IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stefan Merdian
07.06.2024

Outline

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

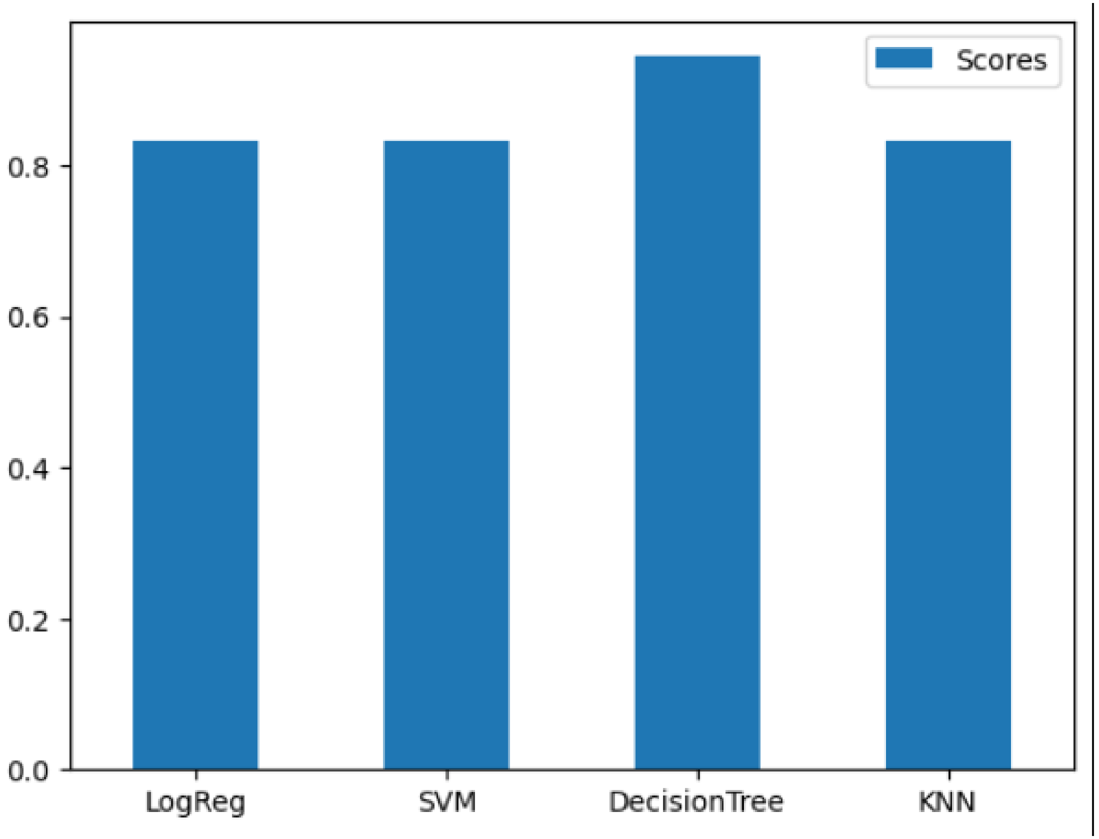
Executive Summary

- Summary of methodologies
  1. **Data Collection:**
     - API Request
     - Data Wrangling and Formatting
  2. **Data Wrangling:**
     - Exploratory Data Analysis (EDA)
     - Label Determination
  3. **Exploring and Preparing Data:**
     - Exploratory Data Analysis (EDA)
     - Feature Engineering
  4. **Machine Learning Prediction:**
     - Pipeline Creation
     - Model Training
     - Model Evaluation

Executive Summary

- Summary of results

We created four different classification models: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Decision Tree. Among these, the **Decision Tree** model achieved the highest accuracy.



Introduction

- Project background and context:

  This project aims to predict the successful landing of SpaceX Falcon 9 first stages to help reduce launch costs.

- Problems you want to find answers:
  - Can we accurately predict the landing success of Falcon 9?
  - Which factors most influence the landing outcome?

Section 1

Methodology

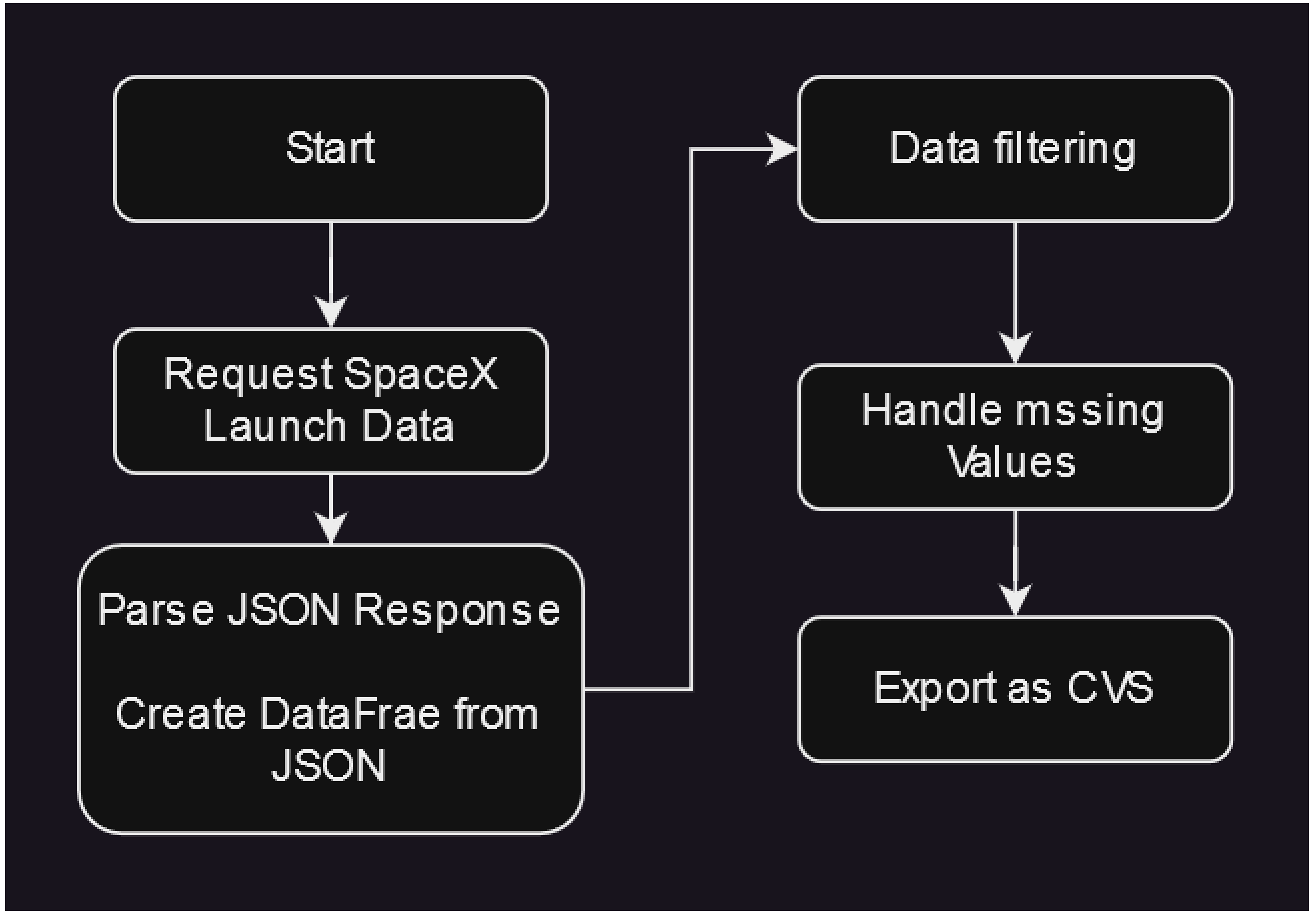
Methodology

Executive Summary

- Data collection methodology
- 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

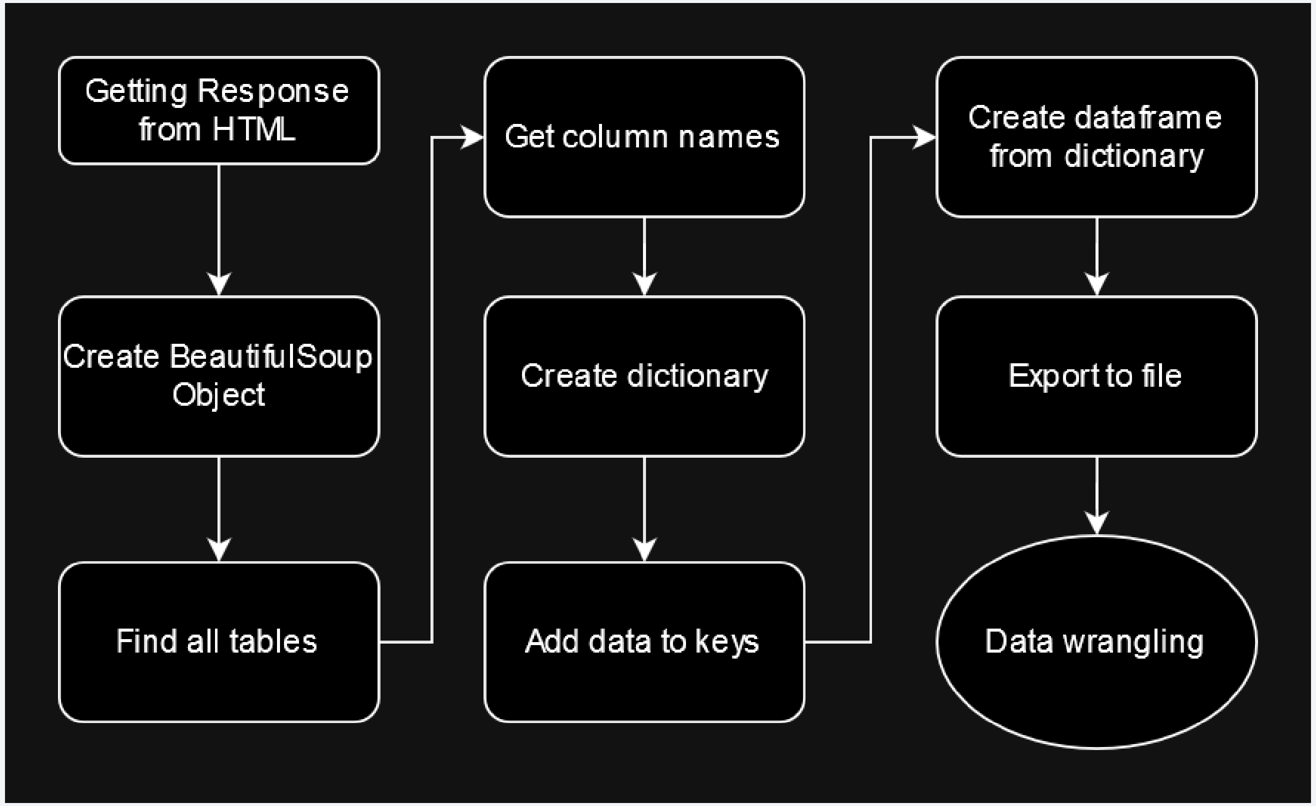
Data Collection – SpaceX API

- **API Request:** Request and parse SpaceX launch data using GET request
- **Observation:** Many fields contain only IDs
- **Detailed Data Retrieval:** Use API to get detailed information using IDs
- https://github.com/yamisukii/Data-Science-Capstone-Predicting-Falcon-9-First-Stage-Landing-Outcomes/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



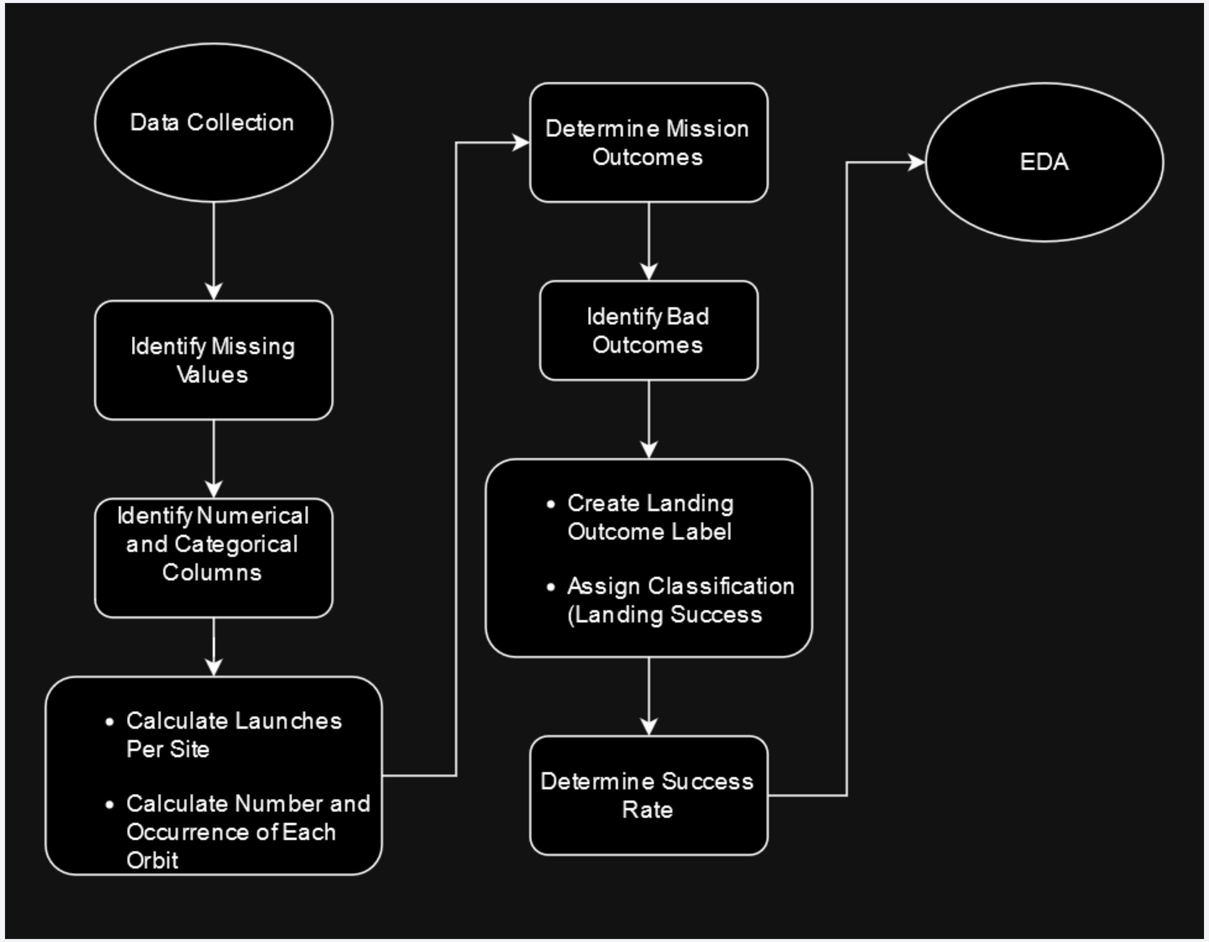
Data Collection - Scraping

- **Web Scraping**: Collect data from web pages using web scraping techniques
- **Observation:** Many fields contain unstructured or semi-structured data.
- **Data Parsing:** Parse HTML content to extract relevant information and convert it into a structured format.
- https://github.com/yamisukii/Data-Science-Capstone-Predicting-Falcon-9-First-Stage-Landing-Outcomes/blob/main/Capstone_Webscraping.ipynb



Data Wrangling

- **Transform:** We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.
- https://github.com/yamisukii/Data-Science-Capstone-Predicting-Falcon-9-First-Stage-Landing-Outcomes/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

We used:

- **Catplot** – How Flight number and Payload would affect the launch outcome, success rate of launch site
- **Barchart** – relationship between success rate of each orbit type
- **Catplot** – relationship between FlightNumber and Orbit type and between Payload and Orbit type
- **Line chart** – to get the average launch success trend over the years

http://github.com/yamisukii/Data-Science-Capstone-Predicting-Falcon-9-First-Stage-Landing-Outcomes/blob/main/edadataviz.ipynb

EDA with SQL

We performed SQL queries to gather and understand data from dataset:

• Displaying 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, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.

• Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Build an Interactive Map with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

• Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).

• Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker,

folium.features.DivIcon).

• The grouping of points in a cluster to display multiple and different information for the same coordinates

• (folium.plugins.MarkerCluster).

• Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing.

• (folium.map.Marker, folium.Icon).

• Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them.

(folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

• These objects are created in order to understand better the problem and the data. We can show easily

all launch sites, their surroundings and the number of successful and unsuccessful landings.

Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, rangeslider and scatter plot components

• Dropdown allows a user to choose the launch site or all launch sites

• (dash_core_components.Dropdown).

• Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie).

• Rangeslider allows a user to select a payload mass in a fixed range

• (dash_core_components.RangeSlider).

• Scatter chart shows the relationship between two variables, in particular 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

- Selection of 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 the best accuracy will be chosen (see Notebook for result)

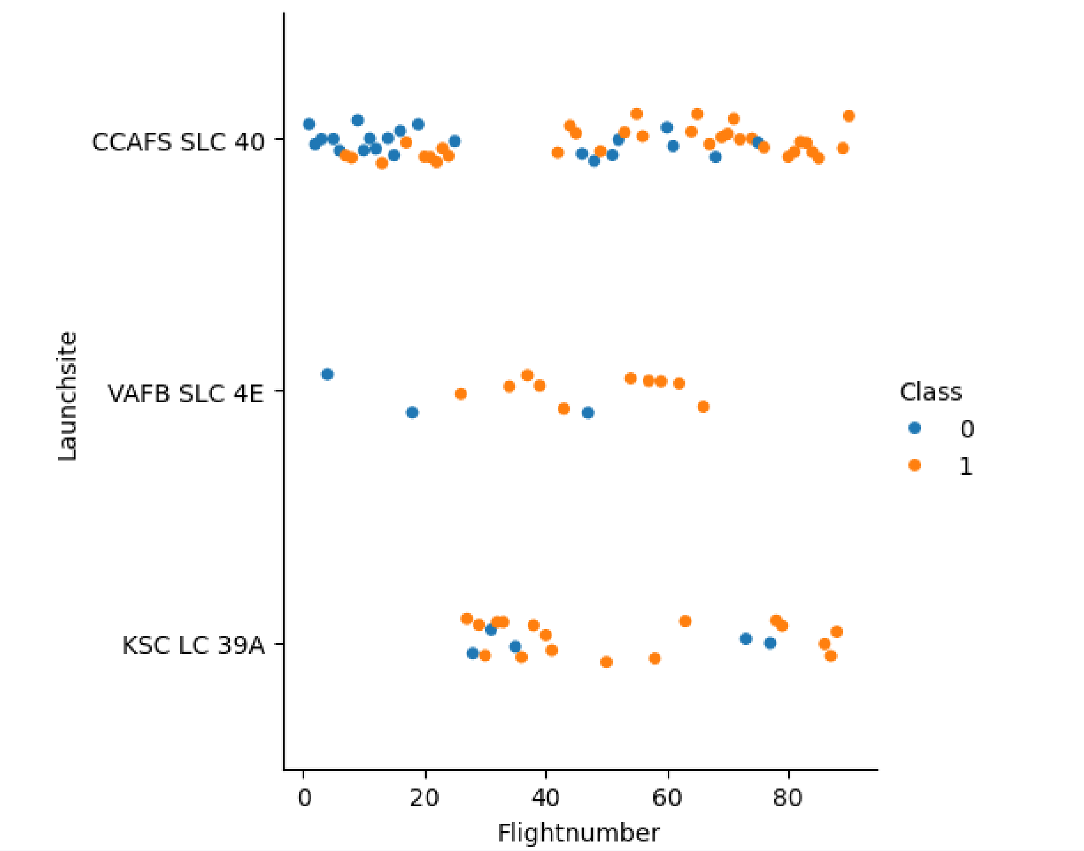
https://github.com/yamisukii/Data-Science-Capstone-Predicting-Falcon-9-First-Stage-Landing-Outcomes/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Section 2

Insights drawn from EDA

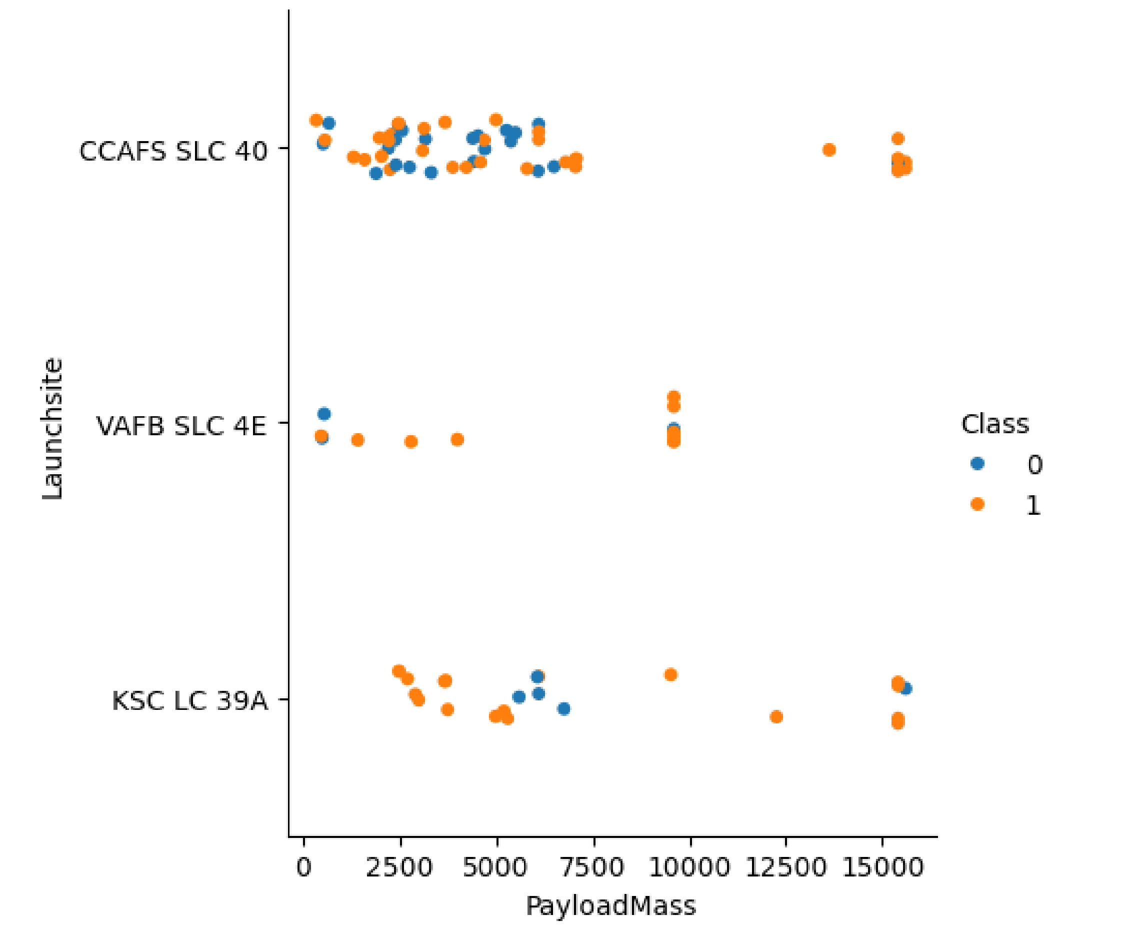
Flight Number vs. Launch Site

We **observe** that, for each site, the success rate is increasing



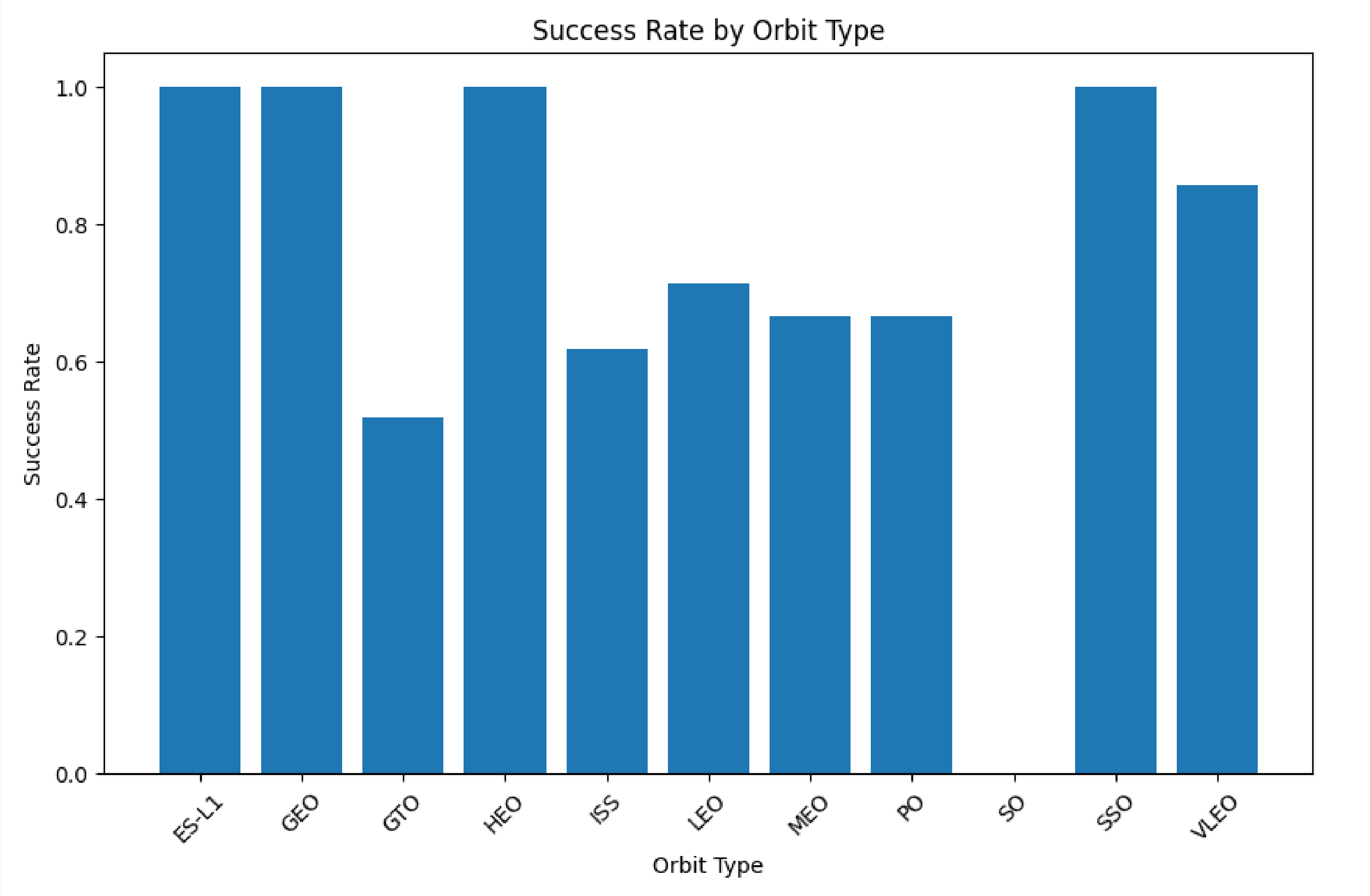
Payload vs. Launch Site

- Depending on the launch site, a heavier payload may be a consideration for a successful landing.



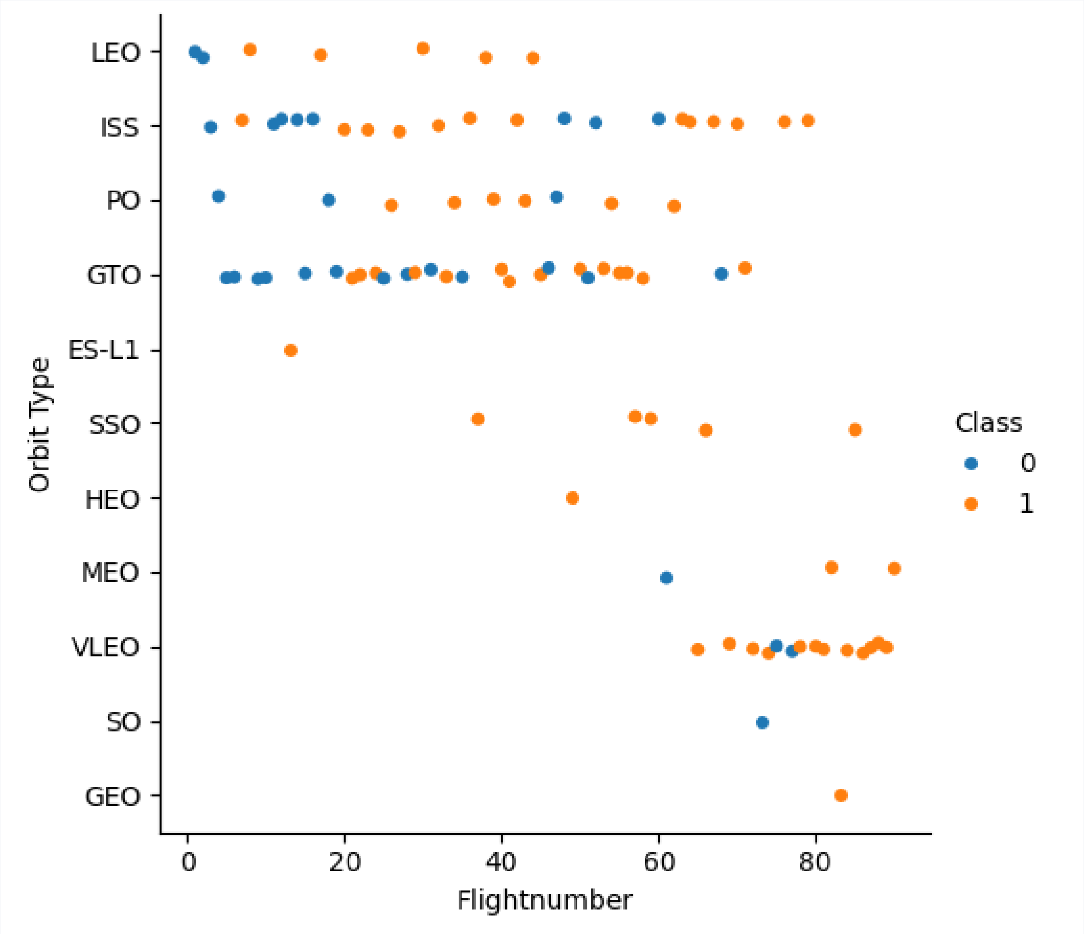
Success Rate vs. Orbit Type

- With this plot, we can see success rate for different orbit types. Wenote that ES-L1, GEO,HEO, SSO have the best success rate.



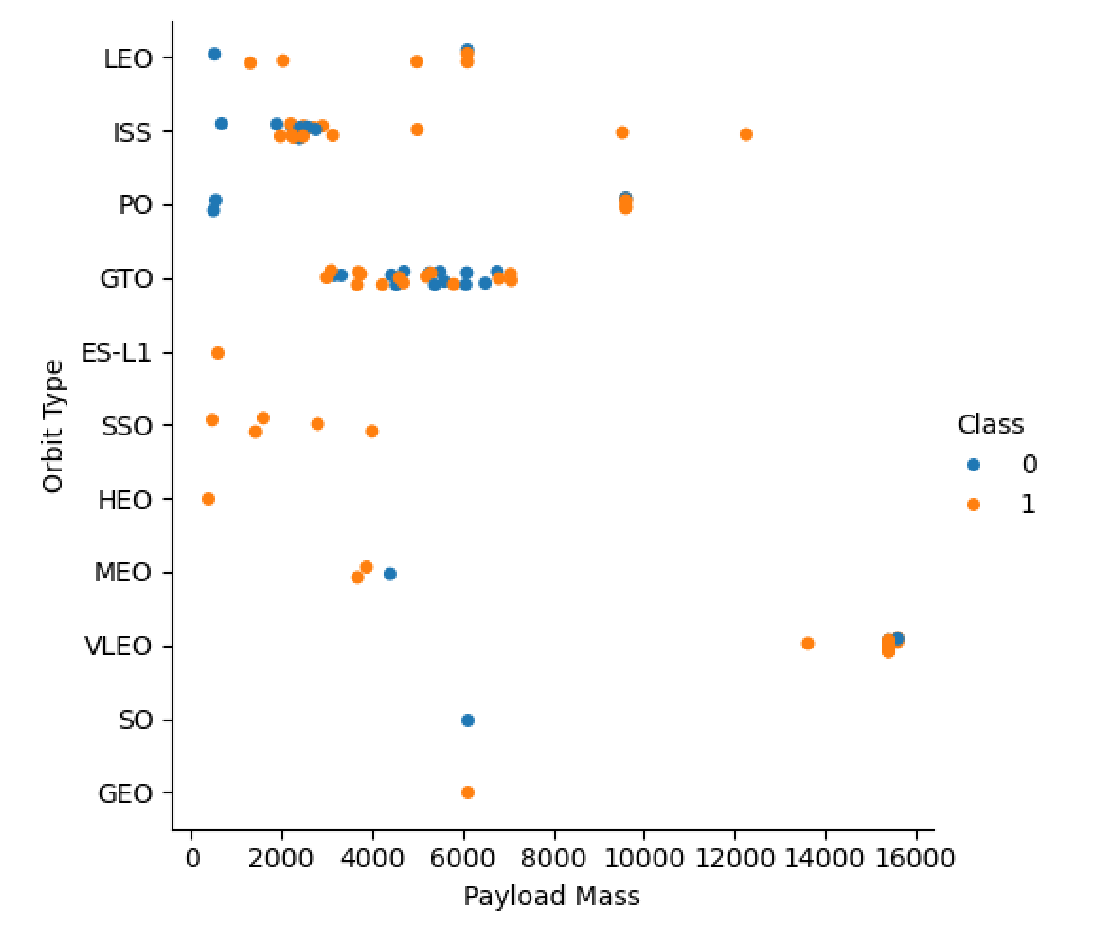
Flight Number vs. Orbit Type

- We notice that the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights.
- But we can suppose that thehigh success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches for other orbits.

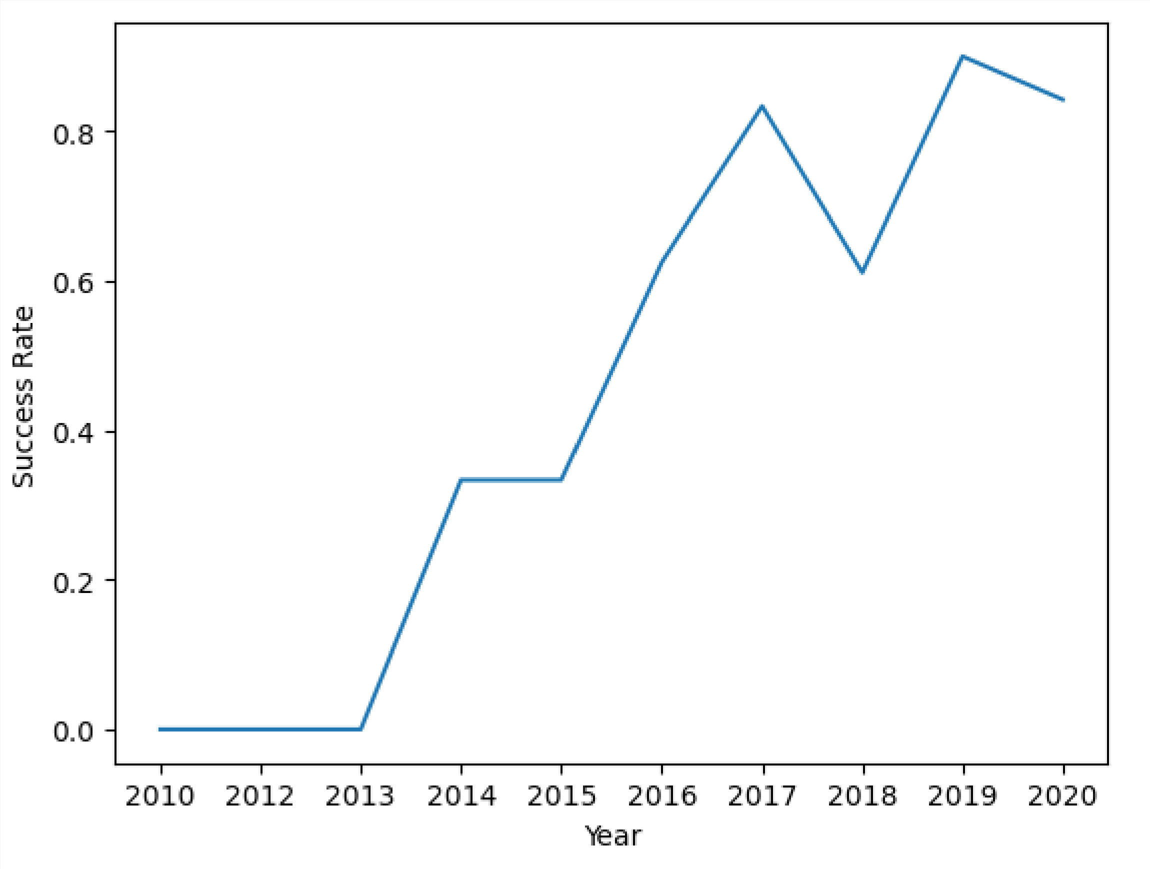


Payload vs. Orbit Type

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



Launch Success Yearly Trend



All Launch Site Names

SQL Query

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

Results

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

Explanation

The use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE.

Launch Site Names Begin with 'CCA'

SQL Query

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

Explanation

The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering.

Results

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

SQL Query

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

Results

SUM("PAYLOAD_MASS__KG_")
45596

Explanation

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

Average Payload Mass by F9 v1.1

SQL Query

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

Results

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

Explanation

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

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

Results

MIN("DATE")
01-05-2017

Explanation

With this query, we select the oldest successful landing.
The WHERE clause filters dataset in order 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

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

Results

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

Explanation

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

```
%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
```

Results

SUCCESS	FAILURE
100	1

Explanation

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

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

Explanation

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.

Results

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

SQL Query

```
%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'
```

Results

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

Explanation

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

```
%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 ;
```

Results

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Explanation

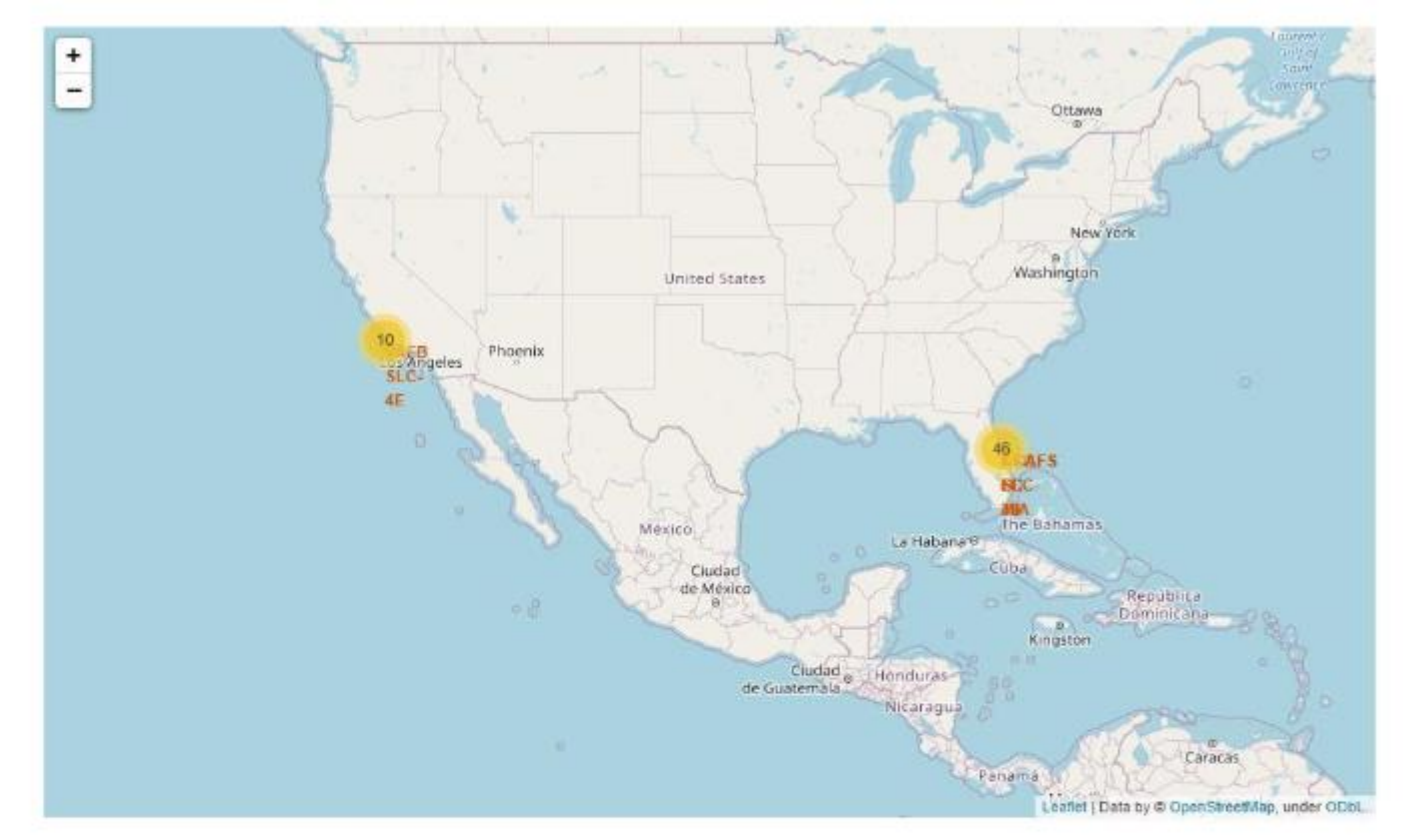
This query returns landing outcomes and their count where mission was successful and date is between 04/06/2010 and 20/03/2017. The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

Section 3

Launch Sites Proximities Analysis

Folium Map – Ground Stations

- We see that Space X launch sites are located on the coast of the United States



Folium Map – Uncuccessful launches

Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSCLC-39A has a higher launch success rate.

Folium Map – Distances between CCAFS SLC-40 and it's Proximities

Section 4

Build a Dashboard with Plotly Dash

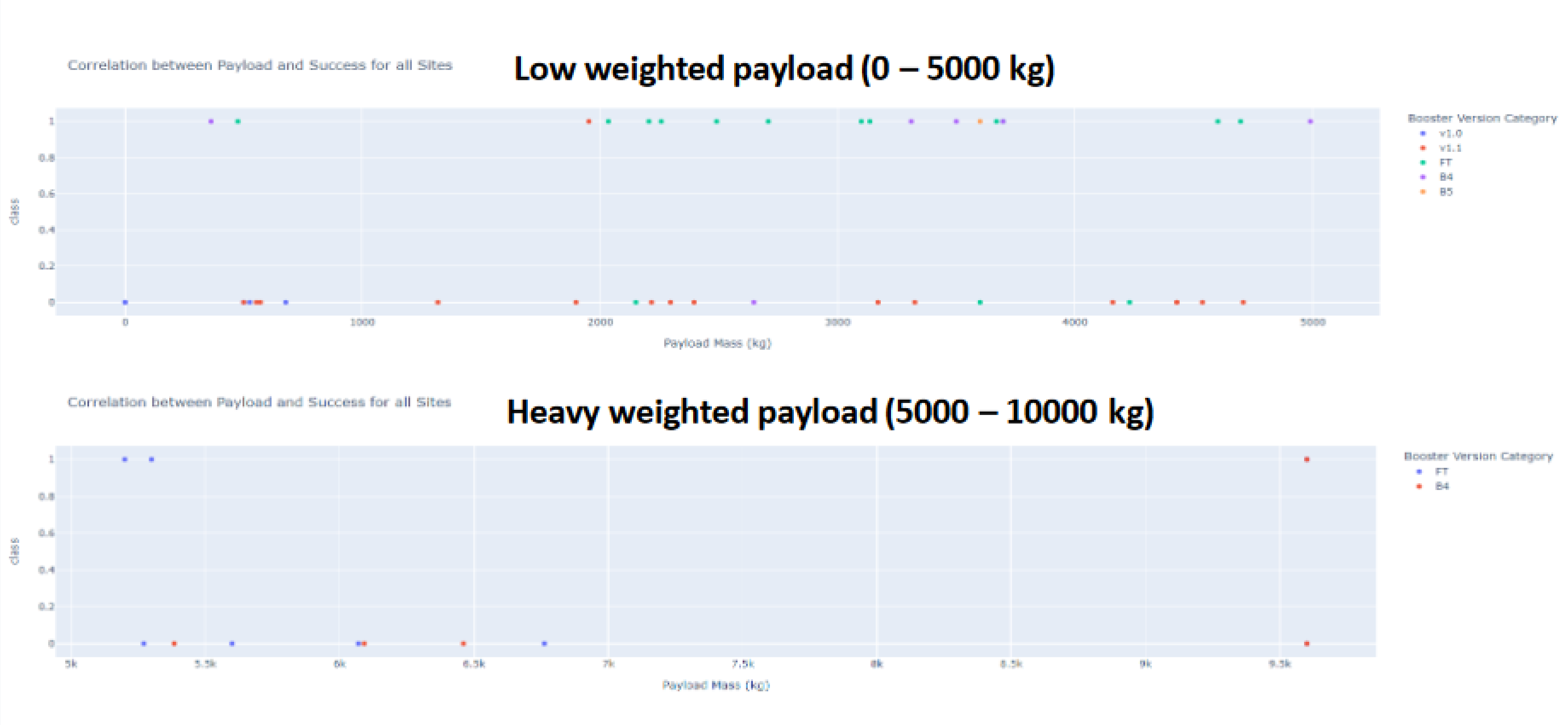
Dashboard - Success rate of launches

- We see that KSC LC-39A has the best success rate of launches.

Dashboard - Success rate of KSC LC-39A

- We see that 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

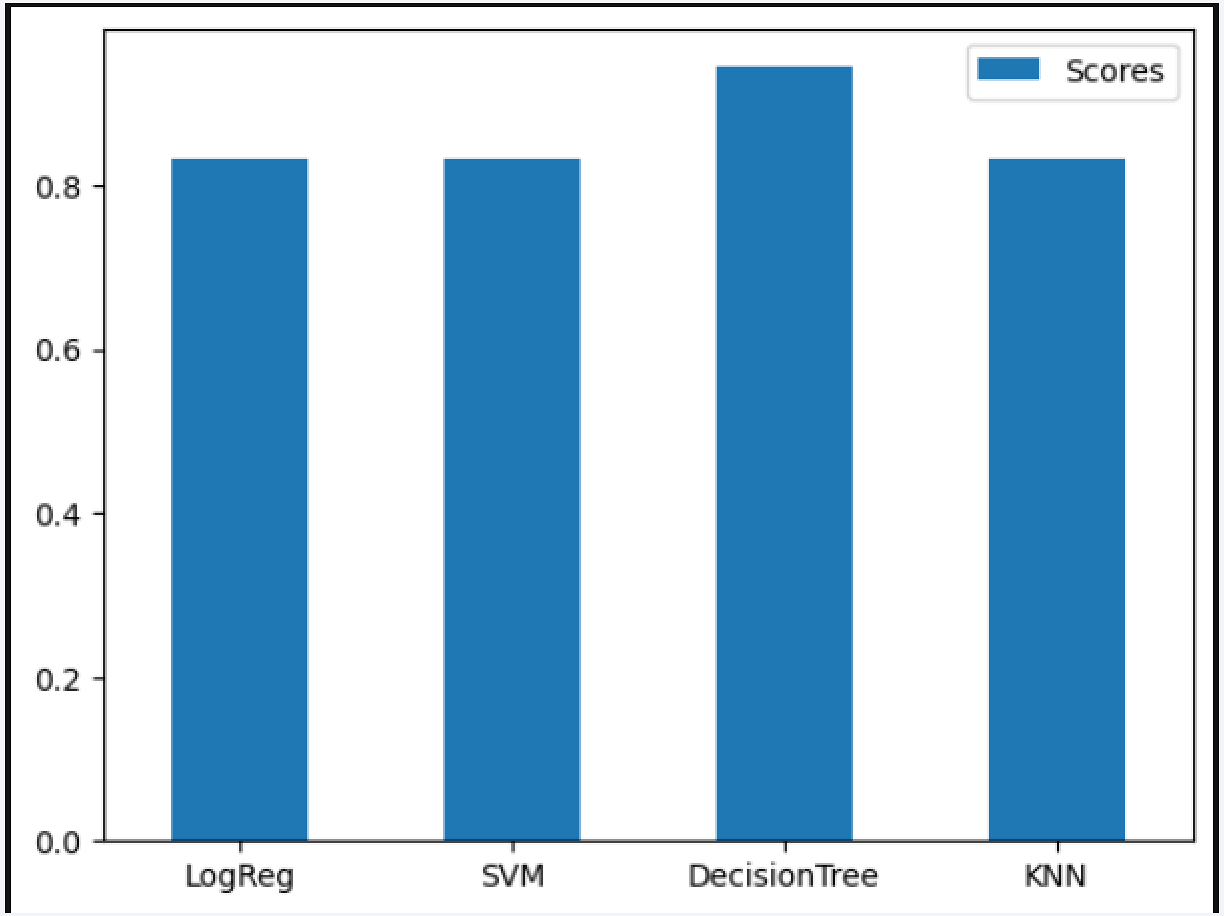


Section 5

Predictive Analysis (Classification)

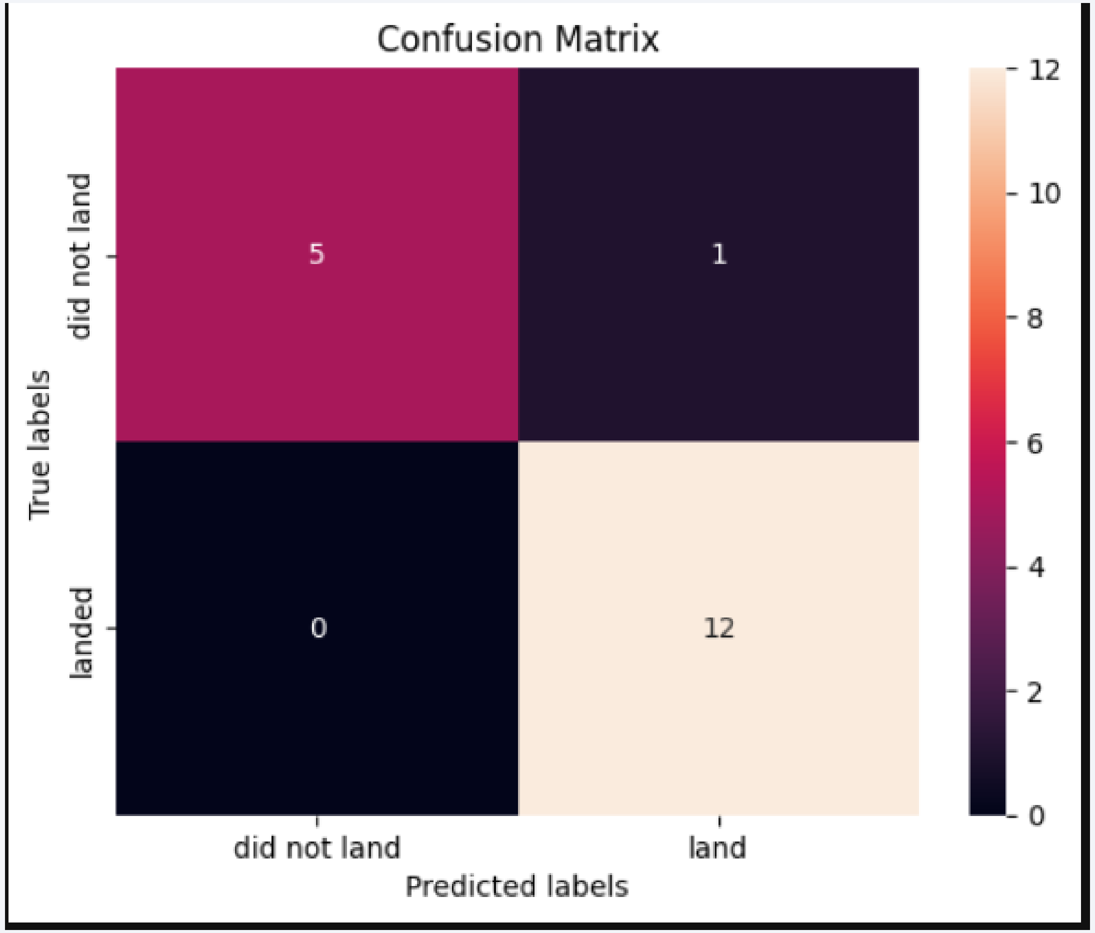
Classification Accuracy

- Decision Tree has the highest accuracy rate for the test data



Confusion Matrix – Decision Tree

- **Strong Prediction:** Model accurately predicts successful landings.
- **False Positives**: Some unsuccessful landings are incorrectly predicted as successful.
- **Improvement Needed:** More data required to reduce false positives and enhance prediction accuracy.



Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose Decision Tree Algorithm because it has the best test accuracy, but more data required to reduce false positives and enhance prediction accuracy.

Thank you!