Winning The Space Race with Data Science

Rebeca Oyola

PRESENTATION

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

R.O

INTRODUCTION

Project background and context

SpaceX, a leader in the space industry, aims to make space travel more affordable and accessible. Its achievements include sending spacecraft to the International Space Station, deploying a satellite network that provides global internet access, and launching manned missions. One key factor behind its relatively low launch costs around $62 million per mission, is the innovative reuse of the Falcon 9 rocket's first stage. In contrast, other providers that cannot reuse this stage face costs exceeding $165 million per launch. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. To do this, we'll use publicly available data and machine learning techniques to forecast whether SpaceX or a competing company can reuse the rocket's first stage.

Through this Capstone project, we'll explore the full data science workflow: from data collection and cleaning to model training and evaluation. Using exploratory data analysis and classification algorithms, we'll build a predictive model that identifies the likelihood of stage reuse. This approach not only helps us understand the factors influencing successful landings but also demonstrates how data science can drive innovation in high-tech industries like aerospace.

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGY:

- **Data collection methodology:** data using SpaceX REST API web scraping techniques
- **Data Wrangling:** by filtering the data, handling missing values and encoding to prepare the data for analysis and modeling
- **Exploratory Data Analysis (EDA) with Visualization:** Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Exploratory Data Analysis (EDA) with SQL:** calculating the following statistics: total payload, payload range for successful launches, and total number of successful and failed outcomes
- **Interactive MAP with FOLIUM:** creating an interactive Maps with Folium
- **Interactive Dashboard with PLOTY :** Creating an interactive dashboard with Ploty
- **Predictive Analysis using classification models**: predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN).

SUMMARY OF RESULTS

- **Exploratory Data Analysis Results:**

Launch success has improved over time
KSC LC-39A has the highest success rate among landing sites
Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

- **Visualization/Analytics:**

Most launch sites are near the equator, and all are close to the coast
Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

- **Predictive Analytics:**

Decision Tree model is the best predictive model for the dataset

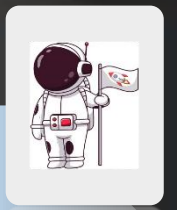
METHODOLOGY

Data Collection - API

For this Capstone project, we collected launch data using the SpaceX REST API, specifically the endpoint https://api.spacexdata.com/v4/launches/past. We used Python's requests library to send GET requests and received the data in JSON format, where each object represents a single launch. These JSON records include key details such as rocket type, payload mass, launch site, orbit, and landing outcome.

To make the data easier to analyze, we transformed the structured JSON into a flat table using json_normalize. Additional launch attribute, such as booster version or payload specifications, were retrieved from other API endpoints using unique identifiers. We filtered out Falcon 1 launches to focus exclusively on Falcon 9 missions. Missing values, especially in payload mass, were handled by replacing them with the mean, while nulls in the landing pad column were kept intact since they indicate no landing pad was used.

Data Collection – SpaceX API



Request datafrom SpaceX API (rocket launch data)

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Request informationabout the launches from SpaceX API using custom functions

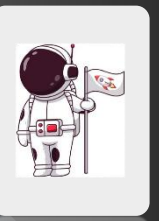
Create a Dictionary from data

Create dataframefrom the dictionary

Replace missing values

Export data to csv file

Data Collection – WEB Scraping



Request data(Falcon 9 launch data) from Wikipedia

Create Beautiful Soup object from HTML response

Extract column namesfrom HTML table header

Collect data from parsing HTML tables

Create dictionary and **Create a data Frame**

Export data to csv file

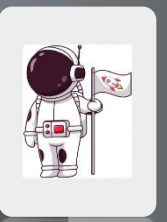
Data Wrangling

In this stage of the project, we focused on organizing and preparing the raw launch data for analysis. The dataset includes key attributes such as flight number, launch date, booster version, payload mass, orbit type, launch site, and landing outcome. We explored orbital classifications like LEO (Low Earth Orbit and GTO (Geostationary Transfer Orbit), which influence mission profiles.

The landing outcome column was especially important it indicates whether the rocket's first stage landed successfully. We transformed this into a binary classification target: 1 for successful landings and 0 for failures. This new variable, Y, will be used to train our machine learning model.

By cleaning, filtering, and converting these attributes into structured formats, we ensured the dataset was ready for predictive modeling and deeper insights into SpaceX's launch performance.

Data Wrangling



Export data to csv file

Create binary landing outcome column (dependent variable)

Calculate: Number & occurrence of Mission outcomes per orbit type.

Calculate: Number & **Occurrence of each orbit**

Calculate: Number of launches for each site

R.O

EDA with SQL

Queries

Display:

- ✓ Names of unique launch sites
- ✓ 5 records where launch site begins with ‘CCA’
- ✓ Total payload mass carried by boosters launched by NASA (CRS)
- ✓ Average payload mass carried by booster version F9 v1.1.

List:

- ✓ Date of first successful landing on ground pad
- ✓ Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- ✓ Total number of successful and failed missions
- ✓ Names of booster versions which have carried the max payload
- ✓ Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- ✓ Count of landing outcomes between 2010

2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

AVG("PAYLOAD_MASS__KG_")
2928.4

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

R.O

EDA with Data Visualization

Charts were plotted:

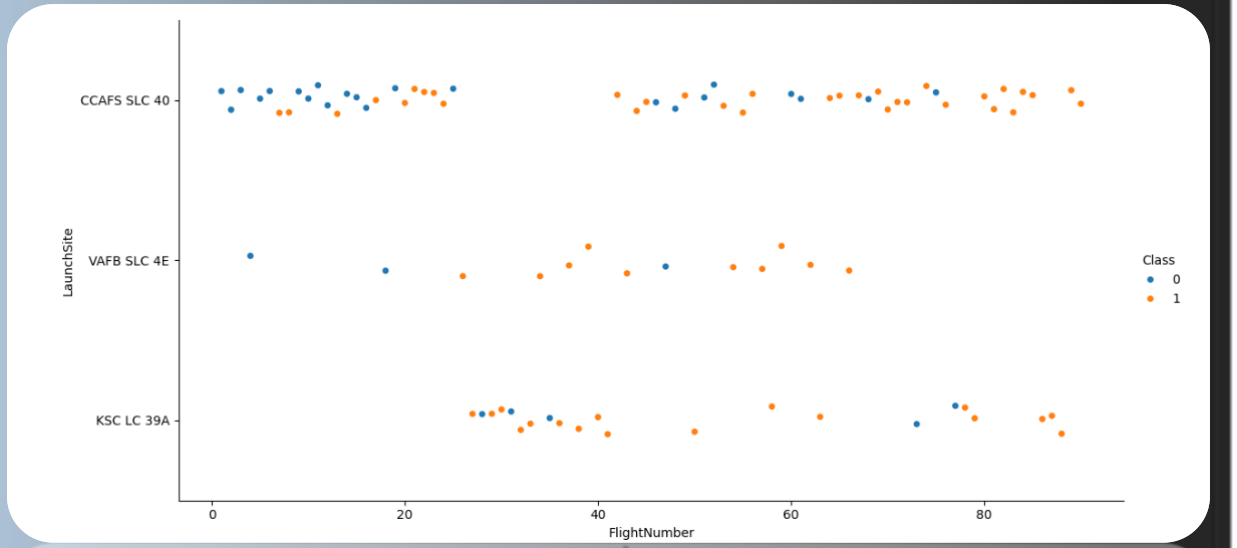
- ✓ Flight Number vs. Payload Mass,
- ✓ Flight Number vs. Launch Site,
- ✓ Flight Number vs. Orbit Type,
- ✓ Payload Mass (kg) vs. Launch Site
- ✓ Payload Mass (kg) vs. Orbit type
- ✓ Success rate vs Orbit Type
- ✓ Success Rate Yearly Trend

Scatter plots: Show the relationship between variables.
Bar charts: show comparisons among discrete categories.
Line charts: show trends in data over time (time series).

R.O

Flight Number vs. Launch Site

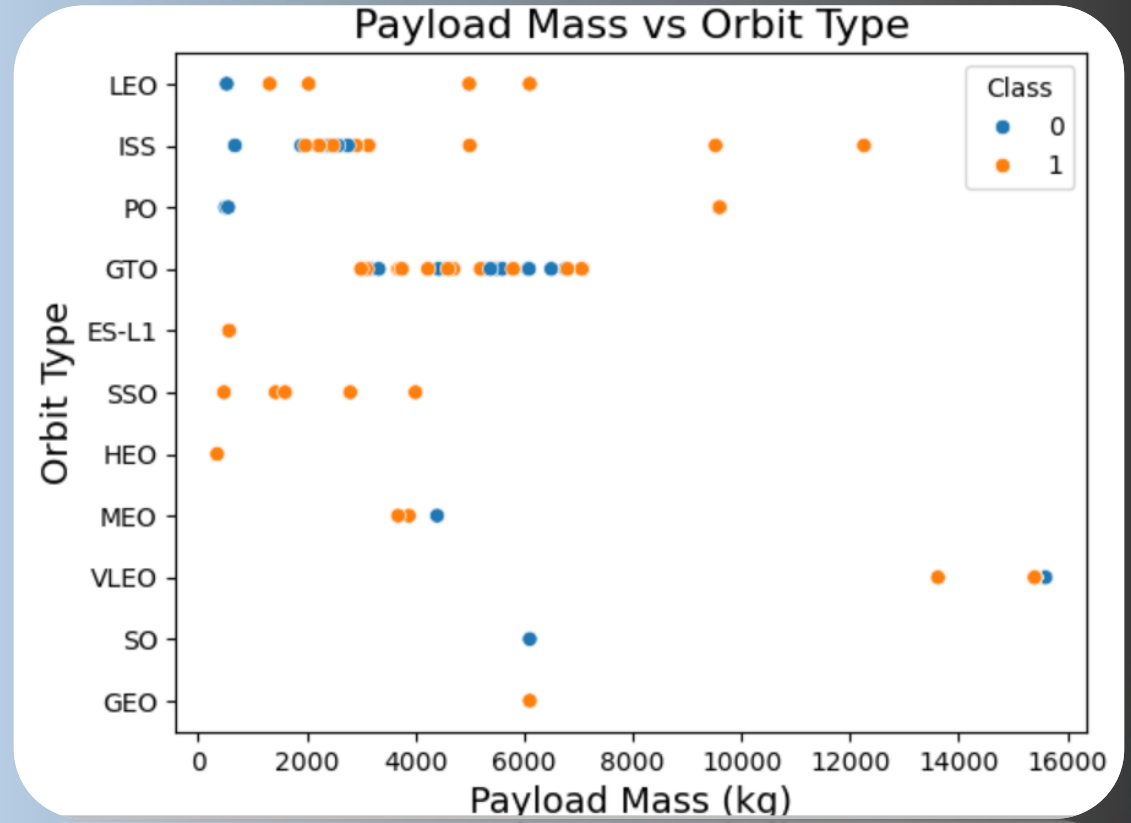
- Exploratory Data Analysis
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

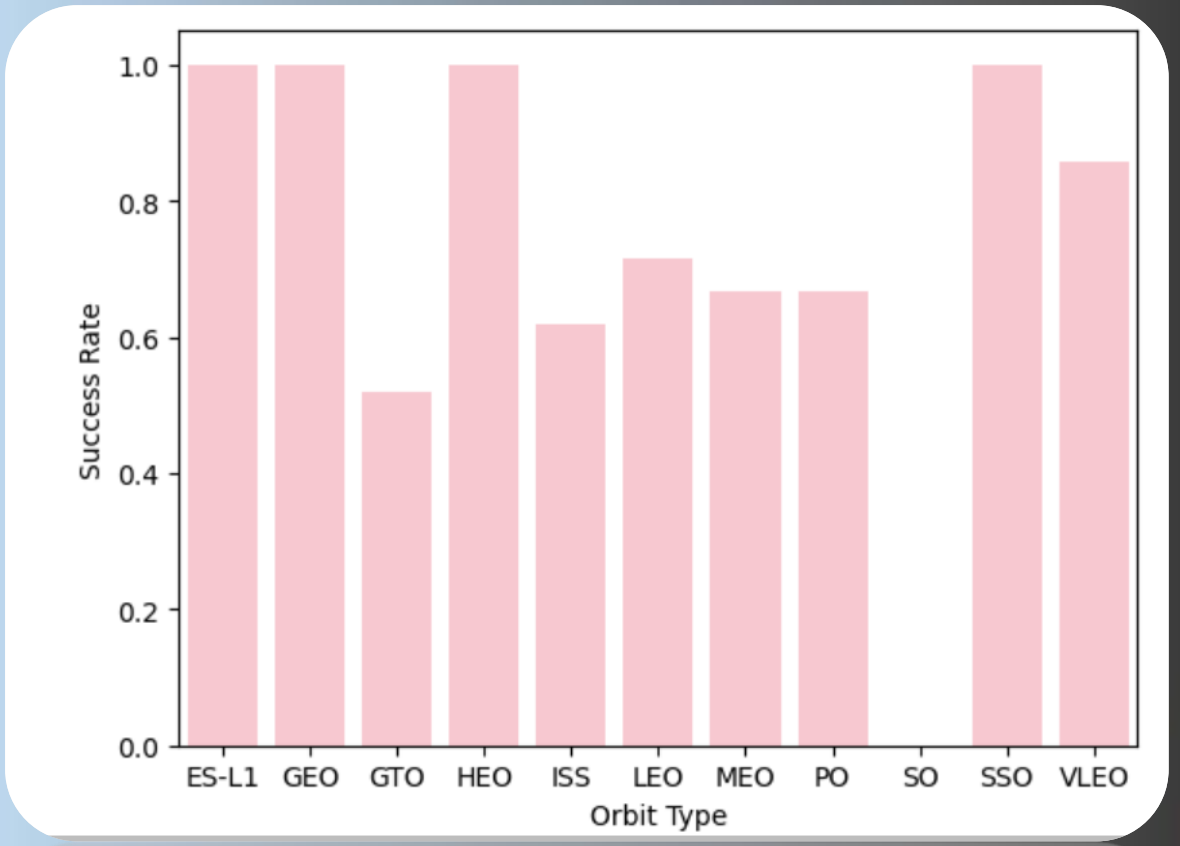
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

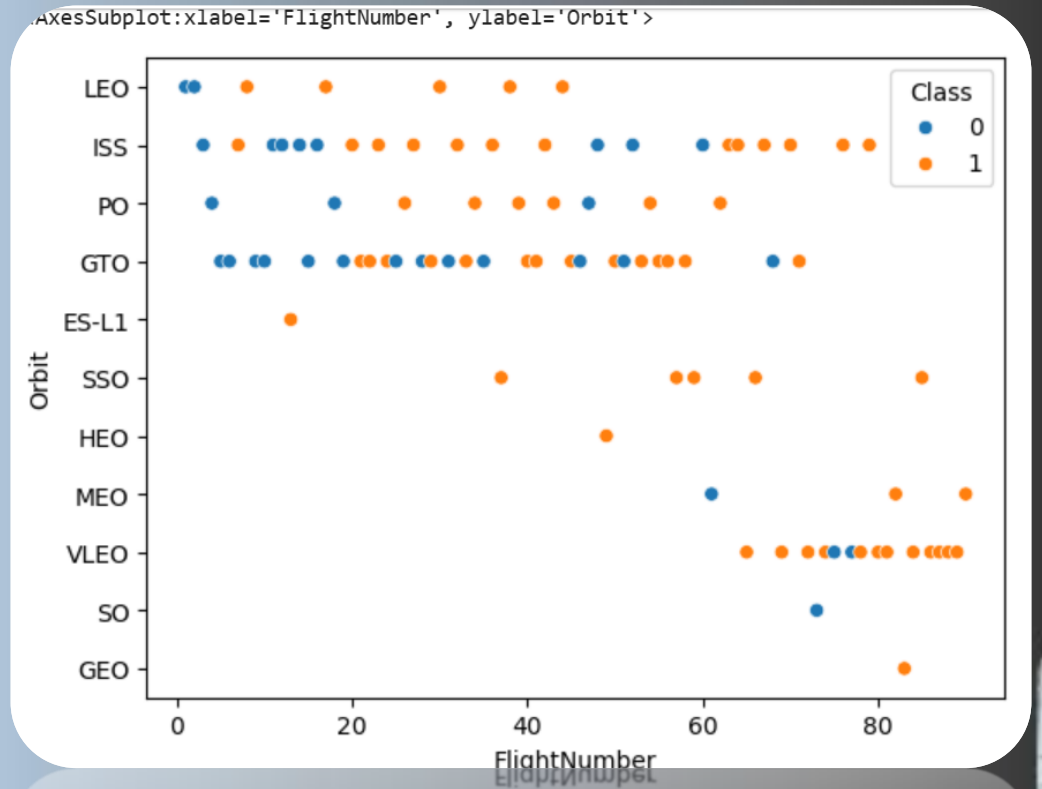
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

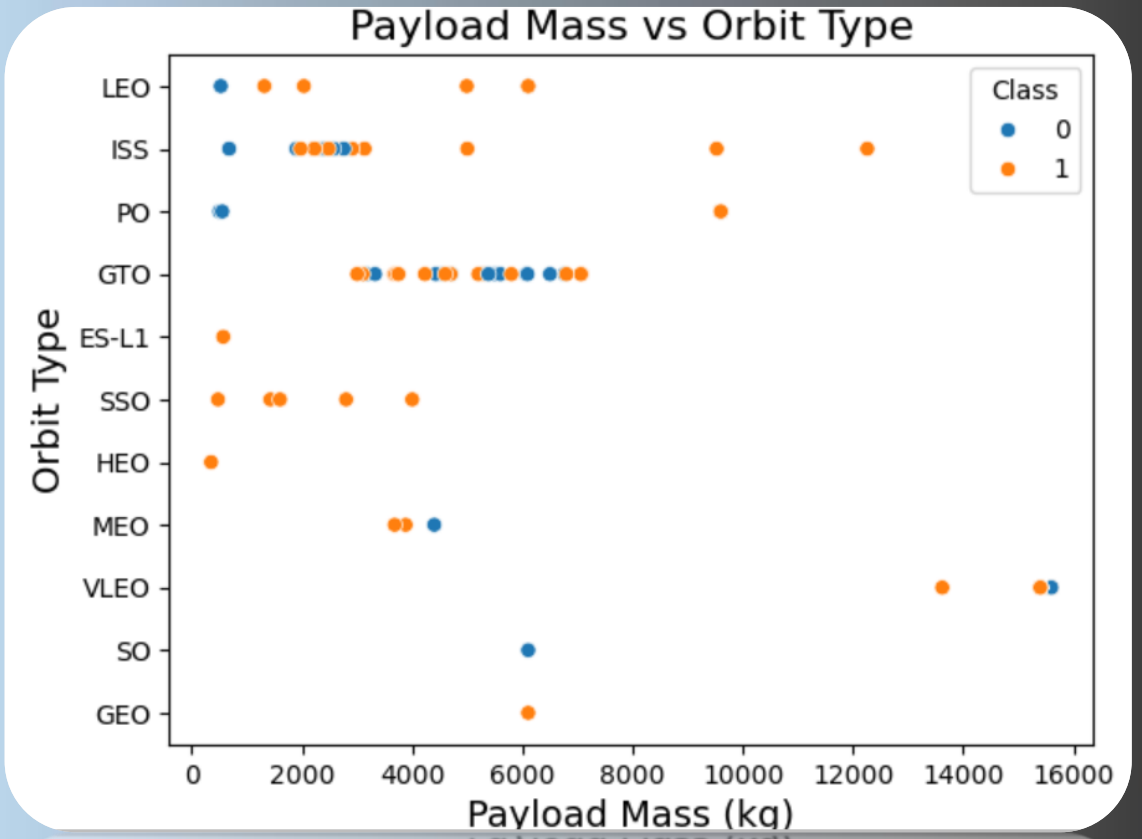
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

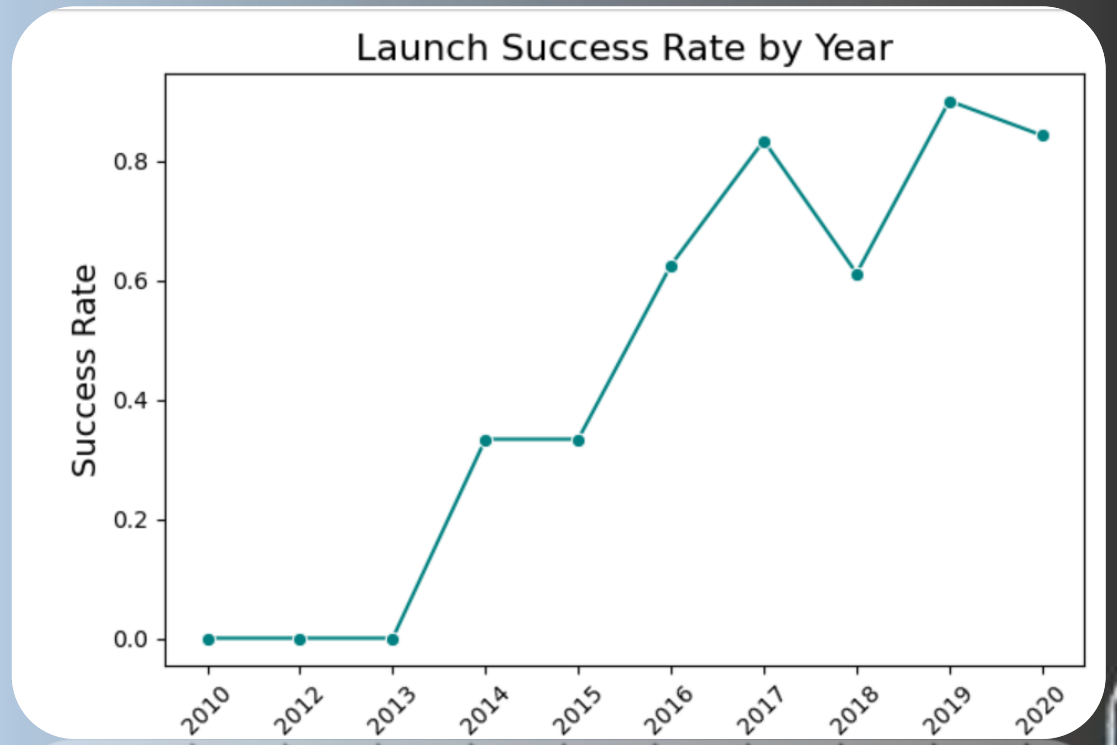
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

Launch Site Names

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

Landing Outcome Cont.

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

RESULTS

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

Launch Site Names Begin with 'CCA'

Records with Launch Site Starting with CCA

- Displaying 5 records below

2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Total Payload Mass

Payload Mass

- **Total Payload Mass**
- **45,596 kg** (total) carried by boosters launched by NASA (CRS)

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "CUSTOMER" = "NASA (CRS)"
```

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

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

Average Payload Mass

2,928 kg (average) carried by booster version F9 v1.1

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "BOOSTER_VERSION" LIKE "%F9 V1.1"
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

[17]:
```
%sql SELECT MIN("DATE") FROM SPACEXTABLE WHERE "MISSION_OUTCOME" LIKE "%SUCCESS"
```

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

[17]:

MIN("DATE")
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)'A
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT (SELECT COUNT(*) FROM SPACEXTBL WHERE "Mission_Outcome" LIKE '%Success%') AS SUCCESS, (SELECT
```

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

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

Carrying Max Payload

- 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

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

R.O

2015 Launch Records

Showing month, date, booster version, launch site and landing outcome

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

R.O

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

Ranked Descending

- Count of landing outcomes between 2010-06-04 and 2017-03- in descending order

Landing_Outcome	Outcome_Count
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

Build an Interactive Map with Folium

R.O

Build an Interactive Map with Folium

Colored Markers of Launch Outcomes

- Added colored markers of successful(green) and unsuccessful(red) launches at each launch site to show which launch sites have high success rates
- Distances Between a Launch Site to Proximities
- Added colored lines to show the distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

Folium Map Screenshot 1

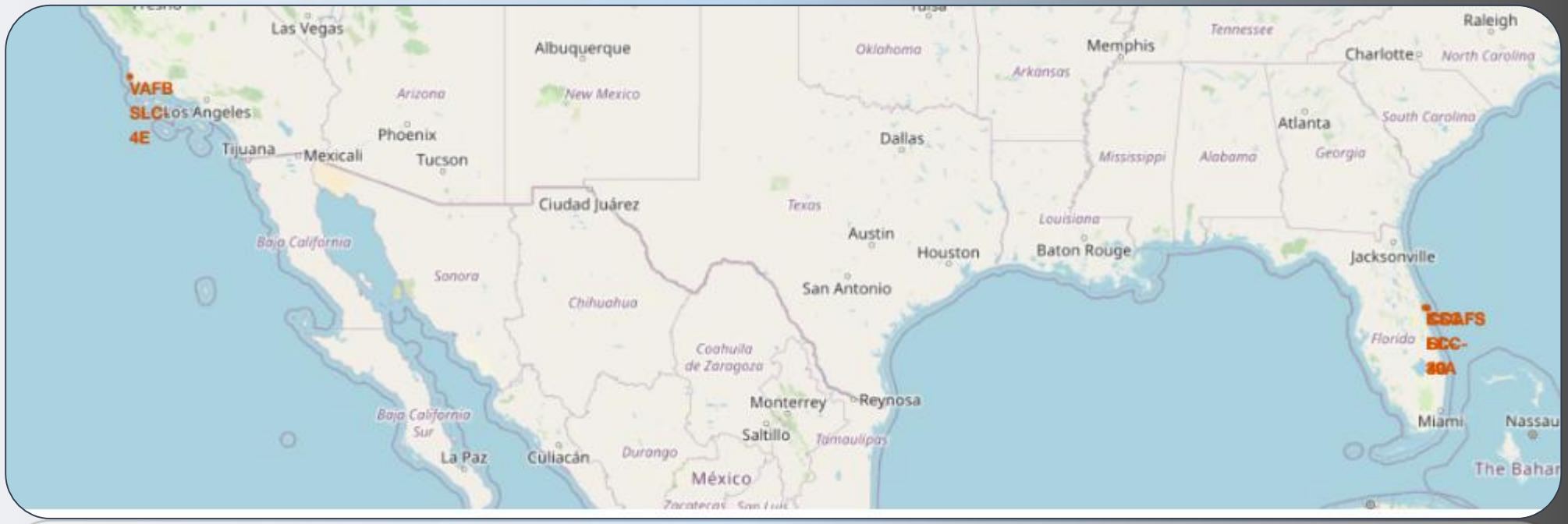
This map displays the location of SpaceX launch sites

Note that all launch sites are in proximity to the equator line

Note that all launch sites are very close to the coast

Rockets launched oven the ocean reduce the risk

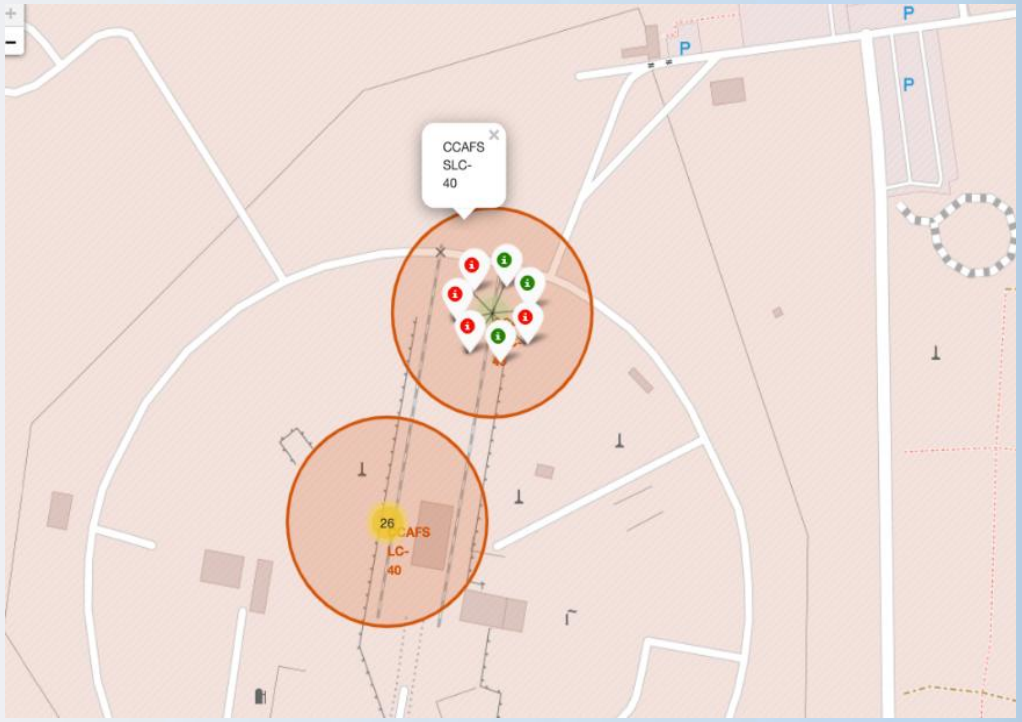
Of dropping or exploding near to the people.

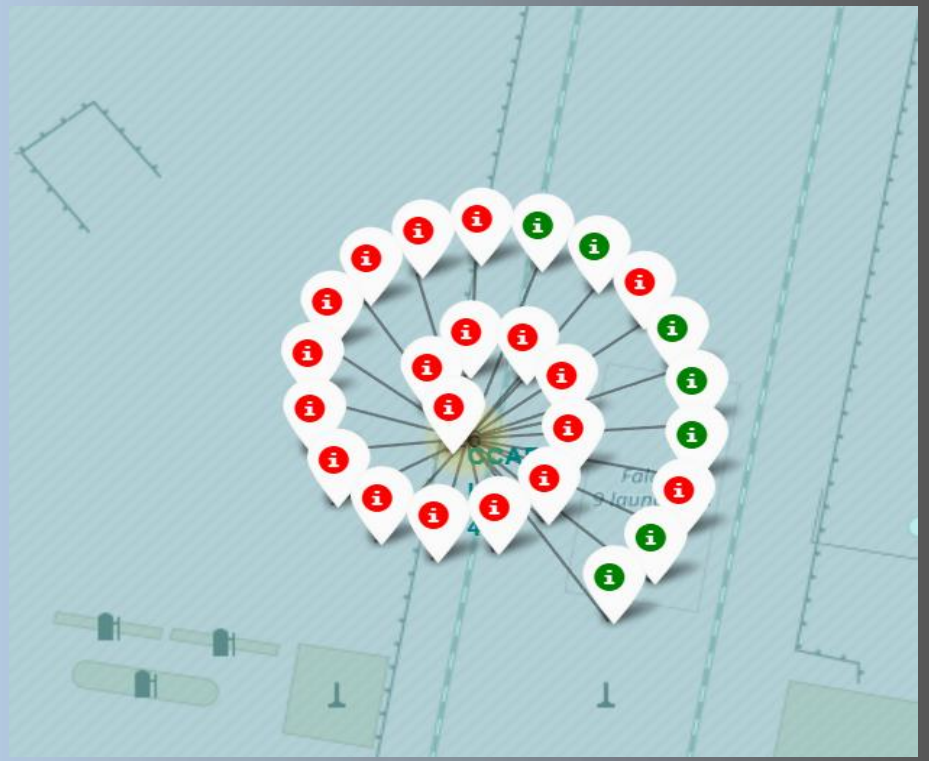


Folium Map Screenshot 2

Outcomes:

- **Green** markers for successful launches
- **Red** markers for unsuccessful launches





Build a Dashboard with Plotly Dash

- Slider of Payload Mass Range
- Pie Chart Showing Successful Launches
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version


Clasification

Predictive Analysis (Classification)

Charts

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** Grid SearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

Classification Accuracy

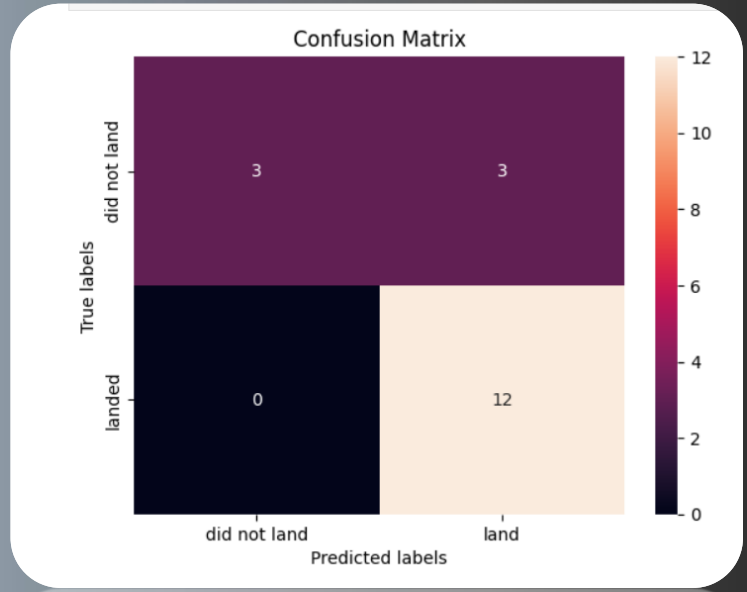
- All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset.
- The Decision Tree modelslightly outperformed the restwhenlookingat.best_score_
- best_score_ is the average of all cv folds for a single combination of the parameters

	Model	Test Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
3	KNN	0.833333
2	Decision Tree	0.777778

Confusion Matrix

Performance Summary

A confusion matrix summarizes the performance of a classification algorithm
All the confusion matrices were identical
The fact that there are false positives (Type 1 error) is not good
Confusion Matrix Outputs:12 True positive
3 True negative
3 False positive
0 False Negative




CONCLUSION

Valuable insights:

•**Model Performance:** The classification models demonstrated comparable results on the test set, with the decision tree model showing a slight edge. However, future work should consider implementing more advanced algorithms such as XGBoost, which may offer improved accuracy and robustness.
•**Geographic Location:** Most launch sites are situated near the equator, leveraging Earth's rotational velocity to reduce fuel requirements and operational costs. Additionally, all sites are located near coastlines, which facilitates logistics and enhances safety protocols.
•**Launch Success Trends:** There is a clear upward trend in launch success over time. Notably, **KSC LC-39A** stands out with a 100% success rate for missions carrying payloads under 5,500 kg.
Orbits and Payload Mass: Orbits such as ES-L1, GEO, HEO, and SSO have achieved perfect success rates. Moreover, a positive correlation was observed between payload mass and launch success, suggesting that heavier missions are often better planned and executed.

Recommendations for Future Research

Dataset Expansion: Incorporating a larger dataset would enhance the generalizability of the findings and strengthen the predictive capabilities of the models.
Feature Analysis: Further exploration of feature importance, including techniques like Principal Component Analysis (PCA), could uncover hidden patterns and improve model performance.
Advanced Model Evaluation: Testing models such as XGBoost could provide a more comprehensive comparison and potentially outperform the current classifiers.

R.O