IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thang Nguyen
19 Sep 2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection and wrangling
  - Exploratory data analysis using SQL, Pandas and Matplotlib
  - Interactive visual analytics with Folium
  - Interactive dashboard with Plotly Dash
  - Predictive analysis (Classification)
- **Summary of all results**
  - Data collection and wrangling results
  - Exploratory data analysis results
  - Interactive analytics results
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches, with a cost of 62 million dollars; 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. This information can be used if Space Y wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - How do variables (payload mass, launch site, number of flights and orbits) impact success of first stage landings?
  - Estimate total cost of launches by predicting success of first stage landings
  - Where is the best place to launch?

Section 1

Methodology

Methodology

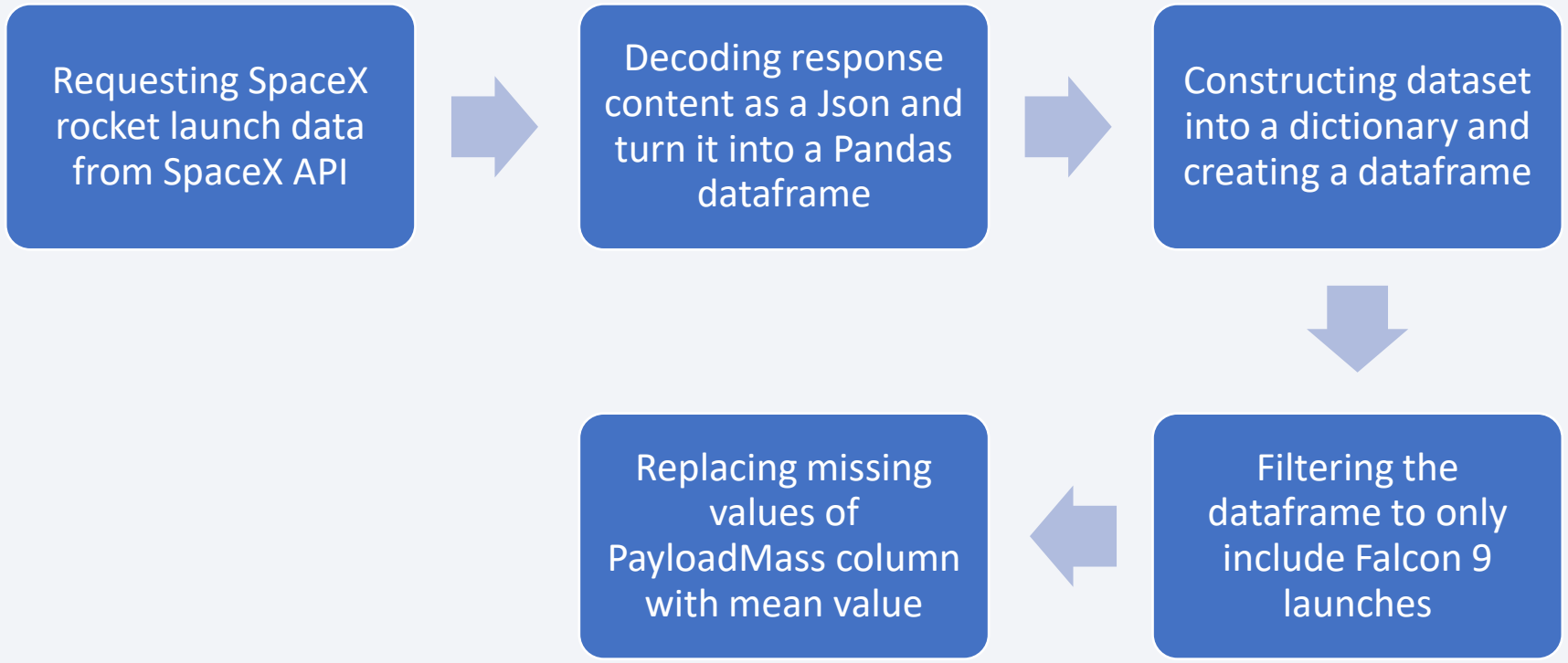
Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web scraping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
- 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 divided in training and test data sets, evaluated by 4 different classification models

Data Collection

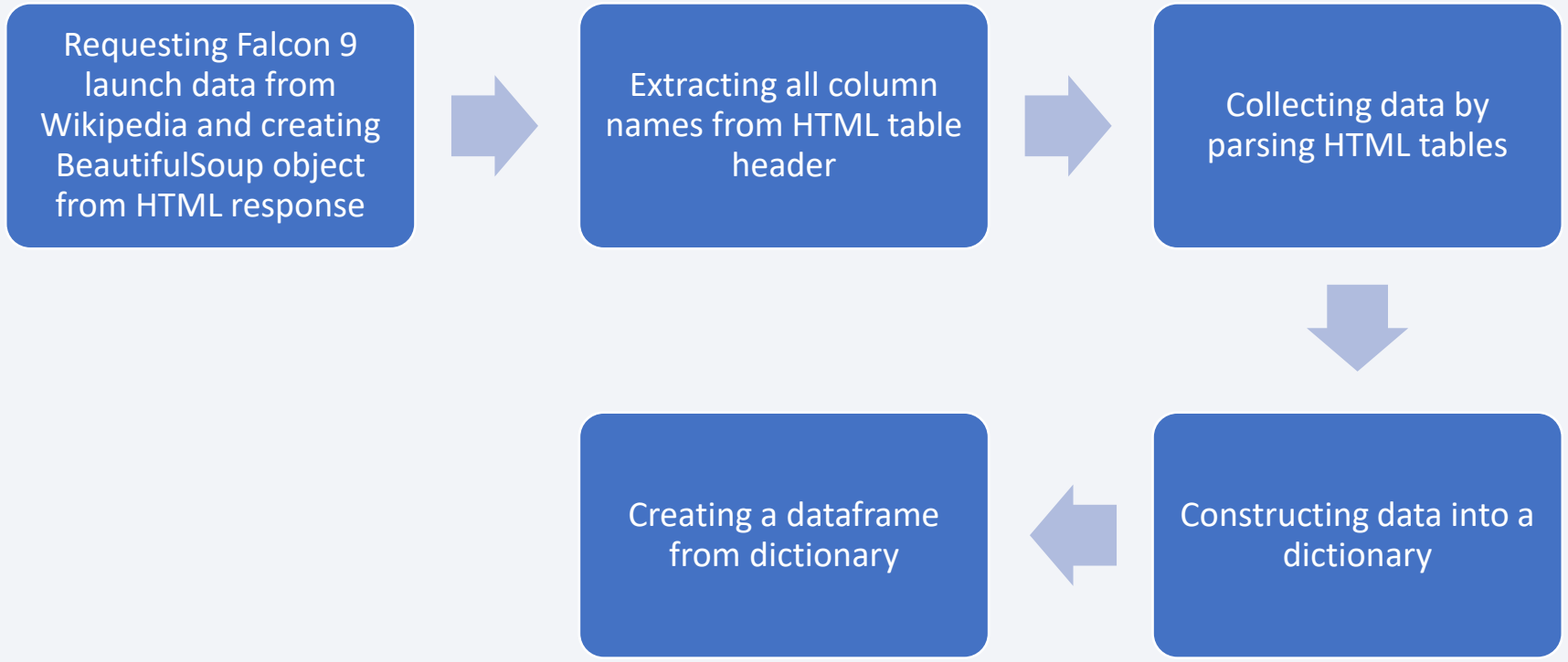
- Data sets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and web scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API



Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Some Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models.
- Calculated summaries on launches per site, occurrences of each orbit, and occurrences of mission outcome of the orbits
- Created landing outcome label from outcome column

Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts were plotted: FlightNumber vs. PayloadMass, FlightNumber vs. LaunchSite, PayloadMass vs. LaunchSite, FlightNumber vs. Orbit, PayloadMass vs. Orbit, and Success Year Trend
- Scatter plots show the relationship between variables.
- Bar charts show comparisons among discrete categories.
- Line charts show trends in data over time.

Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed:

- 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- 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

Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines, and market clusters were used with Folium Map
  - ➢Markers indicate points such as launch sites
  - ➢Circles indicate highlighted areas with specific coordinates
  - ➢Marker clusters indicate groups of markers having the same coordinate
  - ➢Lines indicate distances between 2 coordinates

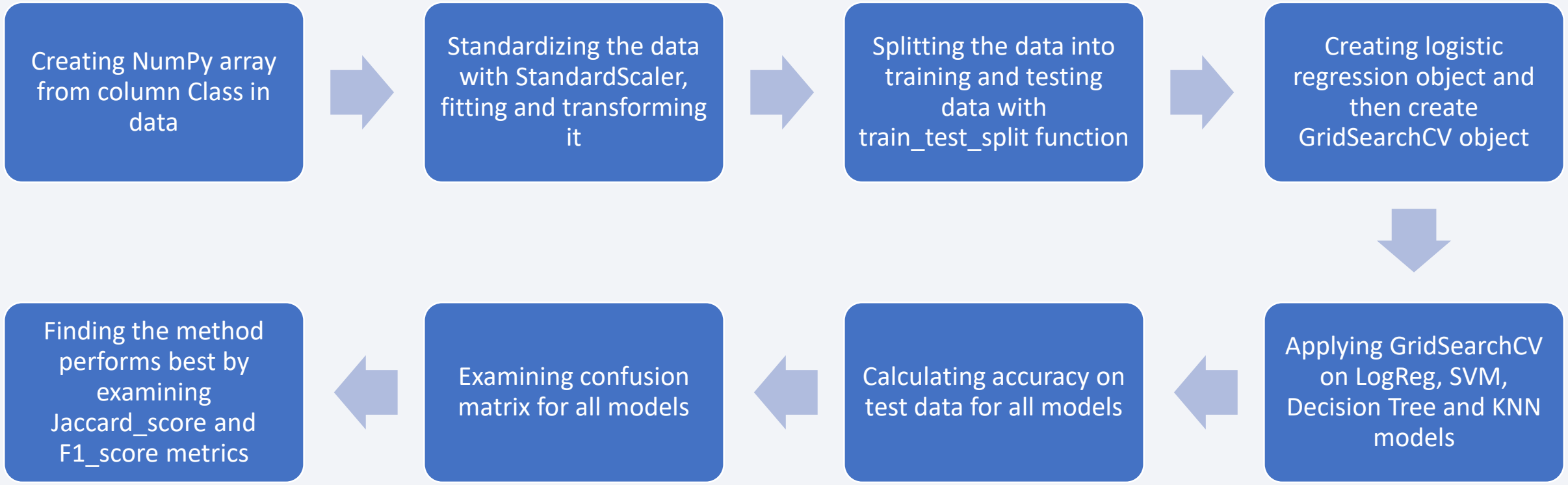
Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added dropdown list to select different Launch Sites
- Pie Chart visualizing launch success counts:
  - Added pie chart to show successful launches count for all sites and the success vs. failed counts for the site
- Slider to select payload:
  - Added slider to select Payload range
- Scatter plot of Payload vs. Success Rate:
  - Added scatter chart to show correlation between Payload and Launch Success for selected site

Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/spacex_dash_app.py

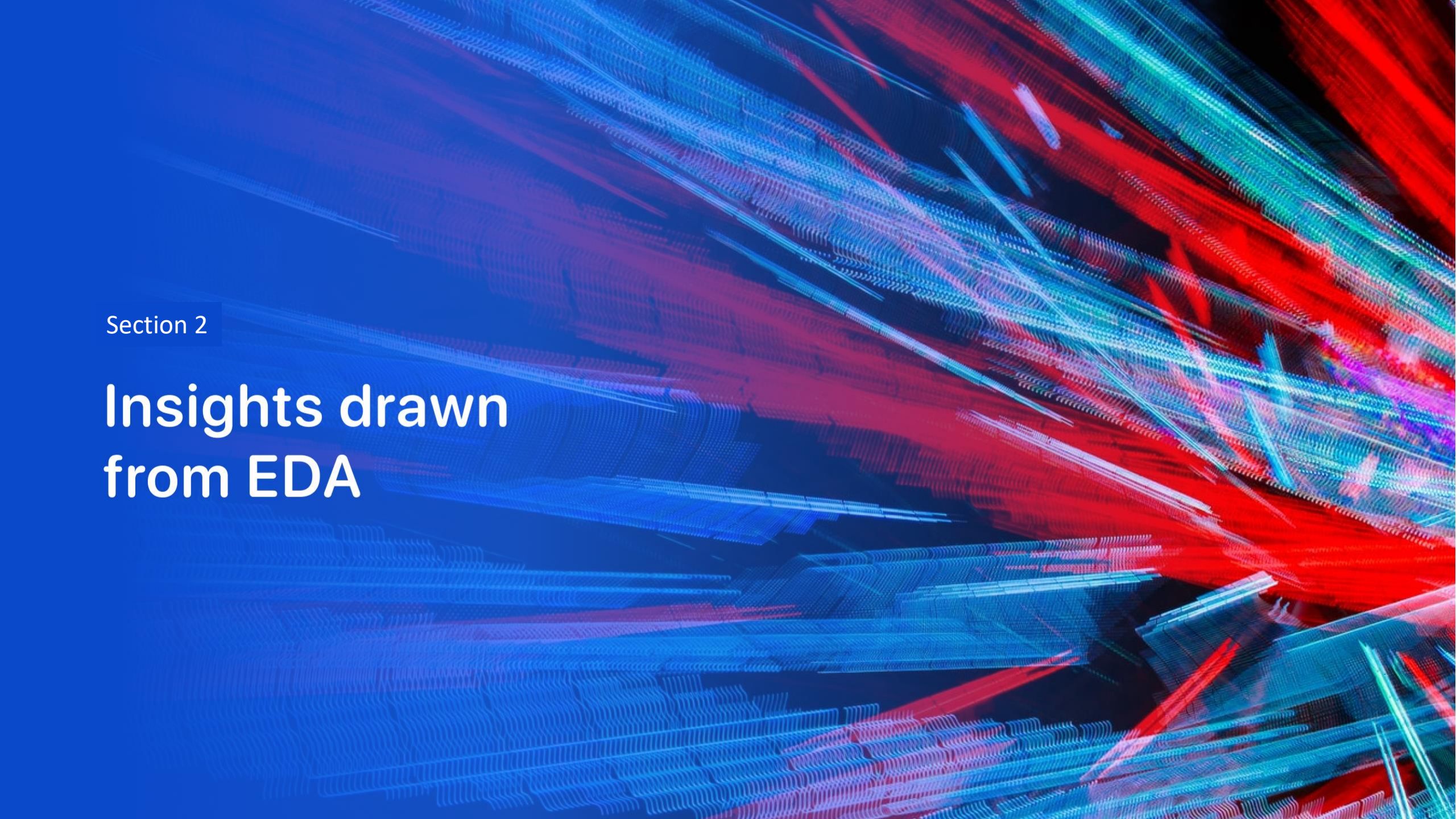
Predictive Analysis (Classification)



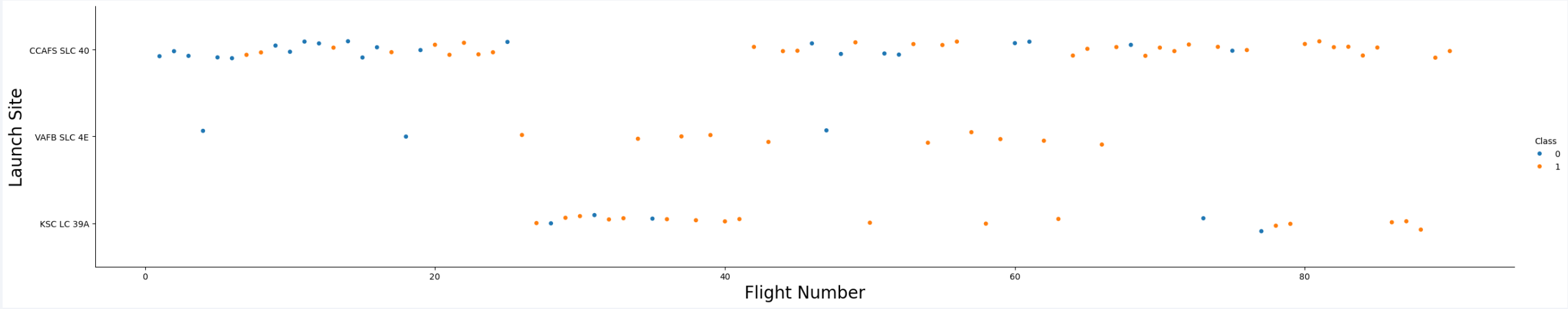
Source code: https://github.com/tnguyen259/data-science-capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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


Section 2
Insights drawn from EDA

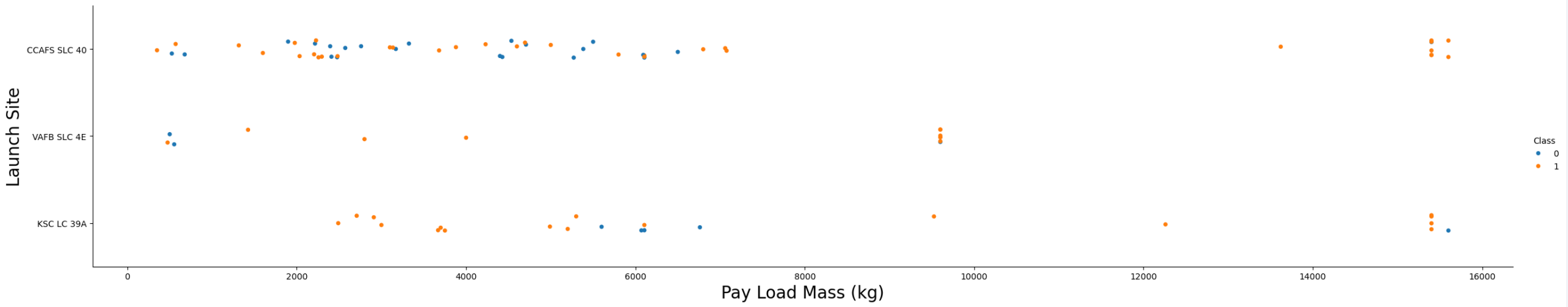
Flight Number vs. Launch Site



Explanation:

- Each new launch has higher rate of success
- CCAFS SLC 40 launch site has approximately half of all launches
- VAFB SLC 4E and KSC LC 39A have higher rate of success

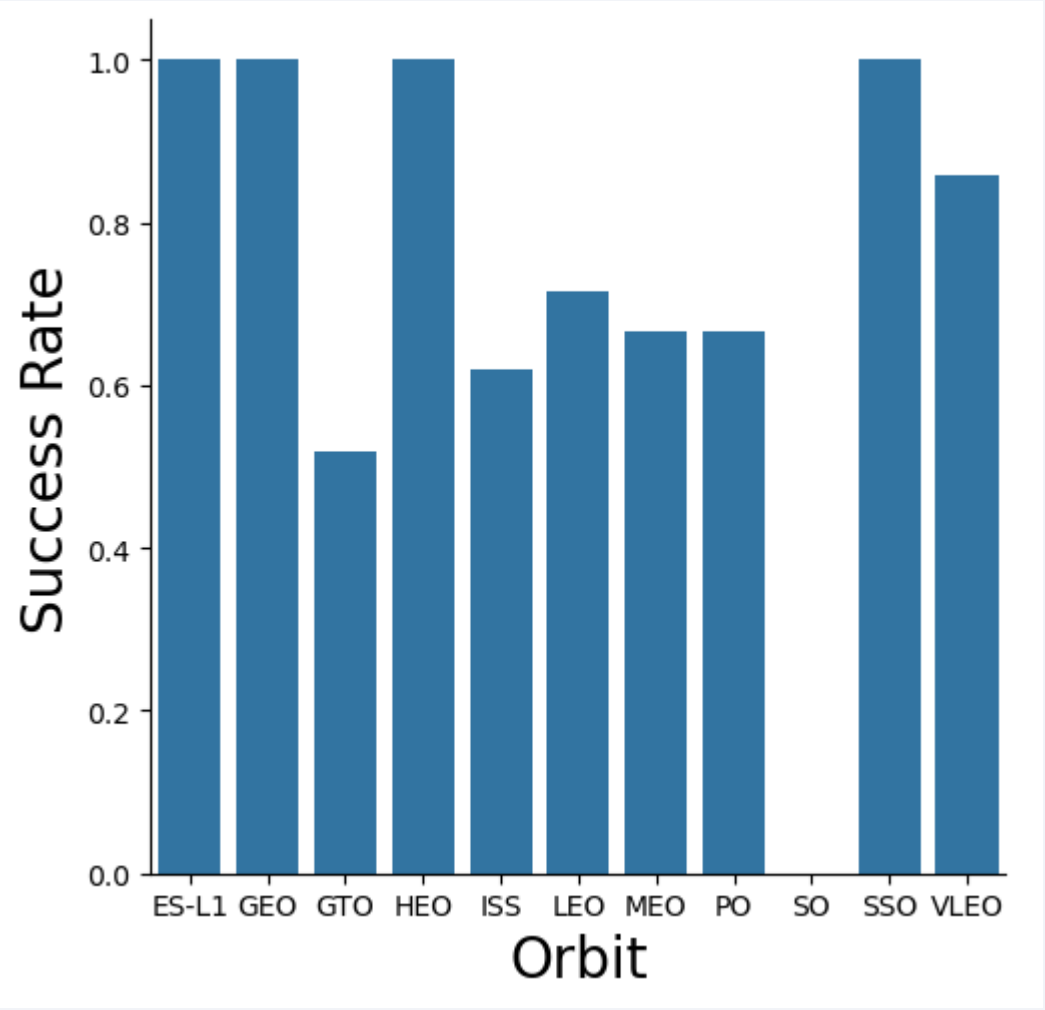
Payload vs. Launch Site



Explanation:

- Majority of launches with payload mass over 9000kg were successful
- Launches with payload mass over 12000kg only happened at CCAFS SLC 40 and KSC LC 39A launch sites

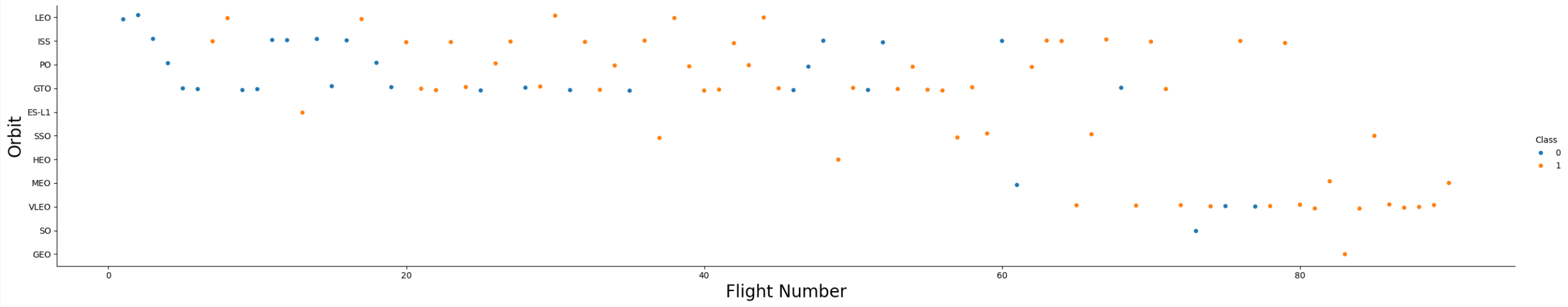
Success Rate vs. Orbit Type



Explanation:

- Orbits with 100% success rate: ES-L1, GEO, HEO, and SSO
- Orbit with 0% success rate: SO

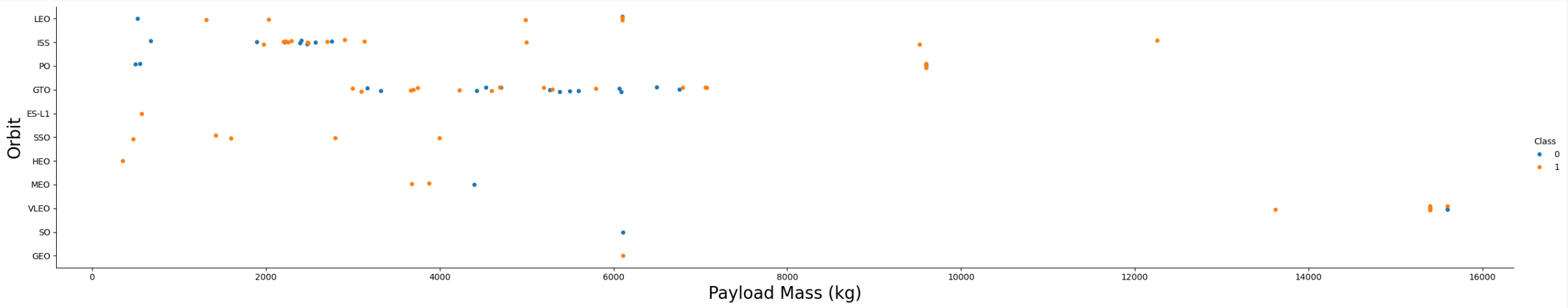
Flight Number vs. Orbit Type



Explanation:

- Success rate improved over time for all orbits
- 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

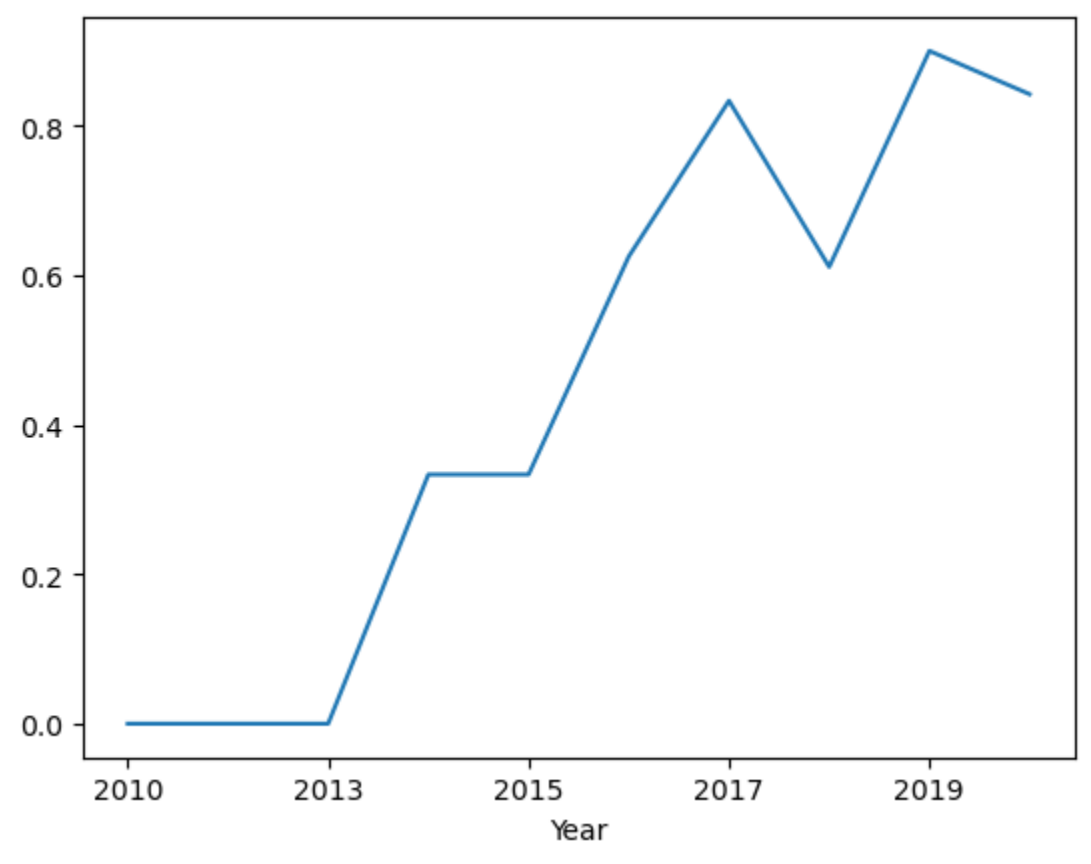
Payload vs. Orbit Type



Explanation:

- With heavy payloads the successful landing or positive landing rate are more for PO, VLEO and ISS
- There were much less launches for SO and GEO

Launch Success Yearly Trend



Explanation:

- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Explanation:

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

Launch Site Names Begin with '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

Explanation:

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

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

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

TOTAL_PAYLOAD
45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE Booster_Version = "F9 v1.1";
```

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

AVG_PAYLOAD
2928.4

Explanation:

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

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_LAND FROM SPACEXTBL WHERE LANDING_OUTCOME = "Success (ground pad)";
```

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

FIRST_SUCCESS_LAND
2015-12-22

Explanation:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = "Su
```

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

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

Explanation:

- List 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

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Explanation:

- List the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%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 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Explanation:

- List the names of the booster_versions which have carried the maximum payload mass.

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) AS MONTH, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTBL where Landing_Outcome =
 * sqlite:///my_data1.db
Done.
```

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

Explanation:

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS TOTAL FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LAND
```

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

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

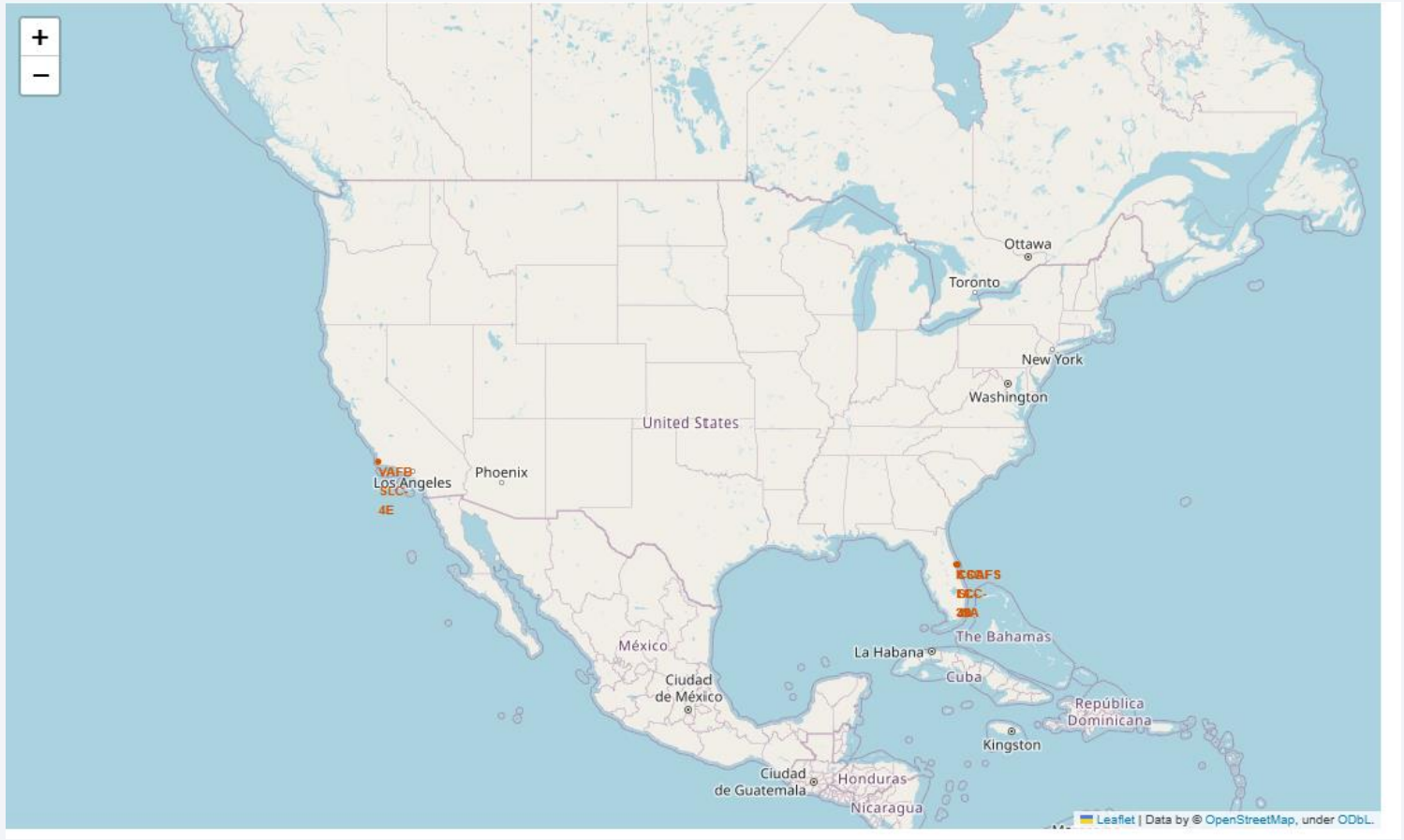
Explanation:

- 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.

Section 3

Launch Sites Proximities Analysis

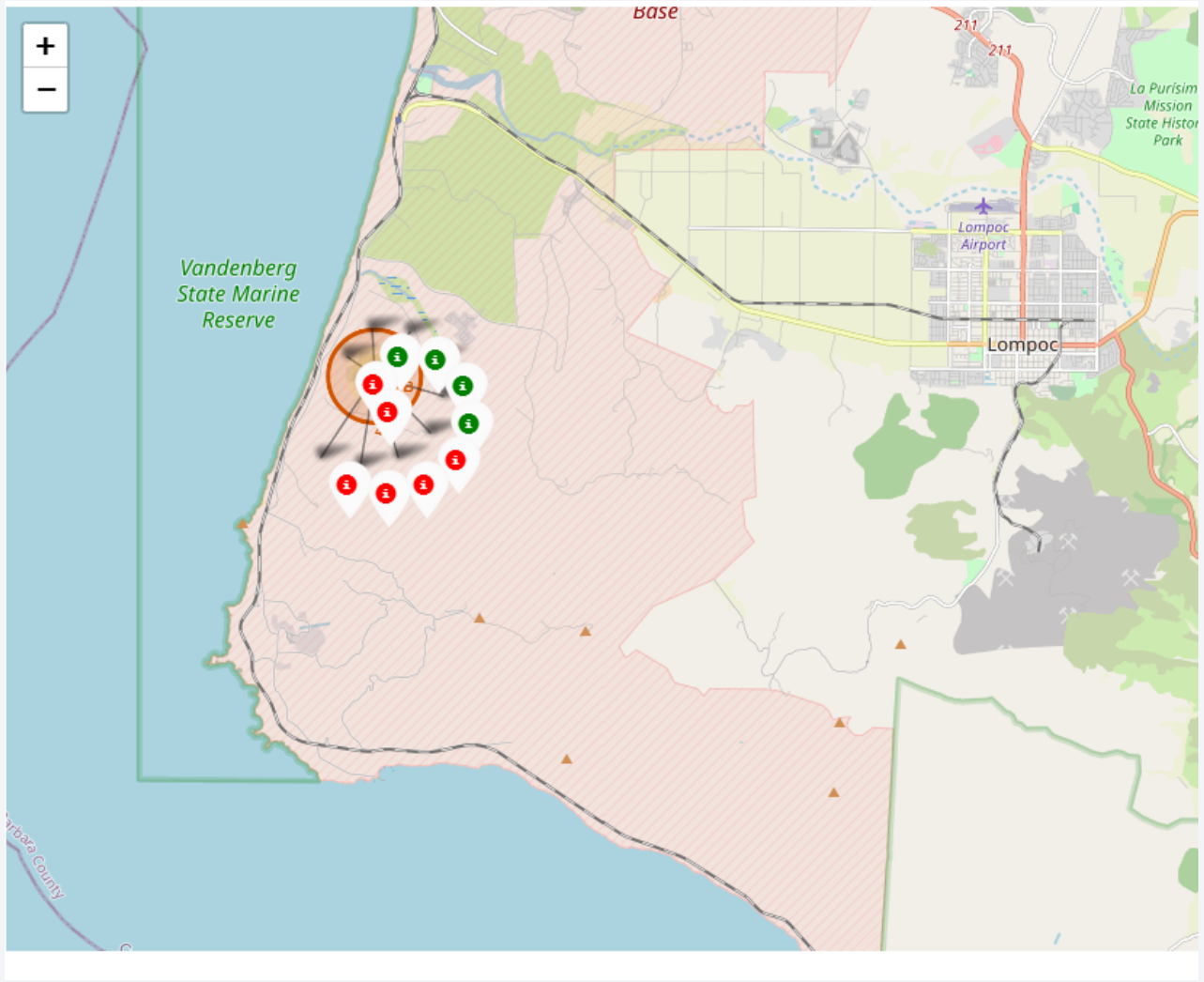
All launch sites



Explanation:

- All launch sites are in proximity to the Equator line
- All launch sites are in very close proximity to the coast

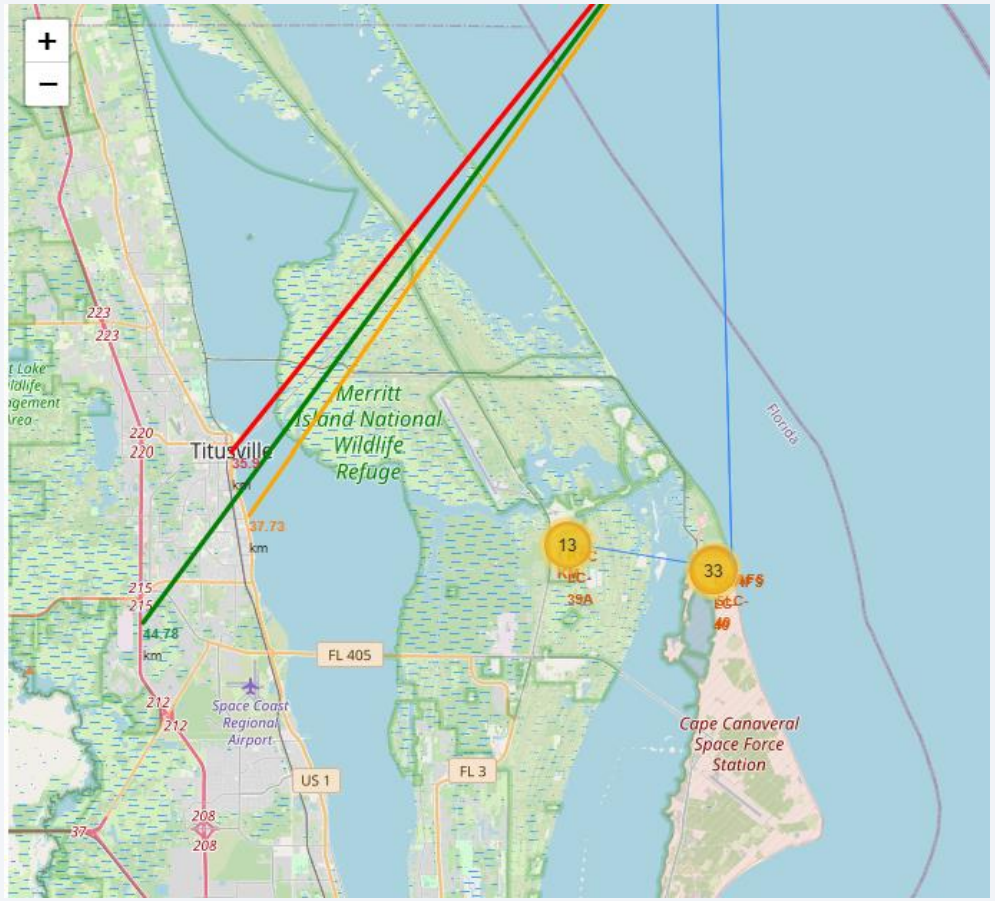
Launch Outcomes by Site



Explanation:

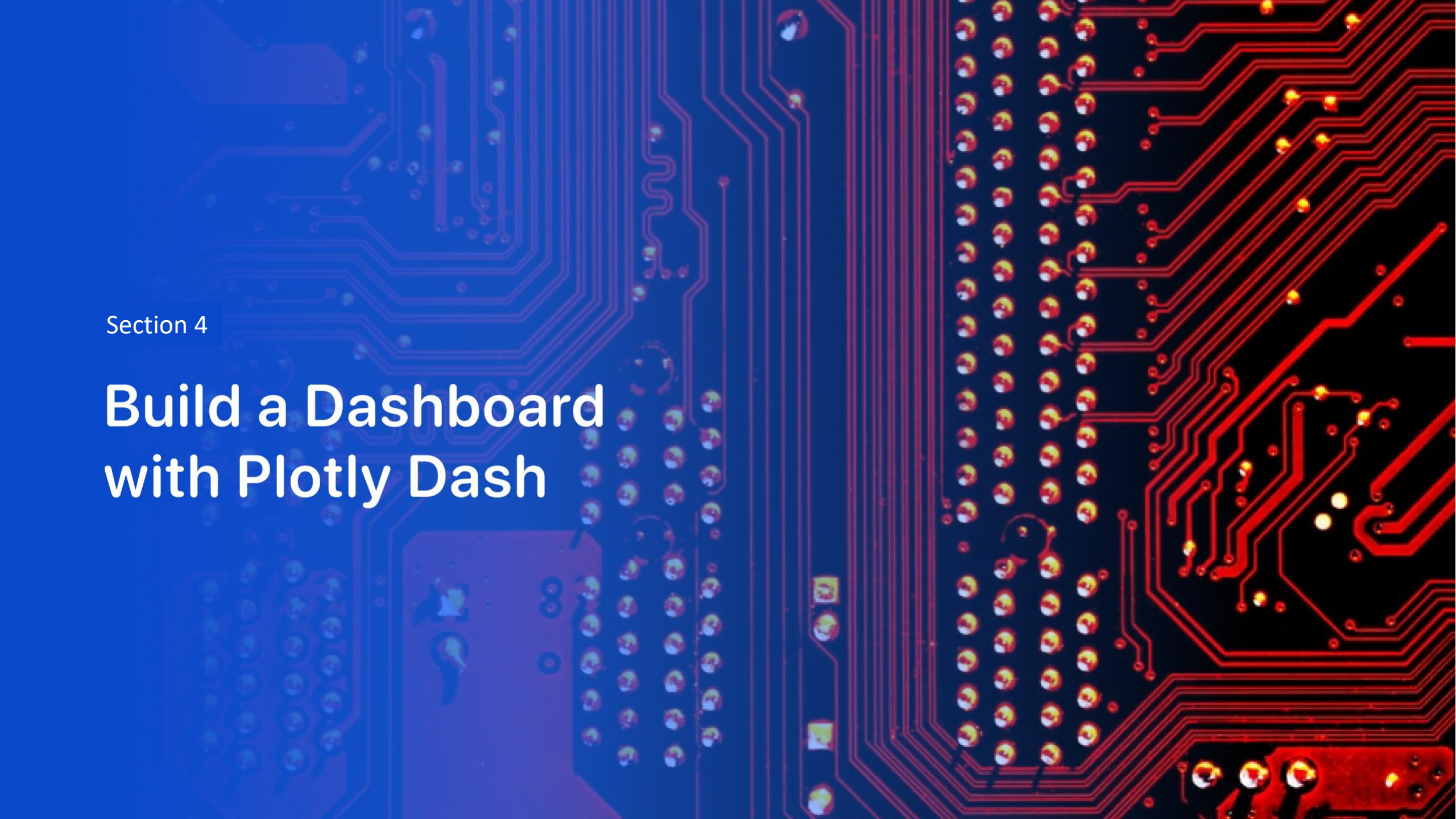
- From the color-labeled markers in marker clusters, it is easy to identify which launch sites have relatively high success rates.
  - Green Marker = Successful launch
  - Red Marker = Failed Launch

Distances between a launch site to its proximities

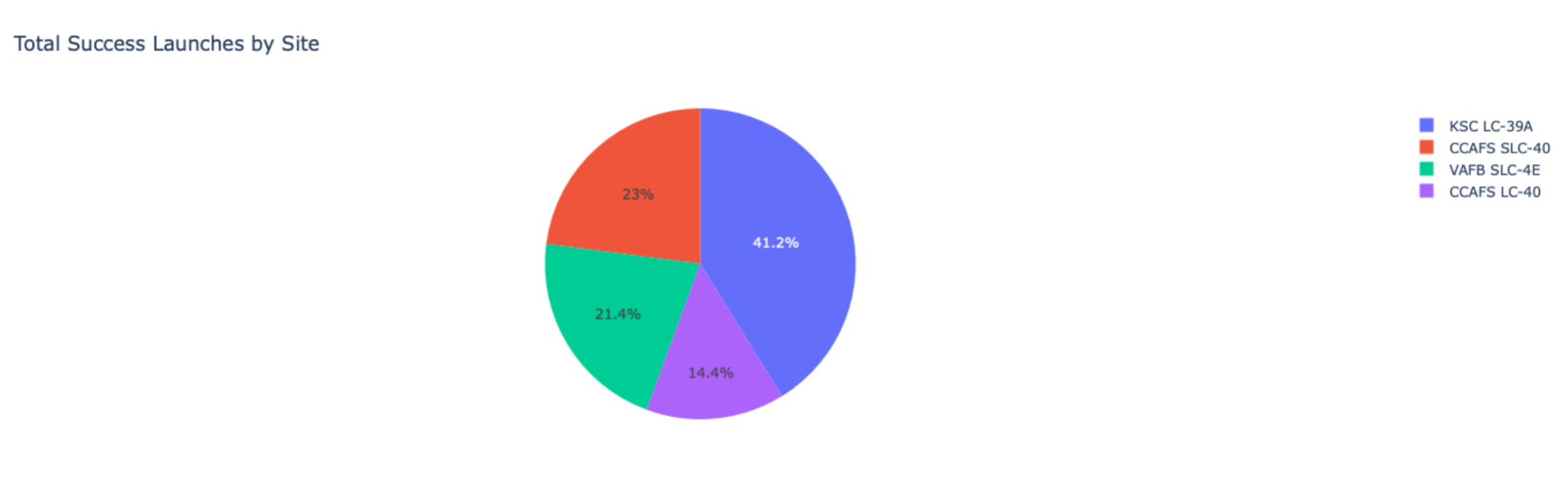


Explanation:

- Launch site KSC LC-39A is close to railway, highway and coastline
- The site is also close to Titusville


Section 4
Build a Dashboard with Plotly Dash

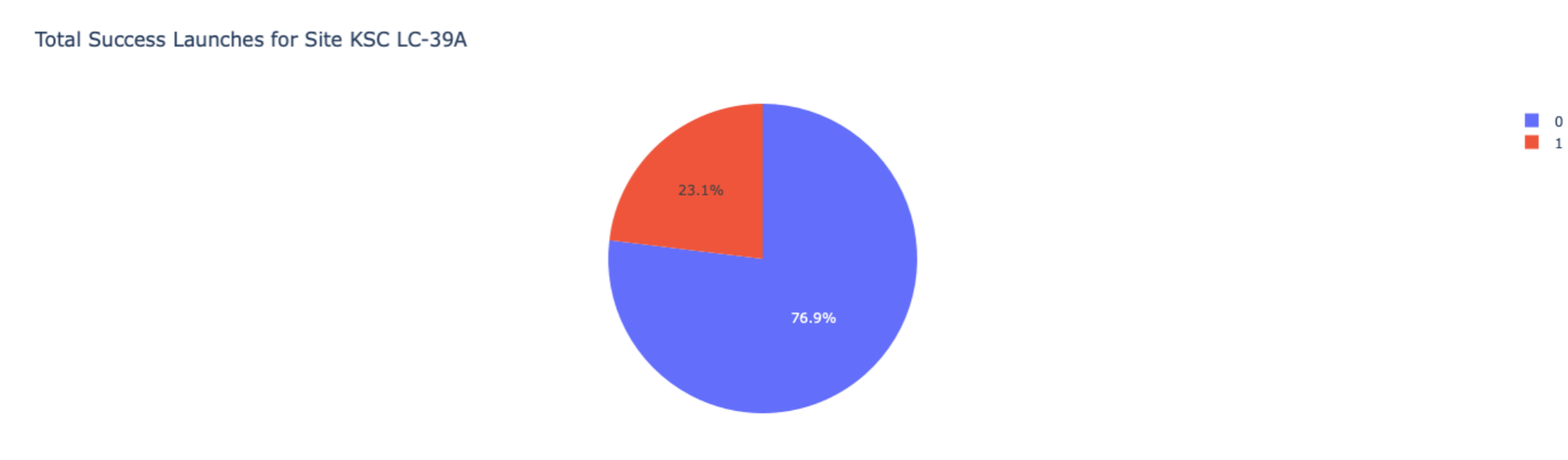
Launch success for all sites



Explanation:

- KSC LC-39A site has the most successful launches while CCAFS LC-40 has the least successful launches

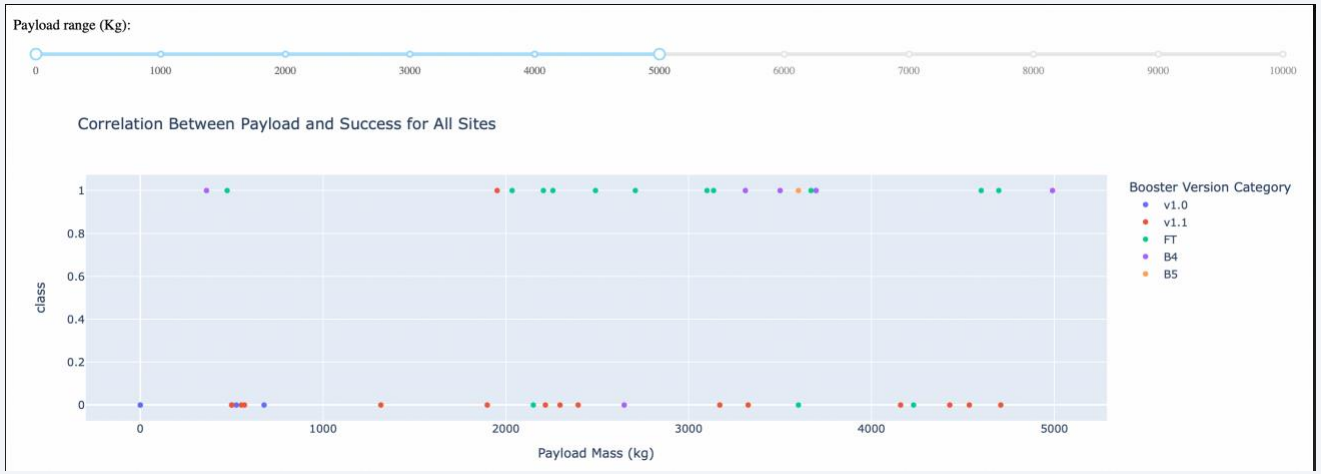
Launch Site with Highest Launch Success Ratio

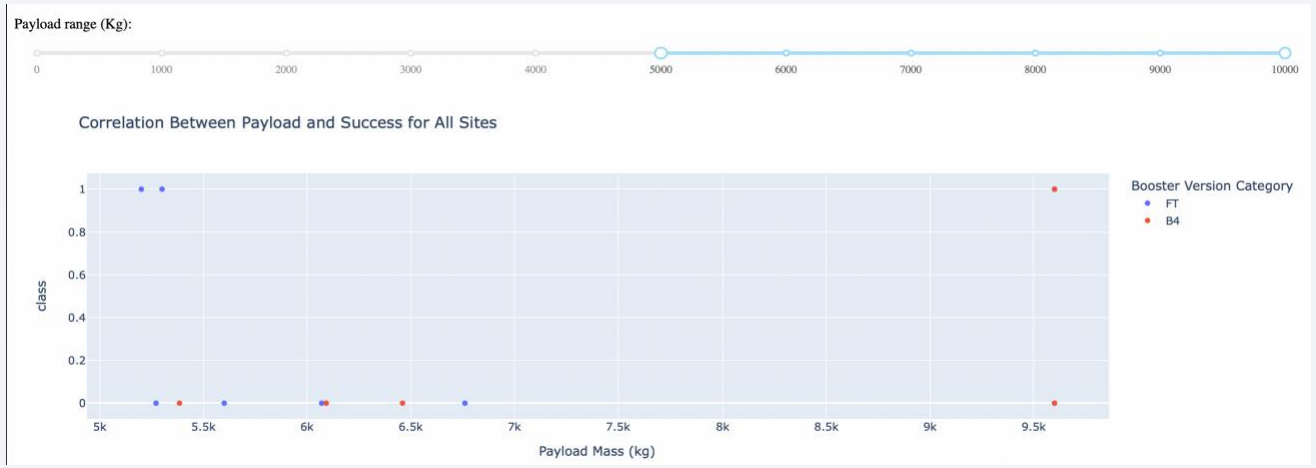


Explanation:

- KSC LC-39A site has the highest launch success ratio (76.9%).

Payload Mass vs. Launch Outcome for all sites





Explanation:

- Payloads between 2000 and 5500kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.903226	0.819444
F1_Score	0.909091	0.916031	0.949153	0.900763
Accuracy	0.866667	0.877778	0.933333	0.855556

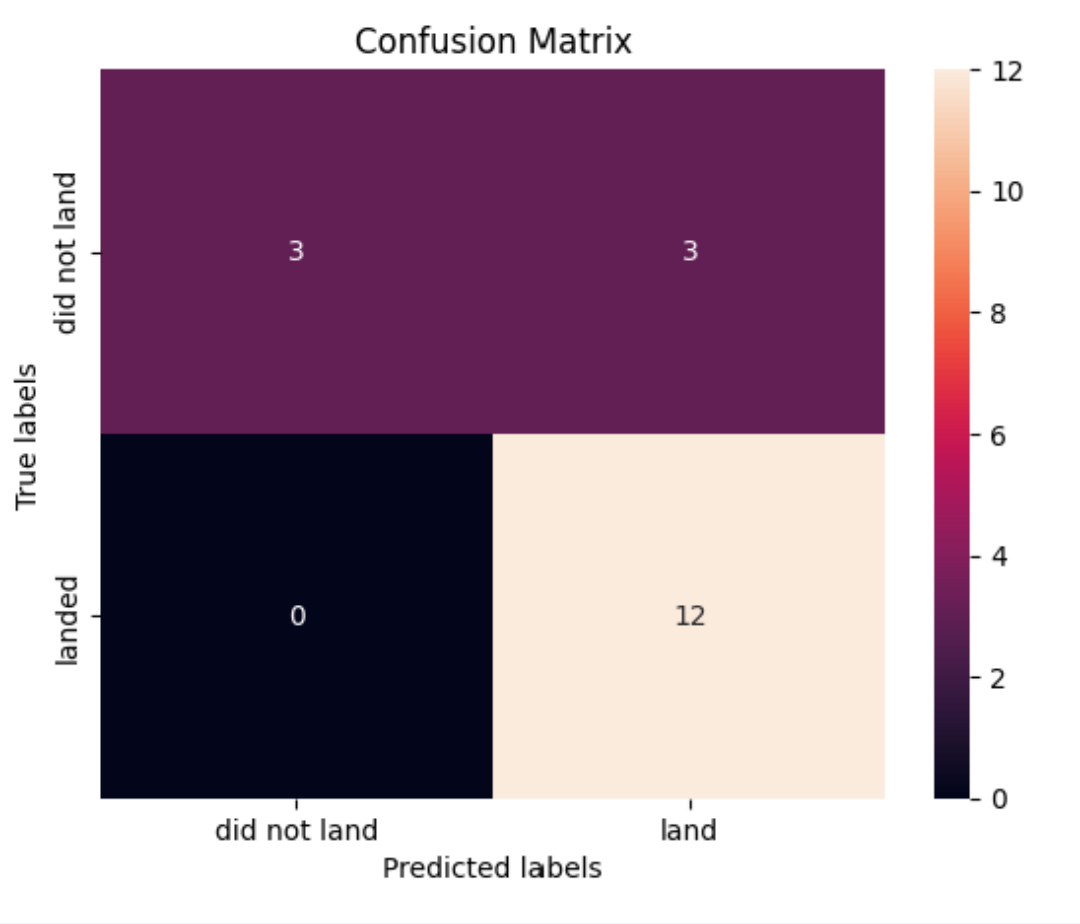
Scores and Accuracy of Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

Explanation:

- Based on the scores of the Test Set, we are unable to confirm the method performs best
- Based on the scores of Data set, Decision Tree Model performs the best

Confusion Matrix



Explanation:

- Confusion matrix of Decision Tree proves its accuracy by showing big numbers of true positive and true negative compared to the false ones.

Conclusions

- Decision Tree is the best model for this dataset
- Success rate of launches increased over time
- KSC LC-39A has the highest success rate of launches
- Launches with low payload mass show better results
- Launch sites are in proximity to the Equator line and the coast

Appendix

- Thank you to the instructors

Thank you!