Winning The Space Race With Data Science

Dominique Wilson

24 March 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- In an era where space exploration is open to private enterprises, the commercial space industry is experiencing a revolution. SpaceX, with its groundbreaking achievements and cost-efficient Falcon 9 rocket launches, stands at the forefront.
- Now, as Space Y seeks to challenge SpaceX's dominance, this report focuses on determining whether the Falcon 9 first stage will successfully land. If we can determine if the first stage will land, we can determine the cost of a launch.
- The data shows positive correlation between the features and outcome of the launches, showing that a decision tree is the most successful machine learning algorithm to determine whether the Falcon 9 first stage will successfully land.

INTRODUCTION

Background and context

SpaceX has revolutionized the space industry with its groundbreaking approach, offering Falcon 9 rocket launches at a remarkable price point of $62 million, significantly lower than competitors' costs, which are priced upwards of $165 million per launch.

It is largely believed that SpaceX's cost-saving strategy is down to their unique concept of reusing the first stage of the Falcon 9, thereby driving down launch expenses with each successful re-landing.

Here at SpaceY we are poised to challenge SpaceX, our mission is clear: to develop robust machine learning capable of predicting the landing outcome of Falcon 9's first stage, a key stage in determining the optimal pricing strategy to compete effectively with SpaceX for rocket launch contracts.

Key challenges

- Identifying the factors that influence the landing outcome.
- Understanding the relationships between variables and their impact on the landing result.
- Uncovering the conditions to enhance the likelihood of a successful landing.

METHODOLOGY

Dominique Wilson

METHODOLOGY

1. Data collection:
   - SpaceX API
   - Web scraping from Wikipedia
2. Data wrangling:
   - One-hot encoding applied to categorical features
3. Exploratory data analysis (EDA):
   - visualisation
   - SQL
4. Data visualisation:
   - Folium
   - MatPlotLib
   - Seaborn
5. Predictive analysis using Machine Learning Models:
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

Data Collection Methodology - API

Data Collection - SpaceX API
1. Get response from API
2. Convert response to JSON file
3. Transform data
4. Create data dictionary
5. Create data frame
6. Filter data frame
7. Export to CSV file

- API https://api.spacexdata.com/v4/rockets/

```
# Takes the dataset and uses the launchpad column to call the API and append the data to the list
def getLaunchSite(data):
    for x in data['launchpad']:
       if x:
         response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
         Longitude.append(response['longitude'])
         Latitude.append(response['latitude'])
         LaunchSite.append(response['name'])
```

- Data was taken from the API relating to SpaceX launches, then filtered to include only Falcon 9 launches

- Data was cleaned with missing values replaced

Full SpaceX API data collection is available on GitHub here

Data Collection Methodology - Scraping

Data Collection - Scraping
1. Request Wikipedia HTML
2. BeautifulSoup
3. Locate launch info HTML
4. Create data dictionary
5. Extract data to dictionary
6. Cast dictionary to data frame

- API
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

- Web scraping was performed to collect Falcon 9 historical launch records

- Data was cleaned with missing values replaced

Scraping data collection is available on GitHub here

Data Wrangling

- Data Wrangling was performed to find patterns in the data and determine the label for training supervised models.
- There are several different outcomes for successful and unsuccessful landings within the data including landing in the ocean, on a ground pad and on a drone ship.
- The data was converted to turn those outcomes into Training Labels where:

1 = booster successfully landed

0 = unsuccessful landing

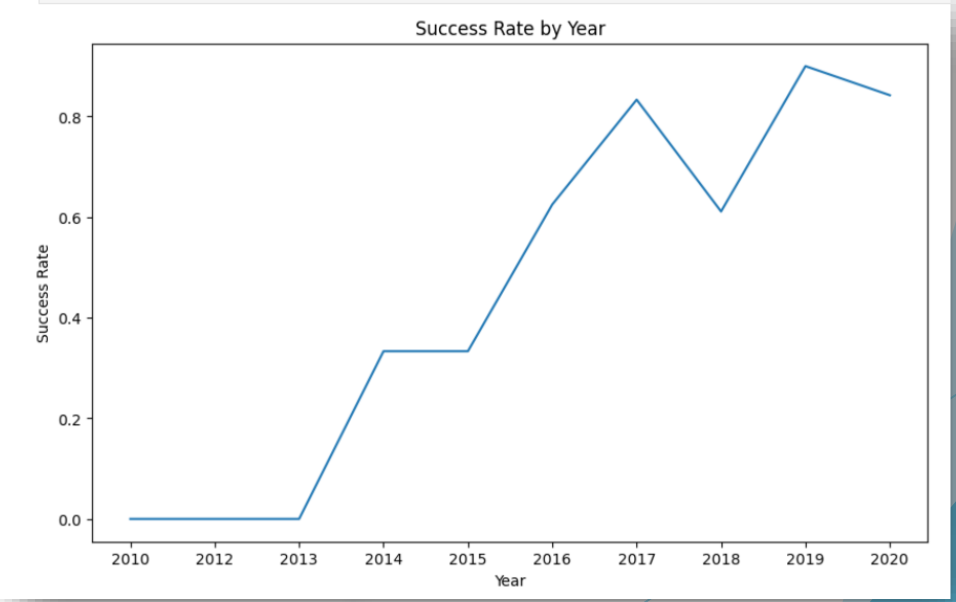
- A landing class label was then introduced to determine the success rate of the landings

Data wrangling is available on GitHub here

EDA & Data Visualisation

Relationships Explored
FlightNumber vs. PayloadMass
FlightNumber vs. LaunchSite
Payload vs. LaunchSite
Payload Mass vs. LaunchSite
Orbit vs. SuccessRate
FlightNumber vs. Orbit
Payload Mass vs Orbit

- Exploratory Data Analysis (EDA) was performed on variables including launch site, orbit information, payload mass and mission outcomes.
- The purpose of the EDA & visualisation was to determine the existence of relationships and whether they could be used to train the machine learning model



EDA is available on GitHub here

EDA with SQL

- Executing SQL queries allowed additional insight into the data including:

- Unique launch sites
- Payload mass by NASA & SpaceX
- Successful landing dates
- Booster names with certain landing conditions
- Landing outcomes

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [25]:

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome LIKE '%Success (drone ship)%'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000;
```

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

Out[25]:

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

EDA with SQL is available on GitHub here

Creating an Interactive Map

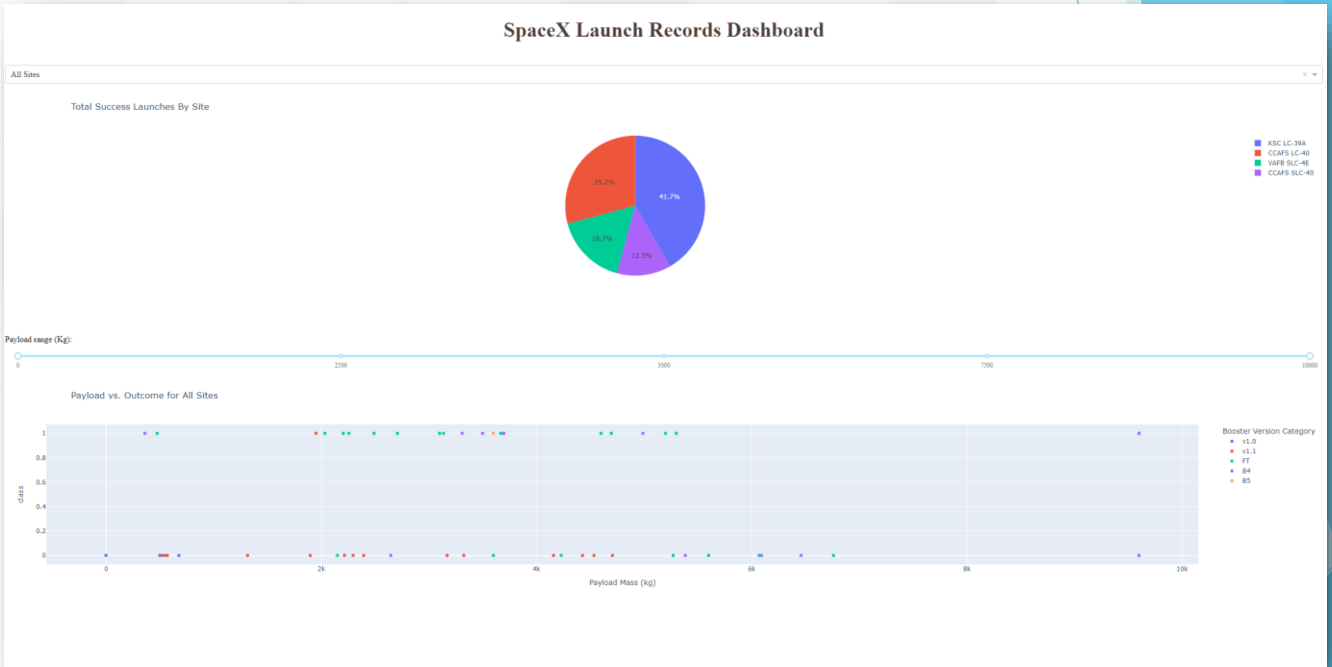
- In order to determine details about the launch site an interactive map was created with Folium
- Markers were added to indicate success/failure of launches for each site on the map
- Distances were calculated to help build a picture of the features around launch sites such as proximity to railways, highways, coastlines and cities

Folium Interactive Map is on GitHub [here]

Building A Dashboard

- Dash by Plotly was used to create an interactive dashboard with site selection and range slider options

- The dashboard displays successful launches per site as well as mission outcomes when influenced by payload mass



Dash App is available on GitHub [here]

Machine Learning – Predictive Analysis

- Scikit-learn was used to create the machine learning model & prediction phase:
  - Standardising the data
  - Splitting the data into training & test
  - Creating machine learning models:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fitting the models to the training set
  - Finding hyperparameters for each model
  - Determining the best performing classification model based on accuracy scores and confusion matrix

RESULTS

Dominique Wilson

Results – SQL

- We begin by visiting the SQL outcomes to tell the story of the data
- Throughout all of these results 1 = successful landing 0 = unsuccessful landing
- There are 4 launch sites for the Falcon 9

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

- Displaying 5 records where launch sites begin with the string CCA

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

Results - SQL

Total_Payload_Mass
45596

Average_Payload_Mass
2928.4

First_Successful_Landing_Date
2015-12-22

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

- The mission outcome counts

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

- Names of boosters with success landing on a drone ship with a payload mass between 4000 to 6000kg

Results - SQL

- Boosters that have carried the maximum payload mass

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- 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.

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

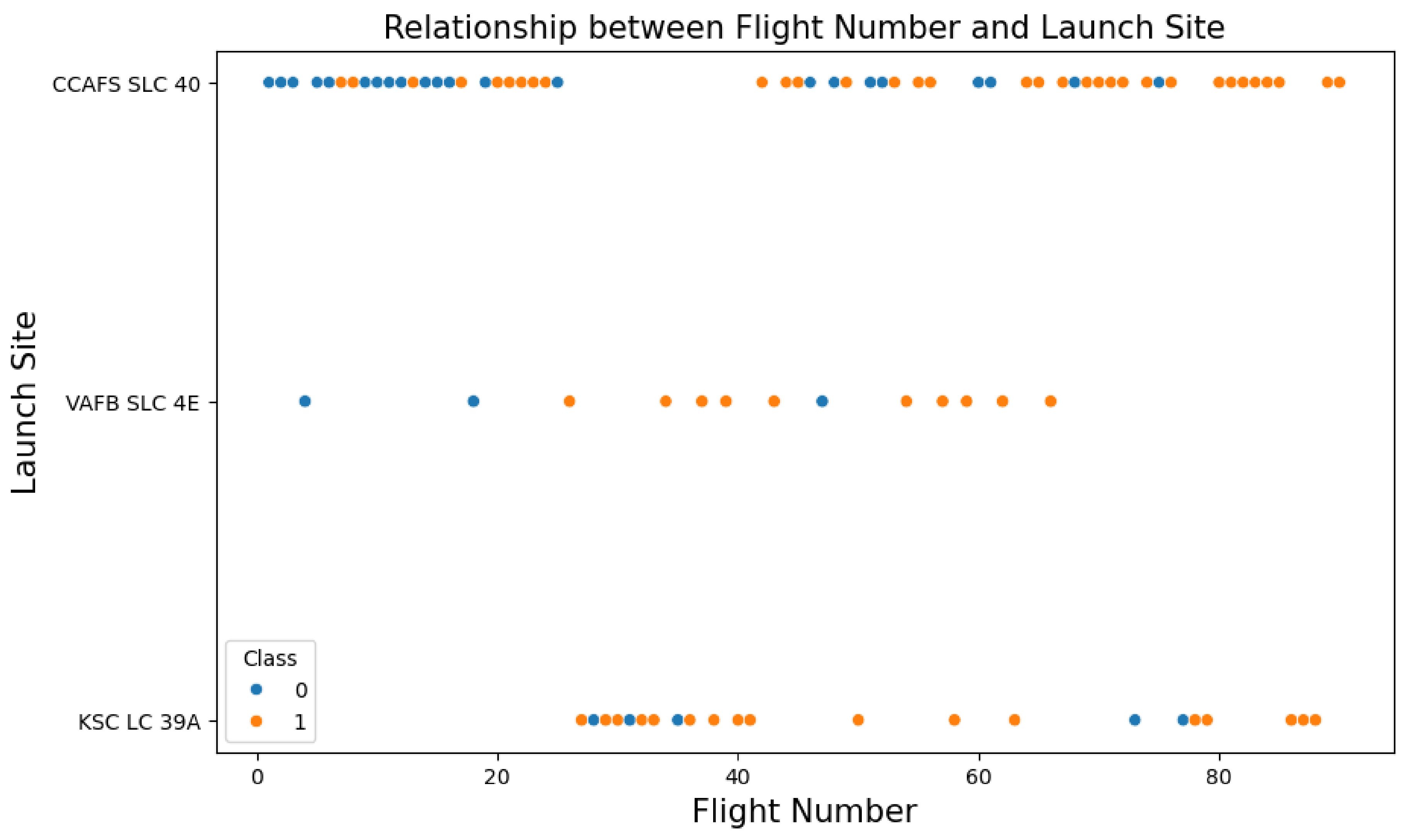
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

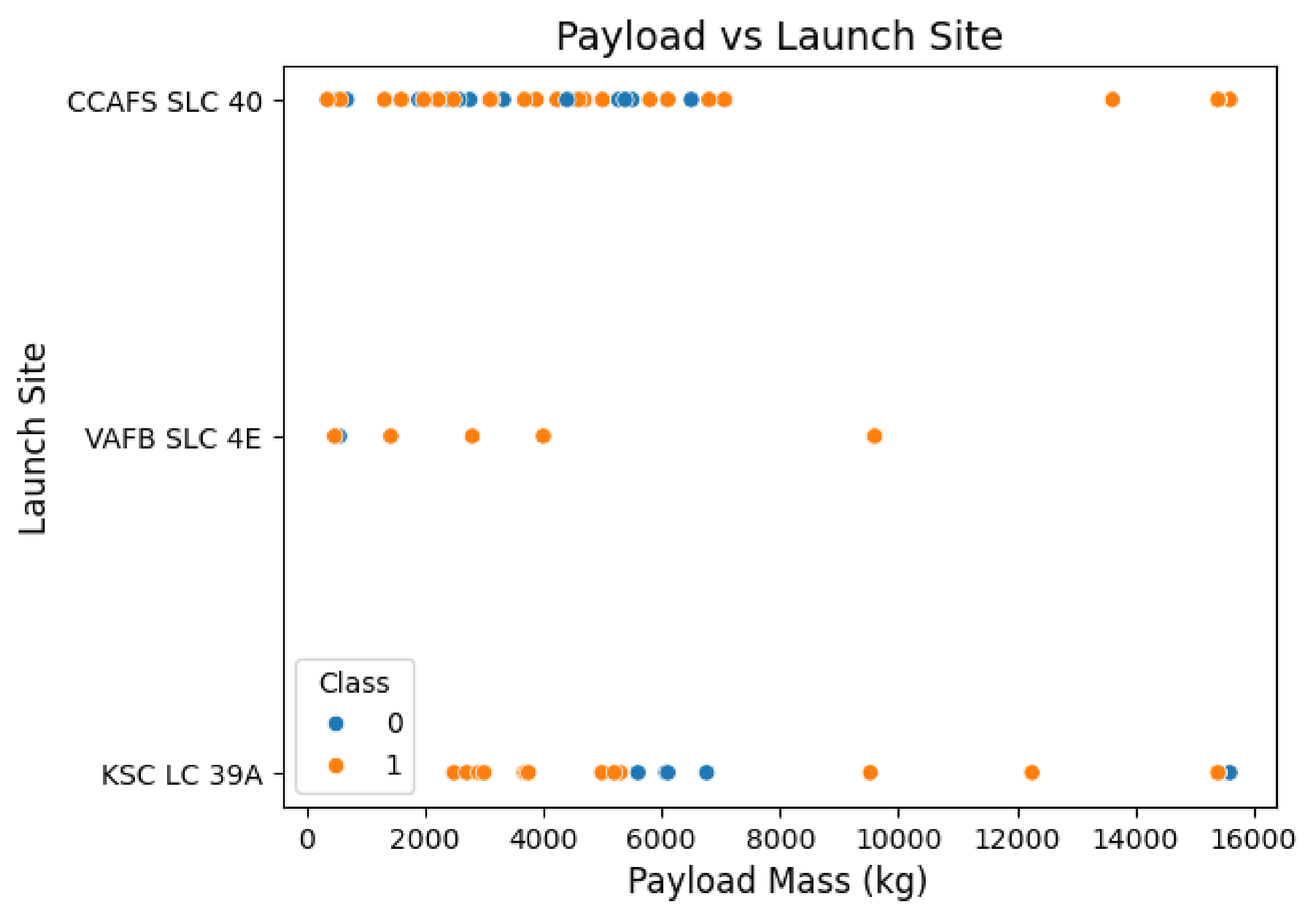
Dominique Wilson

Results – EDA Visualisation Matplotlib and Seaborn

- The relationship between flight number and launch site



Results – EDA Visualisation Matplotlib and Seaborn



- The relationship between payload mass and launch site

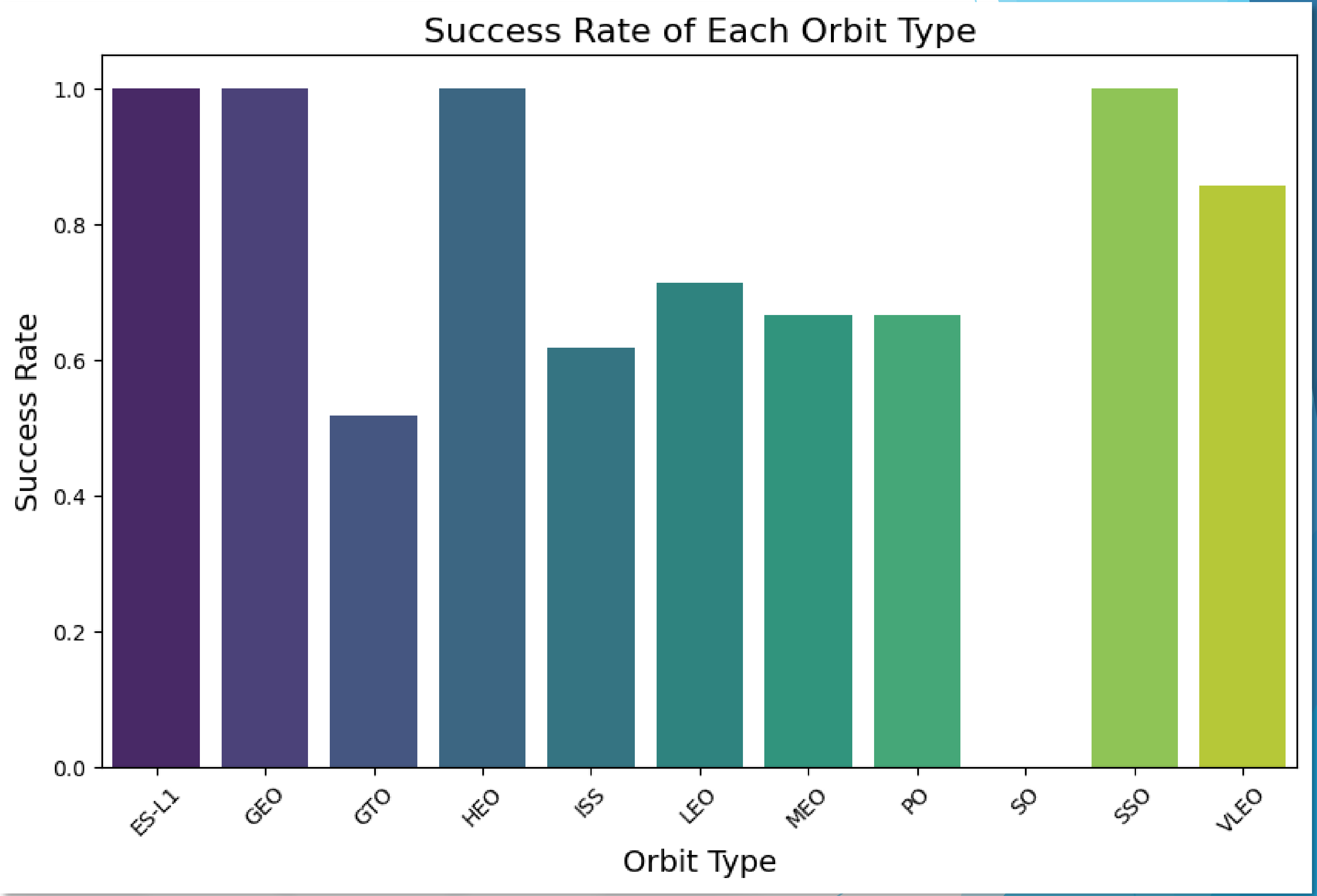
Dominique Wilson

IBM Developer

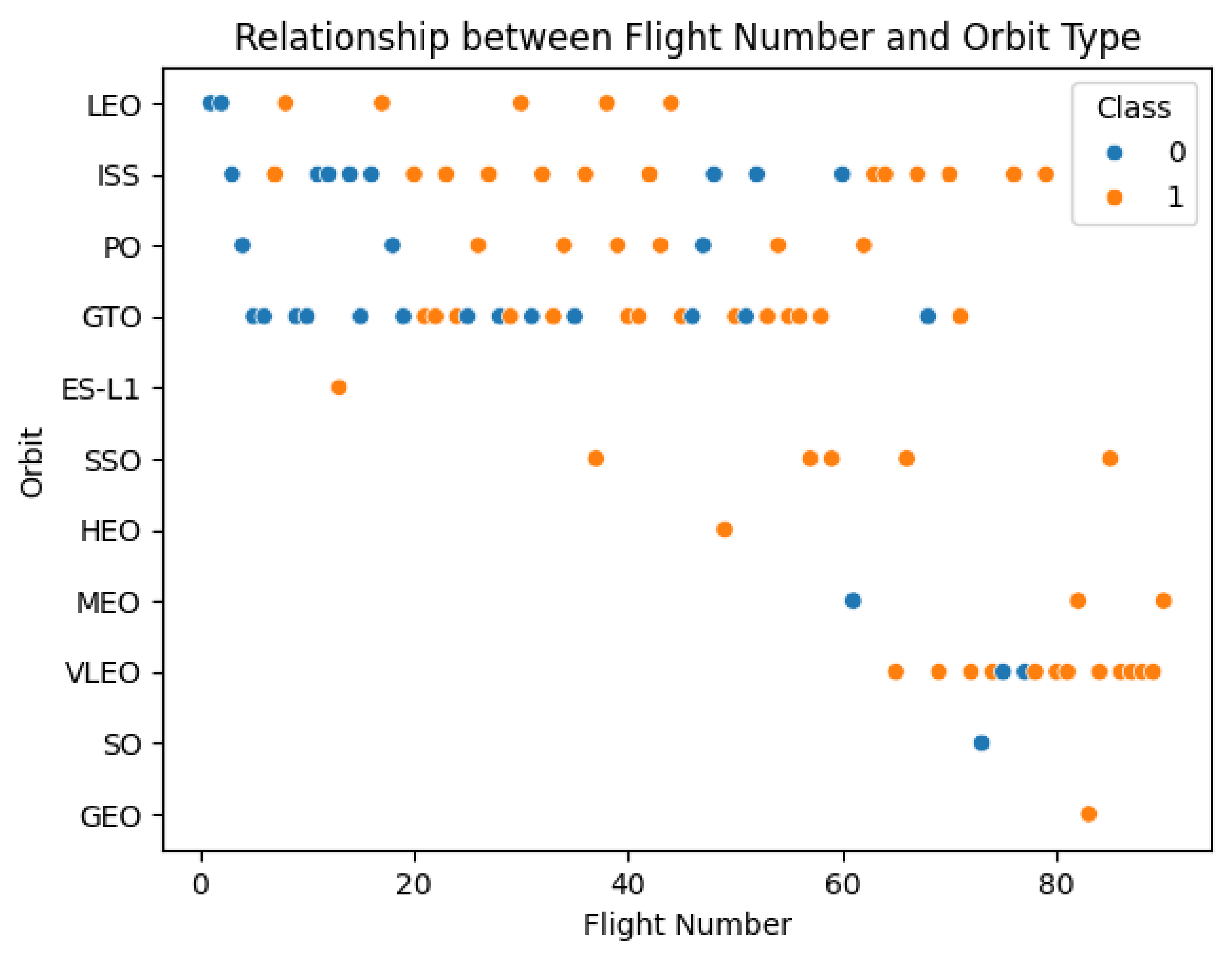
SKILLS NETWORK

Results – EDA Visualisation Matplotlib and Seaborn

- The relationship between orbit type and success rate

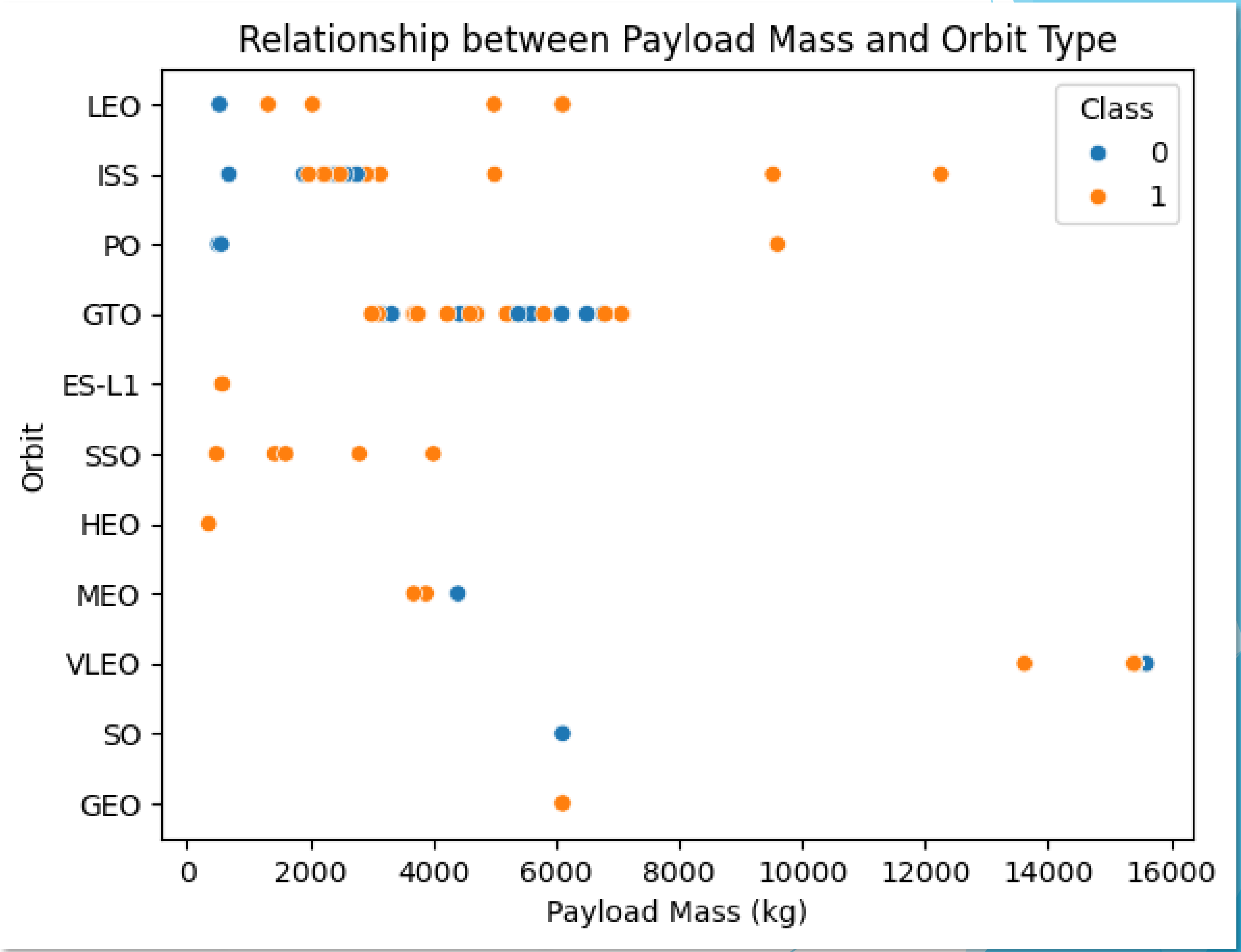


Results – EDA Visualisation Matplotlib and Seaborn



- The relationship between flight number and orbit type

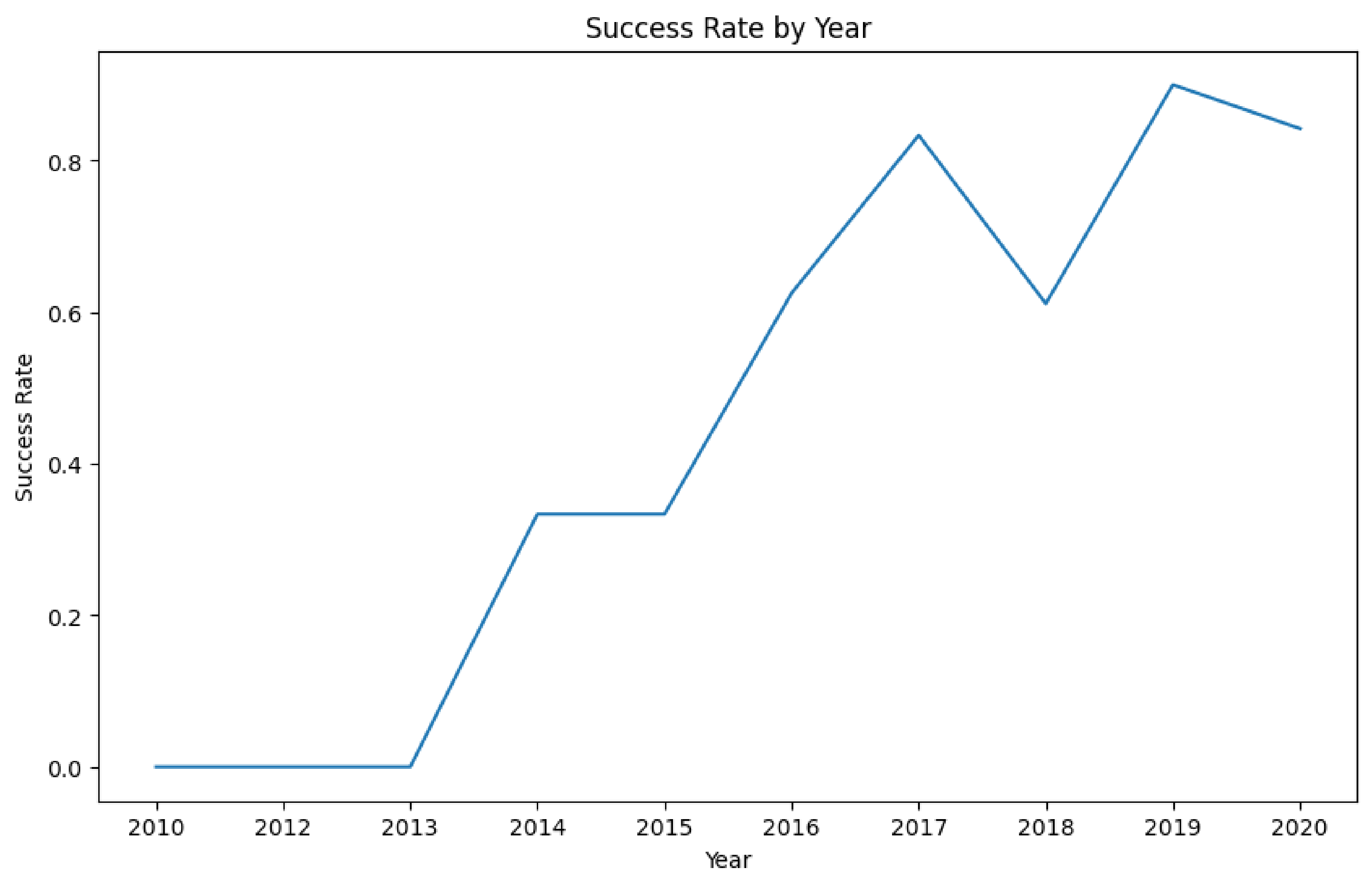
Results – EDA Visualisation Matplotlib and Seaborn



- The relationship between payload mass and orbit type

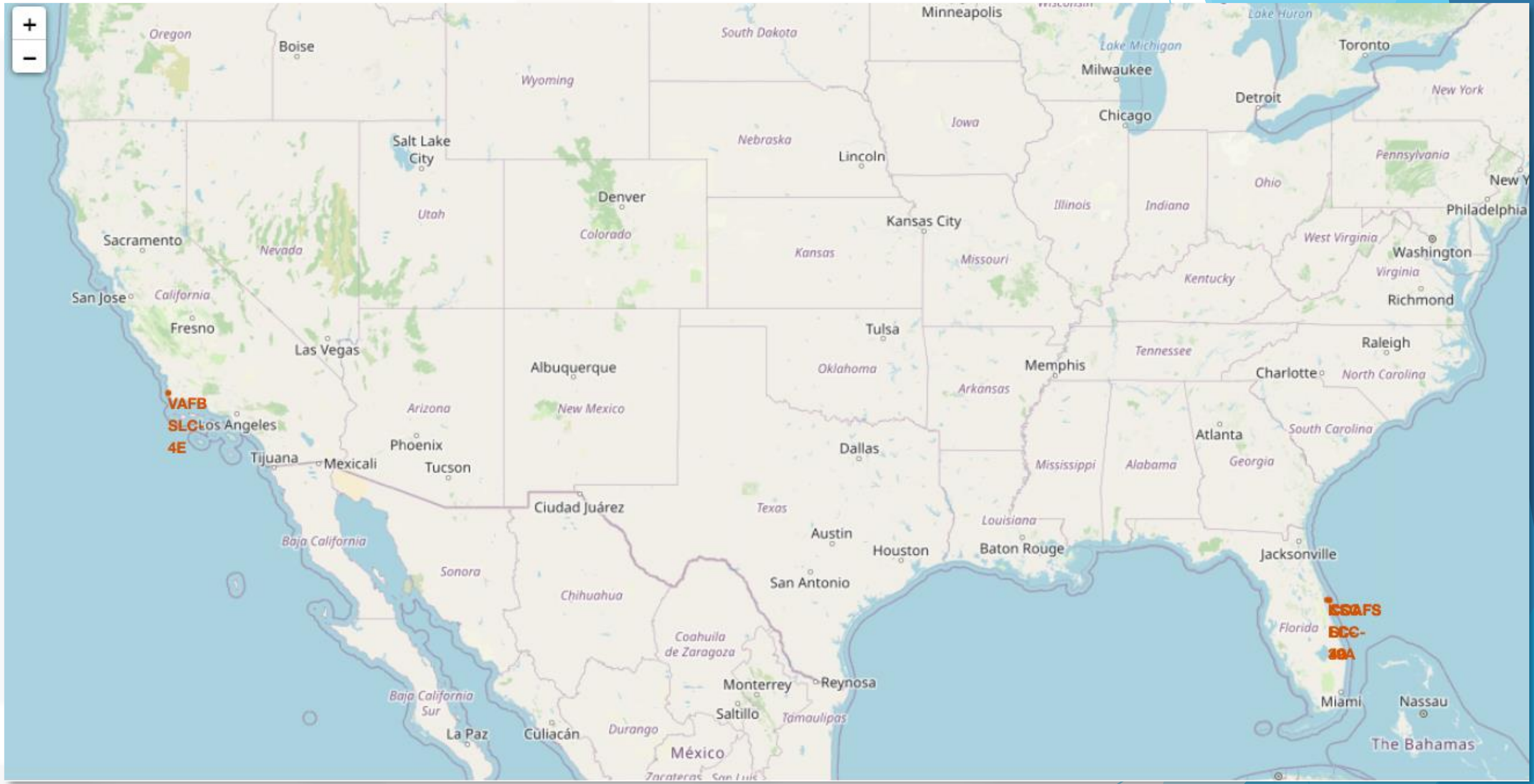
Results – EDA Visualisation Matplotlib and Seaborn

- Success rate by year trend



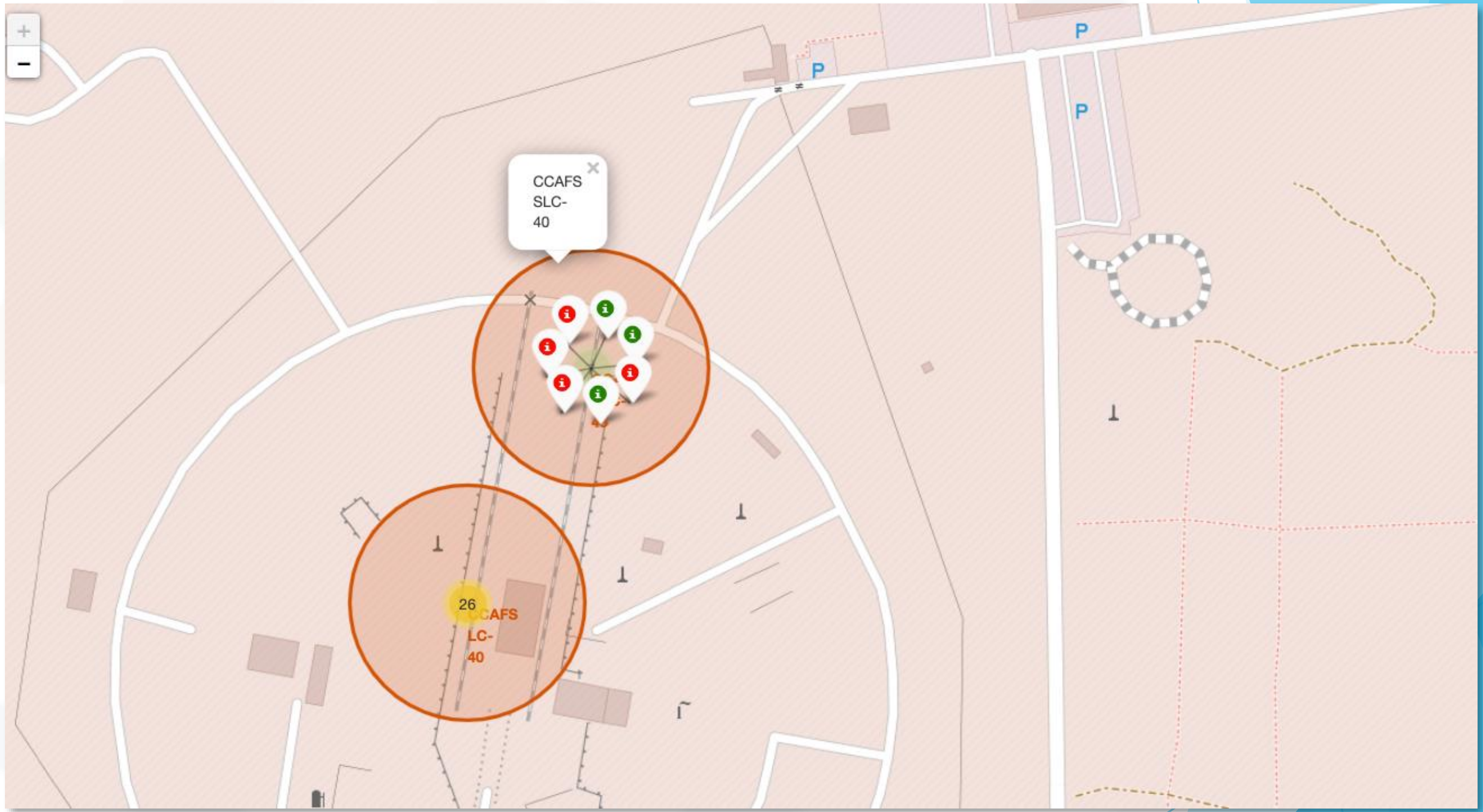
Results – Folium

- Launch sites plotted on a map



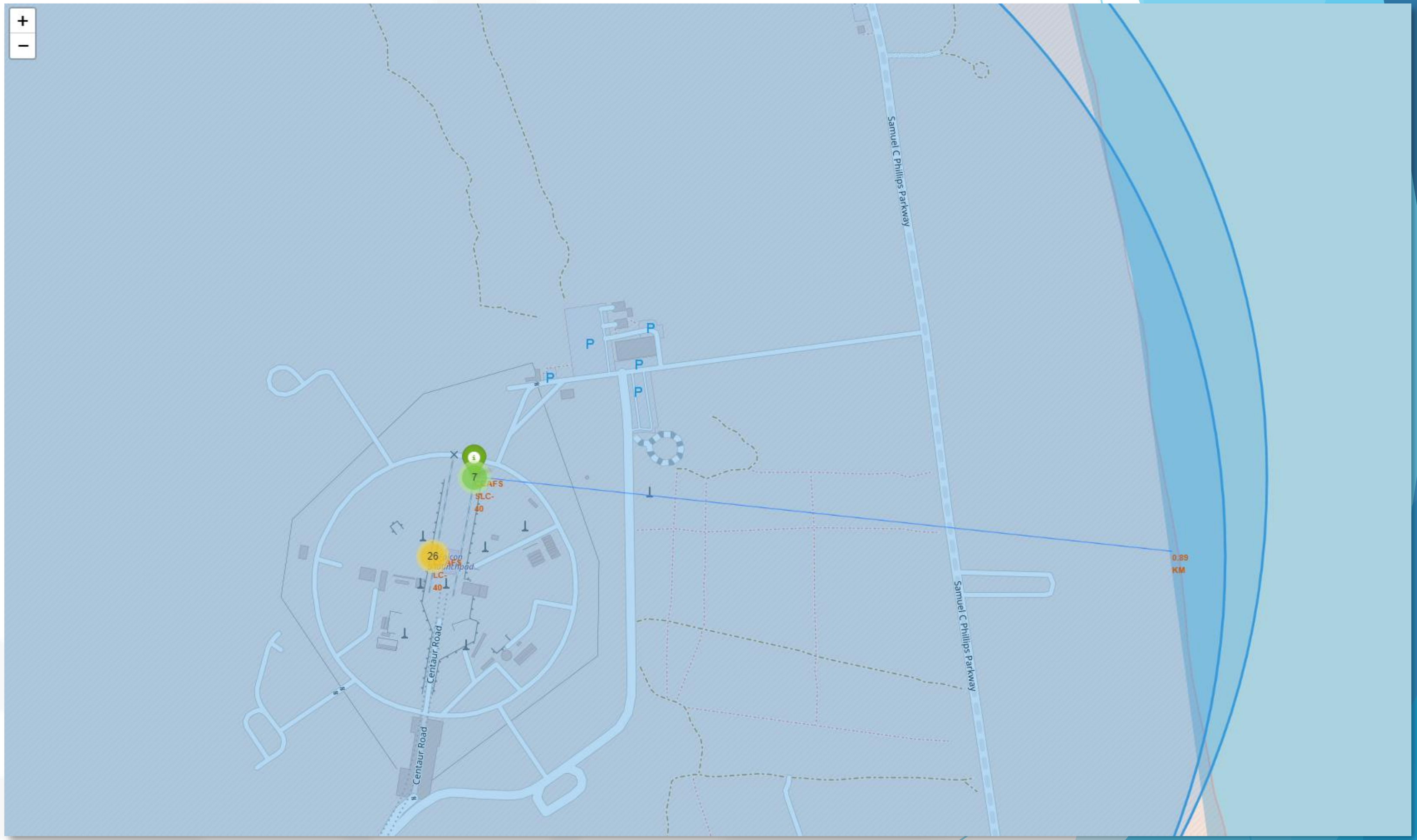
Results – Folium

- Zooming in to the map indicates successful (green) an unsuccessful (red) landings from launch sites



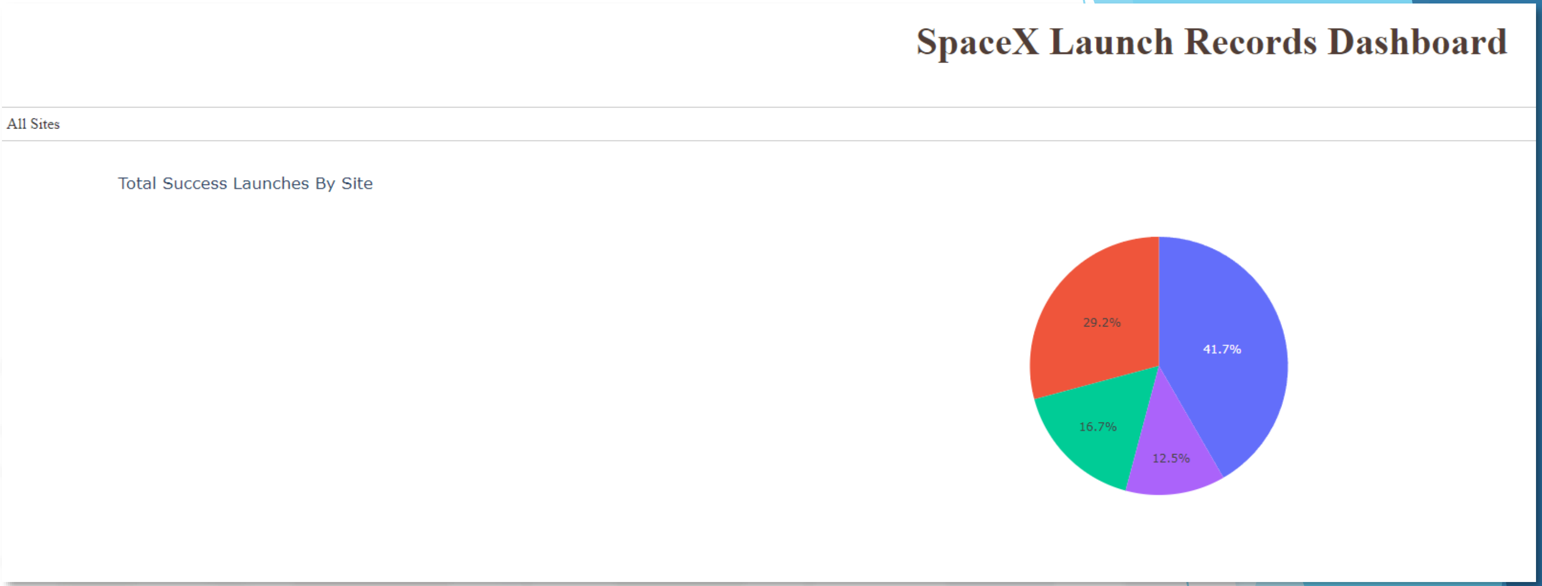
Results - Folium

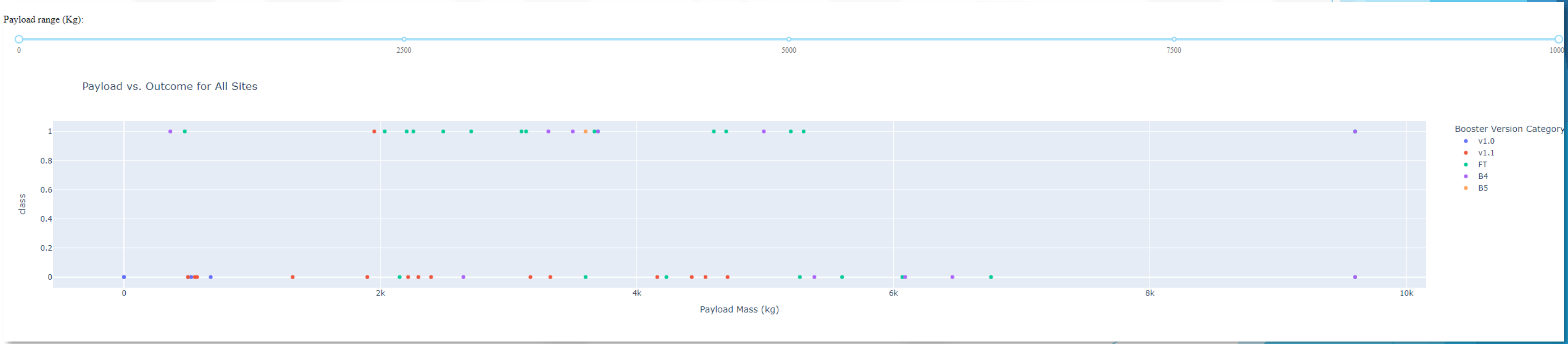
- Distance indicated from launch site to proximities such as coastline (shown here), railway, city & highway



Results - Dash

- Dash with dropdown to select site & view success launches by site
- Payload range slider to view the success launches by site and payload

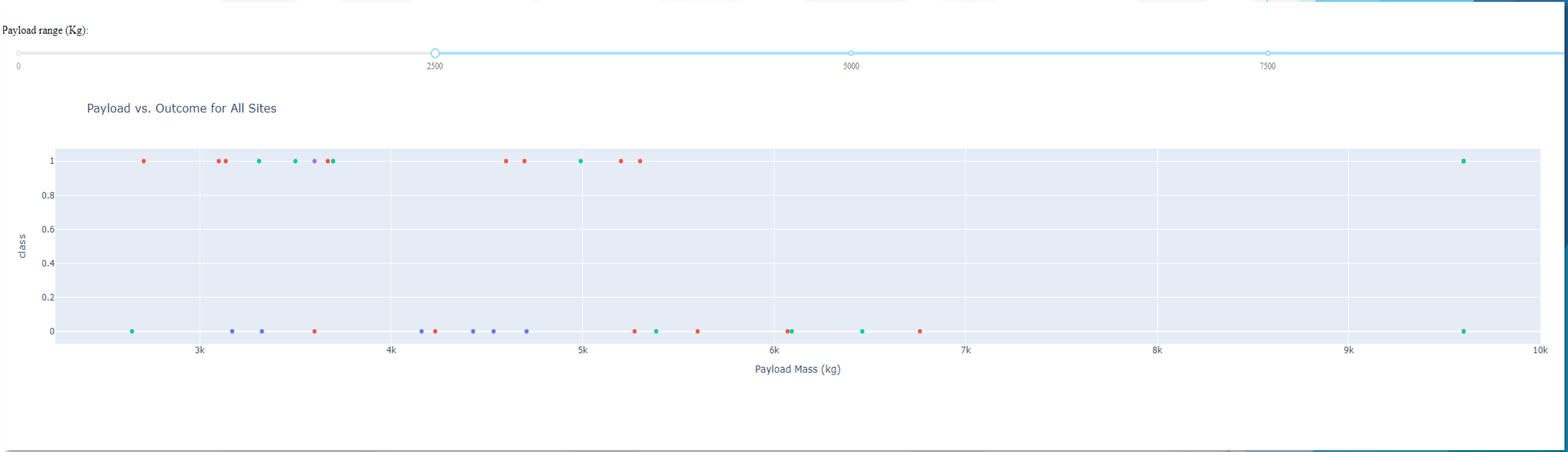




Dominique Wilson

Results – Dash

- Dash showing the change in results when the payload slider is activated



PREDICTIVE ANALYSIS

Dominique Wilson

Predictive Analysis

The main purpose of this study is to be able to predict the landing success of the first stage Falcon 9, therefore using predictive analysis is key.

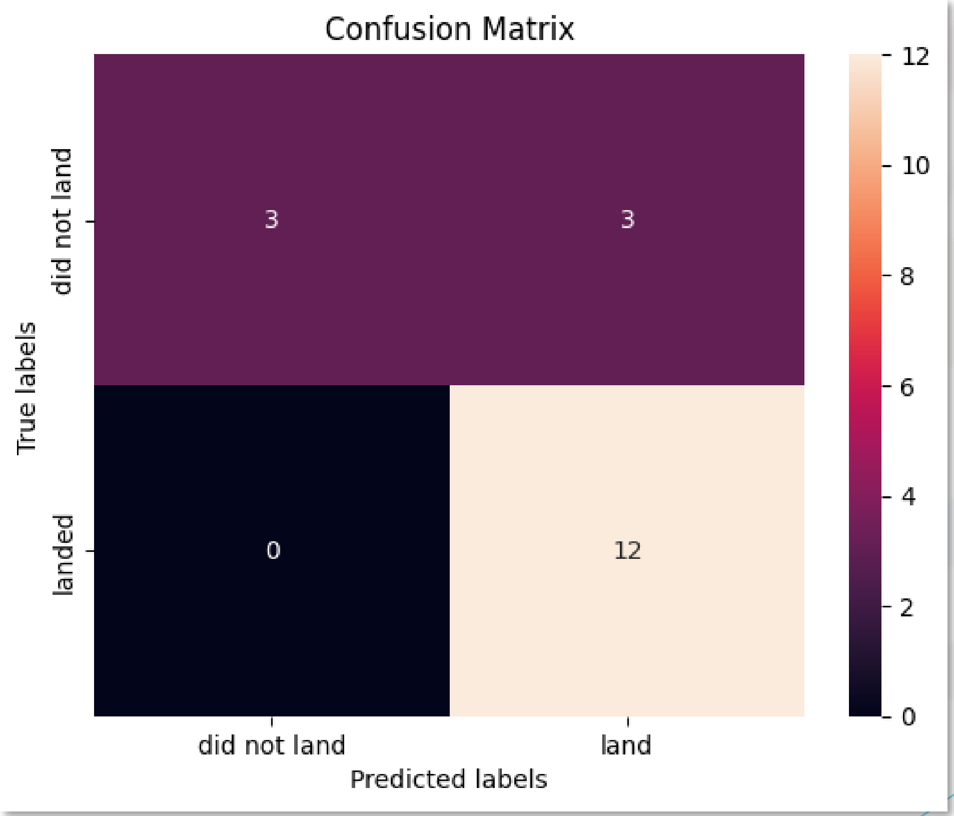
4 models were used:

- Logistic regression
- Support vector machine (SVM)
- Decision tree
- K-nearest neighbors (KNN)

Predictive Analysis

Logistic Regression

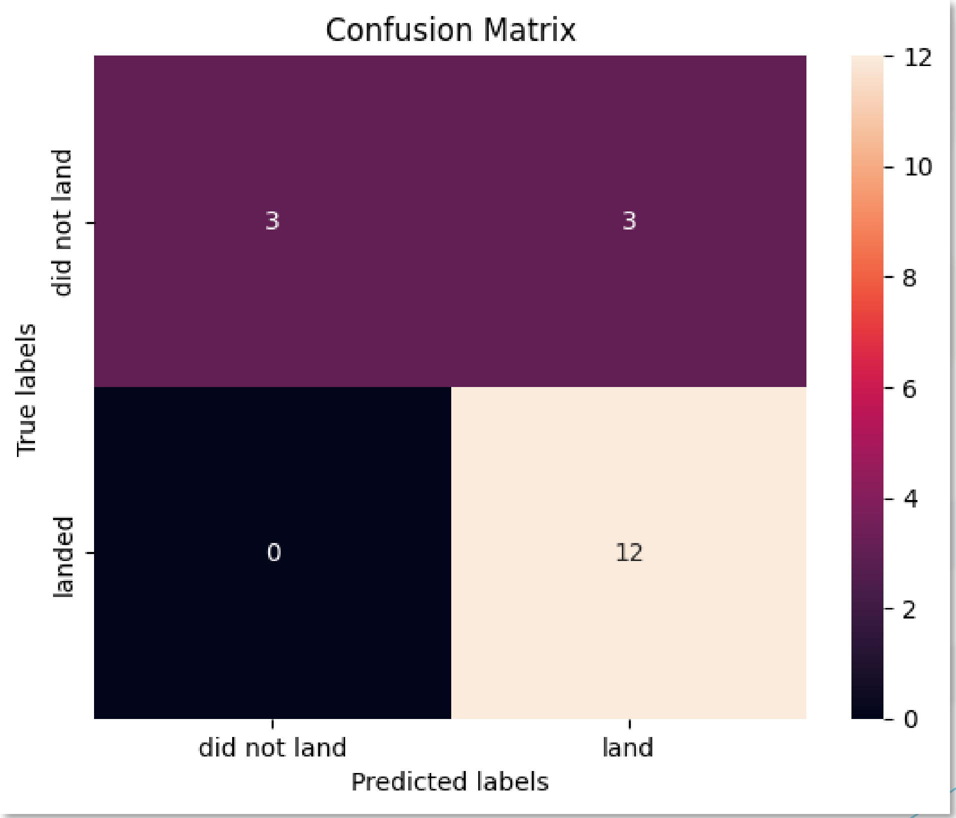
- GridSearchCV best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Predictive Analysis

Support Vector Machine (SVM)

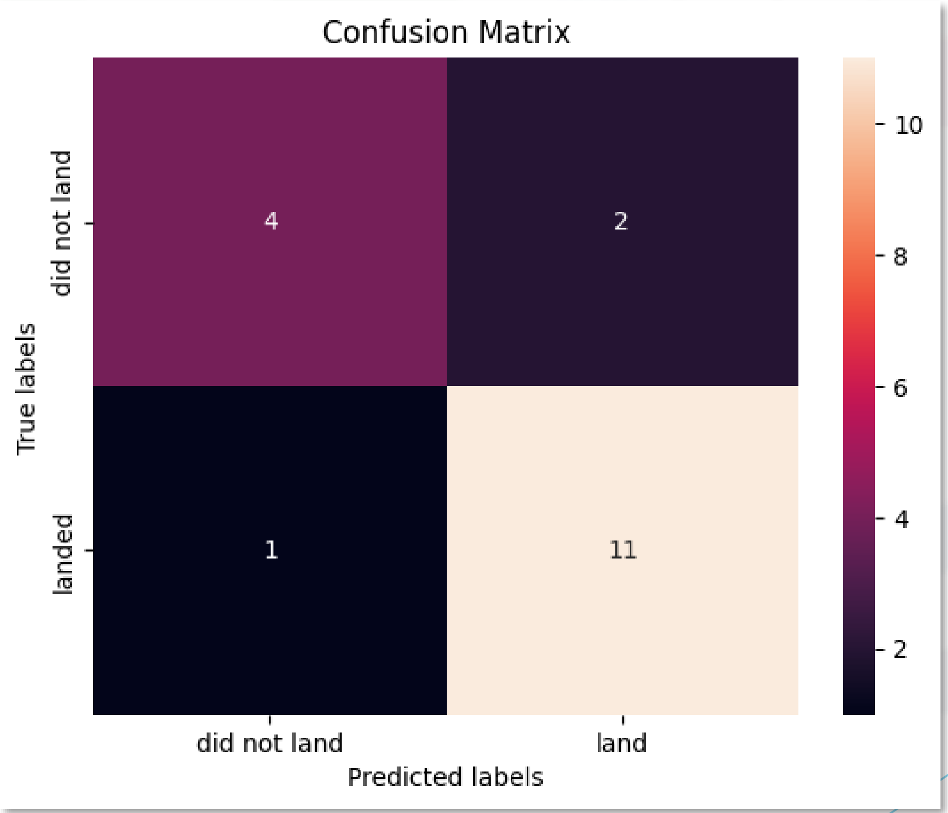
- GridSearchCV best score: 0.8482142857142856
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Predictive Analysis

Decision Tree

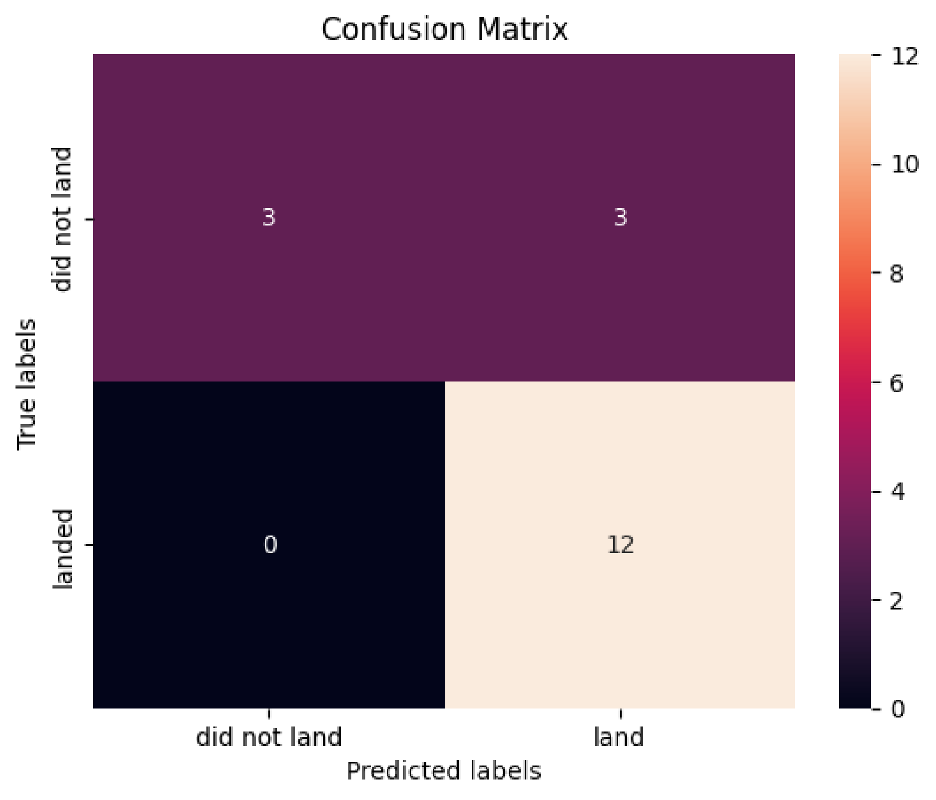
- GridSearchCV best score: 0.8714285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



Predictive Analysis

KNN

- GridSearchCV best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



DISCUSSION

Dominique Wilson

IBM Developer

SKILLS NETWORK

Discussion

It is clear from the data that there are many factors that influence the outcome of a first stage landing. For example it appears that the larger the flight number at a launch site the greater the success rate of the first stage landing. However, when this is then compared against orbit type the results differ slightly again.

Though there may be some clear relationships amongst the data it isn't clear how several influencing factors will impact the landing rate without the use of predictive modelling. The best model is decision tree with a classification accuracy of 0.87 so will be the best model in predicting first stage landing outcomes

CONCLUSION

Dominique Wilson

Conclusion

- The aim of this project was to predict the landing success of the first stage of a Falcon 9 launch in order to determine the cost of a launch
  - There are many factors that appear to impact the mission outcome including orbit type, payload mass, flight number and even launch site
  - 4 machine learning algorithms were deployed to learn from the Falcon 9 launch data available to us with the aim of producing a predictive model for future launches
  - The decision tree algorithm produced the predictive model with the best results

APPENDIX

Dominique Wilson

GitHub Notebooks

- All notebooks that provided date for this report can be found following [this link]