IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Markus Fabian Lang
21.06.2024

Outline

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

Executive Summary

- Summary of Methodologies
  - **Data Collection**: Utilizing SpaceX API and Web Scraping
  - **Data Analysis**: Performing EDA, including Data Wrangling and Data Visualization
  - **Predictive Modeling**: Applying Machine Learning techniques

- Summary of Results
  - **Data Success**: Collected valuable data from public sources
  - **Feature Selection**: Identified key predictors of launch success
  - **Model Performance**: Determined the best machine learning models

Introduction

- Project Background and Context
  - **Commercial Space Age**: Increasing number of companies making space travel affordable
  - **Industry Leader**: SpaceX, provider of space rockets, space missions, and satellite internet
  - **Advantage**: Reusable first stage of Falcon 9 rocket, leading to significant cost reduction
  - **Cost Comparison**: SpaceX's Falcon 9 launch costs $62 million vs. competitors' $165 million

- Problems You want to Find Answers
  - **Cost Analysis**: Determine the cost of each SpaceX Falcon 9 launch
  - **Success Prediction**: Predict if the first stage of SpaceX Falcon 9 rocket will land successfully
  - **Influence Analysis**: Investigate the effect of several outer factors on the landing success rate

Section 1

Methodology

Methodology

Executive Summary

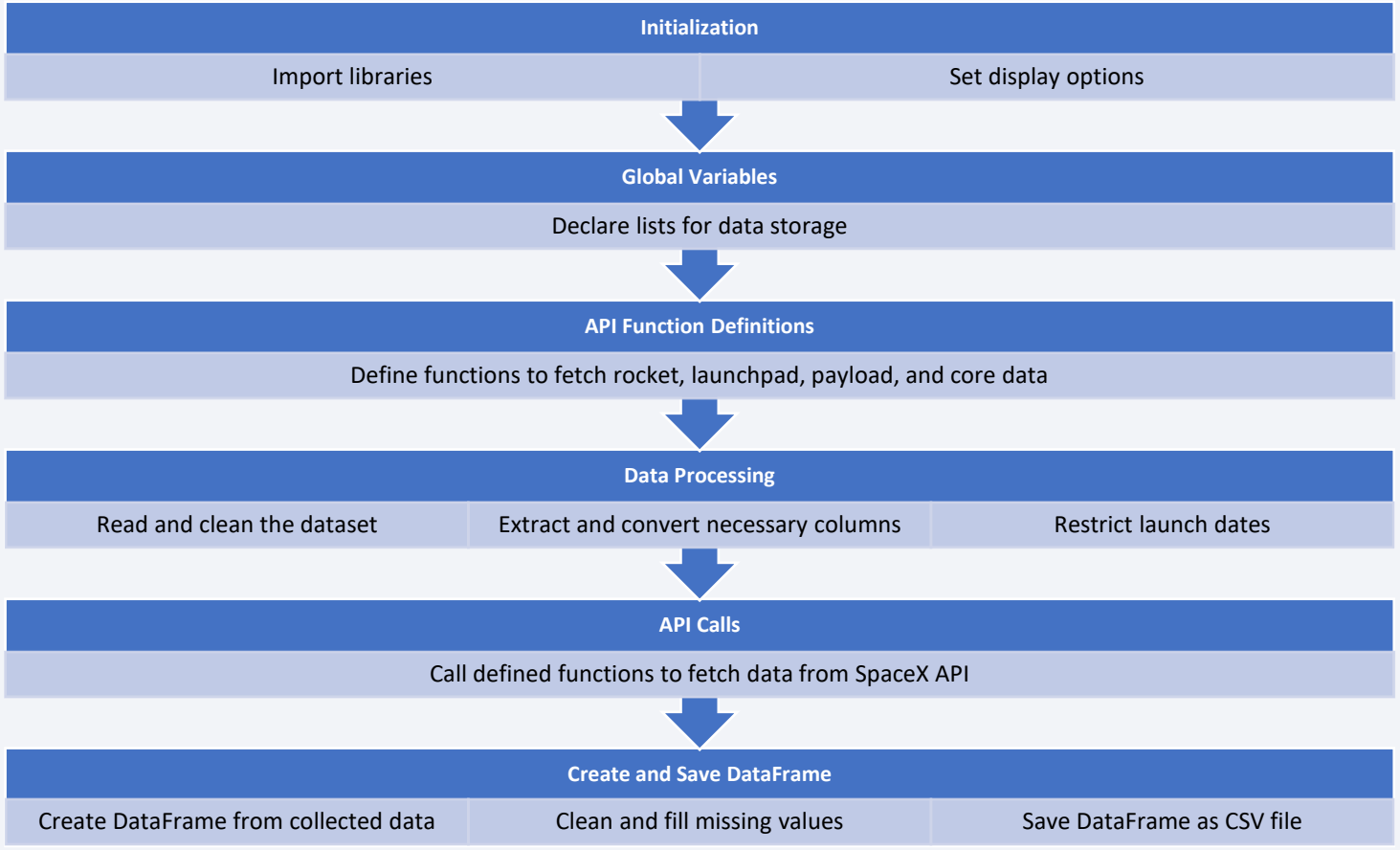
- Data collection methodology:
  - Data was collected from publicly available sources like SpaceX API and Wikipedia.org
- Perform data wrangling
  - Dealt with missing values, removed duplicates, handled outliers, applied feature engineering
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected suitable models, split data into training and test sets, trained and fitted the model, tuned hyperparameters, evaluated model performance with common metrics, applied cross-validation to ensure model stability and avoid overfitting

Data Collection

- Data sets were collected from publicly available sources:
  - Requested Rocket Launch Data from SpaceX API (URL: https://api.spacexdata.com/v4/launches/past)
  - Web Scraped Falcon 9 historical launch records from a Wikipedia page (URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

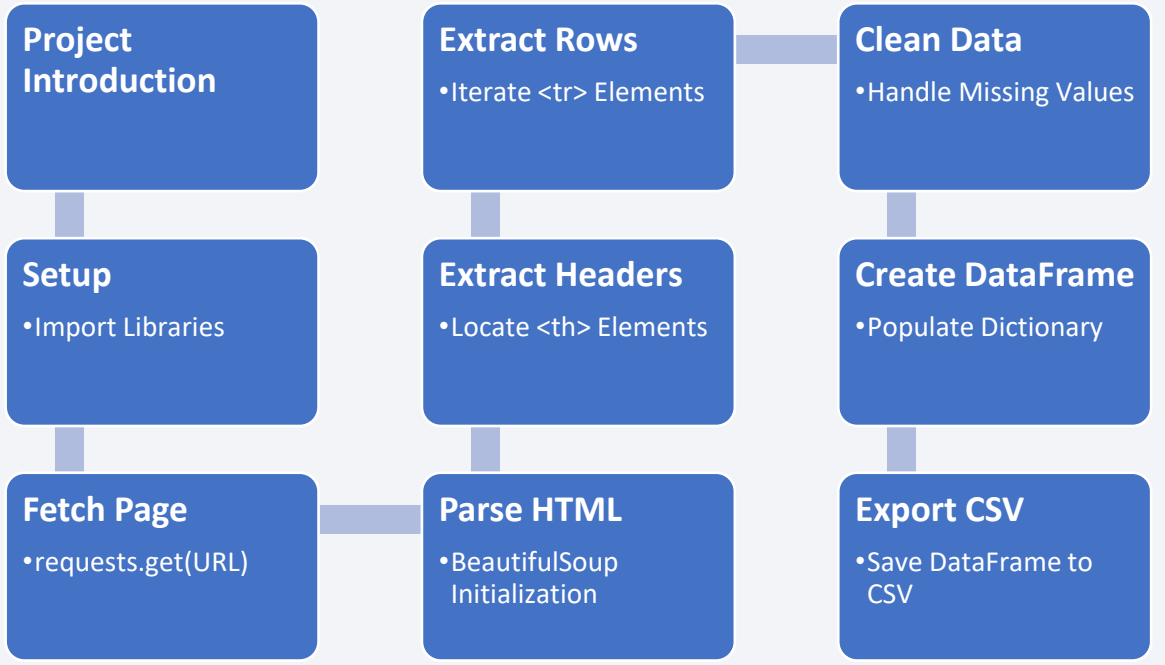
Data Collection – SpaceX API

- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



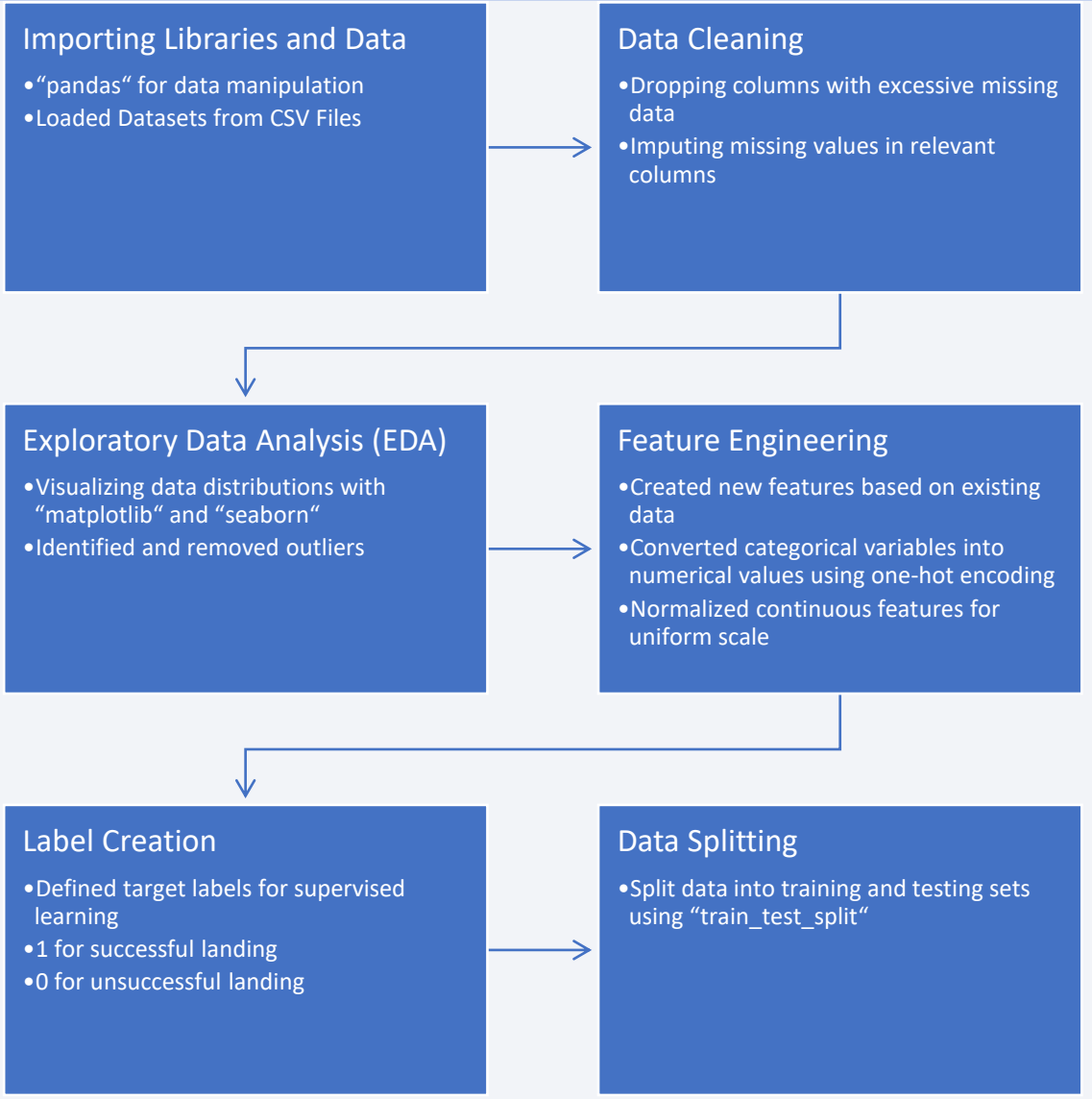
Data Collection - Scraping

- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- FlightNumber vs. PayloadMass (Scatter Plot)
  - Purpose: Analyze the impact of flight number and payload mass on launch success
- FlightNumber vs. LaunchSite (Scatter Plot)
  - Purpose: Explore the influence of launch site on success rate over time
- PayloadMass vs. LaunchSite (Scatter Plot)
  - Purpose: Examine how payload mass affects success rates at different sites
- Orbit vs. Success Rate (Bar Plot)
  - Purpose: Evaluate success rate for different orbits

- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Key SQL Queries:

- **Create Table**: Created a table with non-null dates
- **Distinct Launch Sites**: Retrieved unique launch sites
- **Total Payload for NASA**: Calculated total payload for NASA (CRS)
- **First Successful Landing**: Found date of first ground pad landing success
- **Mission Outcomes**: Counted grouped mission outcomes

- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

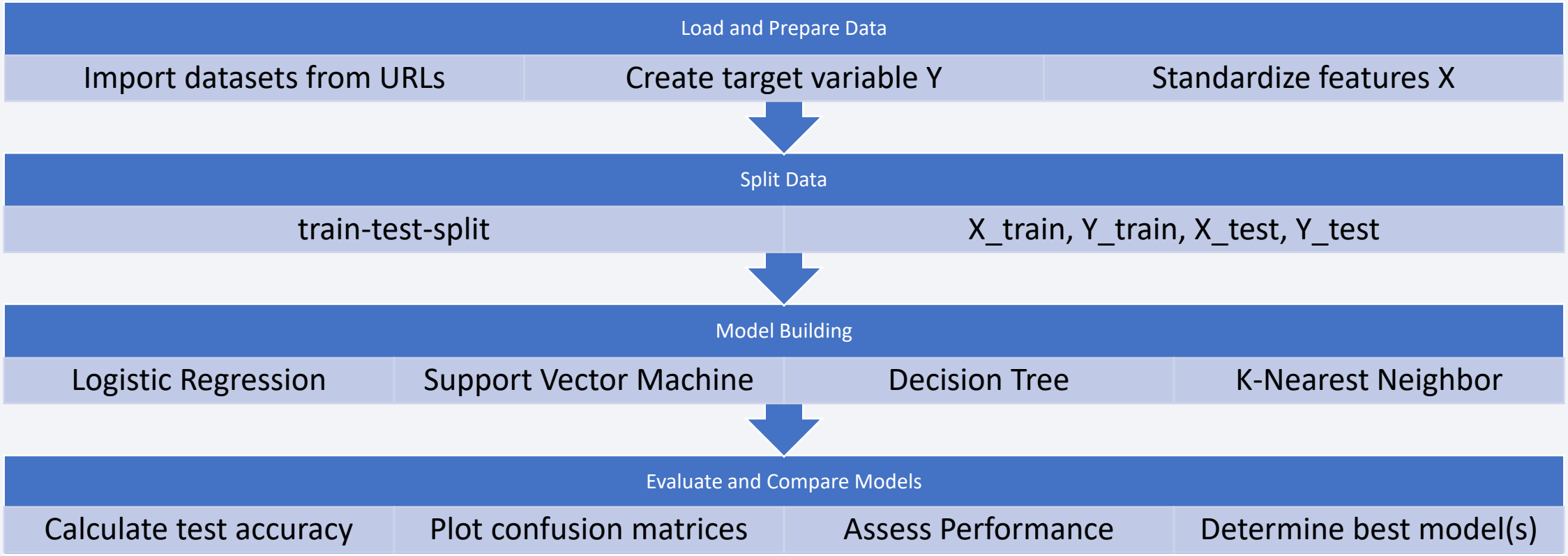
Build an Interactive Map with Folium

- Summary of Map Objects in Folium
  - Markers: Added to indicate the location of launch sites
  - Circles: Used to highlight the areas around the launch sites, indicating regions of interest
  - Lines: Drawn to connect different launch sites, showing the paths or trajectories
- Purpose of Adding Objects:
  - Markers: Help in pinpointing the exact locations of various launch sites on the map
  - Circles: Provide a visual representation of the areas surrounding the launch sites, potentially indicating safety zones or areas of influence
  - Lines: Illustrate connections or routes between different launch sites, aiding in understanding spatial relationships and logistics
- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Interactive Dashboards allow users to explore SpaceX launch data effectively
- Dropdown Menu:
  - Allows users to select a specific launch site or view data for all sites
- Pie Chart:
  - Displays the total success launches by site, updating based on the selected site from dropdown menu
- Range Slider:
  - Enables users to filter data by adjusting payload mass range
- Scatter Plot:
  - Shows the correlation between payload mass and launch success, updating based on selected site from dropdown menu and payload mass from range slider
- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

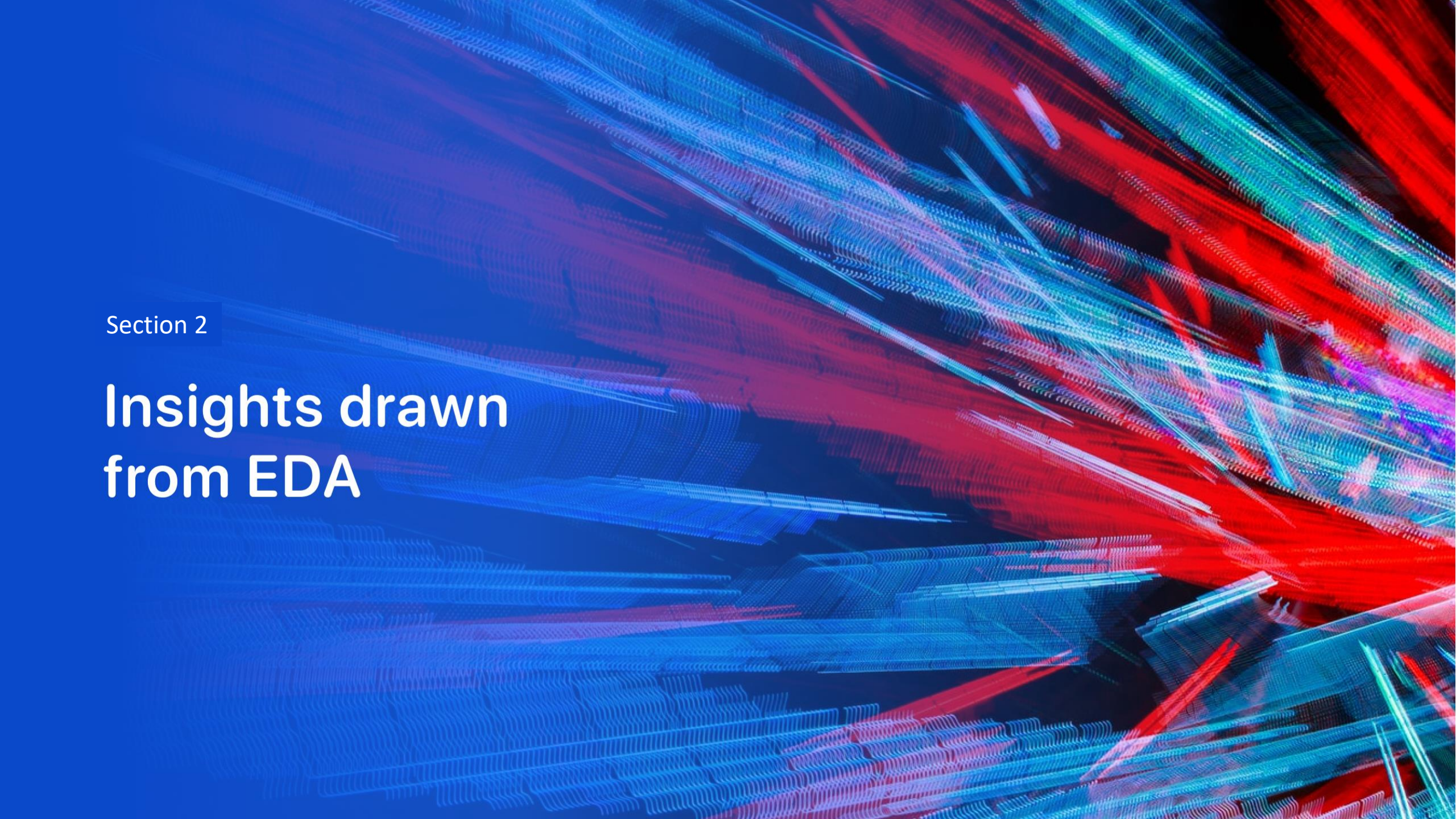
Predictive Analysis (Classification)



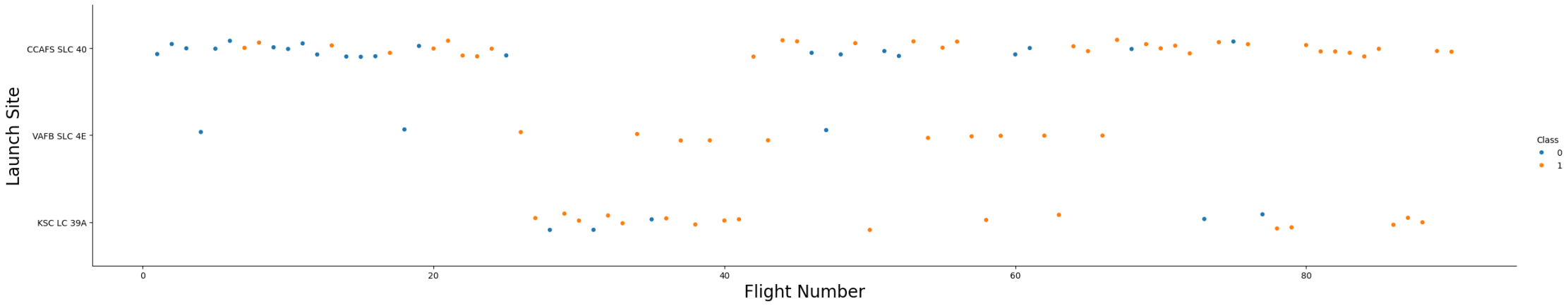
- GitHub URL: https://github.com/markuslangus/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Explanatory Data Analysis reveals an increasing trend in the success rate of SpaceX missions over time, indicating technological advancements
- Interactive Analytics identify launch sites that are near safety zones, seas, and have strong logistical infrastructure
- Various predictive analysis methods perform similarly on test accuracy, suggesting that the dataset might be too small to determine the best model

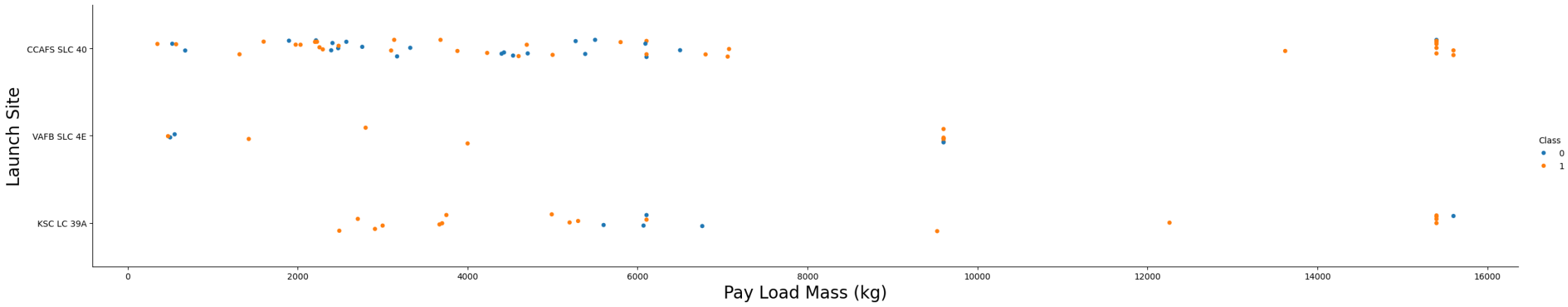

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



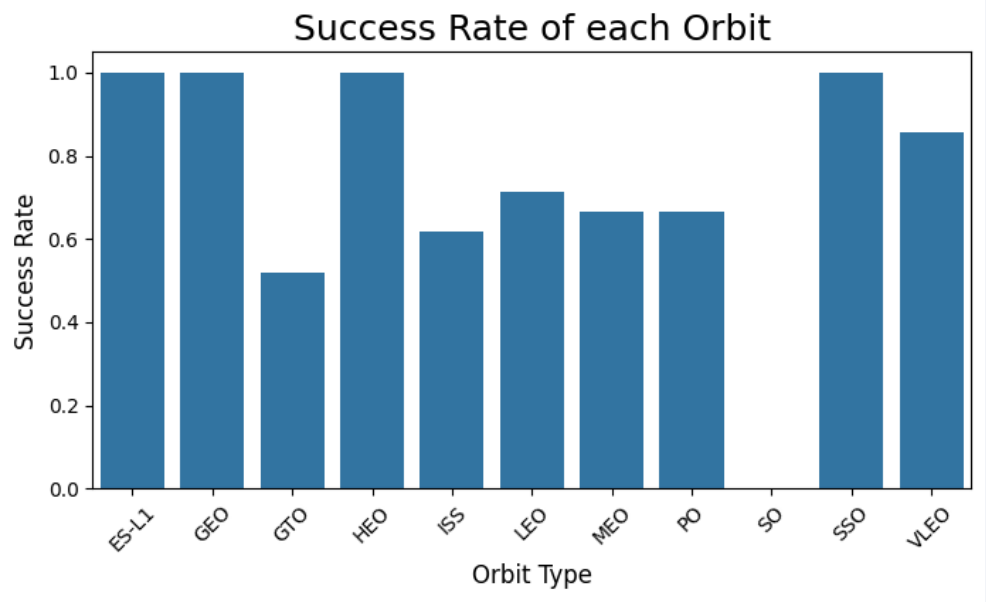
- No clear trend or pattern indicating that a specific launch site consistently results in more successful landings of the first stage
- Each launch site has both successful and unsuccessful landings throughout the range of flight numbers
- All launch sites show similar behavior concerning success and failure rates across flight numbers, with no significant outliers or anomalies in the data

Payload vs. Launch Site



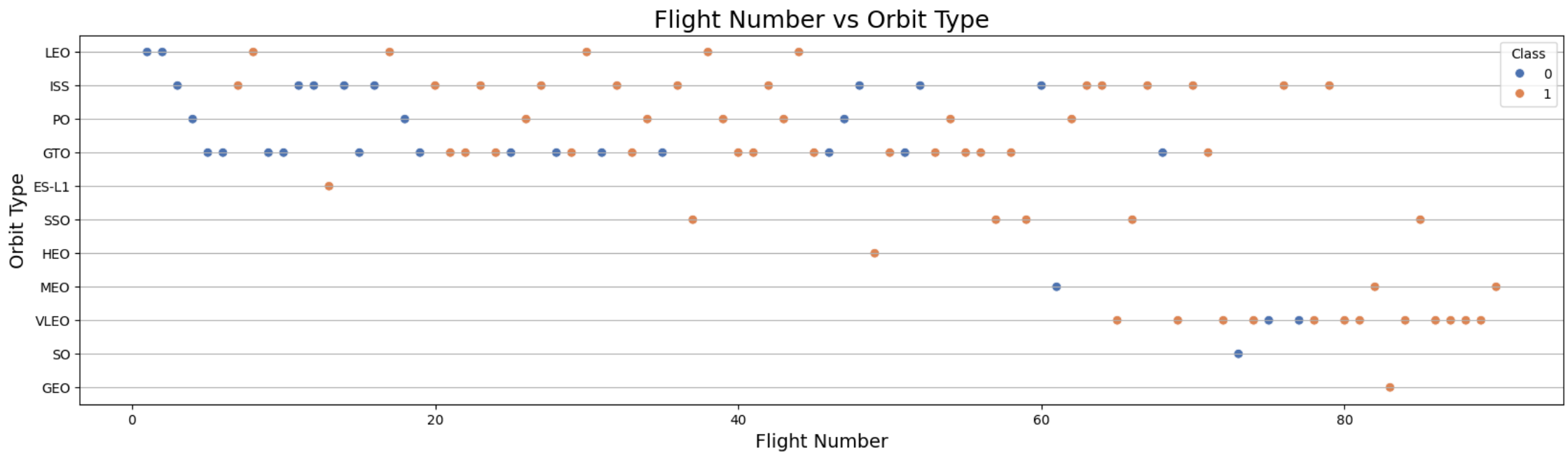
- Visible trend of higher success rates at higher payload masses
- Lower payload masses show more variability in success, with a mix of both successful and unsuccessful landings of the first stage
- Higher payload masses have fewer unique values, suggesting that heavier payloads tend to be more standardized

Success Rate vs. Orbit Type



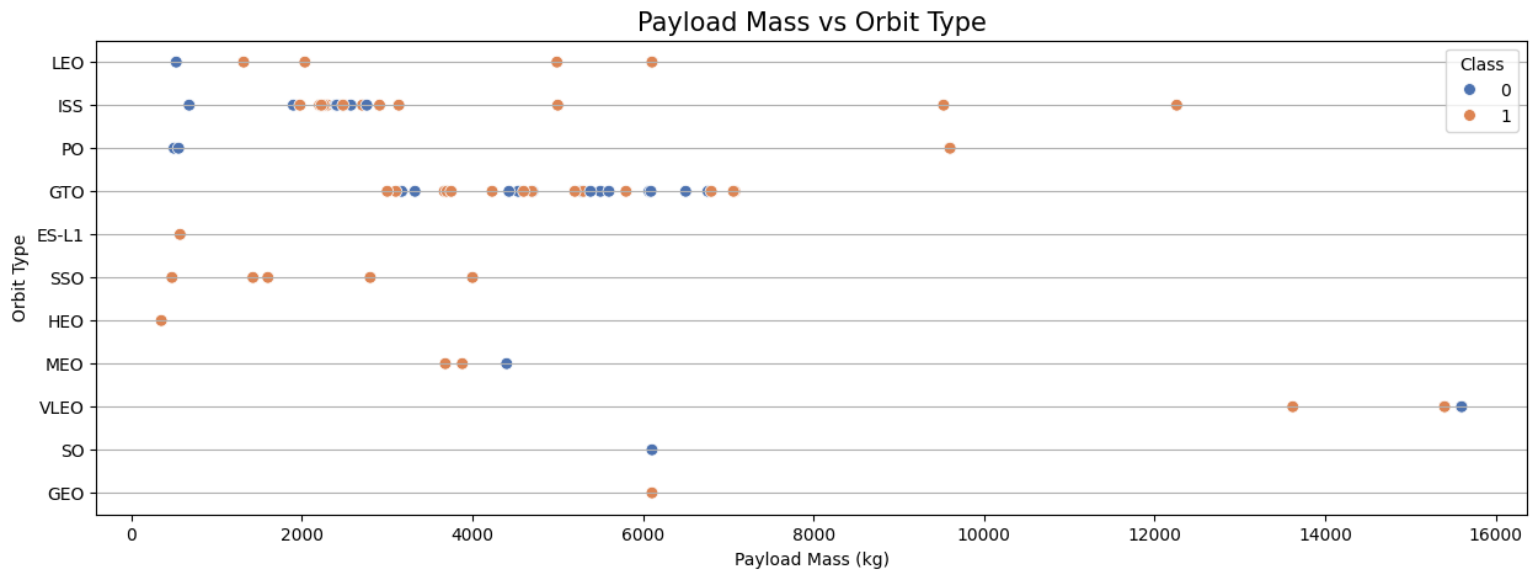
- Perfect success rate among Orbits Earth-Sun Lagrange Point 1, Geostationary Earth Orbit, Highly Elliptical Orbit, and Sun-Synchronous Orbit
- No success in Suborbital, indicating technical challenges or limited attempts
- Very Low Earth Orbit with good landing success rate, all the remaining orbit types with moderate success rates

Flight Number vs. Orbit Type



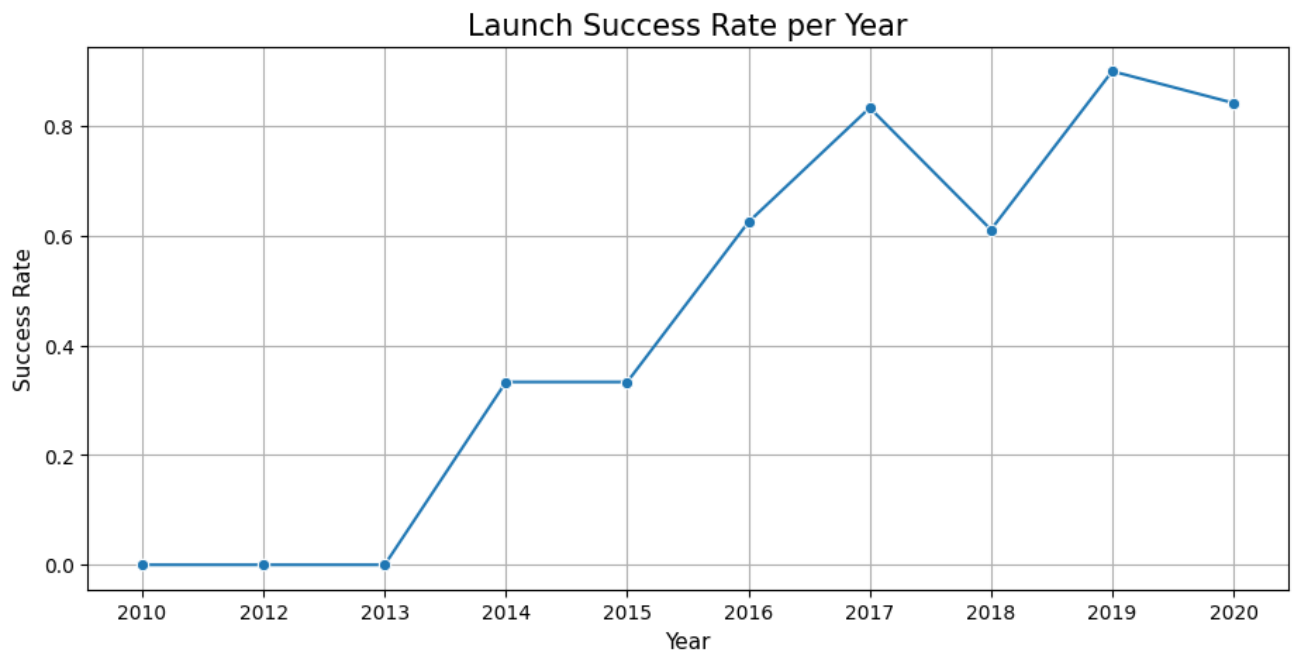
- Most recent flights are concentrated in Very Low Earth Orbit and partly in the International Space Station Orbit, indicating a current focus on satellite and space station missions
- Neglect of Low Earth Orbit and Polar Orbit after high usage in early flight numbers suggest these orbits were primarily used for testing purposes due to their proximity to Earth
- Orbit types with fewer flights, such as Highly Elliptical Orbit, Suborbital, and Geostationary Orbit, likely indicate greater challenges associated with these orbits, leading to less frequent missions

Payload vs. Orbit Type



- Wide range of payload masses across different orbit types, with most payloads concentrated below 6,000 kg, indicating a preference for lighter payloads
- Higher payload masses are primarily associated with Very Low Earth Orbit, Polar Orbit, and partly ISS Orbit, suggesting these orbits are suited for missions requiring larger payloads
- First stage's landing success rate is higher for orbits with lower payload masses and ISS missions, indicating better reliability or experience in these scenarios

Launch Success Yearly Trend



- Initial period from 2010 to 2013 saw no successful landings, highlighting the complexity and challenges involved in developing self-landing rocket technology
- 2014 to 2017 showing marked improvement in landing outcomes, reflecting significant advancements in this innovative technology
- Success rate remained high from 2017 onwards, despite a notable dip in 2018, suggesting a need for further investigation to understand the factors behind this significant drop

All Launch Site Names

```
In [9]: %%sql
        SELECT DISTINCT "Launch_Site"
        FROM SPACEXTABLE;
         * sqlite:///my_data1.db
        Done.
```

Out[9]:

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

- List of all unique launch sites used by SpaceX for their missions, provided by an SQL query
- SQL query selects distinct values from the “Launch_Site” column in the “SPACEXTBL” table, ensuring that each unique launch site is listed only once
- This Table provides insights into the different launch sites, which is useful for understanding the geographical distribution of their launch operations

Launch Site Names Begin with 'CCA'

In [10]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

Out[10]:

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

- This query selects all columns from the “SPACEXTABLE” where the “Launch_Site” column values start with the string “CCA”
- A subset of records is provided by using the “LIKE” and “LIMIT” clauses, which search for the string “CCA” in the “Launch_Site” column and limit the output to the first 5 records that have been filtered

Total Payload Mass

```
In [11]: %%sql
         SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass
         FROM SPACEXTABLE
         WHERE Customer LIKE "%NASA (CRS)%";
          * sqlite:///my_data1.db
         Done.
Out[11]: Total_Payload_Mass
                    48213
```

- This query sums up all the values of the “PAYLOAD_MASS__KG_” column that have been filtered from the “SPACEXTABLE” table
- Filtering happens by searching all the values of the “Customer” column that contain the substring “%NASA (CRS)%”
- The total payload mass is displayed in the output

Average Payload Mass by F9 v1.1

```
In [12]: %%sql
         SELECT AVG("PAYLOAD_MASS__KG_") AS "Average_Payload_Mass"
         FROM SPACEXTABLE
         WHERE "Booster_Version" = "F9 v1.1";
          * sqlite:///my_data1.db
         Done.
Out[12]: Average_Payload_Mass
                      2928.4
```

- The average of the payload mass of all filtered rows is calculated
- Filtering happens by checking whether the row corresponds to booster version “F9 v1.1”
- The average payload mass is displayed in the output cell

First Successful Ground Landing Date

```
In [13]: %%sql
         SELECT MIN(Date) AS First_Succesful_Landing
         FROM SPACEXTABLE
         WHERE "Landing_Outcome" = "Success (ground pad)";
          * sqlite:///my_data1.db
         Done.
Out[13]: First_Succesful_Landing
                      2015-12-22
```

- The minimum value of the “Date” column is selected
- Table rows are filtered by checking if the landing outcome is equal to “Success (ground pad)”
- Output showing the earliest date where a landing outcome was successful
- Therefore every landing outcome before the 22nd December 2015 was not successful

Successful Drone Ship Landing with Payload between 4000 and 6000

In [14]:
```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = "Success (drone ship)"
AND "PAYLOAD_MASS__KG_" > 4000
AND "PAYLOAD_MASS__KG_" < 6000;
```

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

Out[14]:

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

- Query shows the value of every booster version from the filtered table
- Filtering occurs by checking whether the landing outcome was a successful drone ship landing and additionally whether the payload mass was between 4000 and 6000 kg
- Output shows the distinct names of the booster versions that satisfy these conditions

Total Number of Successful and Failure Mission Outcomes

In [15]:
```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS Total_Count
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

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

Out[15]:

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

- Query groups the rows of the table by their mission outcome
- For each unique mission outcome, the value of the "Mission_Outcome“ column is selected and the frequency of rows with that specific value is counted
- Output displays the total frequency of each unique mission outcome in the table

Boosters Carried Maximum Payload

In [16]:
```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

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

Out[16]:

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

- Query selects each value of the “Booster_Version” column from SPACEXTABLE where the "PAYLOAD_MASS__KG_“ equals the value obtained from the subquery
- Subquery selects the maximum payload mass value from SPACEXTABLE
- Output table contains all booster versions that carried the maximum payload mass

2015 Launch Records

In [17]:
```
%%sql
SELECT
    CASE
        WHEN substr(Date, 6, 2) = '01' THEN 'January'
        WHEN substr(Date, 6, 2) = '02' THEN 'February'
        WHEN substr(Date, 6, 2) = '03' THEN 'March'
        WHEN substr(Date, 6, 2) = '04' THEN 'April'
        WHEN substr(Date, 6, 2) = '05' THEN 'May'
        WHEN substr(Date, 6, 2) = '06' THEN 'June'
        WHEN substr(Date, 6, 2) = '07' THEN 'July'
        WHEN substr(Date, 6, 2) = '08' THEN 'August'
        WHEN substr(Date, 6, 2) = '09' THEN 'September'
        WHEN substr(Date, 6, 2) = '10' THEN 'October'
        WHEN substr(Date, 6, 2) = '11' THEN 'November'
        WHEN substr(Date, 6, 2) = '12' THEN 'December'
    END AS Month_Name,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
    AND substr(Date, 0, 5) = '2015';
```

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

Out[17]:

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Query selects the month when the mission occured, the landing outcome, the booster version, and the launch site of all missions that meet both specified conditions
- Conditions check whether the landing outcome was a failure on a drone ship and whether the mission was conducted in the year 2015
- Output table contains the month when the landing outcome was a failure on a drone ship, the booster version used, and the launch site from which the rocket was launched

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

In [18]:
```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS "Count"
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC;
```

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

Out[18]:

Landing_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

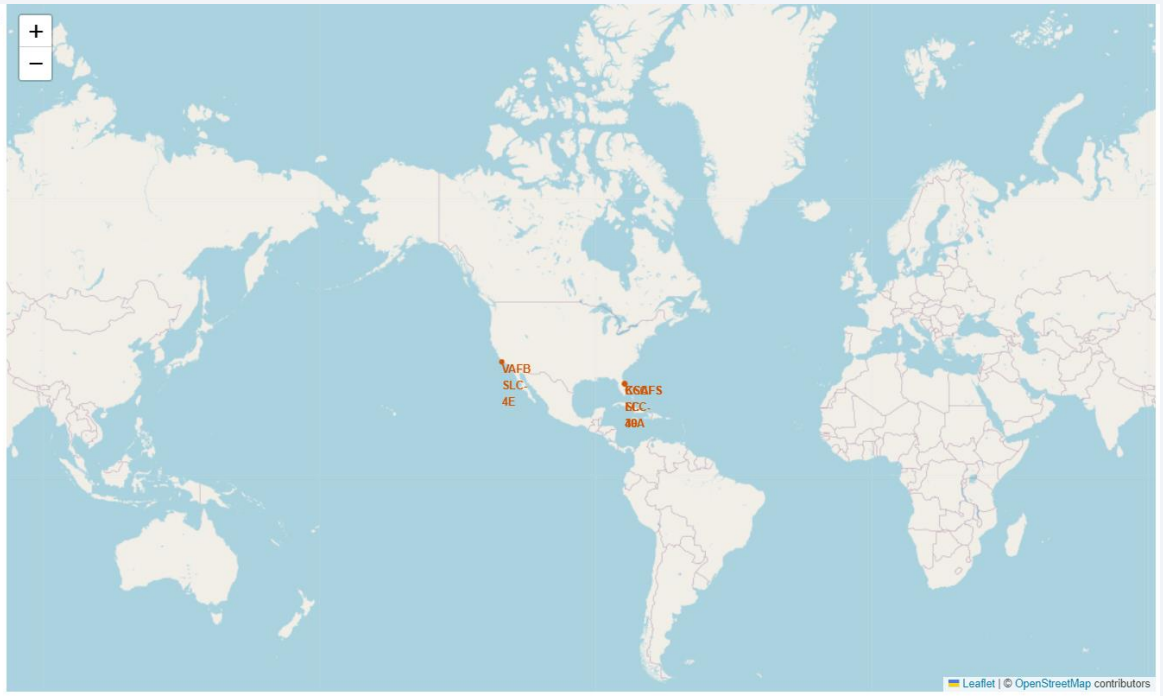
- Query groups the rows by their landing outcome, selects the names of each unique landing outcome, and counts the number of rows for each landing outcome that meets the specified condition
- Condition checks whether the date of the mission launch was between June 4, 2010, and March 20, 2017
- Results are ordered by their count in descending order, with the most frequent landing outcome occurrence listed first

Section 3

Launch Sites Proximities Analysis

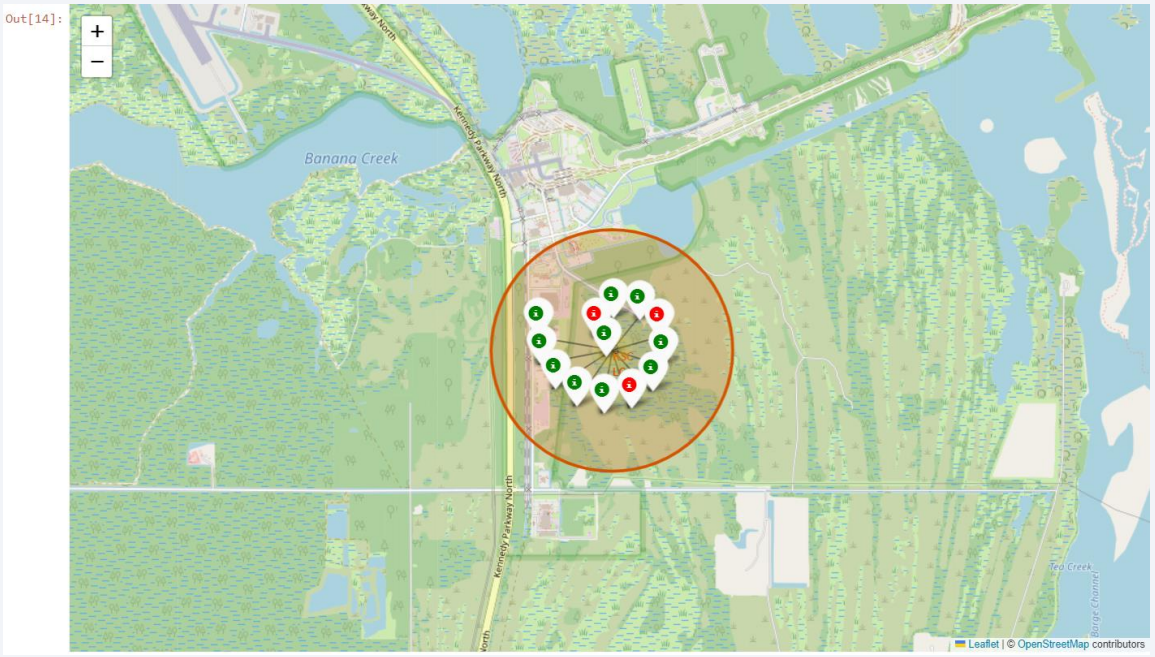
Mark all Launch Sites on a Global Map

Out[10]:



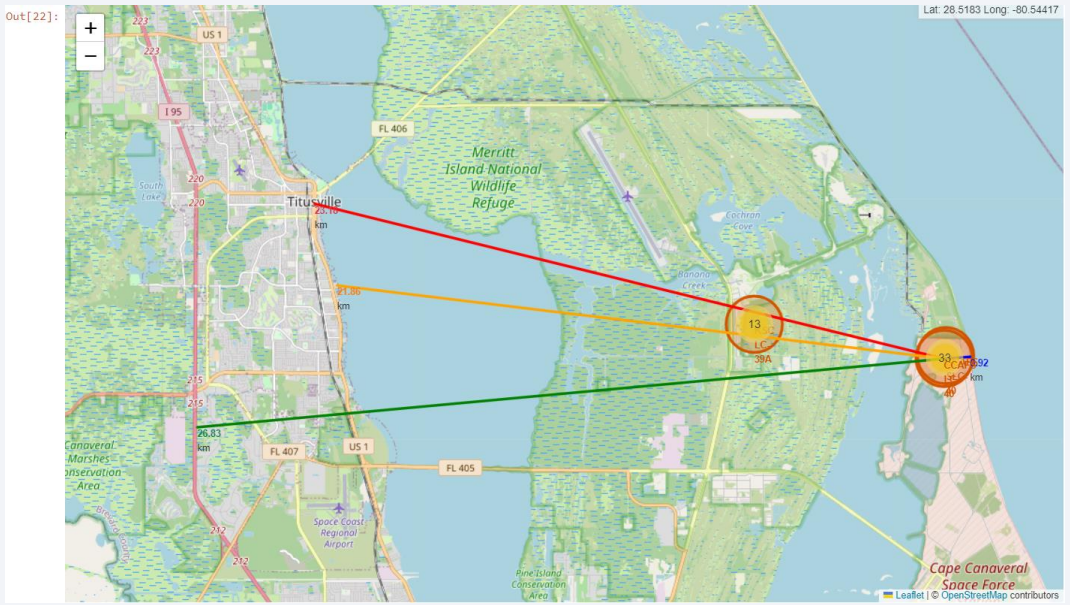
- Launch sites include Vandenberg Air Force Base, Kennedy Space Center, and Cape Canaveral Air Force Station
- All launch sites are located close to the east or west coast, minimizing risks to populated areas in the event of launch failures
- All launch sites are relatively near the equator, offering advantages for reaching certain types of orbits

Mark the Success/Failed Launches for each Launch Site

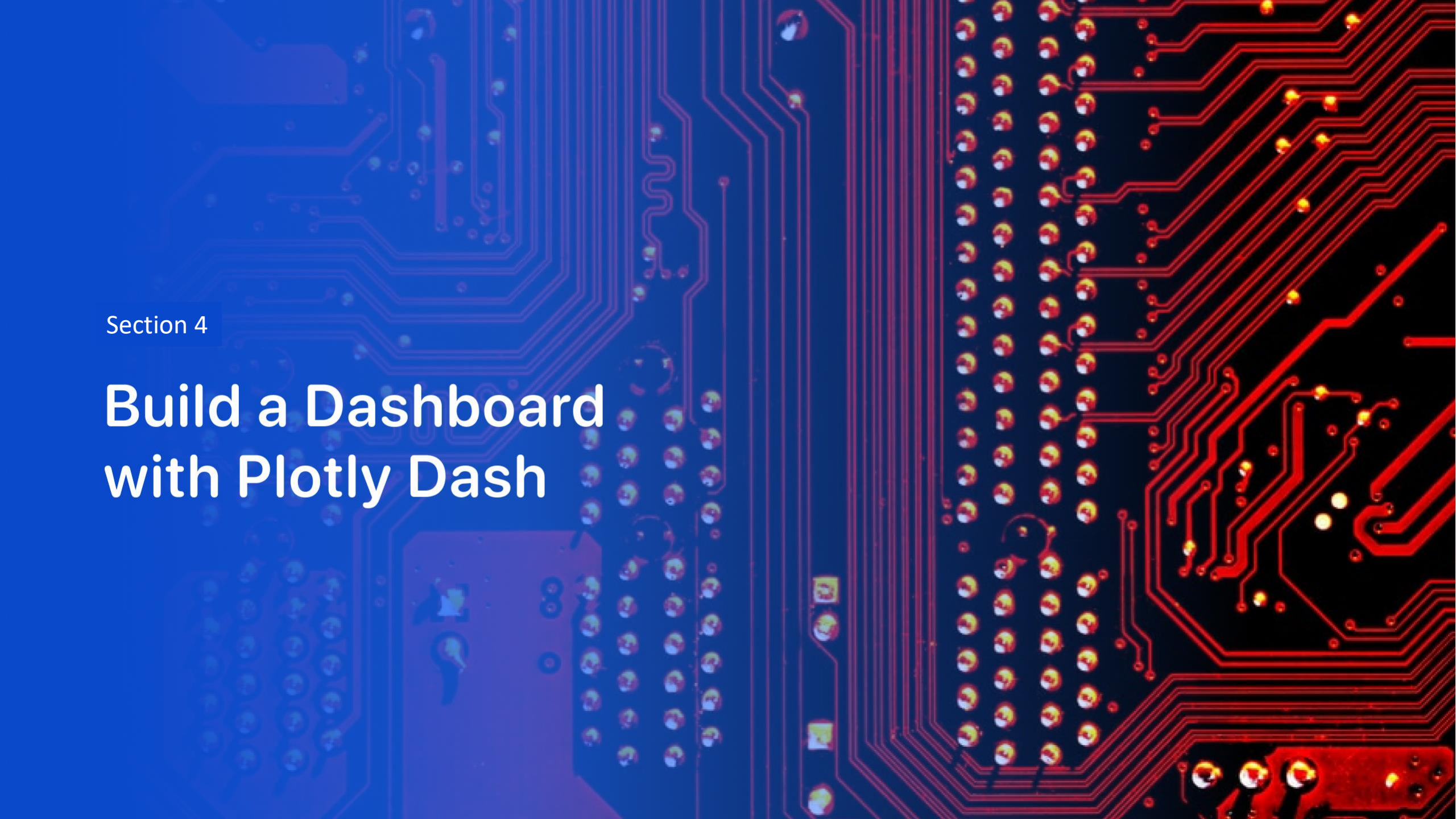


- Markers indicate the landing outcomes of the first stage of SpaceX rockets
  - Green markers represent successful landings
  - Red markers represent failed landings
- Markers are placed on the launch sites from where the rockets were launched; the landings occurred elsewhere

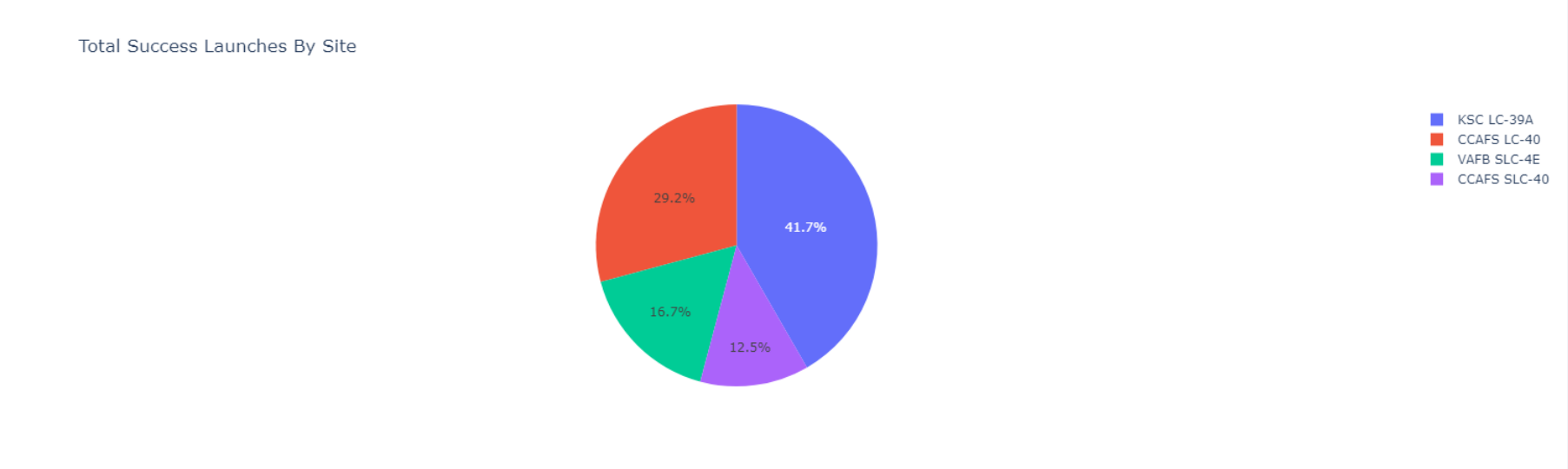
Distances between a Launch Site to key Proximities



- Distances from a selected rocket launch site to key proximities are represented by different colored lines:
  - Red line: City center, 23.16 km
  - Yellow line: Coastline, 21.36 km
  - Green line: Highway, 26.83 km
- The launch site offers good logistical infrastructure and a lower risk of impacting populated areas

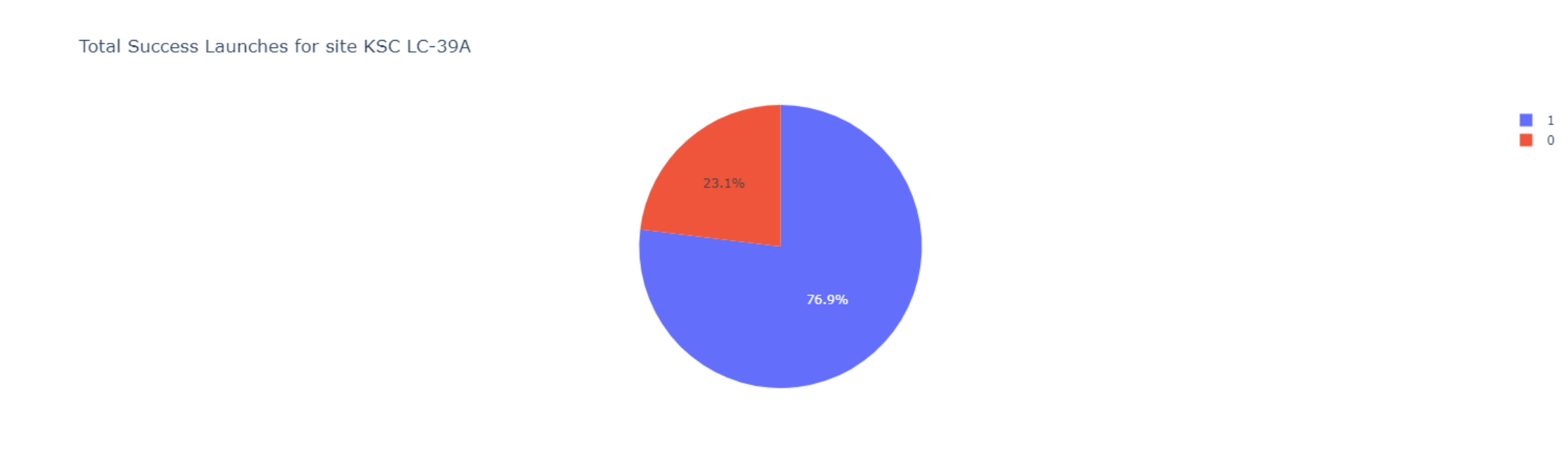

Section 4
Build a Dashboard with Plotly Dash

Distribution of Successful Space Launches by Site



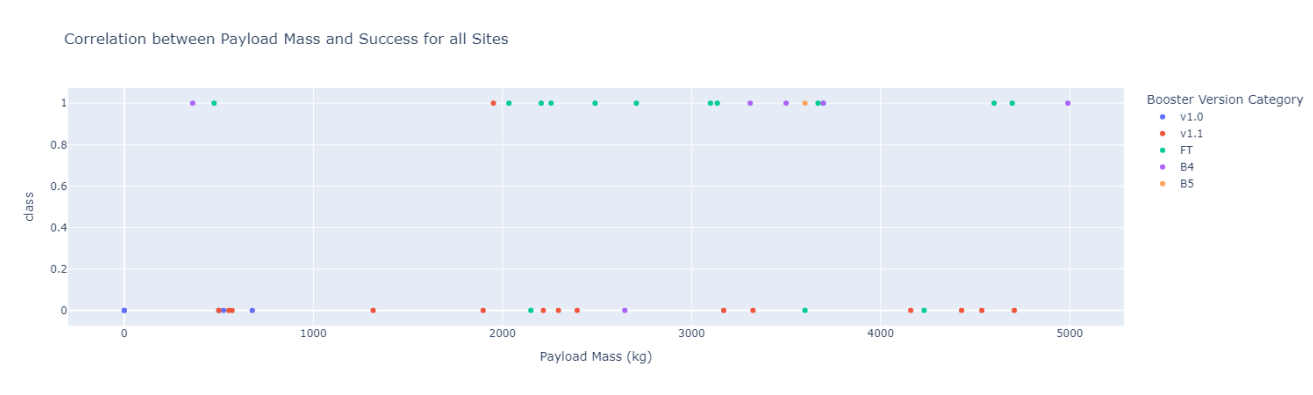
- KSC LC-39A has the highest proportion of successful launches, making up 41.7% of the total
- CCAFS LC-40 follows with 29.2% of successful launches, while VAFB SLC-4E accounts for 16.7%
- CCAFS SLC-40 has the smallest share of successful launches at 12.5%

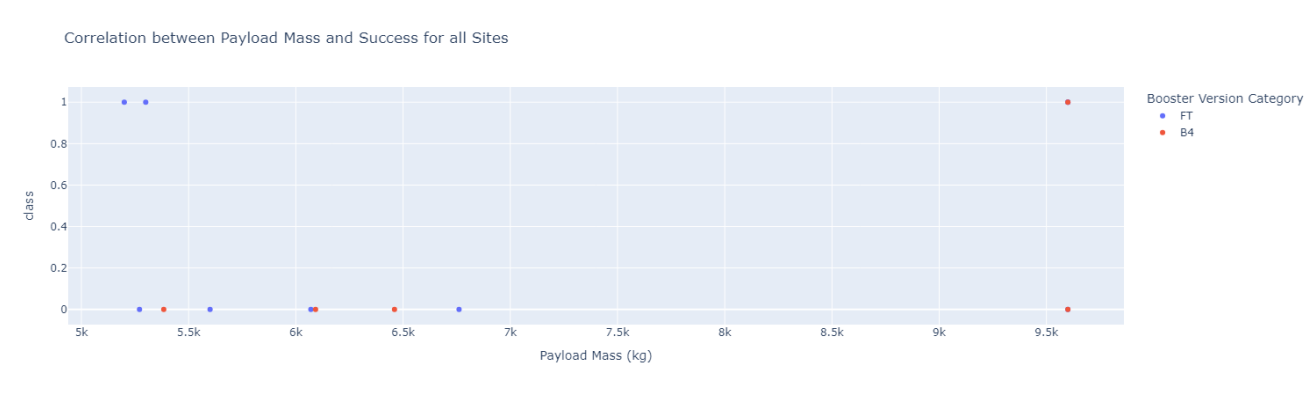
Success Rate of Space Launches at KSC LC-39A



- KSC LC-39A has a high success rate of 76.9% for its launches
- 23.1% of the launches from KSC LC-39A were unsuccessful

Payload Mass and Launch Success across Booster Versions



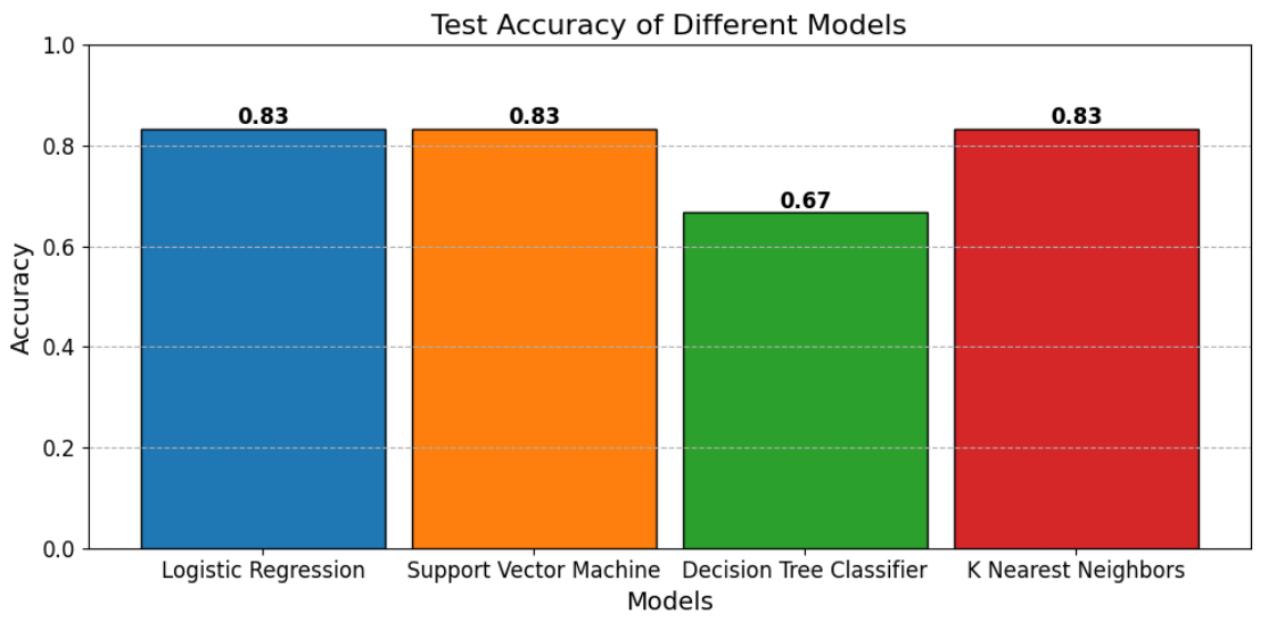


- Successful and unsuccessful launches are spread across different payload masses and booster versions
- Higher payload masses (5000 to 10,000 kg) show a higher ratio of unsuccessful launches
- Lower payload masses (up to 5000 kg) have varied success rates across different booster versions, with many launches in total

Section 5

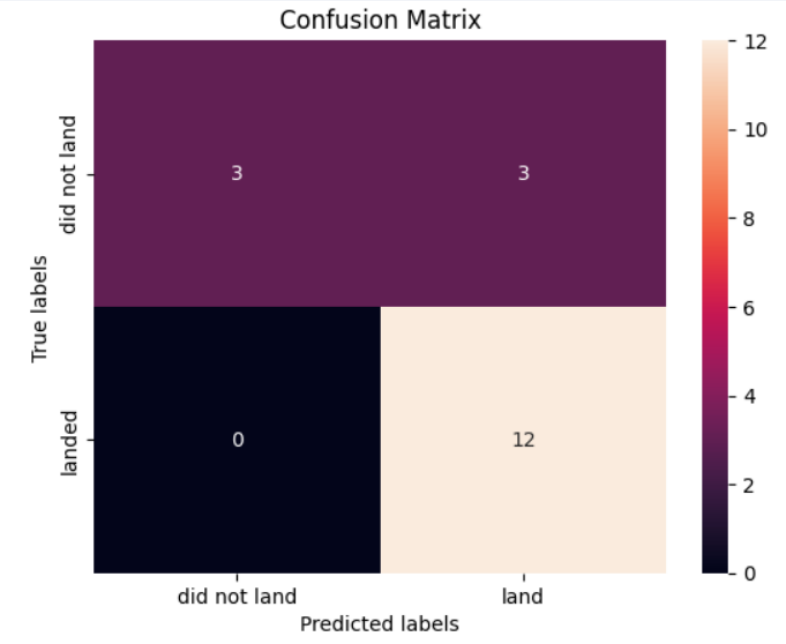
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, Support Vector Machine, and K Nearest Neighbors exhibit the highest classification accuracy
- The current test data set is too limited to conclusively identify the best-performing model
- A larger and more diverse data set is required for a more accurate assessment

Confusion Matrix



Explanation	
True Negatives	False Positives
False Negatives	True Positives

- The confusion matrix is identical for the best performing models: Logistic Regression, SVM, and KNN
- The models exhibit strong performance in identifying the ‘landed’ class (12 out of 12) and moderate accuracy in predicting the ‘did not land’ class (3 out of 6)
- There is an imbalance in misclassification, with all 3 misclassified instances occurring as false positives in the ‘did not land’ class and 0 false negative instances

Conclusions

- Data Collection and Methodology
  - Comprehensive data collection via SpaceX API and web scraping
  - Ensured data accuracy and completeness
- Exploratory Data Analysis (EDA)
  - Identified trends in payload mass, launch success rates, and launch site performance
  - Key insights gained through visualization techniques
- Predictive Modeling
  - Applied Logistic Regression, SVM, Decision Tree and KNN models
  - Similar performance across models; more data needed for conclusive results

Conclusions

- Impact of Variables
  - Payload mass, launch site, and orbit type significantly affect success rates
  - Higher success rates observed for certain orbits like Very Low Earth Orbit
- Temporal Trends
  - Marked improvement in landing success rates over time
  - Occasional dips (e.g., 2018) suggest areas for further investigation
- Launch Site Analysis
  - Effective use of multiple launch sites with no consistent outperformer
  - Coastal launch sites minimize risk and logistical challenges

Recommendations

- Expand dataset to improve predictive model accuracy
- Investigate causes of lower success periods (e.g., 2018)
- Refine feature engineering techniques
- Optimize future missions based on insights from launch site and orbit analyses

Appendix

- For more information and additional materials, please visit my GitHub repository
- For details on viewing the notebooks properly, please refer to the README file

Credits and Acknowledgements

- Primary Instructors
  - Joseph Santarcangelo
  - Yan Luo
- Other Contributors & Staff
  - Project Lead: Rav Ahuja
  - Instructional Designer: Lakshmi Holla
  - Lab Authors: Joseph Santarcangelo, Yan Luo, Azim Hirjani, Lakshmi Holla
  - Technical Advisor: Yan Luo
- Production Team
  - Publishing: Grace Barker, Rachael Jones
  - Project Coordinators: Kathleen Bergner
  - Narration: Bella West
  - Video Production: Simer Preet, Lauren Hall, Hunter Bay, Tanya Singh, Om Singh
- Teaching Assistants and Forum Moderators
  - Malika Singla
  - Duvvana Mrutyunjaya Naidu
  - Lakshmi Holla
  - Anita Verma

Thank you!