IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lauren V.

11 Oct 2024

Outline

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

Executive Summary

- Summary of methodologies

- The project employed a comprehensive data science approach encompassing multiple phases:
  - Data Collection and Integration
  - Data Preprocessing and Quality Assurance
  - Exploratory Data Analysis (EDA)
  - Geospatial Analysis
  - Interactive Dashboards
  - Summary of Results
- The analysis revealed significant patterns in Falcon 9 first stage landing success rates across different launch sites and mission parameters. Through our comprehensive analysis and interactive dashboard, we identified several key findings:
- **Launch Site Performance**
  - Success rates vary significantly among different launch sites
  - Visualization through pie charts revealed site-specific success patterns
  - Interactive filtering enabled detailed site-by-site comparison
- **Payload Impact**
  - Correlation between payload mass and mission success was identified through scatter plot analysis
  - Interactive payload range filtering revealed optimal mass ranges for successful landings
  - Different booster versions showed varying performance across payload ranges
- **Operational Insights**
  - The interactive dashboard enabled rapid identification of successful launch parameters
  - Patterns emerged showing relationship between booster versions and success rates
  - Combined analysis of site location and payload mass revealed optimal launch configurations
  - These findings provide crucial insights for competitive pricing strategies in the commercial space launch market. The interactive nature of our analysis tools allows for dynamic exploration of success factors, enabling more accurate prediction of landing success probability for future launches based on specific mission parameters.
  - The results suggest that launch site selection, payload mass, and booster version selection significantly influence landing success rates, with certain combinations demonstrating higher success probabilities. This intelligence is valuable for companies seeking to compete with SpaceX by understanding the technical and operational factors that contribute to reduced launch costs through first stage recovery and reuse.

Introduction

- Project background and context
  - The commercial space industry has been revolutionized by SpaceX's ability to significantly reduce launch costs through innovative rocket reusability. SpaceX advertises Falcon 9 rocket launches at $62 million per launch, while competitors price similar services at upward of $165 million. This dramatic price difference is primarily attributed to SpaceX's capability to recover and reuse the first stage of their Falcon 9 rockets, a technological achievement that has redefined the economics of space launches.
- Problem Statement
  - This project aims to solve several critical business questions: Can we accurately predict the successful landing of Falcon 9's first stage? What are the key factors influencing landing success? By analyzing launch data and developing predictive models, we seek to understand the probability of successful first-stage recovery for future launches. This analysis will enable potential competitors to better estimate actual launch costs and develop more competitive pricing strategies when bidding against SpaceX for launch contracts.

Section 1

Methodology

Methodology

Executive Summary

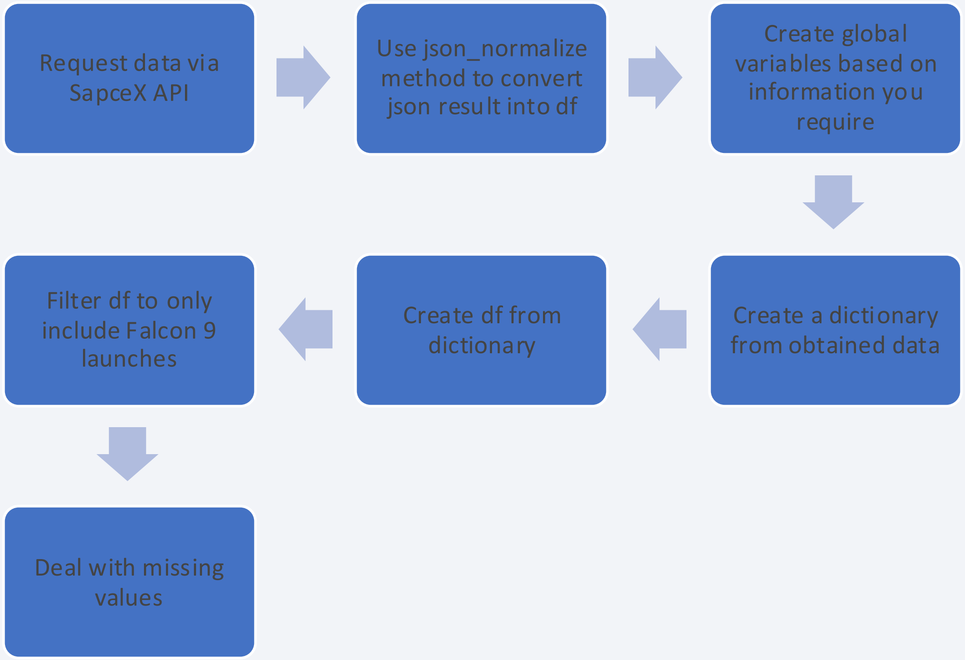
- **Data collection methodology:**
    - Extracted launch data through SpaceX API integration
    - Scraped supplementary Falcon 9 launch records from Wikipedia using BeautifulSoup
    - Consolidated data into a structured Db2 database for analysis
- **Perform data wrangling**
    - Implemented data cleaning protocols to ensure consistency
    - Standardized formatting across multiple data sources
    - Validated data integrity through automated checks
- **Perform exploratory data analysis (EDA) using visualization and SQL**
    - Conducted statistical analysis to identify patterns and correlations
    - Utilized matplotlib and seaborn for visualization of launch success rates
    - Analyzed relationship between launch sites and mission outcomes
- **Perform interactive visual analytics using Folium and Plotly Dash**
    - Mapped all launch sites using geographical coordinates
    - Created visual overlays of successful and failed launches by location
    - Calculated proximity metrics between launch sites and critical infrastructure
- **Interactive Dashboard Development**
    - Created a dynamic web-based dashboard using Dash and Plotly
- Perform predictive analysis using classification models
- How to build, tune, evaluate classification models
    - Perform exploratory Data Analysis and determine Training Labels
    - create a column for the class
    - Standardize the data
    - Split into training data and test data
    - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
    - Find the method performs best using test data

Data Collection

- Extracted launch data through SpaceX API integration
- Scraped supplementary Falcon 9 launch records from Wikipedia using BeautifulSoup
- Consolidated data into a structured Db2 database for analysis

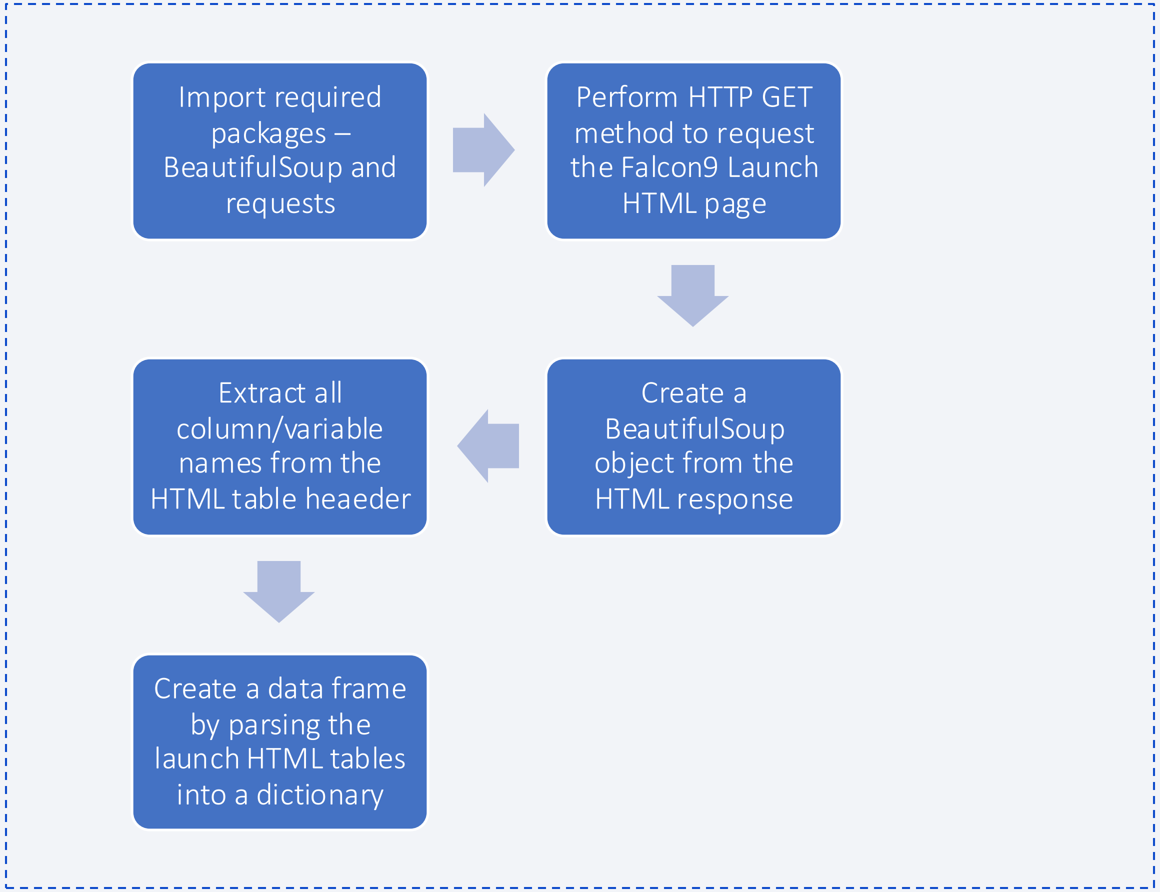
Data Collection – SpaceX API

Github -
https://github.com/llvasquez/testrepo/blob/52d4bdfc619a59e7047f79de50d8f7b0b5e2ef65/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

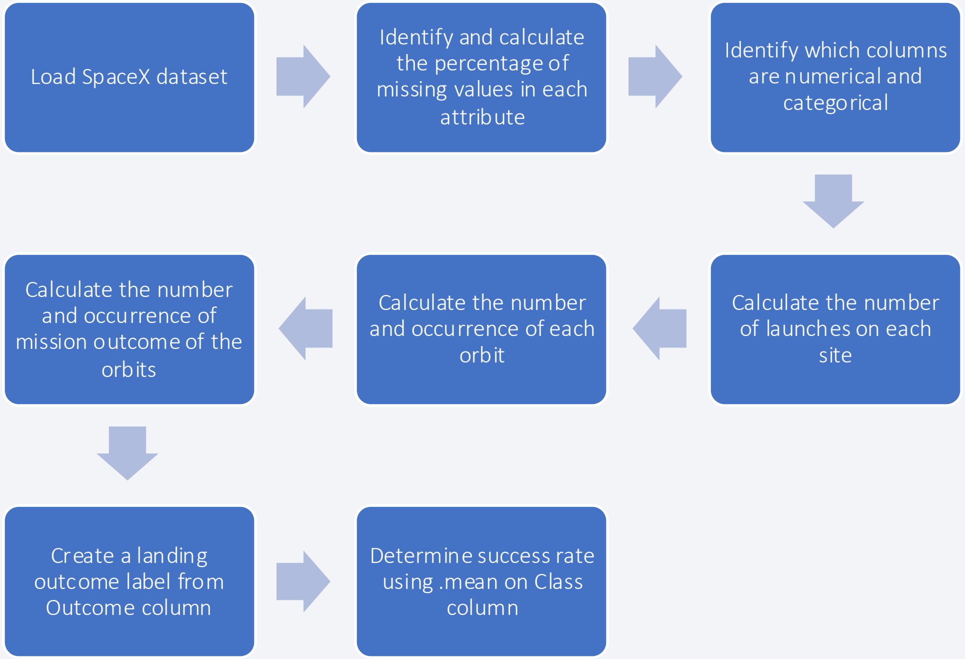
- **Github** - https://github.com/llvasquez/testrepo/blob/52d4bdfc619a59e7047f79de50d8f7b0b5e2ef65/jupyter-labs-webscraping.ipynb



Data Wrangling

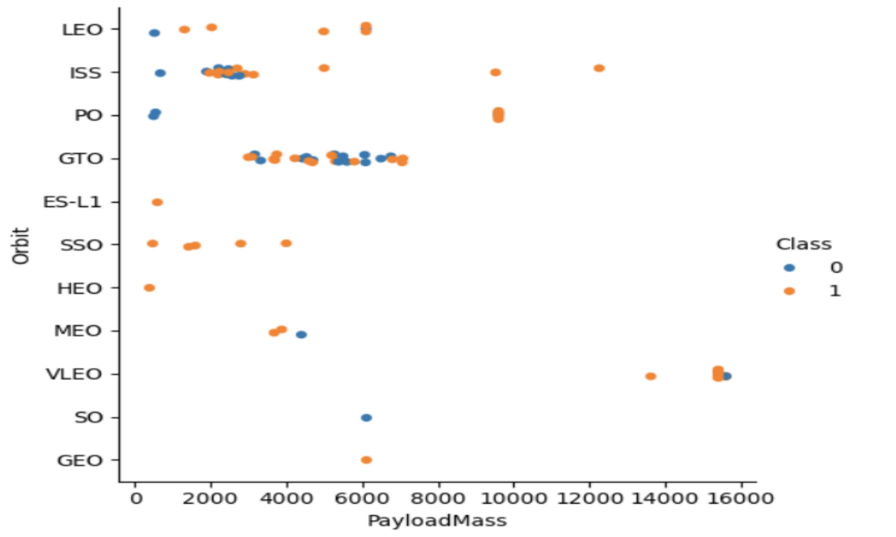
- Implemented data cleaning protocols to ensure consistency
- Standardized formatting across multiple data sources
- Validated data integrity through automated checks

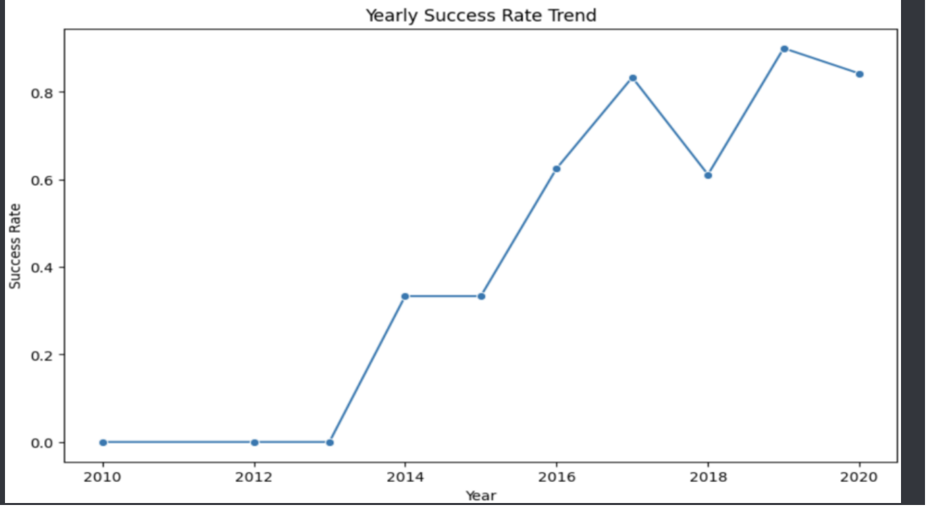
- GitHub Link - https://github.com/llvasquez/testrepo/blob/52d4bdfc619a59e7047f79de50d8f7b0b5e2ef65/labs-jupyter-spacex-Data%20wrangling%20(2).ipynb

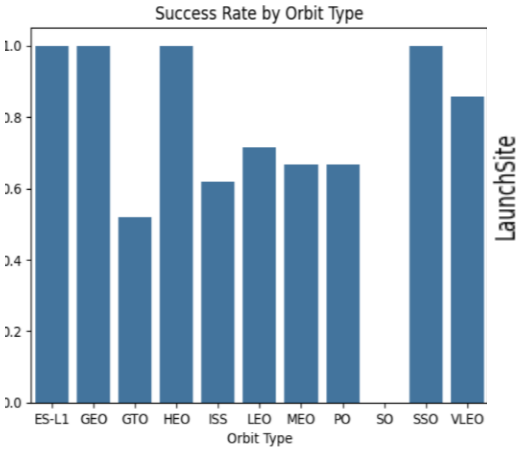


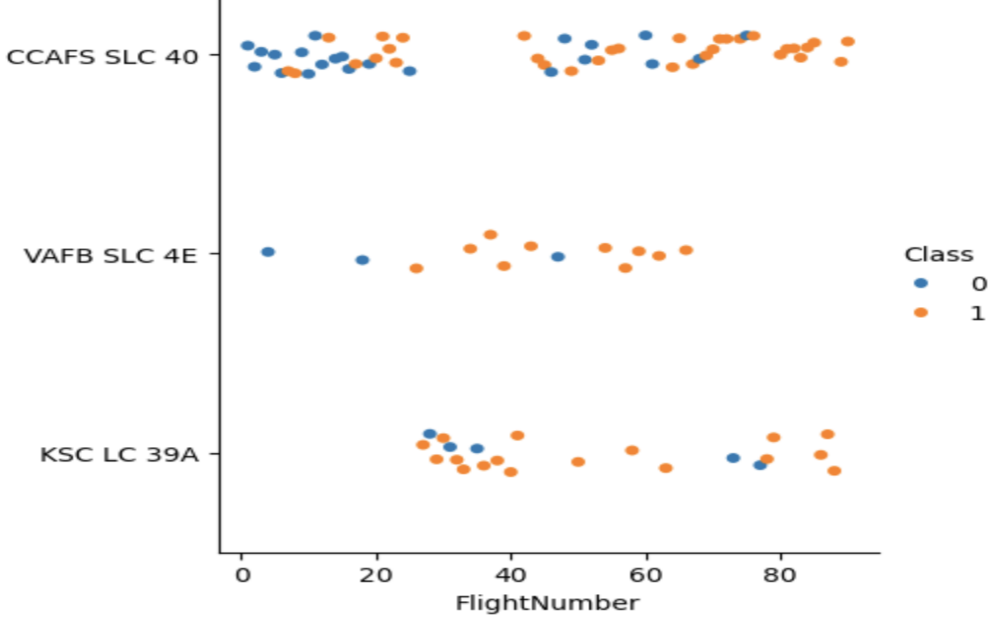
EDA with Data Visualization

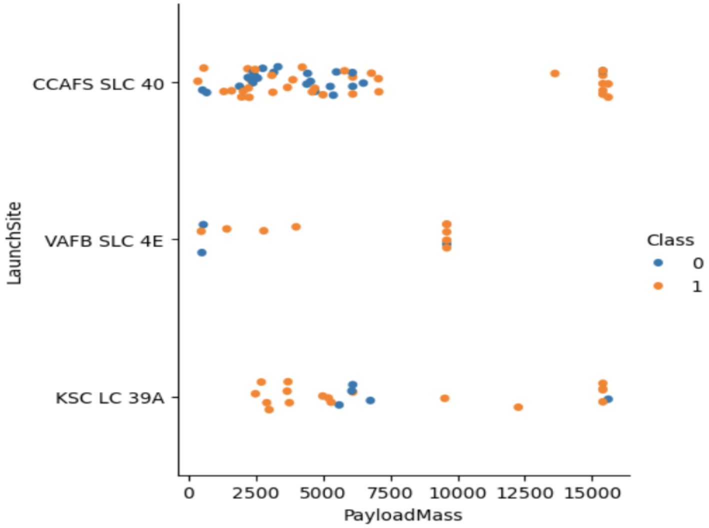
https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/edadataviz.ipynb

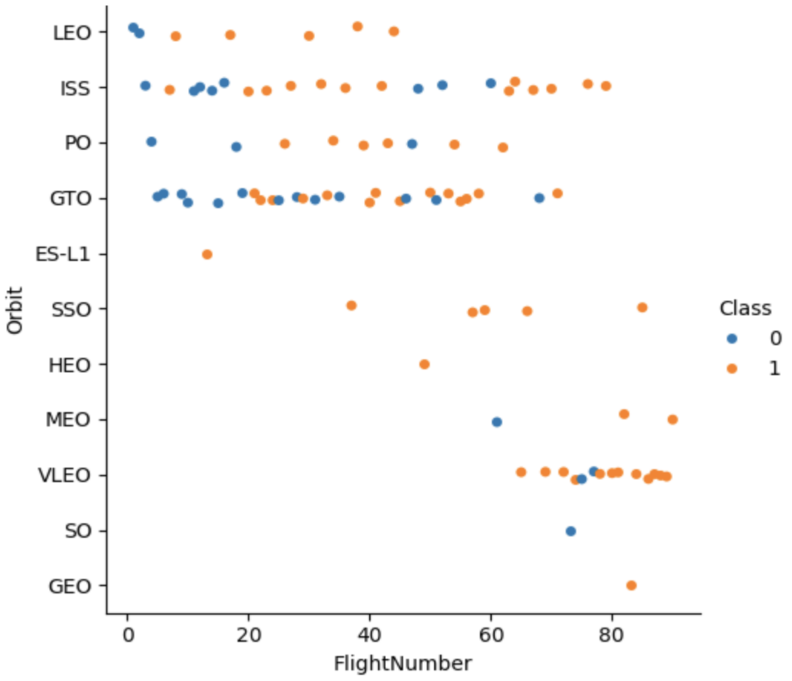












EDA with SQL

- Removed blank rows from table
- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed the average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failed mission outcomes
- Listed the names of the booster_versions which have carried the maxium payload mass using a subquery
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranked 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.
- https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created a folium Map object, with an initial center location of NASA Johnson Space Center at Houston, TX.
- Added a circle to highlight the area with a text label.
- Added a circle and folium Marker for each launch site
- Marked the success/failed launches for each site on the map to determine which sites have high success rates.
    - Created a marker Cluster and marker for all launch records – green marker for success and red for failed
- Added a MousePosiiton on the map tp get coordinates for a mouse over a point on a map to easily find the coordinates of any point of interest.
- Added a Marker with distance to the nearest coastline
    - Drew a PolyLine between the launch site and the coastline
- Added a Marker with distance to the nearest Highway
    - Drew a PolyLine between the distance to the nearest Highway
- https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Initialized a basic dash application
- Added a dropdown list to enable Launch Site selection
- Created a Pie Chart to display successful launches based on Launch Site
- Added a Payload range selector
- Added a slider to select the payload range
- Added a scatter chart to display the success rate based on payload range
- https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/spacex_dash_app.py

Predictive Analysis (Classification)

https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

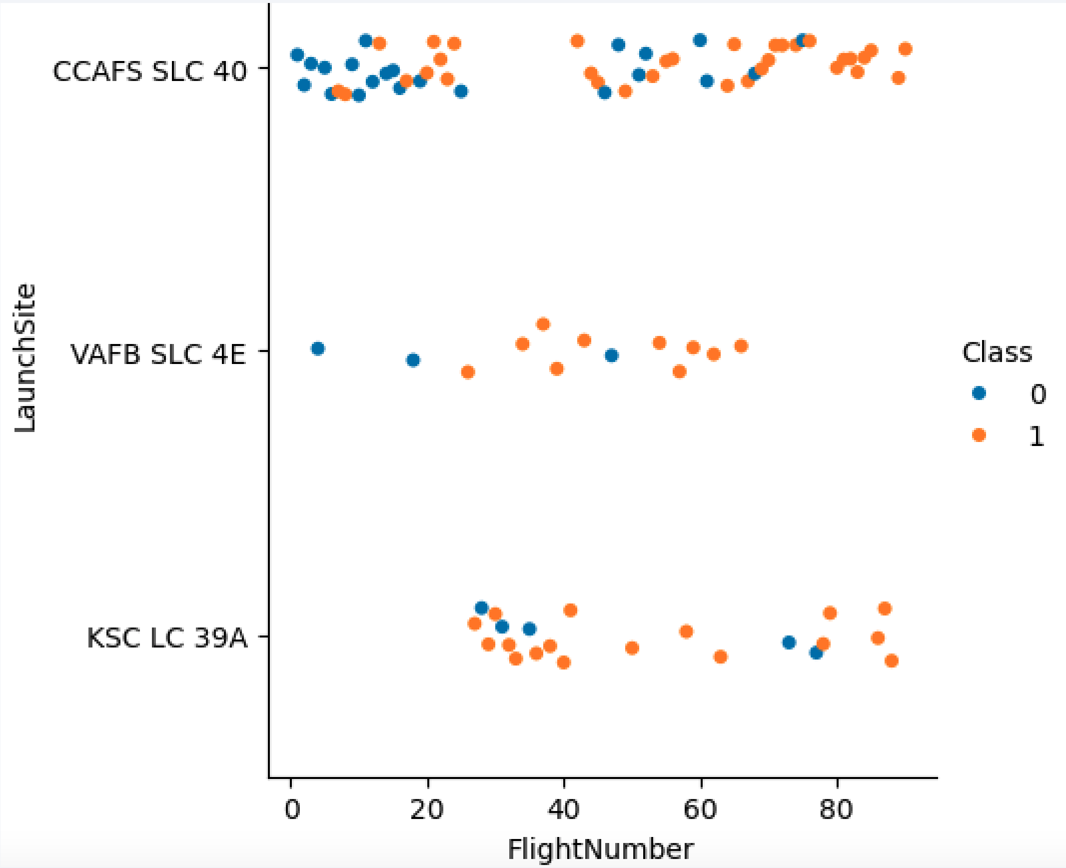
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

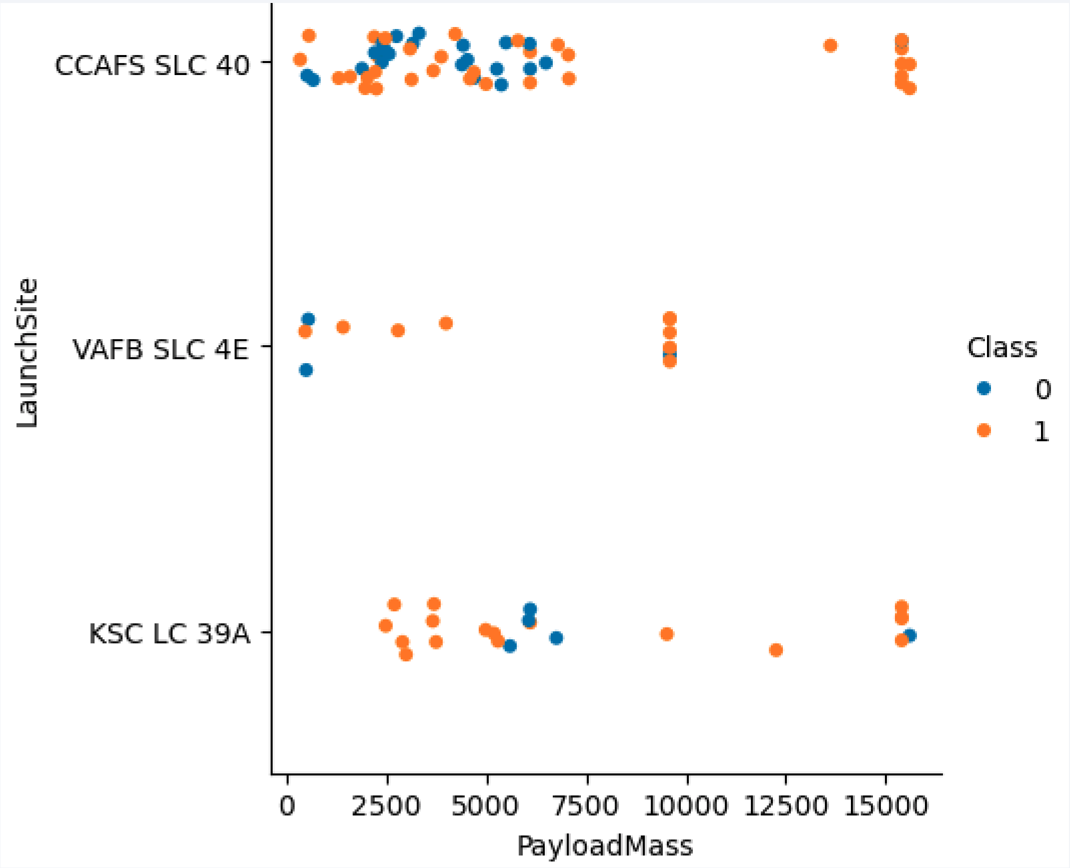
Flight Number vs. Launch Site

- As flight number increases, the first stage is more likely to land successfully.



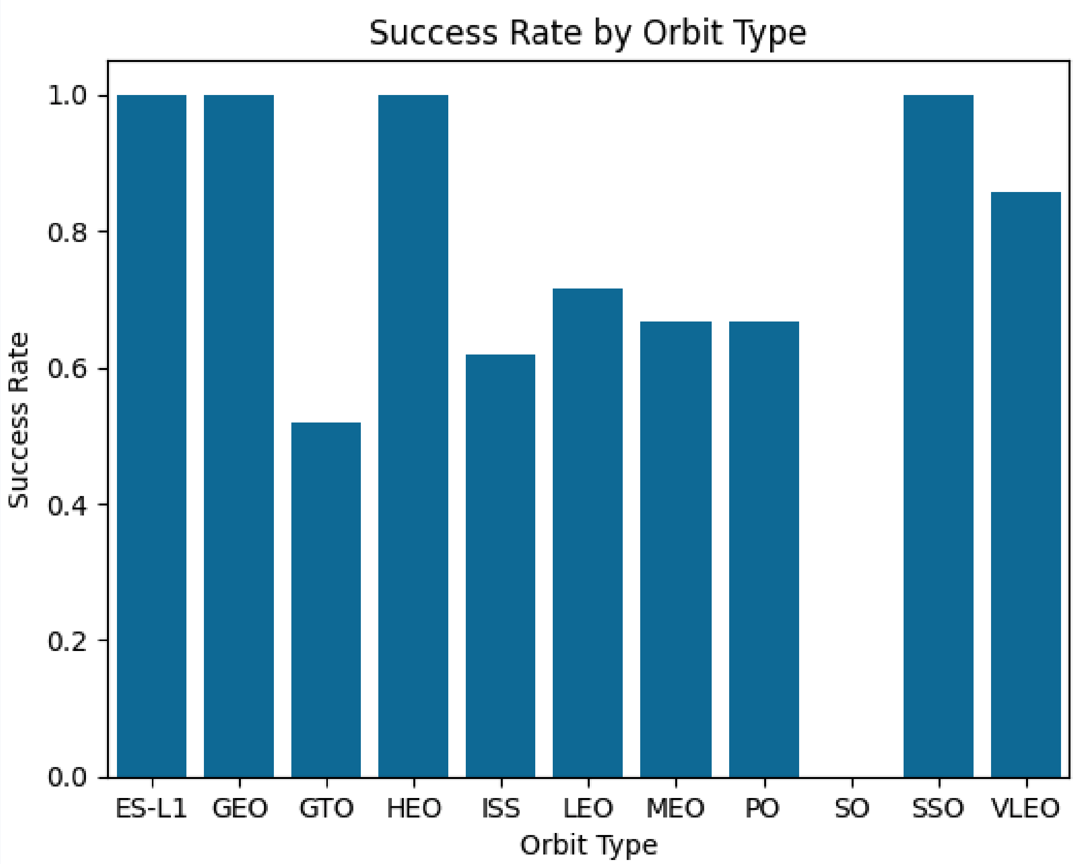
Payload vs. Launch Site

- There appears to be a relationship between launch sites and payload mass
    - VAFB-SLC launch site does not have rockets launched for the heavypayload mass



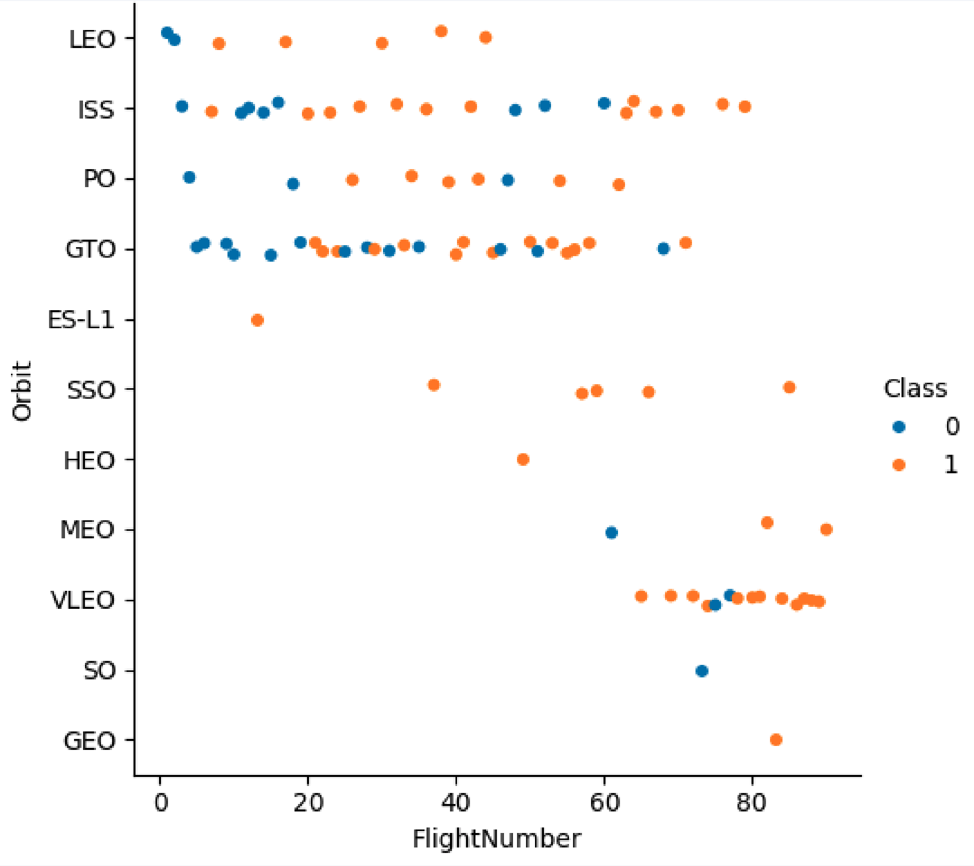
Success Rate vs. Orbit Type

- ES-L1; GEO; HEO; SSO appear to have the highest orbit type success rates



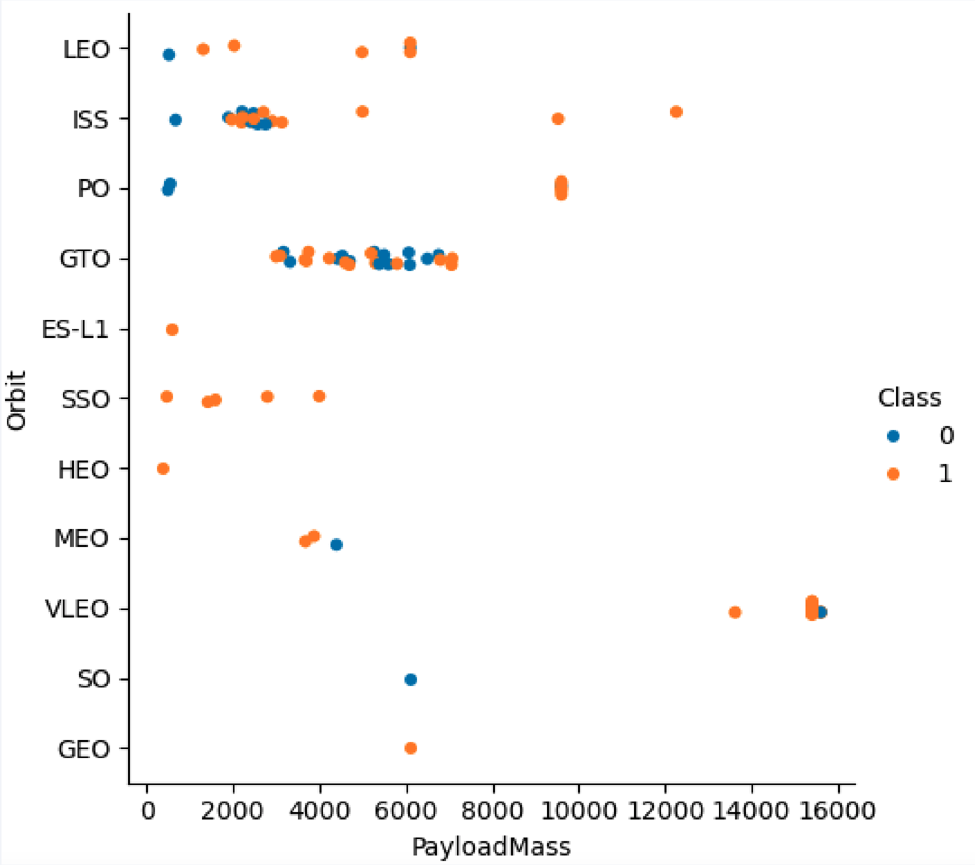
Flight Number vs. Orbit Type

- Success rates seem to be related to the number of flights in the LEO orbit
- In the GTO orbit, there appears to be no relation between flight number and success



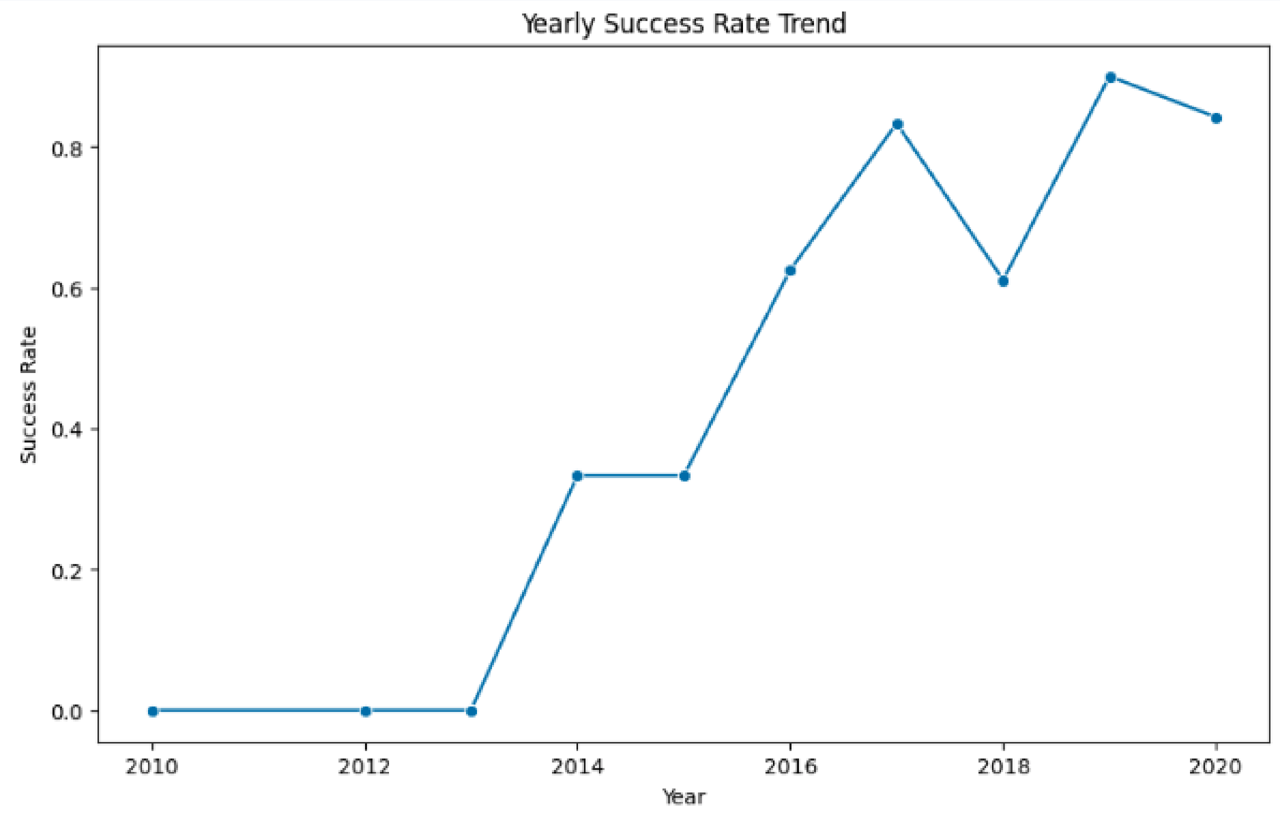
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- Success rates since 2013 kept increasing till 2020



All Launch Site Names

- Displayed the names of the unique launch sites in the space mission

Display the names of the unique launch sites in the space mission

In [24]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Out[24]:

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

Launch Site Names Begin with 'CCA'

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

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT "Launch_Site" FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

Hint:Use min function

```
%sql SELECT MIN(Date) AS First_Successful_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)
```

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

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- `%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;`

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
XTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- `%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);`

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
ECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- ```
  %sql SELECT substr(Date, 6, 2) AS month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE substr(Date, 0, 5)
  = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';
  ```

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
n", "Launch_Site" FROM SPACEXTBL WHERE substr(Date, 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
```

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

- `%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Outcome_Count DESC;`

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.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '20
```

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

Landing_Outcome	Outcome_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

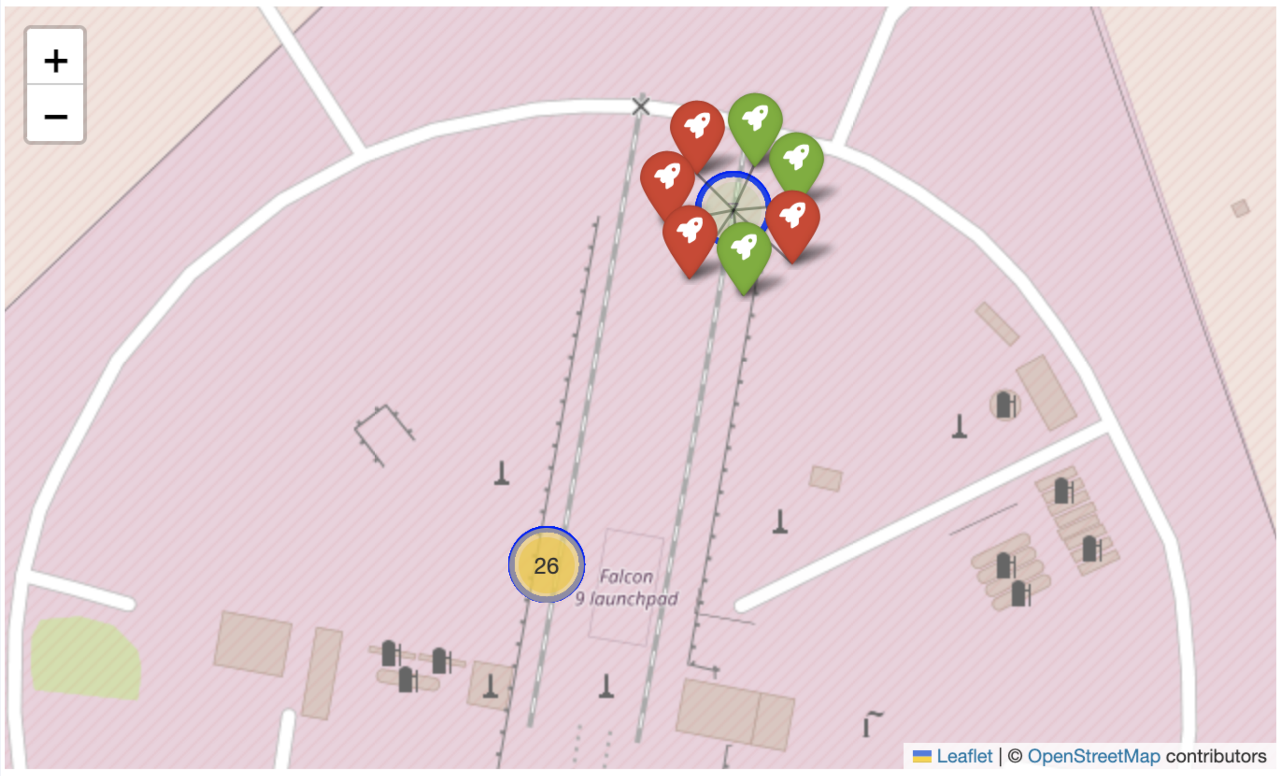
Section 3

Launch Sites Proximities Analysis

Launch Sites Location Analysis with Folium

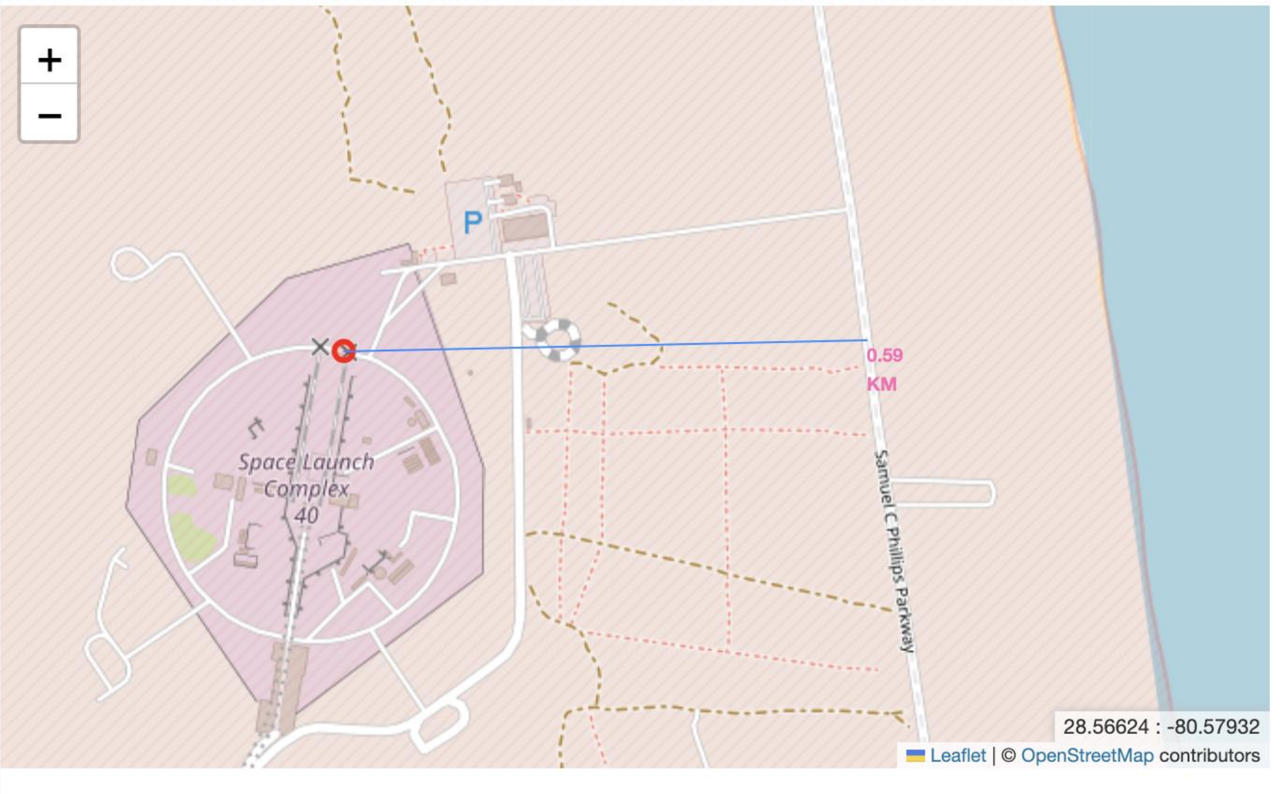
- The map shows three main launch sites:
  - VAFB SLC-4E Vandenberg Space Launch Complex
  - Kennedy Space Center – KSC LC-93A
  - Cape Canaveral Space Force Station -
    - CCAFS SLC-40
    - CCAFS LC-40

Launch Sites Location Analysis with Folium



- Of the 3 main locations, the Kennedy Space Center LC-39A appears to have the best success.

Launch Sites Location Analysis with Folium

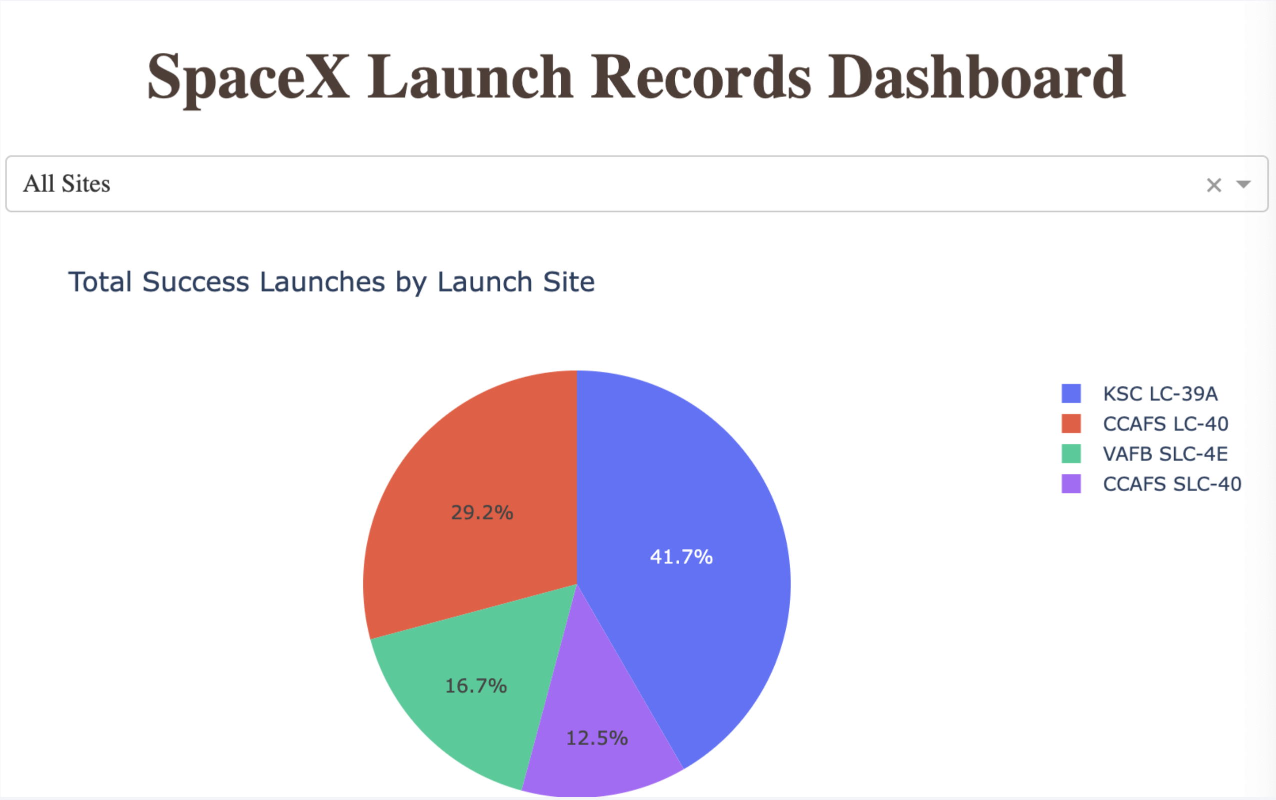


- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

Section 4

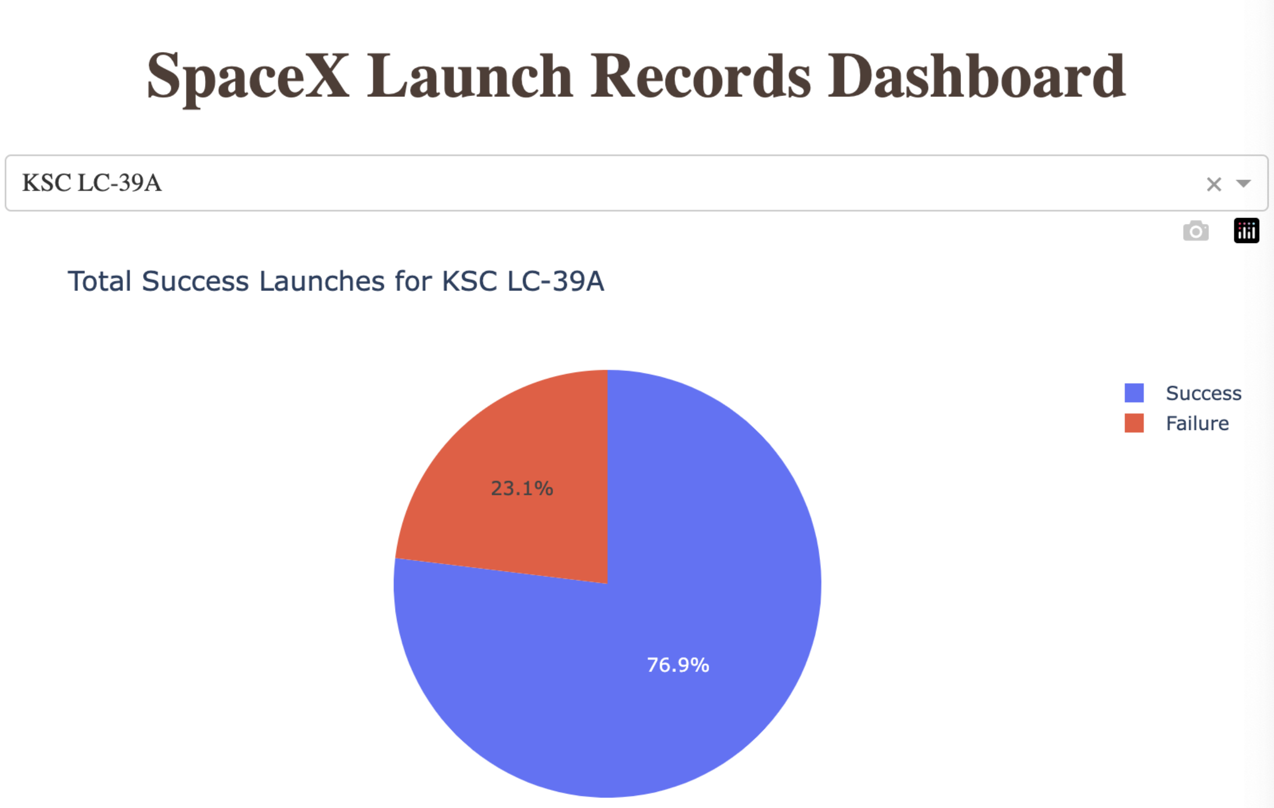
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



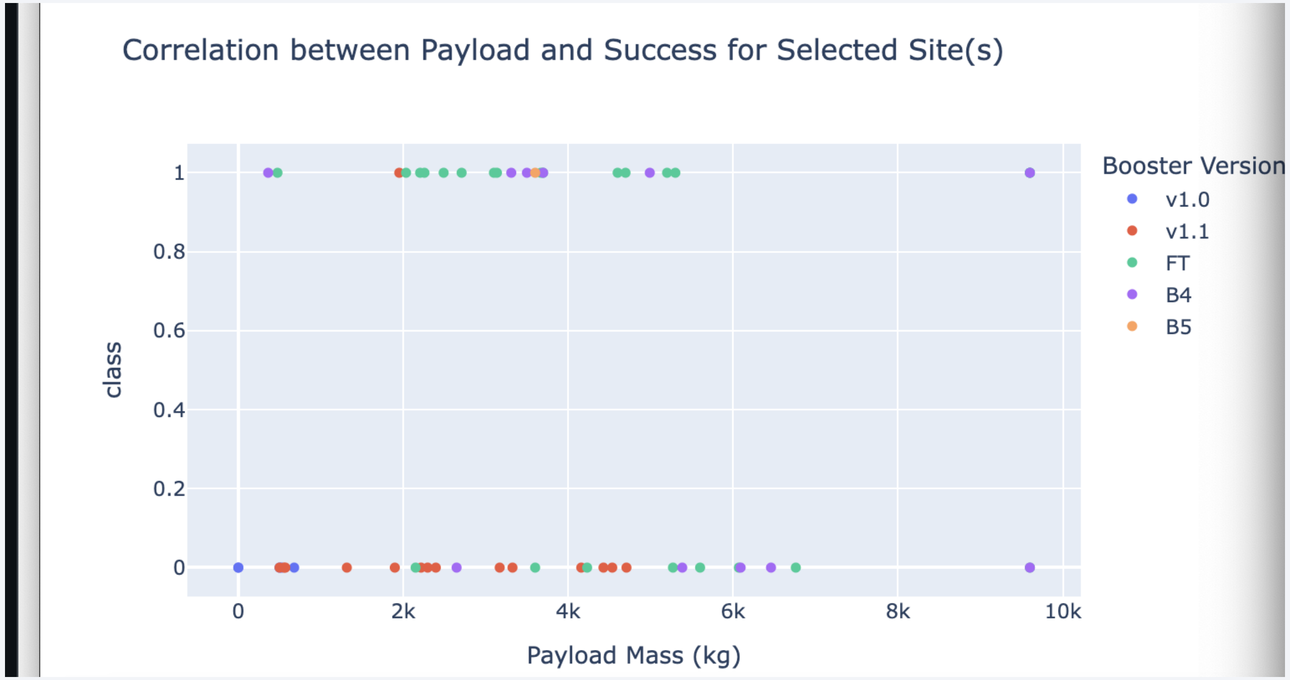
- The provided pie chart shows the breakdown of successful launches by site.
- KSC LC-39A makes up 41.7% of successful launches.

SpaceX Launch Records Dashboard



- The pie chart shows the breakdown of successful vs. Failed launches for the KSC LC-39A site.
  - 76.9% succeeded vs. A 23.1% failure rate.

SpaceX Launch Records Dashboard



- Payload mass between 2k-5k have the highest success rates.

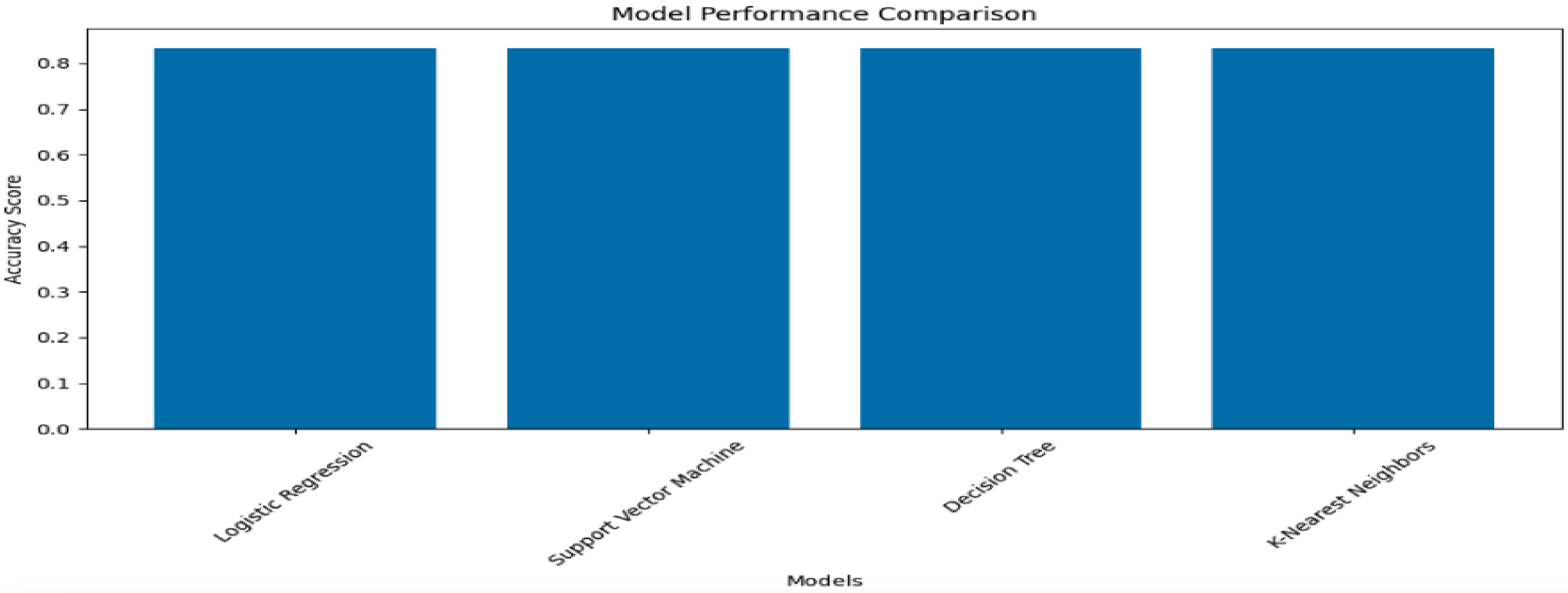
Section 5

Predictive Analysis (Classification)

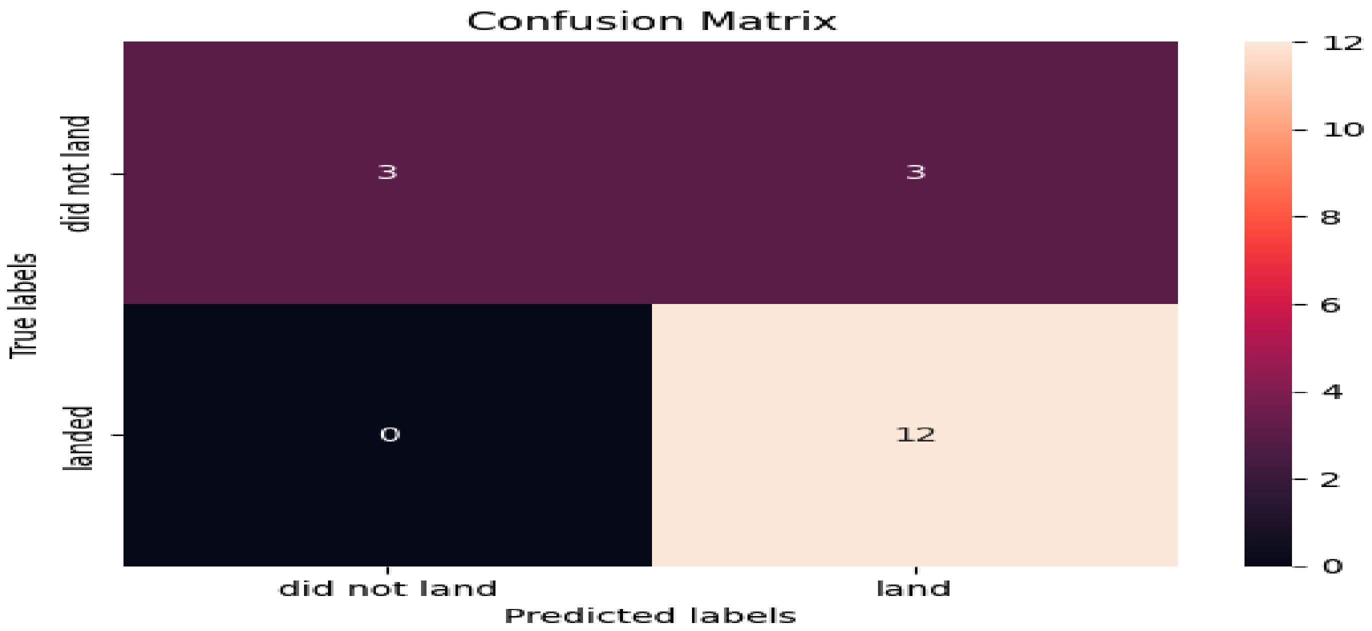
Classification Accuracy

```
Model Performance Comparison:
--------------------------------
Logistic Regression: 0.8333
Support Vector Machine: 0.8333
Decision Tree: 0.8333
K-Nearest Neighbors: 0.8333

Best Performing Model:
Logistic Regression with accuracy: 0.8333
```



Confusion Matrix



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

True Postive - 12 (True label is landed, Predicted label is also landed)

False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- **Comprehensive Analysis**: The project utilized a multi-phase data science approach, including data collection, preprocessing, exploratory data analysis, geospatial analysis, and interactive dashboard development.
- **Key Findings**: Significant variations in Falcon 9 first stage landing success rates were identified, with launch site performance and payload mass being critical factors influencing mission success.
- **Interactive Insights**: The dashboard facilitated dynamic exploration of data, allowing users to visualize site-specific success patterns and optimal payload ranges for enhanced decision-making.
- **Strategic Implications**: Insights gained from the analysis inform competitive pricing strategies in the commercial space launch market, highlighting the importance of launch site selection, payload mass, and booster version in maximizing landing success probabilities.

Appendix

- Github Link to SpaceX Web Scraped Data in CSV format - https://github.com/llvasquez/testrepo/blob/4f533944b80b78663892fd8bc8305ff97571ccd6/spacex_web_scraped.csv

Thank you!