IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rebeka Bhattacharyya
27-07-2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - The methodology involved loading SpaceX launch data using `pandas` and building an interactive dashboard with `Dash` and `Plotly`. The dashboard features dropdowns for selecting launch sites and booster versions, a payload mass slider, and visualizations including a pie chart and scatter chart. Callbacks dynamically update these charts based on user input to show launch success distributions and correlations between payload mass and success rates, filtered by site and booster version.

Executive Summary

- **Summary of Results**
  - The results include a pie chart that visualizes the distribution of successful launches across different sites or specific sites based on user selection, and a scatter chart that shows the correlation between payload mass and success rates. The scatter chart updates dynamically based on the selected payload range, launch site, and booster version, providing insights into performance trends and relationships in the data.

Introduction

- **Project background and context**
  - In the rapidly evolving aerospace industry, analyzing the performance of space missions is crucial for optimizing launch strategies and improving success rates. SpaceX, a leader in space exploration and innovation, provides a rich dataset of launch records that includes various attributes such as launch sites, payload masses, and booster versions. This dataset offers an opportunity to delve into the factors influencing launch success and to identify trends and correlations that can inform future missions.

Introduction

Problems We Want to Find Answers

1. **Launch Site Performance:** How does the success rate of launches vary across different launch sites? What insights can we gain about the effectiveness of each site based on historical data?
2. **Payload Mass Impact:** What is the relationship between payload mass and launch success? Does payload mass influence the likelihood of a successful mission, and if so, how does this correlation vary by launch site or booster version?
3. **Booster Version Influence:** How do different booster versions affect launch success rates? Can we identify any patterns or trends related to specific booster configurations?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - **Sources**: Data was sourced from https://api.spacexdata.com/v4/rockets/
  - **Methods**: Utilized tools and techniques such as API calls, web scraping libraries, SQL queries to gather data. Ensured data was collected at the necessary granularity and temporal resolution.
- Perform data wrangling
  - The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data Collection: Identified and accessed relevant data sources through APIs, web scraping, and public datasets. Extracted and stored the data in structured formats like CSV, JSON, or databases, ensuring accuracy and completeness through initial verification.

- Data Integration: Merged data from multiple sources into a unified dataset, preparing it for subsequent processing and analysis.

Data Collection – SpaceX API

- Data obtained from the API is as below: Github URL For Notebooks https://github.com/Benbucket/IBMCourseNotebooks/tree/main/CourseraIBMAppliedDataScienceCapstoneBenjamin

```
[25] data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
     data_falcon9
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Se
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	E
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	E
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	E
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	E
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	E
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	E

90 rows × 17 columns

Data Collection - Scraping

- Results from Web Scraping: Github Url: https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/2.Data_Collection_with_Web_Scraping.ipynb

```
[ ] df
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

121 rows × 11 columns

Data Wrangling

Handling Missing Values:

- Identified and addressed missing entries using imputation or removal techniques to ensure a complete dataset.

Encoding Categorical Features:

- Applied one-hot encoding to convert categorical variables into binary vectors, making them suitable for modeling.

Feature Engineering:

- Created new features or modified existing ones to enhance the dataset's predictive power.

Data Transformation:

- Standardized or normalized numerical features to bring them onto a comparable scale.

Adding Target Variable:

- Introduced a new column called 'Class' to indicate whether a launch was successful (1) or failed (0).

Final Dataset Preparation:

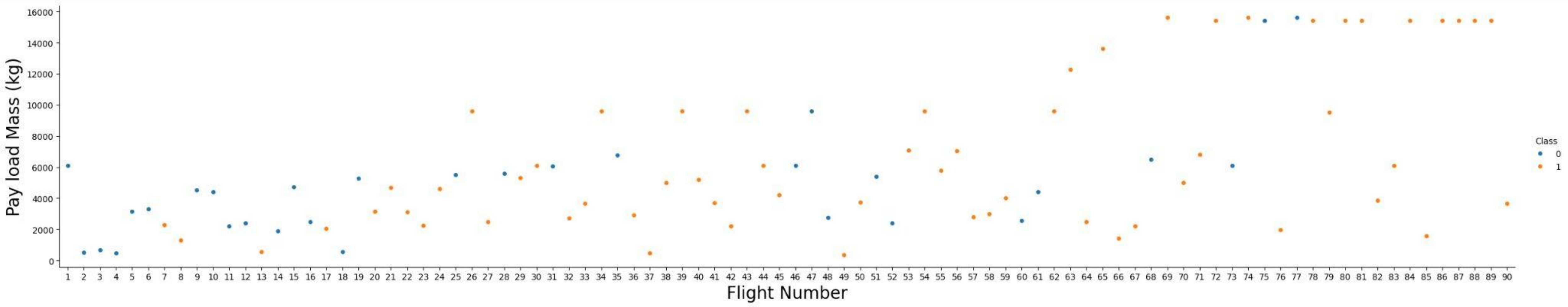
- Resulting in a dataset with 90 rows and 83 columns, ready for exploratory analysis and modeling.

https://github.com/Benbucket/IBMCourseNotebooks

EDA with Data Visualization

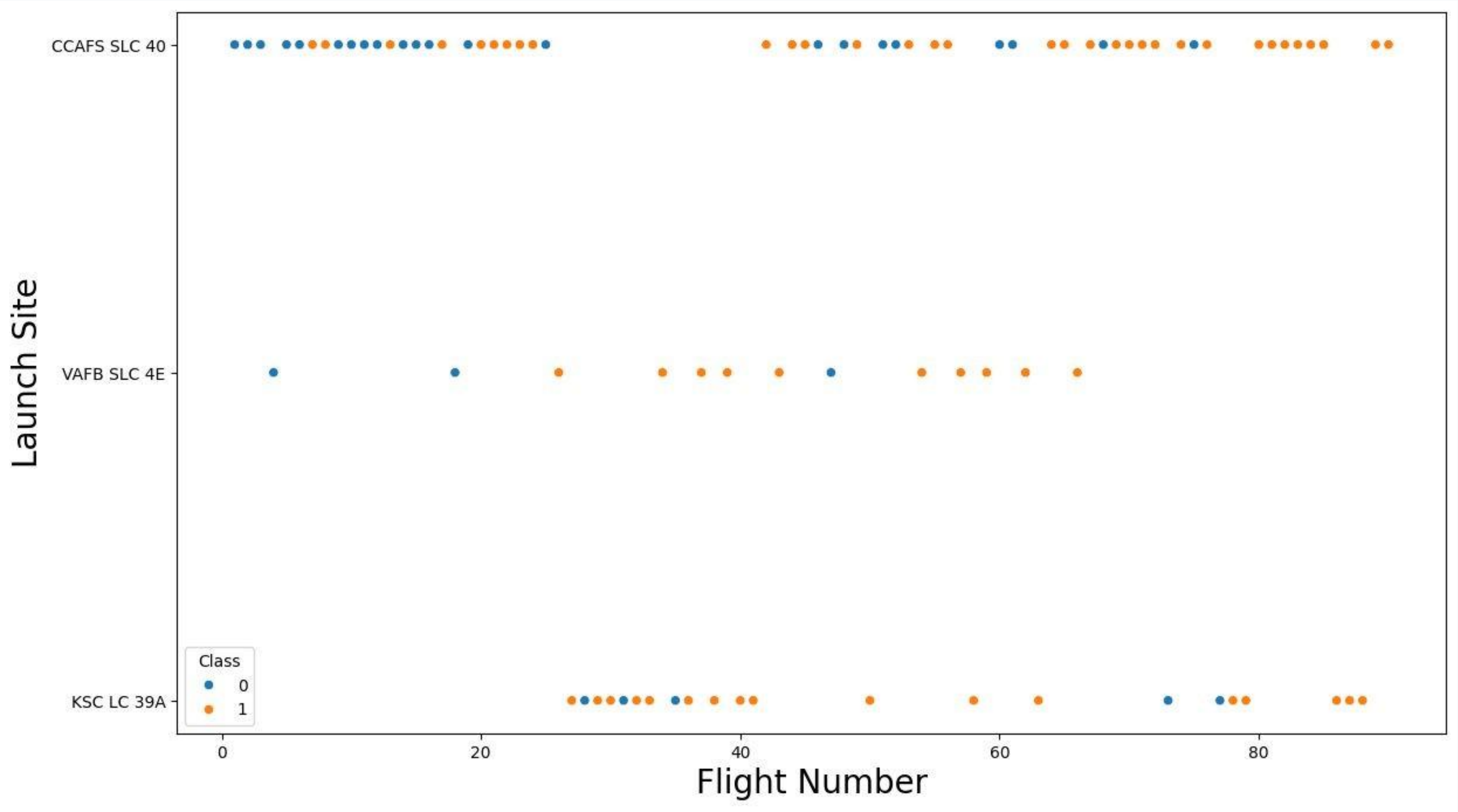
https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/5.EDA_Visualization.ipynb

Scatter Plot:



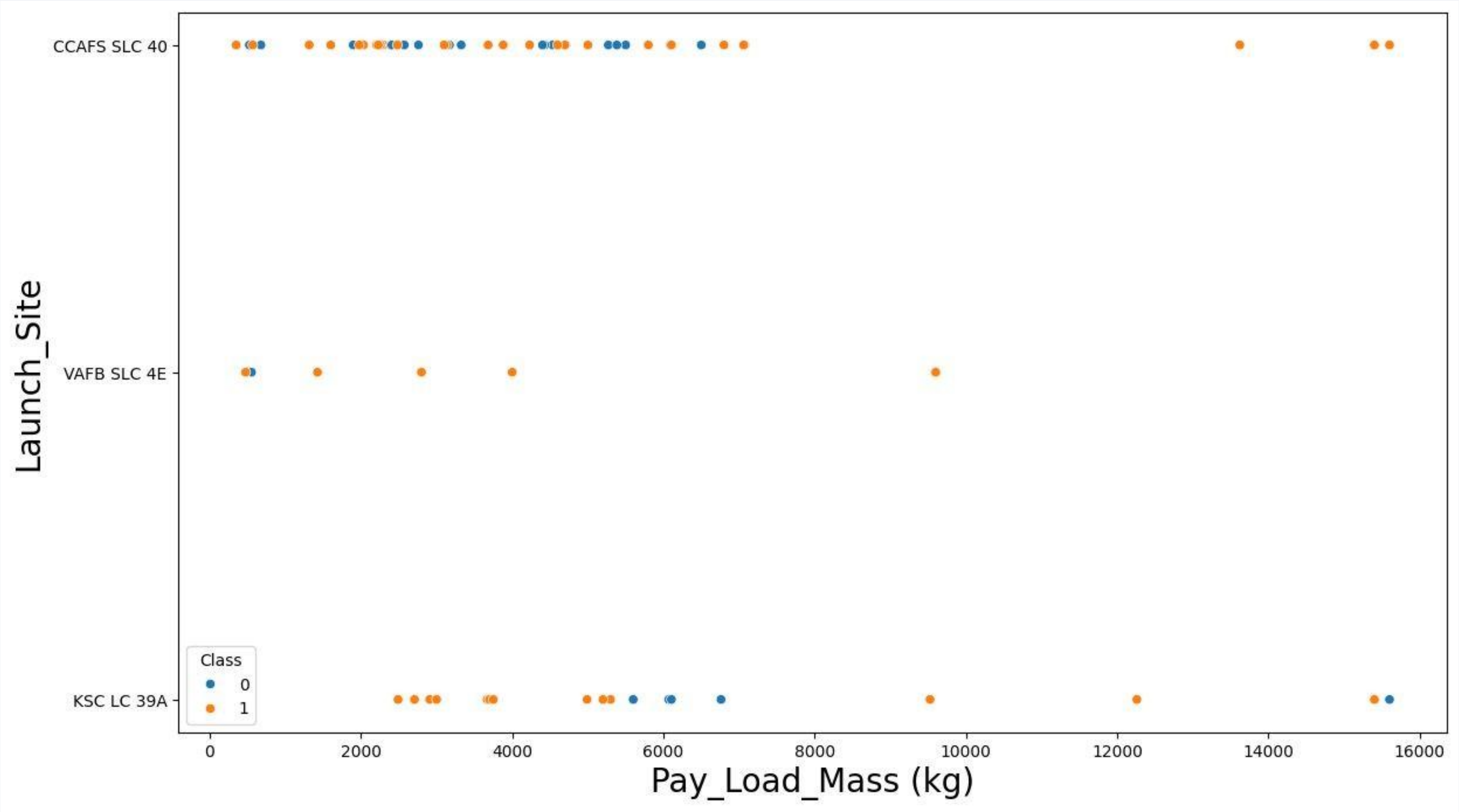
EDA with Data Visualization

Scatterplot chart to Visualize the relationship between Flight Number and Launch Site



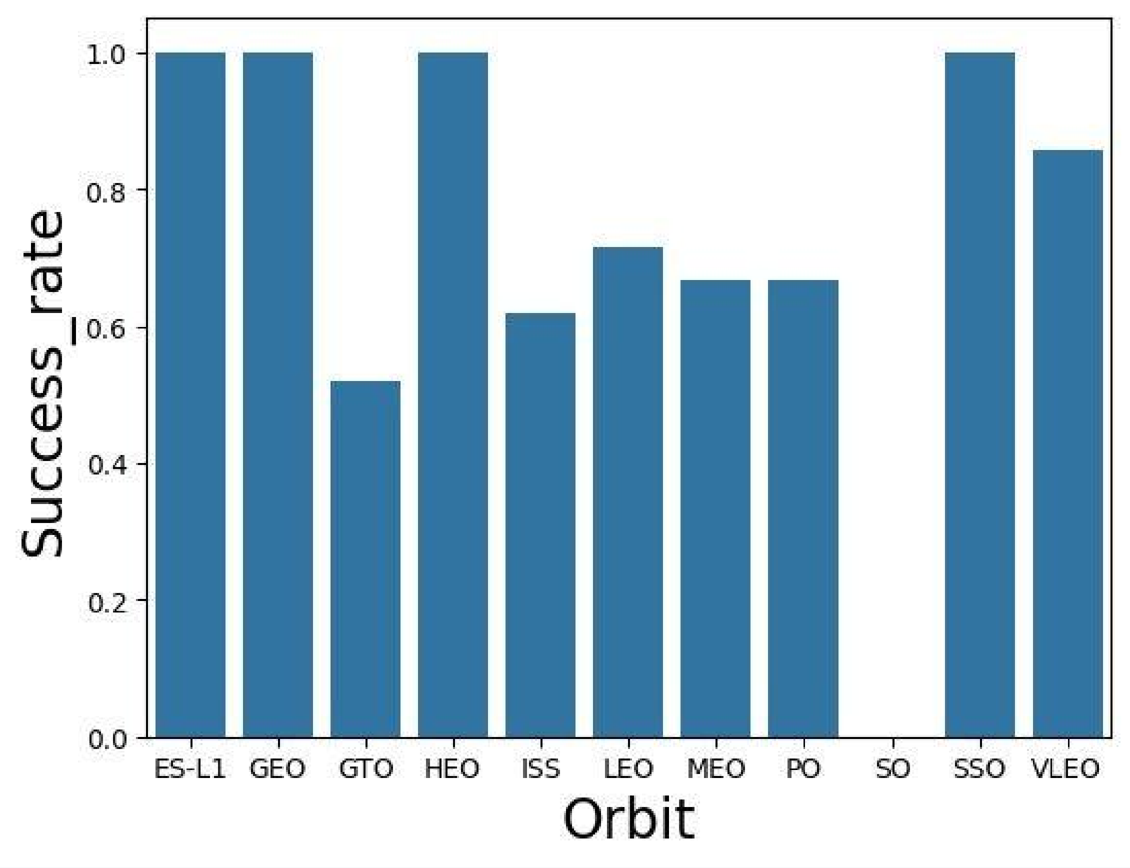
EDA with Data Visualization

Scatterplot chart to Visualize the relationship between Payload and Launch Site



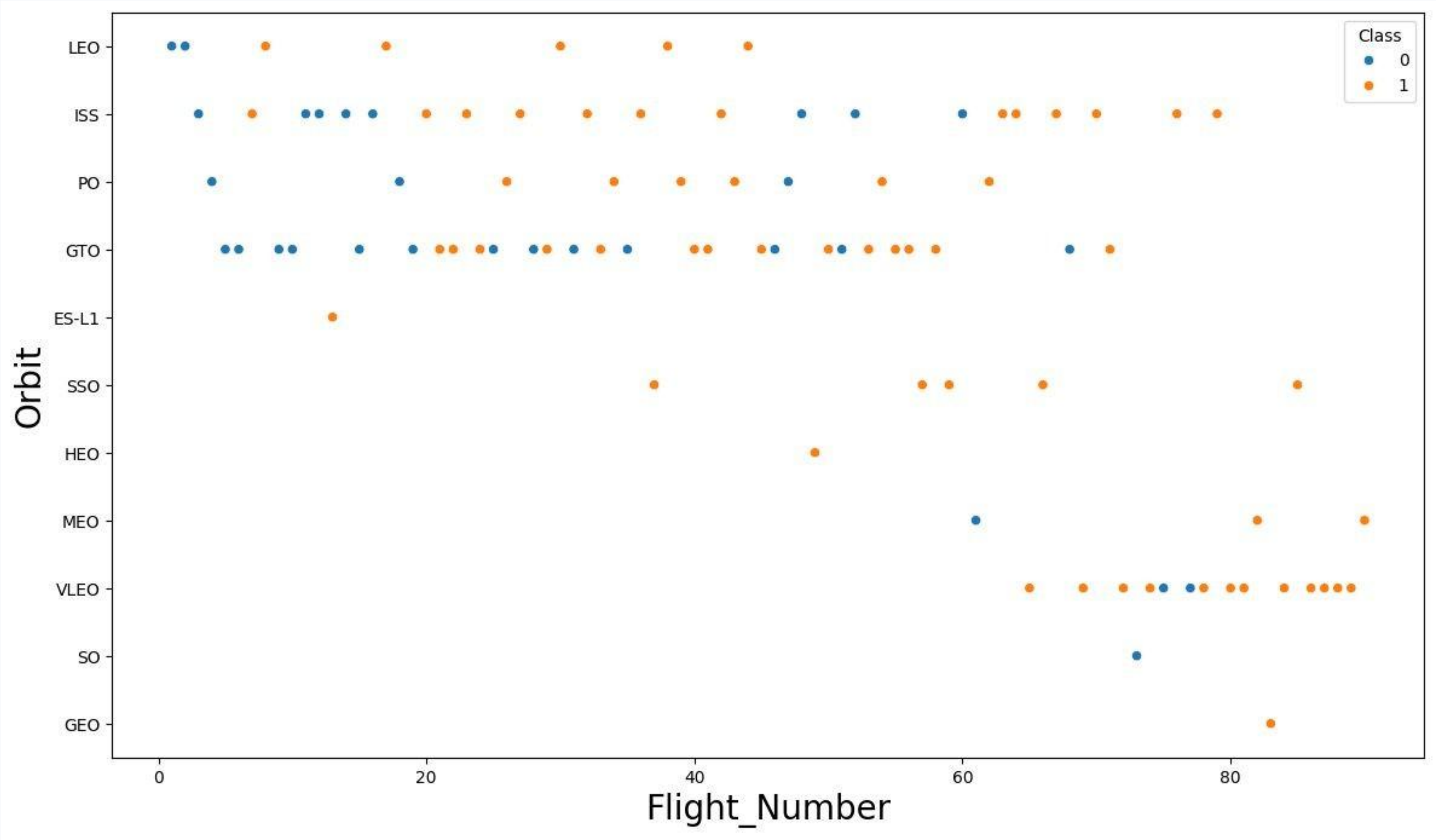
EDA with Data Visualization

Bar Chart to Visualize the relationship between success rate of each orbit type



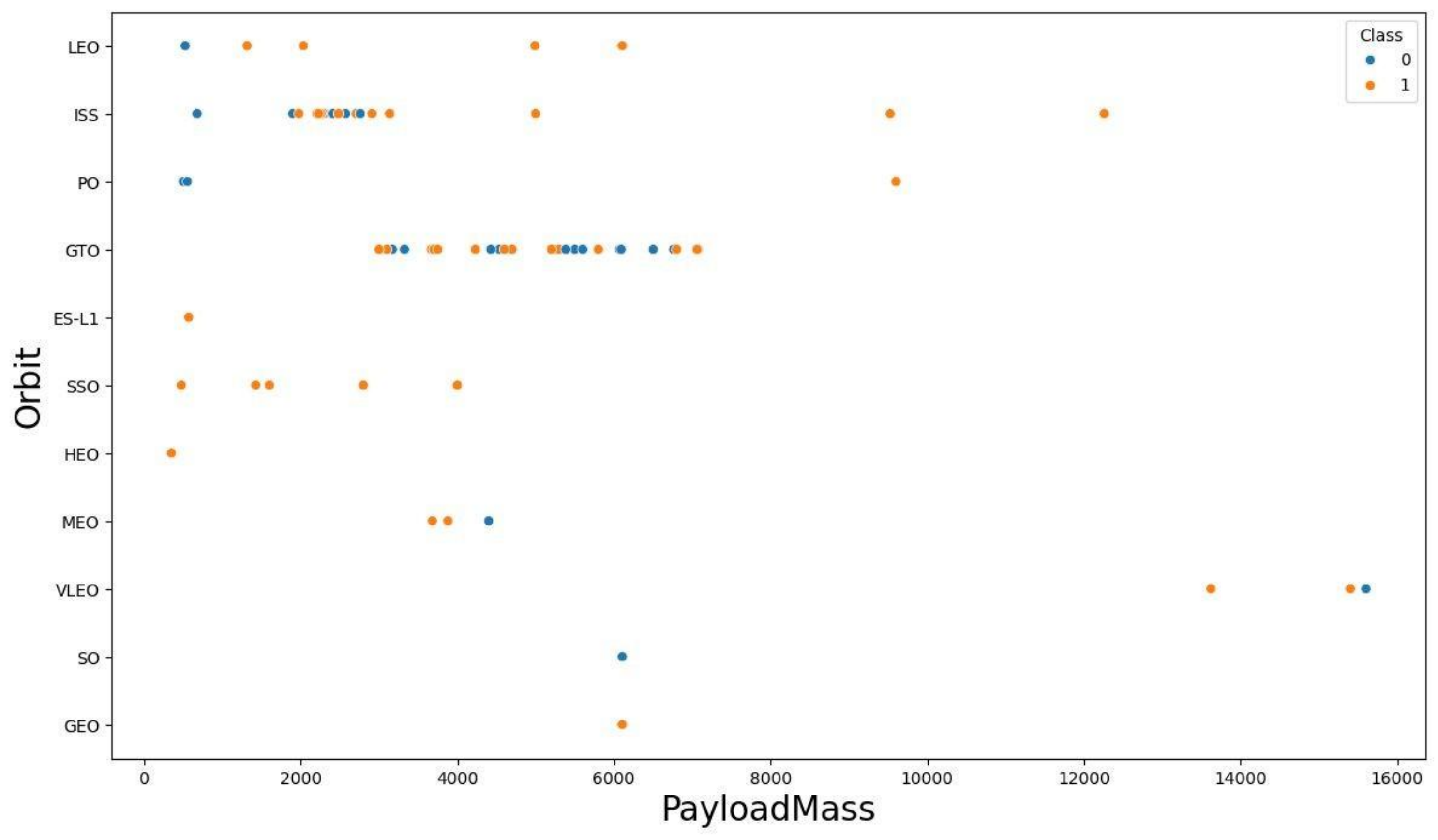
EDA with Data Visualization

Scatterplot to Visualize the relationship between FlightNumber and Orbit type



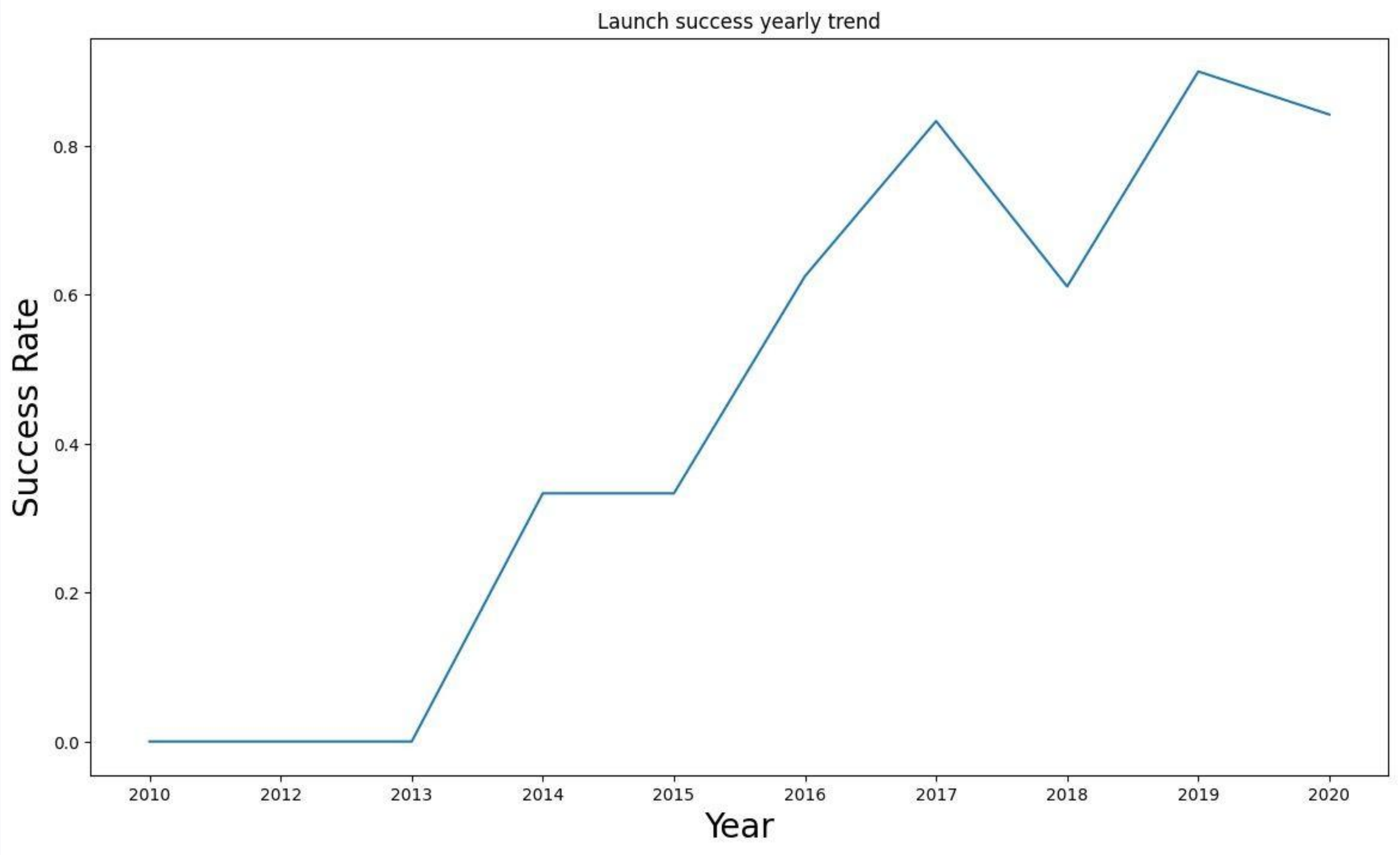
EDA with Data Visualization

Scatterplot to Visualize the relationship between Payload and Orbit type



EDA with Data Visualization

Line chart to Visualize the launch success yearly trend



EDA with SQL

- Unique Launch Sites: List all distinct launch sites used in the space missions.
- Launch Sites Starting with ‘CCA’: Provide 5 records where the launch sites have names beginning with 'CCA'.
- Total Payload Mass for NASA (CRS): Calculate the total payload mass carried by boosters launched under NASA's CRS missions.
- Average Payload Mass for Booster Version F9 v1.1: Compute the average payload mass carried by the booster version F9 v1.1.
- First Successful Ground Pad Landing Date: Identify the date when the first successful landing outcome occurred on a ground pad.
- Boosters with Successful Drone Ship Landings and Specific Payload Mass: List the names of boosters that had successful landings on a drone ship and carried a payload mass between 4000 and 6000.

https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/4.EDA_with_SQL.ipynb

EDA with SQL

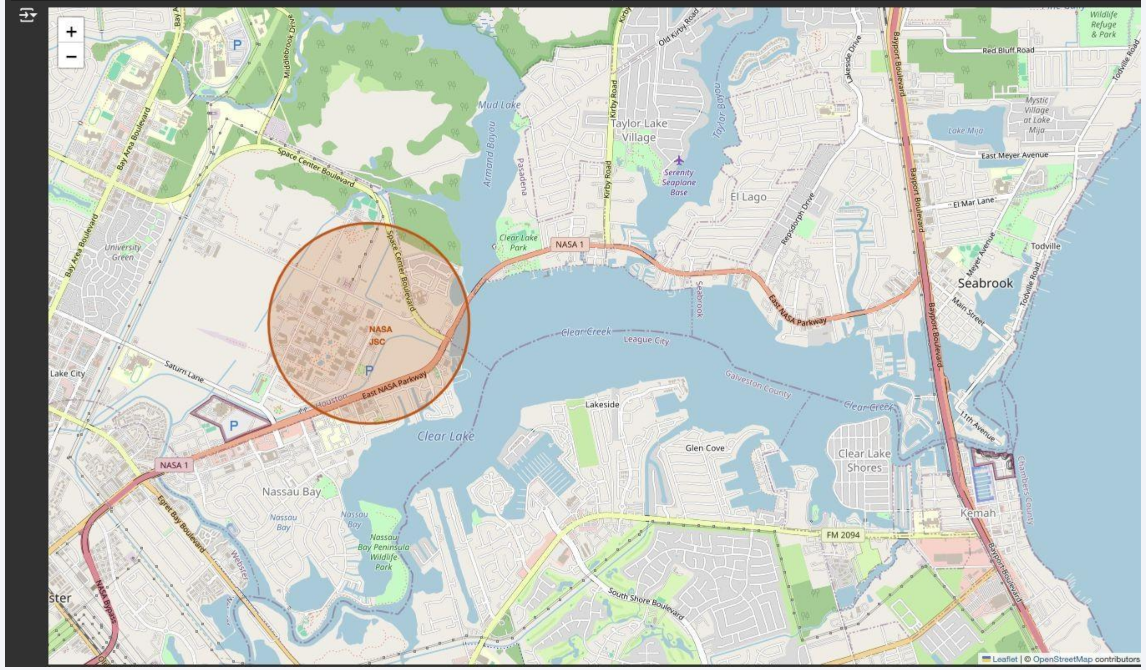
- Booster Versions with Maximum Payload Mass: Identify the booster versions that have carried the highest payload mass.
- Total Successful and Failed Mission Outcomes: Count the total number of successful and failed mission outcomes.
- Failed Drone Ship Landings in 2015: Provide details on failed drone ship landings, including booster versions and launch site names, for the year 2015.
- Landing Outcomes Count (2010-06-04 to 2017-03-20): Count the number of landing outcomes between June 4, 2010, and March 20, 2017, sorted in descending order.

https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/4.EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.

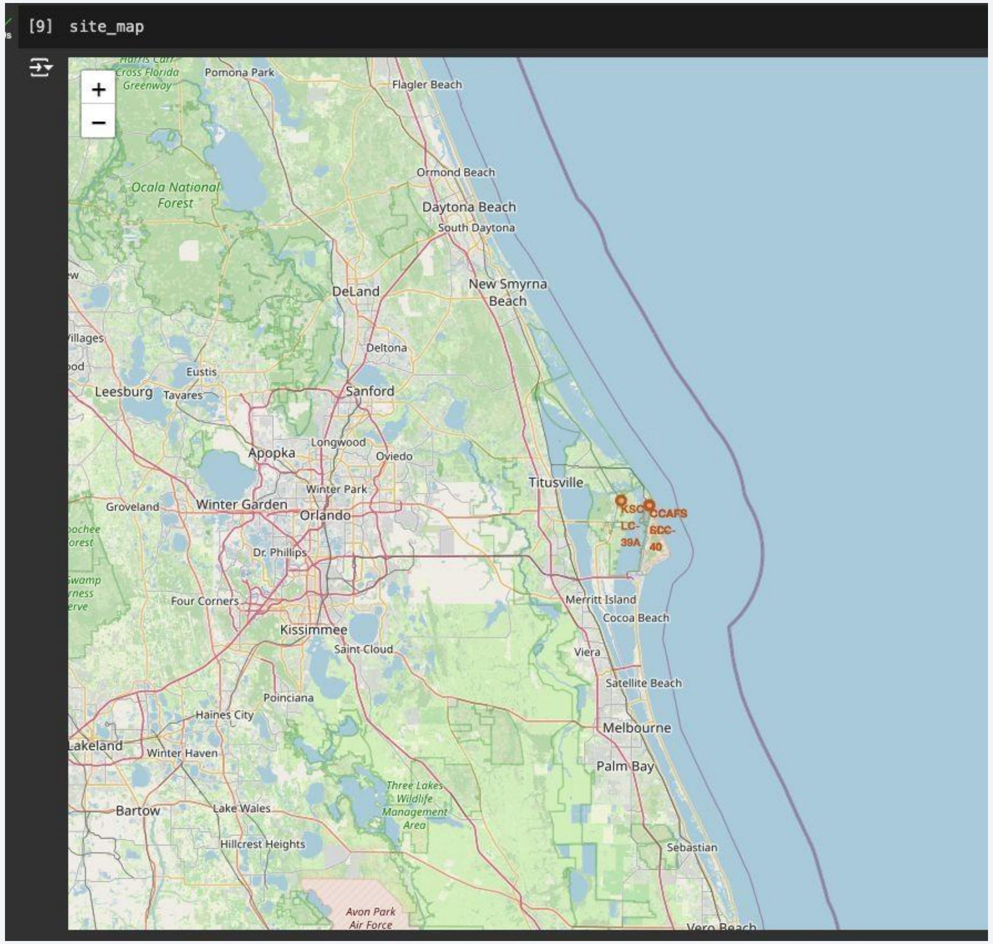
```
[7] # Create a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name
    circle = folium.Circle(nasa_coordinate, radius=1000, color='#d35400', fill=True).add_child(folium.Popup('NASA Johnson Space Center'))
    # Create a blue circle at NASA Johnson Space Center's coordinate with a icon showing its name
    marker = folium.map.Marker(
        nasa_coordinate,
        # Create an icon as a text label
        icon=DivIcon(
            icon_size=(20,20),
            icon_anchor=(0,0),
            html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>' % 'NASA JSC',
            )
        )
    site_map.add_child(circle)
    site_map.add_child(marker)
```



- https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/6.Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build an Interactive Map with Folium

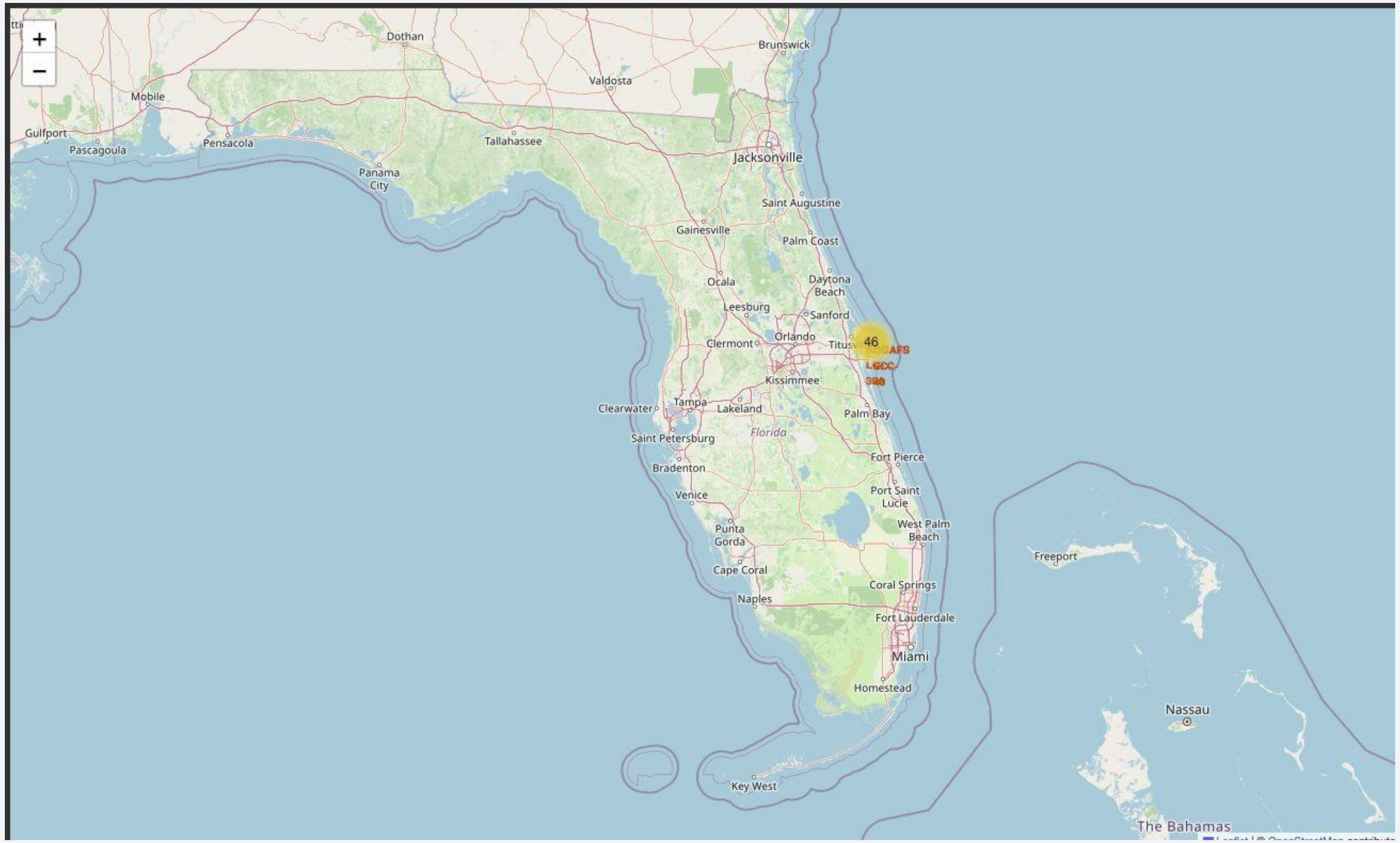
- circle for each launch site in data frame launch_sites



- https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/6.Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build an Interactive Map with Folium

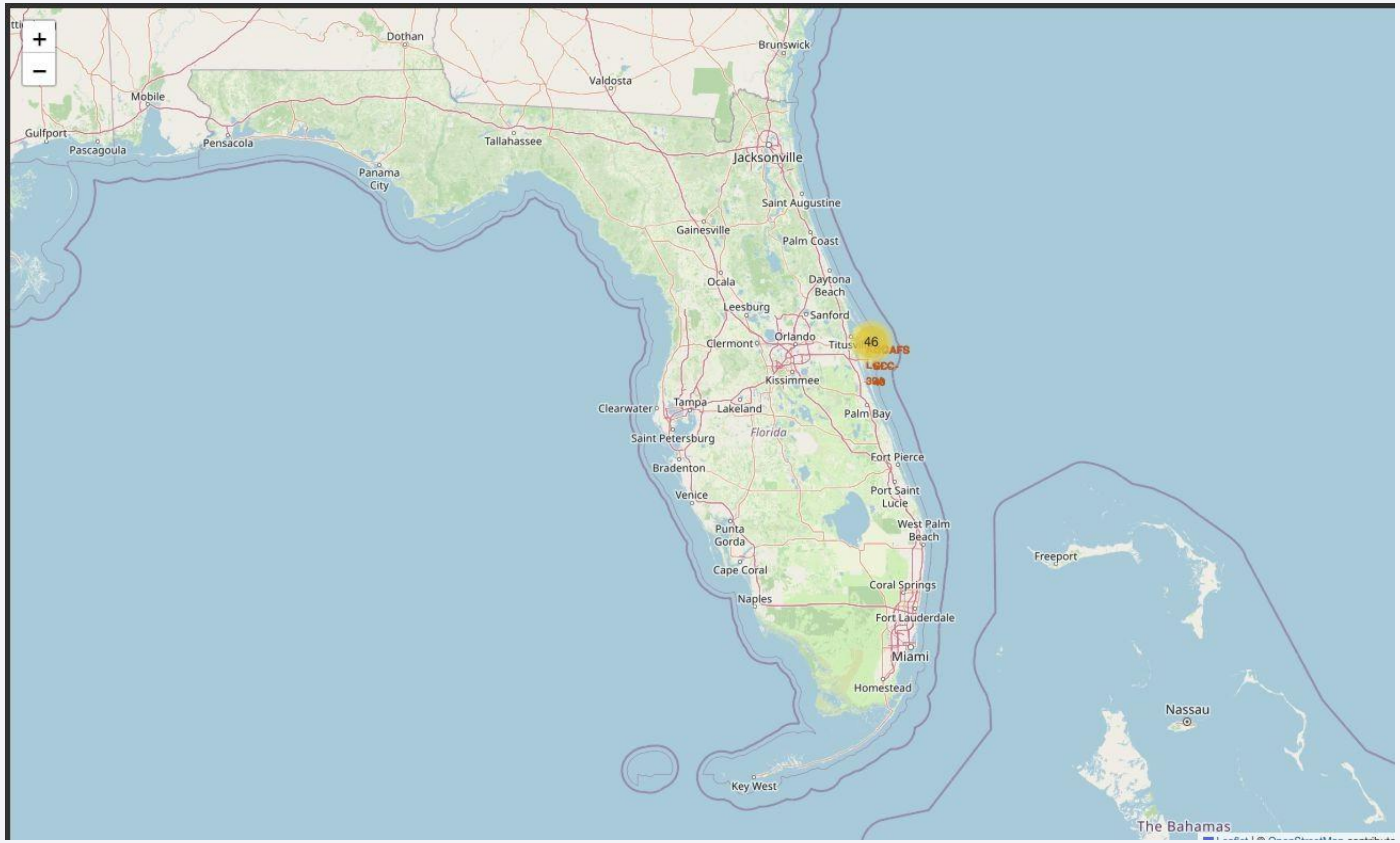
For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster



- https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/6.Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build an Interactive Map with Folium

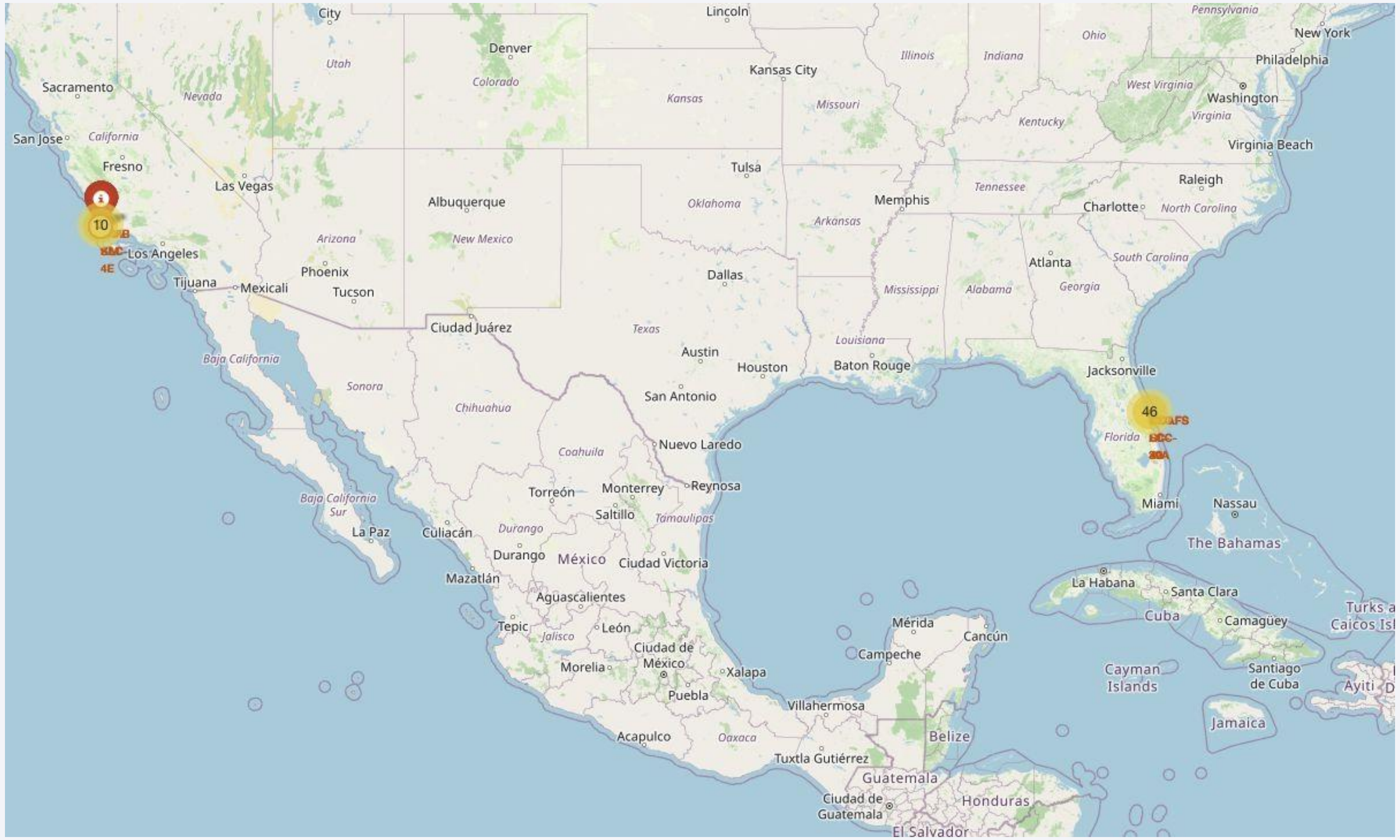
For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster



- https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/6.Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build an Interactive Map with Folium

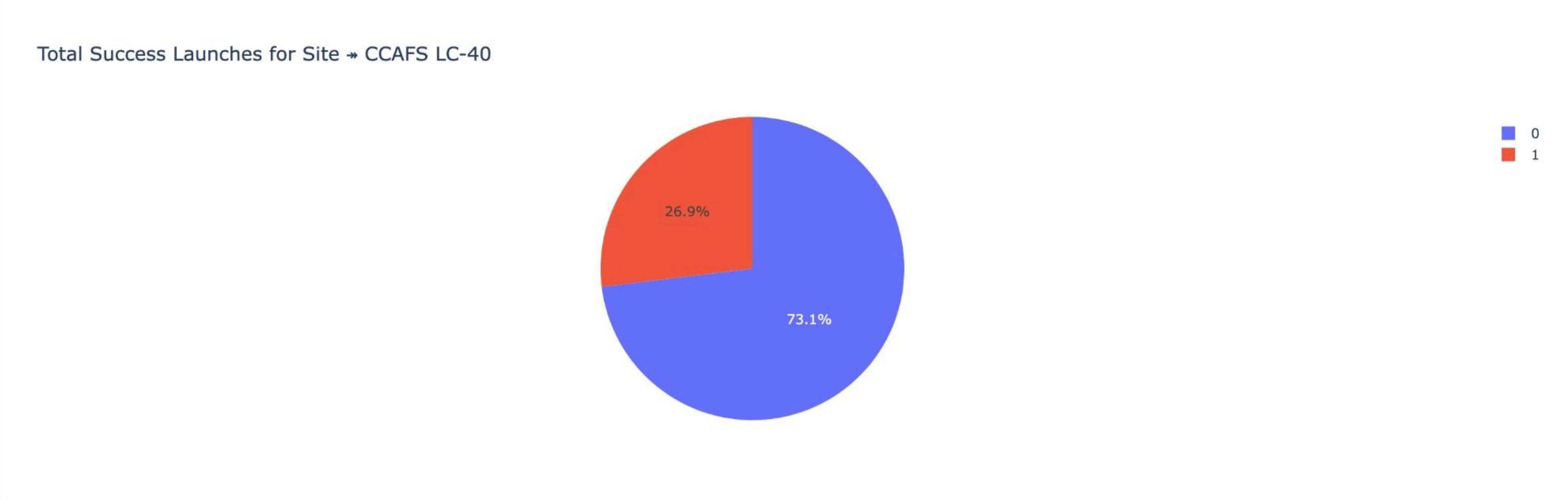
Plot the markers for each site to find the distance



- https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/6.Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build a Dashboard with Plotly Dash

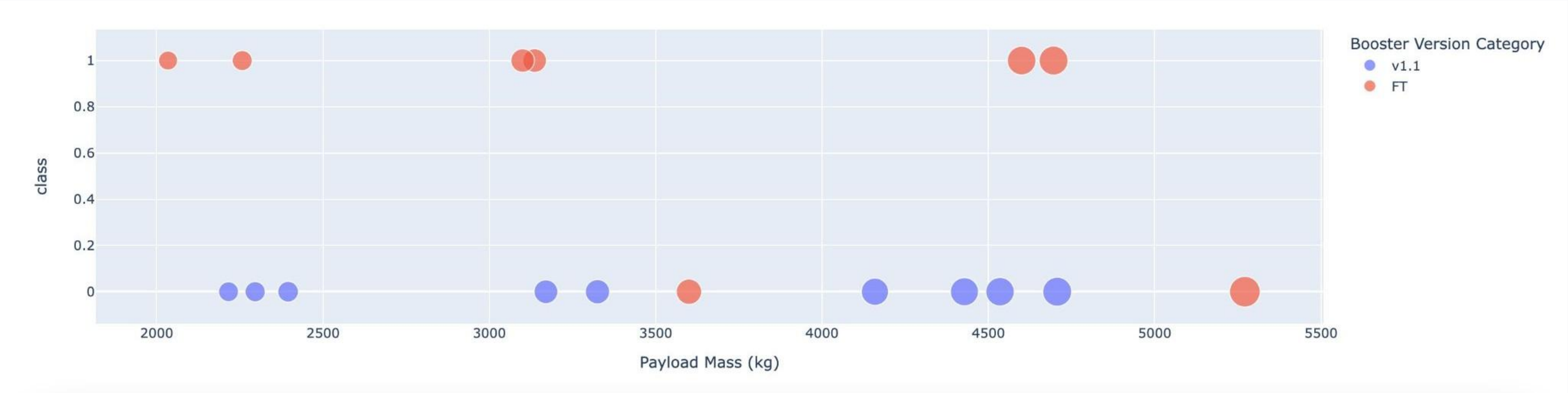
Pie chart for launch site CCAFS LC-40



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/spacex_dash_app.py

Build a Dashboard with Plotly Dash

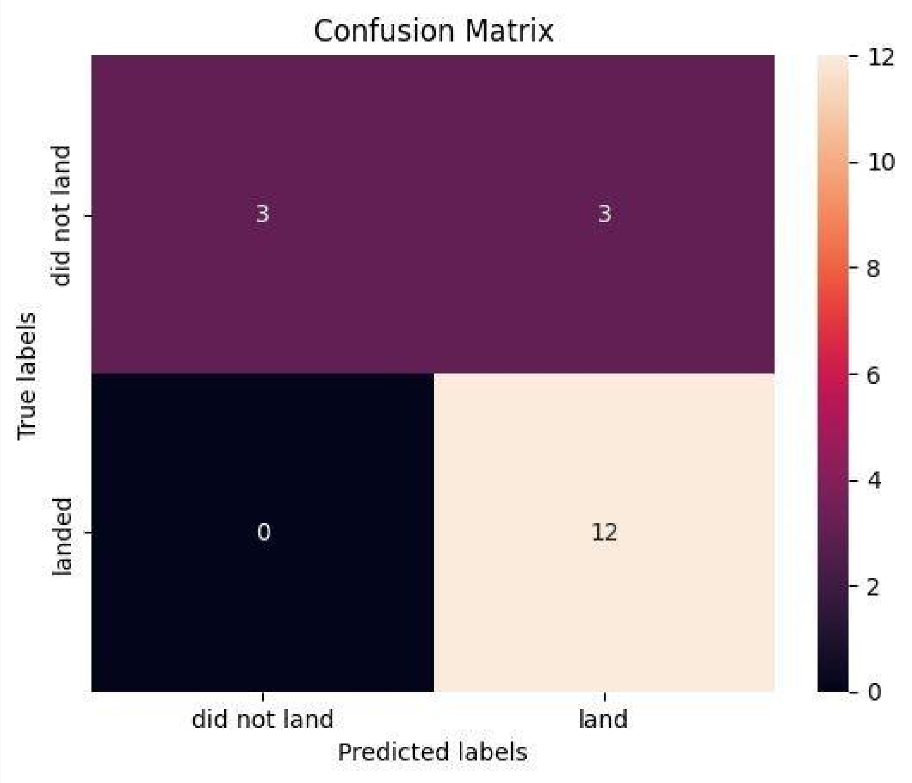
Scatter plot payload mass



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/spacex_dash_app.py

Predictive Analysis (Classification)

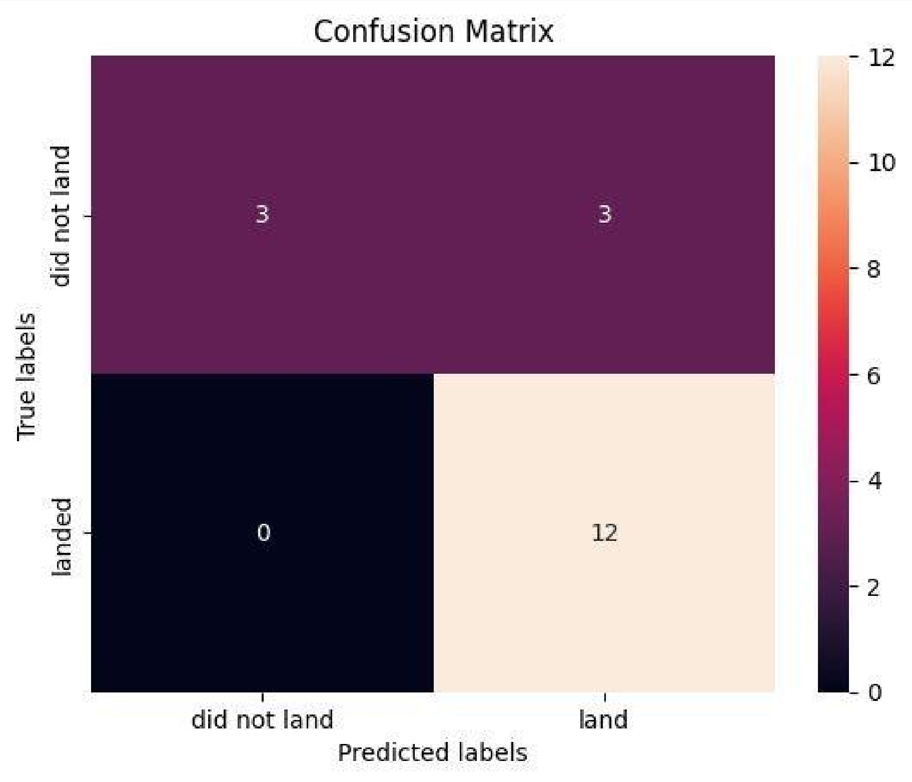
- Confusion Matrix to Analyze performance of predictions (lr)



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/7.Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

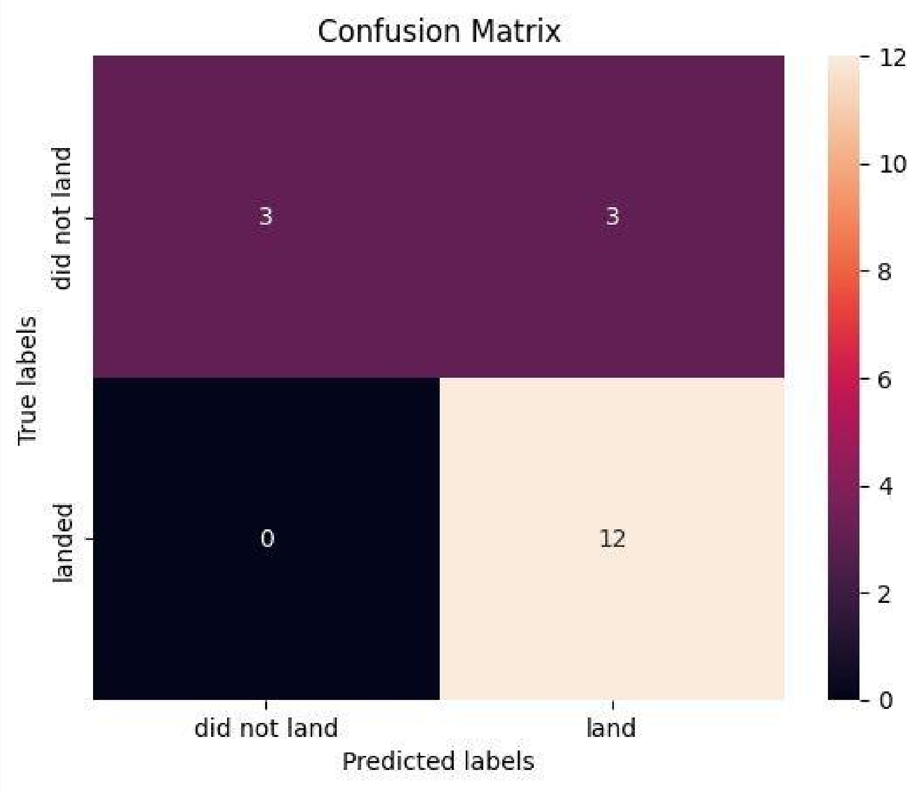
- Confusion Matrix to Analyze performance of predictions (svm)



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/7.Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

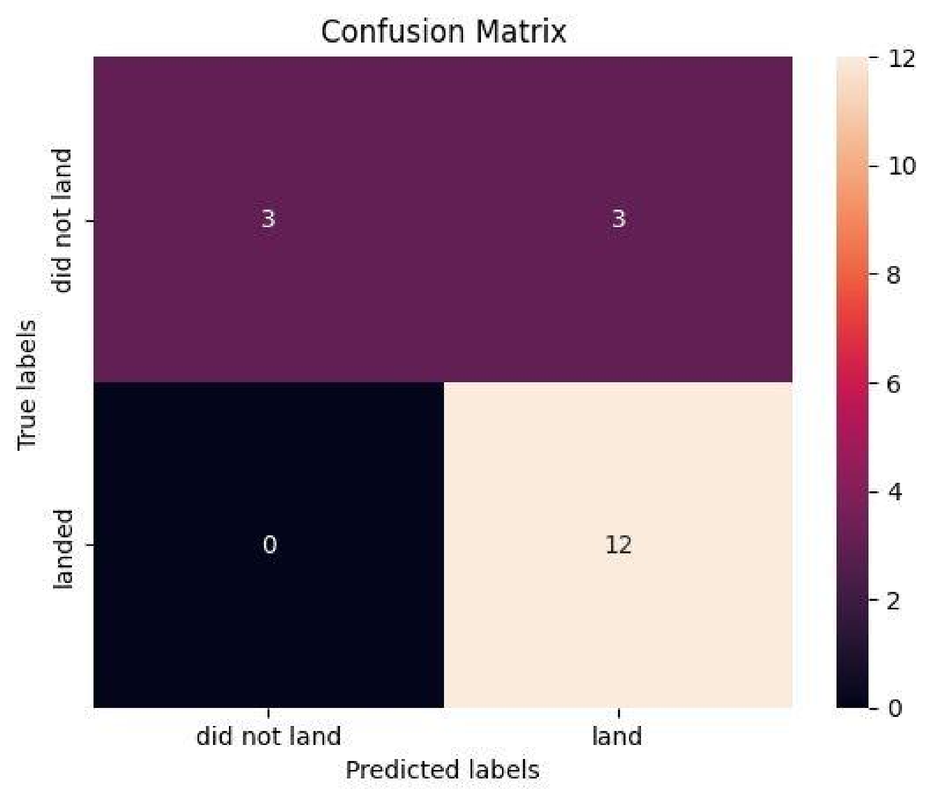
- Confusion Matrix to Analyze performance of predictions (tree_cv)



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/7.Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

- Confusion Matrix to Analyze performance of predictions (knn)



https://github.com/Benbucket/IBMCourseNotebooks/blob/main/CourseraIBMAppliedDataScienceCapstoneBenjamin/7.Machine_Learning_Prediction.ipynb

Results

Decision tree has yielded the best result out of all the models used

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.901786
KNN	0.848214

Conclusion

The analysis of SpaceX launch data has provided valuable insights into the factors influencing mission success. By examining the performance across various launch sites, we have identified which sites consistently achieve higher success rates and which may require further optimization. The correlation between payload mass and launch success has revealed important patterns, indicating that payload size can significantly impact mission outcomes. Additionally, the influence of different booster versions on launch success has highlighted specific configurations that are more likely to achieve successful missions. These findings offer actionable recommendations for improving launch strategies and optimizing resource allocation. Overall, this analysis underscores the importance of data-driven decision-making in advancing aerospace technology and enhancing the reliability of space missions.