IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Chuan Yin>
<07-09-2024>

Outline

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

Executive Summary

- This presentation encompasses a comprehensive data analysis project
- Module 1: Introduction
  - Define and formulate a real-world business problem using data science methodologies
  - Load, clean, and analyze datasets to find interesting insights
- Module 2: Exploratory Data Analysis (EDA)
  - Visualize data to extract meaningful patterns for guiding the modeling process, in SQL and Pandas
- Module 3: Interactive Visual Analytics and Dashboard
  - Build an interactive dashboard with pie charts and scatter plots using Plotly Dash
  - Generate interactive maps, plot coordinates, and mark clusters with Folium
- Module 4: Predictive Analytics (Classification)
  - Build predictive models to enhance business efficiency using machine learning

Introduction

- In this project, we will predict if the Falcon 9 first stage will land successfully.
- Much of SpaceX's competitive pricing is due to the saving from reusing 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.

Section 1

Methodology

Methodology

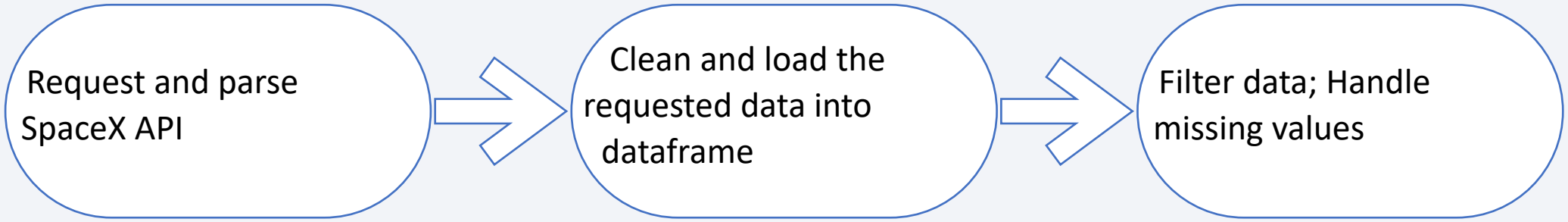
Executive Summary

- Data collection methodology:
  - Data collected from various sources by API and web scraping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classifiers: logistic regression, support vector machine (SVM), decision tree classifier, k-nearest neighbor
  - Tune hyperparameters for each model with GridSearchCV
  - Evaluate each model on test data using `score` and confusion matrix

Data Collection

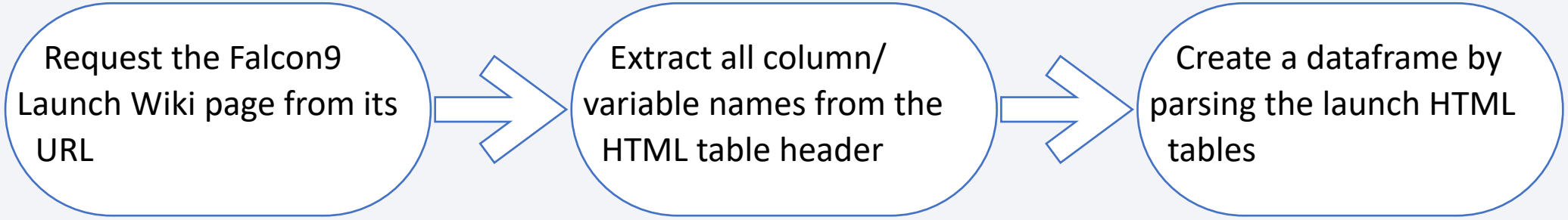
- Data are collected from the following sources
  - Space X API (https://api.spacexdata.com/v4/), via `request` library
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) , via `beautifulsoup4` library

Data Collection – SpaceX API



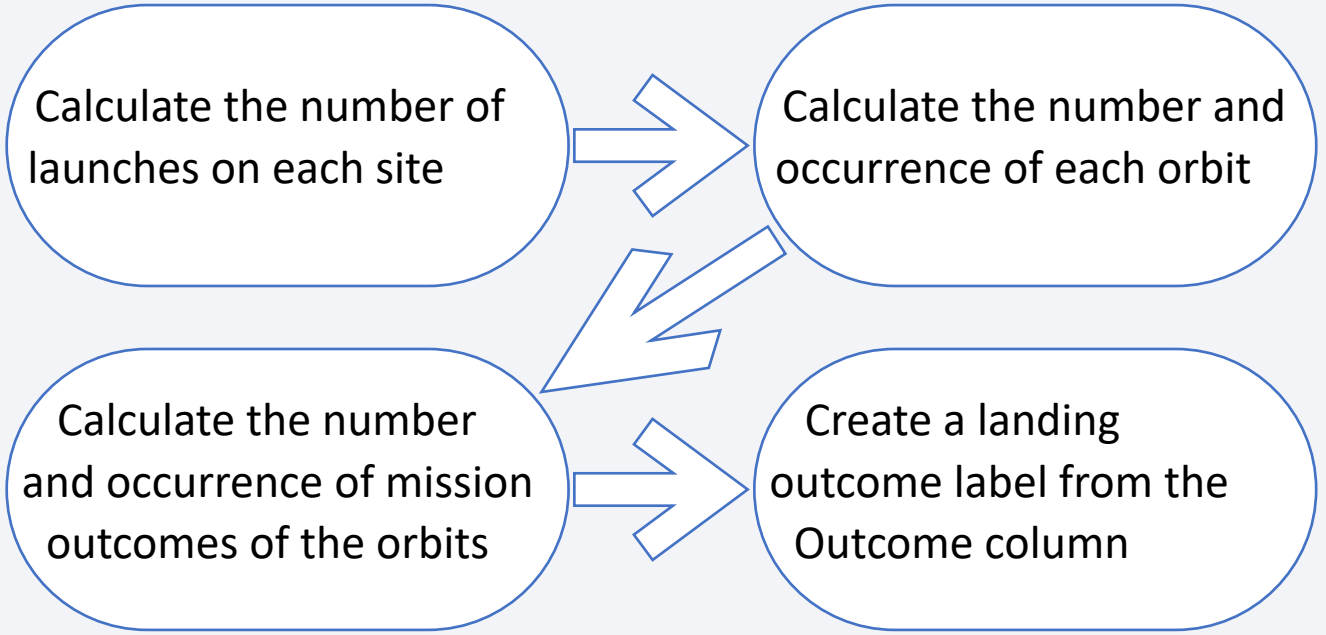
- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/1_Data Collection API Lab.ipynb

Data Collection - Scraping



- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/2_Data Collection with Web Scraping lab.ipynb

Data Wrangling



- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/3_Data Wrangling.ipynb

EDA with Data Visualization

- Scatter: on the relationship between Flight Number and Launch Site
- Scatter: on the relationship between Payload and Launch Site
- Bar: on the relationship between success rate of each orbit type
- Scatter: on the relationship between Flight Number and Orbit type
- Scatter: on the relationship between Payload and Orbit type
- Line: on the launch success yearly trend

- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/5_EDA with Visualization Lab.ipynb

EDA with SQL

- 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
- 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

- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/4_EDA with SQL.ipynb

Build an Interactive Map with Folium

- Markers added for each launch site and each launch result from the spacex_df dataframe. To pinpoint the exact locations of launch sites and individual launch events. This helps in identifying the geographical spread and specific positions on the map.
- Circles added around each launch site. To provide a visual boundary around each launch site, making it easier to spot on the map and giving a sense of the area covered by each site.
- Marker Clusters added. To manage a large number of markers and make the map easier to navigate and interpret.
- Polylines added between launch sites and the nearest coastline points. To visualize the distance and connection between launch sites and the coastline, aiding in understanding the logistical and geographical context.

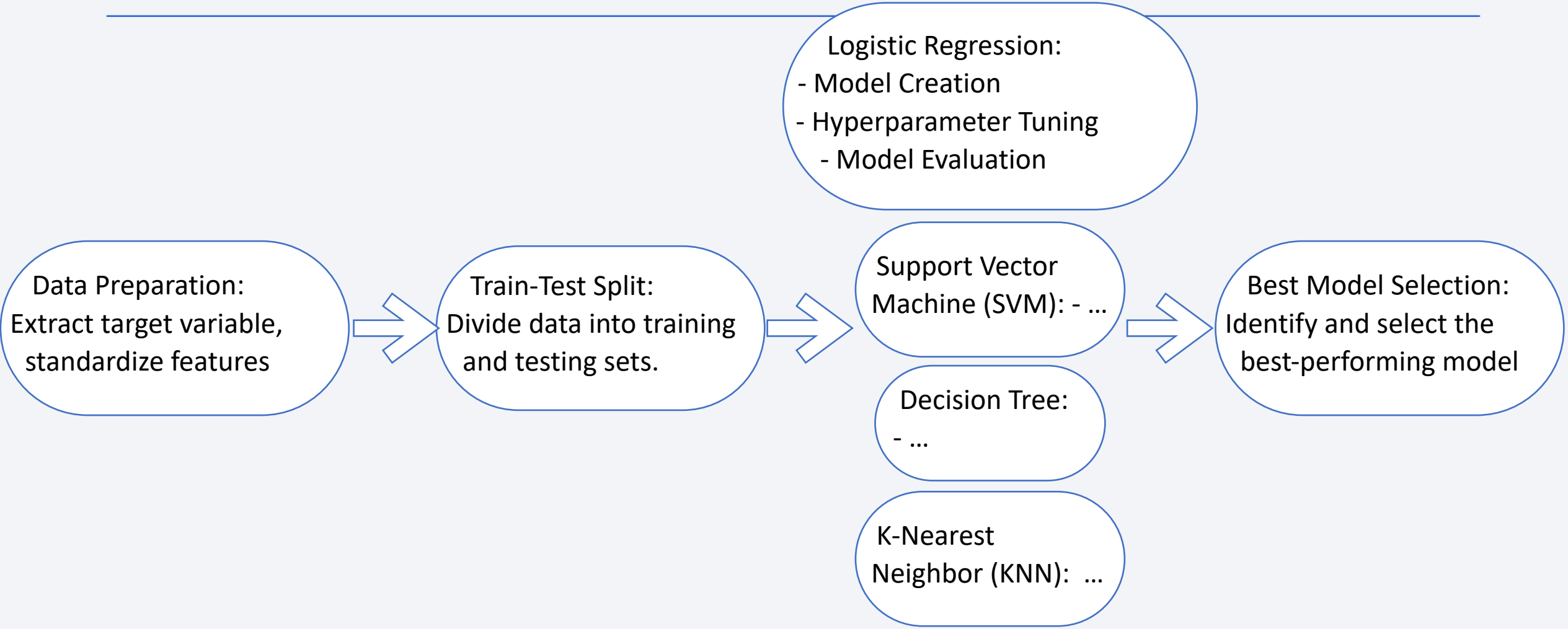
- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/6_Interactive Visual Analytics with Folium lab.ipynb

Build a Dashboard with Plotly Dash

- Launch Site Dropdown: enables specific site analysis, enhancing interactivity.
- Success Pie Chart: provides a quick overview of launch success rates.
  - A callback function for `site-dropdown` as input, `success-pie-chart` as output
- Payload Range Slider: allows focused analysis on specific payload ranges.
- Scatter Chart: identifies trends and correlations between payload and success.
  - A callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output

- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/7_spacex_dash_app.py

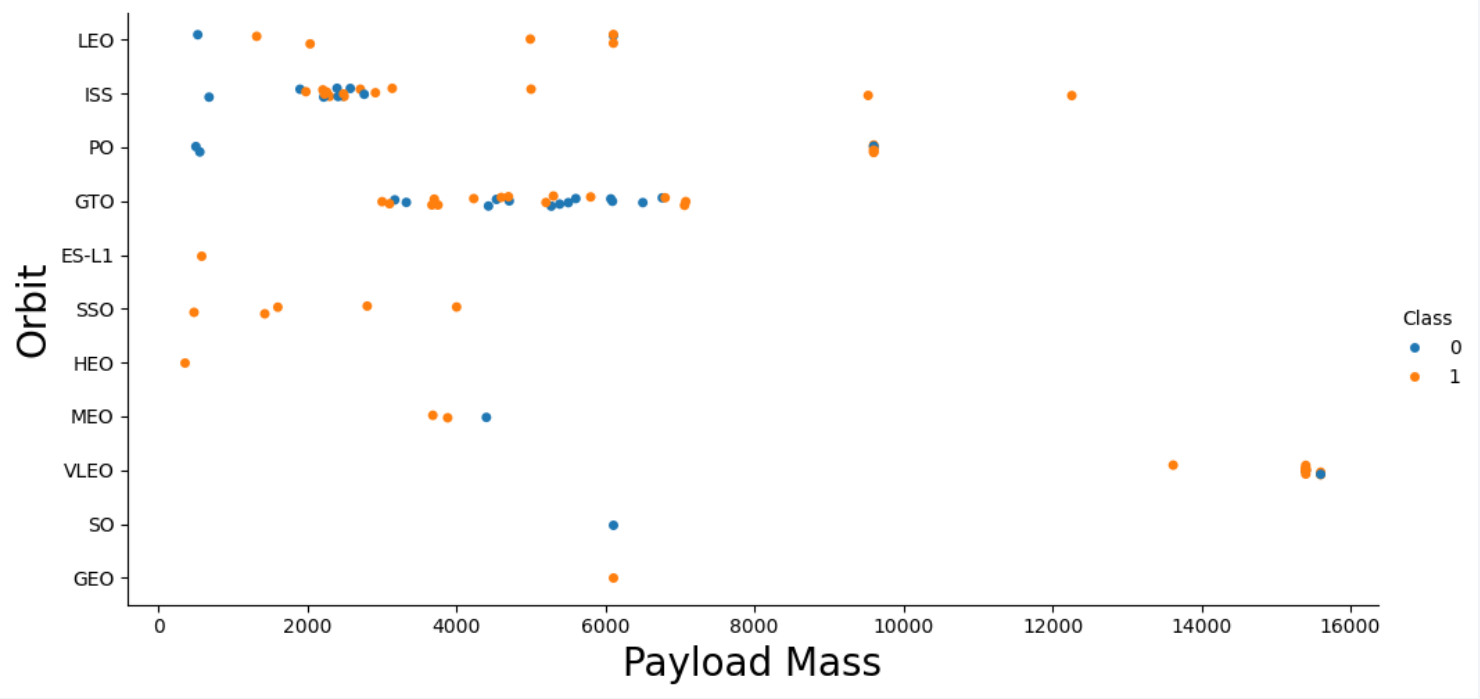
Predictive Analysis (Classification)

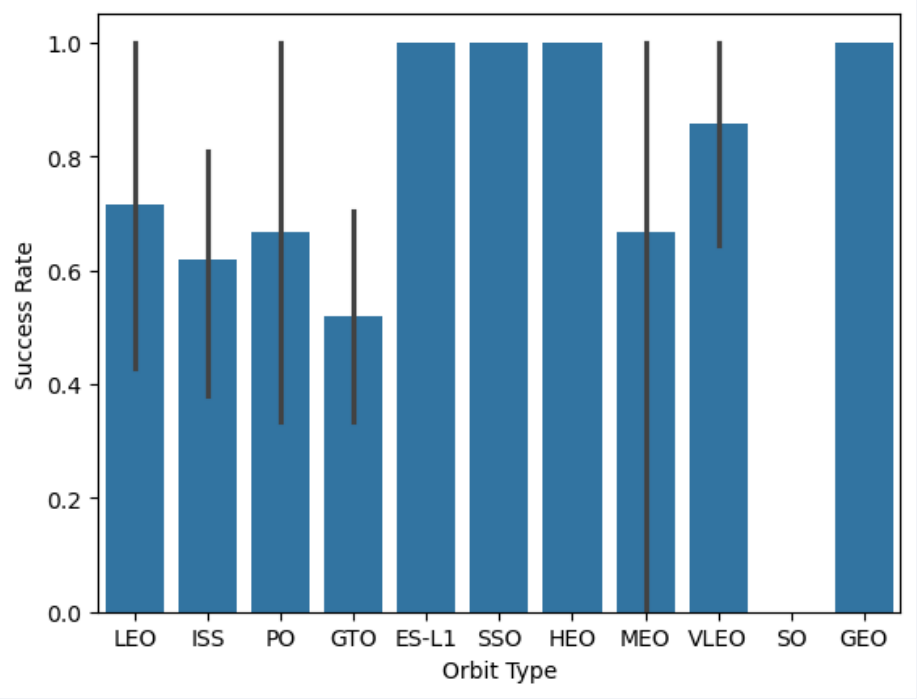


- https://github.com/chuanyinn/coursera/blob/main/0_course_10_Applied-Data-Science-Capstone/8_Machine Learning Prediction lab.ipynb

Results

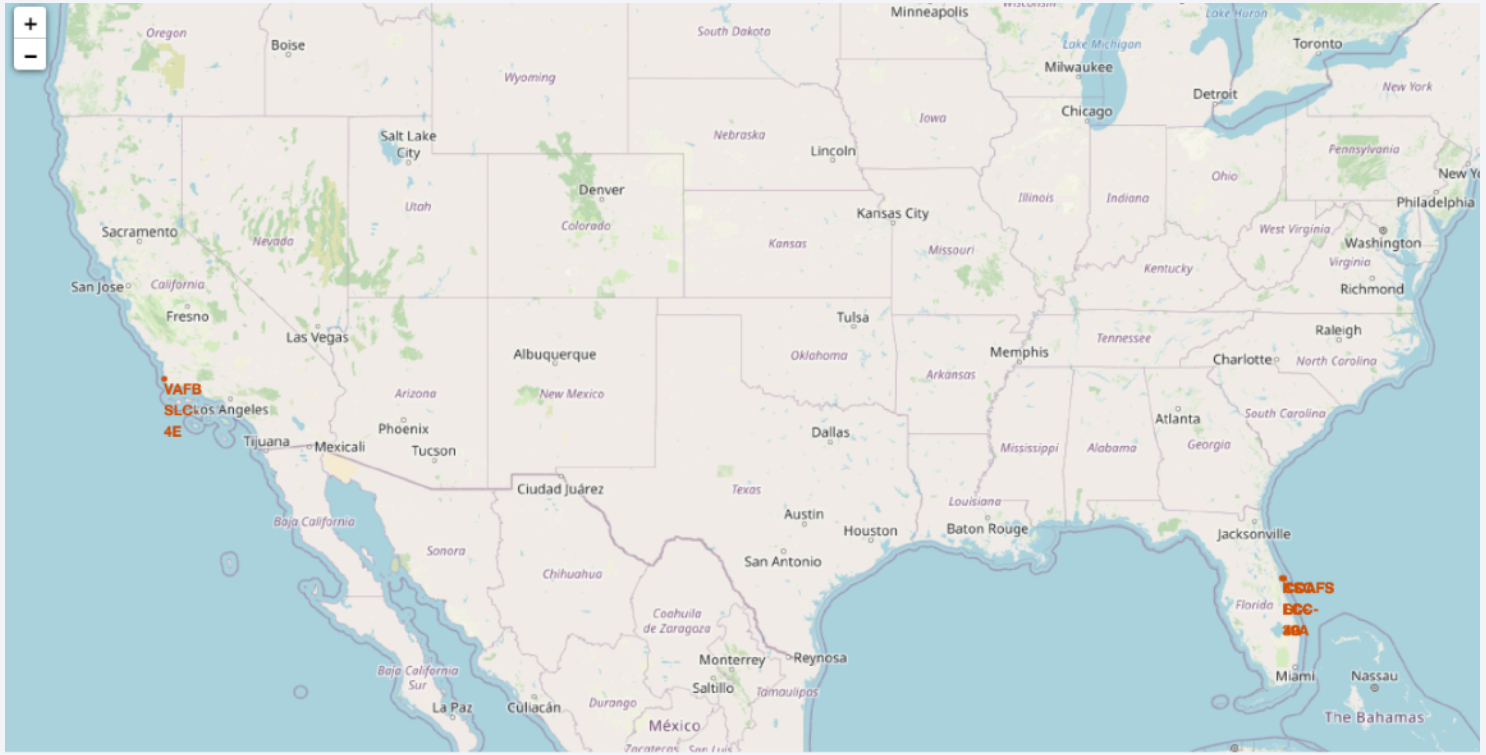
- Exploratory data analysis results

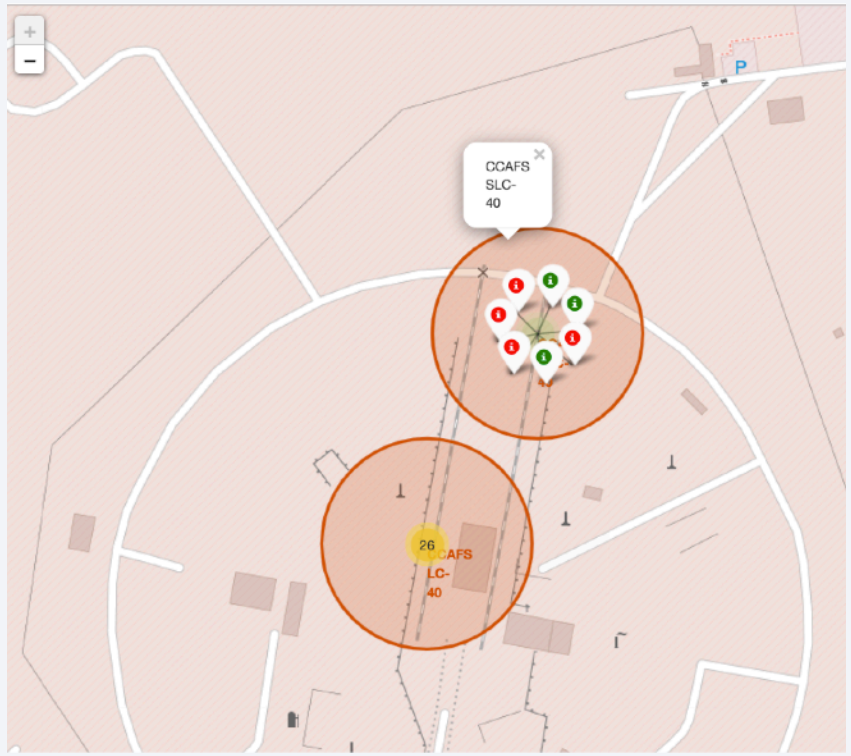




Results

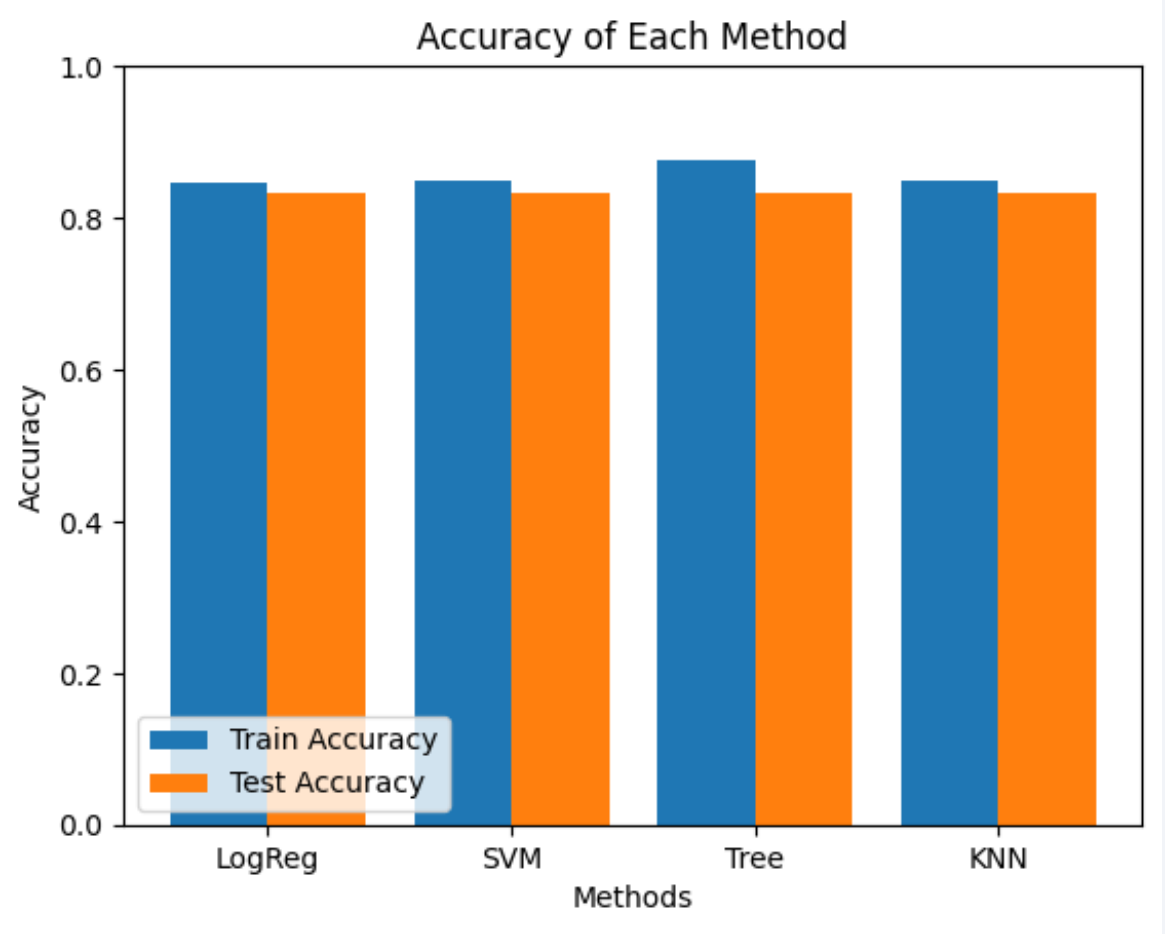
- Interactive analytics demo in screenshots





Results

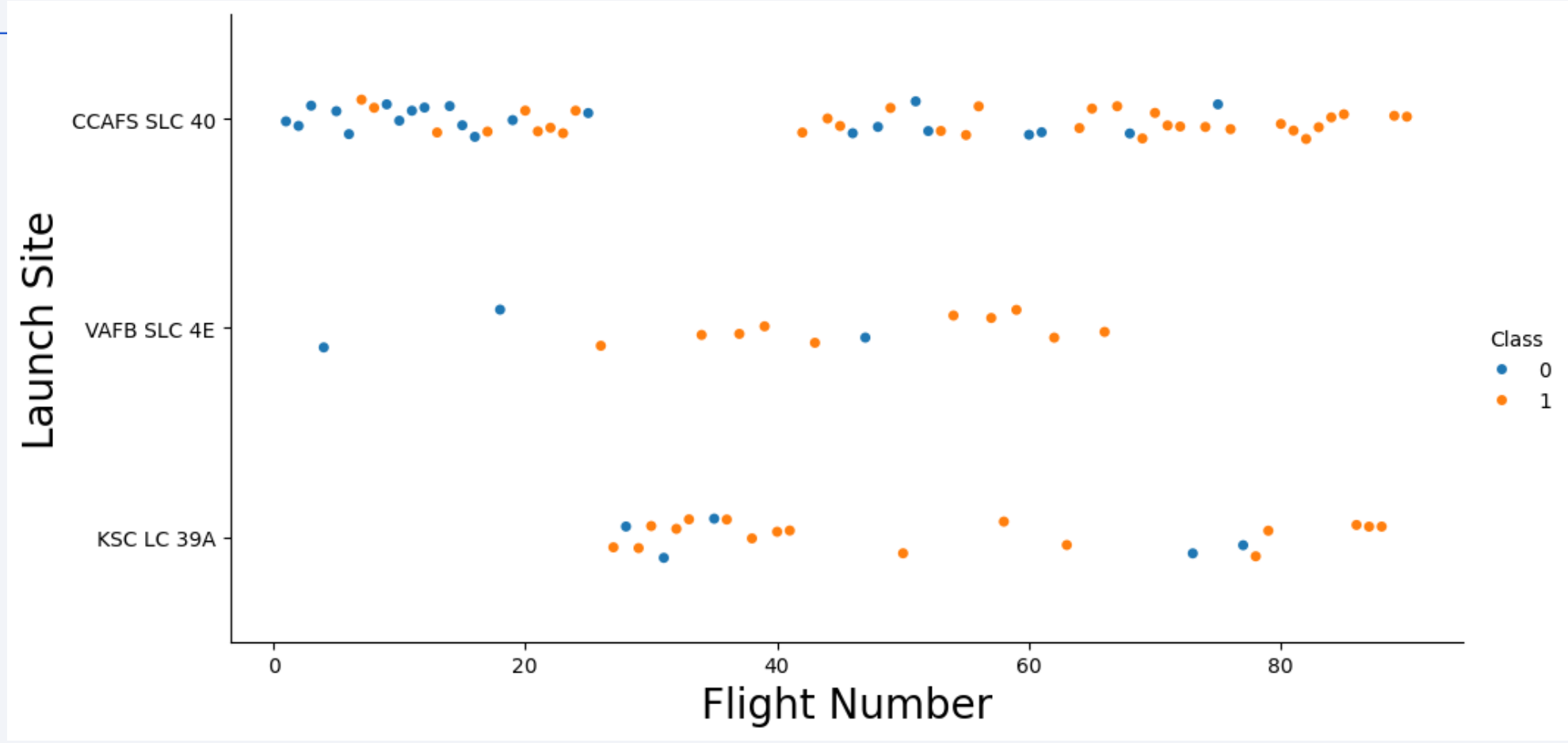
- Predictive analysis results



Section 2

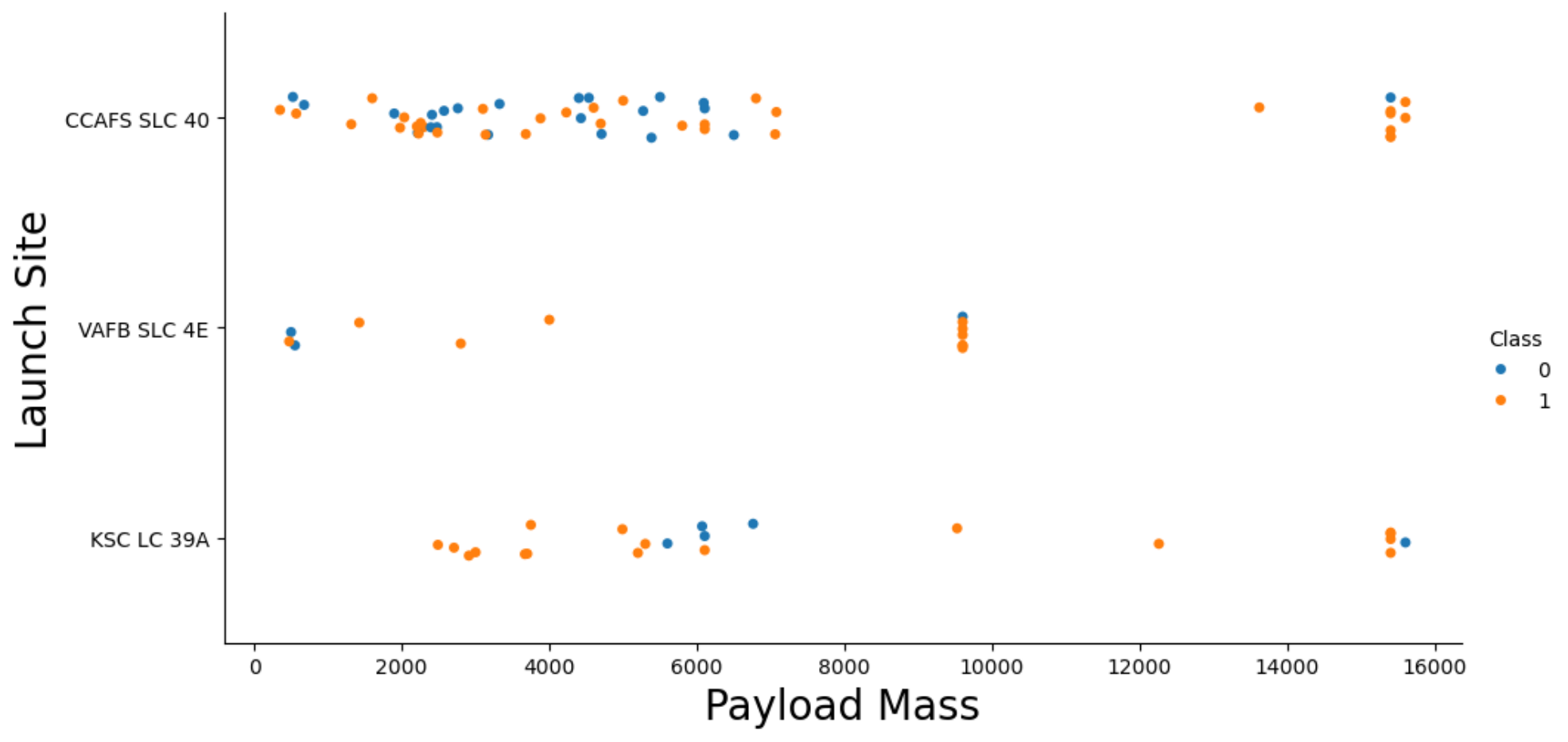
Insights drawn from EDA

Flight Number vs. Launch Site



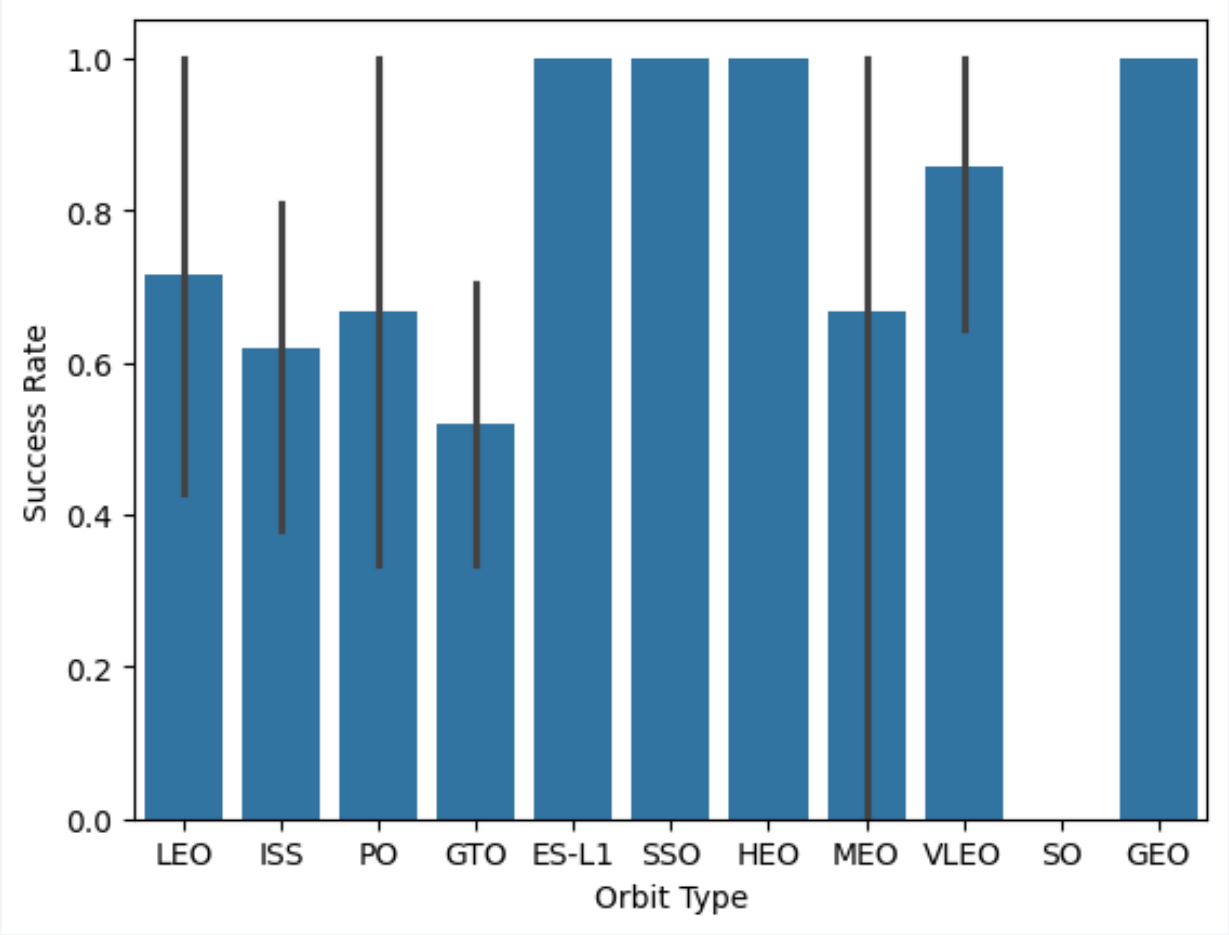
- For each launch site, the higher the flight number, the more likely it is to succeed
- There is no low flight number for site KSC LC 39Am, no medium flight number for CCAFS SLC 40, no high flight number for VAFC SLC 4E

Payload vs. Launch Site



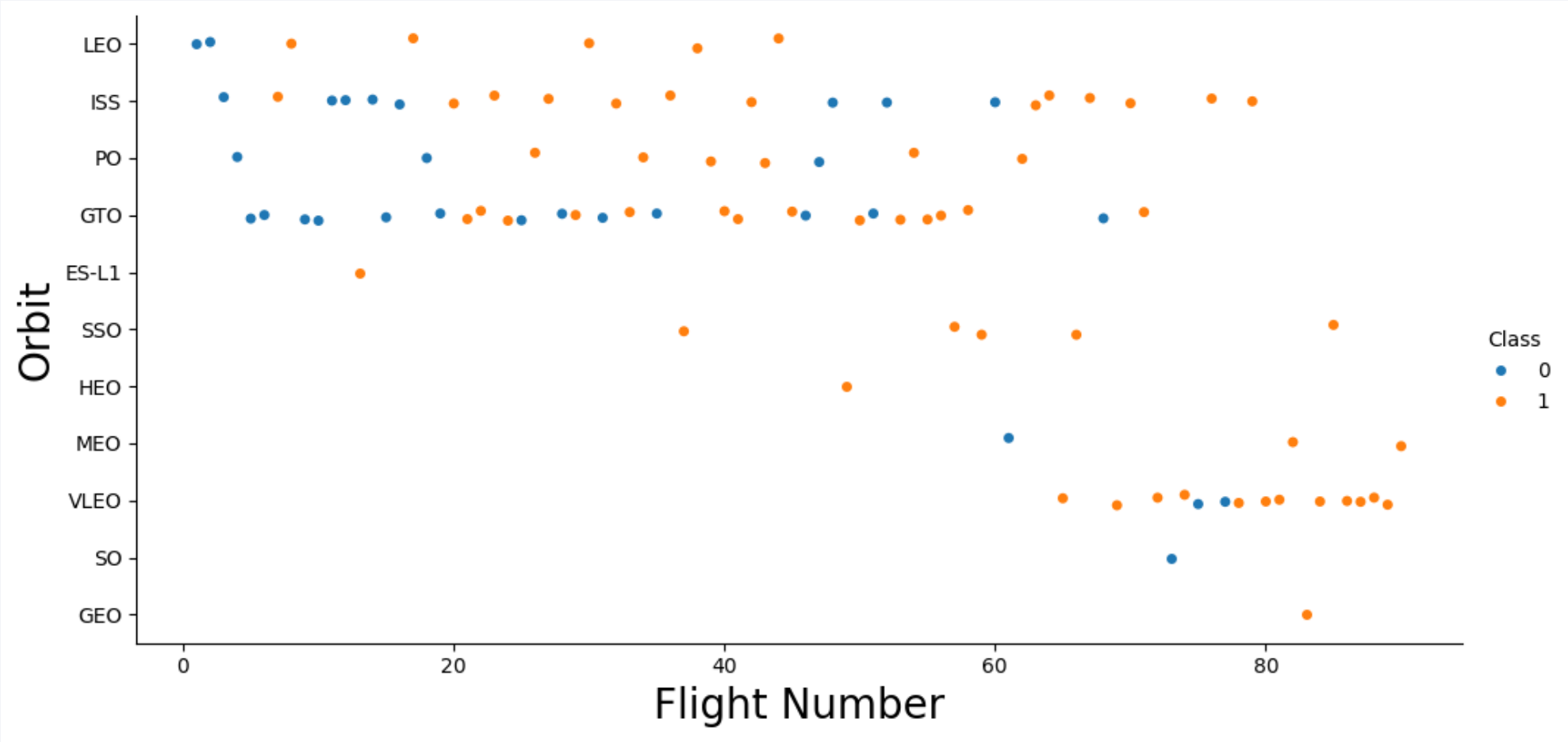
- Payloads over 12000 kg have good success, but only on CCAFS SLC 40 and KSL LC 39A launch sites

Success Rate vs. Orbit Type



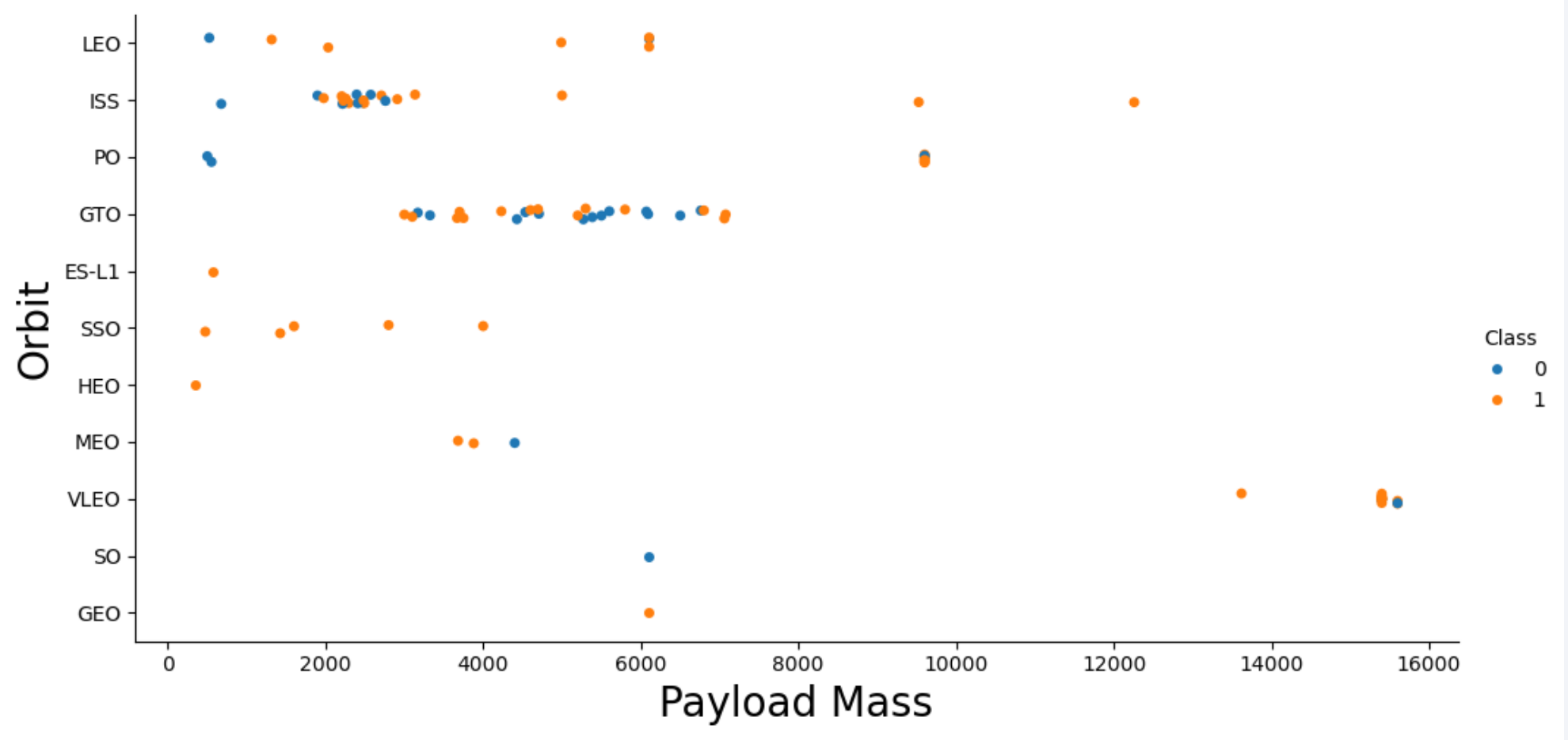
- ES-L1, SSO, HEO, GEO orbits have the best success rate

Flight Number vs. Orbit Type



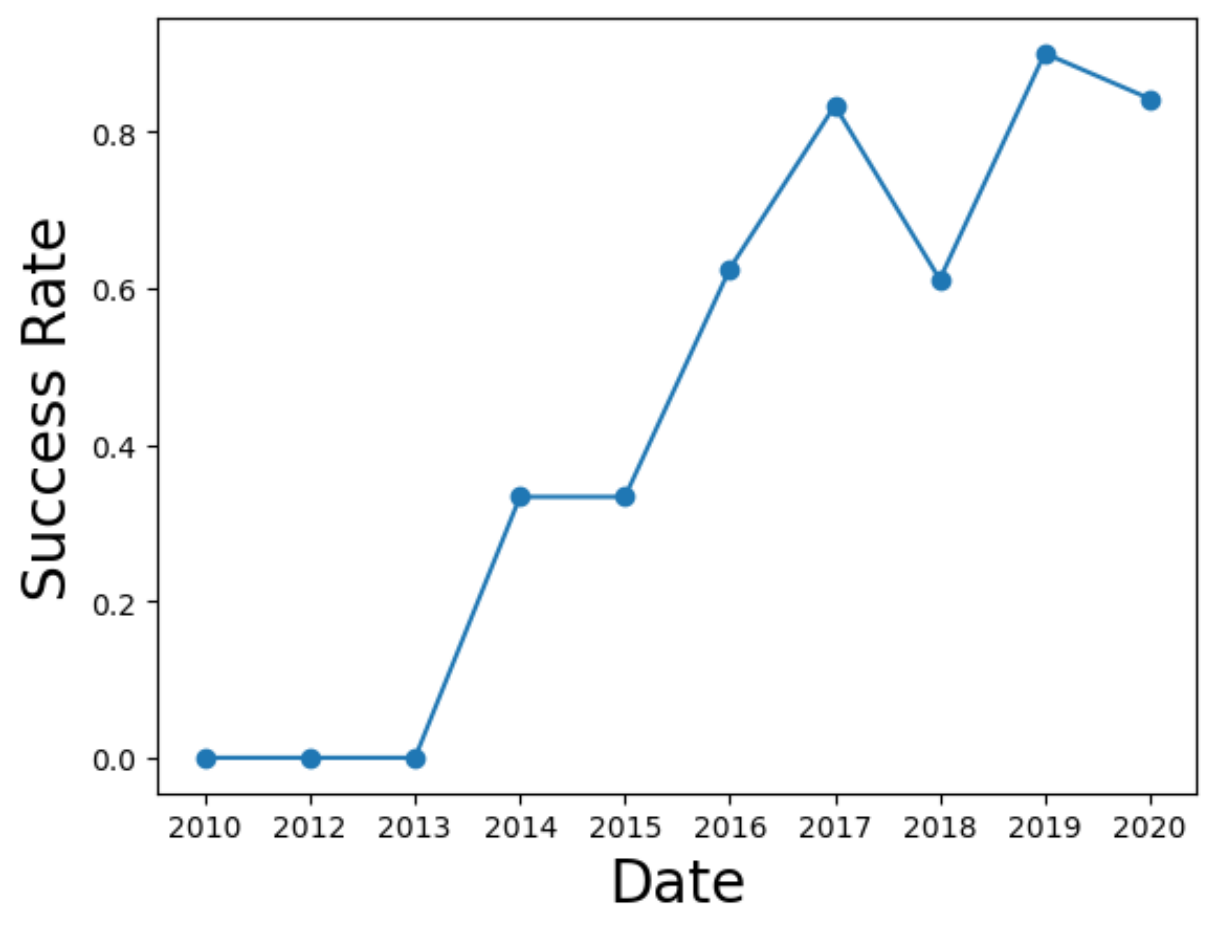
- Higher flight numbers are associated with higher rates of success
- Some orbits are assigned higher flight numbers, potentially due to recency

Payload vs. Orbit Type



- There is no strong relation between payload and orbit type
- GTO and ISS orbits see the most traffic

Launch Success Yearly Trend



- The success rate since 2013 shows an overall increasing trend

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

- The 4 launch sites

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No at
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No at
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No at

- Above shows 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- Above is the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Above is the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (ground pad)";
```

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

MIN(Date)
2015-12-22

- Above is the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT 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

- A list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT COUNT(*) FROM SPACEXTABLE
WHERE Mission_Outcome LIKE "Success%"
```

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

COUNT(*)
100

```
%%sql
SELECT COUNT(*) FROM SPACEXTABLE
WHERE Mission_Outcome LIKE "Failure%"
```

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

COUNT(*)
1

- Above left shows the total number of successful mission outcomes
- Above right shows the total number of failure mission outcomes

Boosters Carried Maximum Payload

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

- Above shows the names of the booster which have carried the maximum payload mass

2015 Launch Records

Month_Name	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- A list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

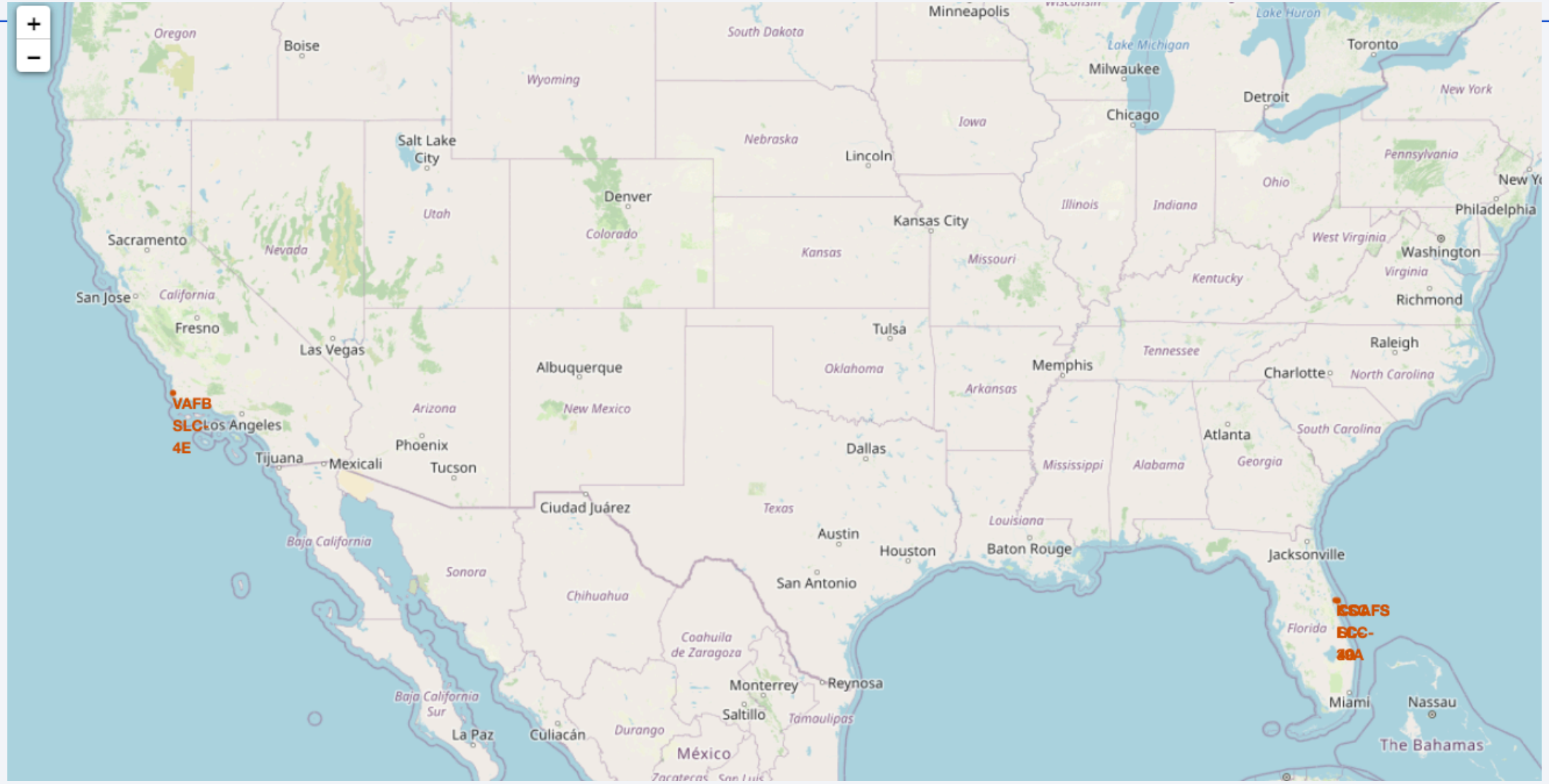
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Above ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

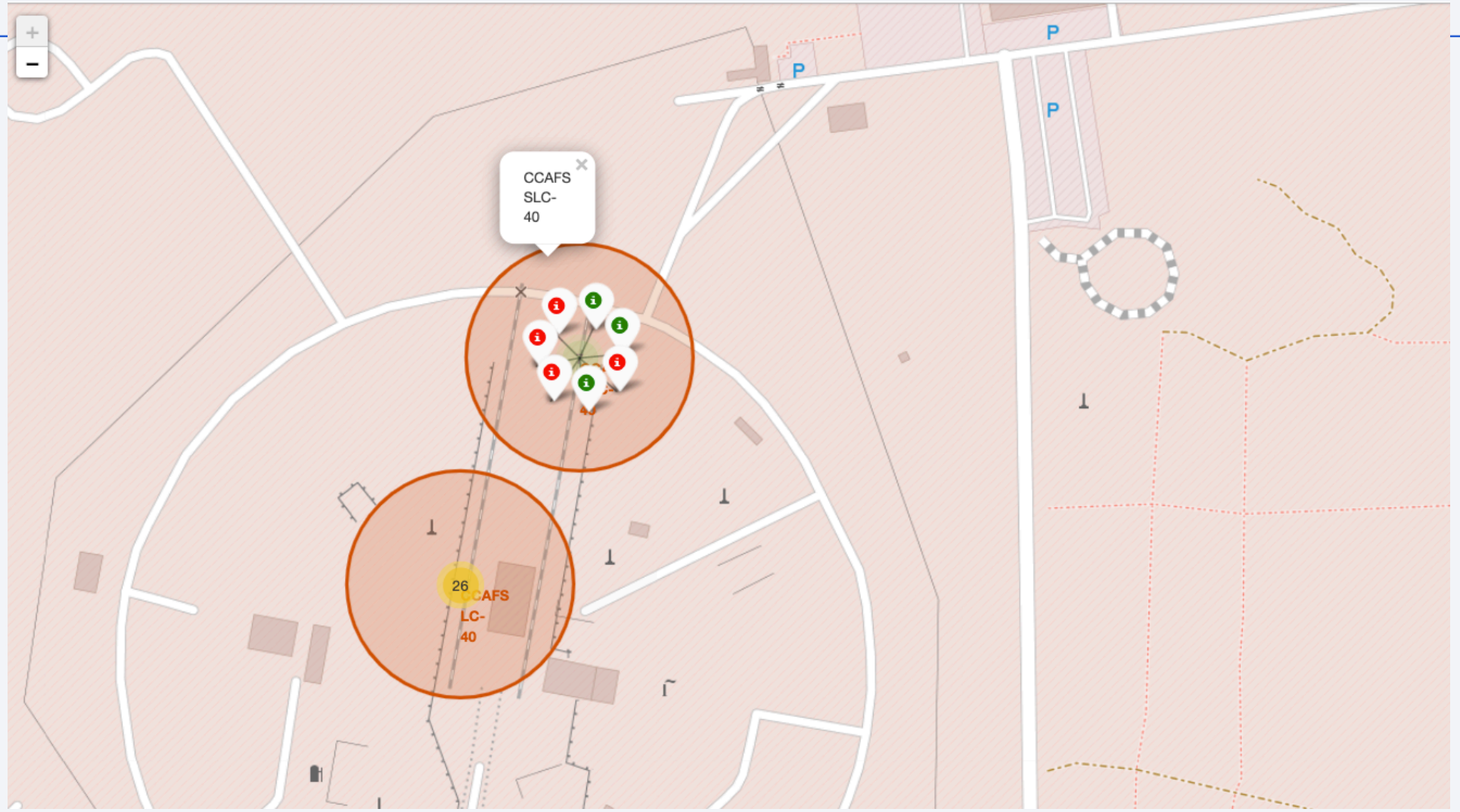
Launch Sites Proximities Analysis

All Launch Sites



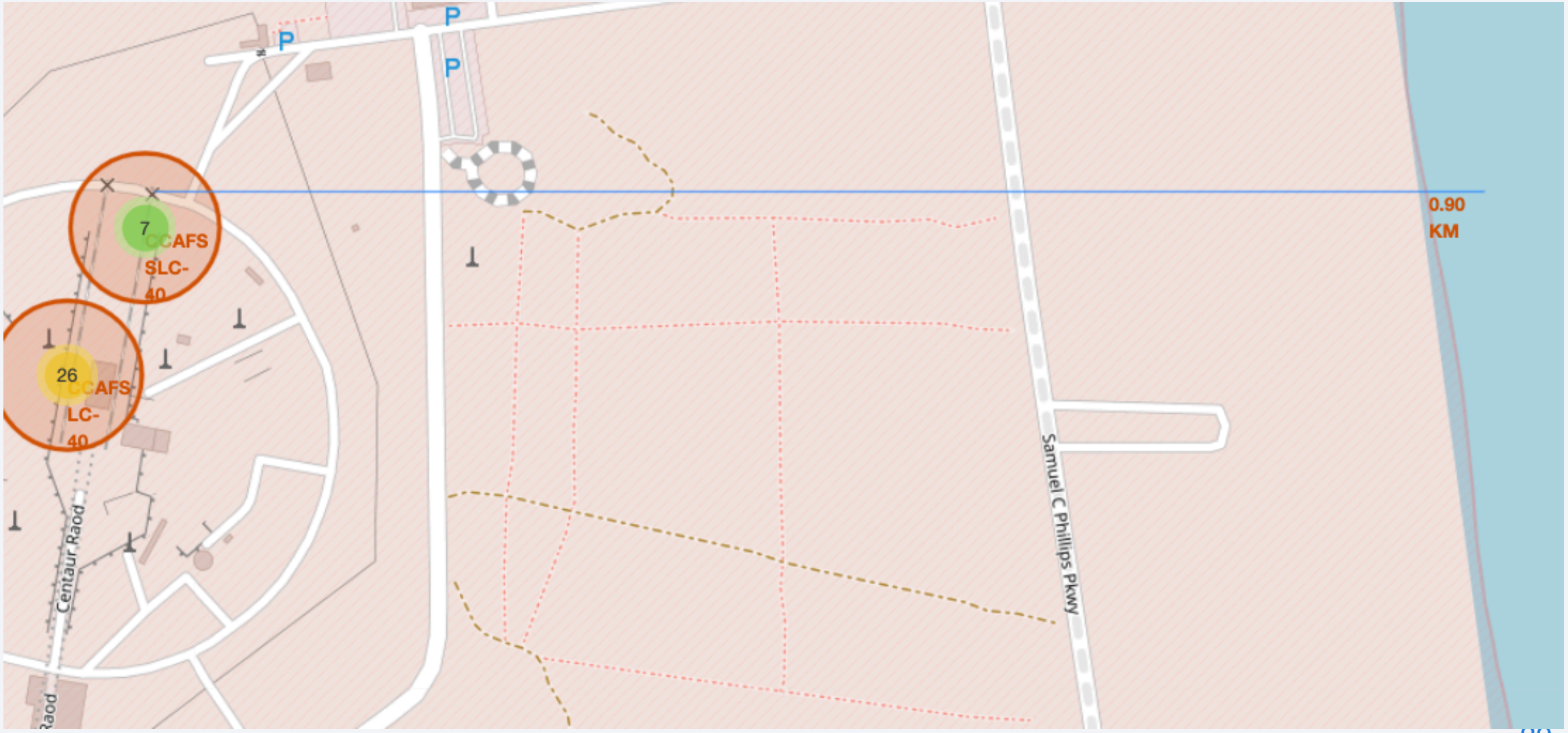
- All launch sites are at the coastal area of US

Launch Outcomes

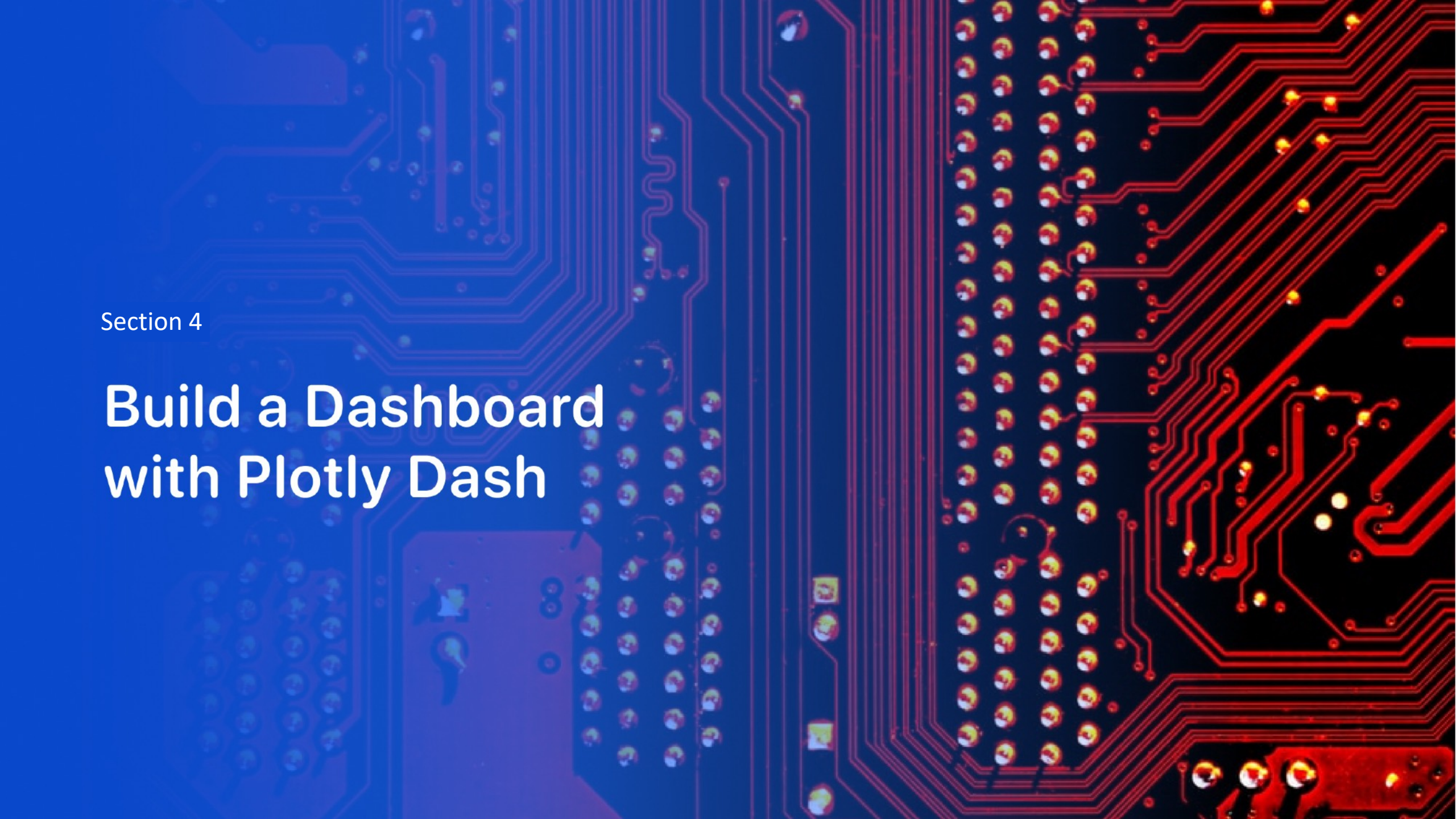


- Example of CCAFS SLC-40 launch outcomes, showing successes and failures

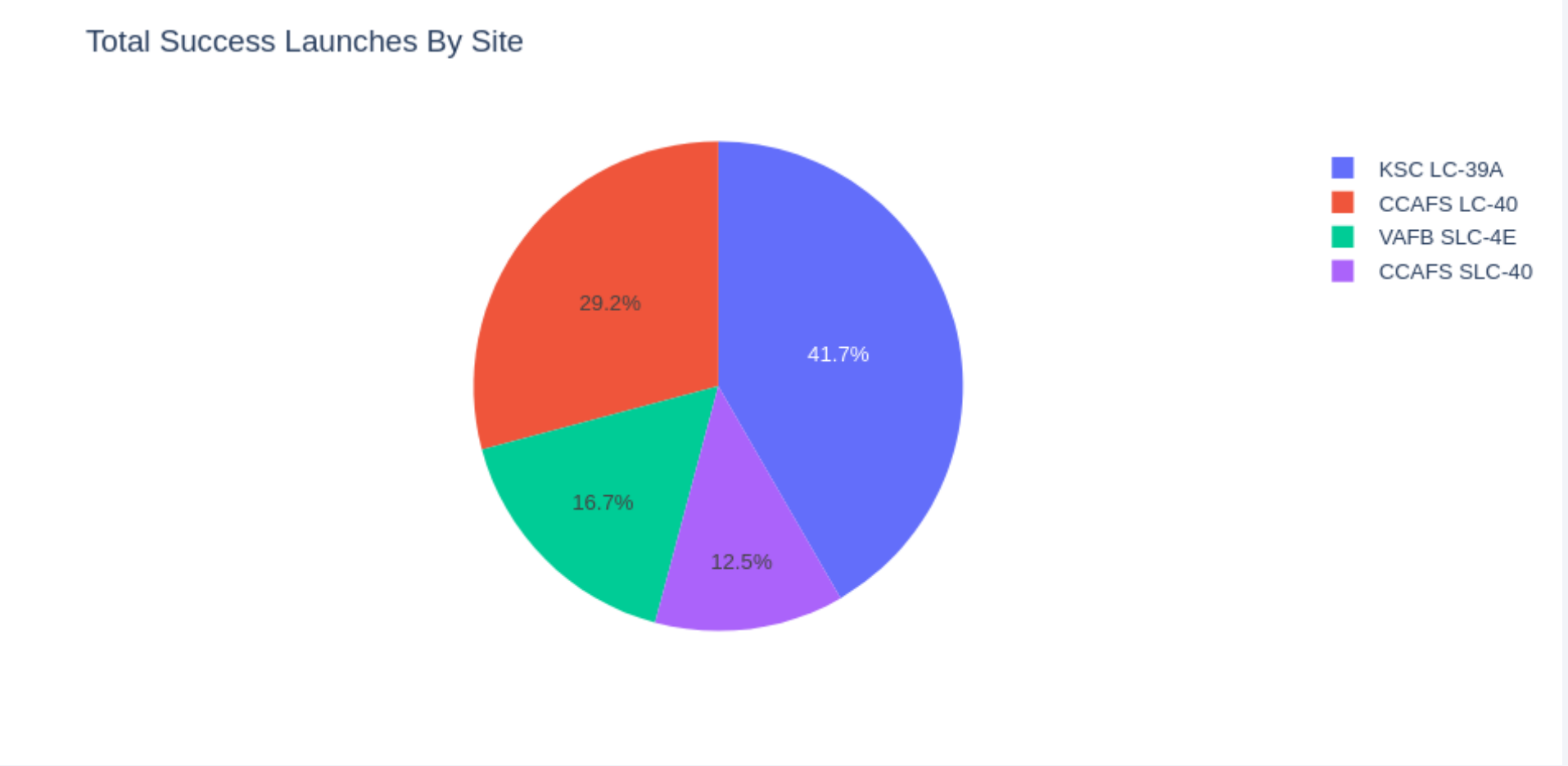
Launch Site Proximity



- The launch site shown is close to the coast line

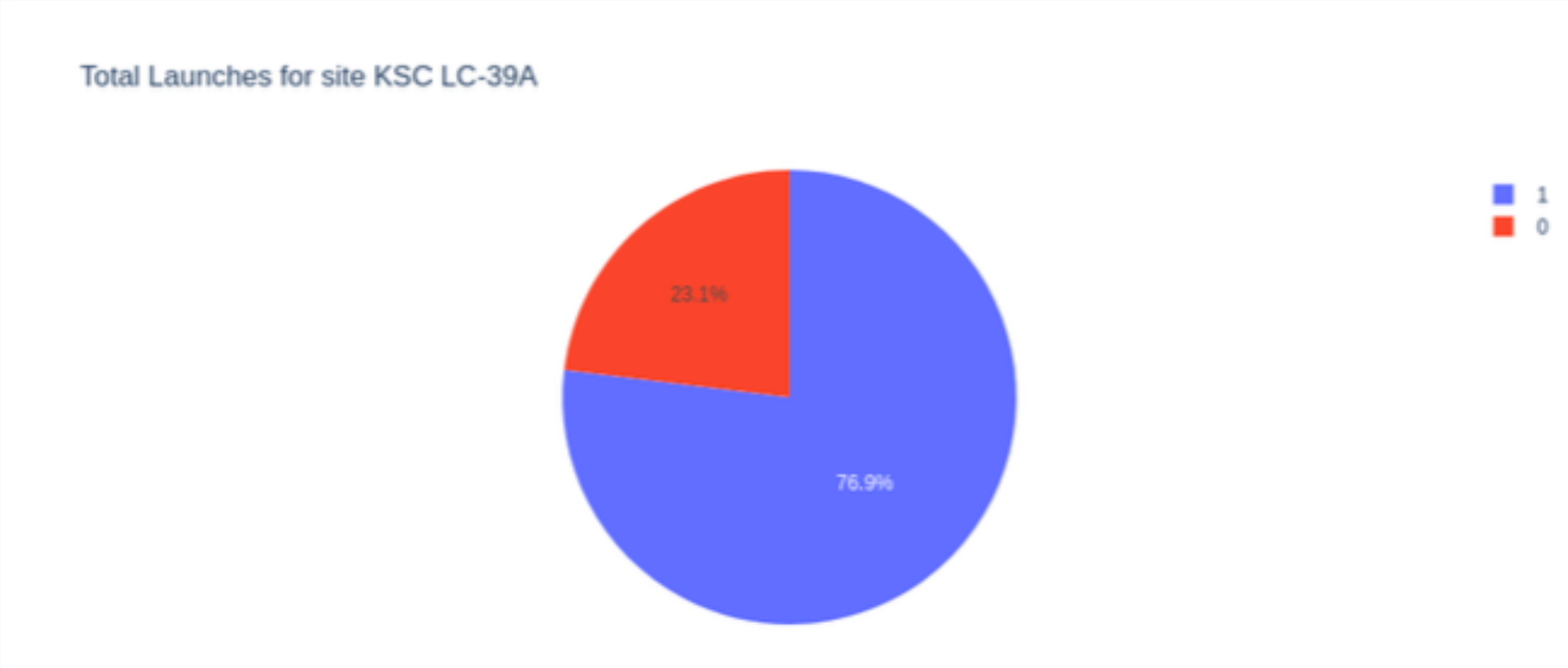

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



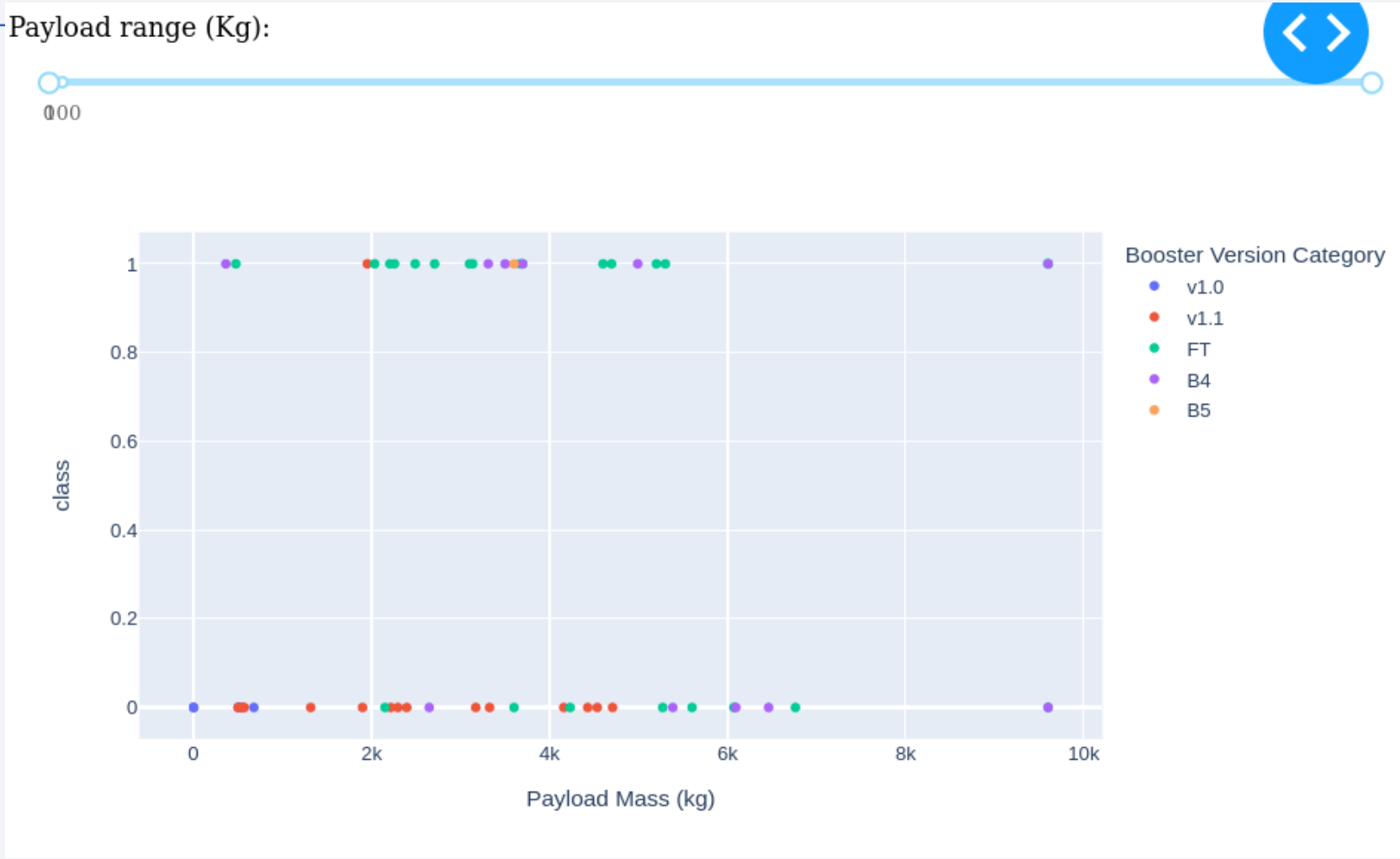
- KSC LC-39A sees the largest number of successful launches

Launch Success Ratio for KSC LC-39A



- About 77% of launches are successful on this site

Payload vs. Launch Outcome

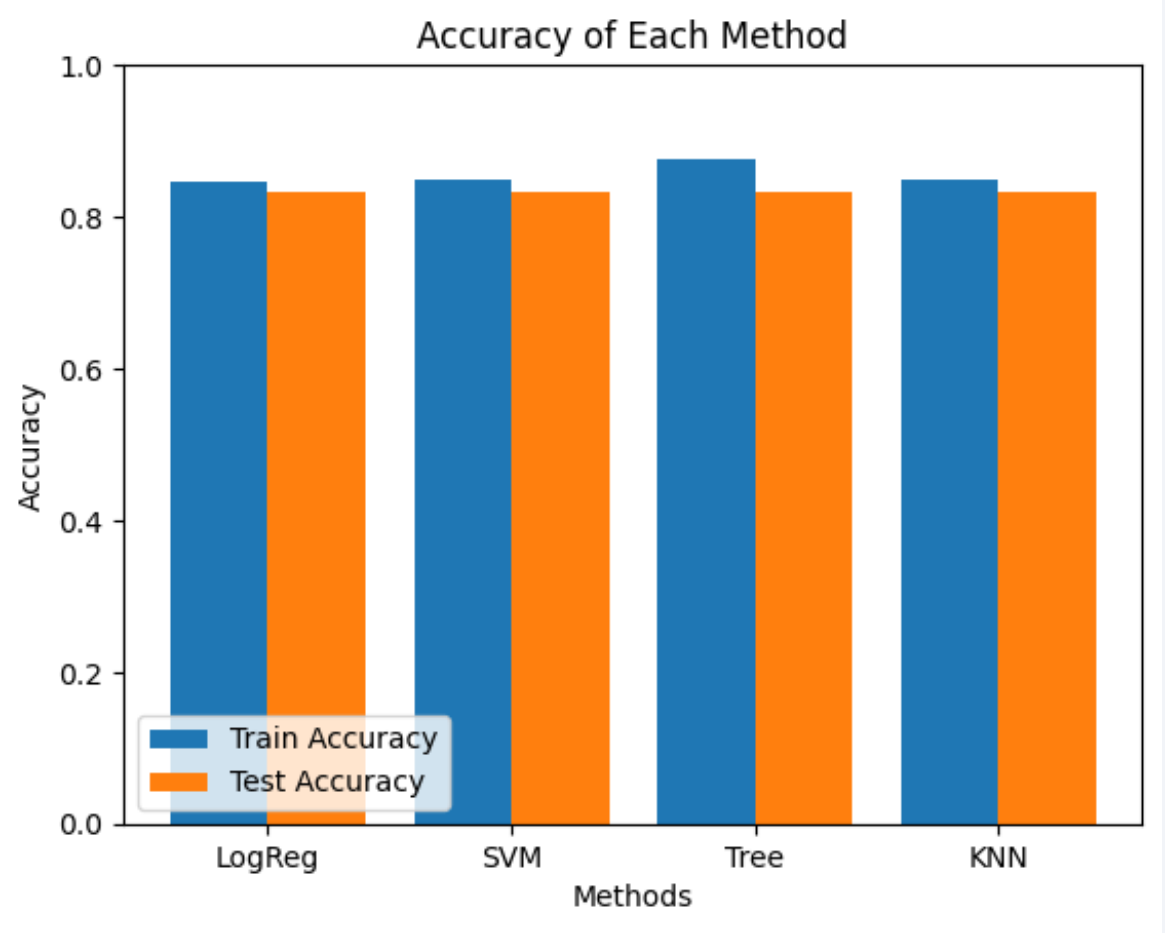


- The combination of under 6k-kg payload mass and FT booster has the largest success rate

Section 5

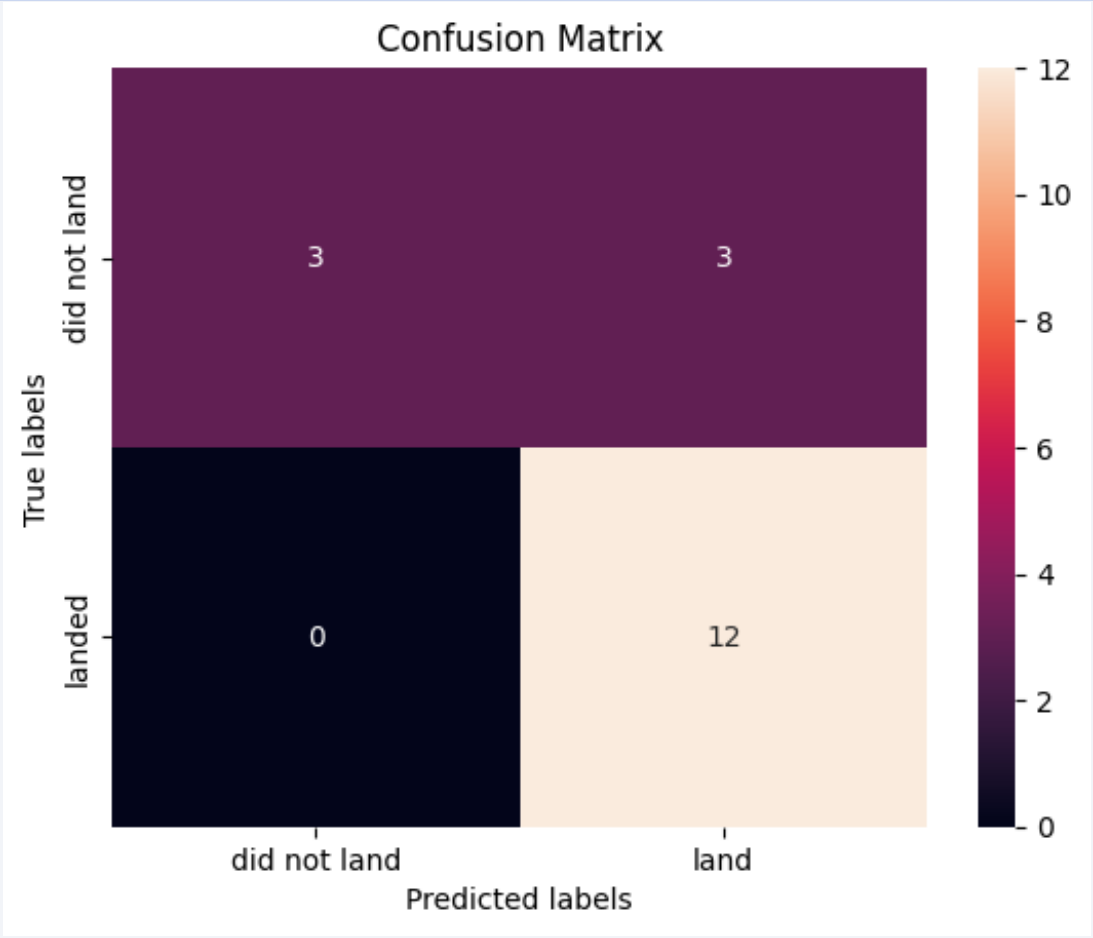
Predictive Analysis (Classification)

Classification Accuracy



- All models have the same test accuracy (though tree model's test accuracy fluctuate a bit), but tree has the highest training accuracy

Confusion Matrix



- The confusion matrix of the best performing model. Most predictions were correct, only 3 did not land cases were incorrectly predicted as landed

Conclusions

- We looked at data with SQL, Python, processed with pandas, displayed with pandas, Folium, and Plotly, and trained with various machine learning models
- The launch site with the biggest share of success is KSC LC-39A
- Average landing rate improves over time
- The test accuracy of different models are similar, but Decision Tree has the highest training accuracy

Appendix

- All data sources: https://github.com/chuanyinn/coursera/tree/main/0_course_10_Applied-Data-Science-Capstone

Thank you!