IBM Developer
SKILLS NETWORK

SPACEX
Space Exploration Technologies

Winning Space Race with Data Science

Fırat Olçum
7 May 2023

Outline

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

Executive Summary

1. Summary of methodologies
   - Data collection
   - Data wrangling
   - Exploratory Data Analysis with Data Visualization
   - Exploratory Data Analysis with SQL
   - Building an interactive map with Folium
   - Building a Dashboard with Plotly Dash
   - Predictive analysis - Classification
2. Summary of all results
   - Exploratory Data Analysis results
   - Interactive analytics demo in screenshots
   - Predictive Analysis results

Introduction

1. Project Background and context

   - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollar; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

   - Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

2. Questions to be answered

   - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

   - Does the rate of successful landings increase over the years?

   - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

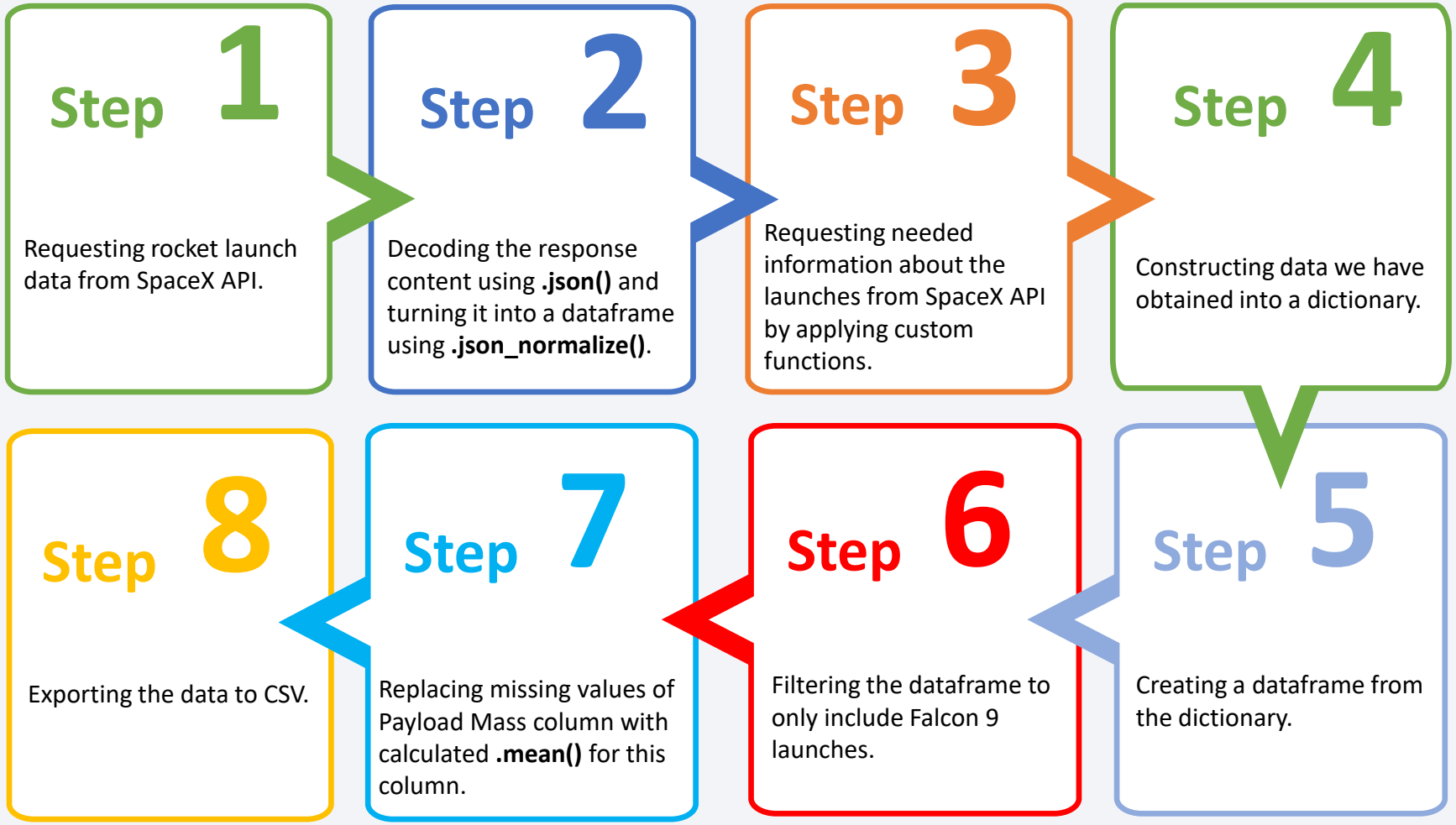
Executive Summary

- Data collection methodology:
  - We will be working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API.
  - We will be using the Python BeautifulSoup package to web scrape some HTM tables that contain valuable Falcon 9 launch records.
- Perform data wrangling
  - The data will be stored in lists and will be used to create our dataset. We will filter the data to remove irrelevant data.
  - We will deal with NULL values in order to make the dataset viable for analysis.
  - We will be using One Hot Encoding to prepare the data to a binary classification.
- 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 will be building, tuning and evaluating of classification models to ensure the best results.

Data Collection

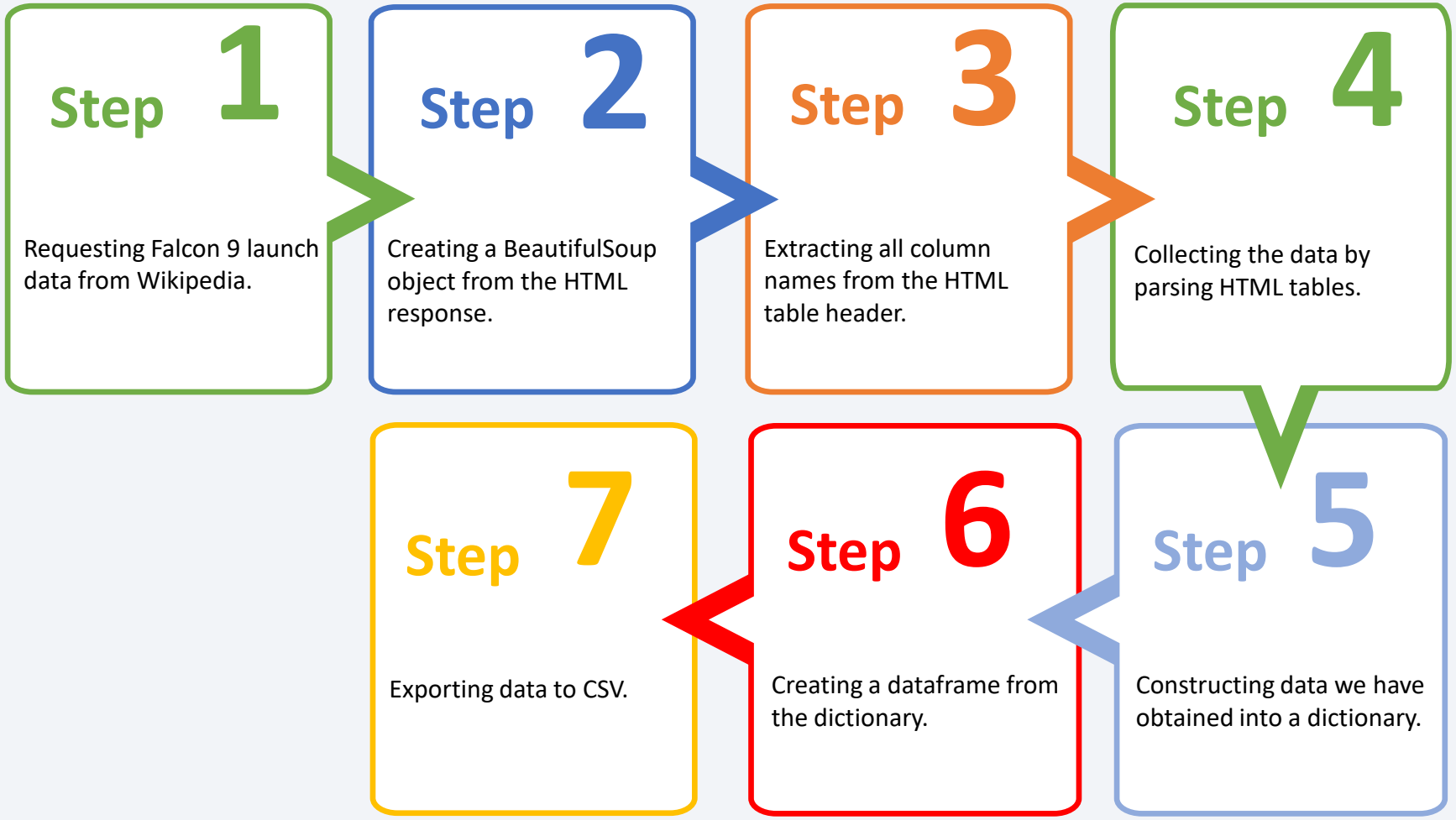
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.
- We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
    - Data Columns are obtained by using SpaceX REST API:
        - FligthNumber, 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



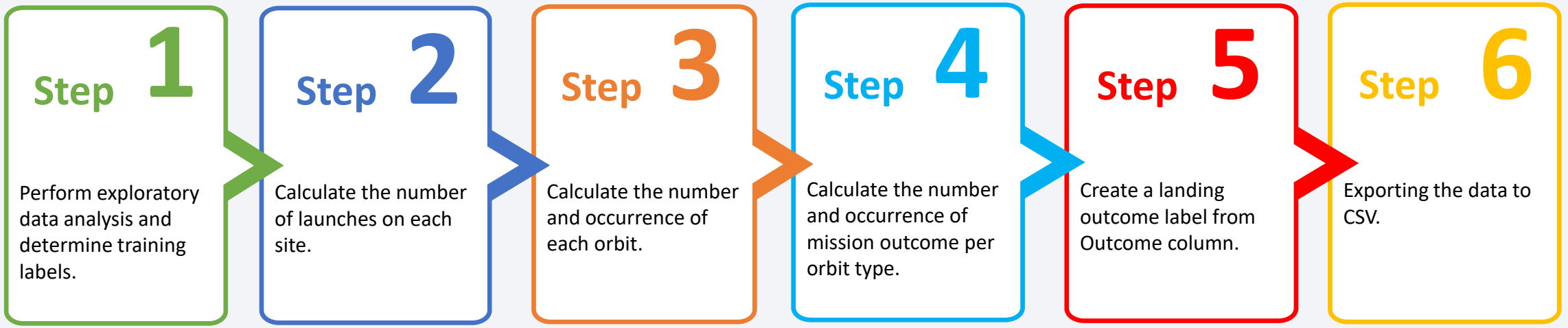
Click to see Jupyter Notebook about Data Collection API

Data Collection – Scraping



Click to see Jupyter Notebook about Web Scraping

Data Wrangling



In the dataset, there are different cases where the booster land or did not land successfully:

- True Ocean means that the mission outcome was successfully landed to a specific region of the ocean, while False Ocean means the opposite.
- True RTLS means that the mission outcome was successfully landed to a ground pad, while False RTLS means the opposite.
- True ASDS means that the mission outcome was successfully landed on a drone ship, while False ASDS means the opposite.

We mainly convert those outcomes into training labels with "1" means the booster successfully landed, "0" means it was unsuccessful.

Click to see Jupyter Notebook about Data Wrangling

EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.

- **Scatter plots** show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- **Bar charts** show comparisons among discrete categories. The goal is to show the relationships between the specific categories being compared and a measured value.
- **Line charts** show trends in data over time.

Click to see Jupyter Notebook about EDA with Data Visualization

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 the 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 between the date 2010-06-04 and 2017-03-20 in descending order.

Click to see Jupyter Notebook about EDA with SQL

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.

Colored Markers of the launch outcomes for each Launch Site:

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

Click to see Jupyter Notebook about Interactive Map with Folium

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:

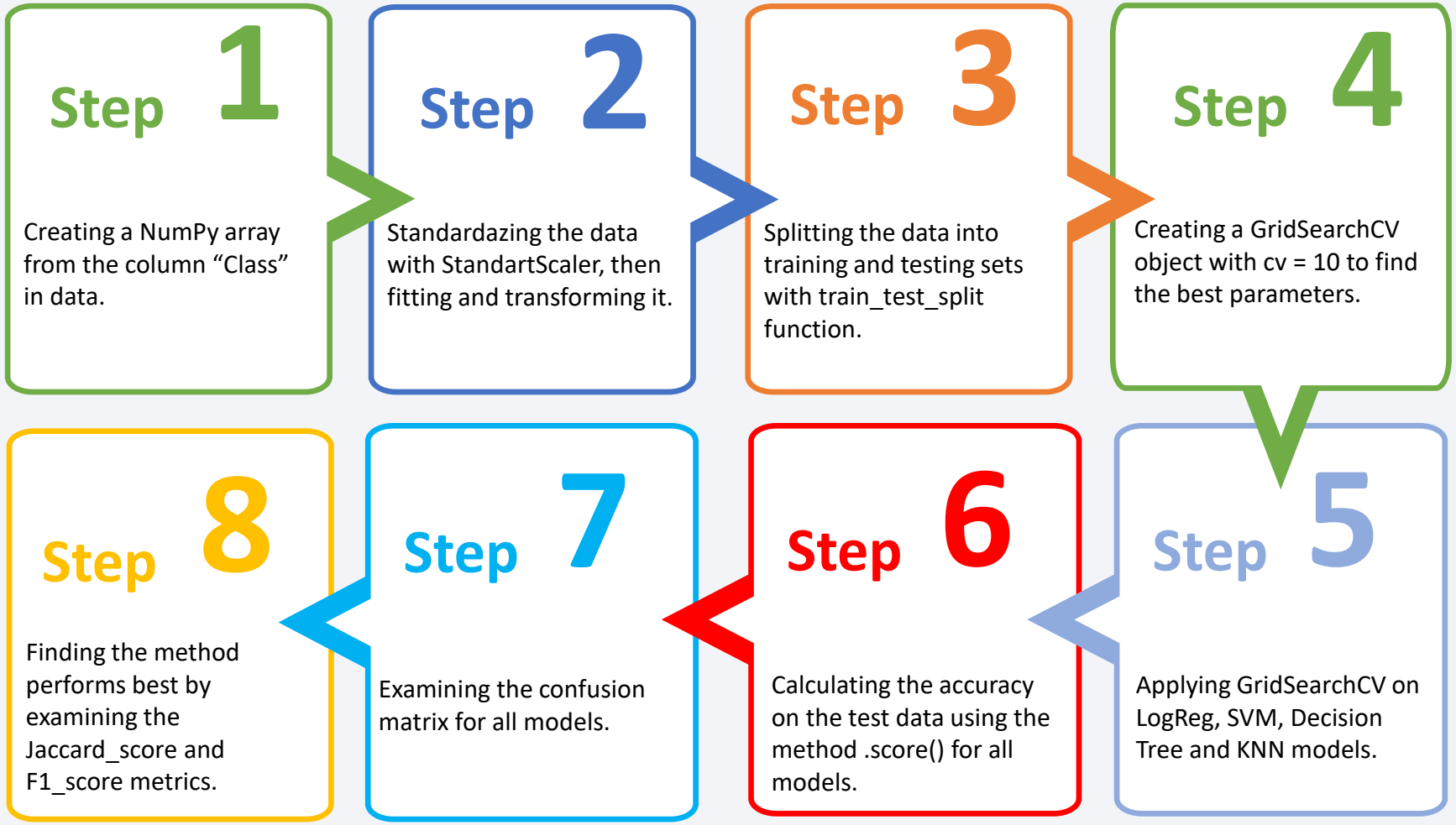
- Added a slider to select Payload range.

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.

Click to see Jupyter Notebook about Dashboard with Plotly Dash

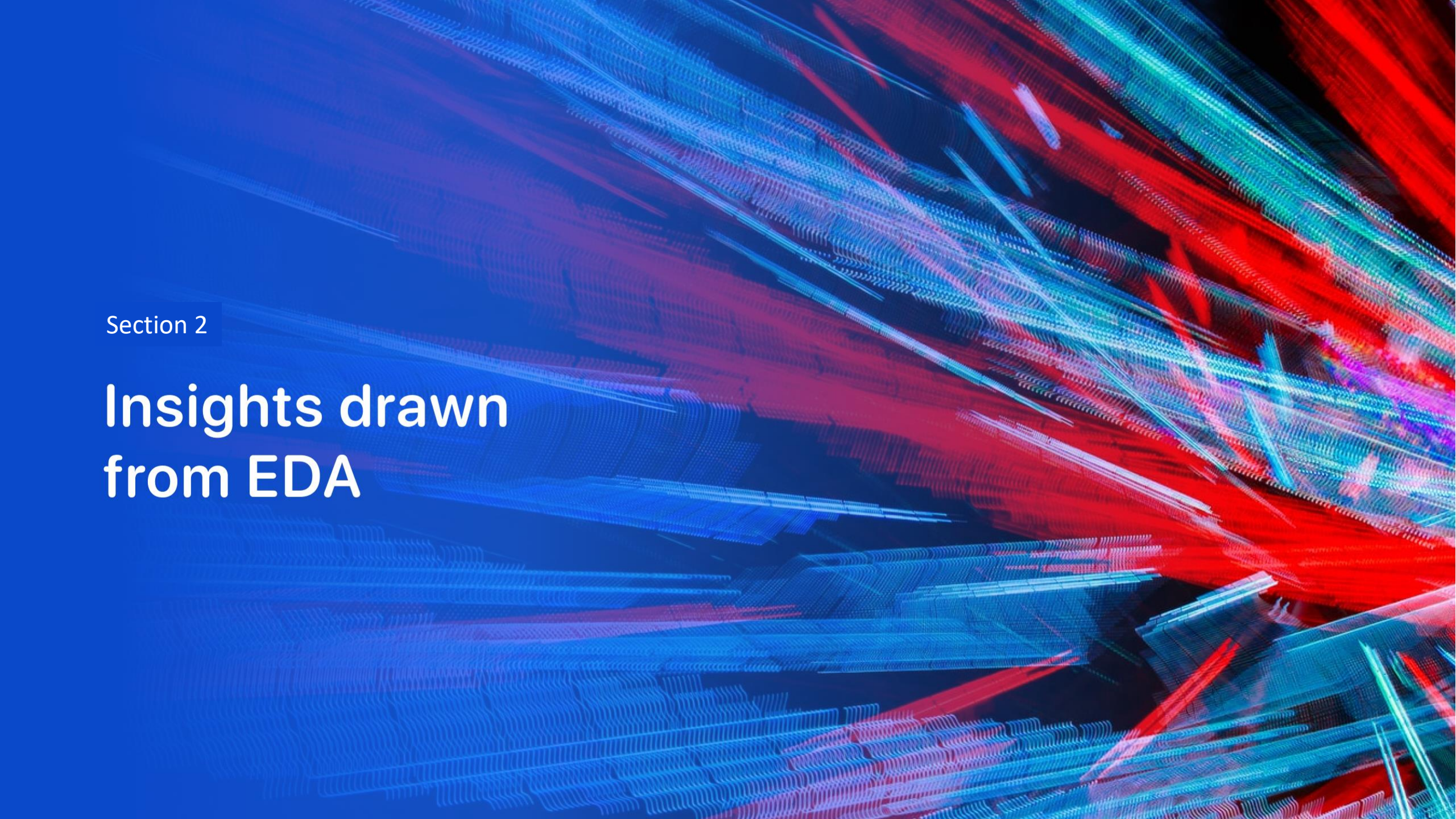
Predictive Analysis (Classification)



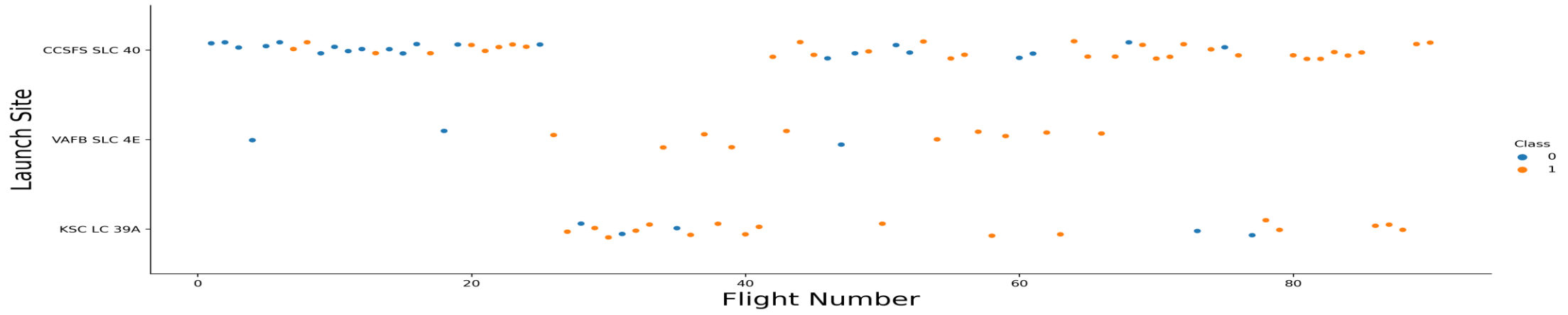
Click to see Jupyter Notebook about Predictive Analysis

Results

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


Section 2
Insights drawn from EDA

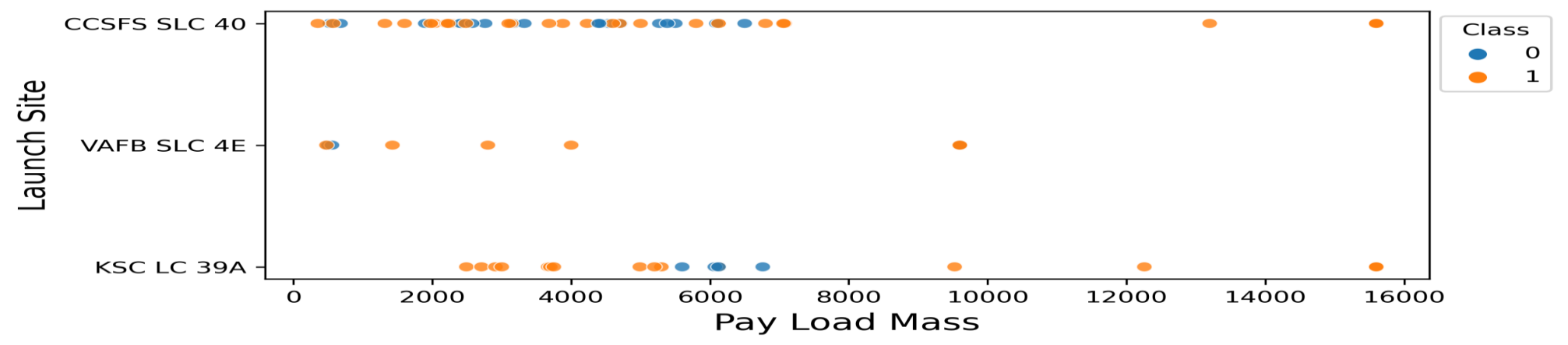
Flight Number vs. Launch Site



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.

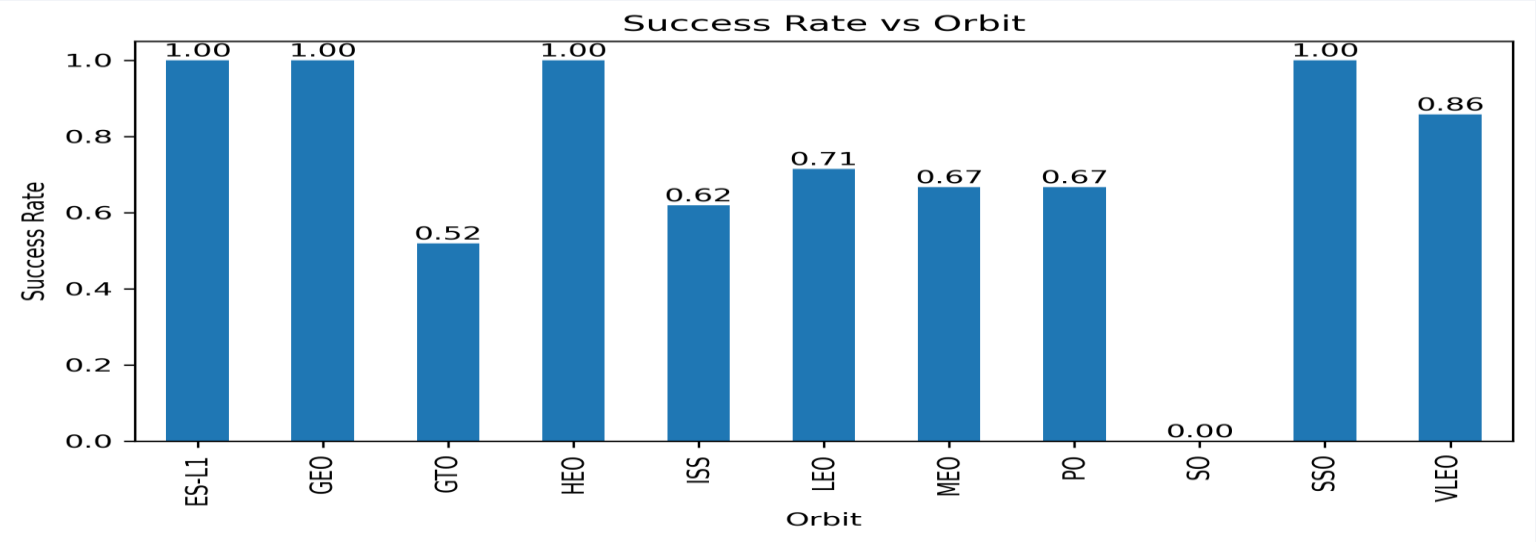
Payload vs. Launch Site



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.

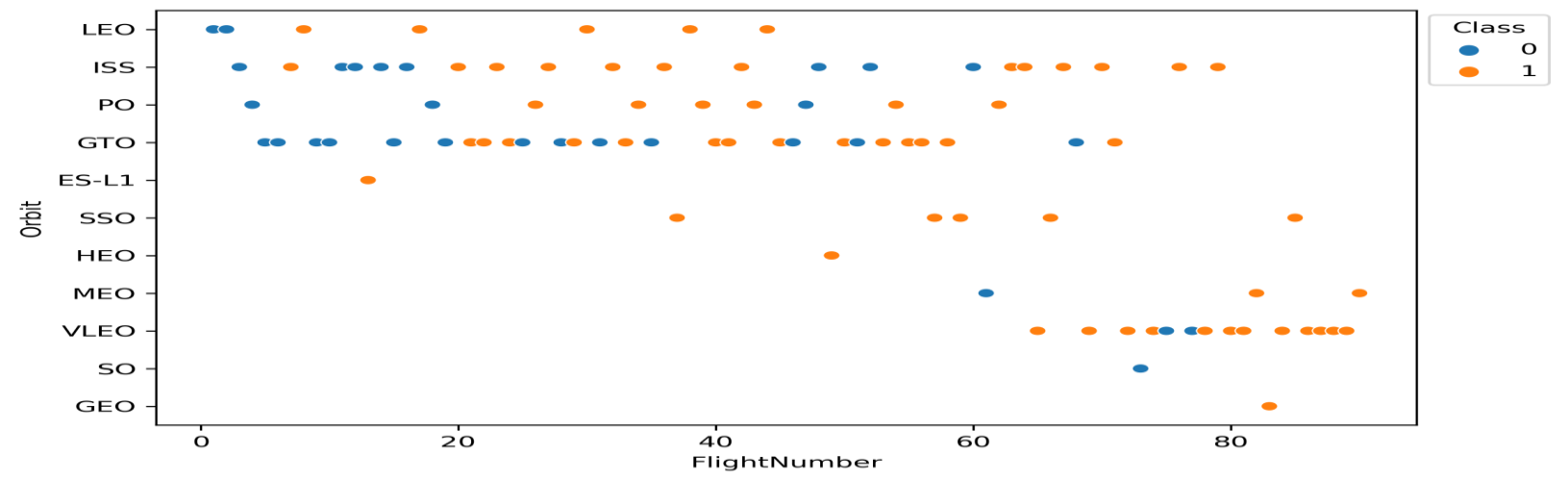
Success Rate vs. Orbit Type



Explanation:

- **Orbits with 100% success rate :** ES-L1, GEO, HEO, SSO
- **Orbits with success rate between 50% and 85% :** GTO, ISS, LEO, MEO, PO
- **Orbits with 0% success rate :** SO

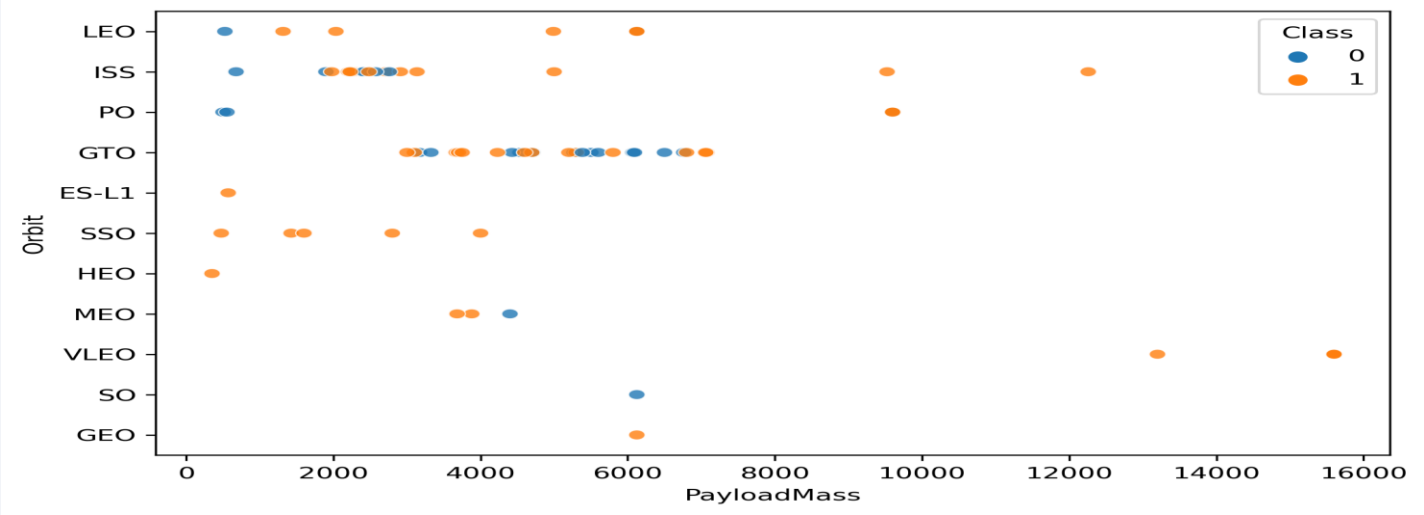
Flight Number vs. Orbit Type



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.

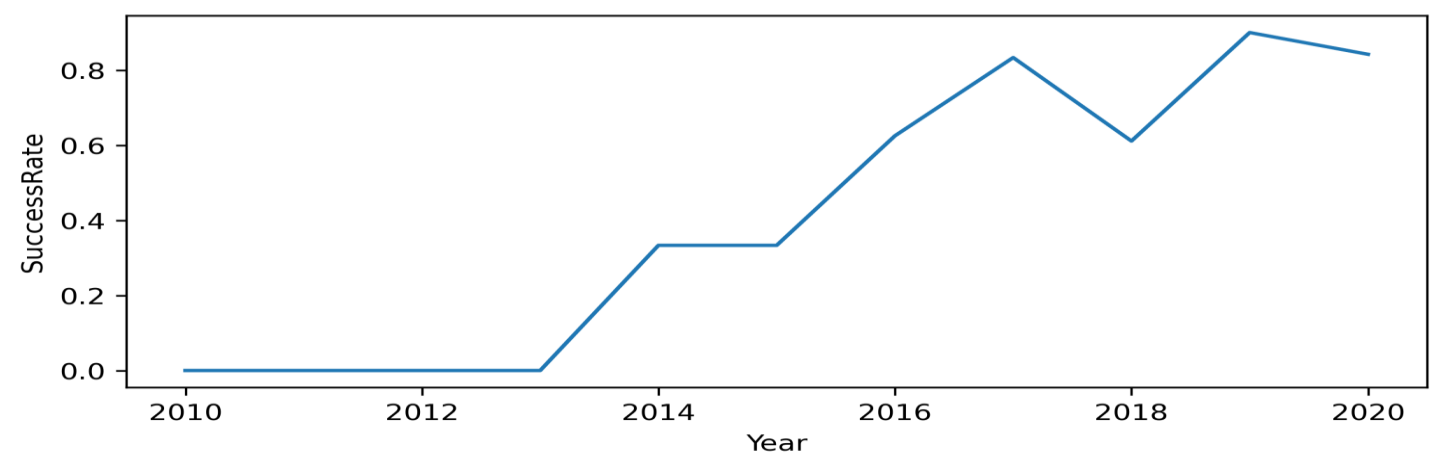
Payload Mass vs. Orbit Type



Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on ISS and LEO orbits.

Launch Success Yearly Trend



Explanation:

- The success rate since 2013 kept increasing until 2020.

All Launch Site Names

In [12]:
```
%%sql

SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

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

Out[12]:

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

Explanation:

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

Launch Site Names Begin with 'CCA'

In [14]:
```
%%sql

SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

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

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Displaying 5 records where launch sites begin with the string ‘CCA’.

Total Payload Mass

```
In [17]: %%sql

         SELECT SUM(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Customer = 'NASA (CRS)'

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

Out[17]:

SUM(PAYLOAD_MASS__KG_)
45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
In [19]: %%sql

         SELECT AVG(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Booster_Version = 'F9 v1.1'

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

Out[19]:

AVG(PAYLOAD_MASS__KG_)
2928.4

Explanation:

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

First Successful Ground Landing Date

In [23]:
```
%%sql

SELECT MIN(DATE)
FROM SPACEXTBL
WHERE "Landing _Outcome" = "Success (ground pad)"
```

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

Out[23]:

MIN(DATE)
01-05-2017

Explanation:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

In [18]:
```
%%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Out[18]:

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

Explanation:

- Listing the 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

```
In [19]: %%sql

         SELECT Mission_Outcome, COUNT(Mission_Outcome)
         FROM SPACEXTBL
         GROUP BY Mission_Outcome

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

Out[19]:

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

Explanation:

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [20]: %%sql

         SELECT DISTINCT Booster_Version
         FROM SPACEXTBL
         WHERE PAYLOAD_MASS__KG_ = ((SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL))

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

Out[20]:

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

Explanation:

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

2015 Launch Records

In [24]:
```
%%sql

SELECT SUBSTR(Date, 4, 2) AS Month, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)' AND SUBSTR(Date, 7, 4) = '2015'
```

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

Out[24]:

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

Explanation:

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

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

In [25]:
```
%%sql

SELECT Date, "Landing _Outcome", COUNT("Landing _Outcome") AS Count
FROM SPACEXTBL
WHERE SUBSTR(Date,7)||SUBSTR(Date,4,2)||SUBSTR(Date,1,2)
      BETWEEN '20100604' AND '20170320' AND "Landing _Outcome" LIKE 'Success%'
GROUP BY "Landing _Outcome"
ORDER BY 3 DESC
```

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

Out[25]:

Date	Landing _Outcome	Count
08-04-2016	Success (drone ship)	5
22-12-2015	Success (ground pad)	3

Explanation:

- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

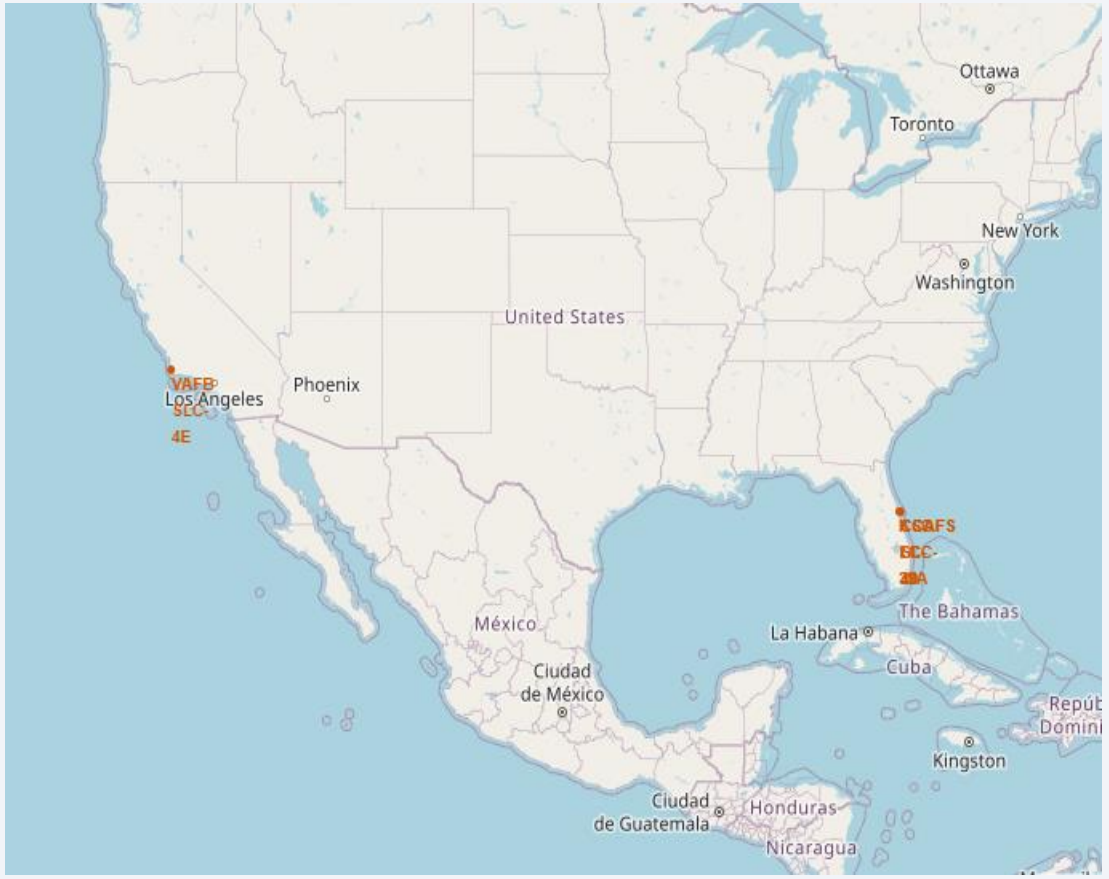
Section 3

Launch Sites Proximities Analysis

All Launch Sites On a Global Map

Explanation:

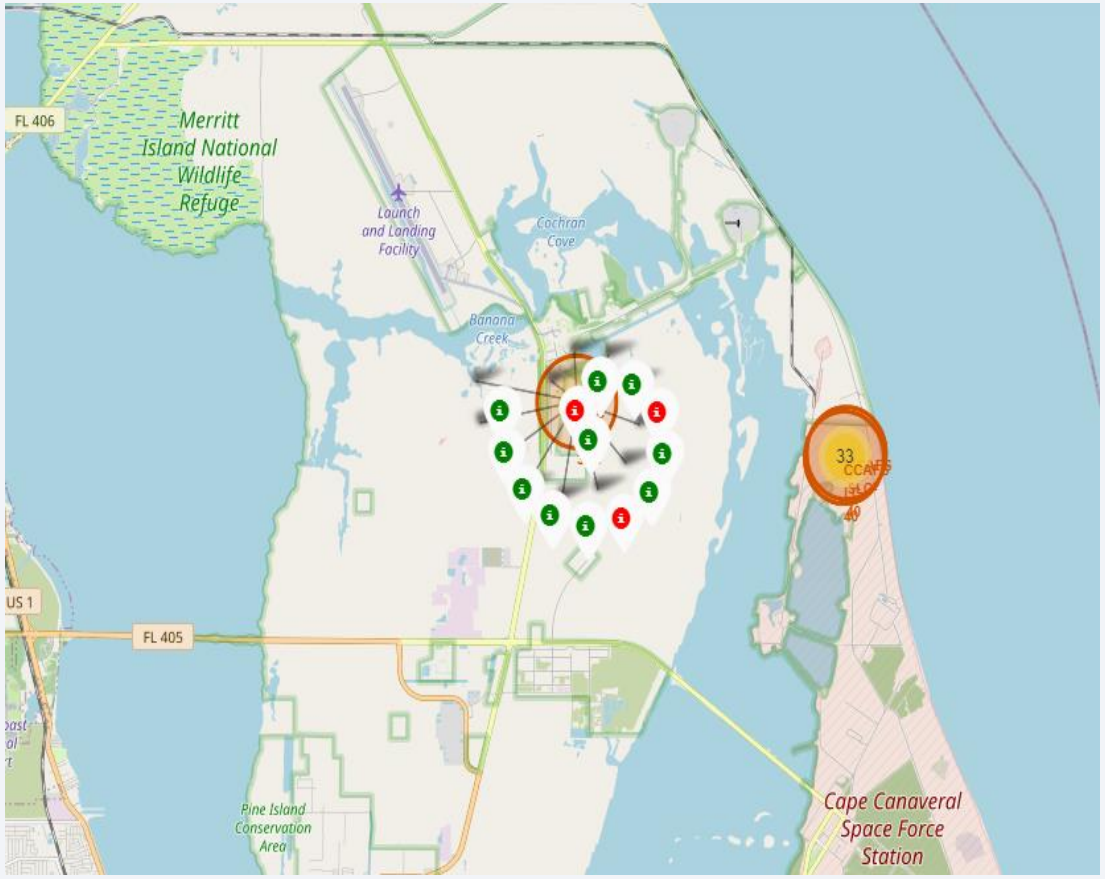
- Most of the Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour.
- If a ship is launched from the equator it goes up into space, and it also moving around the Earth at the same speed it was moving before launching. This is because of inertia. The 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 minimizes the risk of having any debris dropping or exploding near people.



Launch Records on a Global Map

Explanation:

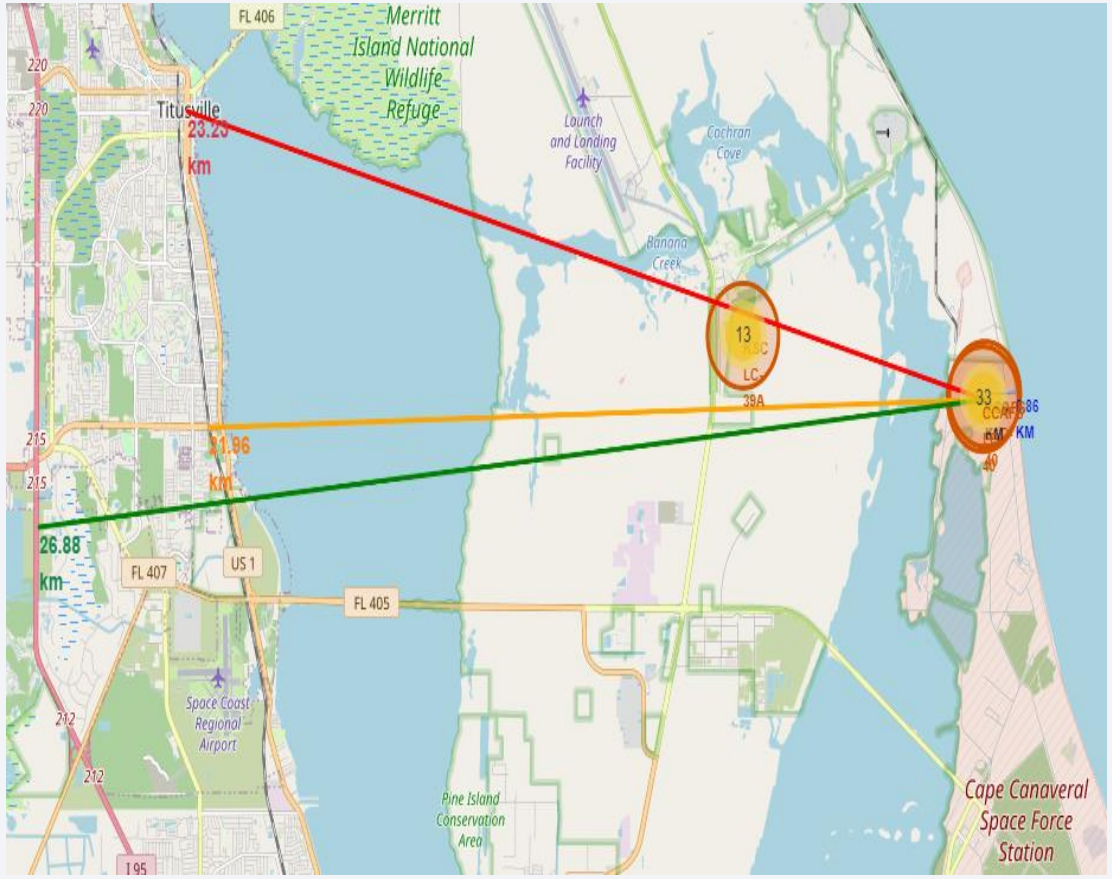
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker : Successful Launch
  - Red Marker : Failed Launch
- Launch site KSC LC-39A has a very high success rate.

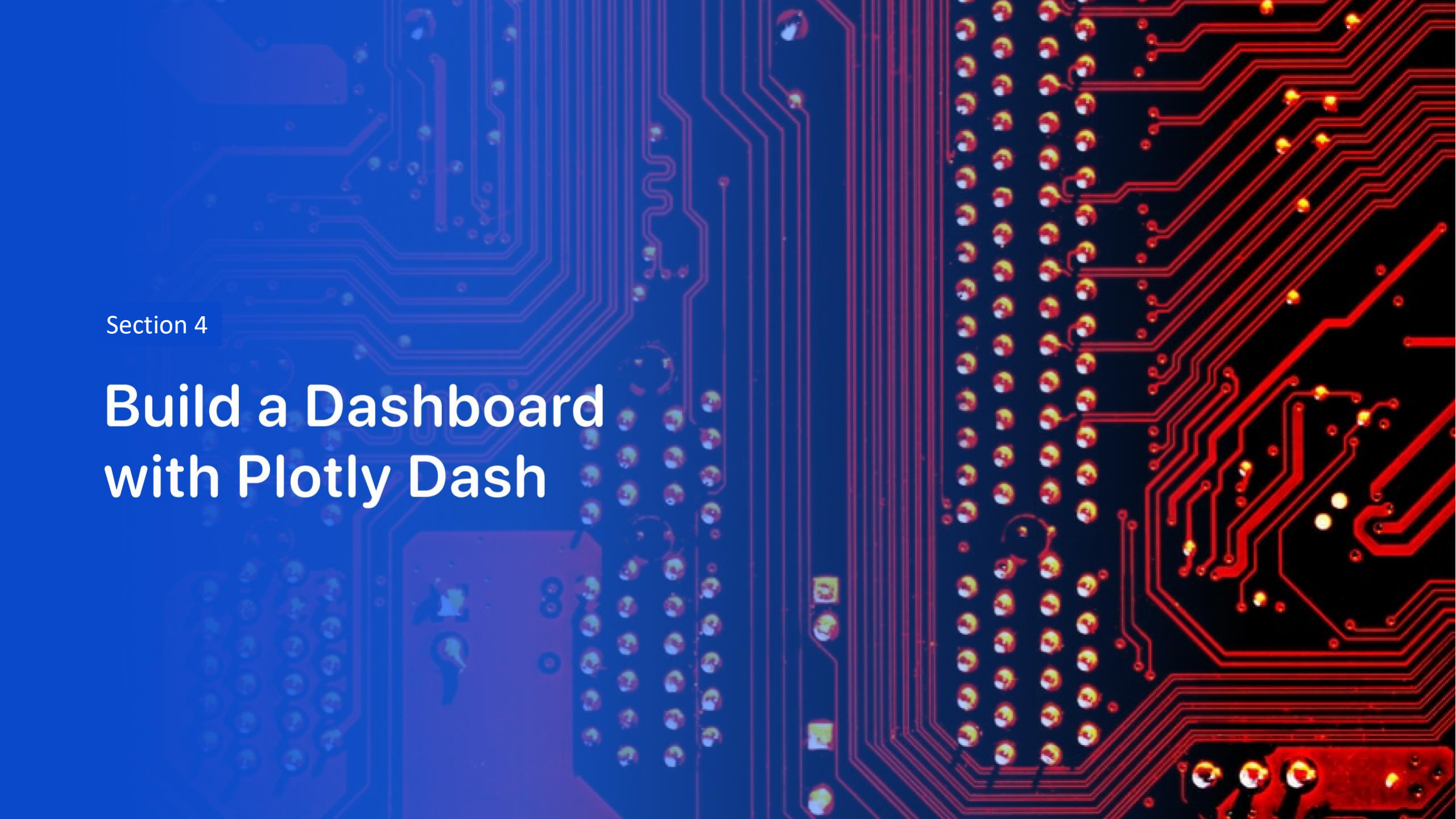


Distance From the Launch Site to Its Proximities

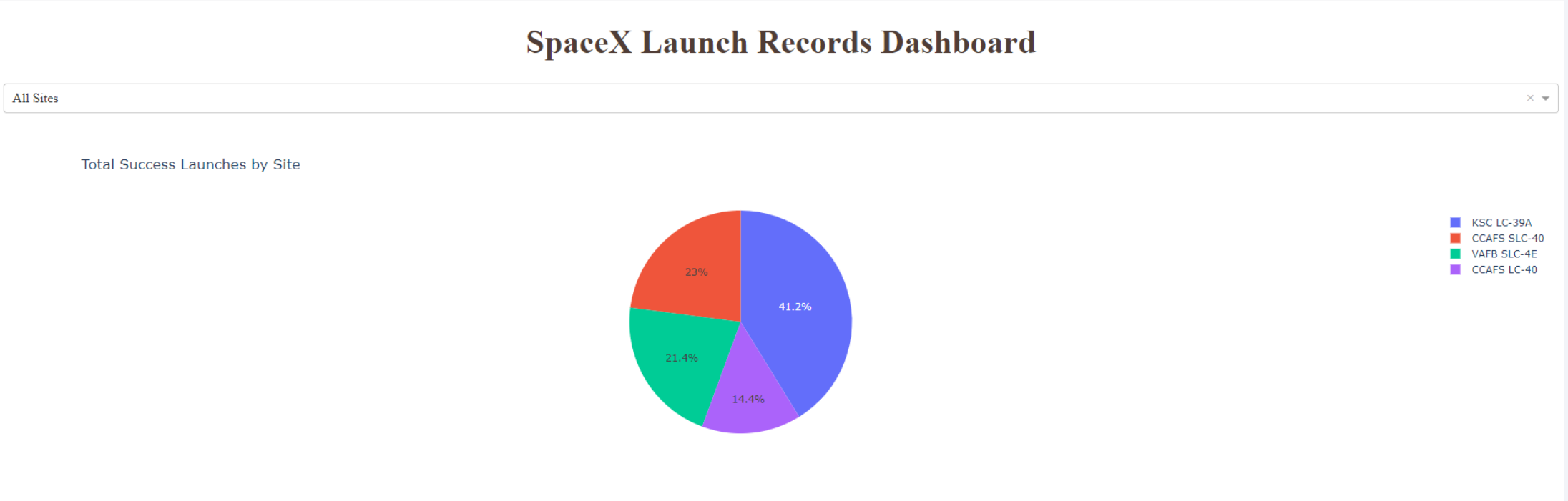
Explanation:

- From the visual analysis of the launch site CCAFS LC-40 we can clearly see that it is:
  - Relatively close to railway : 21.96 km
  - Relatively close to highway : 26.88 km
  - Relatively close to coastline : 0.86 km
- Also, the launch site CCAFS LC-40 is relatively close to its closest city Titusville : 23.23 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

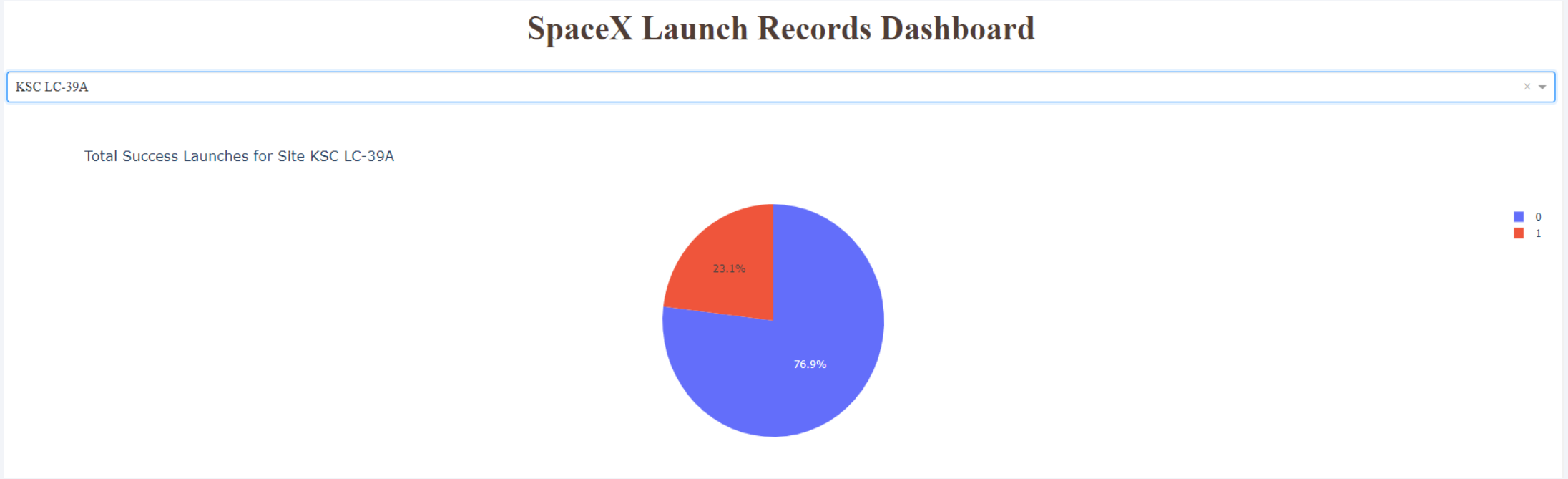
Total Success Launches by Site



Explanation:

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

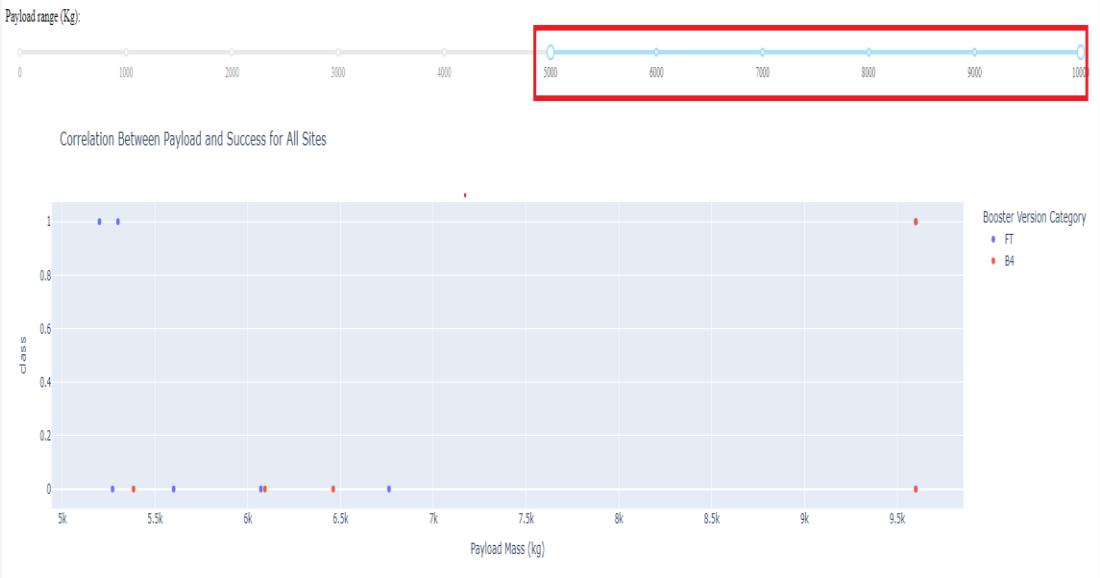
Total Success Launches for Site KSC LC-39A



Explanation:

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for All Sites



Explanation:

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the Test Set

	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

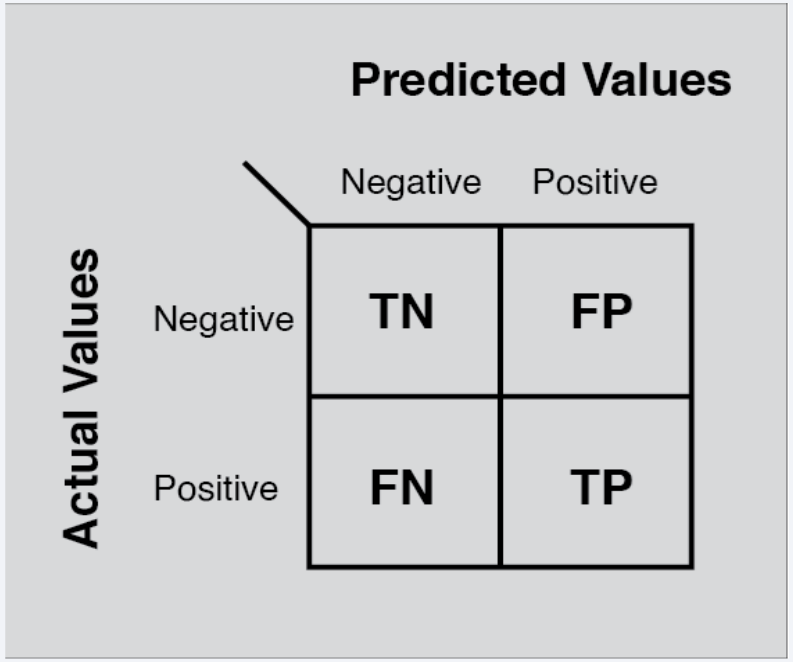
Scores and Accuracy of the Entire Data Set

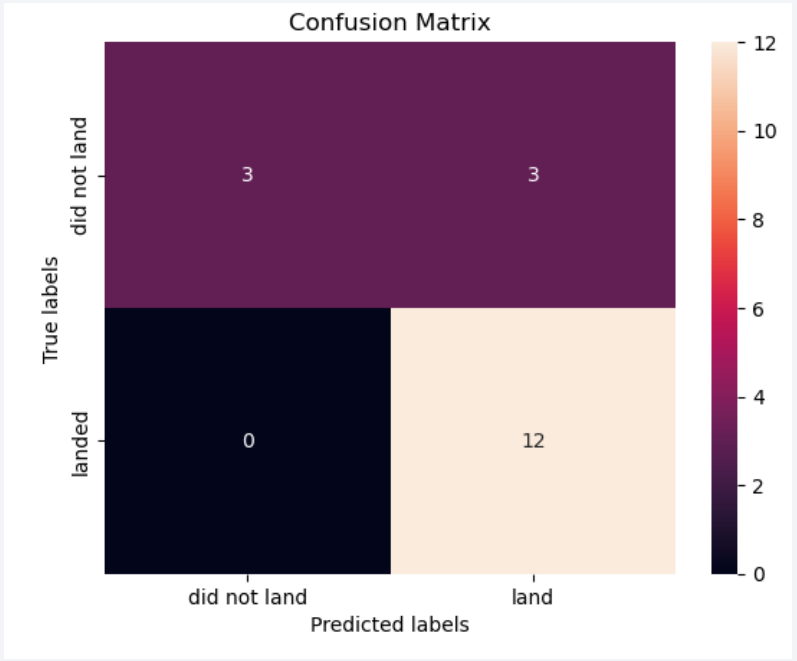
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Explanation:

- Based on the scores of the test set we can not confirm which method performs best.
- Same test 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.

Confusion Matrix





Explanation:

- 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 consists of 10 courses:

1. What is Data Science?
2. Tools for Data Science
3. Data Science Methodology
4. Python for Data Science, AI & Development
5. Python Project for Data Science
6. Databases and SQL for Data Science with Python
7. Data Analysis with Python
8. Data Visualization with Python
9. Machine Learning with Python
10. Applied Data Science Capstone

Thank you!