IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nayanthara James
01/10/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data analysis result
  - Interactive analytics in screenshots
  - Predictive analysis result

Introduction

- Project background and context

  SpaceX offers Falcon 9 rocket launches on its website for 62 million dollars; other companies charge up to 165 million dollars apiece; much of the savings is due to SpaceX's ability to reuse the first stage. As a result, if we can predict whether the first stage will land, we can estimate the cost of a launch. This data can be utilized if another firm wishes to compete with SpaceX for a rocket launch. The goal of the project is to create a machine learning model that can predict whether or not the initial stage will land successfully

Section 1

Methodology

Methodology

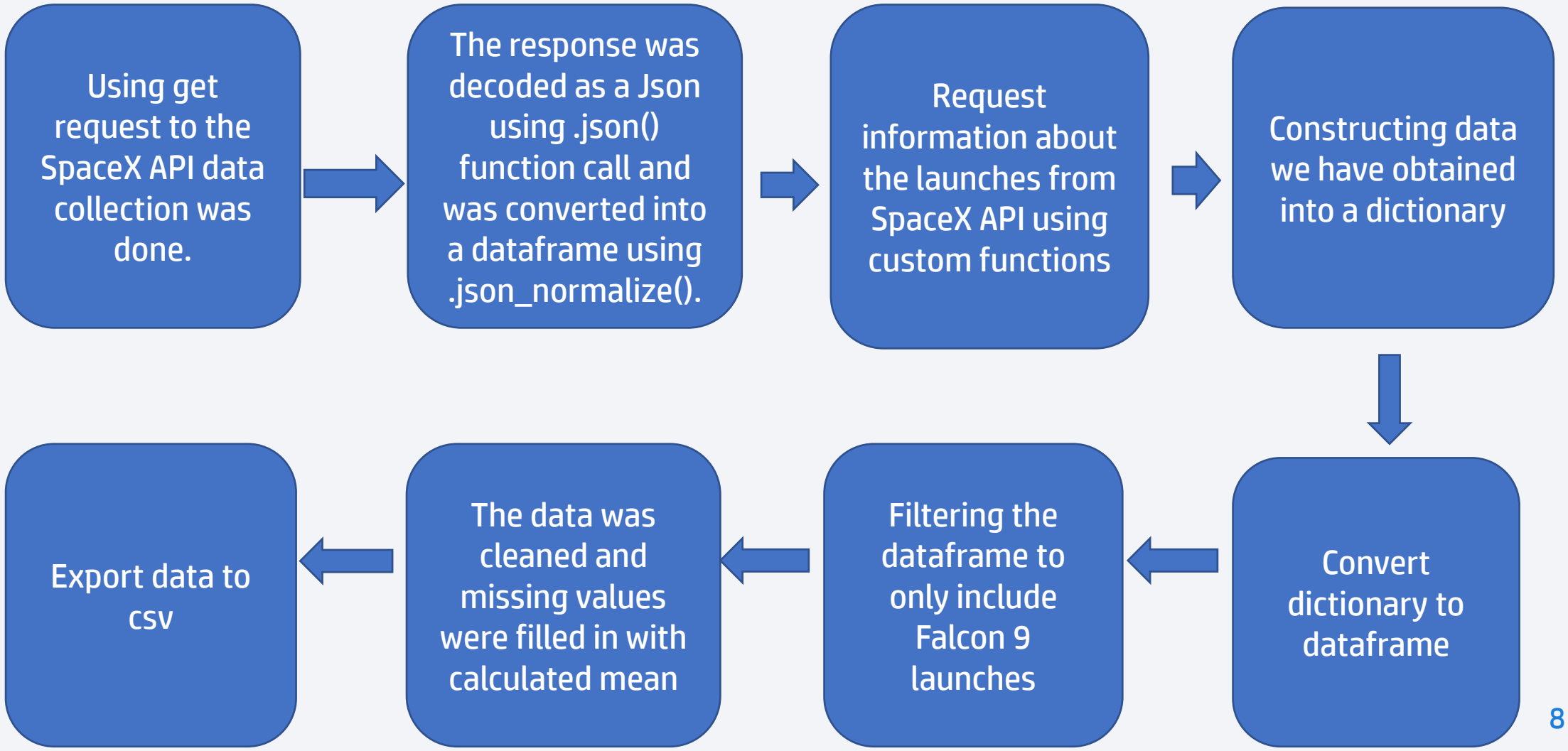
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - Clean the data, fill in missing values and apply one hot coding to prepare the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

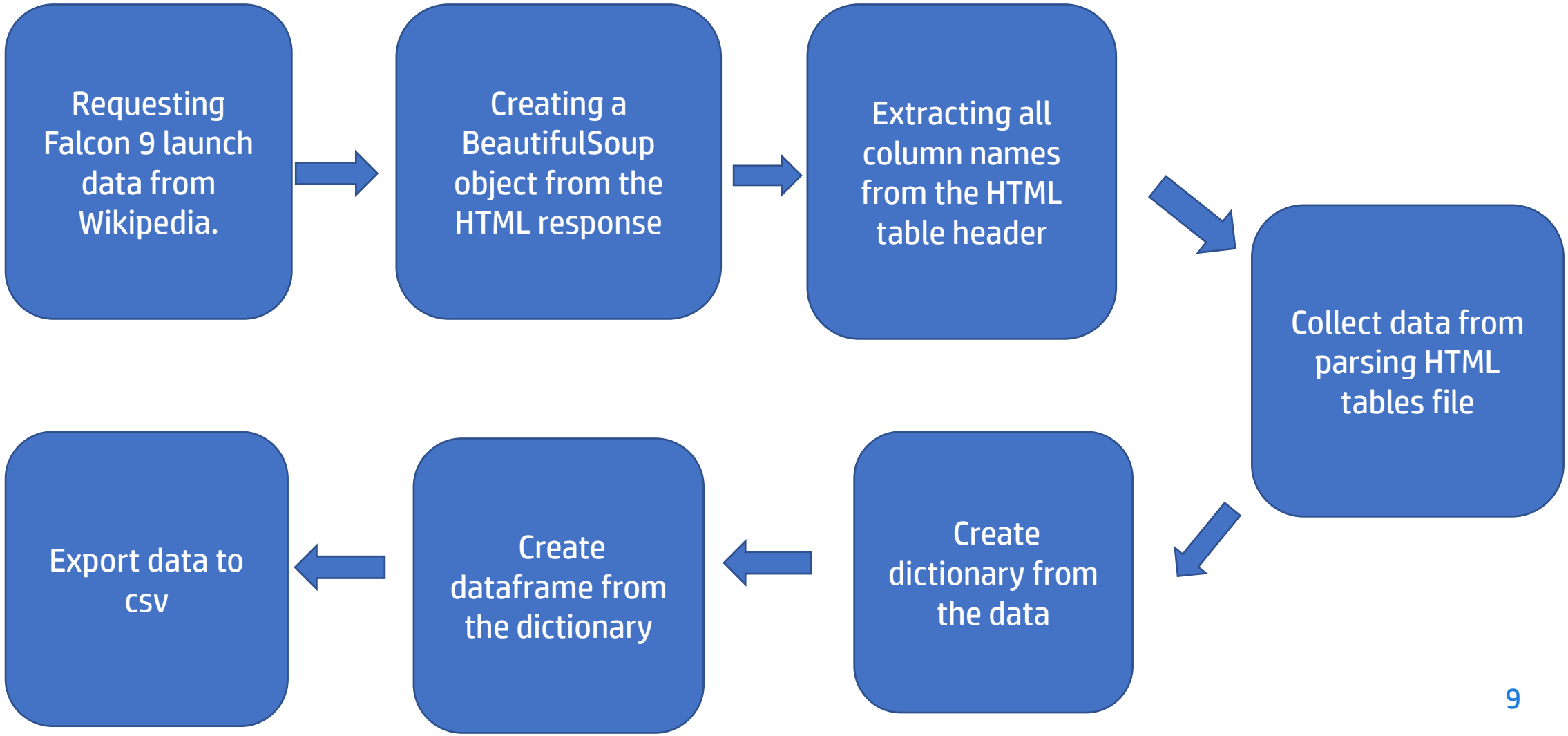
Data Collection

- The data was collected using various methods
  - Using get request to the SpaceX API data collection was done.
  - The response was decoded as a Json using .json() function call and was converted into a dataframe using .json_normalize().
  - The data was cleaned and missing values were filled in with calculated mean.
  - Falcon 9 launch data was scraped from Wikipedia using BeautifulSoup.
  - The launch data was extracted as HTML table, parse the table and convert it to a dataframe.
- The link to the notebook is https://github.com/nj0528/testrep/blob/main/jupyter-labs-spacex-data-collection-api.ipynb.
- The link to the notebook is https://github.com/nj0528/testrep/blob/main/jupyter-labs-webscraping.ipynb.

Data Collection – SpaceX API



Data Collection – Scraping



Data Wrangling

Perform EDA and determine data labels

Calculate the number of
- launches on each site
- each type of orbit and its occurrence
- of mission outcome per orbit type and its occurrence

Create a landing outcome label from Outcome column

Export data to csv file

- Landing was not always successful
- True Ocean: mission outcome had a successful landing to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
- True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.
- True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship
- Outcomes converted into 1 for a successful landing and 0 for an unsuccessful landing
- The link to the notebook is https://github.com/nj0528/testrep/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb.

EDA with Data Visualization

- The relationship that:

  -Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type and Payload Mass vs Orbit Type,

  have on launch outcome were plotted using Scatterplot.
- A bar chart was plotted to identify the relationship between success rate and orbit type.
- A line chart was plotted to visualize the launch success yearly trend.
- The link to the notebook is https://github.com/nj0528/testrep/blob/main/edadataviz.ipynb

EDA with SQL

- Performed EDA with SQL queries to display the:

  - names of the unique launch sites in the space mission 5 records where launch sites begin with the string 'CCA' , total payload mass carried by boosters launched by NASA (CRS), average payload mass carried by booster version F9 v1.1

- To list the

  - date when the first successful landing outcome in ground pad was achieved, names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, total number of successful and failure mission outcomes, names of the booster versions which have carried the maximum payload mass, failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

- and 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

- The link to the notebook is https://github.com/nj0528/testrep/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb.

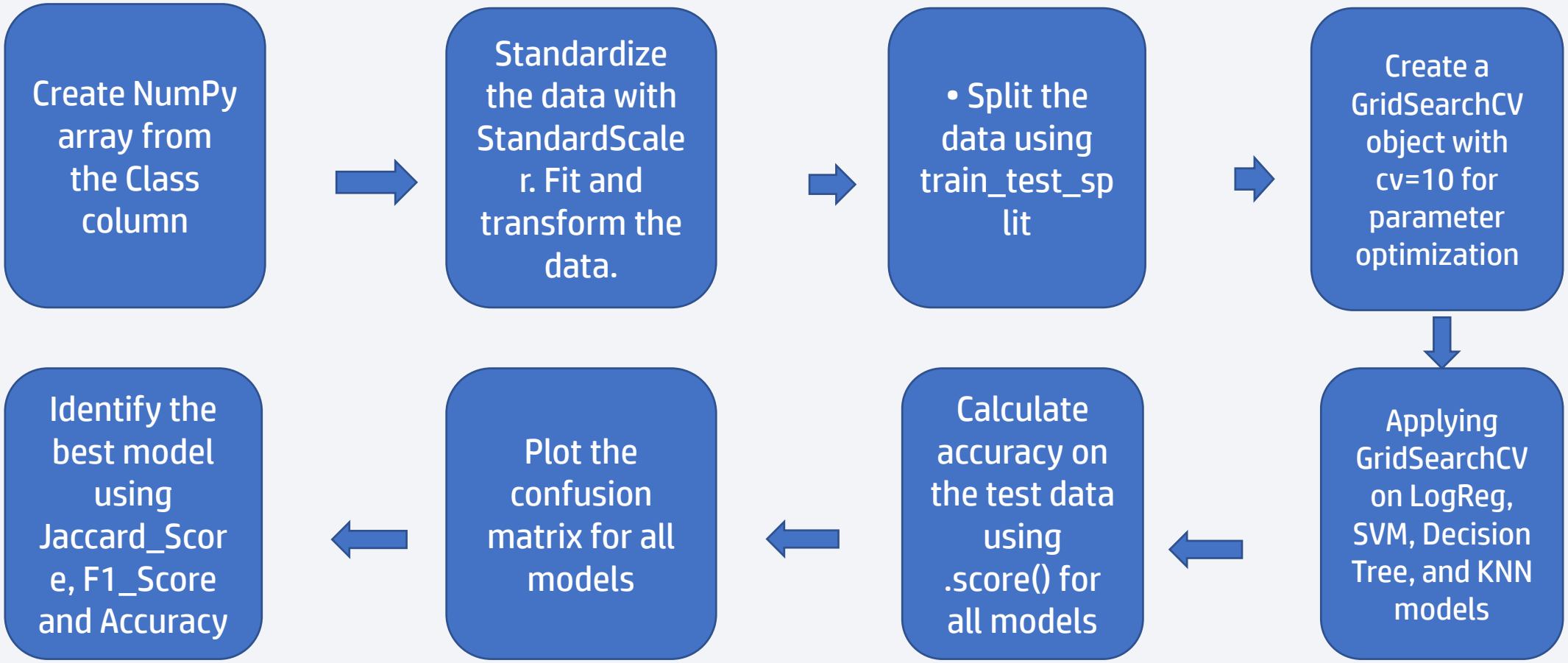
Build an Interactive Map with Folium

- The following are the map objects that were added to the folium map:
- Markers and circles were added for launch sites and for the NASA Johnson Space Center
- Color-labeled marker clusters added to determine which launch sites have comparatively high success rates.
- Colored Lines to show distances between the Launch Site KSC LC-39A and its proximities
- The link to the notebook is :- https://github.com/nj0528/testrep/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- The Launch Sites Dropdown List is utilized to select one or all launch sites for the pie chart.
- The pie chart illustrates the total number of successful launches across all sites and, if a specific launch site is selected, the success vs. failure numbers for that site.
- The correlation between the Launch outcome and the payload mass (kg) for each booster version is illustrated using a scatter plot. In the scatterplot, the Payload Mass Range is selected using a slider.
- The link to the notebook is https://github.com/nj0528/testrep/blob/main/Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb

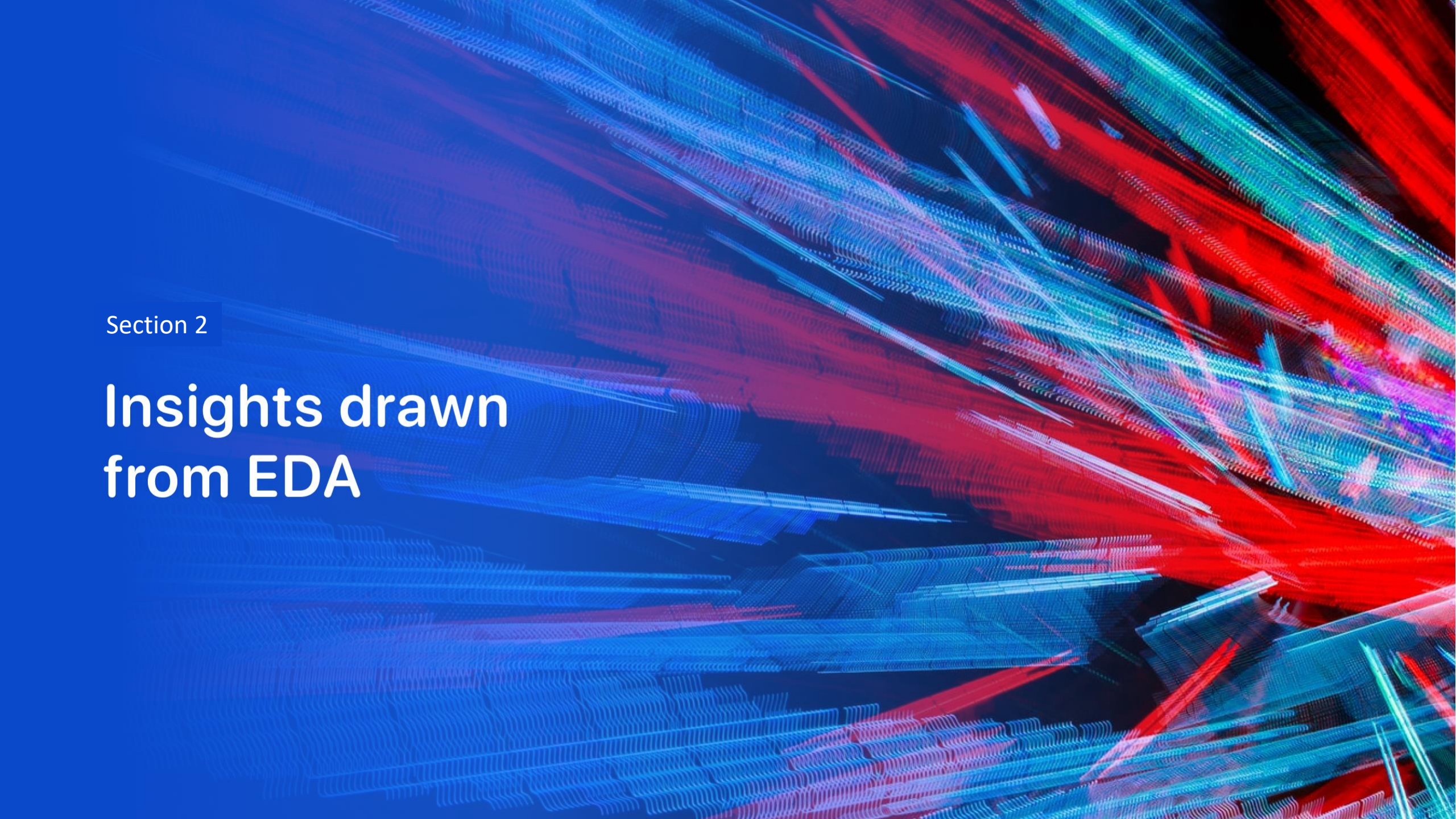
Predictive Analysis (Classification)



- The link to the notebook is https://github.com/nj0528/testrep/blob/main/Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb.ipynb

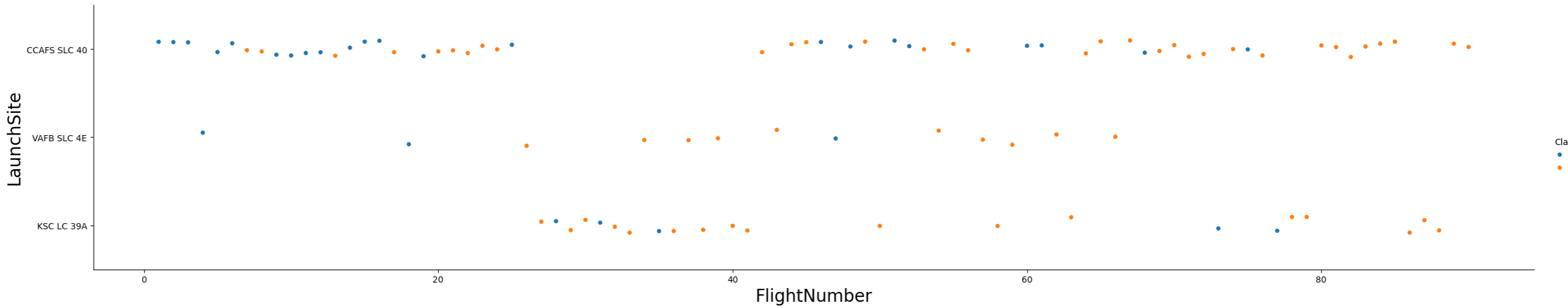
Results

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


Section 2
Insights drawn from EDA

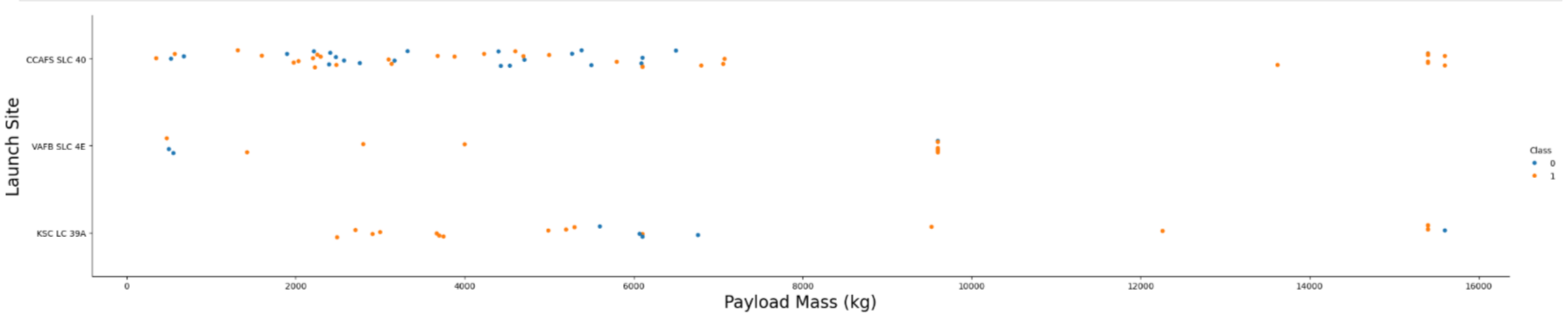
Flight Number vs. Launch Site

- Exploratory Data Analysis
- The success rate of earlier flights was lower (blue = fail).
- The success rate for later flights was higher (orange = success).
- The CCAFS SLC 40 launch site hosted around half of the launches.
- Success rates are greater at KSC LC 39A and VAFB SLC 4E.
- It concludes that the success rate of new launches is higher.



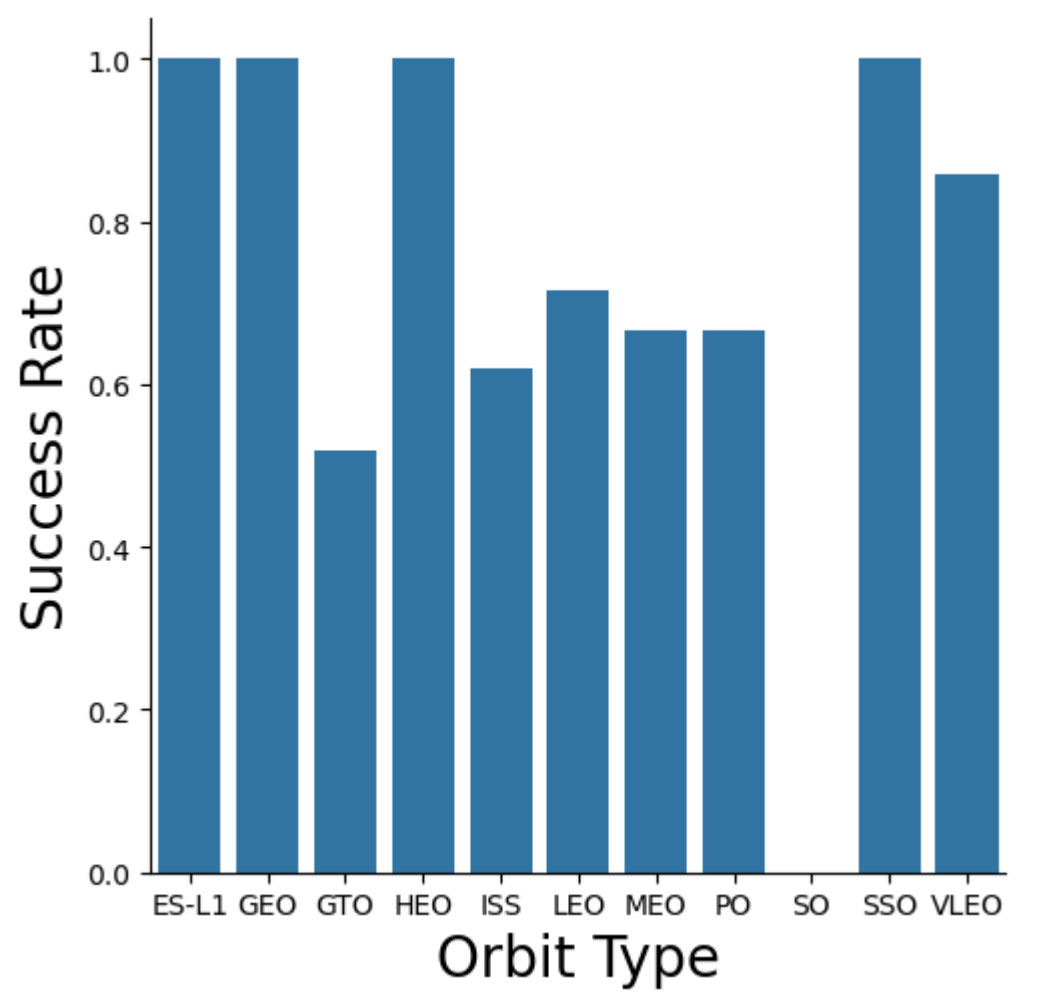
Payload vs. Launch Site

- It can be observed that the success rate predominantly improves as payload mass (kg) increases.
- Almost all launches with payloads more than 7,000 kg were successful.
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than 10,000 kg



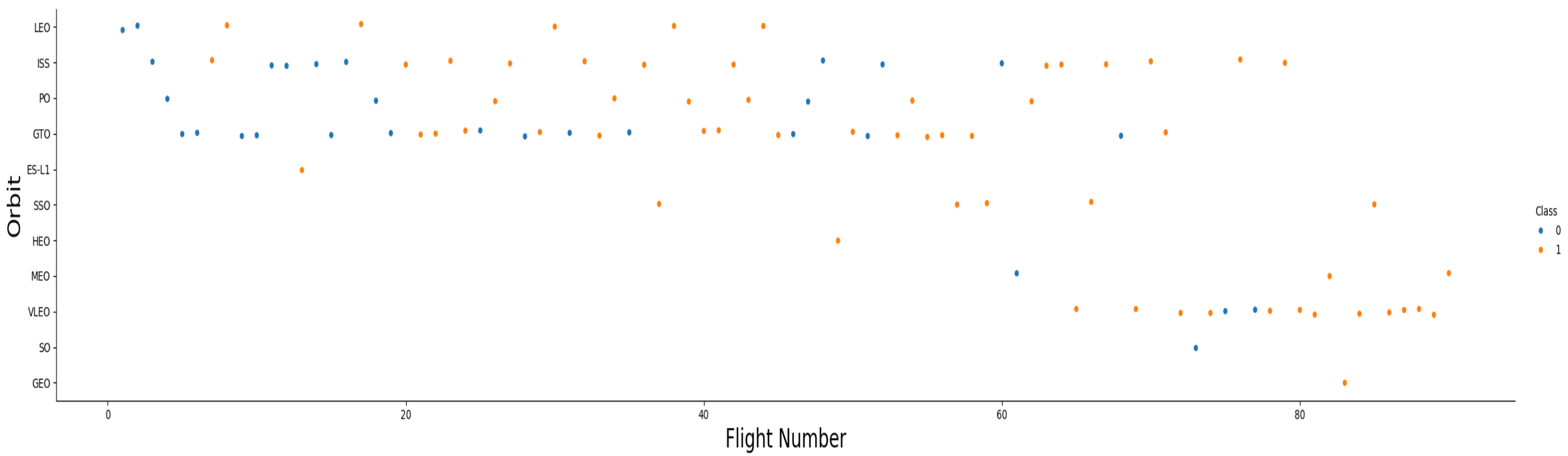
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO are the Orbits with 100% success rate
- SO is the Orbit with 0% success rate
- GTO, ISS, LEO, MEO, PO are the Orbits with success rate between 50% and 80%



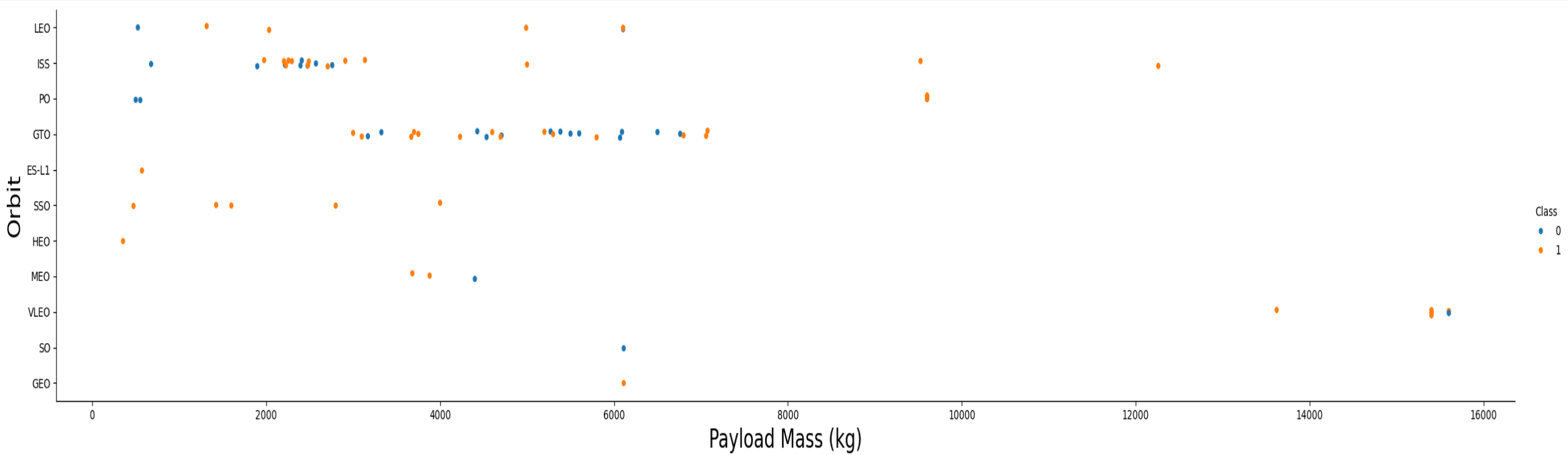
Flight Number vs. Orbit Type

- While in GTO orbit, there doesn't seem to be a relation between the number of flights and success; however, in LEO orbit, success is related to the number of flights.



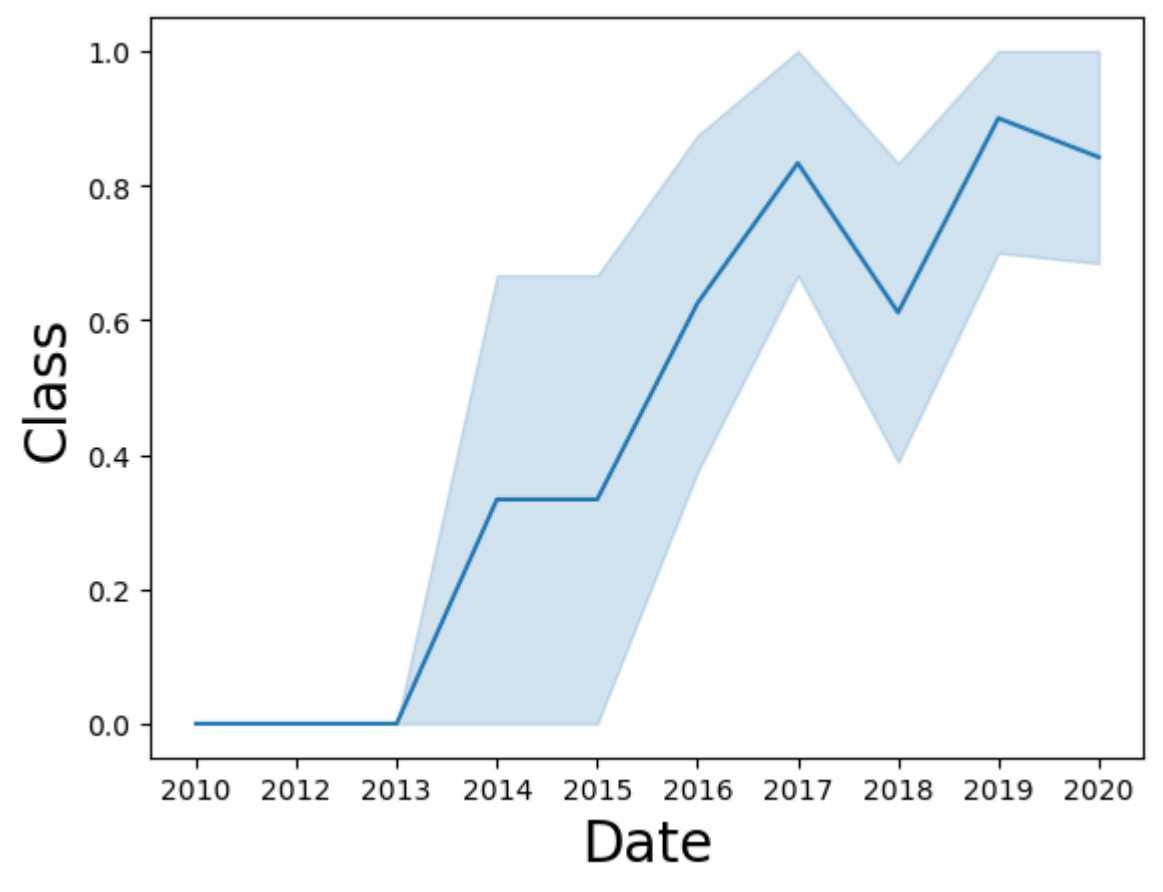
Payload vs. Orbit Type

On GTO orbits, heavy payloads have a negative impact; on GTO and Polar LEO (ISS) orbit, they have a positive impact.



Launch Success Yearly Trend

- The success rate increased between the years 2013–2017 and 2018–2019.
- Between 2017 and 2018 as well as between 2019 and 2022, the success rate fell.
- Since 2013, the success rate has generally increased.



All Launch Site Names

Displaying the names of the unique launch sites for the space project using keyword DISTINCT

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

```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * 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'

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

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

```
%sql select * from SPACEXTBL 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_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

- Displaying the total payload carried by boosters from NASA using the query below

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- 1st Successful Landing in Ground Pad was on 2015-12-22

```
%sql select min(DATE) from SPACEXTBL 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

- The boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
%sql select Booster_Version from SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and F
```

```
 * 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

- The total number of successful and failure mission outcomes are 1 failure during flight, 99 successes, and 1 success (payload status uncertain).

```
%sql select mission_outcome, count(*) as total_number from  SPACEXTBL group by mission_outcome;
```

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

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

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass are listed here

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

```
 * 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

- The failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015 are listed below

```
%sql SELECT SUBSTR(Date,6,2) AS Month, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,0,5) = '2015'
```

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

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

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

- Ranking landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order is listed below

```
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04 ' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Numbers DESC;
```

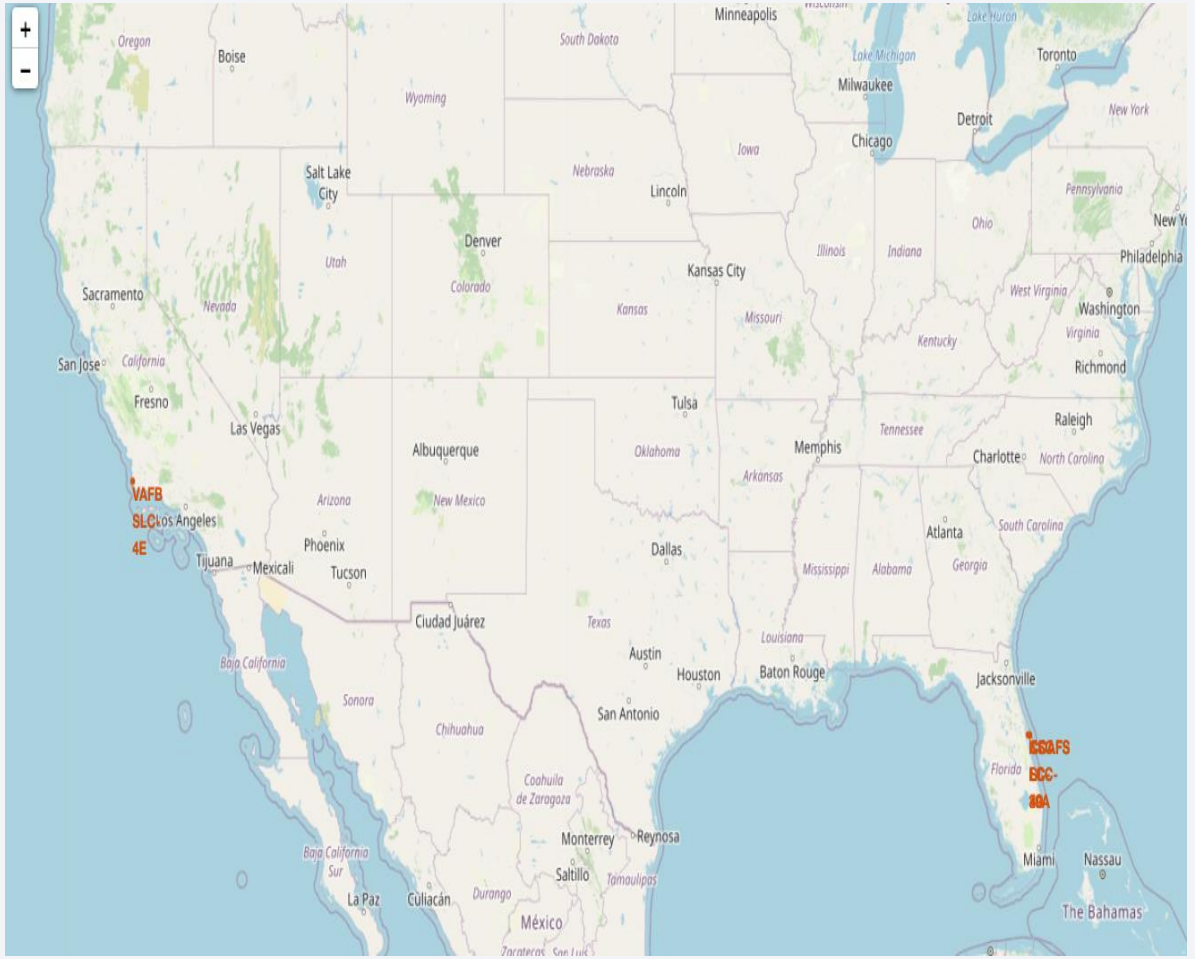
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

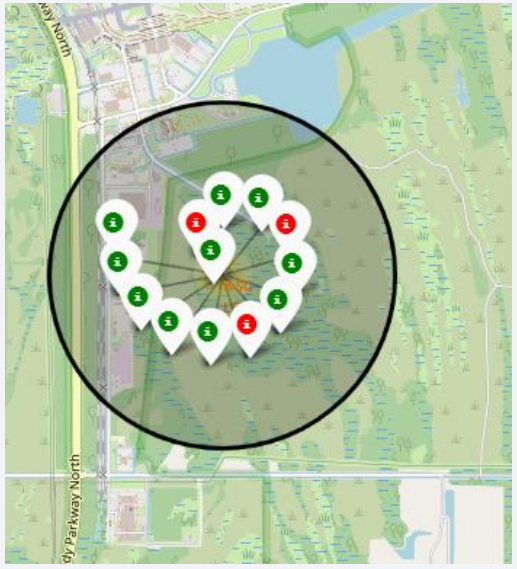
All launch sites global map markers

- Rockets are launched from locations close to the equator for a few different reasons.
- One explanation is that a rocket launched from the east coast receives an extra push from Earth's spinning speed.
- Another reason is that the debris will fall into an ocean rather than a heavily populated area in the event that something goes wrong during the ascent.
- The majority of launch sites are also found close to the equator since the extra natural boost these locations provide helps rockets launch from them consume less fuel and boosters.

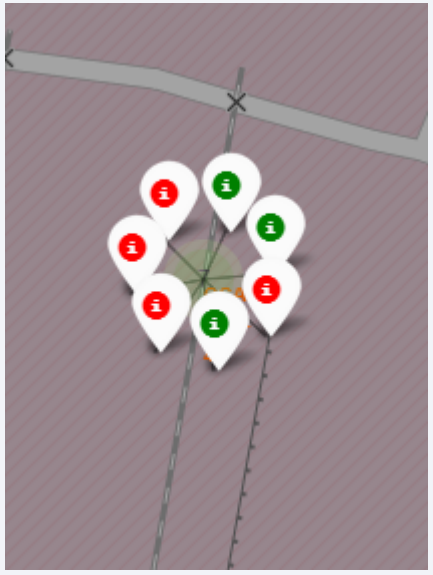


Launch locations identified with colored-label markers.

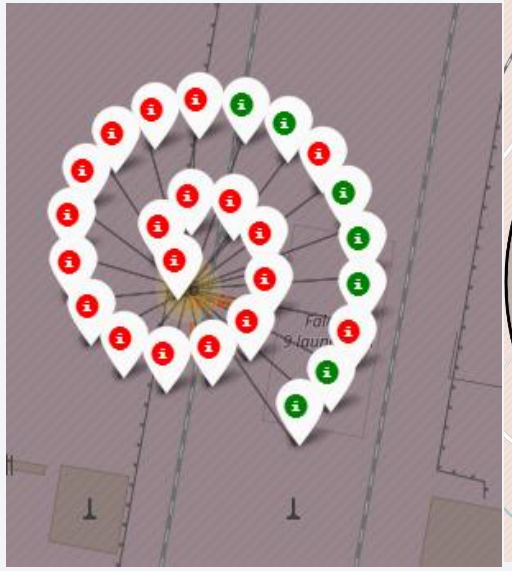
- The colour-labeled markers enable in determining which launch sites have comparatively high success rates
- Green Marker indicates a Successful Launch and Red Marker indicates a Failed Launch.
- KSC LC-39A Launch Site has a higher rate of successful launches.



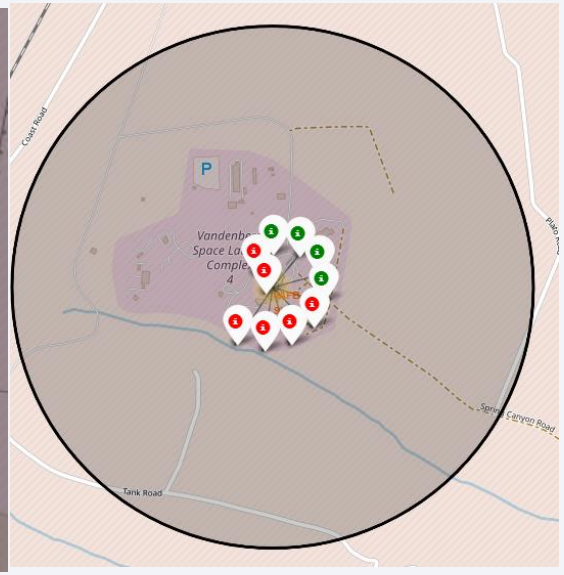
KSC LC-39A



CCAFS SLC-40



CCAFS LC-40



VAFB SLC-4E

Distance between the launch site KSC LC-39A to its proximities

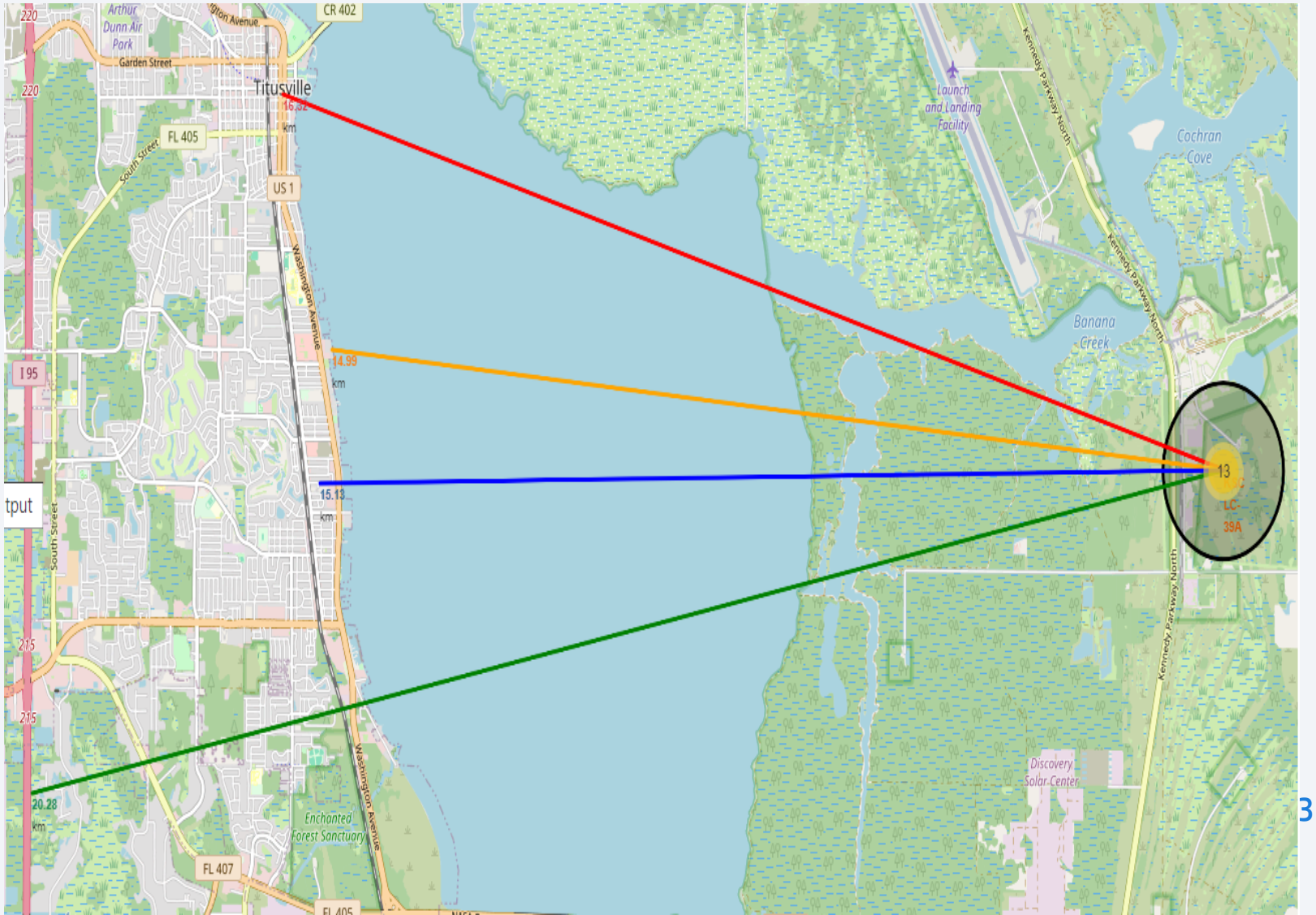
- It can be seen that the launch site KSC LC-39A is at a distance of :

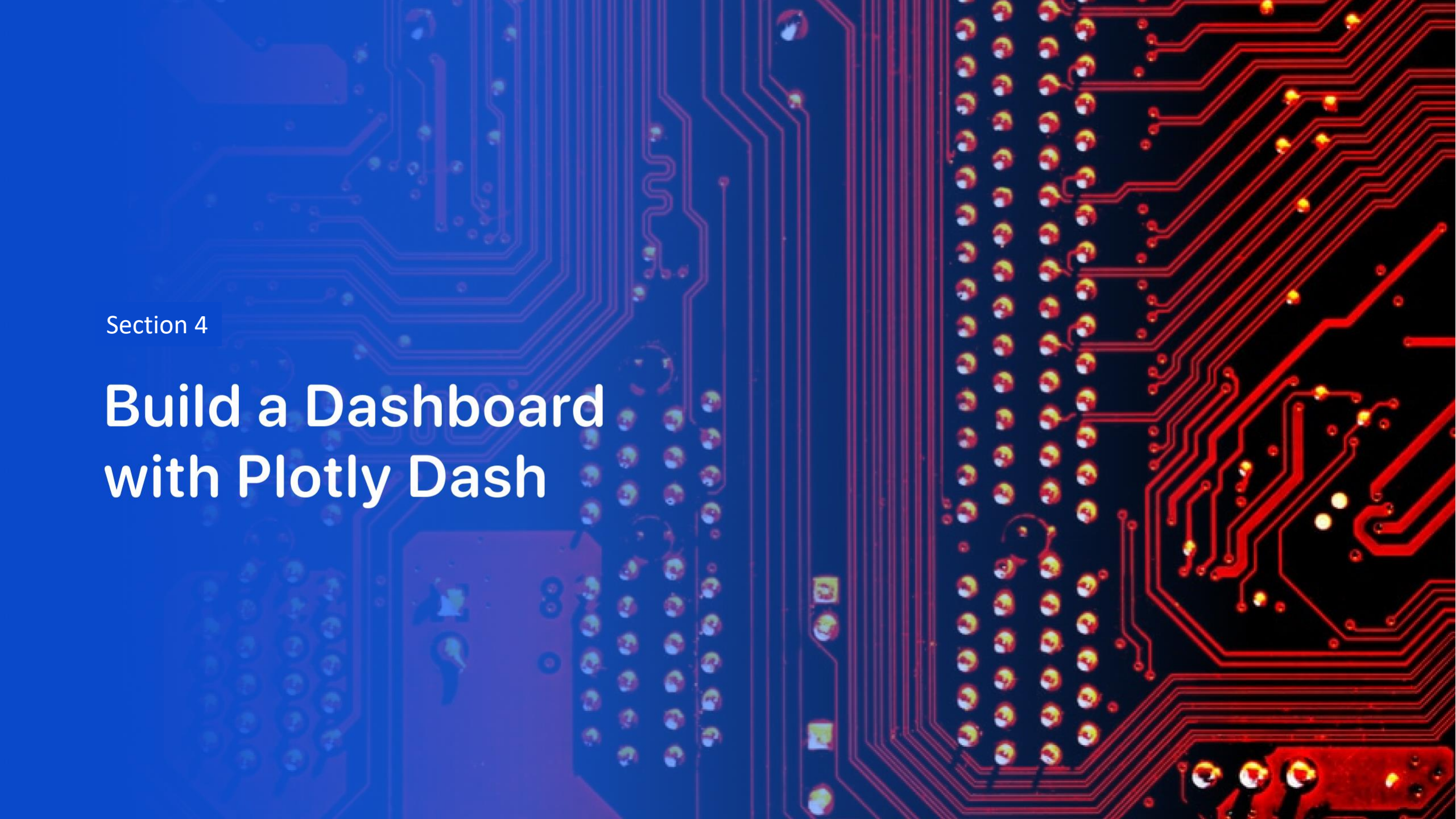
  -15.12 km from railway.

  - 20.28 km from highway.

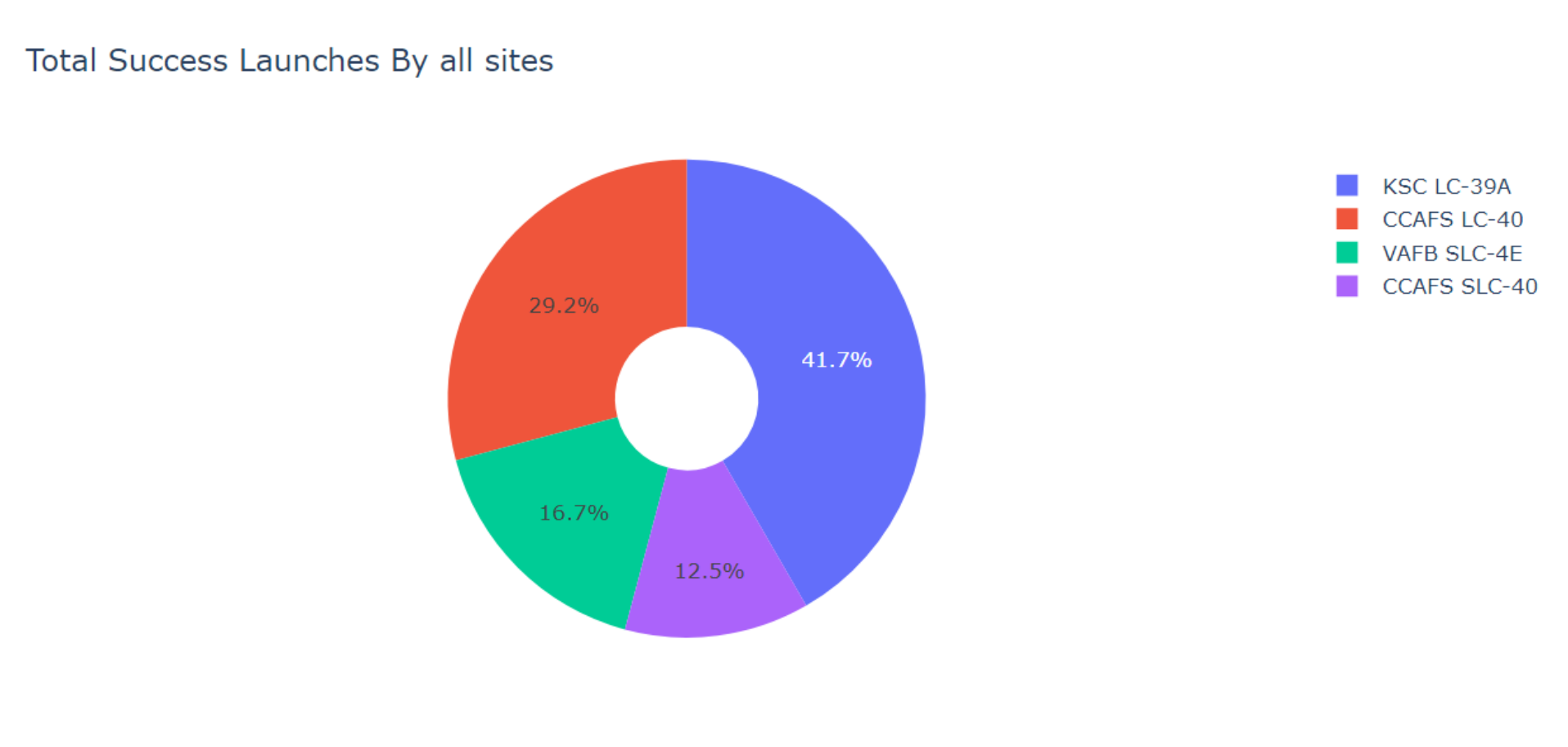
  - 14.99 km from coastline.

  -16.32 km from city Titusville.



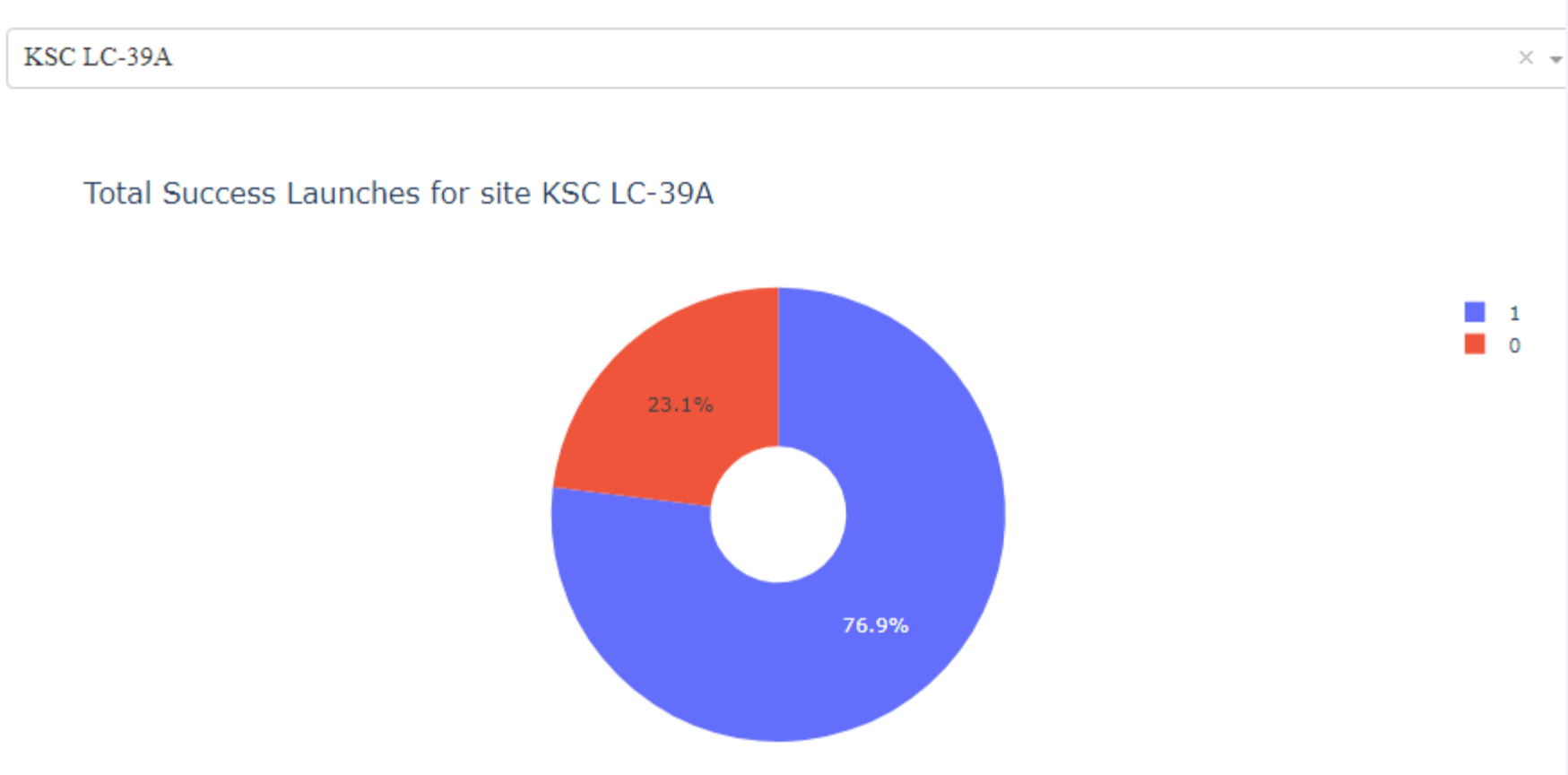

Section 4
Build a Dashboard with Plotly Dash

Pie chart illustrating success percentage of each launch site.



The chart demonstrates that KSC LC-39A has had the most successful launches out of all the locations

Pie chart illustrating the Launch site with highest launch success ratio



- The highest launch success percentage of KSC LC-39A is the highest at 76.9%, with 10 successful and only 3 failed landings

Scatter plot of Payload Mass vs. Launch Outcome for all sites

The payloads with the highest success rate are those weighing between 2,000 and 5,000 kg. A successful outcome is denoted by a 1 and a failed outcome by a 0.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

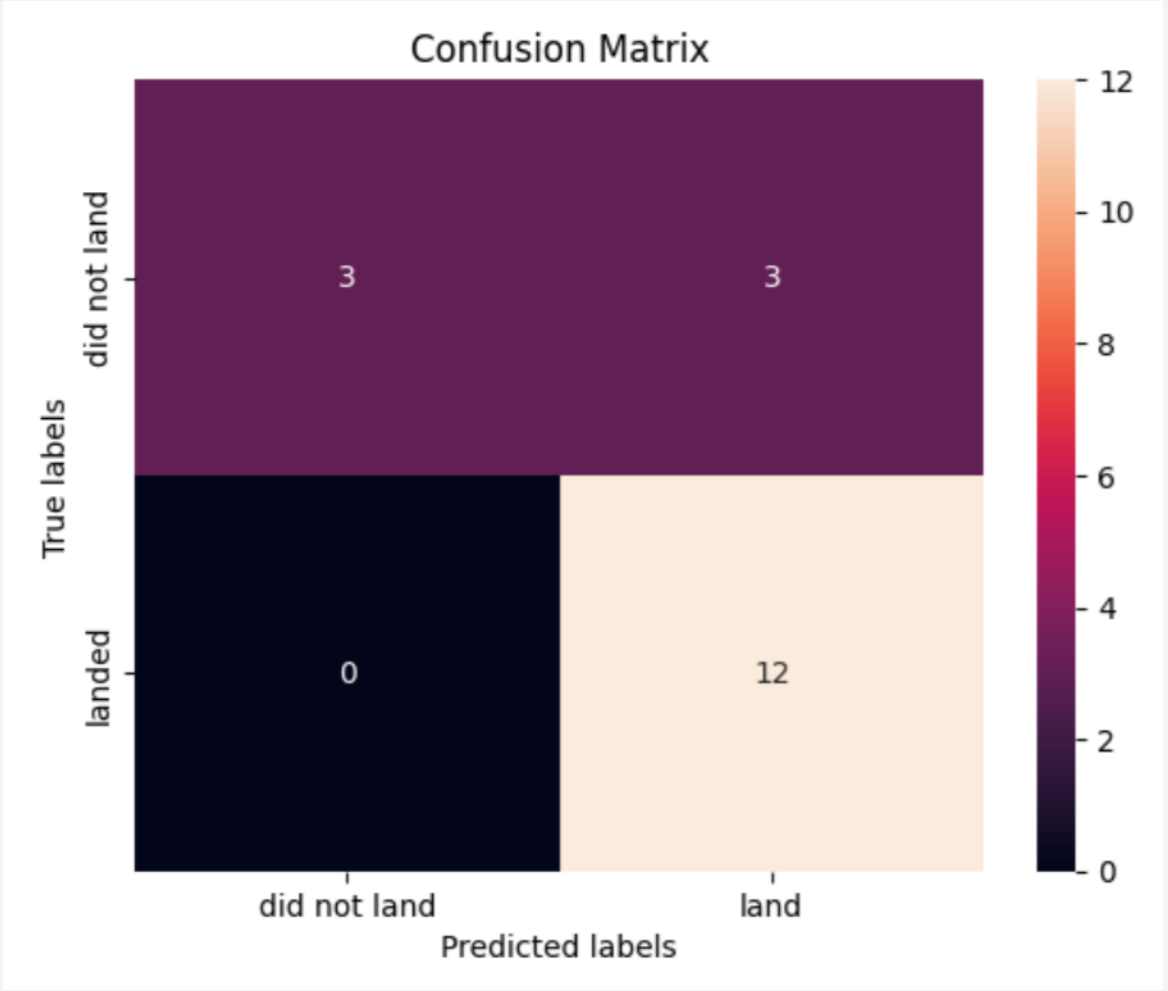
```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.875
Best params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

- Because of the small test sample size, all the models had the same scores and accuracy.
- When analyzing the average of all CV folds for a particular set of parameters, the Decision Tree model appeared slightly better than the others.

Confusion Matrix

- The effectiveness of a classification algorithm is summarized by a confusion matrix.
- The decision tree classifier's confusion matrix demonstrates its ability to differentiate between the various classes.
- The presence of false positives is the primary issue at stake. i.e., the classifier marks an unsuccessful landing as a successful landing.



Conclusions

The following conclusions can be drawn:

- Launches with smaller payload masses perform better than those with bigger payload masses.
- The majority of launch locations are located close to the equator in order to benefit from an extra natural boost caused by the earth's rotation. •
- Over the years, the launch success rate rises.
- Out of all the launch sites, KSC LC-39A has the best success percentage for launches.
- The orbits with the highest success rate were ES-L1, GEO, HEO, and SSO.
- The Decision Tree Model is the most ideal algorithm for this project.

Thank you!