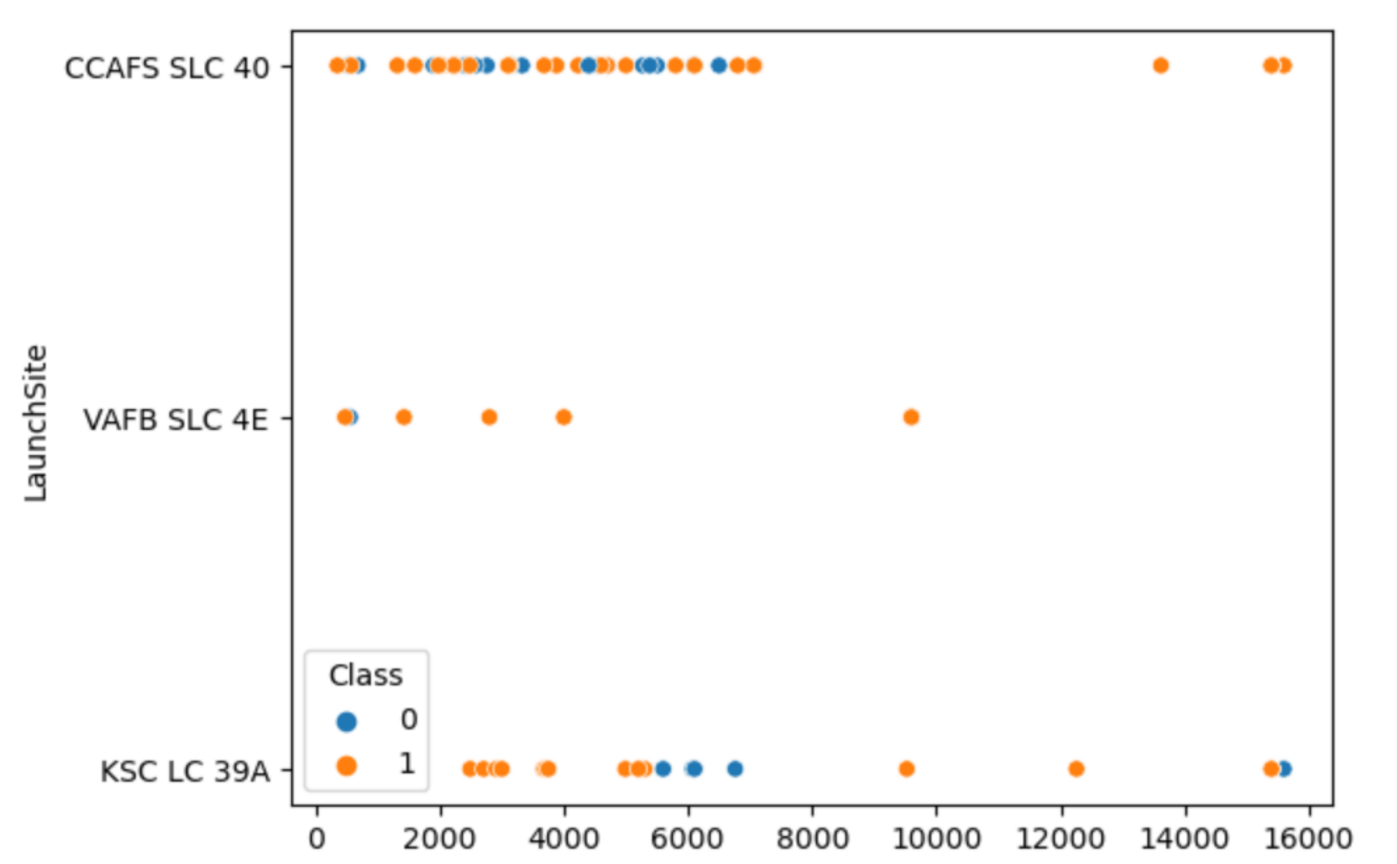
Insights drawn from EDA

Flight Number vs. Launch Site



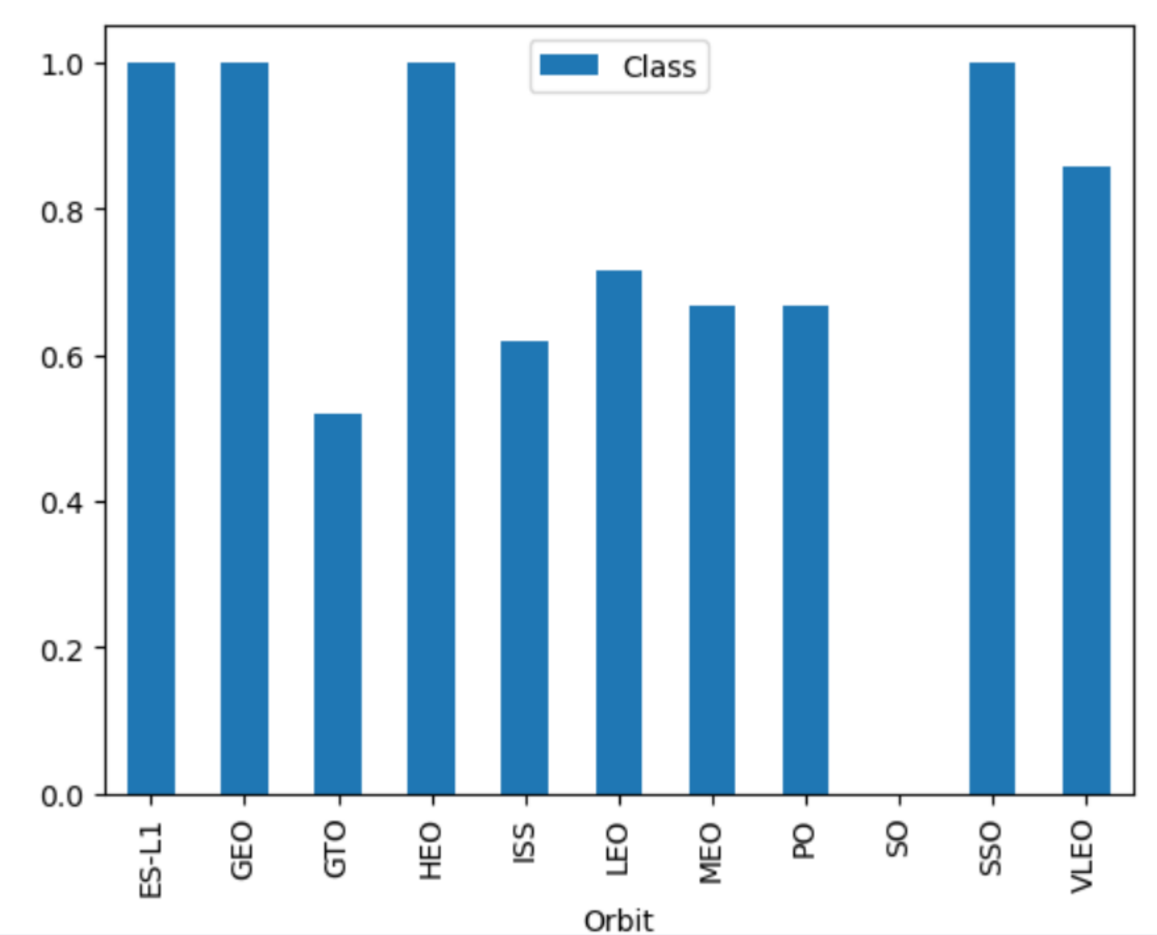
First stage launches success is increased with flight numbers from different launch sites. Launch site CCAFS LS-40 has low success rate in initial flights

Payload vs. Launch Site



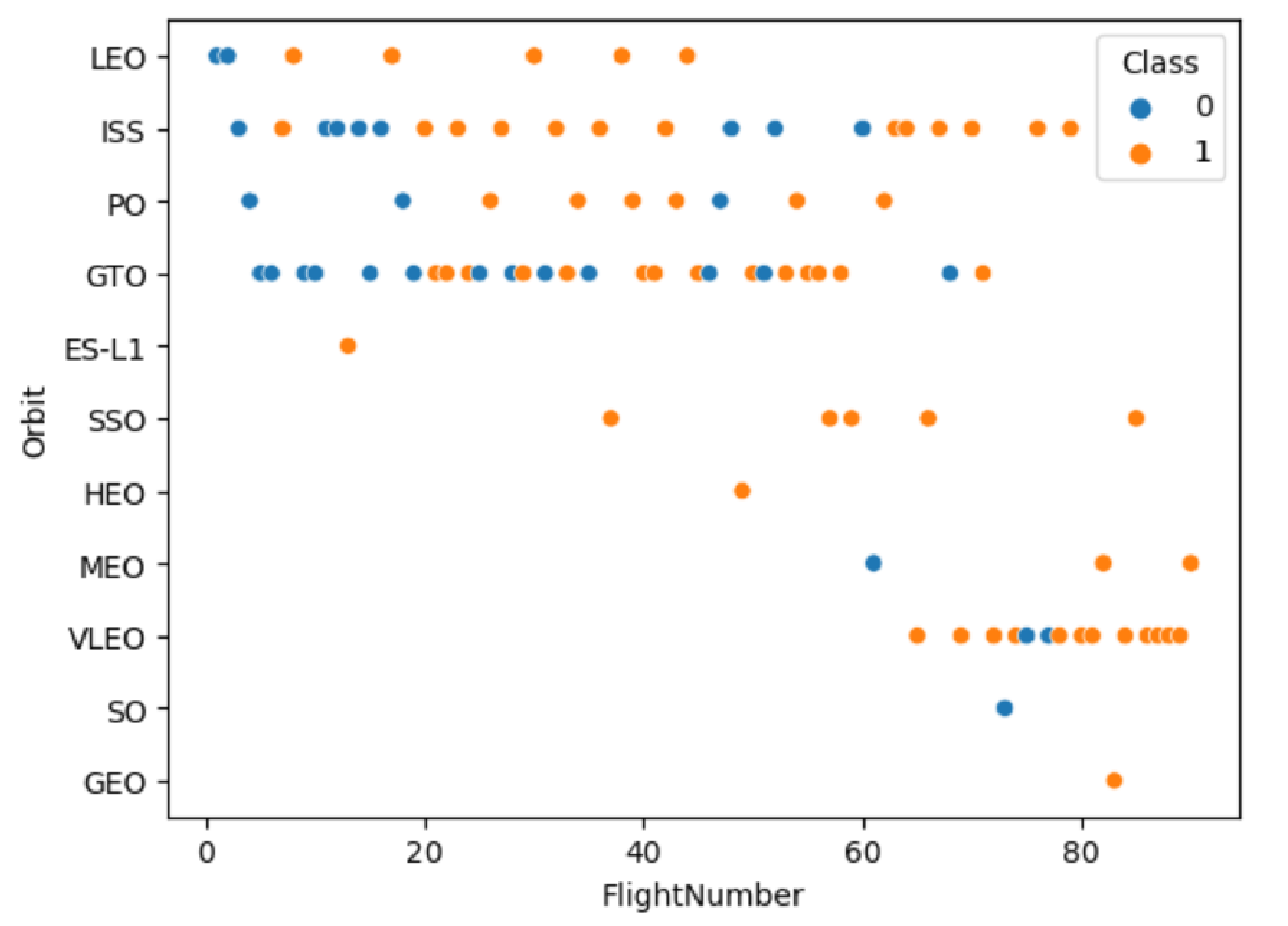
- VAFB-SLC has not launched any rockets having payload more than 10K. CCAFS-CLC 40 has more success rate for heavy payloads more than 12k

Success Rate vs. Orbit Type



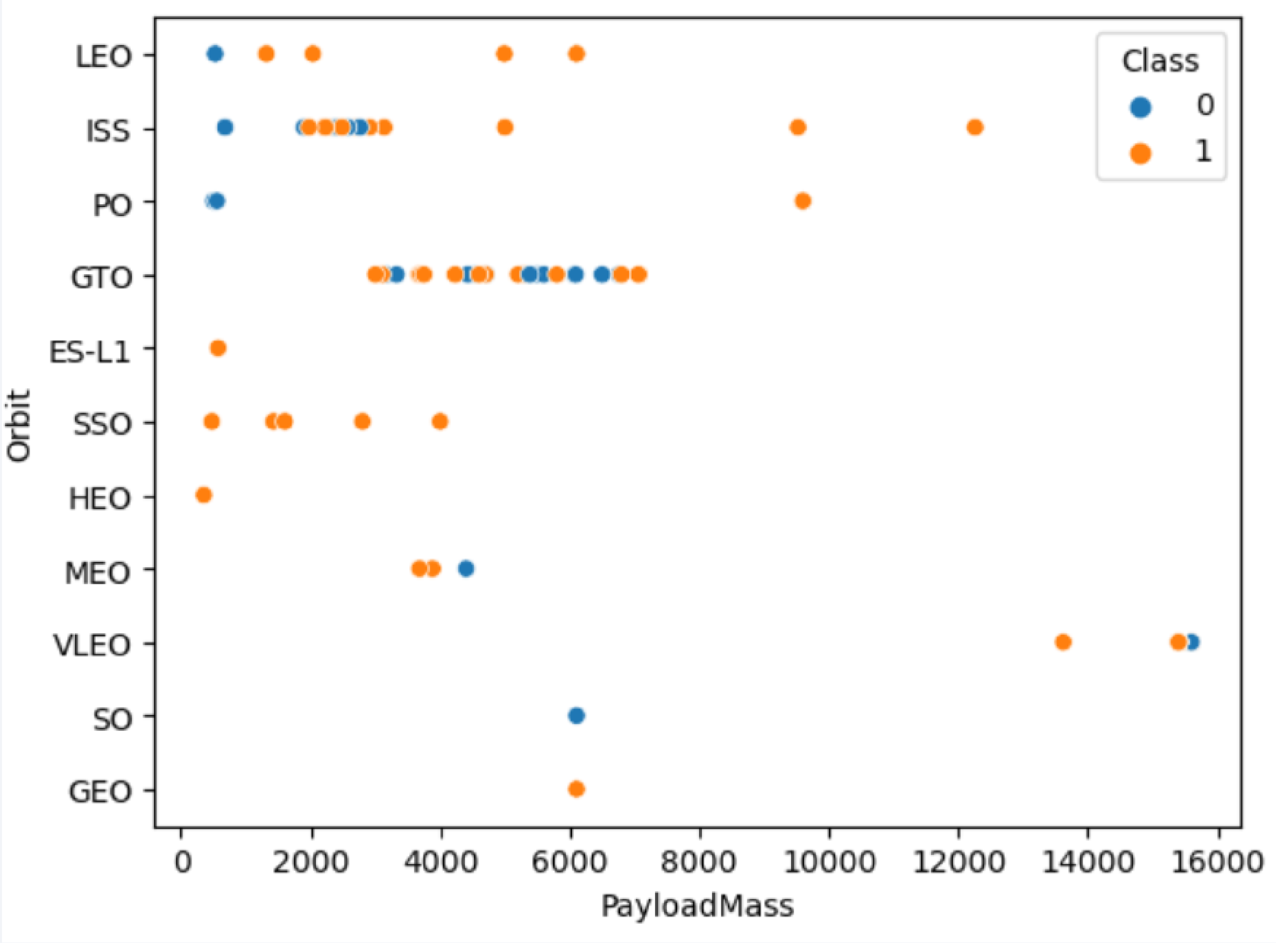
- ES-L1 , GEO , HEO and SSO orbit types have highest success rate and GTO has the lowest success rate.

Flight Number vs. Orbit Type



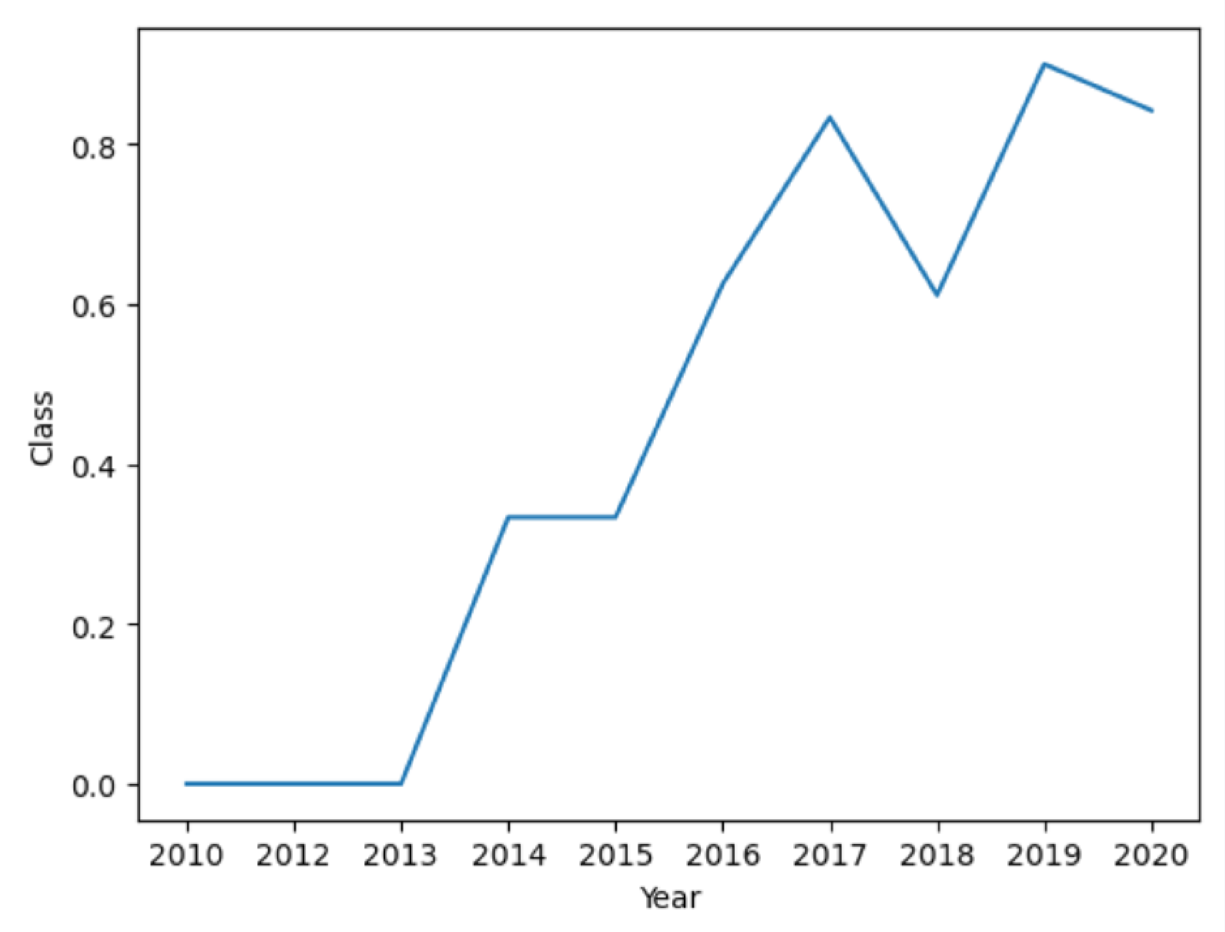
- LEO orbit is having better success rate as flight number increases and GTO orbit has no relation with flight number in case of success rate

Payload vs. Orbit Type



- SSO has 100% success rate. LEO, Polar and ISS performs better with heavy payload

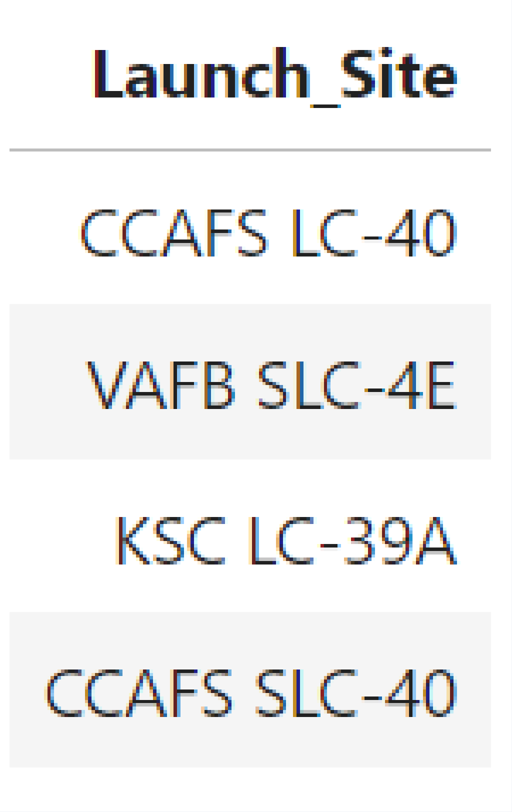
Launch Success Yearly Trend



- Initially, success rate is low till 2013 and then it increases with the years .

All Launch Site Names

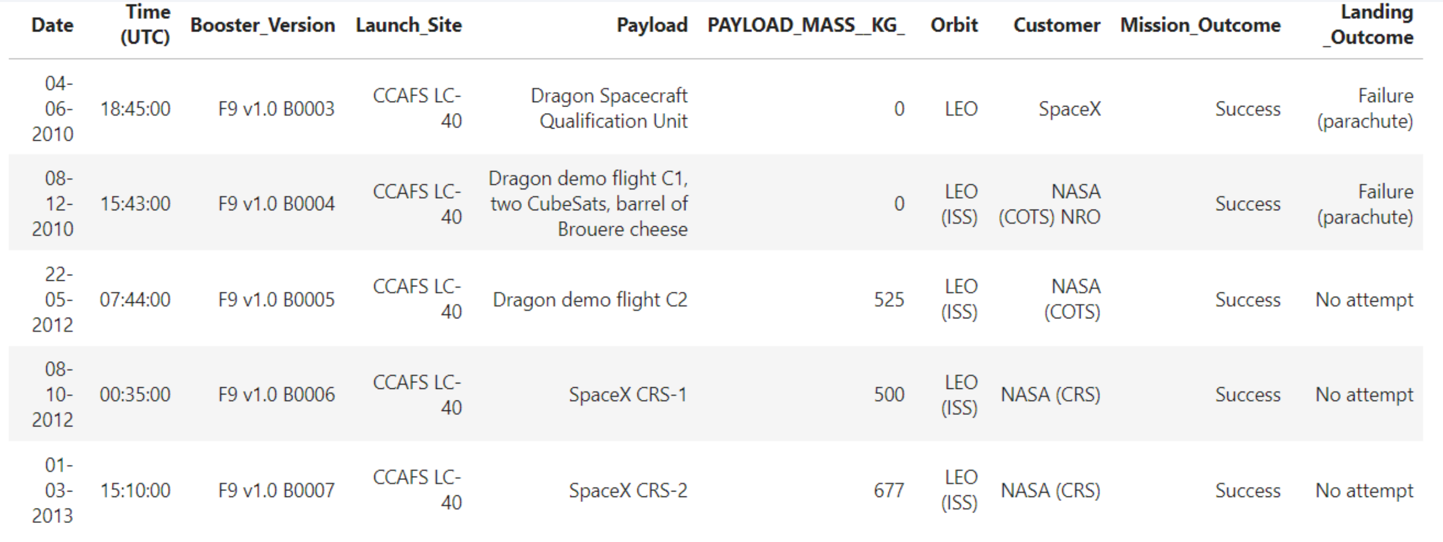
- The unique launch sites
- SQL select query with distinct clause to get all launch sites name. There are 4 launch site in the mission.



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

Launch Site Names Begin with 'CCA'

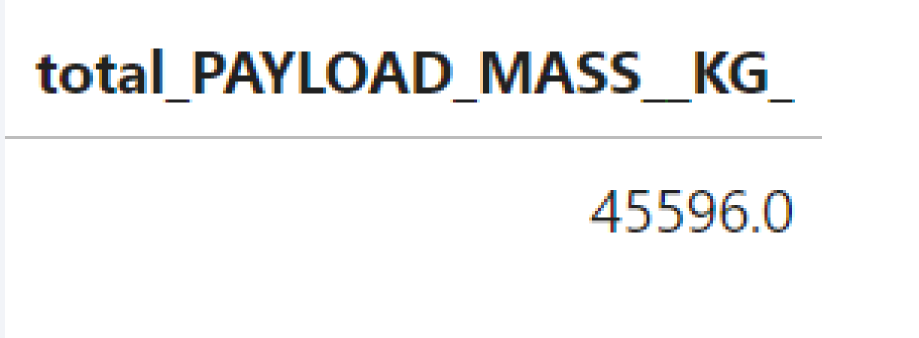
- First 5 records where launch sites begin with `CCA`
- SQL select query to get launch sites using filter and limit operation. All initial 5 record are from same launch site.



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

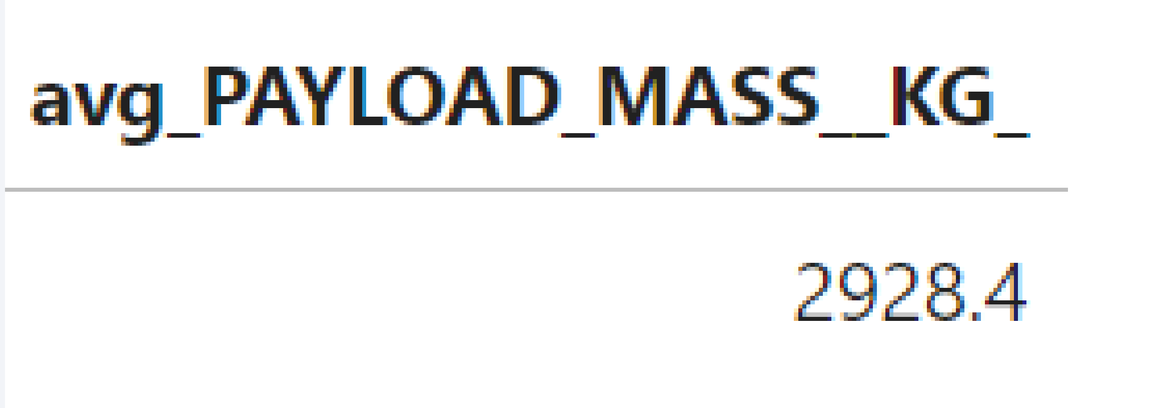
- The total payload carried by boosters from NASA
- SQL query with TOTAL function to get total payload mass kg of customer NASA



total_PAYLOAD_MASS__KG_
45596.0

Average Payload Mass by F9 v1.1

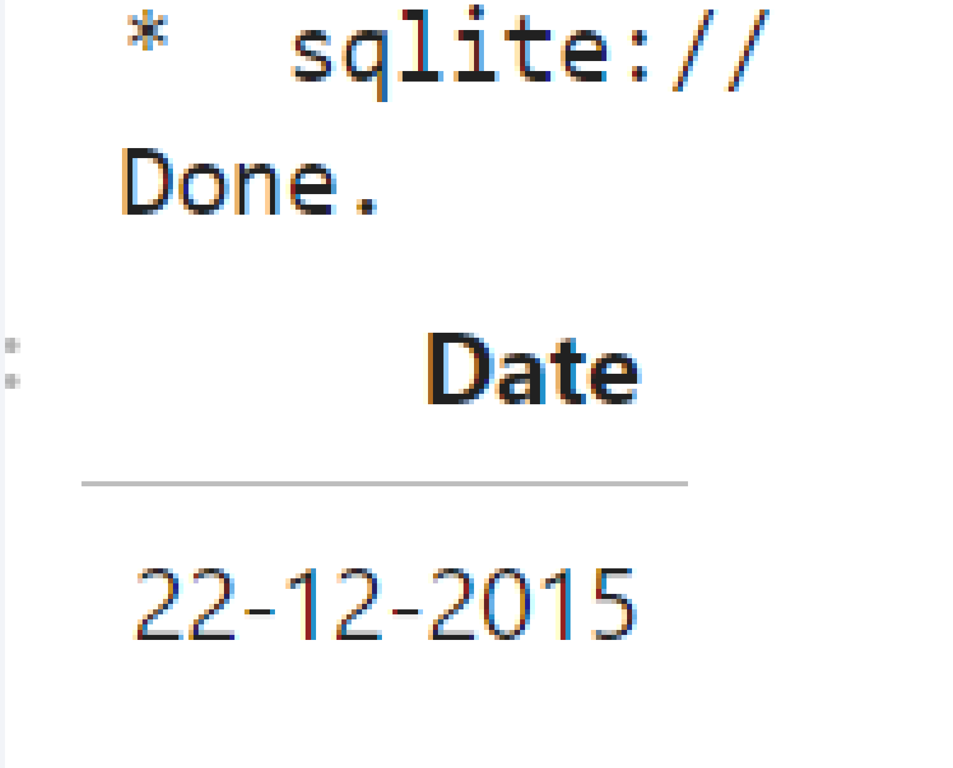
- Average payload mass carried by booster version F9 v1.1
- SQL query with AVG function to get average payload mass kg using filter on booster version F9 v1.1



avg_PAYLOAD_MASS__KG_
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad
- SQL query with MIN function to get first date for landing outcome 'Success (ground pad)' .



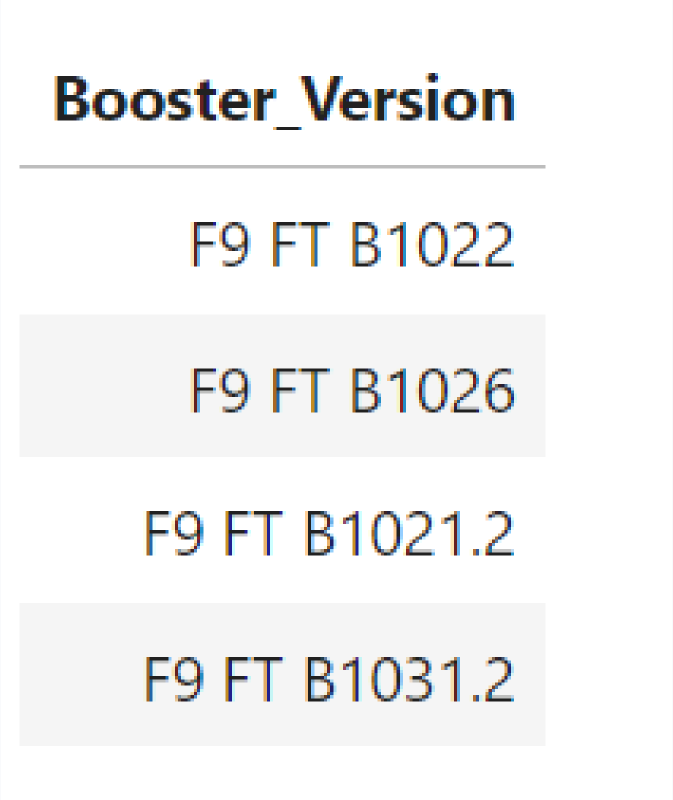
```
*  sqlite://
Done.

      Date
  ----------
  22-12-2015
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

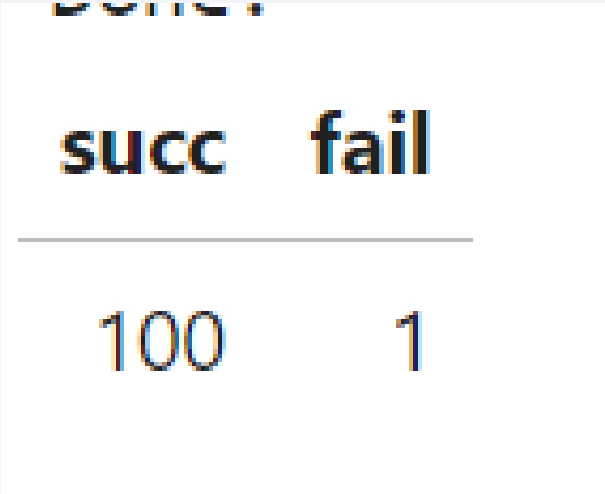
- SQL select query to get booster version where landing outcome is "Success (drone ship)" and payload is between 4k to 6K. All booter version are variant of F9 FT.



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

Total Number of Successful and Failure Mission Outcomes

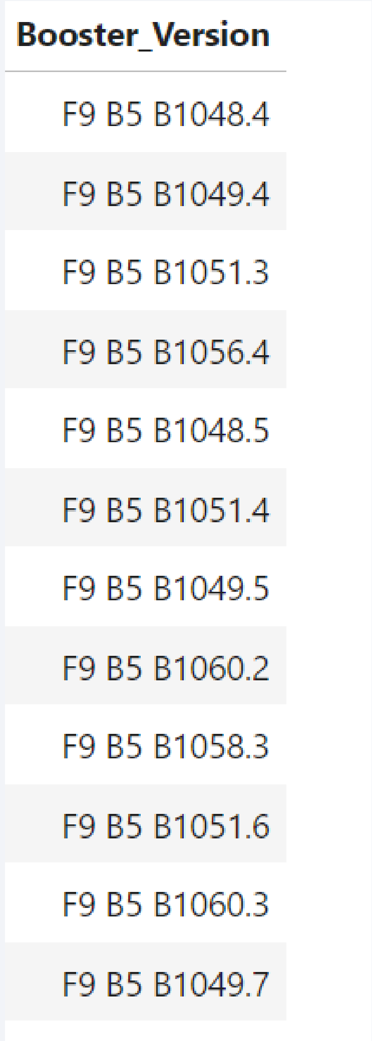
- The total number of successful and failure mission outcomes
- SQL query to count success and failure mission in all mission outcome



succ	fail
100	1

Boosters Carried Maximum Payload

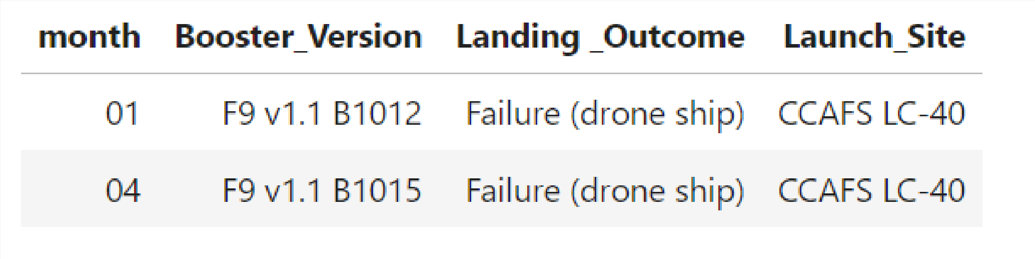
- The names of the booster which have carried the maximum payload mass
- SQL query to get all booster version where PAYLOAD_MASS__KG is greater than equal to maximum PAYLOAD_MASS__KG . Variants of F9 has the highest payload mass.



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

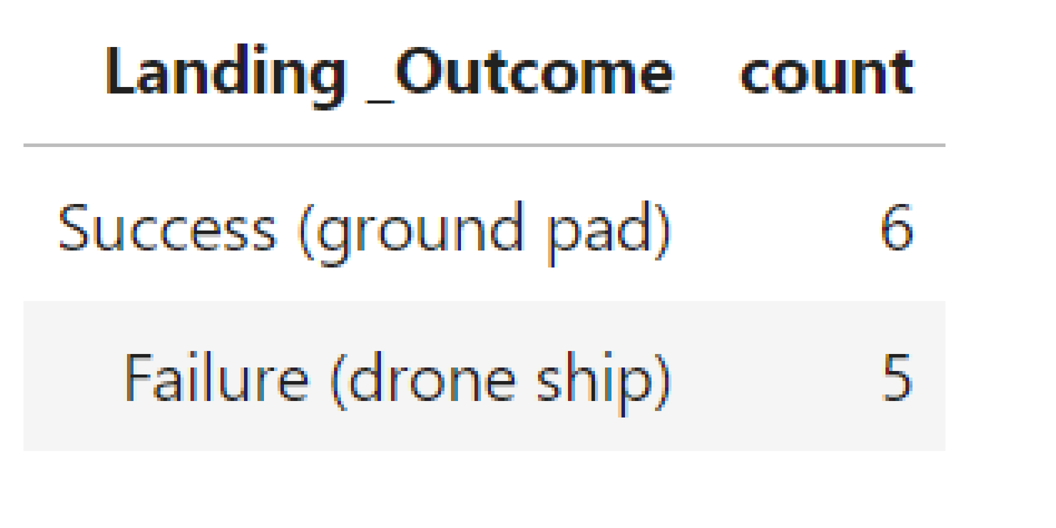
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- SQL select query to find month, landing_outcome, booster version and launch site where year is 2015 and landing outcome is equal to ''Failure (drone ship)' . In 2015 both failure came from the same launch site.



month	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

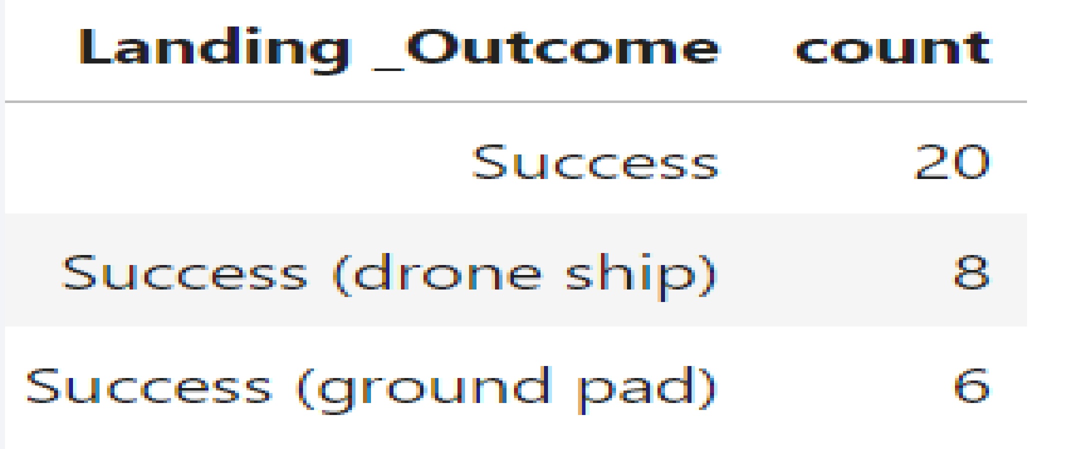
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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
Success (ground pad)	6
Failure (drone ship)	5

- SQL query to get landing outcome from the 2010-06-04 and 2017-03-20, in descending order. Drone ship has the highest success count.

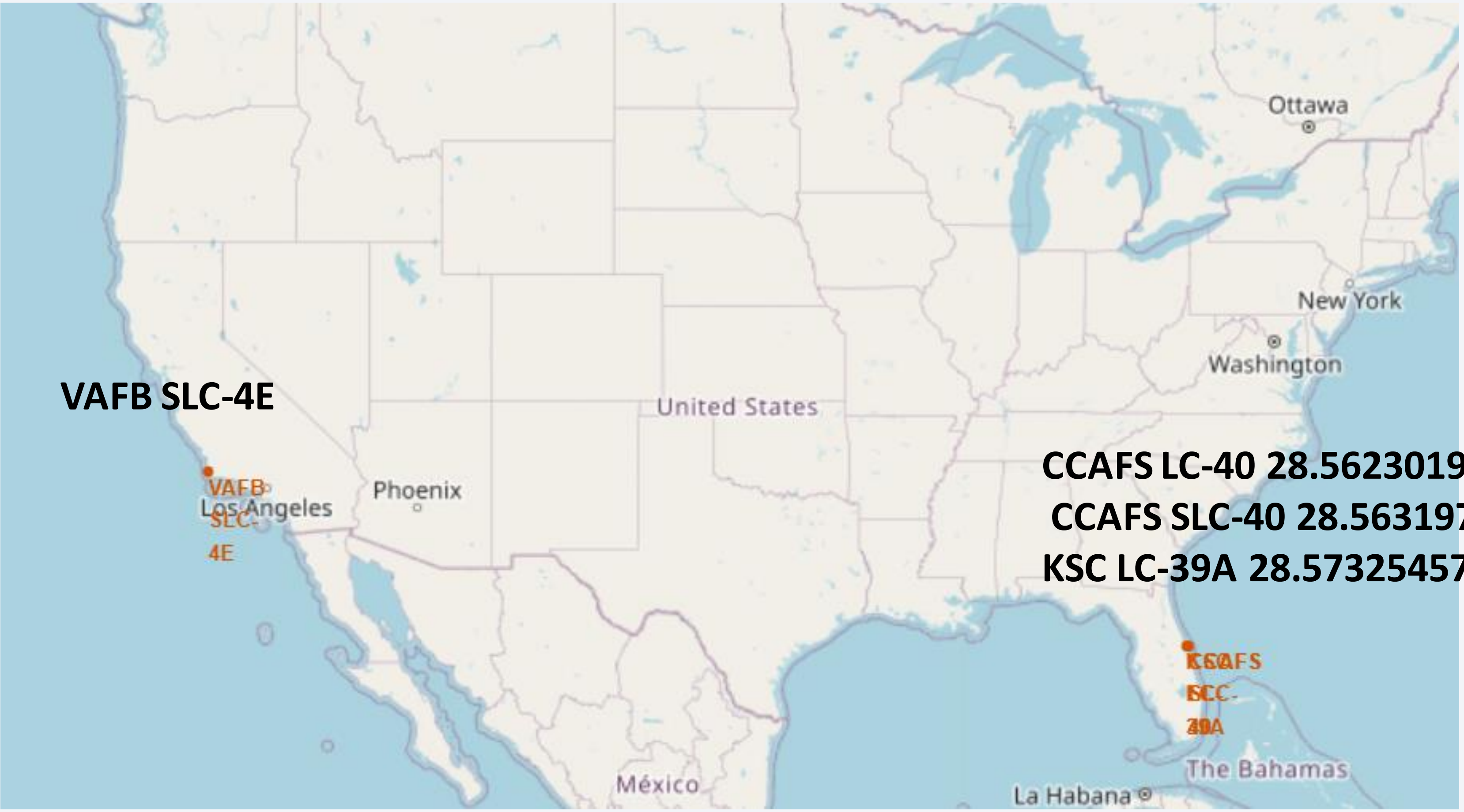


Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

All Launch Sites in the Map

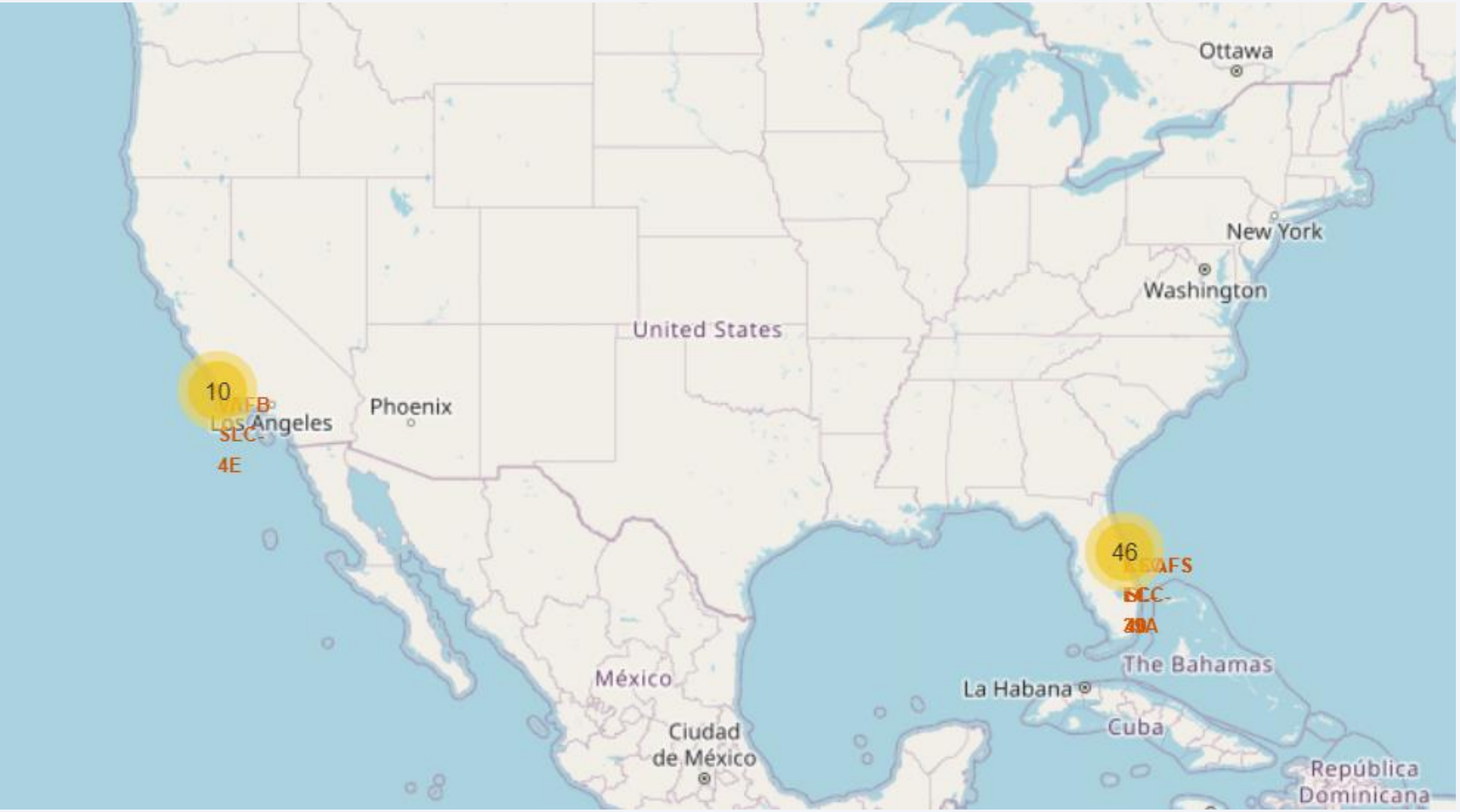


VAFB SLC-4E

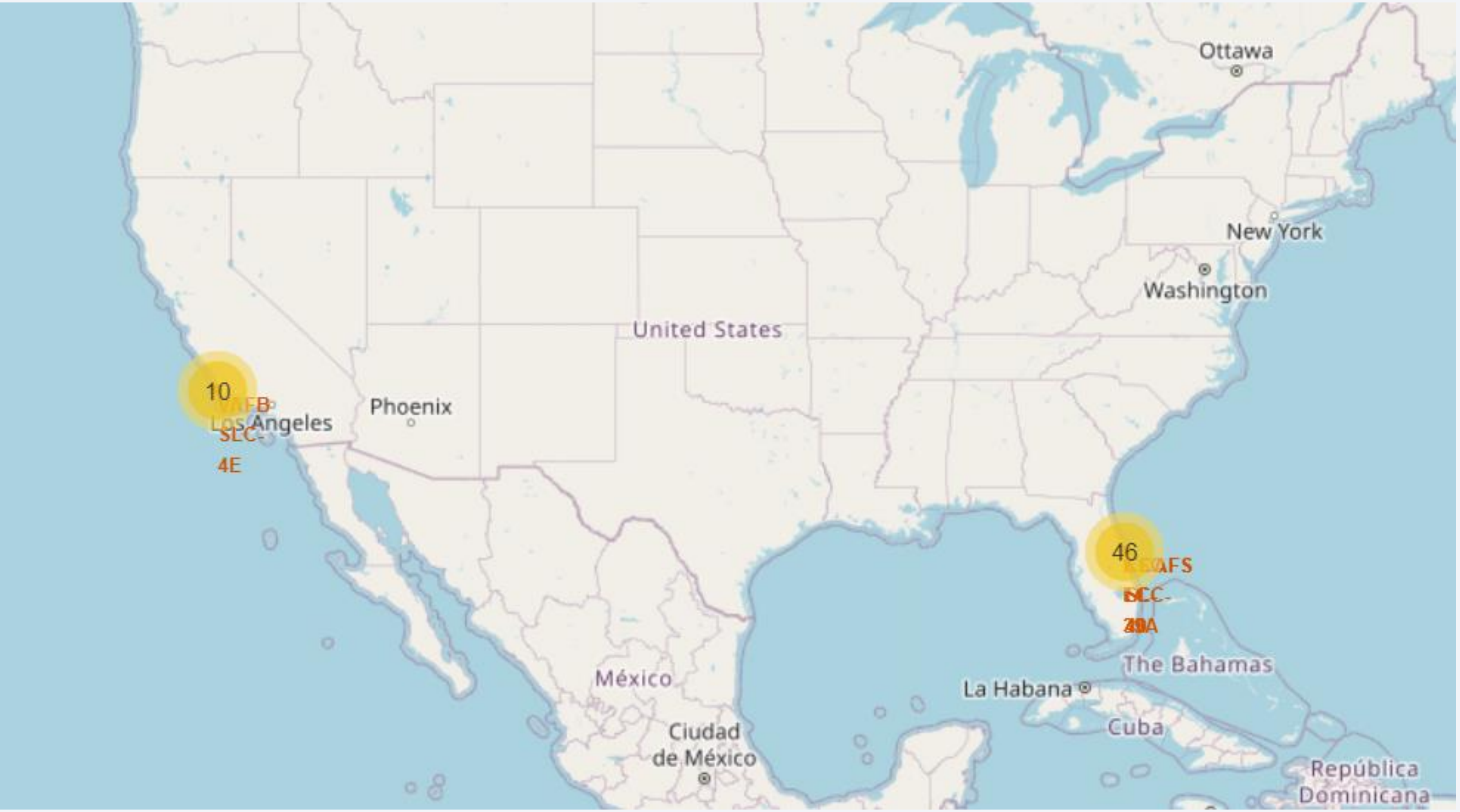
CCAFS LC-40	28.56230197	-80.57735648
CCAFS SLC-40	28.56319718	-80.57682003
KSC LC-39A	28.57325457	-80.64689529

CCAFS LC-40 and CCFAS SLC-40 are very close to each other near east coastline and VAFB SLC-4E is very far from other sites and near to west coastline.

Success/Failure for each site in the map

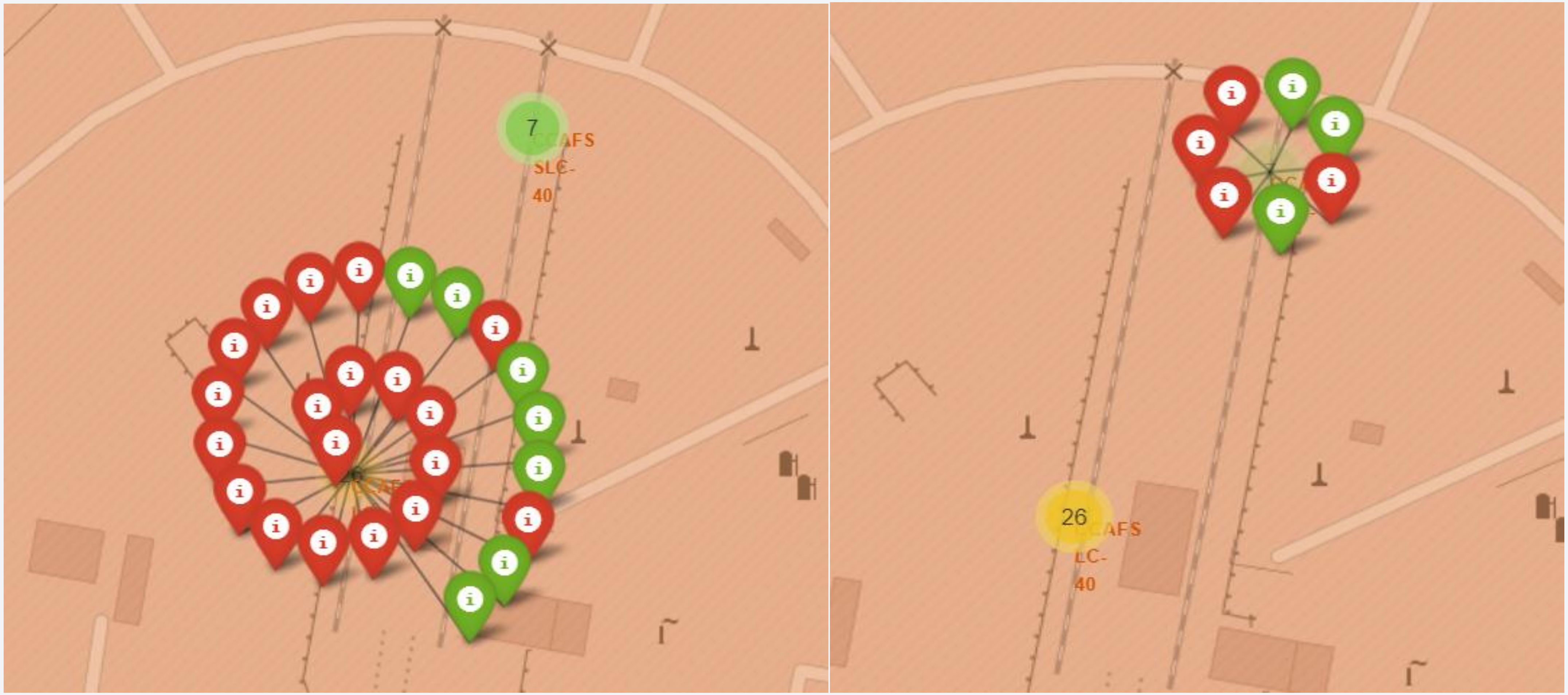


Success/Failure for each site in the map



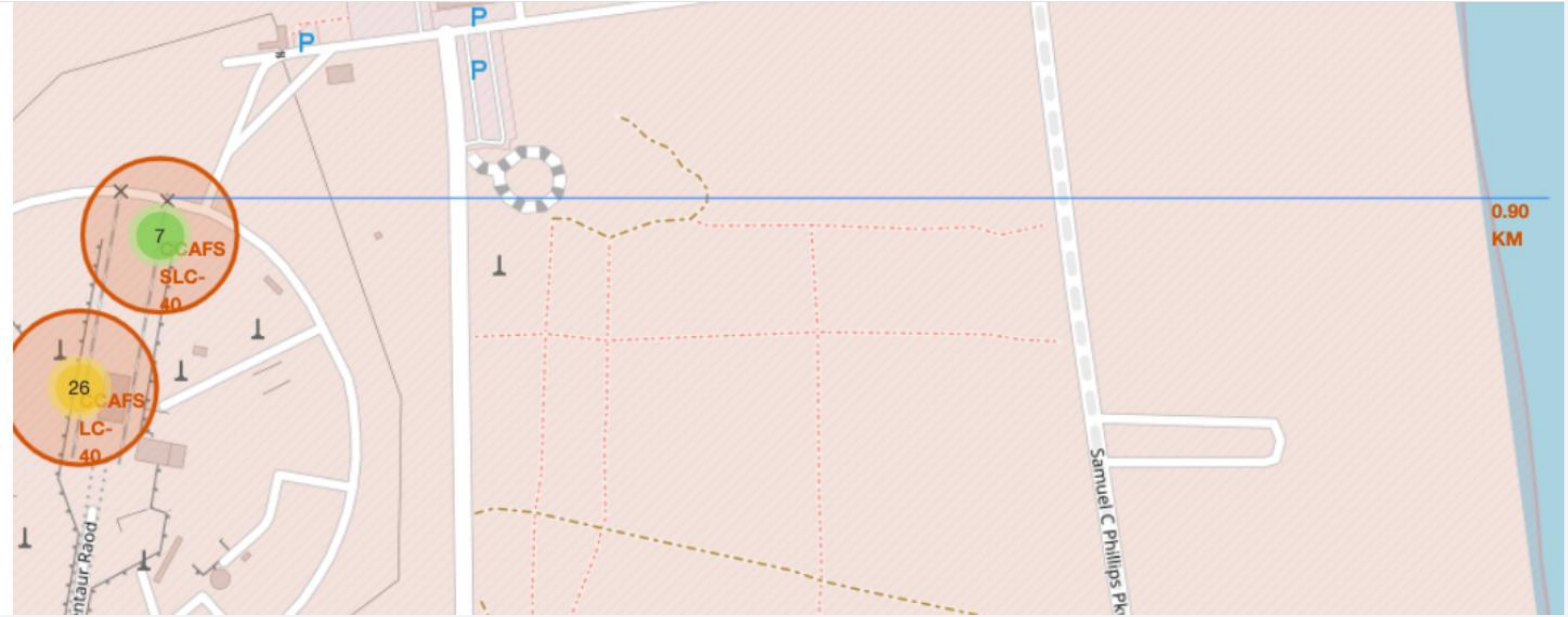
Total number of launches of each sites shown in yellow color

Success/Failure for CCAFS sites in the map



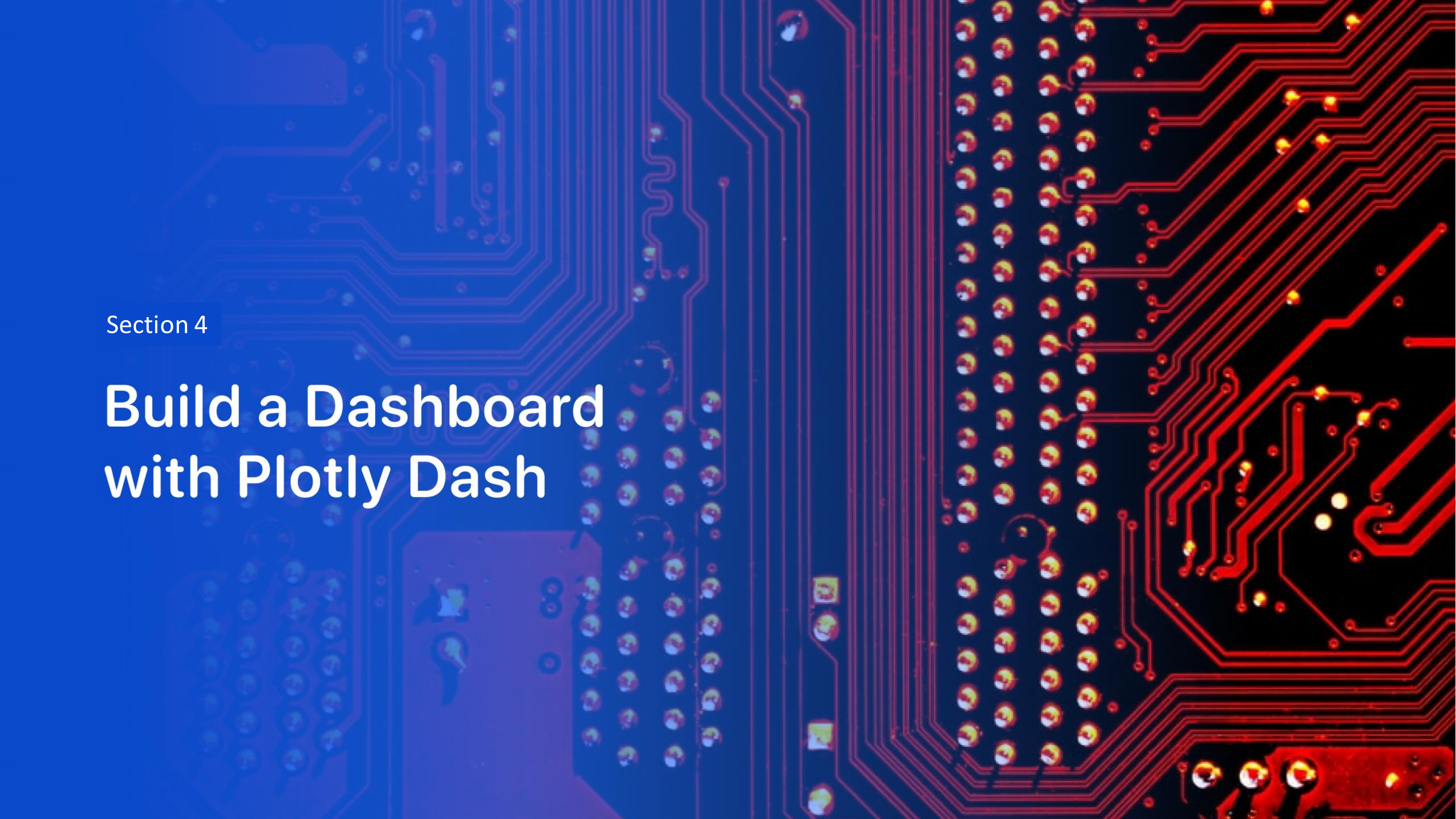
Success for CCAFS sites is in green color and failure is in red color.

The distances between a launch site to its proximities

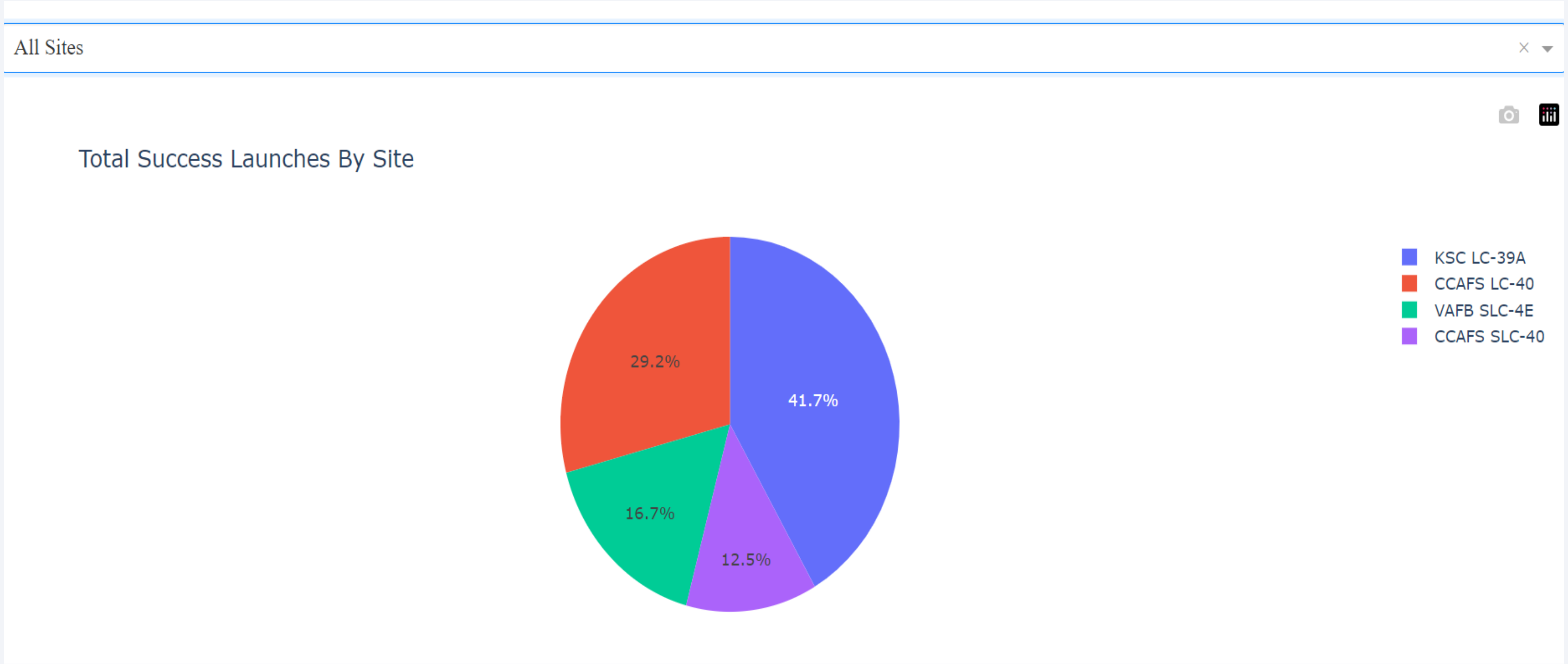


Section 4

Build a Dashboard with Plotly Dash



SpaceX Launch Record Dashboard – Total Success Rate Launches by All Sites

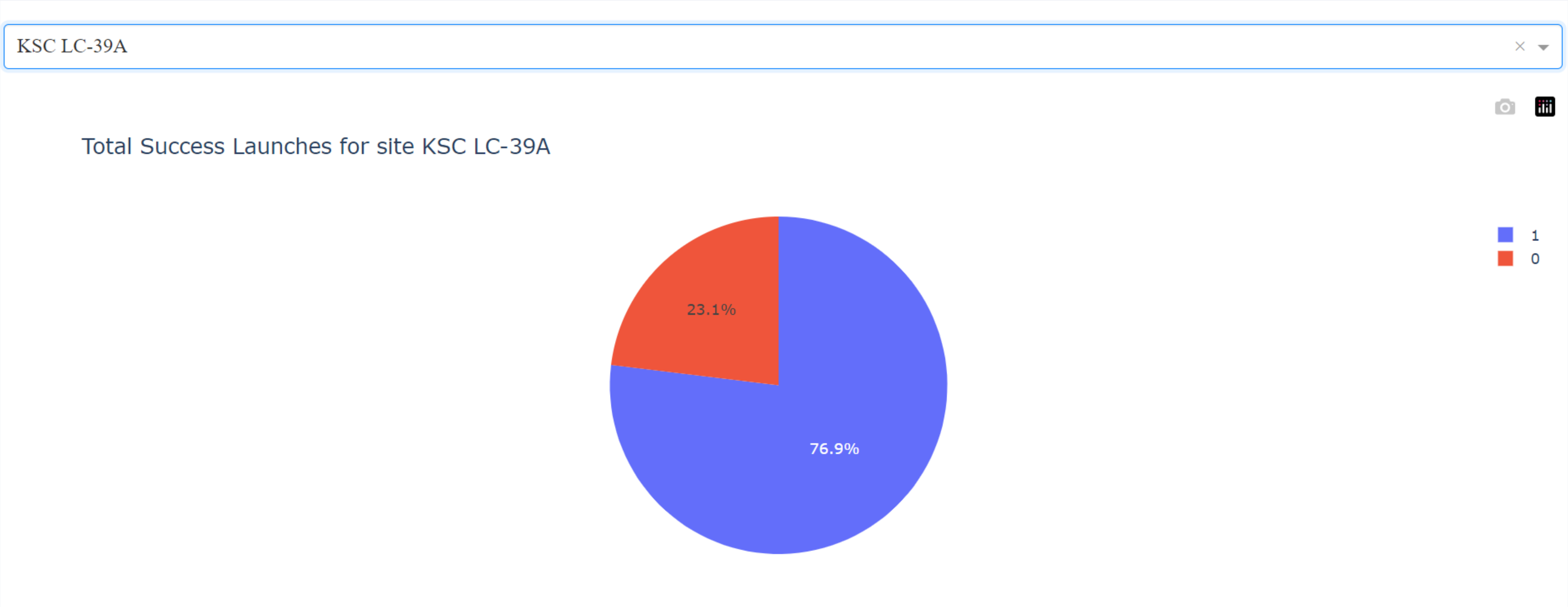


All Sites

Total Success Launches By Site

LSC LC-39A has maximum success launches ratio and CCAFS SLC has the lowest one

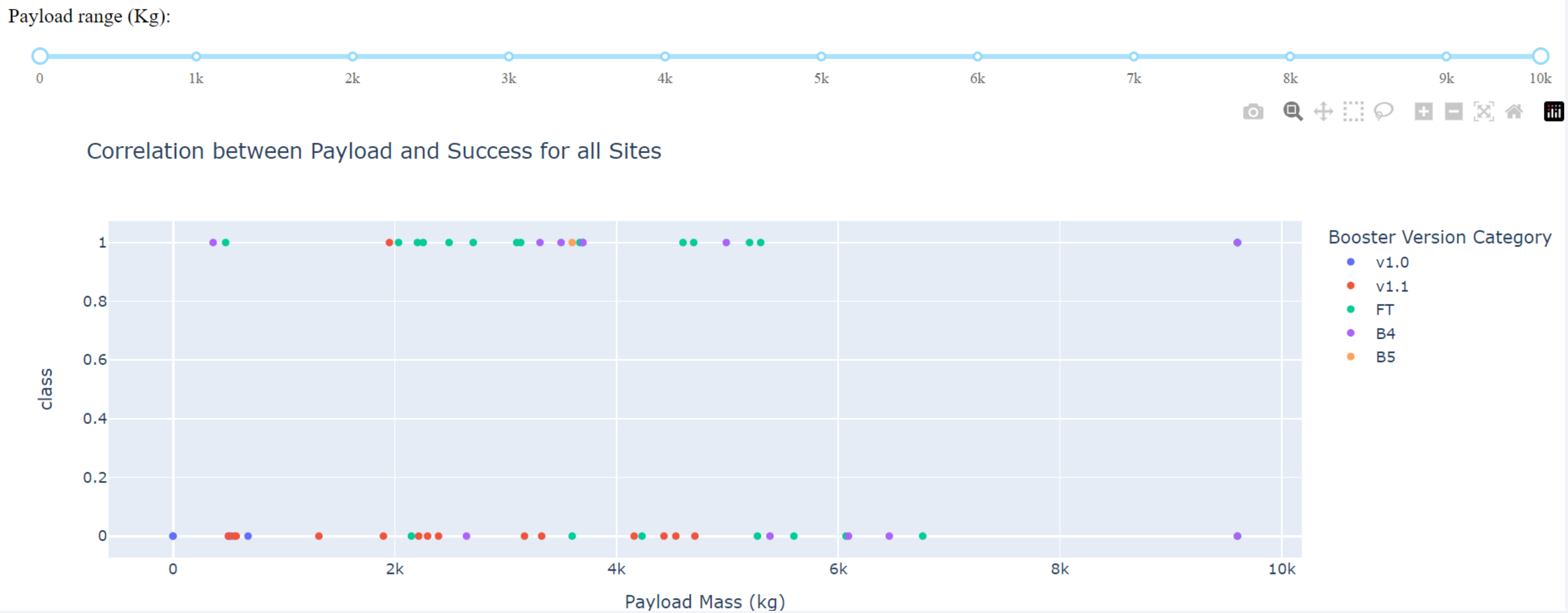
SpaceX Launch Record Dashboard – Percentage of success and failure by site KSC LC 39A



KSC LC-39A

The Pie Chart shows that KSC LC-39A sites has more than 75% success. Class success=1 and failure=0

SpaceX Launch Record Dashboard – Payload vs Launch Outcome



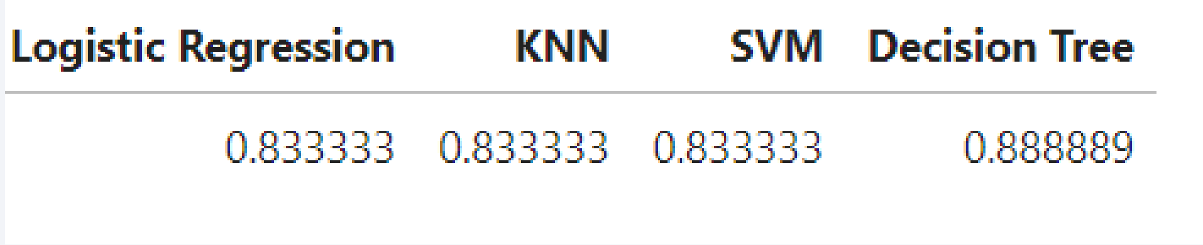
The Scatter plot shows that FT booster has very high success rate and payload mass more than 6k has very less success rate.

Section 5

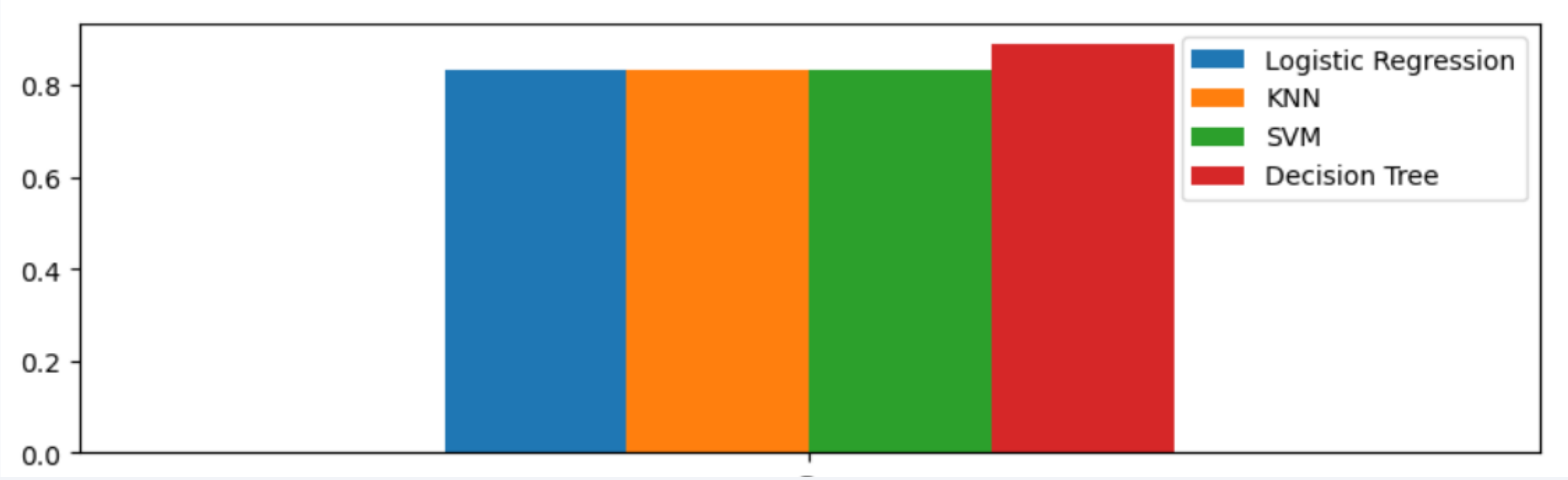
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree model has the highest classification accuracy

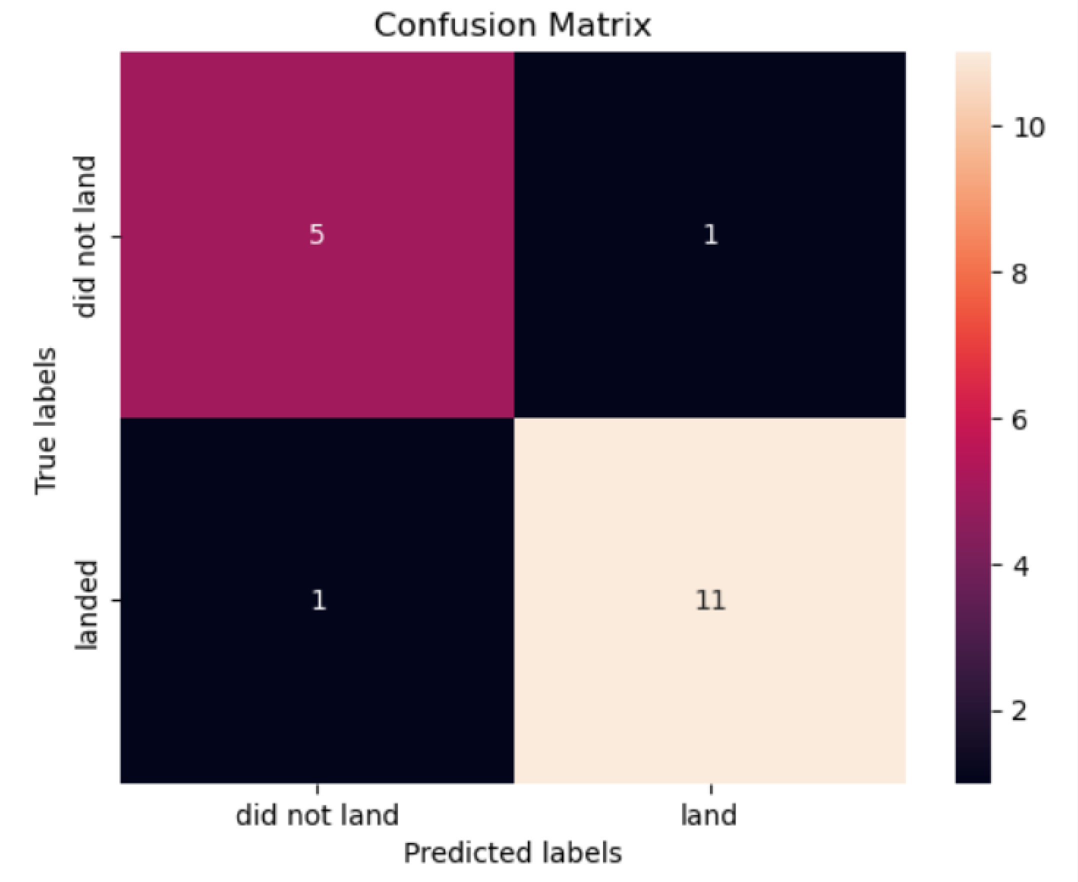


Logistic Regression	KNN	SVM	Decision Tree
0.833333	0.833333	0.833333	0.888889



Confusion Matrix

- Decision Tree model has the best f1-score among all models.

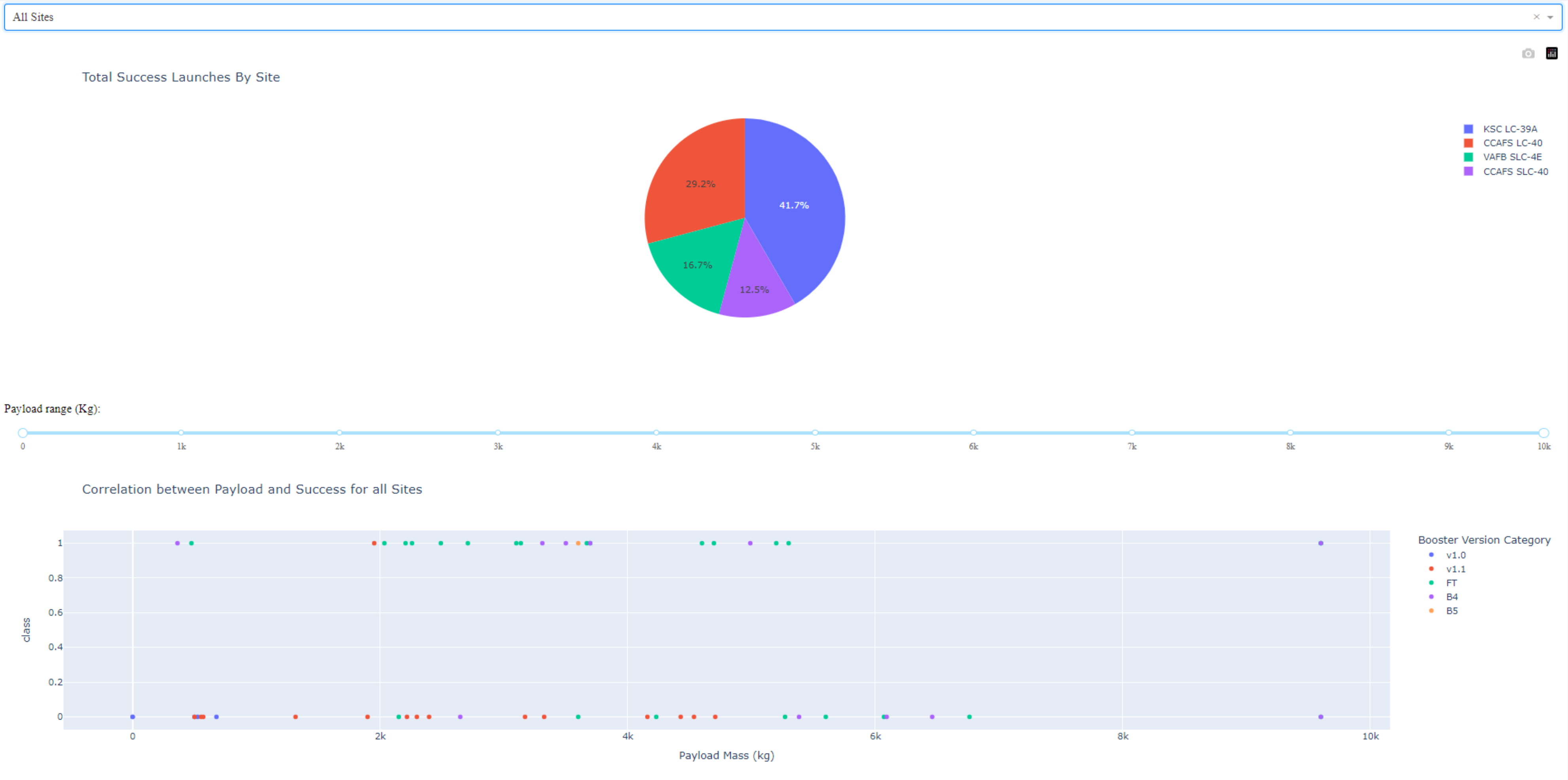


Conclusions

- SpaceX has edge over its competitor for launching satellite at low cost due to the high success rate of first stage landing outcome from 2013 onwards .
- First stage launch is more difficult with heavy payload and also landing outcome has low success rate.
- Launching site KSC LC-39A has the highest success launch.
- Decision tree model performs best due to high f1-score and has accuracy more than 5% compare to other models such as KNN, Logistic Regression and Support Vector Machine.
- ES-L1 , GEO, HEO and SSO orbit types have highest success rate and GTO has the lowest success. It require further inspection and data as other orbit types are considered significantly very less compare to GTO orbit.

Appendix

- Full dashboard Screenshot of SpaceX records



All Sites

Thank you!