# DATA SCIENCE CAPSTONE PROJECT

Ayşenur Tunç
22.09.2024

# Outline

# Executive Summary

## Summary of Methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

## Summary of all Results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

# Introduction

## Project Context

The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes Falcon 9 rocket flights, which start at 62 million dollars; in comparison, other suppliers charge up to 165 million dollars per launch; a large portion of the cost savings are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the launch cost if we can ascertain if the first stage will land. We will forecast if SpaceX will reuse the first stage based on available data and machine learning techniques.

To be answered are the following questions:

- In what ways do factors like payload mass, launch site, and number of It's?

- How do flights and orbits impact the first stage landing's effectiveness?

- Does the number of successful landings rise with time?

- In this scenario, which algorithm works best for binary classification?

# Introduction

## Project Context

The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes Falcon 9 rocket flights, which start at 62 million dollars; in comparison, other suppliers charge up to 165 million dollars per launch; a large portion of the cost savings are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the launch cost if we can ascertain if the first stage will land. We will forecast if SpaceX will reuse the first stage based on available data and machine learning techniques.

To be answered are the following questions:

- In what ways do factors like payload mass, launch site, and number of It's?

- How do flights and orbits impact the first stage landing's effectiveness?

- Does the number of successful landings rise with time?

- In this scenario, which algorithm works best for binary classification?

# Introduction

## Project Context

The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes Falcon 9 rocket flights, which start at 62 million dollars; in comparison, other suppliers charge up to 165 million dollars per launch; a large portion of the cost savings are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the launch cost if we can ascertain if the first stage will land. We will forecast if SpaceX will reuse the first stage based on available data and machine learning techniques.

To be answered are the following questions:

- In what ways do factors like payload mass, launch site, and number of It's?

- How do flights and orbits impact the first stage landing's effectiveness?

- Does the number of successful landings rise with time?

- In this scenario, which algorithm works best for binary classification?

# Methodology

# Methodology

Data collection methodology:
- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Perform data wrangling
- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models
- Building, tuning and evaluation of classification models to ensure the best results

# Data Collection

A combination of web scraping data from a table in SpaceX's Wikipedia entry and API requests from the company's REST API were used in the data collection process.

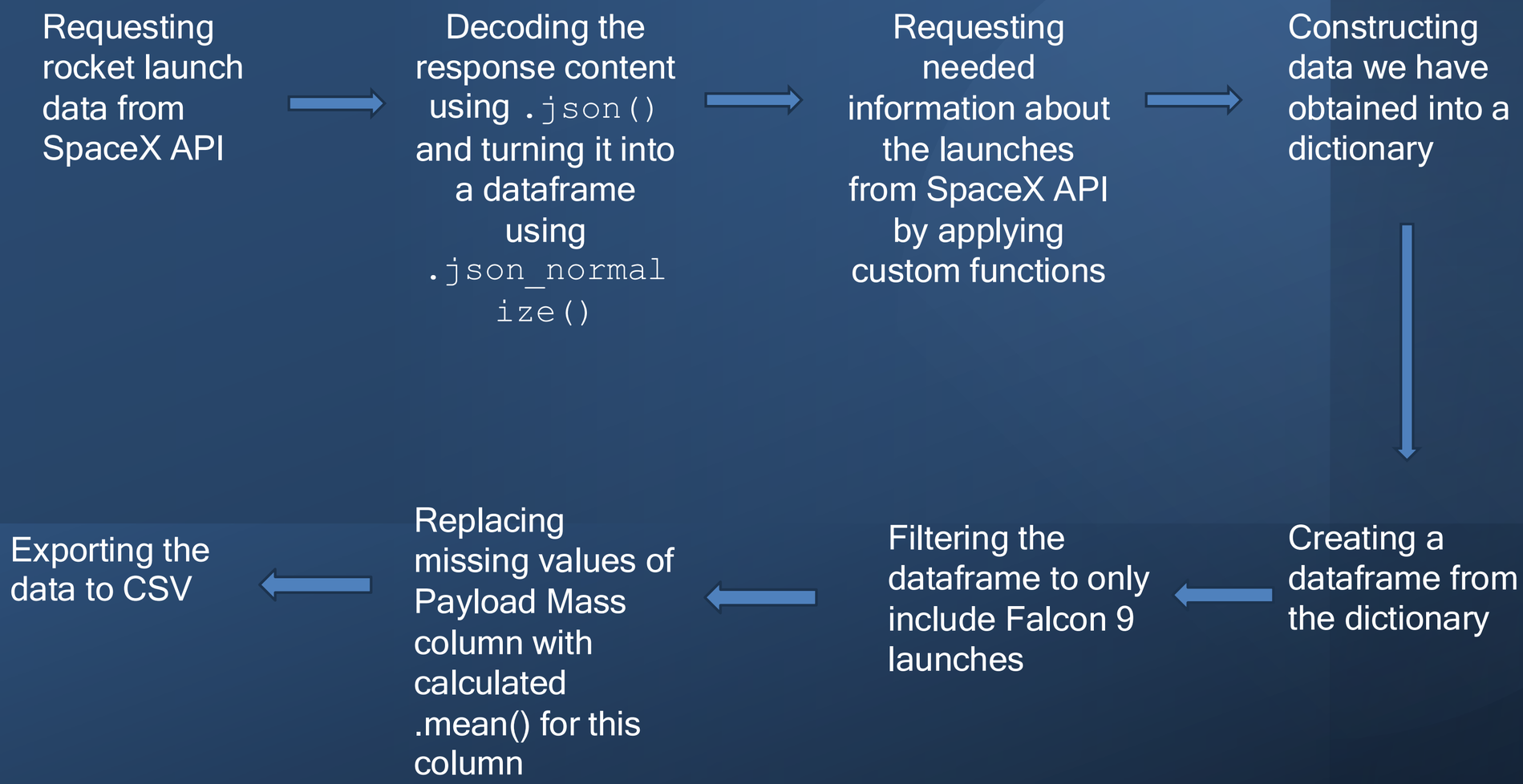
SpaceX REST API is used to obtain Data Columns:

FlightNumber - Date - BoosterVersion - PayloadMass - Orbit - LaunchSite - Outcome - Flights GridFins - Reused - Legs - LandingPad - Block - ReusedCount – Serial - Longitude Latitude

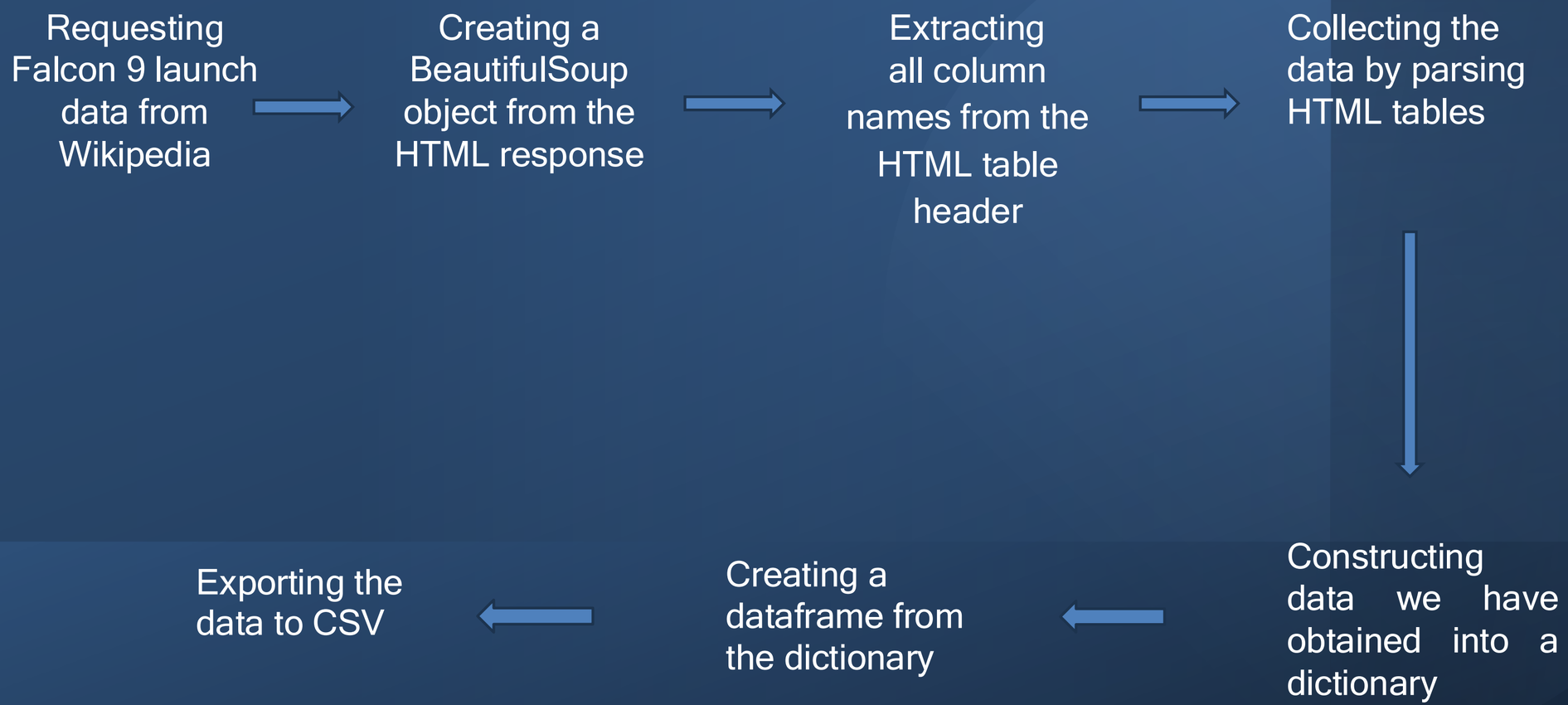
Data Columns are acquired through the use of web scraping on Wikipedia:

Flight No - Launch site - Payload - PayloadMass - Orbit – Customer - Launch outcome - Version Booster - Booster - landing - Date - Time

# Data Collection – SpaceX API



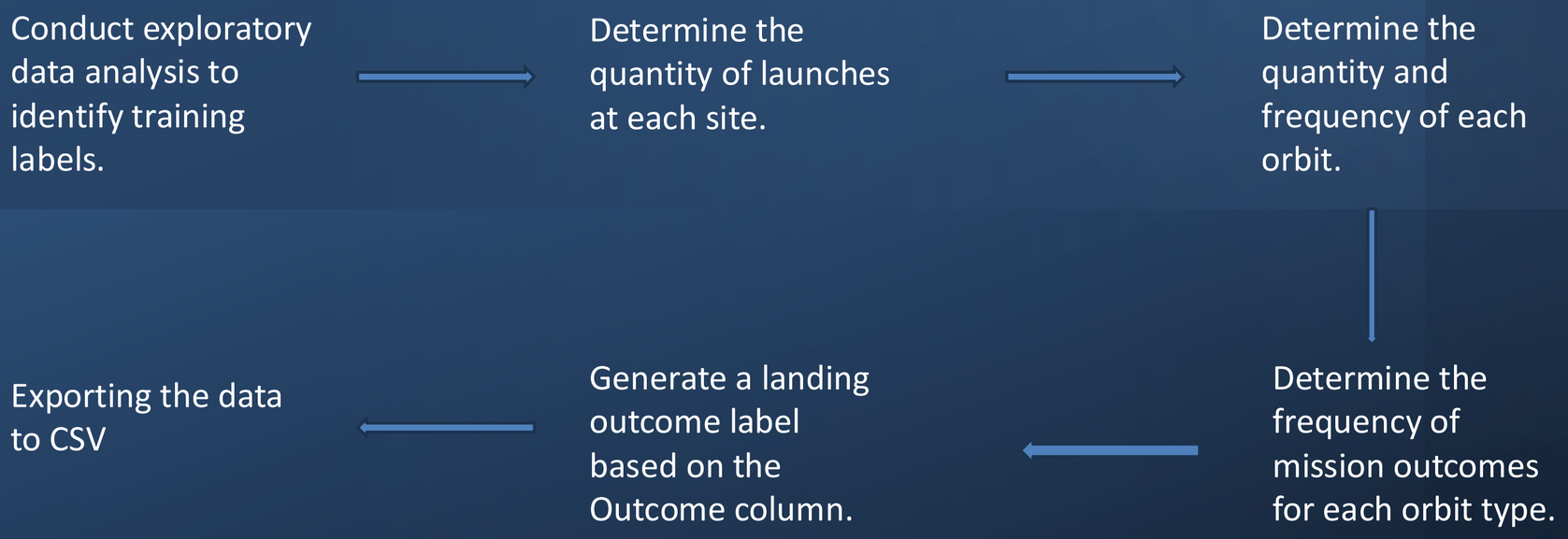
# Data Collection – Scraping



# Data Wrangling

At times, a landing was attempted but unsuccessful due to an accident.

The outcomes are primarily converted into Training Labels where "1" indicates a successful landing of the booster and "0" indicates an unsuccessful landing.



# EDA with Data Visualizations

-Charts were created:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

-Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

-Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

-Line charts show trends in data over time (time series).

# EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

# Build an interactive map with Follium

Markers have been added to all Launch Sites with a Circle, Popup Label, and Text Label for NASA Johnson Space Center, using its latitude and longitude coordinates as the starting location. Similarly, markers have been added for all Launch Sites, displaying their geographical locations and proximity to the Equator and coasts.
Colored markers indicating launch outcomes (Green for success, Red for failure) have been added using Marker Cluster, highlighting launch sites with higher success rates.
Colored lines have been added to show distances between Launch Site KSC LC-39A (for example) and nearby landmarks such as Railway, Highway, Coastline, and the Closest City on GitHub.

# Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:
- Implemented a dropdown list for selecting Launch Sites.

Pie Chart showing Success Launches (All Sites/Certain Site):
- Included a pie chart displaying the total count of successful launches for all sites and the comparison of Success vs. Failure. Failed counts are recorded for the selected launch site.

Slider for Payload Mass Range:
- A slider has been included for selecting the payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- A scatter chart has been included to display the correlation between Payload and Launch Success.

# Predictive Analysis (Classification)

Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Create a GridSearchCV object with the parameter cv set to 10 in order to identify the optimal parameters.

GridSearchCV is applied to Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors models.

Calculating the accuracy on the test data for all models using the .score() method.

Examining the confusion matrix for all models

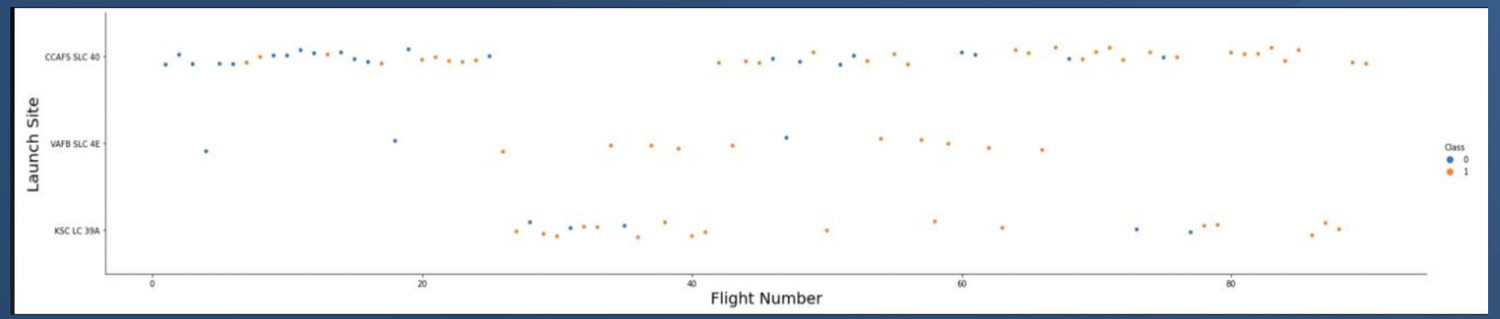
Analyzing the metrics for F1_score and Jaccard_score allows one to determine which method works best.

# Results

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

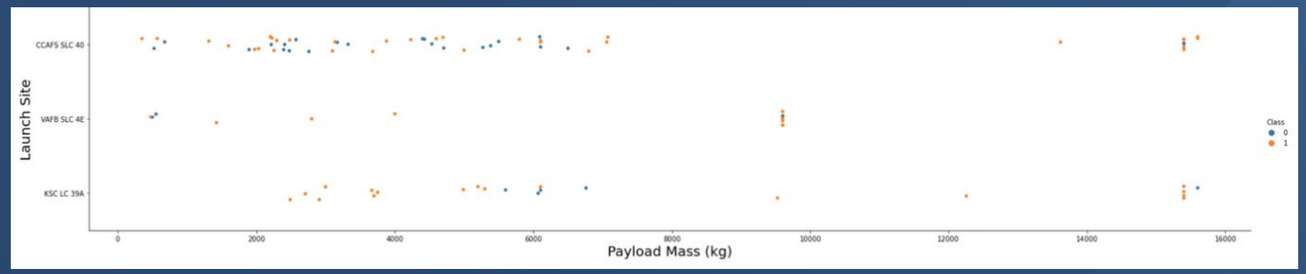
# EDA With Visualization

# Flight Number vs Launch Site



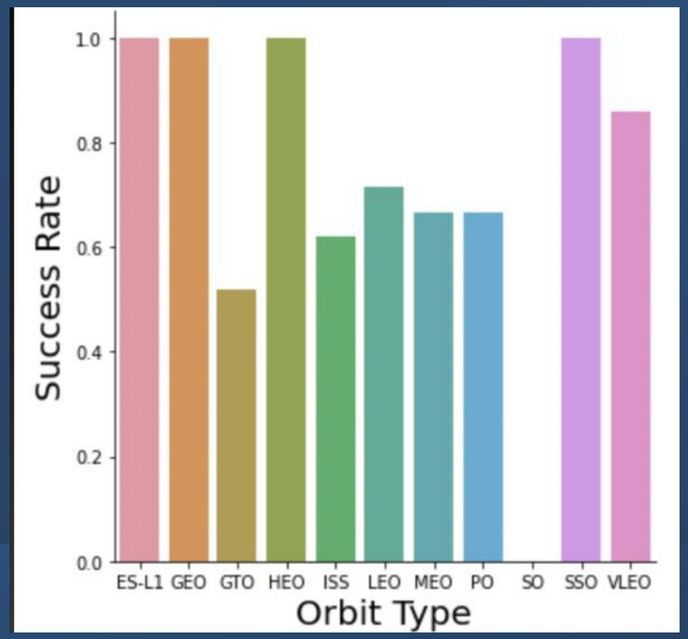

-All of the most recent flights were successful, while none of the earlier ones were.
-It can be assumed that every new launch has a higher success rate.
-VAFB SLC 4E and KSC LC 39A have higher success rates.
-Approximately half of all launches occur at the CCAFS SLC 40 launch site.

# Payload vs Launch Site



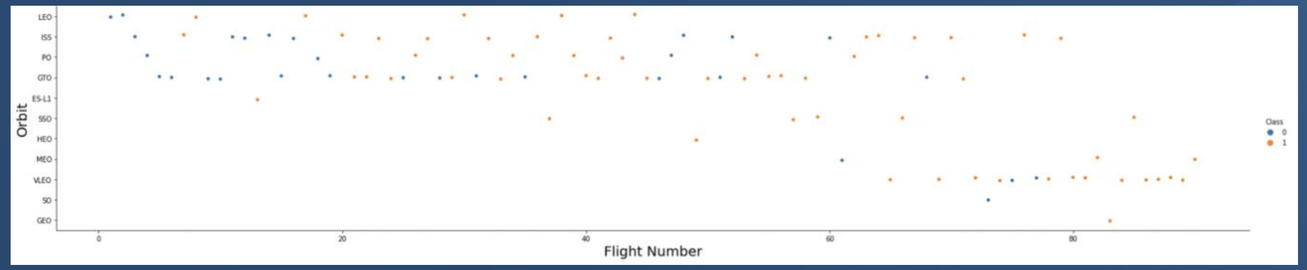

- The majority of launches with payload masses exceeding 7000 kg were successful
- The higher the payload mass, the higher the success rate for each launch site
- Under 5500 kg, KSC LC 39A also has a 100% success rate for payload mass.

# Success Rate vs Orbit Type



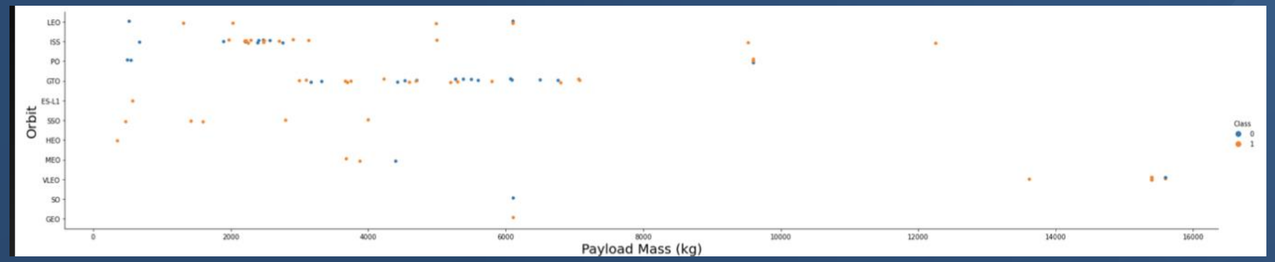

- 100% success rate orbits: GEO, HEO, SSO, and ES-L1
- Orbits with 0% succeed: SO
- Orbits with a 50% to 85% success rate: PO, MEO, ISS, GTO, and LEO

# Flight Number vs Orbit Type



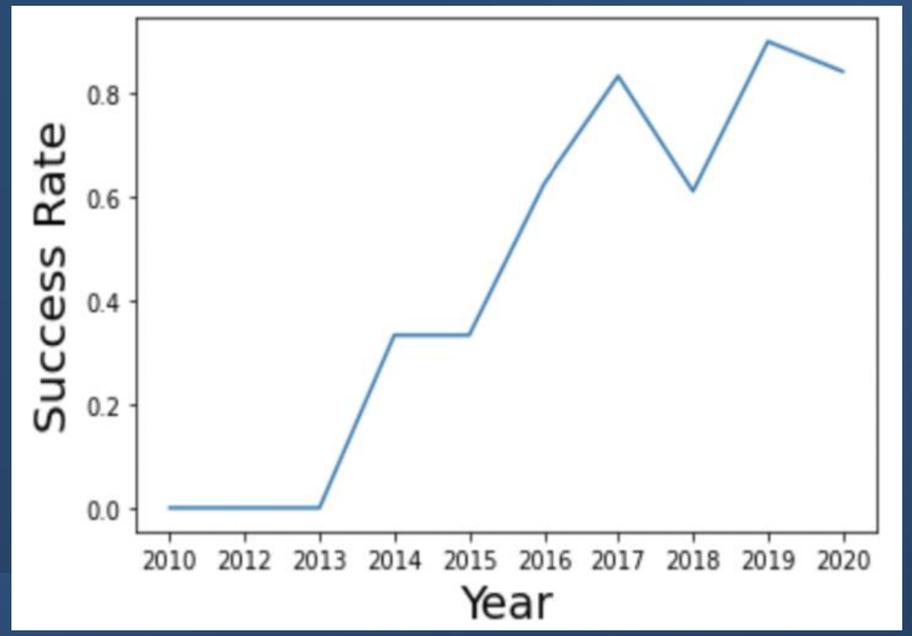

In LEO orbit, success appears to be related to the number of flights; however, in GTO orbit, there appears to be no relationship between flight number.

# Payload Mass vs Orbit Type




Large payloads adversely affect GTO orbits while benefiting GTO and Polar LEO (ISS) orbits.

# Launch Success Yearly Trend




Since 2013, the success rate has increased steadily until 2020.

# All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Displaying the names of each unique launch site in the space mission.

# Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 00:35:00 | 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 |

Displaying 5 records where the launch site begins with the string 'CCA'.

# Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

| total_payload_mass |
| --- |
| 45596 |

Displaying the total payload mass carried by NASA's boosters (CRS).

# Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

| average_payload_mass |
| --- |
| 2534 |

Shows the average payload mass carried by booster version F9 v1.1.

# First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

| first_successful_landing |
| --- |
| 2015-12-22 |

Indicating the date of the ground pad's first successful landing outcome.

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

List of boosters successful on a drone ship with a payload mass between 4000 and 6000.

# Total Number of Successful and Failure Mission Outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

| mission_outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

A list of all mission outcomes, both successful and unsuccessful.

# Boosters carried maximum payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

| 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 |

Naming the booster iterations that have transported the largest payload mass.

# 2015 launch records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

| MONTH | DATE | booster_version | launch_site | landing__outcome |
|---|---|---|---|---|
| January | 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| April | 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

A list of the drone ship's unsuccessful landings along with the names of the launch sites and the versions of their boosters for each month in 2015.

# Rank success count between 2010-06-04 and 2017-03-20

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
             where date between '2010-06-04' and '2017-03-20'
             group by landing__outcome
             order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

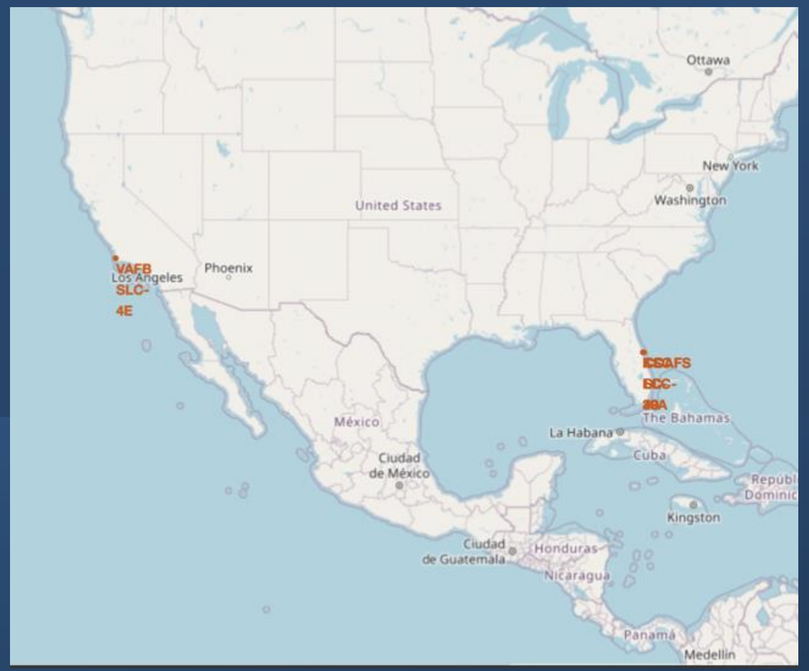
Out[13]:

| landing__outcome | count_outcomes |
|---|---|
| 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 |

Sorting the number of landing results (e.g., ground pad success or drone ship failure) between 2010-06-04 and 2017-03-20 in descending order.

# Interactive map with Folium

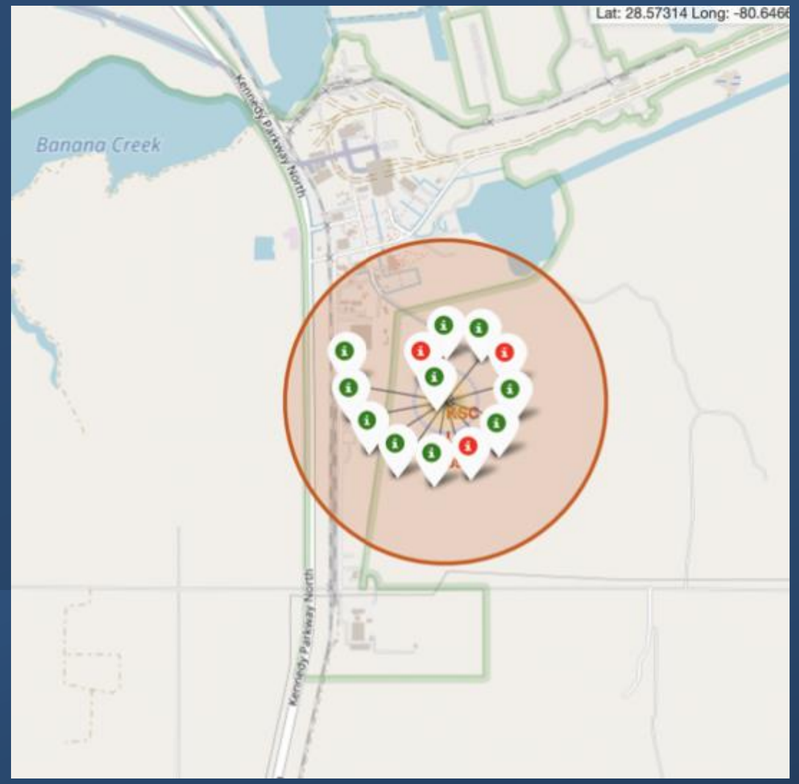
# All launch sites’ location markers on a global map




Most launch sites are located near the Equator. The land at the equator moves quicker than any other location on Earth. Anything located at the equator of the Earth is already moving at a speed of 1670 km/hour. If a ship is launched from the equator, it goes up into space while maintaining the same speed as when it was on the Earth. This is due to inertia. This velocity will enable the spacecraft to maintain a sufficient speed for orbital stability.

• All launch sites are located near the coast to minimize the risk of debris falling or exploding near people when rockets are launched towards the ocean.

# Colour-labeled launch records on the map

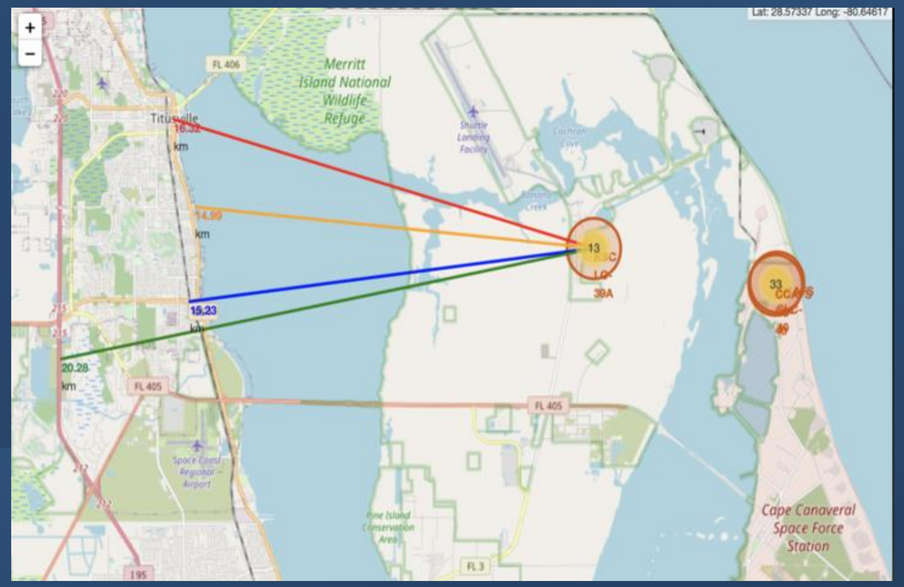



The colour-labeled markers will help us quickly determine the success rates of different launch sites.

- Green Marker = Successful Launch
- Red Marker = Failed Launch

• Launch Site KSC LC-39A has a very high Success Rate.

# Distance from the launch site KSC LC-39A to its proximities



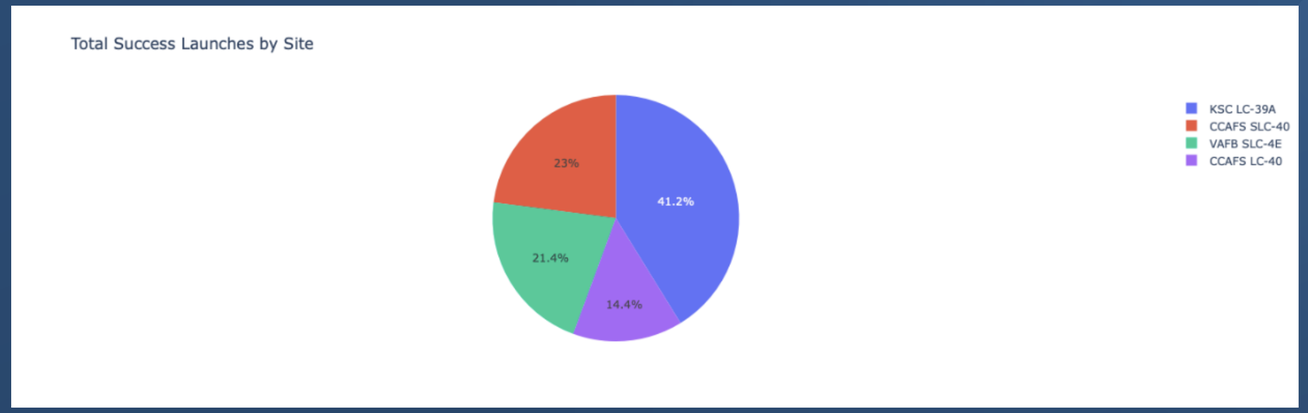

- The visual analysis of the launch site KSC LC-39A reveals its proximity to various transportation routes:
- Railway: 15.23 km
- Highway: 20.28 km
- Coastline: 14.99 km

- The launch site KSC LC-39A is also relatively close to its nearest city, Titusville, at a distance of 16.32 km.

- A failed rocket, traveling at high speed, can cover distances of 15-20 km in a matter of seconds. It has the potential to be hazardous to populated areas.

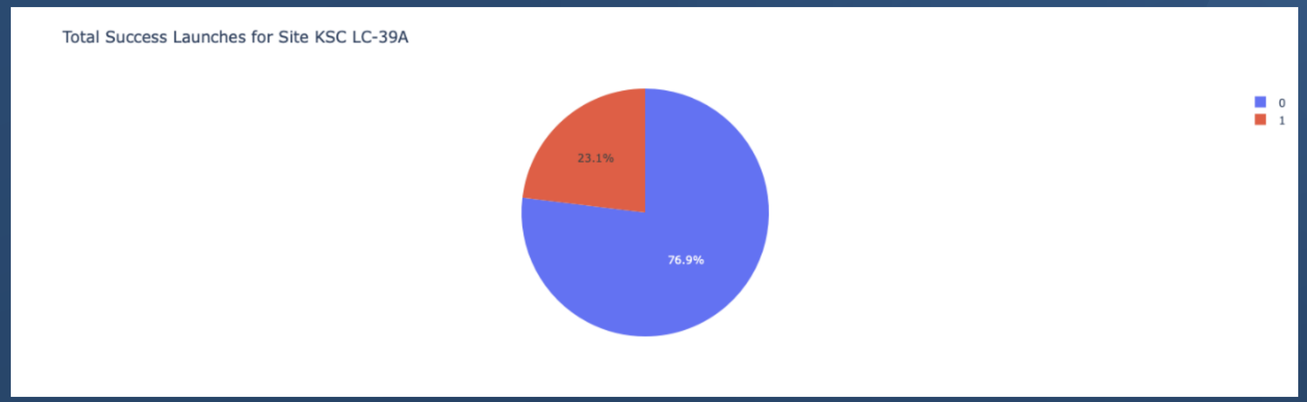
# Build a Dashboard with Plotly Dash

# Launch success count for all sites



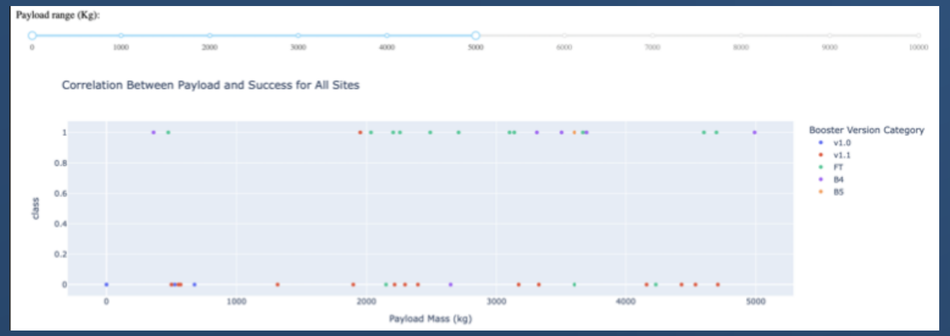

The graph unequivocally demonstrates that KSC LC-39A has the most successful launches out of all the sites.

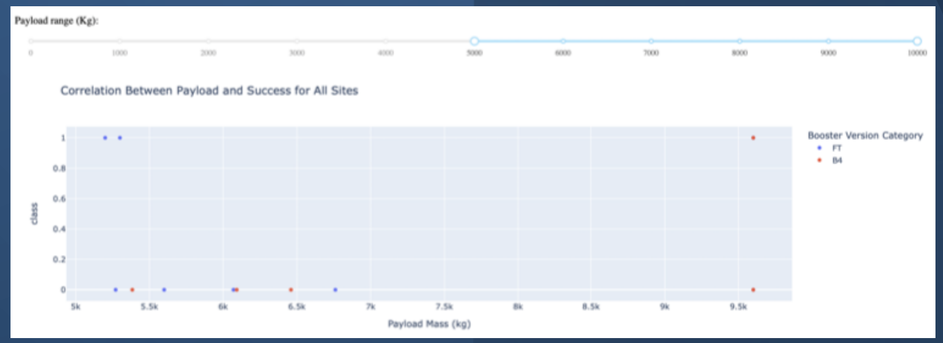
# Launch site with highest launch success ratio




With 76.9% launch success, KSC LC39A has the highest landing success rate, with 10 successful landings and only 3 failures.

# Payload Mass vs. Launch Outcome for all sites




The charts indicate that payloads ranging from 2000 to 5500 kg exhibit the greatest success rate.

# Predictive analysis (Classification)

# Classification Accuracy

Scores and Accuracy of the Test Set

| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy | 0.833333 | 0.833333 | 0.833333 | 0.833333 |

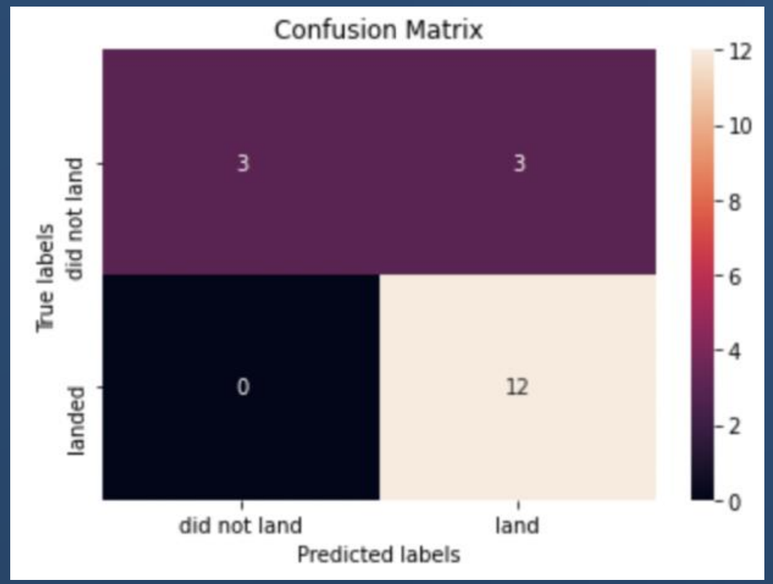
Scores and Accuracy of the Entire Data Set

| | LogReg | SVM | Tree | KNN |
|---|---|---|---|---|
| Jaccard_Score | 0.833333 | 0.845070 | 0.882353 | 0.819444 |
| F1_Score | 0.909091 | 0.916031 | 0.937500 | 0.900763 |
| Accuracy | 0.866667 | 0.877778 | 0.911111 | 0.855556 |

Based on the Test Set scores, it is inconclusive which method is the most effective.

- Similar Test Set scores could be attributed to the small sample size of 18 samples. Hence, we evaluated all methods using the entire dataset.
- The results from the complete dataset indicate that the Decision Tree Model is the most effective. This model achieves superior scores and highest accuracy.

# Confusion Matrix




Upon review of the confusion matrix, it is evident that logistic regression can effectively differentiate between various classes. The primary issue identified is false positives.

# Conclusion

The Decision Tree Model is the most suitable algorithm for analyzing this dataset.
- Launches with a low payload mass yield better results than those with a larger payload mass.
- Most launch sites are located near the Equator line and close to the coast.
- The success rate of launches has been increasing over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

# Appendix

IBM - United States

Coursera