# Winning Space Race with Data Science

Burçak Aydın
02/01/2023

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection via API and Web Scraping
  - Exploratory. Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Machine Learning Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive maps and dashboard
  - Predictive results

# Introduction

- Project background and context
  - The aim of the project is to predict if the Falcon 9 first stage will successfully be landed. SpaceX states that Falcon 9 rocket launch costs $60M and other providers' launches cost $165M.
  - Based on the successful launches and landing, we may predict total cost, and give some recommendations for competing with SpaceX in Space Race.

- Problems you want to find answers
  - What are the main factors of a successful or failed landing?
  - What are the effects of each relationship of the rocket features on landing?
  - What operating conditions that allow SpaceX to achieve a successful landing program?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Drop unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# EDA with Data Visualization

- Scatter Graphs
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass

- Bar Graph
  - Success rate vs. Orbit

- Line Graph
  - Success rate vs. Year

# EDA with SQL

- SQL queries to gather and understand data from dataset
  - Display the names of distinct launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total load mass carried by boosters launched by NASA(CRS).
  - Display average payload mass carried by booster version F9.v1.1.
  - List the successful landing outcome in ground pad
  - List the successful boosters in drone ship with 4000 < payload mass < 6000
  - List the successful and failed mission outcomes
  - Display the month names, failed landing in drone ship, booster versions, launch sites in 2015.
  - Rank successful landing outcomes between the dates 04-06-2019 and 20-02-2017 in descending order.

# Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Texas, Houston
  - Red circle at NASA Johnson Space Center (folium.Circle, folium.map.Marker)
  - Red circle at launch site coordinates (folium.Circle, folium.map.Marker, folium.features.DivIcon)
  - Points in a cluster to display multiple information at the same point (folium.plugins.MarkerCluster)
  - Markers to show successful (Green) and unsuccessful (Red) landings (folium.map.Marker, folium.Icon)
  - Markers to show distance with plotline between launch site and key locations(city, railway, highway, coast way) (folium.map.Marker, folium.features.DivIcon, folium.PolyLine)

- These objects are created to understand the problem and data. We can easily show the all launch sites and surroundings, number of successful and unsuccessful landings.

# Build a Dashboard with Plotly Dash

- Dashboard includes dropdown menu, pie chart, range slider, and scatter plot.
  - Dropdown allows the user to choose the launch site or all launch sites(dash_core_components.Dropdown).
  - Pie chart shows totals success and failure for each launch site (plotly.express.pie).
  - Range slider allows the user to select a payload mass in a fixed range(dash_core_components.Rangeslider).
  - Scatter plot shows the relationship between two variables, for instance, Success vs. Payload Mass (plotly.express.scatter).

# Predictive Analysis (Classification)

1. Data preparation
    - Load dataset
    - Normalize data
    - Split data into training and test sets

2. Model preparation
    - Select machine learning algorithms
    - Set parameters for each algorithm to GridSearchCV
    - Training GridSearchModel models with training dataset

3. Model evaluation
    - Get best hyperparameters for each type of model
    - Compute accuracy for each model with test dataset
    - Plot Confusion Matrix

4. Model comparison
    - Comparison of models according to their accuracy
    - The model with best accuracy will be selected

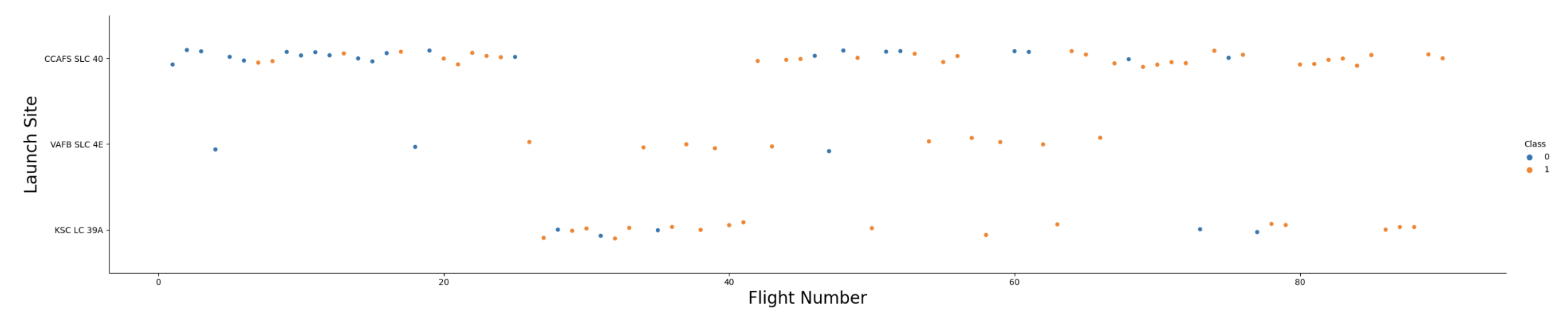
# Results

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

Section 2

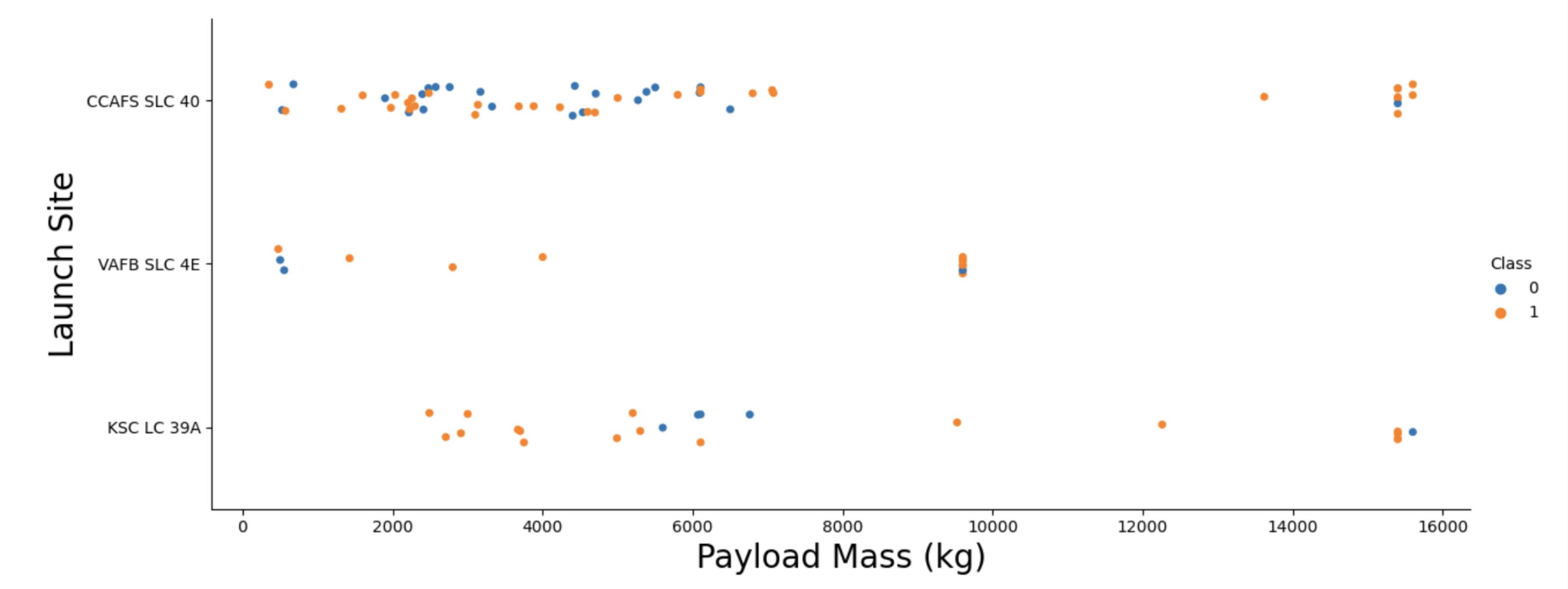
# Insights drawn from EDA

# Flight Number vs. Launch Site



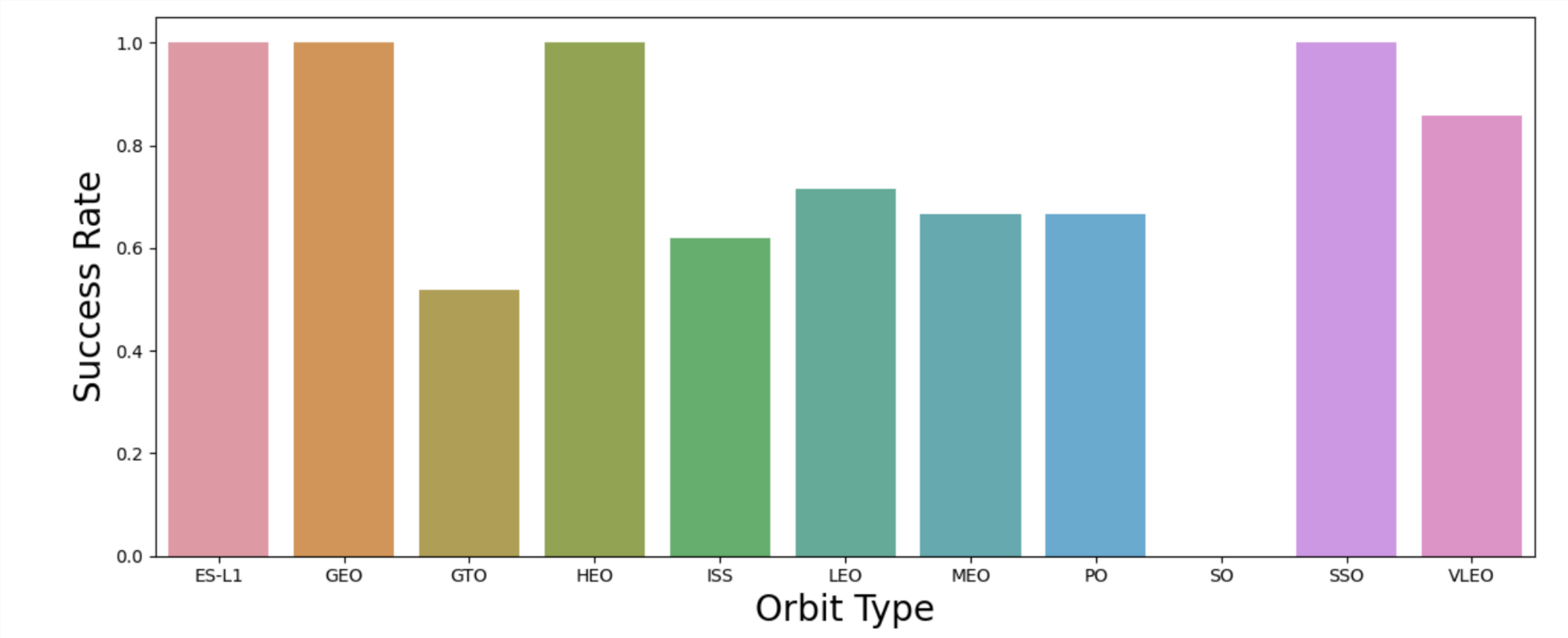

- The success rate is increasing for each site.

# Payload vs. Launch Site



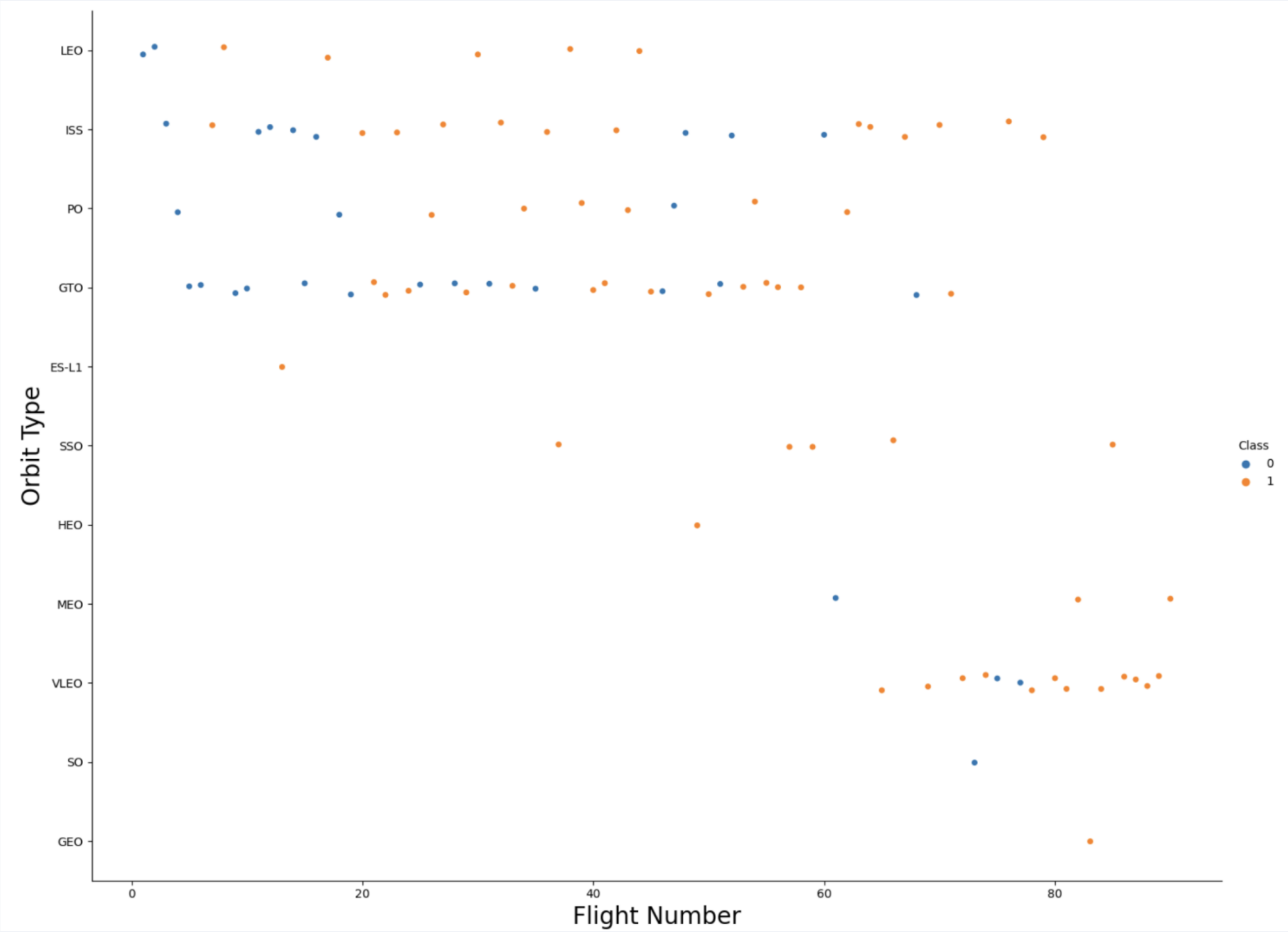

- Based on the launch site,
  - a heavier payload may be a consideration for a successful landing.
  - a too heavy payload can make a landing fail.

# Success Rate vs. Orbit Type



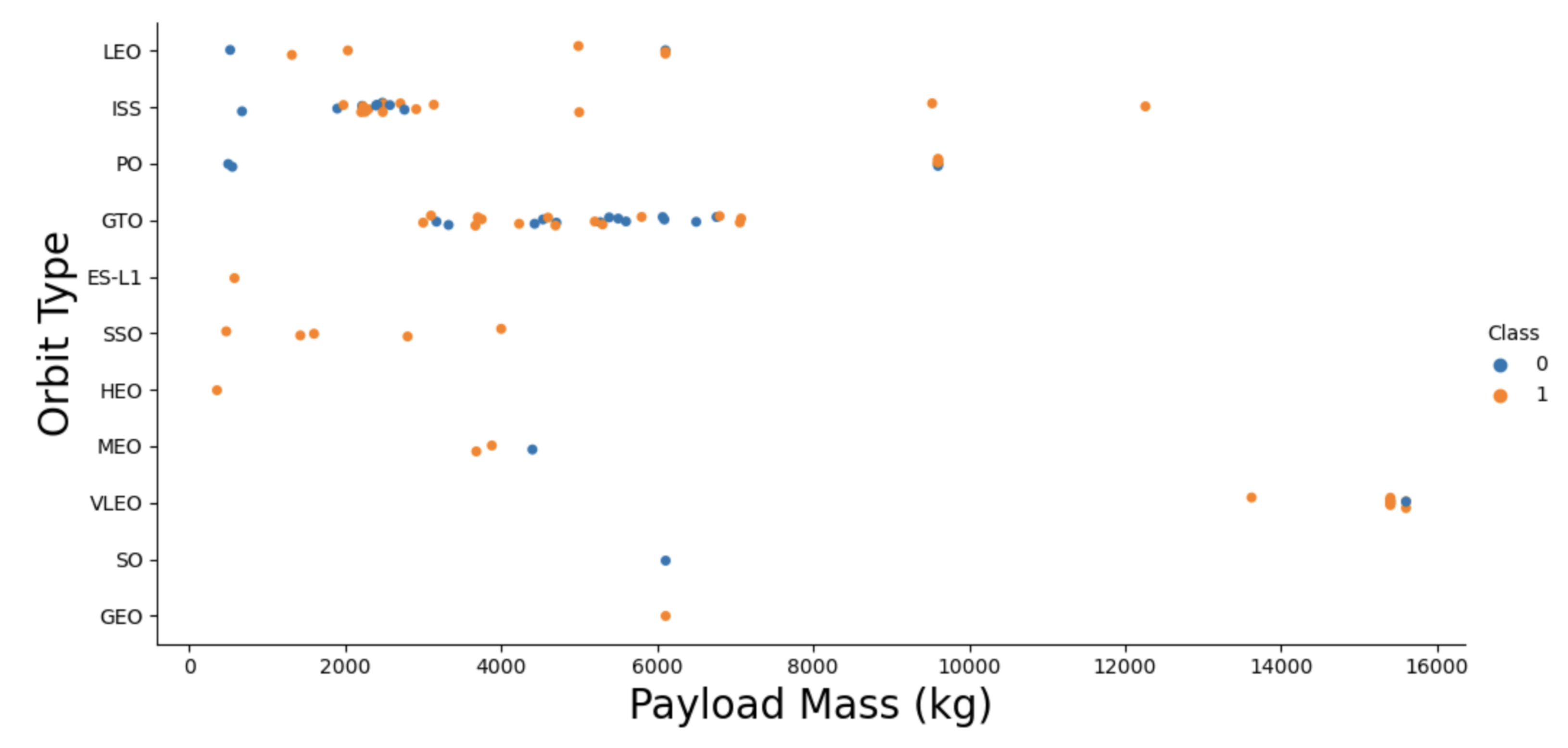

- ES-L1, GEO, HEO, SSO have the best success rate.

# Flight Number vs. Orbit Type




- Success rate increases with the number of flights for the LEO orbit.
- There is no correlation between the success rate and the number of flights.

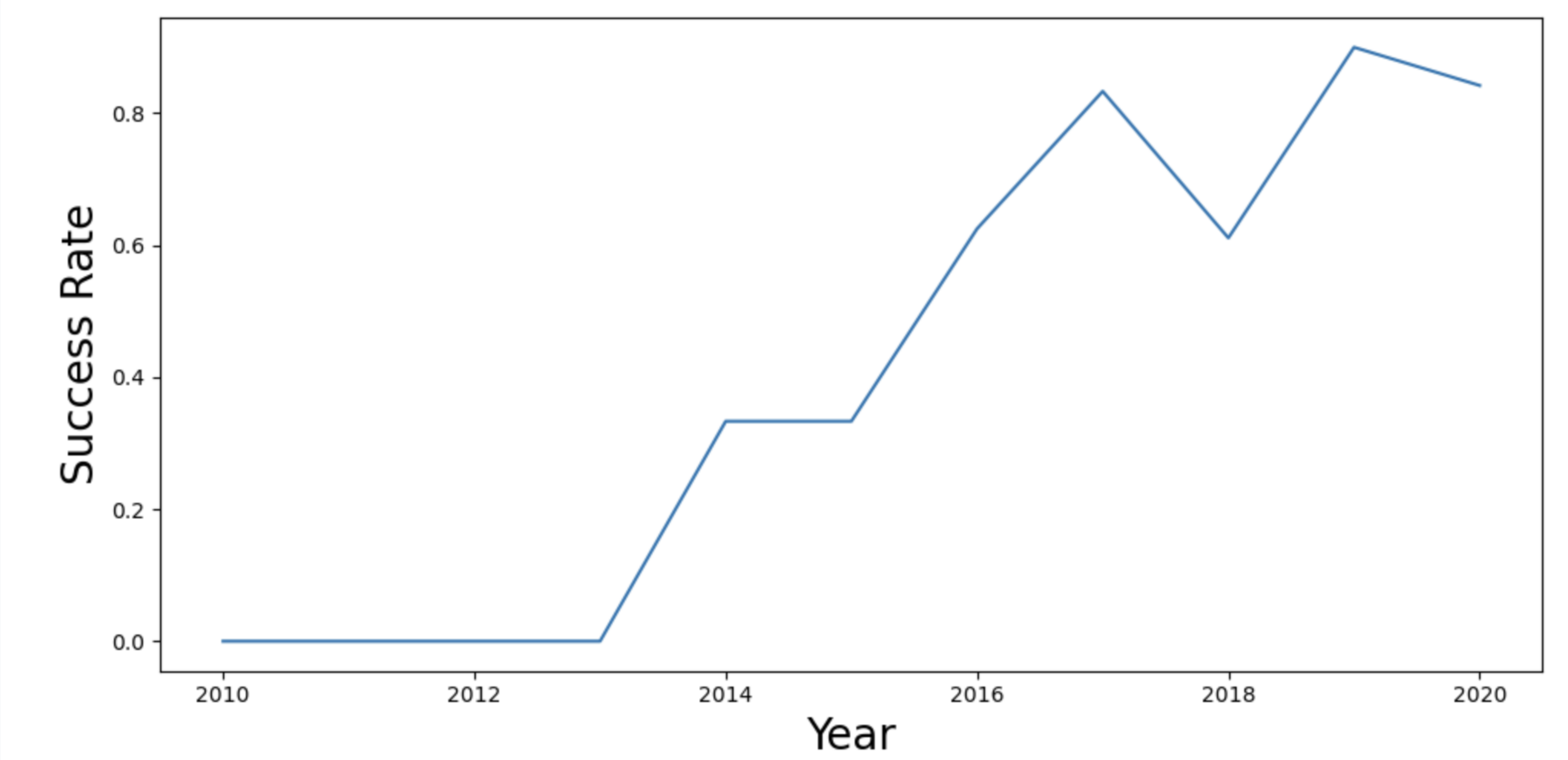
# Payload vs. Orbit Type




The weight of the payloads can have a great influence on the success rate of the launches in certain orbits.

The less payload weight for a GTO orbit improves the success of a launch.

# Launch Success Yearly Trend




- We can see an increase in the Space X Rocket success rate since 2013

# All Launch Site Names

## SQL Query and Result

```
%sql select DISTINCT(LAUNCH_SITE) from SPACEXTBL;
```

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

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

- **DISTINCT in the query was used to remove duplicated LAUNCH_SITE**

# Launch Site Names Begin with 'CCA'

## SQL Query and Result

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

| Launch_Site |
| --- |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

- LIKE clause filters launch sites that contain the substring CCA.
- LIMIT 5 shows only 5 records

# Total Payload Mass

## SQL Query and Result

```
[8]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where "CUSTOMER" = 'NASA (CRS)';

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

[8]:

| sum(PAYLOAD_MASS__KG_) |
|---|
| 45596 |

- This query returns the sum of all payload masses where the customer is NASA (CRS).

# Average Payload Mass by F9 v1.1

## SQL Query and Result

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where "BOOSTER_VERSION" LIKE '%F9 v1.1%';
```

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

| avg(PAYLOAD_MASS__KG_) |
|---|
| 2534.6666666666665 |

- This query returns the average of all payload masses where the booster version contains the substring F9 v1.1.

# First Successful Ground Landing Date

## SQL Query and Result

```
%sql select min(DATE) from SPACEXTBL where "Landing _Outcome" lIKE '%Success%';
```

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

| min(DATE) |
| --- |
| 01-05-2017 |

- Selection of the oldest successful landing
- The WHERE clause filters dataset to keep only records where landing was successful. With the MIN function, we select the record with the oldest date.

# Successful Drone Ship Landing with Payload between 4000 and 6000

## SQL Query and Result

```
%sql select "BOOSTER_VERSION" from SPACEXTBL 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 |

- This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg.
- The WHERE and AND clauses filter the dataset.

# Total Number of Successful and Failure Mission Outcomes

## SQL Query and Result

```
%sql select (select count("MISSION_OUTCOME") from SPACEXTBL where "MISSION_OUTCOME" LIKE '%Success%') as Success, \
(select count("MISSION_OUTCOME") from SPACEXTBL where "MISSION_OUTCOME" LIKE '%Failure%') as Failure;
```

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

| Success | Failure |
|---|---|
| 100 | 1 |

- With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. The second subquery counts the unsuccessful mission.
- The WHERE clause followed by LIKE clause filters mission outcome. The COUNT function counts records filtered.

# Boosters Carried Maximum Payload

## SQL Query and Result

```
%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 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 |

- We used a subquery to filter data by returning only the heaviest payload mass with MAX function.
- The main query uses subquery results and returns unique booster version (SELECT DISTINCT) with the heaviest payload mass.

# 2015 Launch Records

## SQL Query and Result

```
]: %sql SELECT substr("DATE",4,2) as MONTH,"BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
where "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE", 7, 4) = '2015';

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

| MONTH | Booster_Version | Launch_Site |
|---|---|---|
| 01 | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | F9 v1.1 B1015 | CCAFS LC-40 |

- This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015.
- substr() function process date in order to take month or year.
- substr(DATE, 4, 2) shows month. substr(DATE,7, 4) shows year.

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

## SQL Query and Result

```
%sql select "LANDING _OUTCOME", count("LANDING _OUTCOME") FROM SPACEXTBL\
where "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" like '%Success%'\
group by "LANDING _OUTCOME"\
order by count("LANDING _OUTCOME") DESC;
```

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

| Landing _Outcome | count("LANDING _OUTCOME") |
|---|---|
| Success | 20 |
| Success (drone ship) | 8 |
| Success (ground pad) | 6 |

- This query returns landing outcomes and their count where mission was successful, and date is between 04/06/2010 and 20/03/2017.
- The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

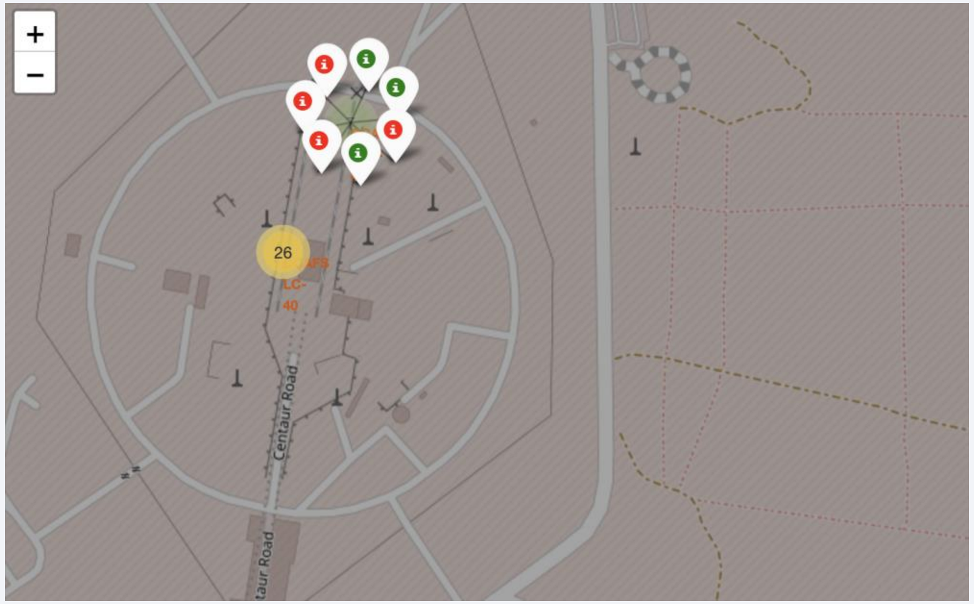
Section 3

# Launch Sites Proximities Analysis

# Folium map – Ground stations


Space X launch sites are located on the coast of the United States

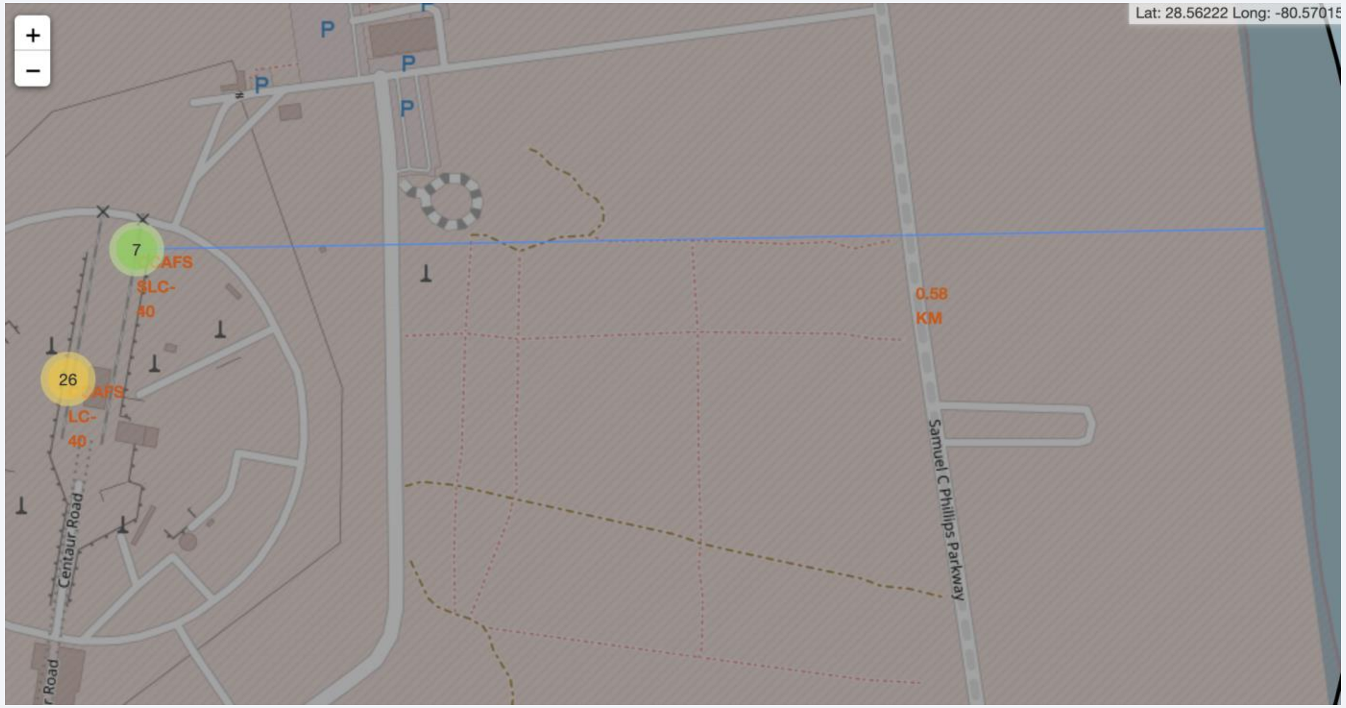
# Folium map – Color Labeled Markers





- Green marker represents successful launches.
- Red marker represents unsuccessful launches.
- We note that KSC LC-39A has a highest launch success rate.

# Folium Map – Distances between CCAFS SLC-40 and its proximities

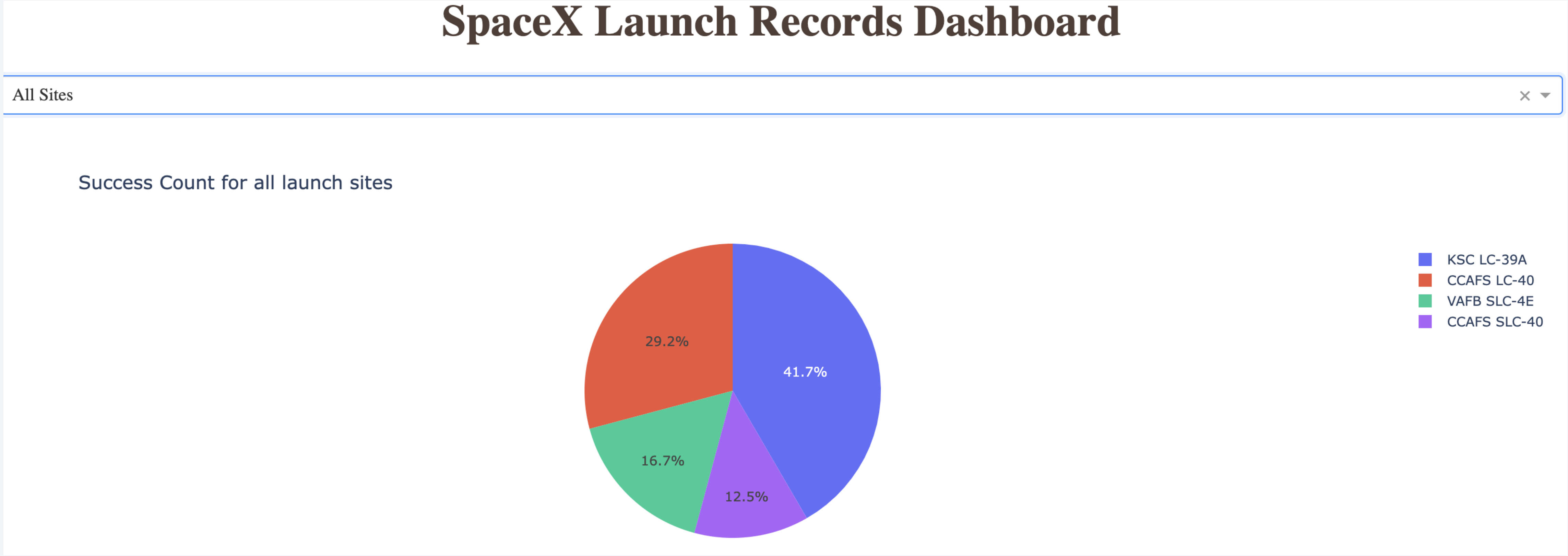



- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? No

Section 4

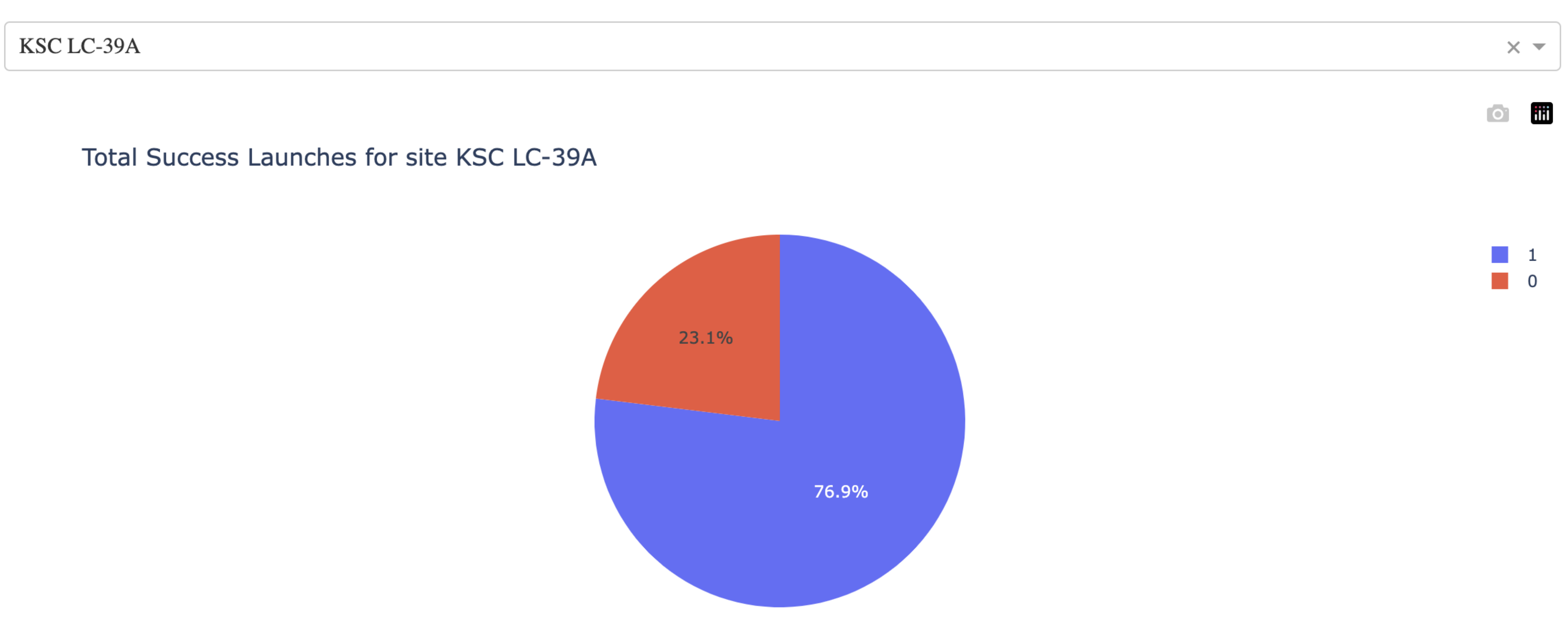
# Build a Dashboard with Plotly Dash

# Dashboard - Success by Site



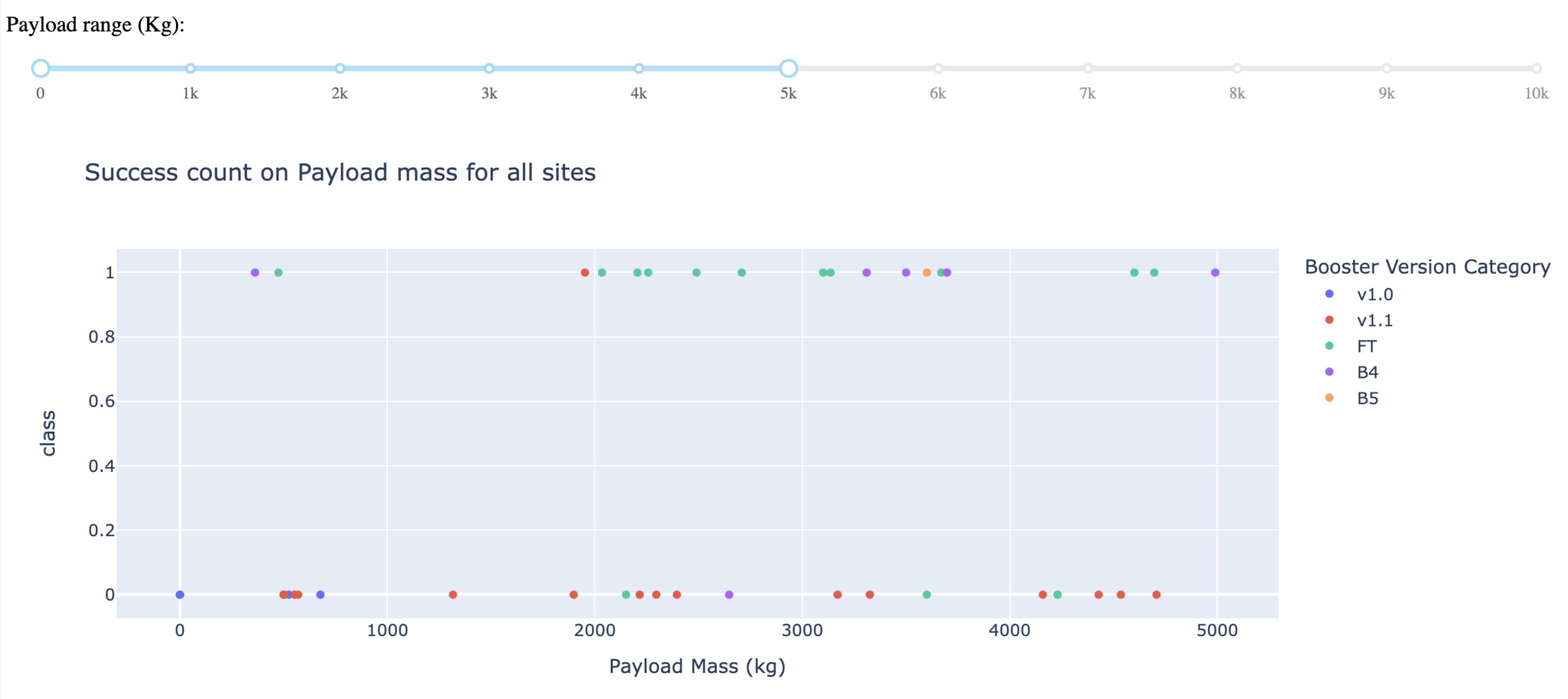

- KSC LC-39A has the best success rate of launches

# Dashboard - Success launches for Site KSC LC-39A



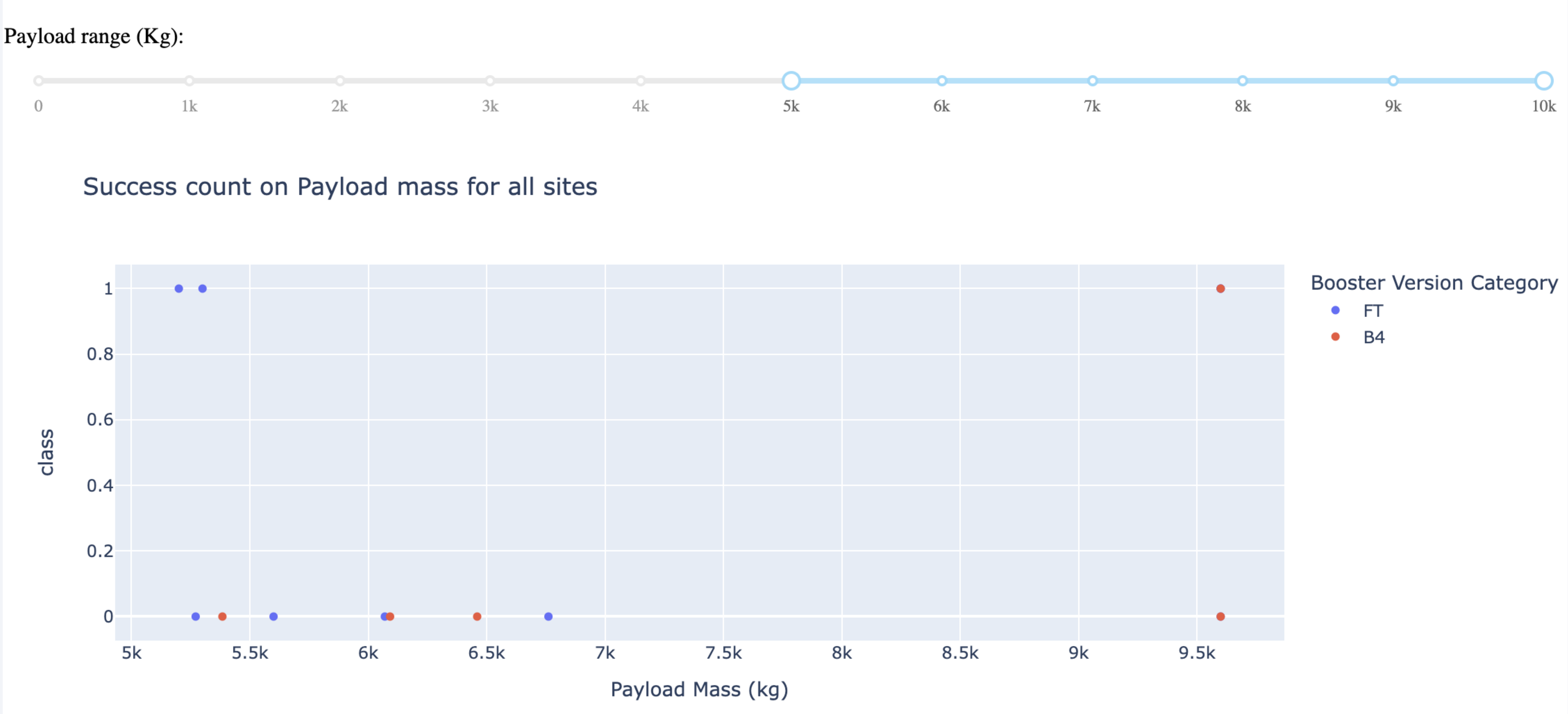

- KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate

# Dashboard – Payload mass vs Outcome for all sites with different payload mass selected




- Low weighted payloads have a better success rate than the heavy weighted payloads.

# Dashboard – Payload mass vs Outcome for all sites with different payload mass selected

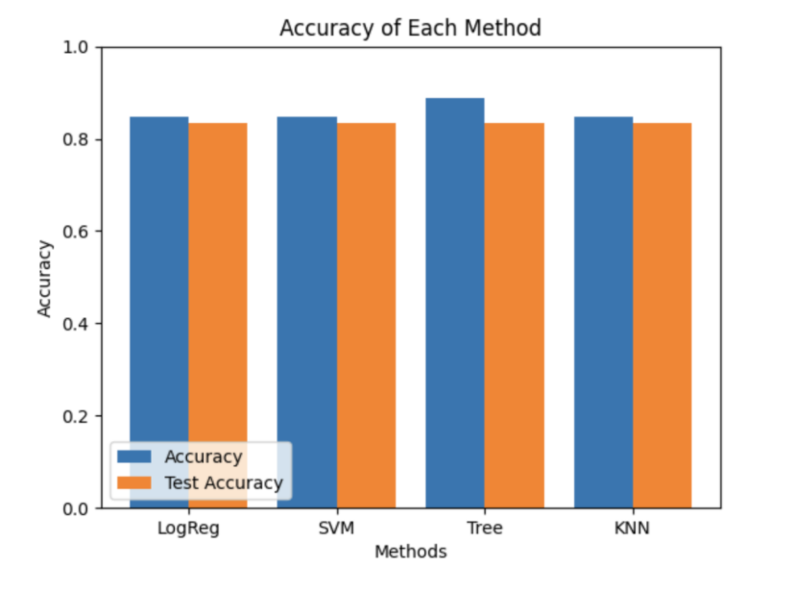


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

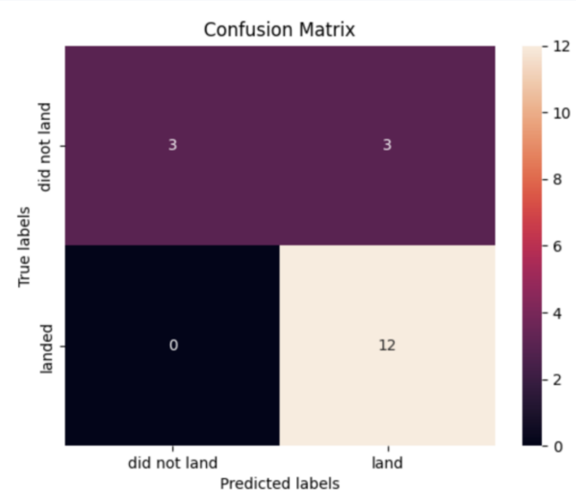
| | Accuracy Train | Accuracy Test |
|---|---|---|
| Logreg | 0.846429 | 0.833333 |
| Svm | 0.848214 | 0.833333 |
| Tree | 0.875000 | 0.833333 |
| Knn | 0.848214 | 0.833333 |



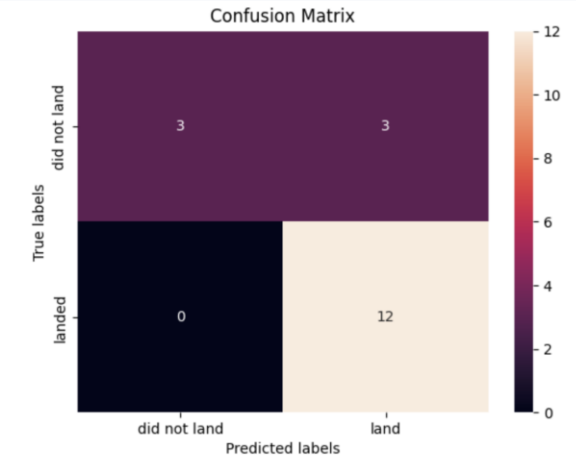

- In accuracy test, all methods performed similar.
- If we really need to choose one right now, we will take the **decision tree**.

# Confusion Matrix

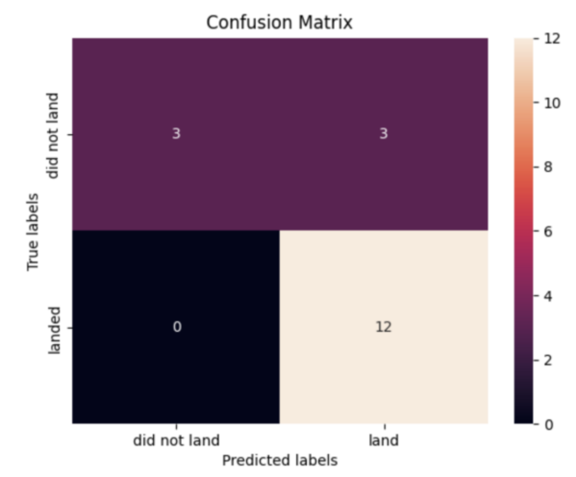
**Logistic regression**



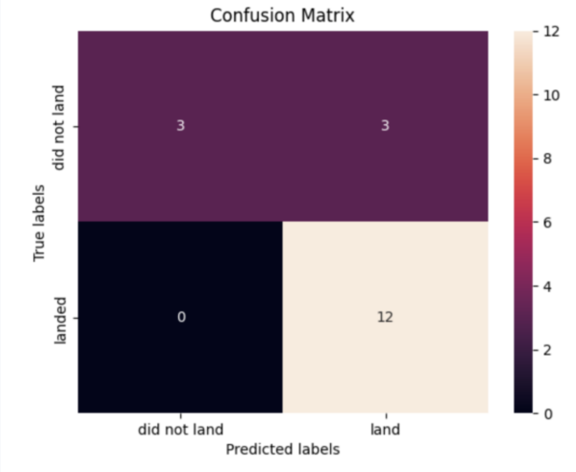

**Decision Tree**



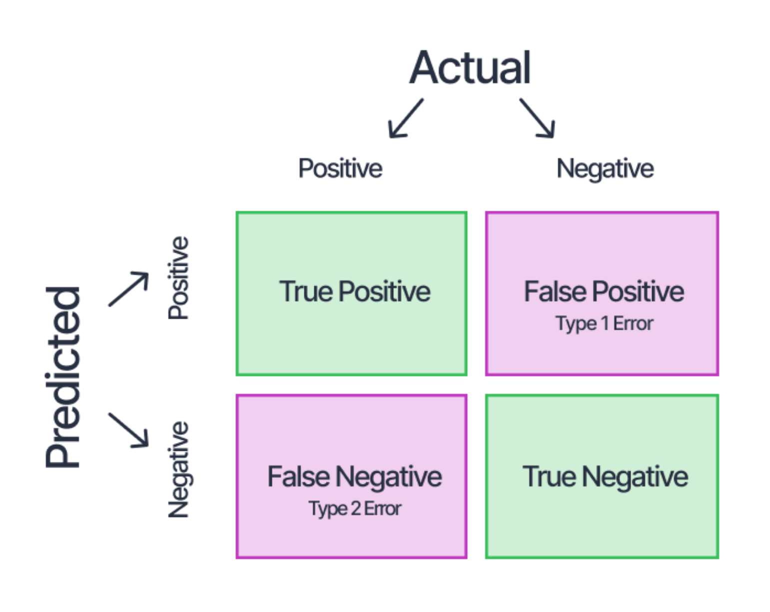

**kNN**




**SVM**




- Show the confusion matrix of the best performing model with an explanation



# Conclusions

- Different data sources were analyzed, refining conclusions along the process;
  - GEO, HEO, SSO, ES-L1 have the best success rates,
  - Launches above 7,000kg are less risky,
  - The best launch site is KSC LC-39A,
  - Low weighted payloads perform better than the heavy weighted payloads.
- According the evolution of processes and rockets,
  - Most of mission outcomes are successful,
  - Successful landing outcomes seem to improve over time.
- The Decision Tree Algorithm has a better train accuracy and test accuracy between all the models used is identical. Decision Tree Classifier can be used to predict successful landings and increase profits.

# Appendix

- Python code snippets, SQL queries, charts, Notebook outputs, and data sets were used to create the project.
- SpaceX Falcon 9 Landing Prediction Project was deployed to GitHub.
- https://github.com/BurcakAydin/ibmcapstone

Thank you!