IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ABDULRAHMAN ALKOJAK ALMASRI

1 April. 2023

Outline

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

Executive Summary

- In this project, we analyzed SpaceX Falcon 9 first stage landing.
- Data Collection performed using SpaceX's API and using webscarping methods.
- Preliminary insights about the data have been derived during EDA process.
- Further visualization techniques were used to geographically determine landing sites.
- Potential features can be used to train ML models, we obtained accuracy of 92%.

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.

Section 1

Methodology

Methodology

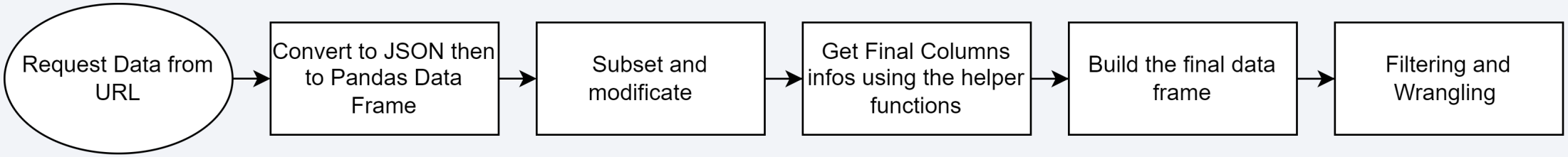
Executive Summary

- Data collection methodology:
  - Using SpaceX API.
  - Using Webscarping.
- Perform data wrangling
  - Using pandas library and orbits, booster versions information.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

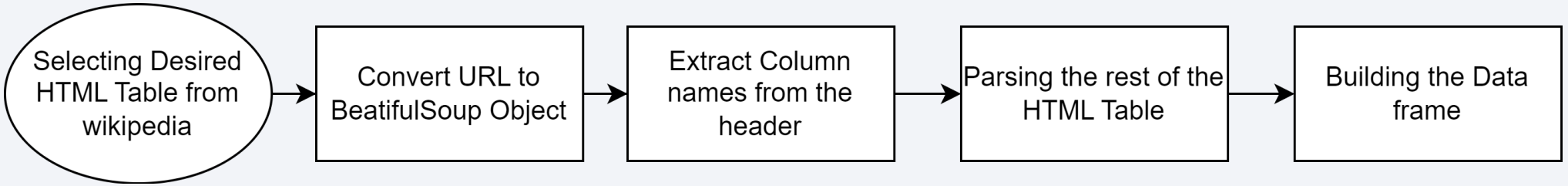
Data Collection

- Data were Collected using two methods which are described below:

Using API

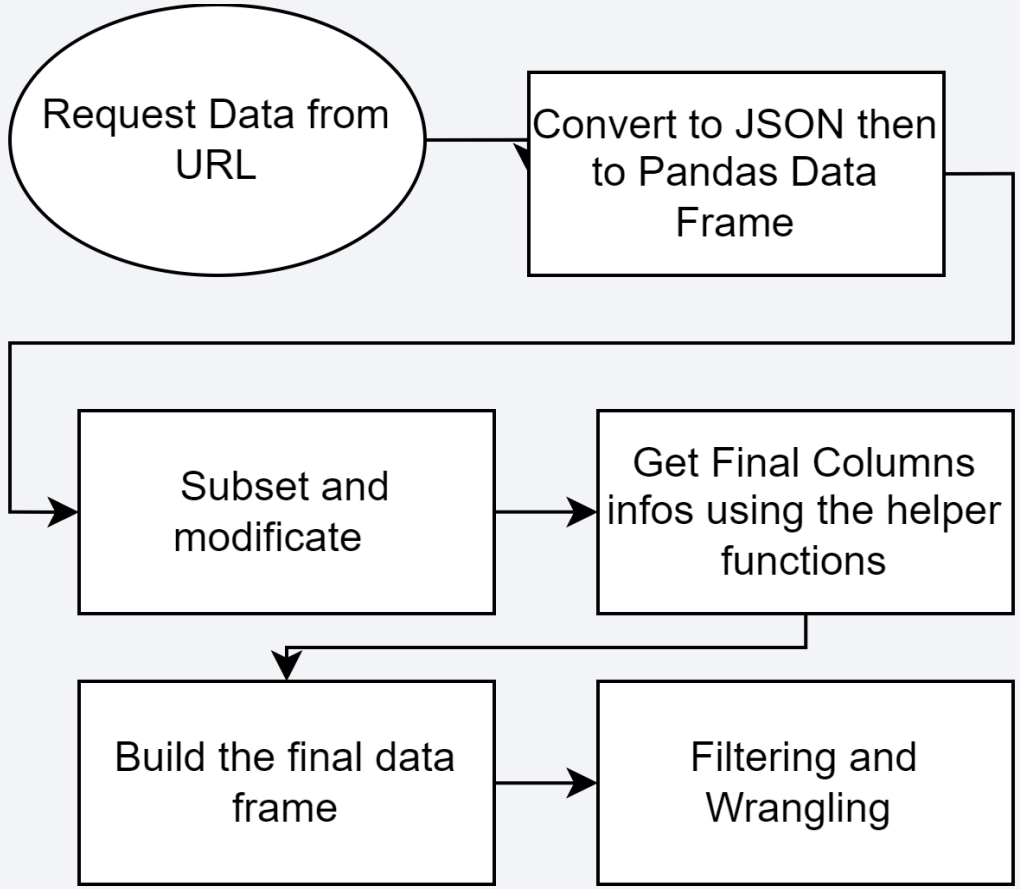


Using Webscaping (BeatifulSoup)



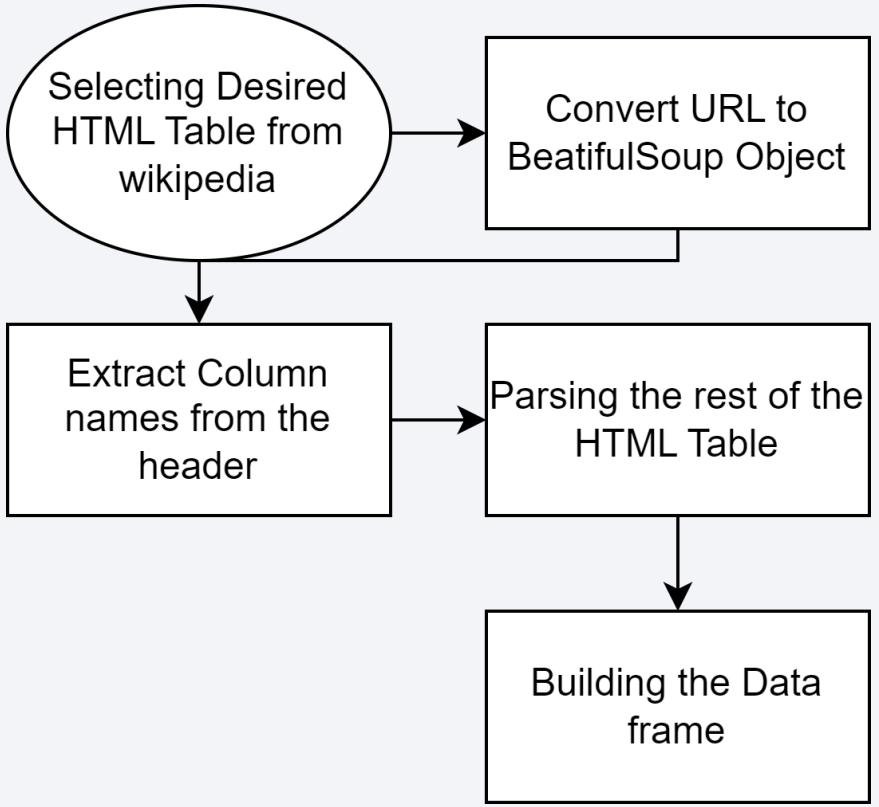
Data Collection – SpaceX API

Check out related Jupyter Notebook here



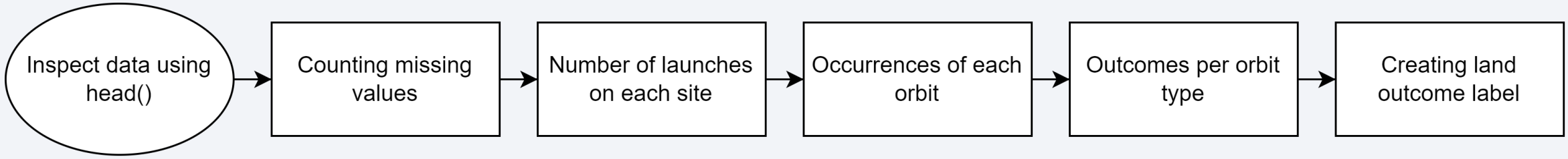
Data Collection - Scraping

Check out related Jupyter Notebook Here.



Data Wrangling

- The main objective of the data wrangling process was to determine the label column and potential, useful variables.
- First process was dealing with the missing values, and checking data types.
- EDA Summary consisted of the following: number of launches on each site, number of orbits, number of outcomes for each orbit.
- Find Jupyter Notebook Here.



EDA with Data Visualization

- Scatter plots: Pay Load Mass vs Flight Number, Launch Site vs Flight Number, Launch site vs Pay Load Mass, Orbit vs Flight Number, Orbit vs Pay Load Mass, (All of the scatter plots were with respect to the launch outcome).
- Success rate vs years , and Success rate for each orbit.
- The goal is to investigate the relationship between the variables.
- Jupyter Notebook can be found Here.

EDA with SQL

- Unique launch sites in the dataset
- 5 records for launch sites begin with CCA
- Total payload mass by booster
- Avg payload mass for booster version F9 v1.1
- Date for the first successful landing in ground pad
- Boosters succeeded in drone ship with payload mass between 4000 and 6000
- Total number of successful and failure mission outcomes
- Maximum payload mass for each booster version
- Ranking successful landings between 2010 and 2017 in descending order
- Find Jupyter Notebook Here.

Build an Interactive Map with Folium

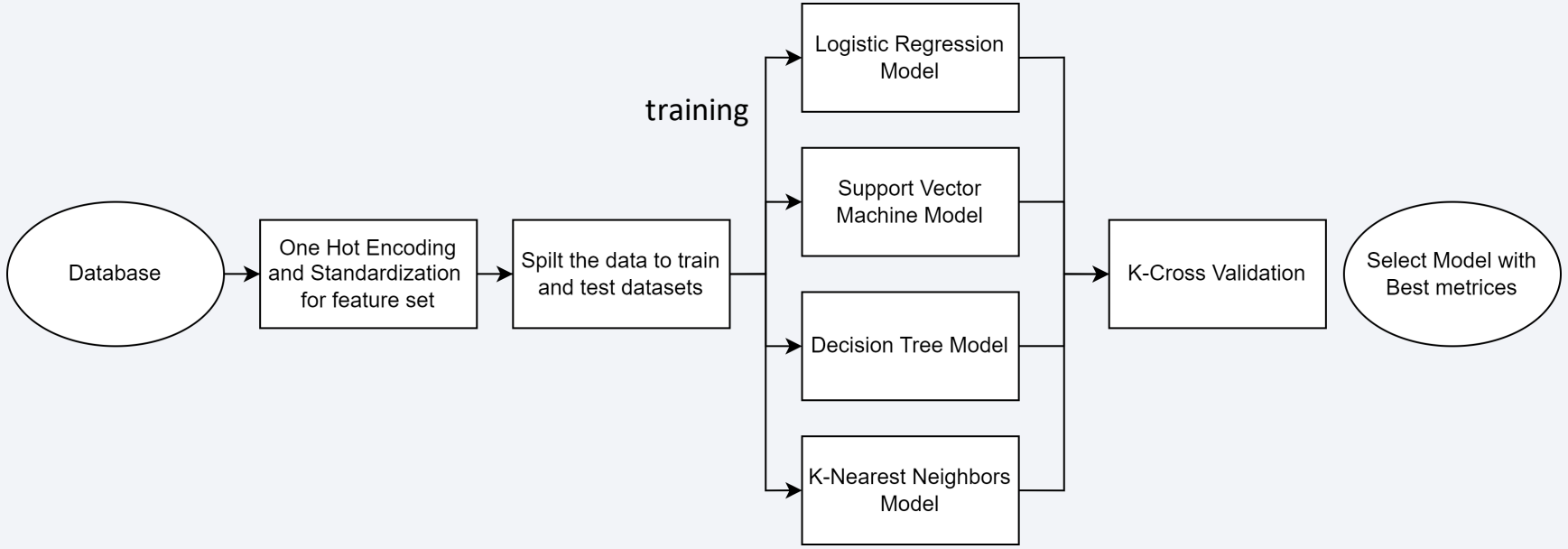
- Circles were added to better visualize the area with a labeled text.
- Marker clusters were used to simplify the process of visualizing markers having the same coordinate.
- Polylines were useful for plotting distances between launch sites and other areas.
- Jupyter Notebook can be found Here.

Build a Dashboard with Plotly Dash

- The dashboard contains two graphs (pie chart and scatter plot) with two interactions (site selection and payload mass range selection).
- The interactions facilitate visualizing booster versions and success rates with regard two the site and payload mass which are the most important features.
- Dash app code Here.

Predictive Analysis (Classification)

- The process of classification involves choosing 4 models.
- Calculating the accuracy and confusion matrix for each model.
- The evaluation process consists of splitting the data into training and testing sets, and performing K cross validation for each model to ensure that no overfitting will occur.
- Jupyter Notebook can be found Here.

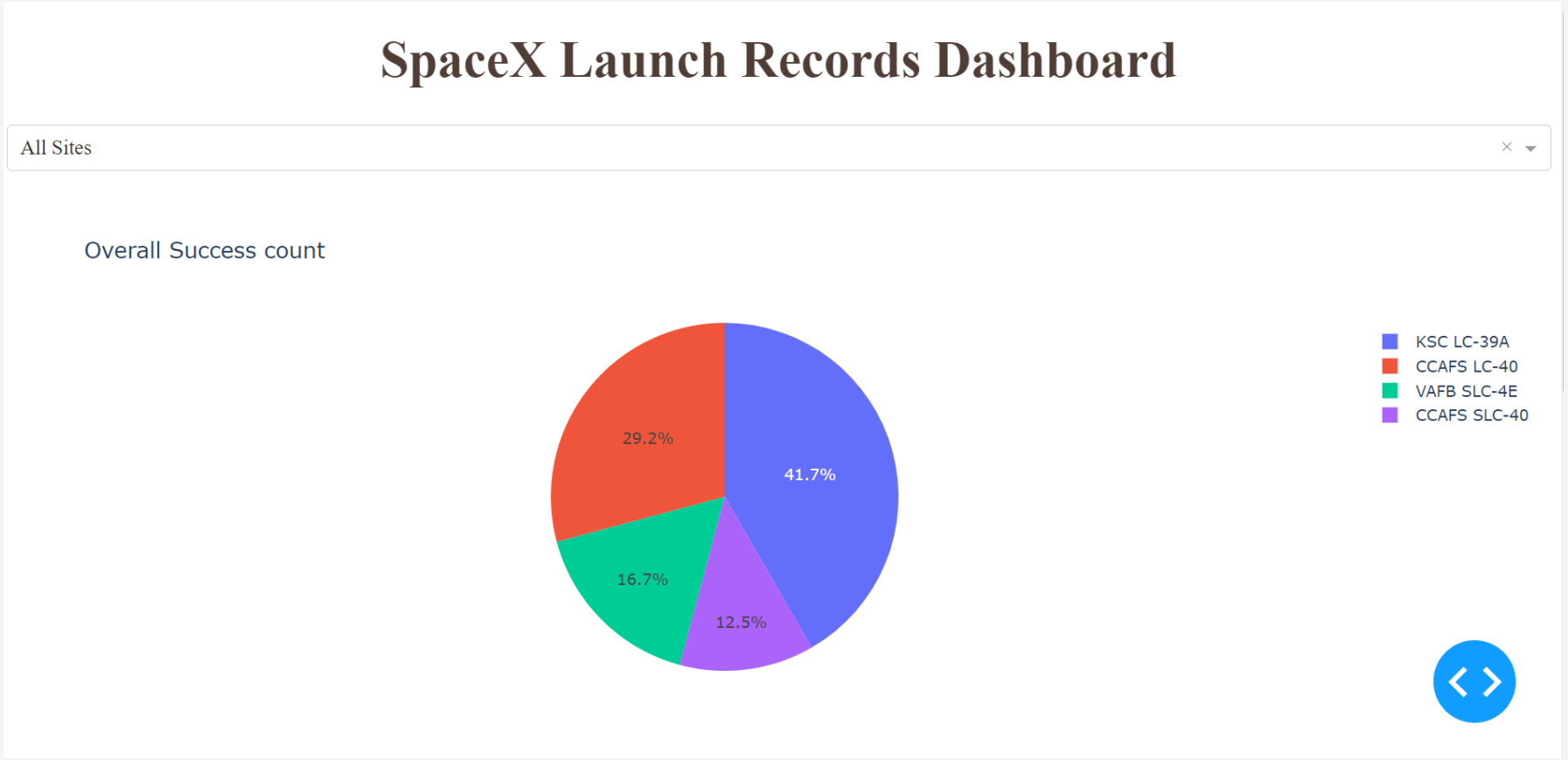


Results

- Exploratory data analysis results:
  - We can consider some variables as Payload mass, Booster version, and Launch site as the most important variables.
  - The relationship between variables are somehow unclear and require obtaining relation between two or more variable at the same time.
  - Some Orbits like SSO has better landing results than the others.
  - The Higher the Payload, The Higher possibility of successful landing.

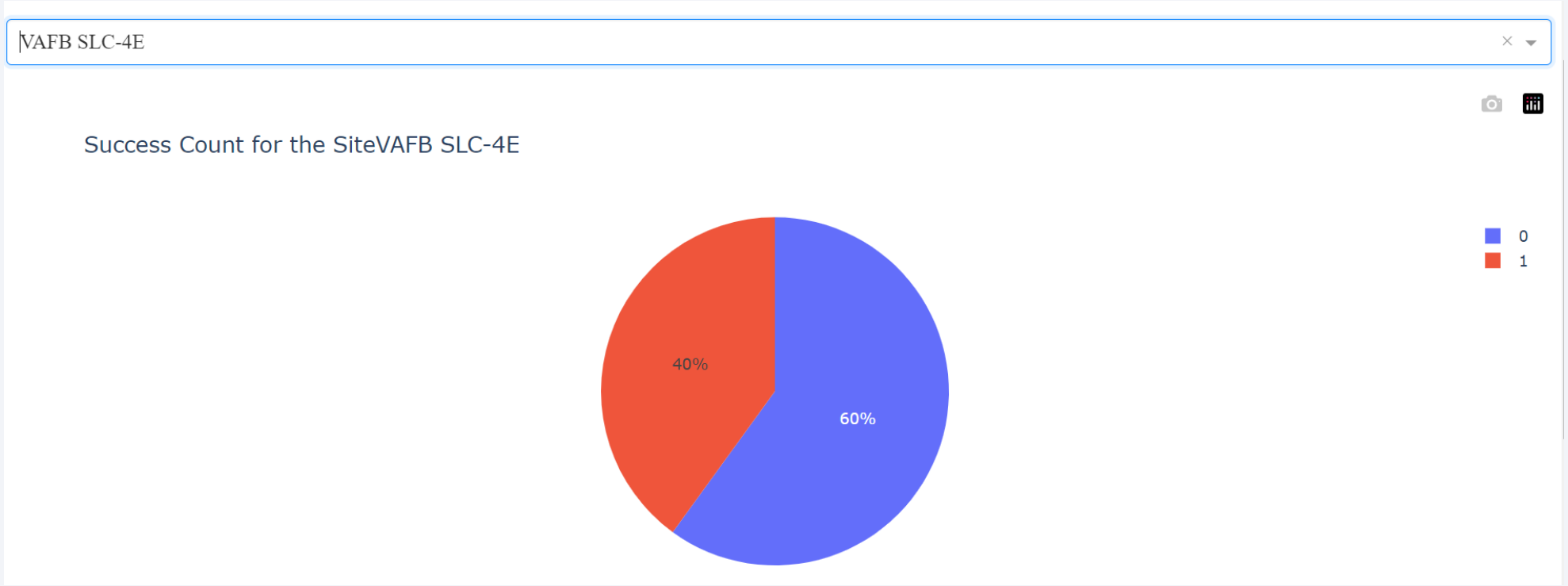
Results

- Interactive analytics demo in screenshots



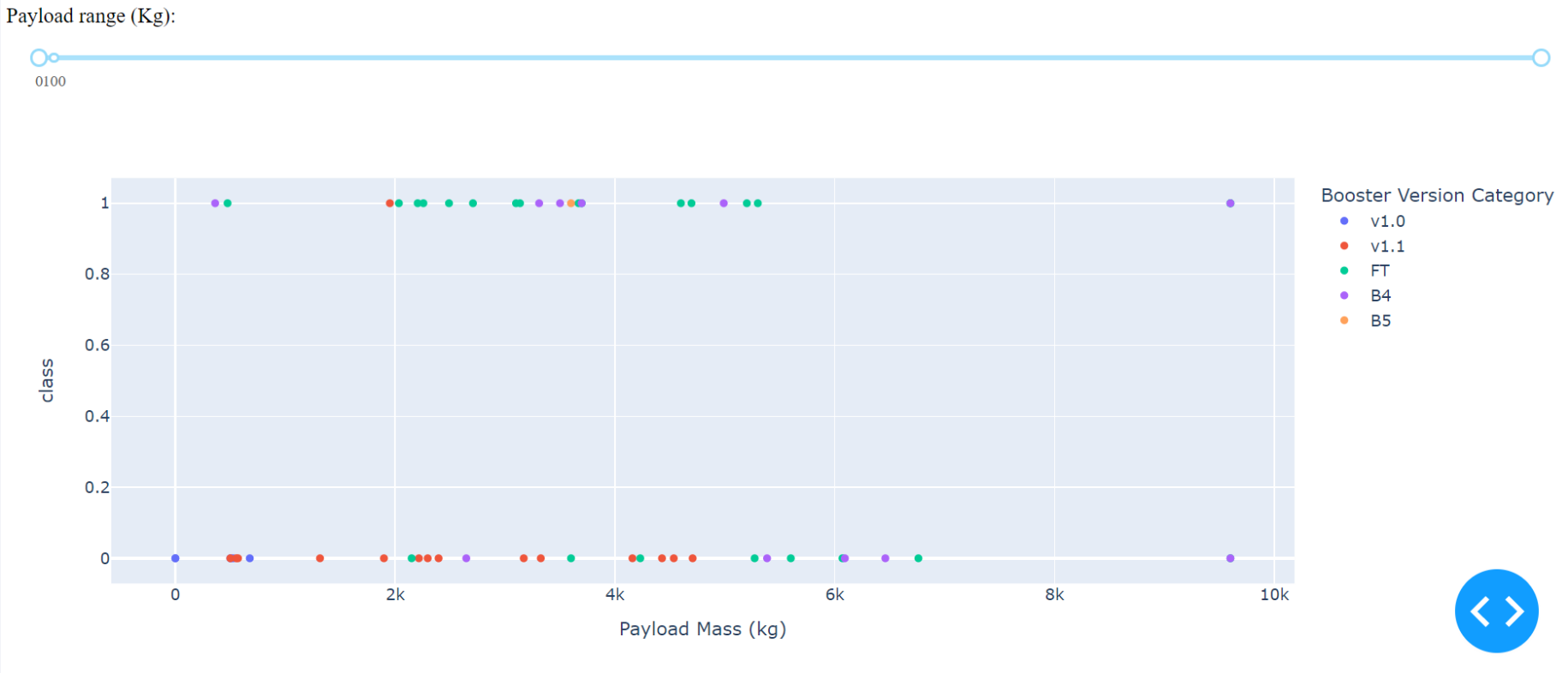
Results

- Interactive analytics demo in screenshots



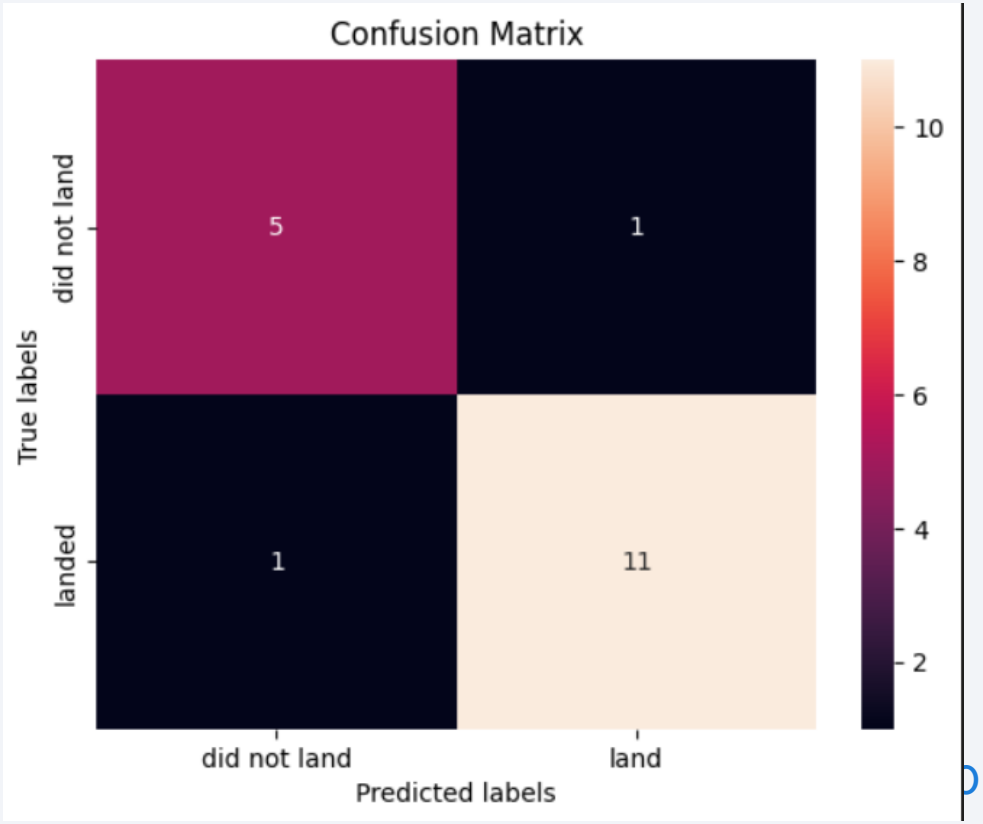
Results

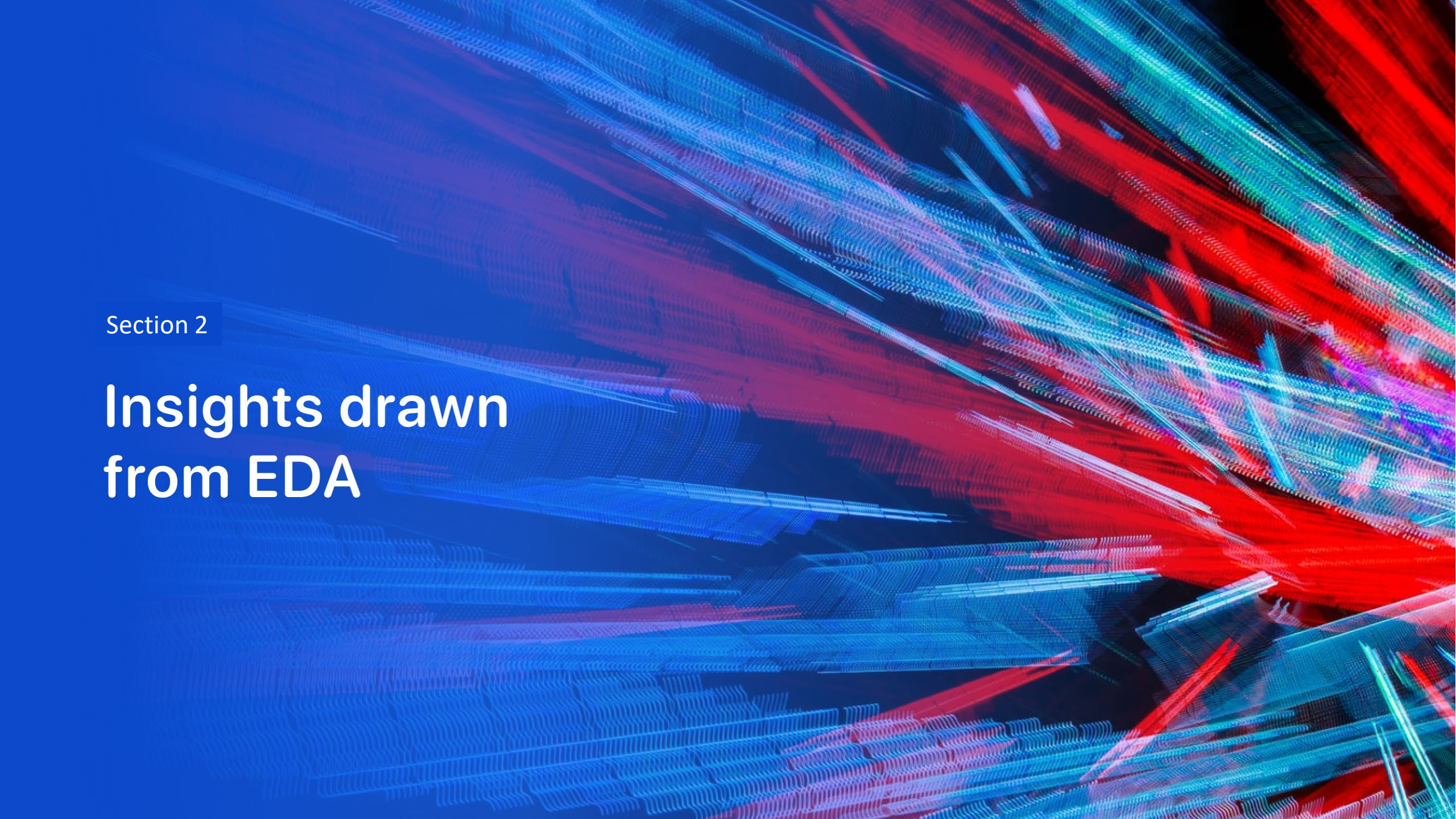
- Interactive analytics demo in screenshots



Results

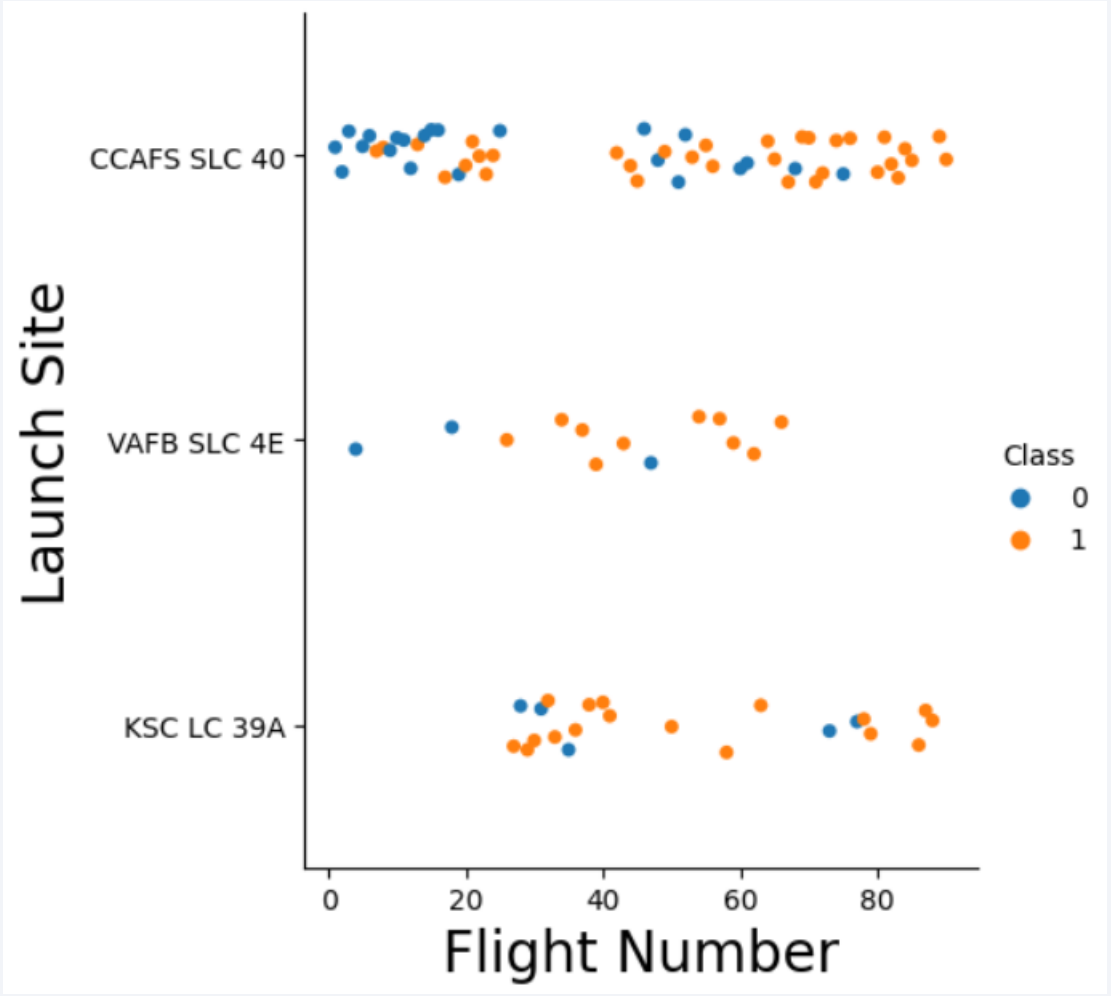
- Predictive analysis results:
  - Decision Tree the best performing model with 88% accuracy




Section 2
Insights drawn from EDA

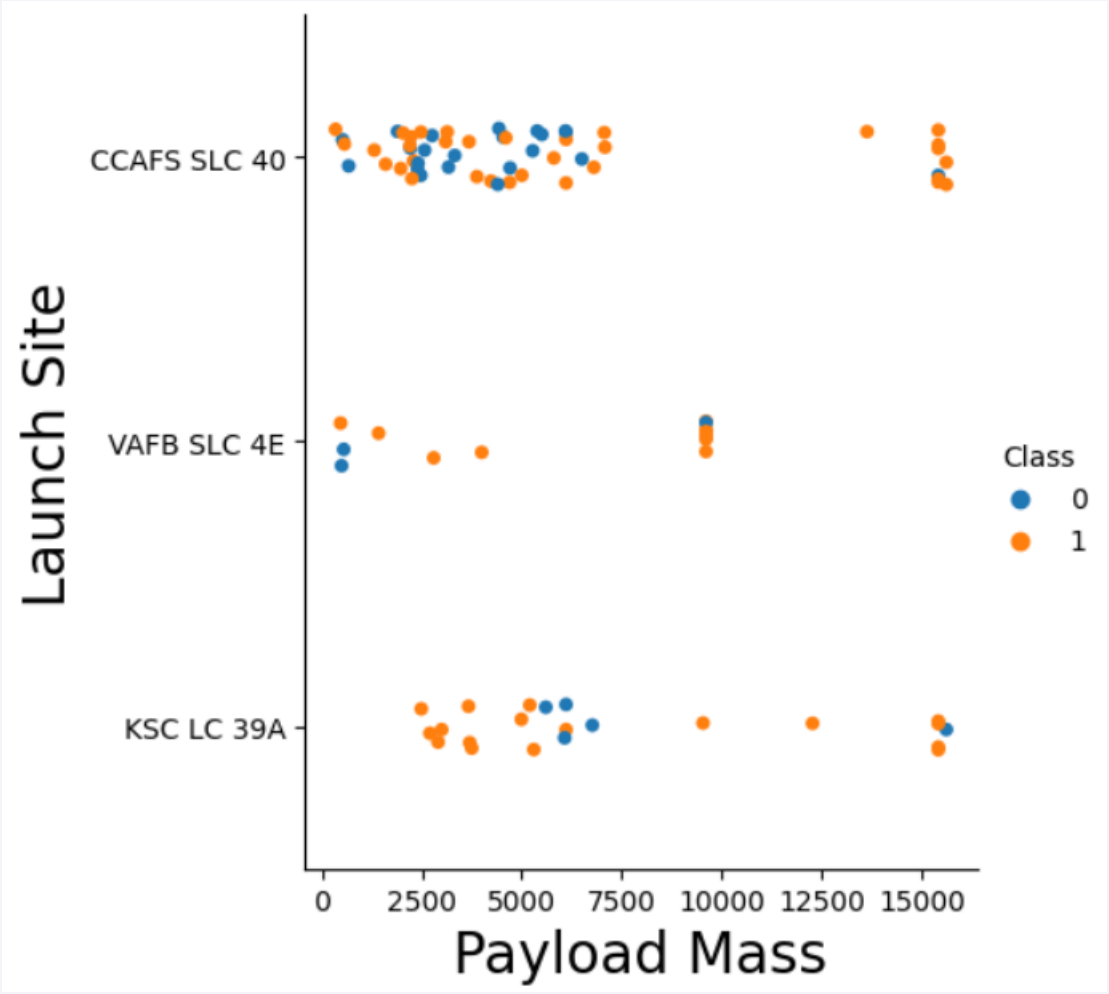
Flight Number vs. Launch Site

- Plot of Flight Number vs. Launch Site:
  - Obviously, the launch site KSC LC 39A has more successful landings.
  - CCAFS SLC 40 has the majority of launches.



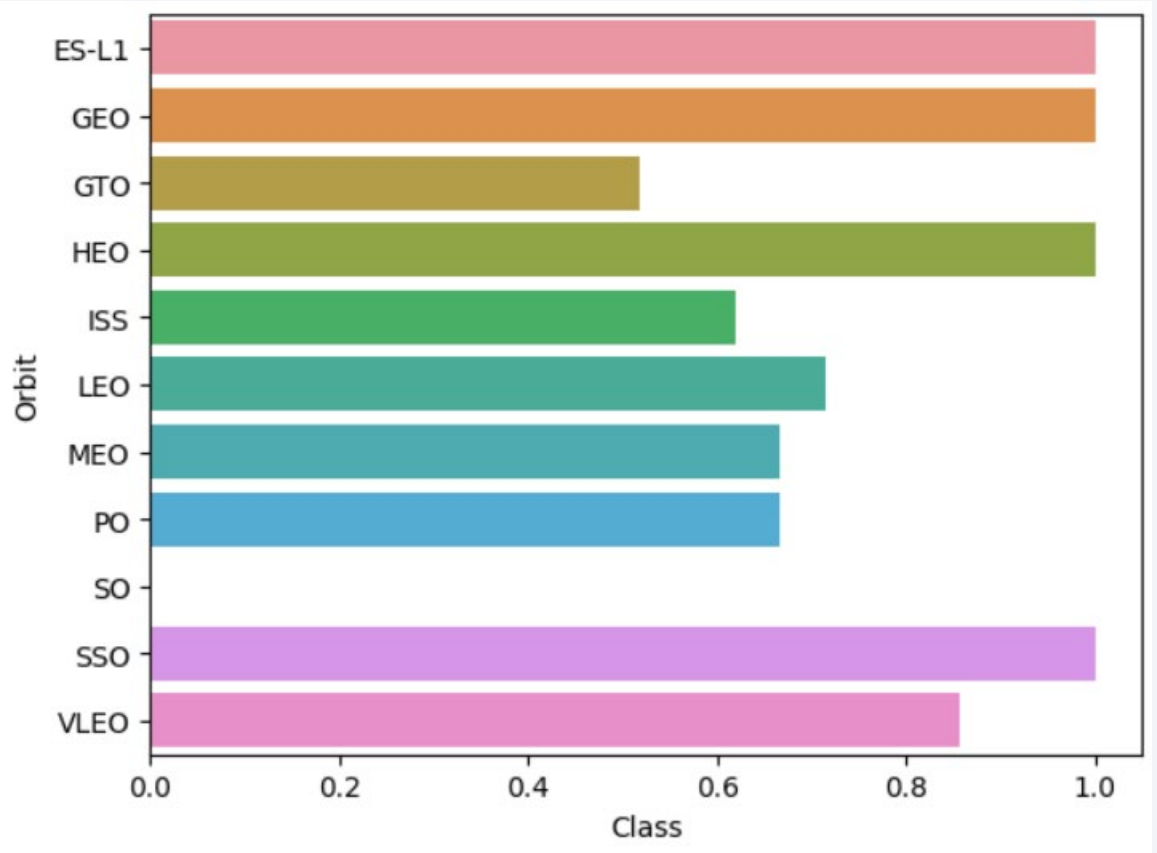
Payload vs. Launch Site

- Plot of Payload vs. Launch Site:
  - The Higher the Payload, The Higher possibility of successful landing.
  - The relationship between the variables is not that storng.



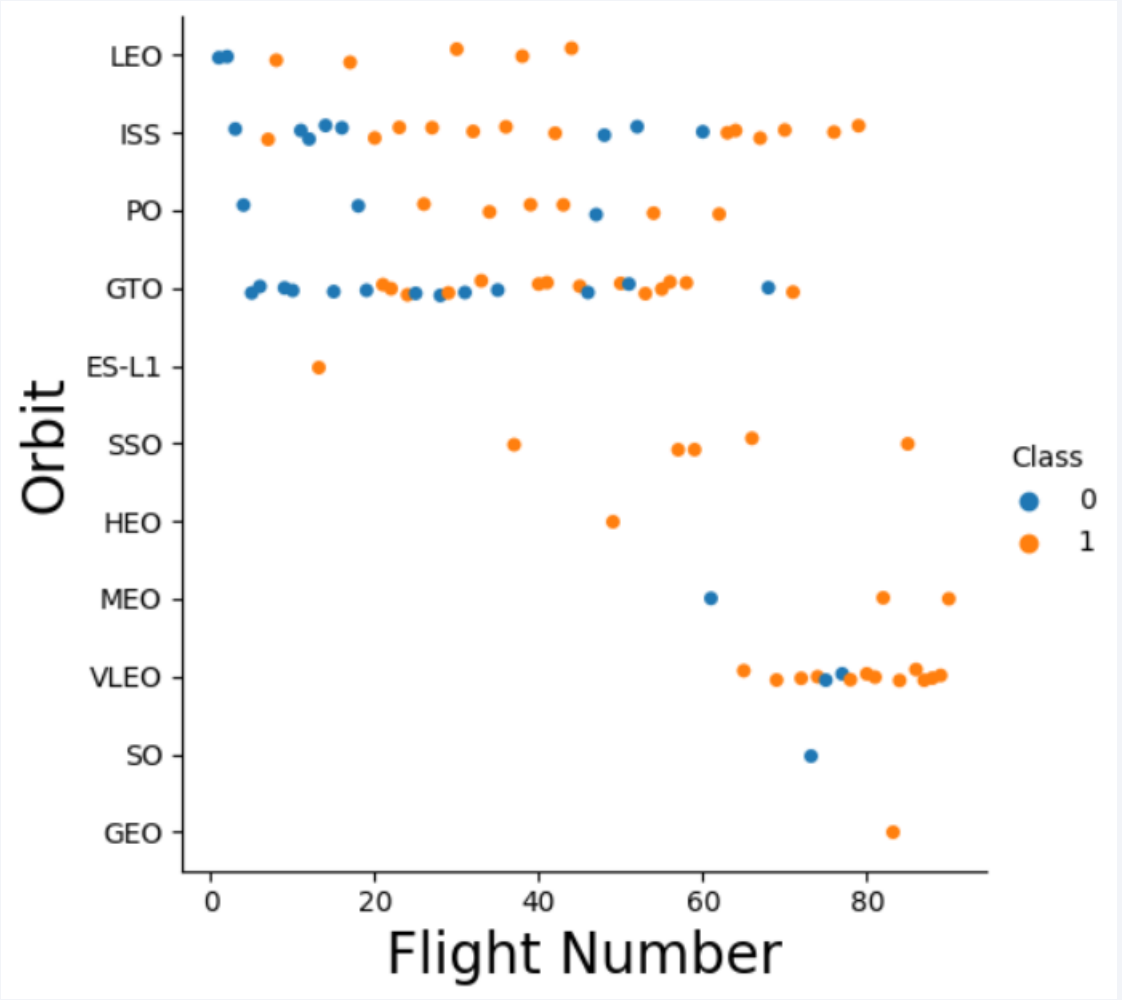
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type:
  - Orbits such as ES-L1, GEO, HEO, and SSO, have success rates of 100%.



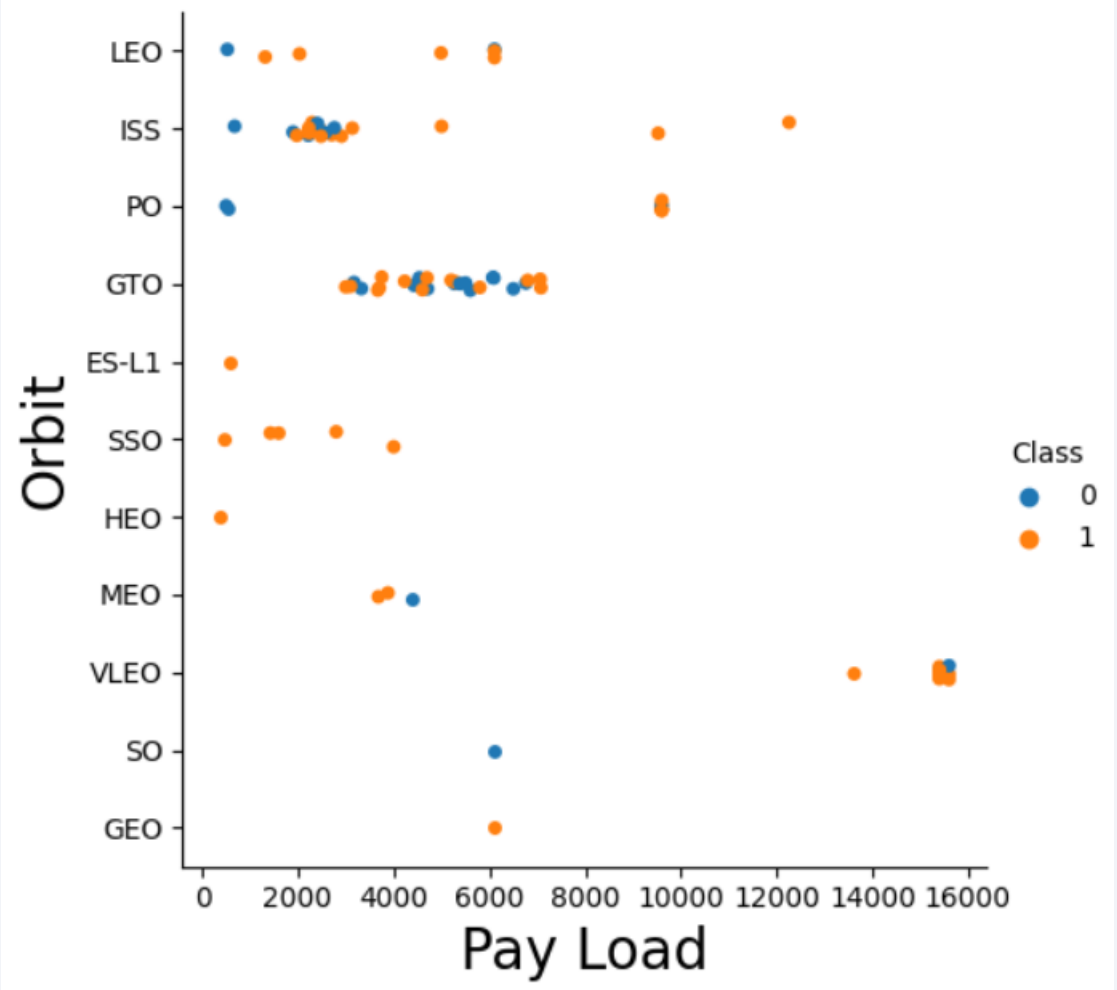
Flight Number vs. Orbit Type

- Scatter plot of Flight number vs. Orbit type:
  - We notice that the orbits with 100% success rates either have 1 flight of don't have any flights at all.
  - This finding makes the bar chart not accurate enough.



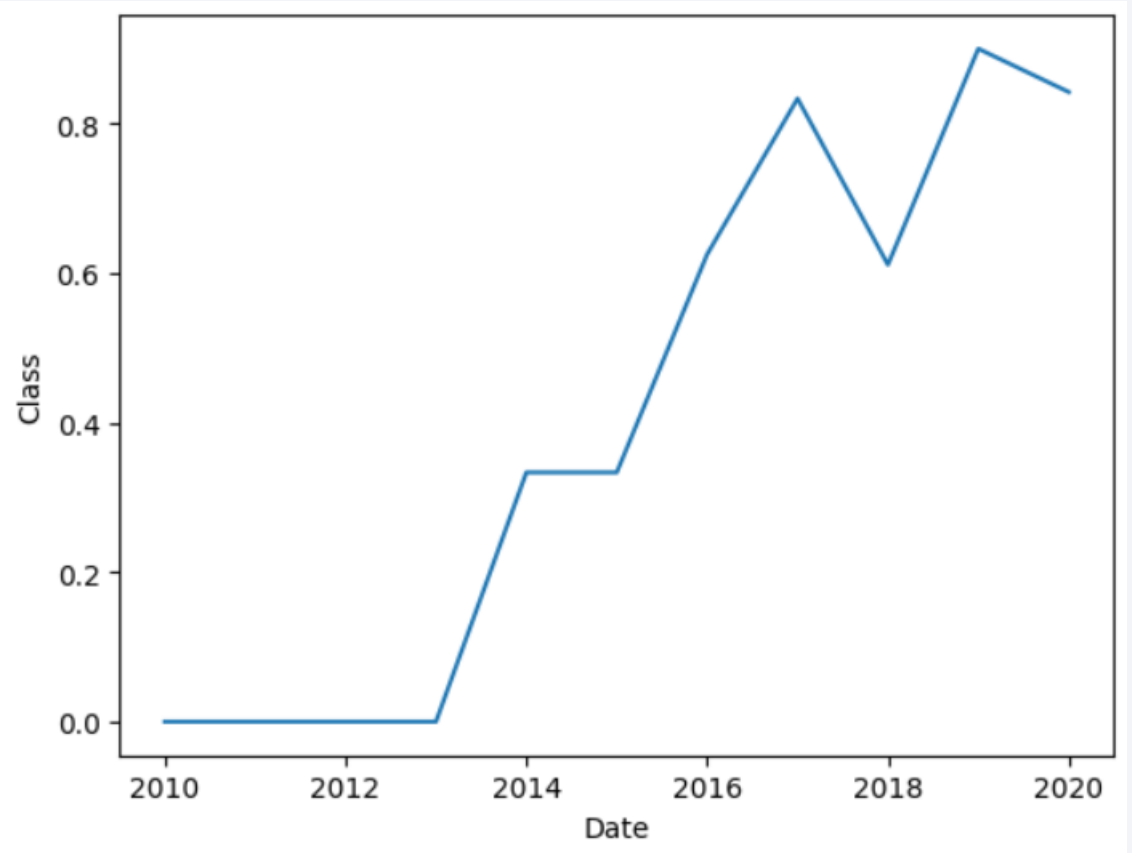
Payload vs. Orbit Type

- Scatter plot of payload vs. orbit type:
  - For some orbits, high payload mass results with high success rates.
  - But for GTO orbit, the rule doesn't apply.



Launch Success Yearly Trend

- Line chart of yearly average success rate:
  - After 2013, the success rate increases rapidly due to technology advancements, etc..



All Launch Site Names

- The Query selects distinct launch sites from SPACEXTBL dataset

```
%sql select distinct "Launch_site" from SPACEXTBL
```

```
[13]
 * 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'

- The Query selects launch site column with a condition that the record will contain CCA with limit to the first 5 records

```
%sql select Launch_Site from SPACEXTBL where Launch_Site like '%CCA%' limit 5
```

[28]

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The Query groups data by booster version then calculate the sum of payload mass for each booster

```
[30]  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL group by Booster_Version
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
2647
5384
9600
6460
3310
4990
9600

Average Payload Mass by F9 v1.1

- The Query calculate the average of payload mass for booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "F9 v1.1%"
```

[31]

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The Query finds the min date for missions that have succussed.

```
[33]  %sql select min(date) from SPACEXTBL where Mission_Outcome == "Success"

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

</>   min(date)
      01-03-2013
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The Query selects distinct booster versions depending on two conditions that have been mentioned.

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

```
[57] %sql select distinct Booster_Version from SPACEXTBL where (PAYLOAD_MASS__KG_ between 4000 and 6000) and ("Landing_Outcome" like 'Success (drone ship)')
```

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

Booster_Version

Total Number of Successful and Failure Mission Outcomes

- The Query displays mission outcomes along with its count after grouping the data by mission outcome.

```
%sql select  Mission_Outcome,count(*) from SPACEXTBL group by Mission_Outcome
```

[60]

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

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

Boosters Carried Maximum Payload

- The Query display booster version whose payload mass equals the maximum payload mass.

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ == (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

[7] ✓ 0.0s

```
 * 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

- Present your query result with a short explanation here

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

- Present your query result with a short explanation here


Section 3
Launch Sites
Proximities Analysis

Launch sites using Folium

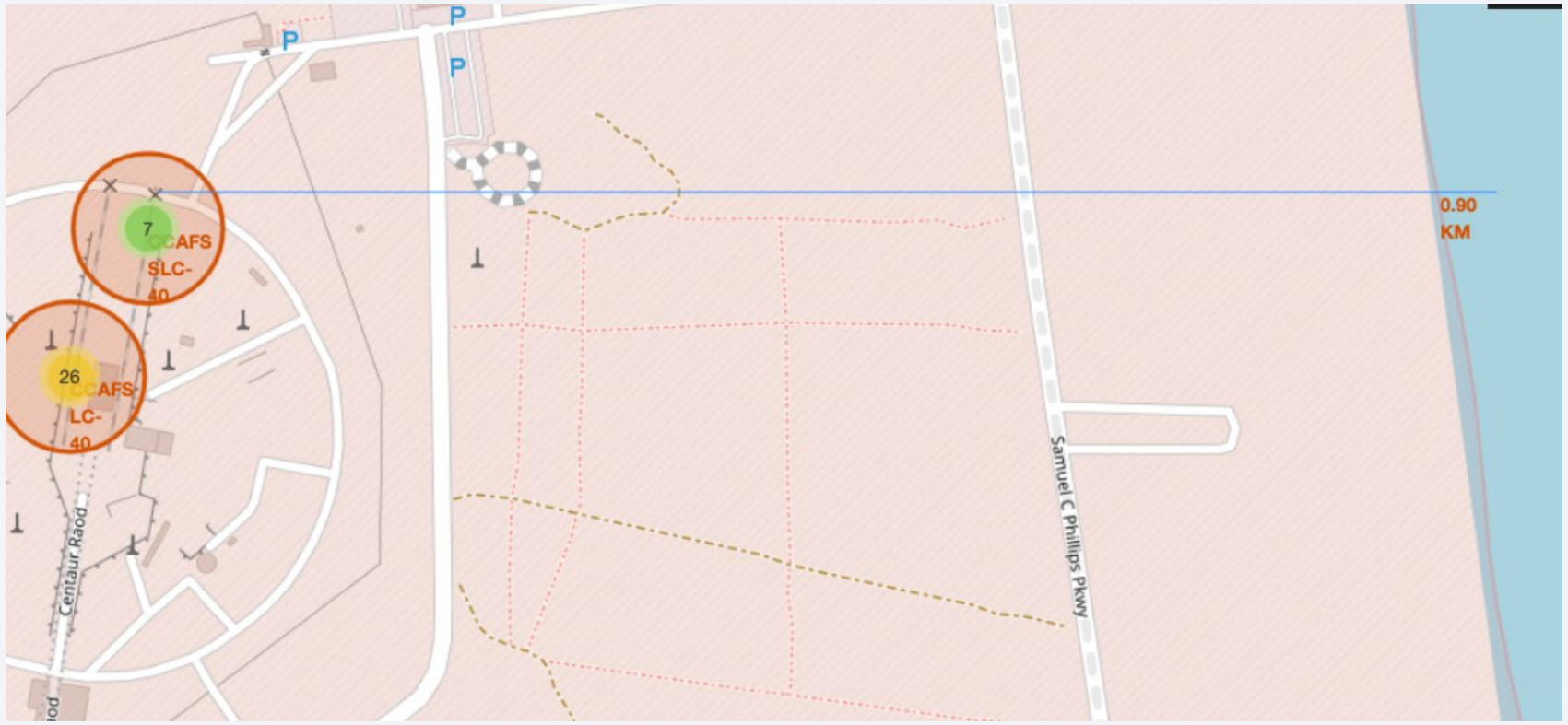
- We used markers and circles as a first try to visualize the launch sites.

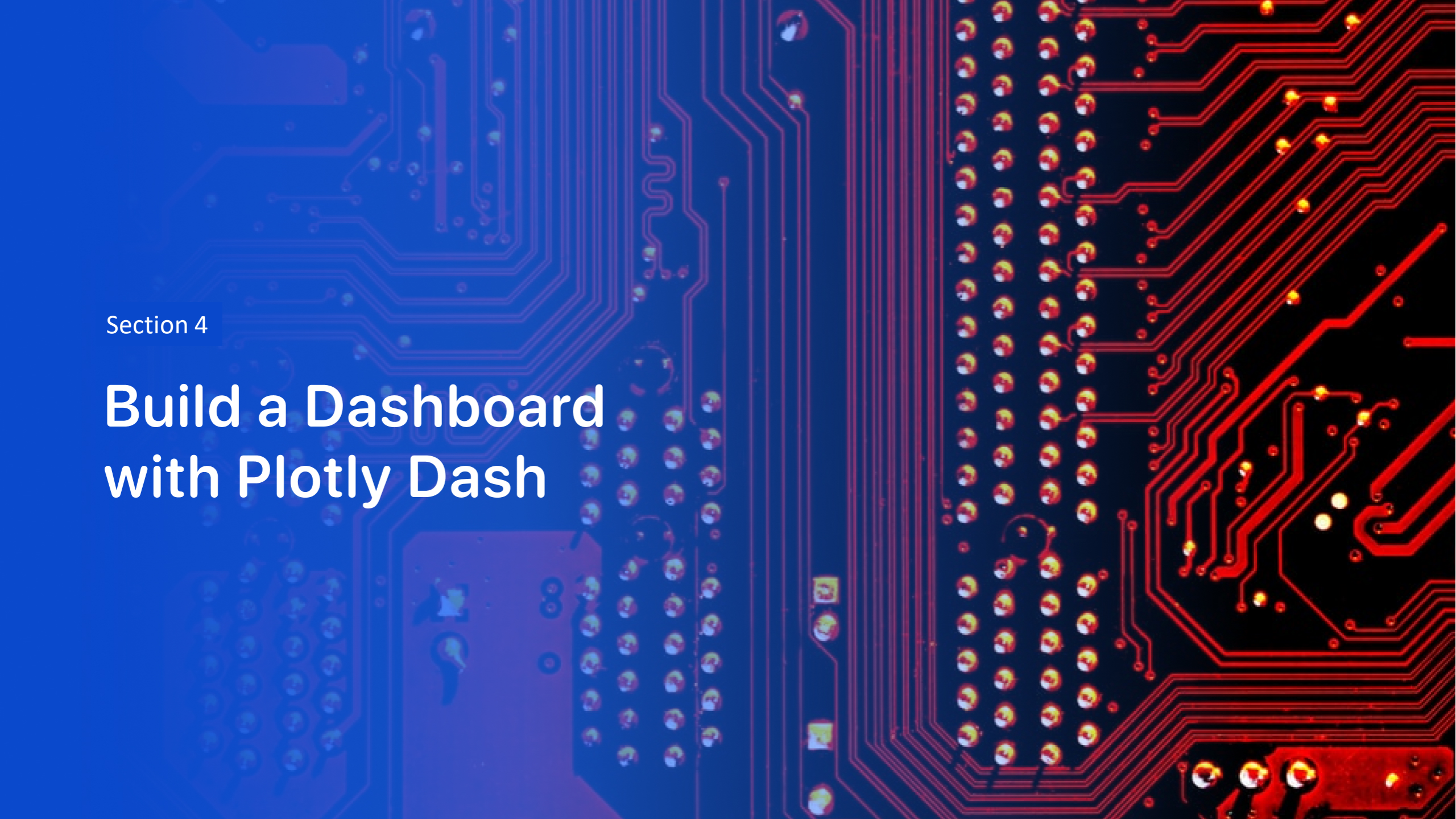
Success/Failure Launches

- As the successed of failed launches for same site has the same location, we used cluster to better visualize the data.

Distances and proximities

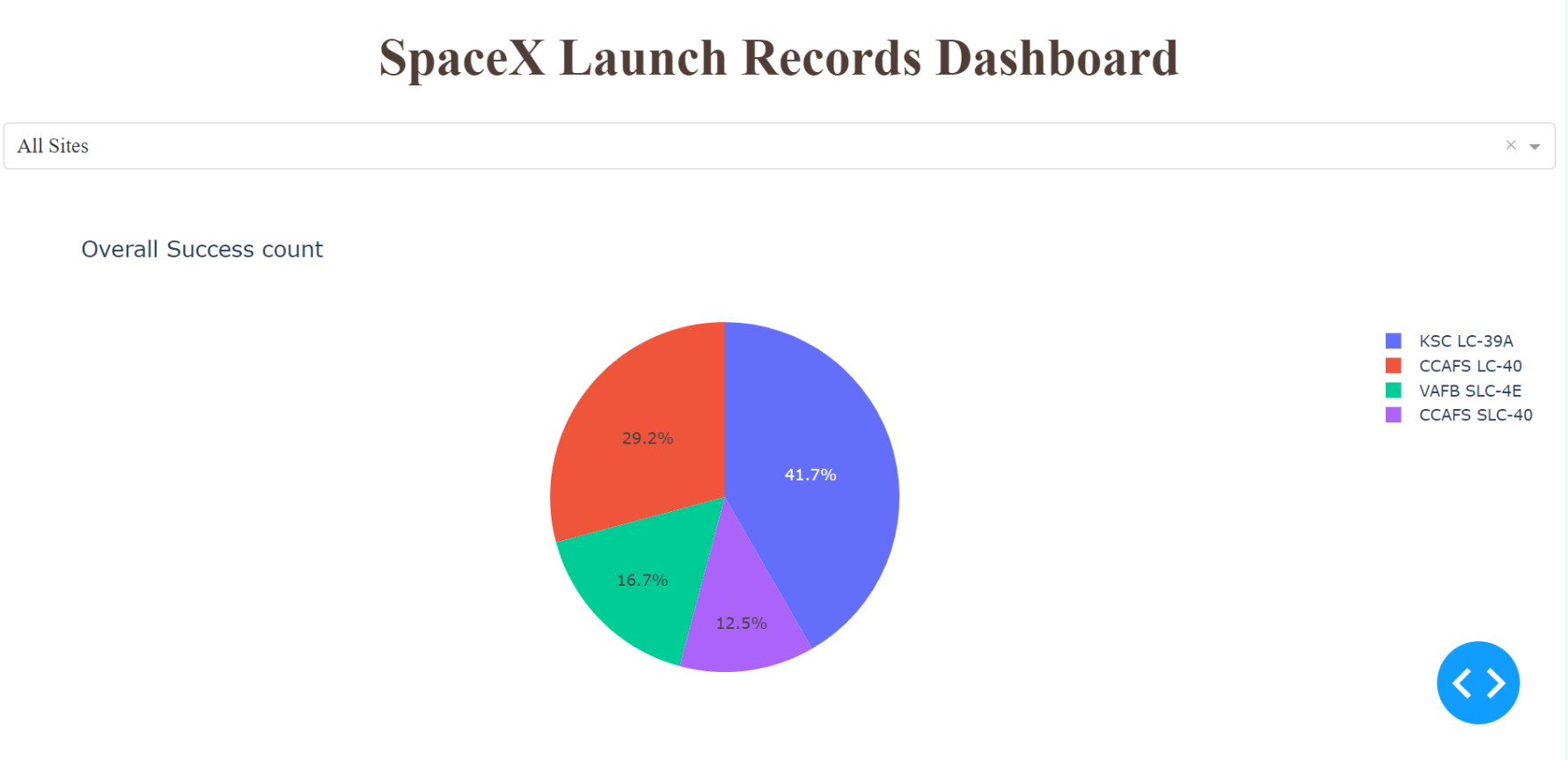
- Polyline is helpful to visualize the distances.




Section 4
Build a Dashboard with Plotly Dash

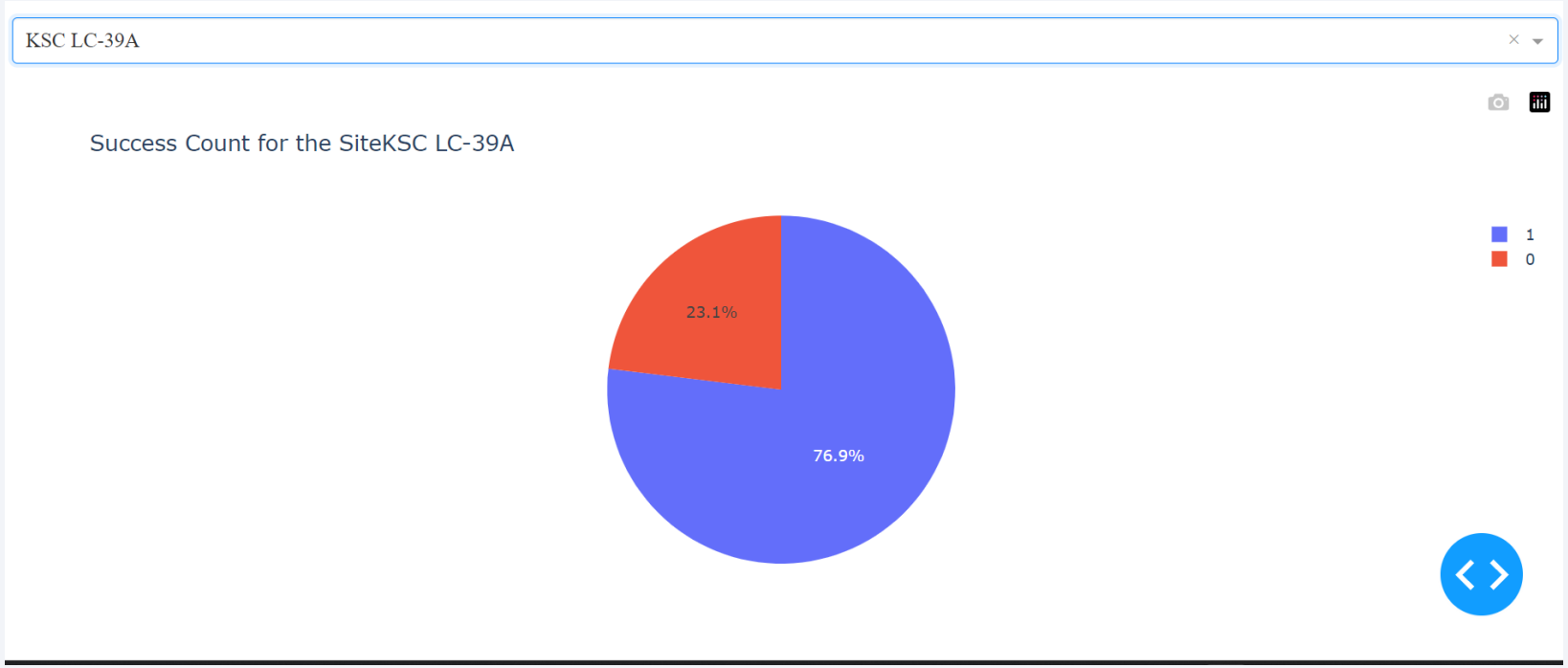
Interactive Pie charts

- KSC LC-39A has the most successful rate



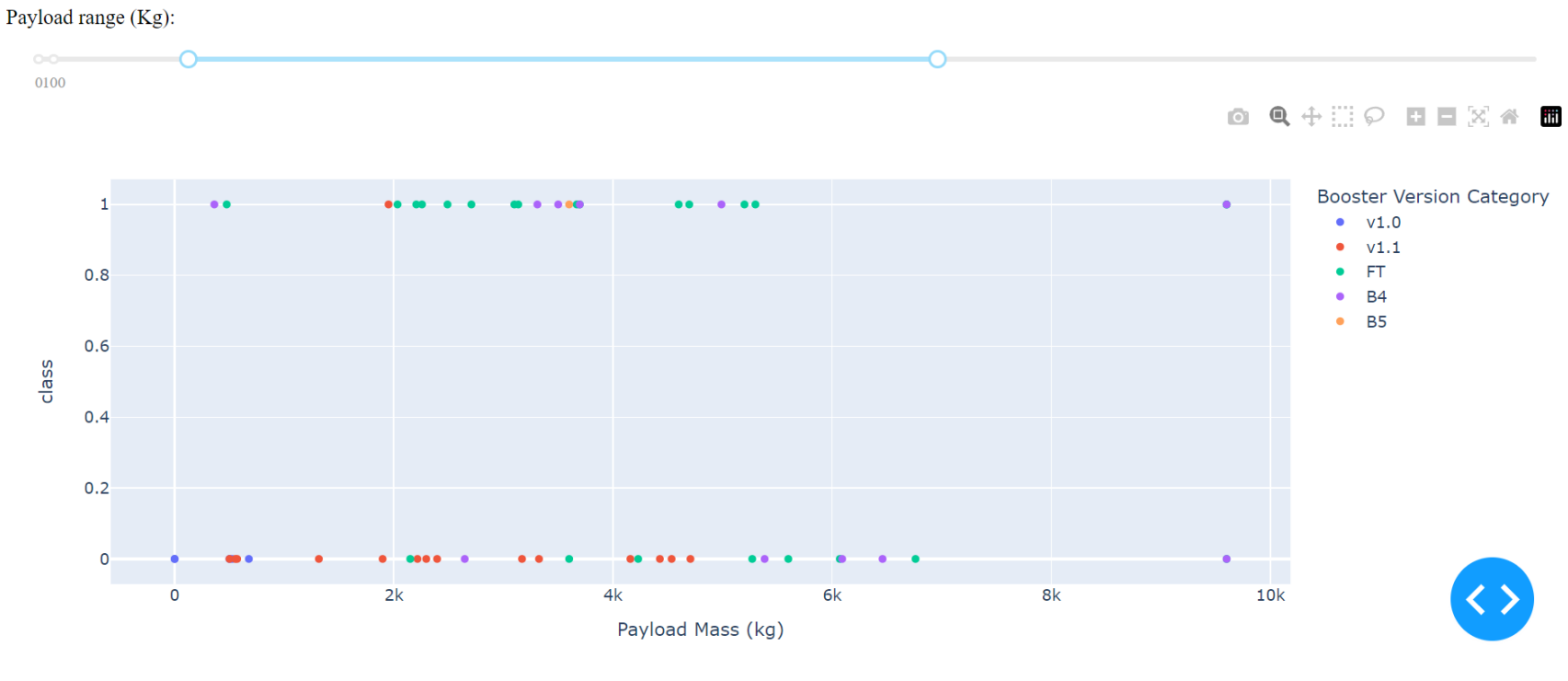
KSC LC-39A

- Its certain that this site has high percentage of successful rate since it has the highest percentage among the other sites



Payload vs. Launch Outcome scatter plot

- Booster version FT performs the best in wide range of payloads

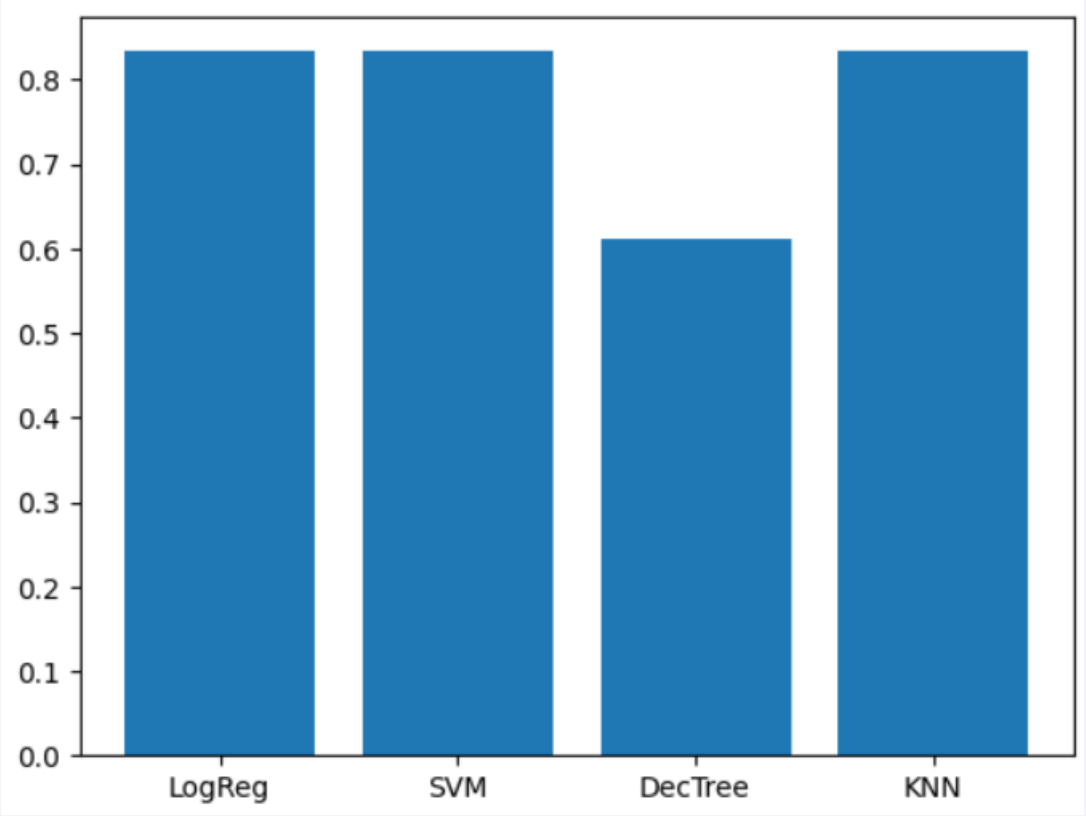


Section 5

Predictive Analysis (Classification)

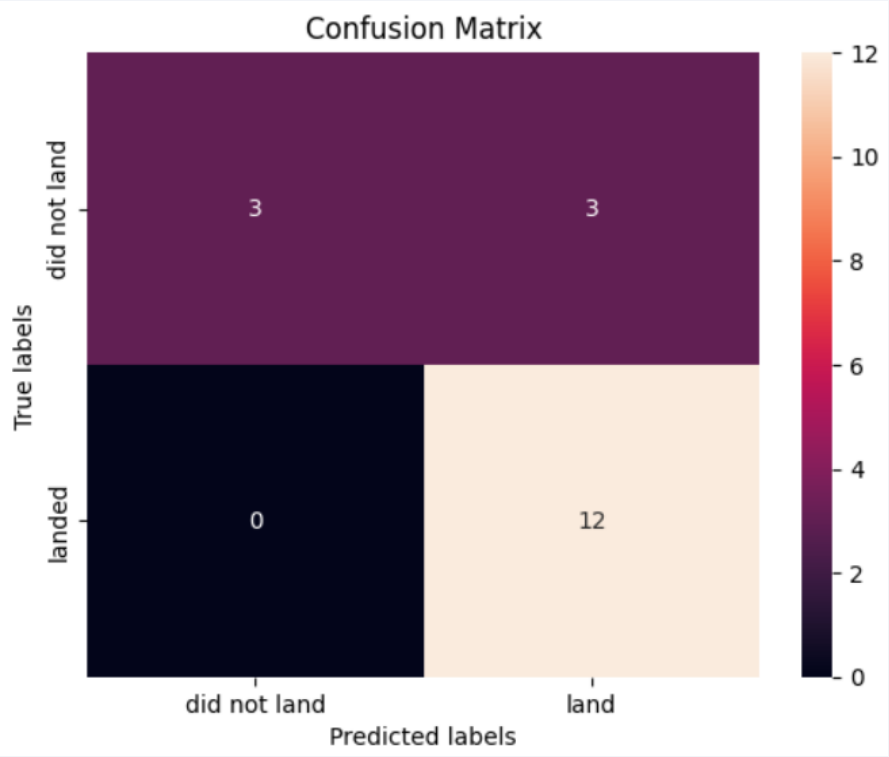
Classification Accuracy

- Logistic Regression and SVM and KNN Performs Equally.



Confusion Matrix

- Logistic Regression's confusion matrix



Conclusions

- We can consider some variables as Payload mass, Booster version, and Launch site as the most important variables.
- The relationship between variables are somehow unclear and require obtaining relation between two or more variable at the same time.
- Some Orbits like SSO has better landing results than the others.
- The Higher the Payload, The Higher possibility of successful landing.

Appendix

- Final Data used for classification

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Thank you!