IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

John Eide
28-September-2023

Outline

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

Executive Summary

Summary of methodologies:

- Gathering and Preparing Data
- Investigating Data Through Visual Exploration
- Investigating Data Using SQL Queries
- Creating an Interactive Map with Folium
- Developing a Dashboard with Plotly Dash
- Forecasting and Categorization Analysis

Summary of all results:

- Findings from Exploratory Data Analysis
- Snapshot of Interactive Analytics Demonstration
- Outcomes of Predictive Analysis

Introduction

Project background and context:

- SpaceX stands out as the preeminent company in the era of commercial space exploration, heralding an era of more accessible space travel. Prominently featured on their website are Falcon 9 rocket launches, offered at a competitive price point of 62 million dollars, a significant departure from the upward of 165 million dollars charged by other providers. This cost disparity owes much of its existence to SpaceX's groundbreaking capability to recover and reuse the first stage of their rockets. Therefore, our objective is to ascertain the likelihood of a successful first-stage landing, as this determination holds the key to estimating the overall launch cost. Leveraging publicly available data and employing advanced machine learning models, our mission is to predict whether SpaceX will indeed opt for first-stage reuse.

Questions to answer:

- What impact do factors like payload mass, launch location, flight count, and orbital parameters have on the probability of a successful first-stage landing?
- Is there an observable trend of improving first-stage landing success rates over the course of time?
- Which binary classification algorithm is most suitable for this particular scenario?

Section 1

Methodology

Methodology

Data collection methodology:

- Data was downloaded via provided api and .csv links within the Coursera chosen integrated development platforms (IDE). Additional data was obtained via web-scraping from a Coursera provided URL.

Performed data wrangling:

- The data was imported into a Pandas DataFrame for further analysis. In addition, data was 'cleaned' or dispensed with based on relevancy.

Performed exploratory data analysis (EDA) using visualization and SQL:

- Extracted various queries from the Spacex MySQL Lite database.

Performed interactive visual analytics using Folium and Plotly Dash

Methodology

Performed predictive analysis using classification models:

- Initially, the gathered data underwent normalization before being divided into training and test subsets.
- Four distinct classification prediction algorithms were applied to the data, utilizing hyperparameters selected by the user.
- All possible combinations of hyperparameters were tested using the out-of-sample test dataset.
- Subsequently, each classification output underwent assessment to determine its accuracy.

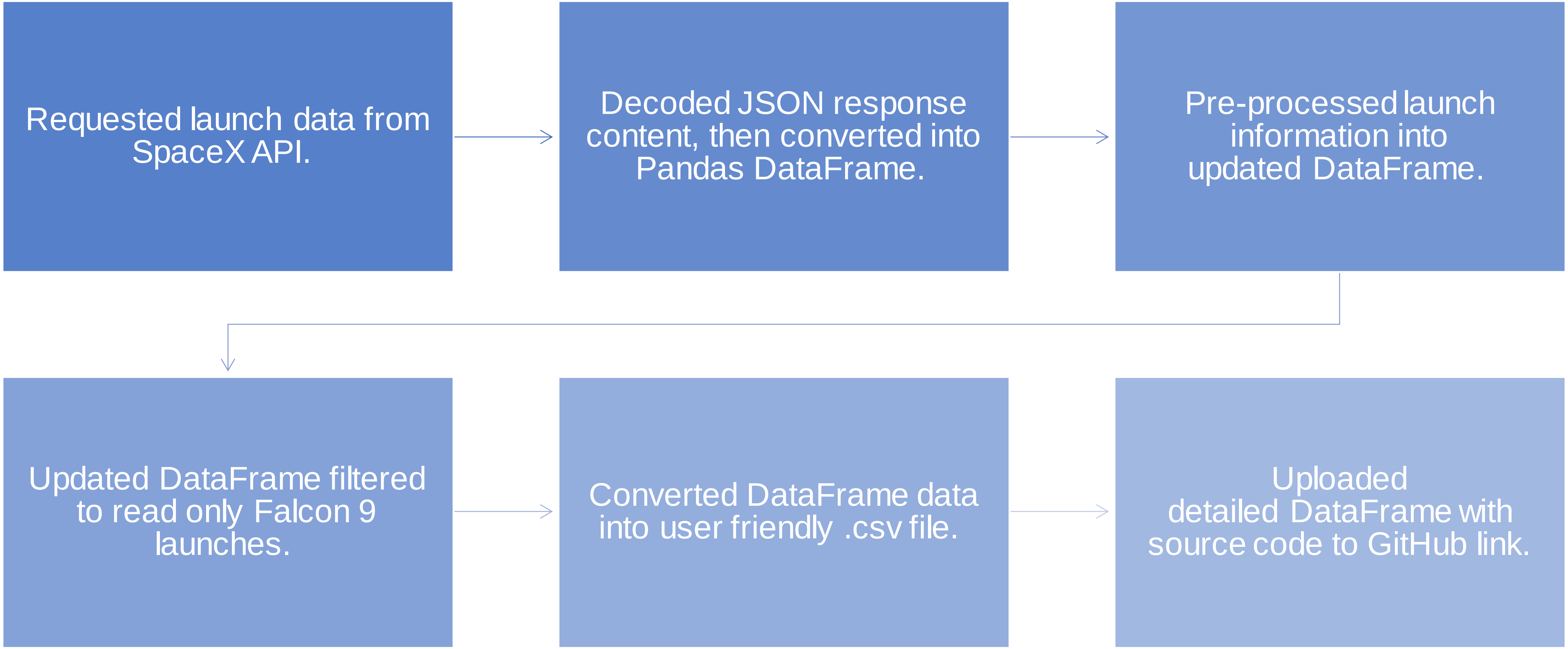
Data Collection

How Data Sets Were Collected:

- *Dataset 1*, SpaceX api: https://api.spacexdata.com/v4/past
  - Rockets used
  - Launch pads used
  - Cores
  - Landing types
- *Dataset 2*, https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Wed scraped Falcon 9 and Falcon Heavy Launches Records

Data Collection – SpaceX API

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Requested launch data from SpaceX API.

Decoded JSON response content, then converted into Pandas DataFrame.

Pre-processed launch information into updated DataFrame.

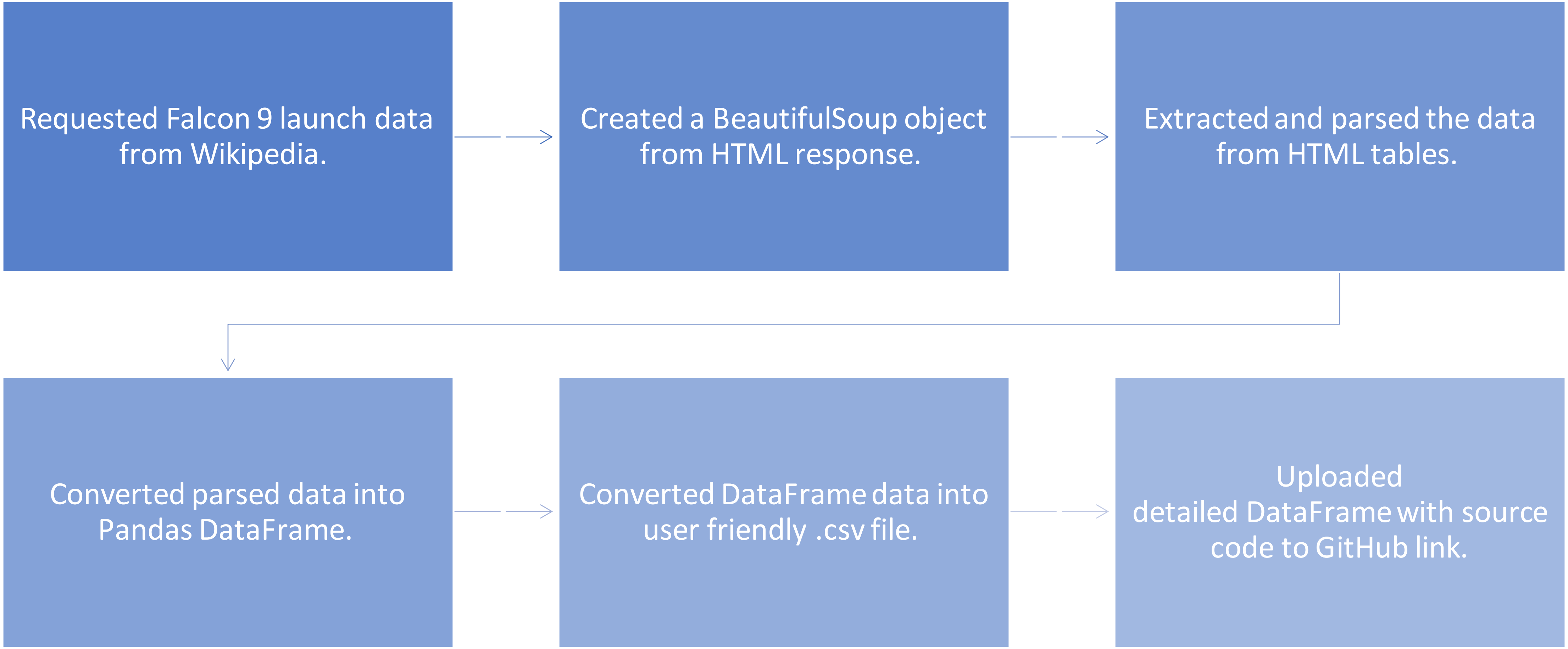
Updated DataFrame filtered to read only Falcon 9 launches.

Converted DataFrame data into user friendly .csv file.

Uploaded detailed DataFrame with source code to GitHub link.

Data Collection – Scraping

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-webscraping.ipynb



Requested Falcon 9 launch data from Wikipedia.

Created a BeautifulSoup object from HTML response.

Extracted and parsed the data from HTML tables.

Converted parsed data into Pandas DataFrame.

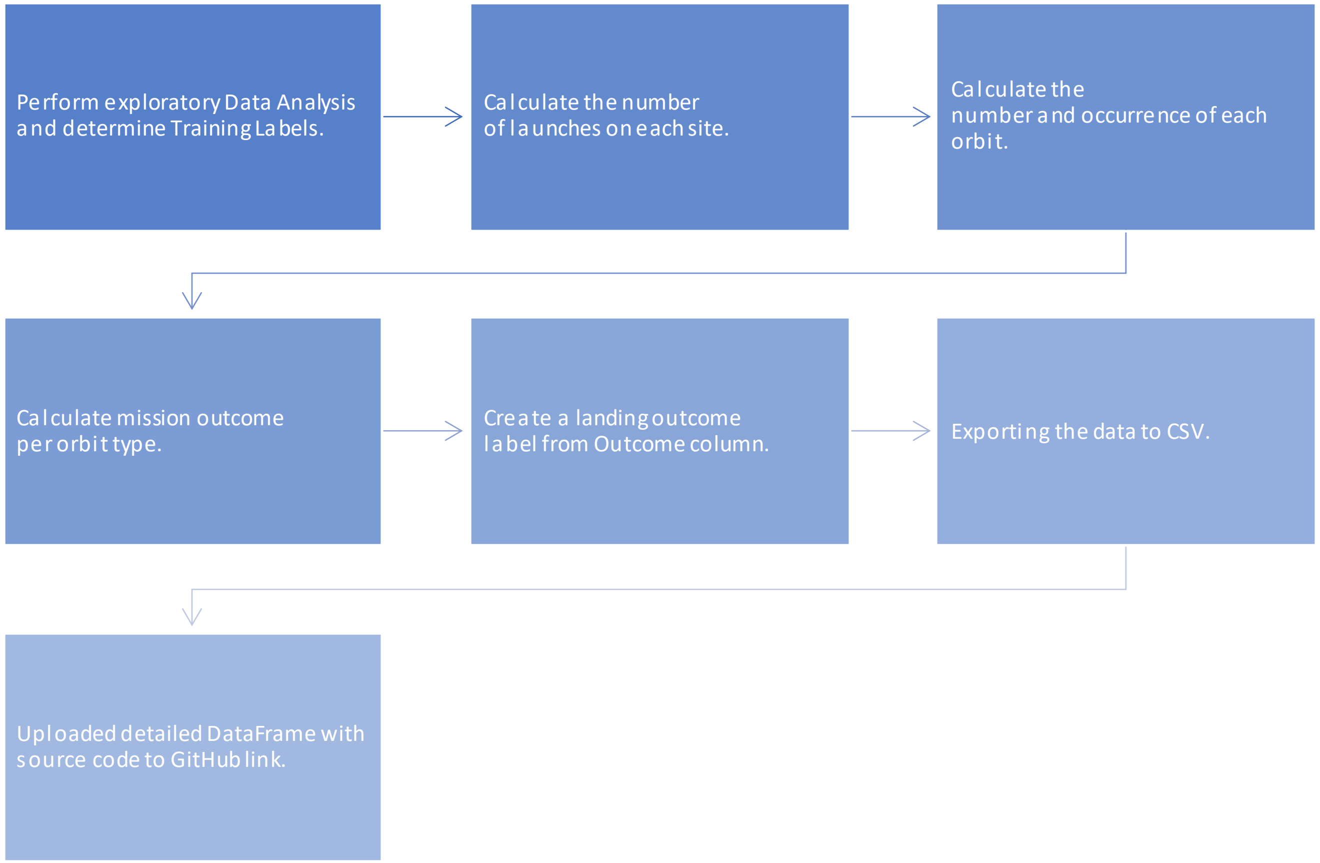
Converted DataFrame data into user friendly .csv file.

Uploaded detailed DataFrame with source code to GitHub link.

Data Wrangling

In the dataset, there are multiple instances where the booster's landing was not successful. On some occasions, landing attempts were made but ended in failure due to accidents. For instance, "True Ocean" indicates a mission that successfully landed in a specific oceanic region, while "False Ocean" denotes an unsuccessful landing attempt in a designated oceanic area. Similarly, "True RTLS" signifies a mission with a successful landing on a ground pad, while "False RTLS" indicates an unsuccessful ground pad landing. Lastly, "True ASDS" represents a mission with a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful drone ship landing.

These outcomes are primarily transformed into Training Labels, where a value of "1" signifies a successful booster landing, while a value of "0" denotes an unsuccessful landing.



Perform exploratory Data Analysis and determine Training Labels.

Calculate the number of launches on each site.

Calculate the number and occurrence of each orbit.

Calculate mission outcome per orbit type.

Create a landing outcome label from Outcome column.

Exporting the data to CSV.

Uploaded detailed DataFrame with source code to GitHub link.

SOURCE: https://github.com/EIDSTER/coursera/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charting was employed to visually represent the following relationships and trends:

- The correlation between Flight Number and Payload Mass
- The association between Flight Number and Launch Site
- The interplay between Payload Mass and Launch Site
- The connection between Orbit Type and Success Rate
- The alignment of Flight Number with Orbit Type
- The juxtaposition of Payload Mass with Orbit Type
- The yearly evolution of the Success Rate

Scatter plots serve as a means to uncover potential relationships between variables, which can subsequently inform the construction of machine learning models.

Bar charts, on the other hand, facilitate comparisons between distinct categories. These charts aim to elucidate the connection between specific categories under consideration and a quantified value.

Lastly, **line charts** are instrumental in revealing temporal trends within the data, highlighting changes and patterns over time.

EDA with SQL

The following SQL queries were as follows:

- Displaying the names of distinct launch facilities involved in space missions.
- Presenting five records where launch sites commence with the string 'CCA.'
- Exhibiting the cumulative payload mass transported by boosters launched under NASA's CRS program.
- Showcasing the average payload mass carried by booster version F9 v1.1.
- Enumerating the date of the inaugural successful ground pad landing outcome.
- Listing the names of boosters that achieved success on drone ships, with payload masses ranging from 4000 to 6000.
- Providing a breakdown of the total counts for successful and unsuccessful mission outcomes.
- Enumerating the booster versions that accommodated the highest payload masses.
- Cataloging failed landing results on drone ships, along with their associated booster versions and launch site names, during the months of 2015.
- Ranking the frequency of landing outcomes, including Failures (drone ship) and Successes (ground pad), between June 4, 2010, and March 20, 2017, in descending order.

Build an Interactive Map with Folium

Markers for All Launch Sites:

- Incorporated a Marker featuring a Circle, Popup Label, and Text Label for NASA Johnson Space Center, utilizing its latitude and longitude coordinates as the starting point.
- Included Markers with Circles, Popup Labels, and Text Labels for all Launch Sites, employing their latitude and longitude coordinates to visually represent their geographical positions in relation to the Equator and coastlines.

Colored Markers for Launch Outcomes at Each Launch Site:

- Introduced colored Markers, denoting successful launches in green and unsuccessful ones in red, utilizing Marker Clusters for easy identification of launch sites with relatively high success rates.

Distances from Launch Sites to Nearby Locations:

- Integrated colored Lines to illustrate the distances between Launch Site KSC LC-39A (as an illustrative example) and nearby features such as railways, highways, coastlines, and the closest city.

Build a Dashboard with Plotly Dash

Dropdown Menu for Launch Site Selection:

- Introduced a dropdown menu to facilitate the selection of Launch Sites.

Pie Chart Displaying Successful Launches:

- Implemented a pie chart to visually represent the cumulative count of successful launches across all sites and to compare the counts of Success versus Failure when a specific Launch Site is chosen.

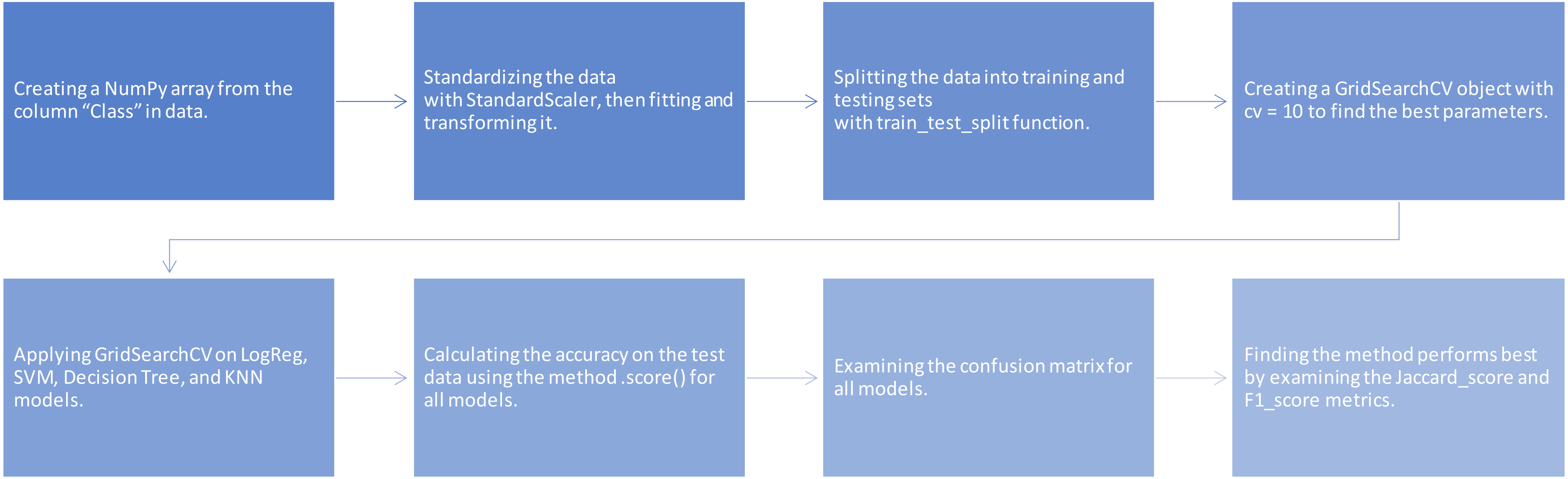
Payload Mass Range Slider:

- Integrated a slider for choosing the range of Payload masses.

Scatter Chart Illustrating Payload Mass vs. Launch Success Rate for Various Booster Versions:

- Included a scatter chart to graphically depict the relationship between Payload Mass and Launch Success across different Booster Versions.

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.

Creating a GridSearchCV object with cv = 10 to find the best parameters.

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models.

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

Examining the confusion matrix for all models.

Finding the method performs best by examining the Jaccard_score and F1_score metrics.

SOURCE: https://github.com/EIDSTER/coursera/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

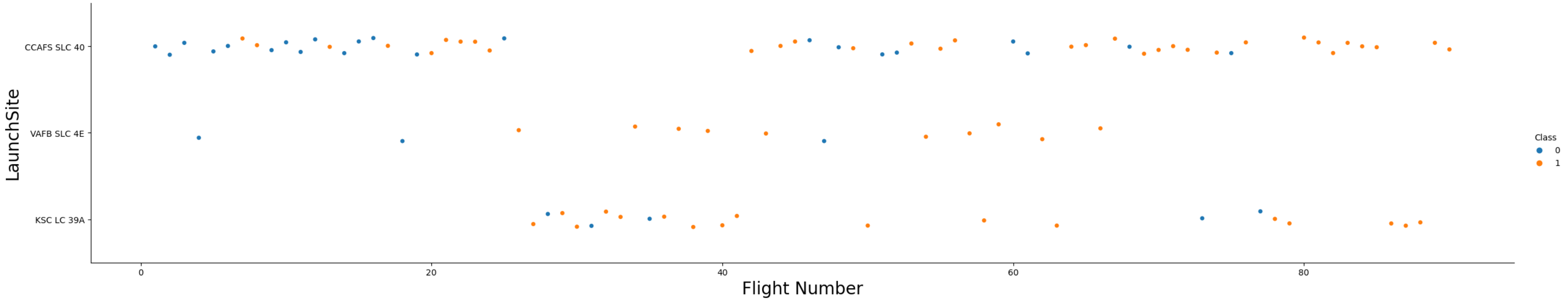
- SpaceX operates from four distinct launch facilities.
- Initial launches were conducted both by SpaceX itself and in collaboration with NASA.
- The average payload capacity of the F9 v1.1 booster stands at 2,535 kg.
- The first successful landing occurred in 2015, five years after the inaugural launch.
- Numerous Falcon 9 booster versions achieved successful drone ship landings with payloads exceeding the average.
- Nearly 100% of mission outcomes resulted in success.
- Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landings on drone ships in 2015.
- Over time, the number of successful landing outcomes improved.
- The use of interactive analytics revealed that launch sites were typically located in secure areas near bodies of water, such as the sea, and possessed robust logistical infrastructure.
- Predictive Analysis determined that the Decision Tree Classifier is the optimal model for forecasting successful landings.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

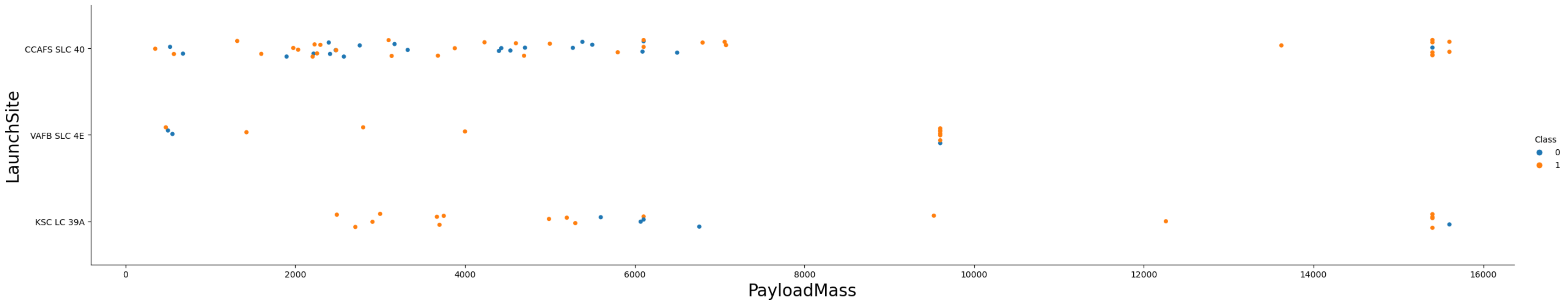
- The initial flights ended in failure, while the most recent ones achieved success.
- Approximately half of all launches originate from the CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A exhibit superior success rates.
- It is reasonable to infer that each successive launch enjoys an improved likelihood of success.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Payload vs. Launch Site

- Across all launch sites, there is a positive correlation between payload mass and success rate.
- A majority of launches with a payload mass exceeding 7000 kg achieved success.
- KSC LC 39A boasts a perfect success rate for payload masses below 5500 kg as well.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Success Rate vs. Orbit Type

Orbits with a flawless 100% success rate:

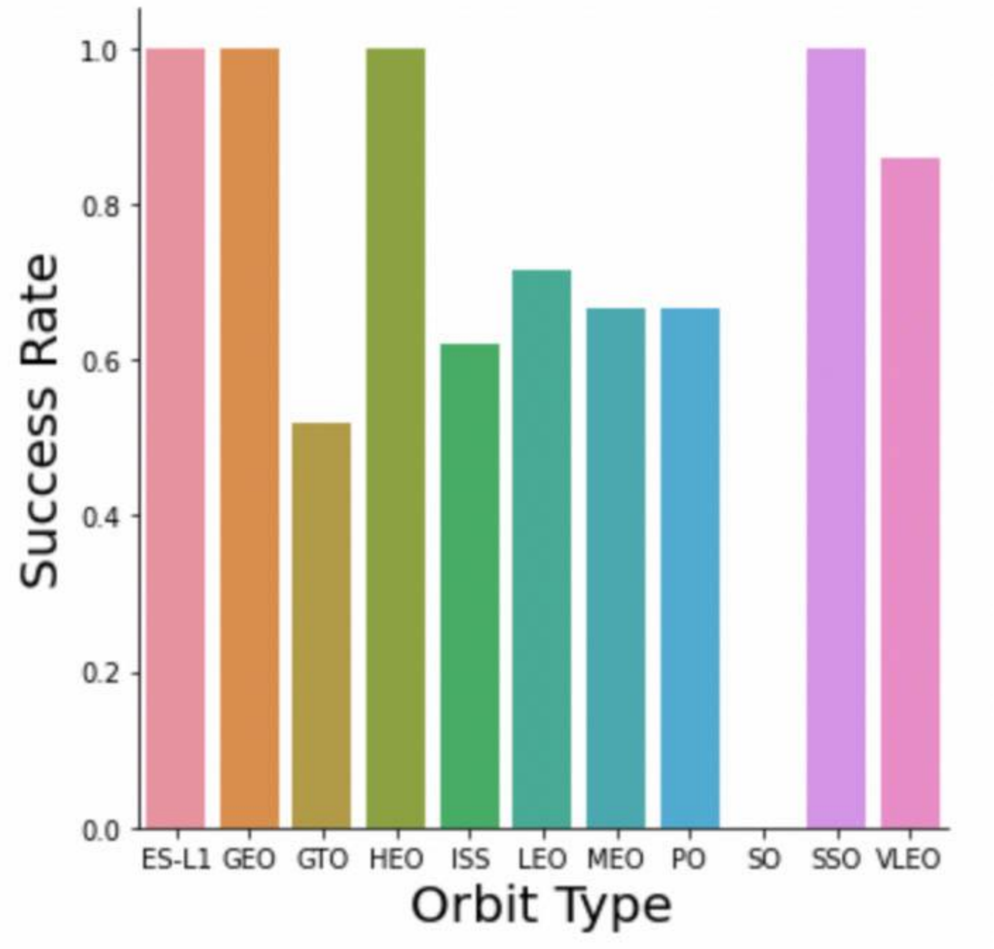
- ES-L1, GEO, HEO, SSO

Orbits with an absolute 0% success rate:

- SO

Orbits with success rates ranging from 50% to 85%:

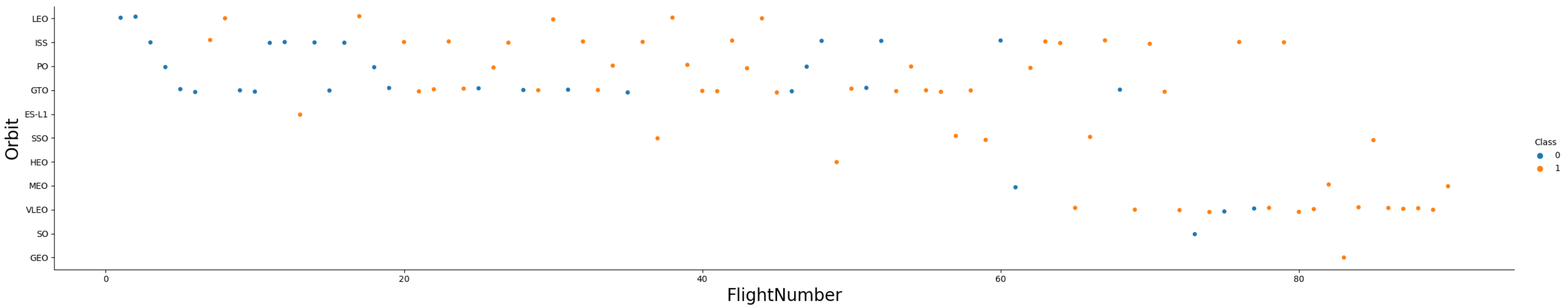
- GTO, ISS, LEO, MEO, PO



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Flight Number vs. Orbit Type

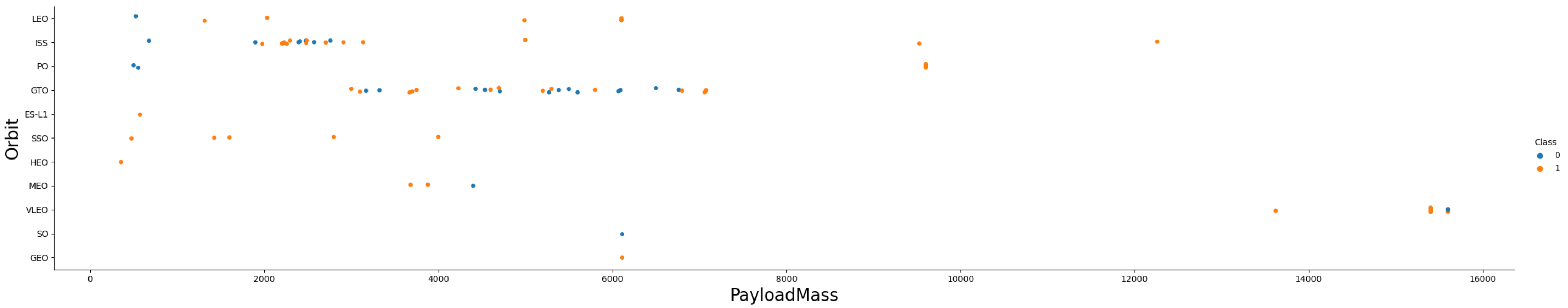
- Within the LEO orbit, there appears to be a correlation between success and the number of flights, while in the GTO orbit, no such relationship seems evident.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Payload vs. Orbit Type

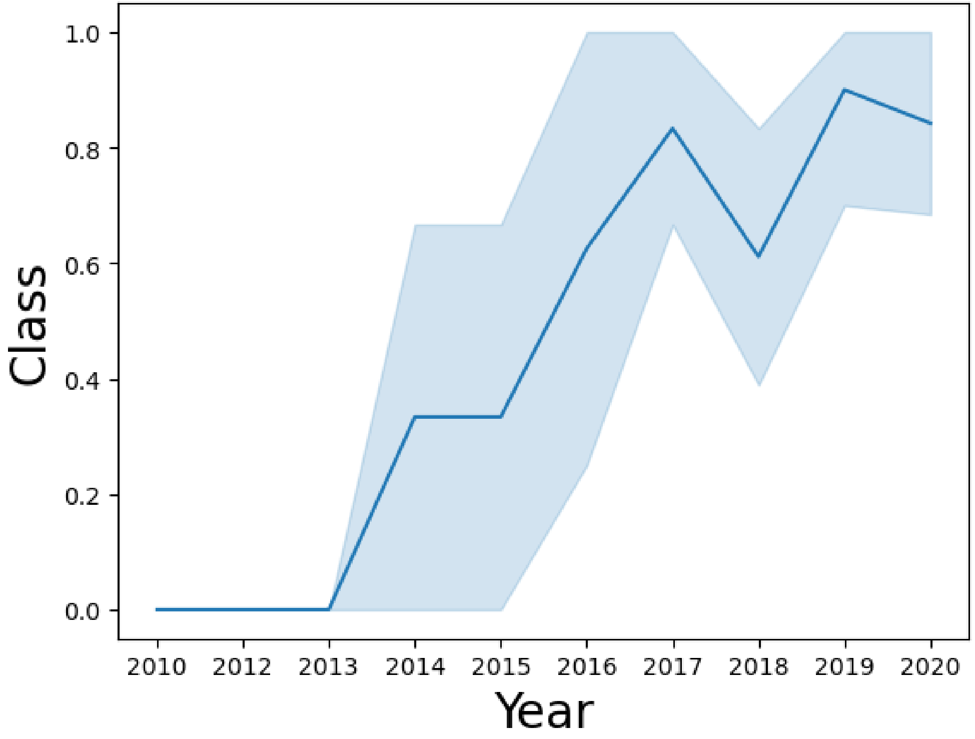
- Larger payloads exhibit a detrimental effect on GTO orbits but a favorable impact on GTO and Polar LEO-ISS orbits.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Launch Success Yearly Trend

- Success increased from 2013 to 2020.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

All Launch Site Names

- A query of the data yielded the following distinct Lauch Sites.

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

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Launch Site Names Begin with 'CCA'

- The following query yielded records from the CCAFS LC-40 launch site:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Payload Carried by Boosters from NASA CRS:

- 45,596 Kg

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Average Payload Mass by F9 v1.1

- 2,535 Kg

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

First Successful Ground Landing Date

- December 22nd, 2015

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Successful Drone Ship Landing with Payload Between 4000 and 6000

- Booster names which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Number of Successful and Failure Mission Outcomes

- The following query shows total number of successful, failure and no attempt mission outcomes:

Landing_Outcome	COUNT(Landing_Outcome)
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

2015 Launch Records

- Failed Drone Ship landing outcomes including their booster versions, and launch site names in 2015:

Booster_Version	Launch_Site	Month	Year
F9 v1.1 B1012	CCAFS LC-40	5	2015
F9 v1.1 B1015	CCAFS LC-40	5	2015

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

- Rank 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

Landing_Outcome	Lcount
No attempt	9
Failure (drone ship)	5
Success (drone ship)	4
Controlled (ocean)	3
Uncontrolled (ocean)	2
Success (ground pad)	2
Precluded (drone ship)	1

SOURCE: https://github.com/EIDSTER/coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

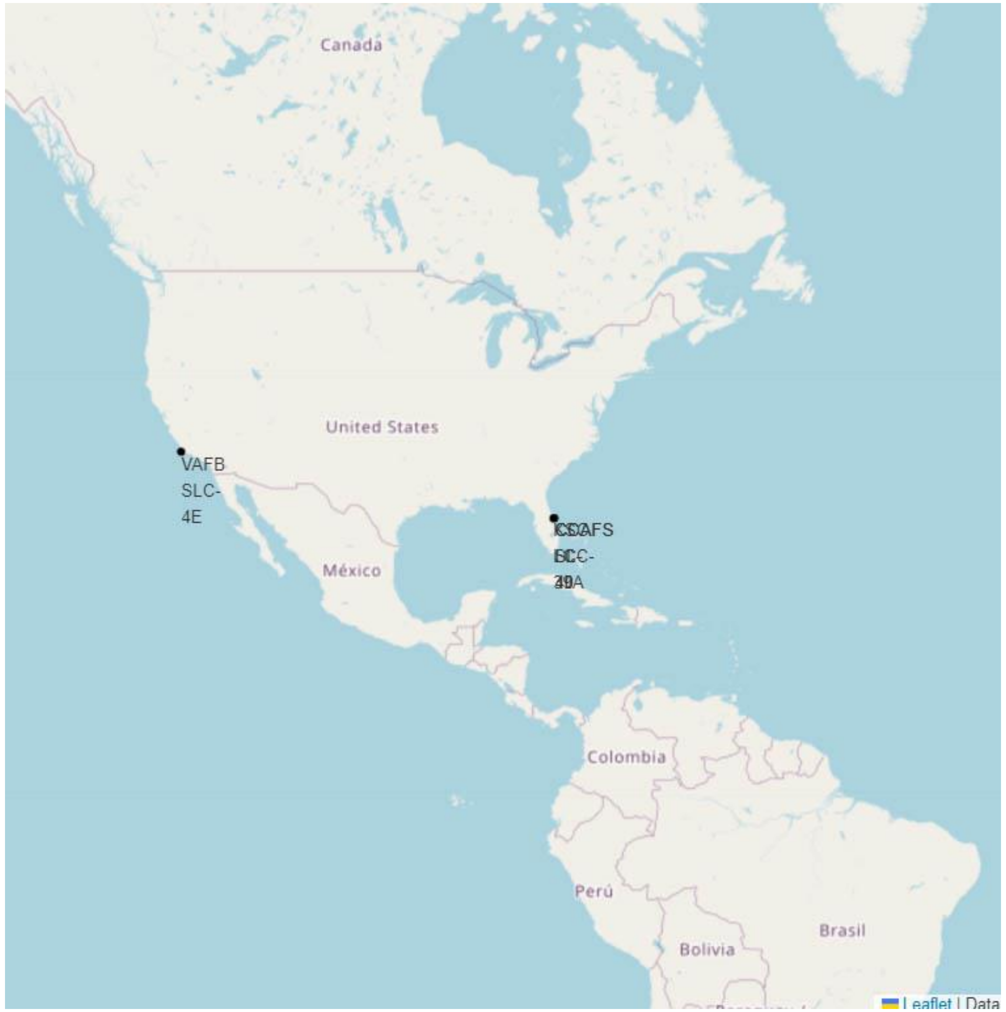
Section 3

Launch Sites Proximities Analysis

Global Map of All Launch sites.

- The majority of launch sites are situated in close proximity to the Earth's Equator. Notably, the Equator is a region where the Earth's surface moves at an astonishing speed of 1670 kilometers per hour. When a spacecraft is launched from the Equator, it not only ascends into space but also maintains the same velocity it had before liftoff due to the principle of inertia. This retained speed is essential for the spacecraft to sustain the necessary velocity for orbital travel.

- All launch facilities are strategically positioned near coastlines. When rockets are launched over the ocean, it serves as a risk mitigation measure to minimize the possibility of any debris falling or exploding near populated areas.

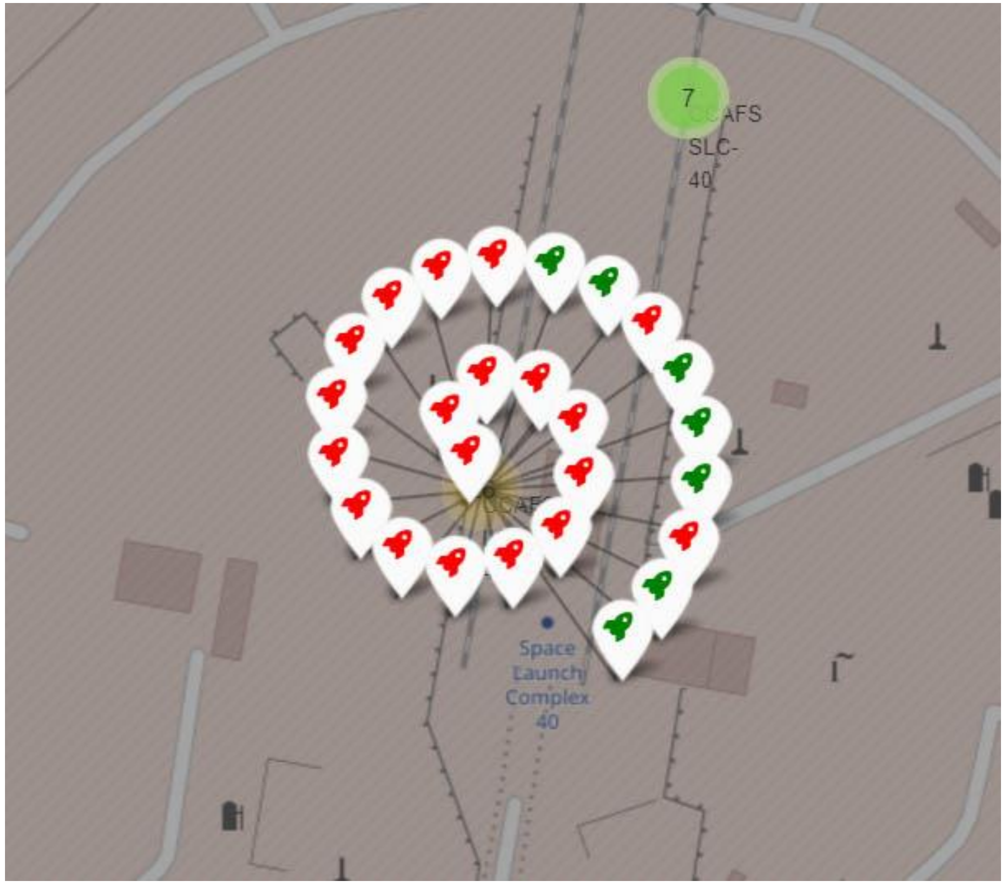


SOURCE: https://github.com/EIDSTER/coursera/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Color-Labeled Launch Records

- By observing the markers color-coded for success outcomes, we can readily discern launch sites with notably elevated success rates.

- Green Markers indicate a successful launch.
- Red Markers indicate a failed launch.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

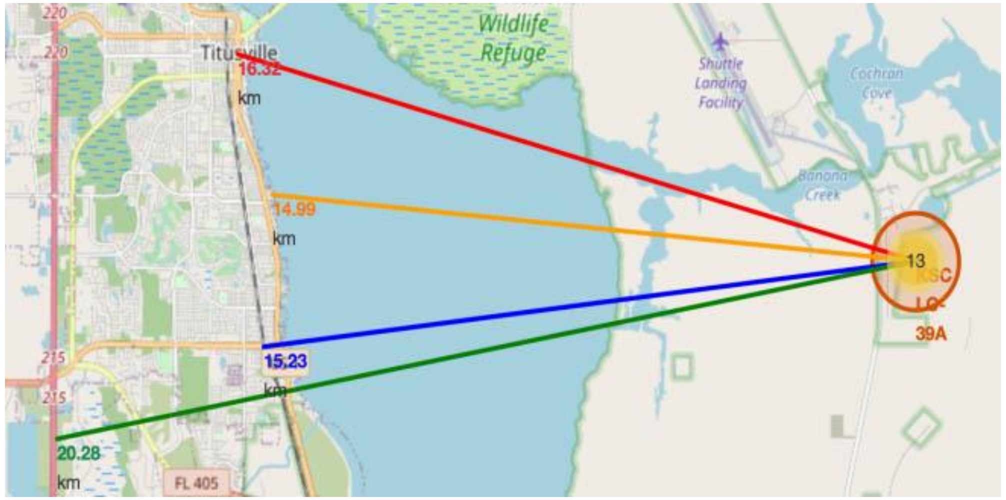
Launch Site KSC LC-39A Distance to Important Locations

Through a visual examination of launch site KSC LC-39A, we can discern the following:

- Its proximity to a railway is approximately 15.23 kilometers.
- It is situated relatively near a highway at a distance of about 20.28 kilometers.
- The coastline is relatively close at a distance of approximately 14.99 kilometers.

Furthermore, launch site KSC LC-39A is also in close proximity to its nearest city, Titusville, at a distance of roughly 16.32 kilometers.

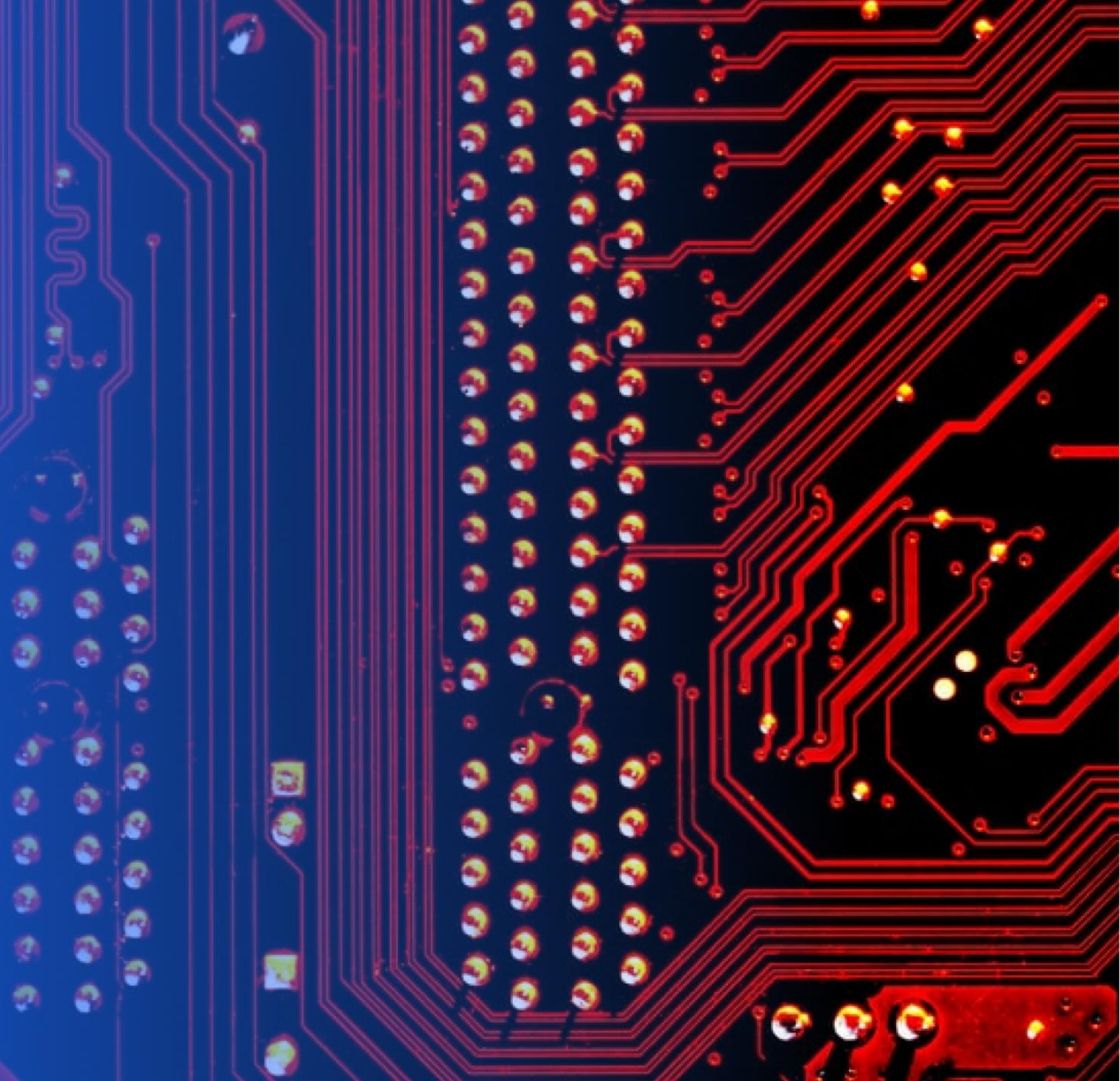
It's worth noting that a malfunctioning rocket, given its high velocity, can traverse distances of 15-20 kilometers in mere seconds. Such circumstances could pose potential hazards to populated areas.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

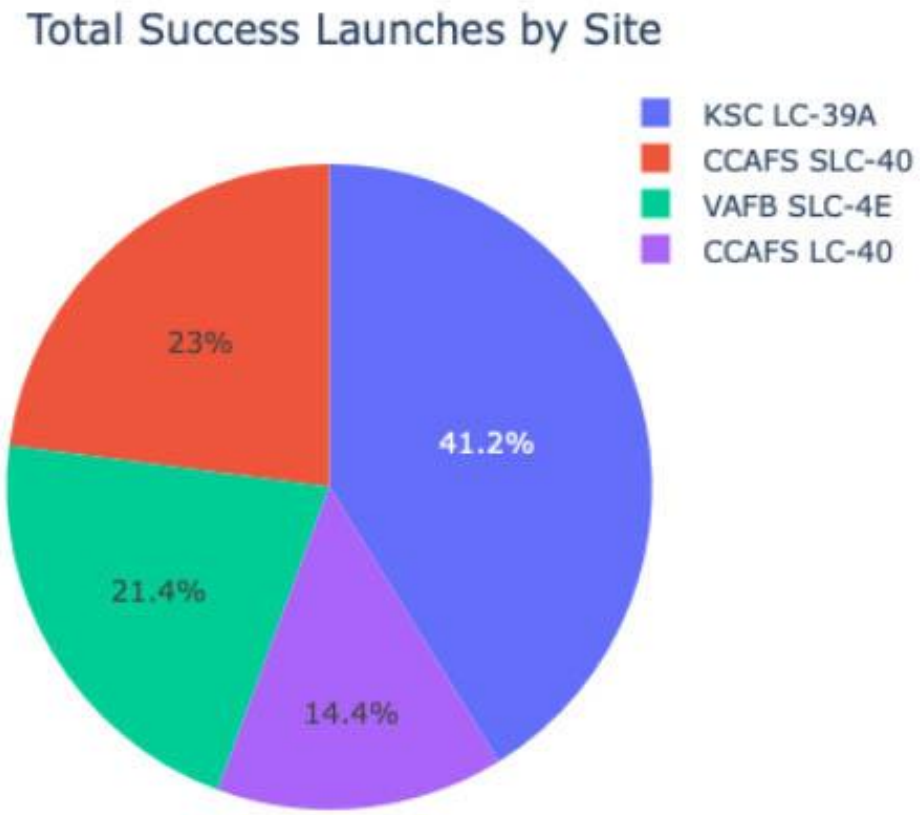
Section 4

Build a Dashboard with Plotly Dash



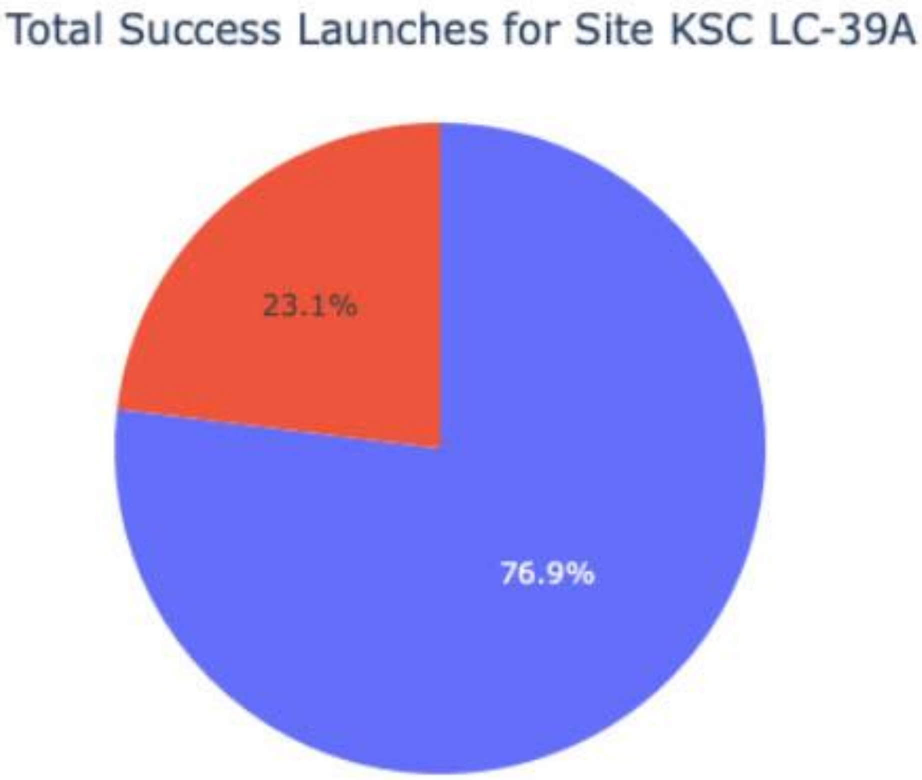
Launch Success Breakdown

- The chart provides a clear indication that among all the launch sites, KSC LC-39A stands out with the highest number of successful launches.



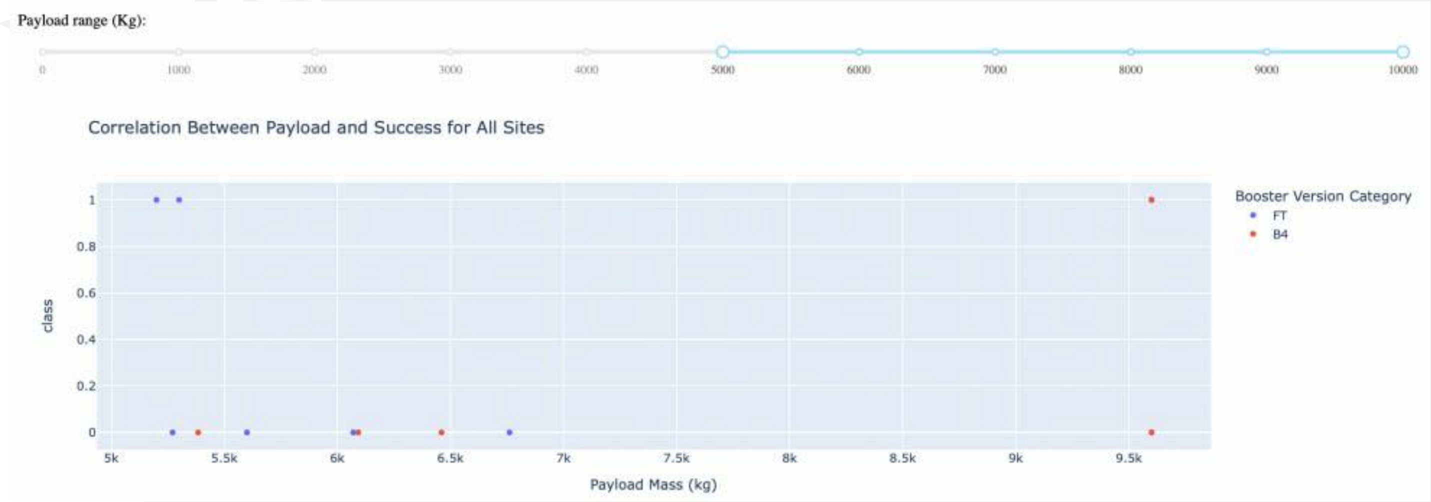
Launch site success ratios

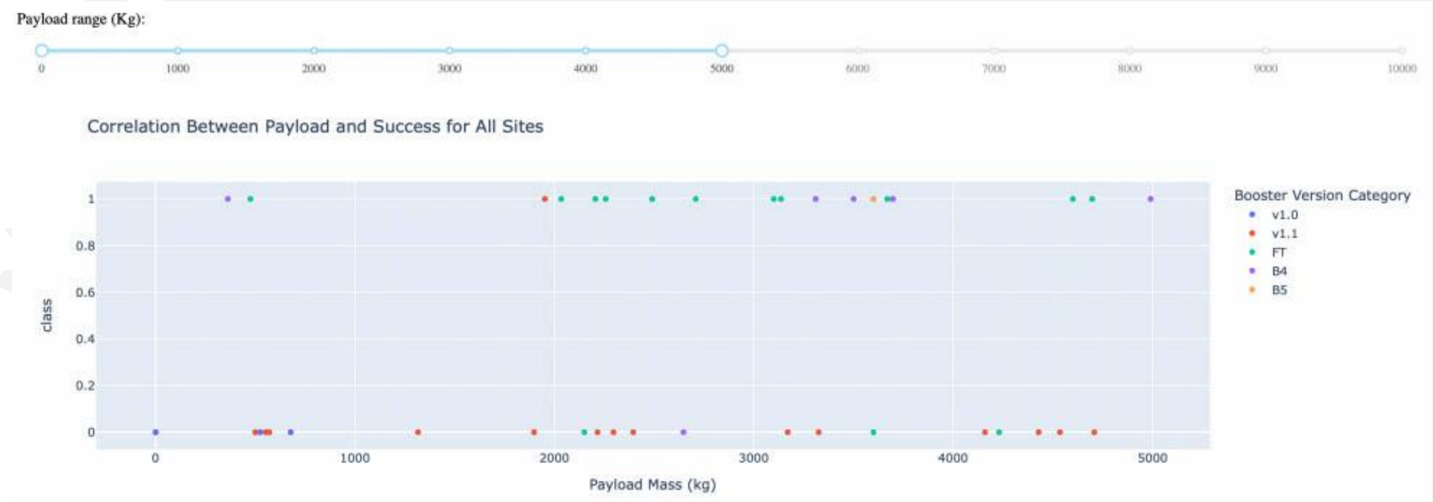
- KSC LC-39A boasts the highest launch success rate at 76.9%, recording 10 successful landings and only 3 failures.



Payload Mass vs Launch Outcomes

- The charts reveal that payloads ranging from 2000 to 5500 kg exhibit the most favorable success rates.



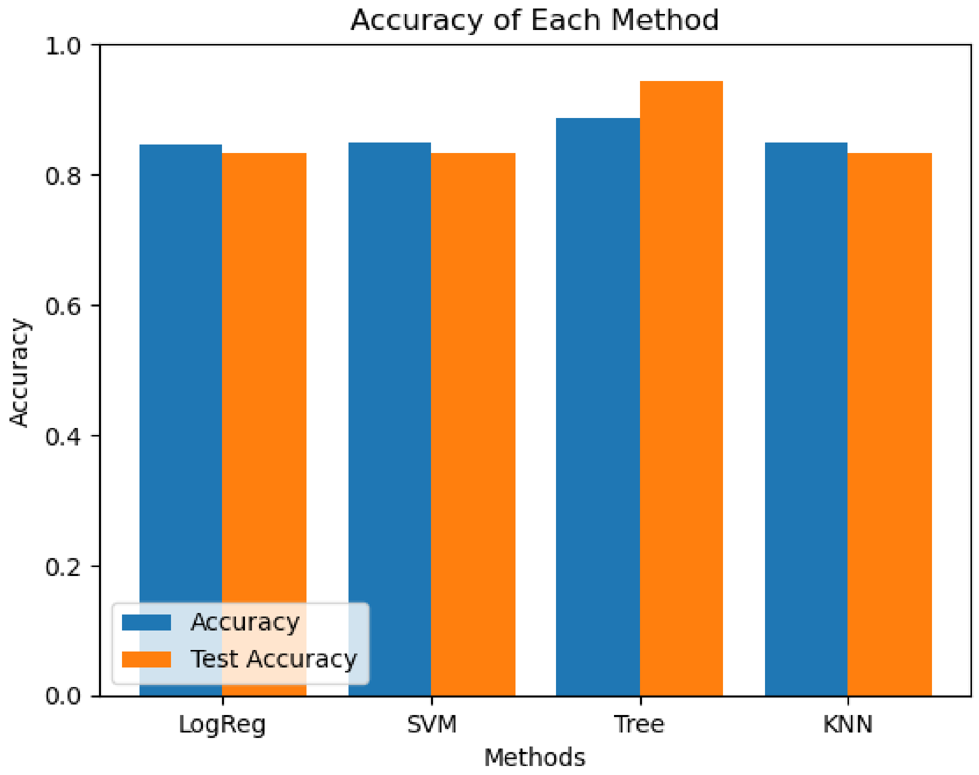


Section 5

Predictive Analysis (Classification)

Classification Accuracy

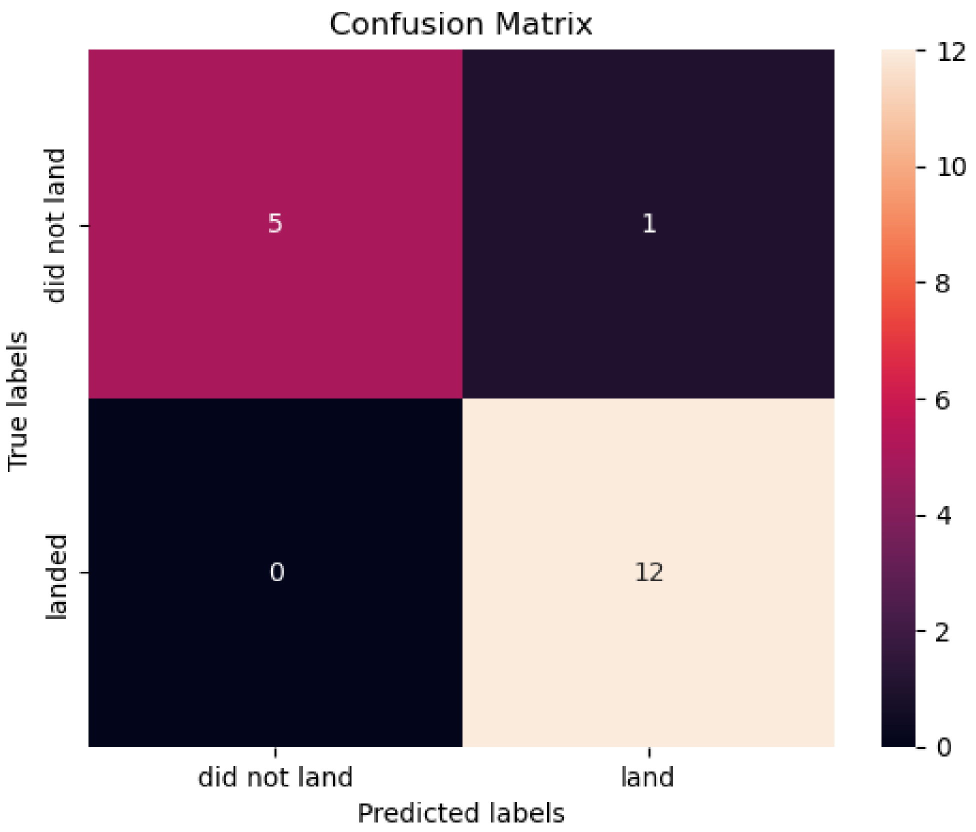
- Four classification models were tested, and their accuracy results are displayed alongside:
- The Decision Tree Classifier exhibits the highest classification accuracy at 88.57% with a test accuracy of 94.44%.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Confusion Matrix

- In the confusion matrix of the Tree model, which performed the best, only 1 false positive is observed.



SOURCE: https://github.com/EIDSTER/coursera/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Conclusions

- The Tree Model emerges as the most suitable algorithm for this dataset.
- Launches with lower payload masses yield better outcomes compared to those with larger payloads.
- The majority of launch sites are situated in proximity to the Equator, and all sites are located very close to coastlines.
- The success rate of launches shows an upward trend over the years.
- KSC LC-39A maintains the highest success rate among all the launch sites.
- Orbits ES-L1, GEO, HEO, and SSO consistently achieve a perfect 100% success rate.
- The analysis involved various data sources, refining conclusions throughout the process, leading to the following key findings:
- KSC LC-39A emerges as the top-performing launch site.
- Launches with payloads exceeding 7,000 kg are associated with lower risks.
- While most mission outcomes are successful, the rate of successful landings appears to improve over time, reflecting advancements in processes and rocket technology.
- Utilizing the Decision Tree Classifier can aid in predicting successful landings and potentially increasing profits.

Appendix

The following python packages were used for analysis:

- **Pandas** is a software library written for the Python programming language for data manipulation and analysis.
- **NumPy** is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays
- **Matplotlib** is a plotting library for python and pyplot gives us a MatLab like plotting framework. We will use this in our plotter function to plot data.
- **Folium** is a geography based plotting package
- **Seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics
- **Sklearn** is a machine learning package with multitudes of pre-processing and train test modules

Thank you!