



Leveraging Data Science for Space Race Success

Jigar Patel
10/20/2024

OUTLINE



- Executive Summary
- Introduction
- Methodology
- Results
 - Visualization – Charts
 - Dashboard
- Discussion
 - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY



- This project leverages publicly available data from the SpaceX API and Wikipedia to analyze and predict the outcomes of Falcon 9 rocket launches. The primary goal is to assist a rival rocket launch company in predicting the success or failure of SpaceX's first-stage landings using advanced data science techniques.
- The data includes flight number, launch date, payload mass, orbit type, launch site, mission outcomes, and other relevant variables. Key steps in the project included:
 - Data collection and wrangling, focusing on extracting launch outcome information as the dependent variable for predictive modeling.
 - Exploratory data analysis through SQL queries, static visualizations, interactive maps, and dashboards to uncover insights.
 - Implementation of various machine learning models, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN), to predict landing success.
- Initial results show that Logistic Regression, SVM, and KNN models performed similarly well in predicting Falcon 9 first-stage landing outcomes. This analysis provides a foundation for future improvements in prediction accuracy and decision-making in the space race competition.

INTRODUCTION



- SpaceX has revolutionized the commercial space industry by drastically reducing the cost of space travel, largely due to its innovative approach to reusing the first stage of its Falcon 9 rockets. While other providers charge upwards of \$165 million per launch, SpaceX advertises launches for \$62 million, with much of the savings attributed to the reusability of the rocket's first stage.
- The ability to predict whether the first stage will land successfully is critical for determining the overall cost and future viability of each launch. By leveraging publicly available data and advanced machine learning models, this project aims to predict whether SpaceX will be able to reuse the first stage of the Falcon 9 rocket after each launch. Accurate predictions could provide valuable insights for competitors and stakeholders in the space industry.
- Key Questions to be Answered:
 - How do factors such as payload mass, launch site, number of flights, and orbit type influence the success of the Falcon 9 first-stage landing?
 - Has the rate of successful first-stage landings improved over time?
 - Which machine learning algorithm is most effective for predicting the binary outcome of a first-stage landing (success or failure)?

METHODOLOGY



- The project employed a multi-step data science process to analyze and predict the success of SpaceX Falcon 9 rocket landings:
- Data Collection: Data for Falcon9 launches was collected from
 - SpaceX API
 - web scraping of Wikipedia launch tables
- Data Wrangling: Data was cleaned and pre-processed, including handling missing values, filtering, and applying One Hot Encoding to prepare the dataset for binary classification models.
- Exploratory Data Analysis (EDA): Visualization and SQL queries were used to explore the data, uncover patterns, and generate initial insights. Interactive visualizations were created using Folium and Plotly Dash to provide a dynamic view of launch locations and outcomes.
- Predictive Analysis: Classification models were built, tuned, and evaluated using various algorithms to predict the success of first-stage rocket landings. These models included logistic regression, SVM, decision trees, and k-Nearest Neighbors to ensure 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 to ensure a comprehensive dataset for detailed analysis. By combining these two methods, we were able to gather complete and accurate information on SpaceX launches.

➤ **SpaceX REST API:**

The following key columns were obtained through API requests, offering detailed technical data on each launch:

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

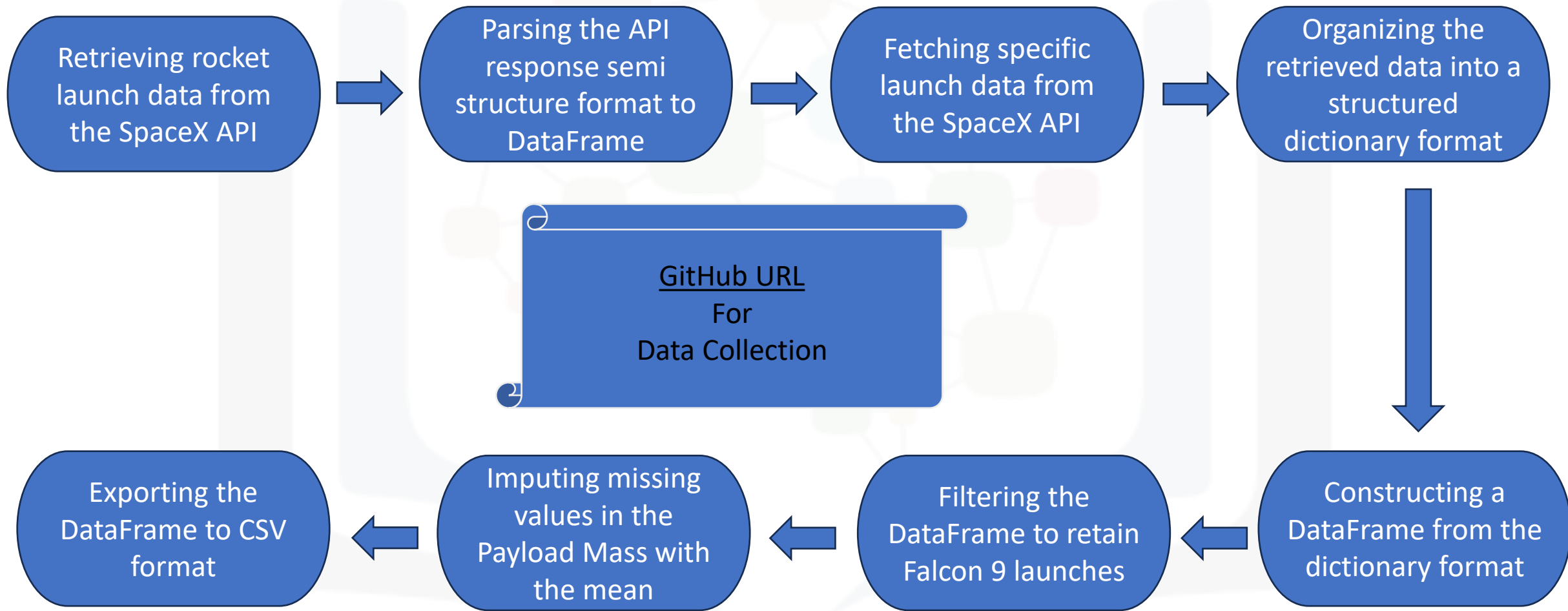
➤ **Wikipedia Web Scraping:**

To fill any gaps and provide historical context, additional columns were extracted via web scraping, adding valuable insights for our analysis:

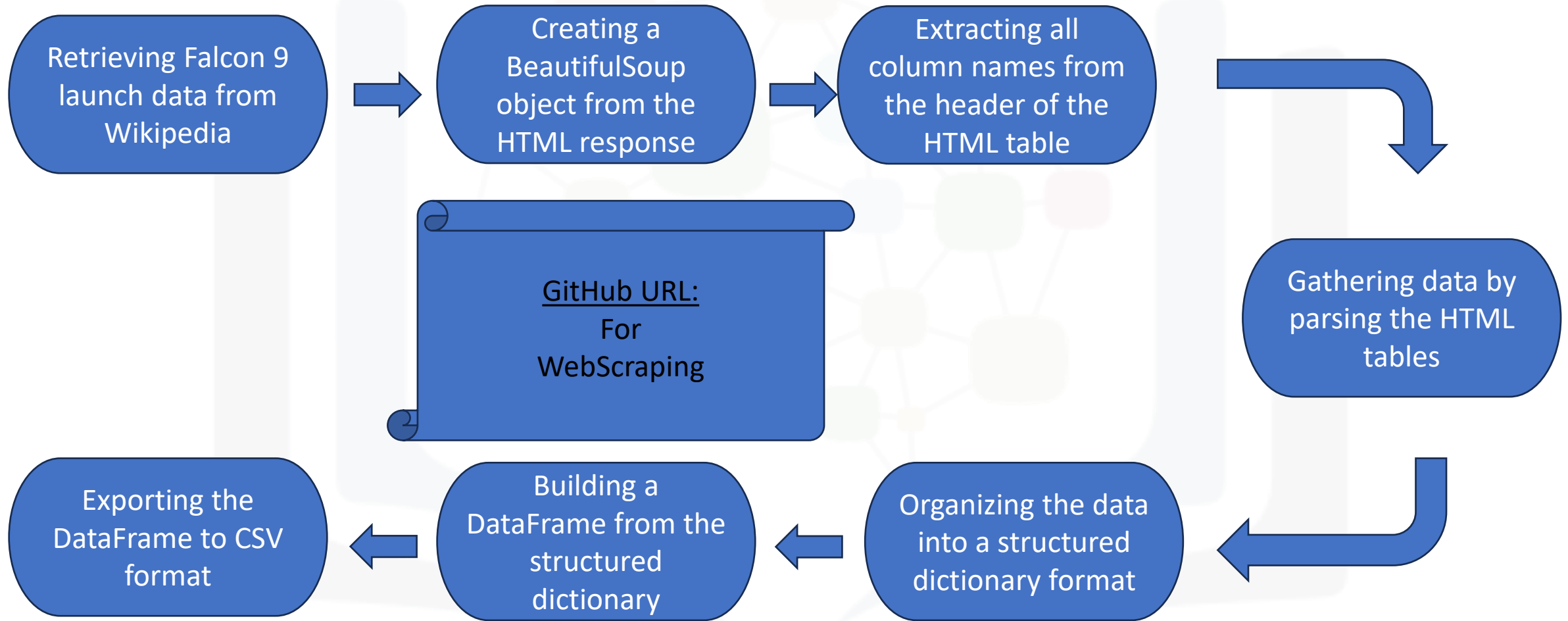
- ❖ Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer
- ❖ Launch Outcome, Version Booster, Booster Landing, Date, Time

By combining both sources, we ensured the dataset contained all critical variables necessary for a more in-depth exploration and predictive analysis of SpaceX Falcon 9 launches.

Data Collection from SpaceX API



Data Collection Using WebScraping



Data Wrangling

In the dataset, there are various scenarios where the booster did not land successfully. For instance, a landing may have been attempted but failed due to an accident. Specific outcomes include:

- **True Ocean:** The booster successfully landed in a designated ocean region.
- **False Ocean:** The booster unsuccessfully attempted to land in a designated ocean region.
- **True RTLS (Return to Launch Site):** The booster successfully landed on a ground pad.
- **False RTLS:** The booster failed to land on a ground pad.
- **True ASDS (Autonomous Spaceport Drone Ship):** The booster successfully landed on a drone ship.
- **False ASDS:** The booster failed to land on a drone ship.

For the purpose of machine learning, these outcomes were converted into training labels, where "1" represents a successful landing, and "0" represents an unsuccessful landing.

GitHub URL:
For Data
Wrangling

EDA and Visualization

Various charts were plotted to explore relationships between key variables in the dataset:

➤ Scatter Plots:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type

These scatter plots reveal potential relationships between variables. If meaningful patterns exist, these relationships could be leveraged in the machine learning models.

➤ Bar Charts:

- Orbit Type vs. Success Rate

Bar charts provide a clear comparison between discrete categories, helping to illustrate how factors like orbit type correlate with the success rate of launches.

➤ Line Charts:

- Success Rate Yearly Trend

Line charts highlight trends in success rates over time, allowing for the analysis of how the rate of successful landings has evolved.

GitHub URL:
For EDA
And Data
Visualization

SQL

To extract meaningful insights from the dataset, various SQL queries were executed, focusing on key aspects of SpaceX rocket launches:

➤ **Unique Launch Sites:**

- Displayed the names of all unique launch sites used in the space missions.

➤ **Launch Sites Starting with 'CCA':**

- Retrieved 5 records where launch sites begin with the string 'CCA'.

➤ **Total Payload Mass for NASA (CRS):**

- Displayed the total payload mass carried by boosters launched by NASA's CRS missions.

➤ **Average Payload Mass for Booster Version F9 v1.1:**

- Calculated the average payload mass for missions using booster version F9 v1.1.

➤ **First Successful Ground Pad Landing:**

- Listed the date when the first successful landing outcome on a ground pad was achieved.

SQL

➤ Boosters with Drone Ship Success and Specific Payload:

- Listed the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg.

➤ Total Successful and Failed Mission Outcomes:

- Displayed the total number of successful and failed mission outcomes.

➤ Boosters with Maximum Payload Mass:

- Listed the booster versions that carried the maximum payload mass.

➤ Failed Drone Ship Landings in 2015:

- Retrieved the failed landing outcomes on drone ships, along with their booster versions and launch site names, for the year 2015.

➤ Ranking Landing Outcomes:

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

These queries helped uncover vital patterns and trends in SpaceX launch data, which will guide further analysis and model development.

GitHub URL:
Data
Exploration
with SQL

Interactive Visualization with Folium

To enhance the understanding of the geographical distribution and proximity of launch sites, several markers and visual elements were added:

➤ **Markers for All Launch Sites:**

- **NASA Johnson Space Center:** A marker with a circle, popup label, and text label was added using its latitude and longitude as the start location.
- **All Launch Sites:** Markers with circles, popup labels, and text labels were added for all SpaceX launch sites, showing their exact locations. The markers also highlight their proximity to the Equator and coastal areas, which are crucial for successful launches.

➤ **Coloured Markers for Launch Outcomes:**

- Launch outcomes were visualized using coloured markers:
 - **Green Markers** for successful launches
 - **Red Markers** for failed launches
- Marker Clustering was employed to clearly identify which launch sites have higher success rates, providing an easy visual comparison across locations.

Interactive Visualization with Folium

➤ Proximity Distances:

- Colored lines were added to illustrate distances from Launch Site KSC LC-39A to key proximities, such as:
 - Railway
 - Highway
 - Coastline
 - Closest City These visual connections help to showcase the logistical and geographical factors that may influence launch operations and success rates.

This approach provides an intuitive visual exploration of SpaceX's launch site locations, outcomes, and surrounding infrastructure.

GitHub URL:
Interactive
Visualization with
Folium

DASHBOARD with Plotly



To enhance user interaction and provide detailed insights, several interactive elements were implemented:

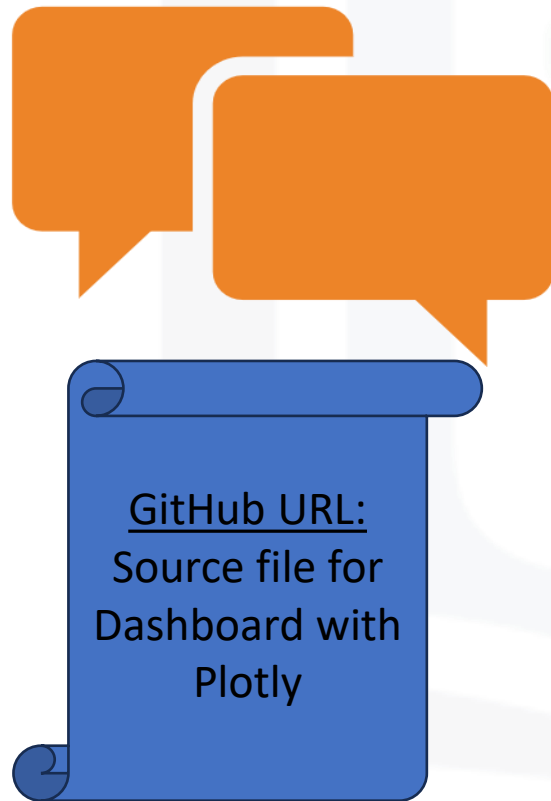
➤ **Launch Sites Dropdown List:**

- A dropdown list was added to allow users to select a specific launch site. This feature enables the exploration of individual sites and their launch outcomes.

➤ **Pie Chart of Success Launches (All Sites or Specific Site):**

- A pie chart was introduced to display the total count of successful launches across all sites.
- When a specific launch site is selected from the dropdown, the pie chart updates to show the breakdown of **Success vs. Failed** launches for that particular site, offering a clear view of the site's performance.

DASHBOARD with Plotly



➤ Payload Mass Range Slider:

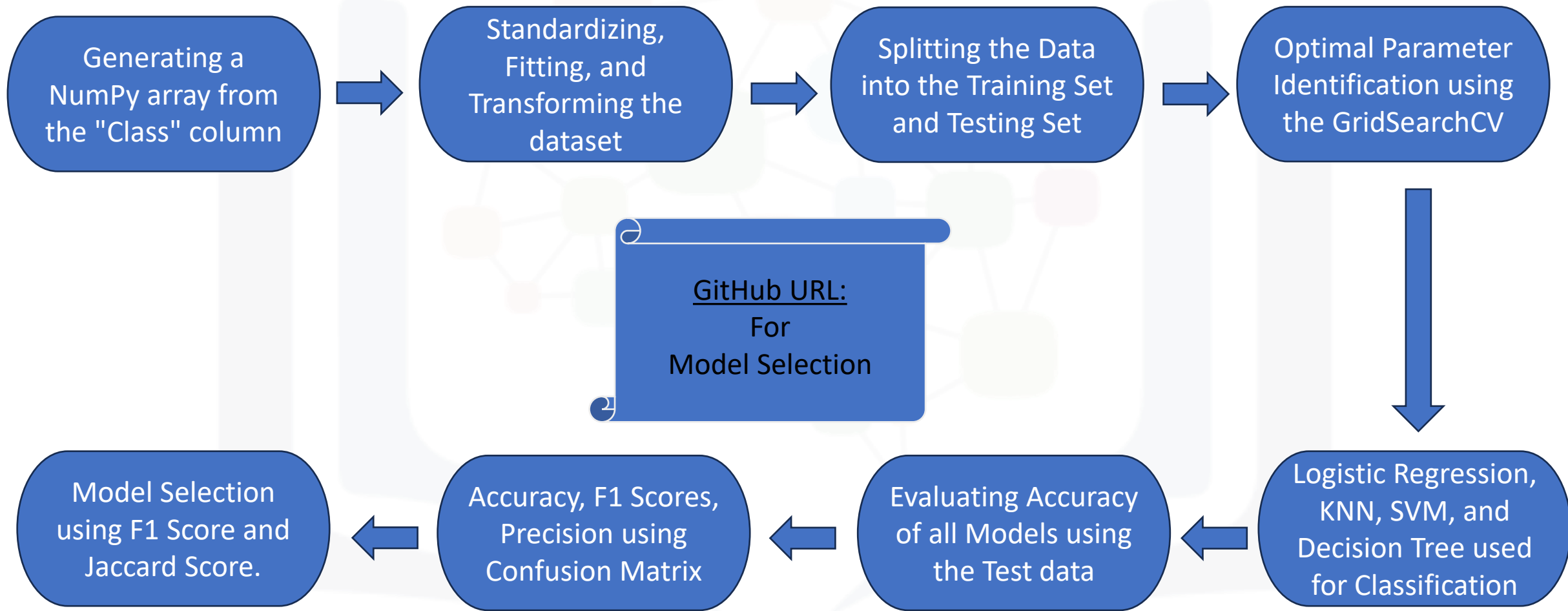
- A slider was implemented to enable users to filter the data based on a specific payload mass range. This allows for targeted analysis of launches within the selected range.

➤ Scatter Chart of Payload Mass vs. Success Rate (Booster Versions):

- A scatter chart was added to visualize the relationship between payload mass and launch success across different booster versions. This chart highlights the correlation between payload size and the success rate of missions, providing valuable insights for further analysis.

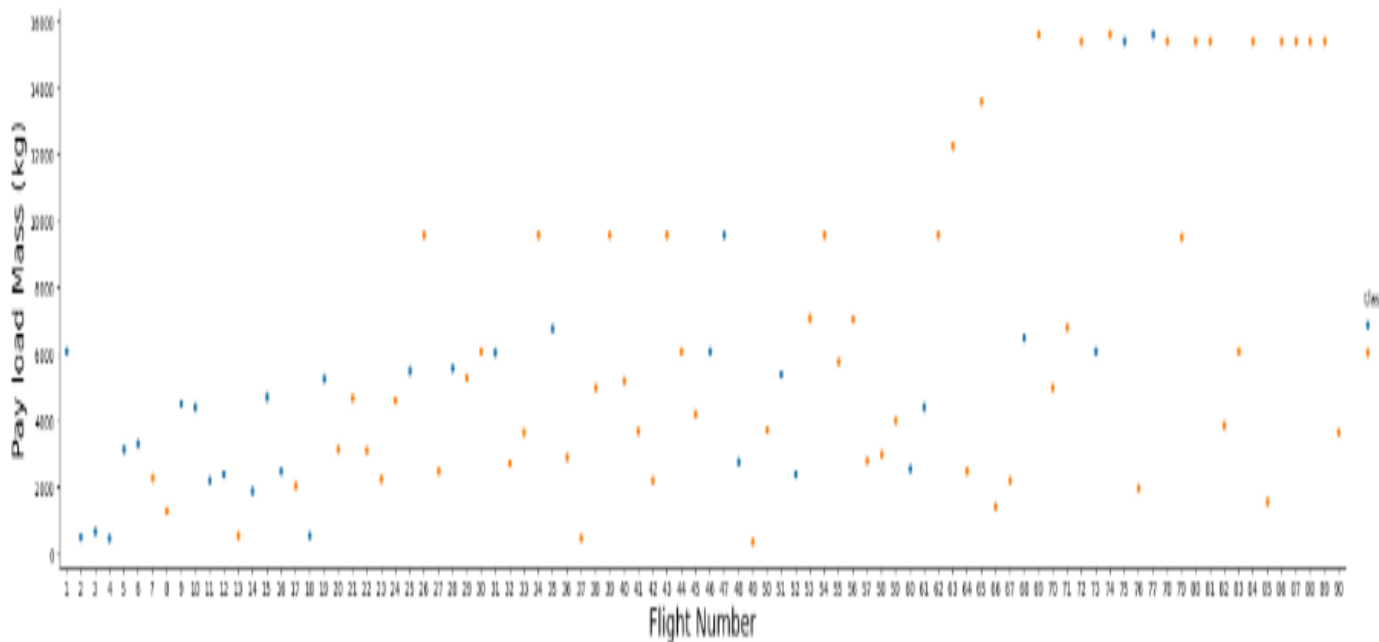
These interactive features allow users to easily navigate and explore the data, offering both a high-level overview and detailed insights into SpaceX's launch performance.

Predictive Analytics



Results (EDA and Visualization)

Flight Number vs Pay Load Mass (in kg)



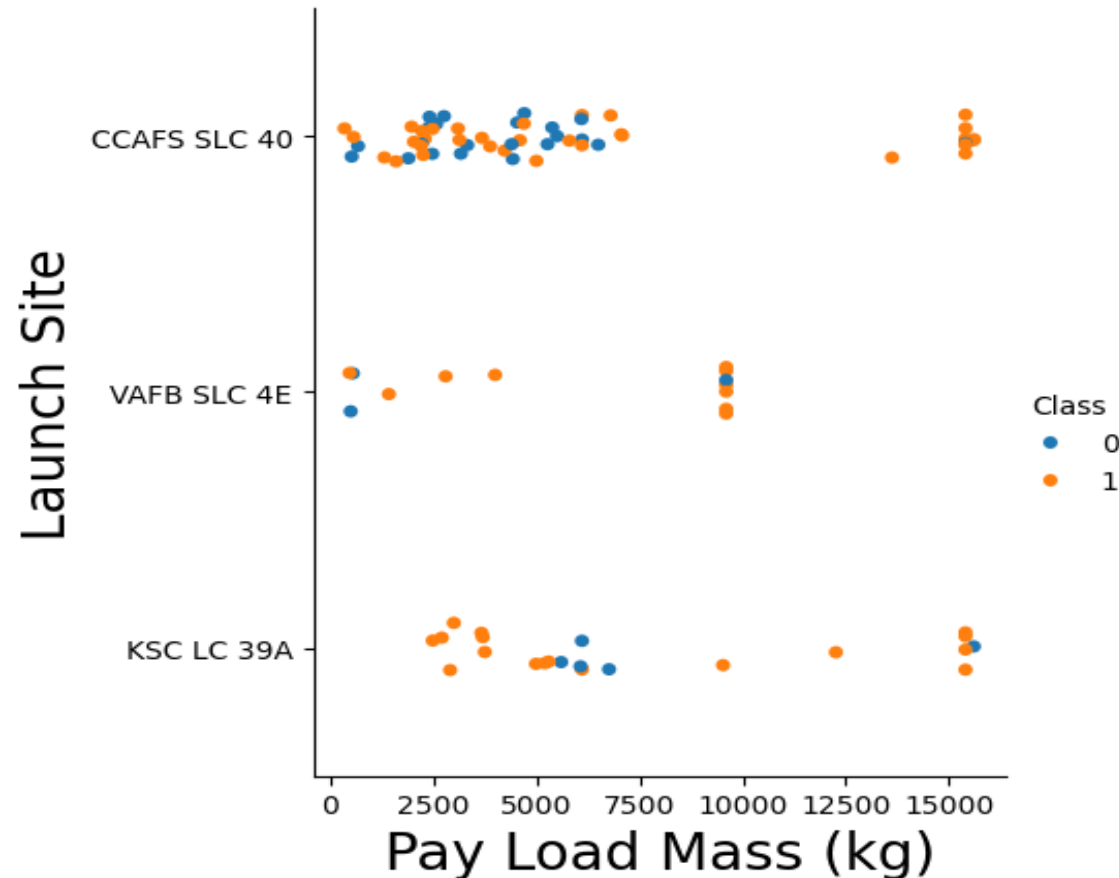
The earliest flights experienced failure, while the most recent flights were all successful.

- The **CCAFS SLC 40** launch site accounts for nearly half of all launches.
- **VAFB SLC 4E** and **KSC LC 39A** have shown higher success rates.

There appears to be a trend of increasing success rates with each new launch.

Results (EDA and Visualization)

Pay Load Mass vs Launch Site

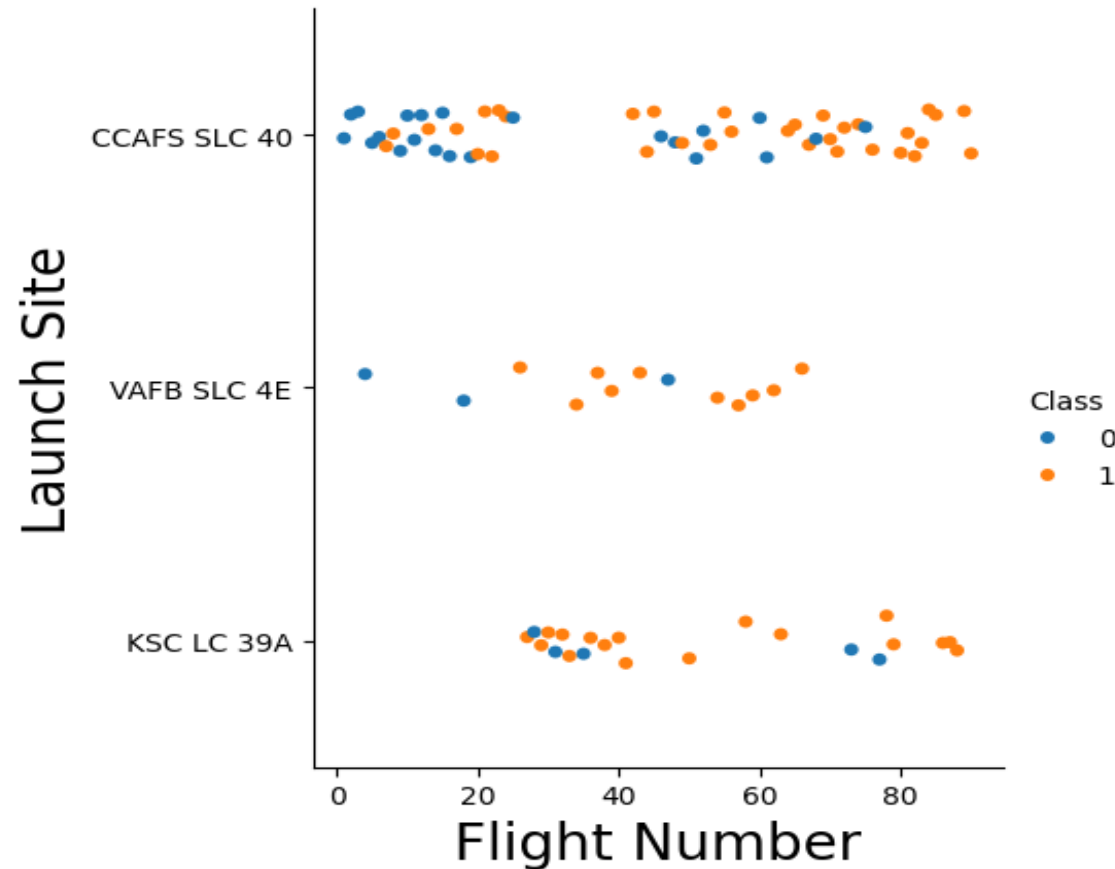


- The Flight with the pay load mass between 7500kg and 16000kg have the highest success rate regardless their launch sites.
- The with pay load mass less than 7500kg has mixed success.
- The launch sites KSC LC 39A and VAFB SLC 4E have higher success in carrying pay load mass less than 7500kg.

There appears to be a trend of increasing success rates with flight carrying higher pay load mass.

Results (EDA and Visualization)

Flight Number vs Launch Site



➤ The Flight numbers between 60 and 90 have the highest success regardless their launch sites.

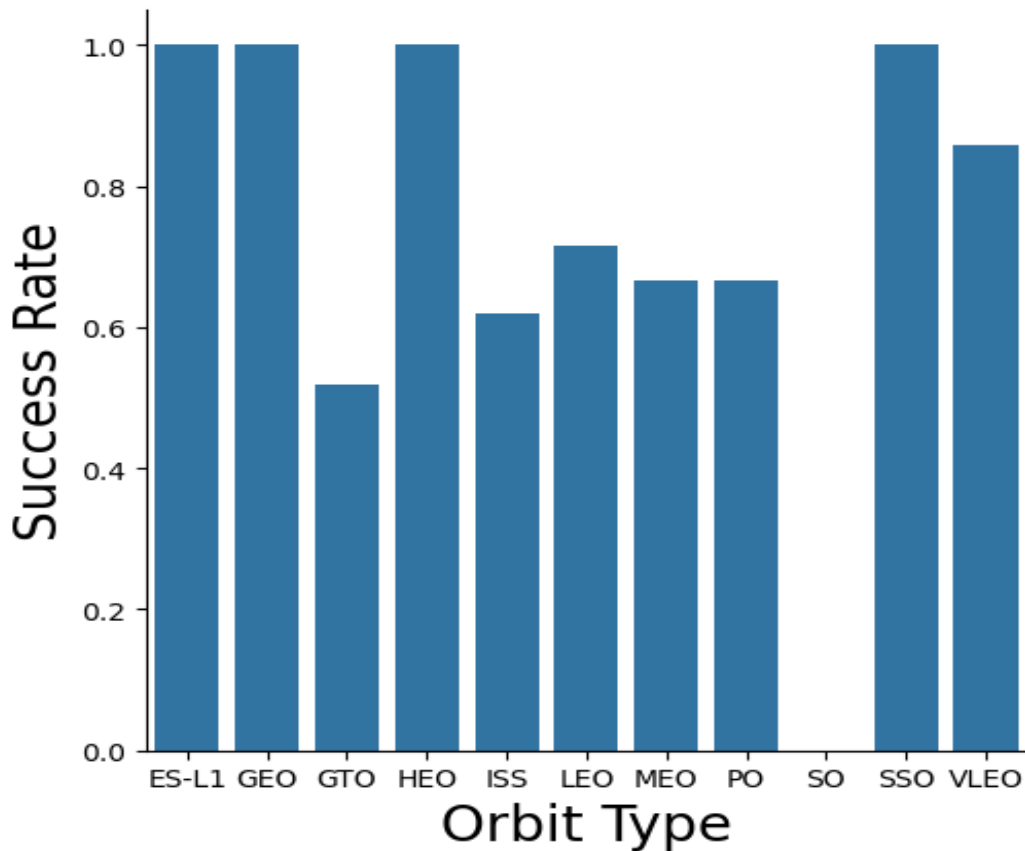
➤ The Flight numbers between 20 and 60 have mid level success.

➤ The Flight numbers less than 20 have higher failure.

There appears to be a trend of increasing success rates with each new launch.

Results (EDA and Visualization)

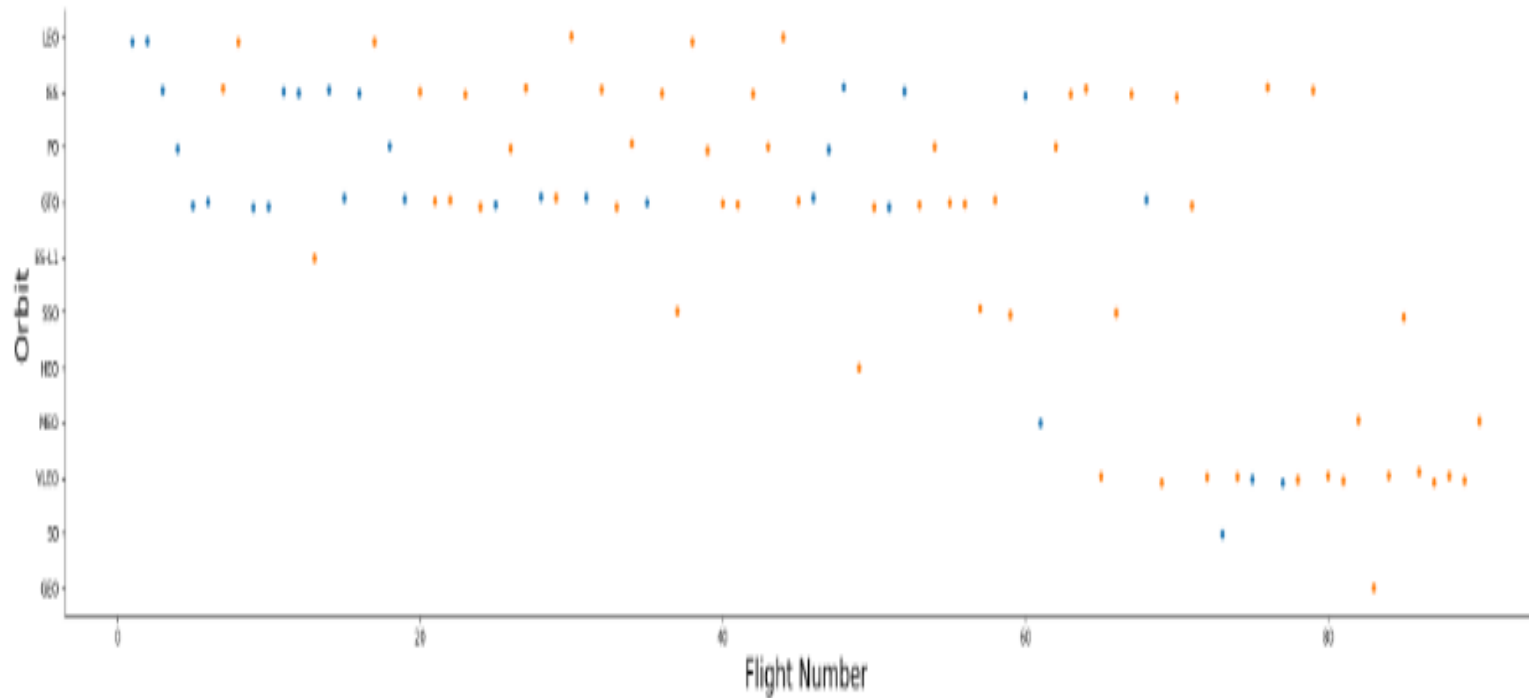
Orbit Type vs Success Rate



- Orbits with a 100% Success Rate:
 - ES-L1, GEO, HEO, SSO
- Orbits with a Success Rate Between 50% and 85%:
 - GTO, ISS, LEO, MEO, PO
- Orbits with a 0% Success Rate:
 - SO

Results (EDA and Visualization)

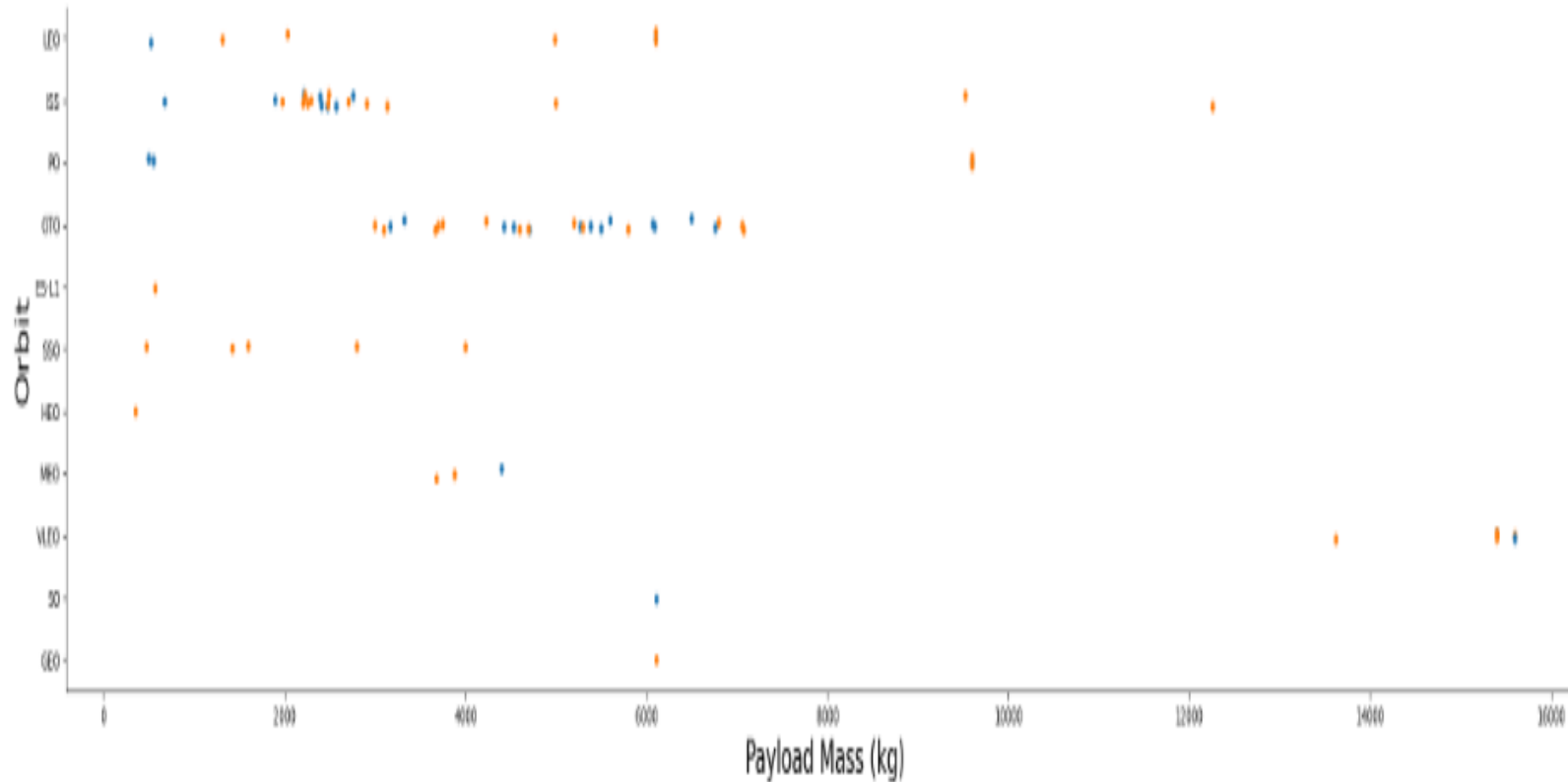
Flight Number Vs Orbit Type



- Newer flight have higher success in due to technical advancements
- In the LEO orbit, success is related to the number of flights.

Results (EDA and Visualization)

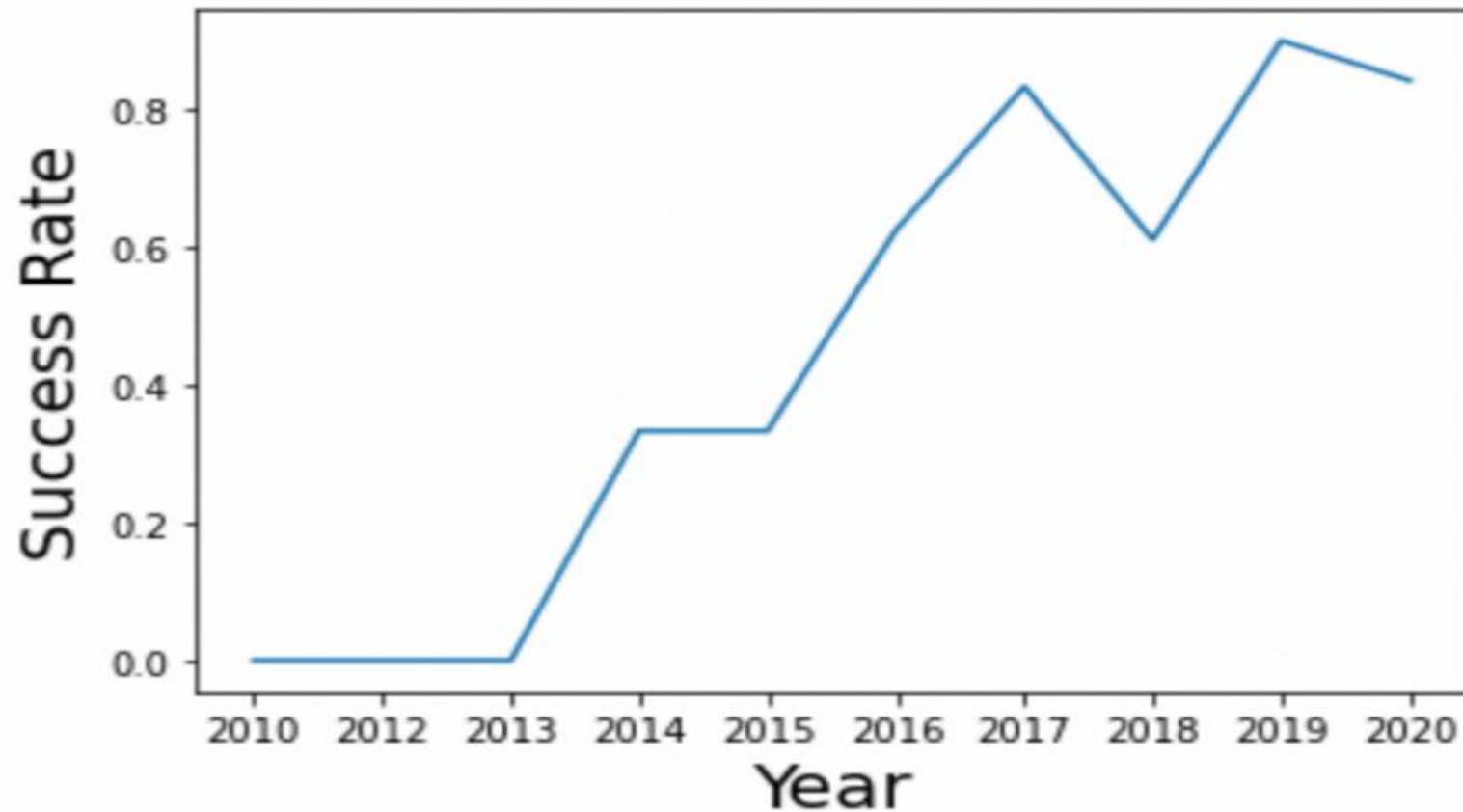
Pay Load Mass Vs Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- With GTO, it is difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Results (EDA and Visualization)

Year Vs Success Rate



➤ Success rate started increasing gradually since 2013 due to advancement in technology.

Results (SQL)

Launch Sites and First 5 Records Using SQL query

In [31]:

```
%sql select distinct Launch_Site from SPACEXTBL
```

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

Done.

Out[31]:

Launch_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

In [32]:

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

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

Done.

Out[32]:

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

Results (SQL)

Total Pay Load Mass and Average Pay Load Mass Using SQL query

```
In [33]: %sql select sum(PAYLOAD_MASS_KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

* sqlite:///my_data1.db
Done.

```
Out[33]: sum(PAYLOAD_MASS_KG_)
         45596
```

```
In [34]: %sql select avg(PAYLOAD_MASS_KG_) from SPACEXTBL where Booster_Version like '%F9 v1.0%'
```

* sqlite:///my_data1.db
Done.

```
Out[34]: avg(PAYLOAD_MASS_KG_)
         340.4
```

Results (SQL)

The first successful landing outcome and names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [35]: %sql select min(Date) as first_successful_landing from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

* sqlite:///my_data1.db
Done.

```
Out[35]: first_successful_landing
```

first_successful_landing
2015-12-22

```
In [36]: %sql select booster_version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS_KG_ between 4000 and 6000
```

* sqlite:///my_data1.db
Done.

```
Out[36]: Booster_Version
```

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

Results (SQL)

Total number of successful and failure mission outcomes and names of the booster_versions which have carried the maximum payload mass.

```
In [37]: %sql select Mission_Outcome, count(*) as total_number from SPACEXTBL group by Mission_Outcome;
```

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

```
Done.
```

```
Out[37]:
```

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

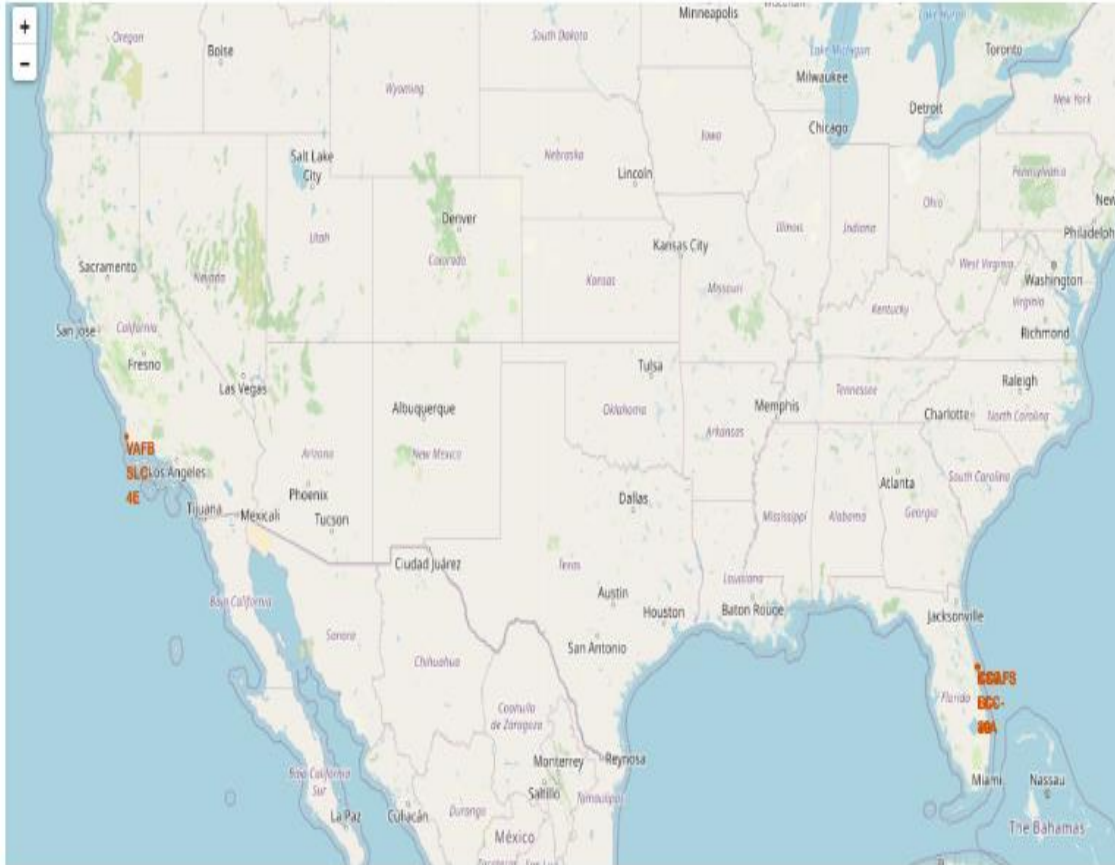
```
In [38]: %sql select Booster_Version from SPACEXTBL where  
* sqlite:///my_data1.db  
Done.
```

```
Out[38]:
```

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

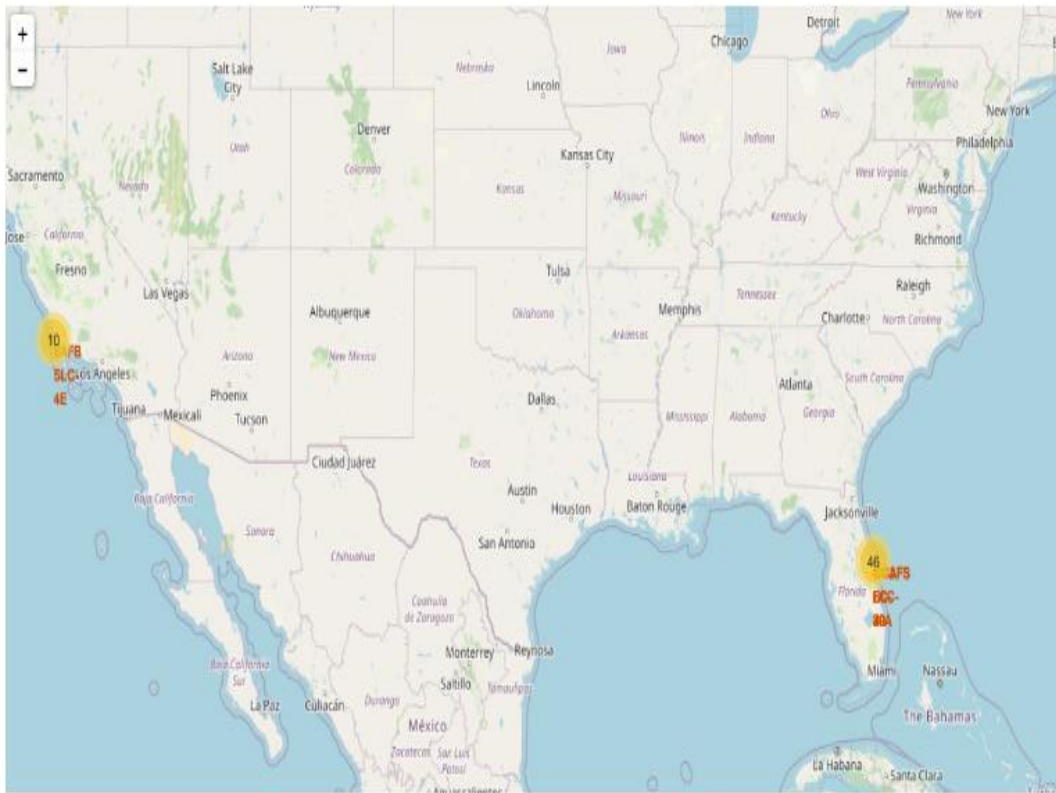
Results (Visualization with Folium)

The generated map with marked launch sites should look similar to the following:



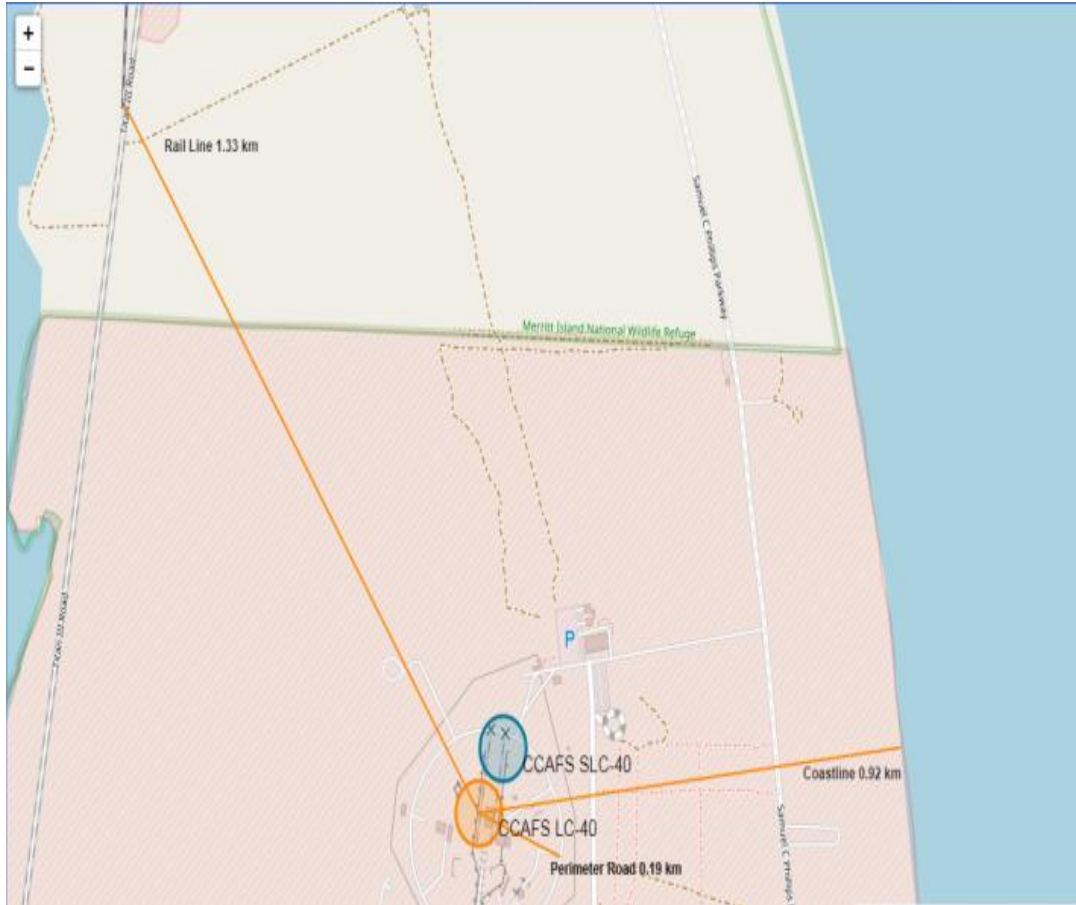
- Most launch sites are located near the Equator, where the Earth's surface moves faster than anywhere else. At the Equator, objects are already traveling at a speed of 1670 km/h. When a spacecraft is launched from this region, it ascends into space while maintaining this rotational speed due to inertia. This initial velocity aids the spacecraft in achieving the necessary speed to remain in orbit.
- Additionally, all launch sites are situated in close proximity to the coast. Launching rockets toward the ocean reduces the risk of debris falling or explosions occurring near populated areas.

Results (Visualization with Folium)



From the color-labeled markers in marker clusters, we should be able to easily identify which launch sites have relatively high success rates.

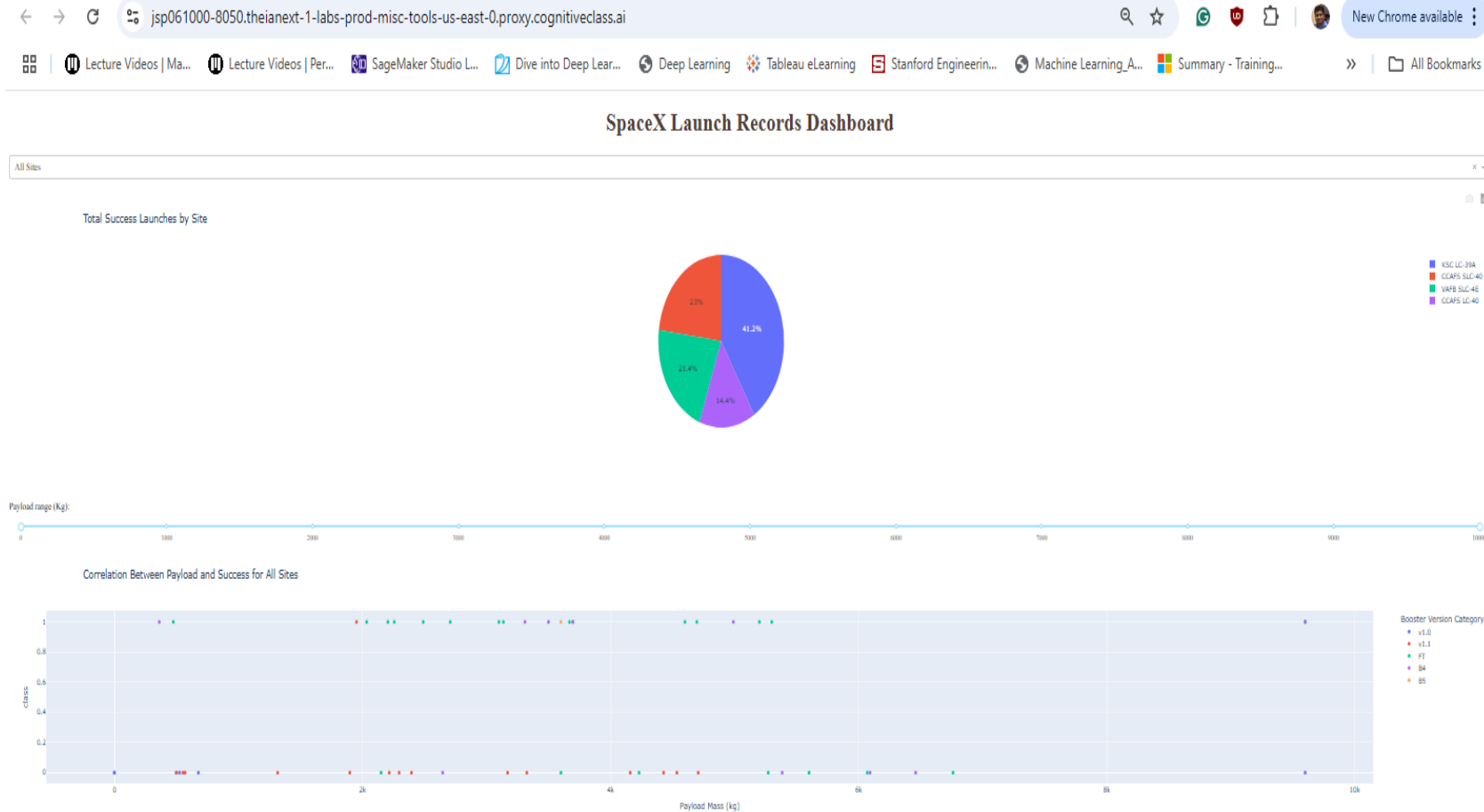
Results (Visualization with Folium)



The **CCAFS LC-40** and **CCAFS SLC-40** launch sites have coordinates that are nearly identical but are not precisely aligned.

- The perimeter road surrounding CCAFS LC-40 is located **0.19 km** from the launch site coordinates.
- The coastline is situated **0.92 km** away from CCAFS LC-40.
- The nearest rail line is **1.33 km** from CCAFS LC-40.

SpaceX Launch Record DASHBOARD



Complete Dashboard including success launches by site and correlation between Payload and Success for all sites.

SpaceX Launch Record DASHBOARD

jsp061000-8050.theianext-1-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai

Lecture Videos | Ma... Lecture Videos | Per... SageMaker Studio L... Dive into Deep Lear... Deep Learning Tableau eLearning Stanford Engineerin... Machine Learning_A... Summary - Training... All Bookmar

SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site



Payload range (Kg):



Total Successful Launches by the Launch Sites

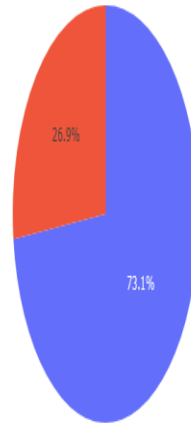


SpaceX Launch Record DASHBOARD

SpaceX Launch Records Dashboard

CCAFS LC-40

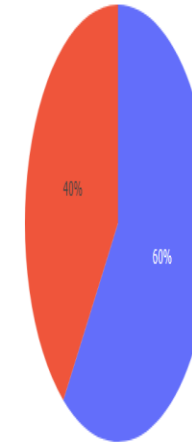
Total Success Launches for Site CCAFS LC-40



SpaceX Launch Records Dashboard

VAFB SLC-4E

Total Success Launches for Site VAFB SLC-4E



Total Success Launches for Sites CCAFS LC40 and VAFB SLC 4E.

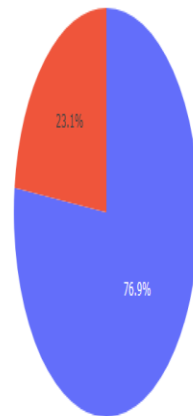
SpaceX Launch Record DASHBOARD

SpaceX Launch Records Dashboard

KSC LC-39A

x ▼

Total Success Launches for Site KSC LC-39A

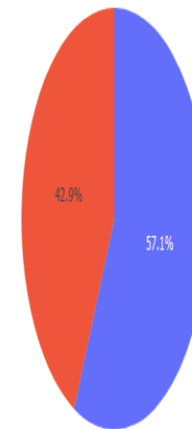


SpaceX Launch Records Dashboard

CCAFS SLC-40

x ▼

Total Success Launches for Site CCAFS SLC-40



Total Success Launches for Sites KSC LC 39-A and CCAFS SLC 40.

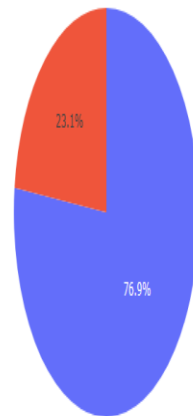
SpaceX Launch Record DASHBOARD

SpaceX Launch Records Dashboard

KSC LC-39A

x ▼

Total Success Launches for Site KSC LC-39A



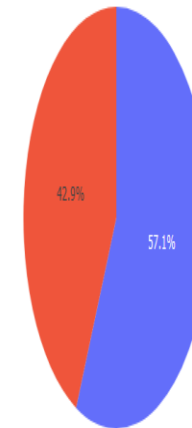
0
1

SpaceX Launch Records Dashboard

CCAFS SLC-40

x ▼

Total Success Launches for Site CCAFS SLC-40



0
1

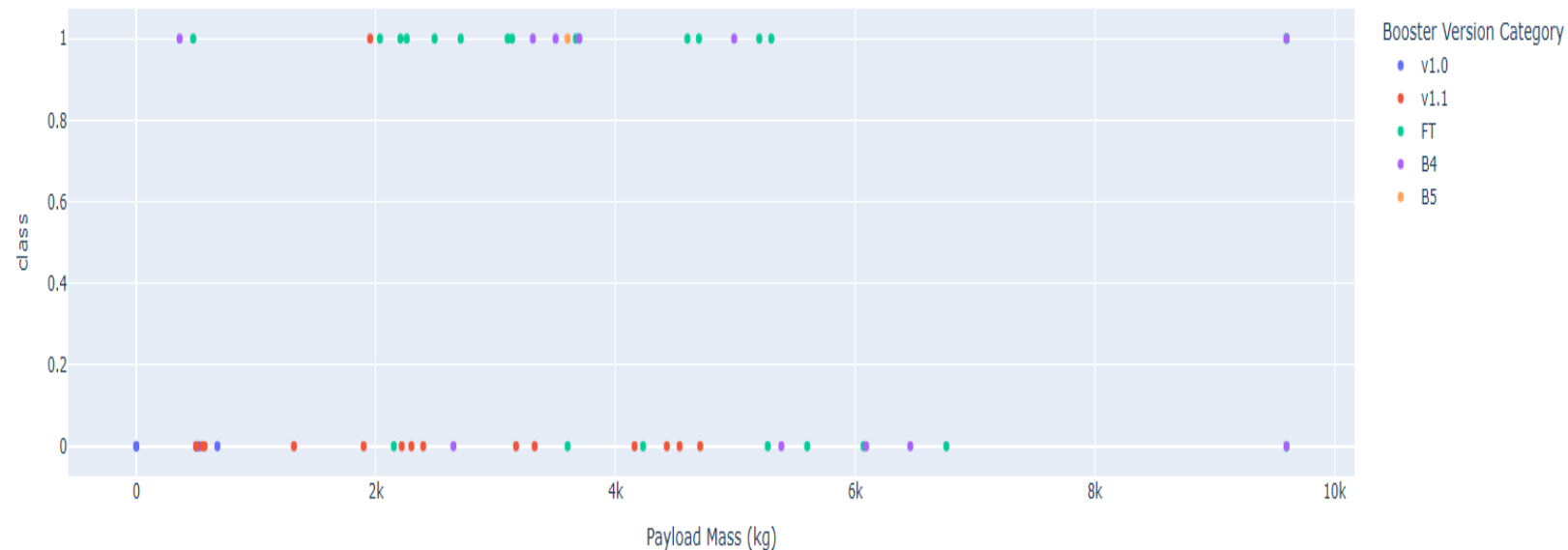
Total Success Launches for Sites KSC LC 39-A and CCAFS SLC 40.

SpaceX Launch Record DASHBOARD

Payload range (Kg):



Correlation Between Payload and Success for All Sites



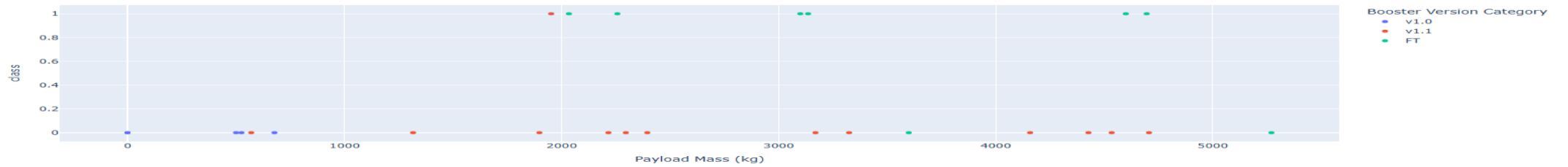
Correlation between payload and success for all sites.

SpaceX Launch Record DASHBOARD

Payload range (Kg):



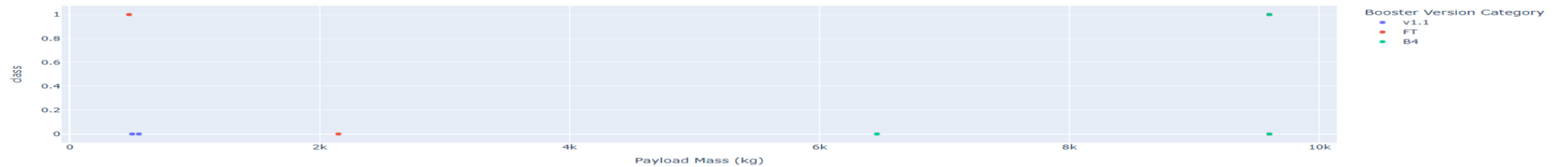
Correlation Between Payload and Success for Site CCAFS LC-40



Payload range (Kg):



Correlation Between Payload and Success for Site VAFB SLC-4E



Correlation between Payload Mass and Success for Sites KSC LC 39-A and CCAFS SLC 40.


SpaceX Launch Record DASHBOARD




Correlation between Payload Mass and Success for Sites KSC LC 39-A and CCAFS SLC 40.

SpaceX Launch Record DASHBOARD

Data-Science-Capstone / spacex_launch_dash.csv 

 jigar1729 Add files via upload

Preview Code Blame 57 lines (57 loc) · 2.42 KB  Code 55% faster with GitHub Copilot

 Search this file

		Flight Number	Launch Site	class	Payload Mass (kg)	Booster Version	Booster Version Cate
1							
2	0	1	CCAFS LC-40	0	0	F9 v1.0 B0003	v1.0
3	1	2	CCAFS LC-40	0	0	F9 v1.0 B0004	v1.0
4	2	3	CCAFS LC-40	0	525	F9 v1.0 B0005	v1.0
5	3	4	CCAFS LC-40	0	500	F9 v1.0 B0006	v1.0
6	4	5	CCAFS LC-40	0	677	F9 v1.0 B0007	v1.0
7	5	7	CCAFS LC-40	0	3170	F9 v1.1	v1.1
8	6	8	CCAFS LC-40	0	3325	F9 v1.1	v1.1
9	7	9	CCAFS LC-40	0	2296	F9 v1.1	v1.1
10	8	10	CCAFS LC-40	0	1316	F9 v1.1	v1.1
11	9	11	CCAFS LC-40	0	4535	F9 v1.1	v1.1
12	10	12	CCAFS LC-40	0	4428	F9 v1.1 B1011	v1.1
13	11	13	CCAFS LC-40	0	2216	F9 v1.1 B1010	v1.1
14	12	14	CCAFS LC-40	0	2395	F9 v1.1 B1012	v1.1

GitHub URL:
CSV file containing
data for the
dashboard.

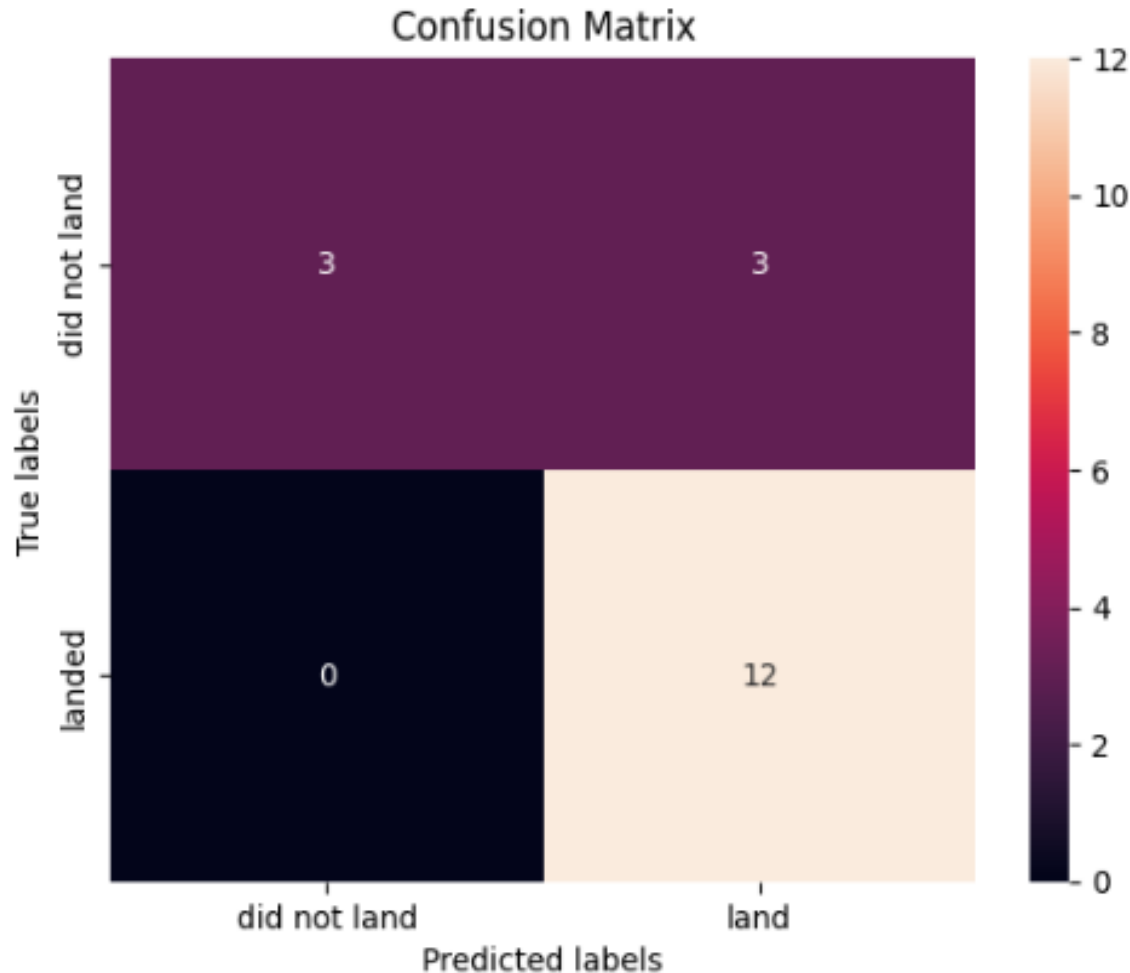
Results Predictive Analytics

Training Accuracies

	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

- Jaccard Score, F1 Scores, and Accuracy for training set remains equal for Logistic Regression, Support Vector Machines, Decision Tree, and K-Nearest Neighbor.
- Therefore we look into the testing accuracies.

Results Predictive Analytics



- Confusion matrix for all classification methods have similar values.
- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)
- There are three misclassification. (did not land predicted as landed)

Results Predictive Analytics

Testing Accuracies

	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

- Among all four classification methods, Jaccard Score, F1 Scores, and Accuracy for Decision Tree are higher compared to the remaining three on the testing dataset.
- Therefore, we select Decision Tree for this task.
- Test accuracy represents an unbiased measure.

CONCLUSION and DISCUSSION



- For this classification task, based on the testing set accuracies our best model selection should be decision tree.
- SpaceX does not have an unblemished record when it comes to the outcomes of Falcon 9 first stage landings.
 - However, the success rate of Falcon 9 first stage landings has been improving with each subsequent launch.
 - Machine learning models can be employed to forecast the outcomes of future Falcon 9 first stage landings by SpaceX.

APPENDIX

Thank you!

All codes and supporting files can be found on my [GitHub](#) Account.

Special Thanks to IBM and Course Instructors.