IBM Developer
SKILLS NETWORK

SPACEX

Winning Space Race with Data Science

Data Science Capstone

Javier Guerrero
October 4th, 2023

Outline

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

Executive Summary

- The prediction of successfully landing of the SpaceX Falcon 9 using some machine learning classification algorithms.
- The steps top follow in this project has been:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- We can resume it on two main results:
  - At the visualization we can conclude that there are a correlation between the rocket features and the outcome launches in the success or failure of the landing.
  - The decision tree has been observe like the most accurate machine algorithm at the predictions.

Introduction

- The objective of this capstone, is to predict the Falcon 9 first stage will land successfully.
- 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, the main advantage in the const is because SpaceX is unique at reuse the first stage rocket. In line with this the prediction of success of recovery will let us determine the cost of a launch.
- Or in order to be more efficient, it can improve in the decision of the parameters at the launch searching for better landing success.
- The searching answer is for a given features on a Falcon 9 rocket launch witch will be the result of the first stage rocket landing.

Section 1

Methodology

Methodology

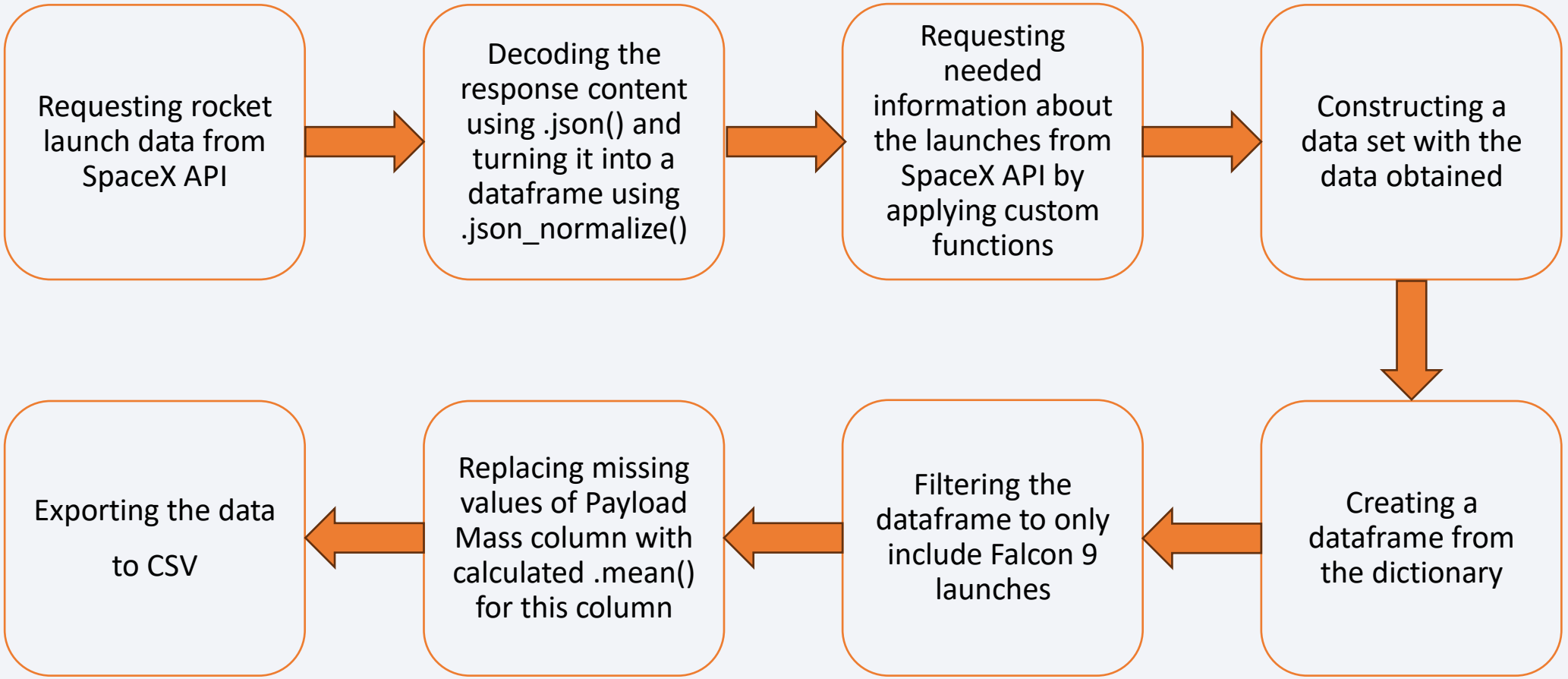
Executive Summary

- Data collection methodology:
  - Web scraping and parsing from the SpaceX API and Wikipedia. into pandas data frames.
- Perform data wrangling
  - Filtering the data, dealing with missing values and classifying 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
  - Create class column, standardize the data and split between test and training.

Data Collection

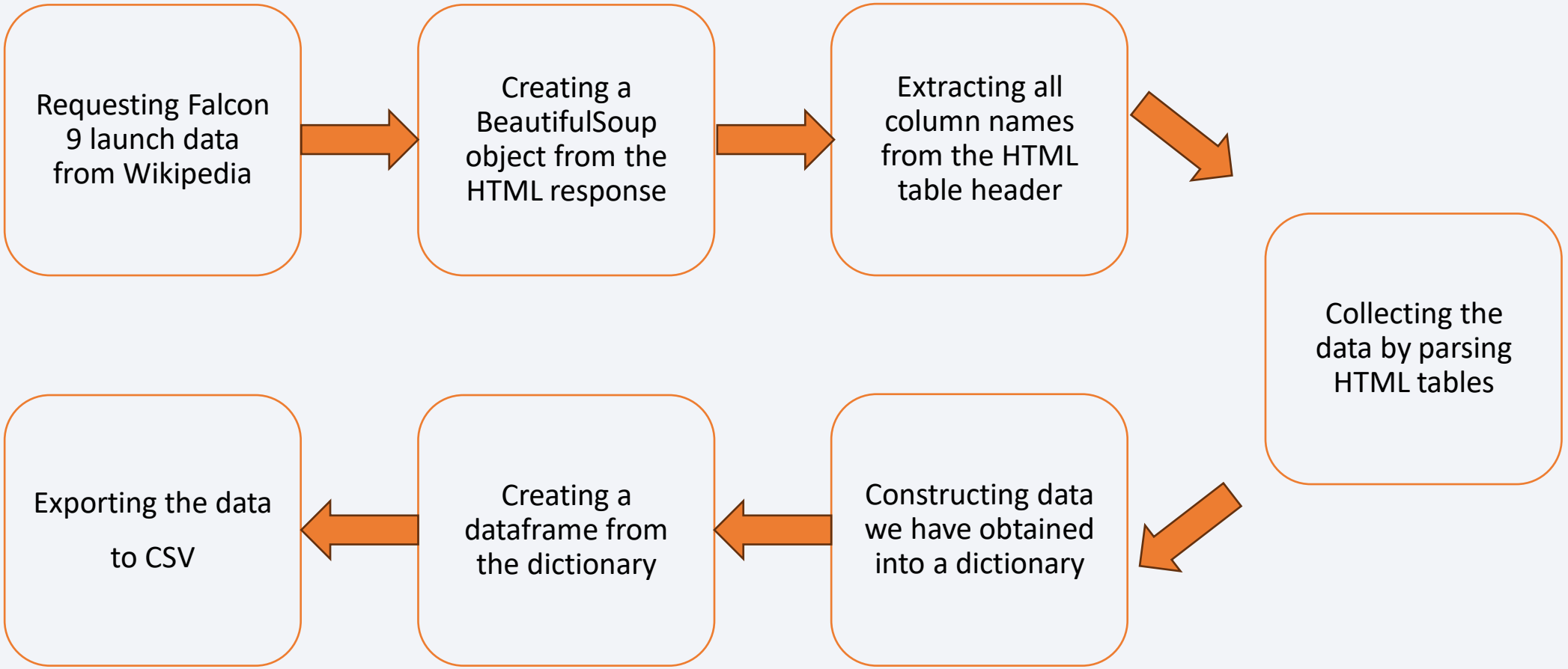
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We need both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- Data Columns are obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, **Flights**, **GridFins**, **Reused, Legs, LandingPad**, **Block**, **ReusedCount, Serial**, **Longitude**, **Latitude**
- Data Columns are obtained by using Wikipedia Web Scraping:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, **Customer**, Launch outcome, Version Booster, **Booster landing**, Date, **Time**

Data Collection – SpaceX API



[Data Collection API URL]

Data Collection – Scraping

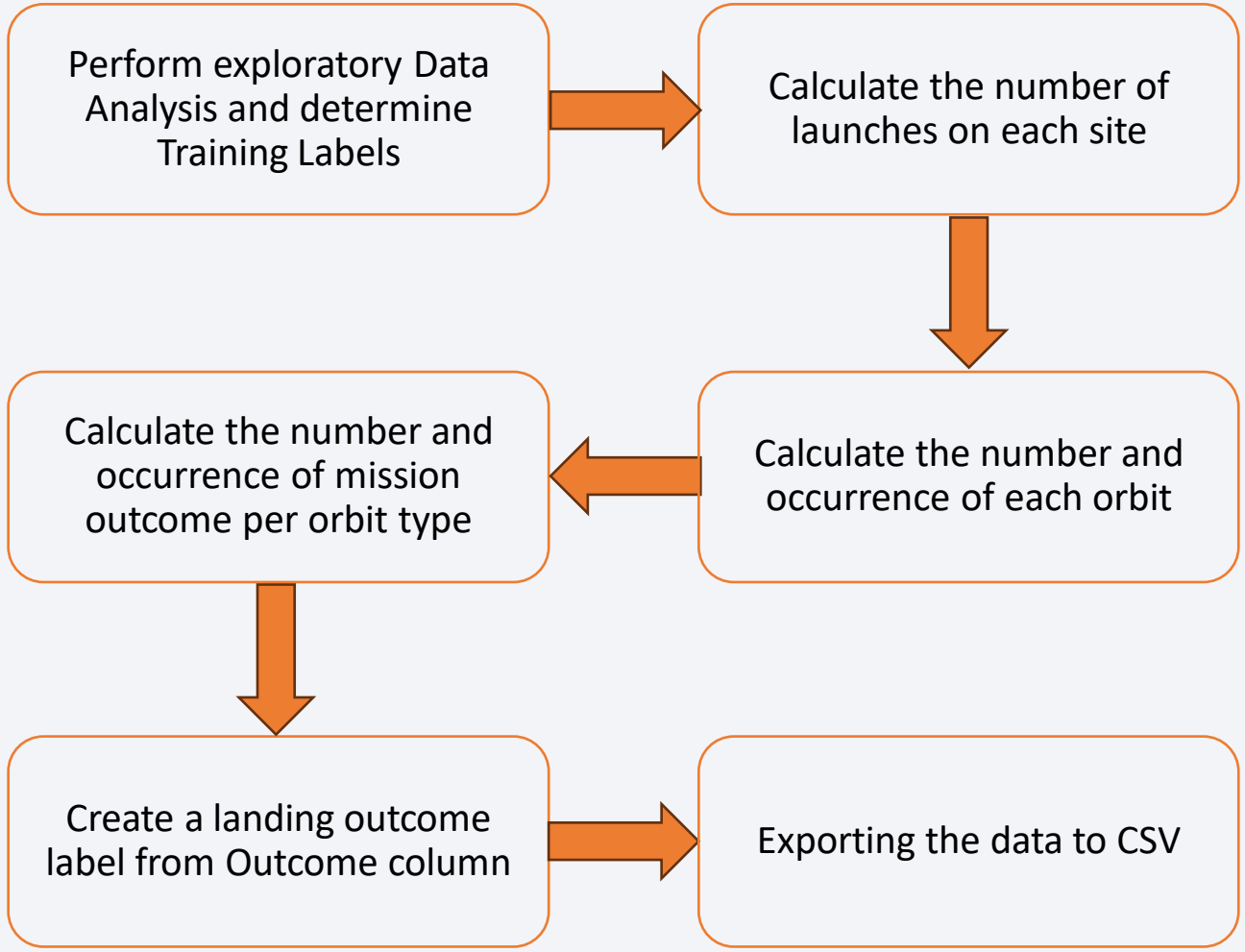


Data Collection Scraping URL

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident represented at outcome.

In order to simplify it takes converted those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



[Data Wrangling URL]

EDA with Data Visualization

- Scatter chart:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
  - A scatter plot shows how much one variable is affected by another. The relationship between two variables is called a correlation. This plot is generally composed of large data bodies.
- Bar chart:
  - Orbit Type vs. Success Rate
  - A Bar chart makes it easy to compare datasets between multiple groups at a glance. One axis represents a category and the other axis represents a discrete value. The purpose of this chart is to indicate the relationship between the two axes.
- Line chart:
  - Year vs. Success Rate
  - A Line chart shows data variables and trends very clearly and helps predict the results of data that has not yet been recorded.

EDA Data Visualization URL

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

EDA SQL URL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- ✓ Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- ✓ Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- ✓ Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- ✓ Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Folium MAP URL

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

✓ Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

✓ Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

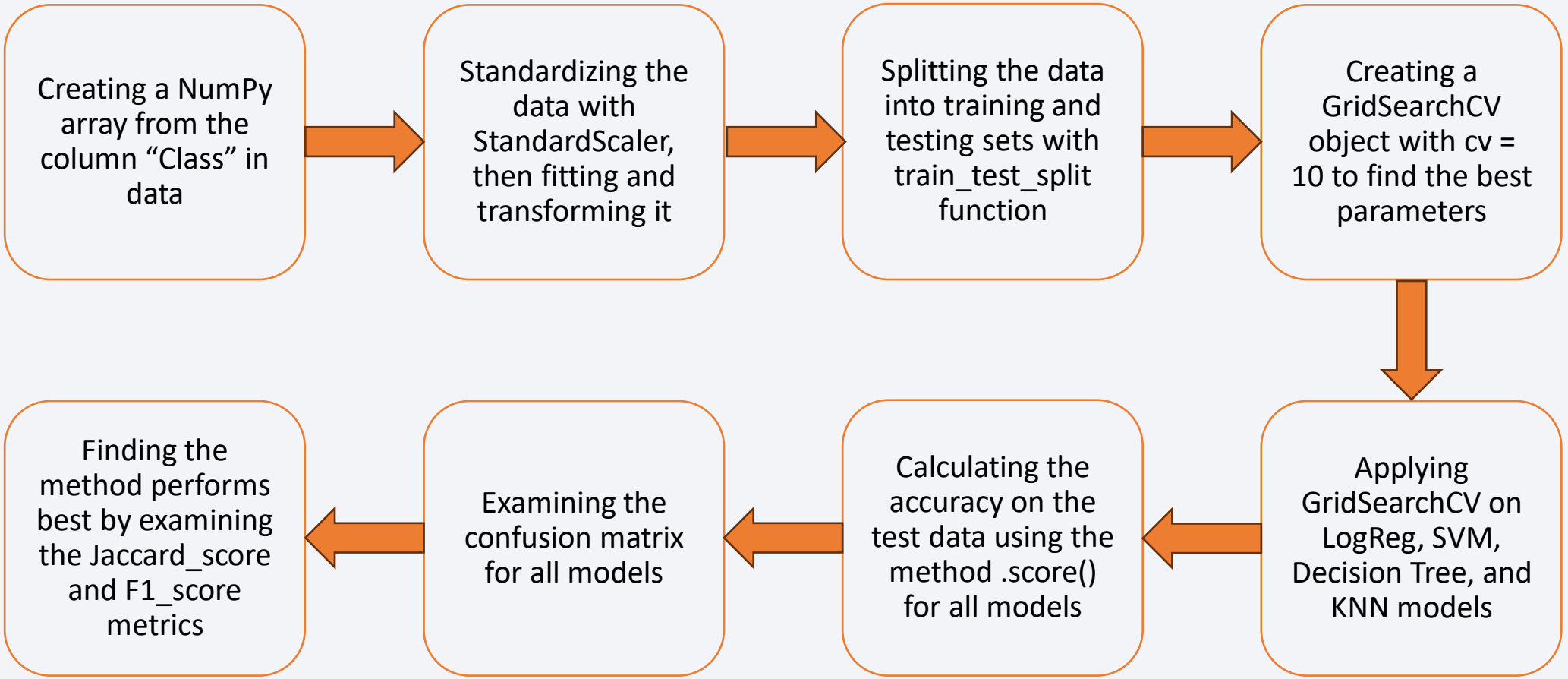
✓ Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

✓ Added a scatter chart to show the correlation between Payload and Launch Success.

Plotly Dashboard URL (code only)

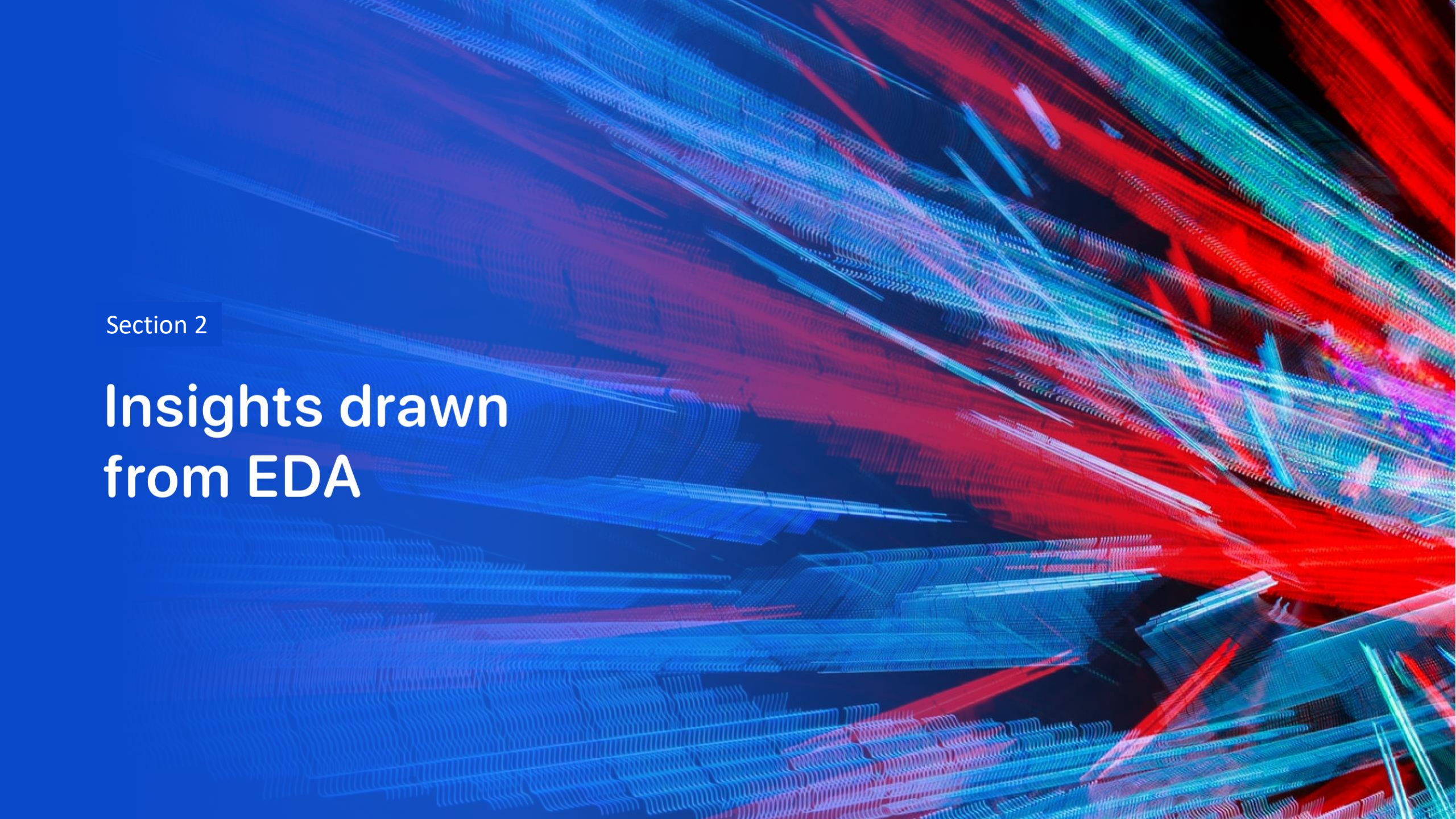
Predictive Analysis (Classification)



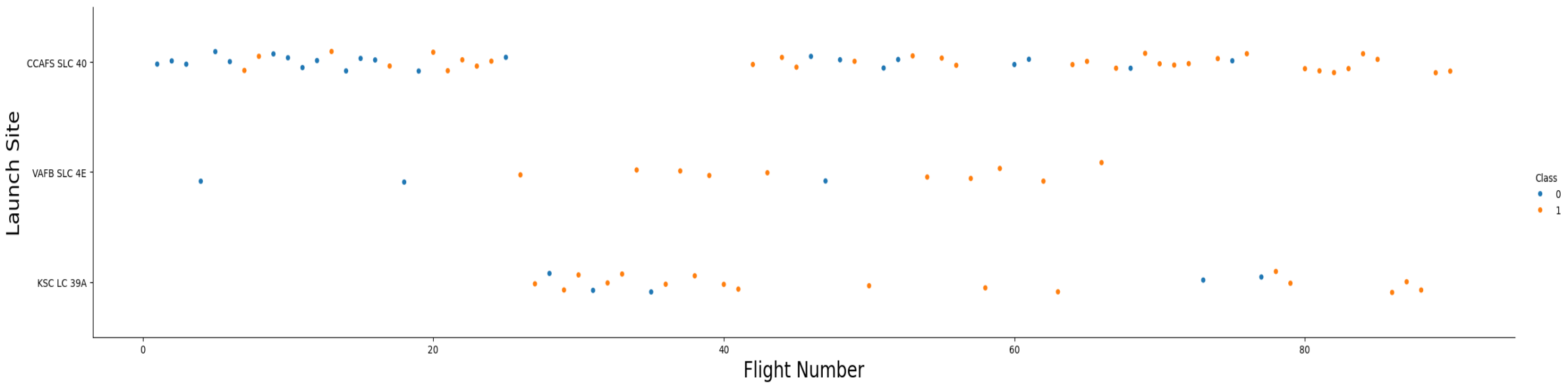
[Machine Learning prediction URL]

Results

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


Section 2
Insights drawn from EDA

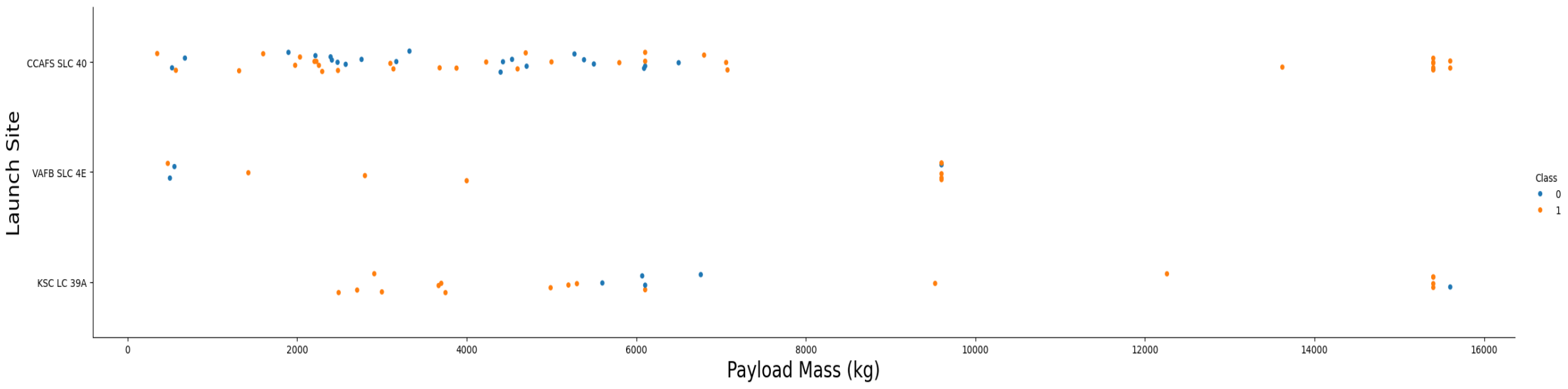
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

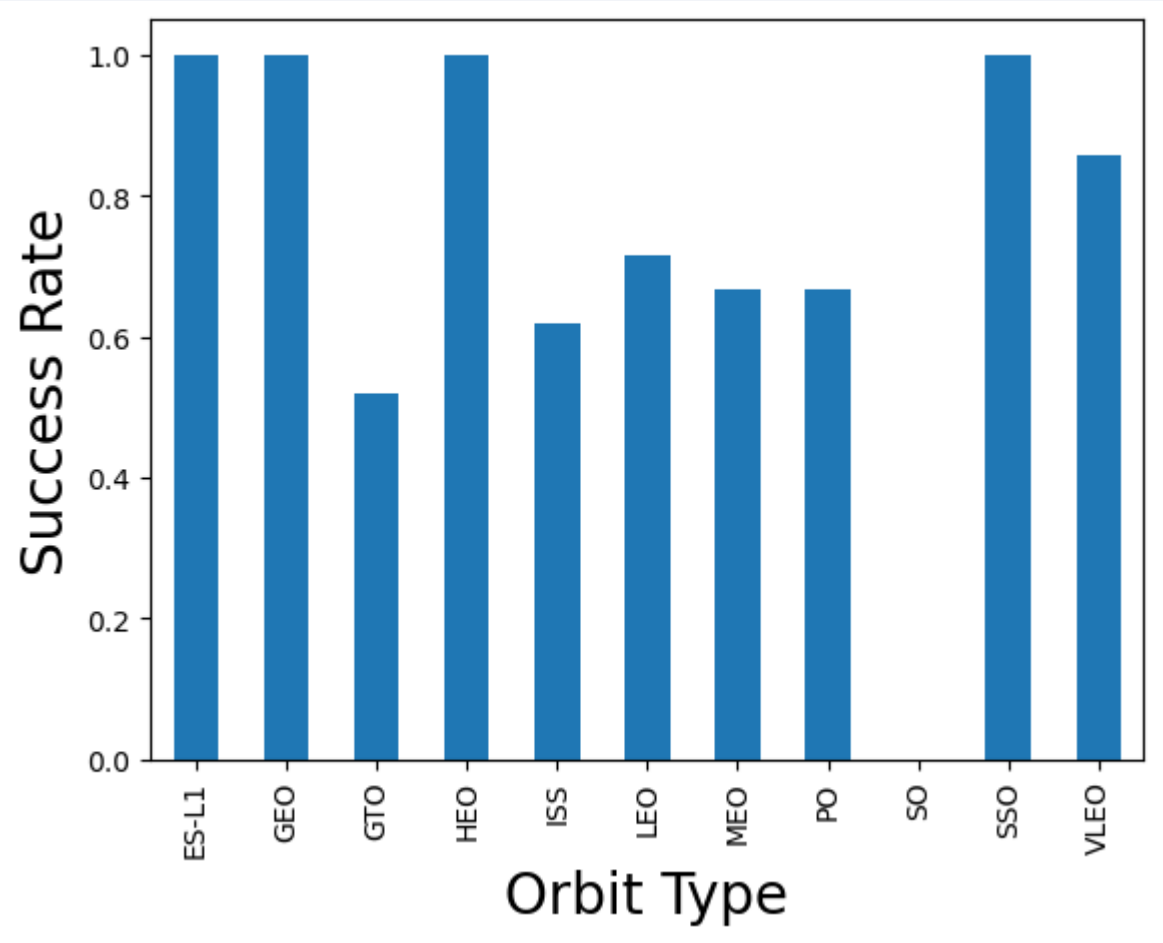
Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



Explanation:

Orbits with 100% success rate:

- ES-L1, GEO, HEO, SSO

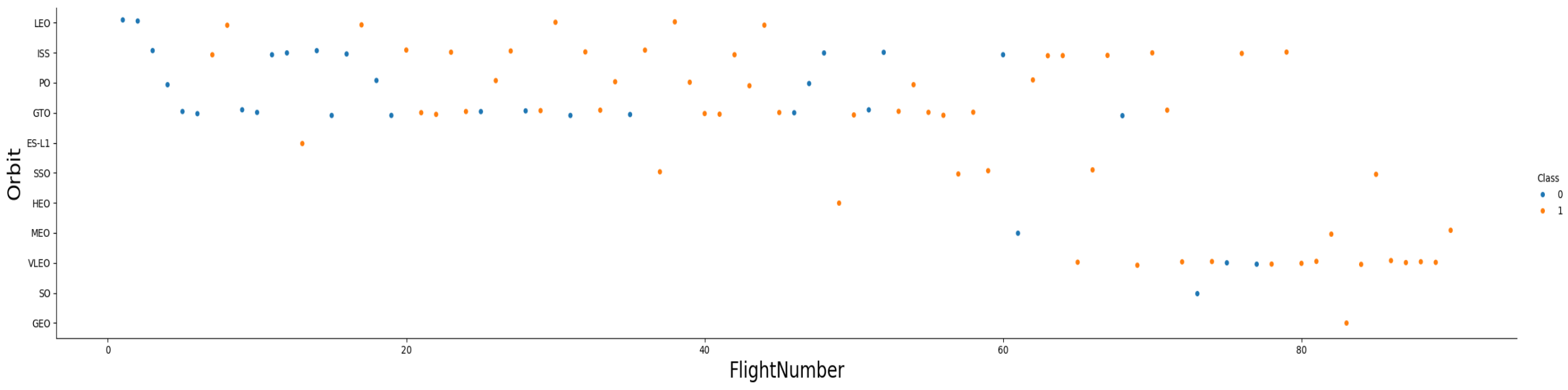
Orbits with 0% success rate:

- SO

- Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO

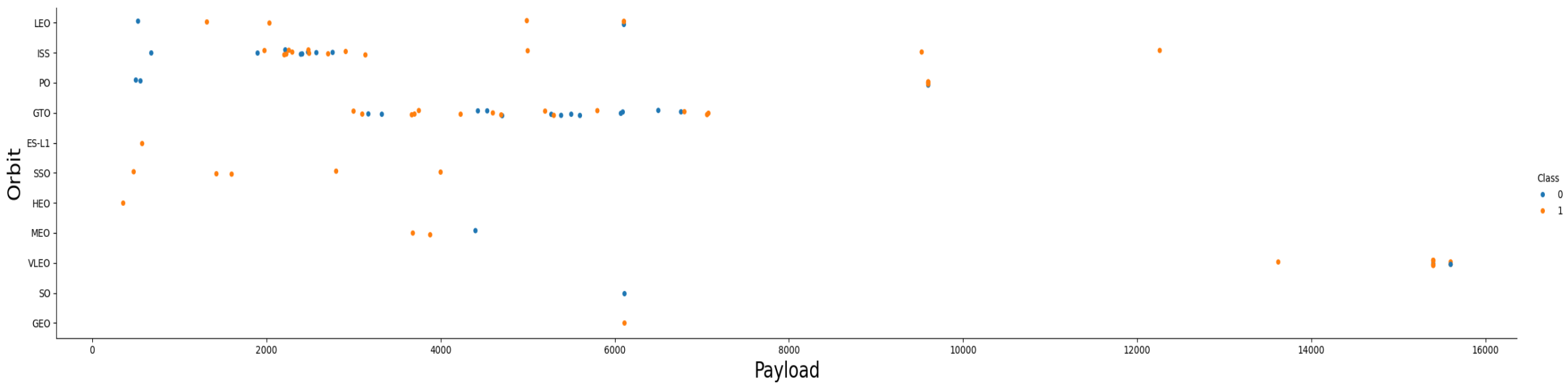
Flight Number vs. Orbit Type



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

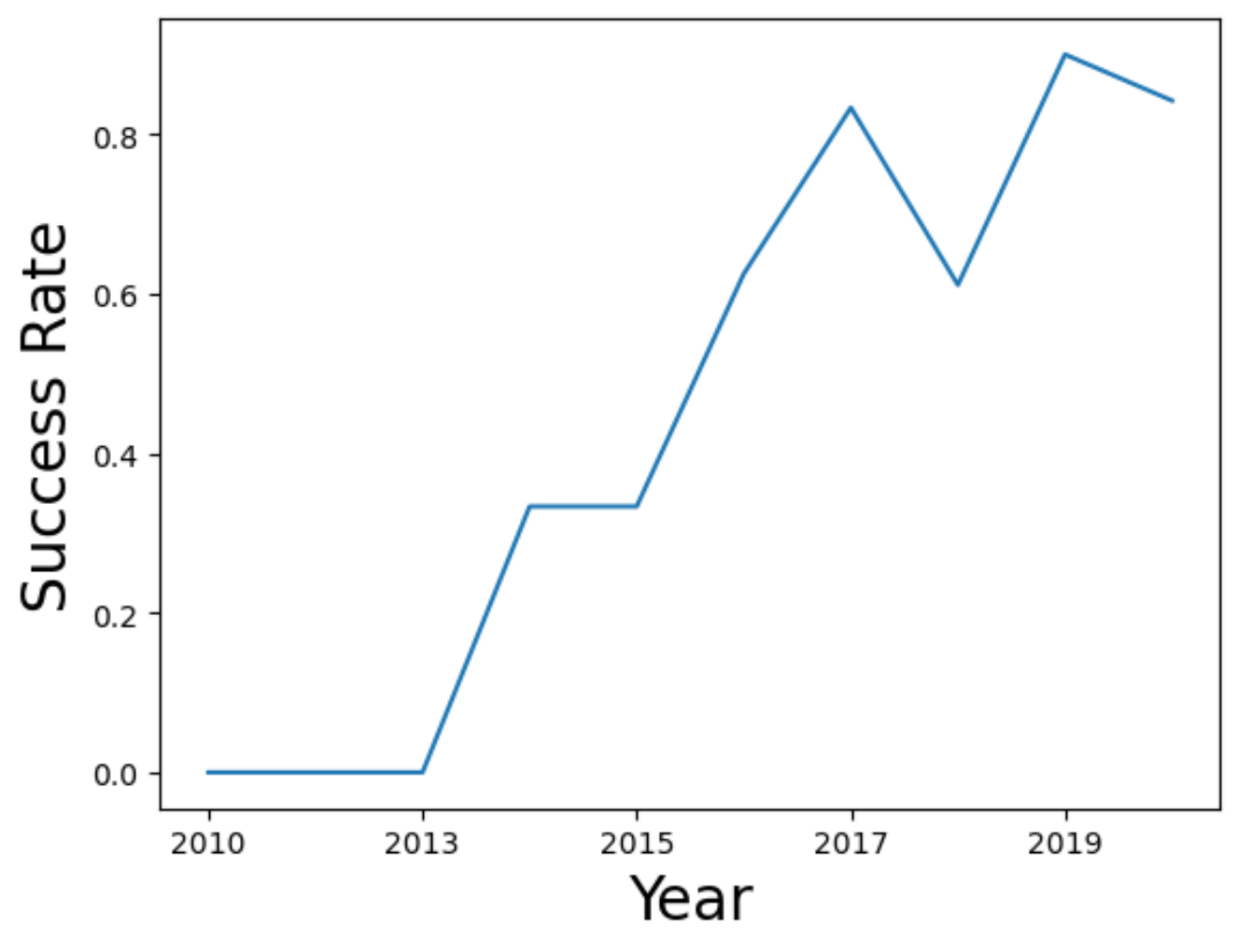
Payload vs. Orbit Type



Explanation:

In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Launch Success Yearly Trend



Explanation:

Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

All Launch Site Names

Launch_Site:

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

Explanation:

Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

Explanation:

Displaying 5 records where launch sites begin with the string 'CCA'.

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

```
 * 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

Explanation:

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

```
%%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

Explanation:

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE '%F9 v1.1%';
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

Explanation:

Listing the date when the first successful landing outcome in ground pad was achieved.

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing_Outcome like ('Success (ground pad)')
;
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

Listing the names of 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 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

Explanation:

Listing the total number of successful and failure mission outcomes.

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) 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

Explanation:

Listing the names of the booster versions which have carried the maximum payload mass.

```
%%sql
SELECT DISTINCT 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

Explanation:

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

```
%%sql
SELECT LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, st.DATE
FROM SPACEXTBL st
WHERE Landing_Outcome = 'Failure (drone ship)'
    AND st.DATE like ('%2015%')
    ;
```

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

Landing_Outcome	Booster_Version	Launch_Site	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-10-01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

Explanation:

Ranking 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.

```
%%sql
SELECT
    LANDING_OUTCOME,
    COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL st
WHERE st.Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

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

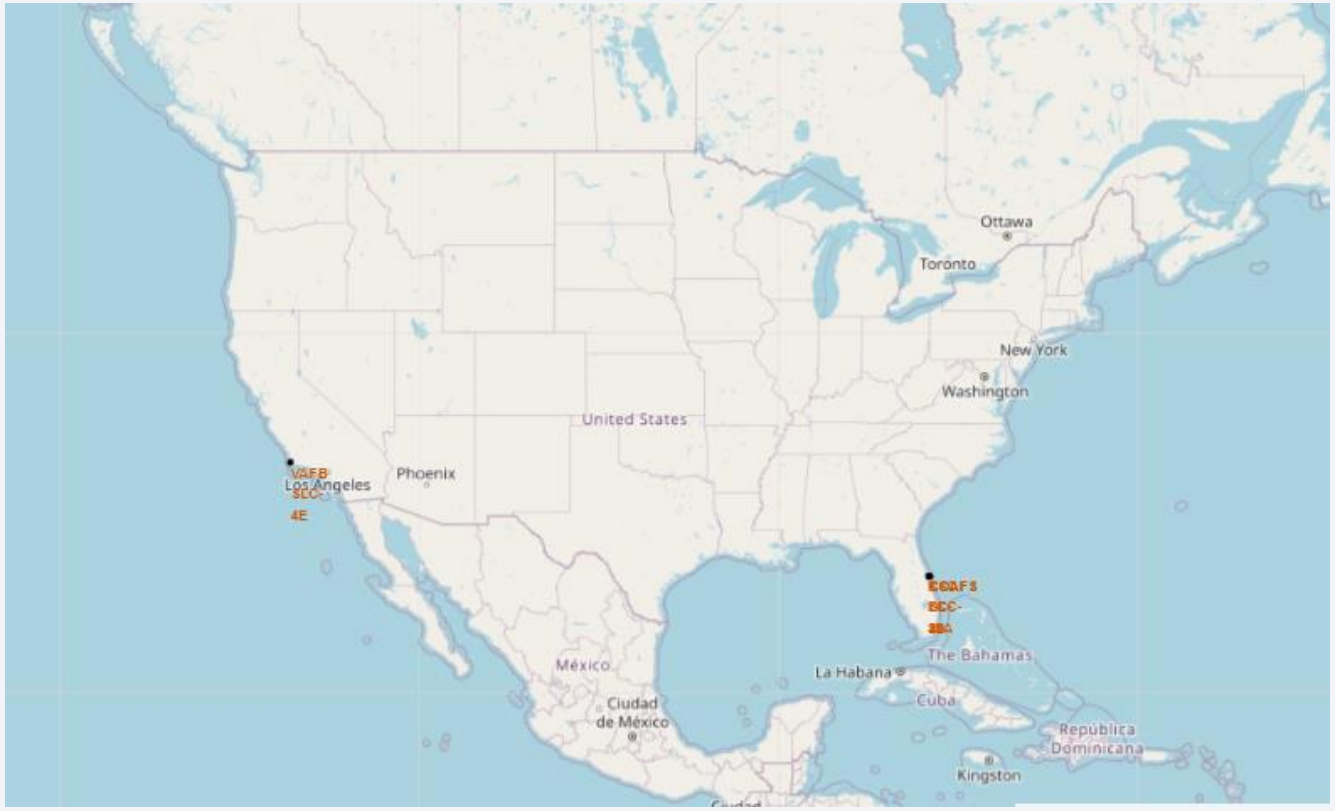
Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

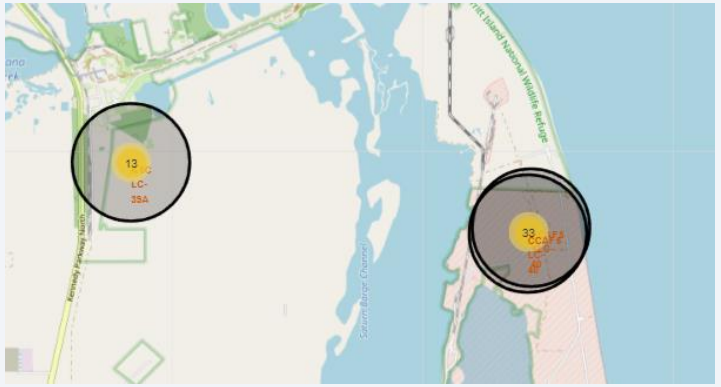
All launch sites' location markers on a global map

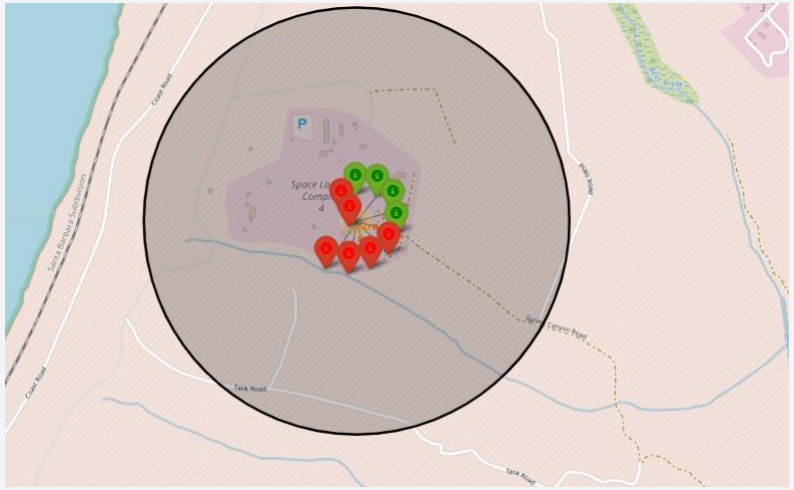
Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

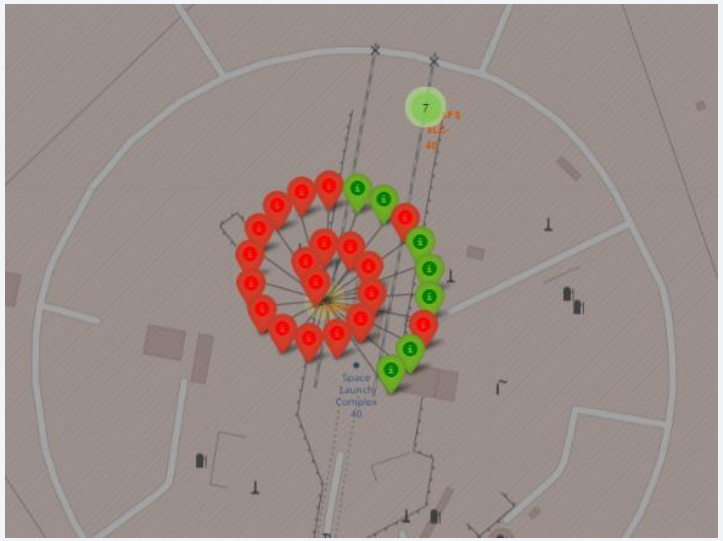
All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Colour-labeled launch records on the map







By clicking on the marker clusters, successful landing (green) or failed landing (red) are displayed.

Proximities of Launch Sites

From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

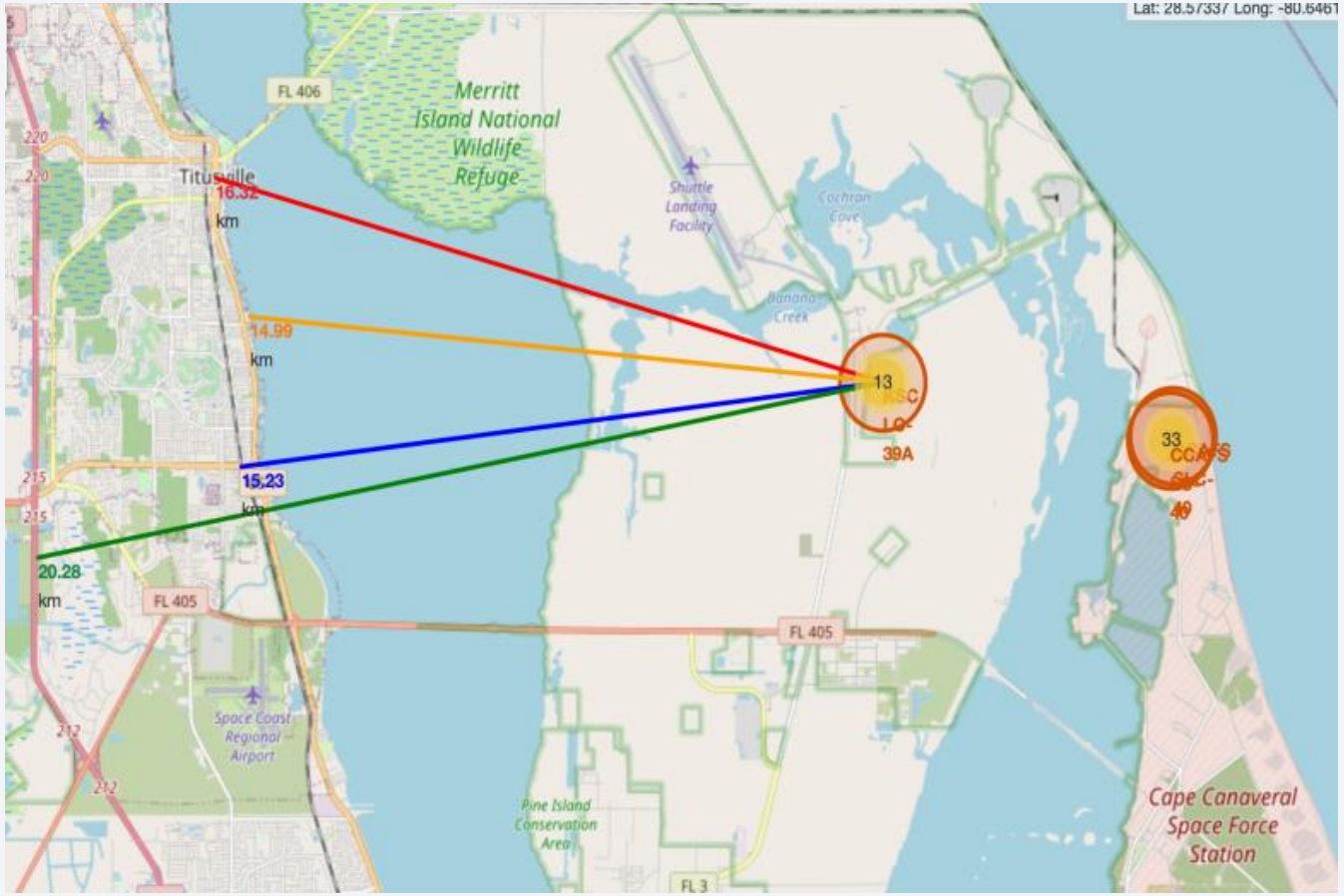
- relative close to railway (15.23 km)

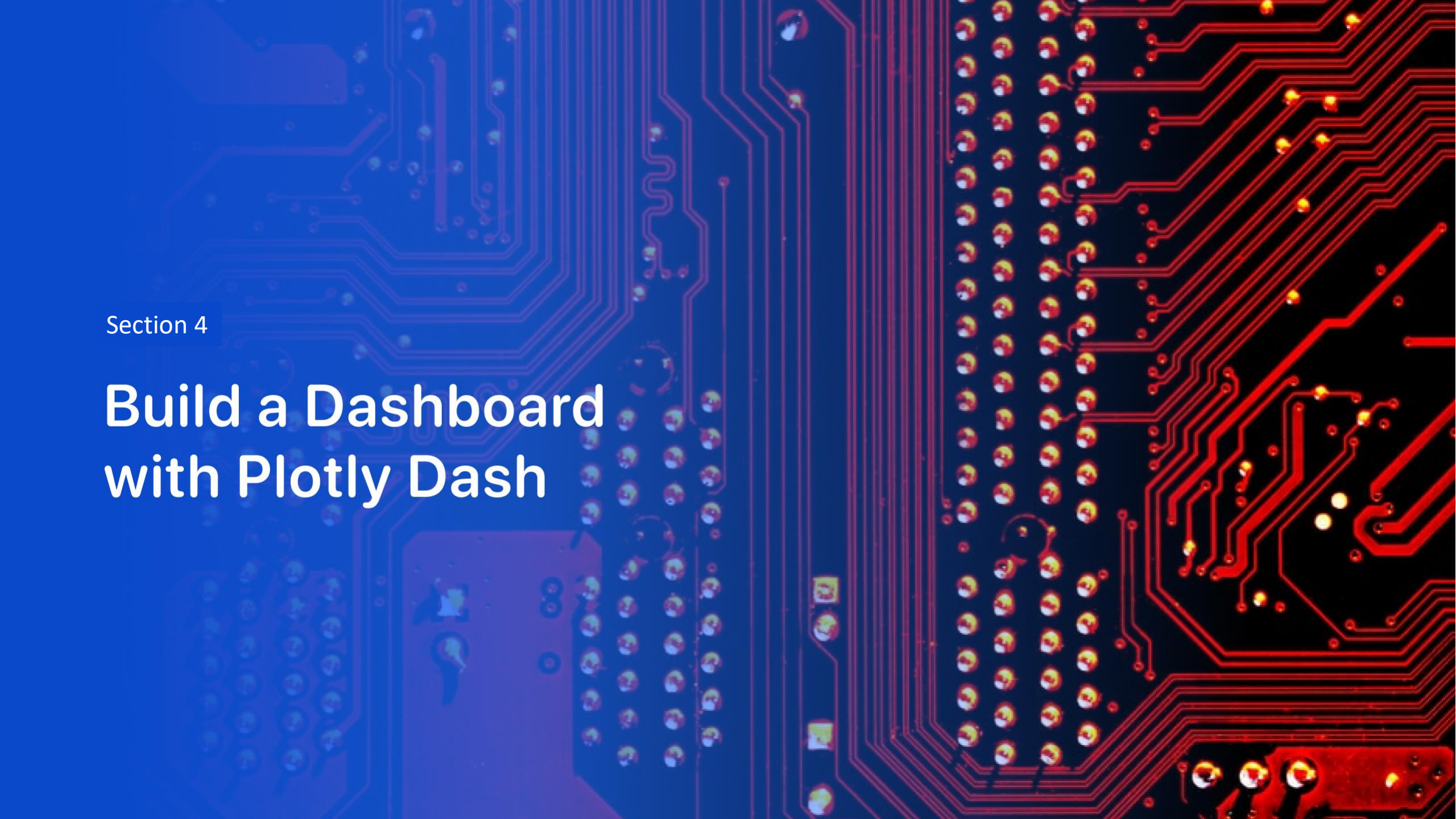
- relative close to highway (20.28 km)

- relative close to coastline (14.99 km)

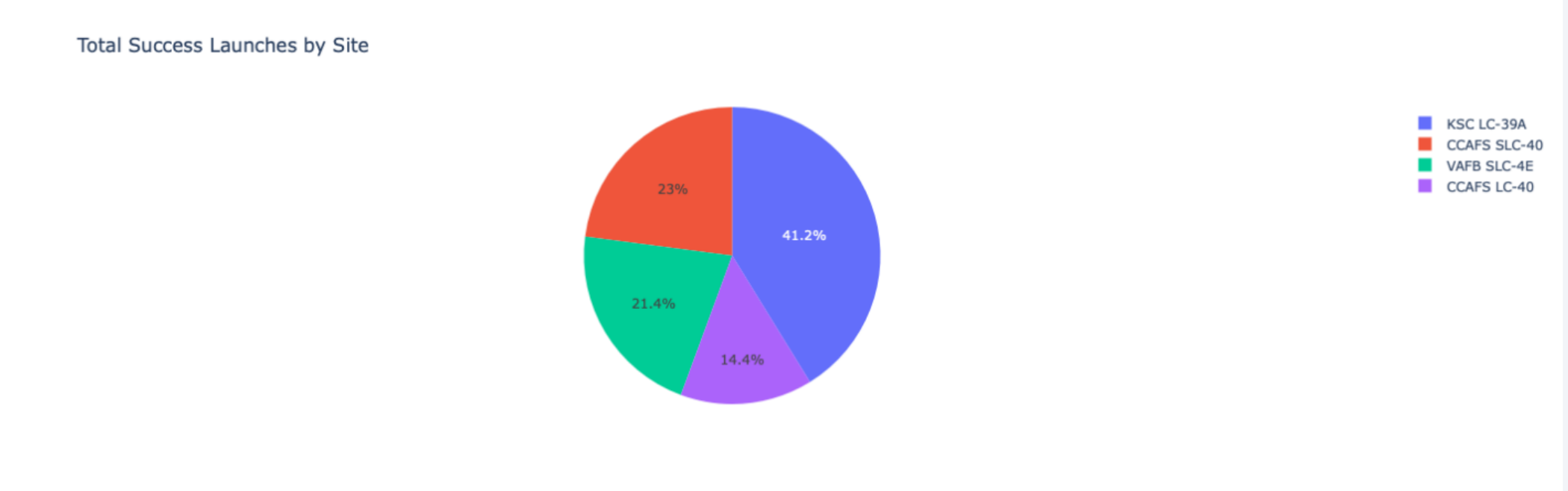
Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas



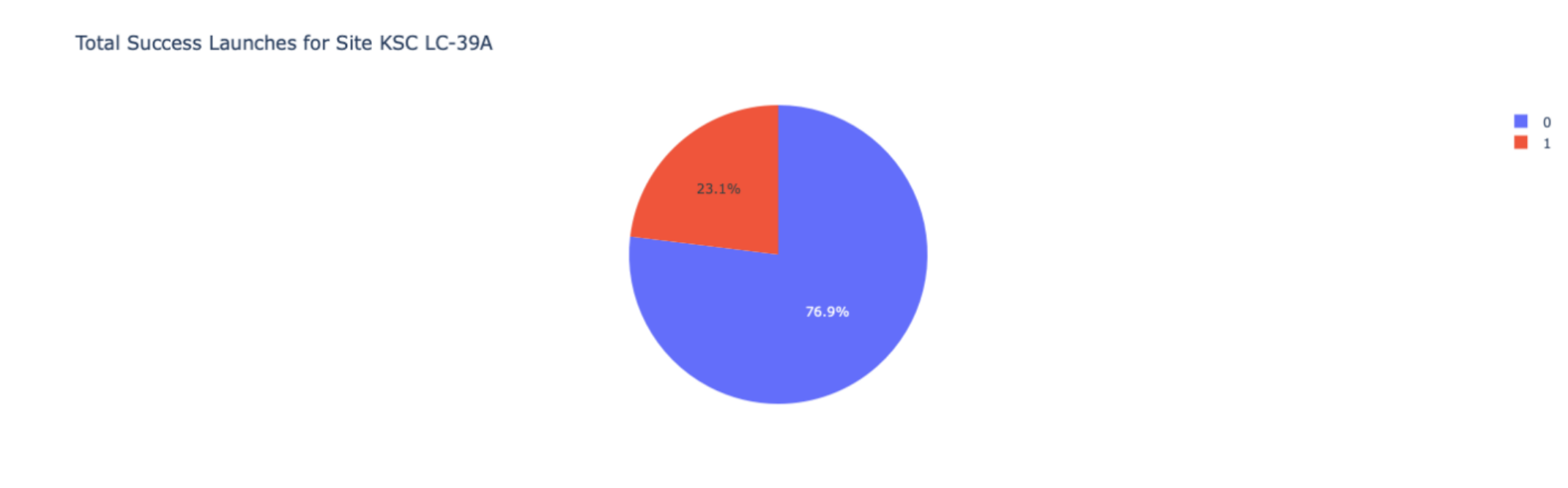

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



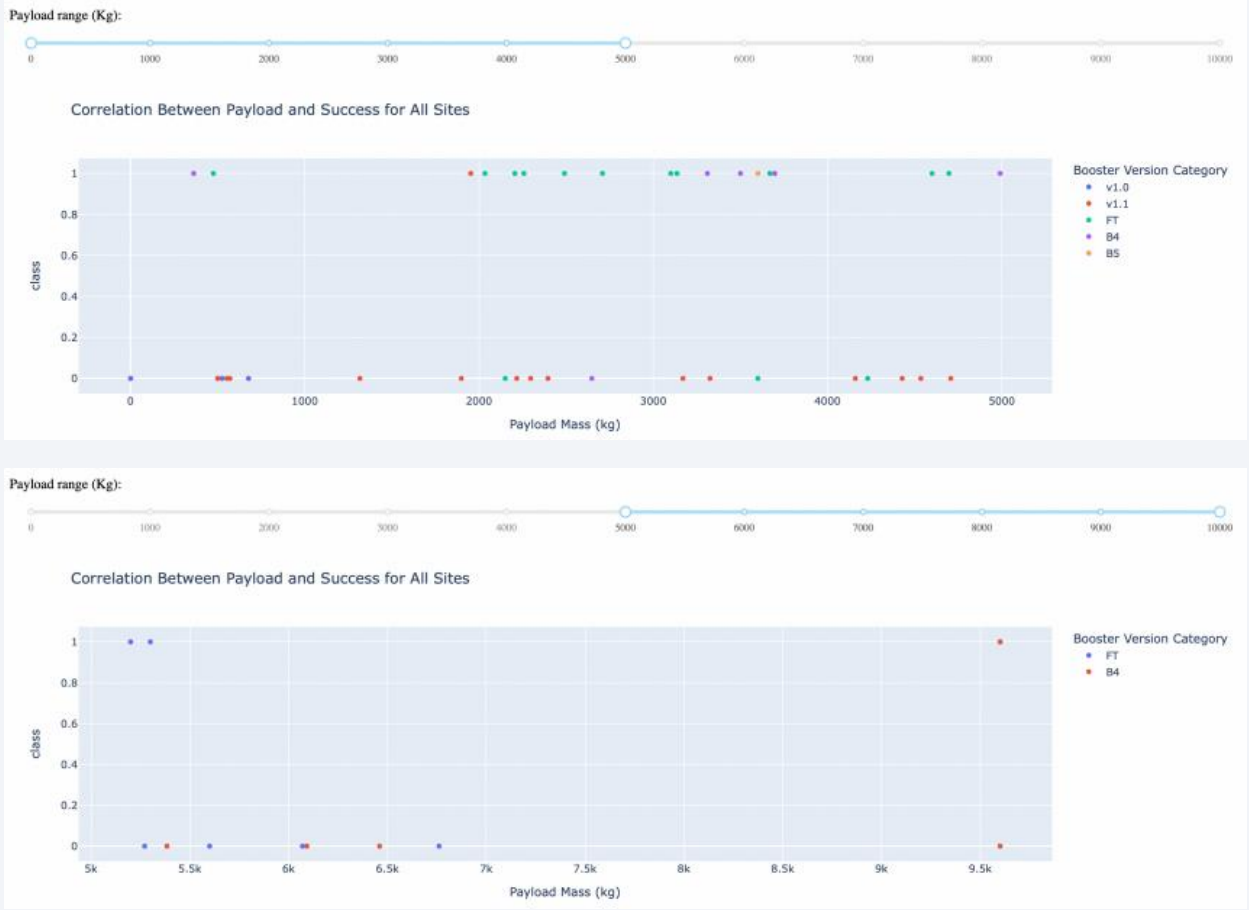
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Payload Mass vs. Launch Outcome for all sites



The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

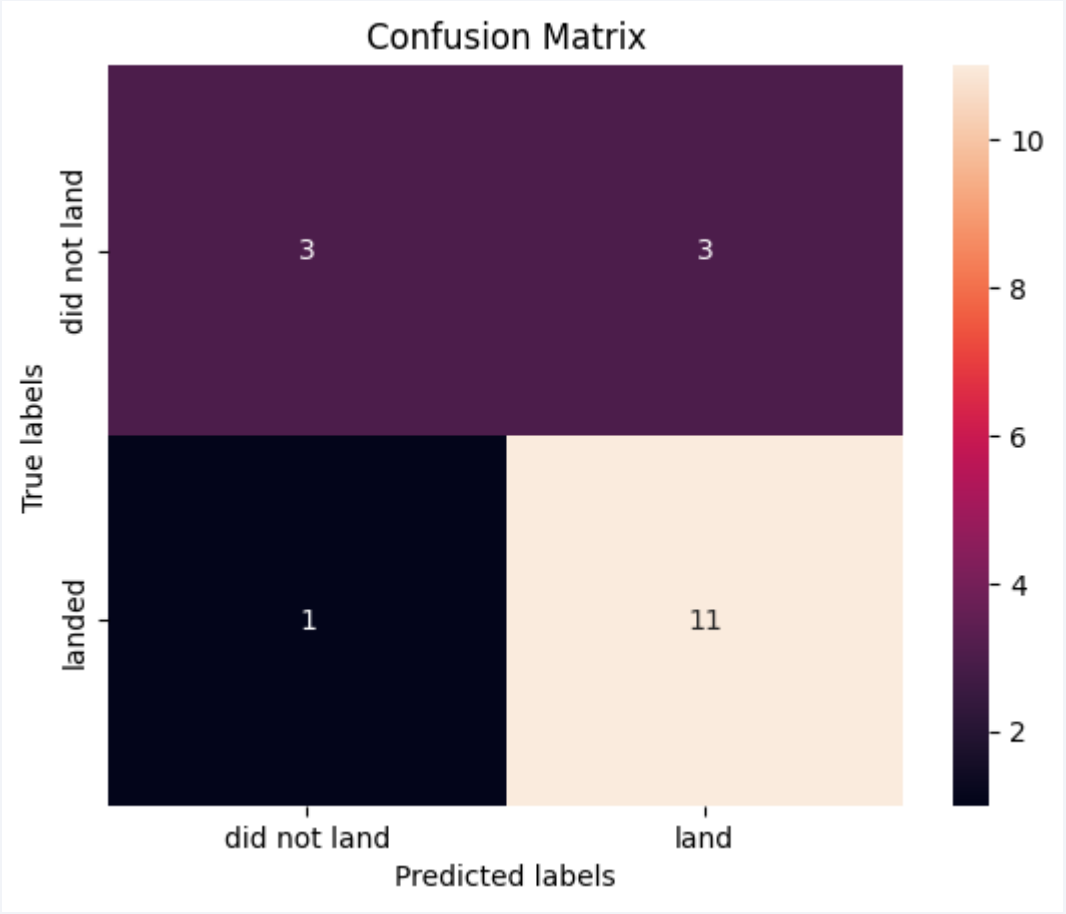
	LogReg	SVM	Tree	KNN
Accuracy	0,83333	0,83333	0,77777	0,83333

Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Accuracy	0,84643	0,84821	0,89107	0,84821

Confusion Matrix

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

IBM Data Science Professional Certificate

IBM

All the original documents at Github:

https://github.com/SalorG1/Applied-Data-Science-Capstone.git

Thank you!