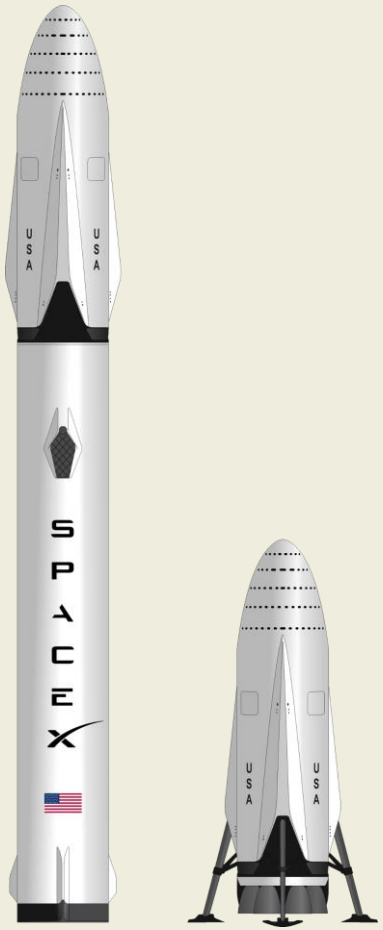
Skills Network

SpaceY: Winning Space Race with Data Science

Final Project Presentation

Capstone Project



16th of November 2023

Presentation done by: João André Pinho

Outline

A. Executive Summary

Project Summary

Through Data Analysis and Predictive Modelling, SpaceY can better estimate which Rocket Landings will be successful.

Executive Summary | Project Summary

Methodology Overview:

- **Data Collection** via SpaceX REST API and web scraping methodologies;
- **Data Wrangling** to establish a success/failure outcome variable;
- **Data Exploration through Visualization**, analysing payload, launch site, flight number, and yearly trends;
- **Data Exploration using SQL** to derive insights such as total payload, payload range for successful launches, and aggregate successful and failed outcomes;
- **Investigation** of launch site success rates and geographical proximities;
- **Development of predictive models** for landing outcome forecasts.

Results:

- The **classification models** exhibited **comparable performance**, leaving room for **refinement** specifically in **reducing Type I errors**.
- **Over time**, there has been a **notable increase in launch success**.
- There is a **positive correlation** between **higher payload** masses and **greater success across all launch sites.**
- **KSC LC-39A** is the **highest success rate** launch site and the best orbits are **ES-L1, GEO, HEO, and SSO.**

B. Introduction

Issue Tree

The competitiveness level and cost effectiveness needed in the space exploration market highlights the necessity for Space Y to better predict its landing outcomes.

Introduction | Issue Tree

Situation

SpaceY, an industry leader in space exploration, has consistently aimed to democratize space travel by making it more cost-effective. Its achievements include missions to the International Space Station, deploying an internet satellite constellation, and conducting manned missions.

Complication

As SpaceY's cost-efficiency stems from its innovative reuse of the Falcon 9 rocket's first stage, a procedure that involves a lot of precision and control, it is of extreme importance that the success of this stage can be properly predicted.

Question

Should SpaceY use Data Science Predictive Modelling techniques to study the likelihood of success of first-stage landings?

Answer

Yes.

Why?

External

New players are joining the space exploration market

Maintain competitive advantage

Internal

For each reusable, the company saves $103 million.

Better estimation of launch costs

SpaceY should use Predictive Modelling

How?

- **Data Collection**
- **Data Wrangling**
- **Data Exploration**
- **Data Visualization**
- **Predictive Analysis**

B. Project Methodology

Scope and Workstreams

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

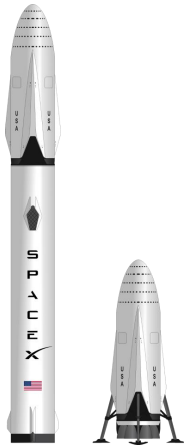
Project Methodology | Scope & Workstreams

STAGES	Data Collection	Data Wrangling	Data Exploration	Data Visualization	Predictive Analysis
OVERVIEW	• Collect data using **SpaceX REST API** and through **Beautiful Soup web scrapping**.	• **Filter the data**, **handle missing values** and **convert categorical features** to prepare data for the exploration stage.	• **Explore data using SQL and Pandas,** as well as **Seaborn** for plotting**.**	• **Visualizing** the data summary using **Folium** for **interactive maps** and **Plotly Dash** for making a **dashboard**.	• **Exploration** of various modelling techniques; • Selecting a **specific model;** • **Testing** to evaluate the accuracy, robustness, and generalization capabilities of the model.
OUTPUTS	Original Dataset	Prediction Dataset	EDA Insights	Maps & Dashboards	Classification Model
STATUS	CONCLUDED	CONCLUDED	CONCLUDED	CONCLUDED	CONCLUDED

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Collection – SpaceX API

Data Collection



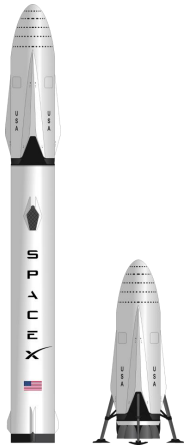
Actions:

1. **Retrieve rocket launch data** from the **SpaceX API** by initiating a data request.
2. **Decode the received response** utilizing the .json() function, **transforming it into a dataframe** through .json_normalize().
3. **Utilize python functions** to **acquire detailed launch information** from the SpaceX API.
4. **Construct a dictionary** based on the obtained data.
5. **Convert the dictionary into a dataframe**.
6. **Filter the dataframe** to exclusively include **Falcon 9 launches**.
7. **Fill in missing Payload Mass values** by computing the **mean** of available data.
8. **Save/export the processed data into a CSV file** for further use.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Collection – Web Scraping

Data Collection



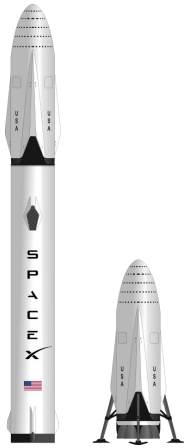
Actions:

1. **Initiate a data request** from Wikipedia to obtain Falcon 9 launch data.
2. **Generate a BeautifulSoup object** by **parsing the HTML response** received.
3. **Extract column names** by **parsing the HTML table header**.
4. **Gather relevant data** by **parsing HTML tables**.
5. **Compile the acquired data into a dictionary**.
6. **Transform the dictionary** into a **structured dataframe**.
7. **Save/export the processed data into a CSV file** for storage or further analysis.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Wrangling

Data Wrangling



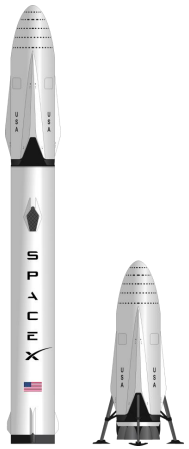
Actions:

1. **Conduct Exploratory Data Analysis (EDA)** to identify and comprehend data labels and patterns.
2. **Compute the following metrics**:
    1. Number of launches for each site.
    2. Quantity and frequency of orbits.
    3. Number and frequency of mission outcomes based on orbit types.
3. **Generate a binary landing outcome column** as the dependent variable.
4. **Save/export the processed data into a CSV file** for further analysis or reference.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Exploration – EDA with Visualization

Data Exploration



Charts:

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

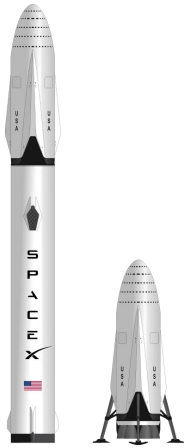
Objective:

Analyse the relationship between variables using different types of plots**,** to figure out **how they affect each other** and **which should be included in the machine learning model.**

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Exploration – EDA with SQL

Data Exploration



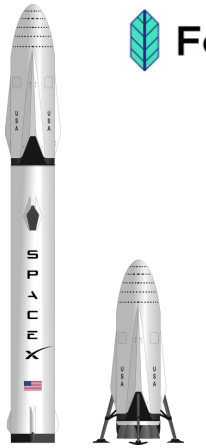
Queries:

- 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.;
- Date of the first successful landing outcome in ground pad;
- Names of the boosters which have successfully drone shipped with a payload between 4000 and 6000 Kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried maximum payload mass;
- Records of failing landings in the year of 2015;
- Ranking of landing outcomes between 04/06/2010 and 20/03/2017.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Visualization – Folium Interactive Maps

Data Visualization



Folium

Markers of Launch Sites:

- Included a blue circle marking the coordinates of NASA Johnson Space Center, featuring a popup label displaying its name derived from its latitude and longitude coordinates.
- Integrated red circles at all launch site coordinates, each with a popup label revealing its name, also based on its latitude and longitude coordinates.

Utilizing Folium for Mapping:

- Visualized launch outcomes with coloured markers distinguishing successful (green) and unsuccessful (red) launches at each site, highlighting sites with notably high success rates.

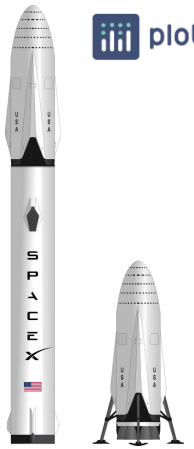
Distances from Launch Sites to Proximities:

- Incorporated coloured lines to demonstrate the distance from launch site CCAFS LC-40 to the nearest coastline, railway, highway, and city proximities.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Data Visualization – Plotly Dash Dashboard

Data Visualization



Launch Site Selection Dropdown:

- Enable users to choose either all available launch sites or specific individual sites from a dropdown menu.

Interactive Dashboard Using Plotly Dash:

- Implement a payload mass range slider, offering users the capability to select desired payload mass ranges.

Visual Representation with Pie Chart:

- Display a pie chart enabling users to observe the percentage representation of successful and unsuccessful launches concerning the total count.

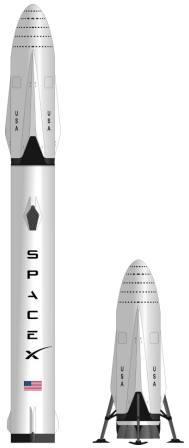
Correlation Visualization via Scatter Chart:

- Utilize a scatter chart to demonstrate the relationship between payload mass and launch success rate across different booster versions, allowing users to explore the correlation between these factors interactively.

The complementarity of the five workstreams will allow Space Y to better predict the success of its flight landings.

Project Methodology | Predictive Analysis

Predictive Analysis



Actions:

1. **Generate a NumPy array** based on the **Class column** data.
2. **Standardize the dataset using StandardScaler** by fitting and transforming the data.
3. **Divide the data into training and testing sets** using train_test_split.
4. **Instantiate a GridSearchCV object** with cv=10 to **optimize parameters**.
5. **Apply GridSearchCV to various algorithms**: Logistic regression, Support Vector Machine (SVM) Decision Tree Classifier, and K-Nearest Neighbor (KNN).
6. **Assess the accuracy of all models** on the test data using **.score().**
7. **Evaluate the confusion matrix** for each model to understand performance.
8. **Determine the best model based on Jaccard_Score, F1_Score, and Accuracy metrics**.

C. Results

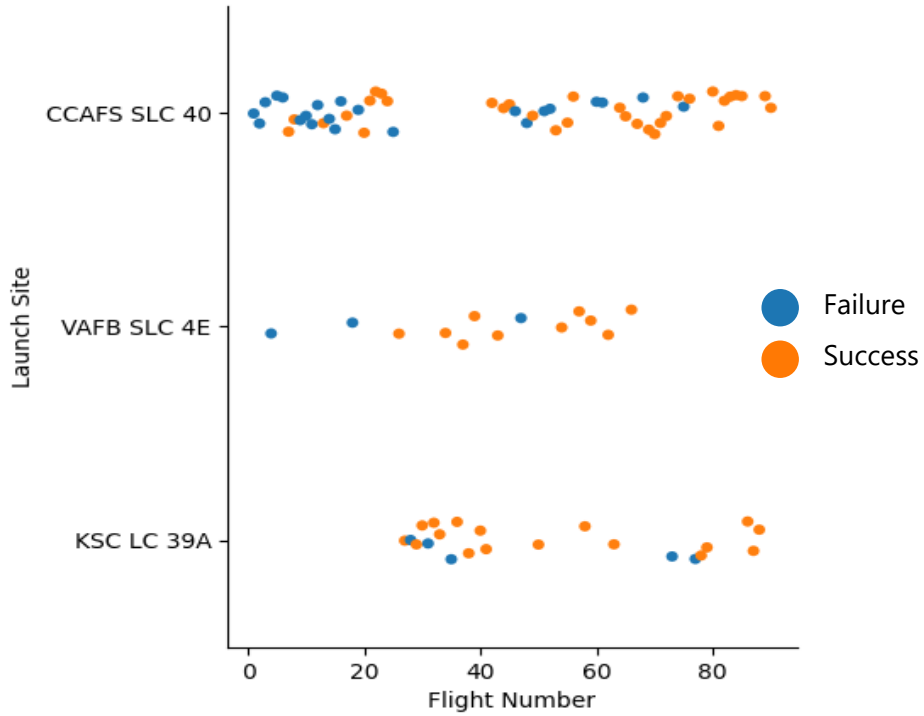
C1. EDA with Visualization

Insights Summary

Despite the success rate disparity across different launch sites, later flights have registered less failures across time.

Results | EDA with Visualization – Flight Number vs. Launch Site

Scatter Plot of Flight Number vs. Launch Site



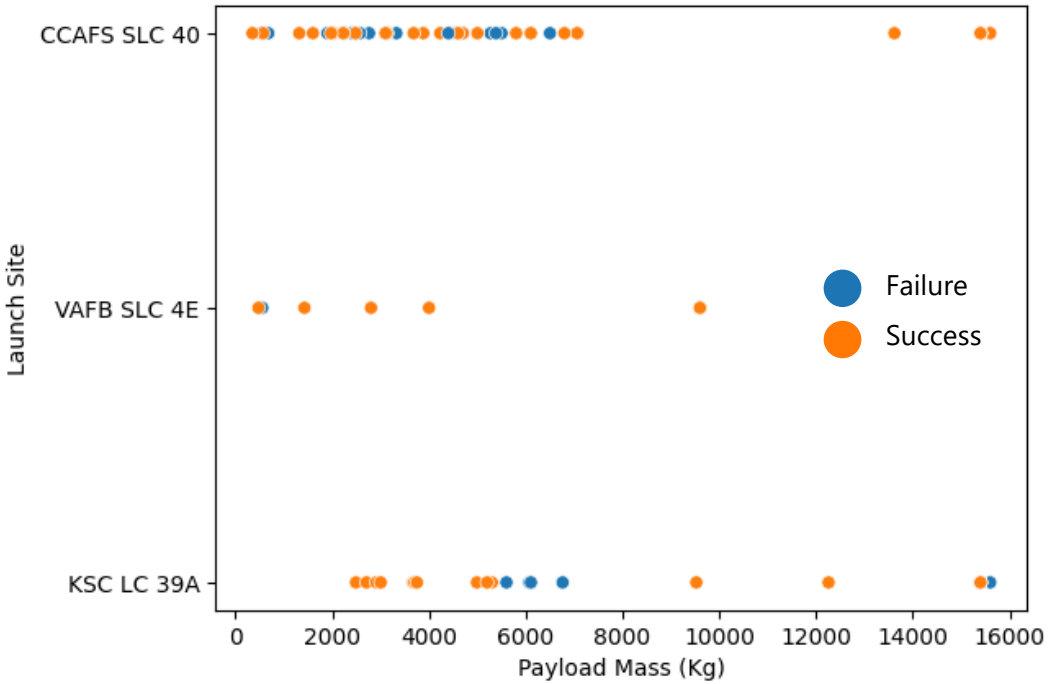
Insights:

- **Later flights** have achieved a **higher success rate** when compared to **earlier ones,** which will **hopefully continue to improve with time.**
- **Most launches (~ 50%)** occur from the **CCAFS SLC 40**.
- **CCAFS SLC 40** has registered a **higher failure rate** than **KSC LC 39A** and **VAFB SLC 4E.**

In terms of Payload Mass, although a higher weight usually represents a higher success rate, the KSC LC 39A launches have performed better when lighter.

Results | EDA with Visualization – Payload vs. Launch Site

Scatter Plot of Payload Mass vs. Launch Site [kg]



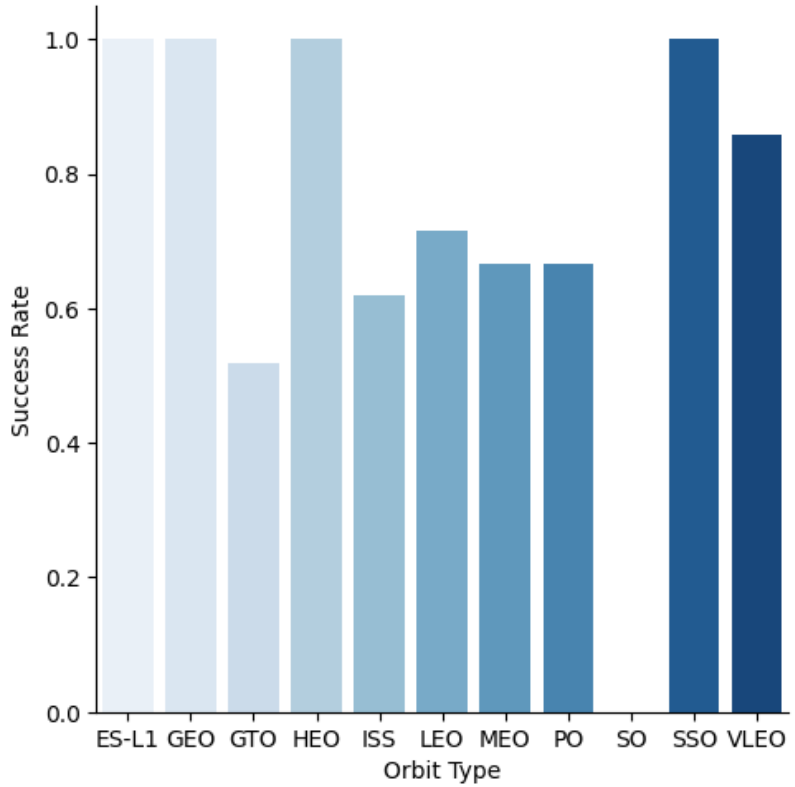
Insights:

- On average, the **higher the Payload mass (kg)**, the **higher the success rate.**
- Most launches with a **Payload mass greater than 8,000 kg were successful.**
- All **CCAFS SLC 40 launches** with a **payload over 12,000 kg** and **KSC LC 39A** with a **Payload under 5,500 kg** have a **100% success rate.**

Across all Orbit Types, the ones that registered a perfect success rate are: ES-L1, GEO, HEO and SSO.

Results | EDA with Visualization – Success Rate vs. Orbit Type

Bar Chart of Orbit Type vs. Success Rate [%]



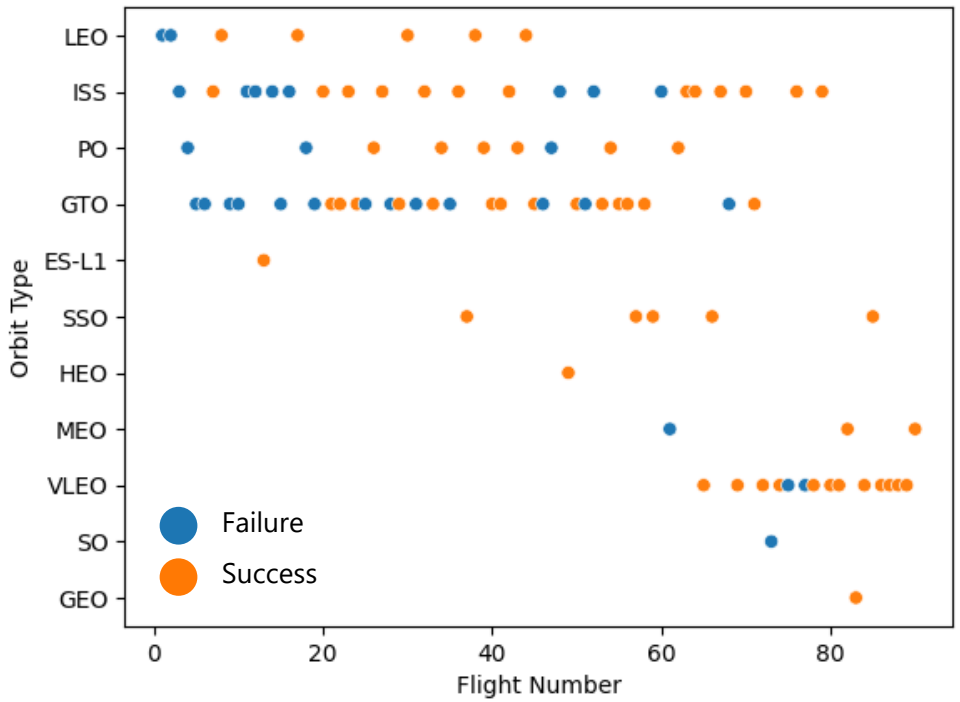
Insights:

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

On average, the flights success rate increases with the number of flights for each orbit.

Results | EDA with Visualization – Flight Number vs. Orbit Type

Scatter Plot of Flight Number vs. Orbit Type

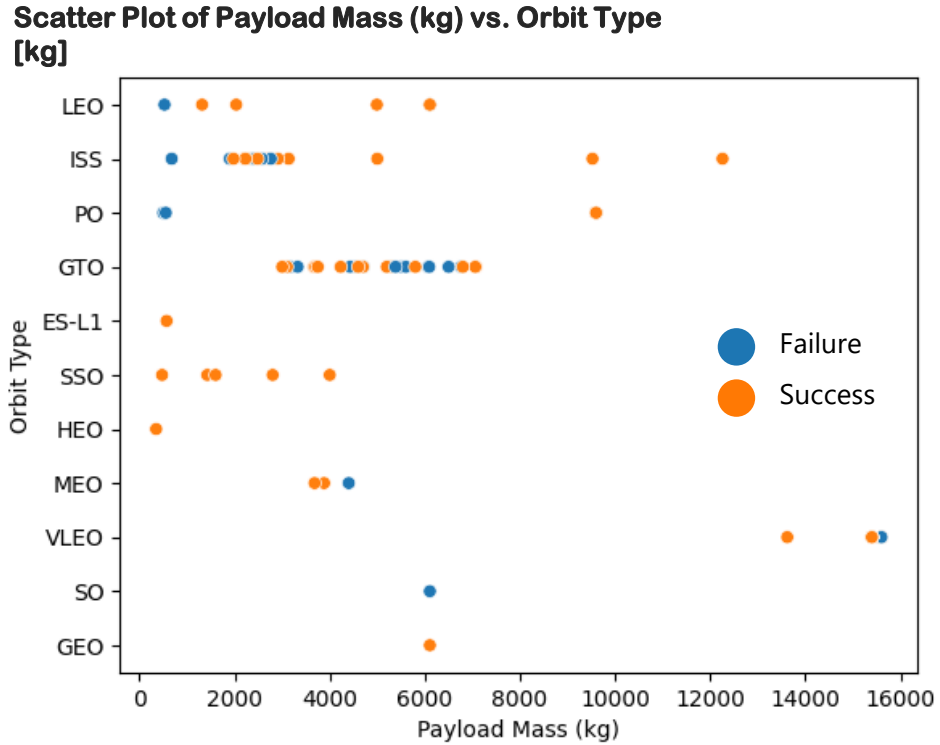


Insights:

- The **flights success rate**, on average, **increases** with the **number of flights** for each orbit.
- The **SSO orbit** showcased a **perfect success rate** on its 5 flights.
- The **GTO orbit** registered the **highest number of failure flights**.

Across all Orbit Types, heavier Payload flights seem to achieve better success rates.

Results | EDA with Visualization – Payload vs. Orbit Type



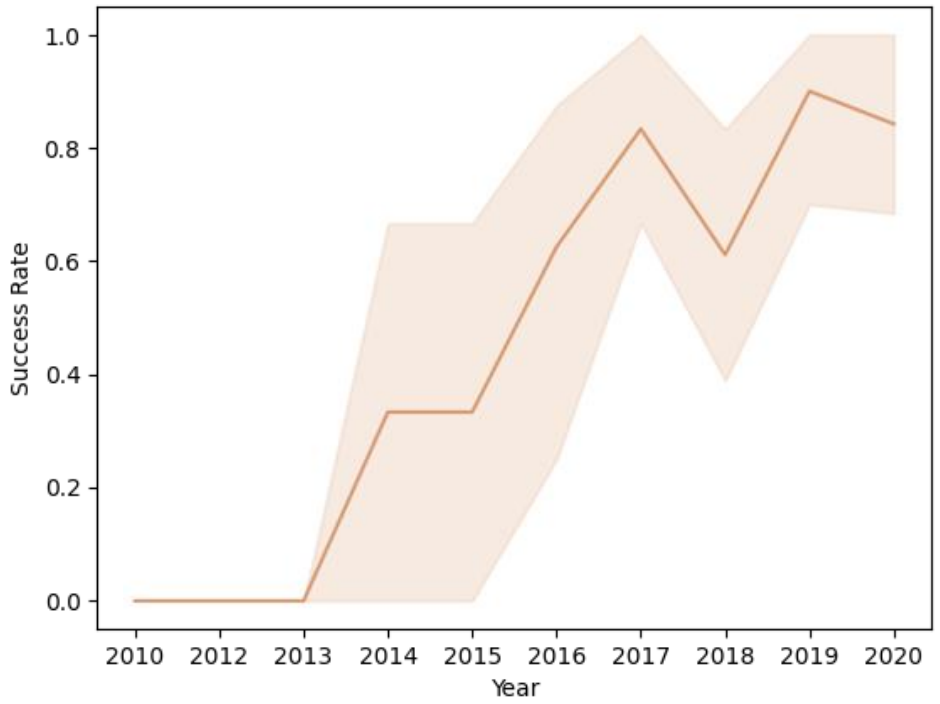
Insights:

- **LEO and ISS** showcased a clear improvement on flights with a **heavier Payload**.
- The **GTO orbit flights' success** is very **inconsistent regardless of flights' Payload mass.**

Flights' success rate has been improving since 2013.

Results | EDA with Visualization – Success Rate vs. Orbit Type

Line Plot of Flight's Success Rate Yearly Trend [%]



Insights:

- Overall, the **success rate has been improving since 2013.**
- Despite this tendency for improvement, there were **two major drops** from **2017-2018** and most recently from **2019-2020**.

C2. EDA with SQL

Insights Summary

There are 5 unique launch site names.

Results | EDA with SQL – All Launch Site Names

Using a SELECT DISTINCT clause, all the unique launch site names were retrieved from the SpaceX Table.

Launch Site Names:

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

SQL Query:

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

```
[53]
 * sqlite:///my_data1.db
Done.
```

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

5 records were retrieved from the launch sites that start with the string “CCA”.

Results | EDA with SQL - Launch Site Names Begin with 'CCA'

Using a WHERE clause with a LIKE operator to search for the matching pattern “CCA%”, 5 unique records were retrieved from the SpaceX Table.

SQL Query:

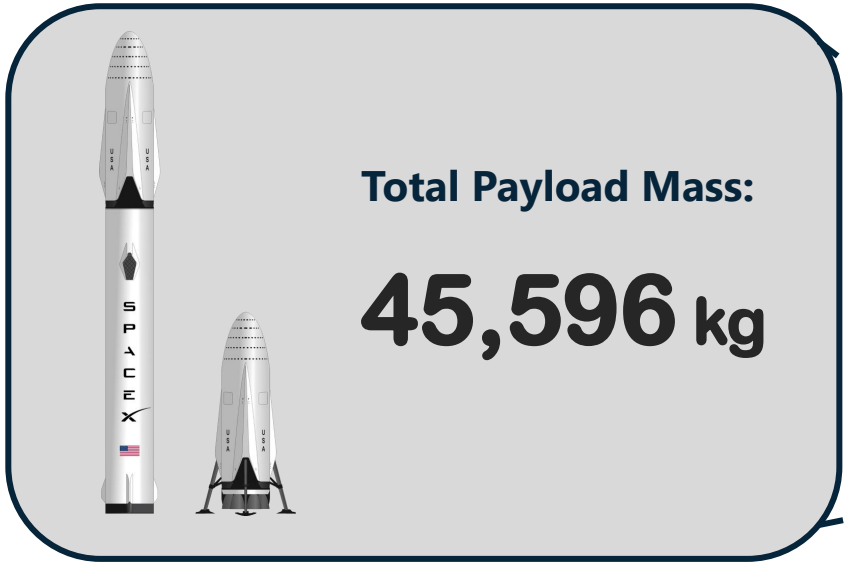
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

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

NASA (CRS) boosters have carried a total of 45,596 kg of Payload mass.

Results | EDA with SQL – Total Payload Mass

Using a SUM operation with a LIKE operator to search for the “NASA (CRS)” boosters, the total payload mass was computed.



Total Payload Mass:

45,596 kg

SQL Query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA (CRS)';
```

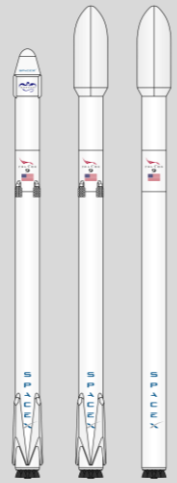
```
[17]
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

The F9 v1.1 has registered an average Payload mass of 2,928 kg.

Results | EDA with SQL – Average Payload Mass by F9 v1.1

Using the AVG operation with a LIKE operator to search for the “F9 v1.1” booster, the average payload mass for this version was computed.



Falcon 9 v1.1

Average Payload Mass by F9 v1.1:

2,928 kg

SQL Query:

```
[18] %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

AVG(PAYLOAD_MASS__KG_)
2928.4
```

The first successful landing in ground pad happened on 22/12/2015.

Results | EDA with SQL – Average Payload Mass by F9 v1.1

Using the MIN operation with a LIKE operator to search for a “Success (ground pad)” landing, the first record dates the 22/12/2015.

22/11/2015

SQL Query:

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (ground pad)';
```

```
[20]
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2015-12-22

Out of 101 Mission Outcomes, ~ 98 % were Successful.

Results | EDA with SQL -

Using the COUNT operation to search for the total number of mission outcomes the GROUP BY statement for aggregating them according to type of mission outcome, the following results were retrieved.

Mission Outcomes:

- **99** Success
- **1** Success (payload status unclear)
- **1** Failure (in flight)

SQL Query:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as Total_Number \
FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

[23]

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

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

12 Booster versions carried the maximum Payload while flying.

Results | EDA with SQL

Using the WHERE clause and a sub SELECT query to filter the booster versions that carried the maximum payload mass, the following results were retrieved.

SQL Query:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

[25]

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

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

Between 04/06/2010 and 20/03/2017 the most common landing outcome was No Attempt, followed by Success and Failure (Drone Ship) .

Results | EDA with SQL

Using the WHERE clause to filter the records within 04/06/2010 and 20/03/2017, the GROUP BY statement for aggregating them according to type of landing outcome, and the ORDER BY to rank in descending order, the following ranking was retrieved.

Ranking between 04/06/2010 and 20/03/2017:

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

SQL Query:

```
%sql SELECT Landing_Outcome, COUNT(*) AS count_outcomes \
FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY count_outcomes DESC;
```

[52]

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

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

C3. Folium Interactive Maps

Launch Site Analysis

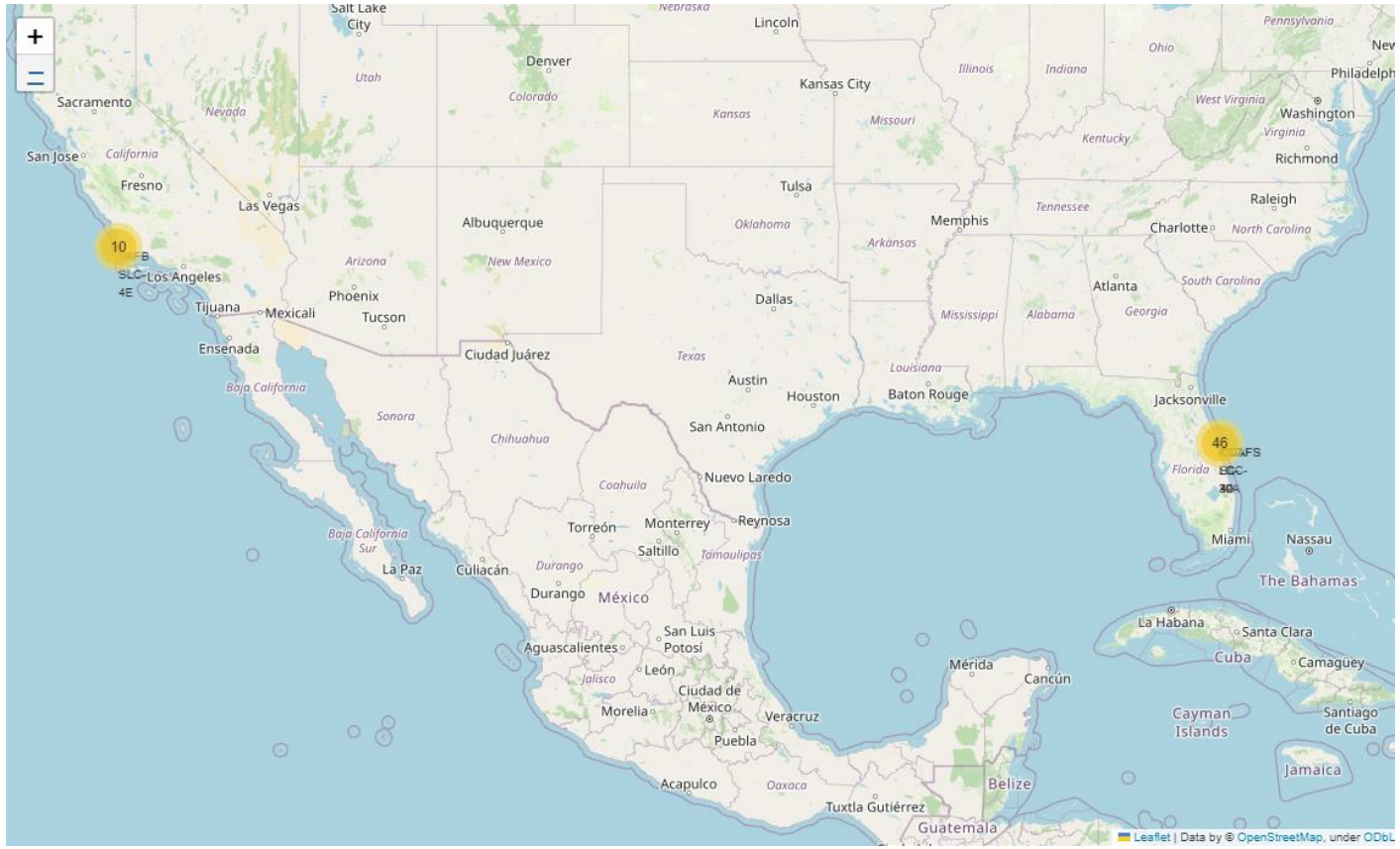
Rockets are often launched near the Equator to benefit from Earth's natural boost, reducing the need for extra fuel and boosters.

Results | Folium Interactive Maps – Launch Sites

Equator Importance:

Closer proximity to the equator offers a launch site **significant advantages** for **achieving equatorial orbits**. The increased velocity from Earth's rotation at these locations aids in launching rockets into prograde orbits, providing a **natural boost that reduces the need for additional fuel** and boosters, consequently **cutting down on costs.**

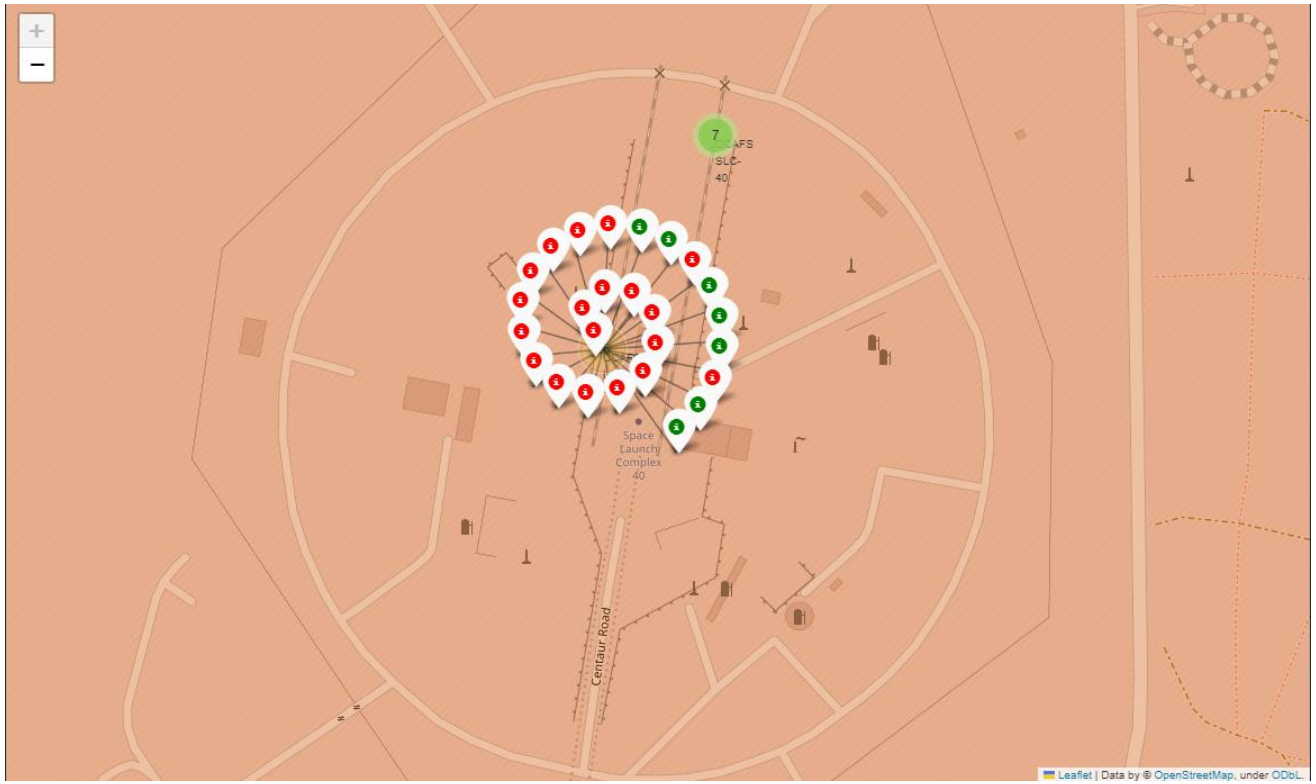
Launch Sites Map Markers [Number]



The flight success rate registered at CCAFS LC-40 Site is around 27%.

Results | Folium Interactive Maps – Launch Outcomes

Launch Outcomes at CCAFS LC-40 Site Map
[Red Markers – Failure, Green Markers - Success]



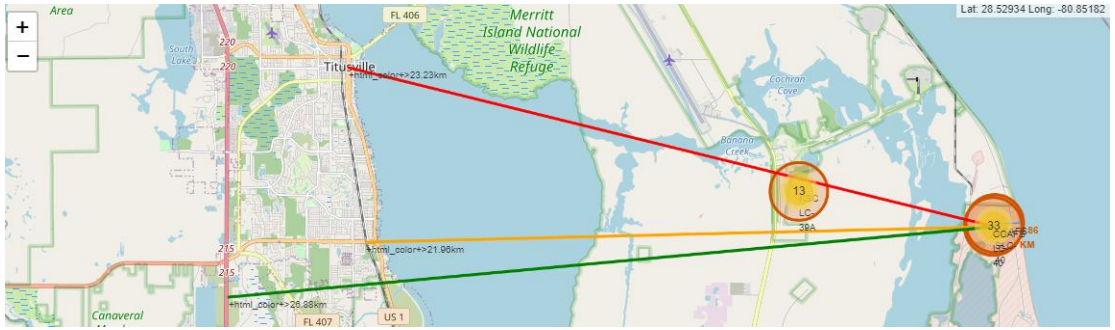
At CCAFS LC-40 Site :

- The **flight success rate** is **~ 27%,** with 19 out of 26 flights registering any sort of failure.

Optimizing the Launch Site location is key for Safety and Operational Logistics.

Results | Folium Interactive Maps – Launch Site Proximities

CCAFS LC-40 Launch Site Proximities Map [Km]



From CCAFS LC-40 Site :

- **0.86 Km** to the **Nearest Coastline** – Brevard County Florida;
- **21.96 Km** to the **Nearest Outside Highway** – Cheney Highway;
- **22.45 Km** to the **Nearest Outside Railway** – Florida East Coast Railway;
- **23.23 Km** to the **Nearest Outside City** – Titusville.

Importance:

Coastal launch sites ensure **safety** by containing debris from spent stages or failed launches, requiring **exclusion zones** for **security** while **strategically locating facilities away from vulnerable areas** yet **near transportation hubs for logistical support**.

Sources: SpaceX Dataset, Open Street Map

C4. Plotly Dash Dashboards

Records Metrics

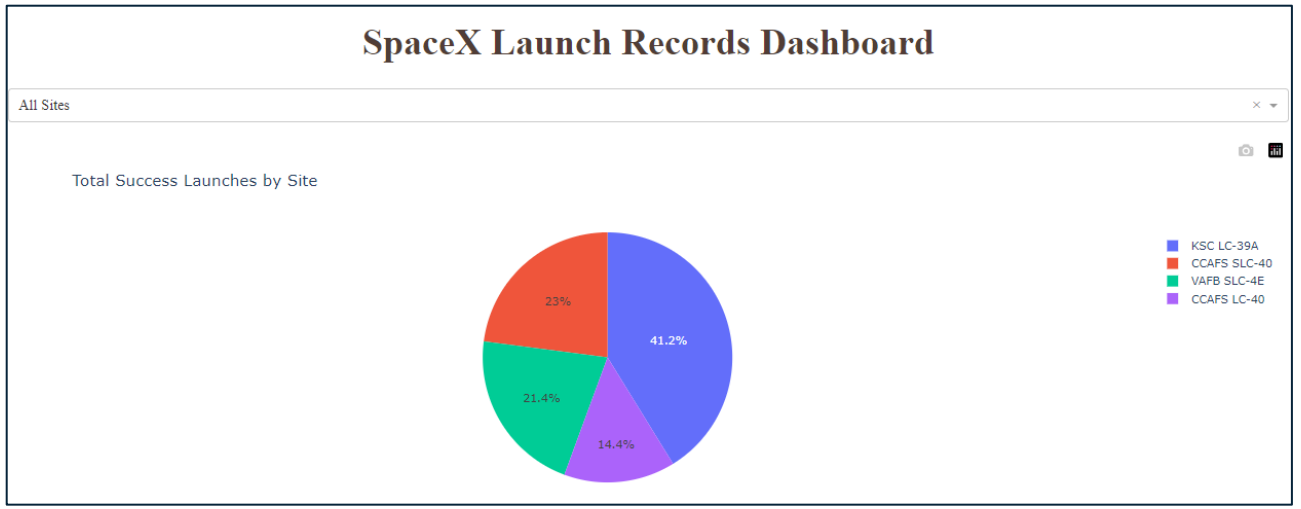
KSC LC-39A registered the most successful launches amongst all launch sites - 41.2%.

Results | Plotly Dash Dashboards – Launches by Site

Insights:

- **KSC LC-39A registered the most successful launches** amongst all launch sites - **41.2%;**
- **CCAFS LC-40 registered the least successful launches** amongst all launch sites **- 14.4%.**

Total Success Launches by Site Dashboard Pie Chart [%]



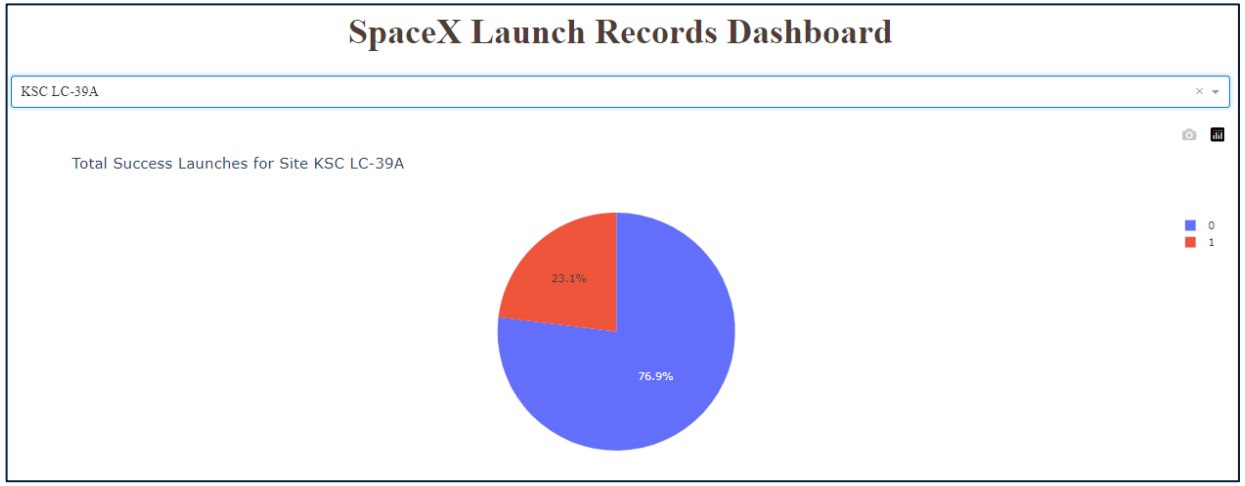
KSC LC-39A registered the highest success rate amongst all launch sites – 76.9%.

Results | Plotly Dash Dashboards – Launches by KSC LC-39A Site

Insight:

- **KSC LC-39A** registered the **highest success rate** amongst all launch sites **– 76.9%.**

Launches for KSC LC-39A Site Dashboard Pie Chart [%]



Payloads between 2,000 Kg and 5,800 Kg register the highest success rate.

Results | Plotly Dash Dashboards – Launches by KSC LC-39A Site

Insight:

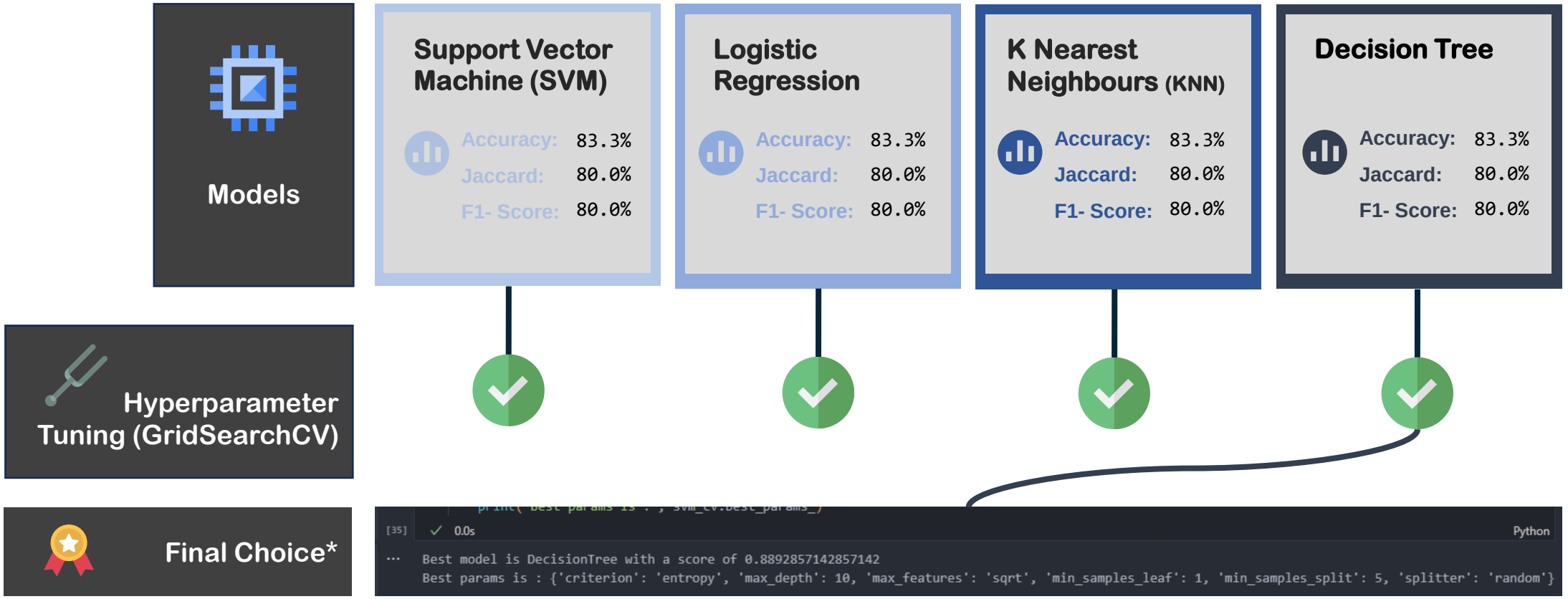
- **Payloads** between **2,000 Kg and 5,800 Kg** register the **highest success rate.**

D. Predictive Analysis

Classification

Despite all models performing at the same level accuracy, likely due to the small dataset, the Decision Tree had a better .best_score.

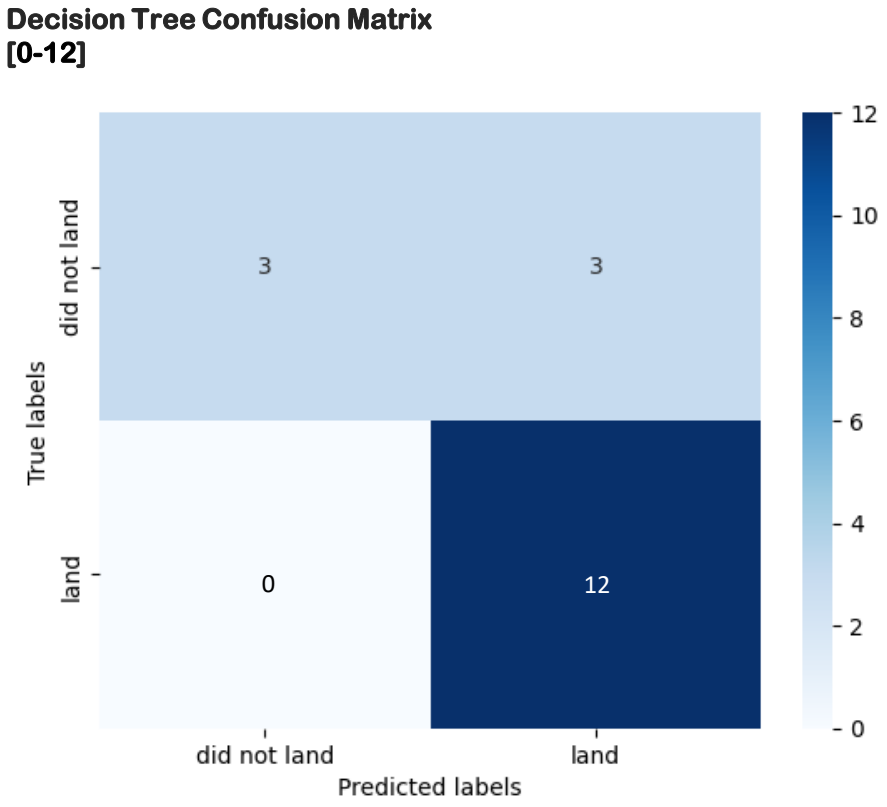
Predictive Analysis | Model Picking - Accuracy



***Note:** The .best_score method represents the average score of all cross-validation scores for a single tested combination of parameters.

The Confusion Matrix showcases a solid level of Precision and Recall, despite the existence of some False Positives (Type I Error).

Predictive Analysis | Model Picking - Accuracy



In terms of classifying the **success or failure of flights' landing**, this **confusion matrix**, sums up the following performance metrics regarding our Decision Tree Model:

- **Precision** = TP / (TP + FP)

$$\frac{12}{15} = 0.8$$

Fraction of relevant instances among the retrieved instances.

- **Recall** = TP/(TP + FN)

$$\frac{12}{12} = 1$$

Fraction of relevant instances that were retrieved

Having 3 False Positives (FP), the model presents **Type I error.**

Sources: Machine Learning Analysis

D. Conclusions & Further Research

Key Takeaways

Regardless of the project's success in classifying successful landings, its potential could be significantly amplified by exploring novel directions.

Conclusions & Further Research

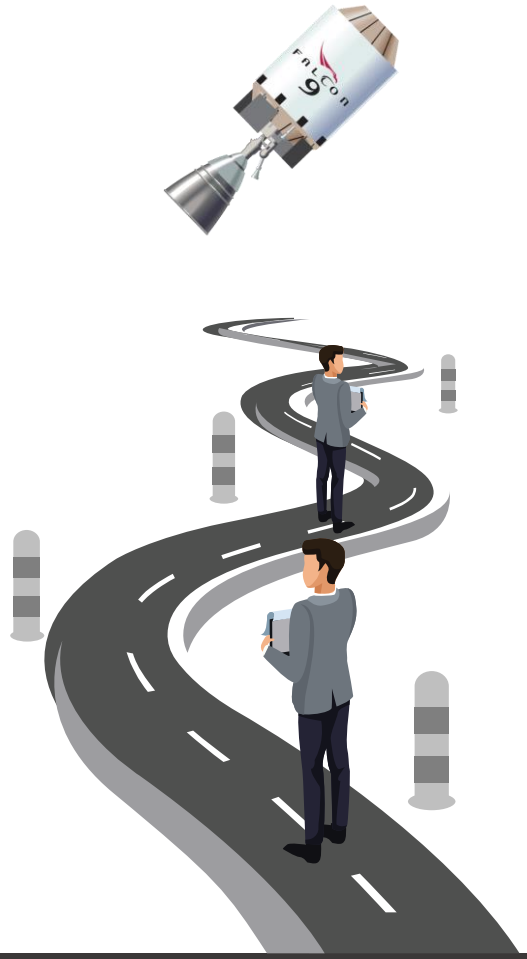
Conclusions

The **classification models** exhibited **comparable performance**, leaving room for **refinement** specifically in **reducing Type I errors**.

Over time, there has been a **notable increase in launch success**.

There is a **positive correlation** between **higher payload** masses and **greater success across all launch sites.**

KSC LC-39A is the **highest success rate** launch site and the best orbits are: **ES-L1, GEO, HEO, and SSO.**



Further Research

As the **dataset becomes larger**, the **model could be refined** with more training data to **support its predictions**.

Light GBM or **XGBoost**, two alternative distributed high-performance framework models **could be used to improve these results.**

Performing **Polynomial Feature Analysis** and **PCA** may possibly refine the **model's accuracy.**

References

Information Sources

References

[1] Coleman, S. (2018, June 14). SpaceX launch data. Kaggle. https://www.kaggle.com/datasets/scoleman/spacex-launch-data;

[2] Navigation. Why is it better to launch a spaceship from near the equator? (n.d.). https://www.qrg.northwestern.edu/projects/vss/docs/navigation/2-why-launch-from-equator.html;

[3] Online courses & credentials from top educators. join for free. Coursera. (n.d.). https://www.coursera.org/learn/applied-data-science-capstone/home/week/5;

[4] SpaceX. (n.d.). https://www.spacex.com/.