Applied Data Science

Capstone Project

Arfath Ahmed Syed

20 Apr 2024

Overview

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

Executive Summary

Methodologies involved:

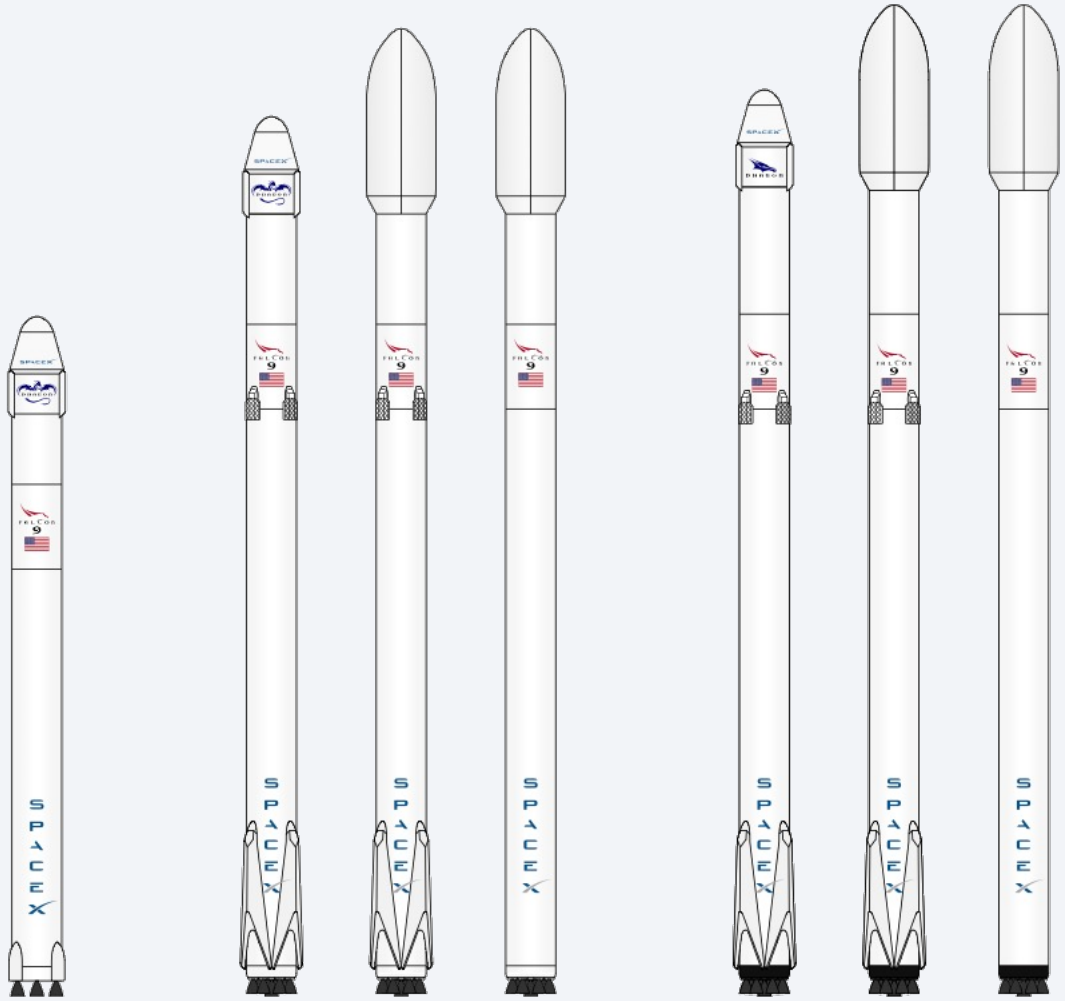
- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Data Visualisation (Interactive)
- Predictive Analysis

Results Generated:

- EDA Visualisations
- Analysis of Geospatial Features
- Interactive Dash Dashboard
- Model Experimentation
- Model Comparison (Predictive Analysis)

Introduction

- Falcon 9 rockets are launched by SpaceX. They cost significantly less that what is cost by the competitors. This cost reduction is primarily due to the capability of recovering and reusing the rocket's first stage.
- The objective of this study is to reliably forecast whether the first stage of the Falcon 9 rocket will successfully land so that SpaceX can accurately project launch costs and losses and stay a step ahead of the competitors



Methodology

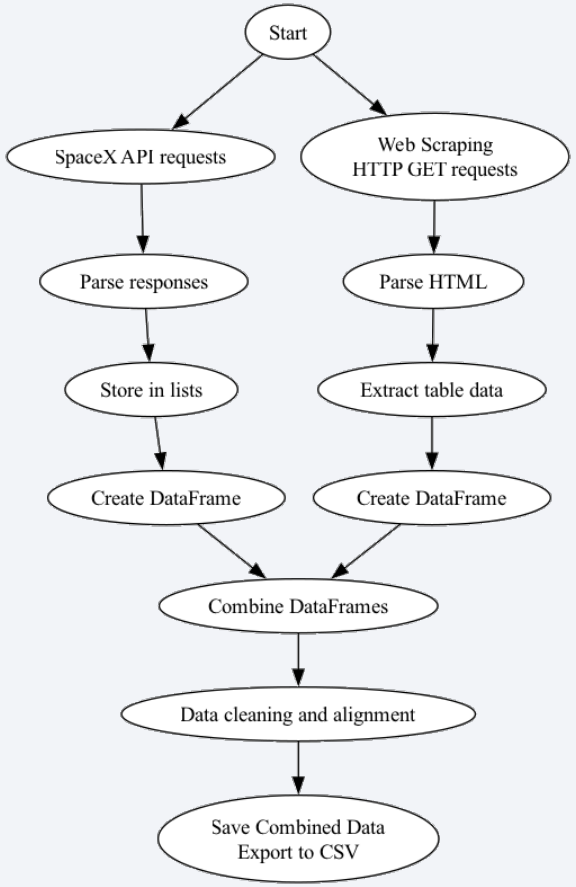
Section 1

Methodology

- Data collection methodology:
  - Web Scraping With BeautifulSoup
  - SpaceX API Request with Requests
- Perform data wrangling
  - Data Imputation of Missing Values
  - Examine Value Counts of important features.
  - New landing_outcome variable created
- Exploratory Data Analysis
  - SQL Queries to examine and manipulate data
  - Descriptive analysis with Pandas & Matplotlib
- Interactive Visualisations
  - Folium GeoSpatial & Cloropleth Maps
  - Dashboard with Dash
- Data Modelling and Evaluation
  - Preprocessing & Standardising Data
  - Train Test Split
  - Train & Evaluate multiple models
  - Hyperparameter Tuning - GridSearchCV

Data Collection – SpaceX API & Web Scraping

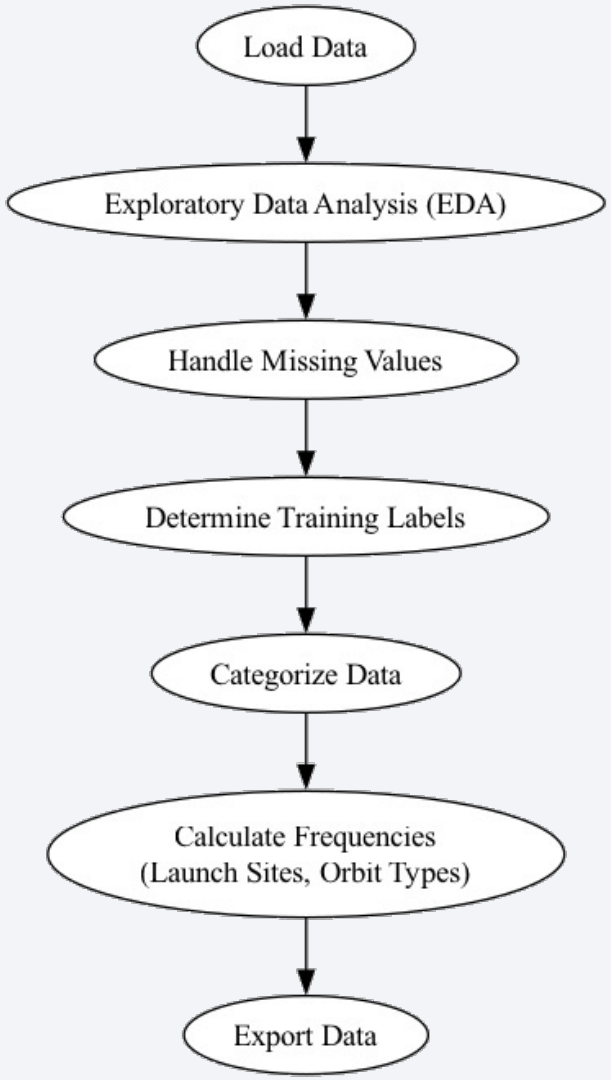
- Data was systematically gathered from the SpaceX API and Web Scraping the Wikipedia Page of Falcon 9 Launches. By doing this we captured detailed launch records including payload, orbit, landing outcomes, Launch sites and Boost versions etc.



- GITHUB LINK – SpaceX API
- GITHUB LINK – Web Scraping

Data Wrangling

- Import the dataset from a previous collection step.
- Convert landing outcomes into binary labels for supervised learning.
- Classify each mission as successful or unsuccessful based on landing data and create a binary classification target variable.
- SpaceX has many launch locations under it. All of these are in the data collected in earlier steps. Calculate Launch site and orbit type frequencies.
- Save the processed data to a CSV file for use in further analysis.
- GITHUB LINK – Data Wrangling



EDA with Data Visualization

Scatter Plots

Flight Number vs. Payload Mass: Used to observe how increases in flight number (indicative of experience and technological advances) and payload mass affect the likelihood of successful landings.

Flight Number vs. Launch Site: Helped to examine if there were correlations between the number of flights and their respective launch sites with successful landings.

Payload Mass vs. Launch Site: Illustrated the distribution of payload masses across different launch sites and how these might correlate with landing success rates.

Flight Number vs. Orbit Type and Payload Mass vs. Orbit Type: These scatter plots were crucial in examining the relationship between the mission's orbit type and both the flight number and payload mass, providing insights into how these factors correlate with success rates in specific orbits.

Bar Charts

Success Rate by Orbit Type: This chart was used to visualize the success rates associated with each orbit type, helping to pinpoint which orbits typically have higher success rates. This is crucial for determining risk levels and expected outcomes based on orbit selection.

Line Charts

Launch Success Yearly Trend: Displayed the trend of launch successes over the years, helping to visualize improvements or declines in success rates over time. This can be particularly useful for spotting trends in the data related to changes in technology or operational practices.

GITHUB LINK – EDA with Visualization

EDA with SQL

Display Unique Launch Sites:
Identify & list all unique launch sites involved in the missions

Display Records for Specific Launch Sites:
Helping to focus analysis on specific geographic locations.

Total Payload Mass by a NASA CRS:
Crucial for understanding the volume of material sent to space.

Average Payload Mass by Booster Version (F9 v1.1):
Insights into the capabilities of different booster iterations.

Date of the First Successful Ground Pad Landing:
1st successful marking a milestone in reusable rocket technology.

List Boosters with Successful Drone Ship Landings and Specific Payload Mass:
Useful for assessing booster performance under certain conditions.

Count Mission Outcomes:
Clear quantitative metric of mission reliability.

List Booster Versions Carrying Maximum Payload Mass:
Highlighting the most capable booster configurations.

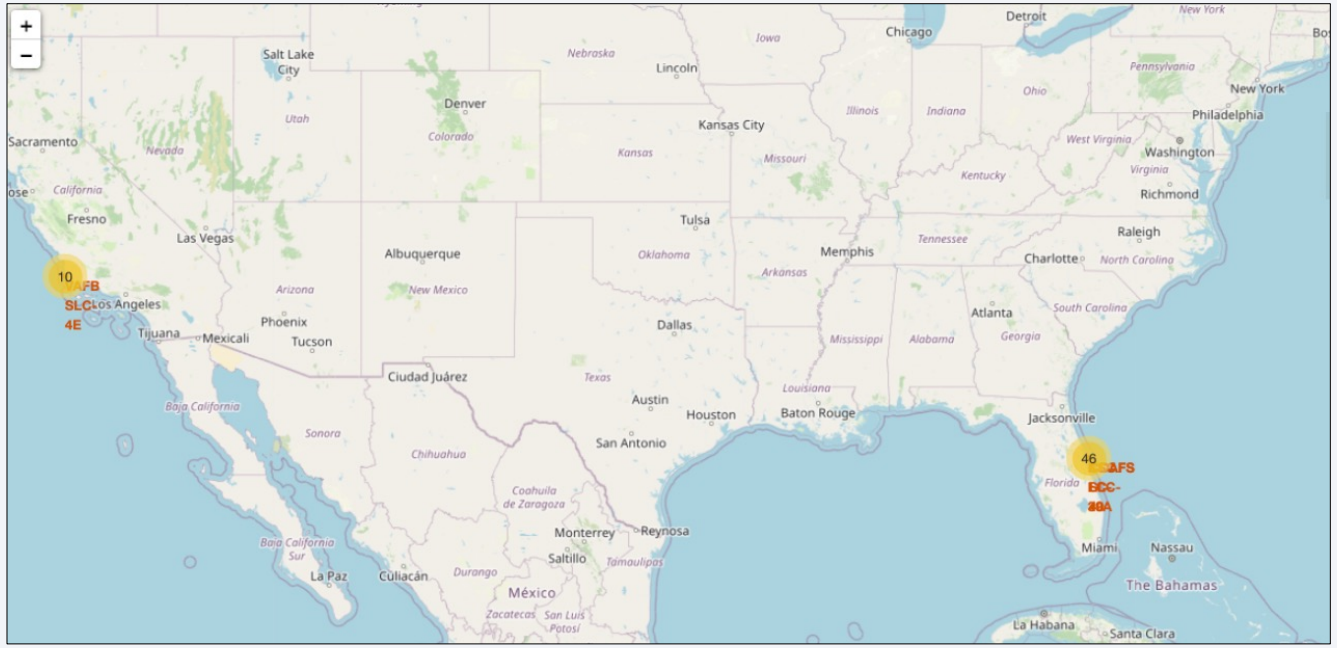
Query for Specific Mission Outcomes over Time:
Helps to evaluate operational trends and effectiveness

GITHUB LINK – EDA with SQL

Build an Interactive Map with Folium

Map Objects added:

- Circles on map: Indicate the location of launch sites.
- Color-coded markers: Show success (green) or failure (red) of each launch.
- Marker clusters: Group markers at the same location to declutter the map.
- Distance lines: Draw lines from launch sites to nearby coastlines or points of interest.
- Polylines: Distances from launch sites to important locations like coastlines, cities, etc.



Map objects such as markers, circles, and lines were added to a folium map to identify launch sites, visualize launch outcomes, and measure distances to key geographical features. This helps in assessing how location influences launch success and operational logistics.

GITHUB LINK - Interactive Map with Folium

Build a Dashboard with Plotly Dash

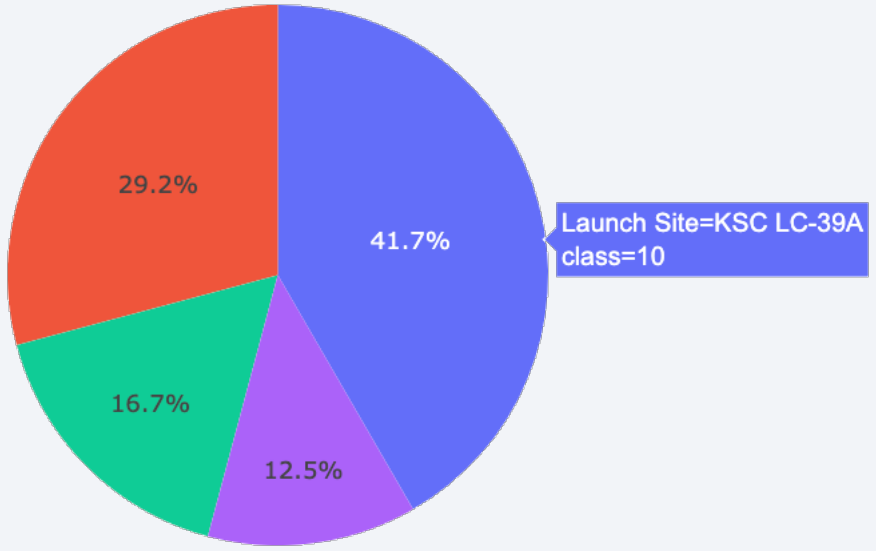
The following plots were included to be a part of the Plotly Dash Dashboard to make the user experience and interface more interactive and intuitive:

Pie Chart

Could be filtered based on dropdown selection

Proportions of total successful launches per site

Makes it clear to see which sites are most successful



Scatter Plot

Could be filtered using the slider provided

To show the correlation between outcome and payload mass (kg)

GITHUB LINK – Plotly Dash Interactive Dashboard

Predictive Analysis (Classification)

Data Loading and Splitting:

Loaded the dataset containing SpaceX Falcon 9 first stage landing data. Separate Feature and Target labels and then into train and test splits.

Standardization:

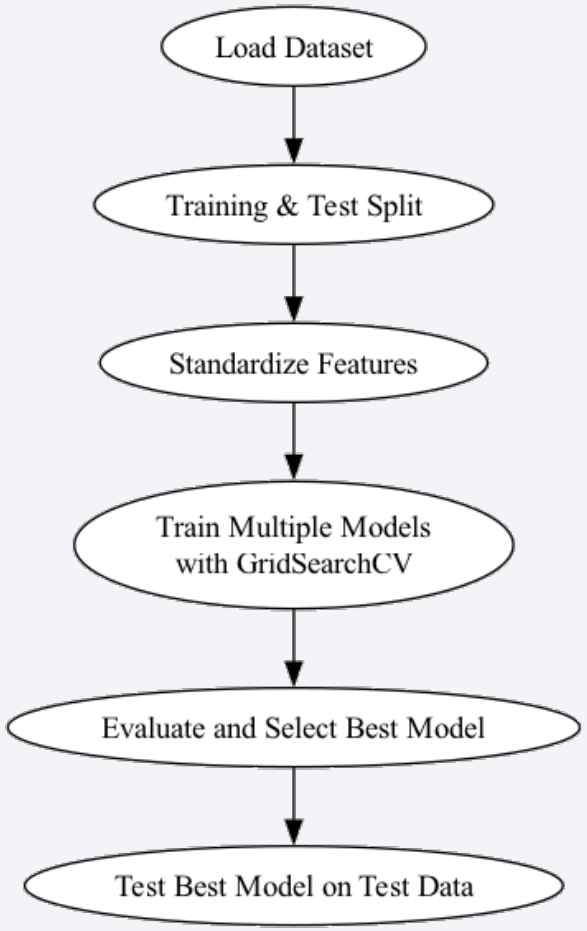
Standard scaling to normalize the feature set, ensuring that each feature contributes equally to the result.

Model Training and Hyperparameter Tuning:

Trained various machine learning models including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (k-NN). Hyperparameter tuning using GridSearchCV to find the optimal parameters for each model.

Evaluation:

Tested the models using the unseen test data to calculate the accuracy scores. Plotted confusion matrices for each model to visualize their performance, paying attention to false positives and negatives.



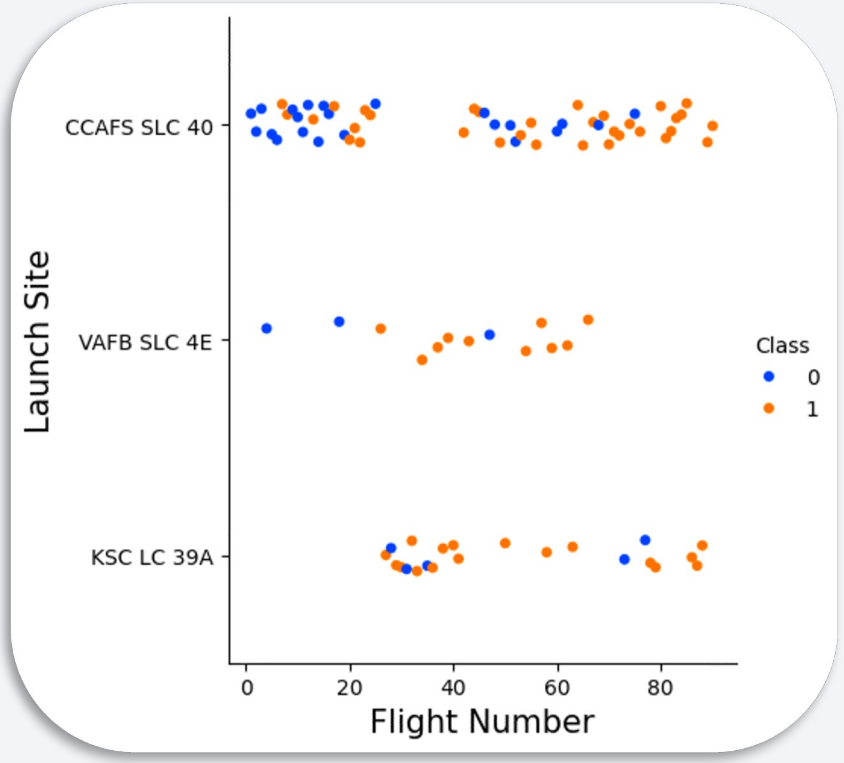
GITHUB LINK – Predictive Analysis (Classification)

Insights From EDA

Section 2

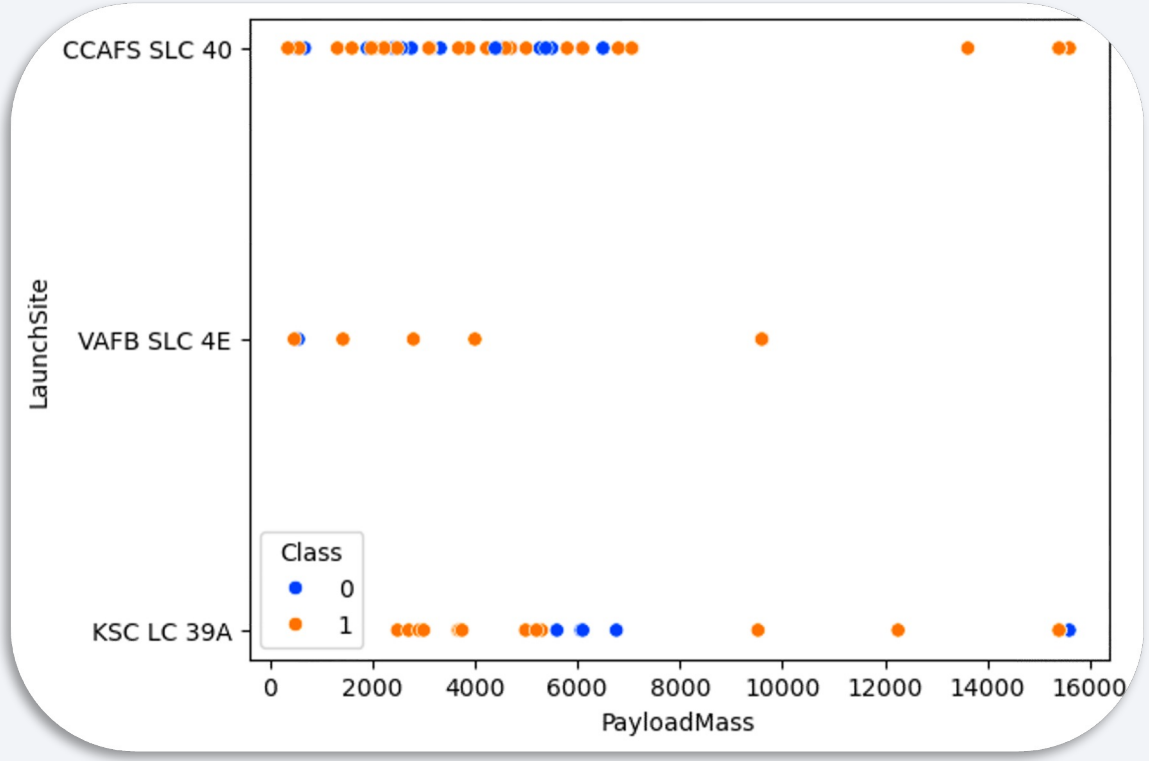
Flight Number vs. Launch Site

- Each point on the plot corresponds to a specific flight, plotted according to the flight number on the x-axis and the launch site on the y-axis.
- For VAFB SLC 4E, there are relatively fewer flights compared to the other sites.
- KSC LC 39A has a mix of successful and unsuccessful outcomes with no clear trend related to the flight number.
- CCAFS SLC 40 shows a concentration of both successful and unsuccessful launches across a wide range of flight numbers.



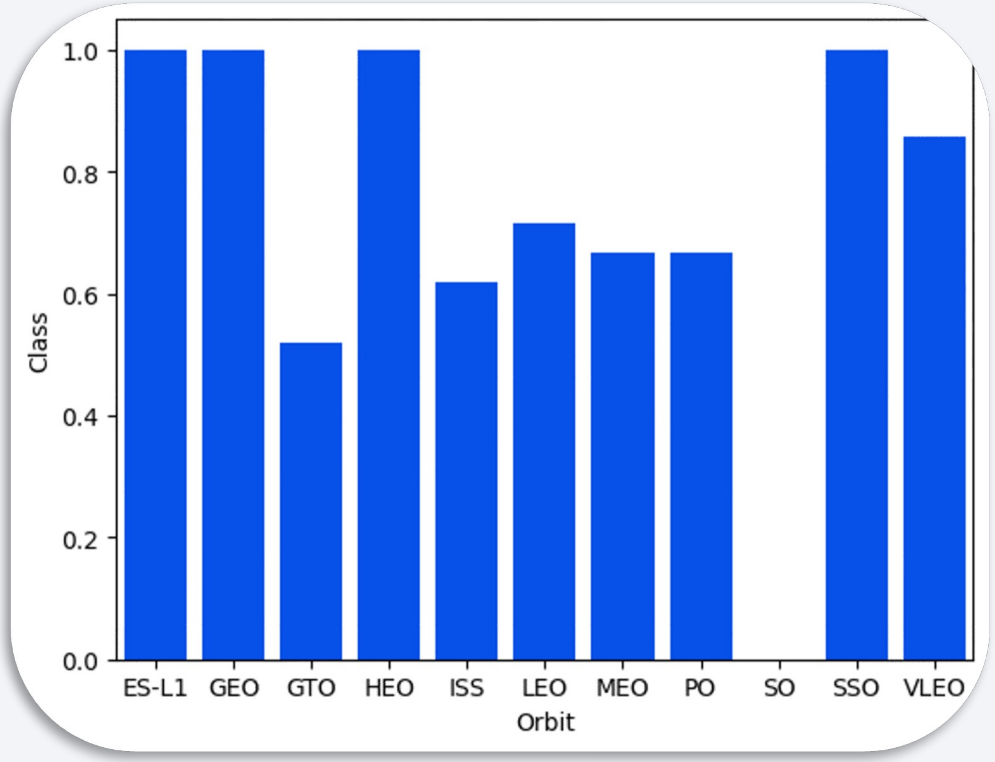
Payload vs. Launch Site

- In the scatter plot, Each dot represents a specific launch, with the payload mass on the x-axis and the launch site on the y-axis.
- Launches from CCAFS SLC 40 cover a wide range of payload masses, including both successful and unsuccessful landings.
- VAFB SLC 4E shows launches with a limited range of payload masses, with some unsuccessful attempts distributed among successful ones.
- KSC LC 39A demonstrates successful landings across various payload masses, with a few unsuccessful landings, notably at lower payload masses.



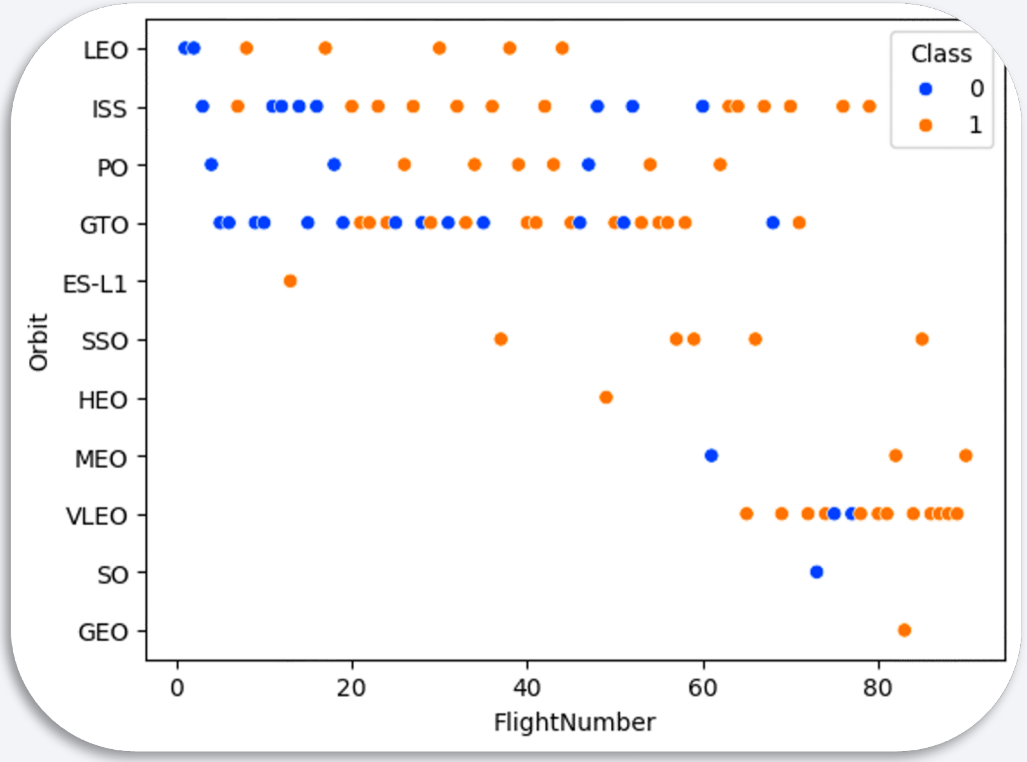
Success Rate vs. Orbit Type

- Bar chart representing the success rate (class) of launches for different orbital classifications:
- Some orbits, like ES-L1, GEO, and SSO (Sun-Synchronous Orbit), show higher success rates nearing 1, which suggests that most launches to these orbits result in successful landings.
- Orbits such as GTO (Geostationary Transfer Orbit) and LEO (Low Earth Orbit) show lower success rates, indicating a higher proportion of unsuccessful landings.
- VLEO (Very Low Earth Orbit) also exhibits a high success rate, similar to ES-L1 and GEO.



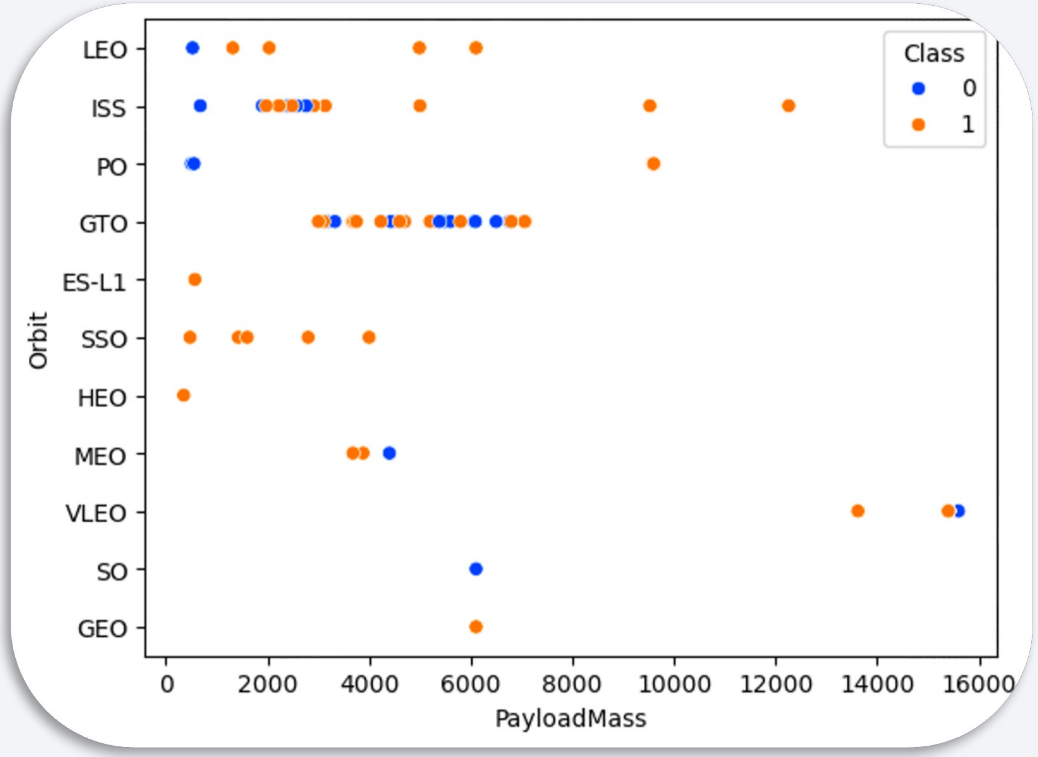
Flight Number vs. Orbit Type

- Scatter plot displays the success of Falcon 9 launches across different orbits as a function of flight number.
- Launches to LEO and ISS, VLEO have higher frequency and with a trend towards more successes in later flights.
- Orbits like ES-L1, SSO, and HEO have fewer data points, suggesting fewer launches to these orbits, with varying success rates.
- Certain orbits such as MEO, and GEO have fewer launches, and in the case of GEO, predominantly successful ones as per the available data.



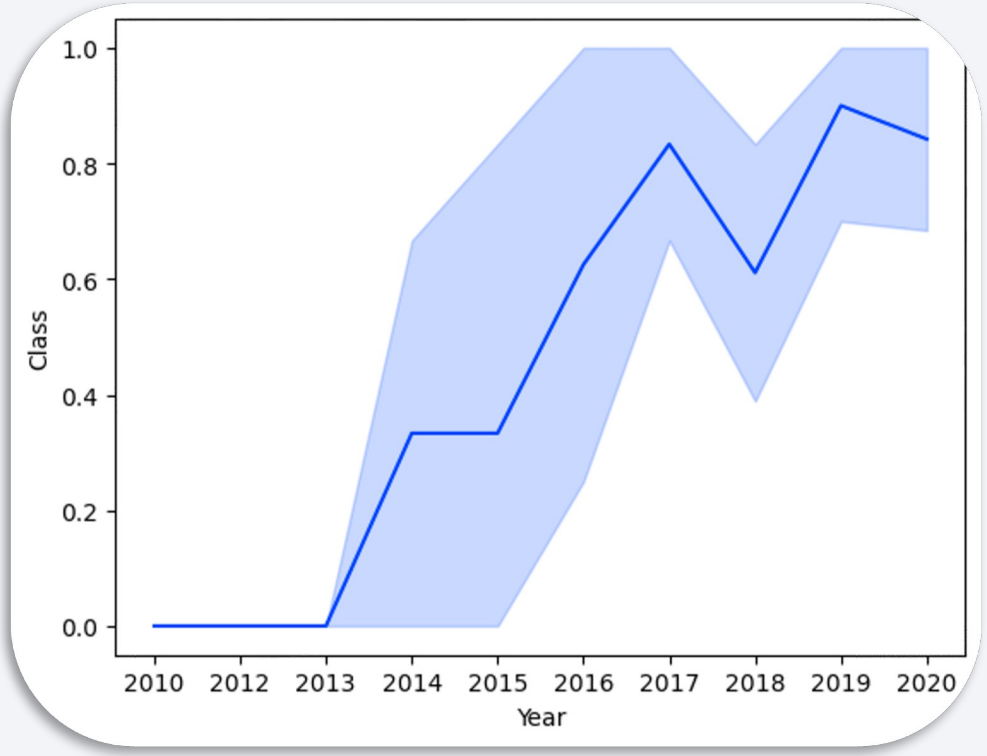
Payload vs. Orbit Type

- Scatter plot comparing the payload mass of Falcon 9 launches to their respective orbital classes.
- There is a cluster of launches to LEO and ISS with varying payload masses, where the success rate appears relatively high given the number of orange dots.
- For GTO, there are both successful & unsuccessful launches across a range of payload masses.
- The chart indicates that higher payload masses are not necessarily linked to unsuccessful landings, as there are successful landings at higher payload masses.



Launch Success Yearly Trend

- Line graph depicting the success rate of Falcon 9 rocket landings over a range of years from 2010 to 2020.
- An initial period with a low success rate (or no landings) around 2010.
- A notable increase in success rate starting around 2013, which continues to rise sharply until around 2016.
- After 2016, the success rate trends upwards more gradually with some fluctuations, indicating a maturation in the landing technology and procedures that allow for more consistent success.



All Launch Site Names

There are 5 unique Launch Sites:

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

Used the DISTINCT method to fetch the unique launch sites.

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

Python

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

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

Launch Site Names Begin with 'CCA'

- 5 records with where the Launch site name begins with CCA.
- Used WHERE and LIKE methods.

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

```
 * 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

Total Payload Mass

- Total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM
SPACEXTBL WHERE Customer = "NASA (CRS)"
```

Python

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM
SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1"
```

Python

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First successful landing outcome on ground pad

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE
Landing_Outcome = "Success (ground pad)"
```

Python

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT(Booster_Version) FROM
SPACEXTBL WHERE Landing_Outcome = "Success
(drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000
AND 6000
```

Python

```
 * 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

- Total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT
(Mission_Outcome) FROM SPACEXTBL GROUP BY
Mission_Outcome
```

Python

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

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX
(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Python

```
 * 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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%sql SELECT Booster_Version, Launch_Site FROM
SPACEXTBL WHERE (Landing_Outcome = 'Failure
(drone ship)') AND (substr(Date,0,5)='2015');
```

Python

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Ranked count of landing outcomes (Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

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

Python

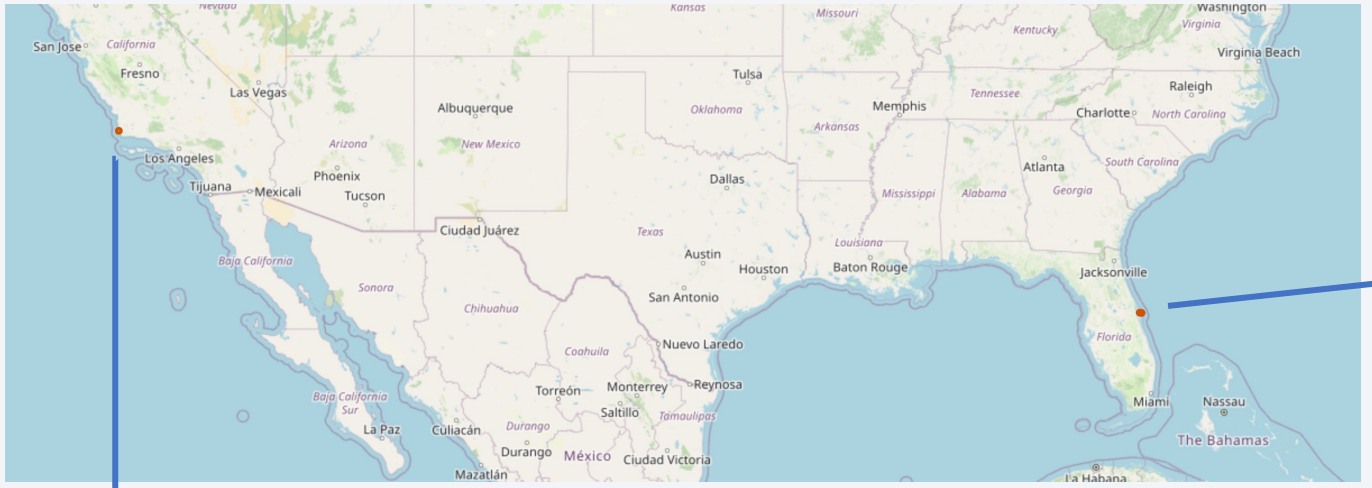
```
 * sqlite:///my_data1.db
Done.
```

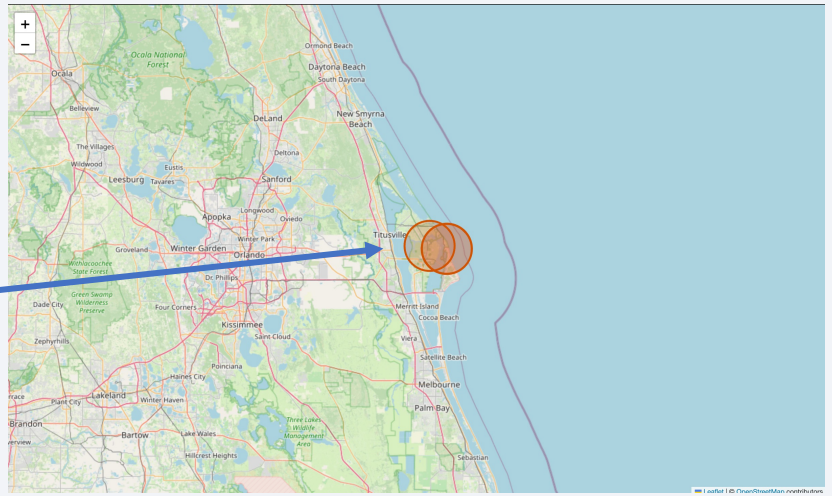
Landing_Outcome	Total_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Launch Sites Proximity Analysis

Section 3

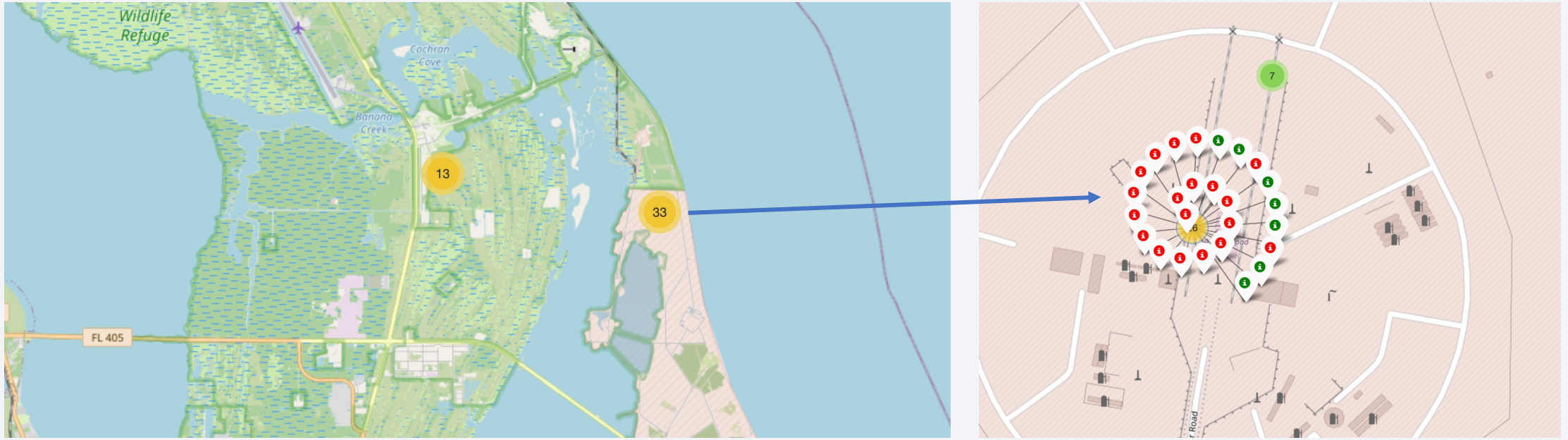
All Launch Site Markers – Folium Map





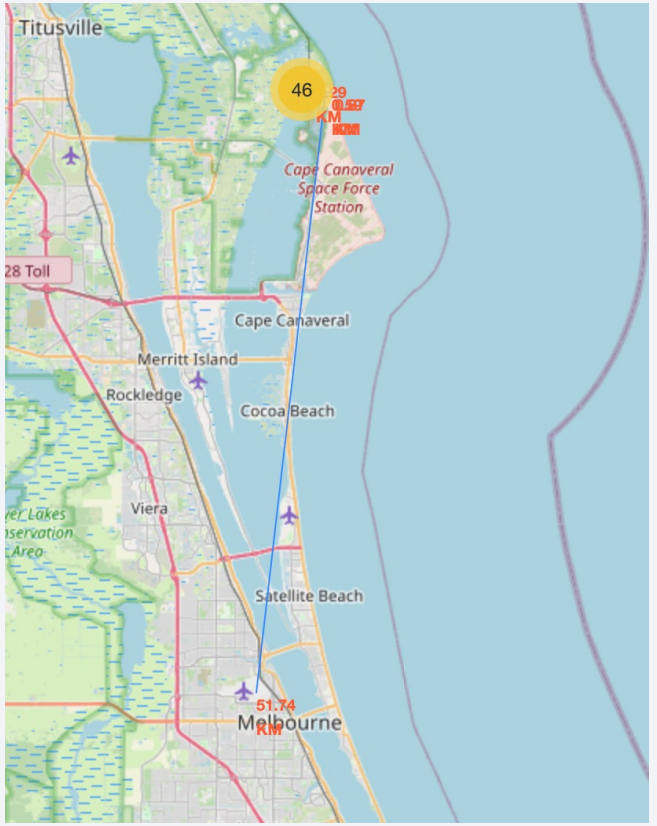
- Red Markers showing all launch Sites in Folium map

Color Labeled Launch Outcomes – Folium Map



- Color abled launch outcomes from the launch sites in Miami

Launch Site Proximities – Folium Maps



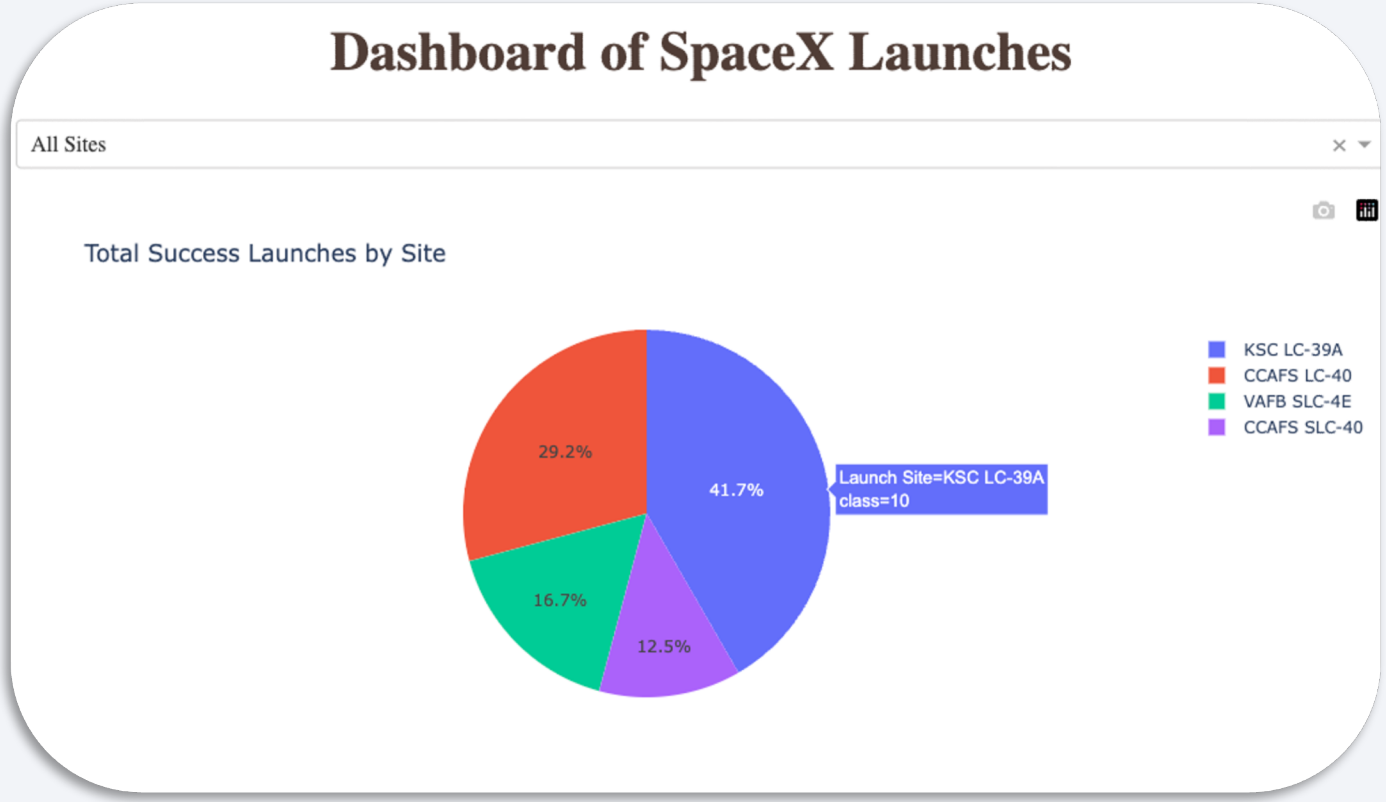
- Distance from the launch site to the coastline – 0.87KM
- From the launch site to the nearest city – 51.74KM

Dashboard with Plotly Dash

Section 4

Total Success Launches – Plotly Dash

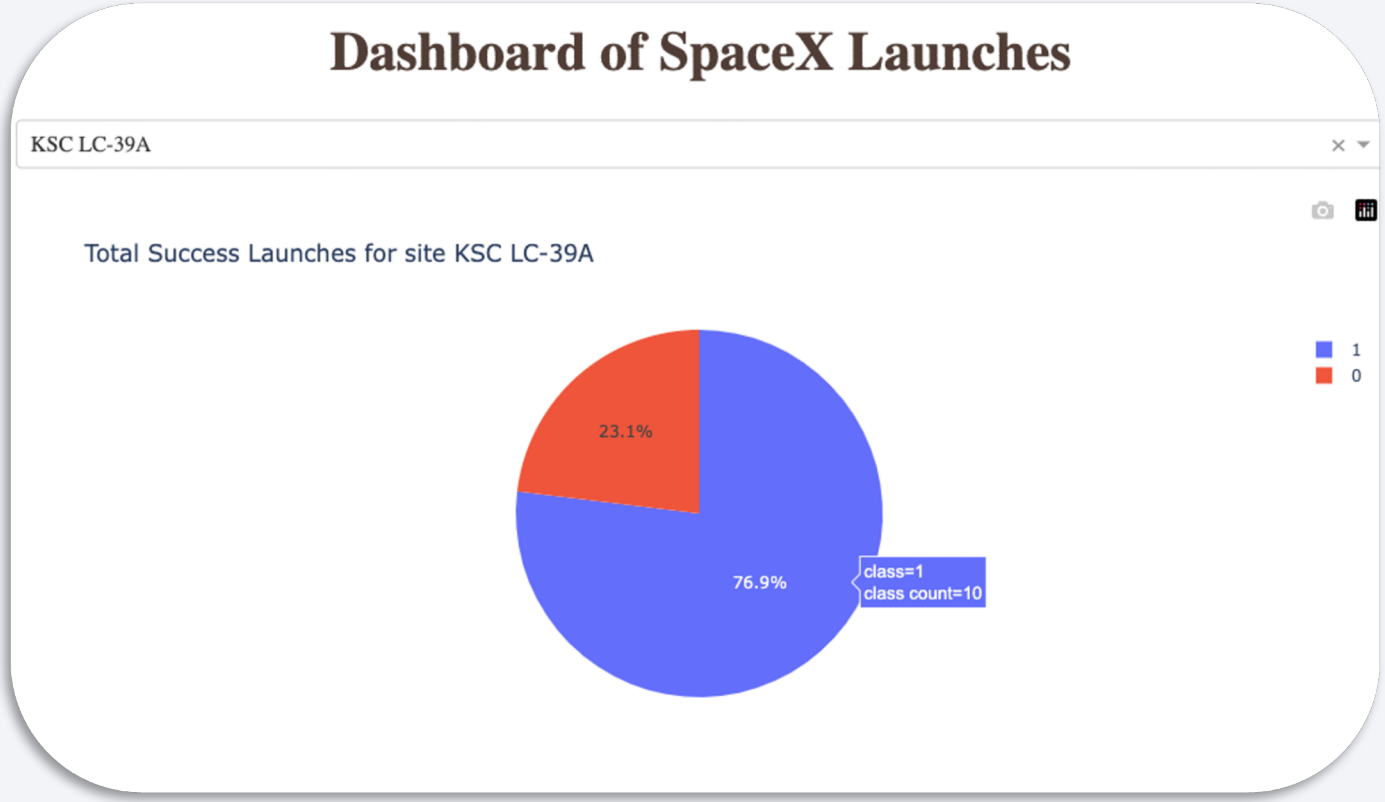
- KSC LC-39A has the largest share of successful launches, making up 41.7% of the total.
- CCAFS LC-40 comes next with 29.2% of successful launches.
- VAFB SLC-4E accounts for 16.7%.
- CCAFS SLC-40 constitutes the smallest portion with 12.5%.



Launches for KSC LC-39A – Plotly Dash

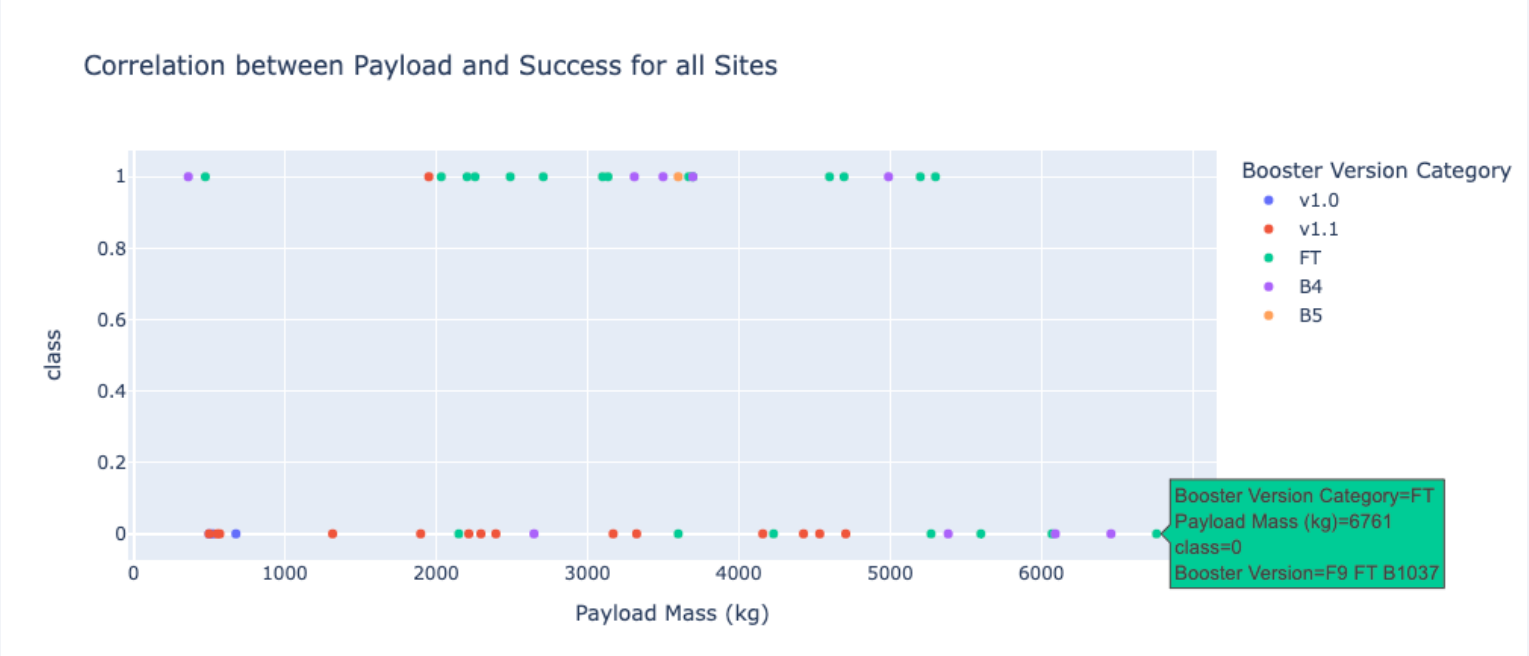
Success rate for launches from the Kennedy Space Center Launch Complex 39A (KSC LC-39A):

- Blue section: Successful launches, which account for 76.9% of the total launches from this site.
- Red section: Unsuccessful launches, making up 23.1% of launches from KSC LC-39A.



Scatter Correlation Payload vs Success – Plotly Dash

- Successful launches (class=1) are distributed across the entire range of payload masses, suggesting that success is not exclusively dependent on payload mass.
- There is a cluster of unsuccessful launches (class=0) at lower payload masses, which are marked by orange and red dots representing the B4 and B5 booster versions.

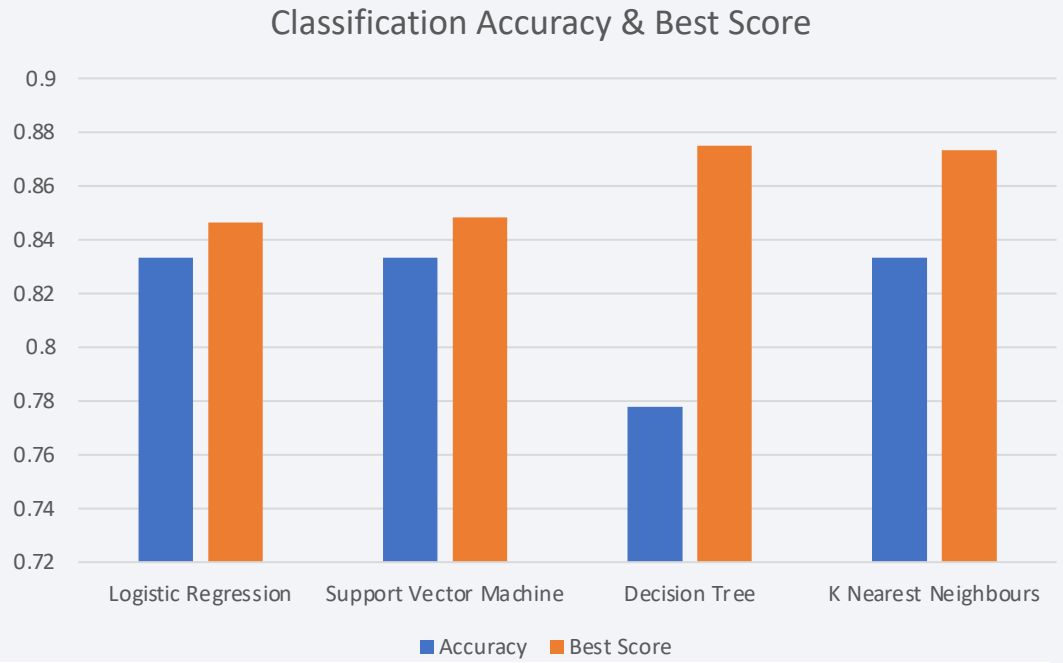


Predictive Analysis

Section 5

Classification Accuracy

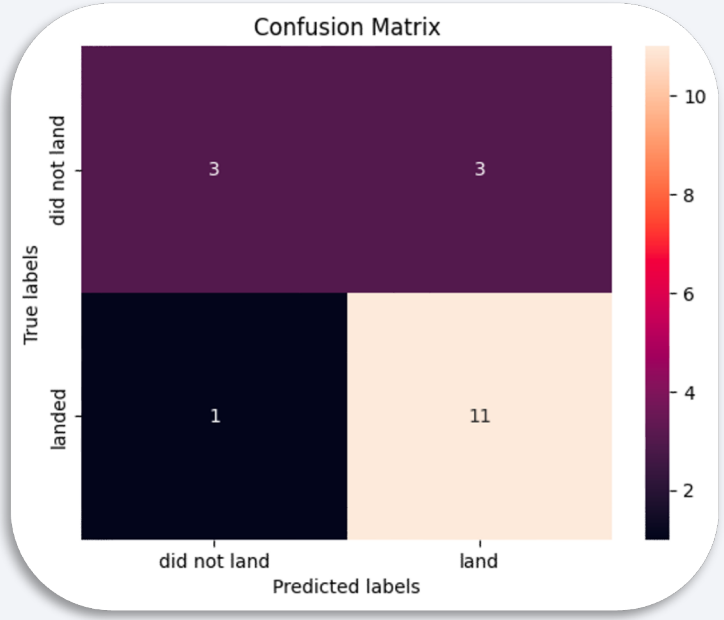
- Decision Tree: Noticeably lower accuracy, suggesting it maybe less effective for this problem or it may have overfit during training.
    - It has the highest best score from grid search, which could indicate that it performed well on the training data but did not generalize as well to the test data.
- SVM & Logistic Regression: Have similar accuracy and best scores.
- KNN: Has a high best score, which, combined with its good accuracy, could make it a reliable model for this task.



	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.777778	0.875000
3	K Nearest Neighbours	0.833333	0.873214

Confusion Matrix – Decision Tree

- The model is more likely to predict a landing (14 predictions) than not (4 predictions).
- The model shows a good number of true positives, indicating that when it predicts a landing, it is often correct.
- However, the false positives and false negatives suggest there is room for improvement. A false positive can be costly if relying on the model for resource allocation or mission planning.
- The overall accuracy is (11 + 3) / 18 = 0.7778, which aligns with the accuracy score reported earlier for the Decision Tree.



Conclusions

Performance of Models:

- Logistic Regression, SVM, and K-NN exhibited similar accuracy scores on test data (~83.33%).
- Decision Tree showed a lower accuracy score on test data (~77.78%) but the highest best score from GridSearchCV (87.5%).

Decision Tree Analysis:

- Confusion Matrix indicated the Decision Tree model had 11 true positives & 3 true negatives, suggesting good predictive power.
- There were 3 false positives, highlighting a tendency towards predicting success (landing) when it was not the case.
- A single false negative indicates that the model seldom missed a successful landing.

Hyperparameter Tuning:

- Each model underwent hyperparameter tuning using GridSearchCV, improving their validation performance.
- The best parameters were identified for each model, tailoring them to the specific dataset characteristics.

Reliability and Model Selection:

- The models demonstrated reliable performance, but the selection for operational use would consider false positives & negatives due to their different costs.
- Considering the trade-offs, the Decision Tree model had the best potential, shown by its highest GridSearchCV score, despite lower test accuracy.


Thank You!
SPACEX