IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robel Equbasilassie Kahsay
August - 2022

Outline

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

Executive Summary

- The data for this analysis are gathered by making a request to the **SpaceX API** and by performing **Web scrapping** on the Wikipedia page to collect and extract Falcon 9 historical launch records.

- Various data wrangling techniques are applied to clean and standardize the data and Exploratory Data Analysis (EDA) is performed to find patterns in the data and determine the best features for training supervised models.

- Visualizations such as scatter plots, bar charts, and maps are used to analyze the relationship between the dependent and independent variables and select the important features that will be used in prediction.

- For predictive analysis models such as Logistic Regression and SVM, DecisionTree classifier, KNeighbors classifier are selected based on their accuracy results.

- KSC LC-39A launch site has the highest success rate of landing and it seems that the more massive the payload, the less likely the first stage will return.

- Interactive dashboards are built to visualize the relation ship of each launch sites with flight number, payload mass, orbit type and success rate.

Introduction

- Space Exploration Technologies Corp. (doing business as SpaceX) is an American spacecraft manufacturer, space launch provider, and a satellite communications corporation headquartered in Hawthorne, California.

- SpaceX was founded in 2002 by **Elon Musk**, with the goal of reducing space transportation costs to enable the colonization of Mars.

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- The goal of this data science capstone project is to **predict if the Falcon 9 first stage will land successfully**. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By making a request to the **SpaceX API** and by performing **Web Scrapping** on the Wikipedia page titled List of Falcon 9 and Falcon Heavy launches to collect and extract Falcon 9 historical launch records.
- Perform data wrangling
  - By Identifying missing data, counting the landing outcomes to rewrite it as classes of 0 and 1 and success rate of Falcon 9 is calculated.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictor features selected from independent variables, data is standardized/scaled, train and test data split is done, best hyperparameter for classification models are calculated by tuning parameters then models are trained.

Data Collection

Data collection is done using two processes

- 1. Making a request to the SpaceX API
  - The following URLs are used to call the SpaceX API using the **requests** library and parse the JSON results to extract information using identification numbers in the launch data;

`spacex_url` = https://api.spacexdata.com/v4/launches/past

`static_json_url` = https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

- 2. Perform Web scrapping on the Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
  - The Falcon 9 launch records are extracted from Wikipedia by parsing the HTML table into a Pandas data frame using Beautifulsoup4 and requests libraries from the following URL;

`static_url = https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922`

Data Collection – SpaceX API

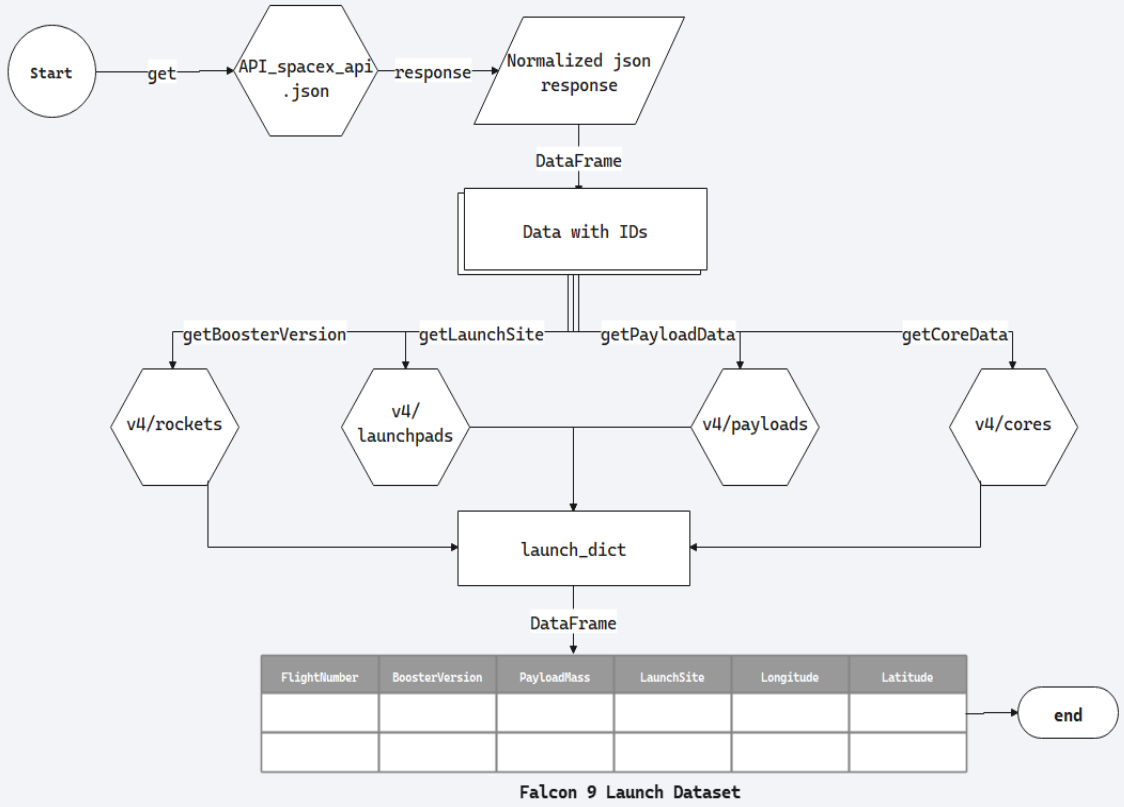
- Request and parse the SpaceX launch data using the GET request

```
response = requests.get("https://api.spacexdata.com/v4/rockets/...").json()

response = requests.get("https://api.spacexdata.com/v4/launchpads/...").json()

response = requests.get("https://api.spacexdata.com/v4/payloads/...").json()

response = requests.get("https://api.spacexdata.com/v4/cores/...").json()

response = requests.get("...")
```

- FlightNumber, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Longitude, Latitude data are extracted from the SpaceX API into a Panda data frame.

- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/01_DataCollection/jupyter-labs-spacex-data-collection-api.ipynb

Flowchart of SpaceX API calls



Data Collection - Scraping

- Request and extracted the Wikipedia Falcon 9 launch records using Beautifulsoup4 and requests libraries

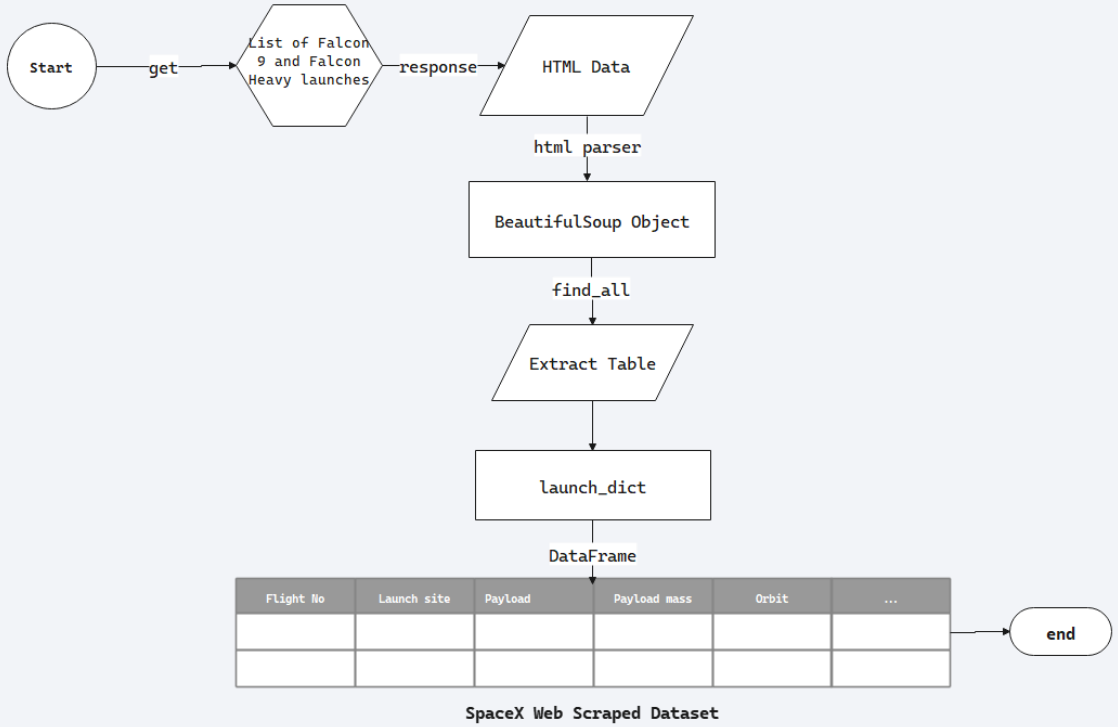
```
response = requests.get("https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922")
soup = BeautifulSoup(response.text, "html.parser")
soup.find_all('table',"wikitable plainrowheaders collapsible")
```

- Falcon 9 launch records are extracted from Wikipedia table rows.

- Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing data are parsed from the Wikipedia table into a Panda data frame.

- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/01_DataCollection/jupyter-labs-webscraping.ipynb

Flowchart of web scraping



Data Wrangling

- Filter the dataframe to only include Falcon 9 launches using the **BoosterVersion** column
  - `data_falcon9 = data[data['BoosterVersion'] == 'Falcon 9']`
- Check for missing values in our dataset.
  - `data_falcon9.isnull().sum()`
  - Identified missing values for **PayloadMass** is **5** and **LandingPad** is **26**
- Dealing with Missing Values
  - Mean values are calculated for the PayloadMass column's payload mass, and each missing value is replaced by the mean values.
  - Calculate the mean value of PayloadMass column

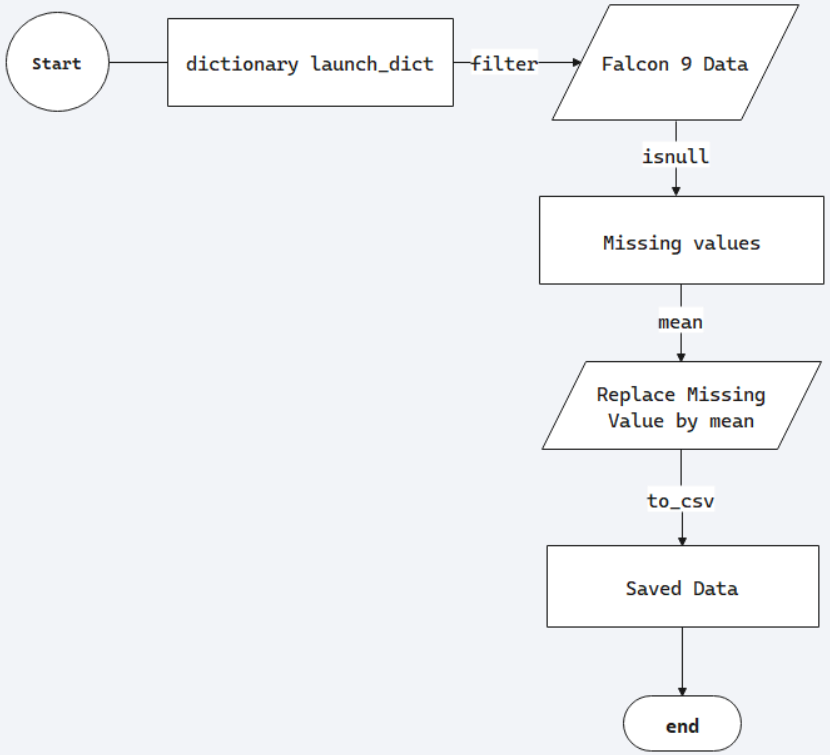
    ```
    mean_PayloadMass = data_falcon9['PayloadMass'].mean()
    ```

  - Replace the np.nan values with its mean value

    ```
    data_falcon9['PayloadMass'].fillna(mean_PayloadMass, inplace=True)
    ```

- **GitHub URL:** https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/02_DataWrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

Flowchart of Data Wrangling



EDA with Data Visualization

- Plots such as **scatter plots** and **bar charts** are used to visualize the relationship between parameters.

Parameters	Chart Type
FlightNumber vs PayloadMass	Scatter Plot
FlightNumber vs LaunchSite	Scatter Plot
PayloadMass vs LaunchSite	Scatter Plot
FlightNumber vs Orbit	Scatter Plot
PayloadMass vs Orbit	Scatter Plot
Orbit vs Success Rate	Bar chart

- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/03_EDA/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- 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 acheived
- List the names of the boosters which have success in drone ship and have 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. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank 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
- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/03_EDA/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Folium objects with their importance are listed below

Map object	Description
Map	To create a base map by passing starting coordinates to Folium
Circle	To clearly visualize the areas centered at the coordinates
Marker	To put location marker with a popup and tooltip HTML
PolyLine	To draw a line between coordinates
MarkerCluster	To simplify a map containing many markers having the same coordinate

- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/04_DataVisualization/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart is used to show the total successful launches count for all sites.
- For a particular launch site, the pie chart is used to display the Success vs. Failed counts.
- Scatter chart is used to show the correlation between payload and launch success
- Dropdown list is used to enable Launch Site selection and interact with the dashboard easily.
- Range slider is used to select payload range and see the success rate of each launch site along with payload mass.

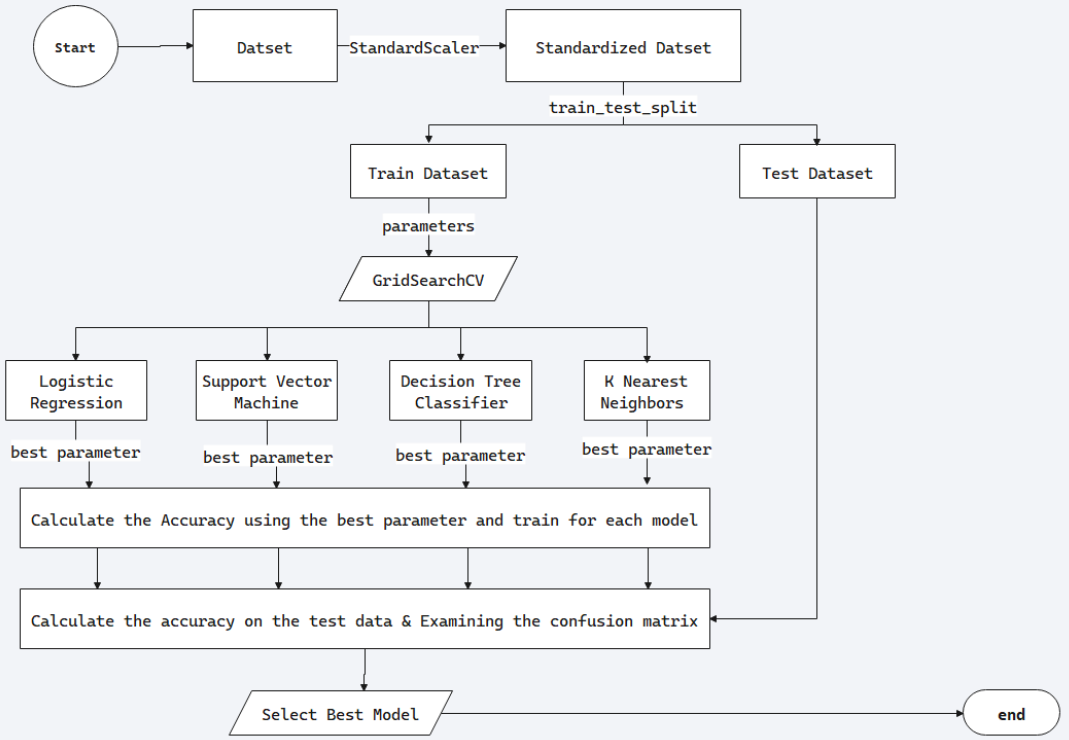
- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/04_DataVisualization/spacex_dash_app.py

Predictive Analysis (Classification)

- Significant predictor features have been selected and the dataset has been standardized to bring all the feature data to the same scale.
- The data is split into training and testing data.
- **Logistic Regression**, **Support Vector Machine**, **Decision Tree Classifier**, and **K Nearest Neighbors** are selected to predict the success class.
- Using the GridSearchCV function hyperparameters are selected and the models are trained.
- By comparing the train test accuracies of each model best model for the data set is identified.
- During the training time SVM and Logistic Regression has best performance

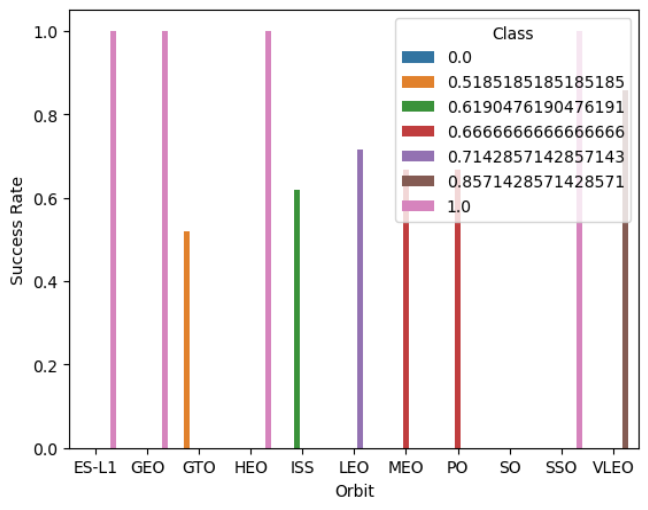
- **GitHub URL**: https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project/blob/main/05_ML_Modeling/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

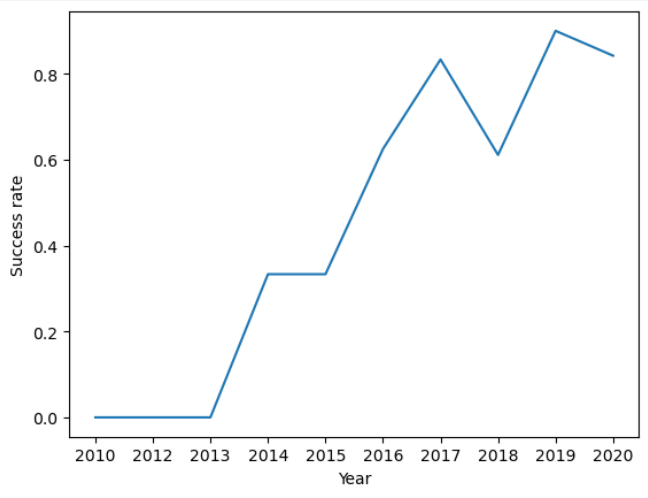
Flowchart of Modeling

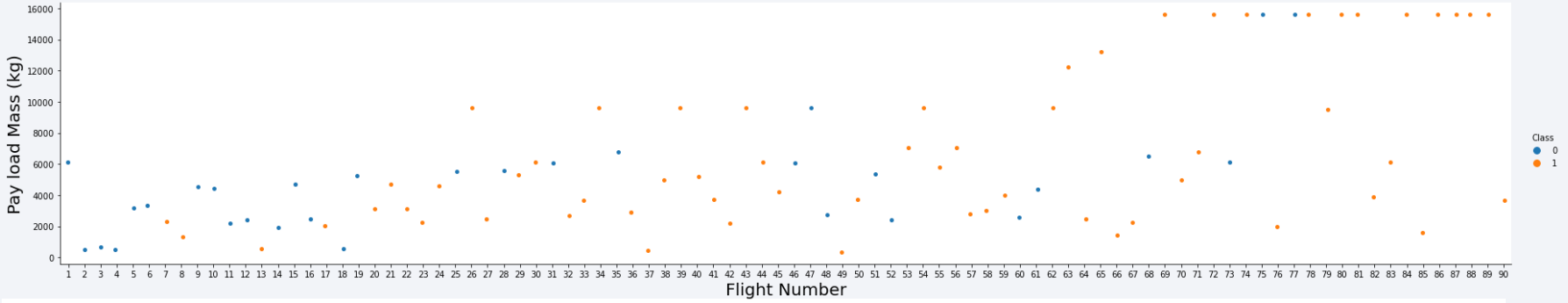


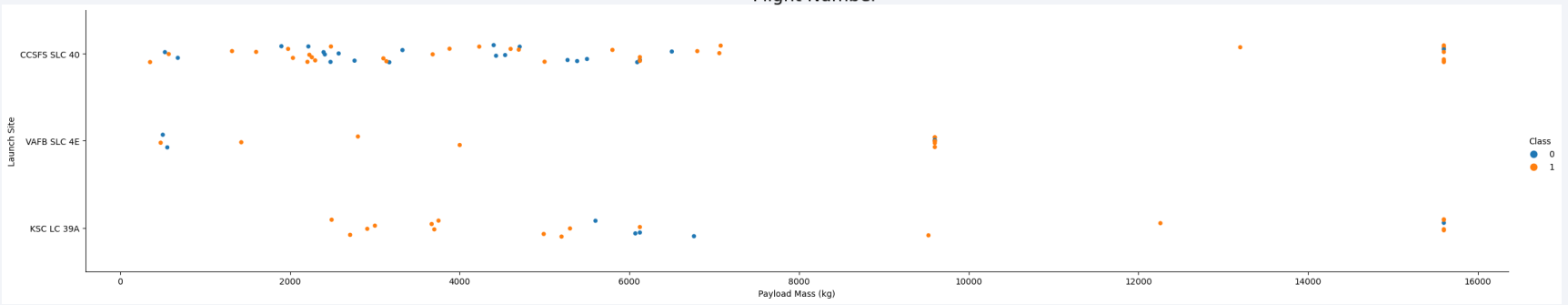
Results

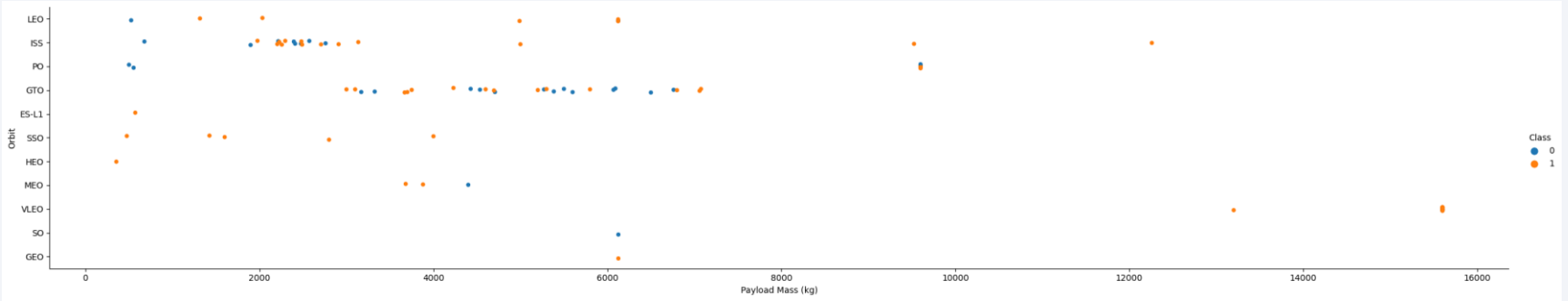
- Exploratory data analysis results





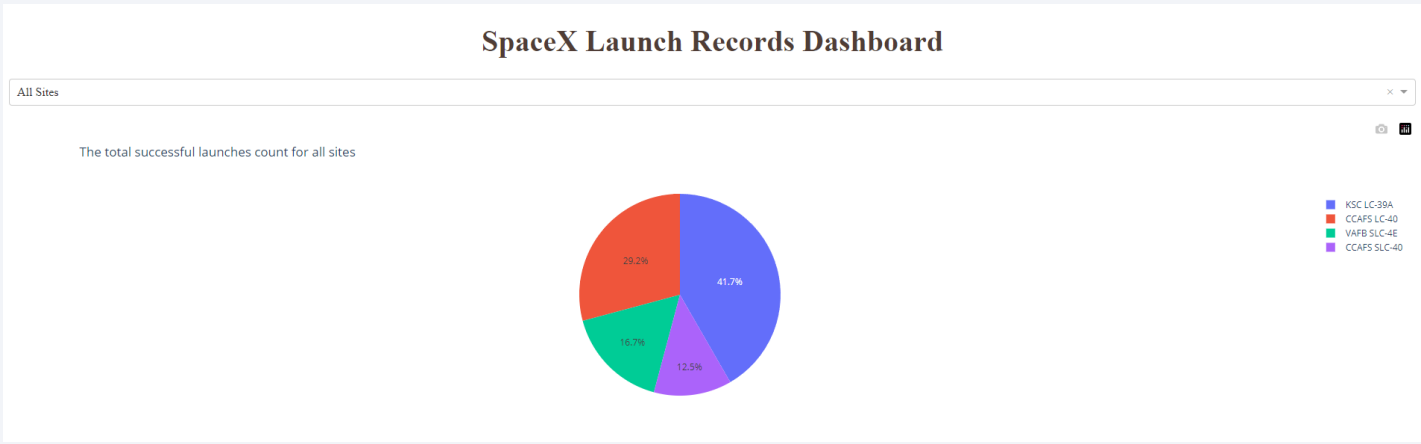


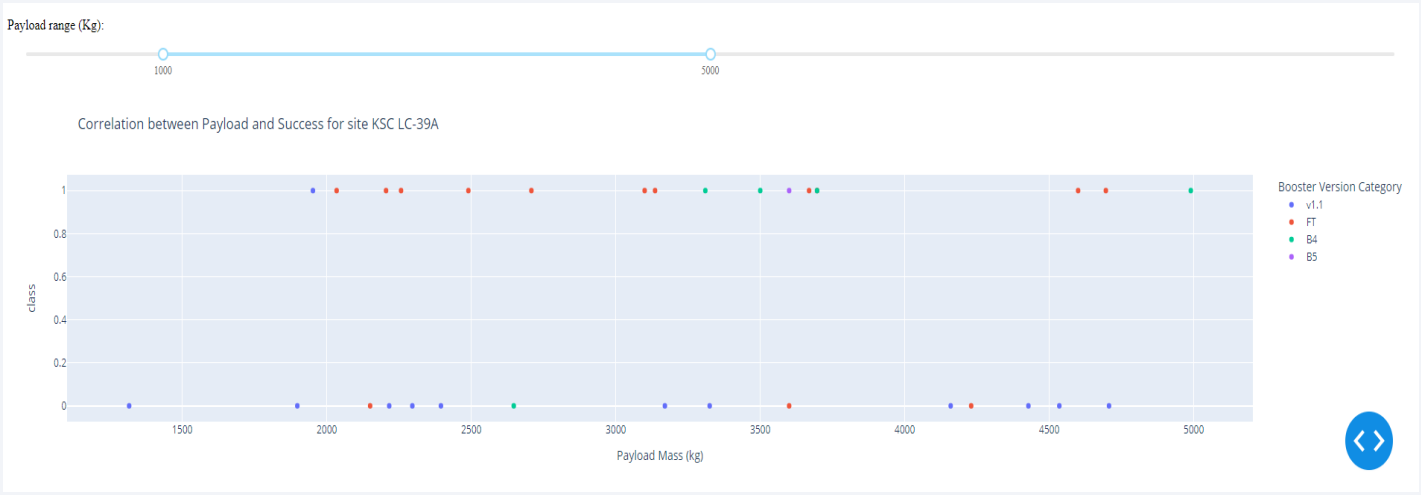




Results

- Interactive analytics demo in screenshots





Results

- Predictive analysis results

```
GridSearchCV(cv=10, estimator=KNeighborsClassifier(),
             param_grid={'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
                         'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
                         'p': [1, 2]},
             scoring='accuracy')
```

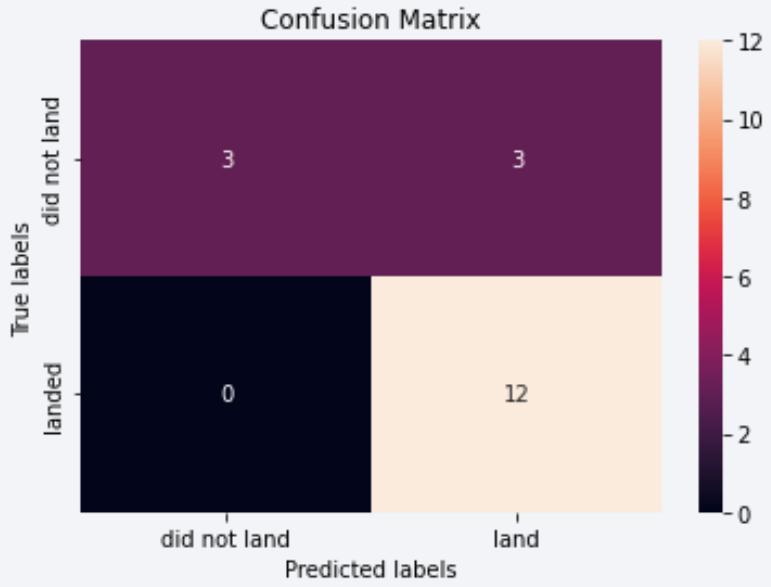
GridSearchCV

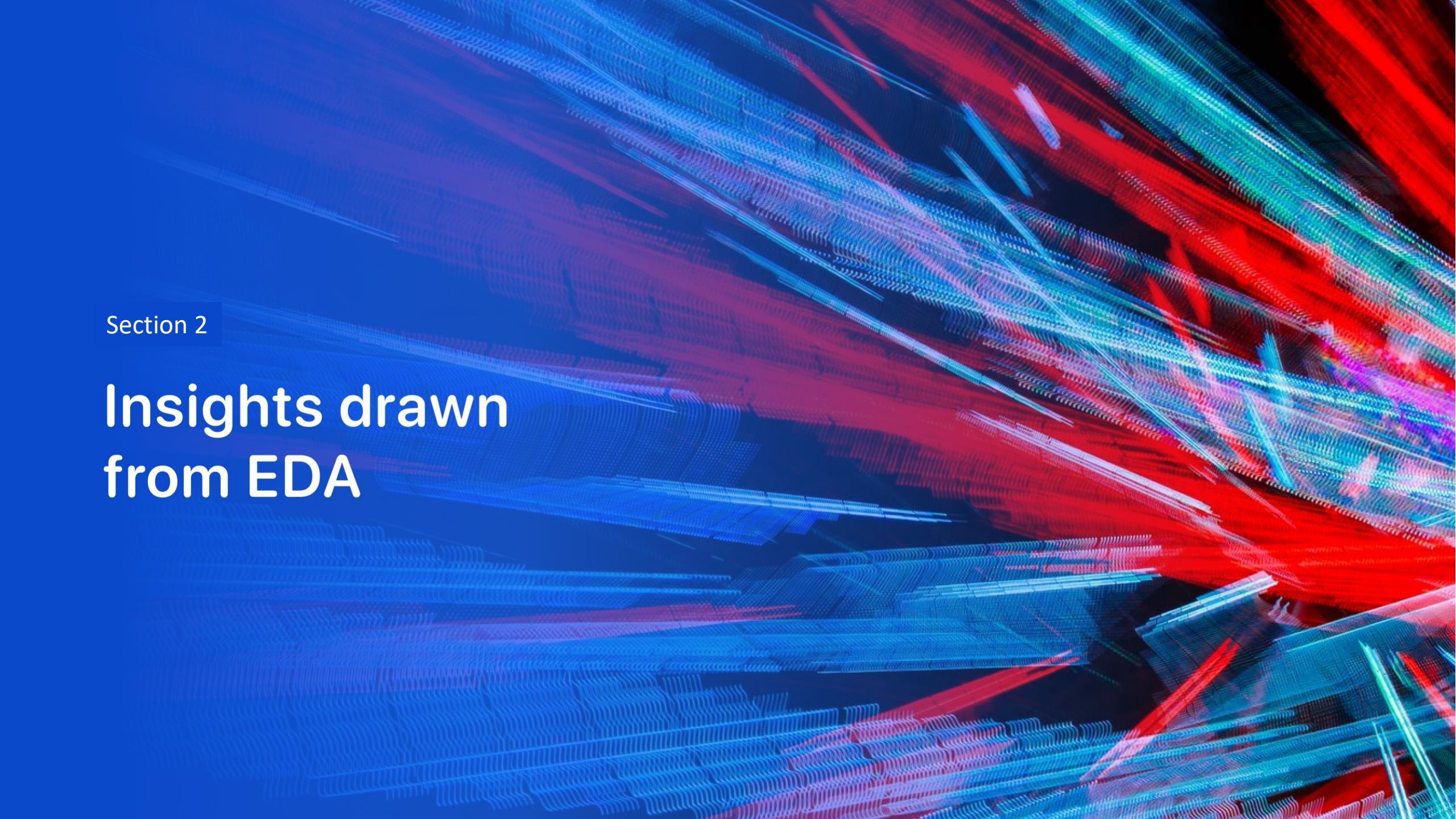
estimator: KNeighborsClassifier

KNeighborsClassifier

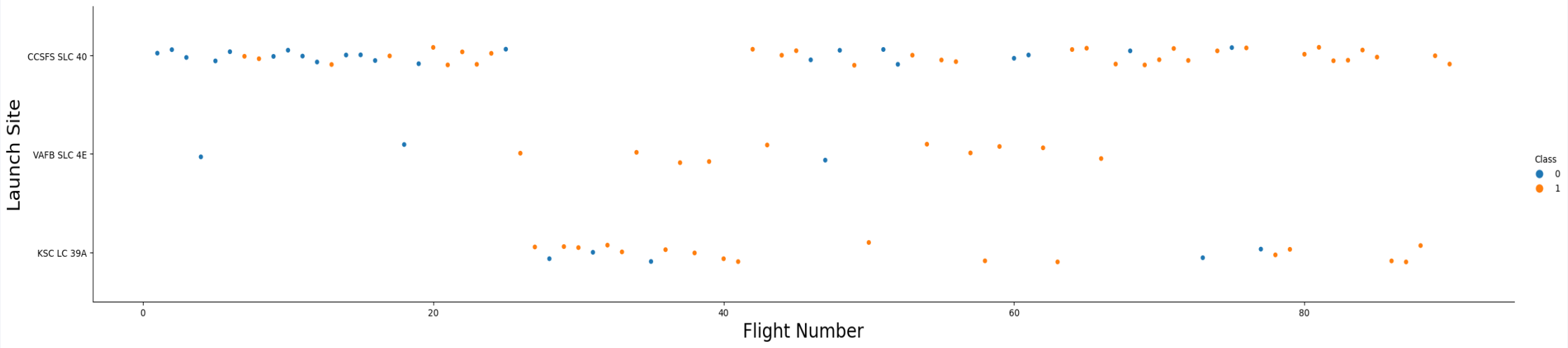
KNeighborsClassifier()

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```




Section 2
Insights drawn from EDA

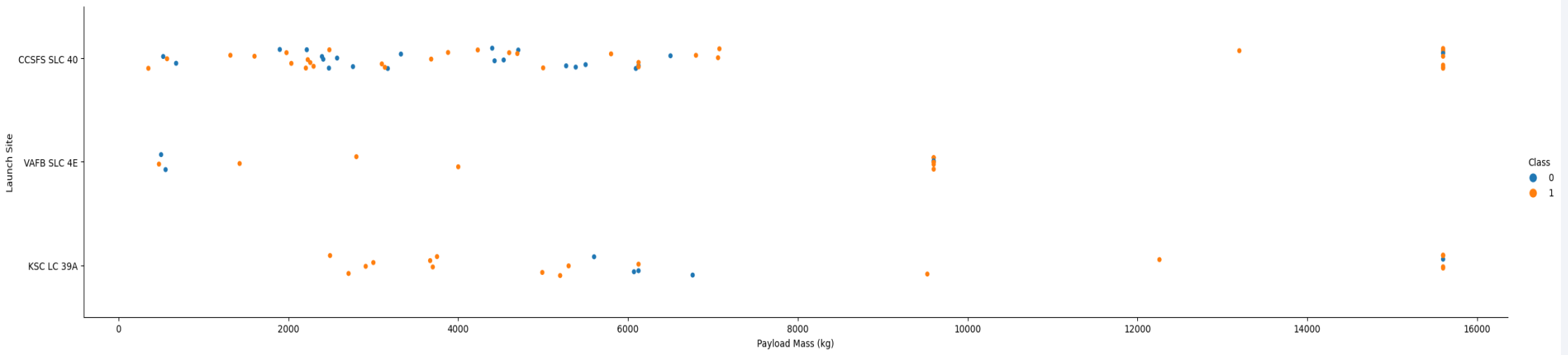
Flight Number vs. Launch Site



- With 26 launches, CCAFS LC-40 launch site has the largest number of launches.

- With 10 successful launches, KSC LC-39A launch site has the largest successful launches.

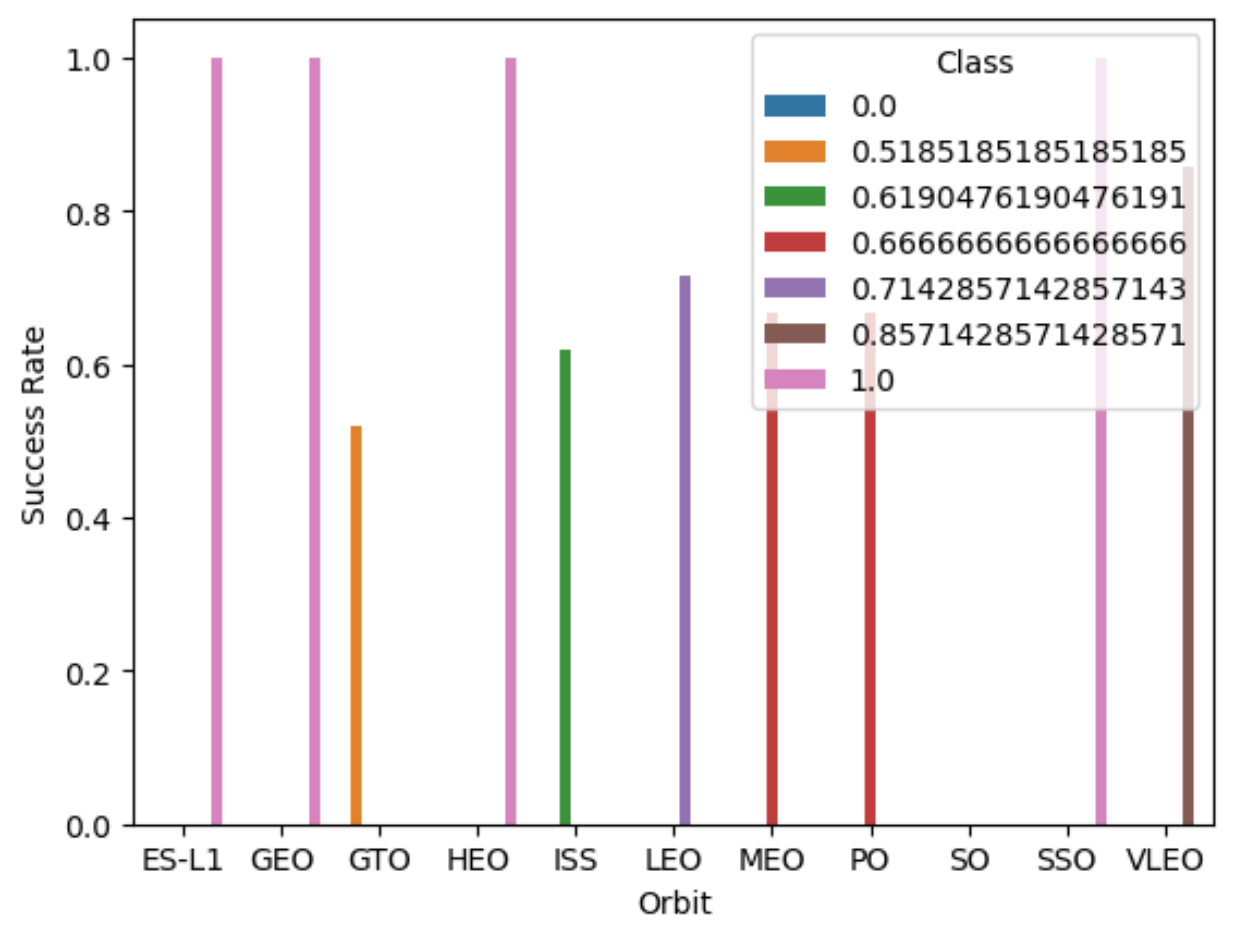
Payload vs. Launch Site



- There are no launches at the VAFB-SLC 4E launch site with heavy payload masses (greater than 10,000).

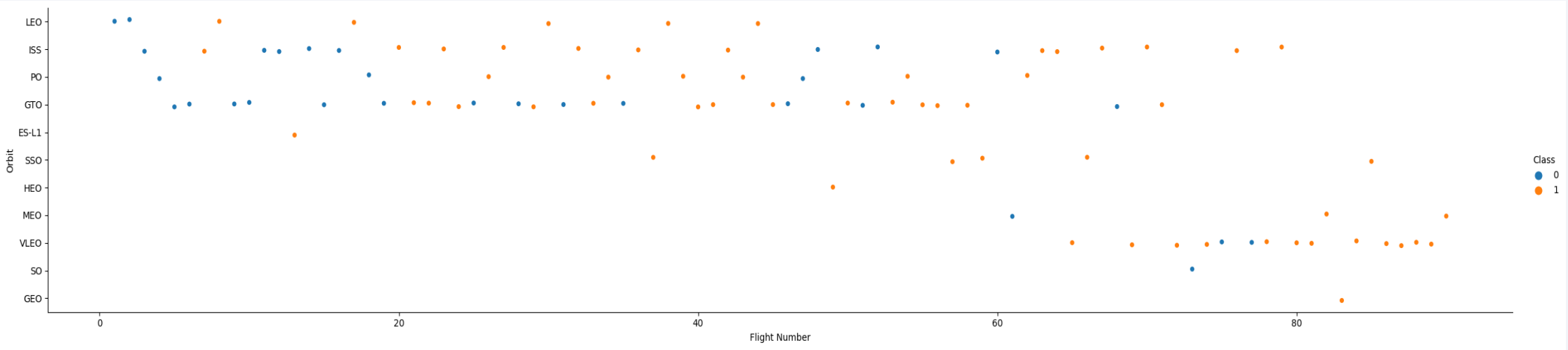
- Launch stations CCAFS SLC 40 and KSC LC 39A have significantly higher success rates for rockets with increased payload mass (greater than 10,000)

Success Rate vs. Orbit Type



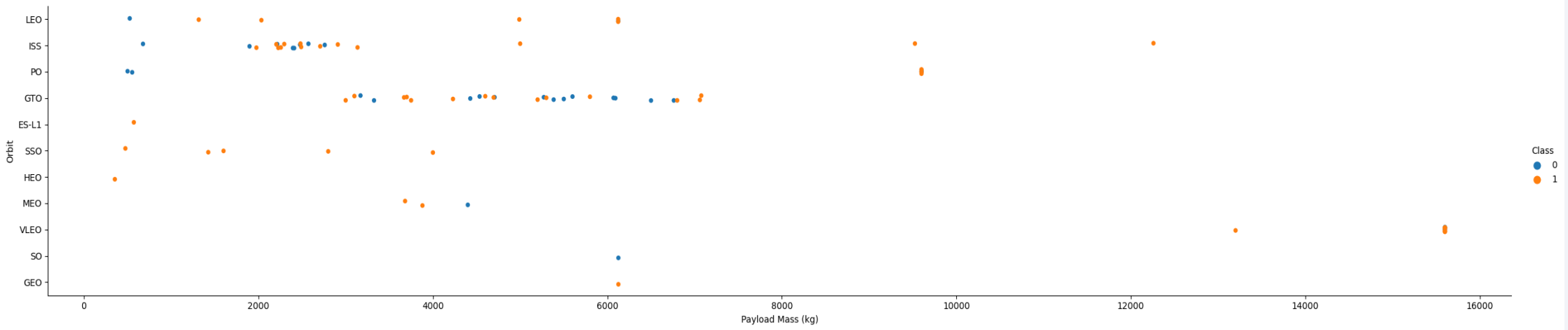
- When compared to other orbits, the ES-LQ, GEO, HEO, and SSO orbits have a higher success rate.

Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights
- VLEO orbits the Success appears to be higher when the number of flights increase.
- There seems to be no relationship between flight number when in GTO orbit.

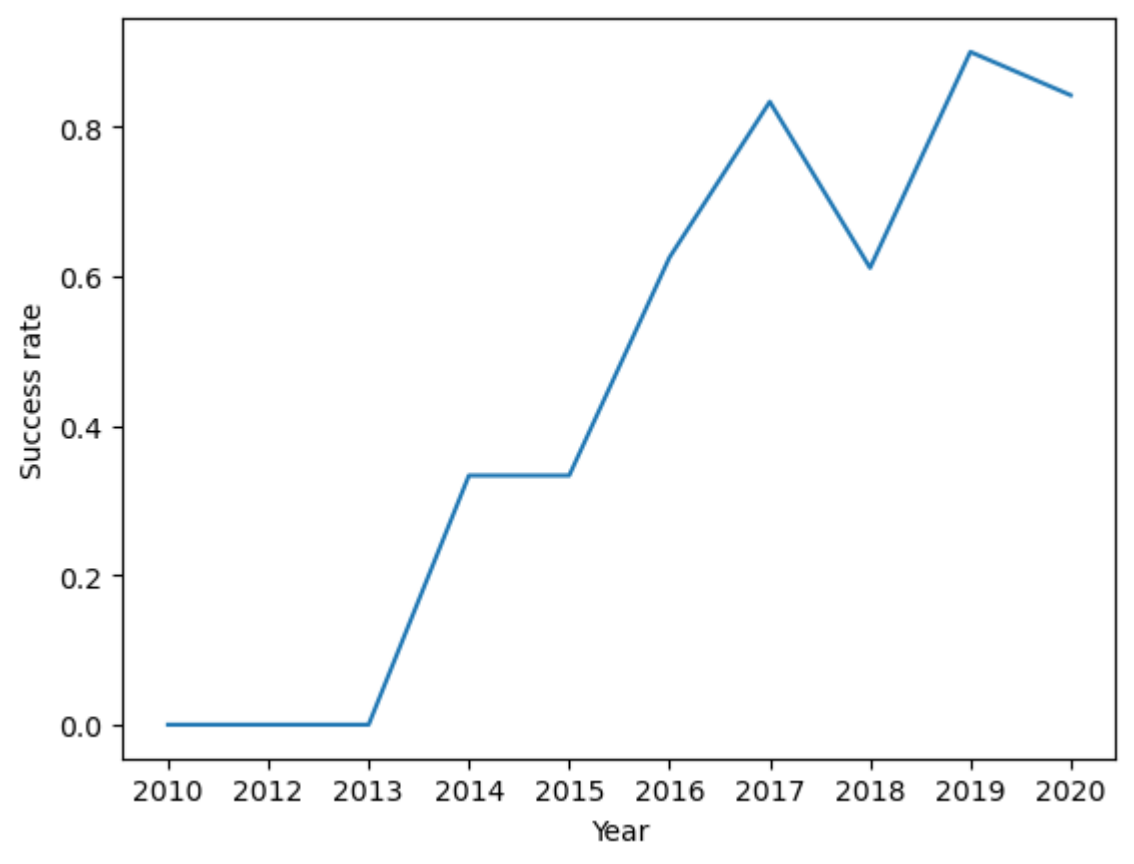
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

Launch Success Yearly Trend



- It can be see that since 2013, the success rate has been rising until 2020.

All Launch Site Names

- The names of the unique launch sites are
  - CCAFS LC-40
  - SC LC-39A
  - VAFB SLC-4E
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- There are only two records where launch sites name begin with `CCA`
  - CCAFS SLC-40
  - CCAFS LC-40

```
%%sql

SELECT distinct(Launch_Site) FROM SPACEXTBL
    WHERE Launch_Site LIKE 'CCA%';
```

✓ 0.1s

```
 * sqlite:///spacex_db.sqlite3
Done.
```

Launch_Site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS) is **45,596 Kg**
- It is calculated by adding the payload mass carried by boosters launched by NASA (CRS)


```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass FROM SPACEXTBL
    WHERE Customer = "NASA (CRS)";
```

```
 * sqlite:///spacex_db.sqlite3
Done.

 Total_Payload_Mass
              45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is **2,928.40 Kg**

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass_Kg FROM SPACEXTBL
    WHERE Booster_Version = "F9 v1.1";
```

```
 * sqlite:///spacex_db.sqlite3
Done.

 Average_Payload_Mass_Kg
                  2928.4
```

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved in **2015-12-22**

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

Total Number of Successful and Failure Mission Outcomes

- Successful mission outcomes are **100**
- Failure mission outcome is **1**

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

  - 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 are listed below

Year	Booster Version	Launch Site	Landing Outcome
2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

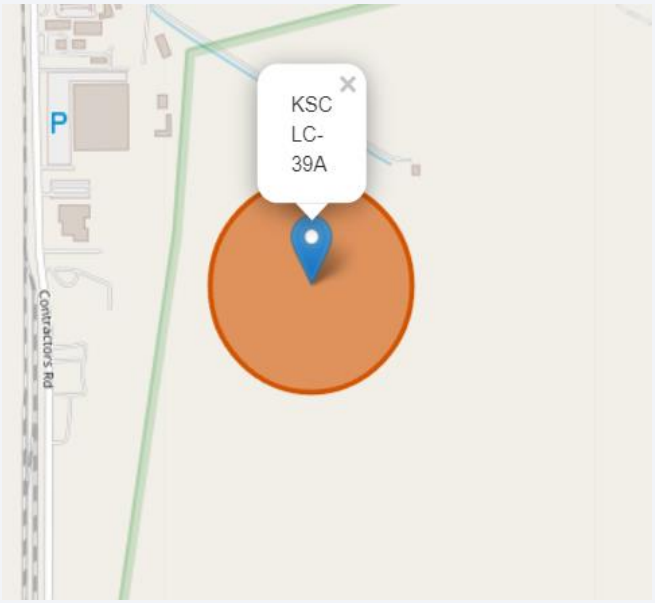
- The rank count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	Count	Rank_no
No attempt	10	8
Failure (drone ship)	5	5
Success (drone ship)	5	5
Success (ground pad)	5	5
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	3
Failure (parachute)	1	1
Precluded (drone ship)	1	1

Section 3

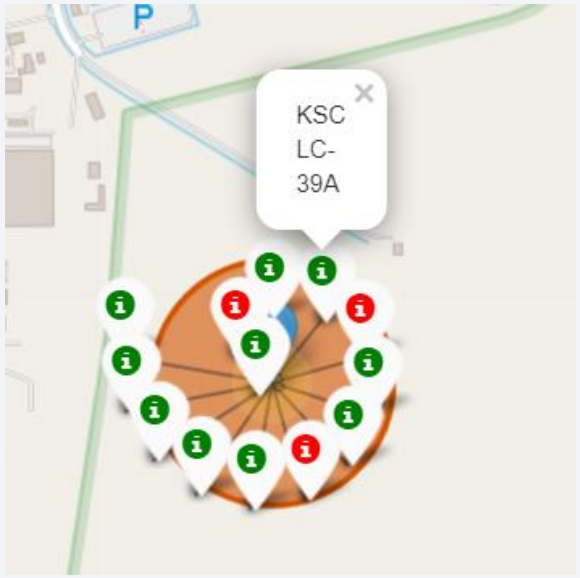
Launch Sites Proximities Analysis

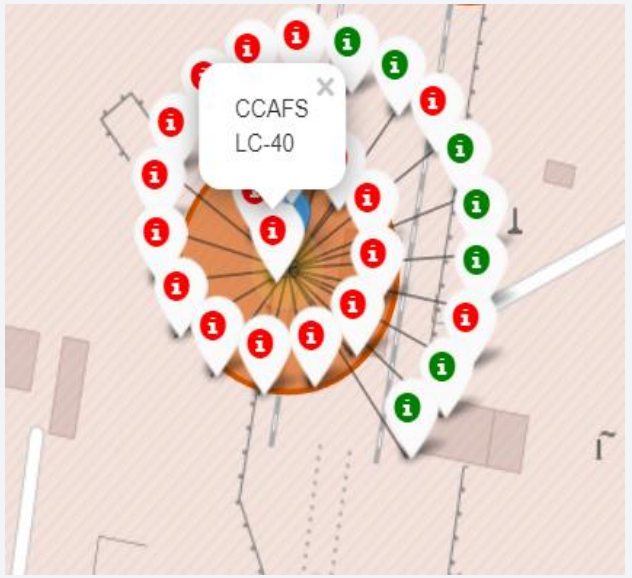
Launch Sites Location Analysis

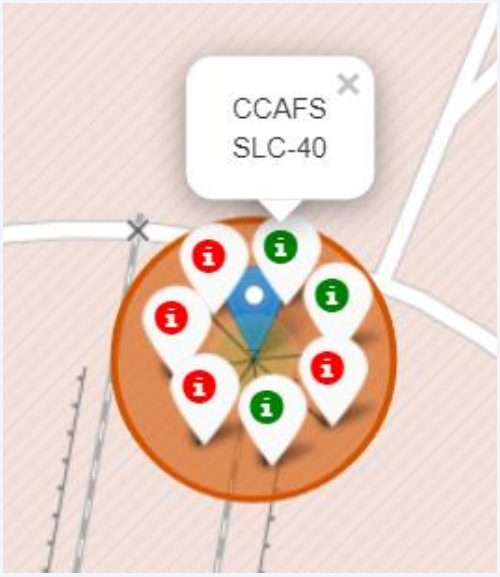


- All launch sites are in proximity to the equator and the coast.
- The launch sites in close proximity to the coast are for safety reasons.

Launch Sites Location and Success Rate

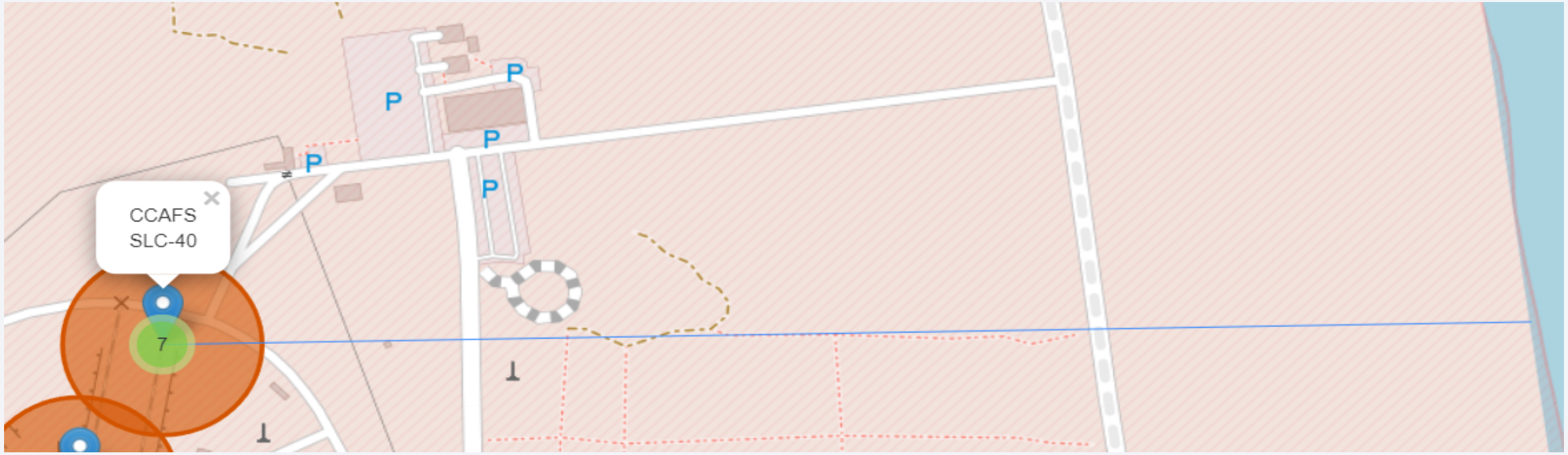






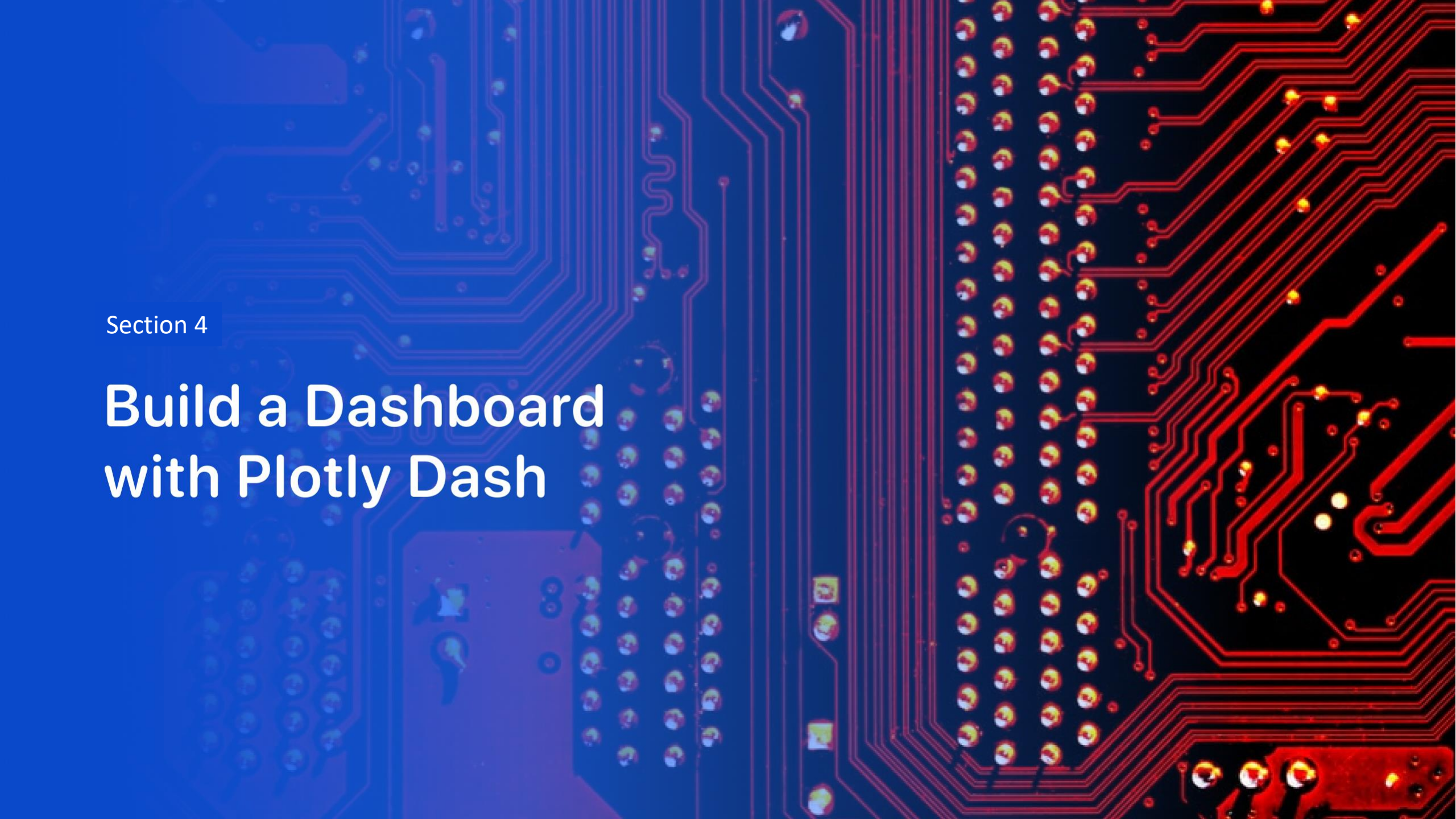
- KSC LC 39A Launch site has the highest success rate compared to all launch sites

Distances between a launch site to its proximities

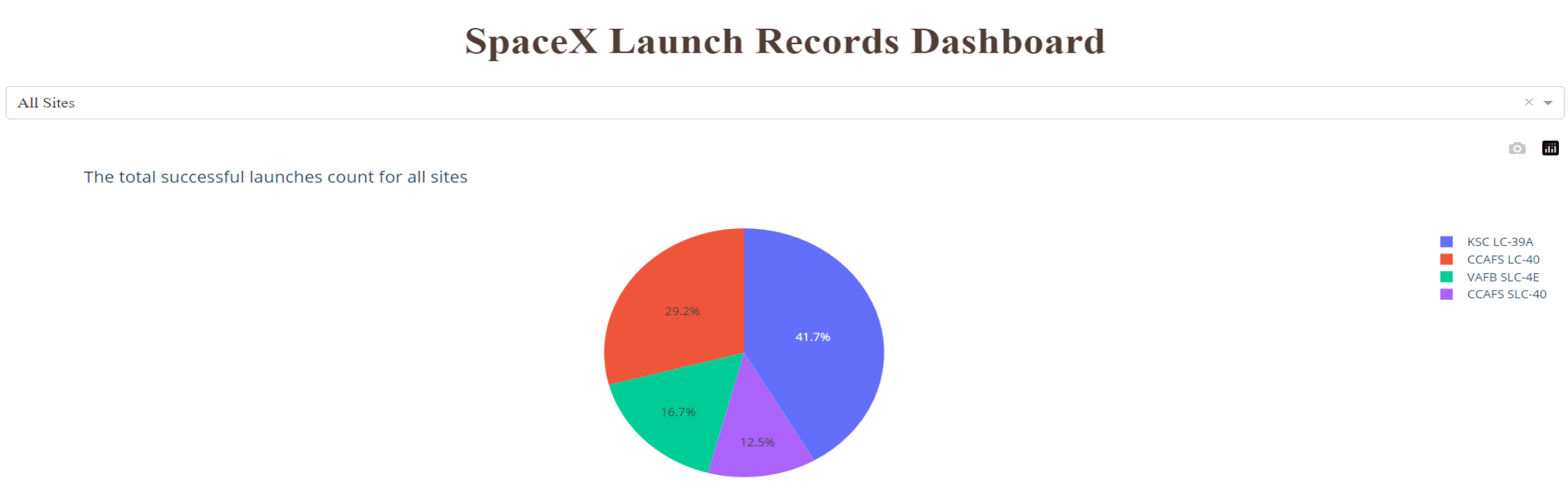


- Launch sites are in close proximity to coastline.

- Launch sites are not in close proximity to cities, which minimizes danger to population dense areas.


Section 4
Build a Dashboard with Plotly Dash

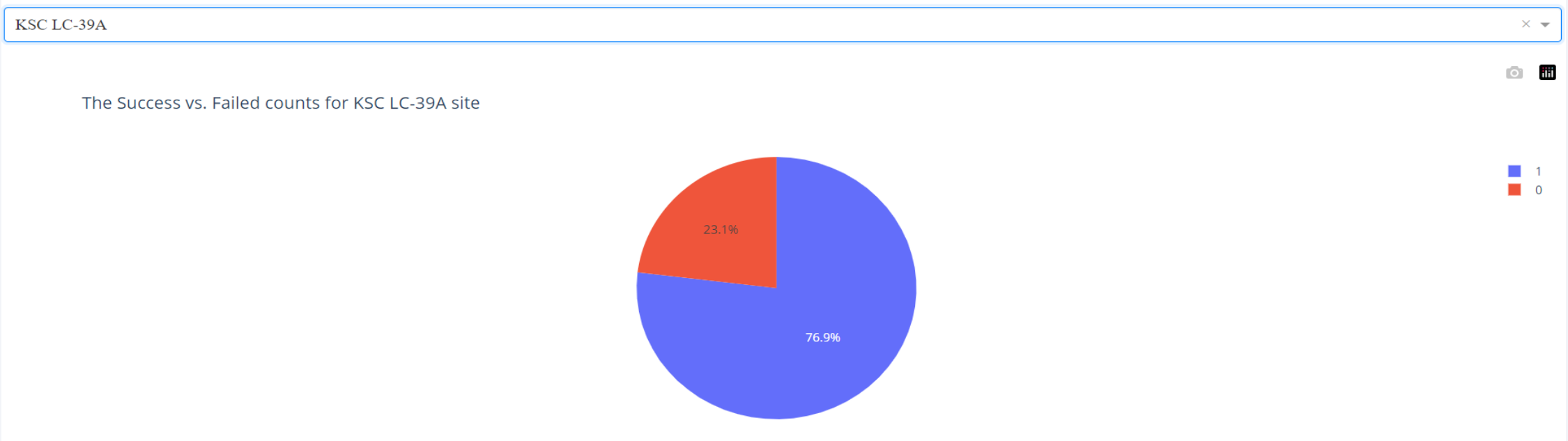
Launch sites success rate dashboard



- The Pie chart shows the total successful launches count for all sites

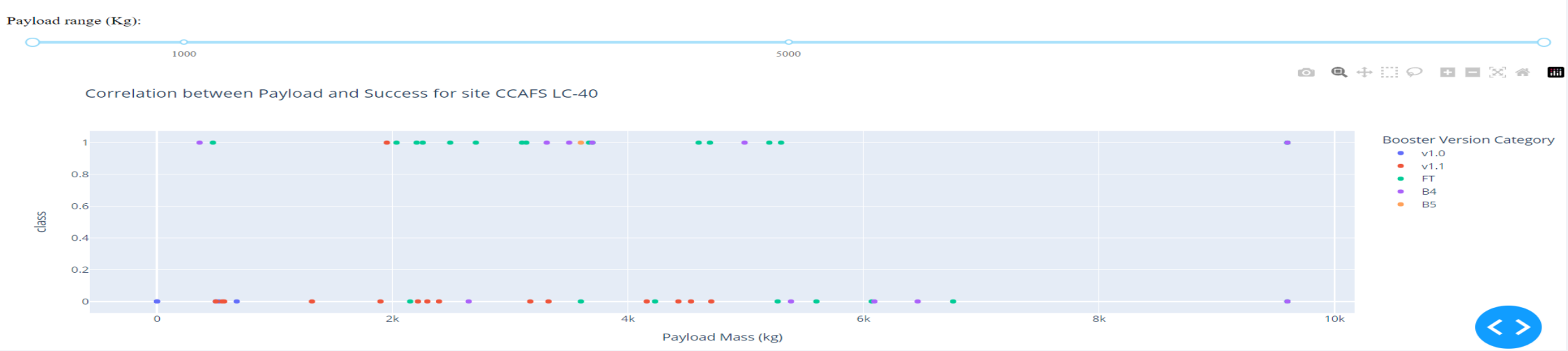
- KSC LC-39A launch site has the largest successful launches

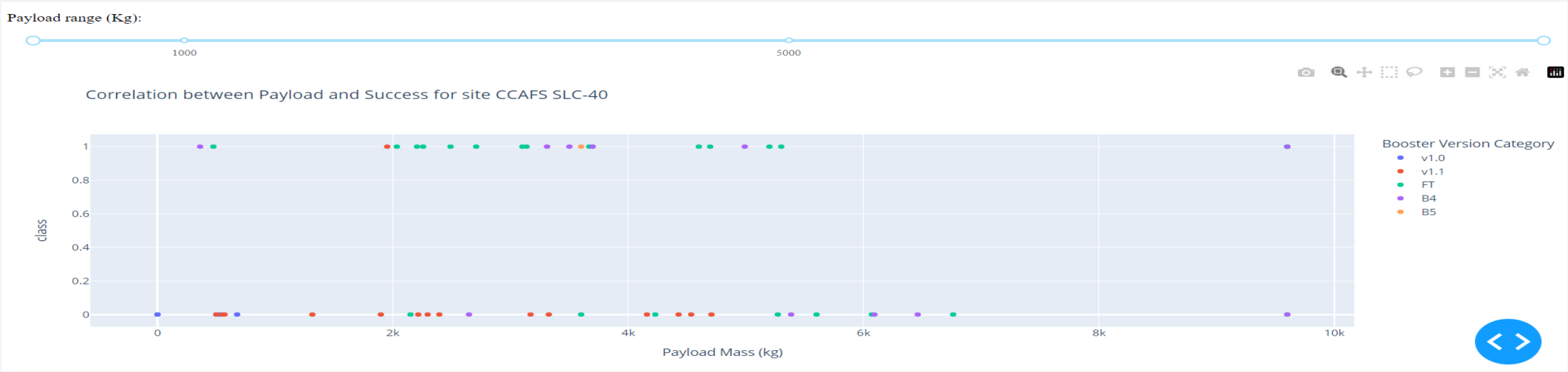
The highest success rate launch site dashboard



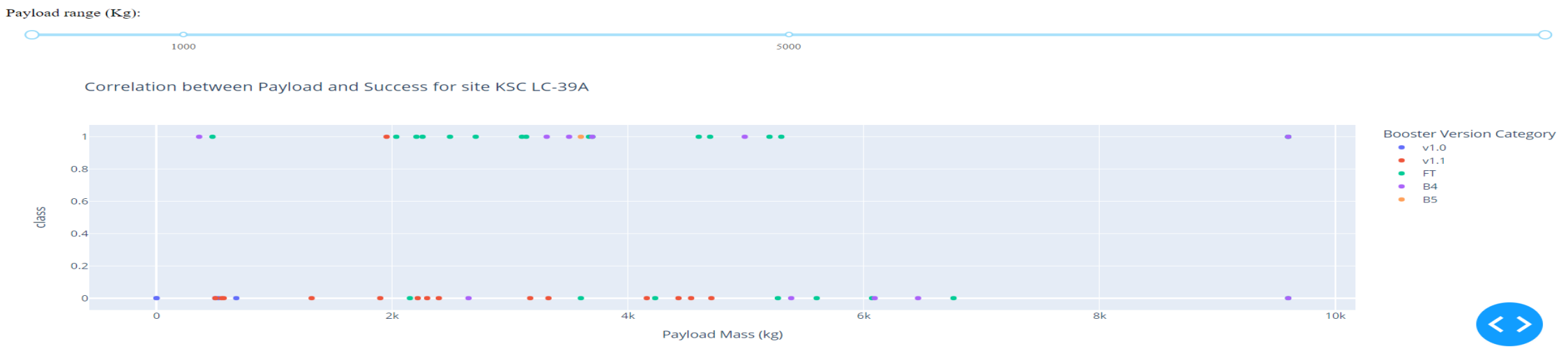
- According to Pichart of the dashboard, the landing at the KSC LC-39A launch site was successful in about 77% of launch cases.

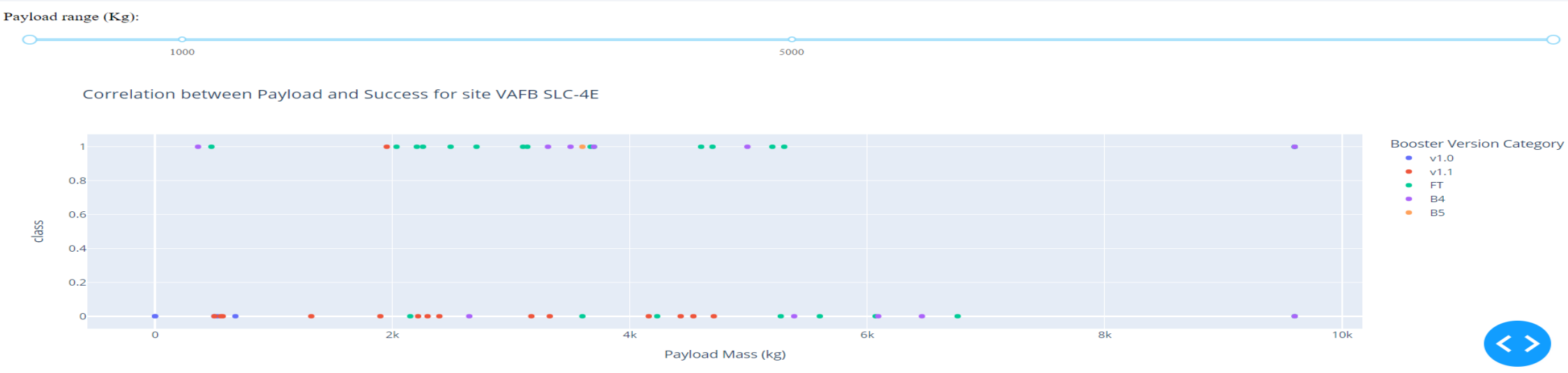
Payload vs. Launch Outcome scatter plot for all sites



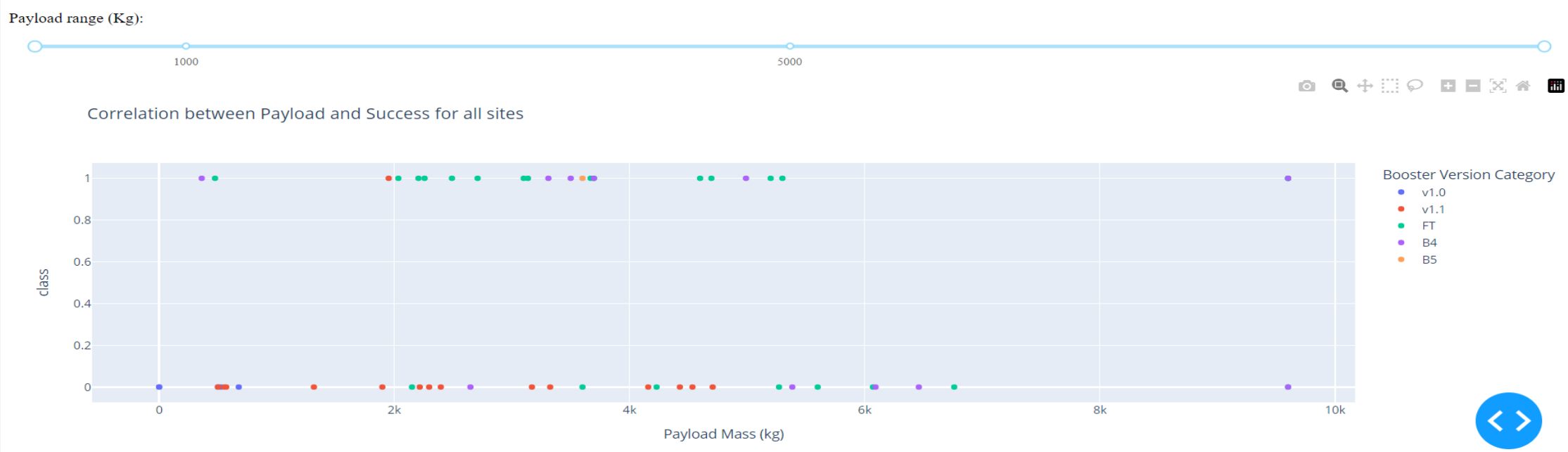


Cont.


Payload range (Kg):
1000
5000
Correlation between Payload and Success for site KSC LC-39A
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5


Payload range (Kg):
1000
5000
Correlation between Payload and Success for site VAFB SLC-4E
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Cont.



- For payload masses between 1000kg and 5000kg, the KSC LC-39A launch site has the highest success rate.
- For payload masses less than 2000kg, almost all launch site has lowest launch success rate.
- Booster version FT has the highest success rate of landing.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- LogisticRegression and SVM has similar and best train and test accuracy

Model	Train Accuracy	Test Accuracy
LogisticRegression	87.50	83.33
SVM	88.88	83.33

Confusion Matrix

- The Confusion Matrix used to evaluate the performance of a each model, it will tell how many correct & wrong predictions are their.

  - Precision = TP/TP+FP = 12/(12+3) = 0.8
  - Recall = TP/TP+FN = 12/(12+0) = 1
  - F1 score = 2(P*R)/(P+R) = 2*(0.8*1)/(0.8+1) = 0.88
  - Accuracy = TP + TN / TP + TN + FP + FN. = (12+3) / (12+3+3+0) = 0.833

Conclusions

- The fact that SpaceX can reuse the first-stage accounts for a large portion of its savings. The success rate of landing the first-stage has been rising Since 2013 until 2020.

- Flight number, payload mass, launch site, Orbit are highly correlated with the success rate of the tests.

- Launch sites are in close proximity to coastline, which minimizes danger to population dense areas.

- KSC LC 39A launch site has the highest success rate about 77% of launch cases compared to all launch sites

- By selecting the above parameters and using recent dataset SVM and Logistic Regression models can be used to predict the success of first-stage landing of new launches.

Appendix

- All the project source codes are documented on GitHub with the following link
  - https://github.com/robeleq/IBM_Applied_Data_Science_Capstone_Project

- External resources used are listed below
  - Falcon_9_first-stage_landing_tests
  - List_of_Falcon_9_and_Falcon_Heavy_launches
  - API SpaceX Data Launches

Thank you!