IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Qian Gao>

<December 1st, 2022>

Outline

- Executive Summary
- Introduction
- Methodology
- Insights from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusions

Executive Summary

- Summary of methodologies
  - Data Collection with API; Data Collection with Web Scraping; Data Wrangling
  - EDA with SQL; EDA with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction.
- Summary of all results
  - EDA results
  - Folium results
  - Machine learning prediction results.

Introduction

- Project background and context

  Using data science, if one can determine if the first stage of SpaceX will land, one can determine the cost of a launch. This information should be useful if another company Space Y wants to compete with space X for a rocket launch. This objective of the project is to evaluate whether Space Y may compete with SpaceX by predicting if the first stage will land successfully through data science.

- Problems you want to find answers
  - The main factors for a successful landing and their interaction between them
  - Way to estimate the cost of landing
  - The best locations for a landing

Section 1

Methodology

Methodology

Executive Summary

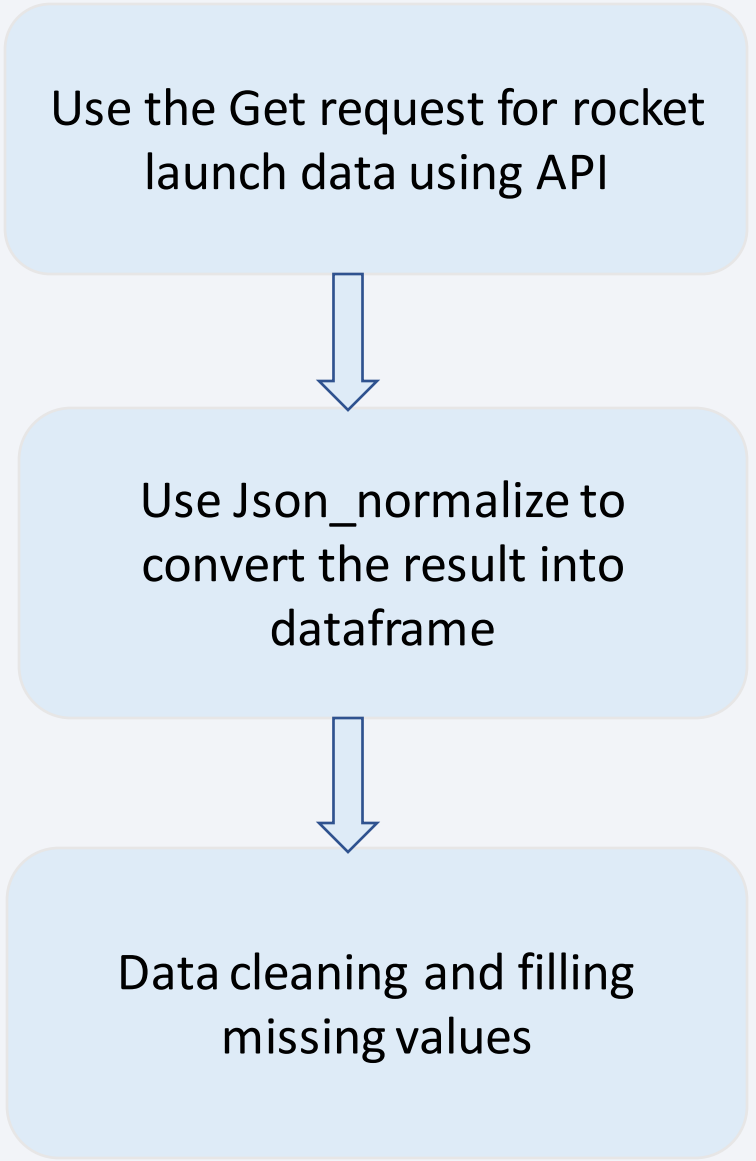
- Data collection methodology:
  - Data collected from SpaceX API and web scraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Creating a landing outcome label from outcome column based on collected
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - Data Collection with API (https://api.spacexdata.com/v4/rockets/)
  - Data is processed into pandas dataframe, cleaned with missing values filled
  - Data web scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

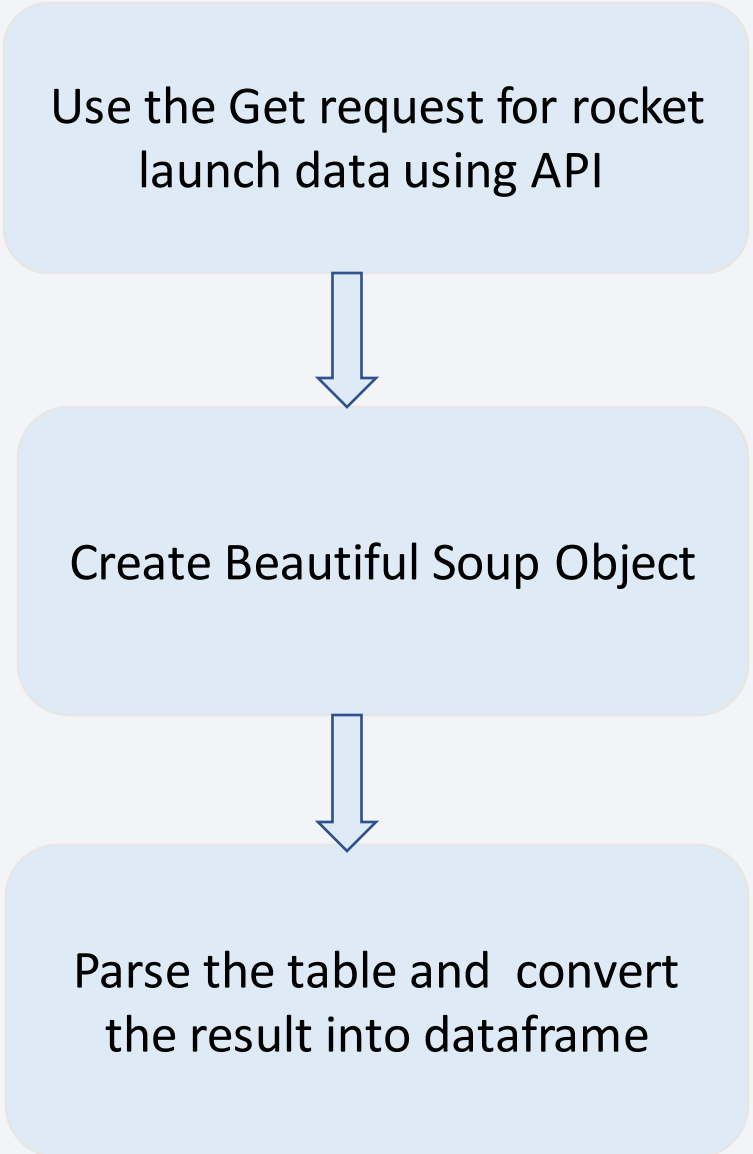
Data Collection – SpaceX API

- Use the Get request to launch data
- Convert data into dataframe
- Data cleaning and filling
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/spacex-data-collection-api.ipynb)



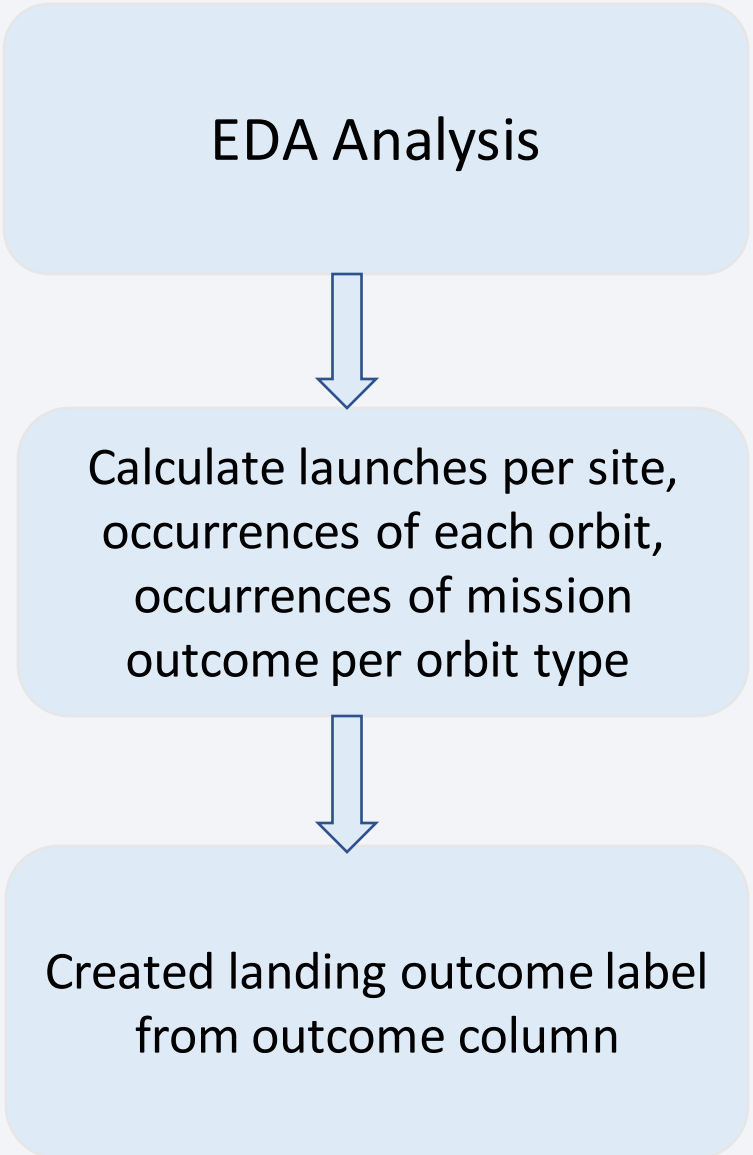
Data Collection - Scraping

- Use the Get request for rocket launch data
- Use web scrapping to Falcon 9 to launch records with BeautifulSoup
- Parse the table and convert the result into dataframe
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/spacex-data-webscraping.ipynb)



Data Wrangling

- Perform Exploratory Data Analysis (EDA) on dataset
- Calculate launches per site, occurrences of each orbit, occurrences of mission outcome per orbit type
- Created landing outcome label from outcome column
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/spacex-data-wrangling.jupyterlite.ipynb)



EDA with SQL

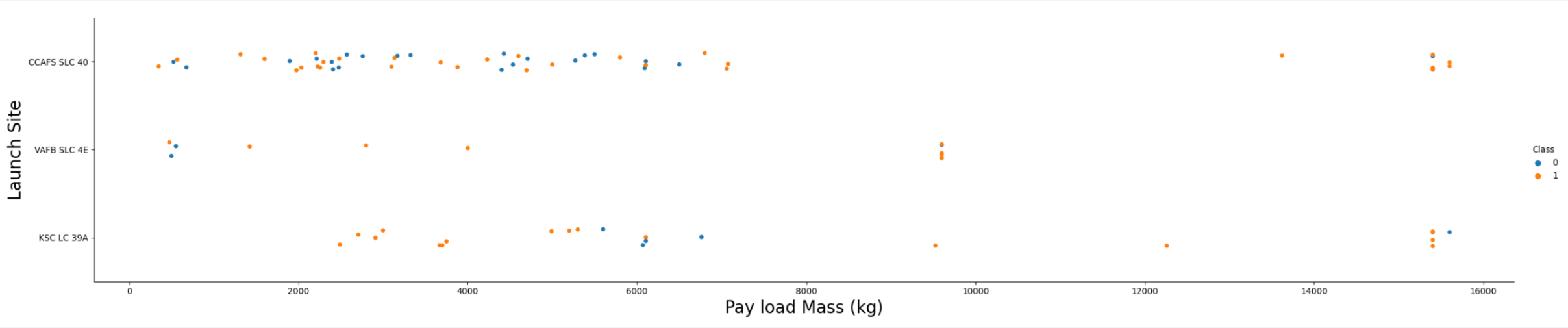
- SQL queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version 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
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship
  - Rank the count of successful landing_outcomes between 04-06-2010 and 20-03-2017 in descending order

EDA with Data Visualization

- Charts plotted include

  Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

- Reason to plot the charts is to explore the relationships between the features



- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/eda-data-visualization.ipynb.jupyterlite.ipynb)

Build an Interactive Map with Folium

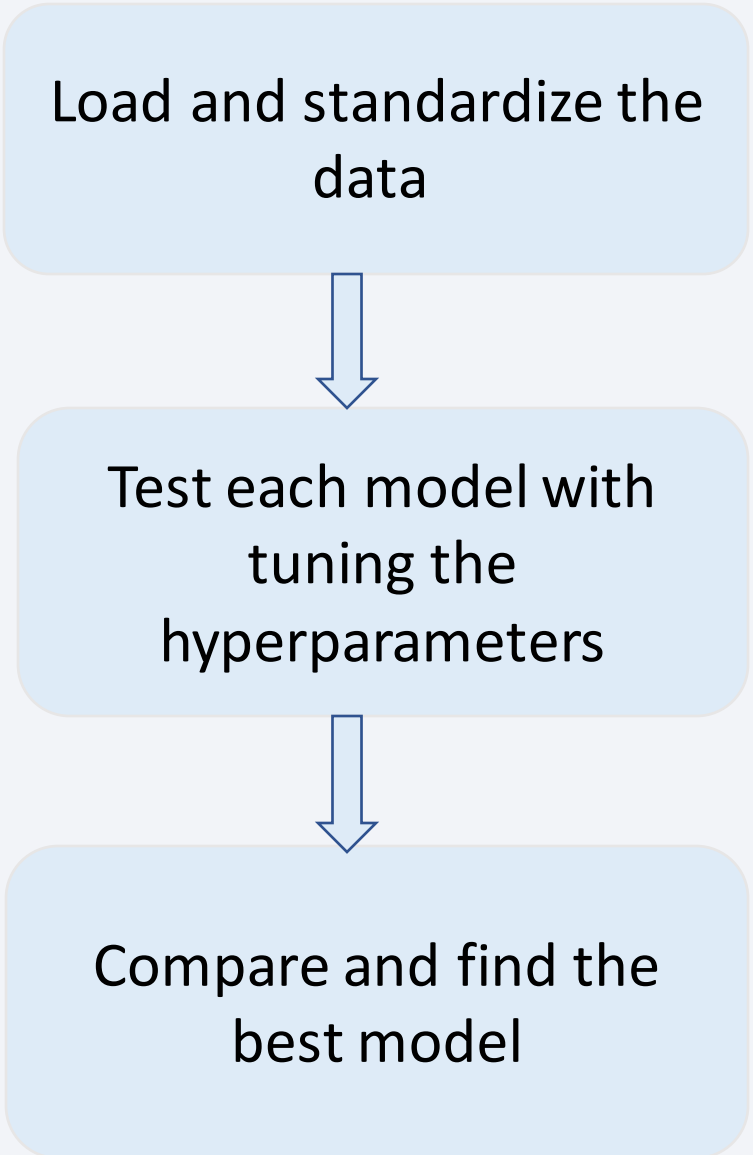
- Markers denotes points such as launch sites; circles denotes highlighted areas around specific coordinates; marker clusters denotes groups of events in each coordinate; lines denotes distances between coordinates
- The objects are added to better answer the questions such as whether launch sites near railways, highways and coastlines; and whether they keep a distance from residential areas in cities.
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/interactive-analytics-folium.ipynb)

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard include chart showing the total launches and graph showing the relationship between outcome and payload mass
- The reason to include the plot is to figure out where is best place to launch based on payloads.
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/plotly-dash.py)

Predictive Analysis (Classification)

- Load and standardize the data by pandas and numpy
- Build and test each model with tuning the hyperparameters by GridSearchCV
- Four models are compared, including logistic regression, support vector machine, decision tree and k nearest neighbors
- GitHub URL of the notebook (https://github.com/qgaoaggienetwork/coursera-capstone-project/blob/master/machine-learning-prediction.ipynb)



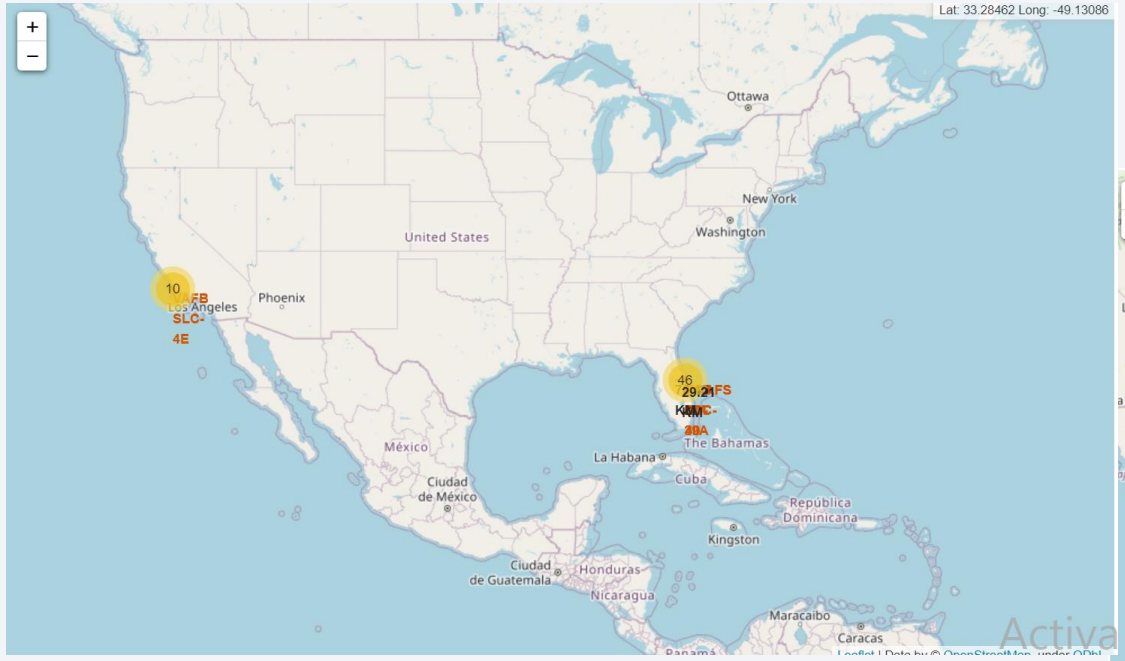
Results

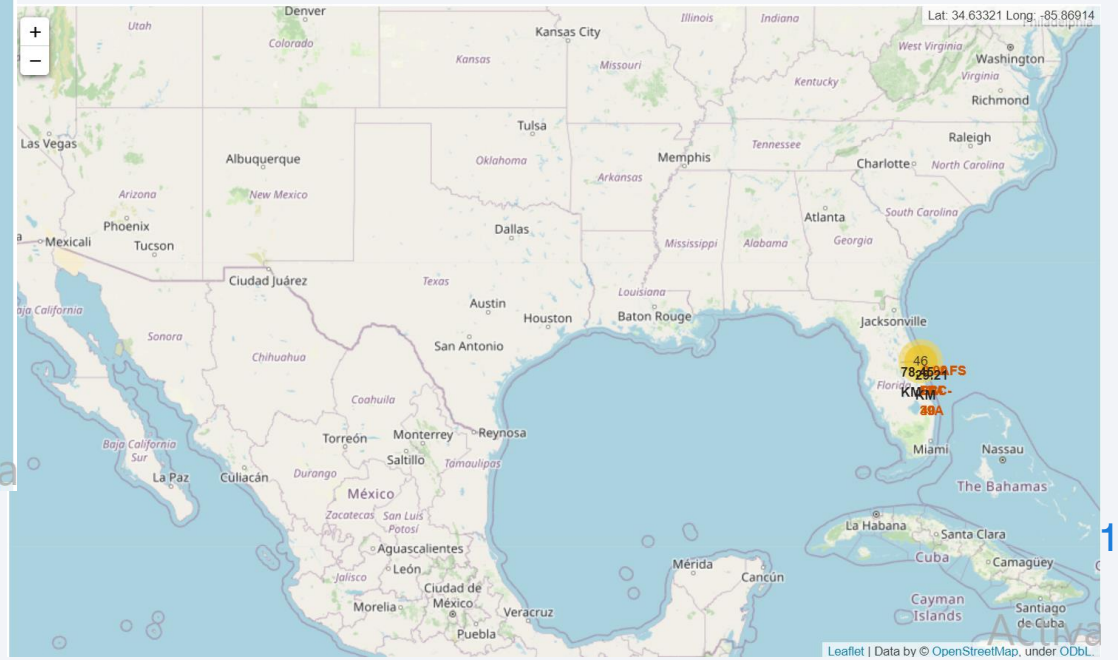
- Exploratory data analysis results
  - The unique launch sites include CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
  - The first 5 launches are done for customer Space X and NASA
  - The total payload mass carried by booster launched by NASA is 111268kg
  - The average payload mass carried by booster version F9 v1.1 is 2928kg
  - The date for first successful landing in ground pad was 2015-12-22
  - The names of boosters which have successfully landed in drone ship and have payload mass greater than 4000 but less then 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, F9 FT B1026
  - The number of failed missions is 1 and successful is 99

Results

- Interactive analytics results

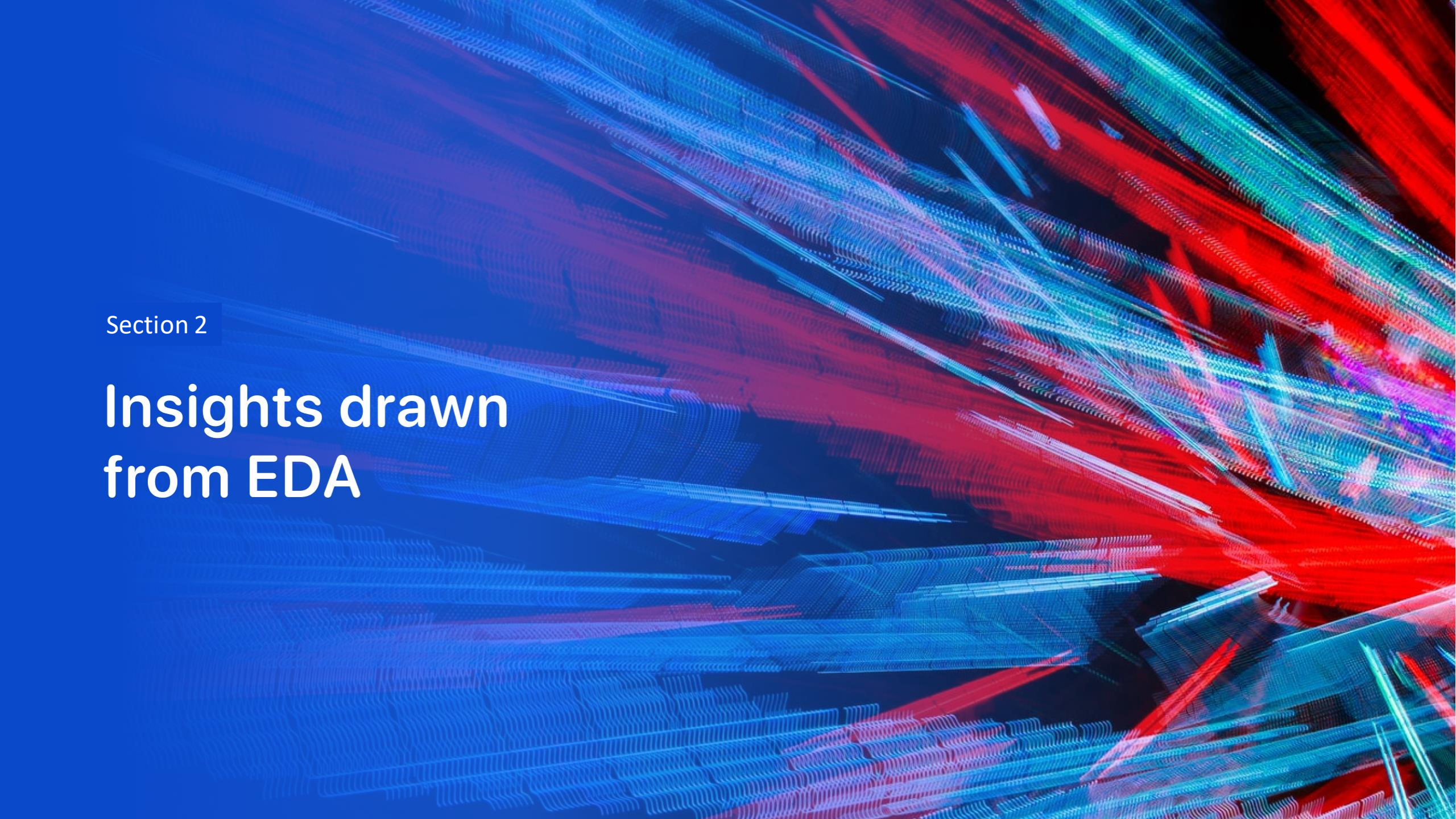
  Interactive analytics is used to find whether launch sites near railways, highways and coastlines; and whether they keep a distance from residential areas in cities





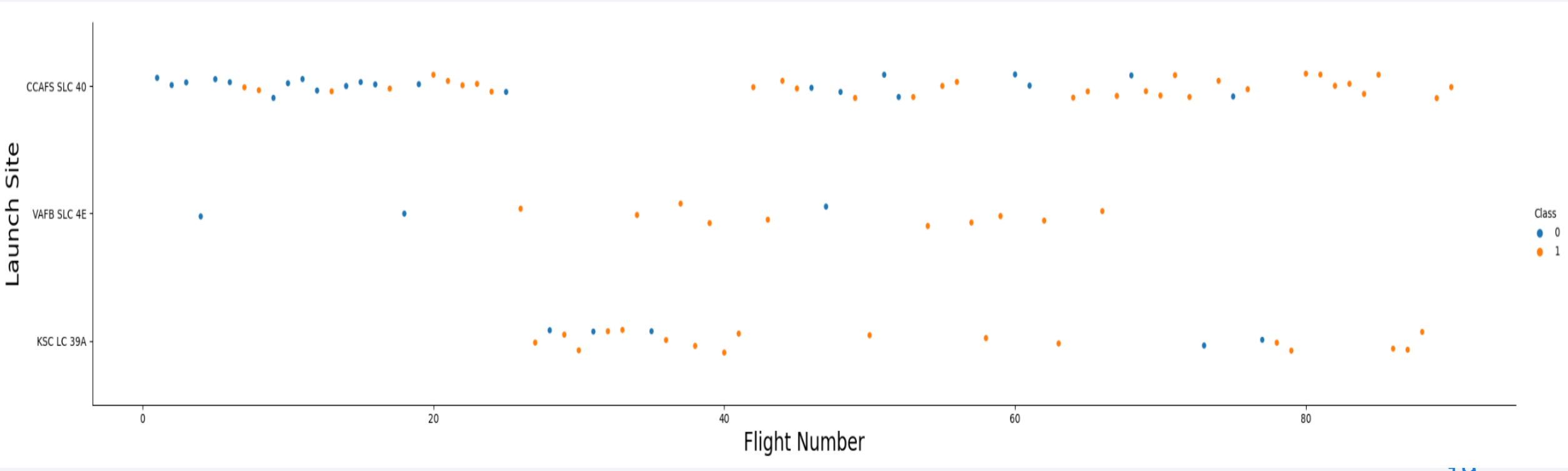
Section 2

Insights drawn from EDA



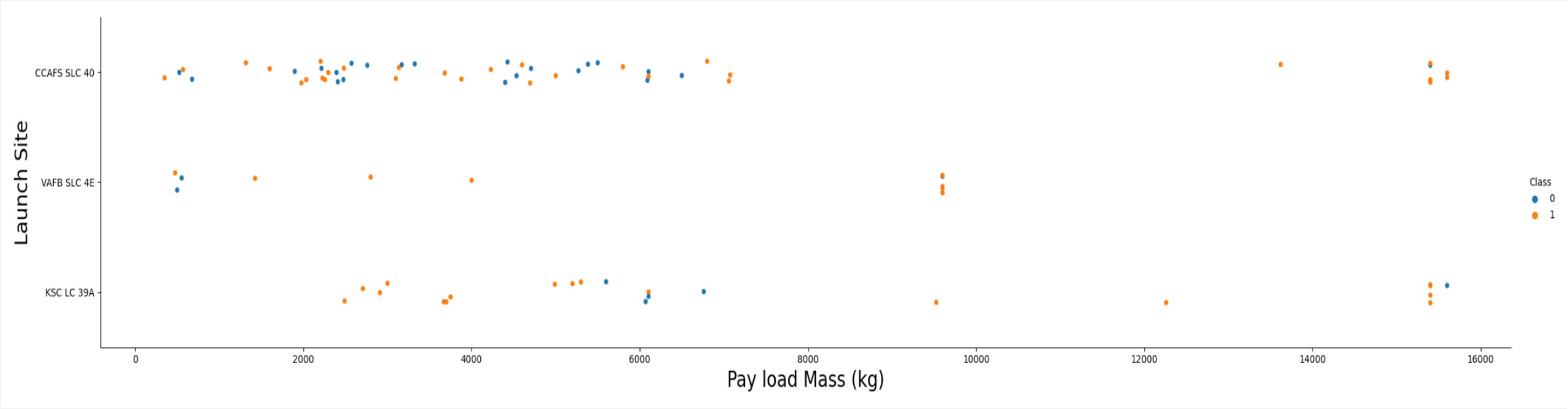
Flight Number vs. Launch Site

- The larger the number of flights, the higher the successful launch rate;
- The best launch site is CCAFS SLC 40.



Payload vs. Launch Site

- The larger the payload mass, the higher the successful launch rate;
- The VAFB SLC 4E launch site can only accommodate payload mass<10000kg.



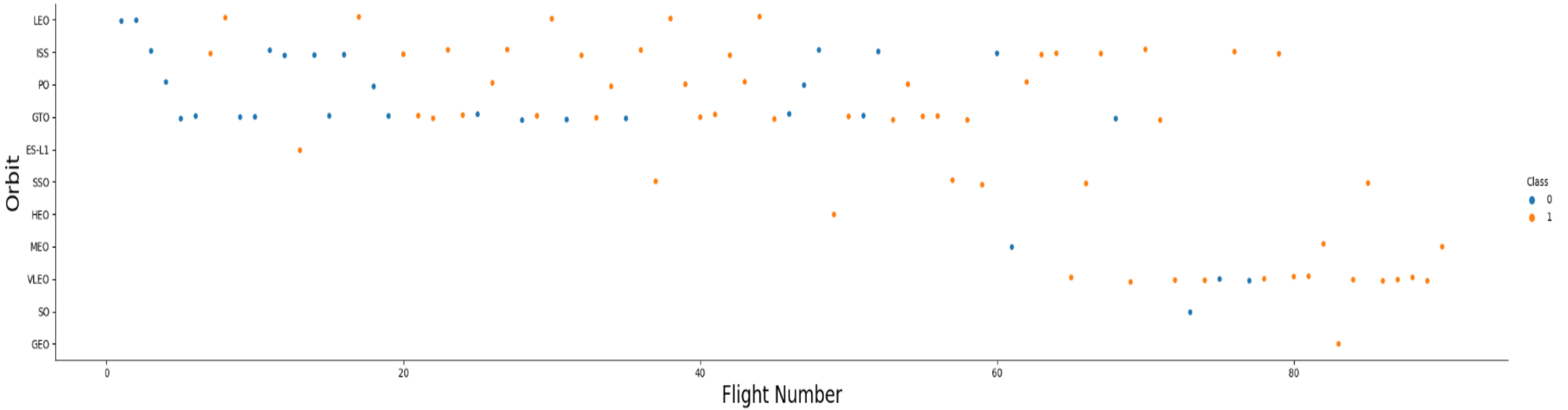
Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, SSO have the higher success rate of 100%, followed by VLEO, LEO, MEO
- The bar chart did not show from my code so I include the original data here.

```
Orbit
ES-L1    1.000000
GEO      1.000000
GTO      0.518519
HEO      1.000000
ISS      0.619048
LEO      0.714286
MEO      0.666667
PO       0.666667
SO       0.000000
SSO      1.000000
VLEO     0.857143
Name: Class, dtype: float64
```

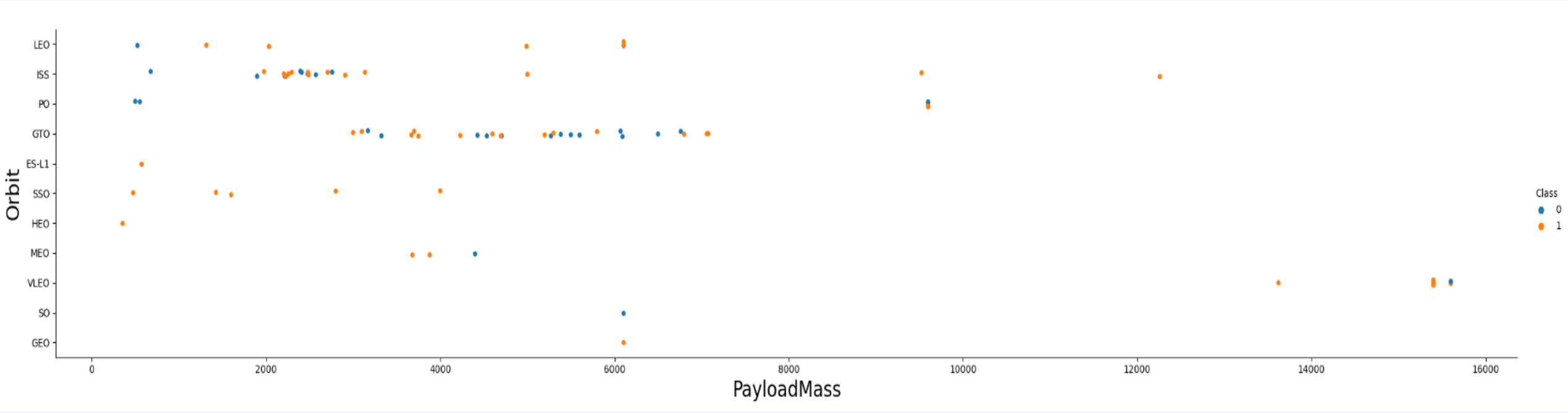
Flight Number vs. Orbit Type

- For most orbit, the successful launch rate increases with the number of flights
- The trend is most obvious with VLEO orbit, but not as obvious for the GTO orbit



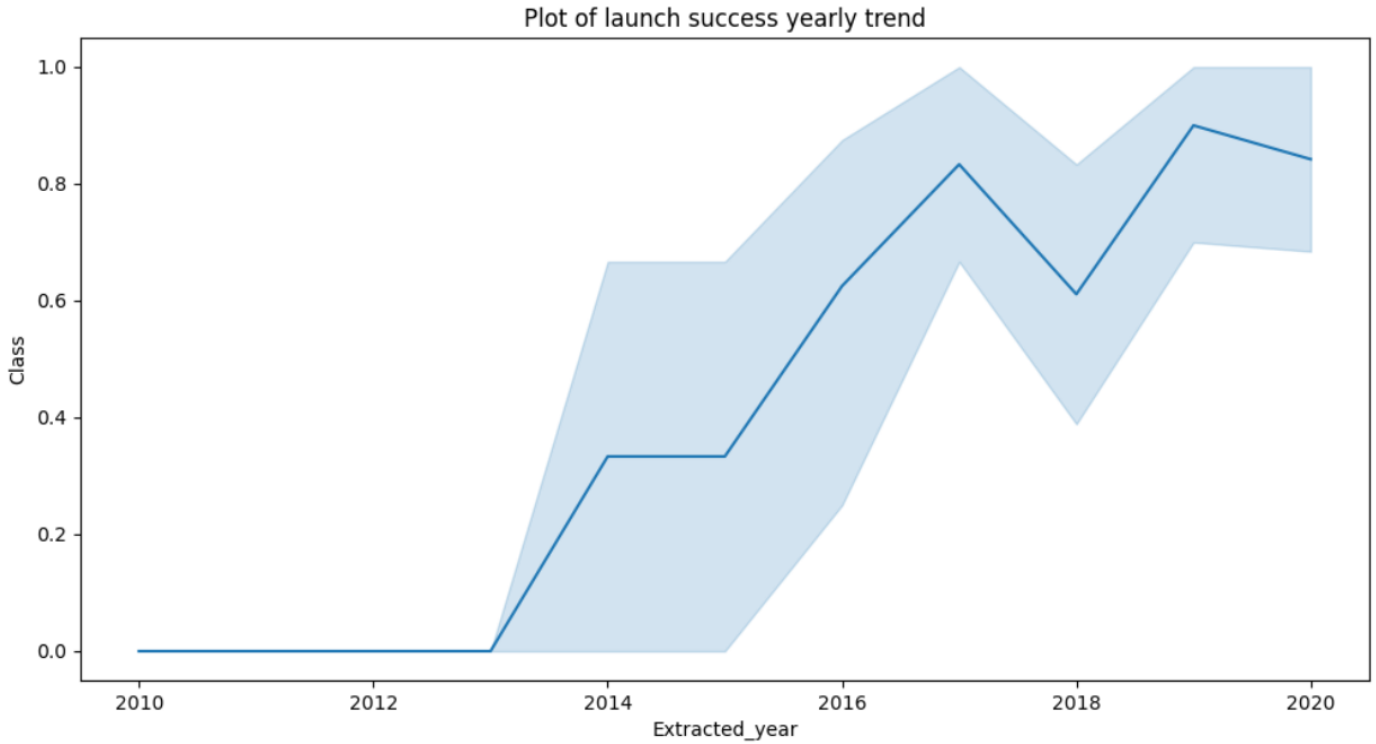
Payload vs. Orbit Type

- With increased payload mass, the success launch rate increases for LEO, ISS
- For GTO, the success launch rate does not notably change.



Launch Success Yearly Trend

- From 2010 to 2013, the success launch rate remains the same at 0%
- From 2013 to 2020, the overall trend is that the success launch rate increases.



All Launch Site Names

- The names of the launch sites are CCAFS LC-40, CCAFS SCL-40, KSC LC-39A, VAFB SLC-4E.
- Data obtained by selecting distinct launch sites from dataset.

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

Launch Site Names Begin with 'CCA'

- Five records of the names of the launch sites that 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 is 111,268 (kg);
- Data obtained by summing all payloads with code 'CRS'.

total_payload
111268

Average Payload Mass by F9 v1.1

- The average payload mass is 2,928 (kg);
- Data obtained by select the booster version 'F9 v1.1' and calculate the average payload mass.

avg_payload
2928

First Successful Ground Landing Date

- The first successful ground landing data is Dec. 22, 2015;
- Data obtained by selecting the minimum data from entries with the successful ground landing.

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The successful drone ship landing with payload between 4000 to 6000 include booster version F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, F9 FT B1026.
- Data obtained by selecting distinct values from dataset where landing outcome is success and payload is between 4000 to 6000

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

Total Number of Successful and Failure Mission Outcomes

- The total number of success is 99+1 and the number of failure is 1;
- Data obtained by counting the number of records for each grouped mission outcomes.

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

Boosters Carried Maximum Payload

- 12 booster versions that carried the maximum payload, e.g., F9 B5 B 1048.4, F9 B5 1060.3;
- Data obtained by selecting booster version entries from dataset that associate with the maximum payload.

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

- Two entries found for failed launches, booster version F9 v1.1 B1012 and F9 v1.1 B1015 at launch site CCAFS LC-40;
- Data obtained by selecting the booster version and launch site from failed launching outcomes.

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

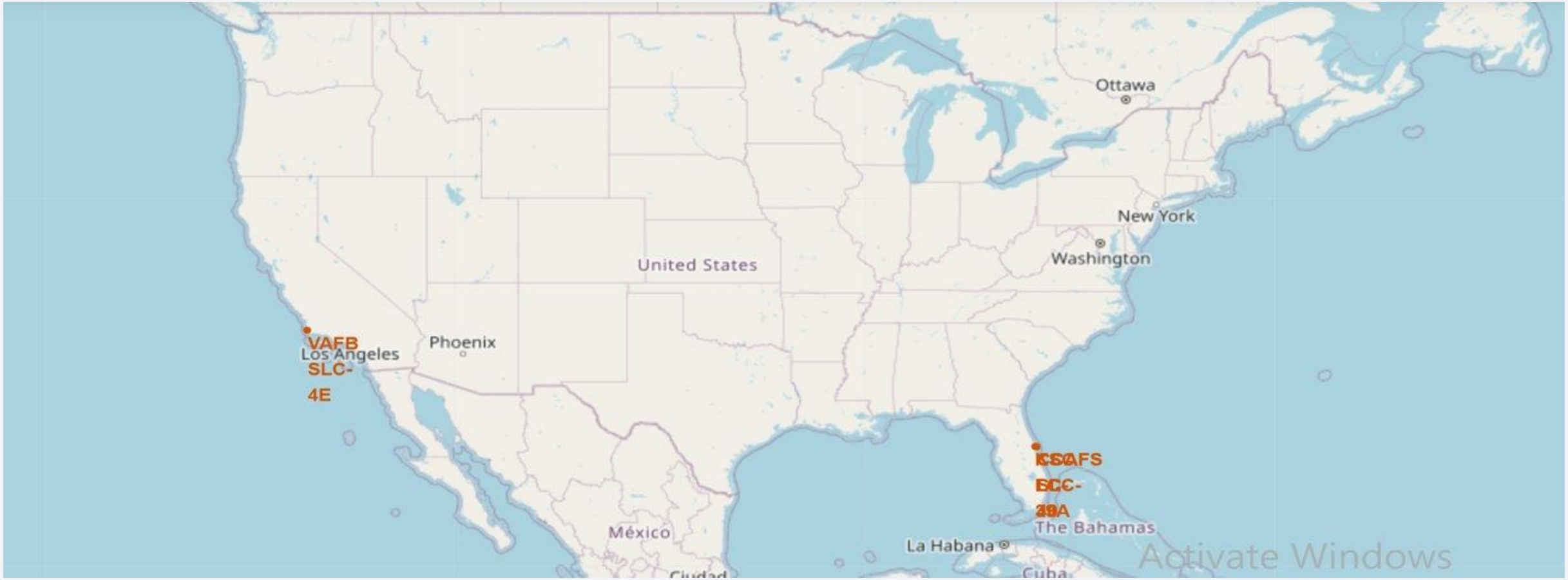
- The landing outcomes are ranked for 8 landing outcomes;
- Data obtained by selecting and counting landing outcomes between 2010-06-04 to 2017-03-20 and list them in descending order based on quantity.

landing__outcome	qty
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

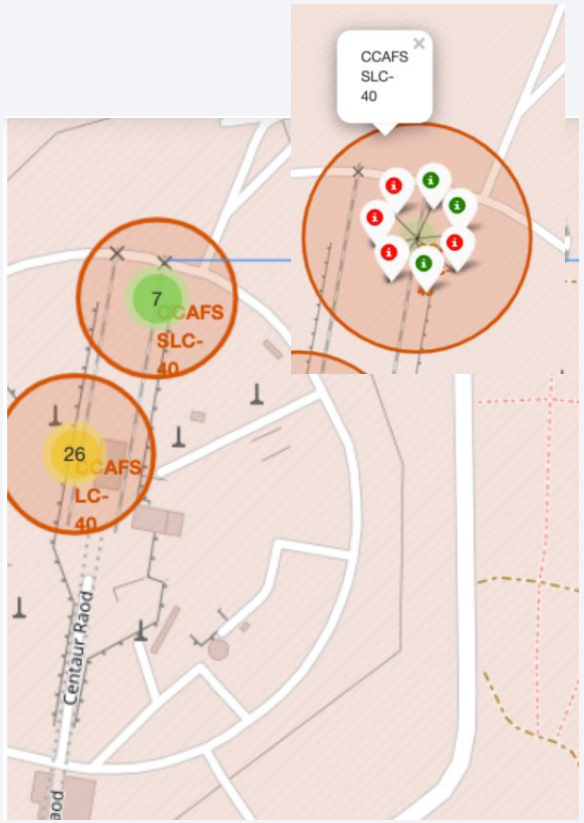
All Launch Sites

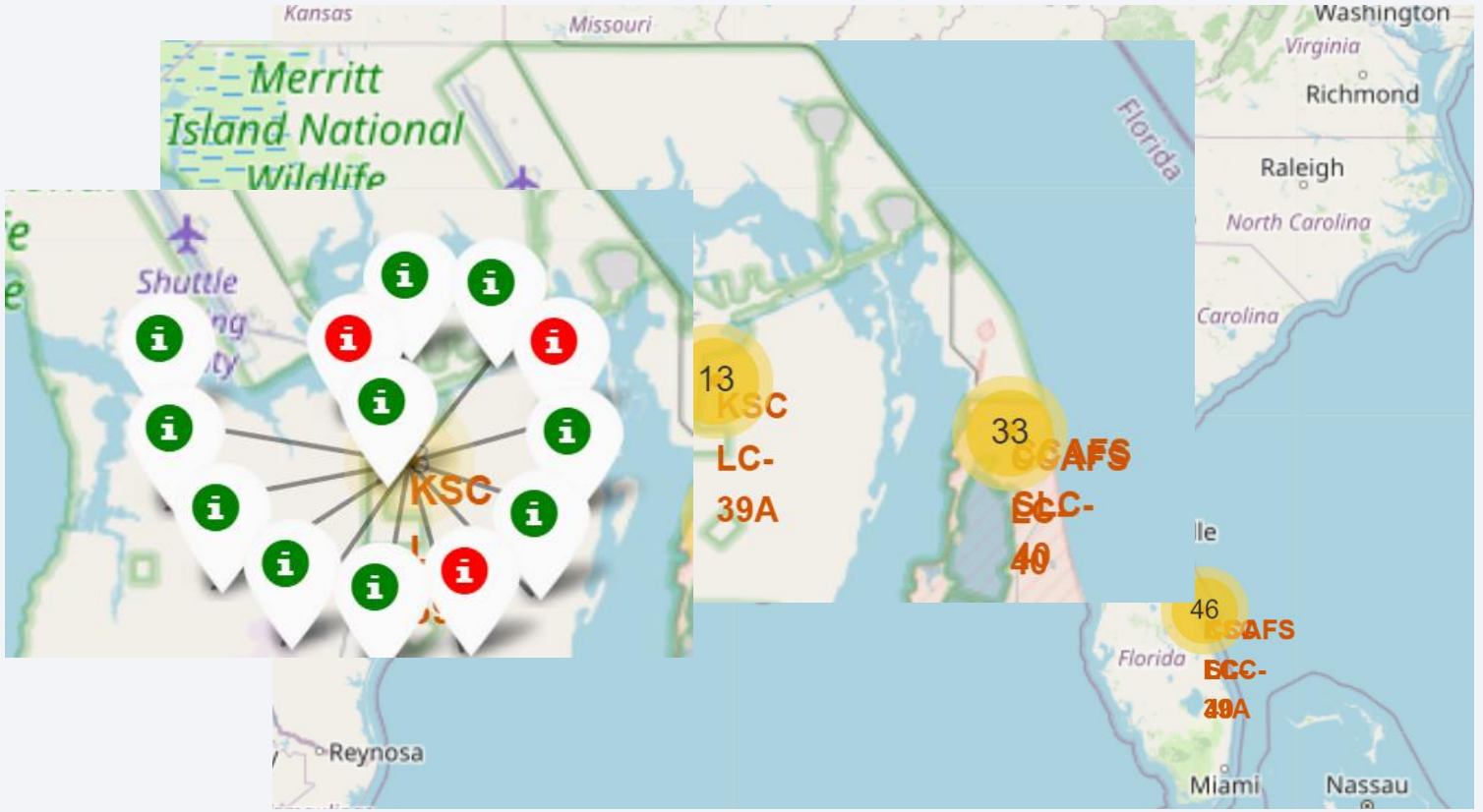
- Launch site of the US are located along the west coast of California and east coast of Florida



Launch Sites with Colored Labels

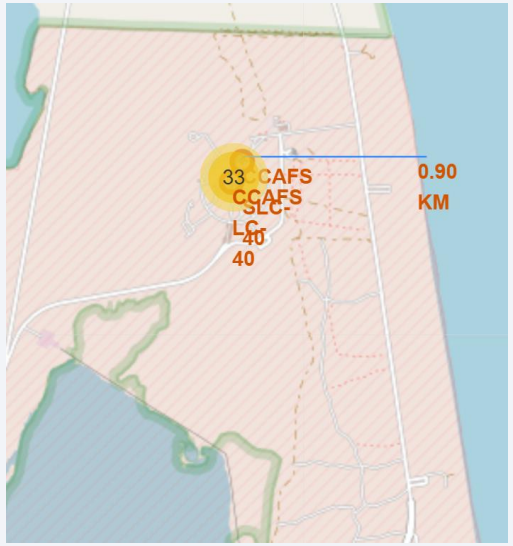
- Below are the zoomed graphs of California and Florida launch results
- Success launches are shown by green marker and failed launches are in red

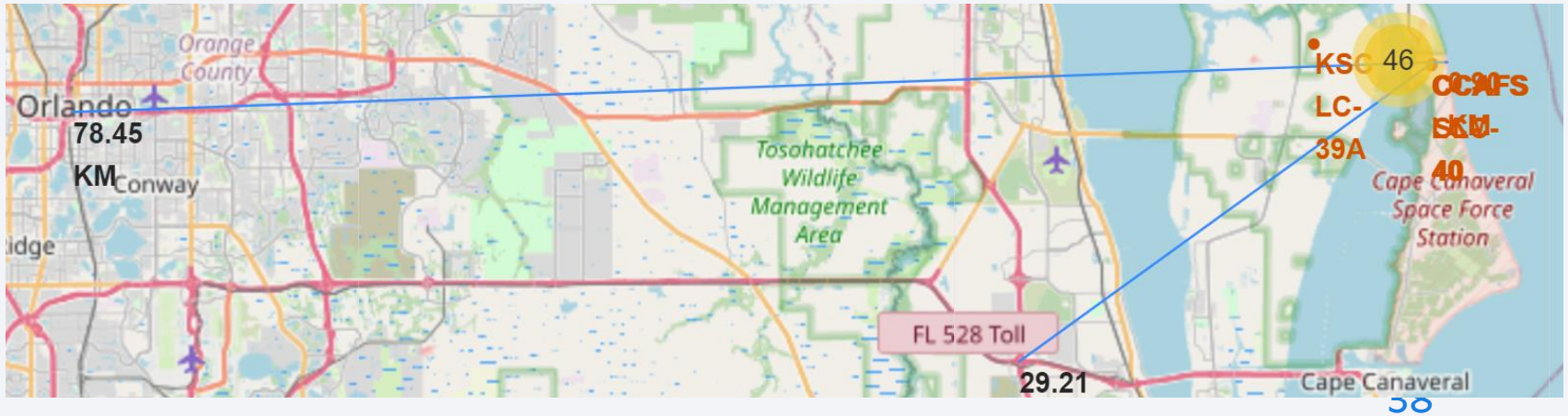


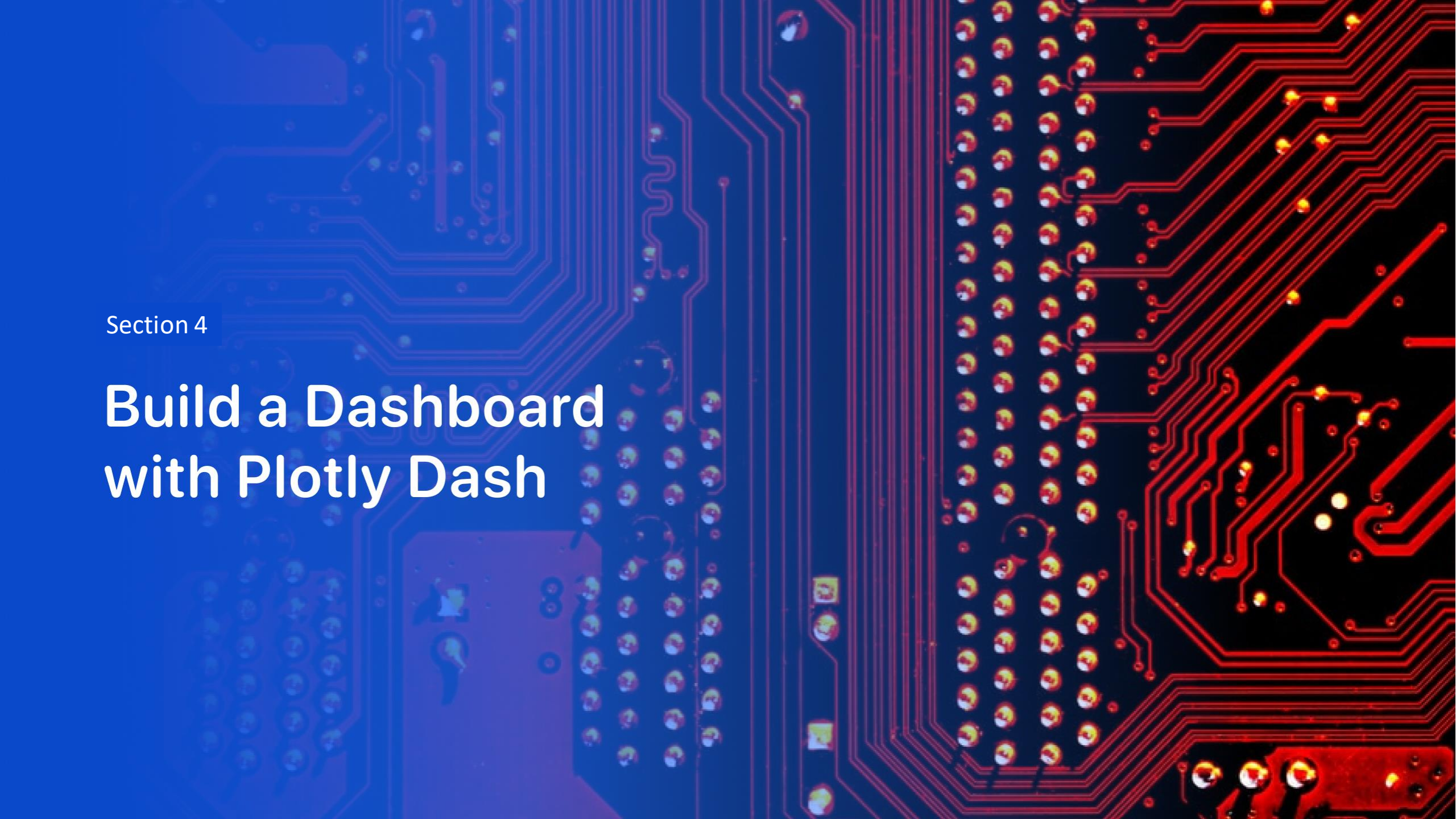


Launch Site Safety Concern

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

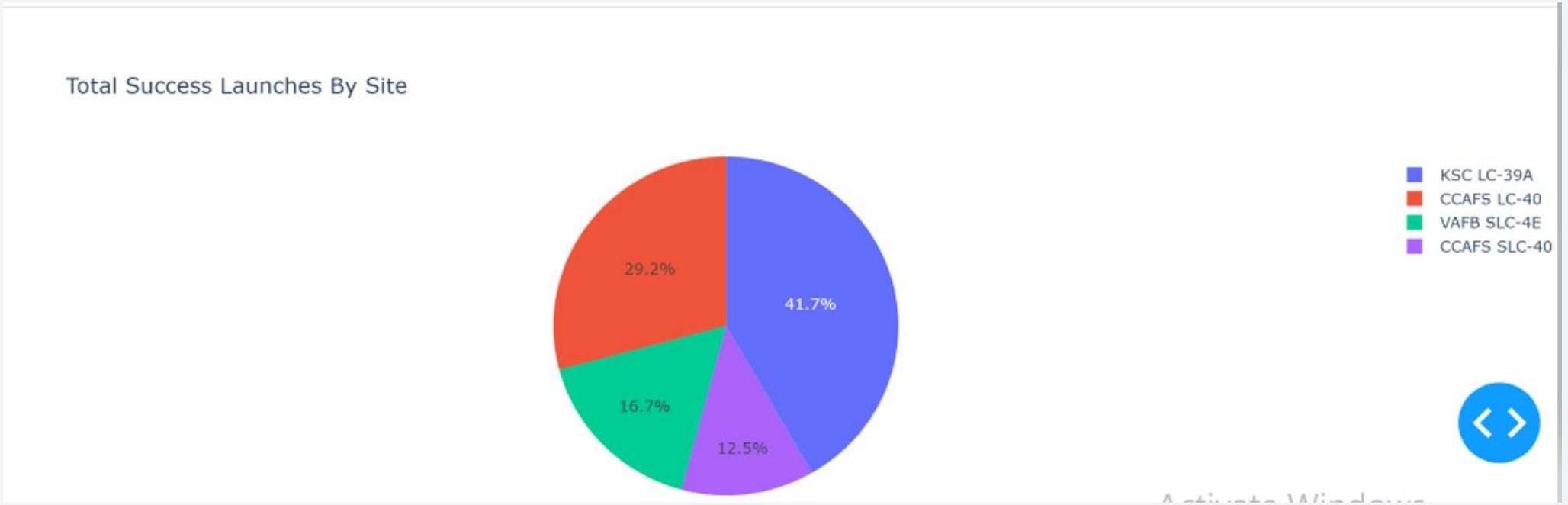





Section 4
Build a Dashboard with Plotly Dash

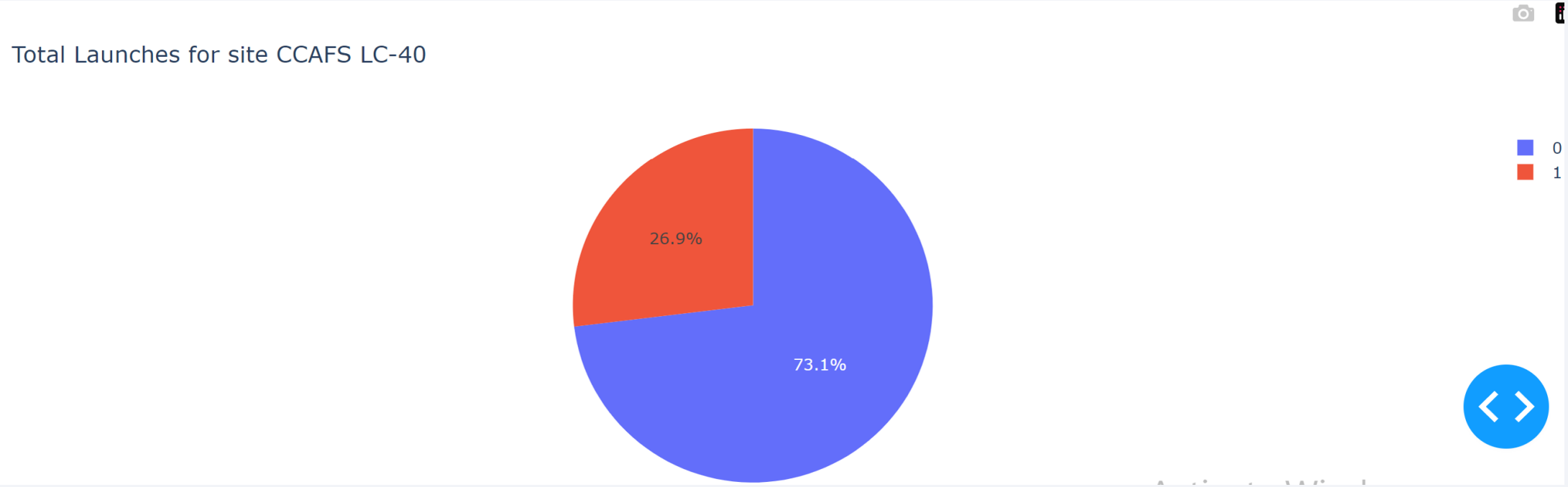
Successful Launch Ratio

- The site with highest launch ratio is KSC LC-39A
- The success of launch is seen to be highly correlated with the Site.



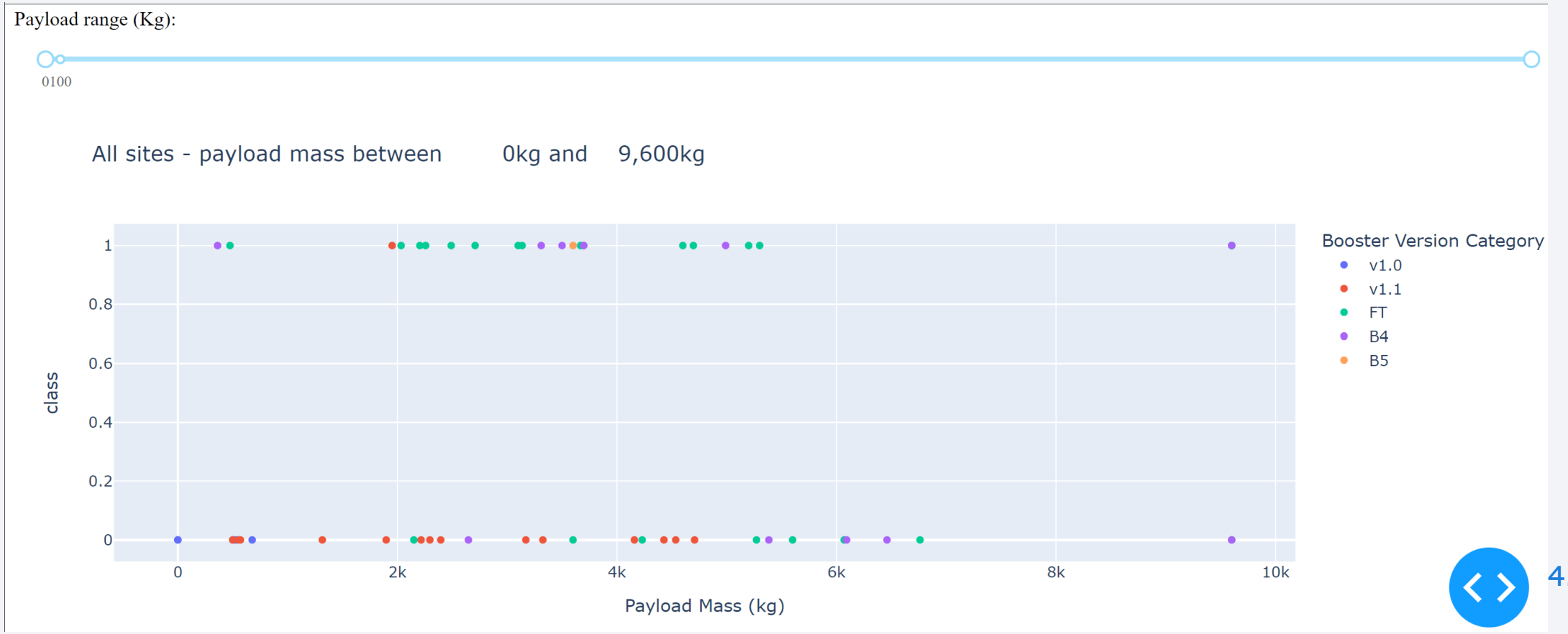
Successful Launch Ratio of CCAF LC-40

- Chart showing the ratio of successful and failed launches of the site CCAF LC-40



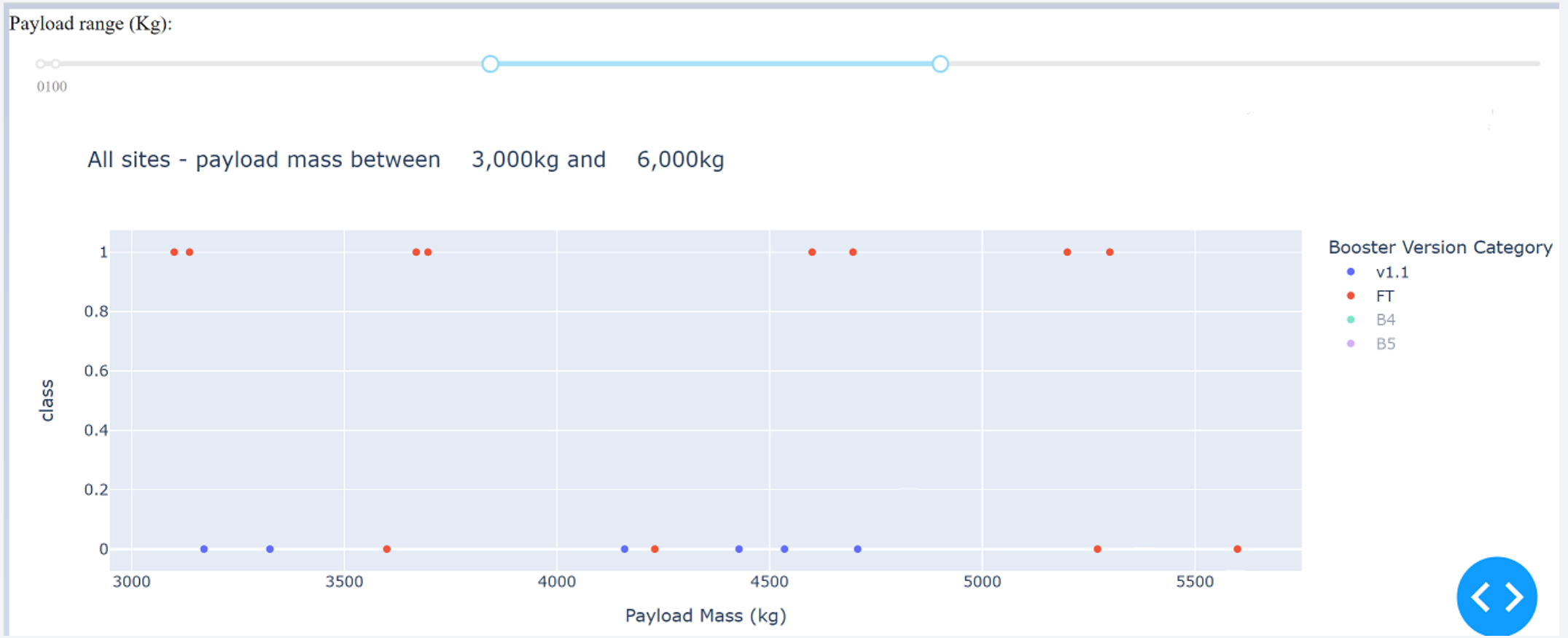
Launch Success vs Payload

- Success ratio increases for B4 for increased payload
- Success ratio decreases for FT for increased payload



Launch Success vs Payload

- Between 3000kg to 6000kg, the successful launch rate for FT increases



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and the most accurate model is Decision Tree Classifier, with the accuracy of 88.75%

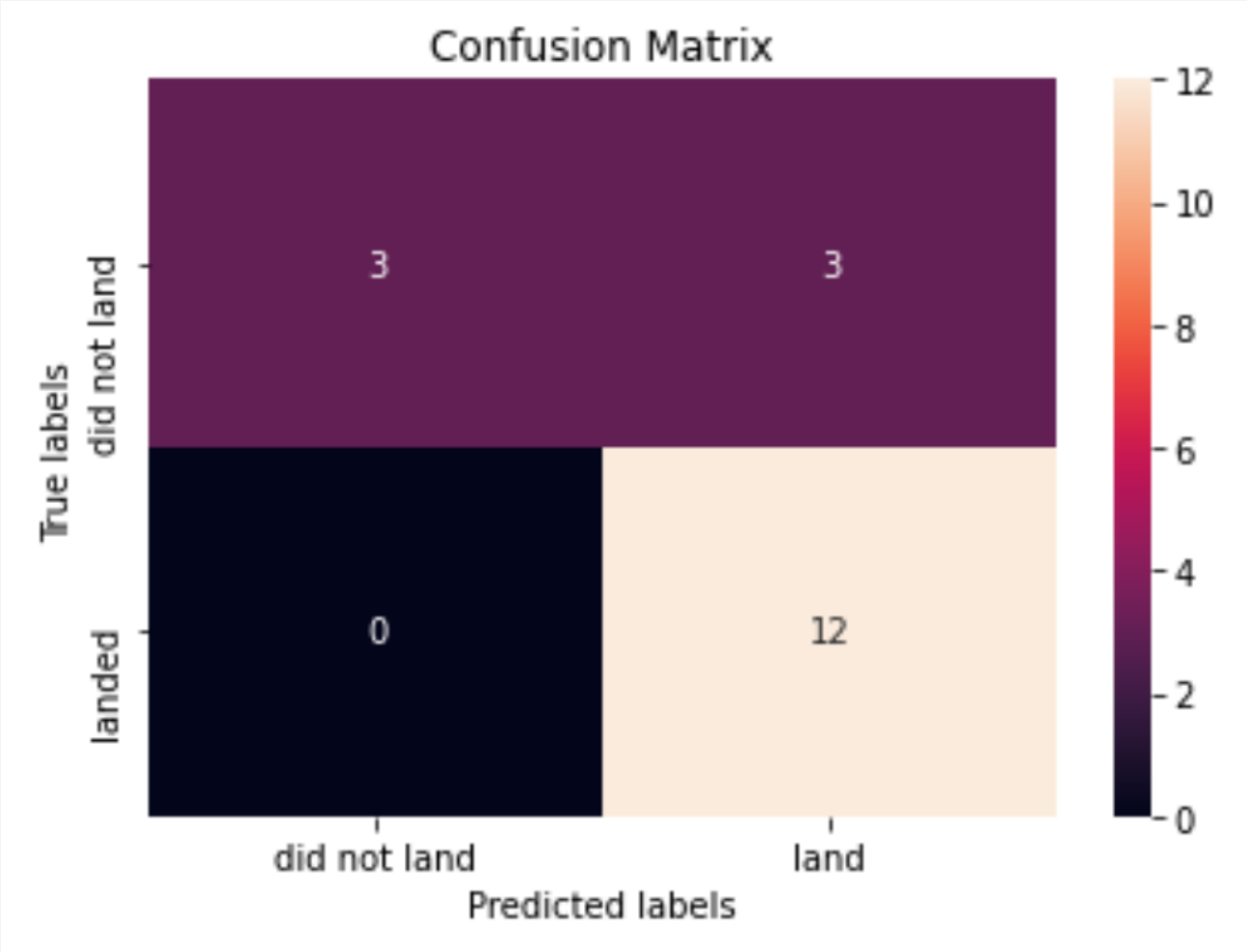
[45]:
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8875
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
```

Confusion Matrix of Decision Tree Classifier

- The number of true positive and true negative shown by the confusion matrix of Decision Tree Classifier is high;
- While there are cases where those successful landed are marked as did not land.



Conclusions

- The ratio of successful launches increases over time between 2013 and 2020, likely due to the improvement of rockets;
- Most of the launch sites are along the west and east coasts, far away from residential areas;
- The ratio of successful launches increases in general with payload;
- The ratio of successful launches increases with the flight amount for each launch site;
- The launch site that has the highest amount of successful launches is KSC LC-39A;
- The ES-L1, GEO, HEO, SSO have the highest ratio of successful launches;
- Decision Tree classifier predicts the successful and unsuccessful launches most accurately.

Thank you!