IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gayathri
05/20/2025

Outline

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

Executive Summary

We will train a machine learning model to predict if SpaceX will reuse the first stage. Listed below are the methodology process and how results are used.

Methodologies :

- Data Collection
  - From API
  - Web Scraping
- Data Wrangling
- Exploratory Data Analysis(EDA)
  - Using SQL
  - Using Matplotlib and Pandas
- Interactive Visual Analytics
  - Using Folium
  - Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Results:

- Finding the best hyperparameters for a machine learning model involves using techniques like GridSearchCV, to explore the space of potential hyperparameter combinations and select the ones that yield the best performance.
- Calculating Accuracy and Plotting Confusion Matrix for Logistic Regression ,SVM,Decision Tree and KNN models can be used to find the model that performs best.

Introduction

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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

We will predict if the Falcon 9 first stage will land successfully using Machine learning models.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from 2 sources.
    - SpaceX API (JSON Object)
    - Web scraping Falcon 9 historical launch data from wiki Pages (HTML Tables)
- Perform data wrangling
  - Exploratory Data analysis(EDA) was performed to find patterns in data.
  - Converted Outcomes in to labels for Training Supervised Models.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was Standardized and Split in to Training and Test data
  - Four different Models was built(Logistic Regression,SVM,Decision Tree & KNN).
  - Best HyperParameters, Accuracy on Training and Test data sets and Confusion Matrix were calculated for all Models.
  - Models were evaluated using the above parameters to find the model with best performance

Data Collection

Data was collected from two sources

- SpaceX Rest API
  - https://api.spacexdata.com/v4/rockets/
  - Request to the SpaceX API
  - Get the JSON Object data and convert it into a Pandas dataframe

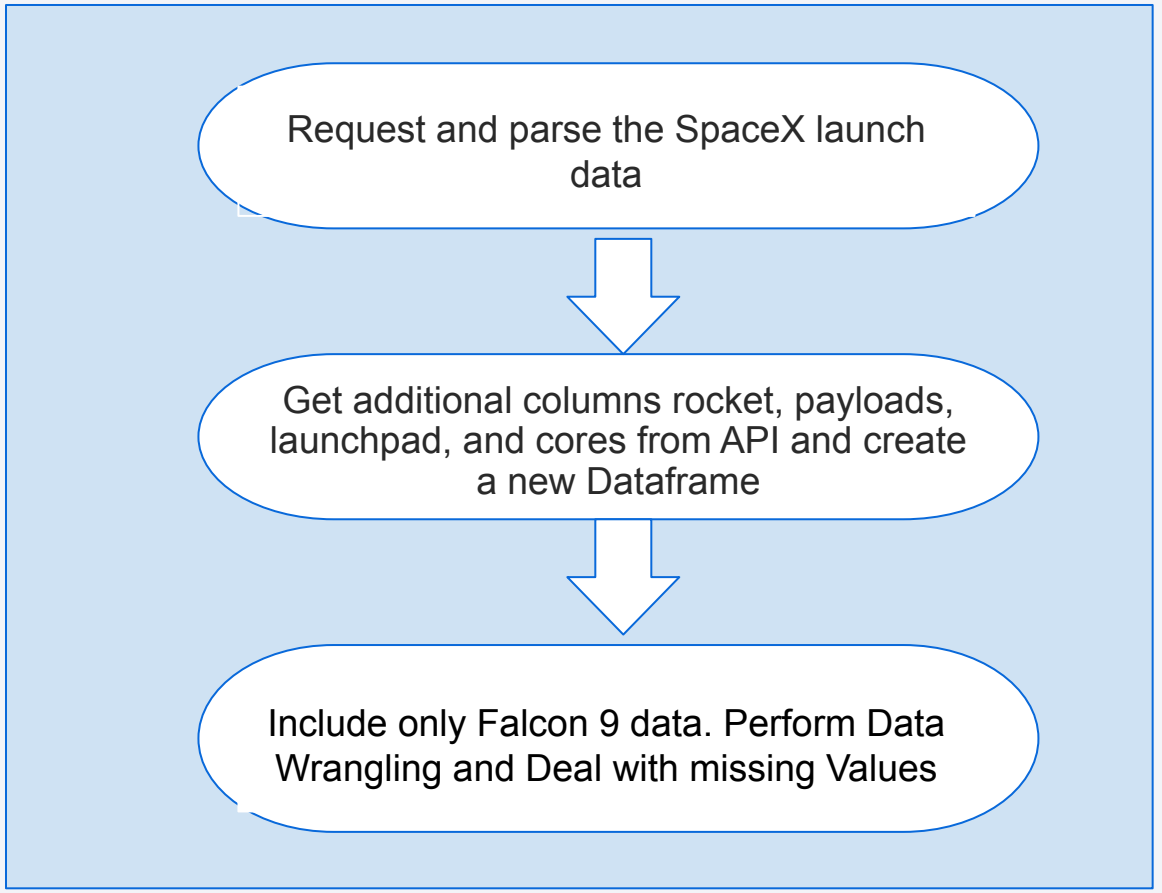
- Web Scraping from Wiki Page
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas dataframe

Data Collection – SpaceX API

SpaceX REST API Data Collection

- Request and parse the SpaceX launch data using the GET request
- Decode the response content as a Json and turn it into a Pandas dataframe
- We can use the API to get information about the launches and will be using columns rocket, payloads, launchpad, and cores
- The data from these requests will be stored in lists and will be used to create a new dataframe.
- Filter the dataframe to only include Falcon 9 launches
- Perform Data Wrangling
- Deal with Missing Values

URL = Data Collection SpaceX API

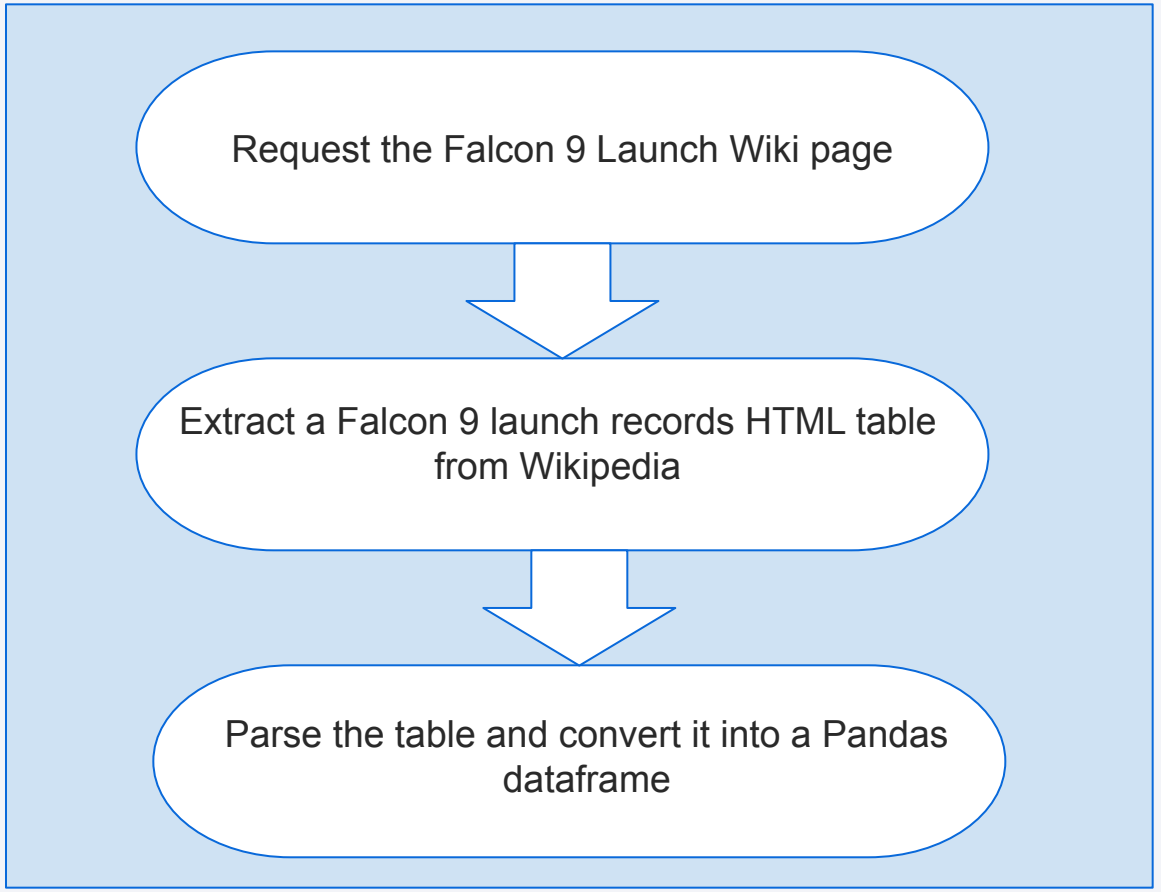


Data Collection - Scraping

Web scraping Data Collection

- Collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches
- Request the Falcon 9 Launch Wiki page from its URL using BeautifulSoup
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

URL = Data Collection Web Scraping

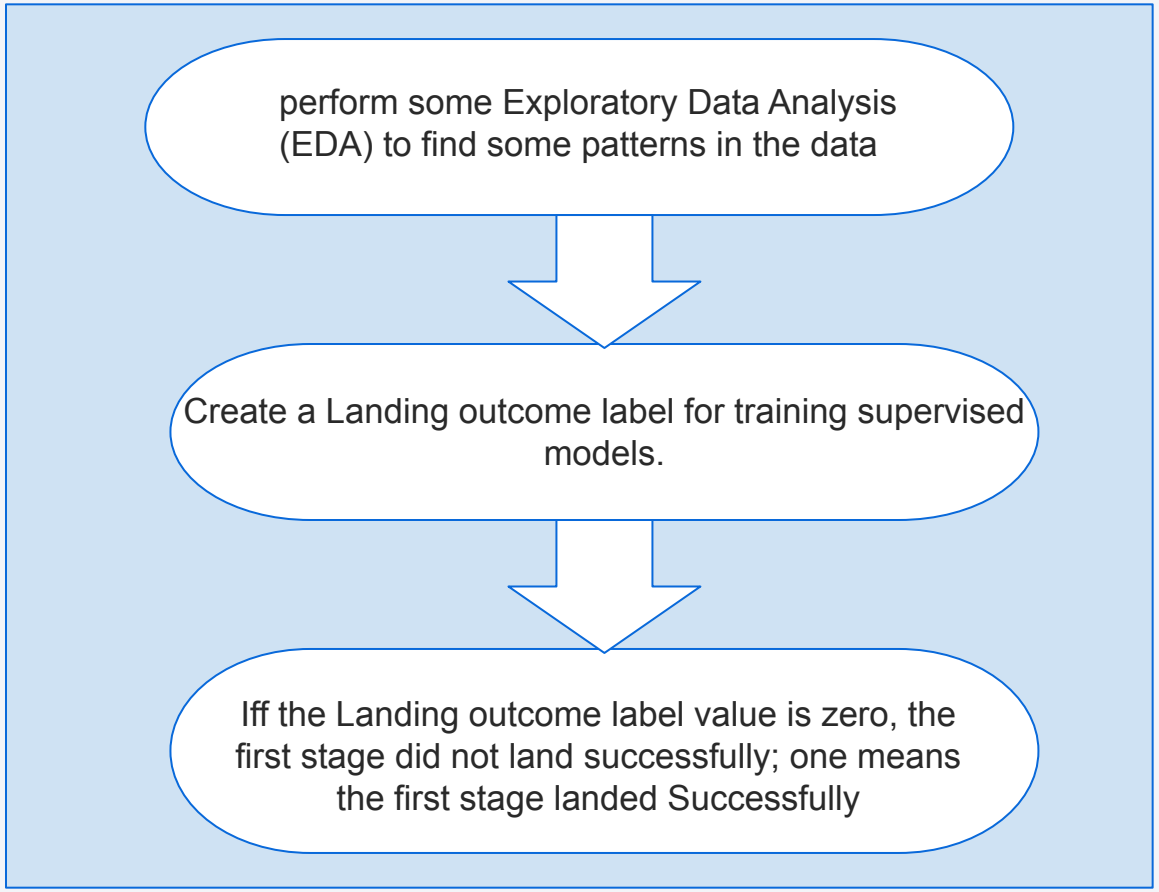


Data Wrangling

Data Wrangling

- perform some Exploratory Data Analysis (EDA) to find some patterns in the data
  - Identify and calculate the percentage of the missing values in each attribute
  - Identify which columns are numerical and categorical
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit and mission outcome of the orbits
- Determine what would be the label for training supervised models.
  - Create a landing outcome label from Outcome column.This variable will represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

URL = Data Wrangling



EDA with Data Visualization

EDA with Data Visualization

- Scatter Plot,Bar Chart,Line Chart were plotted to visualize the relationship between the features.
  - Scatter plot was created to Visualize relationship between Flight Number vs. Payload Mass, Flight Number vs LaunchSite, Payload Mass Vs. Launch Site, Flight Number Vs Orbit type,Payload Mass vs. Orbit
  - Bar chart was created to check if there are any relationship between success rate and orbit type
  - Line chart was created to visualize the launch success yearly trend
- Feature Engineering can be used to select the features that will be used in success prediction in the future Analysis.
  - Based on the above Analysis, We can determine important variable that would affect the success rate and Create dummy variables to categorical columns

URL = EDA with Data Visualization

EDA with SQL

EDA with SQL

SQL queries was created to answer the following Scenarios

- 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 and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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.

URL = EDA with SQL

Build an Interactive Map with Folium

Folium map was created with various objects added to it. These included markers representing specific locations, circles of varying colors and lines connecting different points on the map

- Markers were added to indicate the location of data points on the map.
- Circles were added to visualize data, with their radius specified in meters.
- Folium provides the PolyLine object for drawing lines

Here we created and added f**olium Circle** and **folium Marker** for each launch site on the map.
Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.
MousePosition on the map is used to get coordinate for a mouse over a point on the map.
PolyLine was used to draw a line between a launch site to its closest city, railway, highway, etc.

URL = Analytics with Folium

Build a Dashboard with Plotly Dash

Plotly Dash application was created to perform interactive visual analytics on SpaceX launch data in real-time.
This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

- **Launch Site Drop-down**
  - Drop-down includes 4 Launch Sites and all Sites Options
- Callback function to render **success-pie-chart** based on selected site dropdown
  - Based on the Launch Site dropdown selection, the Pie chart is rendered with Launch success counts
- **Range Slider to Select Payload**
  - Slider is used to find if variable payload is correlated to mission outcome.
  - we can easily select different payload range and see if we can identify some visual patterns.
- Callback function to render the **success-payload-scatter-chart**
  - we can visually observe how payload may be correlated with mission outcomes for selected sites.
  - In addition, we added color-label Booster version on scatter plot so that we may observe mission outcomes with different boosters

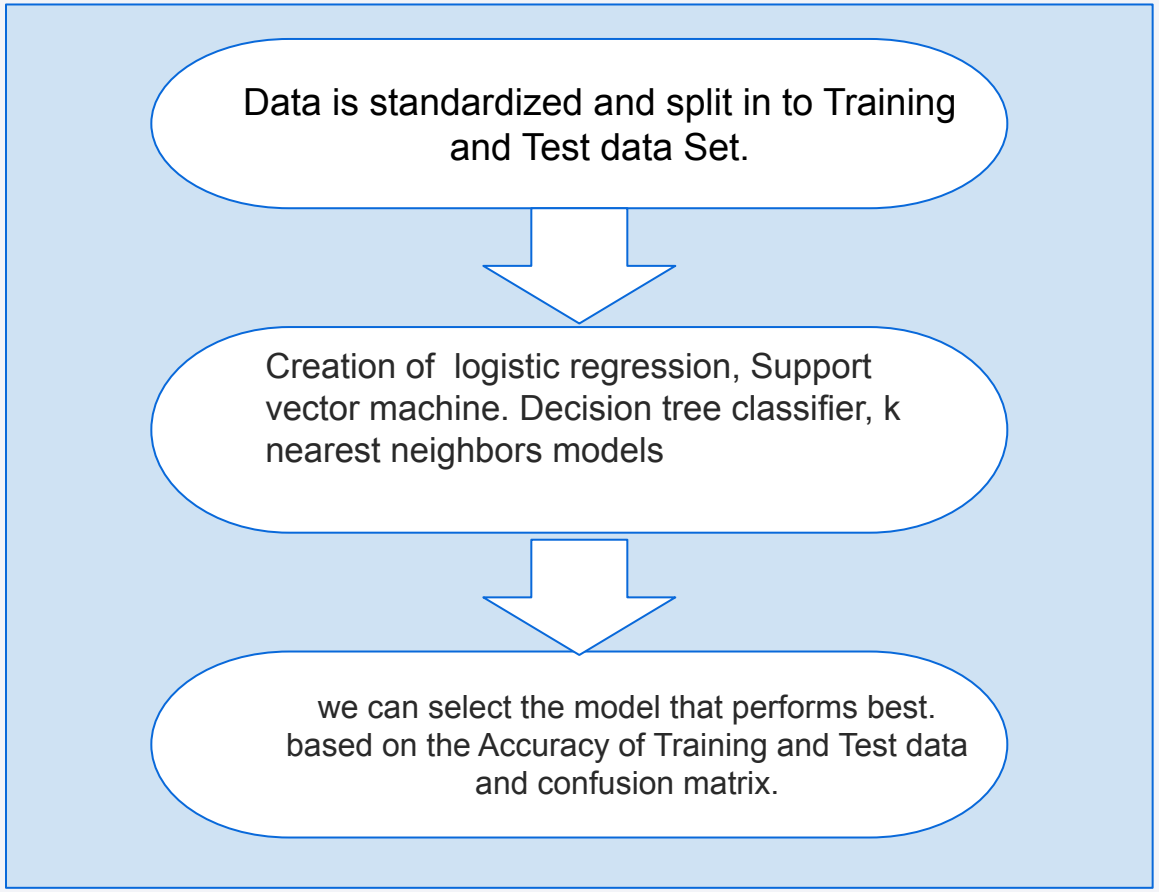
URL = [Dashboard with Plotly Dash]

Predictive Analysis (Classification)

Machine learning pipeline is created to predict if the first stage will land given the data from the preceding labs.

- Creation of NumPy array from the column Class in data
- Standardize the data
- Split the data into training and testing data using the function train_test_split
- Creation of logistic regression, Support vector machine. Decision tree classifier, k nearest neighbors models
- Best Parameters, Accuracy of Training and Test data and confusion matrix is calculated for all models.
- Based on the Results, we can select the model that performs best.

URL = [Machine Learning Prediction]

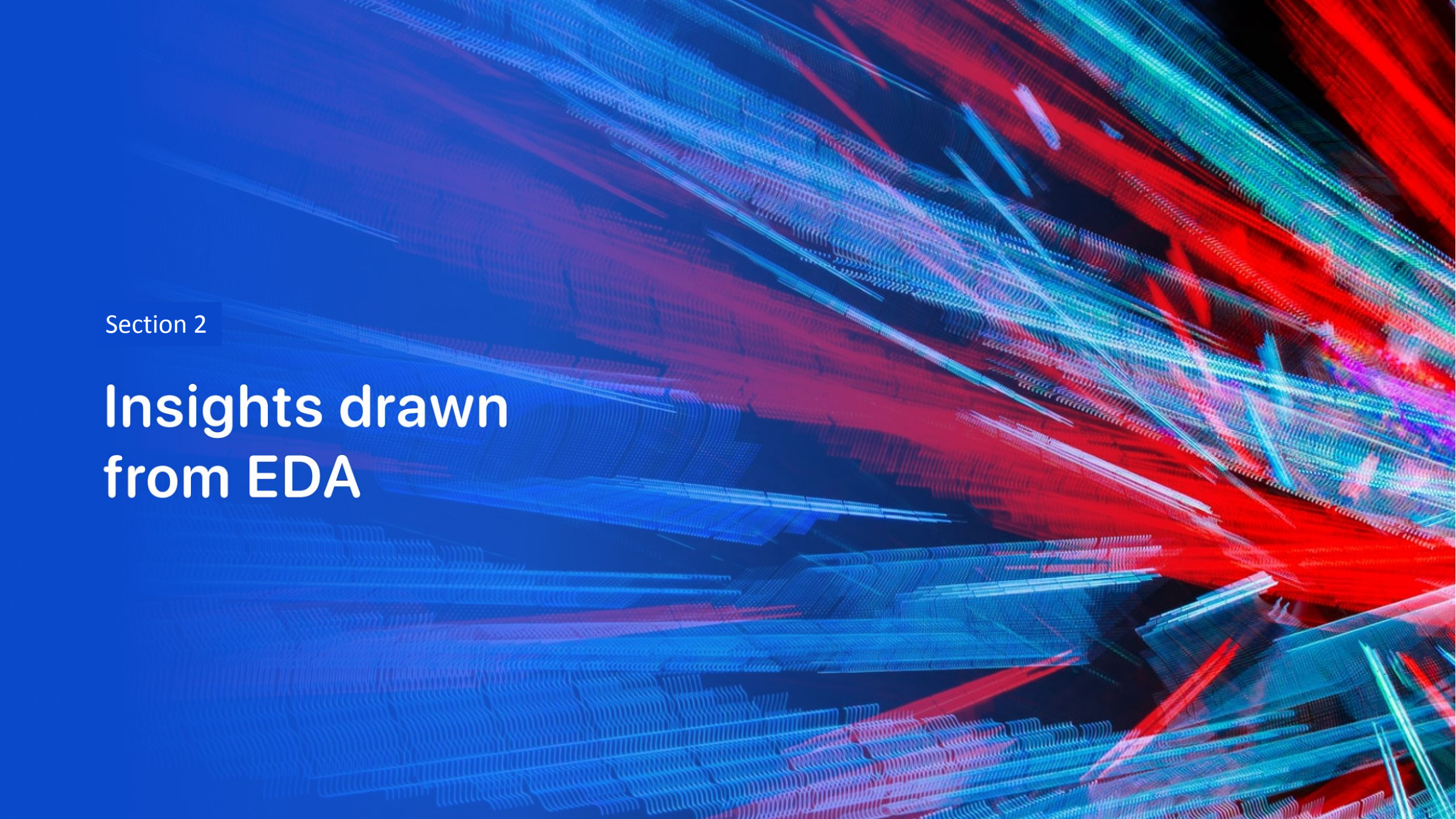


Results

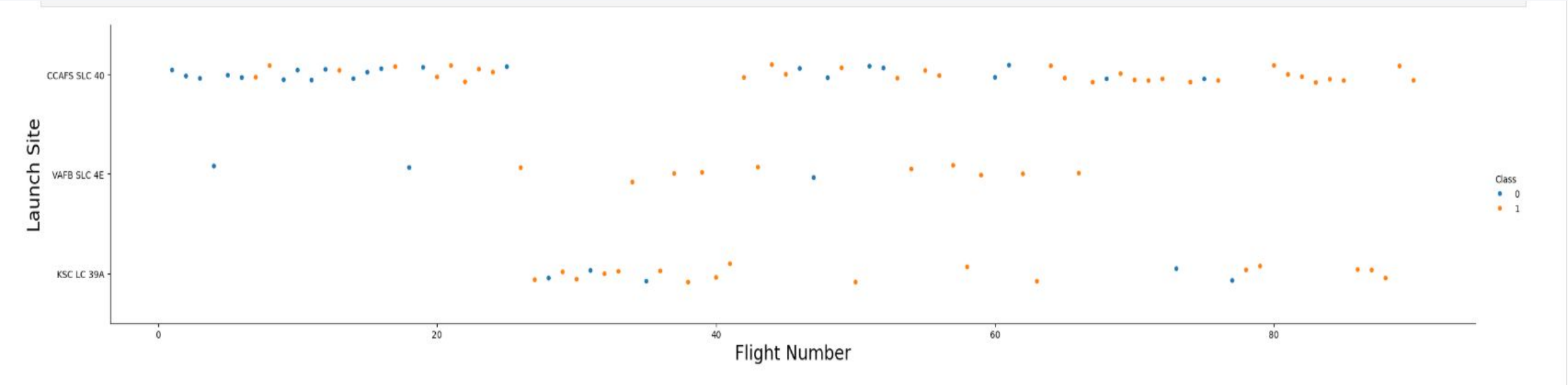
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

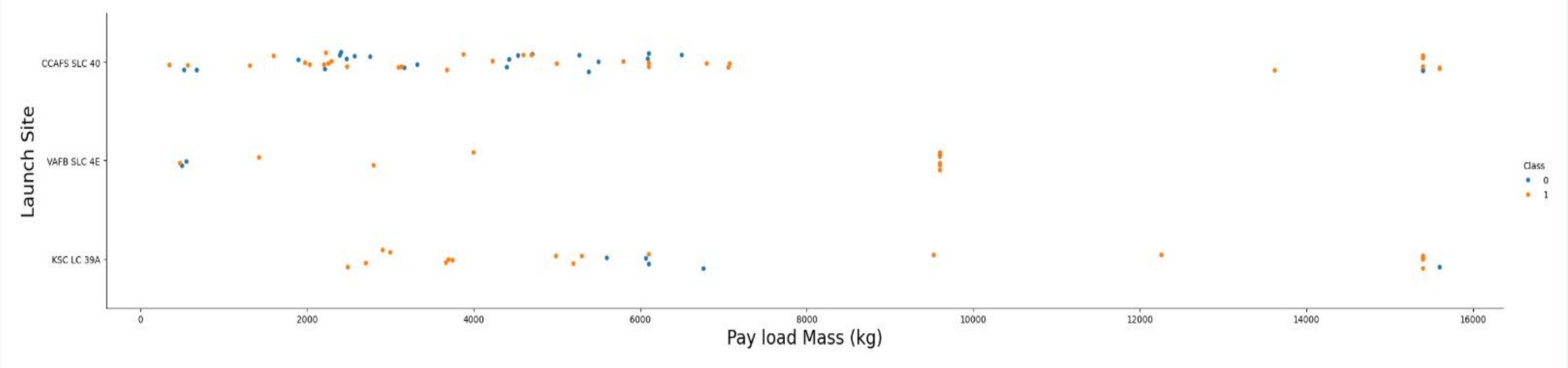


Flight Number vs. Launch Site



- Scatter plot of Flight Number vs. Launch Site
- Success Rate improved over time in most sites.
- CCAFS SLC 40 had more Launches than other sites
- CCAFS SLC 40 had successful landing on most of the latest launches

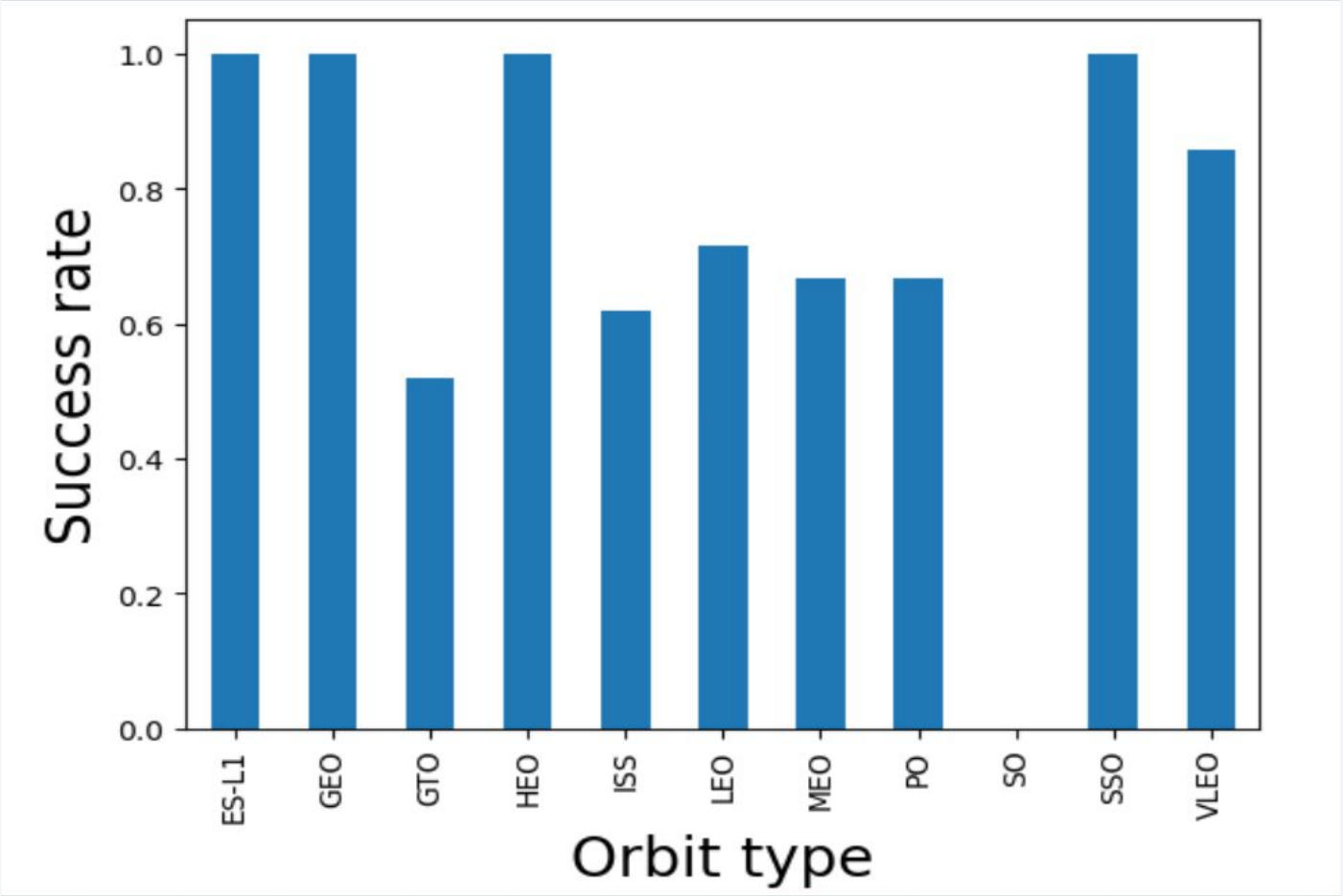
Payload vs. Launch Site



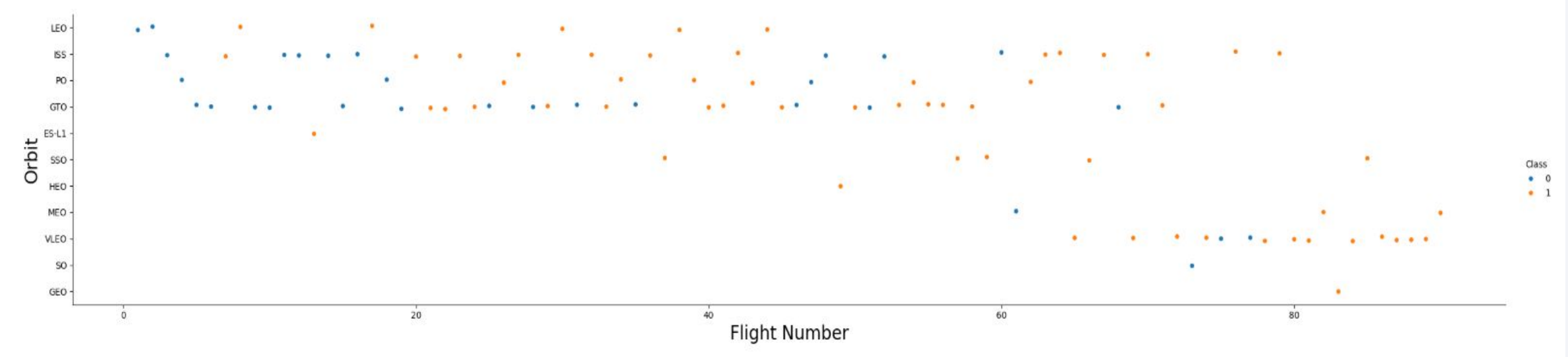
- Scatter plot of Payload vs. Launch Site
- There are no rockets launched for heavy payload mass greater than 10000 for VAFB-SLC launchsite.
- There was higher Success rate for payload mass greater than 8000.

Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO, SSO Orbits has higher Success Rate.
- SO Orbit does not have any Successful Landing

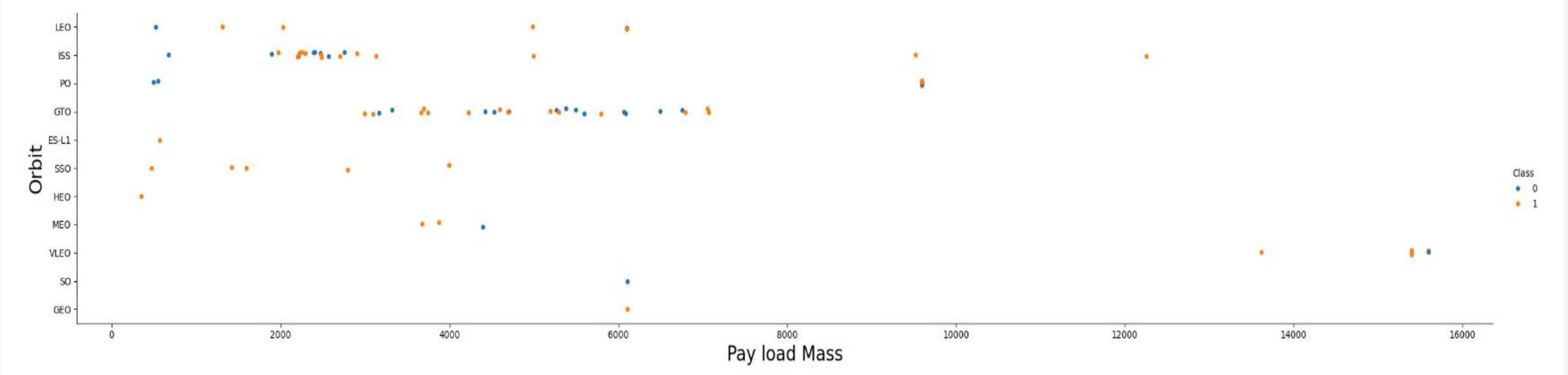


Flight Number vs. Orbit Type



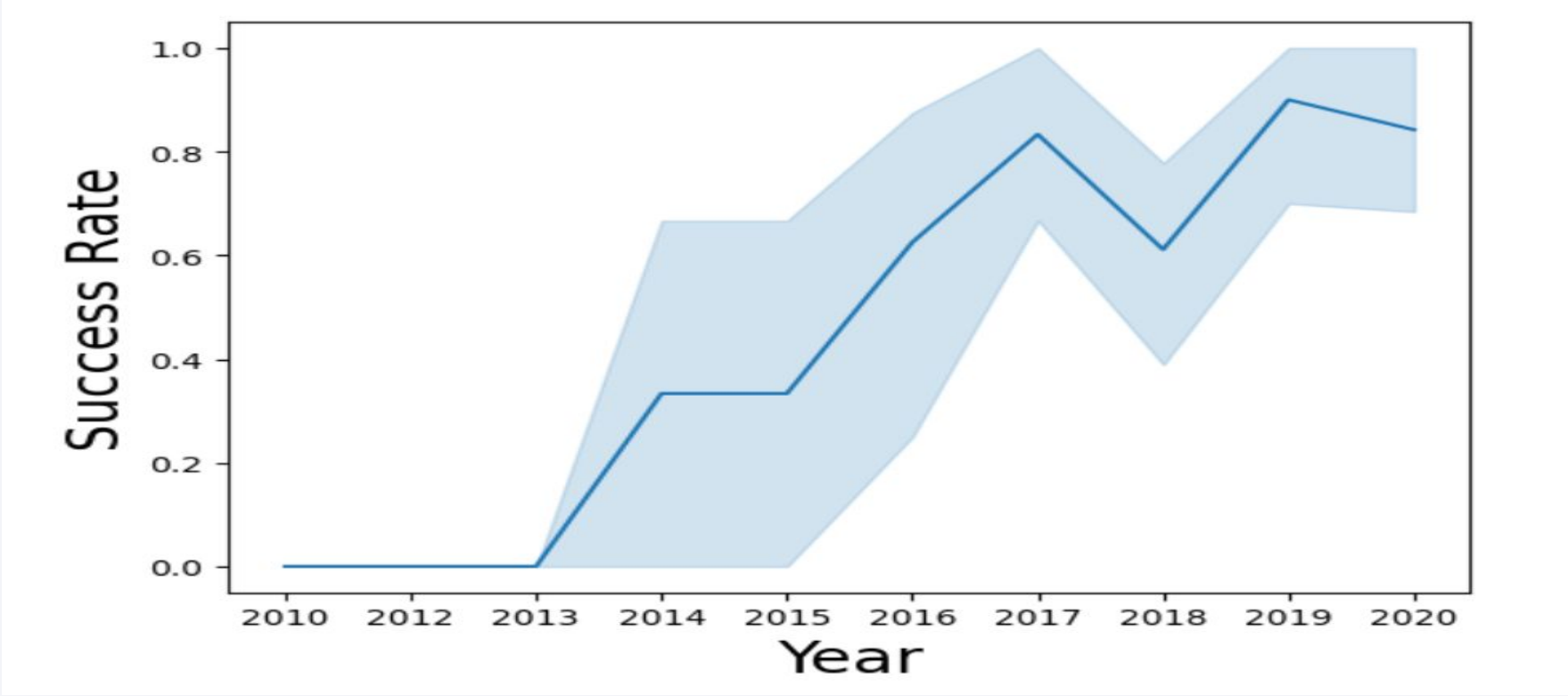
- Scatter plot of Flight number vs. Orbit type
- Success rate increased in most recent flights.
- VLEO orbit has recent increase in frequency and Success
- LEO orbit success seems to be related to the number of flights

Payload vs. Orbit Type



- Scatter plot of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for Po, LEO and ISS.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- Line chart of yearly average success rate
- Success rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites.
  - Distinct Keyword is used in sql query
  - There are 4 Unique Launch Sites

```
%sql Select distinct(launch_Site) from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Query use where,like & limit clause

```
%sql Select * from SPACEXTABLE where launch_Site like 'CCA%' Limit 5 ;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parac
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 (parac
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Query use Sum function and where clause
- Total Payload Mass is 45596 KG

```
%sql Select Sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)' ;
```

```
 * sqlite:///my_data1.db
Done.
```

Sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query use Avg function and where,like clause
- Average Payload is 2928.4

```
%sql Select Avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Query use Min function and where clause
- First successful ground landing date is 2015-12-22

```
%sql Select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query use where clause
- There are 4 successful drone ship landing with payload between 4000 and 6000

```
%sql Select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' AND  (PAYL(
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Query use Count function and group by clause
- There are 100 Successful and 1 Failure mission outcome

```
%sql Select Mission_Outcome, COUNT(Mission_Outcome) from SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
)one.
```

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- Query use where clause and Max function in subquery
- List of boosters carrying Maximum payload are displayed

```
%sql Select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ IN ( SELECT max(PAYLOAD_MASS__KG_
```

```
 * sqlite:///my_data1.db
Done.
```

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query use substr function and where clause
- There are 2 failed landing outcomes in drone ship in year 2015

```
%sql Select substr(Date, 6,2) as month, Landing_Outcome,Booster_Version, launch_site  from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

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
- Query use count function and where,between,group by and order by clause

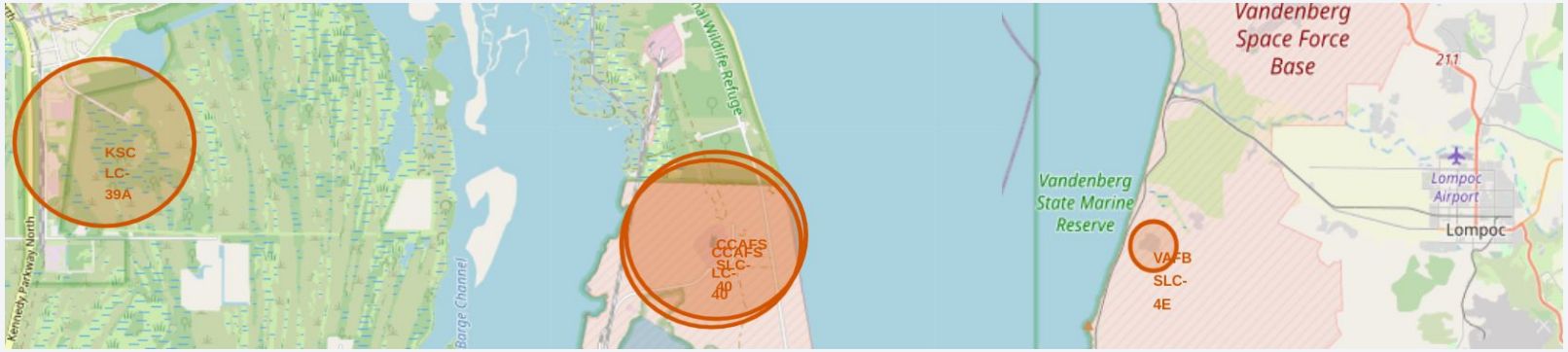
```
%sql Select Landing_Outcome,count(Landing_Outcome) as landingOutcomeCount from SPACEXTABLE where Date b
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	landingOutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

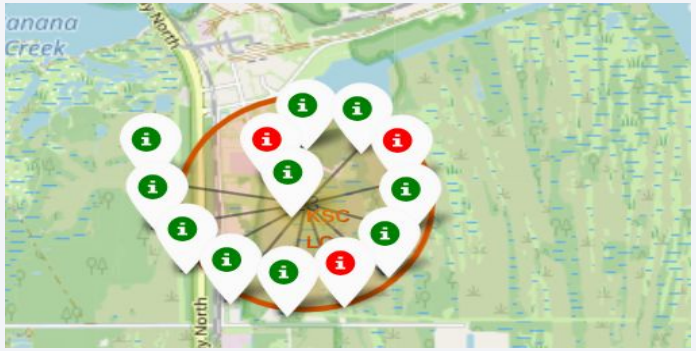

Section 3
Launch Sites
Proximities Analysis

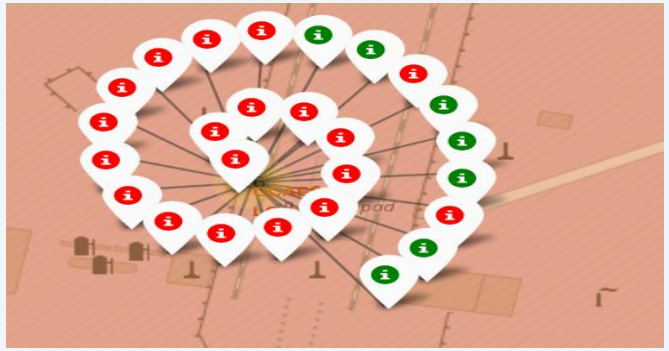
All Launch Sites in Map



All the 4 launch site are displayed in map. These sites are closer to Coast and can be accessed from land also by Road and Railways.

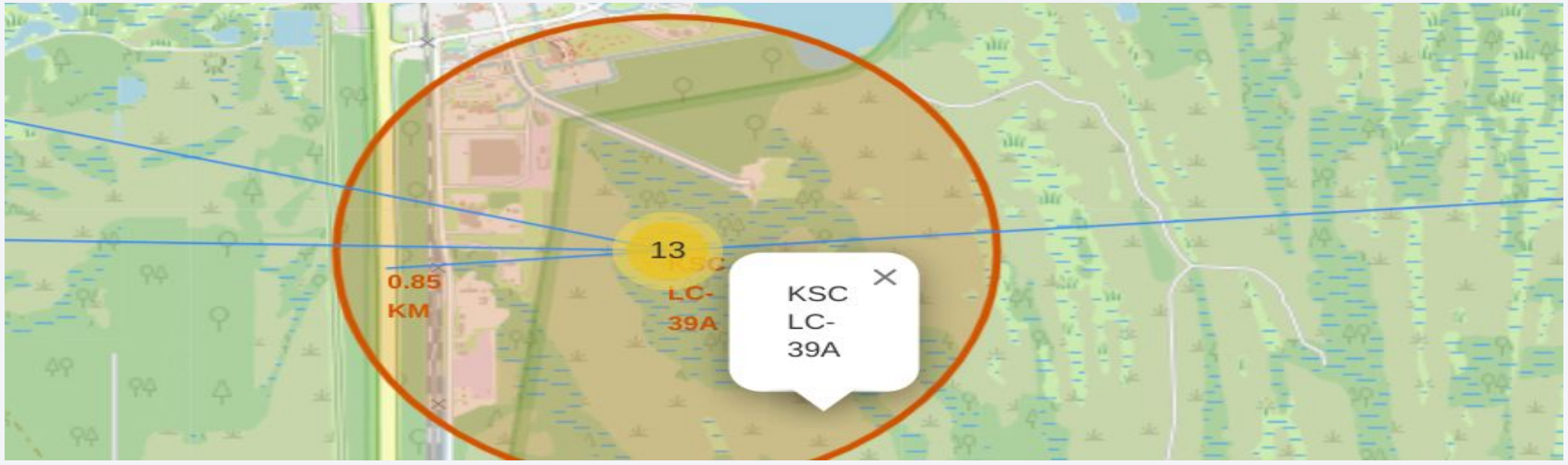
Success and Failure Outcome for all site and Launches





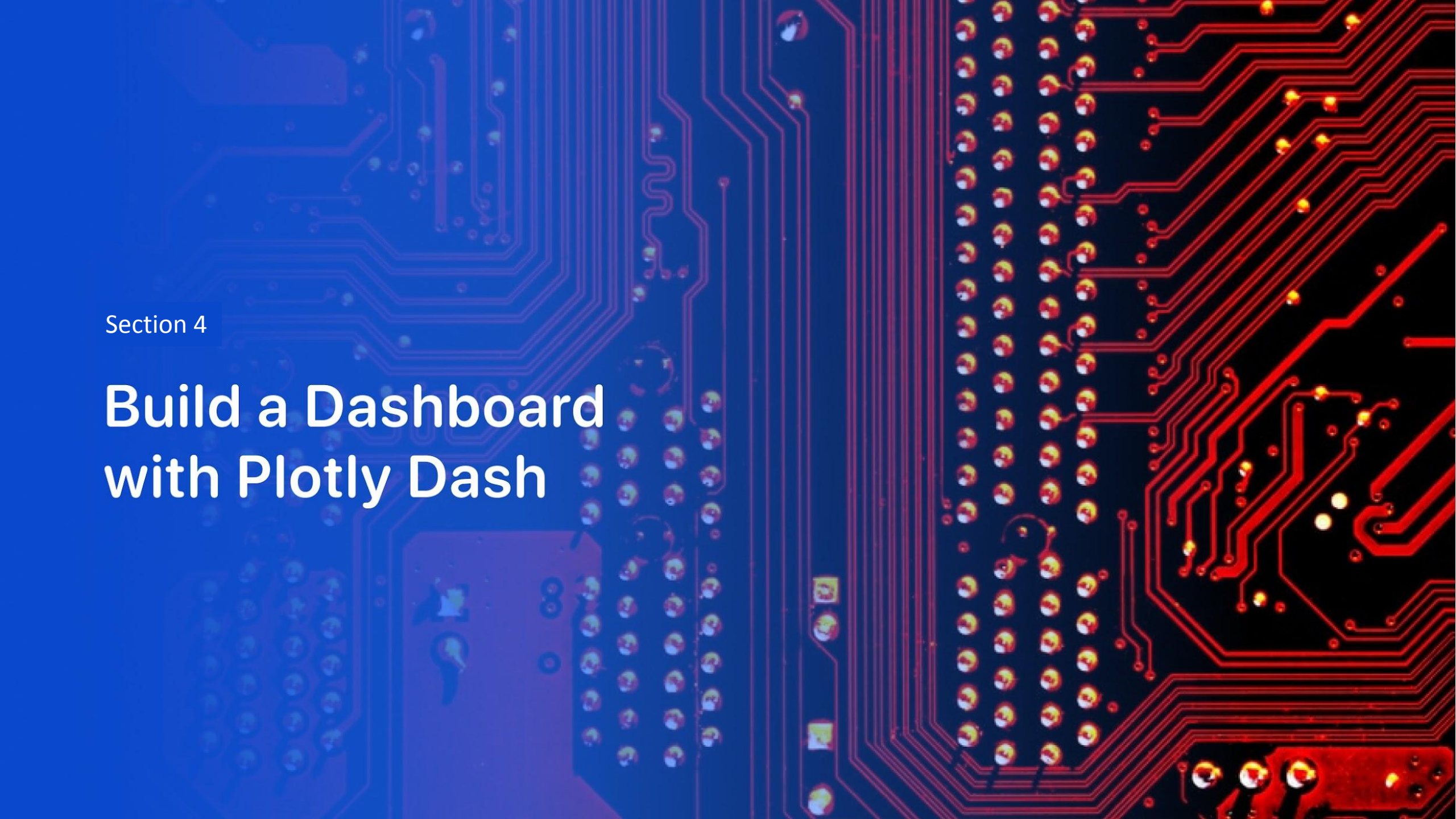
- Top image shows the outcome for sites.
- Bottom image shows clusters for few sites.
- Green marker shows that the launch is successful and Red marker shows the launch is Unsuccessful

KSC LC-39A Launch site and its proximity to city, railway, highway & coastline

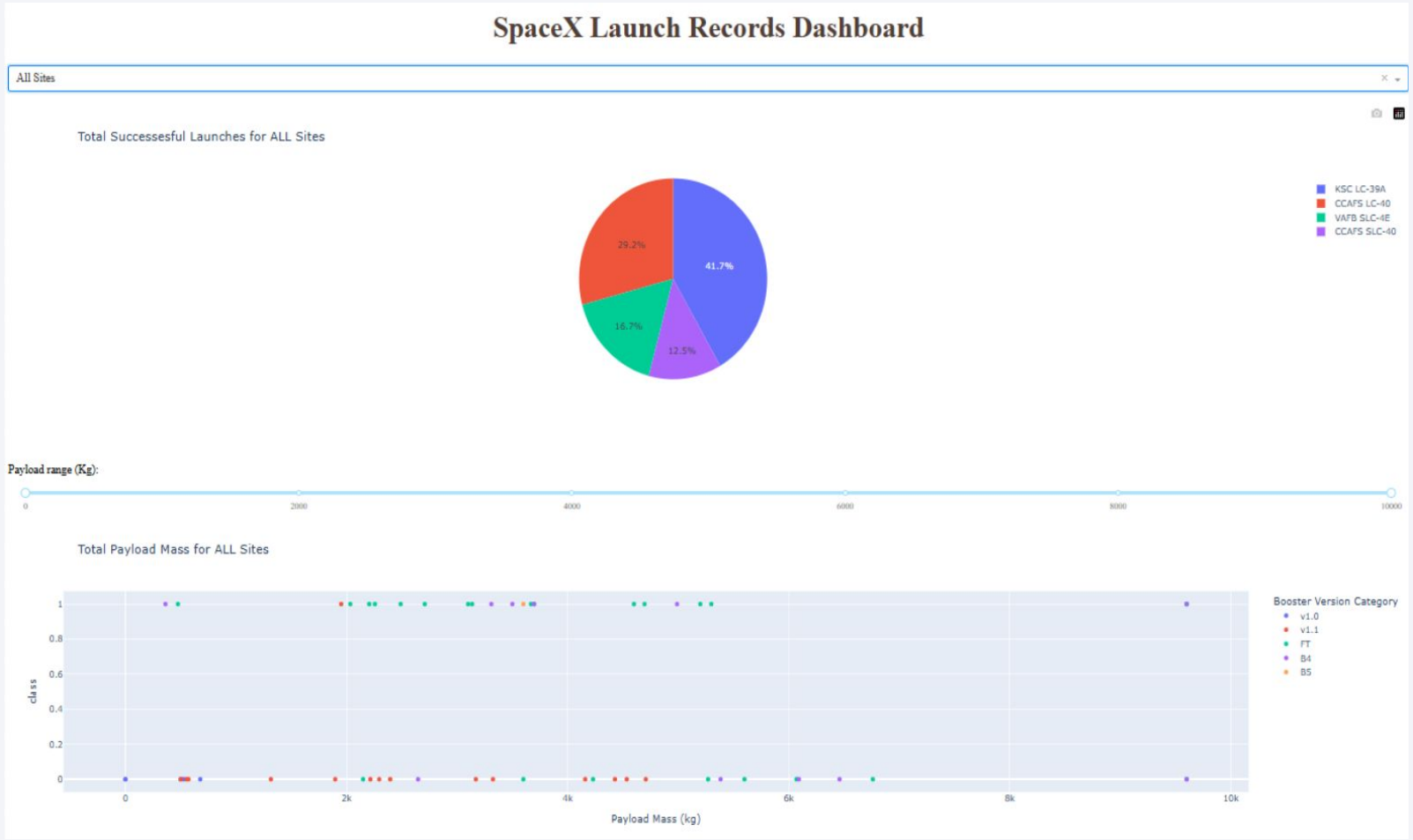


Map shows KSC LC-39A Launch site and its proximity to city, railway, highway & coastline

- Blue line is drawn in between a launch site to its closest city, railway, highway and city.
- Distance is displayed in KM at the end of the line.
- Launch site is in close proximity to
  - Railway(15.53km)
  - Highway(0.85km)
  - City(18.38km)
  - Coast(7.28km)


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



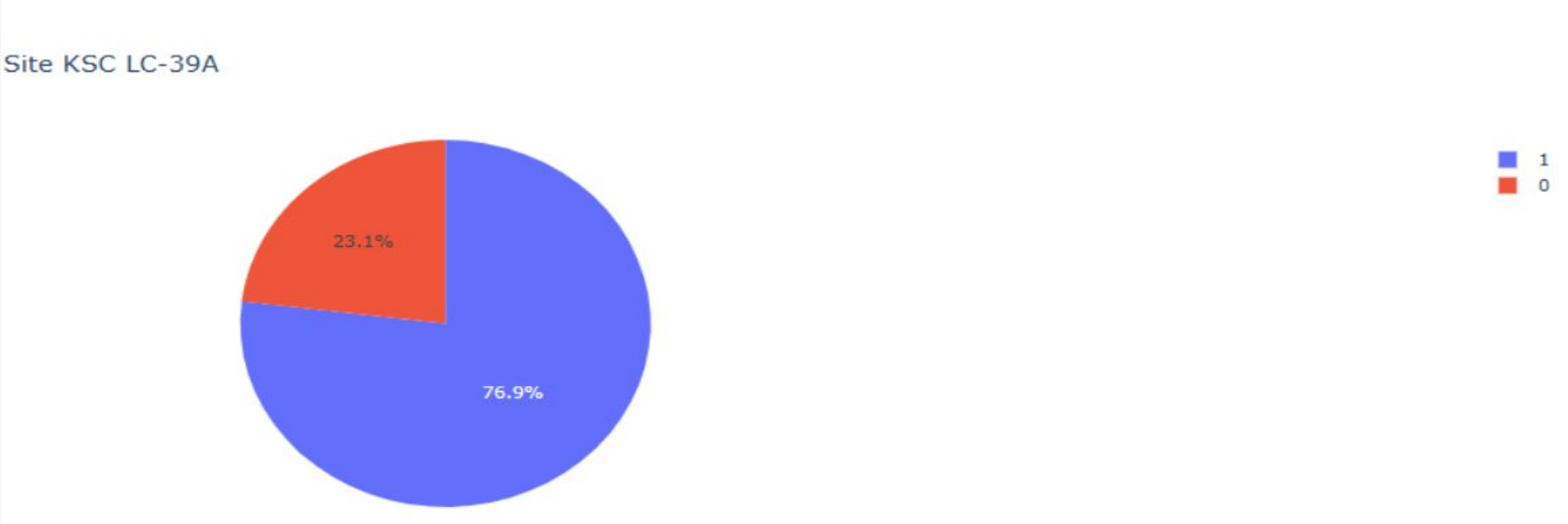
Dashboard includes

- Launch Site Dropdown
- Launch Success Pie chart
- Payload mass Slider
- Payload mass scatter plot

Launch Success percentages for all Sites

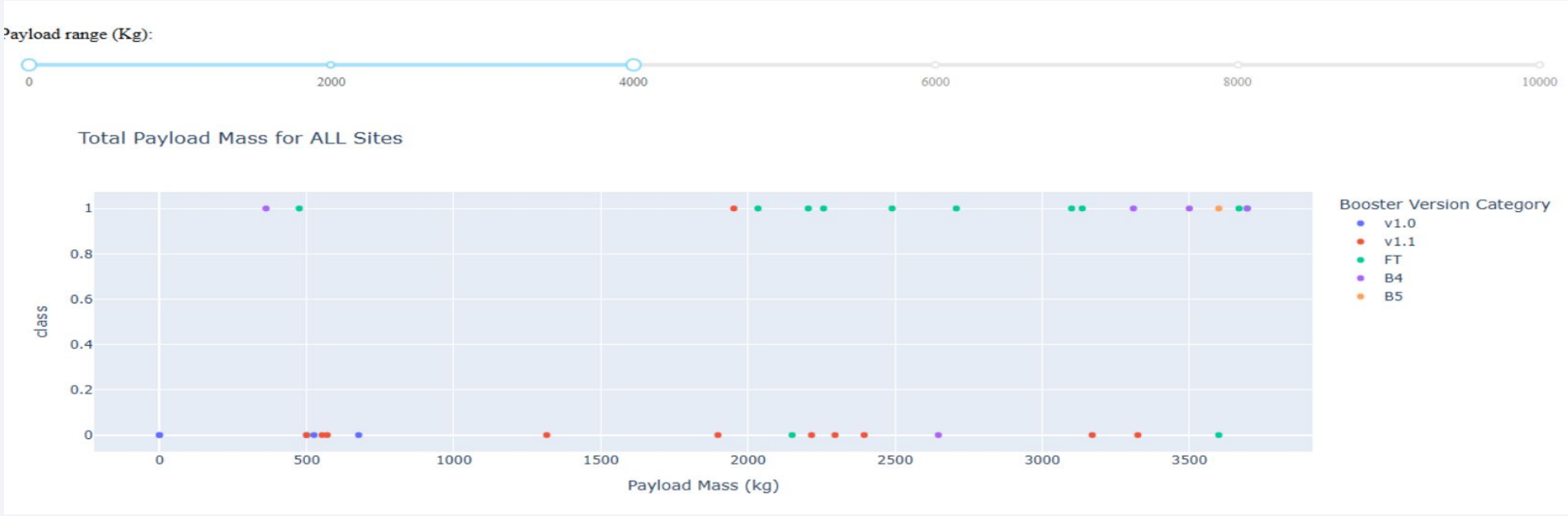
- Pie chart shows the successful launch percentage of all the 4 sites
- KSC-LC-39A Launch site has high success launch of 41.7%

Launch site with highest launch success ratio



- Pie chart shows the launch site with highest launch success ratio
- KSC LC 39A has the high success percentage of 76.9

Payload vs. Launch Outcome scatter plot for all sites



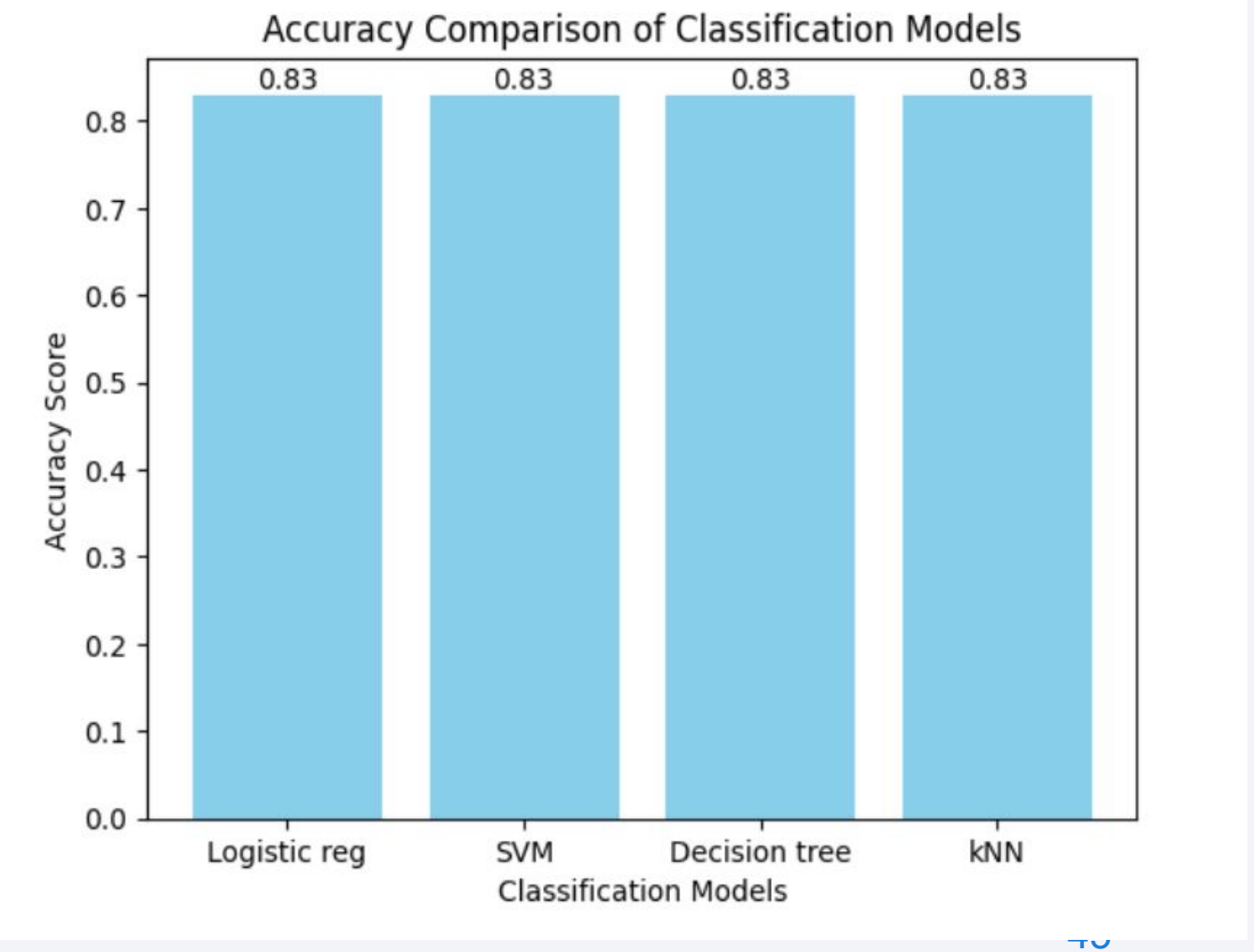
- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider(0 to 4000)
- FT booster version has a more successful between Payload mass of 2000 to 3500

Section 5

Predictive Analysis (Classification)

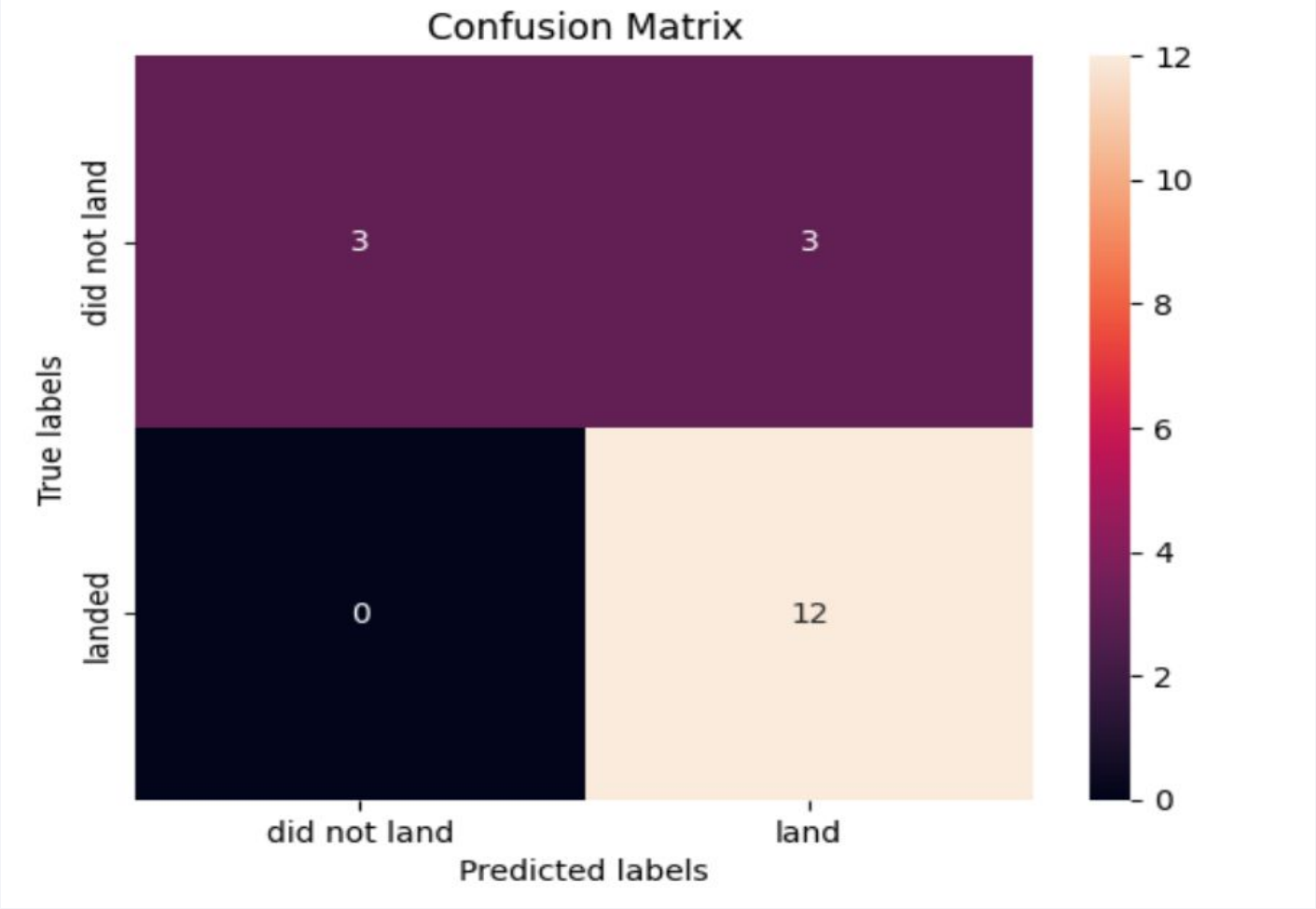
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- All the classification model [Logistic Regression,Support Vector Machine, Decision Tree, K Nearest Neighbors] have the same accuracy of 83%.



Confusion Matrix

- **Confusion matrix** of the best performing model.
- Confusion matrix is same for all the classification models.
- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.
- Overview:
  - True Positive - 12 (True label is landed, Predicted label is also landed)
  - False Positive - 3 (True label is not landed, Predicted label is landed)



Conclusions

Conclusion

- Data Collection, Data Wrangling , EDA using SQL and Visualization,Analytics using Folium and Plotly Dash process was completed to clean & understand the data and also to be used for Machine Learning.
- Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors models was developed.
- HyperParameters,Accuracy and confusion matrix is used to find the best model.
- All the model has same accuracy so they all perform the same
- By using the machine learning model, we can predict if the first stage will land.We can also determine the cost of the launch.
- Based on the model, **SpaceX will successfully land the first stage with 83%**

Appendix

logistic regression :
tuned hyperparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Test Data accuracy : 0.8333333333333334

support vector machine :
tuned hyperparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
Test Data accuracy : 0.8333333333333334

decision tree classifier :
tuned hyperparameters :(best parameters) {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}
Test Data accuracy : 0.8333333333333334

k nearest neighbors :
tuned hyperparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Test Data accuracy : 0.8333333333333334

Github Project Repository Link : Applied Data Science SpaceX Project

Thank you!