IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roman Myskin
2025-07-04

Outline

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

Executive Summary - Methodologies

- We used an analytical approach to analyze SpaceX Landing Outcomes
- We collected, wrangled, conducted EDA, and applied different classification models to the input data to define the most relevant features and make a prediction model
- We used Python Pandas, Numpy, Matplotlib, Seaborn, Sklearn, Folium, Dash libraries to make all programming estimations
- We presented data in Jypyter Notebook format.

Executive Summary - Results

- Launching Sites near the coastlines, roads, and railways has a good outcome for success
- Only conducting more launches may improve our success rate
- ES-L1, GEO, HEO and SSO orbits are the best fit for start since they success rate is high
- We need to pay attention to Booster Version FT of Falcon 9 which has a high success rate compared to others
- The deployed model gives 83% of accuracy based on 90 features

Introduction

- SpaceY is a new growing company that already competes with leading commercial Space Travel providers like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX.
- The company owner, Allon Mask, is sure we need to build new rockets based on the historical data to prevent future mistakes and outbid other companies.
- In this presentation, we'll analyze SpaceX data about Falcon 9 launches and find insights from open available sources to define feature that affect the success of launches.
- With this knowledge, our company will open a new Era in Space Traveling, allowing to refine 1 rocket stage saving millions every year.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX API
- Perform data wrangling
  - We filtered data, filled in null values and determined Successful/ Unsuccessful landing results
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We built, tuned, and evaluated the next classification models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, K-nearest Neighbors

Data Collection

- We collected data using SpaceX API using requests

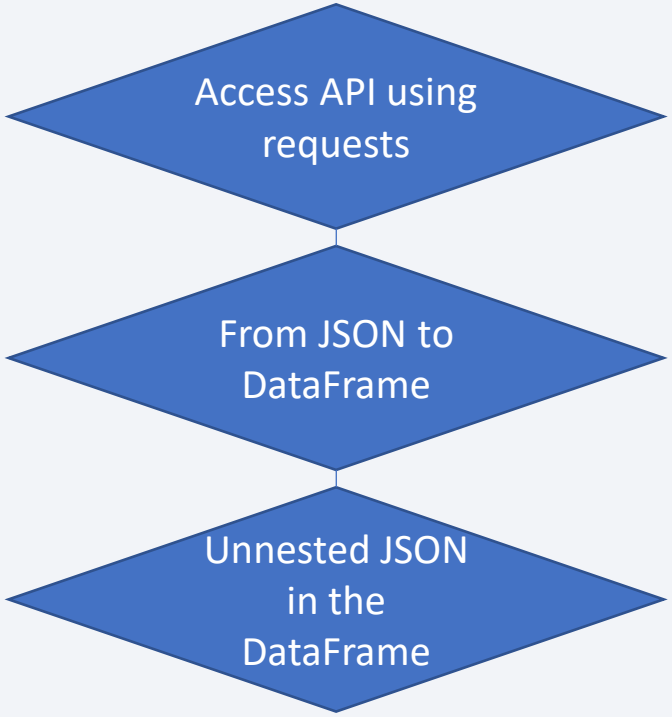
1. Transformed SpaceX API data from JSON to a dataframe
2. Unnested JSON data
3. Filtered data by “Falcon 9” Booster Version

- We collected data using Wikipedia using requests and Beautiful soup:

1. We accessed the Wiki page using requests
2. We parsed the content using BeautifulSoup
3. We found out the target table
4. We formed a DataFrame from the target table

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/1994aba60fd383c2a0cb31540aa64854ee712bb9/Notebooks/Data%20Collection%20%E2%80%93%20SpaceX%20API.ipynb

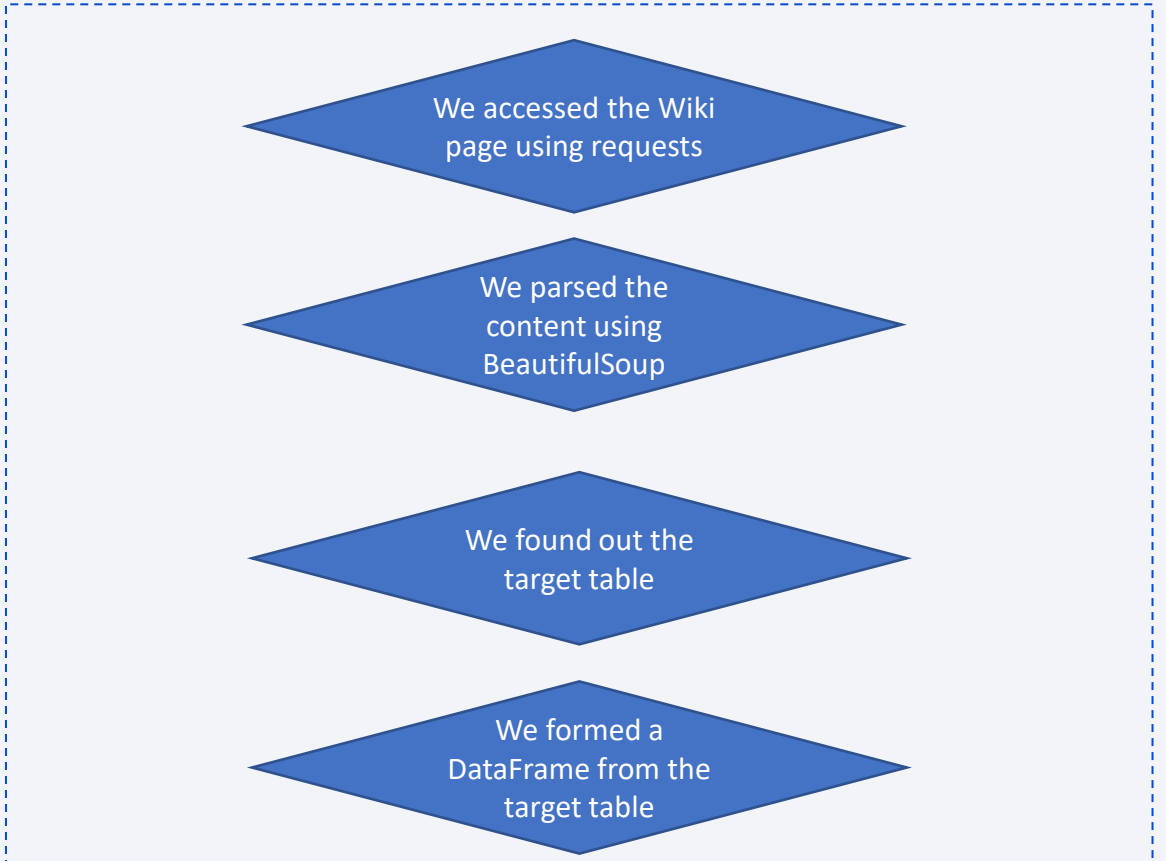


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

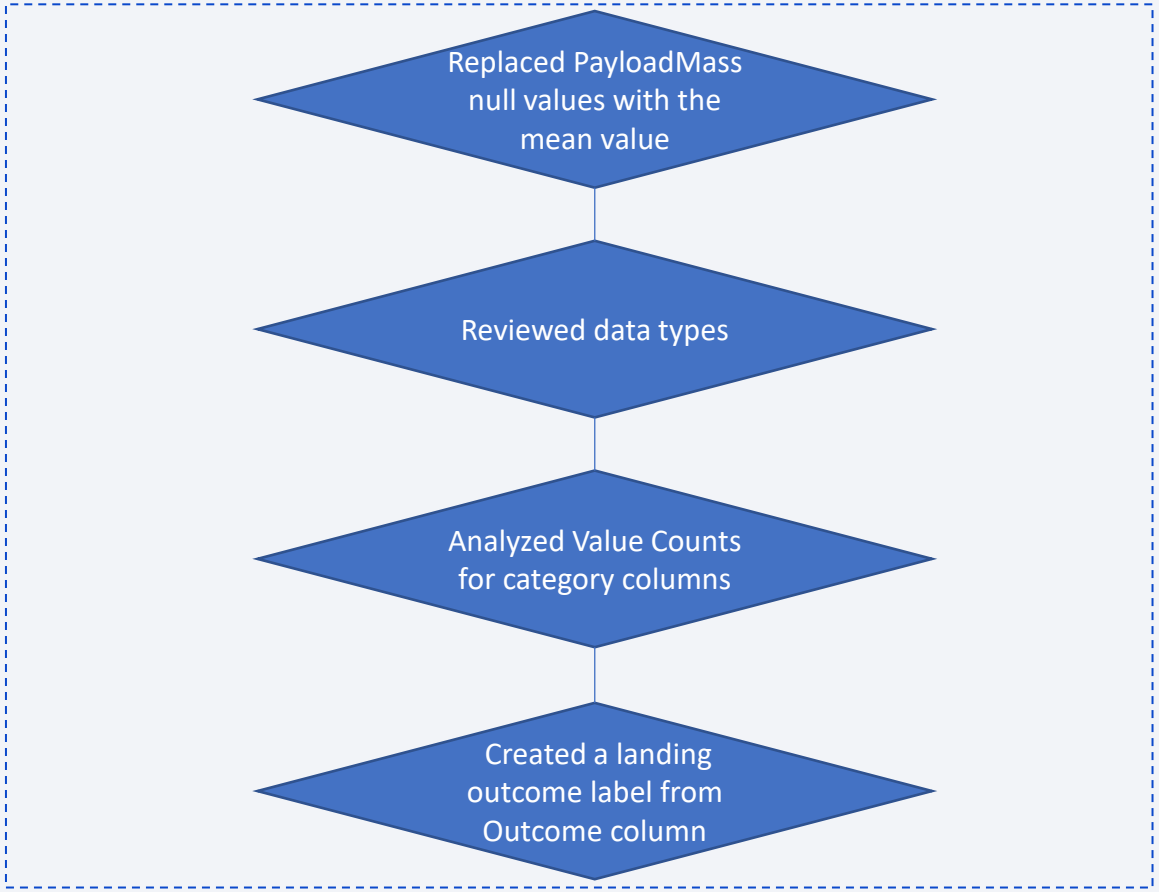
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/164bc04de2d876769490bd6bdbd8bac3924d8d17/Notebooks/Data%20Collection%20-%20Scraping.ipynb



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/e1003a40fdecd826aced3ce248d8c389f14e1789/Notebooks/Data%20Wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/4c1c89d11d746c1c07d7713f286ce1ccf5ab9d62/Notebooks/EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization

- We used a scatter plot to show correlations between Payload Mass, Launch Site and Class features
- We used a bar plot to see Success Rate by Orbits
- We used a scatter plot to show correlations between Orbit, Flight Number and Class features
- We used a line plot to show trend for Successful Landings

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/cfe130ca7a6ed1afc41a43ba212e46c219f48180/Notebooks/EDAwithSQL.ipynb

EDA with SQL

- The unique launch sites
- 5 records where launch sites begin with the string ‘CCA’
- The total payload mass carried by boosters launched by NASA (CRS)
- Average Payload Mass carried by booster version F9 v1.1
- The first successful landing outcome in ground pad
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The booster versions that have carried the maximum payload mass
- The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Count of landing outcomes between the date 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/288fd6ebe2ef8080b8825f4d3bba89afab464a44/Notebooks/Build%20an%20Interactive%20Map%20with%20Folium.ipynb

Build an Interactive Map with Folium

- We marked circle objects to highlight launching sites
- We applied markers with different colors united by clusters to show success/unsuccessful landings
- We added lines to show the distance from coastline, roads and railways

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/9404bec51ab7c6a40200c562c37b331bd0e4b60e/Notebooks/Build%20a%20Dashboard%20with%20Plotly%20Dash.py

Build a Dashboard with Plotly Dash

- We added a pie chart to show Success/Unsuccess allocation with an interactive drop-down filter for Launching Sites
- We added a scatter plot to show the Correlation between Successful landings and Payload Mass for different booster versions

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone/blob/837a58237de96937e227c14eb98081841 68dd577/Notebooks/Predictive%20Analysis.ipynb

Predictive Analysis (Classification)

- We used 90 features to build a model
- We used class (1-success,0-fail) feature as depended value
- We normalized independent features
- We split data by train,test samples (test_size = 0,2)
- We applied the next Classification Models with GridSearch
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-nearest Neighbors

Predictive Analysis (Classification)

Grid Search Parameters (cv=10):

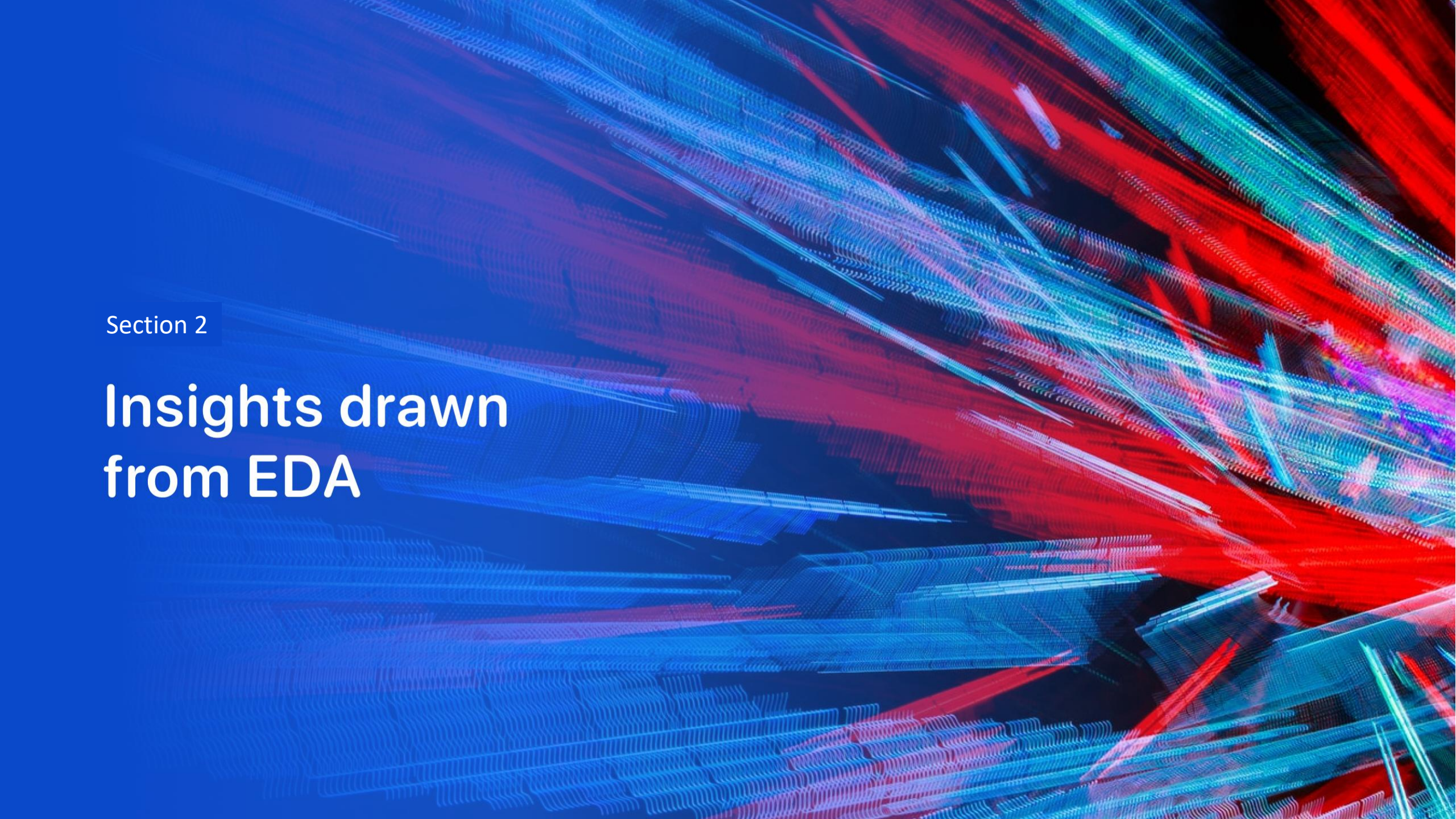
- Logistic Regression: "C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs’]
- Support Vector Machine: 'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'), 'C': np.logspace(-3, 3, 5), 'gamma':np.logspace(-3, 3, 5)
- Decision Tree: criterion': ['gini', 'entropy'], 'splitter': ['best', 'random’], 'max_depth': [2*n for n in range(1,10)], 'max_features': [None,'log2', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]
- K-nearest Neighbors: n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1,2]

Predictive Analysis (Classification)

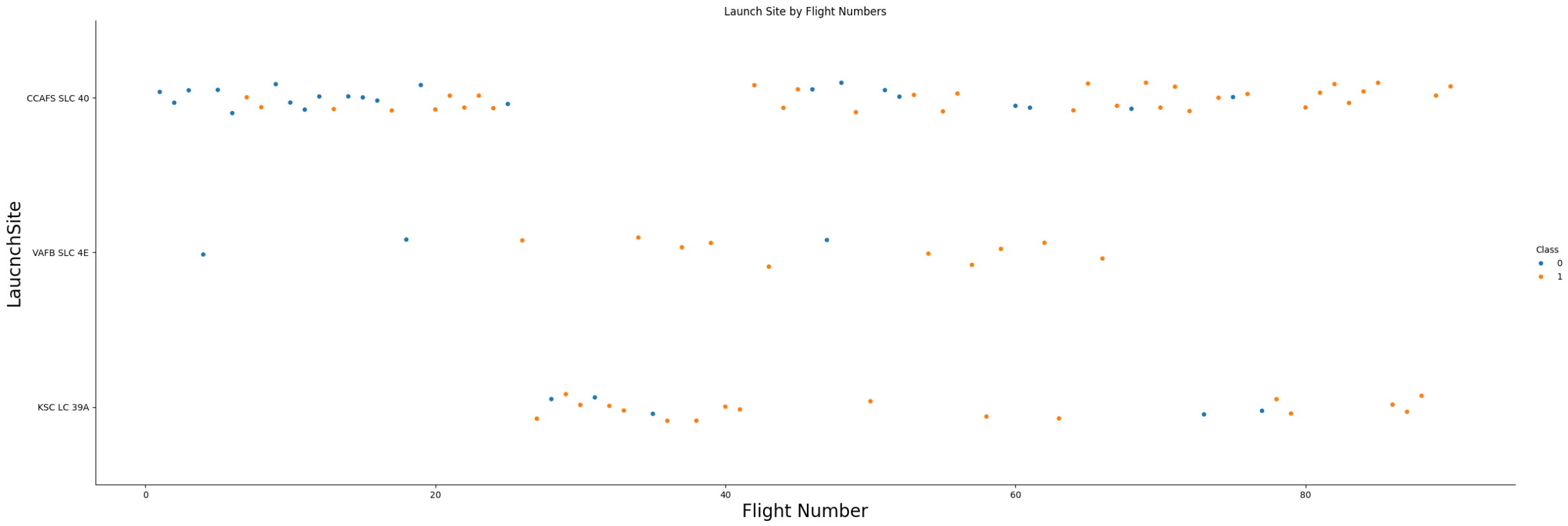
- We found the best parameters and estimated accuracy for every Model
- We plotted a confusion matrix

Results

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

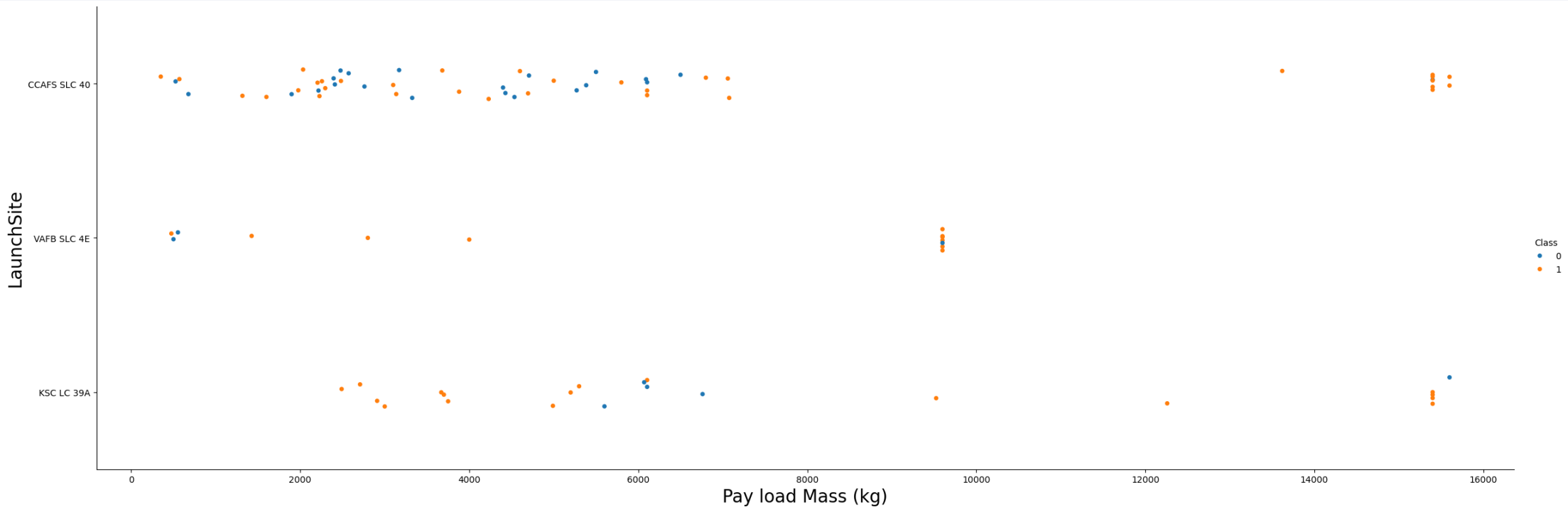

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



We can see that Successful Landings are increasing with the Flight Number

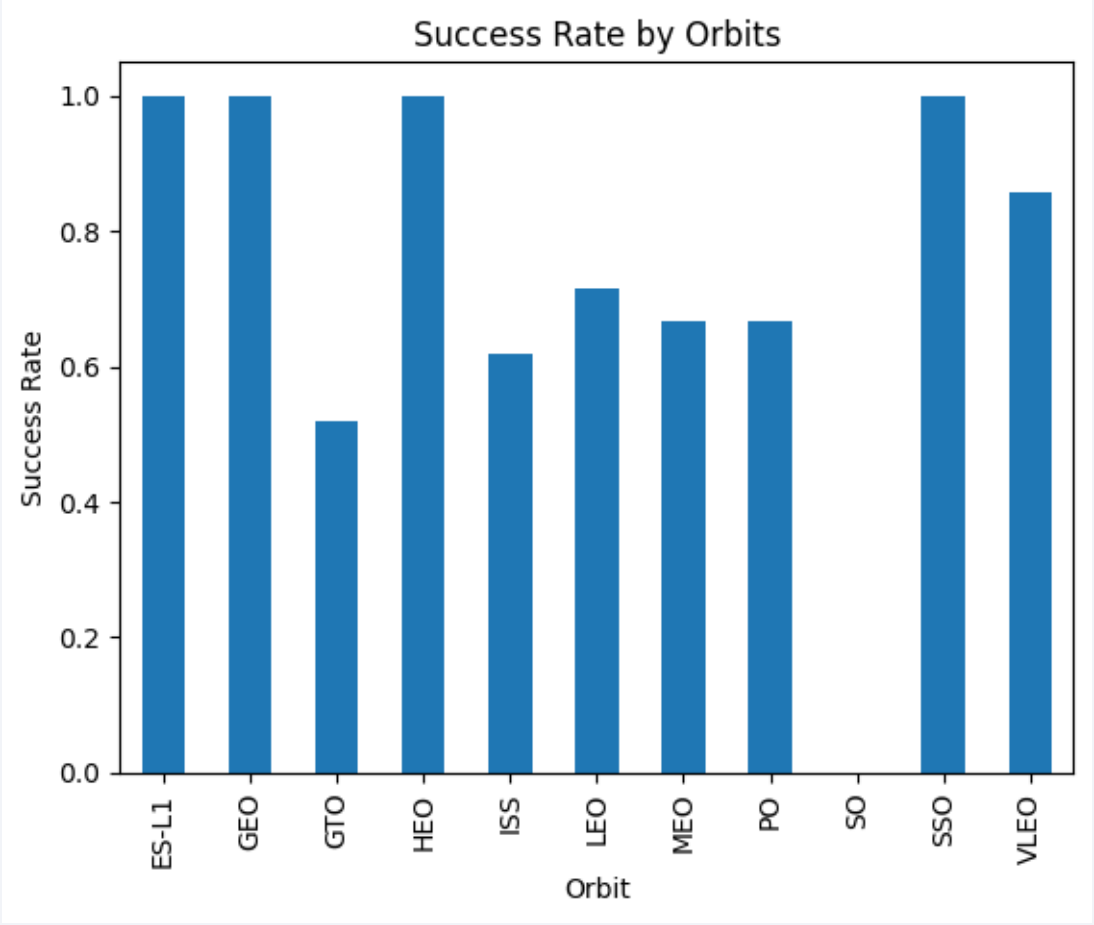
Payload vs. Launch Site



You can see for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

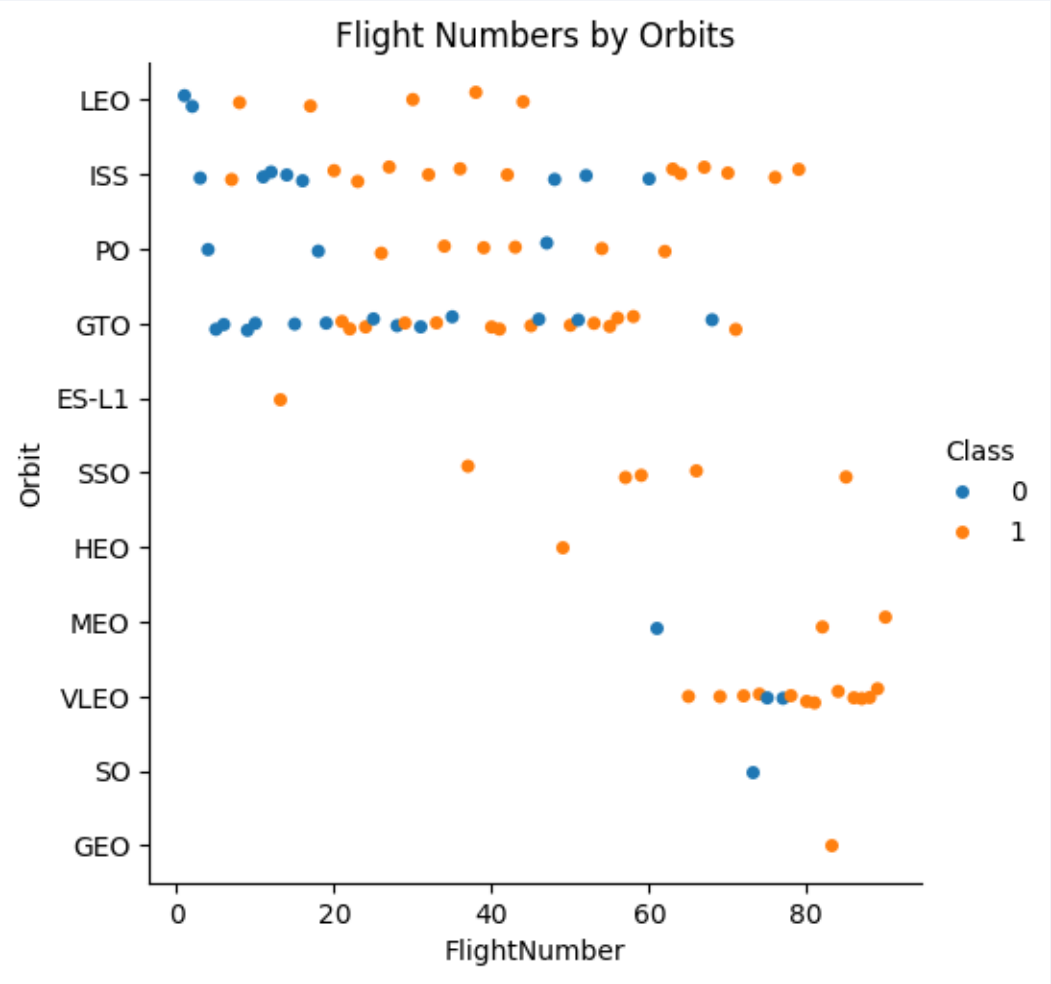
Success Rate vs. Orbit Type

100% of Success Landings were from ES-L1, GEO, HEO and SSO orbits while none of 1 Stages landed from SO orbit



Flight Number vs. Orbit Type

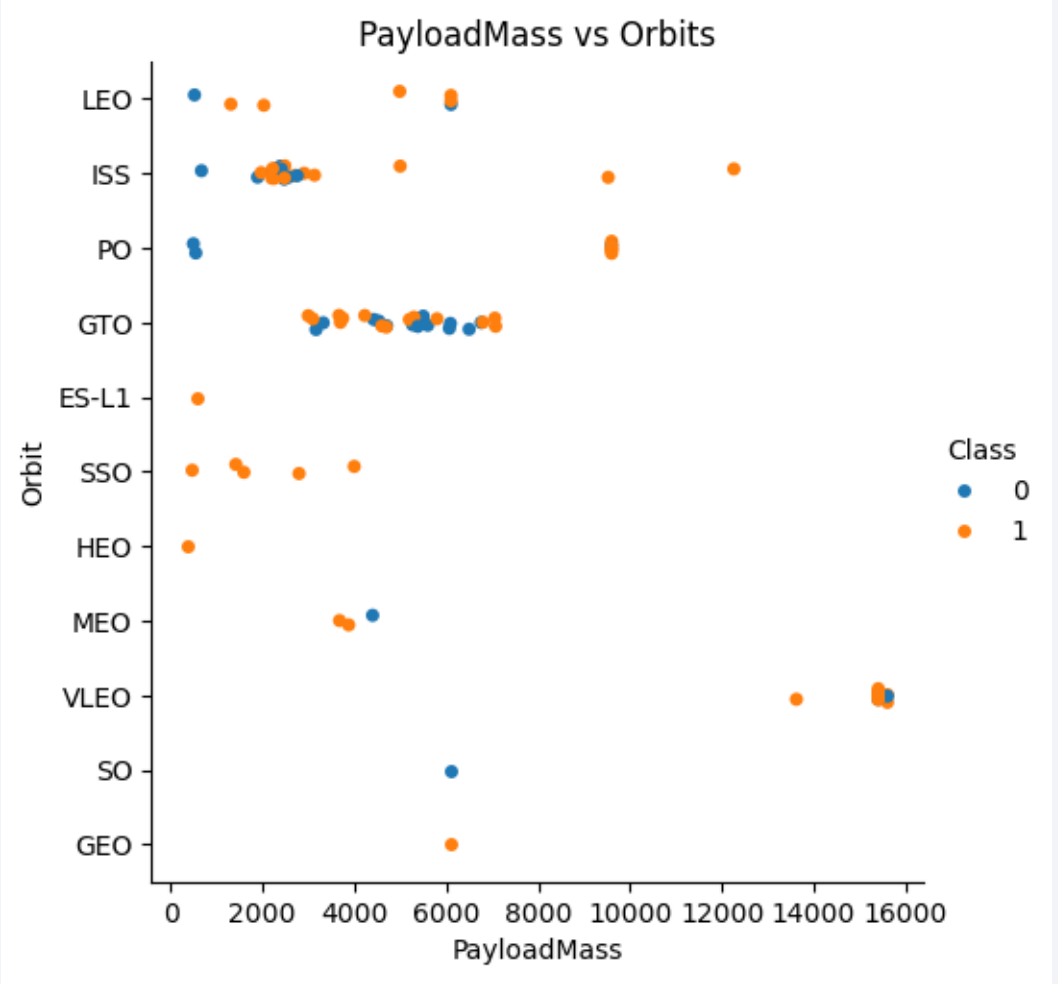
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

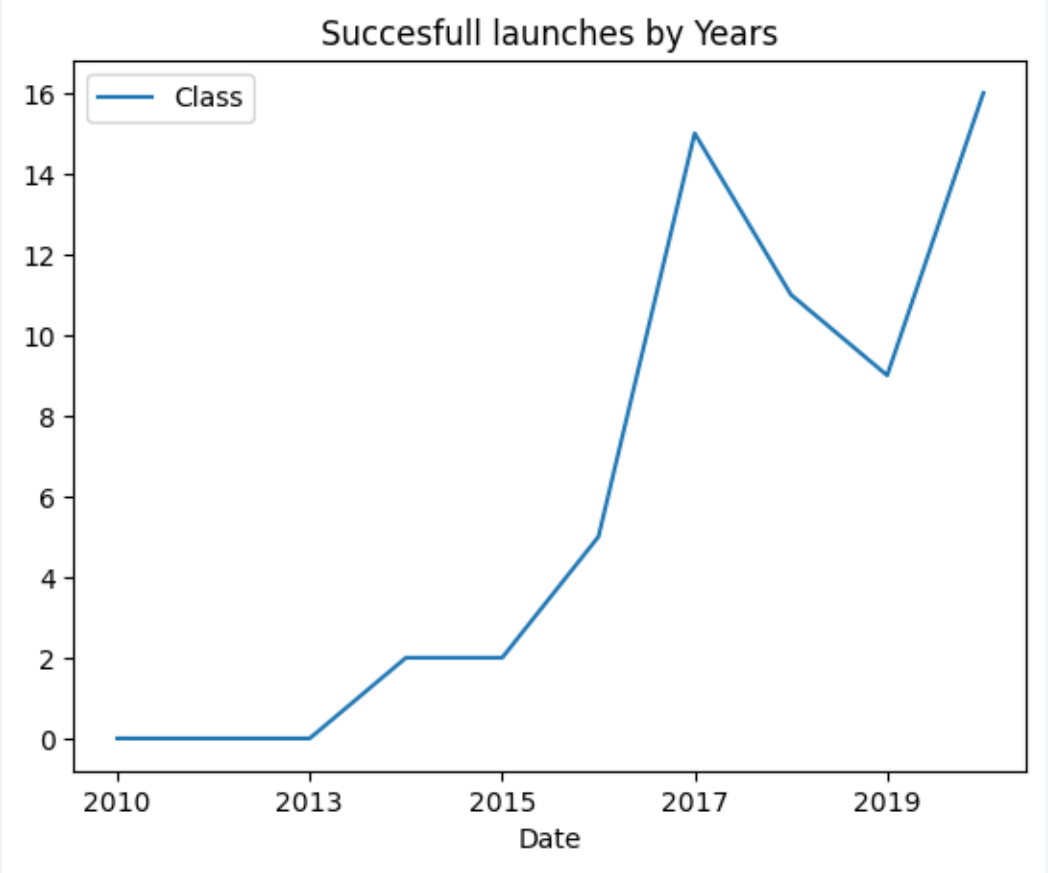
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

You can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here
- There are 4 Launch Sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

```
In [11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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`
- Present your query result with a short explanation here
- Here the example of 5 Records for CCAFS LC-40 Site

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

In [12]:
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here
- Nasa (CRS) launched 45,596 kg

Task 3

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

In [15]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

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

Out[15]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here
- The Average Payload Mass for Booster Version ‘F9 v1.1’ is 2,928.4 kg

```
In [16]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[16]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here
- First successful landing outcome was on 2010-06-40

```
In [19]: %sql SELECT Date FROM SPACEXTABLE WHERE Date = (SELECT MIN(Date) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
Out[19]:
```

Date
2010-06-04

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
- Present your query result with a short explanation here
- Boosters that have success in drone ship and have a payload mass greater than 4000 but less than 6000 are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE "Success%" AND Landing_Outcome LIKE "%drone%" AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
```

```
 * 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
- Present your query result with a short explanation here
- There are 3 fail outcomes and 98 successful outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(Mission_Outcome) AS Success1 FROM SPACEXTABLE WHERE Mission_Outcome = "Success";
```

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

Success1
98

```
%sql SELECT COUNT(Mission_Outcome) AS Fail FROM SPACEXTABLE WHERE Mission_Outcome != "Success";
```

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

Fail
3

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here
- 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 booster versions had the maximum payload mass

```
[43]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

[43]:

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
- Here are the Booster Versions failed in 2015

```
%sql SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome LIKE "%drone%" AND Landing_Outcome NOT LIKE "Succes
```

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

substr(Date, 6,2)	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
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

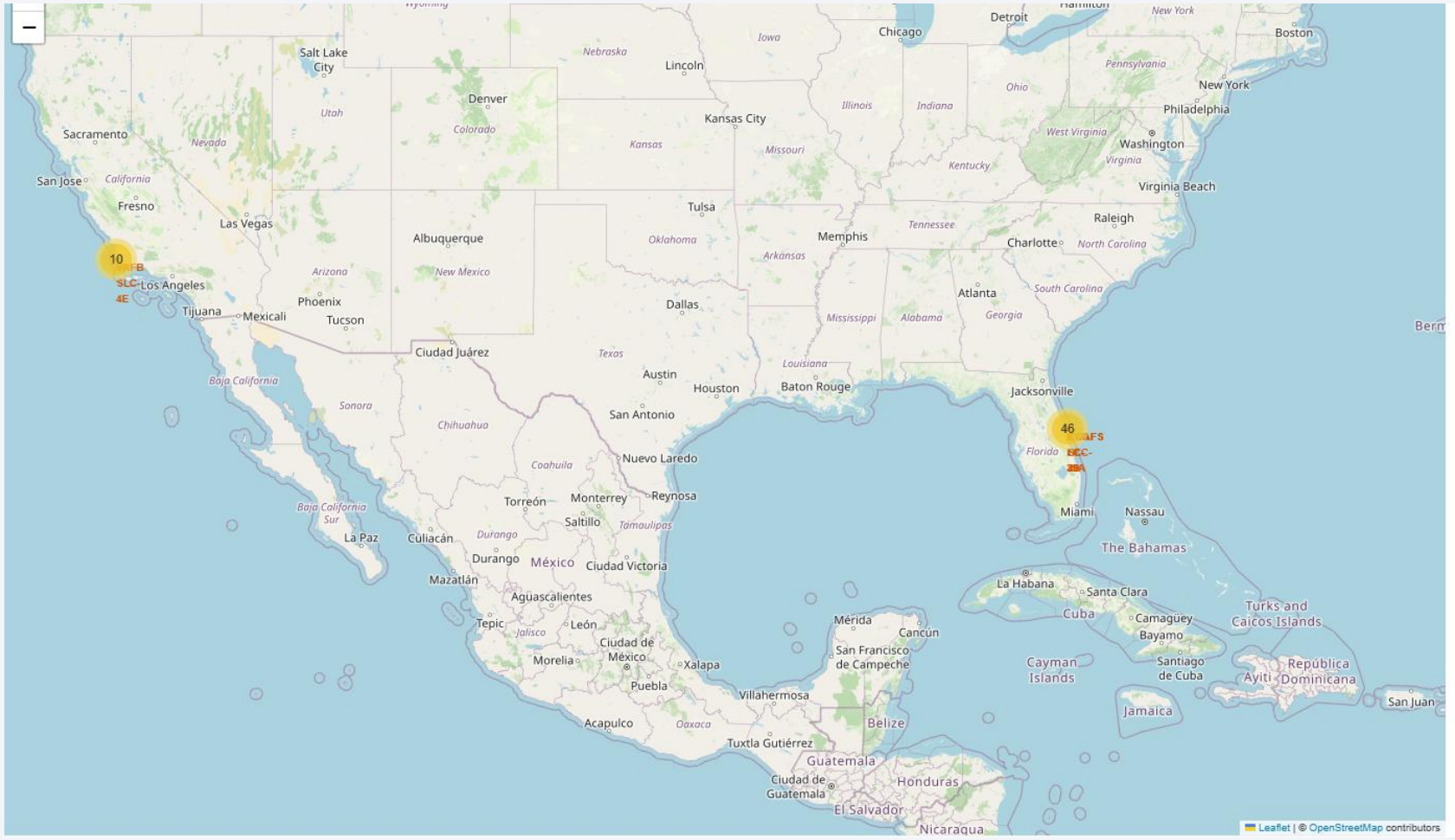
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Present your query result with a short explanation here
- Here are Landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Landing_Outcome	Counts
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

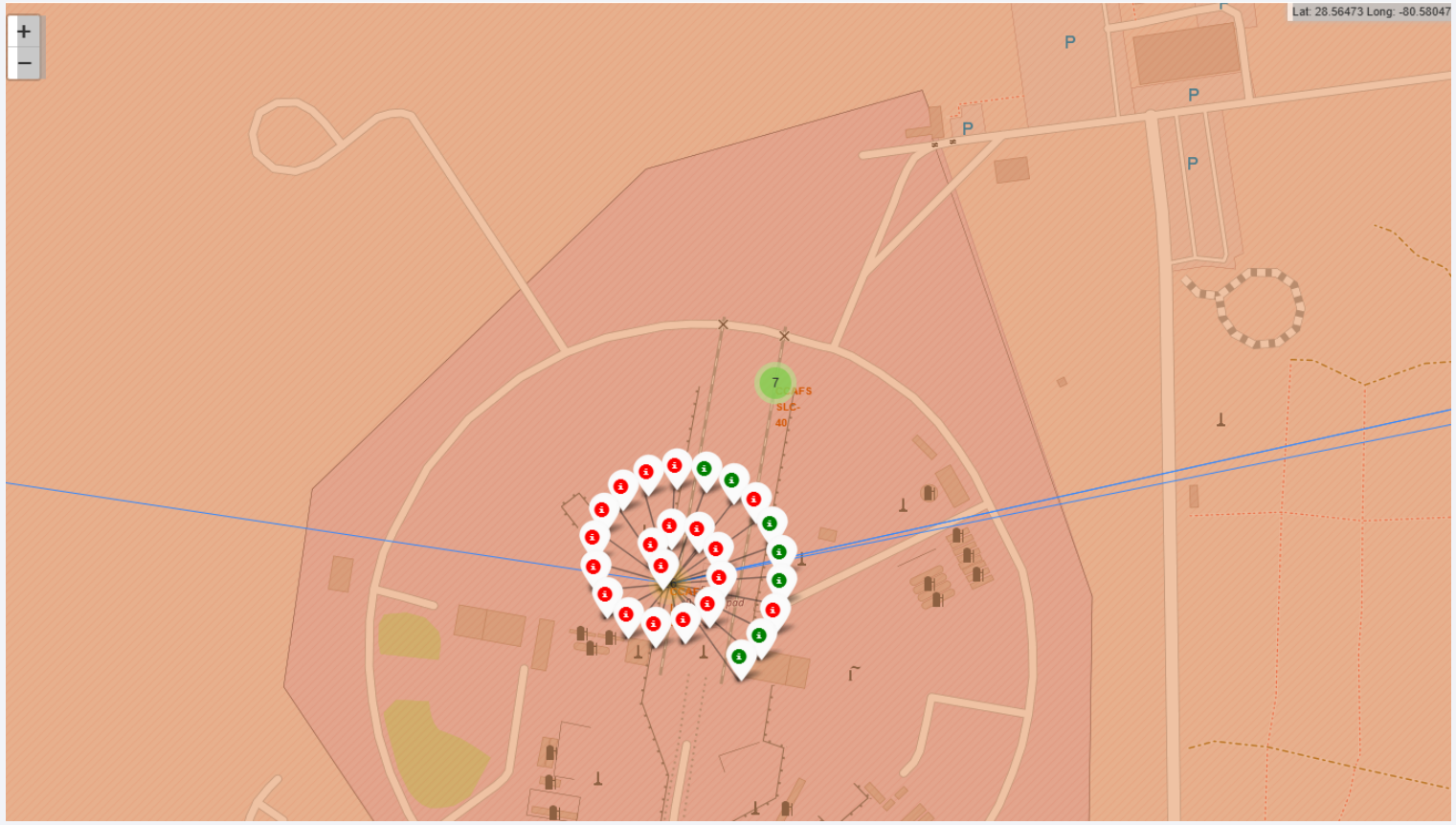
Launch Sites Proximities Analysis

Lauches Sites Locations on a Global Map



- We can see 10 Launches were in California, while 46 were in Florida

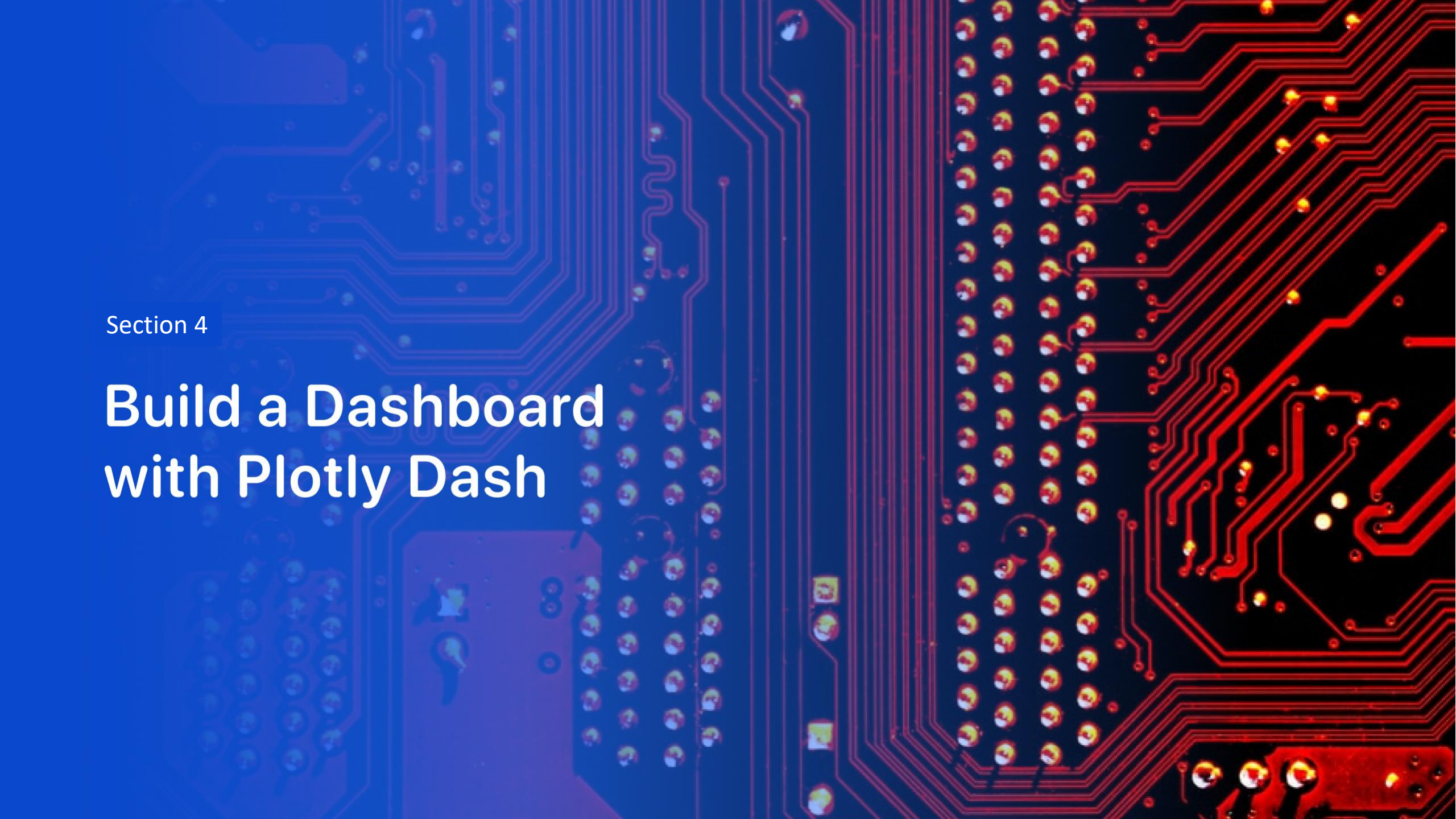
CCAFS LC-40 Landing Results



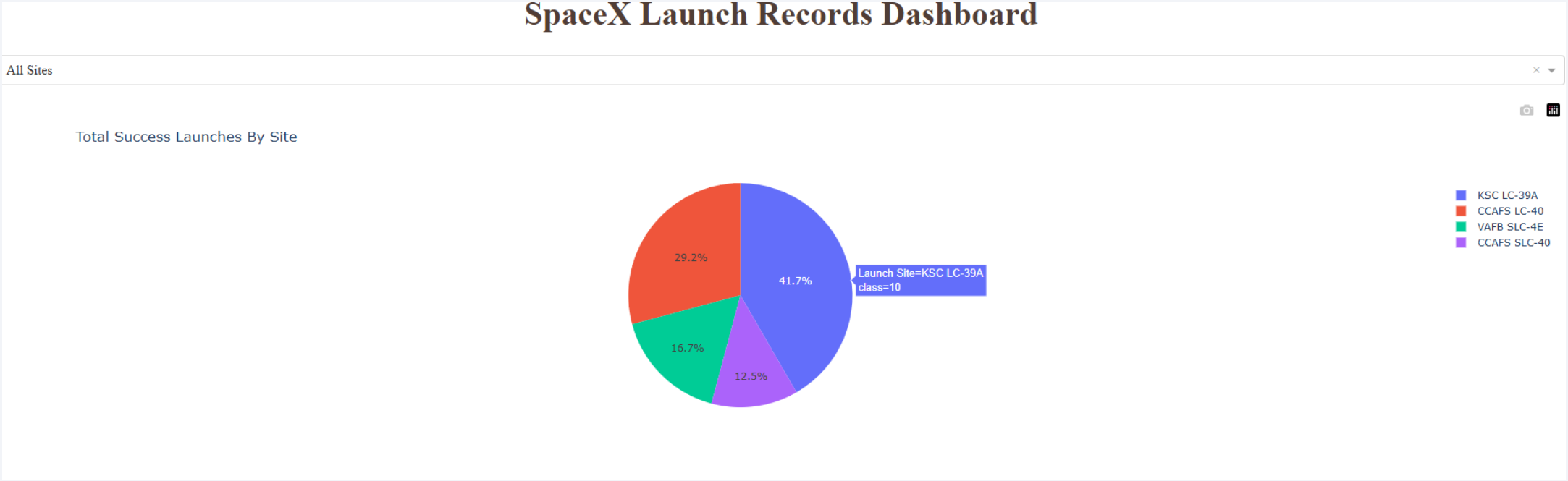
- Markers show Success(green) and Unsuccessful landing results (red)
- We can see most of the launches from CCAFS LC-40 were failure

CCAFS LC-40 Distances from Coastline, Road and Railway

- We can also observe Launching Sites located near the coastline (0.93 km) with a road(0.65 km) and a railway(0.95 km) nearby.

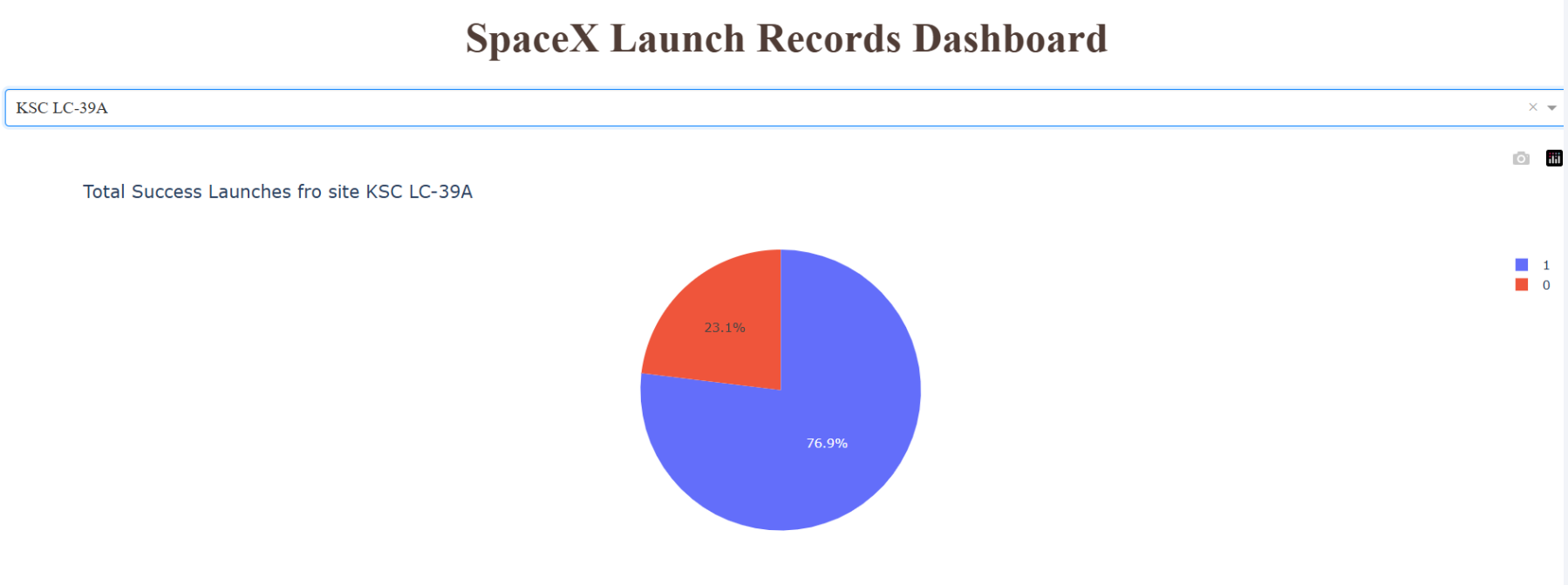

Section 4
Build a Dashboard with Plotly Dash

Total Success Lauches By Site



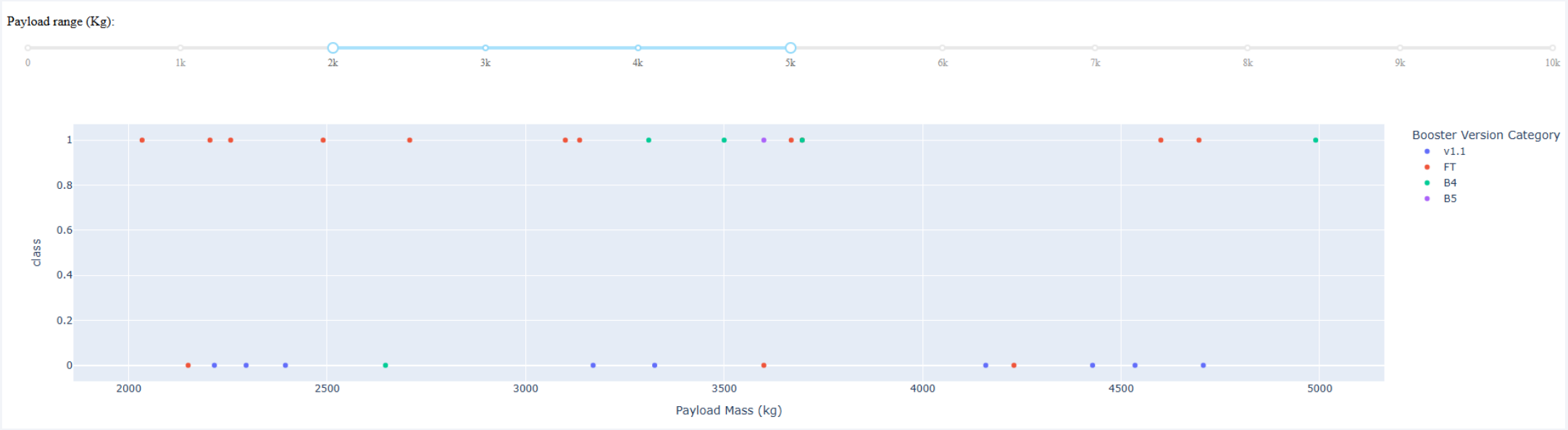
- KSC LC-39A has the highest number of Successful Launches

Total Success Launches for site KSC LC-39A



- KSC LC-39A also has the highest Successful Rate

The number of successful landings for Payload Range from 2000 to 5000

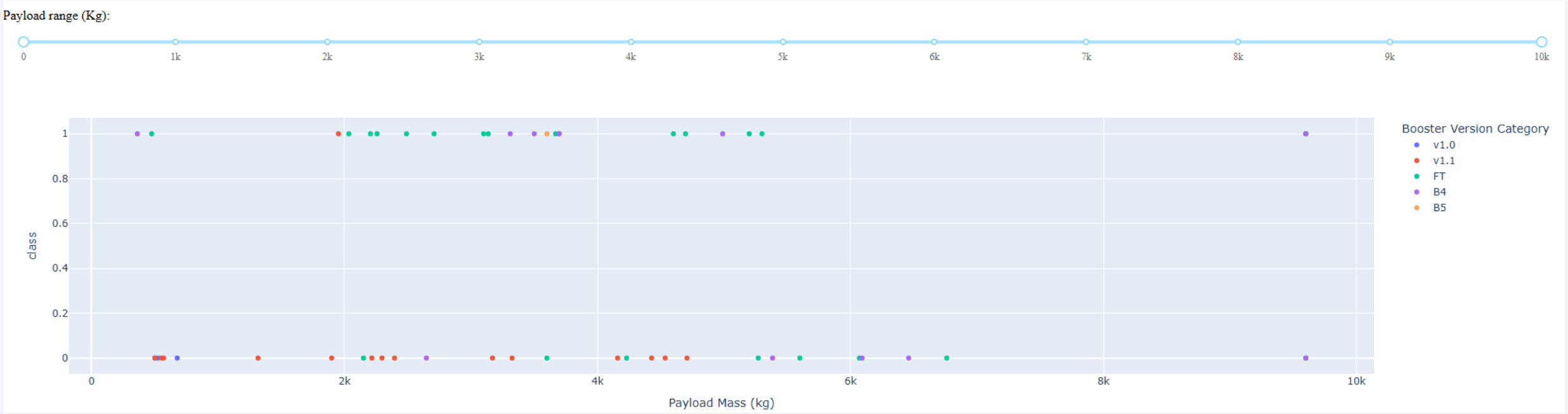


- Payload Range from 2,000 to 5,000 has the highest successful rate

The number of successful landings for Payload Range from 2000 to 5000

- Payload Range from 500 to 6,700 has the lowest successful rate

The number of successful landings for Payload Range from 2000 to 5000



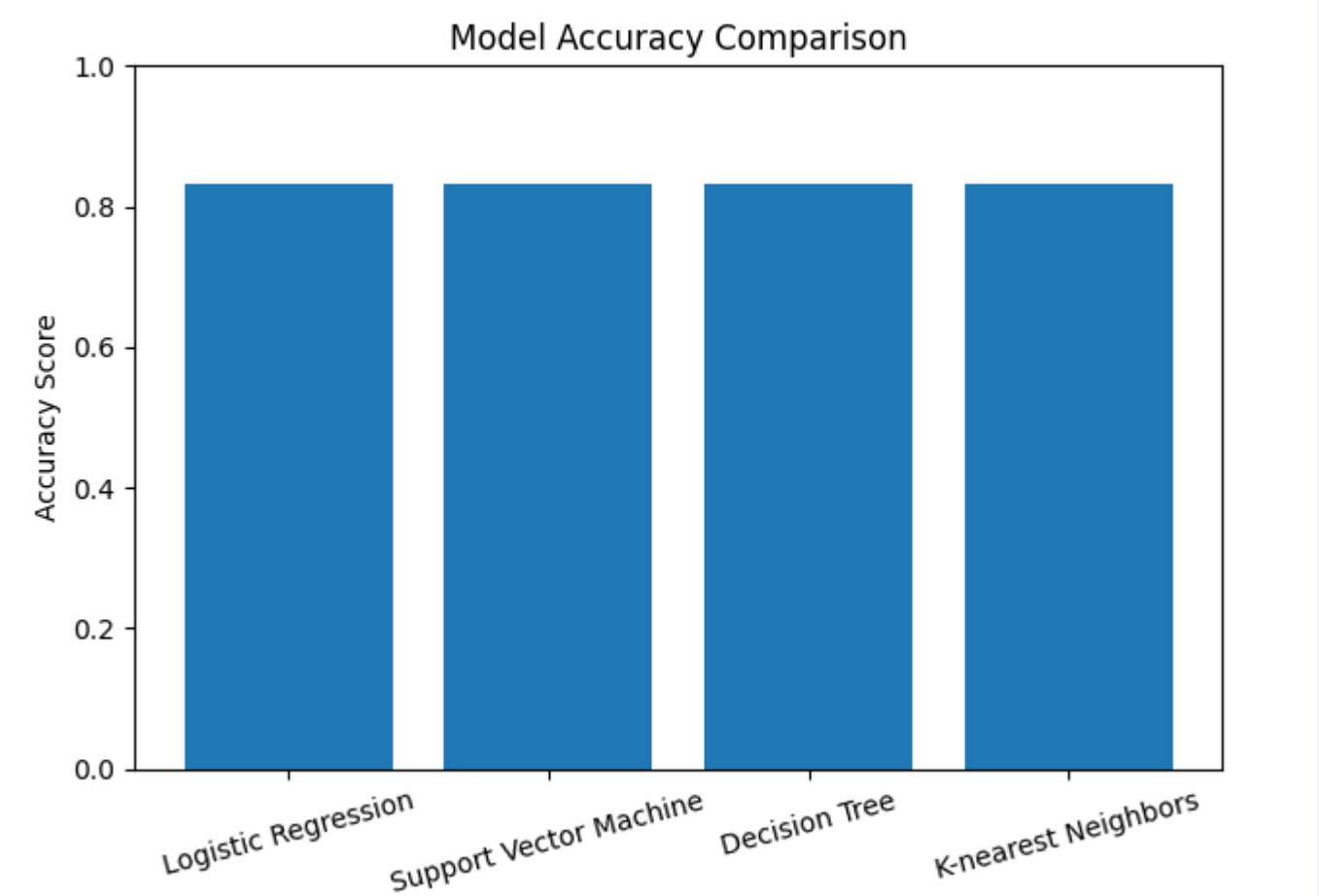
- Booster Version FT has the highest successful rate

Section 5

Predictive Analysis (Classification)

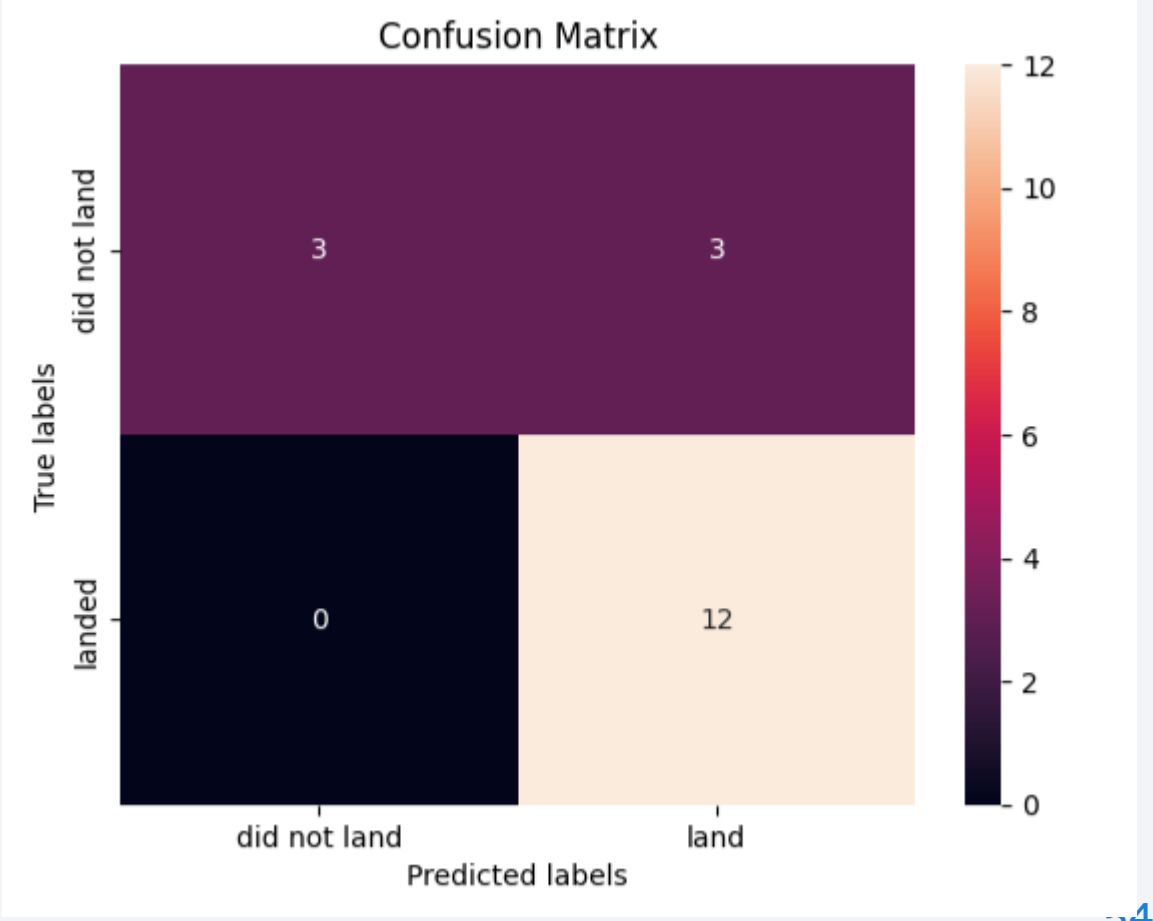
Classification Accuracy

- All models have a high Accuracy of 83%



Confusion Matrix

- Every model showed 12 True positives and 3 True Negatives errors, which makes it precise to predict successes
- Every model showed a good outcome of False Negative predictions, making it good for Failure Prediction.



Conclusions

- Launching Sites near the coastlines, roads, and railways has a good outcome for success
- Only conducting more launches may improve our success rate
- ES-L1, GEO, HEO and SSO orbits are the best fit for start since they success rate is high
- We need to pay attention to Booster Version FT of Falcon 9 which has a high success rate compared to others
- The deployed model gives 83% of accuracy based on 90 features

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- https://github.com/BookerSK/IBM-Applied-Data-Science-Capstone.git

Thank you!