Winning Space Race with Data Science

Yaqun Cai
Dec 04, 2024

Outline

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

Executive Summary

- **Methodologies Used for Data Analysis:**
  - **Data Collection:** Leveraged web scraping and the SpaceX API to gather relevant data.
  - **Exploratory Data Analysis (EDA):** Conducted data wrangling, visualizations, and interactive analytics to explore insights.
  - **Machine Learning:** Applied predictive models to analyze key factors influencing outcomes.
- **Key Results:**
  - Valuable data was successfully gathered from public sources.
  - EDA highlighted the most significant features for predicting launch success.
  - Machine learning predictions identified the best model and key parameters impacting the success rate, utilizing the collected data.

Introduction

- **Objective:**
  - Assess the feasibility of the new company, SpaceY, as a competitor to SpaceX.
- **Key Questions to Address:**
  - What is the most effective method to estimate total launch costs by predicting successful first-stage rocket landings?
  - What is the optimal location for launching rockets?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data Sources:
    - SpaceX API: https://api.spacexdata.com/v4/rockets/
    - Web scraping: Wikipedia - List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - The collected data was processed and enriched by creating a landing outcome label derived from outcome data after summarizing and analyzing relevant features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data was normalized, split into training and test datasets, and analyzed using four distinct classification models. The accuracy of each model was assessed by testing various parameter combinations

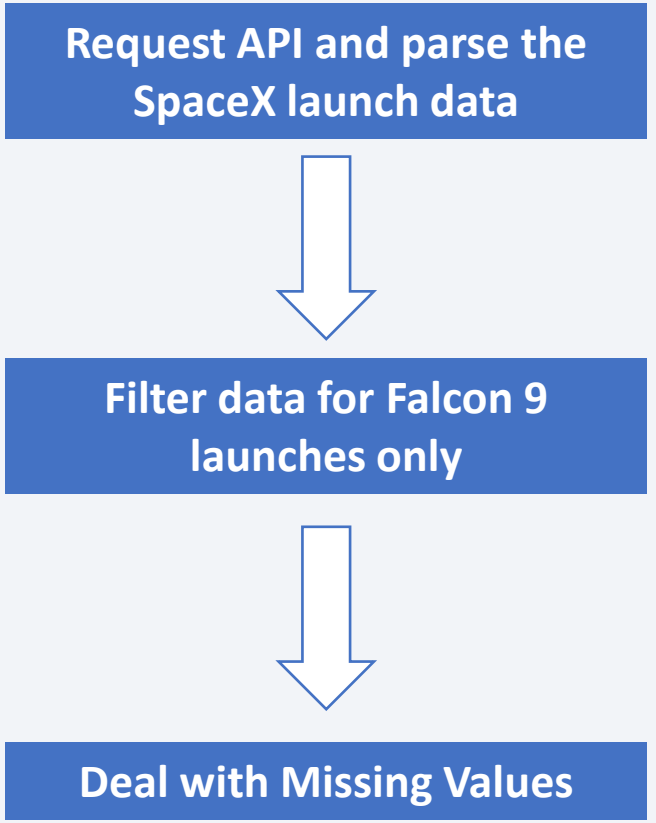
Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/)

and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

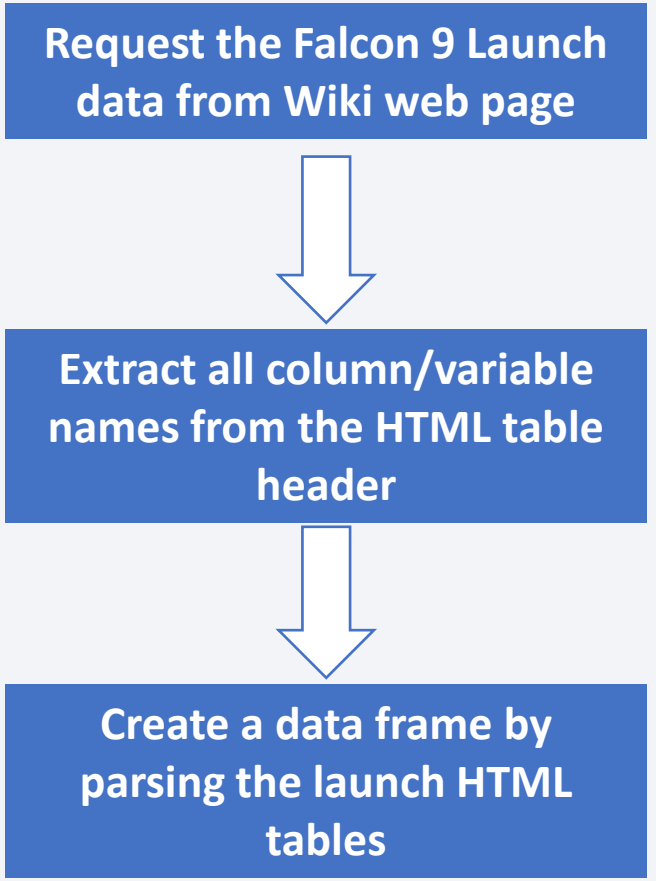
Data Collection – SpaceX API

- SpaceX provides a public API that allows data to be accessed and utilized.
- The API was utilized as outlined in the adjacent flowchart, and the data was subsequently stored.
- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-1-Collecting%20the%20data.ipynb



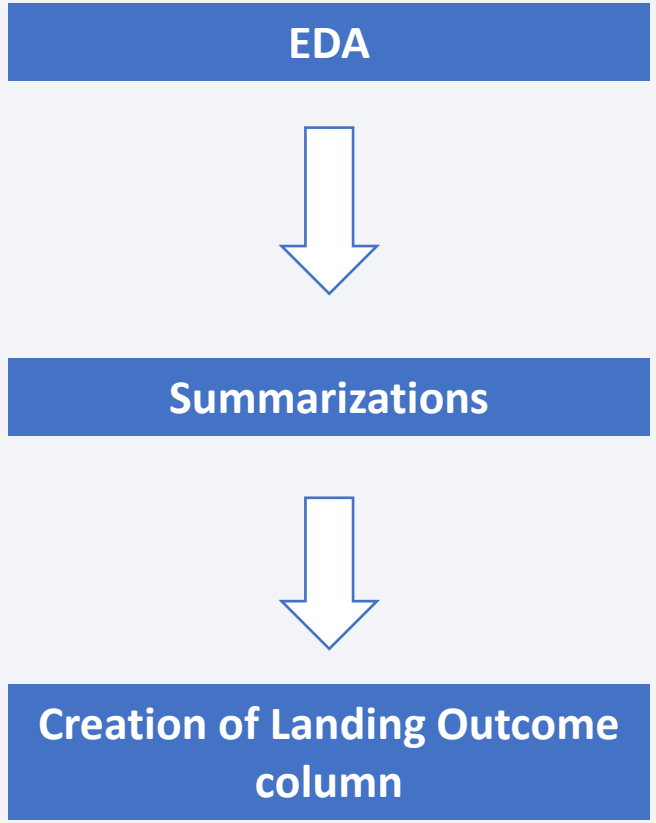
Data Collection - Scraping

- Information about SpaceX launches is also available on Wikipedia
- Data is downloaded from Wikipedia following the outlined flowchart and then stored.
- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-2-Web%20scraping.ipynb



Data Wrangling

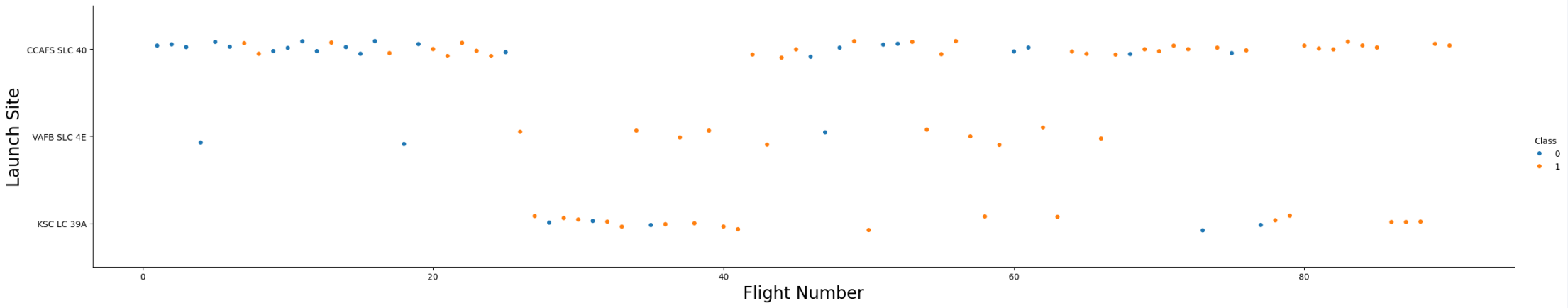
- The dataset was first subjected to Exploratory Data Analysis (EDA).
- Summaries were then generated, including launches per site, occurrences of each orbit type, and mission outcomes by orbit type.
- Lastly, the landing outcome label was derived from the "Outcome" column.



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-3-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, Scatterplots and Barplots were used to visualize the relationship between pair of features
    - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-3-Data%20wrangling.ipynb

EDA with SQL

- The following SQL queries were executed:
  - Retrieve the names of unique launch sites used in space missions
  - Identify the top 5 launch sites whose names start with the string 'CCA.'
  - Calculate the total payload mass carried by boosters launched by NASA (CRS).
  - Determine the average payload mass carried by the booster version F9 v1.1.
  - Find the date of the first successful landing on a ground pad.
  - List the names of boosters that successfully landed on a drone ship and carried payloads between 4,000 and 6,000 kg.
  - Count the total number of successful and failed mission outcomes.
  - Identify booster versions that carried the maximum payload mass.
  - Retrieve failed landing outcomes on drone ships, along with their booster versions and launch site names, for the year 2015.
  - Rank the frequency of landing outcomes (e.g., Failure on drone ship or Success on ground pad) between June 4, 2010, and March 20, 2017

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Build an Interactive Map with Folium

- Identify launch sites and adjacent site features with Folium Maps
  - Markers represent specific points, such as launch sites.
  - Circles highlight areas around specific coordinates, like the NASA Johnson Space Center.
  - Marker clusters group events within each coordinate, such as launches at a particular site.
  - Lines are used to show distances between two coordinates.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb

Build a Dashboard with Plotly Dash

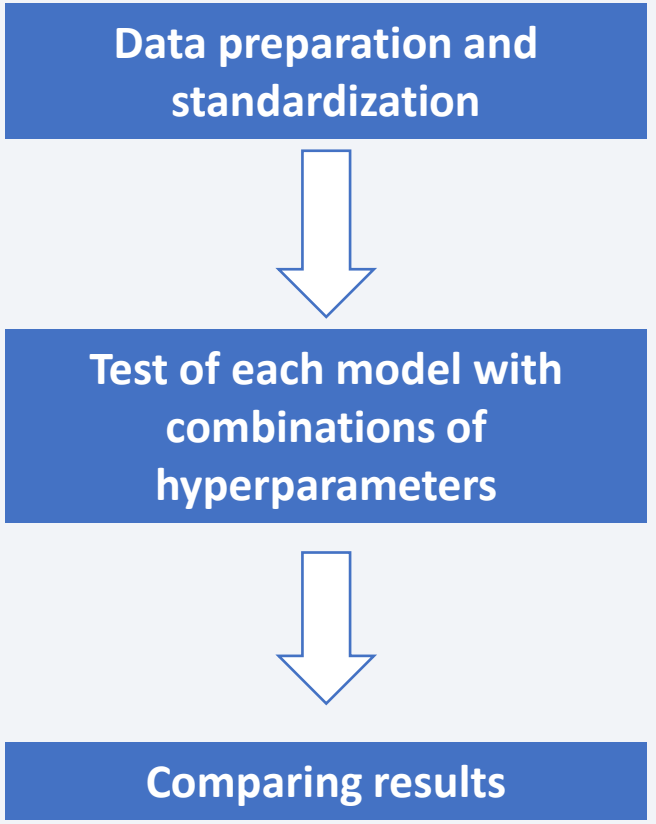
- The following items were used to visualize data
  - Percentage of Successful Launches by Site
  - The relationship between Payload range & Launch Outcome

- These 2 items enabled a quick analysis of the relationship between payloads and launch sites, helping to identify the optimal launch locations based on payloads.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb

Predictive Analysis (Classification)

- Four classification models were evaluated: logistic regression, support vector machines, decision trees, and k-nearest neighbors.



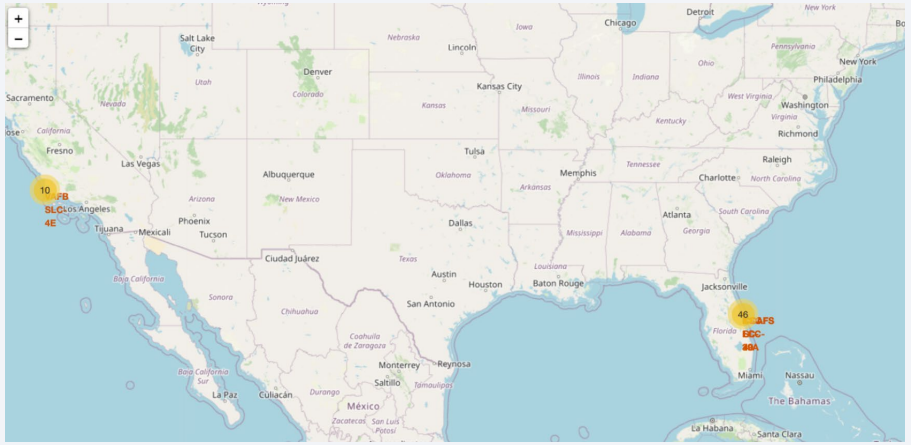
- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-8-ML.ipynb

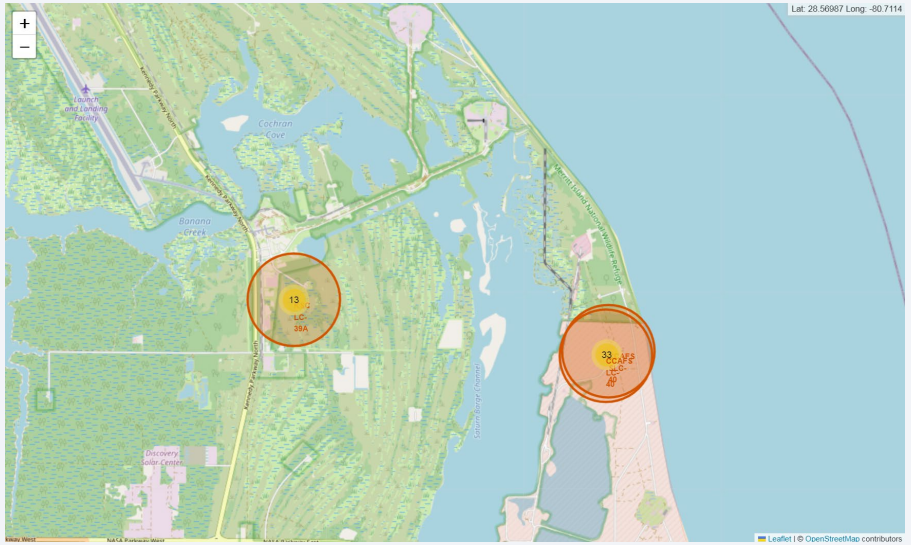
Results

- Exploratory data analysis results
  - SpaceX operates from four distinct launch sites.
  - The initial launches were conducted for SpaceX and NASA.
  - The F9 v1.1 booster has an average payload of 2,928 kg.
  - The first successful landing occurred in 2015, five years after the initial launch.
  - Several Falcon 9 booster versions achieved successful landings on drone ships with payloads exceeding the average.
  - Nearly 100% of mission outcomes were successful.
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015.
  - The success rate of landing outcomes has improved over the years.

Results

- Interactive analytics revealed that launch sites are typically located in safe areas, such as near the sea, and are supported by well-developed logistical infrastructure.
- The majority of launches occur at East Coast launch sites.



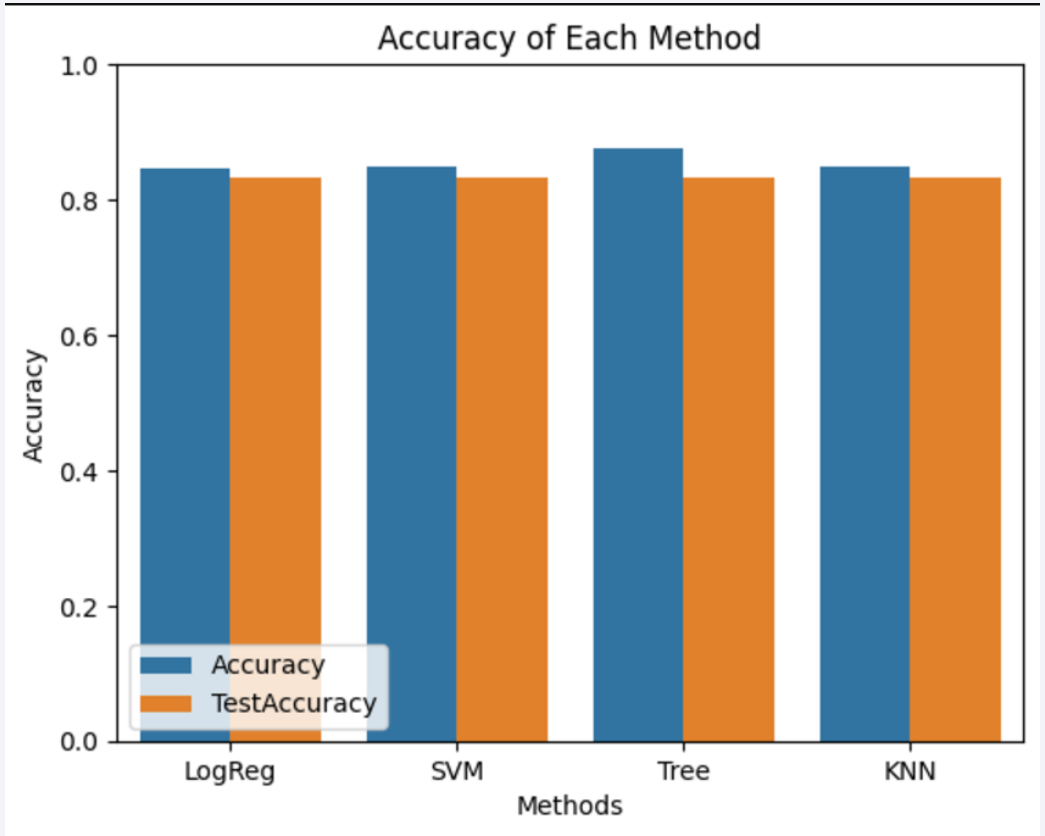


- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb

Results

- Predictive analysis demonstrated that the Decision Tree Classifier is the most effective model for predicting successful landings, achieving over 87% accuracy and exceeding 94% accuracy on test data.

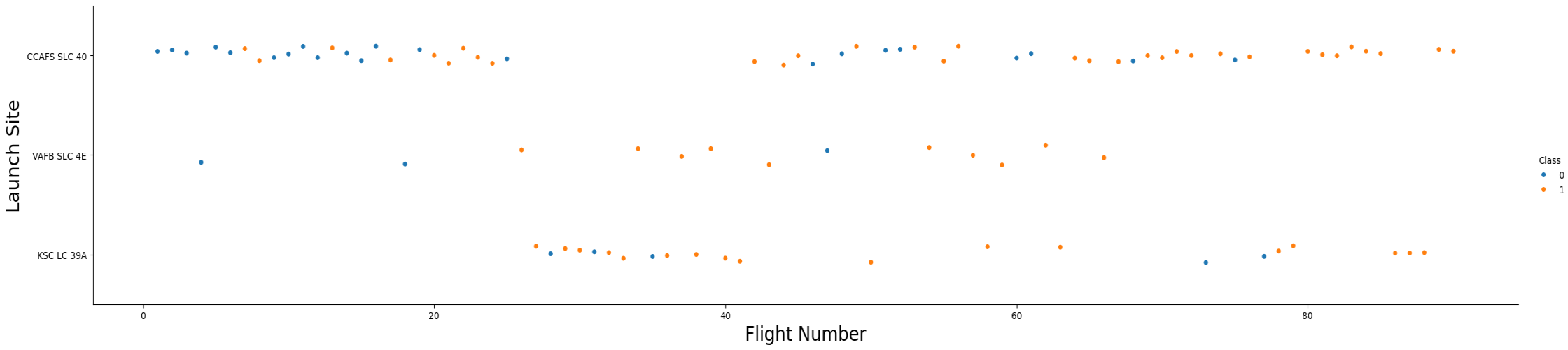
- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-8-ML.ipynb



Section 2

Insights drawn from EDA

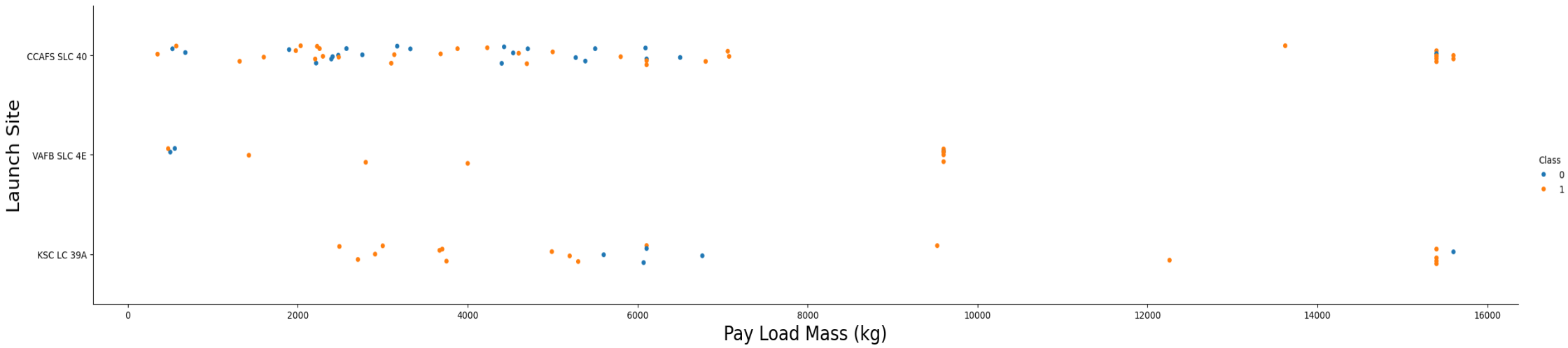
Flight Number vs. Launch Site



- The plot above indicates that the most reliable launch site currently is CCAF5 SLC 40, where the majority of recent launches have been successful.
- VAFB SLC 4E ranks second, followed by KSC LC 39A in third place.
- Additionally, the overall success rate has shown improvement over time.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

Payload vs. Launch Site

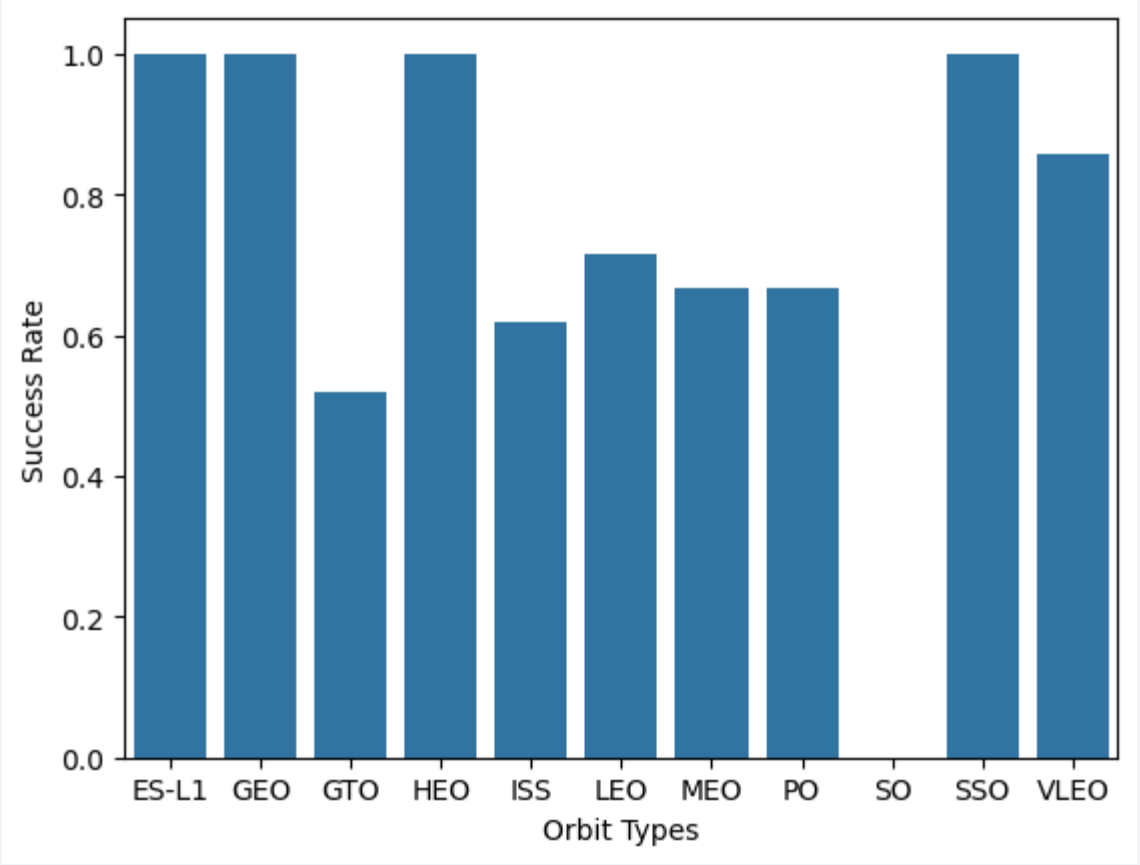


- Payloads exceeding 9,000 kg (approximately the weight of a school bus) have an excellent success rate.
- Payloads over 12,000 kg appear to be feasible exclusively at the CCAFS SLC 40 and KSC LC 39A launch sites.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

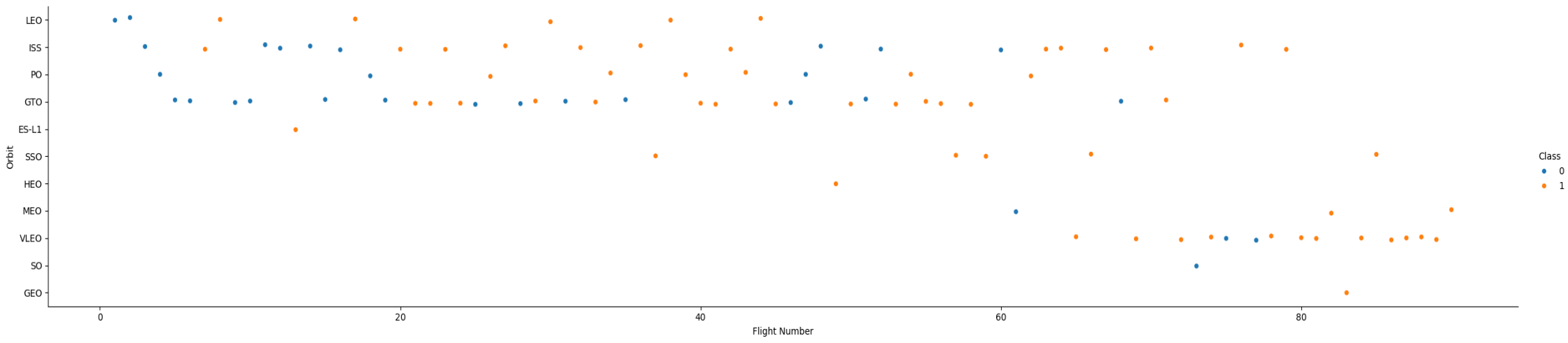
Success Rate vs. Orbit Type

- The highest success rates are achieved in the following orbits:
  - ES L1
  - GEO
  - HEO
  - SSO
- These are followed by:
  - VLEO, with success rates above 80%
  - LFO, with success rates exceeding 70%



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

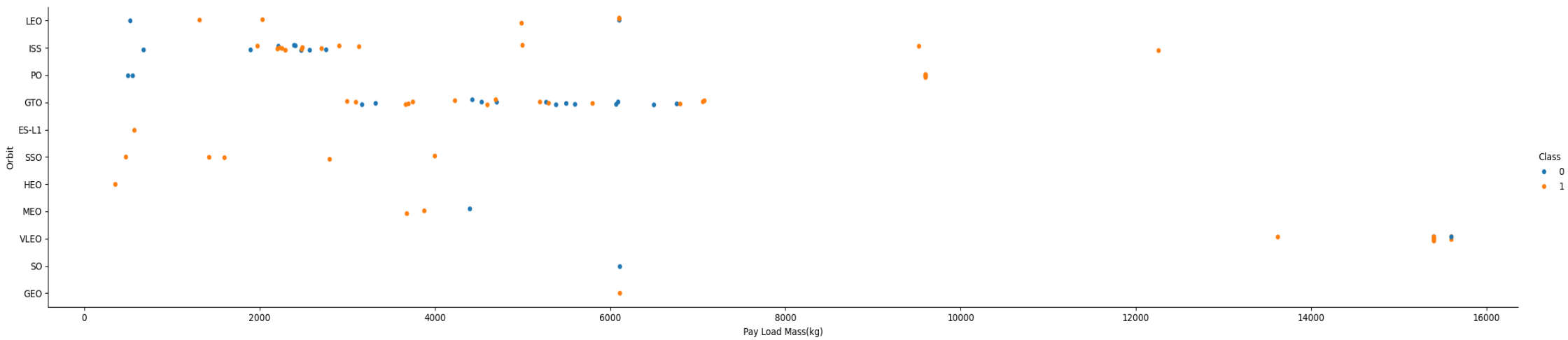
Flight Number vs. Orbit Type



- The success rate improved over time at all orbits
- VLEO orbit shows increasing launch frequency

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

Payload vs. Orbit Type

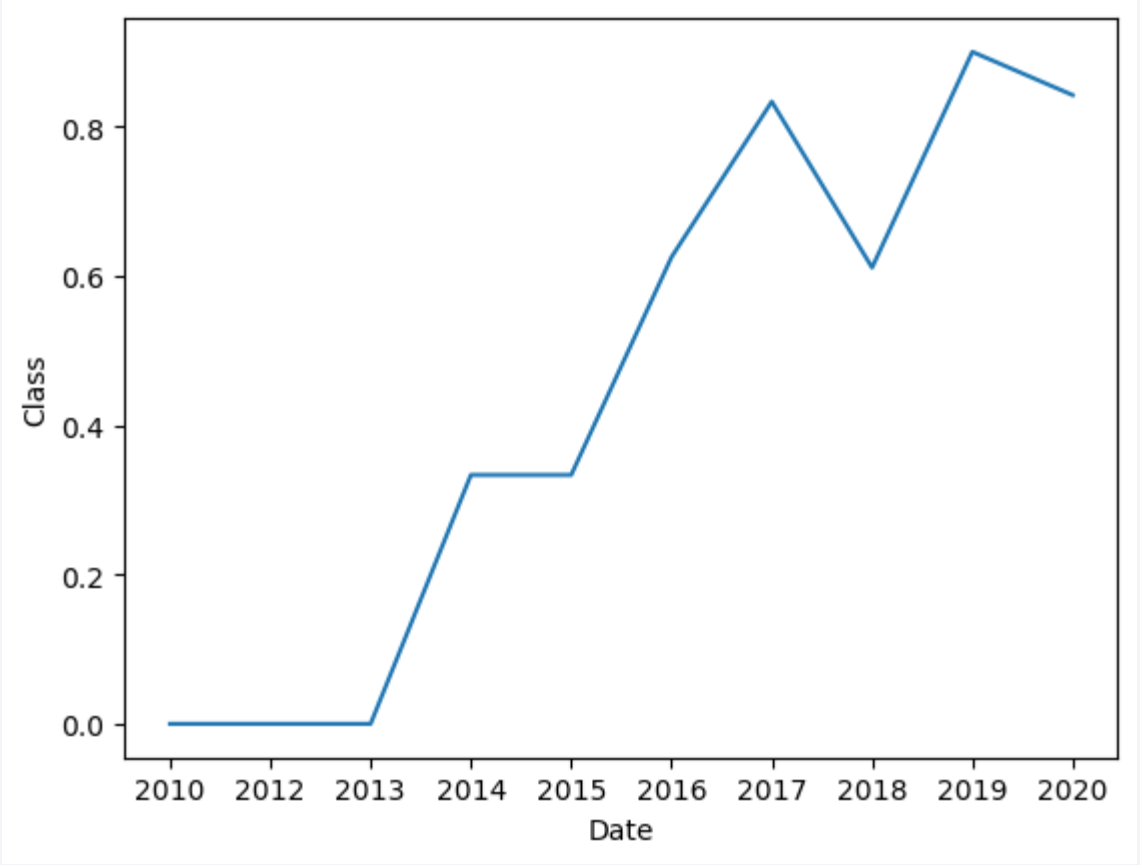


- There appears to be no correlation between payload size and success rate for the GTO orbit.
- The ISS orbit supports the widest range of payloads while maintaining a strong success rate.
- Launches to the SO and GEO orbits are relatively infrequent.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

Launch Success Yearly Trend

- Successful launch rate increases in 2013 until 2020;
- The first three years appear to have been a period of adjustment and technological refinement.



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb

All Launch Site Names

- 4 launch sites:

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

- These are derived by identifying unique occurrences of the "launch_site" values in the dataset.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

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

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Total Payload Mass

- Total payload by boosters from NASA:

Total_Payload_Mass
45596

- Total payload calculated above, by summing all payloads whose codes contain ‘ CRS’, which corresponds to NASA.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Average_Payload_Mass
2928.4

- By filtering the data for the specified booster version and calculating the average payload mass, the average payload mass is 2,928 kg.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

min(Date)
2015-12-22

- Filtering the data for successful landing outcomes on ground pads and identifying the earliest date reveals the first occurrence, which took place on December 22, 2015.

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

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:

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

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- Grouping mission outcomes and total records for each group led to the summary above

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

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

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb

2015 Launch Records

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

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

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.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

count(Landing_Outcome)	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)

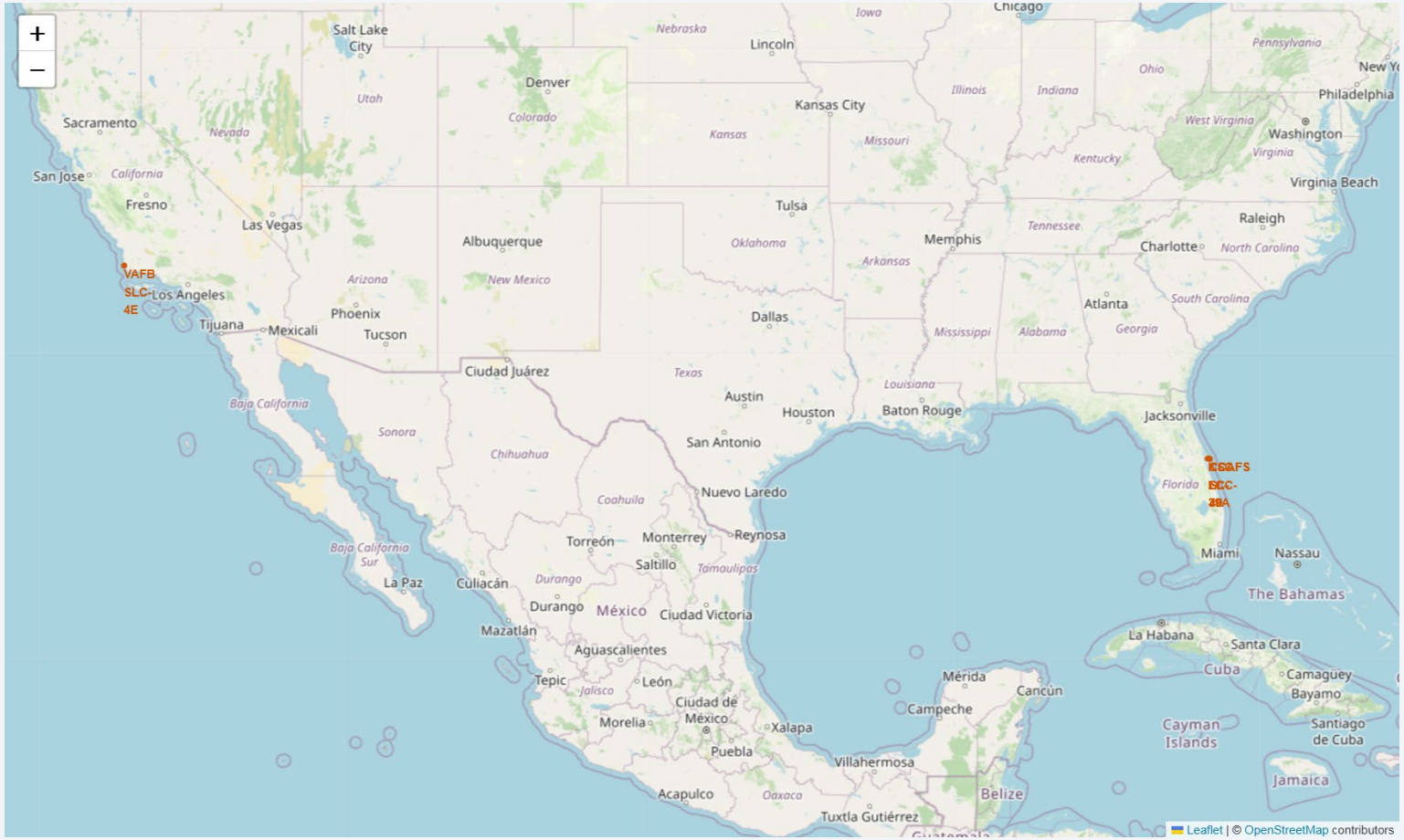
count(Landing_Outcome)	Landing_Outcome
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb


Section 3
Launch Sites
Proximities Analysis

All Launch Sites

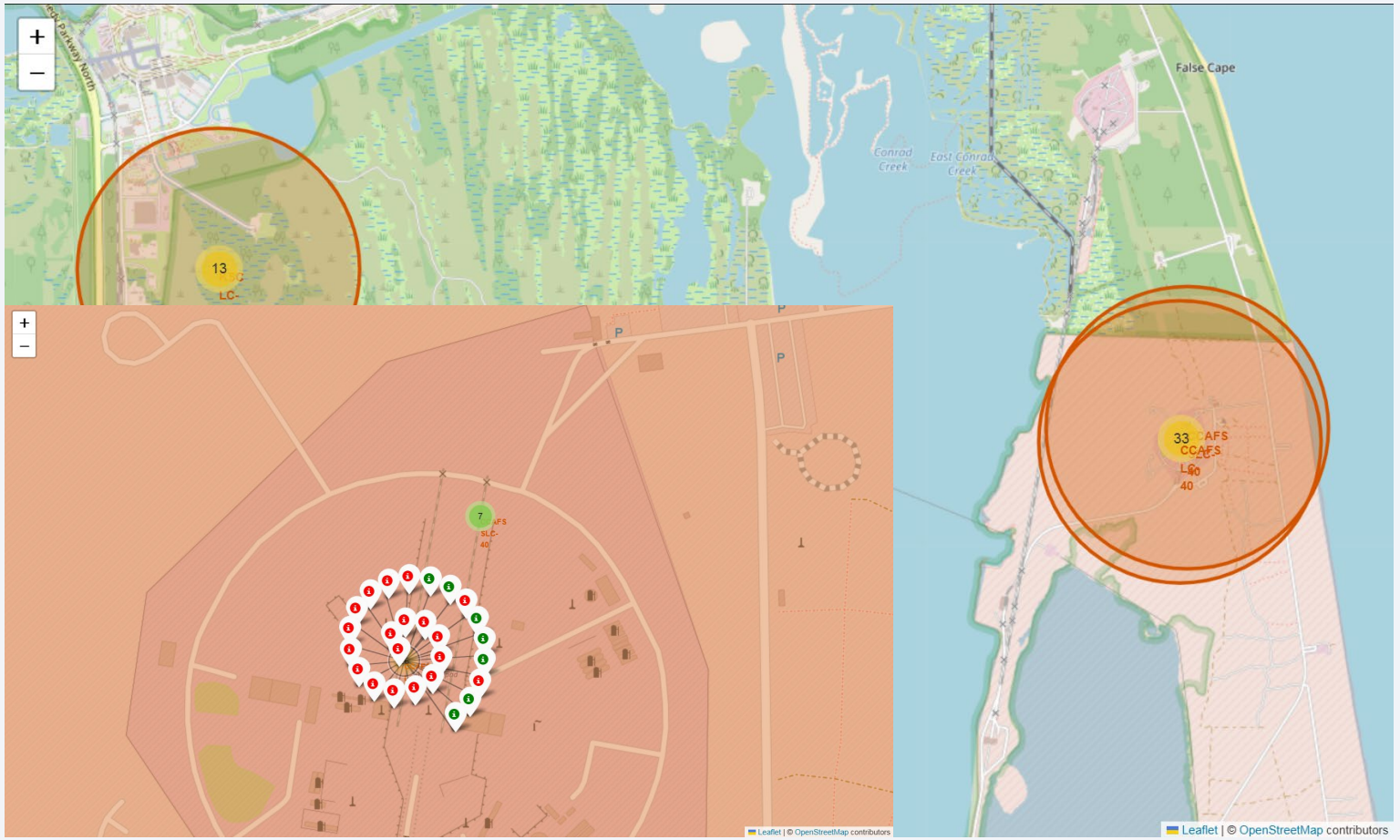
- Launch sites locate near coast, close to roads & railroads



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb

Launch Outcome by Site

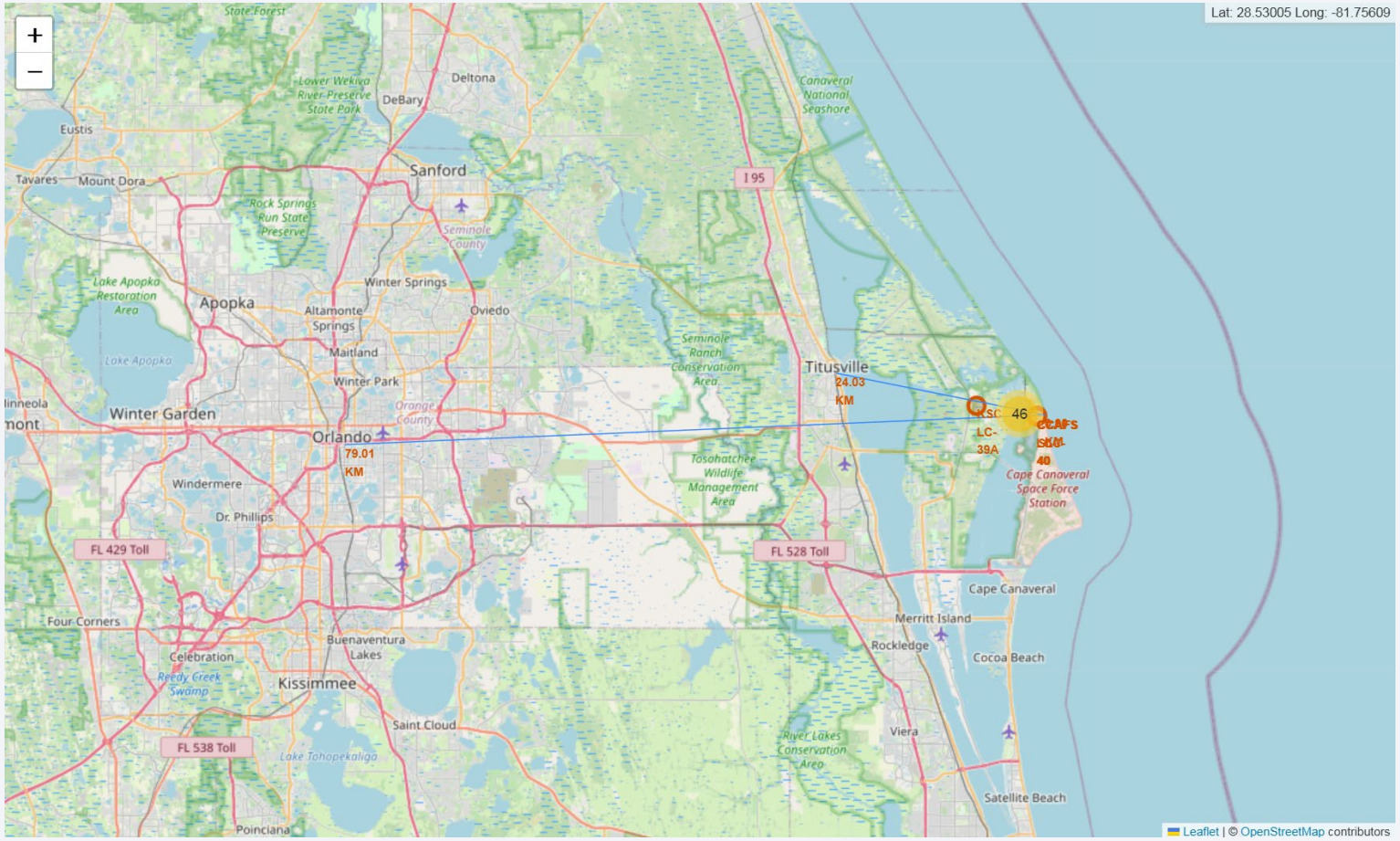
- Example of CCAFS LC-40 launch site
- Green markers show successful outcome and red markers show failure



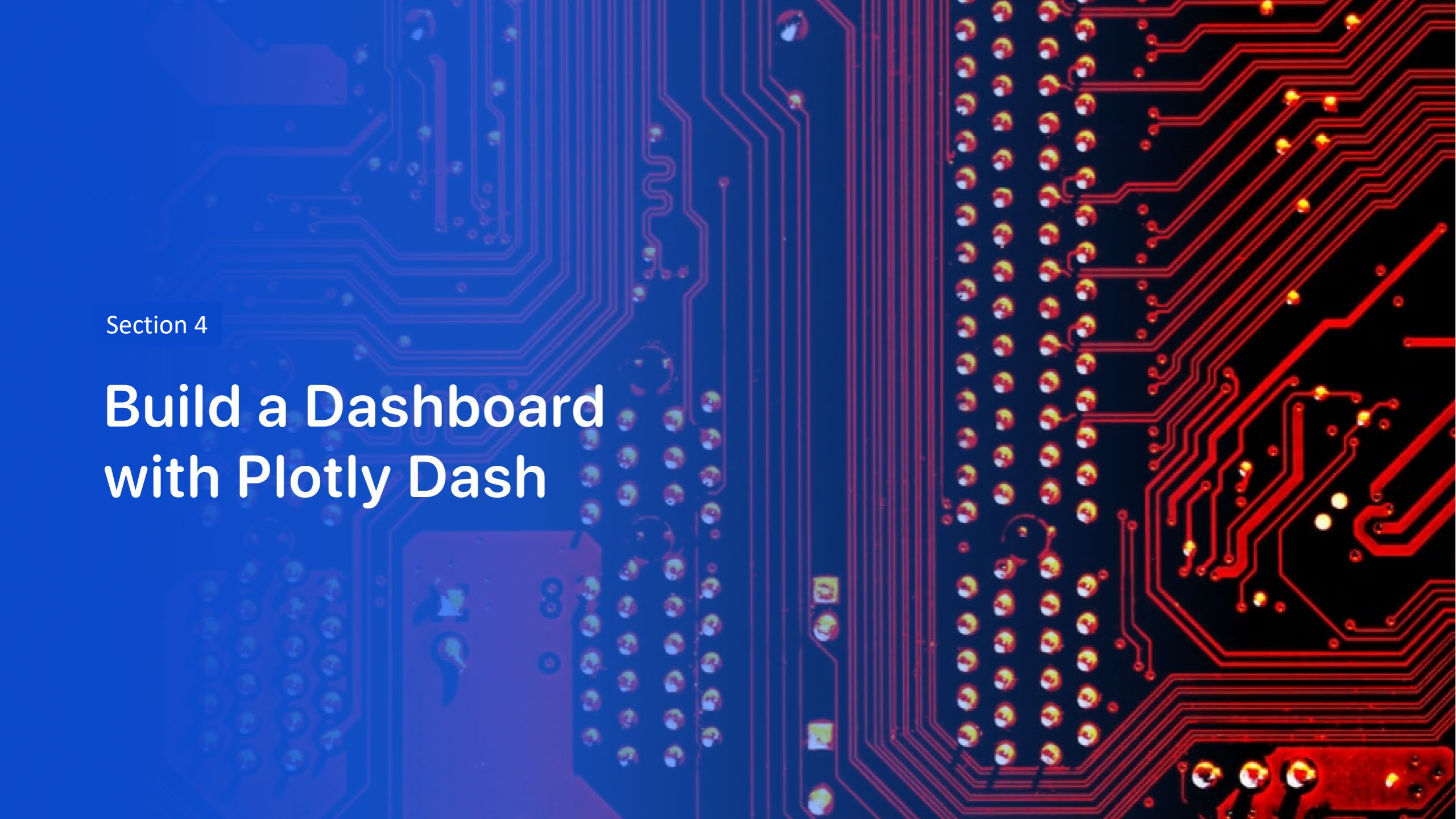
- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb

Site Features

- CCAFS LC-40 launch site locates relatively close to railroad, road, and keeps distance to urban area

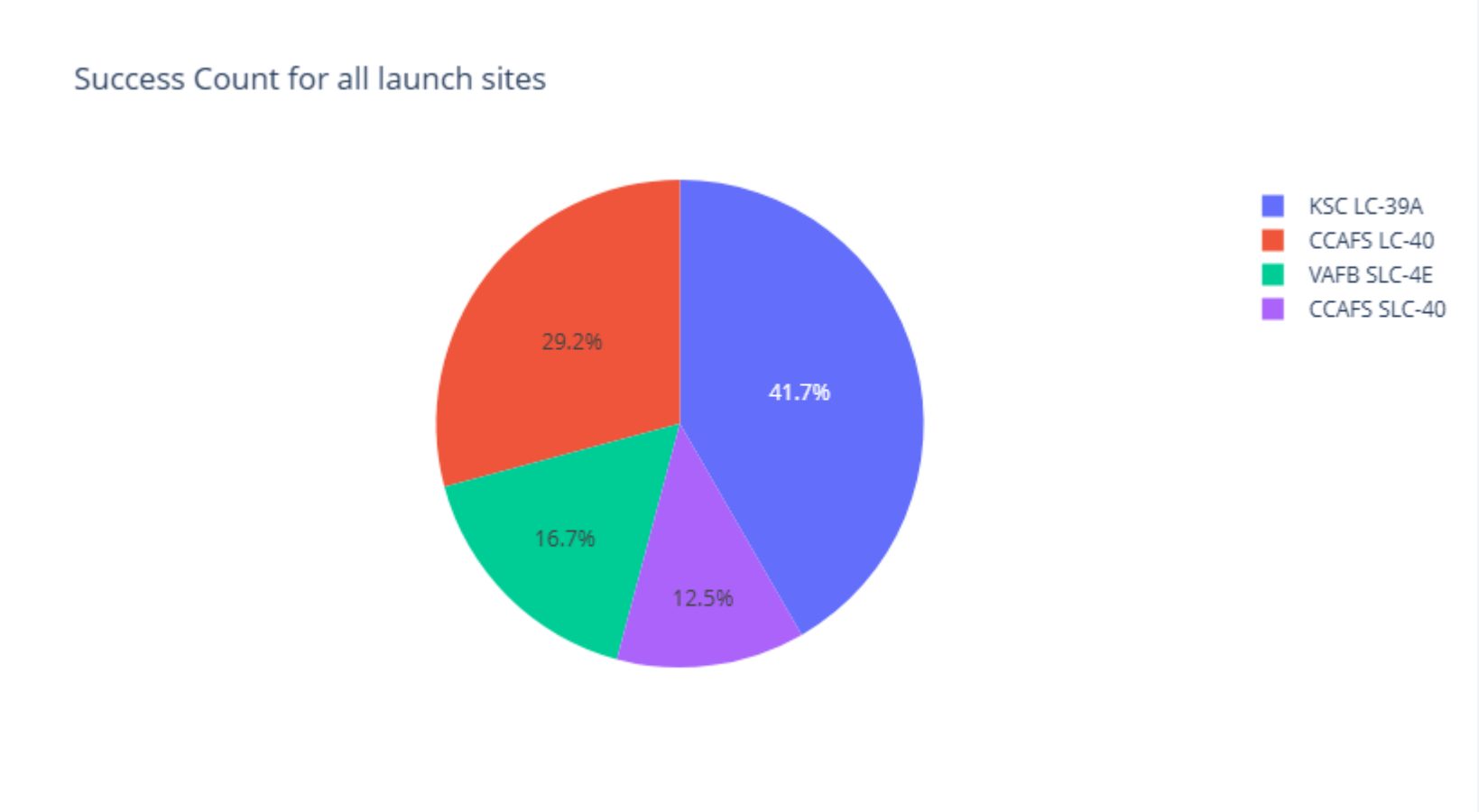


- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb


Section 4
Build a Dashboard with Plotly Dash

Successful Launches Ratio per Site

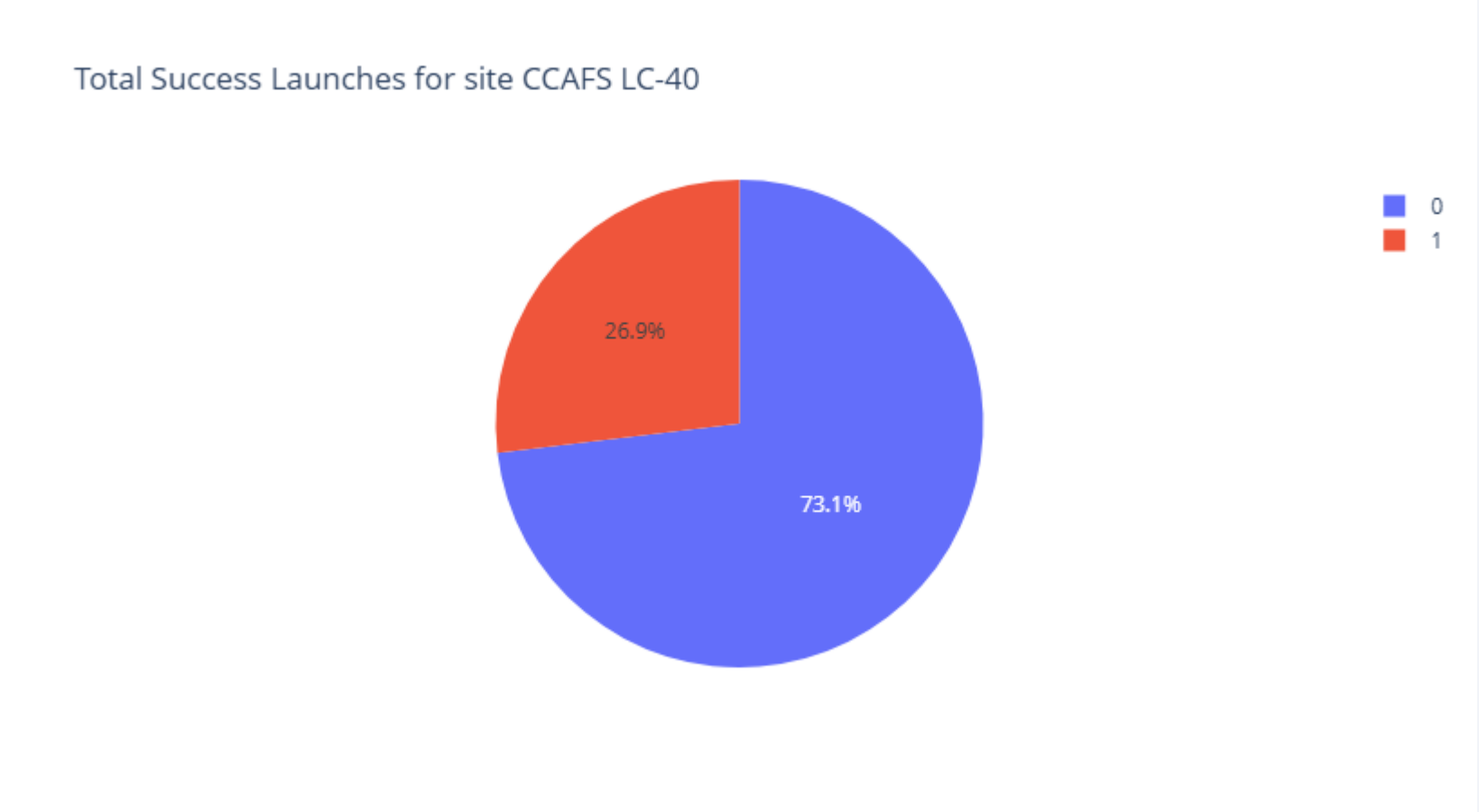
- KSC LC-39A has the highest successful launch rate among all launch sites



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-7-spacex_dash_app.py

Launch Success Rate at CCAFS LC-40

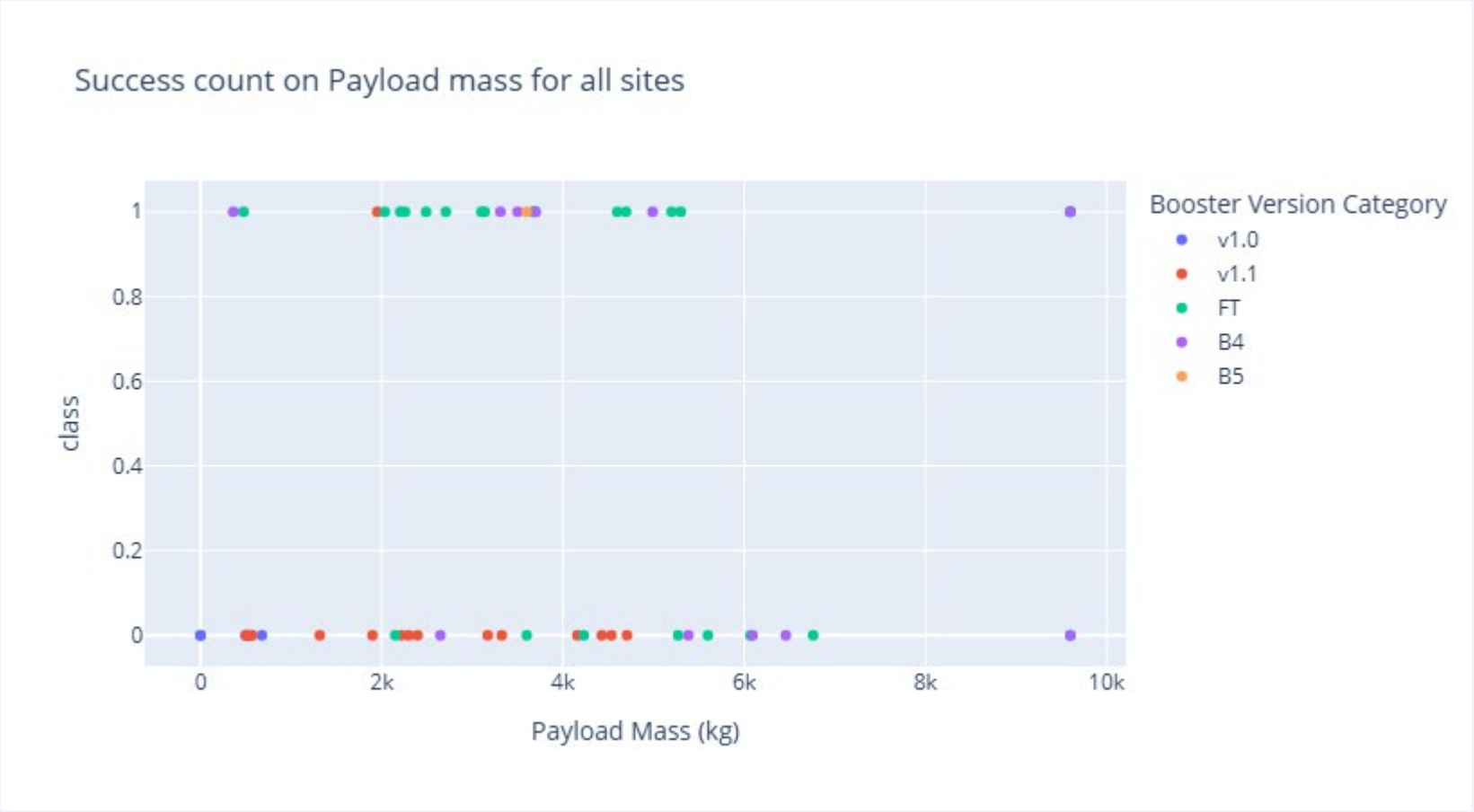
- 73.1% of launches at CCAFS LC-40 are successful



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-7-spacex_dash_app.py

Payload vs Launch Outcome

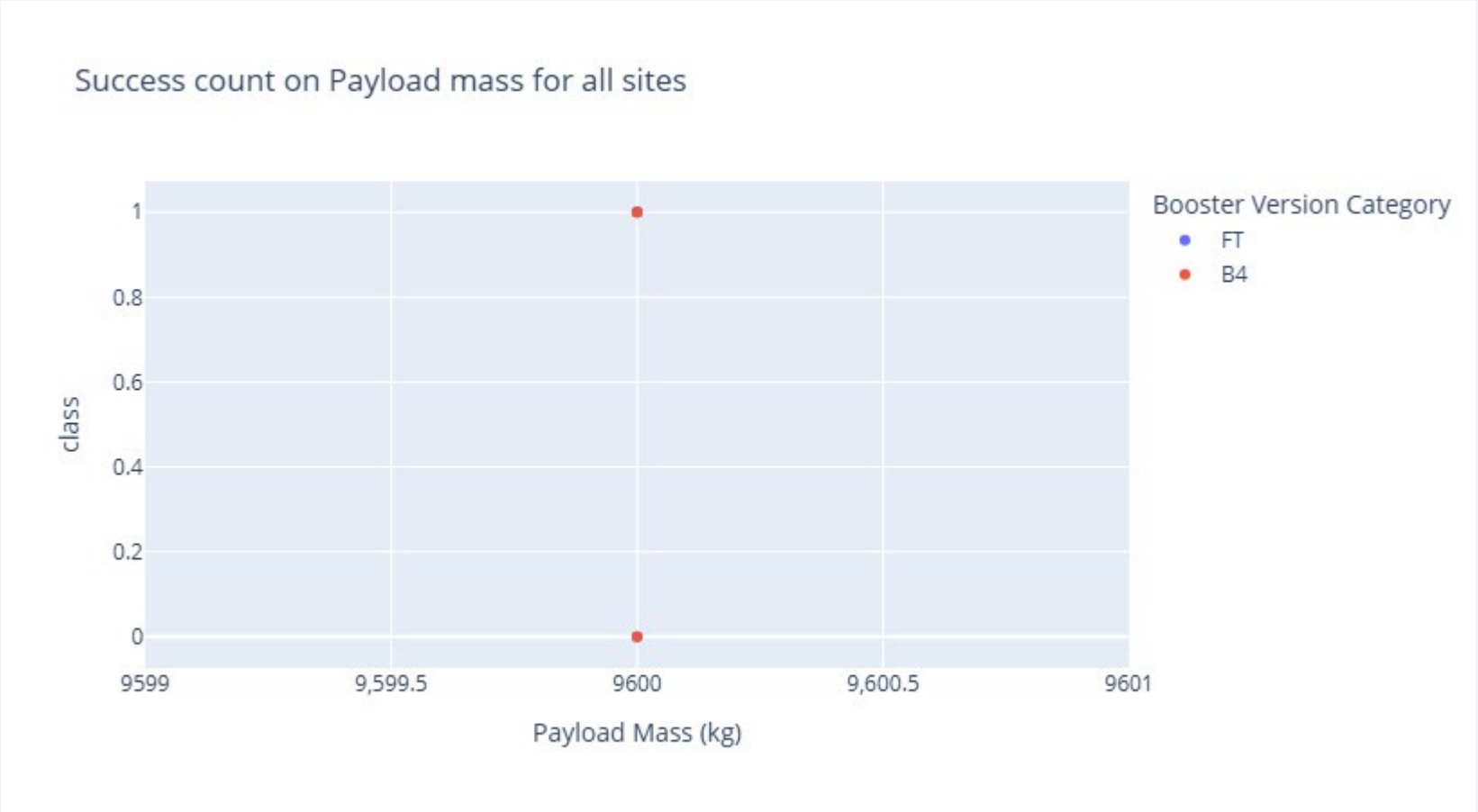
- Payload mass under 6000 kg and FT boosters are the most successful combination



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-7-spacex_dash_app.py

Payload vs Launch Outcome

- Not enough data for estimating the risk of launches over 7000 kg



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-7-spacex_dash_app.py

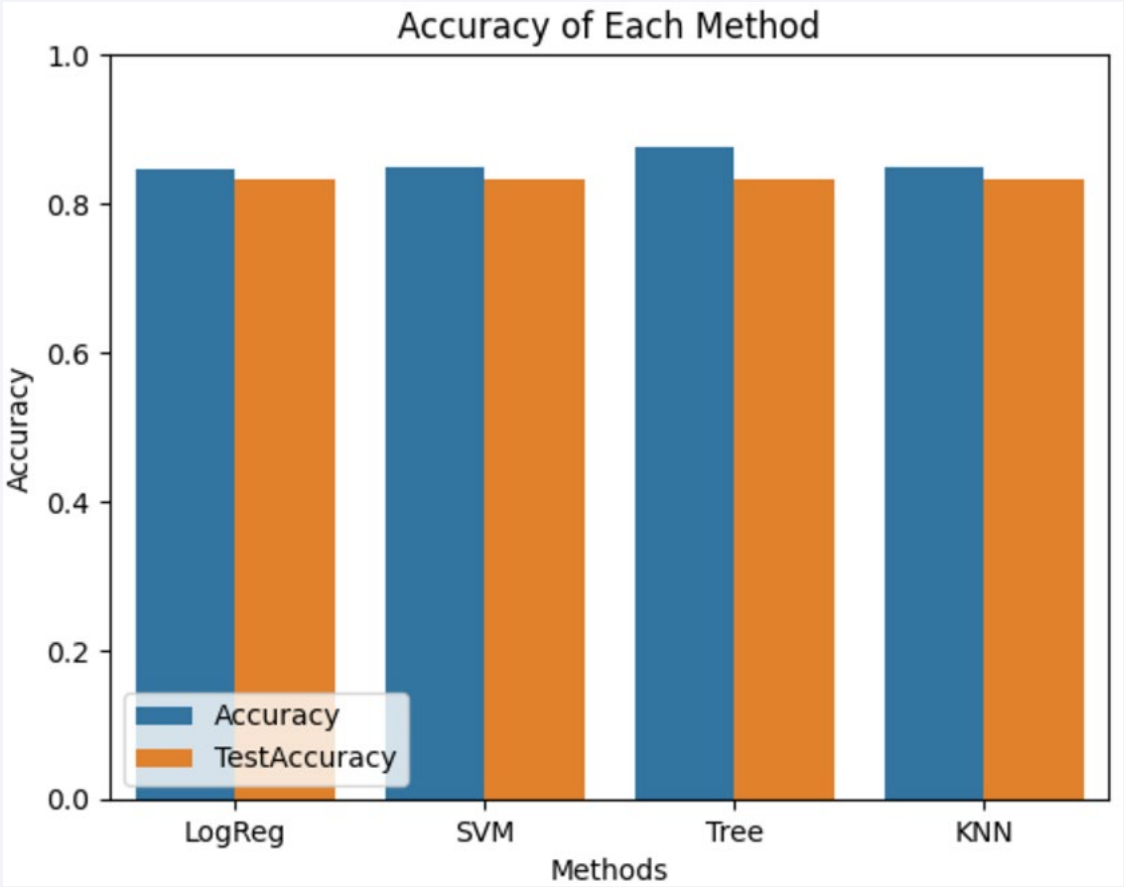
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- LogReg, SVM, Tree, KNN classification models were tested and their accuracies show on right

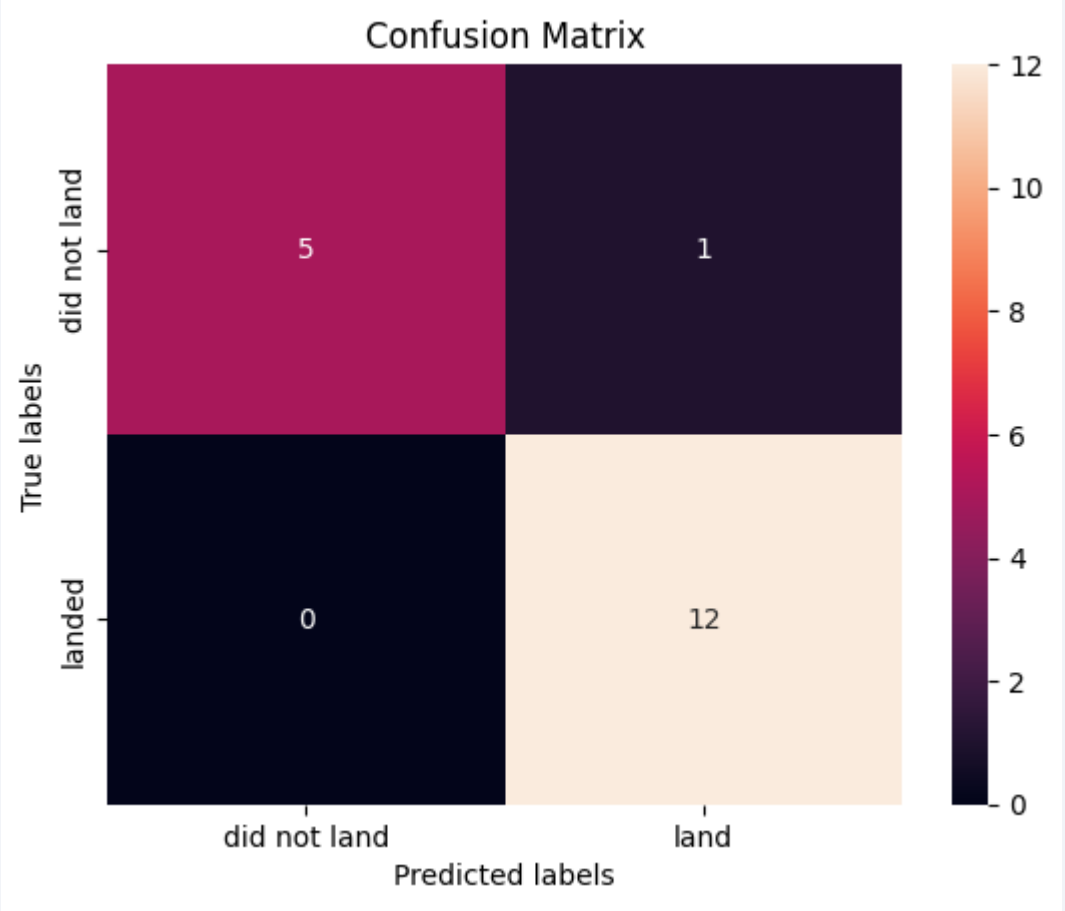
- Decision Tree Classifier has the highest accuracies, 87.5%



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-8-ML.ipynb

Confusion Matrix

- Decision Tree Classifier has the highest accuracies, 87.5%
- Its matrix proves its accuracy by showing the higher true positive and true negative compared to the false ones



- Source code: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-8-ML.ipynb

Conclusions

- Multiple data sources were analyzed, refining conclusions throughout the process.
- The top-performing launch site is KSC LC 39A.Launches with payloads exceeding 7,000 kg are less risky.
- While most mission outcomes are successful, the rate of successful landings has improved over time due to advancements in processes and rocket technology.
- The Decision Tree Classifier proves effective in predicting successful landings, offering potential to boost profits.

Appendix

- Source code:
  - Data Collecting: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-1-Collecting%20the%20data.ipynb
  - Web Scraping: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-2-Web%20scraping.ipynb
  - Data Wrangling: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-3-Data%20wrangling.ipynb
  - EDA Data Visualization: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-4-EDA%20dataviz.ipynb
  - EDA SQL: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-5-EDA%20SQL.ipynb
  - Folium Maps: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-6-Folium%20Maps.ipynb
  - SpaceX DASH app: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-7-spacex_dash_app.py
  - Machine Learning: https://github.com/kaiyakunn/Applied-Data-Science-Capstone-Project/blob/main/Lab-8-ML.ipynb

Thank you!