IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zulfikhar Ali
6/12/2023

Outline

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

Executive Summary

Summary of methodologies

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:

- **Collect** data using SpaceX REST API and web scraping techniques.
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors:
  - payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics:
  - total payload, payload range for successful launches, and total number of successful and failed outcome
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN)

Results

Exploratory Data Analysis

- Launch success improved over time
- KSC LC 39A has highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rate

Visualization/Analytics

- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics

- All models performed similarly on the test set.
- Decision tree model slightly outperformed

Introduction

Background

- **Objective**: Predicting successful landing of Falcon 9 first stage
- **Cost Estimation Importance**: Falcon 9 launch cost ($62 million) significantly lower than competitors' ($165 million)
- **Cost advantage**: SpaceX's reuse of first stage enables cost reduction and competitive pricing
- **Relevance**: Determining landing success for estimating launch cost in bidding against SpaceX

Explore

- Factors influencing first stage landing success:
  - Payload mass
  - Launch site
  - Number of flights
  - Orbits
- Analysis of the rate of successful landings over time
- Identification of the best predictive model for Falcon 9 first stage successful landing through binary classification

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data using SpaceX REST API and Web Scraping Techniques
- Perform data wrangling
  - Filter data, handle missing values and apply one-hot encoding to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build classification models and tune to find best model and parameters

Data Collection

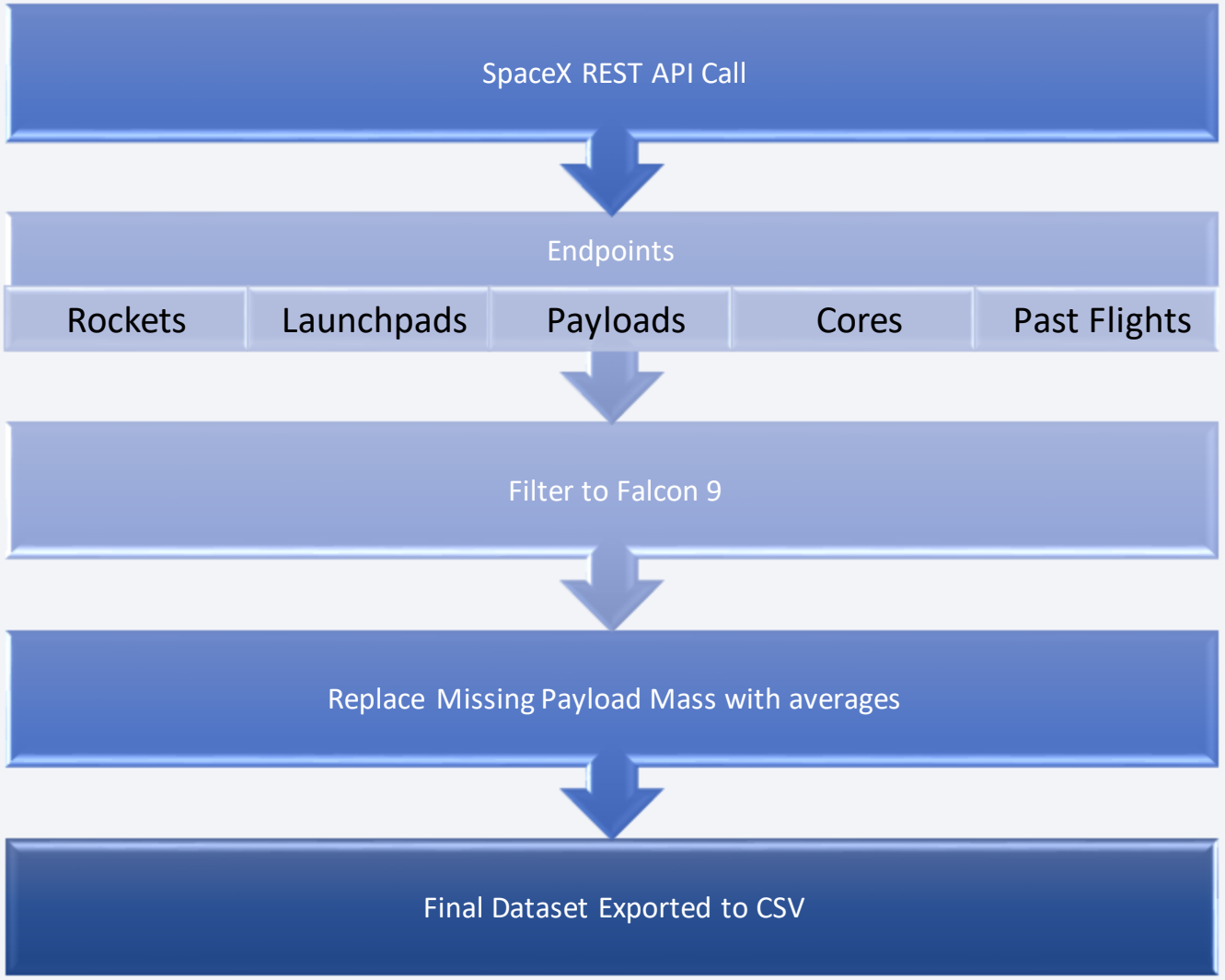
SpaceX REST API

- Data collected from SpaceX REST API: https://api.spacexdata.com/v4
- Data includes information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome

Web Scraping

- Data collected using Beautiful Soup Python Package to parse the Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches"
- Information appended to data set using rocket ID and then normalized to csv

Data Collection – SpaceX API



GitHub Notebook Link

Data Collection - Scraping

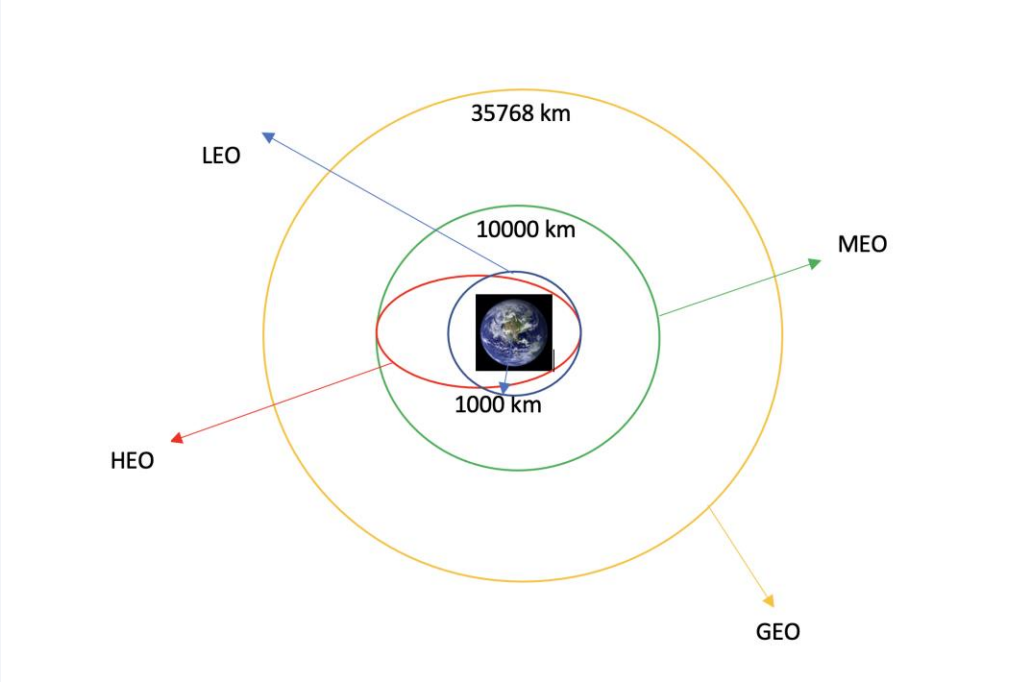
SpaceX Wikipedia Page

Extract following html tables

Dates	Booster Versions	Landing Status	Payload Mass	Launch Site	Orbits	Flight Number

Parse Tables into DataFrame

DataFrame Exported to CSV

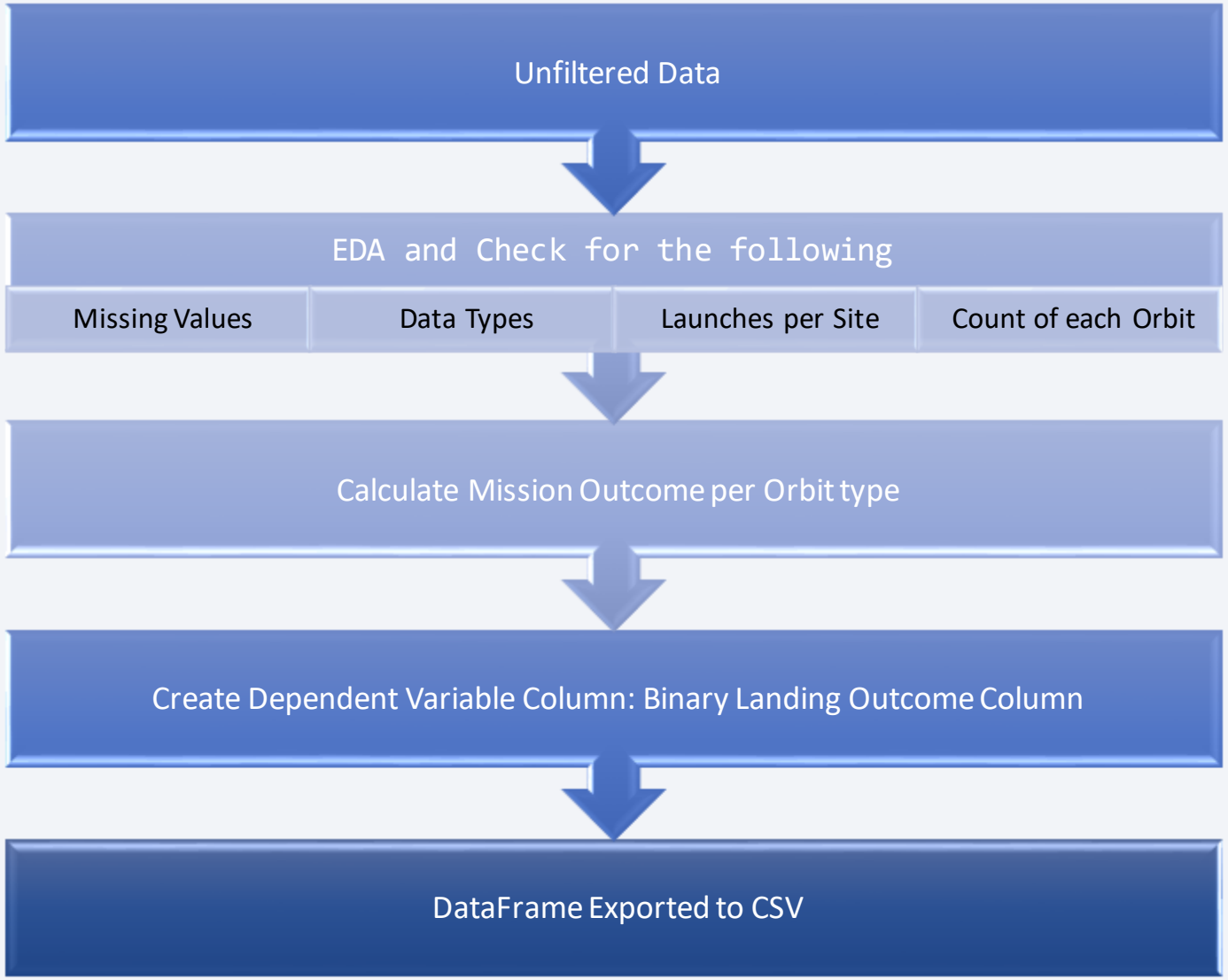


GitHub Notebook Link

Data Wrangling

- Process of gathering, cleaning, transforming, and organizing raw data into a structured and usable format for analysis and modeling

GitHub Notebook Link

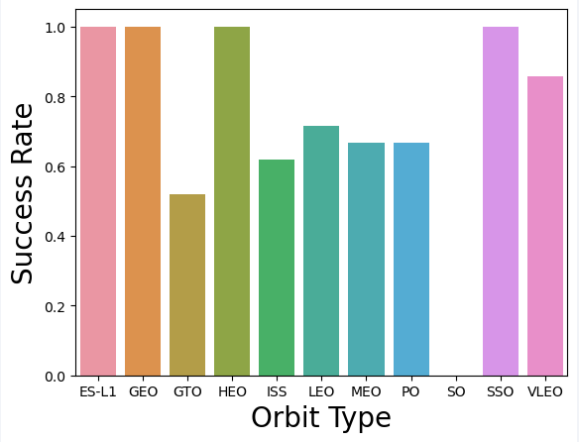


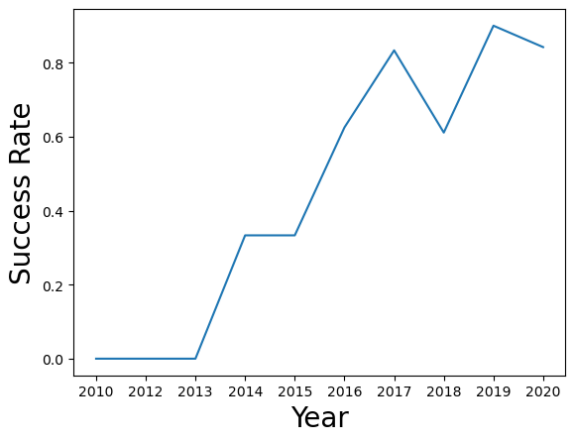
EDA with Data Visualization

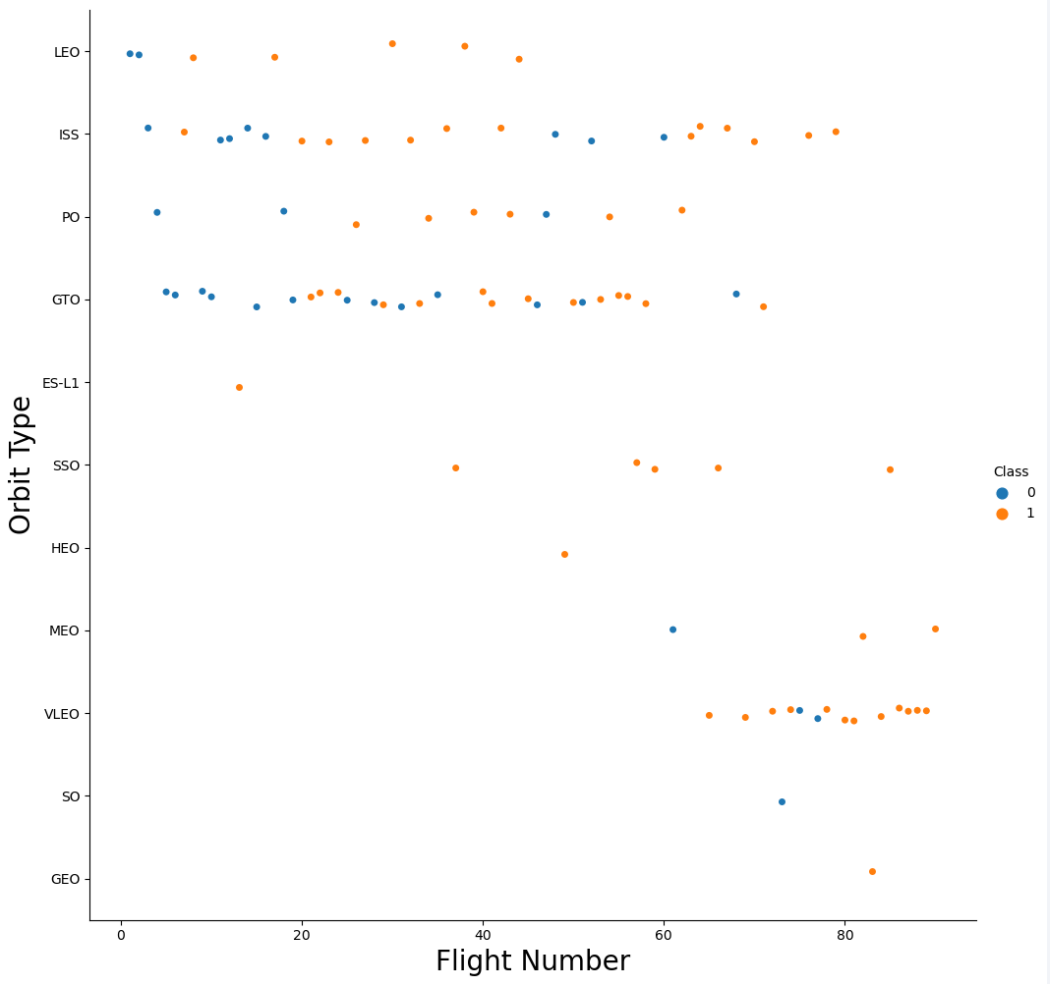
Relationship between the following attributes:

- Flight Number vs.
  - Payload Mass (kg), Launch Site
- Payload Mass (kg) vs.
  - Launch Site, Orbit type
- Success Rate and Yearly Success Rate Trend

GitHub Notebook Link







EDA with SQL

Performed following queries for better understanding of dataset:

- Records where launch sites begin with the string ‘CCA’
- Total payload mass carried by booster launched by NASA (CRS)
  - 45596.0 kg
- Booster Version F9 v1.1 average payload mass carried
  - 2928.4 kg
- Date first successful landing outcome (ground pad)
- Boosters with successful landing in drone ship with payload mass between 4000 – 6000 kg
- Total Successful/Failure mission
  - 100/1 ratio
- Booster Versions that carried the maximum payload mass
- 2015's failed Landing Outcomes in drone ship, their booster versions, and launch sites names
- Rank the count of landing outcomes or success between 2010-06-04 and 2017-03-20
  - Drone ship had the most (and same number of) successful/failed landing outcomes

Github Notebook Link

Build an Interactive Map with Folium

- To visualize the launch data into an interactive map. We took the latitude and longitude coordinates at each launch site and added a circle marker around each launch site with a label of the name of the launch site.

- The dataframes launch_outcomes (failure, success) were assigned to classes 0 and 1 with Red and Green markers on the map in MarkerCluster()

- Haversine's formula to calculate the distance of the launch sites to various landmarks to find the answers to the questions:
  - How close are the launch sites to railways, highways, and coastlines?
  - How close are the launch sites to nearby cities?

Github Notebook Link

Build a Dashboard with Plotly Dash

We built an interactive dashboard with Plotly Dash, which allows the user to play around with the data as they need.

Dropdown List with Launch Sites

- Allow users to select all launch sites or a certain launch site Dashboard with Plotly Dash.

Pie Chart Showing Successful Launches

- Allow users to see successful and unsuccessful launches as a percentage of the total.

Payload Mass Range Slider

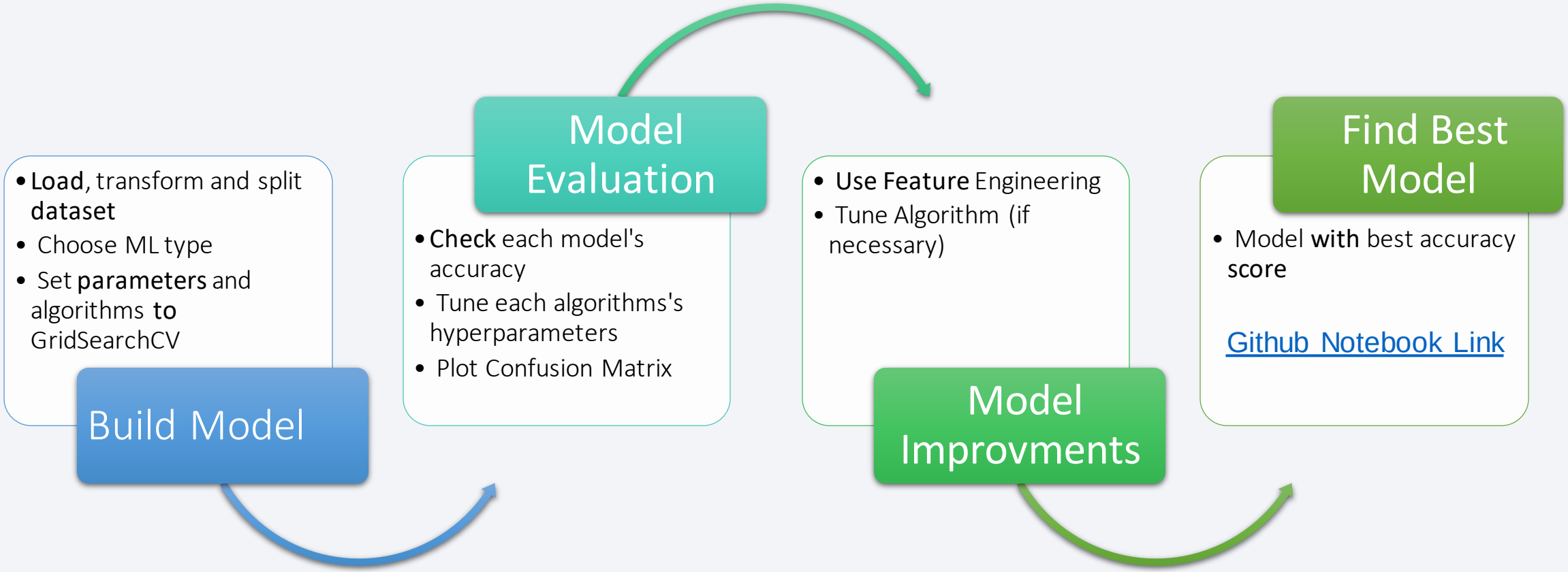
- Allow user to select payload mass range.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow users to see the correlation between Payload and Launch Success.

Github Notebook Link

Predictive Analysis (Classification)



Results

3 main Categories:

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC39A has the highest success rate among landing sites
- Orbits ESL1, GEO, HEO, and SSO have a 100% success rate

Interactive/Visual Analytics

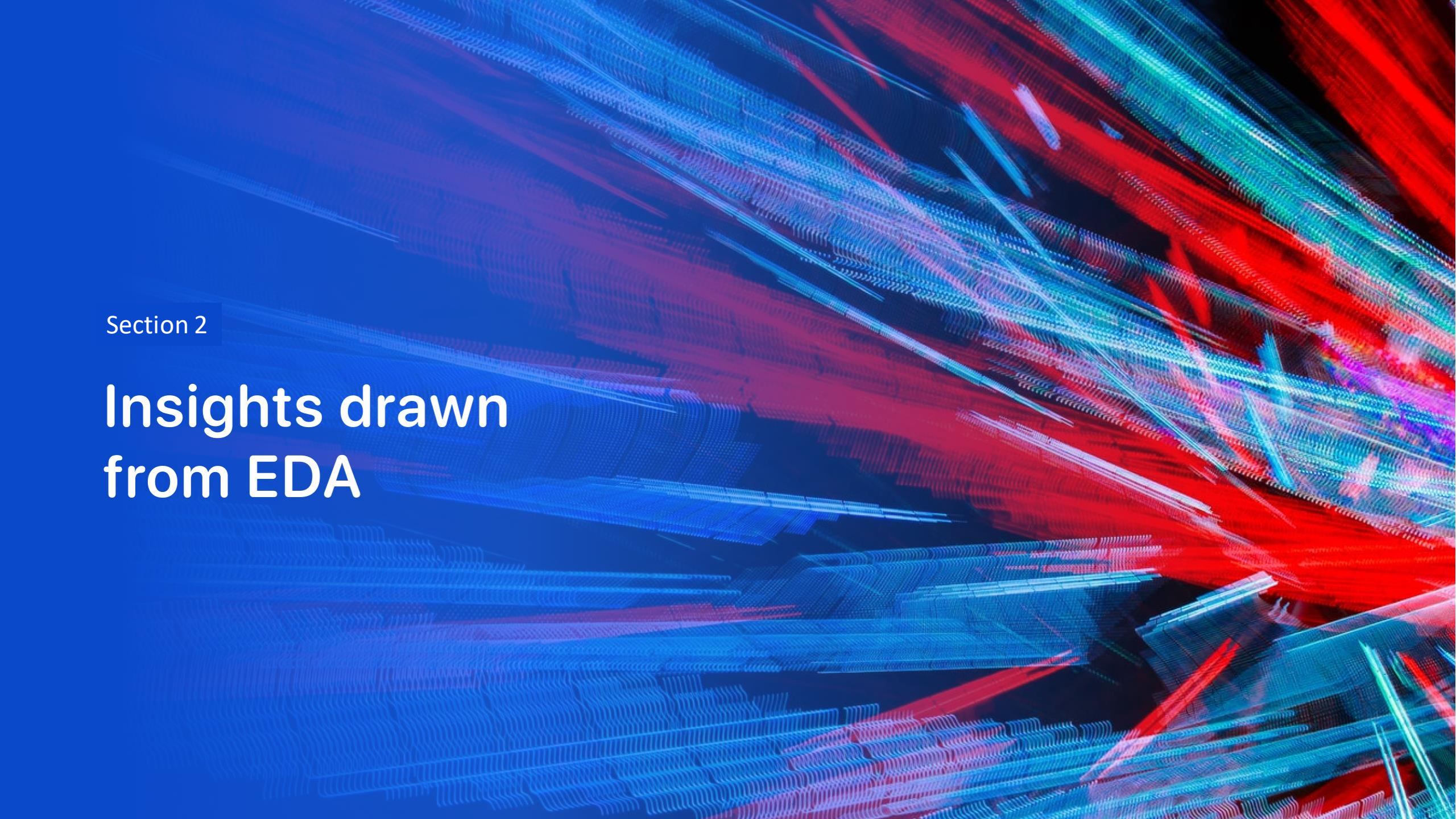
- Most launch sites are near equator; all are close to coast
- Launch sites:
  - Far enough, away from anything a failed launch can damage (city, highway, railway)
  - Close enough to bring people and material to support launch activities

Predictive Analytics

- Decision Tree model was the best predictive model

Section 2

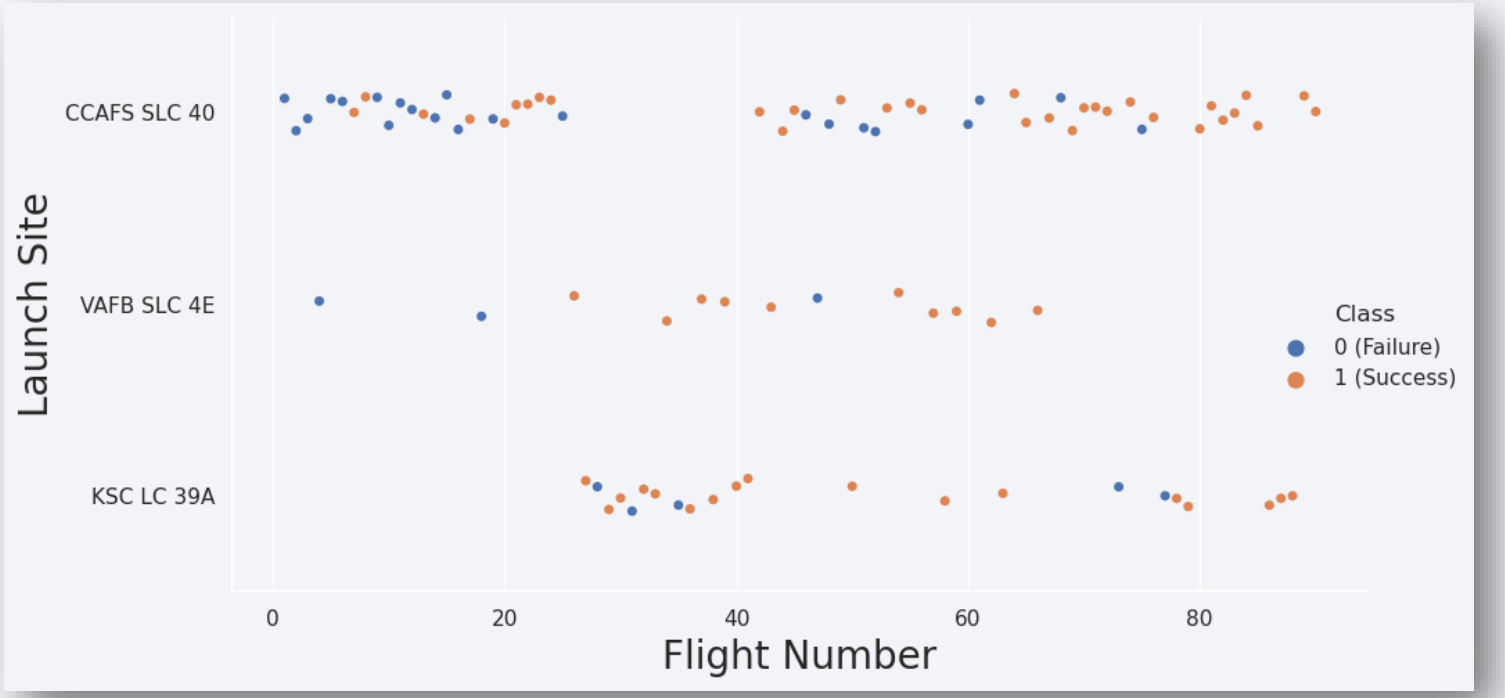
Insights drawn from EDA



Flight Number vs. Launch Site

Later Flight numbers and success rate are ~correlated

- **Earlier Flights** had **lower success rate** (Blue)
- **Later Flights** had **higher success rate** (Orange)
- CCAFS SLC 40: ~50% of launches & 60% success rate
- KSC LC-39A and VAFB SLC 4E: 77%success rate



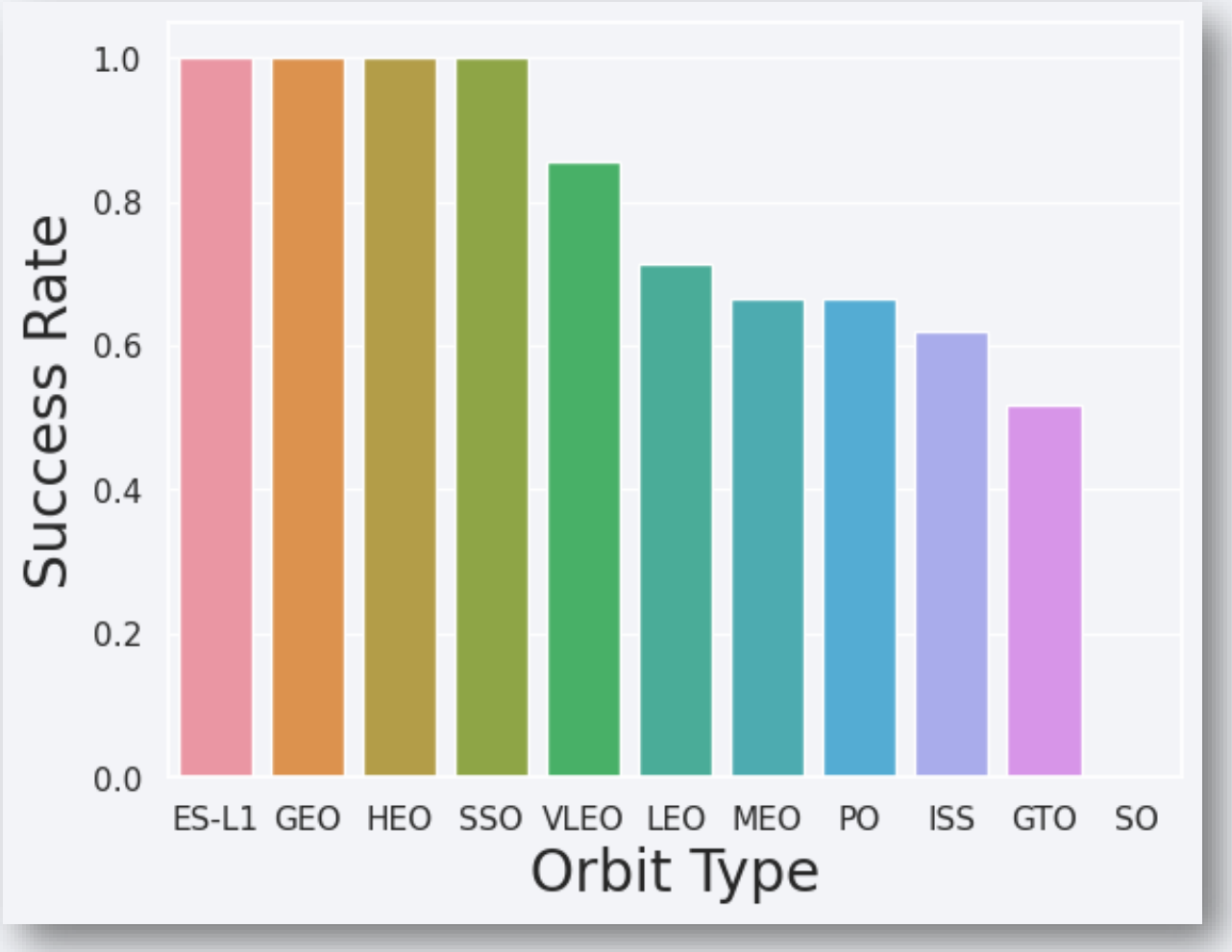
Payload vs. Launch Site

Typically, the higher the payload mass (kg), the higher the success rate

- KSC LC 39A: 100% success rate for launches < 5,500 kg
- VAFB SLC 4E: no rockets launched for heavy payload mass (> 10000 kg)

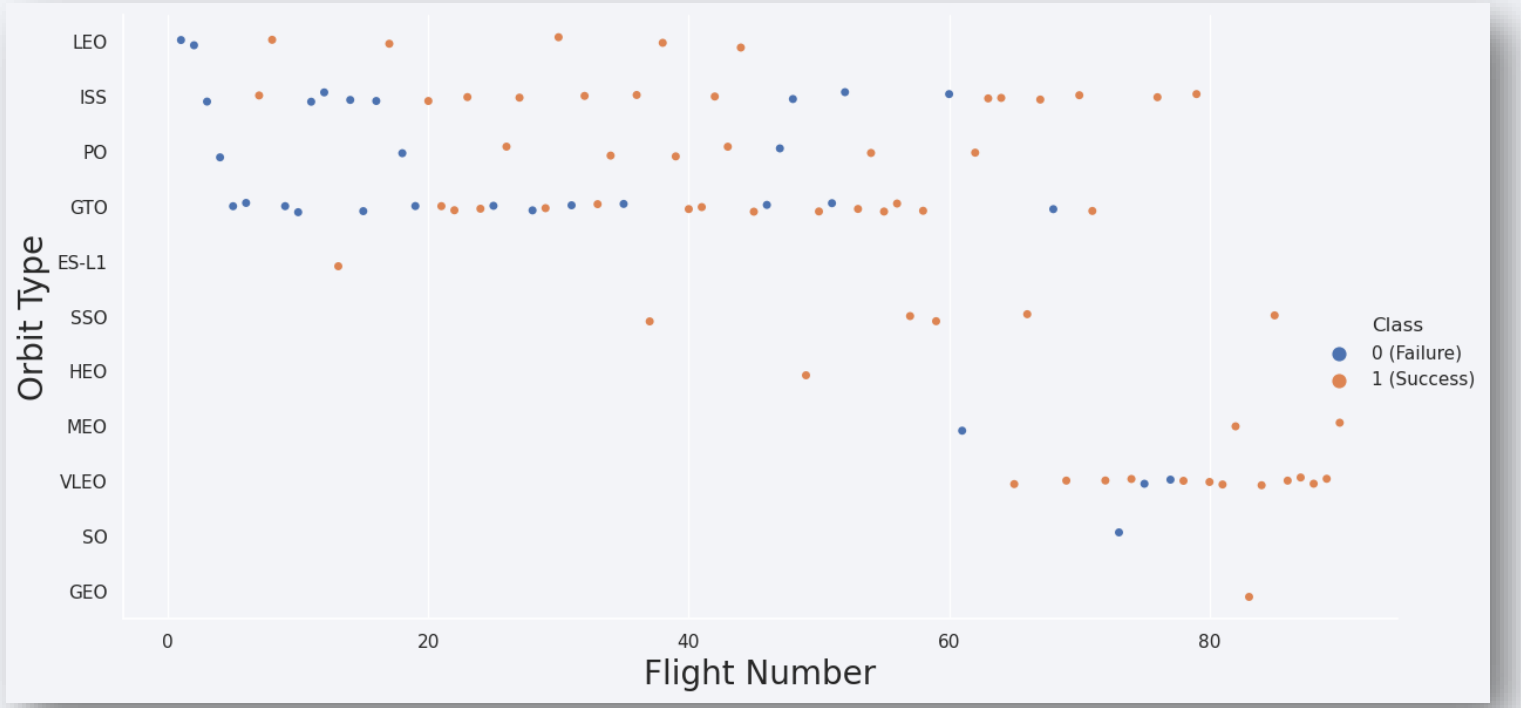
Success Rate vs. Orbit Type

- **100% Success Rate**: ES L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO
- **Depper Analysis:** Some orbits have only 1 occurrence (i.e., GEO, SO, HEO and ES-L1)
- More data needed to see pattern or trend before we draw any conclusion



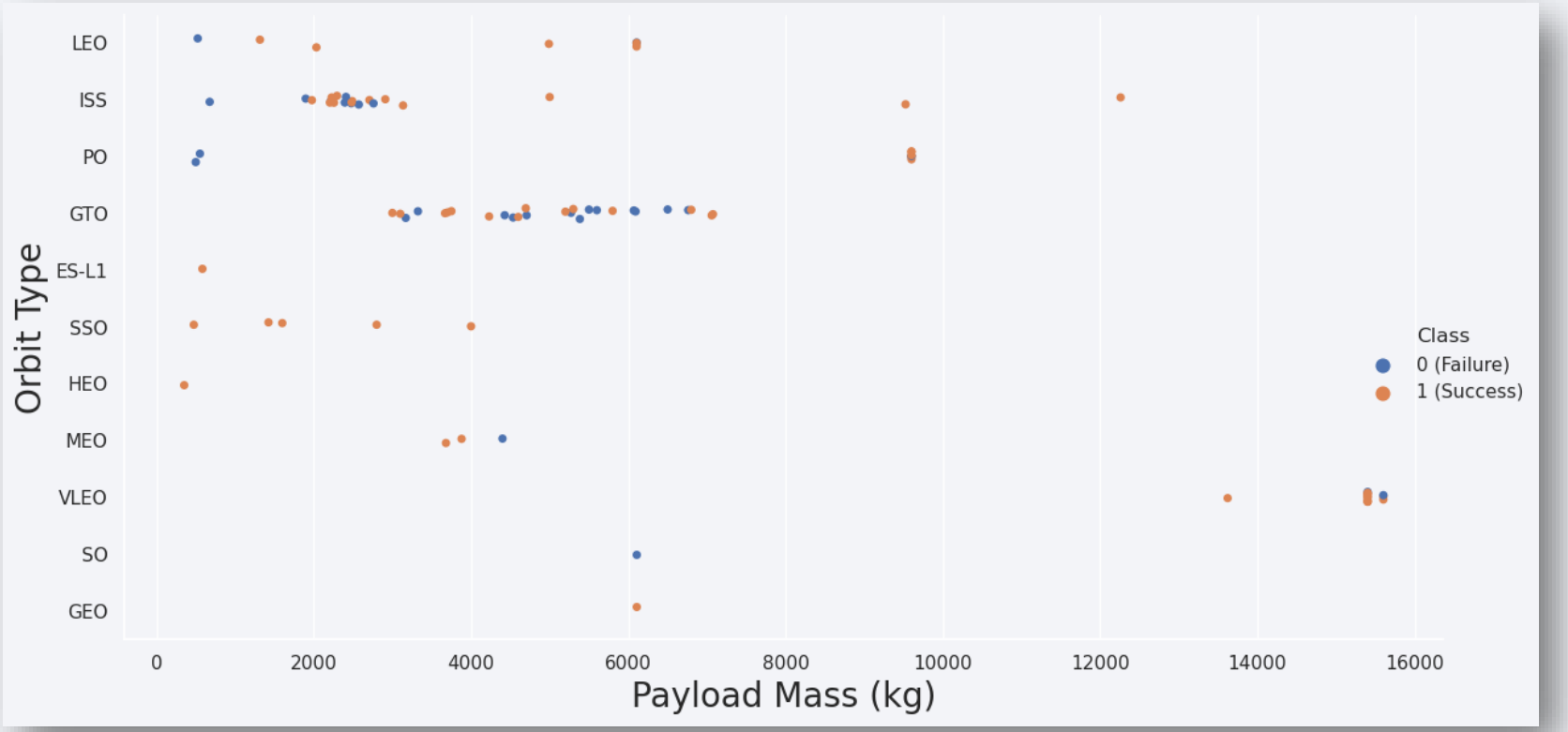
Flight Number vs. Orbit Type

- Success rate typically increases with the number of flights for each orbit
- Relationship is highly apparent for the LEO orbit
- GTO orbit does not follow this trend



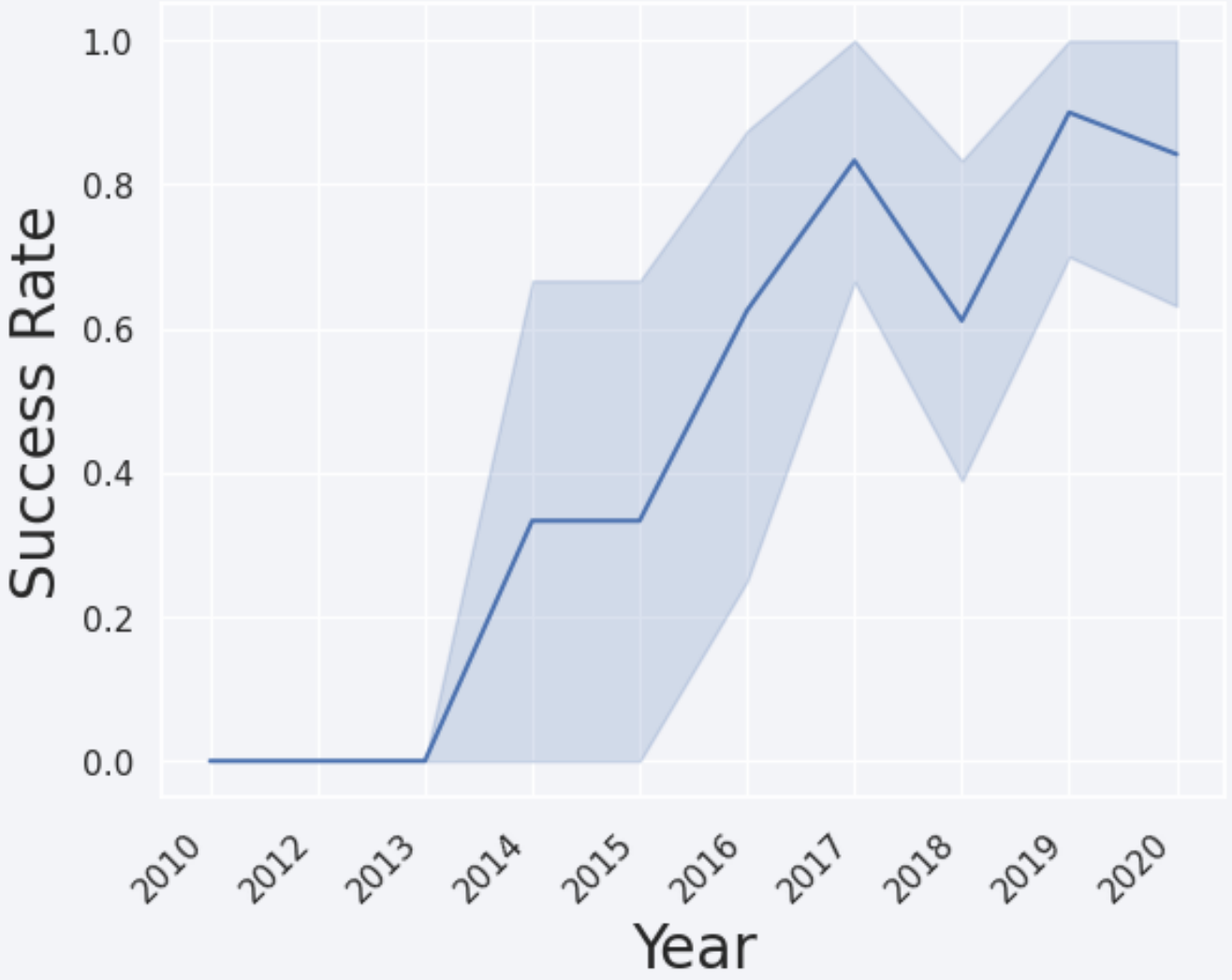
Payload vs. Orbit Type

- Heavy payloads better with LEO, ISS and PO orbits
- GTO orbit has mixed success with heavier payloads
- SO, GEO and HEO orbit need more dataset to see any pattern or trend



Launch Success Yearly Trend

- Success rate
  - **Improved** from 2013 - 2017 and 2018 2019
  - **Decreased** from 2017 - 2018 and 2019 - 2020
- Success rate has improved since 2013
- If trend continues, steadily increase until reaching 1/100% success rate



All Launch Site Names

Used **DISTINCT** to show only unique launch sites from the SpaceX data

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

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

Used the query below to display 5 records where launch sites begin with `CCA`

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

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
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

Total Payload Mass carried by boosters from NASA

```
%sql SELECT SUM(payload_mass_kg_) AS "total payload mass for NASA (CRS)" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

total payload mass for NASA (CRS)
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(payload_mass_kg_) AS "average payload mass for booster F9 v1.1"\
    FROM SPACEX \
    WHERE Booster_Version = 'F9 v1.1';
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

average payload mass for booster F9 v1.1
2928

First Successful Ground Landing Date

First **Successful** landing outcome on ground pad

```
%%sql
SELECT MIN(Date)
FROM SPACEX
WHERE landing_outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters that **Successfully** landed on drone ship with Payload Mass between 4000 and 6000

```
%%sql
SELECT DISTINCT booster_version
FROM SPACEX
WHERE landing_outcome = 'Success (drone ship)'
AND payload_mass_kg_ BETWEEN 4000 and 6000;
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

Total **Successful** and **Failed** Mission Outcomes

- used wildcard like '%' to filter for WHERE Mission Outcome was a success or a failure

```
%%sql
SELECT
    CASE
        WHEN mission_outcome LIKE '%Failure%' THEN 'Failure'
        ELSE 'Successful'
    END AS outcome,
    COUNT(*) AS Count
FROM SPACEX
WHERE mission_outcome != 'None'
GROUP BY CASE
        WHEN mission_outcome LIKE '%Failure%' THEN 'Failure'
        ELSE 'Successful'
    END;
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

outcome	COUNT
Failure	1
Successful	100

Boosters Carried Maximum Payload

Booster that carried the Maximum Payload Mass

```
%%sql
SELECT booster_version FROM SPACEX
WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM SPACEX)
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
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

2015 **Failed** **Landing Outcomes** in drone ship:

- Showing month, landing outcome, booster version and launch site

```
%%sql
SELECT substr(Date, 6, 2) as month, landing_outcome, booster_version, launch_site FROM SPACEX
WHERE substr(Date,1,4)='2015' and landing_outcome = 'Failure (drone ship)';
```

```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

MONTH	landing_outcome	booster_version	launch_site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Count of Landing Outcomes between 2010-06-04 and 2017-03-20 Ranked (Descending Order)

- e.g., **Failure** (drone ship) or **Success** (ground pad)

```
%%sql
SELECT landing_outcome, Count(*) as cnt FROM SPACEX
WHERE DATE between '2010-04-06' and '2017-03-20'
group by landing_outcome
ORDER BY cnt DESC;
```

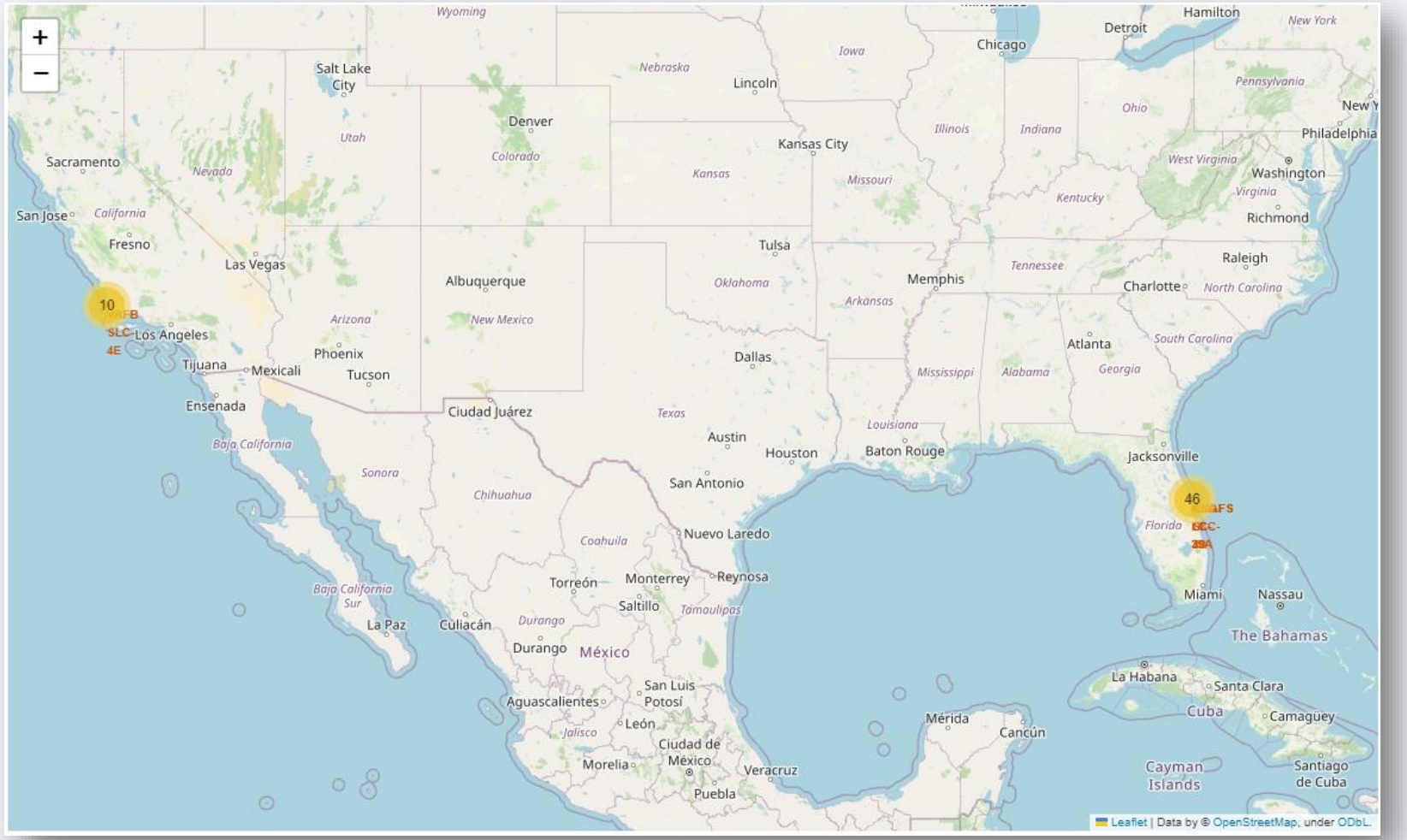
```
 * ibm_db_sa://rls87460:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

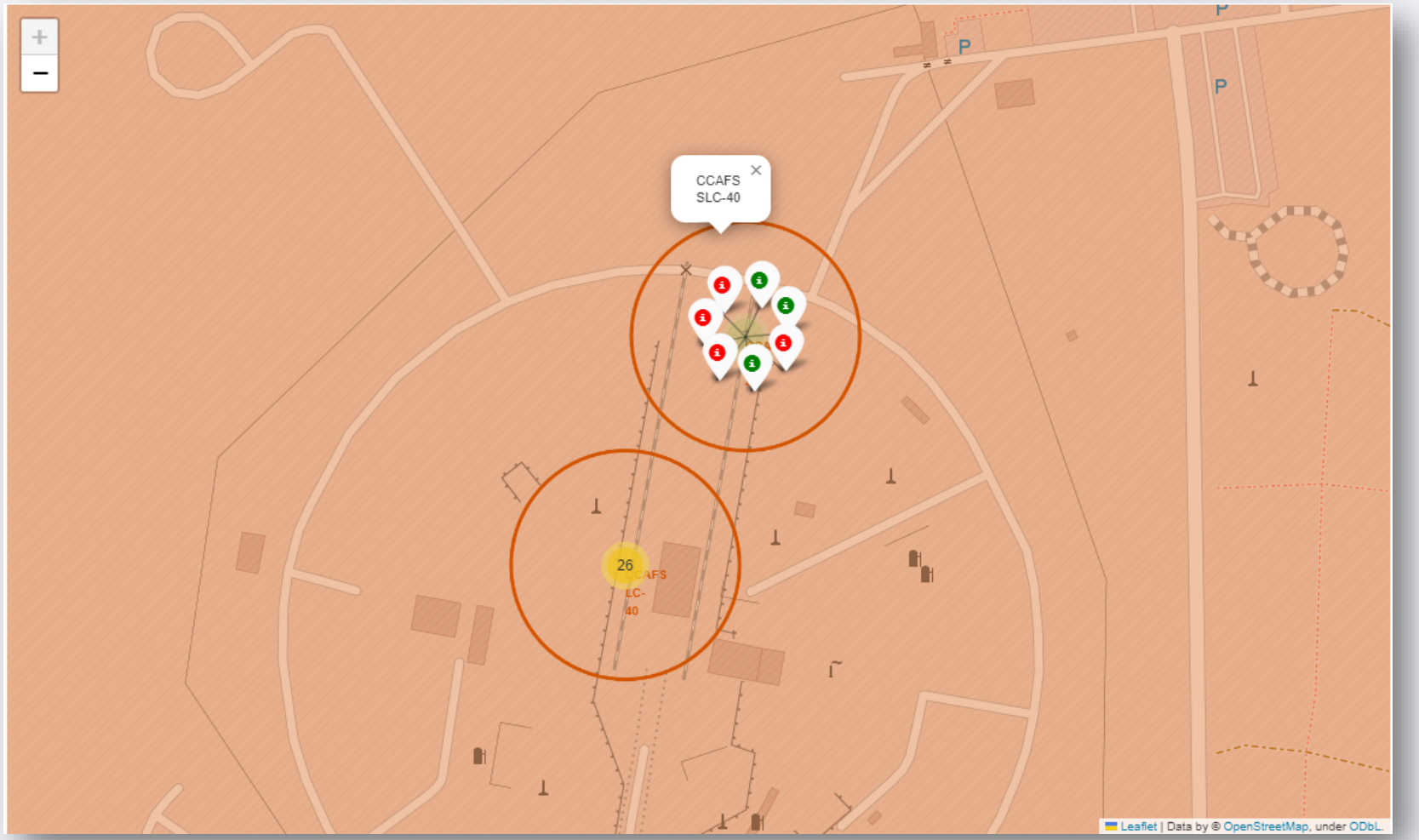
Launch Site

- **Near Equator**: the closer the launch site to equator, the easier to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde trajectory
- Rockets launched from sites near equator get additional boost due to the Earth's rotational speed
  - Helps save cost of putting extra fuel and boosters



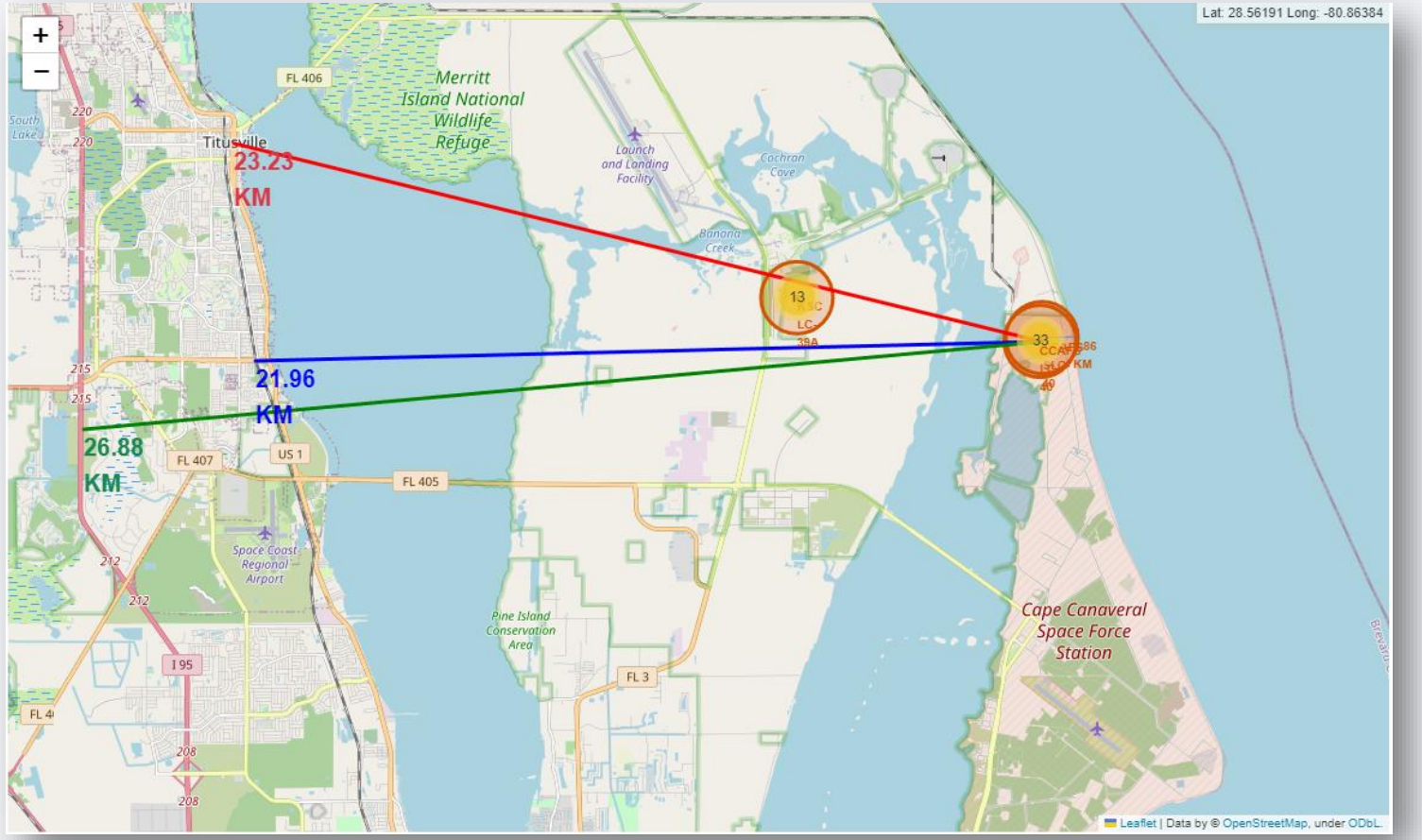
Launch Outcomes

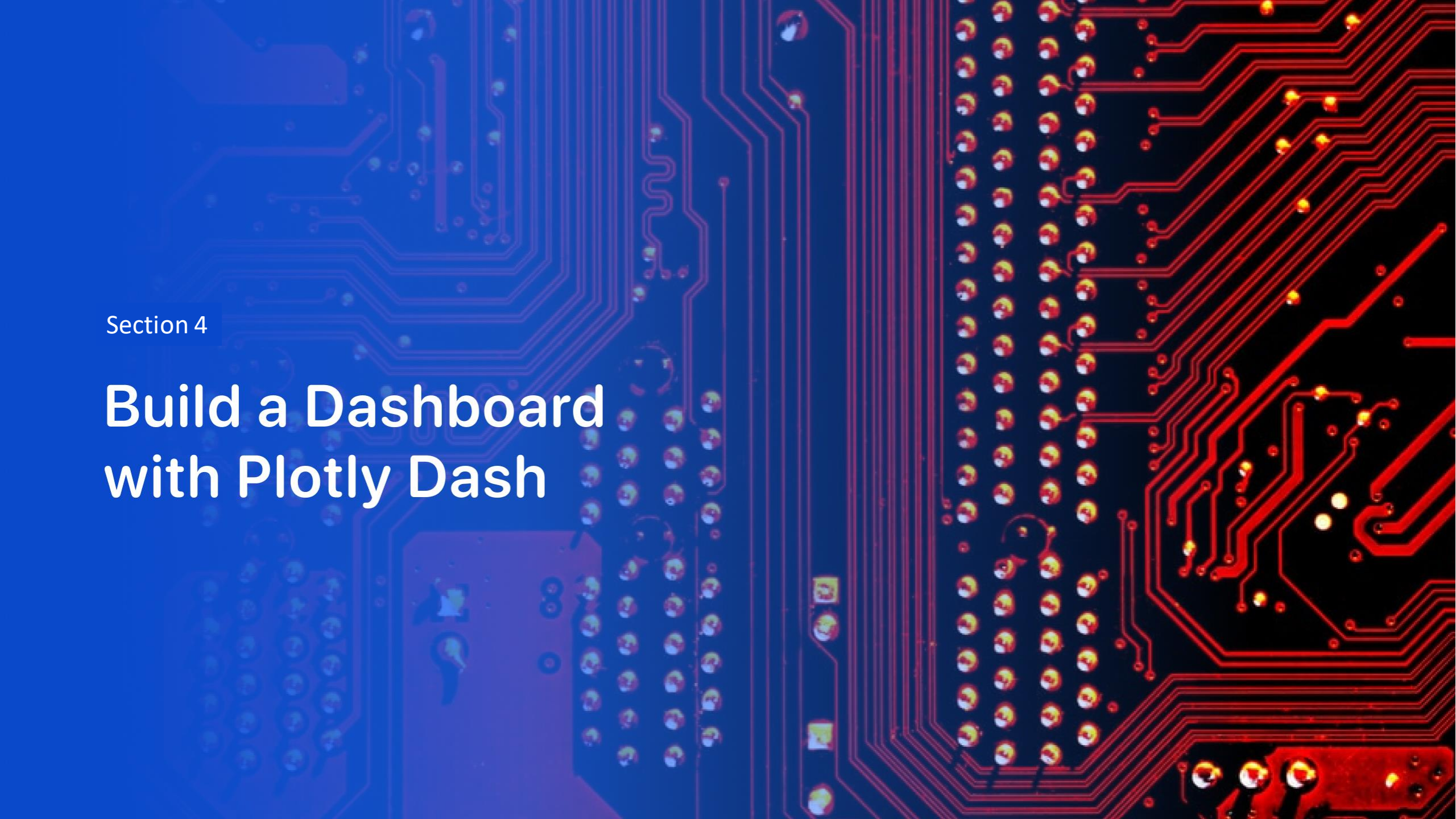
- **Green** markers for **successful** launches
- **Red** markers for **failed** launches
- **CCAFS SLC 40**: 3/7 success rate (i.e., 42.9%)



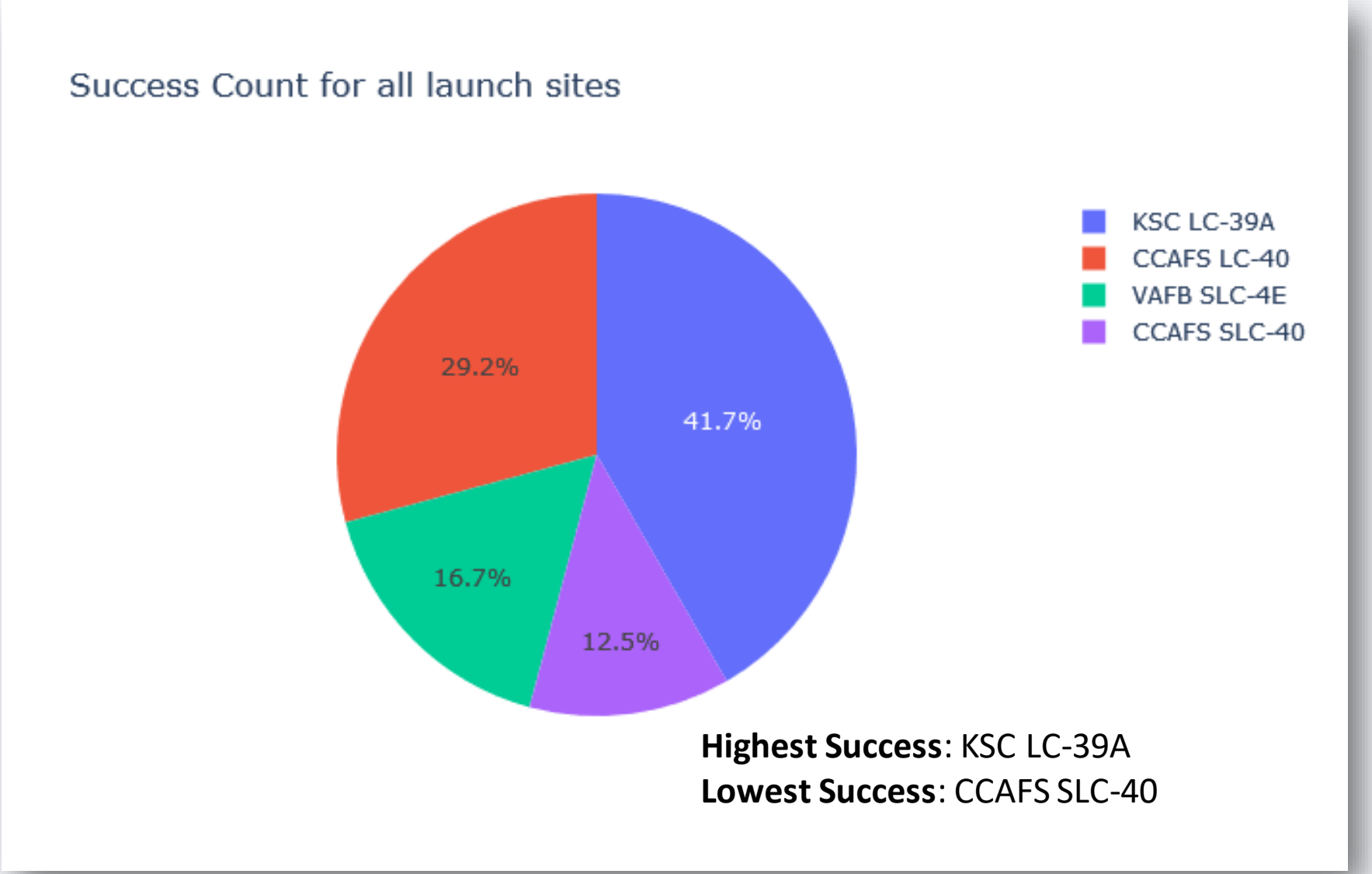
Distances to Proximities

- **Coasts**: help ensure that spent stages dropped along launch path or failed launches don't fall on people or property
- **Safety/Security**: exclusion zone required around launch site to keep unauthorized people away and safe
- **Transportation/Infrastructure and Cities**: needs to be away from anything a failed launch can damage but still close enough to roads/rails/docks to bring people and material to and/or from it to support launch activities



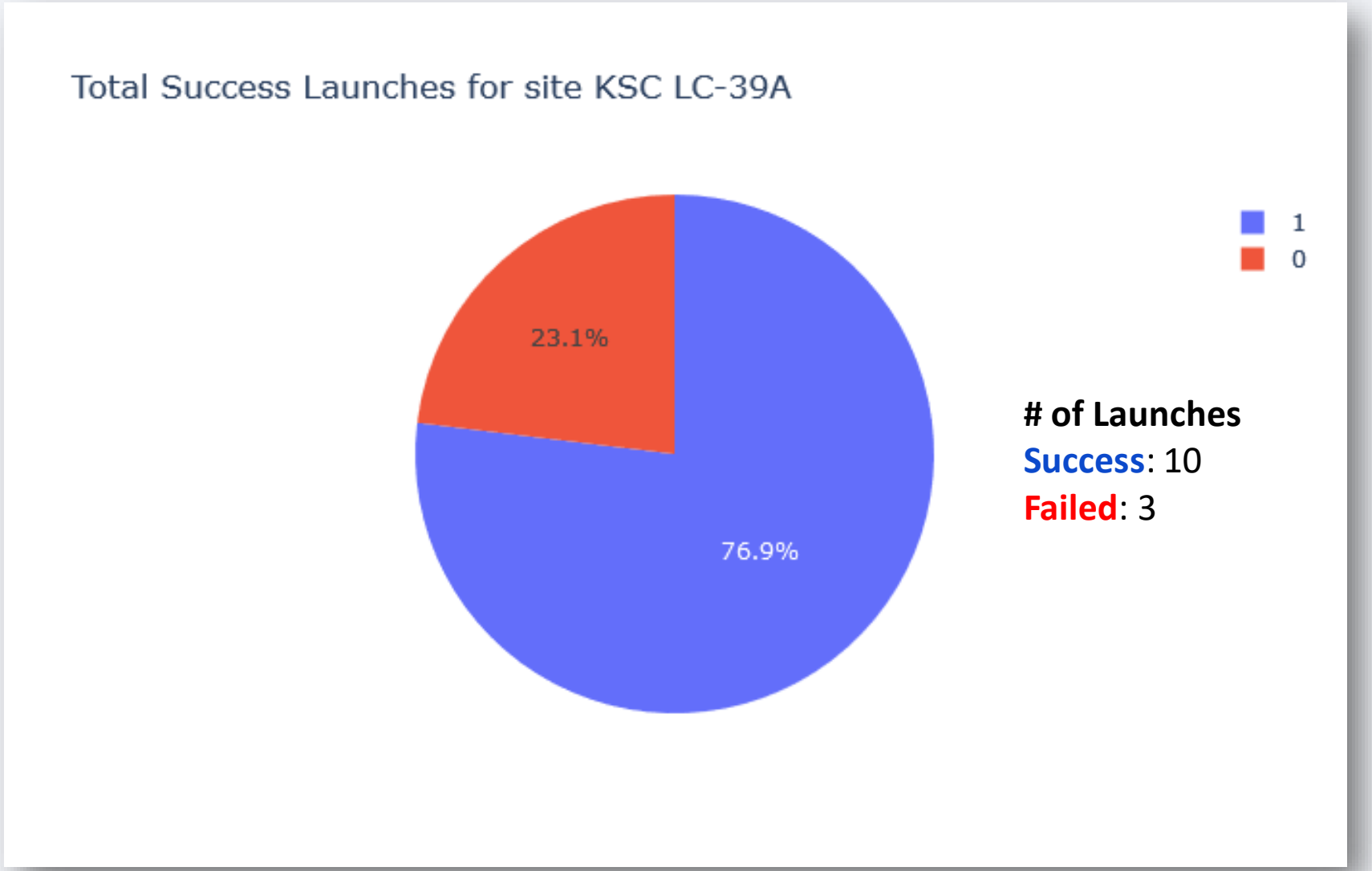

Section 4
Build a Dashboard with Plotly Dash

Launch Success by Site



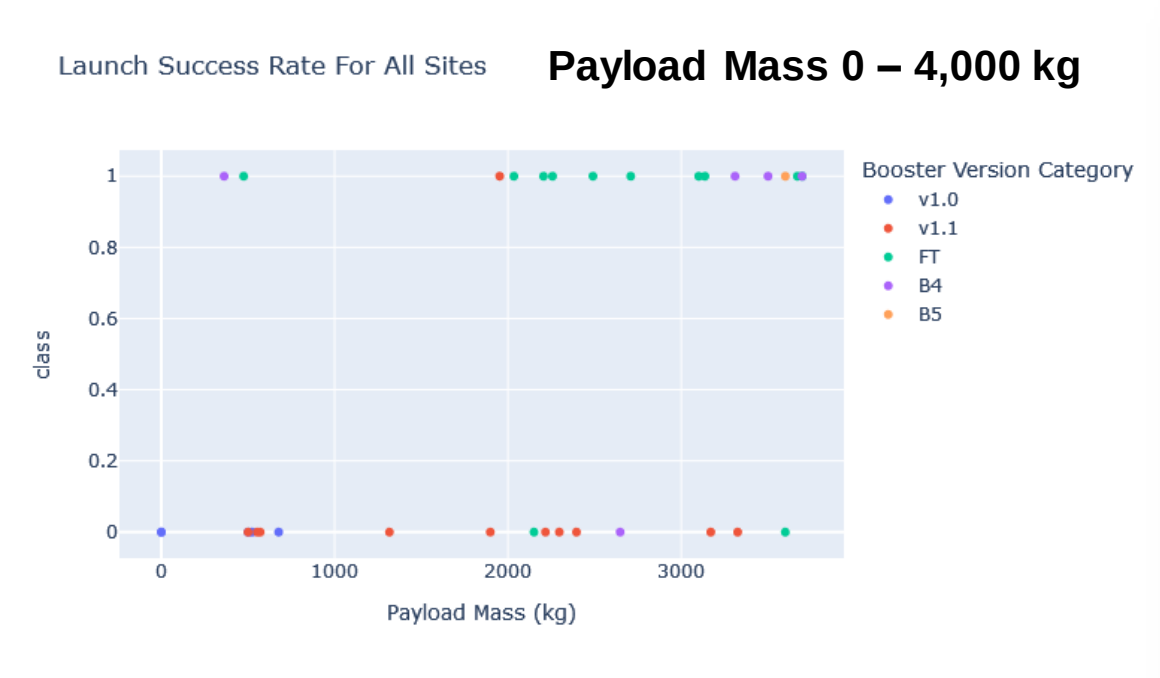
Highest Success: KSC LC-39A
Lowest Success: CCAFS SLC-40

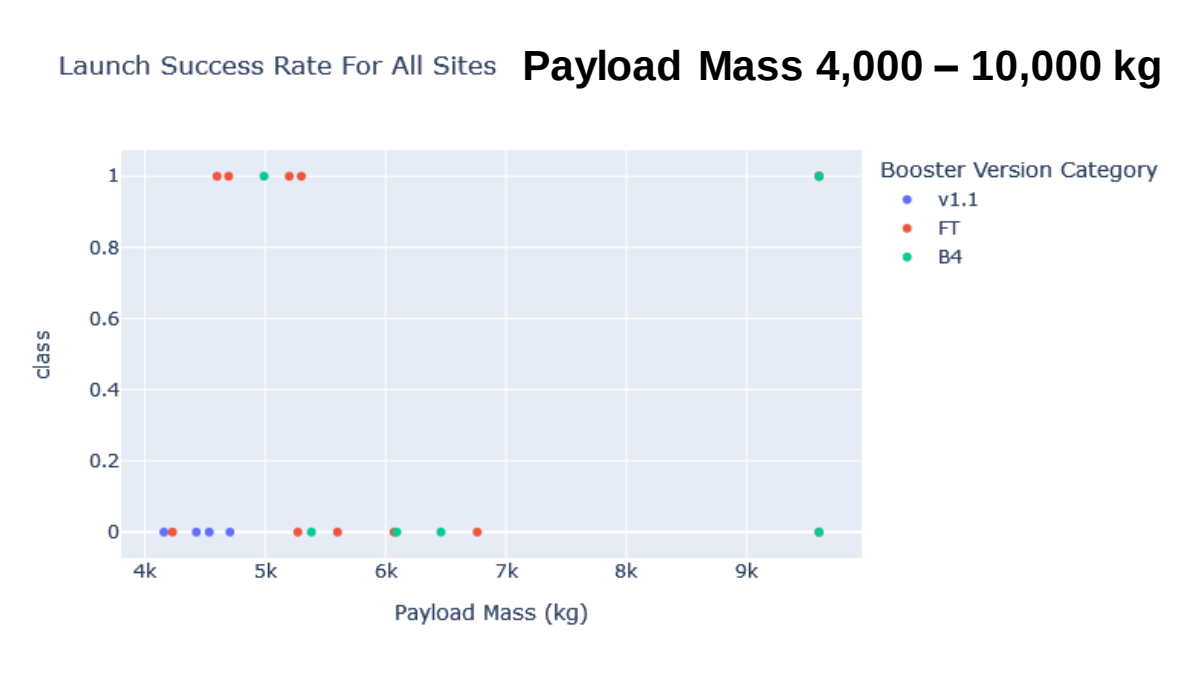
Launch Success (KSC LC-29A)



Payload Mass and Success

Success rate for low weighted payload is higher than heavy weighted payload





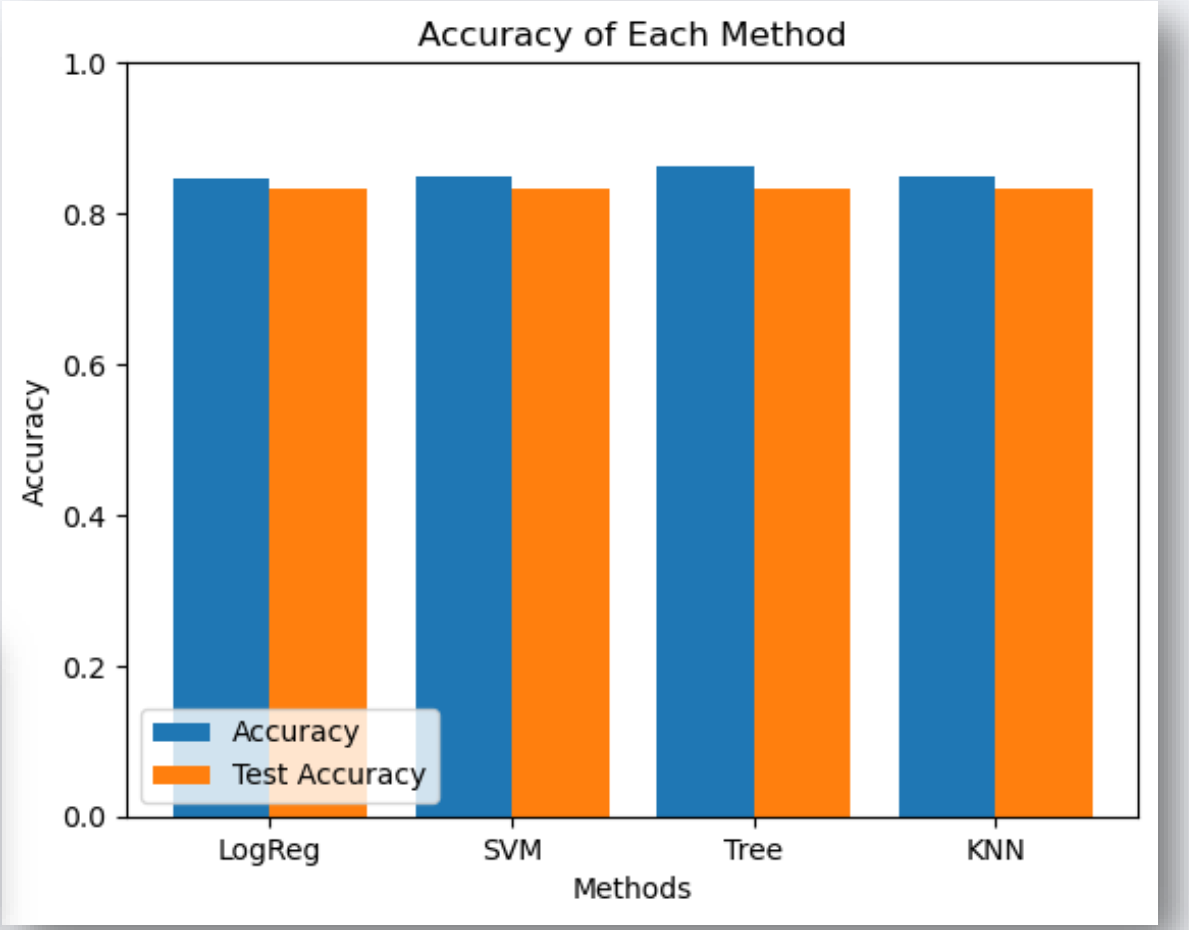
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All models produced ~87% accuracy on the test data
- Decision tree performed the best (slightly) on the training data

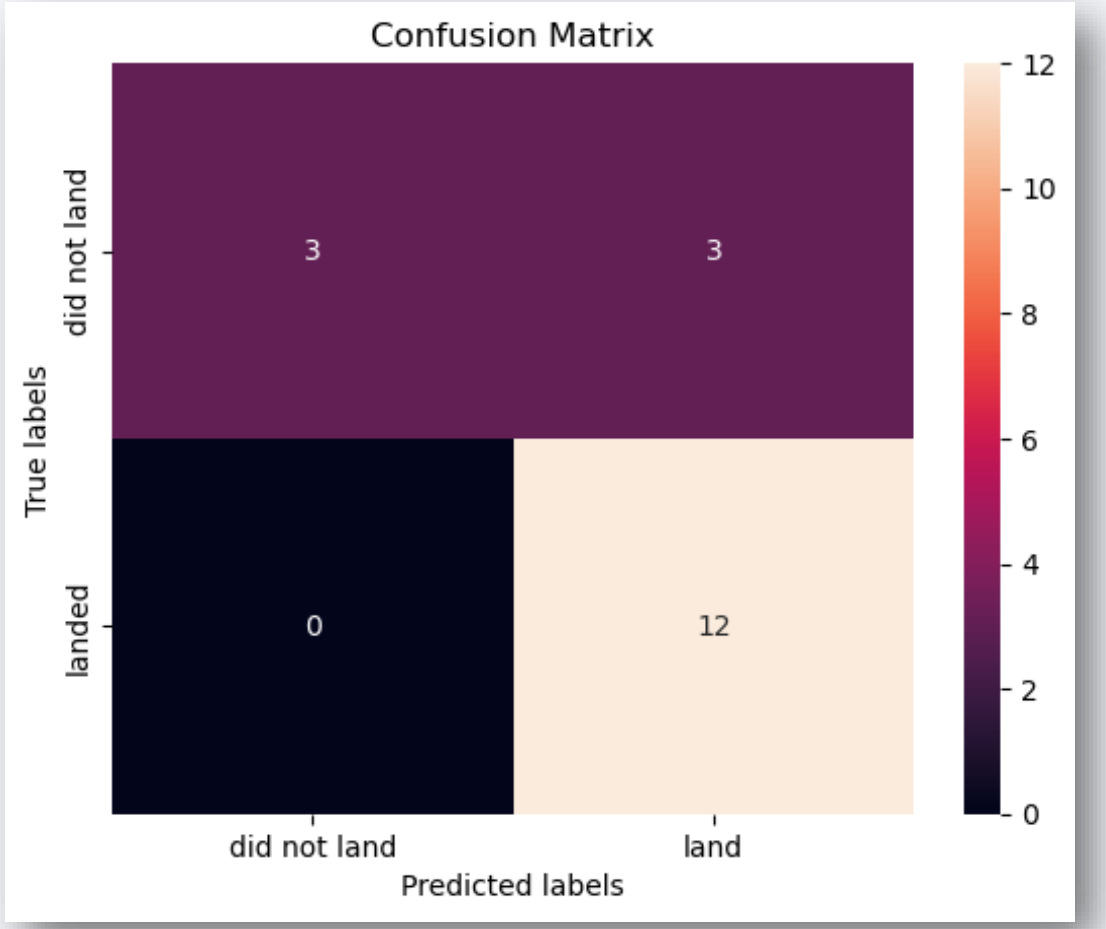
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.8625	0.83333
KNN	0.84821	0.83333



Confusion Matrix

Decision Tree

- Distinguishes between different classes (i.e., Landed or Not)
- **False Positives** (**Type 1 Error**): Major Problem
  - unsuccessful landing marked as successful landing by the classifier
- Outputs:
  - 12 True Positive (TP)
  - 3 True Negative (TN)
  - **3 False Positive (FP)**
  - 0 False Negative (FN)
- **Precision**: TP/ (TP + FP) = **.80**
- **Recall**: TP/ (TP + FN) = **1**
- **F1-score**: 2*(Precision * Recall)/ (Precision + Recall) = **.89**
- **Accuracy**: (TP+TN)/ (TP + TN + FP + FN) = **.833**



Conclusions

- **Model Performance**: Each performed similarly on the test set, with the **decision tree** model slightly outperforming
- **Equator**: Most launch sites are near the equator for an additional natural boost due to the rotational speed of Earth
  - Helps save the cost of putting in extra fuel and boosters
- **Coast**: All launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has highest success rate among launch sites ~76.9%
  - Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES L1, GEO, HEO, and SSO have 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

All Python code, SQL queries, charts, Notebook outputs, and data sets are available at the following GitHub Repository:

SpaceX First Stage Launch GitHub Repository

Thank you!