Applied Data Science Capstone Final Project

Jorge E. Paz C.

Coursera–IBM

13/05/2025

Content

Exectutive Summary

In this project, SpaceX rocket launch data was analyzed and visualized across multiple U.S. launch sites. Using Python and the Folium library, an interactive map was developed with clustered markers indicating the location and outcome (success or failure) of each launch. This visualization helped identify spatial patterns, showing that most launches are concentrated in a few strategic locations. The tool provides an intuitive way to explore SpaceX's historical performance by site, offering value for operational planning and decision-making.

Introduction

SpaceX has conducted numerous rocket launches across different sites in the United States. Understanding where and why launches succeed or fail can inform strategic decisions in space operations. This project aims to explore the spatial distribution of launch outcomes using geospatial tools in Python.

Data Collection and Wrangling Methodology

We used both the SpaceX API and web scraping from Wikipedia to gather the data. Then we performed basic cleaning, handled missing values, and shaped the dataset for the classification models we applied later.

Data was collected using two main methods:

- SpaceX REST API
- Web scraping from Wikipedia

After collection, we cleaned and prepared the data by:

- Handling missing values
- Structuring the dataset for classification tasks

EDA and Interactive Visual Analytics

The following charts were used to explore relationships in the data:

- Scatter plots:
  - Flight Number vs. Payload Mass
  - Payload Mass vs. Launch Site
  - Orbit Type vs. Success Rate
- Bar charts:
  - Launch Site vs. Success Count
  - Orbit Type vs. Success Rate
- Line charts:
  - Yearly trend of launch success

Visualizations helped identify trends and correlations that informed feature selection for modeling.

Predictive Analysis Methodology

We first standardized the features and split the data. Then, using 10-fold GridSearchCV, we tuned parameters for four models: Logistic Regression, SVM, Decision Tree, and KNN. We evaluated each one using accuracy, confusion matrix, and F1 and Jaccard scores to identify the best performer.

Predictive Analysis Methodology

The predictive modeling process included the following steps:

- Converted the target column (Class) to a NumPy array
- Standardized the features using `StandardScaler`
- Split the dataset into training and testing sets
- Applied `GridSearchCV` (10-fold) to optimize hyperparameters for:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- Evaluated models using:
  - Accuracy score
  - Confusion matrix
  - Jaccard score and F1 score

EDA with visualization results

Exploratory Data Analysis

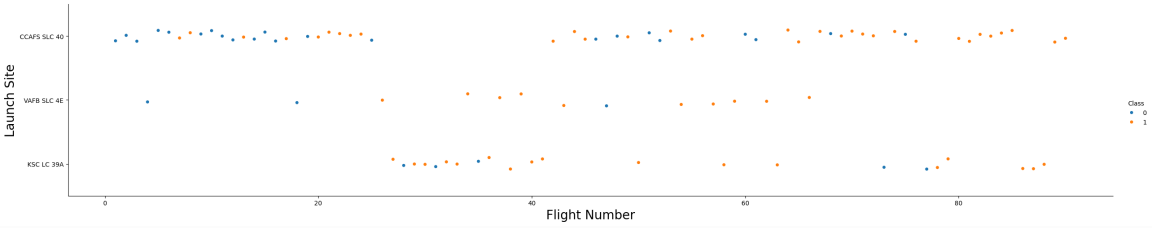


Figura: Flight Number vs Launch Site.

Launch Site vs Flight Success

The scatter plot shows the distribution of SpaceX flight outcomes across different launch sites over time. Most launches occurred at CCAFS SLC 40, with a mix of successes and failures, though success rates seem to improve over time. VAFB SLC 4E had fewer launches but mostly successful ones. KSC LC 39A appears in later flights, also with a majority of successes. Overall, no strong correlation is visible, but trends suggest operational improvements and varying performance by site.

Payload Mass vs Launch Site

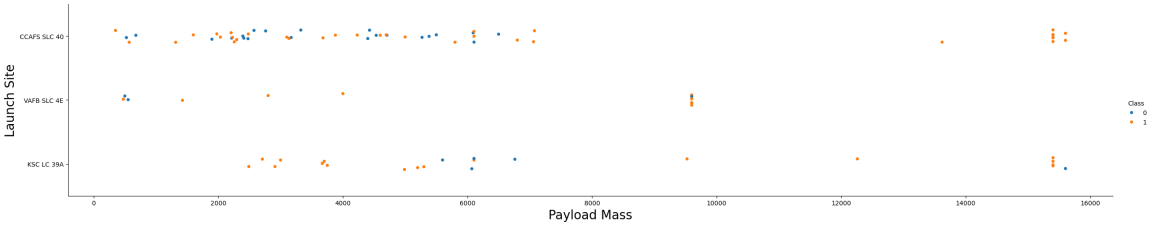


Figura: Payload Mass vs Launch Site

Payload Mass vs Launch Site

1. Most launches occurred at CCAFS SLC 40, followed by KSC LC 39A.
2. Higher payloads (above 10,000 kg) are mainly launched from KSC LC 39A and CCAFS SLC 40.
3. The success rate (Class $= 1$) appears higher at all launch sites across a wide range of payload masses.
4. VAFB SLC 4E handles a narrower range of payload masses, mostly under 10,000 kg.
5. There is no strong visible correlation between payload mass and success.

Orbit Type vs Success Rate

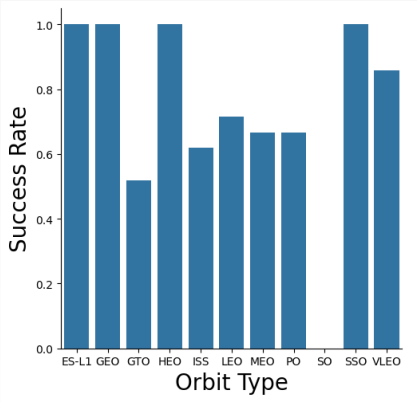


Figura: Orbit Type vs Success Rate

Orbit Type vs Success Rate

We can see that the ES-L1, GEO and SSO has the highest success rates because all landings related was successful.

Another Scatter plots

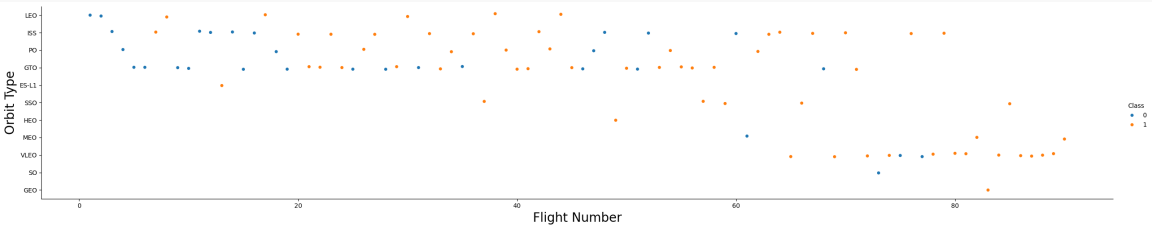


Figura: the relationship between FlightNumber and Orbit type

You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Antoher scatter plots

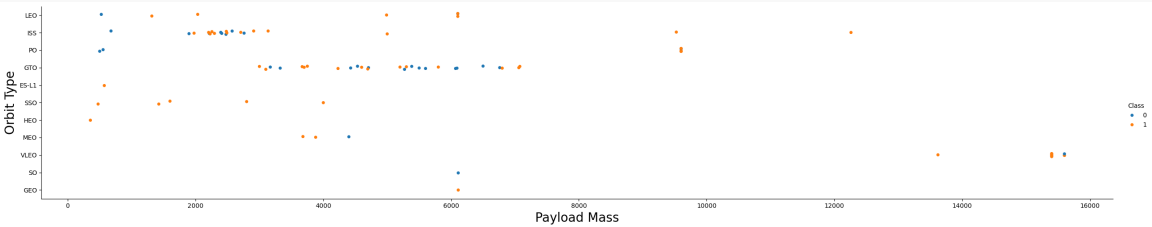


Figura: the relationship between Payload Mass and Orbit type

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present

Launch success yearly trend

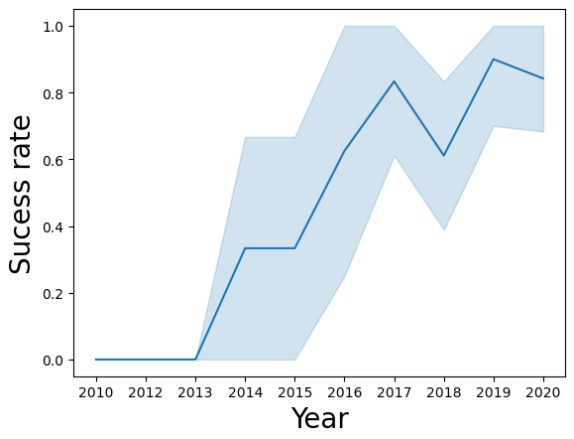


Figura: We can observe that the sucess rate since 2013 kept increasing till 2020

Some results of EDA with SQL results

Display the names of the unique launch sites in the space mission

In [10]:

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL;
```

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

Out[10]:

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

Figura: Unique launch sites in the space mission

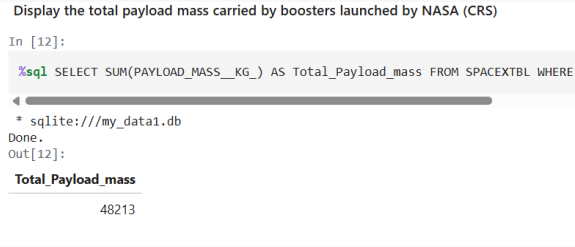
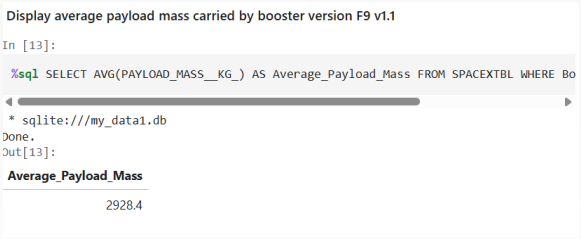


Figura: The total payload mass carried by boosters launched by NASA (CRS)


Display average payload mass carried by booster version F9 v1.1

In [13]:

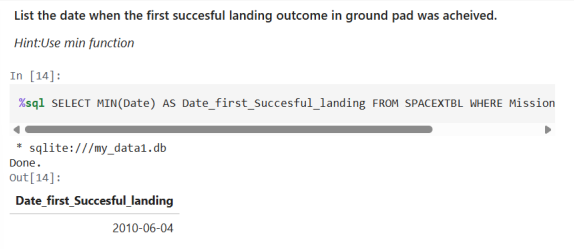
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTBL WHERE Bo
```

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

Out[13]:

Average_Payload_Mass
2928.4

Figura: The average payload mass carried by booster version F9 v1.1


List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
In [14]:
%sql SELECT MIN(Date) AS Date_first_Succesful_landing FROM SPACEXTBL WHERE Mission
 * sqlite:///my_data1.db
Done.
Out[14]:
```

Date_first_Succesful_landing
2010-06-04

Figura: The date when the first succesful landing outcome in ground pad was acheived.

Rank the 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.

```
In [21]:
%sql SELECT landing_outcome, count(*) as count_outcomes from SPACEXTBL WHERE date
 * sqlite:///my_data1.db
Done.
Out[21]:
```

Landing_Outcome	count_outcomes
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

Figura: Ranking the landing outcomes

Interactive map with Folium results

Map with the number of launchings

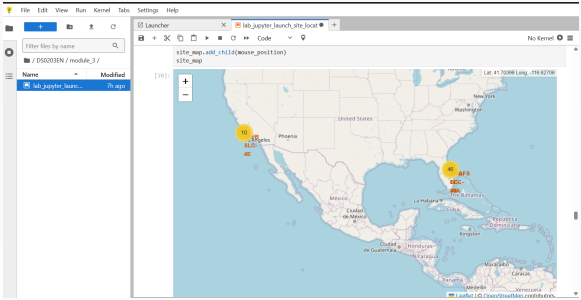


Figura: Number of launchings

Sites where the launches were successful

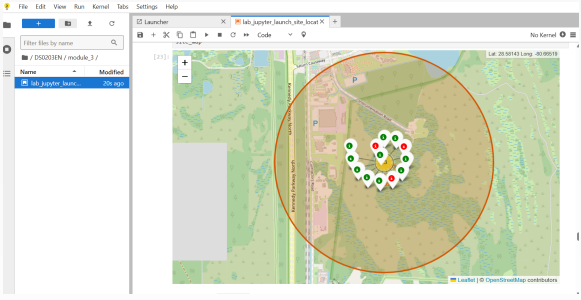


Figura: From the colour-labeled markers we must be able to identify which launch sites have more successful launches rates.

Plotly Dash dashboard results

Dash app results

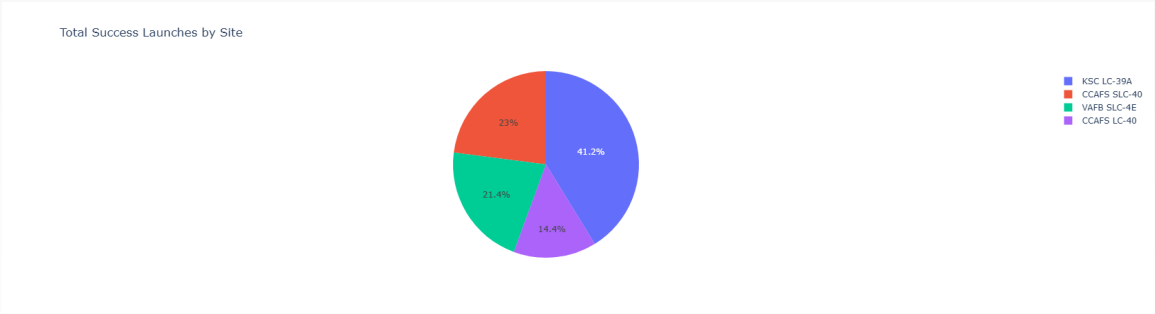


Figura: App created with dash that represents KSC LC39A has the most successful launches in general.

Dash results

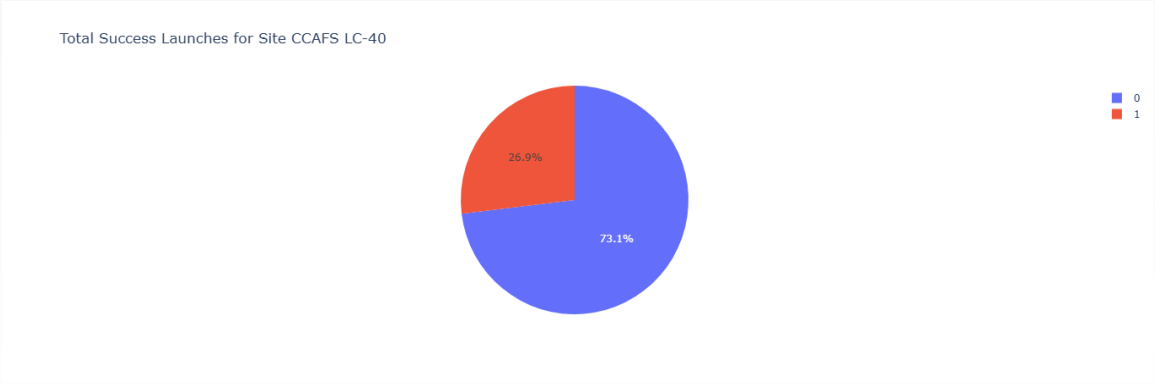


Figura: On the other hand we can see the low successful launches in CCAFS LC40

Dash results

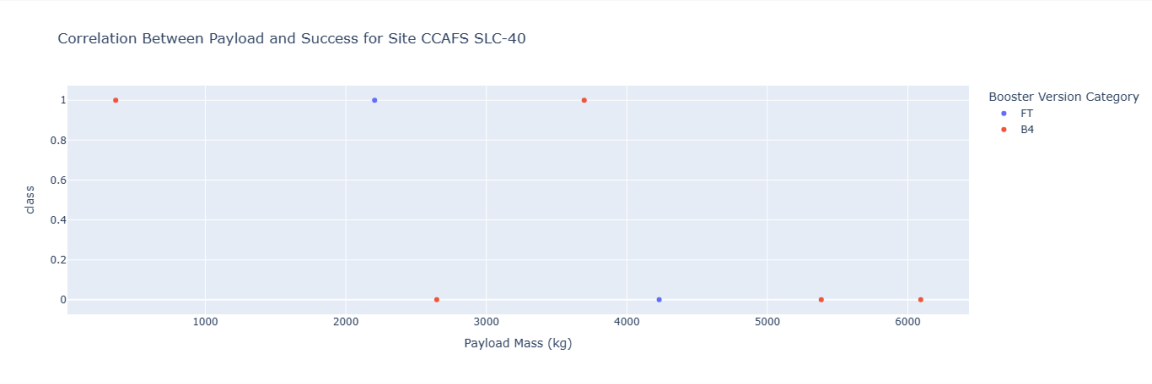


Figura: We can see the correlation between payload an sciccess for CCAFS SLC 40

Machine Learning Prediction

Objetive

If we can determine whether the first stage will land, we can determine the cost of a launch. A machine learning pipeline to predict if the first stage will land given the data on which we worked on it in past notebooks will be useful.

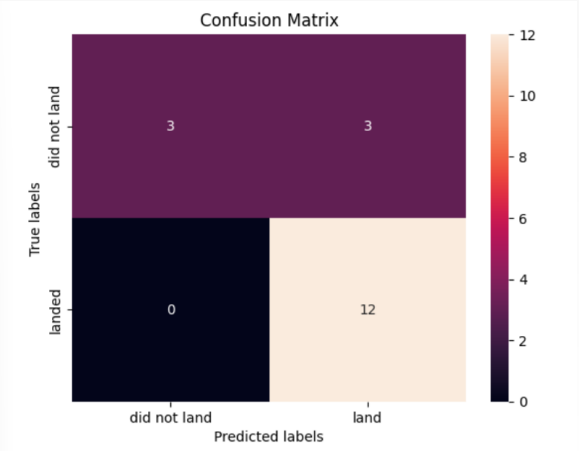
We first standardized the features and split the data. Then, using 10-fold GridSearchCV, we tuned parameters for four models: Logistic Regression, SVM, Decision Tree, and KNN. We evaluated each one using accuracy, confusion matrix, and F1 and Jaccard scores to identify the best performer.

Classification models

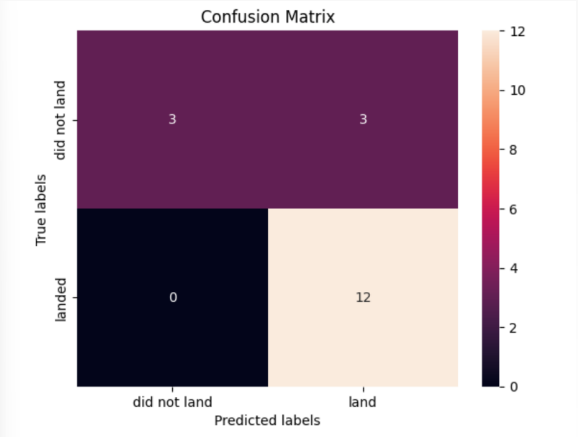
We used 4 different classification models for our purpose:

1. Logistic regression.
2. Support Vector Machine.
3. Tree classifier.
4. KNN.

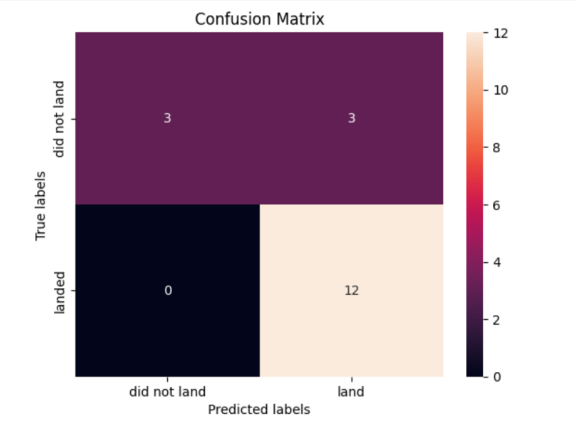
Confusion Matrix (logistic regression)



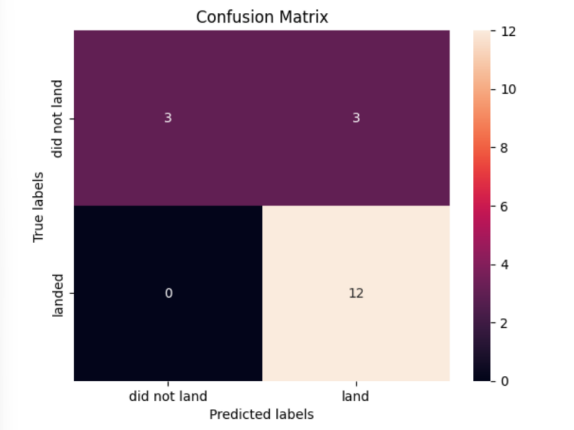
Confusion Matrix (SVM)



Confusion Matrix (Tree classifier)



Confusion Matrix (KNN)



As you can see, all the confusion matrix has the same results. All the classification methods have similar accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.840580	0.819444
F1_Score	0.909091	0.916031	0.913386	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556

Figura: Tree classifier would be our pick to do the classification based on the scores.

Conclusion

Our analysis revealed several key insights. Among the machine learning models tested, the Decision Tree classifier outperformed the others in predicting launch success. We also observed that launches with lower payload mass are more likely to succeed, suggesting a potential correlation between payload weight and mission complexity.

Geospatial analysis showed that most launch sites are located near the Equator and close to coastlines, likely to optimize launch trajectories and reduce risk. Success rates have improved over time, reflecting advancements in technology and operational experience. Notably, the KSC LC-39A site demonstrated the highest success rate, while orbits such as ES-L1, GEO, HEO, and SSO achieved a perfect 100 % success rate.