DATA SCIENCE CAPSTONE PRESENTATION

Nasir Abnathya
11.20.2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - Exploratory Data Analysis w/ Data Visualization
  - Exploratory Data Analysis w/ SQL
  - Predictive Analysis
- Summary of results
  - Exploratory Data Analysis results
  - Predictive Analysis Results

Introduction

Project Background and context:

SpaceX is the most successful company within the commercial space age. The company's newest advertisement, the Falcon 9, costs roughly 100 million less than its competitors, sitting at 62 million dollars. The savings were brought about due to the ability to reuse stage one of the process. If we can determine the success of stage one, the cost of a launch becomes easily estimated. With the help of machine learning models and public information, we can predict whether or not SpaceX will actually reuse stage one.

Methodology

Data Collection Methodology:

- Leverage the SpaceX REST API for data retrieval.
- Conduct web scraping from Wikipedia to gather additional information.

Data Wrangling:

- Filter and cleanse the dataset.
- Handle missing values effectively.
- Apply one-hot encoding to prepare the data for binary classification tasks.

Exploratory Data Analysis (EDA):

- Conduct thorough EDA using visualizations and SQL queries.

Interactive Visual Analytics:

- Create interactive visualizations using Folium and Plotly Dash.

Predictive Analysis:

- Develop, fine-tuned, and evaluated classification models to achieve optimal performance.

Data Collection

The data collection process utilized a combination of API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia entry. Both methods were necessary to gather comprehensive information about the launches, enabling a more detailed analysis.

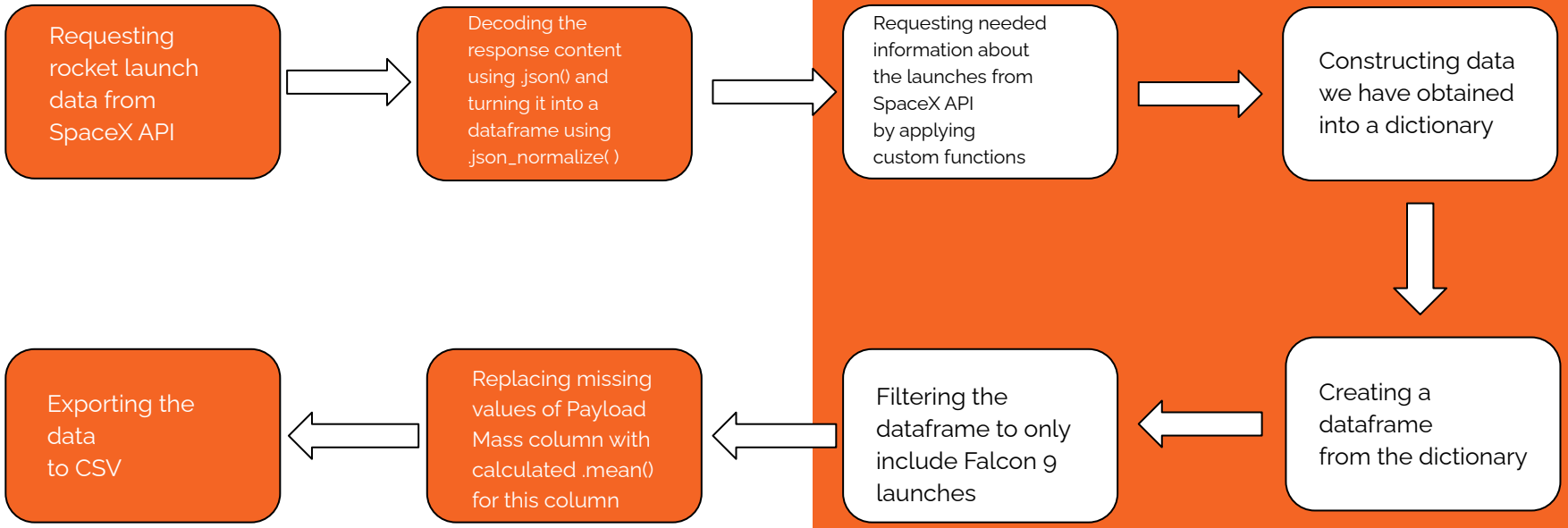
Data columns retrieved via SpaceX REST API:

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

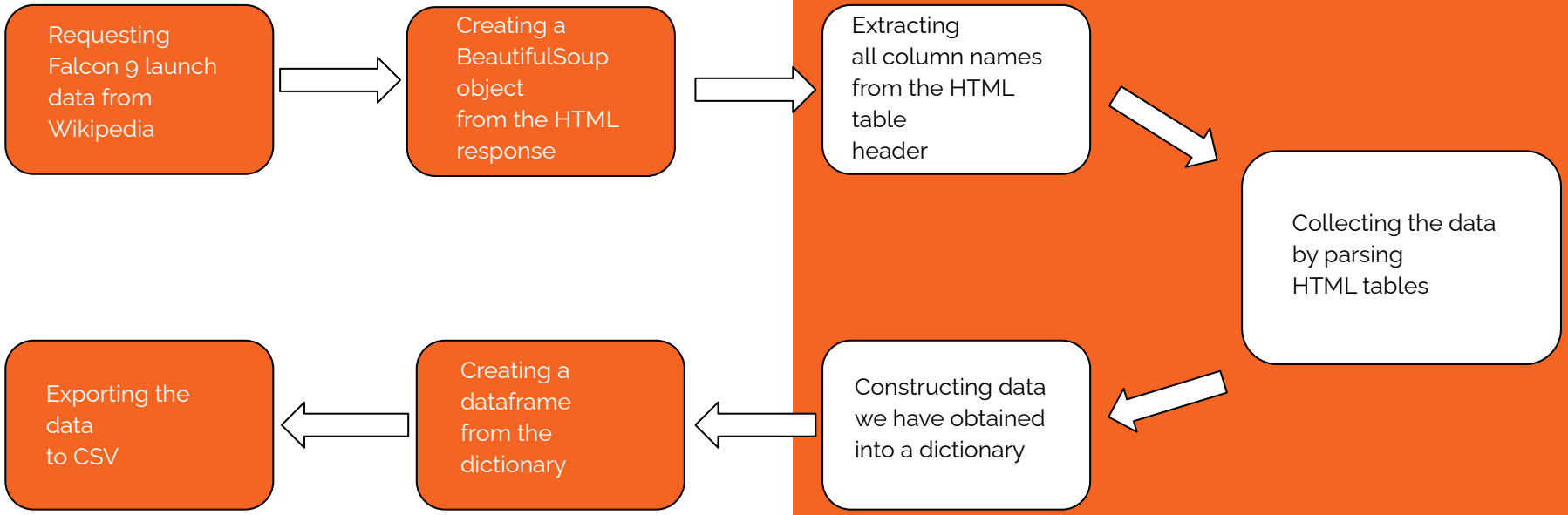
Data columns retrieved via Wikipedia web scraping:

- *Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Date, Time, Launch Outcome, Version Booster, Booster Landing.*

Data collection – SpaceX API



Data collection – Web Scraping



Data Wrangling

The dataset includes various scenarios where the booster failed to land successfully. For instance:

- True Ocean indicates a successful landing in a designated ocean region, while False Ocean indicates an unsuccessful attempt in the same region.
- True RTLS represents a successful landing on a ground pad, while False RTLS denotes an unsuccessful landing attempt.
- True ASDS signifies a successful landing on a drone ship, whereas False ASDS represents a failed landing attempt on a drone ship.

To simplify these outcomes for analysis, they were converted into training labels:

- A value of "1" represents a successful booster landing.
- A value of "0" indicates an unsuccessful landing.

EDA with Data Visualization

The following 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
- Success Rate Yearly Trend

Scatter plots were used to visualize relationships between variables. If significant relationships are identified, these insights could inform machine learning models.

Bar charts were utilized to compare discrete categories, highlighting the relationships between specific categories and measured values.

Line charts were employed to illustrate trends over time, effectively capturing time series patterns.

EDA with SQL

Executed SQL Queries:

- Retrieved the names of unique launch sites involved in the space missions.
- Displayed 5 records where launch site names start with the string ‘CCA’.
- Calculated the total payload mass carried by boosters launched by NASA (CRS).
- Found the average payload mass carried by the booster version F9 v1.1.
- Identified the date of the first successful landing outcome on a ground pad.
- Listed the names of boosters that successfully landed on a drone ship and carried payload masses between 4000 and 6000.
- Counted the total number of successful and failed mission outcomes.
- Retrieved the booster versions that carried the maximum payload mass.
- Listed failed landing outcomes on drone ships, including their booster versions and launch site names, for the months in 2015.
- Ranked the count of landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") between 2010-06-04and 2017-03-20 in descending order.

Building an interactive map using Folium

Markers for All Launch Sites:

- Added a marker with a circle, popup label, and text label for the NASA Johnson Space Center, using its latitude and longitude as the starting location.
- Placed markers with circles, popup labels, and text labels for all launch sites, highlighting their geographical locations and proximity to the equator and coastlines.

Markers with Color for Launch Outcomes at Each Site:

- Added color-coded markers to represent launch outcomes: green for successful launches and red for failed ones. A marker cluster was used to visualize which launch sites have higher success rates.

Distances from a Launch Site to Nearby Landmarks:

- Added colored lines to depict distances from the KSC LC-39A launch site (as an example) to nearby landmarks, such as railways, highways, the coastline, and the closest city.

Building a Dashboard using Dash

Launch Sites Dropdown List:

- Implemented a dropdown menu to allow users to select a specific launch site.

Pie Chart of Successful Launches (All Sites/Specific Site):

- Created a pie chart displaying the total count of successful launches across all sites, or the success vs. failure counts for a selected launch site.

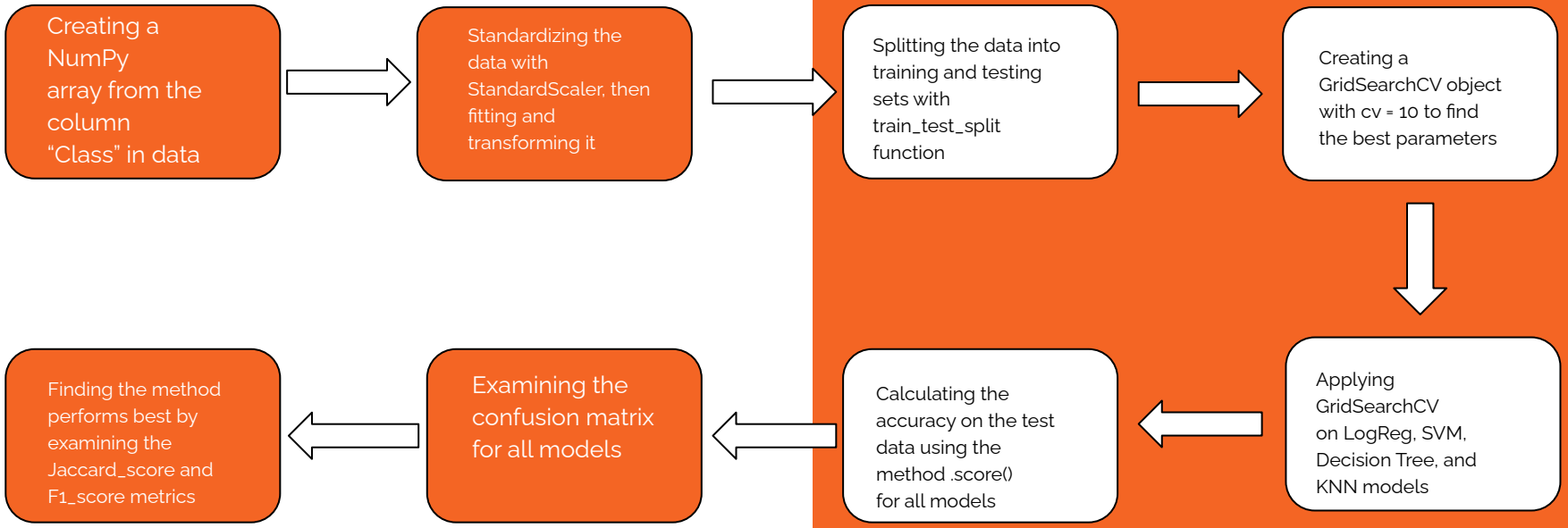
Payload Mass Range Slider:

- Added a slider to enable users to select a specific payload mass range.

Scatter Plot of Payload Mass vs. Success Rate for Booster Versions:

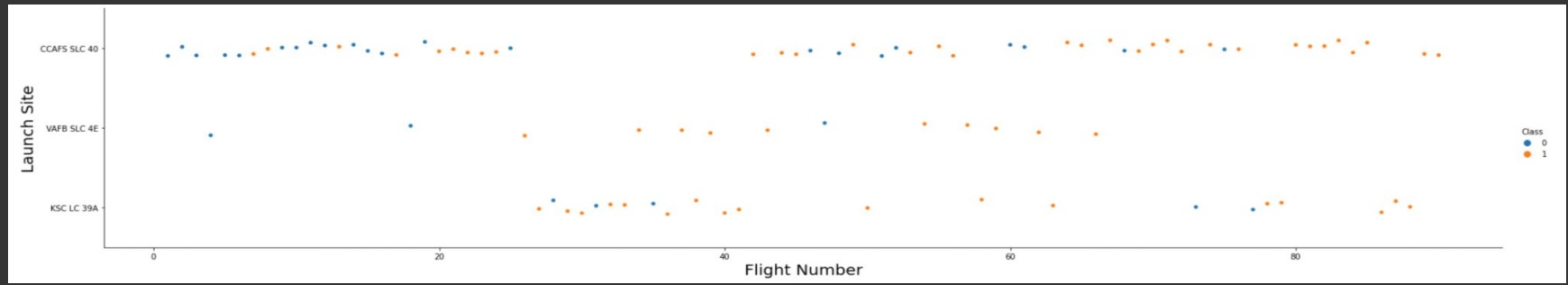
- Developed a scatter plot to visualize the relationship between payload mass and launch success for different booster versions.

Predictive Analysis



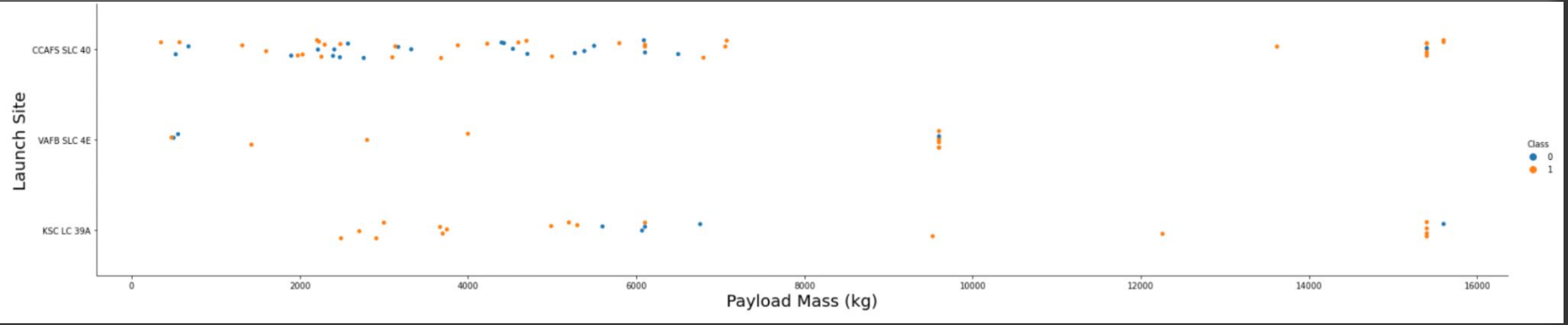
Flight Number vs. Launch Site

- Early flight attempts were unsuccessful, while recent launches have been consistently successful.
- Approximately half of all launches have occurred at the CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A exhibit superior success rates.
- A positive correlation can be observed between launch attempts and the likelihood of mission success.



Payload vs. Launch Site

- A positive correlation exists between payload mass and launch success rates across all launch sites.
- Most launches carrying payloads exceeding 7000 kg have been successful.
- KSC LC 39A maintains a perfect success record for payloads under 5500 kg.



Orbit Type vs. Success Rate

Orbits with 100% Success:

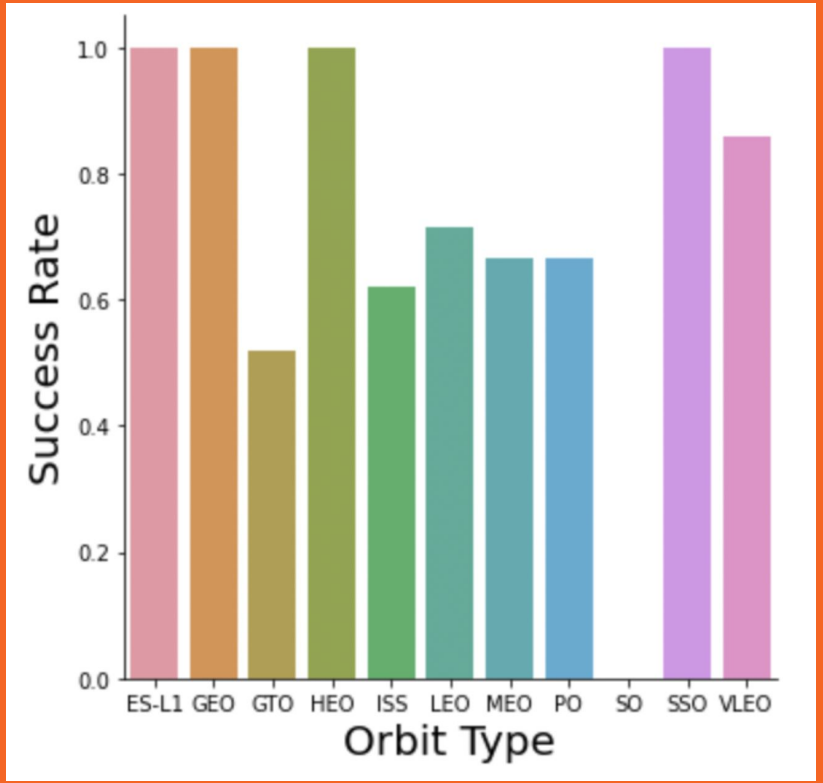
- Earth-Sun L1, Geostationary Earth Orbit (GEO), Highly Elliptical Orbit (HEO), Sun-synchronous Orbit (SSO)

Orbits with 0% Success:

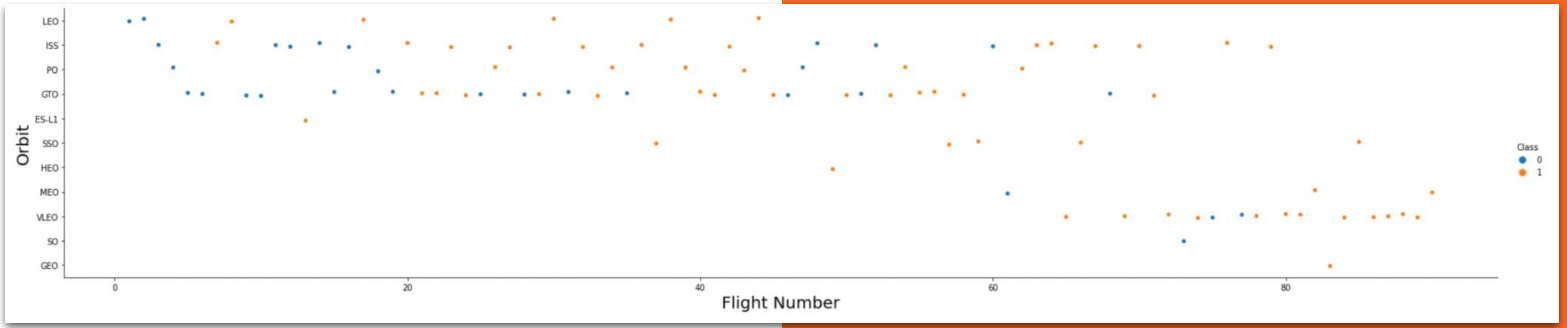
- Solar Orbit (SO)

Orbits with 50-85% Success:

- Geostationary Transfer Orbit (GTO), International Space Station (ISS), Low Earth Orbit (LEO), Medium Earth Orbit (MEO), Polar Orbit (PO)



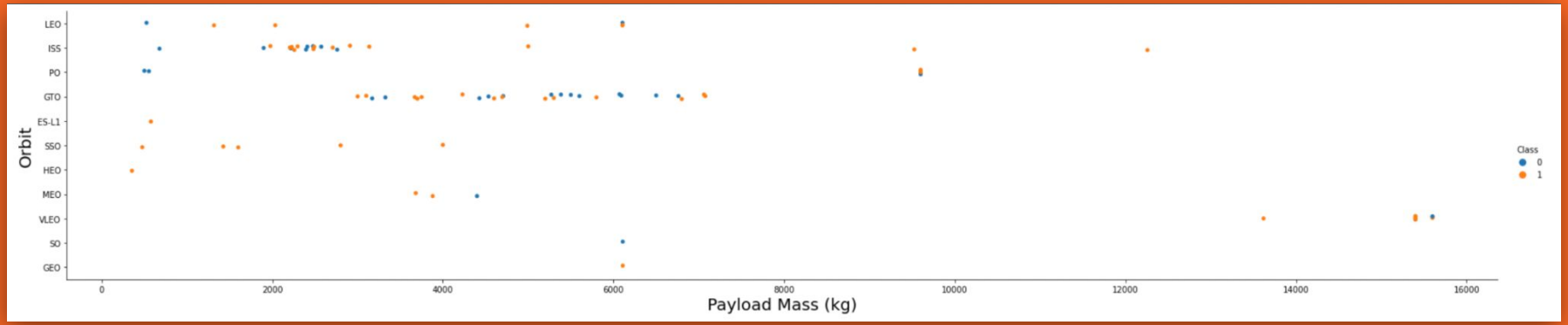
Flight Number vs. Orbit Type



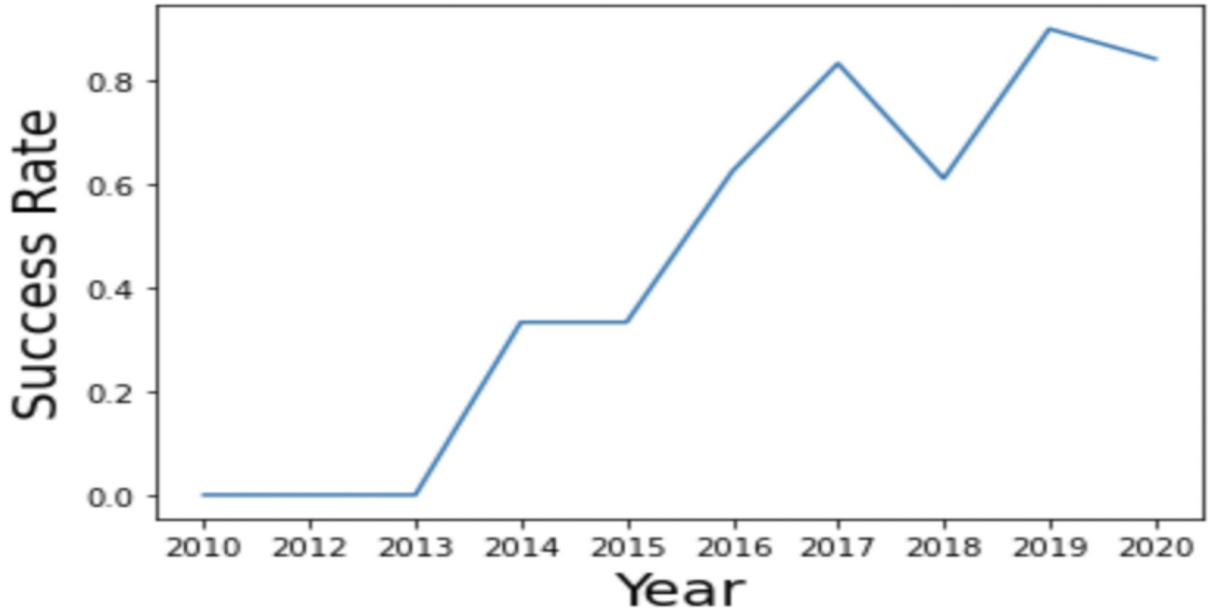
A positive correlation exists between the number of flight attempts and mission success in LEO orbit. In contrast, no discernible relationship is observed between flight attempts and success in GTO orbit.

Payload Mass vs. Orbit Type

Heavy payloads negatively impact GTO orbits, while positively influencing GTO and Polar LEO (ISS) orbits.



Yearly Trend of Launch Success



- The success rate increased at a fairly steady pace from 2013-2020

Names of Launch Site

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

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

Out[4]:

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

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

Average Payload Mass

```
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.
```

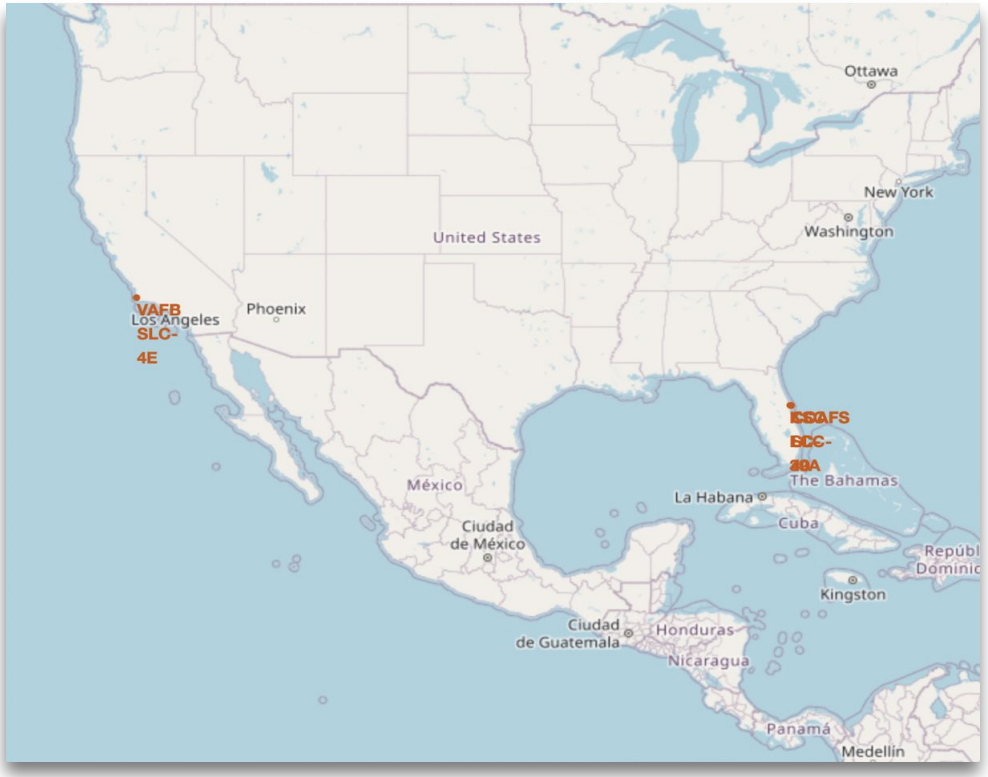
Out[7]:

average_payload_mass
2534

Global Map of Launch Site Locations

The majority of launch sites are situated in close proximity to the equator, where the Earth's rotational velocity is maximized. This inherent eastward velocity provides a significant boost to the spacecraft's orbital velocity.

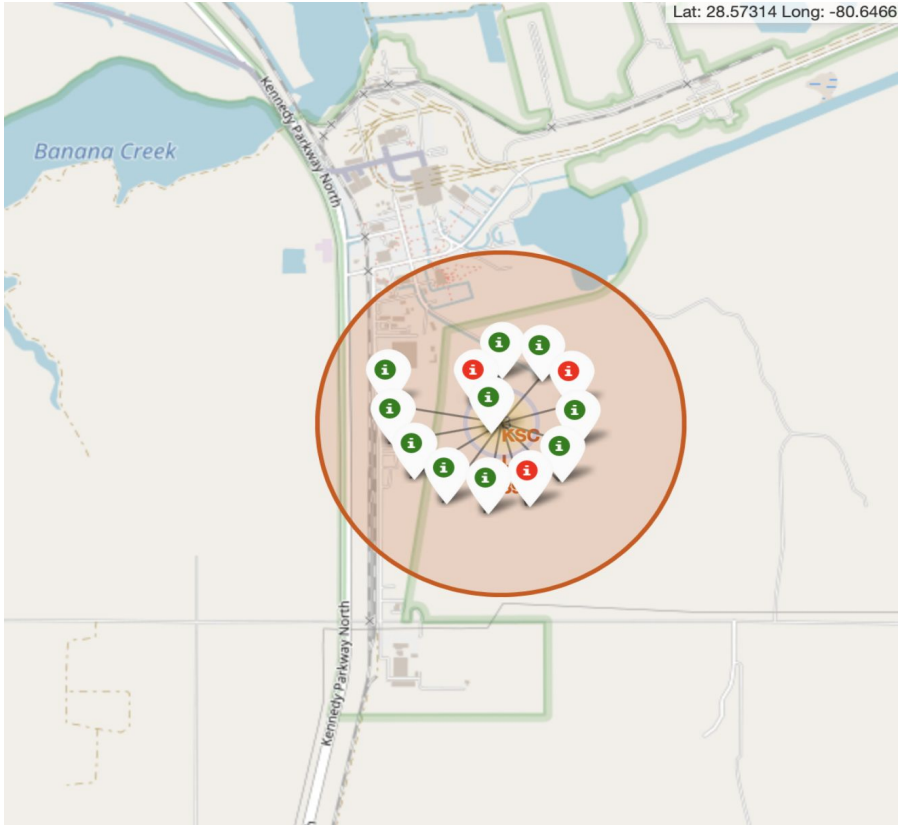
All launch sites are strategically positioned near coastal areas, thereby mitigating the risks associated with rocket failures or debris dispersion over populated landmasses.



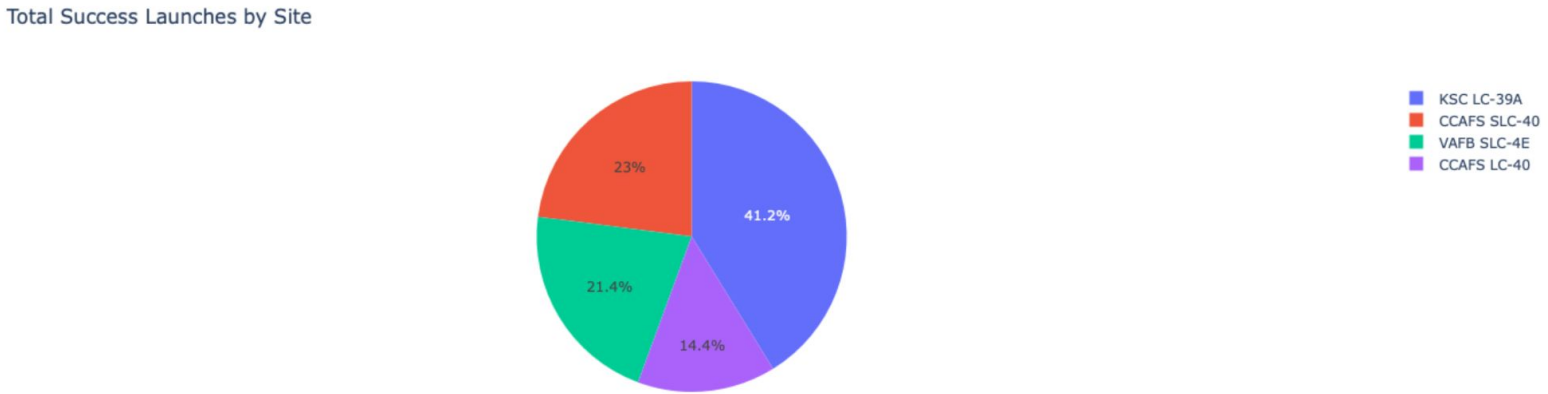
Successful vs Unsuccessful Launch Indicators

Locations marked in green
indicate each successful launch

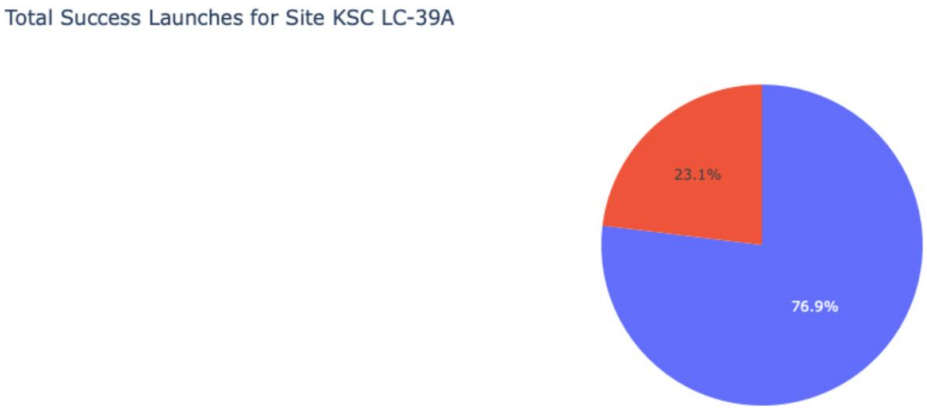
Locations marked in red
indicate each failed launch



Launch Success Rate (All sites)



Launch Site with Highest Success Rate



KSC LC-39A has the highest success rate at 76.9%. This is after thirteen total attempts (10 successful and 3 unsuccessful).

Predictive Analysis (Classification)

Due to the relatively small sample size of the test set (18 samples), a definitive conclusion regarding the superior performance of a particular method is not feasible.

Scores/Accuracy of 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/Accuracy of 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

In Conclusion

- The Decision Tree Model is the optimal algorithm for this dataset.
- Launches with lower payload masses tend to yield better results.
-
- The majority of launch sites are located in close proximity to the equator and coastal regions.
- A positive trend in launch success rates has been observed over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- ES-L1, GEO, HEO, and SSO orbits have achieved perfect success rate (100%).


I'M FINISHED