IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michal Kaminski
31.08.2023

Outline

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

Executive Summary

The following methodologies were used to analyze data:

- Data Collection using web scraping and SpaceX API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics
- Machine Learning Predictions

Summary of all results

- It was possible to collect valuable data from public sources
- EDA allowed to identify which features are the best to predict the success of launches
- Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity the best, using all of the collected data

Introduction

SpaceX is a revolutionary company who has disrupt the space industry by offering a rocket launches specifically Falcon 9 as low as 62 million dollars; while other providers cost upward of 165 million dollar each. Most of this saving thanks to SpaceX astounding idea to reuse the first stage of the launch by re-land the rocket to be used on the next mission. Repeating this process will make the price down even further. As a data scientist of a startup rivaling SpaceX, the goal of this project is to create the machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

The problems included:

- Identifying all factors that influence the landing outcome.
- The relationship between each variables and how it is affecting the outcome.
- The best condition needed to increase the probability of successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping Wikipedia
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters

Data Collection

Data sets were collected from:

- Space X API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia using web scraping technics (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

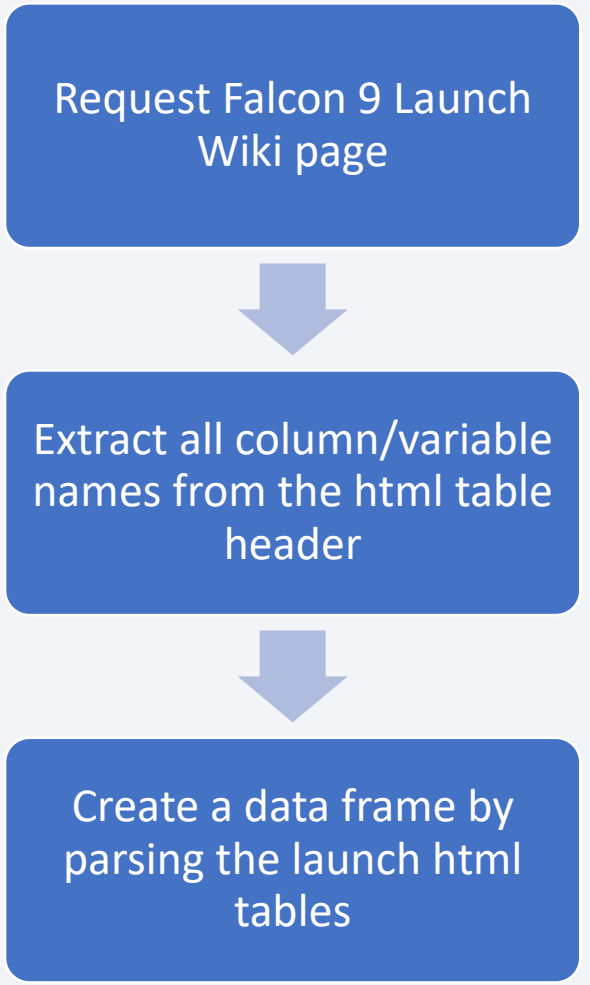
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.

Source code: Applied-Data-Science-Capstone/notebook_Data_Collection_API_Lab_nY4pSxYdh.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)

Data Collection - Scraping

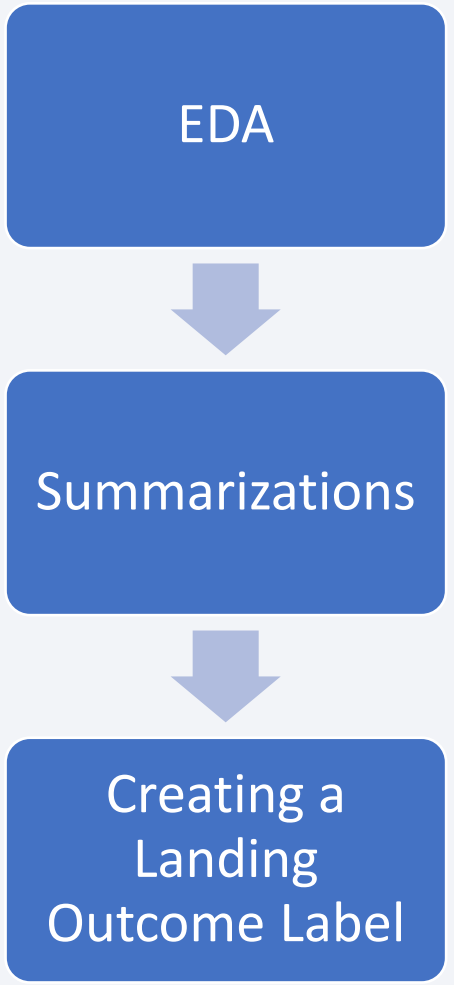
- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted



Source code: [Applied-Data-Science-Capstone/notebook_Data_Collection_with_Web_Scraping_ZeOcMoPOS.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)]

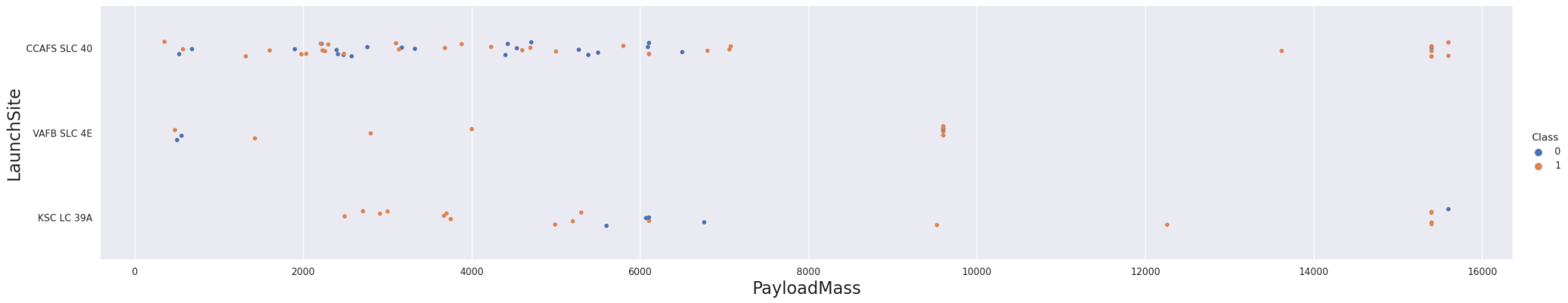
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column



Source code:Applied-Data-Science-Capstone/notebook_Data Wrangling.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)

EDA with Data Visualization



To analyze the data, scatterplots and barplots were employed to visualize connections among various features:

- The connection between Payload Mass and Flight Number,
- The correlation of Launch Site and Flight Number,
- The relationship between Launch Site and Payload Mass,
- The interplay of Orbit and Flight Number,
- The interrelation between Payload and Orbit.

Source Code: [Applied-Data-Science-Capstone/notebook_EDA_with_Visualization_lab_Ly6nPbDn2.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)]

EDA with SQL

Source Code: mickamin/Applied-Data-Science-Capstone (github.com)

Through the utilization of SQL, a multitude of queries were executed to enhance our comprehension of the dataset. These included tasks such as:

- Extracting the names of the launch sites.
- Presenting five records featuring launch sites commencing with the 'CCA' string.
- Calculating the cumulative payload mass transported by boosters launched by NASA under the CRS program.
- Determining the average payload mass carried by booster version F9 v1.1.
- Enumerating the date of the inaugural successful ground pad landing outcome.
- Listing the names of boosters that achieved success on drone ships, with payload mass exceeding 4000 and under 6000.
- Enumerating the total count of both successful and failed mission outcomes.
- Displaying the booster versions responsible for carrying the maximum payload mass.
- Listing failed landing outcomes on drone ships, along with their corresponding booster versions and launch site names, limited to the year 2015.
- Ranking the frequency of landing outcomes or success within the time span from 2010-06-04 to 2017-03-20, arranged in descending order.

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.

Source Code: Applied-Data-Science-Capstone/notebook_Interactive_Visual_Analytics_with_Folium_rgzMRq7YW.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)

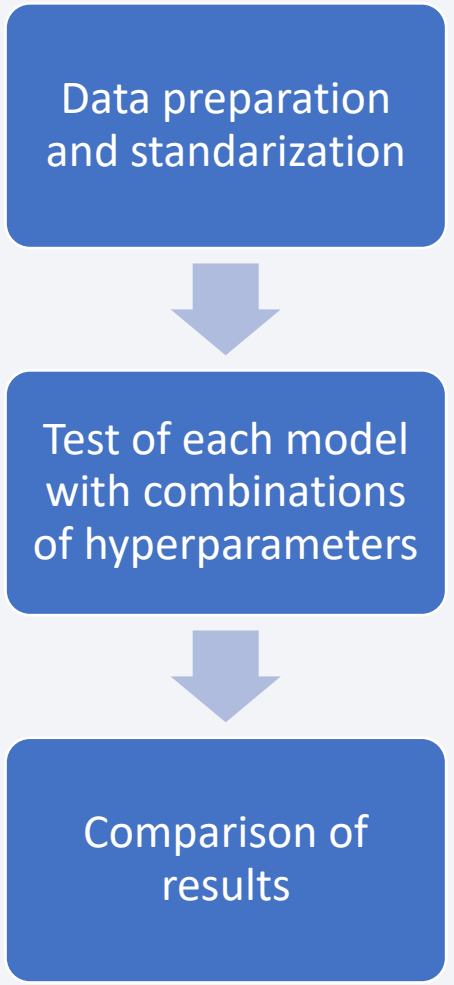
Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash which allowing the user to play around with the data as they need.
- We plotted pie charts showing the total launches by a certain sites.
- We then plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

Source Code: Applied-Data-Science-Capstone/spacex_dash_app.py at main · mickamin/Applied-Data-Science-Capstone (github.com)

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

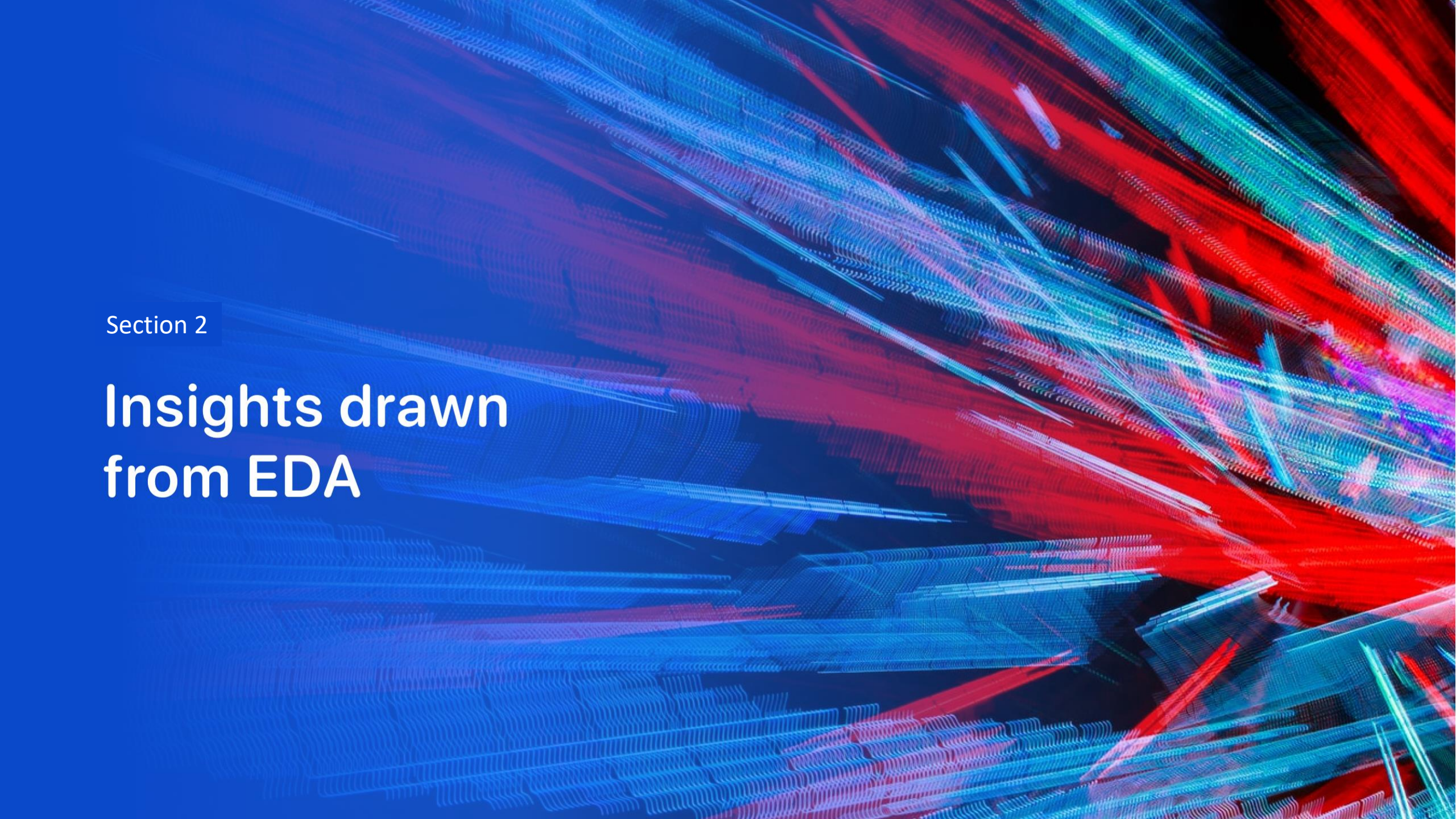


Source Code: [Applied-Data-Science-Capstone/notebook_Machine_Learning_Prediction_qCUY0t8Yh.ipynb at main · mickamin/Applied-Data-Science-Capstone (github.com)]

Results

Results:

- Findings from exploratory data analysis include:
- SpaceX operates from four distinct launch sites.
- Initial launches were directed towards both SpaceX and NASA.
- The average payload for the F9 v1.1 booster amounts to 2,928 kg.
- The first successful landing took place in 2015, five years following the initial launch.
- Numerous iterations of Falcon 9 boosters exhibited successful drone ship landings with payloads surpassing the average.
- Nearly every mission outcome achieved a success rate close to 100%.
- In 2015, two booster versions, namely F9 v1.1 B1012 and F9 v1.1 B1015, experienced landing failures on drone ships.
- Over the course of years, there was a noticeable enhancement in the number of successful landing outcomes.

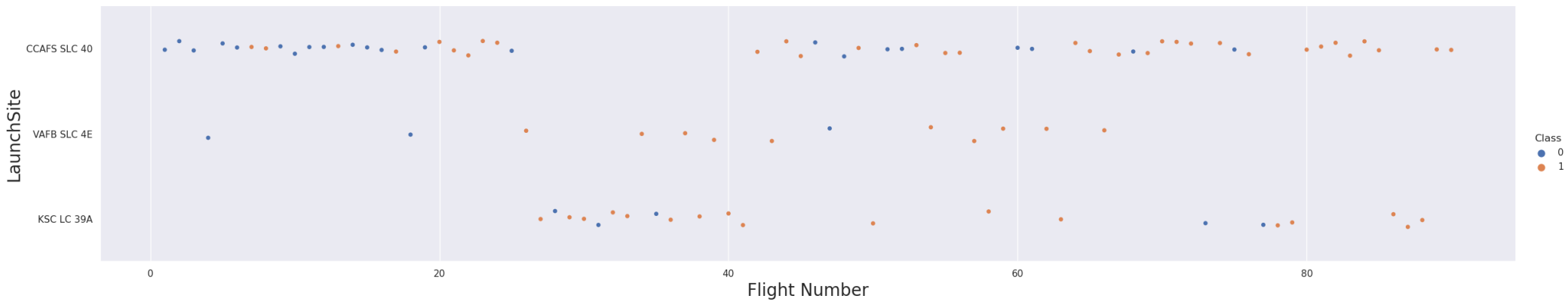

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Based on the presented graph, we can deduce that the current premier launch site is CCAF5 SLC 40, exhibiting a substantial number of recent successful launches.

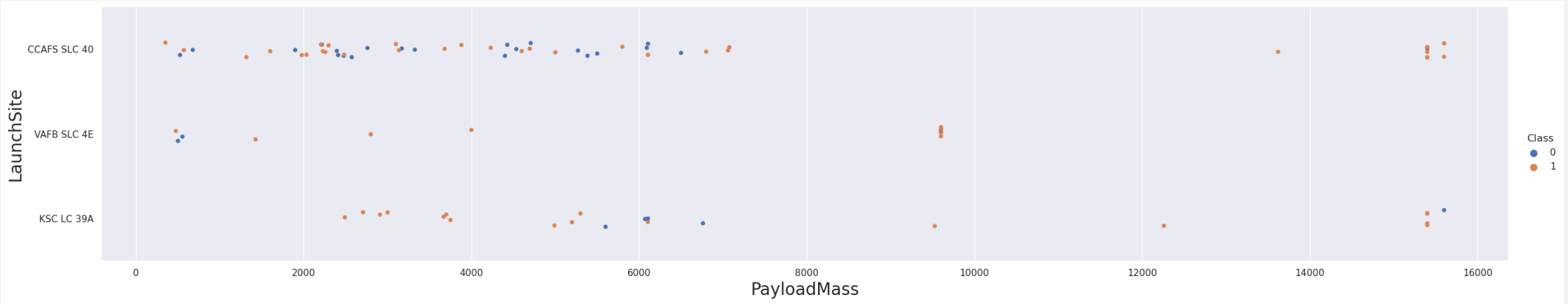
Following closely is VAFB SLC 4E in second place, succeeded by KSC LC 39A in third position.

Furthermore, the graph illustrates a progressive upswing in the overall success rate over the course of time.



Payload vs. Launch Site

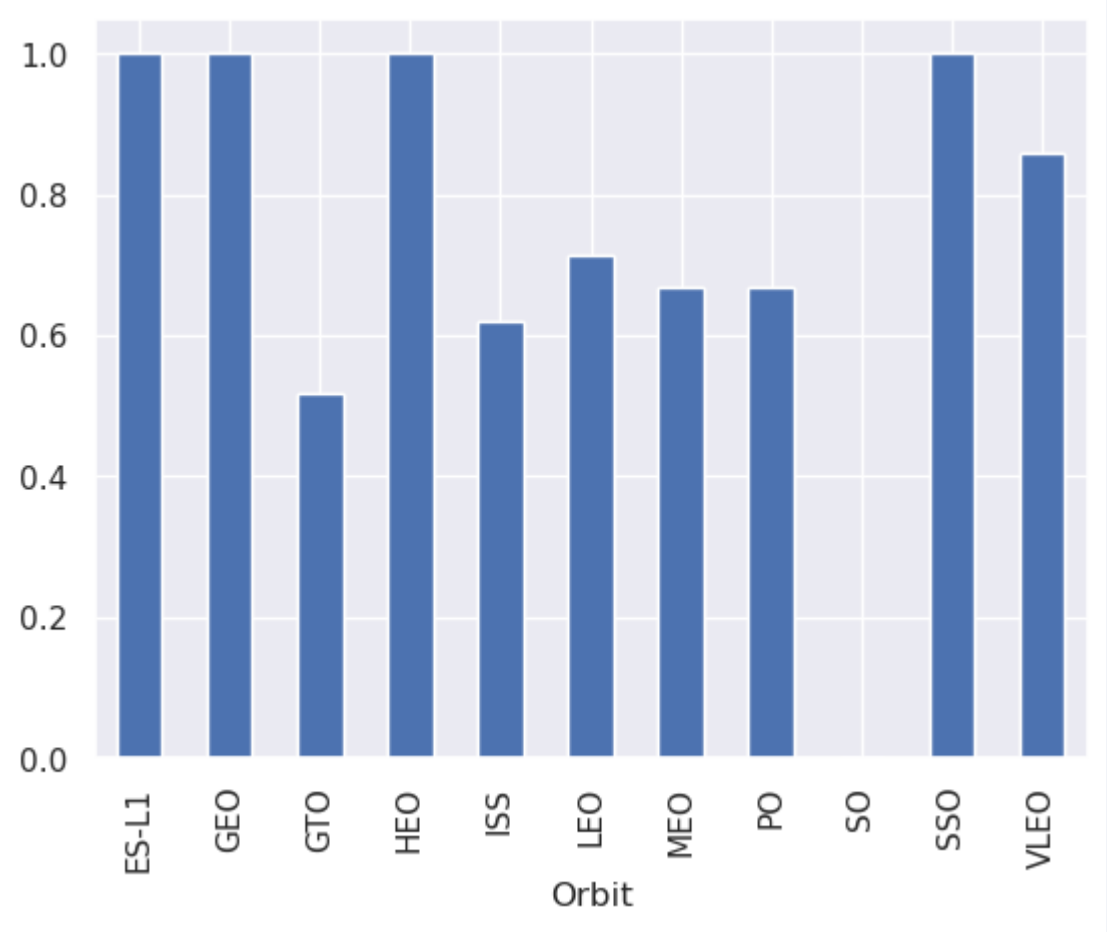
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



Success Rate vs. Orbit Type

The biggest success rates:

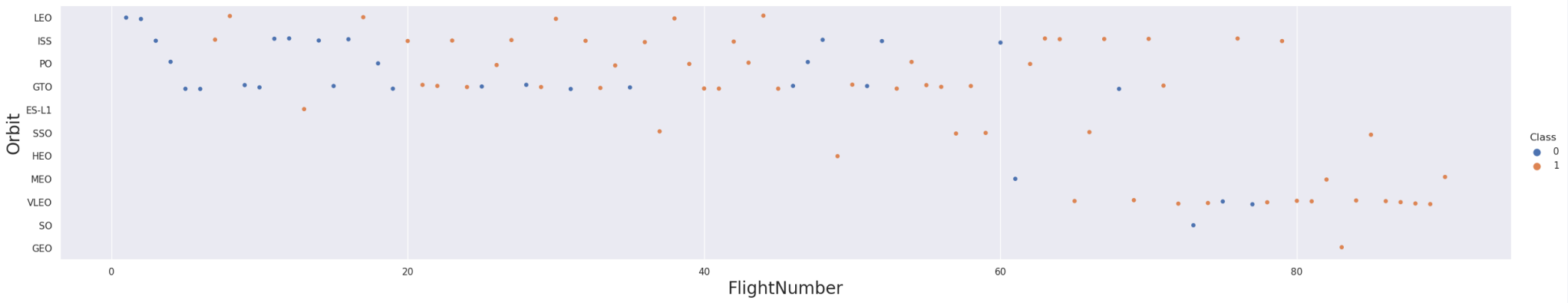
- ES-L1
- GEO
- HEO
- SSO



Flight Number vs. Orbit Type

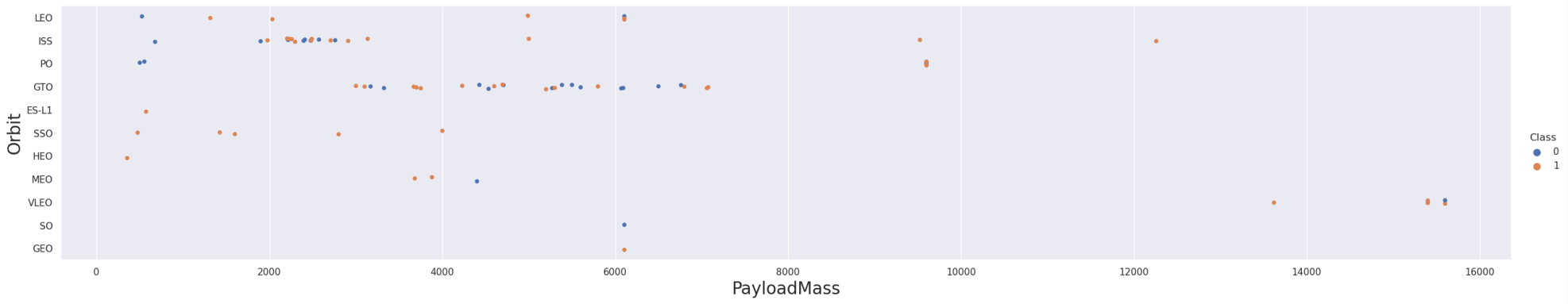
Evidently, there has been a noticeable enhancement in the success rate across all orbits as time has progressed.

Notably, the frequency of VLEO (Very Low Earth Orbit) missions has recently risen, presenting a potential new avenue for business opportunities.



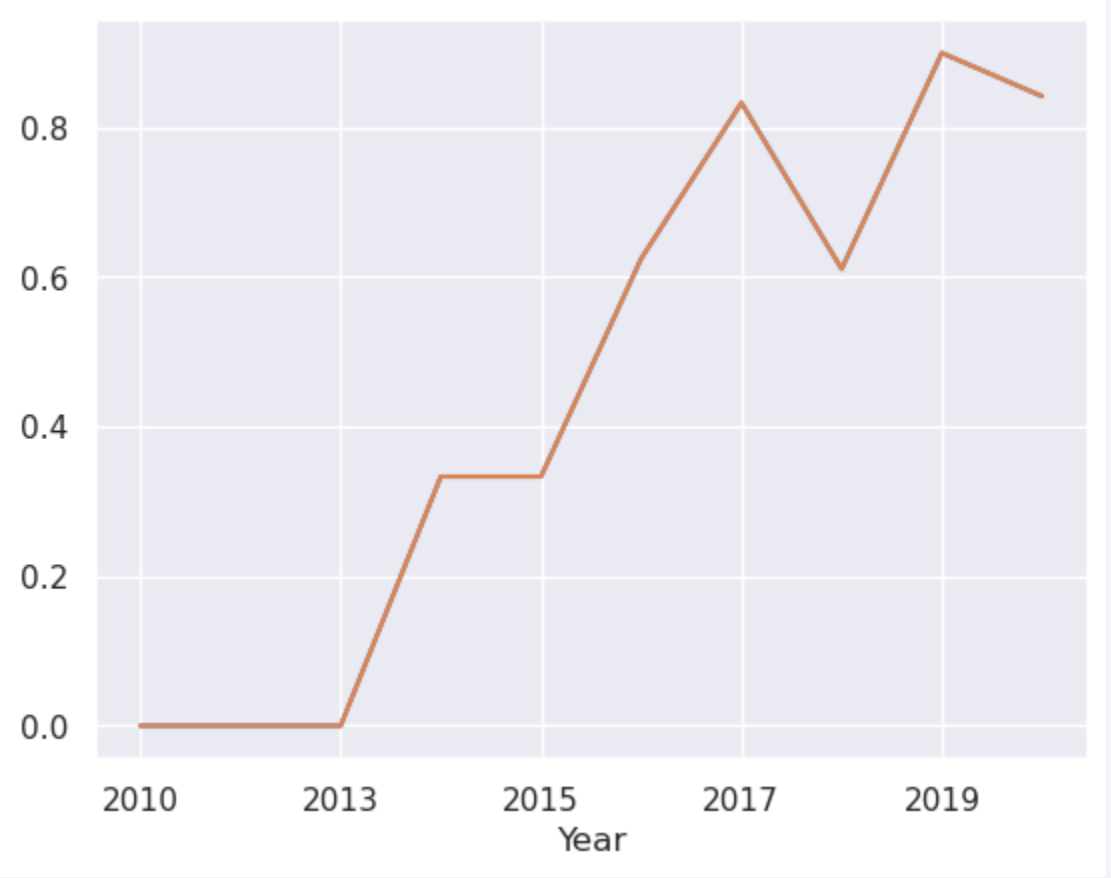
Payload vs. Orbit Type

- It appears that there is no discernible correlation between payload and success rate in the case of GTO (Geostationary Transfer Orbit).
- The orbit associated with the International Space Station (ISS) displays a notable range of payload values along with a commendable success rate.
- A limited number of launches are directed towards the SO (Sun-Synchronous Orbit) and GEO (Geostationary Orbit) categories.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

We used the key word DISTINCT to show only unique launch sites from the SpaceX data:

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

We used the query below to display 5 records where launch sites begin with `CCA`:

```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

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

Total Payload Mass

We calculated the total payload carried by boosters from NASA as 45596 using the query below:

```
In [30]: %sql SELECT SUM (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)' ;
 * sqlite:///my_data1.db
Done.
Out[30]: 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

```
In [40]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[40]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

We use the min() function to find the result We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015:

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

Hint:Use min function

In [32]:
```
%sql SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

Out[32]:

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
In [33]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND
 * sqlite:///my_data1.db
Done.
```

Out[33]:

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

```
In [34]: sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
 * sqlite:///my_data1.db
Done.
```

Out[34]:

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

Boosters Carried Maximum Payload

We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function:

In [35]:
```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[35]:

Booster Versions which carried the Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

We used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015:

```
In [36]: sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND strftime('%Y', DA
 * sqlite:///my_data1.db
Done.
```

Out[36]:

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

Based on the presented graph, we can deduce that the current premier launch site is CCAF5 SLC 40, exhibiting a substantial number of recent successful launches.

Following closely is VAFB SLC 4E in second place, succeeded by KSC LC 39A in third position.

Furthermore, the graph illustrates a progressive upswing in the overall success rate over the course of time.

In [37]:
```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDIN
```

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

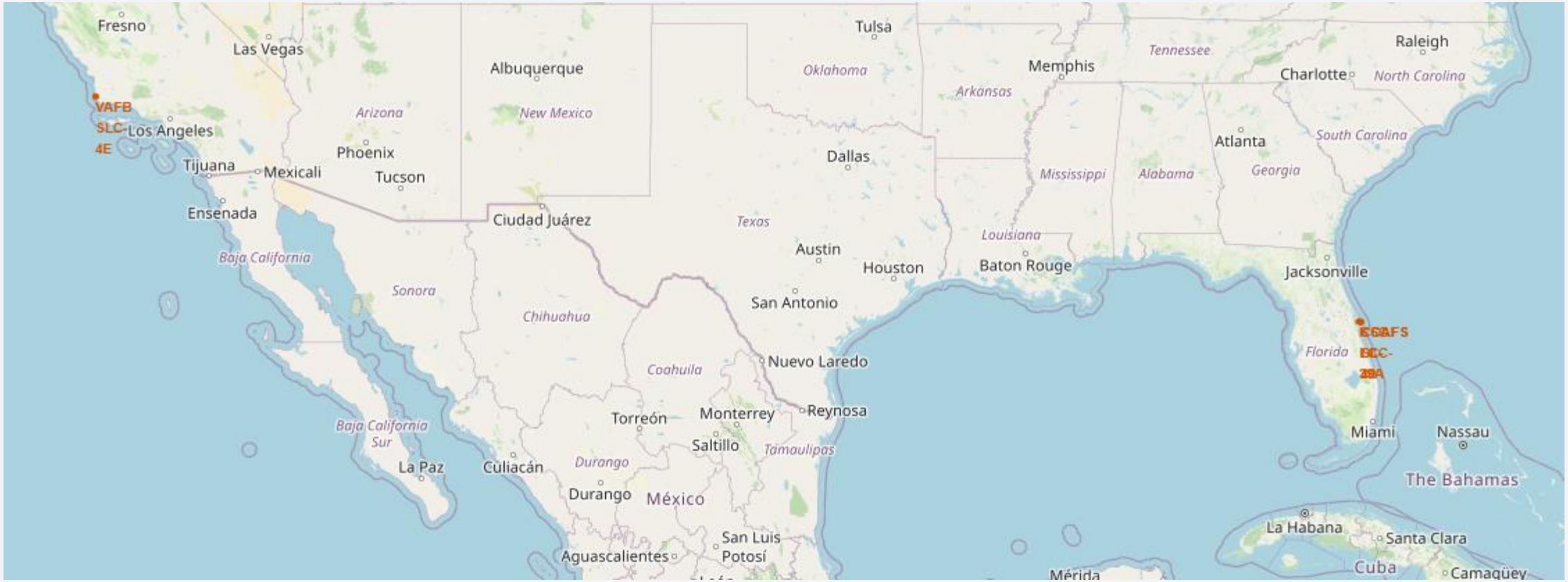
Out[37]:

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

Section 3

Launch Sites Proximities Analysis

All launch sites



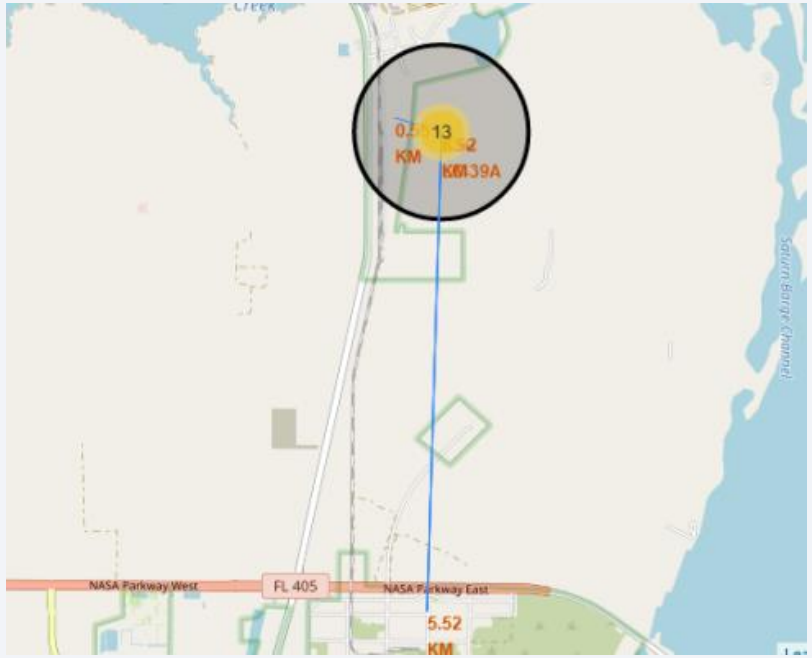
Launch sites are near sea, probably by safety, but not too far from roads and railroads

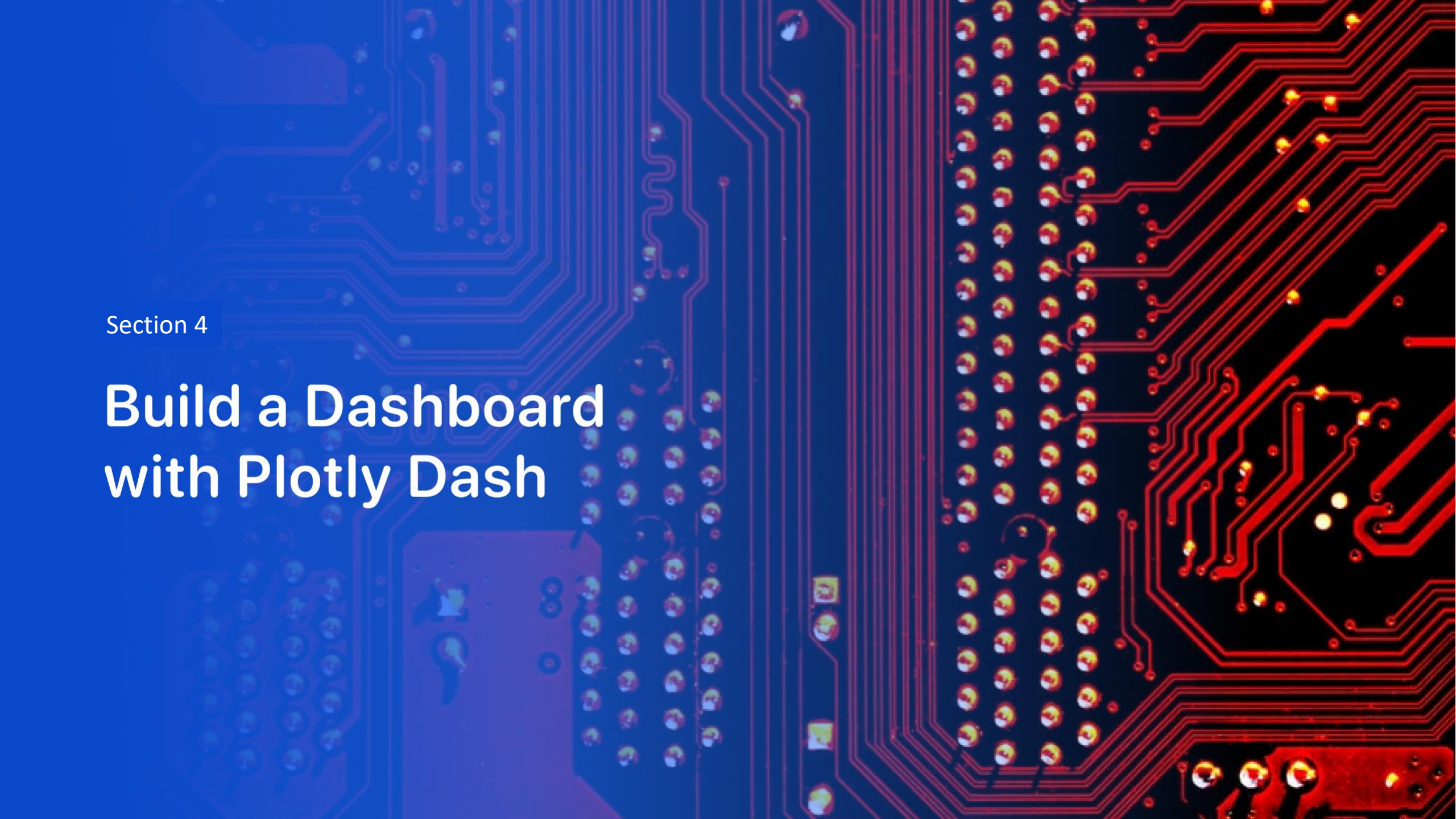
Launch outcomes

Example of KSC LC-39A launch site launch outcomes (green markers- succesful launches, red markers- failure)

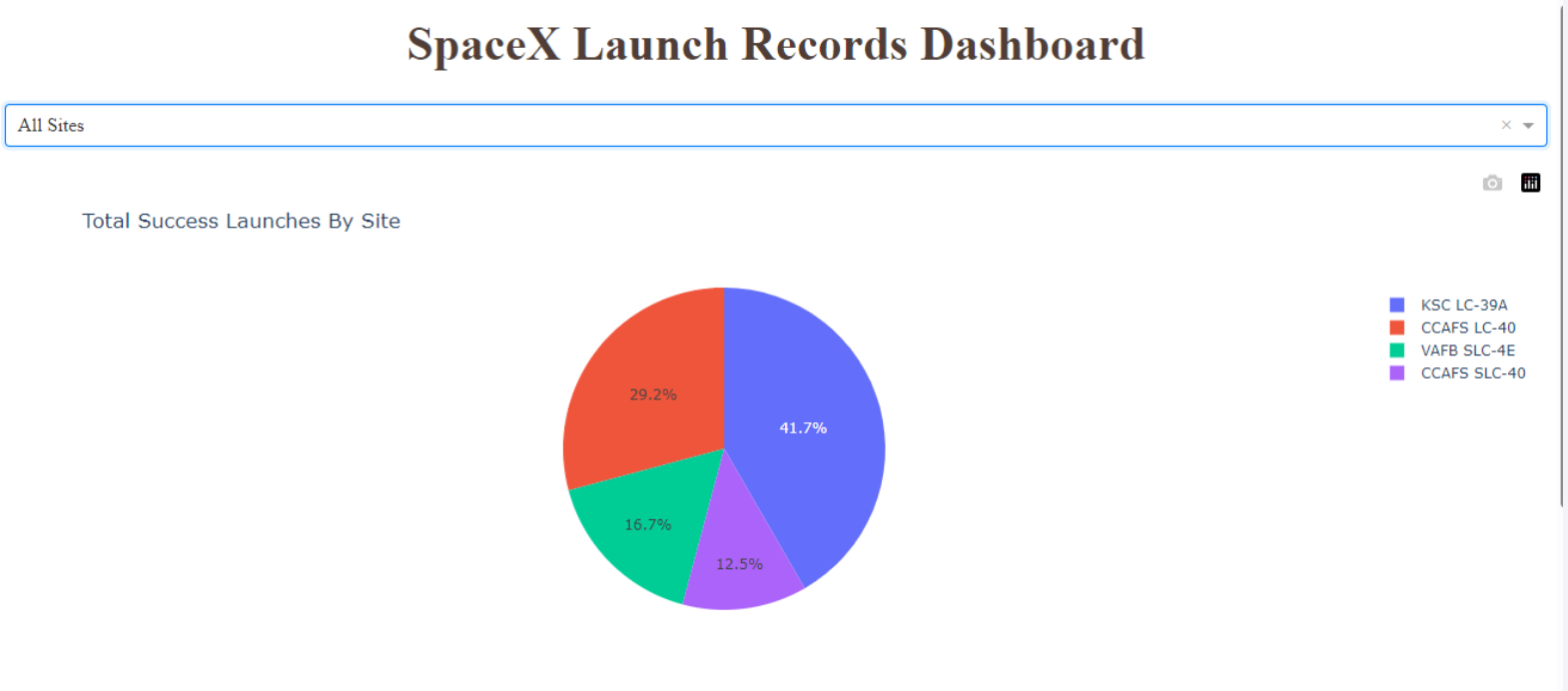
Logistics

Launch site KSC LC-39A has good logistics aspects, being near a railroad/road and relatively far from inhabited areas.



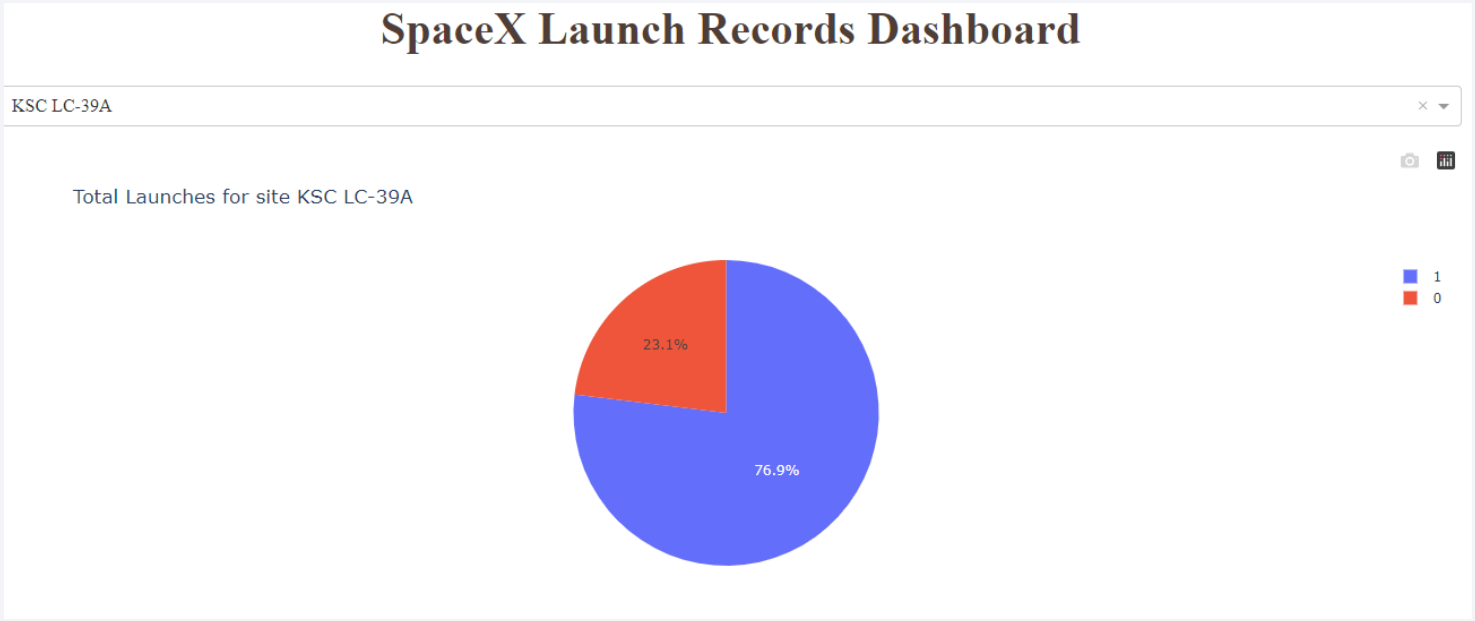

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



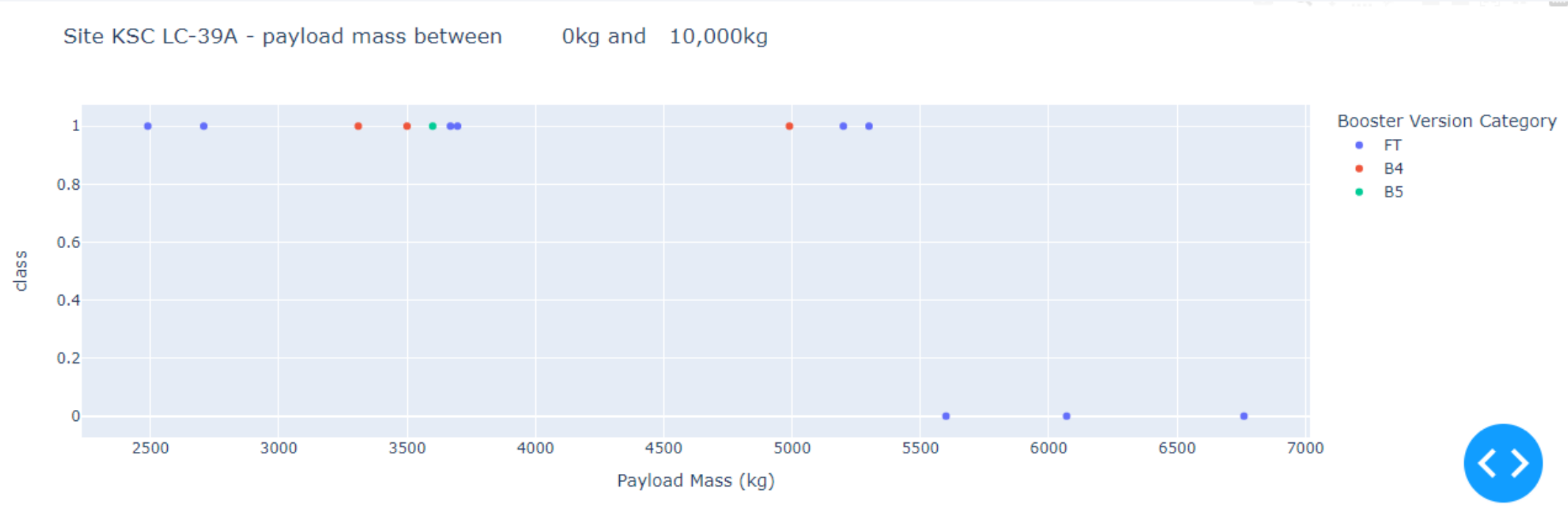
We can see that KSC LC-39A had the most successful launches from all the sites

The highest launch-success ratio for KSC LC-39A



KSC LC-39A had a 76.9% success rate

Launch outcome vs payload scatter plot



We can see that all the success rate for low weighted payload is higher than heavy weighted payload.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Find the method performs best:

In [56]:

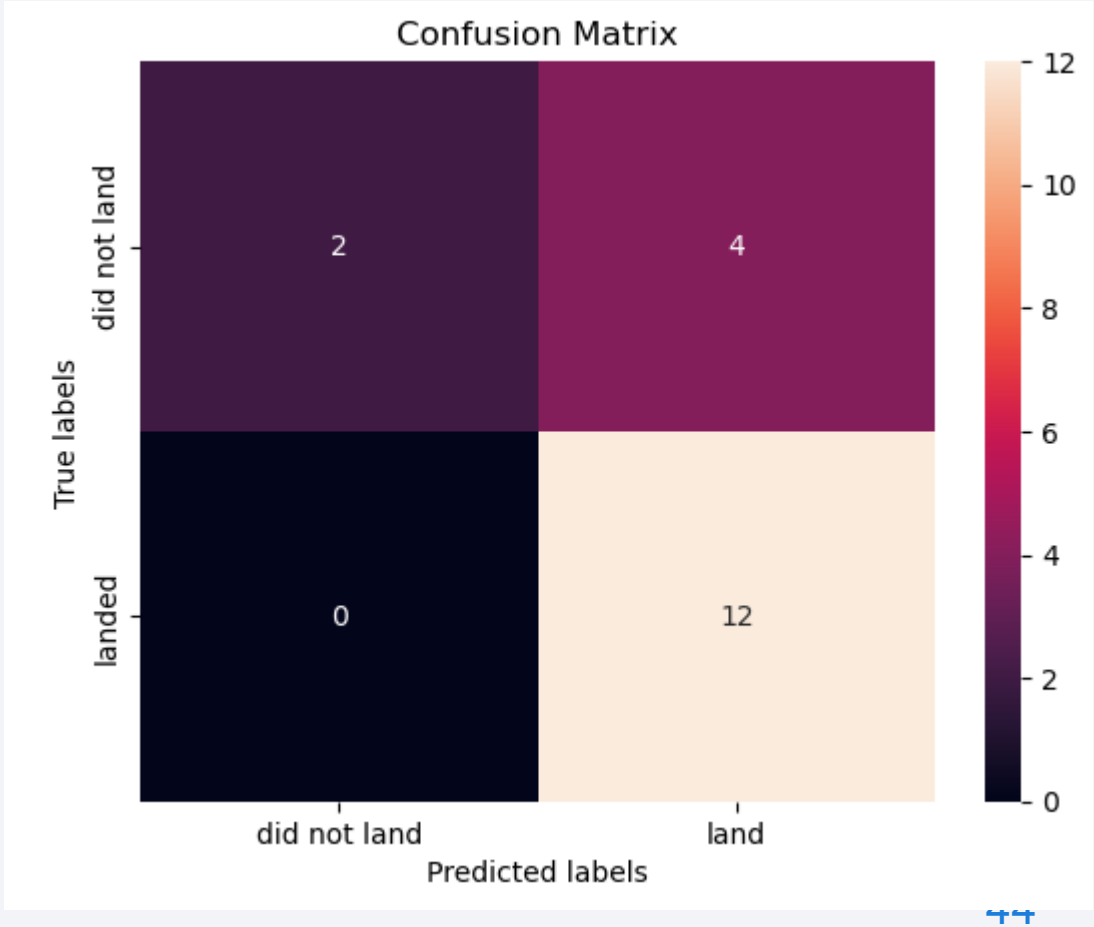
```
print("Model\t\tAccuracy\tTestAccuracy")#,logreg_cv.best_score_)
print("LogReg\t\t{}\t\t{}".format((logreg_cv.best_score_).round(5), logreg_cv.score(X_test, Y_test).round(5)))
print("SVM\t\t{}\t\t{}".format((svm_cv.best_score_).round(5), svm_cv.score(X_test, Y_test).round(5)))
print("Tree\t\t{}\t\t{}".format((tree_cv.best_score_).round(5), tree_cv.score(X_test, Y_test).round(5)))
print("KNN\t\t{}\t\t{}".format((knn_cv.best_score_).round(5), knn_cv.score(X_test, Y_test).round(5)))

comparison = {}

comparison['LogReg'] = {'Accuracy': logreg_cv.best_score_.round(5), 'TestAccuracy': logreg_cv.score(X_test, Y_test).round(5
comparison['SVM'] = {'Accuracy': svm_cv.best_score_.round(5), 'TestAccuracy': svm_cv.score(X_test, Y_test).round(5)}
comparison['Tree'] = {'Accuracy': tree_cv.best_score_.round(5), 'TestAccuracy': tree_cv.score(X_test, Y_test).round(5)}
comparison['KNN'] = {'Accuracy': knn_cv.best_score_.round(5), 'TestAccuracy': knn_cv.score(X_test, Y_test).round(5)}
```

```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.88929         0.77778
KNN             0.84821         0.83333
```

As we can see, by using the code above- the best algorithm isTree Algorithm which has the highest classification accuracy.

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier



Conclusions

We can conclude that:

- Tree Classifier Algorithm excels as a machine learning choice for this dataset.
- Lighter payloads (≤4000kg) consistently outperform heavier ones.
- SpaceX's launch success has notably improved since 2013, indicating a positive trend.
- KSC LC-39A is the top-performing launch site with a 76.9% success rate.
- The Sun-Synchronous Orbit (SSO) boasts a perfect 100% success rate in multiple launches.

Thank you!