WINNING SPACE RACE WITH DATA SCIENCE

IBM DATA SCIENCE CAPSTONE PROJECT

Adli Zakaria
June 2024

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection through API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard using Plotly
- Predictive Analysis using Machine Learning

Result Summary

- Exploratory Data Analysis Results
- Interactive Analytics (Plots and Dashboard)
- Machine Learning Prediction Results

Introduction

Project background and context

A new rocket company named **SpaceY** would like to compete with SpaceX founded by Billionaire industrialist Allon Musk.

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 Space X can reuse the first stage.

In this project we will determine if the first stage will land, so that we can determine the cost of launch. This information can be used by **SpaceY** to bind against SpaceX for a rocket launch.

Problems

What is the features that affect the first landing success?

Section 1

Methodology

Methodology

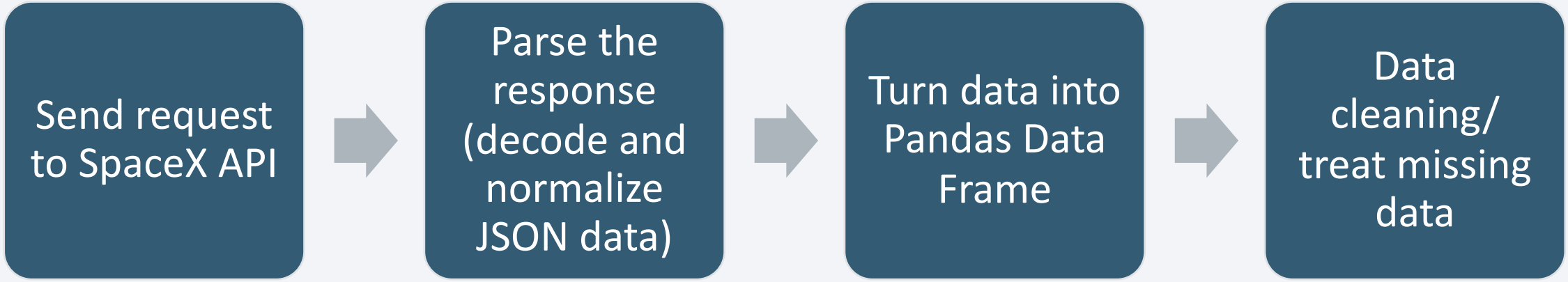
Executive Summary

- Data collection done using the following methodology:
  - Rocket launch data from SpaceX API
  - Falcon 9 historical launch records from Wikipedia using Web Scraping
- Data wrangling performed by cleaning, and preparing data for further analysis.
- Exploratory data analysis (EDA) done using visualization and SQL
- Interactive visual analytics done using Folium and Plotly Dash
- Predictive analysis done using classification models:
  - Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor

Data Collection

- Rocket launch data are collected from the SpaceX API.
- Falcon 9 historical launch records are collected from SpaceX Wikipedia page

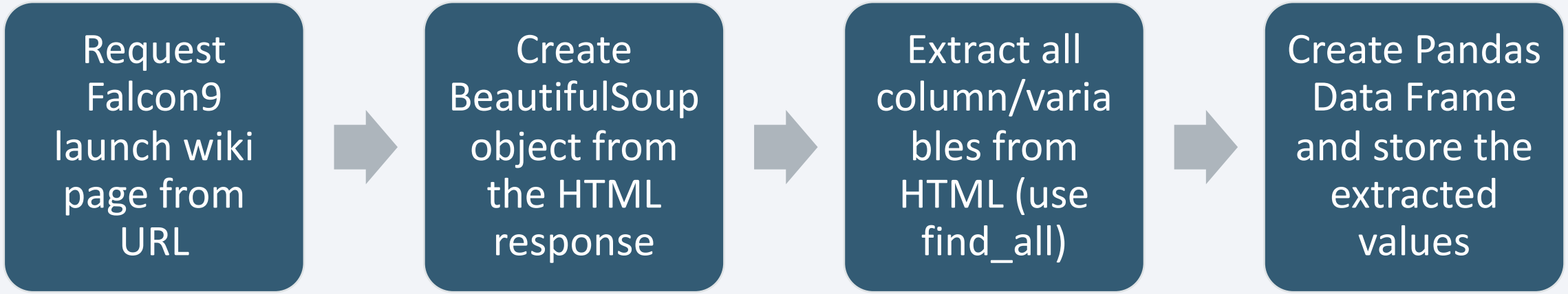
Data Collection – SpaceX API



GitHub URL of the completed SpaceX API calls notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone%201.1%20-%20Data%20collection%20API%20(spaceX).ipynb

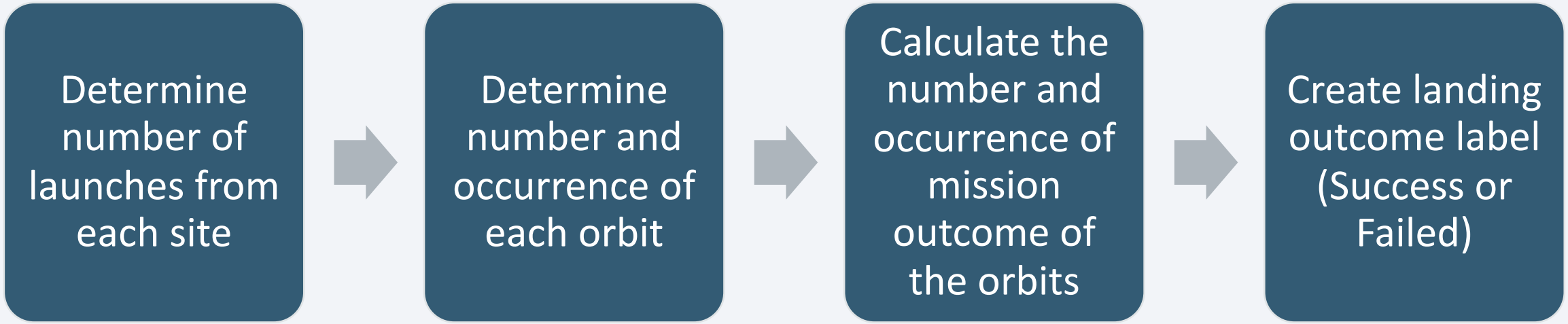
Data Collection - Scraping



GitHub URL of the completed web scraping notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 1.2 - Data collection with webscraping.ipynb

Data Wrangling



GitHub URL of the completed Data Wrangling notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 2.1 - Data wrangling (spaceX).ipynb

EDA with Data Visualization

EDA and Feature Engineering done using 'Pandas' and 'Matplotlib' by visualizing the data to see any trend and relationship between the variables.

To get the insight from the data, the following plots were used:

- Flight Number vs. Pay Load Mass overlaid with the Outcome of the launch
- Scatter plot for relationship between Flight Number and Launch Site
- Scatter plot for relationship between Payload and Launch Site
- Bar plot for success rate of each orbit type
- Scatter plot for relationship between Flight Number and Orbit type
- Line plot for launch success yearly trend (Year vs Average Success Rate)

GitHub URL of the completed EDA with data visualization notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 4.1 - EDA with Visualization.ipynb

EDA with SQL

Load dataset into corresponding tables in database

Use SQL queries to get data about:

- The names of unique launch sites in the space mission.
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The total number of successful and failure mission outcomes
- The failed landing outcomes in drone ship, their booster version and launch site names.

GitHub URL of the completed EDA with SQL notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 3.1 - EDA with SQL.ipynb

Interactive Map with Folium

- Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- All launch sites are marked, and map objects such as markers, circles, lines are added to mark the success or failure of launches for each site on the folium map.
- Using the color-labeled marker clusters, launch sites that have relatively high success rate are identified.
- Distances between a launch site to its proximities are calculated

GitHub URL of the Interactive Map notebook:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 5.1 - Interactive Visual Analytics with Folium (SpaceX).ipynb

Dashboard with Plotly Dash

An interactive dashboard is built with Plotly dash.

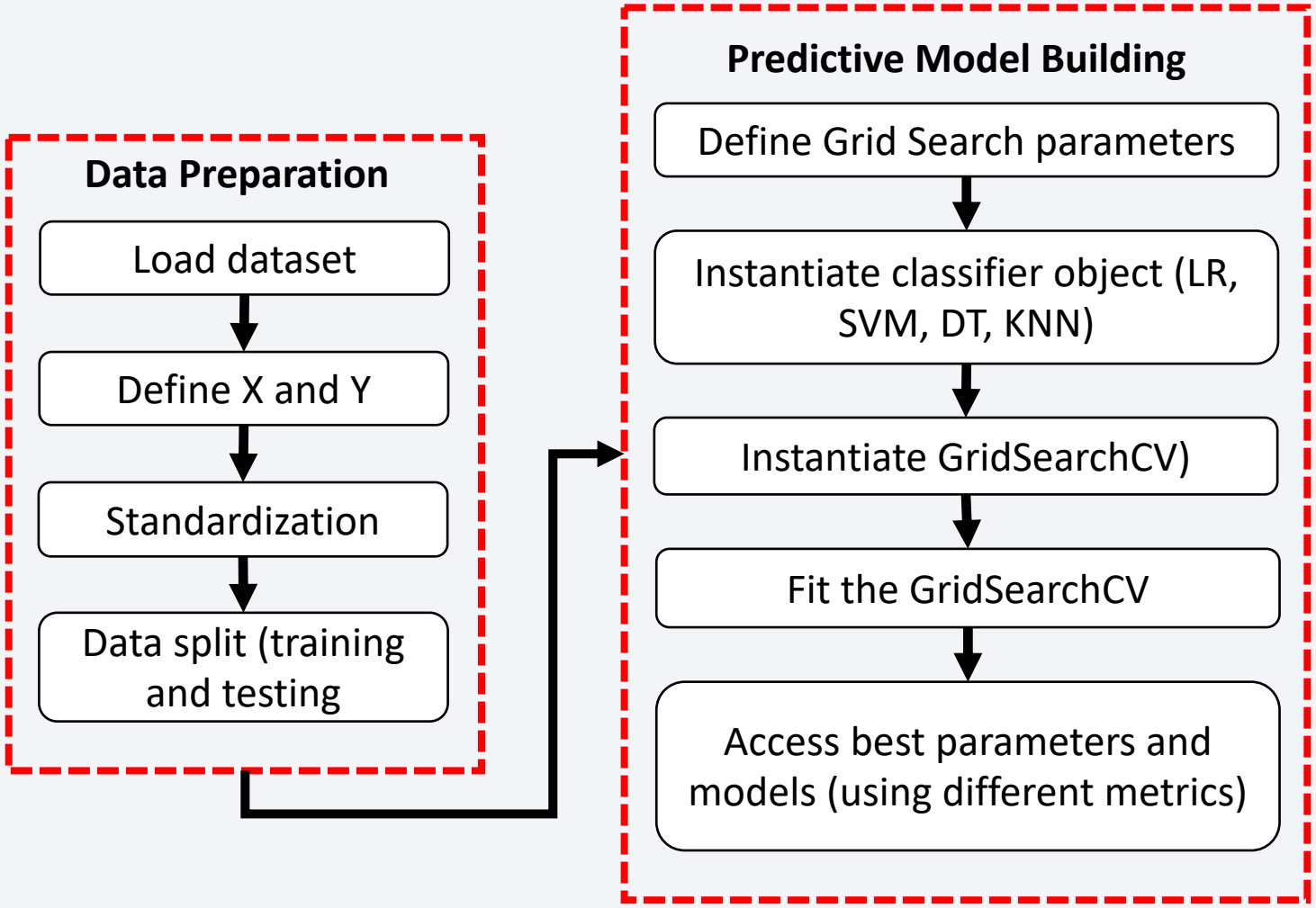
In the dashboard, we present the following:

- Pie charts showing the total launches by a certain sites.
- Scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

GitHub URL:

https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 5.2 - spacex_dash_app.py

Predictive Analysis (Classification)



Classifier models used:

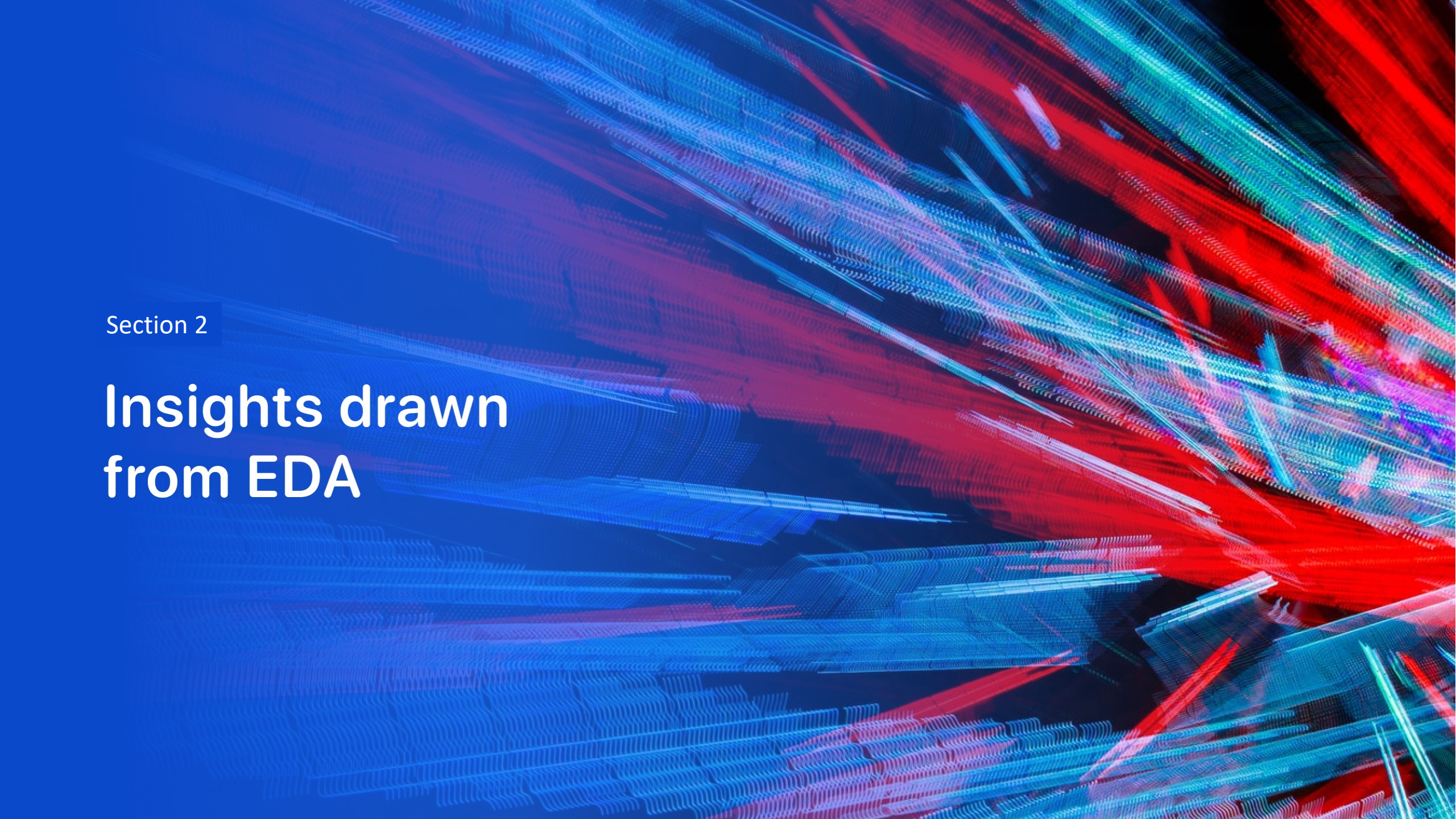
1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. K Nearest Neighbor

The GitHub URL:

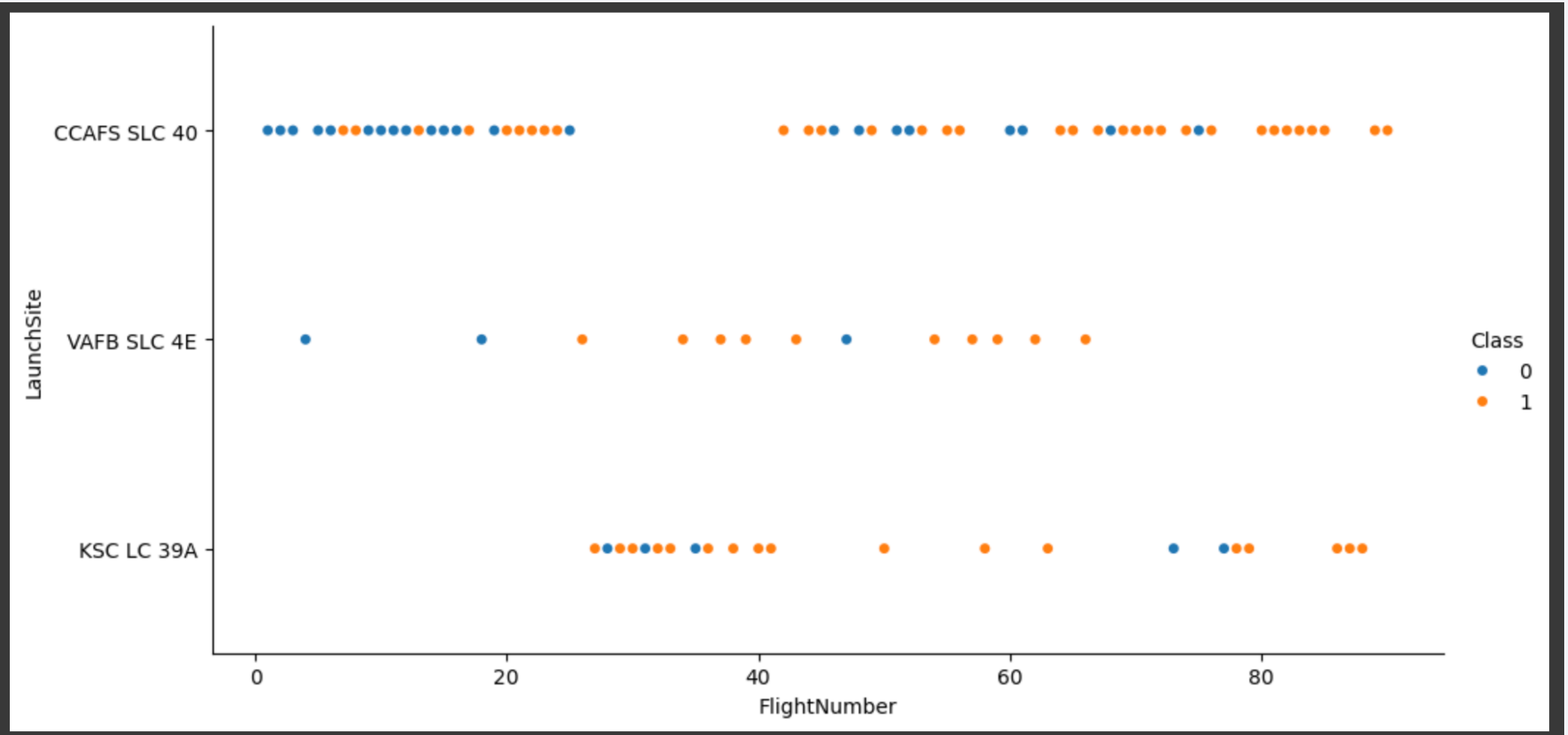
https://github.com/adlizakaria/IBM-Data-Science-Capstone-Project/blob/main/Capstone 6.1 - Machine Learning Prediction (SpaceX).ipynb

Results

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

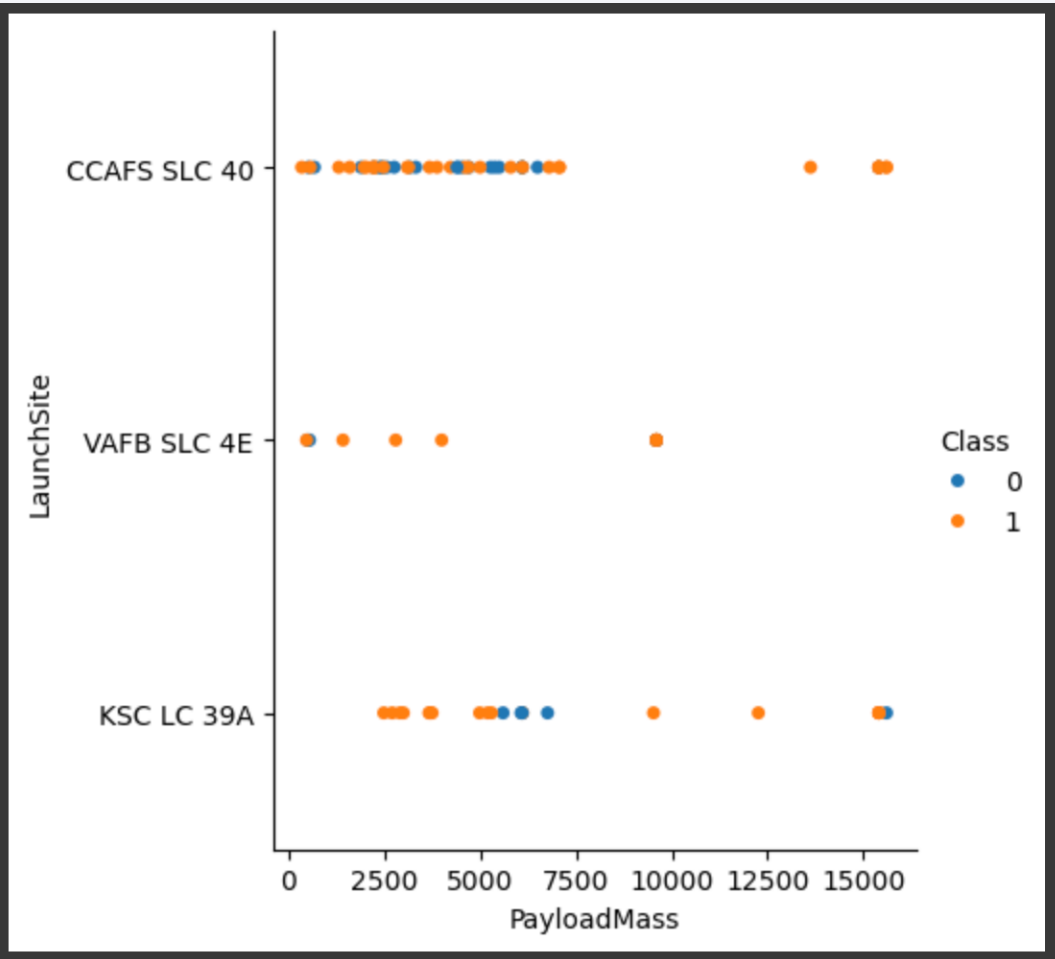

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

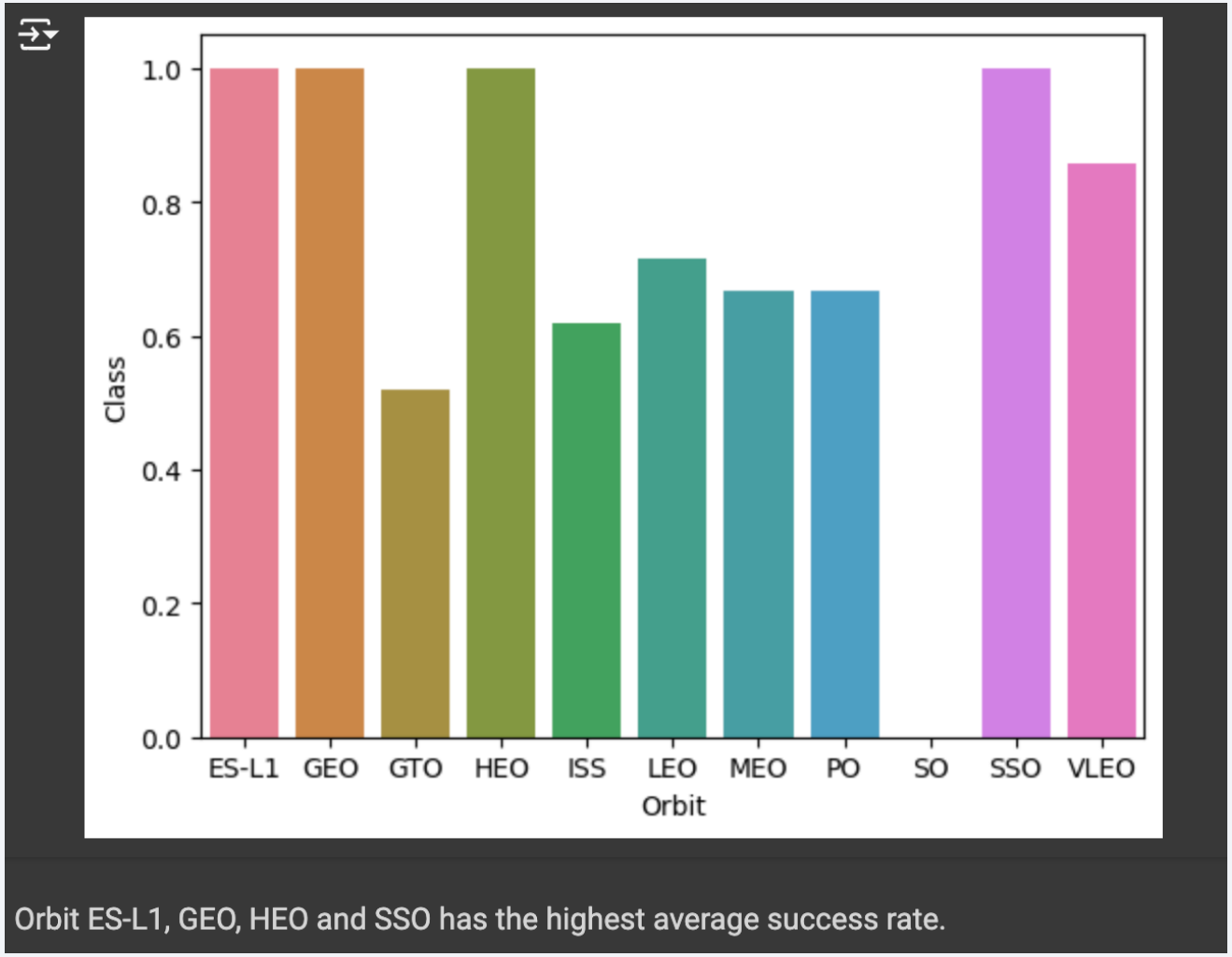
Payload vs. Launch Site



Based on the Payload Vs. Launch Site scatter point chart,for the VAFB-SLC launchsite, there are no rockets launched for heavypayload mass(greater than 10000).

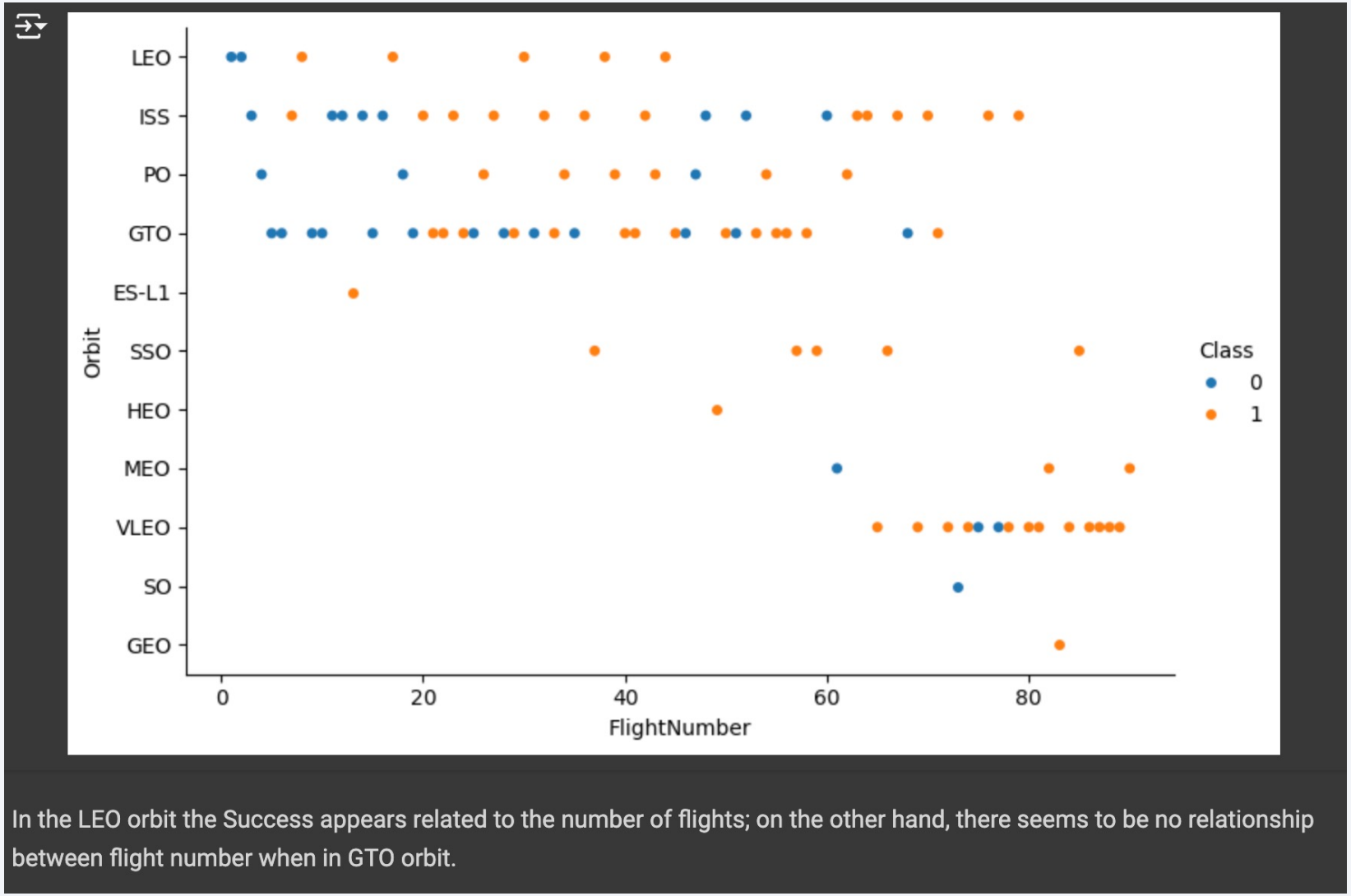
For VAFB-SLC launch site, the are no rockets launched for *heavy payload mass* (greater than 1000).

Success Rate vs. Orbit Type



- Orbit type ES-L1, GEO, HEO and SSO has the highest average success rate.
- Orbit type SO has 0% success rate.

Flight Number vs. Orbit Type

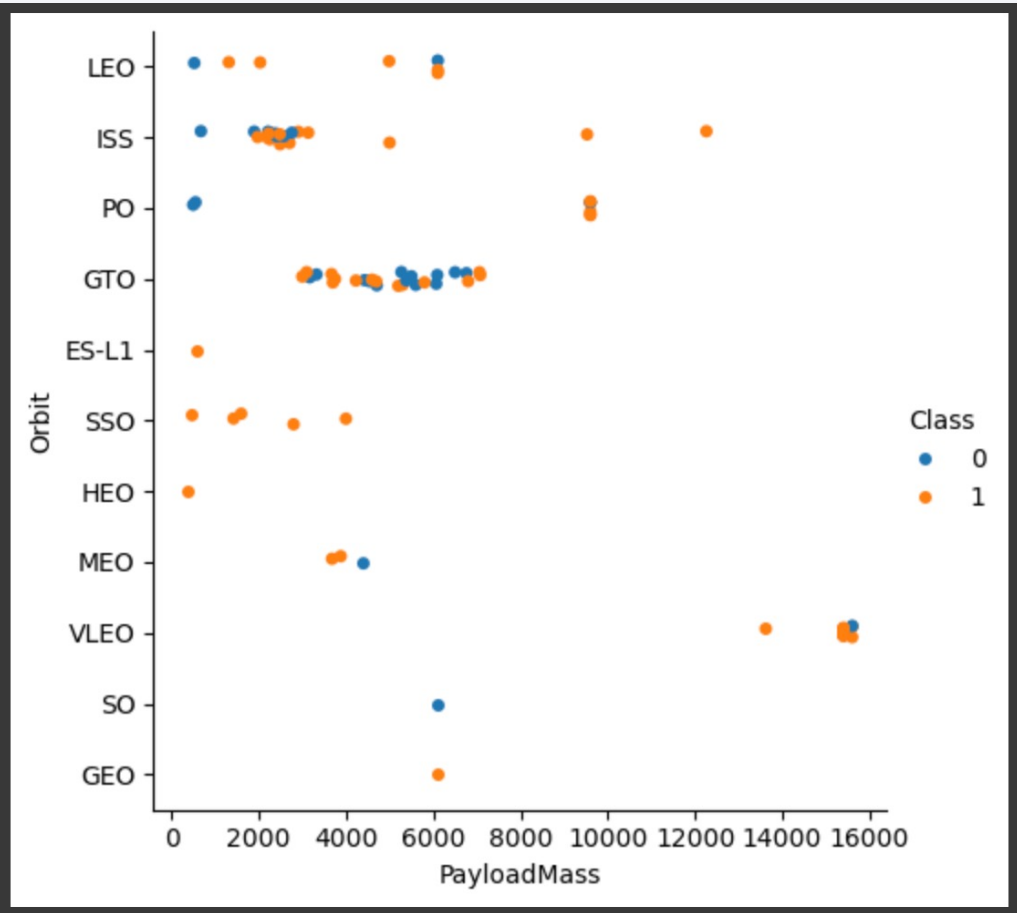


In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

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

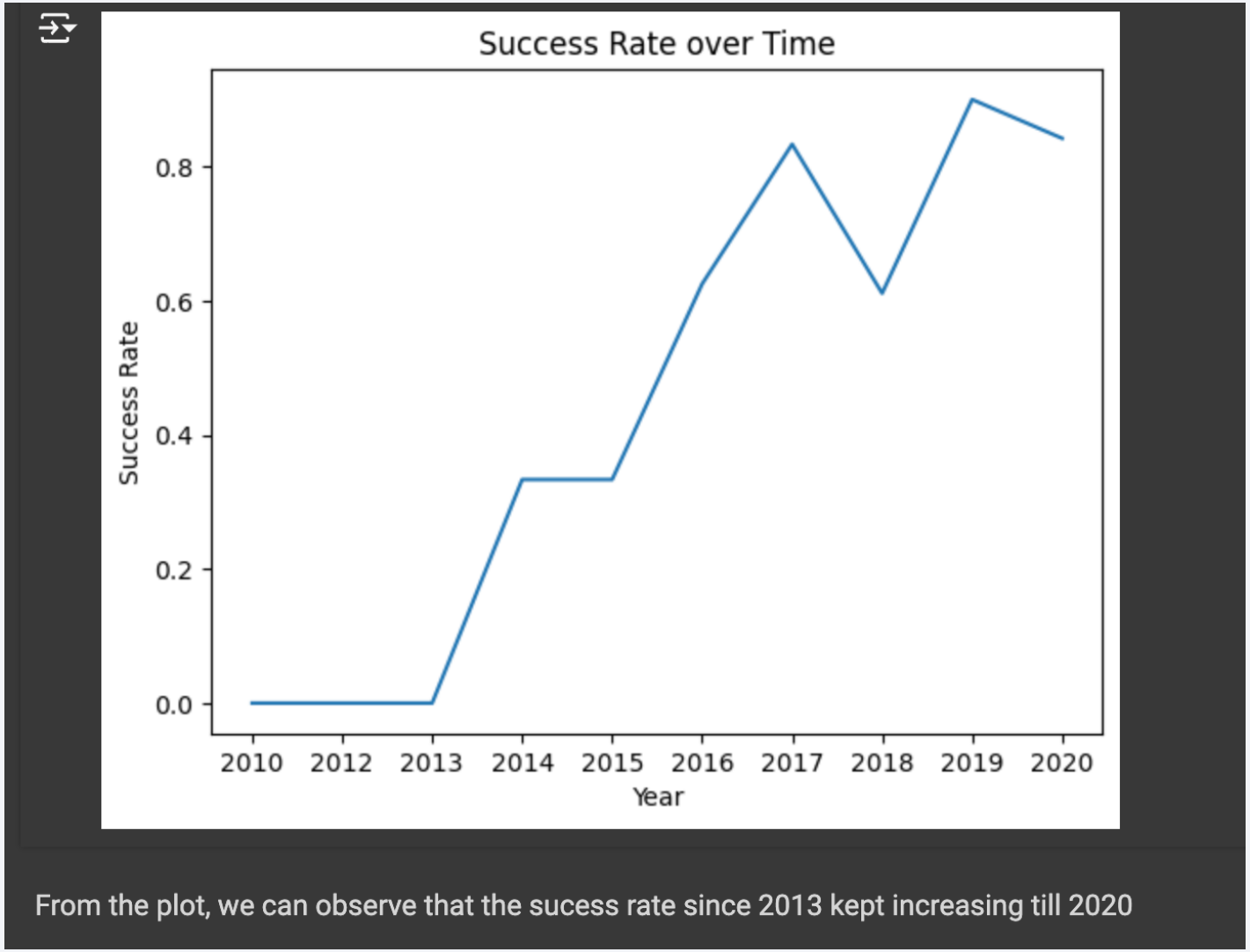


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend



The success rate is increasing since 2013 until 2020

All Launch Site Names

```
[ ]  # to display the names of unique launch sites
     %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

Based on the query result, there are 4 different launch sites available from the record.

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

The 5 earliest launches from site names begin with ‘CCA’ are from CCAAFS LC-40.

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
```

The total payload mass carried by boosters from NASA is 45,596 kg.

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
```

The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg

First Successful Ground Landing Date

```
# date for the first successful landing outcome in ground pad

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

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

MIN(Date)
2015-12-22

Date of the first successful landing outcome on ground pad is on 22 December 2015.

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_>4000 \
AND PAYLOAD_MASS__KG_<6000;
```

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

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

List 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 Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

The total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ \
FROM SPACEXTABLE \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

List of the names of the booster which have carried the maximum payload mass.

2015 Launch Records

```
%sql SELECT SUBSTRING(Date,6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE \
WHERE SUBSTRING(Date,1,4)='2015' AND Landing_Outcome = "Failure (drone ship)"
```

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

List of the 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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Frequency FROM SPACEXTABLE \
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY COUNT(Landing_Outcome) DESC;
```

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

Landing_Outcome	Frequency
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

List of the landing outcomes between 2010-06-04 and 2017-03-20 ranked by the frequency in descending order.

The most common landing outcome is “No attempt”

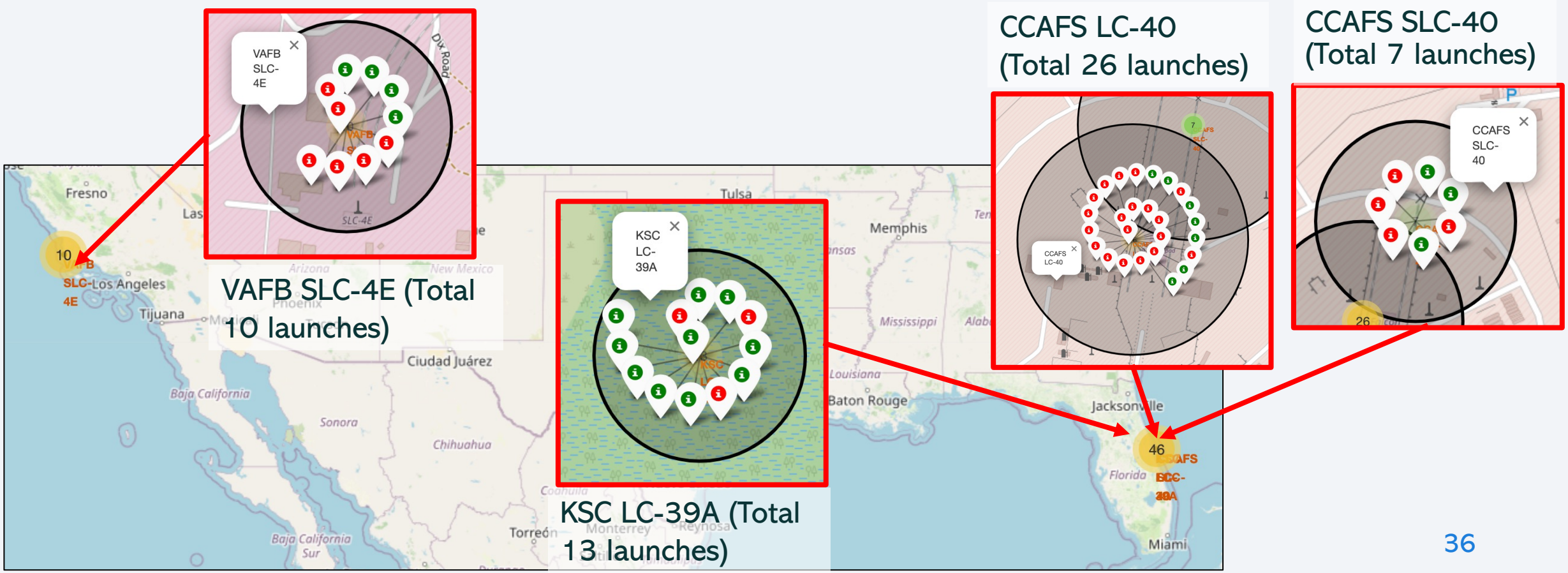

Section 3
Launch Sites
Proximities Analysis

Folium Map: Launch Sites

Map with marked launch sites' location

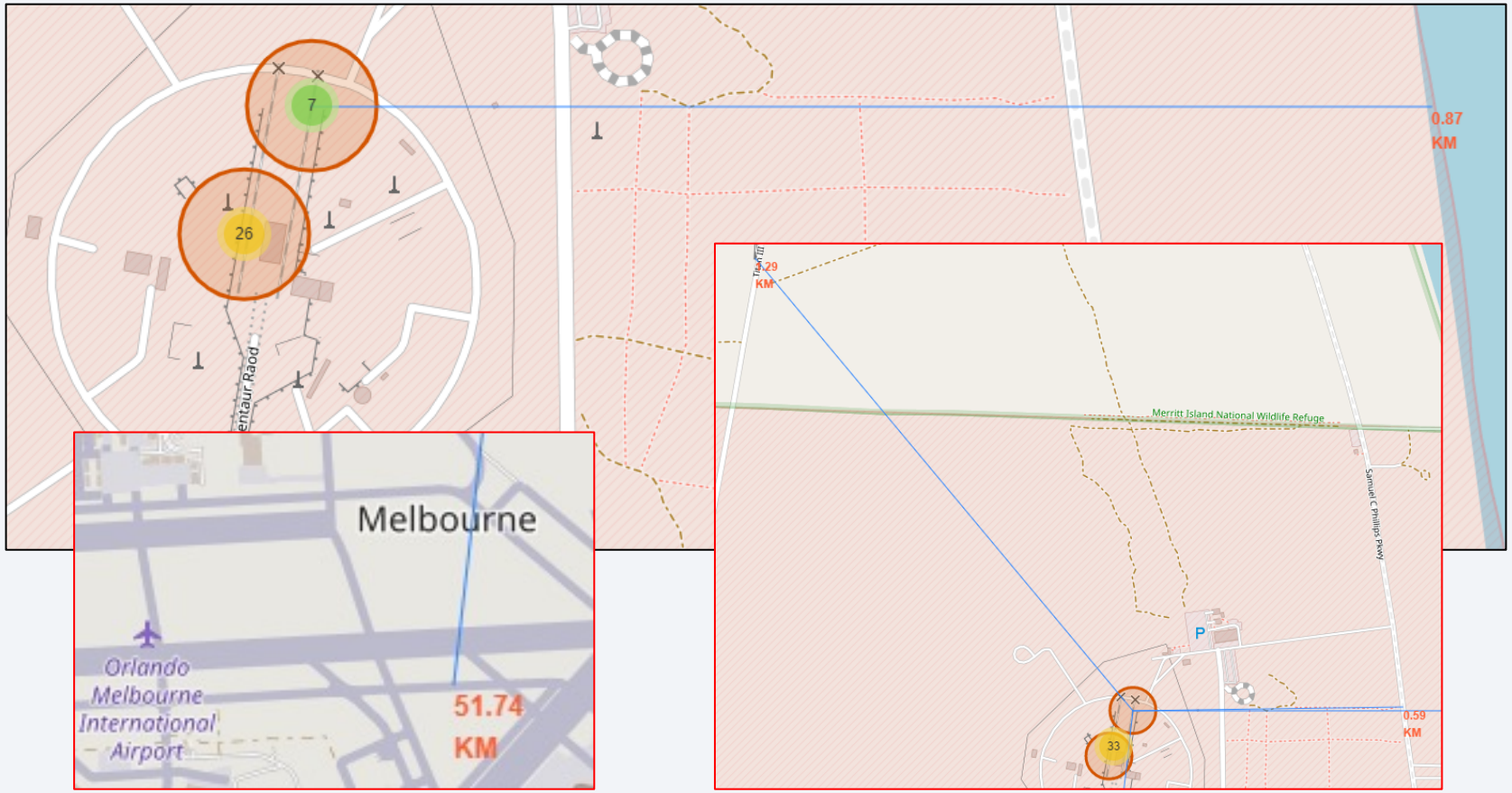
Success and Failed Launches for Each Site

Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.

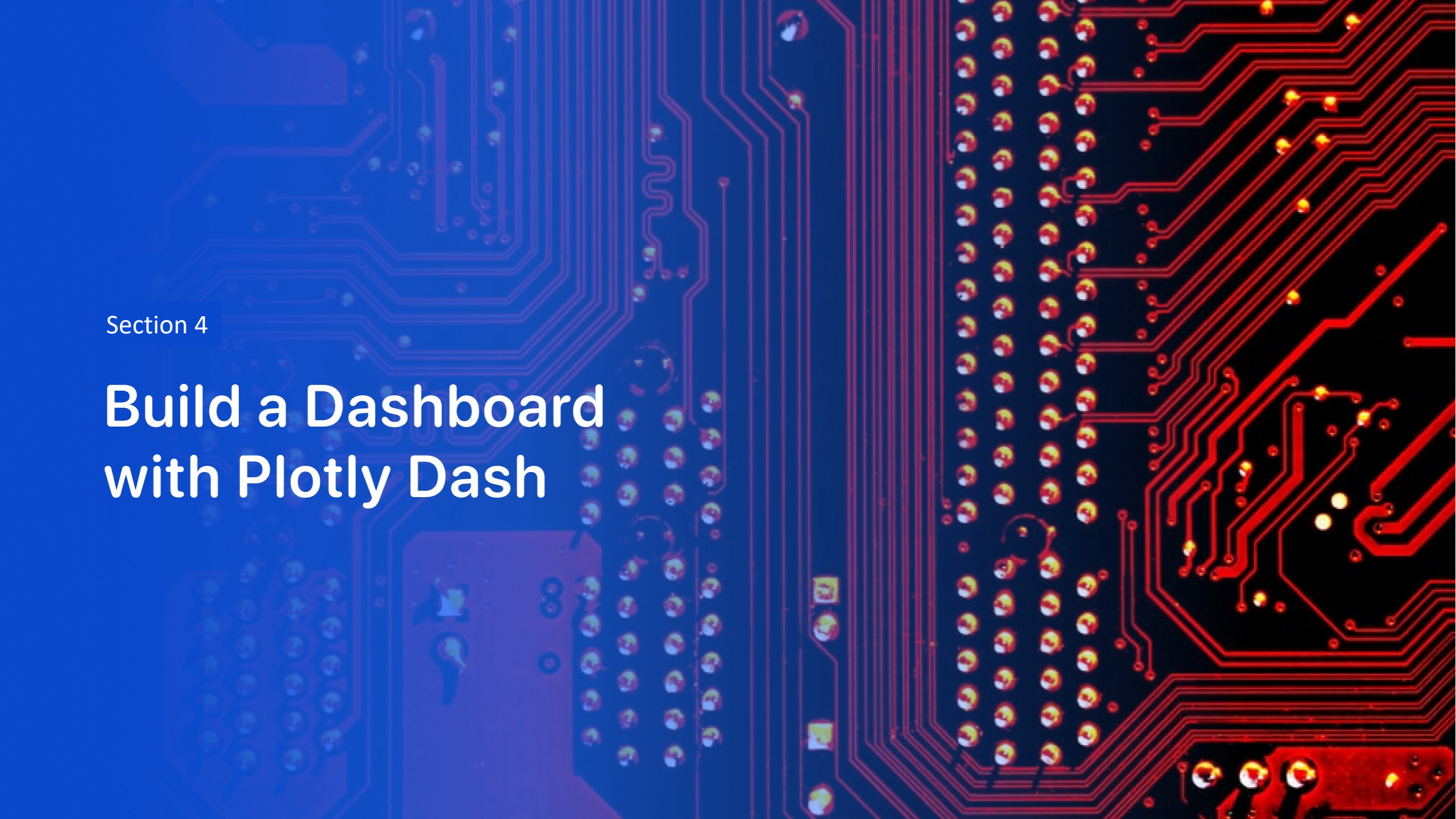


Proximity of Launch Sites to other Point of Interest

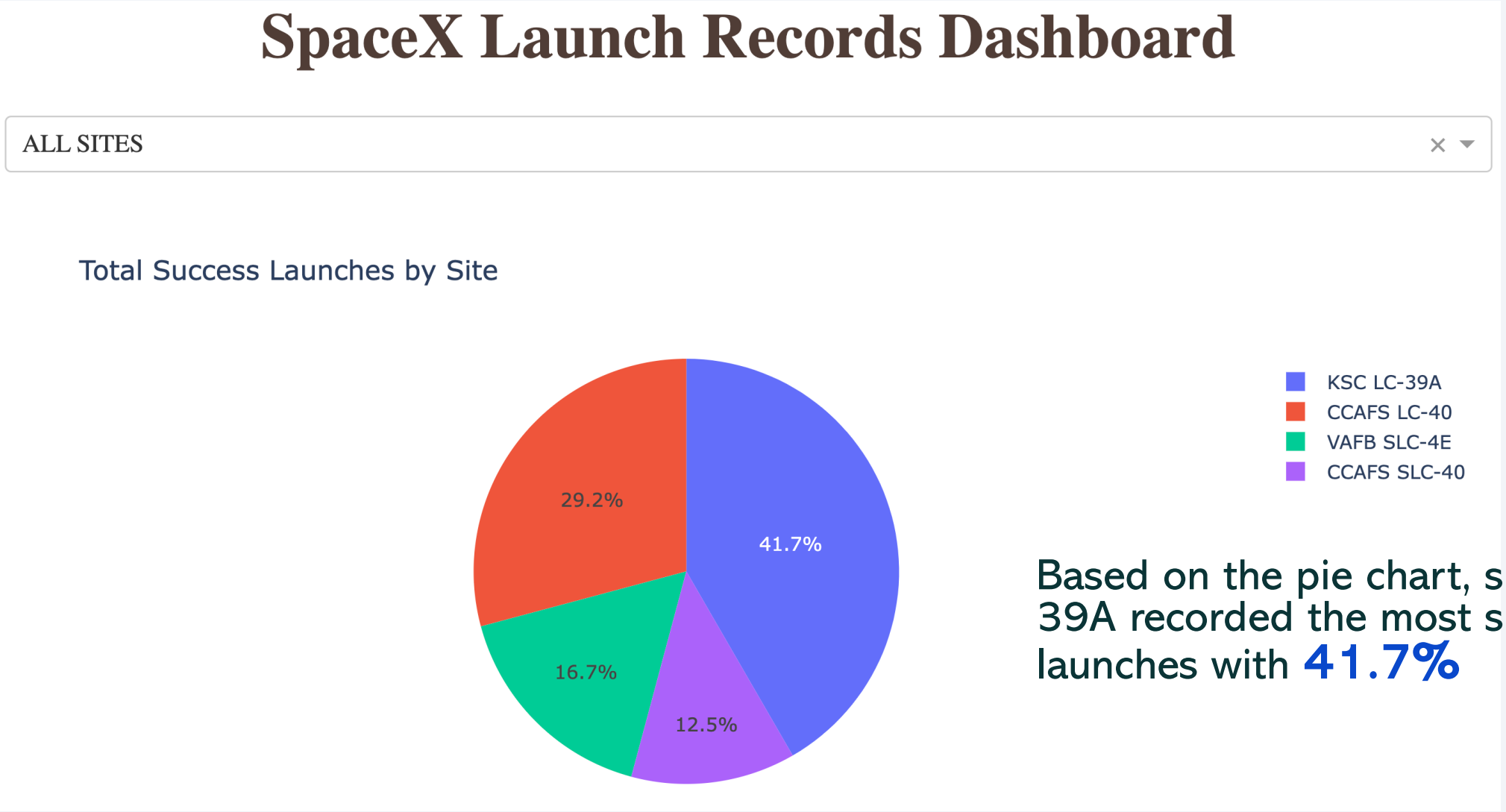
Using the CCAFS SLC-40 launch site as an example site, we can understand more about the placement of launch sites.



- **Are launch sites in close proximity to railways?** *YES. The coastline is only 0.87 km East.*
- **Are launch sites in close proximity to highways?** *YES. The nearest highway is only 0.59km away.*
- **Are launch sites in close proximity to railways?** *YES. The nearest railway is only 1.29 km away.*
- **Do launch sites keep certain distance away from cities?** *YES. The nearest city is 51.74 km away.*

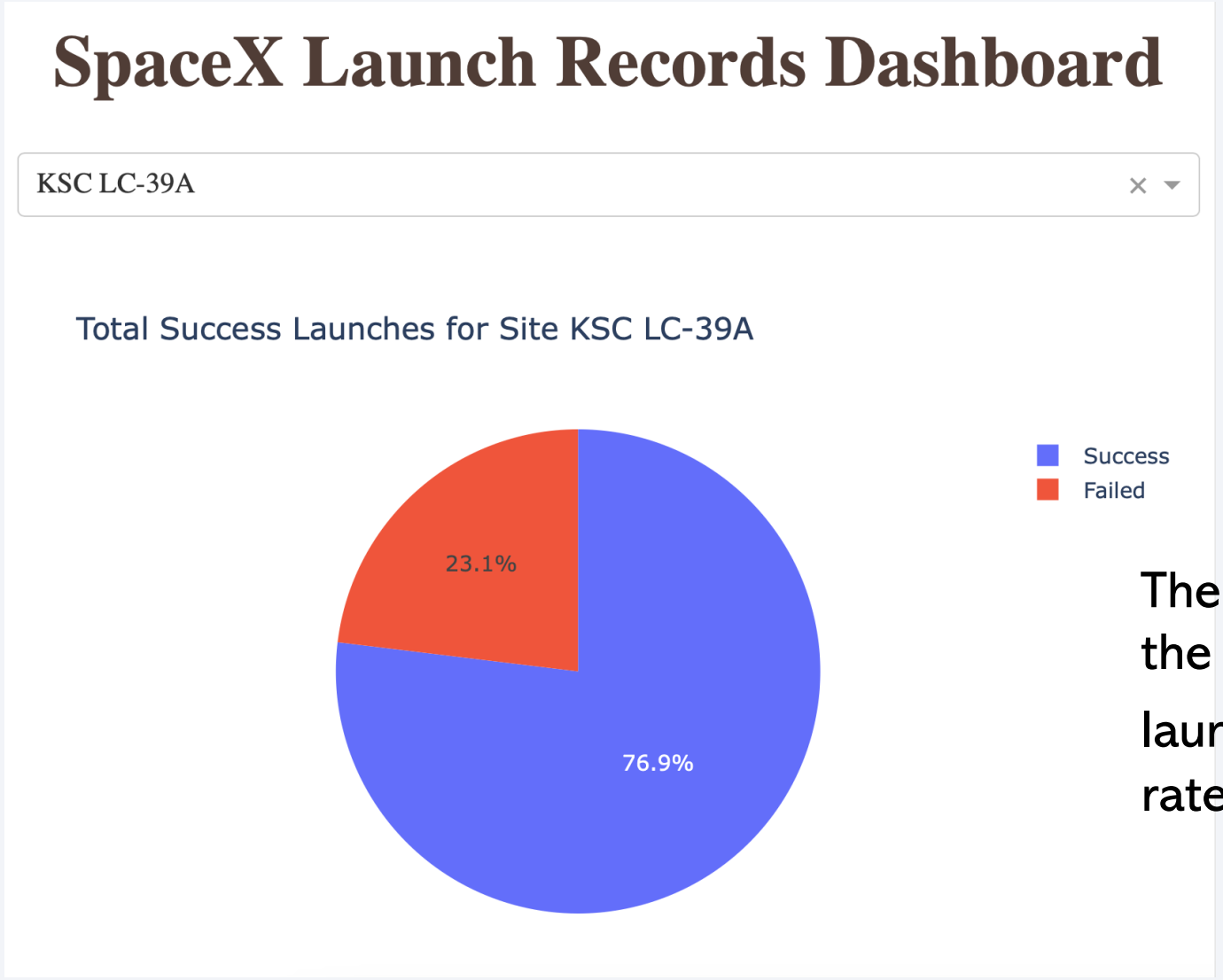

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites



Based on the pie chart, site KSC LC-39A recorded the most success launches with **41.7%**

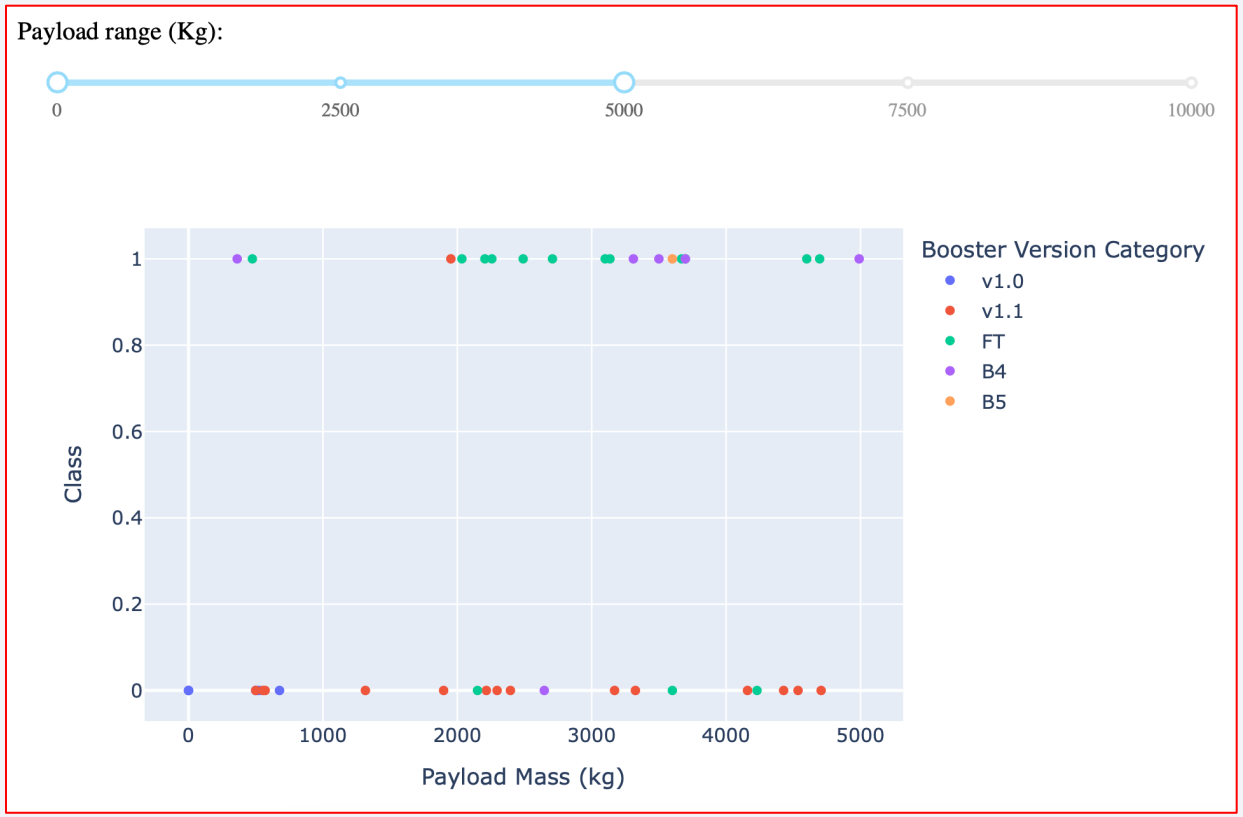
Launch Site with the Highest Launch Success Rate



The launch site KSC LC-39A has the highest rate of successful launches, with a 76.9% success rate.

Payload vs Launch Outcome

The plot is split into 2 ranges: 0 – 5000 kg (low payloads) and 5000 – 10000 kg (massive payloads)



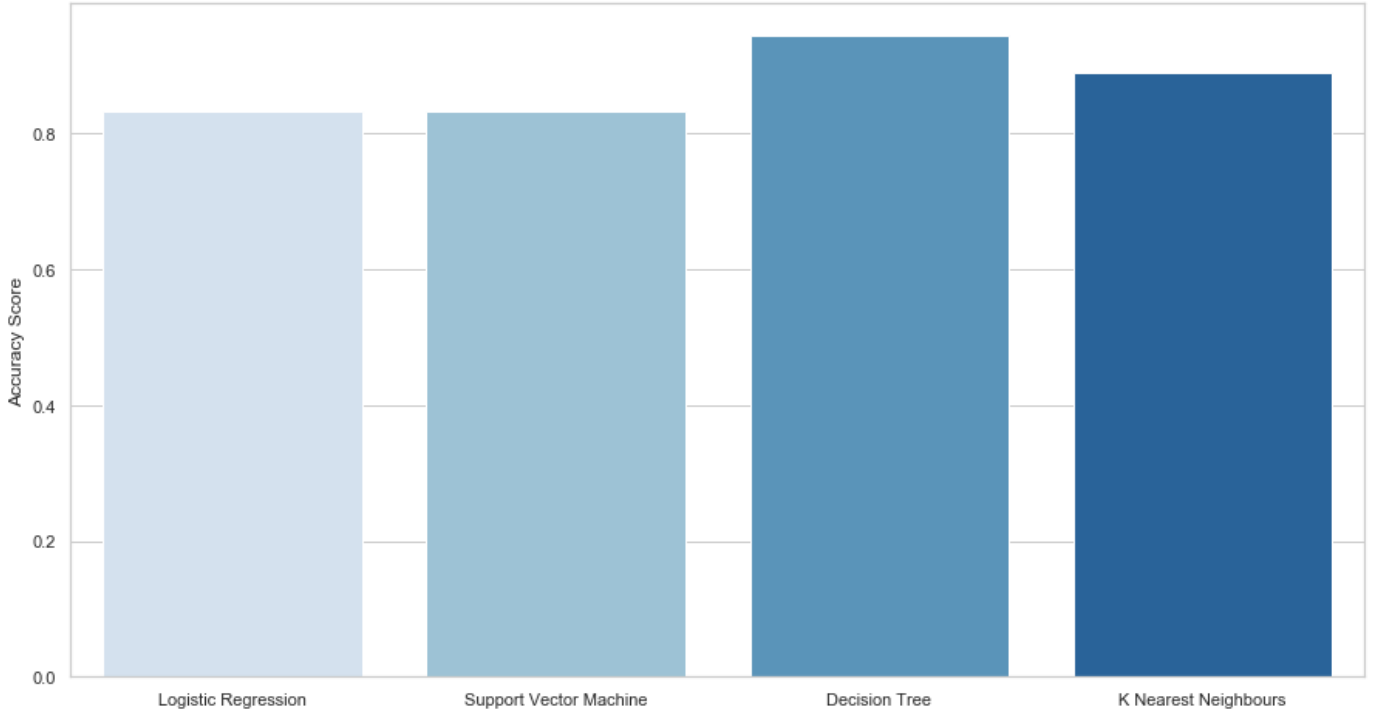
The success rate for massive payloads is lower than that for low payloads.

Some booster types (v1.0 and B5) have not been launched with massive payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Training accuracy score for each model

Training accuracy score

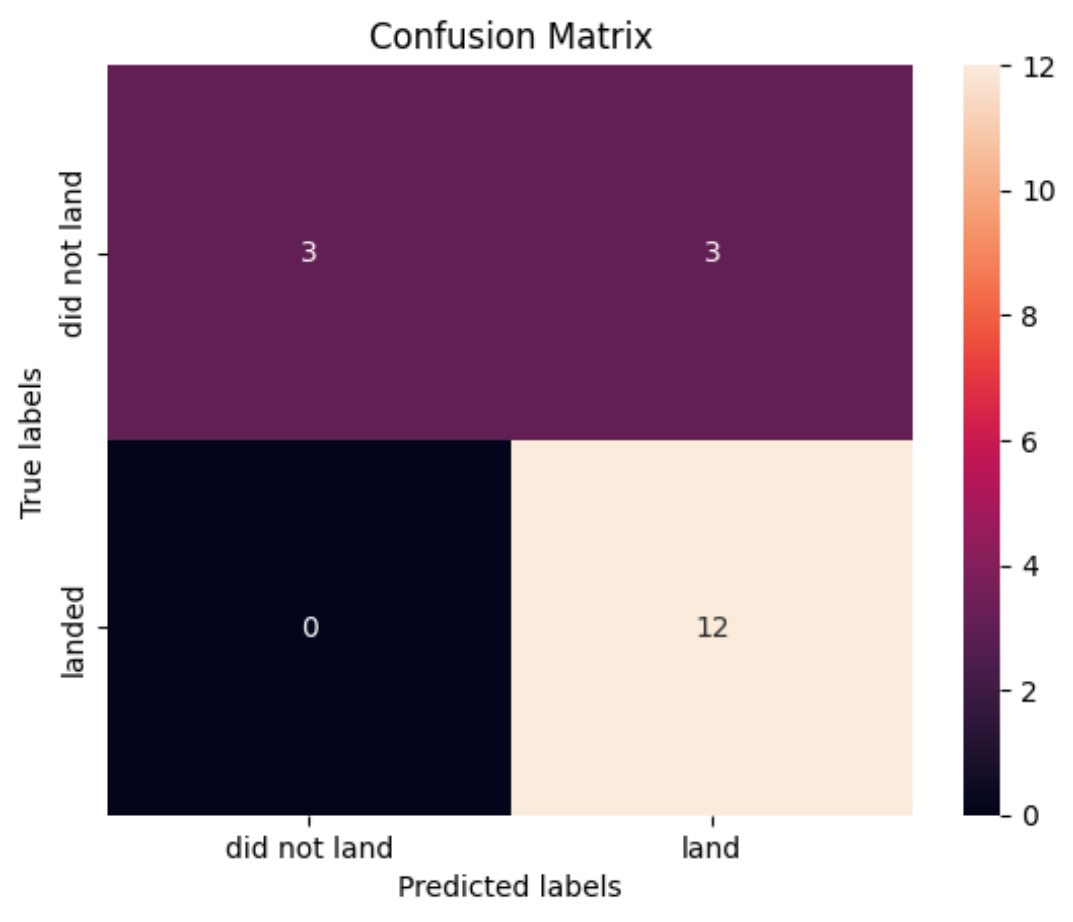
- LR accuracy: 0.8464
- SVM accuracy : 0.848
- DT accuracy : 0.875
- KNN accuracy : 0.848

Test data accuracy

Same accuracy score for Logistic regression, SVM, Decision tree and KNN with 0.8333333333333334

Comparing the performance during training, **Decision Tree** performs best

Confusion Matrix



Confusion matrix of the best performing model: Decision Tree

Only 3 out of 18 total results are classified incorrectly:

- 3 False positive
- 0 False negative

The other 15 results are correctly classified:

- 3 True negative
- 12 True positive

Conclusions

- As the number of flights at a launch site increases, the success rate improves. Initially, many flights may be unsuccessful, but with more experience, the success rate rises..
- Orbit types ES-L1, GEO, HEO, and SSO, have a 100% success rate.
  - ✓GEO, HEO, and ES-L1 orbits each had only one flight, all successful.
  - ✓SSO has a 100% success rate with five successful flights.
  - ✓PO, ISS, and LEO orbits are more successful with heavy payloads.
  - ✓VLEO launches are typically associated with heavier payloads.
- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- The best performing classification model is the Decision Tree model.

Thank you!