SpaceX

Winning Space Race with Data Science

Mandeep Singh
10/08/2024

Presentation Contents

Executive Summary

Summary of Methodologies

The research aims to identify the key factors for successful rocket landings. To achieve this, the following methodologies were employed:

Data Collection : Gathered data using the SpaceX REST API and web scraping techniques.

Data Wrangling : Processed data to create a success/fail outcome variable.

Data Exploration : Used data visualization techniques to examine factors such as payload, launch site, flight number, and yearly trends.

Data Analysis : Analyzed data using SQL to calculate statistics including total payload, payload range for successful launches, and the total number of successful and failed outcomes.

Launch Site Analysis : Investigated success rates of launch sites and their proximity to geographical markers.

Visualization : Mapped launch sites to highlight those with the most success and their successful payload ranges.

Model Building : Developed models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN).

Results

Exploratory Data Analysis

Trend in Launch Success: There has been a noticeable improvement in launch success over time.

Top Landing Site: KSC LC-39A boasts the highest success rate among landing sites.

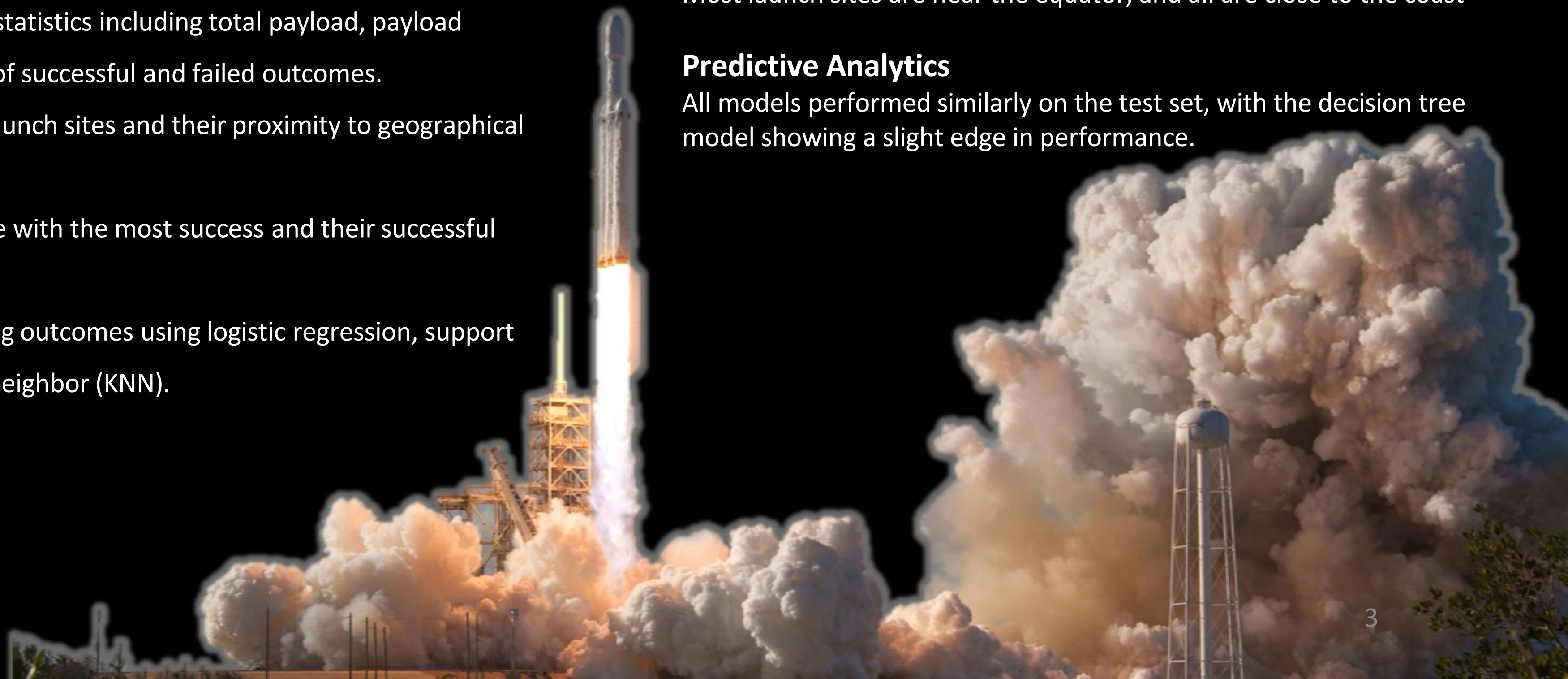
Orbital Success Rates: Orbits ES-L1, GEO, HEO, and SSO each have a 100% success rate

Visualization/Analytics

Most launch sites are near the equator, and all are close to the coast

Predictive Analytics

All models performed similarly on the test set, with the decision tree model showing a slight edge in performance.



Introduction

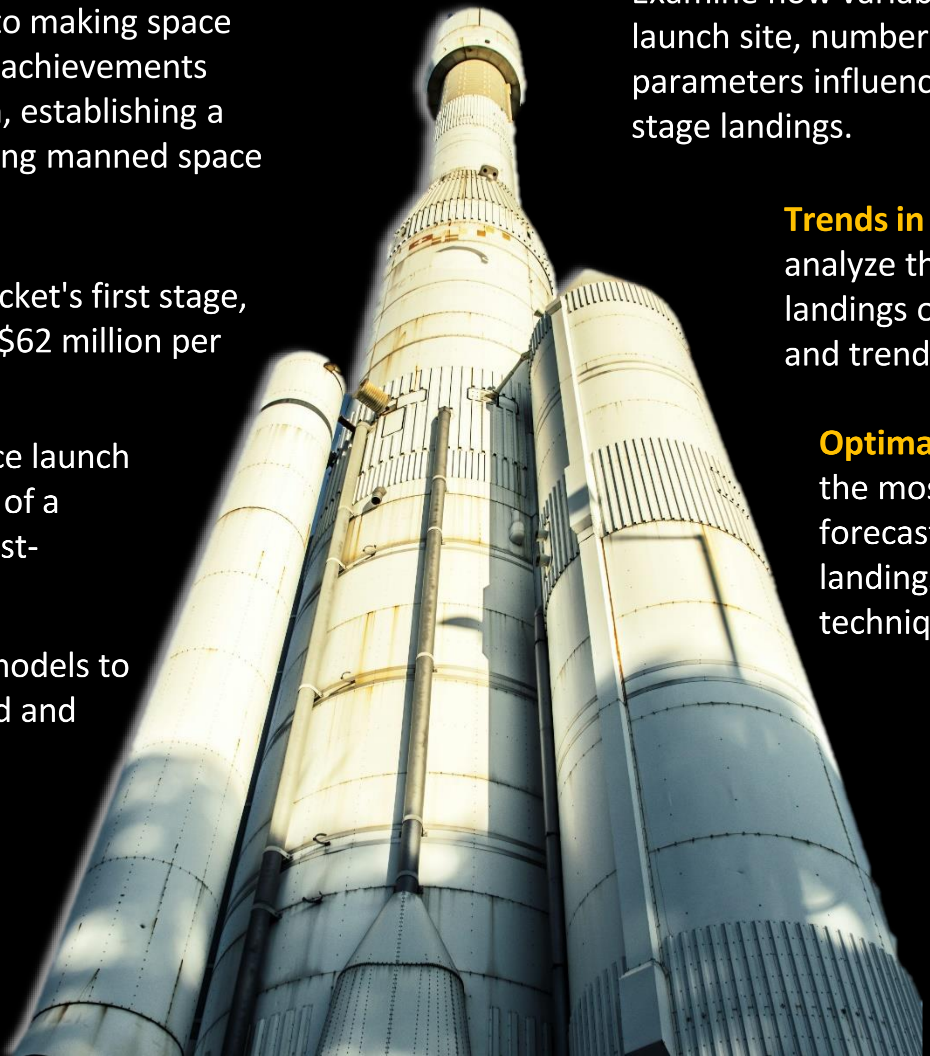
Background

SpaceX, a pioneering force in the space industry, is dedicated to making space travel more accessible and affordable. The company's notable achievements include launching spacecraft to the International Space Station, establishing a global satellite network for internet connectivity, and conducting manned space missions.

One of SpaceX's key innovations is the reuse of the Falcon 9 rocket's first stage, which has significantly reduced launch costs to approximately $62 million per mission.

In contrast, competitors who do not utilize first-stage reuse face launch expenses that exceed $165 million. By analyzing the likelihood of a successful first-stage landing, we can estimate the potential cost-effectiveness of a launch.

This analysis involves using public data and machine learning models to predict whether SpaceX or its competitors can successfully land and reuse the first stage.



Explore

Impact of Factors on First-Stage Landing Success: Examine how variables such as payload mass, launch site, number of flights, and orbital parameters influence the success rate of first-stage landings.

Trends in Landing Success: Track and analyze the success rates of first-stage landings over time to identify patterns and trends.

Optimal Predictive Model : Determine the most effective predictive model for forecasting the success of first-stage landings using binary classification techniques.

Methodology

Methodology

- **Data Collection** : Gather data using the SpaceX REST API and web scraping methods.
- **Data Preparation** : Clean and preprocess the data by filtering, handling missing values, and applying one-hot encoding to make it suitable for analysis and modeling.
- **Exploratory Data Analysis (EDA)** : Analyze the data through SQL queries and data visualization techniques to uncover insights.
- **Data Visualization** : Create visual representations of the data using tools like Folium and Plotly Dash.
- **Model Building** : Develop and train classification models to predict landing outcomes. Optimize and evaluate the models to identify the best-performing model and parameters.

Data Collection – Using API

Steps

- **Request Data from SpaceX API:** Use the SpaceX API to obtain data related to rocket launches.
- **Decode Response:** Parse the API response using .json() and convert the data into a Pandas dataframe with .json_normalize().
- **Request Launch Information:** Utilize custom functions to retrieve detailed information about the launches from the SpaceX API.
- **Create Dictionary:** Organize the extracted data into a dictionary structure.
- **Create DataFrame:** Convert the dictionary into a Pandas dataframe for further processing.
- **Filter Data:** Refine the dataframe to include only Falcon 9 launches by applying appropriate filters.
- **Handle Missing Values:** Replace missing values in the Payload Mass column with the calculated mean of the available data.
- **Export to CSV:** Save the cleaned and processed data to a CSV file for future analysis or modeling.

Data Collection – Web Scraping

Steps

- **Request Data from Wikipedia:** Fetch Falcon 9 launch data from Wikipedia using an HTTP request.
- **Create BeautifulSoup Object:** Parse the HTML response by creating a BeautifulSoup object.
- **Extract Column Names:** Identify and extract the column names from the HTML table headers.
- **Parse HTML Tables:** Collect the relevant data by parsing the HTML tables within the response.
- **Create Dictionary:** Organize the extracted data into a dictionary for easier manipulation.
- **Create DataFrame:** Convert the dictionary into a Pandas dataframe for structured data handling.
- **Export to CSV:** Save the cleaned and processed data to a CSV file for further analysis or use.

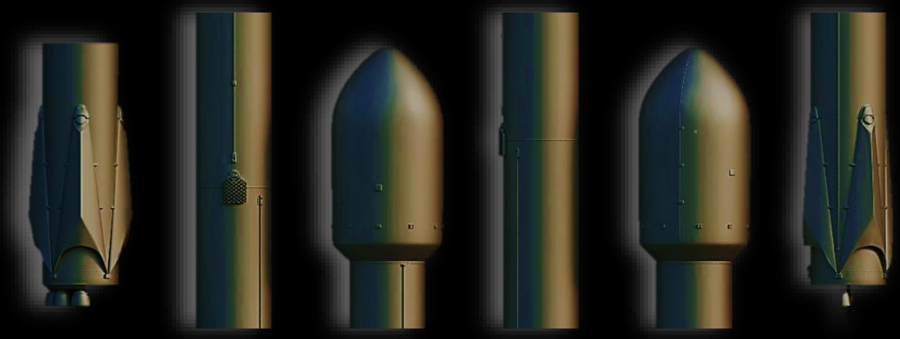
Data Wrangling

Steps

- **Perform EDA and Determine Data Labels:** Conduct Exploratory Data Analysis (EDA) to understand the data and identify relevant labels.
- **Calculate Metrics:**
- 1) Calculate the number of launches per site.
  2) Determine the frequency and occurrence of each orbit type.
  3) Analyze the occurrence of mission outcomes based on orbit type.
- **Create Binary Landing Outcome Column:** Add a column representing the landing outcome as a binary variable (successful or unsuccessful).
- **Export to CSV:** Save the processed data, including all calculated metrics and the binary outcome column, to a CSV file for further analysis.

Launch Sites

1) **CCAFS SLC 40** (Cape Canaveral Air Force Station, Space Launch Complex 40)
2) **CCAFS LC 40** (Cape Canaveral Air Force Station, Launch Complex 40)
3) **KSC LC 39A** (Kennedy Space Center, Launch Complex 39A)
4) **VAFB SLC 4E** (Vandenberg Air Force Base, Space Launch Complex 4E)



Landing Outcome

Outcome	Definition
True Ocean	Successfully landed in a specific region of the ocean.
False Ocean	Unsuccessfully landed in a specific region of the ocean.
True RTLS	Successfully landed on a ground pad near the launch site.
False RTLS	Unsuccessfully landed on a ground pad near the launch site.
True ASDS	Successfully landed on a drone ship in the ocean.
False ASDS	Unsuccessfully landed on a drone ship in the ocean.
None ASDS	Failure to land on the drone ship.
None None	General failure to land, with no attempt to land on either a drone ship or a ground pad.

Final Predictor Column

Outcomes converted into 1 for a successful landing and 0 for an unsuccessful landing

Tools used : pandas

EDA with Visualization & Feature Engg.

Charts –

Understanding Relationship between :

- **Flight Number** vs. **Payload Mass**
- **Flight Number** vs. **Launch Site**
- **Payload Mass (kg)** vs. **Launch Site**
- **Payload Mass (kg)** vs. **Orbit Type**
- **Flight Number** vs. **Orbit Type**
- **Year** vs. **Success Rate**

Analysis

- **Success Rate of Each Orbit Type** : Use bar charts to compare distinct categories and illustrate their relationships with corresponding values. For tracking trends over time, use a line chart to identify the success of landings over last 10 years.
- **Feature Engineering** : Use **scatter plots** to investigate potential relationships between variables, as identifying correlations can be valuable for machine learning. Additionally, convert categorical columns to dummy variables to facilitate machine learning calculations.

Tools used : pandas matplotlib

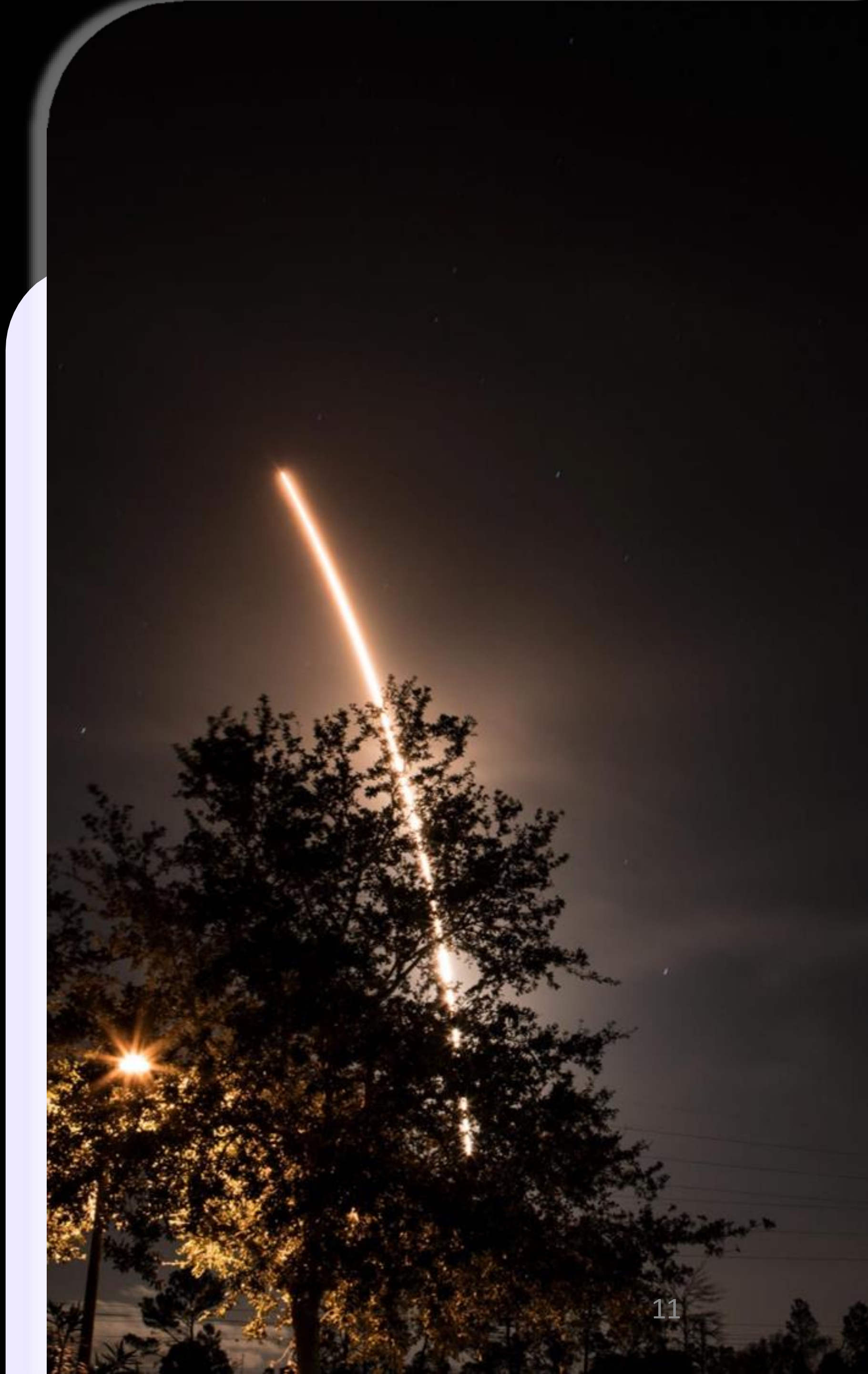
EDA with SQL Queries

Display

- Unique launch site names
- Five records where the launch site starts with 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by the F9 v1.1 booster version

List

- Date of the first successful landing on a ground pad
- Names of boosters that successfully landed on a drone ship and carried a payload mass between 4,000 and 6,000
- Total number of successful and failed missions
- Names of booster versions with the highest payload capacity
- Failed landing outcomes on drone ships, including the booster version and launch site, for the year 2015
- Count of landing outcomes from 2010-06-04 to 2017-03-20 (in descending order) with Ranking



Tools used :

Map with Folium

Mark all Launch Sites on map

- Added **blue circles** at NASA Johnson Space Center's coordinate with a
- popup label showing its name using its latitude and longitude coordinates

- Added **orange circles** at all launch sites coordinates with a popup label
- showing its name using its name using its latitude and longitude coordinates

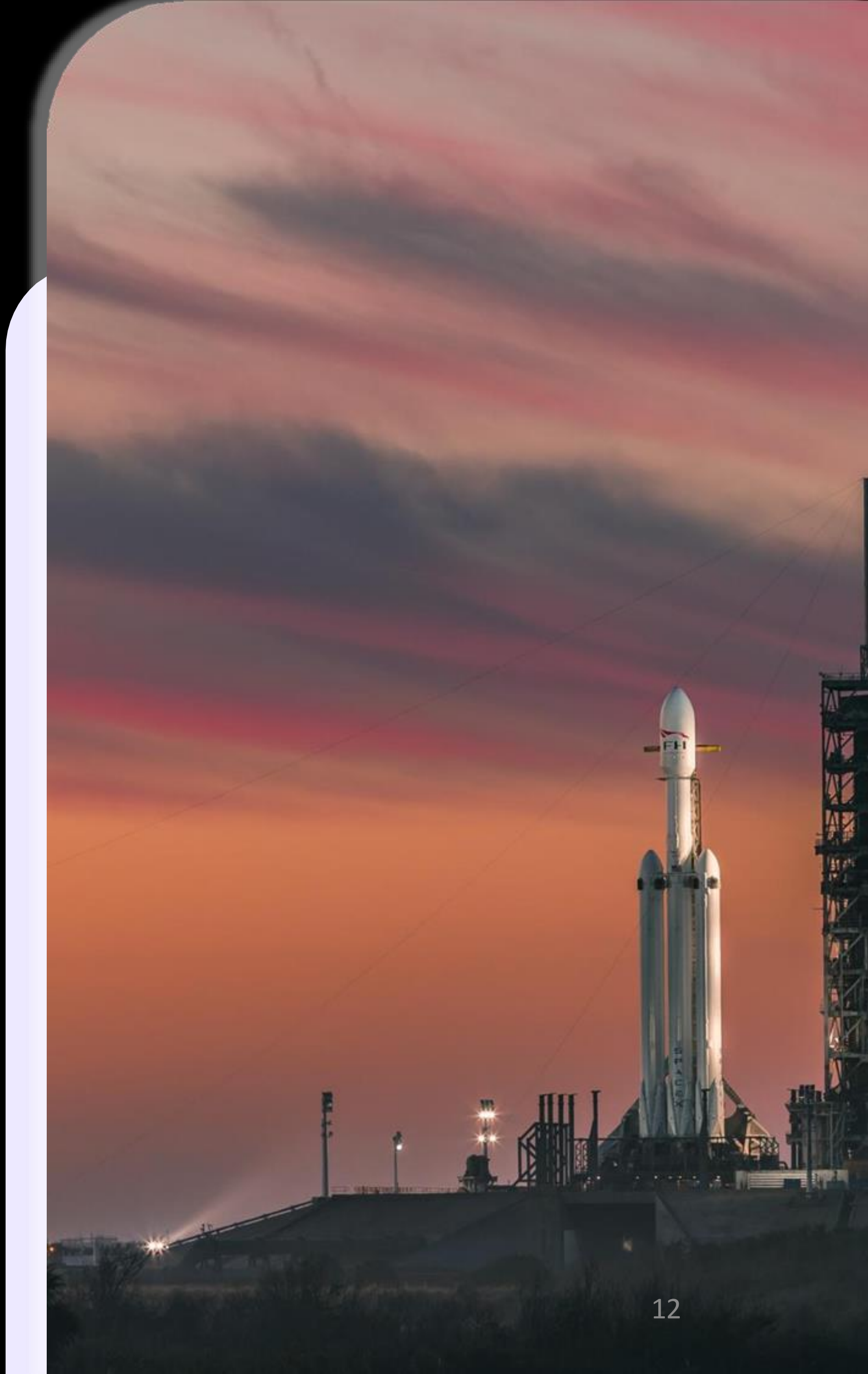
Mark the successful and failed launches for each sites on the map

- Applied colored markers to each launch site to represent launch outcomes: **green** for successful and **red** for unsuccessful, illustrating which sites exhibit higher success rates.

Calculate the Distances between a Launch Site to its Proximities

- Added colored lines to show distance between launch site **CCAFS SLC-40** (Cape Canaveral Air Force Station, Space Launch Complex 40) and its proximity to the nearest coastline, railway, highway, and city

Tools used : pandas Folium



Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

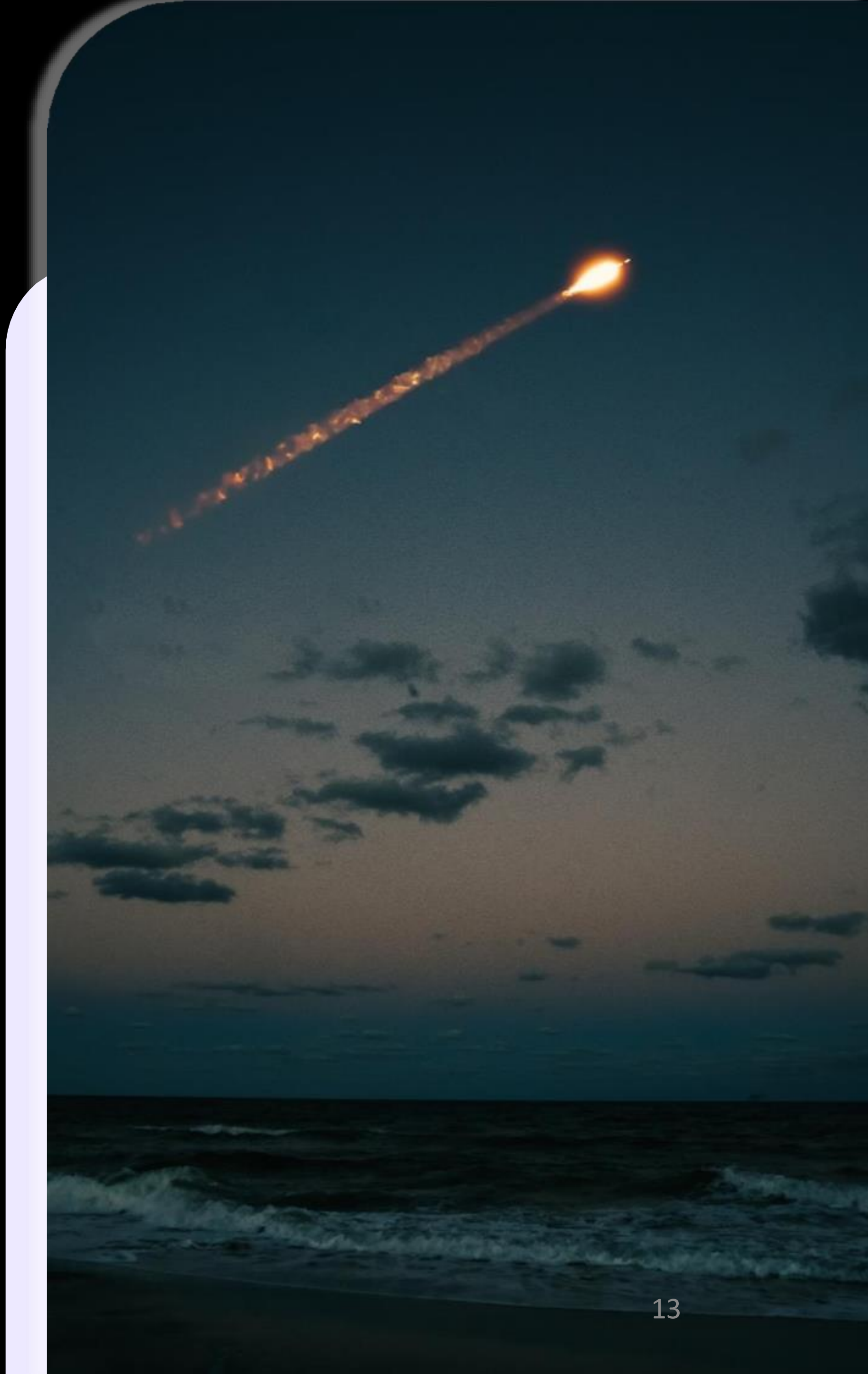
- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Plot Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

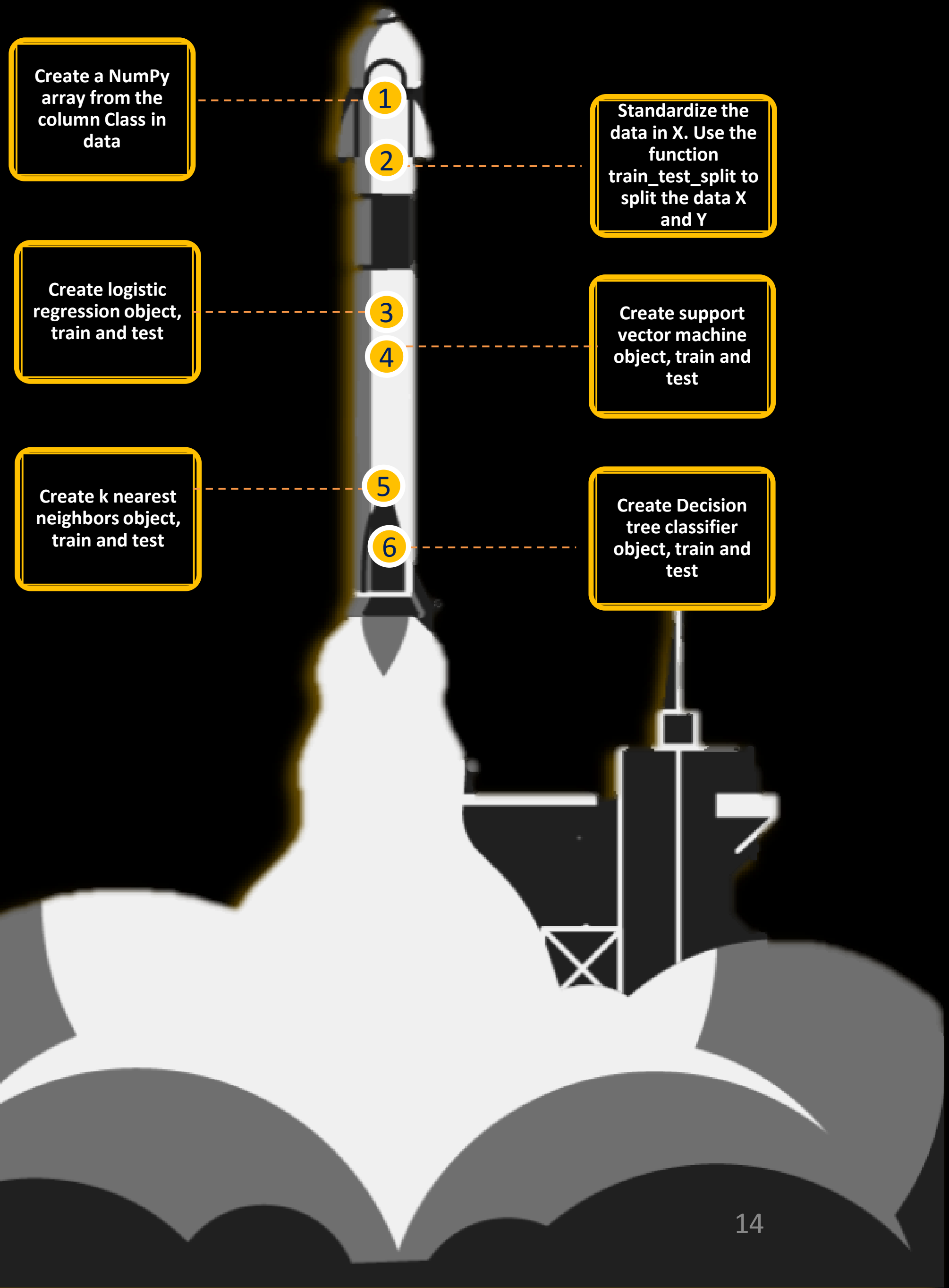


Tools used : plotly

Predictive Analytics

Steps:

- Create **NumPy** array from the Class column.
- Standardize the data using **StandardScaler**, fitting and transforming it.
- Split the data using **train_test_split** with train size 80% from overall data.
- Set up a **GridSearchCV** object with cv=10 for parameter optimization.
- Apply GridSearchCV to different algorithms: **Logistic Regression , Support Vector Machine, Decision Tree,** and **K-Nearest Neighbor**.
- Compute accuracy on the test data using the .score() method for each model.
- Evaluate the **confusion matrix** for each model.
- Determine the best model based on **Jaccard Score, F1 Score,** and **Accuracy**.



Models used :
KNN, Logistic Regression, SVM, Decision Tree with **Hyperparameter Tuning** using **GridSearchCV** object for best parameters for calculations.

Tools used : pandas matplotlib scikit learn NumPy

Results

Results Summary

Exploratory Data Analysis

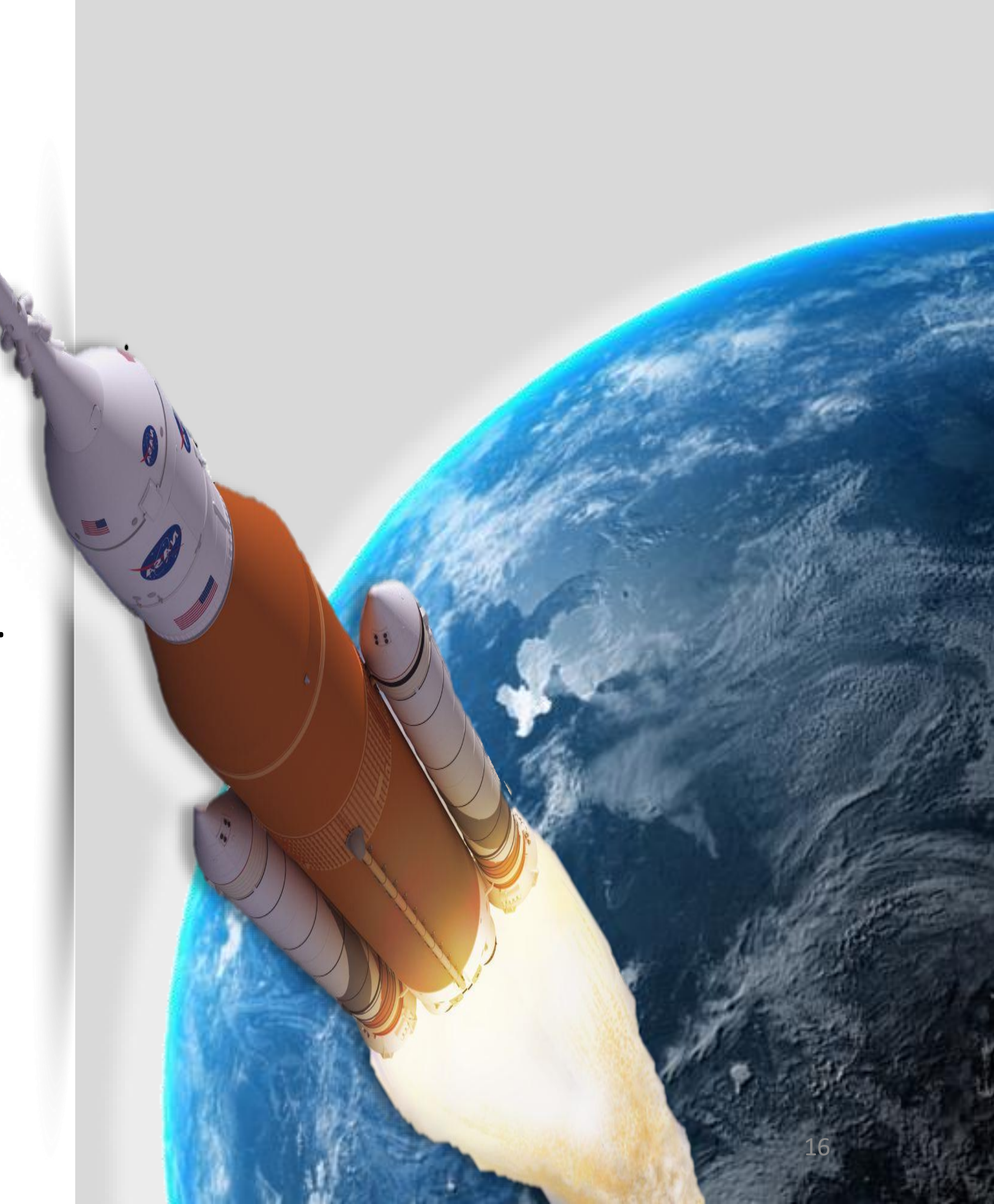
- Launch success rates have improved over time from 2013.
- KSC LC-39A boasts the highest success rate among all landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.
- Heavy payloads have higher successful landing rates for Polar, LEO, and ISS orbits.
- The number of successful landings has risen significantly since 2013.

Visualisation

- 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.
- KSC LC-39A has a 77% success rate for all landings and a 42% success rate compared to all landing sites.
- When the range is from 0-10,000kg Booster version v1.1 has the highest success. However, when the payload range is from 2500-7500kg booster version FT has the highest success.

Predictive Analysis

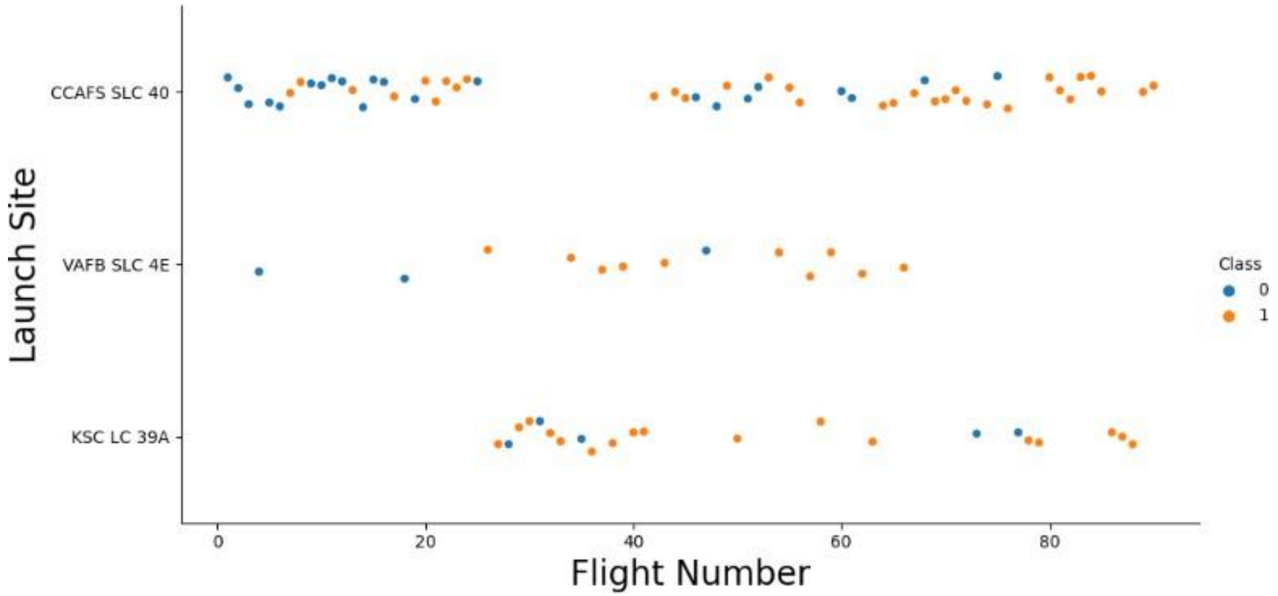
- Decision Tree model is the best predictive model for the dataset



Flight Number vs. Launch Site

Exploratory Data Analysis

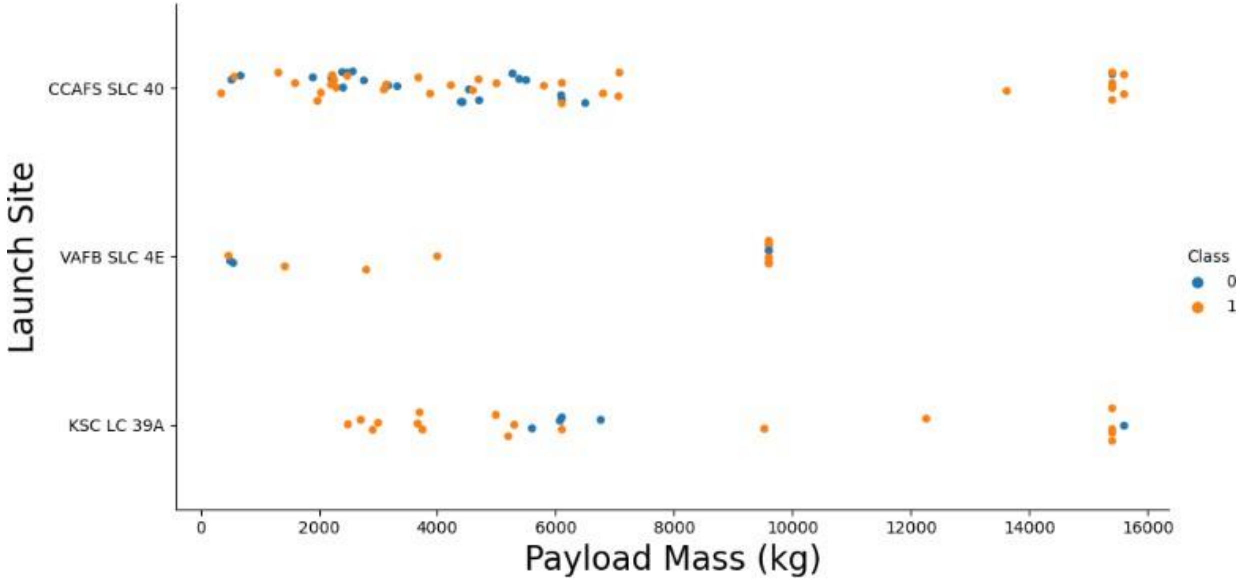
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

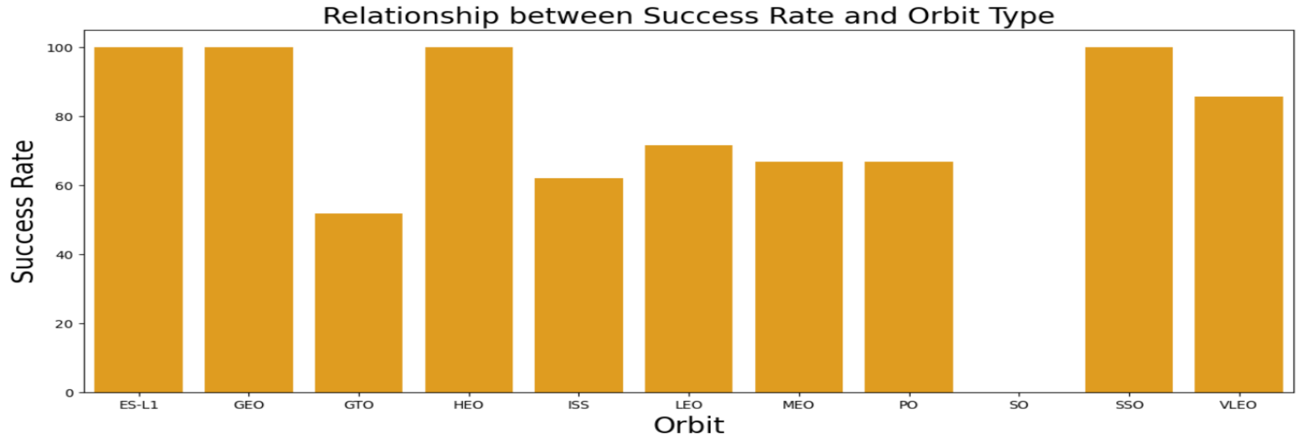
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate by Orbit

Exploratory Data Analysis

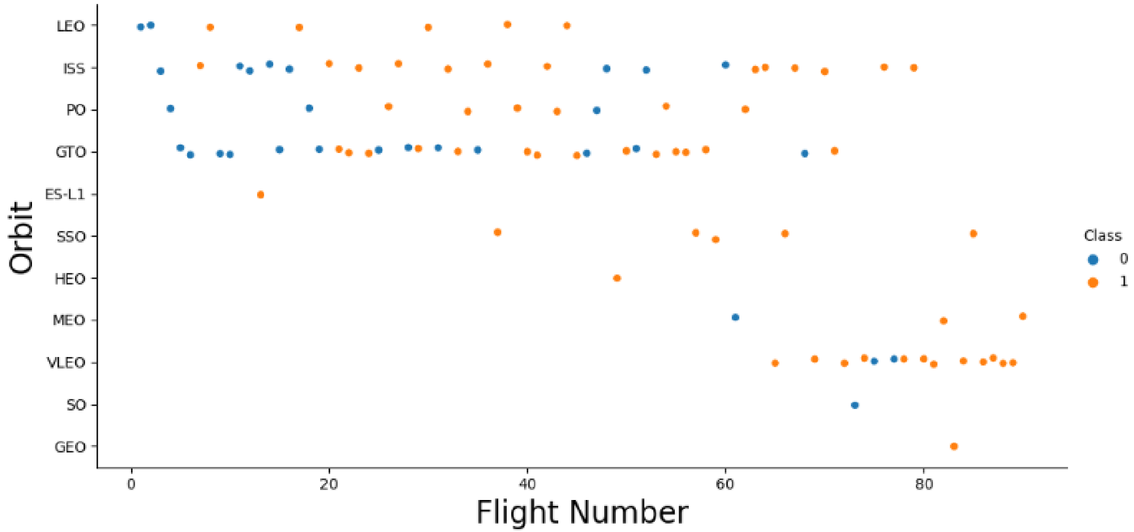
- **100% Success Rate: ES-L1, GEO, HEO and SSO**
- **50%-80% Success Rate: GTO, ISS, LEO, MEO, PO**
- **0% Success Rate: SO**



Flight Number vs. Orbit

Exploratory Data Analysis

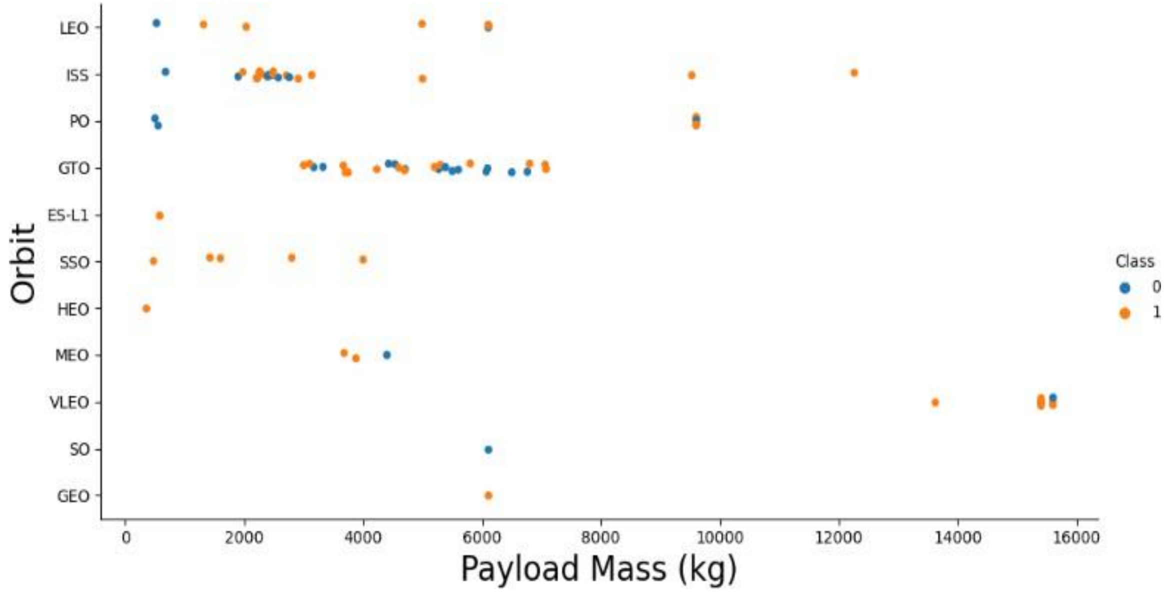
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit

Exploratory Data Analysis

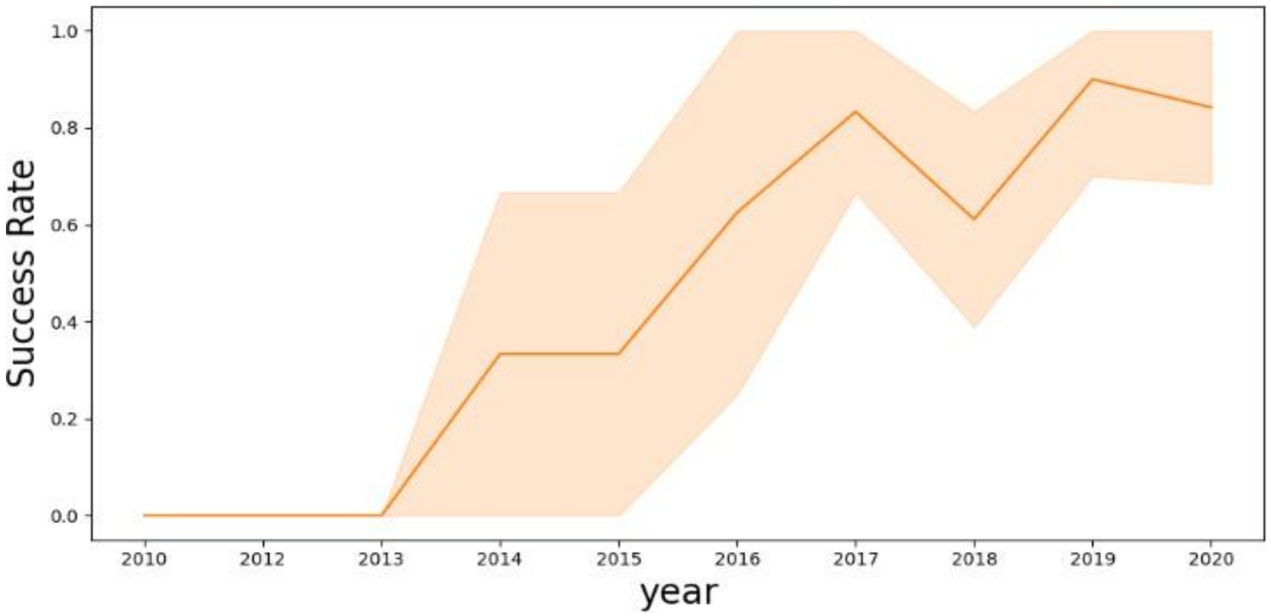
- **Heavy payloads are better with LEO, ISS and PO orbits**
- **The GTO orbit has mixed success with heavier payloads**



Launch Success over Time

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.



Launch Site Information

1) Launch Sites 's Names – Using SQL query

```
#by Magic Command
%sql select DISTINCT(Launch_Site) from SPACEXTABLE
```

* sqlite:///my_data1.db
Done.

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

2) Records with Launch Site Starting with CCA – Using SQL query

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

Python

* sqlite:///my_data1.db
Done.

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

Payload Mass

3) Total Payload Mass carried by boosters launched by NASA (CRS) – Using SQL query

```
1  %sql select sum(PAYLOAD_MASS__KG_) as 'Total payload mass' from SPACEXTABLE where Customer =  'NASA (CRS)'
```

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

Total payload mass
45596

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

4) Average Payload Mass carried by booster version F9 v1.1– Using SQL query

```
1  %sql select round(avg(PAYLOAD_MASS__KG_),2) as 'Average payload mass' from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

✓ 0.0s

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

Average payload mass
2928.4

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

Landing & Mission Information

5) 1st Successful Landing 's Date in Ground Pad – Using SQL query

```
%sql select min(Date) as first_successful_landing_date from SPACEXTABLE where Landing_Outcome like "Success (ground pad)"
```

* sqlite:///my_data1.db
Done.

first_successful_landing_date
2015-12-22

- **22nd December 2015 was first successful landing in ground pad**

6) Booster Drone Ship Landing– Using SQL query

```
%%sql select Booster_Version from SPACEXTABLE
where Landing_Outcome like "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

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

7) Total Number of Successful and Failed Mission Outcomes

```
%%sql
SELECT
    COUNT(CASE WHEN Mission_Outcome like '%Success%' THEN 1 END) AS Successful,
    COUNT(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 END) AS Failed
FROM SPACEXTABLE;
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

Successful	Failed
100	1

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

Boosters

8) Booster's version which carried the maximum payload mass – Using SQL query

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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

Booster Version with Maximum Payload Mass

- **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**

Failed Landings on Drone Ship

9) Showing month, date, booster version, launch site and landing outcome in 2015– Using SQL query

```
%%sql select case substr(Date,6,2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as "Month_Name",
    Date,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTABLE
    where substr(Date,1,4) = '2015' AND landing_outcome LIKE '%Failure (drone ship)%'
```

✓ 0.0s

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

Month_Name	Date	Landing_Outcome	Booster_Version	Launch_Site
January	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Count of Successful Landings

10) Count of landing outcomes between 4th June 2010 and 20th March 2017 in descending order – Using SQL query

```sql
%%sql WITH LandingCounts AS (
    SELECT landing_outcome,
           COUNT(*) AS outcome_count
    FROM SPACEXTABLE
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY landing_outcome
)
SELECT landing_outcome,
       outcome_count,
       DENSE_RANK() OVER (ORDER BY outcome_count DESC) AS rank
FROM LandingCounts
ORDER BY rank;
```

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

landing_outcome	outcome_count	rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	3
Success (ground pad)	3	3
Failure (parachute)	2	4
Uncontrolled (ocean)	2	4
Precluded (drone ship)	1	5

Launch Site Analysis

Launch Sites Map

Launch sites **near the equator** offer significant advantages for space missions, especially when aiming for equatorial orbits.

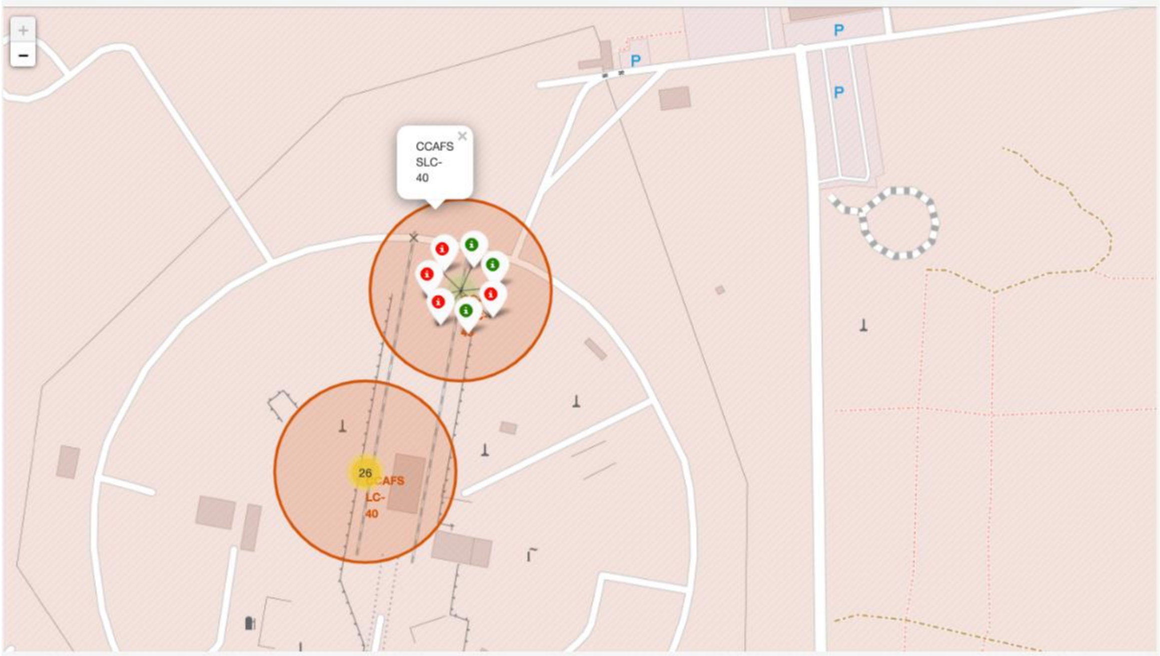
The Earth's rotation provides a **natural boost** to rockets launched from these locations, making it easier to achieve a prograde orbit and reducing the need for additional fuel and boosters.

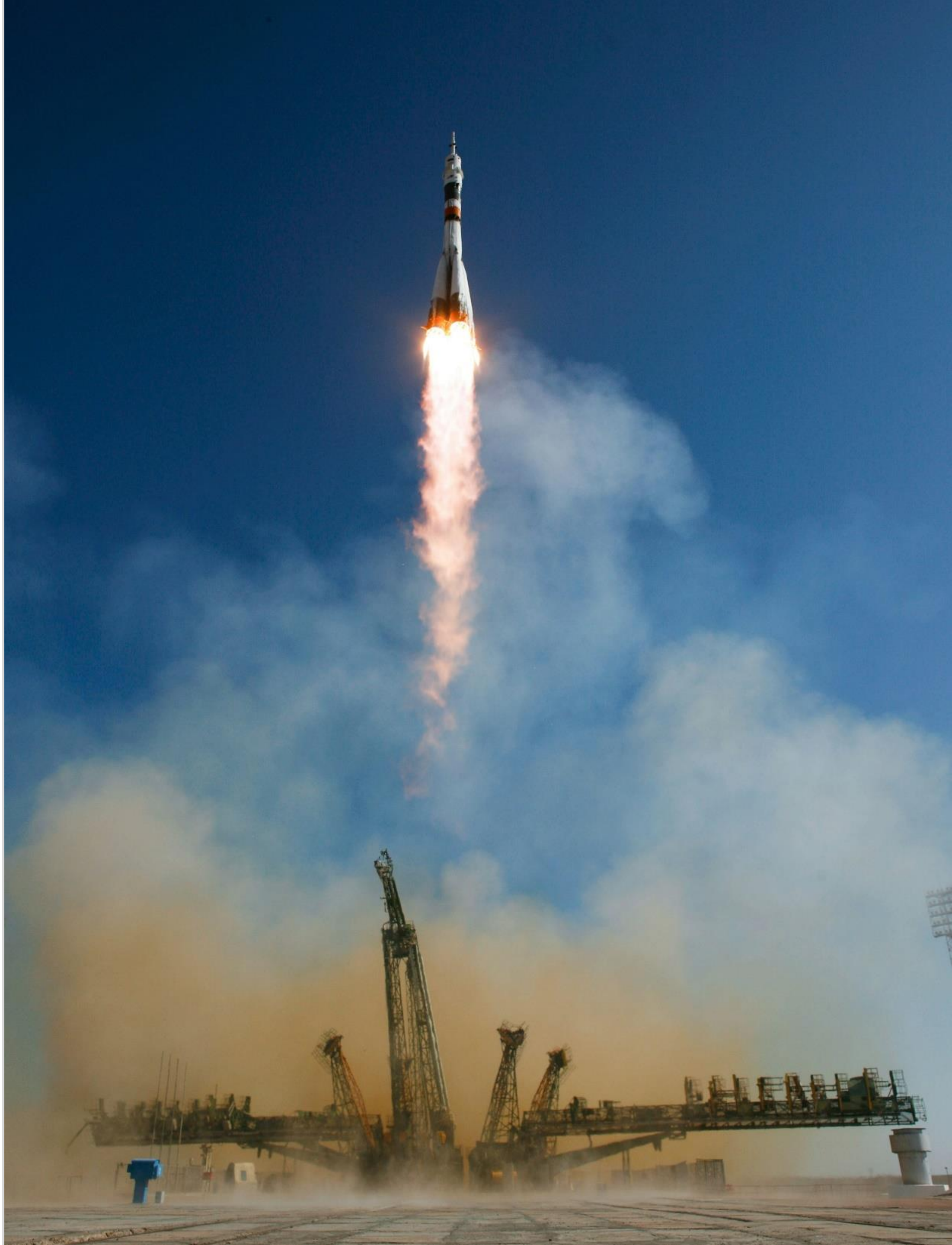
Key coastal launch sites that benefit from this include **CCAFS LC, CCAFS SLC, KSC LC-39A**, and **VAFB SLC-4E**.

Launch Outcomes

Launch site CCAFS SLC-40 has a success rate of 43%, with 3 successful launches and 4 unsuccessful ones.

Successful launches are marked with green markers, while unsuccessful ones are indicated with red markers.



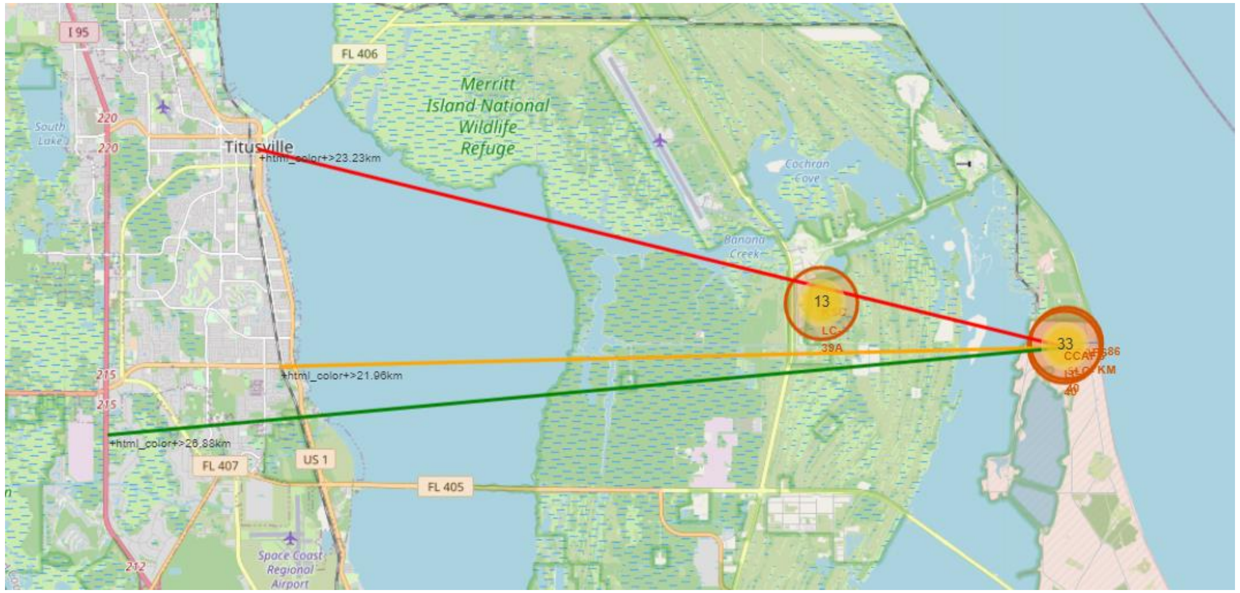


Distance to Proximities

Coasts: Coastal locations ensure that spent rocket stages or debris from failed launches fall safely into the ocean, minimizing the risk to people and property.

Safety & Security: An exclusion zone around the launch site is essential to keep unauthorized individuals at a safe distance and protect public safety.

Transportation, Infrastructure and Cities: The launch site should be sufficiently distant from urban areas and infrastructure to reduce potential damage from a failed launch, while remaining close enough to roads, railways, and docks for efficient transportation of personnel and materials.

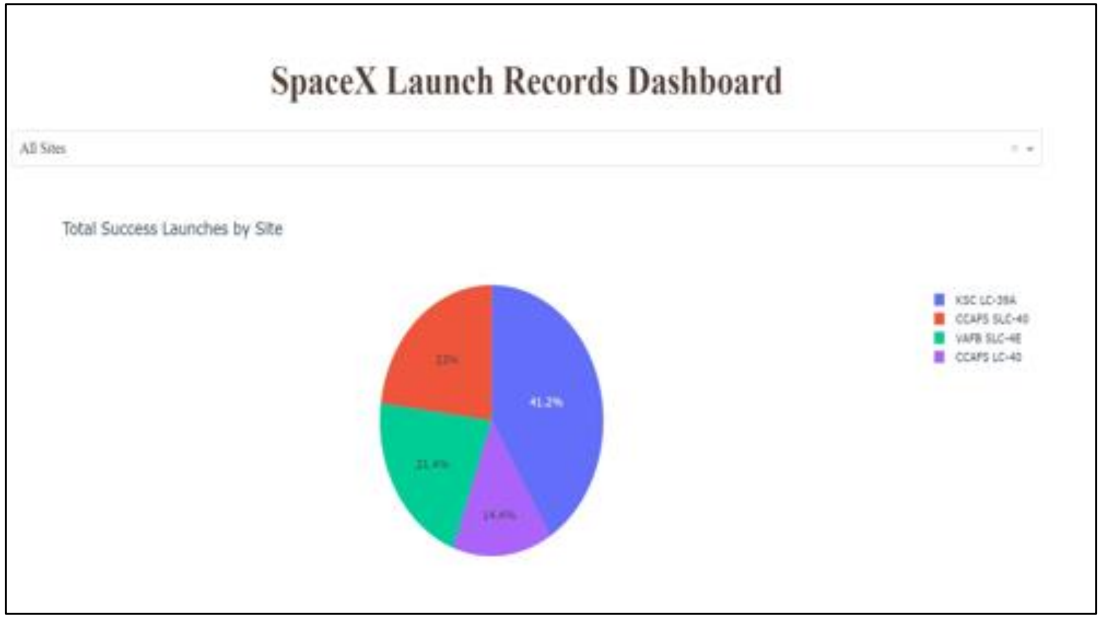


CCAFS SLC-40 is located :-
0.86 km from the nearest **coastline**,
21.96 km from the nearest **railway**,
23.23 km from the nearest **city**,
26.88 km from the nearest **highway**.

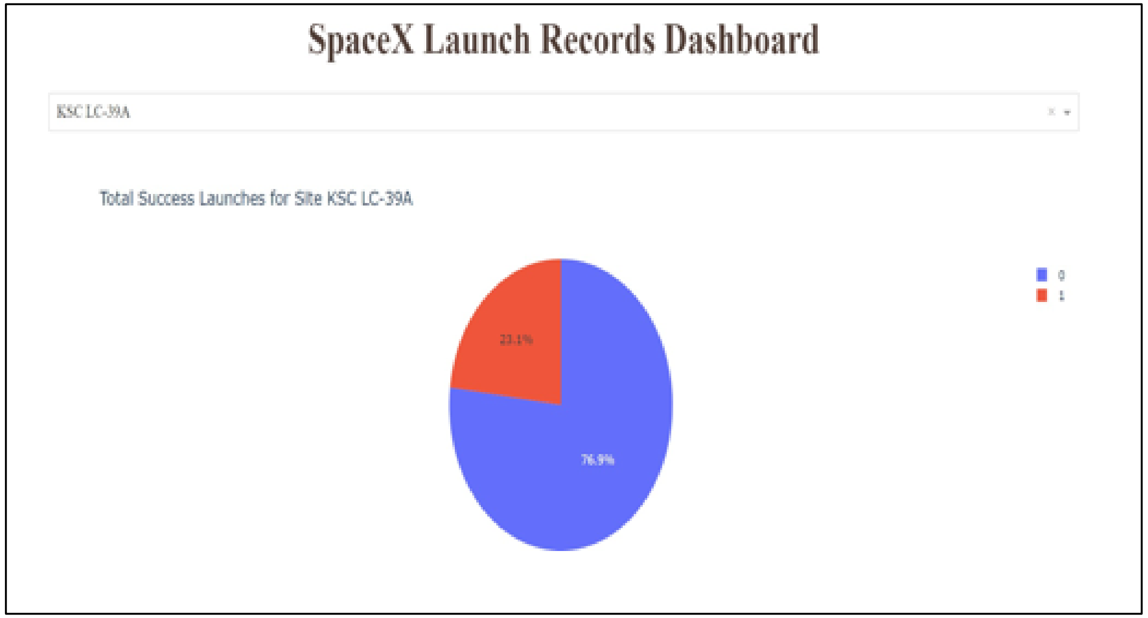
Dashboard with Plotly

Site Launch Success Rate

- KSC LC-39A has the most successful launches amongst launch sites (41.2%) among all 4 landing sites
- KSC LC-39A boasts the highest success rate among launch sites at 76.9%, with 10 successful launches and 3 failed launches.



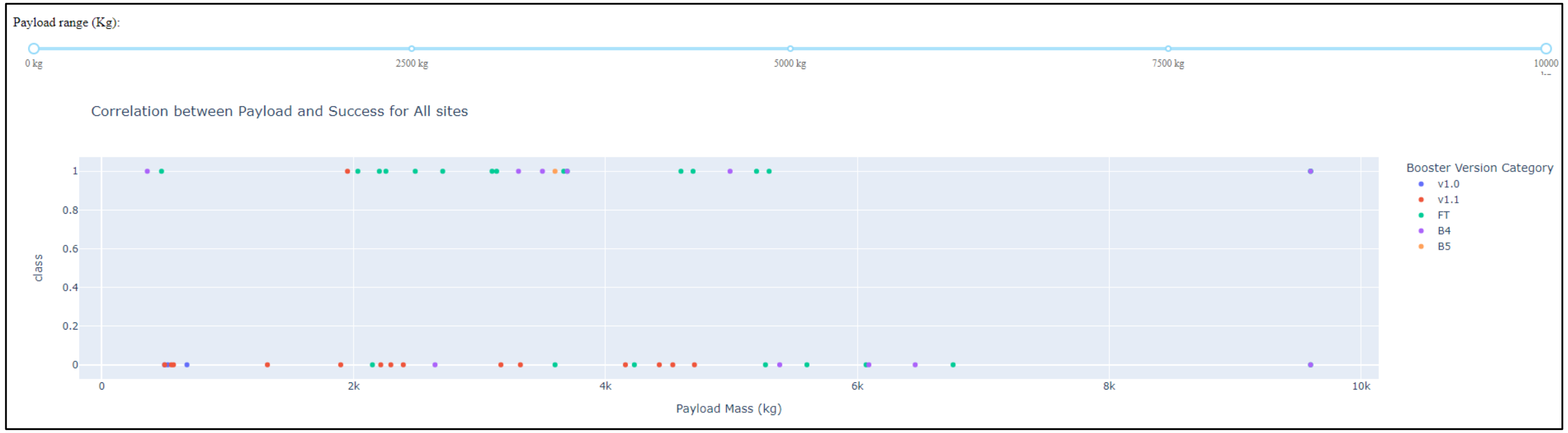
Among all four sites



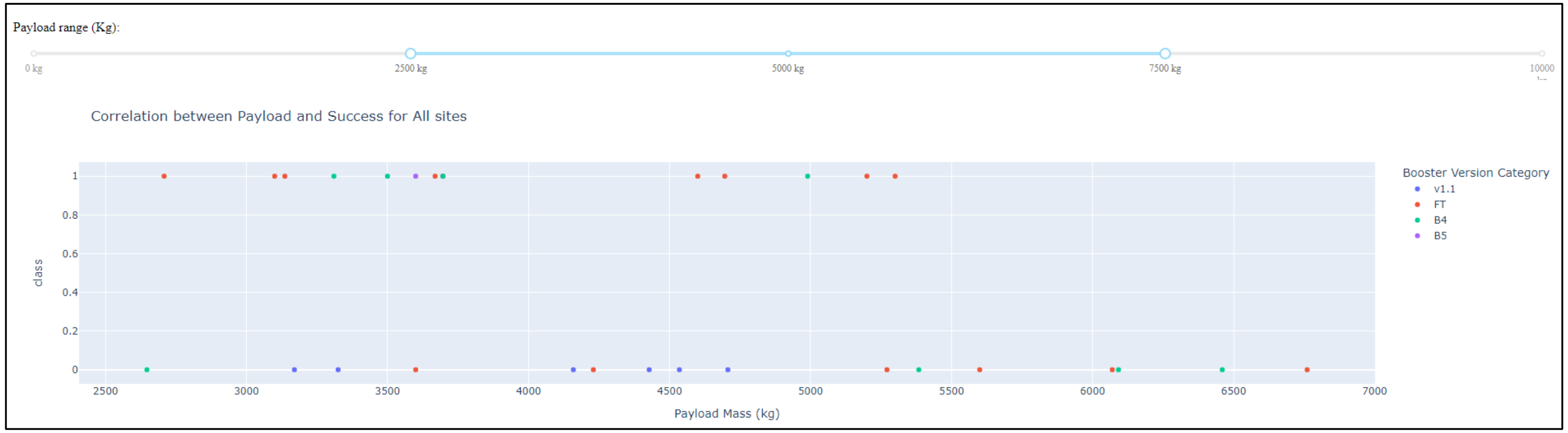
For KSC LC – 39 A

Payload Mass and Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- When the range is from 0 - 10,000 kg Booster version v1.1 has the highest success.
- When the payload range is from 2,500 - 7,500 kg booster version FT has the highest success.



0 - 10,000 kg Payload mass range



2,500 - 7,500 kg Payload mass range

Predictive Analysis

Classification

- All the models performed similarly, with comparable scores and accuracy, likely due to the small dataset.
- However, the **Decision Tree model** slightly outperformed the others when considering the **.best_score_**, which represents the average score across all cross-validation folds for a specific combination of parameters.

Confusion Matrix Outputs:

12 **True Positive**, 3 **True Negative**
3 **False Positive**, 0 **False Negative**

- **Precision =** TP / (TP + FP) = 87%
- **Recall =** TP / (TP + FN) = 83%
- **F1 Score =** 2 * (Precision * Recall) / (Precision + Recall) = 81%
- **Accuracy =** (TP + TN) / (TP + TN + FP + FN) = 83%

Logistic Regression

Best Parameters :-

- **c – 0.01**
- **penalty – L2**
- **solver – 'lbgfs'**

Best Score Accuracy : 84%
F1 Score : 81%
Recall : 83%, Precision: 87%
Accuracy : 83%

Support Vector Machine

Best Parameters :-

- **c – 1**
- **gamma – 0.0316**
- **kernel – 'sigmoid'**

Best Score Accuracy : 84%
F1 Score : 81%
Recall : 83%, Precision: 87%
Accuracy : 83%

Decision Tree

Best Parameters :-

- **criterion – 'entropy'**
- **max_depth – 18**
- **max_features – 'sqrt'**
- **min_samples_leaf – 4**
- **min_samples_split – 10**
- **splitter – 'random'**

Best Score Accuracy : 88%
F1 Score : 81%
Recall : 83% , Precision: 87%
Accuracy : 83%

KNN Classifier

Best Parameters :-

- **algorithm – 'auto'**
- **n_neighbors – 10**
- **p – 1**

Best Score Accuracy : 84%
F1 Score : 81%
Recall : 83% , Precision: 87%
Accuracy : 83%

Conclusion

Conclusion

- **Payload Mass** : The analysis reveals that payload mass is the most crucial factor influencing the success rate of launches. A higher payload mass is associated with a higher success rate across all launch sites.

- **Booster Selection** : Understanding the payload allows for the selection of the most suitable booster version, which in turn maximizes the likelihood of a successful launch.

- **Cost Efficiency** : A higher success rate leads to cost savings for SpaceX. The ability to reuse the first stage of the rocket reduces overall launch costs.

- **Launch Success Trends** : The success rate of launches has improved over time, highlighting advancements and increased reliability in launch operations.

- **Equatorial Advantage** : Launch sites are predominantly near the equator to benefit from the Earth's rotational speed, which provides a natural boost and minimizes the need for extra fuel and boosters.

- **Coastal Locations** : All launch sites are positioned close to the coast to ensure that spent stages or debris fall into the ocean, reducing risk to people and property.

- **KSC LC-39A Performance** : KSC LC-39A stands out with the highest success rate among launch sites and achieves a 100% success rate for payloads under 5,500 kg.

- **Orbit Success Rates** : Orbits such as ES-L1, GEO, HEO, and SSO have consistently achieved a 100% success rate.

Thank You

Access all analysis and visualization notebooks here: GitHub link - MandeepS123/Data_Science_IBM