Applied Data Science Capstone

Tien Chi Lin

23.12.2024

Background Image: **Media Invited to SpaceX Falcon Heavy Launch of Four NASA Missions**

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Based Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis
  - ddd
  - ddd

Introduction

- Project background and context

  SpaceX is an American aerospace manufacturer and space transportation services company with the goal of revolutionizing space technology and making space travel more affordable and accessible. In this capstone, we will predict if 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, 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 an alternate company wants to bid against SpaceX for a rocket launch.

- Questions to be answered

  - How do variables such as payload mass, orbit type, flight number, and launch site lead to the success of the first-stage landing?
  - What is the best machine learning algorithm to predict the success of the first-stage landing?

Section 1

Methodology

Methodology

- Data collection
  - Launch data using SpaceX Rest API
  - Launch records from Wikipedia using Web Scraping
- Data wrangling
  - Filtering data, dealing with missing values
  - Labeling landing outcome as 0 (failure) / 1(success)
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Feature engineering with OneHotEncoder
  - Grid search for hyperparameter tuning
  - Algorithms with Logistic regression, Support vector machine (SVM), Decision tree, and K-nearest neighbor (KNN)

Data Collection

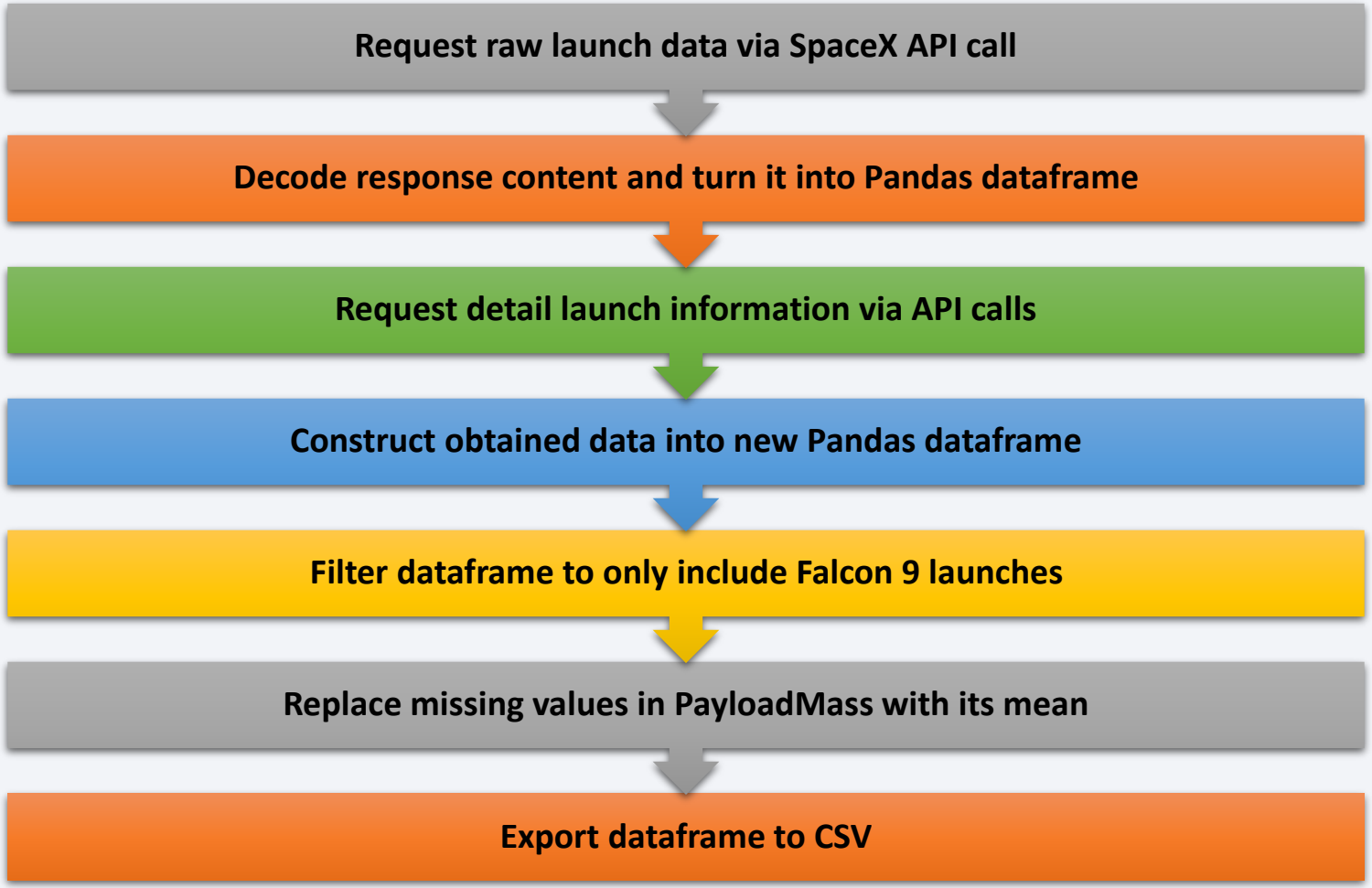
- In this stage, we collected launch data via SpaceX REST API calls and launch records via Web scraping of the SpaceX Wikipedia. Both collected data are crucial for complete launch information and further comprehensive analysis.
- Launch data via SpaceX REST API

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Launch Records via Wikipedia Web scraping

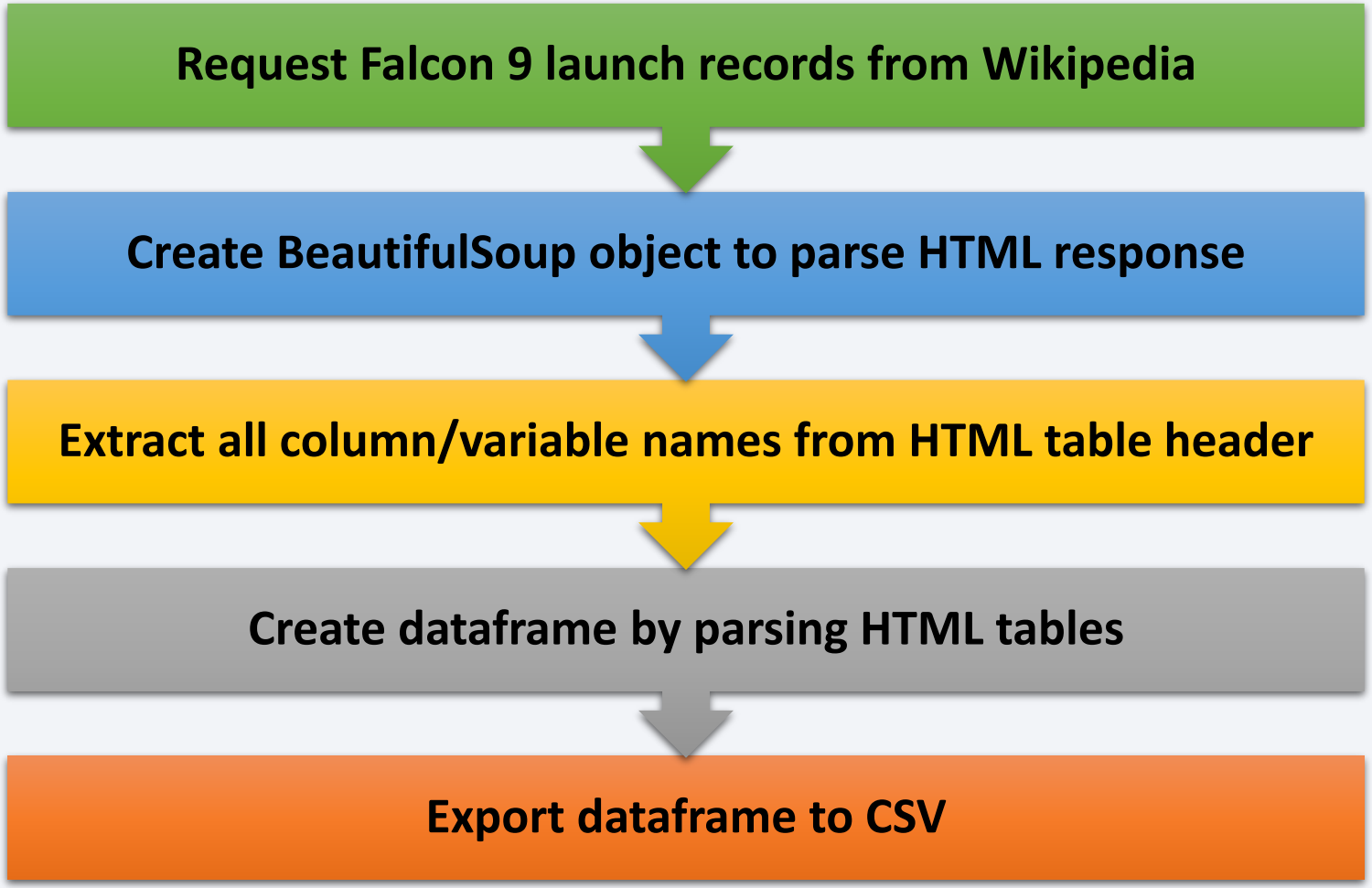
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: Data Collection API

Data Collection - Scraping



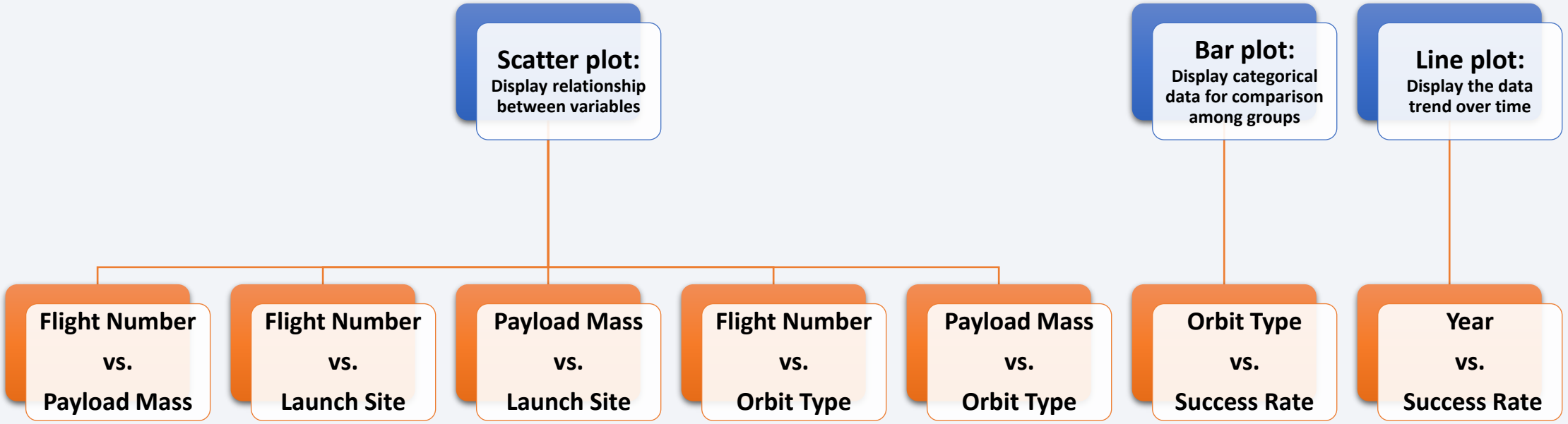
GitHub URL: Data Web Scraping

Data Wrangling

Calculate the number of launches on each site	The data contains several Space X launch facilities. Therefore, we first determine the number of launches based on the column **LaunchSite**.
Calculate the number and occurrence of each orbit	The data contains multiple orbit types for launch tasks. Therefore, we also calculate the number and occurrence of each orbit.
Calculate the number and occurence of mission outcome of the orbits	The data contains detailed outcomes in the column **Outcome**, such as **True Ocean**/ **False Ocean** representing successful/ unsuccessful landing to a specific region of the ocean. Therefore, we determine the number of outcomes based on different outcome type.
Create a landing outcome label from Outcome column	We convert those outcome labels into classification variables 1 and 0, indicating successful and unsuccessful landing, respectively.
Export dataframe to CSV	

GitHub URL: Data Wrangling

EDA with Data Visualization



GitHub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery
- 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 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

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites
  - Obtain intuitive visual insights
- Colored Markers of Launch outcomes for each Launch Site
  - Identify the difference in success rate between launch sites
- Distances between Launch Site to its proximities
  - Identify the proximities of the launch site CCAFS SLC-40 (as an example), such as coastline, railway, highway, and city
  - Obtain further insights from colored lines displaying distances between the launch site and its proximities

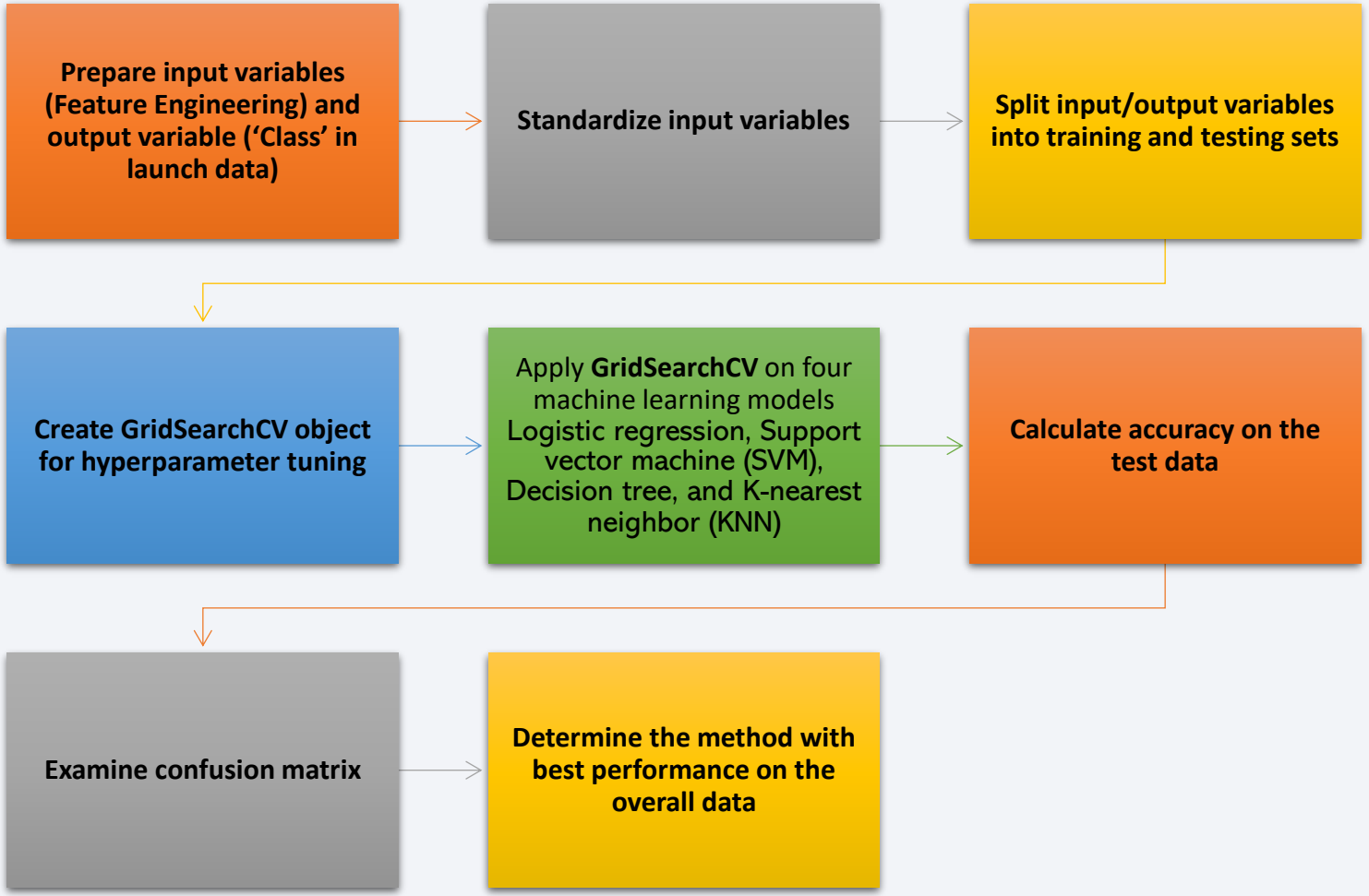
GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Launch Site Dropdown List
  - Select either all sites or a single site for analysis display
- Pie Chart of Success Launches
  - Display each site's proportion of the total successful launches
  - Display the proportion of successful and failed launches for a single site
- Slider of Payload Mass Range
  - Select the payload range to display in the scatter chart
- Scatter Chart of Payload Mass vs. Success Launches
  - Display the correlation between payload and launch success
  - Display the launch success color-labeled by Booster version

GitHub URL: SpaceX Dash App

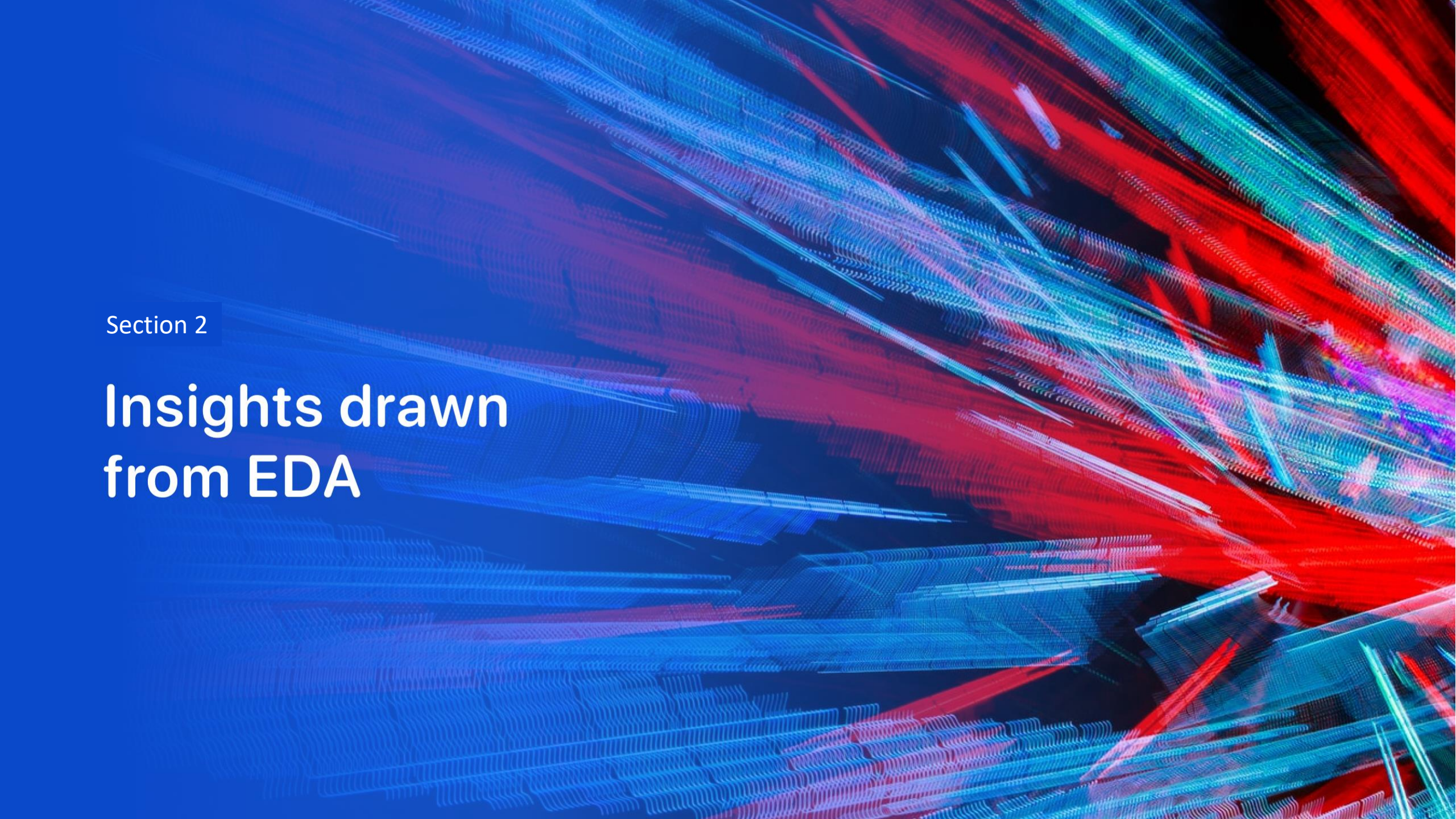
Predictive Analysis (Classification)



GitHub URL: Machine Learning Prediction

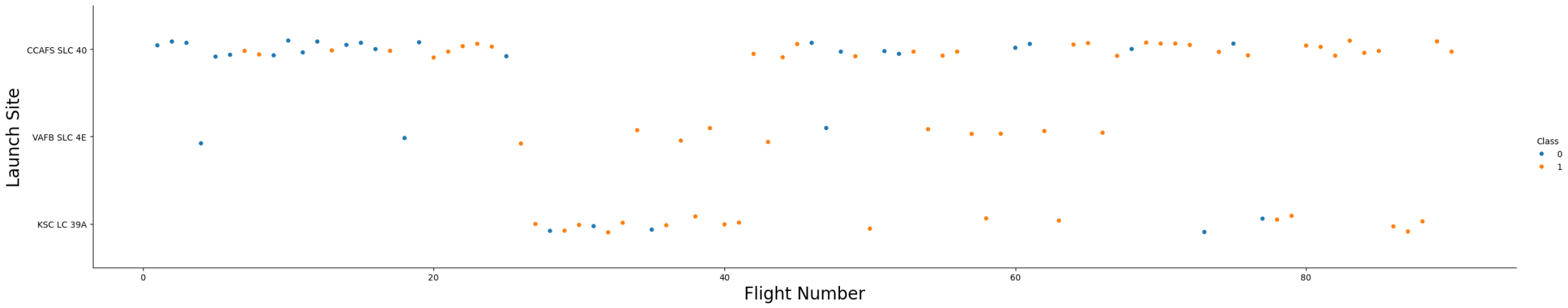
Results

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


Section 2
Insights drawn from EDA

EDA with Visualization

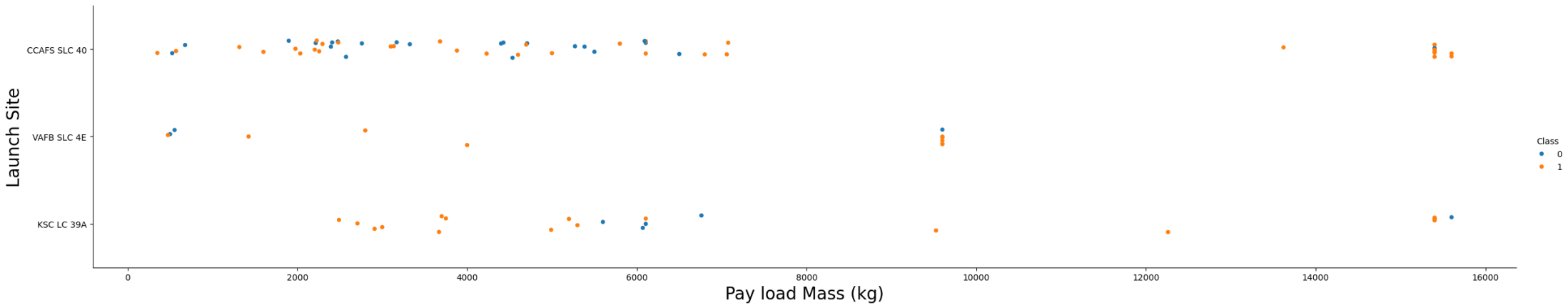
Flight Number vs. Launch Site



Explanation

- The earliest flights tended to fail, while the latest ones tended to succeed, indicating an increase in success rate over time.
- Most of the failures occurred at the CCAFS SLC 40 launch site, resulting in a lower success rate than the other two launch sites.

Payload vs. Launch Site



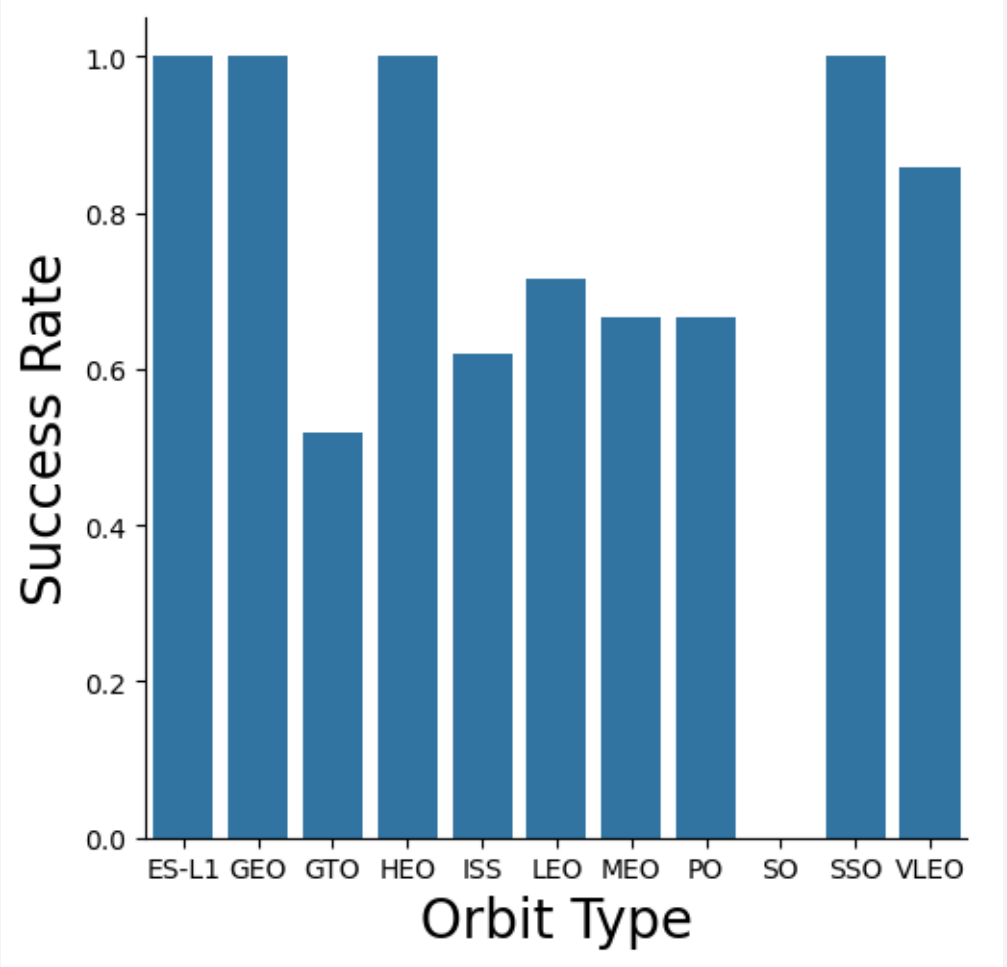
Explanation

- Most launches with payload mass over 8000 kg were successful.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

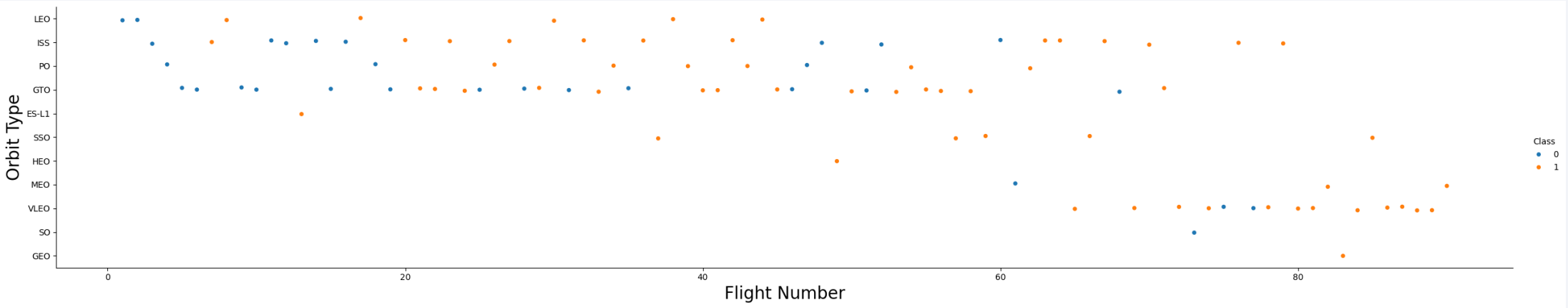
Success Rate vs. Orbit Type

Explanation

- Orbits ES-L1, GEO, HEO, SSO have the highest success rate, 100%.
- Orbit SO has the lowest success rate, 0%.
- The success rate is above 60% for most orbit types, except orbit GEO which is 50% and SO which is 0%.



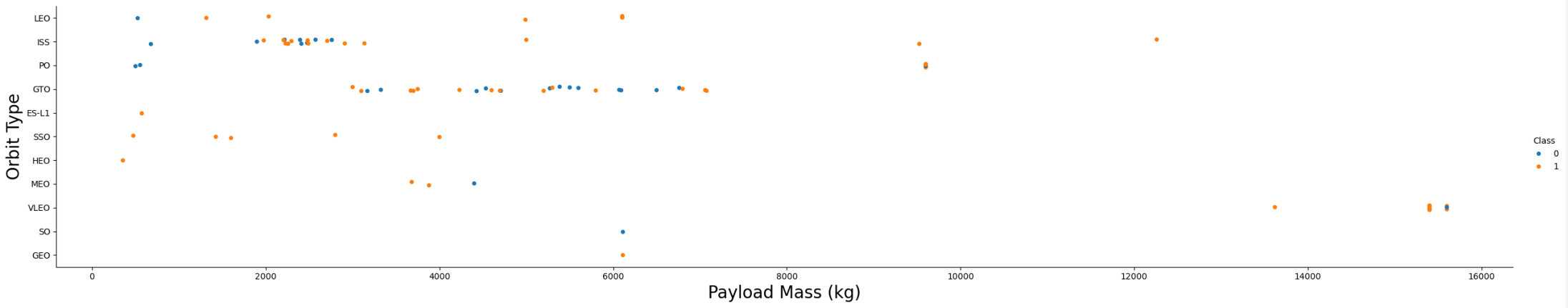
Flight Number vs. Orbit Type



Explanation

- In the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



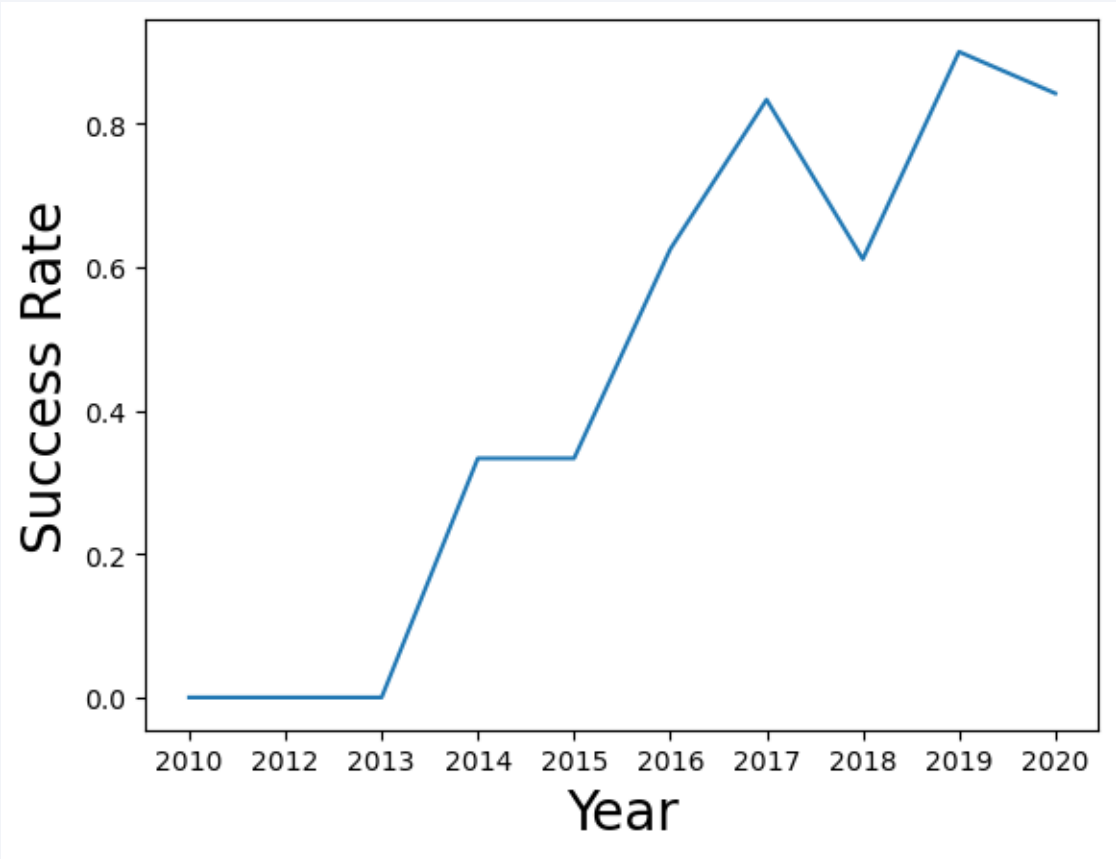
Explanation

- For LEO and ISS, there is positive relationship between payload mass and launch success.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.
- For SSO, there is more than one success, no failures, and only in the low payload range.

Launch Success Yearly Trend

Explanation

- The success rate since 2013 kept increasing till 2020



EDA with SQL

All Launch Site Names

```
%sql Select distinct launch_site From SPACEXTABLE;
```

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

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

Explanation

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

Launch Site Names Begin with 'CCA'

```
%sql Select * From SPACEXTABLE 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_Mass_Kg From SPACEXTABLE Where Customer = 'NASA (CRS)';
```

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

Total_Payload_Mass_Kg
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 Average_Payload_Mass_Kg From SPACEXTABLE Where Booster_version like 'F9 v1.1%';
```

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

Average_Payload_Mass_Kg
2534.6666666666665

Explanation

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

First Successful Ground Landing Date

```
%sql Select Min(Date) as First_Successful_Landing_Date_Ground_Pad From SPACEXTABLE Where Landing_outcome = 'Success (ground pad)'
```

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

First_Successful_Landing_Date_Ground_Pad
2015-12-22

Explanation

- Display 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 Booster_version From SPACEXTABLE Where Landing_outcome = 'Success (drone ship)' and Payload_mass__kg_ Between 4000 and 6000;
```

```
 * 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 distinct Mission_outcome From SPACEXTABLE
```

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

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

```
%%sql Select * From (Select count(*) as Number_Successful_Mission from SPACEXTABLE Where Mission_outcome like 'Success%') as Success,
    (Select count(*) as Number_Failure_Mission from SPACEXTABLE Where Mission_outcome like 'Failure%') as Failure
```

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

Number_Successful_Mission	Number_Failure_Mission
100	1

Explanation

- There are two ‘Success’ when displaying the distinct names of mission outcomes
- List the total number of successful (regardless of details) and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql Select Booster_version From SPACEXTABLE Where Payload_mass__kg_ = (Select Max(Payload_mass__kg_) From SPACEXTABLE);
 * 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

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 SPACEXTABLE
    WHERE substr(Date,0,5) = '2015' and Landing_outcome = 'Failure (drone ship)';
```

```
 * 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 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

```
%%sql Select Landing_outcome, count(*) as Number_outcome From SPACEXTABLE
    Where Date Between '2010-06-04' and '2017-03-20'
    Group by Landing_outcome
    Order by Number_outcome DESC;
```

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

Done.
```

Landing_Outcome	Number_outcome
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.

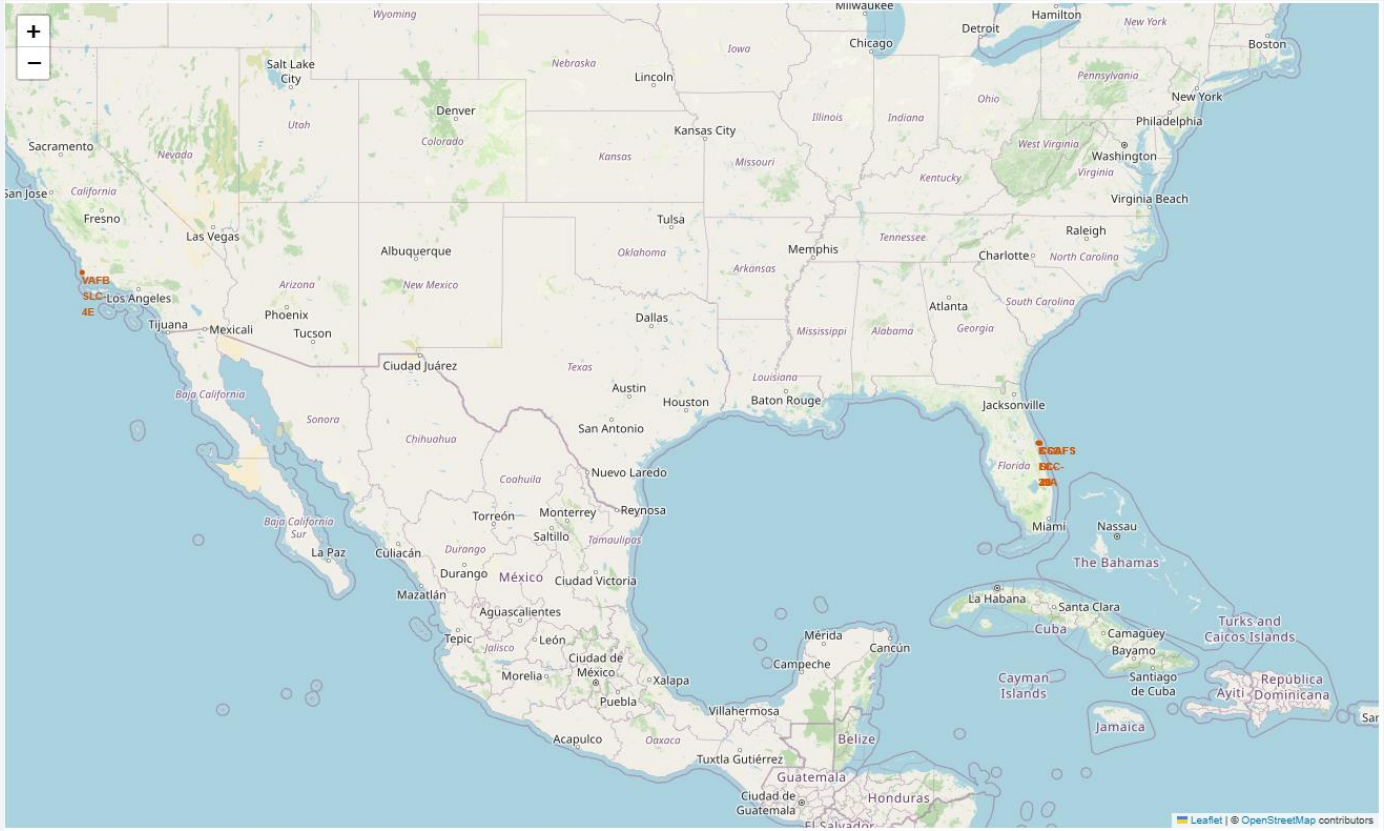
Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

Explanation

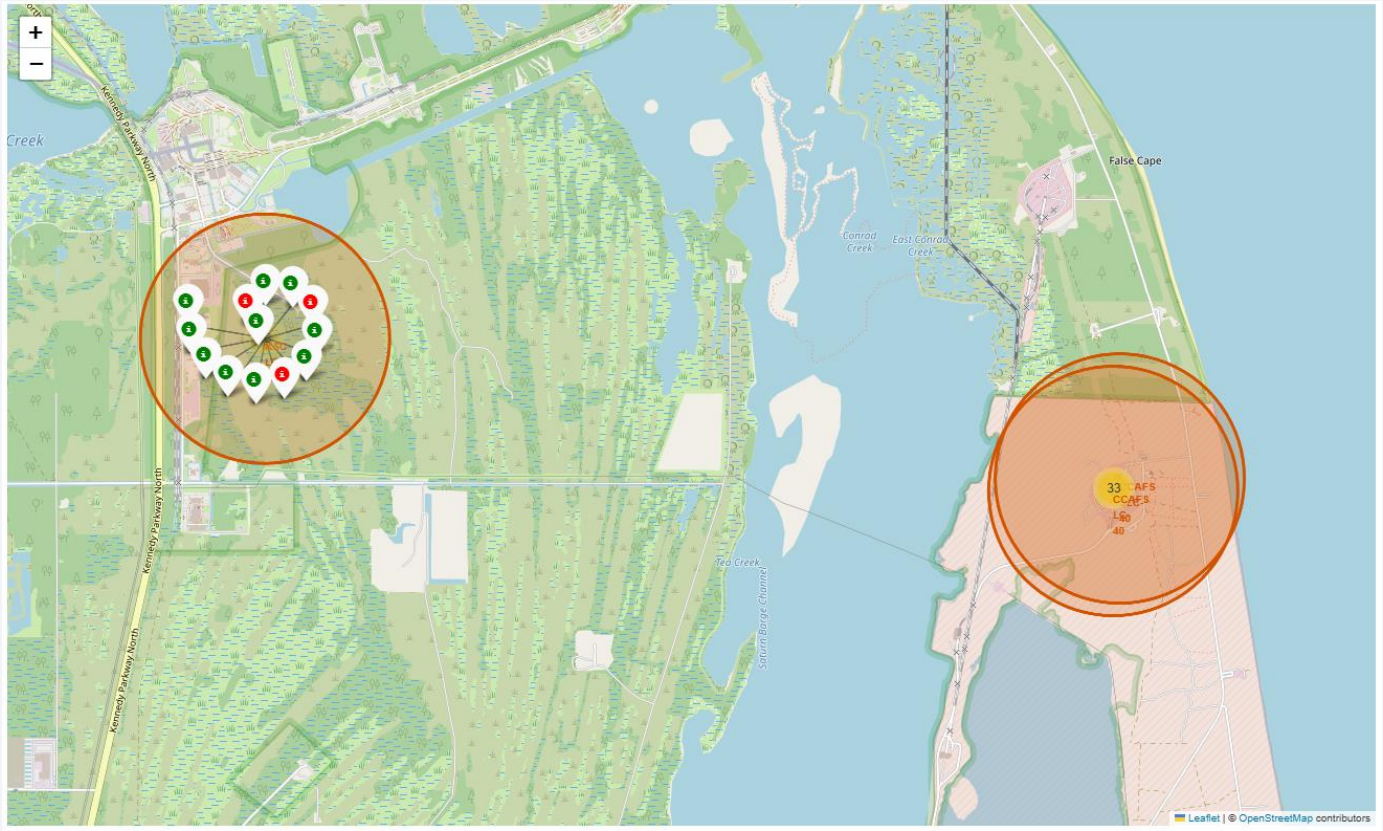
- All launch sites are located as close to the equator as possible within the United States, since they can take advantage of the Earth's substantial rotational speed.
- All launch sites are also in very close proximity to the coast, since we can ensure that no components are shed over populated areas.



Color-labeled launch records on the map

Explanation

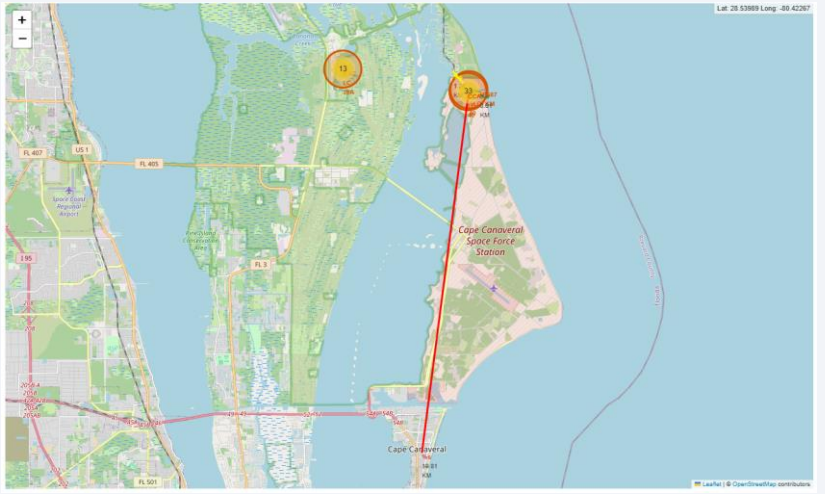
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Launch Success
  - Red Marker = Launch Failure
- Launch Site KSC LC-39A has a very high Success Rate.

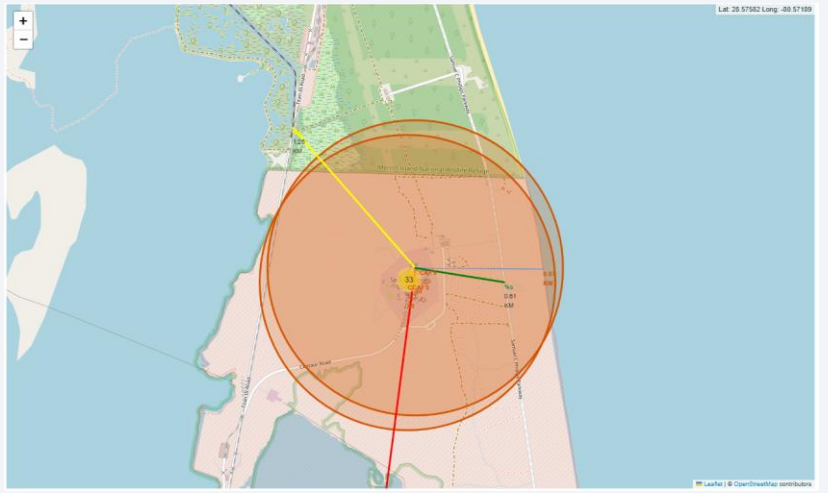


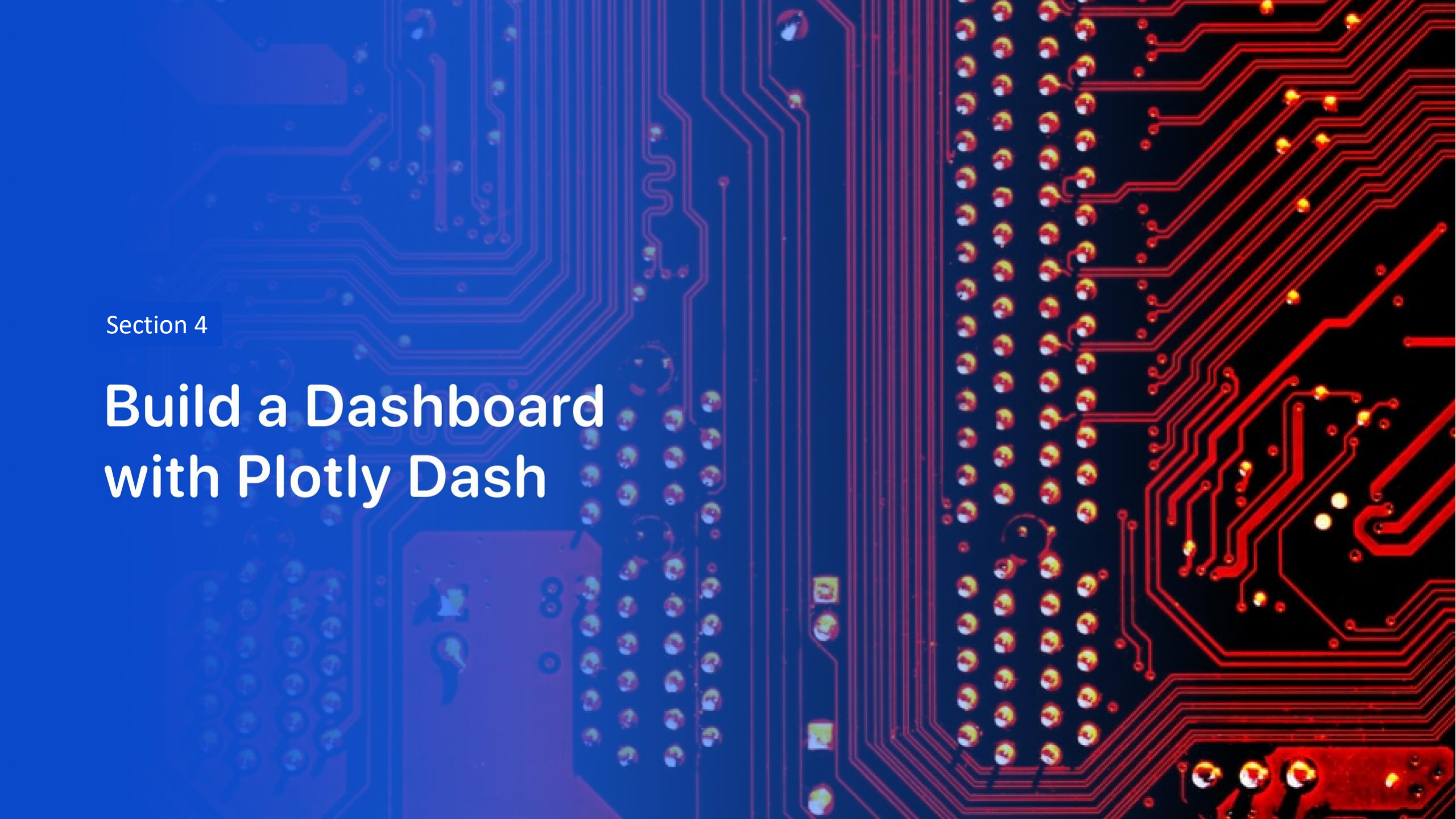
Distance from the launch site CCAFS SLC-40 to its proximities

Explanation

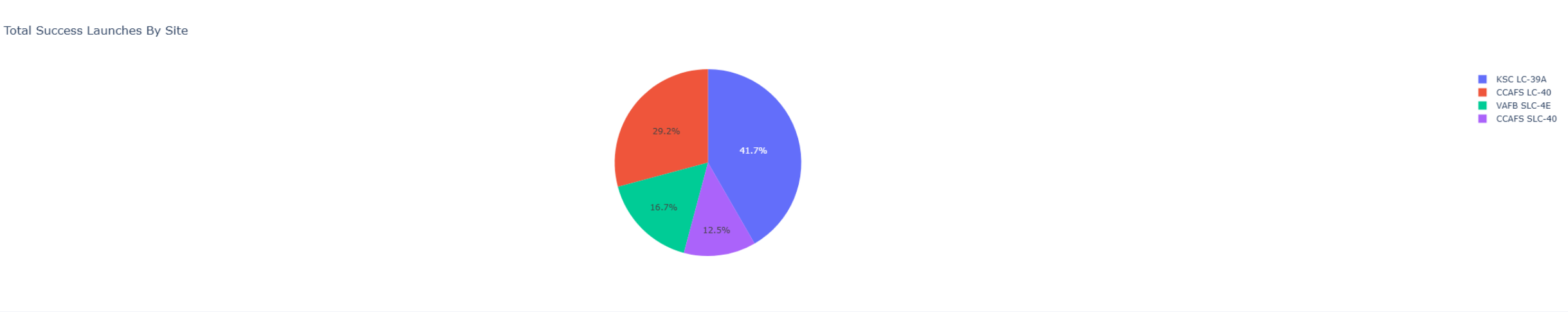
- For the launch site CCAFS SLC-40, we can observe that:
  - Relatively close to railway (yellow line, 1.26 km)
  - Relatively close to highway (green line, 0.61 km)
  - Relatively close to coastline (blue line, 0.87 km)
- The distance to the closest city (red line) is 19.81 km, indicating that launch site is relatively far from the city.






Section 4
Build a Dashboard with Plotly Dash

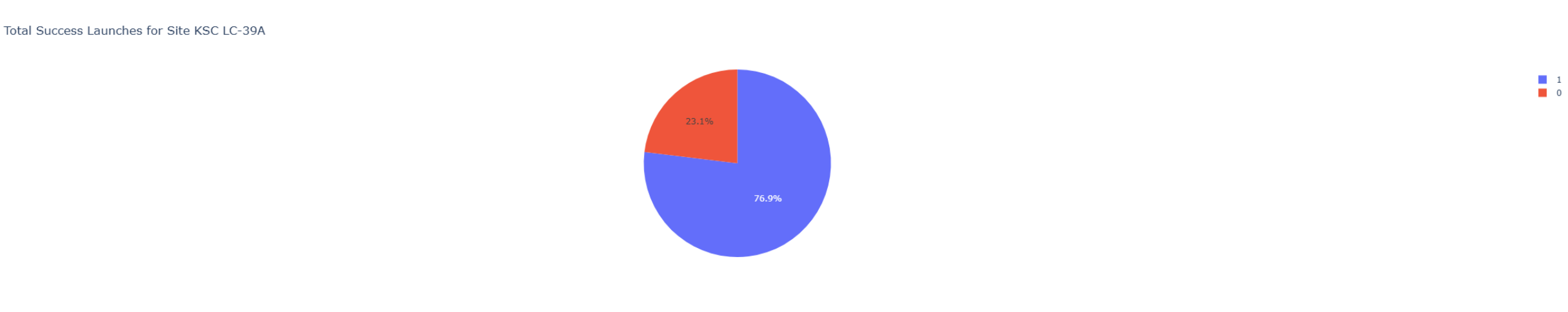
Launch success count for all sites



Explanation

- Among all sites, KSC LC-39A has the most successful launches

Launch site with highest launch success rate



Explanation

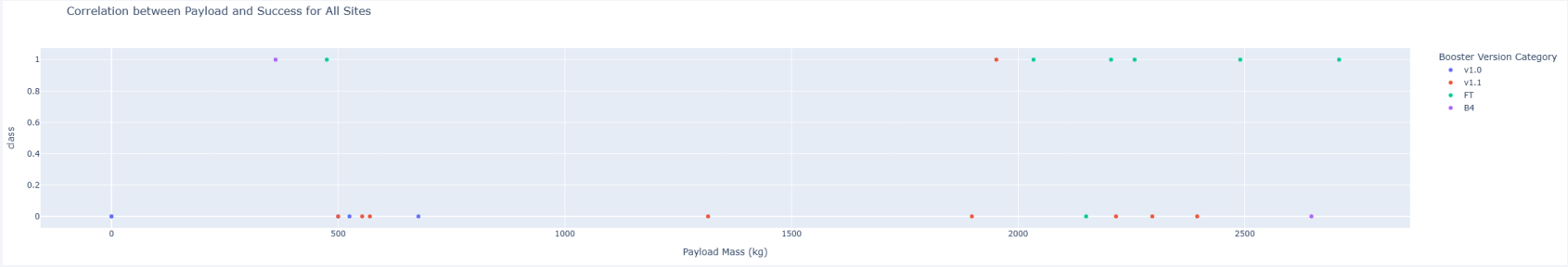
- Among all sites, KSC LC-39A also has the highest launch success rate (76.9%)

Payload Mass vs. Launch Outcome for all sites

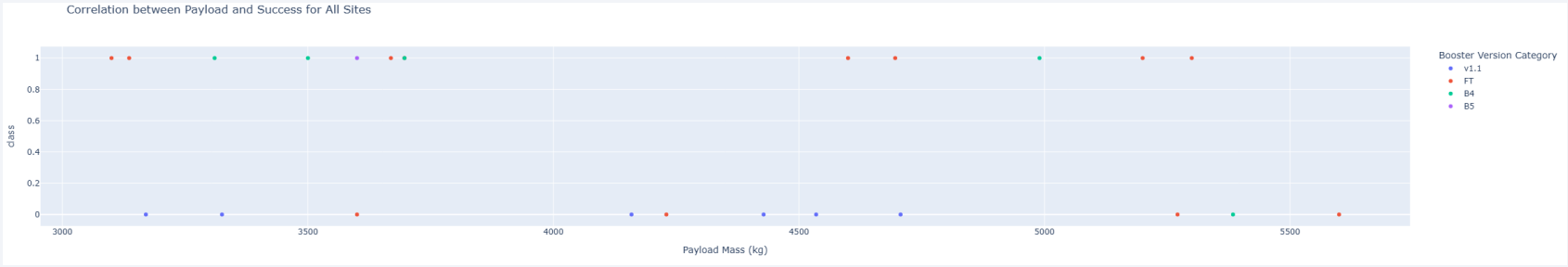
Explanation

- The payload mass range with the highest success rate is 3000~5000 kg

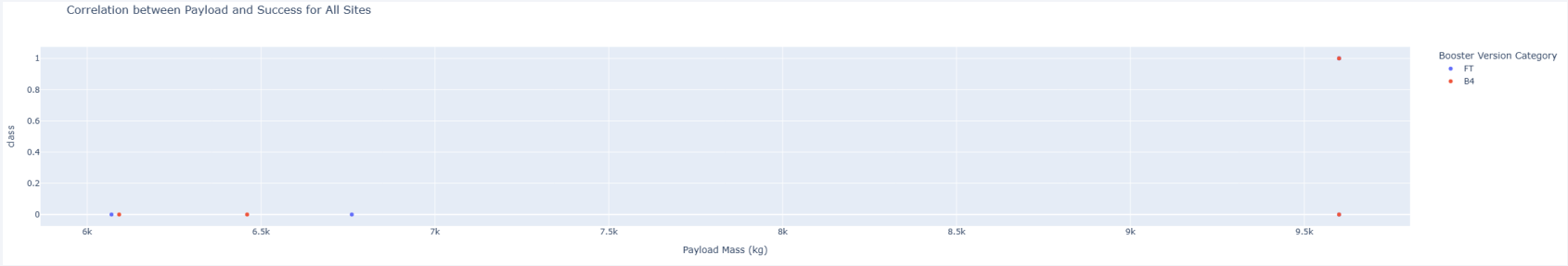
Payload Mass Range: 0~3000 kg



Payload Mass Range: 3000~6000 kg



Payload Mass Range: 6000~10000 kg



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation

- All four methods show the same performance on the testing data. The reason could be the small sample size (18).
- For performance on the entire data set, the SVM model has slightly higher metrics than the other three models, indicating the SVM method performs the best.

Performance Metrics on Test Data 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

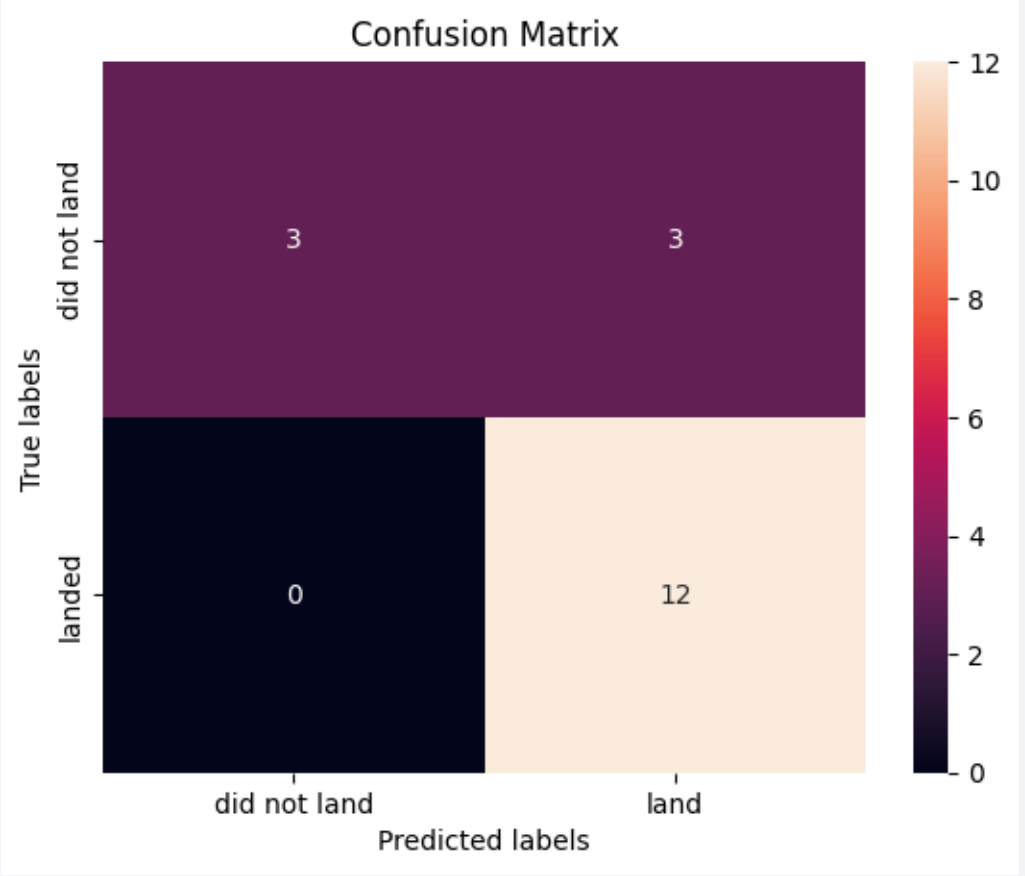
Performance Metrics on Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.826087	0.819444
F1_Score	0.909091	0.916031	0.904762	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

Confusion Matrix for SVM Model

Explanation

- SVM Model has good overall performance
- SVM Model has difficulty distinguishing false positive (predicted label = land, true label = did not land) cases



Conclusions

- Most launches with payload mass over 8000 kg were successful.
- Orbits ES-L1, GEO, HEO, SSO have the highest success rate, 100%, while orbit SO has the lowest success rate, 0%.
- For Orbit LEO and ISS, there is positive relationship between payload mass and launch success.
- The success rate since 2013 kept increasing till 2020.
- All launch sites are in very close proximity to the equator as well as the coast.
- Launch Site KSC LC-39A has the most successful launches as well as the highest launch success rate.
- SVM model is the best algorithm for the launch success classification.

Appendix

- GitHub/TienChiLin/Applied-Data-Science_Capstone
- Applied Data Science Capstone Course

Thank you!