IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Saira Banu>
<11-OCT-2023>

Outline

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

Executive Summary

Summary of methodologies

1. Data Collection with SpaceX API and Web Scraping from Wikipedia page.
2. Data Wrangling
3. Exploratory Data Analysis (EDA) with SQL , Pandans and Matplotlip
4. Interactive Visual Analytics and Dashboard with Folium and Plotly Dash
5. Machine Learning Prediction - Predictive aAnalysis (Classification)

- Summary of all results

1. Results from Interactive Visual Analytics
2. Presenting Best Hyperparameter for SVM, Classification Trees and Logistic Regression
3. Method that performs best using test data with the accuracy score

Introduction

Project background and context

Space X a rocket launch company, helps in rocket launches in an inexpensive cost compare to other companies in the market. This is because Space X can reuse the first stage.

The context here is to determine if first stage will land, and cost of its launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers

1. What factors determine if the rocket will land successfully?
2. The interaction amongst various features that determine the success rate of a successful landing. OR Where is the best place to make launches.?
3. What operating conditions needs to be in place to ensure a successful landing program.
4. The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API(https://api.spacexdata.com/v4/rockets/) •
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - The Collected data was in Json format, After downloading the data from the above sites. The data was parsed using Beautiful soup and later was converted into a Pandas Data Frame.
  - One hot encoding was performed to enrich the data and created respective labels or features in analyzing it further

Methodology

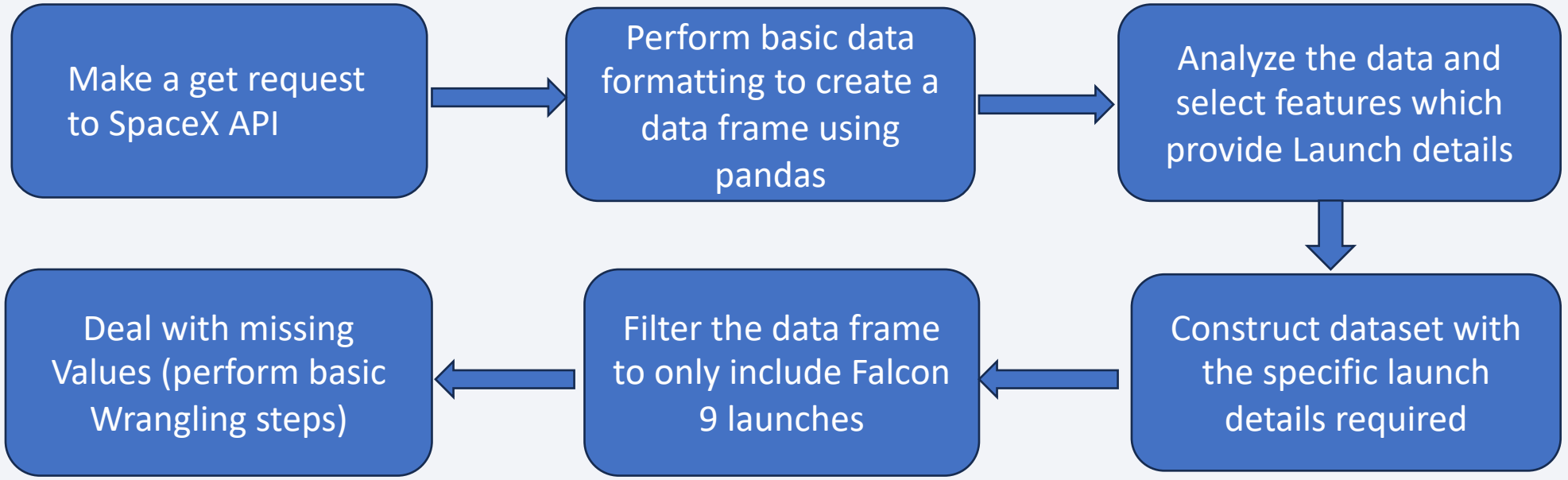
Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis was performed with the help of Machine Learning – Classification Algorithm
  - For this the data was normalized, divided or was splitted into training and testing set along with an dependent variable as a separate feature.
  - Later the training set was fitted based on the algorithms and accuracy was calculated and the best method was selected

Data Collection

Main objective for Data Collection step:

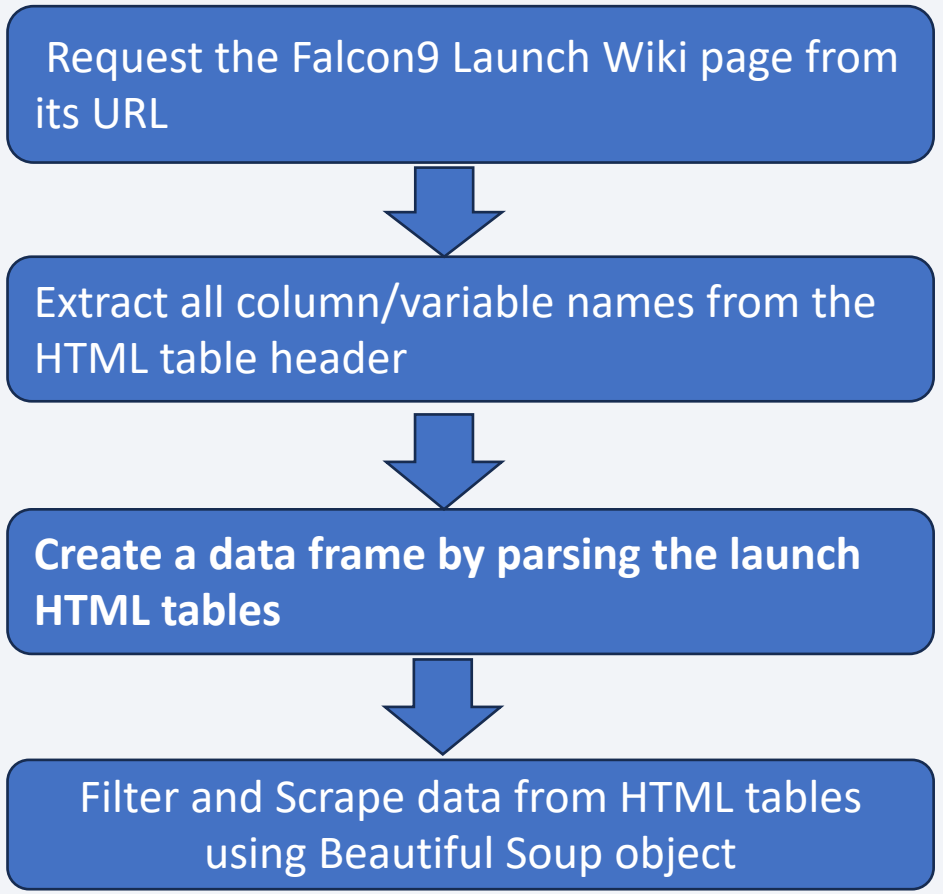
- Request to the SpaceX API
- Clean the requested data



Source Code : https://github.com/Sair07/Applied_data_science_capstone/blob/main/spacex-data-collection.ipynb

Data Collection - Scraping

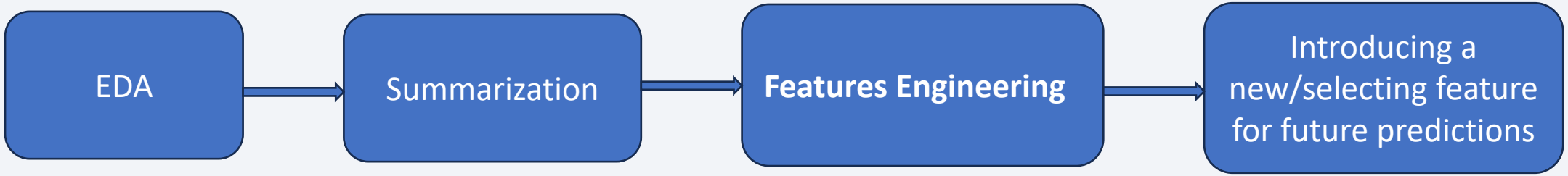
- performing web scraping to collect Falcon 9 historical launch records
- Wikipedia link https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Applying basic BeautifulSoup Python library to scrape webpages for data and filter the data



Source code: https://github.com/Sair07/Applied_data_science_capstone/blob/main/webscraping.ipynb https://github.com/Sair07/Applied_data_science_capstone/blob/main/labs_module%201_Web%20Scraping_Web-Scraping-Review-Lab.ipynb

Data Wrangling

- Performing Exploratory Data Analysis (EDA) on the dataset using Pandas.
- After loading the dataset into pandas check the summary of the data
- drill down to each site visualize its detailed launch records
- Visualize the relationship between features in the dataset
- Finally, select the features that will be used in success prediction in the future module



Source Code : https://github.com/Sair07/Applied_data_science_capstone/blob/main/Spacex-data_wrangling.jupyterlite.ipynb

EDA with Data Visualization

- **Catplot** to visualize the relationship between Flight Number and Payload.
- **Catplot** to visualize the relationship between Flight Number and Launch Site.
- **Catplot** to visualize the relationship between Payload and Launch Site.
- **Bar chart** to visualize the relationship between success rate of each Orbit type.
- **Catplot** to visualize the relationship between Flight Number and Orbit type.
- **Catplot** to visualize the relationship between Payload and Orbit type.
- **Line chart** to visualize the launch success early trend
- Source Code : https://github.com/Sair07/Applied_data_science_capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission

  %sql select distinct(Launch_Site) from SPACEXTABLE

- Display 5 records where launch sites begin with the string 'CCA'

  %sql select * from SPACEXTABLE where Launch_Site like "%CCA%" limit 5;

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

  %sql select sum(PAYLOAD_MASS__KG_) ,Booster_Version,Customer from SPACEXTABLE where Customer like '%NASA%' group by Booster_Version;

- Display average payload mass carried by booster version F9 v1.1

  %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%%F9 v1.1'

- List the date when the first succesful landing outcome in ground pad was acheived.

  %sql SELECT MIN(DATE) AS first_ successful_ landing FROM SPACEXTBL WHERE (landing_outcome)='Success (ground pad)';

EDA with SQL

- Display the names of the unique launch sites in the space mission

  %sql select distinct(Launch_Site) from SPACEXTABLE

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

  %sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;

- List the total number of successful and failure mission outcomes

  %sql select count(Mission_Outcome),Mission_Outcome from SPACEXTABLE group by Mission_Outcome;

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

  %%sqlselect distinct(Booster_Version) from SPACEXTABLE where Booster_Version in (select Booster_Version where PAYLOAD_MASS__KG_ > 15000);

Source Code :
https://github.com/Sair07/Applied_data_science_capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Summary of map objects that were created and added to the Folium map

- folium. Circle and folium.Marker to add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.
- MarkerClusterobject for simplify a map containing multiple markers that has the same coordinate.
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Lines are used to indicate distances between two coordinates.

Source Code :
https://github.com/Sair07/Applied_data_science_capstone/blob/main/Interactive_map_Folium.ipynb

Build a Dashboard with Plotly Dash

The dashboard application contains input components as mentioned below to interact with a pie chart and a scatter point chart.

  - dropdown list
  - range slider

- A launch Site Drop-down Input Component, contains four different launch sites and a dropdown menu let us select different launch sites with a select all option
- A callback function to render success-pie-chart based on selected site from dropdown.This callback function gets the selected launch site from site-dropdown and render a pie chart visualizing launch success counts.
- A range Slider to Select Payload , to easily select different payload range and see if we can identify some visual patterns.

Build a Dashboard with Plotly Dash

- A callback function to render the success-payload-scatter-chart scatter plot.To visually observe how payload may be correlated with mission outcomes for selected sites and also collective for all sites based on the drop down value.

Source Code:
https://github.com/Sair07/Applied_data_science_capstone/blob/main/space_dash_app.py

Predictive Analysis (Classification)

Summary of the model development process

- Creation of a NumPy array from the dependant column Class in data.
- Data standardization using standard scaler preprocessing method
- Use of the function train_test_split to split the data X and Y into training and test data.
- Searching for the best Hyperparameters for Logistic Regression, SVM, Decision Tree and KN classifiers, by calculating the accuracy after training the dataset
- Find the method performs best using test data
- Source Code: https://github.com/Sair07/Applied_data_science_capstone/blob/main/SpaceX_Machine_Learning_Prediction.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015-12-22;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed (14 success and 5 Failure)

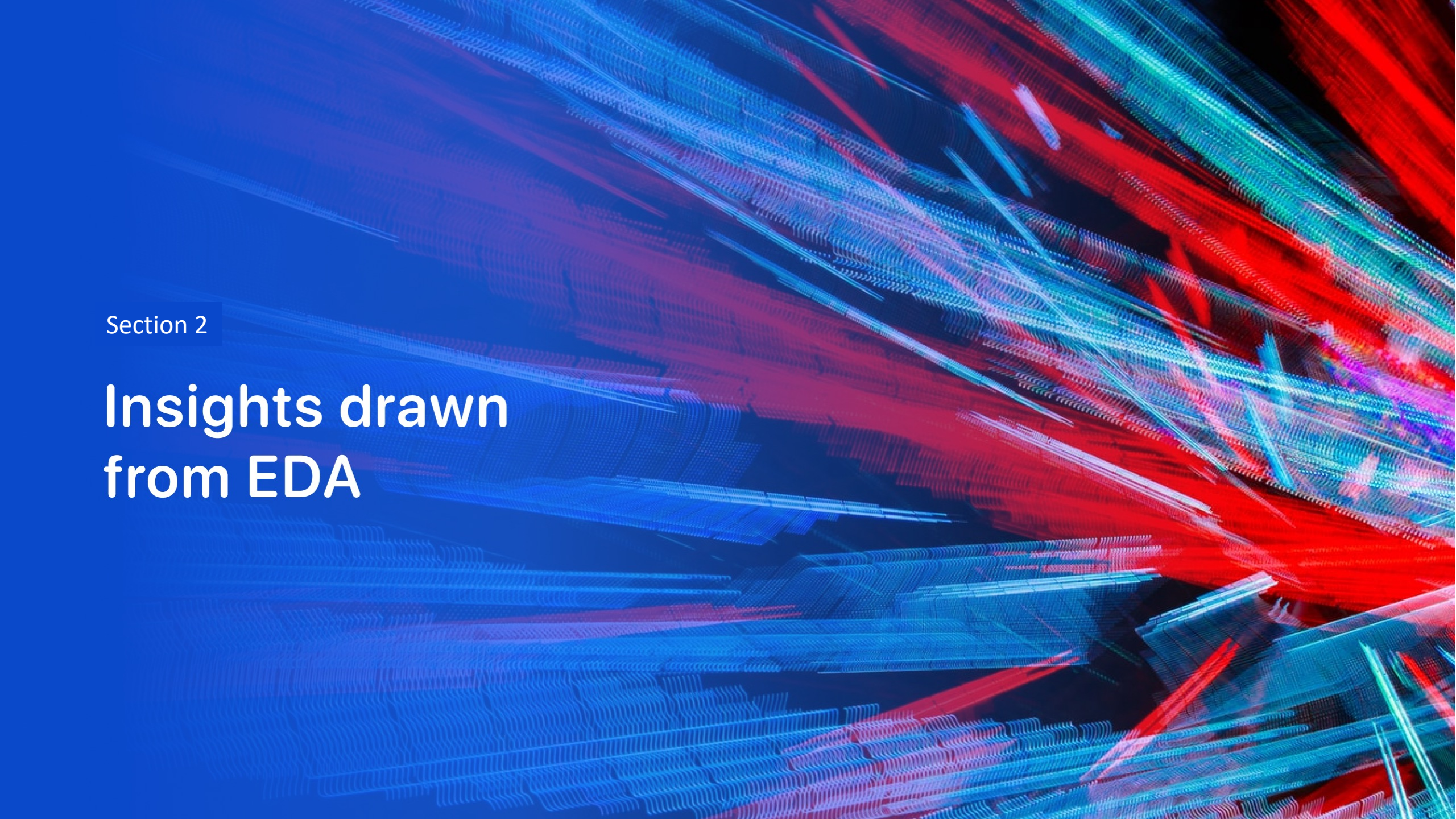
Results

- Interactive analytics demo in screenshots
  - Using interactive analytics was possible to identify that launch sites is safety places, near sea, as shown below with a map derived from Folium

- Predictive analysis results

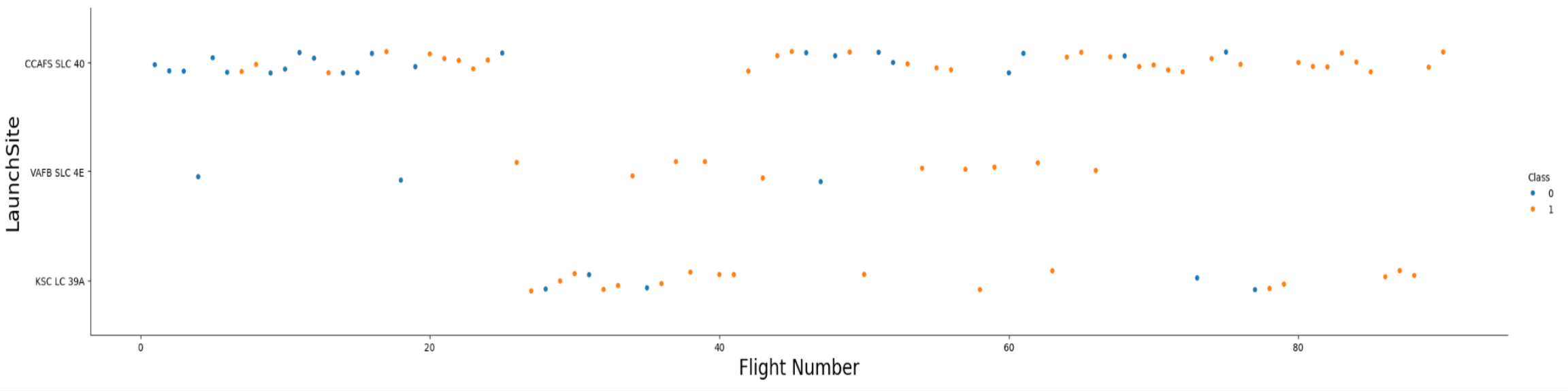
Results

- Predictive analysis results
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.
  - Descision Tree came best out of logistic regression, support vector machine, decision tree and k nearest neighbors.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

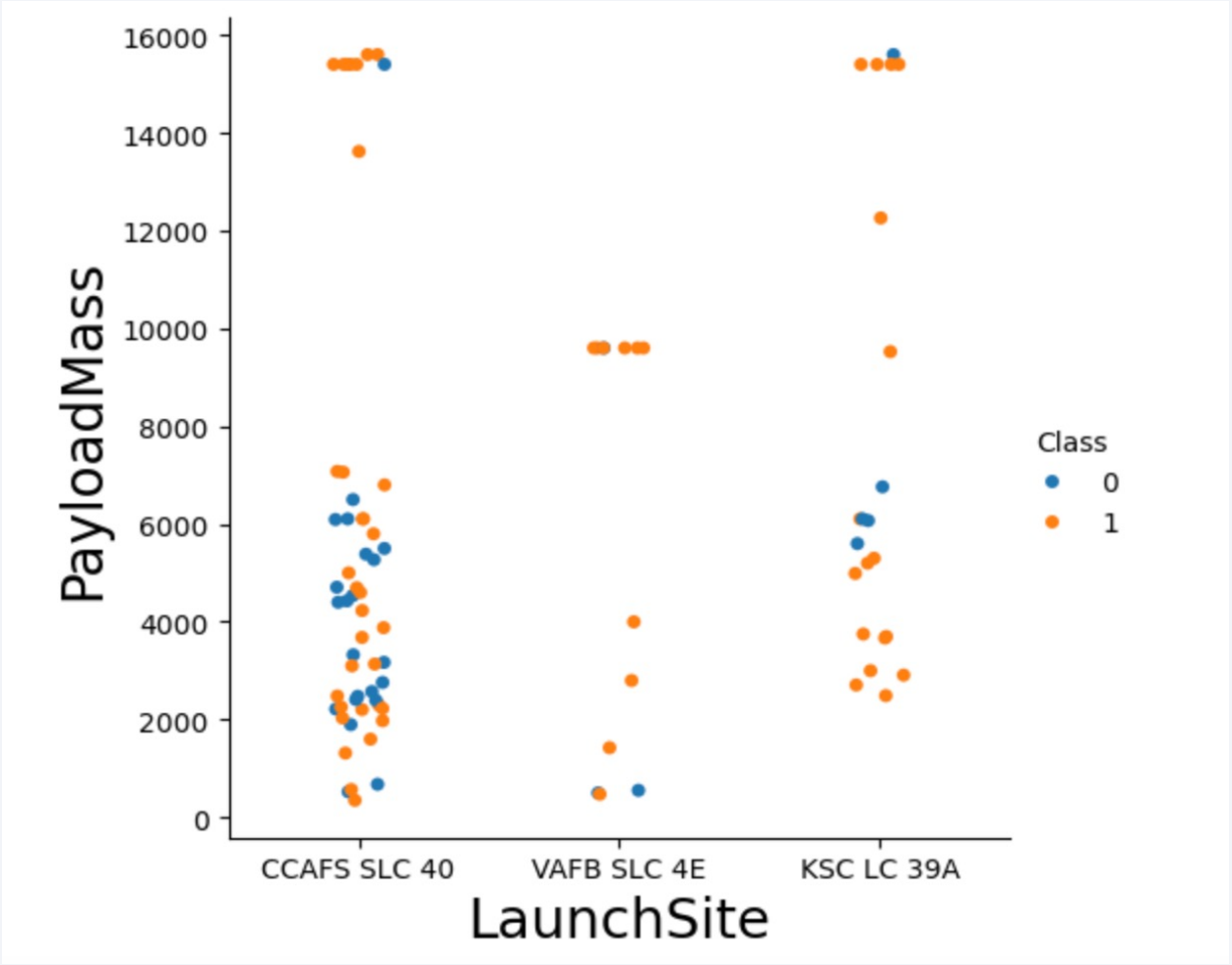
scatter plot of Flight Number vs. Launch Site



- Its clear from the the above graph that success rate has improved over time.
- According to the plot, the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
- VAFB SLC 4E is second and third place KSC LC 39A;

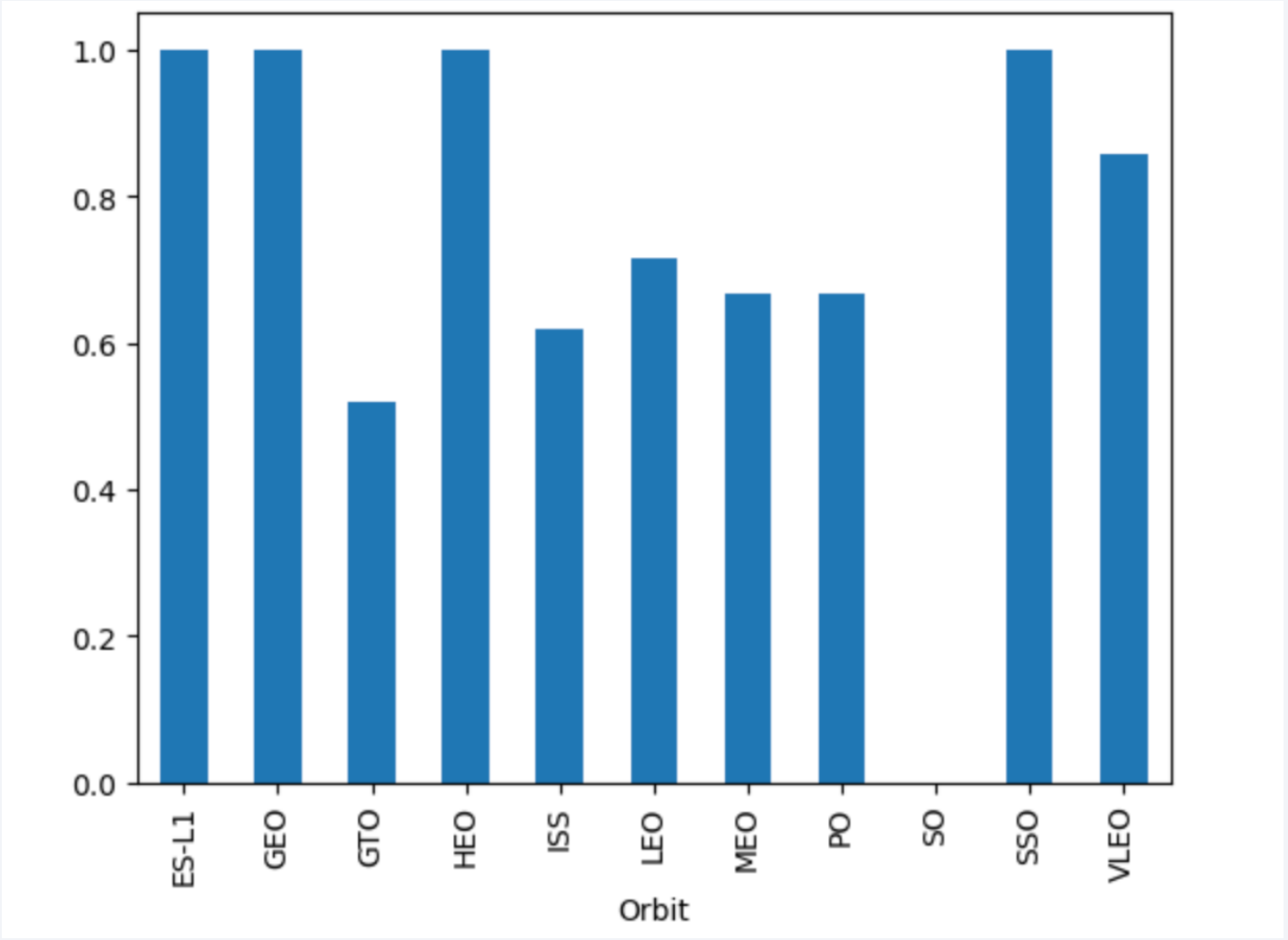
Payload vs. Launch Site

- Payloads over 9,000kg have good success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



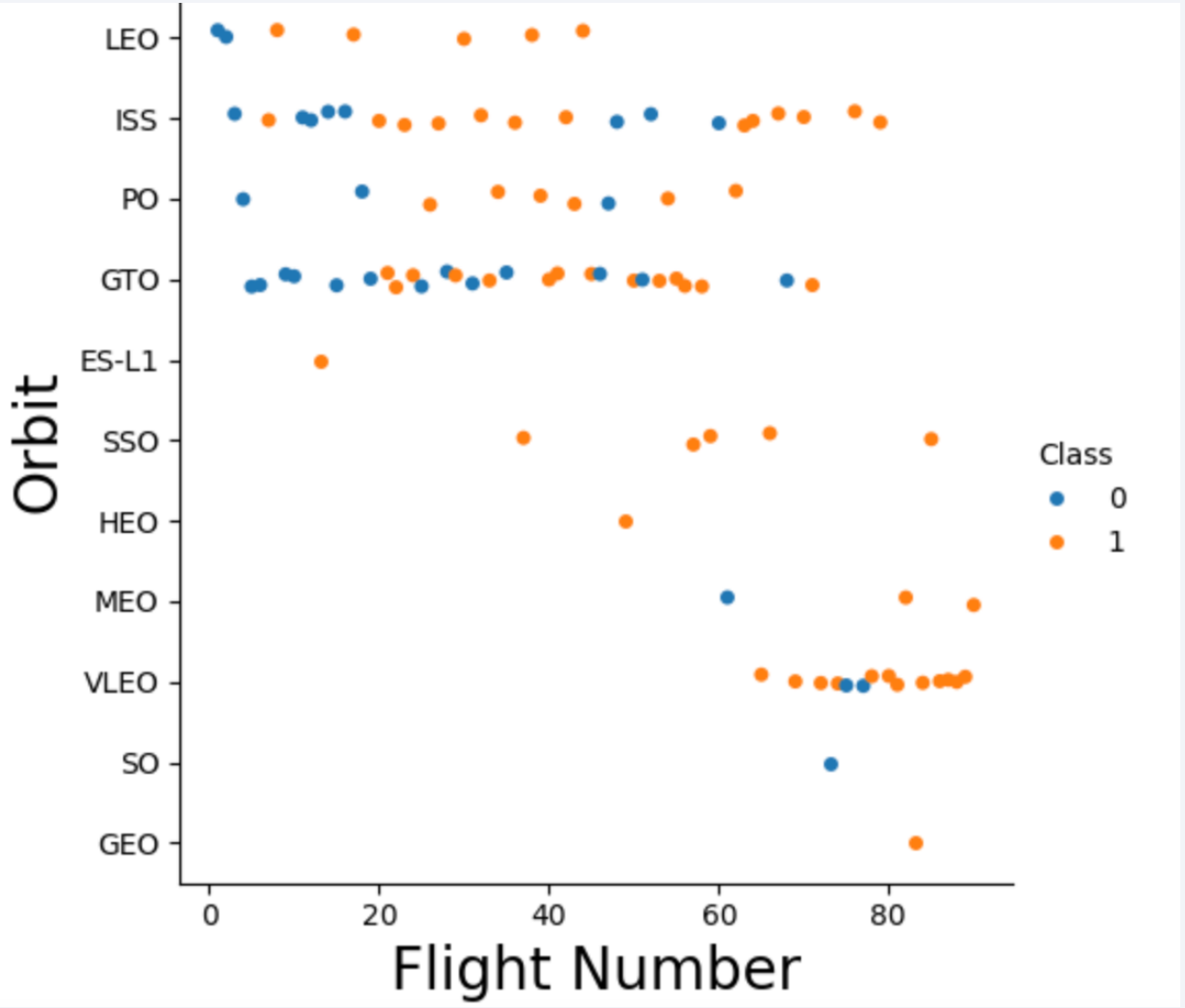
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO



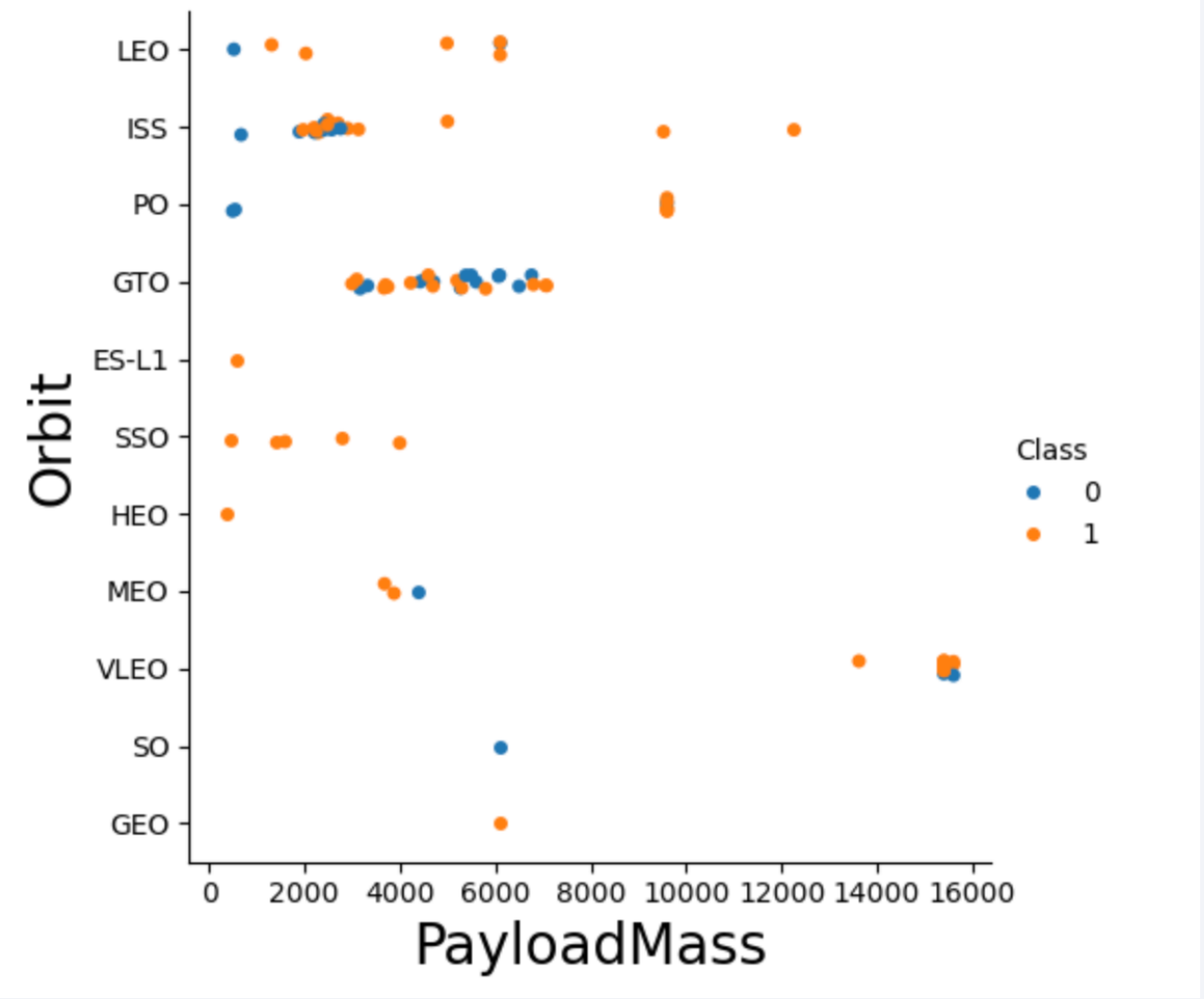
Flight Number vs. Orbit Type

- success rate improved over time to all orbits.
- VLEO orbit looks good as it has good success launches.



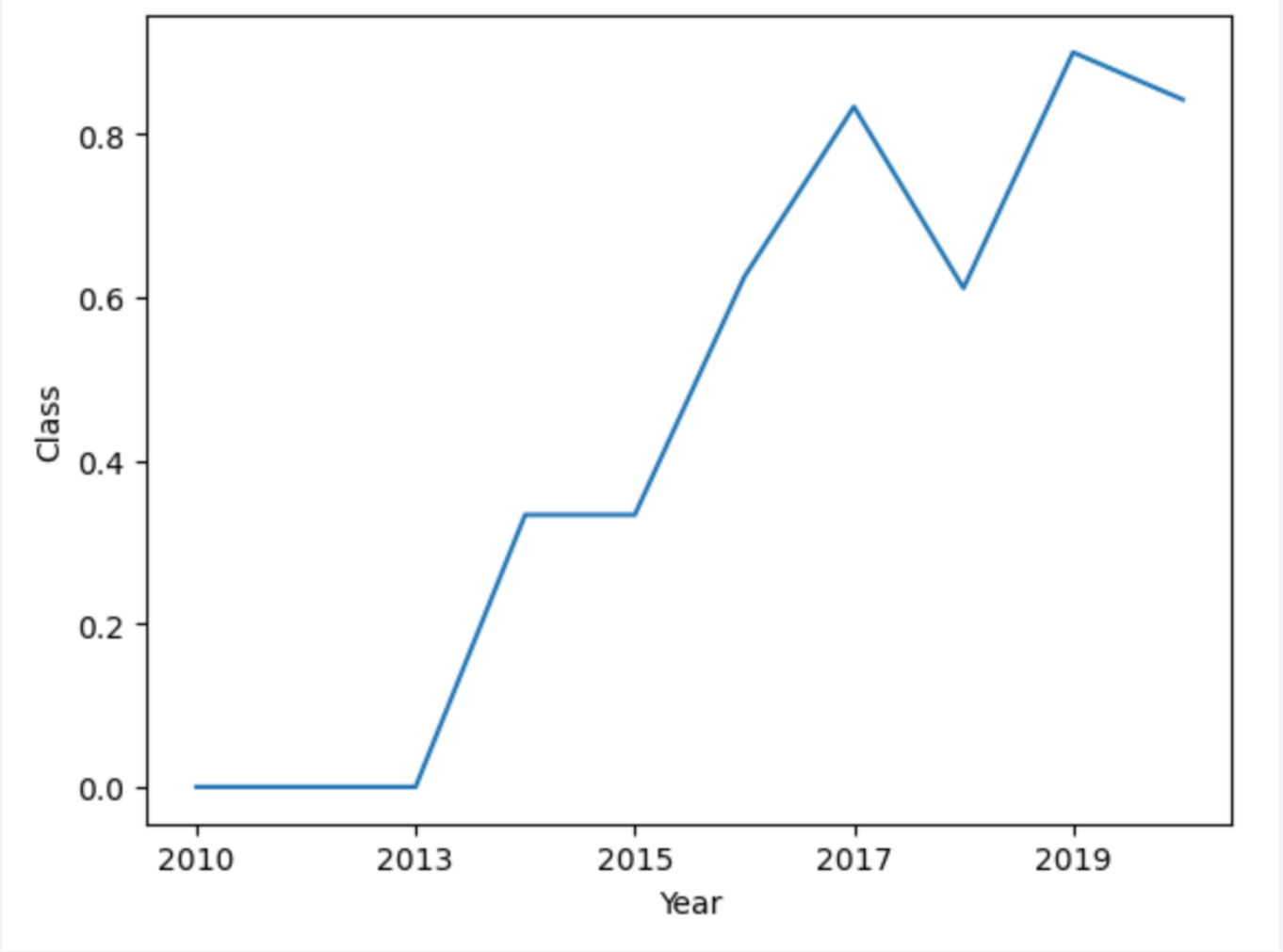
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO as it has mixed results
- ISS orbit has wide range of payload and with high success rate
- There are less launches to the orbits SO and GEO.



Launch Success Yearly Trend

- Success rate started improving from 2013
- First 3 years looks like learning years to get hands on the technology with no success



All Launch Site Names

- Names of the unique launch sites

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

Query: select distinct(Launch_Site) from SPACEXTABLE

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Query: select * from SPACEXTABLE where Launch_Site like "%CCA%" limit 5;

Total Payload Mass

- Total payload carried by boosters from NASA

Query: select sum(PAYLOAD_MASS__KG_) ,Booster_Version,Customer from SPACEXTABLE where Customer like '%NASA%' group by Booster_Version;

- Query to Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

- Query : select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%%F9 v1.1'
- Filtering data by the booster version above and calculating the average payload mass using avg() function we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

Out[29]:

min(Date)
2015-12-22

- Query: SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing_outcome)='Success (ground pad)';
- Using the min function and filtering on the landing outcome.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[36]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- Query: select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
- By filtering the given range on the payload mass Kg column and landing outcome we can see the above 4 booster had success landing

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

count(Mission_Outcome)	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

- Query: select count(Mission_Outcome),Mission_Outcome from SPACEXTABLE group by Mission_Outcome;
- Just by using the group by clause we can get the clear detail on the failure and success outcomes

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Query: select substr(Date, 6,2) as "Month",Landing_Outcome, Booster_Version,Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome ='Failure (drone ship)' limit 10;
- use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

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

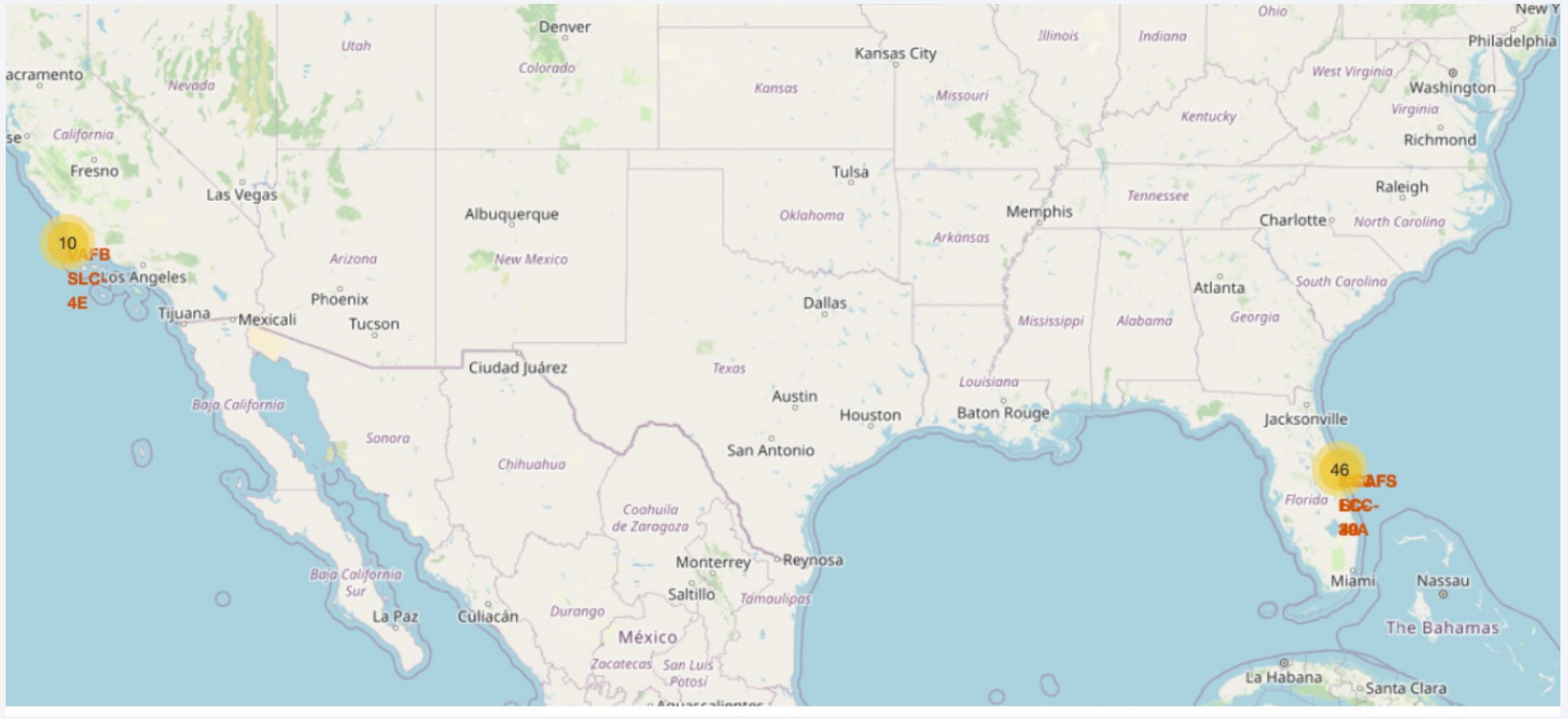
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

counts	Landing_Outcome
14	Success (drone ship)
5	Failure (drone ship)

- QUERY: select count(Landing_Outcome) as counts,Landing_Outcome from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' or Landing_Outcome = 'Success (drone ship)' group by Landing_Outcome order by counts desc;


Section 3
Launch Sites
Proximities Analysis

Success/Failed Launches For Each Site

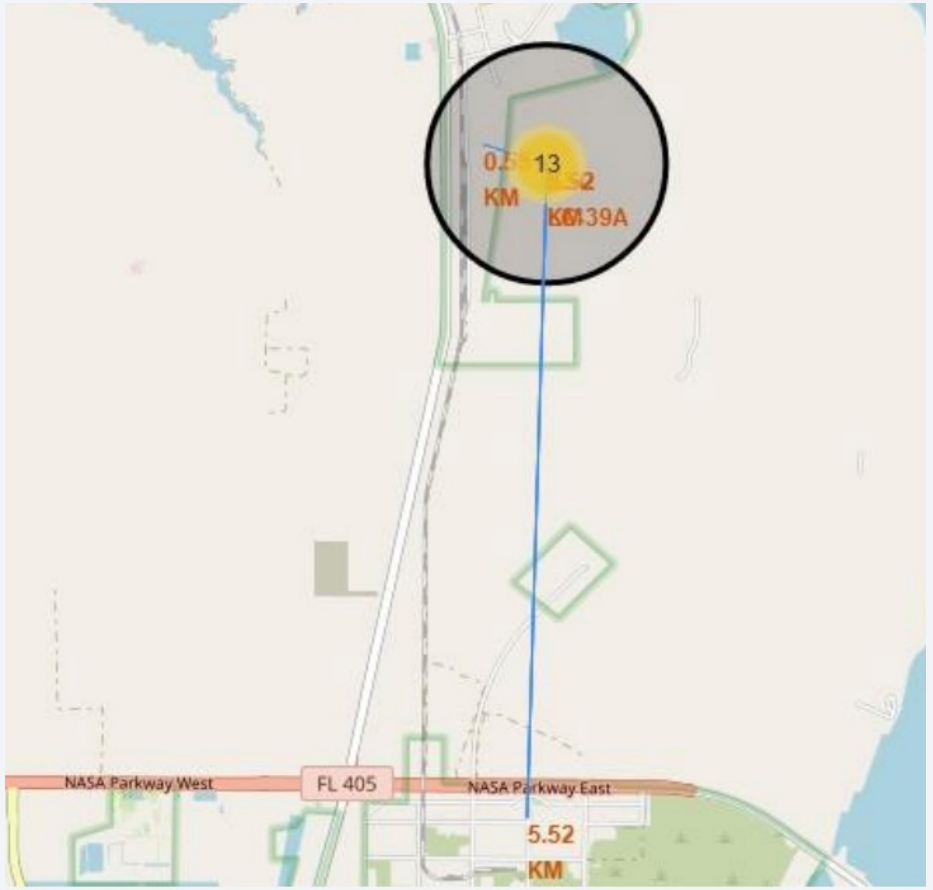


- The map shows clusters for every launch site, the second shows a red marker for a failed launch.

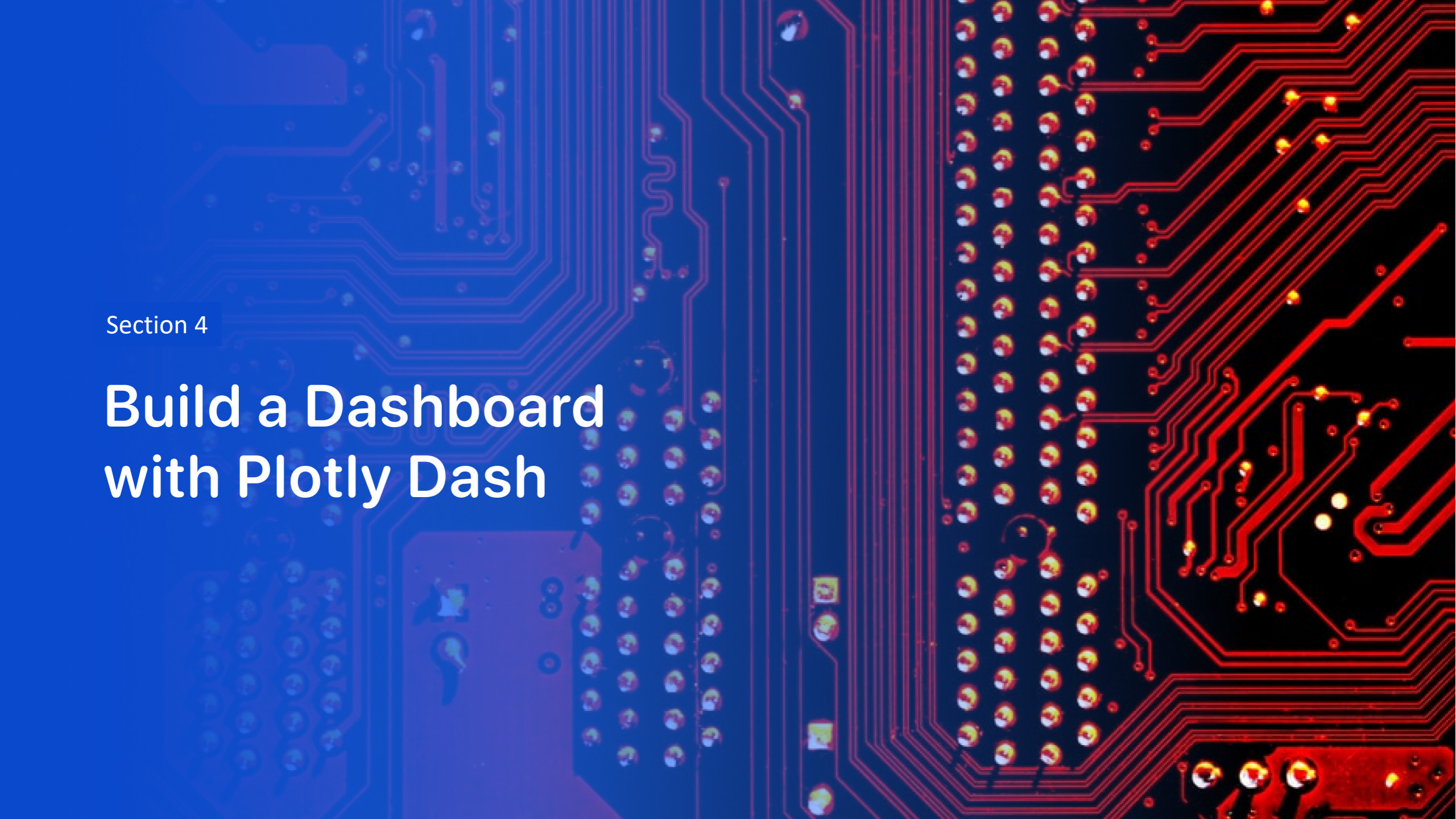
All Launch Site

- All Launch sites are in close proximity to the coast

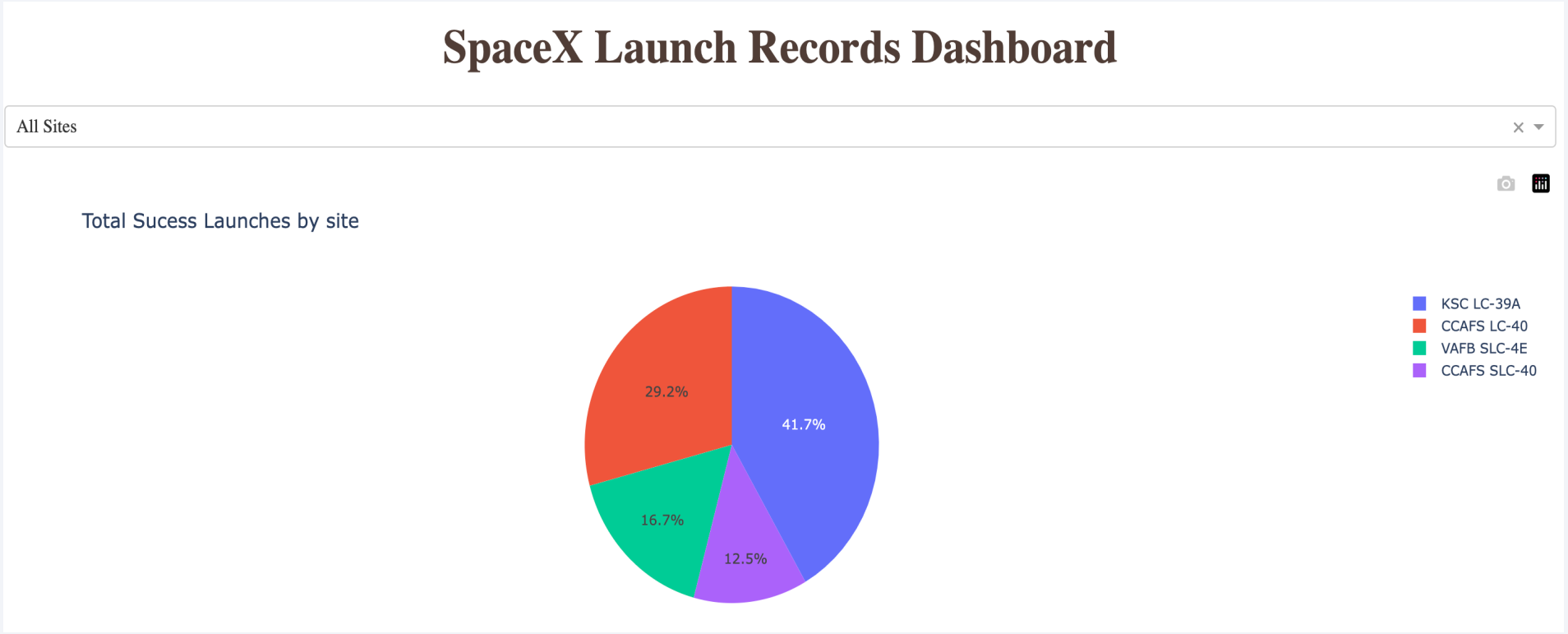
Proximity of Launch Site



- Launch sites are near to railways, roads, highways and coastline.

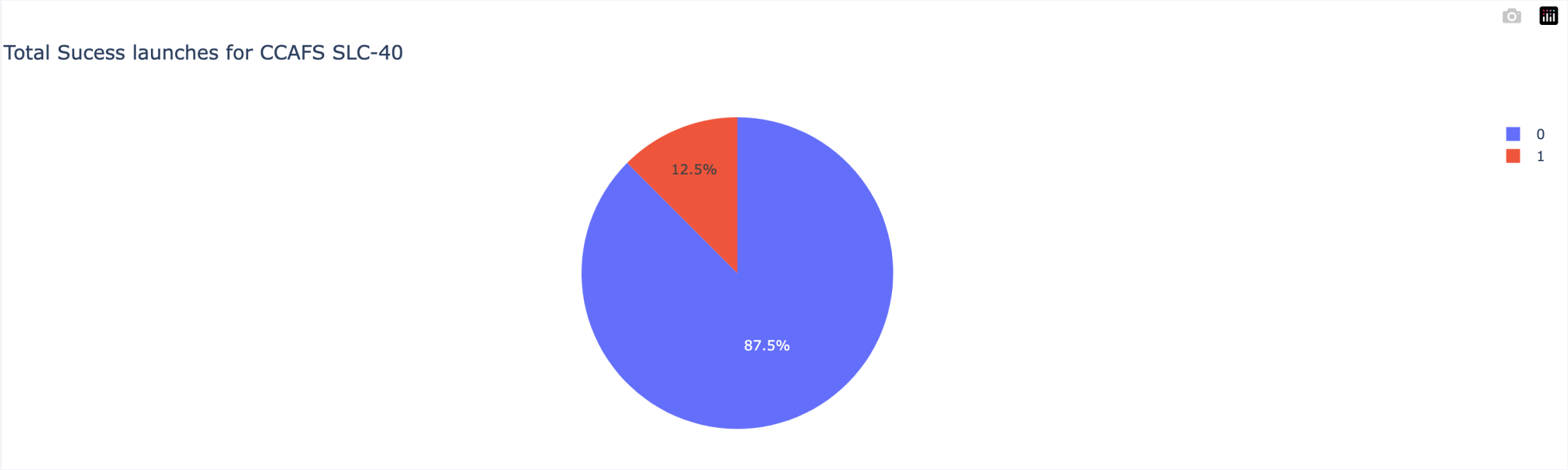

Section 4
Build a Dashboard with Plotly Dash

Success Launches by ALL sites



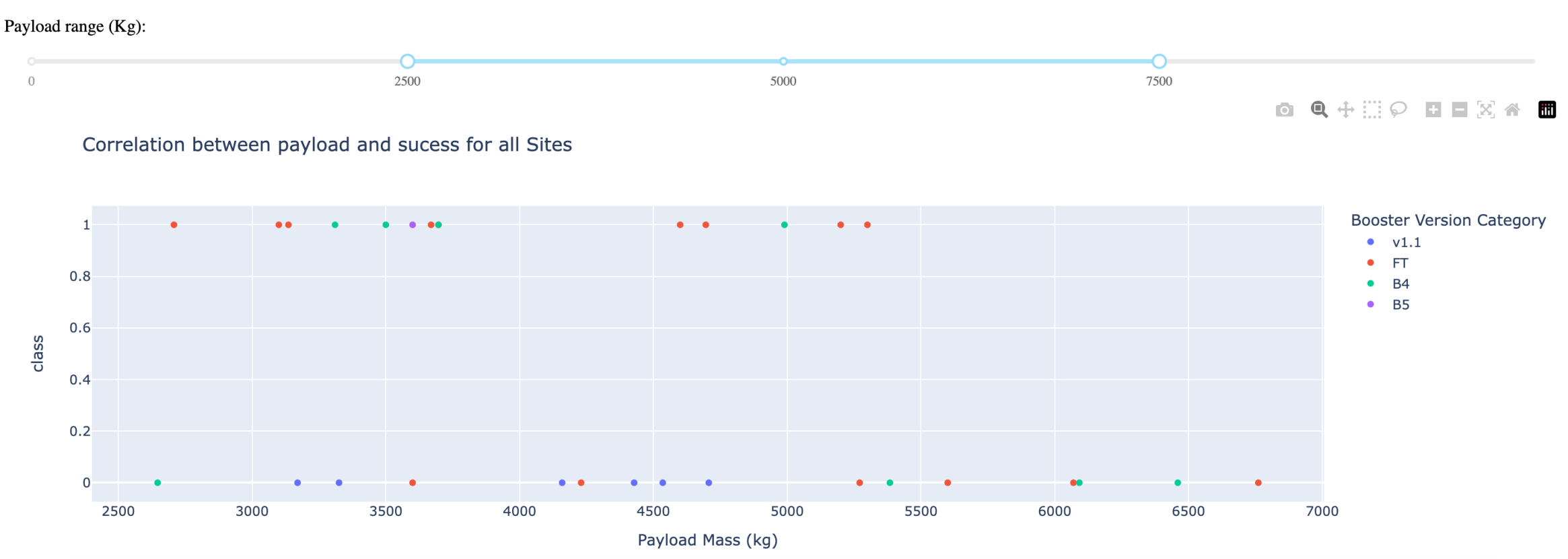
- KSC LC-39A is the site with the higher success launches followed by CCAFS LC-40. Also the place of launch seems to have an important factor in success of missions.

Launch Site with Highest Success Rate



- Launch Site CCAFS SLC-40 has the highest success rate of 87.5%
- Also from this interactive dash site we see that this is very close to the second best launch site VAFB SLC-4E which is having highest success rate of 83.3%

<Dashboard Screenshot 3>

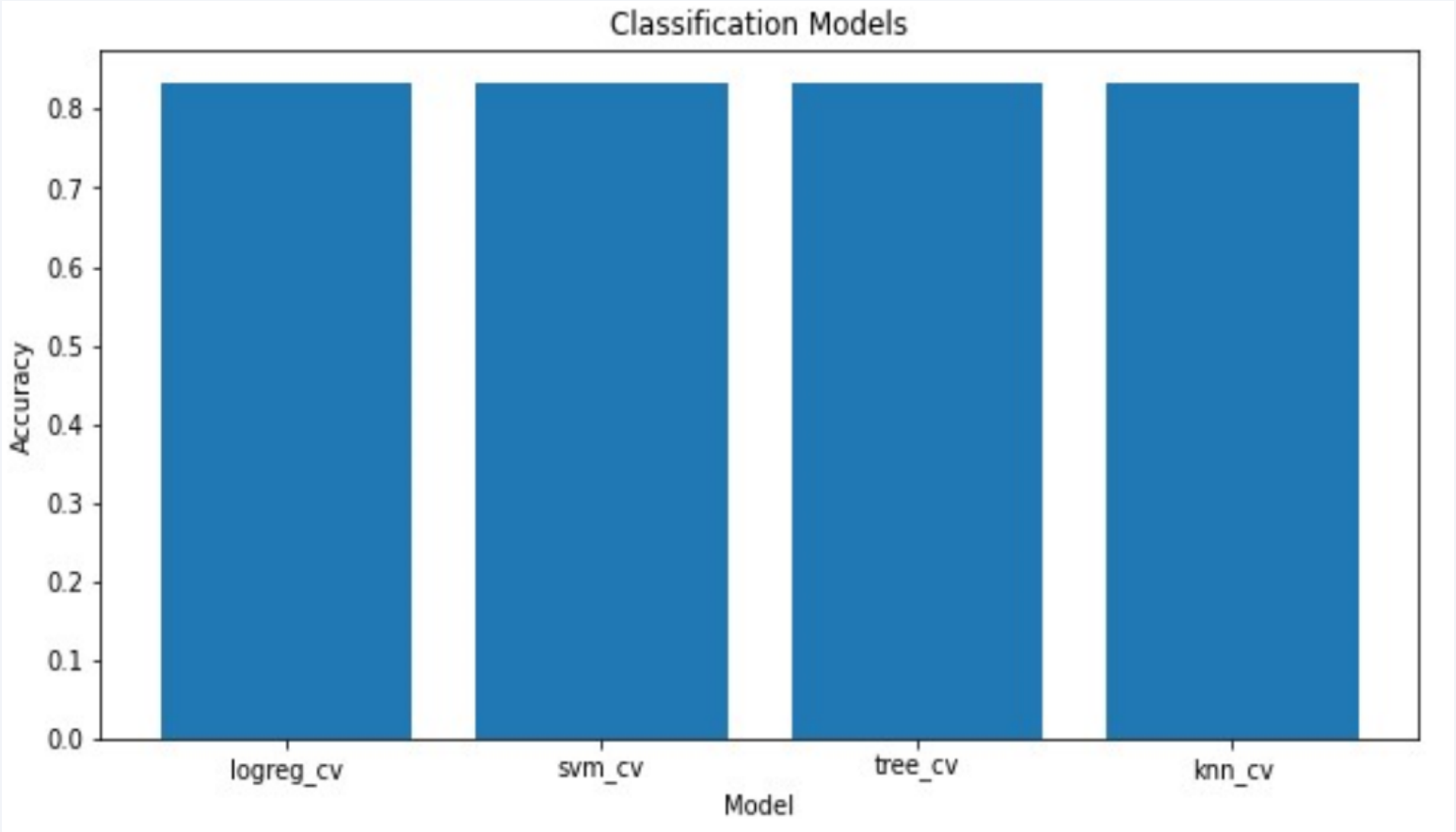


- Payload Vs the Success rate with a range slider to view in the specific range of payload.Scatter plot for all sites with 2500(kg) to 7000(kg) payload ranges.
- The 2500-5000(kg) range concentrate the majority of the successfully launches and has all success rates up to 5500(KG)

Section 5

Predictive Analysis (Classification)

Classification Accuracy

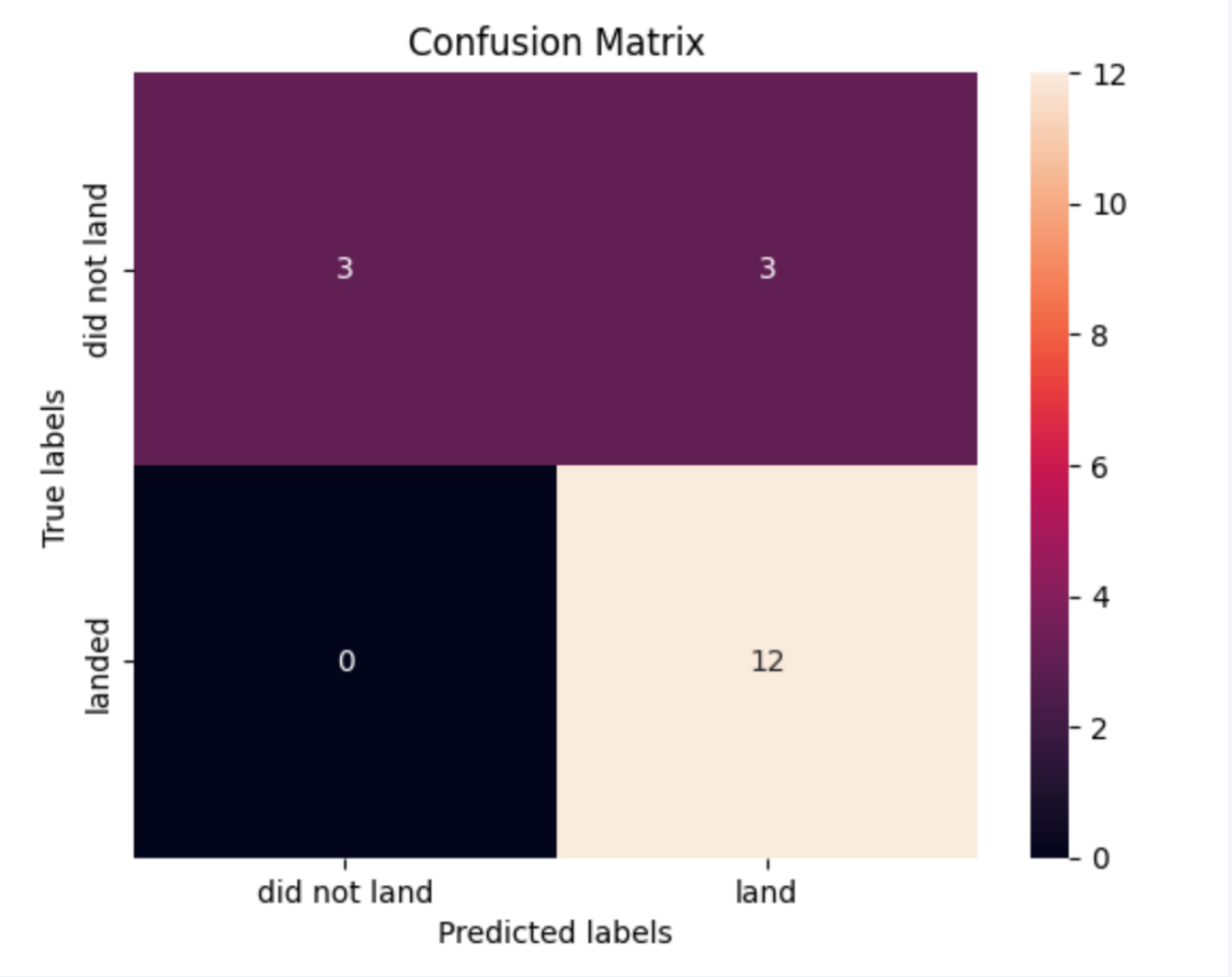


Classification Accuracy

```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix



Conclusions

- Different data sources were analyzed, refining conclusions along the process
- Launches above 7,000kg are less risky
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- As all the algorithms are giving the same accuracy, they all perform practically the same.
- By using our machine learning model, we can predict if the first stage of our competitor will land and determine the cost of a launch.

Appendix

- For Jupyter notebook or code snippets or sql queries access the below Github repository Link

  https://github.com/Sair07/Applied_data_science_capstone

Thank you!