IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammed Ibrahim
9/17/2024

Outline

Executive Summary

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

Introduction

- Project background and context

This project aims to predict whether the Falcon 9 rocket's first stage will successfully land. SpaceX states that launching a Falcon 9 rocket costs $62 million, while other companies' launch costs can exceed $165 million. This cost difference is mainly because SpaceX reuses the rocket's first stage. By predicting the landing success, we can estimate the cost of a launch. This information could be valuable for other companies looking to compete with SpaceX.

- Problems you want to find answers

What are the key factors that lead to a successful or failed landing?

How does each rocket-related variable affect the success or failure of the landing?

Section 1

Methodology

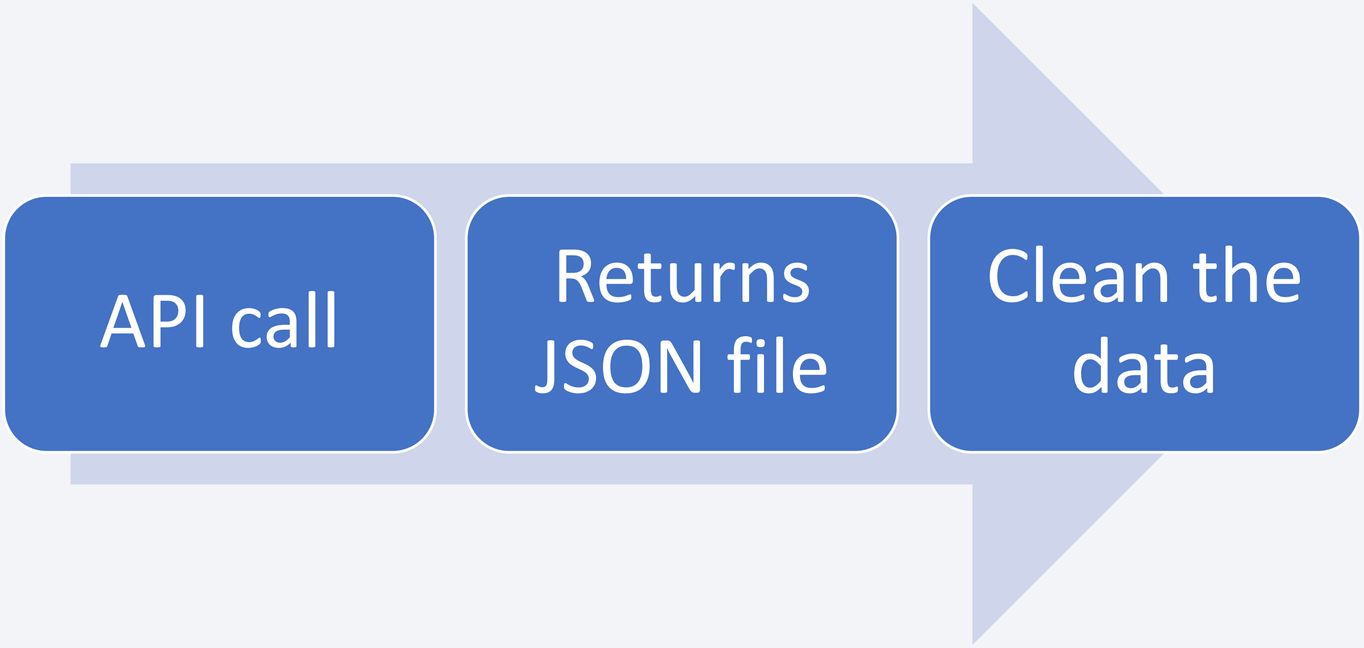
Methodology

Executive Summary

- Data collection methodology:
    - SpaceX REST API
- Perform data wrangling
    - One Hot Encoding
    - Dropping unnecessary columns
- 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

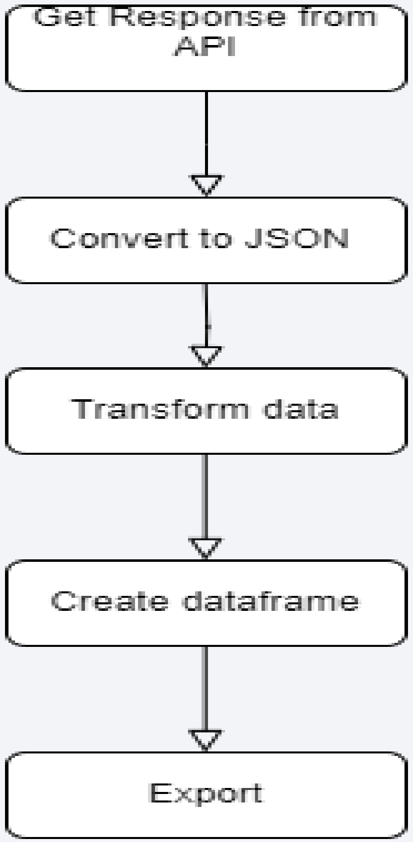
Data Collection

- Datasets are collected from Rest SpaceX API
- The API returns information about rocket, launches, and payload.



Data Collection – SpaceX API

- The step-by-step flowchart data collection from SpaceX REST API



EDA with Data Visualization

- Scatter plots show correlation between variables.

  Flight Number vs Payload Mass vs Launch Site

- Line graphs show data variables and their trends

  Success rate vs Year

- Bar graphs show the relationship between numeric and categoric variables

  Success rate vs. Orbit

EDA with SQL

We used SQL queries to explore and understand the data in the dataset:

- Show the unique names of the launch sites in the space missions.
- Show 5 records where the launch sites start with the letters 'CCA'.
- Show the total payload mass carried by boosters launched by NASA (CRS).
- Show the average payload mass carried by the booster version F9 v1.1.
- List the date when the first successful landing on a ground pad happened.
- List the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000.

Build an Interactive Map with Folium

- The Folium map is centered on NASA Johnson Space Center in Houston, Texas:
- A red circle marks the NASA Johnson Space Center with its name.
- Red circles show each launch site with labels.
- Points are grouped in clusters to display different information at the same location.
- Markers indicate landings: green for success, red for failure.
- Markers and lines show distances from launch sites to key locations (railway, highway, coast, city).
- These elements help visualize all launch sites, surroundings, and landing outcomes.

Build a Dashboard with Plotly Dash

The dashboard includes a dropdown, pie chart, range slider, and scatter plot:

- The dropdown lets users select a specific launch site or all sites.
- The pie chart displays the total successes and failures for the selected launch site.
- The range slider allows users to pick a payload mass within a set range.
- The scatter plot shows the relationship between success and payload mass.

Predictive Analysis (Classification)

Data Preparation

- Load the dataset - Normalize the data - Split data into training and test sets

Model Preparation

- Choose machine learning algorithms - Set parameters using GridSearchCV - Train models with the training data

Model Evaluation

- Find the best hyperparameters for each model - Calculate accuracy using the test data - Plot the confusion matrix

Model Comparison

- Compare models based on accuracy - Choose the model with the highest accuracy

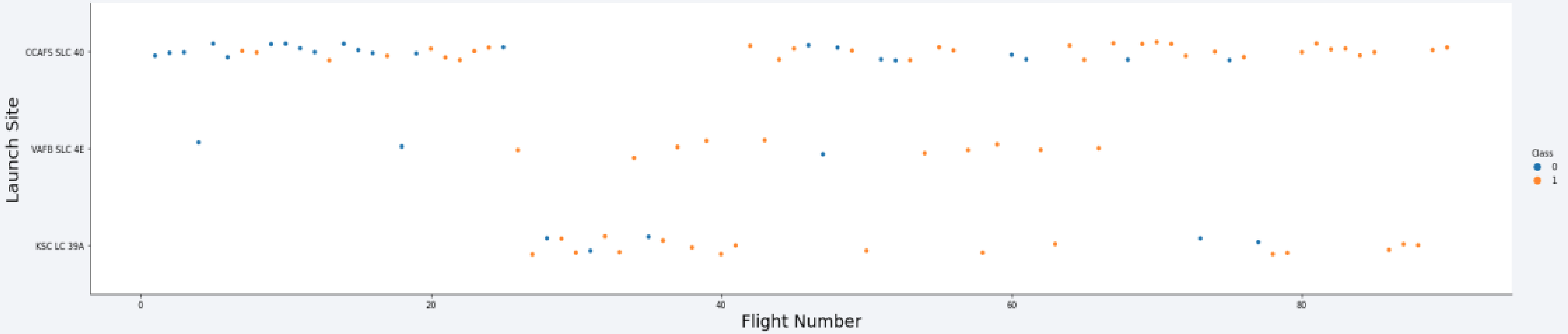
Results

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

Section 2

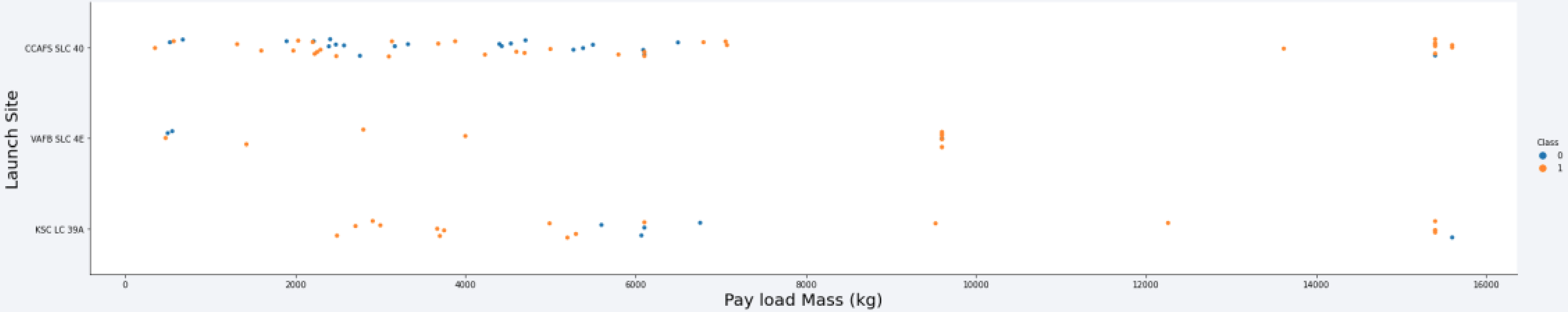
Insights drawn from EDA

Flight Number vs. Launch Site



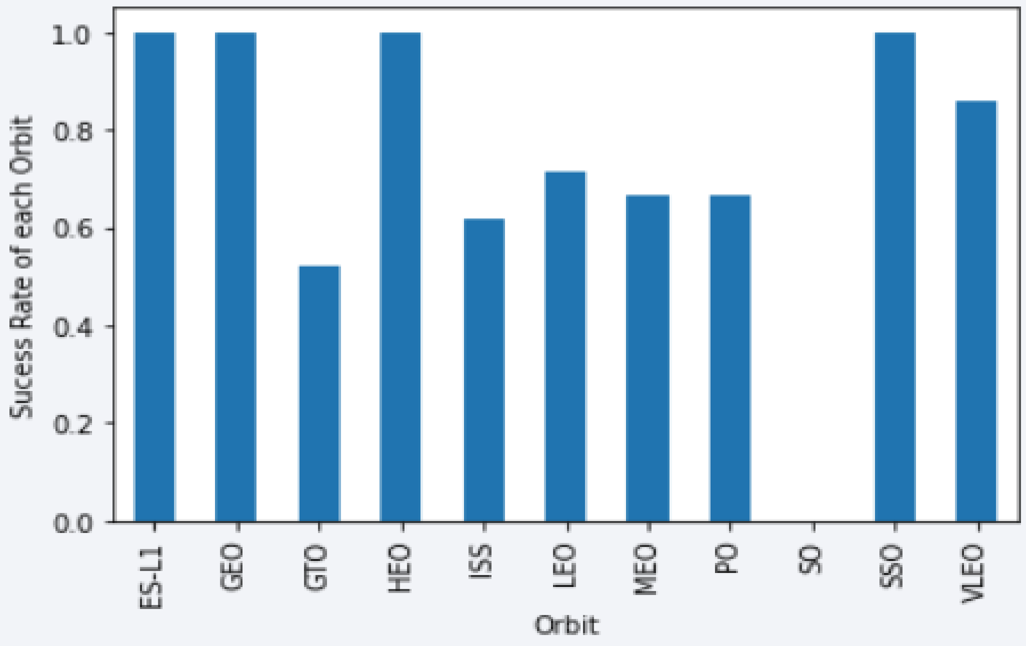
We notice that the success rate is improving for each site.

Payload vs. Launch Site



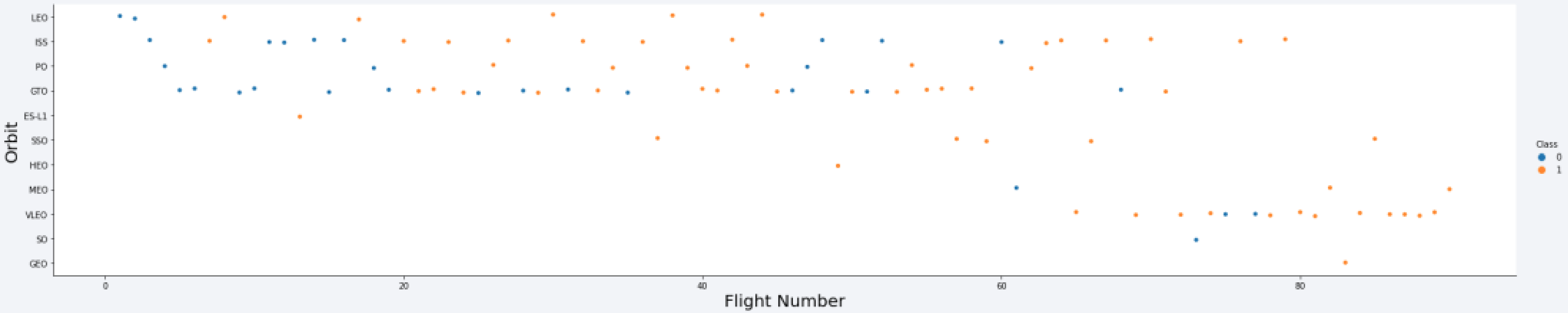
The launch site affects how the payload weight impacts a successful landing. A heavier payload might help, but if it's too heavy, the landing can fail.

Success Rate vs. Orbit Type



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

Flight Number vs. Orbit Type

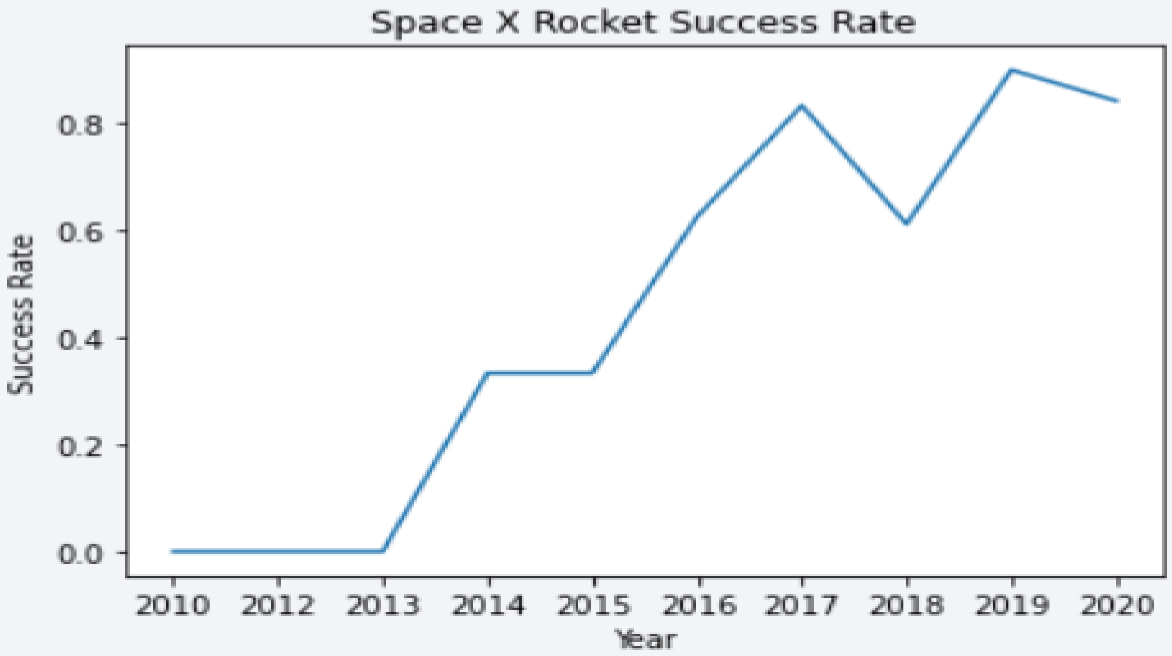


Success rate increases with the number of flights for the LEO orbit

Payload vs. Orbit Type

The weight can influence on the success rate

Launch Success Yearly Trend



Positive success rate.

All Launch Site Names

```
SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

DISTINCT to remove duplicate LAUNCH_SITE.

Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

LIMIT 5 shows 5 records from filtering.

Total Payload Mass

```
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

returns the sum of all payload and the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

```
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

Returns the average of all payload masses where the booster version contains F9 v1.1.

First Successful Ground Landing Date

```
SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

With this query, we select the oldest successful landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

This query returns the booster version where landing was successful, and payload mass is between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%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
```

With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission and the second counts the failure mission outcome

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

We filter data by returning only the heaviest payload mass with MAX function

2015 Launch Records

```
%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'
```

This query returns the month and booster version and launch site where landing was unsuccessful, and landing date took place in 2015.

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

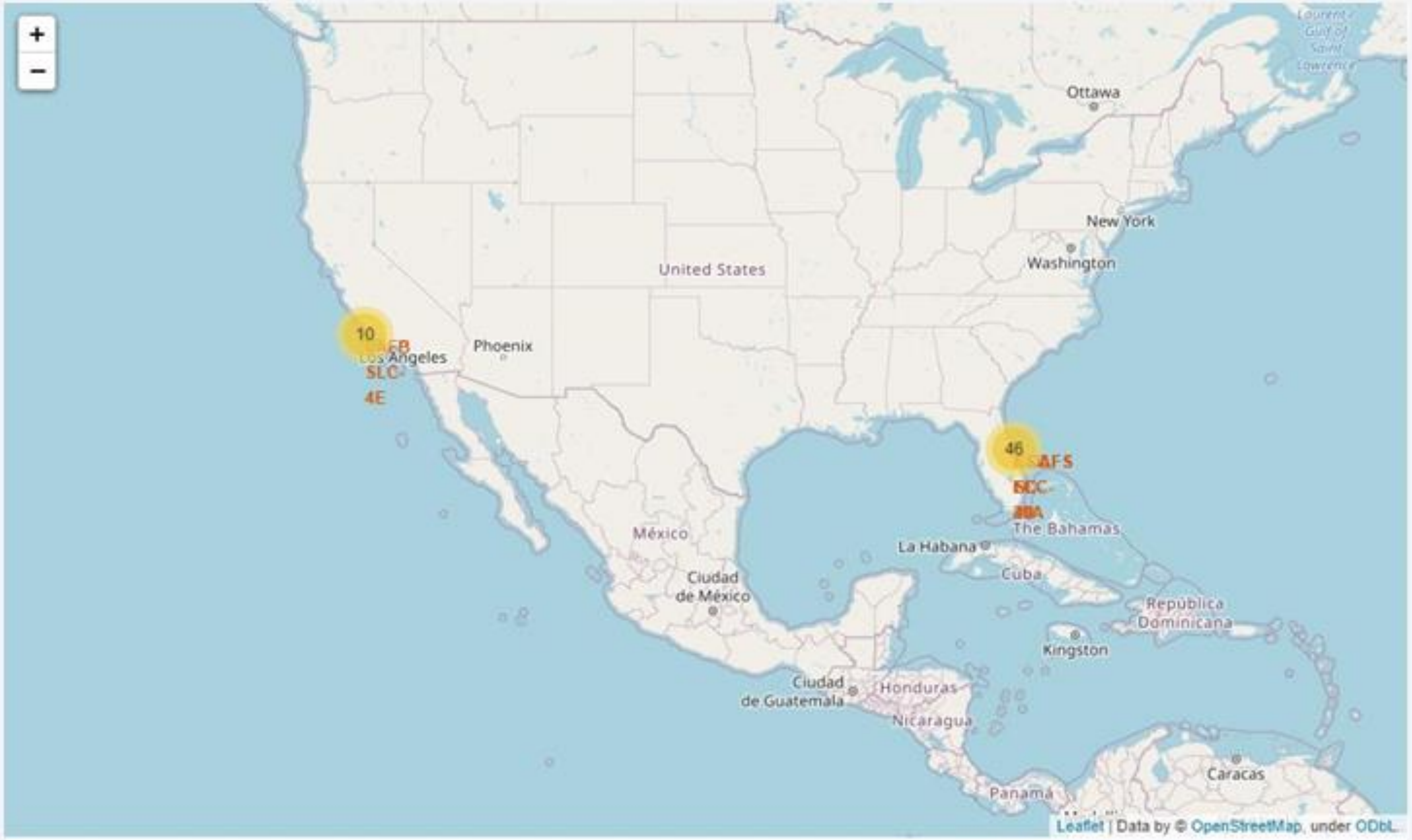
```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

This query returns landing outcomes and their count where mission was successful and date is between 2010 and 2017.

Section 3

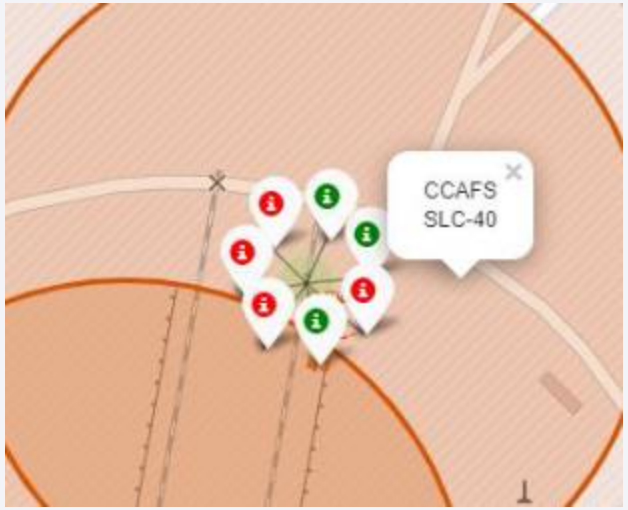
Launch Sites Proximities Analysis

Folium map – Ground stations



Space X launch sites are in United States

Folium map – Color Markers



Green are successful launches. Red are unsuccessful launches.

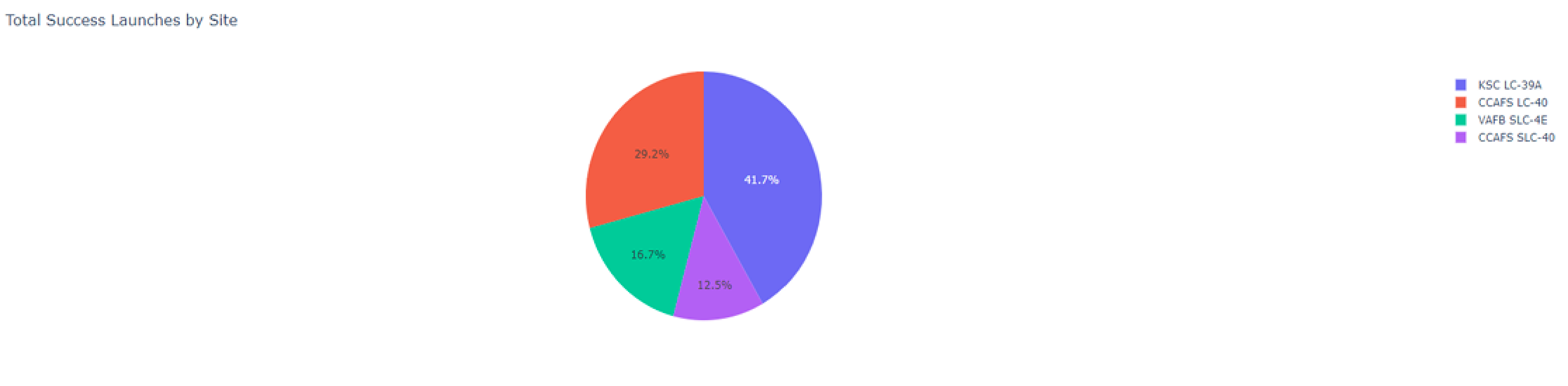
Folium Map – Distances of CCAFS SLC-40 to the proximities

CCAFS SLC-40 is to railways and highways and coastline

Section 4

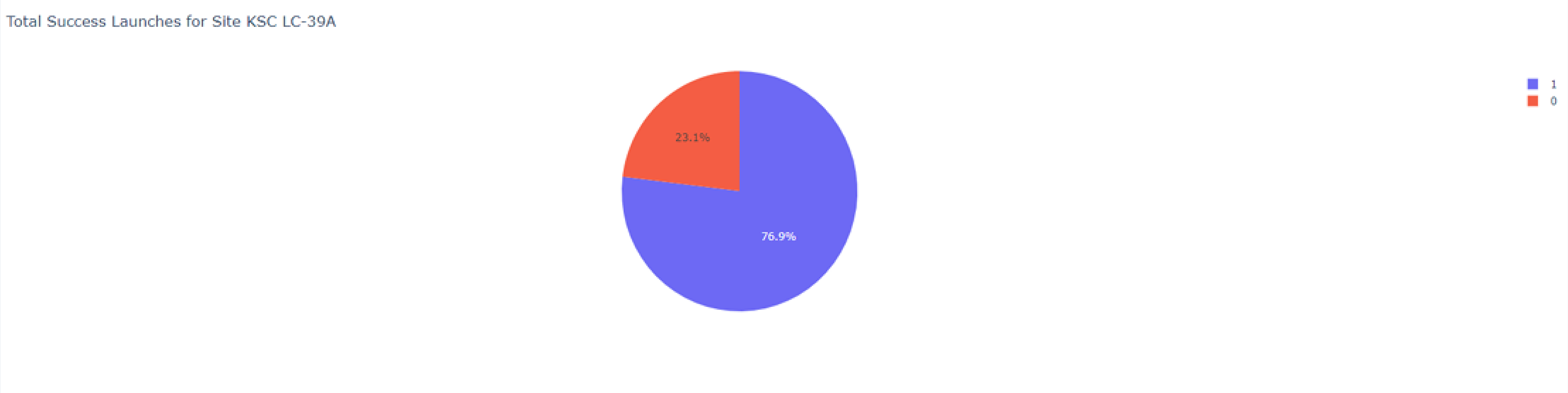
Build a Dashboard with Plotly Dash

Dashboard – Total success by Site



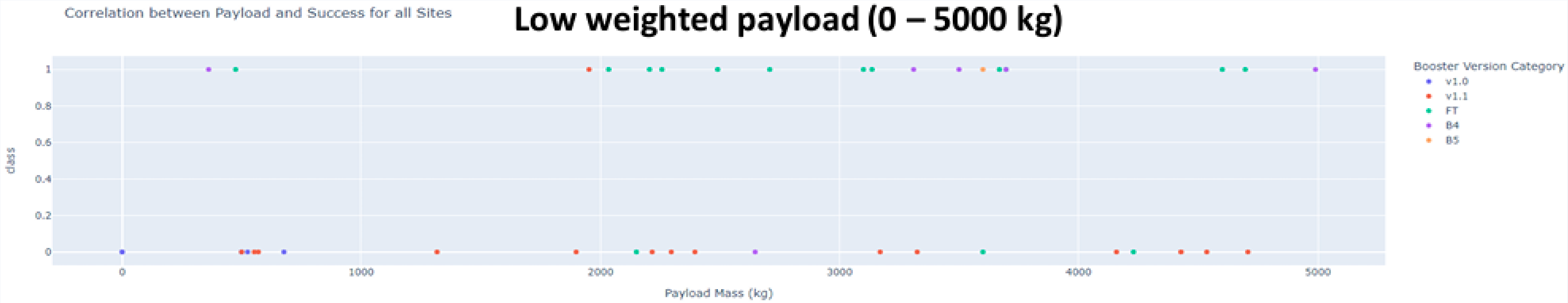
KSC LC-39A has the best success rate

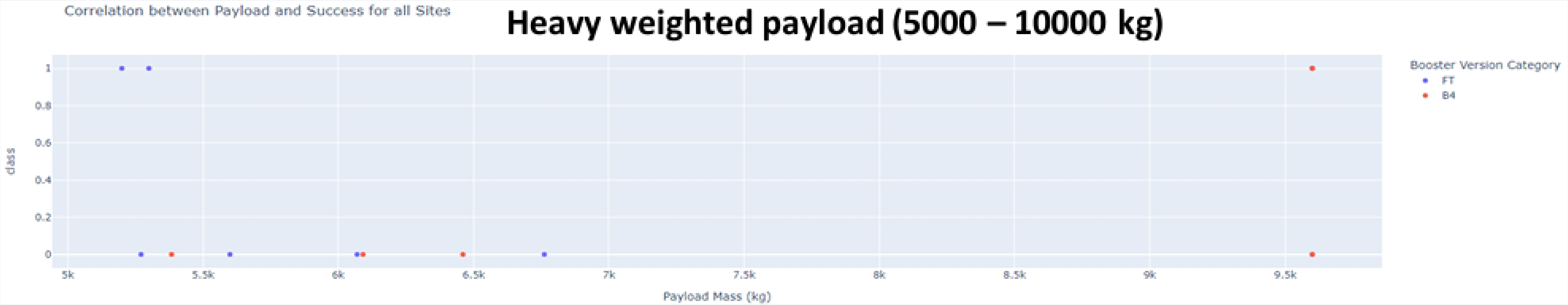
Dashboard – Total success launches for Site KSC LC-39A



KSC LC-39A has achieved a 76.9%success rate

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





Low weighted payloads have a better success rate.

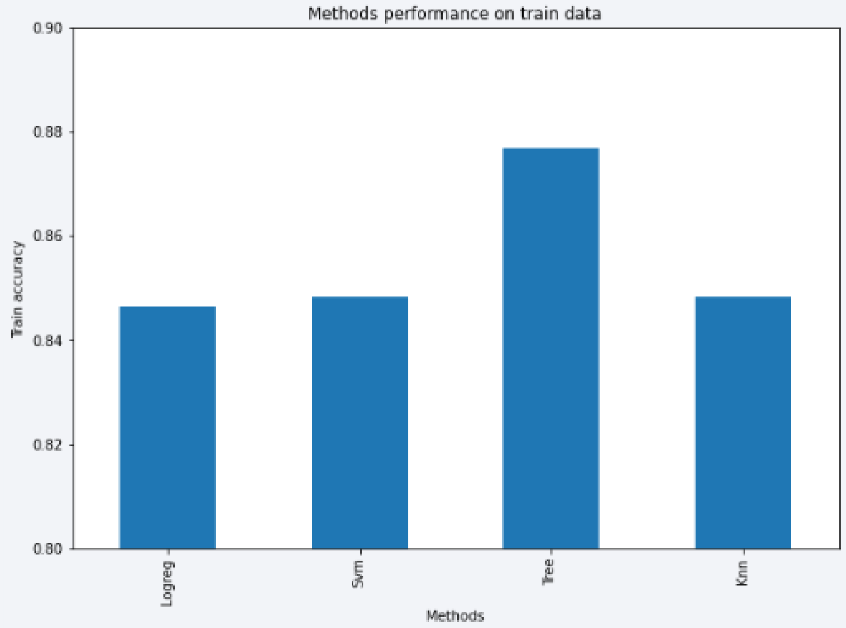
Section 5

Predictive Analysis (Classification)

Classification Accuracy

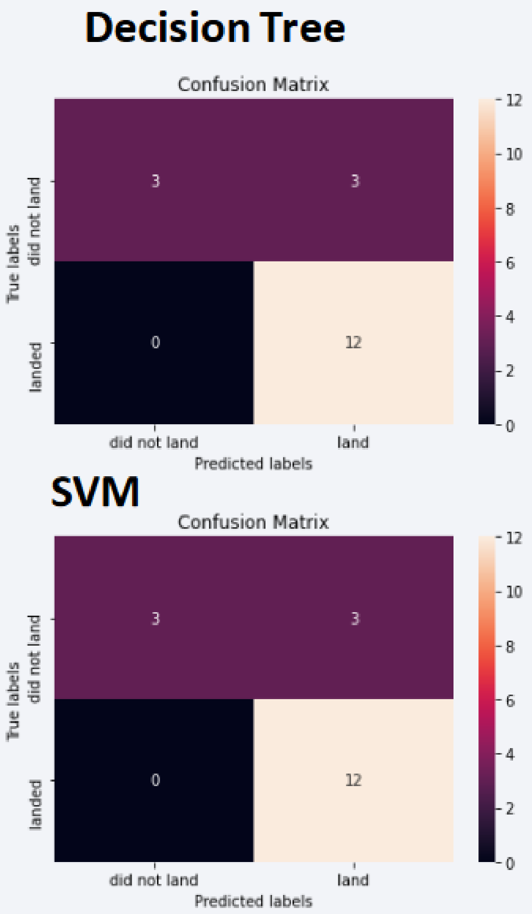
- All methods performed similar in accuracy.

- But we will take the decision tree.



Confusion Matrix

- the confusion matrices for the best two are identical.
- They suffer from false positives



Conclusions

- The success of a space mission depends on factors like the launch site, orbit type, and the number of previous launches. More launches usually lead to better success due to gained knowledge.
- The best orbits for success are GEO, HEO, SSO, and ES-L1. Lighter payloads generally perform better than heavier ones, depending on the orbit.
- We don't know why some launch sites, like KSC LC-39A, are more successful, but we could investigate atmospheric data for answers.
- We chose the Decision Tree Algorithm for our analysis because it showed better training accuracy, even though all models had similar test accuracy.

Thank you!